A Report from the University of Vermont Transportation Research Center

Analysis of Address-Based Employment and Demographic Data Sources For Travel Modeling and Transportation Planning

Final Report

TRC Report 18-001
July 2018

Analysis of Address-Based Employment and Demographic Data Sources For Travel Modeling and Transportation Planning

July 10, 2018

Prepared by:

James Sullivan, Research Projects Director

Transportation Research Center
Farrell Hall
210 Colchester Avenue
Burlington, VT 05405

Phone: (802) 656-1312
Website: www.uvm.edu/trc

Acknowledgements

The author would like to acknowledge VTrans for providing funding for this work.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the UVM Transportation Research Center. This report does not constitute a standard, specification, or regulation.

Table of Contents

List of Tables

Executive Summary

According to the responses received from the travel modeling community nationwide, address-based business and consumer data is overwhelmingly used for point-level employment counts from the business data. The only use of the consumer data identified was for providing a population "frame" and contact information for carrying out initial contacts for a household travel survey.

The use of DnB data appeared to be more common amongst the transportation planning community. This could indicate that DnB are the more "established" source in the field and InfoGroup is trying to penetrate that market, or that DnB is simply a more recognizable name.

For travel demand modeling, the business data was described as allowing the use of flexible geographic areas for sub-area modeling. Many users offered cautions about the need for excessive "cleaning" of the data after purchase, and the potential for "headquarters/branch" employment counts to have errors with their location.

The data quality assessment conducted consisted of a data reduction and an assessment of field error, resulting in the removal of many of the data fields delivered, particularly by InfoGroup. Many of the field provided were either missing, unexplained, or not relevant to transportation planning. Through the data reduction, a final set of fields that provide useful, valid, defined data was determined:

Summary of the Data Reduction

	InfoGroup				DnB
No. of Fields (Variables)...	2015 Consumer	2015 Business	2017 Consumer	2017 Business	2017 Business
Delivered	63	89	145	172	27
Blank, Unexplained, Undefined, or Unavailable	8	2	72	42	0
Related to Geo-coded Location	32	27	36	62	10
Not Relevant to Transportation Planning	8	18	11	20	0

No. of Fields (Variables)...	InfoGroup				DnB
	2015 Consumer	2015 Business	2017 Consumer	2017 Business	2017 Business
Redundant or Unacceptable Quality	6	33	11	22	8
Final Set	**9**	**9**	**15**	**26**	**9**

The final set of fields are:

Final Set of Data Fields

IG 2015 Consumer		
Household Income	Year Home was Built	Marital Status
Household Wealth	Age of Head of Household	Primary Family at the Address?
Household Purchasing Power	Length of Current Residence	Vacancy?
IG 2015 Business		
Headquarters, Branch, or Sole Location	Location Name	Secondary SIC Codes
Company Name	Parent Company Employee Size	Square Footage
Location Employee Size – Modeled and Range	Primary SIC/NAICS Code	Year 1st Appeared
IG 2017 Consumer		
Adult Age Range	Home Age	Mean Years of Schooling
Delivery Unit Size	Home Equity Estimate	Number of Trade (Credit) Lines
Early Internet Adopter?	Household Income	Residence Ownership
Expendable Income	Loan-to-Value Ratio	Residence Type
Heavy Internet User?	Marital Status	Household Wealth
IG 2017 Business		
Corporate Employment Size	Foreign Parent Company?	Public Company?
Corporate Sales Volume	High-Income Executives?	Bankruptcy Filing?
Credit Score	High-Tech Business?	Secondary SIC
Location Employment Size	Headquarters or Branch	Small Business?
Affluent Neighborhood?	Import/Export Activity?	Square Footage
Asset Size	Individual or Firm	White Collar?
Big Business	Medium Size Business?	White Collar Percentage
Female Owner Executive?	Modeled Employment Size	Work At Home Business?

Fleet Size	NAICS Code	
DnB 2017 Business		
Business Name	Small Business?	Primary NAICS Code
Location Employee Size	Manufacturing?	Sales Volume
Year Started	Primary SIC	3-Year & 5-Year Growth in Employment and Sales Volume

Subsequent requests for data should only include these variables. Other consumer variables of interest for transportation modeling and planning include the number of household vehicles, the household size and composition (number of children, other adults, etc.), student status of children, and worker status of adults. Other business variables of interest for transportation modeling and planning include fleet size, fleet type (vehicle size), and shipment information (incoming/outgoing weight, frequency, mode, vehicle size, etc.).

The geo-coding quality of the data varied considerably. For the Dun & Bradstreet data, the tolerances of the geo-coding quality compromise its use for detailed spatial analysis. Correcting over 20% of the geo-coded locations is not feasible. Geo-coding of the 2017 InfoGroup data is considerably better, with the consumer data in particular achieving a high rate of matching to the Parcel level. However, the 2015 InfoGroup data does not achieve nearly the same level of quality. It is unclear if all data before 2017 will be compromised in the same way, or if geo-coding of any data that is not "current" loses quality. In either case, subsequent requests for data should stipulate geo-coding quality that meets the following standards:

- 80% or more of the geo-coded locations from each data set (measured independently) matched to the PARCEL
- 90% or more of the geo-coded locations from each data set (measured independently) within 0.31 miles of an associated point in the E911 point shapefile

Providing the current E911 point shapefile may enhance the vendor's ability to geocode and comply with the second standard.

1 Introduction

Marketing support firms compile consumer and business data to provide business-to-business and business-to-consumer products. These firms sell the most up-to-date data to companies for marketing campaigns, through mailing, emailing, and other solicitations to potential customers. Two of the largest and most established providers of this type of data for the private sector, InfoGroup and Dun & Bradstreet, offer address-based employment and consumer (demographic) data to transportation planning agencies, which can supplement traditional sources of employment and demographc data like the U.S Census and the Quarterly Census of Employment from a state Department of Labor. These new data sources offer significant opportunities for travel modeling and transport analysis for the Vermont Agency of Transportation (VTrans) for the following types of analyses:

- Economic growth/impacts modeling and calculation
- Disaggregate travel modeling calibration and validation
- Accessibility calculations
- Vulnerable populations identification

However, the fact that there are only two known providers for this type of high-resolution data and that they do not have a long history of supporting transportation agencies presents some risk for the Agency. The goal of this project was to reduce that risk by assessing samples of the data and gathering information on its uses from the experience of others. The purpose of this project was to obtain samples of the data being offered from both vendors, to conduct an evaluation of its quality and accuracy for travel modeling and transportation analysis, and to solicit other users in the travel/transport modeling community for experiences with this type of data.

The consumer data includes household-level and person-level demographic and economic information about individuals aged 18 years and older, from a variety of sources such as public tax records, credit reporting agencies, credit card transaction data, and internet purchasing indicators. The business data includes location-based information about corporations' physical and economic size including number of employees, sales volume, assets, building size, and industry for the specific location and for its parent company, if any. For both the consumer and business data, perhaps the most critical feature is the address-based location that is provided for every record in the data sets.

This report was prepared under the Project Assignment / Work Authorization No. 004 under EA No. 0001055-302. The tasks included in this work are:

1. Representative sample of data
2. Summarize vendor data
3. Data assessment
4. Determine data fields for planning
5. Final report & presentation

Section 2 of this report provides testimonial evidence from practitioners and other planning agencies of the use of address-based marketing data for travel modeling and transportation planning. Section 3 provides a detailed description of the 5% samples of the data that were obtained for review, and Section 4 contains the detailed review of the data. Section 5 contains a comprehensive review of the geographic quality of the address-based data, and Section 6 contains the conclusions and recommendations.

2 Testimony from Actual Users

The first step in the project was to determine how others in the U.S have used the type of address-based data being investigated here. In order to find out how other agencies might be using this type of data, a solicitation was distributed to the listserv of the FHWA Travel Model Improvement Program.

2.1 Testimony from Users

The following are the actual responses received from the TMIP solicitation:

Mike Aronson, P.E. | Principal Engineer, Transportation Engineering / Planning, Kittelson & Associates, Inc.

I have used the InfoUSA data as my primary source for non-residential land use inputs for many models. It requires some manual effort for the following:

- *It tends to fall short on government employment, and reports employees at the contact site rather than the individual employment sites. Therefore, we always contact each government agency and school district by e-mail/phone to get addresses and numbers of employees for each actual building location. We then add that to the InfoUSA database and replace some of the aggregate entries.*
- *To a lesser extent, the same issue comes up with franchise operations such as fast food – the employment may all be reported at a management office rather than at individual restaurant sites.*
- *Sometimes the number of employees reported is for the entire company rather than the individual site. We manually screen the largest employee totals and make sure they make sense.*
- *We check control totals by employee type (retail, education, manufacturing, etc…) against state employment reports. This often helps us to identify problems or omissions in the database, or correct business type classifications.*
- *We keep the records at the address level as long as possible in the process, so that they can be geocoded to any scale of transportation analysis zone system. I have not gone through the process of trying to relate the addresses to parcels, but I think others have. It gets complicated to develop any kind of one-to-one correspondence because a shopping center and its parking lots may cover multiple parcels, but the InfoUSA will have multiple records to represent each business in the shopping center.*

- *The additional detail on business type in the business database, compared to a source such as LEHD, allows for more detailed attraction trip generation methodologies (for example, to distinguish between high-generating and standard retail uses).*

Jami Dennis, GISP | Senior Information Services Project Manager, Maricopa Association of Governments, Phoenix, Arizona

Here in the Phoenix metro area at the Maricopa Association of Governments, we purchase Dun & Bradstreet data every year for use in our Employer Database. This is a database of all employers in the region. The D&B data is used to augment our existing database that includes employer data from our annual Trip Reduction Program survey (mandated for air quality purposes from all employers with 50 or more employees at one location). The Employer database is used in our socioeconomic model which is also an input to our travel demand model. The employer database has also been used extensively in economic development projects as well.

Every 5 years we evaluate different sources for the data and that has included the InfoGroup data. We have found D&B to be the best for our use – though it is still not perfect and we spend a lot of time cleaning the data. You can see our employer data in one of our 2 employment map viewers, one for our region and one statewide - on our website: http://maps.azmag.gov/*. We recently expanded to a statewide database, in collaboration with the other COGs and MPOs in the state.*

Ben Gruswitz, AICP | Senior Planner, Office of Long-Range Planning, Delaware Valley Regional Planning Commission, Philadelphia, Pennsylvania

A forthcoming NCHRP report may be of interest to you and others: NCHRP 08-36/Task 127 "Employment Data for Planning: A Resource Guide" (not completed as of Paril 18, 2018).

Many MPOs purchase one of the products you mentioned to use as base-year employment in their employment forecasting process. Purchasing 2 sources is a plus, if you've got the funds to do so. With time and resources, you can figure out if one is more complete/accurate, do some reasonableness checks, and see if one is better suited for your purposes. Warning: attempting to "clean" these datasets will induce many headaches. It's very easy to get lost in the weeds and at some point you just have to declare you did what you could and move on.

We use a derivative of Dun & Bradstreet data called the National Establishment Time-Series (NETS) Database which is supposed to be further "cleaned" by Walls & Associates and then we go and clean it some more. We used to use the 2000 CTPP from the long-form decennial census, which had employment data aggregated to TAZ geographies. By purchasing these proprietary, point-level datasets we've been able to aggregate to whatever geography we want, and after

our forecasting process is over, we have a GIS resource that planners continue to use for various local studies and reports on the state of various industries in the region.

We've flirted with the idea of exploring alternative sources than NETS but purchasing and evaluating these products takes a lot of time and resources. I believe most people will tell you that any of these sources have their flaws and it's hard to invest in picking a clear winner when it's unclear if there will be a winner at all. I am, however, looking forward to seeing the NCHRP report I mentioned to see if any further clarity is gained from a focused examination of these sources.

Josie Kressner, Ph.D., Transport Foundry, LLC

A synthetic population with synthetic travel diaries was built from two different types of third-party "big" data - consumer marketing data and passive location data from mobile phones – for the four-county planning region of the Puget Sound Regional Council in metropolitan Seattle, Washington. Refer to the IDEA Program Final Report for NCHRP-184: Synthetic Household Travel Data Using Consumer and Mobile Phone Data for more details.

Aditya Katragadda | Transportation Planner (Modeling), The Corradino Group

Basically this kind of data is used to conduct household travel surveys. It helps in designing the sampling plan to target households by socio-economic characteristics and geographic location (usually referred as strata/bin).

Tom Worker-Braddock, AICP, PTP, LCI | Multimodal Transportation, Olsson Associates

We purchased data a few times several years ago for travel modeling purposes, and discovered it was quite good, and relatively inexpensive. Something like $900 for an entire state worth of how many employees work in block-group level data and where they live. It was originally the Nielsen company, but another company bought them.

Aichong Sun | Transportation Modeling Manager, Pima Association of Governments, Tucson, Arizona

We've had the opportunities to work on both Dun & Bradstreet and InfoUSA business listing data to develop and update our own employment database at PAG. Working with either of them is by no means a trivial undertaking, but it has become a lot easier with a quite sophisticated procedure that we developed to help us walk through the process which initially had been several months long.

Jill Hough | Principal Project Manager, CHS Consulting Group

In my experience working with Dun & Bradstreet data, among the largest of issues to reconcile is where employees within a zone of interest to you are employed by a company whose

corporate headquarters is located elsewhere, in which case those employees don't show up in the local dataset. Incidentally, the reverse was also true: if a company HQ happened to be located locally, the total number of employees was ascribed to the local zone, even if only half or some fraction were employed locally. I'm going back some years and don't know if D & B have improved their data to reconcile these issues. To the extent they have not, it does seem necessary as others have reported, to employ a secondary data set that doesn't have such shortcomings.

Jami Dennis, GISP | Senior Information Services Project Manager, Maricopa Association of Governments, Phoenix, Arizona

We use Dun & Bradstreet's Hoovers data for our employer data, though we augment it with data from other sources and do spend a lot of time cleaning the data each year. The D&B data that we purchase does have employment by location, not just aggregated at the HQ. You just need to query on Single Location rather than HQ. One of the bigger issues (same with InfoGroup) is that closed businesses do not get removed very quickly from the database, often staying in the data for over a year. We do provide D&B with input on closed and changed businesses each year, but that doesn't always make it back in to their database either. At any rate, we have been fairly happy with the D&B data and every 5 years we review other data sources, so far D&B continues to be the better choice for us.

Timothy G. Reardon | Director of Data Services, Metropolitan Area Planning Council, Boston, Massachusetts

We acquired InfoGroup point-level establishment listings in 2011 and again in 2016. While the vendor does claim to have site-level employment estimates (not all at the HQ), the data do benefit from manual efforts to disaggregate headquarters and distribute employment from major institutions (esp universities and hospitals, of which we have quite a few.) Also keep in mind that in many cases the job counts are estimated based on some combination of NAICS code and square footage-it's not all from surveys or administrative records.

Our partner agency for transportation modeling did use the data to establish base-year job counts, and we've used it for estimation of our land use allocation model. However, we've gotten an equal amount of mileage using the data for local planning projects, as a way to understand the local business mix, track change over time (we are testing out measures of commercial gentrification) and contact business owners for engagement purposes [see Attachment A]. ESRI business analyst lets you do some of those things as well, but it's useful to have the comprehensive dataset.

3 Sample Data Obtained for Review

For both vendors, a 5% random sample of all address-based data available for Vermont was requested for the year 2015. InfoGroup offered to provide 2015 data and 2017 (current) data for both businesses and consumers, whereas Dun & Bradstreet offered to provide data for businesses only. The files received with the InfoGroup order are detailed in Table 1.

Table 1 Files Received from InfoGroup

File name(s)	Description
2015 Business Data	
Historical Data-Business P2.pdf	An InfoGroup marketing publication that includes the field names and descriptions for its historical business data
Decode for Field-[Field Name].txt (21 files)	Text descriptions of the valid values and their meanings for 21 of the 89 fields included in the data
Format.txt	A text description of the data contained in the order, including the number of records, each field's character length and each full field name
Order_625150.xlsx	2,400 business data records with 89 fields
PackingSlip.rtf	
Report-Counts.txt	
2015 Consumer Data	
Historical Data-Residential P2.pdf	An InfoGroup marketing publication that includes the field names and descriptions for its historical residential data
VT-30k.csv	30,000 residential data records with 66 fields (no field names)
2017 Business Data	
ver01001.txt	2,400 business data records with 172 fields
business cass.doc	"Coding Accuracy Support System, Summary Report" for Zip+4 coding
business ncoa.xls	Results of NCOA Link Processing

us bus gui field decode.doc	Text descriptions of the valid values and their meanings for 34 of the 172 fields included in the data
ver01lay.txt	A text description of the data contained in the order, including the number of records, each field's character length, and each full field name
2017 Consumer Data	
q2023901.txt	
ver01001.txt	30,000 residential data records with 145 fields
consumer 2000 field decode.doc	Text descriptions of the valid values and their meanings for 35 of the 145 fields included in the data
consumer cass.doc	"Coding Accuracy Support System, Summary Report" for Zip+4 coding
consumer ncoa.xls	Results of NCOA Link Processing
ver01lay.txt	A text description of the data contained in the order, including the number of records, each field's character length, and each full field name

The files received with the Dun & Bradstreet order are detailed in Table 2.

Table 2 Files Received from Dun & Bradstreet

File name	Description
Sample File Order For 1167321	2,000 business data records with 27 fields
Vermont Sample File Data Dictionary	Field names and descriptions for all 27 fields included in the data

The 2017 Business Delivery from InfoGroup contained the following disclaimers and warnings:

- CREDIT DISCLAIMER: Our Business Credit Score Codes are indicators of probable ability to pay. They are based on business demographic factors such as number of employees, years in business, industry stability, barriers to entry, and government data. We

recommend that these ratings be used primarily as a starting point and should not be the sole factor used in making a credit decision. You must obtain more information from bank and trade references, local credit bureaus, or other sources before extending credit. We are not a financial advisor and make no representations or warranties as to the accuracy, timeliness or completeness of the rating codes, and as such will not be responsible for any losses resulting from the use of this information. Furthermore, our liability, if any, will be limited to the initial cost of the credit rating fee paid by the purchaser.

- NOTICE TO ALL USERS OF FACSIMILE INFORMATION: It is a violation of both federal and state law to transmit an unsolicited advertisement to a facsimile machine. Any person violating such laws may be subject to civil and criminal penalties which may exceed $500 for each transmission of any unsolicited facsimile. We provide our business information for lawful purposes only and expressly forbid the use of our business information in any unlawful manner.
- WARNING!! DO NOT USE THIS INFORMATION AFTER 6 MONTHS FROM PRODUCTION DATA: Our Business Database changes by over 70% in just one year. New companies start up, others go out of business, and many move or change their phone number. And key executive names can change even faster. Using this product after the Expiration Date may result in wasted time and effort, since much of the information will be out of date. Please call us for an updated product.

InfoGroup claims to have over 100 different contributing sources to its data, and to make 100,000 calls a day to verify the accuracy of the data. Their focus seems to be on having the most accurate current data, with the target sectors being marketing, search-engine optimization and in-car navigation. They are careful to point out, however, that they do NOT undertake “online tracking”. Among the obvious sources for this data are the yellow pages, Claritas, the U.S. Postal Service, and the US Census. Both data sets contained very specific contact information for an individual adult in the household (for consumer data) and an individual contact (for business data). Since contact information is not needed for travel modeling, these fields were removed, to eliminate the possibility of identifying any individuals in the data set.

4 Data Quality Assessment

For the data quality assessment, key fields were identified whenever possible to facilitate review. Attempts were made to identify the definition and validation (acceptable responses) of each data field, by matching field names in the data table to variable names in the data dictionaries. For the Dun & Bradstreet data, the correspondence between field names in the data table and variable names in the data dictionary was 100%, but validation information was lacking. For the InfoGroup data, several data dictionaries were provided with both definitions and validation information, many of the field names in the data tables were lacking a corresponding definition in the data dictionary, making it difficult to review those variables.

Blank cells presented a problem in the review of both data sets. Blank cells are assumed to indicate that the value is either (1) unknown, (2) not available for this record, or (3) not available for this data set. Very few of the dictionary definitions for the variables in these data sets clarified the meanings of blank cells, making the use of fields with many blank cells infeasible. Therefore, field errors were measured as the fraction of unexplained blank responses to the total number of opportunities for a response.

4.1 InfoGroup 2017 Consumer Data

4.1.1 Data Reduction / Cleaning

FAMILYID was confirmed as a key field. 19 fields called or containing the word “Filler” were removed from the data set. The following 28 fields contain no data (100% blank responses):

- Batch_Number
- Buyer Behavior Code
- Container_Number
- CSA_Code
- Entry_Point_Number
- IMB Barcode
- Key_Code
- Line_of_Travel_Code
- Marriage_Date
- Mortgage Loan Amount
- Mortgage_Finance_Type
- Mortgage_Loan_Amount
- MSA_Code
- Net_Worth_Rank_Code
- Number_Mortgages
- Package_Number
- Pallet_Number
- Presort_Endorsement_Line
- Presort_Package_Destination
- Presort_Pass_Code
- Presort_Pricing_Tier
- Prizm Code
- Prizm Description
- Religion Of Household
- Sequence_Number
- Sort_Sequence_Control

- MSA_Desc
- Title_Address

The significance of the variables in the sample whose entries are entirely blank is unclear. These variables could be infrequent enough that a 5% sample might not have a non-blank response or they could have been excluded from the data set but inadvertently included in the data delivery. Therefore, these fields were also removed from the data set.

The definitions of the following 25 fields could not be located in the field decode sheet or in a subsequently delivered "External US Consumer Data Dictionary":

- Batch_Number
- Buyer Behavior Code
- Container_Number
- Entry_Point_Number
- High_Tech_Indicator
- IMB Barcode
- Key_Code
- Line_of_Travel_Code
- MSA_Desc
- Net_Worth_Rank_Code
- Number_Mortgages
- Package_Number
- Pallet_Number
- Population_Code
- Presort_Endorsement_Line
- Presort_Package_Destination
- Presort_Pass_Code
- Presort_Pricing_Tier
- Prizm Code
- Prizm Description
- Nielsen_Region_Code
- Sequence_Number
- Sort_Sequence_Control
- Sub_HH_Indicator
- Title_Address

Residence Ownership, Residence Type, Occupation, and Vehicle Manufacturer, were not included in the data delivery but were listed in the data dictionary. The field labeled "Own_Rent" seems to correspond to the coding for Residence Ownership. The field labeled "House_or_Apartment" seems to correspond to the coding for Residence Type. It is unclear if the other fields were intended for delivery, or if they were intentionally omitted. "High_Tech_Indicator" is assumed to correspond to "DM High Tech Household" in the data dictionary.

Eight additional fields defined in the "External US Consumer Data Dictionary" (delivered later) were added to the dictionary of selected fields provided in the "consumer 2000 field decode.doc" file. The 36 following variables are directly related to the geographic location of the record or the need to contact:

- County_Name
- CBSA_Code
- Census_Block_Group
- Census_Tract
- Pallet_Number
- Phone
- Carrier Route Code
- Population_Code

- City
- County_Code
- Address
- Contact_Name
- CSA_Code
- Delivery_Point_Bar_Code
- Delivery_Service_Type_Code
- Congressional_District
- Metro/Micro Indicator
- MSA_Code
- MSA_Desc
- Nielsen Population Area
- Nielsen_Region_Code
- Package_Number
- Presort_Endorsement_Line
- Presort_Package_Destination
- Presort_Pass_Code
- Presort_Pricing_Tier
- Prizm Code
- Prizm Description
- Sequence_Number
- Sort_Sequence_Control
- State
- State_Code
- Title_Address
- ZIP_Code
- ZIP10
- ZIP4

The geographic location quality is reviewed in Section 5 using the Match Level Code variable, so these individual geographic variables are not reviewed separately.

4.1.2 Field Error

The reduction leaves 37 variables with defined valid data. 26 of these variables have relevance to transportation planning. These fields are listed, along with their field error, in Table 3. Field errors were measured as the fraction of blank or undefined responses to the total number of opportunities for a response.

Table 3 Field Errors in the InfoGroup2017 Consumer Data

Variable	Description	Field Error
Adult Age Range	Age	0%
Delivery Unit Size	Used to indicate single family and multifamily delivery. For a given street and house number address, families at the address are counted. A typical scheme for carrying delivery size uses 1-9 to represent dwelling unit size 1-9. A delivery unit size of 10 means 10 to 19. A delivery unit size of 20 means 20-29 etc.	0%

Variable	Description	Field Error
Early Internet Adopter	Score of 1 to 10, where "1" indicates most likely and "10" indicates least likely an early adopter of the Internet	0% (modeled)
Expendable Income / Net Worth	This DBA "model" predicts household monthly expendable income (FIND/12) by subtracting out the monthly mortgage payment (reported or inferred) for homeowners, or estimated rent value for non-homeowners. The result is compared against 14 predefined $ amounts that represent a continuum of 15 ranks, from top to bottom.	0% (modeled)
Female_Occupation_ Code	Denotes occupation of a female present in the household. Code will be replaced if pertinent information becomes available from any source, else previous occupation will be retained.	84%
Gender		3.4%
Heavy Internet User	Score of 1 to 10, where "1" indicates most likely and "10" indicates least likely to be a heavy Internet user	0% (modeled)
High Tech Household	Interest (early adopters) in new, cutting edge products gathered from product purchases, subscriptions or survey response and blended with modeled data.	47%
Home Age		32%, corresponds to Residence Ownership of 0 (unknown) or 1 (rent)

Variable	Description	Field Error
Home Equity Estimate		32%, corresponds to Residence Ownership of 0 (unknown) or 1 (rent)
Home Sale Price		74%
Home Value		22%
Household Income		0%
Language Spoken in Household		100%
Loan-to-Value Ratio		32%; corresponds to Residence Ownership of 0 (unknown) or 1 (rent)
Location Type		12% blanks and "T" response not in codebook
Lot_Size	Number of acres associated with property address (nn.nnn) Maximum value is 30.000 acres.	45% blank or "0"
Male_Occupation_Code	Denotes occupation of a male present in the household. Code will be replaced if pertinent information becomes available from any source, else previous occupation will be retained.	91%
Marital Status		0%

Variable	Description	Field Error
Mean Years of Schooling		0.1%
Number_Of_Trade_Lines	Count of trade lines associated with this household. This in no way reflects the credit worthiness of this household.	0%
Occupation		Not provided
Political Party Affiliation		94%
Religion Of Household		100%
Residence Ownership		0%
Residence Type		0%
Vehicle Manufacturer		Not provided
Wealth Finder	Composite variable using a group of other consumer and demographic variables intended to measure wealth of household. The result is a single character ranging from A through T (20 ranks), with the top rank (A) representing households with the highest estimated wealth	0%

4.2 InfoGroup 2017 Business Data

4.2.1 Data Reduction / Cleaning

INFOUSA ID was confirmed as a key field. 7 fields with "Filler" in the title were removed, along with "End of Record Marker", "Source", "Production Date" and "Obsolescence Date", leaving a total of 145 fields in the data. The following 23 fields contain no data (100% blank responses):

- Actual_Corporate_Sales_Volume
- Corporate_Sales_Volume_Code
- Corporate_Sales_Volume_Desc
- Fax_Number
- CSA_Code
- CSA_Descr
- Fortune_Ranking
- IMB Barcode
- Key_Code
- MSA_Code
- MSA_Desc
- Presort Line Of Travel
- Presort_Bag_Number
- Presort_Bundle_Number
- Presort_Endorsement_Line
- Radial_Distance_From_Target_Element
- Secondary SIC Description
- Stock_Exchange_Code
- Stock_Exchange_Desc
- Stock_Ticker_Symbol
- Tertiary_Carrier_Route_Code
- Tertiary_ZIP4
- Title_Address

The significance of the variables in the sample whose entries are entirely blank is unclear. These variables could be infrequent enough that a 5% sample might not have a non-blank response or they could have been excluded from the data set but inadvertently included in the data delivery. Therefore, these fields were also removed from the data set.

The definitions of the following 12 fields could not be located in the field decode sheet or in a subsequently delivered "External US Consumer Data Dictionary":

- Asset_Size
- BookNO
- IMB Barcode
- Presort Line Of Travel
- Presort_Bag_Number
- Presort_Bundle_Number
- Presort_Endorsement_Line
- Radial_Distance_From_Target_Element
- Selected_SIC_Code
- Selected_SIC_Desc
- Sequence_Number
- Title_Address

Many field names in the data do not match the field names provided in the dictionary files. Therefore, the matching dictionary definition had to be assumed from the responses available.

The following 62 variables are directly related to the geographic location of the record or the need to contact:

- CBSA_Code
- CBSA_Descr
- Census_Block_Group
- Census_Tract
- Contact Ethnic Code
- Contact Ethnic Description
- Contact_Name
- County_Code
- County_Name
- CSA_Code
- CSA_Descr
- Delivery_Point_Bar_Code
- Fax_Number
- Key_Code
- Last_Name
- Latitude
- Longitude
- Metro_Micro_Indicator
- MSA_Code
- MSA_Desc
- Phone
- Population_Code
- Population_Desc
- Presort Line Of Travel
- Presort_Bag_Number
- Presort_Bundle_Number
- Presort_Endorsement_Line
- Primary_Address
- Primary_Carrier_Route_Code
- Primary_City
- Primary_SIC_Code
- Primary_SIC_Desc
- Primary_State
- Primary_State_Code
- Primary_Zip_Code
- Primary_ZIP10
- Primary_ZIP4
- Secondary_Address
- Secondary_Carrier_Route_Code
- Secondary_City
- Secondary_SIC_Code_1
- Secondary_SIC_Code_2
- Secondary_SIC_Code_3
- Secondary_SIC_Code_4
- Secondary_SIC_Desc_1
- Secondary_SIC_Desc_2
- Secondary_SIC_Desc_3
- Secondary_State
- Secondary_State_Code
- Secondary_Zip_Code
- Secondary_ZIP10
- Secondary_ZIP4
- Tertiary_Address
- Tertiary_Carrier_Route_Code
- Tertiary_City
- Tertiary_State
- Tertiary_Zip_Code
- Tertiary_ZIP10
- Tertiary_ZIP4
- Title_Address
- Toll_Free_Number
- Yellow_Page_Code

The geographic location quality is reviewed in Section 5 using the Match Level Code variable, so these individual geographic variables are not reviewed separately.

4.2.2 Field Error

The reduction left 96 variables with defined, valid data. 48 of these variables have relevance to transportation planning and their data quality is described

in Table 4. Field errors were measured as the fraction of blank or undefined responses to the total number of opportunities for a records.

Table 4 Field Errors in the InfoGroup 2017 Business Data

Variable	Description	Field Error
ACTNUMBUS Multitenant Location	None, but appears to align with MTACTN (Multi-Tenant Number of Tenants); values range from 1 to 602	43%
Actual Corporate Employment Size	None, but appears to align with PACTEM (Corporate Employee Size)	99%
Actual Corporate Sales Volume	None, but appears to align with PACTSL (Corporate Sales Volume)	99%
Actual Credit Score	None, but aligns with CRDTSC (Business Credit Score)	0%
Actual Location Employment Size	None	3%
Actual Location Sales Volume	None, but aligns with SLSVDT (Sales Volume)	20%
Affluent Neighborhood Location	None, but aligns with WEALTH (Wealth Code)	0%
Asset Size	None	0%
Big Business	Indicates big business segment (yes/no)	0%
Building Num Multi Tenant	None, but appears to align with MTBLDN (Multi-Tenant Building Number)	43%
Business Size Change	Growing (+) or shrinking (-) business	98%
Corporate Employment Size	None, but appears consistent with PEMPSZ (Corporate Employee Size Code)	99%
Corporate Sales Volume		100%

Variable	Description	Field Error
Credit Score	Business credit score range	0%
Female Owner Exec	FEXOWN (Female Executive-Owner)	0%
Fleet	Modeled fleet size range	0% (blanks defined)
FORGNPAR	Foreign/international parent code	0% (blanks defined)
Fortune Ranking		100%
Franchise Specialty 1…6	Numeric franchise SIC codes to identify franchise/brand affiliation	88% to 99%
Gender	Male/female	27%
GOVSEGCD	Government segment code – Federal, State, County, Municipal	93%
HighIncomeExec	The primary contact is inferred to be a high income executive (yes/no)	0%
HighTechBusiness	The business is in the high tech segment (yes/no)	0%
HQ Branch	Denotes whether the business is a headquarter (1), a branch (2), or a subsidiary headquarter (3), or a single location (9).	0%
IMB Bar	None	100%
IMPEXPCD	Import/export code denotes whether business provides export services (E), import services (I), or both (B)	99%
Individual/Firm	Indicates if location is an individual (1) or a firm/business (2)	0%

Variable	Description	Field Error
Industry Specific	Industry-specific codes, classifying the location according to parameters specific to its industry	95%
Location Employment Size	Range of reported or modeled number of employees working at this location	3%
Location Sales Volume	Range of reported or modeled sales volume at this location	20%
Medium Size Business Entrepreneur	None. Could be MBUSIN (medium-sized business indicator, yes/no)?	0%
Modeled Employment Size	Determines how employment size was determined – actual (A), modeled by business name (B), modeled by SIC code (C), or modeled by calculated sales volume (D)	3%
Multitenant	Identifies the number of tenants at the location - 2 to 4 (A), 5 to 9 (B), and 10+ (C)	53%
NAICS	NAICS code of business	0%
Office Size	Number of professionals working at the location – 1 (A), 2 (B), 3 (C), 4 (D), 5-9 (E), and 10 or more (F)	85%
Own Lease	None, but looks like OWNFLG, which indicates whether the business owns (O) or leases (L) its premises, or if the status is unknown (U)	94%
Public Company Indicator	Indicates if the business is a private company (0), a public company (1), or a branch of a public company (2)	0%
Public Filing Indicator	Looks like PUBFLG, indicating that a bankruptcy filing is available in Public Record Data (yes/no)	0%

Variable	Description	Field Error
Secondary SIC 1…3	Identifies up to 4 additional activities of the business. Approximately 23% of the records in the database have 2 activities, 8% have 3, 3% have 4, and 1% have 5 or more	59% to 96%
Site Number	If Infogroup Number and Site Number are the same, then the record is the primary business at the site. If they are different, then the record is a secondary business at the site.	0%
Small Business Entrepreneur	None. Could be SBUSIN (small-sized business indicator, yes/no)?	0%
Square Footage	The modeled area range of the firm's location, in square feet, using the 2015 model.	0%
True Franchise	Franchise code for the primary business or an additional line of business is a true franchise	97%
Web Site	The primary web address (URL) of the business	38%
White Collar Indicator	Indicates if 50% or More Employees are White Collar (1) or Less than 50% of Employees are White Collar (0)	0%
White Collar Percentage	Why do we need the indicator if we have the exact percentage?	0%
Work At Home Business	Indicates if this is a "Work at Home" business (blanks are defined as not a "Work at Home" business)	0%
Year Established	Year business was established	77%

4.3 InfoGroup 2015 Consumer Data

4.3.1 Data Reduction / Cleaning

There are a total of 63 fields in the data set. FAMILYID was confirmed as a key field – LOCATIONID is blank for many of the records. The following 6 fields contain no data:

- Bathroom_Cnt,
- Bedroom_Cnt,
- Construction_Type_Code,
- CSACode,
- Room_Cnt,
- Tele_Restricted_Ind

The significance of the variables in the sample whose entries are entirely blank is unclear. These variables could be infrequent enough that a 5% sample might not have a non-blank response or they could have been excluded from the data set but inadvertently included in the data delivery. Therefore, these fields were also removed from the data set.

Two other fields not available for 2015 were removed. The following 32 variables are directly related to the geographic location of the record or the need to contact:

- Addresstype
- Box_Type
- CBSAtype
- Census2010CountyCode
- CensusBlockGroup
- CensusStateCode
- City
- DPBC
- House_Num_Fraction
- Latitude
- Route_Num
- State
- Street_Post_Dir
- Street_Suffix
- Unit_Type
- ZIP
- Box_Num
- CBSACode
- Census2010Block
- Census2010Tract
- CensusCountyCode
- CensusTract
- CSACode
- House_Num
- HouseholdStatus
- Longitude
- Route_Type
- Street_Name
- Street_Pre_Dir
- Unit_Num
- USPSNoStats
- ZIP4

The geographic location quality is reviewed in Section 5 using the MatchLevel variable, so these individual geographic variables are not reviewed separately.

4.3.2 Field Error

The reduction left 23 variables with defined, valid data. 15 of these variables have relevance to transportation planning and their data quality is described in Table 5. Field errors were measured as the fraction of blank or undefined responses to the total number of opportunities for a records.

Table 5 Field Errors in the InfoGroup 2015 Consumer Data

Variable	Description	Field Error
Estmtd_Home_Val_Div_1000*	Estimated home value. When more than one household (including non-fulfillment records) is at the same address (as defined by LocationID), the best home value is chosen and stored for all of them.	39% of the records are $0, and many of these are indicated as being owned in "Owner Renter Status"
Find_Div_1000*	FIND is a prediction of HH income.	0%
Wealth_Finder_Score*	Modeled prediction of household wealth	0%
PPI_Div_1000*	Estimate of relative purchasing power of a HH, derived by adjusting FIND with the appropriate cost of living index for the county in which the HH resides.	0%
Building_Area	Square footage of dwelling.	99% (only 4 valid responses)
Built_Year	Year (yyyy) dwelling built.	Valid values only for 51% of likely and confirmed home owners
Children_Ind	Indicates children are present in HH.	Only 10% of the responses are indicated as having children in the household? ACS indicates 23%. 572 of those indicate the CHILDRENHHCOUN
ChildrenHHCount	Number of HH members determined to be children	

Variable	Description	Field Error
		T = 0, which would bring the households with children down to 8%.
Head_HH_Age_Code	Age of head of household	0%
Length_Of_Residence	The difference (in months) between arrival date at a residence and current (system) date, converted to number of years. Range is limited to current year minus 1959.	0%
Marital_Status	Score indicating likelihood Head of HH is married.	0%
Owner_Renter_Status	Score indicating likelihood that the HH either owns their home or is renting.	Only 4% renters? ACS indicates 29%
Primary_Family_Ind	Indicates this record is considered to be the primary family at this address.	0%
Tradeline_Count	Indicates number retail credit lines and/or bank/oil company credit cards linked to HH	88% have 0 credit lines?
Vacant	Indicates vacant	0%

4.4 InfoGroup 2015 Business Data

4.4.1 Data Reduction / Cleaning

There are a total of 89 fields in the data set. ABI was confirmed as a key field. Two variables were removed because they are not available for 2015. None of the variables are entirely blank. The following 27 variables are directly related to the geographic location of the record or the need to contact:

- Address Line 1
- Address Type Indicator
- CBSA Code
- CBSA Level
- Census Block
- Census Tract
- City
- County Code
- CSA Code
- FIBS Code
- Landmark Address
- Landmark City
- Landmark State
- Landmark Zip4
- Landmark Zip Code
- Latitude
- Longitude
- Mailing Address Flag
- Mailing Address
- Mailing City
- Mailing State
- Mailing ZIP4
- State
- Unit Number
- Unit Type
- ZIP4
- Zipcode

The geographic location quality is reviewed in Section 5 using the Match Code variable, so these individual geographic variables are not reviewed separately.

4.4.2 Field Error

The reduction left 60 variables with defined, valid data. 42 of these variables have relevance to transportation planning and their data quality is described in Table 6. Field errors were measured as the fraction of blank or undefined responses to the total number of opportunities for a records.

Table 6 Field Errors in the InfoGroup 2015 Business Data

Variable	Description	Field Error
Business Status Code	Indicates if record is headquarters, subsidiary, branch, or sole location	0%

Variable	Description	Field Error
Company	Name of business - will have blanks for sole proprietors, like dentists	0%
Company Holding Status	Indicates if company is a public company	Only 1 public company? Or could there be others?
Employee Size (5) – Location*	Number of employees at that location, could be modeled	0%
Employee Size (6) – Corporate*	Actual number of corporate employees for the entire company	Poor quality – only 3 of the 8 businesses identified as "Headquarters" have a valid value.
Industry Specific First Byte	Contains "number of" info. (# beds for nursing homes, # rooms for hotels)	Not defined and 81% of the values are "-1"
Location Employee Size code	Code indicating range of employees at that location - range categories for Employee Size (5) - Location	0%.
Location Name	Name of business - backfills with individual contact name for professionals	0%
Location Sales Volume Code	Corporate sales volume code (ranges) represents the total sales company wide	22% of the values are missing, including all of the 110 values that also don't have Employee Size (5) - Location
Modeled Employee Size	Indicates how Employee Size (5) – Location was determined	0%
NAICS Code	North American Industry Classification System code – assigned	69.4% of the records do not have a valid NAICS or SIC, but they all have a valid Primary NAICS and

Variable	Description	Field Error
	through a match against an SIC crosswalk table.	Primary SIC? Most these have a valid NAICS8 Description?
NAICS8 Description	Description for the NAICS Code	How can 1,642 of these have a valid value when NAICS Code does not?
Parent Actual Employee Size*	Parent actual employee size refers to the parent ABI record only	0%. Coordinates well with Business Status Code – only Headquarters, Subsidiaries or Branches have a valid value for this field (except for 2 Branches)
Parent Actual Sales Volume*	Parent actual sales refers to the parent ABI record only	Only 2 valid values in this field. Corresponds to only 0.9% of Headquarters, Subsidiaries or Branches
Parent Employee Size Code	Code indicating range of employees for the Parent ABI - range categories for Parent Actual Employee Size	0%
Parent Number	The parent number identifies the corporate parent of the business and also serves as the ABI number for the headquarters site of the parent. Since all location of a business have the same ultimate parent number, this field provides 'corporate ownership' linkage information. This information is not collected or maintained for the types organization for which ownership is ambiguous. churches and schools, in	0%. Coordinates well with Business Status Code – only Headquarters, Subsidiaries or Branches have a valid value for this field

Variable	Description	Field Error
	particular, are not linked in the file for this reason	
Parent Sales Volume Code	Code indicating range of sales volume for the Parent ABI - range categories for Parent Actual Sales Volume	Only 2 valid values in this field.
Primary NAICS Code	The description for the primary NAICS code	0%
Primary SIC Code	This field contains the 6-digit SIC code for the business' primary activity	0%
Sales Volume (9) – Corporate*	Actual corporate sales volume represents the total sales company wide. (in thousands)	99%
Sales Volume (9) – Location*	Sales volume at that location (in thousands)	22%
SIC Code	Line of business that company engages	69%; unclear how this differs from Primary SIC Code
SIC Code 1…4	This field identifies an additional activity of the business. If there is no additional activity, this field will be blank.	0%; blanks are defined
Site Number	Designates related business at one site, identifying the primary business. If ABI and this field are the same, then the record is the primary business at the site. If ABI and Site Number are different, then the record is a secondary business at the site – determined through relationships between multiple data elements.	78% of these records are empty. Does that mean the record is NOT the primary business at the site, or that its level of primacy is unknown? Some of this 78% identify a distinct Primary NAICS Code, different from NAICS Code

Variable	Description	Field Error
Square Footage	Indicates the square footage of location firm operates in	5%
Subsidiary Number	The subsidiary parent number identifies the business as a regional or subsidiary headquarters for a corporate family. The subsidiary will always have a parent and may or may not have branches assigned to it.	94%; unclear about use of this variable
Year 1st Appeared	Year first appeared in source record obtained (ccyy) (new adds only)	0%
Year Established	Year the business began operating	94%

4.5 Dun & Bradstreet 2017 Business Data Assessment

4.5.1 Data Reduction / Cleaning

There are a total of 27 fields in the data set, and all of them contain data. Field descriptions are available for each field, but field types and valid values are lacking. DUNS NUM is a unique key field.

The following 10 variables are directly related to the geographic location of the record or the need to contact:

- Street Address
- City Name
- State
- Postal Code
- Mailing Address
- Mailing City Name
- Mailing State
- Mailing Postal Code
- Latitude
- Longitude

The geographic location quality is reviewed in Section 5, so these individual variables are not reviewed separately.

4.5.2 Field Error

The reduction left 17 variables with defined, valid data, all of which have relevance to transportation planning. Their data quality is described in Table 7. Field errors were measured as the fraction of blank or undefined responses to the total number of opportunities for a records.

Table 7 Field Errors in the Dun & Bradstreet 2017 Business Data

Variable	Description	Field Error
Business Name	The Primary or Registered name of the business.	0%
Trade Name	A trading style name used by a business. It is an additional name used by a company. Also referred to as "Doing Business As" (DBA) and "Also Known As" (AKA).	79%

Variable	Description	Field Error
Employees Here	The number of employees at this location.	0%, although six records show 0 employees?
Year Started	The year when current ownership or management assumed control of the business or the year established if no control change has taken place. This is not provided for branch records.	9% of the records show a year of "0"
Small Business Indicator	Indicates whether the business is designated a small business as defined by the Small Business Administration of the US government.	0%
Manufacturing Indicator	Indicates whether or not manufacturing is done at this location.	0%
Primary SIC	The US 1987 Standard Industrial Classification (SIC) code system categorizes business establishments based upon the type of activity done by that business at that location. A business can have up to six SIC codes and each SIC can have four extensions. The first-listed SIC code represents the primary operations of the business. Then, SIC codes are assigned in descending order according to the percentage of the revenue contributed by each function of the business. The SIC code of a parent/ultimate may include the activities of its subsidiaries.	0%
Primary SIC Description	A narrative description of the operations or activities of the business. Relates to the primary eight-digit 1987 US SIC.	0.1%
Primary NAICS Code	The NAICs code used to categorize the business establishment. Code is translated using the US 1987 Standard Industrial Classification code system.	0.1%

Variable	Description	Field Error
Primary NAICS Code Description	A narrative description of the operations or activities of the business. Relates to the NAICs code.	0.1%
Sales Volume (US Dollars)	The total annual sales/revenue for this business, expressed in US dollars as a signed, decimal field.	10% of the records have "$0.00"
3 Year Growth Percentage on Employees	The percentage increase or decrease in the number of employees over a three year period. Includes a + or - sign denoting positive or negative growth in sales.	0%
5 Year Growth Percentage on Employees	The percentage increase or decrease in the number of employees over a five year period. Includes a + or - sign denoting positive or negative growth in sales.	0%
3 Year Growth Percentage on Sales Volume	The percentage increase or decrease in the sales volume over a three year period. Includes a + or - sign denoting positive or negative growth in sales.	0%
5 Year Growth Percentage on Sales Volume	The percentage increase or decrease in the sales volume over a five year period. Includes a + or - sign denoting positive or negative growth in sales.	0%

.

5 Geographic Data Quality Assessment

All InfoGroup and Dun & Bradstreet records were provided with a latitude and longitude, to approximate the geographic location of the record address of record. The quality of this geographic data can be a significant determinant of the usefulness of the data for transportation planning and travel modeling.

The variables related to the geo-coding quality in the InfoGroup data share the same coding (Match Level, Match Code, or Match Level Code):

- P – address was matched to the parcel
- 0 – address was matched to the "Site", or location
- 4 – address was matched to the centroid of the ZIP+4 area
- 2 – address was matched to the centroid of the ZIP+2 area
- X – address was matched to the centroid of the 5-digit ZIP code area

The 5-digit ZIP code area is the delivery area of the post office responsible for delivery to this location. The ZIP+4 identifies a specific delivery route within that overall delivery area. Table 8 provides the frequency of each matching type for each of the 4 InfoGroup datasets.

Table 8 Geo-Coding Quality of InfoGroup Data

Geo-Coding Level		2015 Consumer		2015 Business		2017 Consumer		2017 Business	
P	Parcel	15,054	50%	1,661	69%	24,815	83%	1,941	81%
0	Site Level	992	3%	107	4%	908	3%	82	3%
4	ZIP+4	219	1%	26	1%	3,626	12%	105	4%
2	ZIP+2	377	1%	19	1%	221	1%	12	1%
X	ZIP	13,358	45%	587	24%	430	1%	260	11%

Between 50% and 83% of the household locations were matched to the parcel, indicating a highly variable degree of quality in the geo-coding effort. Lacking a variable to indicate the geo-coding quality, the degree of quality for the Dun & Bradstreet data could not be determined.

Vermont's E911 point data was used to validate the quality of the geo-coding results. A minimum error for every point in the dataset was determined as the distance from each address point to the nearest structure whose type matches the type of address. For assessing the consumer data, which should

correspond with residential households in Vermont, the following E911 site types were used:

- Commercial w/Residence
- Condominium
- Institutional Residence / Dorm / Barracks
- Mobile Home
- Multi-Family Dwelling
- Other Residential
- Residential Farm
- Seasonal Home
- Single Family Dwelling

E911 points corresponding to these residential structures comprised 255,633 points. For assessing the business data, which should correspond to commercial locations in Vermont, the following E911 site types were used:

- Airport Terminal
- Auditorium / Concert Hall
- Bank
- Brewery
- Bus Station / Dispatch Facility
- Campground
- Cemetery
- College / University
- Commercial
- Construction
- Farm
- Garage
- Comm. w.Res.
- Community / Rec.
- Court House
- Day Care Facility
- Educational
- Fair Grounds
- Fish Hatchery
- Fitness Facility
- Food Distribution
- Gas Station
- Gated w.Building
- Golf Course
- Gravel Pit / Quarry / Mine
- Greenhouse
- Grocery Store
- Harbor / Marina
- Health Clinic
- Hospital / Med.
- House Of Worship
- Hydroelectric
- Ice Arena
- Industrial
- Landfill
- Law Enforcement
- Library
- Lodging
- Lumber/Saw Mill
- Mnufctrg Facility
- Morgue
- Museum
- National Guard / Armory
- Nursing Home
- Office Building
- Oil / Gas Facility
- Other Comm.
- Outpatient Clinic
- Pharmacy
- Post Office
- Correctional Facility
- Express Shipping
- Race Track
- Railroad Station
- Restaurant
- Retail
- Shooting Range
- Ski/Alpine Resort
- Sports Arena
- Sugarhouse
- Transfer Station
- Veterinary Hosp.
- Visitor Info. Cntr
- Warehouse
- Waste / Biomass
- Wastewater Treatment Plant
- Youth Camp

E911 points corresponding to these commercial businesses comprised 26,052 points. Summary statistics of the minimum-error values for each dataset are provided in Table 9.

Table 9 Minimum Error Summary of Geo-Coding Quality

	All Points			Parcel Match			Other Match		
Data Set	Max	Mean	> 0.31	Max	Mean	> 0.31	Max	Mean	> 0.31
IG 2015 Consumer	0.52	0.03	0.0%	0.52	0.01	0.1%	0.48	0.05	0.0%
IG 2015 Business	3.37	0.13	12.8%	2.87	0.09	8.5%	3.37	0.21	**22.3%**
IG 2017 Consumer	0.86	0.03	0.3%	0.66	0.03	0.3%	0.86	0.04	0.3%
IG 2017 Business	2.11	0.11	10.4%	2.11	0.11	10.3%	1.33	0.11	10.9%
DnB 2017 Business	2.69	0.22	**22.5%**	NA					

Generally, the InfoGroup data locations that had been matched only to the nearest zip code (or "Site") had a higher mean minimum-error than those which had been matched to the parcel. For all of the data sets but the IG 2015 Business data and the DnB 2017 Business data, 10% or fewer records fall greater than 0.31 miles from the nearest potentially matching E911 address point. 0.31 miles is half the theoretical maximum acceptable walking distance for a trip, so geo-coding errors beyond this threshold are less useful for travel and impacts modeling.

Both the 2015 and 2017 business data sets from InfoGroup had a higher mean minimum error in geo-coding than their consumer data counterparts. The 2017 data sets had little difference between those points matched to the parcel and those matched to the zip code or "Site". However, the 2015 Business data had a significantly compromised accuracy for the points that were not matched to the "Parcel". For VTrans' purposes, these points should be considered unsuccessfully geo-coded.

The disparity in the geographic quality of the recent (2017) and historical (2015) business data from InfoGroup may be a testament to the company's traditional focus on current, up-to-date data. However, it may also be possible to improve the geo-coding accuracy by matching addresses from the historical data to the E911 points. A cursory inspection of the data sets revealed that approximately two-thirds of the 2015 business records with an address could be matched successfully to an E911 point. Unfortunately, 82 of the records in the 2015 business data contain no street address, making this improvement impossible. **For address-based point data, records lacking an address are unacceptable.**

17 of the 2,000 records in the DnB 2017 business data lacked valid coordinates, and were not able to be mapped in a GIS. Complete addresses

were provided for each of these records, so it is not clear why they were not geo-coded. Nonetheless, the geo-coding quality of the records that were successfully geo-coded is still very poor, with over 22% of the points falling more than 0.31 miles from the nearest business point from the E911 data. This level of geo-coding accuracy is unacceptable, particularly without a separate field indicating the quality of the geo-coding (like “Match Level” from the InfoGroup data).

6 Conclusions and Recommendations

According to the responses received from the travel modeling community nationwide, address-based business and consumer data is overwhelmingly used for point-level employment counts from the business data. The only use of the consumer data identified was for providing a population "frame" and contact information for carrying out initial contacts for a household travel survey.

The use of DnB data appeared to be more common amongst the transportation planning community. This could indicate that DnB are the more "established" source in the field and InfoGroup is trying to penetrate that market, or that DnB is simply a more recognizable name.

For travel demand modeling, the business data was described as allowing the use of flexible geographic areas for sub-area modeling. For measuring impacts more generally, the business data allows for more precise results, as shown in Attachment A. Many users offered cautions about the need for excessive "cleaning" of the data after purchase, and the potential for "headquarters/branch" employment counts to have errors with their location.

The data quality assessment conducted consisted of a data reduction and an assessment of field error, resulting in the removal of many of the data fields delivered, particularly by InfoGroup. Many of the field provided were either missing, unexplained, or not relevant to transportation planning. Through the data reduction, a final set of fields that provide useful, valid, defined data was determined. A summary of the data reduction is provided in Table 10.

Table 10 Summary of the Data Reduction

No. of Fields (Variables)...	InfoGroup				DnB
	2015 Consumer	2015 Business	2017 Consumer	2017 Business	2017 Business
Delivered	63	89	145	172	27
Blank, Unexplained, Undefined, or Unavailable	8	2	72	42	0
Related to Geo-coded Location	32	27	36	62	10
Not Relevant to Transportation Planning	8	18	11	20	0

Redundant or Unacceptable Quality	6	33	11	22	8
Final Set	**9**	**9**	**15**	**26**	**9**

The final set of fields are summarized in Table 11.

Table 11 Final Set of Data Fields

IG 2015 Consumer		
Household Income	Year Home was Built	Marital Status
Household Wealth	Age of Head of Household	Primary Family at the Address?
Household Purchasing Power	Length of Current Residence	Vacancy?
IG 2015 Business		
Headquarters, Branch, or Sole Location	Location Name	Secondary SIC Codes
Company Name	Parent Company Employee Size	Square Footage
Location Employee Size – Modeled and Range	Primary SIC/NAICS Code	Year 1st Appeared
IG 2017 Consumer		
Adult Age Range	Home Age	Mean Years of Schooling
Delivery Unit Size	Home Equity Estimate	Number of Trade (Credit) Lines
Early Internet Adopter?	Household Income	Residence Ownership
Expendable Income	Loan-to-Value Ratio	Residence Type
Heavy Internet User?	Marital Status	Household Wealth
IG 2017 Business		
Corporate Employment Size	Foreign Parent Company?	Public Company?
Corporate Sales Volume	High-Income Executives?	Bankruptcy Filing?
Credit Score	High-Tech Business?	Secondary SIC
Location Employment Size	Headquarters or Branch	Small Business?
Affluent Neighborhood?	Import/Export Activity?	Square Footage
Asset Size	Individual or Firm	White Collar?
Big Business	Medium Size Business?	White Collar Percentage
Female Owner Executive?	Modeled Employment Size	Work At Home Business?
Fleet Size	NAICS Code	
DnB 2017 Business		
Business Name	Small Business?	Primary NAICS Code

Location Employee Size	Manufacturing?	Sales Volume
Year Started	Primary SIC	3-Year & 5-Year Growth in Employment and Sales Volume

Frequencies of responses for all of these variables in the data provided can be found in Appendix B. Subsequent requests for data should only include these variables. Other consumer variables of interest for transportation modeling and planning include the number of household vehicles, the household size and composition (number of children, other adults, etc.), student status of children, and worker status of adults. Other business variables of interest for transportation modeling and planning include fleet size, fleet type (vehicle size), and shipment information (incoming/outgoing weight, frequency, mode, vehicle size, etc.). Subsequent communications with the vendors should point out that this additional information would be welcomed.

The geo-coding quality of the data varied considerably. For the Dun & Bradstreet data, the tolerances of the geo-coding quality compromise its use for detailed spatial analysis. Correcting over 20% of the geo-coded locations is not feasible. Geo-coding of the 2017 InfoGroup data is considerably better, with the consumer data in particular achieving a high rate of matching to the Parcel level. However, the 2015 InfoGroup data does not achieve nearly the same level of quality. It is unclear if all data before 2017 will be compromised in the same way, or if geo-coding of any data that is not "current" loses quality. In either case, subsequent requests for data should stipulate geo-coding quality that meets the following standards:

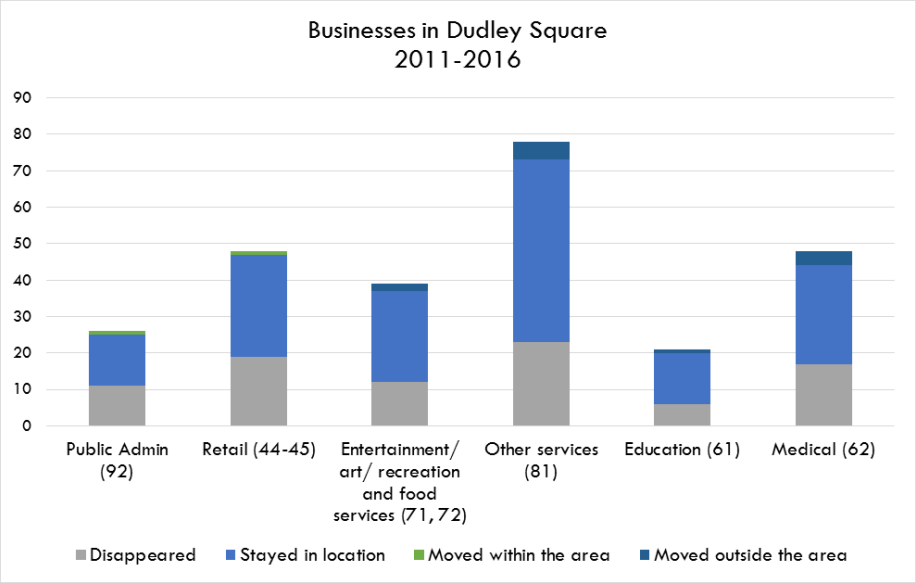
- 80% or more of the geo-coded locations from each data set (measured independently) matched to the PARCEL
- 90% or more of the geo-coded locations from each data set (measured independently) within 0.31 miles of an associated point in the E911 point shapefile

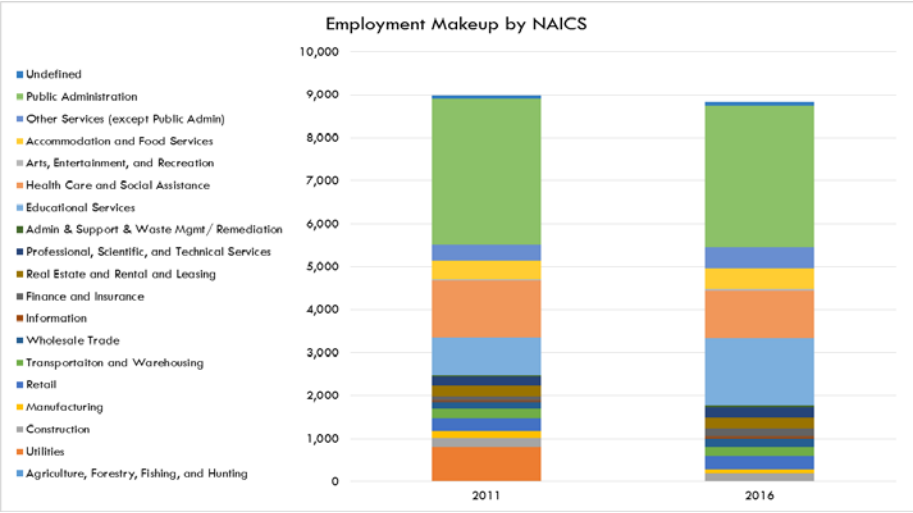
Providing the current E911 point shapefile may enhance the vendor's ability to geocode and comply with the second standard.

Appendix A – Neighborhood Business Change Analysis

NEIGHBORHOOD BUSINESS CHANGE ANALYSIS

While many methods are available to help understand demographic changes at a neighborhood level, few exists to help understand economic change at that scale. MAPC conducted an analysis to explore the possibility of using geo-located business data from 2011 and 2016 to understand the change in a neighborhood's businesses. MAPC looked at businesses in the ½-mile radius around the MBTA Dudley Square station for this exploration.

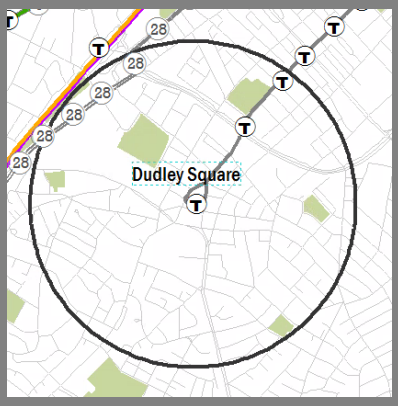




The data includes a unique business ID, company name, address, number of employees, and NAICS code.

Overall, there is significant turnover of businesses in the area. 57% of businesses in Dudley at 2011 stayed in the area until 2016. 41% of all businesses in 2016 were new businesses.

However, a detailed eye with local knowledge of the area is necessary to supplement the information and make sure the analysis reflects what's happening in the area.



Methodology and notes:

Our analysts extracted all of the business points from 2011 within the ½ mile surrounding Dudley Square which we define as our study area. All businesses with only 1 employee were removed from the analysis. The unique business IDs for businesses in 2011 were joined to the same business ID in the 2016 data. All businesses in 2011 that did not have a match in 2016 were labeled "disappeared." If the 2011 point and corresponding 2016 point were within 50 meters from one another, the business was labeled "stayed."

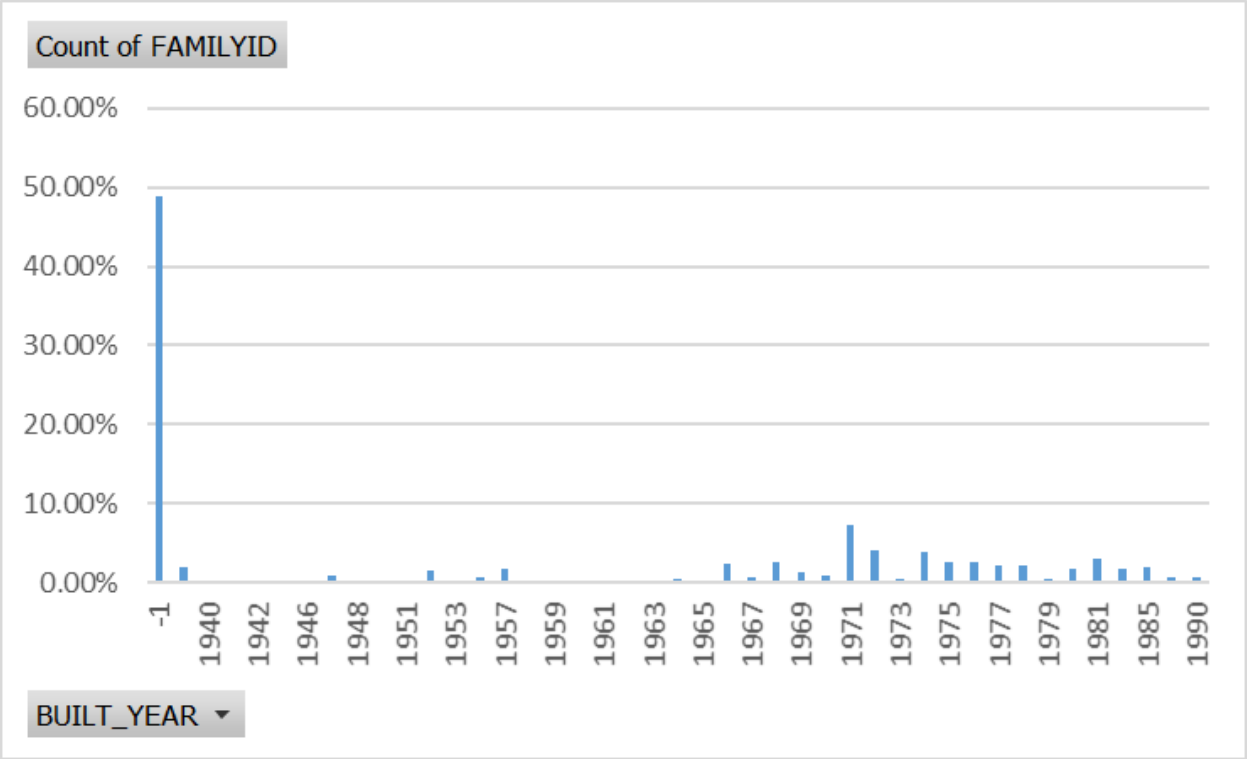
For any point in which the 2016 business was outside of the study area and greater than 50 meters from the 2011 point, the business was labeled "moved outside the area". Any other 2016 matched point was labeled "moved within the area." Our analysts then extracted all of the business points from 2016 within the study area. 2016 points that fall within the study area, but were not matched to a unique ID seen in 2011 were labeled "new." After reviewing the businesses individually, we believe this analysis captures a variety of businesses at different sizes and revenues.

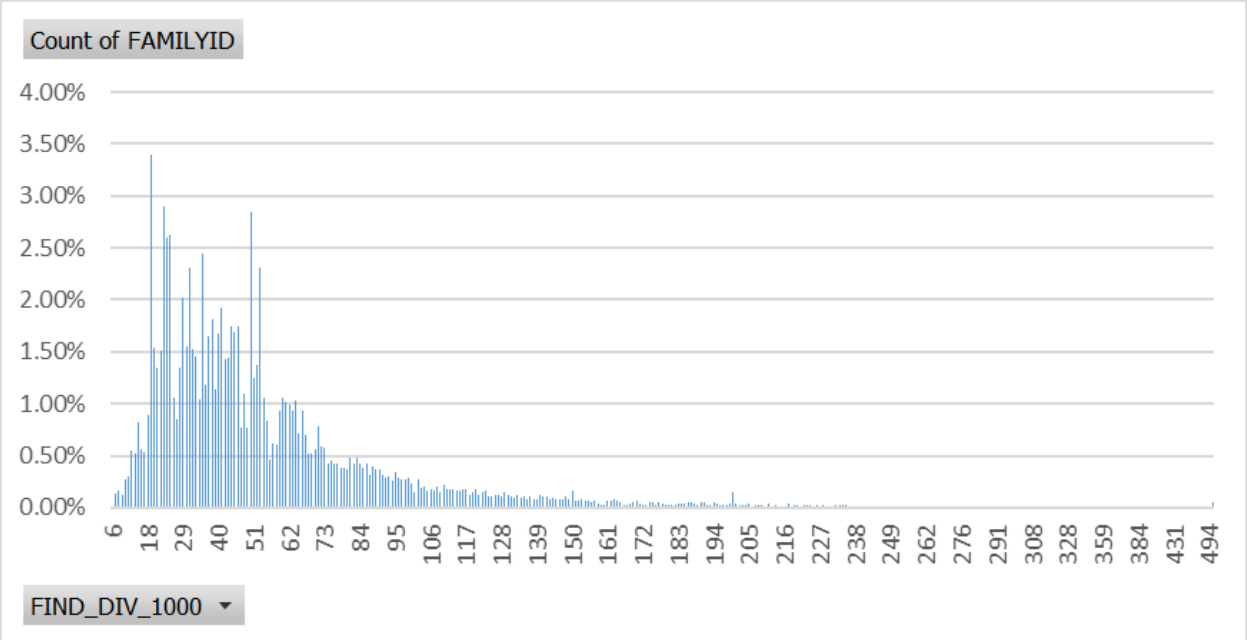
A detailed eye with local knowledge of the area is necessary to supplement the information and make sure the analysis reflects what's happening in the area. We saw a pronounced drop in the Utilities sector (804 jobs in 2011 to 4 jobs in 2016) and an increase in educational services (886 jobs in 2011 to 1,561 jobs in 2016). The drop in utilities was due to the Boston Water Commission reported as 804 employees in 2011 and only 4 in 2016, which was found to be unsubstantiated upon further investigation. This points to need for expert analysts in extracting the information from data while cleaning out inadvertent errors in the dataset.

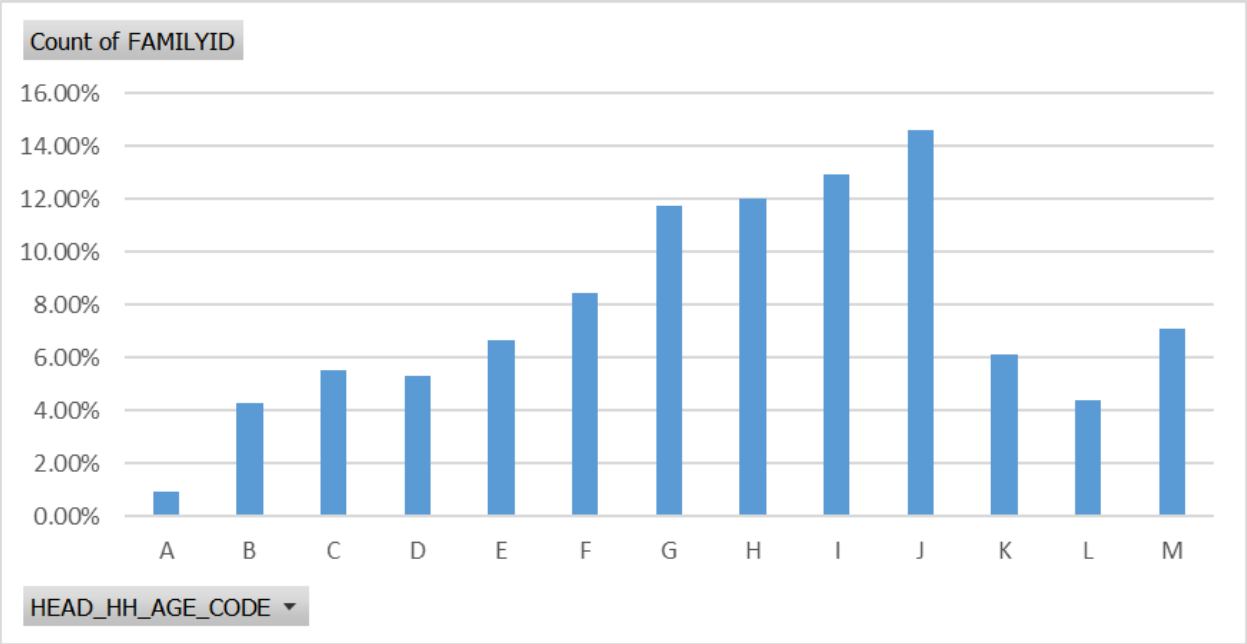
MAPC purchased data from Info-group for this analysis. The data contains access to point-level business data from 2011 and 2016. The data includes a unique business ID, company name, address, number of employees, and NAICS code. Due to confidentiality agreement any raw data shared with the constituents and public, will be either at the 250m grid level or at a block level.

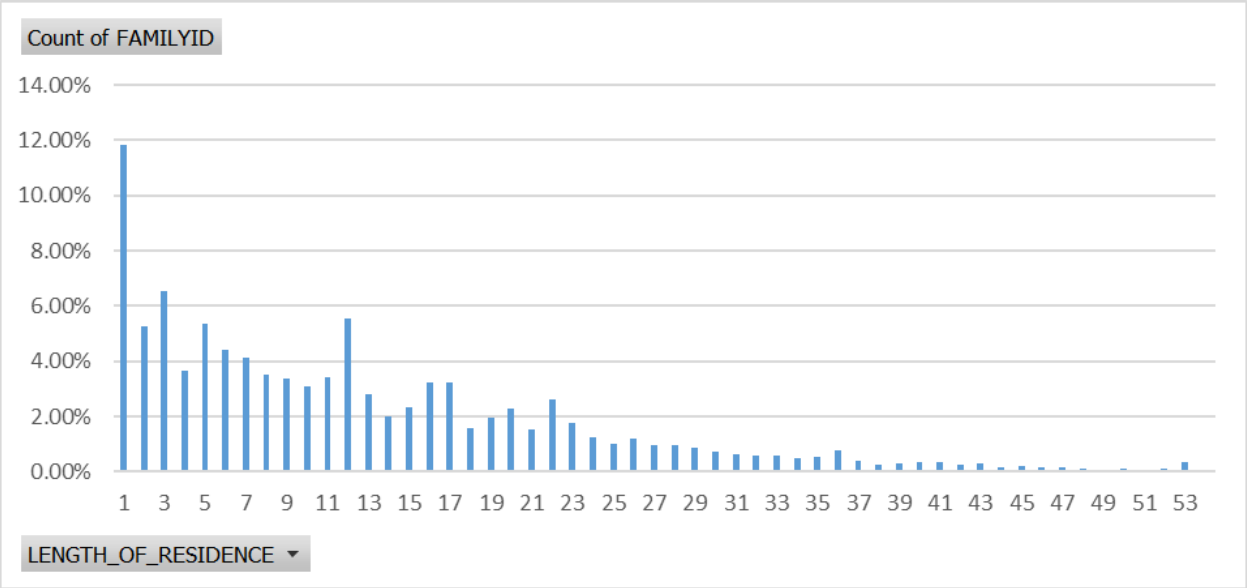
Appendix B – Response Frequencies for Final Set of Variables

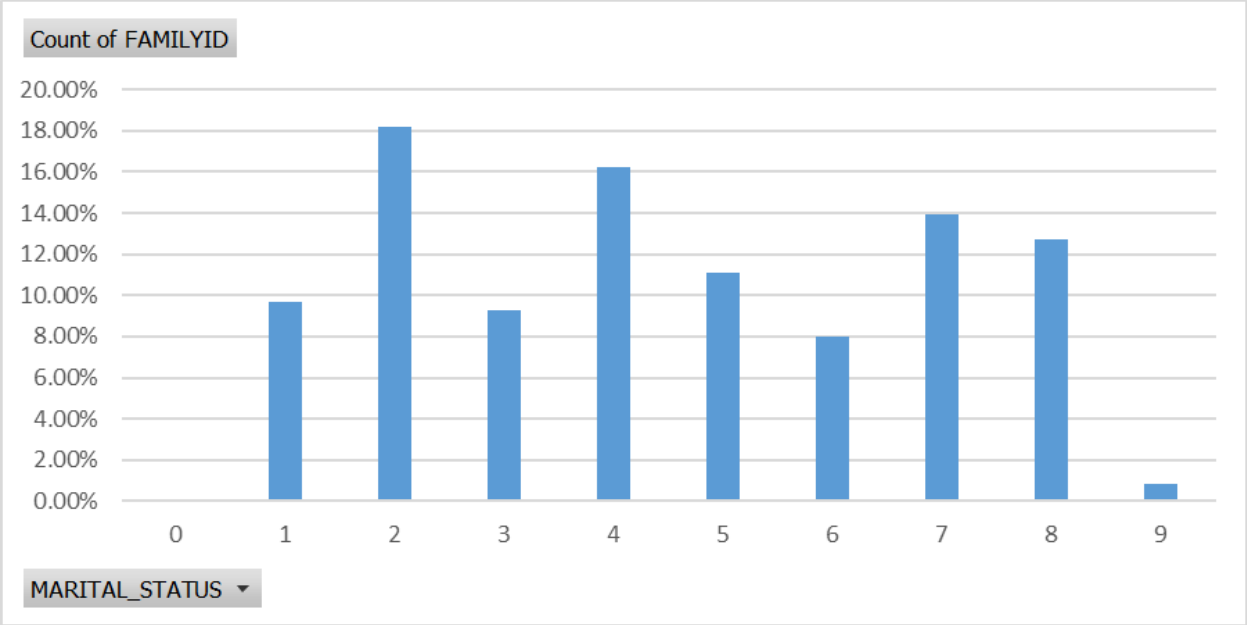
InfoGroup 2015 Consumer Data Charts

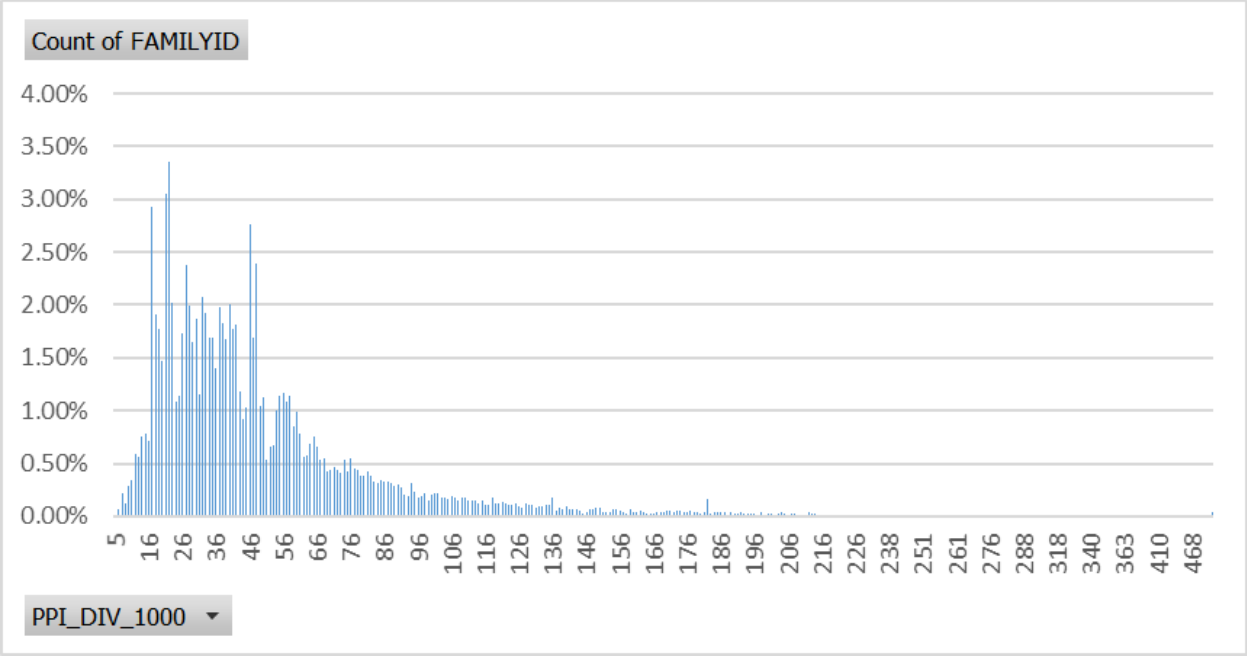


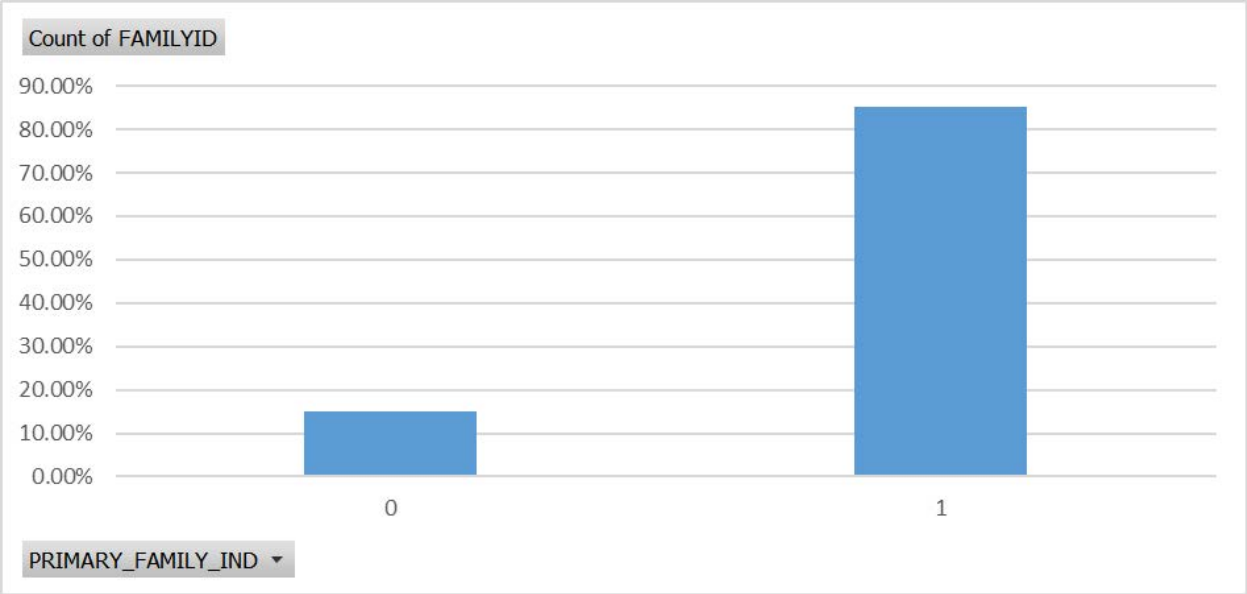


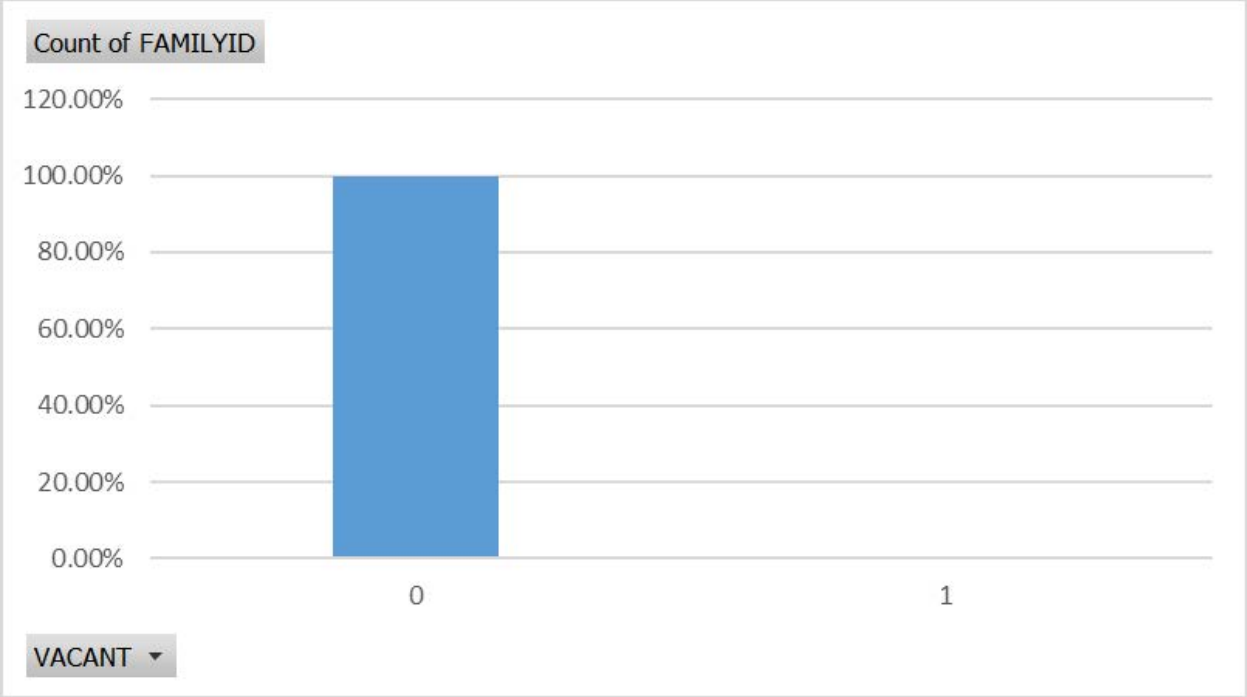

Count of FAMILYID
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
A
B
C
D
E
F
G
H
I
J
K
L
M
HEAD_HH_AGE_CODE

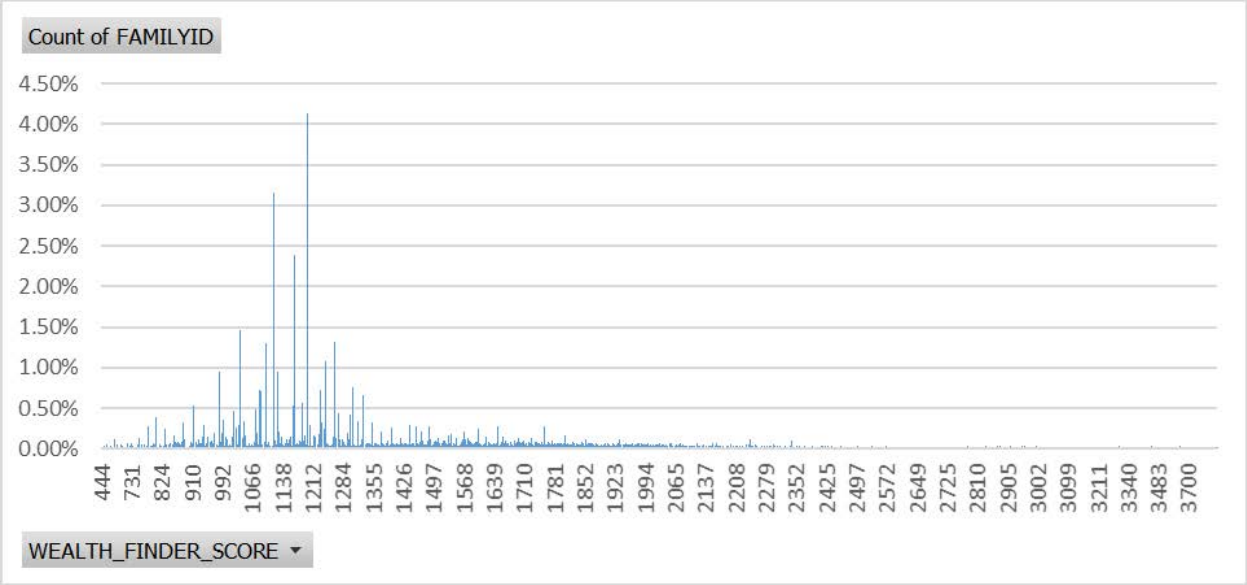

Count of FAMILYID
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
LENGTH_OF_RESIDENCE


Count of FAMILYID
20.00%
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
0
1
2
3
4
5
6
7
8
9
MARITAL_STATUS


Count of FAMILYID
4.00%
3.50%
3.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
5
16
26
36
46
56
66
76
86
96
106
116
126
136
146
156
166
176
186
196
206
216
226
238
251
261
276
288
318
340
363
410
468
PPI_DIV_1000

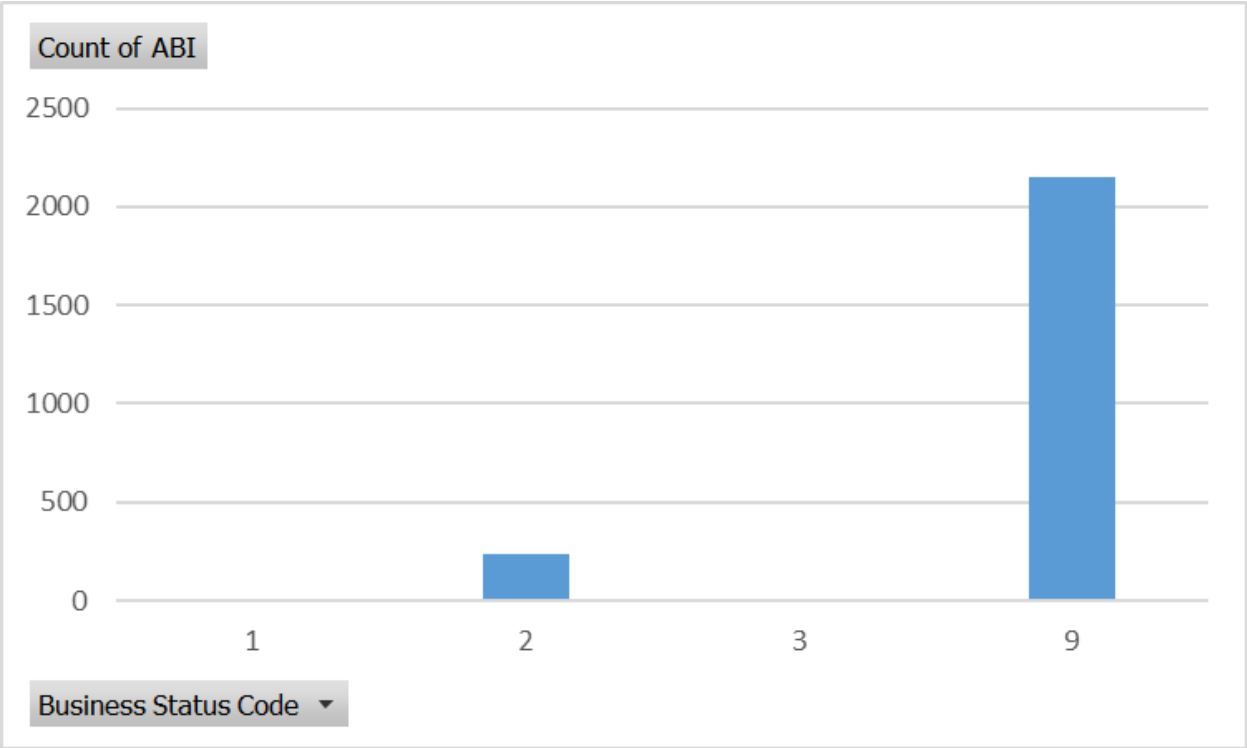

Count of FAMILYID
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
1
PRIMARY_FAMILY_IND

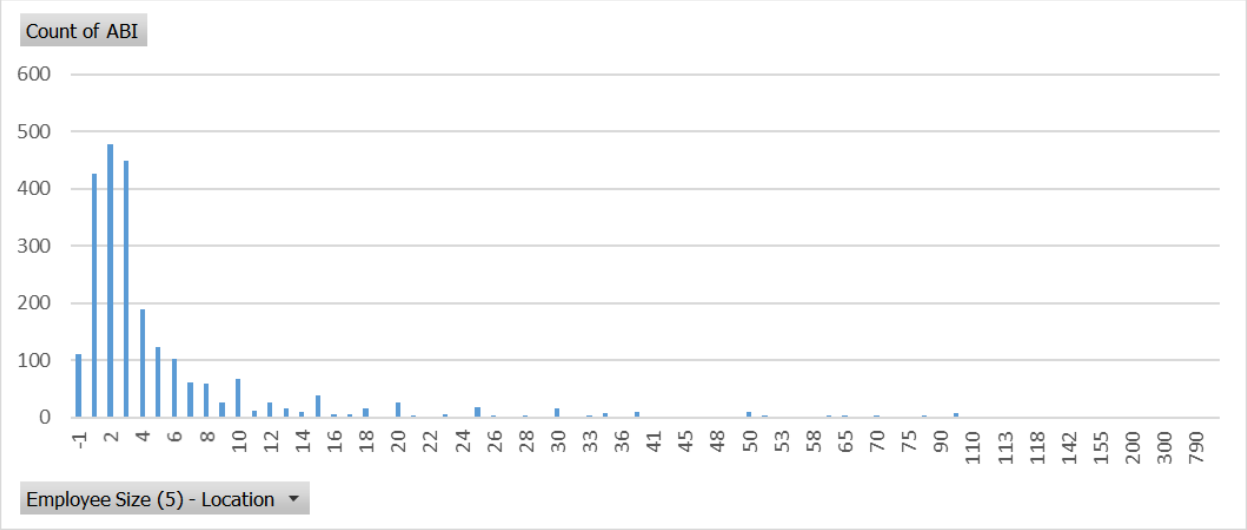

Count of FAMILYID
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
0
1
VACANT

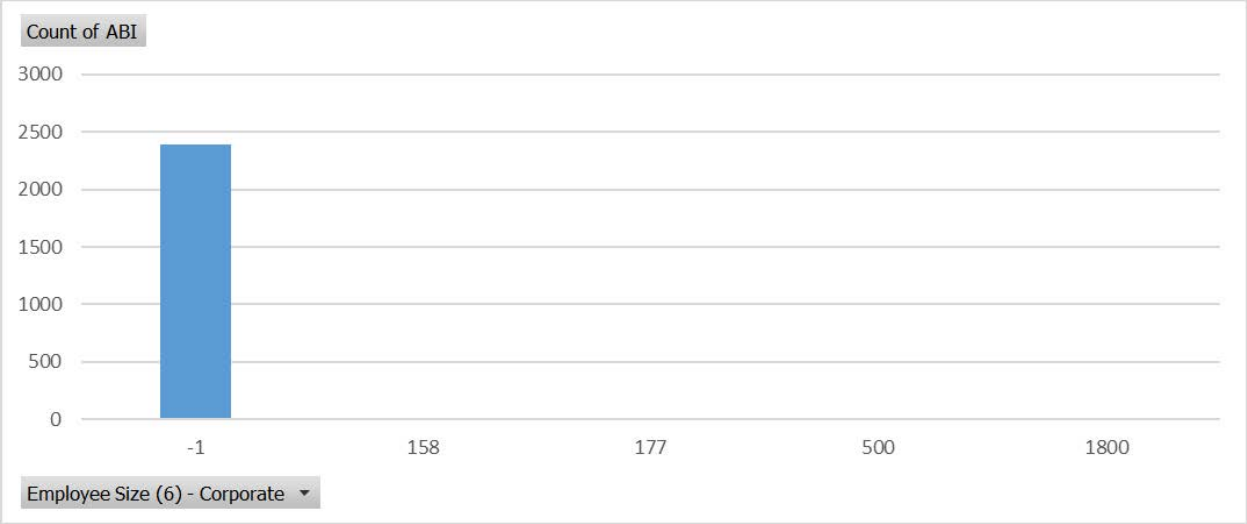

Count of FAMILYID
4.50%
4.00%
3.50%
3.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
444
731
824
910
992
1066
1138
1212
1284
1355
1426
1497
1568
1639
1710
1781
1852
1923
1994
2065
2137
2208
2279
2352
2425
2497
2572
2649
2725
2810
2905
3002
3099
3211
3340
3483
3700
WEALTH_FINDER_SCORE

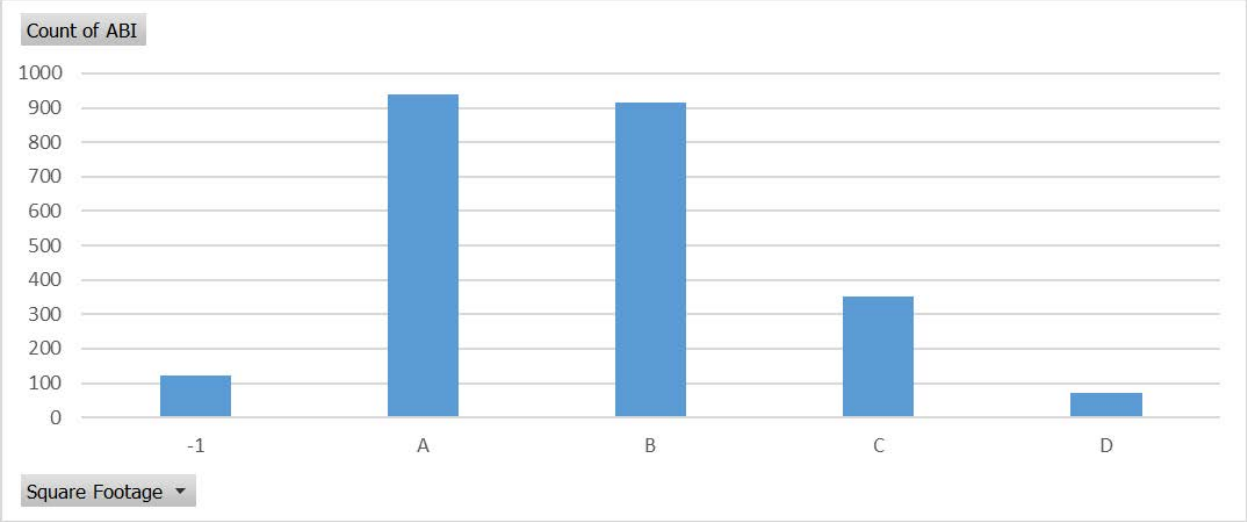
InfoGroup 2015 Business Data Charts

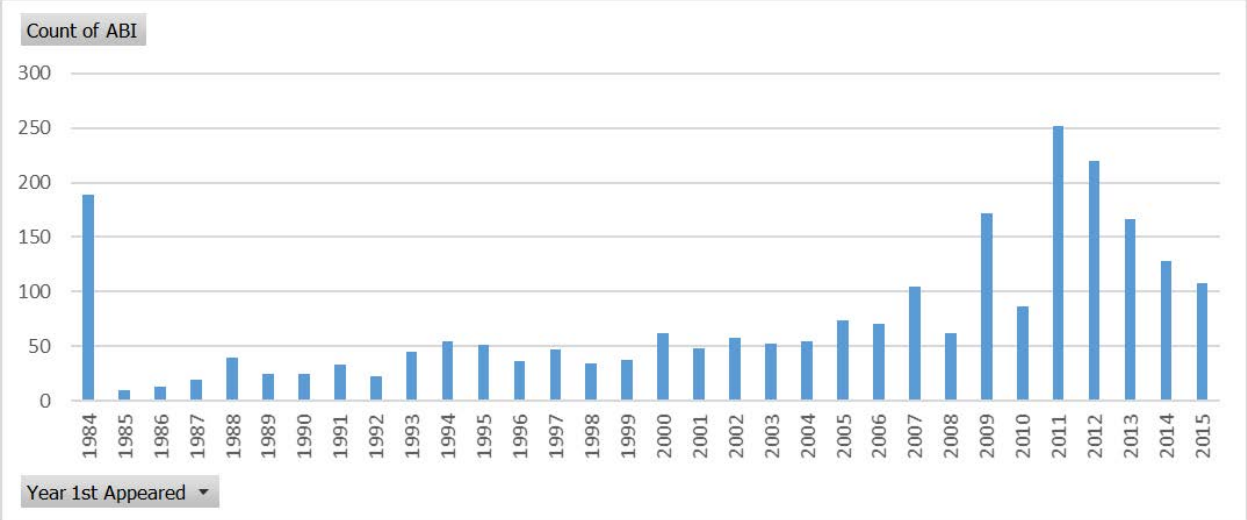
Company Name, Primary NAICS/SIC Codes not shown



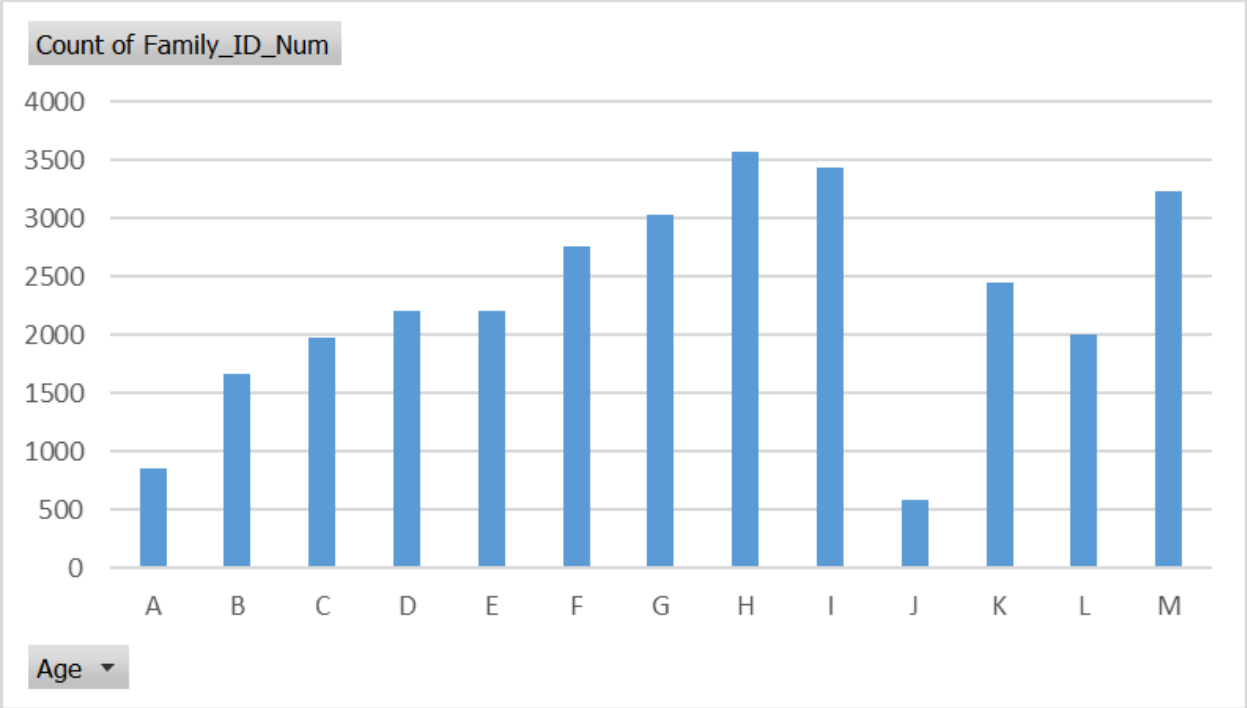


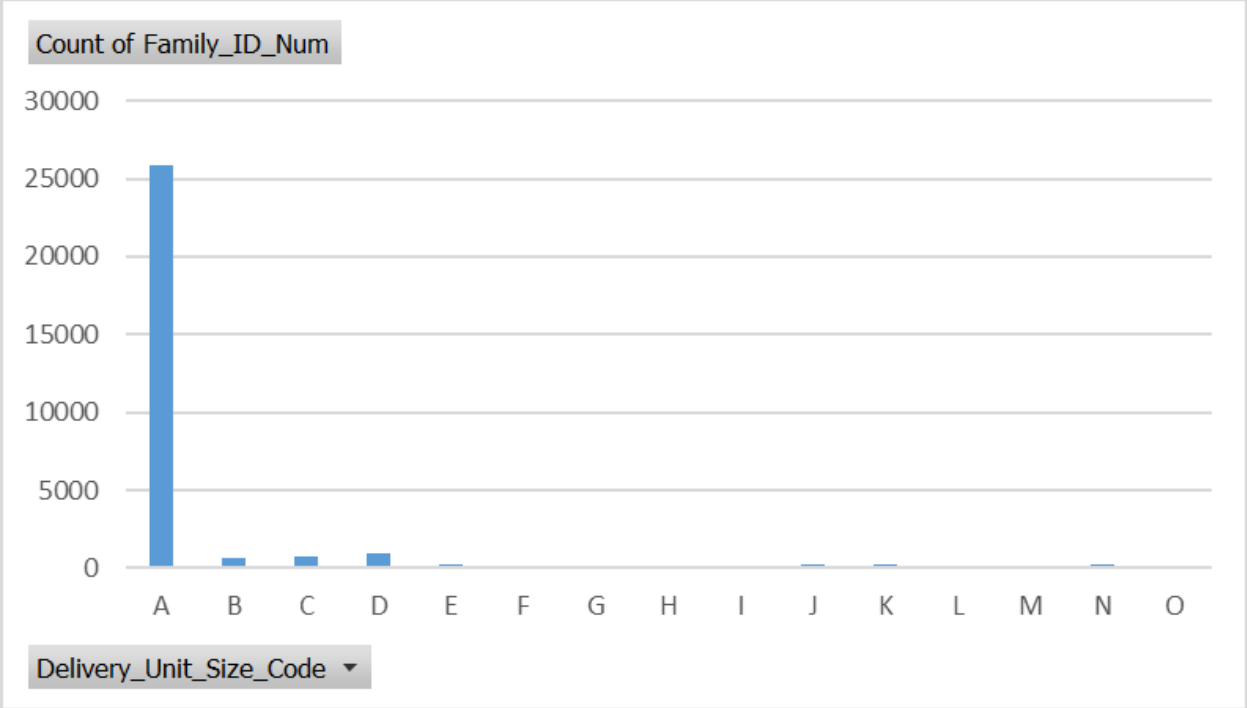

Count of ABI
3000
2500
2000
1500
1000
500
0
-1
158
177
500
1800
Employee Size (6) - Corporate

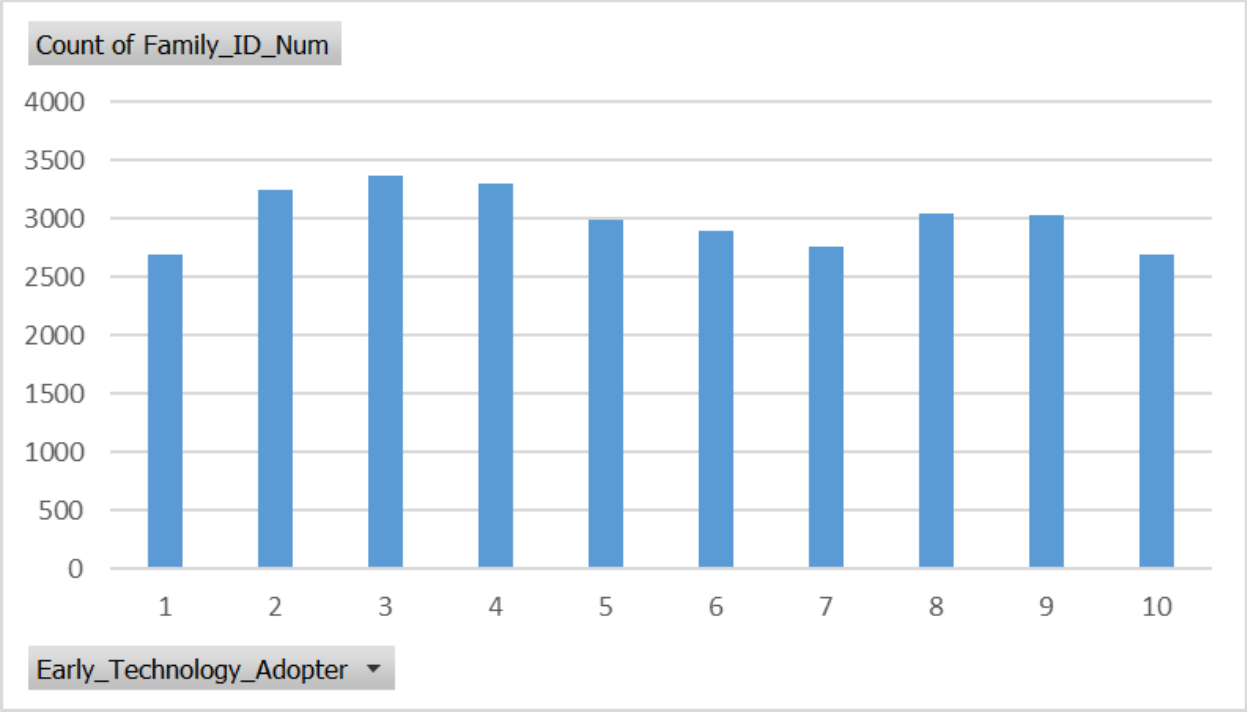

Count of ABI
1000
900
800
700
600
500
400
300
200
100
0
-1
A
B
C
D
Square Footage

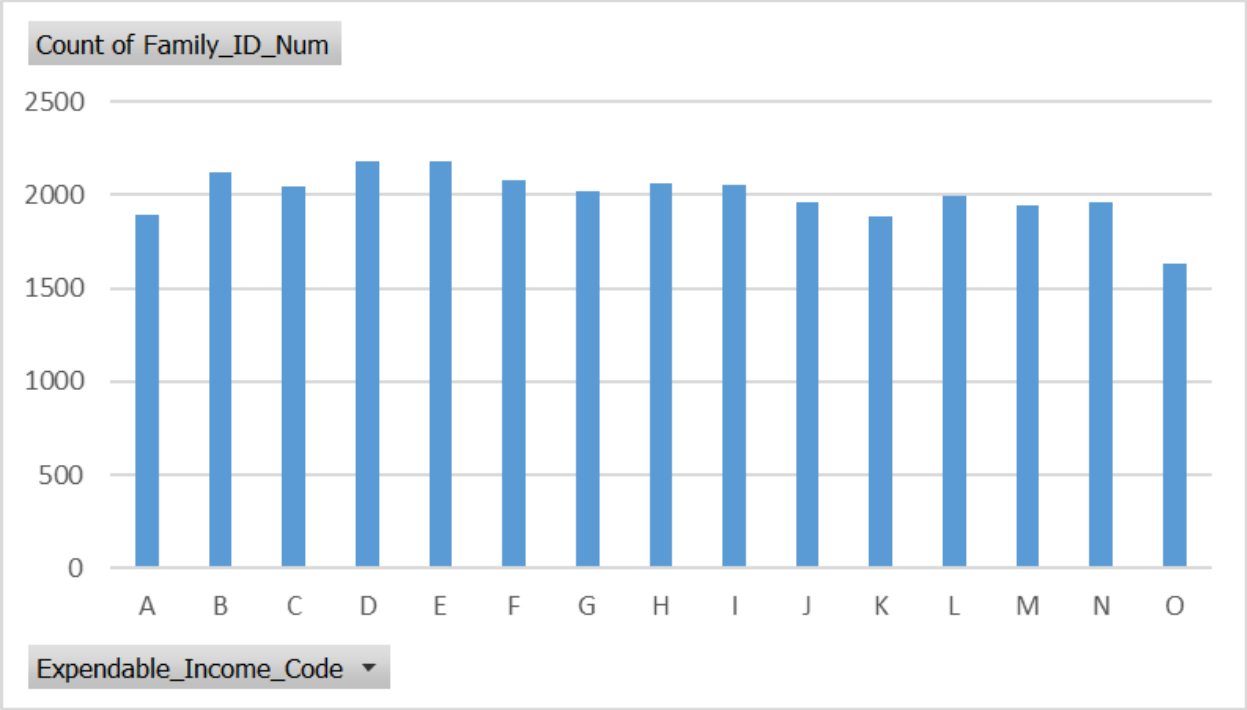

Count of ABI
300
250
200
150
100
50
0
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year 1st Appeared

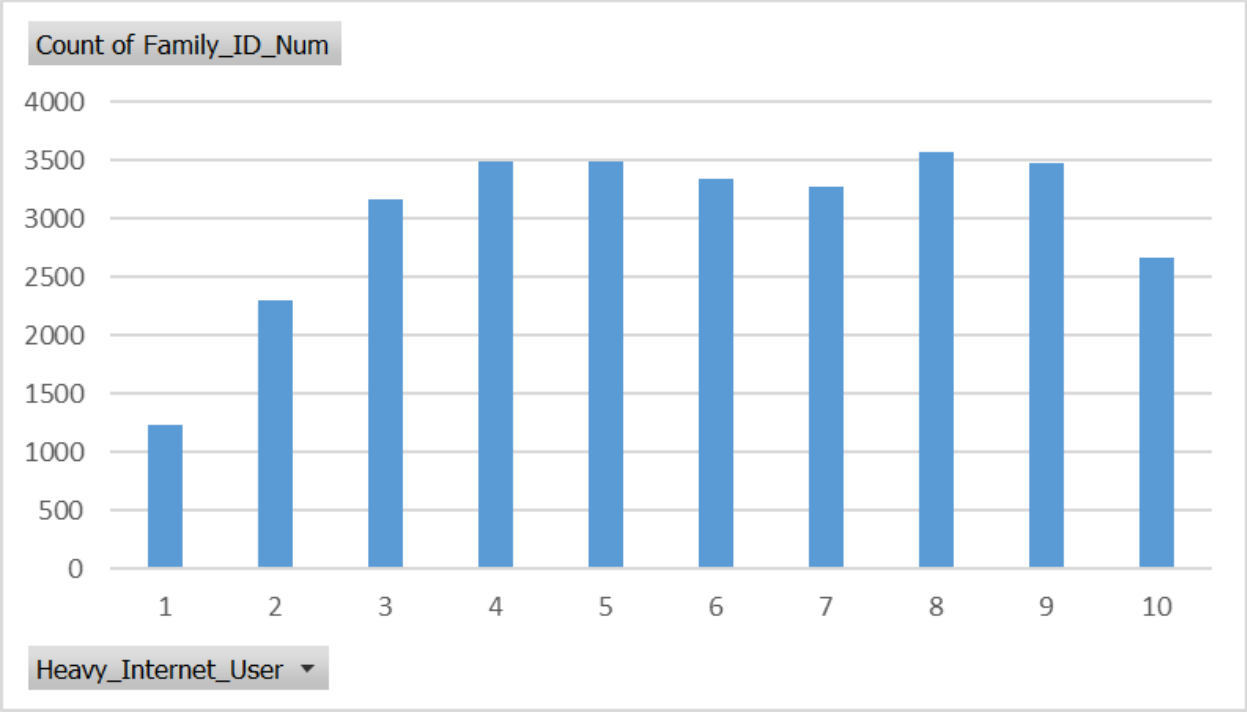
InfoGroup 2017 Consumer Data Charts

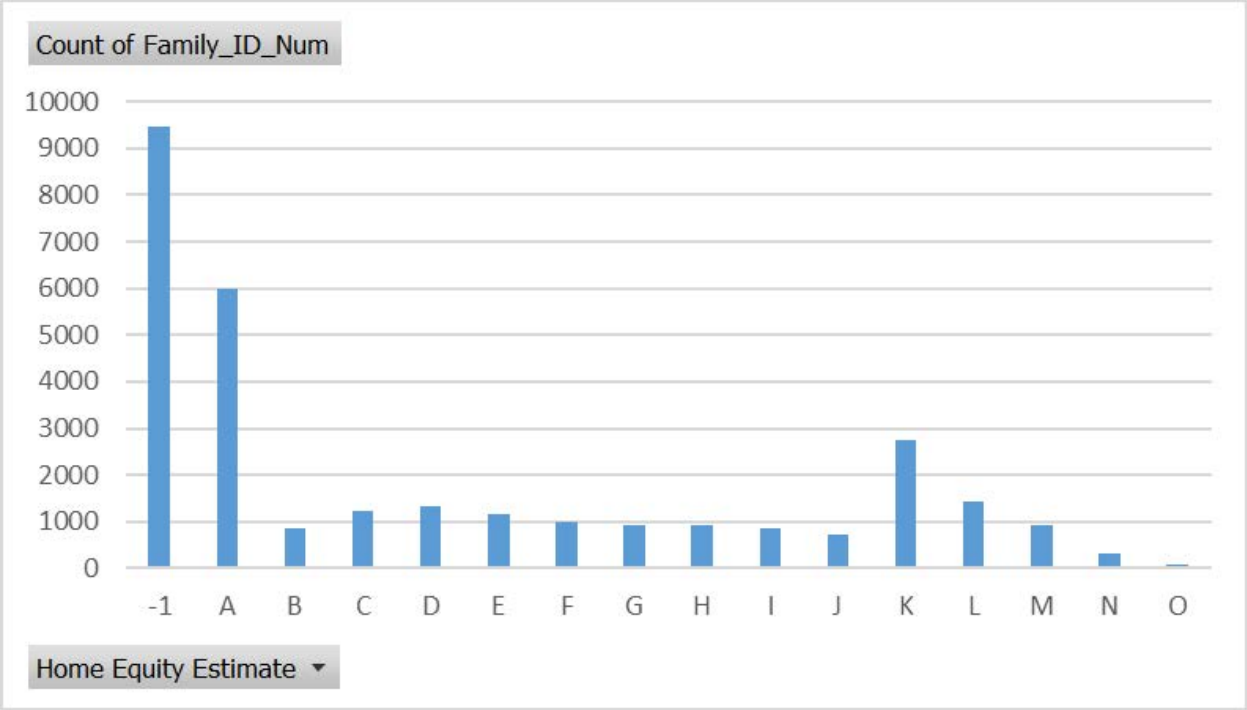


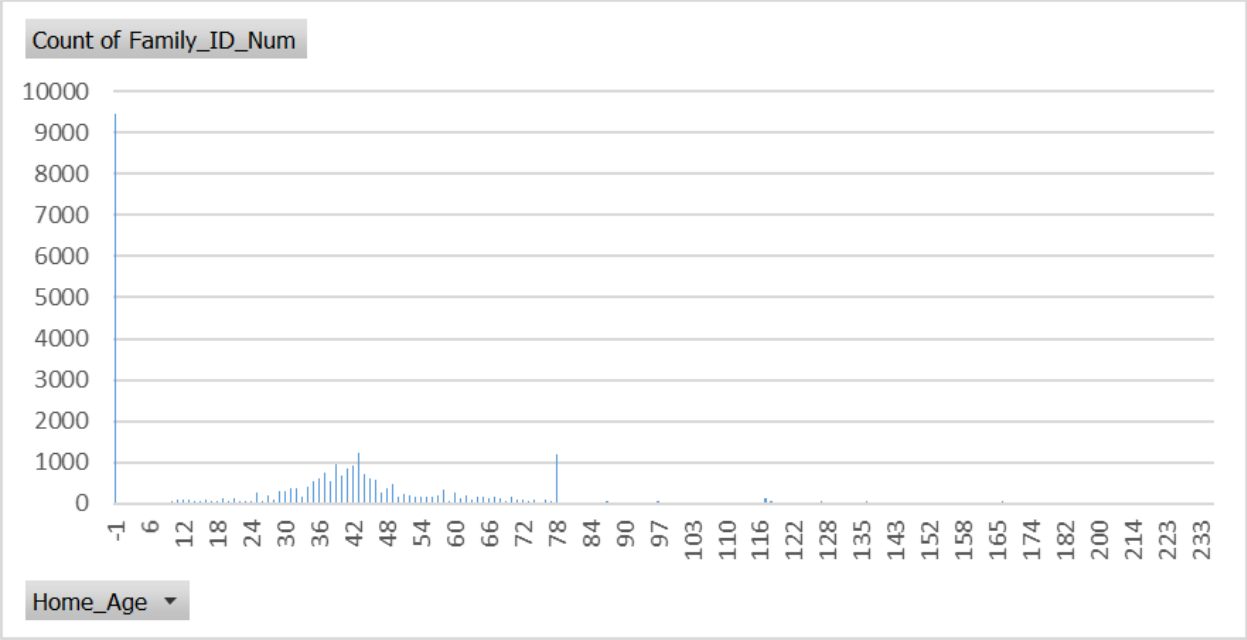


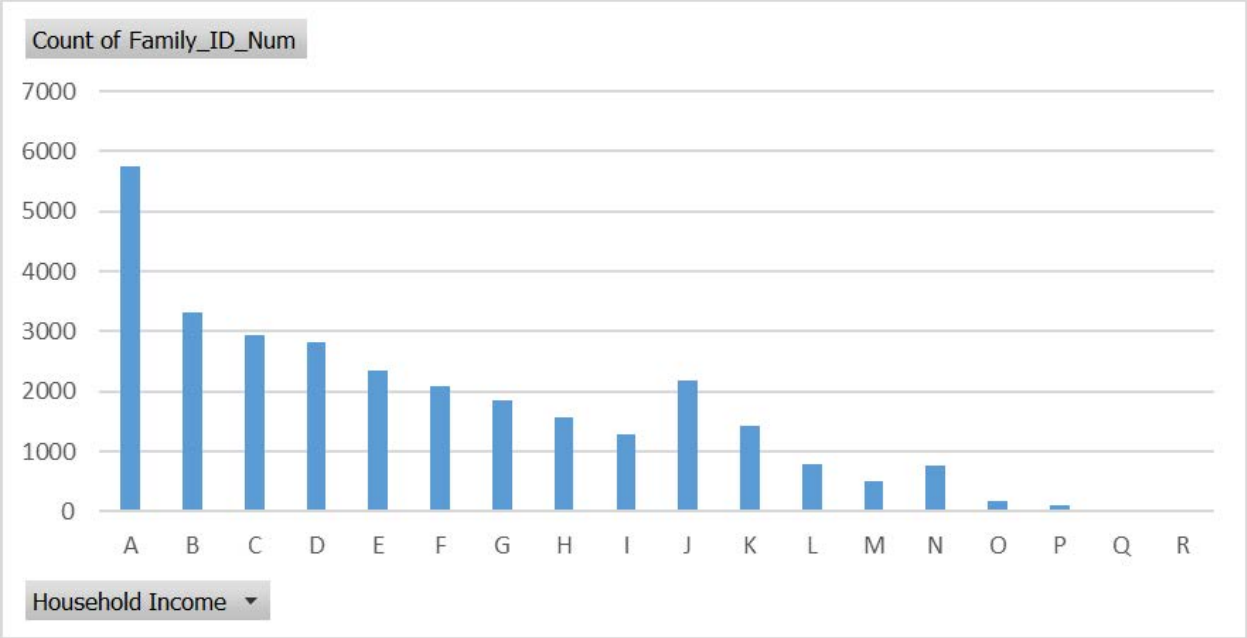

Count of Family_ID_Num
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
Early_Technology_Adopter

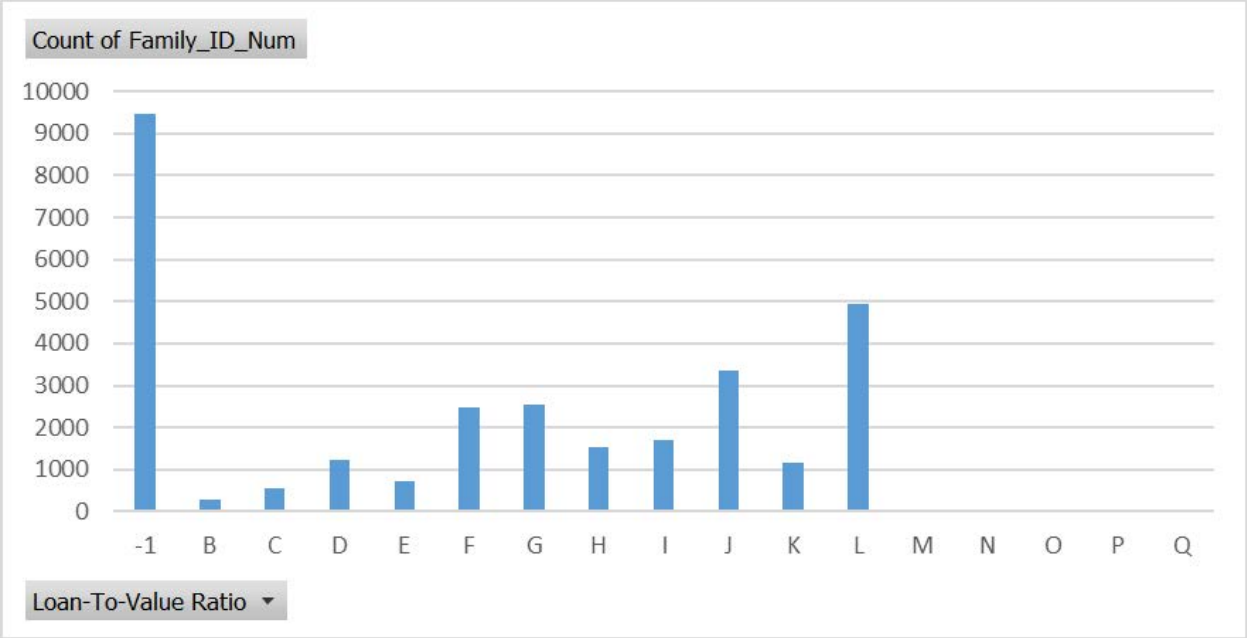

Count of Family_ID_Num
2500
2000
1500
1000
500
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Expendable_Income_Code

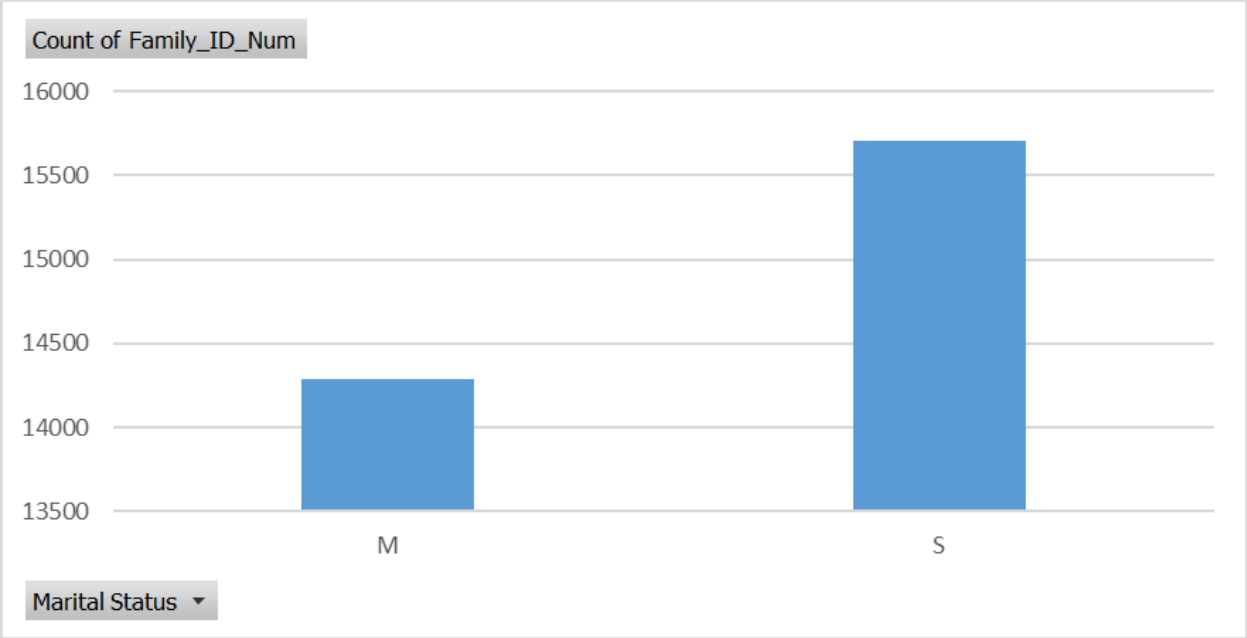

Count of Family_ID_Num
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
Heavy_Internet_User

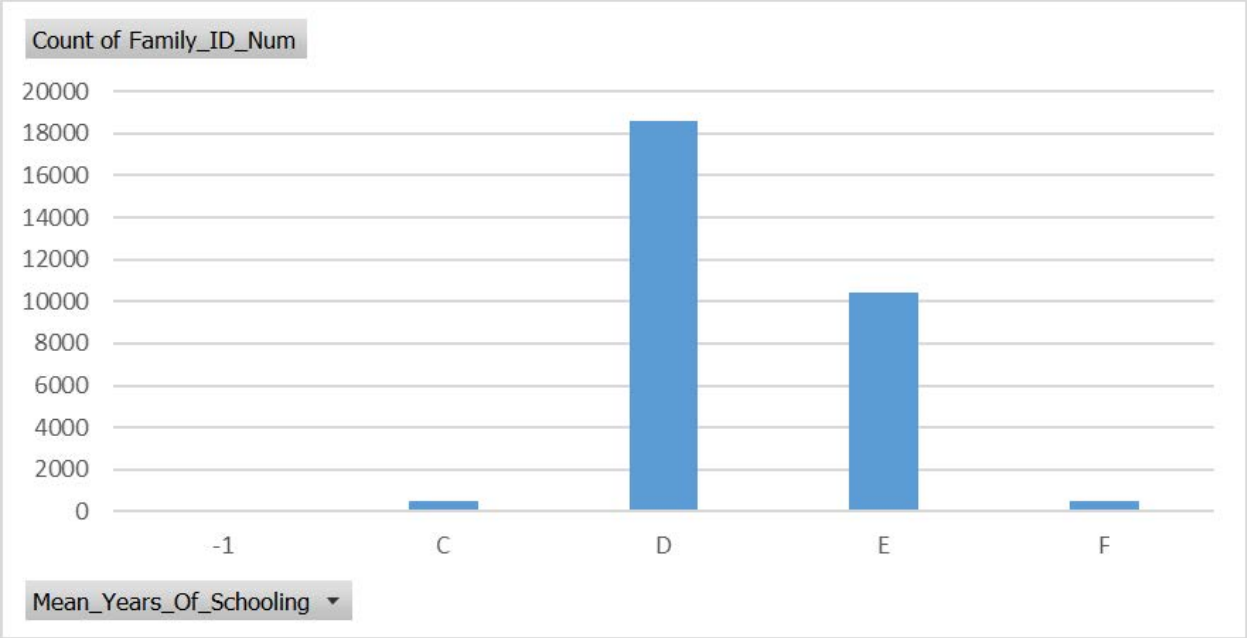

Count of Family_ID_Num
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
-1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Home Equity Estimate

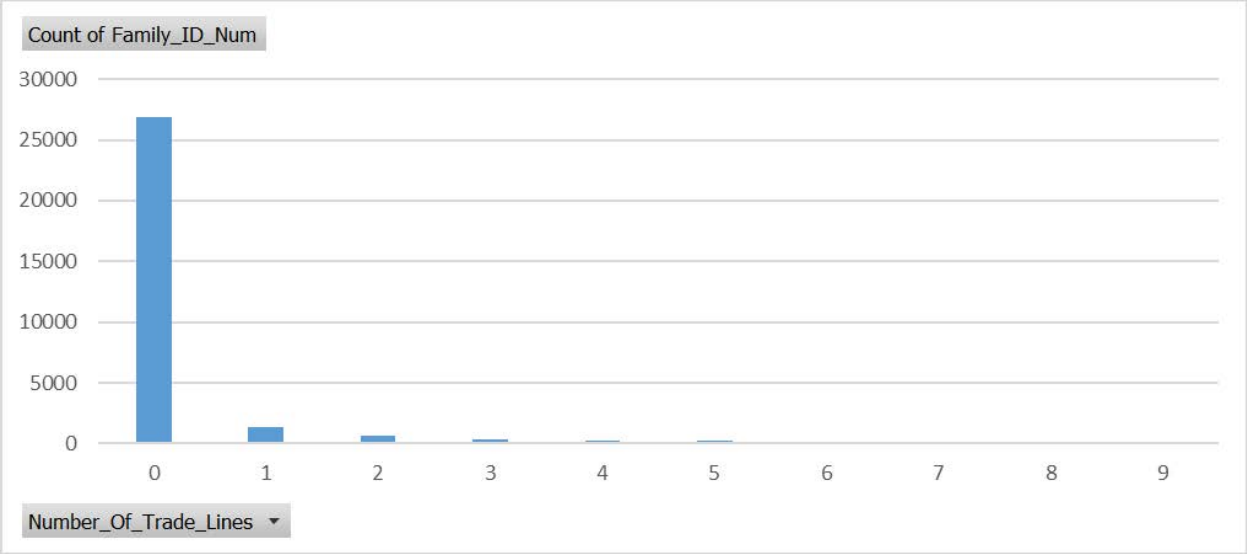

Count of Family_ID_Num
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
-1
6
12
18
24
30
36
42
48
54
60
66
72
78
84
90
97
103
110
116
122
128
135
143
152
158
165
174
182
200
214
223
233
Home_Age

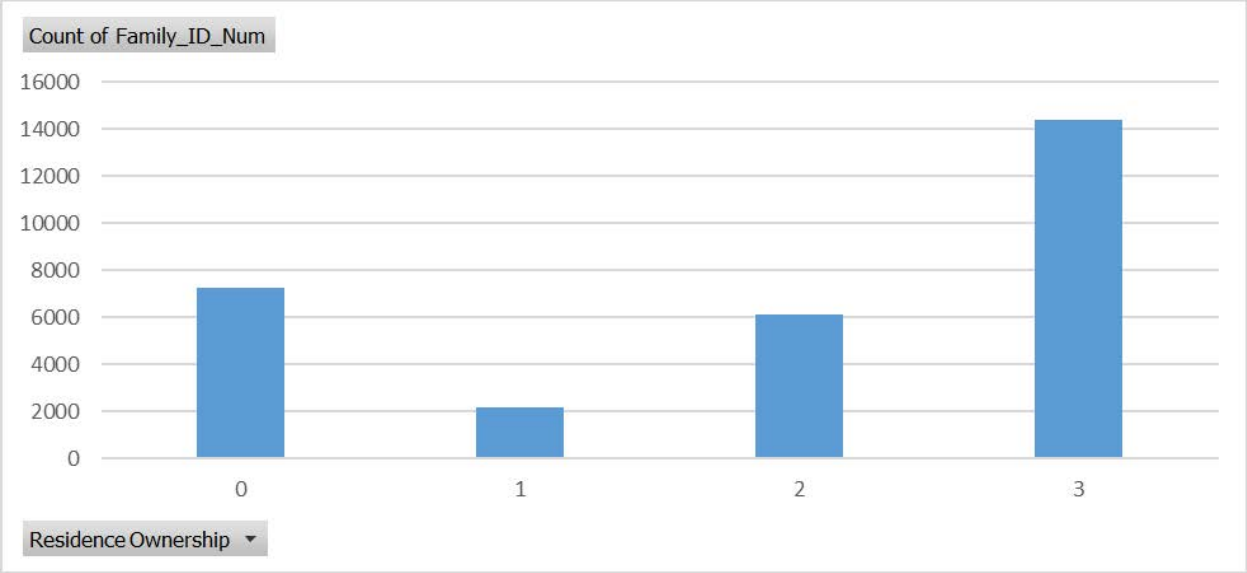

Count of Family_ID_Num
7000
6000
5000
4000
3000
2000
1000
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
Household Income

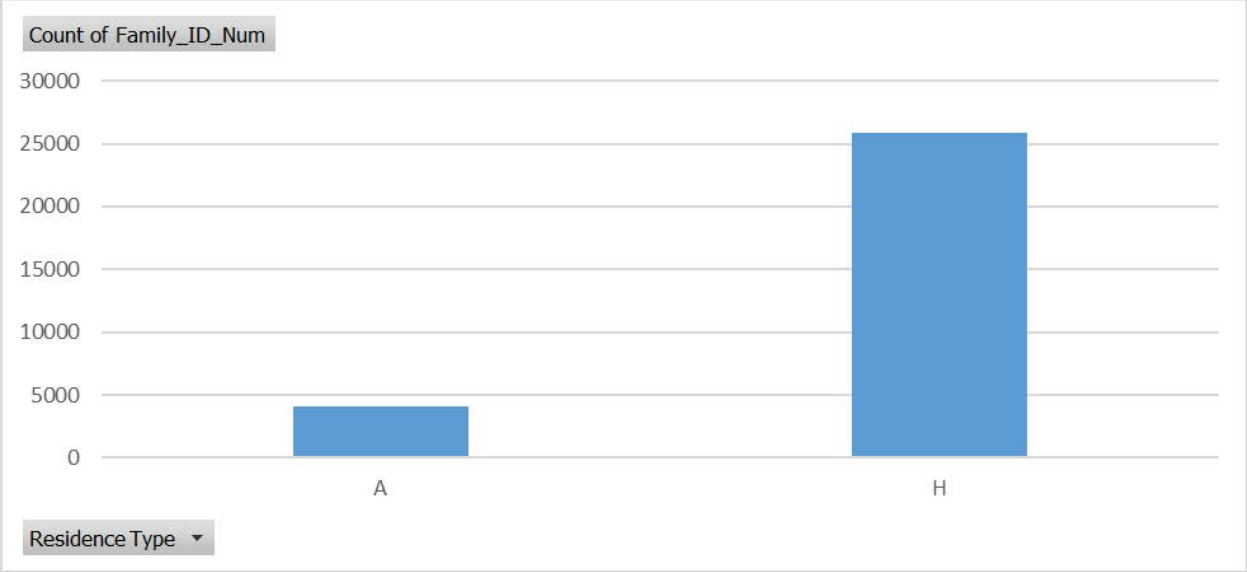

Count of Family_ID_Num
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
-1
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
Loan-To-Value Ratio

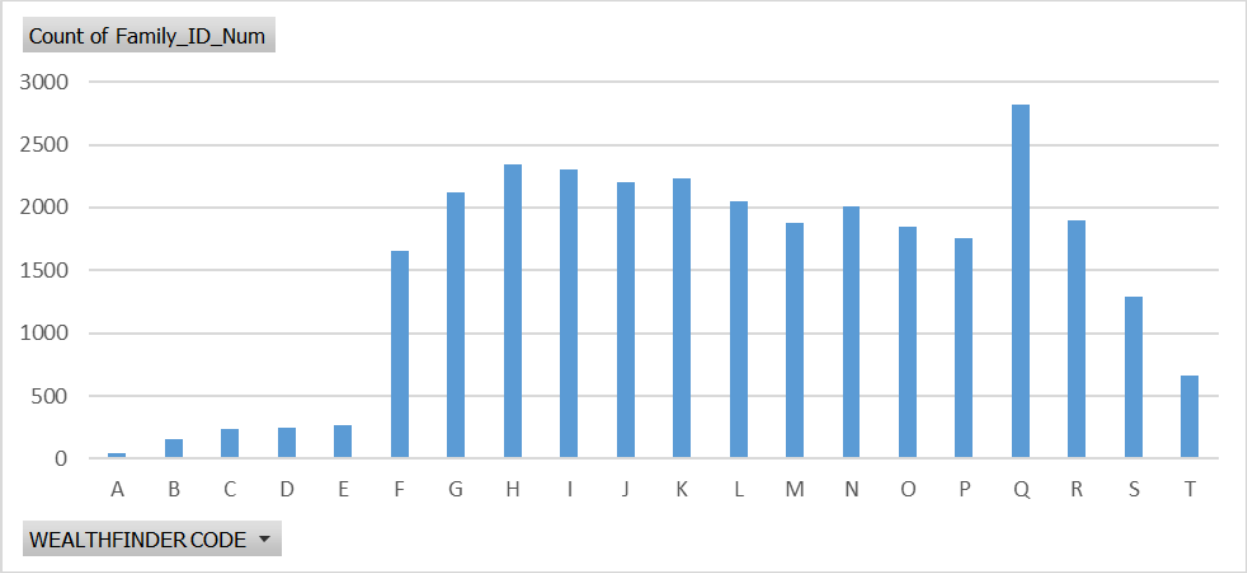

Count of Family_ID_Num
16000
15500
15000
14500
14000
13500
M
S
Marital Status


Count of Family_ID_Num
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
-1
C
D
E
F
Mean_Years_Of_Schooling


Count of Family_ID_Num
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
6
7
8
9
Number_Of_Trade_Lines

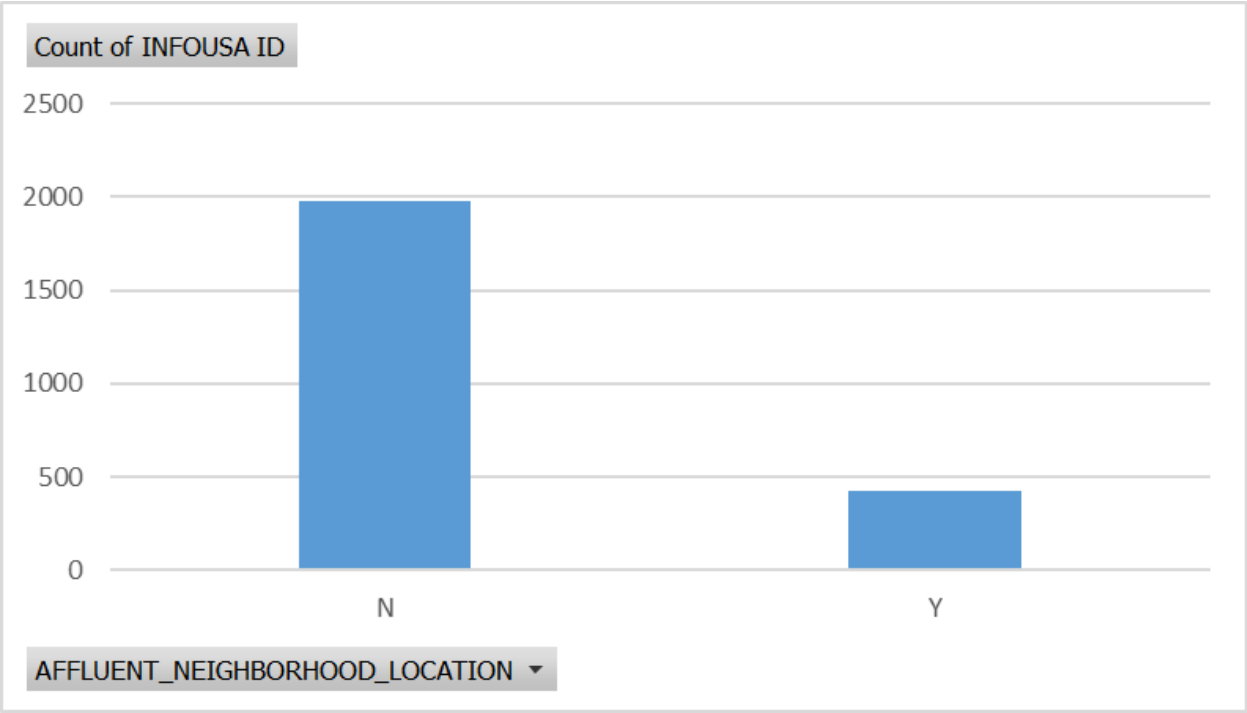

Count of Family_ID_Num
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
Residence Ownership

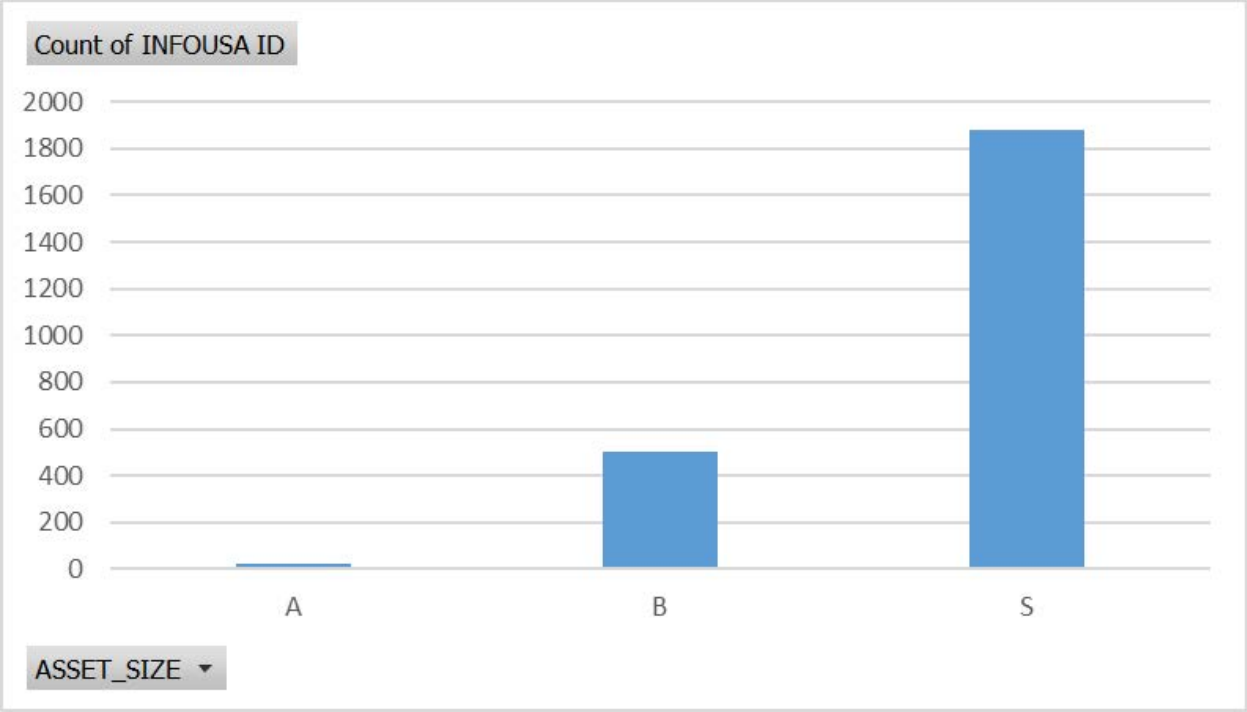

Count of Family_ID_Num
30000
25000
20000
15000
10000
5000
0
A
H
Residence Type

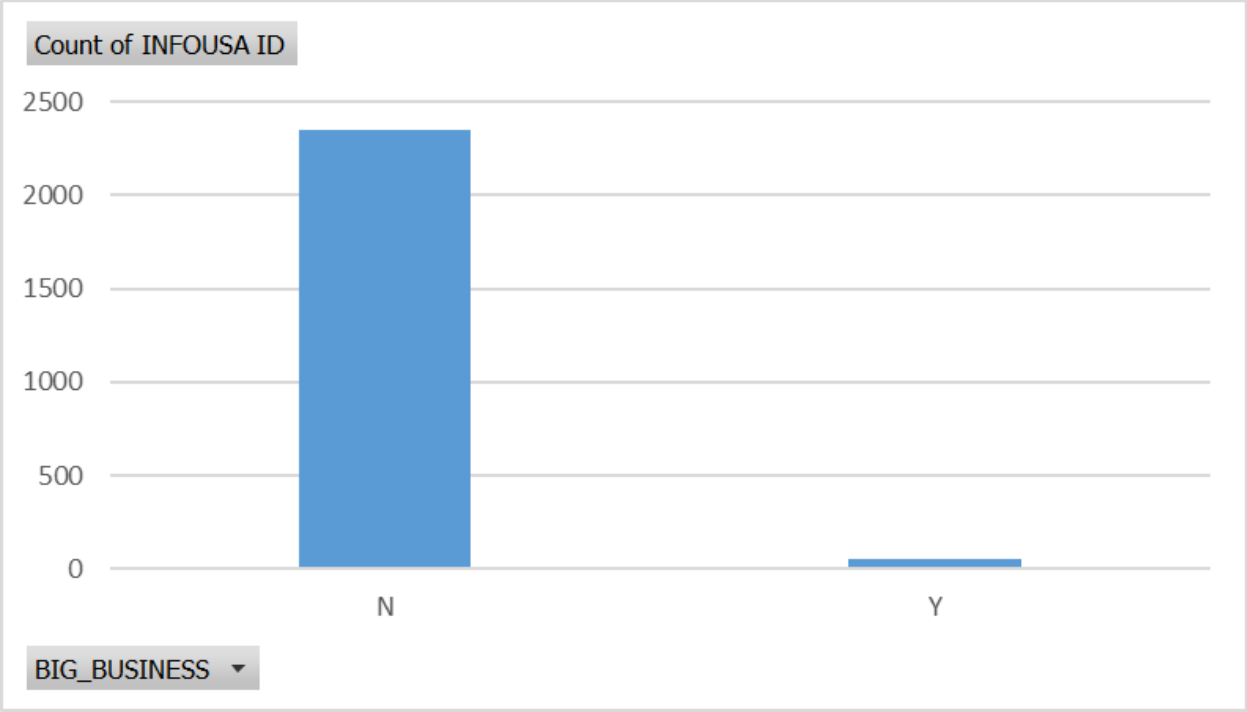

Count of Family_ID_Num
3000
2500
2000
1500
1000
500
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
WEALTHFINDER CODE

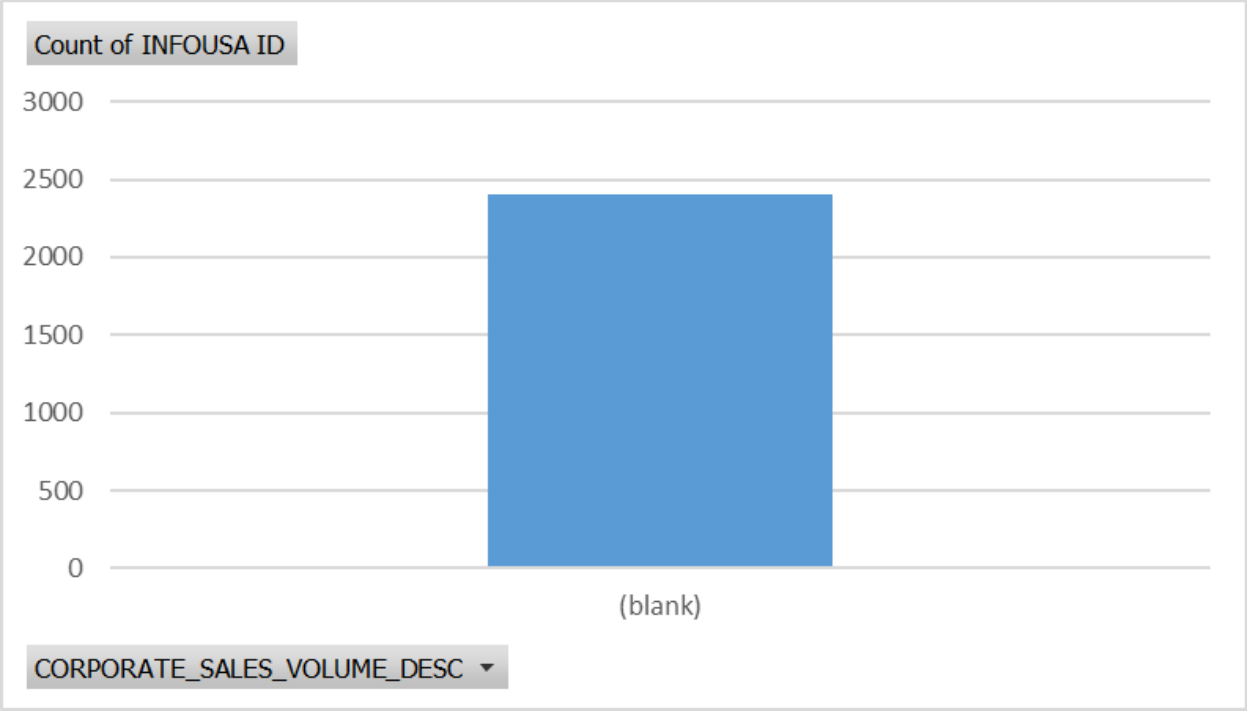
InfoGroup 2017 Business Data Charts

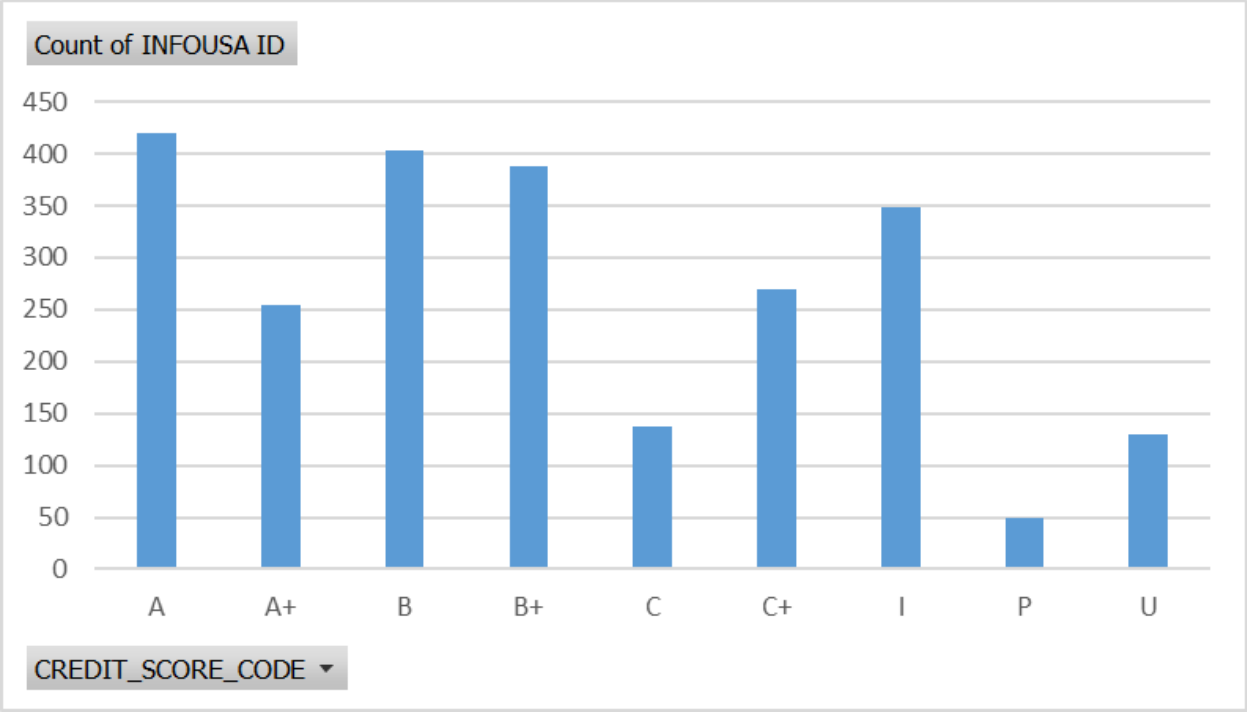
NAICS Code not shown

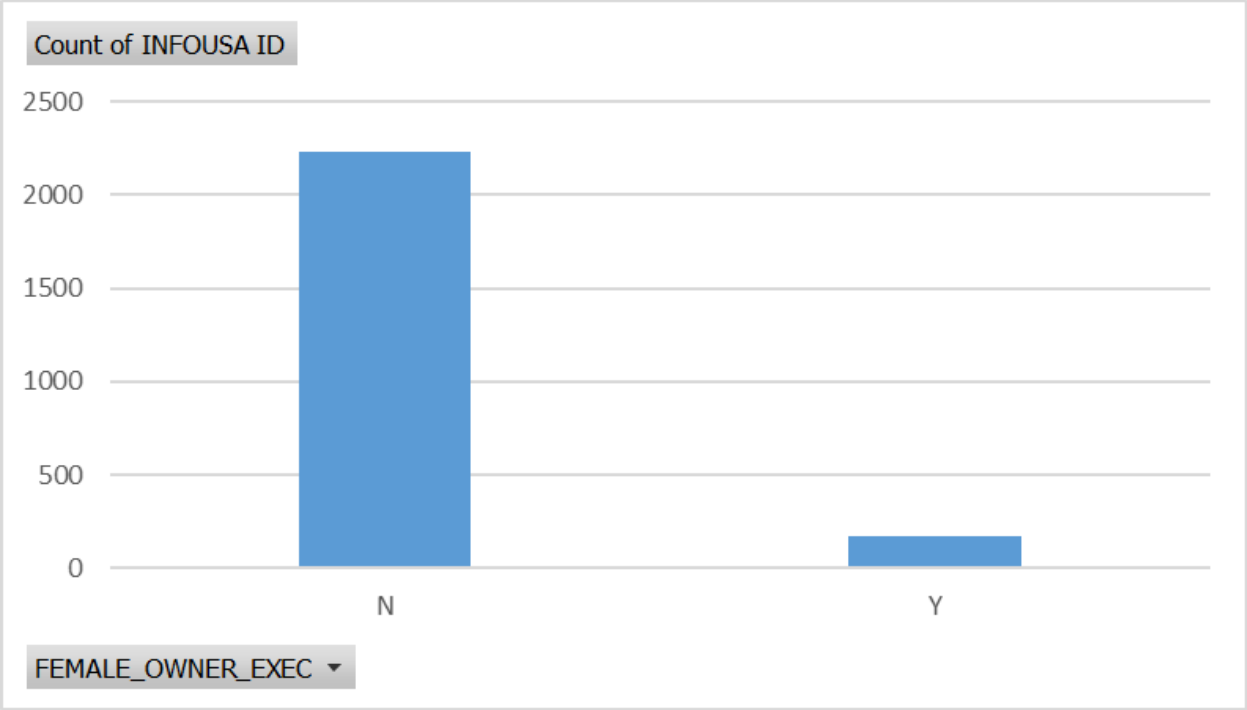


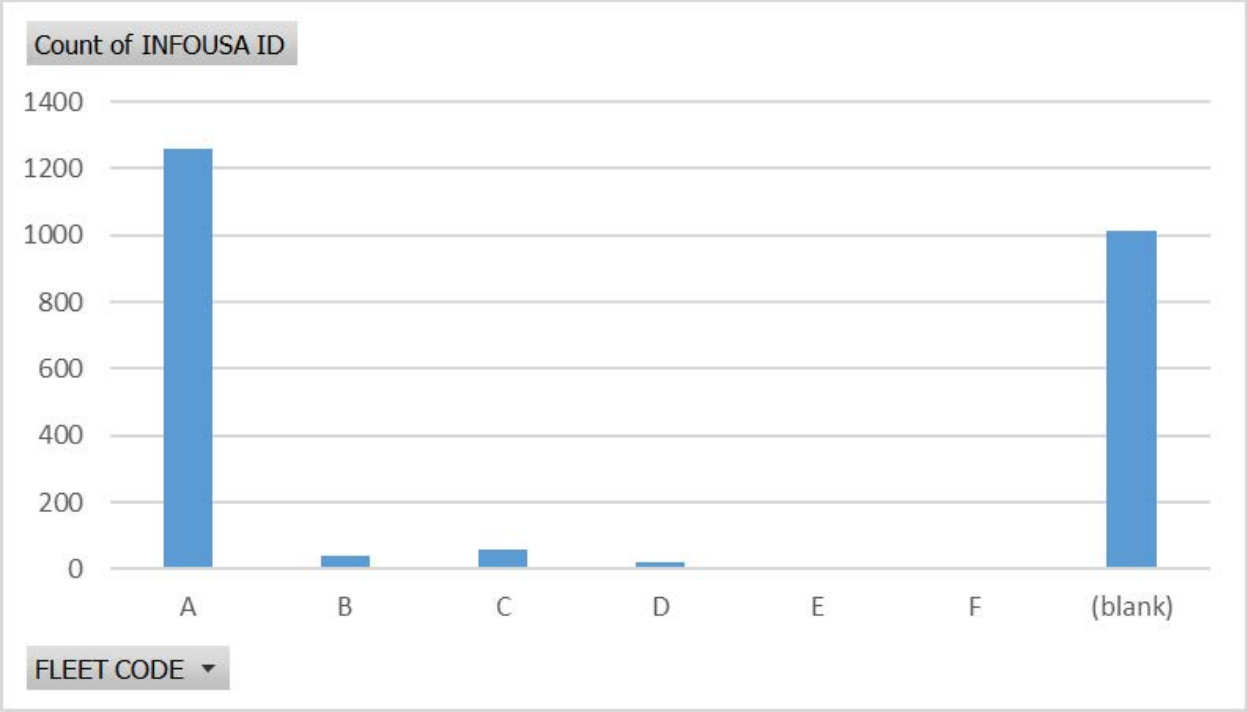


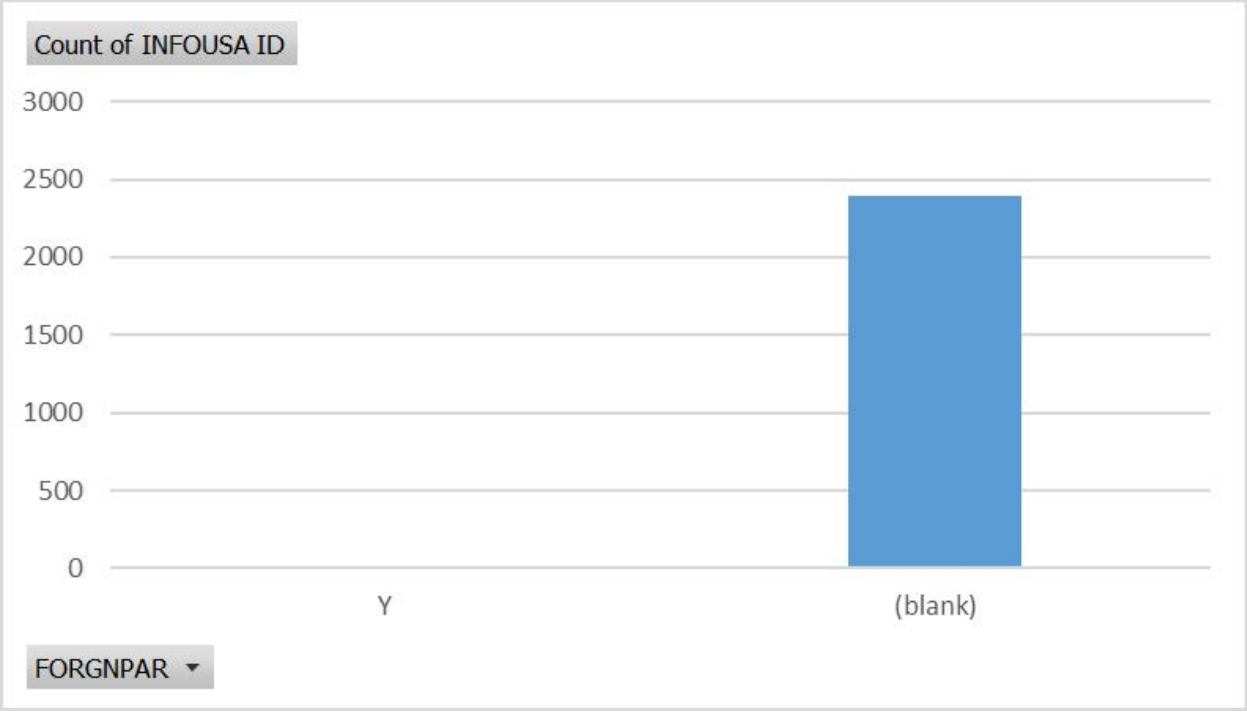

Count of INFOUSA ID
2500
2000
1500
1000
500
0
N
Y
BIG_BUSINESS

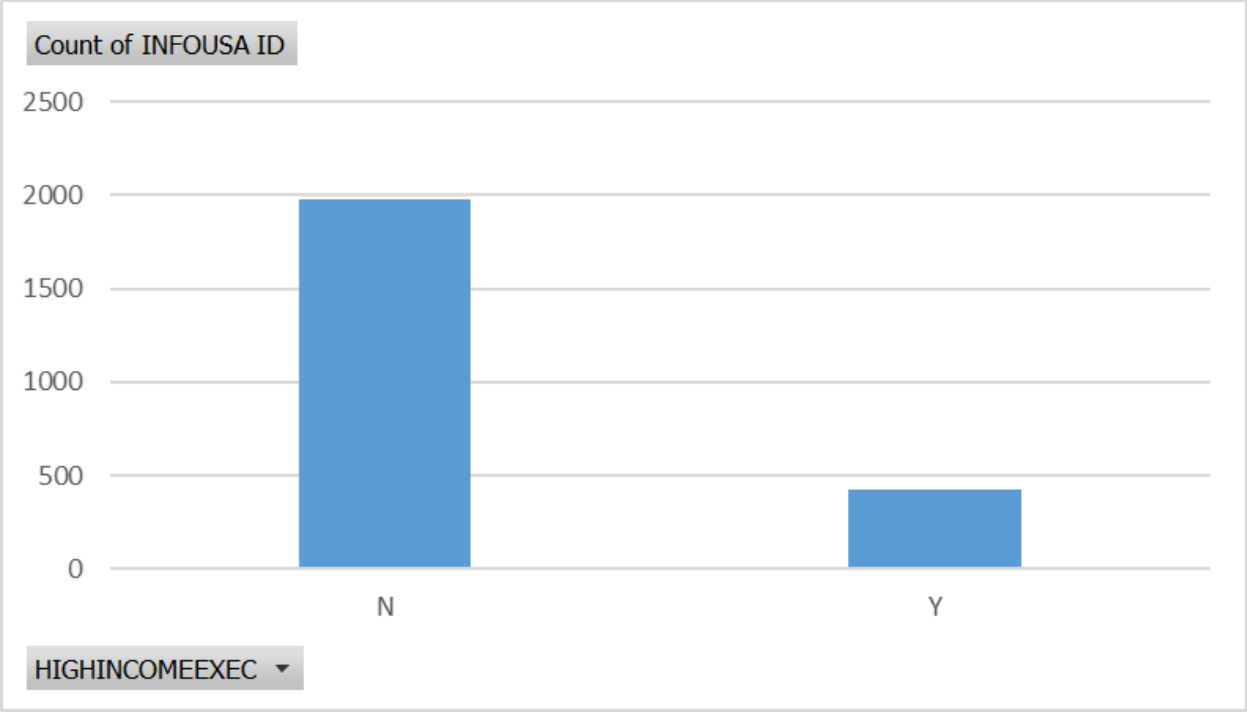

Count of INFOUSA ID
3000
2500
2000
1500
1000
500
0
(blank)
CORPORATE_SALES_VOLUME_DESC

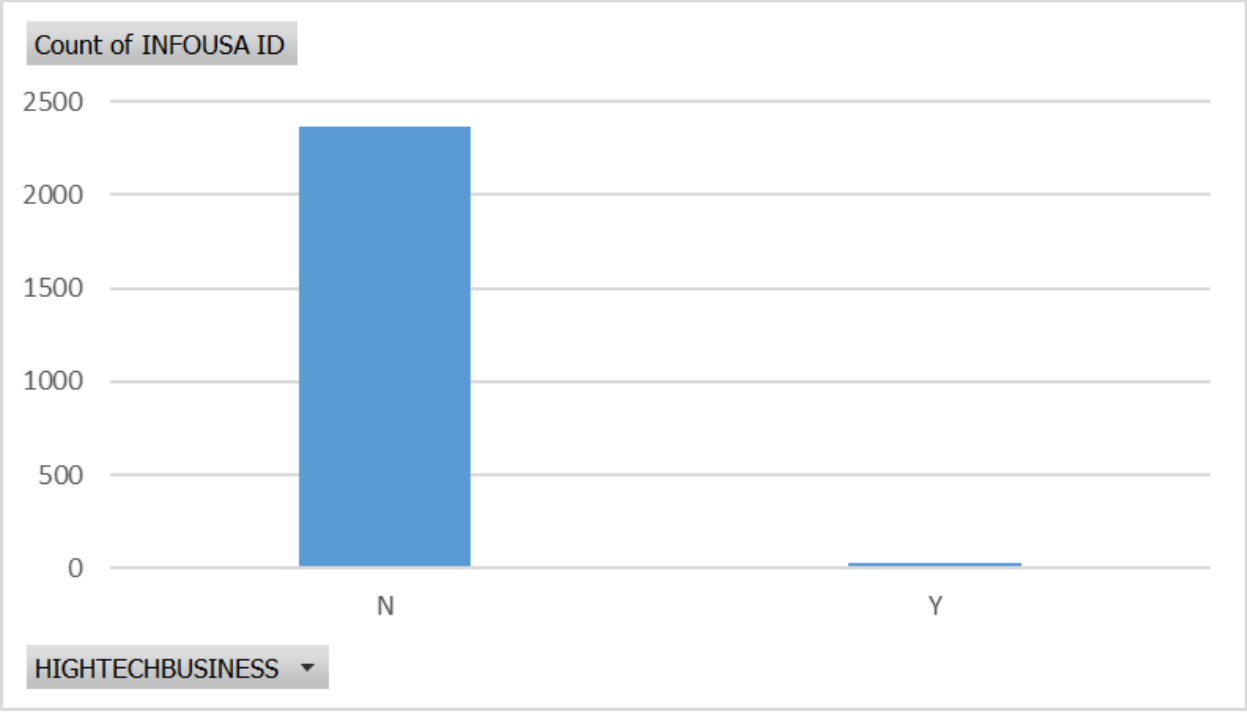

Count of INFOUSA ID
450
400
350
300
250
200
150
100
50
0
A
A+
B
B+
C
C+
I
P
U
CREDIT_SCORE_CODE

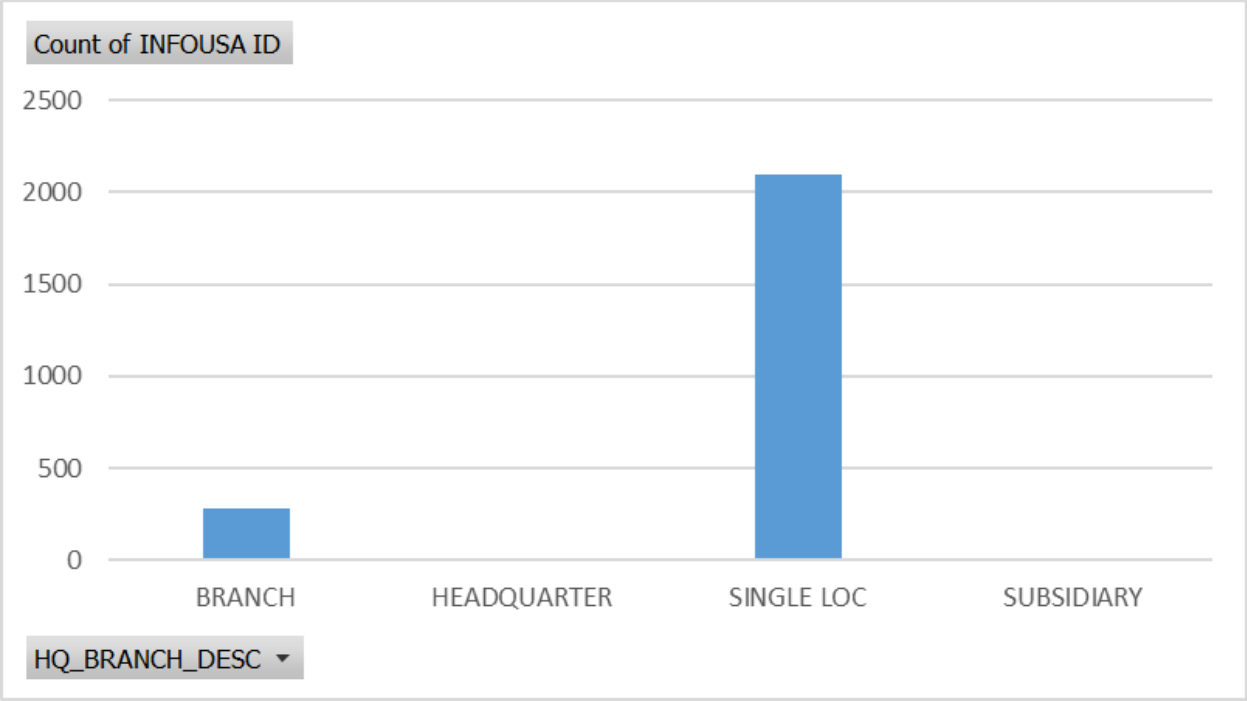

Count of INFOUSA ID
2500
2000
1500
1000
500
0
N
Y
FEMALE_OWNER_EXEC

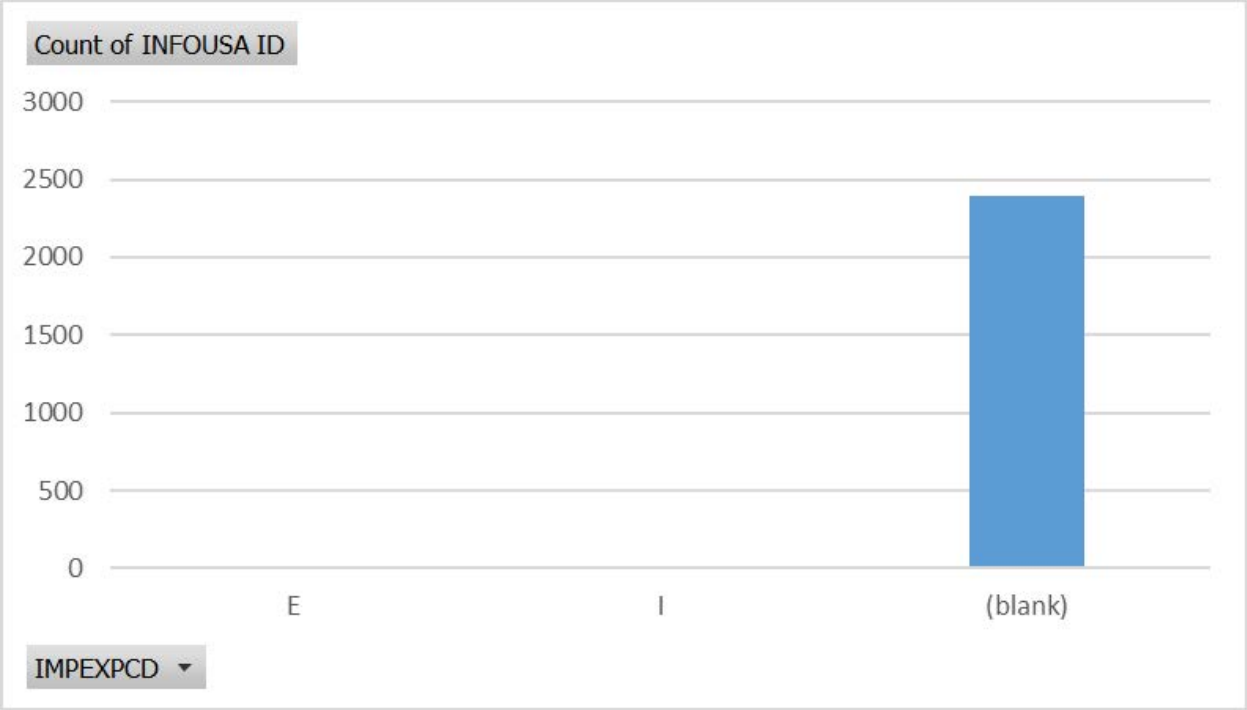

Count of INFOUSA ID
1400
1200
1000
800
600
400
200
0
A
B
C
D
E
F
(blank)
FLEET CODE

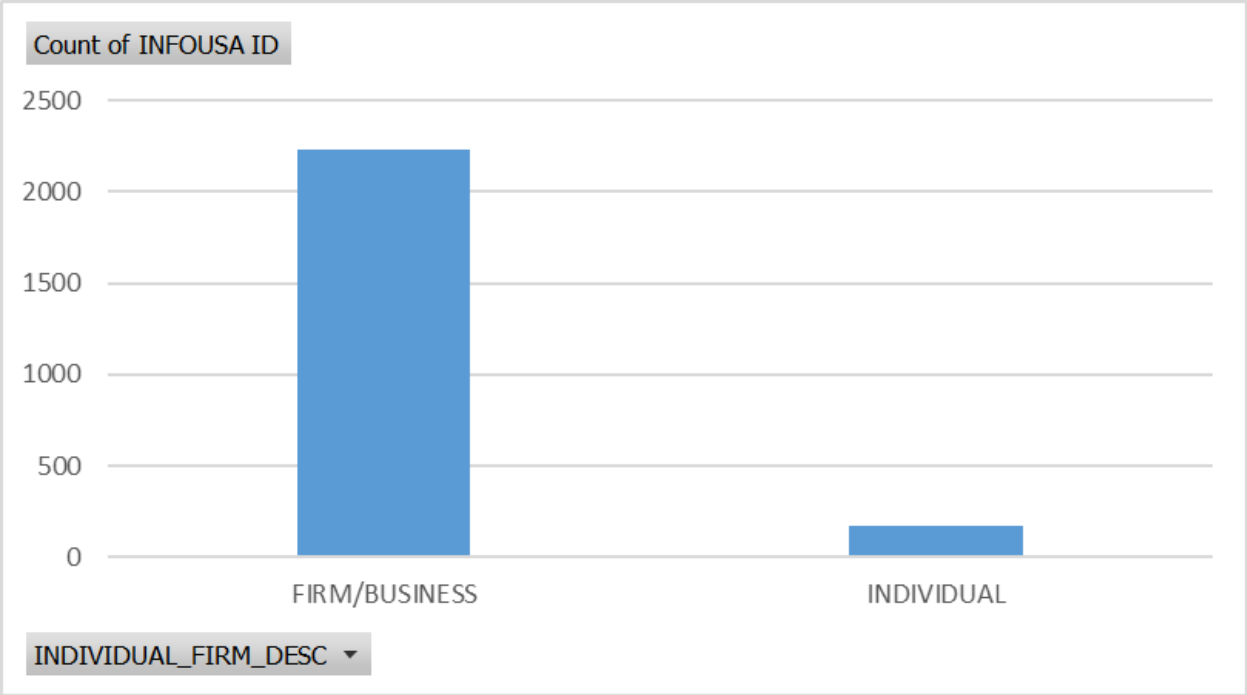

Count of INFOUSA ID
3000
2500
2000
1500
1000
500
0
Y
(blank)
FORGNPAR

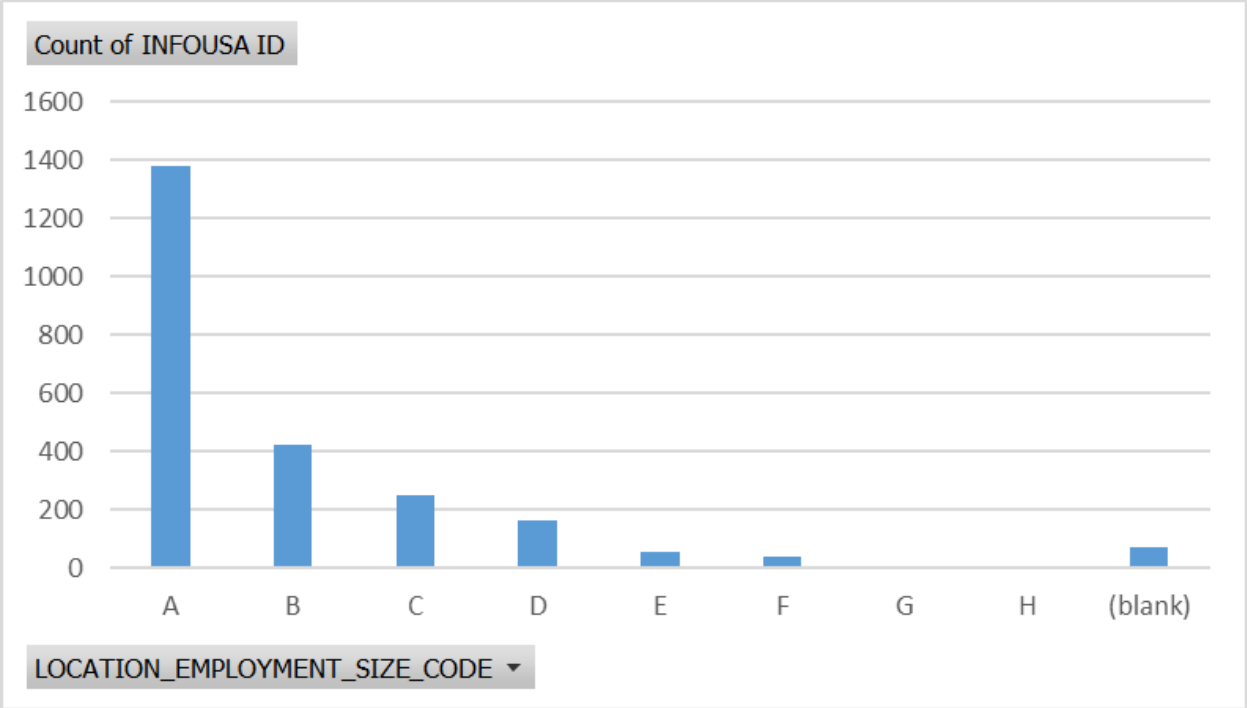

Count of INFOUSA ID
2500
2000
1500
1000
500
0
N
Y
HIGHINCOMEEXEC

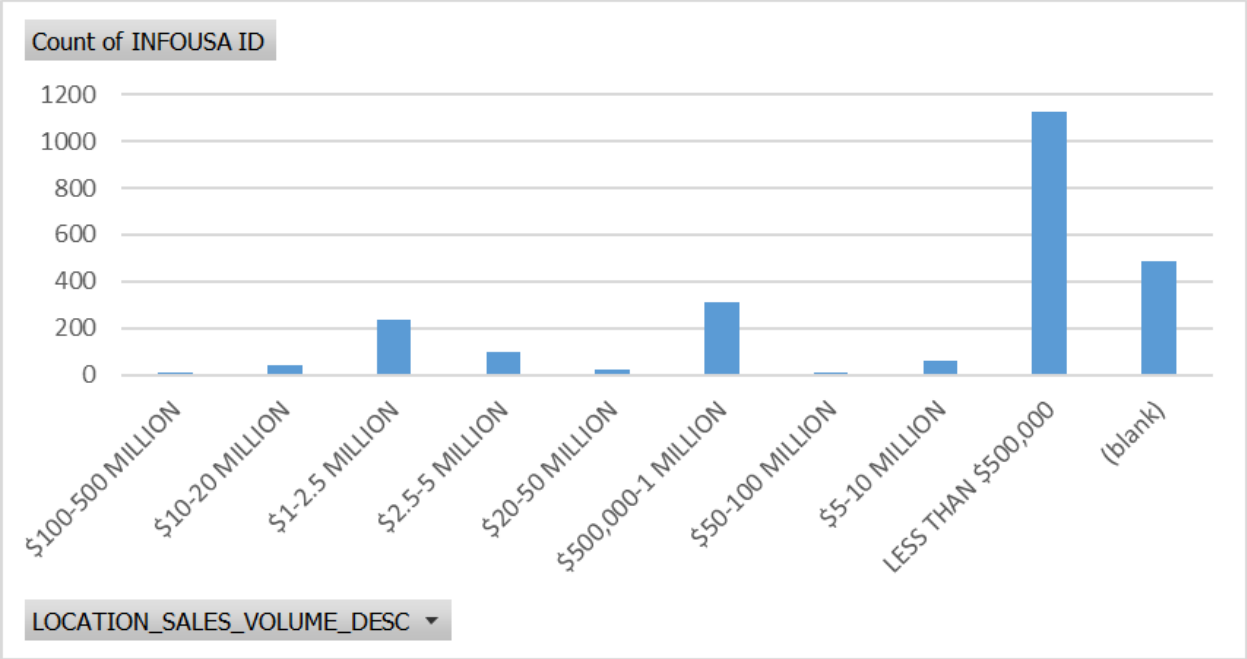

Count of INFOUSA ID
2500
2000
1500
1000
500
0
N
Y
HIGHTECHBUSINESS

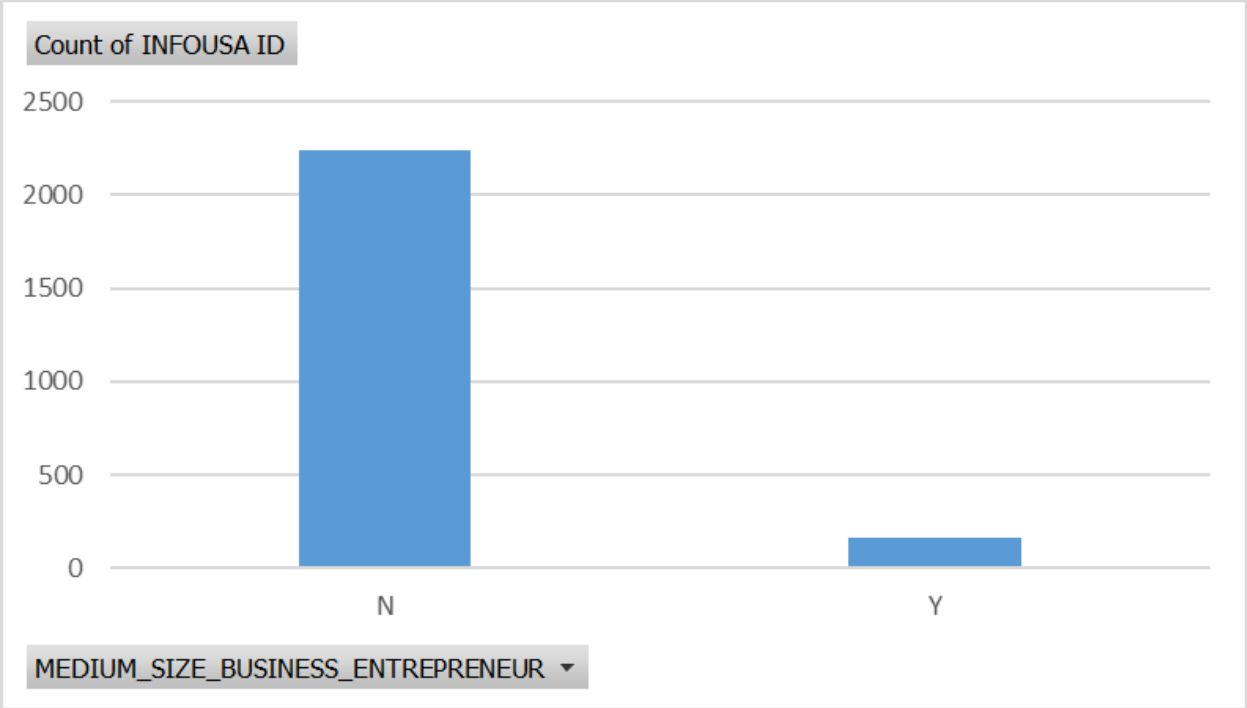

Count of INFOUSA ID
2500
2000
1500
1000
500
0
BRANCH
HEADQUARTER
SINGLE LOC
SUBSIDIARY
HQ_BRANCH_DESC

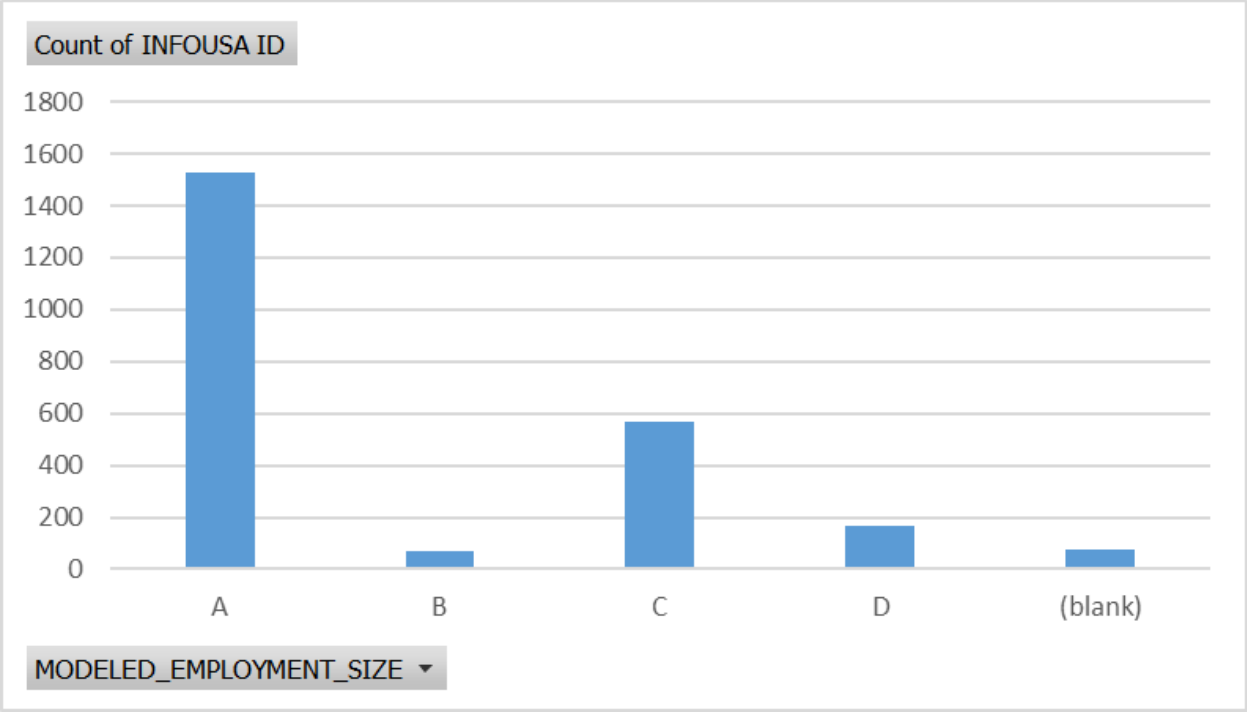

Count of INFOUSA ID
3000
2500
2000
1500
1000
500
0
E
I
(blank)
IMPEXPCD

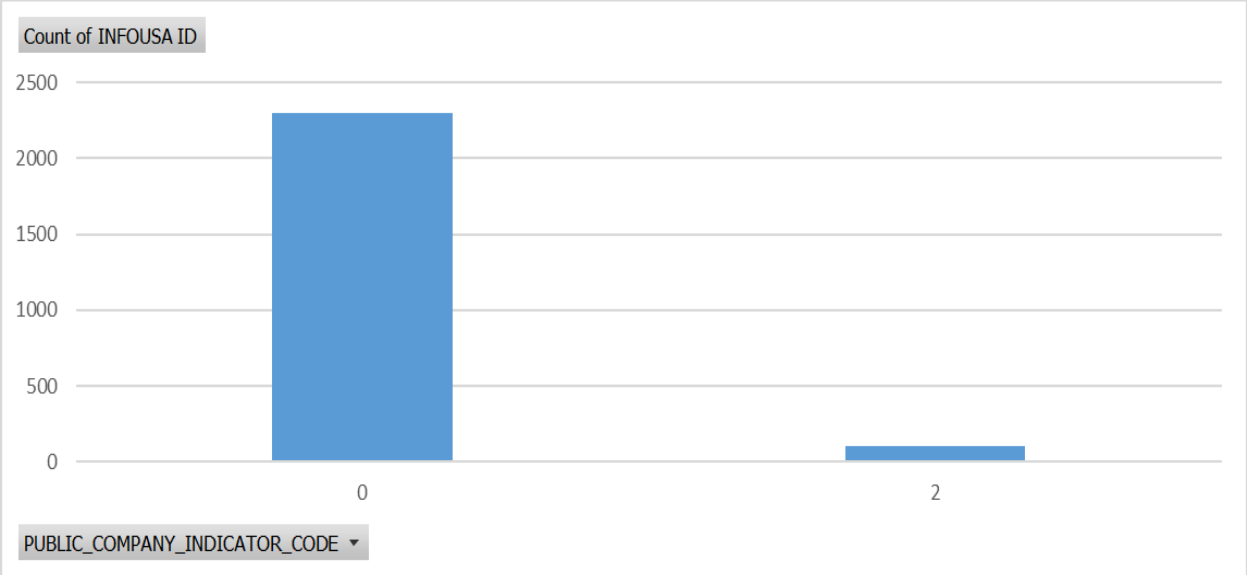

Count of INFOUSA ID
2500
2000
1500
1000
500
0
FIRM/BUSINESS
INDIVIDUAL
INDIVIDUAL_FIRM_DESC

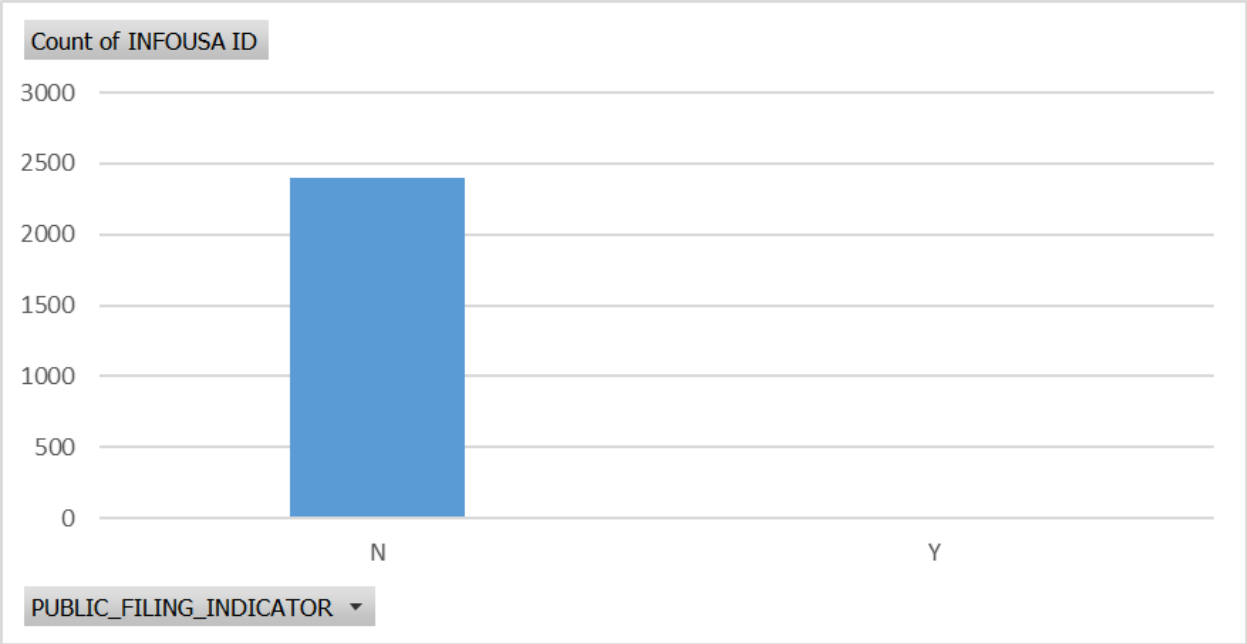

Count of INFOUSA ID
1600
1400
1200
1000
800
600
400
200
0
A
B
C
D
E
F
G
H
(blank)
LOCATION_EMPLOYMENT_SIZE_CODE

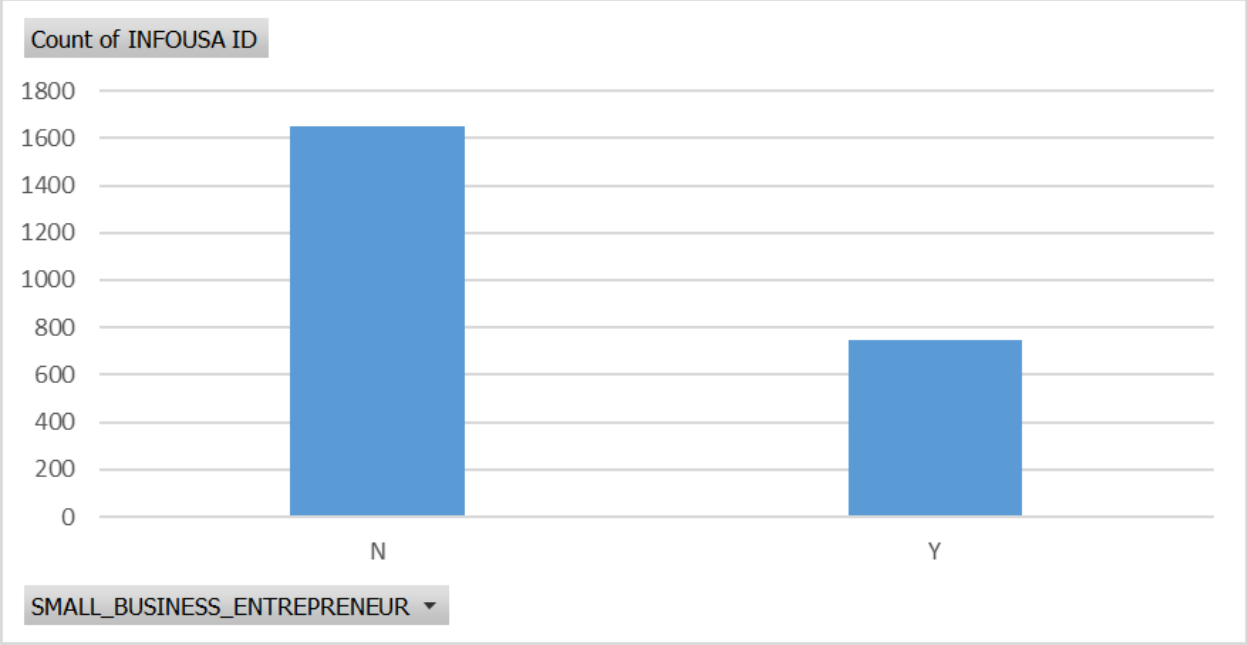

Count of INFOUSA ID
1200
1000
800
600
400
200
0
$100-500 MILLION
$10-20 MILLION
$1-2.5 MILLION
$2.5-5 MILLION
$20-50 MILLION
$500,000-1 MILLION
$50-100 MILLION
$5-10 MILLION
LESS THAN $500,000
(blank)
LOCATION_SALES_VOLUME_DESC

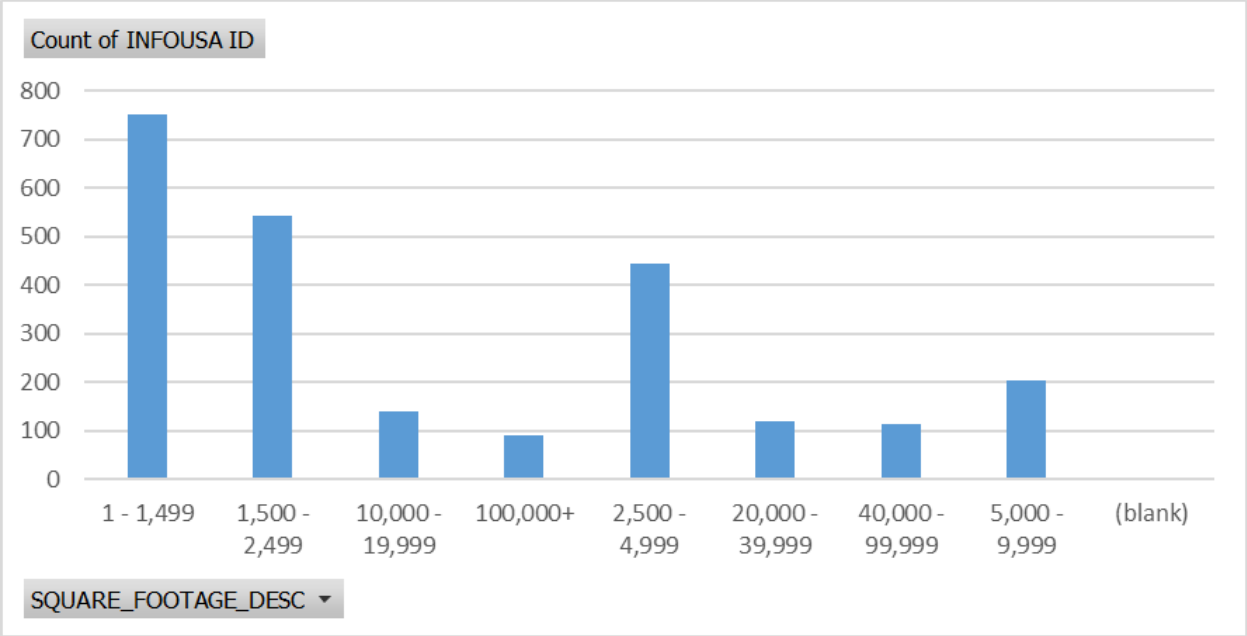

Count of INFOUSA ID
2500
2000
1500
1000
500
0
N
Y
MEDIUM_SIZE_BUSINESS_ENTREPRENEUR

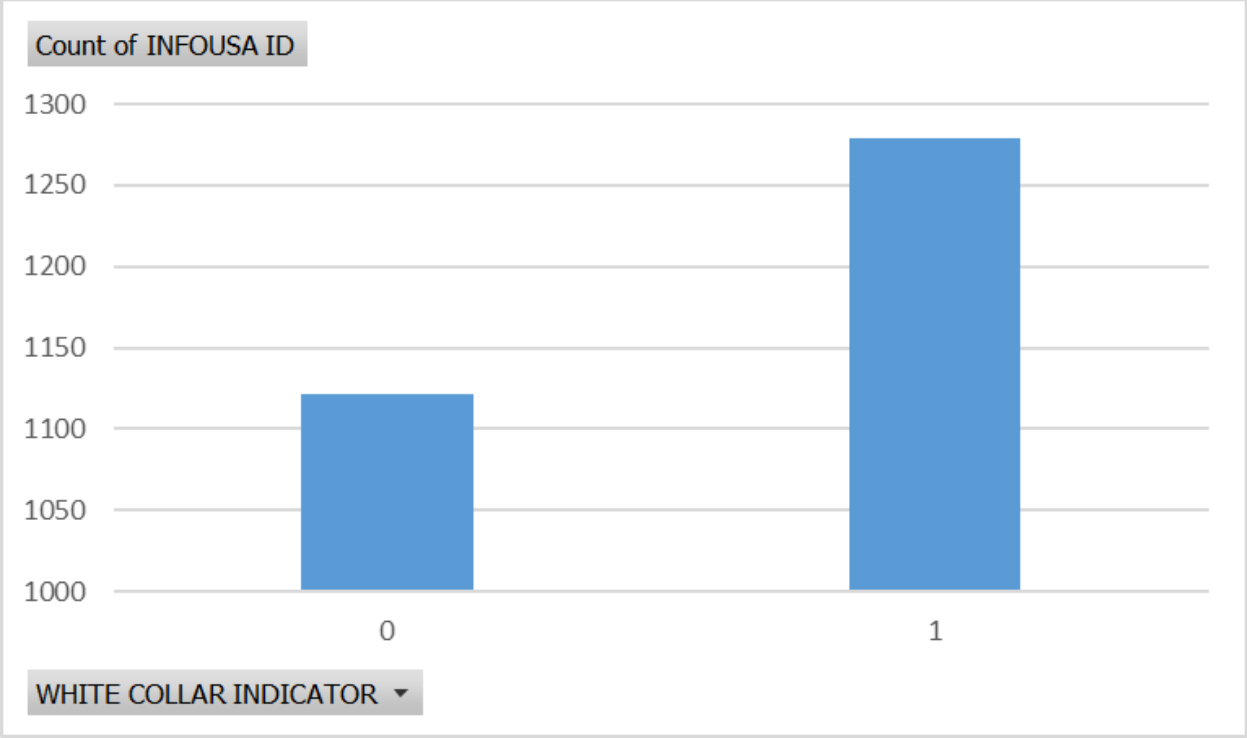

Count of INFOUSA ID
1800
1600
1400
1200
1000
800
600
400
200
0
A
B
C
D
(blank)
MODELED_EMPLOYMENT_SIZE

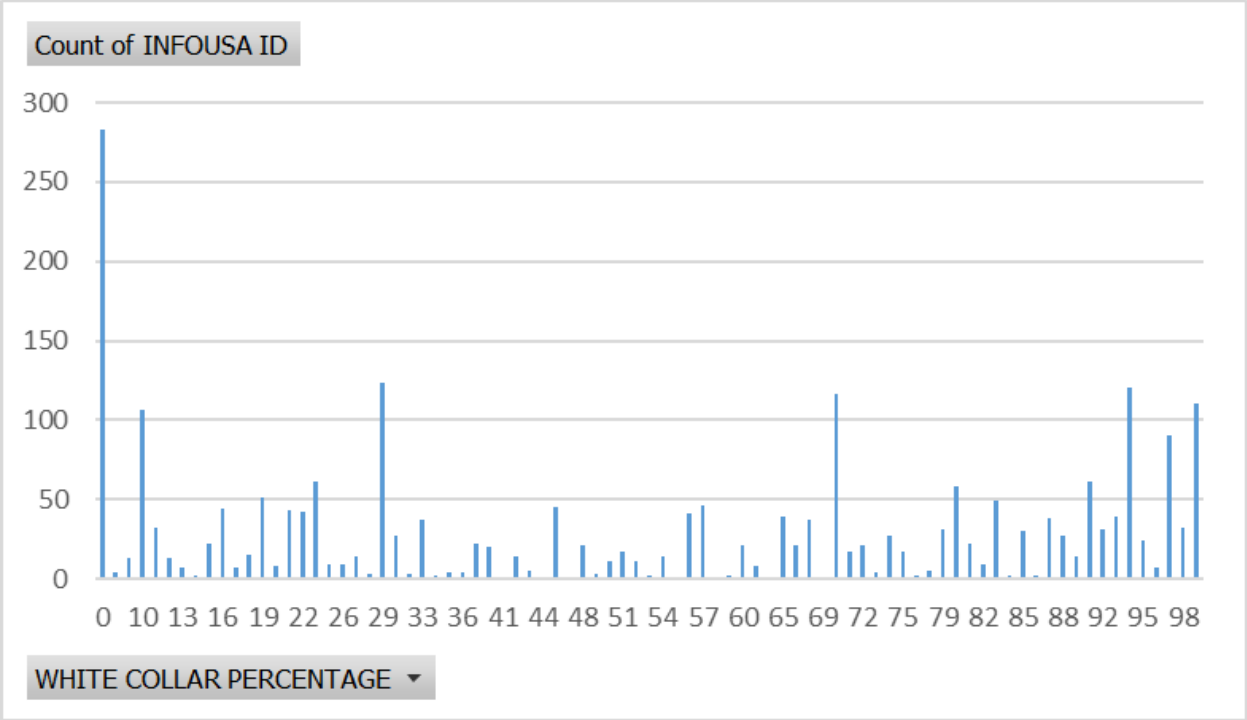

Count of INFOUSA ID
2500
2000
1500
1000
500
0
0
2
PUBLIC_COMPANY_INDICATOR_CODE

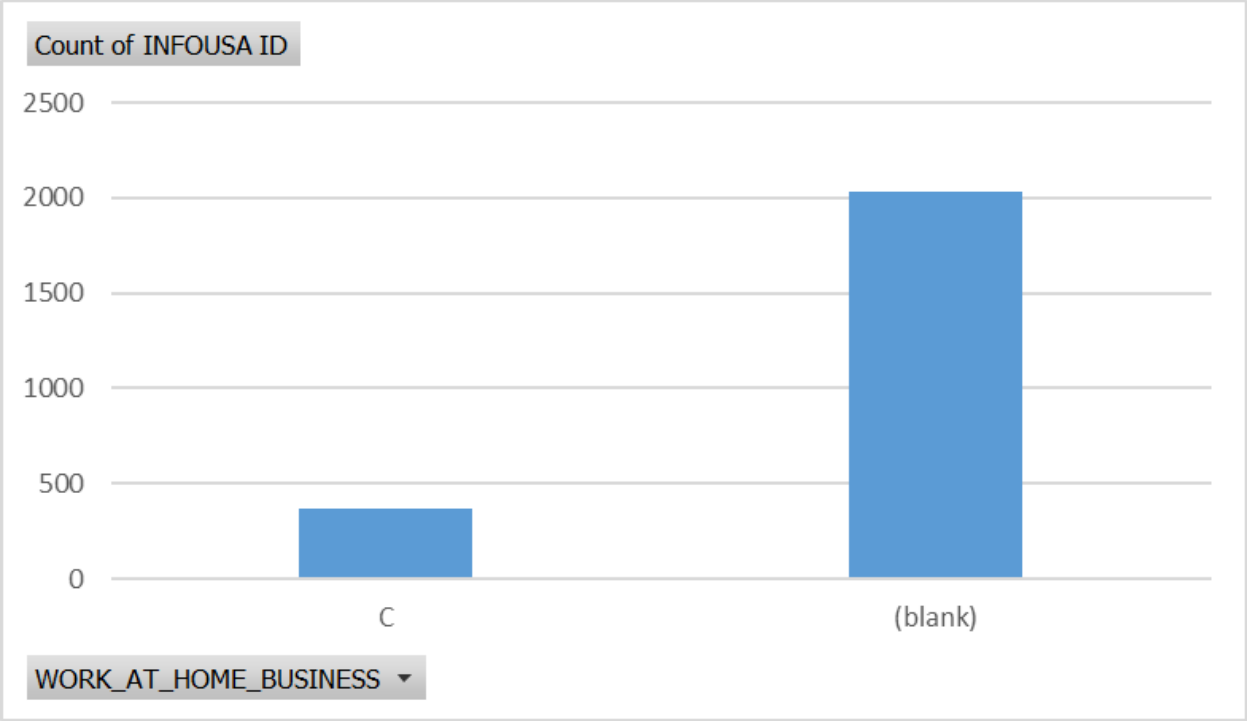

Count of INFOUSA ID
3000
2500
2000
1500
1000
500
0
N
Y
PUBLIC_FILING_INDICATOR


Count of INFOUSA ID
1800
1600
1400
1200
1000
800
600
400
200
0
N
Y
SMALL_BUSINESS_ENTREPRENEUR


Count of INFOUSA ID
800
700
600
500
400
300
200
100
0
1 - 1,499
1,500 - 2,499
10,000 - 19,999
100,000+
2,500 - 4,999
20,000 - 39,999
40,000 - 99,999
5,000 - 9,999
(blank)
SQUARE_FOOTAGE_DESC

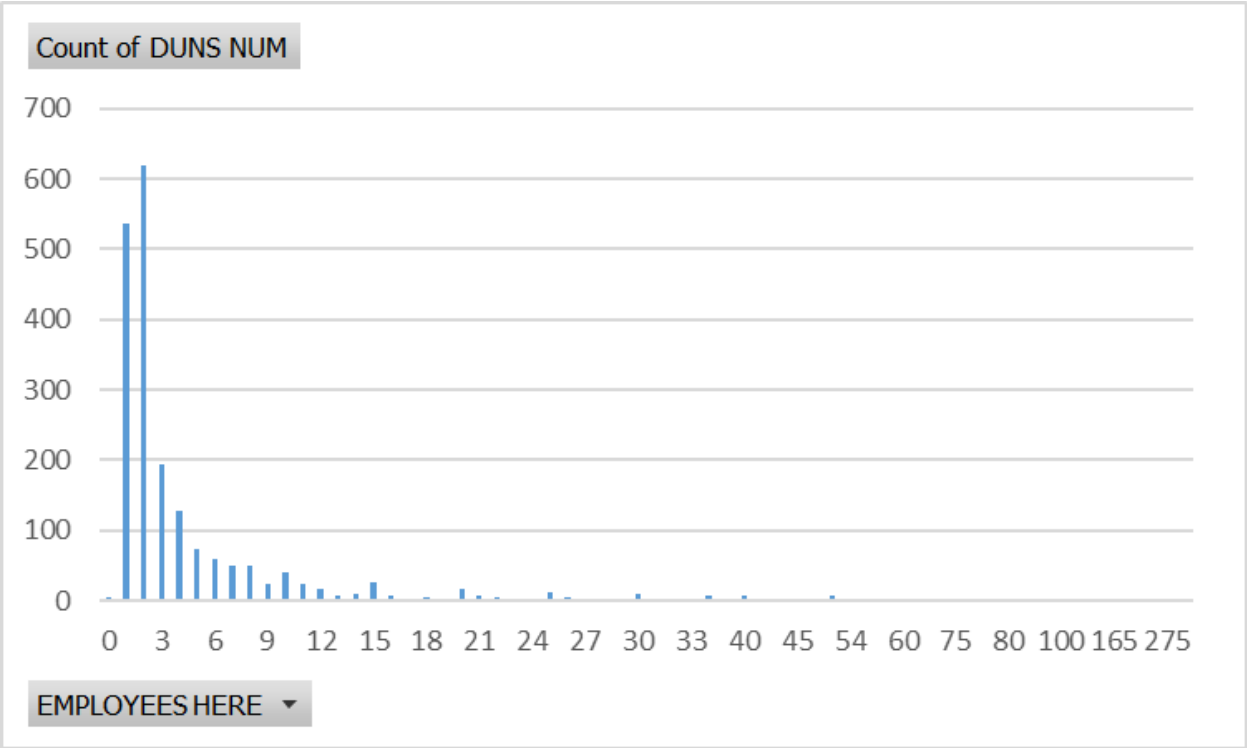

Count of INFOUSA ID
1300
1250
1200
1150
1100
1050
1000
0
1
WHITE COLLAR INDICATOR

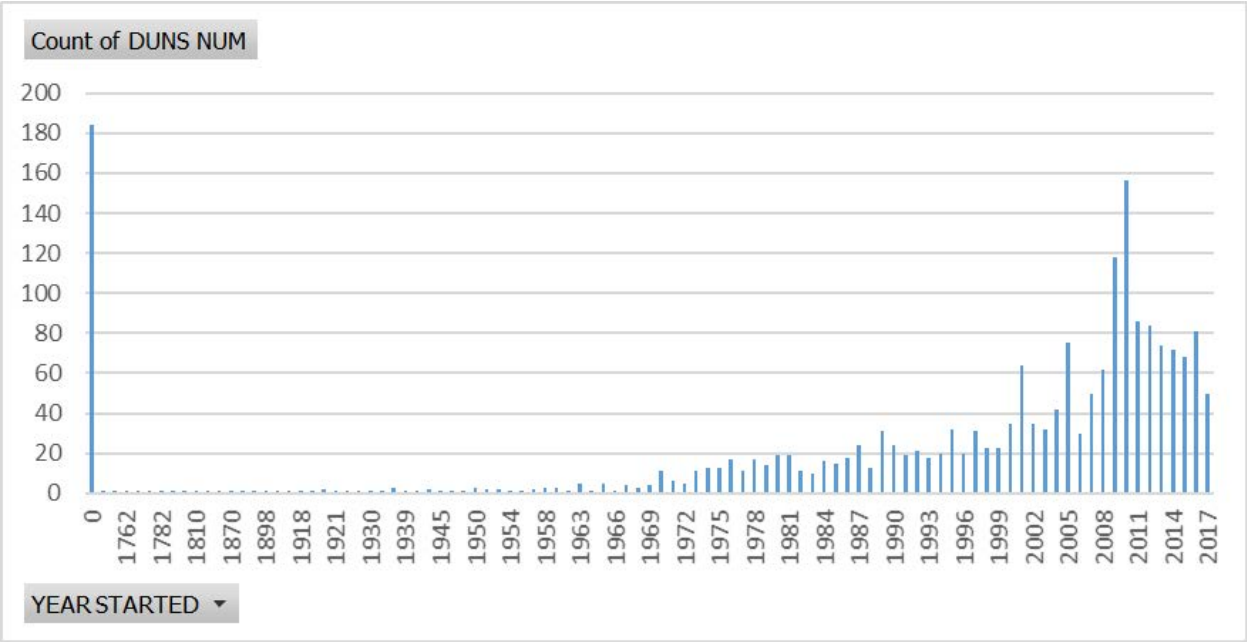

Count of INFOUSA ID
300
250
200
150
100
50
0
0 10 13 16 19 22 26 29 33 36 41 44 48 51 54 57 60 65 69 72 75 79 82 85 88 92 95 98
WHITE COLLAR PERCENTAGE

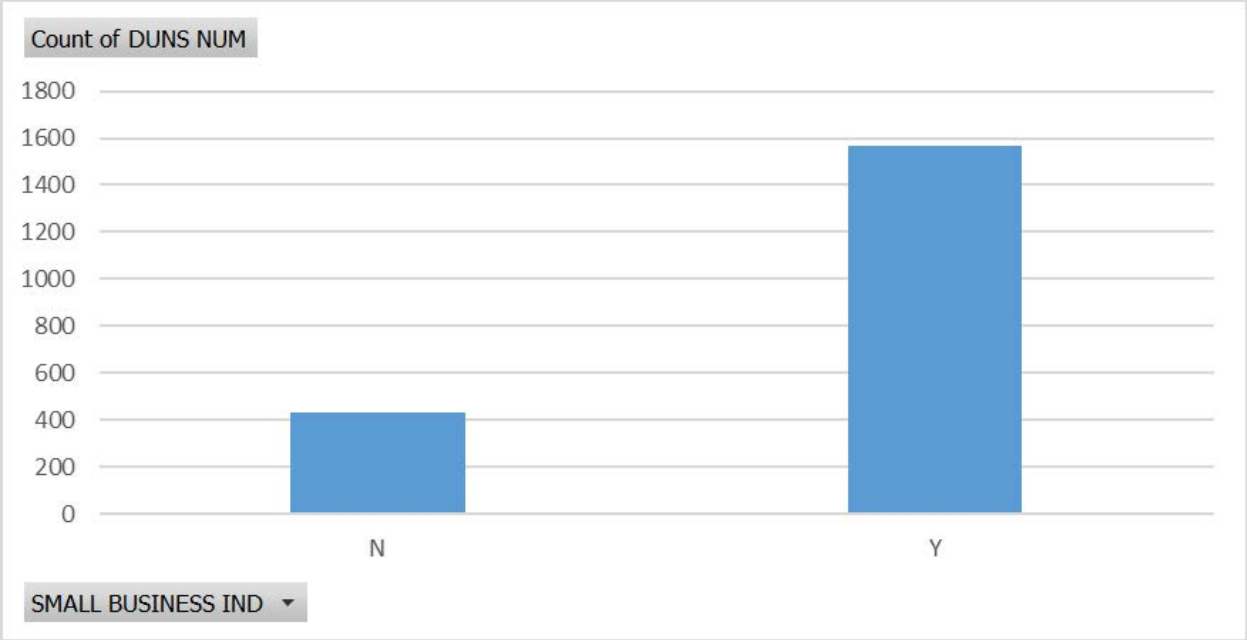

Count of INFOUSA ID
2500
2000
1500
1000
500
0
C
(blank)
WORK_AT_HOME_BUSINESS

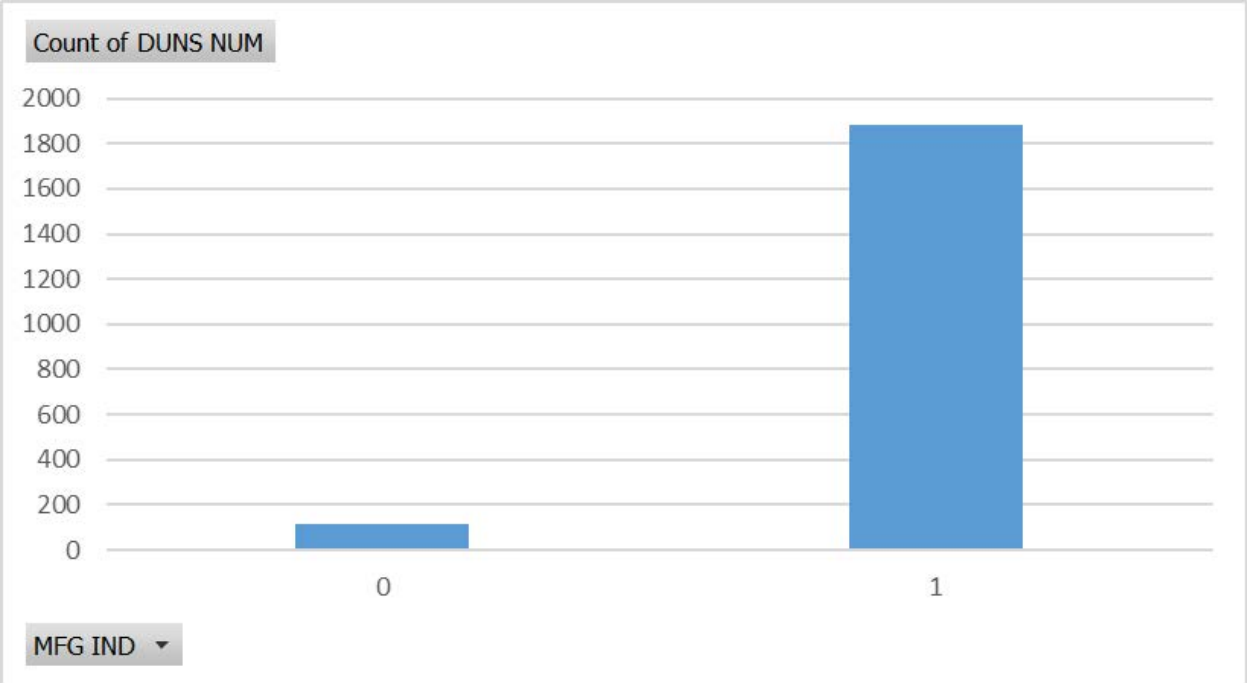
Dun & Bradstreet 2017 Business Data Charts

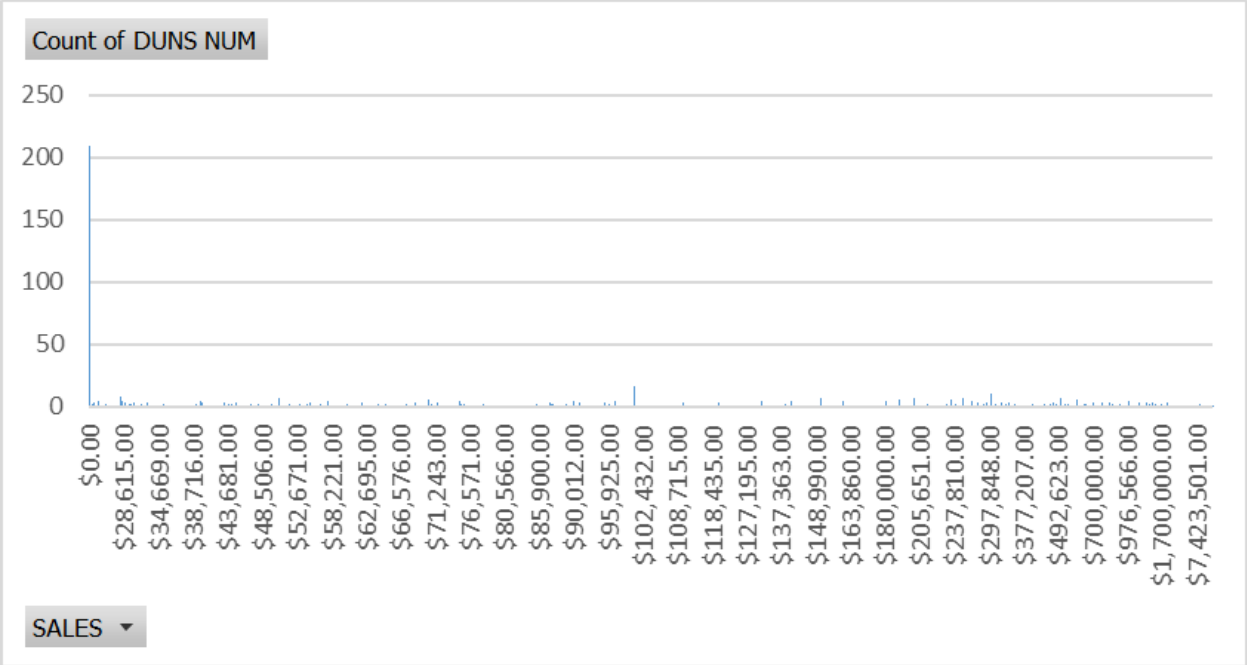
Company Name, Primary NAICS/SIC Codes not shown

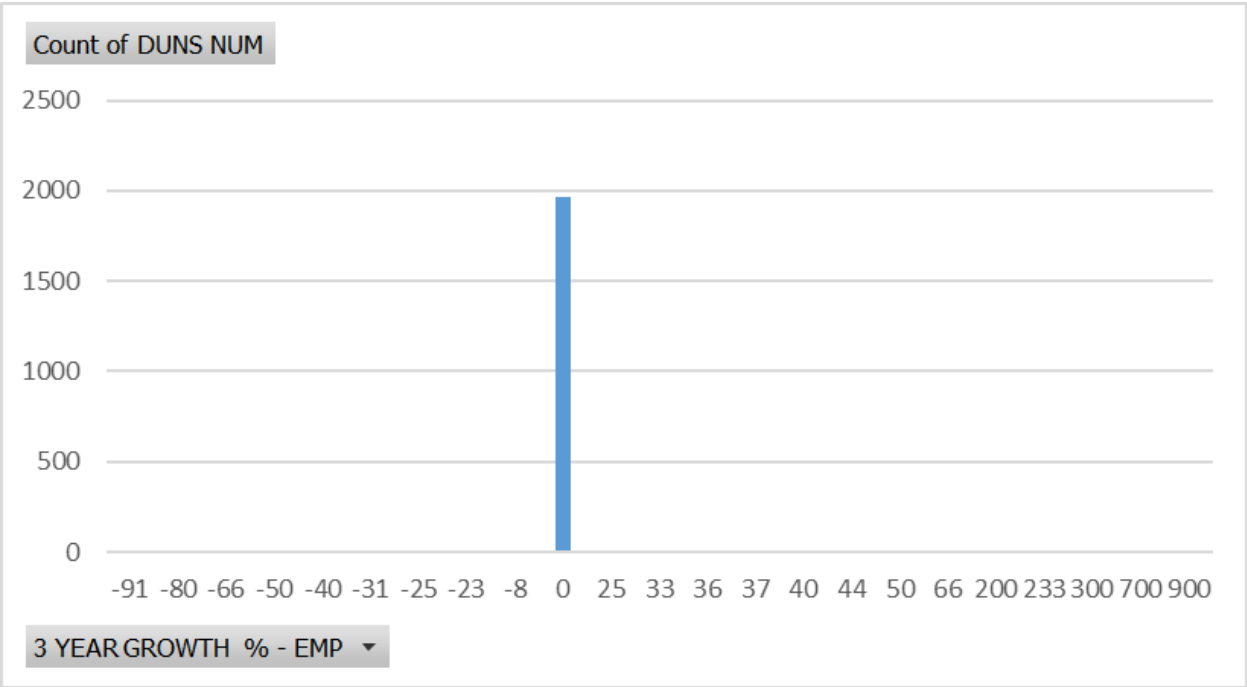


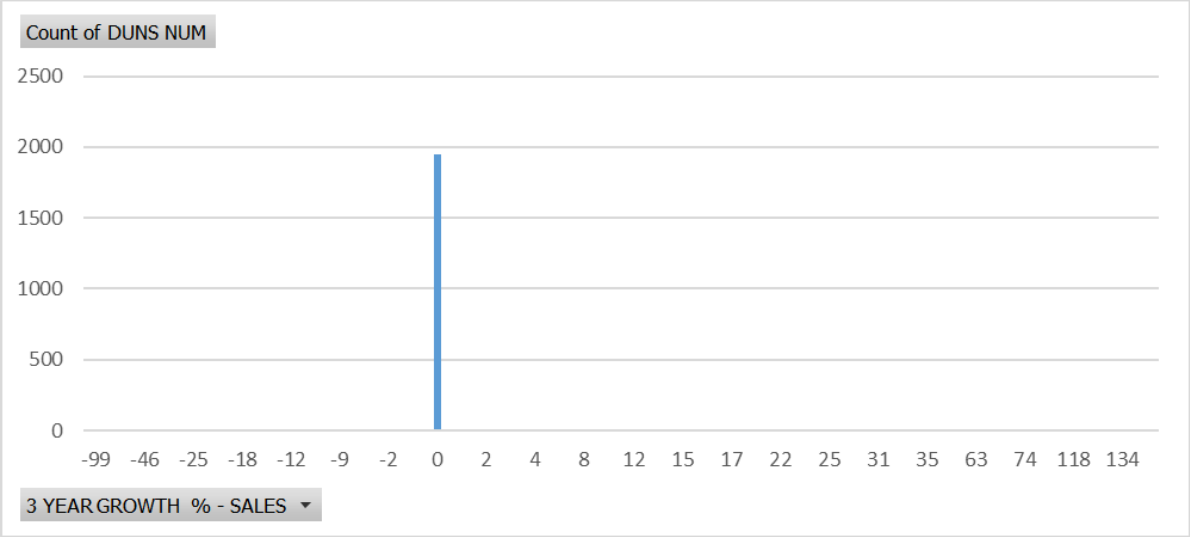


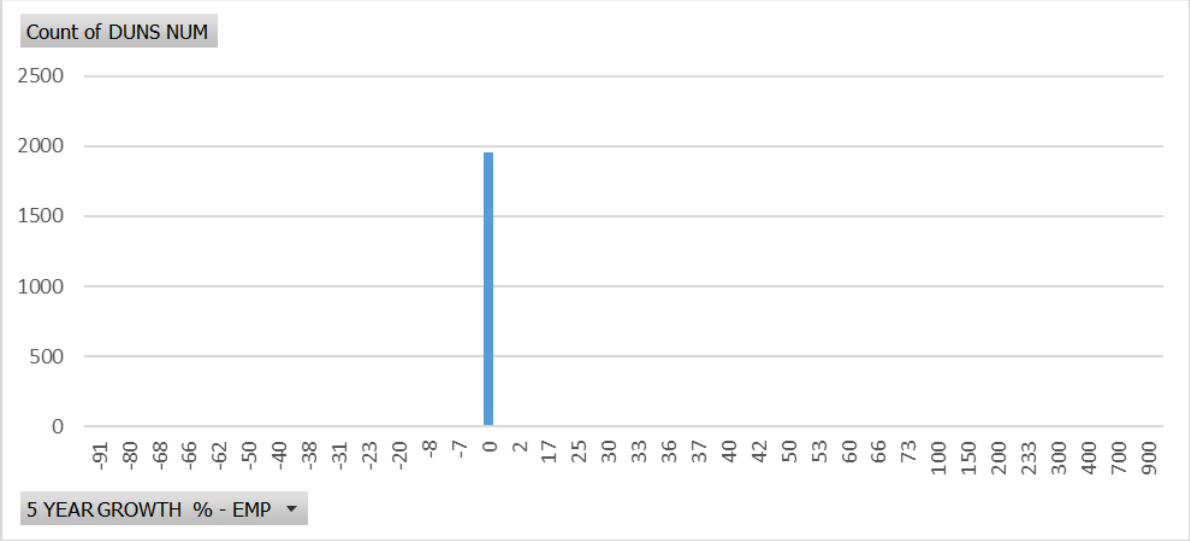

Count of DUNS NUM
1800
1600
1400
1200
1000
800
600
400
200
0
N
Y
SMALL BUSINESS IND

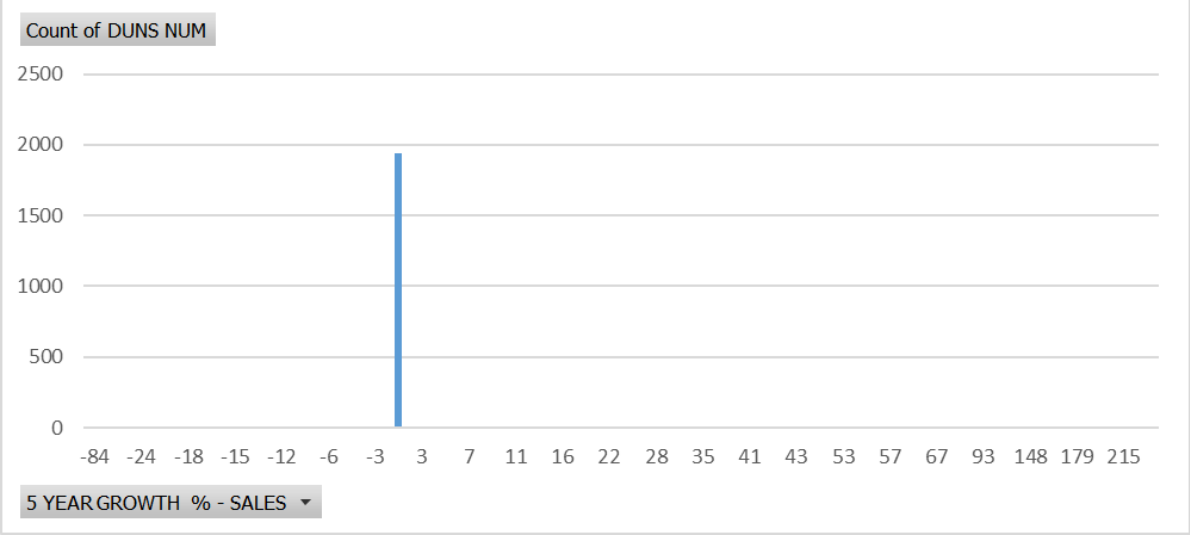

Count of DUNS NUM
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
1
MFG IND


Count of DUNS NUM
250
200
150
100
50
0
$0.00
$28,615.00
$34,669.00
$38,716.00
$43,681.00
$48,506.00
$52,671.00
$58,221.00
$62,695.00
$66,576.00
$71,243.00
$76,571.00
$80,566.00
$85,900.00
$90,012.00
$95,925.00
$102,432.00
$108,715.00
$118,435.00
$127,195.00
$137,363.00
$148,990.00
$163,860.00
$180,000.00
$205,651.00
$237,810.00
$297,848.00
$377,207.00
$492,623.00
$700,000.00
$976,566.00
$1,700,000.00
$7,423,501.00
SALES


Count of DUNS NUM
2500
2000
1500
1000
500
0
-91 -80 -66 -50 -40 -31 -25 -23 -8 0 25 33 36 37 40 44 50 66 200 233 300 700 900
3 YEAR GROWTH % - EMP


Count of DUNS NUM
2500
2000
1500
1000
500
0
-99 -46 -25 -18 -12 -9 -2 0 2 4 8 12 15 17 22 25 31 35 63 74 118 134
3 YEAR GROWTH % - SALES


Count of DUNS NUM
2500
2000
1500
1000
500
0
-91 -80 -68 -66 -62 -50 -40 -38 -31 -23 -20 -8 -7 0 2 17 25 30 33 36 37 40 42 50 53 60 66 73 100 150 200 233 300 400 700 900
5 YEAR GROWTH % - EMP


Count of DUNS NUM
2500
2000
1500
1000
500
0
-84 -24 -18 -15 -12 -6 -3 3 7 11 16 22 28 35 41 43 53 57 67 93 148 179 215
5 YEAR GROWTH % - SALES