**Personal Project: Chase Bank Data Analysis**

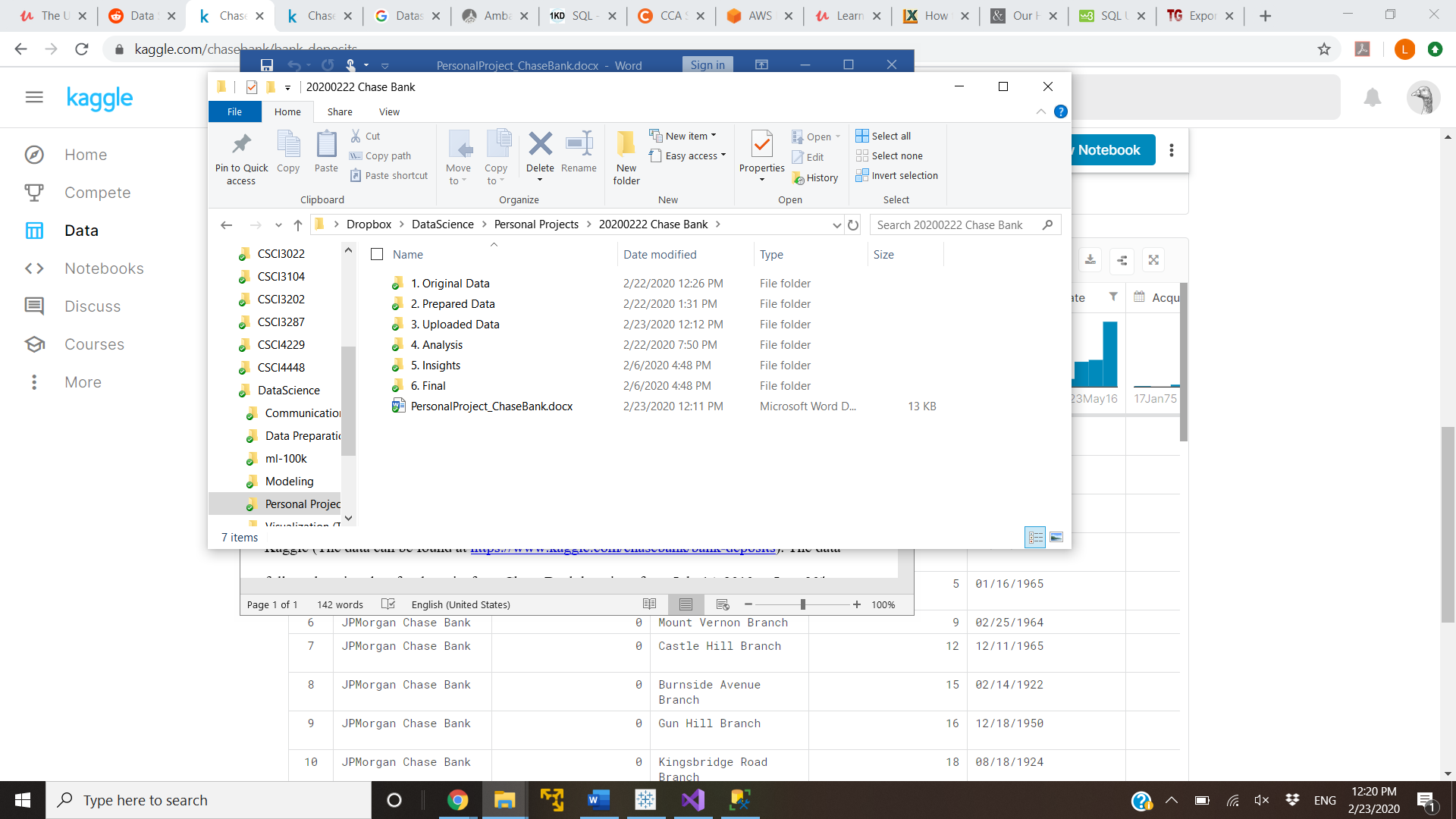
Luna McBride

**Introduction**

This project is an analysis practice project using a dataset from Chase Bank found on Kaggle (The data can be found at <https://www.kaggle.com/chasebank/bank-deposits>). The data follows location data for deposits from Chase Bank locations from July 1st, 2010 to June 30th, 2016 with the main branch keeping track of all data from all branches. Other data included in the file is the address of the branch, where the branch is (city, state, zip code, latitude and longitude of the area), and other key metrics. It has 5413 rows of data.

The purpose of this project is to practice key practices for analyzing data, such as ETL to ensure data quality and gaining insights through visualization via Tableau. I intend to analyze the company based on the data provided for this project.

**File Structure**



The project folder is structured like the image above to keep organized in the steps taken. This file, the write-up, is held outside the folders as it describes the file structure as well as the project in general. It also holds the final insights, which might leave the final folder unused. The other folders are structured as followed:

1. Original Data: This folder holds database.csv, which is the original data, unopened. This is kept for the purpose of copying and using elsewhere, as even the act of opening a file somewhere like Excel could alter the data in the files.
2. Prepared Data: This folder holds the data as it is prepared to be brought into SQL Server. There are two files within this one:
   1. dataset.txt: The original dataset, just changed to a txt file in order to further control how it goes into Excel.
   2. dataset.csv: This is the txt file pushed through Excel to have all the fields as text rather than being automatically converted by Excel. Normally dates should be fixed in Excel in this step as well, however, Excel had problems working with the established dates before 1900.
3. Uploaded Data: this contains a folder for the date the data was brought to SQL. That folder contained 2 files:
   1. dataset.csv: This is the same file as the dataset.csv in the processed data folder. It is simply copied here for organization purposes, as this was the version that was pushed through visual studio SSIS into SQL Server.
   2. Chase\_DataUpdate\_Procedure.sql: this is the procedure file used to bring the raw database brought in via visual studio into a working table. This is set up in a ‘highlight what you want to do’ sort of way. There is are functions for dropping the table, building the table, truncating the table, and filling the table from the RAW table. After those there are various queries to explore the data further, mostly for my sake before bringing the data into Tableau.
4. Analysis: This would normally hold data that was set aside for further analysis after some sort of corruption (perhaps a field being split into two and adding another column or an incorrect delimiter being used and making the number of fields less than the number of columns). Luckily, there were none of those. There are two files in here despite this:
   1. 20200222 ChaseProcessedData.csv: this file holds the data exported from the working table from the SQL Server database. I am using Tableau Public, which does not have the functionality to pull from an SQL Server database. It does allow csv, so I exported the data from SQL Server into a csv to be injected into Tableau Public.
   2. 20200223 Date-Deposit.csv: this file was made to make Tableau create the overall increase in deposits graph. This is a simple graph putting overall deposits to the respective year date given. Tableau did not want to link the deposits to a year and thus create a line graph, hence why I took the data and put it into this Excel file.
5. Insights: This folder holds all the Tableau graphs. I will analyze all of them further on, so they will not be listed out and explained until later on. However, since I am using Tableau Public, I have to save the graphs to the Tableau site and then download them to put them into word, so this is where they are being placed.

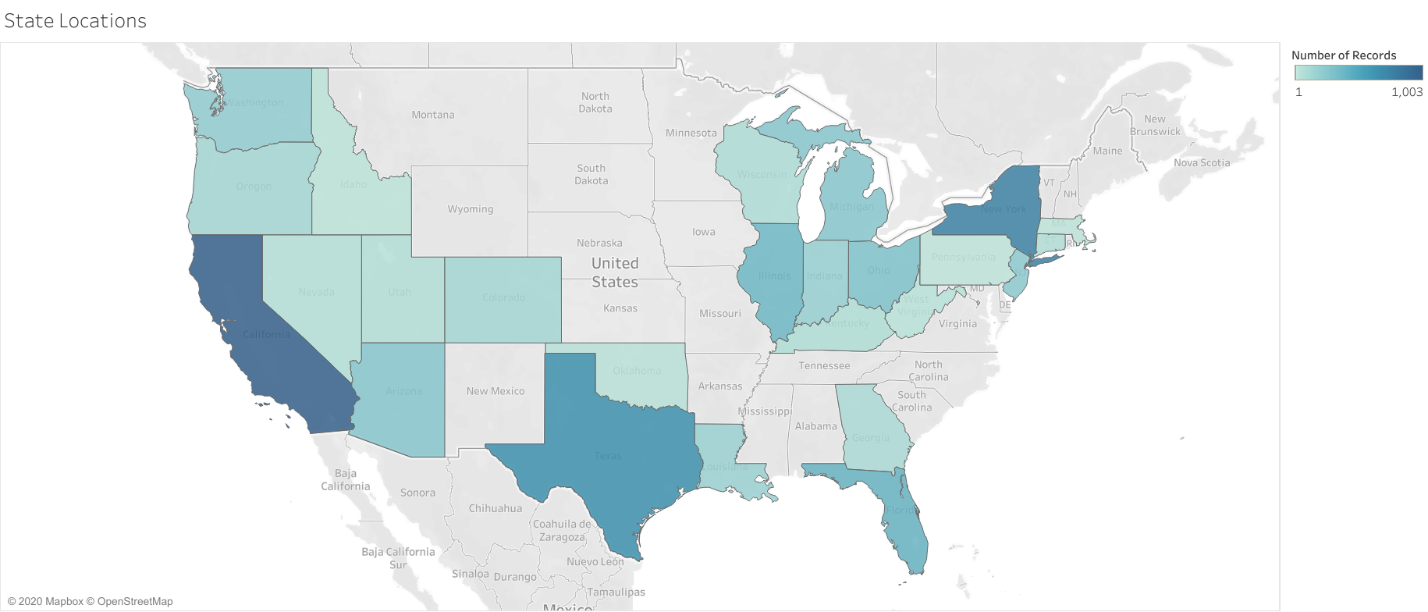
**Notes on ETL**

This area is to emphasize parts of the ETL process that caused problems and issues with the data. It should be emphasized that there were no major issues with the data in terms of corruption or other items that would have been removed by a conditional split in visual studio. There were, however, issues that would not be considered issues unless someone was looking specifically to analyze the data. Note that not all of these are important and needed change, but it is important to know what the data is doing. The issues were:

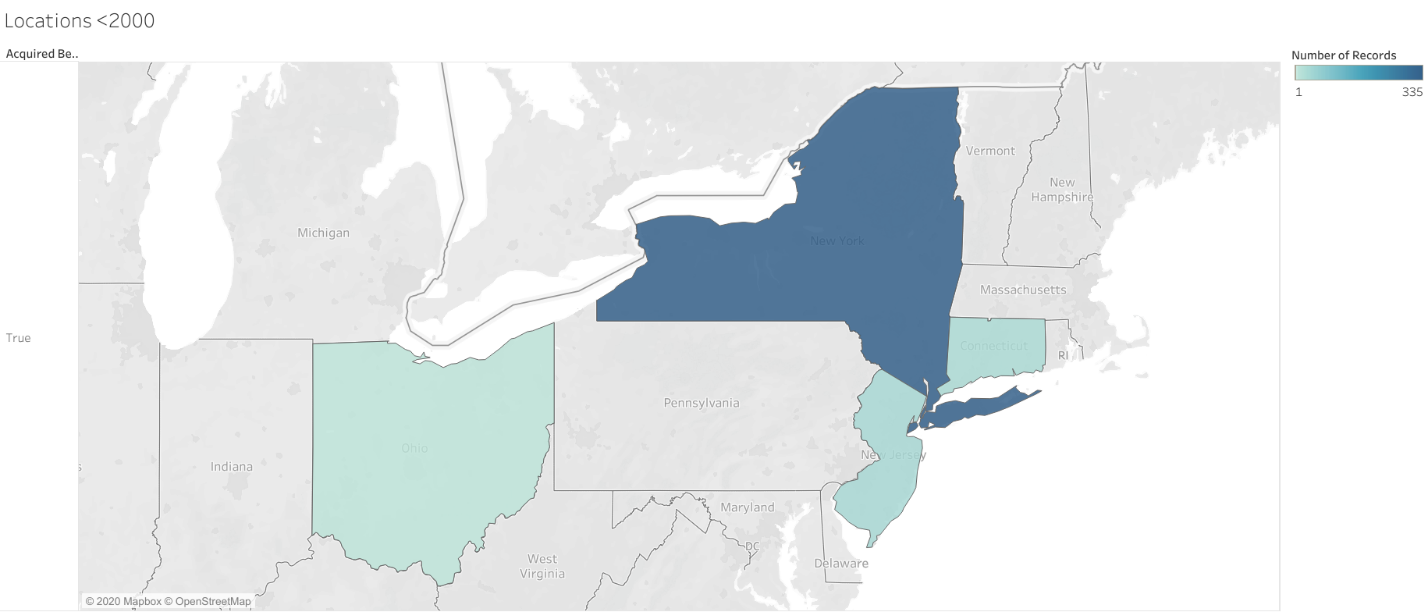
1. When trying to fix the data in Excel to the format yyyy-mm-dd, Excel could not track dates before 1900-01-01. There are dates before 1900, so Excel ignored the dates before 1900 and kept them in the previous mm/dd/yyyy format. After looking through the data, there did not look to be any other issue in that field besides the 1800’s dates, so I undid the date formatting and allowed SQL later down the line to fix it through implicit conversion of the field from text. This worked just fine.
2. In the database, many locations had an Acquired date empty because they did not necessarily need to acquire those locations from someone else. These values were turned into the default date when the Acquired Date field was converted to dates (1900-01-01). This would be dishonest when considering the importance of these dates, so I had it instead insert the Established Date into the Acquired Date. That would be more honest to when the company obtained ownership, as they acquired the location when they bought it.
3. There were several dates in the Established Date that had the month and day fields as the default 01-01. I cannot go up to Chase and ask about the accuracy of dates. The issue does mostly occur with dates in the 1800’s, so it is more likely that the records were not properly held from that long ago for the exact day the location was established. In the real world, I likely would have better access to the company and could ask, but in this case, it does not necessarily matter to analysis.
4. Various locations had the default 0,0 for latitude and longitude. This is really not a bad issue as I am not specifically looking into exact locations, but that would cause issue if that information was absolutely needed.
5. There some locations with zip codes with only 4 digits. I checked each of these on google and the state and county information fit perfectly with zip codes with just a 0 on the front, so these are really just zip codes with the front 0 dropped. These could have been big errors, so I had to check.

**Tableau Graphs and Analysis**

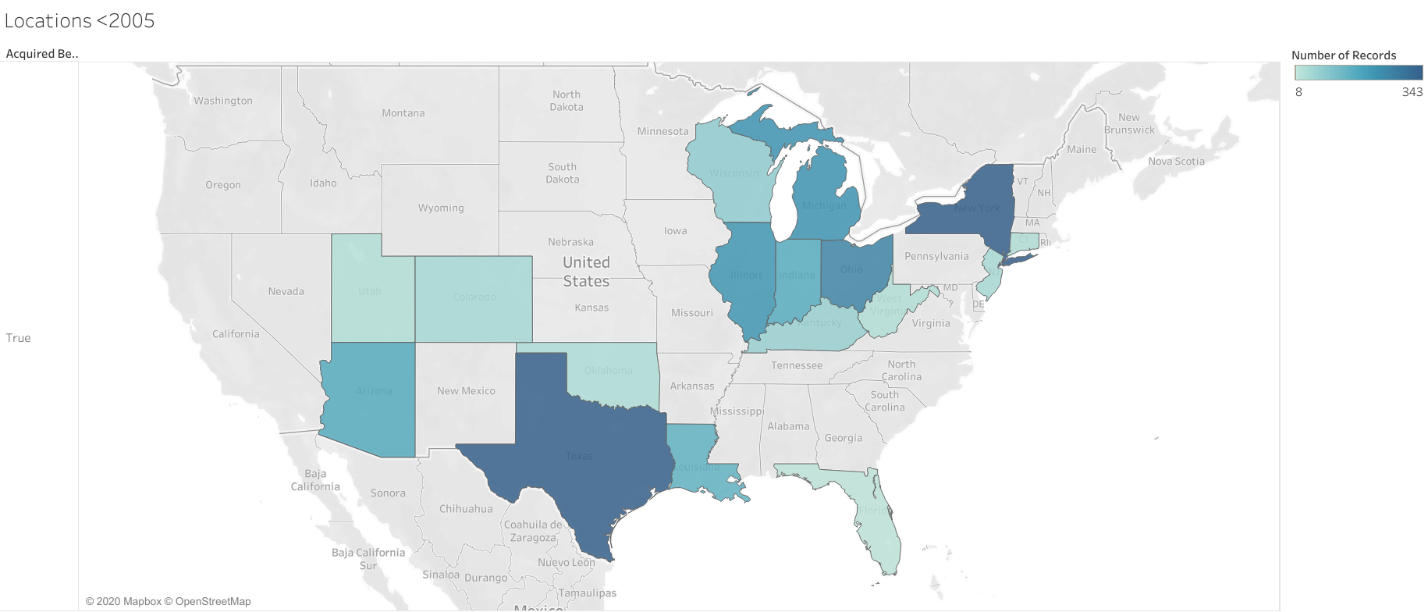
The given data did not have a direct instruction of what to look for, so I used the data to analyze the Chase company.



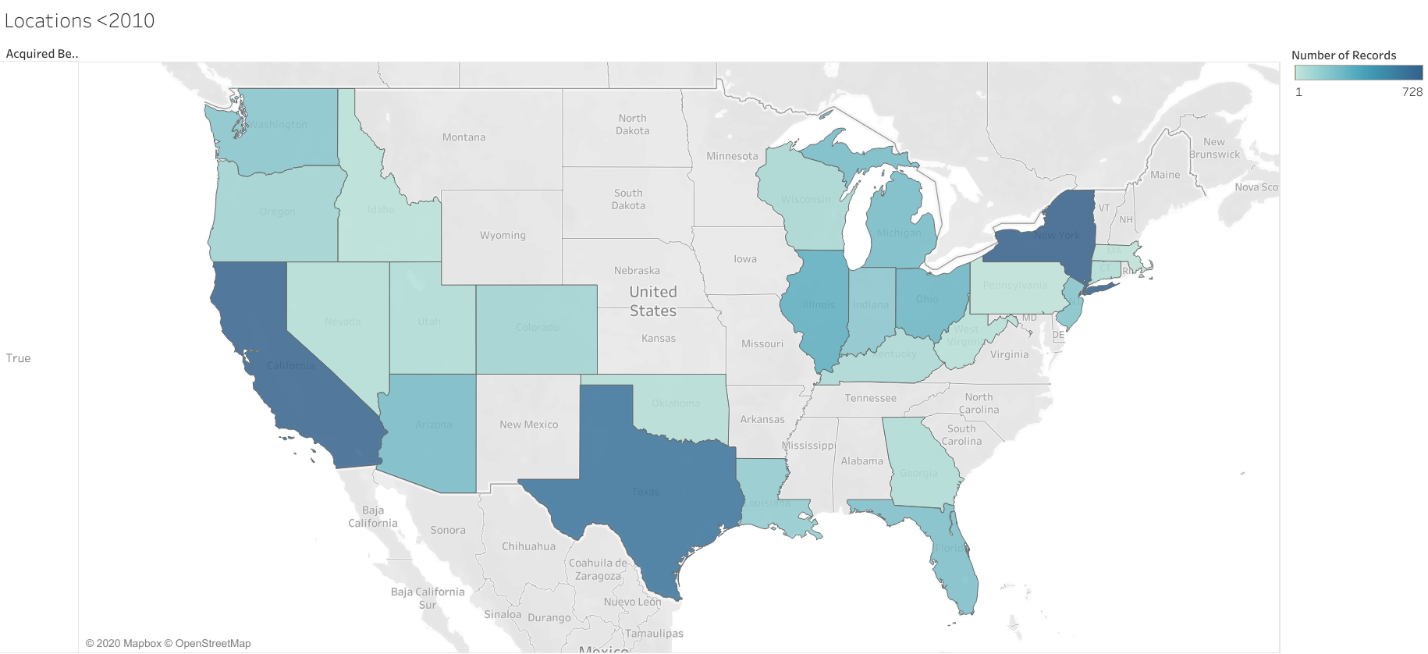
This is an image of the amount of locations per state. The company appears to have heavy focus on typical urban centers, mainly New York, California, and Texas. The company mainly started off in New York, as displayed in the next graph:



This is a map of the company ownership as of 2000. Maps before this were mostly listing New York locations, as up until 2000, this company was split as Chase and JP Morgan. A heavy focus was put on New York, as that is where the company was founded.

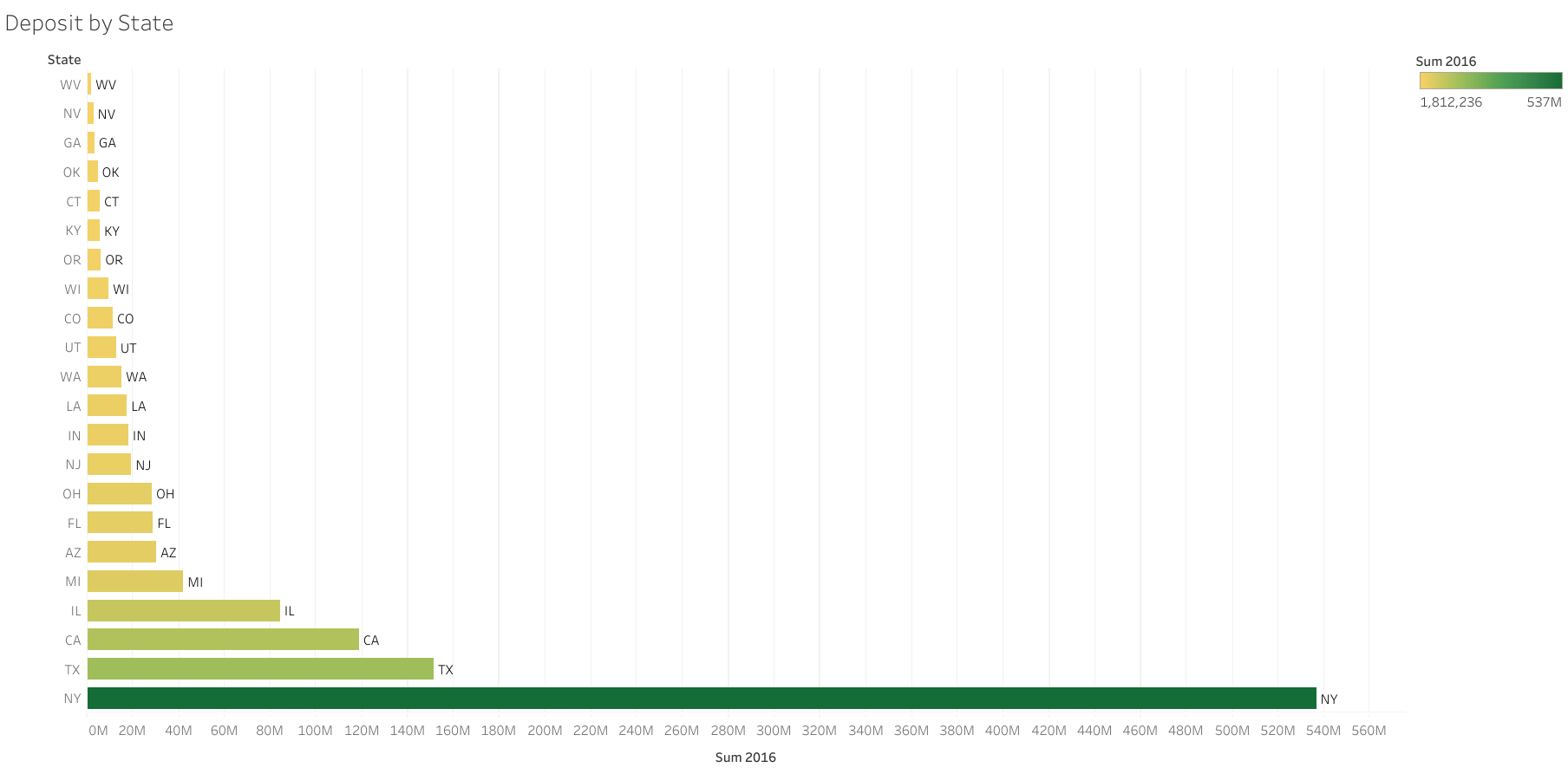


This map is the acquired locations as of 2005. This is likely due to merging with yet another company, just one with higher influence than JP Morgan/Chase had. According to the History section on the Chase site, this merger was with Bank One, thus leaving influence in 17 states.

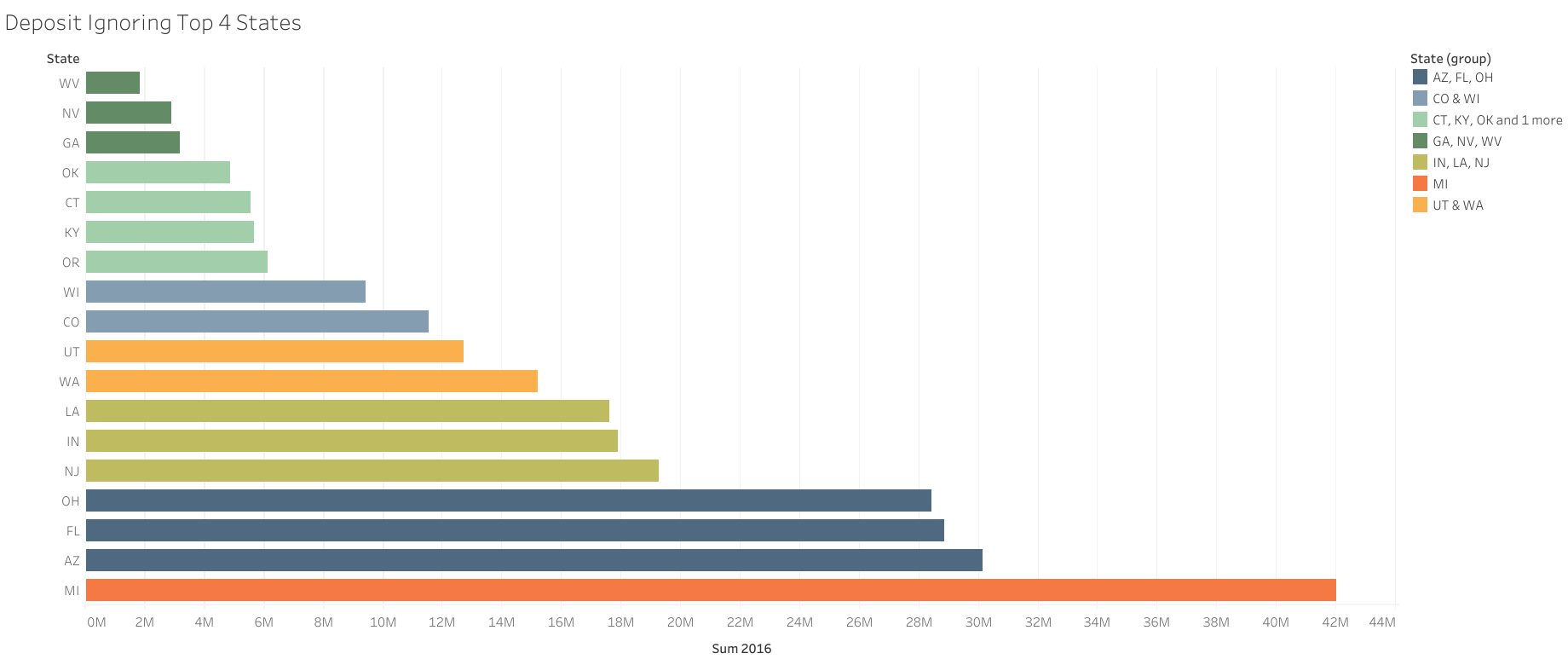


This is the map of locations as of 2010. The company has now stretched from coast to coast. According to the history site, this is due to a merger with Bear Stearns and Washington Mutual. With this, this means in the span of 10 years, the company has gone from 4 to 25 states. That is more than was done in the previous 200 years combined, as the company mostly kept focus in New York before then. Though, many of the newest additions were caused by acquisitions rather than direct expansion, which provides insight into the strategy of Chase/JP Morgan Co.

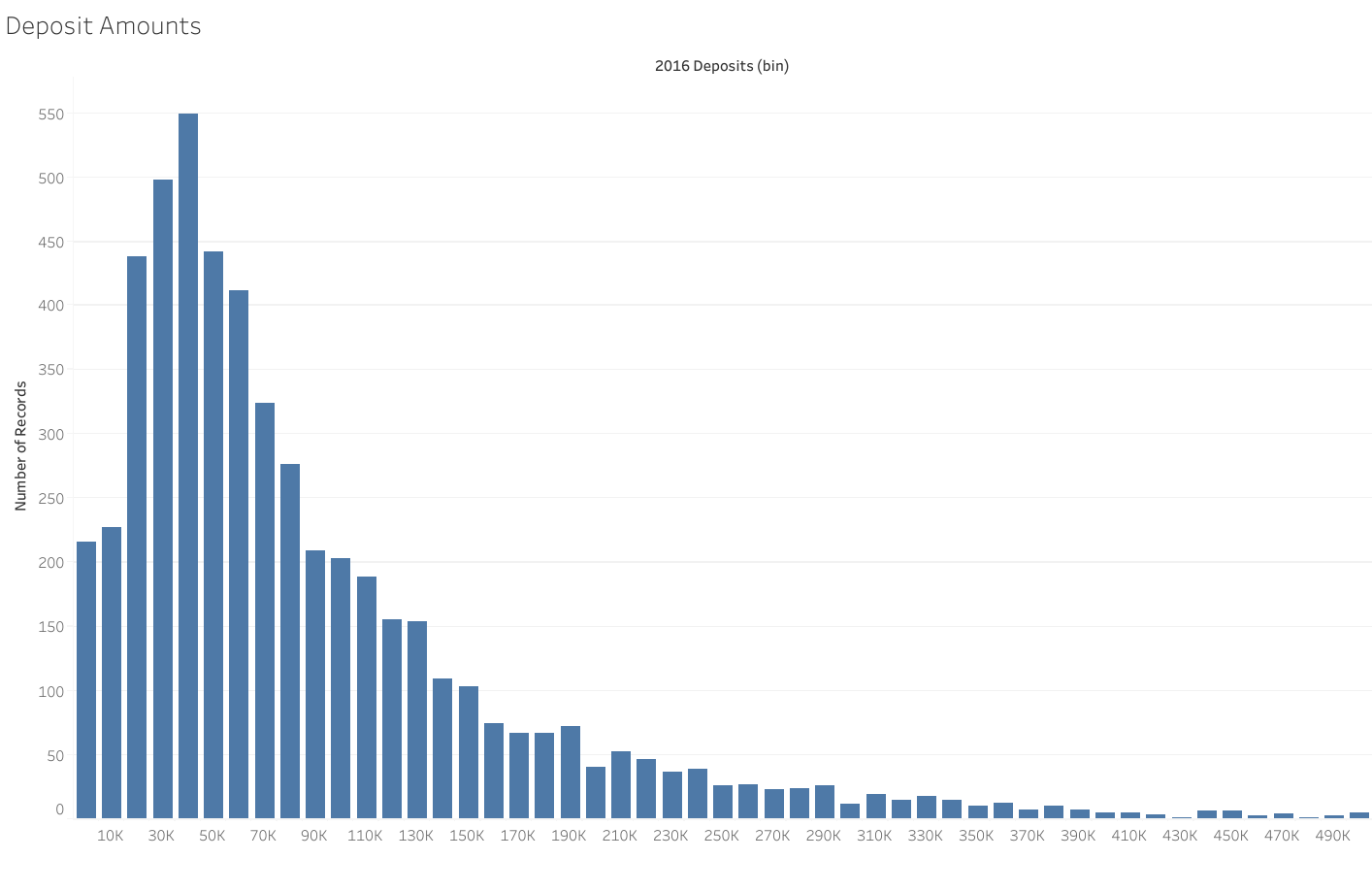
Now, to look at deposits in recent years:



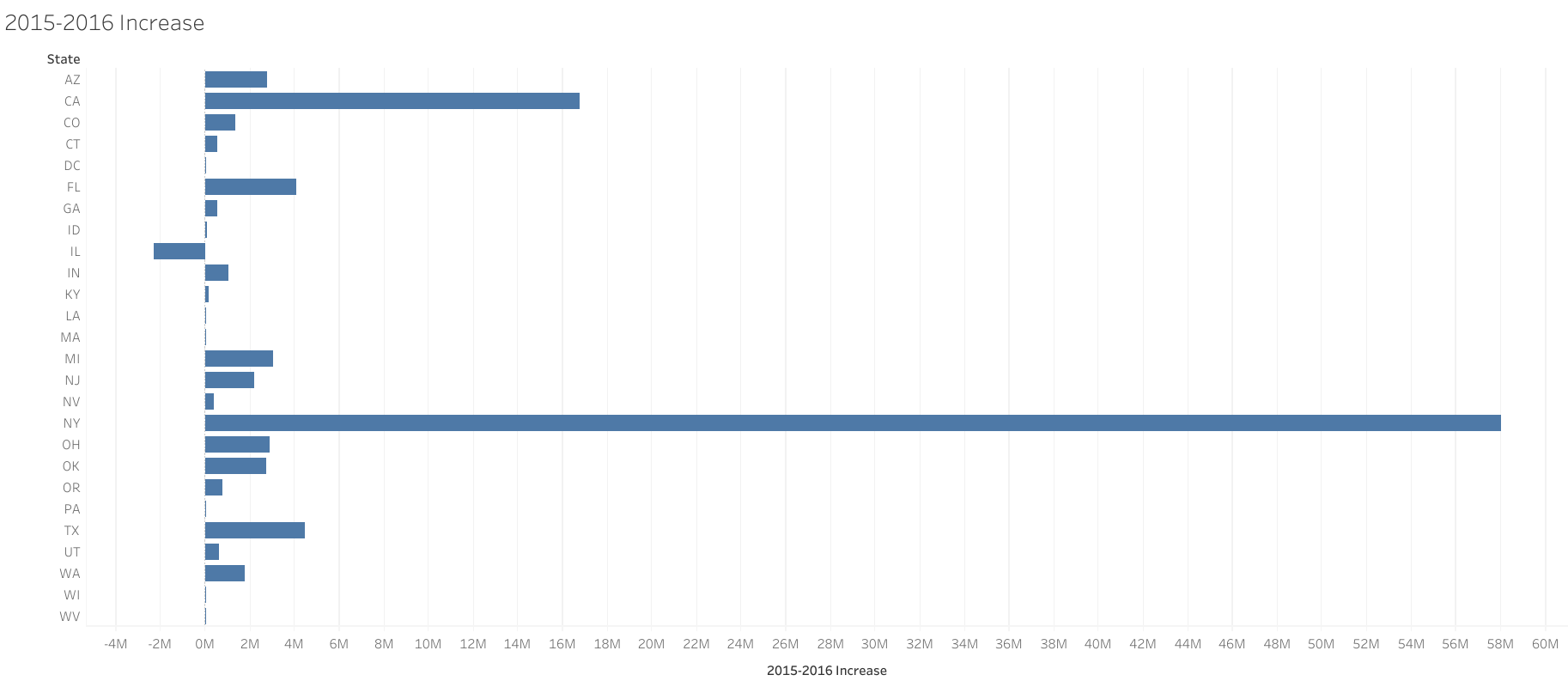
This is the 2016 deposits by state. Even now, New York is clearly an important factor to traffic, as deposits in New York are more than double any other state. As seen in the evolution of locations, New York has always been a key location to business. Otherwise, the other two in the top three are the usual suspects, being the highly-populated Texas and California. In fact, compared to these huge juggernaut states, the smaller states look like a blip on the graph. It is hard to tell the importance of these states, hence why the next graph is here:



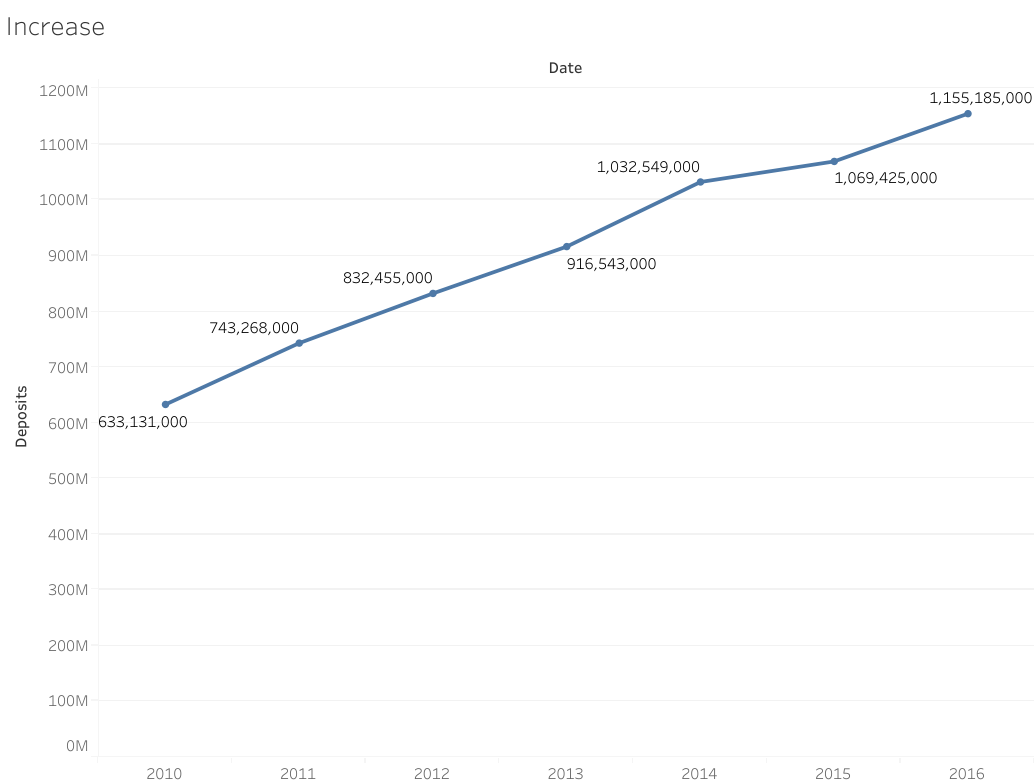
This is the same graph, just ignoring the four largest states. Deposits represent traffic, so it is important to look even into the smaller deposit amounts. I have colored the states by size, containing them in segments of 4 Million of deposits. Many of the lower numbers are associated with smaller states. However, it would still be good to look into them as a company, as it is also possible there are other factors at play, like lack of brand knowledge or a stronger competitor. The data given does not have fields that could potentially be used to check exit rates, but looking into this data, it might be smart to look into exit rates via a subsample to better know the retention in the audience. The data does, however, allows analysis of deposits per individual stores above that of a state level.



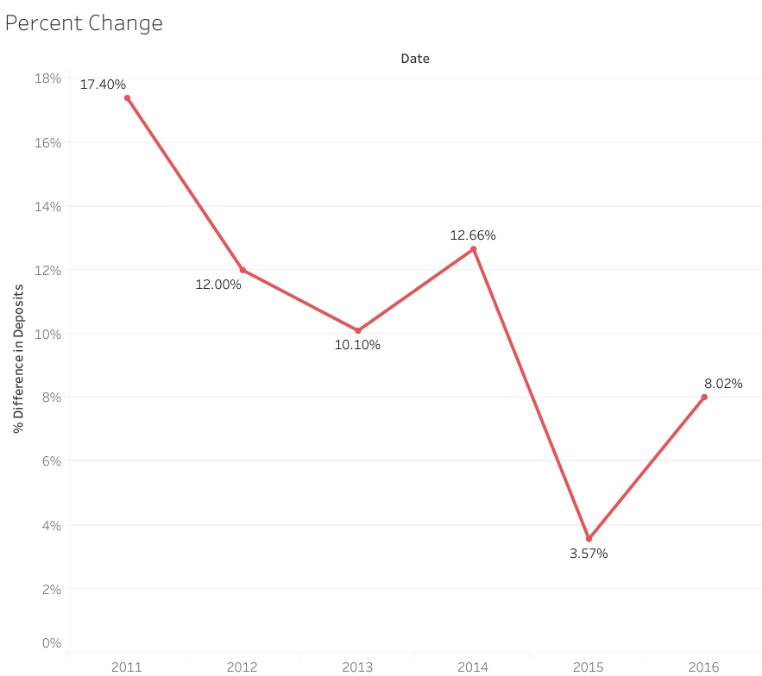
This represents the deposit amounts per store. The tail on the right side extended further than this, however, the tail span does not represent the average, and thus it was cut down. The average store appears to take around 40k in deposits in a given year with the overall majority taking up to 150k per year. These are likely the average stores catered to the community, as there are many individual stores with very large deposit amounts less likely held by the general population, as shown by the tail trailing to the right.



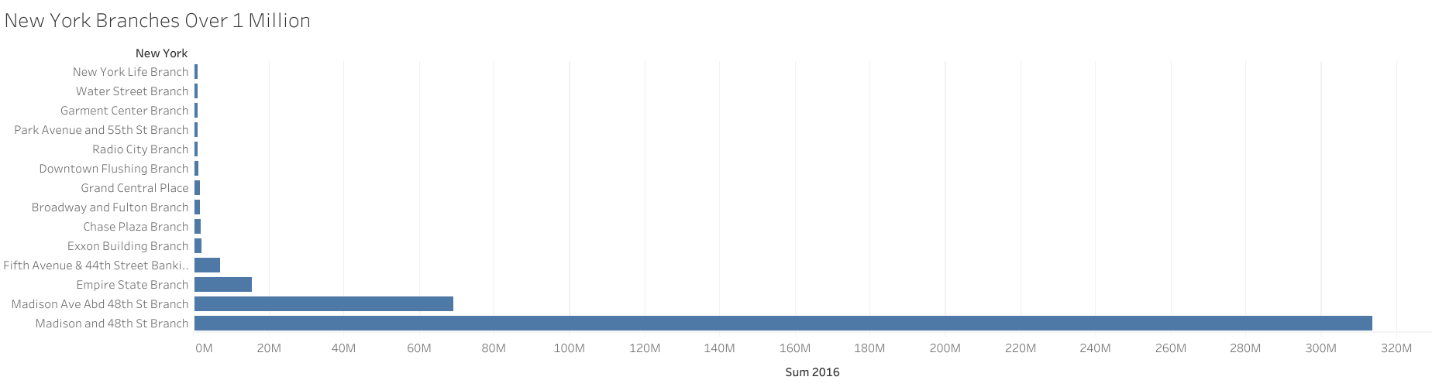
Now, for increase from the previous year, there were heavy increases from the usual suspects. Interestingly, there are several smaller states with pretty large increases for their sizes, such as Washington and New Jersey. They are not nearly as large as New York, but they are impressive in comparison to their relative size. In fact, all fields except one has seen an increase in comparison to the previous year. The one that did not, however, is Illinois. This is the fourth biggest overall, which means there might potentially be something going on with the company in that state. This data does not have enough columns to be able to look further into the issue with this state and I do not have direct access to the company, but the lowering in deposits in such a big state in terms of deposits could be something to look into.



Now in terms of overall increases in deposits, the overall company has an overall increase by nearly half a billion in the span of 6 years. This may potentially be a result of the expansions the years before and the increase in locations in the previous 10 years. The rate of increase, however, as been a bit variable:



The rate of increases has had some fluctuation, but the overall trend is downward. This would also be something to look into if I had access to more Chase data systems, however, for now I can just point out potential problems.



There is one more interesting factor to look at, and that is New York. It has been huge on every deposit amount on the other charts. This chart holds every store that has deposit numbers over 1 Million. Interestingly, there are two stores that that take in way more than the rest, being Madison and 48th and another one called Madison Ave Abd 48th. This could be the same address with an error in spelling, but they have different id numbers and data. This could potentially be an error within the data for a name, but it may also be the actual name of a very nearby location. Either way, these two bring in much more than any other individual location. It is highly likely that these are the locations the company routes large New York companies through just by the massive amount of money in deposits these locations specifically bring in. This information means that these two locations specifically create an anomaly in the data skewing toward New York.

When calculating without those two included, the New York number is 154,021,162 as calculated quickly by a calculator, which is still the highest, beating Texas by about 30 million. That still makes logical sense to the company’s history and business, just not counting the absurd load brought by the largest individual locations.

**Conclusion**

Looking into the data of a company, no matter how little given, can say large amounts about the company. This data did not give much in terms of items that could be tested through AB testing or typical statistical models, most likely so other companies could not draw too much insight to use in their own companies. Despite this, with what was given, I could walk through the company’s history just by the maps Tableau could create. I could also see the approximate size of the average locations, the importance of each state to the company, and see the most important location to the overall operation (that being the Madison Avenue and 48th Street location). I could extrapolate further from the data, but I do not want to dig too deep to the point where I try to make claims unfounded by the current data.

Yet, without the need for more data, there are various points that could be worked on within the company; namely, the decrease in deposits in Illinois. With more access to Chase data, it is likely I could find a reason and extrapolate to help bring forth a better route for the company in the area. Even without the additional data, some of the cracks still become apparent. This exemplifies the power of the tools used and my ability to wrangle them.