Jacob Fullmer

Final Data Viz Project

10-24-2017

Executive summary

Given the recent growth of the the business community in California, Chase Bank deposits in California have grown faster than any other of their key markets. That could be huge news and something the regional manager over California could rightly be proud of. Afterall, those businesses could go anywhere but a number of them and their emplmoyees have chosen Chase. However, the overall financial impact to the national bank is much smaller when compared to their deposits in other markets, specificaly New York. In fact, Texas performs better than California and Illinois is right on its tail.

The final graph illustrates the growth rate experienced in each of these four key markets but provides necessary context against growth rate by showing overall impact on the bank by comparing total deposits for these states over the last six years. California's regional vice-president should be proud of his stewardship, but not get too carried away. To provide additional perspective on thes key market's business impact, I've also included the comparable deposits for states in Chase's secondary or non-key markets, representing 22% of all deposits.

Data background

I chose the Chase Bank data accessed on Kaggle. The source states that it is "a record for every branch of Chase Bank in the United States, including the branch's name and number, date established as a bank office and (if applicable) acquired by JP Morgan Chase, physical location as street address, city, state, zip, and latitude and longitude coordinates, and the amount deposited at the branch (or the institution, for the bank's main office)" from 2010 to June 30, 2016.

I did not need the detailed address information, latitude and longitude coordinates, or acquisition information - though I did contemplate using the lattter. I was furhter encouraged when the collector of this data said it came from the Chase website, which compiled the data from the FDIC (Federal Deposit Insurance Corporation) annual Summary of Deposits reports.

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(stringr)
library(sf)
```

```
## Warning: package 'sf' was built under R version 3.4.2
```

```
## Linking to GEOS 3.6.1, GDAL 2.2.0, proj.4 4.9.3
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 3.4.2
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.4.2
```

```
library(forcats)
```

Data cleaning

To clean the data, I knew I needed to fix two problems with the year. First, it would be not use to me as a variable if not an integer or number. So, I needed to strip it of the "xYYYY.Deposits" through renaming. Then, I chose to gather all the deposit information and match it to the branch, so we have a column for "year" instead of individual years with their own column. That, btw, was some fun coding and learning from watching the refugee_raw to clean conversion. Oy vey…Good learning in the website explanation though. Thanks.

Last, I decided to take out the Main Office data from this data set. While some of the large New York deposits seemed justifiably large because of Wall Street, etc. I didn't think that Columubs, Ohio having $6.3 in deposits seemed uncharacteristic of regular branch activities and was more likely a factor of it being the main office funneling funds off of other branches for whatever reason. One branch in Ohio having 3.5 times the next closest branches deposit totals just seemed fishy.

```
raw_banks <- read.csv("data/Banking.csv")

clean_bank <- raw_banks %>%
  rename("2010" = "X2010.Deposits",
         "2011" = "X2011.Deposits",
         "2012" = "X2012.Deposits",
         "2013" = "X2013.Deposits",
         "2014" = "X2014.Deposits",
         "2015" = "X2015.Deposits",
         "2016" = "X2016.Deposits") %>%
  filter(Main.Office != 1) %>%
  mutate(Branch.Number  = as.numeric(Branch.Number),
         Main.Office = as.factor(Main.Office),
         Zipcode = as.factor(Zipcode)) %>%
  gather(year, deposits, -Institution.Name, -Main.Office, -Branch.Name, -Branch.Number, -Established.Da
  mutate(year = as.numeric(year),
         year_date = ymd(paste0(year, "-01-01")),
         deposits = as.numeric(deposits))

head(clean_bank)
```

```
##      Institution.Name Main.Office                          Branch.Name
## 1 JPMorgan Chase Bank           0        Vernon Hills Scarsdale Branch
## 2 JPMorgan Chase Bank           0 Great Neck Northern Boulevard Branch
## 3 JPMorgan Chase Bank           0               North Hartsdale Branch
## 4 JPMorgan Chase Bank           0             Lawrence Rockaway Branch
## 5 JPMorgan Chase Bank           0                  Mount Vernon Branch
## 6 JPMorgan Chase Bank           0                   Castle Hill Branch
##   Branch.Number Established.Date Acquired.Date           Street.Address
## 1             2       03/20/1961                   676 White Plains Road
## 2             3       09/09/1963                  410 Northern Boulevard
## 3             4       02/19/1966                353 North Central Avenue
## 4             5       01/16/1965                   335 Rockaway Turnpike
## 5             9       02/25/1964                    22 West First Street
## 6            12       12/11/1965                  784 Castle Hill Avenue
##           City      County State Zipcode Latitude Longitude year deposits
## 1    Scarsdale Westchester    NY   10583 40.97008 -73.80670 2010   293229
## 2   Great Neck      Nassau    NY   11021 40.77944 -73.72240 2010   191011
## 3    Hartsdale Westchester    NY   10530 41.02654 -73.79168 2010    87110
## 4     Lawrence      Nassau    NY   11559 40.62715 -73.73675 2010   172608
## 5 Mount Vernon Westchester    NY   10550 40.91144 -73.83804 2010   146820
## 6        Bronx       Bronx    NY   10473 40.82292 -73.84887 2010    75131
##    year_date
## 1 2010-01-01
## 2 2010-01-01
## 3 2010-01-01
## 4 2010-01-01
## 5 2010-01-01
## 6 2010-01-01
```

Individual figures

Figure 1: States Leading in Total Deposits

I chose the geographic mapping of this data because bars and lines just get boring sometimes. When properly executed, the map can be one of the most beautful and context heavy elements available to us. We have drilled into our brains what the states are and where they are; so why not borrow some of that knowledge.

I first needed to make a simpler data base with just the information I needed (states and deposits) in order for things to be processed. I kept trying it with all the data from Clean_bank but it was to much and took forever. I joined that data with the shape file from Census.gov to get the shapes, and chose an appropriate map curve to be lovely to the eye. Originally, I had the legend on the bottom but learned that I needed to move it to fit all of my other charts onto the same page. So, it went to the top. I chose gradient scale and capped the maximum limit so as to have a grouping of their top performing states. If I hadn't done that, then we would only see New York.

Truthful: The data is clear and we convey the Main Office data that I took out.

Functional: It's easy to organize and compare the data visually by state. It was important to set a limit to the maximum deposit amount or else we wouldn't have any contrast at all.

Beautiful: Seriously, It's lovely. I just dig it. The colros are easy to organize and group by. It's a familiar shape. It's not a bar or line graph. Huge win and drives interest.

Insightful: We can clearly see that chase is a HUGE bank, but if they didn't have their business from their four largest states, it would be a loss of 88% of their deposits! HUGE!

Enlightening: If a manager had this information, they would know they should make sure the top states had the resources and investment they needed to keep growing and maintain their business.

Contrast: The colors do a great job of showing what's more and less important to the business. Showing the distance of all the rest of America in between their markets is big.

Repitition: Shades and each state's shape. Not a lot of repitition other than known shapes.

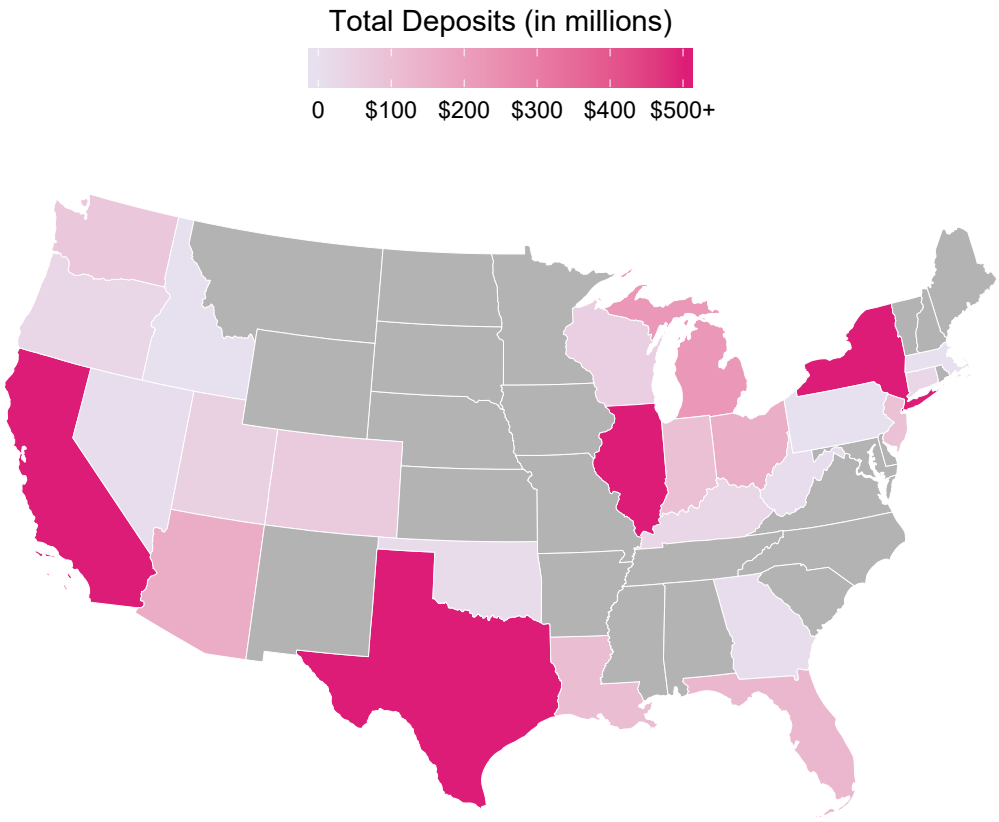
Allignment: The shape file does a perfect job of alligning things. The Albers Equal Area Conic projection is commonly accepted, understood, and beuatiful to patriotic americans everywhere (even if we don't understand map spatial referencing!).

Proximity: Keeping the legend close is important.

```
state_shapes <- st_read("data/cb_2015_us_state_20m.shp",
                        stringsAsFactors = FALSE) %>%
  rename("State" = "STUSPS") %>%
  filter(!(State %in% c("AK","HI", "PR")))
```

```
## Reading layer `cb_2015_us_state_20m' from data source `C:\Users\jakef\Documents\DataViz\Final Projec
## Simple feature collection with 52 features and 9 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
```

```
#figure out what is in this to match the state.
#head(state_maps)

#oh, look. It's STUSPS for state.

state_deposits <- clean_bank %>%
  group_by(State) %>%
  summarize(deposits = sum(deposits, na.rm = TRUE)) %>%
  mutate(trunc.deposits = ifelse(deposits > 500000000, 500000000, deposits))

state_maps <- state_shapes %>%
  left_join(state_deposits, by = "State")

map <- ggplot(state_maps) +
  geom_sf(aes(fill = trunc.deposits), color = "White", size = .05) +
  coord_sf(crs = st_crs(102003)) +
  scale_fill_gradient(low = "#e7e1ef", high = "#dd1c77", na.value = "grey70", breaks=c(0,100000000, 200
  guides(fill = guide_colorbar(title.position = "top",
                               title.hjust = "0.5",
                               title = "Total Deposits (in millions)",
                               barwidth = 10, barheight = 1)) +
  theme_void() +
  theme(legend.position = "top")

map
```




```
ggsave(map, filename = "output/USmap-bankdeposits.pdf", width = 7.29, height = 4.5)
```

Figure 2: Silicone_Valley_Growthrate

I chose the slopegraph because it is the best way to show the most essential information I wantd to convey: deposit growth rate.

Truthful: It's so simple and clear that it can't not be truthful. If the reader fails to realize this is just a percentage rate, they may be tempted to believe each state started with similar deposit amounts. but it was important to show that on the on the other side of the big graph.

Functional: Quick and easy.

Beautiful: It's not the most beautilfy but using colors to contrast important information from the rest does lend some level of aesthetic to it.

Insightful: California is doing great and growing at a great rate.

Enlightening: As mentioend above, this helps upper management know where to put their resources to continue their good winning streak.

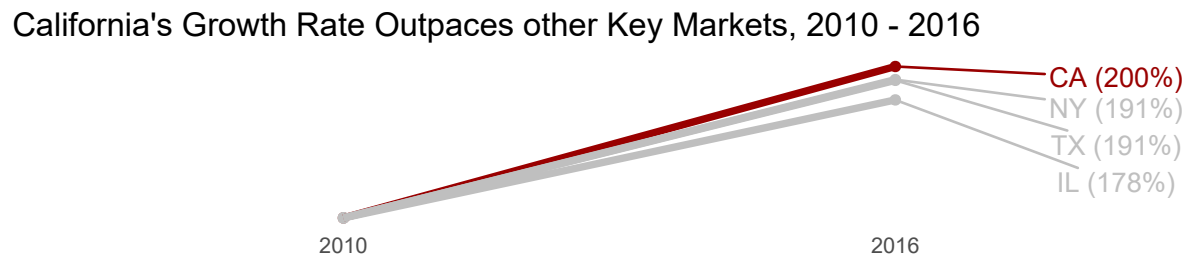
Contrast: Color of lines contrasting.

Repitition: The whole thing is a simple and repetitive style.

Allignment: Alligned well to the data start and end years.

Proximity: Each line is close in proximity to show the differnces. Labels must be close as well.

```
valley_effect <- clean_bank %>%
  group_by(State, year) %>%
  summarize(deposits = sum(deposits, na.rm = TRUE)) %>%
  filter(State %in% c("CA", "TX", "NY", "IL"), year %in% c("2010", "2016")) %>%
  spread(year, deposits) %>%
  mutate(start = `2010`/`2010`, end = `2016`/`2010`) %>%
  gather(period, growth_rate, c(start, end)) %>%
  mutate(period = factor(period, levels = c("start", "end"), ordered = TRUE)) %>%
  mutate(period = fct_recode(period, "2010" = "start", "2016" = "end")) %>%
  mutate(label_end = paste0(as.character(State)," (",round(growth_rate*100, 0),"%)")) %>%
  mutate(label_end = ifelse(period == "2010", NA, label_end)) %>%
  mutate(colorize = ifelse(State == "CA", TRUE, FALSE))

## I had to filter out for only the states and years of data I wnated. We needed to rename (and rename a

valley_slope <- ggplot(valley_effect, aes(x = period, y = growth_rate, group = State, color = colorize))
  geom_point() +
  geom_line(size = 1.4) +
  geom_text_repel(aes(label = label_end), direction = "y", nudge_x = .4, seed = 12) +
  scale_color_manual(values = c("grey75", "#990000")) +
  guides(color = FALSE) +
  labs(title = "California's Growth Rate Outpaces other Key Markets, 2010 - 2016", x = NULL, y =NULL) +
theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y = element_blank())

valley_slope
```




```
ggsave(valley_slope, filename = "output/valley_slope.pdf", width = 7, height = 1.5)
```

Figure 3: Key Market Deposit Summary

I chose bar graphs for the key and secondary market summary comparisons because it is the most data rich comparison chart here.

Truthful: It was difficult to add additional data labels in R but I waited to add them in Illustrator to highlight some of the summary differences.

Functional and Proximity/Alignment: It conveys so much with contrasting each state's performance right next to each other.

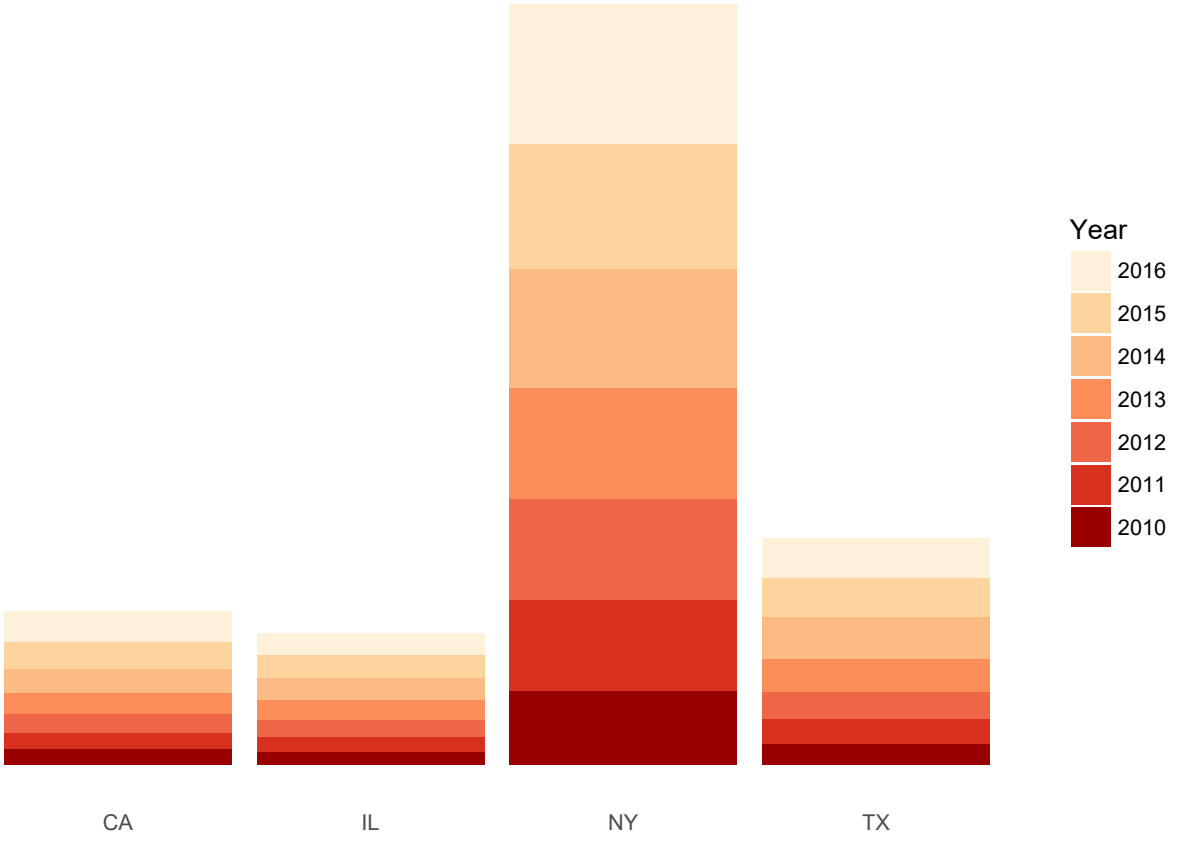
Beautiful: The colors are soft on the eye and yet provide enough contrast.

Insightful and Contast: Though not customary, I selected the year as the fill type for these graphs becasue it

was such a stark contrast between New York's best/most recent year compared against the past SIX years for the next closest three states. New York's financial centers drive so much business it cannot be underestated.

Enlightening: As with all of these charts, management can help set their priorities (or reset them if need be) by viewing some of these graphs simple sumary of information.

```
top_markets <- clean_bank %>%
  group_by(State, year) %>%
  summarize(deposits = sum(deposits, na.rm = TRUE)) %>%
  filter(State %in% c("CA", "TX", "NY", "IL"))

graph_top_markets <- ggplot(top_markets, aes(x = State, y = deposits, fill = fct_rev(as.factor(year))))
  geom_col() +
  labs(y = NULL, x = NULL) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y = element_blank()) +
  guides(fill=guide_legend(title="Year")) +
  scale_fill_manual(values = c("#fef0d9", "#fdd49e", "#fdbb84", "#fc8d59", "#ef6548", "#d7301f", "#9900

graph_top_markets
```

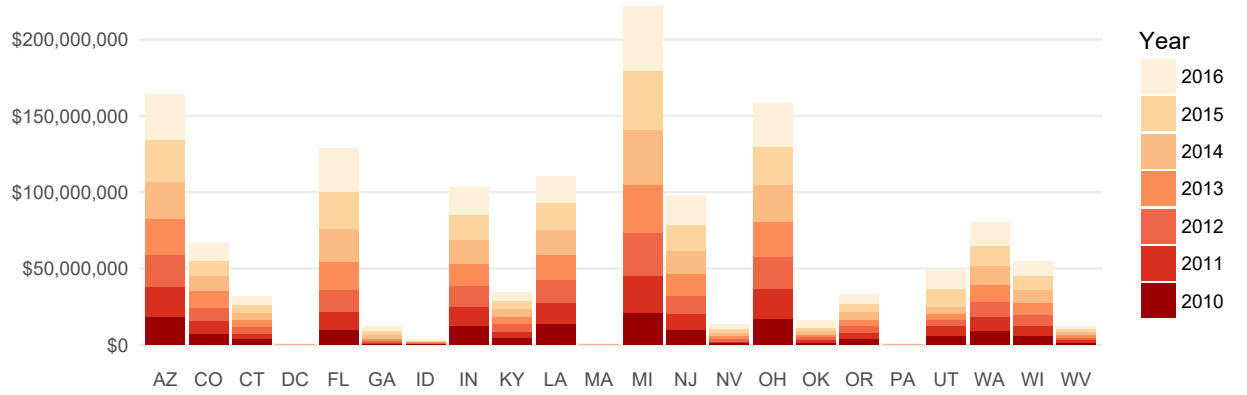



```
ggsave(graph_top_markets, filename = "output/top_markets.pdf", width = 7, height = 5)
```

Figure 4: sum of mid america markets

All similar reasons here for why I did another bar graph for the lower performing states. In addition to the above, however, I felt it important to show the tail-end spread of their business when compared to the key market performers. For this graph, because I didn't need to show just one or two elements (i.e. New York compared against California), I wanted to include some of the major gridlines in a clean and simple format. That way readers can get more than just state by state performance comparisons. They can get the actual numbers, too.

```
mid_markets <- clean_bank %>%
  group_by(State, year) %>%
  summarize(deposits = sum(deposits, na.rm = TRUE)) %>%
  filter(!State %in% c("CA", "TX", "NY", "IL"))

graph_mid_markets <- ggplot(mid_markets, aes(x = State, y = deposits, fill = fct_rev(as.factor(year))))
  geom_col() +
  labs(y = NULL, x = NULL) +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor = element_blank()) +
  guides(fill=guide_legend(title="Year")) +
  scale_y_continuous(labels = scales::dollar) +
  scale_fill_manual(values = c("#fef0d9", "#fdd49e", "#fdbb84", "#fc8d59", "#ef6548", "#d7301f", "#9900

graph_mid_markets
```




```
ggsave(graph_mid_markets, filename = "output/mid_markets.pdf", width = 8, height = 2.7)
```

Final Compilation Figure

Describe why you designed it the way you did? Why did you choose those colors, fonts, and other design elements? Does it convey truth? I wanted the map to the be center of focus because it conveys an immediate message: "These are your priorities. You can't survive as a business without them." Putting the growth rate on the left and the key market comparisons on the right, allowed me to comfortably align and "bookend" the important information. The story in the middle is important, but it's not complete without the rest of the library on each side. Calling out critical numbers and year comparisons in the key markets (and providing some context with the gridlines on the secondary markers) allows this to be more than just a pretty image. This is a fully functioning, illustration that provides a wealth of information.

I worked to keep the fonts and font sizes consistent, as well as alligned all of the elements (right to left call outs, same distances from the edge, differnt types of text set to one style and size, etc.). I did need to adjust the text of the main message. It highlights what needs to be viewed and focused on.

I also wanted to preserve truth by giving he source and confesion about the data omission, should anyone ever want to replicate this information.

I'm really quite proud of this graphic. Here it is:

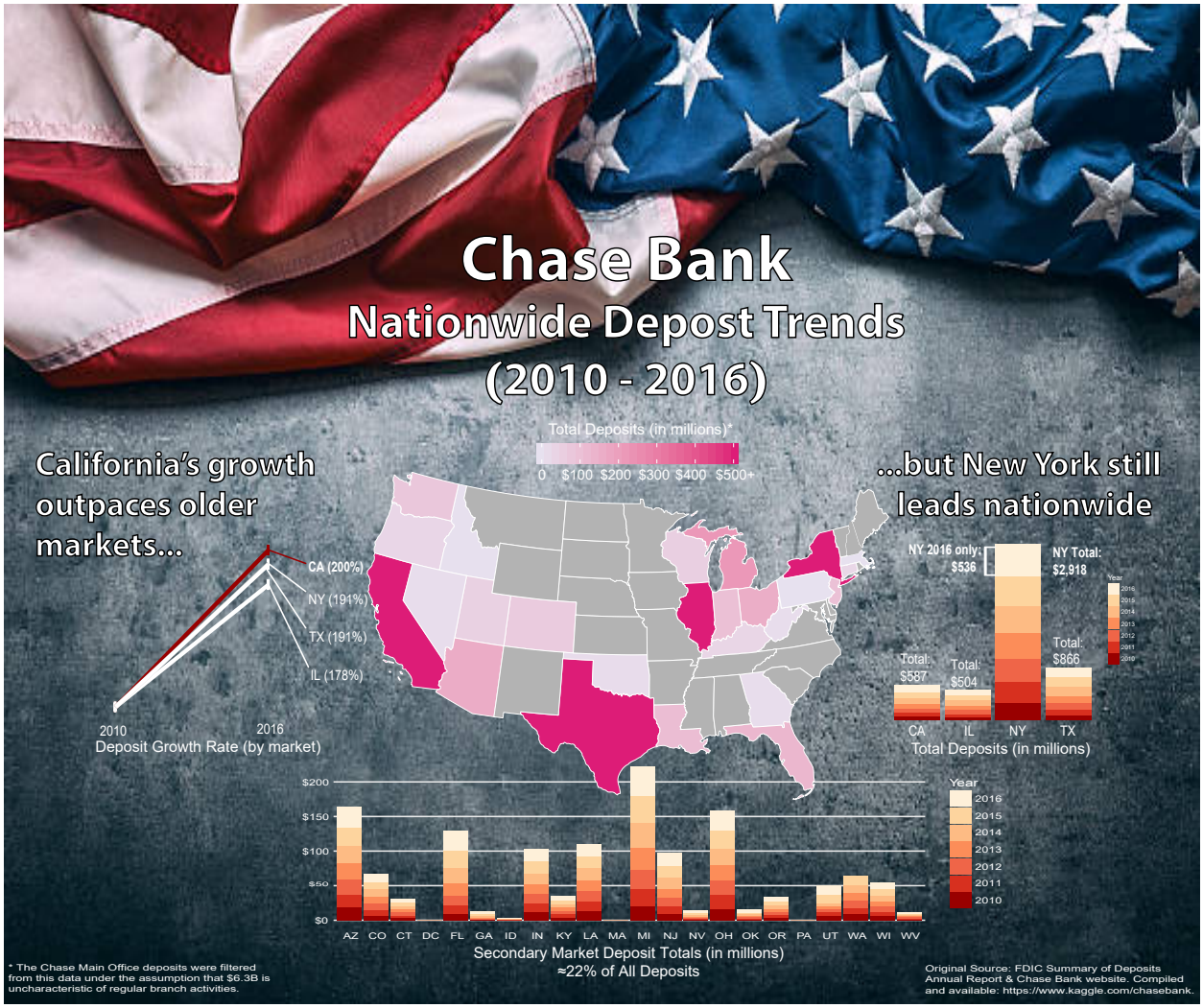


Figure 1: Final Project