Start-up Location in San Francisco

by

Joseph Barjis

Table of Content

PROBLEM DESCRIPTION

For years, I was observing that Data is the new oil, new natural resources. However, to make this resource useful, accessible, and easy to understand, a special service is needed to train and consult companies on their own data. Building in-house capacity takes longer to do so without initial external consulting help. Therefore, I have been building such a service already for years. Now, we (a team of 6 people) want to set up our office in an area with a rich concentration of Fortune 500 companies. The choice is California. In California, we will see which part contains most of the Fortune 500 companies. Our hypothesis is that would be in Bay Area, between San Francisco and San Jose.

However, to choose the right place in either city, one has to study a set of problems: Proximity to major companies, i.e., Fortune 500, affordability of office space, ease of access. In this project, the aim is to:

- Create a list of major companies (Fortune 500) in the US
- Reduce the list to only the ones with headquarter in California
- In CA, find out an area with most concentration of Fortune 500 headquarters

This insight already will help to decide on a suitable location for our start-up company.

Who would Benefit

The results of this project would benefit new start-ups or established companies seeking struggling with a location that has long-lasting strategic importance for their growth.

DATA DESCRIPTION

The data used in this project were acquired from two sources, each list below. From a Wiki list, I was able to get a general list of large companies in the US, over 500 entries. The list did not contain the companies location other than name of the city and state. I used the second source to get the Latitude and Longitude for each headquarter.

Based on these data, I will draw a map of locations that are favorable to other big companies (Fortune 500).

This exploration of locations should help me to recommend the right location for our start-up Data Science company.

DATA SOURCES

https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue
https://www.latlong.net/category/cities-236-15.html

Methodology

From this point starts the main body of the work as I will explain sub-section by sub-section how I completed this project from data collection, to cleaning, processing, and visualization.

Data Collection

The project used a Wiki page with a list of largest companies in the US. The web page processing resulted in three dataframes. The second one was the one with data I needed, i.e., name of the companies and their headquarter location.

```
#Use the Wiki page provided as an input
#url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

url = 'https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue'

df = pd.read_html(url) #use url readign method
print(len(df))
```

```
3
```

```
df = pd.DataFrame(df[1])
df.head()
```

	Rank	Name	Industry	Revenue(USD millions)	Revenue growth	Employees	Headquarters
0	1	Walmart	Retail	523964	NaN	2200000	Bentonville, Arkansas
1	2	Amazon	Retail	280522	NaN	798000	Seattle, Washington
2	3	ExxonMobil	Oil and Gas	264938	NaN	74900	Irving, Texas
3	4	Apple	Electronics	260174	NaN	137000	Cupertino, California
4	5	CVS Health	Healthcare	256776	NaN	290000	Woonsocket, Rhode Island

Select only those companies that have headquarter in CA

```
df_ca = df[df['Headquarters'].str.contains("California")]
df_ca
```

	Name	Industry	Headquarters
3	Apple	Electronics	Cupertino, California
10	Alphabet	Technology	Mountain View, California
14	Chevron	Oil and Gas	San Ramon, California
29	Wells Fargo	Financials	San Francisco, California
45	Intel	Technology	Santa Clara, California
46	Facebook, Inc.	Technology	Menlo Park, California
49	The Walt Disney Company	Media	Burbank, California

Create a dataset of Lat and Long of the US cities from available resources on the web

Notes:

- It could have been done through a lop as well, which I will do later and replace the multiple lines
- geolocator and foursquare does not return the right Lat and Long for big cities, while it does well with small cities

	Place Name	Latitude	Longitude
0	Port Hueneme, CA, the US	34.155834	-119.202789
1	Auburn, NY, USA	42.933334	-76.566666
2	Jamestown, NY, the US	42.095554	-79.238609
3	Fulton, MO, USA	38.846668	-91.948059
4	Bedford, OH, the US	41.392502	-81.534447

Results

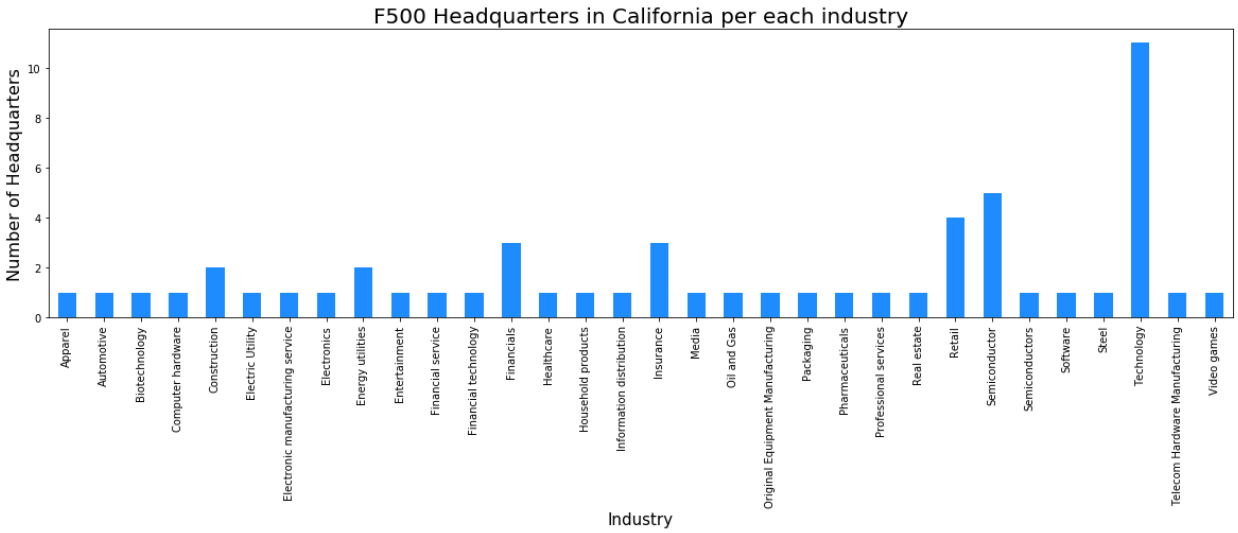
First of all, finally, I was able to create one dataframe that contains the list of all Fortune 500 companies headquartered in California and added to each company their Long and Lat coordinates. I will also build a few visual reports that shows the final results.

	Rank	Name	Industry	Headquarters	Latitude	Longitude
0	3	Apple	Electronics	Cupertino, CA	37.323000	-122.032200
1	10	Alphabet	Technology	Mountain View, CA	37.386051	-122.083855
2	14	Chevron	Oil and Gas	San Ramon, CA	37.764400	-121.954000
3	29	Wells Fargo	Financials	San Francisco, CA	37.773972	-122.431297
4	45	Intel	Technology	Santa Clara, CA	37.354107	-121.955238

Data Visualization

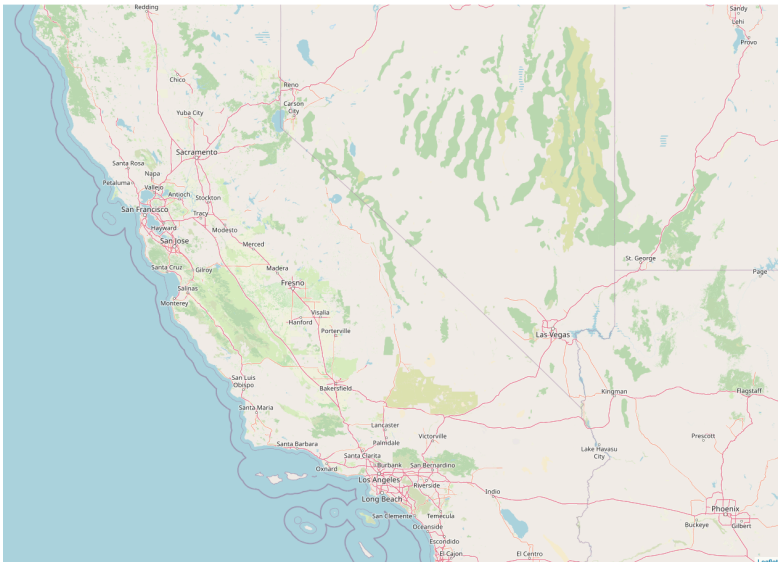
Now, having one single Dataframe that contain all the attributes I need, I do some useful visualizations. But first, I have to get all the libraries in place.

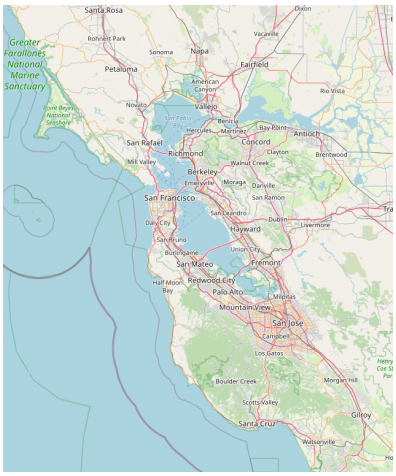
Lest have a look what industry represents most of the Fortune 500 companies in CA. As seen below, it is Technology and second on eis Semiconductor



Map Generation

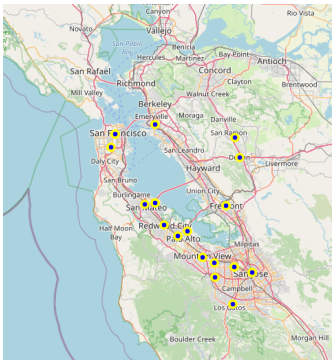
Here, I do one final important visualization. First, drawing CA map. Then, Bay Area Map. Then, show concentration of Fortune 500 headquarters in the Bay Area (between San Francisco and San Jose)





Superimpose Fortune 500 Headquarters onto the Map of Bay Area

As a final step, now we superimpose the headquarter of Fortune 500 companies that are located in the Bay Area, which was our primary goal for this project.



Discussion

This capstone project allowed to dive much deeper than any previous courses. It revealed that Data Science as much hard and technicla skill as it is soft skill, imagination, and creativity. While of course accurate data collectiona nd avilability is crucial, but the many ways of presenting and visualizing it is even more important.

For me, it is a strat of a new career. I took these courses to shift my career from Agile Development of Complex Enterprise Solutions to Data Science or a hybrid of the two.

Conclusion

This project is rather a strat of much longer journey, in which I want to collect and visualize all type of data about Fortune 500 companies. They year by year performance, relocation of headquarters and their reasons, moving up and down the list of ranking, changes in industry, etc. However, here, I focused on California and specifically on Bay Area as we have vested interest in it.

One special thanks to the fellow leaner for gradign this humble work!