**

A picture containing clipart

Description automatically generated

# GitHub URL

<https://github.com/kd-ucd/UCDPA_kierandowling.git>

# Abstract

A brief analysis of the Fortune Top 1000 using Python, featuring visualisations and geospatial representations of the data.

# Introduction

In this project I set out to analyse data of the Fortune Top 1000 companies. I chose this project because I wanted to learn more about the largest companies in the world. These companies have become such common household names, from Alphabet to Apple to Berkshire Hathaway and many more. As marketers, we often look to these champions of industry for inspiration, such as product and marketing ideas, strategies and positionings. It is beneficial to analyse these companies from many different angles and implement any learnings into our own organisations.

# Dataset

The dataset I choose is a list of the top 1000 companies in the world, with the data collated from Fortune.com. The dataset contains 1000 rows and 18 columns. The columns describe key information about each company as well as some details as to how they relate to previous iterations of the list (e.g. Newcomer and previous rank).

* Company
* Rank
* Rank Change
* Revenue
* Profit
* Num. of Employees
* Sector
* City
* State
* Newcomer
* CEO Founder
* CEO Woman
* Profitable
* Prev. Rank
* CEO
* Website
* Ticker
* Market Cap

# Implementation Process

To begin, I installed the Python packages I knew I wanted to use in the project. To do this, I used the package manager in Pycharm. The packages I expected to use were Numpy, Pandas, Matplotlib, Seaborn, Bokeh and Geopandas. The installation process was seamless though I did have difficulties installing GeoPandas, which I will discuss in due course.

With the packages installed, I imported the Fortune 1000 dataset as a data frame using the pd.read\_csv function in Pandas. I stored the data frame in a variable called ‘f1000’. I examined the data frame using the .info() function. Next I got the sum of all of the null values per column to assess what degree of cleaning and manipulation was required. There were few nulls in some columns, e.g 62 nulls in ‘Ticker’ due to these being private companies and not possessing a ticker symbol. However there were 500 nulls in the ‘newcomer’ column, which describes whether or not a company is new to the list. As I would not be using this column for further analysis, I decided to drop it from the data frame. I chose to do this in place with the “inplace” argument of the drop function, rather than create a new variable and use memory unnecessarily.

It can be difficult to get context when large datasets so I decided to generate some basic summary statistics. To do this, I calculated the mean and median revenues of the companies using the respective Numpy functions. Revenue in the dataset is listed in millions of dollars so I needed to format the text to make it more legible. I did this with the '{:,}'.format() string formatting.

Next I decided to visualise the relationship between company revenue by sector. To visualise this, I created a horizontal bar chart using Matplotlib.

Each company is ranked according to its revenue in the Fortune dataset. While this is an important metric, I wanted to understand whether these companies are profitable. I initially attempted to use a bar chart in Matplotlib but ultimately used a countplot in Seaborn.

From here I decided to focus on the 10 most profitable companies for further analysis. I thought it would be interesting to compare the revenues of the 10 companies by means of a horizontal bar chart. I again used Matplotlib for this.

Given that these are the most successful companies in the world, I wanted to extract the individual CEOs to learn more about who these leaders are. This was achieved by looping through the data frame with a for loop and the .iterrows() function in Pandas.

Next, I decided to create an interactive visualisation using Bokeh. To do this, I imported figure and ColumnDataSource from Bokeh.plotting. I set my ColumnDataSource to my dataframe and stored it as a variable called source. I then created a figure of a scatter plot comparing market cap vs revenue.

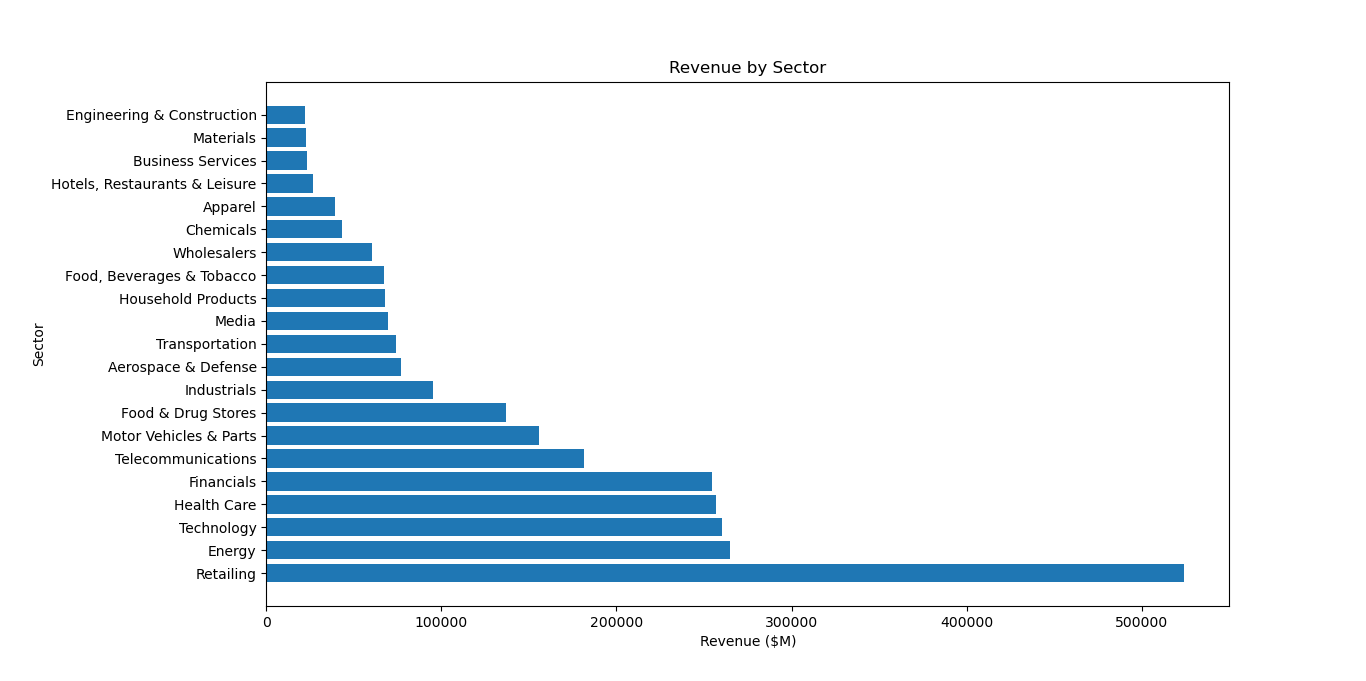
The most challenging aspect of this project, from an implementation perspective, was installing GeoPandas and running it from within the PyCharm IDE. Initially I tried to install GeoPandas through the in-built package manager in PyCharm. This approach was unsuccessful as GeoPandas requires a number of other supporting packages. Next I installed Anaconda, a Python distribution platform. Trying to install GeoPandas through either the command line or Anaconda Navigator threw up conflicts with other packages in the base environment, likely due to my earlier installation attempts.

Finally, I was able to successfully install GeoPandas by creating a new environment in Anaconda Navigator and installing Geopandas there. I set this new environment as my Python Interpreter in the Pycharm settings menu. Once I validated that Geopandas had been installed correctly (by calling it in the Python console), I reinstalled all of the other packages that I had installed previously and was able to resume working on this project.

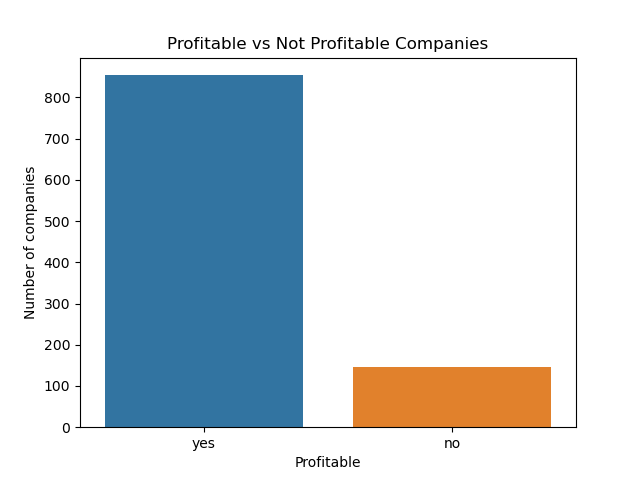
Once I was able to get GeoPandas working I decided to plot the GPS locations of the top 10 companies on a map. I manually collected GPS coordinates of each company from Google Maps and collated this information into a dictionary. To combine this data with top companies data, I converted the dictionary in a Pandas dataframe using the pd.toDataFrame() function. I then performed a merge with the .merge() function. With the data merged, I was then able to plot the coordinates of each company with a Seaborn scatter plot. I layered this scatter plot over a map using GeoPandas and a shapefile obtained from the United States Census Bureau.

# Results

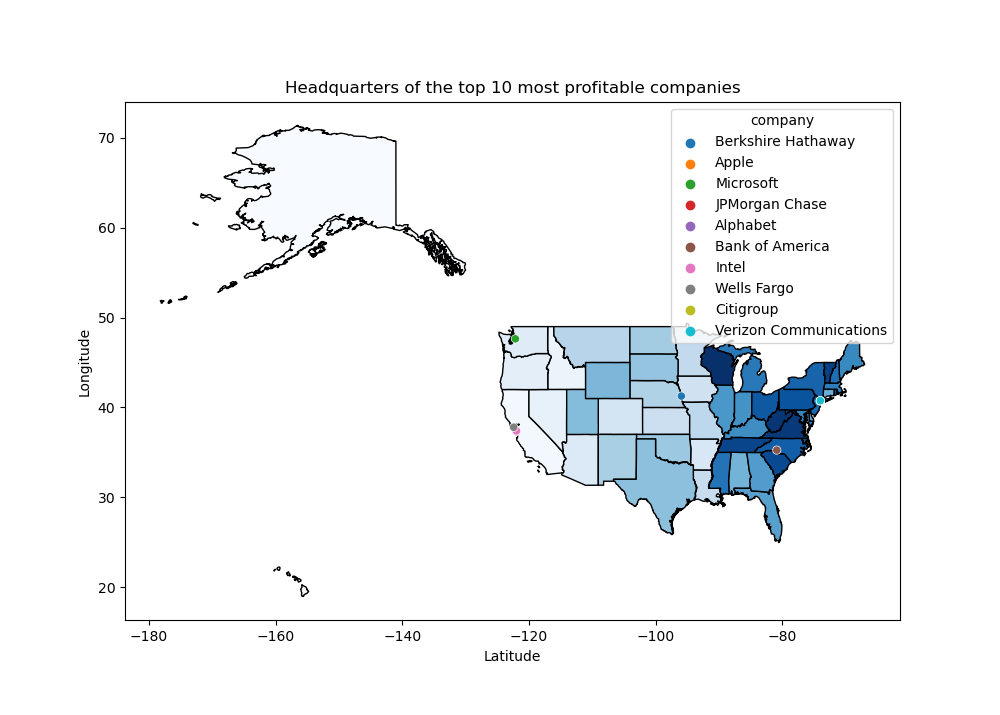
The average revenue is $15,902,339,000 and the median is $5,647,000,000

As can be seen in the figure below, the retail sector generates the most revenue of all the companies in the dataset. Engineering, Construction and Materials generate the lowest. 

To answer the question of how many companies are profitable, 854 of the top companies are profitable while 146 companies are not. The ratio can clearly be seen in the figure below.



The interactive visualisation of Market Cap vs Revenue has limited functionality. This aspect of the project requires further refinement to generate a figure that is more interactive, more practical and more insightful.

The figure below shows the GPS locations of the top 10 most profitable countries on map. As can be seen in the image, all 10 companies are based in America, primarily on the coasts. 

# Insights

1. The average revenue is $15,902,339,000 and the median is $5,647,000,000
2. CEOs for the Top 10 most profitable companies are:

|  |  |  |
| --- | --- | --- |
| 1 | Berkshire Hathaway | Warren E. Buffett |
| 2 | Apple | Timothy D. Cook |
| 3 | Microsoft | Satya Nadella |
| 4 | JPMorgan Chase | James Dimon |
| 5 | Alphabet | Sundar Pichai |
| 6 | Bank of America | Brian T. Moynihan |
| 7 | Intel | Patrick P. Gelsinger |
| 8 | Wells Fargo | Charles W. Scharf |
| 9 | Citigroup | Jane Fraser |
| 10 | Verizon Communications | Hans E. Vestberg |

1. The retail sector generates the most revenue
2. The majority of companies are profitable
3. The top 10 most profitable companies are headquartered in America.

# References

Fortune 1000 - Kaggle

<https://www.kaggle.com/winston56/fortune-500-data-2021>

Google Maps

[maps.google.com/](http://maps.google.com/)

United States Census Bureau - Cartographic Boundary Files - Shapefile

<https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html>

GitHub Repro

<https://github.com/kd-ucd/UCDPA_kierandowling.git>