**Final Project Report**

**IST 652 – Scripting for Data Analysis**

**M002: Spring 2020**

**Kshitij Sankesara (913789324)**

**Rishabh Agarwal (876424625)**

**Data and its Source:**

This dataset is about the Indian Startup Ecosystem. The dataset can answer questions like what type of startups are getting funded, who all are the important investors, and which fields get a lot of funding these days. The dataset is obtained from Kaggle (Indian Startup Funding) and is originally uploaded by trak.in. They have shared the data publicly for everyone to know more about the Indian startups. This dataset has all the information related to startup fundings starting from 2015 till 2020. Some of the significant variables are Date funded, Names of the funders, City the startup is based out of, and Amount invested. The dataset has 10 variables and around 3000 records. One row of data records one startup and their respective funding amount which they received. The dataset also has information on the Industry vertical of each startup and the type of investment made by each investor.

From this data, we can learn about the fundings secured by the startups in the last few years, the important investors and the start-up fields which are getting most fundings. We can also learn about the location where the highest funded start-ups are based. By analyzing the startup data, we can explore these things and can derive insights from it. We can learn about how well a startup is performing based on the funding’s received.

We have analysed another dataset about the startup ecosystem in India from Crunchbase and it contains 36 columns and 849 rows. The dataset contains funding received at each stage from 1999 to 2014. The most significant attributes are status (operating, closed or acquired), founding date, last funding date, number of funding rounds, first funding date and quarter and funding received at each stage (seed, venture, angel, equity, round A to H, etc.) This dataset contains less records as the startup culture had not picked up until 2013 in India. This dataset contains more information to understand factors affecting the funding amount received.

**Data Preprocessing:**

The original data set was of 10 variables and 3044 records. All the records have some data in it, so we didn’t remove any of the record. We removed couple of columns from the data. The Remark column had around 90% null values, so it wasn’t helping for our analysis. Another column which we deleted was the Serial number column as this was not significant for our analysis. We checked data types of the remaining columns to have a better idea of them while analyzing it.

We found out that there are null values in our data set. For ‘Industry Vertical’ and ‘SubVertical’ columns, we replaced all the null values with ‘Not Specified’. We did the same with the ‘City Location’ column. For the ‘Investor Name’ and ‘Investment Type’ we replaced all the null values with ‘Other’. For our target variable, that is ‘Amount’ we replaced all the null values with the mean of the column. By doing this, we made sure that this thing isn’t affecting our overall analysis.

Later, we renamed all the columns in our data set. The previous names were long with some spelling errors and also had special characters which didn’t make much sense. We replaced all the names using proper naming conventions. For example, we replaced ‘InvestmentnType’ with ‘InvestmentType’. While we started to analyze the data, we realized that the Date column had timestamp which weren’t proper. So, we fixed the timestamp and converted the Date column to datetime type using Pandas. We extracted the Year and Month from the Date column and made two new columns for it. We wanted to use both these columns for our data models. The Date column had some NA values too. So, we replaced them using 0 in our Year and Month columns.

There were some strings in the Amount column which didn’t make sense. We replaced all the strings with mean value of the Amount column. All the variables which had text data were not properly recorded. We checked the unique values in each column to understand these types of variables. Some strings were capitalized whereas some weren’t. Some had two names for the same city as ‘Bengaluru’ and ‘Bangalore’. This is one city but was named in multiple ways. We replaced such strings with one proper name. Then there were strings which has any random character in it. We tried understanding each of them and replaced them with their closest meaning. Some data were written in many different ways like E-commerce, e-commerce, E-Commerce, etc. This was an issue when we used the groupby function to analyze the data. So, we replaced each of them with one proper name for our analysis. This was one of the major issues while analyzing the data. We used the describe function to understand our quantitative variables like Amount. We knew their mean, percentiles, minimum and maximum using the describe function in Python.

The second dataset taken from Crunchbase originally contained 849 rows and 36 columns from which the country code and founded quarter column is removed. The founded column quarter column is removed as it contains another similar column found quarter.

The market (33 values), status (17 values), city (5 values), founded date (166) and founded quarter (167) contained missing values which haven’t been replaced with another value as it would create inconsistencies in the data if it is replaced with another value.

The founded date, first funding and last funding date columns are in the object type which is converted to datetime. From the first funding date column, the year and month have been extracted into separate columns for our analysis. The amount\_raised column was also in the object type which has been converted into integer type.

**Method of Analysis:**

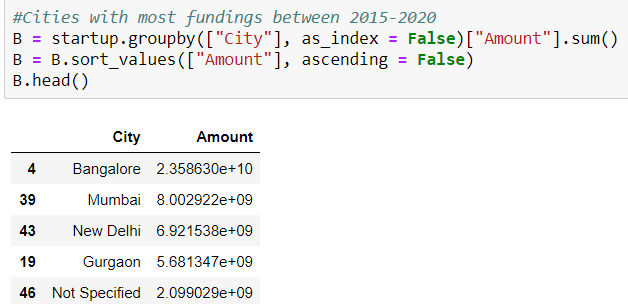
1. *Which Start up received the highest funding amount between 2015-2017 and 1999-2014?*

For our first business question, we started by analyzing the startup, their Industry and the funding amount which they received. By doing these, we came to know that companies like Flipkart, Paytm, Rapido Bike were the startups which received the most funding between 2015-2017. We also came to know the fundings were for which Industry. We grouped the data using Startup and Industry and then took the sum of the Funding Amount. We arranged the data by decreasing funding amount. By doing this, we came to know that Flipkart has received the highest funding amount followed by Rapido Bike Taxi and Paytm.

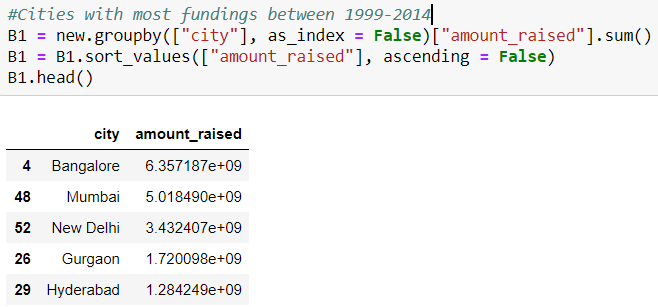
Similar steps were followed to compute the same between 1999-2014. The companies which had received the highest funding were Etable, DecideQuick and Travel Desiya respectively. The first two companies belonged to the social media segment and the last one in the travel segment.

1. *Which City recorded the most funding amount between 2015-2020 and 1999-2014?*

We wanted to analyze if city plays an important role in funding amounts. Here, we are finding cities and funding amount received by the startups of that city. Between 2015-2020, after analyzing the results, we can see that startups in Bangalore received the most funding amount. This is the reason why Bangalore is called as an IT Hub. Bangalore was followed by Mumbai which is another huge city with lots of start ups and then by New Delhi which too has a lot of companies.

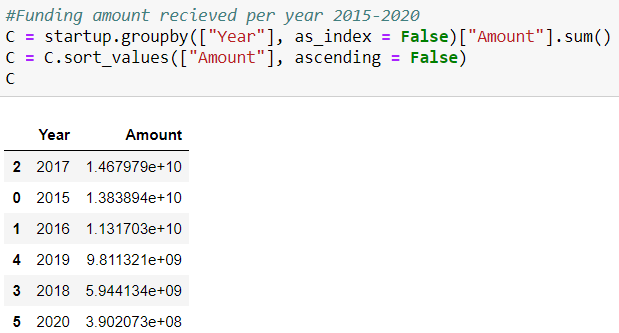


There was a similar trend existed between the time period 1999-2014.



1. *How funding amount fluctuated over the years (2015-2020)?*

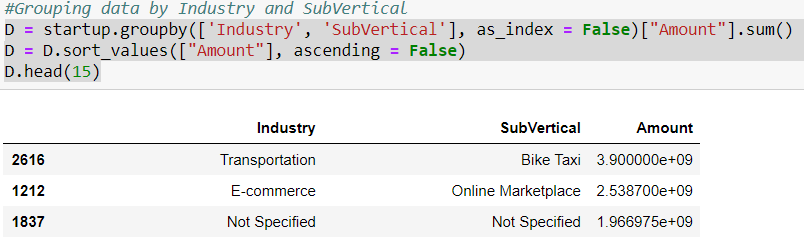
Here, we analyzed how the funding amount fluctuated over the years. We understood that the economic conditions play a major role in the overall fundings. The economy was booming during 2015-17 so the total funding amount received at that time were high as compared to 2018 and 2019. We do not have the entire data of 2020, so we won’t include it in our analysis. The highest amount received was in the year 2017 followed closely by 2015 and 2016.

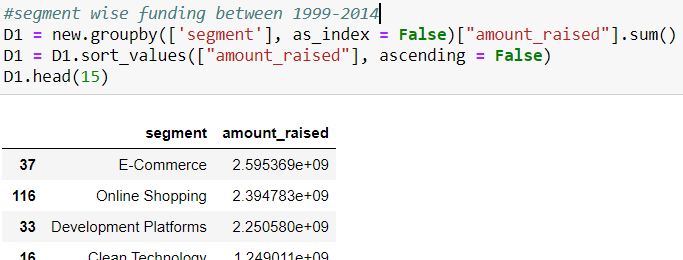


1. *Which Industry is favourable for Funding Amount for the time periods between 2015-2020 and 1999-2014?*

Knowing which Industry is likely to get more funding amount is very important. For the time period 2015-2020, we analyzed the different Industry and their SubVertical to understand the funding ecosystem. Transportation and E-commerce startups have received the most fundings. This is followed by FinTech and Online Marketplace companies.

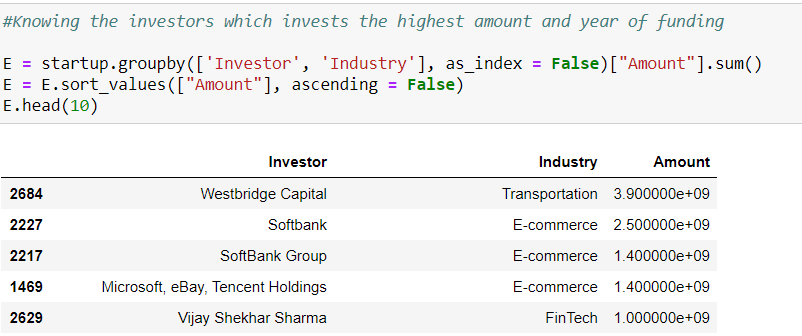
Similar steps were followed to understand the most favorable industry between 1999-2014, E-commerce, development platforms, clean technology and software respectively had received the most funding.





1. *Who are the top Investors?*

Westbridge Capital is the investor with the highest funding. We analyzed the top investors to know about the Industry and the amount funded. Top companies like Softbank, Microsoft, eBay, and Alibabe are the investors which invest heavily in Indian startups.

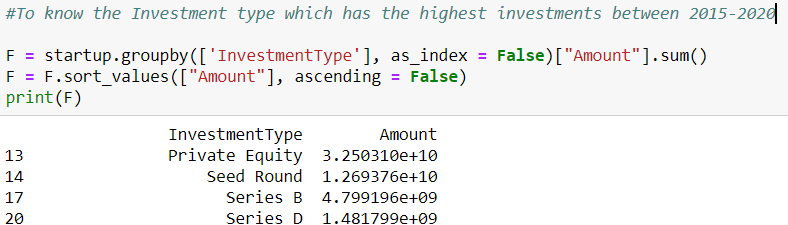


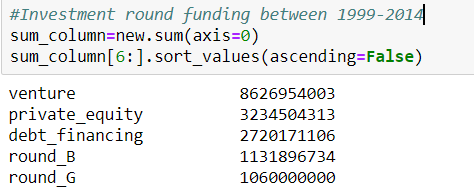
1. *Which Investment Type receives the major fundings between the 2015-2020 and 1999-2014?*

Between the years 2015-2020, major amount is funded during the Private Equity, Seed Round, and Series Rounds like Series B, C, and D. These are the times when investors invest heavy amount into the startup. Private Equity and Seed Round gets the most money.

We later analyzed the average amount of funding received by each company throughout the years. We have plotted some visualizations too showing our analysis. Our first graph is the histogram showing the months and the sum of the amount funded during each month. August and November are the times when heavy sum of money is invested. We later plotted a graph to show which type of Industry receives fundings the greatest number of times. Consumer Internet, Technology and E-Commerce are funded the most number of times. We have also displayed this information in a table and a Word Cloud too. Then, we displayed the correlation matrix to see if the Year and Month are highly correlated with the Funding Amount. All the three variables are not related to each other.

Between the time period 1999-2014, the highest funding has been received in the Venture, Private Equity, Debt financing and Round B respectively.



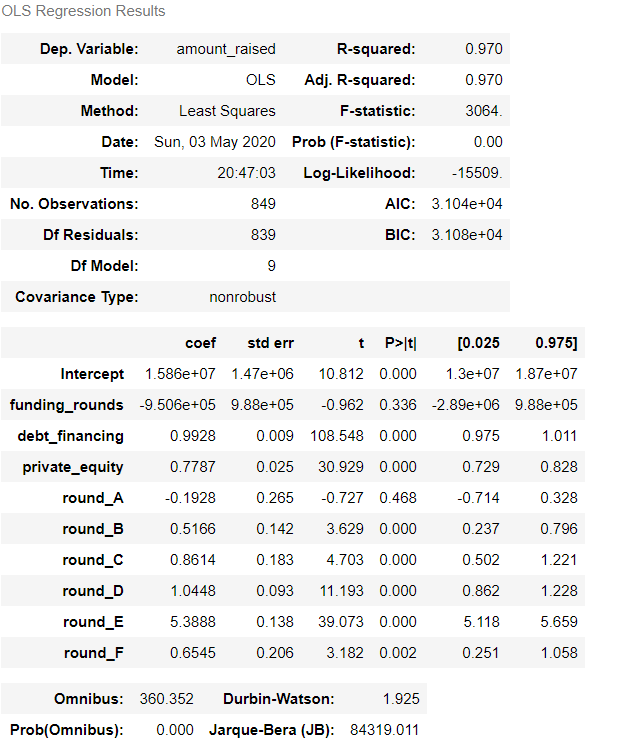


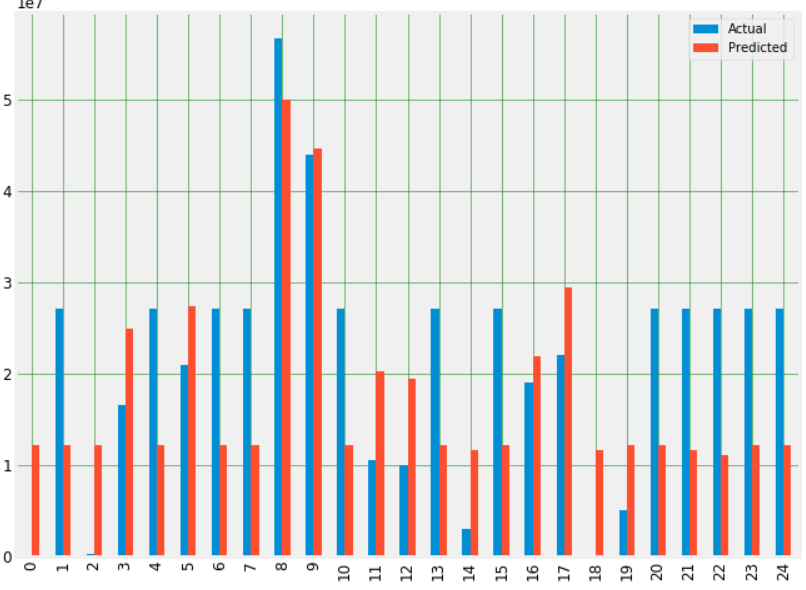
Regression Model:

We implemented a multiple linear regression model for our dataset contain information between the time period 1999-2014. Our target variable is the Amount and our independent variables funding\_rounds, debt\_financing, private\_equity, round\_A, round\_B, round\_C, round\_D, round\_F.

We splitted our data into training and test data sets where 80% of our data is training and 20% is testing data. We implemented the model on the training data and later tested in on our testing data set. We compared our actual values to our predicted values and have plotted a graph showing the difference between the two. For evaluating our model, we used the Mean Absolute error. Our mean absolute error was a bit high because of low number of independent variables and less data. We got the mean absolute error as 11981150, which is not bad considering the large range of Amount variable. We can improve our model by changing some hyperparameters and giving more data to the model.

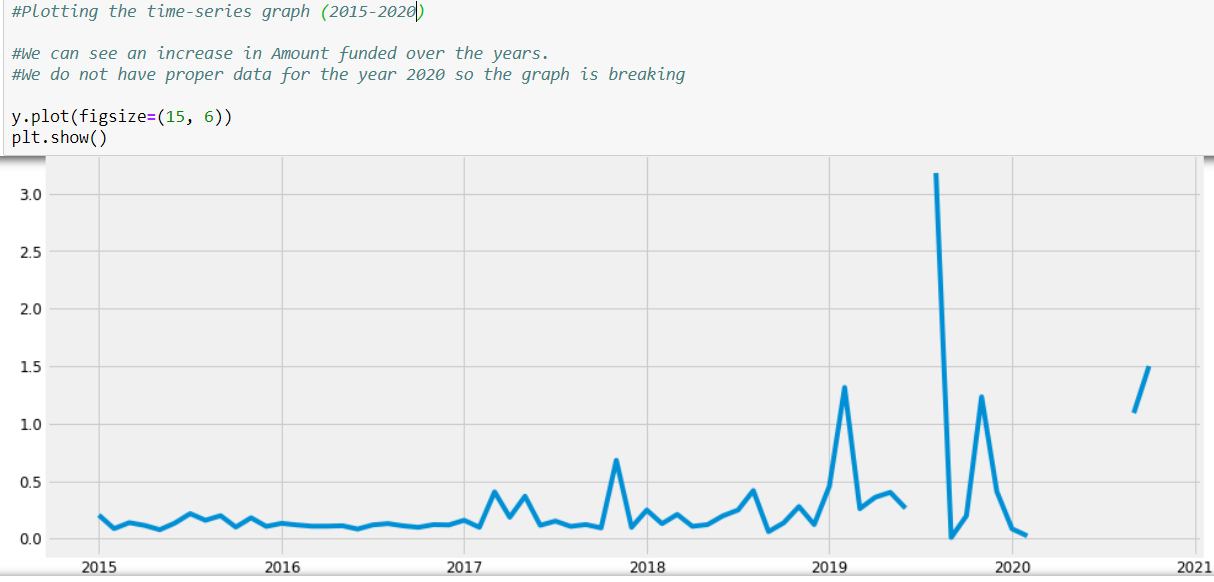
We couldn’t run the same for the time period 2015-2020 as the issue with Regression model is that it only works on numerical values. In that, we just have 3 variables with numerical values. We couldn’t convert our other variables to numerical data as there were a lot of unique values in each column. Converting all of them will be a tedious task.





Time-Series Model:

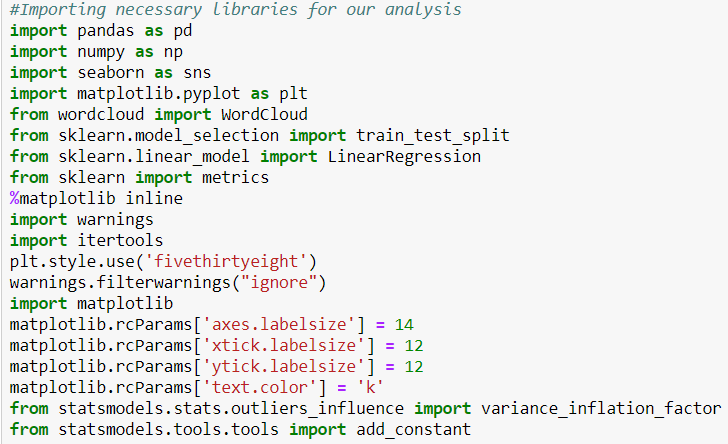
For our Time Series model, we used our date column and the Amount variable. We indexed our Date column to prepare the data for our model. We then plotted a time series plot to analyze the fluctuations of the Funding Amount throughout the years. Our data has some missing values, so the plot isn’t a proper line and have some gaps in between.



**Description and Output of Python Program:**

1. Libraries and Packages used

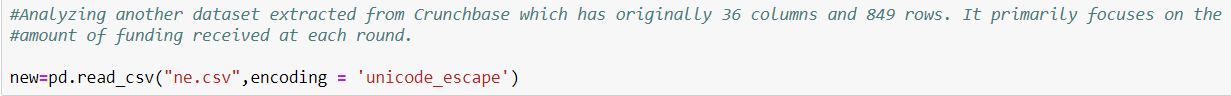
We have used the following packages for our analysis



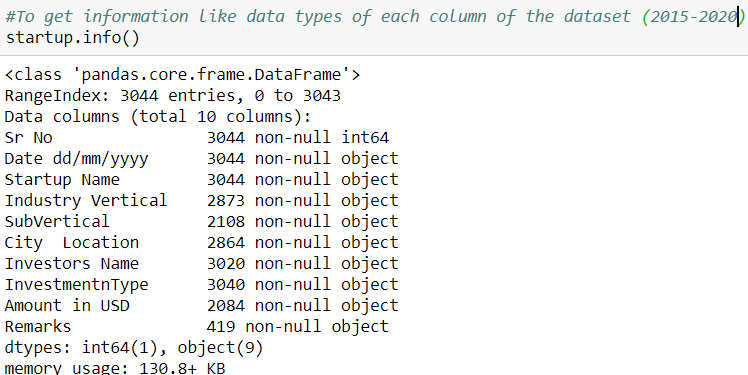
1. Data Munging

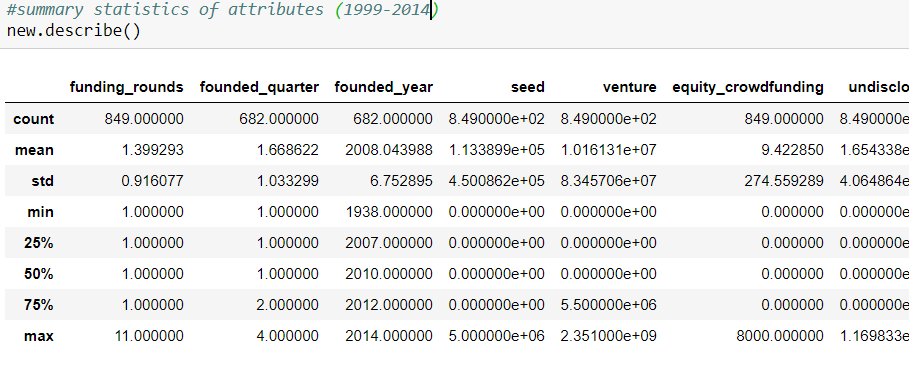
We first started by loading both of our datasets which are in the csv format using pandas to create a dataframe.



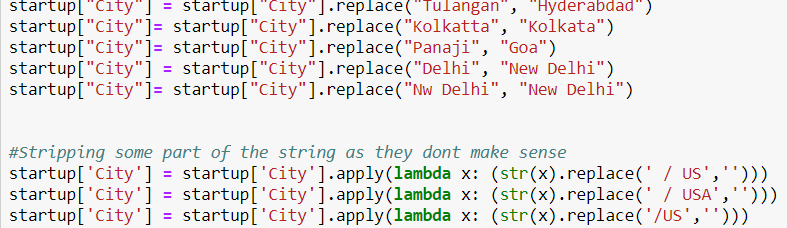


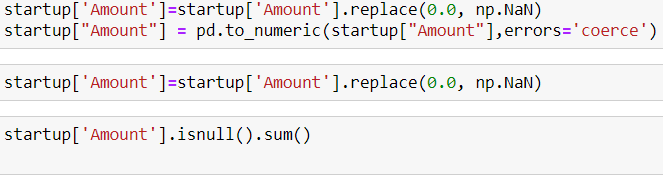
Then we found the information about the attributes in the dataset using pd.shape, pd.info(), pd.describe() and pd.isnull().sum() to handle the missing values and change the datatype accordingly.

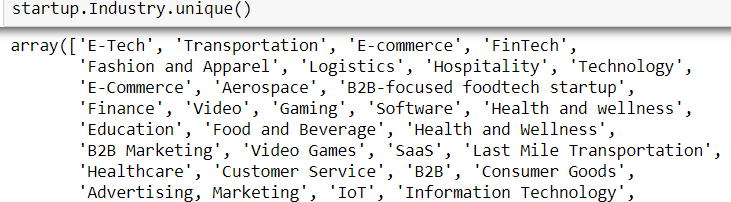




The inconsistencies in the values of the attributes are found by using the pd.nunique() and pd.unique() functions. The inconsistent values were changed by using the replace(), .loc() and apply() functions. The datatypes for the numerical and date attributes were changed using the pd.to\_datetime and pd.astype() functions. The NA values were changed using the fillna() function and for the numerical attributes the NA values were changed to np.nan and then replaced with their mean values.



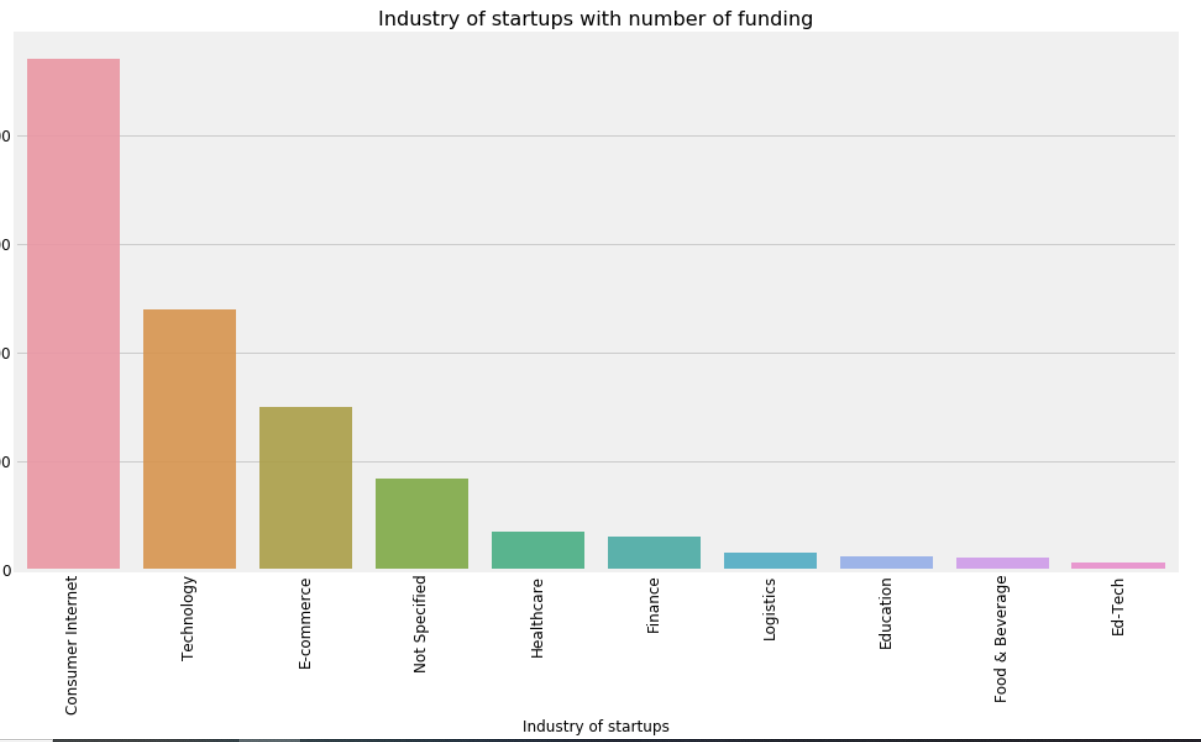


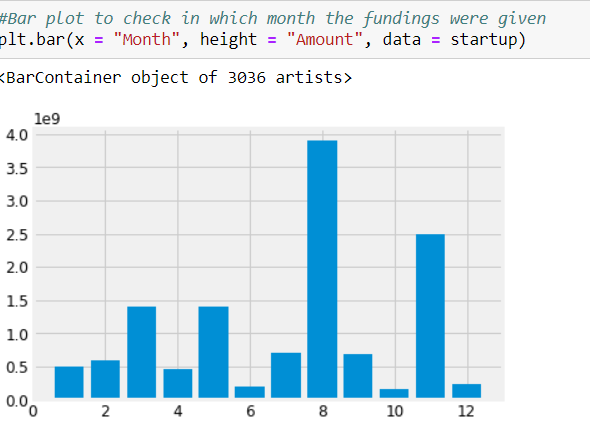


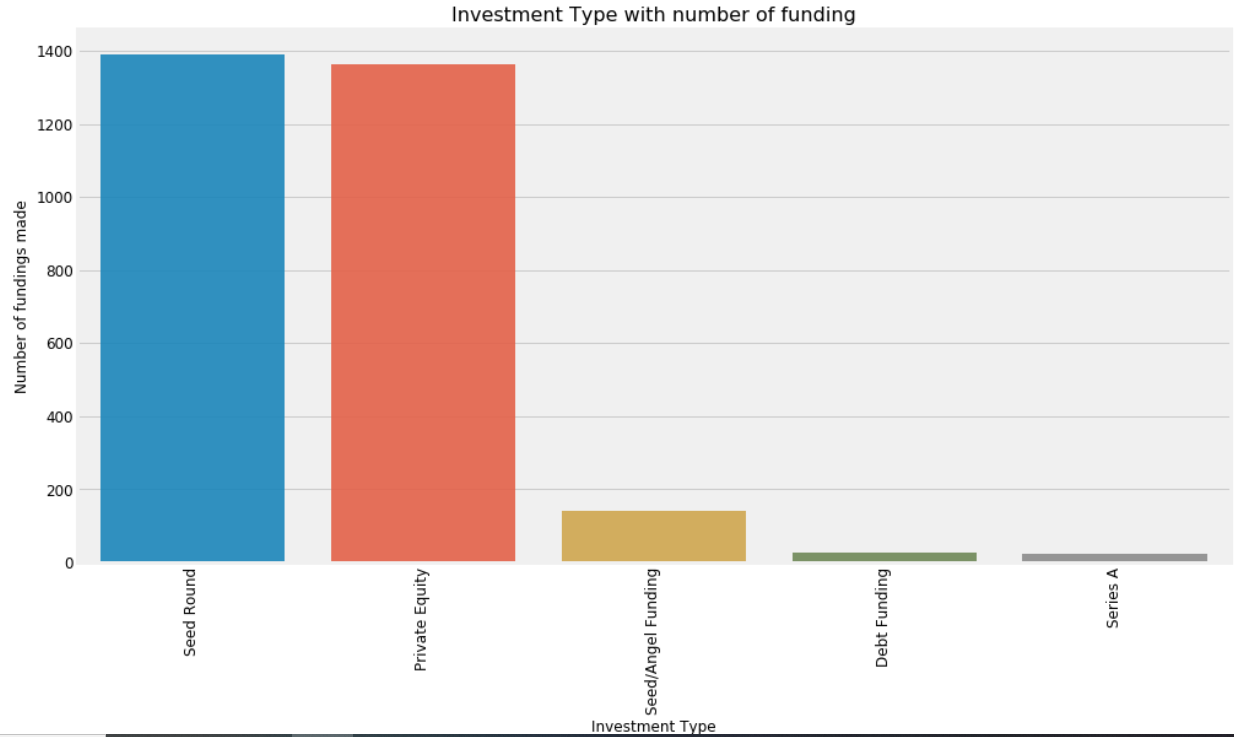
1. Visualizations

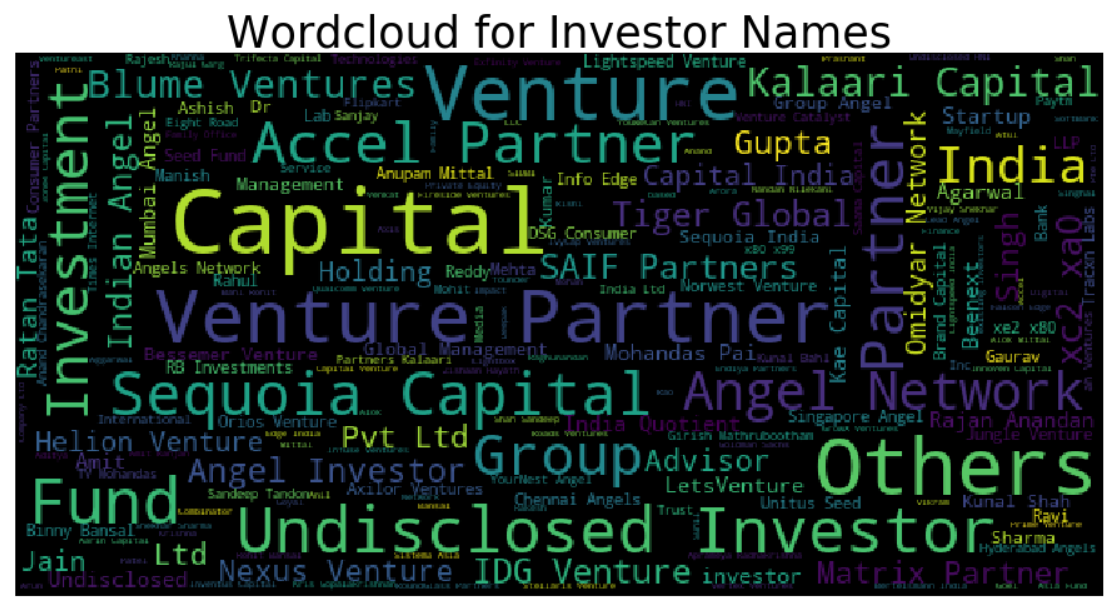
Visualizations were created using matplotlib, seaborn and wordcloud. Visualizations help in analyzing the data in a more easier way and they were created to analyze industries were the maximum funding were received, investment type fundings, month were the fundings were mostly made, etc.

Wordcloud was also created to analyze the text data particularly the investor names which would give us an visual depiction of the most common investors.







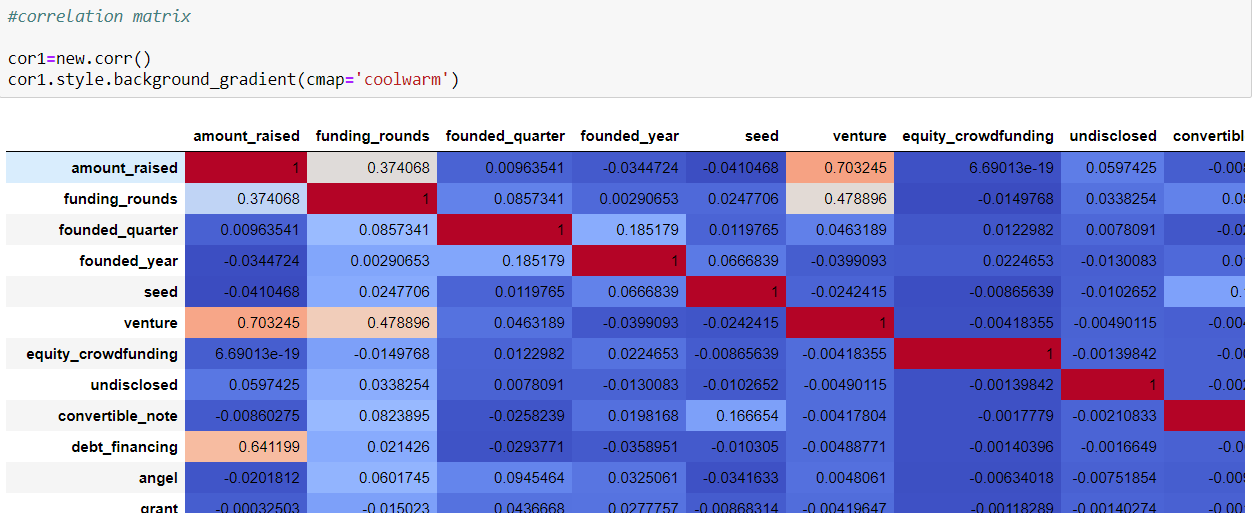


1. Business Questions

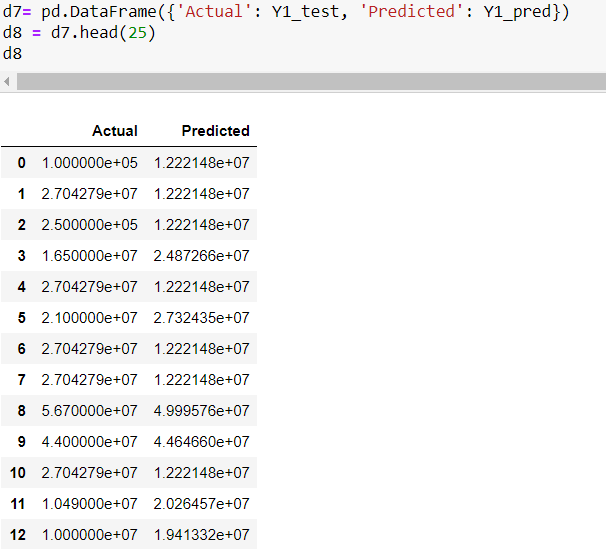
As discussed in the method of analysis section, the business questions were answered by primarily using the groupby() function which is an efficient way of grouping the data based on multiple columns. Questions such as year of maximum funding, segment wise funding, investors who have invested the maximum amount, etc. were answered.

1. Linear Regression and Time Series Model

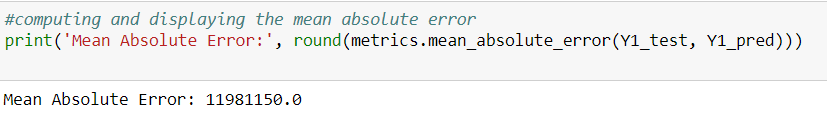
A correlation matrix was generated to analyze the variables most closely correlated with the amount received.



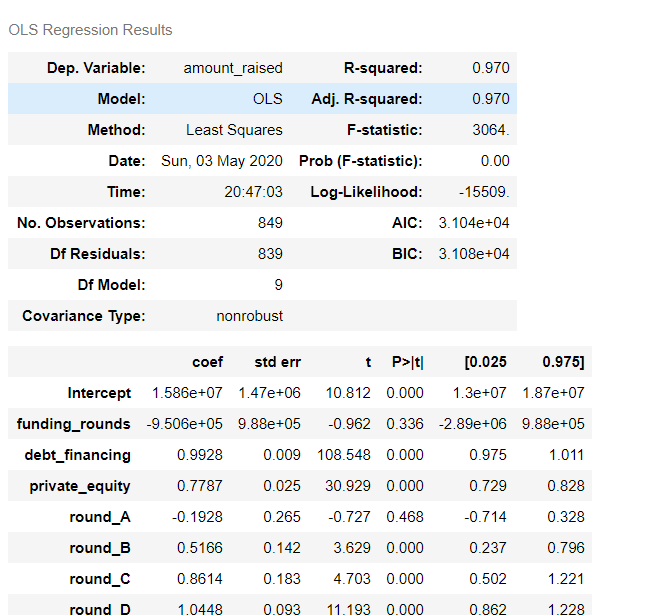
Then the linear regression model was run using the scikit learn package and the data was split into train and test data into a 80:20 ratio and then the predicted funding amount was found.



The mean absolute error was also computed



The OLS regression results were computed as well.



**Conclusions:**

* The period between 1999-2014 doesn’t contain many records as the startup culture had gathered momentum around 2014. The year 2017 and 2016 had seen the maximum amount of funding. Between 1999-2014 the total funding received was $ 22,959,329,135 and 2015-2020 was $ 55,981,419,364
* The transportation, e-commerce and software verticals are most preferred by investors with Flipkart seeing the maximum funding received as it was acquired by Walmart ($16 billion).
* The most fundings are received in venture, private equity and seed rounds.
* The investors who have invested the maximum amount are Westbridge capital, Softbank, Microsoft and Tencent.
* The cities with the highest number of startups are Bangalore, Mumbai and New Delhi. Bangalore is rightfully known as the Silicon Valley of India and Mumbai being the financial capital has the second highest number of startups.

**Group Tasks:**

Rishabh came up with this amazing dataset and it was his idea to analyze the data to know more about the Indian startups. Kshitij analyzed the data during the project proposal and formulated a plan regarding the siginificant variables and the models which we will use on this data. Later, we both worked remotely and equally contributed to the coding program. We made sure we covered all the parts which we had thought about during the project proposal. We both worked on the Report together on Google Docs.

**References:**

1. <https://www.kaggle.com/sudalairajkumar/indian-startup-funding>
2. <https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673>
3. <https://realpython.com/python-data-cleaning-numpy-pandas/>
4. <https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>
5. <https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b>