# **Newswire.com coding project - Report**

Total Run time = approx. 300 sec.

**Installation and execution Instructions:**

1. To run this code, you need Anaconda Spyder 3(preferred). which can be downloaded at "https://www.anaconda.com/download/".

1.1. To download Spyder 3 IDE, click on above link and download "Python 3.6 version".

1.2. upon downloading, start installation and follow the instructions given by install wizard for successful installation.

1.3. navigate to the directory where Anaconda3 is installed, open the directory and launch "Spyder".

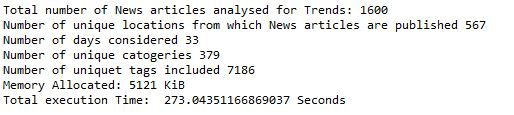
2. Clone the repo on to your PC and open "newswire\_trends.py" file in Spyder.

3. Run the "newswire\_trends.py" file by pressing F5 or Run button.

4. 4 plots will be created in the directory where you save the code to execute it.

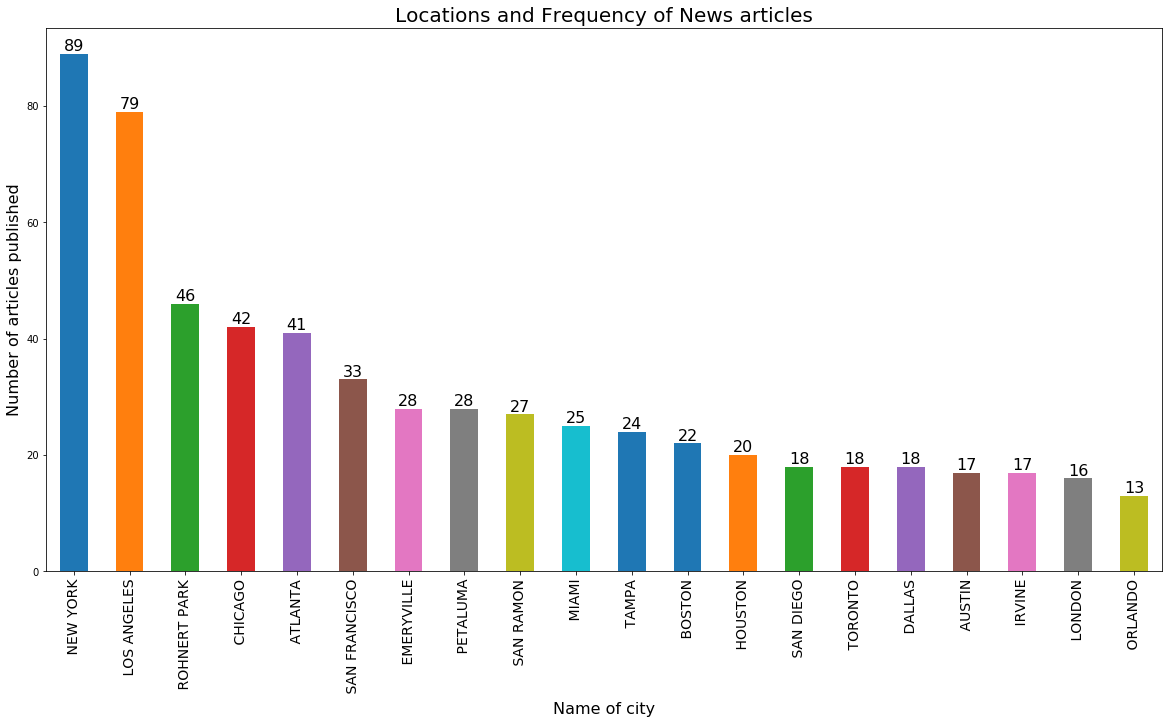
**Summary of output:**

1. Output is a result of analysis of 1600 News articles present in 1-50 pages of "https://www.newswire.com/newsroom" which in turn are 1600 separate web pages or documents.



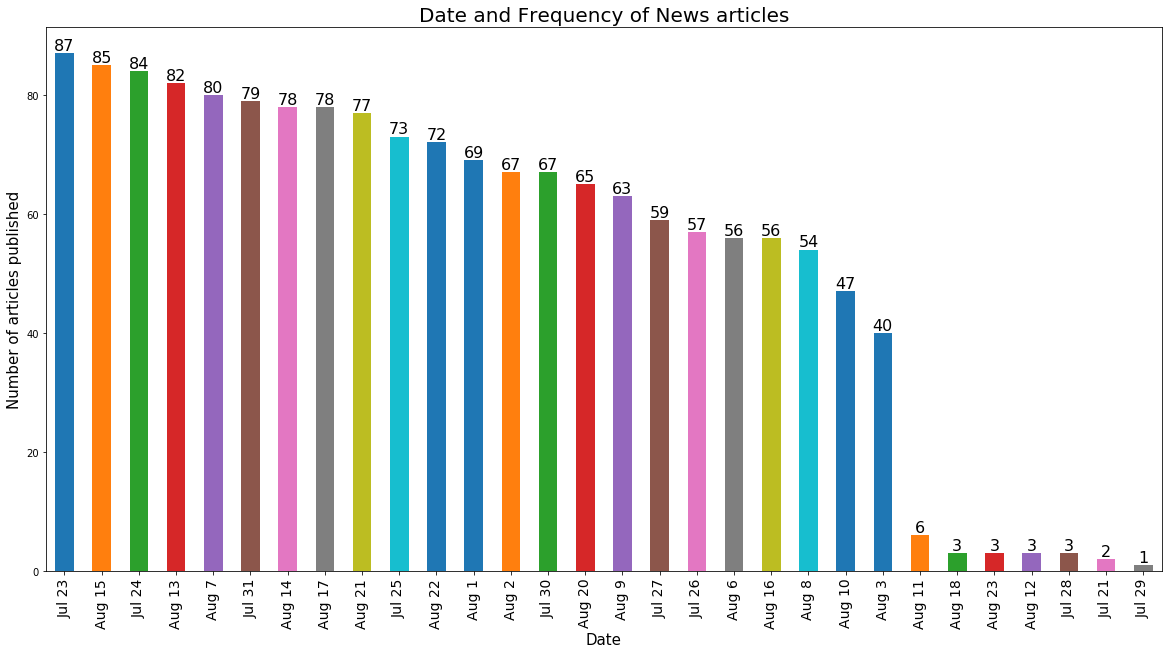
2. News articles were published from 567 locations, top 20 locations and the number of articles from the respective locations is plotted.

2.1. Highest number of the articles are published from New York, followed by Los Angeles.



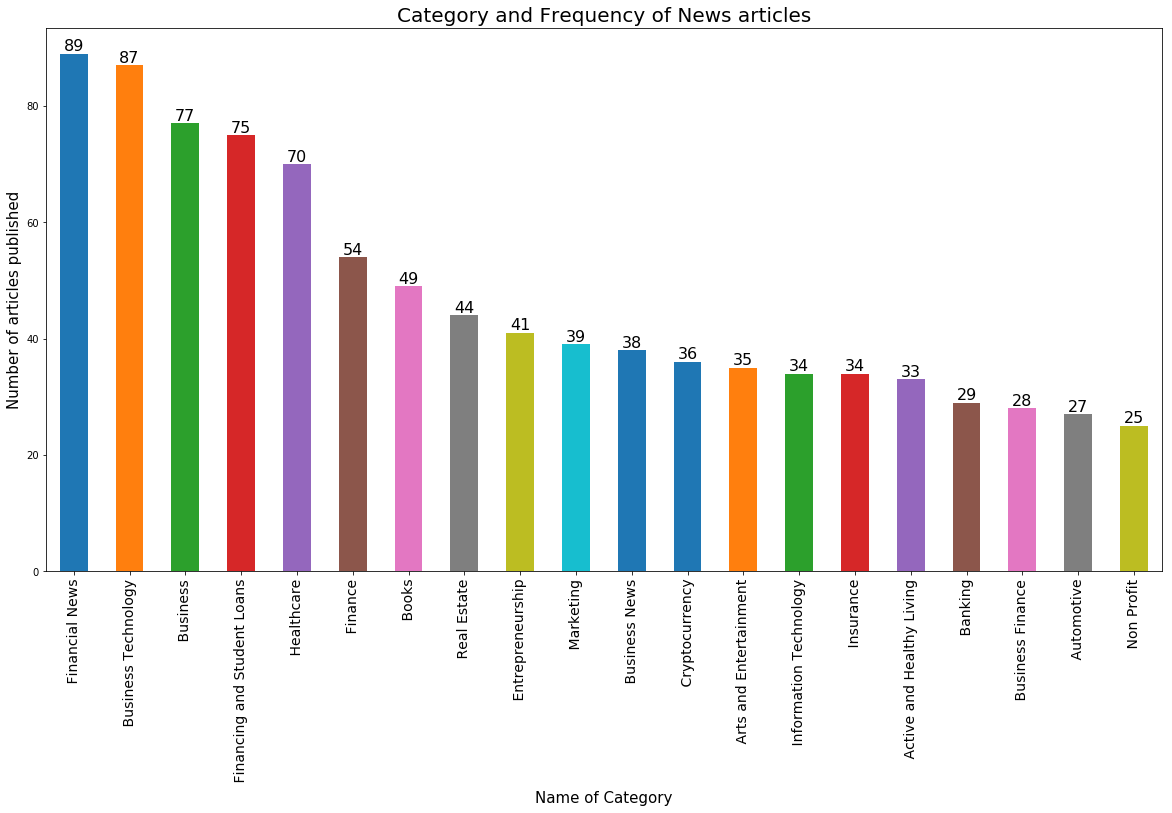
3. News articles from past 33 days (as of Aug 23, 2:50 AM) are considered under analysis, top 30 days and the number of articles published on those days is plotted.

3.1. Very few articles are being published on weekends.



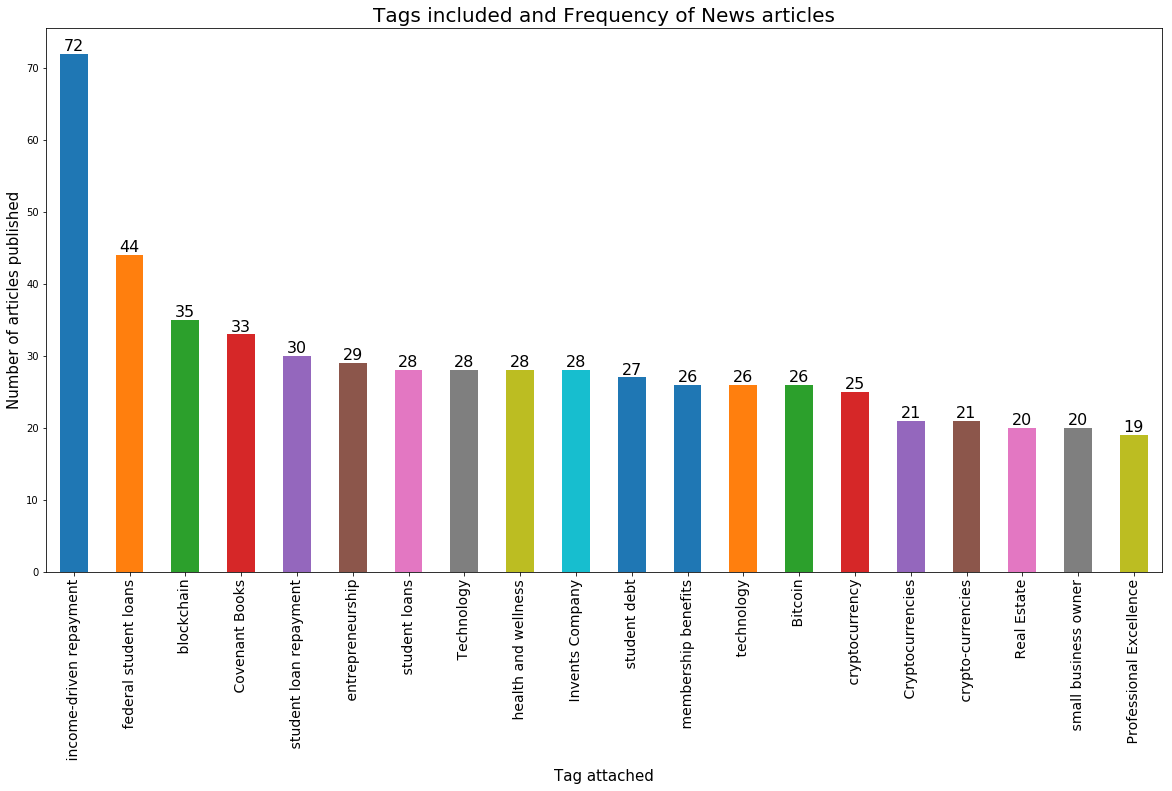
4. News articles were classified into 379 unique categories, top 20 categories and the number of articles published under them is plotted.

4.1. Finance and Business-related articles take a major share among all articles.



5. 7186 unique Tags were mentioned overall, top 20 tags and the number of articles published under them is plotted.

5.1. Student loans and Crypto currency related technologies were trending.



**Challenges Faced:**

1. missing data types.

2. lot of unstructured data.

3. Categories and Tags were mentioned under same HTML tag which demanded significant time for data cleaning.

4. significant time was spent on plots to get the desired labels in appropriate locations.

**Future Scope:**

There is a lot of scope for analysis from the obtained data. Though there are several plots available, Bar plots are best for understanding basic demographics. provided significant time, In-depth analysis can be carried out which could include

1. Sentiment analysis depending on variables (location specific, day specific, based on categories and tags mentioned).

2. Clustering locations and creating zones to carryout analysis.

3. Categories and tags can also be clustered based on type and can be used for better understanding of business.

4. Predictive model creation using various Machine learning techniques.

5. In-depth statistical analysis for various insights.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\* THE END \*\*\*\*\*\*\*\*\*\*\*\*\*\***