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

Total Run time = approx. 300 sec. (Spyder)

Total Run time = approx. 960 sec. (Command line)

**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.

**Alternative execution instructions using Command prompt:**

1. To run the code you should have Python 3 installed on your PC. If you don’t have Python 3 installed, download from "https://www.python.org/downloads/".

2. Follow the instructions and install Python 3.

3. Clone the repo on to your PC.

4. Open command prompt and open the downloaded directory.

5. There are various libraries used in the code. so, we have to install them first if we are using command prompt to run the code.

5.1. make sure you have pip installed which comes bundled with python 3. check it by simply typing "**pip**" in command prompt.

5.2. install pandas using "**py -3 -m pip install pandas**" command

5.3. install matplotlib using "**py -3 -m pip install matplotlib**" command

5.4. install bs4 using "**py -3 -m pip install bs4**" command

5.5. install lxml using "**py -3 -m pip install lxml**" command

5.6. use the above syntax and install any other missing modules.

5.7. once you are done with all required installations, enter "py newswire\_trends.py" on command line and hit enter to start execution.

6. you will see the plots generated as execution reaches end, close the generated plot to get next plot (Don’t worry plots will be saved in the current directory from where you are running the code.)

**Alternative execution instructions using default python IDE:**

1. follow above mentioned steps 1-6 to install Python 3 and required modules.

2. Instead of running through command prompt, you can open python 3 IDLE, it will open ‘Python 3 shell’ in that go to 'file' menu / tab present at the top. In file menu select open.

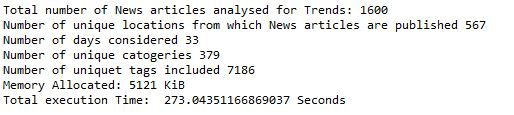
3. A window opens where you should navigate to the cloned repository and open "newswire\_trends.py" file.

4. The "newswire\_trends.py" file will open in a new IDE window where you can run the code by pressing 'F5' function key or selecting 'Run Module' option from Run menu on top.

5. Execution will start, and results will be displayed in python 3 Shell. output images will be saved in the current directory.

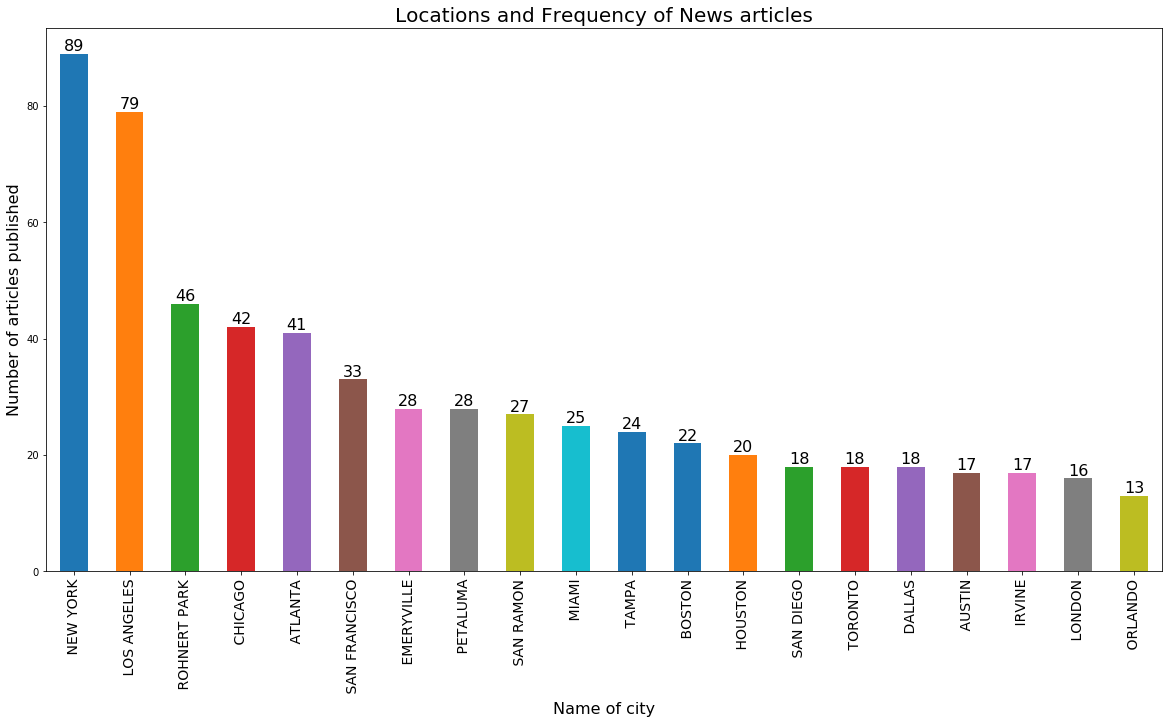
**Summary of output:**

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.



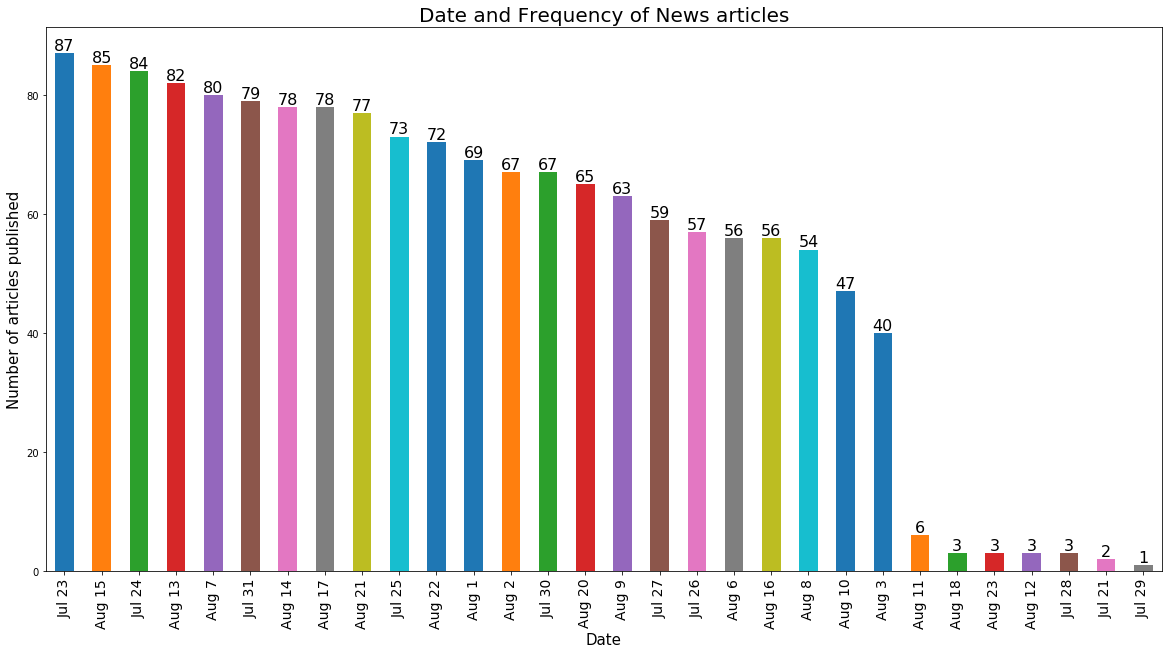
2. News articles were published from 567 locations, top 20 locations and the number of articles from the respective locations is plotted.

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



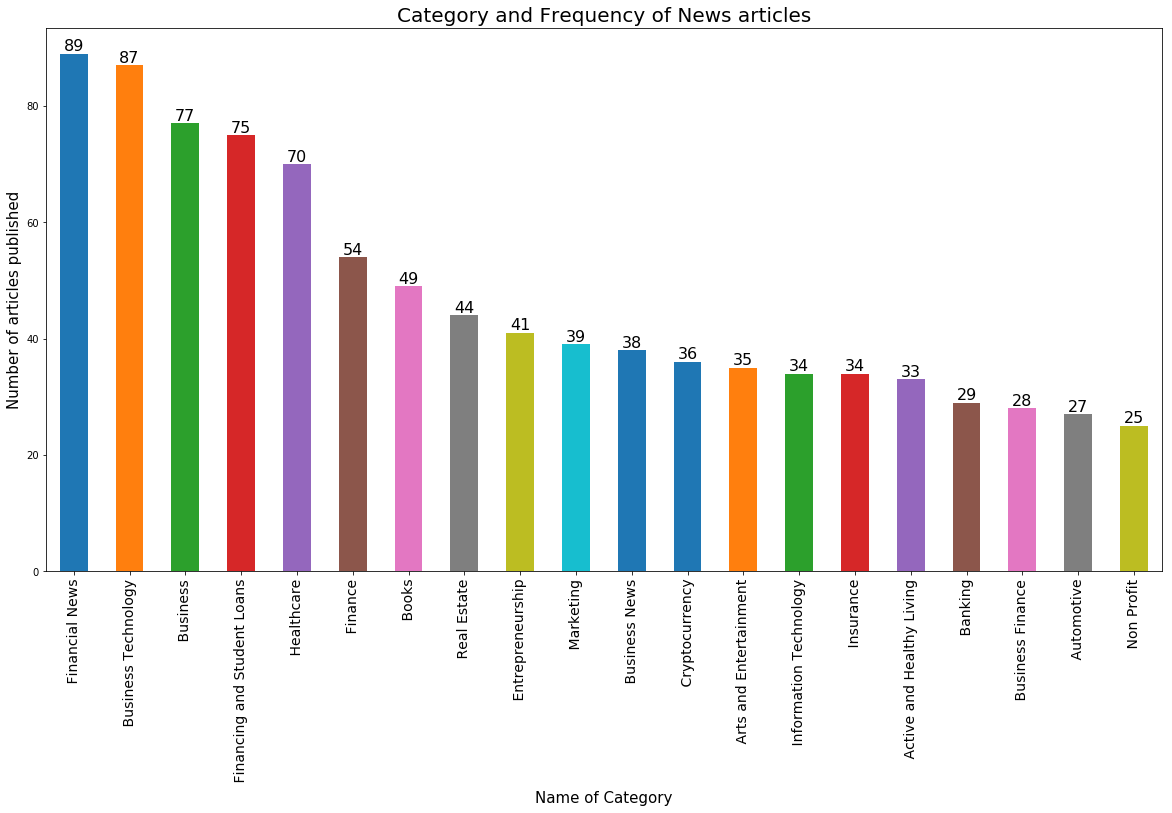
3. News articles from past 33 days (as of Aug 23, 2:50 AM) are considered under analysis, top 30 days and the number of articles published on those days is plotted.

3.1. Very few articles are being published on weekends.



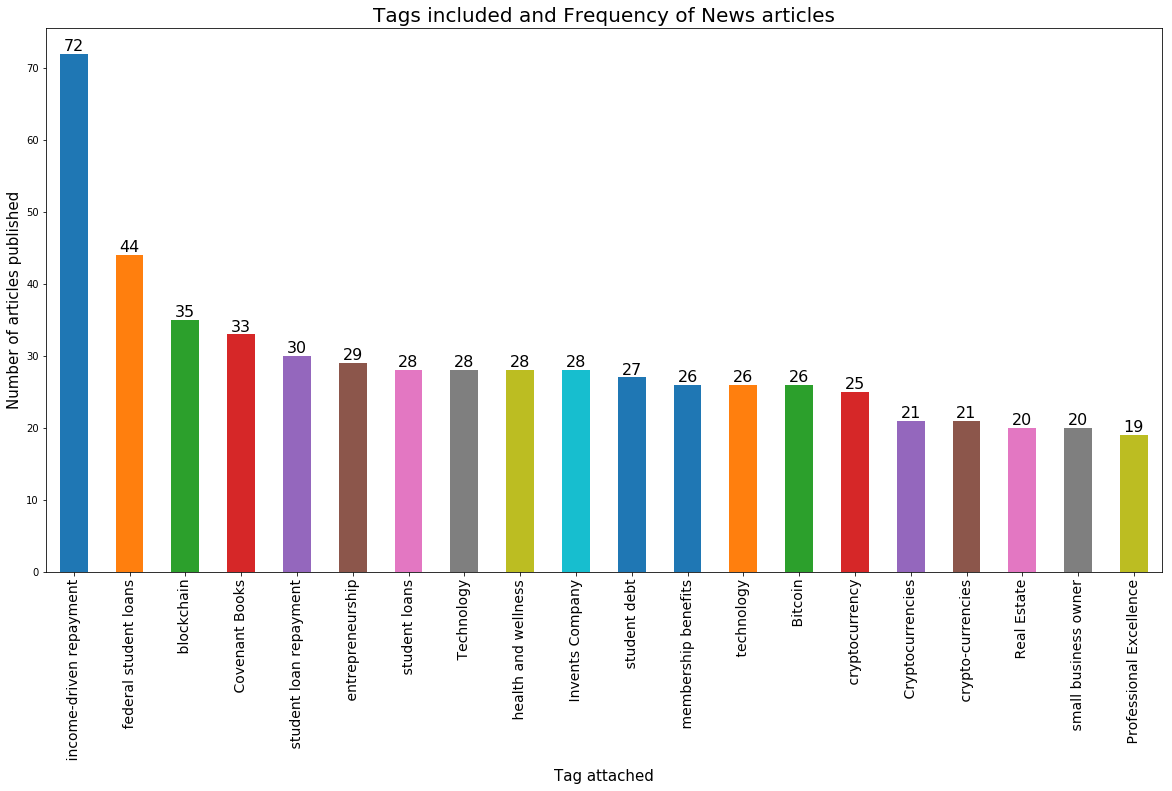
4. News articles were classified into 379 unique categories, top 20 categories and the number of articles published under them is plotted.

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.

**Possible code enhancements:**

There are many possibilities to enhance the submitted code. Based on business requirements, it can be modified to generate

1. Reports for a specific period of time.
2. Reports for a specific location
3. Report for a specific News tag coming from different locations and so on….

Code can be optimized to write a single plot function and display various plots (downside being all plots tend to have similar parameters and limiters)

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