**Data Analysis of Satellites**

Smit Jagdish Chheda, 18186319

Database and Analytics Programming, Group B

Master of Science in Data Analytics

**Abstract**

Importance of Satellite dependability was long recognized by technical literature, but, a statistical analysis is inadequate. The survival or non-existence of a satellite leads to inconsistency of literature. The commenced work will thus solve and fill the gap in technical literature by providing 1) Nonparametric statistical analysis of satellite data, 2) Insights of data analysed that can be used ahead while implementing in future. The data that is analysed in this study consists of 2062 satellites records which are revolving around earth either in LEO, GEO, MEO or in Elliptical region. The output presented here should prove convenient for space or satellite research analysis, redundant allocation and while growing dependability.

**Introduction**

Considering the dependability of the class of satellites has become not only prime but relatively unconsumed research area. This is because of the increasing awareness of satellites. There are only limited studies which are established on statistical analysis of satellites in orbits although responsible model for satellite analysis may exist.

The principal feature is a change in physical characteristic of the surface owing to the replacement of vegetation. The decrease in surface available changes the radiative fluxes. These physical processes would be tough to handle and hence satellite born instruments are analysed so that they can keep a track at the activities which are performed.

Analysis of satellites here includes what purpose are the satellites majorly used for? i.e. Communication, Earth Observation, Navigation, Technology Development, etc. and in which orbit does major chuck of satellites belong i.e. LEO, GEO, MEO, Elliptical. It also includes which is the launch vehicle highly used?

**Related work**

Multiple approaches were proposed for analysis of satellite data in which a data with infrared oriented satellite did not have direct contact with the ocean. Electro magnetic waves are detected by the sensors where they get limited. Data that is received from the satellites have a combination of signals from atmospheric and ocean. A second concept is required for the separation of these signals which will be demonstrated.

Due to this second concept, vast amount of data is produced, and computing analysis techniques are required so that the data can be manipulated. This makes the process expensive. [Paper 1]

There was another approach where data was specialized and, in an orbit, the specific satellites were considered. Here, the analysis of the dependability was on the type of satellite. But the disadvantage was the reduction in sample size which leads to results becoming uncertain while visualizing the data. The reliability value would also vary after the reduction in sample size. [Paper 3].

**Data Source**

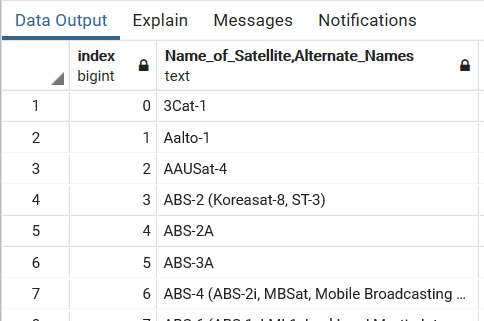
Data obtained in .csv format from the Union of Concerned Scientists (UCS) Satellite Database containing both numerical values as well as String.

Source: <https://www.ucsusa.org/resources/satellite-database>

**Methodology**

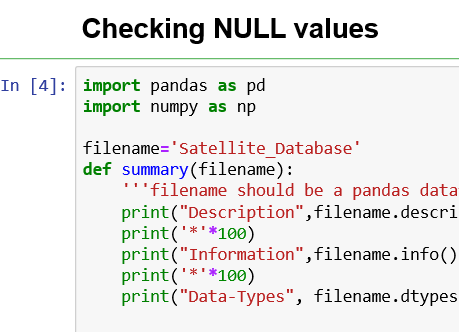
UCS provided with a dataset having 2062 rows and 65 columns consisting of numerical, string and alpha numeric data.

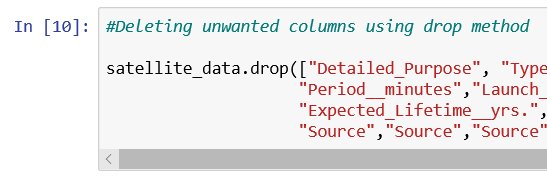
In postgres, a repository was created after which the extracted data from UCS were proficiently pushed and uploaded.



Considering efficiency of the data, there were many columns which weren't needed for our data analysis and could had been neglected or omitted.

In order to go on with the analysis, we first checked for Null and Duplicate column entries but did not find any.



This was followed by removing the columns which were not necessary in Data Visualization. 

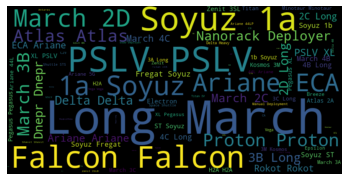
We were left with 12 columns while others were dropped.



The snippets below show the before and after of Data cleaning respectively. The dataset was then stored on Cassandra as Cassandra allows column storage and helps in managing tabular database easily. The data set was then pulled from Cassandra in the form of .csv file and data visualization were performed. All the actions taken were performed using Python.

**Results**

For visualization and plotting the graphs, three different libraries were used i.e. plotly, seaborn and matplotlib. Figure 1 shows Word Cloud where the column used was Launch vehicle to analyse which vehicle is majorly used. The insight that we can get from this graph is What launch vehicle can be used for a specific purpose that satellite is used for i.e. Military, Communication, Earth observation, etc.



**Figure 1**

We also plotted a pie chart (Figure 2) and bar chart (Figure 3) where we analysed which are the countries operating maximum number of satellites and a relation of purpose of the satellite and the orbit in which the satellite is present respectively.

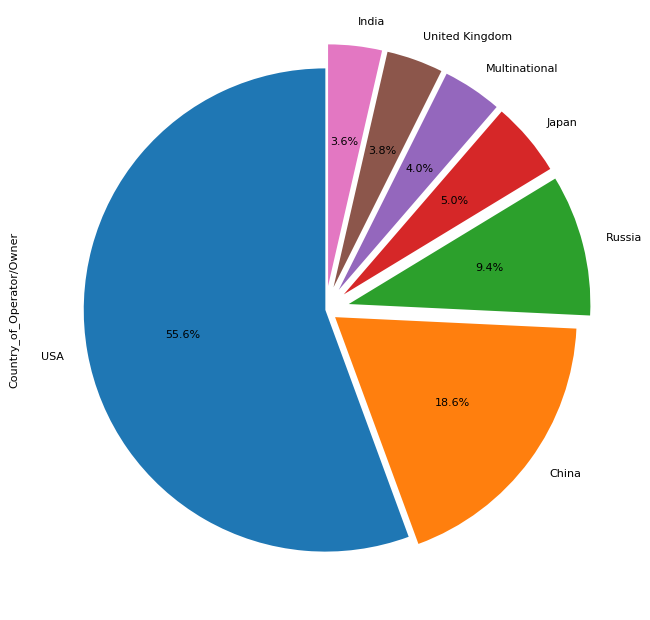


Figure 2

The relation of Purpose and orbit gives us an insight that a Satellite ‘X’ with purpose ‘Y’ should be placed in orbit ‘Z’.

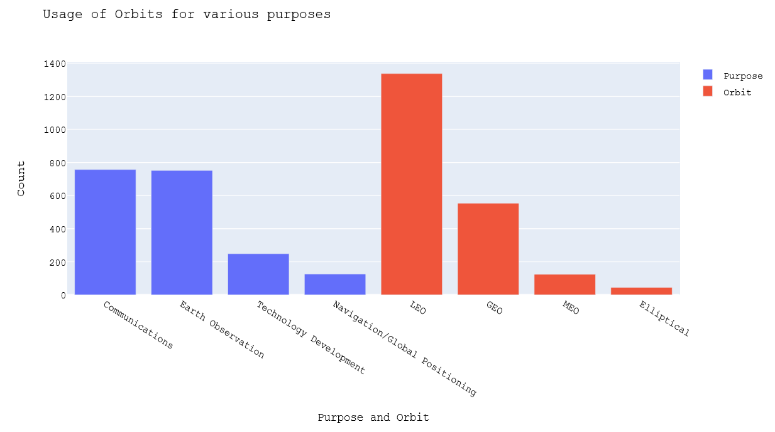


Figure 3

In figure 4, values of Perigee and Apogee were considered to plot univariate distribution of observations among these two values.

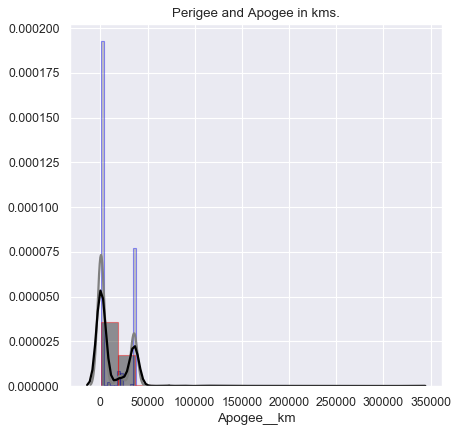


Figure 4

While in figure 5 displayed below, we have considered Country of contractor to display number of contracts that top 20 countries have received. This was done using Custom Lollipop graph which has a lollipop like structure on its line. Using this graph along with the Purpose that is displayed in figure 2, we get an insight that satellites with Purpose ‘P’ have contacted the contractor ‘C’.

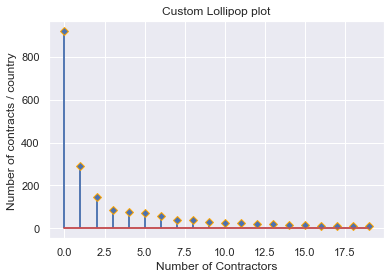


Figure 5

**Conclusion and Future Scope**

In the analysis of Satellites from the data provided by Union of Concerned Scientists, we were able to analyse data and get some insights from the data being 1) Which Launch vehicle to be used as per the purpose of satellite 2) Orbit on which the satellite should be present on basis of its use 3) Type of users using the satellite for a particular purpose.

However, this data being gathered from open source operational sources, could have been easily manipulated for analysis and research work and hence could not be completely trusted. But, the outcome from analysis could be taken into consideration to further analyse what sort of launch vehicle can be considered for a specific purpose and at which orbit it should be placed. Data mining and machine learning methods can also be used to predict what would be the total Mass of Satellite before it has been released depending on the launch vehicle and other contents