Project 1 Final Report

Proposal

Perform data analysis on a satellite database, for our client who wants to start a space program. My analysis will give us an insight into the history and trends of the type of satellites that orbit earth today. We will help to answer questions such as; How many satellites a country has and what are they used for? What are the characteristics that distinguish satellites from each other? What kind of satellites the client should launch to be with current trends?

Data Aquisition

To be able to answer these questions we will use the 2017 UCS satellite database. This dataset is acquired through the website of an American non-profit organisation known as ‘Union of Concerned Scientists’. It contains 1738 entries of various international satellites as well as 68 columns of information associated with each of these entries. The dataset is downloaded in Excel form.

- Satellite information.
    - Name of Satellite
    - Country/Org of UN Registry
    - Country of Operator/Owner
    - Operator/Owner
    - Users
    - Purpose
    - Class of Orbit
    - Longitude of GEO (degrees)
    - Perigee (km)
    - Apogee (km)
    - Eccentricity
    - Period (minutes)
    - Date of Launch

Note that the above list of features is not complete and only lists some of the more interesting and relevant features.

Data Cleaning

The Excel file is imported as a pandas data frame and is used in that form throughout the project.

Dropping features

- The dataset contains 20 unnamed columns with zero non-null values. All of these are dropped because they are of no use to us.
- The ‘Name’ column contained duplicate entries and so the extra copies were dropped.
- The following columns were dropped because they were simply not useful for this project and/or contained too many NA values:
    - 'DetailedPurpose', 'Comments', 'Unnamed: 27', 'Source Used for Orbital Data', 'Source', 'Source.1', 'Source.2', 'Source.3', 'Source.4', 'Source.5', 'Unnamed: 58', 'Unnamed: 59', 'Unnamed: 60', 'Unnamed: 64'

Renaming features

- The following column names are renamed to be more functional by shortening names and removing spaces between words.
    - 'Name of Satellite, Alternate Names' : 'Name'
    - 'Country/Org of UN Registry' : 'RegistryCountry'
    - 'Detailed Purpose' : 'DetailedPurpose'
    - 'Perigee (km)' : 'Perigee(km)'
    - 'Apogee (km)' : 'Apogee(km)'
    - 'Inclination (degrees)' : 'Inclination(degrees)'
    - 'Period (minutes)' : 'Period(minutes)'

- 'Launch Mass (kg.)' : 'LaunchMass(kg)'
- 'Dry Mass (kg.)' : 'DryMass(kg)'
- 'Power (watts)' : 'Power(W)'
- 'Date of Launch' : 'DateofLaunch'
- 'Country of Operator/Owner' : 'CountryofOwner'
- 'Class of Orbit' : 'ClassofOrbit'
- 'Type of Orbit' : 'TypeofOrbit'
- 'Expected Lifetime' : 'ExpectedLifetime'
- 'Country of Contractor' : 'CountryofContractor'
- 'Launch Site' : 'LaunchSite'
- 'Launch Vehicle' : 'LaunchVehicle'
- 'COSPAR Number' : 'COSPAR#'
- 'NORAD Number' : 'NORAD#'

- The ‘Users’ column contained similar entries such as ‘Civil/Governement’ and ‘Government/Civil’. Entries such as these are merged together and result in a ‘Users’ column with the following values and counts:

```
Commercial                      768
Government                      337
Military                        257
Civil                           128
Government/Commercial           114
Military/Commercial              69
Military/Government              31
Government/Civil                 25
Military/Civil                    2
Commercial/Government/Military    1
```

- Similarly, the ‘Purpose’ column contained similar features which were merged together Before merging:

```
Communications                                    725
Earth Observation                                 579
Technology Development                            179
Navigation/Global Positioning                      98
Space Science                                      65
Earth Science                                      22
Technology Demonstration                           13
Communications/Technology Development              11
Navigation/Regional Positioning                    10
Space Observation                                   9
Earth Observation/Technology Development            6
Communications/Maritime Tracking                    5
Earth Observation/Communications                    2
Space Science/Technology Development                1
Communications/Navigation                           1
Earth Science/Earth Observation                     1
Earth Science/Space Science                         1
Earth Observation/Space Science                     1
Earth Observation                                   1
Earth Observation/Communications/Space Science      1
Technology Development/Educational                  1
```

After merging similar values:

```
Communications              725
Earth Observation           580
Technology Development      179
Navigation                  108
Space Science                65
Multipurpose                 31
Earth Science                22
Technology Demonstration     13
Space Observation             9
```

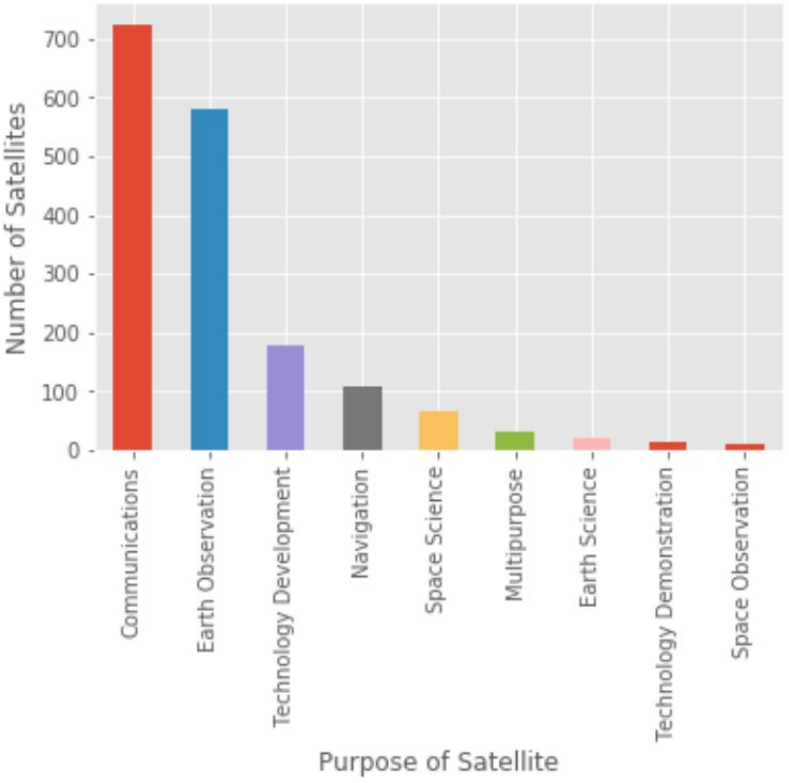
- The ‘ClassofOrbit’ column contained duplicate values because of unecessary spaces in some of the values. The duplicate values are renamed and result in the following values and counts.

```
LEO           1065
GEO            531
MEO             97
Elliptical      39
```

This cleaned dataframe is saved under the name ‘clean_data.pkl’

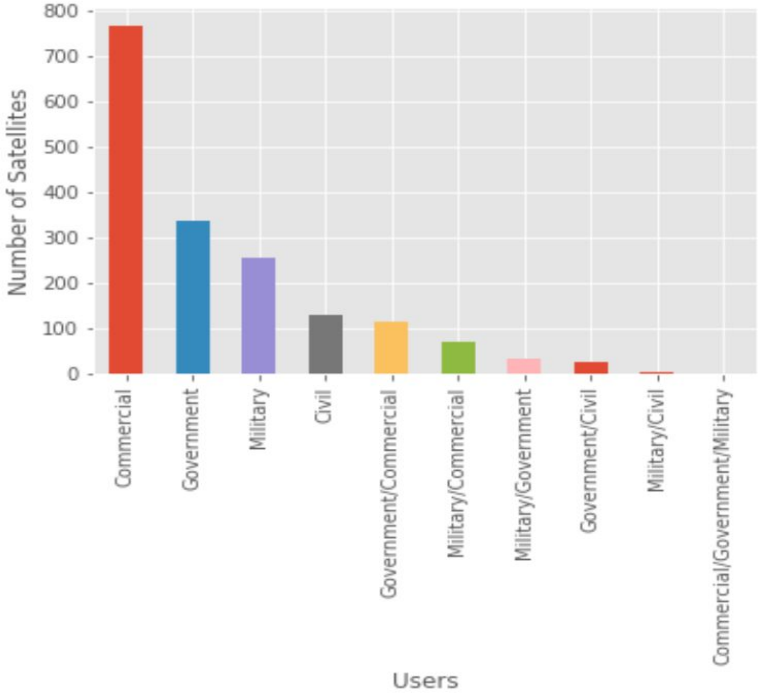
Exploratory Data Analysis

Number of satellites by purpose



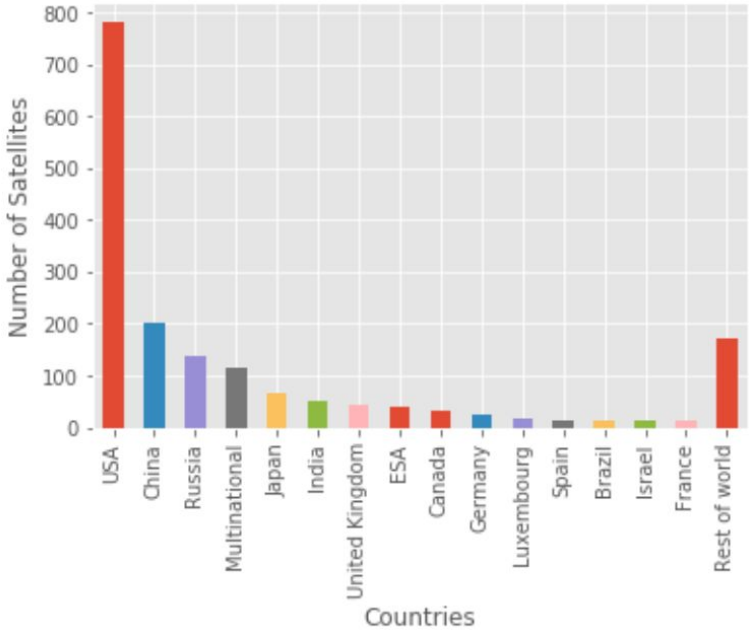
- Communicatons and earth observation satellites are the most common coming in at 725 and 580 satellites respectively.

Number of satellites by users



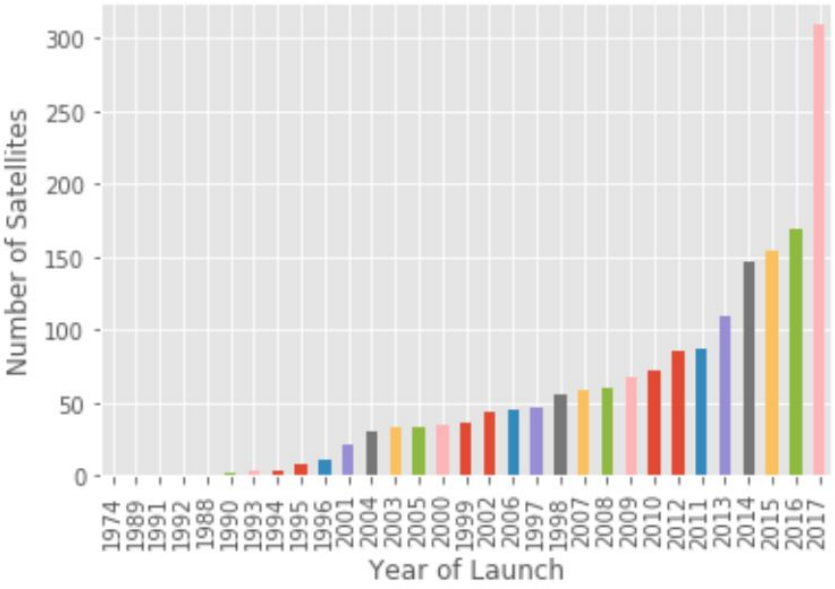
- Commercial and Government satellites are most common with 768 and 337 satellites repectively.
- It is also interesting to note that government and military users are also included along side the other users that appear on the graph

Number of satellites by country



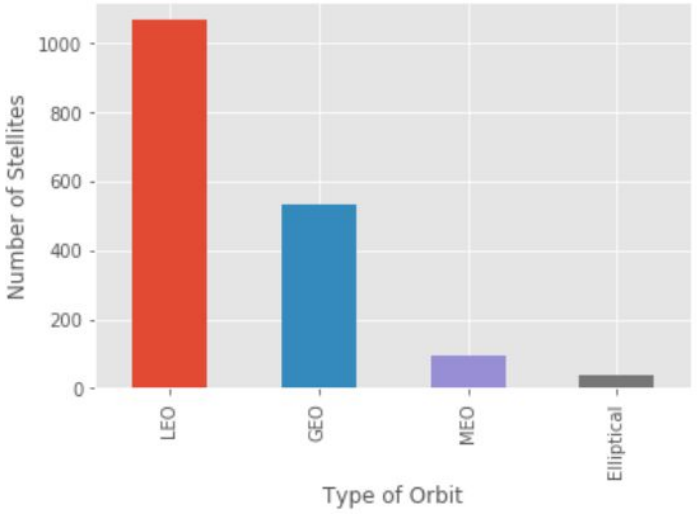
- For this visualization we grouped all countries with less than 12 satellites into a single country labeled ‘Rest of World.
- The USA has the most satellites at 780 followed by China and Russia at 203 and 139 respectively.
- In fourth place is the ‘MultinItional’ category and includes satellites that are co-owned by multiple countries

Number of satellites by year of launch



- As expected we see an overall increase in satellite launches from 1974 to 2017.
- Only one satellite was launched in 1974, 1988, 1989 and 1991
- The number of satellite launches increased tenfold in the 17 years between 1996 and 2013.
- And in more recent years the number of satellite launches almost doubled just between 2016 and 2017, from 169 to 309 launches respectively!

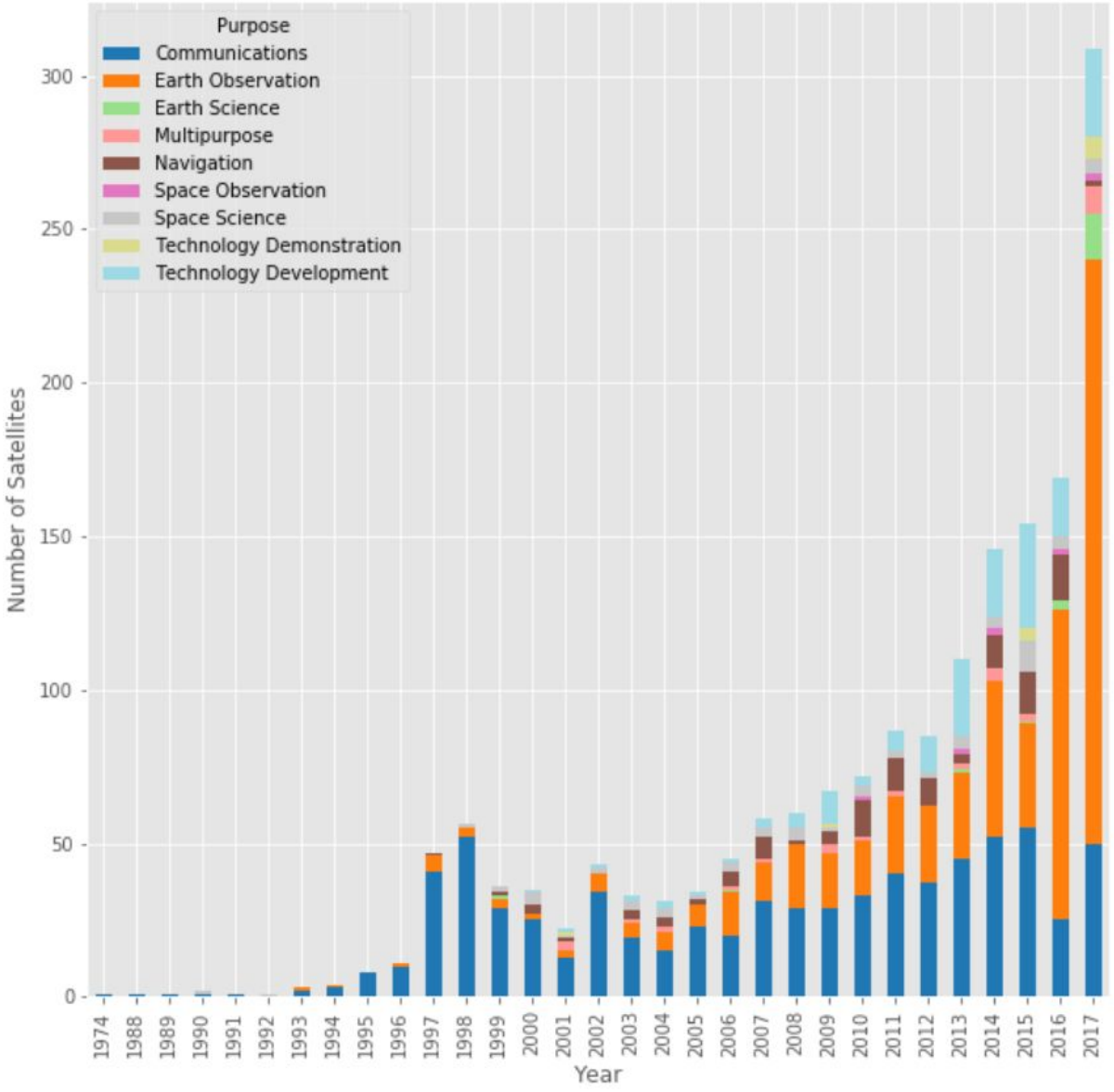
Number of satellites by type of orbit



- There are 4 categories of satellite orbits in this dataset. They are:
    - Low Earth orbit (LEO) - 1,065 satellites
    - Geostationary orbit (GEO) - 531 satellites
    - Medium Earth orbit (MEO) - 97 satellites
    - Elliptical orbit - 39 satellites

- The type of orbit plays an important role in what the orbit is going to be used for, and this is further studied in the later part of the EDA section.

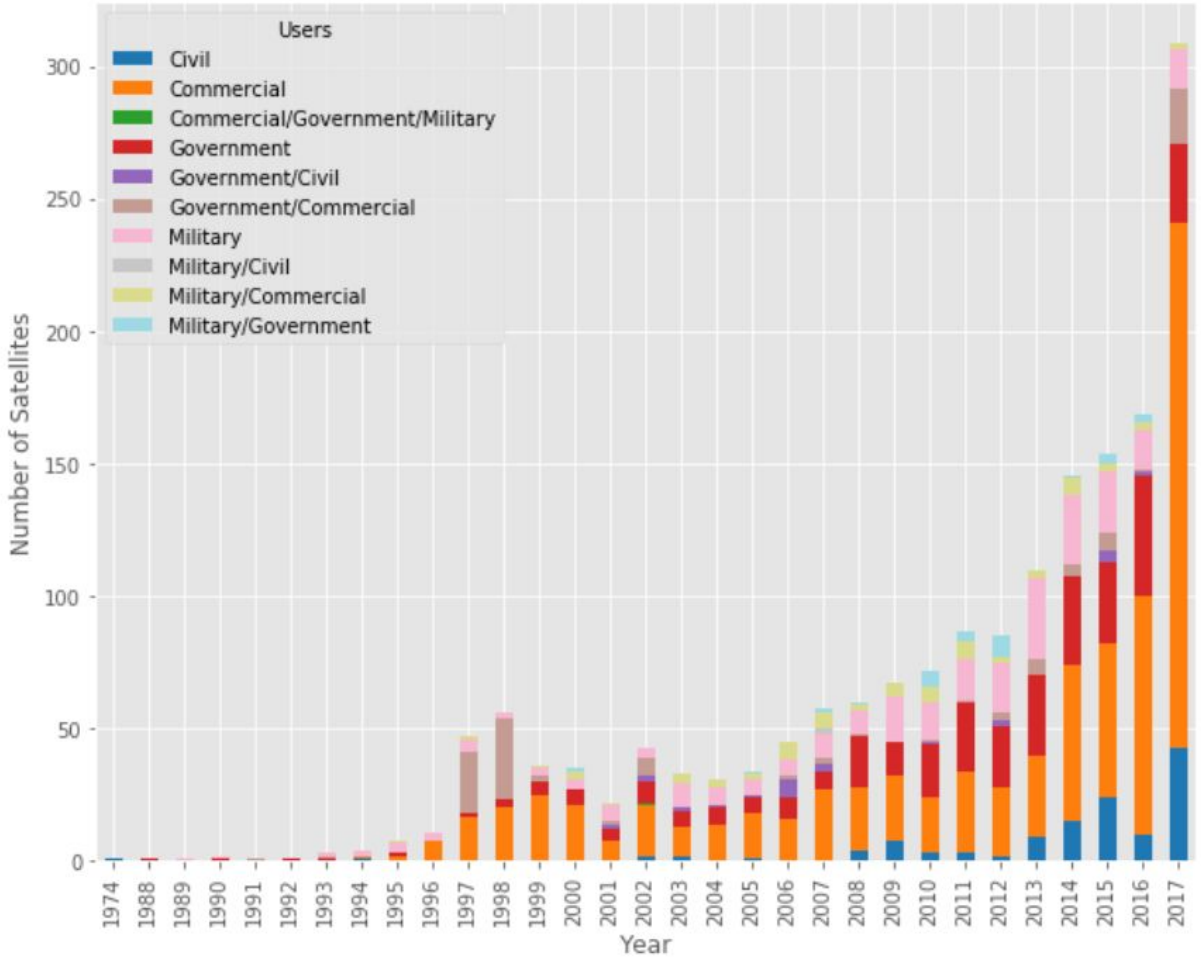
Distribution of the purpose of satellites by the year of launch



- This graph represents the change in the purpose of the satellites being launched from 1974 to 2017. Missing years represnt years when no satellites were launched.
- The most visible change can be seen in the shift from, communication satellites to Earth observation satellites over the years. Communicaton satellites dominated the purpose category until 2005 after which we can see a gradual increase in in Earth observation satellites with significant Earth observation satellite launches in 2014, 2016 and, 2017.

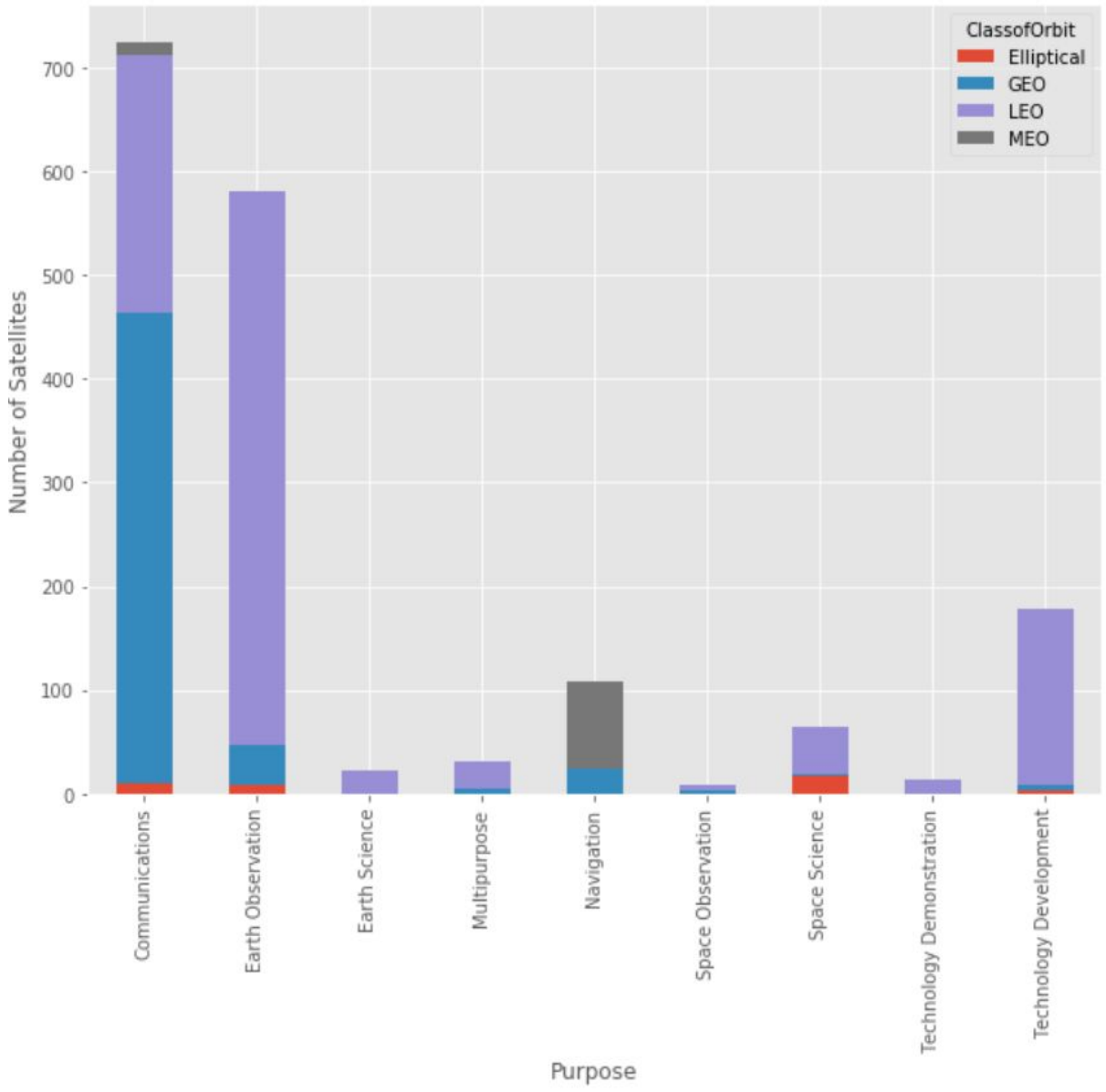
- There is also an increase in Earth science satellites in 2017. And interestingly Navigation satellites occupied a decent share of the purpose category since 2009, but seem to have significantly decreased in 2017 especially considering the overall increase in satellite launches.
- Technology development satellite launch numbers seem to have stayed relatively constant since 2013.

Distribution of the users of satellites by the year of launch



- The main user types here are ‘commercial’, ‘government’ and ‘military’. Wher the commercial satellites are the largest share from 2014 to 2017. There is also a steady increase in commercial satellite with the general increse in satellite launches.
- It is important to note that apart from the largest 3 categroies and the civil category all other users are just combinatioons of the latter and were kept that way as it was the best possible way to justly represent the data.

Distribution of the class of orbit of the satellites for each purpose.



- In this visualization we can see that most communication satellites have either geostationary orbits or low earth orbits. But can sometimes also have elliptical or medium earth orbits. Communication satellites are also the only satellite type that can have any of the 4 types of orbits.
- Navigation satellites are the only satellite type that never have low earth orbits. Navigation satellites have mostly medium earth orbits and can sometimes have geostationary orbits.

Inferential Statistics

We performed **chi 2 contingency** test using the statistical function from scipy.

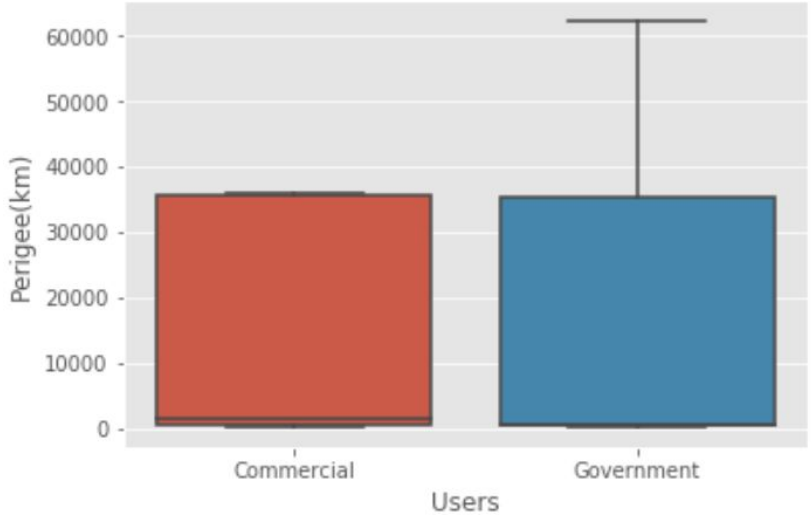
1.We use the test to find a dependence between the purpose of a stellite and it's type of orbit. We get a p-value of 0.0 for this test which tells us that null-hypothesis, that there is no dependence between the two variables, holds true.

2.We use the test to find a dependence between the purpose of a stellite and it's users. We get a p-value of 0.0 for this test which tells us that null-hypothesis ,that there is no dependence between the two variables, holds true.

3.We use the test to find dependence between the country of the satellite and it's type of users. We get a p-value of 1.229 for this test which tells us that null-hypothesis, that there is no dependence between the two variables, is false. There is a dependence between the type of Users a satellite has and what country it is owned by.

We performed **T-tests** to see if different variables had identical average values.

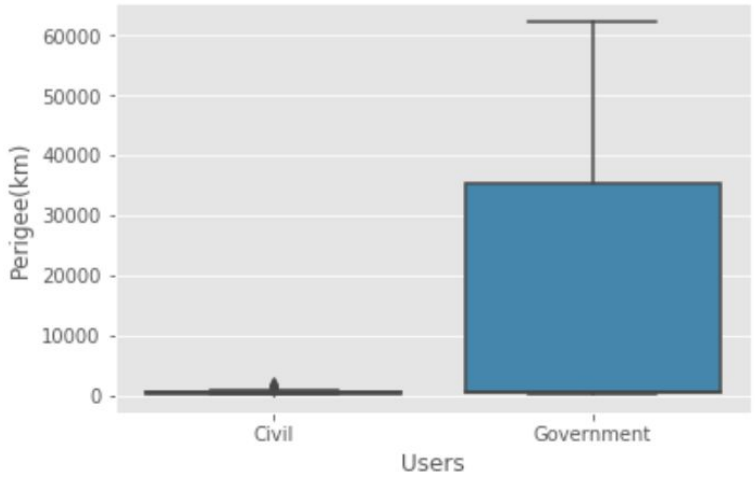
1.Government Perigee and Users Perigee



p-value = 4.8E-7

Since the p-value < 0.05 we reject the null-hypothesis that commercial and government perigee of satellites have identical averages.

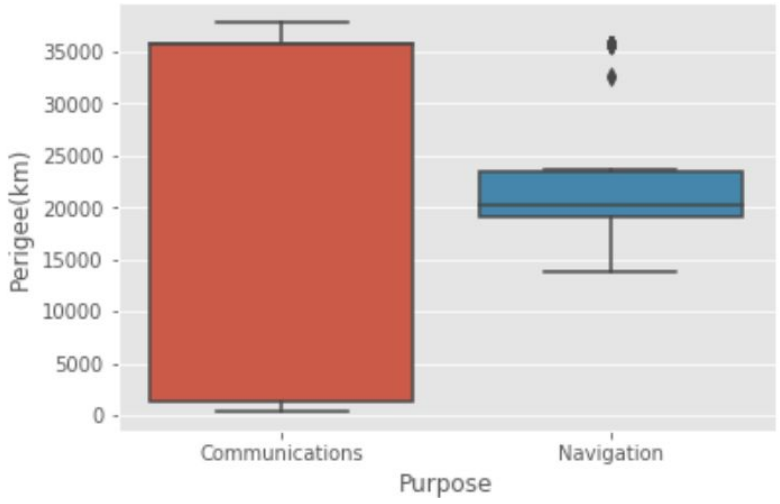
2.Civil perigee and government perigee



p-value = 1.189E-11

Since the p-value < 0.05 we reject the null-hypothesis that civil and government perigee of satellites have identical averages.

3.Communications and navigation perigee



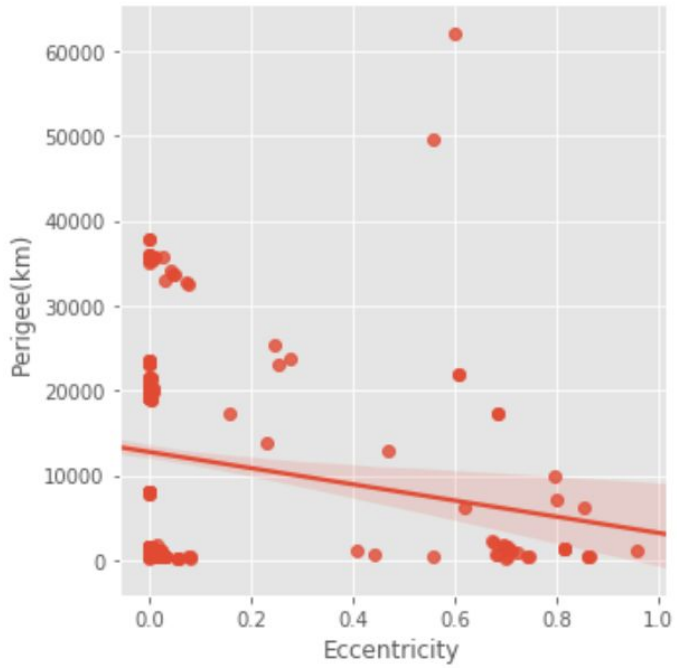
p-value = 0.56

Since the p-value > 0.05 we accept the null-hypothesis that communication and nvigation perigee of satellites have identical averages.

Pearson r correlation

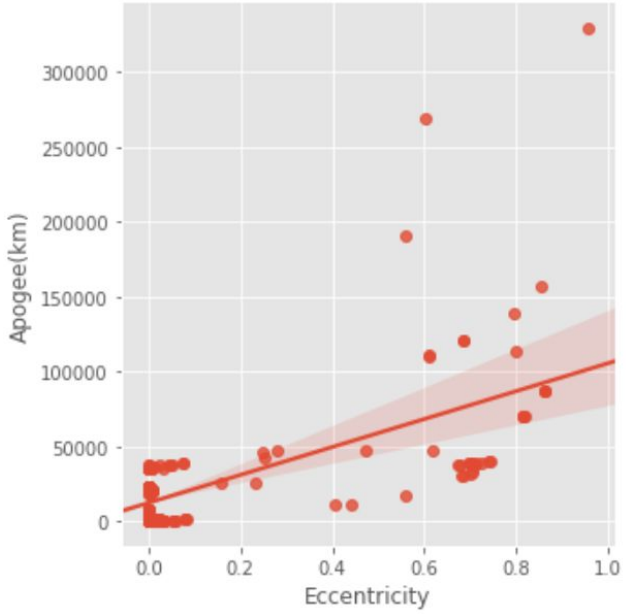
The Pearson correlation coefficient measures the linear relationship between two datasets.

1. Perigee vs. eccentricity



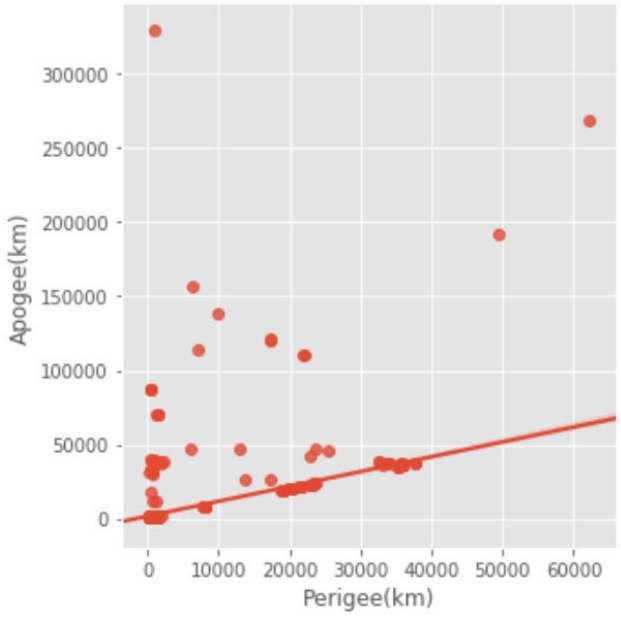
We get a correlation value of -0.06. This means that there is a weak inverse relatinship between perigee and eccentricity and it is statistically significant since it has a p-value of 0.0094 which is less than 0.05.

2. Apogee vs. eccentricity



We get a correlation value of 0.46. This means that there is a positive correlation between apogee and eccentricity and it is statistically significant since it has a p-value of 2.45E-93 which is less than 0.05.

3. Apogee vs. perigee



We get a correlation value of 0.76. This means that there is a strong positive correlation between apogee and eccentricity and it is statistically significant since it has a p-value of 0.0.

Machine Learning

In this part of the project we use machine learning techniques to be able to predict the purpose of the satellite using all the other features in the dataset.

At first we only use the numerical features in the prdiction. So we are only left with the following features; 'Purpose', 'Longitude of GEO (degrees)', 'Perigee(km)', 'Apogee(km)', 'Eccentricity', 'Inclination(degrees)', 'Period(minutes)' and, 'LaunchMass(kg)'. We do this so that we don't have to use dummy variables to use logistic regression or the KNN classifier.

We however, need to use a label encoder on the purpose column to convert all the different purposes in the column to integer values to be able to use it in either of our models. To do this we use the 'Label Encoder' module from the scikit learn library.

With the encoded purpose column we do a train test split on the data and fit it on a KNN model with 3 neighbors and are get an accuracy score of 81.4%. We then fit it on a logistic regression model and get an accurcay score of 62.5%.

We then use both the KNN and logistic regression models again but this time we include features that have non numerical entries. These added features are; 'CountryofOwner', 'Users', 'Purpose', 'ClassofOrbit'. To be able to use these features we use the pandas function ‘get_dummies’ to make dummy variables with values of 0 or 1. These dummy variable lists are added as columns to the data frame.

Once we fit this data to the two models we get accuracy scores of 53.2% and 70.8% for the logistic regression and KNN classifier respectively.

Confusion Matrix

```
[[190  24   3   1   1   0   1   0  10]
 [ 29 131   0   0   1   0   2   0   5]
 [  1   1   0   0   0   0   0   0   3]
 [  5   2   0   1   0   0   0   0   1]
 [  6   0   0   0  21   0   0   0   0]
 [  2   0   0   0   0   0   0   0   0]
 [  8   5   1   0   0   0   3   0   3]
 [  2   2   0   0   0   0   0   0   1]
 [ 16  13   0   1   0   0   3   0  16]]
```

Classification Report

	precision	recall	f1-score	support
0	0.73	0.83	0.78	230
1	0.74	0.78	0.76	168
2	0.00	0.00	0.00	5
3	0.33	0.11	0.17	9
4	0.91	0.78	0.84	27
5	0.00	0.00	0.00	2
6	0.33	0.15	0.21	20
7	0.00	0.00	0.00	5
8	0.41	0.33	0.36	49
avg / total	0.67	0.70	0.68	515

We then made the decision to use the KNN classifier and the dataset with only numerical value for the remainder of the machinle learning analysis because that combination gves the best results. The following results are using the KNN model and dataset with only numerical values.

Confusion Matrix

```
[[198   5   0   2   0   0   0   0   2]
 [ 19 138   0   0   0   0   2   1   1]
 [  0   2   3   0   0   0   0   0   1]
 [  4   3   0   0   0   0   0   0   2]
 [  6   0   0   0  26   0   0   0   0]
 [  0   1   0   0   0   0   0   0   0]
 [  2   9   0   0   0   0   6   0   0]
 [  1   0   1   0   0   0   0   1   1]
 [  8  15   1   0   0   0   1   0  21]]
```

Classification Report

	precision	recall	f1-score	support
0	0.83	0.96	0.89	207
1	0.80	0.86	0.83	161
2	0.60	0.50	0.55	6
3	0.00	0.00	0.00	9
4	1.00	0.81	0.90	32
5	0.00	0.00	0.00	1
6	0.67	0.35	0.46	17
7	0.50	0.25	0.33	4
8	0.75	0.46	0.57	46
avg / total	0.80	0.81	0.80	483

Feature Selection

From the feature selection we conlcuded that the ‘Apogee(km)’ was the most innfluetial feature with a score of 0.68 followed by ‘Period(minutes)’ and Inclination(degrees) with scores of 0.62 and 0.60 respectively.

Grid Search CV

The grid search gave us an interesting results. The best parameter from the grid search was n_neighbors = 43 and the best score was 0.68. This is strange becasue by just setting the n_neighbors parameter in the KNN model we got an ccuracy score of 0.81.

Conclusion

We’ve seen that our machine learning model at best can predict the purpose of a satellite using numerical features with 81% accuracy at best. We learnt from feature selection that apogee, period and inclination ar the most effective features in the prediction.

Through data visualizatin we learned that commercial satellites used for earth observation are the most common type of satellites today. And maybe the client should launch similar satellites to be with current trends. We also learnt that communications and earth observation satellites have mostly LEO and GEO orbit types but if the client is interested in launching satellits for navigation purpose the MEO orbit is the most common.

Future work in this project includes improving the machine learning model by using only the top scoring features from the feature selection in the prediction model. And performing more sattistical tests between different features of the satellite dat to find more interesting and useful insights for the client.