Winning Space Race with Data Science

Deivith Enrique Amaya Lopez

22 february 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The methodology of this project focuses on the analysis of the characteristics of the space x company flights. For this purpose, information collected from the space x API and information in Wikipedia is used. Subsequently, a cleaning is made, an exploratory analysis and culminates with the creation of models to make the prediction of the success of the flights.
- It was possible to define a series of characteristics that could increase the possibilities of success in each mission. In addition, we have a model that will help to make decisions about the different scenarios that arise in each mission.

Introduction

- The objective of this project is to analyze the information on the characteristics and conditions present in the aerospace launches and landings of the SPACE X company. The objective of this project is to analyze the variables that have an effect on the success or failure of the missions in order to increase the possibilities of replicating these scenarios and achieve successful missions. For this purpose, the information obtained from the SPACE X API and from Wikipedia pages with relevant information was used. In addition, taking into account that the flight conditions, such as the orbit to be accessed or the cargo to be transported, are not entirely controllable, it is necessary to train a machine learning model to help us make inferences about the possible outcomes of the different scenarios presented to us.
- Is there a relationship between the launch site, the orbit to be accessed and the mission outcome?
- What are the rocket characteristics to increase the probability of success in the different scenarios presented for the missions?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collection consists of 2 parts: 1. use the SPACE X API to get the information from the releases and use id to relate this information to other tables in other endpoints. 2. use web scraping to collect the data from the Wikipea page.
- Perform data wrangling
  - Clean the data by replacing null values of variables that do not "Landing_Outcome" as the mean of the variable. For the null values of the result variable, Landing_Outcome, they are treated as failed results.

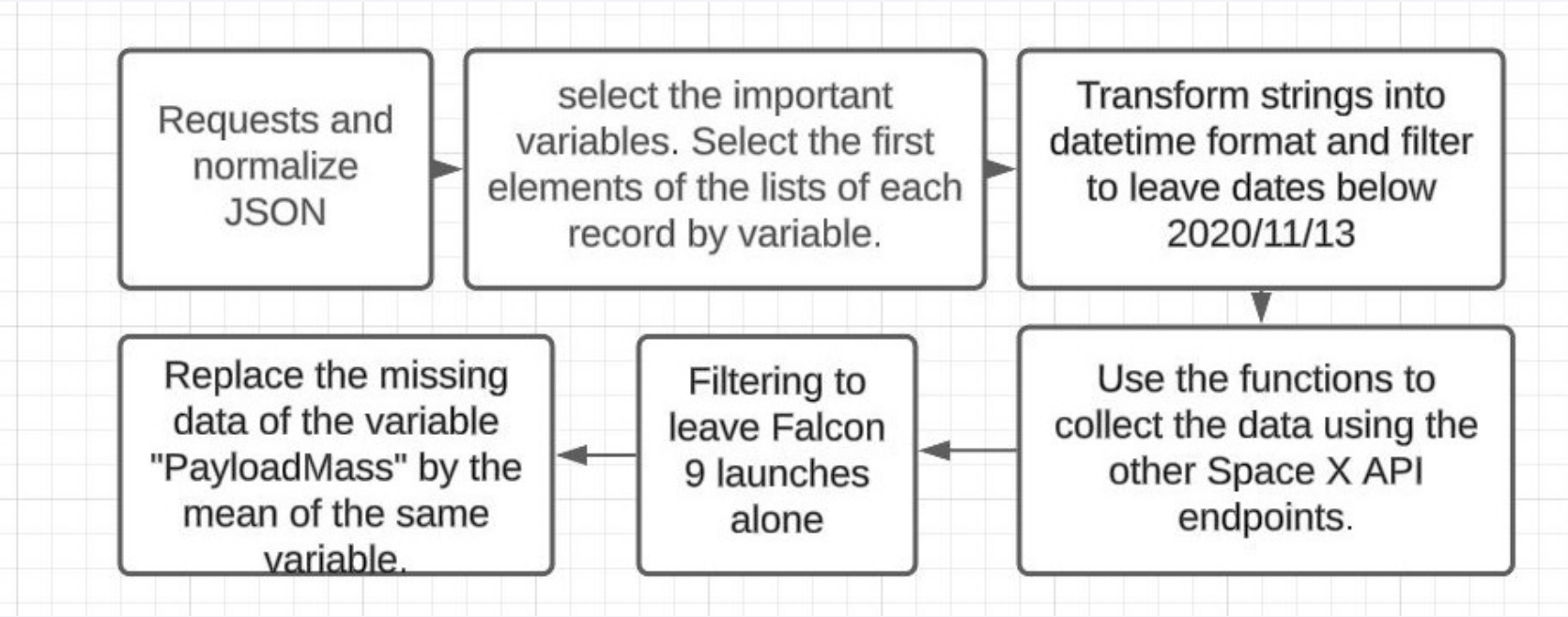
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform an exploratory analysis to determine the correlation of the independent variables with the dependent variable "class", which is the processed Landing_Outcome variable. Also review the collinearity between the independent variables to avoid overfitting.
  - Persist tables on a local SQL server. The database manager used to store the information is MYSQL SERVER.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Create dashboards to be able to present information and hypotheses interactively.
- Perform predictive analysis using classification models
  - Perform 4 different scoring models, evaluate their performance and choose the model with the score.

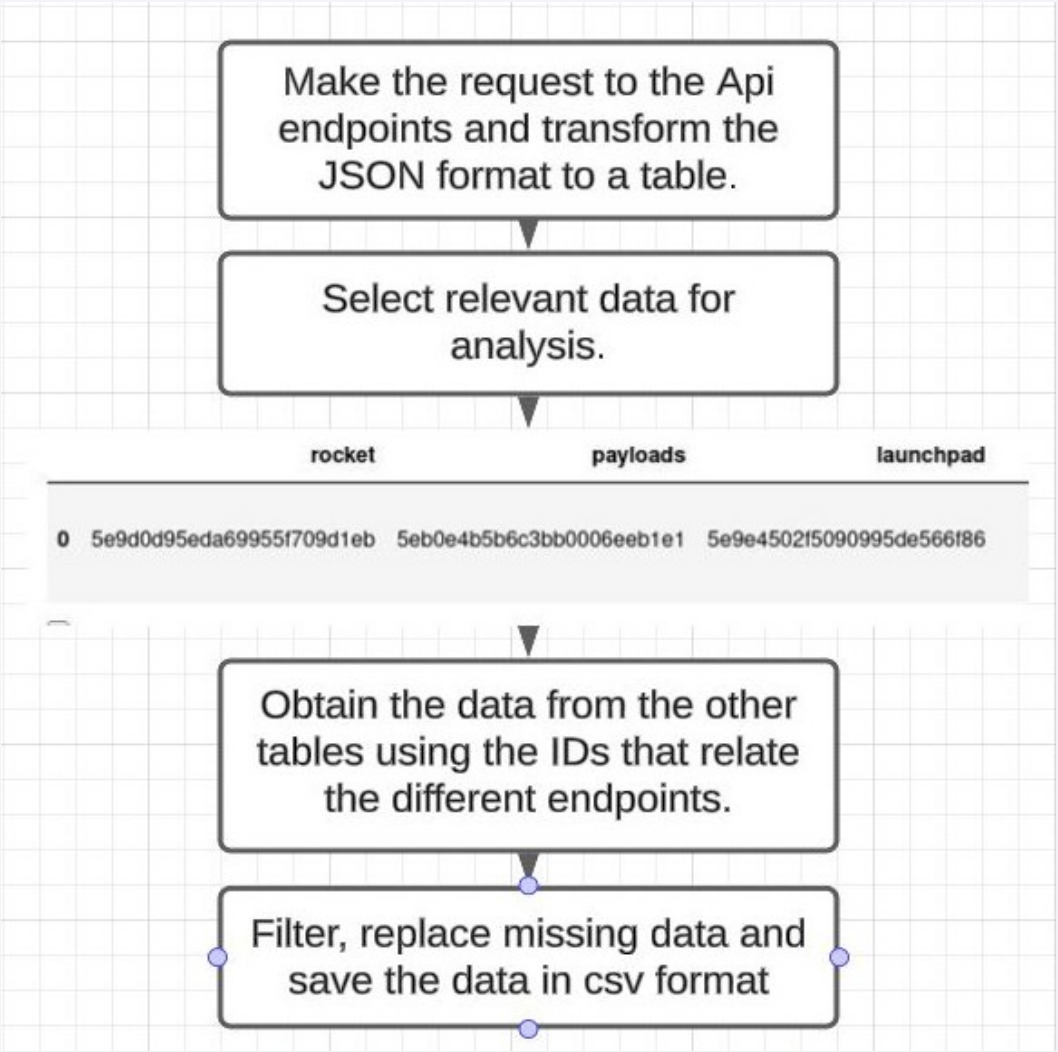
Data Collection

Collect data with the Space X API. This has 3 endpoints. Primarily, they are obtained from the Launches endpoint. This information has variables that relate the missions to the other 2 endpoints.



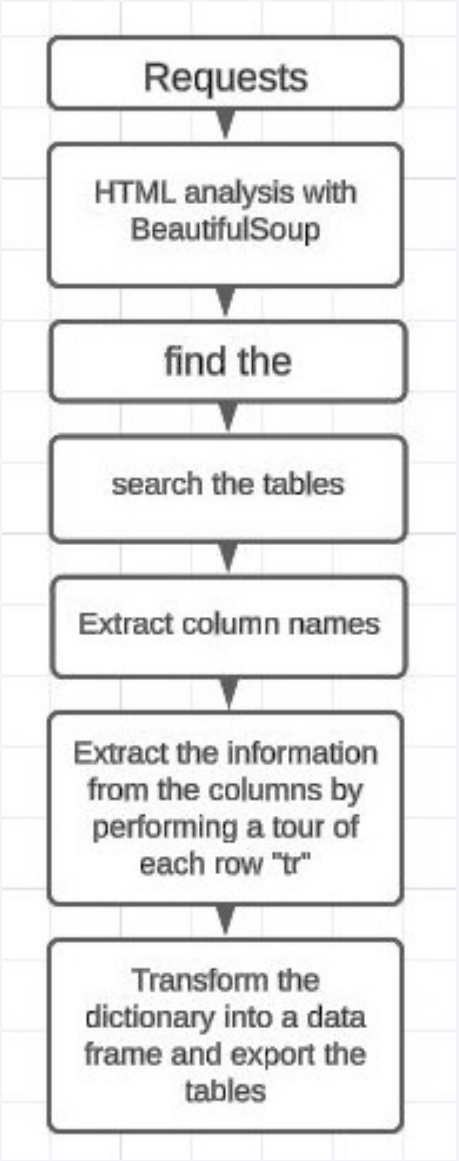
Data Collection – SpaceX API

- By means of a request the data is obtained from the Space X API, as the data is in JSON format it must be transformed to a table.
- Relevant data is separated from data that has no relevant information. In addition, the date data is transformed into a datetime.
- Finally, the information for each launch is obtained by relating the data to the other endpoints.
- https://github.com/deivithamaya/space_Y/blob/main/jupyter_notebooks/gettingData.ipynb



Data Collection - Scraping

- The web scraping technique is used. First the request to the web page must be made.
- Then the html format is parsed with the BeautifulSoup object. Then the tables and variable names are extracted.
- The information is extracted from the tables by traversing the records and from each record, obtaining the information of each variable. Each variable is cleaned of noise and formatted to its respective data type.
- https://github.com/ deivithamaya/space_X/blob/



Data Wrangling

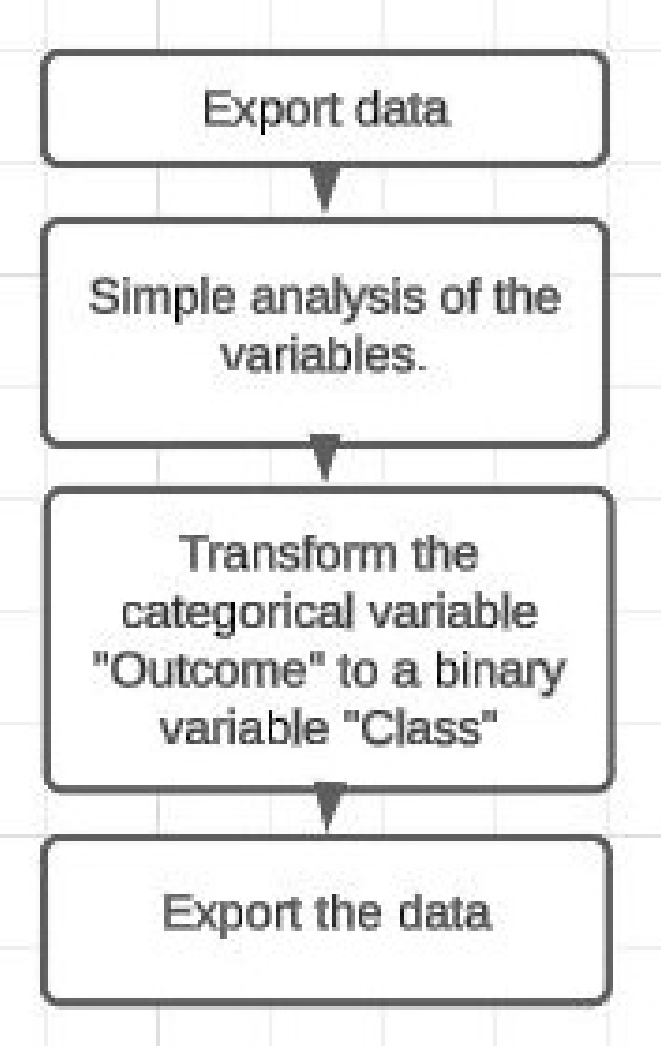
Because one of the objectives is to create a classification model, a target variable must be created. In this case, the "Outcome" variable is processed so that the result will be a binary variable of 1 for a successful mission and 0 for a failed mission.

The exported data set is loaded.

A simple analysis is done to observe the distribution of the data.

The categorical variable is transformed to a binary class 1 for any type of successful mission and 0 for any failure.

https://github.com/deivithamaya/space_Y/blob/main/



EDA with Data Visualization

A scatter plot is made that relates the variables of flight number, load and flight result, in order to observe if there could be a linear relationship between these variables.

Show the distribution of the number of flights per launch station, also relating these 2 variables to the outcome of the mission.

Make a scatter plot to observe the relationship between launch site, payload weight and mission outcome.

Make a histogram of mission results by launch site.

The last 2 graphs represent the relationship between the weight of the cargo, the orbit to access and the result of the mission. As well as the increasing rate of successful missions as time goes by, represented by a line graph of year vs success rate.

https://github.com/deivithamaya/space_Y/blob/main/jupyter_notebooks/EdaVisualization.ipynb

EDA with SQL

Importing mysql-connector and sqlalchemy
To be able to persist in a database from the dataframe you must have a connection created with compute_engine

- Create engine with compute_engine

- Use the to_sql method of the dataframe.

Make a new table removing the null values.
Perform the queries.
https://github.com/deivithamaya/space_Y/blob/main/jupyter_notebooks/EdaWithSql.ipynb

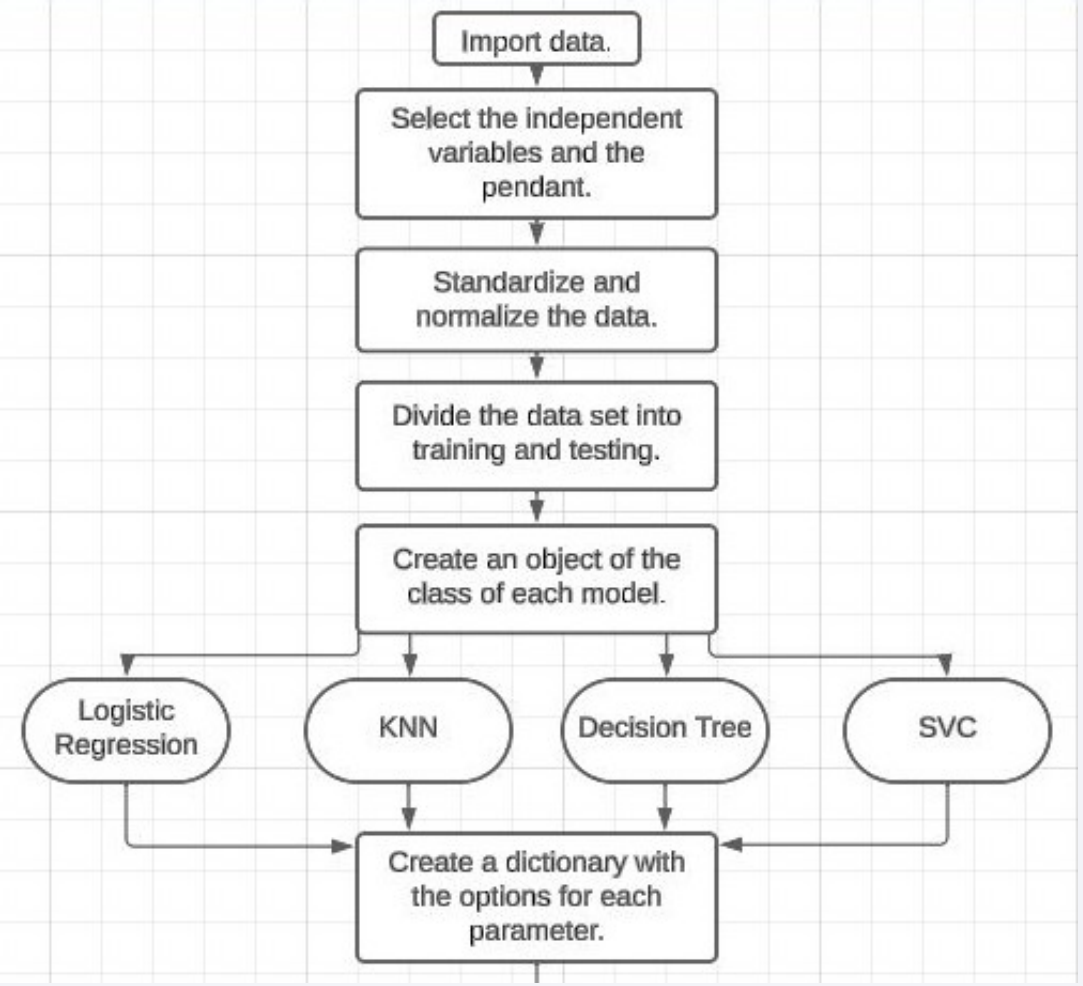
Build a Dashboard with Plotly Dash

Two types of graphs are made relating the launch site to:

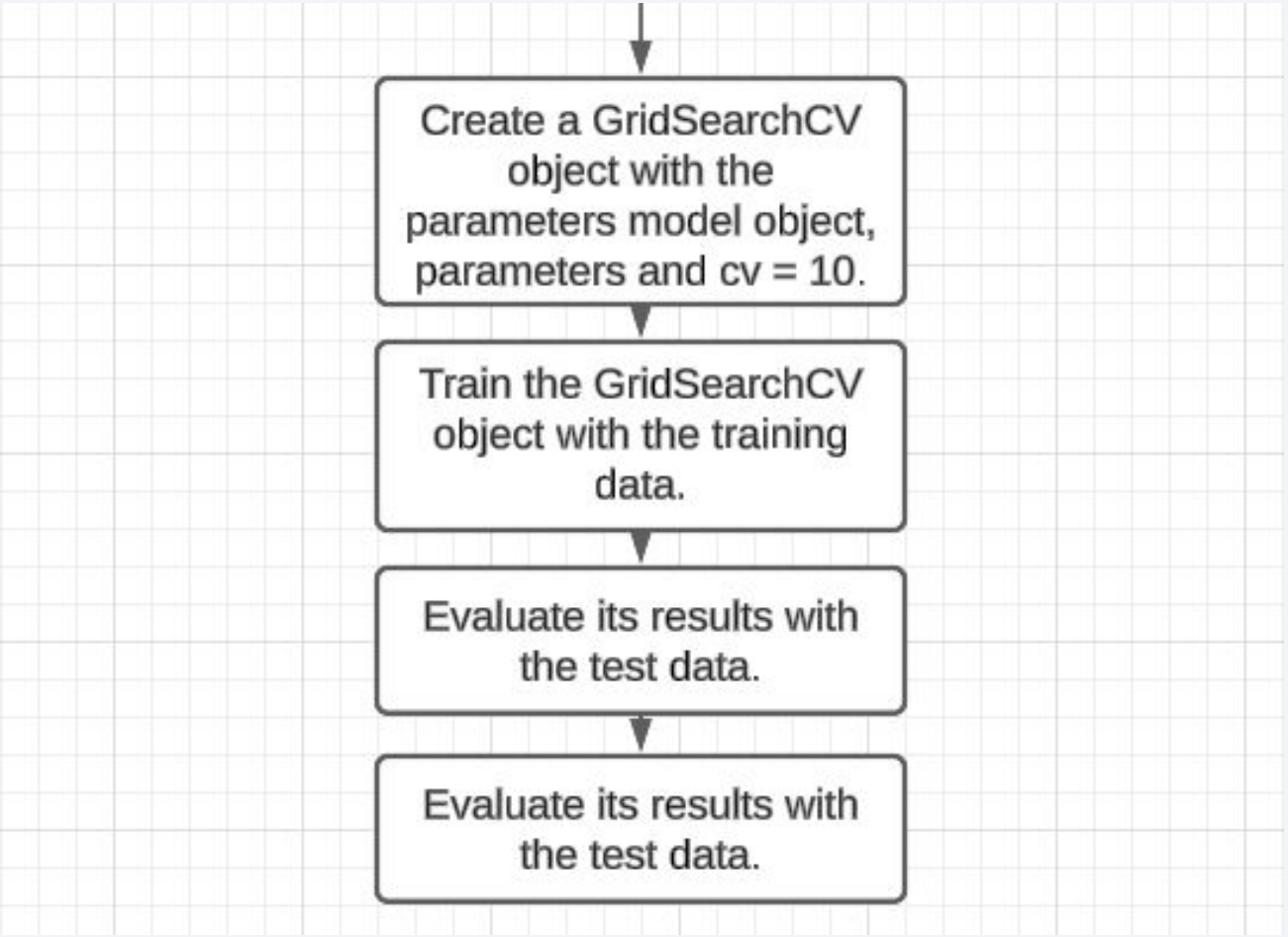
- The success rate of the missions.
- The relationship of payload weight, booster version and class.

These graphs allow to easily present the hypothesis and the results expected to be obtained when selecting the characteristics and launch sites.
https://github.com/deivithamaya/space_Y/blob/main/jupyter_notebooks/spacex_dash_app.py

Predictive Analysis (Classification)



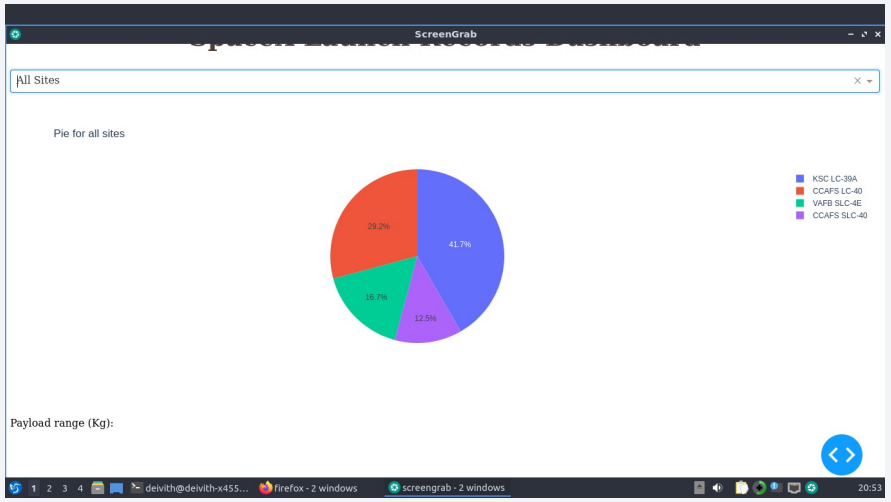
Predictive Analysis (Classification)



Results

Using dash, 2 interactive graphs were created to represent the information and obtain the characteristics of the missions that will increase the probability of success in the mission.

Features	Value
LaunchSite	KSC LC 39A
Outcome	ASDS
PayloadMass	(10000, 15000)
Serial	(B1051, B1058, B1060)
Orbit	VLEO



Results

Of the 4 models trained, 2 were obtained with the same accuracy: decision tree and support vector machine. However, when training the decision tree, the training results are not constant, but this model has an accuracy of 100% with the data having the selected features and the support vector machine only has an accuracy of 50%.Then, based on the results obtained with the data containing the selected features, the

```
In [75]: yy.values
Out[75]: array([1, 1, 1, 1])

In [76]: yhat = tree_cv.predict(xx)
         yhat
Out[76]: array([1, 1, 1, 1])

In [77]: yhat = svm_cv.predict(xx)
         yhat
Out[77]: array([0, 1, 1, 0])
```

```
In [22]: models
Out[22]: logistic regression    0.666667
         SVM                    0.833333
         Tree classifier        0.833333
         KNN                    0.777778
         dtype: float64
```

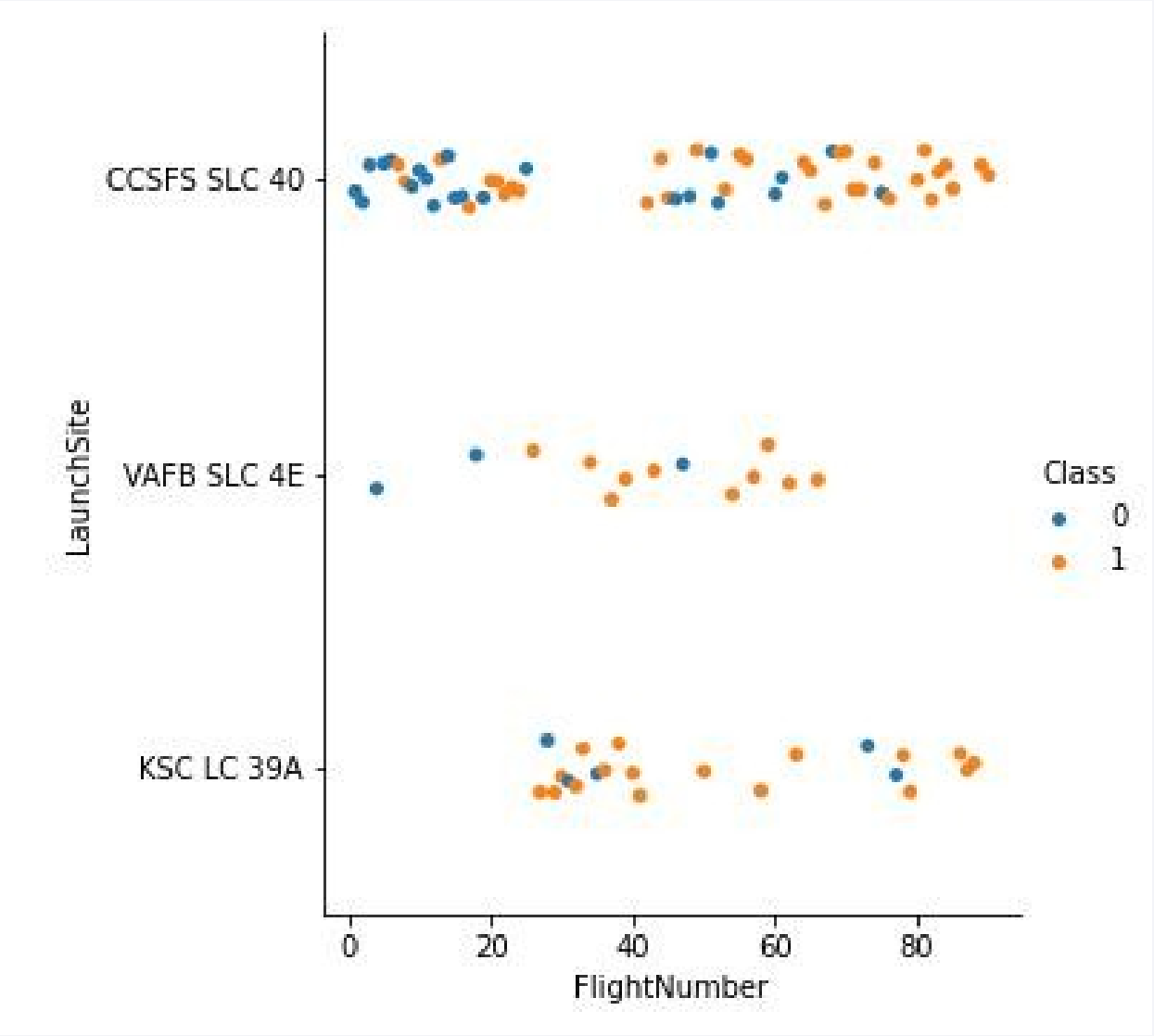
```
In [30]: models
Out[30]: logistic regression    0.666667
         SVM                    0.833333
         Tree classifier        0.722222
         KNN                    0.777778
         dtype: float64
```

Section 2

Insights drawn from EDA

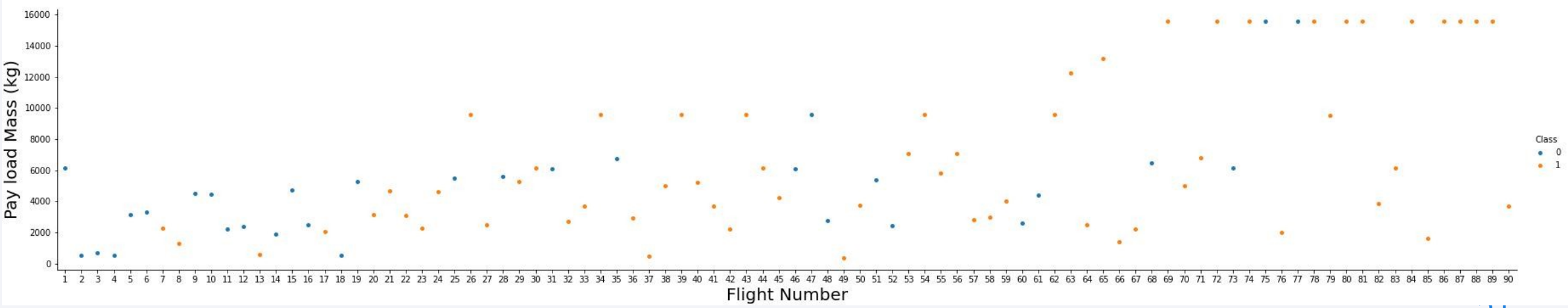
Flight Number vs. Launch Site

It represents the distribution of flights by each launch site, along with the mission result. From this graph it can be seen that launches from VAFB SLC 4E have been discontinued.



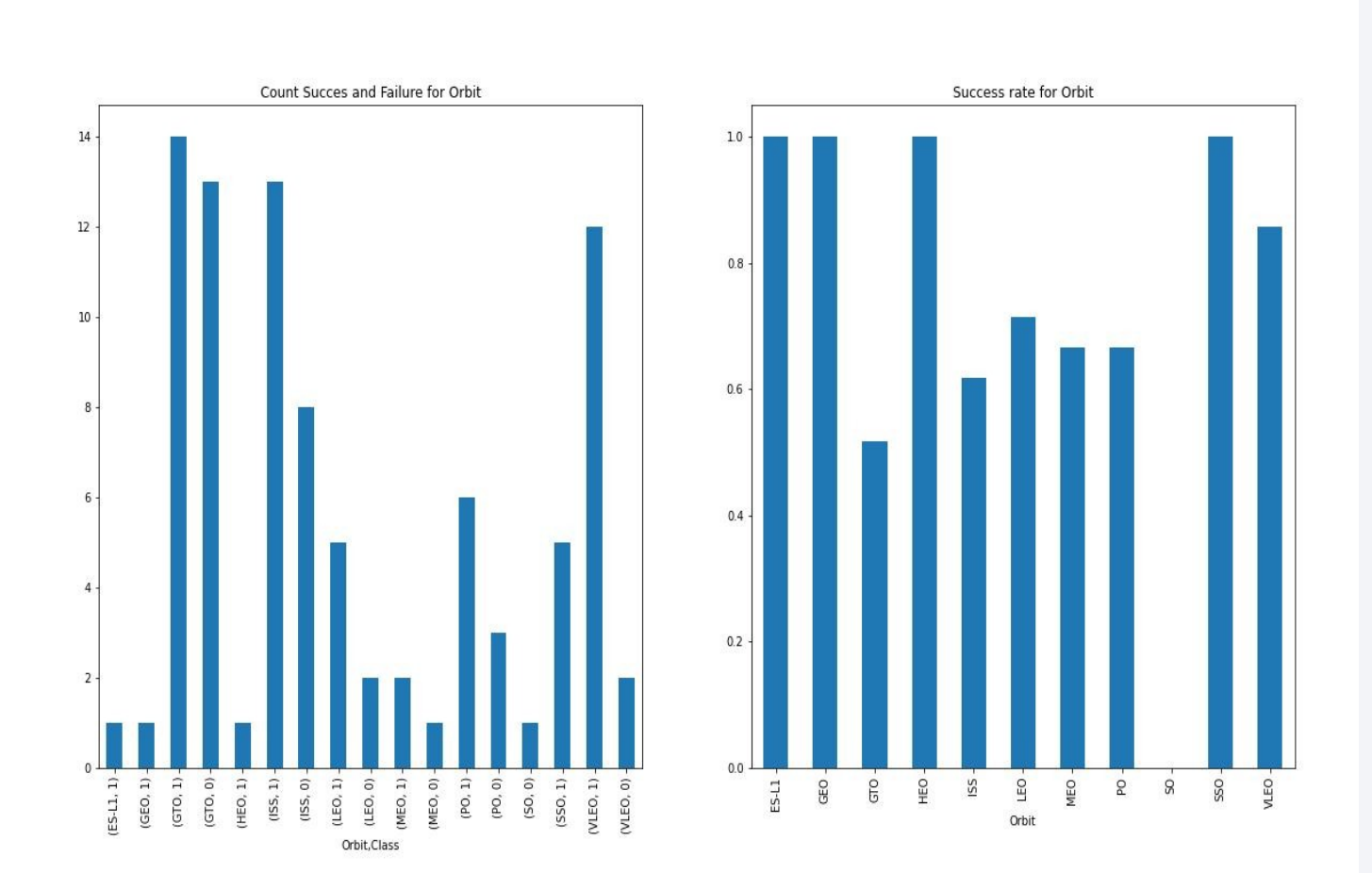
Payload vs. Launch Site

It represents the distribution of flights with the weight of the cargo, along with the result of the mission. From this graph it can be seen that the weight of the cargo has increased, plus the more weight the higher the percentage of successful missions.



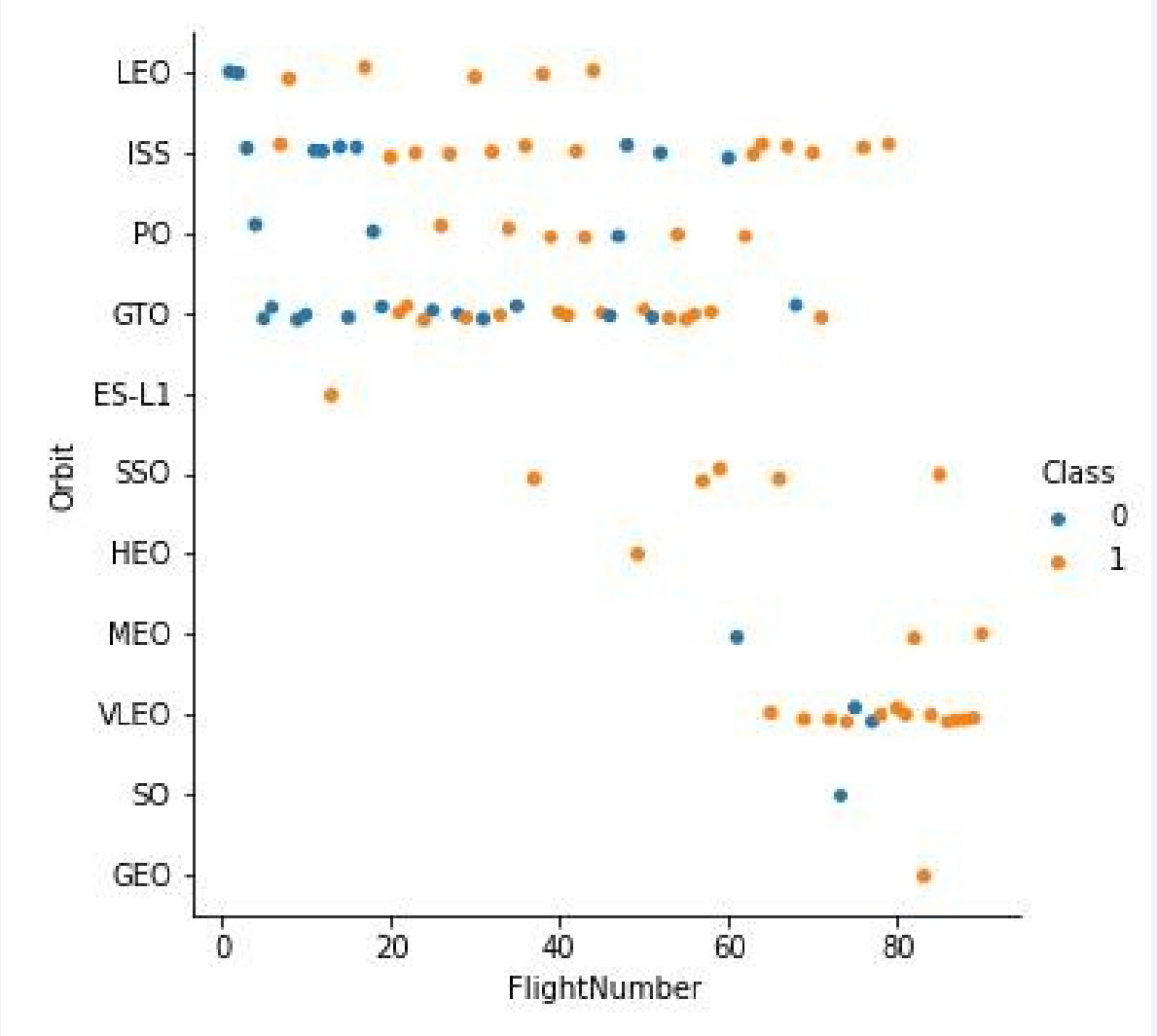
Success Rate vs. Orbit Type

It represents the success rate of the missions per orbit. Since there were several that had a 100% success rate, a histogram was created to compare the success rate with the number of missions per orbit.



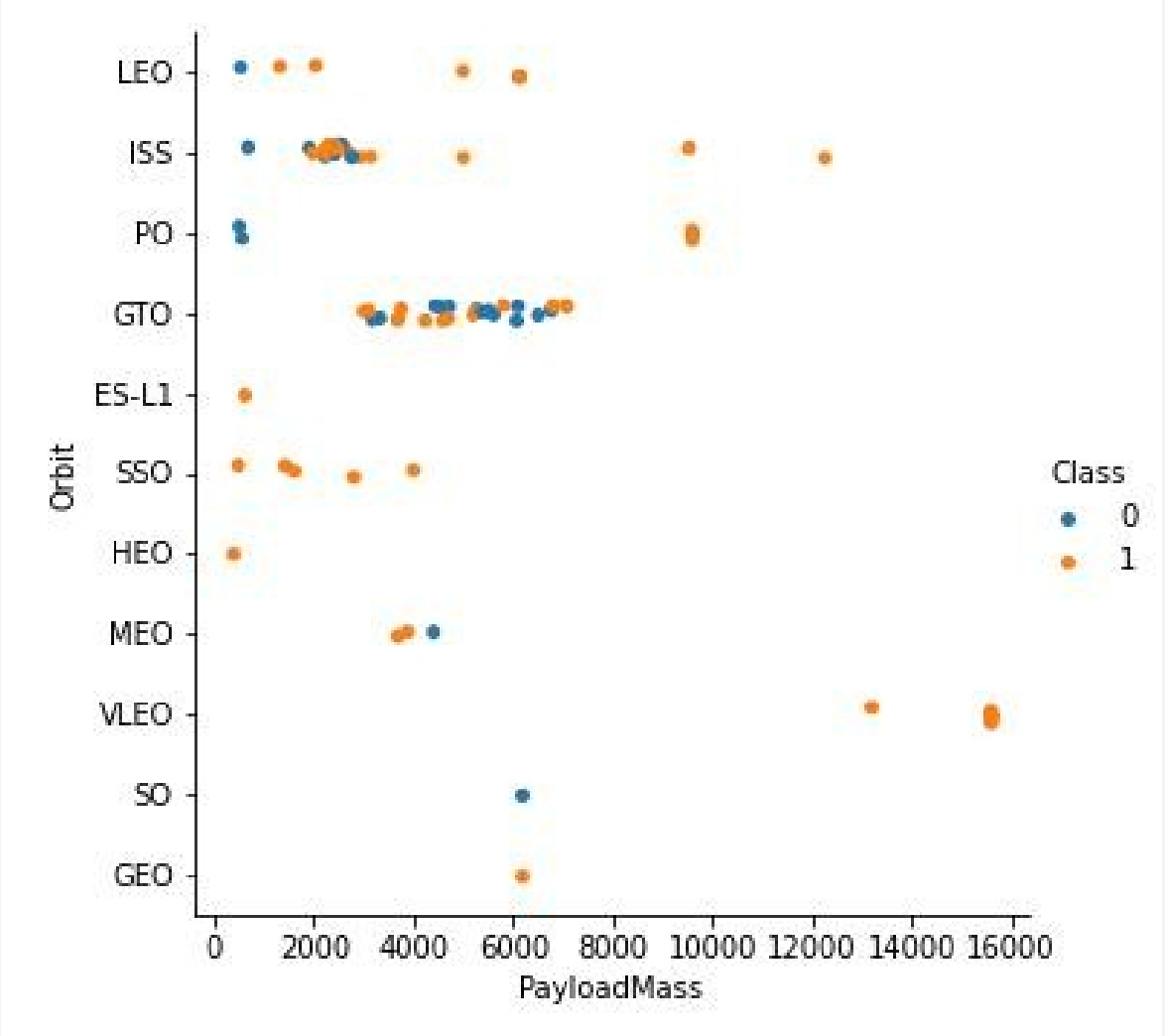
Flight Number vs. Orbit Type

It represents the orbit of each mission. It can be seen that, although missions have been sent to all orbits, the main objective of the latest missions has been the VLEO orbit and that of all the missions to this orbit, only 2 have failed.



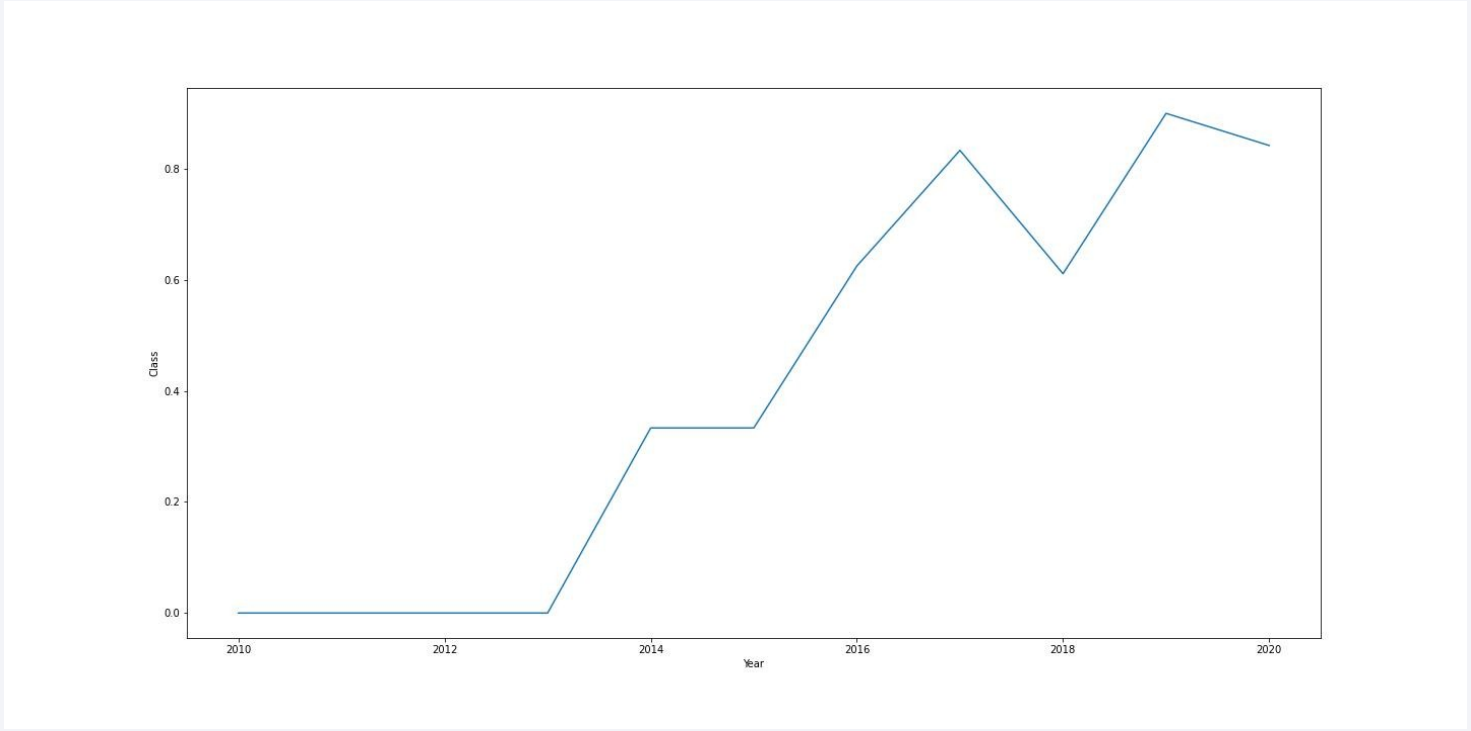
Payload vs. Orbit Type

It represents the weight sent to each orbit. Most missions do not exceed 8000 kg, but all missions with payload weights above 8000 kg have been successful.



Launch Success Yearly Trend

Line graph representing the success rate in each year. As can be seen, the success rate is increasing.



All Launch Site Names

The different categorical values of the Launch_site variable, the launch sites.

```
In [44]: %sql select distinct(Launch_Site) from SPACEXTABLE;
 * mysql+mysqlconnector://root:***@localhost/SPACE_Y
4 rows affected.
```

Out[44]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

The first 5 records matching the word CCA.

```
In [46]: %sql select Launch_Site from SPACEXTABLE where Launch_Site like "CCA%" limit 5
 * mysql+mysqlconnector://root:***@localhost/SPACE_Y
5 rows affected.
```

Out[46]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

The total weight sent by NASA.

```
In [52]: %sql select sum(PAYLOAD_MASS__KG_) as \
         "Mass total for NASA (CRS)" from \
         SPACEXTABLE where Customer = "NASA (CRS)"

          * mysql+mysqlconnector://root:***@localhost/SPACE_Y
         1 rows affected.

Out[52]:
```

Mass total for NASA (CRS)
45596

Average Payload Mass by F9 v1.1

The average weight of the load shipped with the booster version "F9 v1.1".

```
In [55]: %sql select avg(PAYLOAD_MASS__KG_) as \
         "Mean for booster version F9 v1.1" from \
         SPACEXTABLE where Booster_Version like "F9 v1.1%"

          * mysql+mysqlconnector://root:***@localhost/SPACE_Y
         1 rows affected.

Out[55]:
```

Mean for booster version F9 v1.1
2534.6667

First Successful Ground Landing Date

The date of the first falcon 9 launch.

```
In [56]: %sql select min(Date) from SPACEXTABLE \
         where Mission_Outcome ="Success"

          * mysql+mysqlconnector://root:***@localhost/SPACE_Y
         1 rows affected.

Out[56]:  min(Date)
          ----------
          2010-06-04
```

Successful Drone Ship Landing with Payload between 4000 and 6000

All records of the Booter_version variable where the mission was successful and the cargo shipped is between 4000 and 6000 kg.

```
In [64]: %sql select Booster_Version  from \
SPACEXTABLE where Mission_Outcome ="Success" \
and (PAYLOAD_MASS__KG_ between 4001 and 5999)

 * mysql+mysqlconnector://root:***@localhost/SPACE_Y
22 rows affected.
```

Out[64]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2

Total Number of Successful and Failure Mission Outcomes

Counting mission results between successful and unsuccessful.

```
In [114]: %sql select count(Mission_Outcome) as \
          Success,(select count(Mission_Outcome) from \
          SPACEXTABLE where Mission_Outcome like "%Fa%") as Failure \
          from SPACEXTABLE where Mission_Outcome like "%Suc%"

           * mysql+mysqlconnector://root:***@localhost/SPACE_Y
          1 rows affected.

Out[114]:
```

Success	Failure
100	1

Boosters Carried Maximum Payload

Selects all booster version values where the load sent was the maximum load sent of all missions.

```
In [110]: %sql select Booster_Version, PAYLOAD_MASS__KG_ \
          from SPACEXTABLE where PAYLOAD_MASS__KG_ = \
          (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

           * mysql+mysqlconnector://root:***@localhost/SPACE_Y
          12 rows affected.
```

Out[110]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Select the year, result, booster version and launch site. Where the mission failed and the launch year was 2015.

```
In [126]: %sql select monthname(Date), Landing_Outcome, \
          Booster_Version, Launch_Site from SPACEXTABLE  where \
          Landing_Outcome="Failure (drone ship)" and year(Date) = 2015

           * mysql+mysqlconnector://root:***@localhost/SPACE_Y
          2 rows affected.
```

Out[126]:

monthname(Date)	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

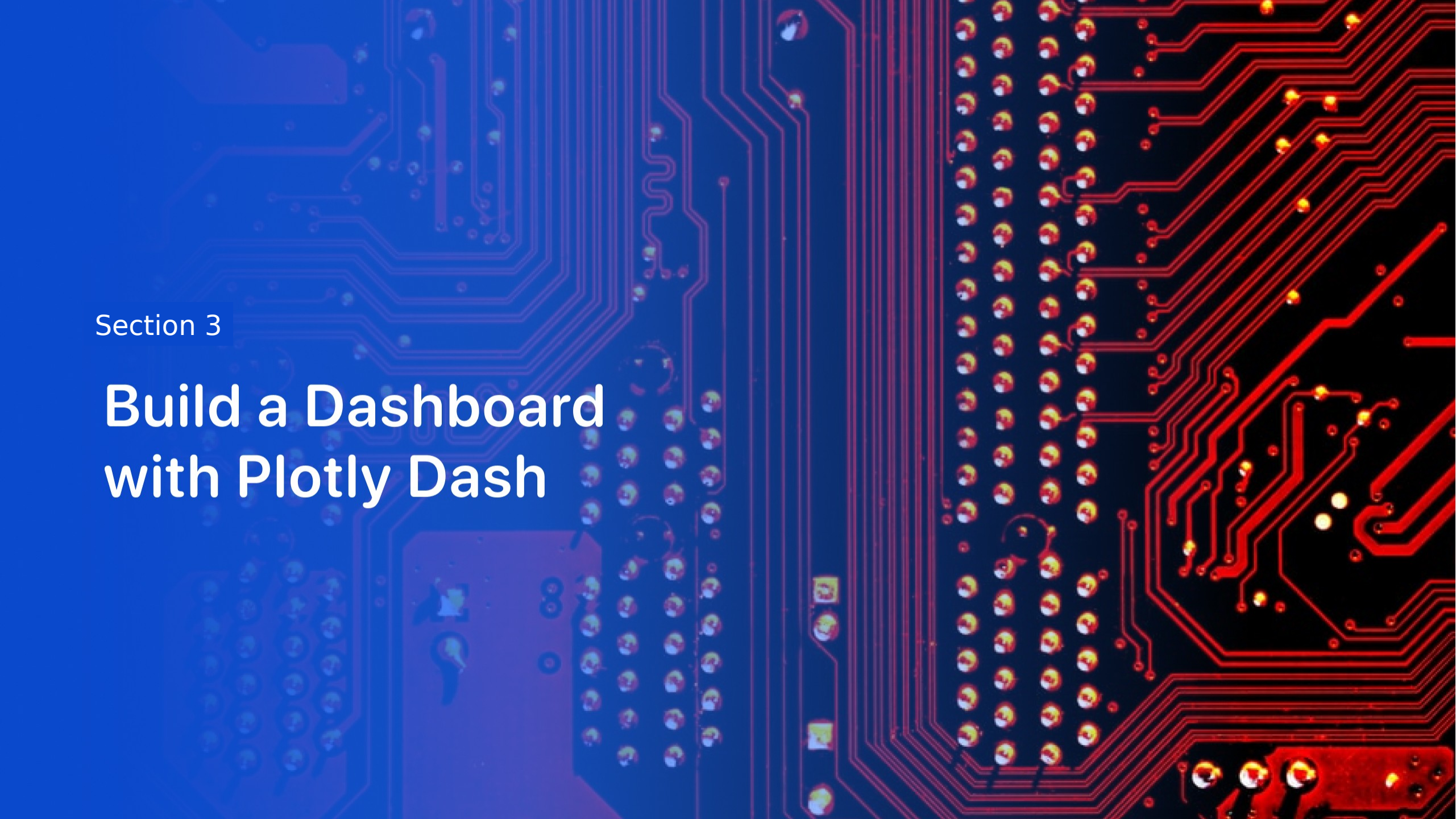
Perform a count of the missions with Failure (drone ship) and Success (ground pad) results that have been launched between 2010/06/04 and 2017/03/20. In addition, the count should be sorted in descending order.

```
In [186]: %sql select Landing_Outcome, count(Landing_Outcome) \
          as conteo from SPACEXTABLE where (Date between "2010-06-04" \
          and "2017-03-20") and Landing_Outcome in ("Failure (drone ship)", \
          "Success (ground pad)") group by Landing_Outcome order by conteo desc

           * mysql+mysqlconnector://root:***@localhost/SPACE_Y
          2 rows affected.
```

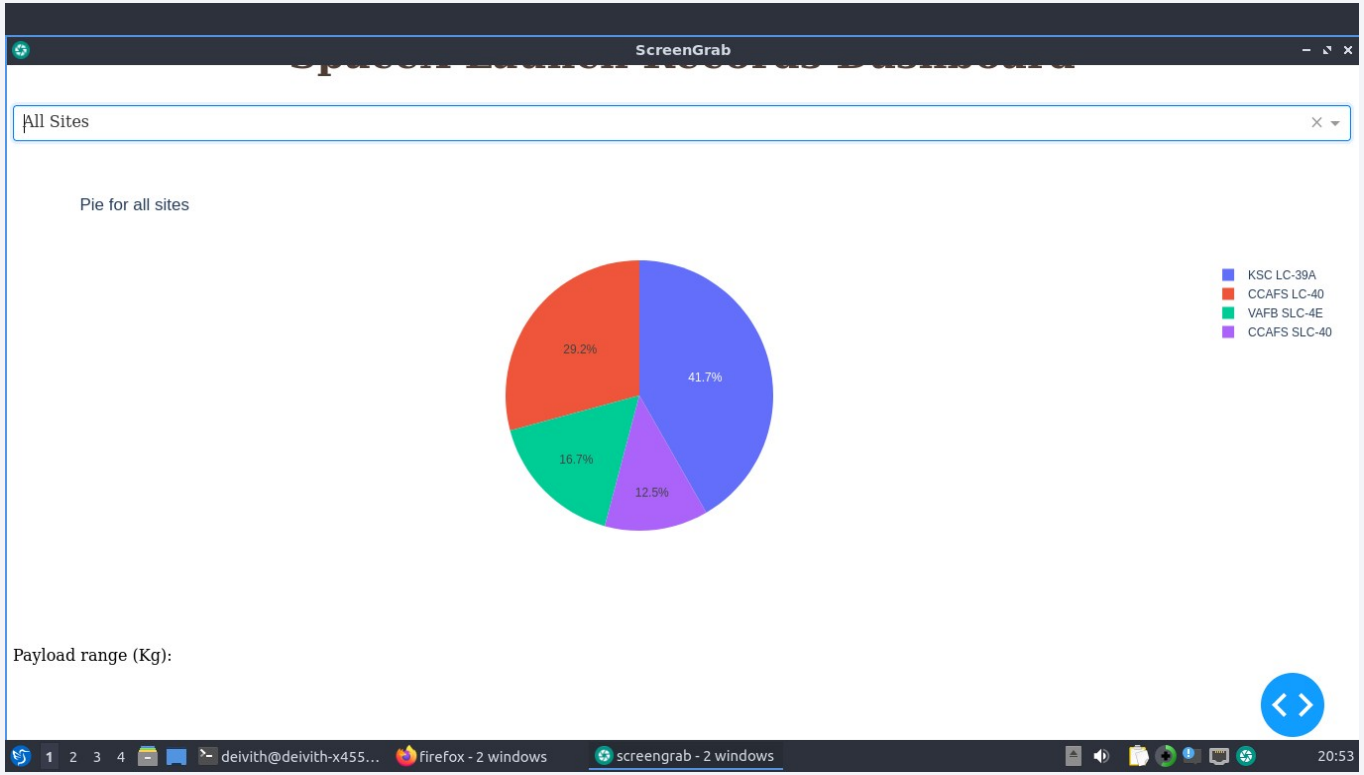
Out[186]:

Landing_Outcome	conteo
Failure (drone ship)	5
Success (ground pad)	3


Section 3
Build a Dashboard with Plotly Dash

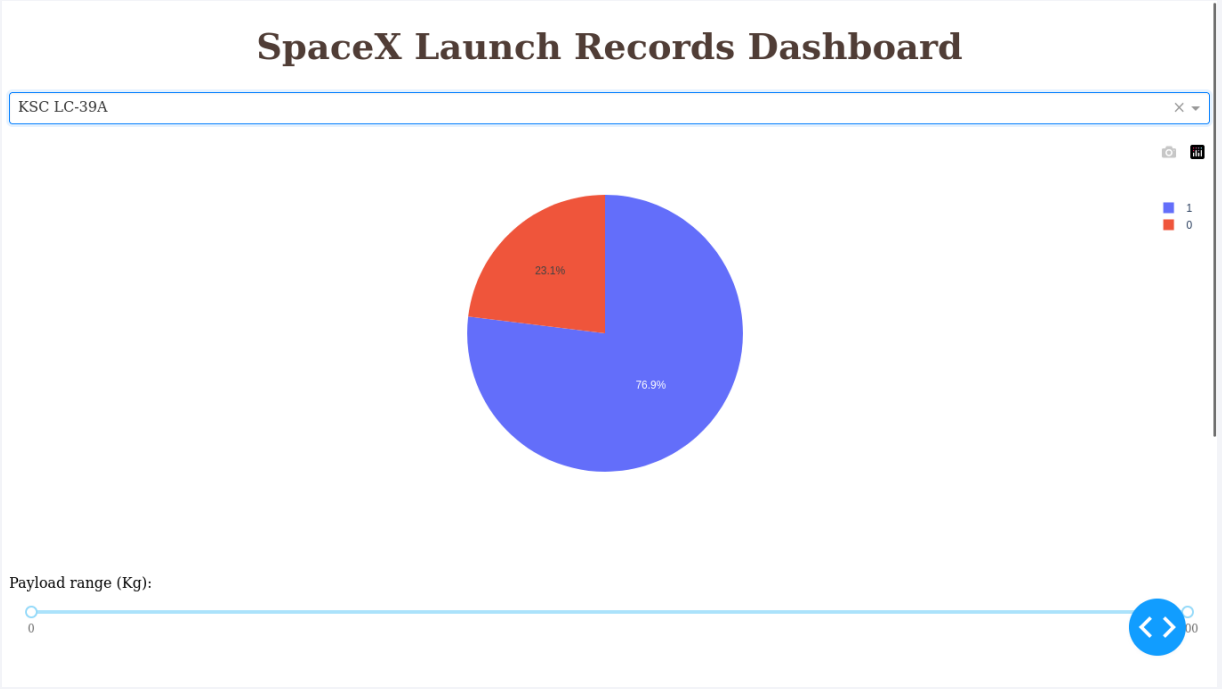
Success rate pie chart

This pie chart represents the success rate of all launch sites. You can easily see which of these locations have the highest success rate.



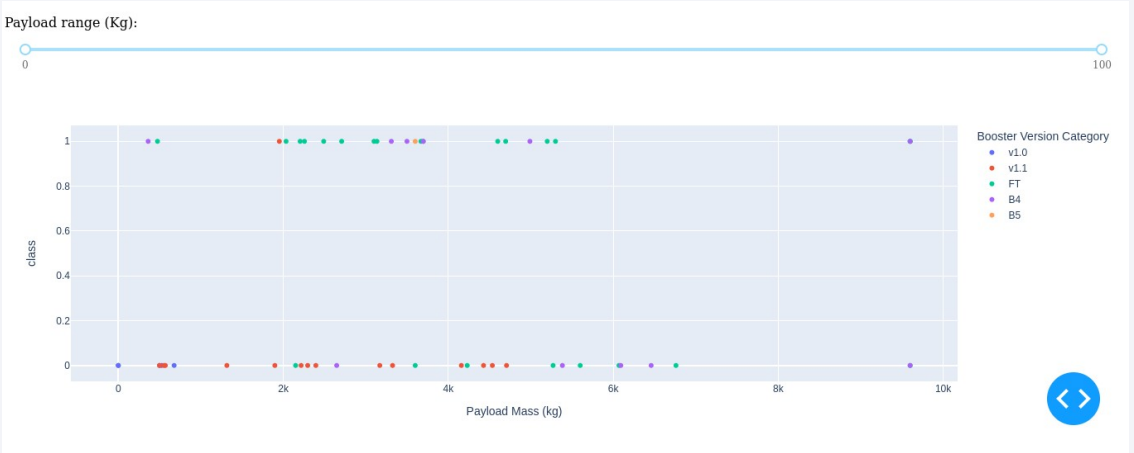
Percentage of results

The percentage of launch site results with the highest success rate.

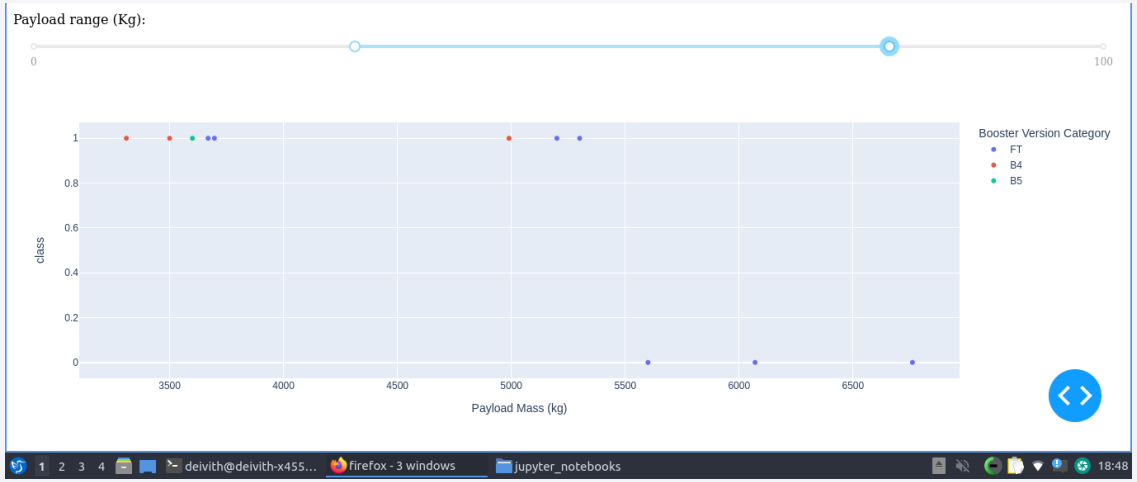


Scatter plot of results by year

Scatter plot of mission results by year. In addition, data are categorized by booster version.



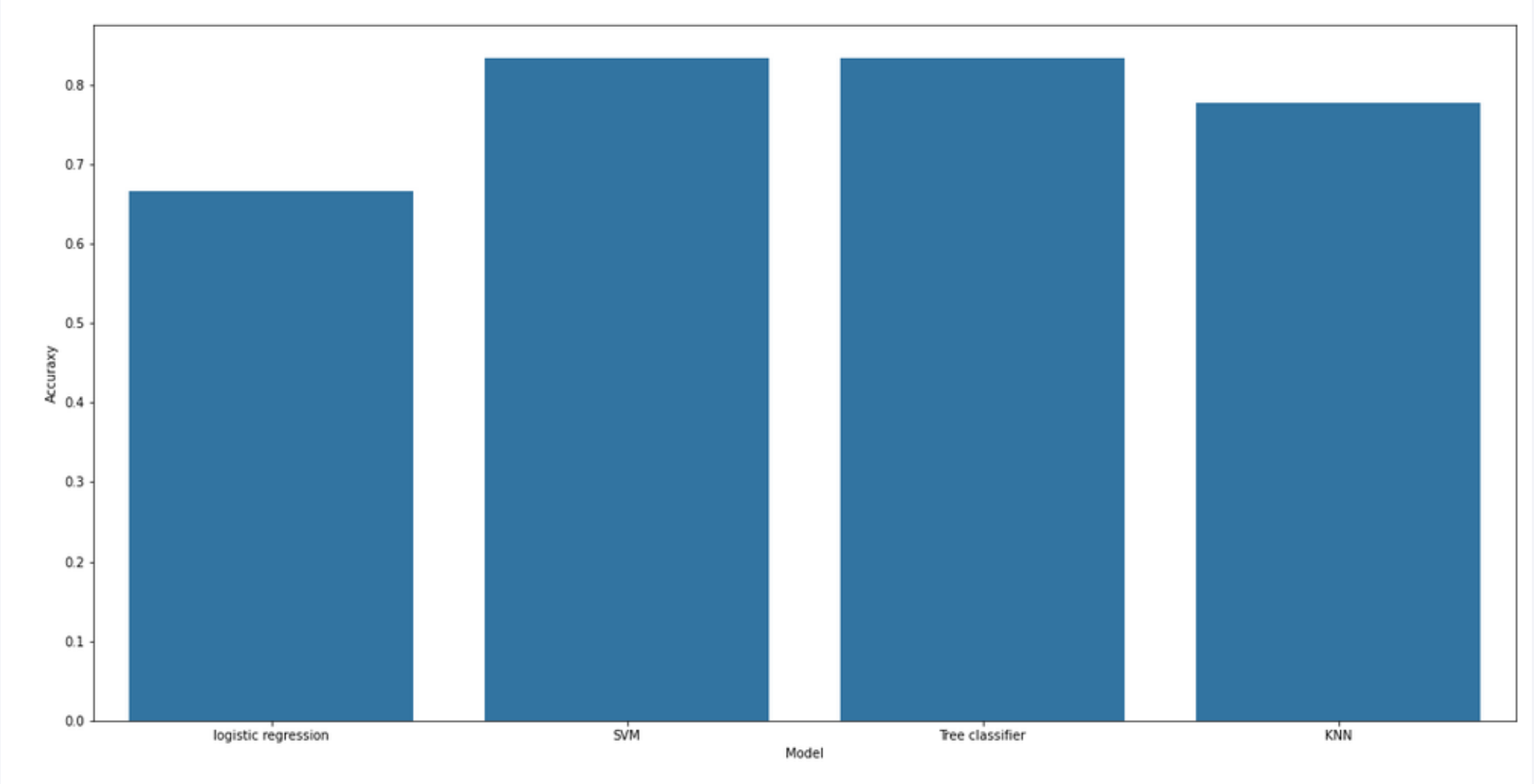
The scatter plot can filter the load weight values by displaying only those records that fall within the selected range.



Section 4

Predictive Analysis (Classification)

Classification Accuracy

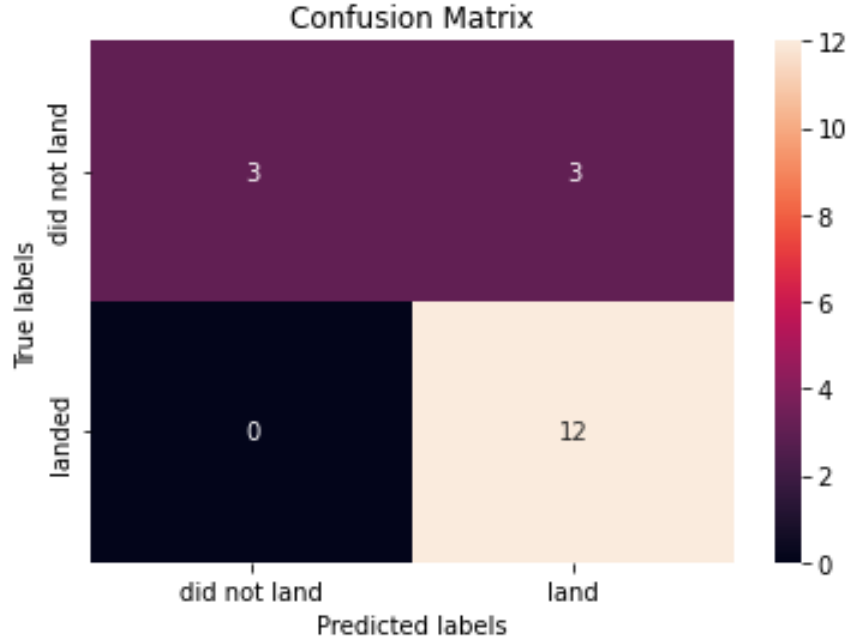


Confusion Matrix

The selected model was the decision tree, with an accuracy of 0.83 for the test data, but for the data of the records with the selected characteristics, the accuracy was 100%.

```
yhat = tree_cv.predict(x_test)
score = tree_cv.score(x_test,y_test)
print("Score for y test  = {}".format(score))
models["Tree classifier"] = score
plot_confusion_matrix(y_test,yhat)
```

```
Score for y test  = 0.8333333333333334
```



Conclusions

A number of features were obtained to maximize the chances of succeeding in the missions.

It was possible to create a model to make predictions about the possible outcome of a mission. But it lacks data on failed missions with the selected characteristics. This could be due to the fact that these characteristics and conditions are the best and will not fail, except for errors or problems during the mission that are not under our control.

It is recommended to look for more data on these missions or to improve the model by adding significant variables.

Thank you!