IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

Executive Summary

Summary of methodologies:

- Data collection – Web scraping
- Data wrangling
- EDA with statistics and SQL queries
- Visualization with charts, graphics and Folium maps.
- Prediction with DT, kNN, SVM, LogReg.

Summary of all results:

- Importance of the Booster version and the payload mass.
- Accuracy of 0,83 in prediction.

Introduction

Project background and context:

SpaceX advertises Falcon 9 rocket launches on its website, costing $62 million; other providers cost upwards of $165 million each, much of the savings due to SpaceX being able to reuse the first stage. Thus, predicting whether the first stage will land successfully allows one to determine the cost of a launch

Problems to find answers:

- Predict whether the first stage will land successfully.
- Find the relationship among the variables involved.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data is collected using an API: the SpaceX REST API.
- Perform data wrangling
  - Use of Pandas tools and Python functions.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Use of statistic parameters and graphics for EDA, and SQL queries.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Use of Folium and Plotly Dash tools.
- Perform predictive analysis using classification models
  - Use of different algorithms in scikit-learn.

Data Collection – SpaceX API

PROCESS OF DATA COLLECTION WITH SPACEX REST CALLS

1. Make a request using the Requests library.

2. Use URL to point to a specific API endpoint:

requests.get(spacex_url)

3. Decode JSON response content in a dataframe.

pd.json_normalize(response.json())

GitHub URL of completed SpaceX API calls notebook
https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod1_Lab1_Collecting%20de%20data_FINALIZADO.ipynb

Data Collection - Scraping

WEB SCRAPING PROCESS

1. Make an HTTP GET method to request the Falcon9 launch HTML page, as an HTTP response

requests.get(static_url).text

2. Create BeautifulSoup object for web scraping on Wiki pages containing HTML tables with Falcon 9 launch records.

BeautifulSoup(response, 'html.parser')

3. Convert data into a Pandas dataframe for better visualization and analysis.

pd.DataFrame.from_dict(launch_dict)

GitHub URL of the completed web scraping notebook:

https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod1_Lab2_Web%20scraping_FINALIZADO.ipynb

Data Wrangling

PROCESS OF DATA WRANGLING TO DETERMINE THE LANDING SUCCESS RATE

1. **Calculate the number of launches on each site.**

df.LaunchSite.value_counts()

2. **Calculate** the number and occurrence of each orbit .

df.Orbit.value_counts()

3. **Calculate the number and occurence of mission outcome per orbit type.**

df.Outcome.value_counts()

4. **Create a landing outcome label from Outcome column.**

foroutcome**in**df['Outcome']:**if**outcome**in**bad_outcomes:landing_class.append(0)**else**:landing_class.append(1)

GitHub URL of the completed data wrangling notebook:

https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod1_labs-spacex-Data%20wrangling_FINALIZADO.ipynb

EDA with Data Visualization

SUMMARIZE OF THE CHARTS PLOTTED

Type of chart	Variables	Objective
Scatter plot	PayloadMass vs. FlightNumber with overlay of the outcome of the launch.	To see how FlightNumber and Payload variables would affect the launch outcome.
Scatter plot	LaunchSite vs. FlightNumber.	To visualize the relationship between Flight Number and Launch Site.
Scatter plot	LaunchSite vs. payload mass.	To visualize the relationship between Payload and Launch Site.
Bar chart	Sucess rate of each orbit.	To visualize relationship between success rate of each orbit type.
Scatter plot	Orbit vs. FlightNumber.	To visualize the relationship between FlightNumber and Orbit type.
Scatter plot	Orbit vs. Payload.	To visualize the relationship between Payload and Orbit type.
Line chart	Average success rate vs. Year.	To Visualize the launch success yearly trend.

GitHub URL of the completed EDA with data visualization notebook:

https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod2_jupyter-labs-eda-datavisualization_FINALIZADO.ipynb

EDA with SQL

SUMMARIZE OF THE SQL QUERIES PERFORMED (first five)

- Display of the names of the unique launch sites in the space mission
- Display of 5 records where launch sites begin with the string 'KSC'
- Display of the total payload mass carried by boosters launched by NASA (CRS)
- Display of average payload mass carried by booster version F9 v1.1
- List of the date where the first successful landing outcome in drone ship was achieved

GitHub URL of completed EDA with SQL notebook:
https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod2_jupyter-labs-eda-SQL-edx_FINALIZADO.ipynb

EDA with SQL

SUMMARIZE OF THE SQL QUERIES PERFORMED (last five)

- List of the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- List of the total number of successful and failure mission outcomes
- List of the names of the booster_versions which carried the maximum payload mass.
- List of the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking of the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

GitHub URL of completed EDA with SQL notebook:
https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod2_jupyter-labs-eda-SQL-edx_FINALIZADO.ipynb

Build an Interactive Map with Folium

SUMMARIZE OF MAP OBJECTS CREATED AND ADDED TO A FOLIUM MAP

- Create a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name to identify it on the map
- Add circlec for each launch site in data frame to identify them on the map.
- Add marker cluster to current site map.
- Add mouse position to get the coordinate for a mouse over the map.
- Create a `folium marker and PolyLine to show distances.
- Draw lines between sites to study proximities between them.

GitHub URL of completed interactive map with Folium map:
https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod3_lab_Launch_site_location_Finalizado.ipynb

Build a Dashboard with Plotly Dash

Dashboard included:

- Pie chart: to show the success ratio of launches sites.
- Dropdown list: to enable Launch Site selection.
- Slider: to change the payload range.
- Scatter plot: to see the relationship between payload, launch success and booster version used,

GitHub URL of completed Plotly Dash lab:
https://github.com/mauriciohv/conception/blob/af45eb145fd9de39653d86d3bf21c2ed77a55fec/SpaceX_launch_records_dashboard.ipynb

Predictive Analysis (Classification)

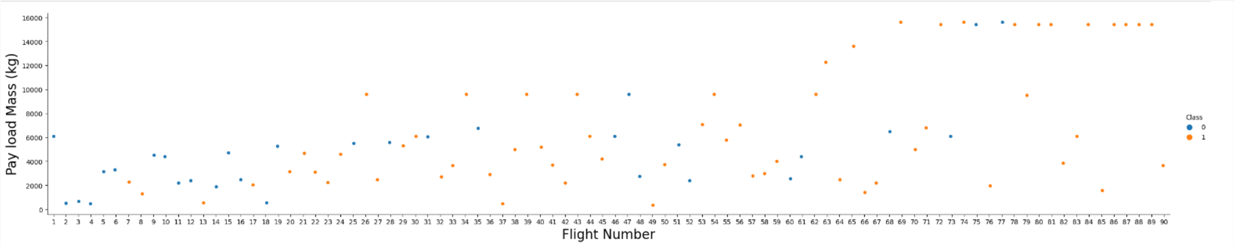
SUMMARIZE OF CLASSIFICATION PROCESS

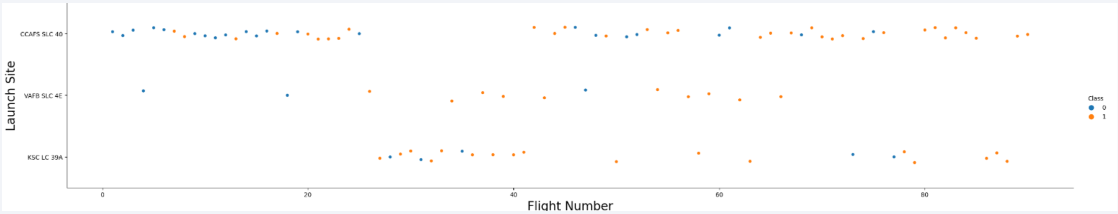
- Create a NumPy array for to the variable Y applying the method to_numpy().
- Standardize the data in X using StandardScaler().
- Split the data X and Y into training (80%) and test data (20%) using train_test_split.
- Fit a logistic regression, SVM, DT and kNN model.
- Improve model tuning hyperparameters with iterative process GridSearchCV.
- Evaluate the accuracy on the test data using the method score.
- Find the best performing model based on maximum accuracy.

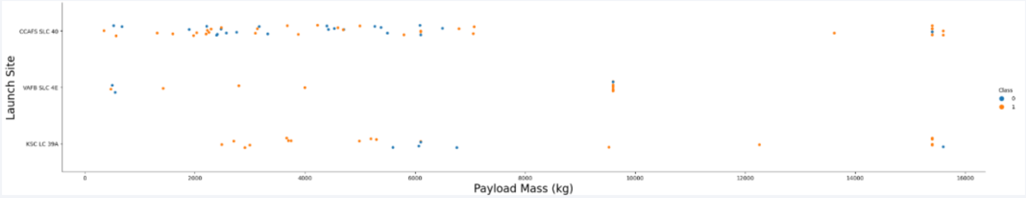
GitHub URL of completed predictive analysis lab:
https://github.com/mauriciohv/conception/blob/bb5ecb49f0127e6bc3a7664036e7f49502c5e4e2/Mod4_SpaceX_ML%20Prediction_Part_5%20(FINALIZADO).ipynb

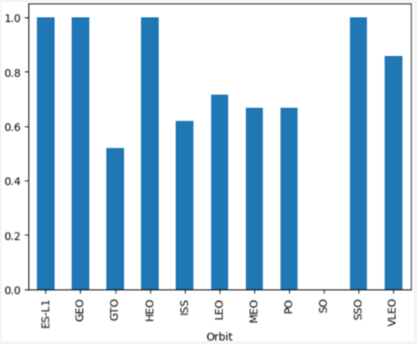
Results

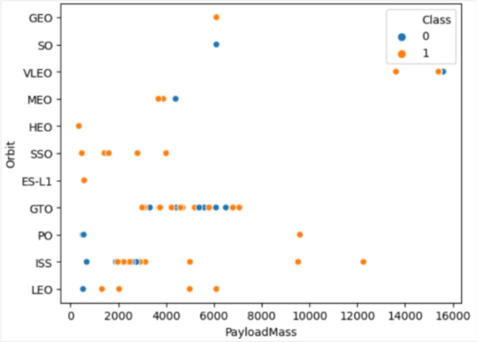
EXPLORATORY DATA ANALYSIS RESULTS

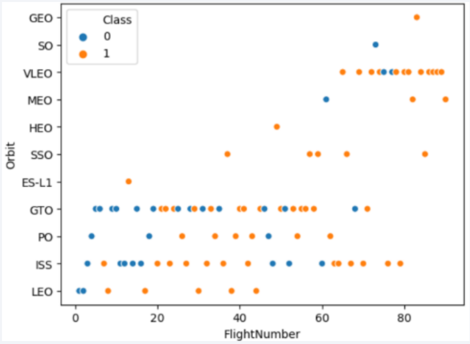


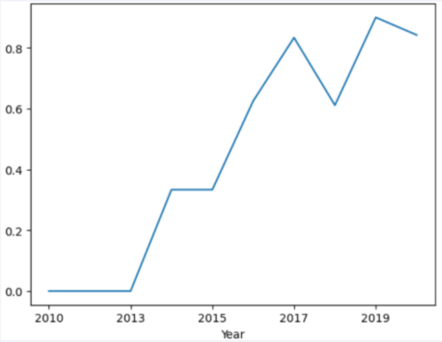






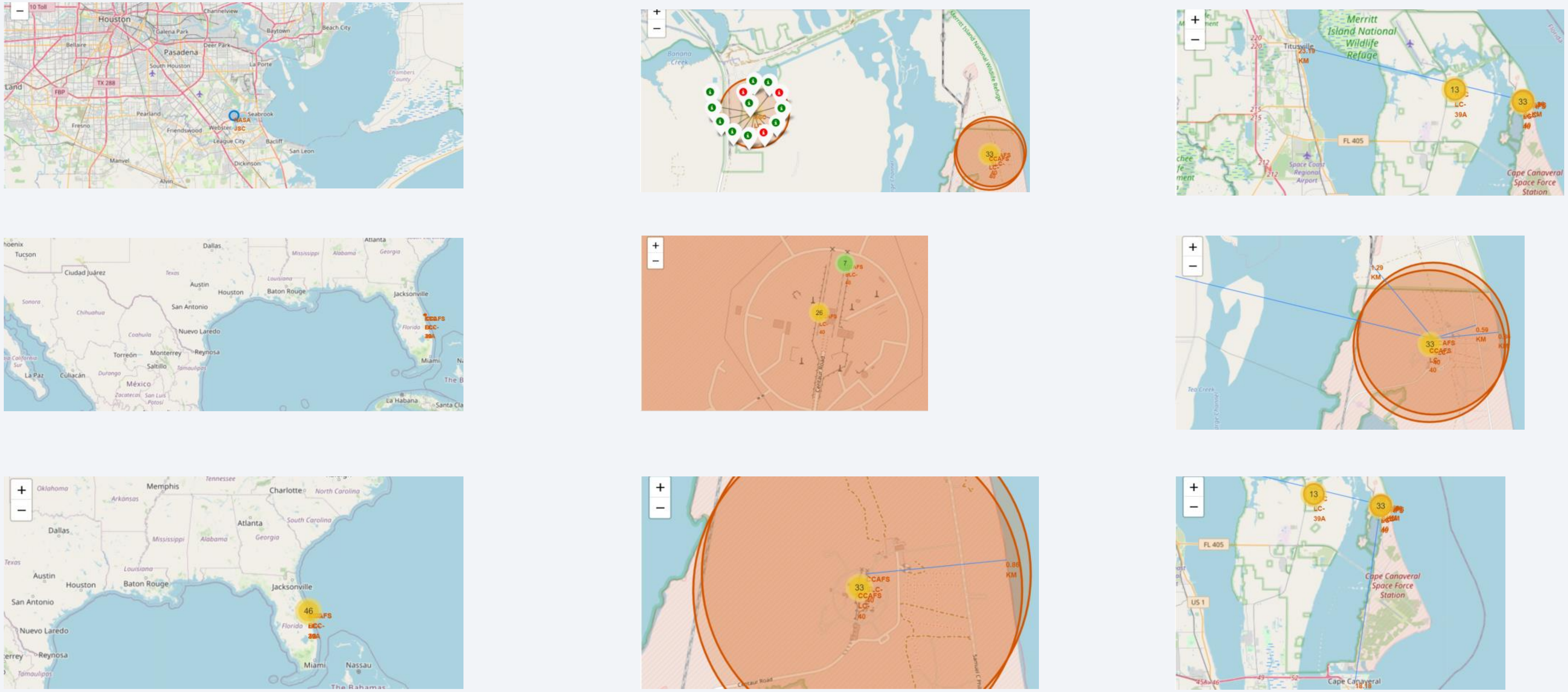






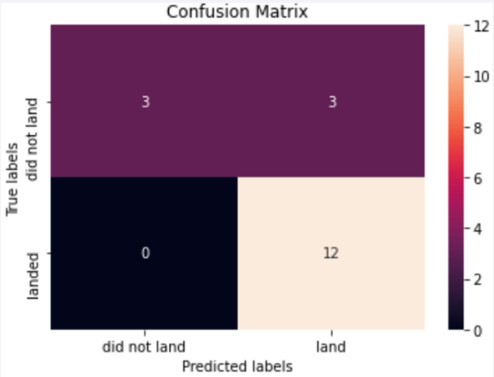
Results

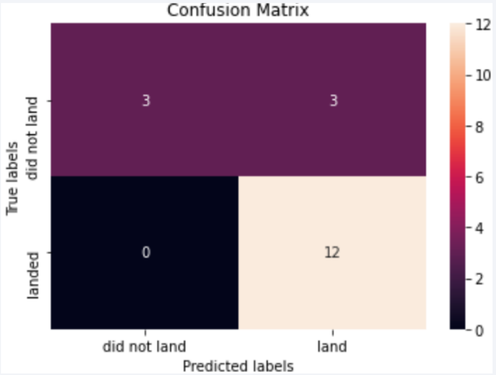
INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

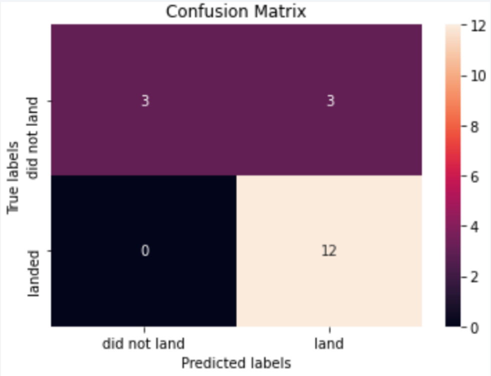


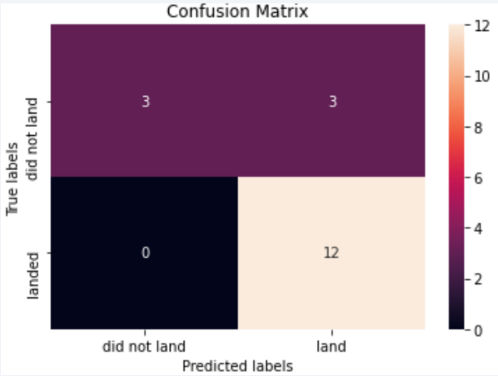
Results

PREDICTIVE ANALYSIS RESULTS










```
scores = [lr_score,svm_score,tree_score,knn_score]
print(scores)
print(max(scores))
```

```
[0.8333333333333334, 0.8333333333333334, 0.7777777777777778, 0.8333333333333334]
0.8333333333333334
```

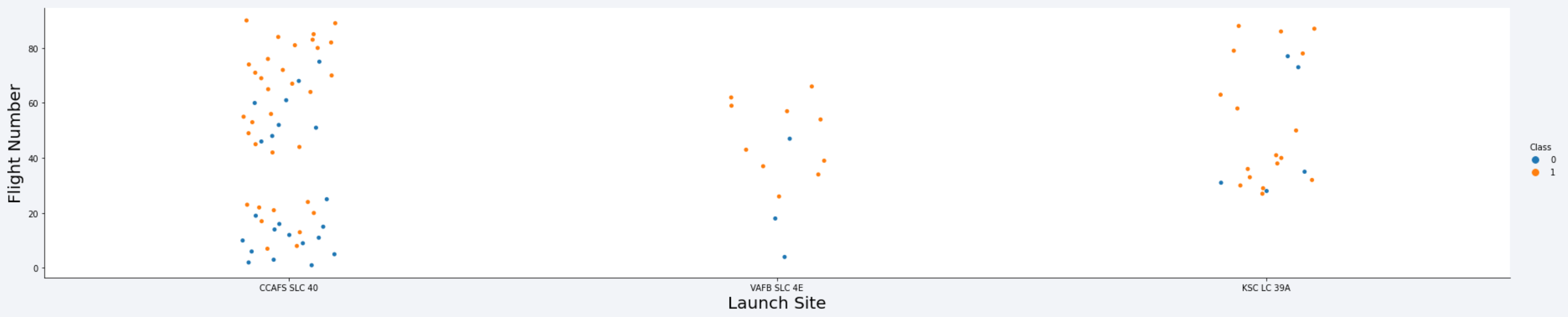
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

SCATTER PLOT OF FLIGHT NUMBER VS. LAUNCH SITE.

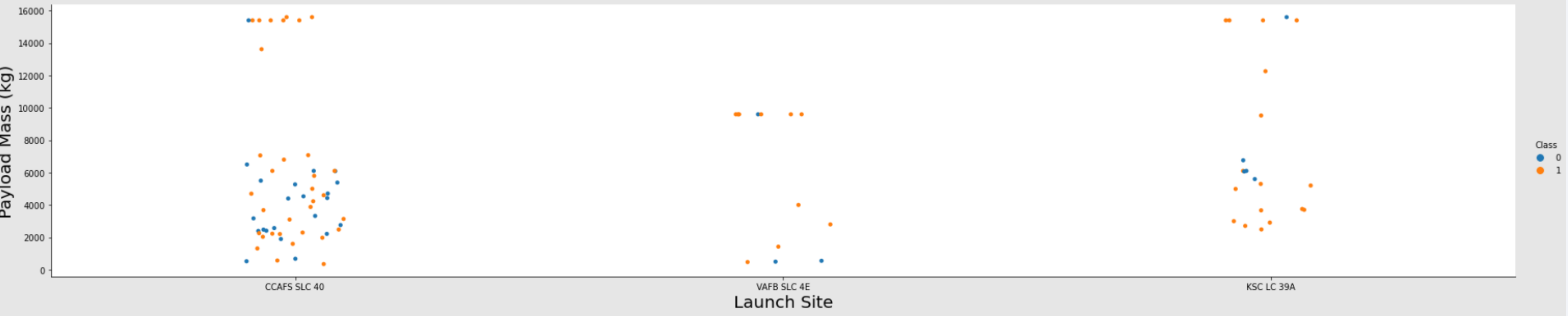
- The greater the number of launches, the greater the number of success launches.
- There were much more launches in CCAFS than the other sites.



Payload vs. Launch Site

SCATTER PLOT OF PAYLOAD VS. LAUNCH SITE

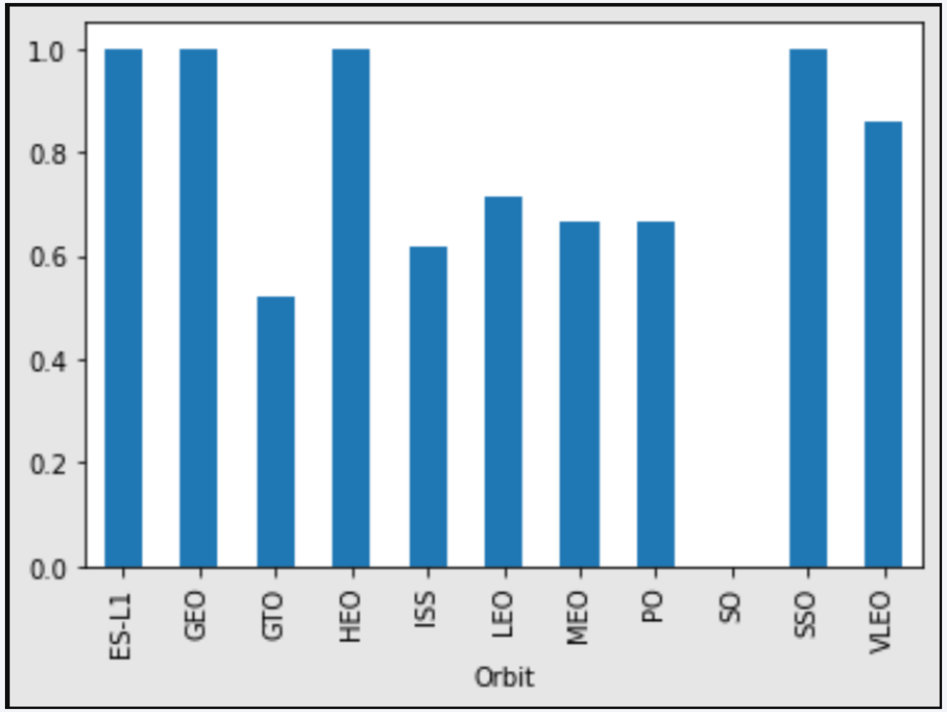
There were more fails with less payload mass.



Success Rate vs. Orbit Type

BAR CHART FOR THE SUCCESS RATE OF EACH ORBIT TYPE

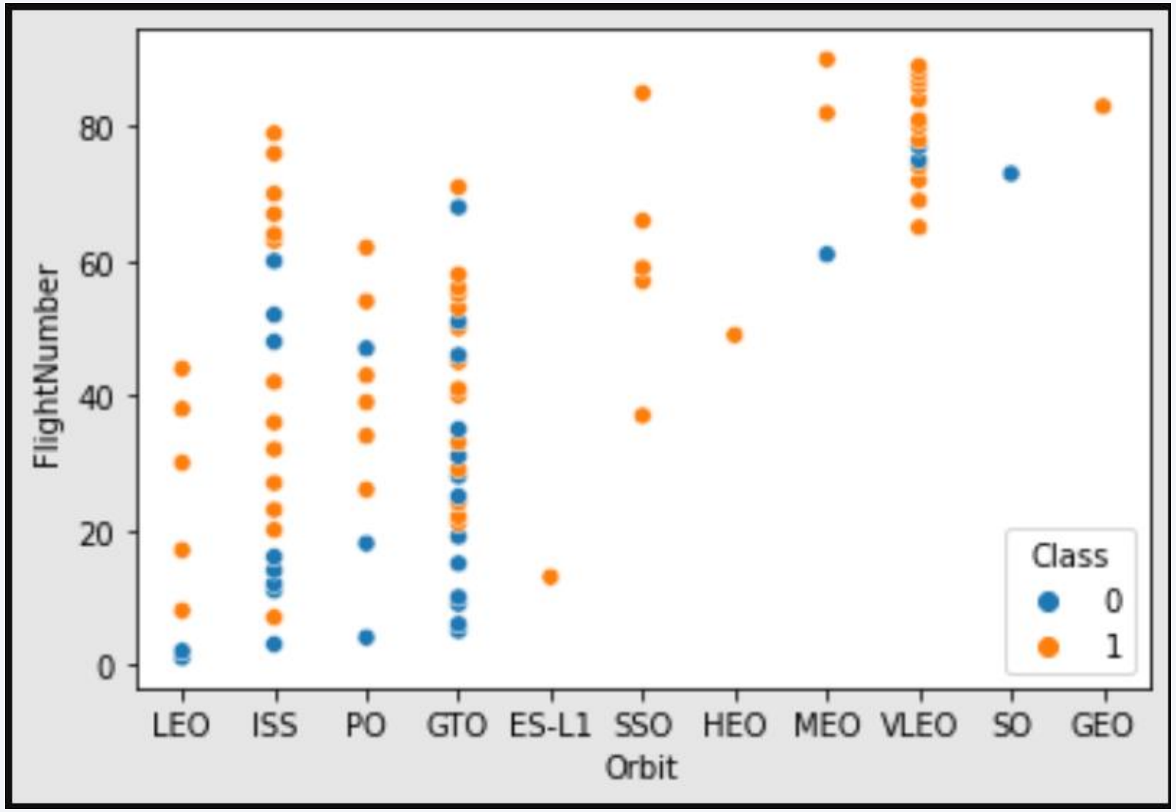
- There are four orbits with a perfect success rate.



Flight Number vs. Orbit Type

SCATTER PLOT OF FLIGHT NUMBER VS. ORBIT TYPE

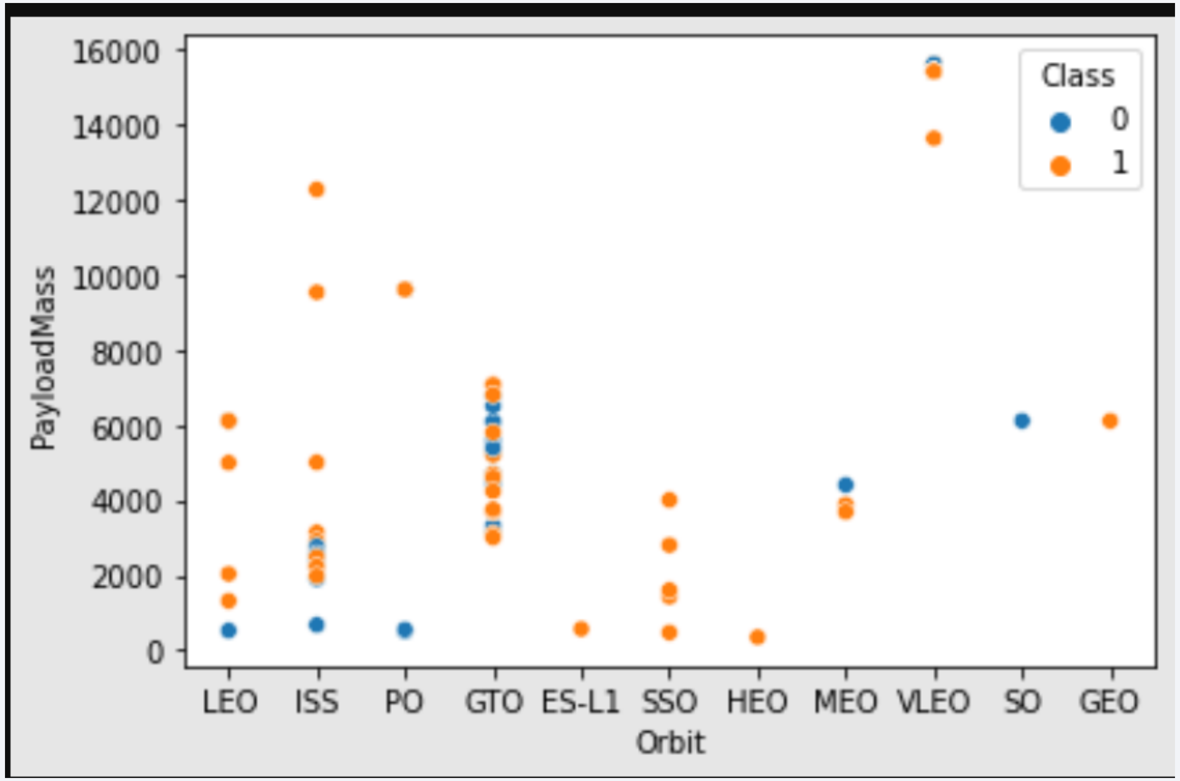
- The greater the number of flights, the greater the success in launches.



Payload vs. Orbit Type

SCATTER PLOT OF PAYLOAD VS. ORBIT TYPE

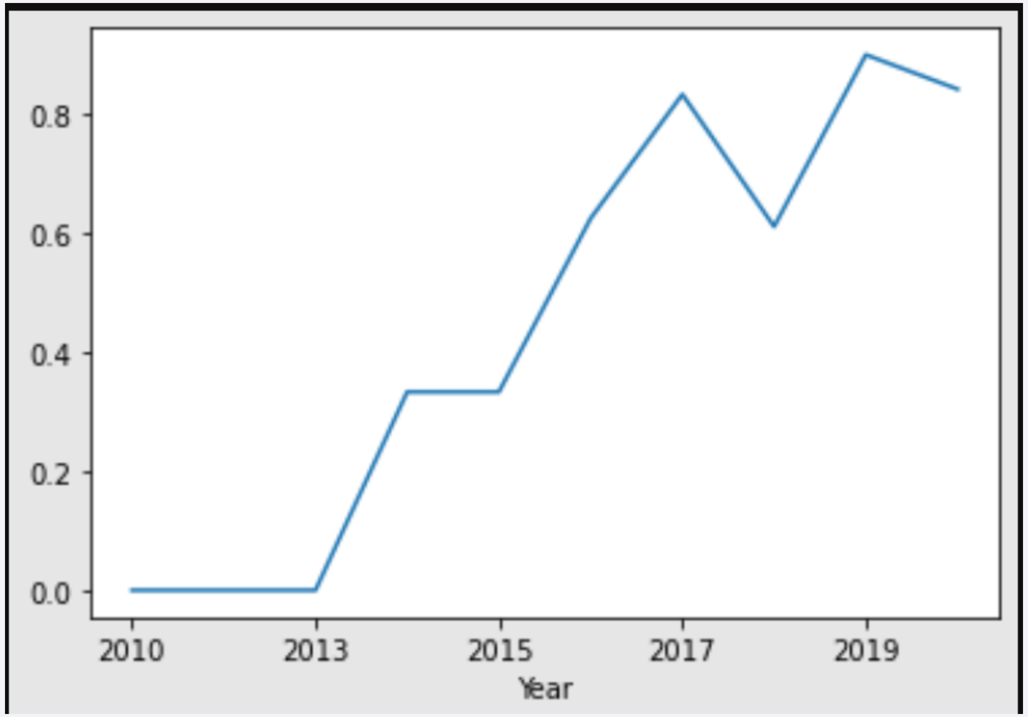
- There is not a clear relationship between the variables and categories.



Launch Success Yearly Trend

LINE CHART OF YEARLY AVERAGE SUCCESS RATE

- There is a clear correlation between the passing of the years and the increase in the success rate.



All Launch Site Names

The next code displays the names of the four unique launch sites.

%sql select Unique(LAUNCH_SITE) from SPACEXTBL

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

Launch Site Names Begin with 'KSC'

The next code shows the five records where launch sites' names start with `KSC':

%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'KSC%' LIMIT 5;

launch_site
KSC LC-39A
KSC LC-39A
KSC LC-39A
KSC LC-39A
KSC LC-39A

Total Payload Mass

The next code calculates the total payload carried by boosters from NASA:

%sql select sum(PAYLOAD_MASS__KG_) as Totalpayloadmass from SPACEXTBL where CUSTOMER='NASA (CRS)';

totalpayloadmass
45596

Average Payload Mass by F9 v1.1

The next code calculates the average payload mass carried by booster version F9 v1.1:

```
%sql select avg(PAYLOAD_MASS__KG_) as Averagepayloadmass from SPACEXTBL
where BOOSTER_VERSION='F9 v1.1';
```

averagepayloadmass
2928

First Successful Ground Landing Date

The next code lists the dates of the first successful landing outcome in drone ship:

%sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)';

1
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

The next code lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (ground pad)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

The next code calculates the total number of successful and failure mission outcomes:

%sql select count(MISSION_OUTCOME) as missionoutcome from SPACEXTBL GROUP BY MISSION_OUTCOME;

missionoutcome
1
99
1

Boosters Carried Maximum Payload

The next code lists the names of the booster which have carried the maximum payload mass:

%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- The next code lists the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017:
- %sql SELECT MONTH(DATE),LANDING__OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(MONTH FROM DATE)='2017';

1	landing__outcome	booster_version	launch_site

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

The next code ranks the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%sql select LANDING__OUTCOME from SPACEXTBL where DATE between
'2010-06-04' and '2017-03-20' order by DATE DESC;
```

landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 3

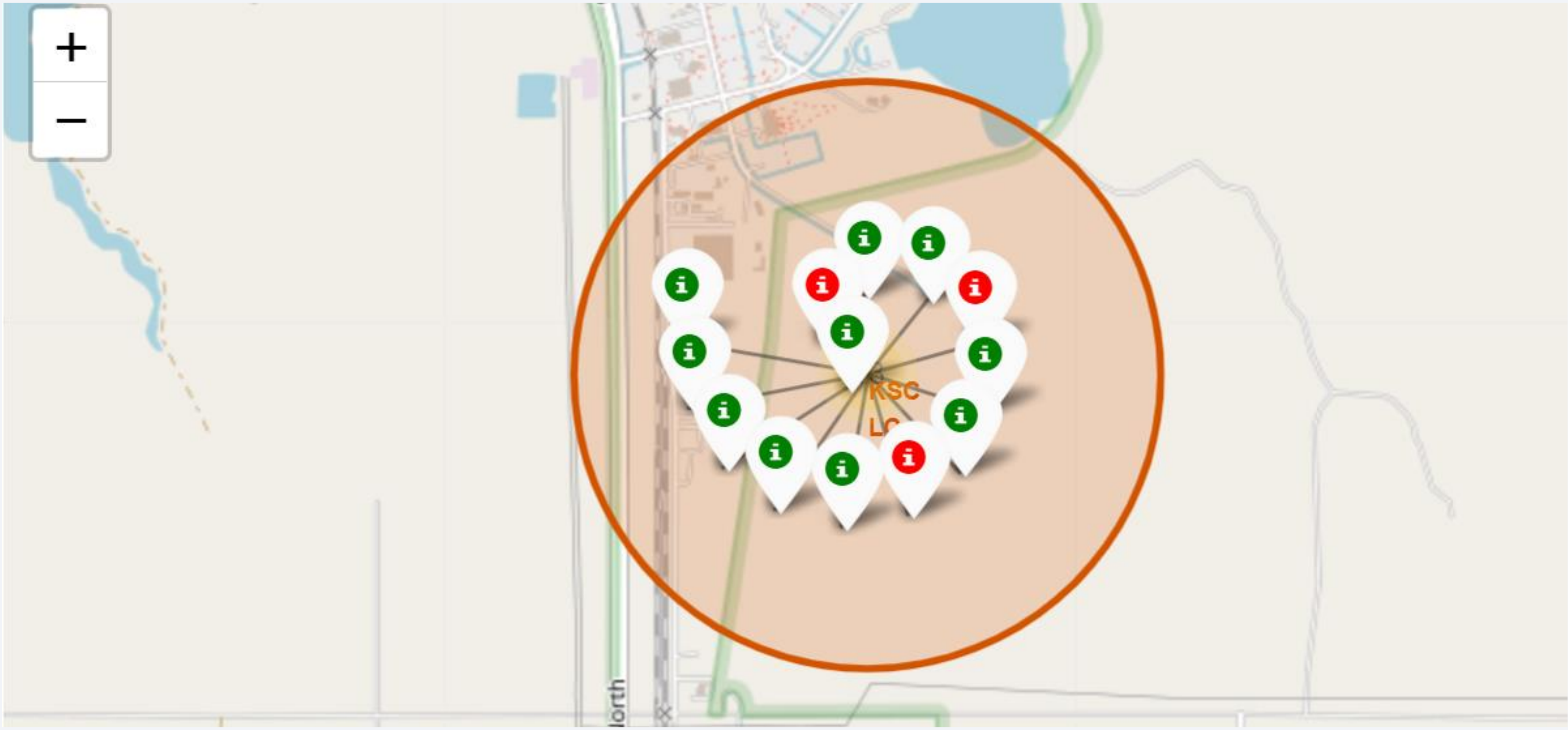
Launch Sites Proximities Analysis

Identification of all launch sites in Folium Map

Launh sites are shown with folium circle function and add_child method.

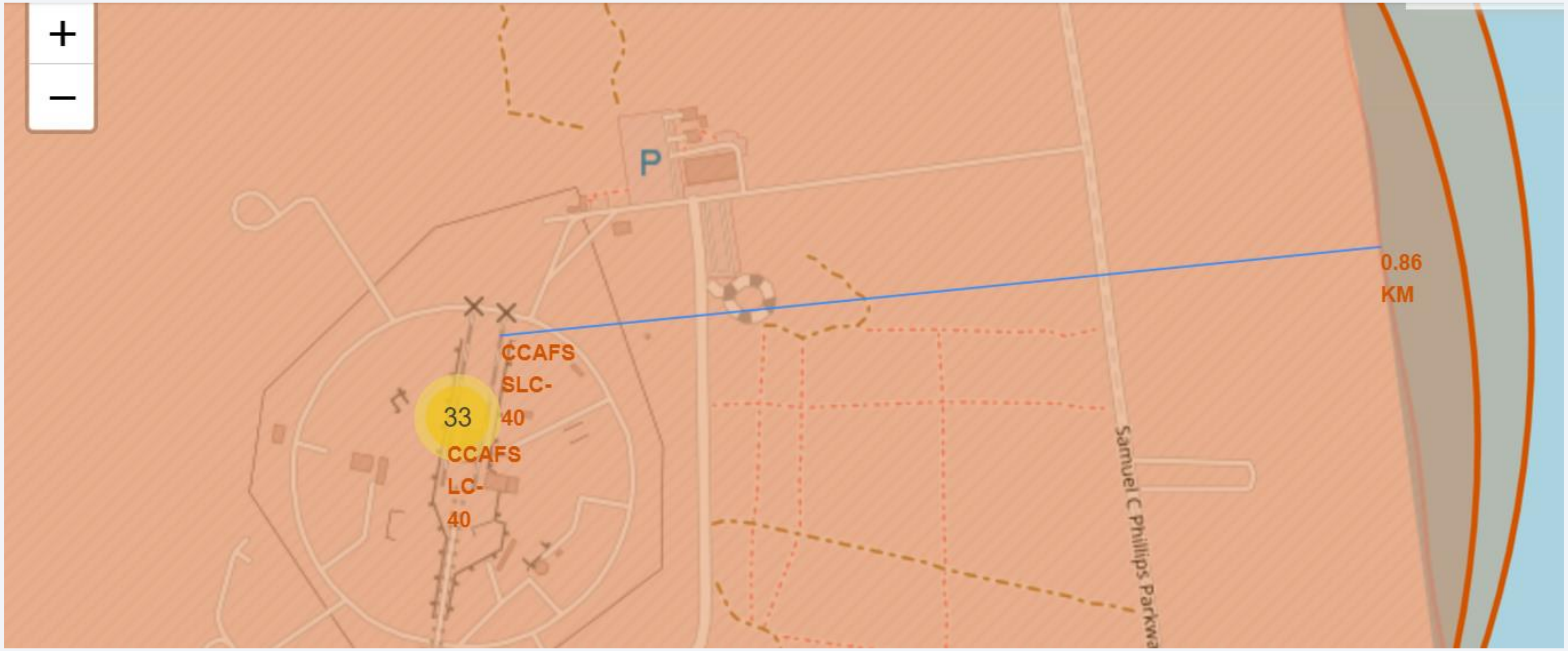
Color-labeled launch outcomes

Successful launches (in green) and failed launches (in red) are shown with the function MarkerCluster().



Important distances on the map

- Distance between launch sites and places such as railway, highway, coastline, are calculated and displayed on the map based on their Lat and Long values.

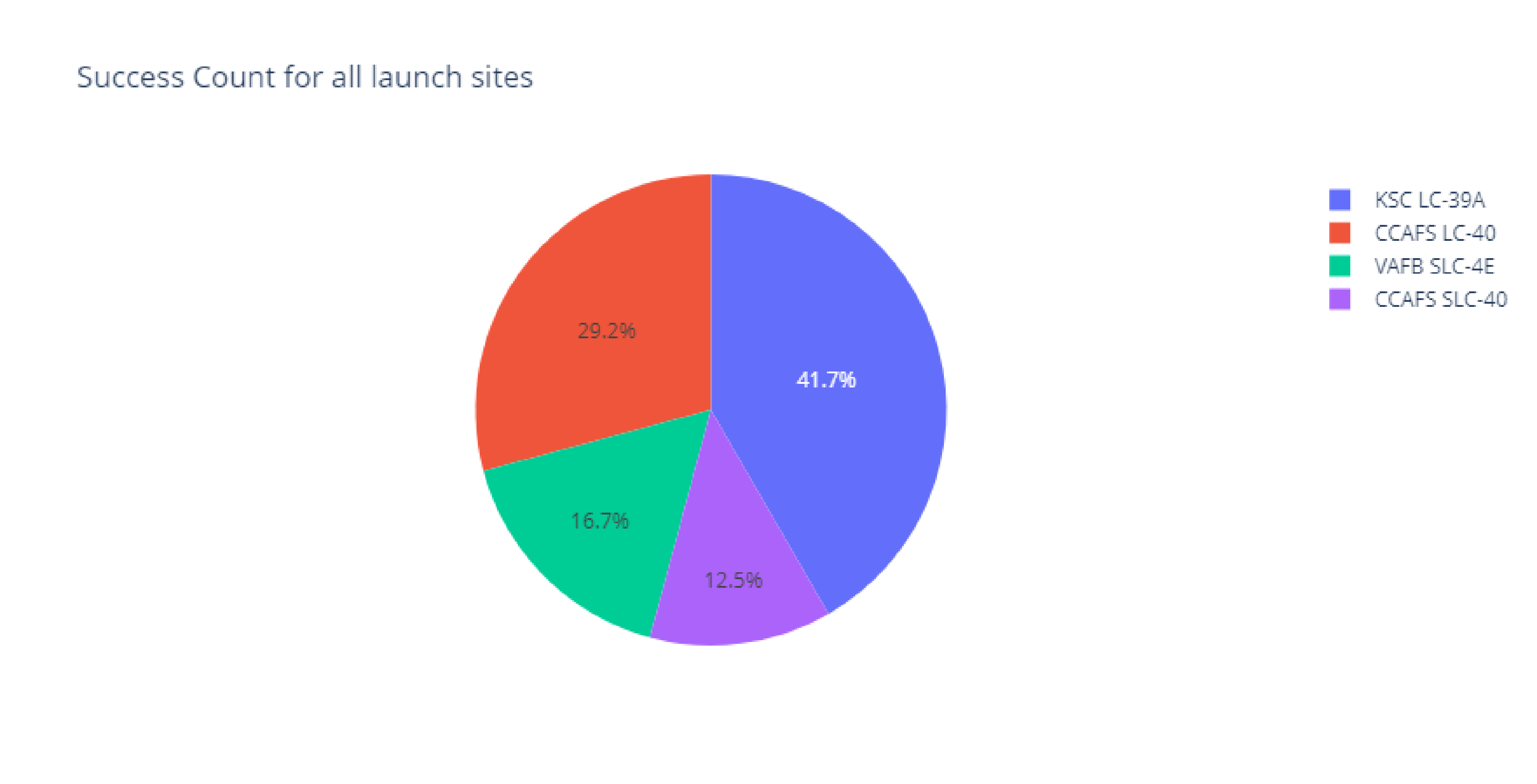


Section 4

Build a Dashboard with Plotly Dash

Dashboard chart for launch success count

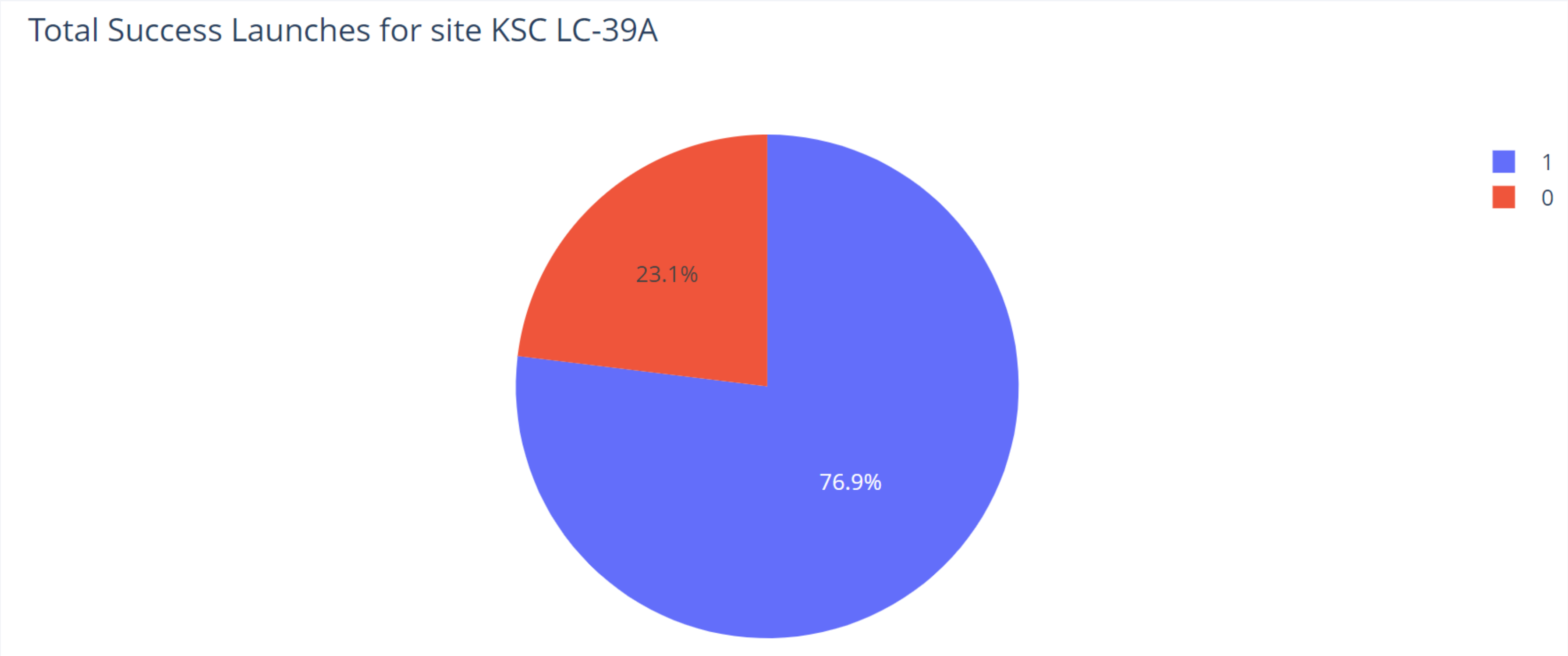
The next pie chart in the Dashboard shows the count for launch success in each launch site.



Dashboard chart for the highest launch success

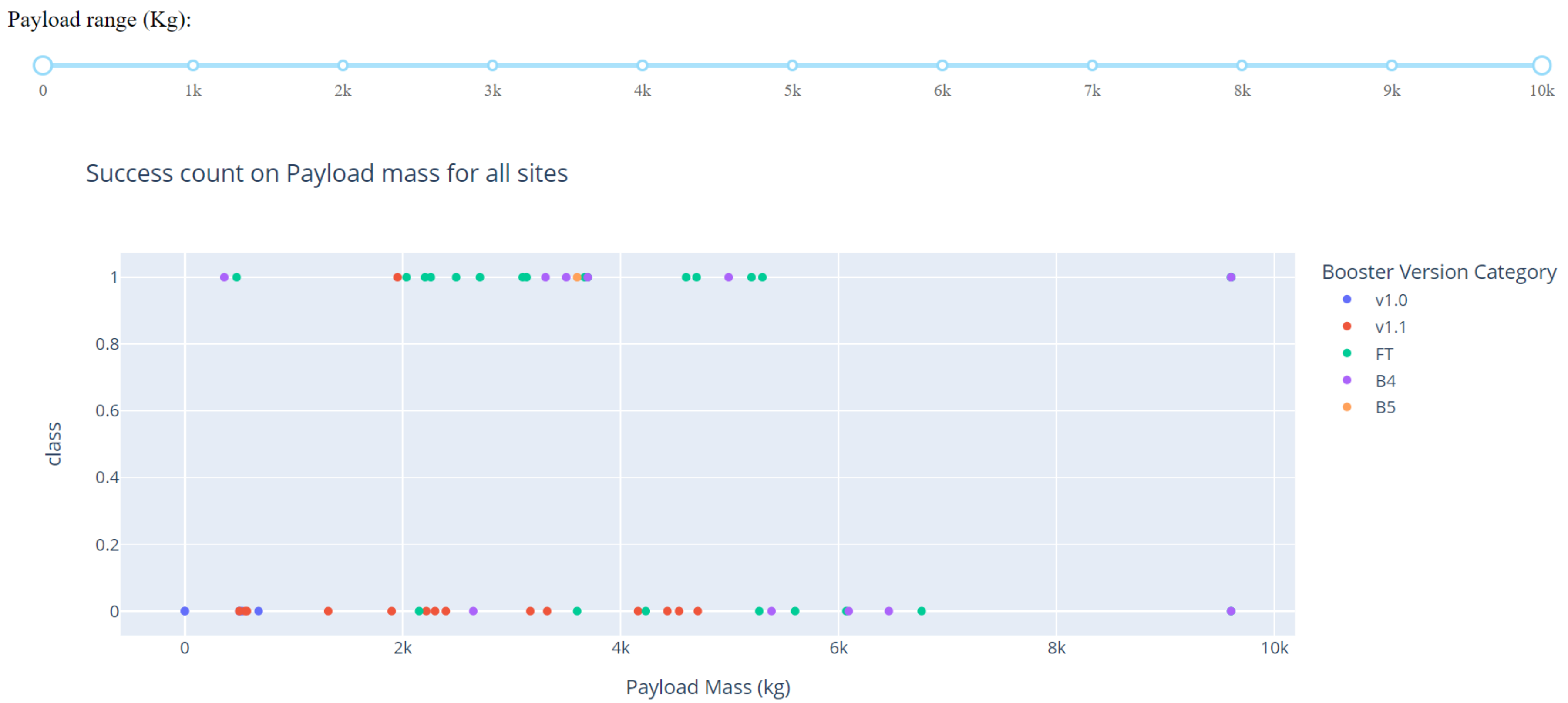
The next pie chart shows the launch ratios for the launch site with the highest success count, KSC LC-39A.

The blue zone represents the success ratio and the red one the failure ratio.



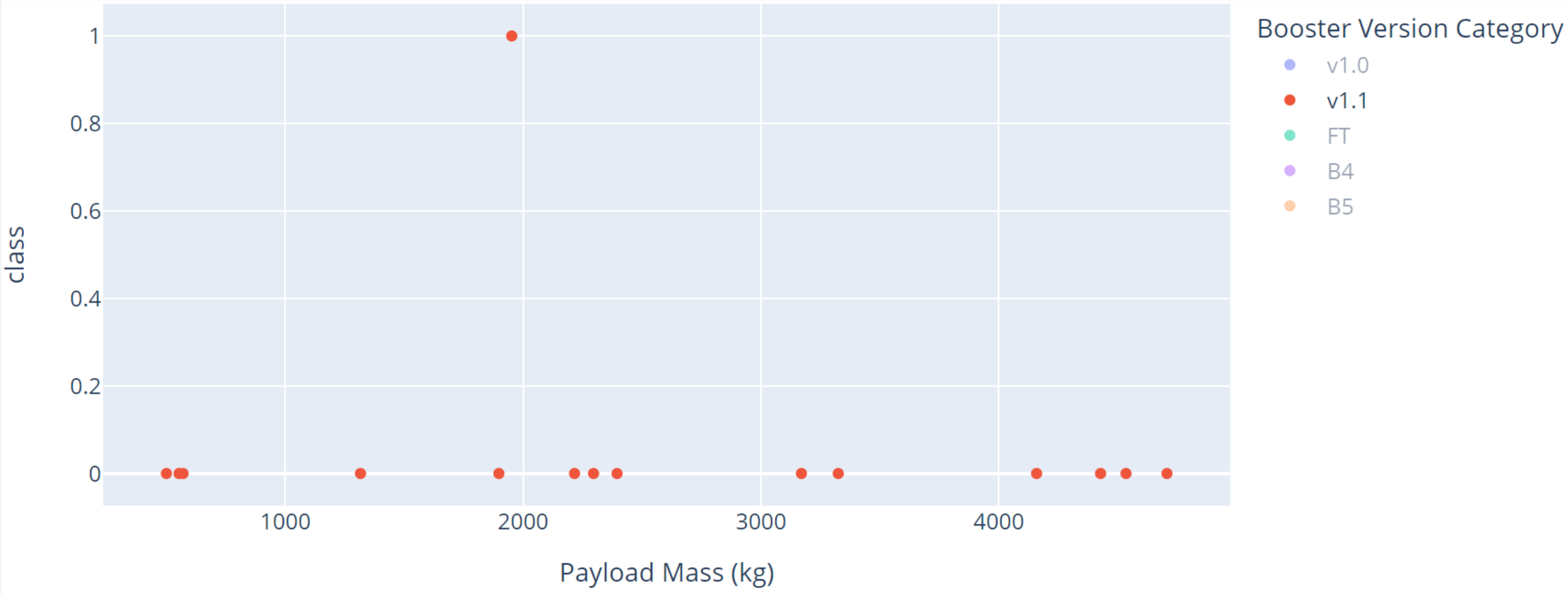
Dashboard of Payload vs. Launch Outcome

The next scatter plots show Payload vs. Launch Outcome for all sites, with different payload selected in a range slider, with different Boosters selected in the legend.



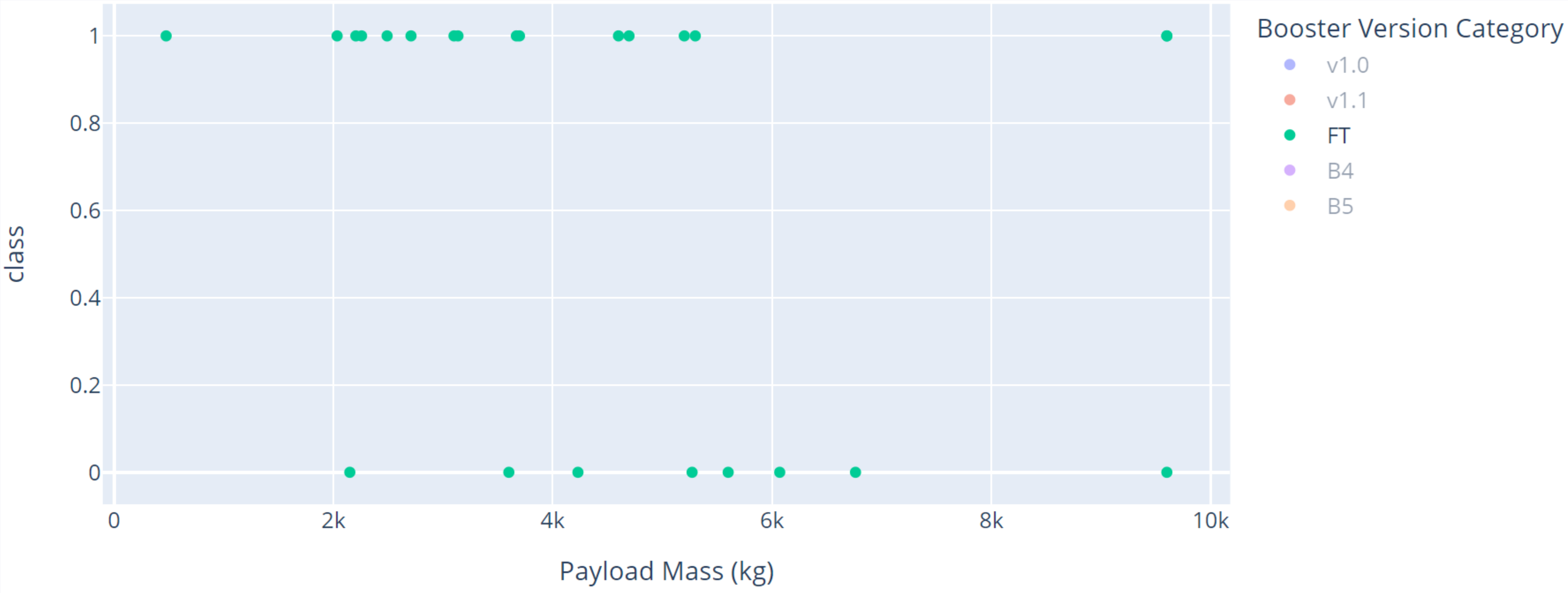
Dashboard of Payload vs. Launch Outcome – Analysis 1

The selection of only Booster version 1.1. in the scatter plot shows that it has the worst success rate.



Dashboard of Payload vs. Launch Outcome – Analysis 2

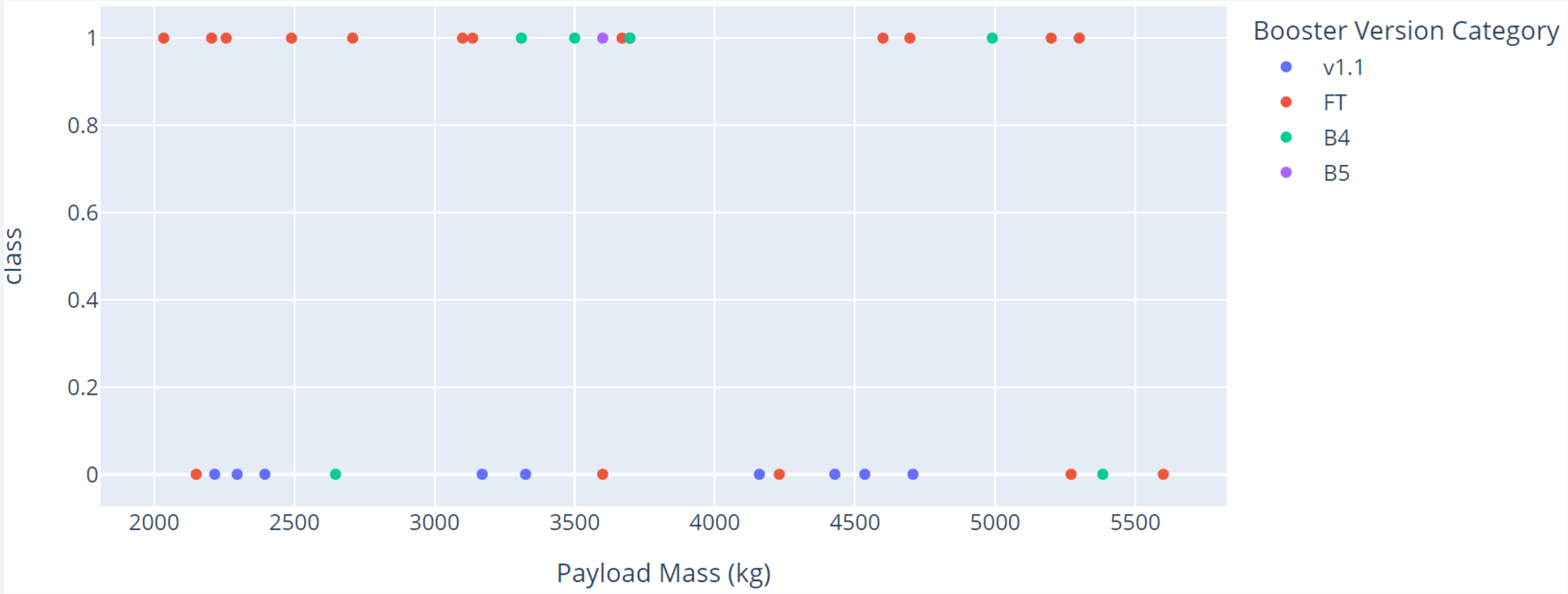
The selection of only Booster version FT in the scatter plot shows that it has the largest success rate. Booster B5 has only one datapoint so it is not reliable.



Dashboard of Payload vs. Launch Outcome – Analysis 3

Changing the Payload range with a slider, the scatter plot shows that success launches concentrate in the Payload range 2kg-6kg and only in four of the five Boosters.

FT has the best success ratio in this payload range.

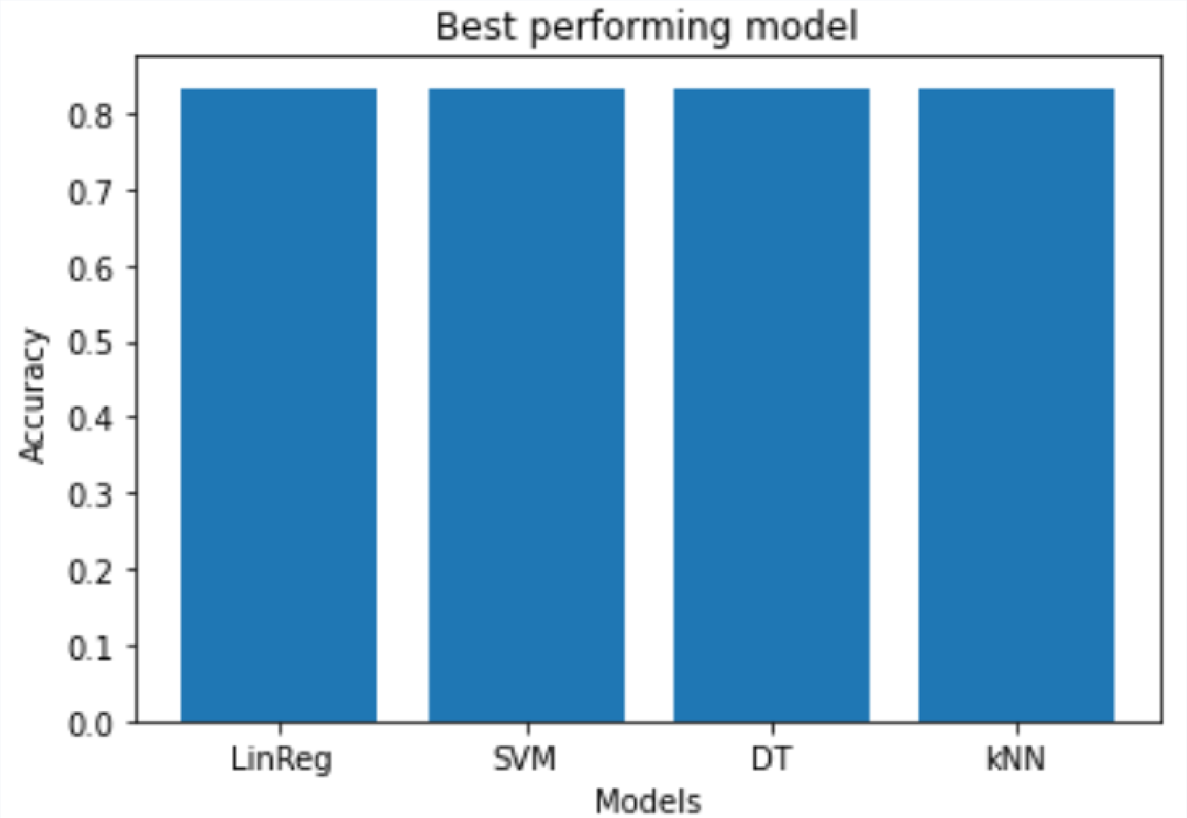


Section 5

Predictive Analysis (Classification)

Classification Accuracy

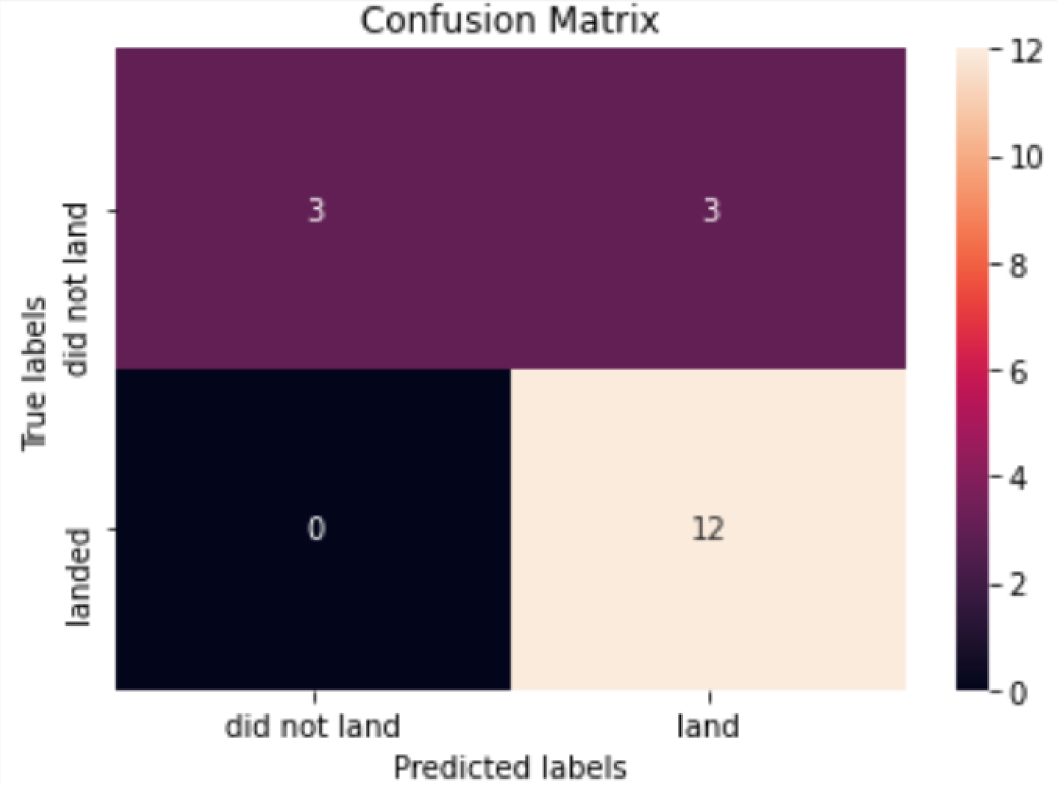
- The bar chart shows that the best performance was the same for all the models, an accuracy of 0,83.



Confusion Matrix for the best performing model

The confusion matrix shows that the model:

- Predicts correctly that twelve first stage will land.
- Wrongly predicts that three failed launches would launch successfully.
- Has a precision of 0,8 (12/15).



Conclusions

- There is an increase trend through the years in the successful rate of launches.
- The FT Booster has the largest success ratio.
- The prediction of success launches can be improved including other variables, varying the hyperparameters and/or fitting other algorithms like neural networks.
- Algorithmic approach is a valid methodology to estimate the success of the launch first stage.

Appendix

These are the versions of Python and some of the most important libraries used in the project:

- Python 3.8.8
- iPython 7.22.0
- Pandas 1.2.4
- Matplotlib 3.6.2
- Numpy 1.23.4
- Dash 2.7.1
- Folium 0.14.0
- Scikit-learn 1.1.3
- Seaborn 0.11.1
- Wget 3.2
- Scipy 1.8.1
- Pip 22.3.1
- tensorflow_intel 2.11.0
- Ipympl 0.9.2

Thank you!