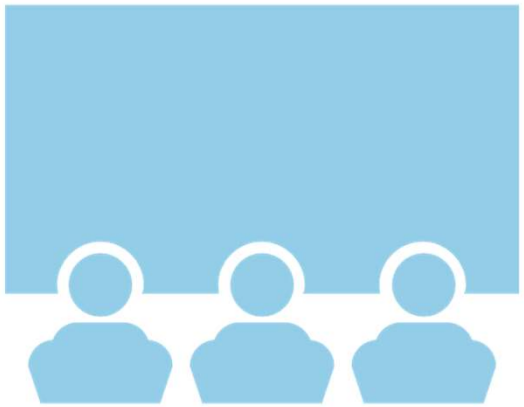
Data Science Capstone project
Predict if the SpaceX's Falcon 9 first stage will land successfully

<Marco D'Amato>

<02/09/2021>

Outline

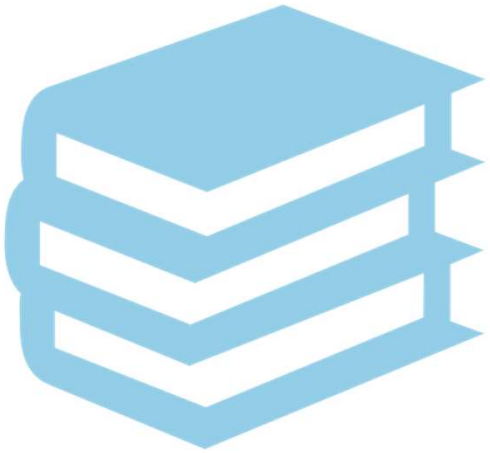


- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

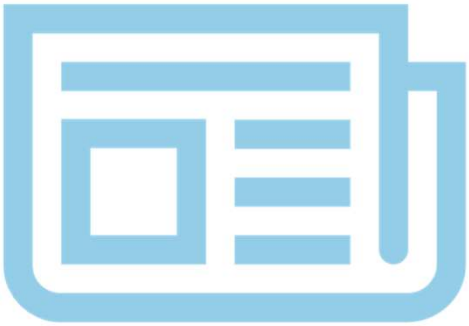
- Summary of methodologies
- Summary of all results

Introduction



- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - This project is focused on determining if the first stage of the Falcon 9 will land successfully

Methodology



- Data collection methodology:
  - Data were collected by using SpaceX REST API (api.spacexdata.com/v4/launches/past)
- Perform data wrangling
  - Using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records
  - Parsing the data from tables and convert them into a Pandas data frame for further visualization and analysis.
  - Transforming raw data into a clean dataset which provides meaningful data
  - Wrangling Data using an API
  - Sampling Data and Dealing with Nulls
  - Converting landing outcomes into a class (0-1)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Methodology

Data collection

Perform data wrangling

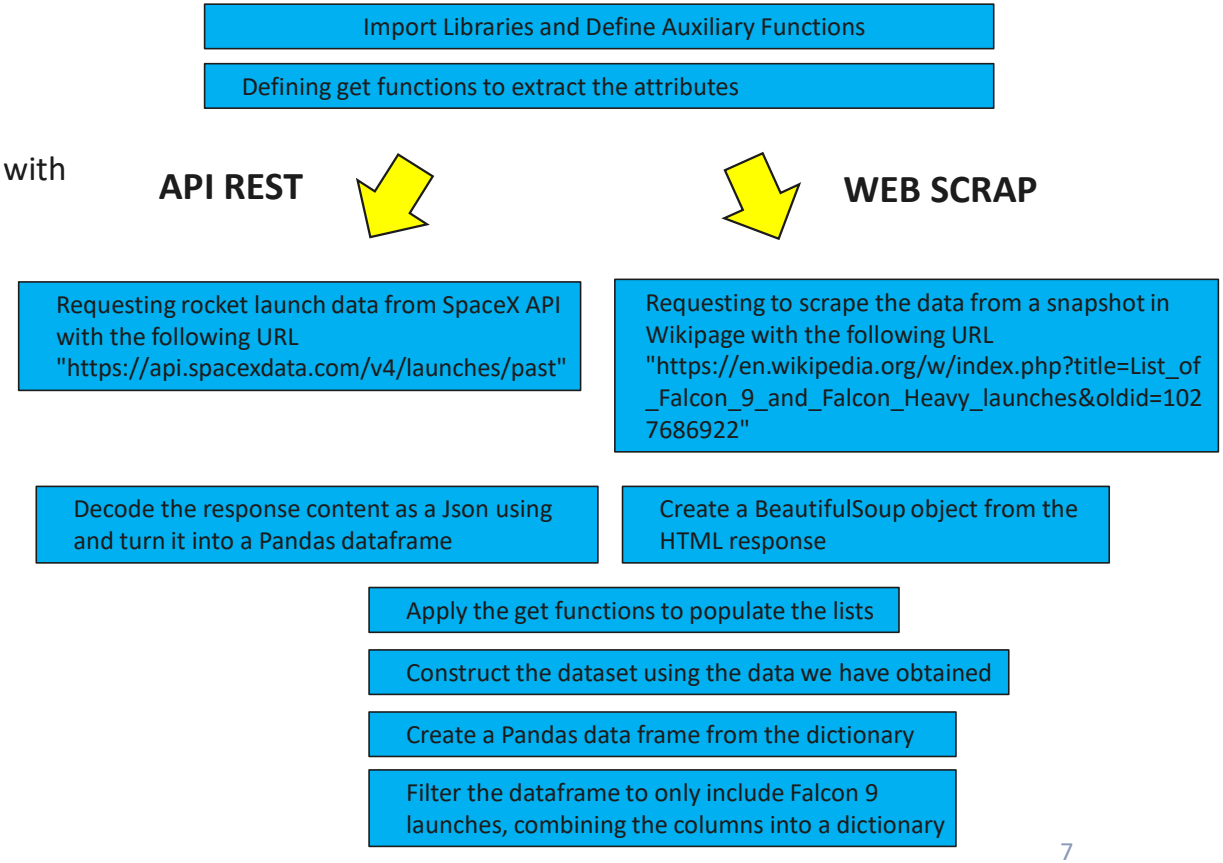
Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data collection

- Data were collected in two ways:
  - by using SpaceX REST API (api.spacexdata.com/v4/launches/past)
  - Through web scraping Falcon 9 launch records with BeautifulSoup

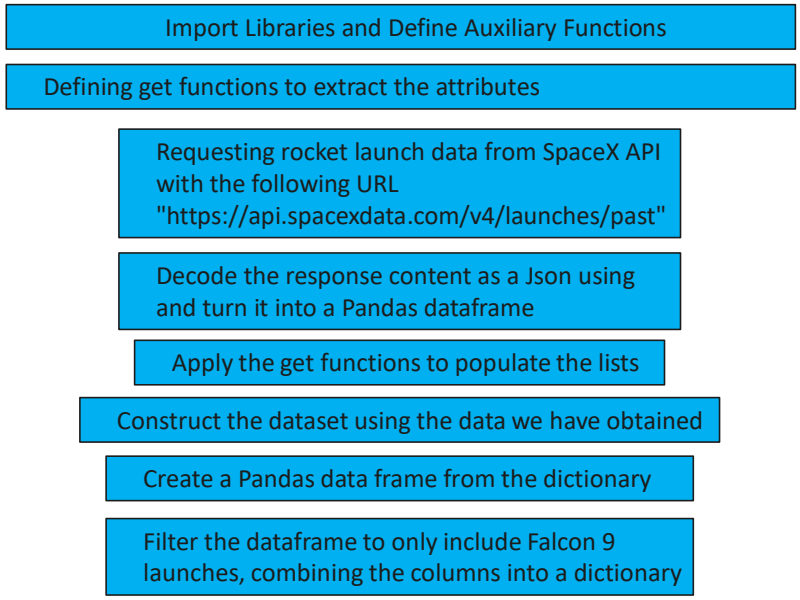


Data collection – SpaceX API

GitHub URL of the completed SpaceX API calls notebook:

https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Added a flowchart of SpaceX API calls here

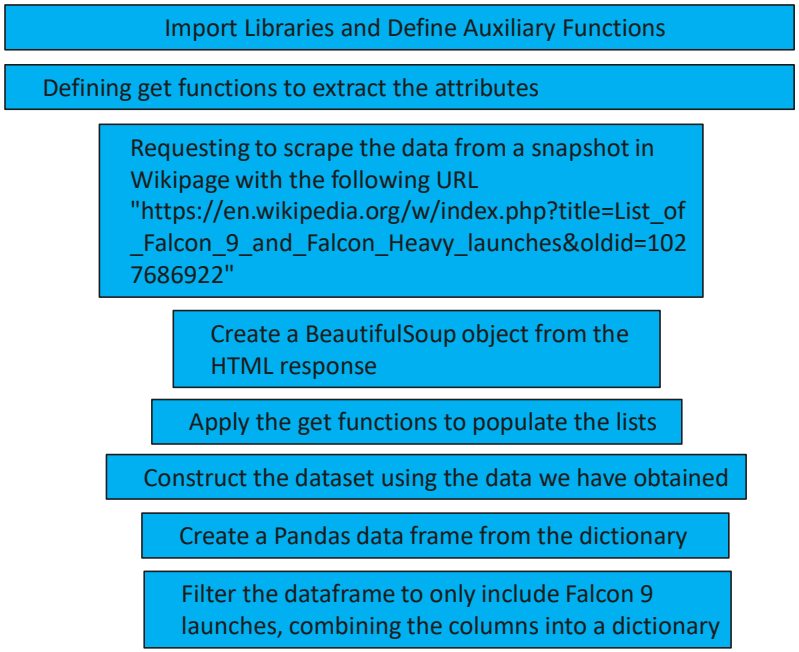


Data collection – Web scraping

GitHub URL of the completed SpaceX API calls notebook:

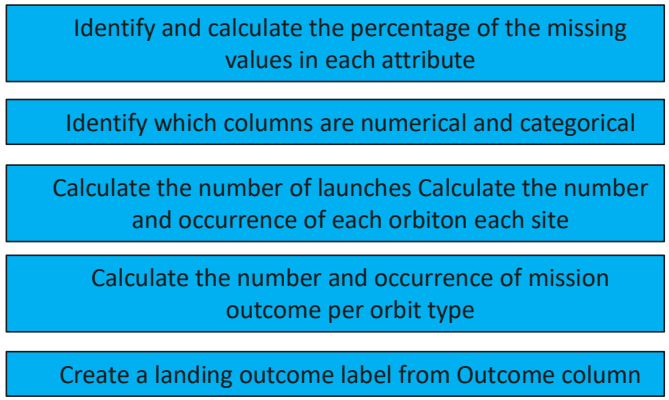
https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Add a flowchart of web scraping here



Data wrangling

- Data were processed by loading Space X dataset from CSV ("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv")
- Sampling Data and Dealing with Nulls
- Converting landing outcomes into a class (0-1)



GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

EDA with data visualization

In order to understand the most relevant attributes the Preparing Data Feature Engineering is performed. This is done by plotting charts to understand which are the most relevant features.

The carts that were plotted are:

- scatter plot of Flight Number vs. Launch Site
- scatter plot of Payload vs. Launch Site
- bar chart for the success rate of each orbit type
- scatter point of Flight number vs. Orbit type
- scatter point of payload vs. orbit type
- line chart of yearly average success rate

GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

EDA with SQL

List of the SQL queries:

- the names of the unique launch sites
- all launch sites begin with `CCA`
- the total payload carried by boosters from NASA
- the average payload mass carried by booster version F9 v1.1
- the date when the first successful landing outcome in ground pad
- the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- the total number of successful and failure mission outcomes
- the names of the booster which have carried the maximum payload mass
- the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Build an interactive map with Folium

This is the list of map object created with folium. Those were chosen in order to see the location of the launch sites in the Country, together with their rate of success, including their proximity to important sites like railway. This could help identifying a relation between success rate and geographical position.

- all launch sites' location markers on a global map
- color-labeled launch records on the map
- selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Build a Dashboard with Plotly Dash

This is the list of plots/graphs and interactions I have added to a dashboard

- Show the screenshot of launch success count for all sites, in a pie chart
- Show the screenshot of the pie chart for the launch site with highest launch success ratio
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Predictive analysis (Classification)

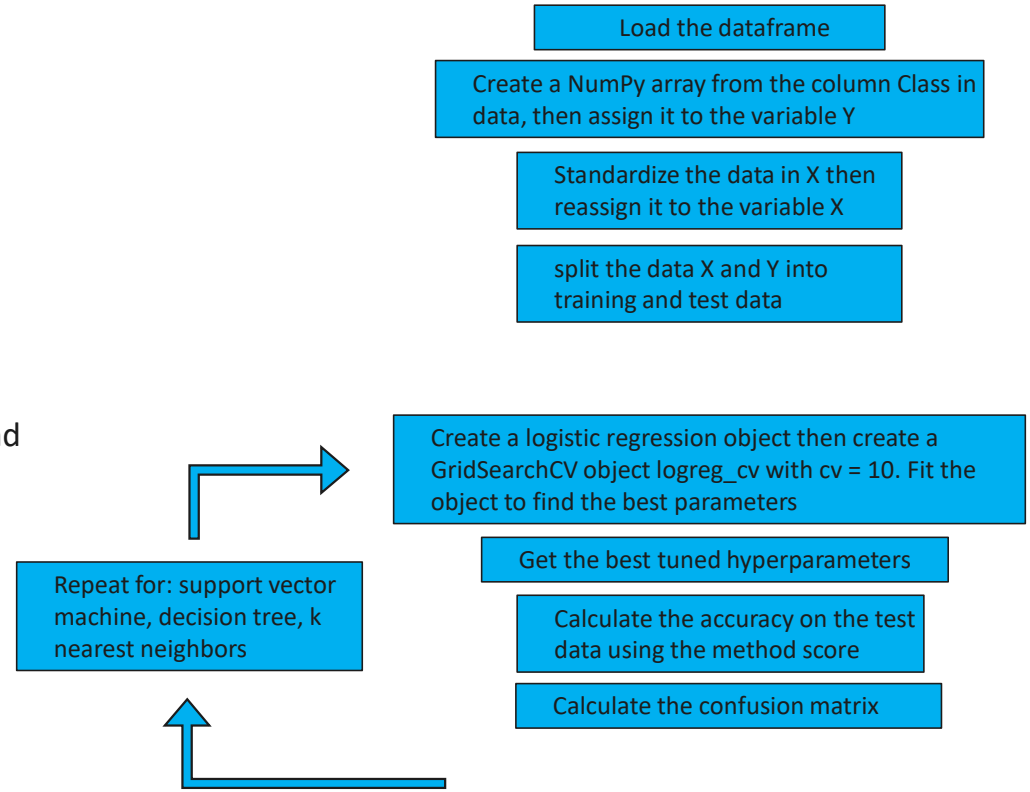
In order to build the model, it is necessary to standardize the data to avoid biased model, then split the data into train and test.

Once the model has been chosen, the optimization of the hyperparameters of the model needs to be done to calibrate the model to best fit the data training.

Interrogating the model by using the test data shall reveal the model outcome.

The calculation of the confusion matrix and accuracy shall reveal the over all model performances

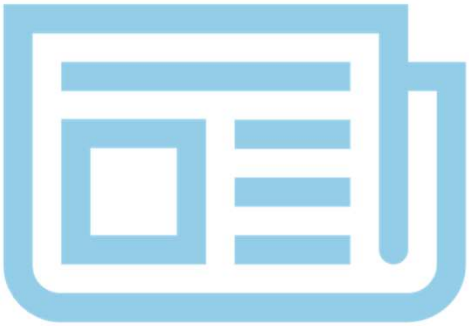
Various model could be selected and evaluated in order to understand which one is the most appropriate for the data



GitHub URL of the completed SpaceX API calls notebook:

- https://github.com/ukitcode/Applied_Data_Science_Capstone.git

Results



- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

scatter plot of Flight Number vs. Launch Site

scatter plot of Payload vs. Launch Site

bar chart for the success rate of each orbit type

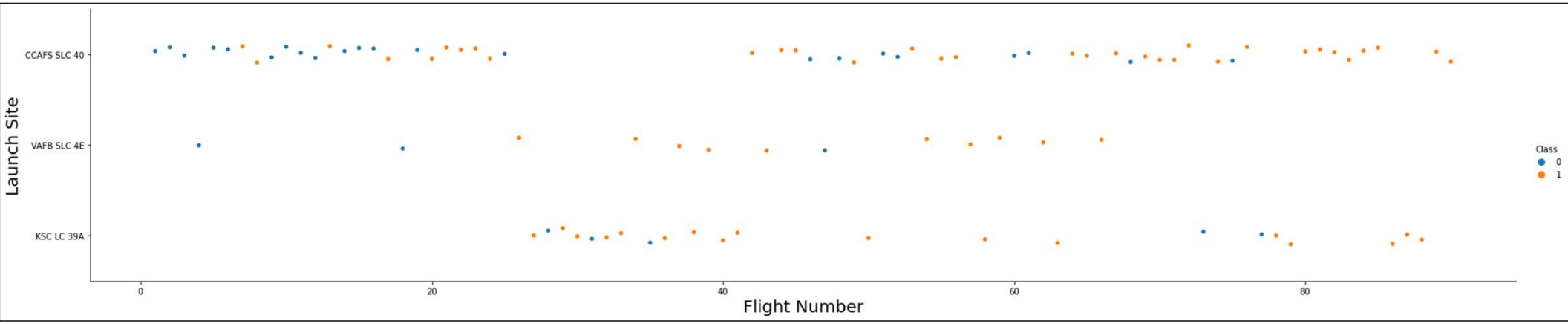
scatter point of Flight number vs. Orbit type

scatter point of payload vs. orbit type

line chart of yearly average success rate

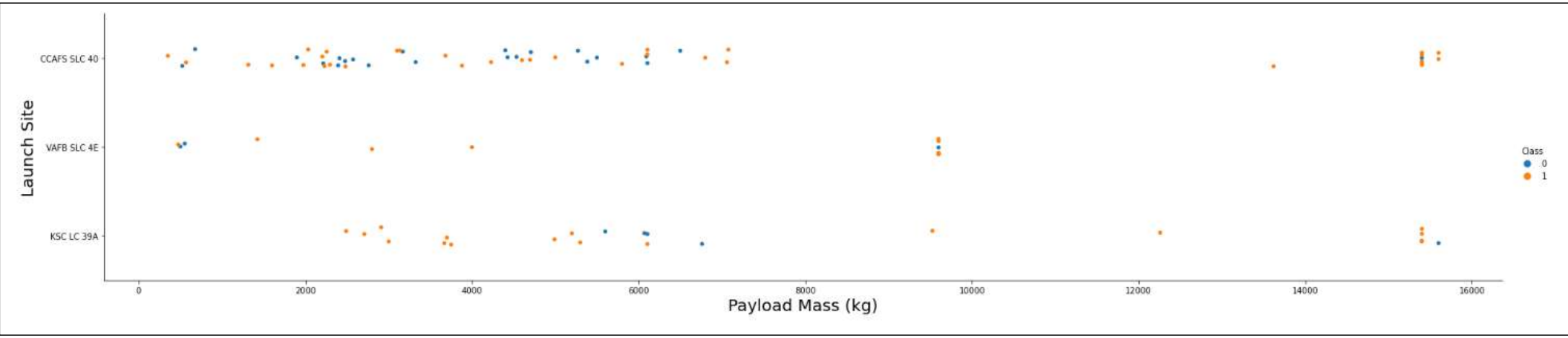
Flight Number vs. Launch Site

Every launch site seems to have an increase in success when flight number increases. However, the site named KSC LC-39A appears to be the one with higher failure rate.



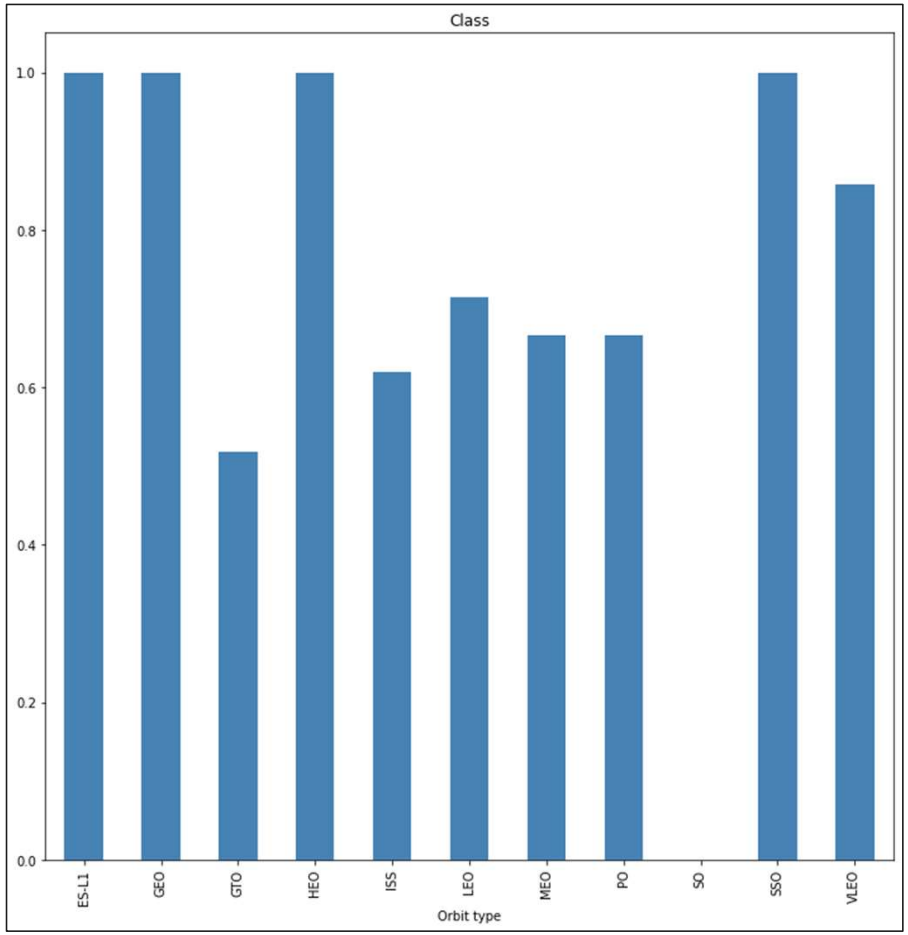
Payload vs. Launch Site

Most of the launches are with payload weight below 7000kg It appears that, in all the launch sites, the success rate decreases with the increase of payload mass, certainly between 0 and 7000kg, and it seems to stabilize above 7000kg



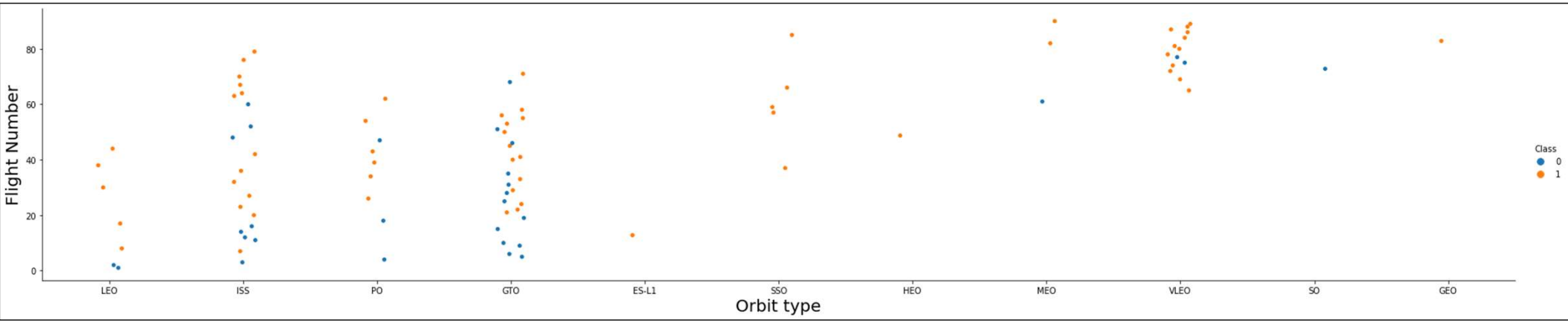
Success rate vs. Orbit type

The orbits with the highest success rate are: ES-L1, GEO, HEO, SSO with 1 and VLEOF with around 0.85



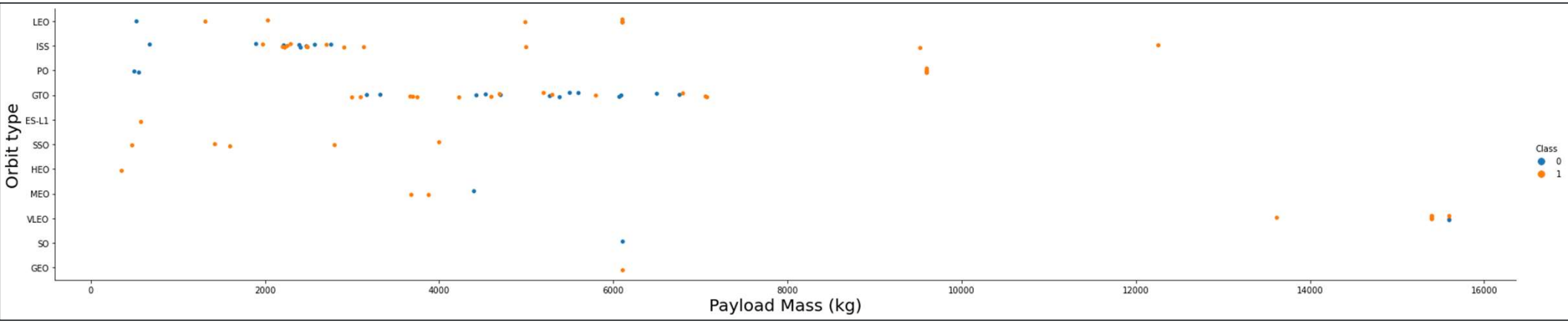
Flight Number vs. Orbit type

The LEO orbit success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



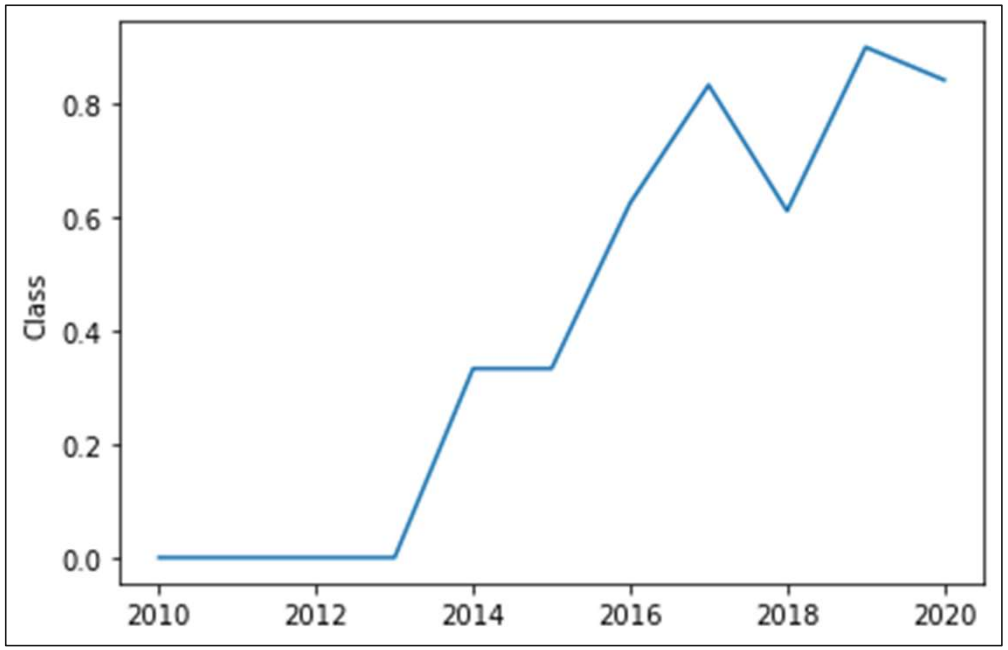
Payload vs. Orbit type

Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch success yearly trend

The success rate since 2013 kept increasing till 2020



EDA with SQL

the names of the unique launch sites

all launch sites begin with `CCA`

the total payload carried by boosters from NASA

the average payload mass carried by booster version F9 v1.1

the date when the first successful landing outcome in ground pad

the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

the total number of successful and failure mission outcomes

the names of the booster which have carried the maximum payload mass

the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

All launch site names

The names of the unique launch sites in the space mission are:

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Launch site names begin with `CCA`

The names of the unique launch records where launch sites begin with the string 'CCA’ are:

CCAFS LC-40

CCAFS SLC-40

Total payload mass

The total payload carried by boosters from NASA (CRS):

45596

Average payload mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1:

2534

First successful ground landing date

The date when the first successful landing outcome in ground pad

2015-12-22

Successful drone ship landing with payload between 4000 and 6000

The names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total number of successful and failure mission outcomes

The total number of successful mission outcomes:

61

The total number of failed mission outcomes:

10

Boosters carried maximum **payload**

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

F9 FT B1029.1
F9 FT B1034
F9 FT B1036.1
F9 FT B1037
F9 B4 B1041.1
F9 FT B1036.2
F9 B4 B1044
F9 B4 B1041.2
F9 B4 B1043.2
F9 B5B1047.1
F9 B5B1048.1
F9 B5B1049.1
F9 B5 B1049.2
F9 B5B1051.1
F9 B5 B1049.3
F9 B5 B1047.3
F9 B5 B1048.4
F9 B5 B1056.3
F9 B5 B1049.4
F9 B5 B1046.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5B1058.1
F9 B5 B1049.5
F9 B5 B1059.3
F9 B5 B1051.5
F9 B5 B1049.6
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5B1061.1
F9 B5 B1049.7

2015 launch records

The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

The count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

landing__outcome	total
Success (drone ship)	5
Success (ground pad)	3

Interactive map with Folium

all launch sites’ location markers on a global map

color-labeled launch records on the map

selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

Folium map: all launch sites' location markers on a global map

Most of the launch sites are located on the USA east coast (tree of them), and only one is located on the west coast.

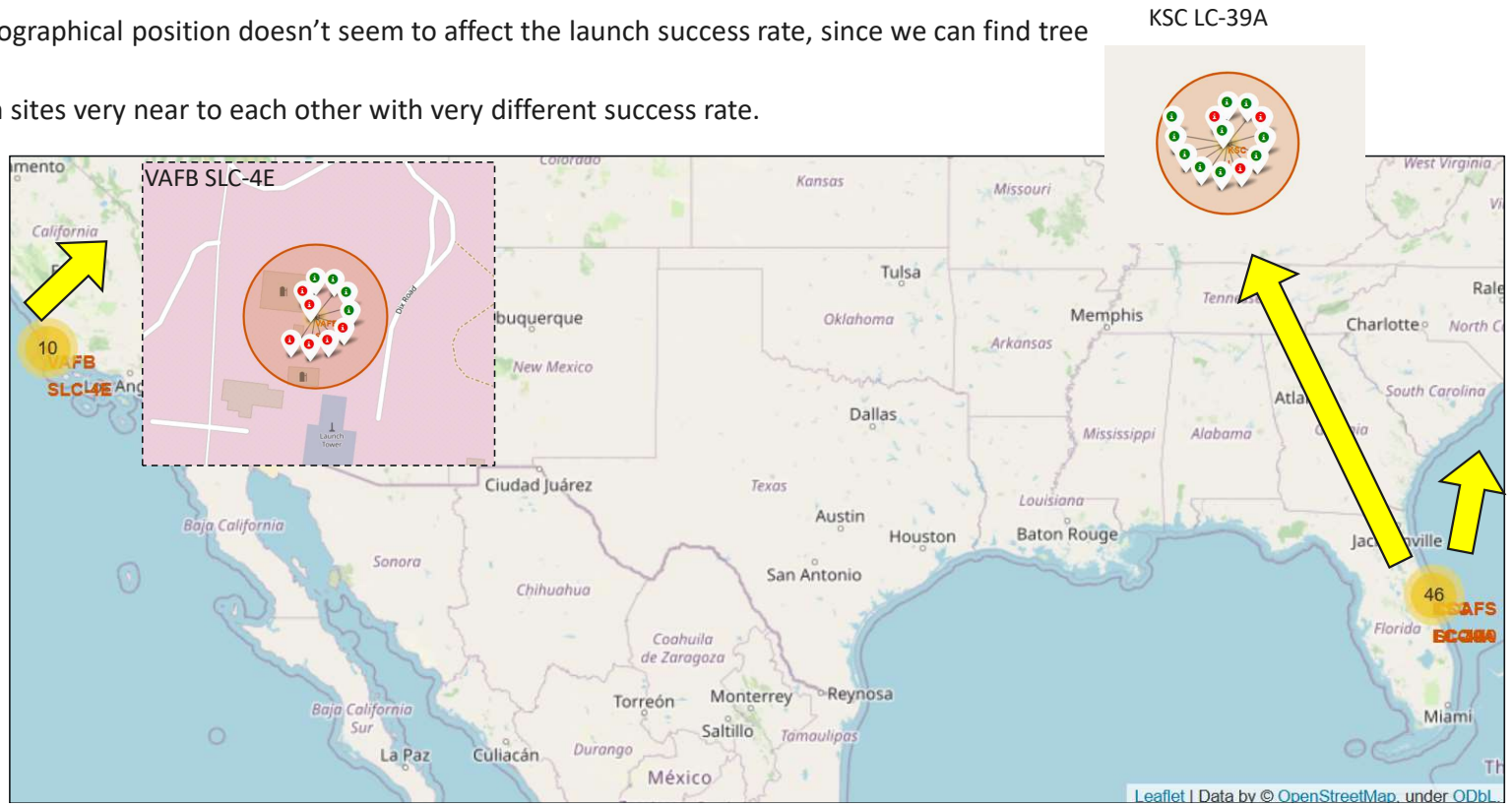
Anyhow, all of them are in the proximity of the coastline

Folium map: color-labeled launch records on the map

CCAFS LC-40 launch site appears to be the one where most of the launches were made, and also the site with the lowest success rate.

CCAFS SLC-40 launch site appears to be the one with the least number of launches but with the highest success rate

All in all, the geographical position doesn't seem to affect the launch success rate, since we can find tree different launch sites very near to each other with very different success rate.



Folium map: selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

VAFB SLC-4E launch site appears to be very near to the railway, approximately 1.25km

Looking at the other launch sites it can be observed that:

- Launch sites are in close proximity to railways
- Launch sites are in close proximity to highway
- Launch sites are in close proximity to coastline
- Launch sites keep some distance from cities, but not very much

Build a Dashboard with Plotly Dash

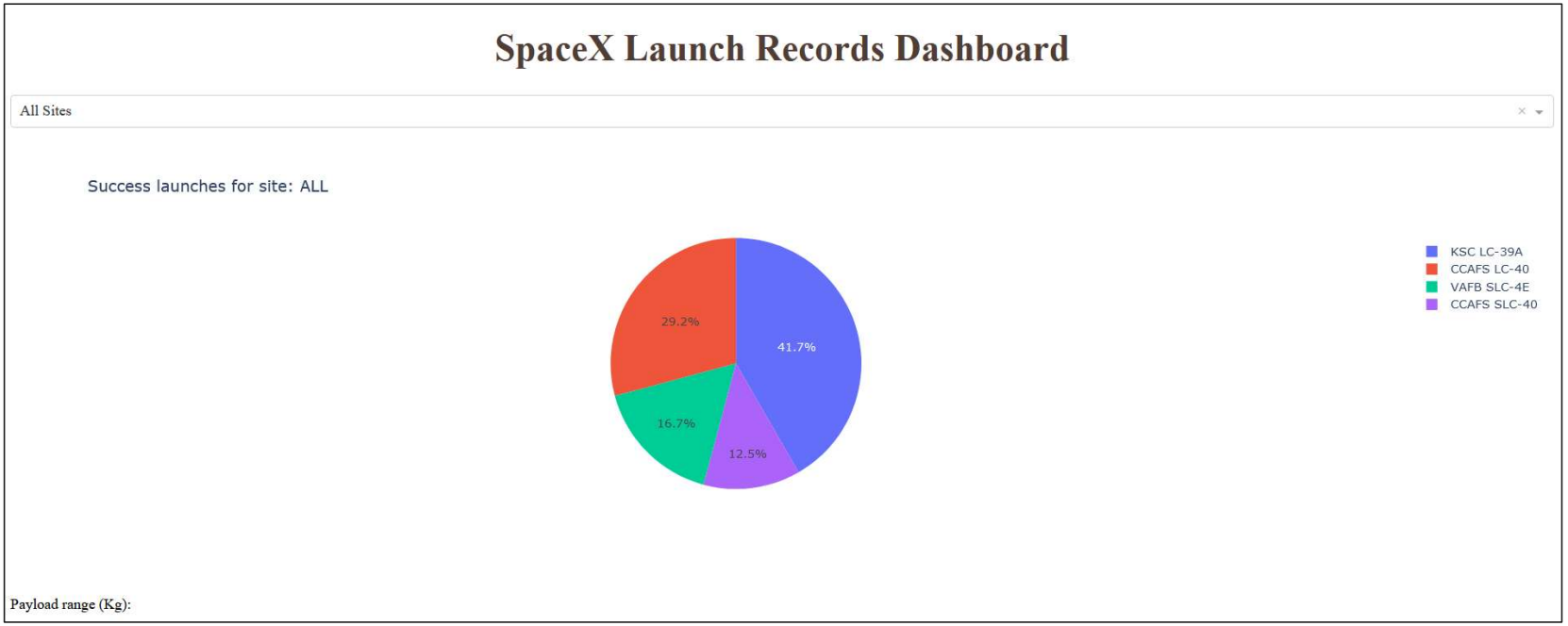
Show the screenshot of launch success count for all sites, in a pie chart

Show the screenshot of the pie chart for the launch site with highest launch success ratio

Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

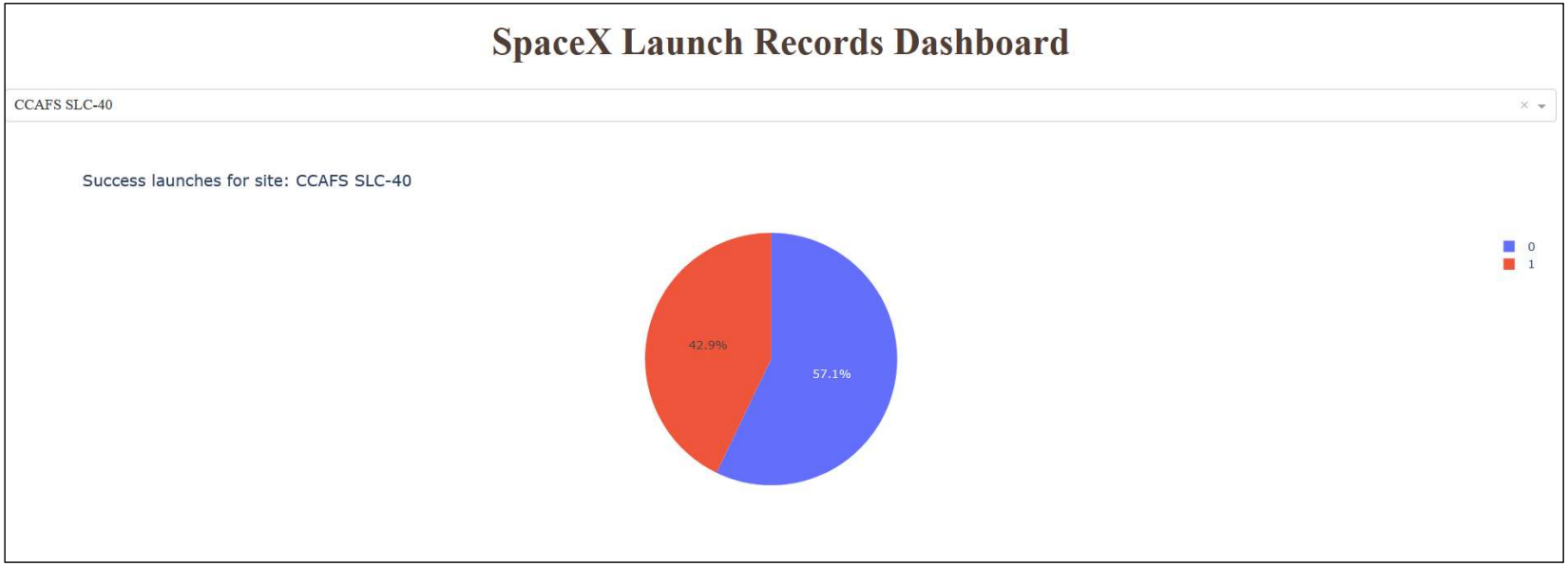
Dashboard: launch success count for all sites in a pie chart

Explain the important elements and findings on the screenshot



Dashboard: the pie chart for the launch site with highest launch success ratio

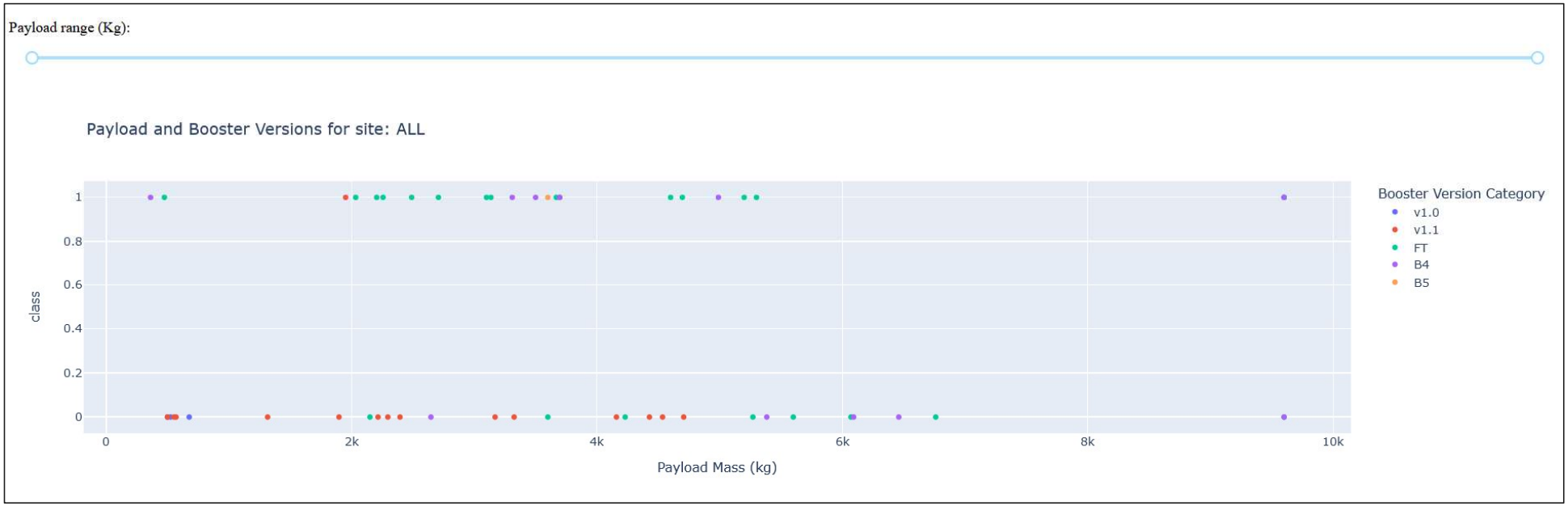
Explain the important elements and findings on the screenshot



Dashboard: Payload vs. Launch Outcome scatter plot for all sites with different payload selected in the range slider

This is the Dashboard Payload vs. Launch Outcome scatter plot for all sites with the entire range of payload selected in the range slider.

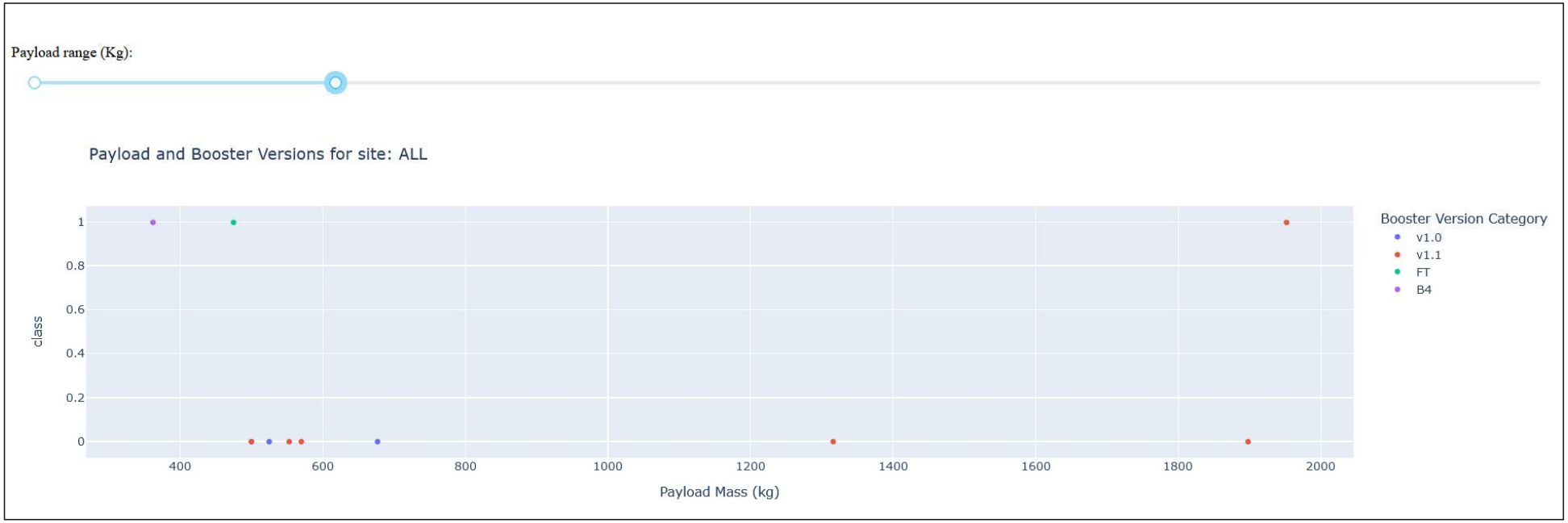
Looking at the entire payload range, it is observable that there is a relation between payload and success rate, hence grouping the data per payload group could reveal some insights



Dashboard: Payload vs. Launch Outcome scatter plot for all sites with different payload selected in the range slider

This is the Dashboard Payload vs. Launch Outcome scatter plot for all sites with the lowest range of payload

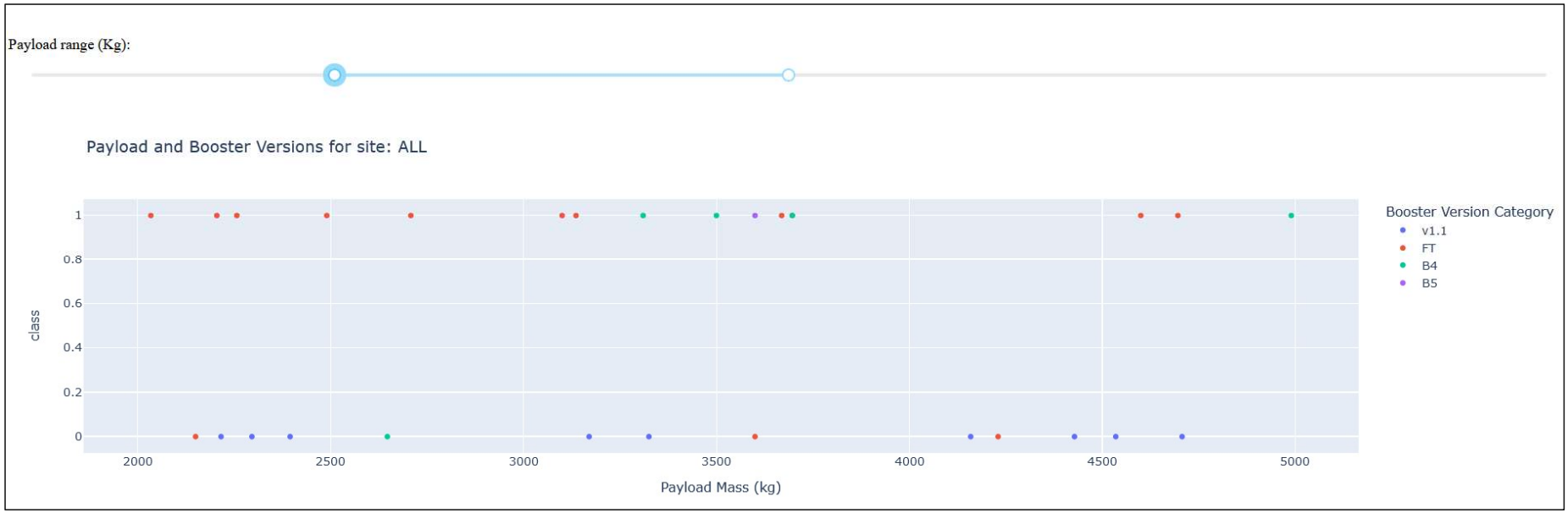
At this payload range it is observable a poor launch success rate.



Dashboard: Payload vs. Launch Outcome scatter plot for all sites with different payload selected in the range slider

This is the Dashboard Payload vs. Launch Outcome scatter plot for all sites with the medium range of payload.

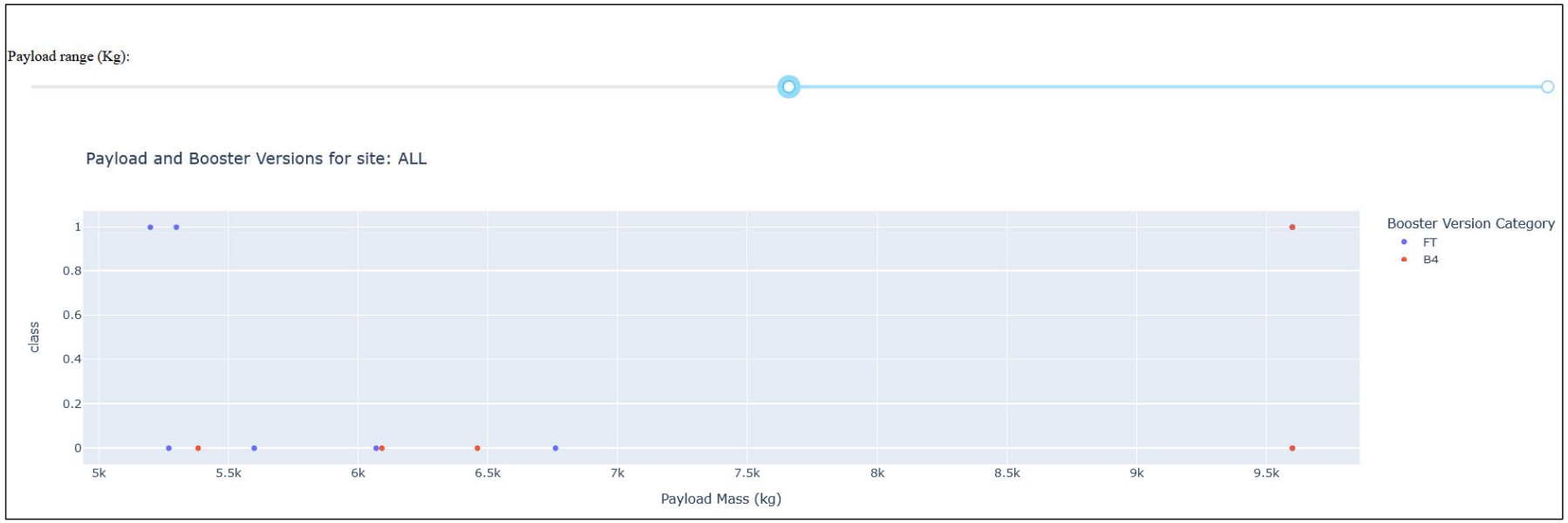
At this payload range it is observable a very good launch success rate.



Dashboard: Payload vs. Launch Outcome scatter plot for all sites with different payload selected in the range slider

This is the Dashboard Payload vs. Launch Outcome scatter plot for all sites with the highest range of payload.

At this payload range it is observable a clear detriment of the launch success rate.



Predictive analysis (Classification)

Logistic Regression

Support Vector Machine

Decision tree

K nearest neighbors

Classification Accuracy

The models under analysis:

Logistic Regression

Supp. Vector Mach

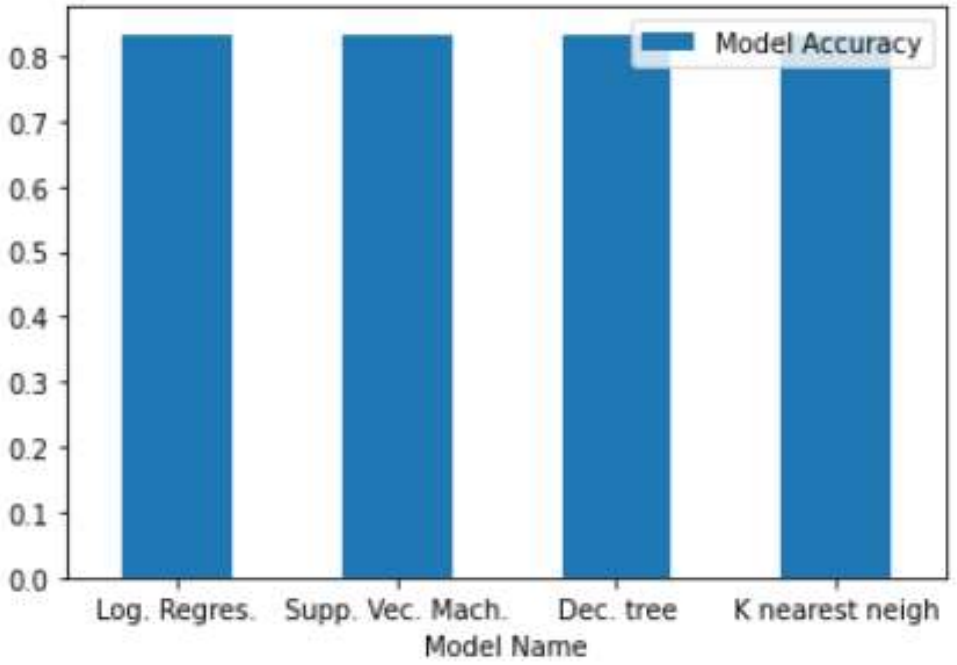
Decision tree

K nearest neighbors

Have shown the same test accuracy of

0.8333333333333334

Hence, those models have the same performance in predicting the landing outcome



It appear that all models are good to predict the most likely landing outcome and hence the cost of the launch.

Confusion Matrix

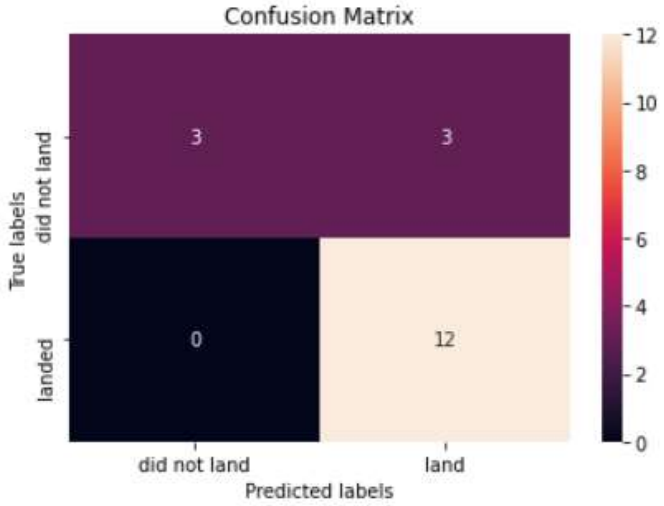
The models under analysis:

Logistic Regression

Supp. Vector Mach

Decision tree

K nearest neighbors



Have shown the same confusion matrix results.

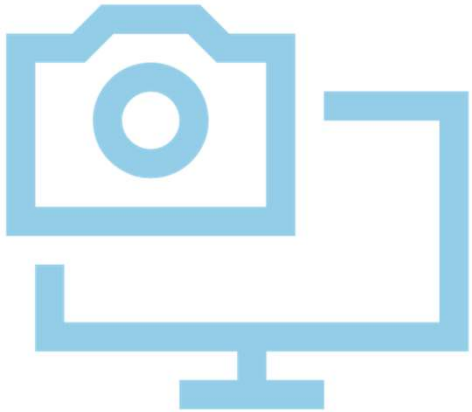
Hence, those models have the same performance in predicting the landing outcome

It appear that all models are good to predict the most likely landing outcome and hence the cost of the launch.

CONCLUSION

- the success rate decreases with the increase of payload mass, especially at very high payload
- Every launch site seems to have an increase in success when flight number increases
- The LEO orbit success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits
- Launch sites are in close proximity to railways
- Launch sites are in close proximity to highway
- Launch sites are in close proximity to coastline
- Launch sites keep some distance from cities, but not very much
- All in all, the geographical position doesn't seem to affect the launch success rate, since we can find tree different launch sites very near to each other with very different success rate.
- The launch site with highest launch success ratio is CCAFS SLC-40
- The orbits with the highest success rate are: ES-L1, GEO, HEO, SSO with 1 and VLEOF with around 0.85
- Every classification model under analysis (Logistic Regression, Supp. Vector Mach, Decision tree and K nearest neighbors) have shown the same accuracy and confusion matrix results which it seems to be so high to be a good way of predicting the most likely landing outcome and hence the cost of the launch.

APPENDIX



- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project