IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SAYAN ROY
21/04/2024

Table of Contents

Executive Summary

- All the data used in the project has been collected from SpaceX API and Wikipedia through API calls and web scraping.
- The collected data is cleaned by filling the missing values, removing unnecessary columns and labelling the outcome as 0 for failure and 1 for success for the first stage landing.
- We perform some EDA with SQL queries and get some insights as the name of each launch site, the total success/failure mission outcomes, ranking the different landing outcomes.
- We also perform EDA with visual tools to find relations between flight number and payload on the success, how a particular launch site has been more successful in first stage, in which orbit type success rates are high and how the success rate has increased since 2013.

Executive Summary

- We then perform some site location analysis through geospatial data and create interactive dashboards to show relations between percentage of successes for each launch site and percentage of success and failure for individual launch sites based on user's choice.
- The dashboard also enables us to find how payload mass for each booster version category affects success rate.
- Predictive analysis is performed on the data using various classification algorithms.
- The SVM classification technique yields best results with about 83% correctly predicted outcomes.

Introduction

- With the emergence of commercial space age, companies are making space travel affordable for everyone.
- Some mention worthy companies who are major manufacturers and producers of sub orbital space flights, reusable rockets and satellite providers are Virgin Galatic, Rocket Lab and Blue Origin, SpaceX.
- The most successful among them is SpaceX. Its major accomplishments are:
  a) Starlink- provider of satellite internet access.
  b) Sending manned missions to space.
  c) Sending spacecrafts to international space stations.
- SpaceX's Falcon 9 rocket launches are relatively inexpensive due to the fact that they can reuse the first stage.

Introduction

- If we can determine whether first stage will land, we can determine the cost of a launch.
- In this project we will try to estimate the cost of a launch, we will gather information from SpaceX and other publicly available resources.
- We will study the correlation of different parameters like payload mass, booster version, number of flights for a spacecraft, the orbits with the outcome of successful landing.
- We will try to predict if the first stage can land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data has been collected from multiple sources like SpaceX API and Wikiepedia through API calls and web scraping.
- Perform data wrangling
  - The missing values in the collected data such as landingpad and payloadmass columns were filled with appropriate values.
  - categorical data was one-hot encoded and the final outcome was labelled as 1 or 0 for success and failure.
  - The final set of features was scaled down before it can be trained with the classification algorithms.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - The correlation between different pairs of features were examined.
  - Various scatter plots and line plots were generated to view trends and get more insight.
  - Lists of top performing launch sites and booster versions were found out.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive maps were created to analyze how a launch site's geographical location can impact a successful launch example: initial trajectory, closeness to equator and to gain more insights from their geographical locations.
  - Interactive dash boards were also created to analyze final outcome with respect to launch sites and how varying payload mass affects the success rate.

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - The cleaned and transformed dataset was split into training and test sets.
  - Four classification algorithms- Logistic Regression, Decision Tree Classification, Support Vector Machine(classifier) and KNN were used to train the data.
  - 10 fold cross validation was performed to validate the accuracy of models and grid search was applied to choose the optimum hyperparameters.

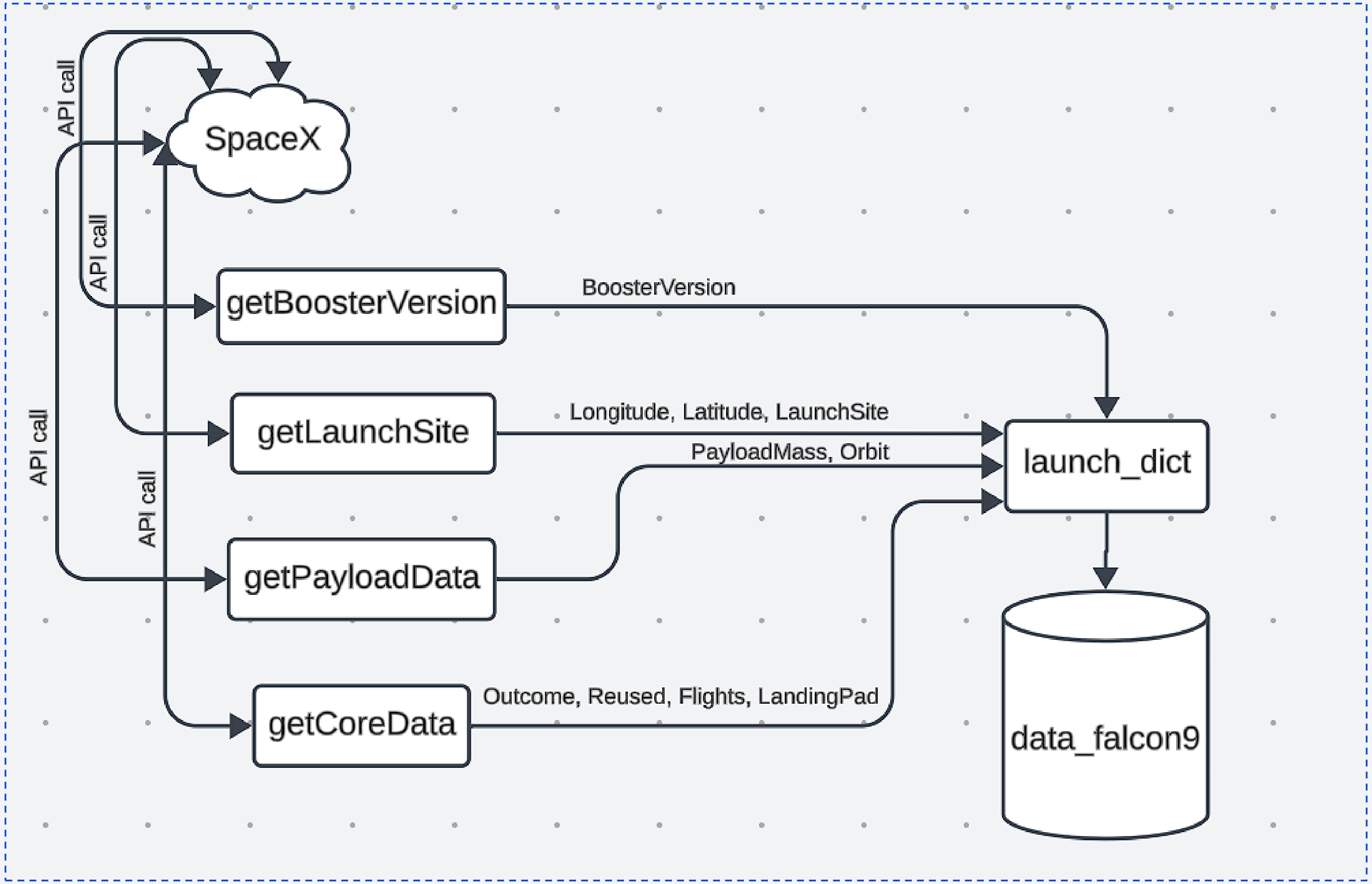
Data Collection

- The data is collected from multiple sources through API calls to SpaceX and data scraping using BeautifulSoup a popular web scraping library for python.
- In case of API calls, we create a bunch of helper function where the API calls are made to get the respective data like orbit, booster version, launch site details from SpaceX API.
- Post that the data gets appended to individual lists for each of the variables like orbit, outcome, payload mass, booster version and so on.
- Once the lists are populated with the respective data, they are loaded in a dataframe which is further used for data wrangling.

Data Collection

- For web scraping we use this static URL: [Wikipedia] and send a get request. We use BeautifulSoup to store the contents from the response in a variable.
- We find the desired table from the list of tables and extract the column names into a variable.
- Then a dictionary is created with those column names.
- We then extract the data for respective columns from the BeautifulSoup object and append it to the dictionary using the respective column name as keys.
- Once all the data is populated, it is loaded in a dataframe.

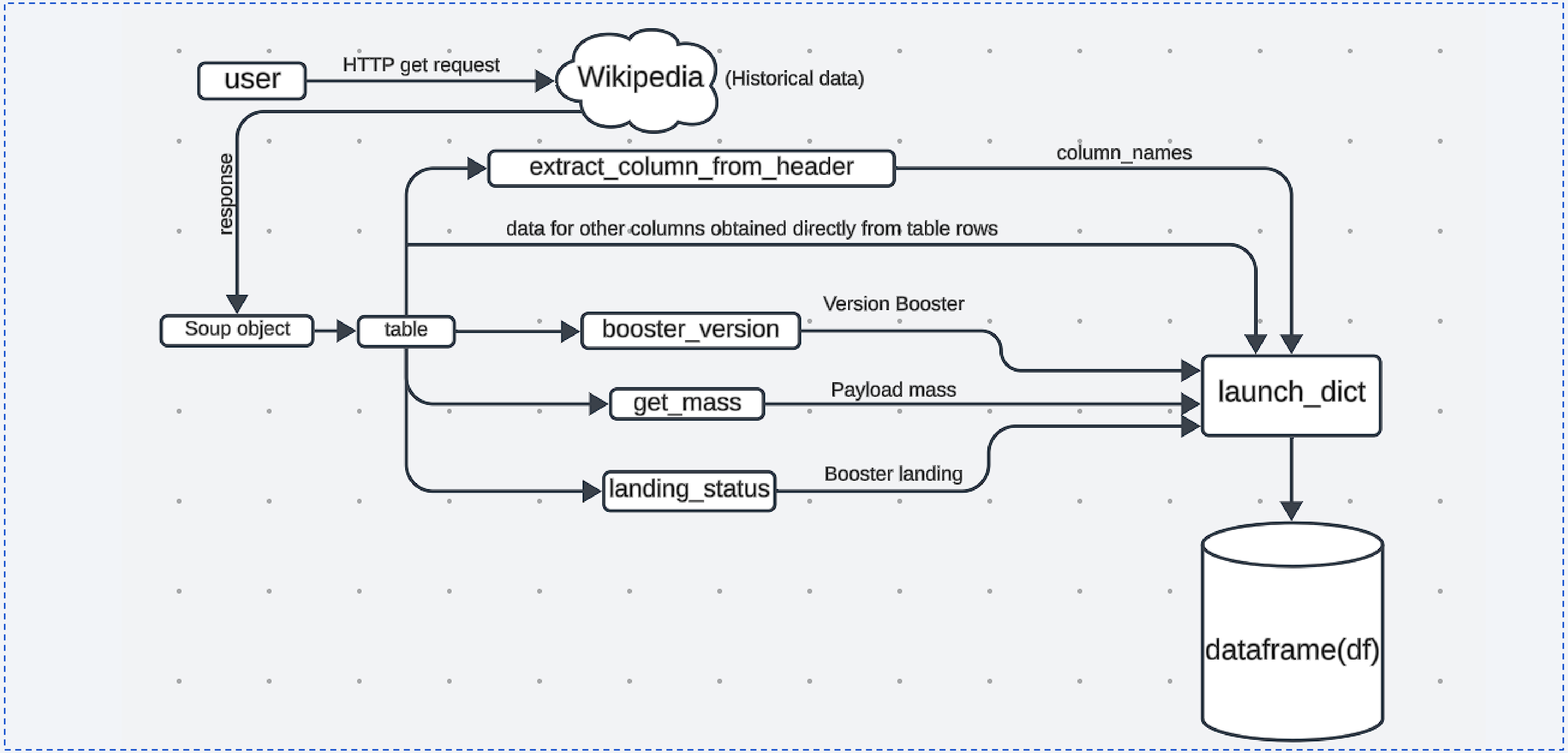
Data Collection – SpaceX API



Data Collection – SpaceX API

- We get our initial data from this static url: link in JSON format and load it to a dataframe.
- We take the relevant columns from here and perform some preprocessing to clean the data since most of the data contains IDs and lists.
- We call the API through helper methods to get specific data like booster name, launch site, mass of payload, outcome, type of landing.
- We will append this data in respective lists created for each column.
- Finally we will combine the lists in a dictionary and create our dataframe.
- The GitHub URL of the completed SpaceX API calls notebook is in this link.

Data Collection - Scraping



Data Collection - Scraping

- Initially we extract historical data from this page by sending a get request and storing the response in a BeautifulSoup object.
- Using the soup object we find the relevant table and extract the column names.
- We create a dictionary using these column names as keys.
- Then we start iterating through each table and extract data from the table rows to get year wise data.
- We also use some helper methods for converting data that is not in standard form and append it to the dictionary.
- Once all the data is stored in the dictionary, it is loaded in a dataframe.
- The GitHub link for the notebook can be accessed here.

Data Wrangling

- First we load the dataframe we created in SpaceX API section and check for null values in each column.
- The missing payload mass values were filled with the mean value in that section.
- We perform some basic data analysis like number of launches from each site, number and types of mission outcomes etc.
- We find that there are 8 different mission outcomes. Instead of keeping 8 different outcomes, we label them as 1 for success and 0 for failure.
- We call the labels as ‘class’ and add the column ‘class’ to the dataframe.
- Click here to access the notebook from GitHub.

EDA with Data Visualization

- We have used scatter plots to find any correlation between flightnumber vs payload on the outcome of the launch.
- Scatter plots are also made for the launch sites vs payload masses, Flightnumber vs orbit. Scatter plots help us identify if there is a general increasing or decreasing trend and how strong the correlation is between 2 given points. It also helps understand how the data is distributed.
- A line plot is created between year and average success rate for that year.
- A bar chart was also used to show which orbits had highest success rates.
- The GitHub URL for the EDA with data visualization notebook is here.

EDA with SQL

- The first step is to load the SQL extension and connect with the database.
- We then got the list of all the distinct launch sites.
- We displayed the total payload mass carried by boosters launched by NASA (CRS)
- We got the average payload mass carried by booster version F9 V1.1.
- We found the date of first successful landing on ground.
- We got the number of success and failure mission outcomes.
- We got the list of booster versions that carried the maximum payload.
- We ranked the various landing outcomes between 2 given dates based on the count.
- The GitHub link for the notebook is here.

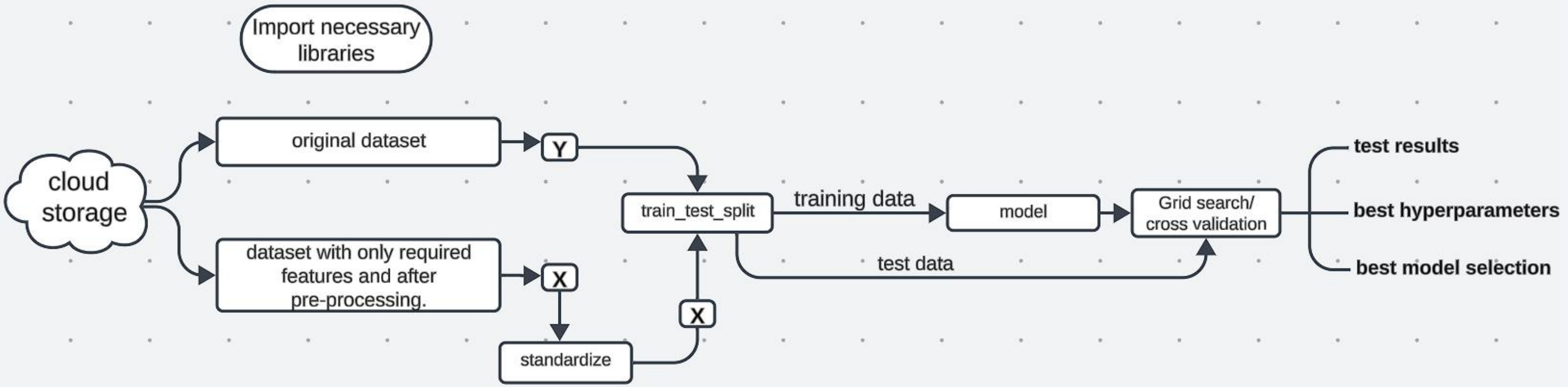
Build an Interactive Map with Folium

- In the launch site analysis, we used markers to mark all 4 launch sites on the map. Markers make it easy to identify the points on the map even when zoomed out.
- We then create a marker cluster with a bunch of markers and icons to depict the outcomes for each launch site. The successes were marked with green and failures with red.
- Various distances are measured from the launch site for coastline, highway, city and railways.
- These distances (in KM) are drawn on the map with a marker and a blue line is drawn between the points. This makes it easy to understand which two points are used to calculate the distance.
- The GitHub link for the notebook is here.

Build a Dashboard with Plotly Dash

- From the dashboard, the user can choose to view analytics of either all sites or an individual launch site.
- The first plot shows a pie chart of total success percentage of all the sites or success to failure percentage of an individual site as chosen by user.
- The second plot is a scatter plot showing the relationship between success/failure to the payload mass for each booster version category.
- For further analysis, user can set the ranges of payload mass using a slider.
- The GitHub link for the python file can be found here.

Predictive Analysis (Classification)

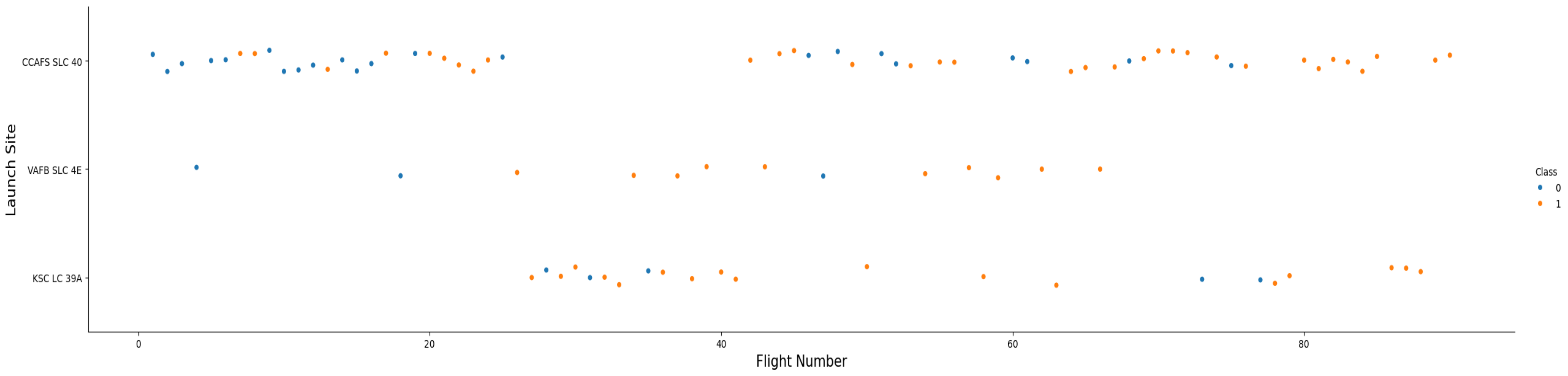


Predictive Analysis (Classification)

- The data is loaded from the two data sets we created earlier. The independent variables(X) and the outcome(Y) are selected.
- The independent variables are standardized to bring all the values within a common range.
- The dataset is split into training and test sets.
- The classification models are built and trained.
- Grid search is applied to get the optimum hyperparameters for the given model.
- The prediction scores and accuracies are calculated and confusion matrix is generated.
- By comparing the results, the best model is selected.
- The GitHub link for the notebook is here.

Results

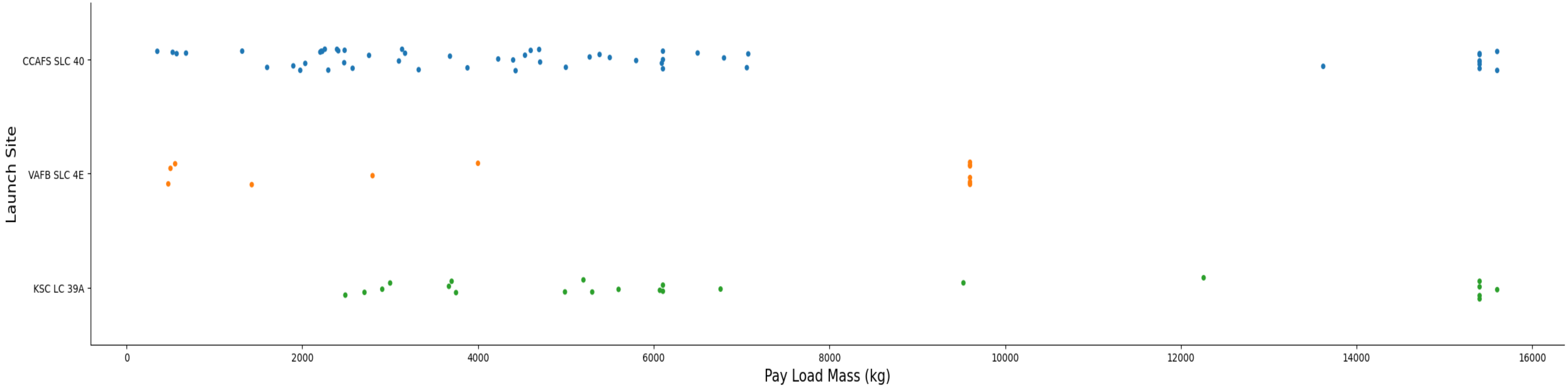
Exploratory data analysis with visualization



Effect of flight number and launch site on launch outcome

Results

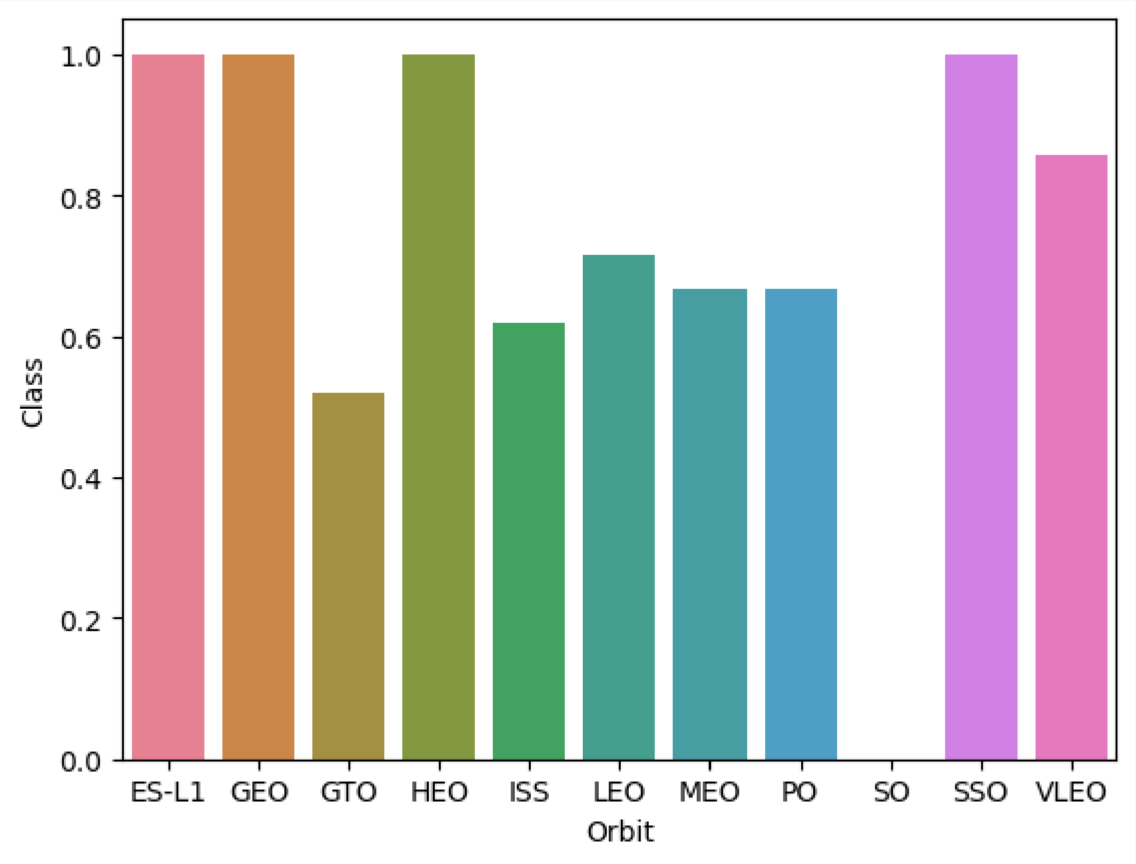
Exploratory data analysis with visualization



Scatter plot to show relation between launch sites and their payload mass

Results

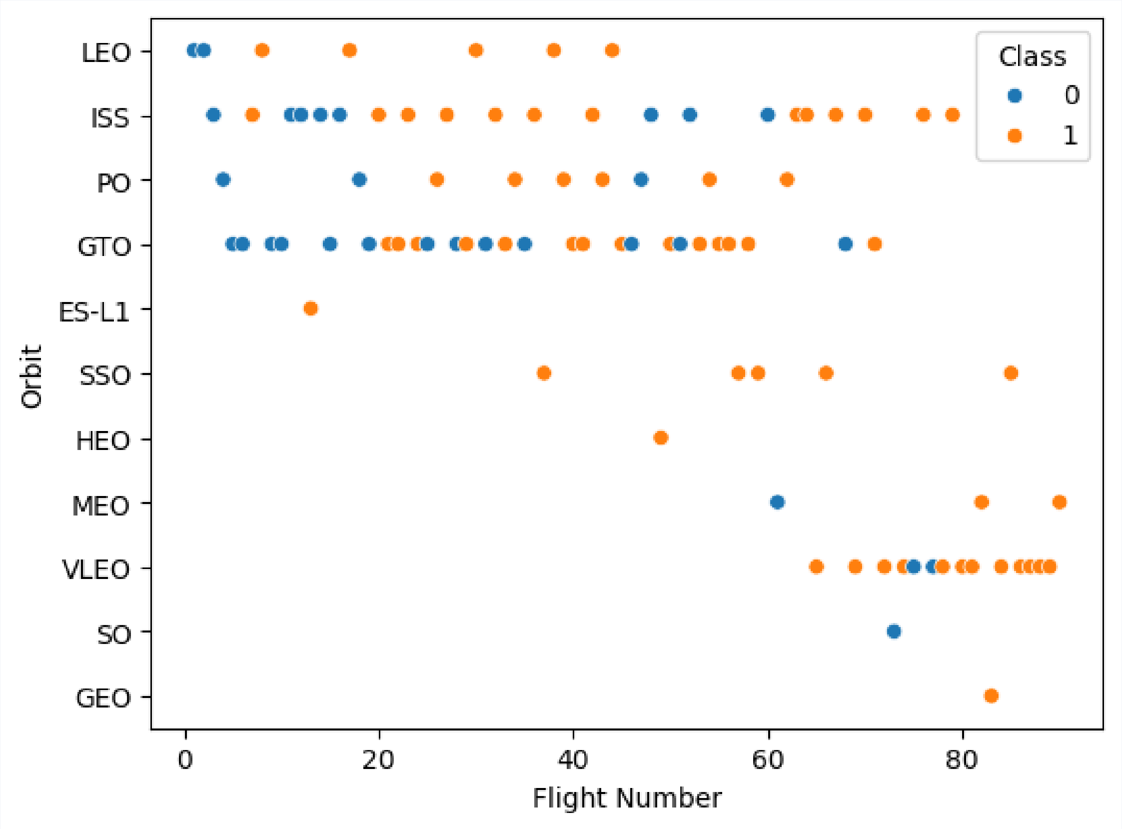
Exploratory data analysis with visualization



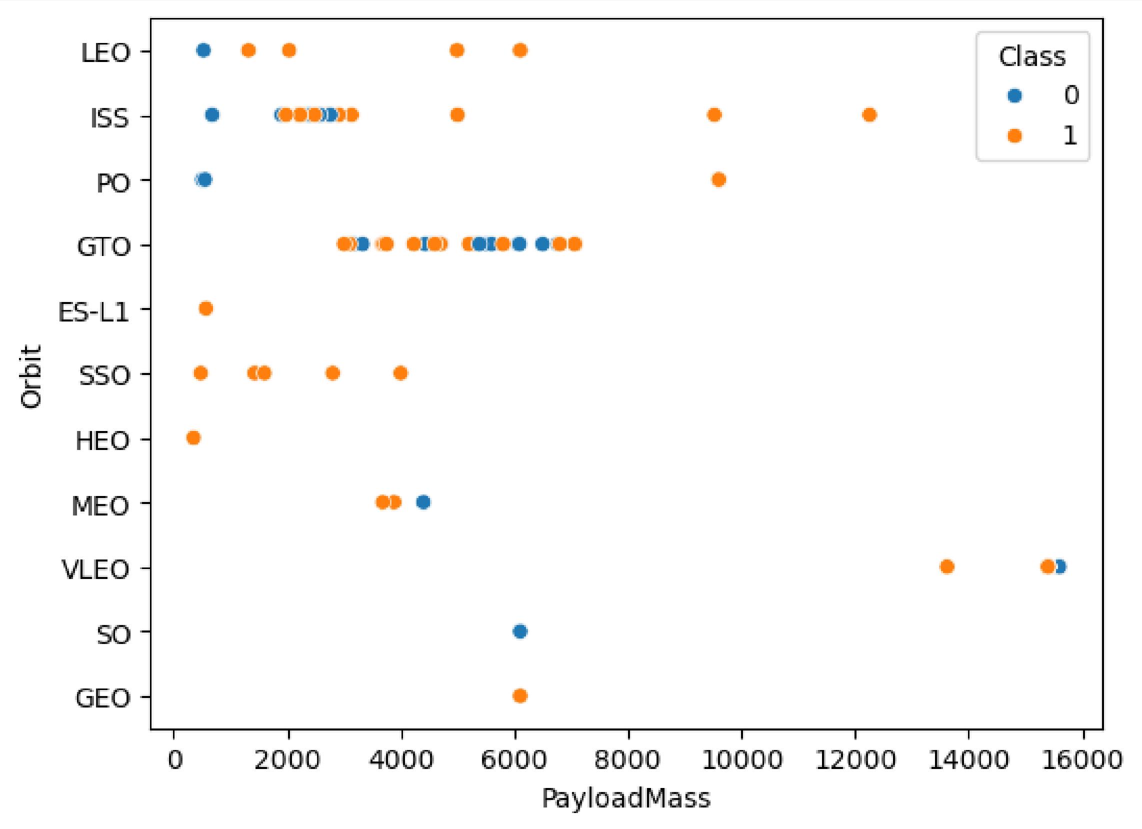
Bar plot showing success rate for each orbit

Results

Exploratory data analysis with visualization



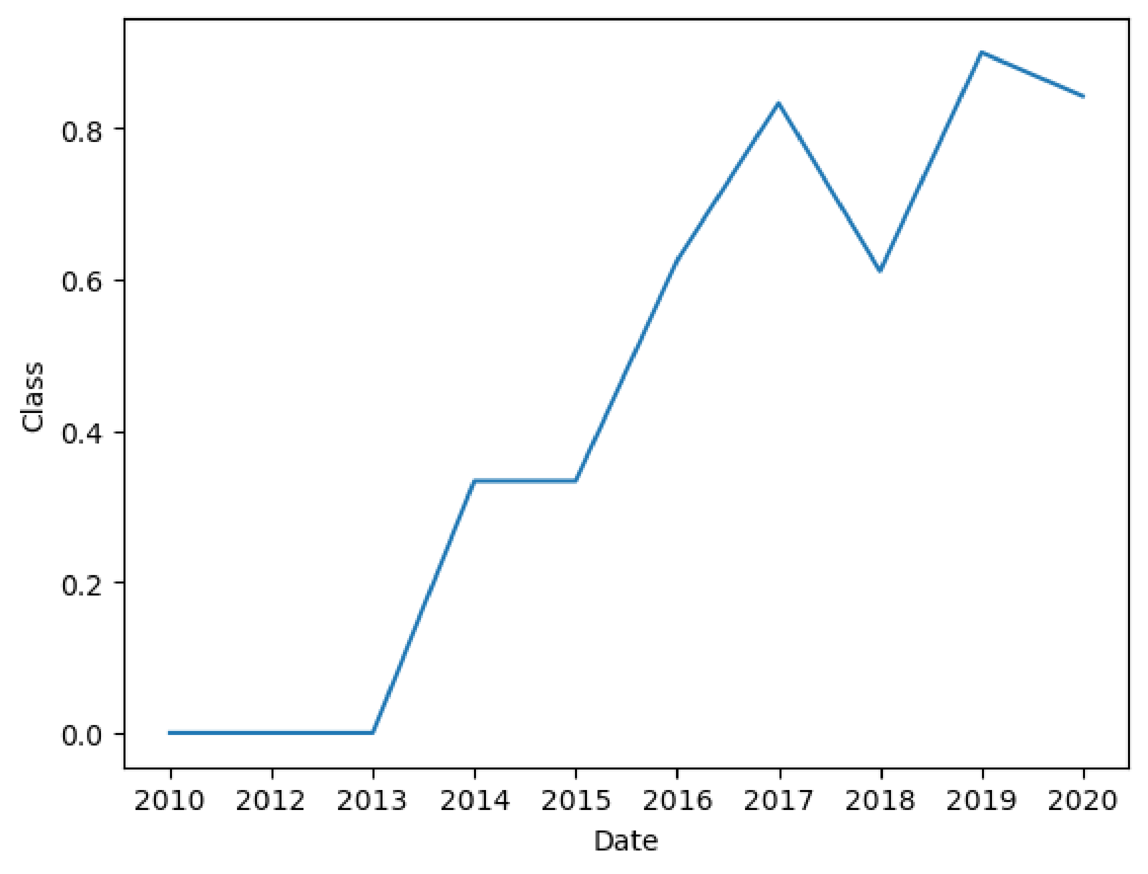
Scatter plot between flight number and orbit type



Scatter plot between payload mass and orbit type

Results

Exploratory data analysis with visualization



Line chart showing average success rate over the years

Results

Exploratory data analysis with SQL

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

The list of all the distinct launch sites

sum(PAYLOAD_MASS__KG_)
45596

Total payload mass carried by boosters launched by NASA(CRS)

Results

Exploratory data analysis with SQL

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The list of 5 records where launch site name begins with ‘CCA’

Results

Exploratory data analysis with SQL

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

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

min(Date)
2015-12-22

Date of first successful landing outcome in ground pad

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

List of boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.

Results

Exploratory data analysis with SQL

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The list of various mission outcomes along with their total count

Booster_Version
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

The list of booster versions that have carried the maximum payload.

Results

Exploratory data analysis with SQL

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

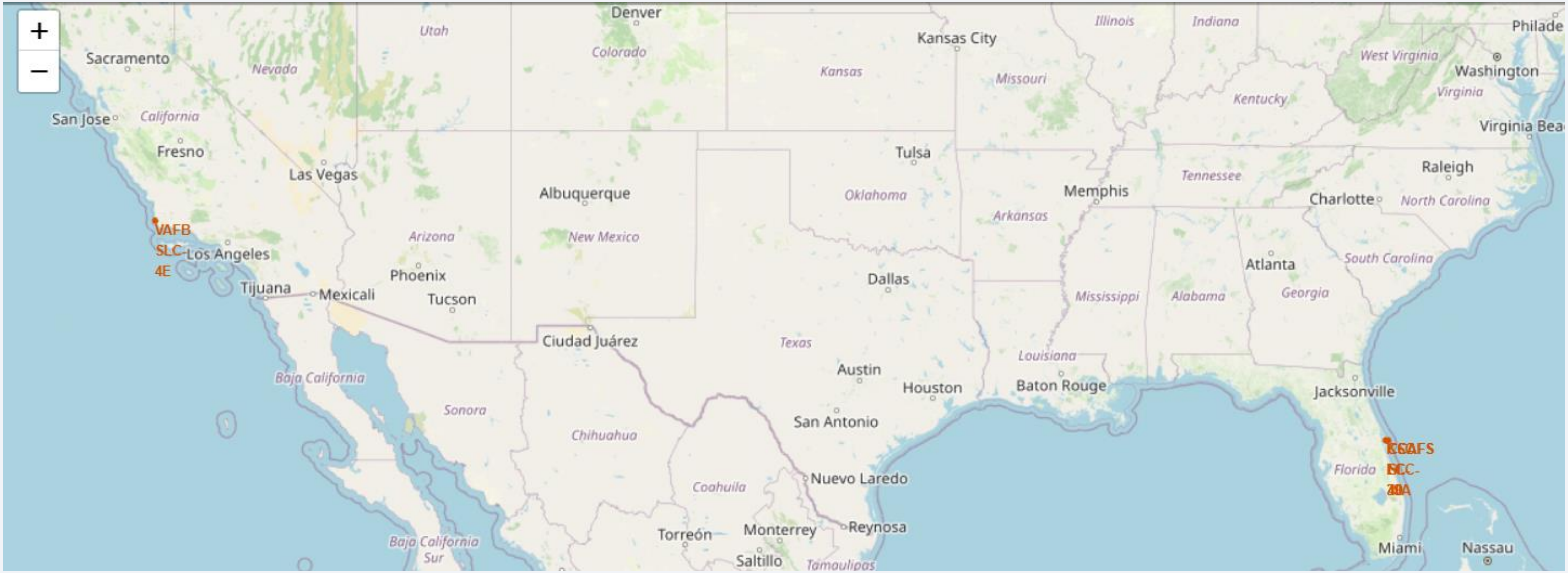
Failure landing outcomes in drone ship in the months of 2015

Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The list of landing outcomes between 2010-06-04 and 2017-03-20 ranked by the total count.

Results

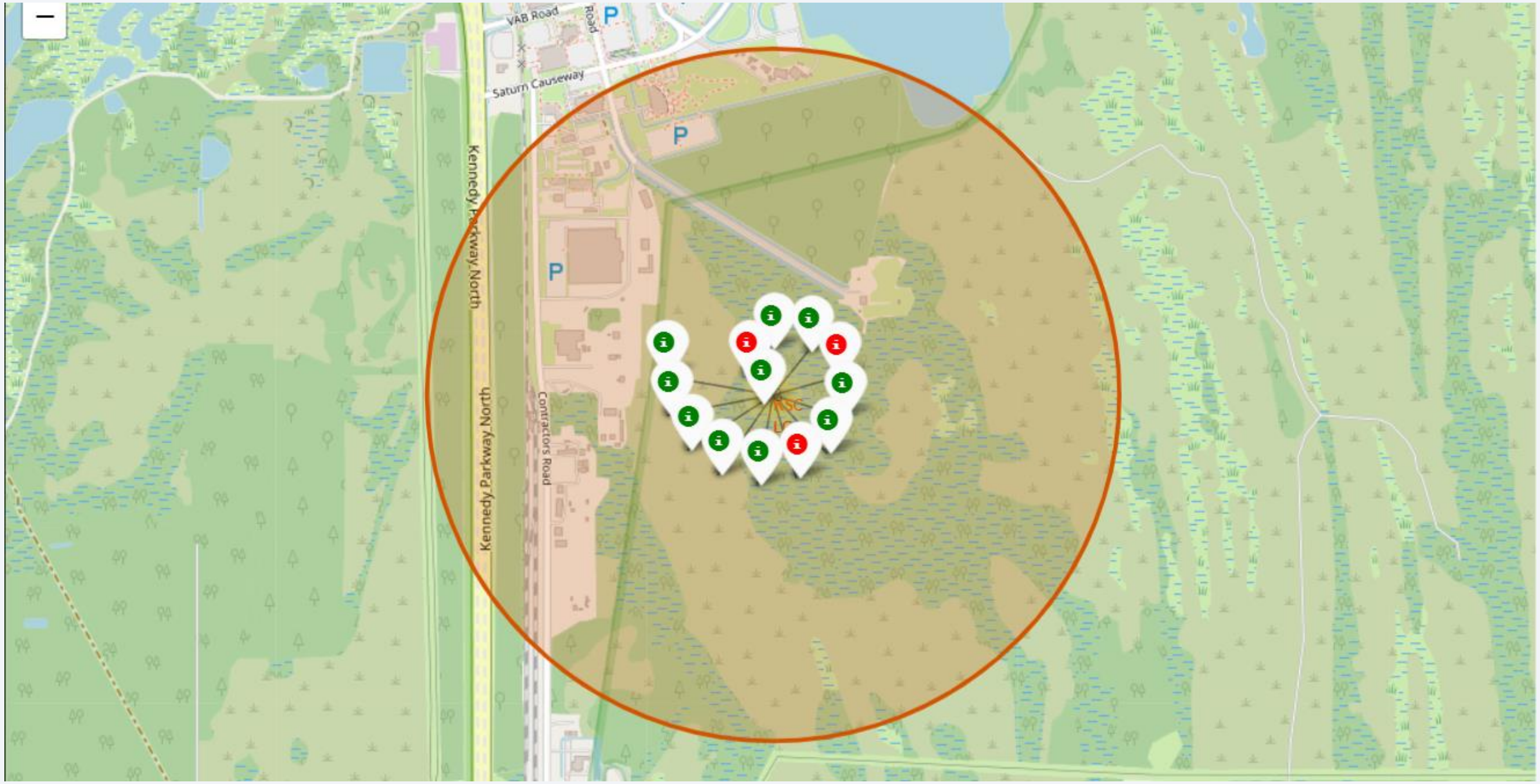
Interactive geospatial analysis with Folium map



Map with different launch sites shown with markers

Results

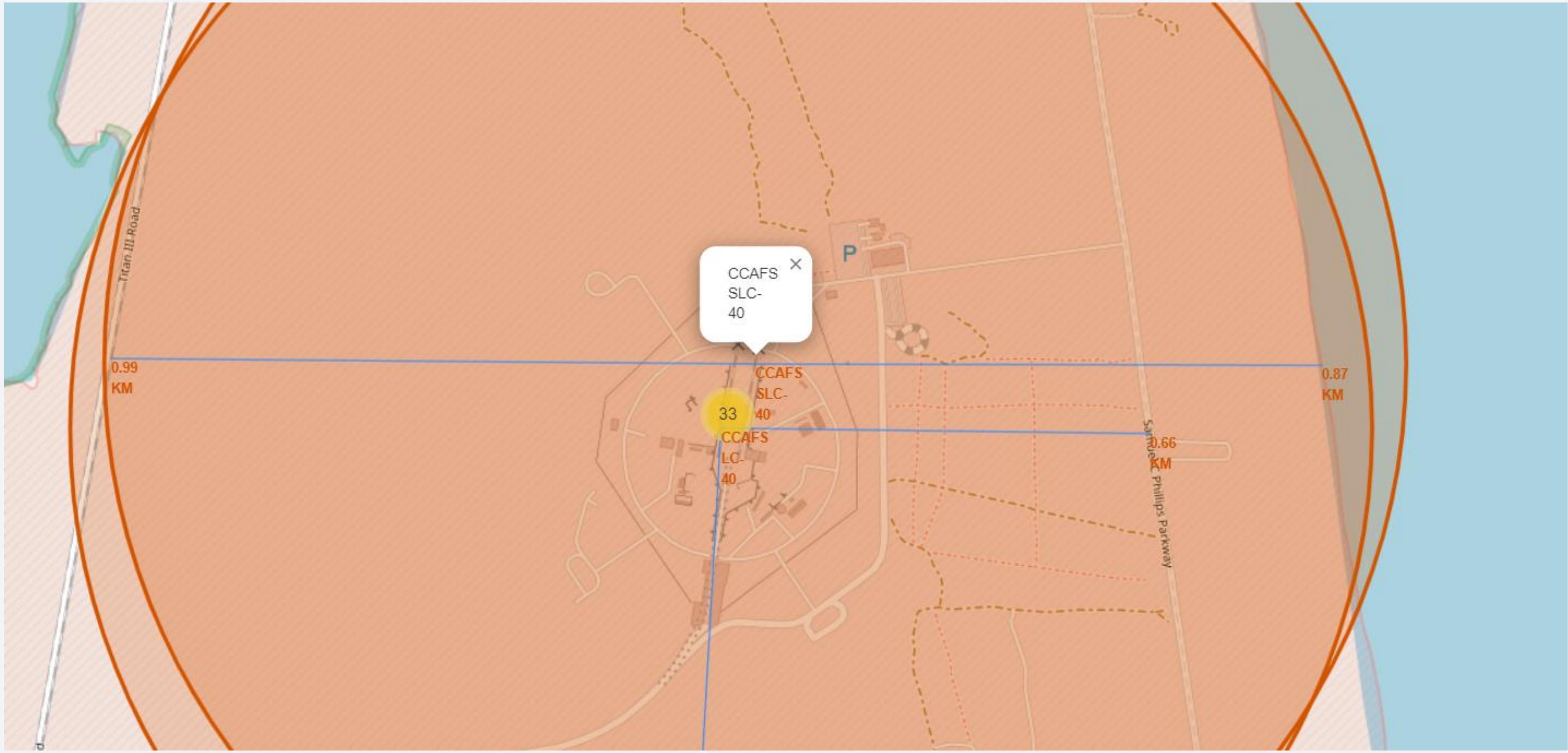
Interactive geospatial analysis with Folium map



Map showing successful and failed launches for launch site KSC LC-39A

Results

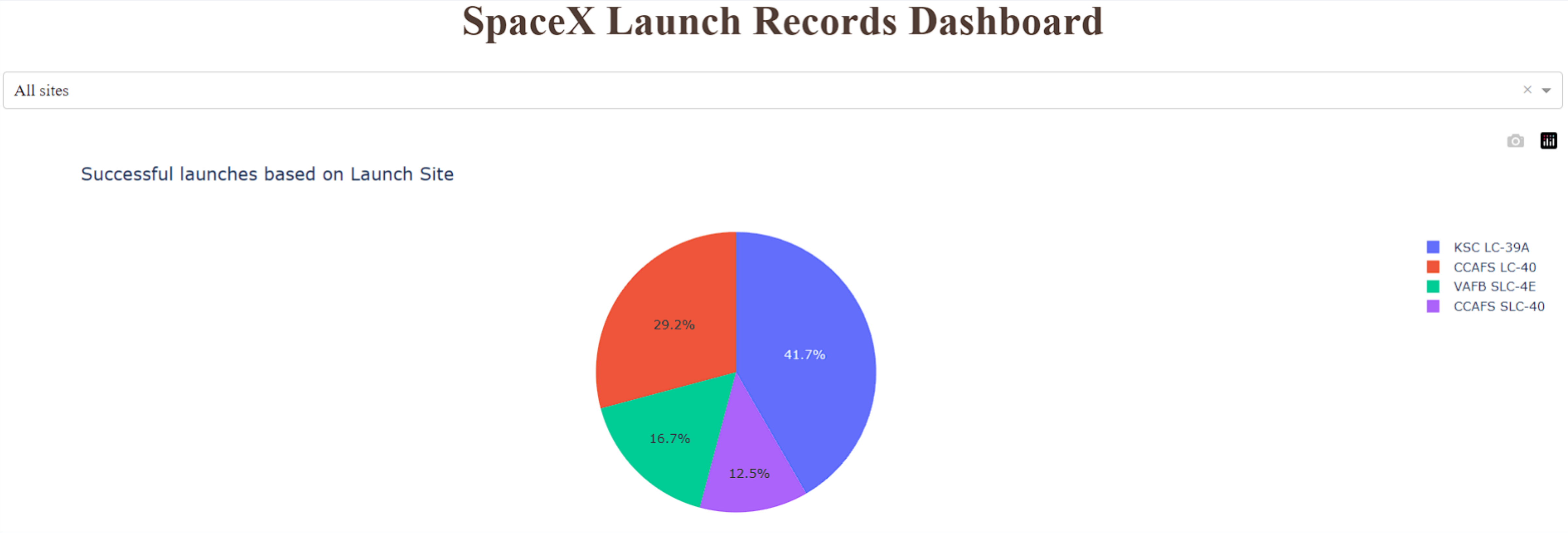
Interactive geospatial analysis with Folium map



Map showing distance of launch sites from nearest highway, coastline and railway.

Results

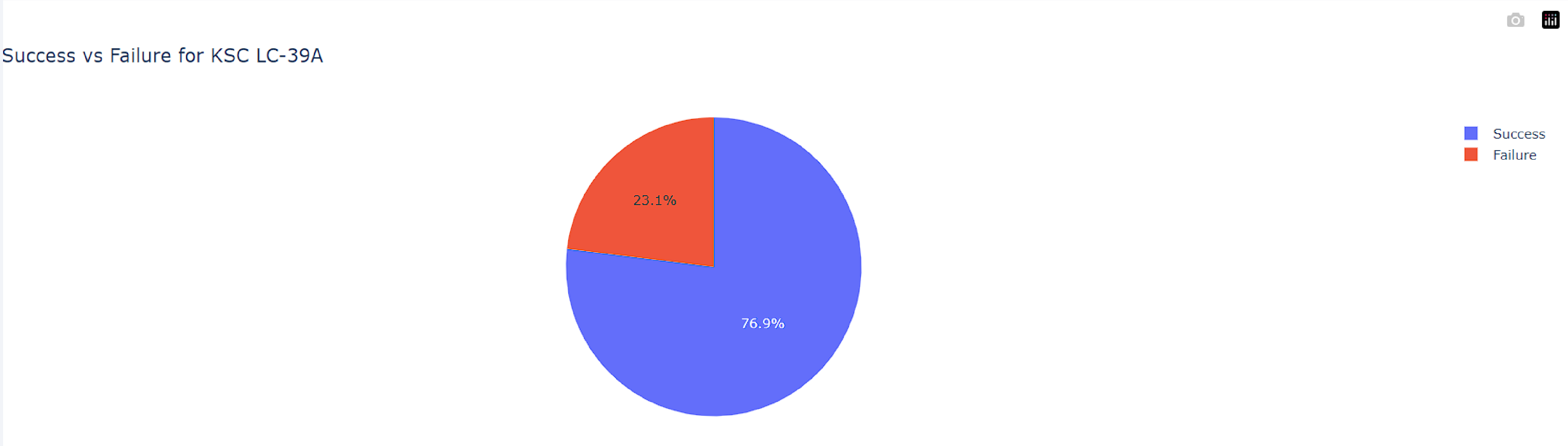
Interactive dashboard using Plotly Dash



Pie chart for successful launches for all the launch sites

Results

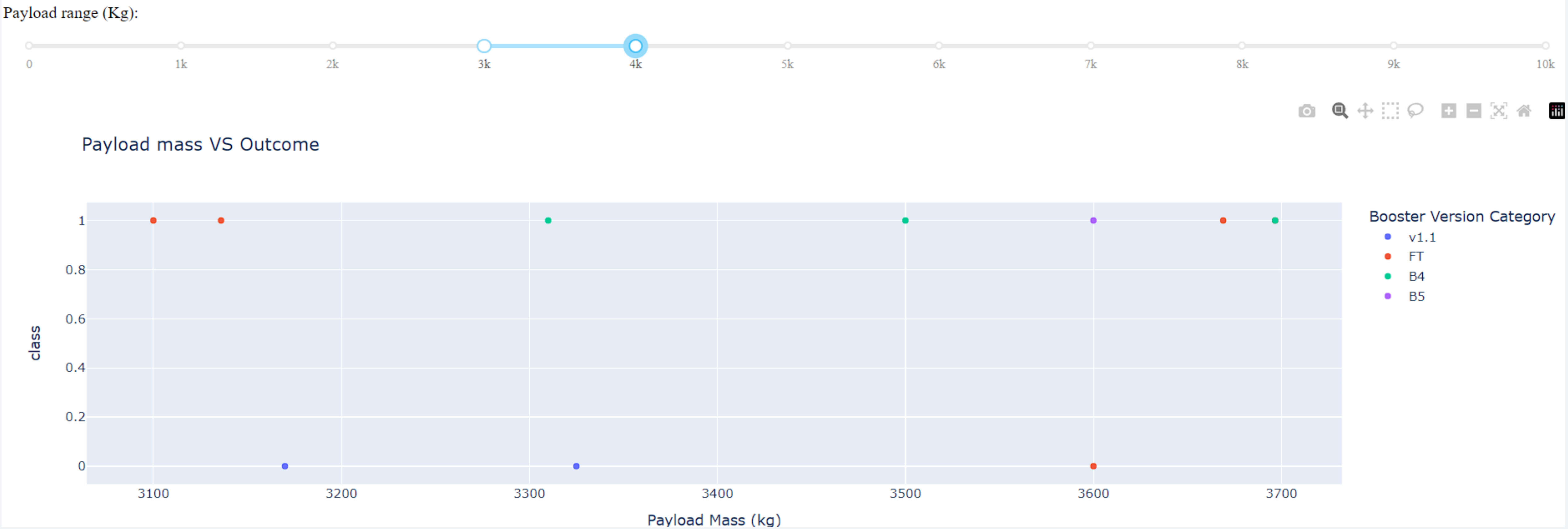
Interactive dashboard using Plotly Dash



Pie chart showing success/failure percentage of launches
for launch site KSC LC-39A

Results

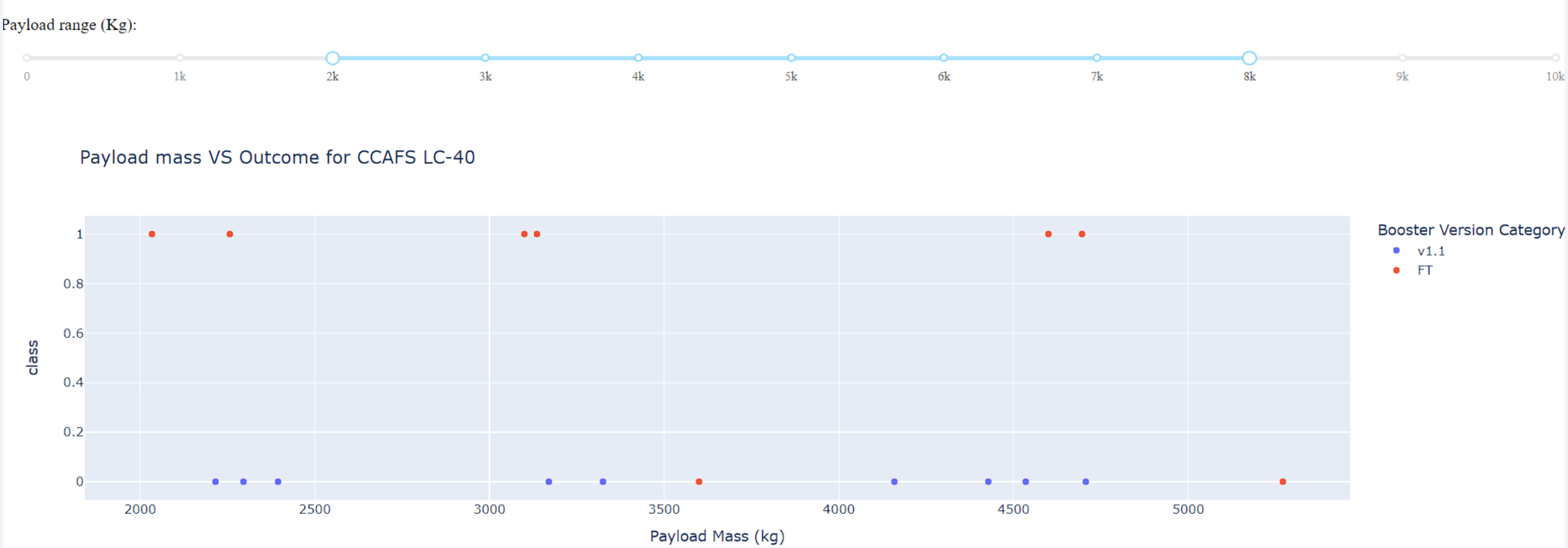
Interactive dashboard using Plotly Dash



Scatter plot to show relationship between payload mass(3000-4000kg) and launch outcome for different booster versions when all sites are selected

Results

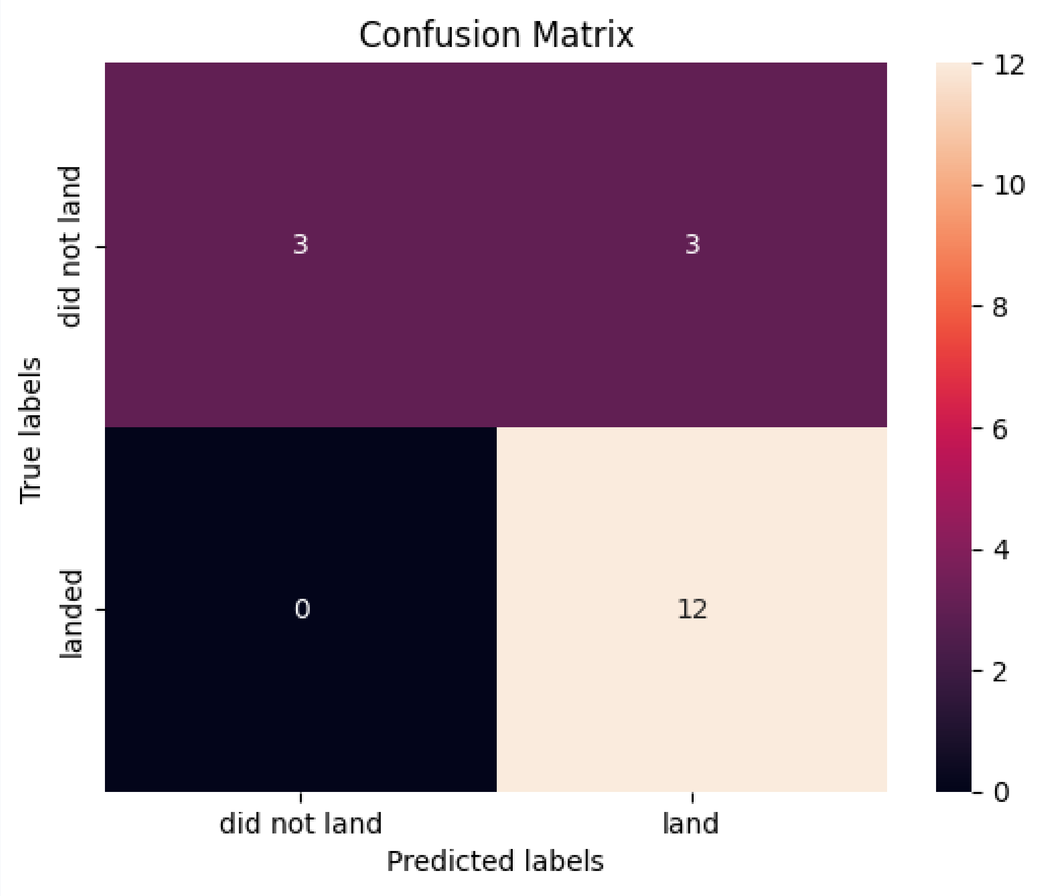
Interactive dashboard using Plotly Dash



Scatter plot to show relationship between payload mass(2000-8000kg) and launch outcome for different booster versions for launch site CCAFS LC-40

Results

Predictive analysis using classification



Confusion matrix for logistic regression

Best hyperparameters from Grid Search

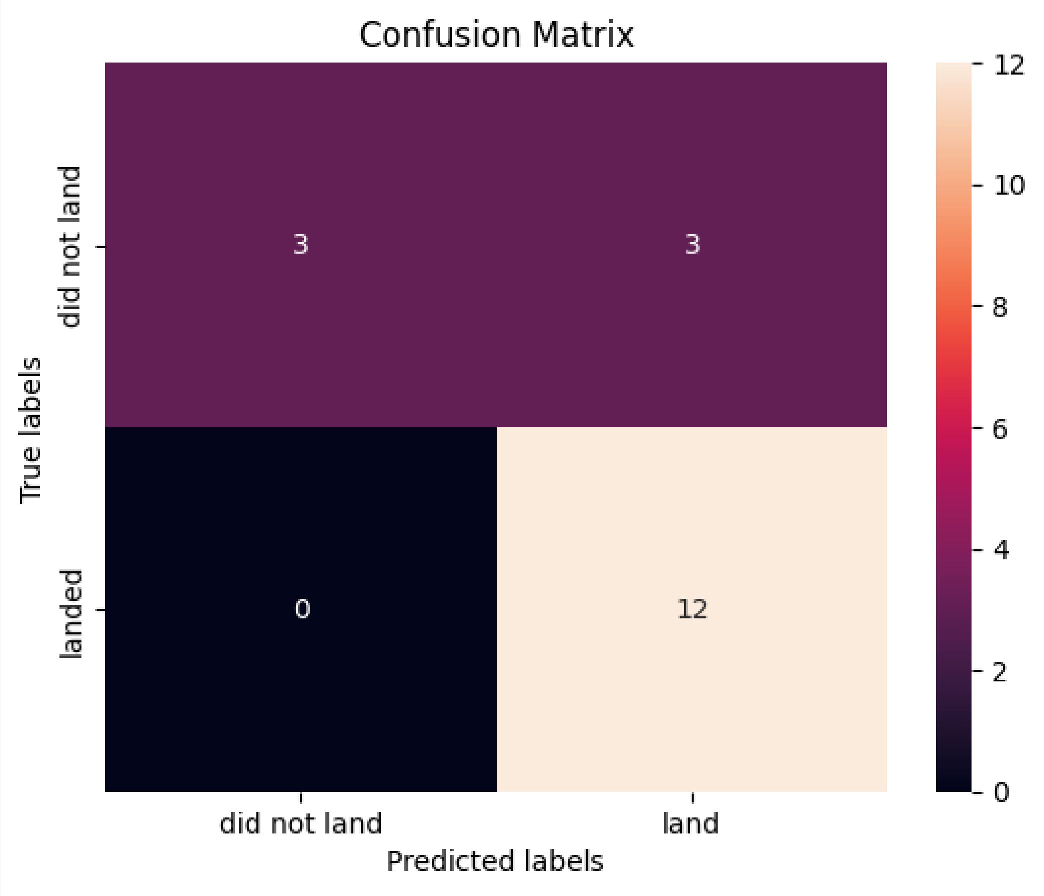
{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}

Accuracy with test data

0.83

Results

Predictive analysis using classification



Confusion matrix for SVC

Best hyperparameters from Grid Search

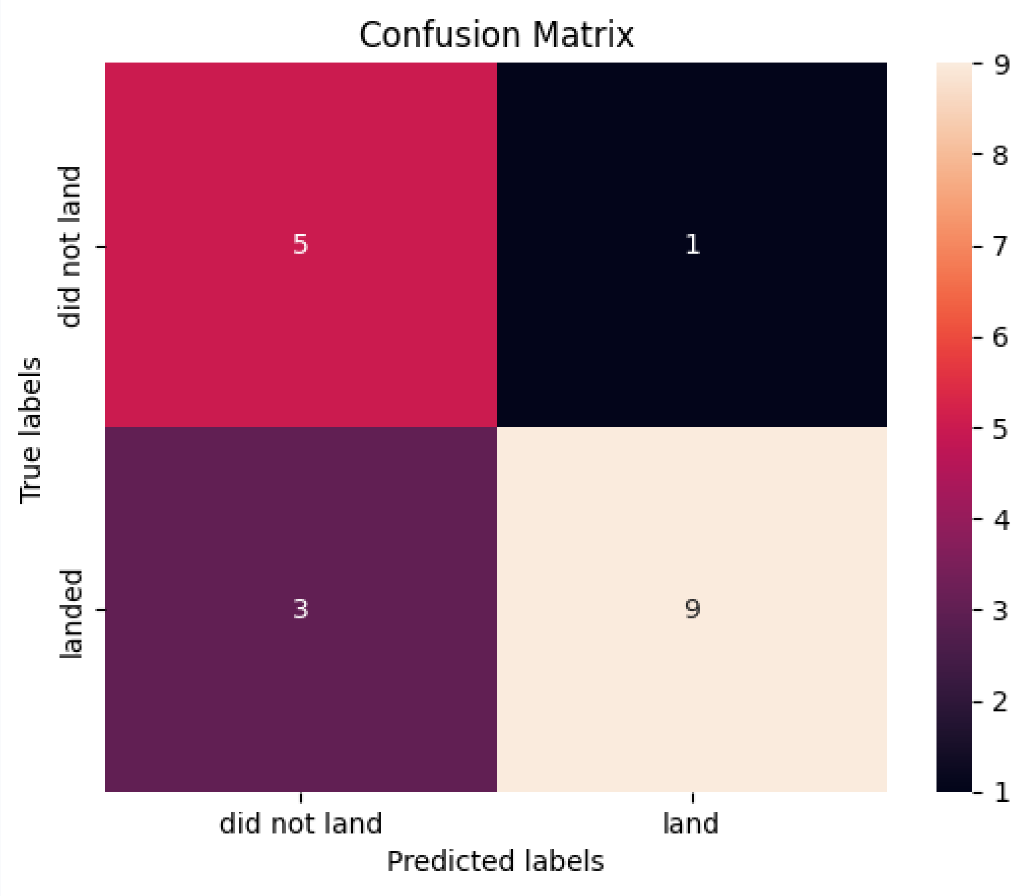
{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}

Accuracy with test data

0.83

Results

Predictive analysis using classification



Confusion matrix for DecisionTreeClassifier

Best hyperparameters from Grid Search

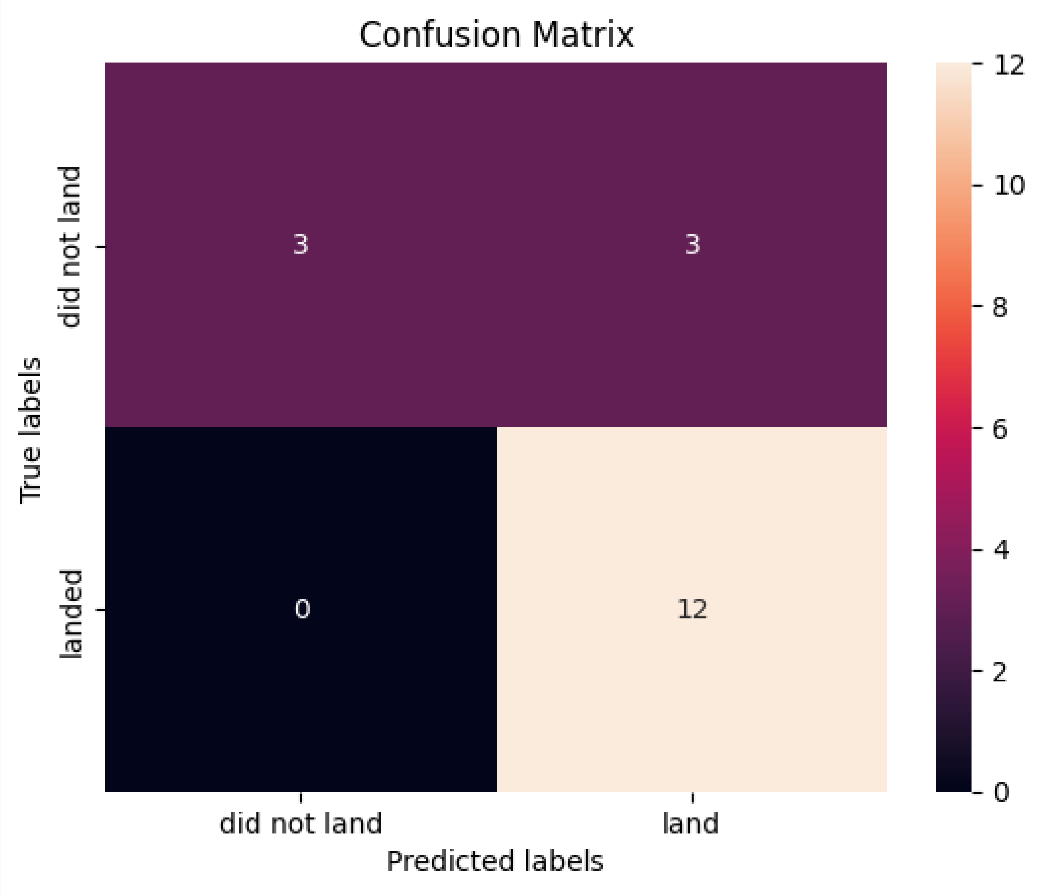
{'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}

Accuracy with test data

0.77

Results

Predictive analysis using classification



Confusion matrix for KNN

Best hyperparameters from Grid Search

{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

Accuracy with test data

0.83

Results

Predictive analysis using classification

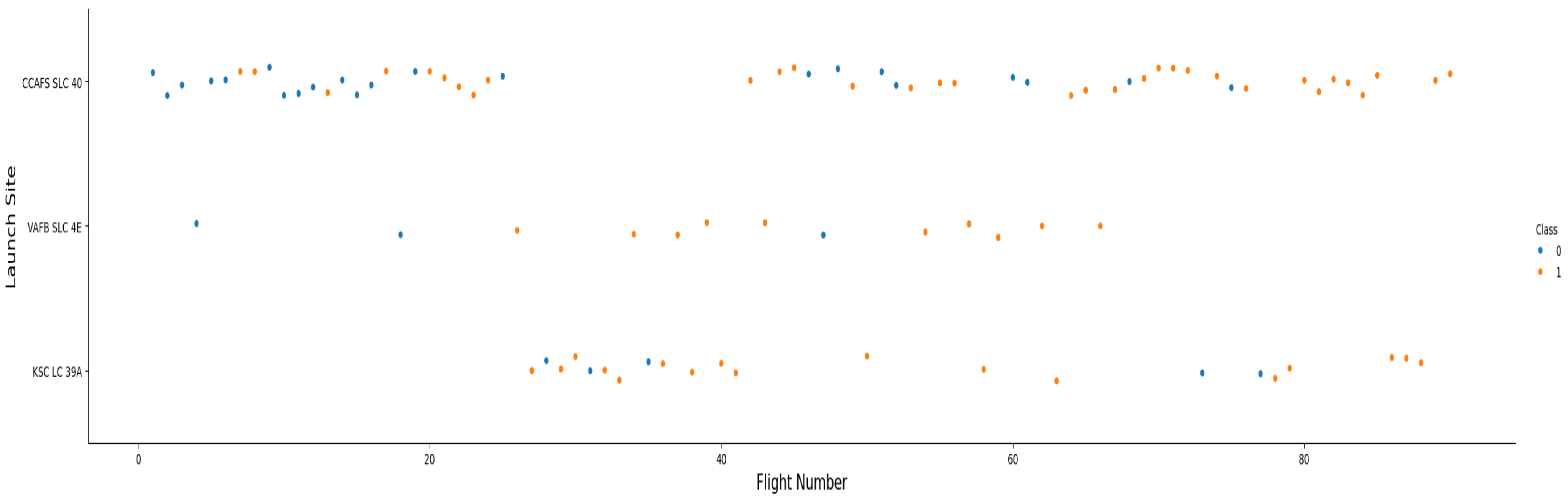
Summary of training and test accuracies for the 4 models:

	Prediction accuracy	
Model	Training	Test
Logistic Regression	0.846	0.833
Support Vector Classification	0.848	0.833
Decision Tree Classifier	0.861	0.777
KNN	0.848	0.833

Section 2

Insights drawn from EDA

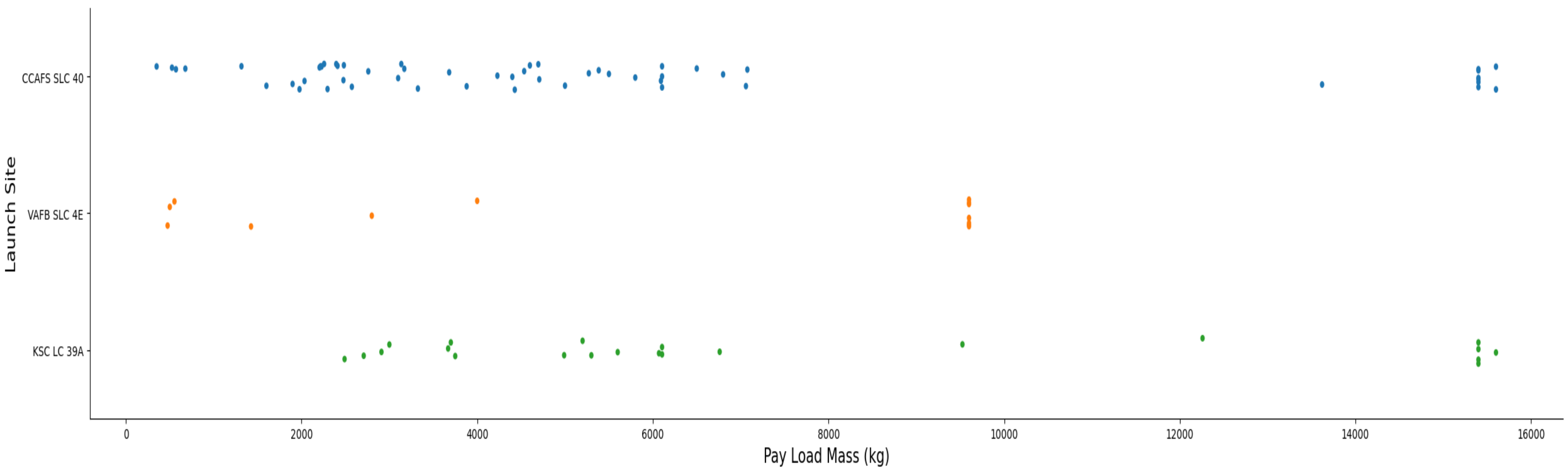
Flight Number vs. Launch Site



Flight Number vs. Launch Site

- The scatterplot of Flight number and launch site is shown.
- We observe an overall increasing trend of successful landing as the flight number increases.
- In launch site CCAFS SLC-40, we find an erratic pattern in success/failure rates when flight numbers are between 40 to 60.

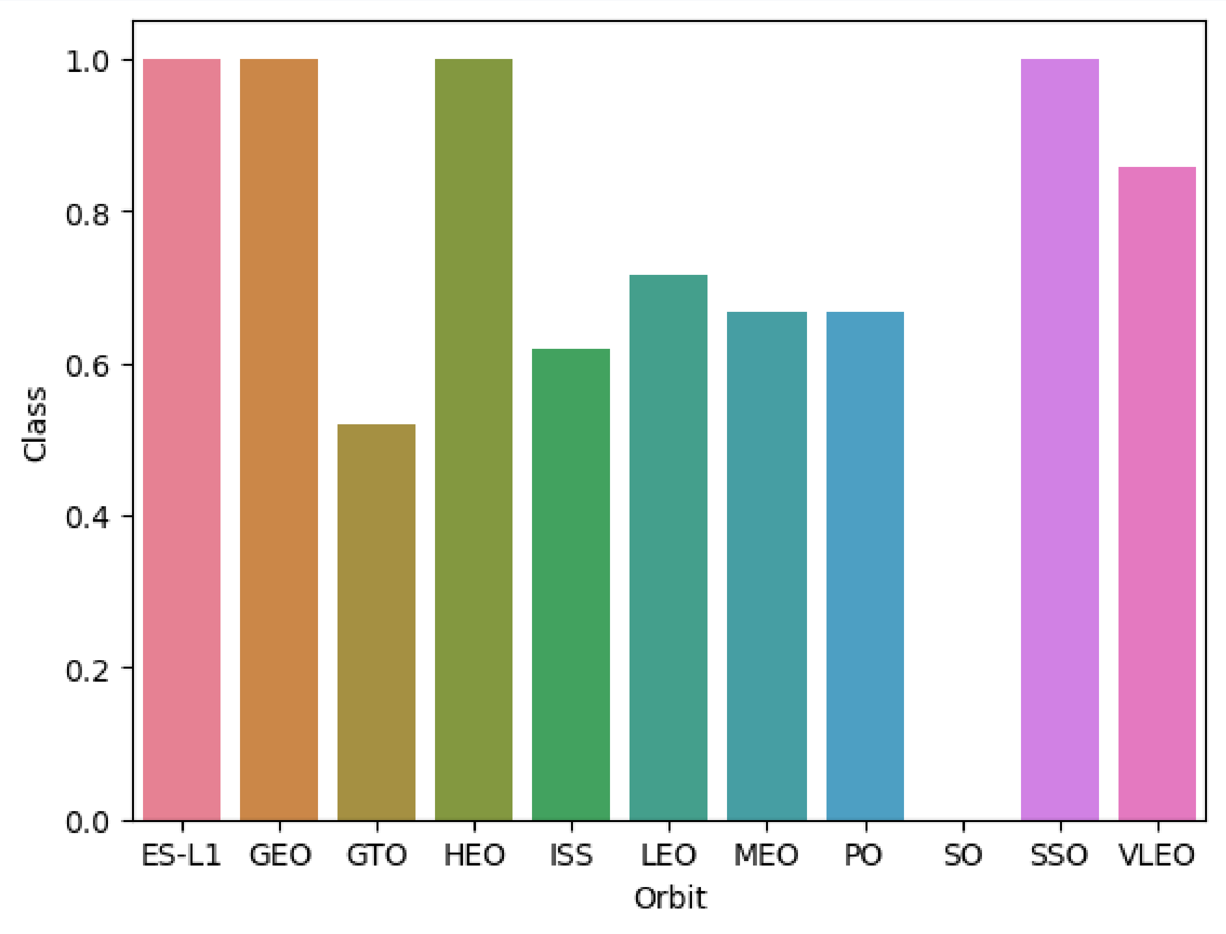
Payload vs. Launch Site



Payload vs. Launch Site

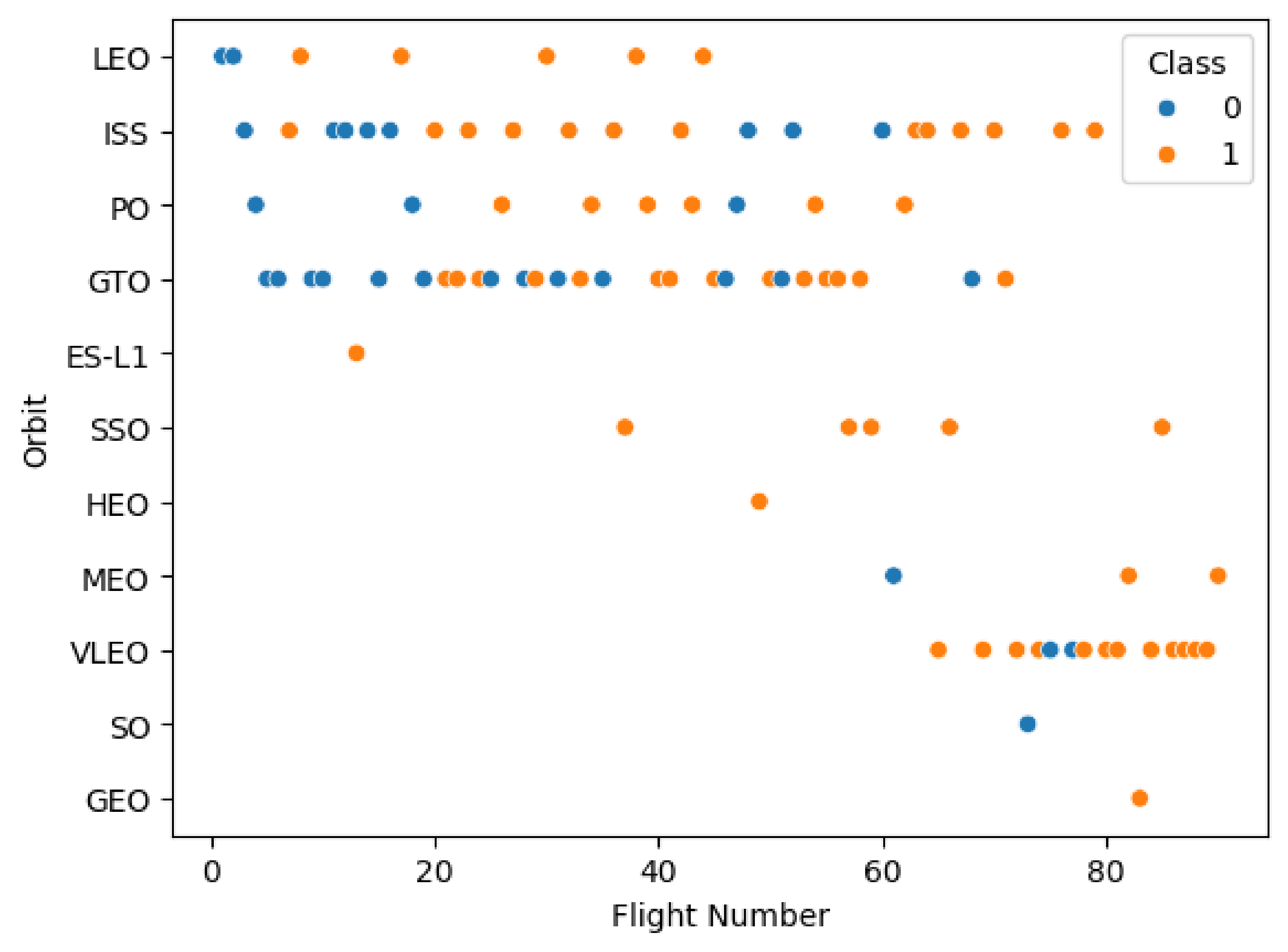
- The scatter plot of payload vs launch sites has been shown.
- We observe that for site VAFB-SLC, no rockets are launched beyond 10000 kg payload.
- For launch site CCAFS SLC-40, majority of the rockets were launched with a payload in the range of 2000 – 6000 kg.
- The launch site VAFB-SLC didn’t launch rockets in the mid ranges between 5000-9000 kg.

Success Rate vs. Orbit Type



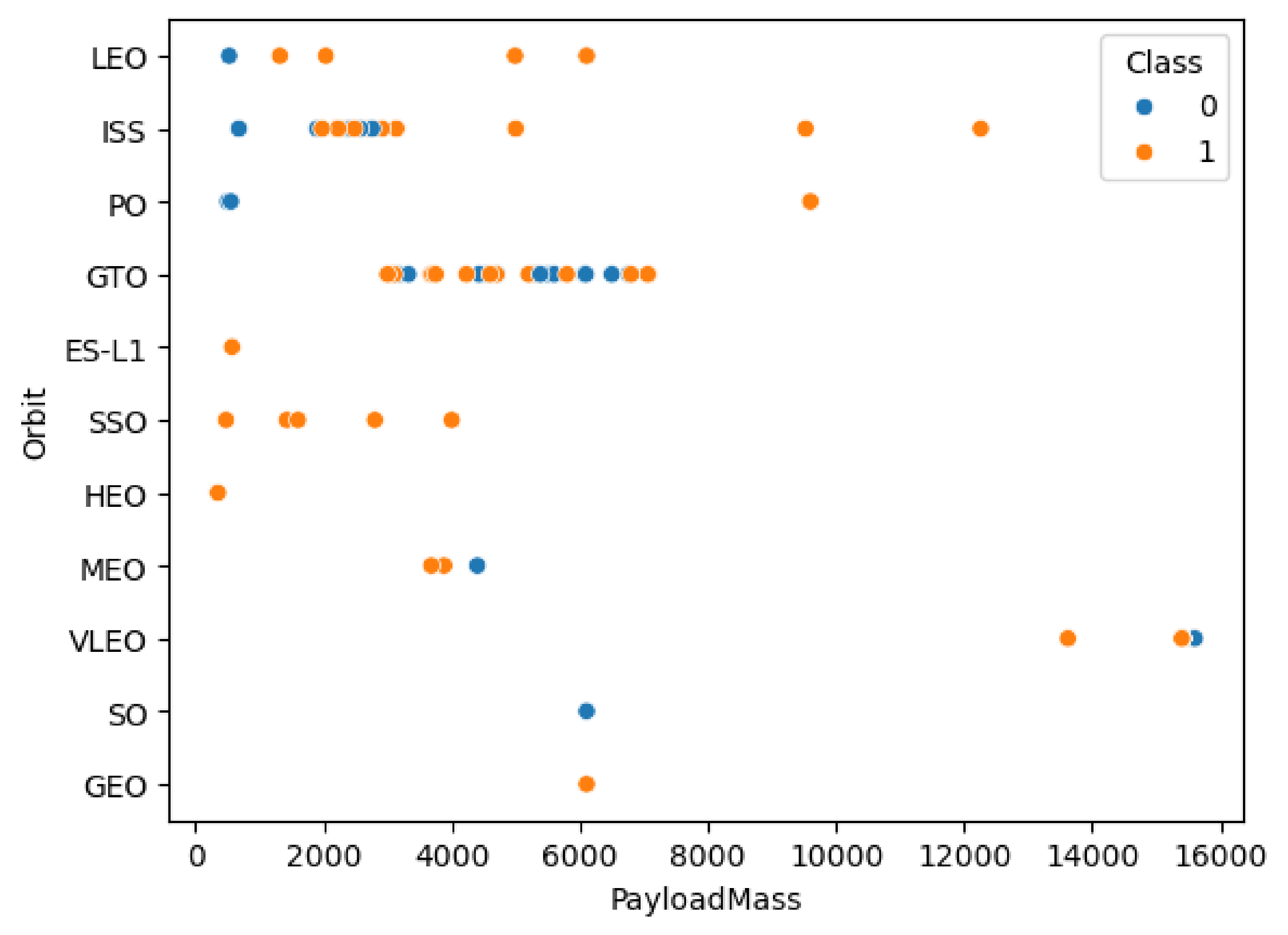
- The bar chart shows the success rate inn various orbits.
- It is seen that orbits ES-L1, GEO, HEO, SSO have the highest success rates.
- The high success rate for ES-L1 can be related to the fact that a small object in that orbits is in equilibrium relative to center of mass of large bodies.

Flight Number vs. Orbit Type



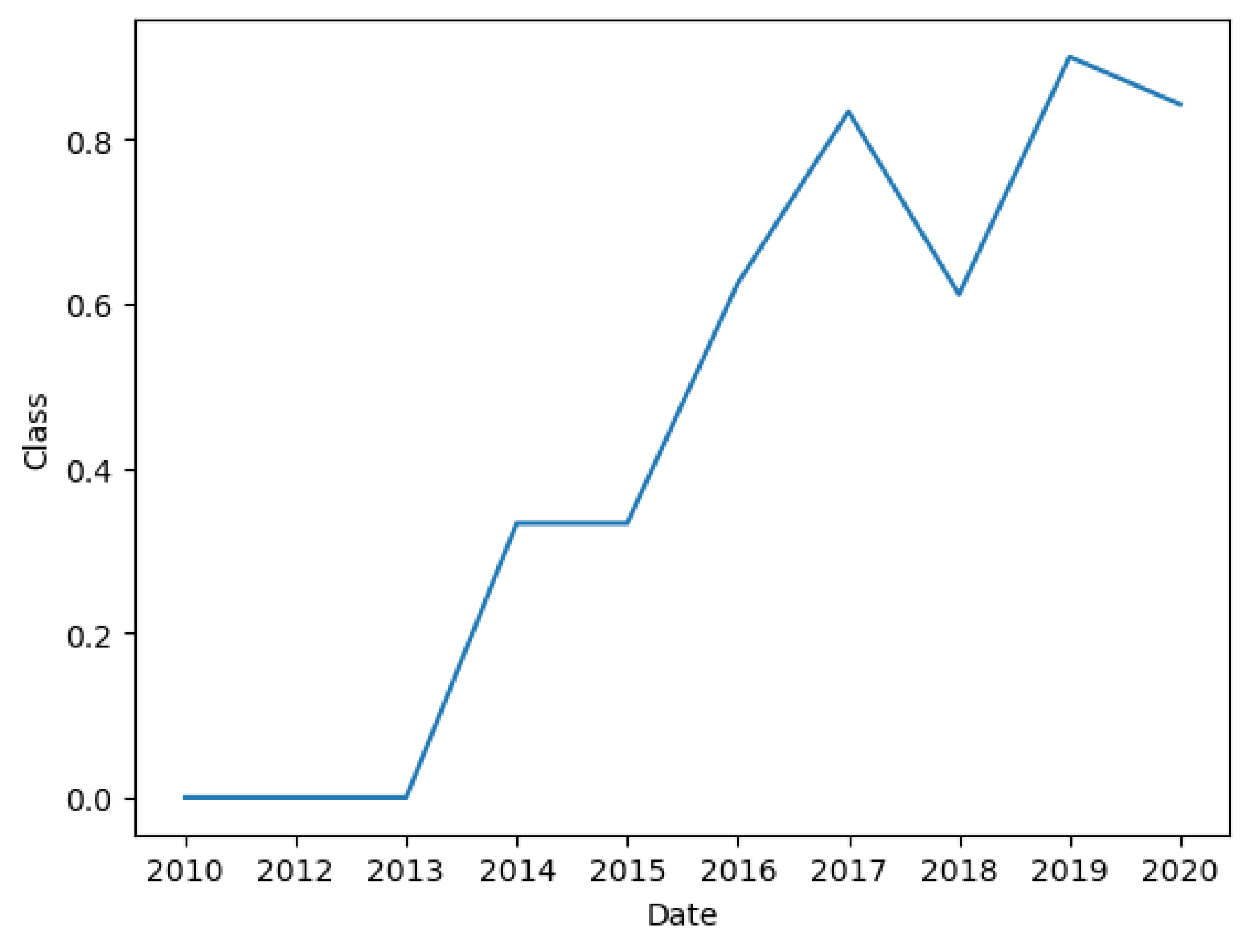
- From the scatter point orbit LEO has an increasing rate of success with the number of flights.
- A similar trend is observed for VLEO and MEO.
- However for other orbits no apparent relationship exists between flight number and orbit.

Payload vs. Orbit Type



- It can be seen that with increasing payload the success rate increases for orbits LEO, ISS and PO.
- However no strong relation is observed between success rate and payload mass for orbits GTO, VLEO and MEO.

Launch Success Yearly Trend



- The line chart depicts the yearly average success rates.
- We can see that there is an overall increase in success from the year 2013 onwards.
- There was a sharp decline in 2017-18. The causes for the same require further study.

All Launch Site Names

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

- The unique launch sites have been shown on the left.
- The following query can be used to get this list.

```
%sql select distinct Launch_Site from SPACEXTABLE
```

- It selects the distinct launch sites present in the SPACEXTABLE table.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The following query can be used to get the launch site names starting with ‘CCA’

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

- It selects the data from the table where the launch site names begin with ‘CCA’ followed by 0 or more characters.

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

- The total payload mass carried by boosters launched by NASA (CRS) is shown.
- The following query generates the result.

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE
where Customer = 'NASA (CRS)'
```

- We get the sum of payload mass from the table keeping customer as NASA (CRS).

Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 is shown here.
- The query to get this result is:

```
%sql select avg(PAYLOAD_MASS__KG_) from
SPACEXTABLE where Booster_Version like 'F9
v1.1%’
```

- We use the avg() method on payload mass and set booster version as F9 v1.1

First Successful Ground Landing Date

min(Date)
2015-12-22

- Figure shows the date of first successful ground landing.
- The query for this is:

```
%sql select min(Date) from SPACEXTABLE
where Landing_Outcome like '%ground pad%’
```

- We select the minimum date after filtering landing outcome to ground pad to get the result.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Here we have the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- The query used to generate the list:

```
%sql select Booster_Version from SPACEXTABLE where
Landing_Outcome = 'Success (drone ship)'\
    and PAYLOAD_MASS__KG_ between 4000 and 6000
```

- We select the booster versions and filter our results using the 2 conditions of drone ship landing and payload mass in the range 4000 to 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Figure shows comprehensive study of number of successful and failure mission outcomes.
- Following query is used to get this result:

```
%sql select Mission_Outcome,count(Mission_Outcome)
from SPACEXTABLE group by Mission_Outcome
```

- We group the mission outcomes and then find the total count for each type.
- The result can be summarized as 100 successes and 1 failure.

Boosters Carried Maximum Payload

Booster_Version
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

- The image shows the list of booster versions that have carried the maximum payload.
- The following query was used to get the list:

```
%sql select distinct(Booster_Version) from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)
from SPACEXTABLE)
```

- We select the booster versions where payload is maximum. A subquery was used to get the maximum value of payload from the table.

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The list of failed landing outcomes in drone ship, their booster versions, and launch site names for months in year 2015 is shown in the figure.
- The query that was used:

```
%sql select substr(Date, 6, 2) as
Month,Landing_Outcome,Booster_Version,Launch_Site
from SPACEXTABLE\
```

- ```
      where Landing_Outcome = 'Failure (drone ship)'
  and substr(Date, 0, 5) = '2015’
  ```

- Since sqlite does not support month(), we use substr() to extract the month from the date and we display other details and filter the year to 2015.
  ```

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

Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The figure shows the ranked count of various landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order.
- The following query gives the desired result:

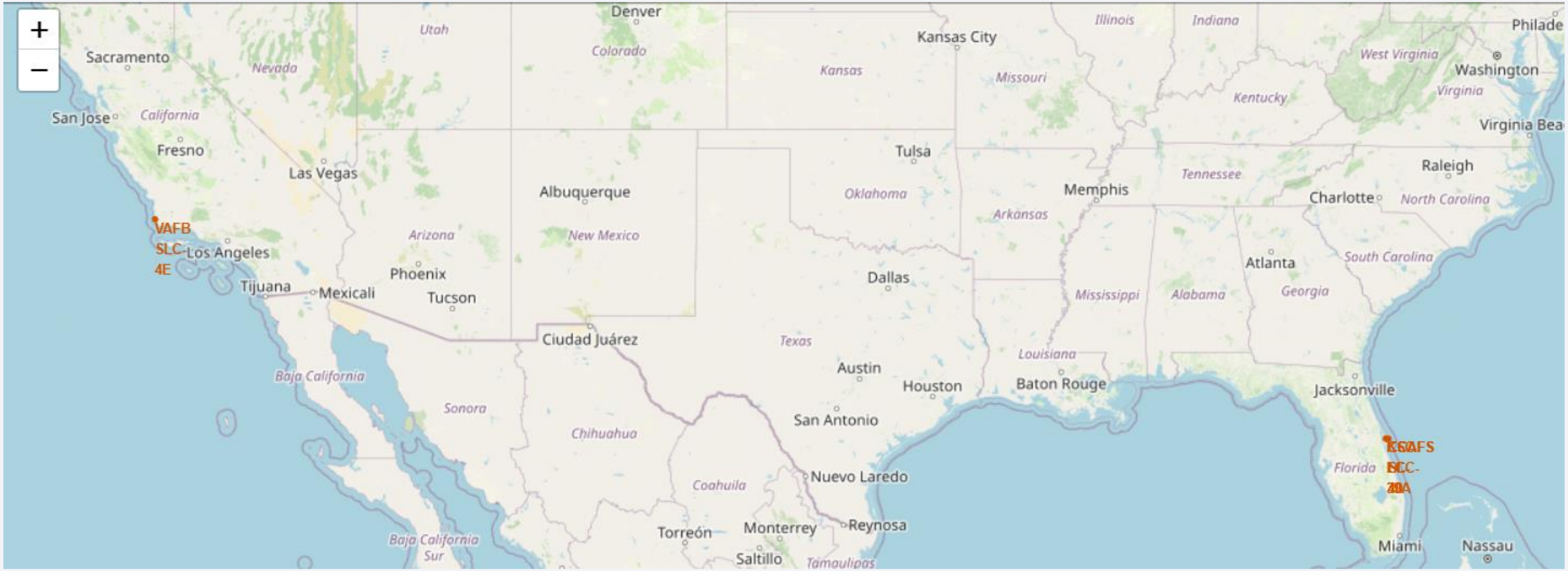
```
%sql select Landing_Outcome, count(Landing_Outcome)
as Total from SPACEXTABLE\
     where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by Total desc
```

- We group the data by landing outcomes, get the count for each type of landing outcome and arrange them in descending order.

Section 3

Launch Sites Proximities Analysis

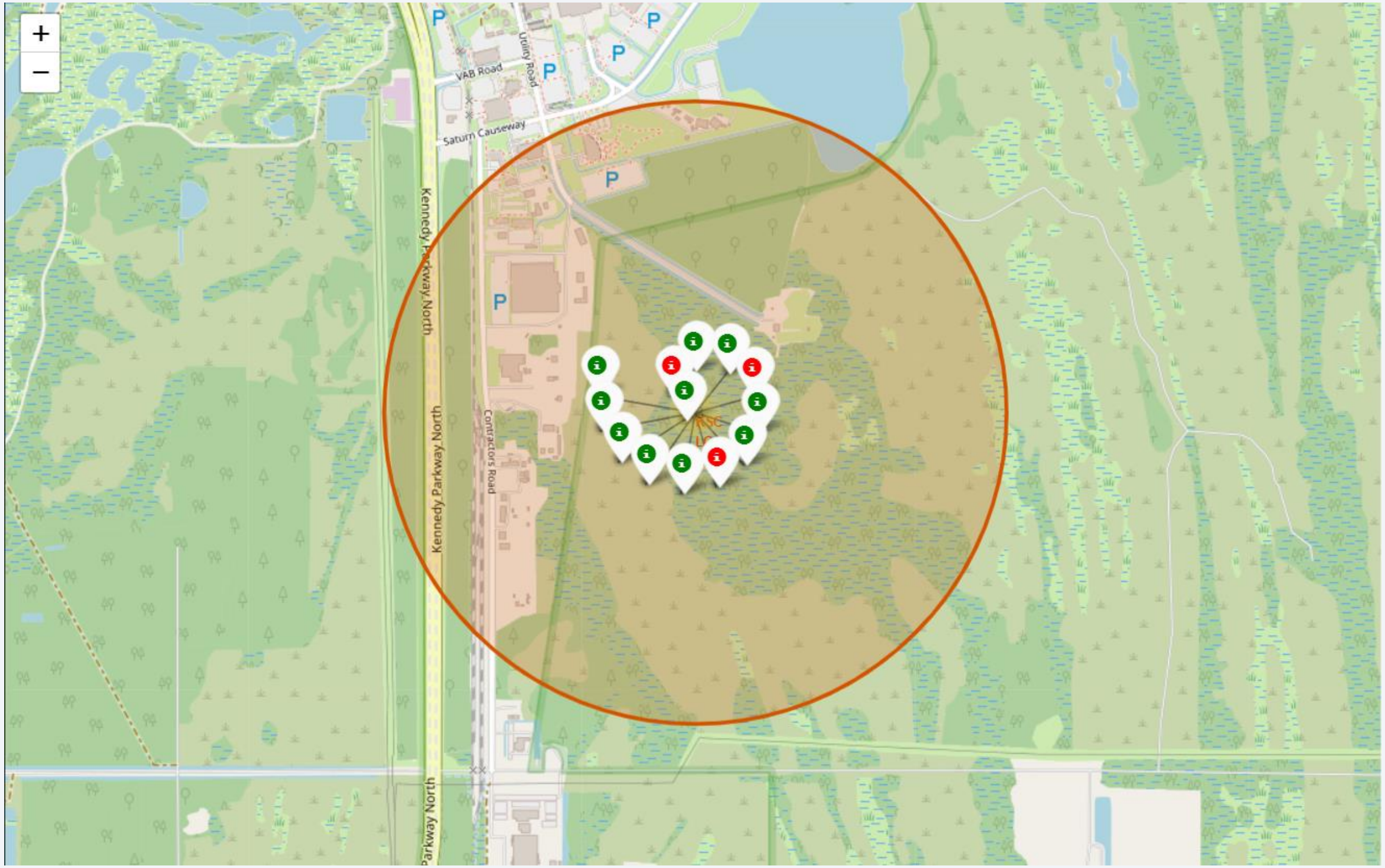
Launch site locations



Launch site locations

- From the map we can see that one of the launch sites is located near Santa Maria and the other three are close to Titusville.
- All the launch sites are situated in a close proximity to the coastline.
- The launch sites are located close to the equator line.
- This is advantageous since land near the equator moves at about 1650 km/hr while it is about 1180 km/hr elsewhere so the rocket moves 500km/hr faster once it is launched near the equator.
- This would also mean that the launched spacecraft will need less propellant or a more massive spacecraft can be launched.

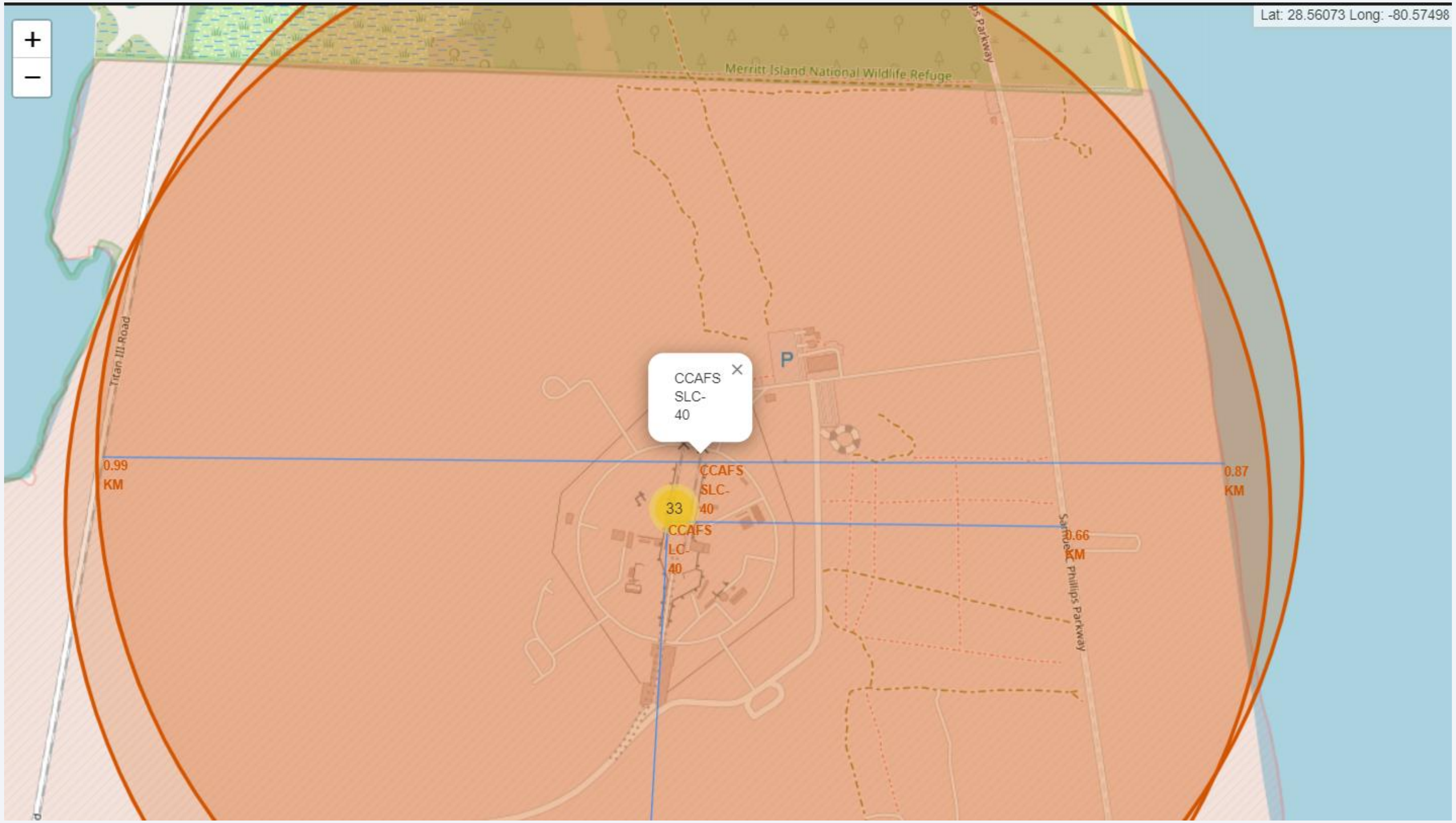
Launch outcomes



Launch outcomes

- The given map shows the launch outcomes for the site KSC LC-39A.
- It is located on the Merritt Island in Florida having the coordinates (28.573255, -80.646895)
- It shows the highest success rate (about 77%) of launches compared to the other launch sites.

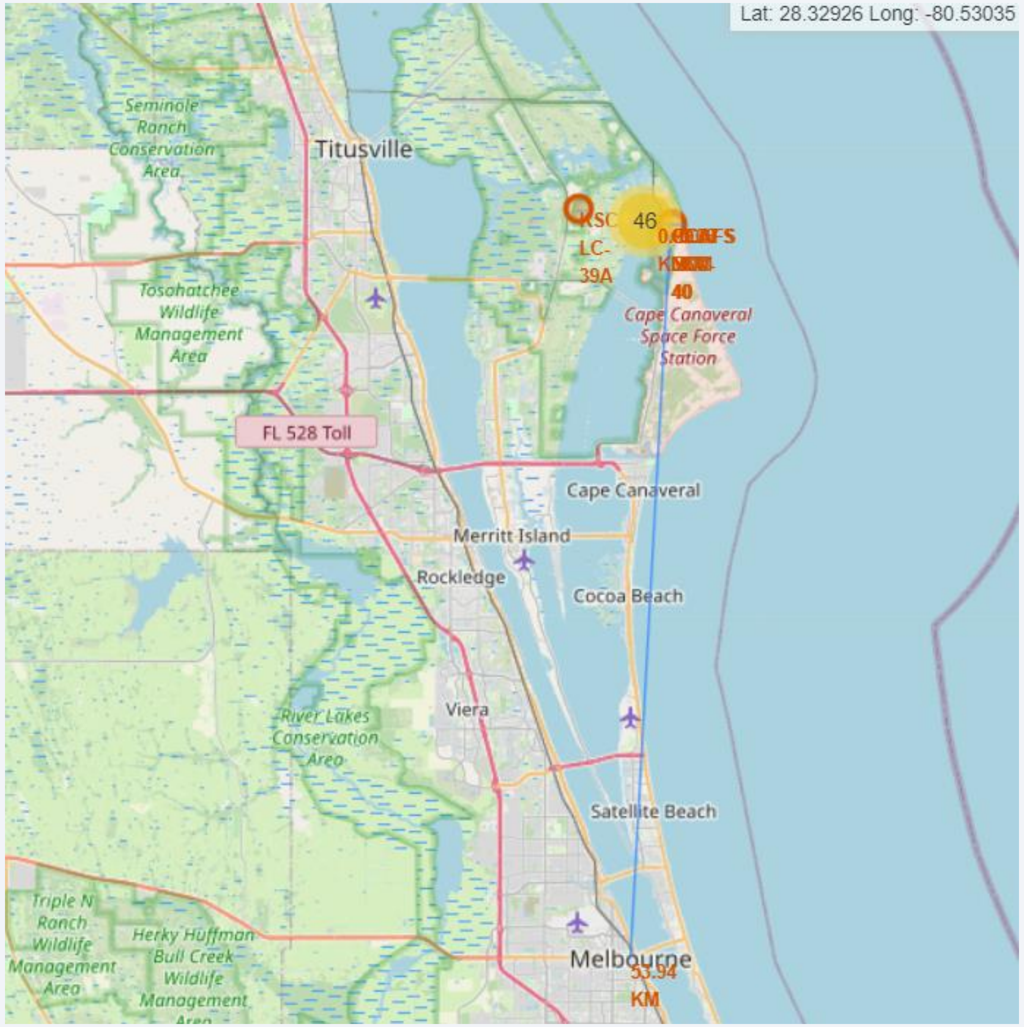
Measuring distances from the coastline, railways and city



Measuring distances from the coastline, railways and city

- The map shows the various distances calculated from the launch sites CCAFS SLC-40 and CCAFS LC-40.
- The measured distances have been indicted with orange numerics and a blue colored line has been drawn between each pair of points.
- It is found that the distance from the coastline is about 0.87km, the distance from the railways is about 1km and that from the nearest highway is 0.66km.
- From these findings, it is evident that the launch sites are located at a close proximity to the coastline, railways and the highway.
- The proximity to coastline is due to the fact that in the event of a crash there is no major impact to inhabited areas.
- The proximity to highway and railways is due to the fact that large number of rocket parts have to be brought in for assembly before they can be launched.

Measuring distances from the coastline, railways and city



- The map shows distance calculated from the launch sites CCAFS LC-40 and Melbourne.
- The measured distance is approximately 54kms.
- Launch sites are located far away from the city to mitigate the risks in the event of a catastrophic failure.

Section 4

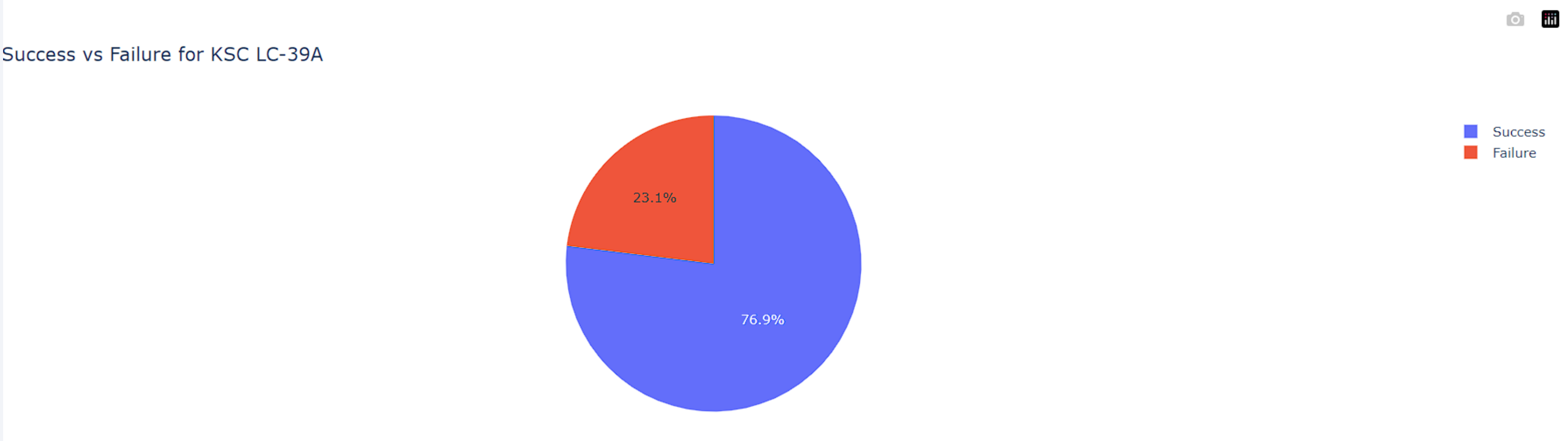
Build a Dashboard with Plotly Dash

Successful Launches (Overall)

Successful Launches (Overall)

- The pie chart shows the percentage of successful launches of the respective launch sites.
- It can be seen that launch site KSC LC-39A has the highest number of successful launches contributing to about 42 % of total successes.
- On the contrary, launch site CCAFS SLC-40 has a low success rate of 12.5%.

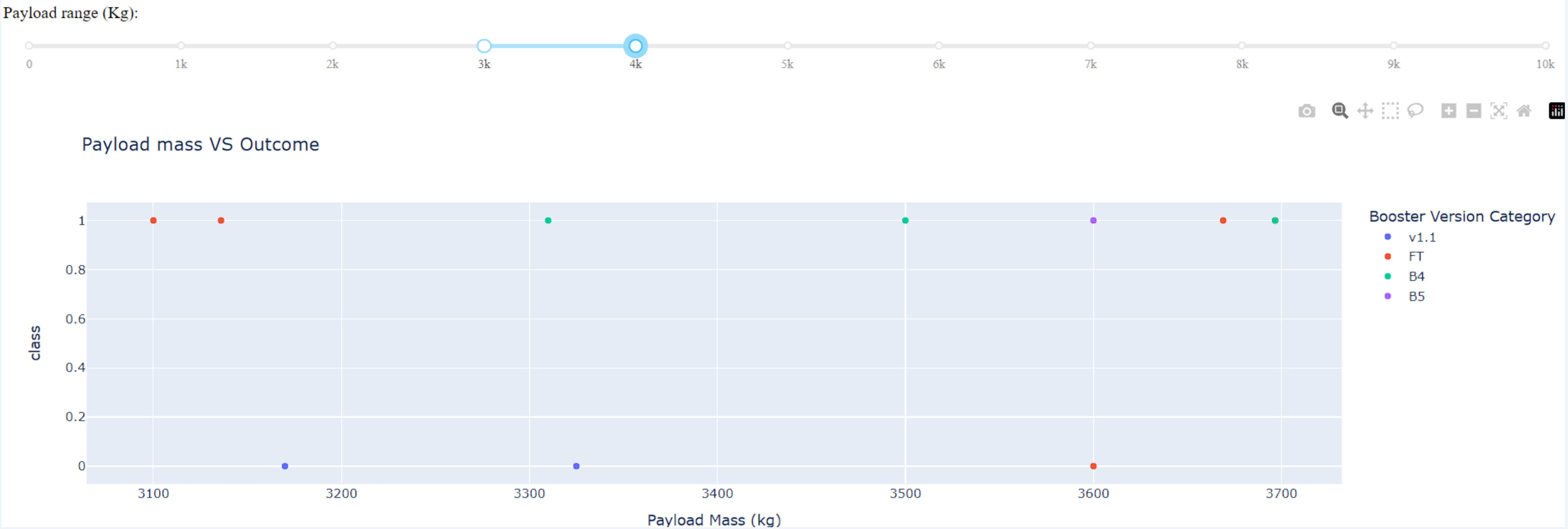
Highest Launch Success



Highest Launch Success

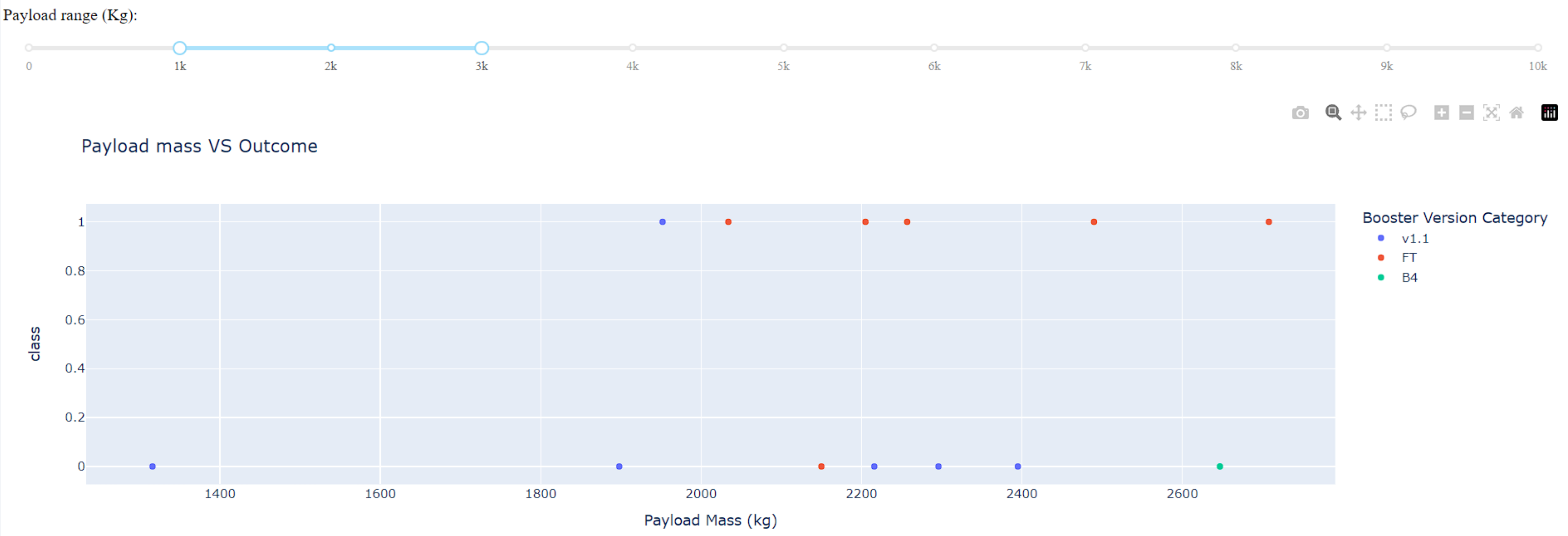
- The pie chart shows the launch success to failure ratio for the launch site KSC LC-39A.
- It has nearly a 77% success rate which is almost twice the success rate for other sites making it the most favorable one.
- It is found to have a successful launch when majority of its payload mass lies between 3000 and 5000 kg.
- The booster versions used are- FT, B4 and B5.

Relation of payload to launch success



Relation of payload to launch success

Relation of payload to launch success



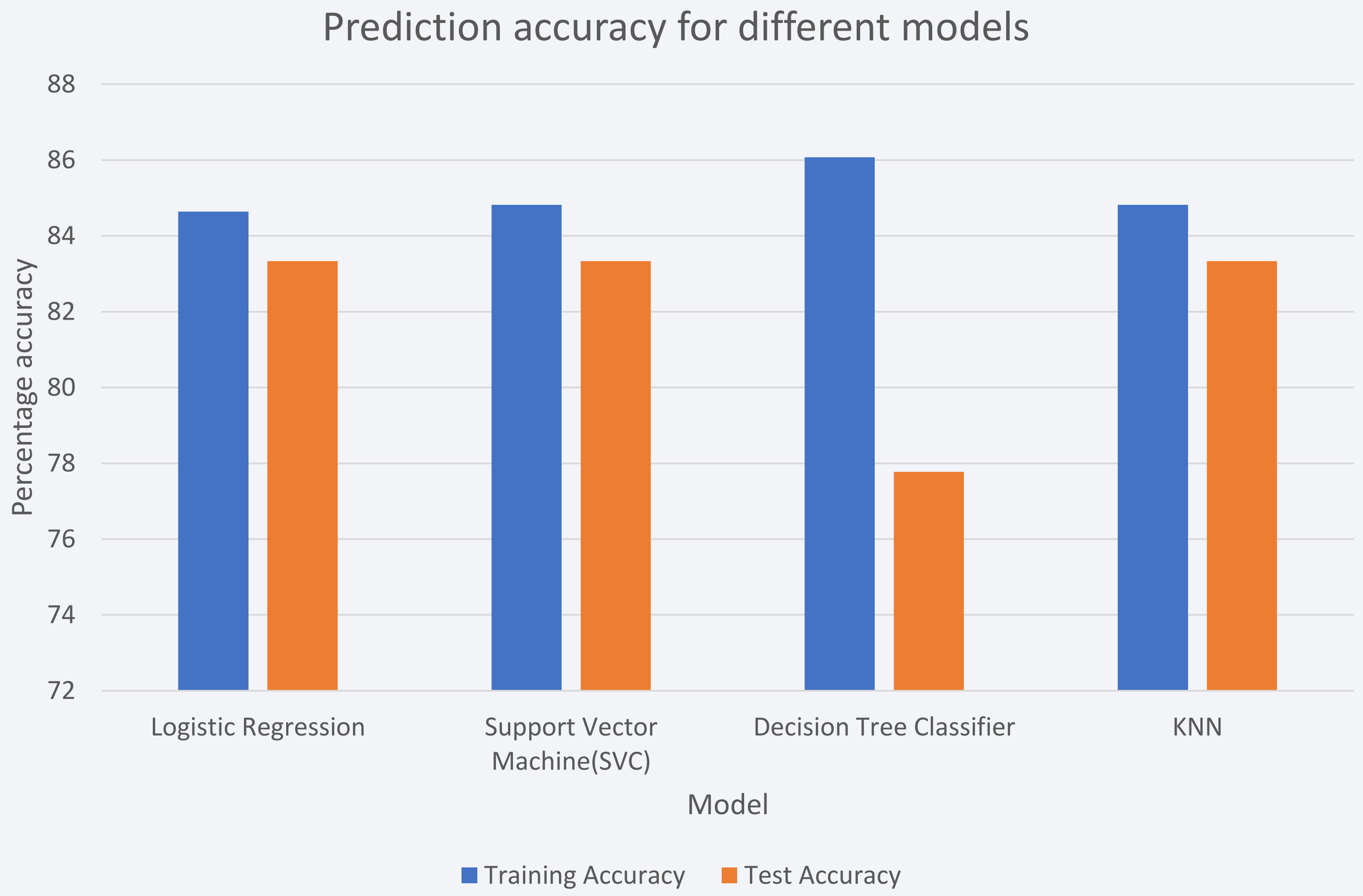
Relation of payload to launch success

- From the figures, we see that when payload mass lies in the range 3000 kg to 4000 kg, the success rate is the highest.
- When payload mass is between 4000 to 7000 kg, the success rate is very low.
- And there is a balance between launch success and failure in the range of 1000 to 3000 kg.

Section 5

Predictive Analysis (Classification)

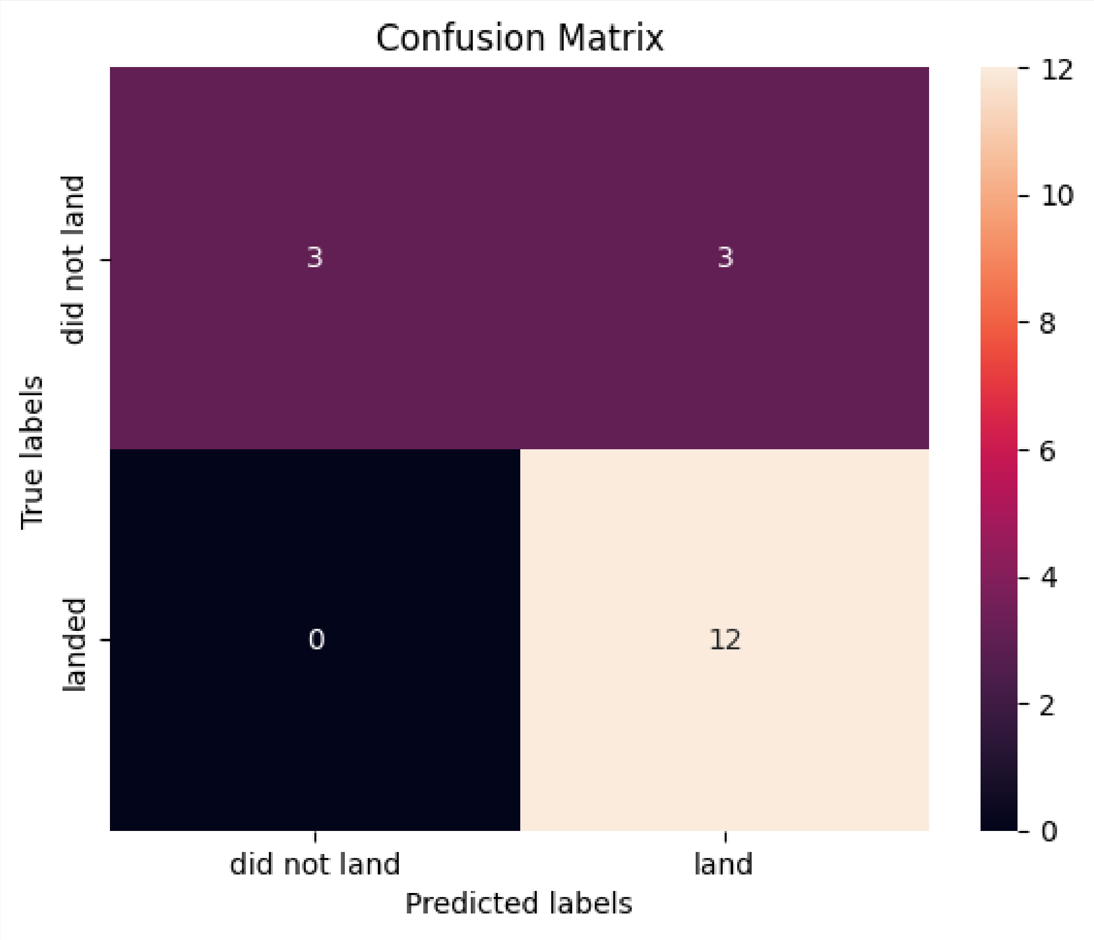
Classification Accuracy



Classification Accuracy

- From the bar chart, it is evident that all the models perform fairly well with a training accuracy of over 80 % and a test accuracy > 75%.
- Two models have the same highest training and test scores of 84.82 % and 83.33% respectively.
- Since SVM's in general perform well with small datasets given that proper hyperparameters have been selected, we will select the SVC as the best model for this project.

Confusion Matrix



From the confusion matrix of the SVC model, the following can be inferred:

- TP - 12 correct predictions were made that the rocket will land where the rocket did actually land.
- TN - 3 correct predictions were made that the rocket won't be able to land and it did not land.
- FP – 3 incorrect predictions where the rocket was predicted to land but it could not land successfully.
- FN – 0 incorrect predictions where the rocket was expected not to land but it landed in truth.

Conclusions

- Using predictive analysis, we can get an accuracy of 83.33% in determining whether a Falcon 9 rocket can land successfully in the first stage or not.
- Based on the prediction, the cost of each launch can be estimated.
- We can also determine if the first stage can be reused from the prediction.
- Significant cost reduction may be possible if the first stage can be reused.
- Rockets can be bought at relatively lower prices and more space operations can be conducted.
- The launch site KSC LC-39A has the highest success rate for launches and the success rate depends largely on the payload mass.

Acknowledgements

- Joseph Santarcangelo
- Yan Luo
- Saishruthi Swaminathan
- Hima Vasudevan
- Alex Aklson
- Polong Lin

Appendix

The datasets used in this project are listed.

- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv
- https://api.spacexdata.com/v4/launches/past
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv

Appendix

The notebooks, py files, SQL queries, charts and outputs of this project can be viewed from the following link.
https://github.com/sayanroy11/IBM-data_science-capstone

Please browse through each notebook for detailed understanding of the methodology.

Thank you!