IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Laura M.
2024-02-04

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected via the SpaceX API.
  - Data was also obtained from the Wikipedia site, through the use of webscraping and the Beautiful Soup libraries.
  - Data was placed into Pandas data frames to be normalized and queried.
- Summary of all results
  - Exploratory data analysis was performed using visualizations and SQL queries.
  - Folium and Plotly Dash were used to create visual and interactive analysis.
  - Classification models were created and compared to perform predicitive analytics.

Introduction

This project aims to predict Falcon 9 first stage launch success or failure. The Falcon 9 rocket launches are publicized on the SpaceX website. While other providers' cost to launch can be upwards of 62 million dollars, SpaceX boasts a cost of only 62 million dollars, because SpaceX is able to reuse the first stage of the launch.

The goal is to determine whether or not the first stage will land, so that the cost to launch can be estimated. This information will be useful to alternate companies that may want to bid against SpaceX.

Section 1

Methodology

Methodology

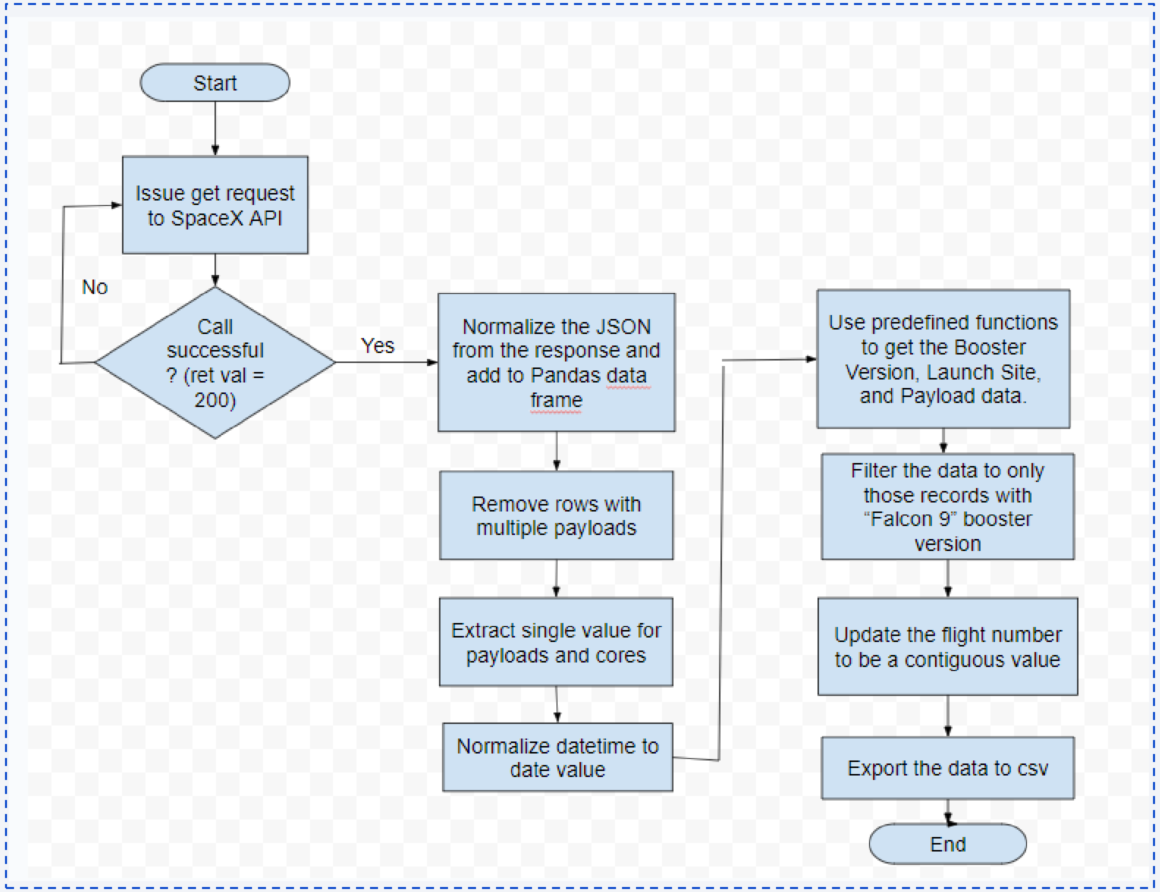
- Data collection methodology:
  - The data for this project was collected via REST API and web scraping. The data was converted to Pandas data frames for analysis.
- Perform data wrangling
  - The collected data in the Pandas data frame was analyzed and manipulated through python code and sql queries, and meaningful patterns were extracted to be visualized through scatter plots and charts.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The known data was split into training and testing data. Four different classification models where then used: SVM, Classification Trees, Logistic Regression, and K Nearest Neighbor.

Data Collection

- Data was first collected via REST API through the SpaceX website. The JSON from the response was normalized and then added to a Pandas data frame. Rows that had multiple payloads were removed. Payloads and cores were extracted to a single value, and the datetime value was normalized to just the date value. Predefined functions were used to get the Booster Version, Launch Site, and Payload data. The data was filtered to just those records using the "Falcon 9" booster version, and the flight number was reset to be a contiguous value.
- In the webscraping collection, data was obtained through a wikipedia page that lists Falcon 9 and Falcon Heavy launches. Beautiful Soup was used to parse the html response from the page and extract the data. The column names were extracted from the <th> elements, and then a Pandas data frame was created and populated. Null values were replaced with "unknown."

Data Collection – SpaceX API

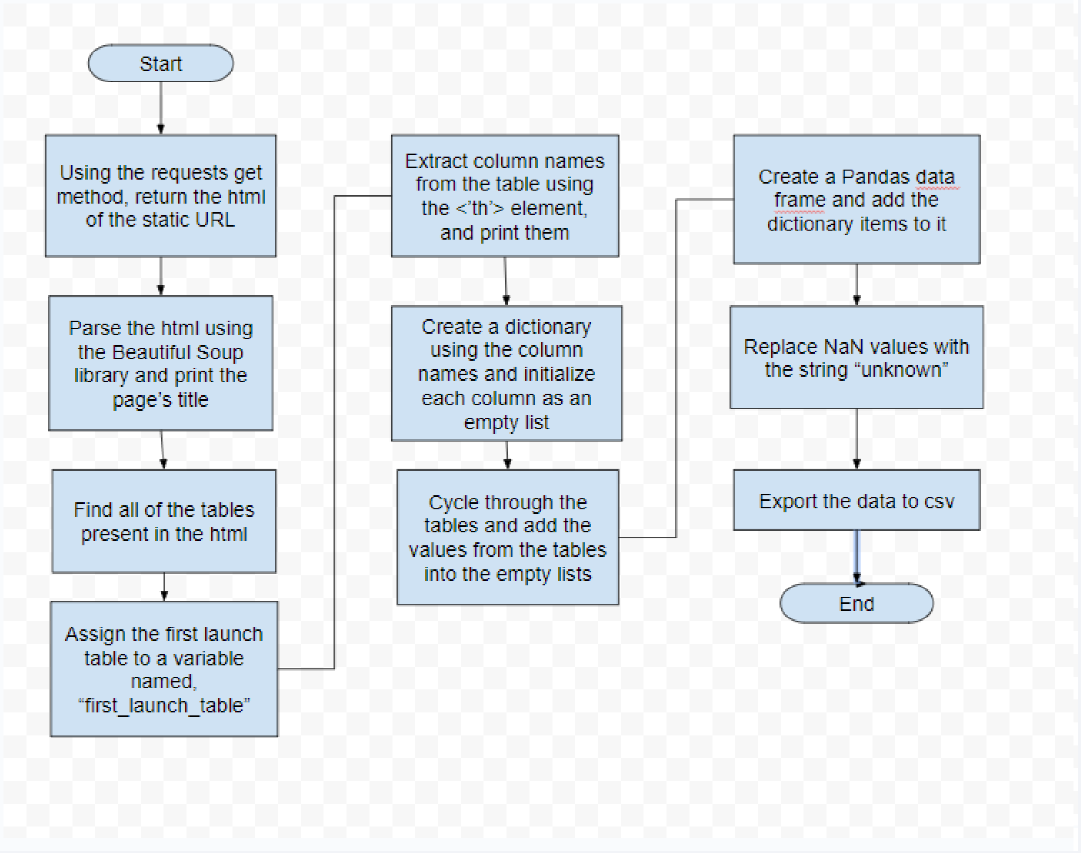
- The SpaceX data was collected using a get request to the SpaceX API. The methodology and data manipulations are outlined in the flow chart on the right.
- External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Start
Issue get request to SpaceX API
No
Call successful ? (ret val = 200)
Yes
Normalize the JSON from the response and add to Pandas data frame
Remove rows with multiple payloads
Extract single value for payloads and cores
Normalize datetime to date value
Use predefined functions to get the Booster Version, Launch Site, and Payload data.
Filter the data to only those records with "Falcon 9" booster version
Update the flight number to be a contiguous value
Export the data to csv
End

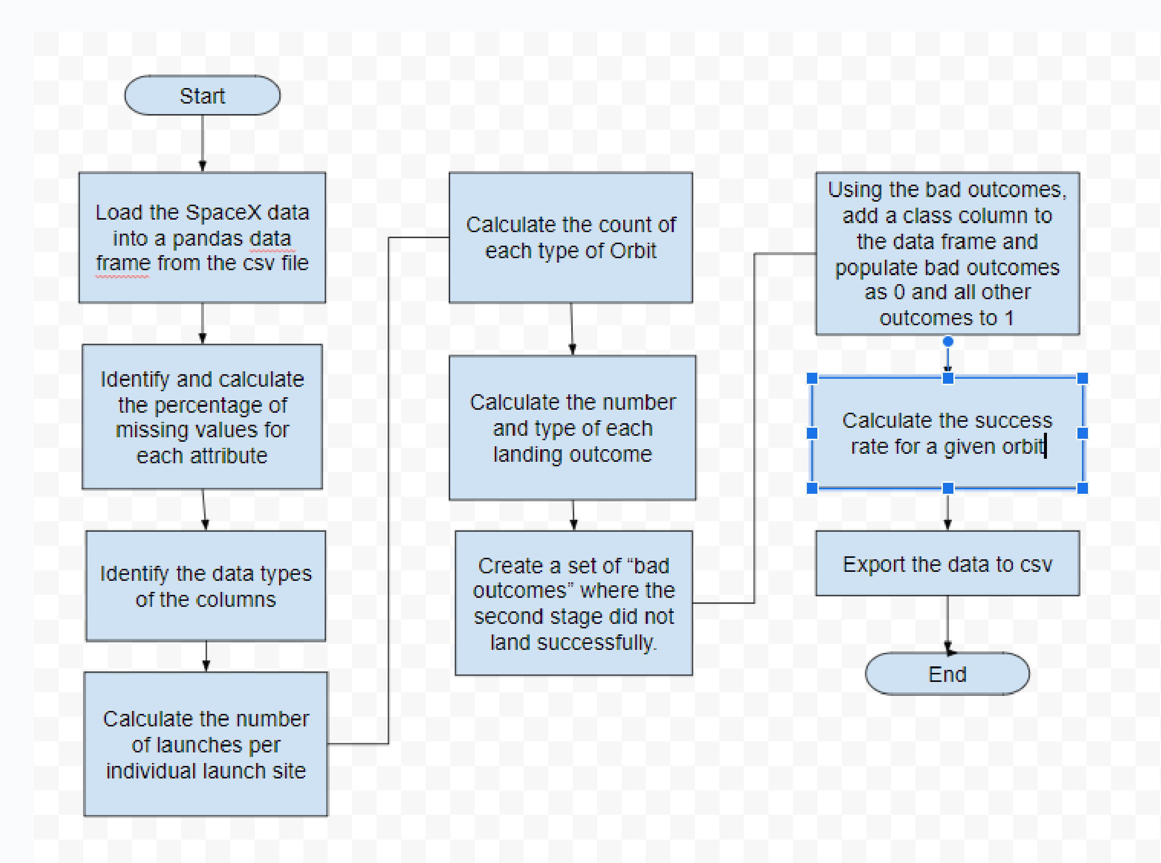
Data Collection - Scraping

- The wikipedia data was collected and its html was parsed using the Beautiful Soup library. The data was extracted to a Pandas data frame, the NaN values were updated, and the data was exported to csv, as outlined in the flowchart on the right.
- External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/jupyter-labs-webscraping.ipynb


Start
Using the requests get method, return the html of the static URL
Parse the html using the Beautiful Soup library and print the page's title
Find all of the tables present in the html
Assign the first launch table to a variable named, "first_launch_table"
Extract column names from the table using the <'th'> element, and print them
Create a dictionary using the column names and initialize each column as an empty list
Cycle through the tables and add the values from the tables into the empty lists
Create a Pandas data frame and add the dictionary items to it
Replace NaN values with the string "unknown"
Export the data to csv
End

Data Wrangling

- After the SpaceX data was loaded from a csv, calculations were made to determine missing attribute values by percentage. A number of characteristics were summed by their key value, and a class (of successful or not [1,0]) was calculated and added to the data frame, as outlined in the flow chart on the right.
- External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb


Start

Load the SpaceX data into a pandas data frame from the csv file

Identify and calculate the percentage of missing values for each attribute

Identify the data types of the columns

Calculate the number of launches per individual launch site

Calculate the count of each type of Orbit

Calculate the number and type of each landing outcome

Create a set of "bad outcomes" where the second stage did not land successfully.

Using the bad outcomes, add a class column to the data frame and populate bad outcomes as 0 and all other outcomes to 1

Calculate the success rate for a given orbit

Export the data to csv

End

EDA with Data Visualization

- Scatter plots were created to determine any correlation between launch outcome and the relationship between: flight number and payload mass, flight number and launch site, and payload and launch site. A bar chart was created to examine the success rates of different Orbit types. Additional scatter plots were created to examine success by flight number and orbit, as well as payload and orbit. Also, the average launch success rate by year was charted in a line chart.

- External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Sql queries were performed for the following result sets:

- Displayed the names of unique launch sites.
- Displayed 5 records where the launch site name began with 'CCA.'
- Displayed the total payload mass carried by boosters launchd by NASA (CRS).
- Displayed the average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome in a ground pad was achieved.
- Listed the names of the boosters which had success in drone ship and had a payload mass greater than 4,000 kg, but less than 6,000 kg.
- Listed the total number of successful and failed mission outcomes.
- Listed the names of the booster versions which had carried the maximum payload mass.
- Listed the months, failed landing outcomes in drone ship, booster versions, and launch sites for the year 2015.
- Listed the total of each type of landing outcome between the dates 2010-06-04 and 2017-03-20, ranked by descending count.

External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

With the Folium maps, the following were accomplished:

- Marked all launch sites on the map: These were marked with circles and labels so that the launch sites could be easily identified and distinguished from one another.
- Marked all successful/failed launches for each site: Marker clusters were created using the class field to designate red markers for failed launches and green markers for successful launches.
- Calculated distances between a launch site to its proximities: Lines were created from one of the launch sites to the nearest coastline, railway, highway, and city, and the distances of these points was displayed as a label, to show the proximity of the launch site to each of these items.

External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_Final Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

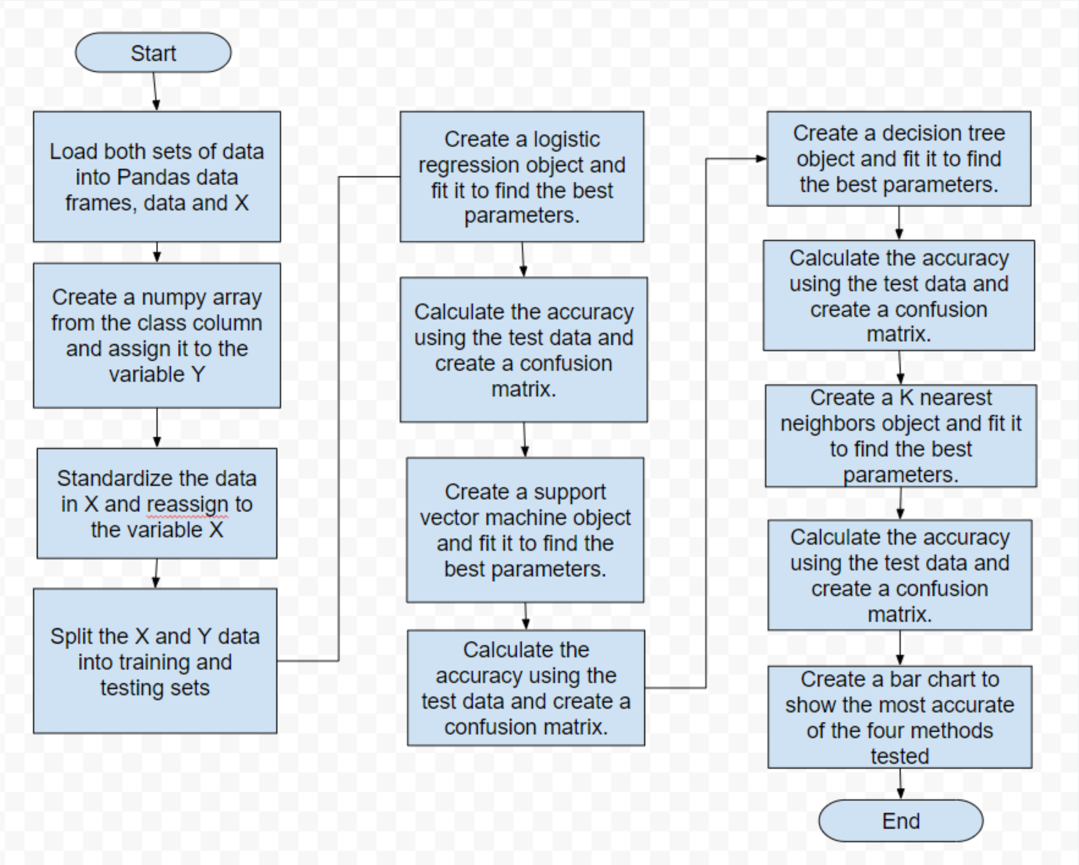
- A pie chart was added to the dashboard, which was driven by a drop-down to select either "All sites" or one of the 4 specific site names. This would help determine which site had the largest percentage of successful launches, and also which specific site hasd the highest launch success rate.

- Additionally, a scatter plot was added showing success rate by payload mass, where the booster version category was designated by color. The payload range was driven by a range slider above it, which determined the payload range shown in the scatter plot. Through this scatter plot, it can be determined which payload ranges had the highest and lowest success rates, as well as which booster version had the highest success rate.

- External reference to complete Plotly Dash code: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The classification models were built and evaluated using the methods in the flow chart on the right. The best parameters were determined using a GridSearchCV object, and the best score was also determined via the GridSearchCV. The test data was evaluated using the score method and the confusion matrices were ploted from the test data.
- External reference to complete Jupyter notebook: https://github.com/theLamaMoos/SpaceX_FinalProject/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Start

Load both sets of data into Pandas data frames, data and X

Create a numpy array from the class column and assign it to the variable Y

Standardize the data in X and reassign to the variable X

Split the X and Y data into training and testing sets

Create a logistic regression object and fit it to find the best parameters.

Calculate the accuracy using the test data and create a confusion matrix.

Create a support vector machine object and fit it to find the best parameters.

Calculate the accuracy using the test data and create a confusion matrix.

Create a decision tree object and fit it to find the best parameters.

Calculate the accuracy using the test data and create a confusion matrix.

Create a K nearest neighbors object and fit it to find the best parameters.

Calculate the accuracy using the test data and create a confusion matrix.

Create a bar chart to show the most accurate of the four methods tested

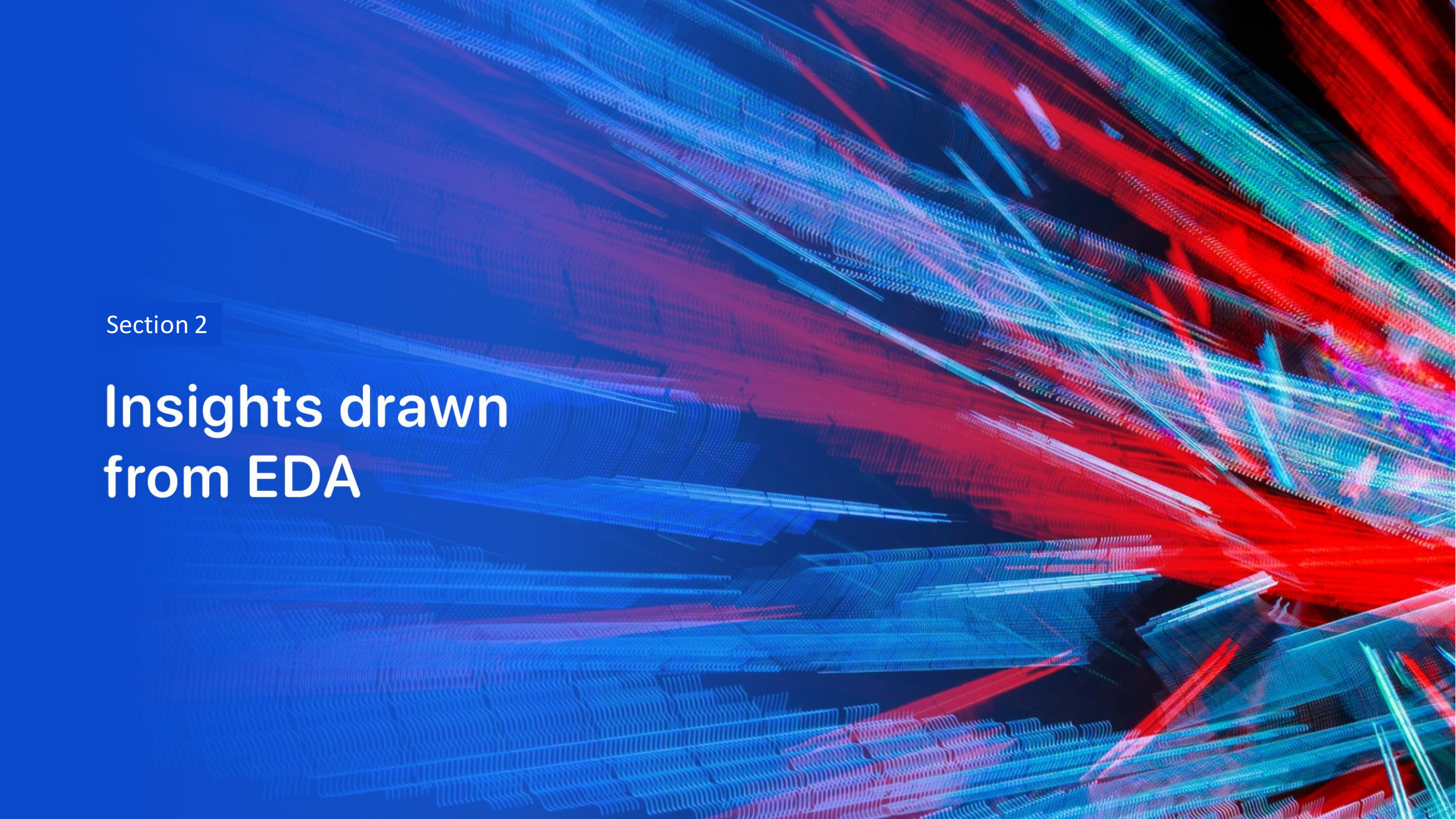
End

Results

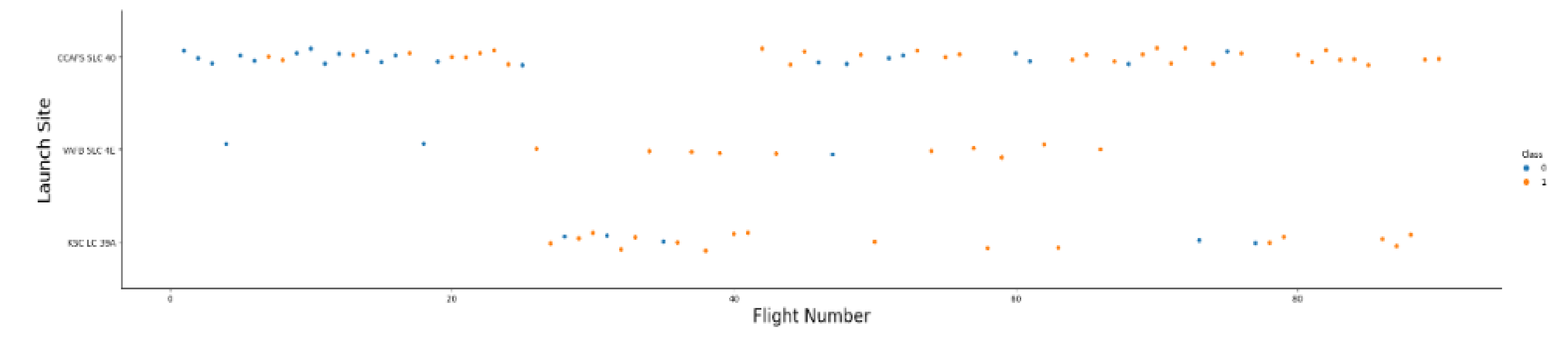
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

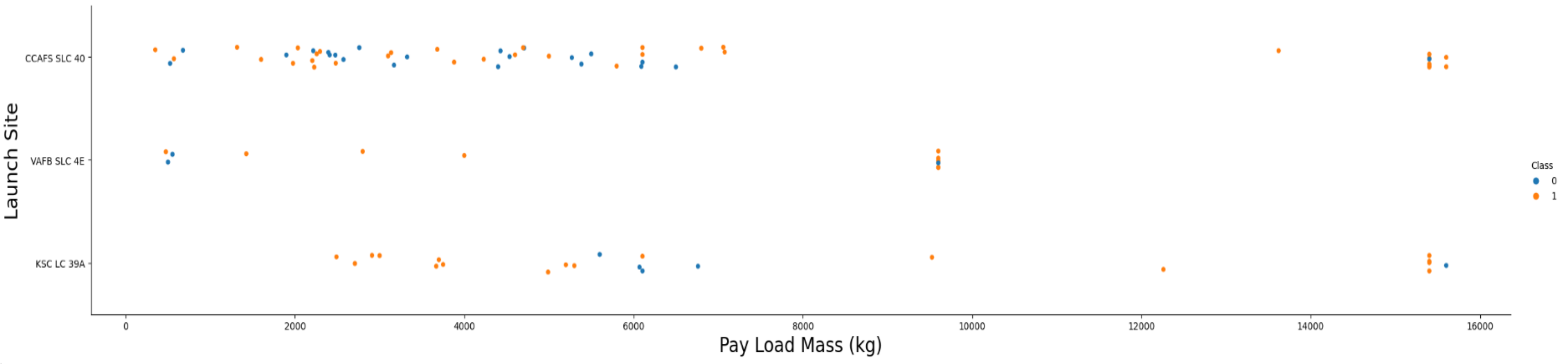


Flight Number vs. Launch Site



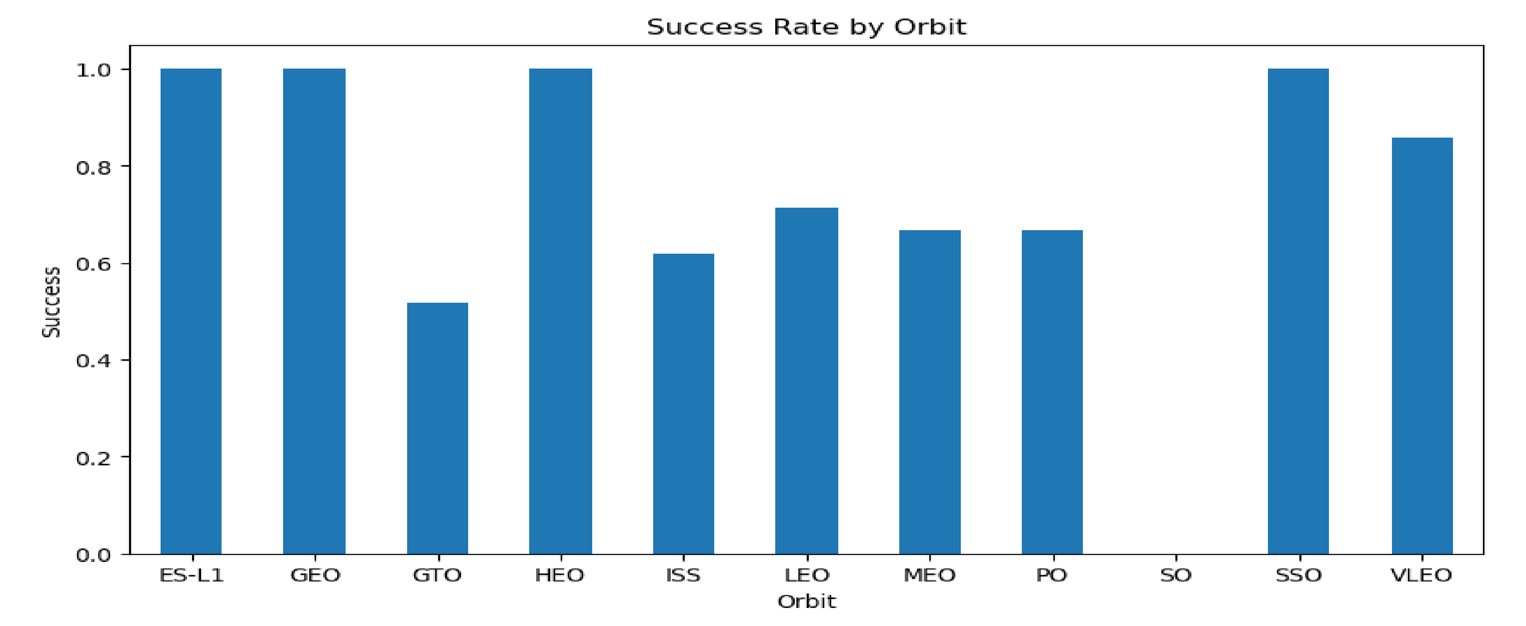
- For all three sites, there seems to be some correlation between the increase of flight number and the increases of successful launches.
- However, for the site, CCAFS SLC 40, there are a fair number of failed launches in the increased flight number, and it should be noted that all three sites have some rate of failure as the flight number increases.

Payload vs. Launch Site



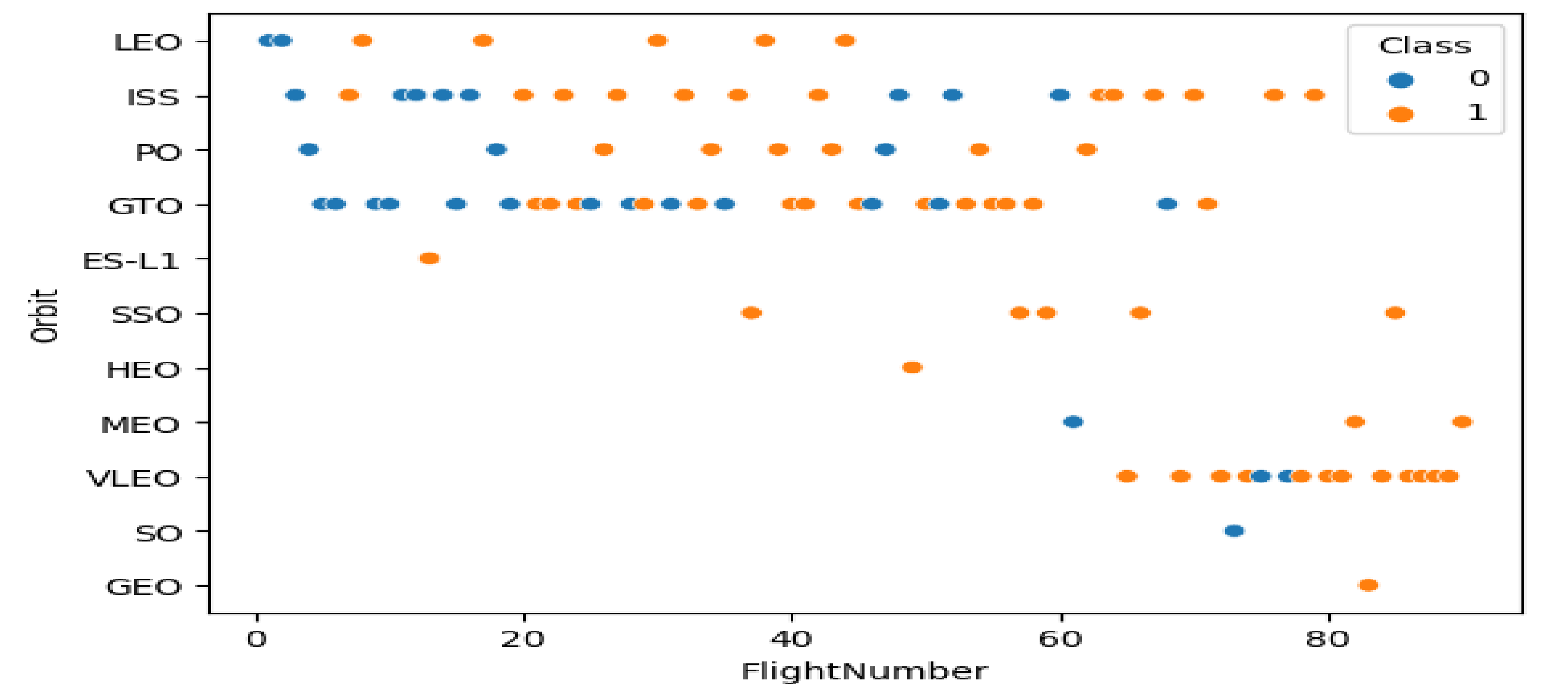
- For the site CCAFS SLC 40, there does not seem to be a correlation for success or failure based on the Payload Mass.
- For the site VAFB SLC 4E, the majority of Payload Mass that were higher than 1,000 kg succeeded, with the exception of 1 failure close to 10,000 kg.
- For the site KSC LC 39A, we see mostly successful launches, with the exception of 4 failures around the 6,000 kg mark, and one failure close to 16,000 kg.

Success Rate vs. Orbit Type



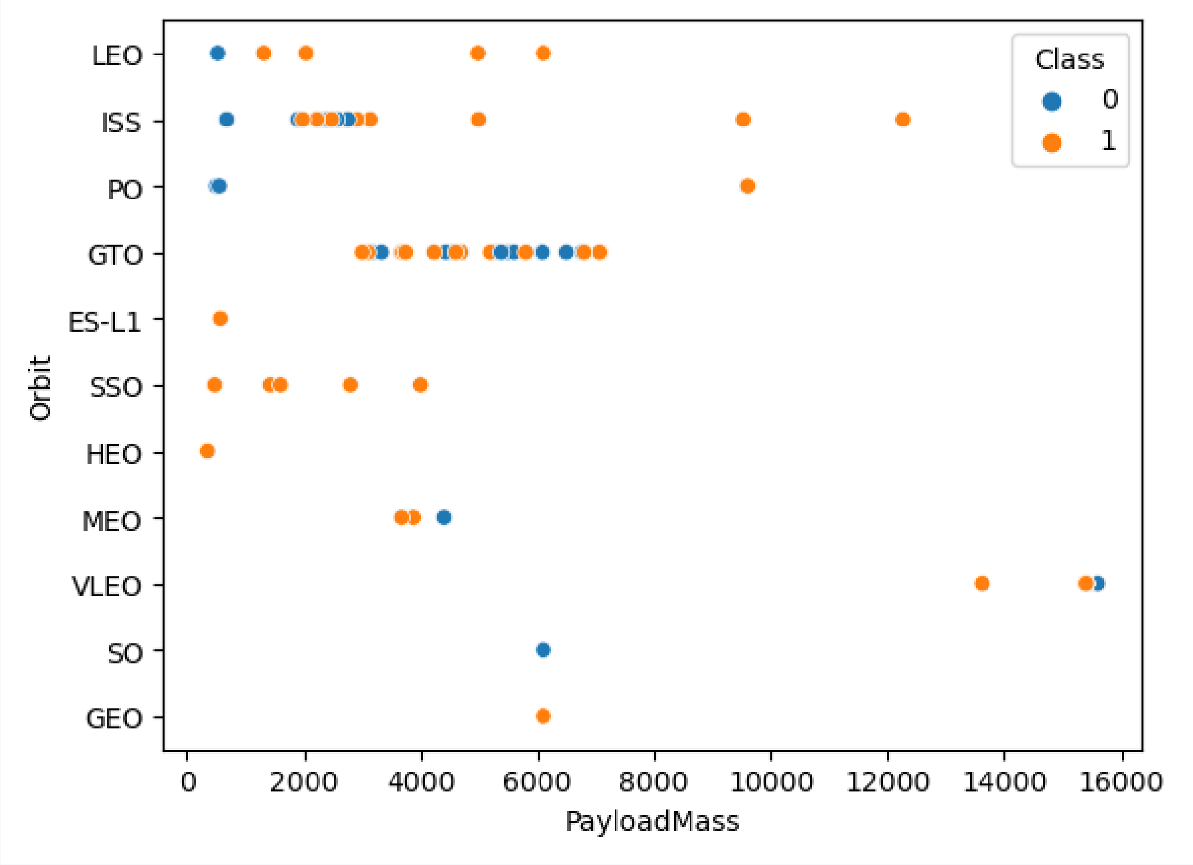
- The following Orbits had the highest success rate: ES-L1, GEO, HEO, and SSO. VLEO had a moderately high success rate, just above 80%.
- Most other Orbits had a mid-range success rate, between about 50% and 70%.
- The Orbit SO had no recorded outcomes.

Flight Number vs. Orbit Type



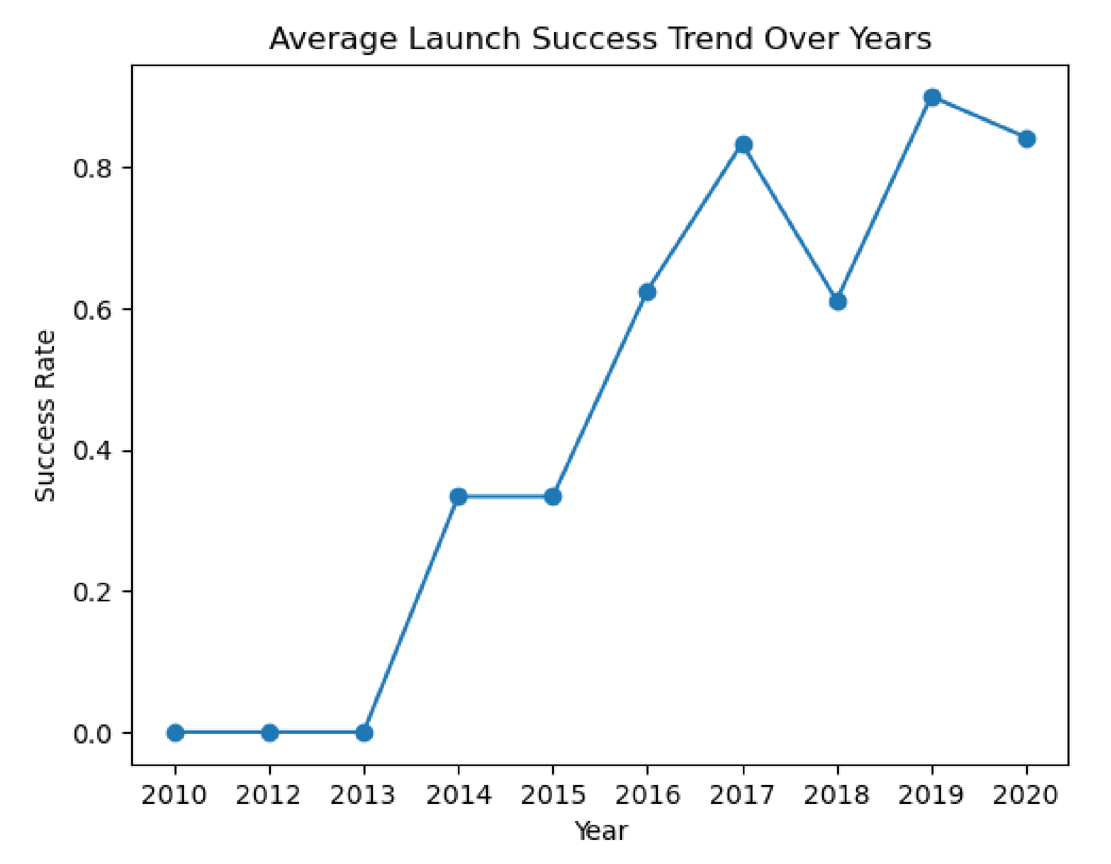
- Some of the Orbit types seem to have an increased success rate as the flight number increases. These include: LEO, PO, and MEO.
- Some Orbit types have only successes, such as ES-L1, SSO, HEO, and GEO.
- Those Orbit types having a distributed mix of failures and successes include: ISS, GTO and VLEO.
- The Orbit type SO does not have any successful outcomes.

Payload vs. Orbit Type



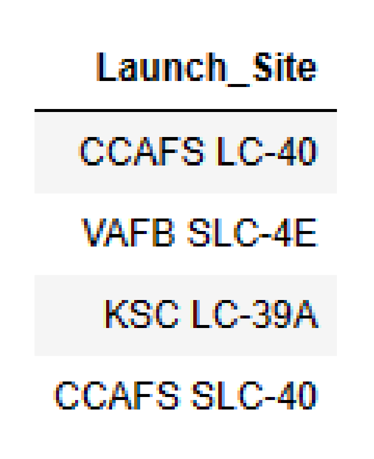
- For the following Orbits, the success rate appears to increase as the payload mass increases: LEO, ISS, and PO.
- The following Orbits had decreased successes as the payload mass increased: MEO and VLEO.
- The following orbits had only successes: ES-L1, SSO, HEO, and GEO.
- The SO orbit had only failures.
- The GTO orbit had a mixture of success and failure, regardless of Payload mass.

Launch Success Yearly Trend



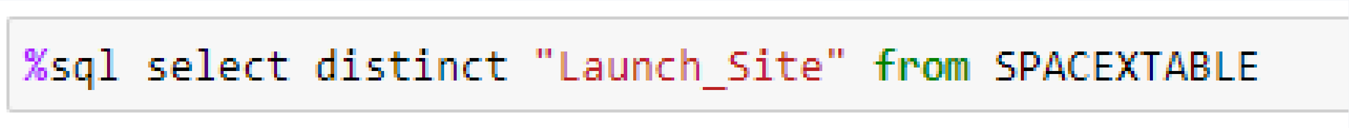
- The general trend for launch successes appears to increase with the progression of time.
- While there are some dips in this trend, it has remained over 50 % since 2016.

All Launch Site Names



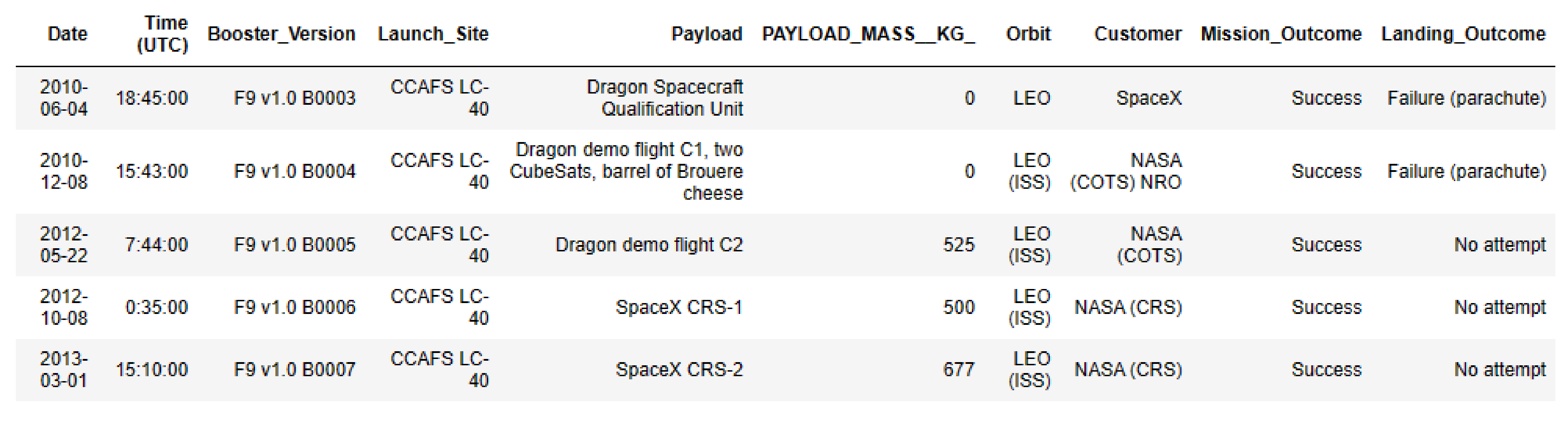
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- These are all the launch sites from the data. These were compiled using a sql statement with a "distinct" clause.



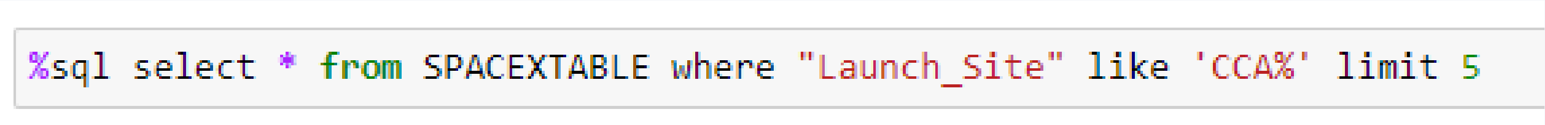
```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

Launch Site Names Begin with 'CCA'



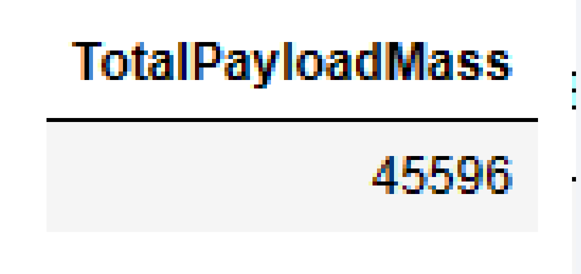
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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 first 5 records for launch sites beginning with CCA. Records were obtained using a sql query with the 'limit' clause and a 'where' condition.



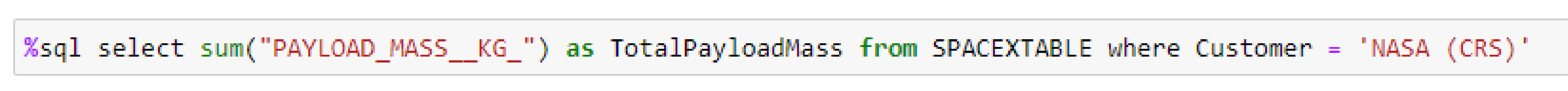
```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

Total Payload Mass



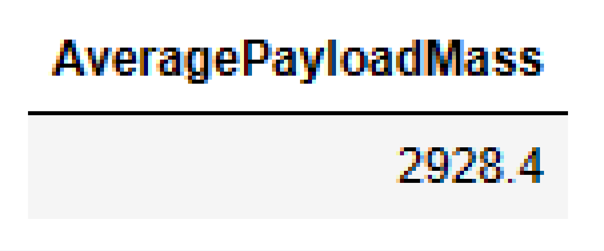
TotalPayloadMass
45596

- This is the calculated payload mass for boosters from NASA. The amount was obtained with sql using a 'sum' function and a 'where' clause to limit the records to those belonging to NASA.



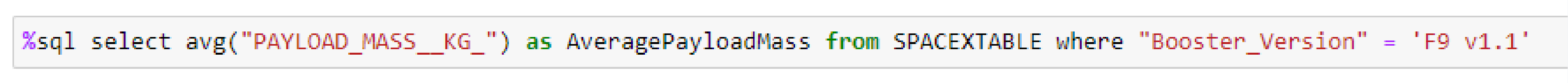
```
%sql select sum("PAYLOAD_MASS__KG_") as TotalPayloadMass from SPACEXTABLE where Customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1



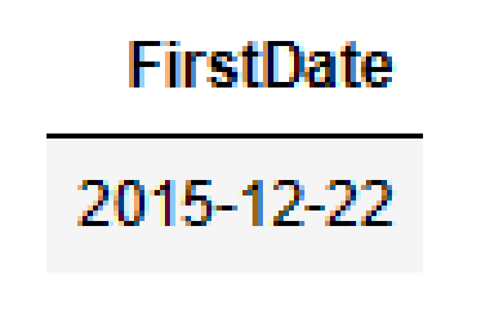
AveragePayloadMass
2928.4

- This is the calculated average payload mass carried by booster version F9 v1.1. This was calculated with a sql query using the 'avg' function, and was filtered to the booster type with a 'where' clause.



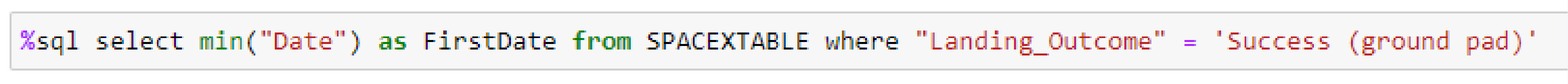
```
%sql select avg("PAYLOAD_MASS__KG_") as AveragePayloadMass from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'
```

First Successful Ground Landing Date



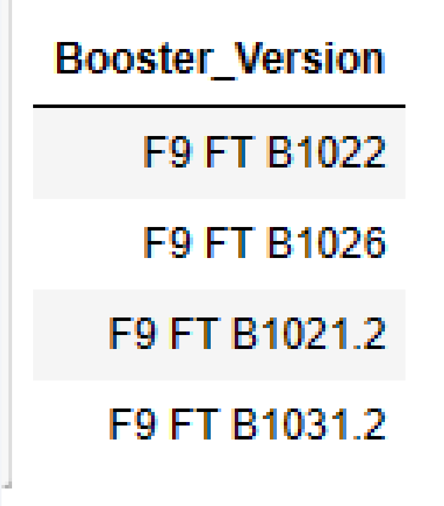
FirstDate
2015-12-22

- The date of the first successful landing outcome on ground pad was December 22nd, 2015. This result was obtained by a sql query, using the 'min' function, and was filtered with a 'where' clause to only return successful landings on a ground pad.



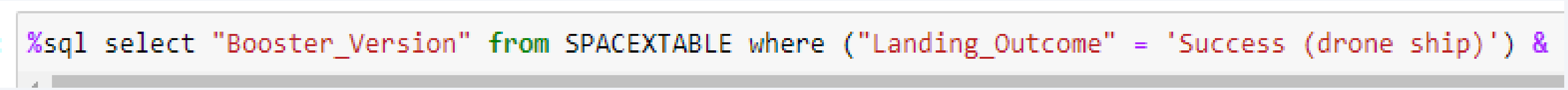
```
%sql select min("Date") as FirstDate from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

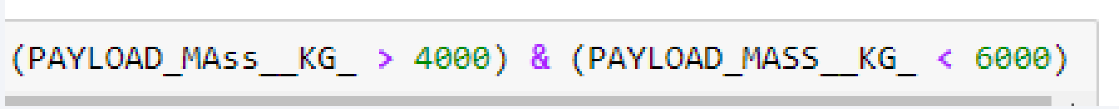


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 kg but less than 6,000 kg. The results were obtained with a sql query specifying multiple conditions in the 'where' clause.

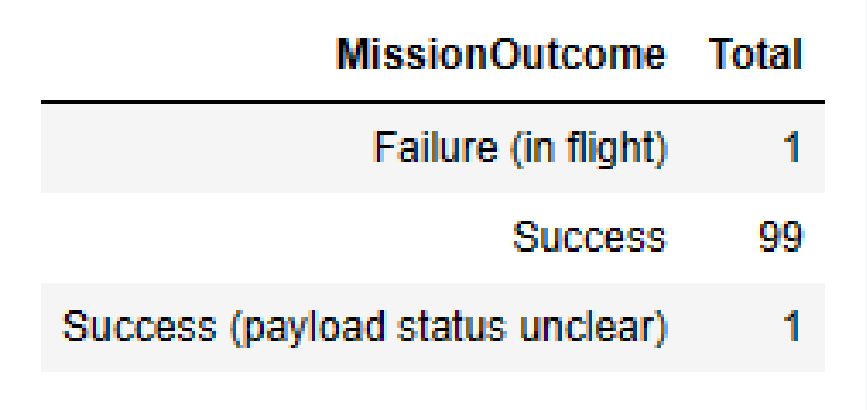


```
%sql select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" = 'Success (drone ship)') &
```



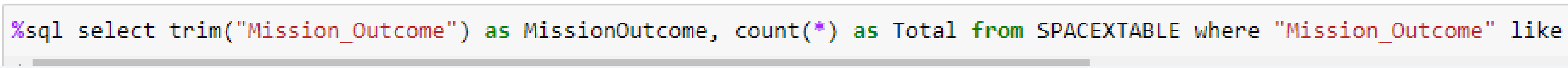
```
(PAYLOAD_MAss__KG_ > 4000) & (PAYLOAD_MASS__KG_ < 6000)
```

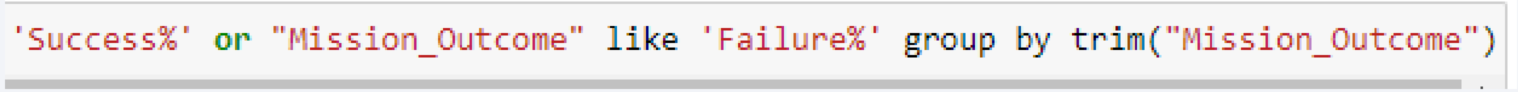
Total Number of Successful and Failure Mission Outcomes



MissionOutcome	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

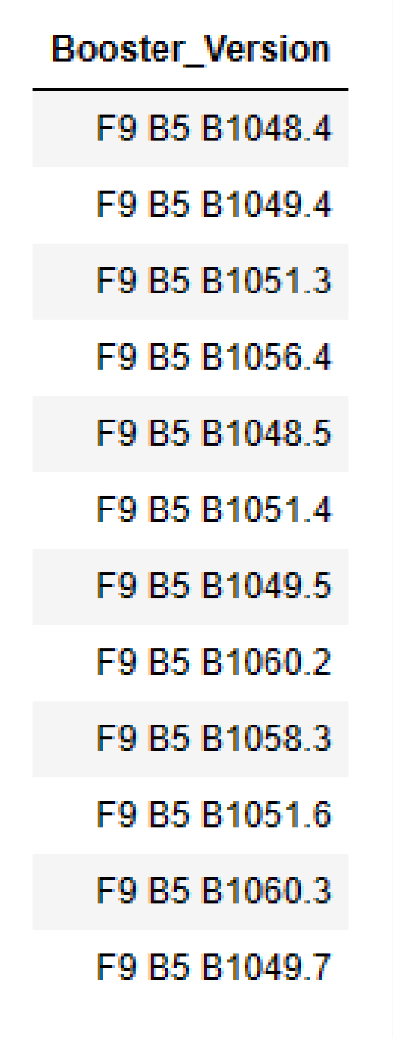
- Presented here are the totals for the types of mission outcomes, which are predominantly successful. The data includes one failure that occurred in flight, with all other missions reporting success.





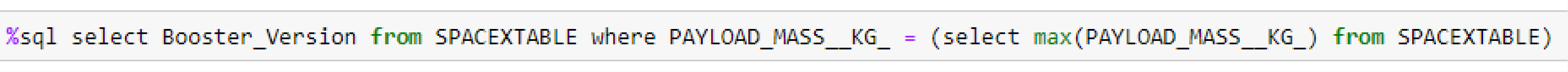
```
%sql select trim("Mission_Outcome") as MissionOutcome, count(*) as Total from SPACEXTABLE where "Mission_Outcome" like
'Success%' or "Mission_Outcome" like 'Failure%' group by trim("Mission_Outcome")
```

Boosters Carried Maximum Payload



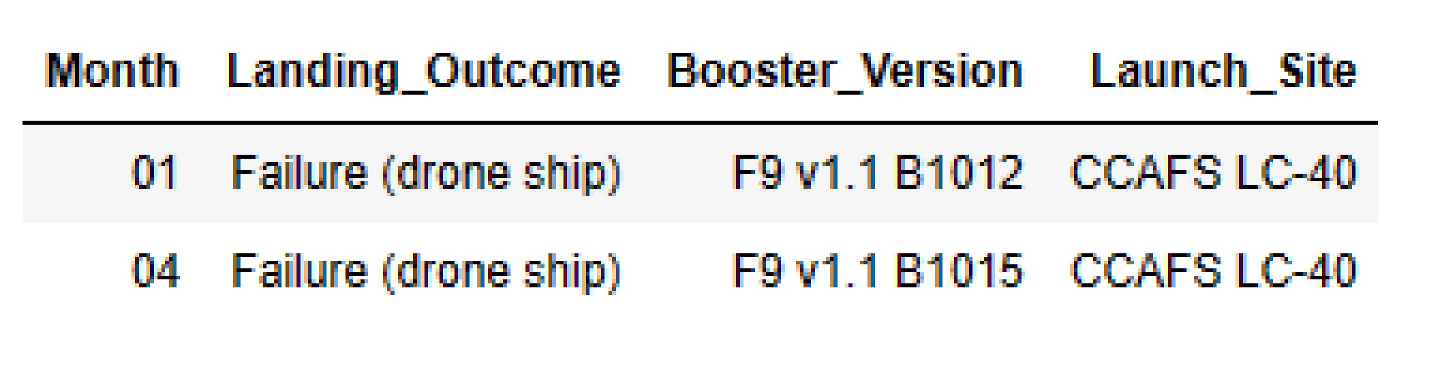
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- These are the boosters that have carried the maximum payload. This data was obtained through a sql statement, which used a subquery to determine the maximum value of payload.



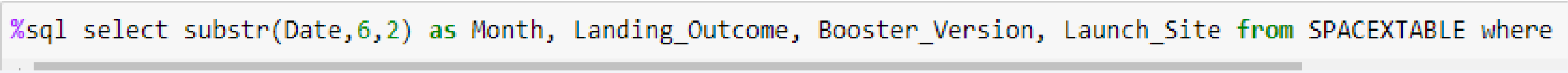
```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

2015 Launch Records

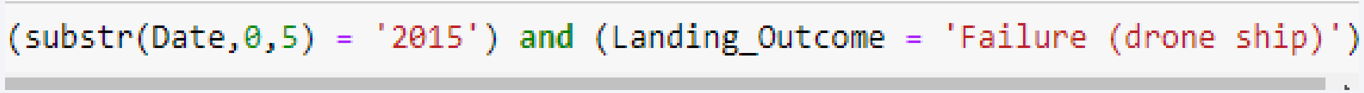


Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Listing of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, along with the month of occurrence. This data was obtained through a sql query, using a substring function to produce month and year from the original date field, and which used a 'where' clause to limit records to those from 2015 and with a landing outcome of 'Failure (drone ship)'.

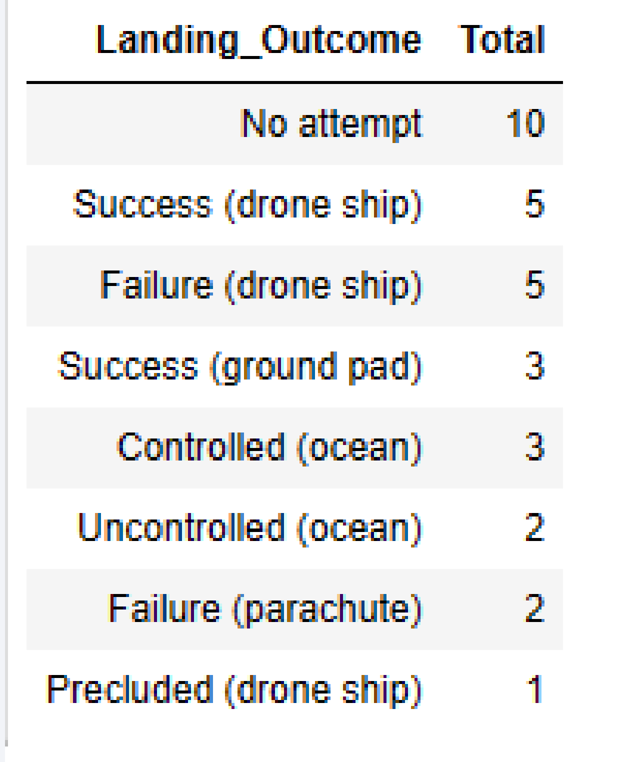


```
%sql select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where
```



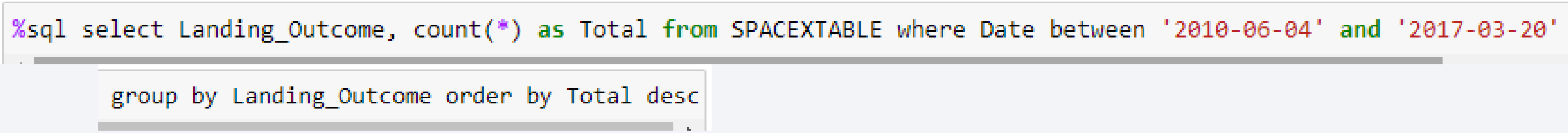
```
(substr(Date,0,5) = '2015') and (Landing_Outcome = 'Failure (drone ship)')
```

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

- This is the count of landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order
- This data was compiled using a sql query with a count function, a where clause to limit the dates, and both a group by and order by to ensure both proper aggregation as well as the desired descending order of counts.

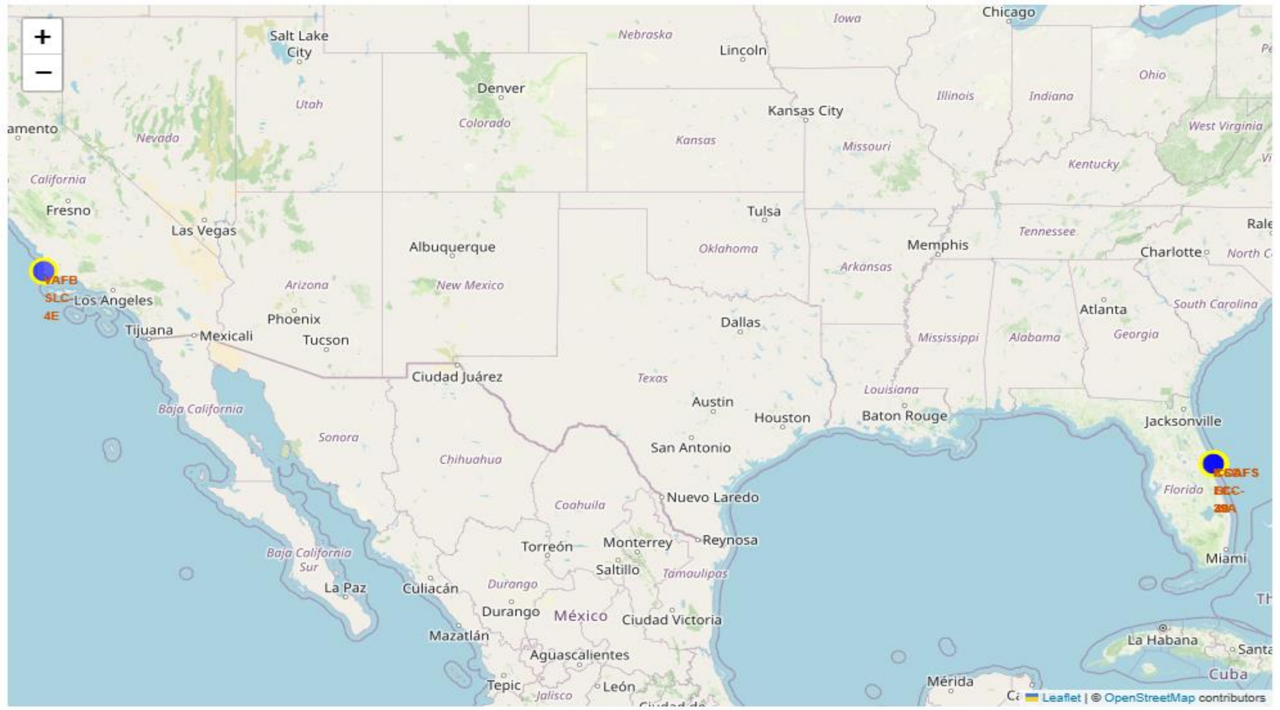


```
%sql select Landing_Outcome, count(*) as Total from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by Total desc
```

Section 3

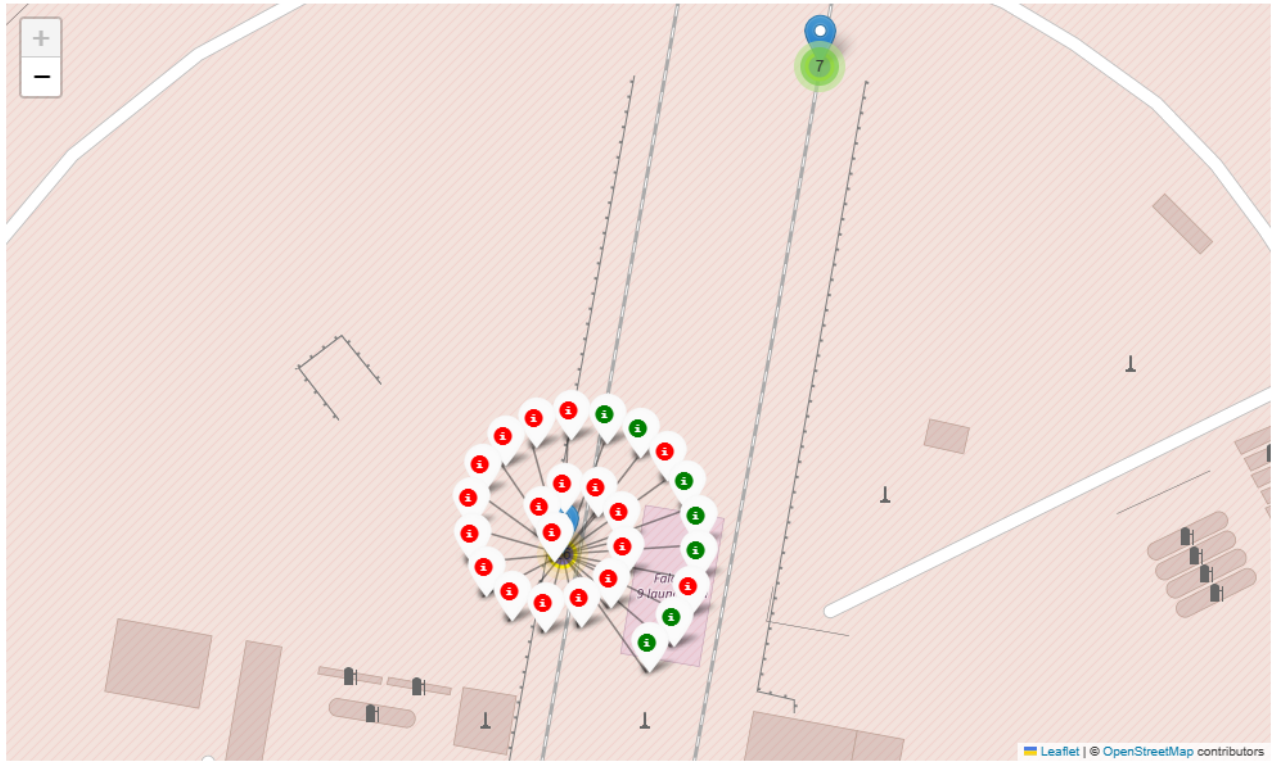
Launch Sites Proximities Analysis

Launch Site Locations



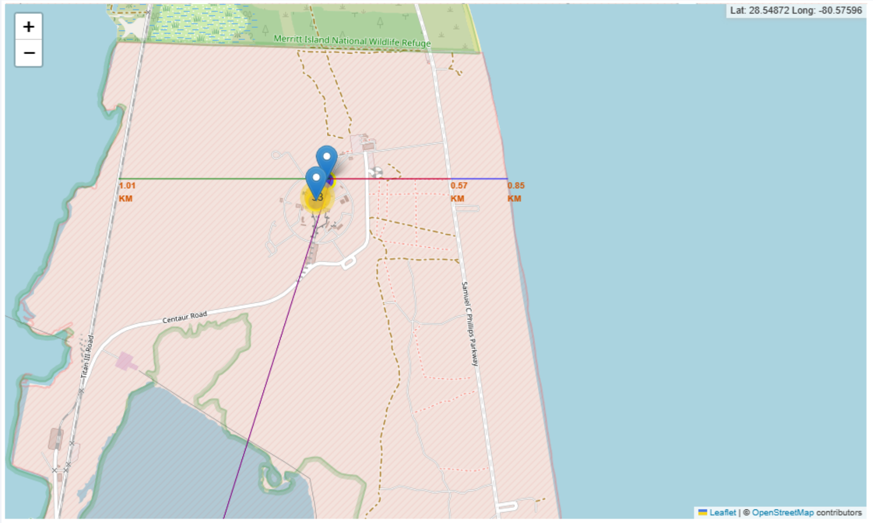
- Each launch site location has been marked with a blue circle, as well as a red label with the name of the site. There is a total of 3 launch sites on the east coast in Florida, and one launch site on the west coast in California.

Site Detail with Color-labeled Launch Outcomes

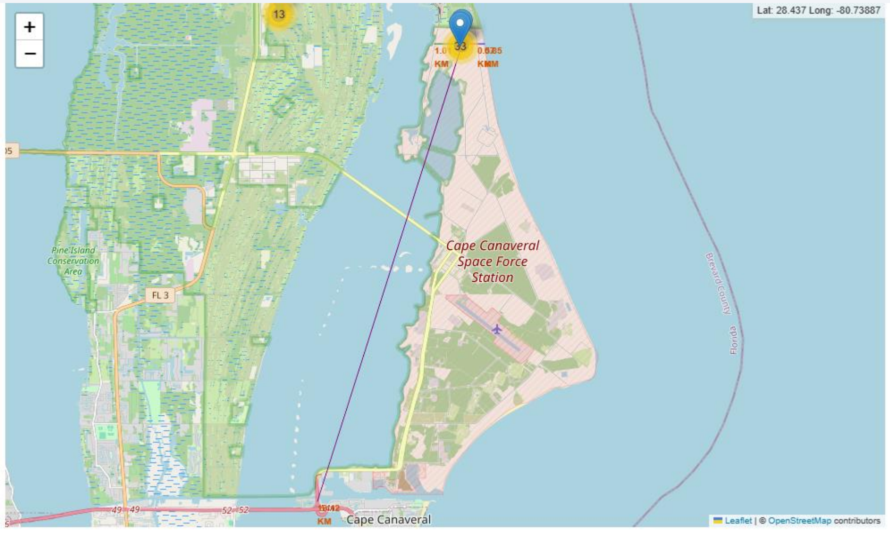


- This is a close up view of one of the launch sites, showing the color-labeled launch outcomes on the map. This demonstrates a site where there were more unsuccessful launches than successful ones.

Launch Site Proximity

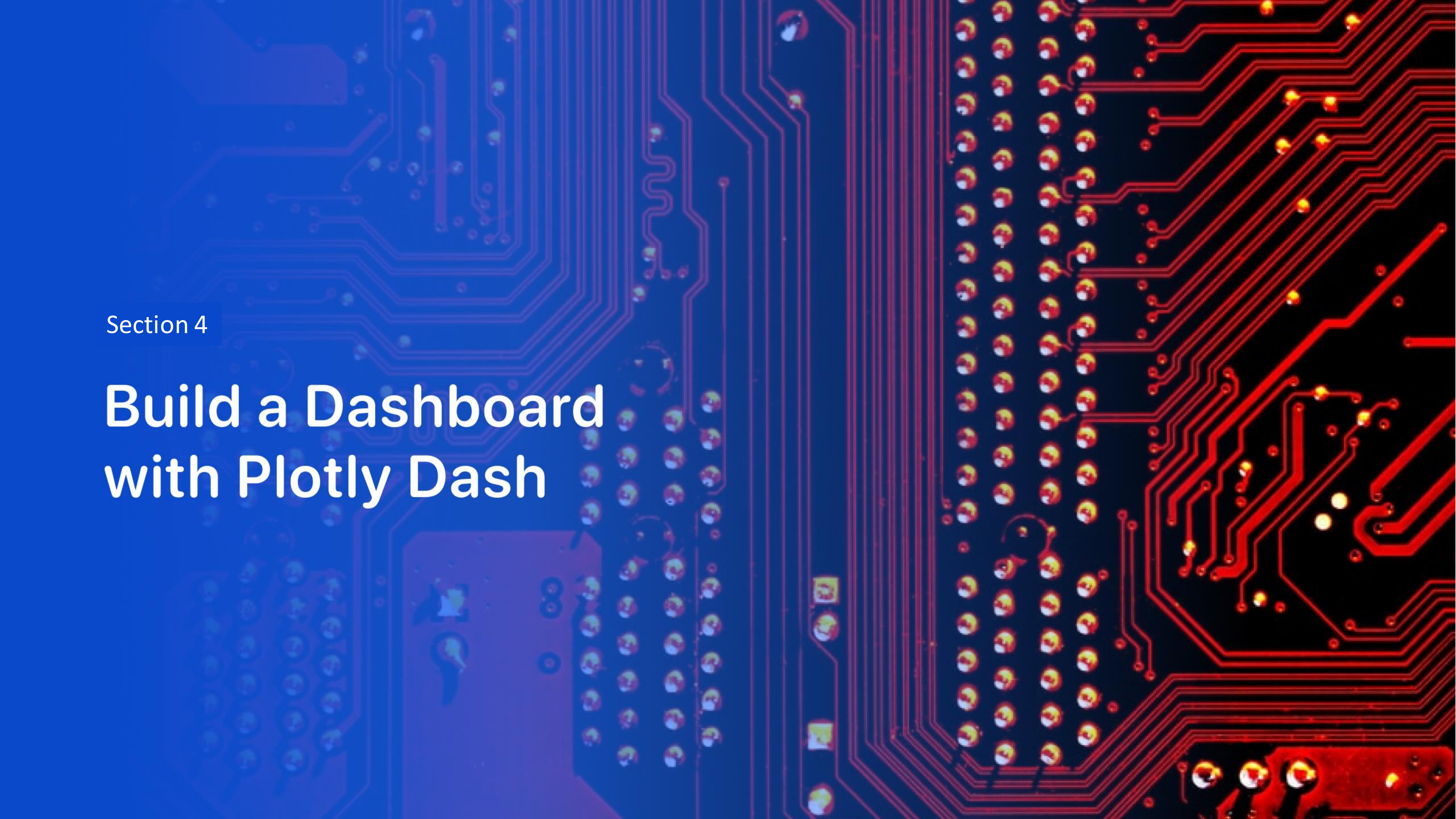


+
−
Lat: 28.54872 Long: -80.57596
Merritt Island National Wildlife Refuge
1.01 KM
0.57 KM
0.85 KM
33
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors



+
−
Lat: 28.437 Long: -80.73887
13
33
KM
KM
Pine Island Conservation Area
FL 3
Cape Canaveral Space Force Station
Brevard County
Florida
49
52
KM
Cape Canaveral
Leaflet | © OpenStreetMap contributors

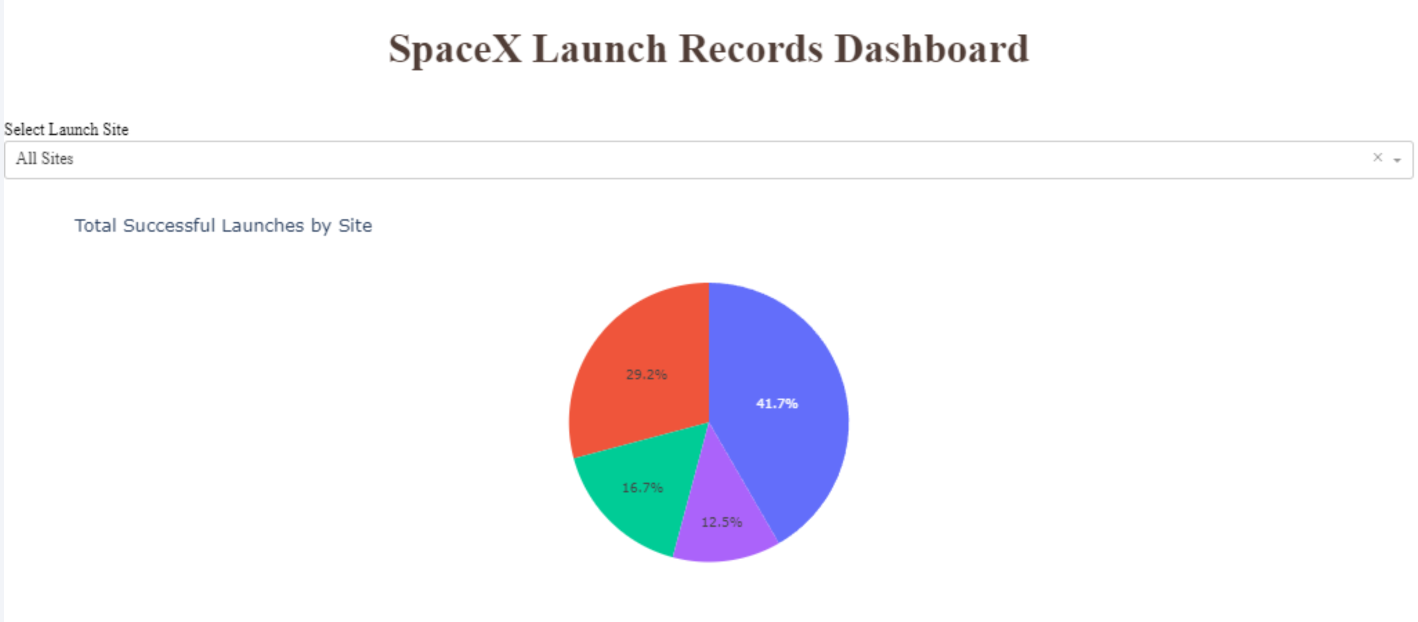
- This map shows one of the launch sites and its proximity to the nearest railway, highway, coastline, and city. The calculated distance is shown in the labels. While the highway, railway, and coastline are relatively near the site, the nearest city is over 10 miles away.



Section 4

Build a Dashboard with Plotly Dash

Launch Success Counts for All Launch Sites



SpaceX Launch Records Dashboard

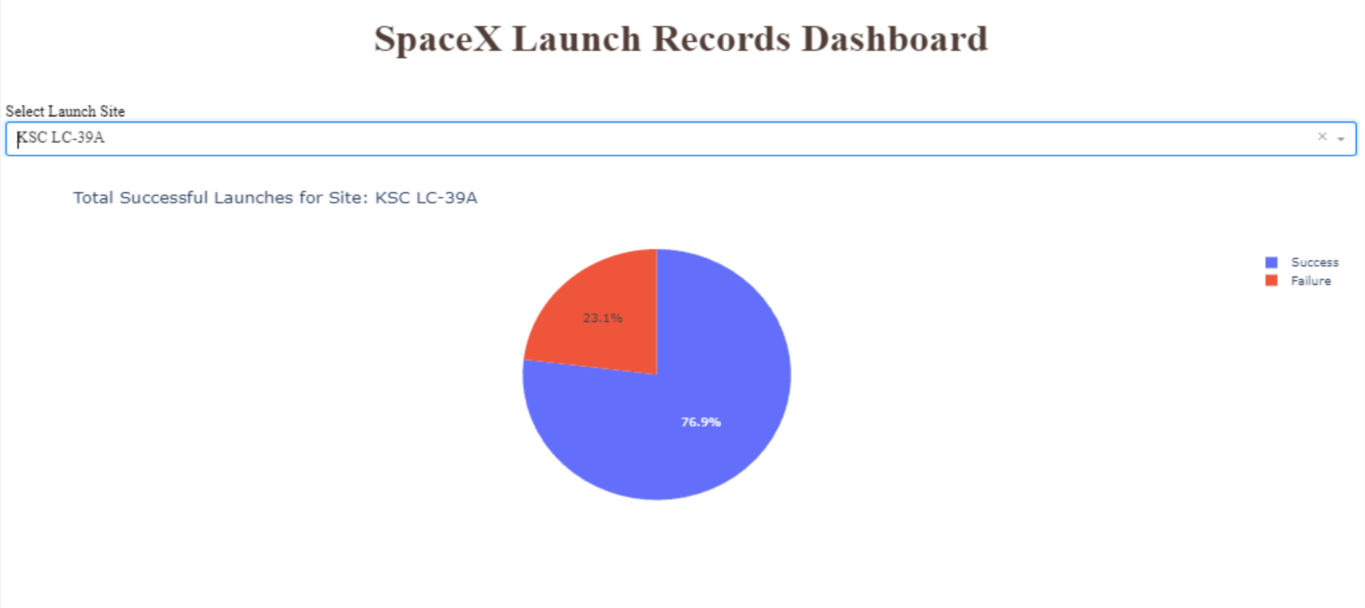
Select Launch Site

All Sites

Total Successful Launches by Site

- This pie chart shows the launch success count for all sites, and their percentage.
- The drop-down selector allows us to further filter this result to individual sites, and will display a pie chart for the successes vs. failures at the individual site level.

Launch site KSC LC-39A (Highest Launch Success Ratio)



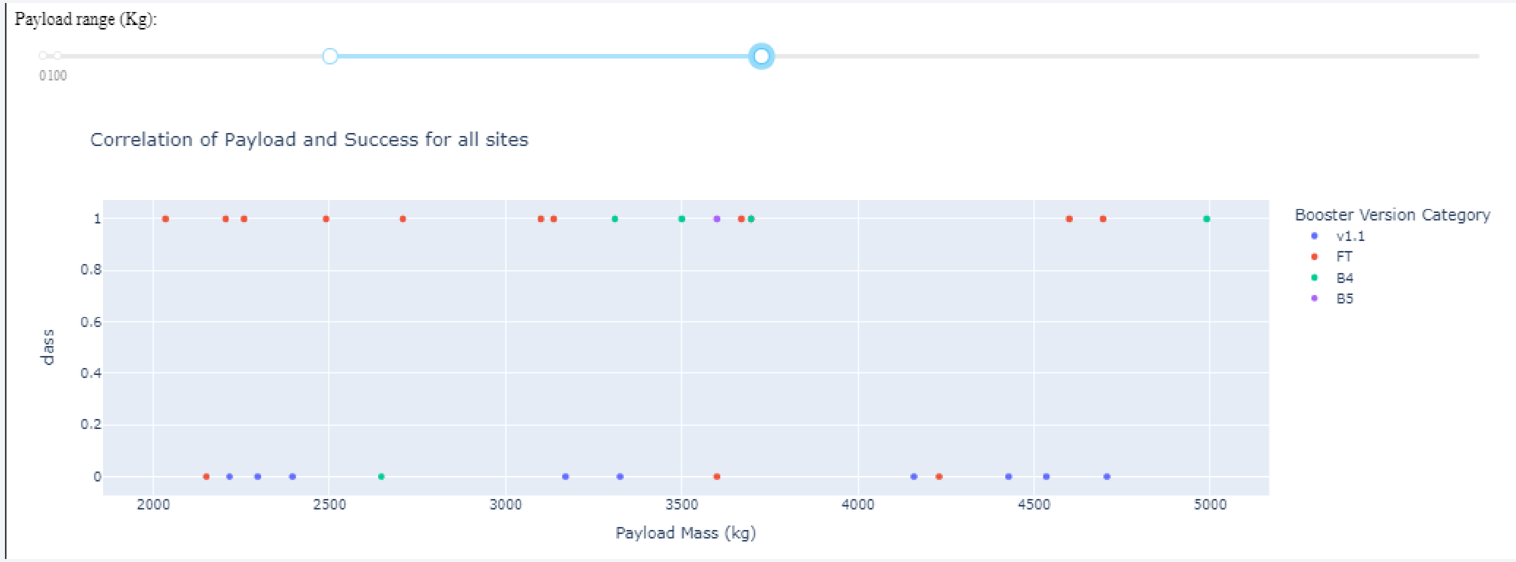
SpaceX Launch Records Dashboard

Select Launch Site

KSC LC-39A

- This chart shows the launch site with the highest launch success ratio, KSC LC-39A.
- The success rate for this site is 76.9%, and is the highest success rate among the sites.

<Dashboard Screenshot 3>



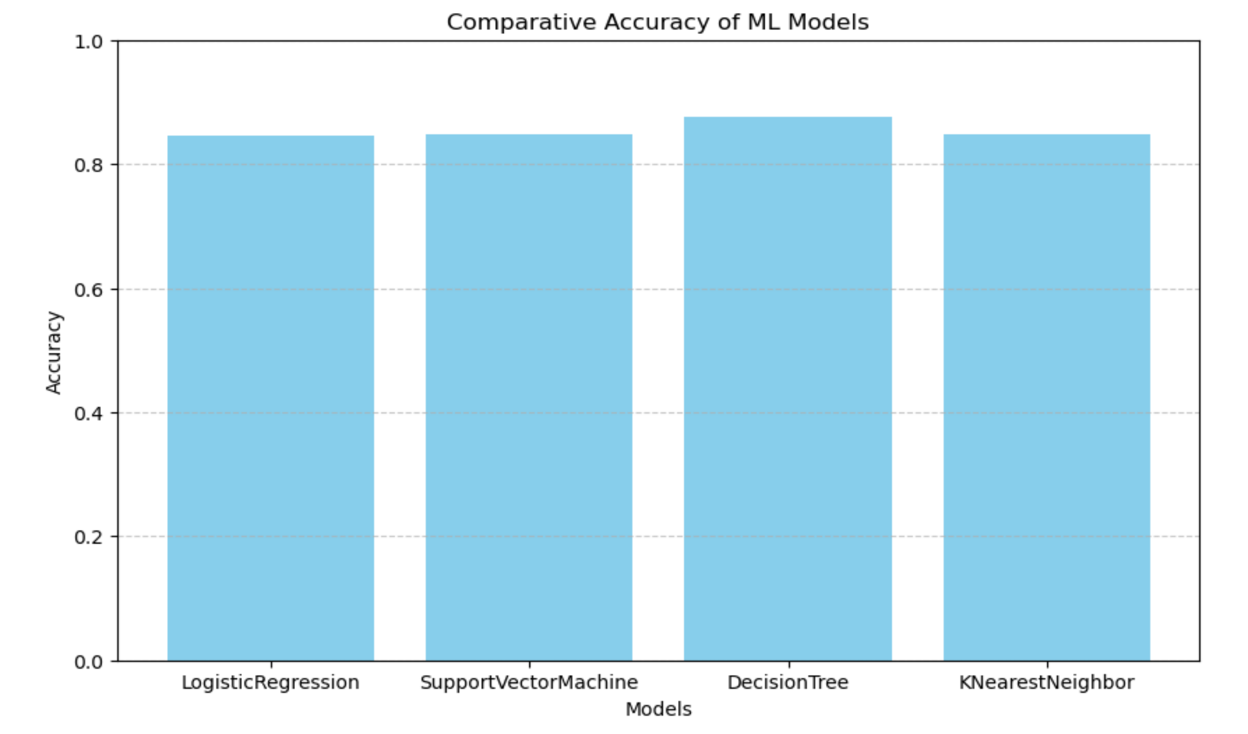
- This is the scatter plot for Payload vs. Launch Outcome for all sites, and the selected payload in the range slider is 2,000 kgs to 5,000 kgs. The booster version category is color-coded.

- It should be noted that there are quite a few failed launches with the v1.1 booster version, while the FT, B4, and B5 booster categories have a much greater majority of success.

Section 5

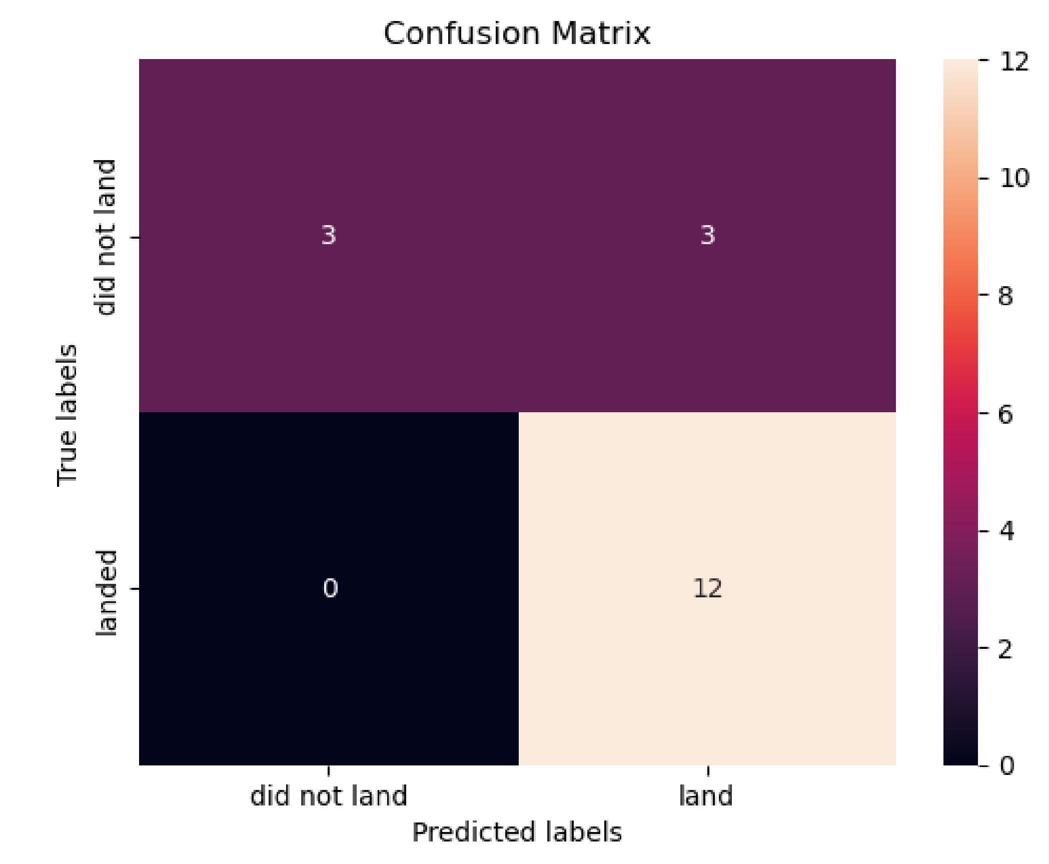
Predictive Analysis (Classification)

Classification Accuracy



- Displayed here is a bar chart for the model accuracy for all built classification models. We can see that they all produced similar accuracies, with the Decision Tree having a slight edge over the other models.

Confusion Matrix



Due to the small sample size of the test data, the models all produced similar confusion matrices. One such matrix is shown here, with a small amount of false positives, and no false negatives.

Conclusions

- Success was dependent on multiple factors, and included:
- Higher flight numbers tended to have better success rates.
- Certain orbit types had better success rates. In our data, the orbit types ES-L1, SSO, HEO, and GEO had only successes, while the orbit types LEO, ISS, and PO had higher success rates as the payload mass increased.
- Success rate also increased year over year.

Appendix

- All completed notebooks for this project can be found at: https://github.com/theLamaMoos/SpaceX_FinalProject/tree/main

Thank you!