IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ritik Kulkarni
09/12/2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results:
  - EDA result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context:
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, where other providers cost upward of 165 million dollars each. SpaceX saves so much money 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.
- The goal of this project is to find out if the Falcon 9 first stage will land successfully. To that end, as a data scientist, we apply many of the processes in data science methodology, ranging from Data collection to Model Evaluation, to provide the solution to this problem.

Section 1

Methodology

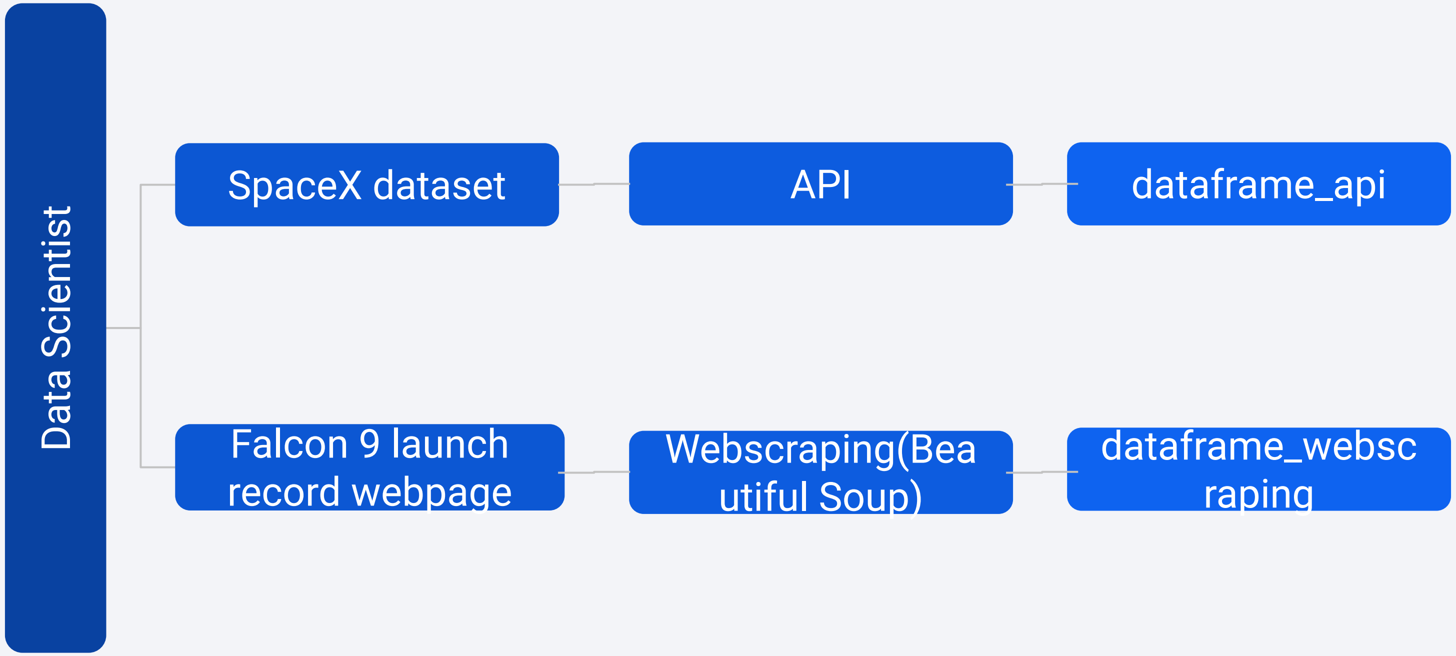
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were created using the help of scikit-learn library and their best parameter values were discovered with the help of GridSearchCV.

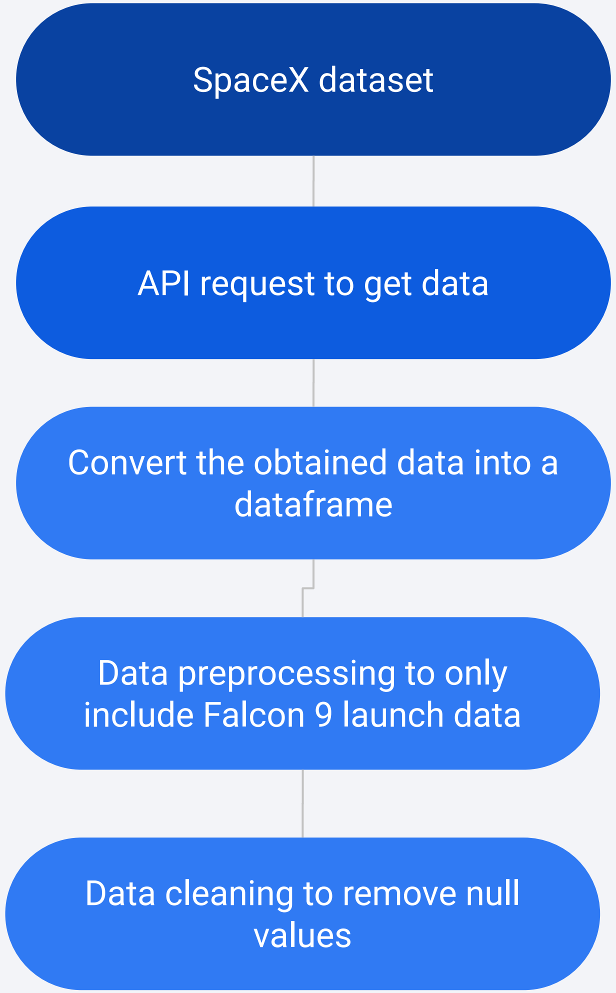
Data Collection

- Data was collected using two methods, one was loading the SpaceX dataset into a dataframe with the help of an API, and the other was scraping the web page using beautiful soup for the table with launch history and converting it into a dataframe for further processing and use.



Data Collection – SpaceX API

- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



SpaceX dataset

API request to get data

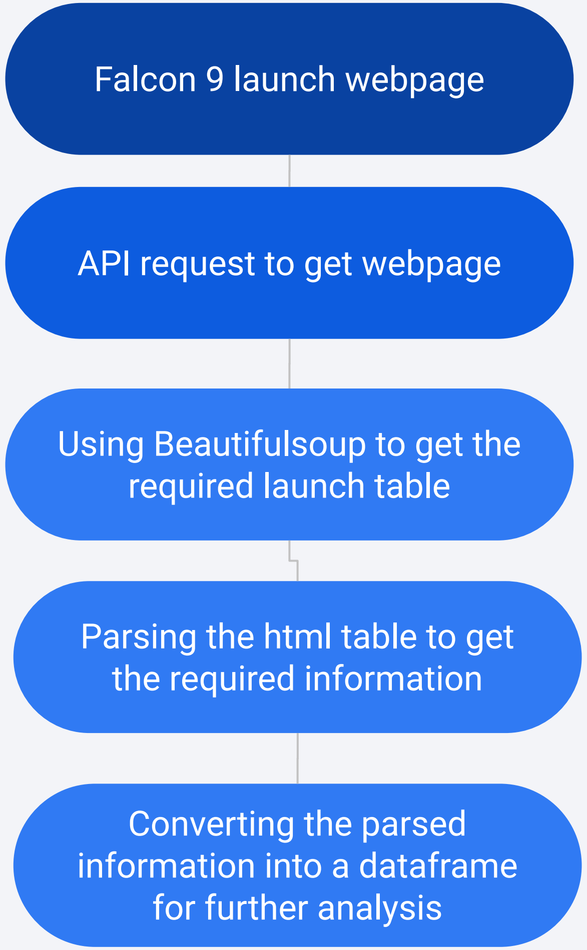
Convert the obtained data into a dataframe

Data preprocessing to only include Falcon 9 launch data

Data cleaning to remove null values

Data Collection - Scraping

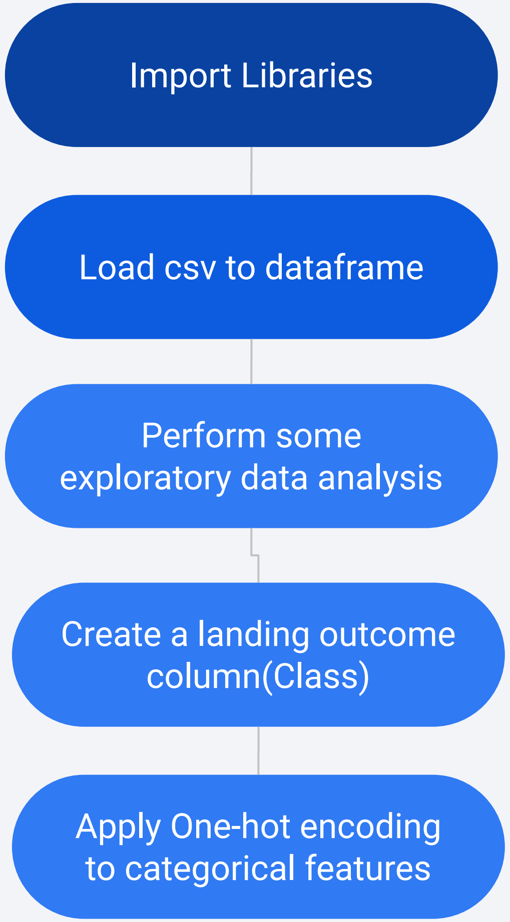
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Falcon 9 launch webpage
API request to get webpage
Using Beautifulsoup to get the required launch table
Parsing the html table to get the required information
Converting the parsed information into a dataframe for further analysis

Data Wrangling

- Initially, we perform some preliminary analysis on the data to get an idea of some factors which we believe would be important and may impact the final prediction.
- We get an idea of how launches varied by launch site, orbit type, as well as how many were successful vs unsuccessful and what types of landing they were.
- Finally, we create an landing outcome column(Class) to simplify the landing outcomes into successful(1) and unsuccessful(0).
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



Import Libraries

Load csv to dataframe

Perform some exploratory data analysis

Create a landing outcome column(Class)

Apply One-hot encoding to categorical features

EDA with Data Visualization

- Scatter point charts were used to visualize the relationships between various factors such as Flight Number, Payload Mass, Launch Site, Orbit Type.
- Bar charts to visualize the success rate for each Orbit Type.
- Line charts to visualize the yearly launch success trend.
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/eda_pandas_matplotlib.ipynb

EDA with SQL

- Queries to find out the names of launch site in the space mission.
- Queries to find the payload mass using different aggregate functions carried by the various boosters.
- Queries to find which boosters had successful landing outcomes, as well as finding out the count and date of the successful and unsuccessful outcomes.
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

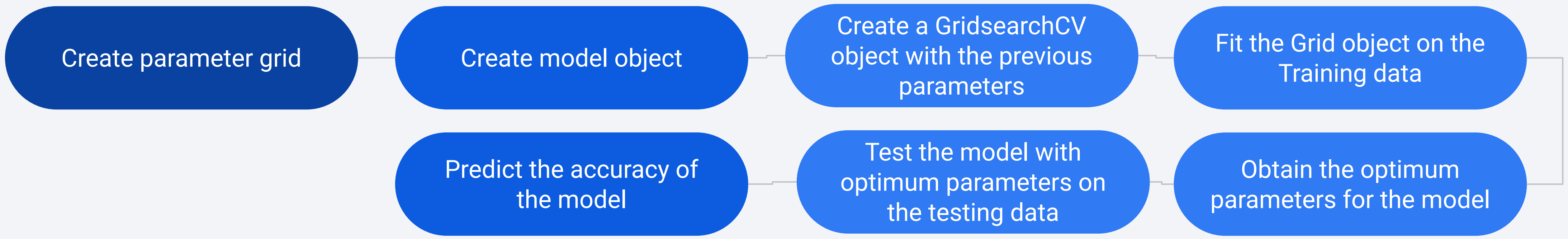
- We used map objects such as markers and circles initially to mark all the launch sites on a map.
- Afterwards, we further added markers to a map object known as MarkerCluster to mark the success/failed launches for each site on the map.
- Finally, we calculated the distance between the launch sites and its proximites(such as railways, cities, coastlines, etc.) and added the map object lines to represent this distance on the map. This helped us understand whether the launch sites were kept close to such proximites or away from them and why.
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Created an interactive dashboard with a dropdown for selecting the launch site, to either get an overview of all sites or target a specific launch site.
- Also, included a payload slider to test out the difference in success rates for various boosters for different payload ranges.
- Added a pie chart to get a ratio of the successful launches for each launch site.
- Finally, added a scatter plot of payload mass vs success rate to get an idea of the performance of various boosters for different payload values.
- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

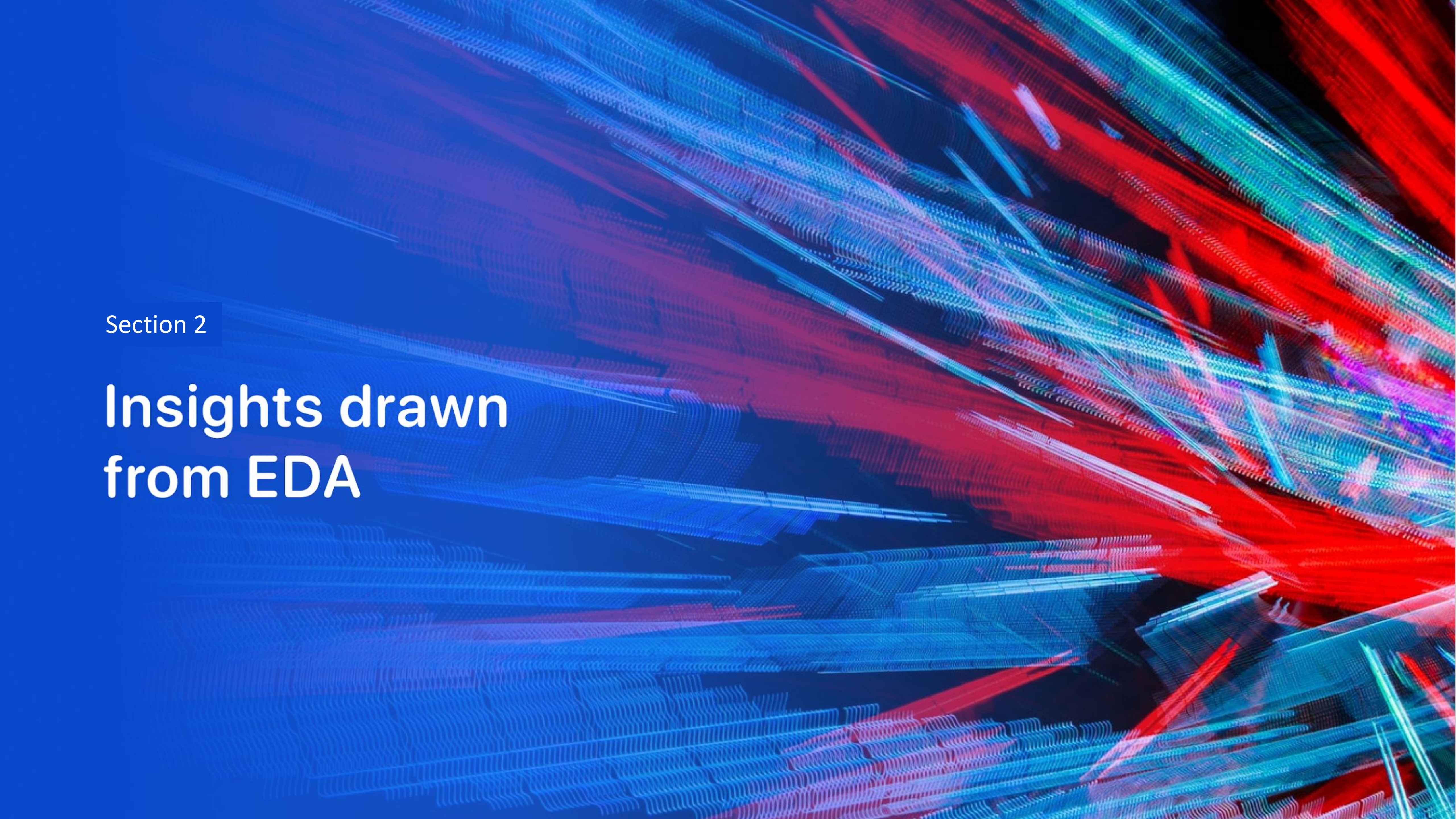
- The classification models that we used for this project were Logistic Regression, SVM, Decision Tree and K nearest neighbours.
- Initially, we normalize the input data and then proceed to create a train test split, where we use the training data to train the model and testing data to obtain the trained model accuracy.
- For each classification model, the model development would look like this:


Create parameter grid
Create model object
Create a GridsearchCV object with the previous parameters
Fit the Grid object on the Training data
Obtain the optimum parameters for the model
Test the model with optimum parameters on the testing data
Predict the accuracy of the model

- GitHub URL: https://github.com/RK-LAYZEE/Data_Science_Professional_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

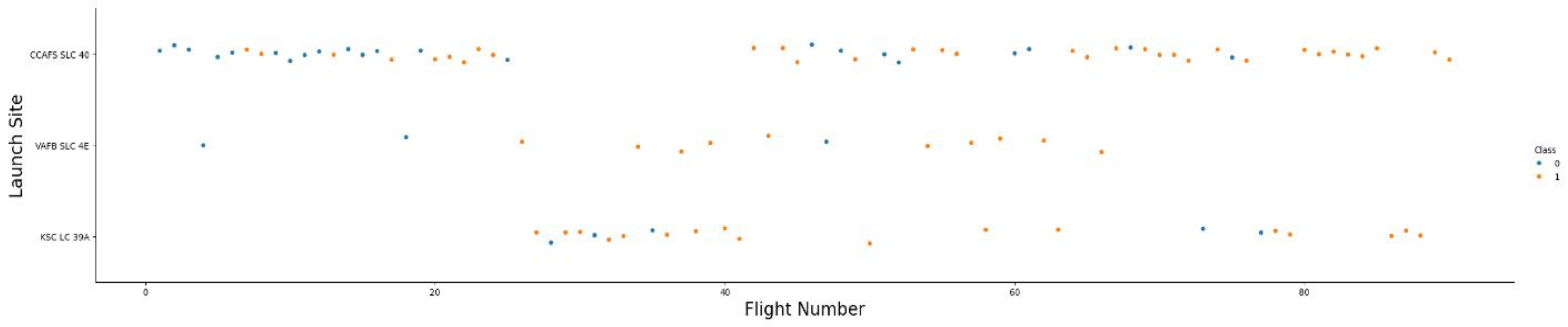
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

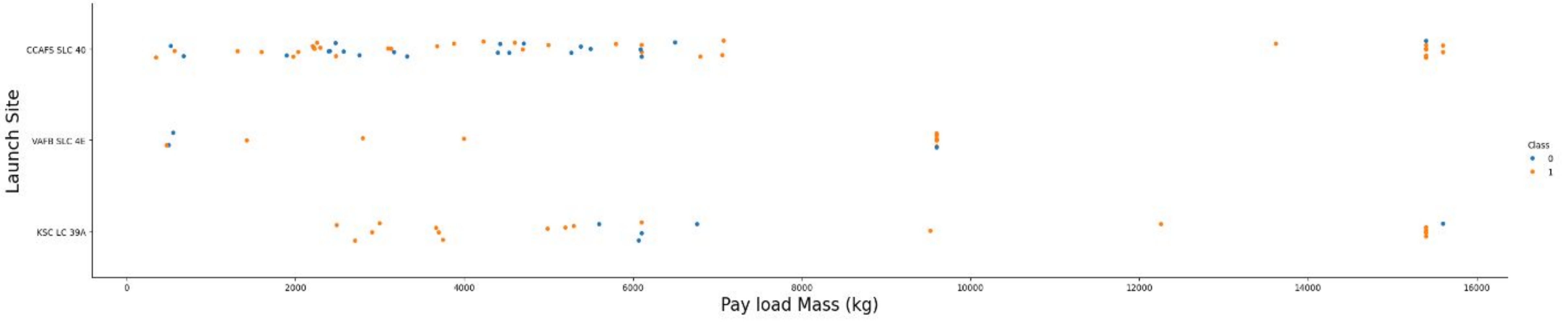
Insights drawn from EDA

Flight Number vs. Launch Site



- The scatter plot shows that as Flight Number increases, the number of successful outcomes also increases at each Launch Site.
- We can see that though successful, the number of flight attempts at VAFB SLC-4E are very few. This could be due to the launch site not having that close of a proximity to the other two launch sites.

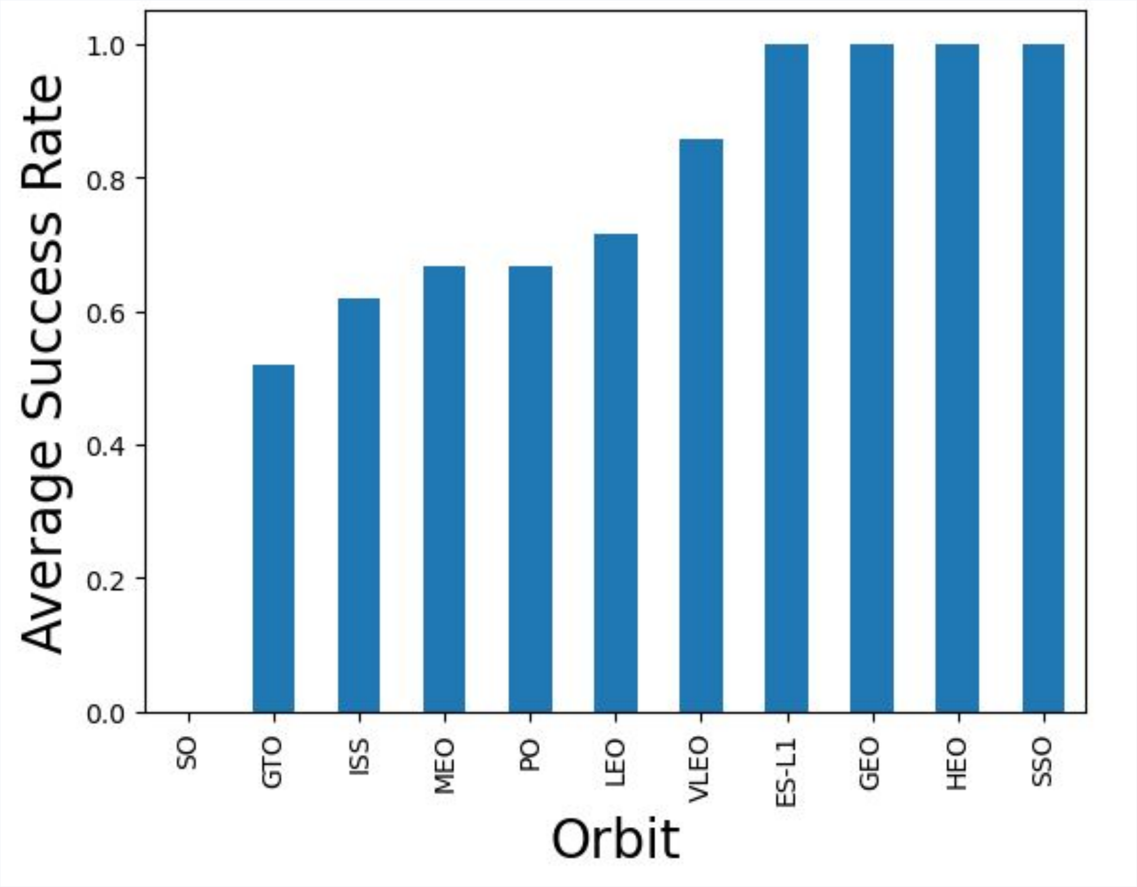
Payload vs. Launch Site



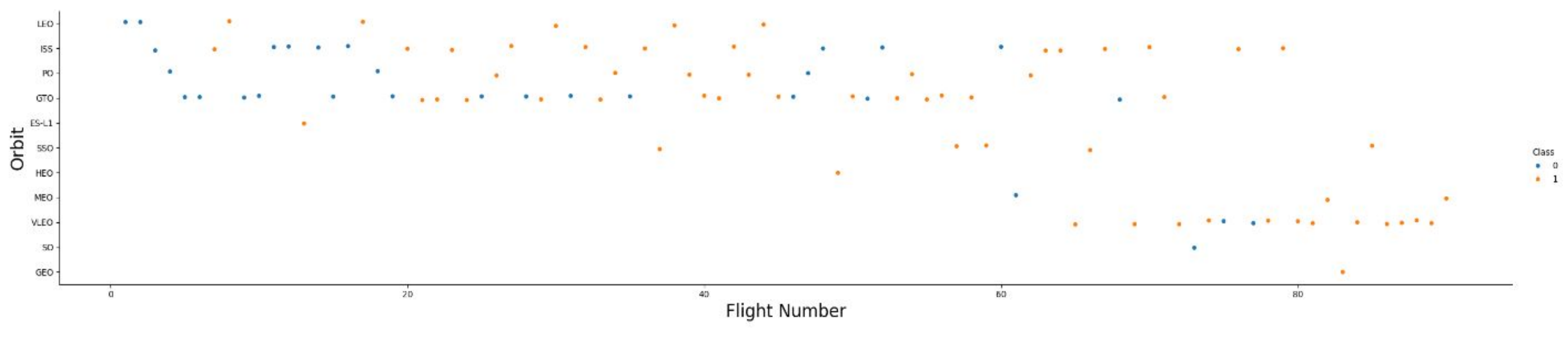
- We can infer from the scatter plot that for heavier payloads(greater than 8000) there are more successful outcomes than failures, possibly based on learning achieved from the lighter payloads.
- Additionally, we can observe that for VAFB-SLC 4E launch site there are no rockets launched for heavy payloads(greater than 10000).

Success Rate vs. Orbit Type

- For orbits ES-L1, GEO, HEO, SSO, the success rate is almost 1, indicating that they have the highest success rates.
- For orbit SO, unfortunately the success rate is very low, possibly indicating that future launches with SO orbit in mind might have a very low chance of a successful outcome.



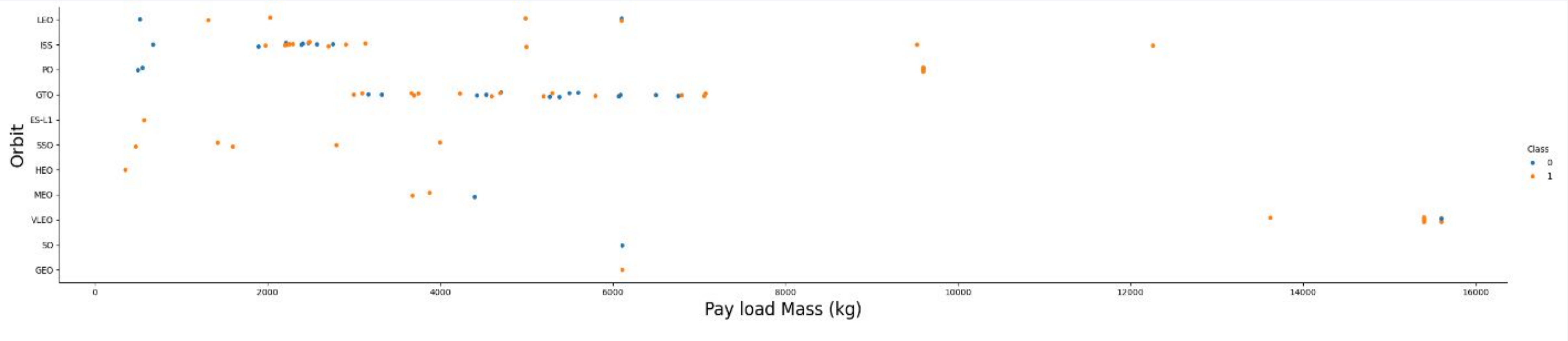
Flight Number vs. Orbit Type



- We can infer from the plot that in the LEO orbit, success seems to be related to the number of flights.

- Conversely, in the GTO orbit, there appears to be no relationship between flight numbers and success.

Payload vs. Orbit Type

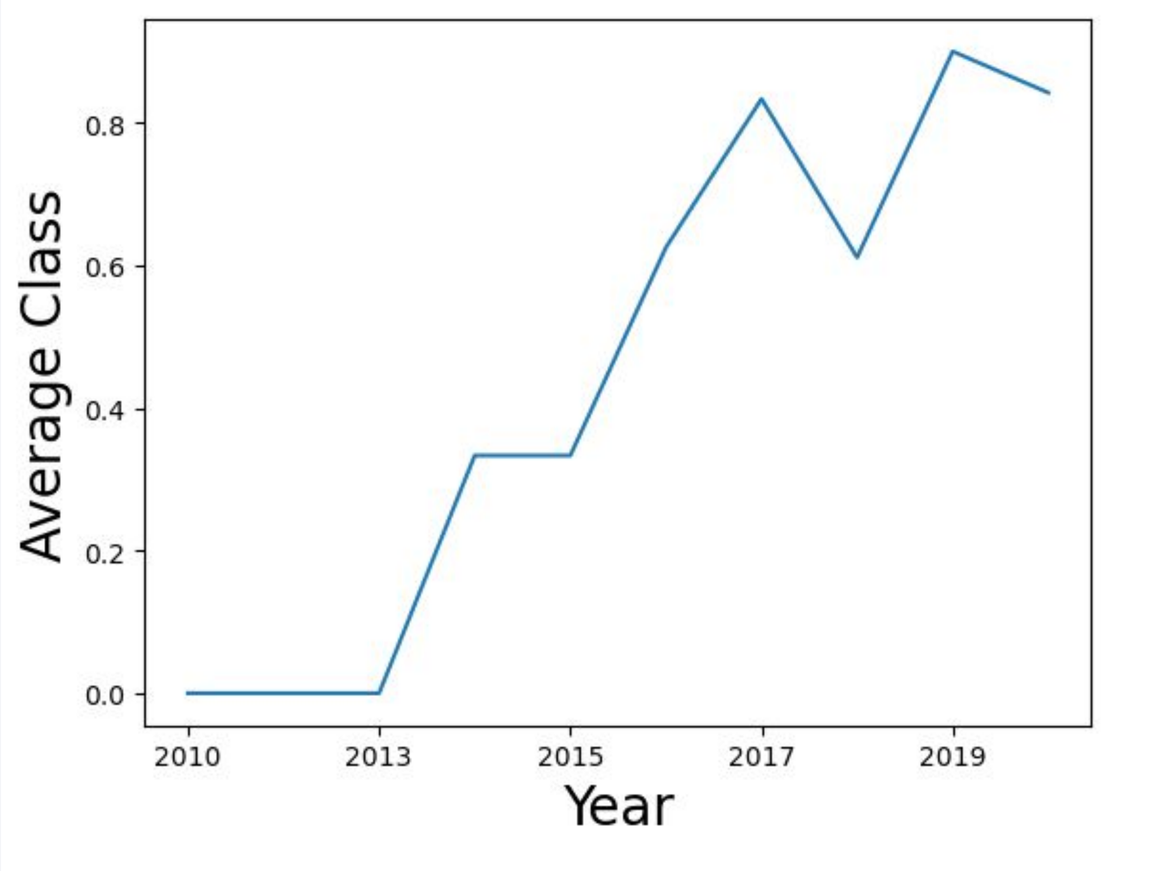


- From the graph we can infer that, with heavy payloads, the successful landing rate are more for Polar, LEO and ISS.

- However, for GTO, it’s difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- The line chart shows that the success rate or rate of successful landings saw an increase or an upward trend starting in the year 2013.

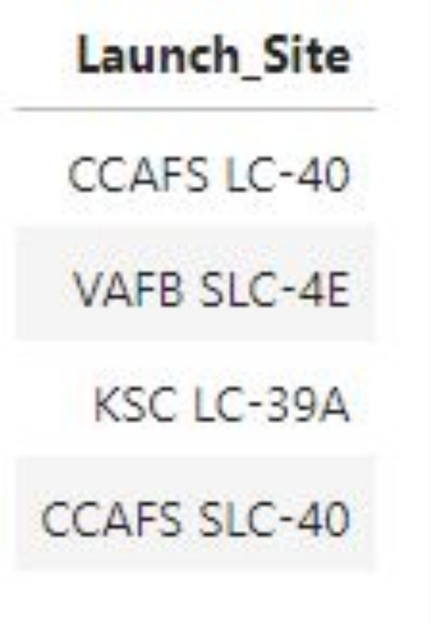


All Launch Site Names

- Find the names of the unique launch sites

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
```

- Result:
    - We used the Distinct keyword to find the unique launch sites.
    - We are able to see that SpaceX conducted all launches from four launch sites.



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

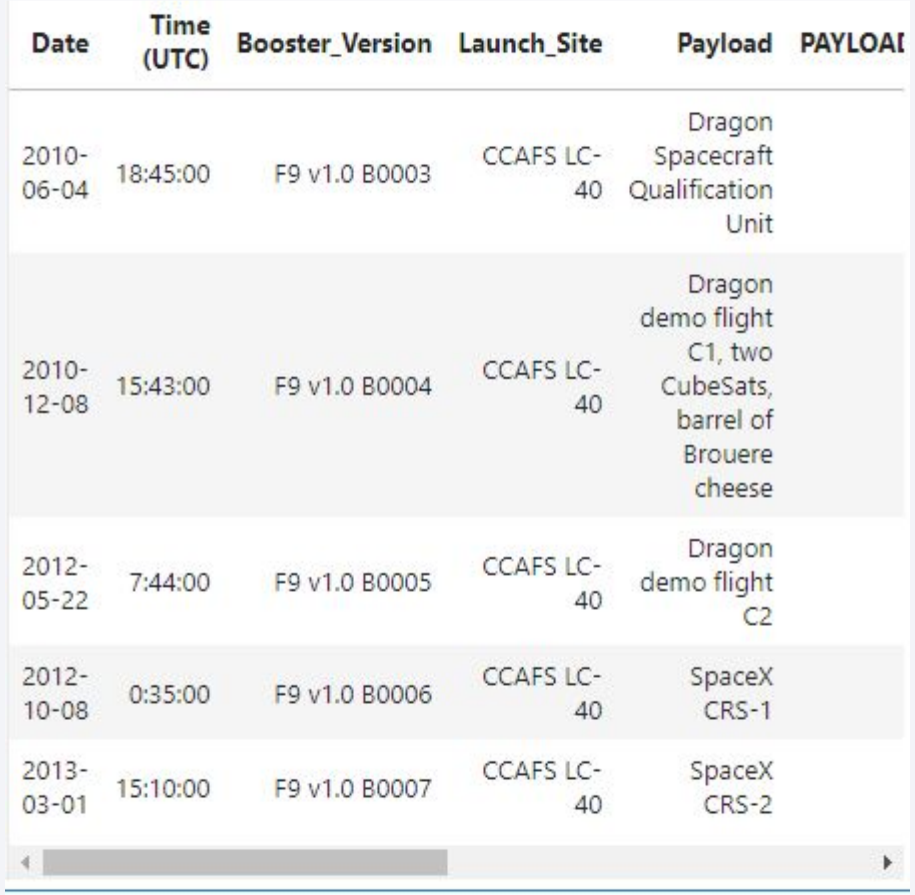
Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with \`CCA\`

  %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like "CCA%" LIMIT 5;

- Result:

  - Using the ‘like’ operator, we were able to find all records that started with CCA, and the we took the first 5 records.



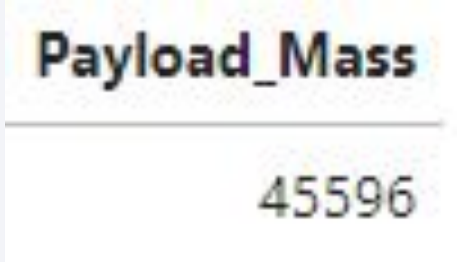
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as Payload_Mass FROM SPACEXTABLE WHERE Customer =
"NASA (CRS)";
```

- Result:
  - The result was achieved by performing the Sum operation on the Payload mass retrieved from the rows where the Customer was ‘NASA (CRS)’



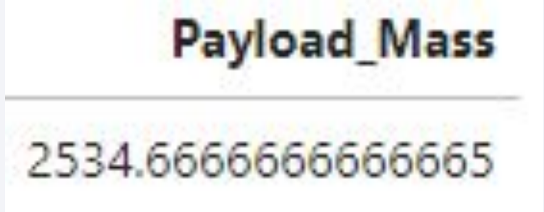
Payload_Mass
45596

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version like "F9 v1.1%";
```

- Result:
    - We used the AVG operation to get the average or mean of all payload masses retrieved from rows where booster version started with ‘F9 v1.1’.



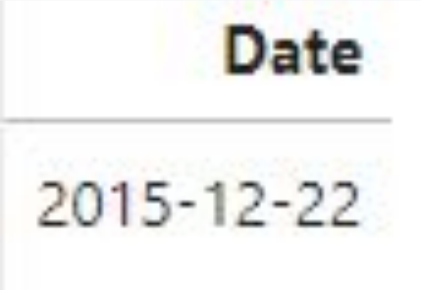
Payload_Mass
2534.666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT min(Date) as Date FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
```

- Result:
  - First, we retrieved the dates from the rows where there was a successful landing outcome on the ground pad.
  - Secondly, we used the Min operation to get the smallest or the earliest date.



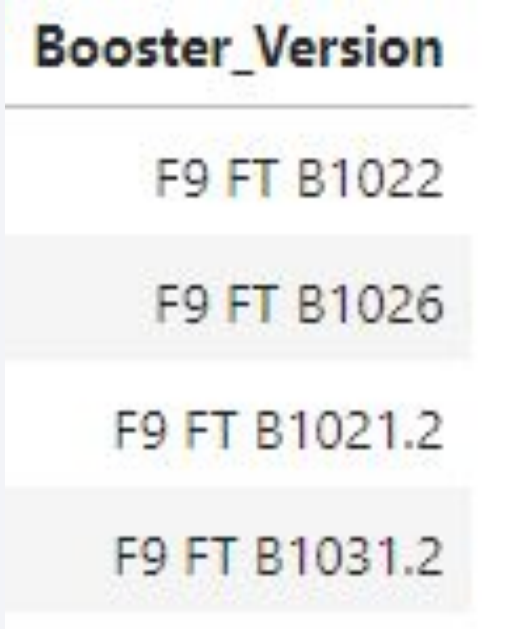
Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND
PAYLOAD_MASS__KG_ < 6000;
```

- Result:
    - The booster versions of those rows were retrieved where it had a successful drone ship landing outcome and the payload was between 4000 and 6000.



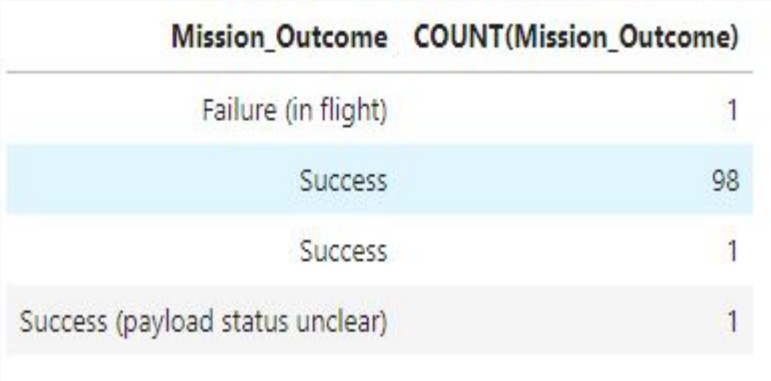
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome)
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

- Result:
  - We group by the rows of the table based on the mission outcome, and we see that the total failures is 1 and total success is 100.



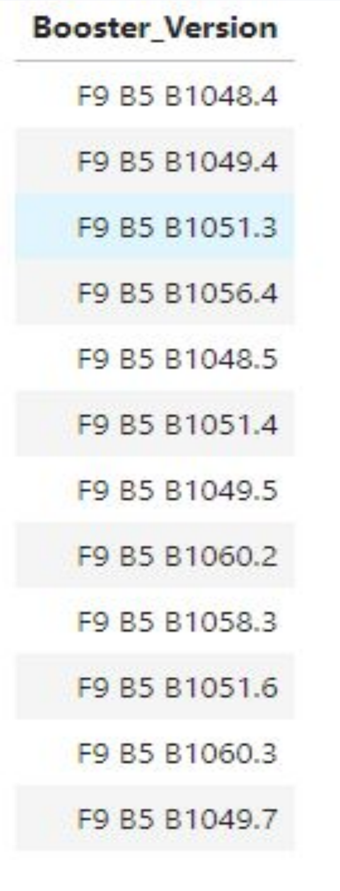
Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE);
```

- Result:
  - Here, we retrieve the booster versions where their payload mass is the maximum.



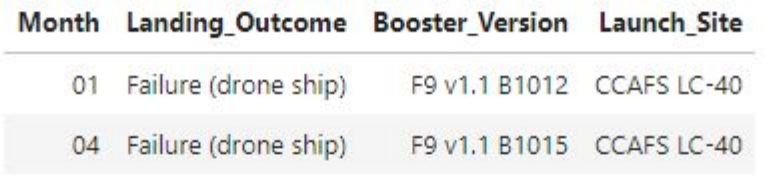
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%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)";
```

- Result:
  - Here, we retrieve the rows where the landing outcome is a failed drone ship landing for the year 2015.



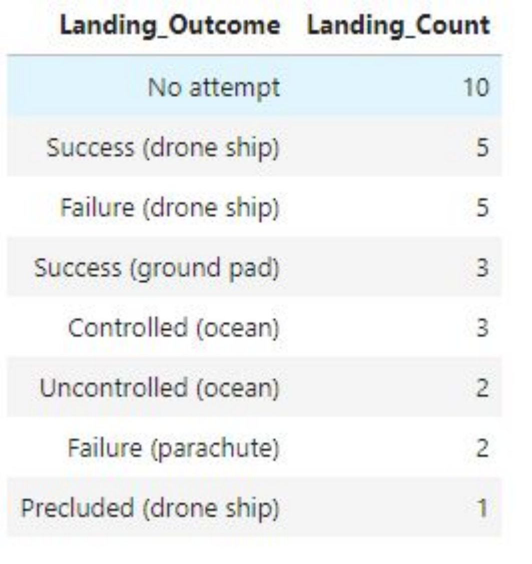
Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Landing_Count
FROM SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
GROUP BY Landing_Outcome
ORDER BY Landing_Count DESC;
```

- Result:
  - Here, we have ordered the counts of landing outcomes in descending order.



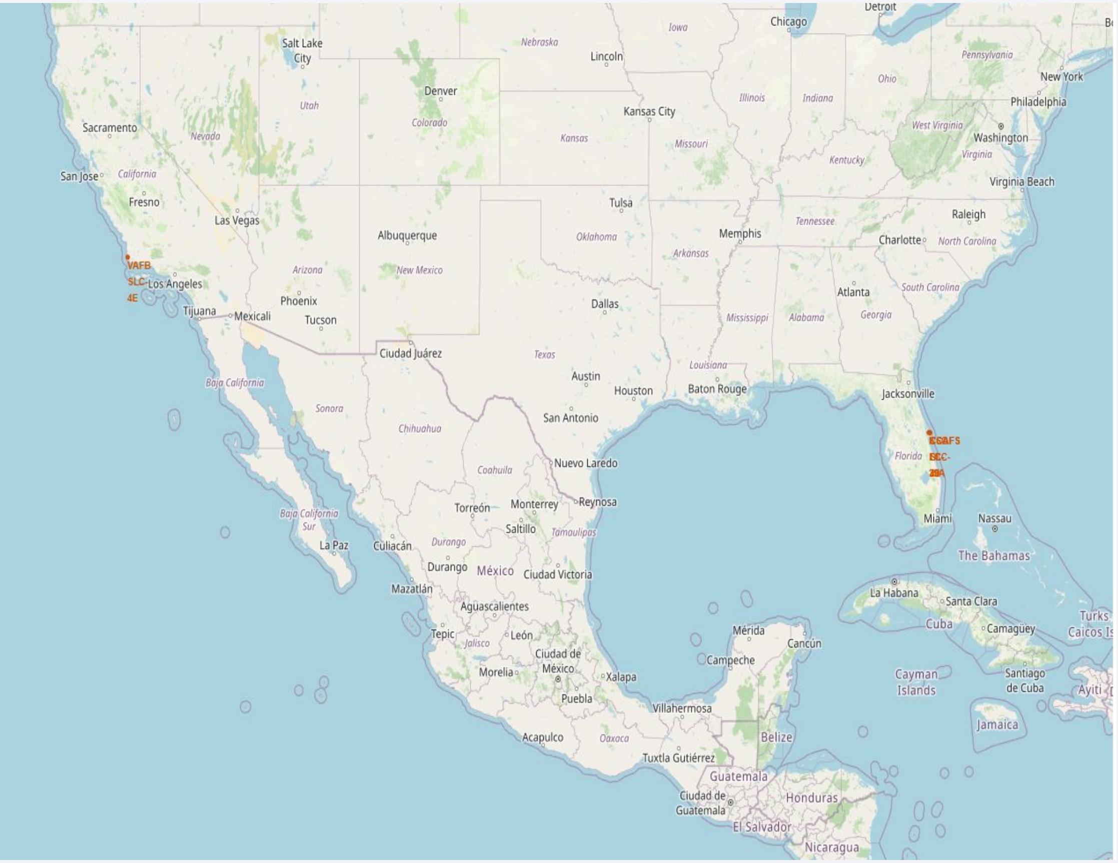
Landing_Outcome	Landing_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

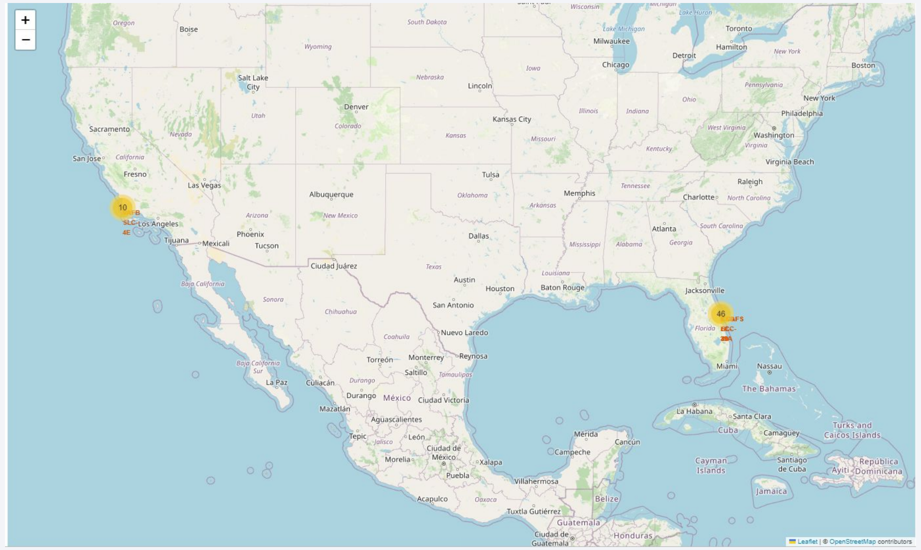
Folium map with all launch sites

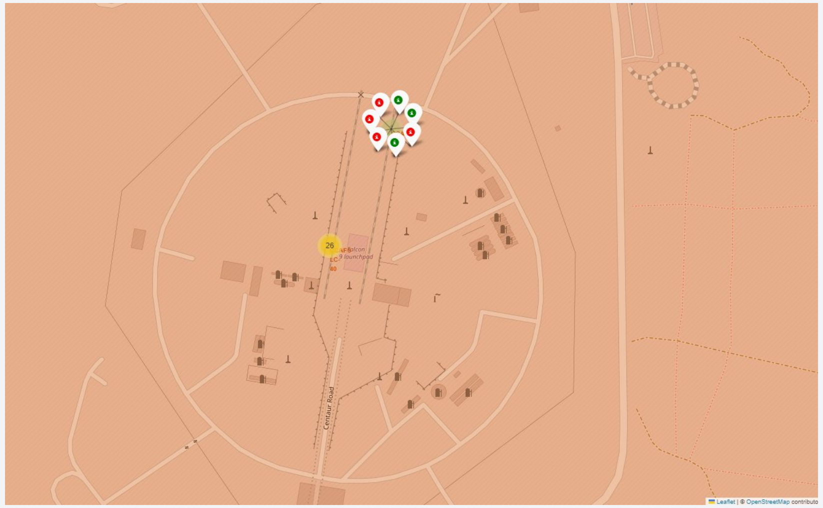
- We can see all 4 launch sites marked out on the Folium map.
- Additionally, we can see that three of the launch sites, other than VAFB-SLC 4E are in certain proximity to one another as evidenced by their closeness shown on the map.
- We can also see that all the launch locations are located near the coast.



Folium map with launch outcomes

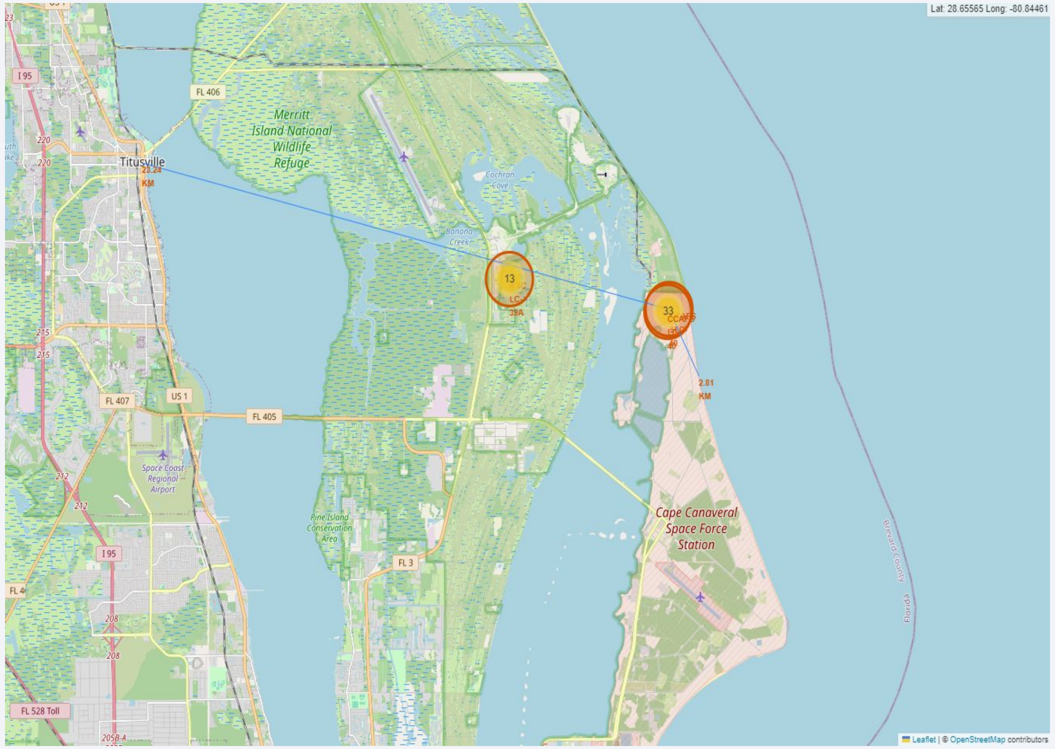
- The value in each yellow marker circle, tells anyone who sees it how many launches occurred on that site. This tells them which launch site is considered more impactful.
- Using the launch outcome markers, anyone can easily infer which launch sites have a higher launch success rate





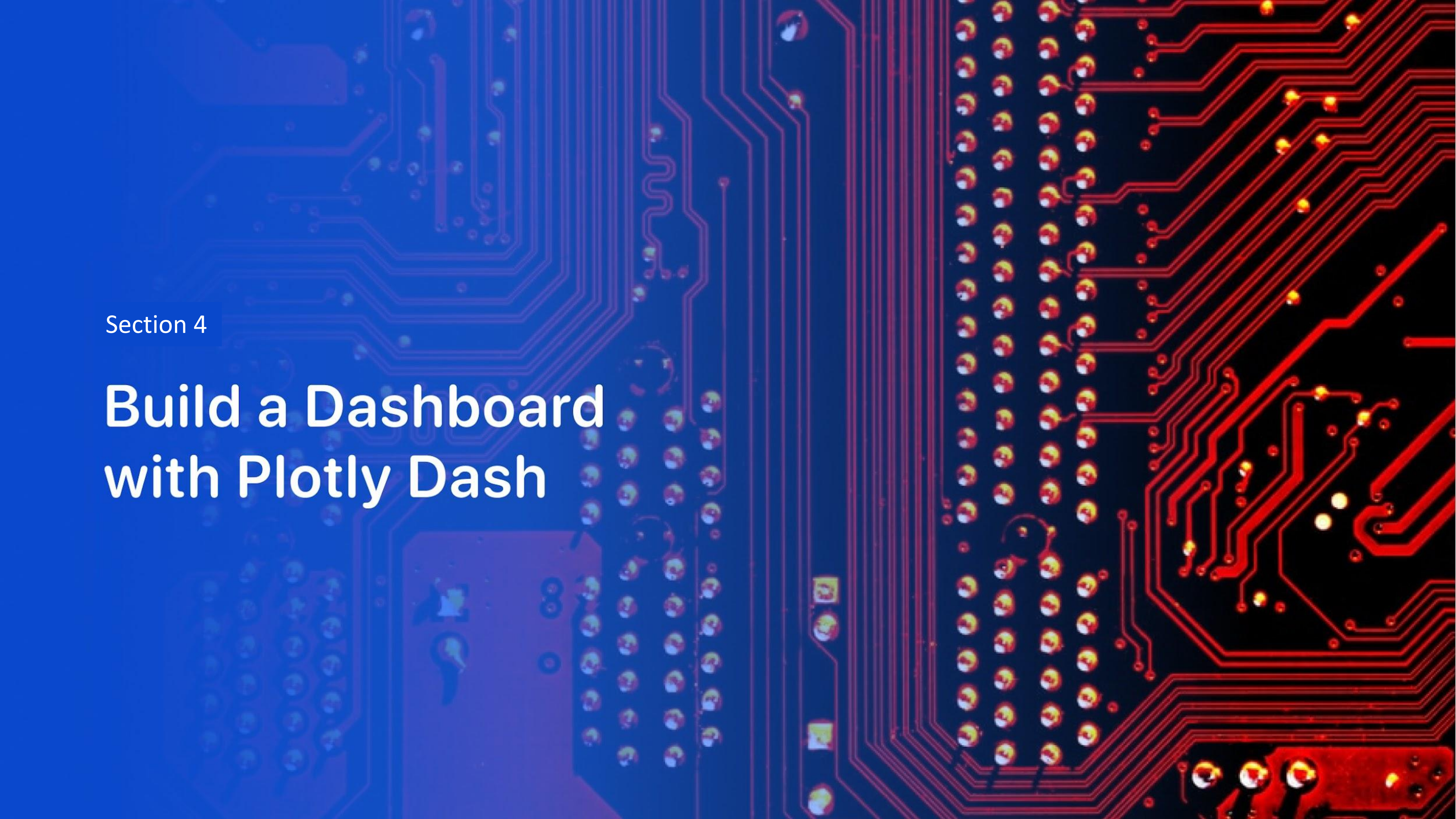
Folium map with proximity markings

- We can see two proximity lines in the given image.
- The first line is to a point near the coastline, the close proximity to the coast is because SpaceX launches their rockets over the sea to avoid danger to the general populace.
- The second line is to the city of Titusville. The large distance between the launch site and the city, indicate that they want to keep the general populace away from danger in case of an issue, as well as away from general concerns that may rise by having a rocket launch nearby.

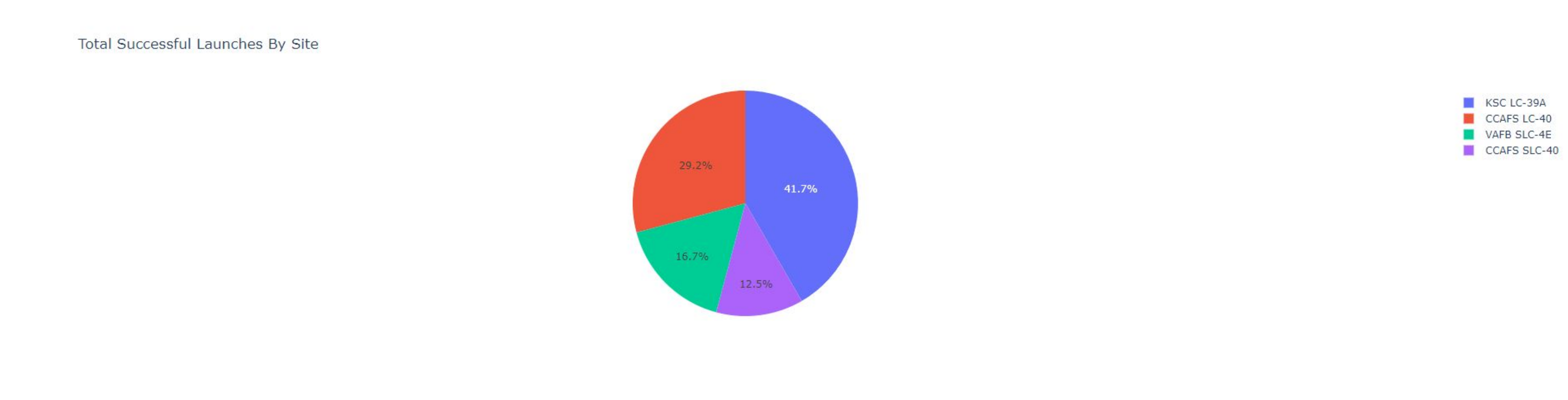

Lat: 28.65565 Long: -80.84461
FL 406
Merritt Island National Wildlife Refuge
Titusville
23.24 KM
13
33
2.81 KM
FL 407
US 1
FL 405
Space Coast Regional Airport
Pine Island Conservation Area
FL 3
I 95
Cape Canaveral Space Force Station
FL 528 Toll
Leaflet | © OpenStreetMap contributors

Section 4

Build a Dashboard with Plotly Dash

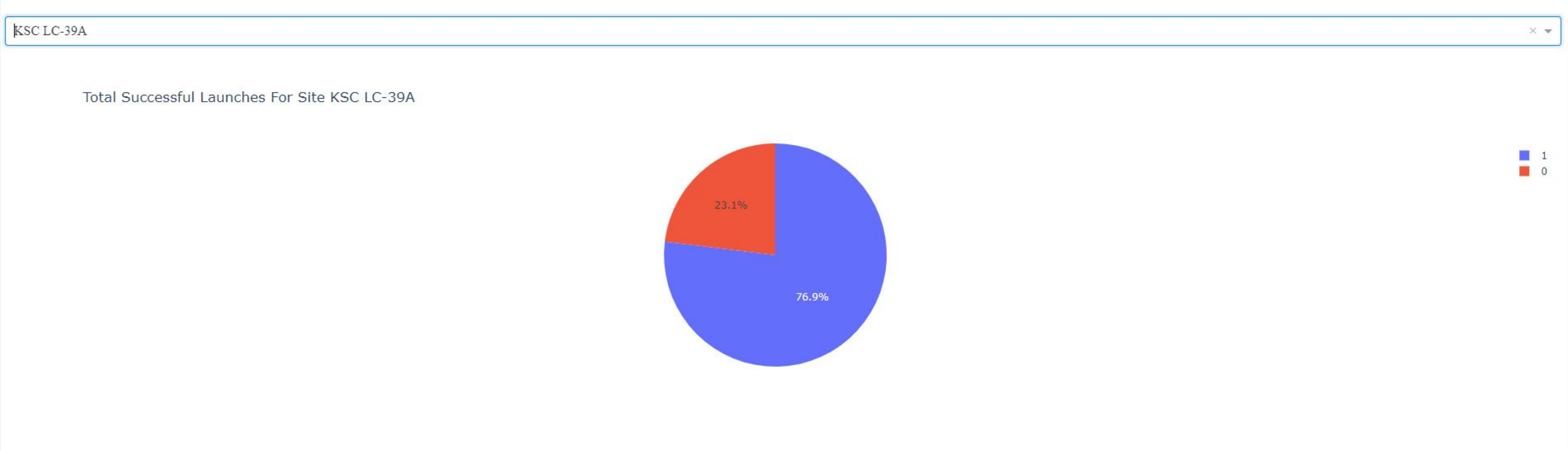


Successful Launches by launch site



- The pie chart shows that the most number of successful launches happened on the KSC LC-39A.
- Another observation is that the success rate of launch site KSC LC-39A is more than the cumulative success rate of VAFB SLC-4E and CCAFS SLC-40.

Successful Launches for a specific site

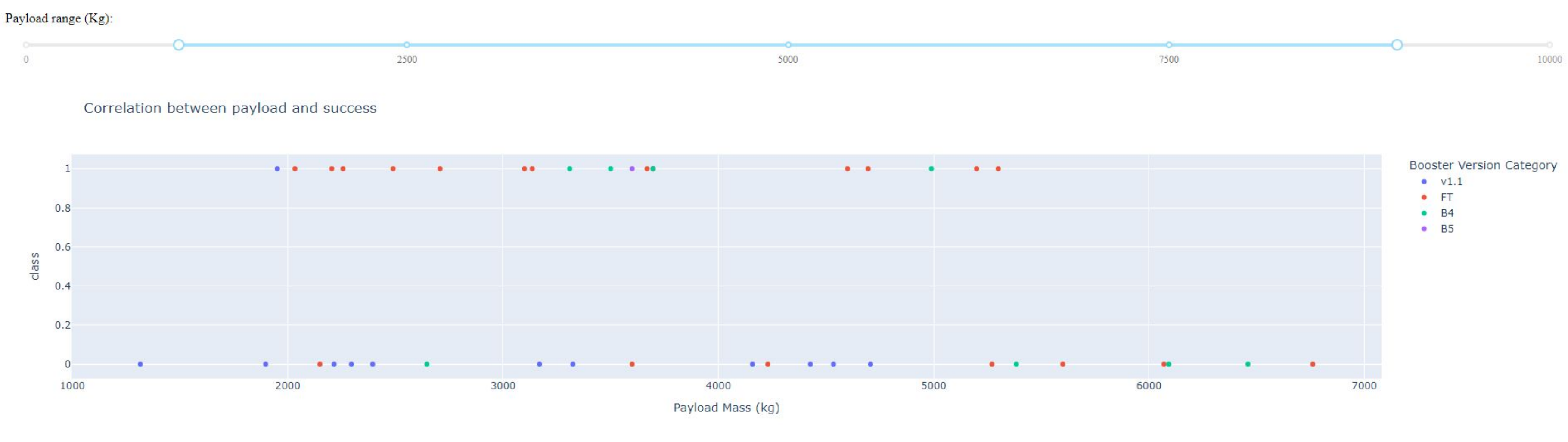


KSC LC-39A

Total Successful Launches For Site KSC LC-39A

- We see through the pie charts that launch site KSC LC-39A has the highest launch success ratio.
- The ratio itself shows that more than 75%(¾ ths) of launches have succeeded for launch site KSC LC-39A.

Correlation between Payload vs Launch success rate



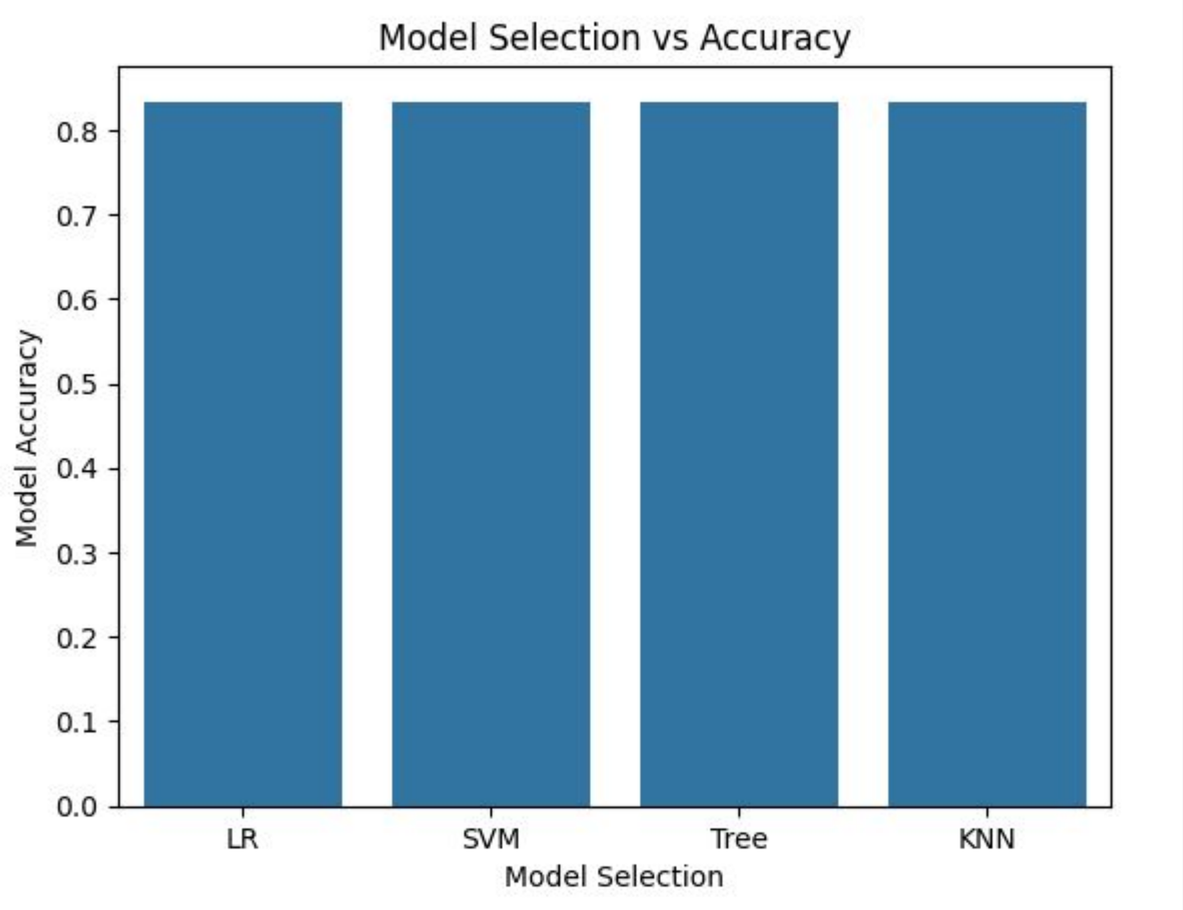
- The graph displays booster success rate for payload range 1000 to 7000.
- We can see that FT boosters have a higher success rate for low to medium payloads, falling off only towards the heavy payloads. Also, FT boosters, seem to perform best among all boosters.
- We also see that v1.1 boosters show a general trend of a low rate of success across all values of payloads.
- B4 boosters have a more balanced success rate for the given payload range.

Section 5

Predictive Analysis (Classification)

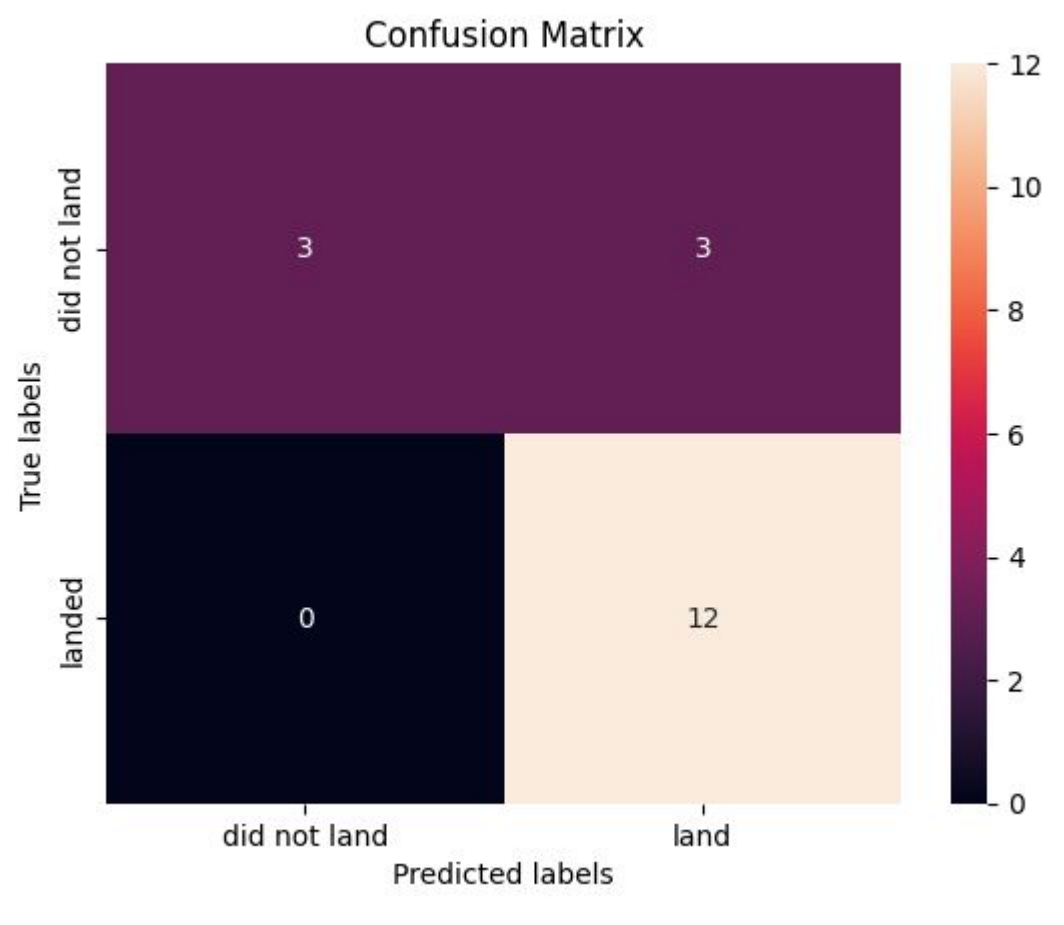
Classification Accuracy

- We discover that all models have the same classification accuracy, so for the given data, all models performed equally well.



Confusion Matrix

- Since all models performed the same, they all produced the same confusion matrix.
- Looking at the matrix we can see that the True label ‘landed’ which has a count of 12 have all been correctly predicted.
- For the True label ‘did not land’ or failed landing, we see that only half the labels have been correctly predicted, the other half show as landed when they did not land(False Positive).



Conclusions

- Exploratory Data Analysis helped us figure out that there are a lot of variables that have a significant impact on whether the launch and subsequent Falcon 9 first stage landing is successful.
- Visualizations such as scatter point charts made understanding the relationships between the variables a lot easier.
- Dashboarding helped us present out data to other people who would be looking at it in a informative and visually appealing manner.

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

Thank you!