IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zain Naseer
26/11/2022

Outline

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

Executive Summary

- The steps followed in this project closely abided with the CRSIP-DM framework.
- Initially the data was collected, then data wrangling was carried out. Wrangling involved exploratory data analysis, cleaning, and preparation of the data for modelling.
- In the modelling stage four classification models were built, and the models were evaluated to choose the best classification model.

Introduction

SpaceX can launch rockets at a comparatively cheaper price to other providers. SpaceX's Falcon 9 rockets cost 62 million dollars; other providers cost upwards of 165 million dollars. Much of the savings is because SpaceX can reuse the first stage. Therefore, determining whether the first stage will land can help determine the cost of a launch.

In this capstone , I will be taking the role of a data scientist working for SpaceY. The challenge is to determine the price of each launch and to determine if SpaceX will reuse the first stage. Machine learning will be used to determine if SpaceX will reuse the first stage.

Section 1

Methodology

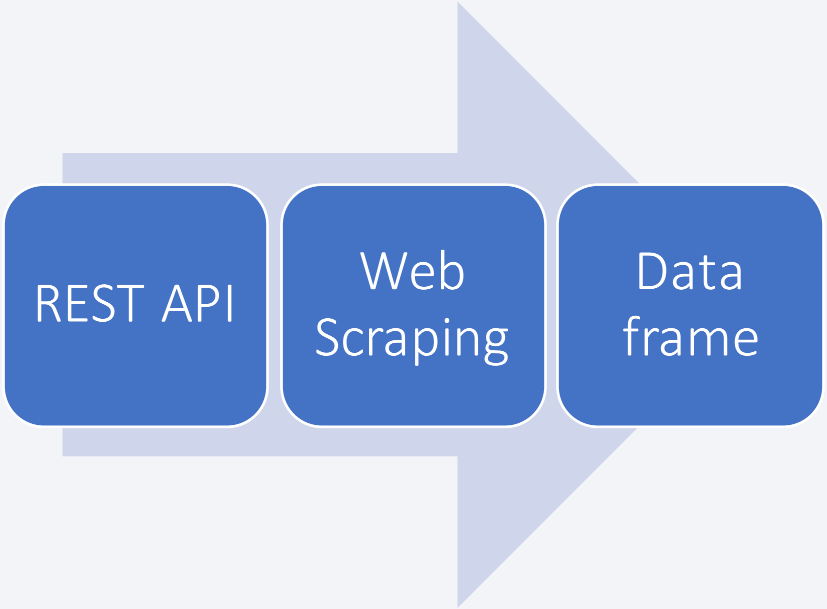
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through web scraping and the use of a REST API.
- Perform data wrangling
  - Exploratory data analysis was performed, and training labels were created.
- 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 classification models were trained on training data and tested on test data - with the use of confusion matrices and the mean accuracy metric.

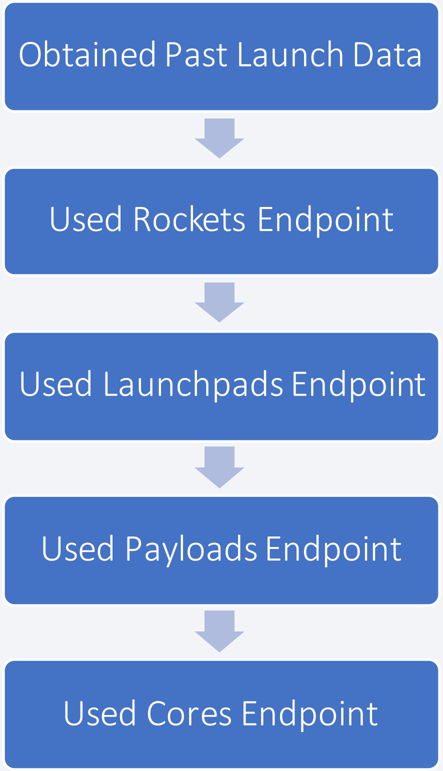
Data Collection

- Web scraping and a REST API was used to collect the data. Then the data was converted into a Pandas data frame.



Data Collection – SpaceX API

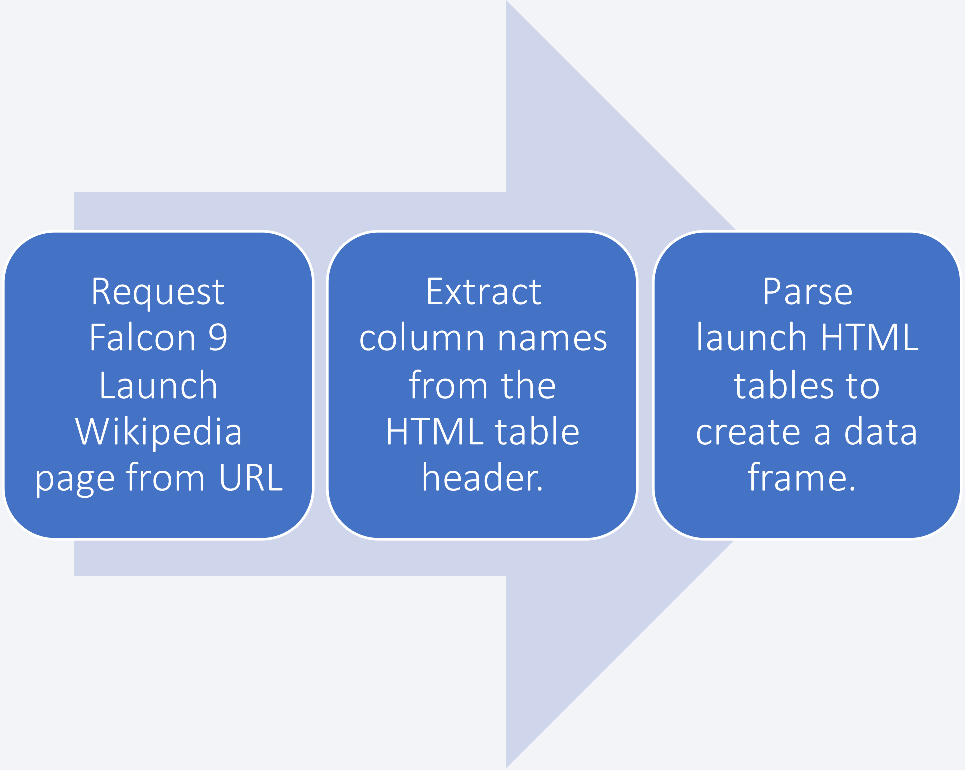
- A particular URL was used to target a specific endpoint of the API to get past launch data.
- In some columns there was not actual data but identification numbers; so, helper functions were used to use the API again, targeting a different endpoint to gather specific data.
- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

- In this lab a Falcon 9 launch records HTML table was extracted from Wikipedia. The table was parsed and then converted into a Pandas data frame.

- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Request Falcon 9 Launch Wikipedia page from URL

Extract column names from the HTML table header.

Parse launch HTML tables to create a data frame.

Data Wrangling

- Exploratory data analysis was performed. The value_counts() method was used to calculate the number of launches on each site, calculate the count of each orbit type , and to tally the mission outcomes.
- Landing outcome training labels were formed from the outcome column.

Exploratory Data Analysis
Form Training Labels

- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots were beneficial for seeing the relationship between two variables. Also, colouring by class label allowed further insight.
- Bar charts were plotted to visualise any association between orbit type and average success rate of the booster (first stage) landing successfully.
- Line plots are useful for visualising changes in data over time, so a line plot was used to visualise any association between year and average success rate.
- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Pandas%20and%20Matplotlib.ipynb

EDA with SQL

- Some of the following SQL queries were used:
  - Simple SELECT queries
  - SELECT queries with functions
  - SELECT queries in conjunction with functions and multiple clauses such as WHERE, GROUP BY and ORDER BY
  - Subqueries for more complex data extraction

- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- All launch sites were marked on a folium map using a circle object.
- Successful and failed launches for each launch site were marked using a cluster object.
- A mouse position object was added to the folium map to discover coordinates of points of interest.
- The distance between a launch site and its proximities was visualised with a polyline object.

- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

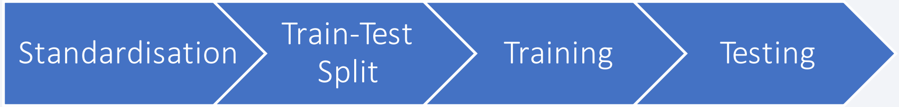
Plots and interactions that have been added to the dashboard:

- A launch site drop down input component was added to the dashboard
- A callback function to render a pie chart showing total successful launches by launch site
- A range slider is added to select payload
- A callback function to render a scatter plot of payload vs launch outcome, with each data point coloured by the booster version.

https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/tree/main/Interactive%20Dashboard%20with%20Plotly%20Dash

Predictive Analysis (Classification)

- The data was standardised, and a train-test data split was performed.
- Grid search was used to choose the best hyperparameters for each machine learning (ML) algorithm.
- Each of the ML algorithms were trained on the training data, and then the algorithms were evaluated on the test data using the mean accuracy metric.
- A confusion matrix was also formed for each algorithm, for deeper evaluation.



Standardisation

Train-Test Split

Training

Testing

- https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

Results

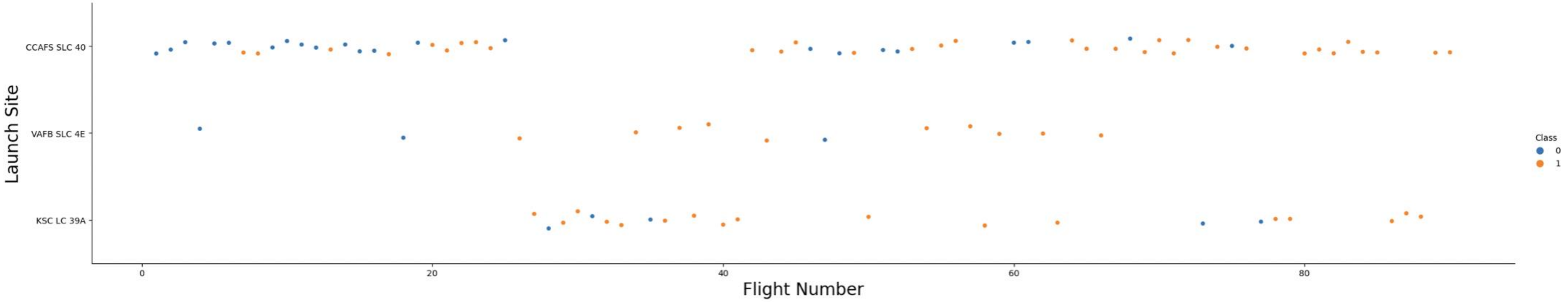
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

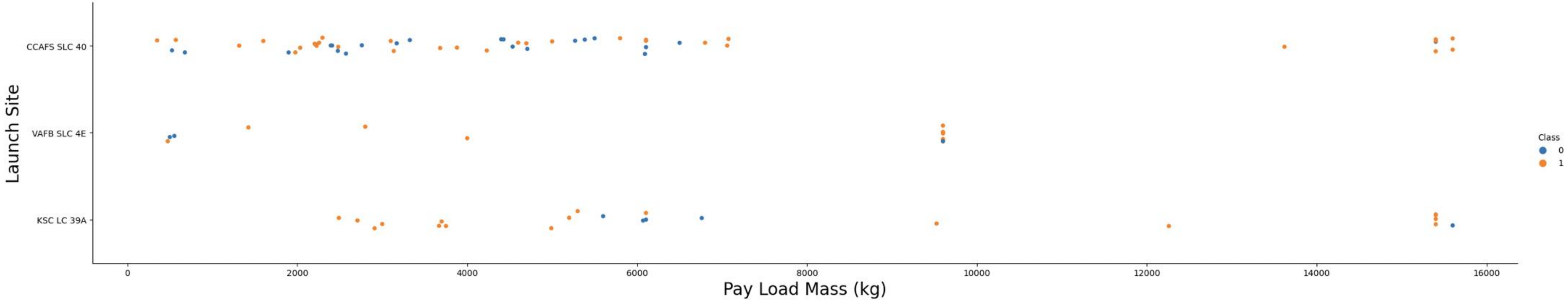
Flight Number vs. Launch Site

- For each launch site as flight number increases the chance that the first stage lands successfully is greater.
- CCAFS SLC 40 has an overall lower success rate compared to the other launch sites.



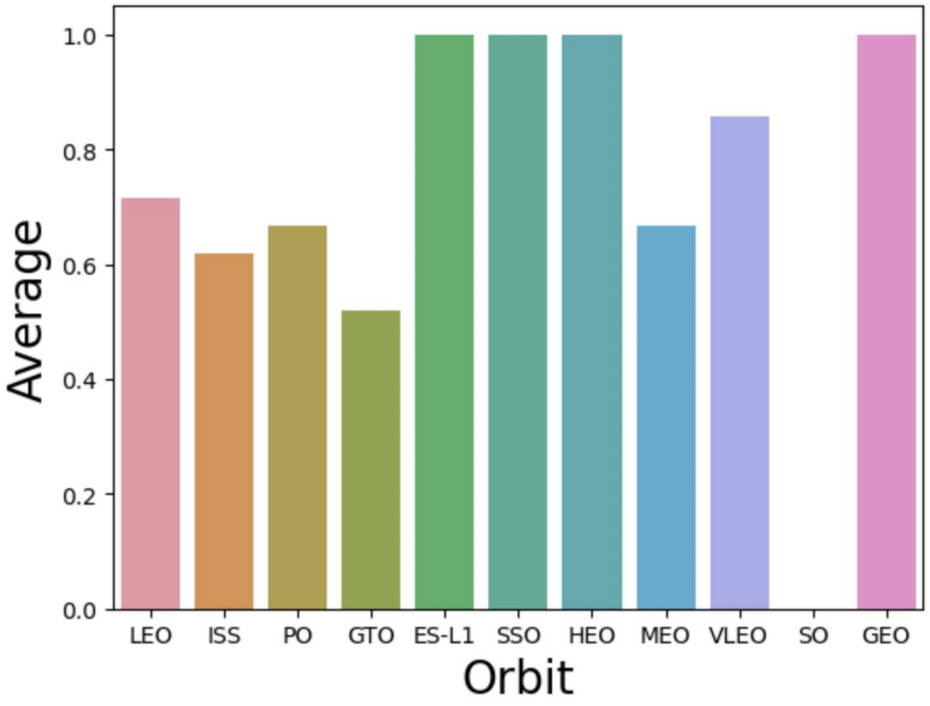
Payload vs. Launch Site

- For the VAFB SLC 4E launch site there were no rockets launched with pay load mass greater than 10000kg.
- For other launch sites rockets of various pay load mass were launched.



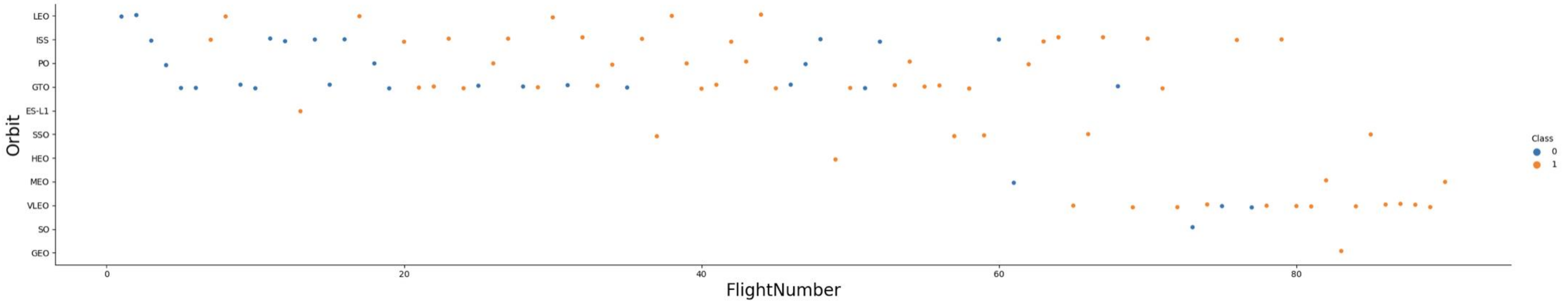
Success Rate vs. Orbit Type

- Orbit type GEO,ES-L1, SSO, HEO have the highest average success rate.
- Orbit type GTO has the lowest average success rate.



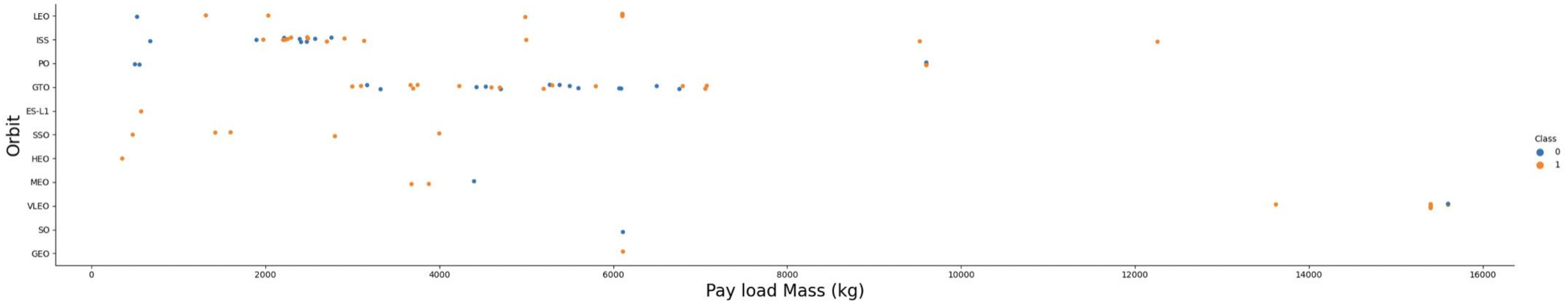
Flight Number vs. Orbit Type

- As flight number increases for orbit type LEO, the chances of success is higher.
- As flight number increases above 60, orbit type VLEO is more common.



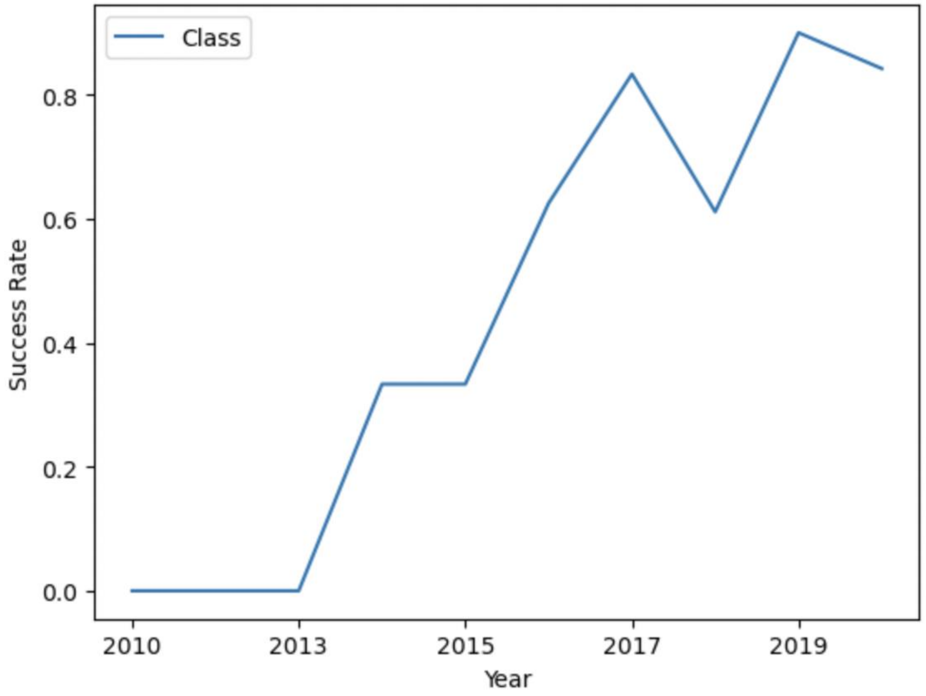
Payload vs. Orbit Type

- For launch site LEO, success rate is greater as pay load mass increases.
- For orbit type GTO, as pay load mass increases there is no effect on success rate (for better or worse).



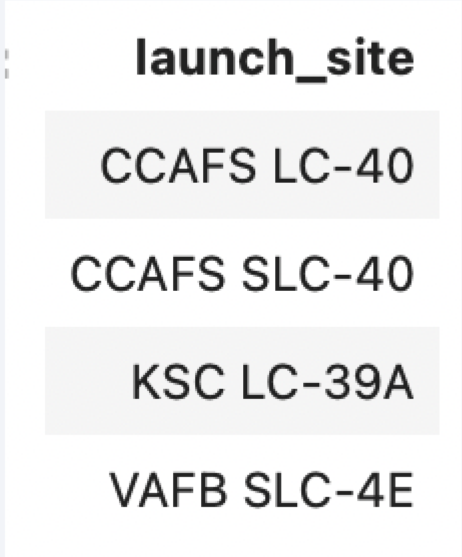
Launch Success Yearly Trend

- Yearly average success rate has increased since 2013, although there was a dip in success rate between 2017 and 2018.



All Launch Site Names

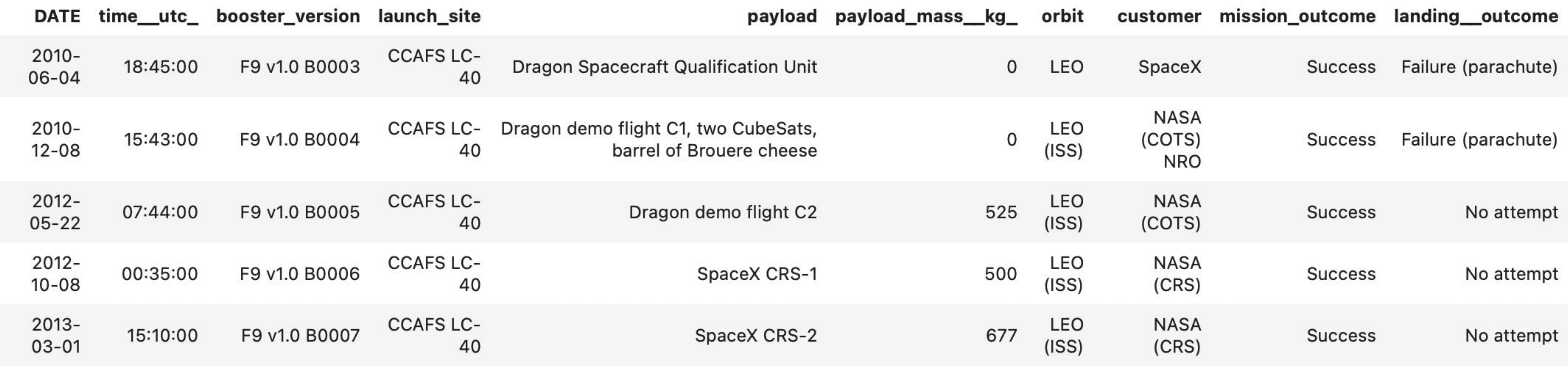
- Four unique launch sites were identified:



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

Launch Site Names Begin with 'CCA'

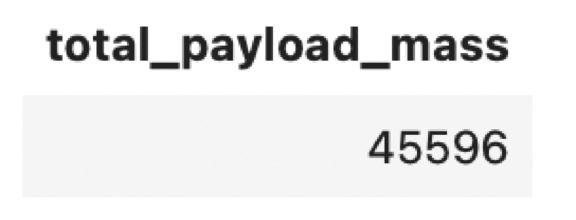
- A SELECT statement was used with a WHERE clause to identify launch site names which begin with 'CCA'. The result set was restricted to five rows.



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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

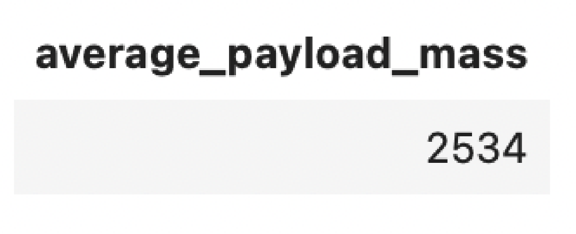
- A SELECT statement with a SUM function and WHERE clause was used to find the total payload mass.



total_payload_mass
45596

Average Payload Mass by F9 v1.1

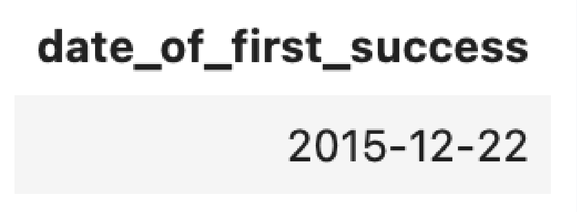
- A SELECT statement with the AVG function and a WHERE clause was used to calculate the average pay load mass by booster version F9v1.



average_payload_mass
2534

First Successful Ground Landing Date

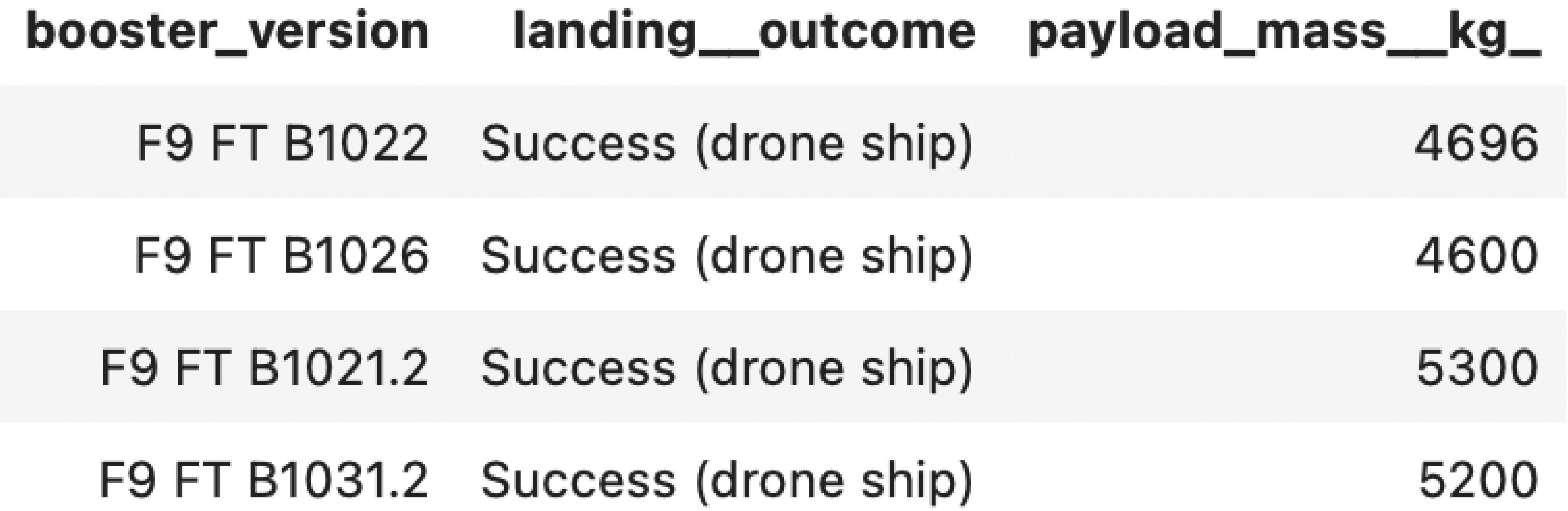
- By using the MIN function and WHERE clause in the SELECT statement, the date of the first successful landing outcome in ground pad was obtained.



date_of_first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

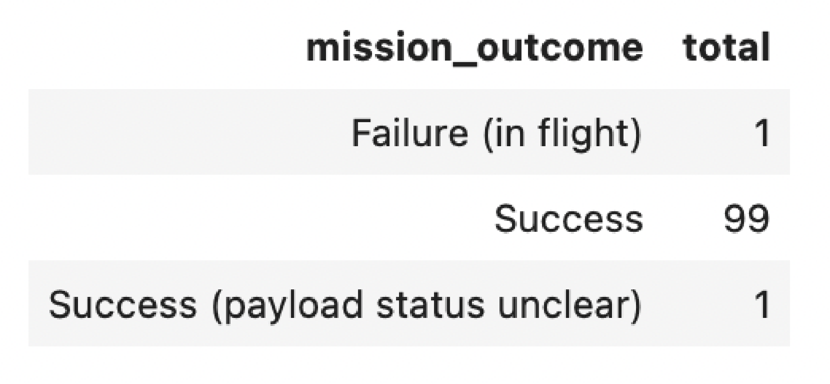
- A SELECT statement with a WHERE clause was used to obtain successful drone ship landings with payload between 4000 and 6000.



booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

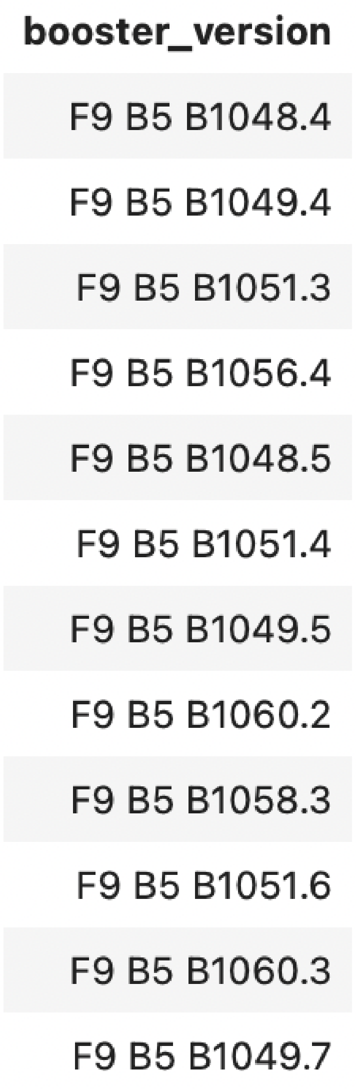
- A GROUP BY clause within the SELECT statement was used to obtain the total number of successful and failure mission outcomes.



mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters that Carried Maximum Payload

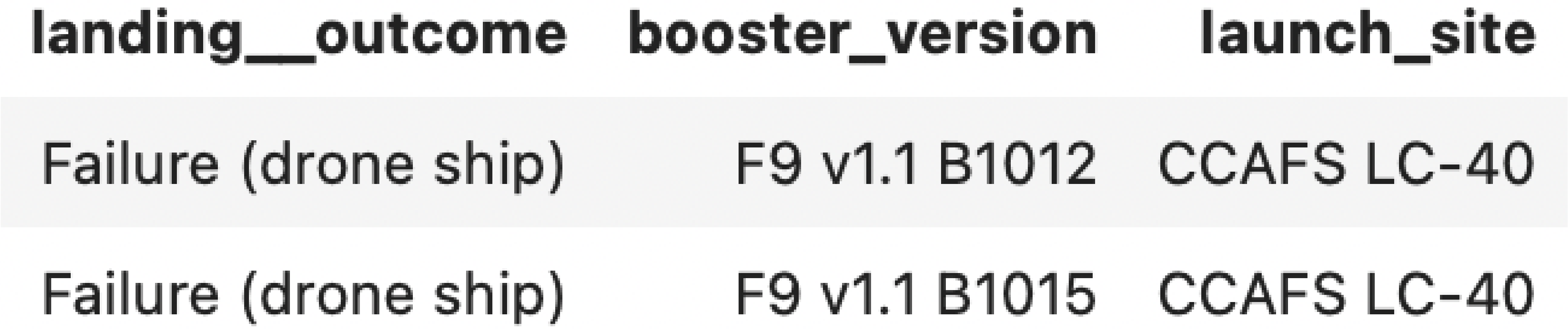
- A subquery was used to obtain the names of the booster versions that have carried the maximum payload mass.



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

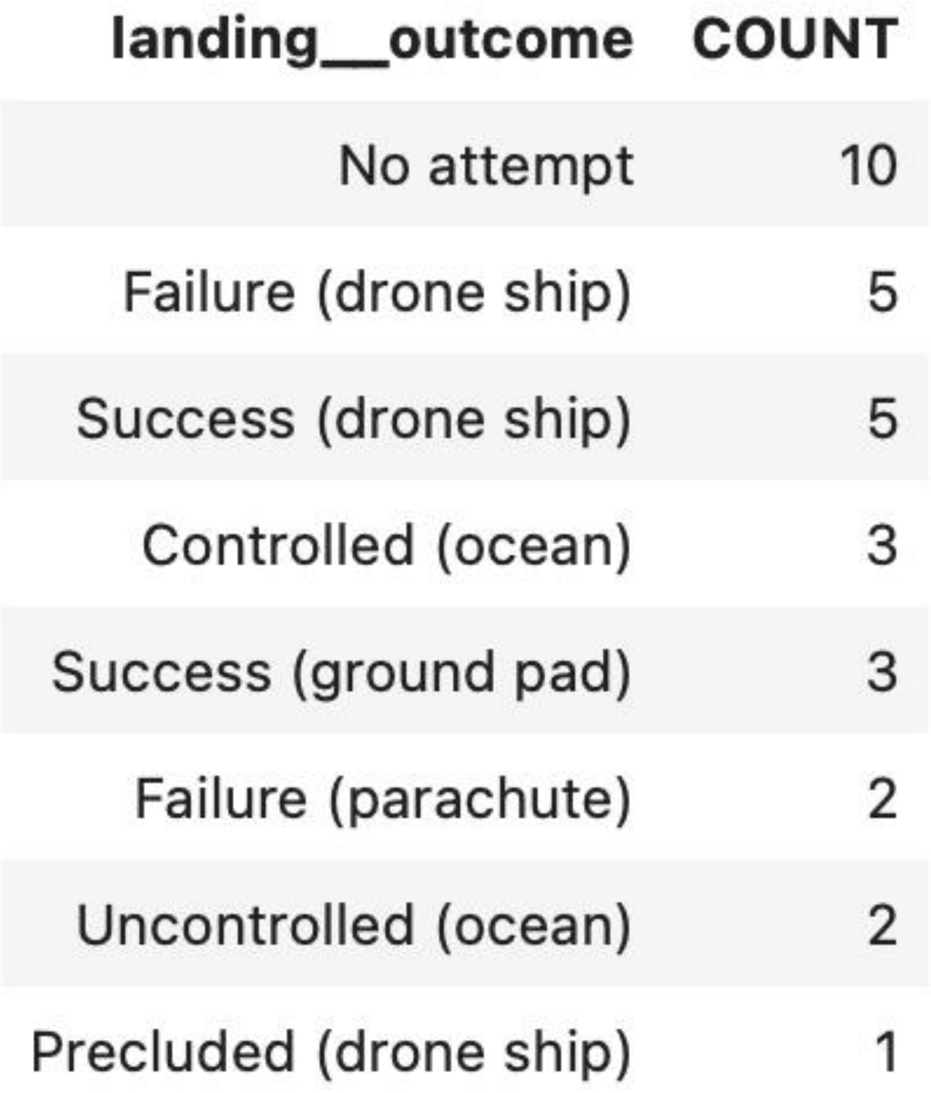
- A SELECT statement was used with the required column names and WHERE clause to list the failed landing outcomes in 2015.



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

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

- A SELECT statement with the WHERE, GROUP BY, and DESC commands was used to rank the landing outcomes between 2010-06-04 and 2017-03-20.



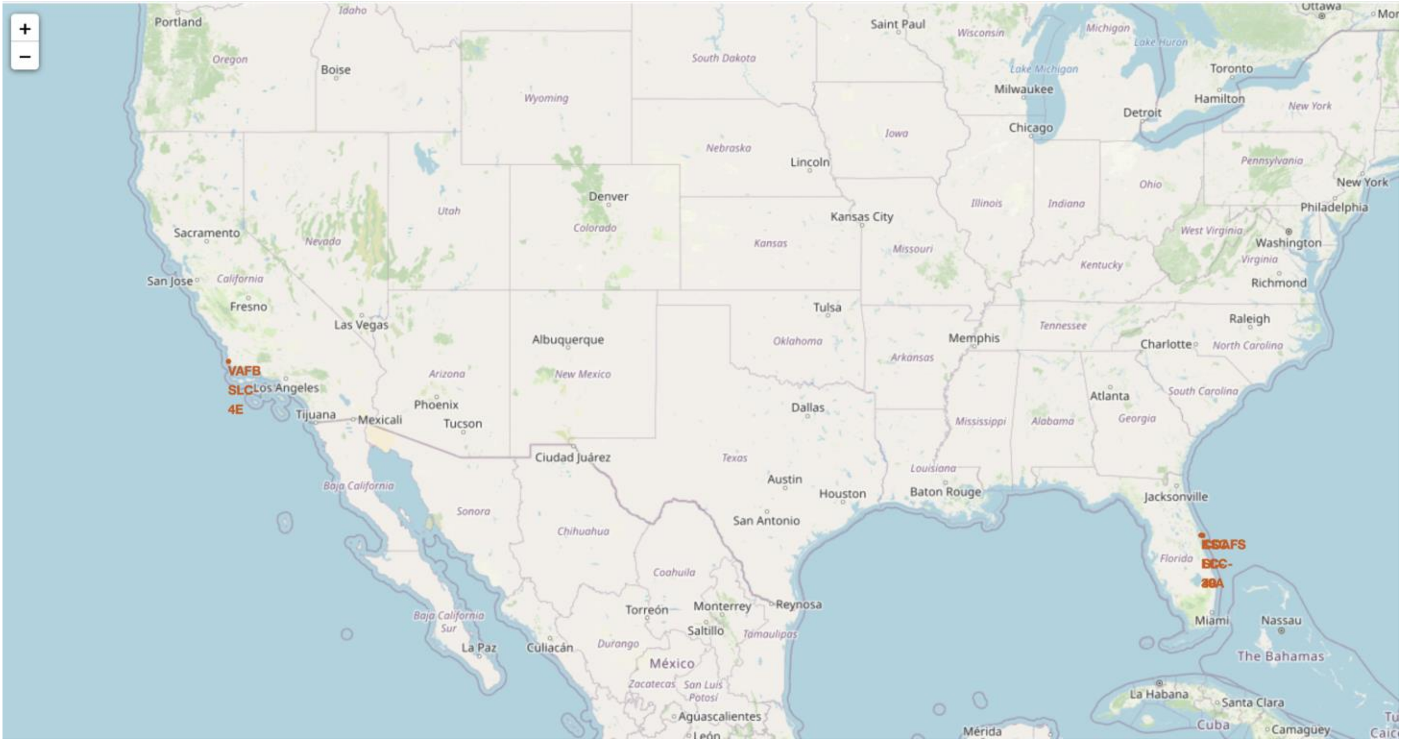
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

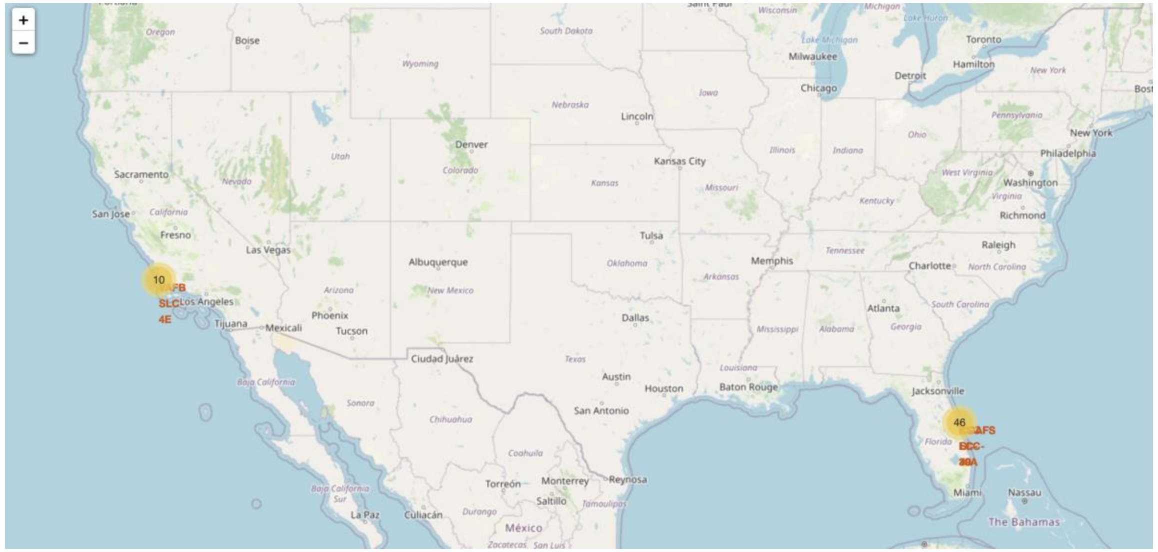
Launch Site Locations

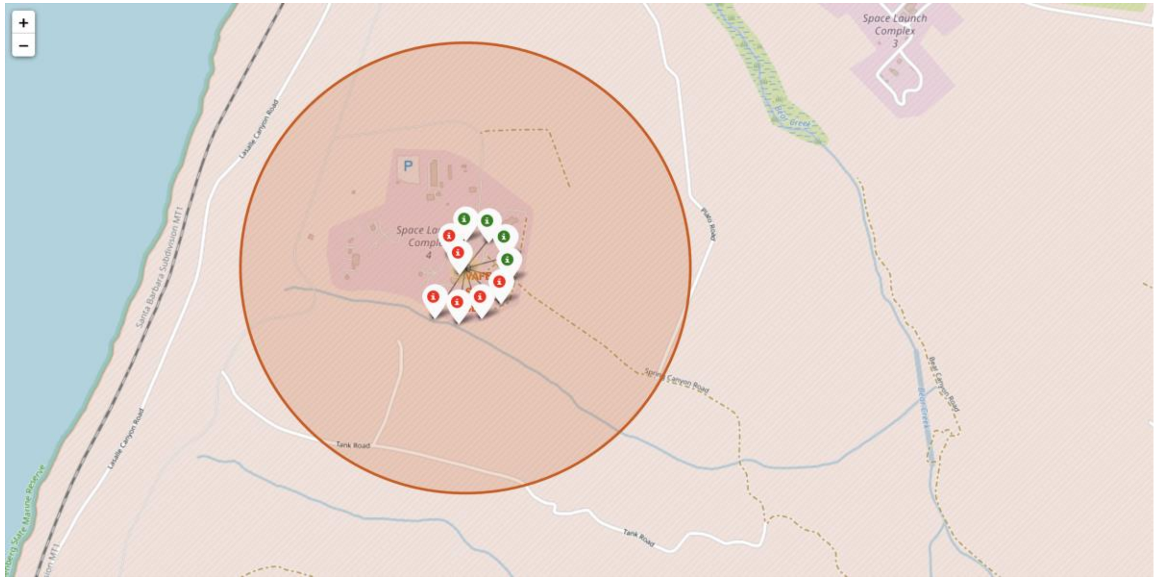
- This Folium map showed that launch sites were always near the coast and close to the equator line.



Successful and Failed Launches for each Launch Site

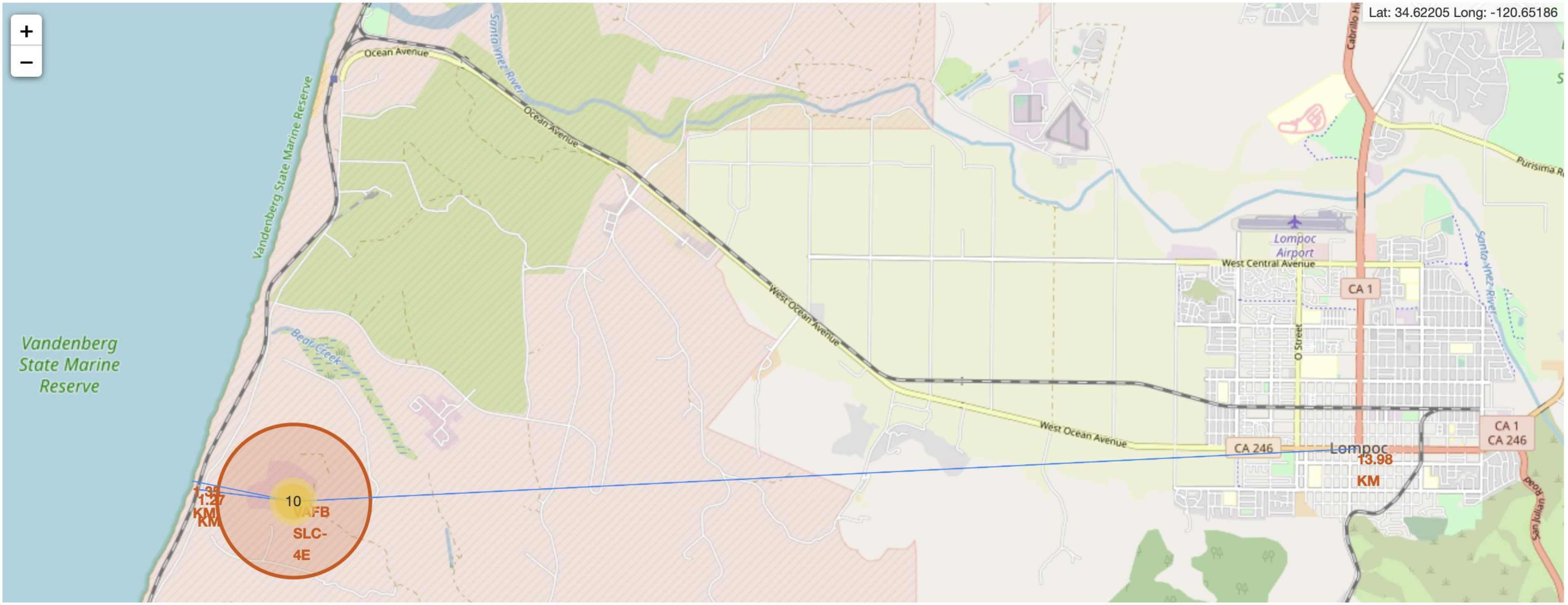
- The zoomed-out Folium map displays the number of overall launches from each launch site.
- For launch site VAFB SLC-4E the successful and unsuccessful launch sites are visible in green and red, respectively.

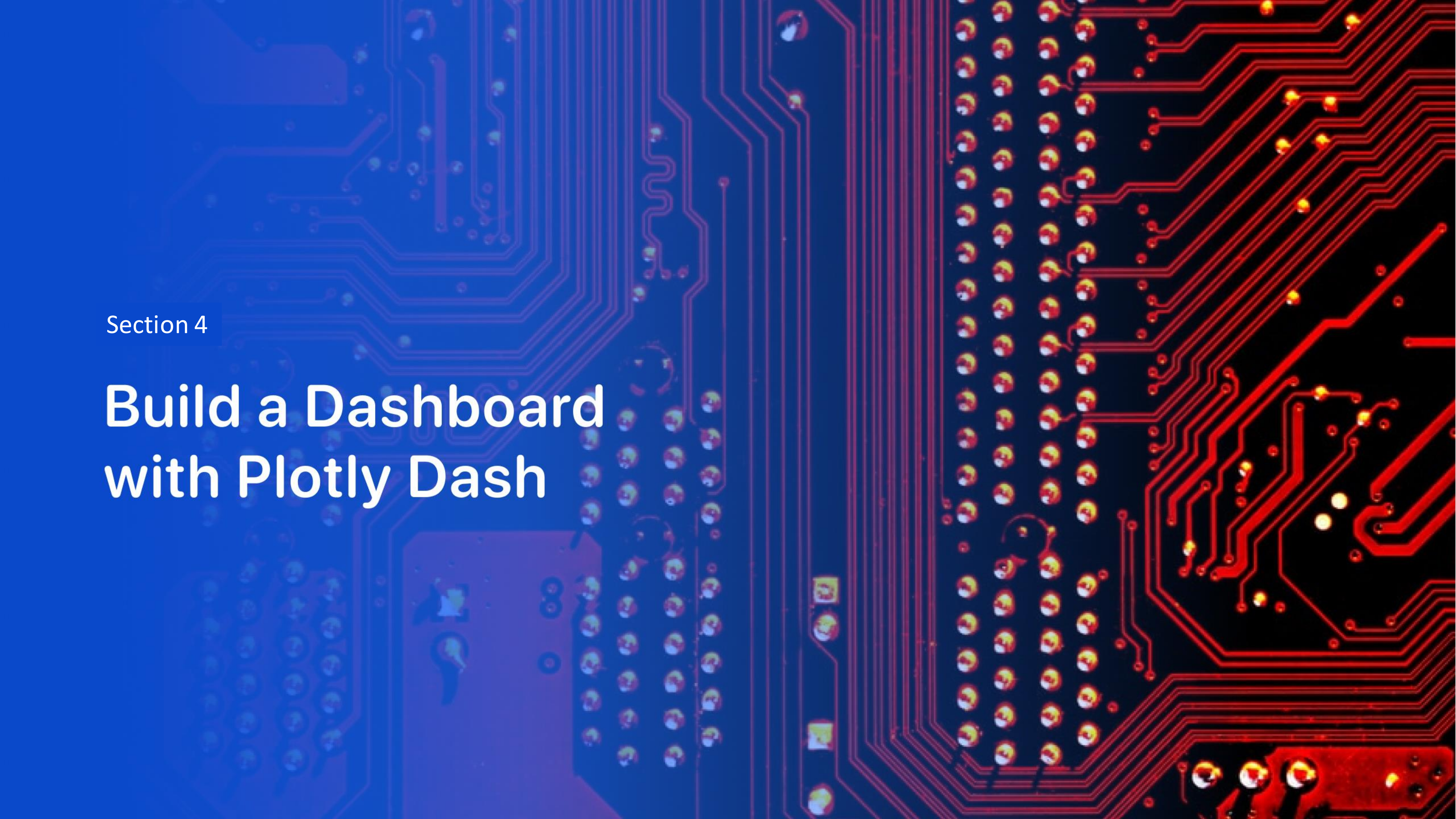




Distance from Launch Site to Proximities

- The Folium map displays the distance of the VAFB SLC-40 launch site to its proximities.
- The map showed that the launch site was away from cities, but close to the coastline. Also, the launch site was close to the railway.
- In general, all launch sites were near the railway or/and highway.



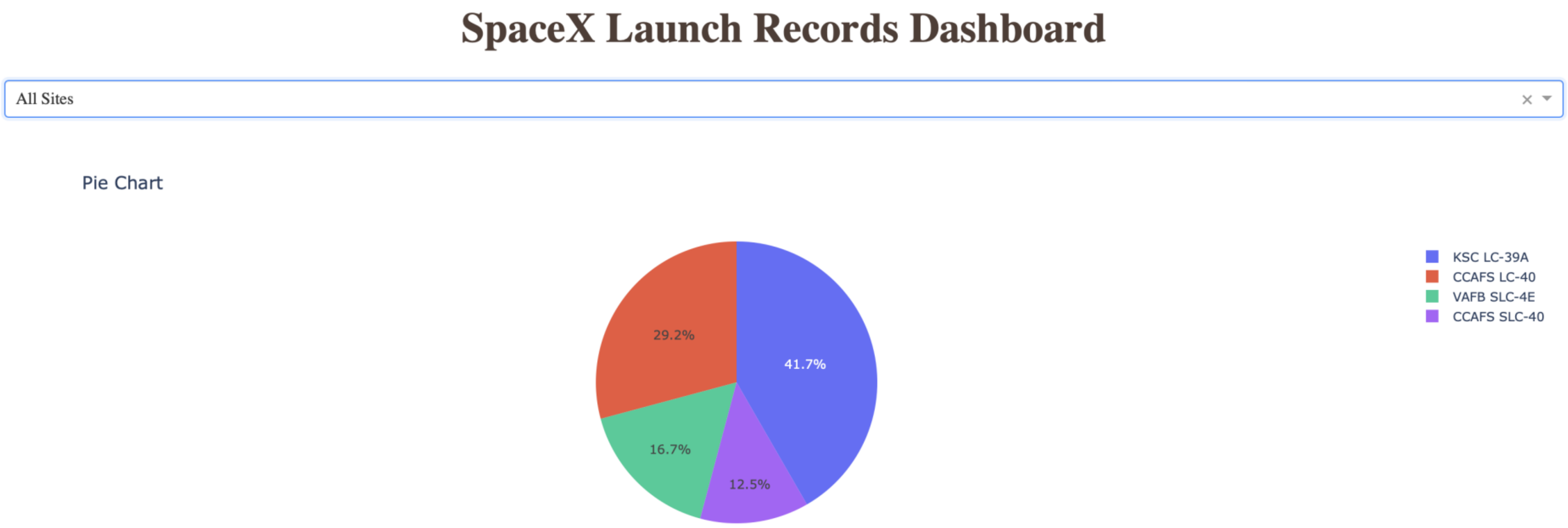


Section 4

Build a Dashboard with Plotly Dash

Pie Chart of Launch Success for all Sites

- The percentage of launch success for the launch site KSC LC-39A was 41.7%, which was greater than any other launch site.
- The percentage of launch success for CCAFS SLC-40 was 12.9%, which was lower than any other launch site.



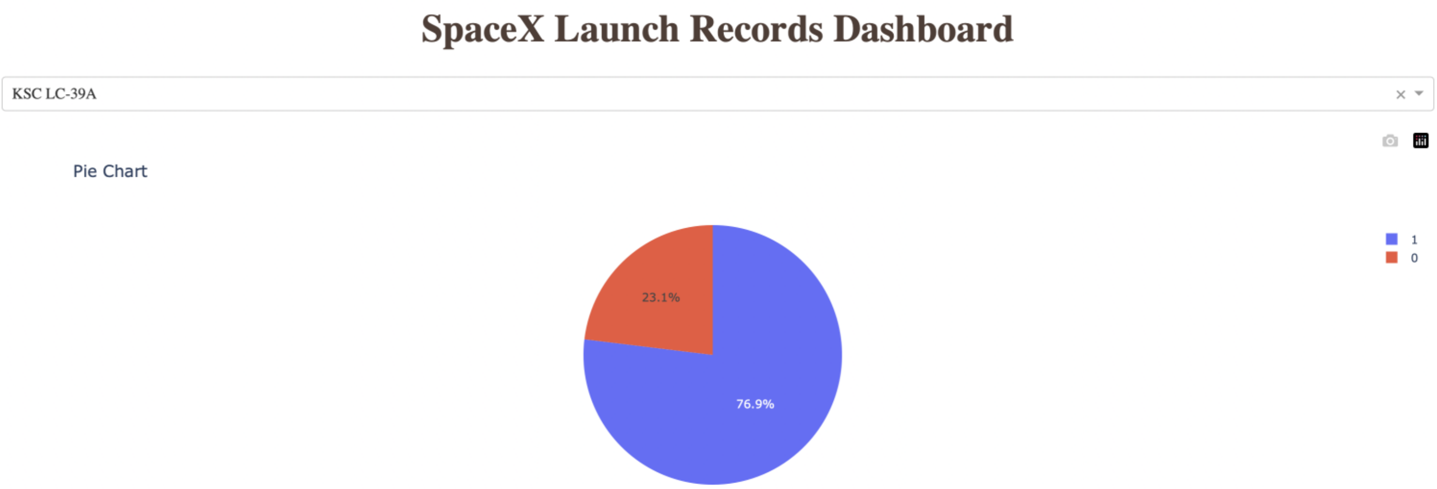
SpaceX Launch Records Dashboard

All Sites

Pie Chart

Pie Chart of Launch Outcome for KSC LC-39A

- The launch site KSC LC-39A had the highest launch success percentage, with a percentage of 76.9%.

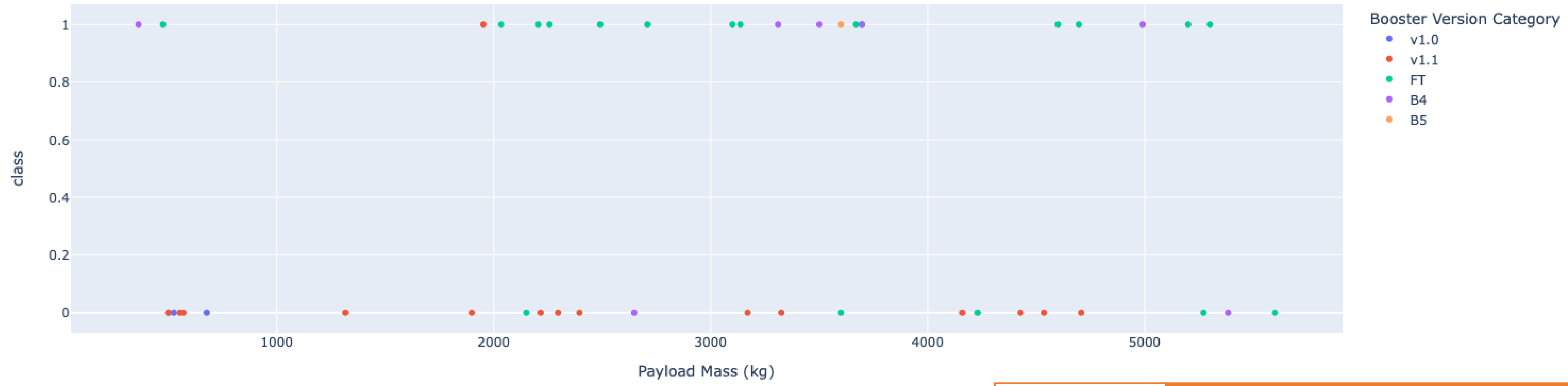


SpaceX Launch Records Dashboard

KSC LC-39A

Scatter Plot of Payload vs Launch Outcome for all Sites.

- Given below is the scatter plot for the *Payload Mass (kg)* between 0 and 6000.
- From the graph it is seen that *Payload Mass (kg)* between 2000 and 3500 gives a high chance of a successful launch for *Booster Version Category* 'FT'.



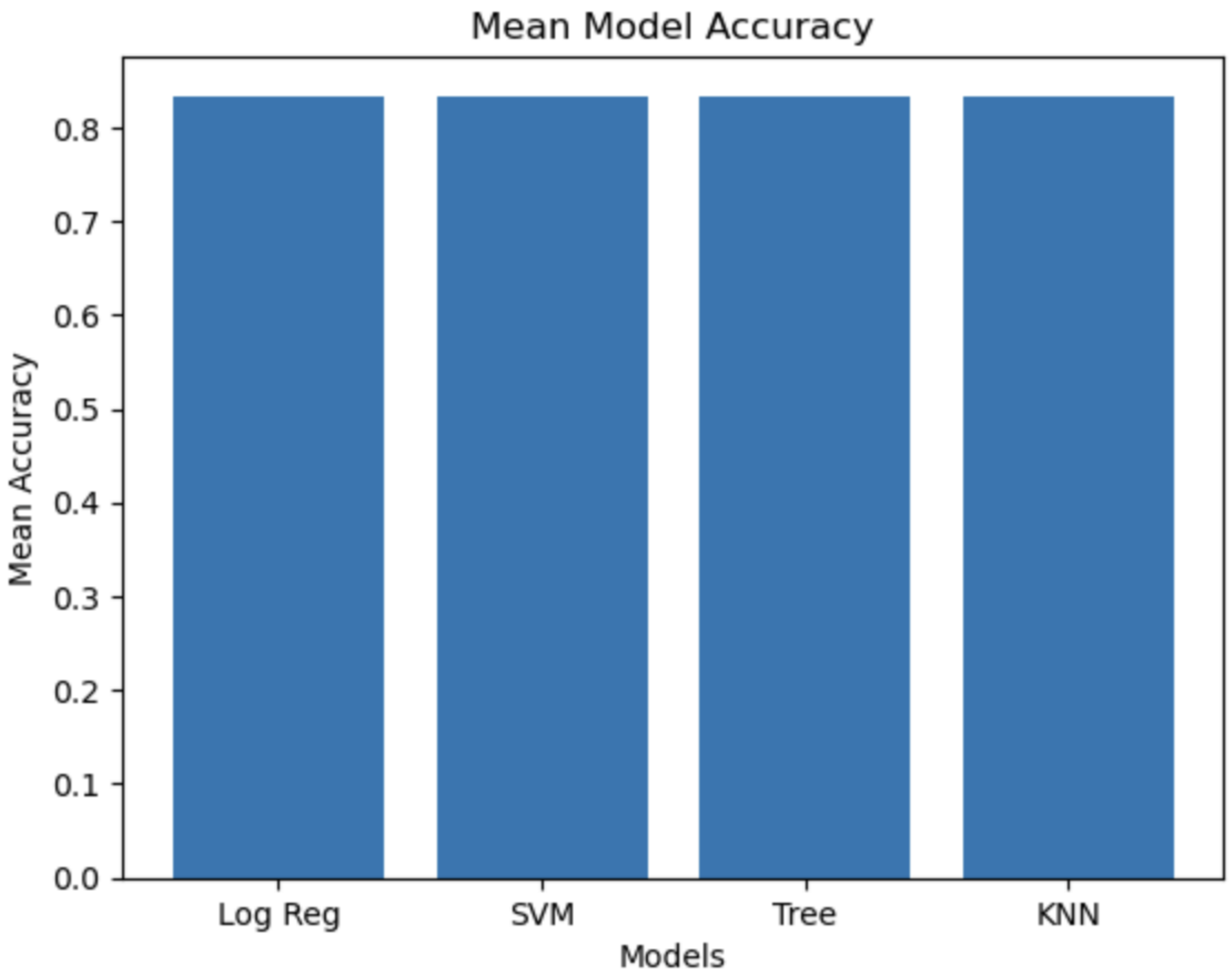
Replace	Replace <Dashboard screenshot 3> title with an appropri
Show	Show screenshots of Payload vs. Launch Outcome scatte different payload selected in the range slider

Section 5

Predictive Analysis (Classification)

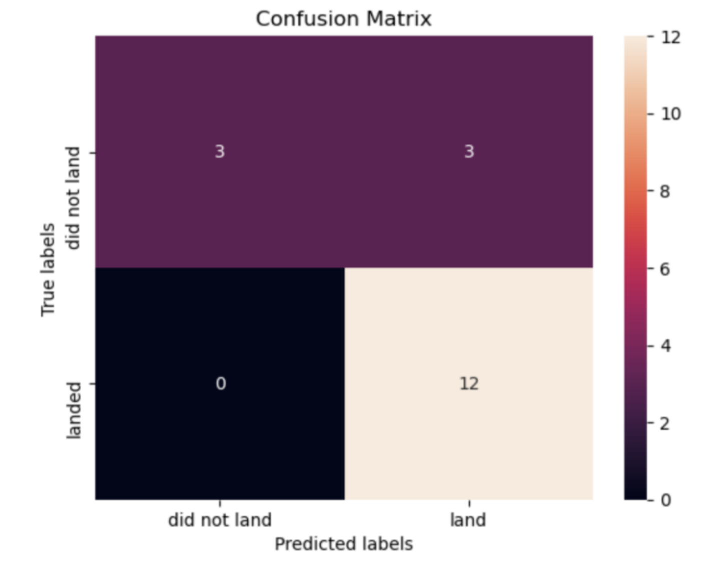
Classification Accuracy for Each Model

- All the models had the same mean accuracy, due to the test dataset being small.



Confusion Matrix

- The same confusion matrix was obtained for all models.
- Out of six failed landings, all classifiers correctly predicted three as failed landings and incorrectly predicted three as successful landings.
- All twelve successful landings were correctly predicted by all classifiers.



Conclusions

- Overall, the models that were built were fairly accurate.
- The dataset given for this project was not very large, this could mean that not enough variation was accounted for in the training of the models.
- Even better performing models might have been obtained through using Artificial Neural Networks.

Appendix

- Link to project repository:

https://github.com/ZainN123/IBM-Course-10-Applied-Data-Science-Capstone

Thank you!