IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jean-Baptiste GOMEZ

25/03/2023

Outline

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

Executive Summary

- Summary of methodologies :
  - Data Collection using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- Summary of all results :
  - Valuable data was collected from public sources.
  - EDA, the best features for predicting the success of launchings were identified.
  - The most effective model for predicting important characteristics and maximizing opportunities was determined Machine Learning Prediction using all collected data.

Introduction

- The objective is to assess the feasibility of the new company Space Y to compete with Space X.
- Desired answers:
  - The most effective way to estimate the total cost of launches by predicting successful landings of the first stage of rockets.
  - The best location for conducting launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API : https://api.spacexdata.com/v4/rockets/

      https://api.spacexdata.com/v4/launchpads/

      https://api.spacexdata.com/v4/payloads/

      https://api.spacexdata.com/v4/cores/

    - WebScraping: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Methodology

Executive Summary

- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing 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
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

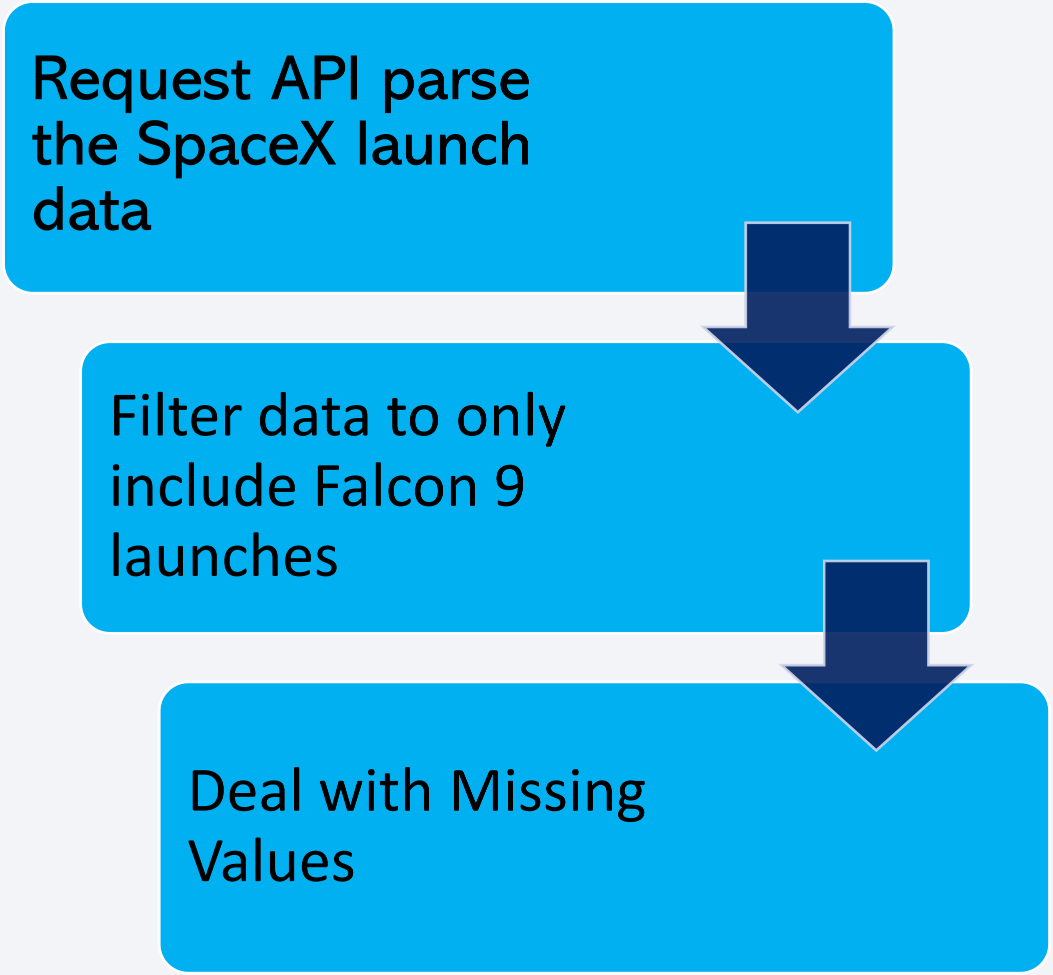
Data sets were collected from Space X API and from Wikipedia using web scraping technics like BeautifulSoup.

Data Collection – SpaceX API

- SpaceX has made available a public API that allows for the retrieval and utilization of data;
- This API was used according to the flowchart beside and then data is persisted.

Source code:

https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Request API parse the SpaceX launch data

Filter data to only include Falcon 9 launches

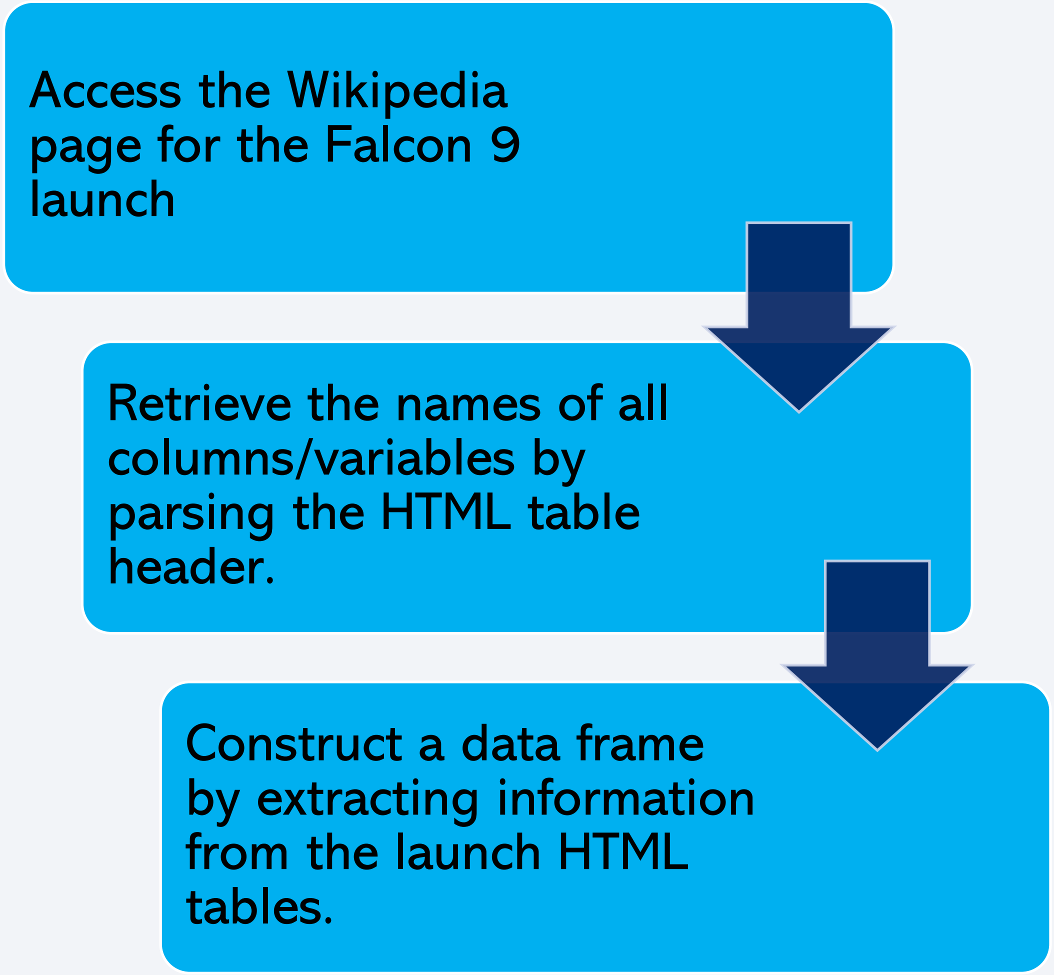
Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia for scraping;
- The flowchart specifies the process of retrieving data from Wikipedia, which is then saved for future reference.

Source code :

https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/jupyter-labs-webscraping.ipynb



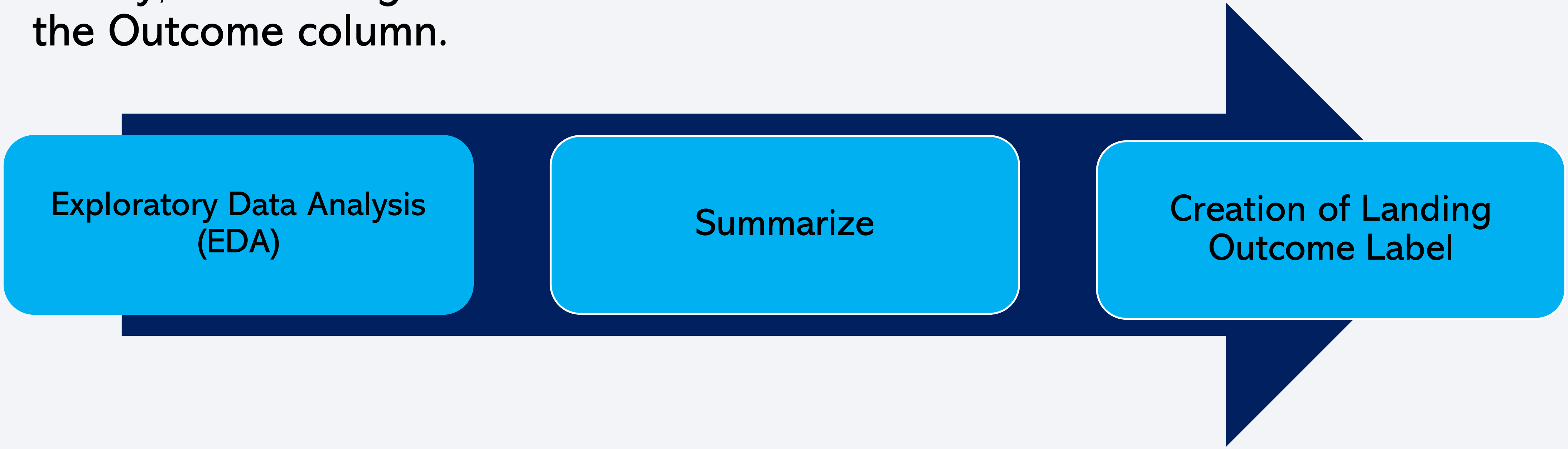
Access the Wikipedia page for the Falcon 9 launch

Retrieve the names of all columns/variables by parsing the HTML table header.

Construct a data frame by extracting information from the launch HTML tables.

Data Wrangling

- First, an Exploratory Data Analysis (EDA) was conducted on the dataset.
- Then, the number of launches per site, the occurrences of each orbit, and the occurrences of mission outcomes per orbit type were calculated.
- Finally, the landing outcome label was created based on the information in the Outcome column.



Exploratory Data Analysis (EDA)

Summarize

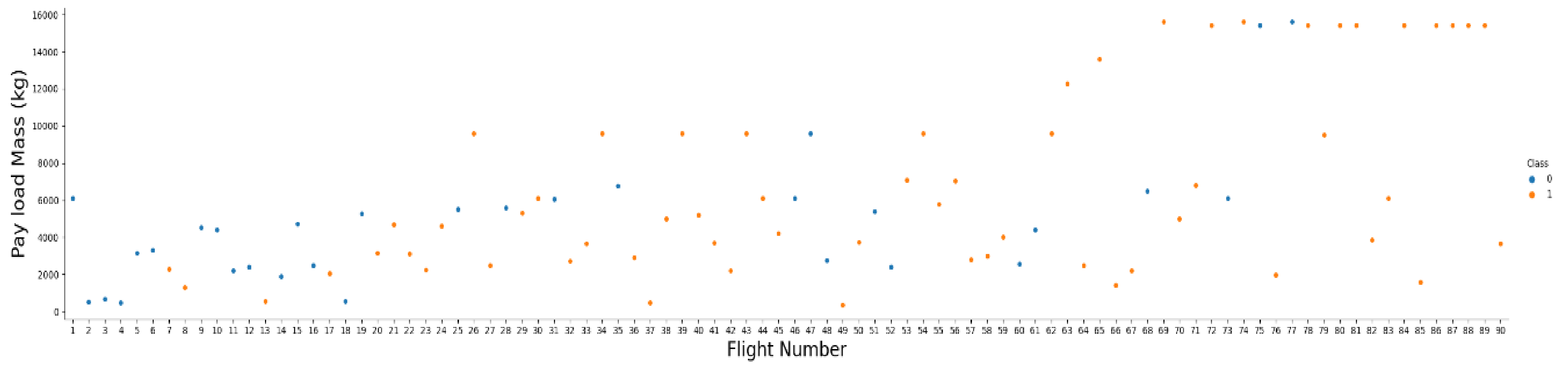
Creation of Landing Outcome Label

Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Scatterplots and barplots were utilized for visualizing the correlation between pairs of features in order to conduct data exploration:

- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit.



Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA’;
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions;
- launch site names for in year 2015; and Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

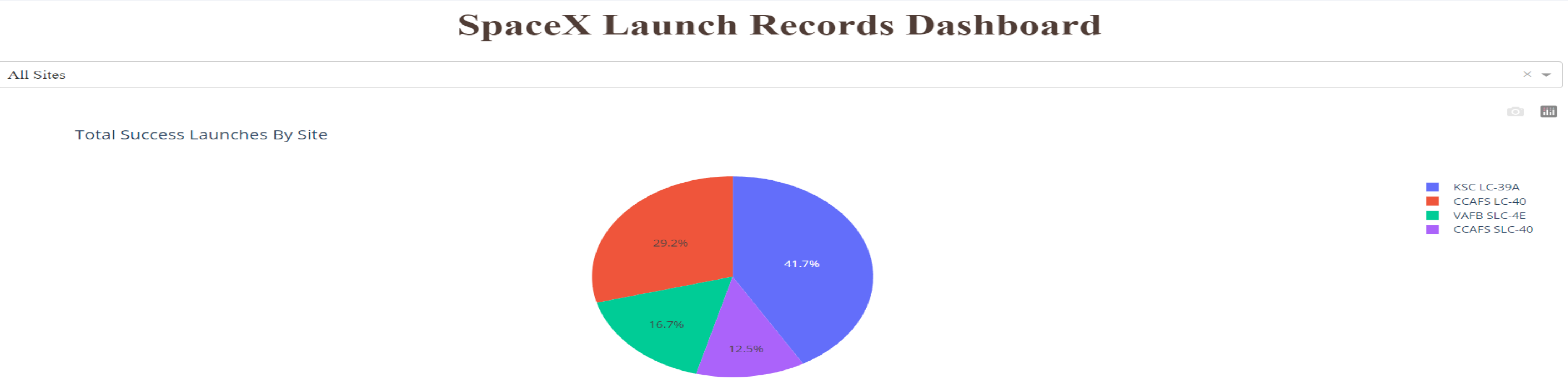
Folium Maps were employed with markers, circles, lines, and marker clusters to convey information:

- Markers were utilized to represent points such as launch sites;
- Circles were employed to highlight specific areas around particular coordinates, such as NASA Johnson Space Center;
- Marker clusters were used to group events in each coordinate, such as launches at a given launch site;
- Lines were utilized to display distances between two coordinates.

Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The data was visualized through graphs and plots such as:
  - The percentage of launches by site
  - Payload range.
- This approach enabled a prompt analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch site based on payload.



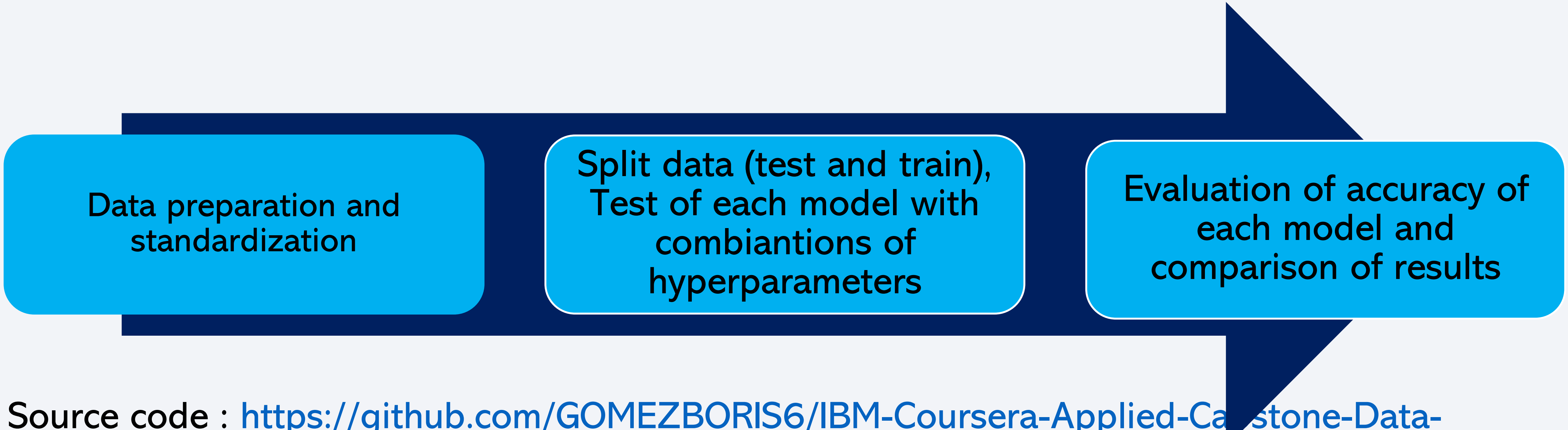
SpaceX Launch Records Dashboard

All Sites

Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/Dash_with_plotly.py

Predictive Analysis (Classification)

Four distinct classification models were evaluated and compared, namely logistic regression, support vector machine, decision tree, and k nearest neighbors.


Data preparation and standardization
Split data (test and train), Test of each model with combiantions of hyperparameters
Evaluation of accuracy of each model and comparison of results

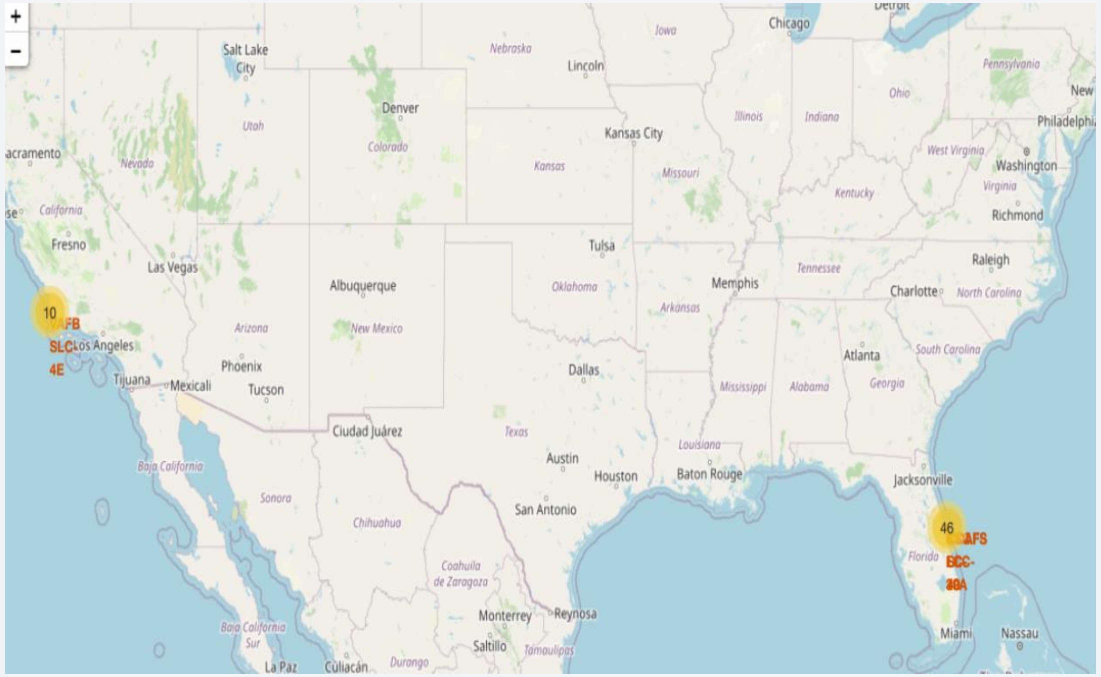
Source code : https://github.com/GOMEZBORIS6/IBM-Coursera-Applied-Capstone-Data-Science/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

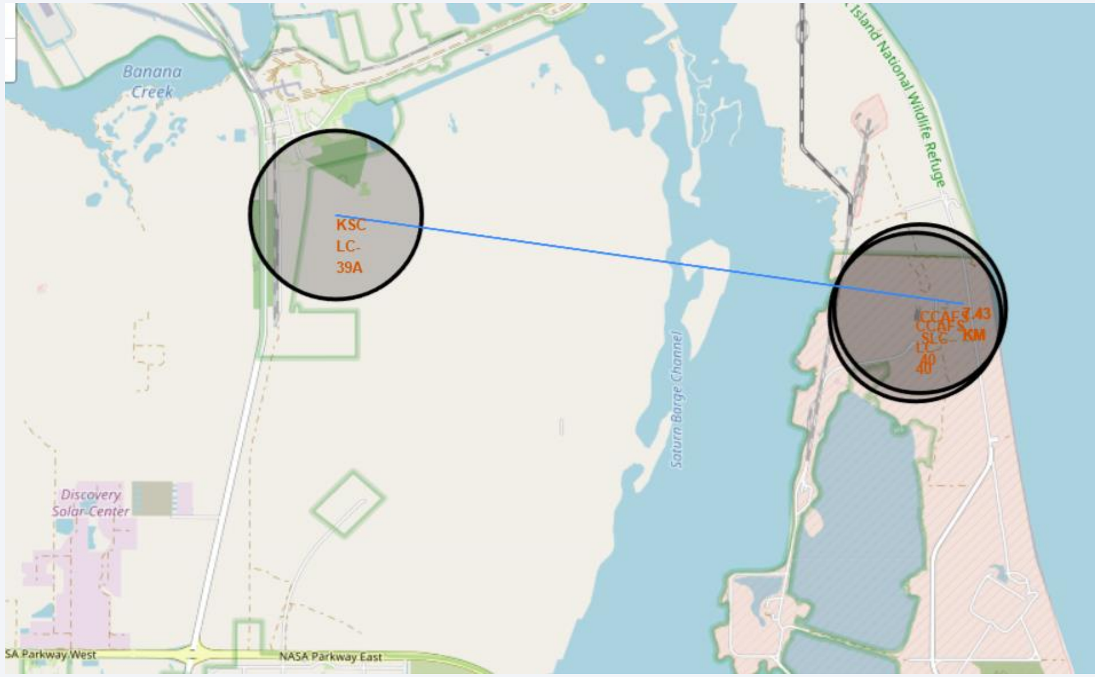
Results

- Exploratory data analysis results
  - Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

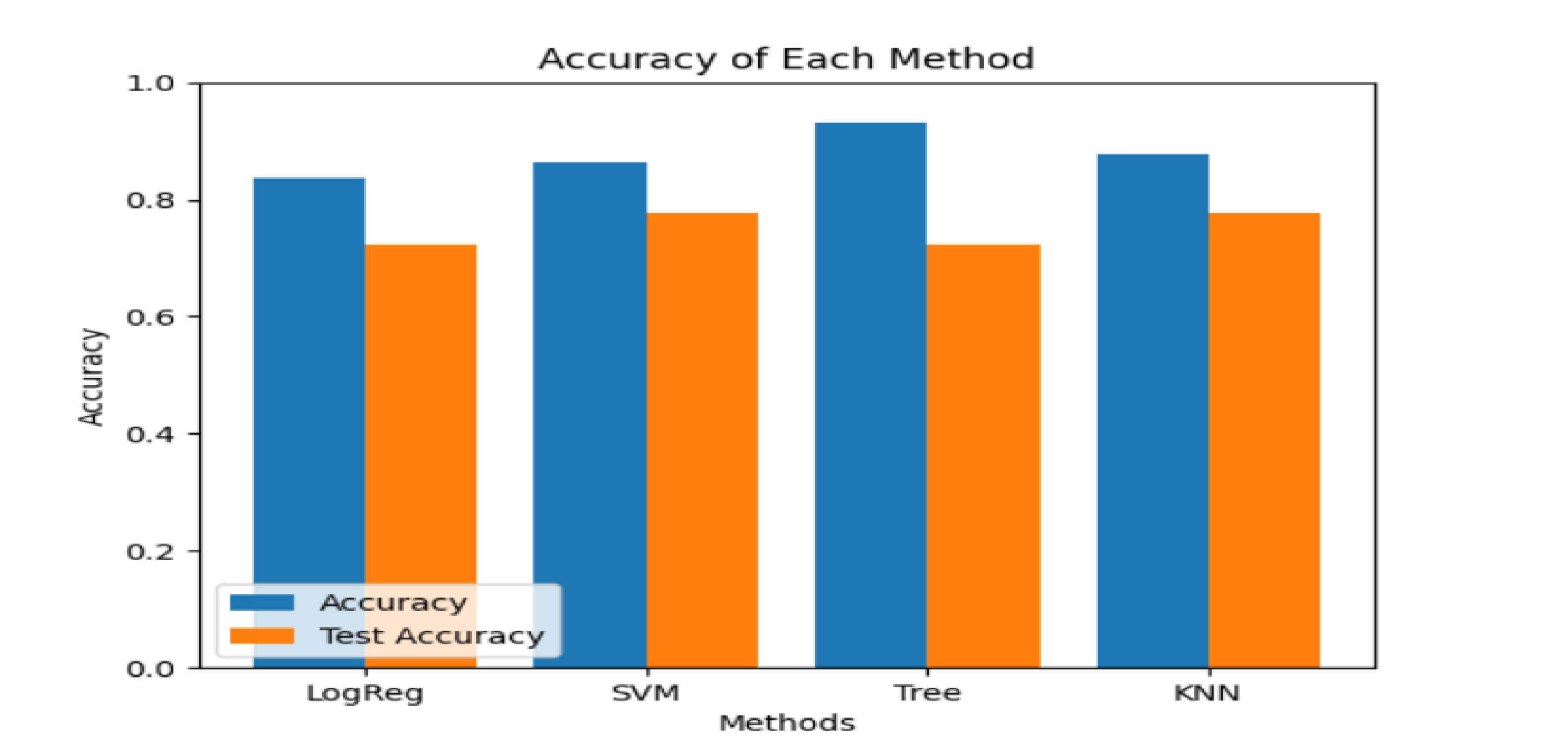
- Interactive analytics facilitated the identification of safety-oriented launch sites, typically located near the sea and equipped with reliable logistic infrastructure.
- Furthermore, the majority of launches are observed to occur at launch sites situated along the east coast.





Results

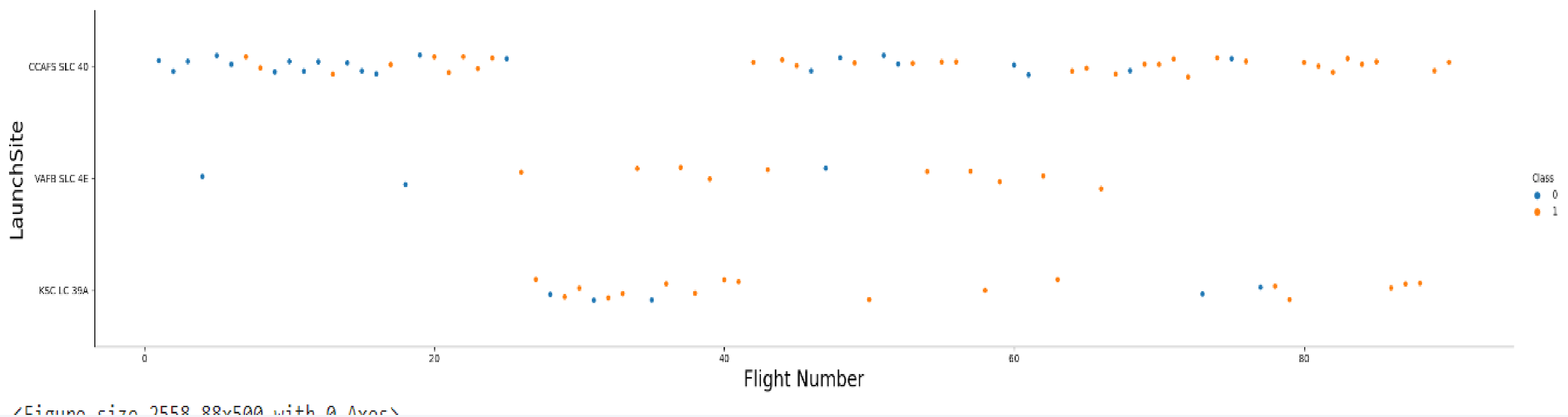
- The outcome of Predictive Analysis revealed that the Decision Tree Classifier model is the most effective in predicting successful landings, with an accuracy exceeding 96%, and a test data accuracy of over 77%.



Section 2

Insights drawn from EDA

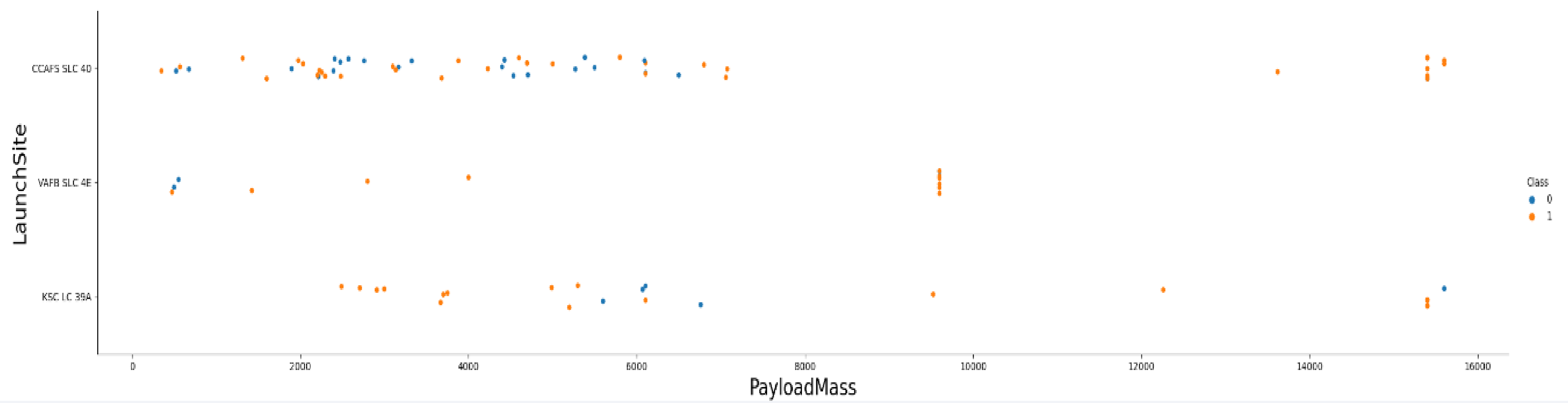
Flight Number vs. Launch Site



```
<Figure size 2558.88x500 with 0 Axes>
```

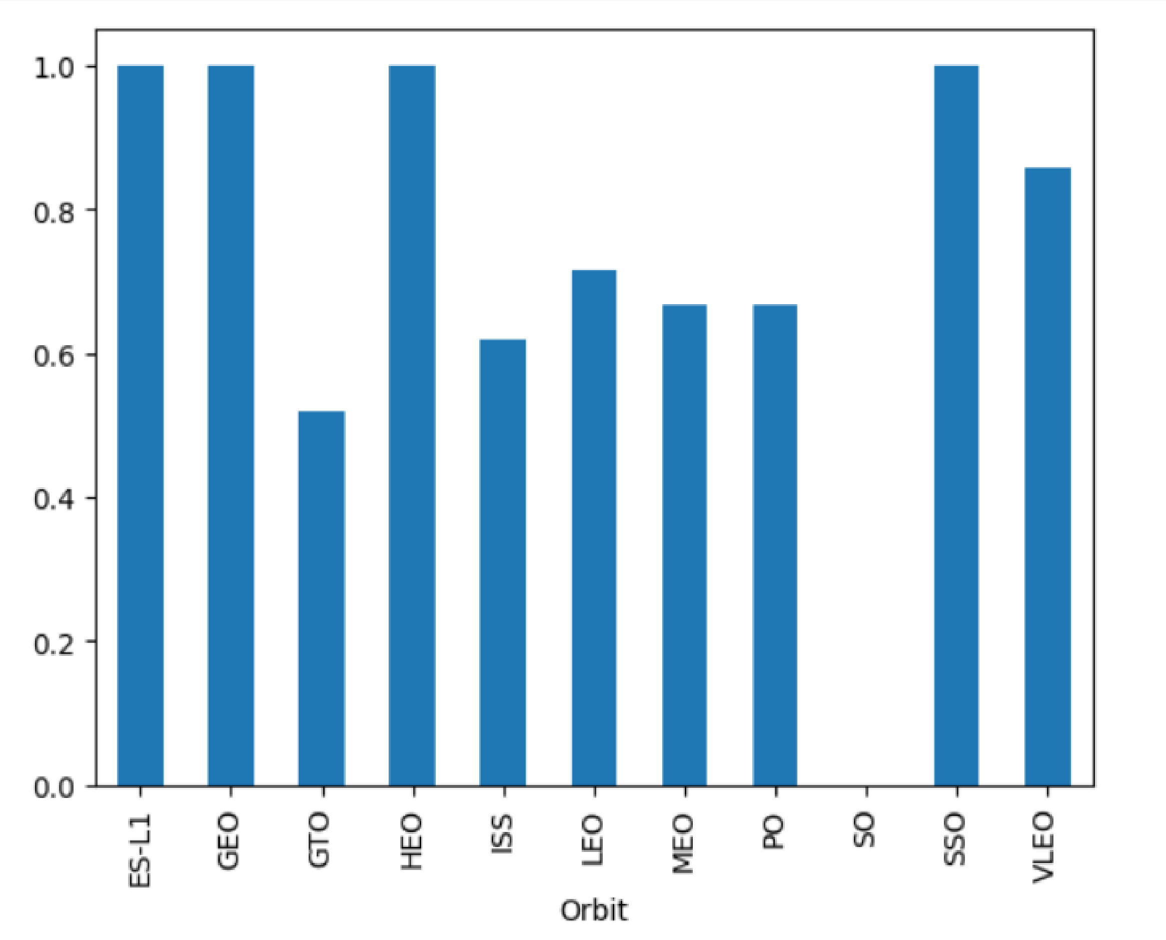
Based on the graph presented, it can be observed that the optimal launch site at present is CCAF5 SLC 40, which has experienced the highest success rate among recent launches. The second and third places are occupied by VAFB SLC 4E and KSC LC 39A, respectively. Additionally, it is evident that the overall success rate has improved over time.

Payload vs. Launch Site



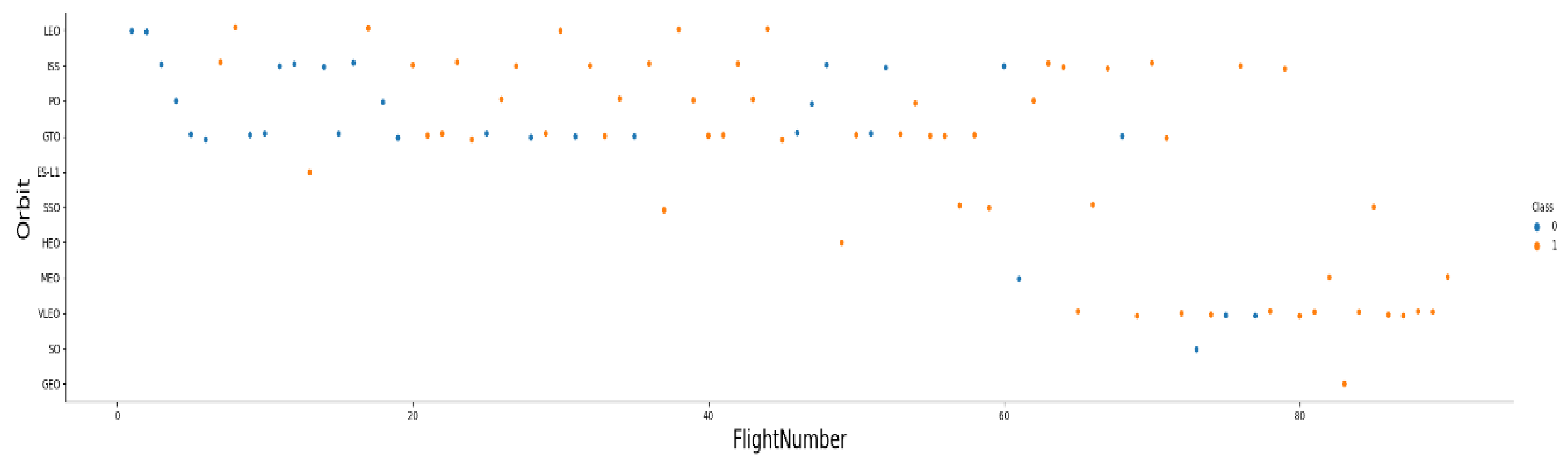
Payloads weighing over 9,000kg, approximately equivalent to the weight of a school bus, demonstrate a significantly high success rate. However, payloads weighing over 12,000kg can only be feasibly launched from CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



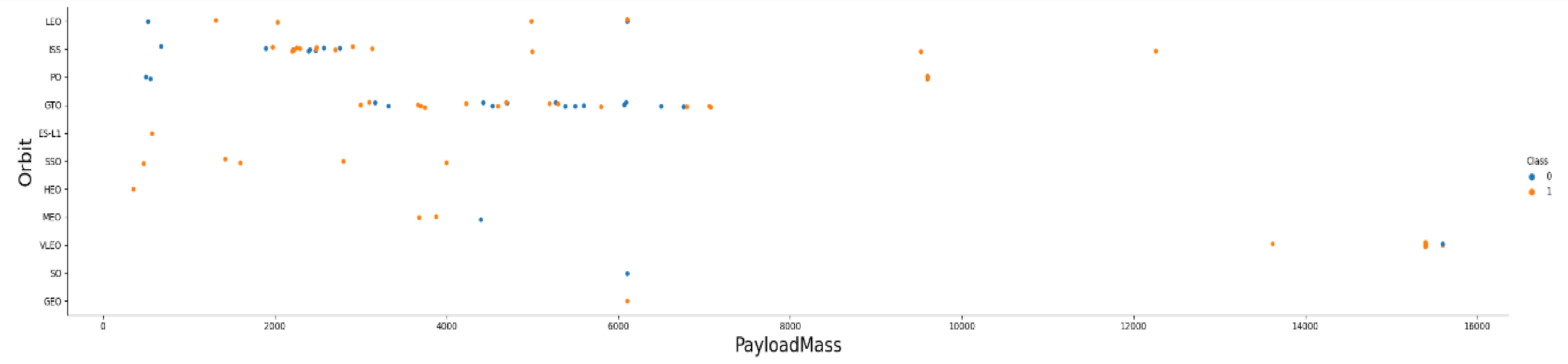
- The most successful orbits are:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Additionally, VLEO exhibits a success rate exceeding 80%, while LFO boasts a success rate above 70%.

Flight Number vs. Orbit Type



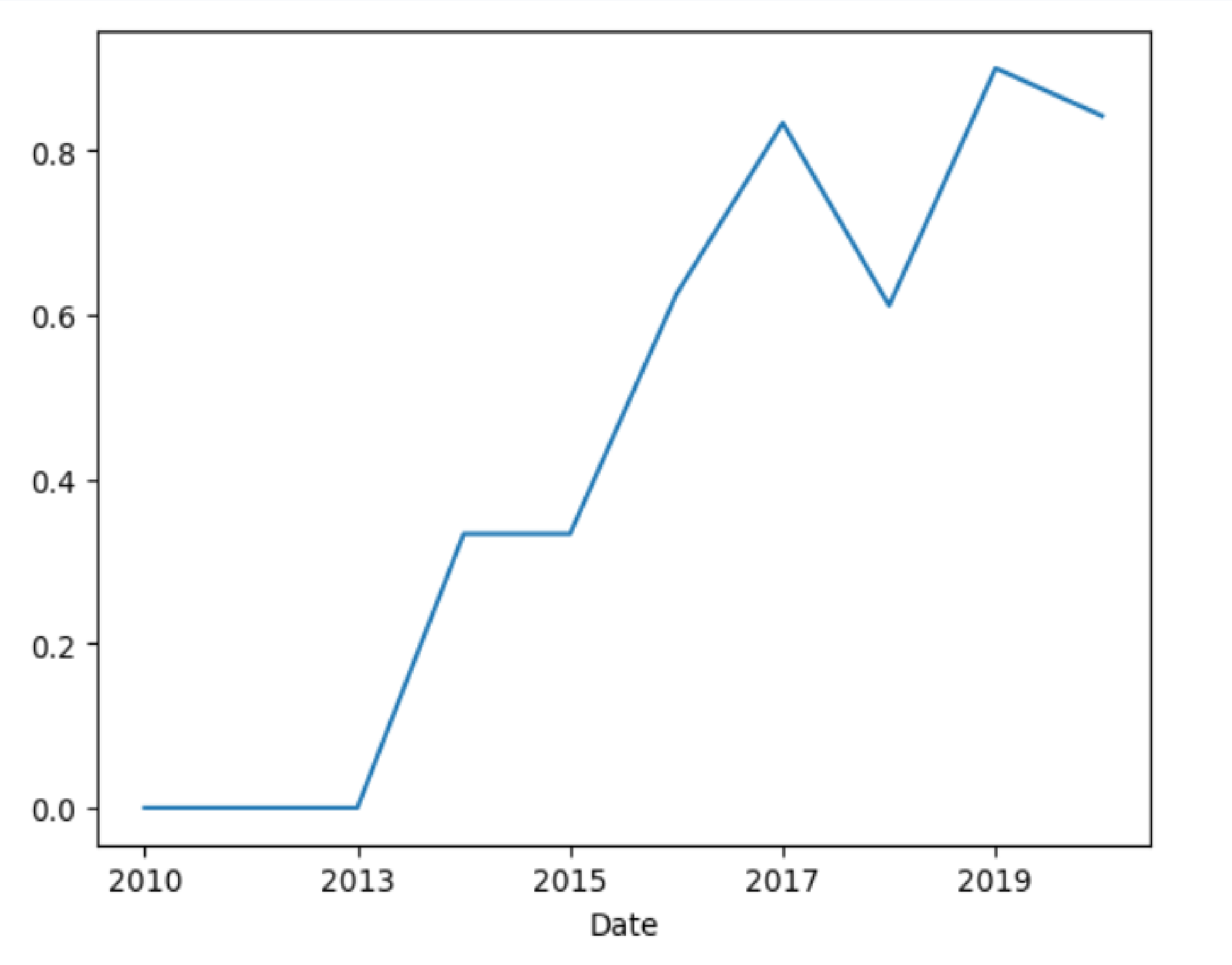
Evidently, the success rate has improved over time for all orbits. Furthermore, the recent increase in the frequency of VLEO orbit launches signifies a new business opportunity.

Payload vs. Orbit Type



We can see that there is no correlation between payload and success rate for the GTO orbit. The ISS orbit, on the other hand, boasts the widest payload range and an impressive success rate. Conversely, there have been relatively few launches to the SO and GEO orbits.

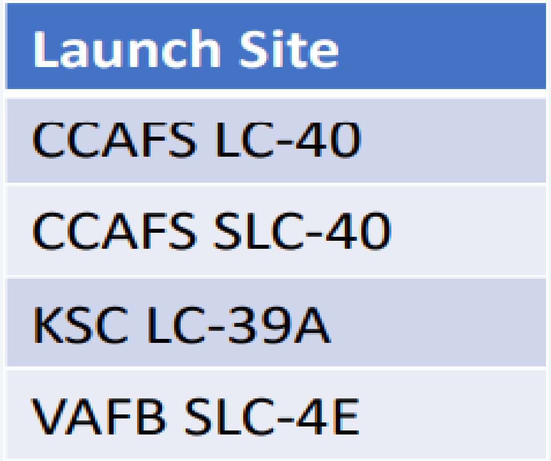
Launch Success Yearly Trend



Beginning in 2013, the success rate has been steadily increasing and has continued to do so until 2020. It is likely that the first three years of this period were focused on making necessary adjustments and refining technological advancements.

All Launch Site Names

- Based on the available data, there are four launch sites that have been identified.

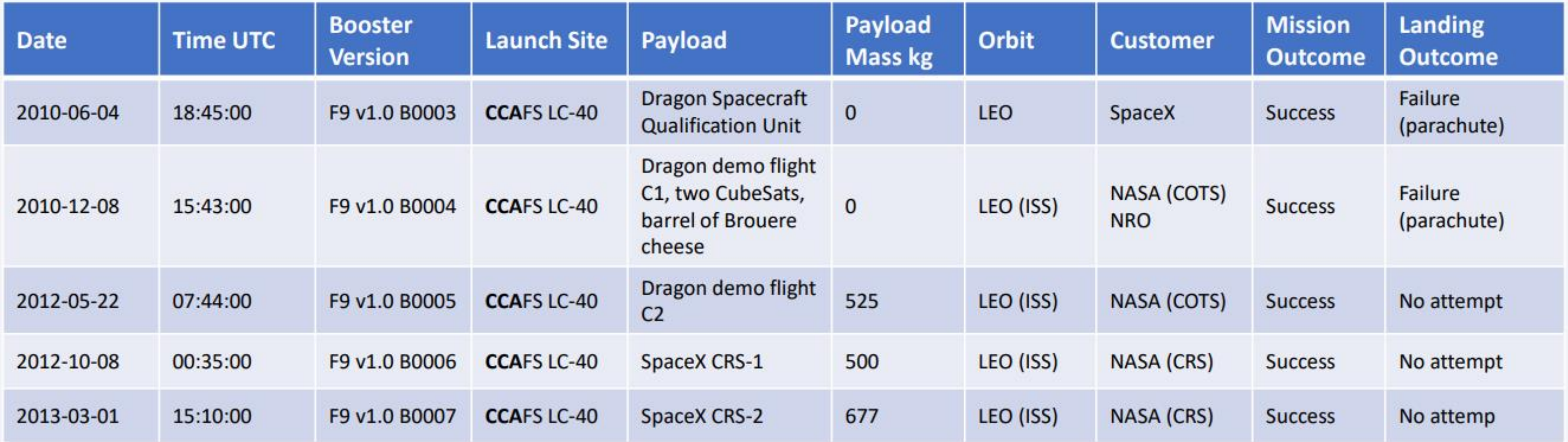


Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- These launch sites have been distinguished by selecting the distinct instances of "launch_site" values present within the dataset.

Launch Site Names Begin with 'CCA'

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

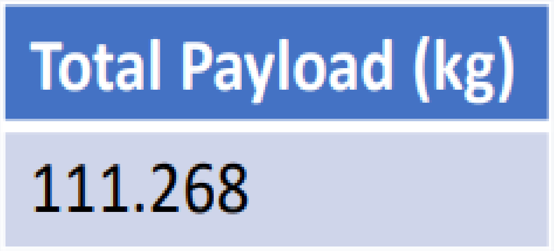


Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Presented here are five examples of launches that occurred at Cape Canaveral.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

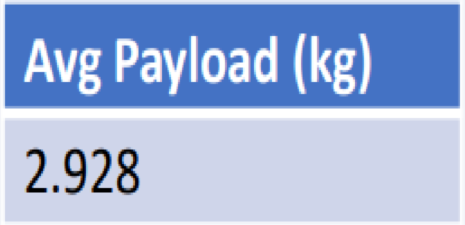


Total Payload (kg)
111.268

- The total payload displayed above is determined by summing the payloads associated with codes containing "CRS," which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

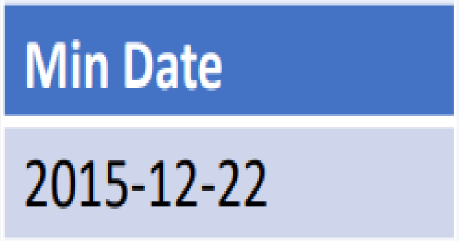


Avg Payload (kg)
2.928

- By selecting data based on booster version and computing the mean payload mass, we arrived at a value of 2,928 kg.

First Successful Ground Landing Date

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

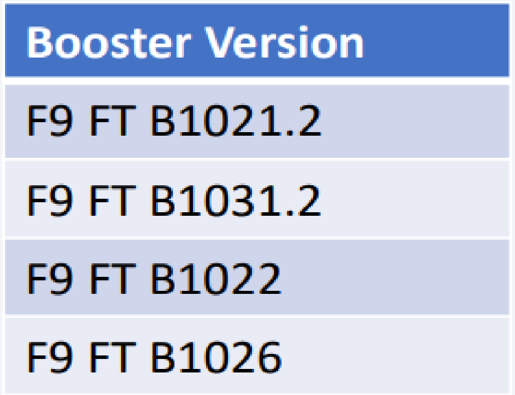


Min Date
2015-12-22

- By applying a filter to select only data with successful landing outcomes on a ground pad and finding the earliest date in the dataset, we can identify the first occurrence which happened on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

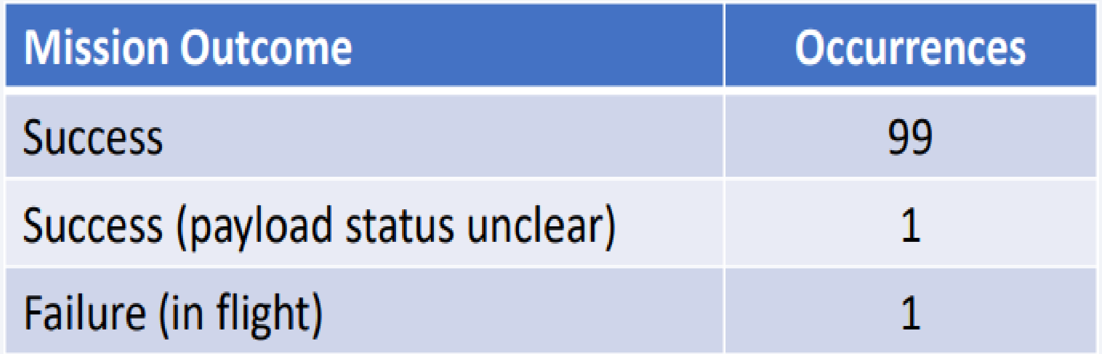


Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- After applying the aforementioned filters to select unique booster versions, the resulting set comprises of the following four versions.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

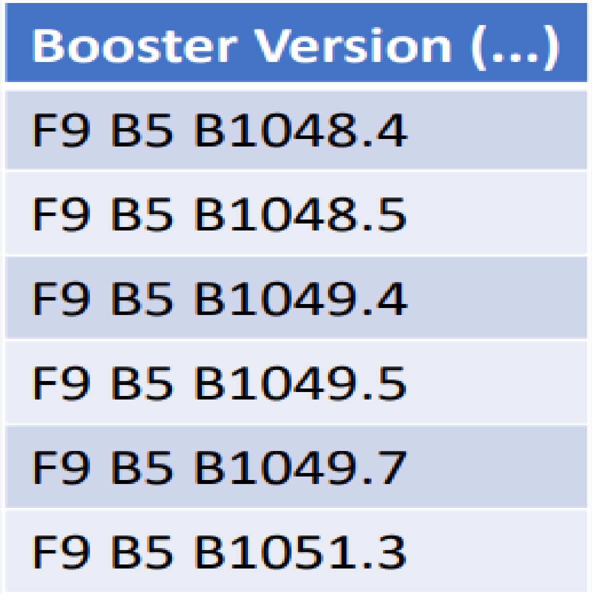


Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

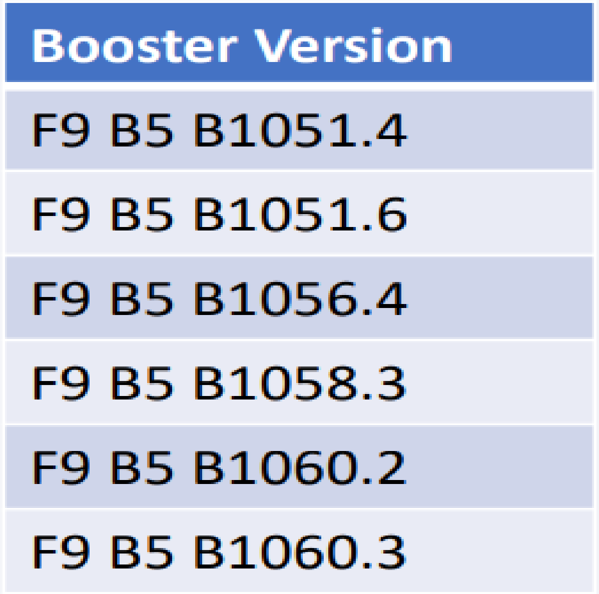
- By grouping mission outcomes and tallying the number of records in each group, we obtained the summary presented above.

Boosters Carried Maximum Payload

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



Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

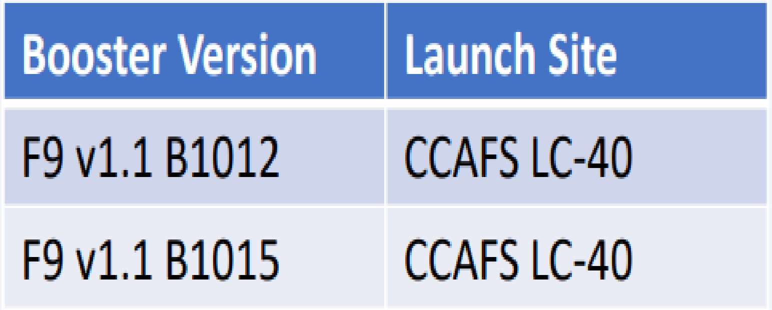


Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The following boosters have transported the highest recorded payload masses in the dataset.

2015 Launch Records

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

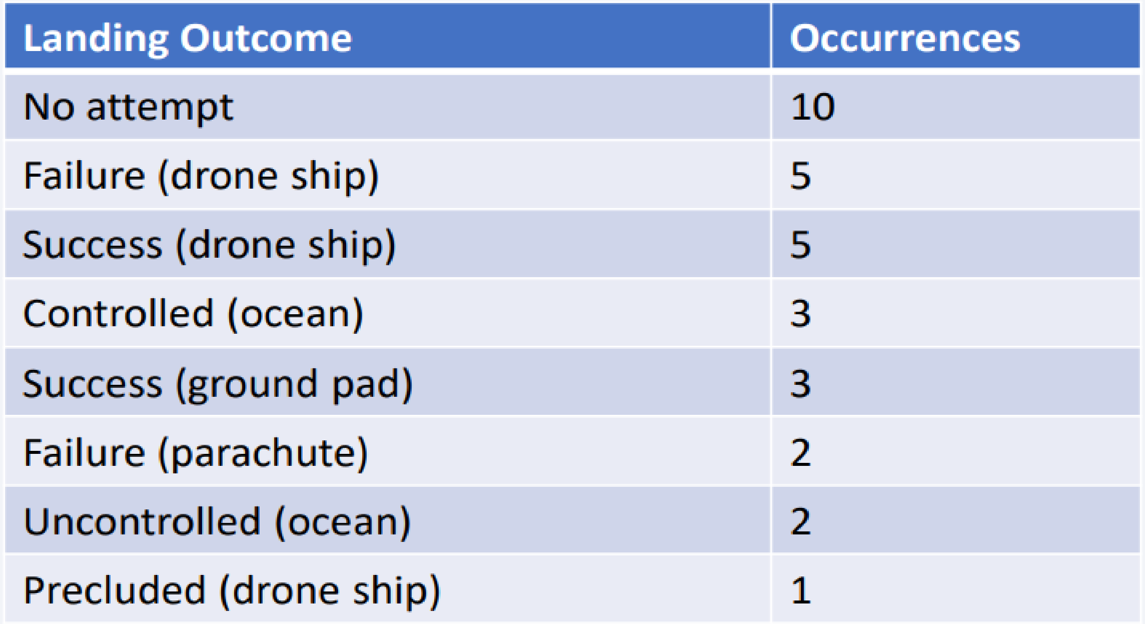


Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The aforementioned list contains only two instances.

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



Landing Outcome	Occurrences
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

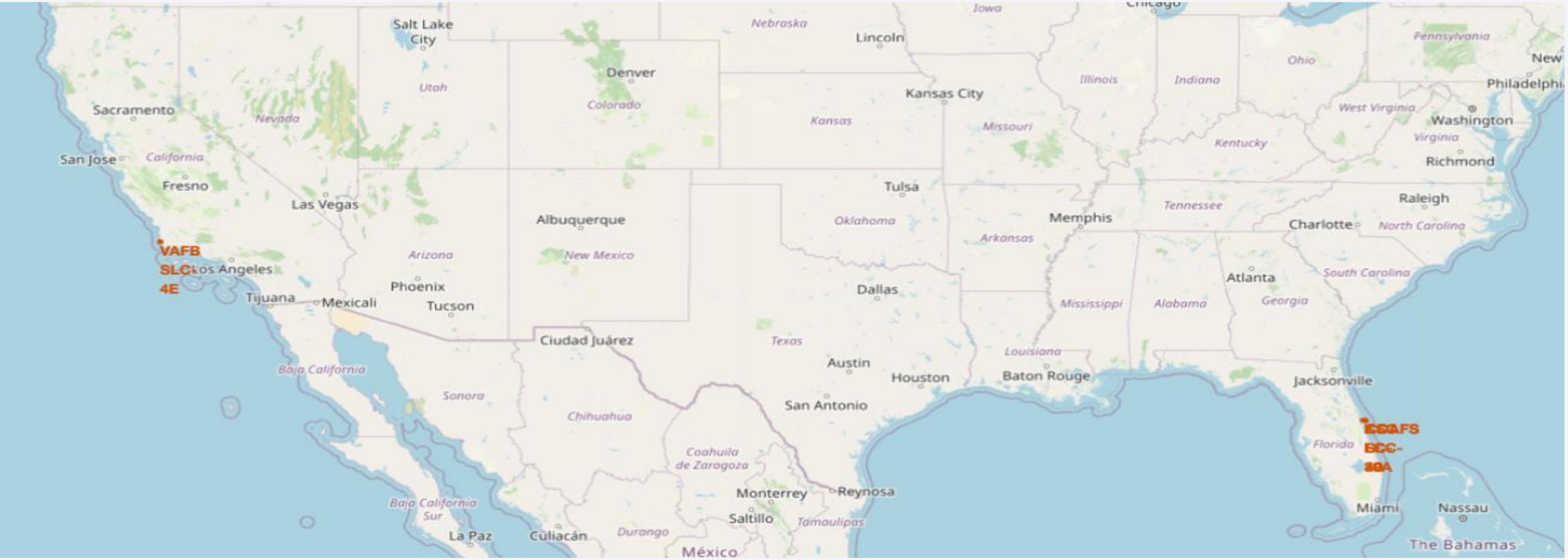
- This perspective of the data emphasizes the importance of accounting for the instances where "No attempt" was made.

Section 3

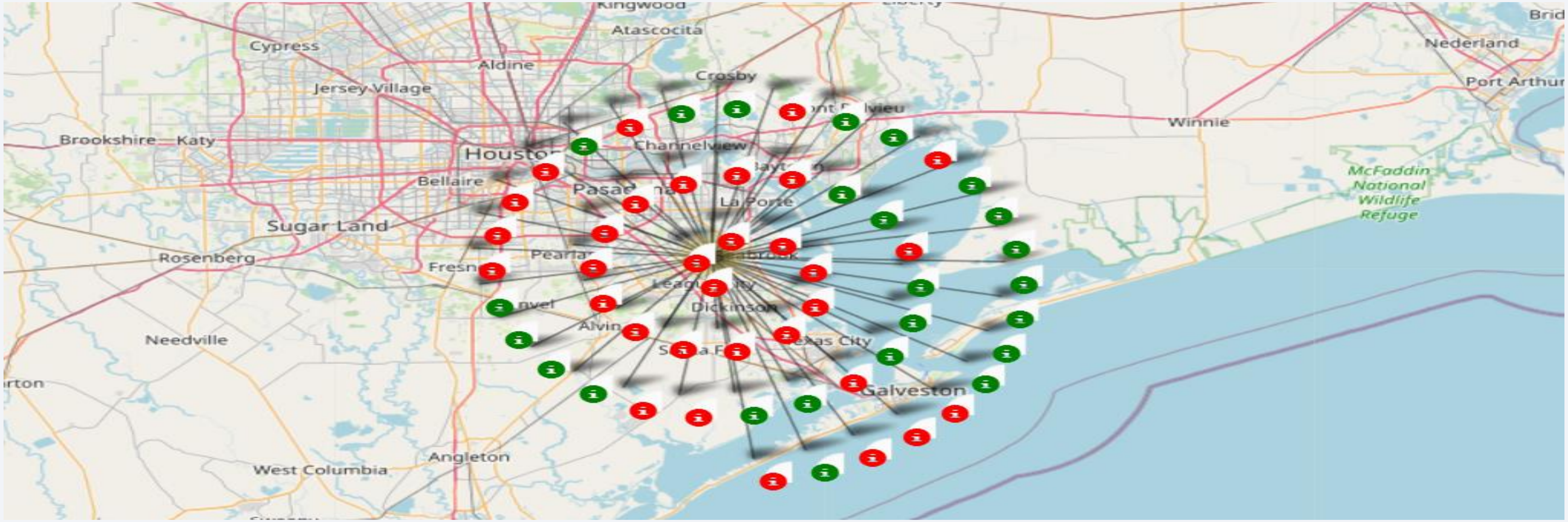
Launch Sites Proximities Analysis

All launch sites

The launch sites are located in proximity to the sea, presumably for safety reasons, yet they are not too far away from roads and railroads.

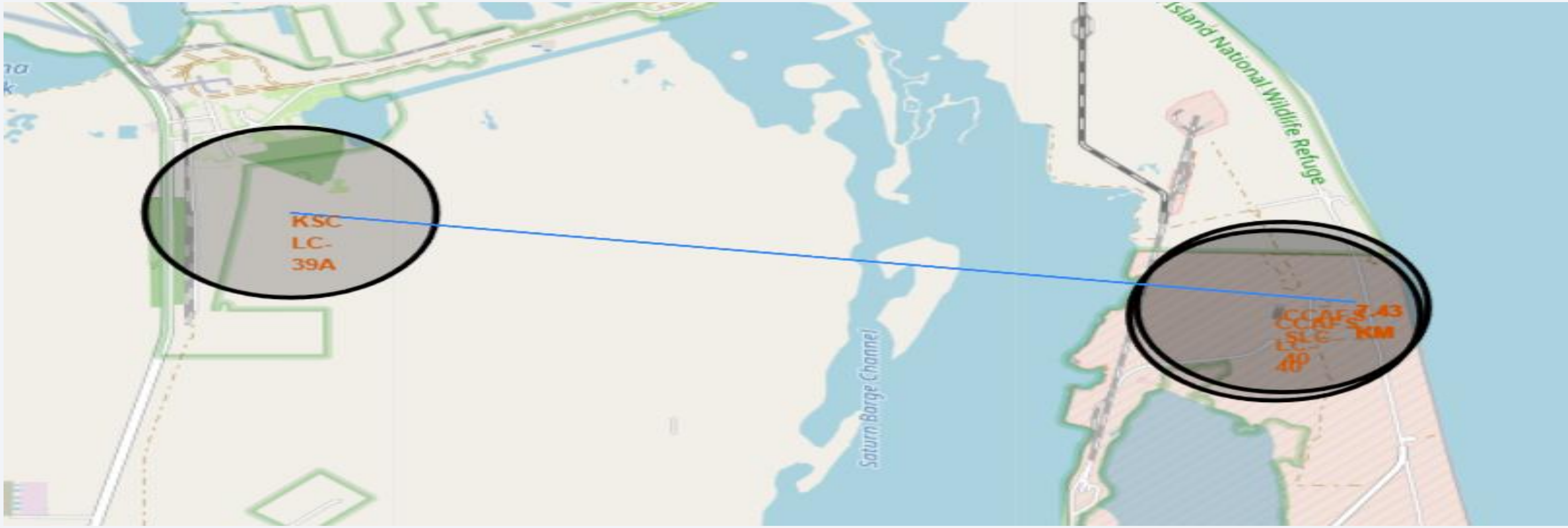


Launch Outcomes by Site

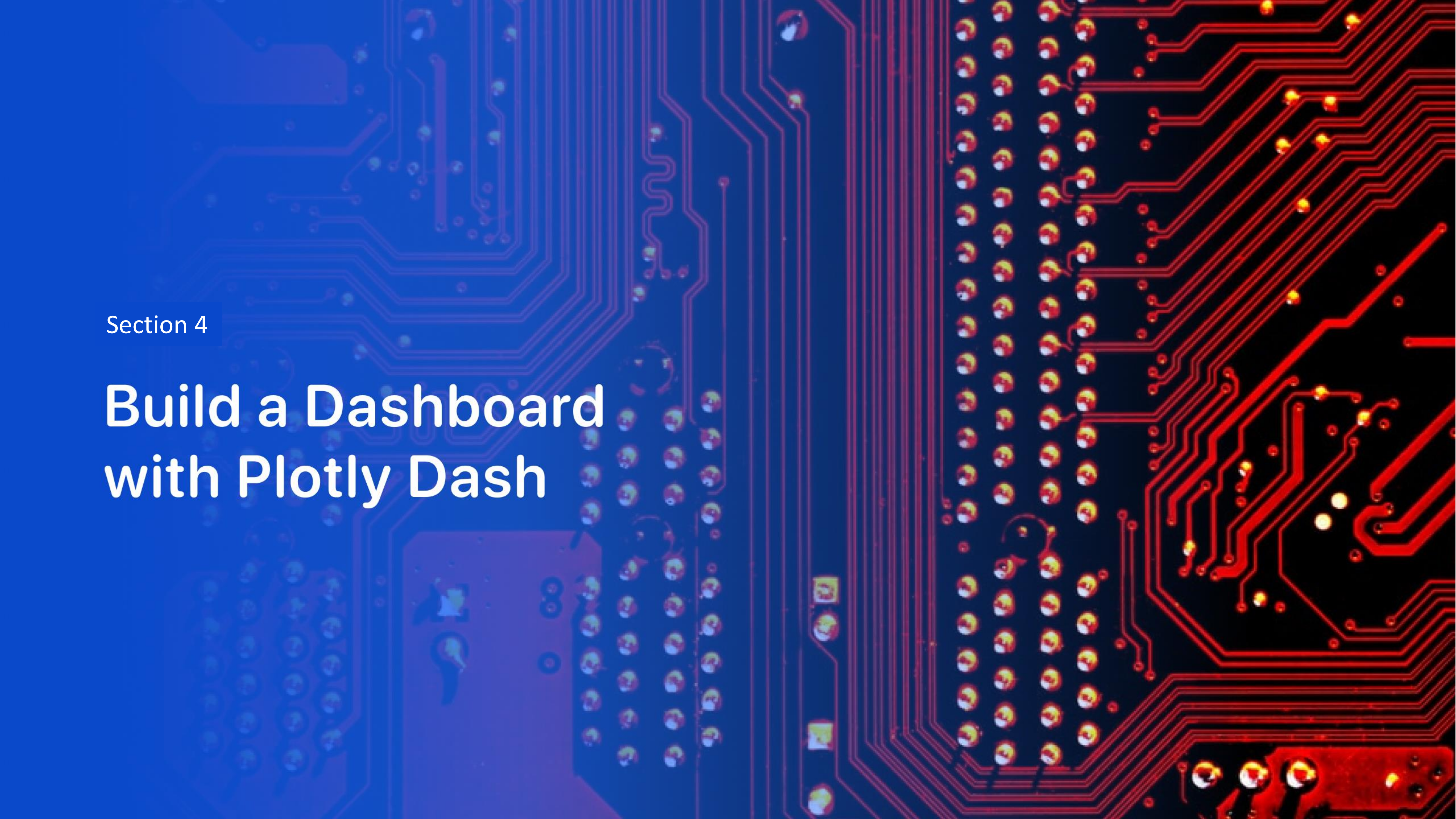


Successful attempts are denoted by green markers, while red markers represent failed attempts.

Logistics and Safety



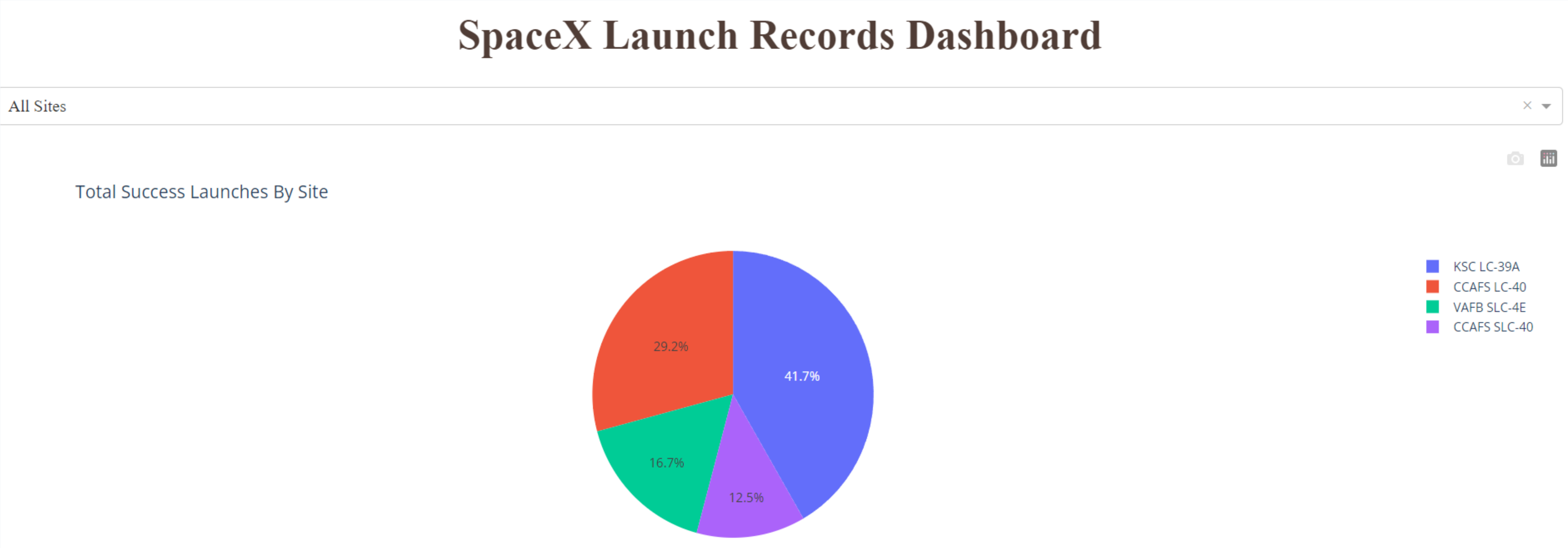
The KSC LC-39A launch site boasts favorable logistical attributes, given its close proximity to both rail and road networks, and its relative distance from residential areas.



Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

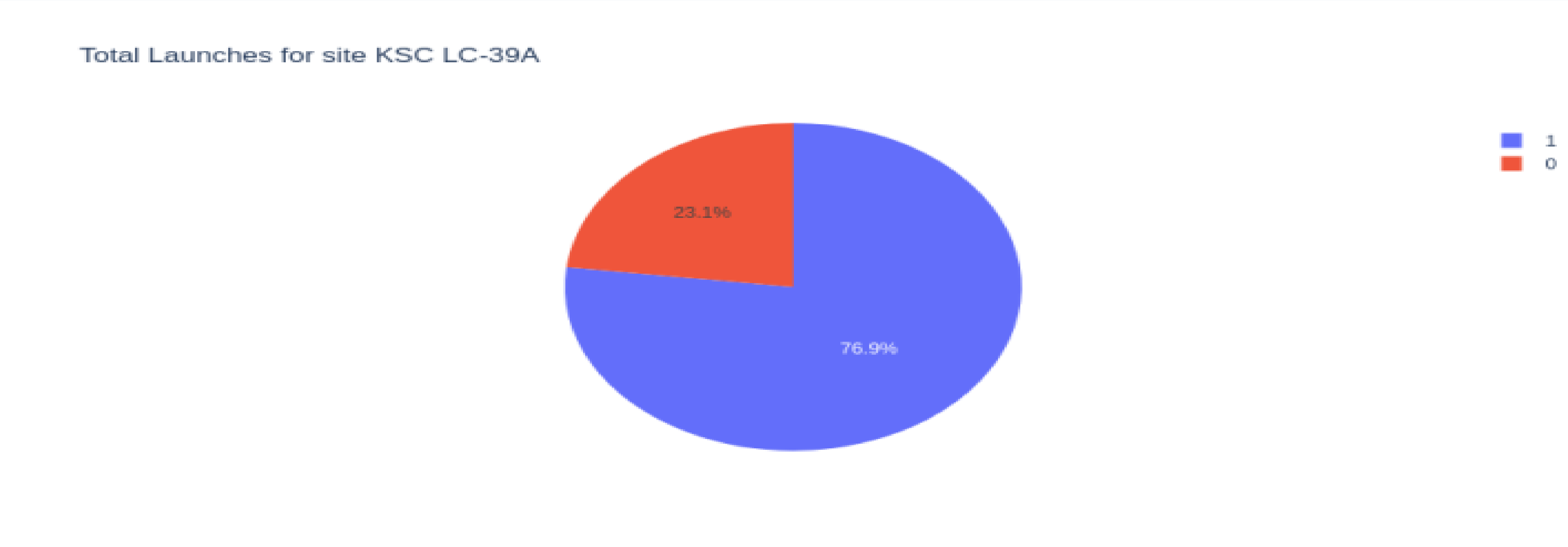


SpaceX Launch Records Dashboard

All Sites

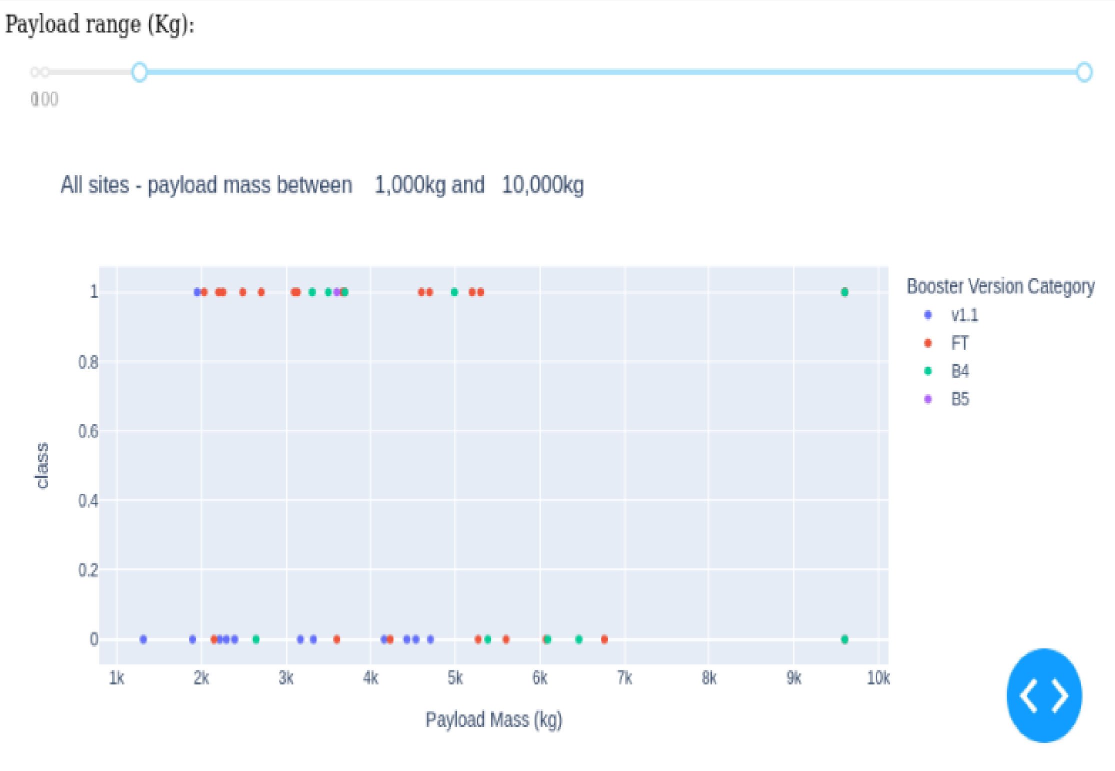
The location from where launches are conducted appears to be a crucial determinant of mission success.

Launch Success Ratio for KSC LC-39A

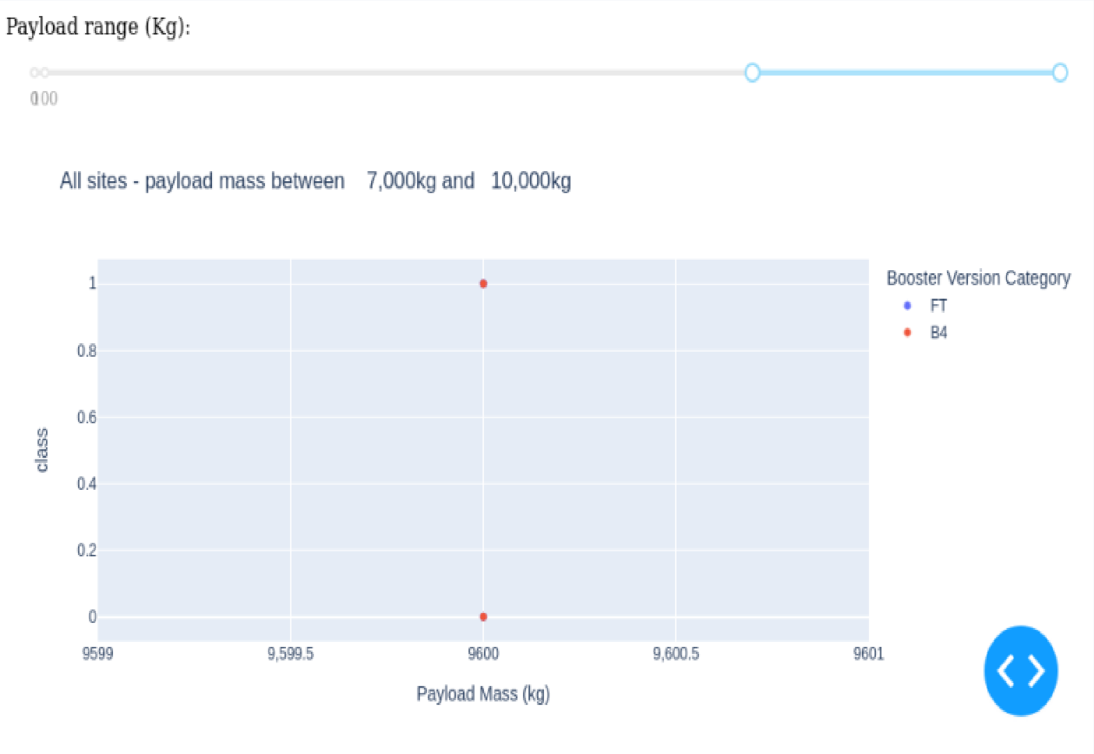


The success rate of launches at this site is 76.9%.

Payload vs. Launch Outcome


Payload range (Kg):

All sites - payload mass between 1,000kg and 10,000kg


Payload range (Kg):

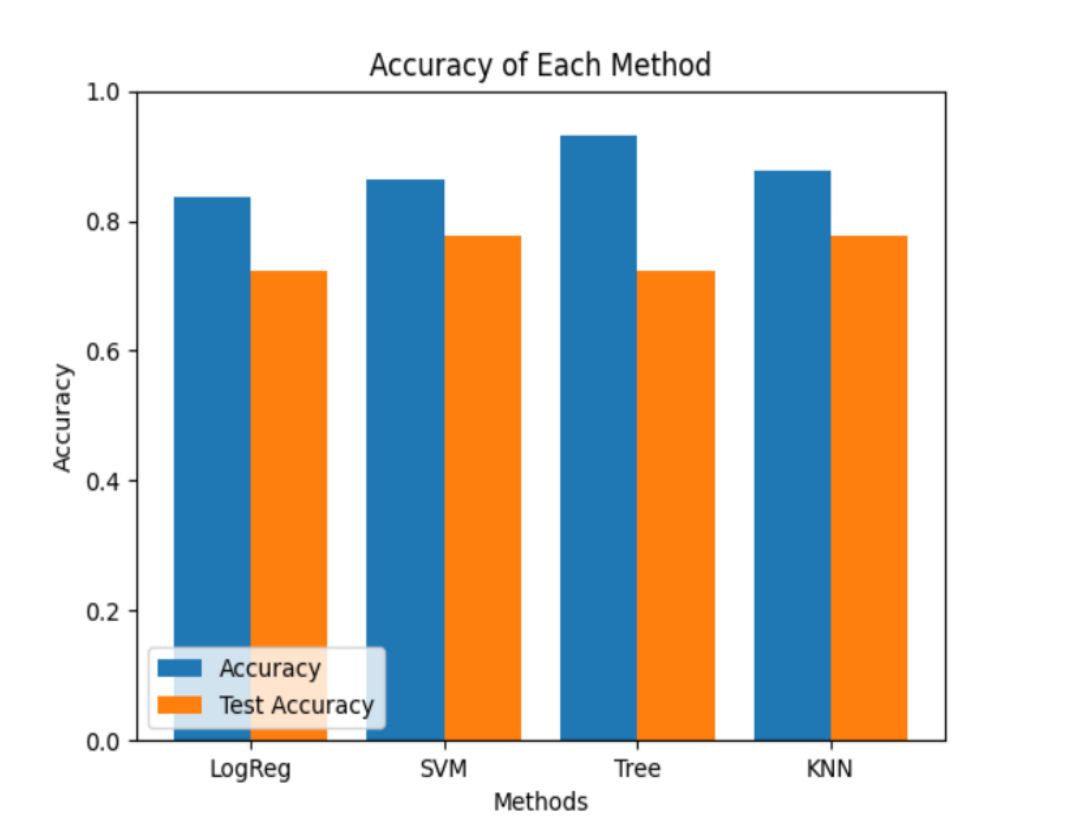
All sites - payload mass between 7,000kg and 10,000kg

The combination of payloads weighing less than 6,000 kg and FT boosters is the most successful. The available data is insufficient to accurately assess the risk associated with launches of payloads exceeding 7,000 kg.

Section 5

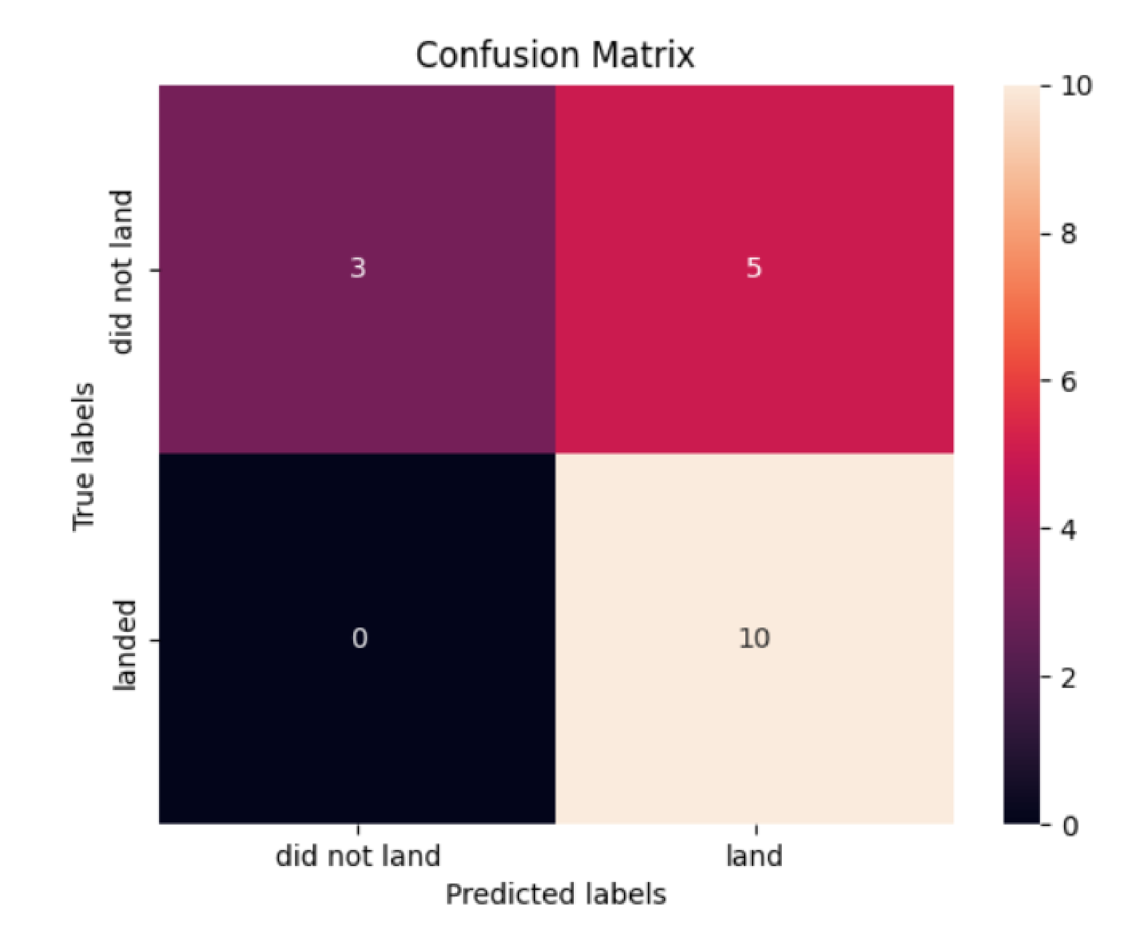
Predictive Analysis (Classification)

Classification Accuracy



Four classification models were evaluated and their corresponding accuracies are presented in the accompanying plot. The Decision Tree Classifier exhibited the highest classification accuracy, surpassing 96%.

Confusion Matrix of Decision Tree Classifier



The accuracy of a Decision Tree Classifier can be demonstrated by examining the Confusion Matrix, which reveals a high number of false negatives and true negatives relative to True positives and false positives.

Conclusions

Throughout the analysis process, various data sources were examined, leading to refined conclusions. The data suggests that the optimal launch site is KSC LC-39A and that launches of payloads exceeding 7,000kg are less hazardous. While most missions are successful, successful landings appear to be improving over time due to advancements in processes and rockets. To enhance profits, a Decision Tree Classifier could be utilized to predict successful landings.

Appendix

To enhance the accuracy of model testing, it is crucial to establish a value for the "`np.random.seed`" variable.

Additionally, due to Folium's inability to display maps on Github, screenshots were captured instead.

Thank you!