IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ruan
23 October 2024

Outline

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

Executive Summary

Executive Summary

1. Overview: The notebooks analyze SpaceX's Falcon 9 rocket launches, focusing on predicting the success of first-stage landings and highlighting cost-effectiveness compared to competitors.
2. Data Preparation: Datasets are loaded into Pandas DataFrames, with data wrangling techniques applied to clean and prepare the data for analysis.
3. Exploratory Data Analysis (EDA): Various EDA techniques are used to visualize launch success rates, payload distributions, and correlations between payload mass and launch outcomes.
4. Machine Learning: Machine learning models are developed to predict landing success based on features like payload mass and launch site, with interactive Dash applications for visualization.

Executive Summary

5. SQL Integration: SQL queries extract insights from SQLite databases, including unique launch sites and total payload mass.
6. Visualization: Visualizations using Plotly and Folium present data effectively, emphasizing the importance of visual representation in analysis.
7. Conclusion: The notebooks provide insights into factors influencing landing success and overall booster performance, aiming to inform future SpaceX missions.

Objective

- The goal is to leverage data science techniques to analyze SpaceX launch data, predict landing success, and derive actionable insights for future strategies.

Introduction

- This project was done for the space exploration company SpaceY
- The aim of this project is to look at the effectivity of competitor SpaceX and to learn from their mistakes
- This includes looking at:
    - Booster types
    - Launch sites
    - Orbit types
    - Payload mass

Section 1

Methodology

Methodology

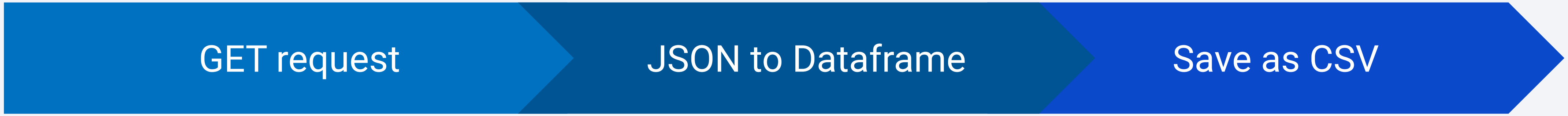
Executive Summary

- Data collection methodology:
  - SpaceX API calls
  - Wikipedia page was web scraped
- Perform data wrangling
  - Missing values were replaced and some variables were modified
- 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 tuned with grid search

Data Collection

- Data sets were collected in 2 ways:

  1. Using the SpaceX API



GET request

JSON to Dataframe

Save as CSV

Use a get request to retrieve the data in JSON format from the SpaceX API.

Convert the data from JSON format to a Pandas dataframe. Do this by processing the JSON.

Save the dataframe as a csv file to make it easily available for future use.

Data Collection – SpaceX API

01

GET
"https://api.spacexdata.com/v4/launches/past"

- Gets all data for past launches
- Most of data contains only ID values

02

GET
"https://api.spacexdata.com/v4/rockets/{ID}"

- Retrieve all the rocket names from using the ID's

03

GET
"https://api.spacexdata.com/v4/launchpads/{ID}"

- Retrieve all the data about the launchpads from their ID's

04

GET
"https://api.spacexdata.com/v4/payloads/{ID}"

- Retrieve all the data about the payloads from their ID's

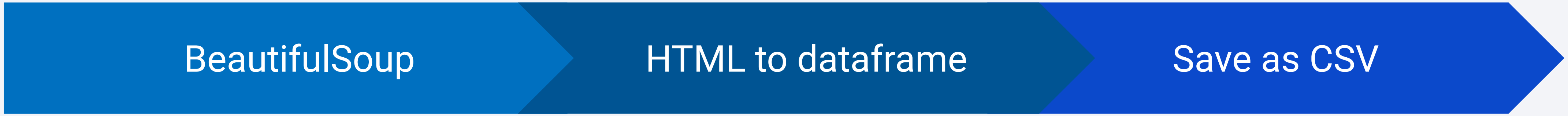
05

GET
"https://api.spacexdata.com/v4/cores/{ID}"

- Retrieve all the data about the cores from their ID's
- This includes outcome of the retrieval of the boosters

Data Collection

2. Scraping the data from the SpaceX wikipedia page



BeautifulSoup

HTML to dataframe

Save as CSV

Use beautifulsoup to extract the data from the table, in HTML format.

Convert the data from HTML format to a Pandas dataframe. Do this by processing each line of the HTML table.

Save the dataframe as a csv file to make it easily available for future use.

Data Wrangling

- All null values for Payload Mass were replaced with mean of payload mass

- The outcomes of the launches were split into two types
  - Good outcomes (booster successfully landed)
  - Bad outcomes (booster not landed successfully

EDA with Data Visualization

- Charts were plotted that explored all the possible relationships between:
  - Launch Site
  - Payload mass
  - Flight Number
  - Orbit type
  - Outcome of launch

- The Success rate of launches over the years were also plotted.
- Dummy variables for each of the categorical attributes were created for future analysis

EDA with SQL

- SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
- SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
- SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Landing_Outcome;

EDA with SQL

- SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

- 9. SELECT substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015';

- SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC;

Build an Interactive Map with Folium

- The map was created with Folium with various objects added to the map
    - Circles were added to each of the launch sites to indicate their location
    - Markers were added at the location of each launch
        - The markers were in clusters to improve readability
    - Lines were drawn from launch sites to nearby points of interest to indicate the distance between them

Build a Dashboard with Plotly Dash

- Pie Chart
  - Displays the total successful launches by site.
  - If a specific launch site is selected, it shows the success vs. failed counts for that site.
- Scatter Chart
  - Illustrates the correlation between payload mass (in kg) and launch success.
  - The data points are colored based on the booster version category.
- These visualizations help users analyze the performance of SpaceX launches based on different criteria, such as launch site and payload mass.

Predictive Analysis (Classification)

1. Building the Model

    - Data Preparation:
        - Loaded the dataset and created a target variable `Y` from the `Class` column.
        - Standardized the feature set `X` using `StandardScaler`.
        - Split the data into training and testing sets using `train_test_split`.

    - Model Selection:
        - Implemented multiple classification algorithms, including:
        - Logistic Regression
        - Support Vector Machine (SVM)
        - Decision Tree
        - K-Nearest Neighbors (KNN)

Predictive Analysis (Classification)

2. Evaluating the Model

- Cross-Validation:
    - Used `GridSearchCV` to perform hyperparameter tuning for each model with cross-validation (cv=10).
    - Evaluated models based on their accuracy on the validation set.

- Performance Metrics:
    - Printed the best parameters and accuracy for each model.
    - Used confusion matrices to visualize the performance of the models on the test data.

Predictive Analysis (Classification)

3. Improving the Model

- Hyperparameter Tuning:
  - For each model, a set of hyperparameters was defined and optimized using `GridSearchCV`.
  - This process involved testing various combinations of parameters to find the optimal settings for each algorithm.

- Model Comparison:
  - Compared the accuracy and confusion matrices of all models to identify strengths and weaknesses.

Predictive Analysis (Classification)

4. Finding the Best Performing Model:

- Final Evaluation:
  - After tuning, the accuracy of each model was calculated on the test set.
  - The model with the highest accuracy and the best performance metrics (e.g., lowest false positives) was selected as the best-performing model.
- Results Presentation:
  - The best model's parameters and accuracy were printed, and its confusion matrix was plotted for a clear understanding of its performance.
- This systematic approach ensured that the best classification model was identified based on rigorous evaluation and optimization processes.

Results

- Exploratory data analysis results

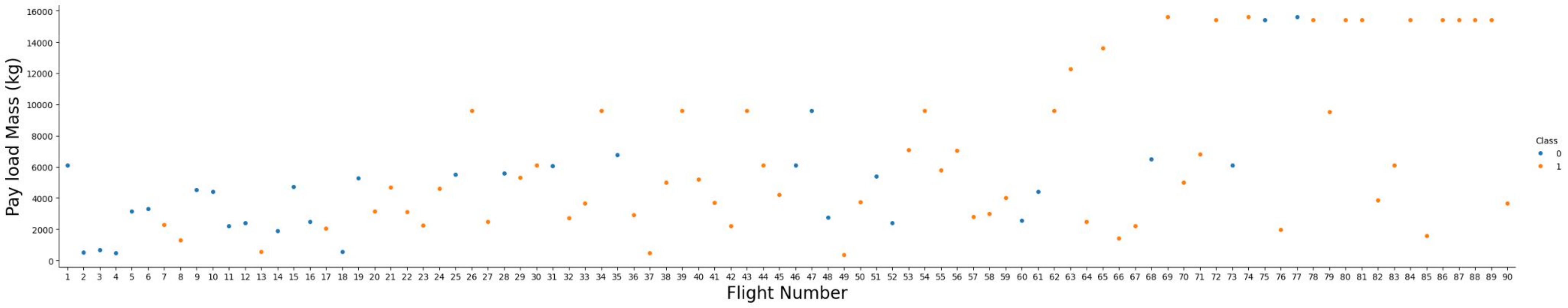
- Interactive analytics demo in screenshots

- Predictive analysis results

Section 2

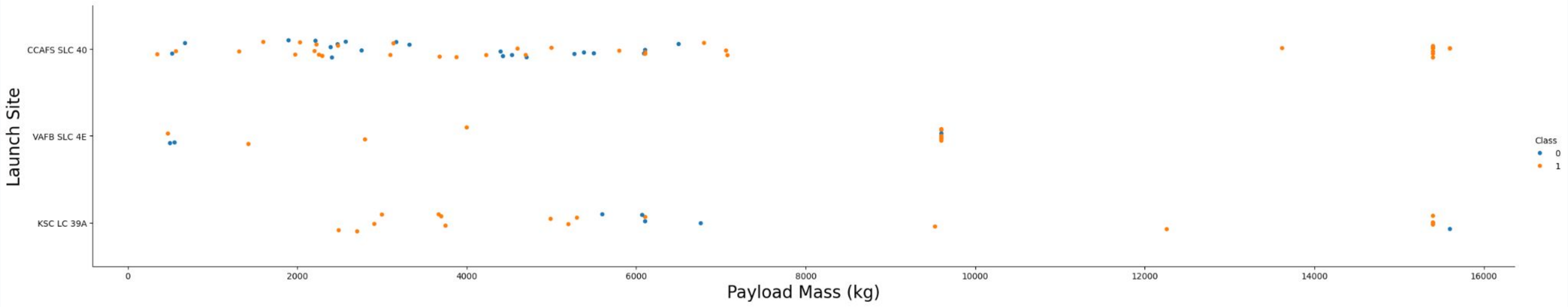
Insights drawn from EDA

Flight Number vs. Launch Site



This scatter plot shows the relationship between the number of flights (Flight Number) and the mass of the payload (PayloadMass). The hue represents the launch outcome (Class), indicating whether the landing was successful or not. The trend suggests that as the flight number increases, the likelihood of a successful landing also increases. Additionally, it appears that heavier payloads are associated with a lower success rate for landings.

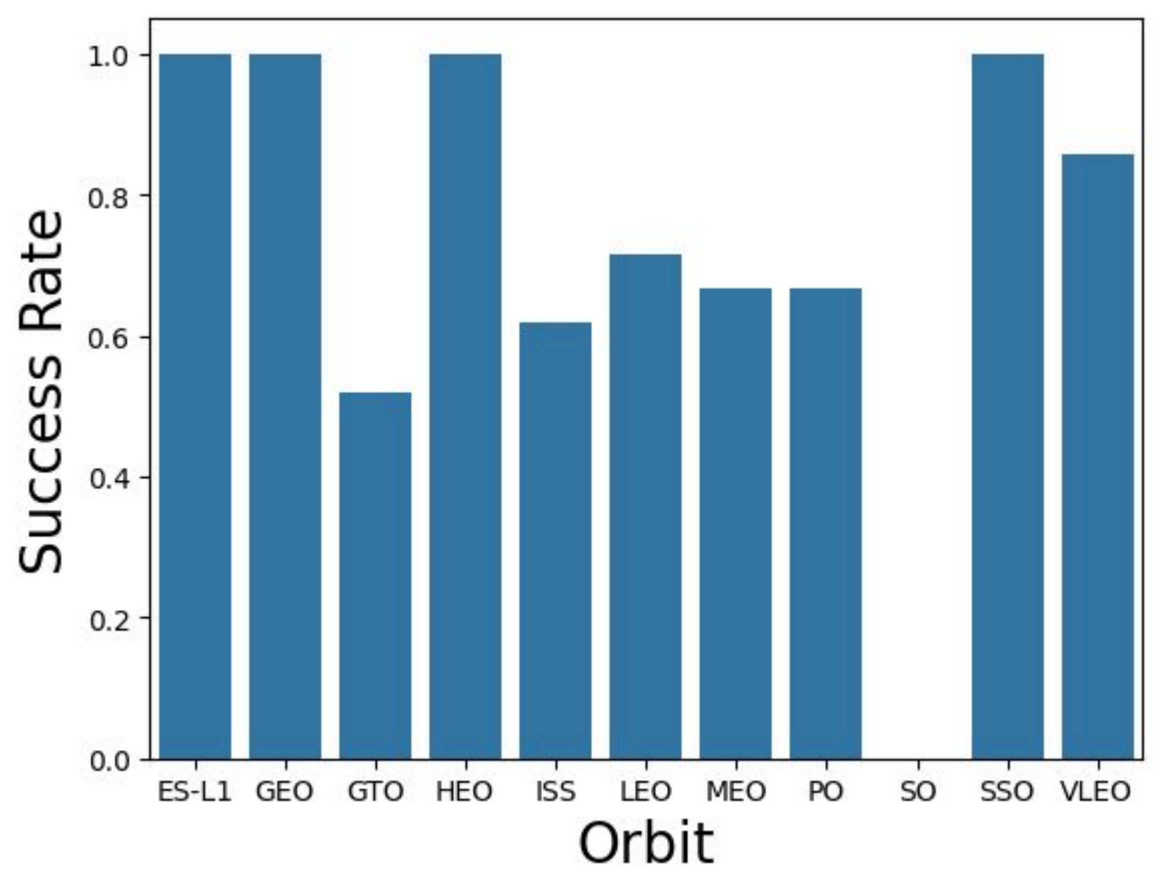
Payload vs. Launch Site



This scatter plot examines the relationship between Payload Mass and Launch Site. It highlights how different launch sites handle varying payload sizes and whether there are any trends in success rates based on the payload mass.

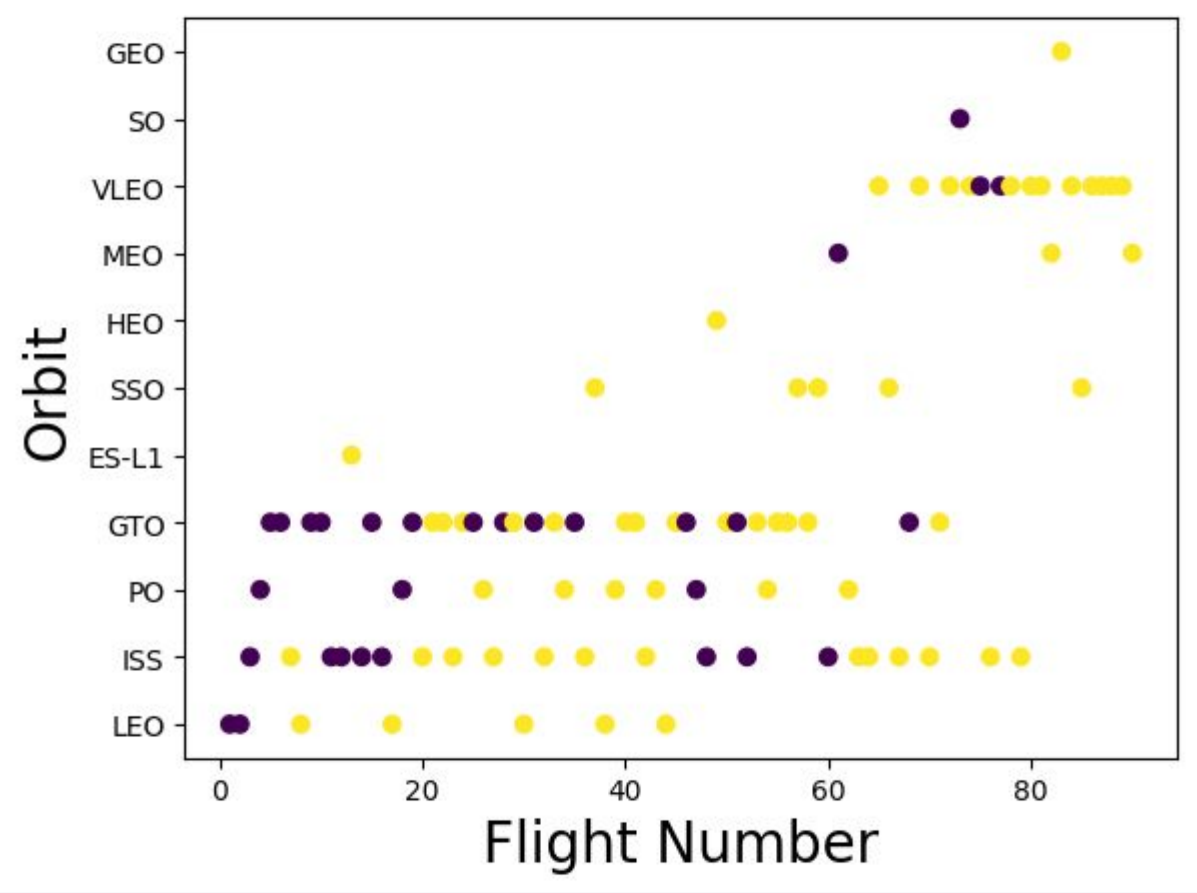
Success Rate vs. Orbit Type

The bar chart displays the average success rate for each orbit type. This visualization allows for a quick comparison of which orbits have higher success rates, indicating that certain orbits may be more favorable for successful landings.



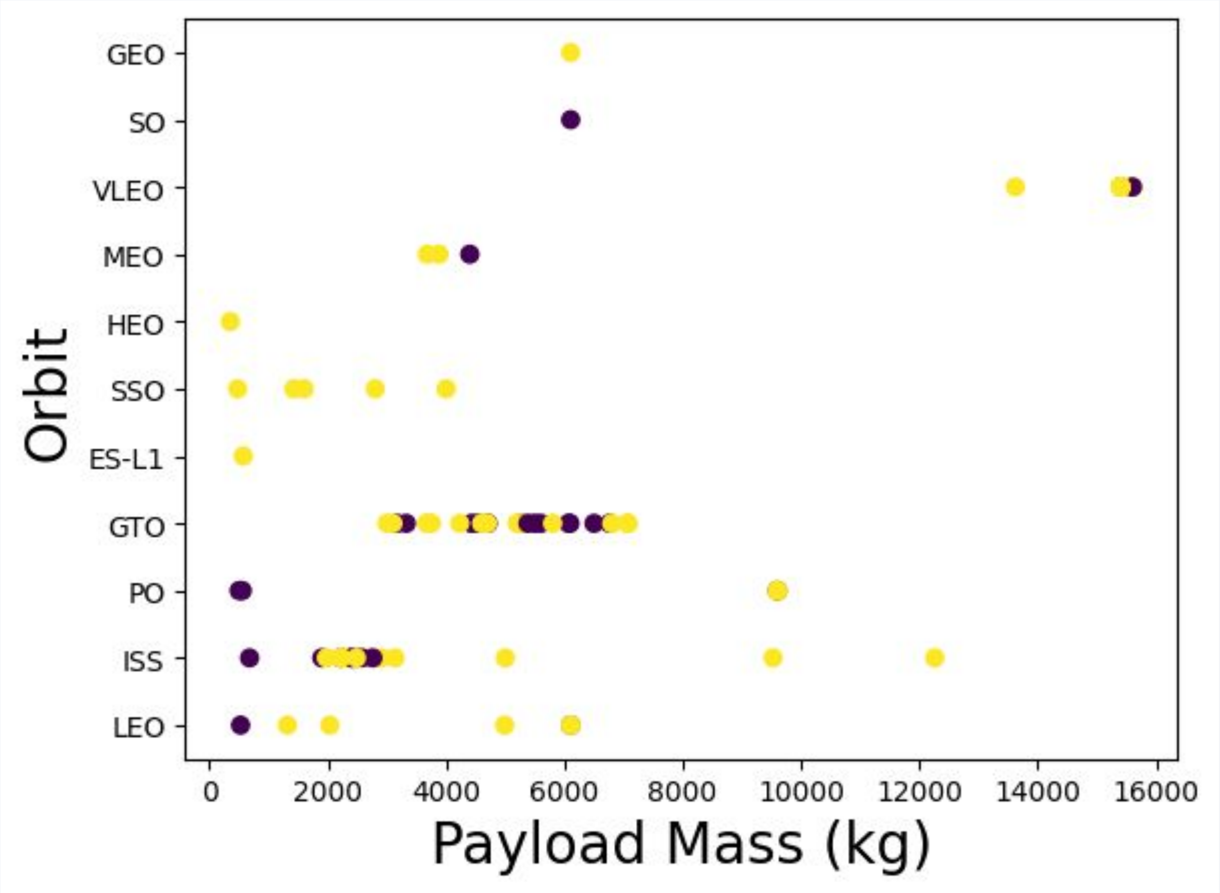
Flight Number vs. Orbit Type

This scatter plot shows the relationship between Flight Number and Orbit type, colored by the success of the landing. It indicates that for LEO (Low Earth Orbit), there is a positive correlation between the number of flights and success rates, while GTO (Geostationary Transfer Orbit) shows no clear relationship.



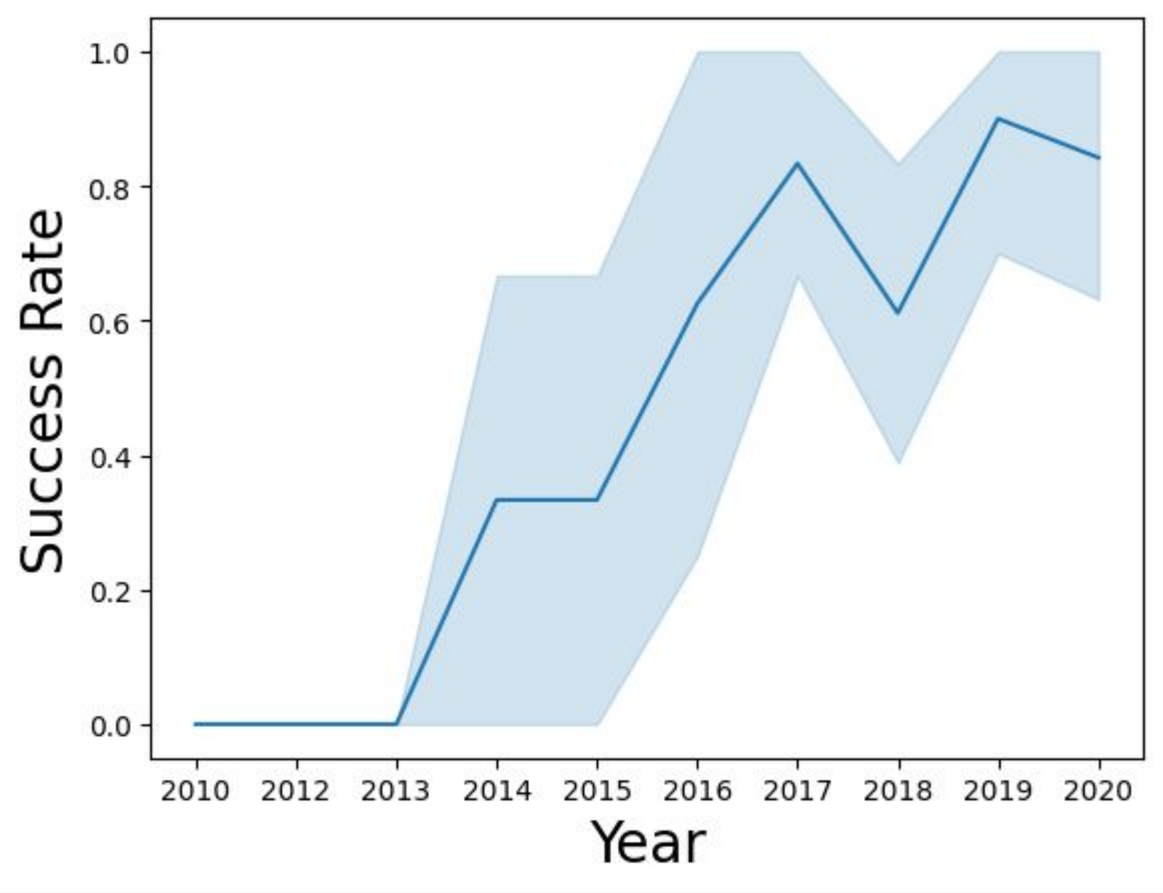
Payload vs. Orbit Type

This scatter plot examines the relationship between Payload Mass and Orbit type, again colored by landing success. It suggests that heavier payloads tend to have higher success rates in certain orbits, such as Polar and LEO, while GTO shows mixed results.



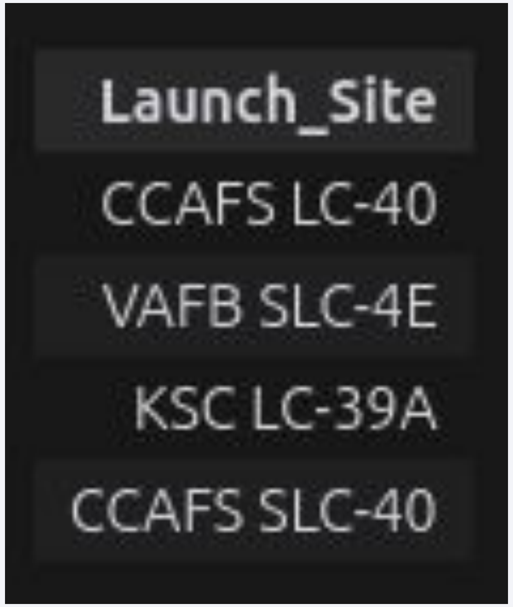
Launch Success Yearly Trend

The line chart illustrates the trend of launch success rates over the years. It shows that the success rate has generally increased since 2013, with a notable rise after 2015, indicating improvements in launch technology and operational practices.



All Launch Site Names

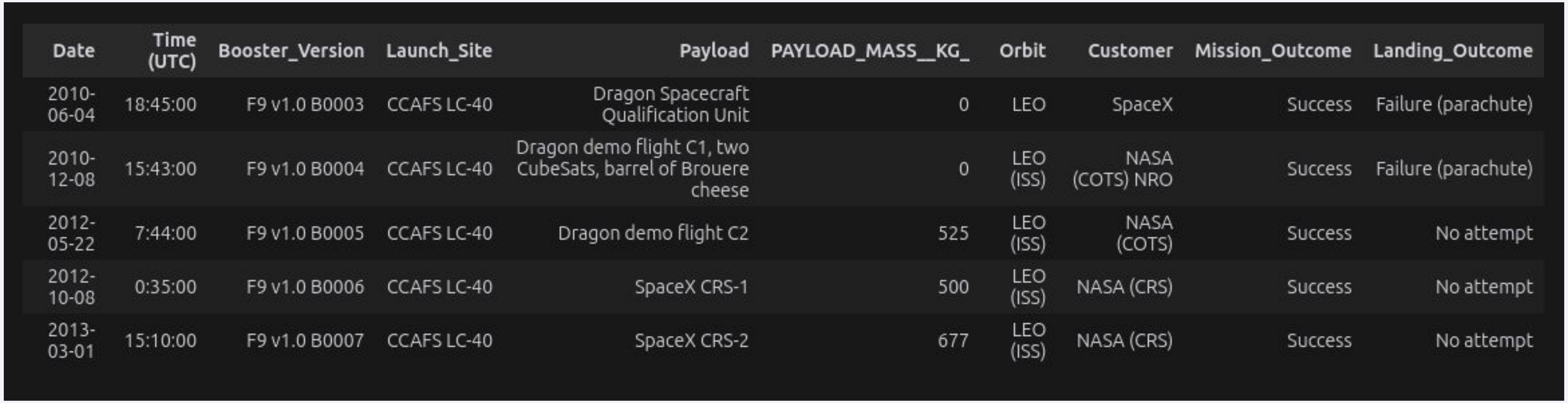
Lists the distinct launch sites used by SpaceX for missions



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

Launch Site Names Begin with 'CCA'

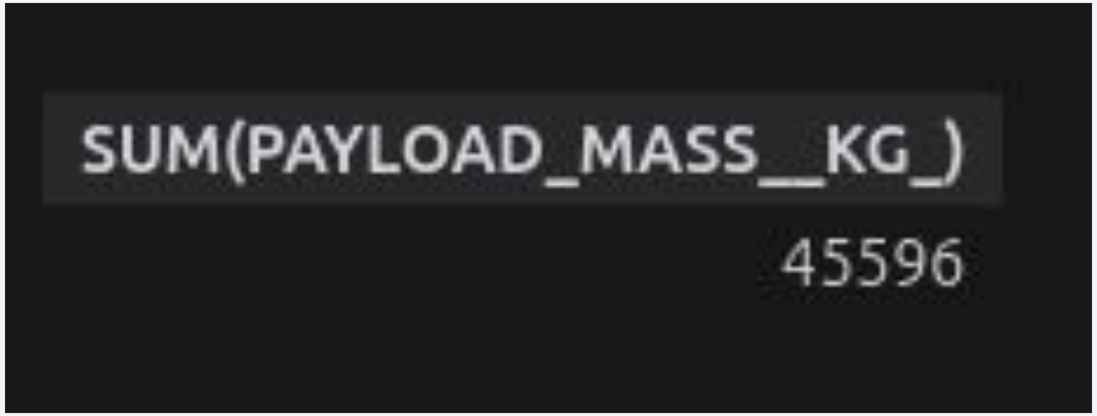
- Find 5 records where launch sites begin with `CCA`
- Shows the first 5 launches from sites that begin with 'CCA', indicating the frequency of launches from these sites



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

Total Payload Mass

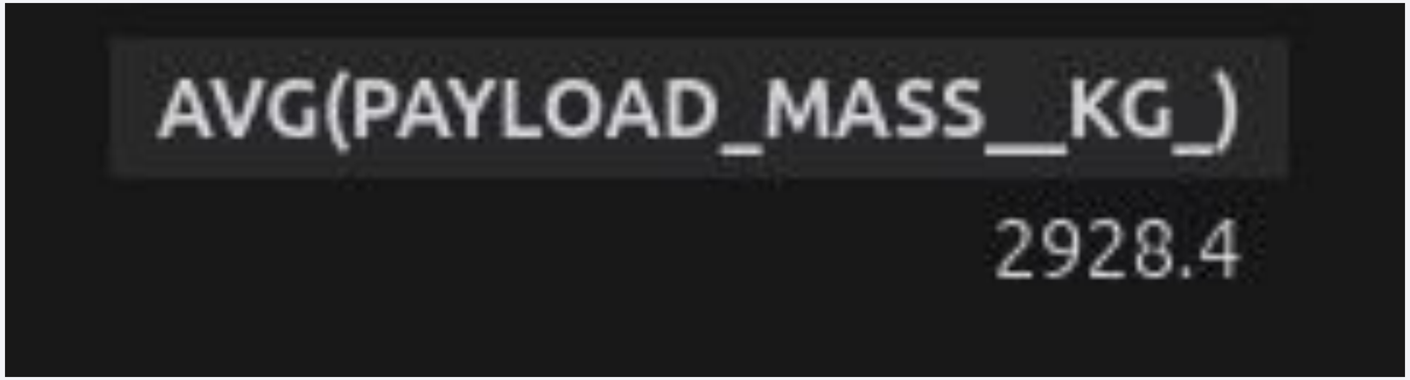
- Calculate the total payload carried by boosters from NASA
- The total mass of payloads carried by SpaceX missions for NASA's CRS program is 45,596 kg.



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

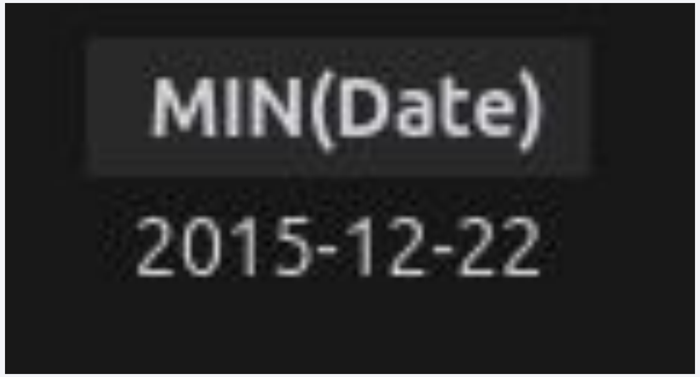
- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass for missions using the F9 v1.1 booster is approximately 2,928.4 kg.



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

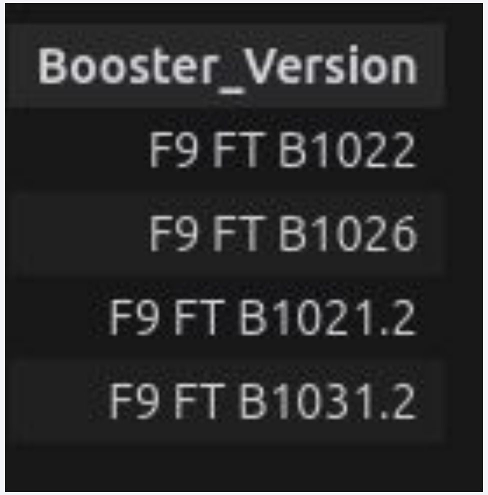
- Find the dates of the first successful landing outcome on ground pad
- The earliest date when a successful landing outcome on a ground pad was achieved is December 22, 2015.



MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

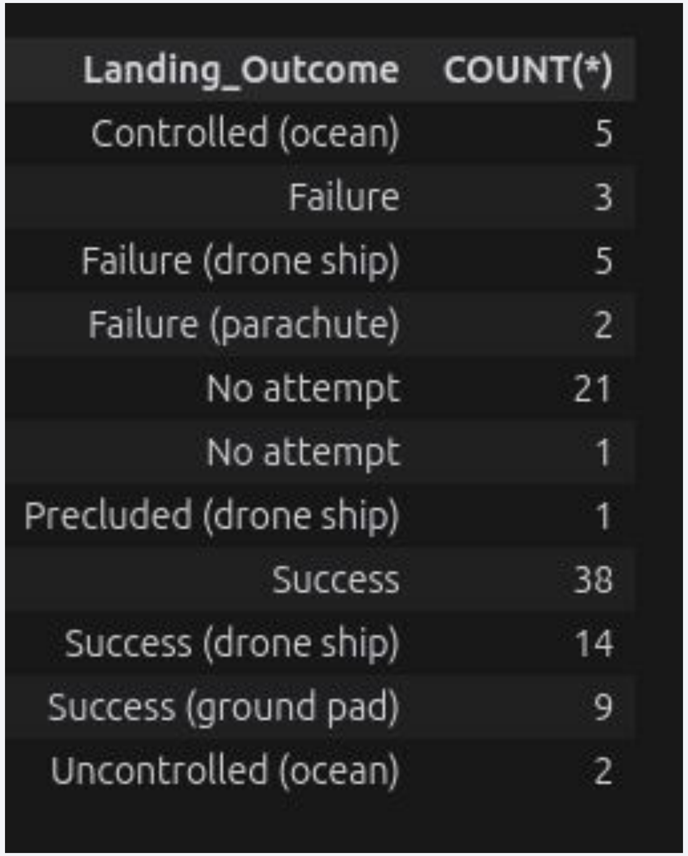
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Identifies boosters that successfully landed on a drone ship with payload masses between 4,000 kg and 6,000 kg.



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

Total Number of Successful and Failure Mission Outcomes

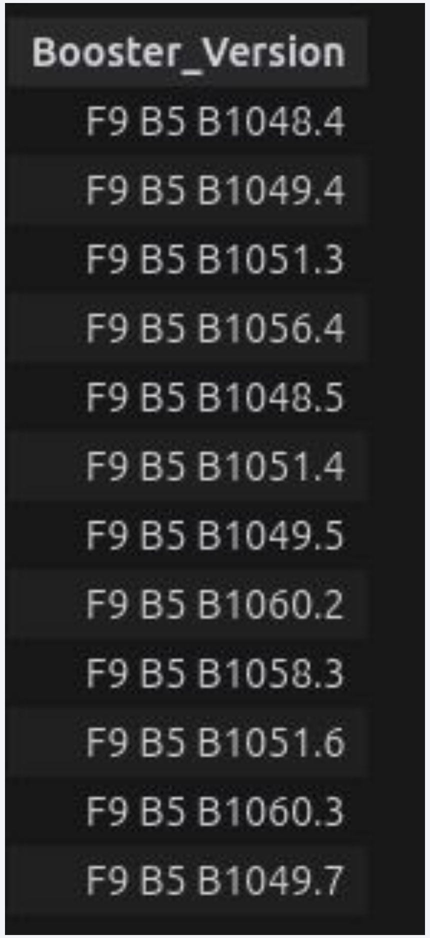
- Calculate the total number of successful and failure mission outcomes
- Displays the total number of missions categorized by their landing outcomes, showing the distribution of successes and failures.



Landing_Outcome	COUNT(*)
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

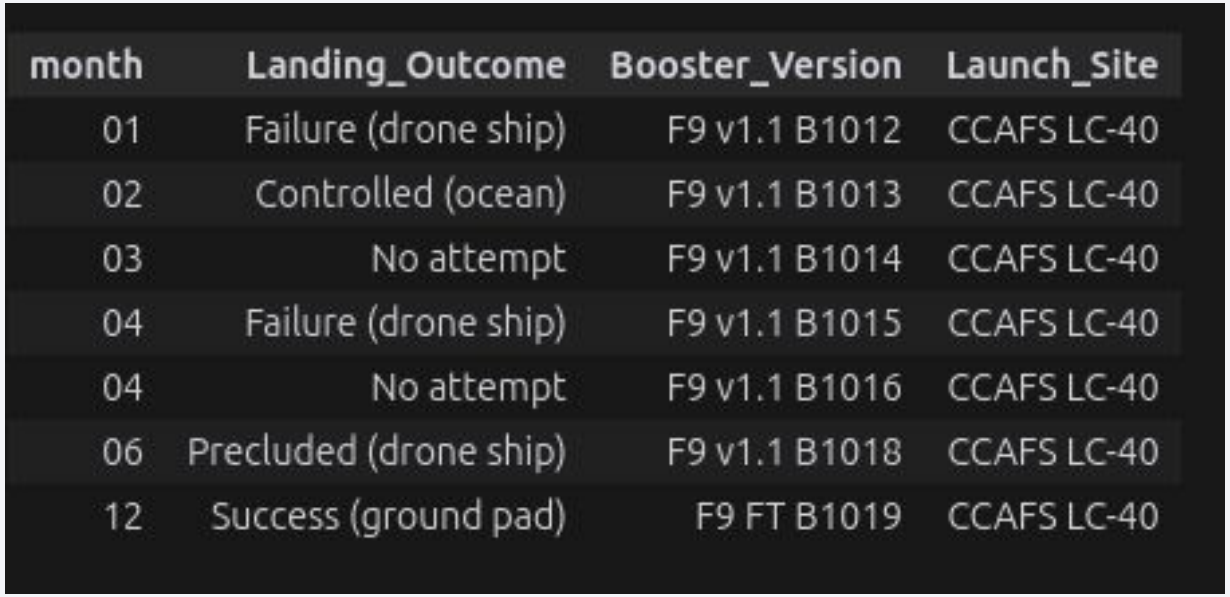
- List the names of the booster which have carried the maximum payload mass
- Identifies the booster versions that have carried the maximum payload mass recorded in the dataset.



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
- Lists landing outcomes categorized by month for the year 2015, focusing on failures and other outcomes.

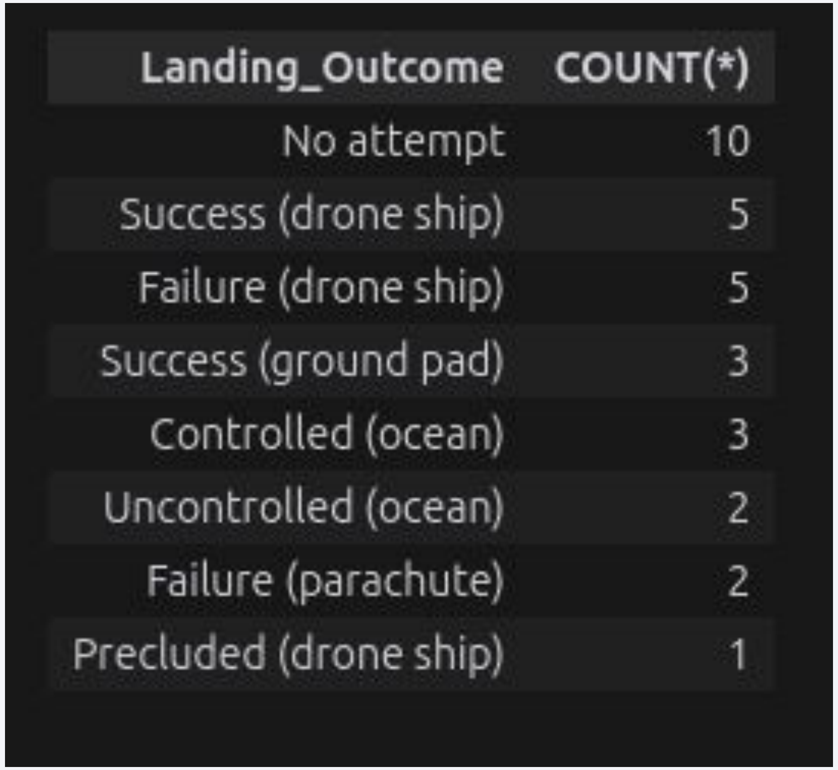


month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	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

- Ranks the count of landing outcomes between the specified dates, showing the most common outcomes in descending order.

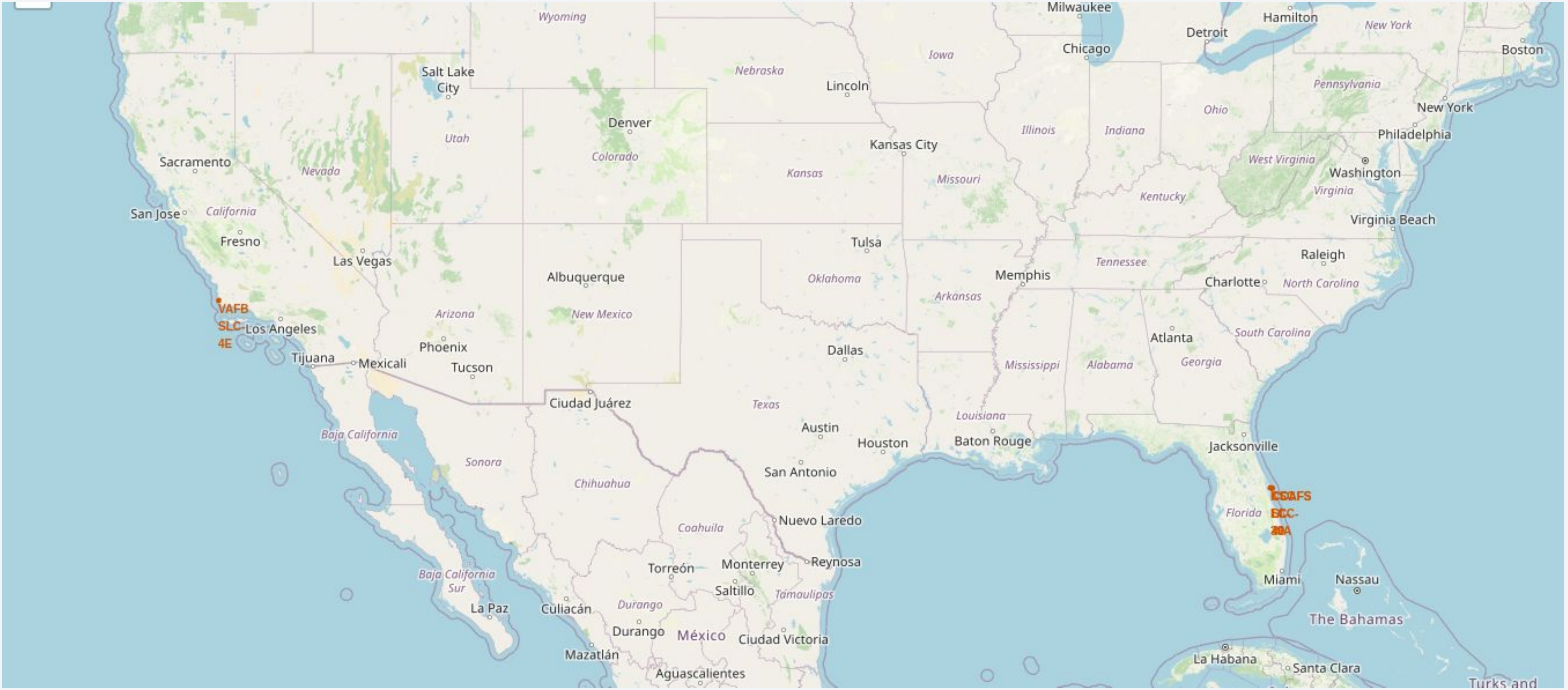


Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

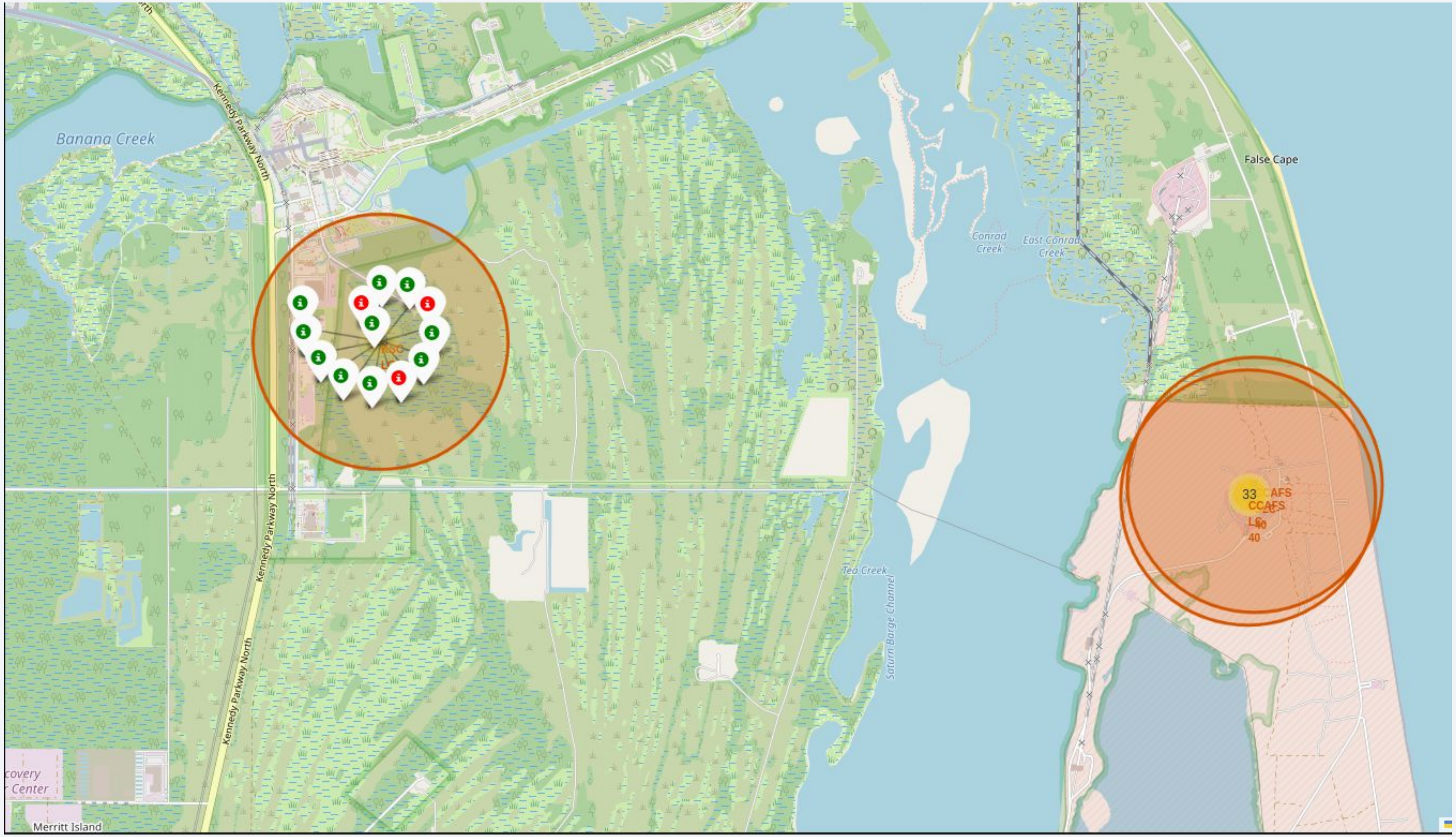
Launch Sites Proximities Analysis

Launch site markers



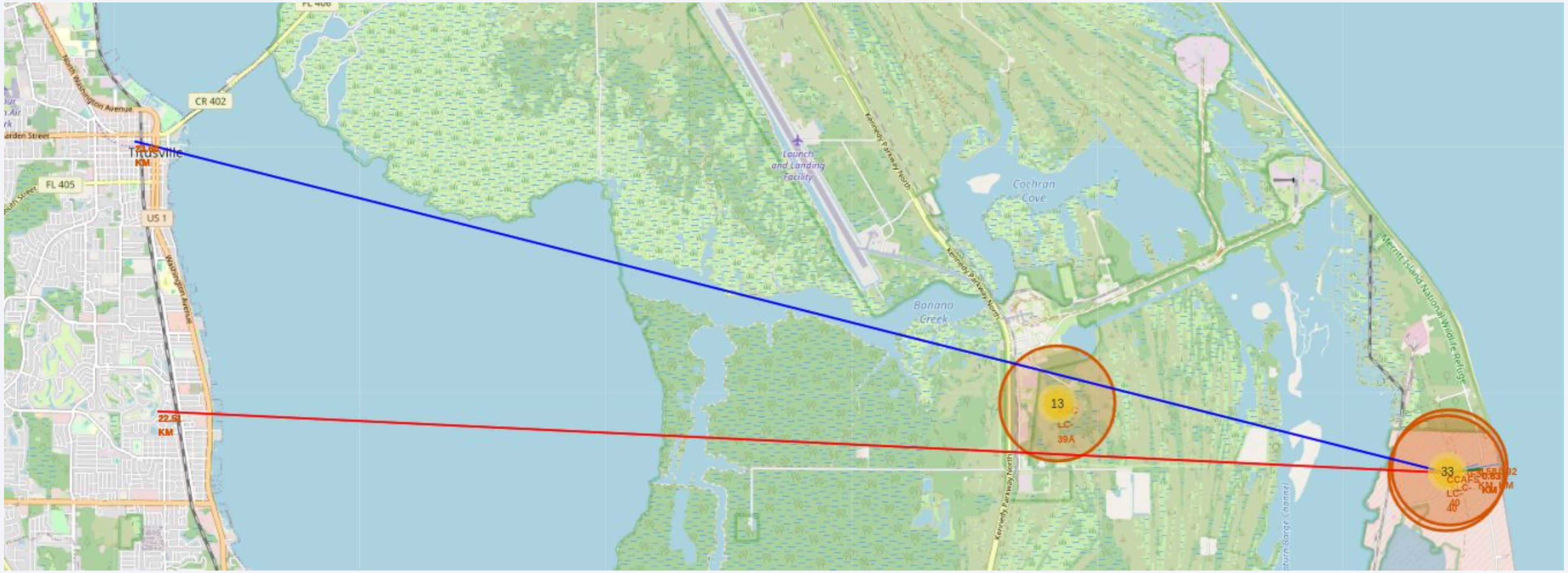
A Map displaying markers of all the launch sites SpaceX have used

Launch markers

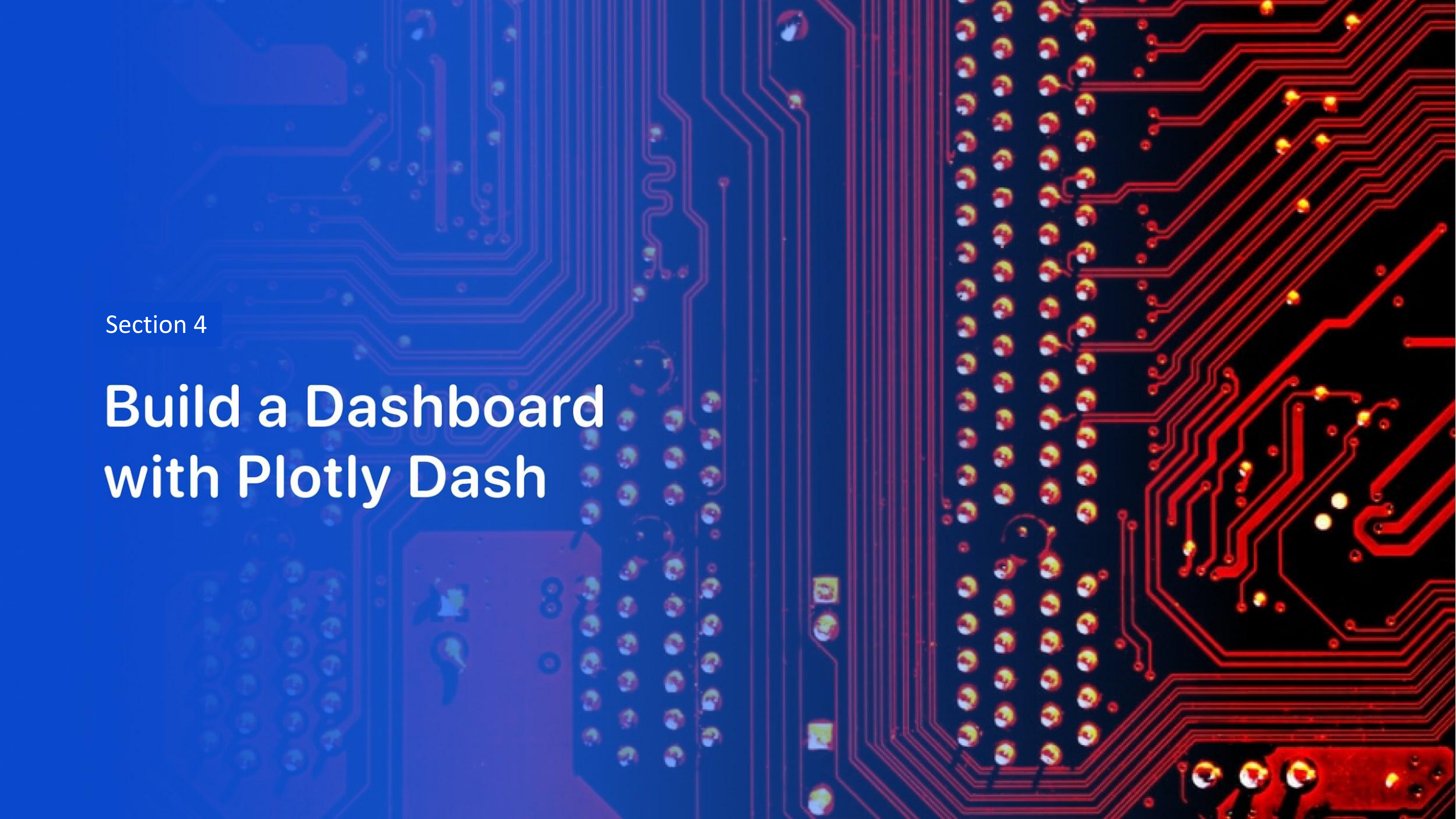


Map displaying markers showing each launch at the site. Green indicates successful launch and red indicates a failed launch.

Distances from Cape Canaveral launch site



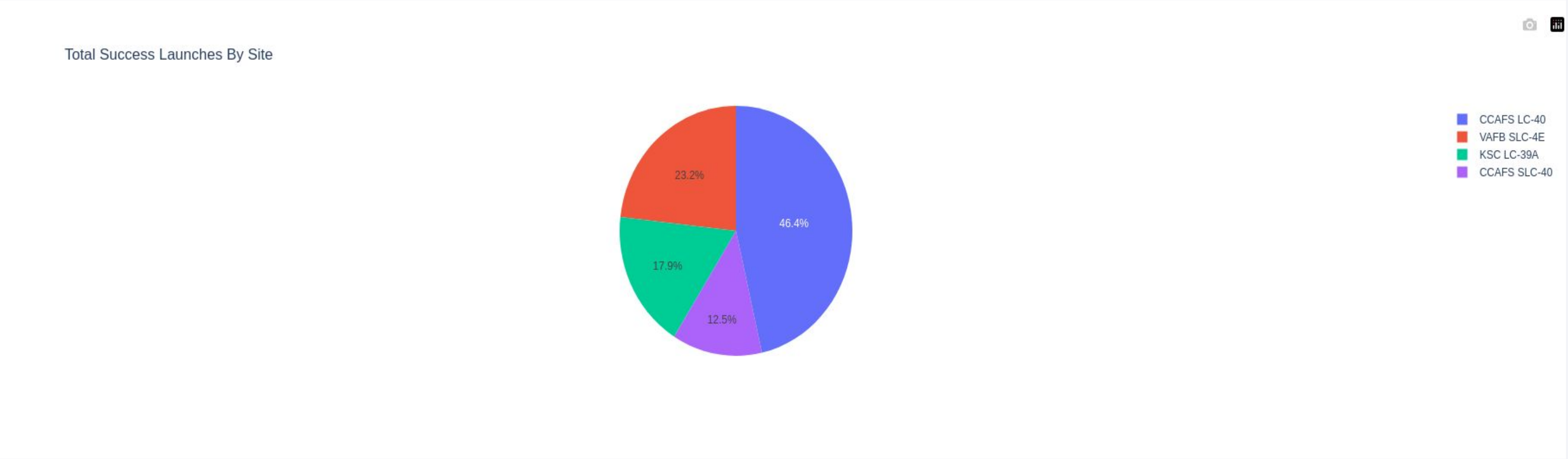
This map displays key points of interest and their relevant distance from the launch site.



Section 4

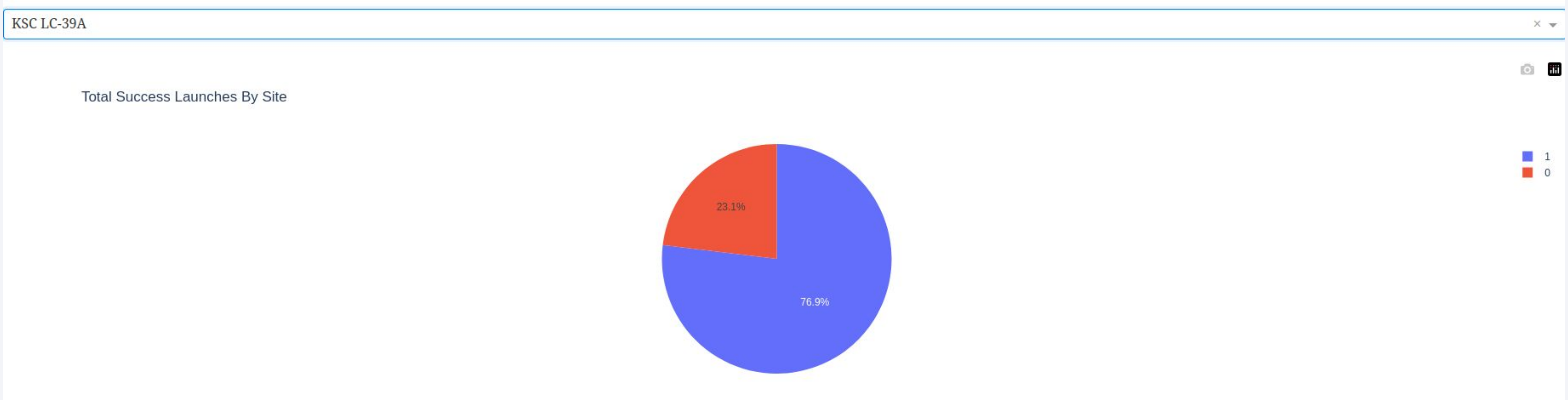
Build a Dashboard with Plotly Dash

Launch site successes



We can see from this chart that the majority of successful launches came from CCAFS LC-40 Launch Site.

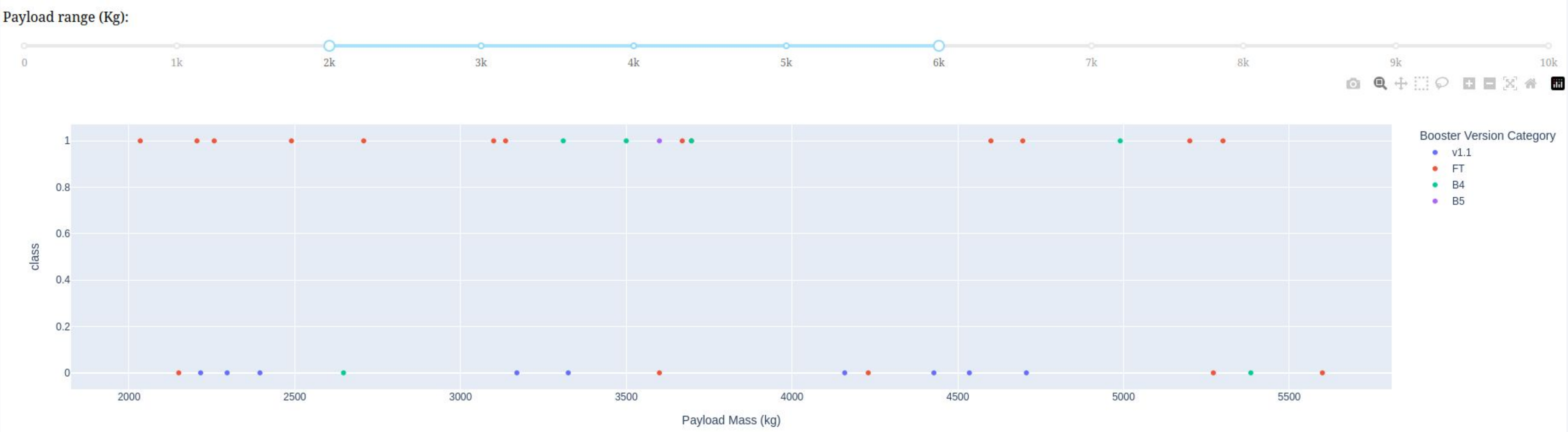
Launch site successes cont.



KSC LC-39A

Here we can see that the KSC LC-39A has the highest success rate of all the launch sites.

Payload success rates

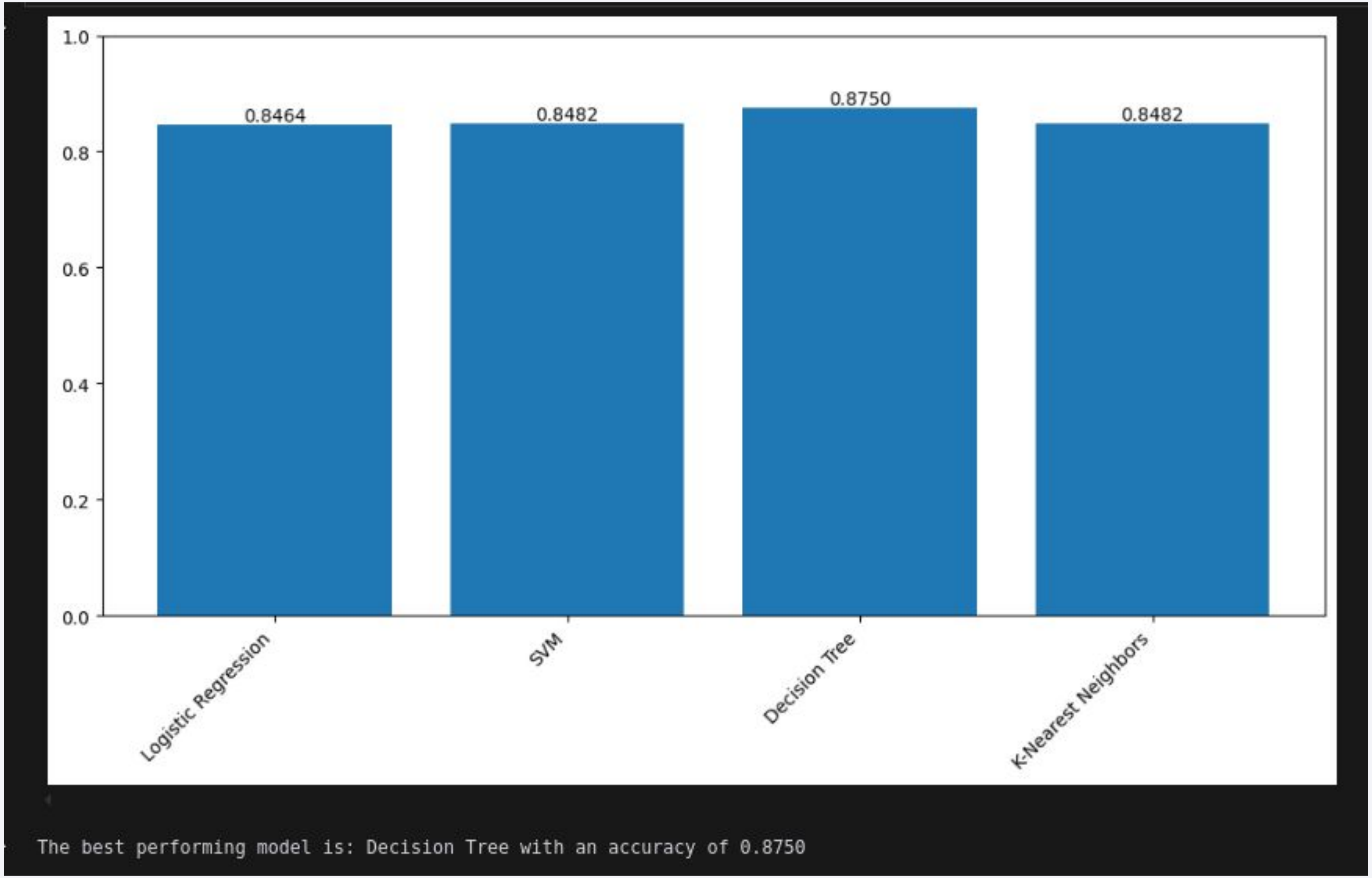


From this graph we can see that the FT booster version has a very good success rate for payloads between 2000 and 6000 kg. Whereas the v1.1 booster has a bad success rate.

Section 5

Predictive Analysis (Classification)

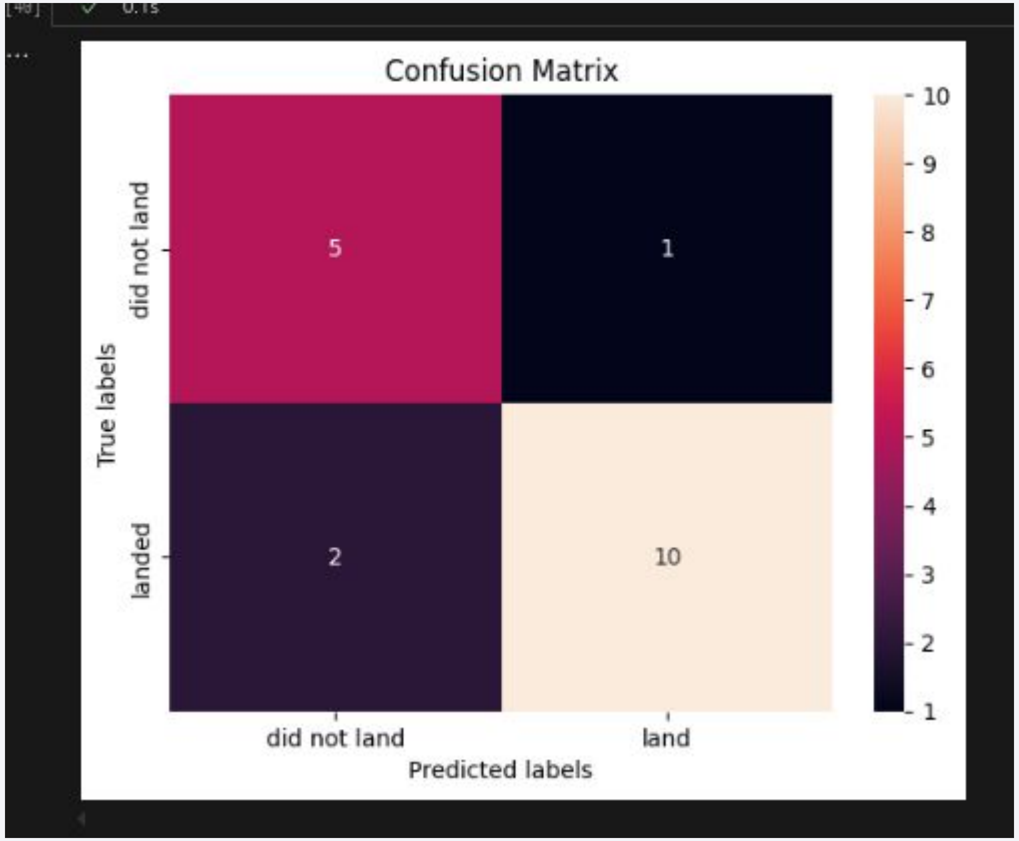
Classification Accuracy



```
The best performing model is: Decision Tree with an accuracy of 0.8750
```

Confusion Matrix

- This shows the decision tree model correctly predicted 15 test cases.



Conclusions

- Decision tree is the best classifier to determine which launches will be successful
- Kennedy space center has the best success rate for launches
- The FT booster is a lot more successful than the others

Thank you!