IBM Developer
SKILLS NETWORK

Data Science Capstone Project

Cedric W.

March 11, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis for Classification

Summary of All Results

- Exploratory Data Analysis Results
- Interactive Analytics
- Predictive Modeling Evaluation

Introduction

Project Background

SpaceX, an American aerospace manufacturer and space transportation company founded in 2002, is headquartered in Hawthorne, California. The company has revolutionized the space industry with its cutting-edge technologies and ambitious vision of reducing space transportation costs.

One of SpaceX’s key innovations in cost reduction is the reuse of the Falcon 9 rocket’s first stage. While a typical rocket launch from other manufacturers may cost upwards of 165 million dollars, SpaceX advertises Falcon 9 launches at just 62 million dollars.

Predicting the successful landing of the first stage is crucial for estimating the cost of a SpaceX launch. This project aims to develop a machine learning model to predict whether the first stage will land successfully and be reused.

Business Problem

- How do dependent variables, such as payload mass, launch site, number of flights, and orbital parameters, affect the success of the first-stage landing?
- Has the rate of successful landings improved over the years?
- Which algorithm is most suitable for binary classification in this context?

Section 1

Methodology

Methodology

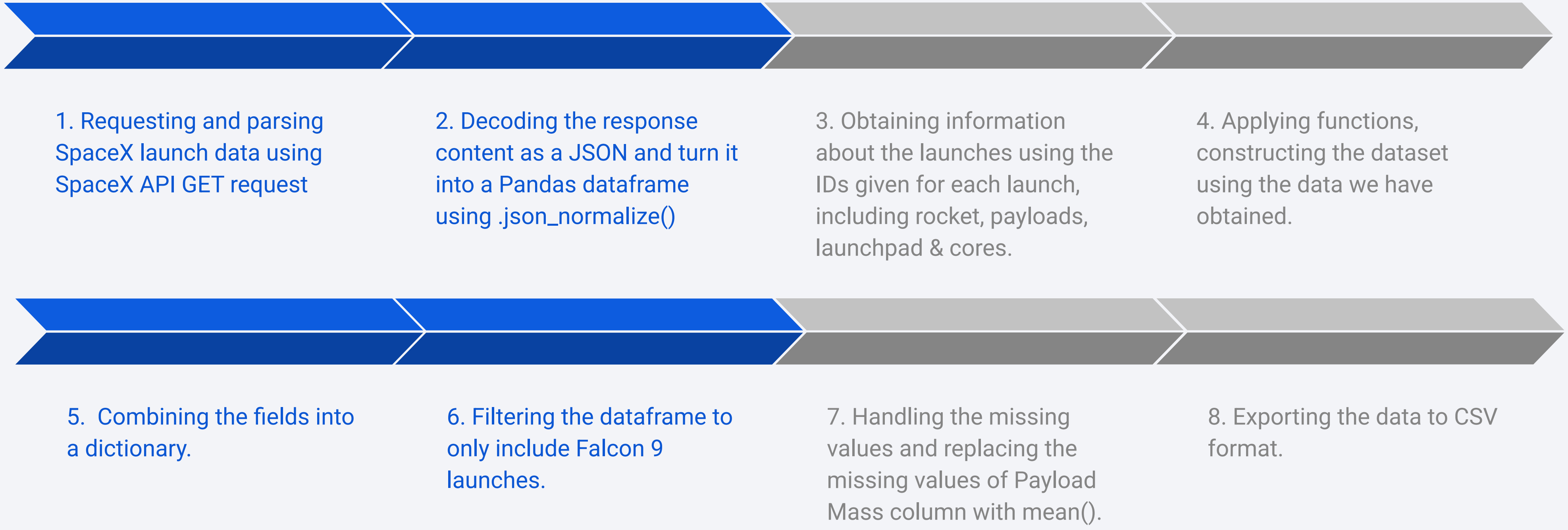
Executive Summary

- Data collection methodology:
  - Using SpaceX API
  - Using Web scraping from Wikipedia page
- Perform data wrangling
  - Filtering the dataset
  - Handling missing values
  - Applying one-hot encoding to prepare the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating classification models to identify the best-performing model

Data Collection

The data collection process combined API requests from the SpaceX API with web scraping from SpaceX’s Wikipedia to obtain complete and comprehensive information about the launches, enabling more detailed analysis.

Data Collection – SpaceX API



1. Requesting and parsing SpaceX launch data using SpaceX API GET request

2. Decoding the response content as a JSON and turn it into a Pandas dataframe using .json_normalize()

3. Obtaining information about the launches using the IDs given for each launch, including rocket, payloads, launchpad & cores.

4. Applying functions, constructing the dataset using the data we have obtained.

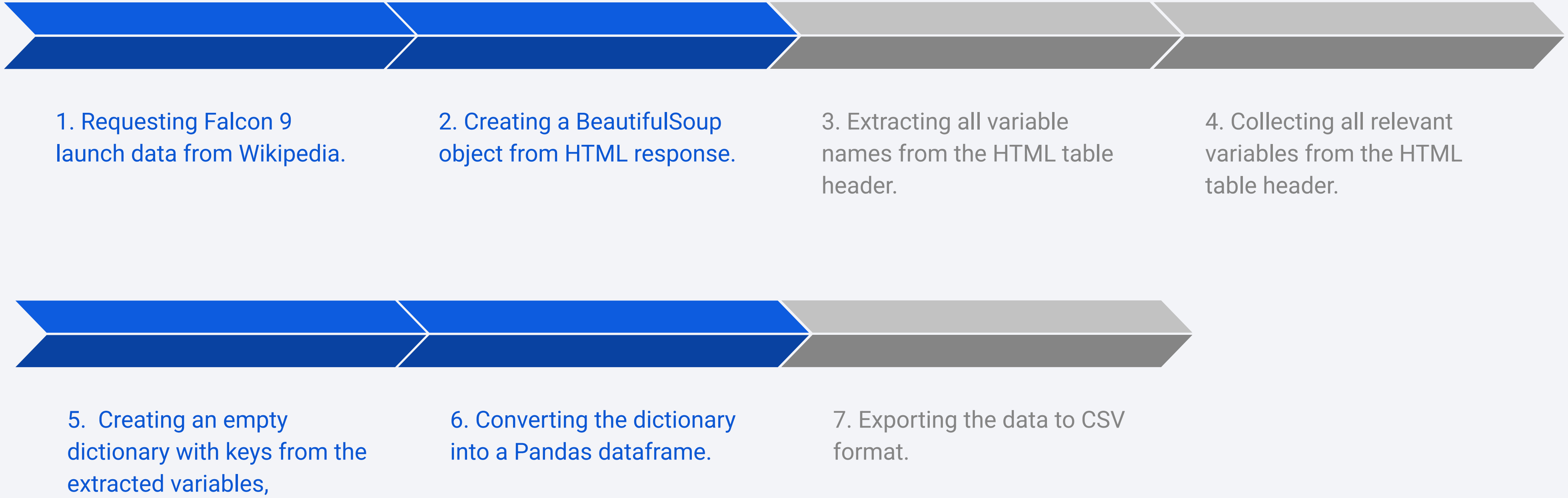
5. Combining the fields into a dictionary.

6. Filtering the dataframe to only include Falcon 9 launches.

7. Handling the missing values and replacing the missing values of Payload Mass column with mean().

8. Exporting the data to CSV format.

Data Collection - Web Scraping



1. Requesting Falcon 9 launch data from Wikipedia.

2. Creating a BeautifulSoup object from HTML response.

3. Extracting all variable names from the HTML table header.

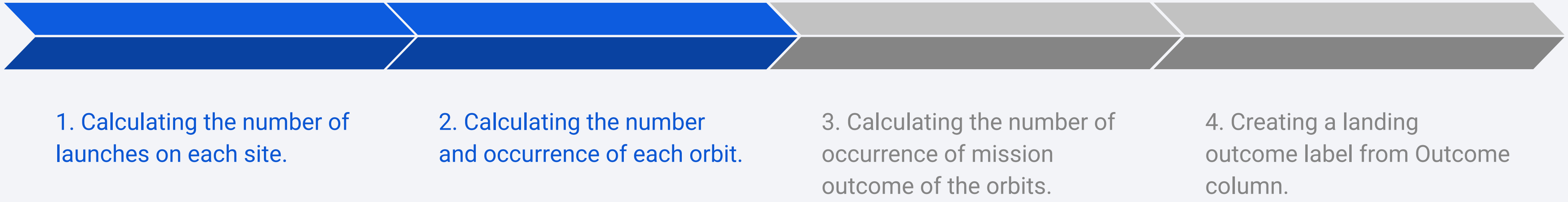
4. Collecting all relevant variables from the HTML table header.

5. Creating an empty dictionary with keys from the extracted variables,

6. Converting the dictionary into a Pandas dataframe.

7. Exporting the data to CSV format.

Data Wrangling



1. Calculating the number of launches on each site.

2. Calculating the number and occurrence of each orbit.

3. Calculating the number of occurrence of mission outcome of the orbits.

4. Creating a landing outcome label from Outcome column.

The dataset contains various outcomes related to booster landings, including both successes and failures.

- True Ocean and True RTLS indicate successful landings on the ocean and ground pad, respectively.
- True ASDS denotes a successful landing on a drone ship.
- On the other hand, False Ocean, False RTLS, and False ASDS represent failed landings in the respective locations.

We convert these outcomes into binary labels: 1 for successful landings and 0 for failed landings.

EDA with Data Visualization

Scatterplots, bar charts, and line charts are used to gain a deeper understanding of our data.

Scatterplots: These plots display the relationship between x and y variables, helping to determine any correlations.

- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type

Bar Charts: These charts are useful for comparing data across different categories with measured values.

- Orbit Type vs. Success Rate

Line Charts: These plots illustrate trends and changes over time in time series data.

- Success Rate Yearly Trend

EDA with SQL

The following SQL queries were performed:

- Displaying the names of the unique launch sites in the space mission.
- Displaying five records where launch sites begin with the string ‘CCA’.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by boosters version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Ranking 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.

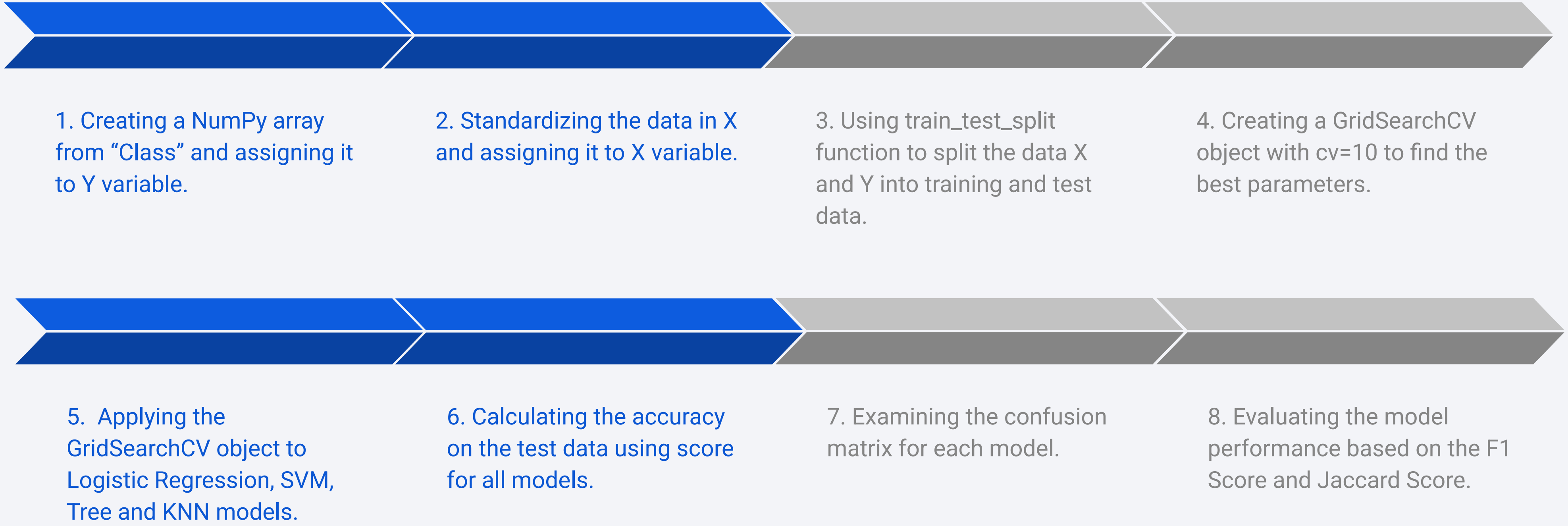
Build an Interactive Map with Folium

- Placed a marker with a circle, popup label, and text label at NASA Johnson Space Center using its latitude and longitude as the starting location.
- Placed markers with circles, popup labels, and text labels at all launch sites to indicate their geographical locations and proximity to the equator and coastlines.
- Used colored markers to represent launch outcomes: green for success and red for failure.
- Utilized marker clustering to identify launch sites with relatively high success rates.
- Added colored lines to display distances between the KSC LC-39A launch site (as an example) and nearby features such as railways, highways, coastlines, and the closest city.

Build a Dashboard with Plotly Dash

- Implemented a dropdown menu to select a specific launch site.
- Displayed a pie chart showing the total number of successful launches across all sites.
- When a specific launch site is selected, the chart shows the count of successful vs. failed launches for that site.
- Included a slider to adjust and select the desired payload mass range.
- Created a scatter plot to visualize the relationship between payload mass and launch success across different booster versions.

Predictive Analysis (Classification)



1. Creating a NumPy array from “Class” and assigning it to Y variable.

2. Standardizing the data in X and assigning it to X variable.

3. Using train_test_split function to split the data X and Y into training and test data.

4. Creating a GridSearchCV object with cv=10 to find the best parameters.

5. Applying the GridSearchCV object to Logistic Regression, SVM, Tree and KNN models.

6. Calculating the accuracy on the test data using score for all models.

7. Examining the confusion matrix for each model.

8. Evaluating the model performance based on the F1 Score and Jaccard Score.

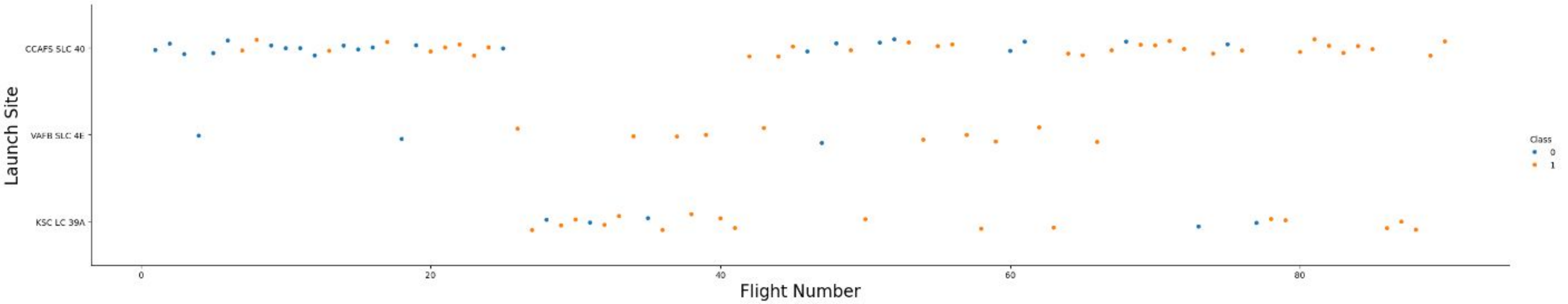
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

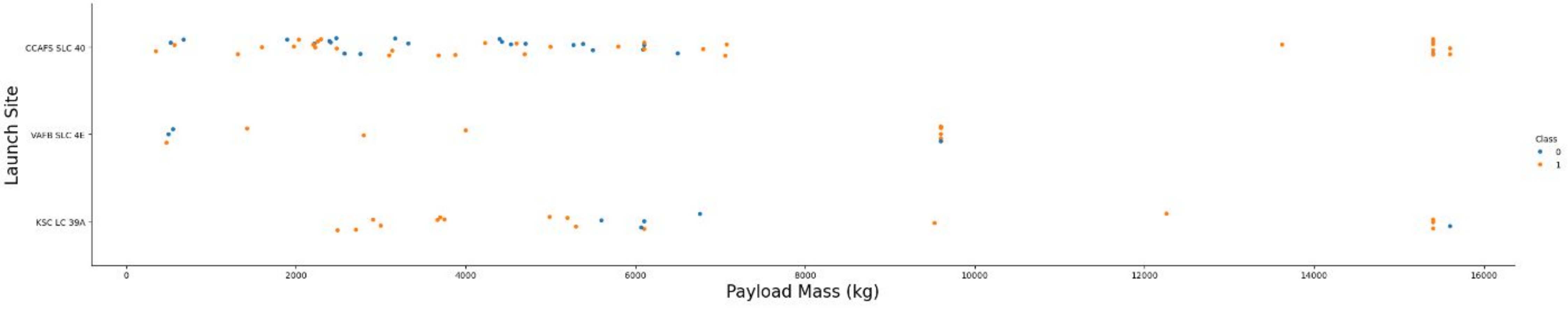


- Based on the scatterplot, most of the earliest flights failed, while all of the more recent flights were successful.

- The majority of flights were launched from CCAFS SLC 40.

- Compared to other launch sites, VAFB SLC 4E and KSC LC 39A have higher success rates.

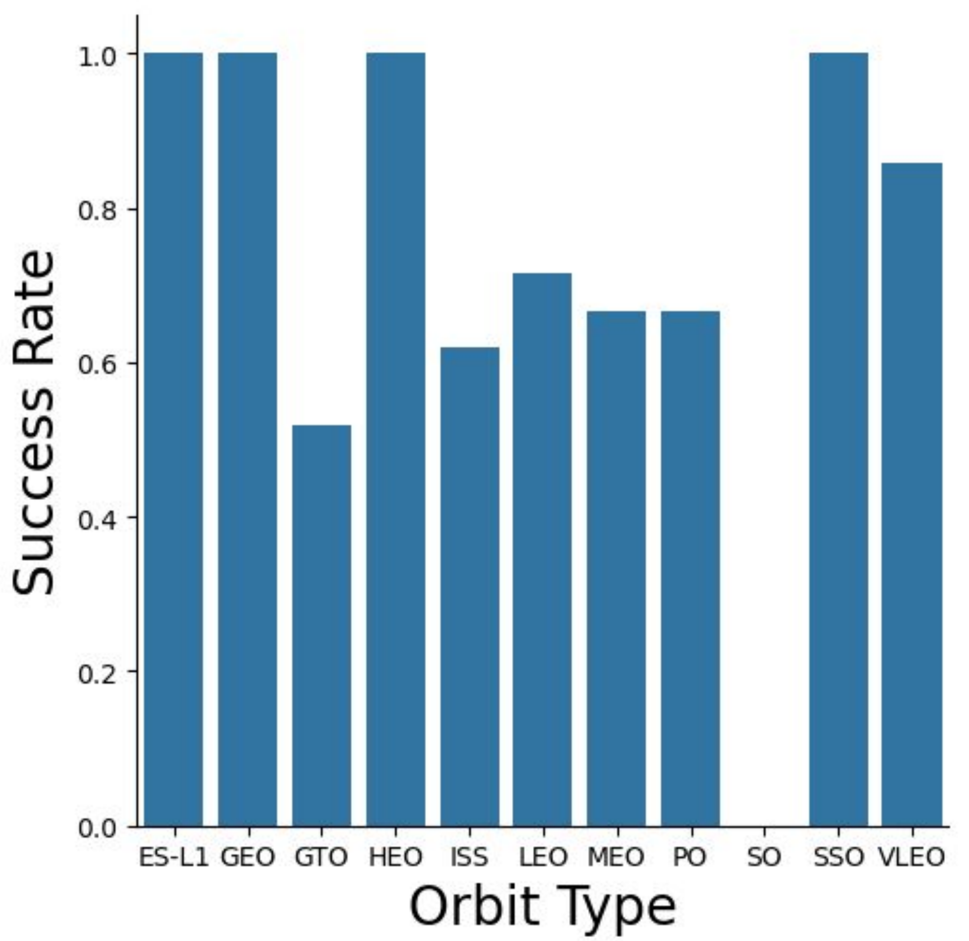
Payload vs. Launch Site



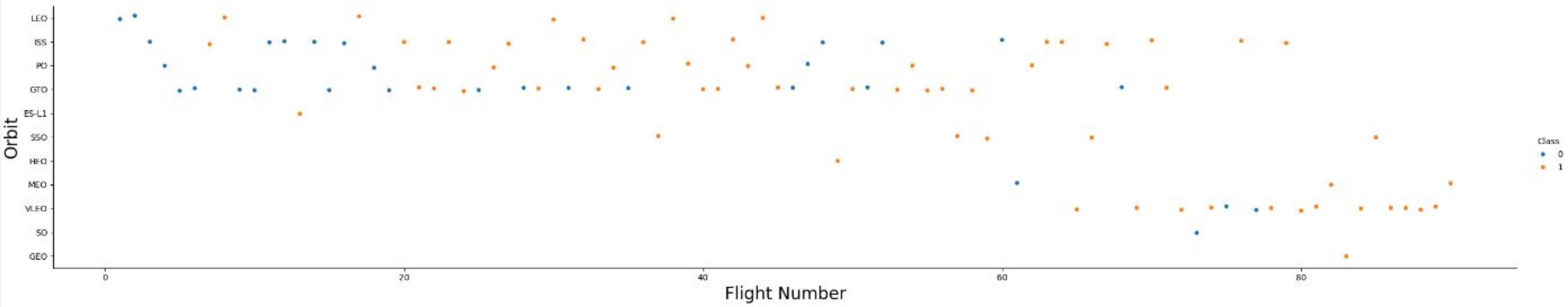
- Launches with higher payload masses tend to show higher success rates across all launch sites.
- In the scatterplot, most landings with a payload mass exceeding 7,000 kg are successful.
- KSC LC 39A has a 100% success rate for landings with payload masses under 5,550 kg.

Success Rate vs. Orbit Type

- The orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate.
- The orbit type SO has a 0% success rate.
- The orbit types GTO, ISS, LEO, MEO, PO, and VLEO have success rates ranging from 50% to 85%.

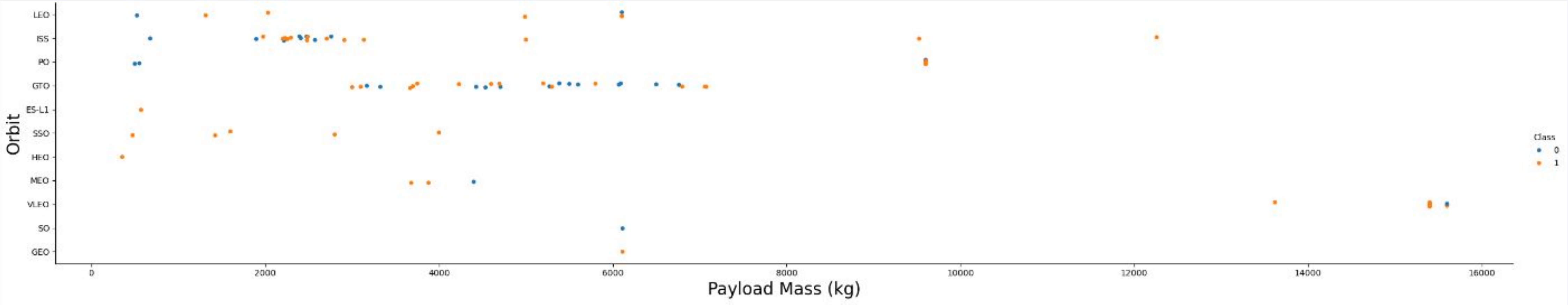


Flight Number vs. Orbit Type



- Most successful landings occur in LEO, ISS, and PO.
- Failures are more frequent in GTO and some SSO launches.
- As the flight number increases, the proportion of successful landings also increases, indicating improved reliability over time.
- Earlier flights show more failures, especially in GTO.
- Success rates vary significantly with orbit type, with higher success in LEO and ISS compared to GTO and GEO.

Payload vs. Orbit Type



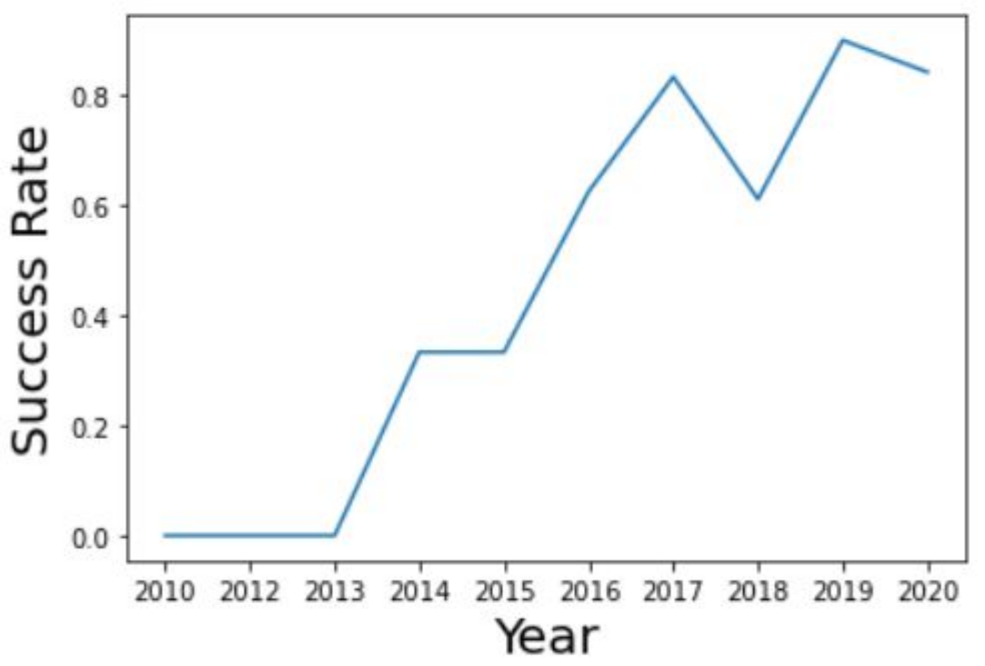
- The majority of successful landings occur with payload masses below 6,000 kg, while higher payload masses (above 10,000 kg) show a greater tendency for landing failures.
- Most successful landings are observed in LEO, ISS, and PO orbits, while GTO and SSO show more failures, especially at higher payload masses.
- A few heavy payload launches (above 14,000 kg) have predominantly failed, suggesting that higher payloads significantly increase landing difficulty.

Launch Success Yearly Trend

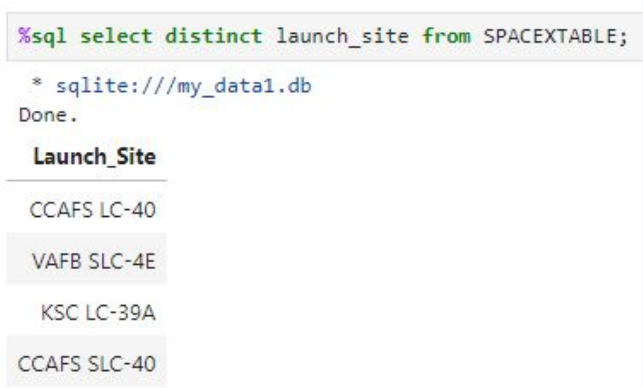
- The success rate has shown an upward trend from 2010 to 2020, indicating significant improvements in landing technology and techniques.

- There was a sharp increase in the success rate around 2015, marking a pivotal point where landing reliability improved significantly.

- After 2017, the success rate stabilized at around 80% to 90%, with minor fluctuations.



All Launch Site Names



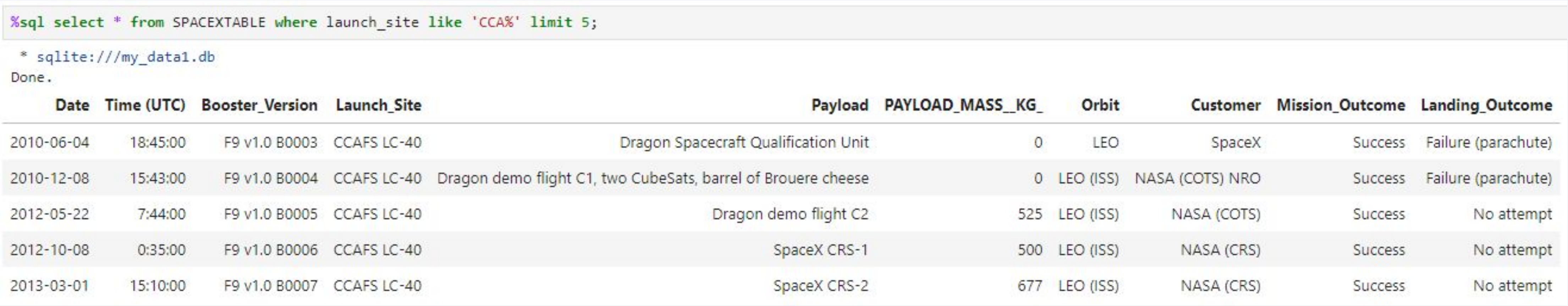
```
%sql select distinct launch_site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Listing the unique launch site names from the SpaceX space mission dataset.

Launch Site Names Begin with 'CCA'



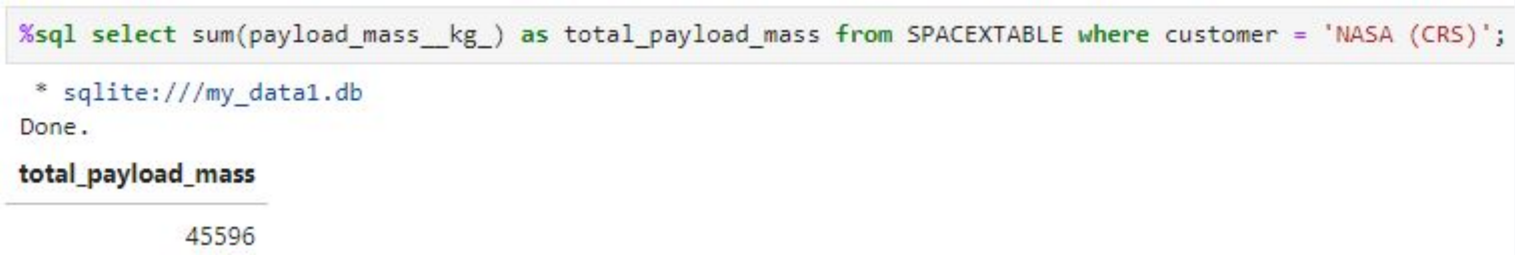
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Displaying 5 records where launch sites begin with ‘CCA’.

Total Payload Mass



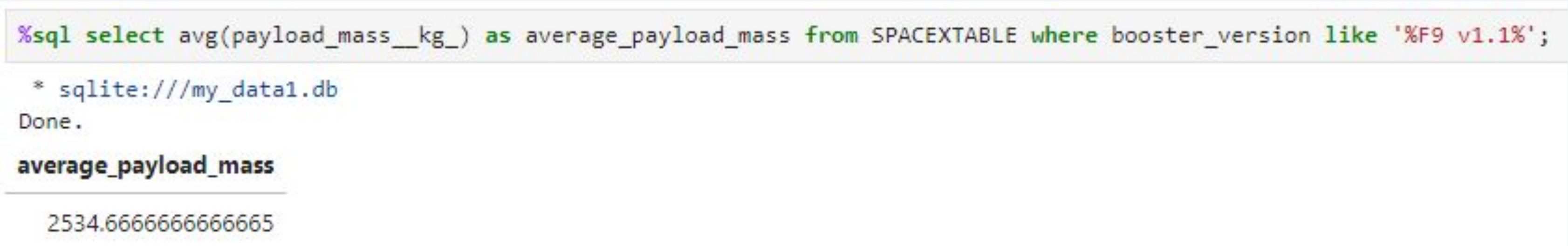
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- Displaying the total payload mass carried by boosters from NASA (CRS).

Average Payload Mass by F9 v1.1

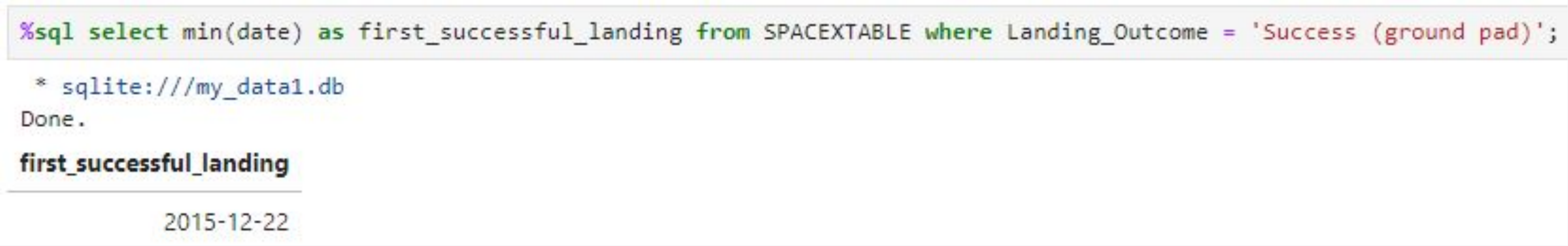


```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
average_payload_mass
   2534.6666666666665
```

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

First Successful Ground Landing Date



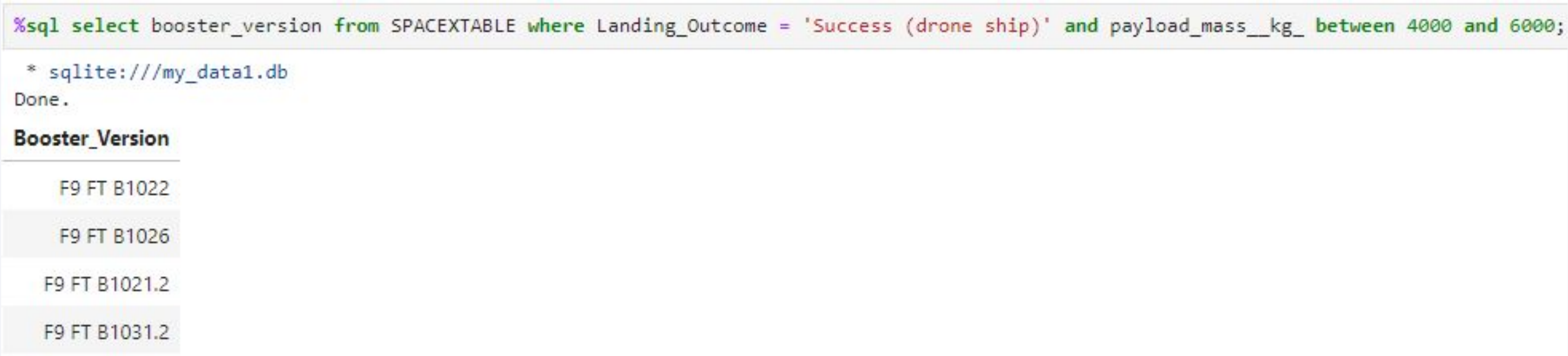
```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

- Listing the dates of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000



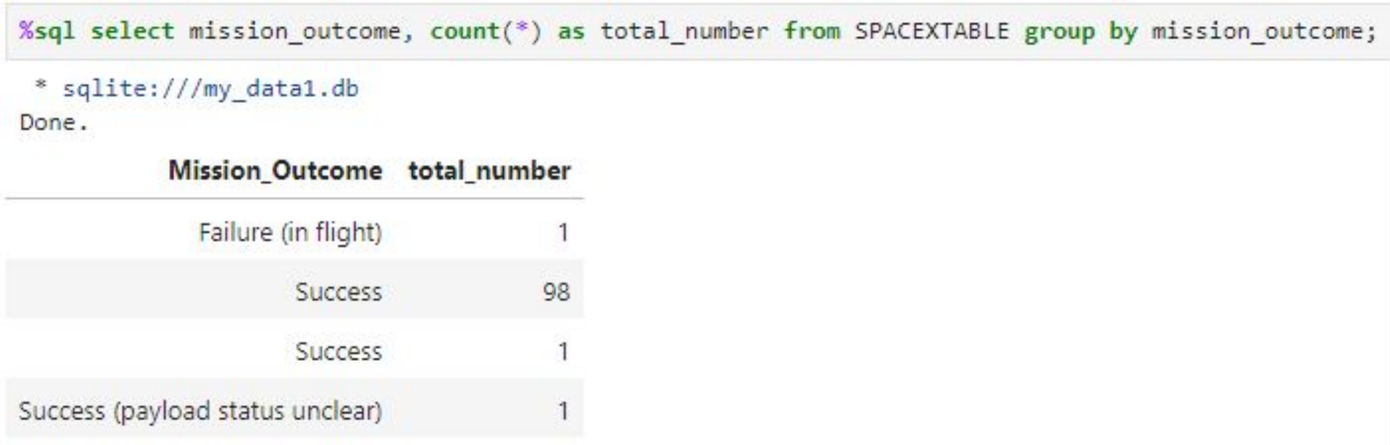
```
%sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

Total Number of Successful and Failure Mission Outcomes



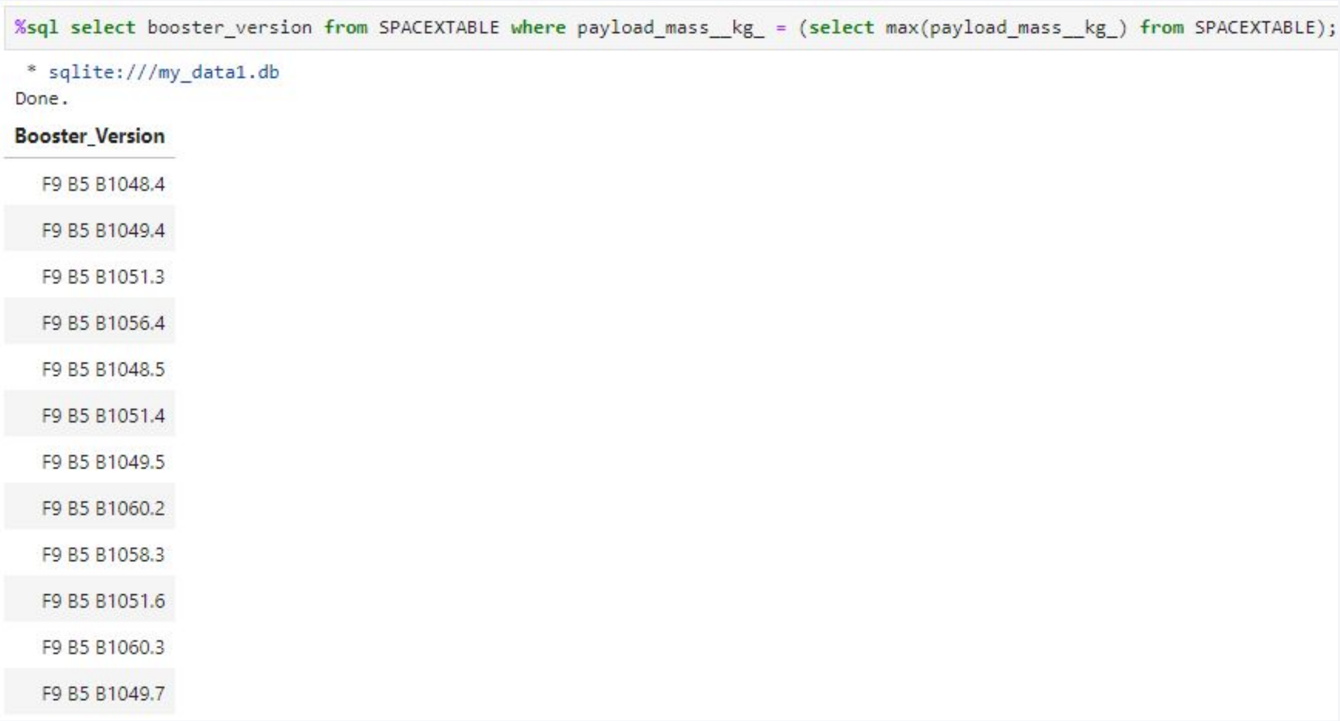
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload



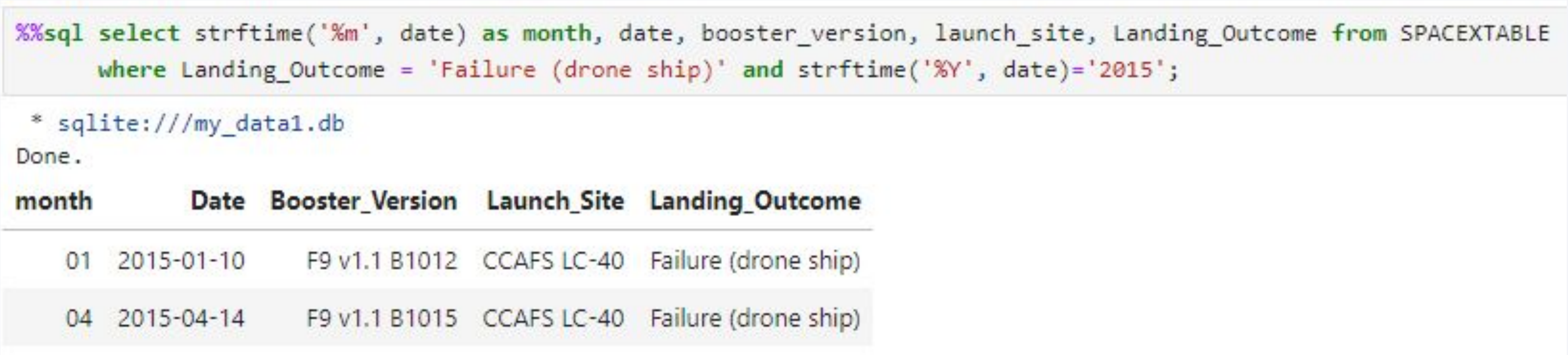
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Listing the names of the booster which have carried the maximum payload mass.

2015 Launch Records



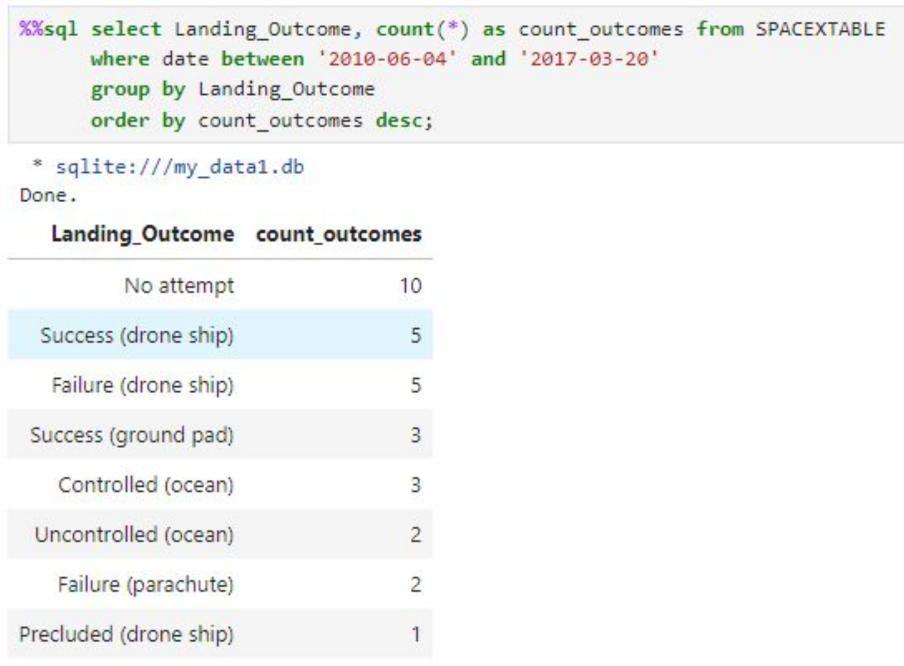
```
%%sql select strftime('%m', date) as month, date, booster_version, launch_site, Landing_Outcome from SPACEXTABLE
    where Landing_Outcome = 'Failure (drone ship)' and strftime('%Y', date)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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



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

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
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

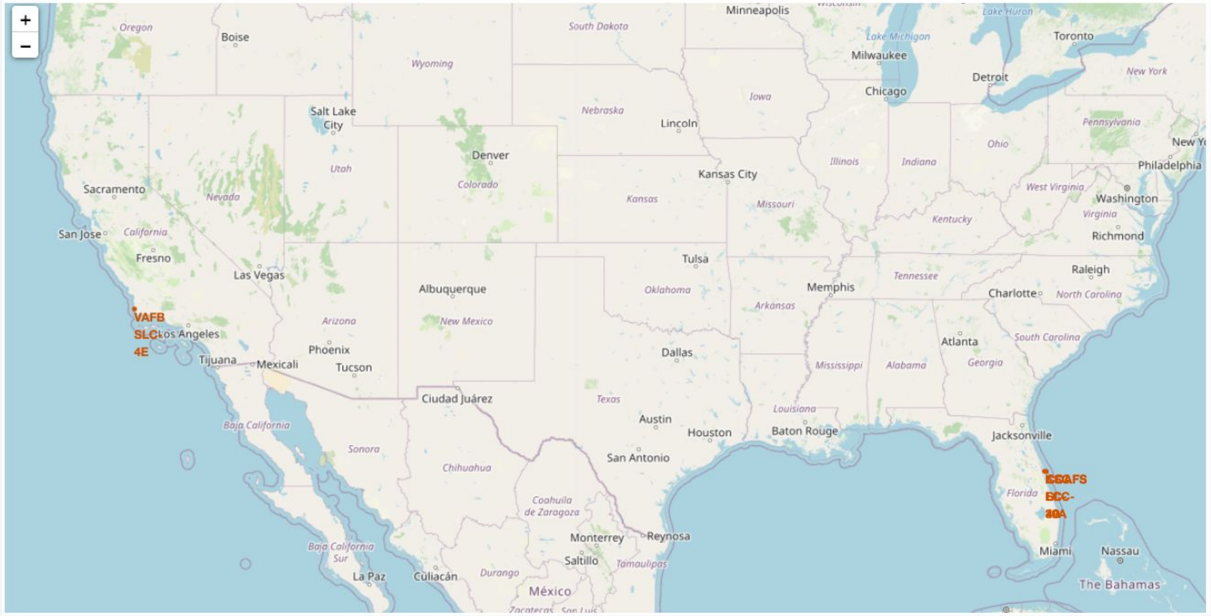
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Section 3

Launch Sites Proximities Analysis

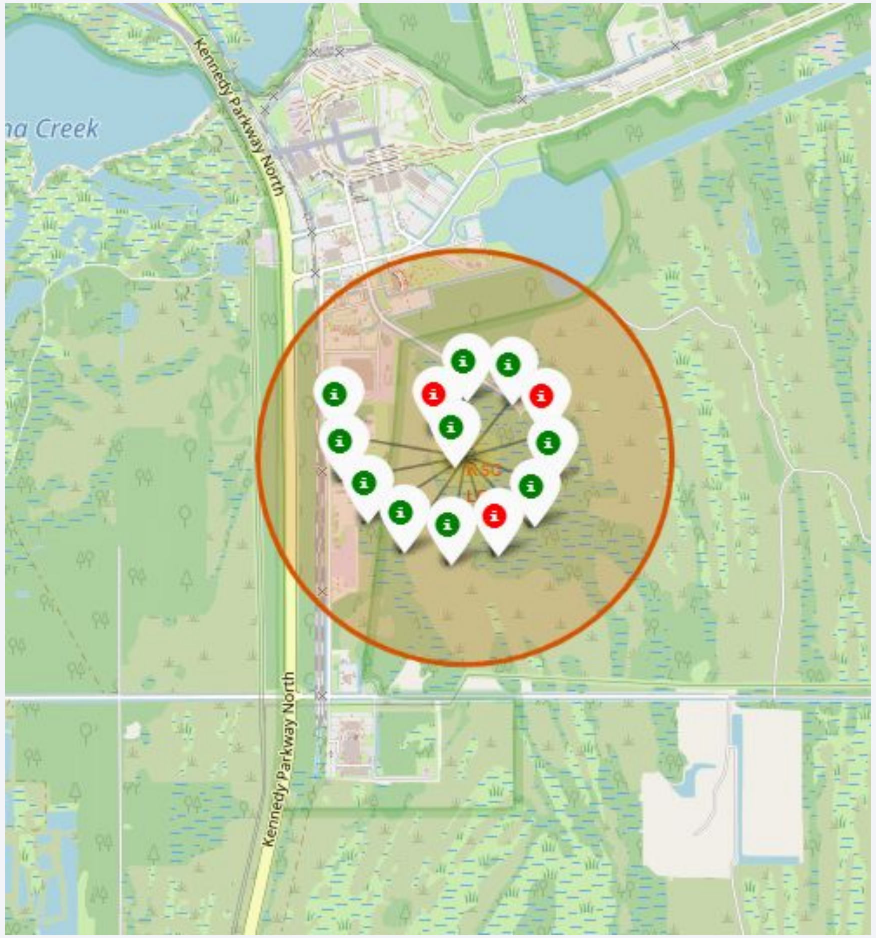
All Locations for Each Launch Sites on a Map

- The East Coast launch sites (CCAFS, KSC, SLC-40) in Florida are much closer to the Equator compared to the West Coast sites. Florida's latitude (~28.5°N) is relatively low, making it favorable for equatorial and geostationary launches.
- Launching closer to the Equator provides an additional speed boost due to the Earth's rotational velocity, which makes it more efficient for achieving geostationary orbits.
- All the marked launch sites are located very close to the coast. Launching over the ocean reduces the risk to populated areas in case of a launch failure or debris fall.



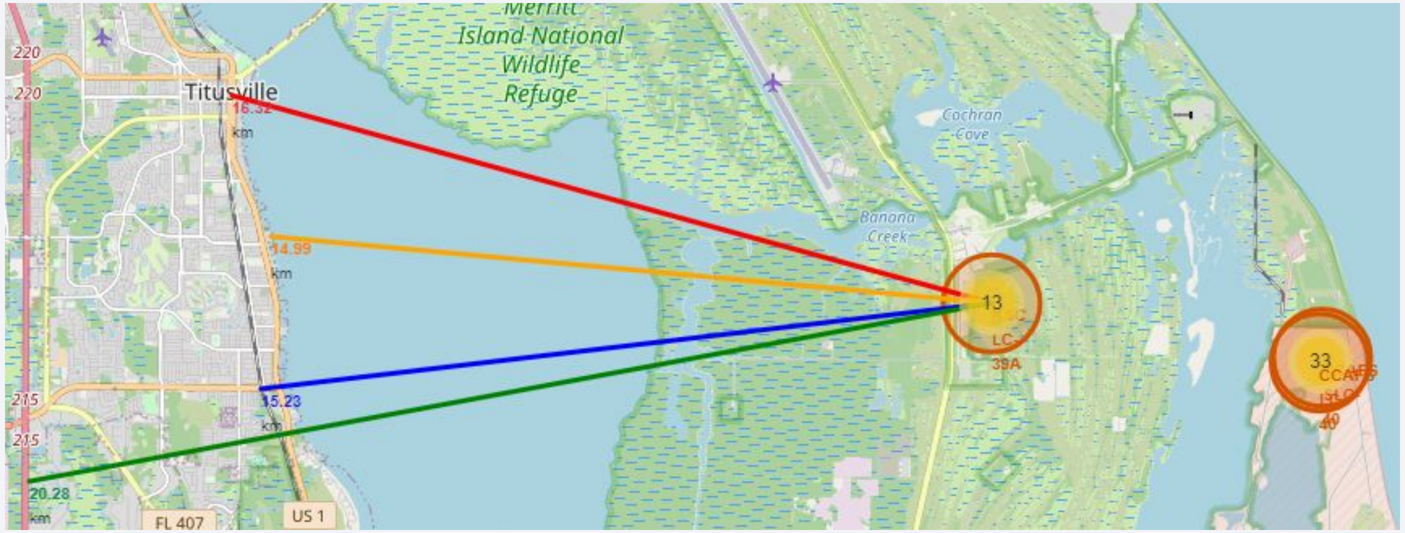
Success/Failed Launches for Each Site on the Map

- Based on the color-labeled markers, the KSC LC-39A launch site has the highest success rate.
- Green markers indicate successful launches, while red markers indicate failed launches.

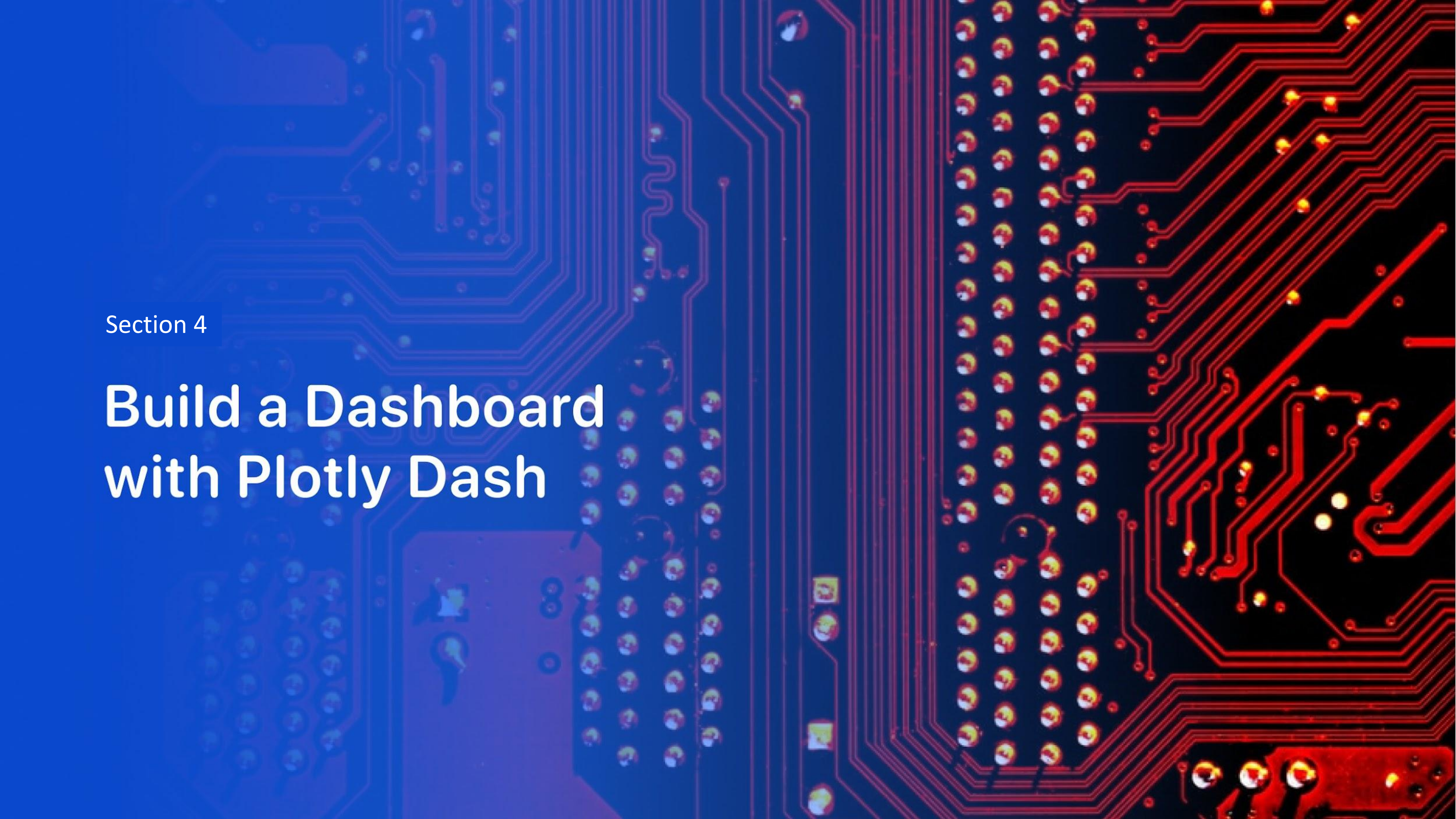


Distances Between KSC LC-39A to Its Proximities

- Red Line: Represents the distance from KSC LC-39A to Titusville (approx. 16.32 km).
- Yellow Line: Represents the distance to Banana Creek (approx. 14.99 km).
- Blue Line: Represents the distance to Cochran Cove (approx. 15.23 km).
- Green Line: Represents the distance to FL 405 Highway (approx. 20.28 km).



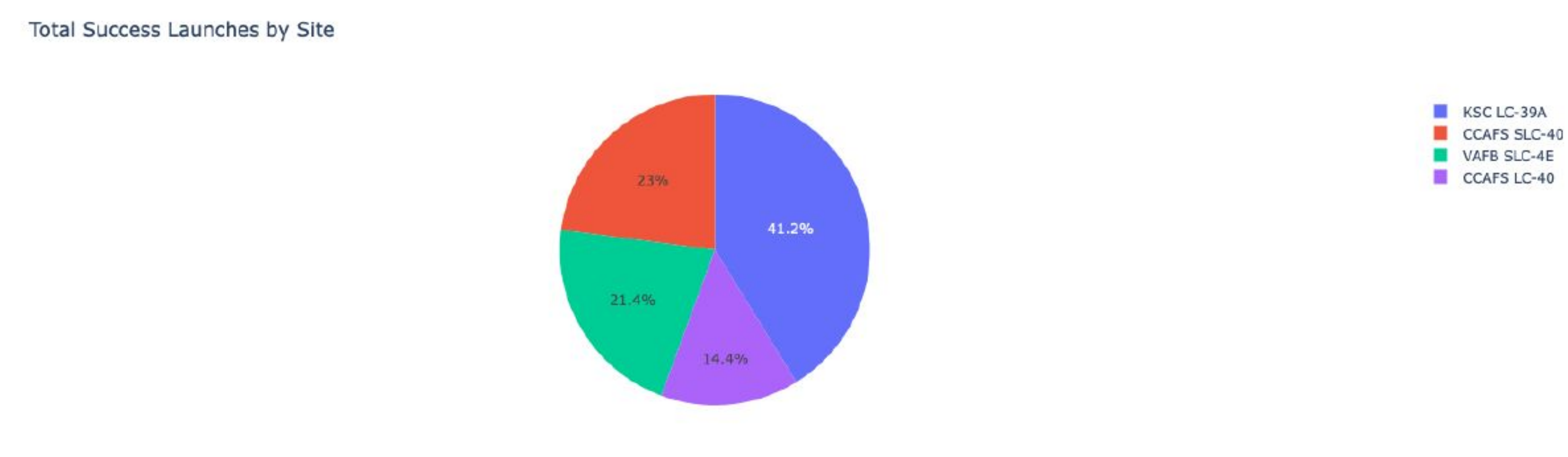
- The map visually demonstrates the geographical proximity of the KSC LC-39A launch site to nearby landmarks and infrastructure. It helps assess the accessibility and logistical considerations related to launch operations.



Section 4

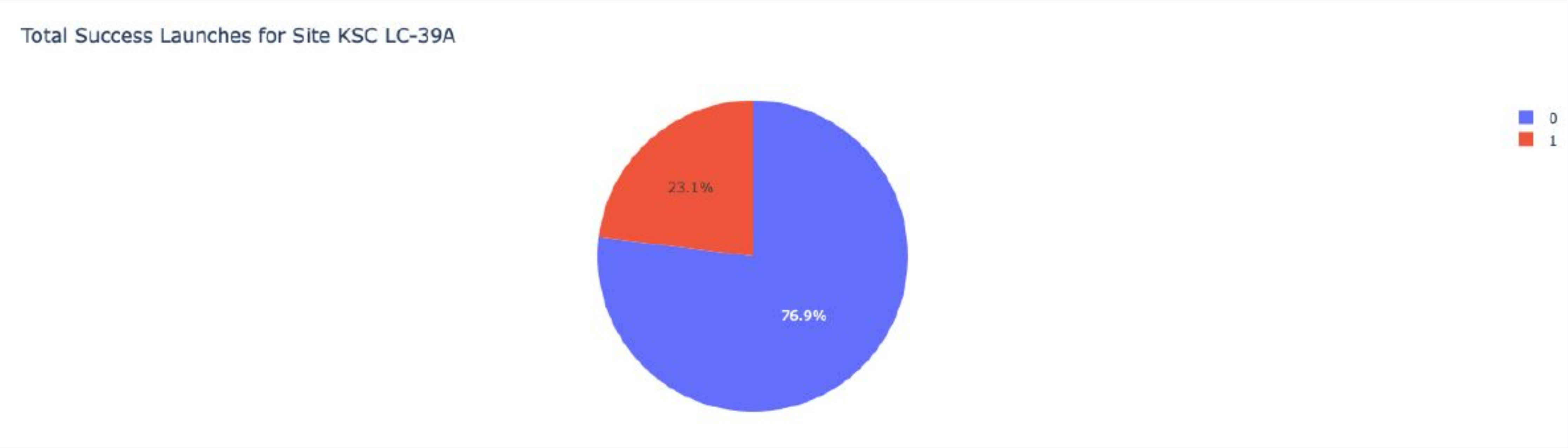
Build a Dashboard with Plotly Dash

Total Success Launches by Site



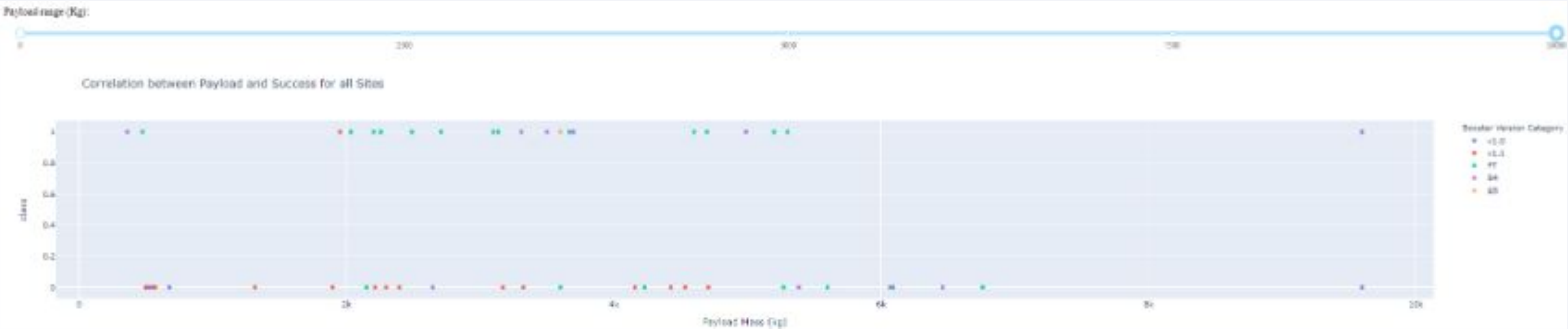
- The launch site KSC LC-39A has the highest success rate, with 41.2% of successful launches among all launch sites shown in this pie chart.

Total Success Launches for KSC LC-39A



- The launch site KSC LC-39A has the highest success rate, with 76.9%—10 successful launches out of 13—shown in this bar chart.

Payload vs. Launch Outcome for All Sites



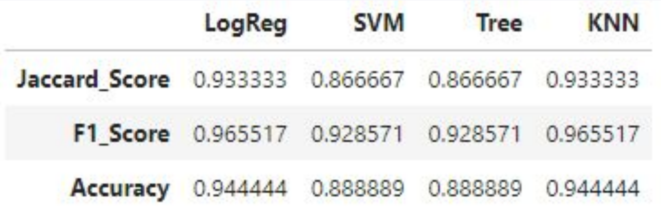
- The scatterplot shows that successful launches are more common with mid to high payload masses (around 3,000 to 6,000 kg), while failures are more frequent with lower payloads. Different booster versions exhibit varying success rates, with some being more reliable at higher payloads.

Section 5

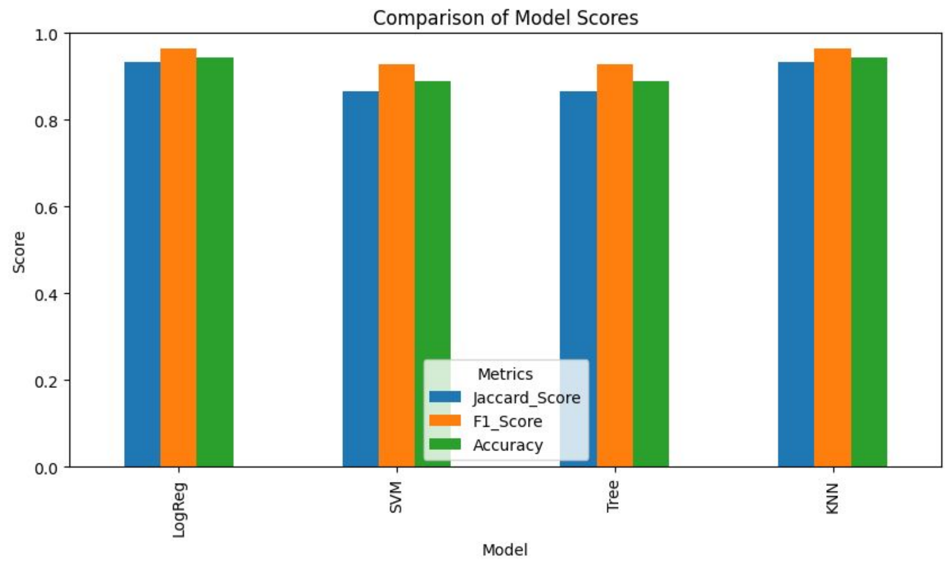
Predictive Analysis (Classification)

Classification Accuracy

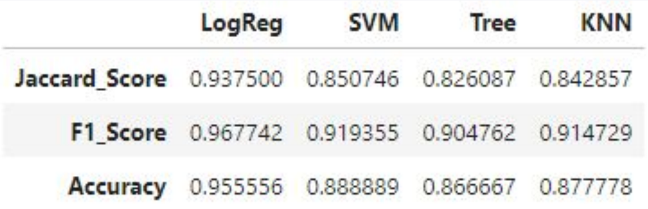
Accuracy & Scores for the Test Set



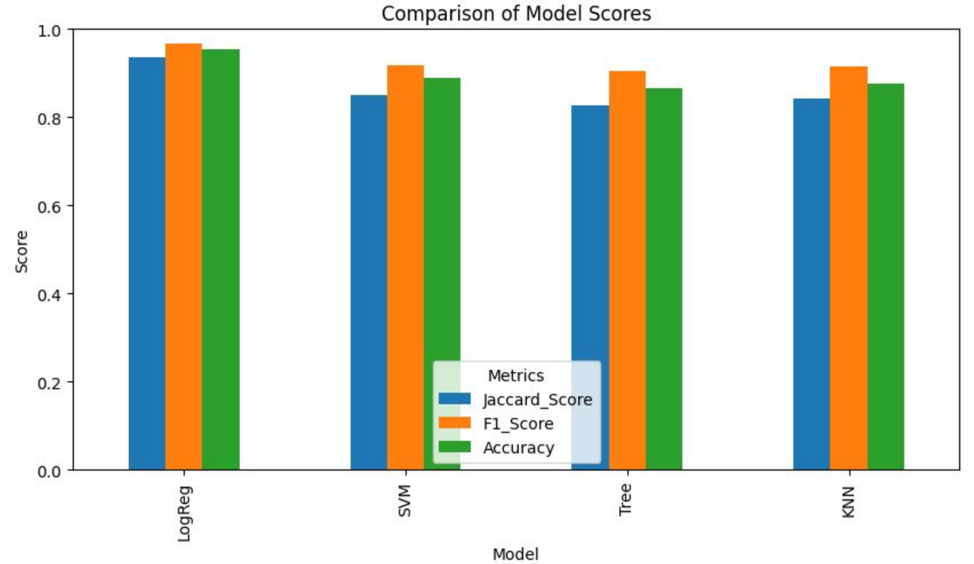
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.933333	0.866667	0.866667	0.933333
F1_Score	0.965517	0.928571	0.928571	0.965517
Accuracy	0.944444	0.888889	0.888889	0.944444



Accuracy & Scores for the Entire Dataset



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.937500	0.850746	0.826087	0.842857
F1_Score	0.967742	0.919355	0.904762	0.914729
Accuracy	0.955556	0.888889	0.866667	0.877778

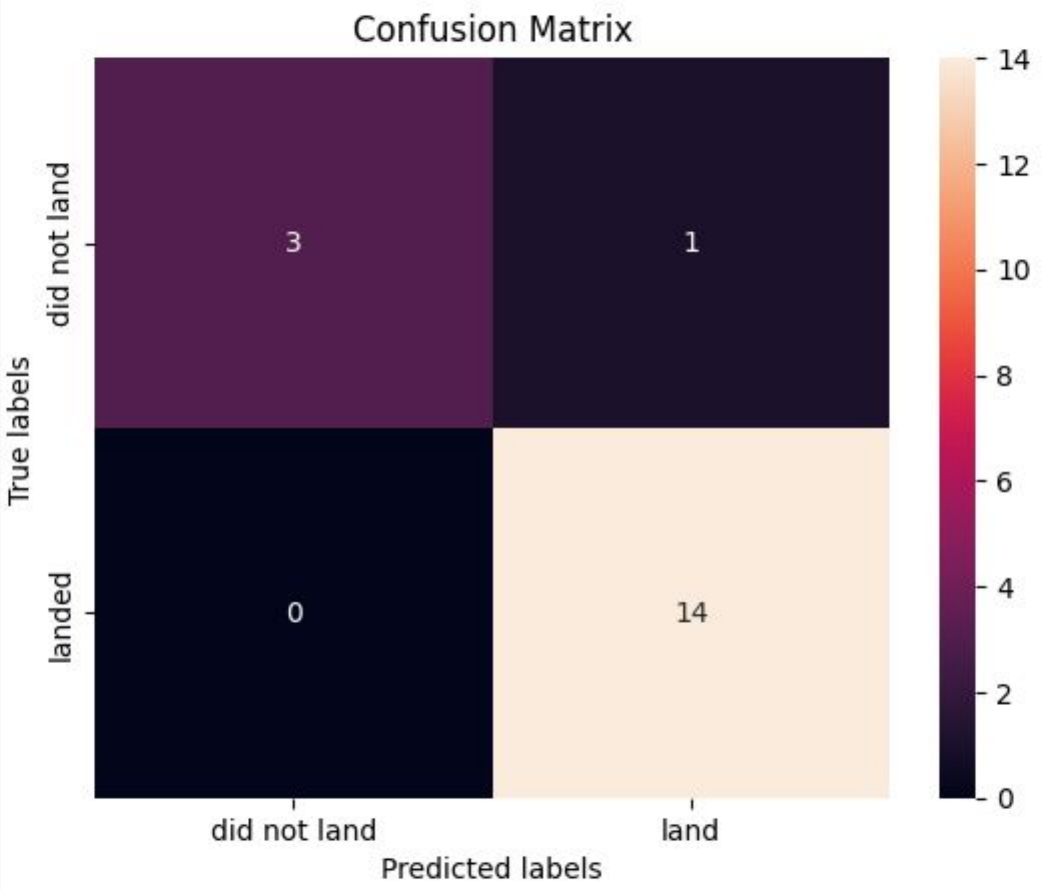


Classification Accuracy Analysis

- Based on the metrics of the test set, Logistic Regression and KNN models perform the best among all models.
- Since only 18 samples are used in the test set, we further evaluate all models using the entire dataset.
- According to the results of evaluating the entire dataset, we confirm that Logistic Regression is the best model, achieving the highest accuracy and scores.

Confusion Matrix - Logistic Regression

- Based on the confusion matrix of the Logistic Regression model, it is evident that the model can distinguish between both classes effectively, with only one false positive classification.



Conclusions

- The Logistic Regression model is the best-performing algorithm for this dataset.
- Launches with mid to high payloads have better outcomes compared to those with lower payloads.
- The launch success rate has increased over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- The orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate.

Thank you!