IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Willmann, Sebastian
28.11.2023

# Outline

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

# Executive Summary

Data was gathered from the public SpaceX API and SpaceX Wikipedia page. This information was structured with a 'class' column to denote successful landings. Utilizing SQL, visualization tools, Folium maps, and dashboards, an extensive exploration was conducted. Pertinent features were extracted for analysis, and categorical variables were converted into binary using one-hot encoding. Standardization was applied to the dataset, and GridSearchCV was employed to optimize machine learning models' parameters. The resulting accuracy scores of these models were effectively visualized, providing a comprehensive view of their performance.

Four distinct machine learning models were developed: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Each model demonstrated parallel performance, achieving an accuracy rate of approximately 83.33%. Notably, these models consistently exhibited a tendency to overpredict successful landings. To enhance model precision and efficacy, acquiring additional data is imperative for improved determination and accuracy.

# Introduction

SpaceX is the most successful company of the commercial space age, making space travel more affordable. The company advertises Falcon 9 rocket launches on its website for a cost of $62 million. Other providers cost up to $165 million per launch, with a large portion of the savings coming from SpaceX's ability to reuse the first stage. However, if we can determine whether the first stage will land, we can determine the cost of a launch.

***Questions to be Answered***

•How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

•Does the rate of successful landings increase over time?

•What algorithm is best suited for binary classification in this case?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

# Data Collection

Our data collection process was a fusion of SpaceX REST API queries and web scraping techniques from SpaceX's Wikipedia entry. This hybrid approach was necessary to ensure a comprehensive dataset for our analysis, as each method provided unique and valuable information.

The SpaceX REST API granted us access to critical columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

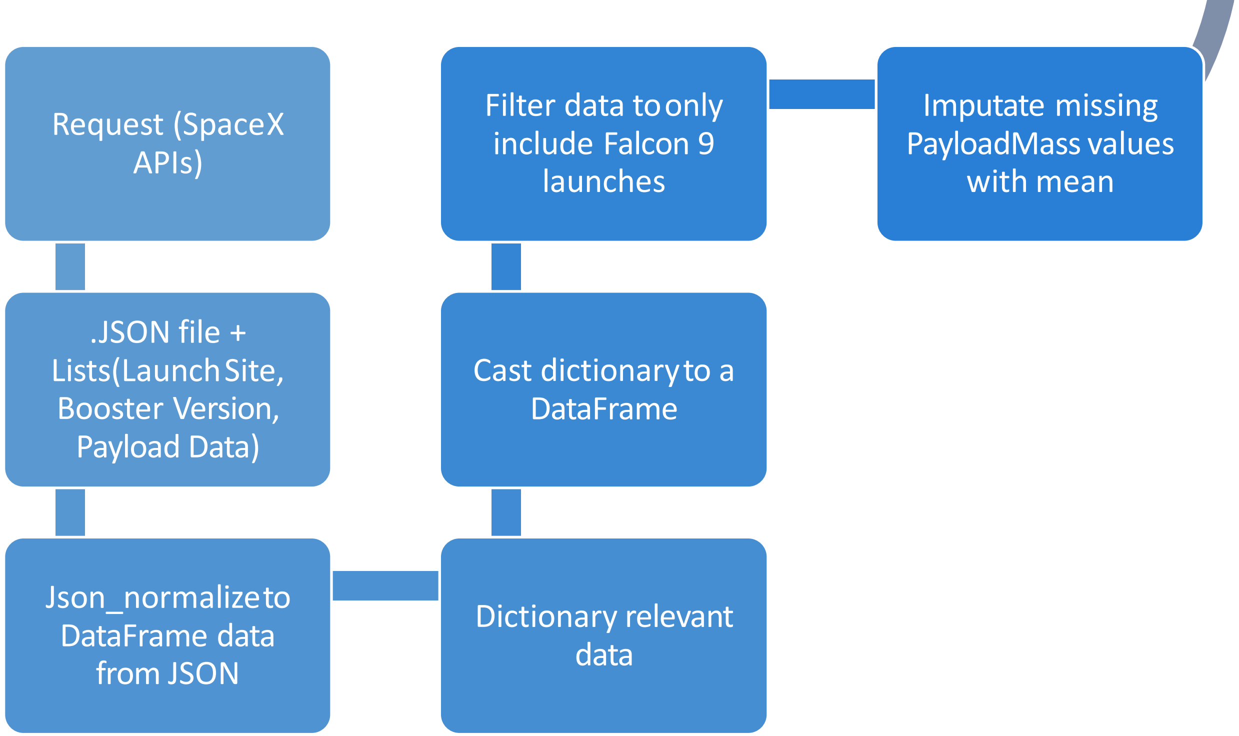
Simultaneously, web scraping from Wikipedia supplied us with additional columns, including Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time. This combined dataset enabled a more detailed and comprehensive analysis of rocket launches for predictive modeling.

# Data Collection – SpaceX API

**Data Collection**

**SpaceX API**

https://github.com/SWillm/DataScienceCaspstone/blob/main/1%20-%20Introduction/Spacex%20Data%20Collection%20API.ipynb



Request (SpaceX APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Dictionary relevant data

Cast dictionary to a DataFrame

Filter data to only include Falcon 9 launches

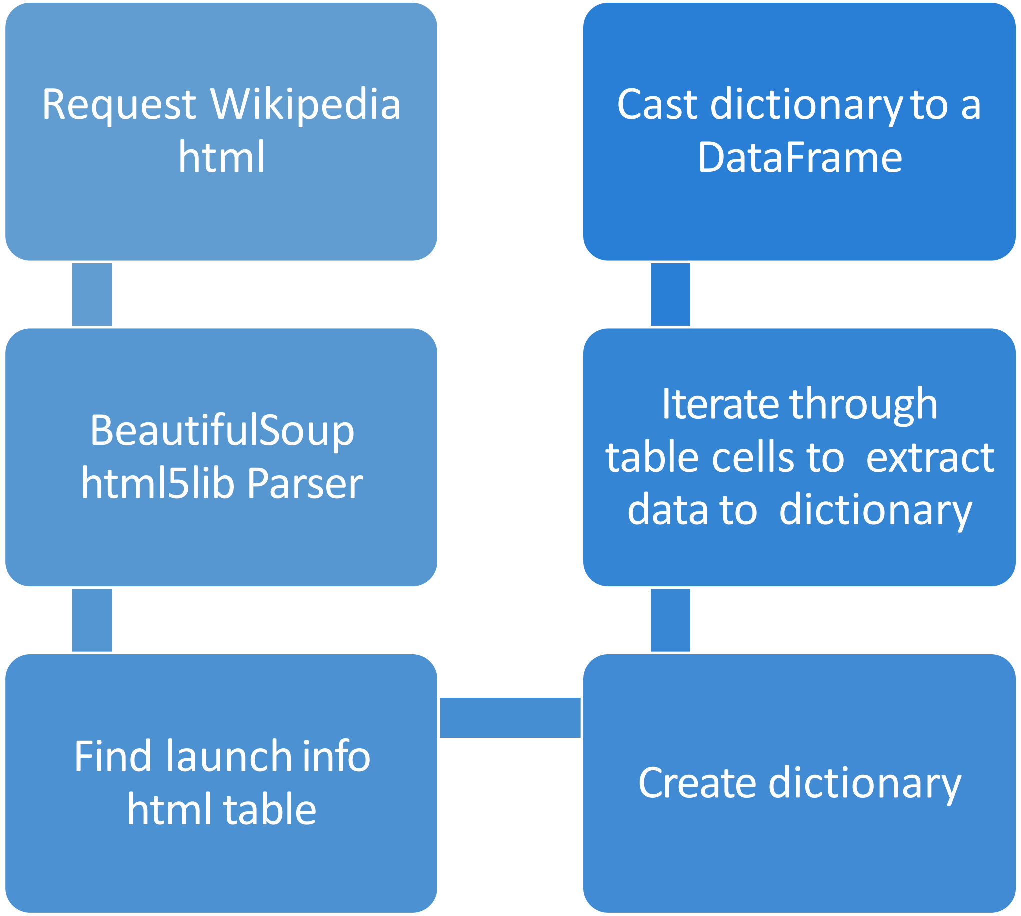
Imputate missing PayloadMass values with mean

# Data Collection – SpaceX Webscraping

**Data Collection**

**Web Scraping**

https://github.com/SWillm/DataScienceCaspstone/blob/main/1%20-%20Introduction/Spacex%20Data%20Collection%20Webscraping.ipynb



Request Wikipedia html

BeautifulSoup html5lib Parser

Find launch info html table

Create dictionary

Iterate through table cells to extract data to dictionary

Cast dictionary to a DataFrame

# Data Wrangling

In the dataset, various scenarios emerged where the booster failed to achieve successful landings. Different outcomes were observed, such as 'True Ocean' signifying a successful landing in a specific oceanic region, and 'False Ocean' indicating an unsuccessful ocean landing. 'True RTLS' and 'False RTLS' denoted successful and unsuccessful ground pad landings, respectively. Similarly, 'True ASDS' and 'False ASDS' represented successful and unsuccessful landings on a drone ship.

These outcomes were standardized into training labels for analysis. A label of '1' was assigned to successful booster landings, while '0' denoted unsuccessful landings. This classification framework provided clear distinctions for our predictive modeling.

1 .Perform exploratory Data Analysis and determine Training Labels

2. Calculate the number of launches on each site

3. Calculate the number and occurrence of each orbit

4. Calculate the number and occurrence of mission outcome per orbit type

5. Create a landing outcome label from Outcome column

6. Exporting the data to CSV

https://github.com/SWillm/DataScienceCaspstone/blob/main/1%20-%20Introduction/Spacex%20Data%20Wrangling.ipynb

# EDA with Data Visualization

Scatter plots illustrated relationships between Flight Number and Payload Mass, Flight Number and Launch Site, Payload Mass and Launch Site, Orbit Type and Success Rate, Flight Number and Orbit Type, Payload Mass and Orbit Type, and the yearly trend of Success Rate. These scatter plots identify potential correlations between variables that could be beneficial for machine learning models.

Bar charts were employed to compare discrete categories, providing insights into connections between specific categories and their corresponding measured values.

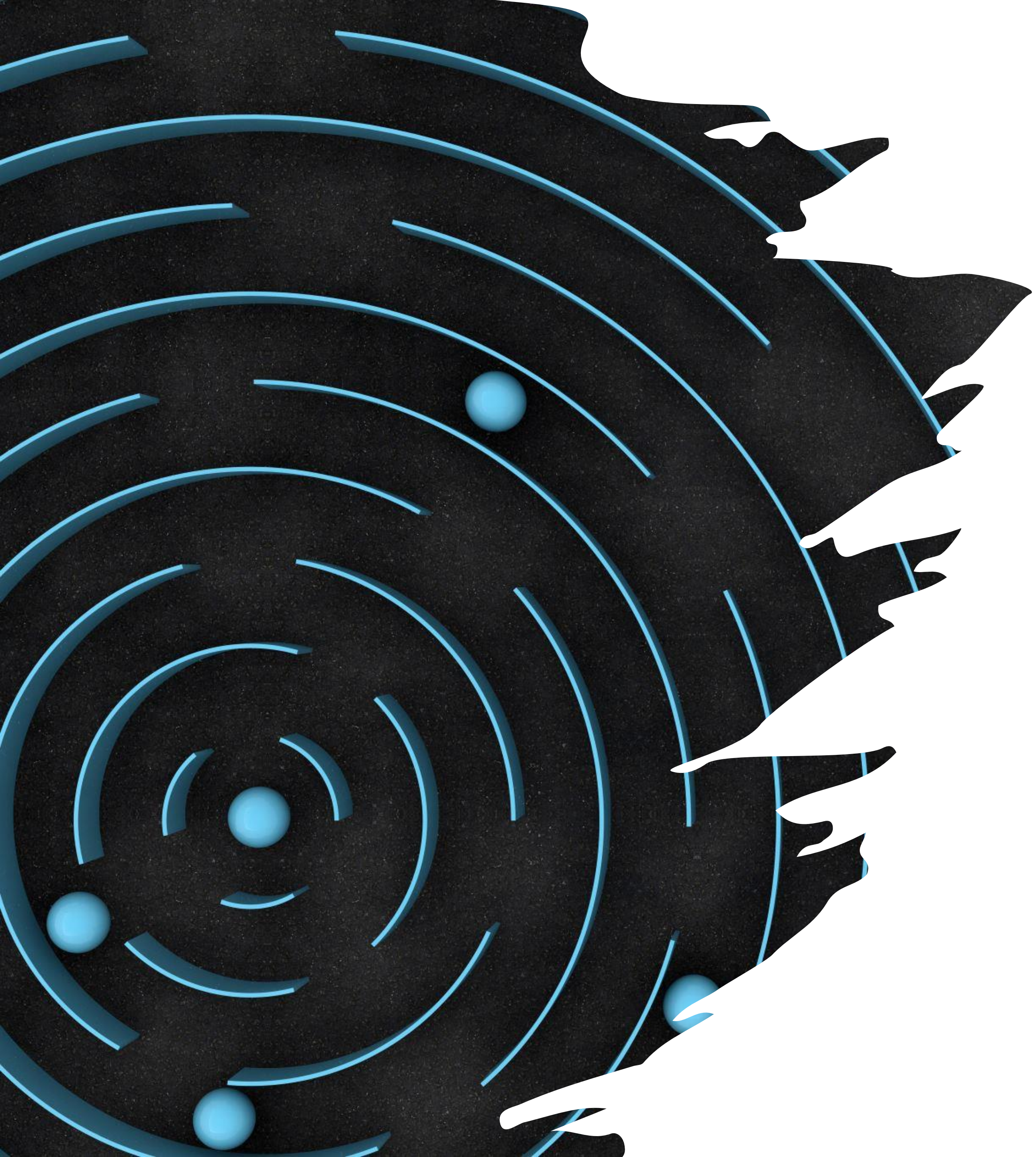
Line charts were utilized to visualize data trends over time, particularly useful for observing time series patterns.

https://github.com/SWillm/DataScienceCaspstone/blob/main/2%20-%20Exploratory%20Data%20Analysis/EDA%20Data%20Visualisation.ipynb

# EDA with SQL

- Retrieving unique launch site names within the space missions.
- Displaying 5 records where launch sites start with the prefix ‘CCA’.
- Calculating the total payload mass carried by boosters launched by NASA (CRS).
- Determining the average payload mass carried by the booster version F9 v1.1.
- Identifying the date of the first successful landing outcome achieved on a ground pad.
- Listing the names of boosters with successful drone ship landings and a payload mass between 4000 and 6000.
- Summarizing the total count of successful and failed mission outcomes.
- Identifying the booster versions that carried the maximum payload mass.
- Listing failed landing outcomes on drone ships, including their booster versions and launch site names, specifically for the months in the year 2015.
- Ranking the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

https://github.com/SWillm/DataScienceCaspstone/blob/main/2%20-%20Exploratory%20Data%20Analysis/Eda%20SQL.ipynb



# Build an Interactive Map with Folium

For the NASA Johnson Space Center, a marker with a Circle, Popup Label, and Text Label was added, utilizing its latitude and longitude coordinates as the start location.

Markers featuring Circles, Popup Labels, and Text Labels for all Launch Sites were included, showcasing their geographical positions in relation to the Equator and coastlines.

Additionally, colored Markers were implemented to denote launch outcomes: green for success and red for failures, using a Marker Cluster to discern Launch Sites with notably high success rates.

Furthermore, colored Lines were integrated to visualize distances between Launch Site KSC LC-39A (as an example) and its surrounding areas such as the Railway, Highway, Coastline, and Closest City.

https://github.com/SWillm/DataScienceCaspstone/blob/main/3%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/Launch%20Site%20Location%20Visualisation.ipynb

# Build a Dashboard with Plotly Dash

- Dropdown List for Launch Sites:
- Implemented a dropdown menu to facilitate Launch Site selection.

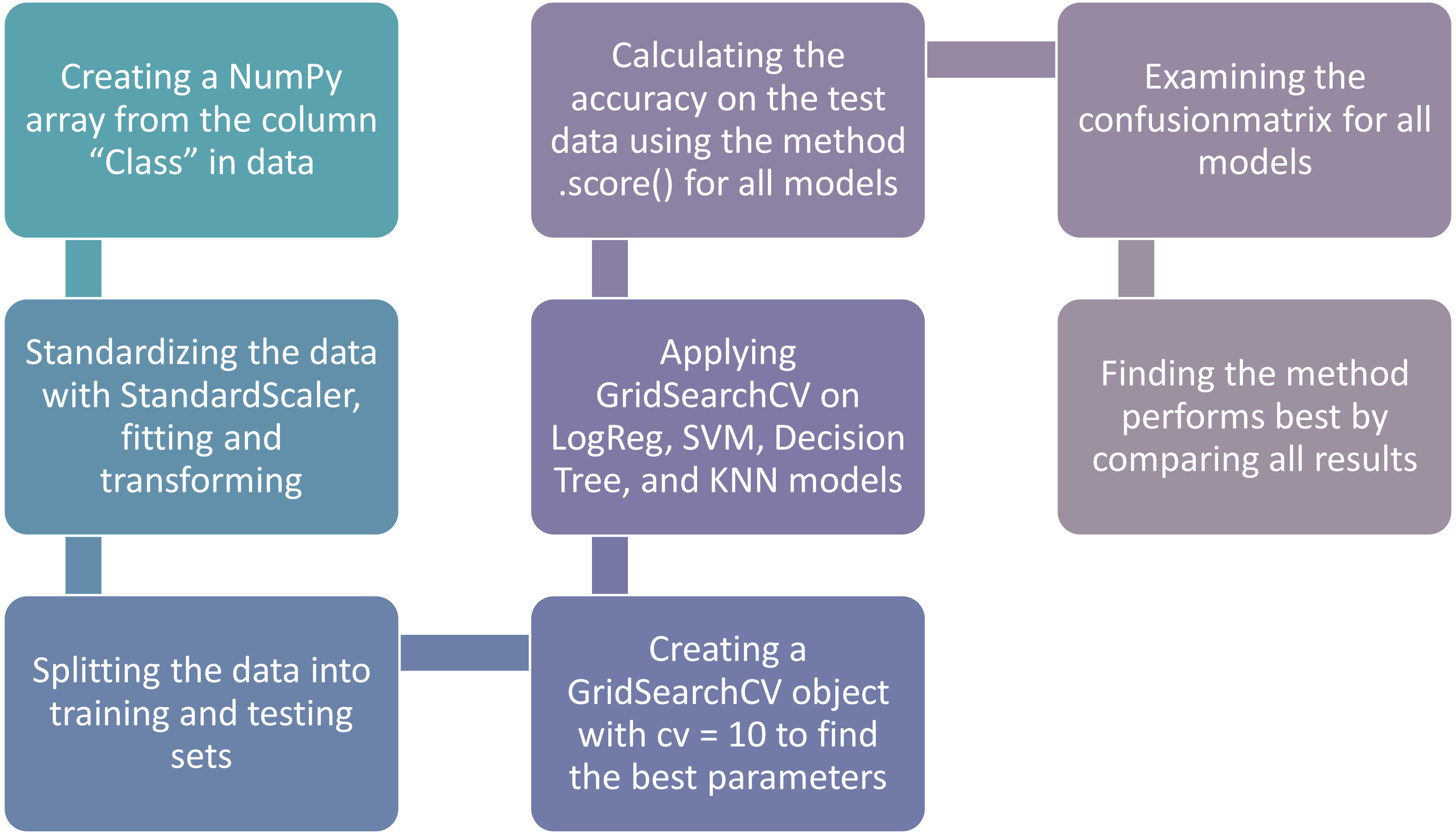
- Pie Chart illustrating Successful Launches:
- Introduced a pie chart showcasing the total count of successful launches across all sites. If a specific Launch Site was chosen, the chart displayed the Success versus Failed counts for that site.

- Payload Mass Range Slider:
- Included a slider enabling the selection of Payload ranges.

- Scatter Chart depicting Payload Mass vs. Success Rate:
- Integrated a scatter chart to demonstrate the relationship between Payload and Launch Success across various Booster Versions.

https://github.com/SWillm/DataScienceCaspstone/blob/main/3%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/Spacex_Dash_App.py

# Predictive Analysis (Classification)



Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, fitting and transforming

Splitting the data into training and testing sets

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

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusionmatrix for all models

Finding the method performs best by comparing all results

# Results

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

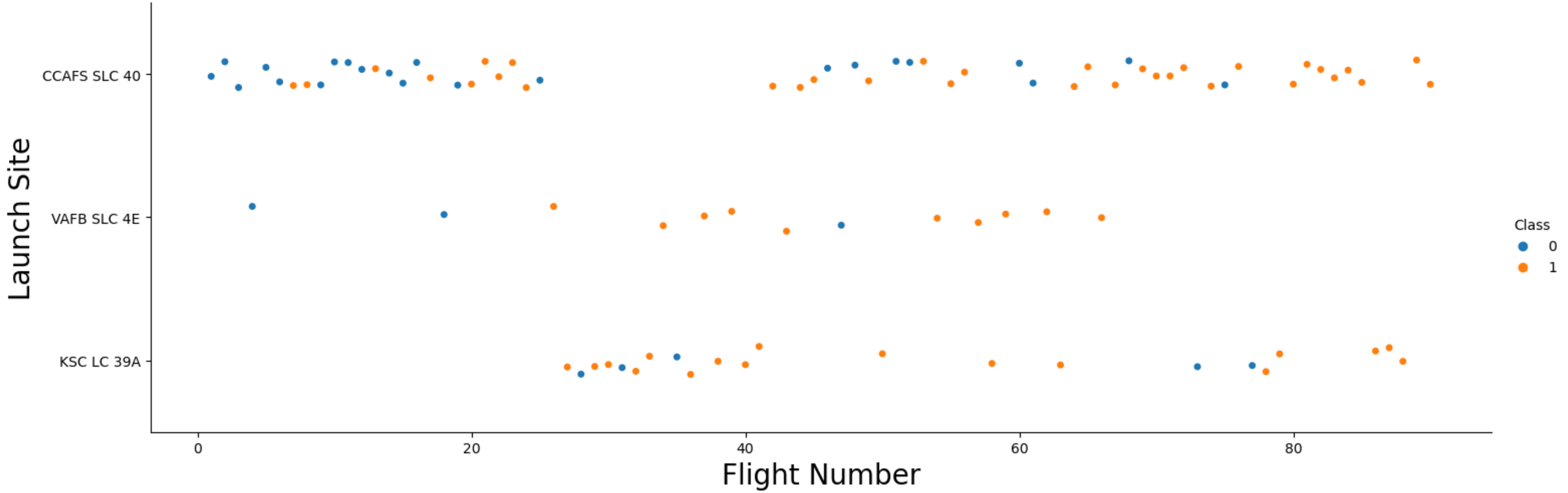
PREDICTIVE ANALYSIS RESULTS

Section 2

# Insights drawn from EDA

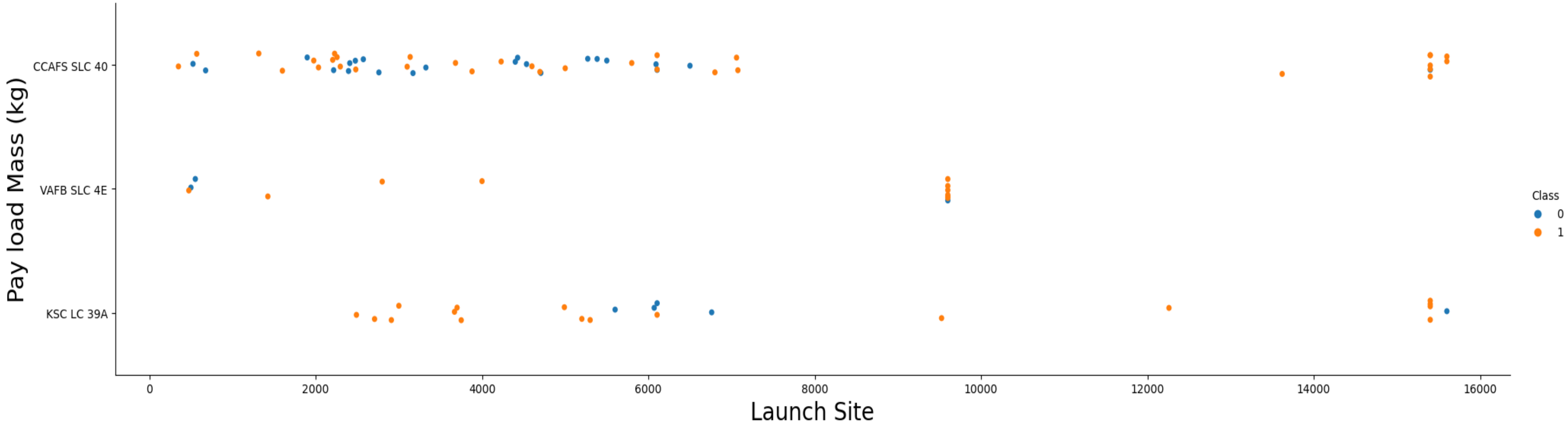
# Flight Number vs. Launch Site

- Early flights experienced failures, contrasting with consistent success in the latest launches.
- Nearly half of all launches originated from the CCAFS SLC 40 site.
- Higher success rates are evident at VAFB SLC 4E and KSC LC 39A launch sites.
- There appears to be a pattern suggesting an increasing success rate with each successive launch, indicating a potential trend toward improved success rates over time.



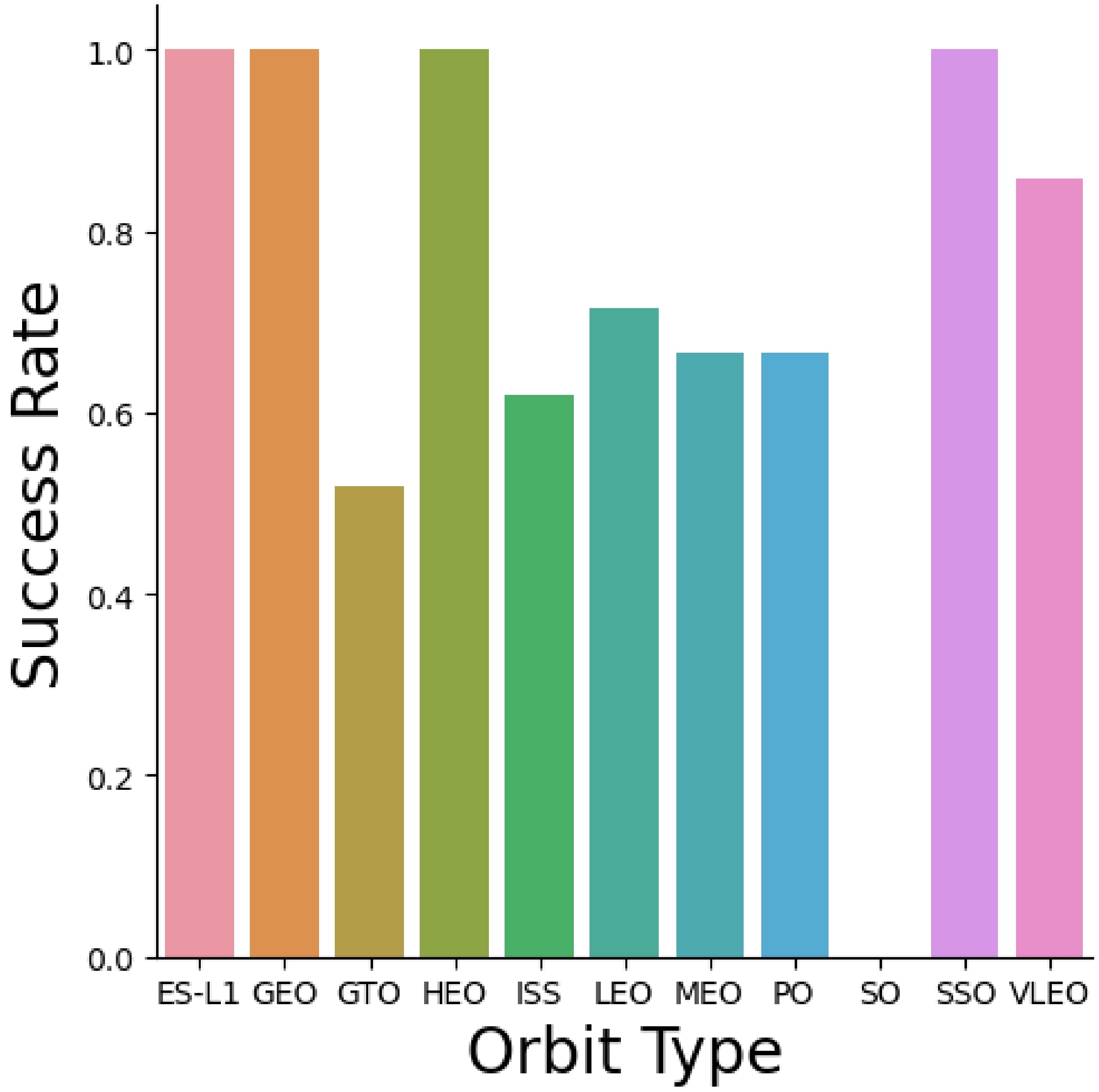
# Payload vs. Launch Site

- There's a direct relationship observed between payload mass and success rate across every launch site.
- Launches with payload masses surpassing 7000 kg predominantly achieved success.
- Different launch sites also seem to use different payload mass.



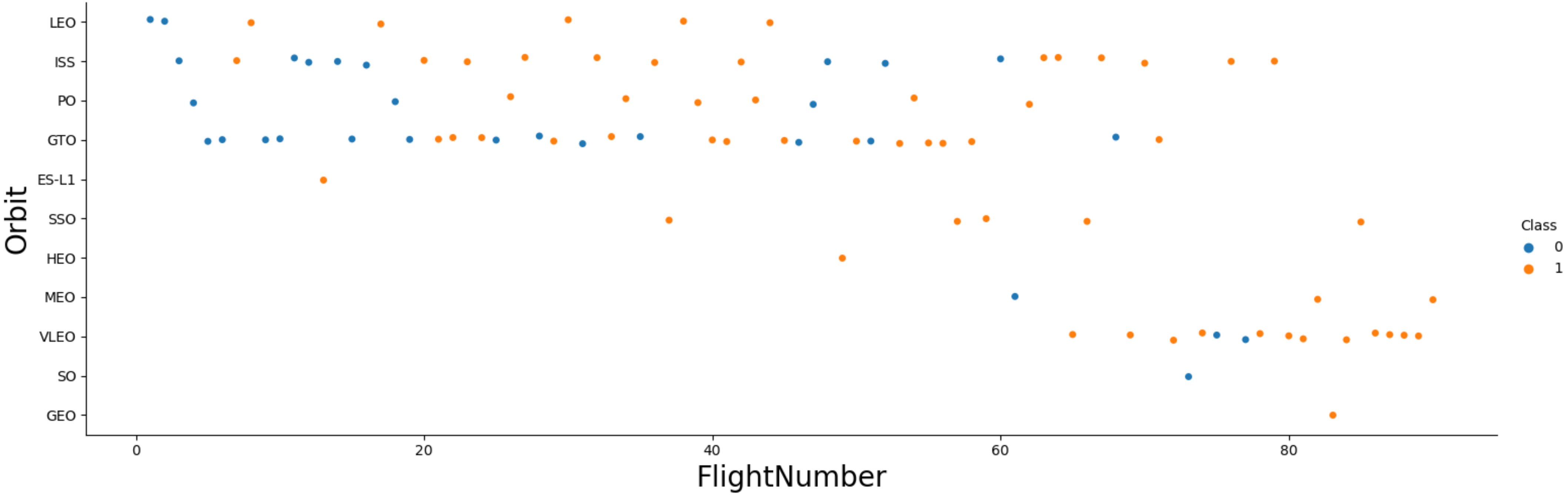
# Success Rate vs. Orbit Type

- Orbits with a 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with a 0% success rate: SO
- Orbits with success rates between 50% and 85%: GTO, ISS, LEO, MEO, PO



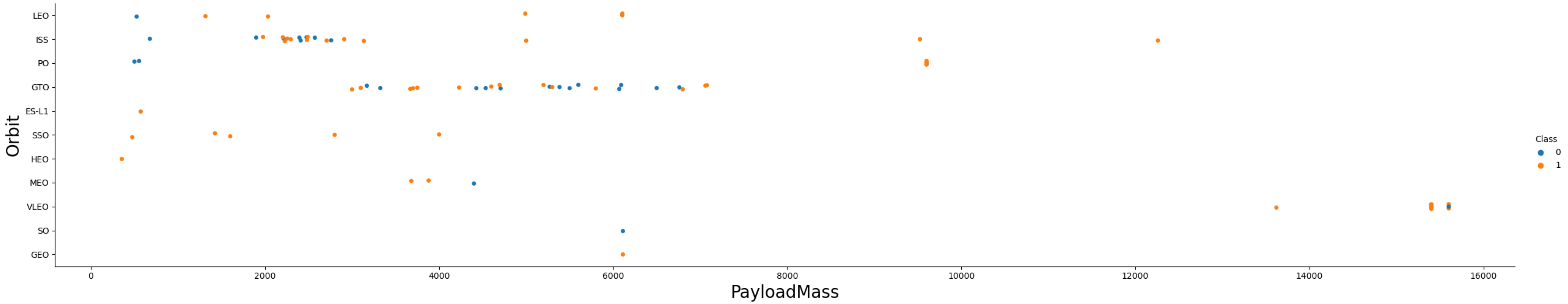
# Flight Number vs. Orbit Type

- For the LEO orbit, success appears to correlate with the number of flights conducted. However, in the GTO orbit, there seems to be no discernible relationship between flight numbers and success rates.



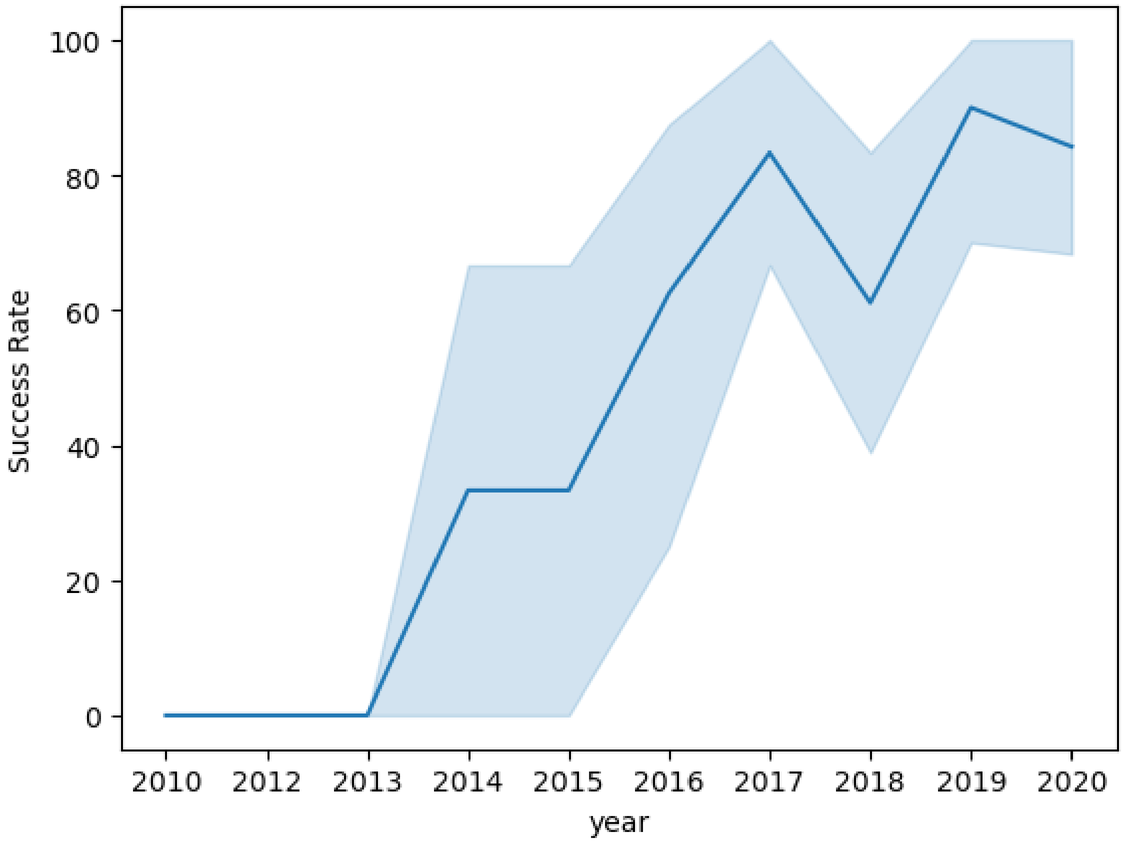
# Payload vs. Orbit Type

- Heavy payloads exhibit a negative impact on GTO orbits while displaying a positive influence on GTO and Polar LEO (ISS) orbits.



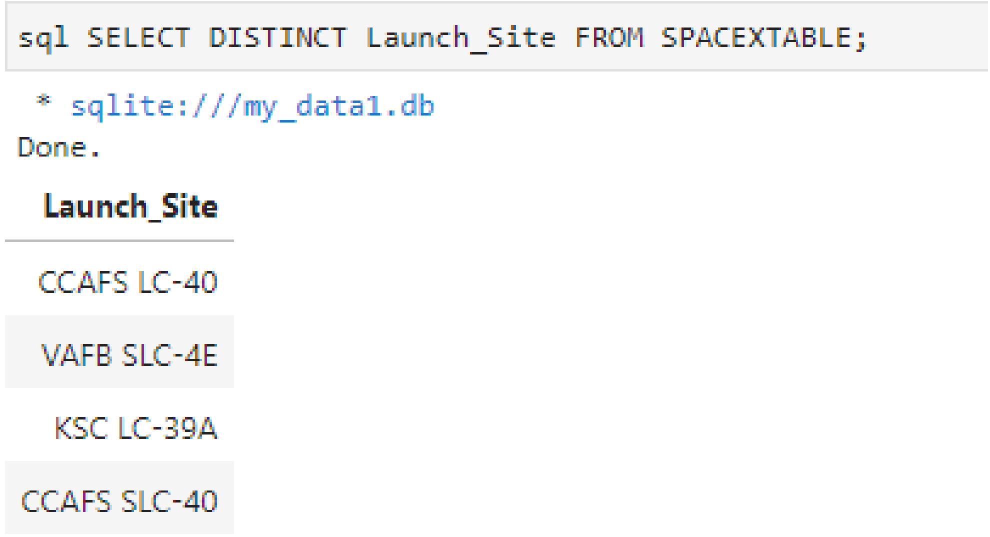
# Launch Success Yearly Trend

- The trend in success rates has been notably ascending, progressively increasing from the year 2013 up until 2020, showcasing a continuous and steady upward trajectory.



# All Launch Site Names

- Showing the distinct names of the launch sites involved in the space missions.



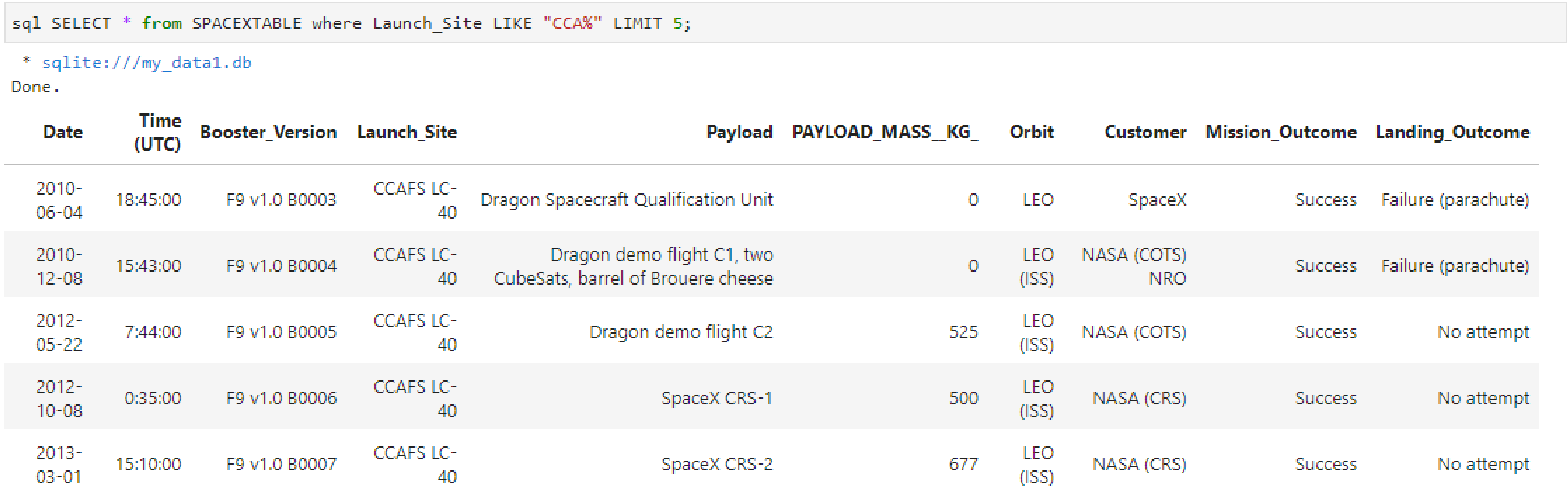
```
sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

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

# Launch Site Names Begin with 'CCA'

Showing the distinct names of the launch sites involved in the space missions.



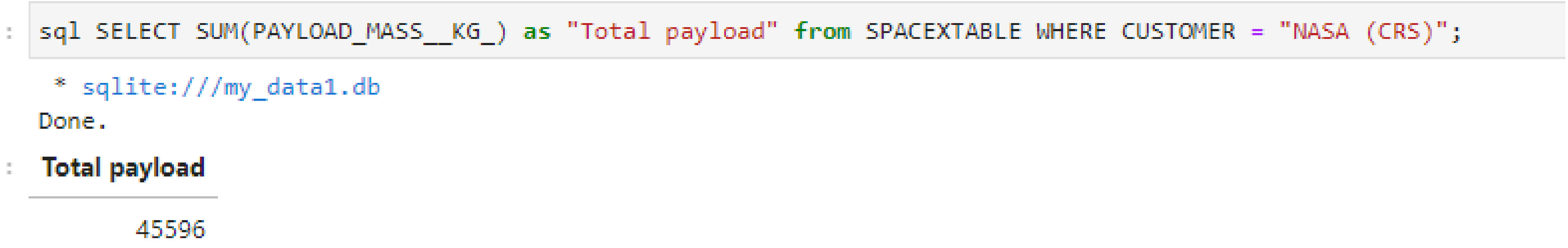
```
sql SELECT * from SPACEXTABLE where Launch_Site LIKE "CCA%" LIMIT 5;
```

```
 * 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 |

# Total Payload Mass

Presenting the cumulative payload mass transported by NASA-launched boosters (CRS).



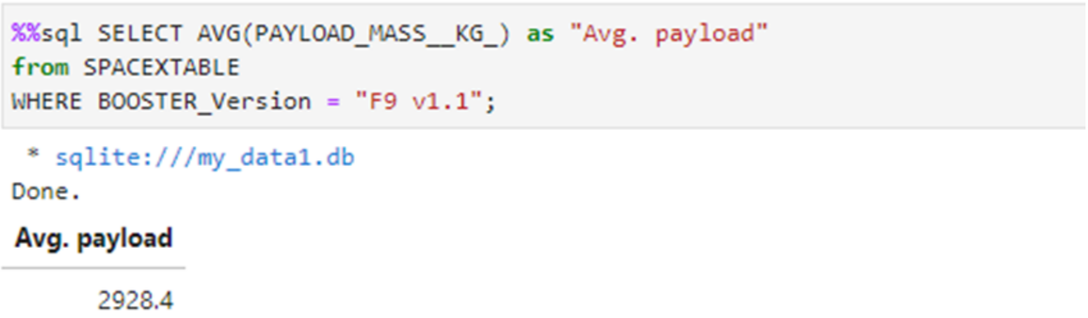
```
sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total payload" from SPACEXTABLE WHERE CUSTOMER = "NASA (CRS)";
```

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

| Total payload |
|---|
| 45596 |

# Average Payload Mass by F9 v1.1

Demonstrating the mean payload mass transported by the booster version F9 v1.1.



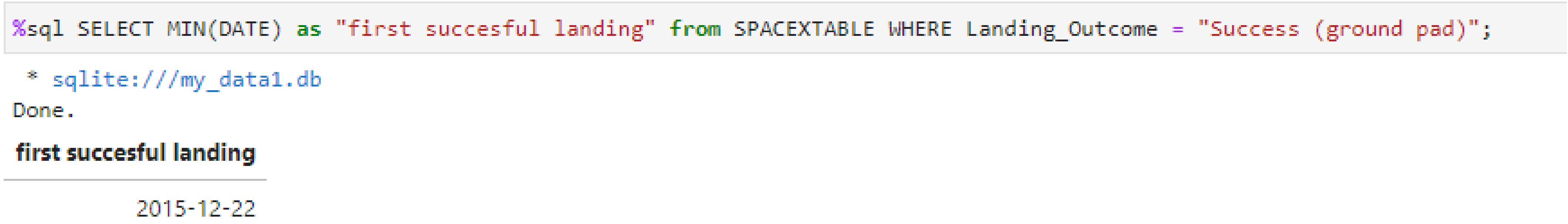
```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Avg. payload"
from SPACEXTABLE
WHERE BOOSTER_Version = "F9 v1.1";
```

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

| Avg. payload |
|---|
| 2928.4 |

# First Successful Ground Landing Date

Demonstrating the mean payload mass transported by the booster version F9 v1.1.



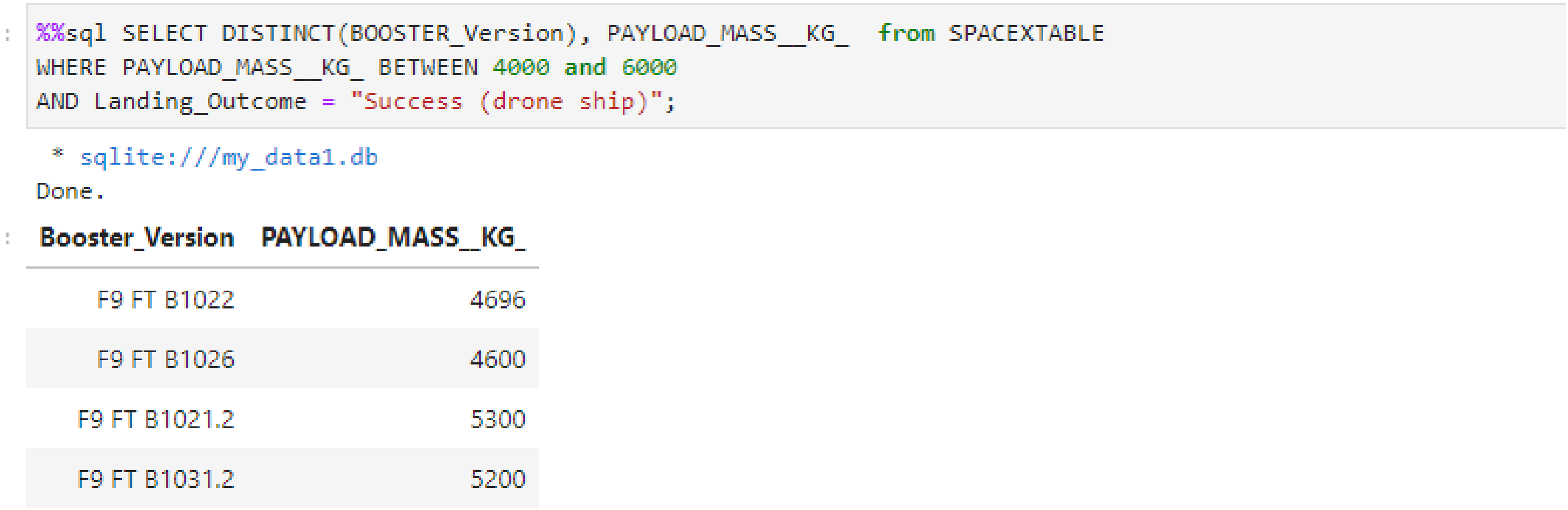
```
%sql SELECT MIN(DATE) as "first succesful landing" from SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
```

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

| first succesful landing |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

Enumerating the names of boosters that achieved success on a drone ship and carried a payload mass greater than 4000 but less than 6000.



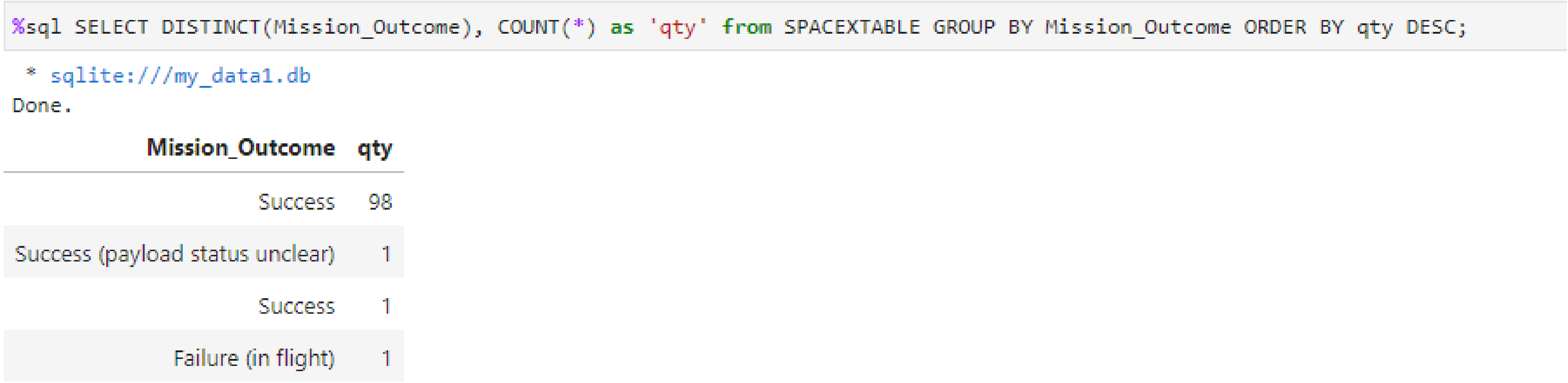
```
%%sql SELECT DISTINCT(BOOSTER_Version), PAYLOAD_MASS__KG_  from SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
AND Landing_Outcome = "Success (drone ship)";
```

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

| Booster_Version | PAYLOAD_MASS__KG_ |
|---:|---:|
| F9 FT B1022 | 4696 |
| F9 FT B1026 | 4600 |
| F9 FT B1021.2 | 5300 |
| F9 FT B1031.2 | 5200 |

# Total Number of Successful and Failure Mission Outcomes

Compiling the total count of successful and failed mission outcomes.



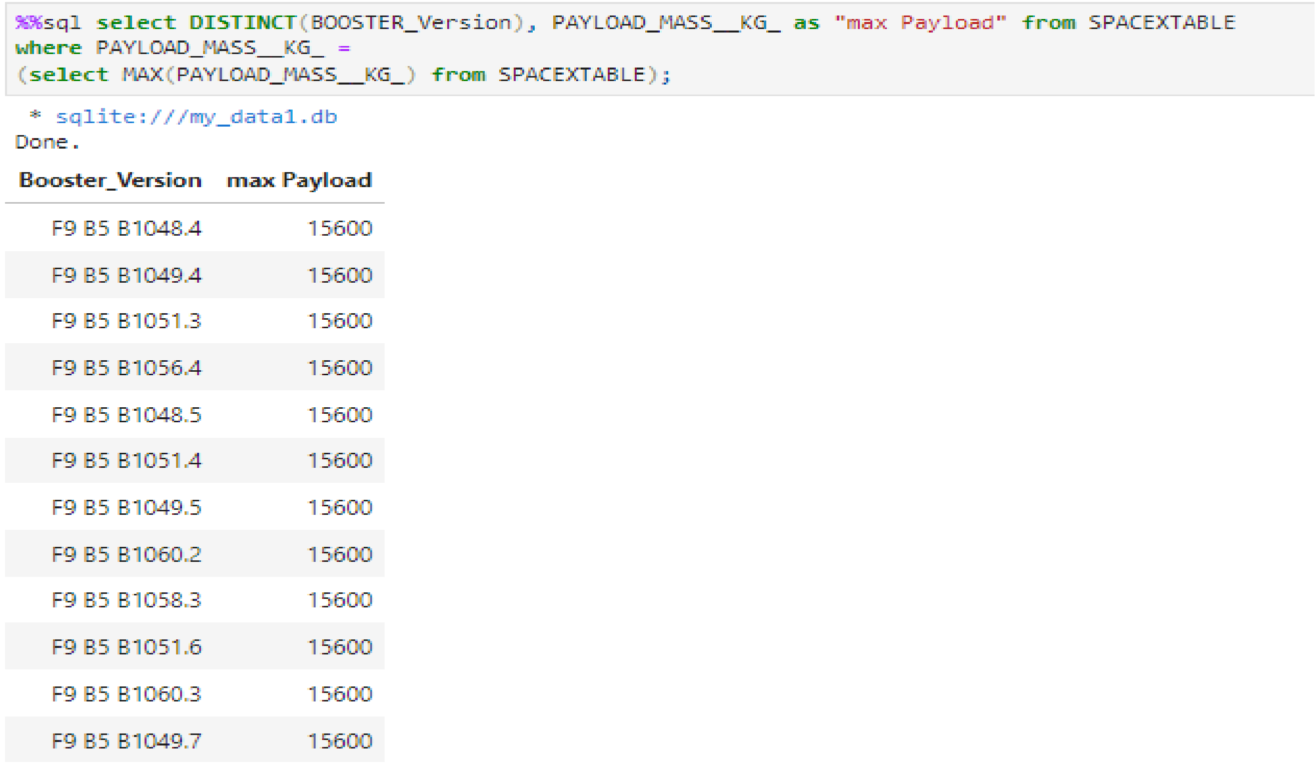
```
%sql SELECT DISTINCT(Mission_Outcome), COUNT(*) as 'qty' from SPACEXTABLE GROUP BY Mission_Outcome ORDER BY qty DESC;
```

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

| Mission_Outcome | qty |
|---:|---:|
| Success | 98 |
| Success (payload status unclear) | 1 |
| Success | 1 |
| Failure (in flight) | 1 |

# Boosters Carried Maximum Payload

Listing the names of the booster versions which have carried the maximum payload mass



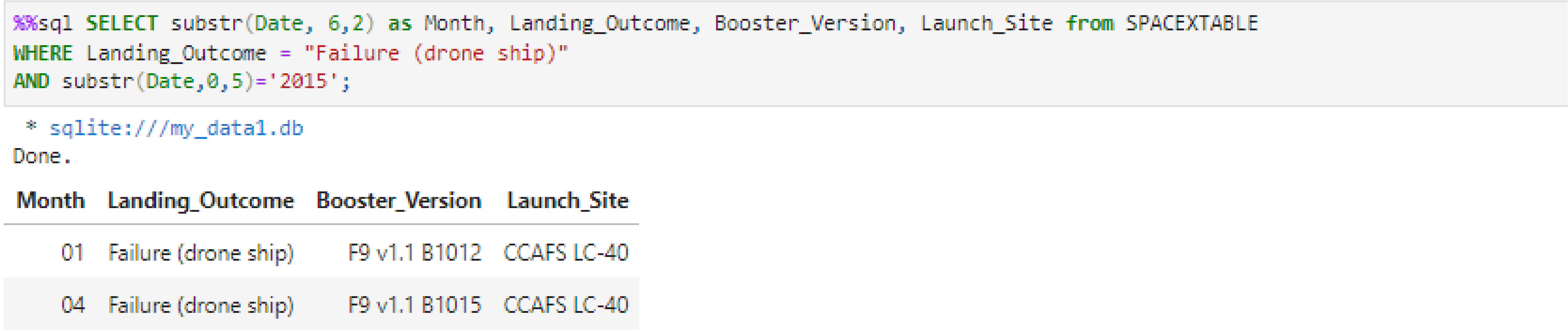
```
%%sql select DISTINCT(BOOSTER_Version), PAYLOAD_MASS__KG_ as "max Payload" from SPACEXTABLE
where PAYLOAD_MASS__KG_ =
(select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

| Booster_Version | max Payload |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1060.3 | 15600 |
| F9 B5 B1049.7 | 15600 |

# 2015 Launch Records

Enumerating the failed landing outcomes on drone ships, along with their respective booster versions and launch site names specifically for the months within the year 2015.



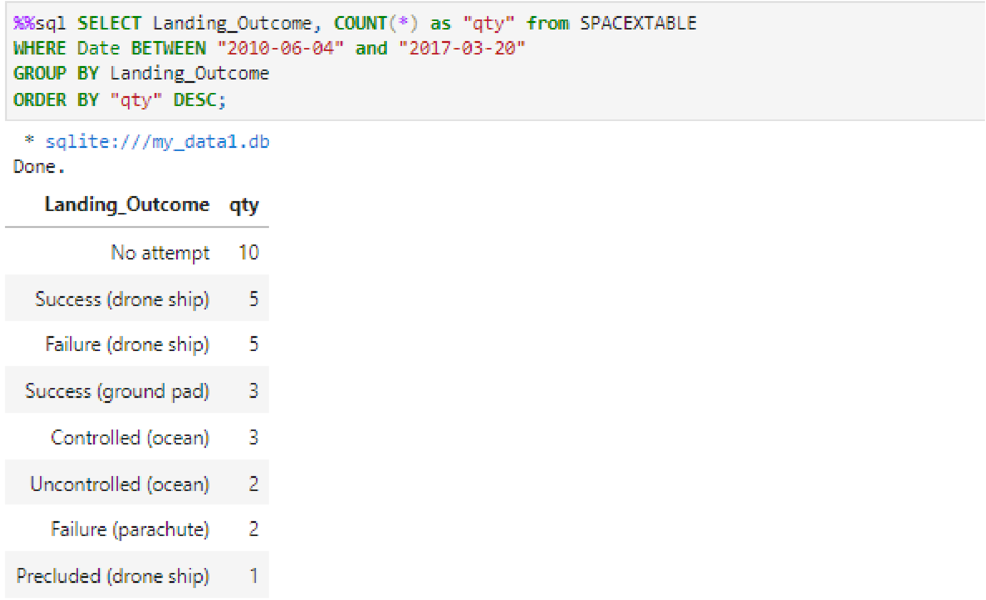
```
%%sql SELECT substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE
WHERE Landing_Outcome = "Failure (drone ship)"
AND substr(Date,0,5)='2015';
```

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

| 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

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



```
%%sql SELECT Landing_Outcome, COUNT(*) as "qty" from SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome
ORDER BY "qty" DESC;
```

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

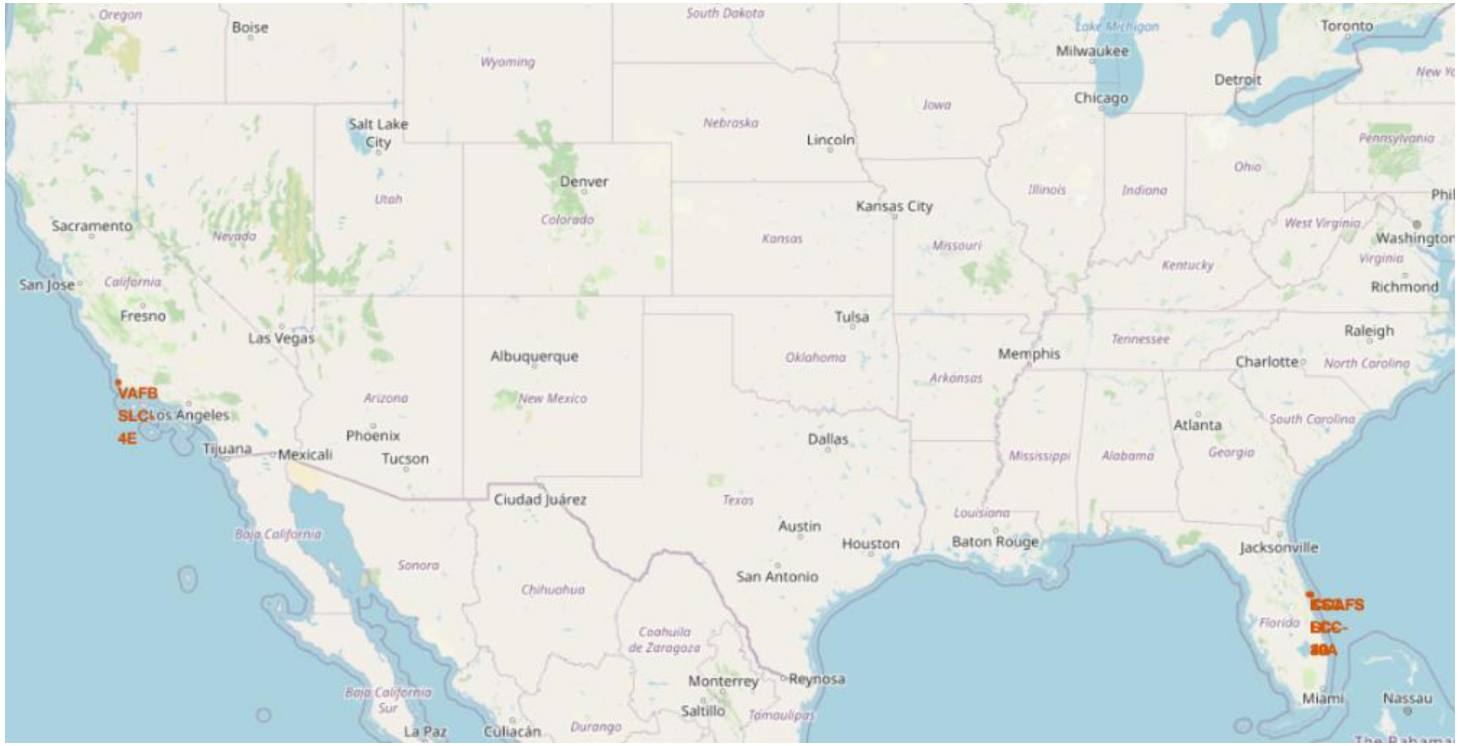
| Landing_Outcome | qty |
|---:|---:|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

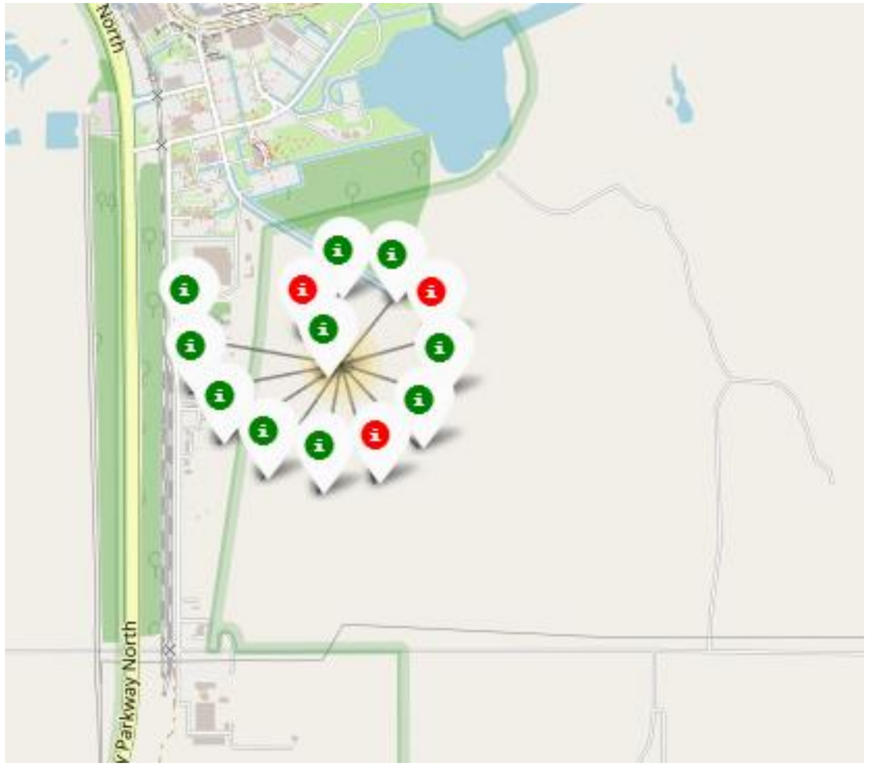
# Launch Sites Locations

Each launch site is strategically located in close proximity to coastlines, enabling rockets to be launched towards the ocean. This practice significantly mitigates the risk of debris falling or exploding near populated areas, prioritizing safety by minimizing potential hazards to people and property.

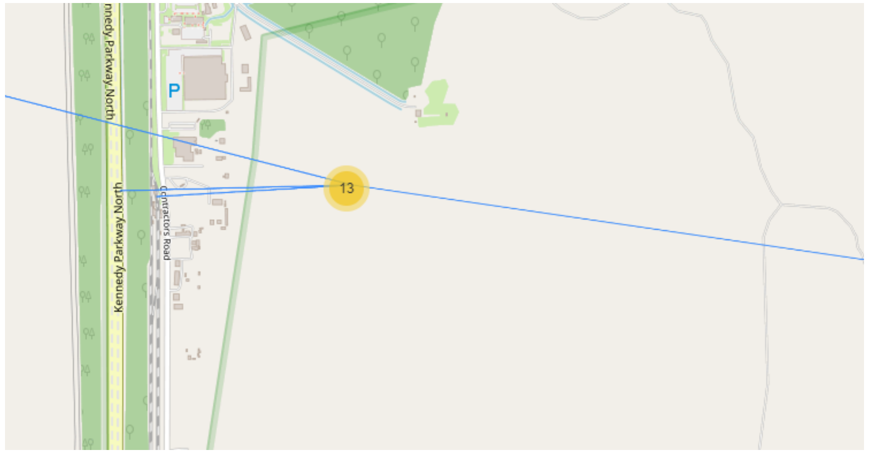


# Colour-labeled launch records on the map

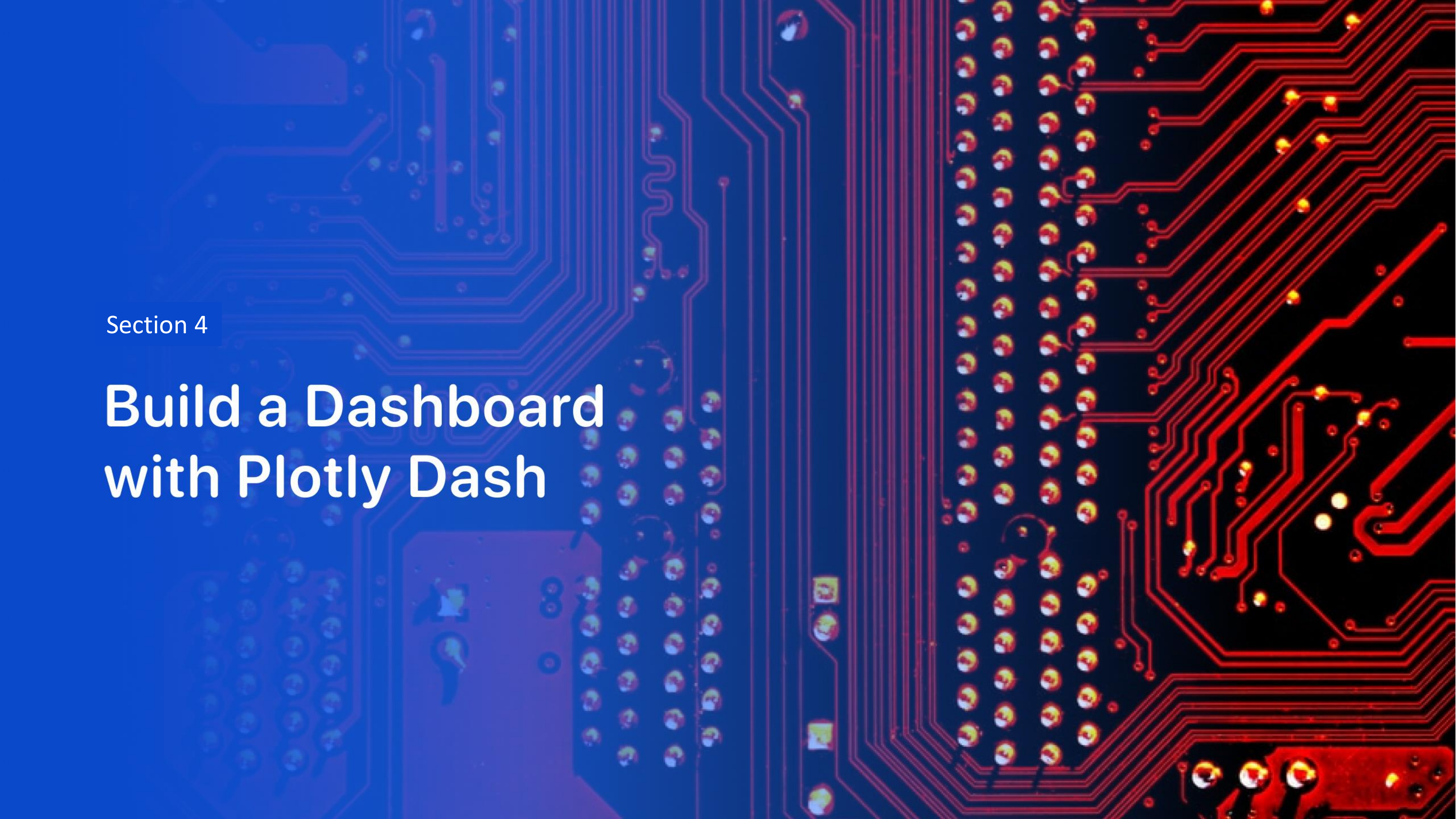
The color-coded markers, specifically green indicating successful launches and red signifying failed launches, allow for quick identification of launch sites with notably high or low success rates.



# Distance from the launch site KSC LC-39A to its proximities



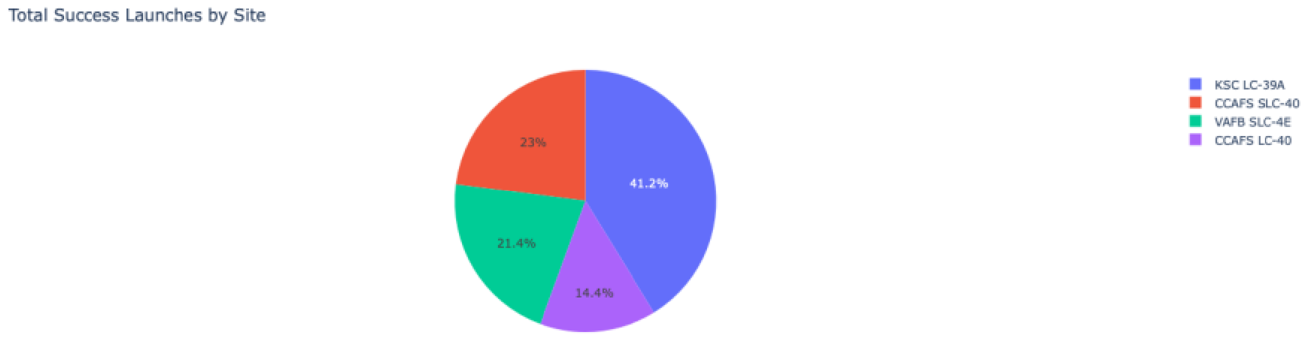
The visual analysis of launch site KSC LC-39A reveals its proximity to various elements: it's relatively close to the railway (15.23 km), highway (20.28 km), and coastline (14.99 km). Moreover, this launch site is also in relative proximity to its nearest city, Titusville (16.32 km).



Section 4

# Build a Dashboard with Plotly Dash

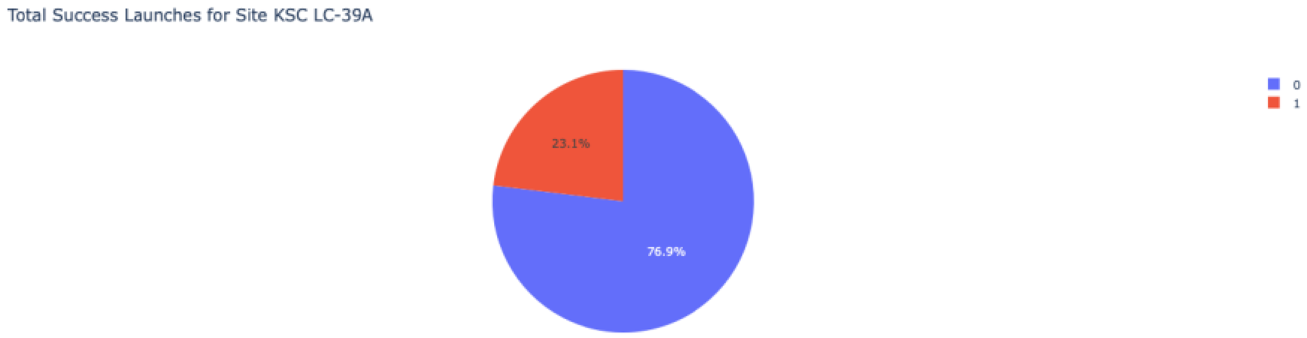
# Overall Launch Success Count



The chart distinctly illustrates that among all sites, KSC LC-39A stands out with the highest number of successful launches.

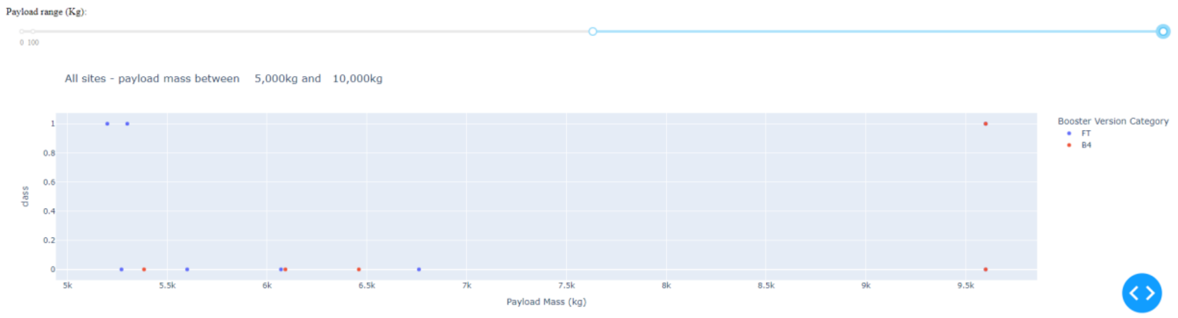
# Top Launch Success Ratio Site

- Among the sites, KSC LC-39A boasts a 76.9% success rate, having seen 10 successful landings and only 3 failed attempts.

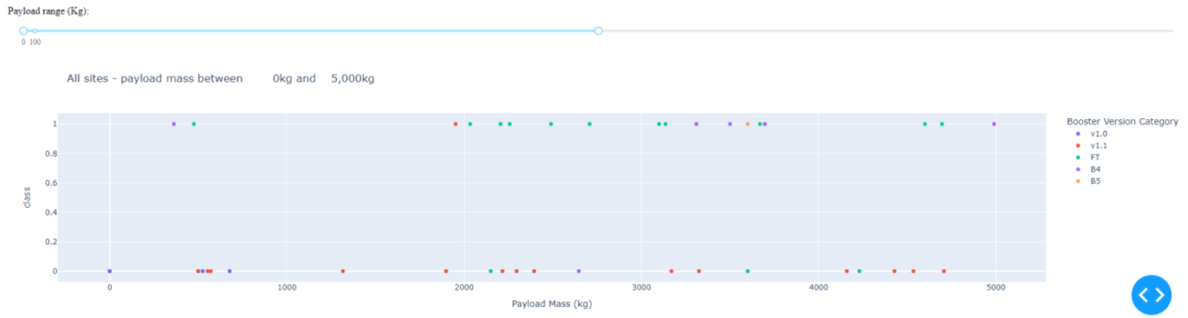


# Payload Mass vs. Launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate


Payload range (Kg):

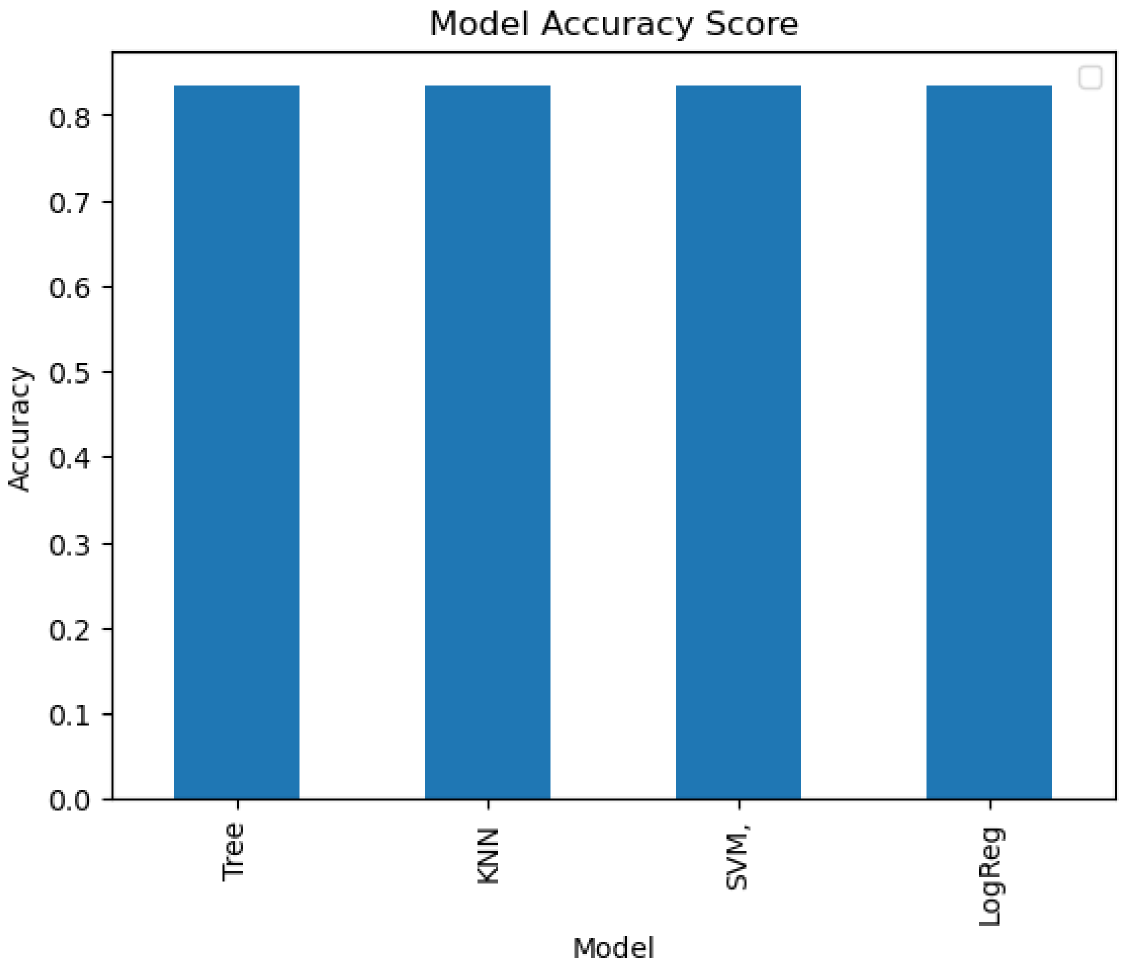
All sites - payload mass between 5,000kg and 10,000kg


Payload range (Kg):

All sites - payload mass between 0kg and 5,000kg

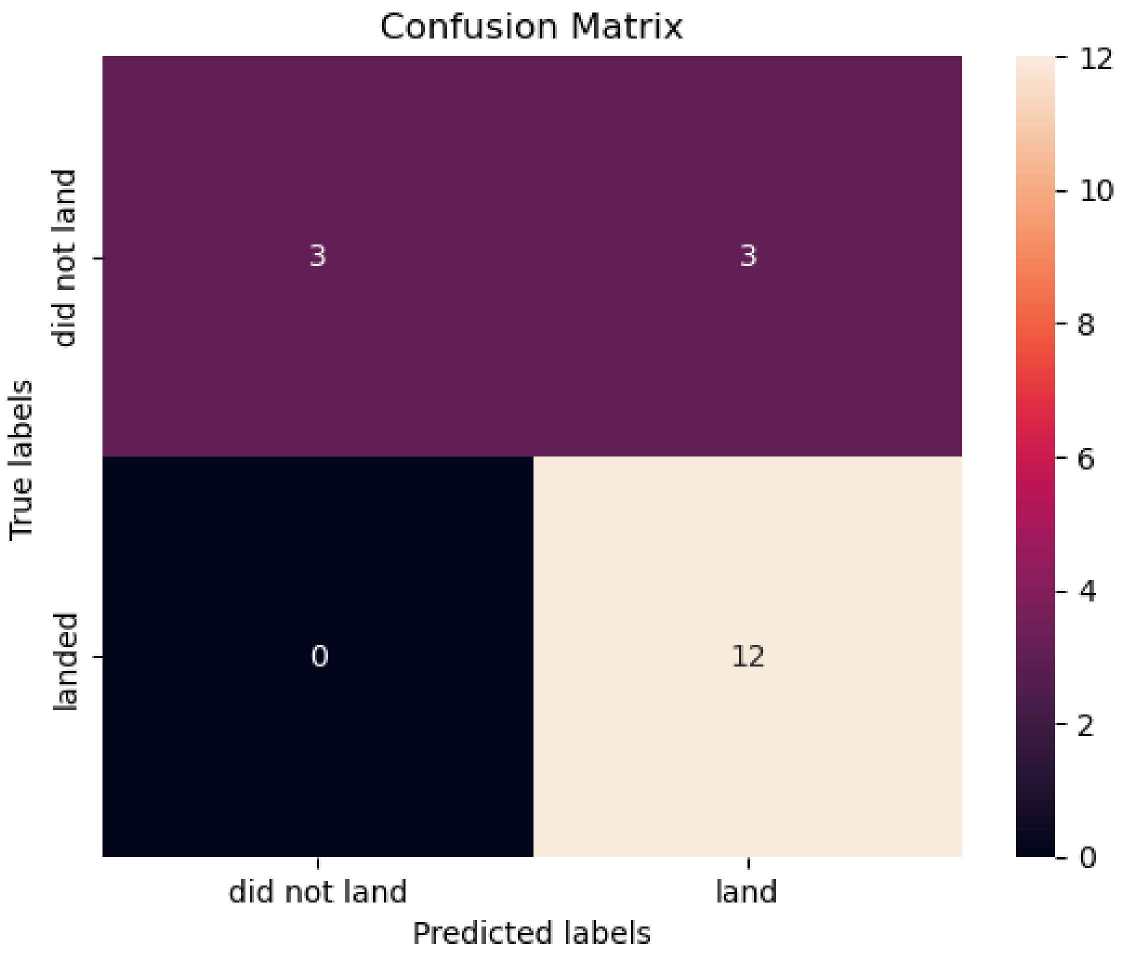
Section 5

# Predictive Analysis (Classification)



# Classification Accuracy

- The models demonstrated consistent accuracy rates of 83.33% on the test set. However, it's essential to note that the test size was relatively small, consisting of only 18 samples. This small sample size can lead to significant variance in accuracy results, as observed in the Decision Tree Classifier model during repeated runs. To ascertain the best model, it's probable that we require a more substantial dataset.



# Confusion Matrix

- All models exhibited identical performance on the test set, resulting in consistent confusion matrices across all models. Specifically, the models correctly predicted 12 successful landings when the true label indicated a successful landing. Additionally, they correctly predicted 3 unsuccessful landings when the true label indicated an unsuccessful landing.

- However, an issue arose where the models predicted 3 successful landings when the true label actually indicated unsuccessful landings (false positives). This suggests that our models tend to overpredict successful landings.

# Conclusions

- Determining the optimal model is inconclusive at present. To enhance accuracy and better identify the most suitable machine learning model, collecting additional data is recommended.Launches with lower payload masses yield superior results compared to those with larger payloads.
- The majority of launch sites are near the Equator, while all sites are extremely close to coastlines.
- Success rates for launches demonstrate an upward trajectory over the years.
- KSC LC-39A stands out with the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO exhibit a flawless 100% success rate.

# Appendix

GitHub repository url:
https://github.com/SWillm/DataScienceCaspstone

Special Thanks to All Instructors:
https://www.coursera.org/professional-certificates/ibm-data-science?#instructors