Winning Space Race with Data Science

Ivan Rudyk
11.02.2025

Outline

- Executive Summary
- Introduction
- Methodology
- Insights drawn from EDA
- Launch Sites Proximities Analisys
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusions
- Appendix

Executive Summary

- **Summary of methodologies**

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

- **Summary of all results**

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

Introduction

- **Project background and context**

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Common problems to explore:**

- Which elements determine whether the rocket will successfully land?

- Ways of different elements interact to affect the likelihood of a successful landing.

- Has the success rate of first-stage landings improved over the years?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API.
  - Web Scraping from SpaceX's Wikipedia page.
- Perform data wrangling
  - Input missing value, encode categorical data, using only relevant columns of data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build several model (SVM, Classification Trees, KNN, and Logistic Regression).
  - Find the best hyperparameter for each model.
  - Find the method performs best using test data.

Data Collection

Hybrid Data Collection Strategy

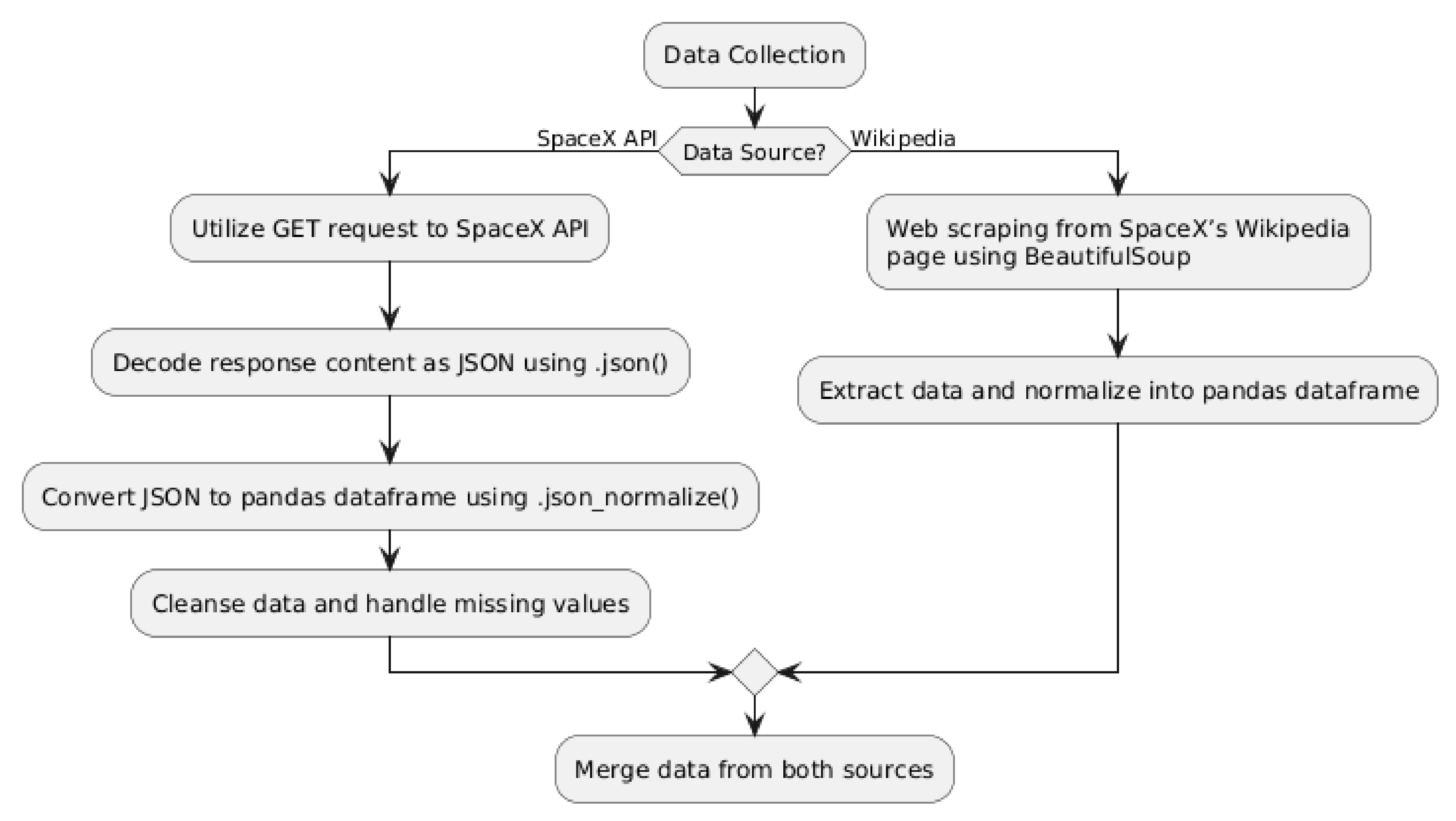
The data gathering process integrated multiple methodologies to ensure a holistic and detailed analysis:

- 1. SpaceX REST API: Leveraged to perform programmatic API requests, obtaining structured and real-time data on key launch parameters.
- 2. Wikipedia Web Scraping: Applied advanced web scraping techniques to extract tabular data from SpaceX's Wikipedia page, supplementing the dataset with historical context.

Purpose of Dual Methods

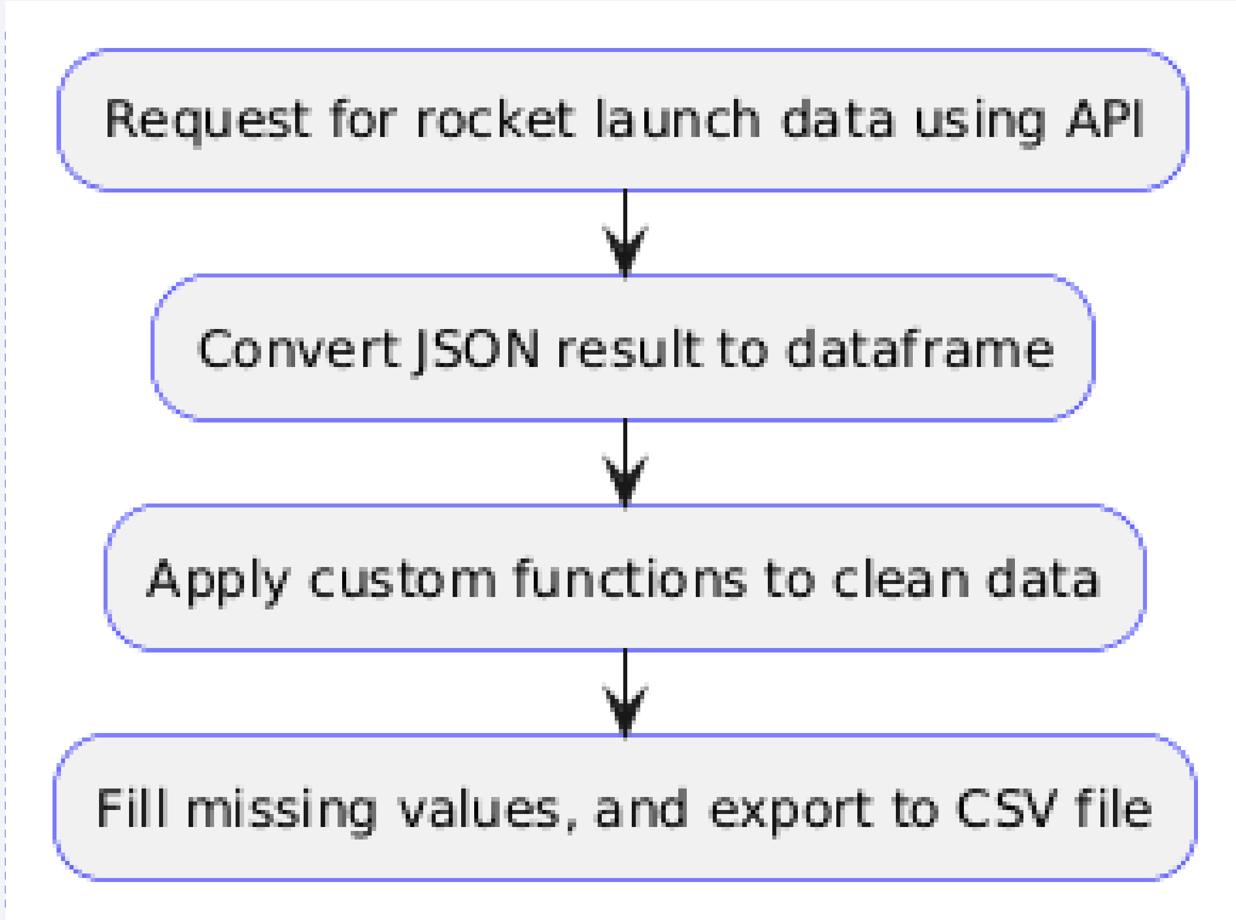
- By combining API requests and web scraping, a complete and enriched dataset was curated to support in-depth exploration and robust modeling.

Data Collection


Data Collection
Data Source?
SpaceX API
Wikipedia
Utilize GET request to SpaceX API
Decode response content as JSON using .json()
Convert JSON to pandas dataframe using .json_normalize()
Cleanse data and handle missing values
Web scraping from SpaceX's Wikipedia page using BeautifulSoup
Extract data and normalize into pandas dataframe
Merge data from both sources

Data Collection – SpaceX API

- This API provides data about launches, including information about booster , launch specifications, payload, and landing details.
- SpaceX API endpoints starts with https://api.spacexdata.com/v4/
- https://github.com/piekn/coursera_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Request for rocket launch data using API

Convert JSON result to dataframe

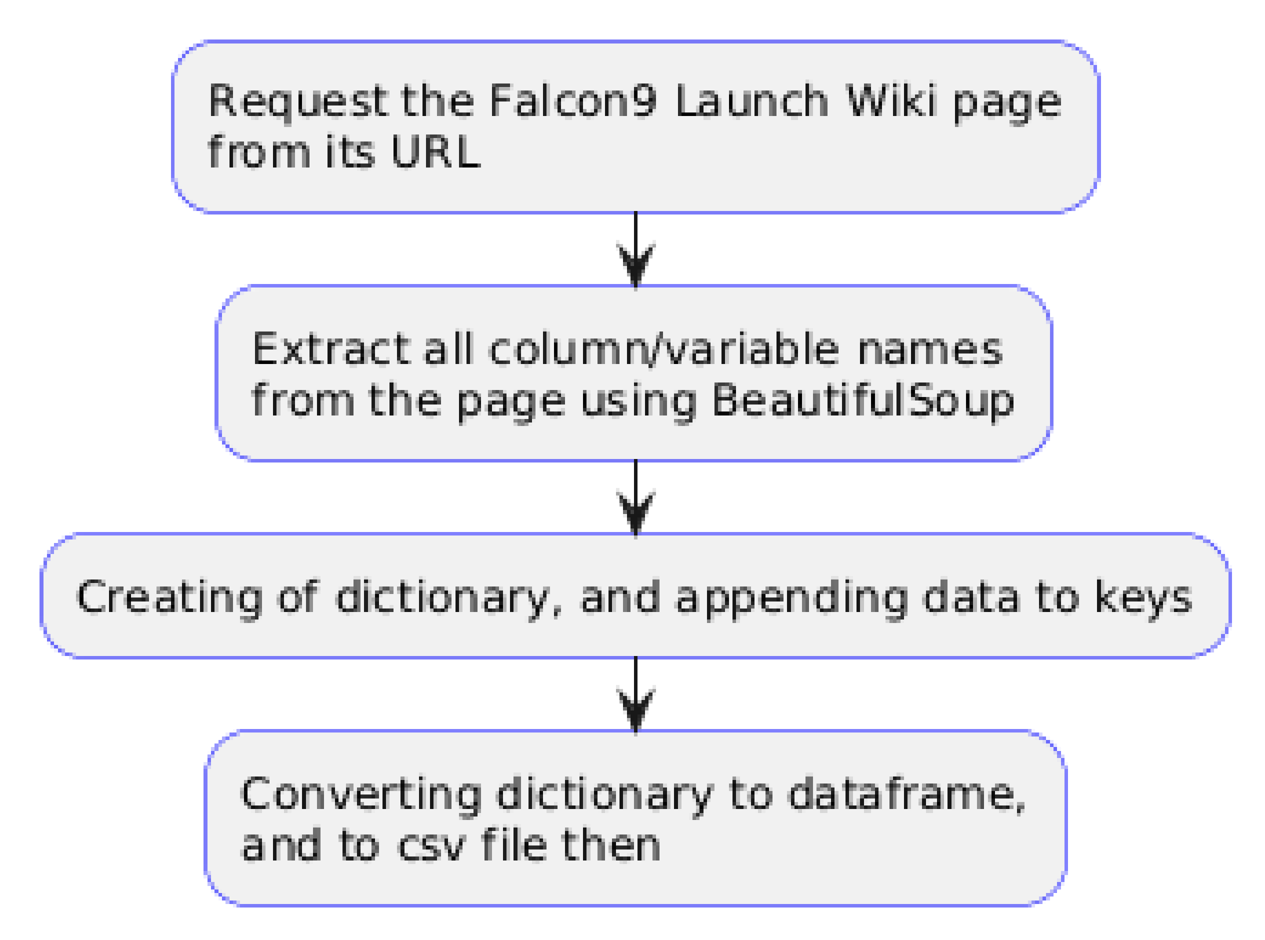
Apply custom functions to clean data

Fill missing values, and export to CSV file

Data Collection - Scraping

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia
- Using BeautifulSoup library
- Parsing the table and convert it into a Pandas data frame

- https://github.com/piekn/coursera_capstone/blob/main/jupyter-labs-webscraping.ipynb



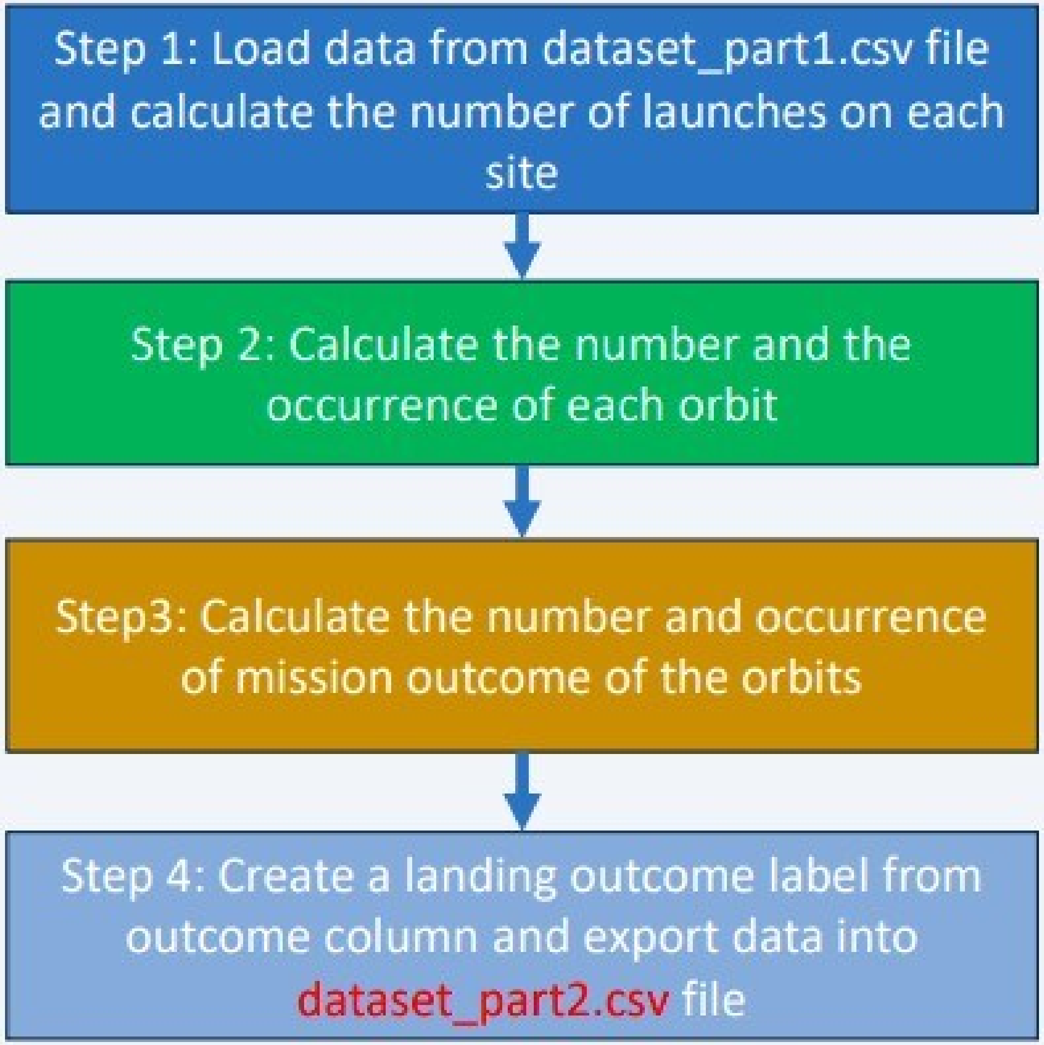
Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the page using BeautifulSoup

Creating of dictionary, and appending data to keys

Converting dictionary to dataframe, and to csv file then

Data Wrangling

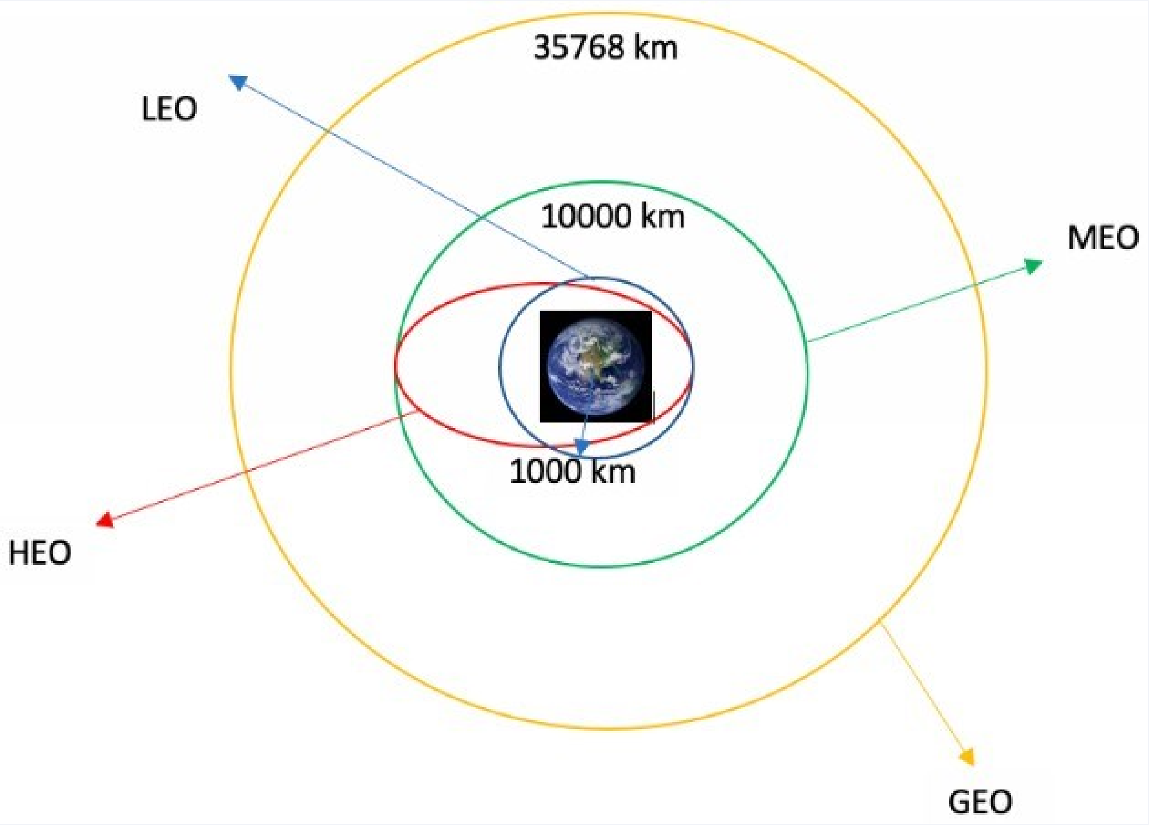


Step 1: Load data from dataset_part1.csv file and calculate the number of launches on each site

Step 2: Calculate the number and the occurrence of each orbit

Step3: Calculate the number and occurrence of mission outcome of the orbits

Step 4: Create a landing outcome label from outcome column and export data into dataset_part2.csv file



https://github.com/piekn/coursera_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter Graphs being drawn:

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

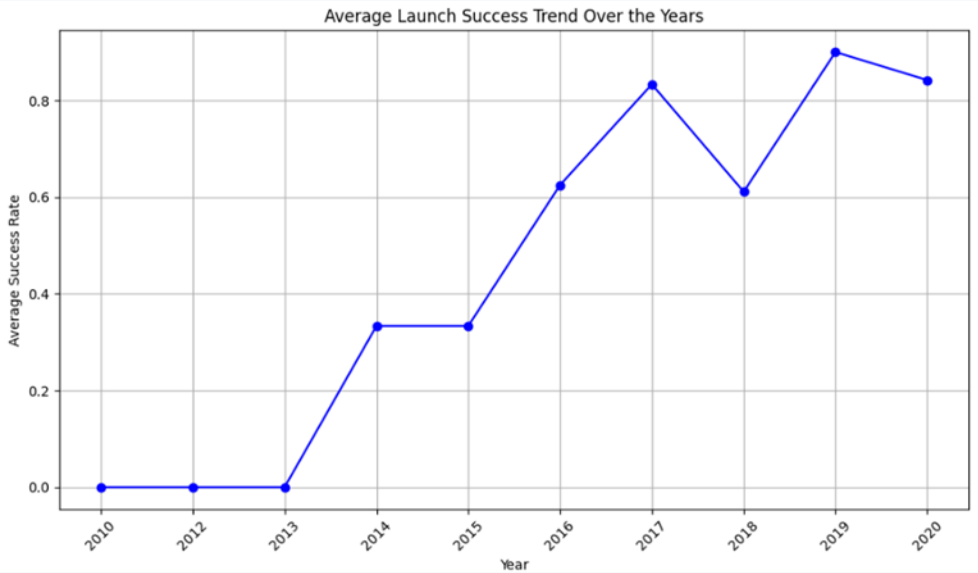
Bar Graph being drawn:

- Success Rate by Orbit Type

https://github.com/piekn/coursera_capstone/blob/main/edadataviz.ipynb

Line Graph being drawn:

- Success Rate VS. Year



EDA with SQL

- The names of the unique launch sites in the space mission
- Records where launch sites begin with the string 'CCA‘
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- The date when the first succesful landing outcome in ground pad was acheived.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes

https://github.com/piekn/coursera_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The markers (and their clusters), circles, lines were created and added to this folium map.
- Dataset `spacex_launch_geo.csv` with coordinates of the launches was added.
- Markers indicate the points like launch sites.
- Marker clusters consolidate the groups of points, for better overview at map.
- Circles highlight the areas around specific coordinates, like NASA Johnson Space Center.
- Lines were used to show the shortest path between two coordinates at the map.

- https://github.com/piekn/coursera_capstone/blob/main/lab_jupyter_launch_site_location%20(2).ipynb

Build a Dashboard with Plotly Dash

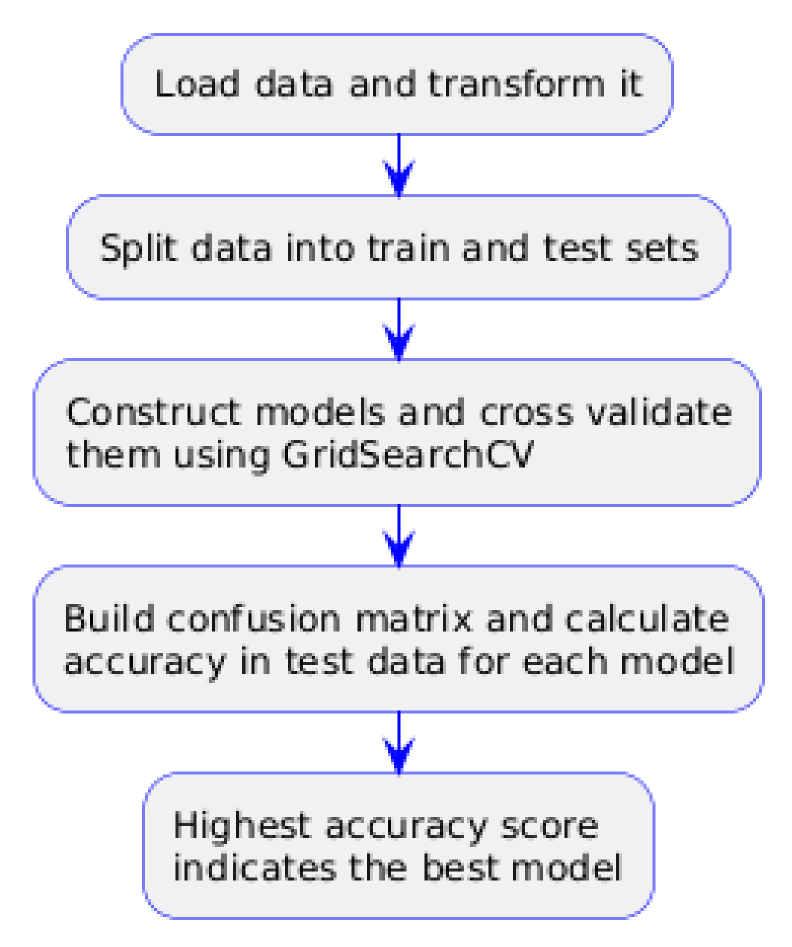
- Were added 2 kinds of the charts: pie chart, and scatter point chart.
- Pie chart shows a distribution of success launches.
- Pie chart was supplemented with dropdown list of the launch sites.
- Scatter point chart demonstrates a correlation between payload and success of the launch.
- Range Slider is used to select Payload range

- https://github.com/piekn/coursera_capstone/blob/main/spacex_dash_app%20(2).py

Predictive Analysis (Classification)

- Load the data, transform it using StandardScaler, and split it into train and test data.
- Construct machine learning models and tune their respective various hyperparameters.
- Calculate accuracy in test data.
- The most effective classification model was discovered by comparing accuracy score.
- 4 models were compared.

https://github.com/piekn/coursera_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Load data and transform it

Split data into train and test sets

Construct models and cross validate them using GridSearchCV

Build confusion matrix and calculate accuracy in test data for each model

Highest accuracy score indicates the best model

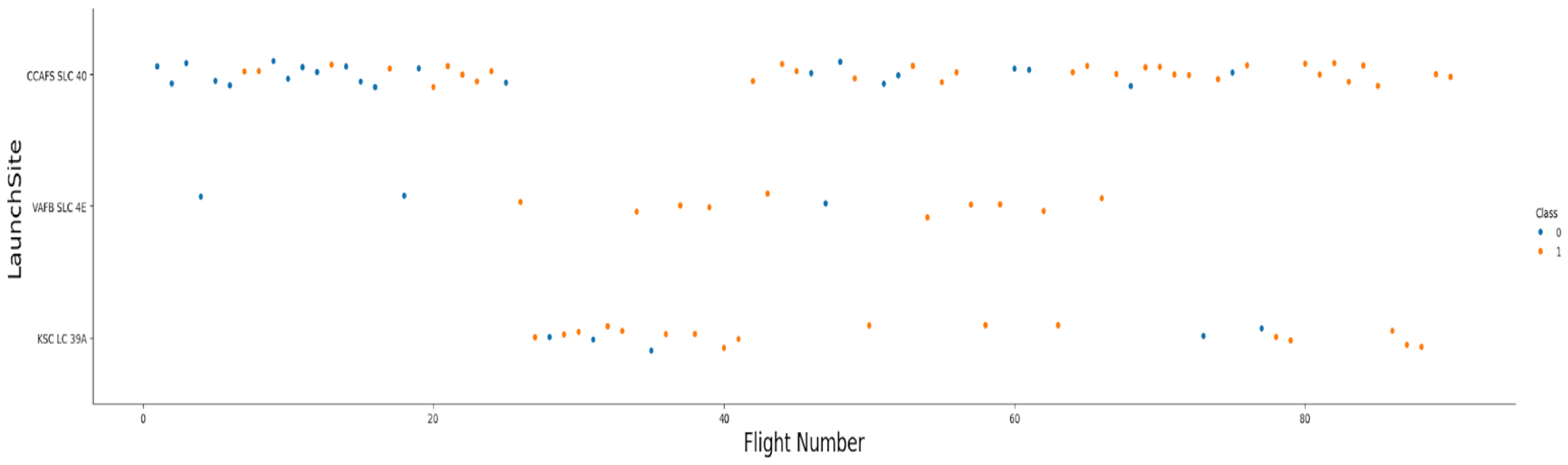
Results

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

Section 2

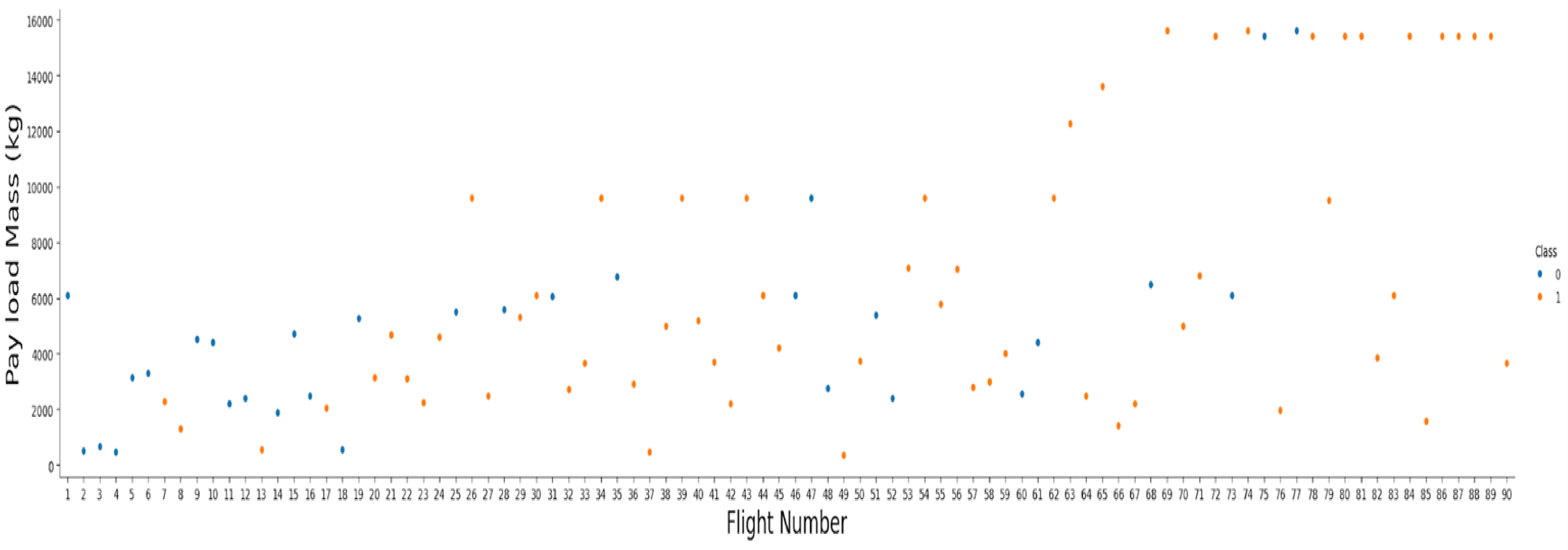
Insights drawn from EDA

Flight Number vs. Launch Site



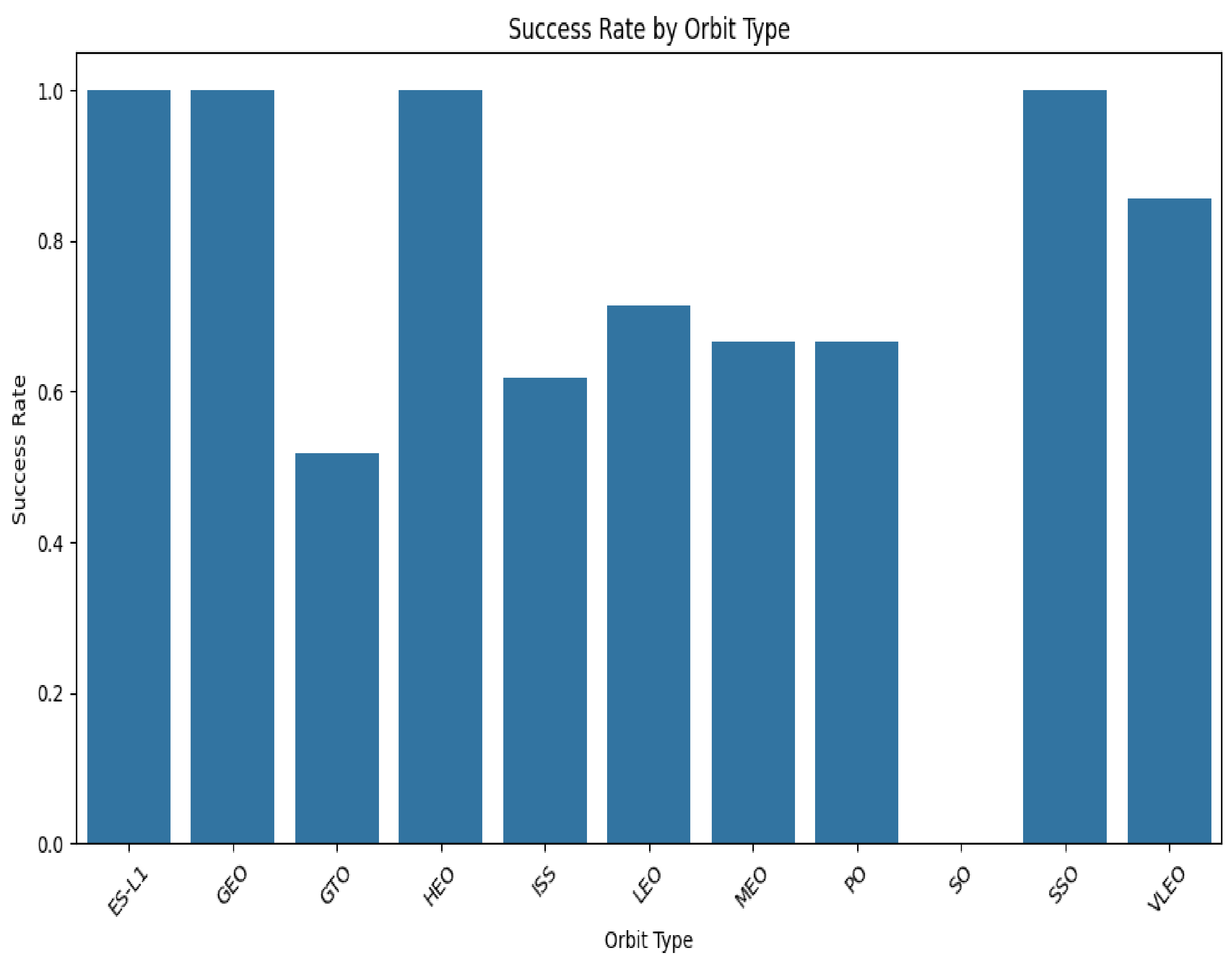
- Count of the launches from site CCAFS SLC 40 is significantly higher than from other sites.
- Early flights experienced failures, but recent flights are consistently succeeded, demonstrating progress.

Payload vs. Launch Site



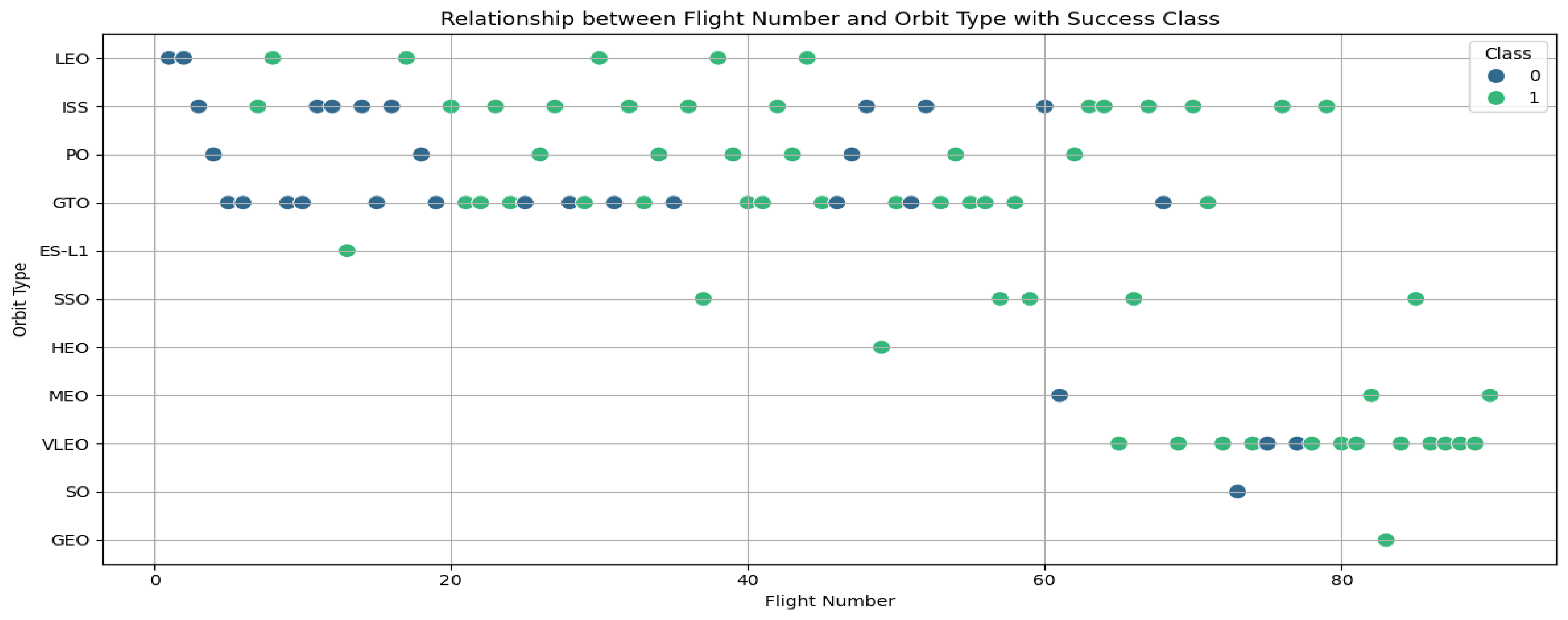
- As flight numbers increase, the likelihood of a successful first-stage landing improves, indicating SpaceX's refinement over time.

Success Rate vs. Orbit Type



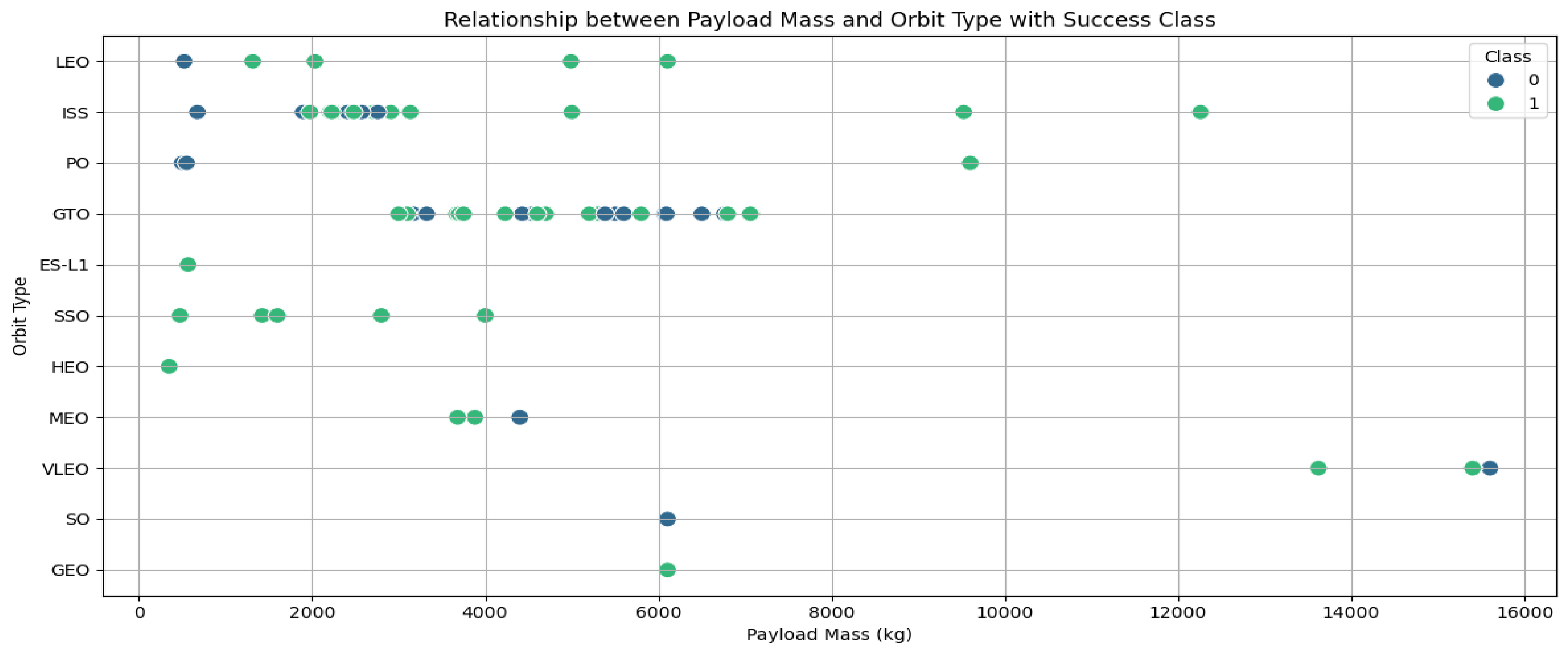
- **ES-L1, GEO, HEO, SSO**: These orbits consistently achieved successful landing, reflecting their reliability and mission execution.
- **SO**: Missions targeting this orbit have experienced no successful landings.
- **GTO, ISS, LEO, MEO, PO**: These orbits show moderate success rates, indicating variability in mission complexity or external factors.

Flight Number vs. Orbit Type



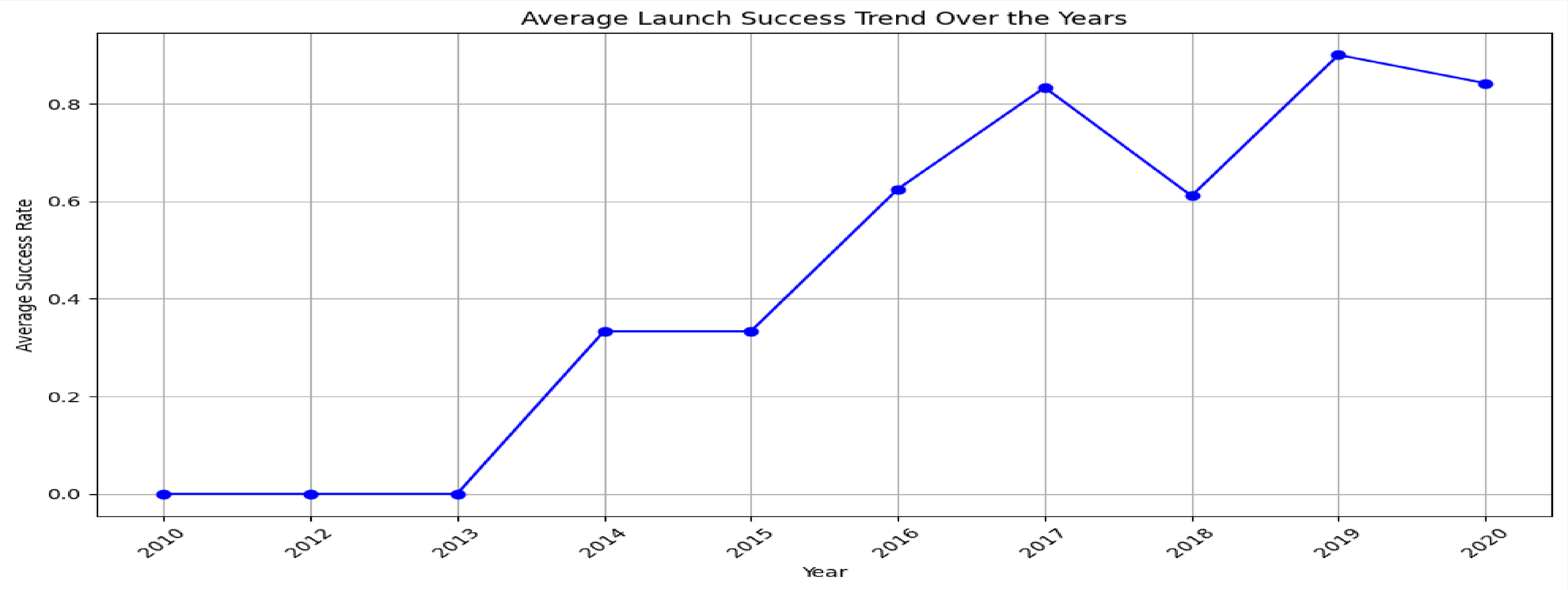
- In general, a positive correlation is observed between the number of flights and success rate. As flight numbers increase, missions are more likely to succeed, suggesting operational improvements or experience accumulation.

Payload vs. Orbit Type



- LEO and ISS orbits demonstrate high success rate for high Payload Mass.
- GTO orbit has average success level for average Payload Mass range.

Launch Success Yearly Trend



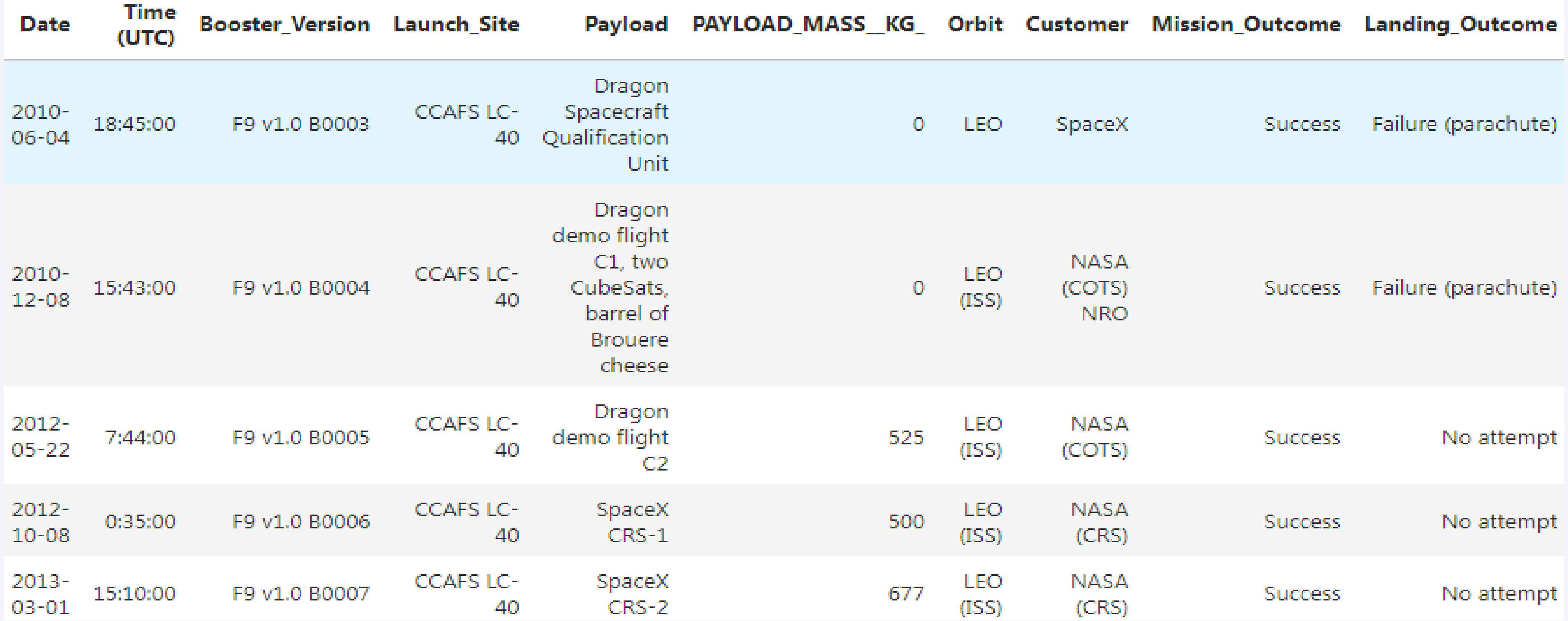
- Starting from 2013, the success rate has shown a consistent increase, reflecting improvements in technology, processes, and mission execution strategies.
- By 2020, the success rate approaches near-optimal levels, demonstrating SpaceX's growing expertise and reliability in achieving successful landings

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

- This query identifies all distinct launch sites used in the SpaceX missions.
- The DISTINCT keyword retrieves unique entries from the launch_site column in the dataset.
- Four unique launch sites were identified:
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40
- This information is crucial for understanding the geographical distribution of SpaceX's launch operations.

Launch Site Names 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	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

```
%sql SELECT DISTINCT * FROM SPACEXTABLE where Launch_Site LIKE 'CCA%' limit 5
```

- This selection focuses on SpaceX's activities at a specific launch complex for easier analysis.

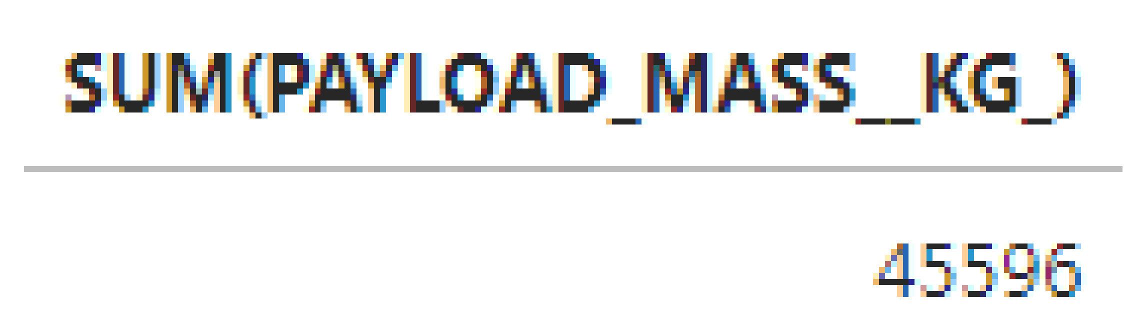
Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE where Customer =
'NASA (CRS)'
```

- This query calculates the total payload mass launched by SpaceX for NASA (CRS) missions.
- Key Elements:

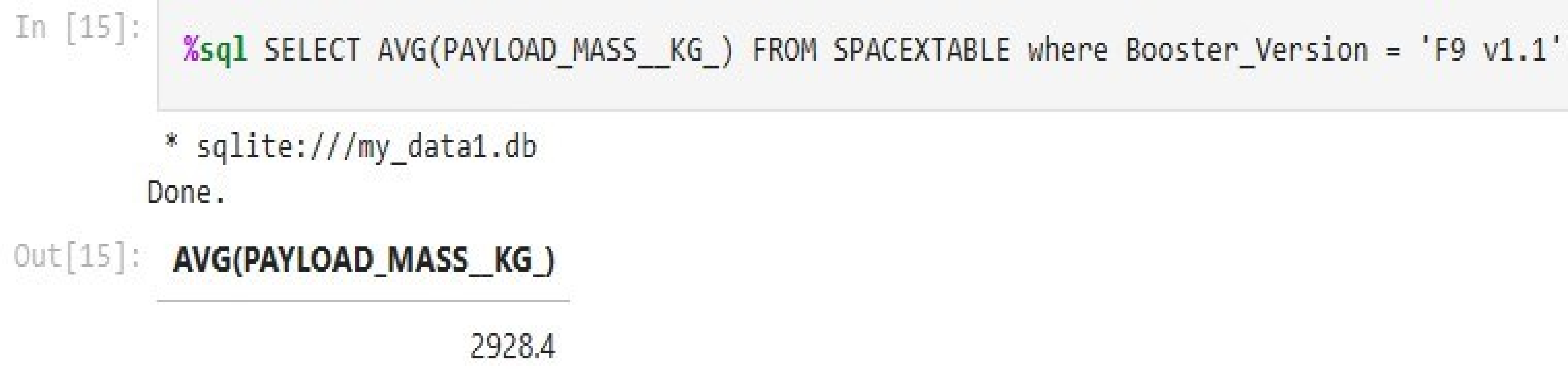
- SUM(payload_mass__kg_) : Computes the sum of payload masses.

- WHERE clause filters the dataset to only perform calculations on Customer NASA (CRS)



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

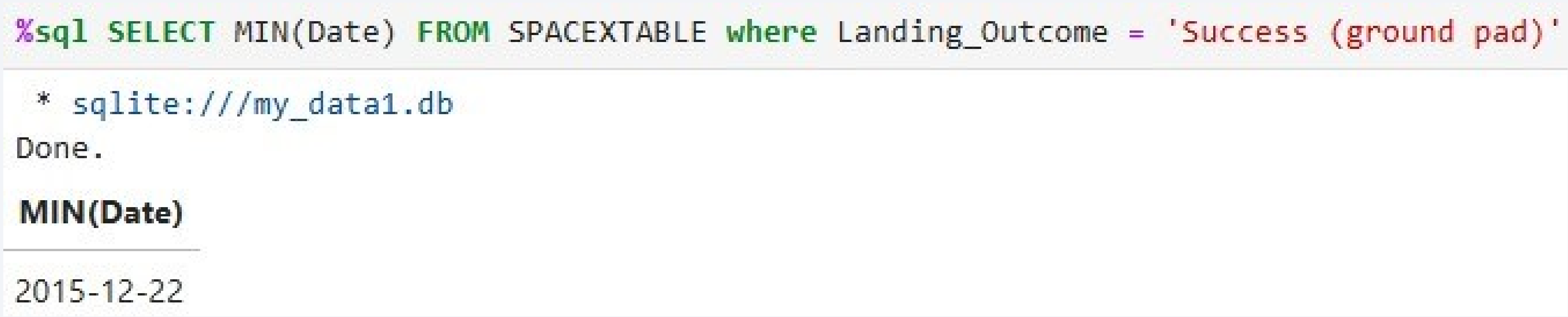


```
In [15]:  %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE where Booster_Version = 'F9 v1.1'

           * sqlite:///my_data1.db
          Done.

Out[15]:  AVG(PAYLOAD_MASS__KG_)
          ----------------------
                          2928.4
```

Key Elements:

- AVG(payload_mass__kg_) Computes the average payload mass.
- WHERE booster_version LIKE '%F9 v1.1%': filters for launches where F9 v1.1 booster was used.

First Successful Ground Landing Date



```
%sql SELECT MIN(Date) FROM SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

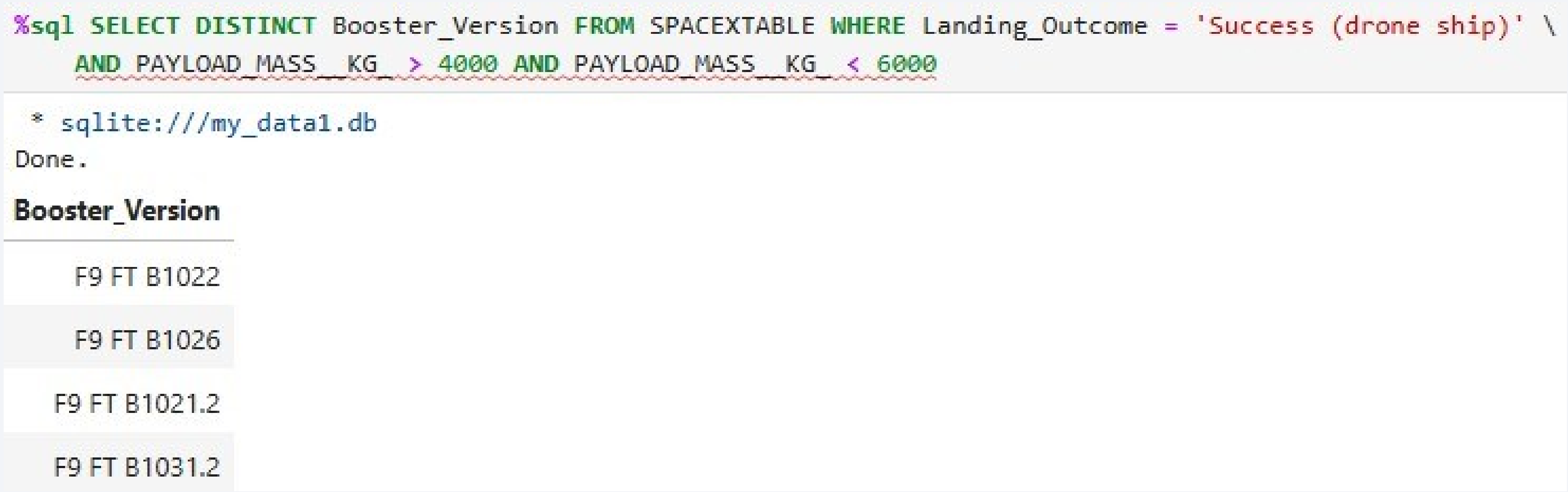
```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

Key Elements:

- MIN(date): finds the earliest date in the dataset.
- WHERE landing_outcome = 'Success (ground pad)': Filters for successful ground pad landings.

Successful Drone Ship Landing with Payload between 4000 and 6000



```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' \
    AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

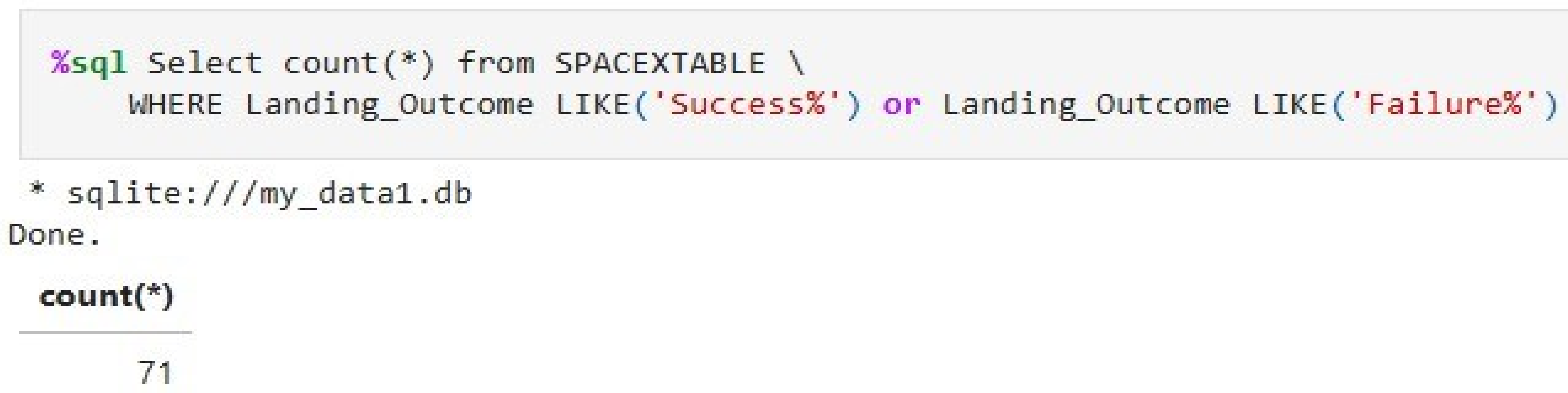
```
 * sqlite:///my_data1.db
Done.
```

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

This query lists the boosters that:

- 1. Successfully landed on a drone ship (landing_outcome = 'Success (drone ship)').
- 2. Had a payload mass between 4000 kg and 6000 kg (payload_mass__kg_ BETWEEN 4000 AND 6000).

Total Number of Successful and Failure Mission Outcomes



```
%sql Select count(*) from SPACEXTABLE \
    WHERE Landing_Outcome LIKE('Success%') or Landing_Outcome LIKE('Failure%')
```

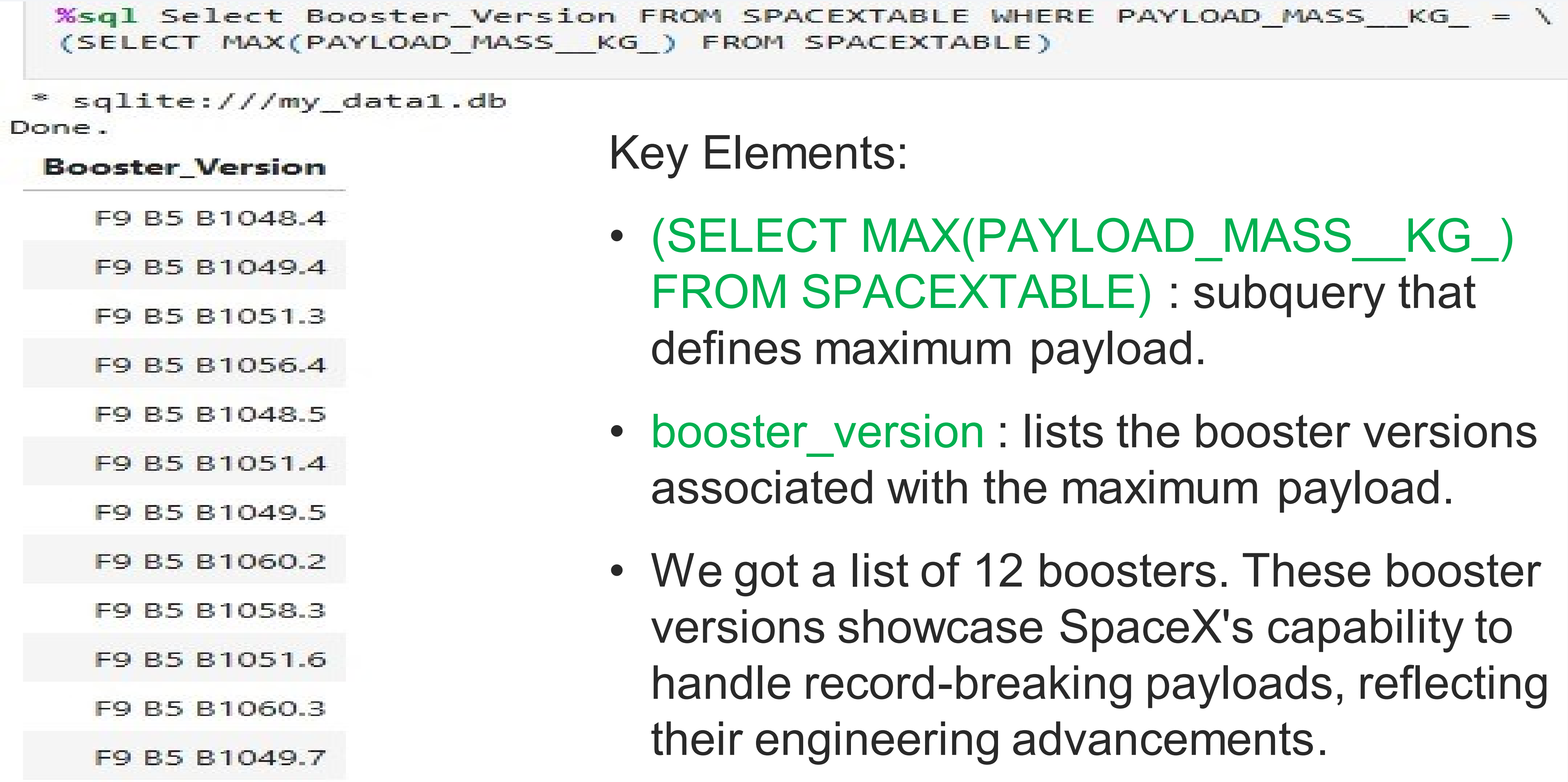
```
 * sqlite:///my_data1.db
Done.
count(*)
      71
```

Key Elements:

- count (*) : calculates count of outcomes.
- WHERE Landing_Outcome LIKE('Success%') or Landing_Outcome LIKE('Failure%') : filters for proper types of outcomes only.

WHERE Landing_Outcome LIKE('Success%') or Landing_Outcome LIKE('Failure%')

Boosters Carried Maximum Payload



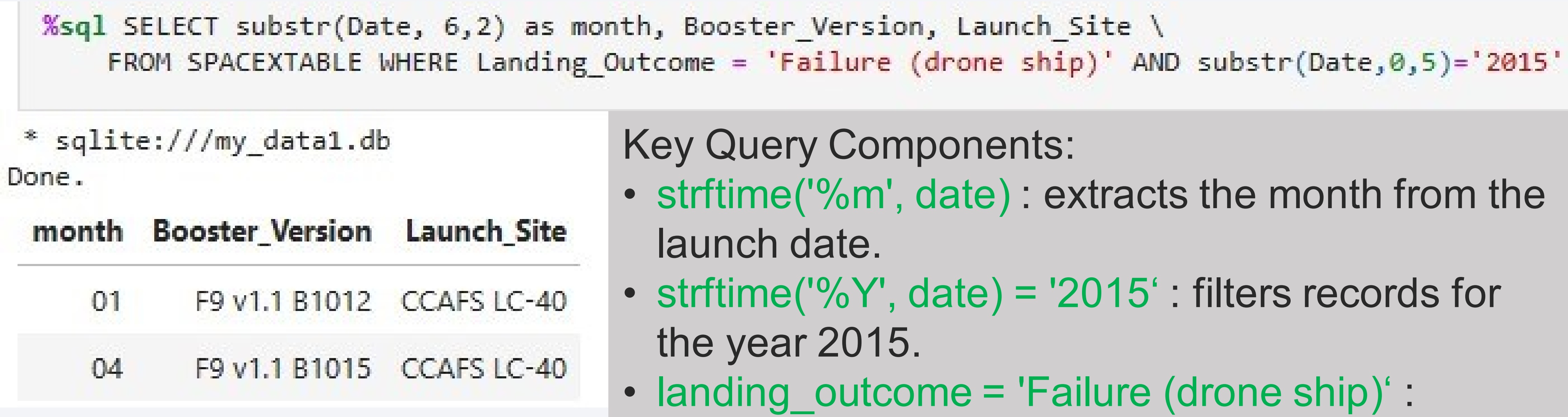
```
%sql Select Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = \
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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
```

Key Elements:

- (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) : subquery that defines maximum payload.
- booster_version : lists the booster versions associated with the maximum payload.
- We got a list of 12 boosters. These booster versions showcase SpaceX's capability to handle record-breaking payloads, reflecting their engineering advancements.

2015 Launch Records



```
%sql SELECT substr(Date, 6,2) as month, Booster_Version, Launch_Site \
    FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015'
```

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

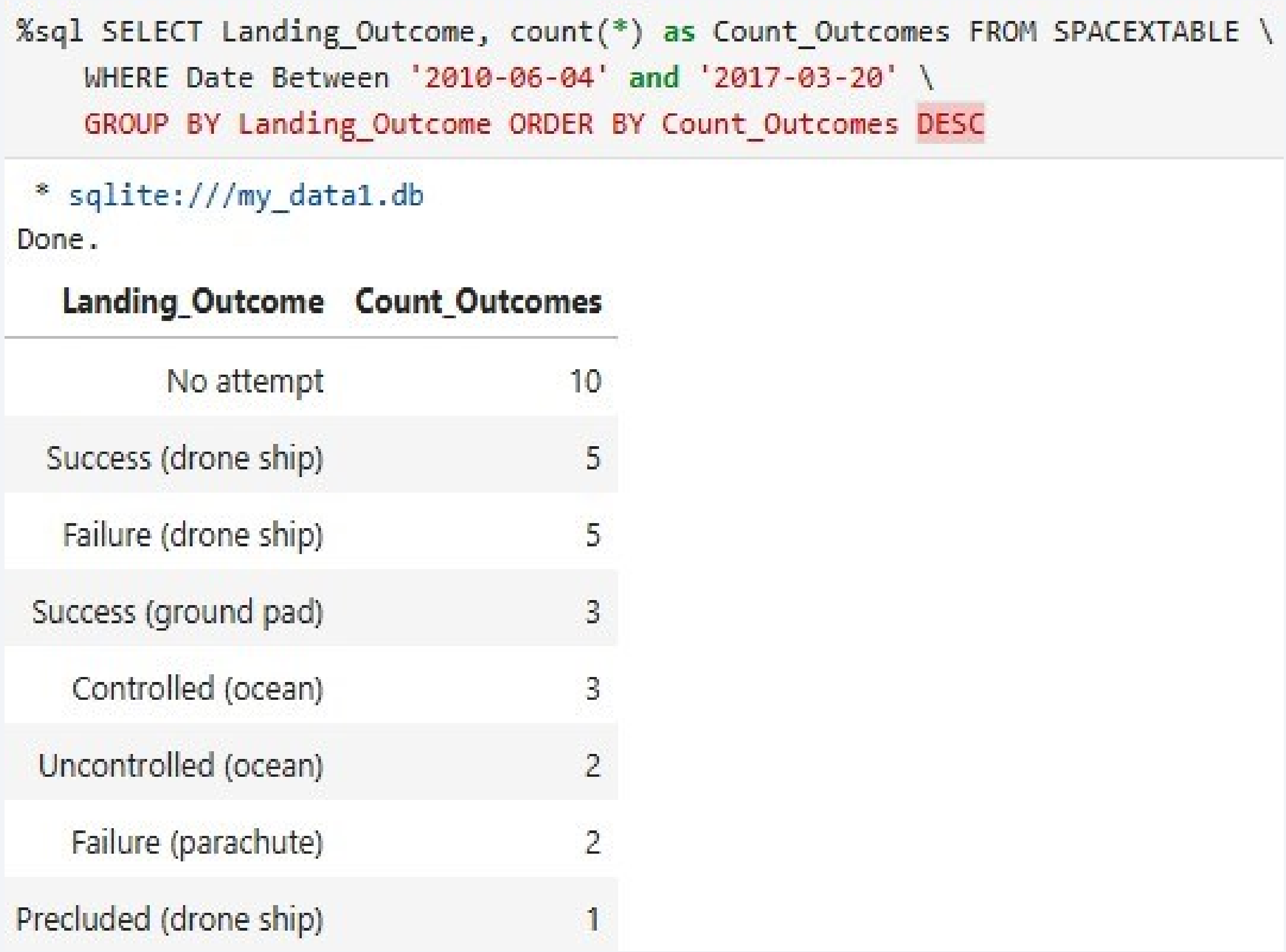
month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Key Query Components:

- strftime('%m', date) : extracts the month from the launch date.
- strftime('%Y', date) = '2015‘ : filters records for the year 2015.
- landing_outcome = 'Failure (drone ship)‘ : ensures only failed landings on drone ships are included.

This analysis provides insights into landing challenges faced by SpaceX during 2015, helping evaluate improvements in drone ship landing success rates over time.

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

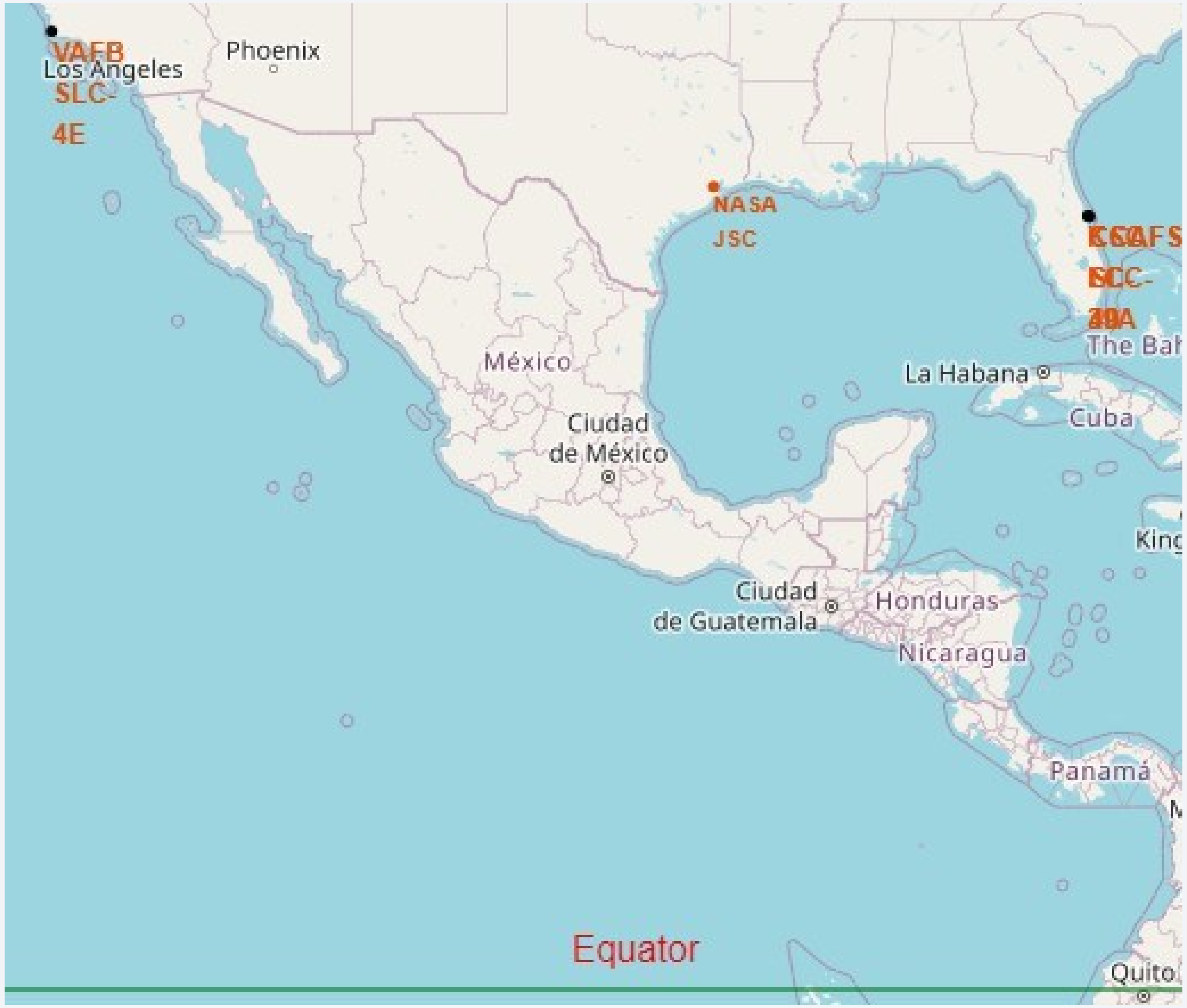
Key Query Details:

- Date Range Filter: Date BETWEEN '2010-06-04' AND '2017-03-20' ensures only launches within this timeframe are analyzed.
- Grouping: GROUP BY Landing_Outcome aggregates records by distinct landing outcomes.
- Ordering: ORDER BY Count_Outcomes DESC ranks outcomes by their frequency in descending order.

Section 3

Launch Sites Proximities Analysis

Analysis of Launch Site Locations

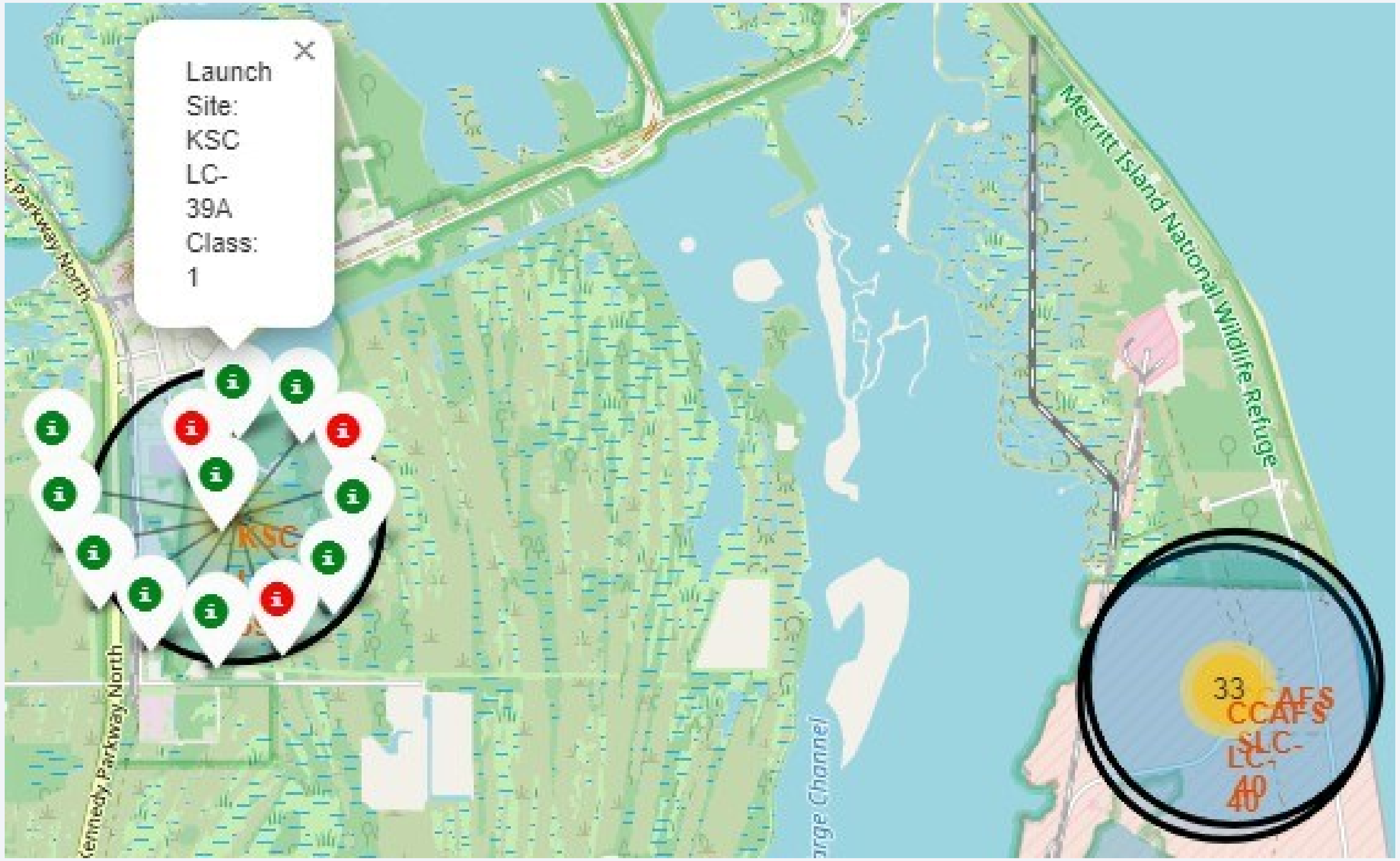


This visualization highlights SpaceX launch sites on an interactive map, providing key insights into their strategic positioning:

- Proximity to the Equator: rockets launched close the equator get benefit from inertia, conserving fuel by using the Earth's momentum to help reach orbit.

- Proximity to Coastlines: launch sites are situated close to coastlines to minimize risks for populated areas.

Visualization of Launch Site Success Rates

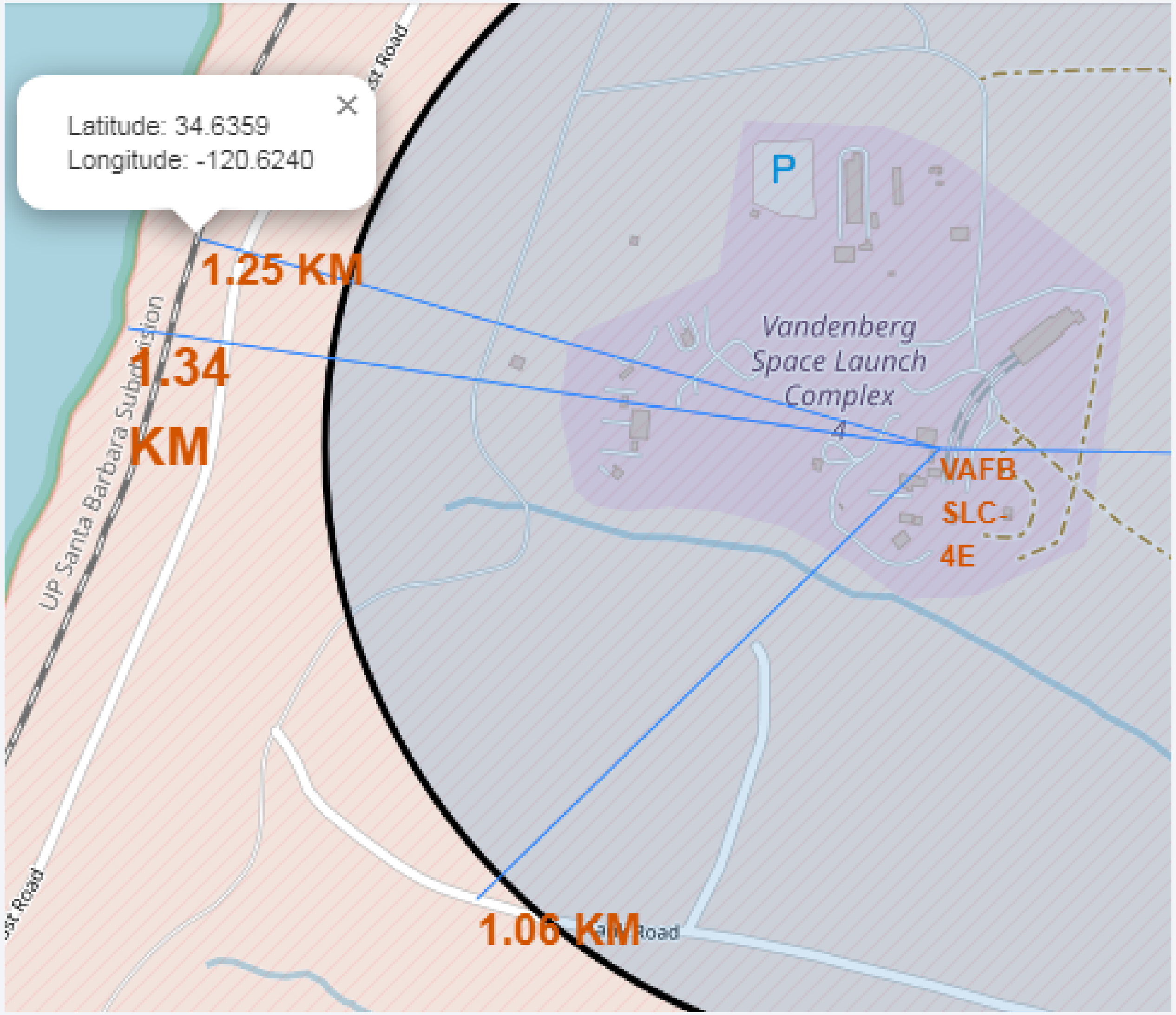


Launch Site: KSC LC-39A Class: 1

This map uses color-coded markers to represent the success and failure outcomes of launches at various SpaceX sites:

- Color representation: green marker indicates a successful launch, red marker: denotes a failed launch.
- Insights: launch site KSC LC-39A stands out with a high success rate, evident from the predominance of green markers.
- Interactive Features: marker clusters group the markers for better visualization, allowing users to zoom in for a closer view of individual launch outcomes.

Analysis of Proximity for Launch Site VAFB SLC-4E



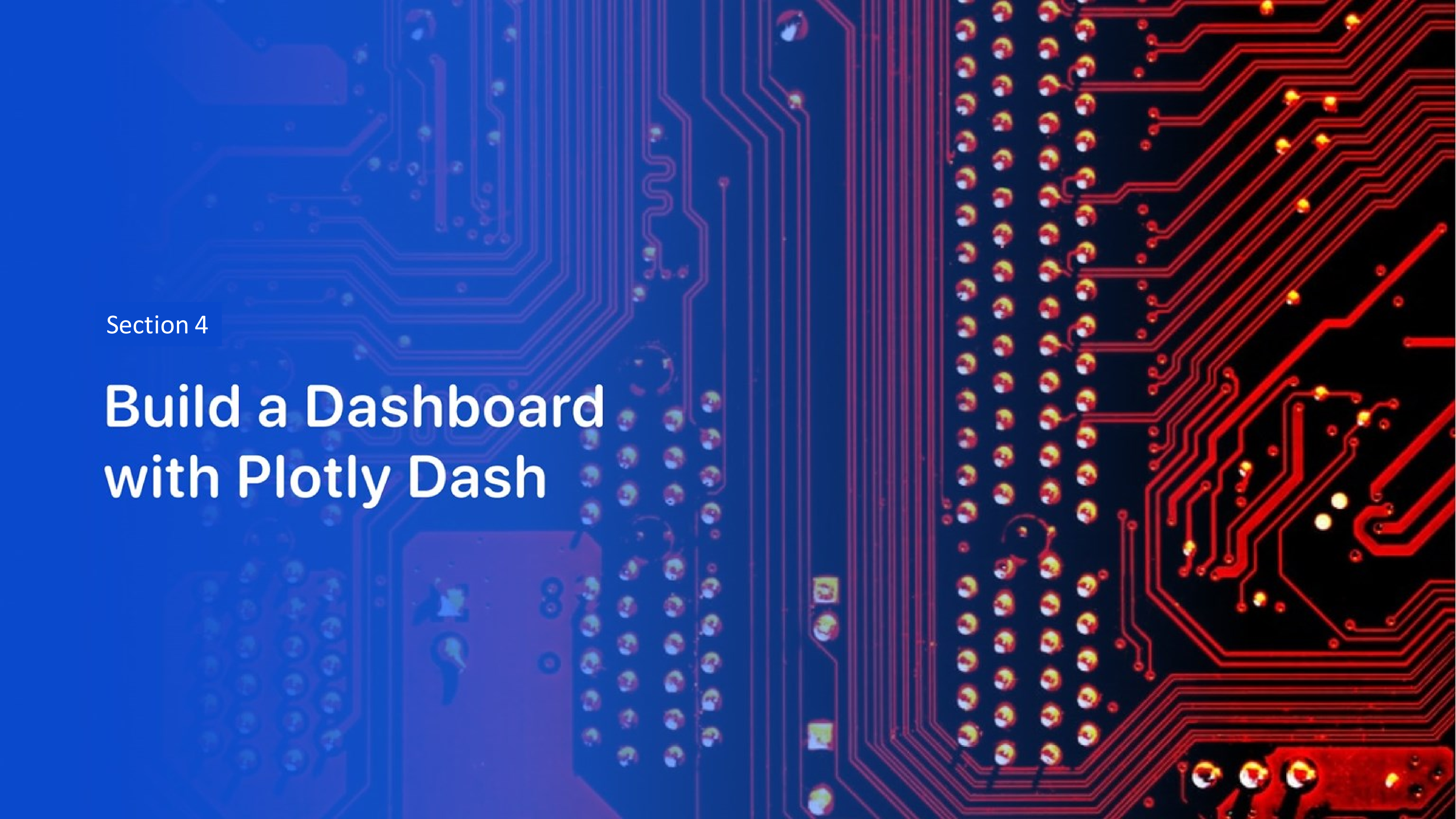
Proximity to Key Infrastructure:

- Railway: Located 1.25 km away.
- Highway: Found at a distance of 1.06 km.
- Coastline: Situated just 1.34 km away.

Closest City:

- The launch site is 11.8 km from Lompoc, the nearest city.

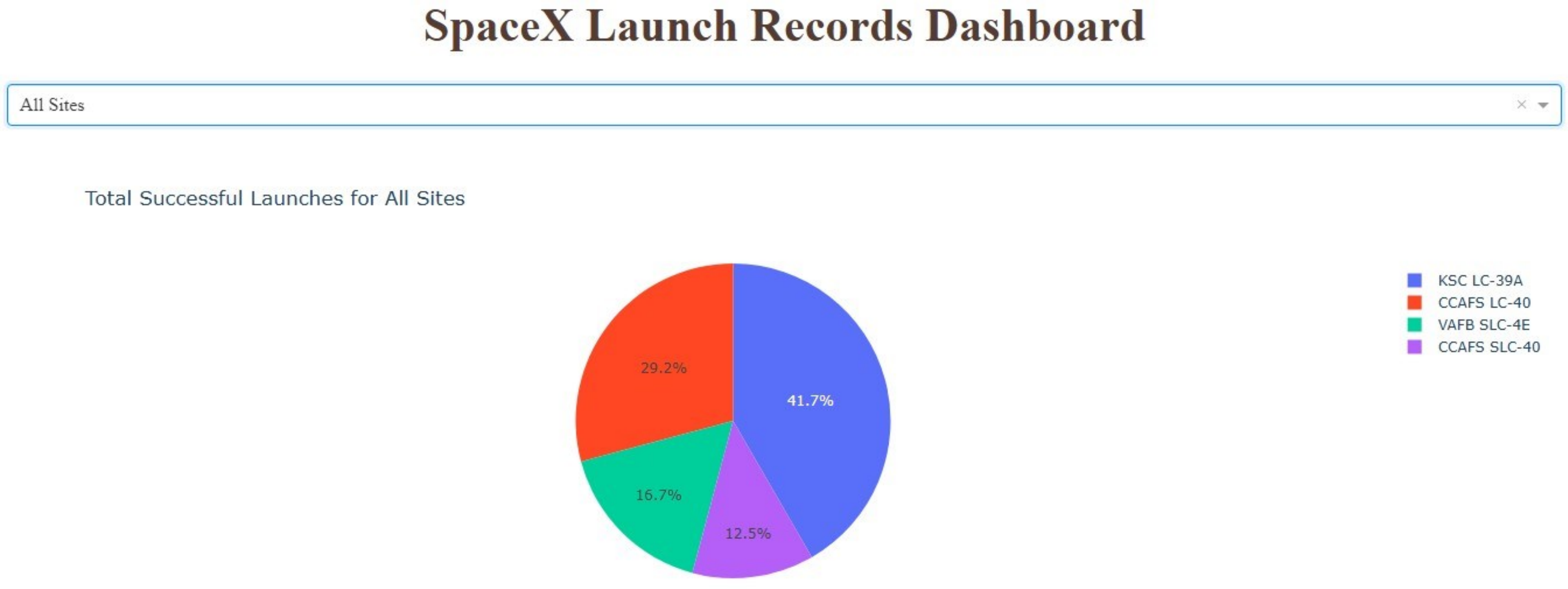
This analysis emphasizes the importance of strategic site placement to balance launch efficiency with safety considerations.



Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

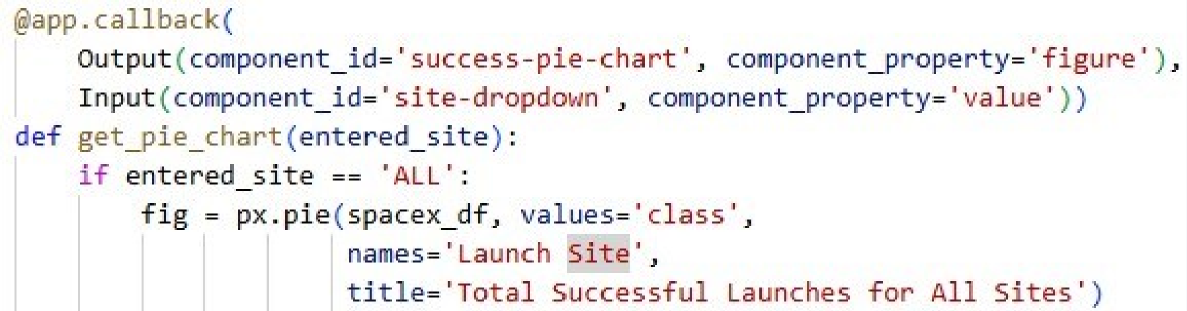


SpaceX Launch Records Dashboard

All Sites

Total Successful Launches for All Sites

This chart is dynamically generated through the following Dash callback function:



```
@app.callback(
    Output(component_id='success-pie-chart', component_property='figure'),
    Input(component_id='site-dropdown', component_property='value'))
def get_pie_chart(entered_site):
    if entered_site == 'ALL':
        fig = px.pie(spacex_df, values='class',
                     names='Launch Site',
                     title='Total Successful Launches for All Sites')
```

Key insights:

KSC LC-39A: 41.7% of successful launches.

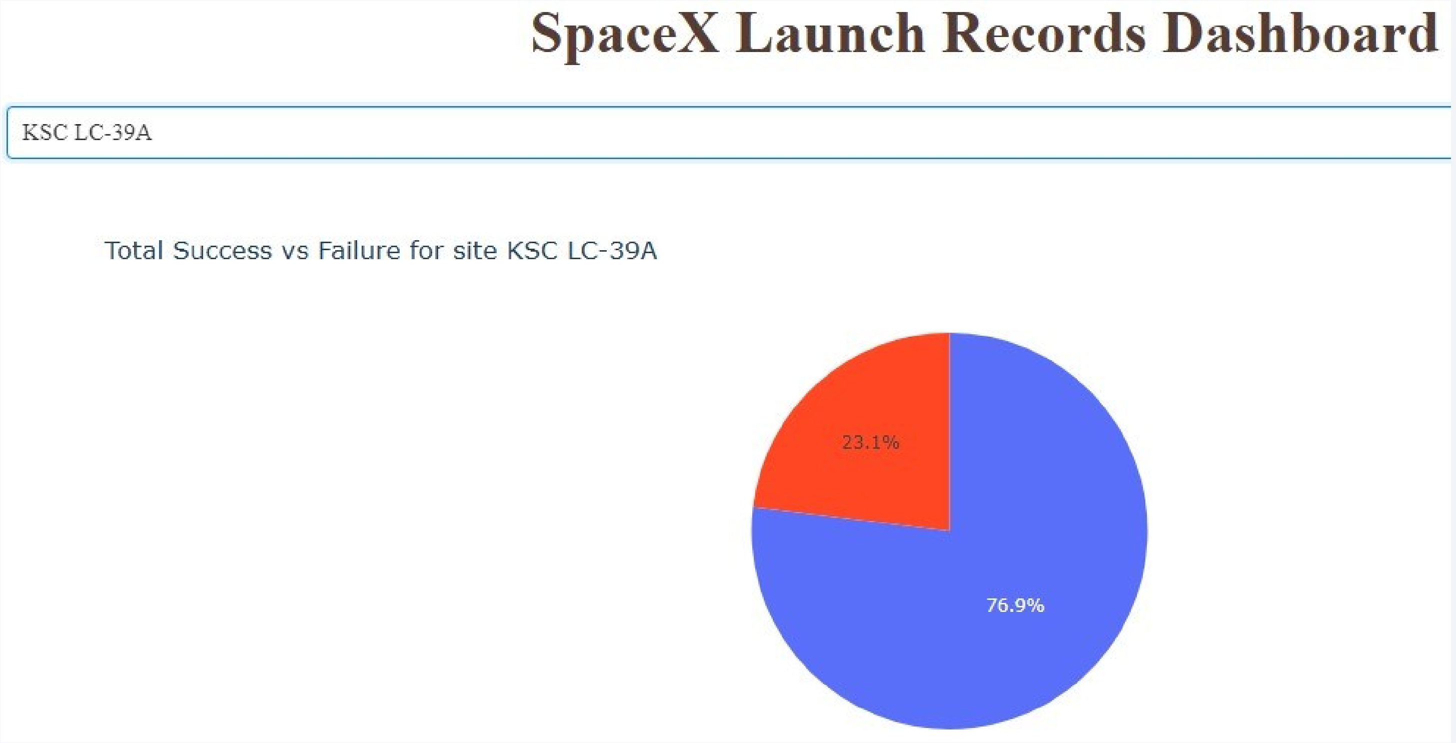
CCAFS LC-40: 29.2%.

VAFB SLC-4E: 16.7%.

CCAFS SLC-40: 12.5%.

The analysis highlights **KSC LC-39A** as the most efficient site for successful launches, providing valuable insights to optimize future launch strategies.

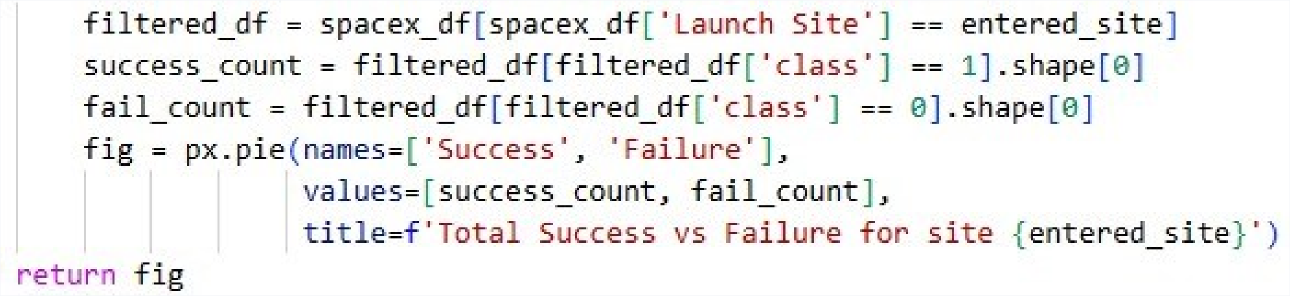
Launch Success Rate at KSC LC-39A



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success vs Failure for site KSC LC-39A



```
    filtered_df = spacex_df[spacex_df['Launch Site'] == entered_site]
    success_count = filtered_df[filtered_df['class'] == 1].shape[0]
    fail_count = filtered_df[filtered_df['class'] == 0].shape[0]
    fig = px.pie(names=['Success', 'Failure'],
                 values=[success_count, fail_count],
                 title=f'Total Success vs Failure for site {entered_site}')
return fig
```

KSC LC-39A demonstrates the highest success rate among all SpaceX launch sites, with an impressive 76.9% success rate.

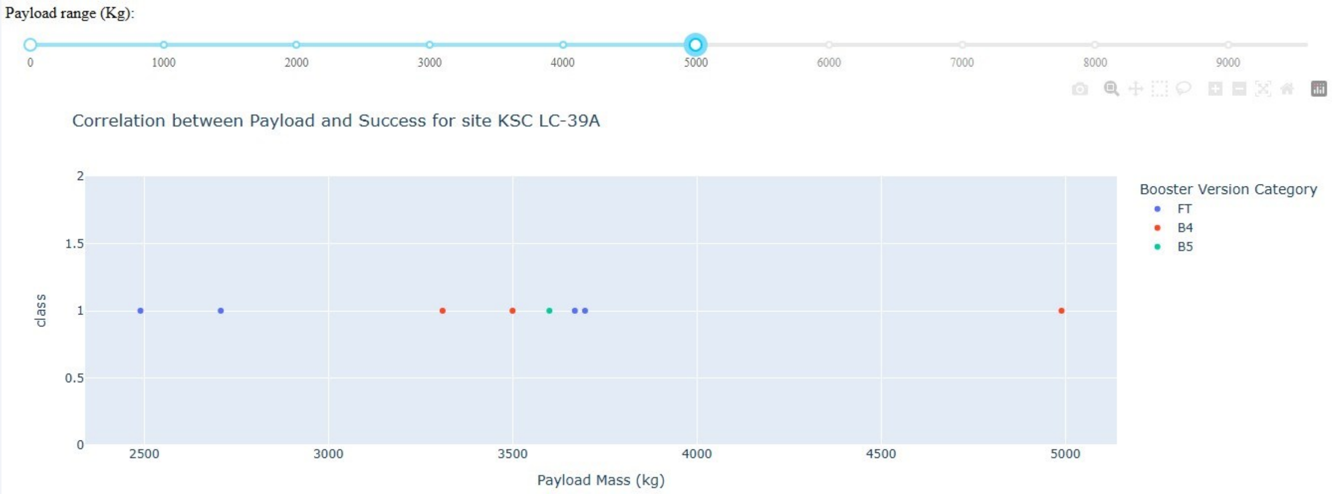
Total launches: 13

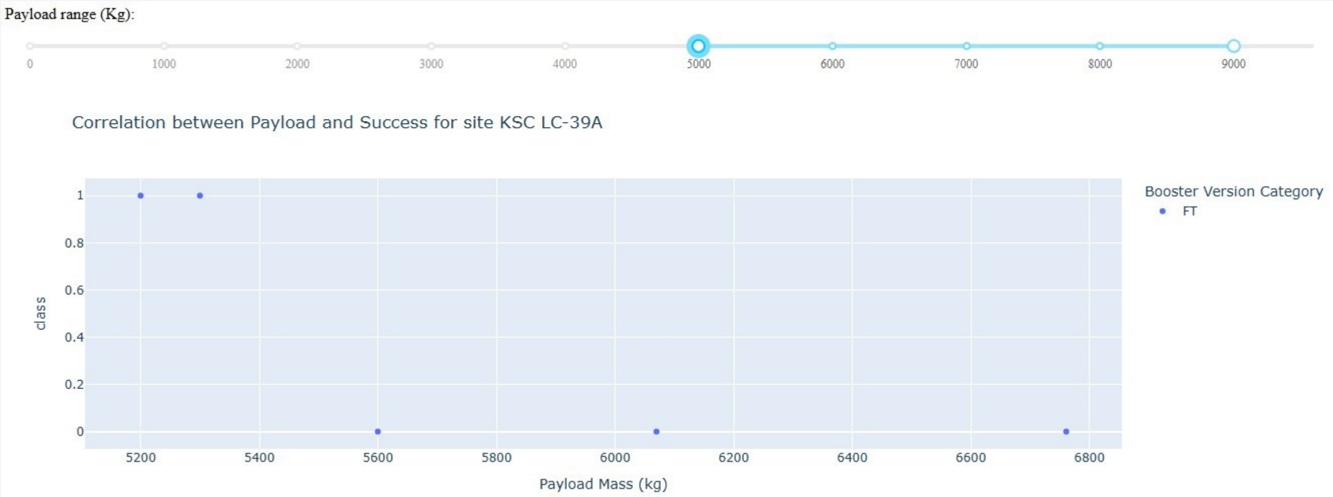
Successful landings: 10

Failed landings: 3

This chart is dynamically generated using filtered data

Correlation Between Payload Mass and Launch Success





Observations and Insights:

- First Chart (0–5000 kg Payload Range):

Success Rates: All the launches have successful outcomes (class = 1).

Booster Versions: FT, B4, B5 are represented in this range

- Second Chart (5000–9000 kg Payload Range):

Success Rates: Successes are less frequent compared to lighter payloads. Only 40% (2 of 5) of the launches were successful.

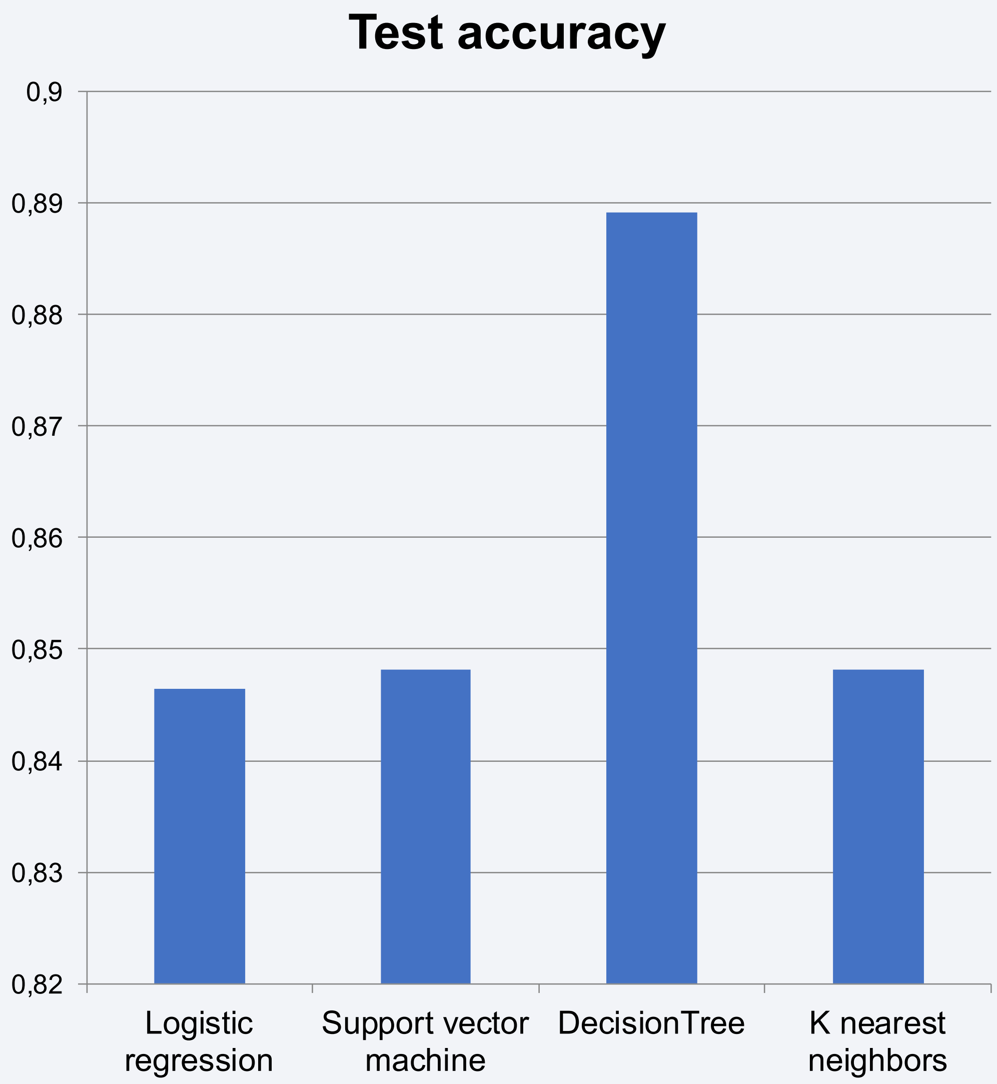
Booster Versions: FT only is represented in this mass range.

Trend: Lightweight payloads show a higher likelihood of success across all sites.

Section 5

Predictive Analysis (Classification)

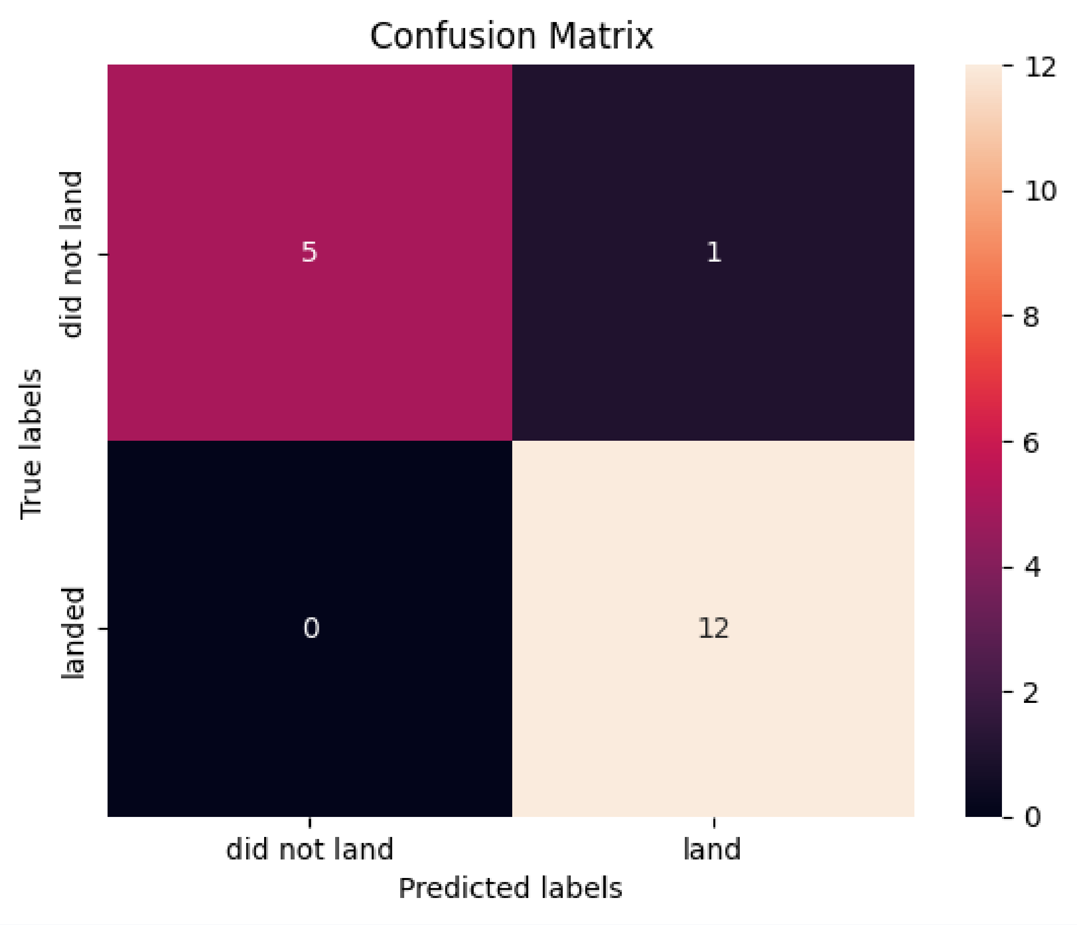
Classification Accuracy



Key Findings:

- Logistic regression : average performance: Test accuracy 0.8464. Best parameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
- Support Vector Machine : average performance: Test accuracy 0.8482. Best parameters: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
- Decision Tree : best performance: Test accuracy 0.8892. Best parameters: {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}
- K nearest neighbors : average performance: Test accuracy 0.8482. Best parameters: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

Confusion Matrix of the best model



Confusion Matrix of Decision Tree

Explanation :

1. True Positives (Bottom Right - 12): correctly predicted landings (successful outcomes). Represents 12 launches where the first stage successfully landed.
2. True Negatives (Top Left - 5): correctly predicted failures to land. Reflects 5 launches where the first stage did not land, as expected.
3. False Positives (Top Right - 1): incorrectly predicted a successful landing when it actually failed. Highlights one misclassified failed landing.
4. False Negatives (Bottom Left - 0): incorrectly predicted a failed landing when it actually succeeded. No false negatives, indicating the model performs exceptionally well in identifying successful landings.

Key Observations:

- High Precision: Only one false positive, indicating strong reliability in predicting successful landings.
- Perfect Recall for Successes: No false negatives, ensuring all actual landings are identified.
- Overall Performance: The model demonstrates robust classification ability for both successful and unsuccessful landings, making it the most effective among the tested models.

Conclusions

- 1. Orbit Success Trends: orbits ES-L1, GEO, HEO, SSO demonstrate 100% success rate.
- 2. Launch Success Rates: The KSC LC-39A site achieved the highest success rate of 76.9%. Proximity to the equator and coastlines contributes significantly to success.
- 3. Payload Impact: LEO and ISS orbits demonstrate high success rate for high Payload Mass launches.
- 4. Advancements Over Time: Success rates have shown consistent improvement since 2013, driven by SpaceX's technological advancements and operational experience.
- 5. Predictive Model Performance: Among tested models, Decision Tree outperformed others. The confusion matrix highlighted strong precision and recall, particularly for predicting successful landings.

Appendix

The complete codebase, data, and visualizations for this project can be found in the GitHub repository.

This repository contains:

- All Python scripts used for data processing, visualization, and modeling.
- Jupyter Notebooks documenting the analysis, SQL queries, and machine learning pipelines.

Thank you!