IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeffrey Rietzler
03.10.2025

Outline

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

Executive Summary

- **Summary of methodologies**
  - **Data Collection:** Retrieved SpaceX launch data via API and web scraping from Wikipedia.
  - **Data Wrangling:** Cleaned, formatted and stored data; prepared features for modeling.
  - **Exploratory Data Analysis (EDA):** Analyzed trends, patterns and key factors affecting launch outcomes.
  - **Interactive Visual Analytics and Dashboards:** Created maps of launch sites, visualized success rates with Folium and built Plotly Dash dashboards.
  - **Predictive Analysis (Classification):** Implemented Logistic Regression, SVM, Decision Tree and K-Nearest Neighbors models; tuned hyperparameters using GridSearchCV and evaluated model accuracy.
- **Summary of all results**
  - Analyzed factors affecting Falcon 9 first-stage landings and visualized success rates across different launch sites.
  - All models (Logistic Regression, SVM, K-Nearest Neighbors, Decision Tree) performed similarly with high accuracy.
  - Launch site and payload mass significantly impact landing success.

Introduction

- **Project background and context**

  The project focuses on predicting whether the Falcon 9 first stage will land successfully. Since SpaceX reuses the first stage, its launch cost is significantly lower (about $62M vs. $165M+ from other providers). Accurately predicting landing success helps estimate launch costs and provides valuable insights for competitors bidding against SpaceX. The task involves collecting launch data via an API, preparing it and building models for prediction.

- **Problems you want to find answers**
  - How do features such as payload mass, launch site, number of flights and orbit affect the success of the first stage landing?
  - Has the rate of successful landings increased over the years?
  - Which machine learning model is best suited for predicting the landing success (binary classification)?

Section 1

Methodology

Methodology

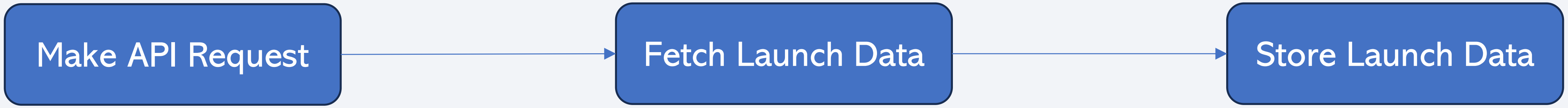
Executive Summary

- Data collection methodology:
    - Collect launch data using the SpaceX REST API
    - Supplement data using web scraping from Wikipedia
- Perform data wrangling
    - Filter the data to keep only the relevant features for analysis.
    - Handle missing values to ensure data consistency and accuracy.
    - Apply One-Hot Encoding to convert categorical variables into a numerical format suitable 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
    - Build, tune and evaluate classification models to identify the algorithm that delivers the most accurate predictions for the binary classification task.

Data Collection

Data was collected using both SpaceX REST API and web scraping from Wikipedia to obtain complete and detailed information on all launches for thorough analysis.

- Space X API Request

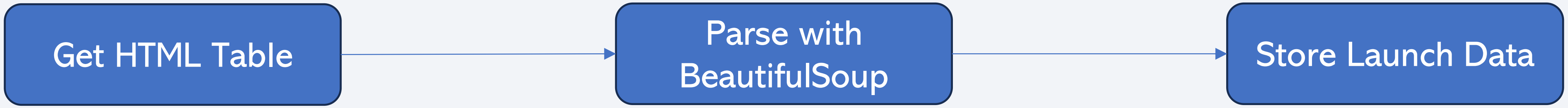


Make API Request

Fetch Launch Data

Store Launch Data

- Space X Web Scraping Wikipedia



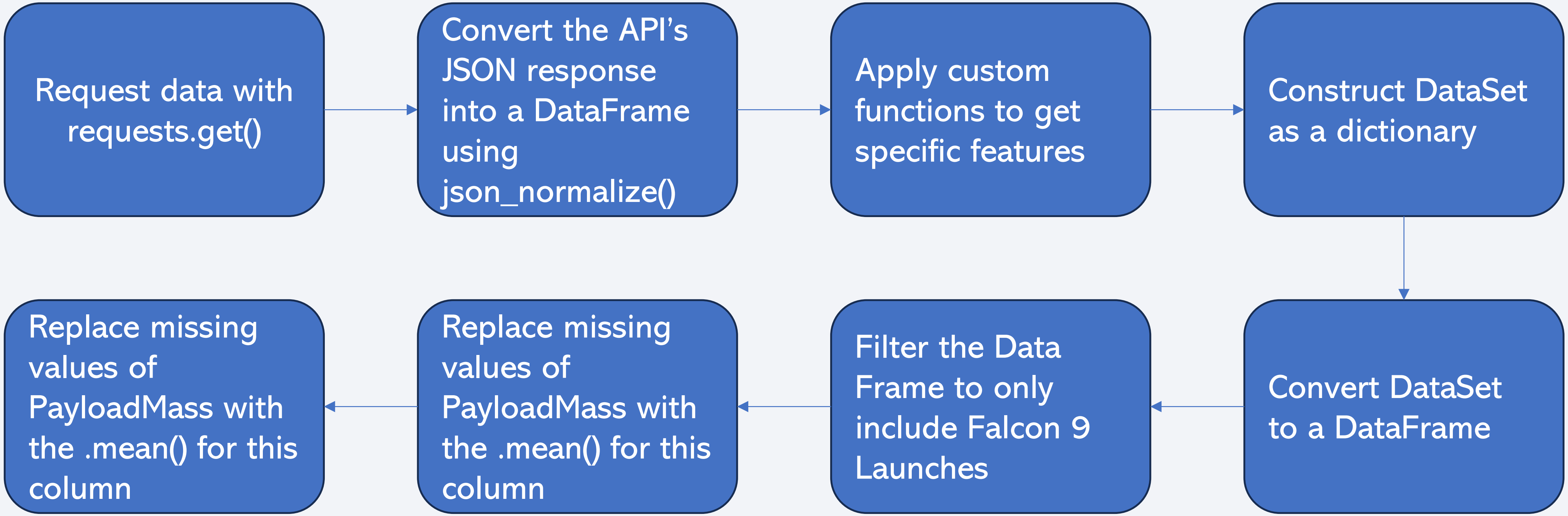
Get HTML Table

Parse with BeautifulSoup

Store Launch Data

Data Collection – SpaceX API

Launch data was retrieved from the SpaceX REST API using Python's requests library, parsed from JSON and key fields were extracted. The data was then stored in a pandas DataFrame and saved for further analysis.



Request data with requests.get()

Convert the API's JSON response into a DataFrame using json_normalize()

Apply custom functions to get specific features

Construct DataSet as a dictionary

Convert DataSet to a DataFrame

Filter the Data Frame to only include Falcon 9 Launches

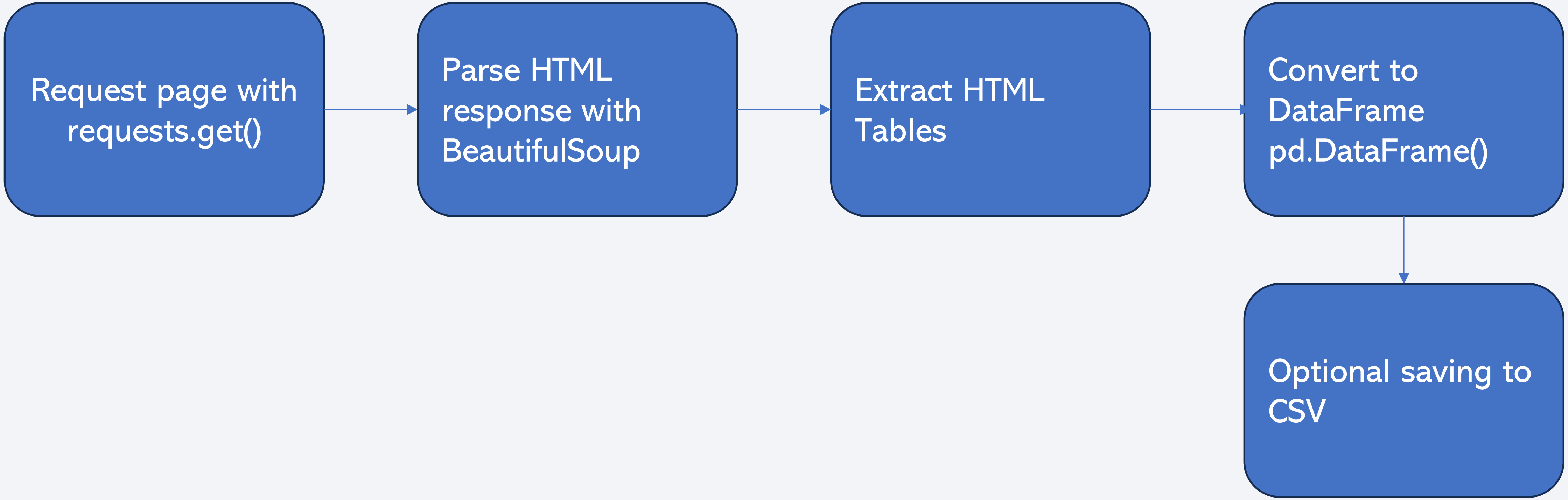
Replace missing values of PayloadMass with the .mean() for this column

Replace missing values of PayloadMass with the .mean() for this column

GitHub URL: jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Launch records were scraped from Wikipedia using Python's requests to fetch HTML and BeautifulSoup to parse it. The launch table was extracted and converted into a pandas DataFrame.



Request page with requests.get()

Parse HTML response with BeautifulSoup

Extract HTML Tables

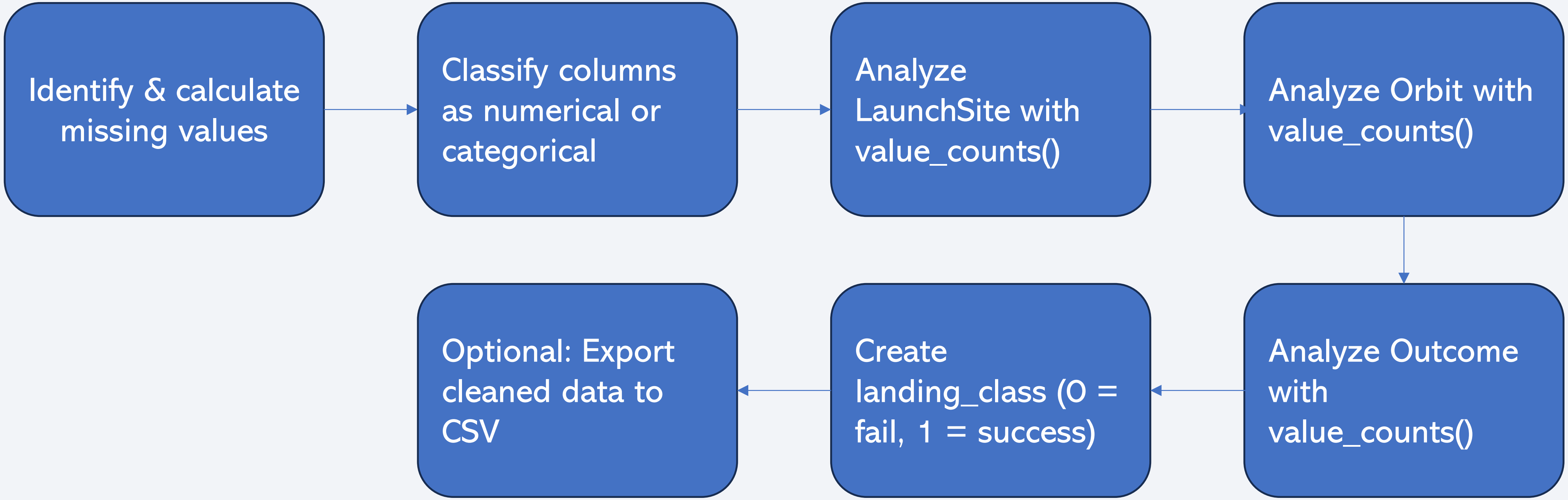
Convert to DataFrame pd.DataFrame()

Optional saving to CSV

GitHub URL: jupyter-labs-webscraping.ipynb

Data Wrangling

Launch records were cleaned and explored by handling missing values, classifying numerical and categorical data, analyzing launch sites, orbits, and outcomes, and creating a binary classification variable to represent landing success.



Identify & calculate missing values

Classify columns as numerical or categorical

Analyze LaunchSite with value_counts()

Analyze Orbit with value_counts()

Analyze Outcome with value_counts()

Create landing_class (0 = fail, 1 = success)

Optional: Export cleaned data to CSV

GitHub URL: labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

Several visualizations were created to explore relationships in the SpaceX dataset, including flight number vs. payload mass, flight number vs. launch site, payload mass vs launch site, orbit type vs success rate, flight number vs orbit type, payload mass vs orbit type and launch success yearly trend.

Chart Type Choice:

- Scatter plots are used to show the relationship between two variables. If a meaningful relationship exists, these patterns can later be leveraged in building machine learning models.
- Bar charts are used to compare values across discrete categories, helping to reveal differences or similarities between groups.
- Line charts are used to show trends over time, making it easier to identify patterns, fluctuations or long-term progressions in the data.

GitHub URL: [edadataviz.ipynb]

EDA with SQL

Performed SQL Queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- 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

GitHub URL: jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

An interactive map was developed to visualize SpaceX launch sites and their surroundings.

- **Markers with Labels and Popups:** Indicate the exact geographic location of each launch site.
- **Circles:** Highlight operational or safety zones around the launch sites.
- **Colored Markers:** Show launch outcomes—green for successful launches and red for failures—using clustering to reveal patterns across sites.
- **Lines:** Display distances from launch sites (e.g., KSC LC-39A) to nearby locations such as railways, highways, coastlines and cities, providing spatial context and relationships.

GitHub URL: lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Components Added

- Launch Sites Dropdown List: Allows selection of a specific launch site for focused analysis.
- Success Pie Chart: Displays overall successful launches for all sites or the success vs. failed launches for a selected site.
- Payload Mass Range Slider: Enables dynamic selection of payload mass ranges for exploration.
- Scatter Chart of Payload Mass vs. Success Rate: Shows the correlation between payload mass and launch success across different booster versions.

Purpose of Plots/Graphs

- Pie Chart: Visualizes the distribution of successful and failed launches, revealing overall success rates.
- Scatter Plot: Explores the relationship between payload mass and launch outcomes, helping identify potential trends or patterns.

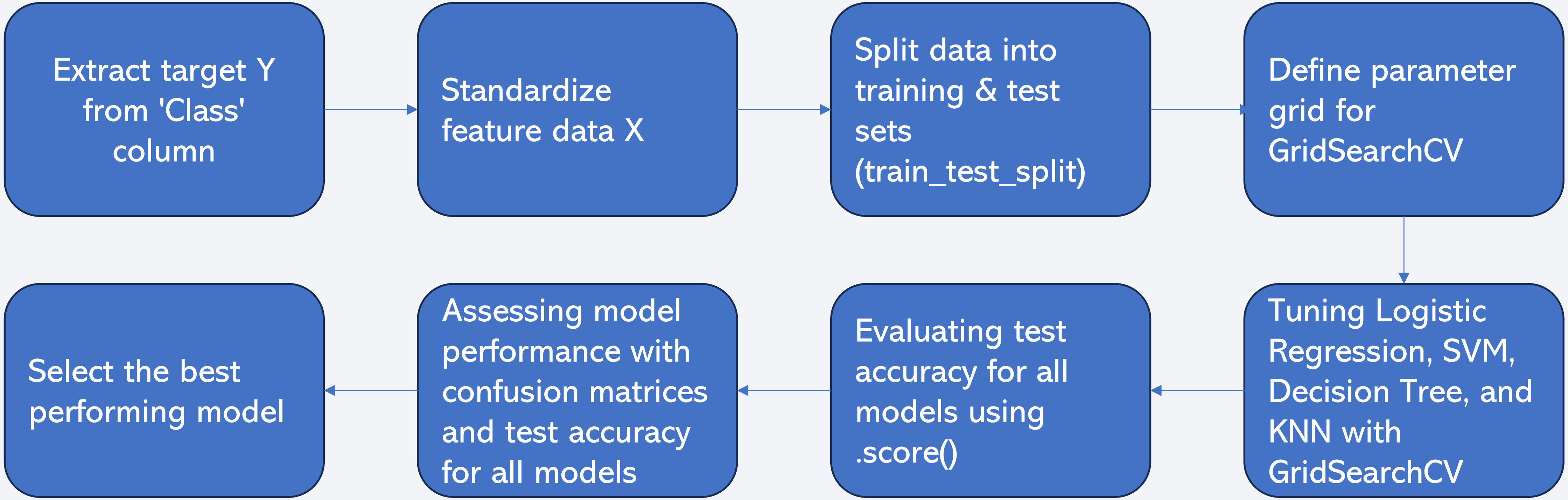
Interactions Added

- Launch Site Dropdown: Facilitates filtering by launch site for targeted analysis.
- Payload Range Slider: Allows dynamic adjustment of payload mass to examine its impact on launch success.

GitHub URL: spacex-dash-app.py

Predictive Analysis (Classification)

Features are standardized and the dataset is split into training and test sets. Multiple models are trained and tuned using cross-validation to find the best hyperparameters. The models are then evaluated on the test data to determine the one with the highest accuracy.



Extract target Y from 'Class' column

Standardize feature data X

Split data into training & test sets (train_test_split)

Define parameter grid for GridSearchCV

Tuning Logistic Regression, SVM, Decision Tree, and KNN with GridSearchCV

Evaluating test accuracy for all models using .score()

Assessing model performance with confusion matrices and test accuracy for all models

Select the best performing model

GitHub URL: SpaceX_Machine Learning Prediction_Part_5.ipynb

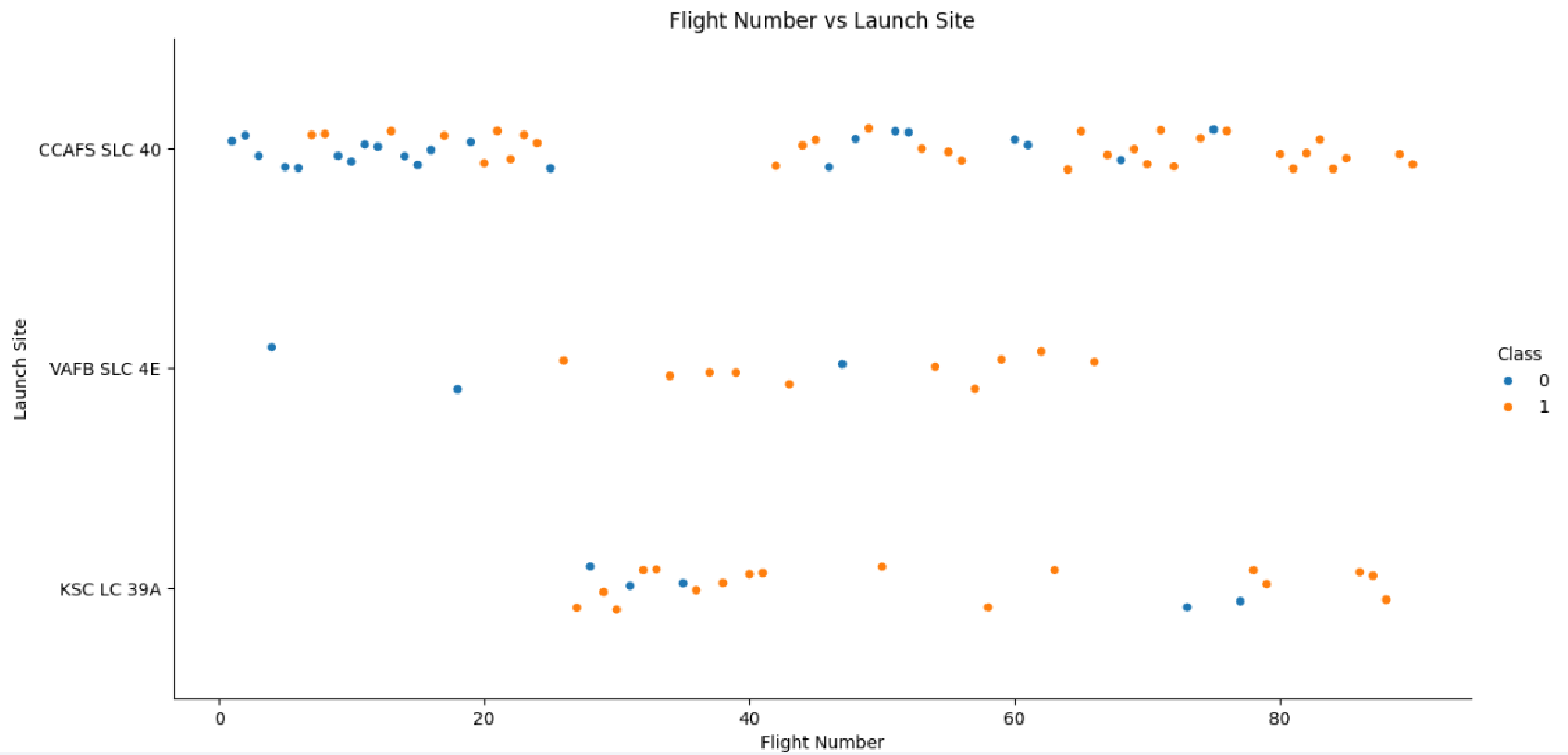
Results

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

Section 2

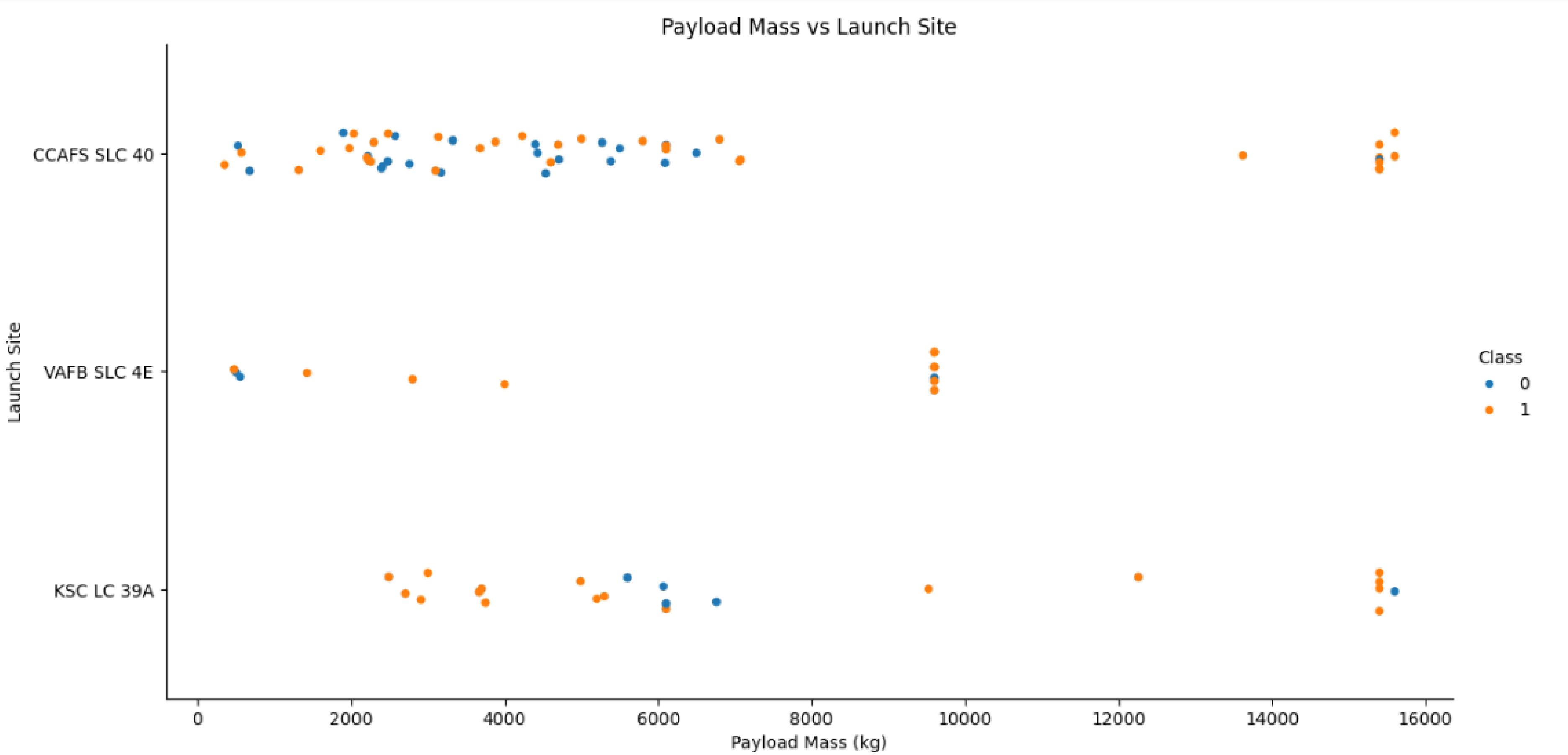
Insights drawn from EDA

Flight Number vs. Launch Site



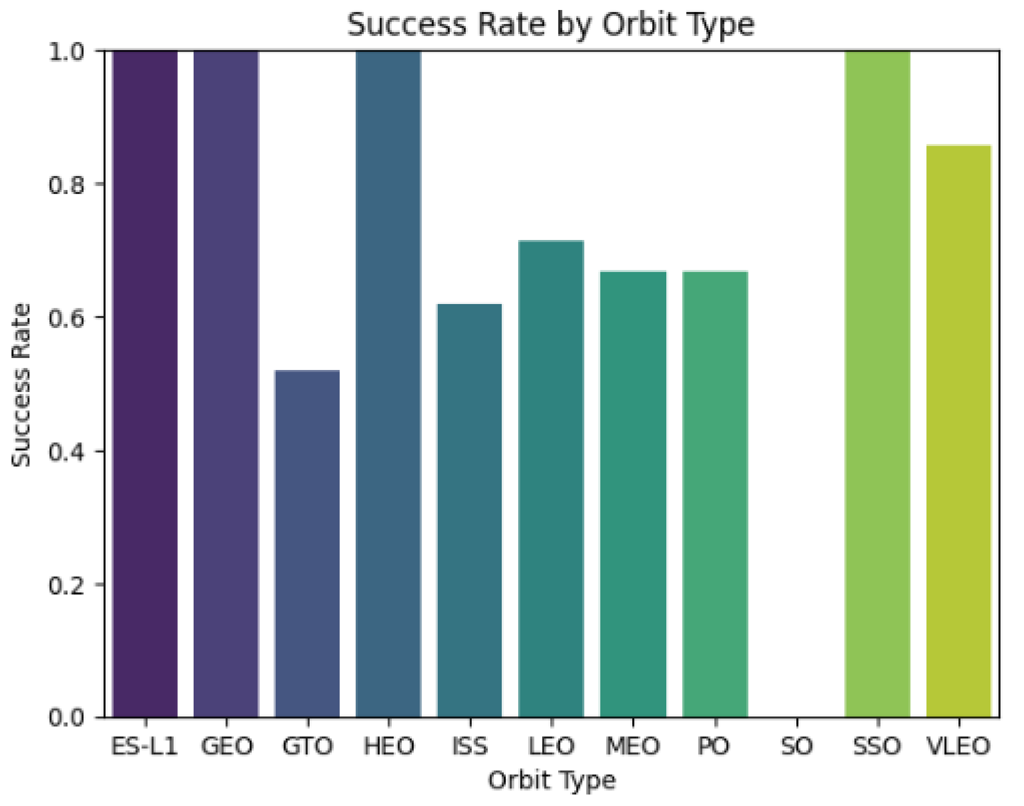
All launch sites experienced both successful and failed first-stage landings, with the success rate improving over time. The initial missions were mostly unsuccessful, suggesting that later advancements in technology and procedures played a key role.

Payload vs. Launch Site



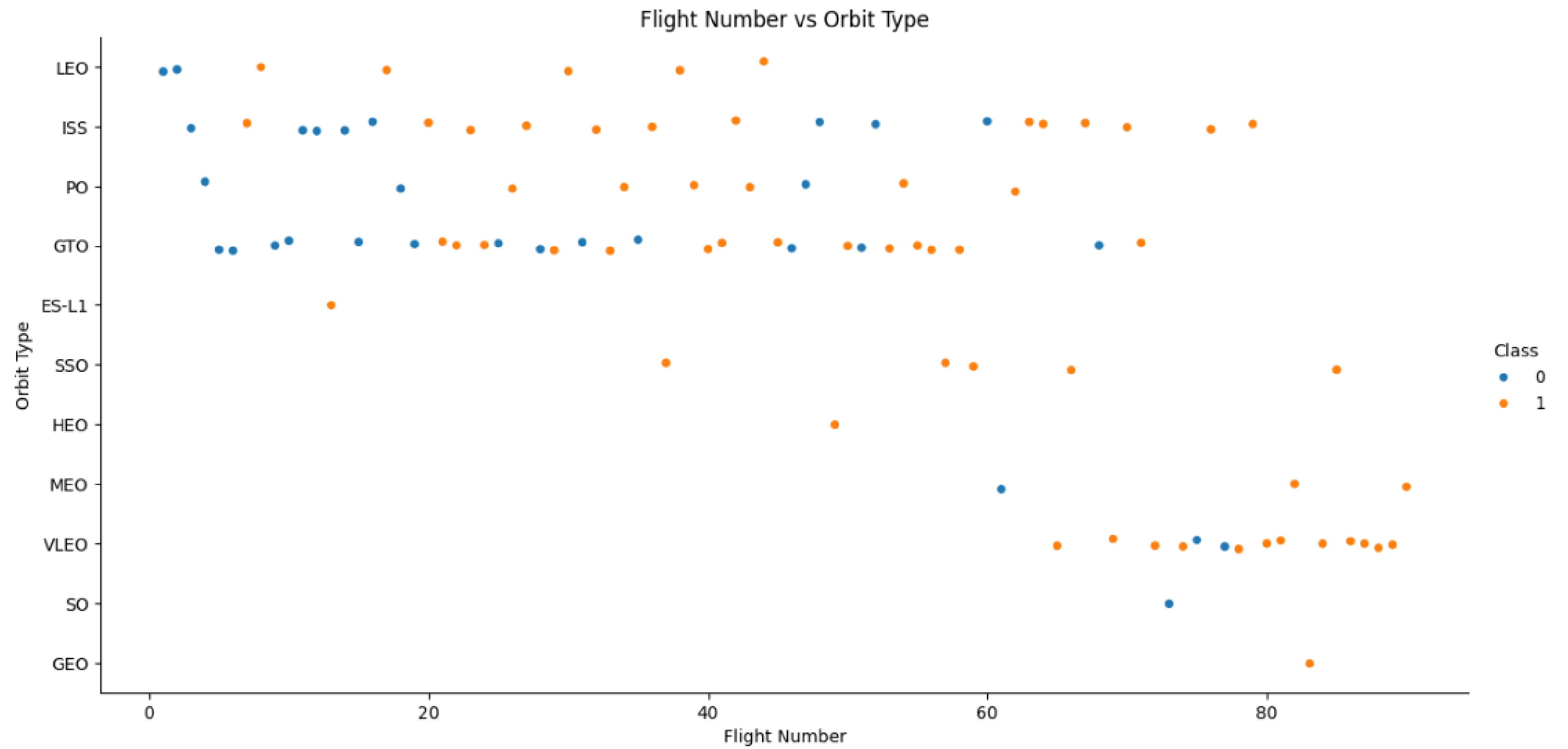
Across all sites, lighter payloads in early flights often failed, while heavier payloads especially those above 7,000 kg showed much higher landing success rates, reflecting technological and operational progress.

Success Rate vs. Orbit Type



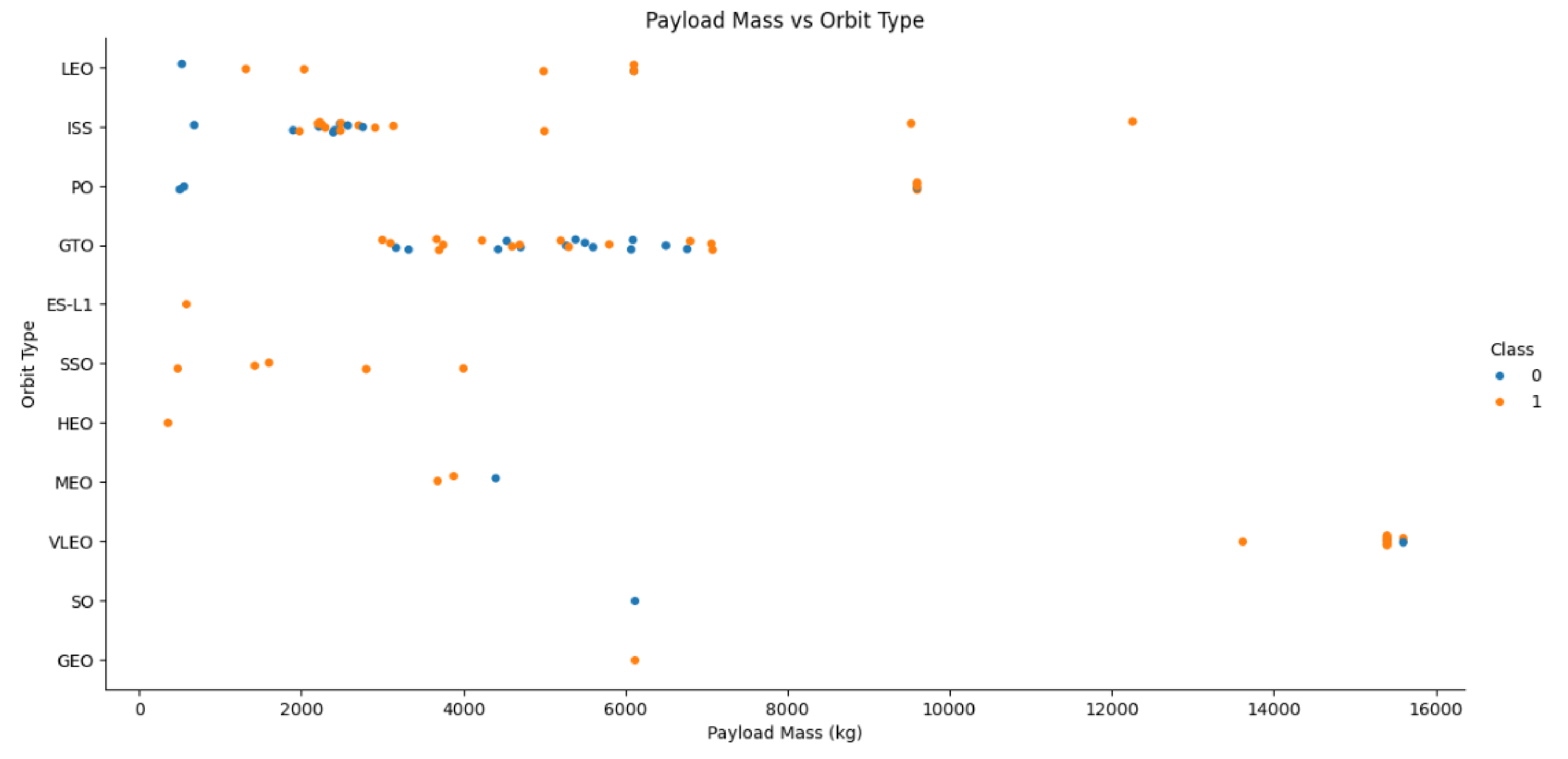
100% success: ES-L1, GEO, HEO, SSO
0% success: SO
50–85% success: GTO, ISS, LEO, MEO, P

Flight Number vs. Orbit Type



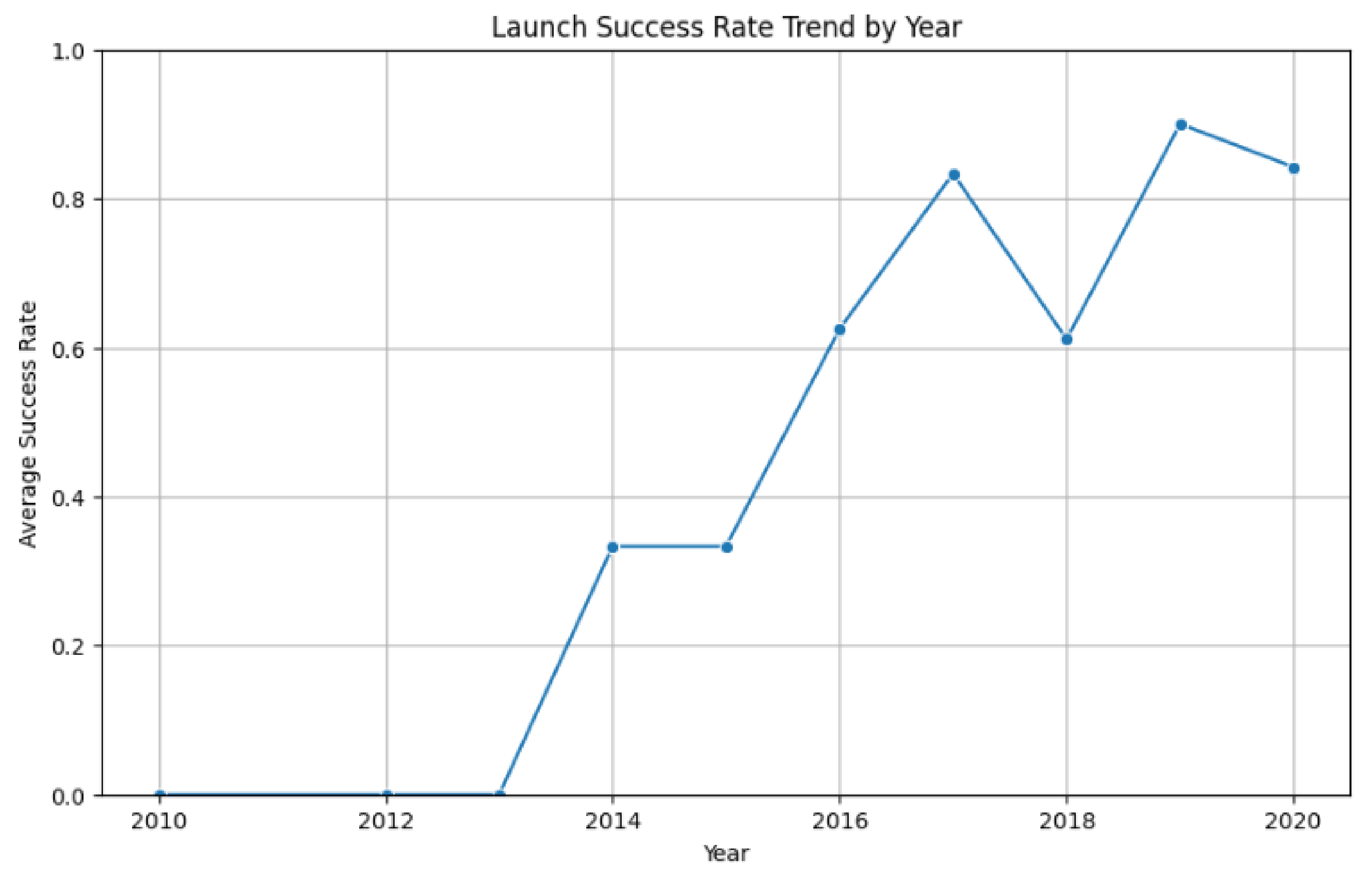
Multiple orbits appear across the flight history, with some only attempted in later missions.
Landing success improves with higher flight numbers, reflecting growing experience and refinements.

Payload vs. Orbit Type



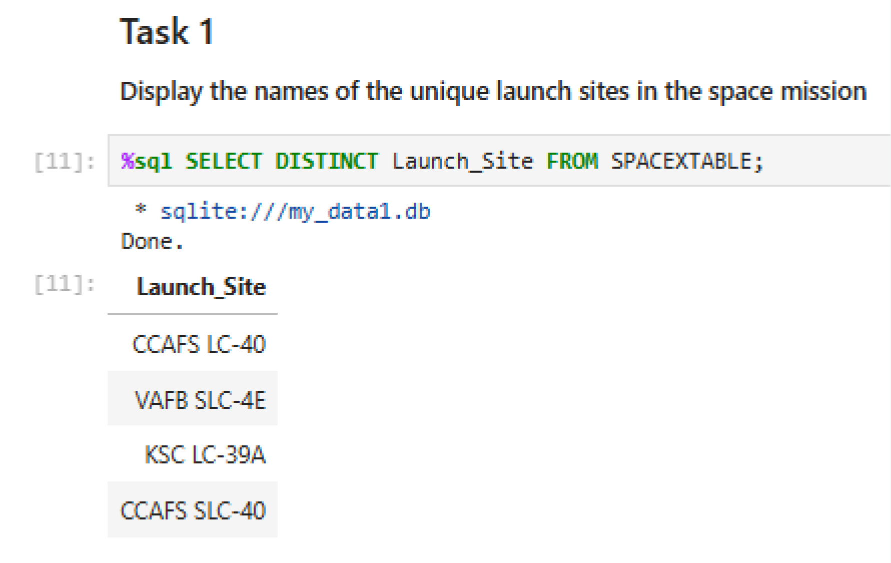
Many orbits cover a broad payload range, while SSO, MEO, HEO, and GEO are narrower.
Limited payload variation often aligns with higher landing success.
Payload mass impacts success differently by orbit: negative for GTO, positive for GTO and Polar LEO (ISS).

Launch Success Yearly Trend



The yearly trend shows steady progress from early difficulties to high landing reliability.
Since 2016, success rates have improved annually, with only a small dip in 2018.

All Launch Site Names



Task 1

Display the names of the unique launch sites in the space mission

```
[11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

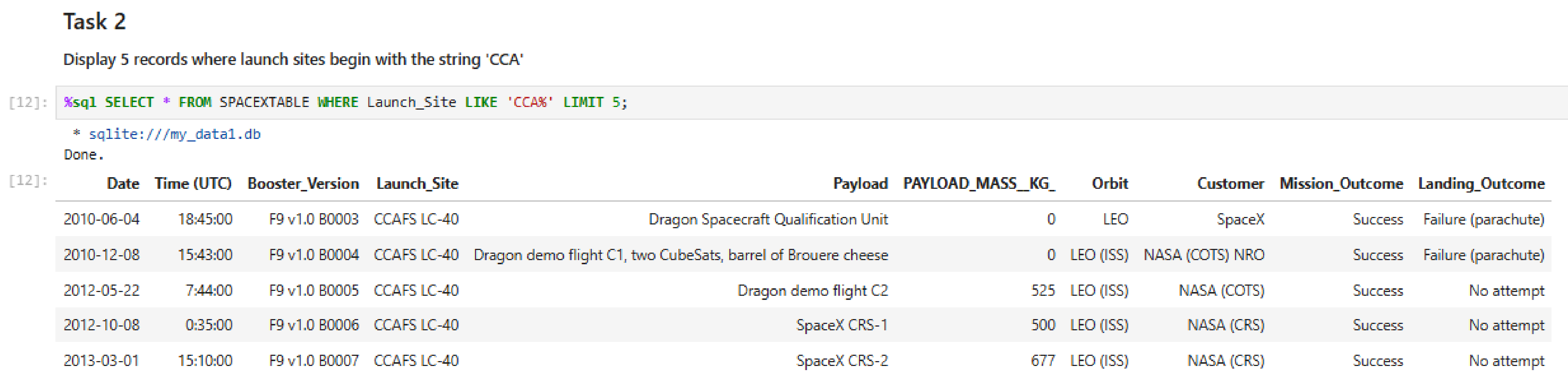
```
 * sqlite:///my_data1.db
Done.
```

[11]:

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

The space missions use four distinct launch sites as listed above.

Launch Site Names Begin with 'CCA'



Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
[12]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

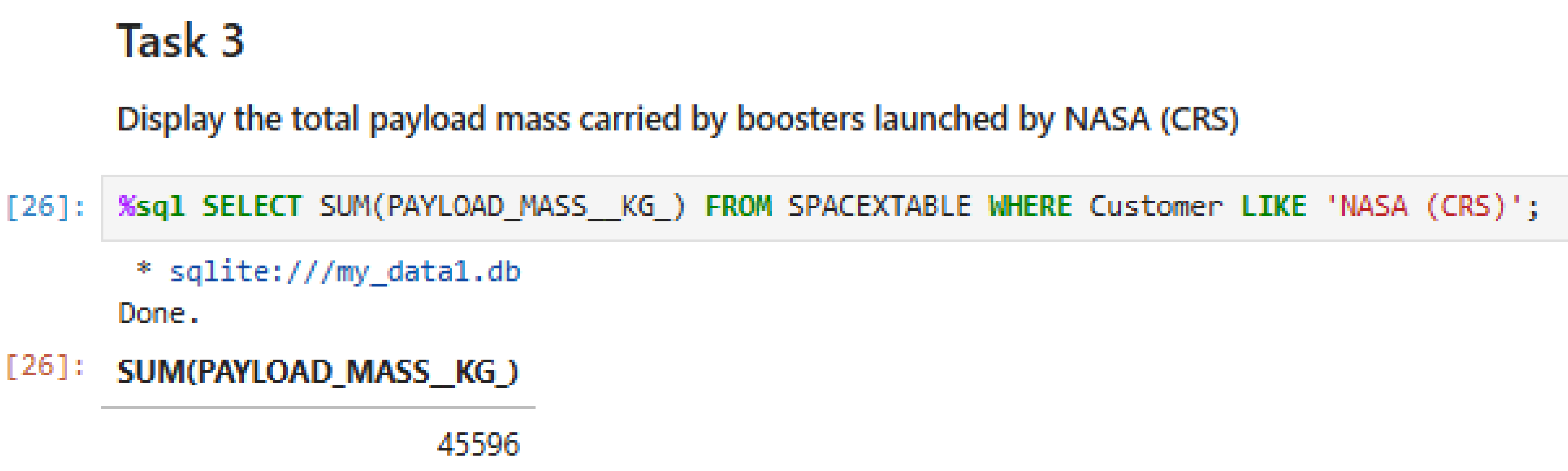
```
 * sqlite:///my_data1.db
Done.
```

[12]:

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

Display the first five records where the launch site name starts with ‘CCA’.

Total Payload Mass



Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[26]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)';
```

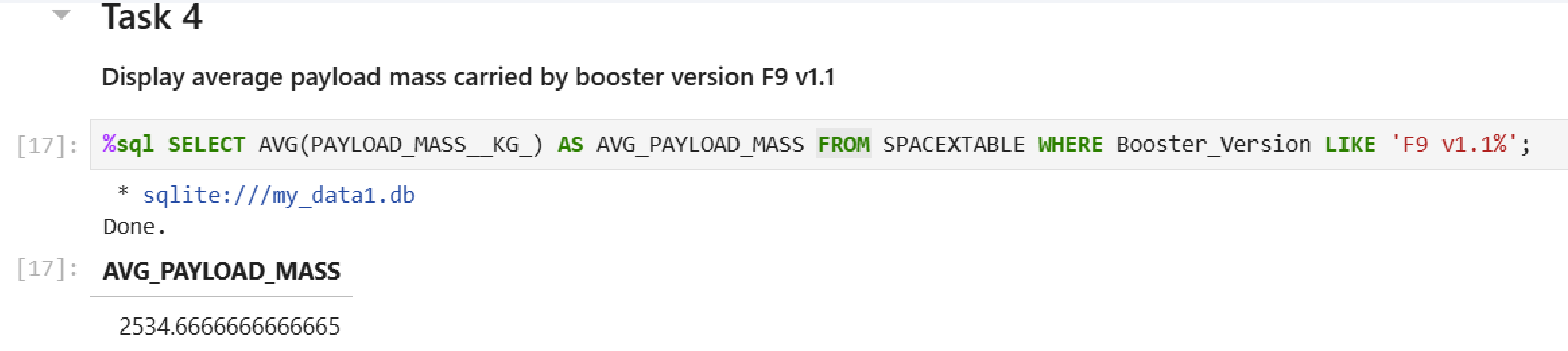
```
 * sqlite:///my_data1.db
Done.
```

[26]:

SUM(PAYLOAD_MASS__KG_)
45596

The cumulative payload delivered by NASA (CRS) boosters amounts to 45,596 kg.

Average Payload Mass by F9 v1.1



Task 4

Display average payload mass carried by booster version F9 v1.1

```
[17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
```

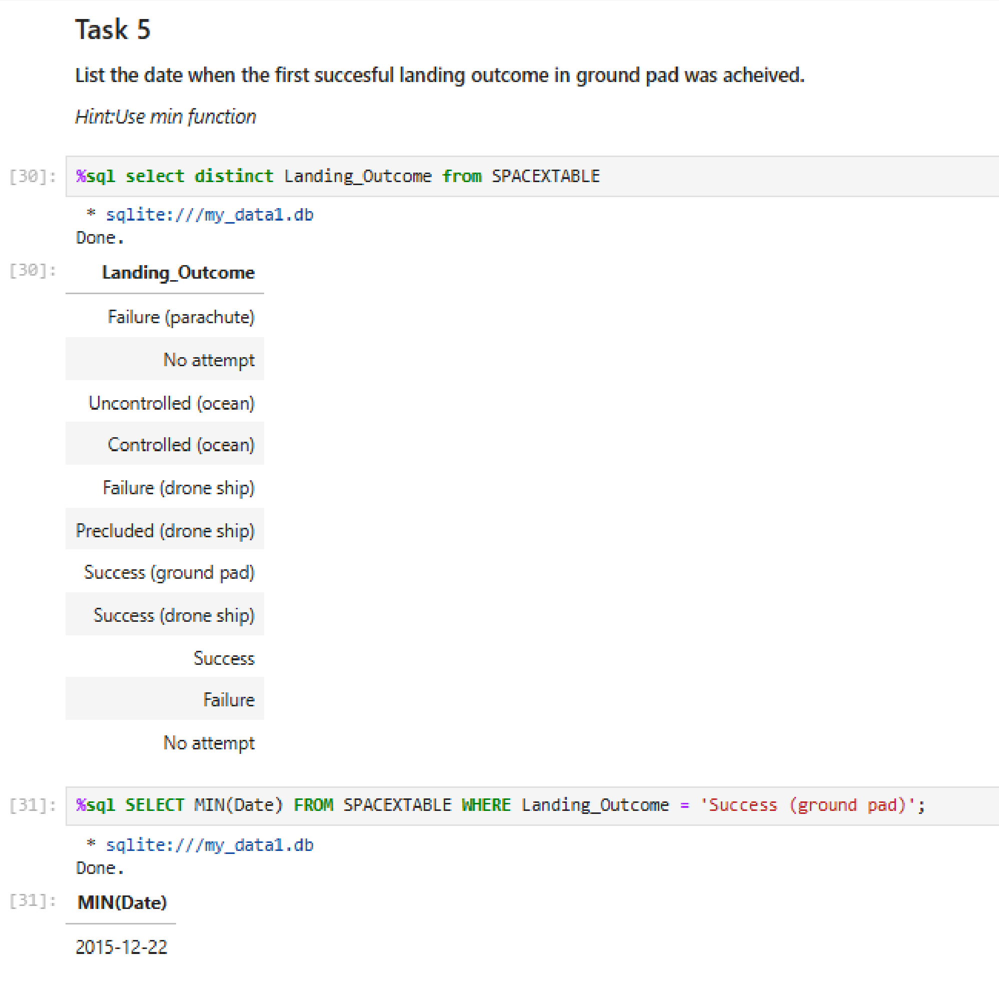
```
 * sqlite:///my_data1.db
Done.
```

[17]:

AVG_PAYLOAD_MASS
2534.6666666666665

On average, the F9 v1.1 booster carries a payload of 2,534.67 kg.

First Successful Ground Landing Date



Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[30]: %sql select distinct Landing_Outcome from SPACEXTABLE
```

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

[30]:

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

```
[31]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

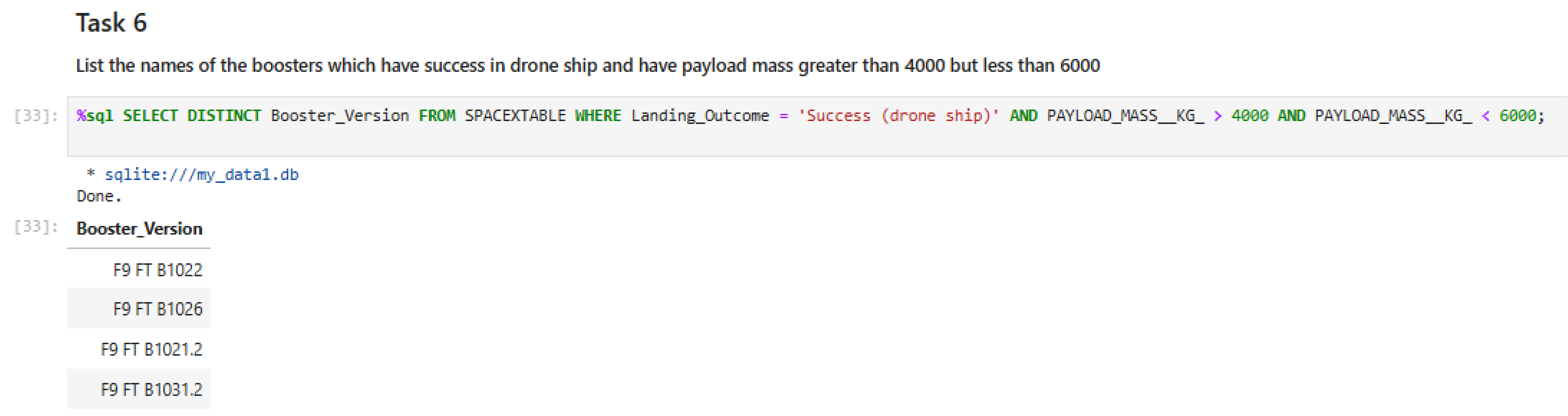
```
 * sqlite:///my_data1.db
Done.
```

[31]:

MIN(Date)
2015-12-22

December 22, 2015, marks the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000



Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[33]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

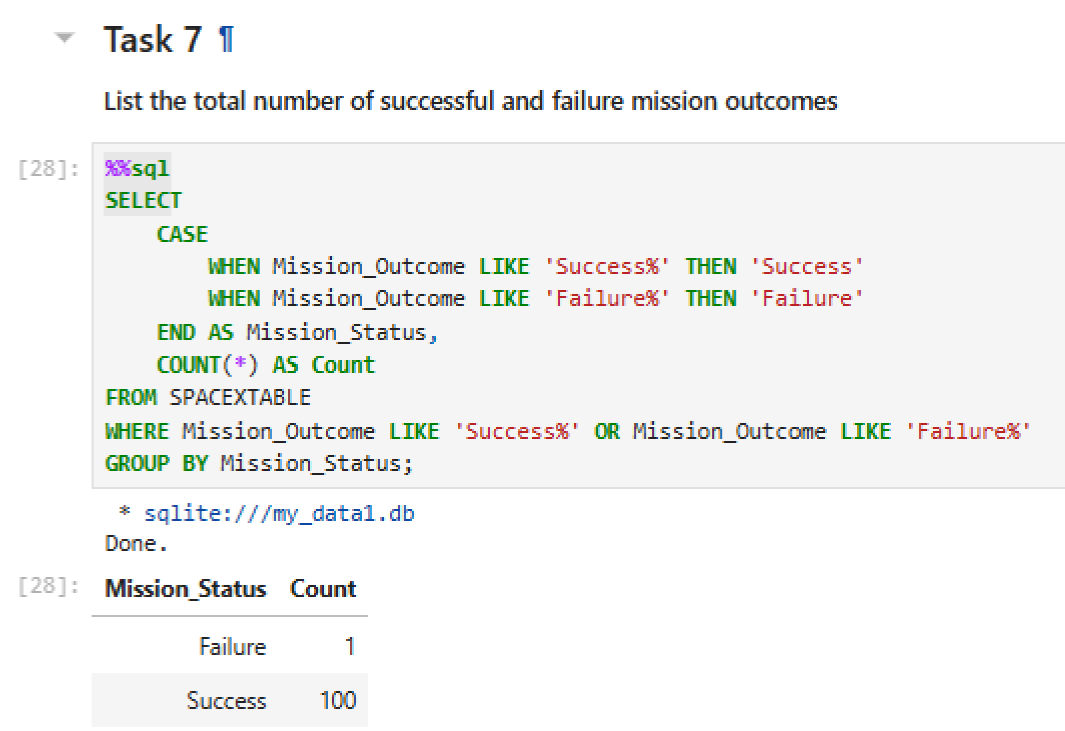
```
 * sqlite:///my_data1.db
Done.
```

[33]:

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

The four listed boosters achieved successful drone ship landings while carrying payloads in the 4,000–6,000 kg range

Total Number of Successful and Failure Mission Outcomes



Task 7 ¶

List the total number of successful and failure mission outcomes

```
[28]: %%sql
      SELECT
          CASE
              WHEN Mission_Outcome LIKE 'Success%' THEN 'Success'
              WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
          END AS Mission_Status,
          COUNT(*) AS Count
      FROM SPACEXTABLE
      WHERE Mission_Outcome LIKE 'Success%' OR Mission_Outcome LIKE 'Failure%'
      GROUP BY Mission_Status;
```

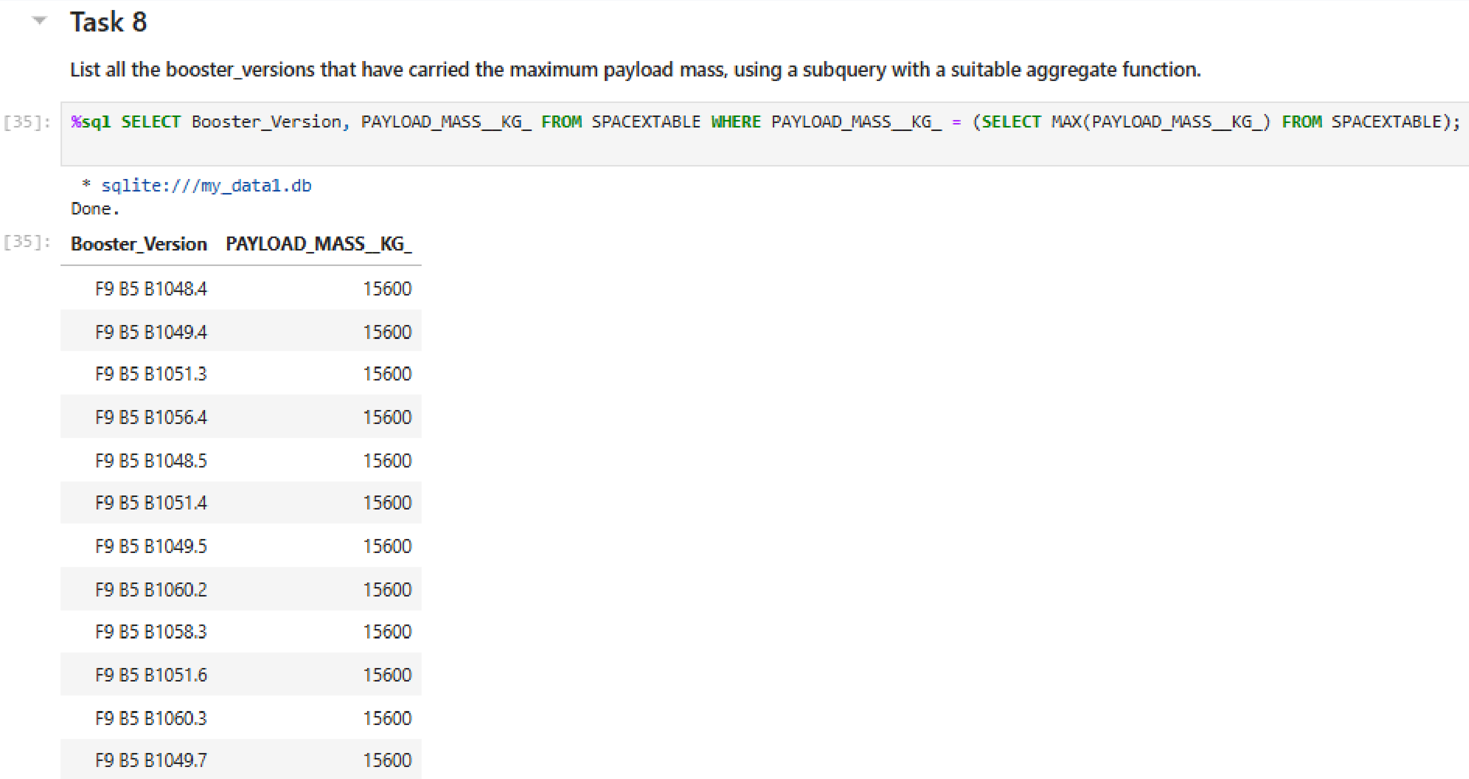
```
 * sqlite:///my_data1.db
Done.
```

[28]:

Mission_Status	Count
Failure	1
Success	100

There was 1 failed mission and 100 successful missions.

Boosters Carried Maximum Payload



Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
[35]: %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

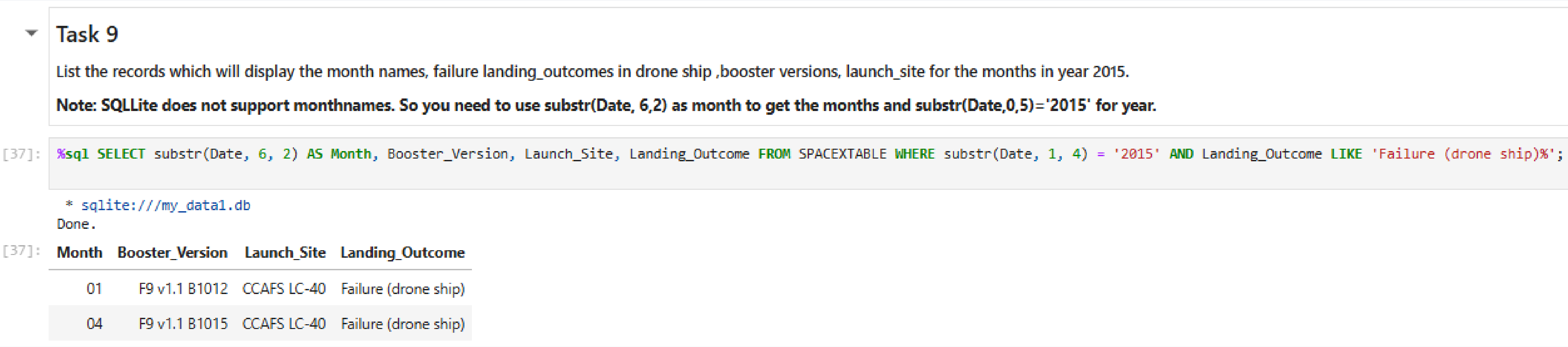
```
 * sqlite:///my_data1.db
Done.
```

[35]:

Booster_Version	PAYLOAD_MASS__KG_
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

The maximum payload of 15,600 kg was carried by the boosters listed above.

2015 Launch Records



Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

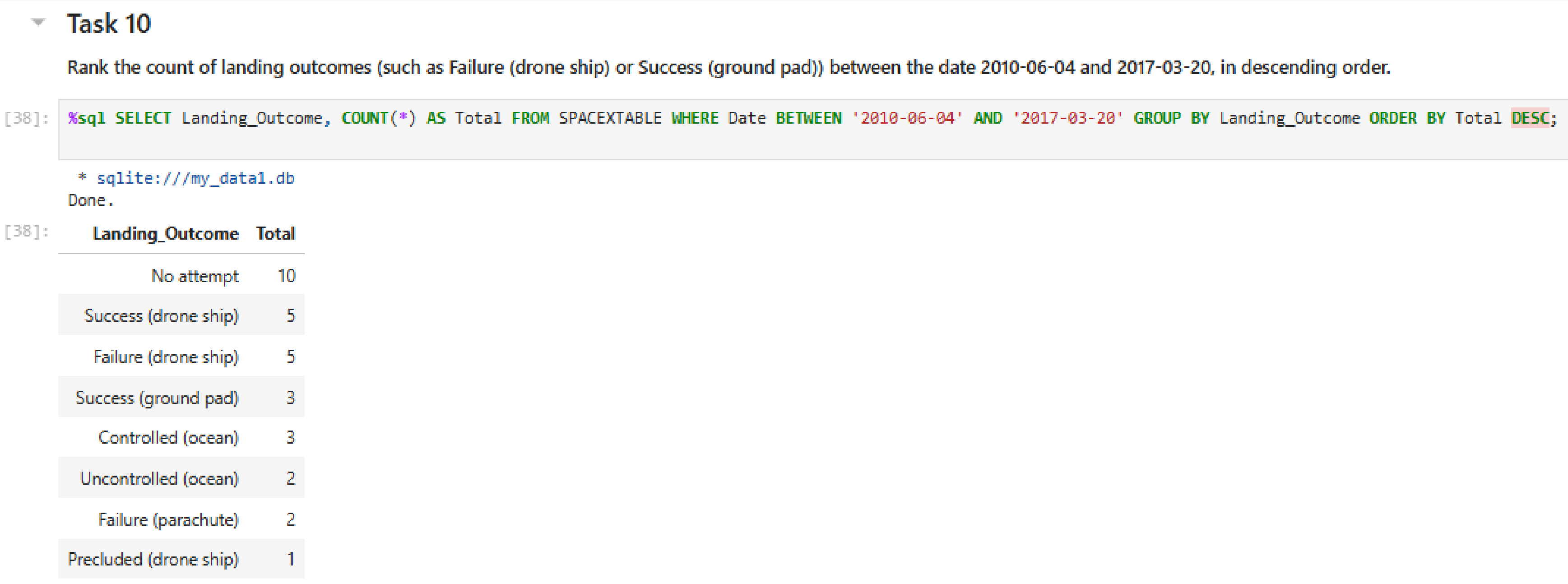
```
%sql SELECT substr(Date, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome LIKE 'Failure (drone ship)%';
```

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

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

In 2015, there were two failed landings: one in January and one in April.

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



Task 10

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.

```
[38]: %sql SELECT Landing_Outcome, COUNT(*) AS Total FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Total DESC;
```

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

[38]:

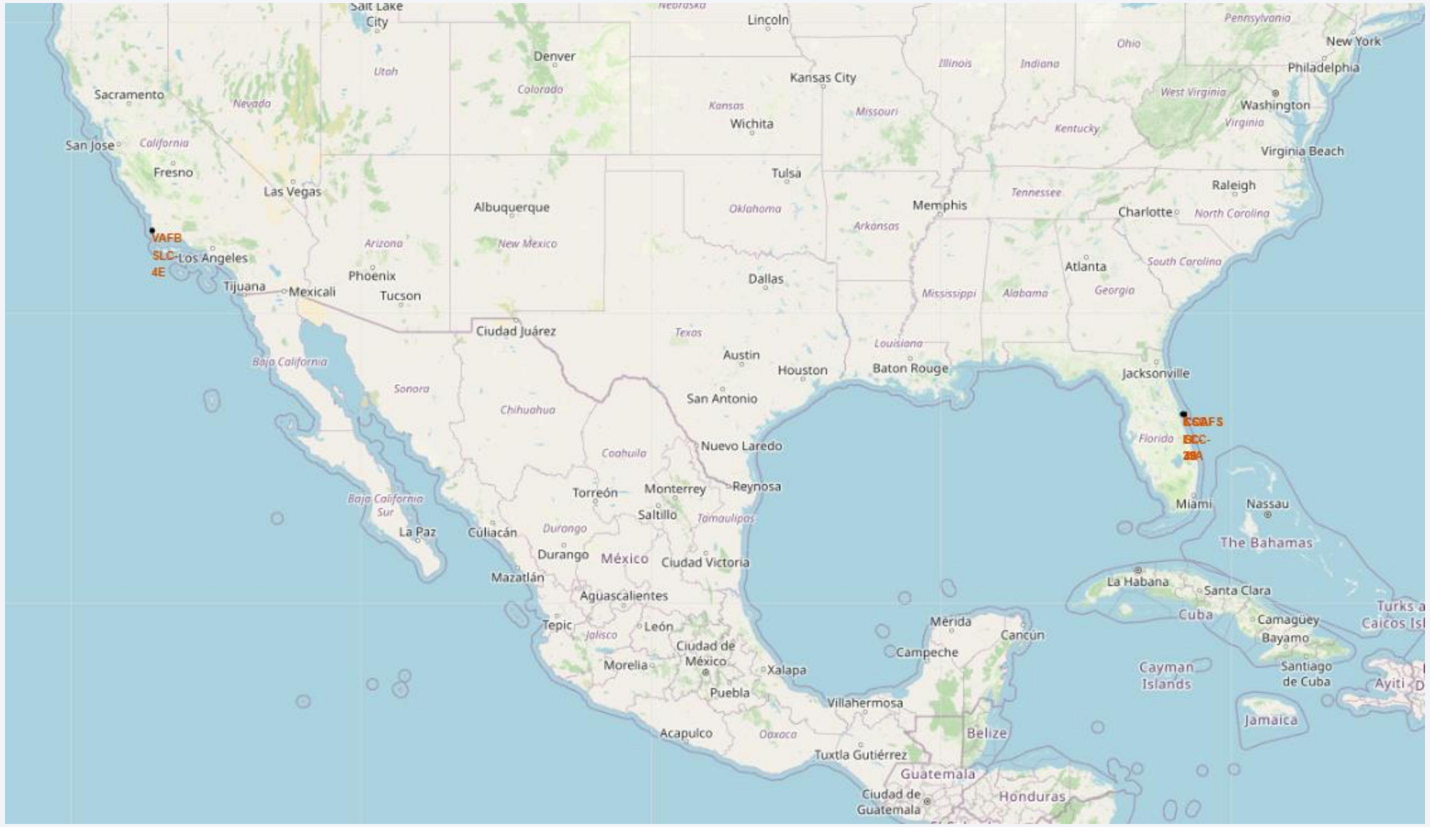
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Landing outcomes were mostly drone ship or ground pad successes/failures, with few ocean or parachute cases.

Section 3

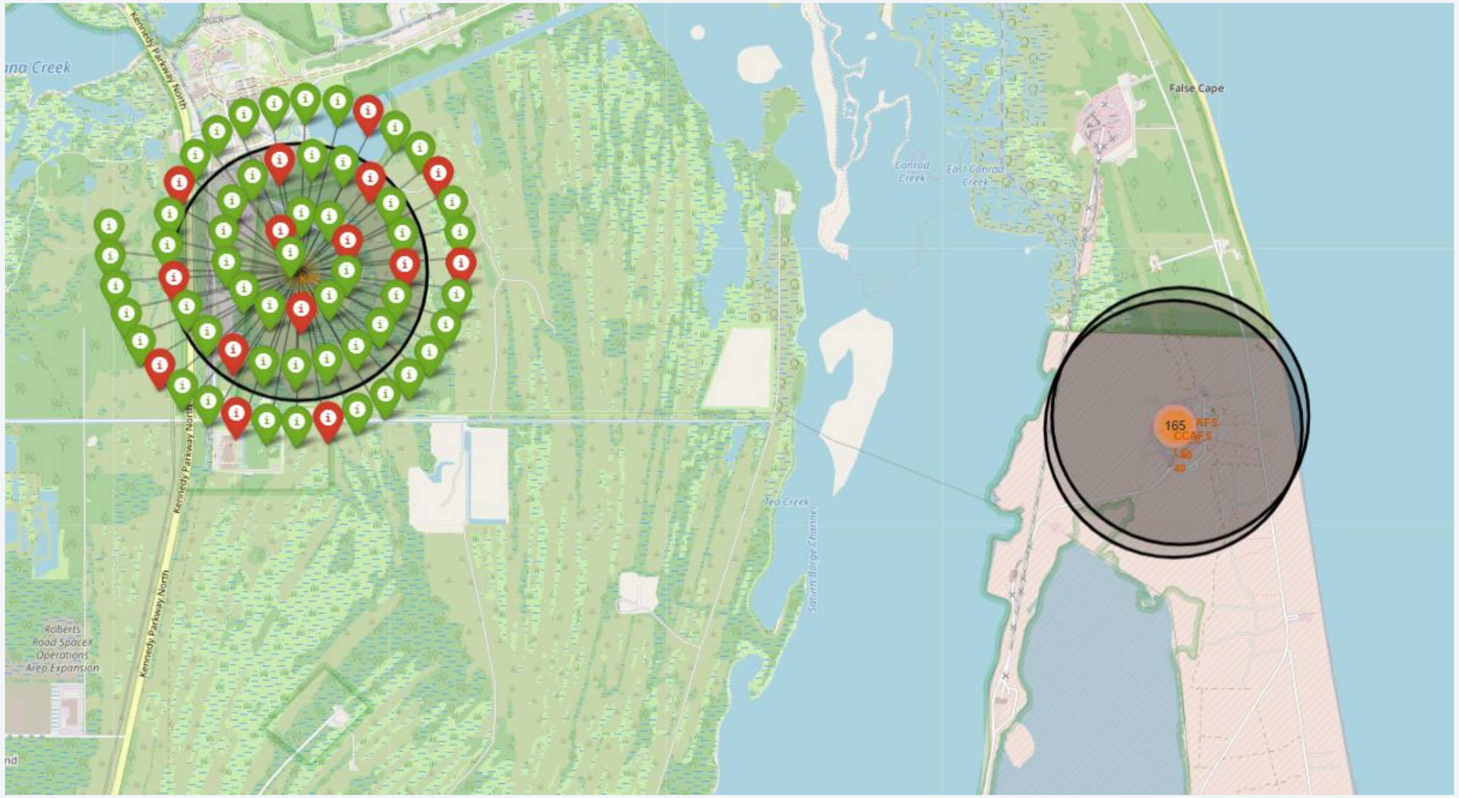
Launch Sites Proximities Analysis

All launch sites on a map



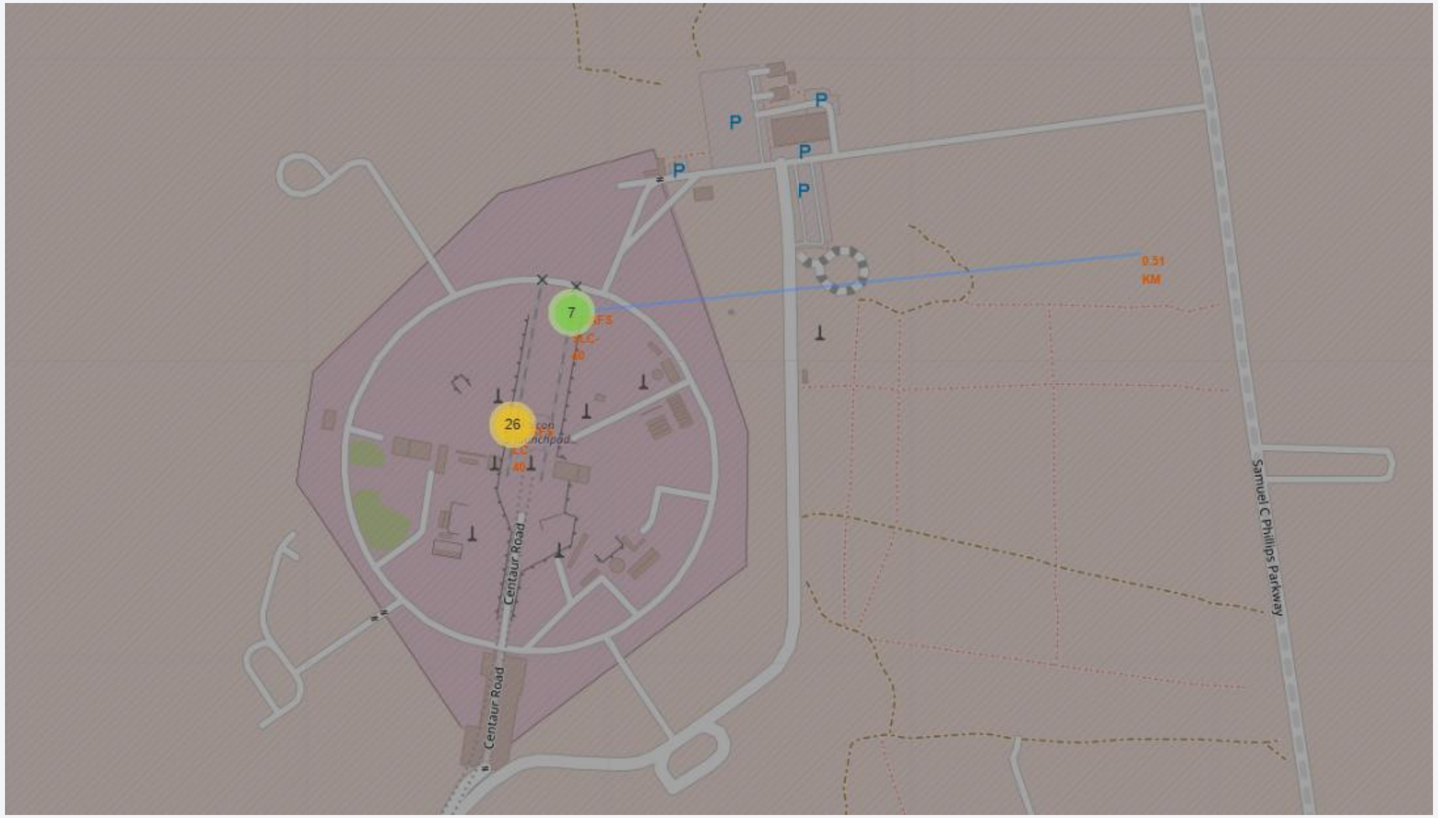
Launch sites in Florida and California are near the coast to minimize risk to populated areas in case of failures.

Successful/failed launches for each site on the map

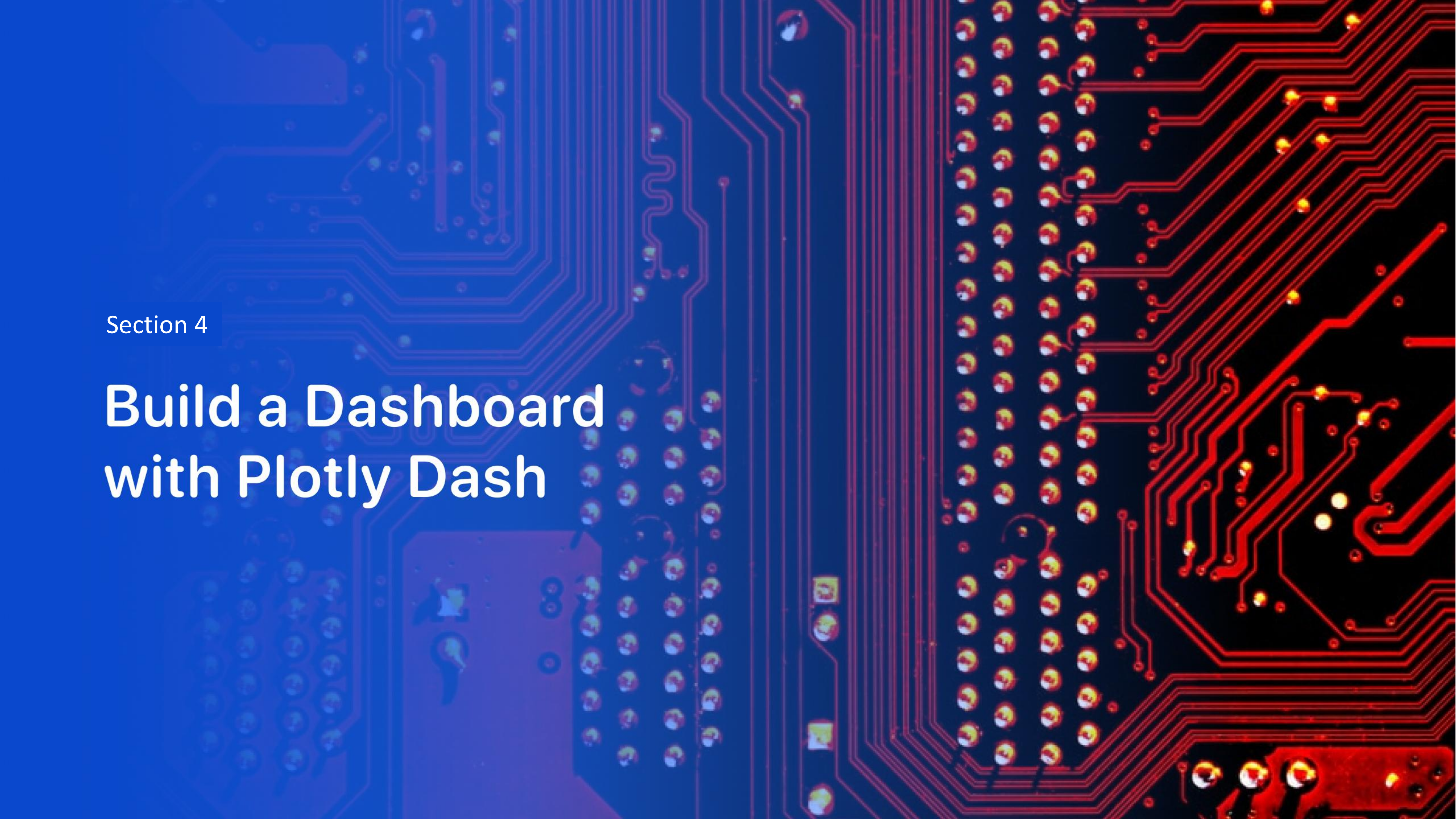


Color-coded markers highlight launch success: green for success, red for failure.

Distances between a launch site to its surrounding proximities



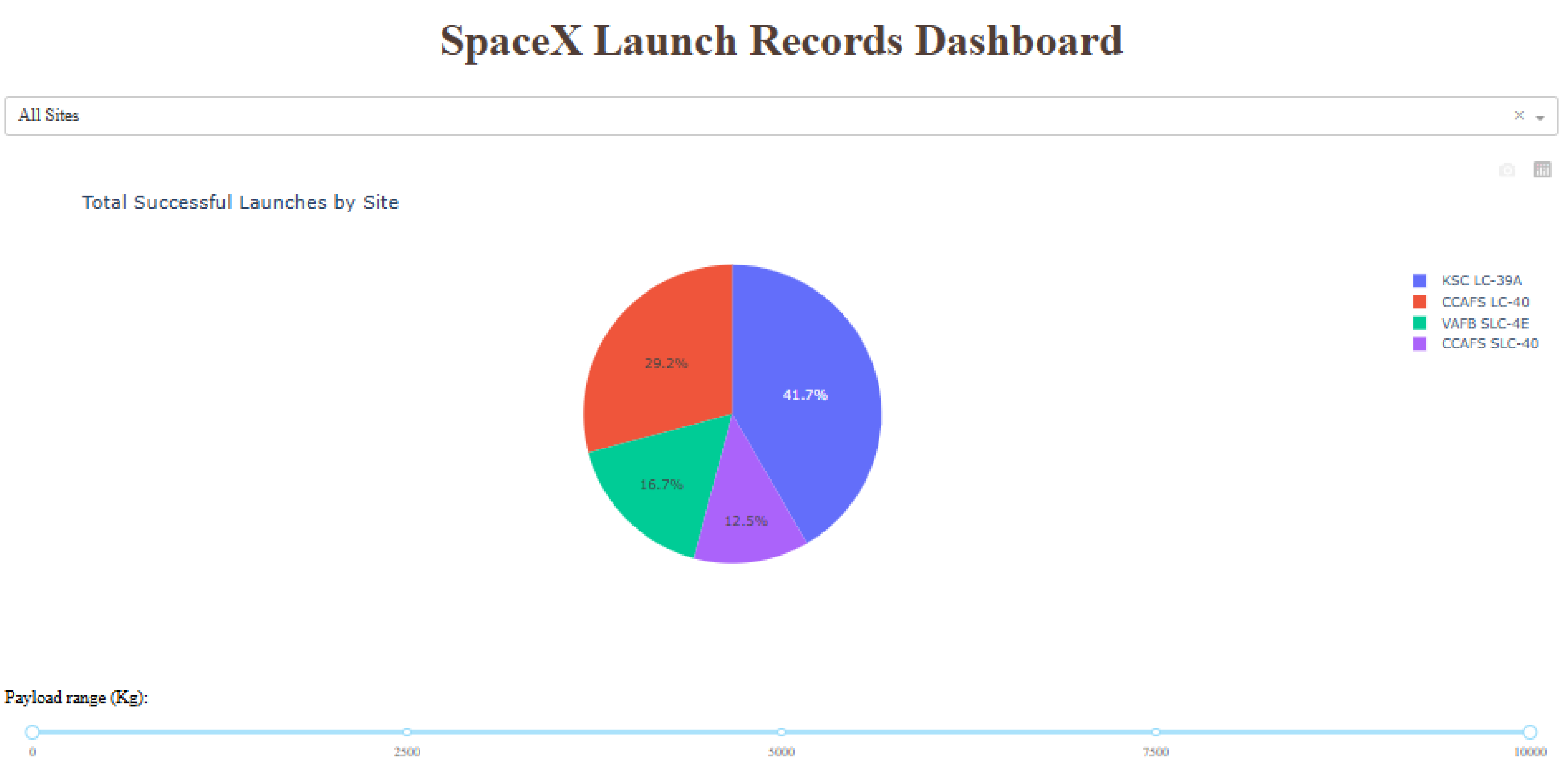
Proximity to railways, highways and the coastline supports logistical access and reduces risk to populated areas near the launch sites.



Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites



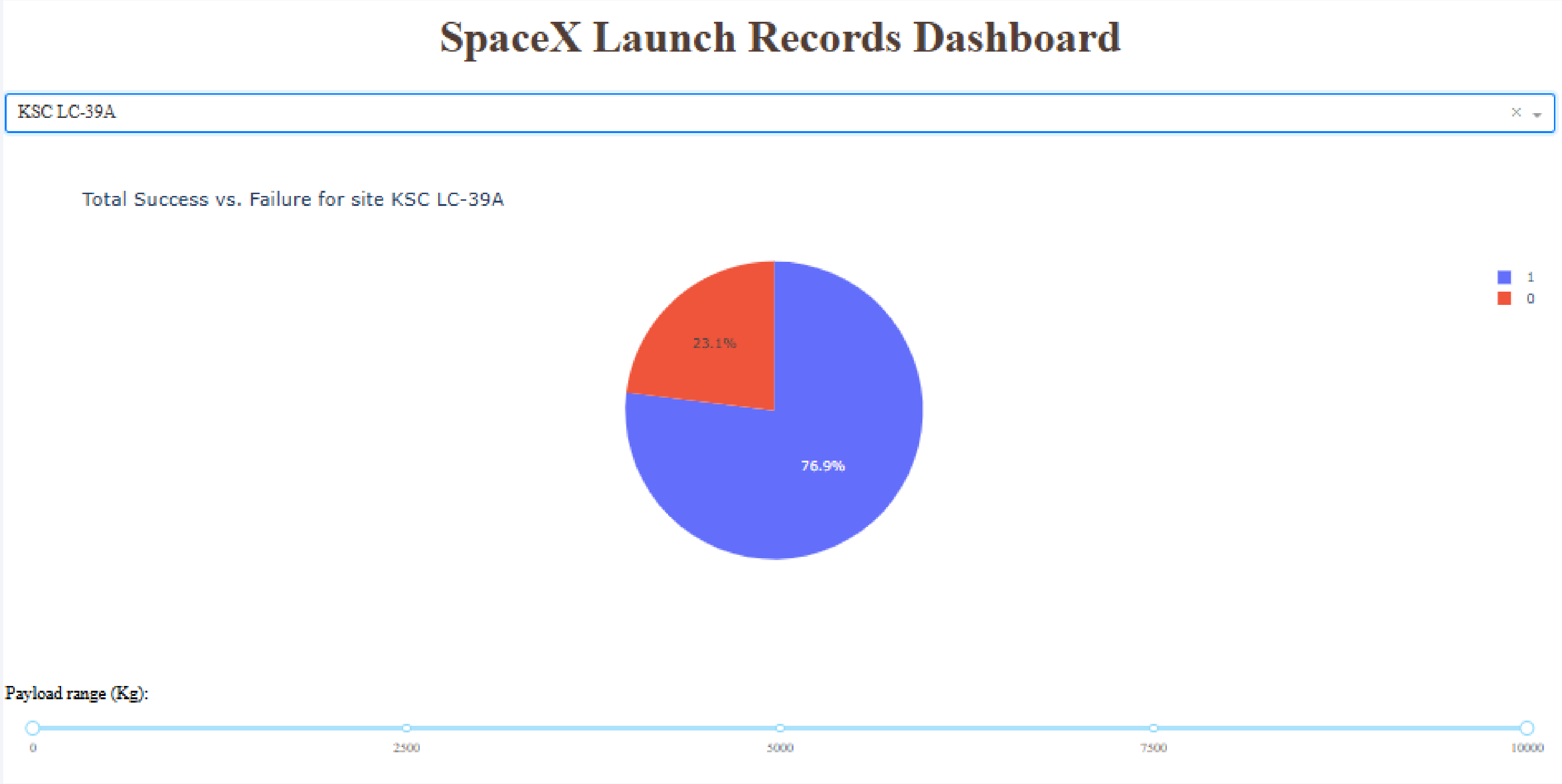
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

KSC LC-39A had the highest landing success rate, followed by CCAFS LC-40, while VAFB SLC-4E and CCAFS SLC-40 had the lowest.

Launch site with the highest launch success rate



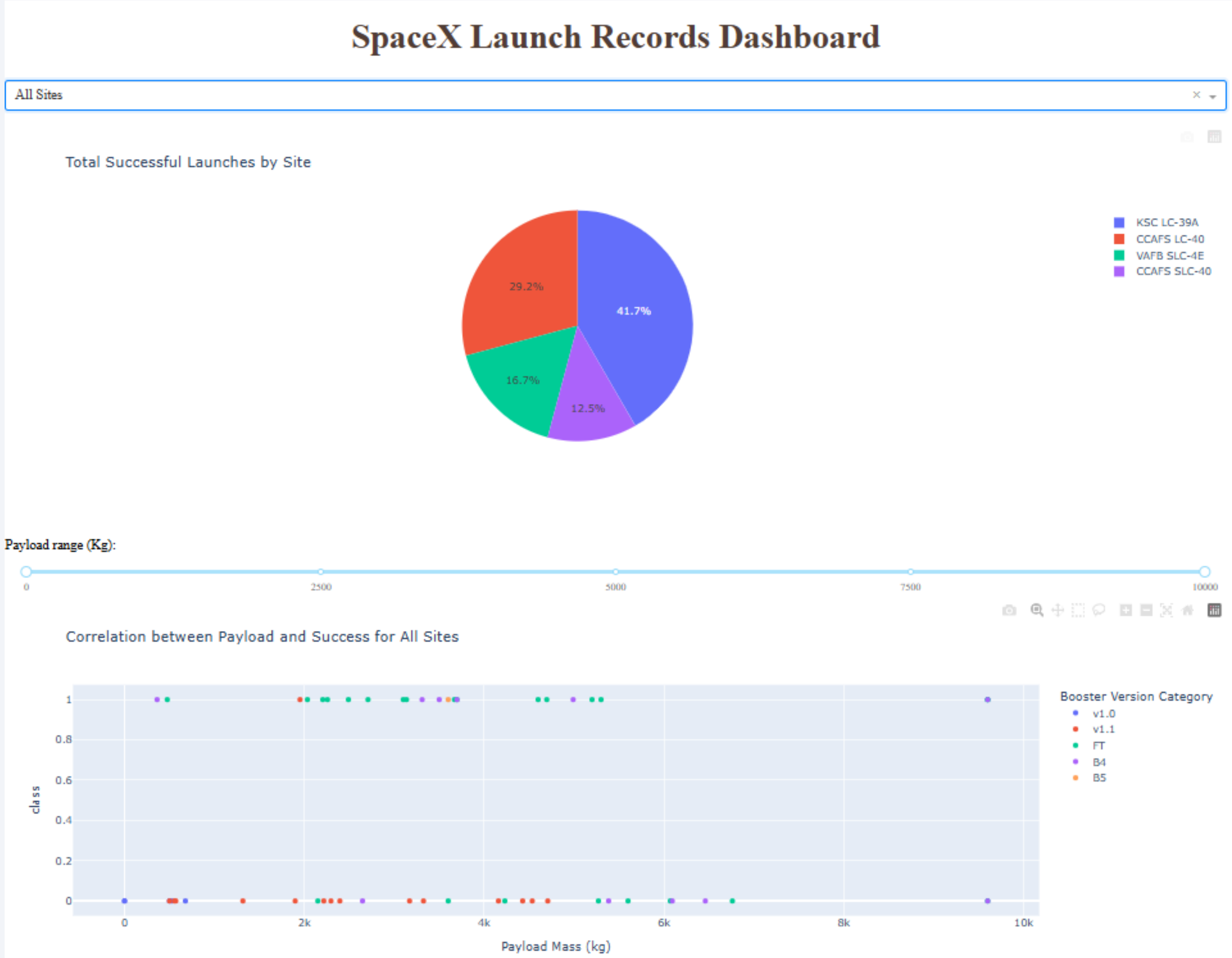
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success vs. Failure for site KSC LC-39A

KSC LC-39A had the highest success rate for landings.

Payload mass vs Launch outcomes for all launch sites



SpaceX Launch Records Dashboard

All Sites

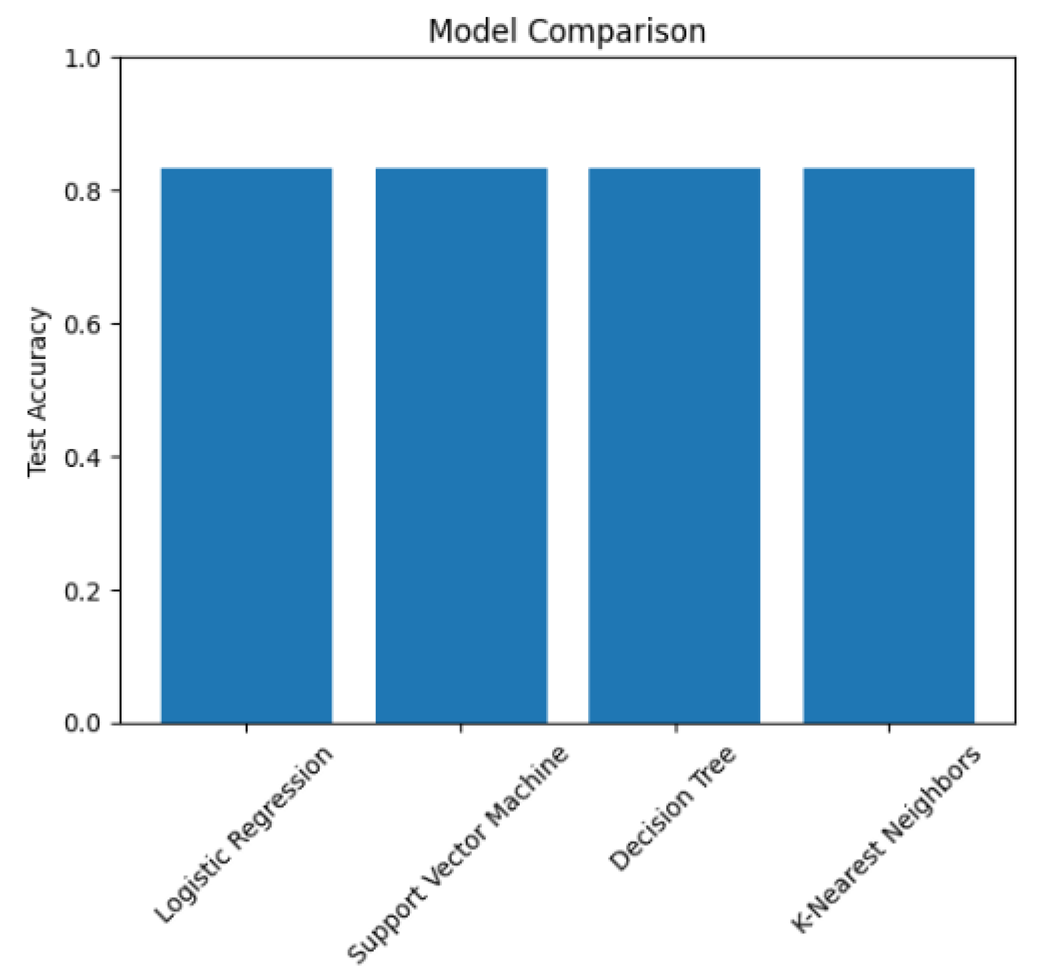
Payload range (Kg):

For payloads between 3,000–5,000 kg, v1.1 boosters had the lowest success, while B4 and B5 performed best, followed by FT.

Section 5

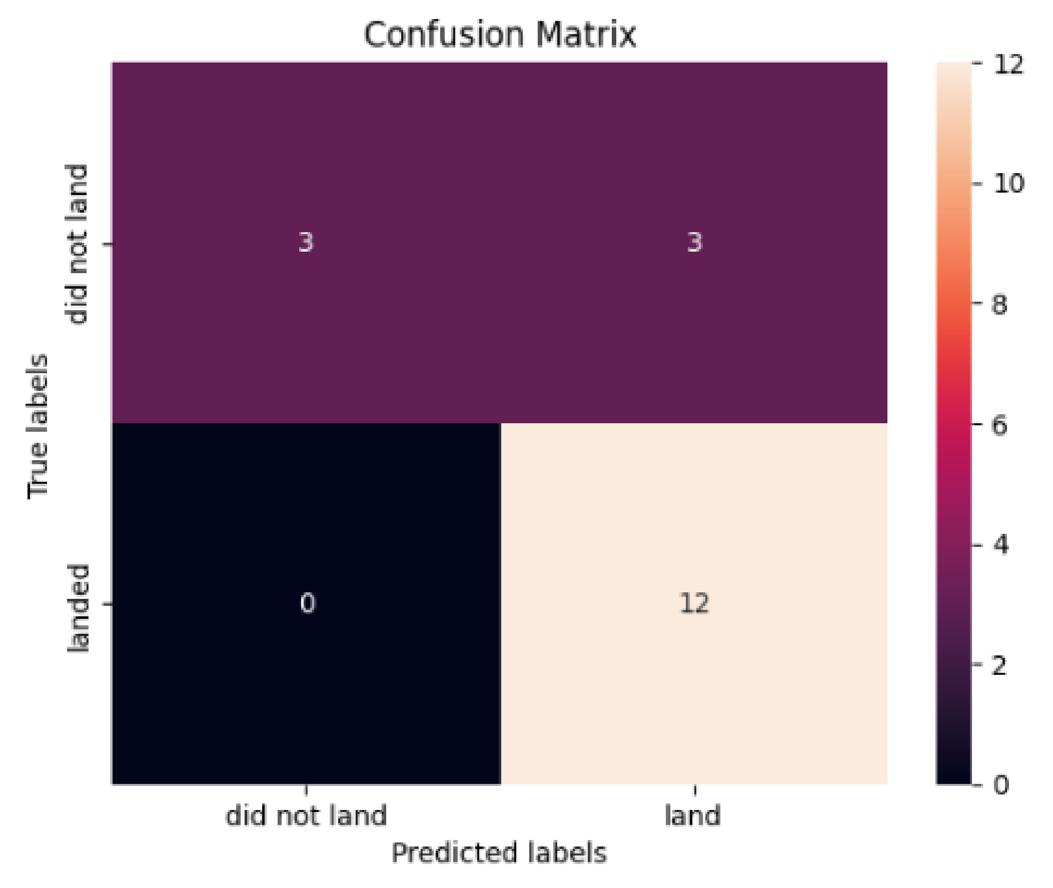
Predictive Analysis (Classification)

Classification Accuracy



All four models demonstrated equal accuracy in predicting successful landings.

Confusion Matrix

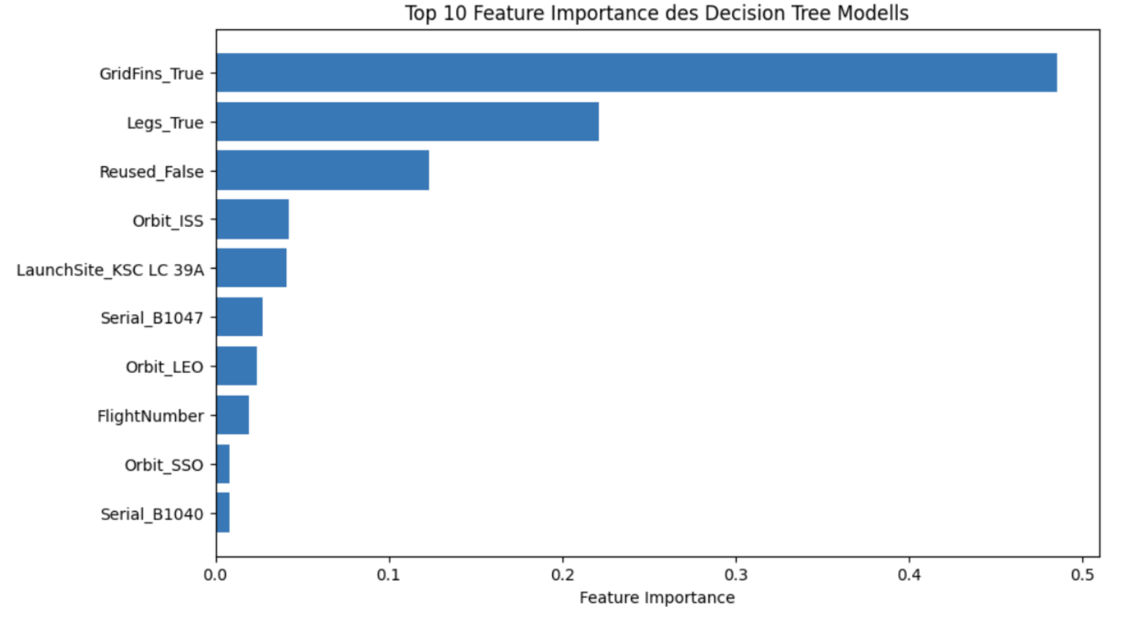


The model correctly predicted 12 successful landings and 3 failures, misclassified 3 failures as successes, and had no missed failures.

Conclusions

- All Models performed the same
- The launch sites are in close proximity to the coast
- The success rates increased over the years
- Most successful landings are made at the KSC LC-39A launch site
- Orbits ES-L1, GEO, HEO and SSO have the highest success rate with 100%

Appendix



- Legs and GridFins are the strongest predictors. Grid fins control descent, and landing legs deploy only at the final stage – so their presence indicates high success probability, but absence doesn’t guarantee failure.

Thank you!