IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neil Watson
2025-02-17

Outline

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

Executive Summary

This project analyzes SpaceX launch data to investigate trends in launch success rates, landing outcomes, and predictive modeling of successful landings. Data was acquired from multiple sources, including the SpaceX API, Wikipedia web scraping, and CSV datasets. Extraction involved Python functions to structure launch attributes, followed by cleaning processes to remove duplicates, standardize formats, and handle missing values. Exploratory data analysis (EDA) used visualizations in Matplotlib, Seaborn, and Plotly to examine trends, such as the impact of payload mass, launch site, and orbit type on success rates. SQL queries further revealed key statistics, including the most frequently used launch sites, average payload masses, and mission success rates by orbit. Interactive analytics with Folium mapped launch sites and infrastructure proximity, while a Plotly Dash dashboard enabled dynamic filtering of launch success factors. Predictive modeling applied classification algorithms—Logistic Regression, SVM, Decision Trees, and KNN—to classify landing success. KNN achieved the highest accuracy (86%), outperforming other models. Results show that launch success rates have significantly improved over time, particularly at CCAFS LC-40 and KSC LC-39A. Certain orbits, such as ES-L1 and GEO, demonstrated near-perfect success rates, while GTO and ISS showed more variability. Payload mass was a critical determinant, with launches exceeding 10,000 kg achieving higher success rates. The findings highlight SpaceX's technological advancements, optimization in orbital missions, and the increasing reliability of its landing procedures.

Introduction

SpaceX has revolutionized the aerospace industry with its reusable rocket technology, significantly reducing launch costs and increasing the frequency of space missions. Understanding the factors influencing launch success and landing outcomes is crucial for optimizing future missions. This project analyzes historical SpaceX launch data to uncover trends in launch site performance, payload mass effects, and orbit-specific success rates. By leveraging data from the SpaceX API, web scraping, and structured datasets, the study aims to answer key questions: Which launch sites have the highest success rates? How does payload mass impact mission success? Are certain orbits more prone to failures? Additionally, predictive modeling is applied to determine whether machine learning can accurately classify landing outcomes. By exploring these questions through data-driven analysis, this project provides insights into SpaceX's operational improvements and helps identify factors that contribute to successful launches and landings.

Section 1

Methodology

Methodology

Contents

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

1. Data Acquisition

- The SpaceX API (https://api.spacexdata.com/v4/) was used to fetch real-time data.
- HTTP requests (GET method) were sent to different endpoints to extract details about:
- Launches (Flight number, date, payload mass, landing success)
- Rockets (Booster version, thrust, fuel type)
- Launch Sites (Coordinates, site name)

2. Data Extraction and Processing

- Custom Python functions (getBoosterVersion(), getLaunchSite()) were used to extract specific attributes from JSON responses.
- Data such as booster version names, launchpad coordinates, and mission details were appended to structured lists.

3. Data Storage and Structuring

- Extracted data was stored in a Pandas DataFrame.
- Columns were checked for completeness and consistency.
- Missing or null values were identified and handled accordingly.

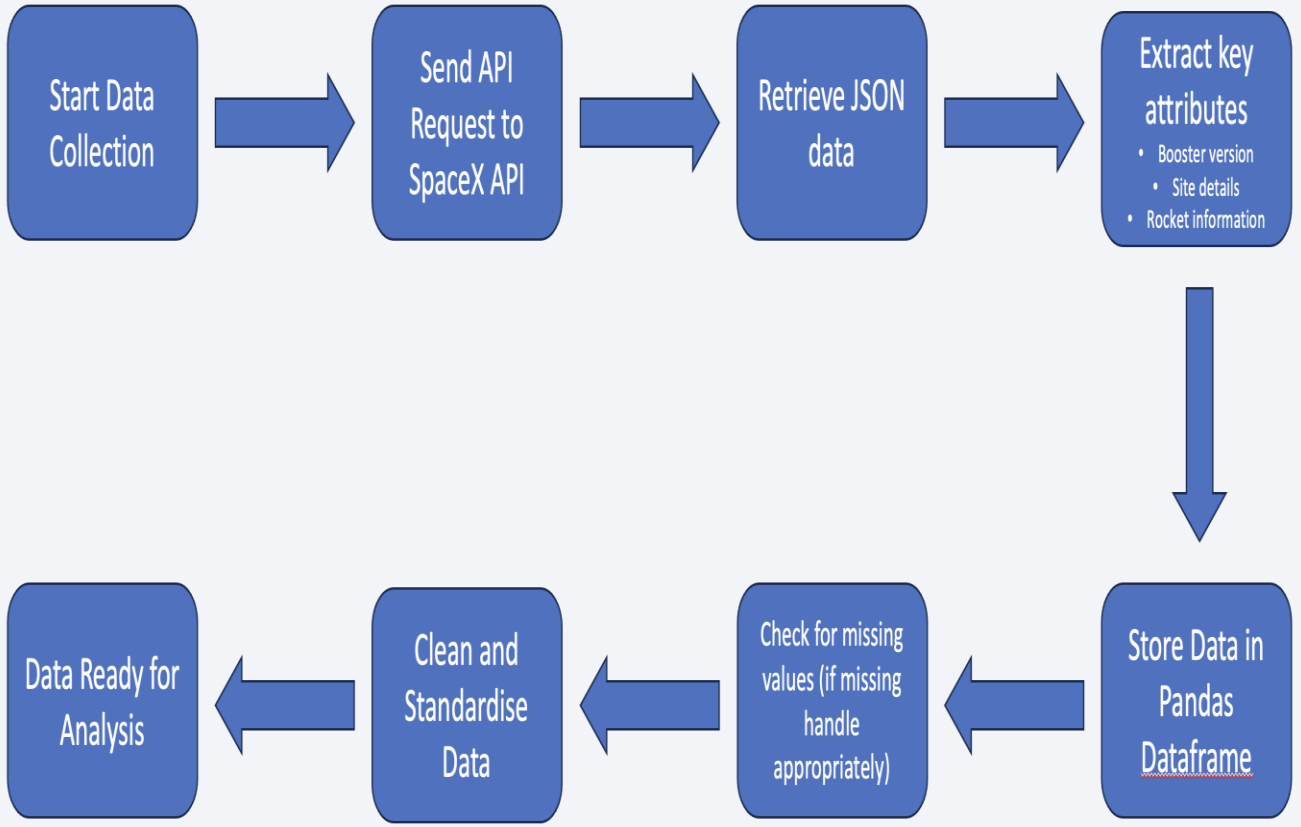
4. Data Cleaning

- Duplicate values were removed.
- Data types were standardised (e.g., dates converted to datetime format).
- Outliers and inconsistencies were checked.

Data Collection – SpaceX API

GitHub URL:

https://github.com/NwtsN/IBM_course_capstone/blob/main/data-collection.ipynb



Data Collection - Scraping

1. Data Acquisition

- The script accesses Wikipedia's List of Falcon 9 and Falcon Heavy Launches using a static URL:
- https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- An HTTP GET request is sent using the requests library.
- The page content (HTML) is retrieved for further processing.

2. Parsing the Web Page

- BeautifulSoup is used to parse the HTML structure.
- The script locates the table containing launch records.

3. Extracting Key Data Fields

- Custom Python functions extract specific attributes:
- Launch Date and Time (date_time() function)
- Booster Version (booster_version() function)
- Landing Status (landing_status() function)
- Payload Mass (get_mass() function)

4. Storing and Structuring Data

Extracted data is stored in a Pandas DataFrame.

Headers are cleaned, and missing values are handled.

Unicode normalization is performed to standardize text.

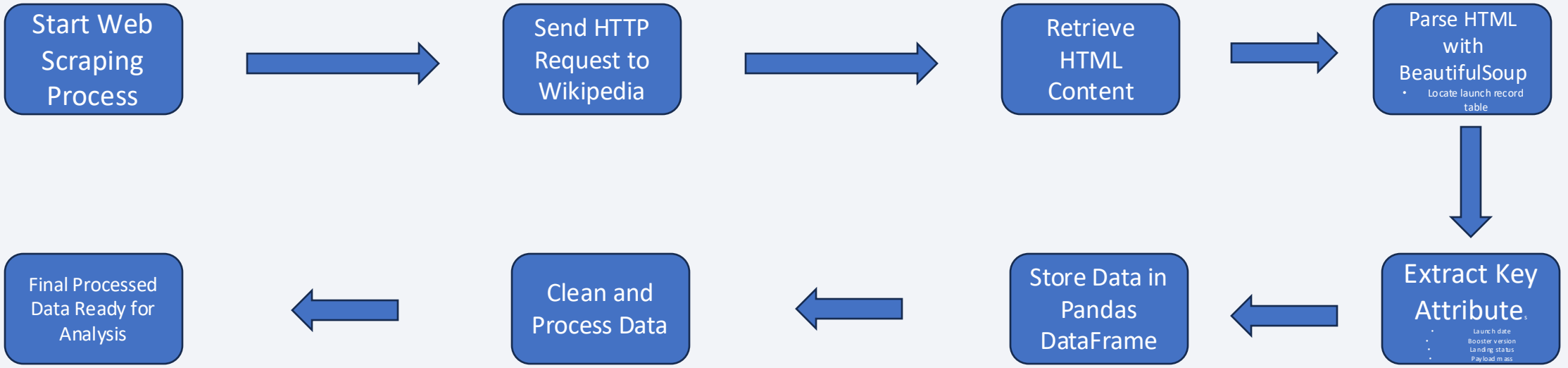
5. Data Cleaning & Preprocessing

Empty or malformed values are corrected.

Column headers are processed to remove unnecessary elements like
, <sup>, and hyperlinks.

Numeric data such as payload mass is converted to a structured format.

Data Collection - Scraping



GitHub URL:
https://github.com/NwtsN/IBM_course_capstone/blob/main/webscraping.ipynb

Data Wrangling

1. Data Acquisition

- The dataset was loaded using Pandas from an external CSV file
  - df = pd.read_csv("dataset_url.csv")
- A quick preview of the first 10 rows was performed to understand the structure
  - df.head(10)

2. Data Inspection & Cleaning

- Checking for Missing Values:
  - The percentage of missing values was computed to assess data completeness
  - df.isnull().sum() / len(df) * 100
- Identifying Data Types:
  - Column data types were checked to ensure correctness
  - df.dtypes

3. Exploratory Data Analysis

- Understanding Distributions:
  - value_counts() was applied to LaunchSite and Orbit columns to identify unique categories
  - df["LaunchSite"].value_counts() df["Orbit"].value_counts()
  - The Outcome column was examined to understand landing results
  - landing_outcomes = df["Outcome"].value_counts()

4. Data Transformation

- Classifying Landing Outcomes:
  - A set of unsuccessful landing outcomes was defined
  - bad_outcomes = set(landing_outcomes.keys()[[1,3,5,6,7]])
- A new binary column landing_class was created, classifying landings as successful (1) or failed (0)
  - df["landing_class"] = df["Outcome"].apply(lambda x: 0 if x in bad_outcomes else 1)

5. Feature Engineering

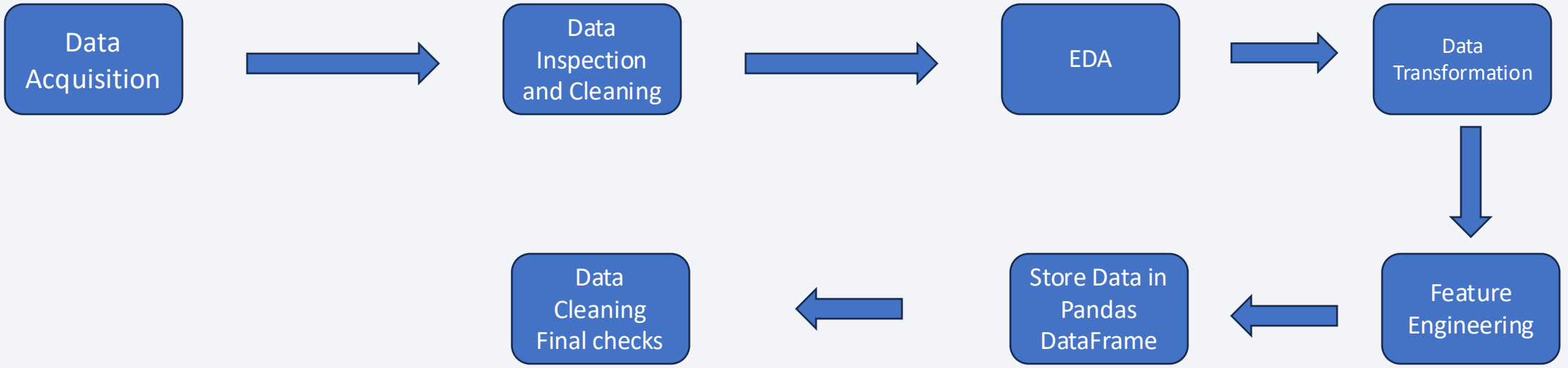
- Extracting Useful Information:
  - BoosterVersion was further analyzed to check its variations.
  - Data was grouped based on key parameters.

6. Data Cleaning Final Checks

- The cleaned dataset was reviewed before proceeding with analysis or modeling.

https://github.com/NwtsN/IBM_course_capstone/blob/main/data-wrangling.ipynb

Data Wrangling - Flowchart



EDA with Data Visualization

Scatter Plot: Flight Number vs. Payload Mass (Colored by Success Class)

- Purpose: To examine the relationship between payload mass and flight number, while highlighting whether the launch was successful or not.

Scatter Plot: Flight Number vs. Launch Site (Colored by Success Class)

- Purpose: To visualize how launch sites vary with flight number and identify if there is a trend in launch success based on location.

Scatter Plot: Payload Mass vs. Launch Site (Colored by Success Class)

- Purpose: To explore how payload mass varies across different launch sites and whether launch success is influenced by payload weight.

Bar Chart: Success Rate by Orbit Type

- Purpose: To display the average success rate for different orbit types, helping to identify which orbits have higher or lower launch success probabilities.

Scatter Plot: Flight Number vs. Orbit Type (Colored by Success Class)

Purpose: To analyze how the flight number relates to different orbit types and how successful launches are distributed among orbits.

Scatter Plot: Payload Mass vs. Orbit Type (Colored by Success Class)

Purpose: To assess how payload mass varies for different orbit types and its impact on the success of launches.

Line Plot: Success Rate Over Time (Implicit from trends in flights and success rates)

Purpose: To analyze trends in launch success rates over time, observing whether SpaceX has improved its launch success rate as the number of flights has increased.

https://github.com/NwtsN/IBM_course_capstone/blob/main/eda-data-vis.ipynb

EDA with SQL

- Query 1: Retrieved the names of unique launch sites in the dataset.
- Query 2: Displayed five records where the launch site names begin with "CCA".
- Query 3: Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Query 4: Found the average payload mass carried by the booster version "F9 v1.1".
- Query 5: Identified the earliest date of a successful landing on a ground pad.
- Query 6: Listed the booster versions that successfully landed on a drone ship and carried payloads between 4000 kg and 6000 kg.
- Query 7: Counted the number of successful and failed mission outcomes.
- Query 8: Found the booster versions that carried the maximum payload mass using a subquery.
- Query 9: Displayed the month, booster version, launch site, and landing outcome for failed landings on a drone ship in 2015.
- Query 10: Ranked landing outcomes by count (e.g., failure or success) for launches between June 4, 2010, and March 20, 2017, in descending order.

https://github.com/NwtsN/IBM_course_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Launch Site Circles and Markers

- Circle Objects: Added for each launch site, using folium.Circle with a 1000m radius.
- Markers with Labels: Each launch site also has a folium.Marker with a DivIcon displaying the site's name.
- Purpose: Highlights the locations of different launch sites, making them visually distinct.

Marker Cluster for Individual Launches

- A MarkerCluster is used to display individual launch markers from spacex_df.
- Each marker uses folium.Icon, colored to indicate launch success (green) or failure (red).
- Purpose: Groups launch points together dynamically, making the map easier to navigate.

https://github.com/NwtsN/IBM_course_capstone/blob/main/launch-site-location.ipynb

Distance Markers to Nearby Features

- Coastline Distance Marker: A folium.Marker at the closest coastline point, displaying the distance from the launch site.
- Railway Distance Marker: A folium.Marker for the nearest railway, also labeled with the distance.
- Road Distance Marker: A folium.Marker for the closest road, labeled with distance.
- Purpose: Provides insight into launch site accessibility and safety by showing proximity to key infrastructure.

Connecting Polylines

- Launch Site to Coastline: A folium.PolyLine connecting the launch site to the nearest coastline.
- Launch Site to Railway: A folium.PolyLine connecting the launch site to the nearest railway.
- Launch Site to Road: A folium.PolyLine connecting the launch site to the nearest road.
- Purpose: Visualizes spatial relationships between the launch sites and key infrastructure, aiding in analysis.

Build a Dashboard with Plotly Dash

1. Dropdown (Launch Site Selection)

- Filters data for all sites or a specific site.
- Updates both the pie chart (success rate per site) and scatter plot (payload vs. success).

2. Pie Chart (Success Rate by Site)

- Displays total successful launches across all sites or
- Success vs. failure rate for a selected site.

https://github.com/NwtsN/IBM_course_capstone/blob/main/plotly_dashboard.py

3. Payload Mass Range Slider

- Filters launches based on payload weight (kg).
- Updates the scatter plot to analyze payload impact on success.

4. Scatter Plot (Payload vs. Success)

- Shows the relationship between payload mass and success.
- Filters data by launch site and payload range.
- Uses colour-coded Booster Versions for comparison.

Predictive Analysis (Classification)

1. Data Preparation & Exploration

- Created a classification target column indicating landing success.
- Standardised the dataset for better model performance.
- Split data into training and testing sets using train_test_split().

2. Model Selection & Hyperparameter Tuning

- Used multiple classification algorithms:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbours (KNN)
- Implemented GridSearchCV for hyperparameter tuning with cross-validation (cv=10) to optimise:
  - Logistic Regression: Tuned C (regularisation strength).
  - SVM: Tuned C and kernel type.
  - Decision Tree: Tuned max_depth and criterion.
  - KNN: Tuned n_neighbors.

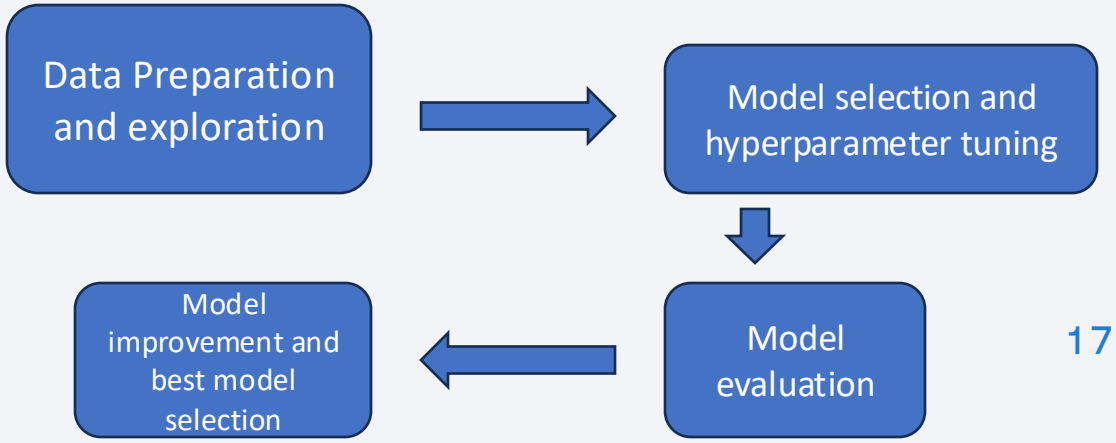
https://github.com/NwtsN/IBM_course_capstone/blob/main/ml-pred.ipynb

3. Model Evaluation

- Evaluated models using:
- Accuracy on validation set (best_score_ from GridSearchCV).
- Accuracy on test set (score(X_test, Y_test)).
- Confusion Matrix to visualise model predictions.

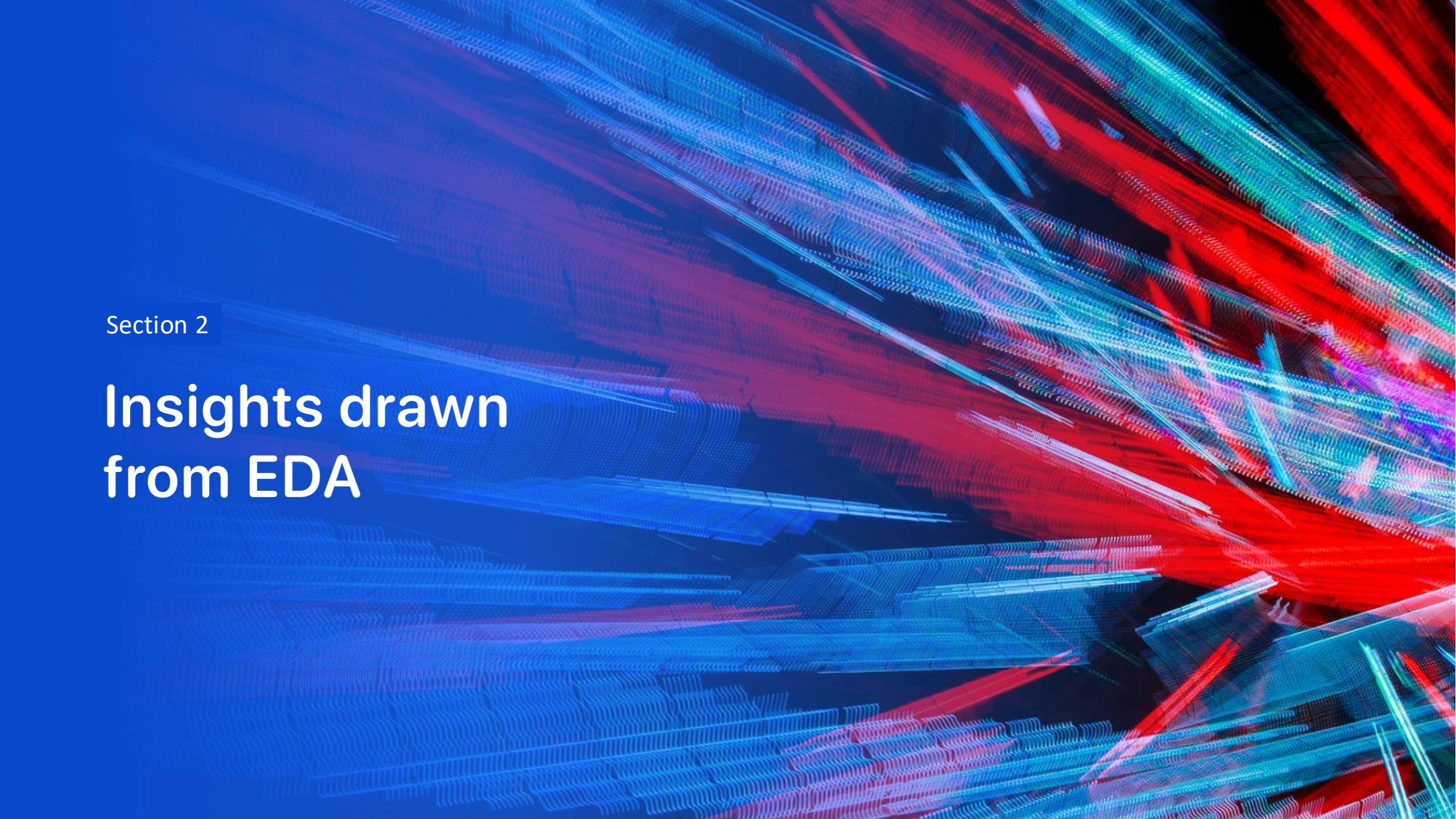
4. Model Improvement & Best Model Selection

- Compared accuracy of different models.
- Selected the best-performing model based on test accuracy.
- Final model evaluated with precision, recall, and F1-score.

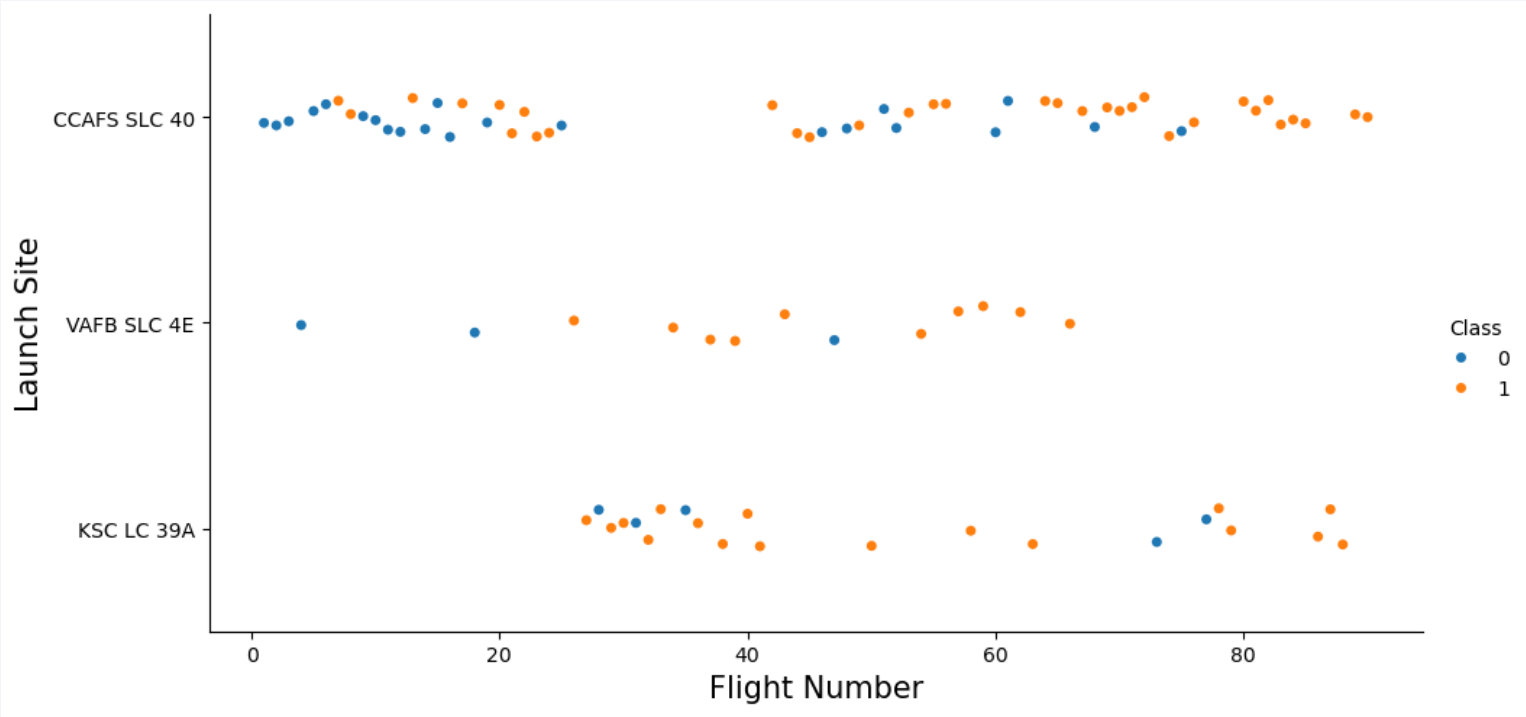


Results

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


Section 2
Insights drawn from EDA

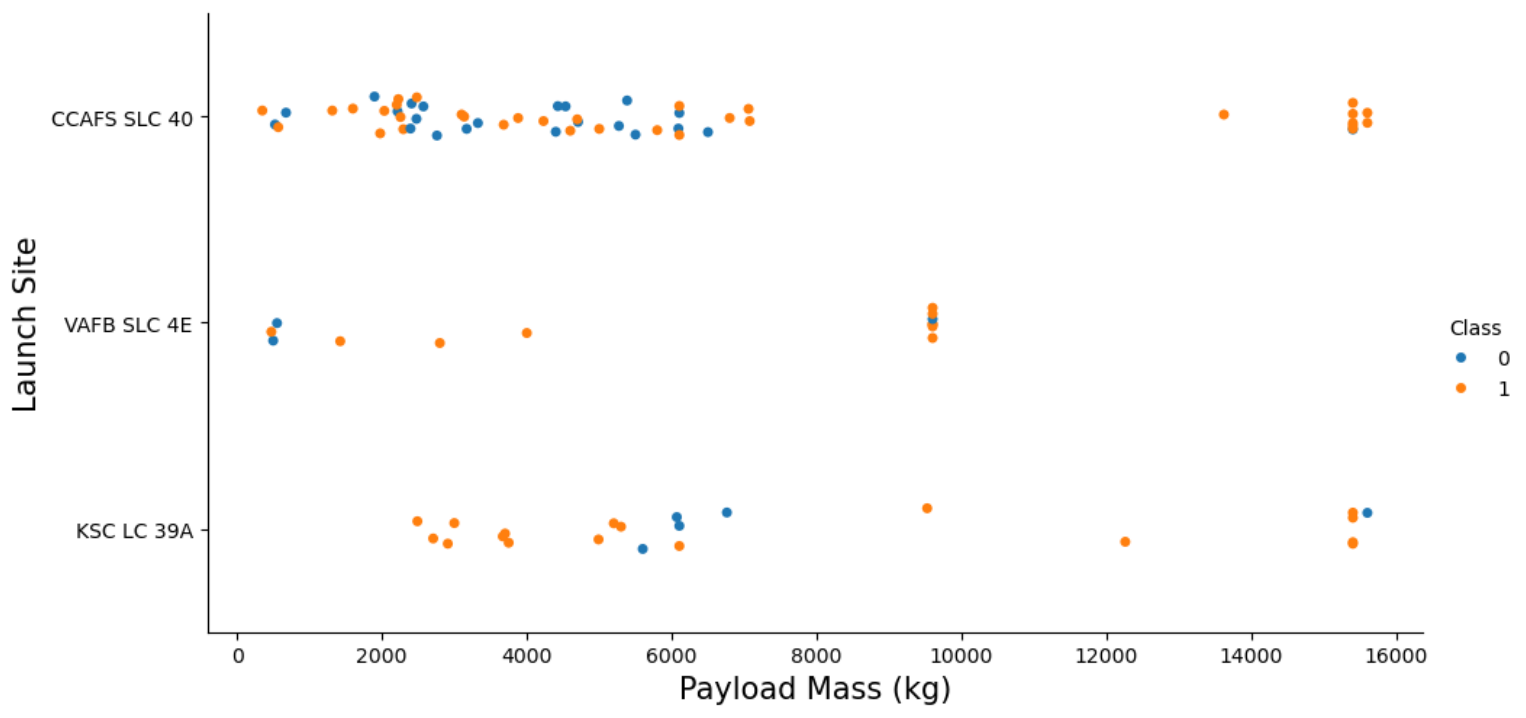
Flight Number vs. Launch Site



Explanation:

Launch sites CCAFS SLC 40 and KSC LC 39A have a higher frequency of launches compared to VAFB SLC 4E. Successful landings (Class = 1, orange) become more common as the flight number increases.

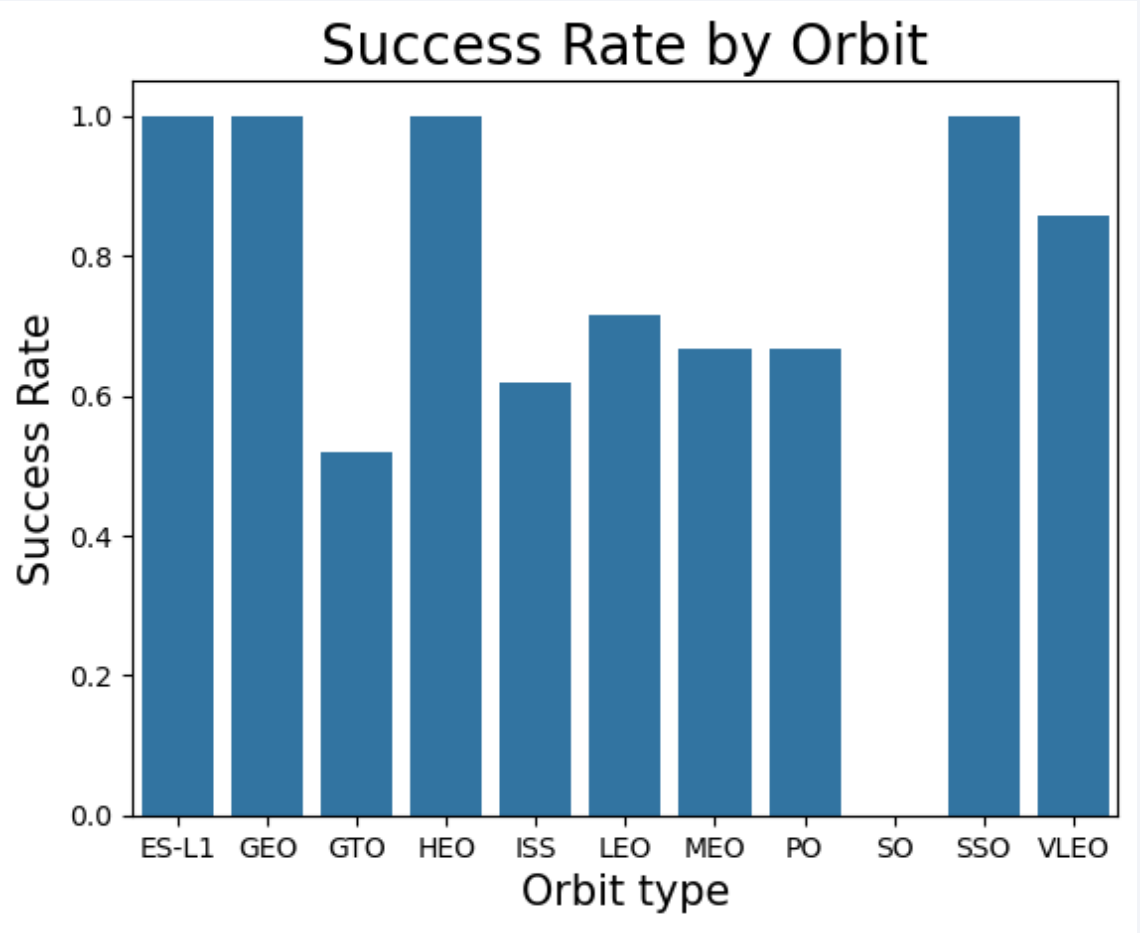
Payload vs. Launch Site



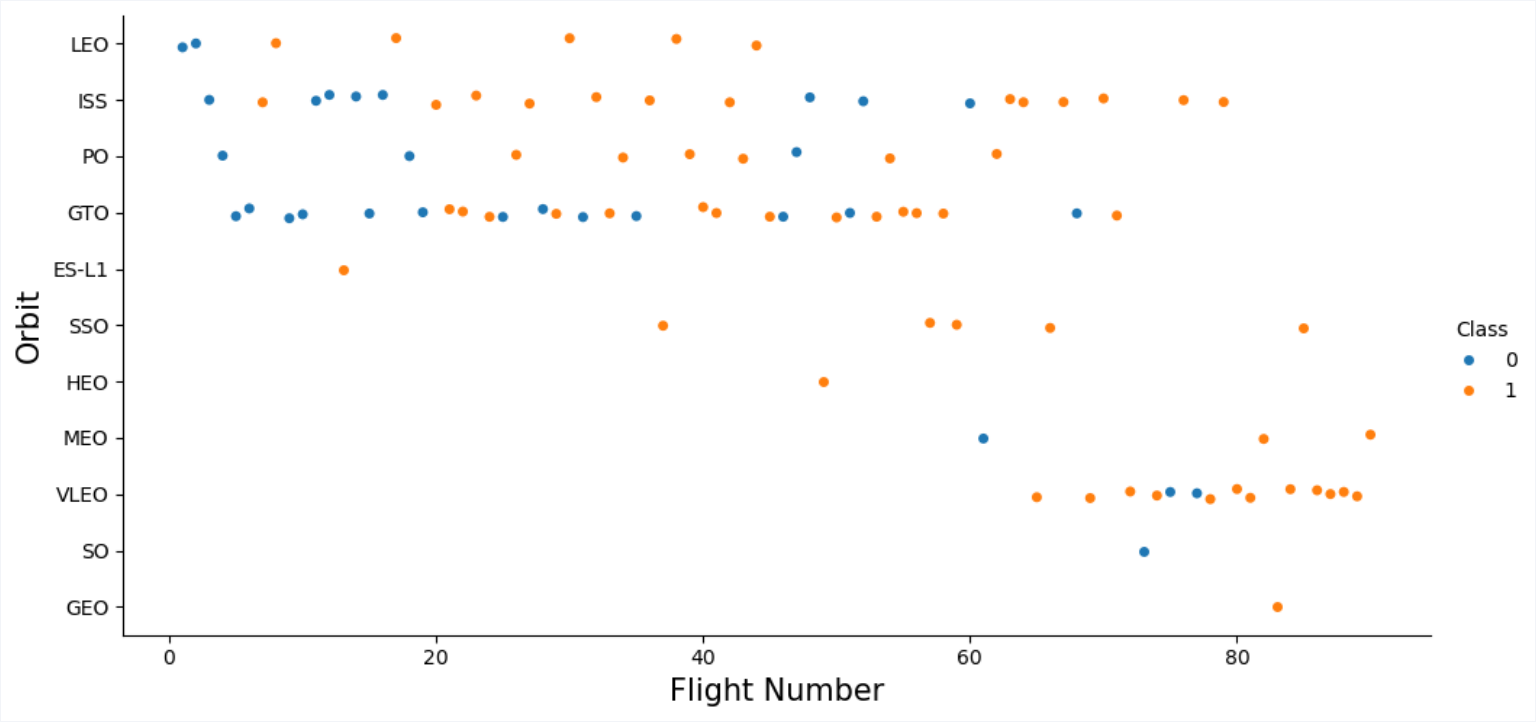
CCAFS SLC 40 appears to have the most launches, with a mix of successes and failures. KSC LC 39A has a good number of successful launches, especially in the mid-range of payloads (around 5,000 - 7,000 kg). VAFB SLC 4E has relatively fewer launches but still shows a mix of success and failure. Some launch sites, particularly CCAFS SLC 40, have been used frequently for lower to mid-range payloads. Heavier payloads appear to have higher success rates (orange dots) while failures (blue dots) seem more common at lower payload ranges, suggesting these might be test launches or earlier missions.

Success Rate vs. Orbit Type

Certain orbits, such as ES-L1, GEO, HEO, and SSO, boast a perfect 100% success rate, indicating high mission reliability and well-optimized launch procedures. Similarly, VLEO also exhibits a strong success rate, around 90%, further demonstrating SpaceX's proficiency in reaching low-altitude operational orbits. In contrast, orbits like LEO, ISS, MEO, and PO show moderate success rates between 60% and 70%, likely reflecting the increased complexity and frequency of launches to these destinations. Notably SO has not yet successfully completed a mission to solar orbit and around only half of GTO missions have been a success.



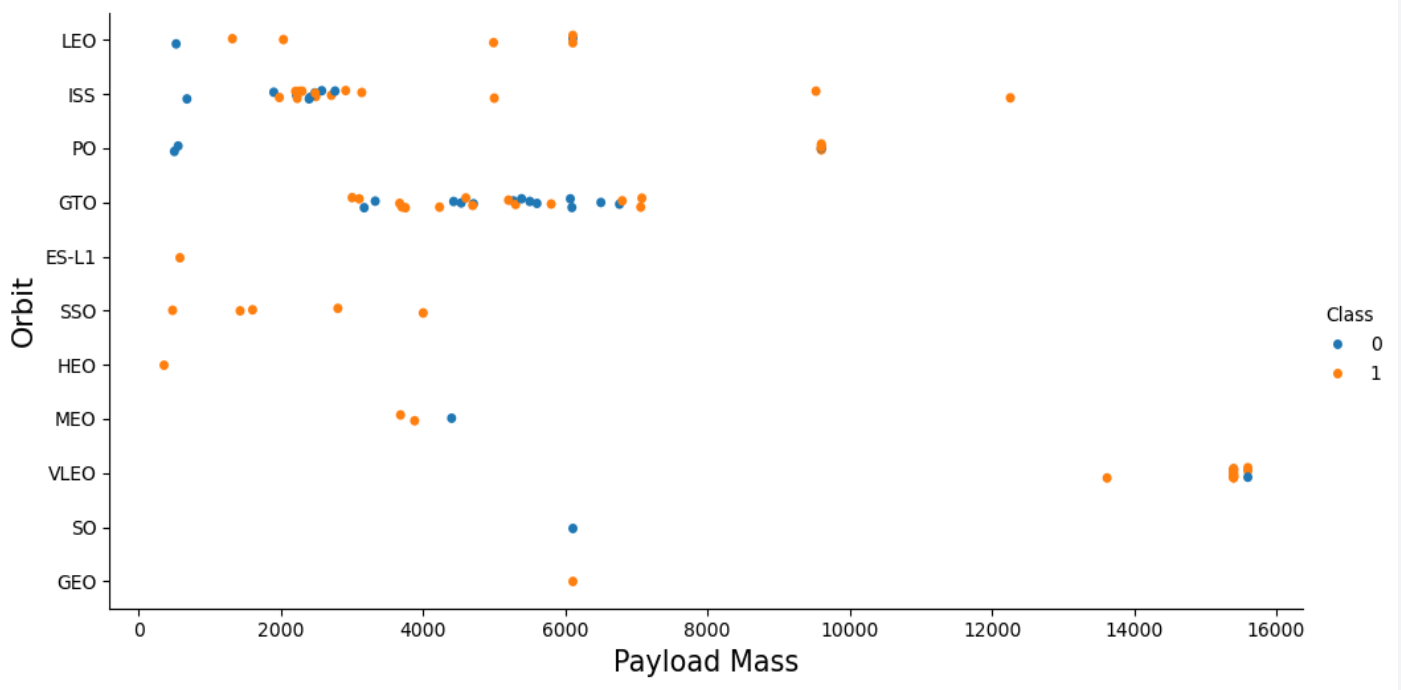
Flight Number vs. Orbit Type



In the early missions, there is a noticeable higher frequency of failures (blue dots) across multiple orbits, particularly in ISS, and GTO, indicating initial technical challenges. However, as the flight numbers increase, the proportion of successful launches (orange dots) grows significantly, reflecting SpaceX's improving reliability. Orbits such as VLEO, and ISS show a high success rate in later missions, while GTO continues to have a mixed record, suggesting persistent challenges in these missions. The sparse data points for GEO, ES-L1, SO and HEO indicate fewer attempts, with some failures. Overall, the trend highlights SpaceX's increasing proficiency in reaching different orbits, demonstrating technological advancements and better mission execution over time.

Payload vs. Orbit Type

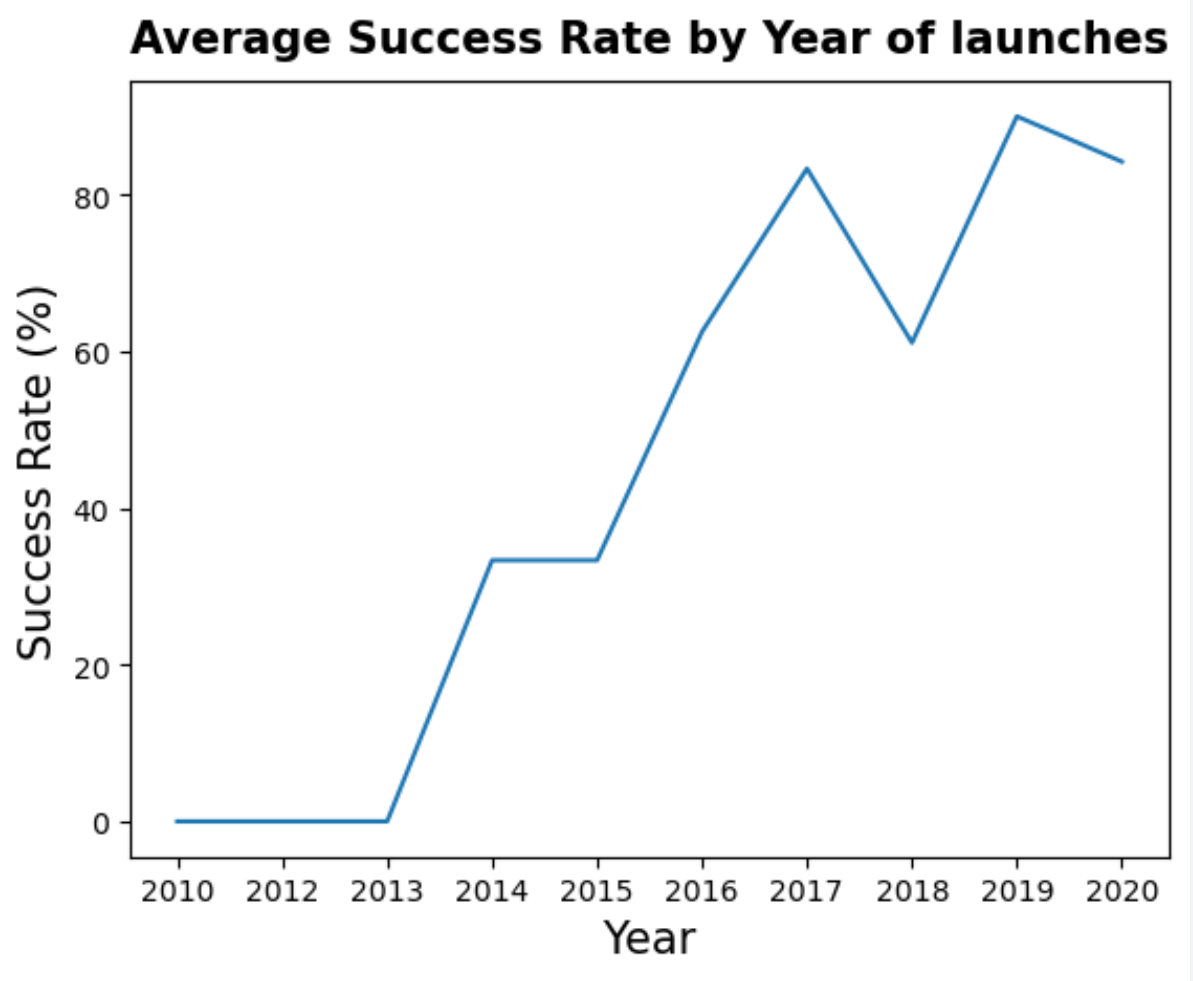
The distribution of points shows that ISS missions accommodate the widest range of payload masses, with both successes and failures occurring more frequently at lower payloads. GTO and VLEO have a large concentration of payloads in the 2,000 – 8,000 kg and 13,000 – 16,000 kg ranges, respectively, both with mixed success, suggesting challenges in these specific payload categories. ES-L1, SSO, HEO, and MEO payloads are in the lower range and all exhibit high success rates, indicating that these mission profiles are well-optimized.



The LEO orbit shows an increasing trend of success as payload mass rises, reinforcing SpaceX's improved reliability with heavier payloads. Additionally, heavier payloads exceeding 10,000 kg tend to have higher success rates, likely due to more mission-critical launches receiving greater planning and precision. Overall, the plot highlights SpaceX's ability to successfully handle a broad range of payloads while revealing persistent challenges in certain transfer orbits like GTO.

Launch Success Yearly Trend

This line chart illustrates the average success rate of SpaceX launches over time, showing a clear upward trend from 2013 onwards. In the early years (2010–2012), the success rate was zero, indicating either failed attempts or no recorded launches. However, from 2013 to 2015, there was a sharp increase in success rate, reaching around 35%, likely due to improvements in technology and mission execution. The trend continued to rise steadily, surpassing 60% by 2016 and peaking above 80% in 2017, reflecting SpaceX's growing reliability in orbital launches. While 2018 saw a slight dip, the success rate rebounded in 2019, reaching its highest level, before experiencing a minor decline in 2020. Overall, the plot highlights SpaceX's rapid progress in launch success, demonstrating continuous refinement in rocket design, operational efficiency, and mission reliability over the years.



All Launch Site Names

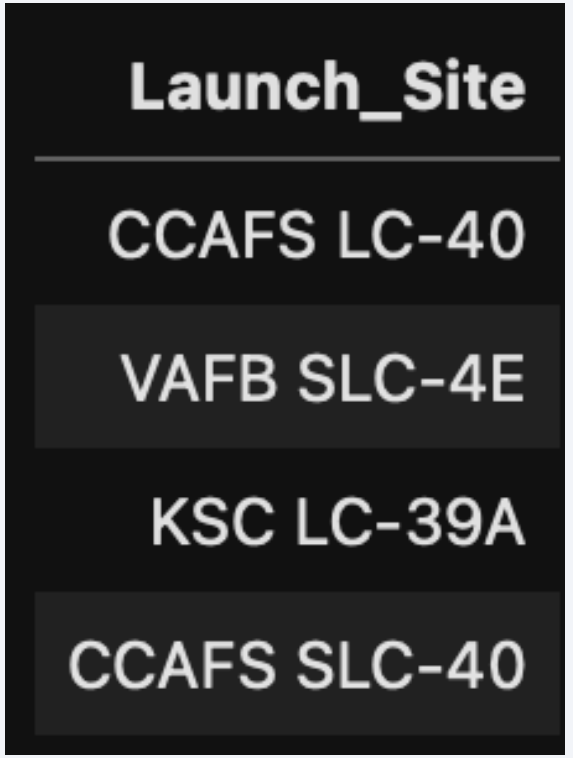
The SQL query below retrieves a list of unique SpaceX launch sites from the table named SPACEXTABLE:

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

Explanation:

- SELECT DISTINCT: This clause is used to eliminate duplicate entries in the result set, ensuring that each launch site appears only once.
- Launch_Site: Refers to the column containing the names of the launch sites.
- FROM SPACEXTABLE: Specifies the table from which the data is being queried.
- This query is useful for quickly identifying all the distinct locations from which SpaceX has conducted launches.

Result:



Launch Site Names Begin with 'CCA'

Result:

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

The SQL query below retrieves up to five launch site names from the SPACEXTABLE where the launch site name starts with 'CCA':

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

Explanation:

- SELECT * : Retrieves all records from the table.
- FROM SPACEXTABLE: Specifies the table containing the data.
- WHERE Launch_Site LIKE 'CCA%': Filters results to include only those launch sites whose names begin with 'CCA' (e.g., sites at Cape Canaveral).
    - The LIKE operator is used for pattern matching.
    - 'CCA%' means that the name must start with 'CCA', followed by any number of characters (% is a wildcard).
- LIMIT 5: Restricts the output to at most five results to improve efficiency.

Total Payload Mass

Result:

TotalPayloadMassLaunched
48213

The SQL query below calculates the total payload mass carried by boosters launched for NASA (CRS missions) from the table SPACEXTABLE:

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMassLaunched FROM SPACEXTABLE WHERE Customer LIKE '%CRS%';
```

Explanation:

- SELECT SUM(PAYLOAD_MASS__KG_): Computes the total payload mass across all relevant launches.
- AS TotalPayloadMassLaunched: Assigns a meaningful alias to the computed sum.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- WHERE Customer LIKE '%CRS%': Filters the records to include only those where the Customer column contains 'CRS', which corresponds to NASA Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

Result:

average_payload_mass_BV_F9v1point1
2534.6666666666665

The SQL query below calculates the average payload mass carried by booster version F9 v1.1 from the SPACEXTABLE table:

```
SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass_BV_F9v1point1 FROM SPACEXTABLE
WHERE Booster_Version LIKE '%F9 v1.1%';
```

Explanation:

- SELECT AVG(PAYLOAD_MASS__KG_): Computes the average payload mass across all relevant launches.
- AS average_payload_mass_BV_F9v1point1: Assigns a meaningful alias to the computed average.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- WHERE Booster_Version LIKE '%F9 v1.1%': Filters the records to include only those launches where the Booster_Version contains 'F9 v1.1', ensuring we calculate the average for the correct booster model.

First Successful Ground Landing Date

Result:

Date
2015-12-22

The SQL query below retrieves the date of the first successful landing on a ground pad from the SPACEXTABLE table:

```
SELECT Date FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%Success (ground pad)%' ORDER BY Date ASC LIMIT 1;
```

Explanation:

- SELECT Date: Retrieves the date of the successful landing.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch and landing data.
- WHERE Landing_Outcome LIKE '%Success (ground pad)%': Filters the records to only include landings that were successful on a ground pad.
- ORDER BY Date ASC: Sorts the results in ascending order, ensuring the earliest landing appears first.
- LIMIT 1: Restricts the output to only the first successful landing event.

Successful Drone Ship Landing with Payload between 4000 and 6000

The SQL query below retrieves the names of unique booster versions that have successfully landed on a drone ship and carried a payload mass between 4000 kg and 6000 kg from the SPACEXTABLE table:

```
SELECT DISTINCT Booster_Version FROM SPACEXTABLE

WHERE Landing_Outcome = 'Success (drone ship)'

AND PAYLOAD_MASS__KG_ > 4000

AND PAYLOAD_MASS__KG_ < 6000;
```

Explanation:

- SELECT DISTINCT Booster_Version: Retrieves the unique booster names to avoid duplicates in the result.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- WHERE Landing_Outcome = 'Success (drone ship)': Filters the records to include only launches that successfully landed on a drone ship.
- AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000: Ensures that only launches with a payload mass greater than 4000 kg but less than 6000 kg are selected.

Result:

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

Total Number of Successful and Failure Mission Outcomes

The SQL query below retrieves the total number of successful and failed mission outcomes from the SPACEXTABLE table:

```
SELECT Mission_Outcome, COUNT(*) AS Total_Count FROM
SPACEXTABLE GROUP BY Mission_Outcome;
```

Explanation:

- SELECT Mission_Outcome: Retrieves the outcome of each mission (e.g., success or failure).
- COUNT(*) AS Total_Count: Counts the number of occurrences for each mission outcome.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- GROUP BY Mission_Outcome: Groups the results by each unique mission outcome, ensuring that the total count is calculated separately for successes and failures.

Result:

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

Boosters Carried Maximum Payload

The SQL query below retrieves the names of booster versions that have carried the maximum payload mass from the SPACEXTABLE table:

```
SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE);
```

Explanation:

- SELECT DISTINCT Booster_Version: Retrieves a unique list of booster versions that match the criteria.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE):
  - The subquery (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) identifies the highest payload mass recorded in the dataset.
  - The main query then filters the results to include only those booster versions that carried this maximum payload mass.

Result:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

This SQL query below retrieves records displaying the month names, booster versions, launch sites, and failure landing outcomes on a drone ship for all SpaceX launches that occurred in 2015:

```
SELECT DISTINCT CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May' WHEN '06' THEN 'June' WHEN '07' THEN 'July' WHEN '08' THEN 'August' WHEN '09' THEN 'September' WHEN '10' THEN 'October' WHEN '11' THEN 'November' WHEN '12' THEN 'December' END AS Month_Name, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

Explanation:

- SELECT DISTINCT: Ensures that only unique records are retrieved.
- CASE statement with substr(Date, 6, 2): Extracts the month from the Date column (assuming Date is in YYYY-MM-DD format) and converts it into a month name.
- Booster_Version: Retrieves the booster version used in each failed landing.
- Launch_Site: Specifies the launch site for the corresponding mission.
- Landing_Outcome: Filters for failures on drone ships, meaning unsuccessful landings at sea.
- FROM SPACEXTABLE: Queries data from the SpaceX launch table.
- WHERE substr(Date, 1, 4) = '2015': Restricts results to launches from the year 2015.
- AND Landing_Outcome = 'Failure (drone ship)': Filters only failed landings on drone ships, excluding successful landings or failures on ground landing sites.

Result:

Month_Name	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The SQL query below ranks the count of landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") for SpaceX launches between 4th June 2010 and 20th March 2017, sorting them in descending order based on frequency:

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

Explanation:

- SELECT Landing_Outcome, COUNT(*) AS Outcome_Count: Retrieves each unique landing outcome and counts how many times it occurred.
- FROM SPACEXTABLE: Specifies the table containing SpaceX launch data.
- WHERE Date BETWEEN '2010-06-04' AND '2017-03-20': Filters the records to include only landings that occurred within the specified date range.
- GROUP BY Landing_Outcome: Groups the results by each distinct landing outcome, ensuring that the count applies to each category separately.
- ORDER BY Outcome_Count DESC: Sorts the results in descending order, so the most frequent landing outcomes appear at the top.

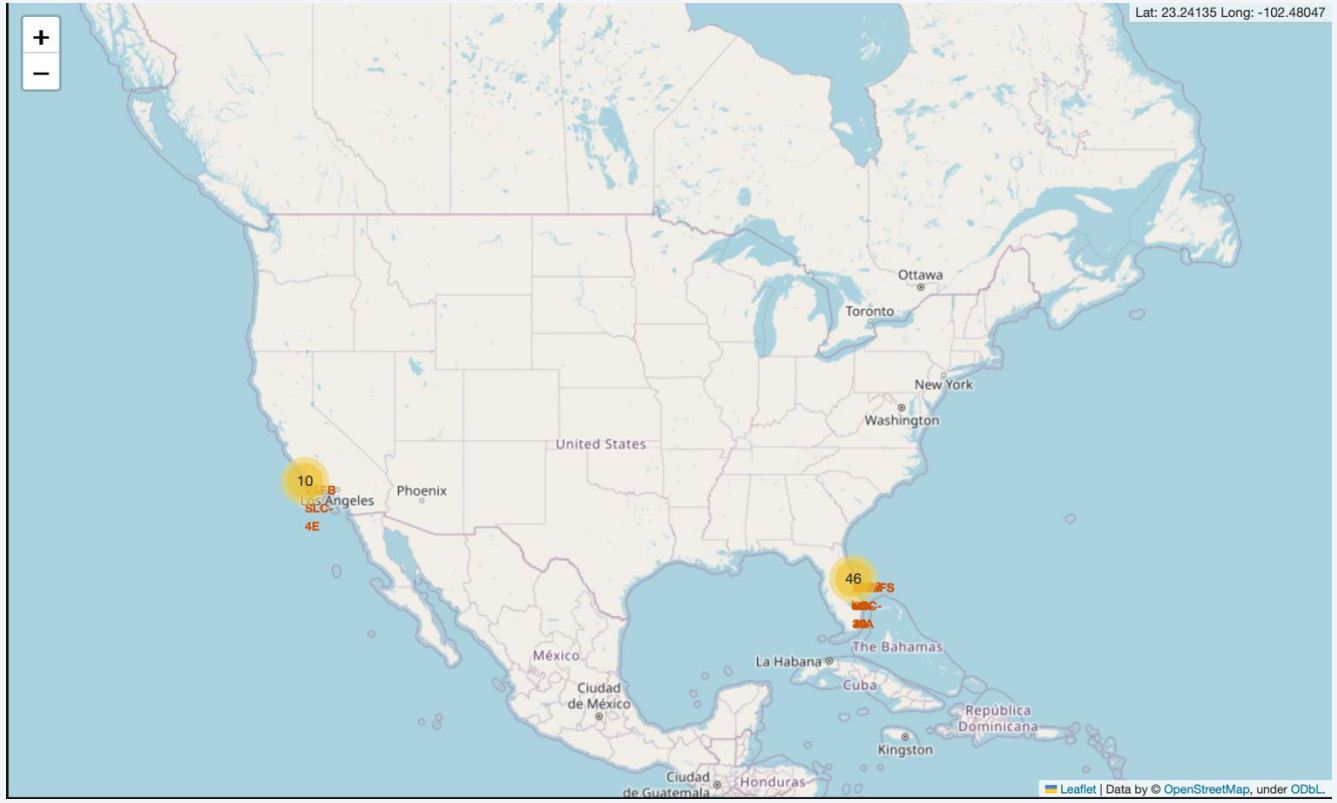
Result:

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

Section 3

Launch Sites Proximities Analysis

Global SpaceX Launch Site Locations



The map highlights two primary clusters of SpaceX launch sites:

- Florida (Cape Canaveral/Kennedy Space Center): A large cluster is observed in Florida, where SpaceX conducts most of its launches.
- California (Vandenberg Space Force Base): A smaller cluster is present on the West Coast, representing the launch site in California.

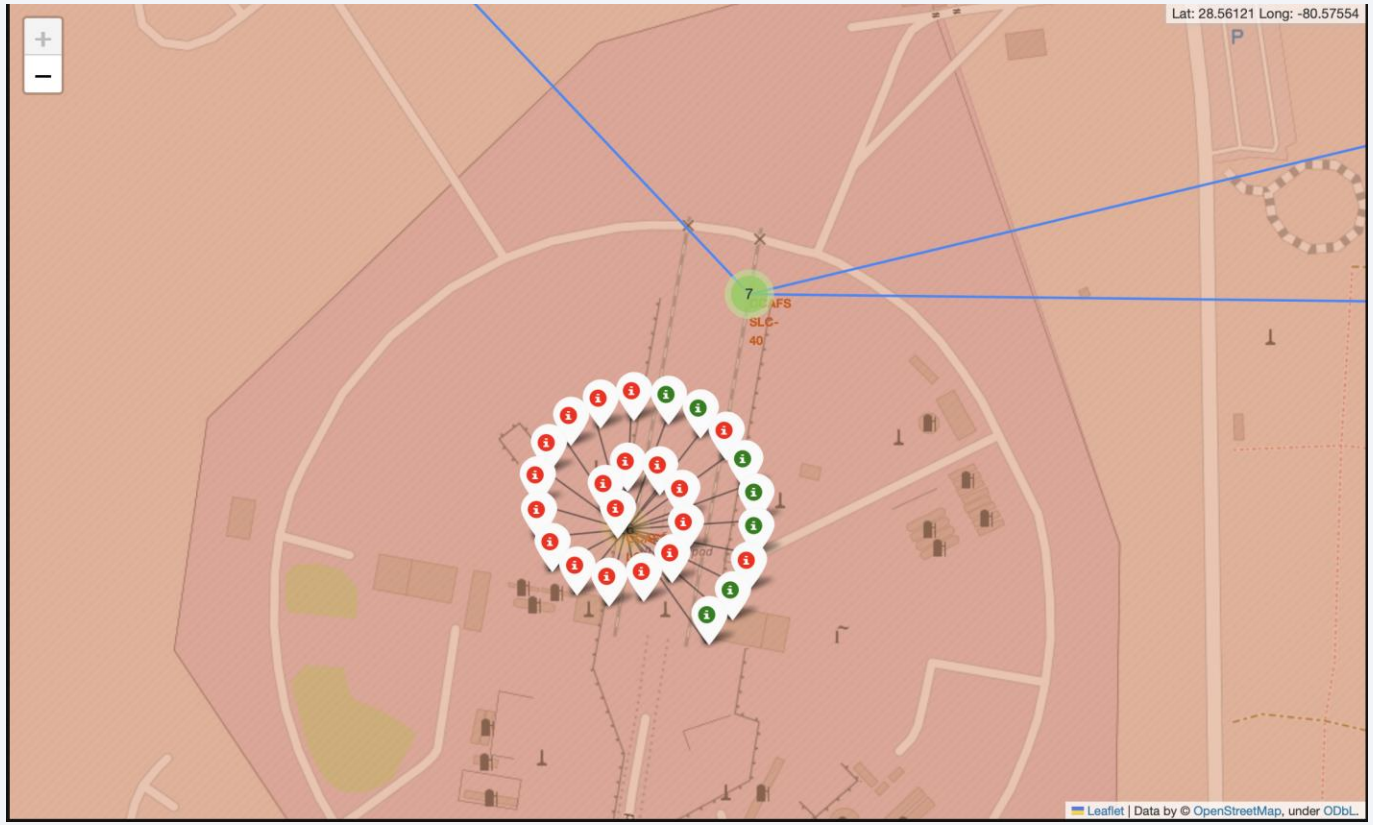
The yellow markers with numbers indicate clusters of launch sites. For example:

- "46" near Florida indicates a high concentration of launches.
- "10" near California suggests fewer launches from that location.

The clustering approach helps in visualizing the density of launches and prevents overlapping markers.

- Florida is the Dominant Launch state: The concentration of markers in Florida confirms that the Kennedy Space Center and Cape Canaveral Space Force Station are the primary locations for SpaceX launches.
- Vandenberg as a Secondary Site: The smaller cluster in California (Vandenberg) indicates its role in launching polar orbit missions.
- Geographic Distribution of Launches: SpaceX mainly launches from coastal regions, ensuring safe rocket trajectories over the ocean.

Launch Success Status by Launch Site



Launch Breakdown for LC-40 and SLC-40:
From the spiral of markers at LC-40, we can clearly see that there were 26 launches from this pad, with:

- 7 successful launches (green markers)
- 19 unsuccessful launches (red markers)

The green marker above the LC-40 spiral (with "7") represents SLC-40, which can be expanded to show:

- 7 total launches
- 4 unsuccessful launches (red markers)
- 3 successful launches (green markers)

Spiral of Markers Showing Launch Success at CCAFS LC-40:

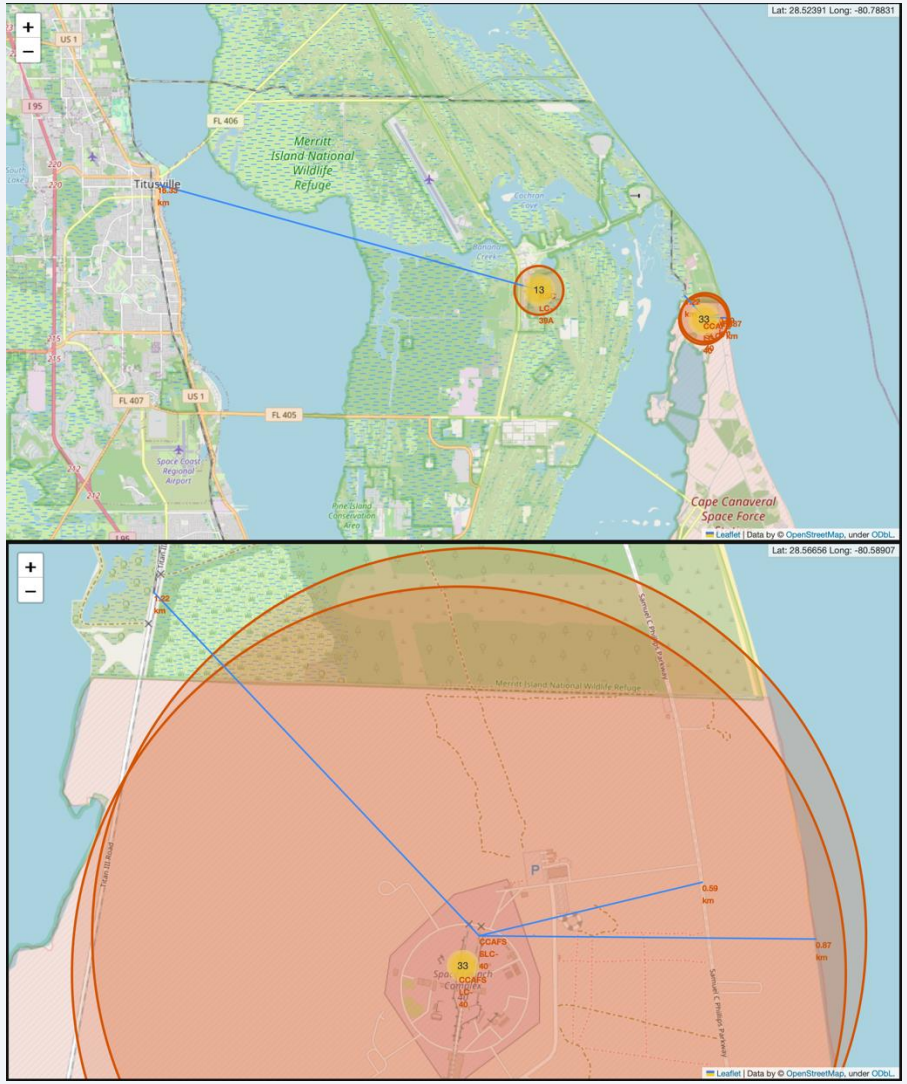
- Each marker in the spiral represents a past SpaceX launch from LC-40.
- Red icons likely indicate failed launches.
- Green icons represent successful launches.
- The spiral layout helps with visualization, avoiding overlap while still showing all historical launch outcomes.

Expandable Launch History for CCAFS SLC-40 (7 Launches):

- The green marker with "7" north of launch site LC-40 represents launches from SLC-40.
- Expanding it would reveal a similar success/failure breakdown in a spiral, just like LC-40.
- This method allows a clear comparison between all launch sites based on historical performance.

The large red shading is due to the site marker icon size staying constant regardless of zoom level

Launch Site Proximity to Notable Locations

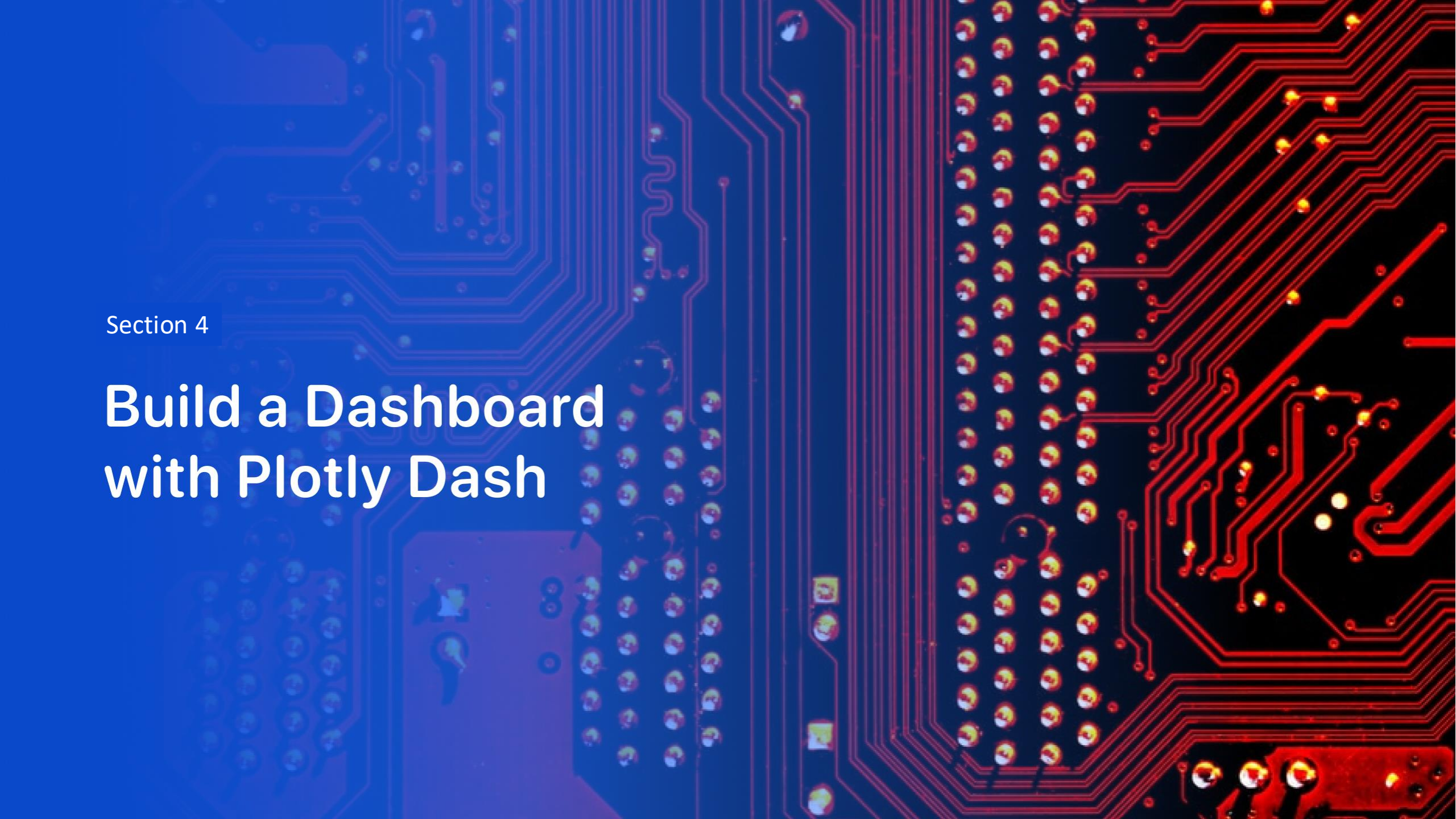


Two Primary Launch Site Clusters:

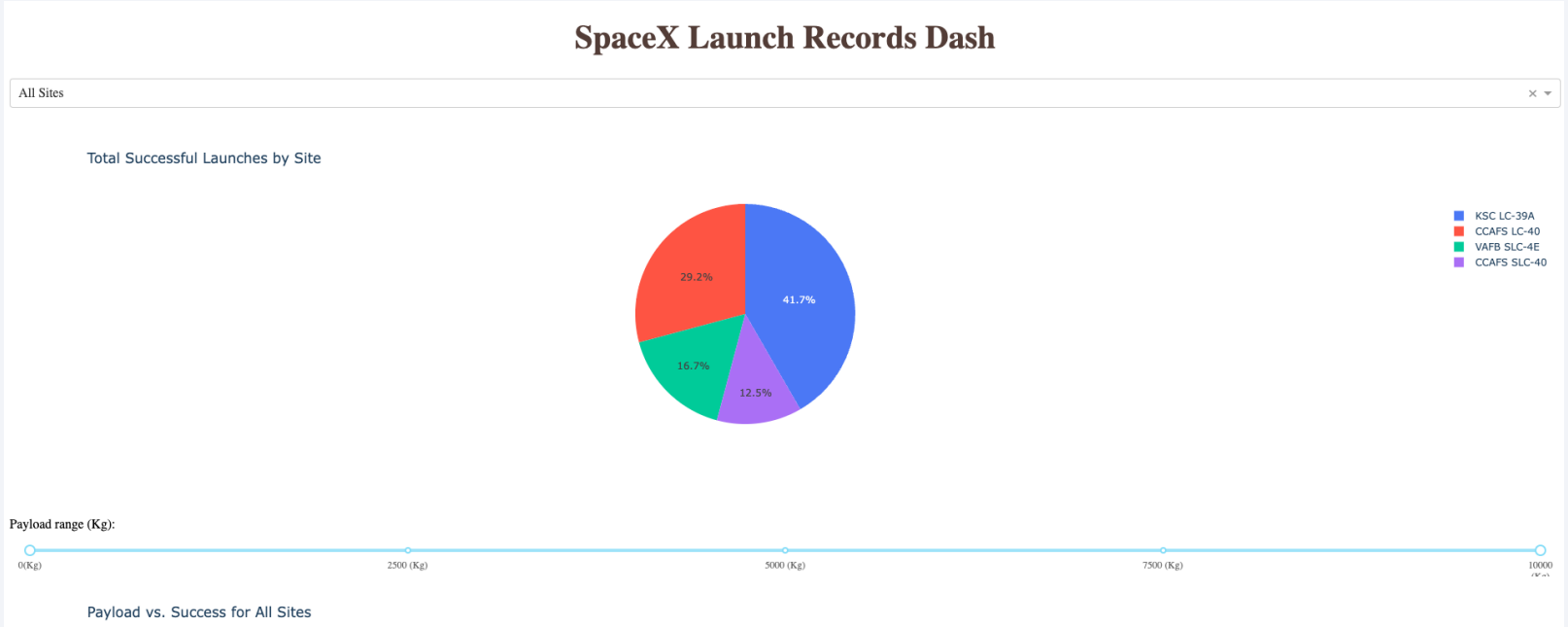
- Large cluster (39 markers) at Cape Canaveral Space Force Station (CCAFS), showing LC-40 and SLC-40 launch sites.
- Smaller cluster (13 markers) to the west, near Titusville, representing the Kennedy Space Center.

Blue Lines Representing Key Proximity Measurements:

- One line extends from launch site KSC LC-39A to the city of Titusville (Nearest City to a launch site).
- Another line connects the nearest coastline to a launch site (site CCAFS SCL-40).
- Additional blue lines indicate the distances to the nearest highway and railway from the closest launch site (again site CCAFS SLC-40).
- Proximities in kilometers are provided in the map screenshots.


Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



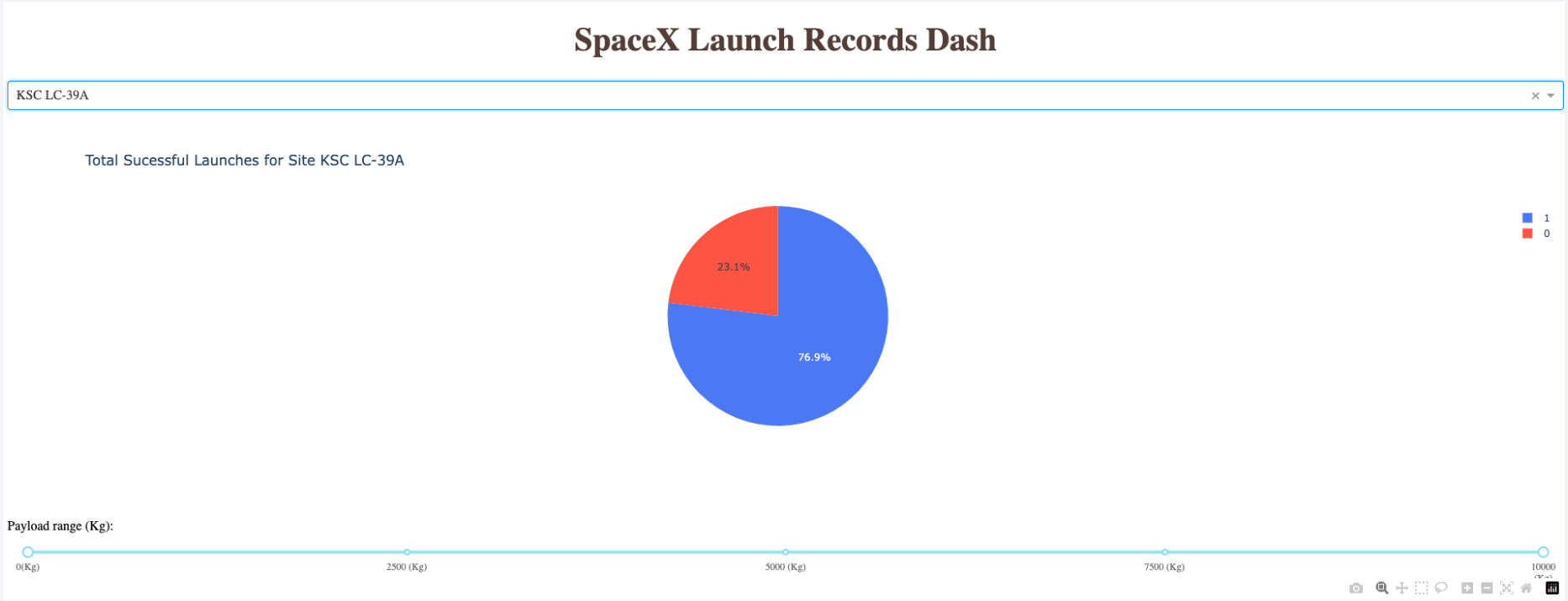
The screenshot provided displays a SpaceX Launch Records Dashboard. It contains a dropdown currently set to "All Sites". Users can also filter the data by different launch sites. This interactivity allows users to analyze launch statistics for specific locations. For each launch site selection a pie chart showing the proportion of successful launches is displayed. Currently displayed is the pie chart for proportions of successful launches for all of the sites.

The legend on the right maps colors to specific launch sites:

- Blue (41.7%): KSC LC-39A
- Red (29.2%): CCAFS LC-40
- Green (16.7%): VAFB SLC-4E
- Purple (12.5%): CCAFS SLC-40

Interpretation: KSC LC-39A has the highest percentage of successful launches, followed by CCAFS LC-40. VAFB SLC-4E and CCAFS SLC-40 have fewer successful launches.

Launch Success Proportion (Site KSC LC-39A)

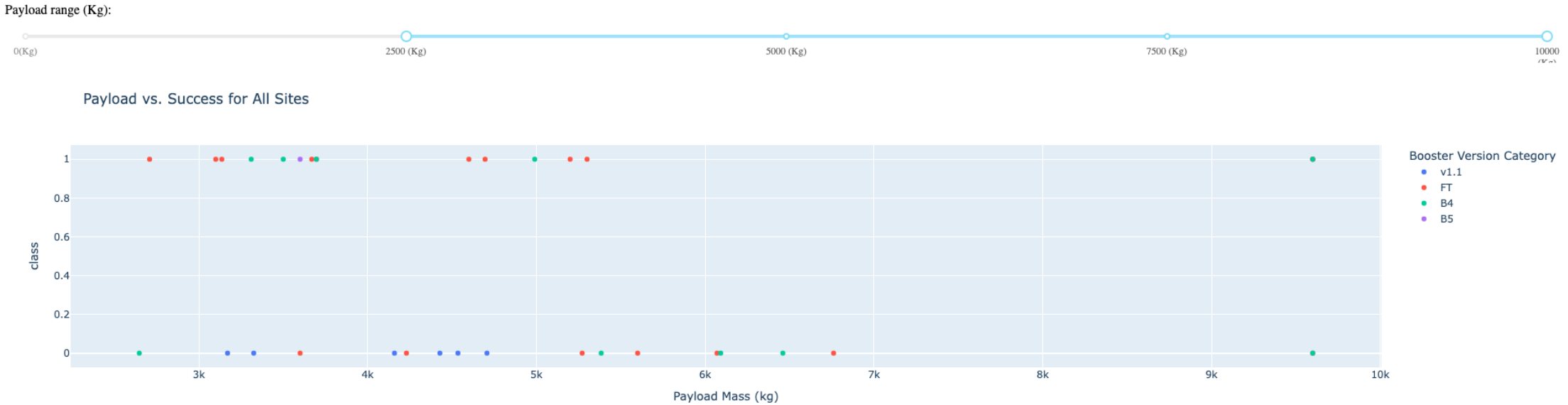


The dropdown has been set to "KSC LC-39A", filtering the dashboard to show data only for this launch site. The pie chart shown here is proportion of Success vs Failure for site KSC LC-39A.

Legend: Blue (76.9%) → Successful launches (labelled as 1) and Red (23.1%) → Failed launches (labelled as 0)

i.e. KSC LC-39A has a high success rate (76.9%), making it a reliable launch site with relatively low failures (23.1%), but failure is still present.

Payload Mass vs Success (every site) by Booster Version



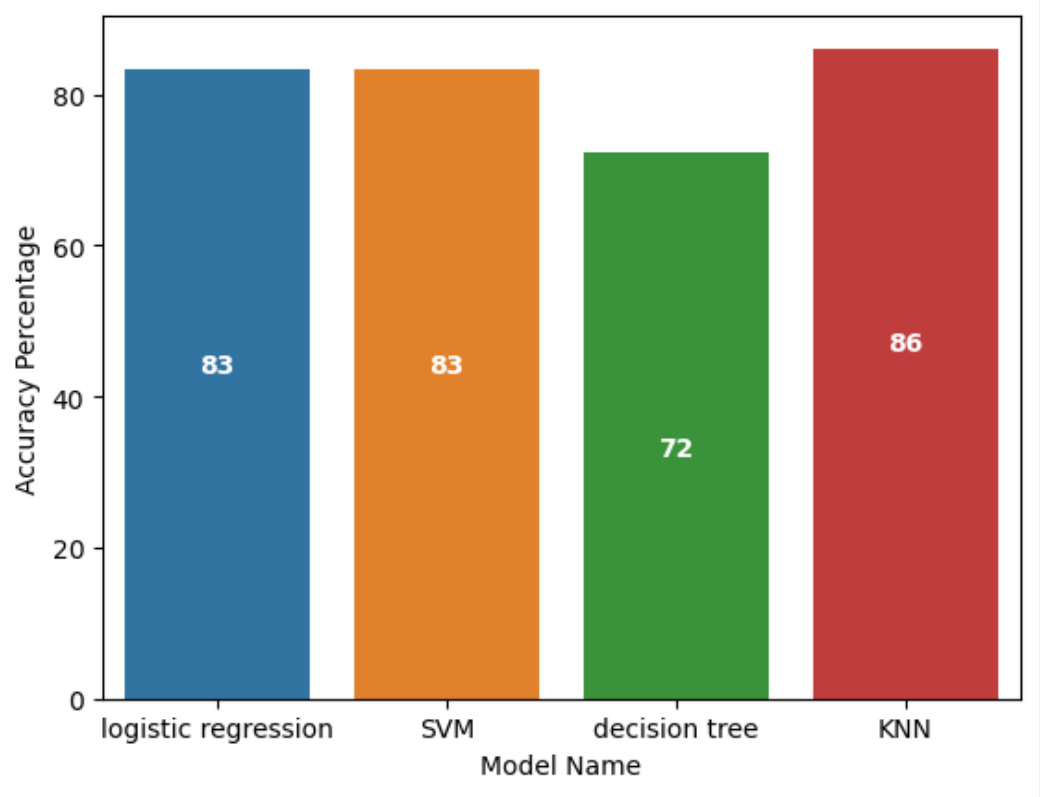
This section of the SpaceX Launch Records Dashboard focuses on the relationship between payload mass and launch success for all sites, with the payload range filter set between 2,500 kg and 10,000 kg. The scatter plot visualises launches across this weight range, where the x-axis represents payload mass and the y-axis represents launch outcome (1 for success, 0 for failure). Each dot corresponds to an individual launch, and the legend on the right categorises launches by booster version, including v1.1 (blue), FT (red), B4 (green), and B5 (purple). The data reveals that launches with payloads between 3,000 kg and 5,000 kg tend to have the highest success rates, while failures become more frequent for payloads exceeding 6,000 kg. The FT booster version (red) appears in many successful launches, suggesting it has strong reliability, whereas v1.1 (blue) shows a mix of successes and failures, indicating lower consistency. The B4 and B5 boosters (green and purple) appear less frequently but show a higher proportion of success, implying that these newer boosters are more reliable. Failures are scattered across all payloads but become more concentrated above 5,000 kg, with FT and v1.1 boosters contributing to most of these failures. This analysis suggests that SpaceX achieves the highest launch success with FT boosters and payloads under 5,000 kg, while higher payloads and older boosters show increased failure risk.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The bar chart visualises the accuracy percentage of the four models.
- The KNN model has the highest accuracy (86%), making it the best classifier among the four.
- Logistic Regression and SVM have the same accuracy (83%), which is slightly lower than KNN.
- The Decision Tree model performs the worst (72%), indicating it may not generalise well for this problem.



Confusion Matrix

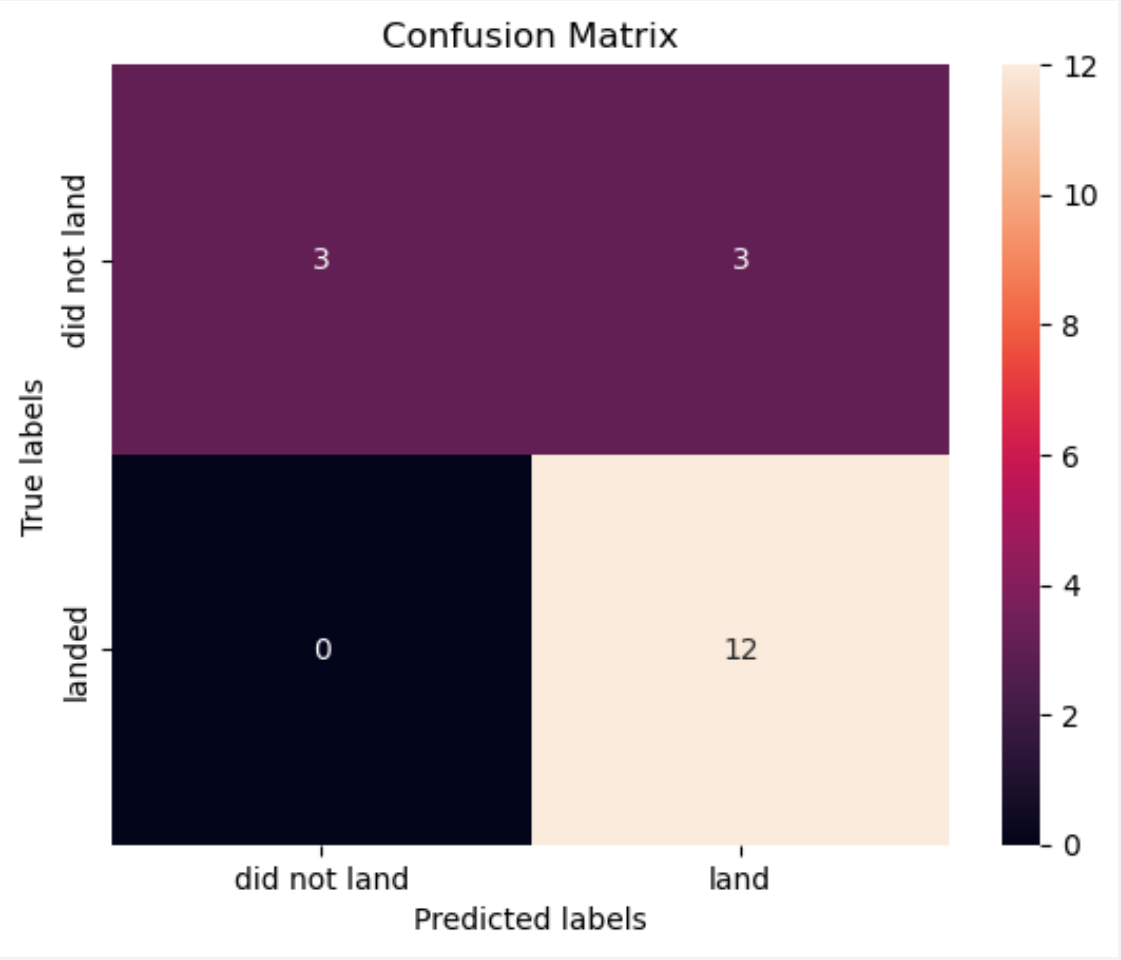
This confusion matrix represents the performance of the K-Nearest Neighbours (KNN) model in predicting whether a SpaceX rocket will successfully land.

Interpretation of the Confusion Matrix:

- True Positive (TP) = 12: The model correctly predicted 12 instances where the rocket successfully landed.
- True Negative (TN) = 3: The model correctly predicted 3 instances where the rocket did not land.
- False Positive (FP) = 3: The model incorrectly predicted that the rocket would land when it actually did not.
- False Negative (FN) = 0: The model did not misclassify any actual landings as failures.

Performance Insights:

- The accuracy of the model seems high, as there are more correct predictions (TP + TN) than incorrect ones.
- The false negative rate is 0, meaning the model never missed a successful landing.
- However, the false positive rate is moderate (3 cases), meaning it sometimes incorrectly predicts a landing when the rocket does not land.
- Since landing predictions are critical, having 0 false negatives is a strong positive, as the model reliably identifies successful landings.



Conclusions

This project provided a comprehensive analysis of SpaceX launch data, revealing key factors that influence mission success and landing outcomes. Through exploratory data analysis, we identified that launch sites such as CCAFS LC-40 and KSC LC-39A have the highest success rates, with KSC LC-39A particularly excelling in mid-range payloads. Payload mass plays a significant role in determining mission outcomes, with heavier payloads generally exhibiting higher success rates, while failures are more common at lower payload ranges. Orbit-specific analysis highlighted that ES-L1, GEO, HEO, and SSO have near-perfect success rates, while LEO, ISS, MEO, and PO show moderate success rates between 60% and 70%, and GTO remains a challenge with only around 50% success. SQL queries and interactive visualizations further contextualized trends in launch performance, infrastructure proximity, and historical mission reliability. Predictive modeling using machine learning demonstrated that classification algorithms can effectively determine landing success, with K-Nearest Neighbors (KNN) achieving the highest accuracy at 86%. Over time, SpaceX has shown significant improvements in launch reliability, with success rates rising sharply from 2013 onward and surpassing 80% by 2017, reflecting advancements in technology and operational efficiency. These findings not only provide valuable insights into SpaceX's mission success factors but also offer a data-driven foundation for optimizing future launches and enhancing predictive capabilities in aerospace analytics.

Appendix

- Full GitHub Repo found at:

https://github.com/NwtsN/IBM_course_capstone/tree/main

Thank you!