IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

This capstone project focuses on predicting the success of Falcon 9 first stage landings using real-world data from SpaceX. The objective is to analyze historical launch information to uncover patterns that influence successful landings and to build a predictive model using supervised machine learning techniques.

Data was collected from the official SpaceX API and supplemented with mission metadata via web scraping from Wikipedia. The dataset was cleaned, merged, and enriched with relevant features including launch site, payload mass, orbit, and booster version. Exploratory Data Analysis (EDA) was conducted using Python visualizations and SQL queries, followed by interactive visual analytics through Folium and Plotly Dash.

Several classification models were developed, including Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors. Hyperparameter tuning was performed using GridSearchCV with cross-validation. The best-performing model was the Decision Tree Classifier, achieving a validation accuracy of 90.36% and consistent performance on test data. The project also includes an interactive dashboard and geospatial visualizations to enhance interpretability and stakeholder engagement.

Introduction

SpaceX has revolutionized space transportation by developing reusable rocket technology, particularly with the Falcon 9 rocket. The ability to land the first stage after launch is critical for reducing mission costs and ensuring sustainable operations. Predicting the success of these landings can provide strategic insights for mission planning, risk management, and design improvements.

Key Questions Addressed:

- What are the main factors that influence the success or failure of Falcon 9 first stage landings?
- Can a predictive model be developed using historical mission data to estimate landing outcomes?
- How do launch site, payload mass, orbit type, and booster version affect landing success rates?

Section 1

Methodology

Methodology

Executive Summary

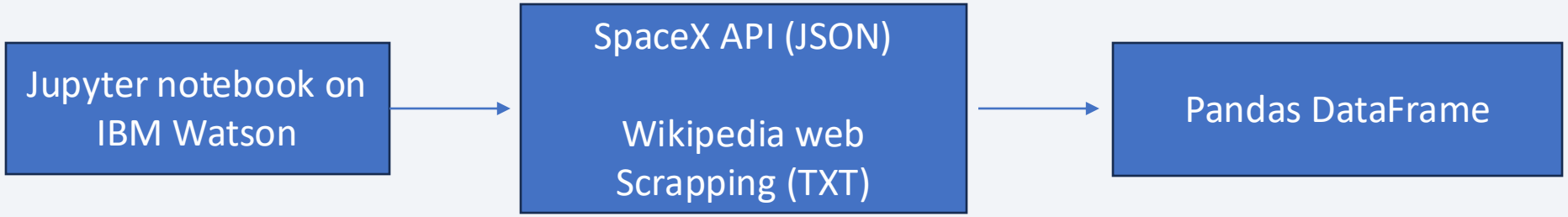
- Data collection methodology: Data was collected using SpaceX rest API, and web scrapping from Wikipedia.
- Perform data wrangling: The data was preprocessed using pandas. Some techniques are: OneHot encoding, data normalization and standardization.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Splitting data into train and test sets.
    - Identifying the best algorithm and parameters.
    - Adopting the best algorithm and parameters for model deployment.

Data Collection

Data sources:

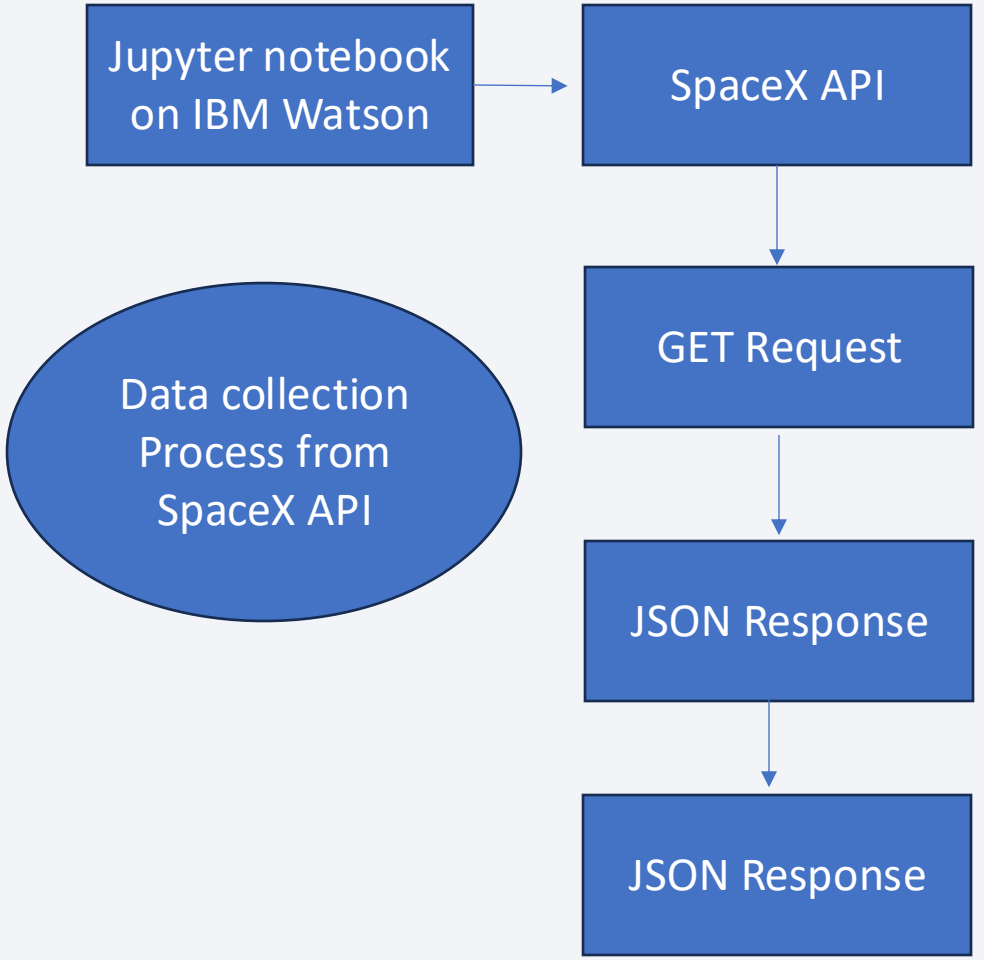
- SpaceX API: Open-source REST API for various rocket landing data.
- Wikipedia: Free online encyclopedia, created and edited by volunteers.

- The process of data collection was the following:



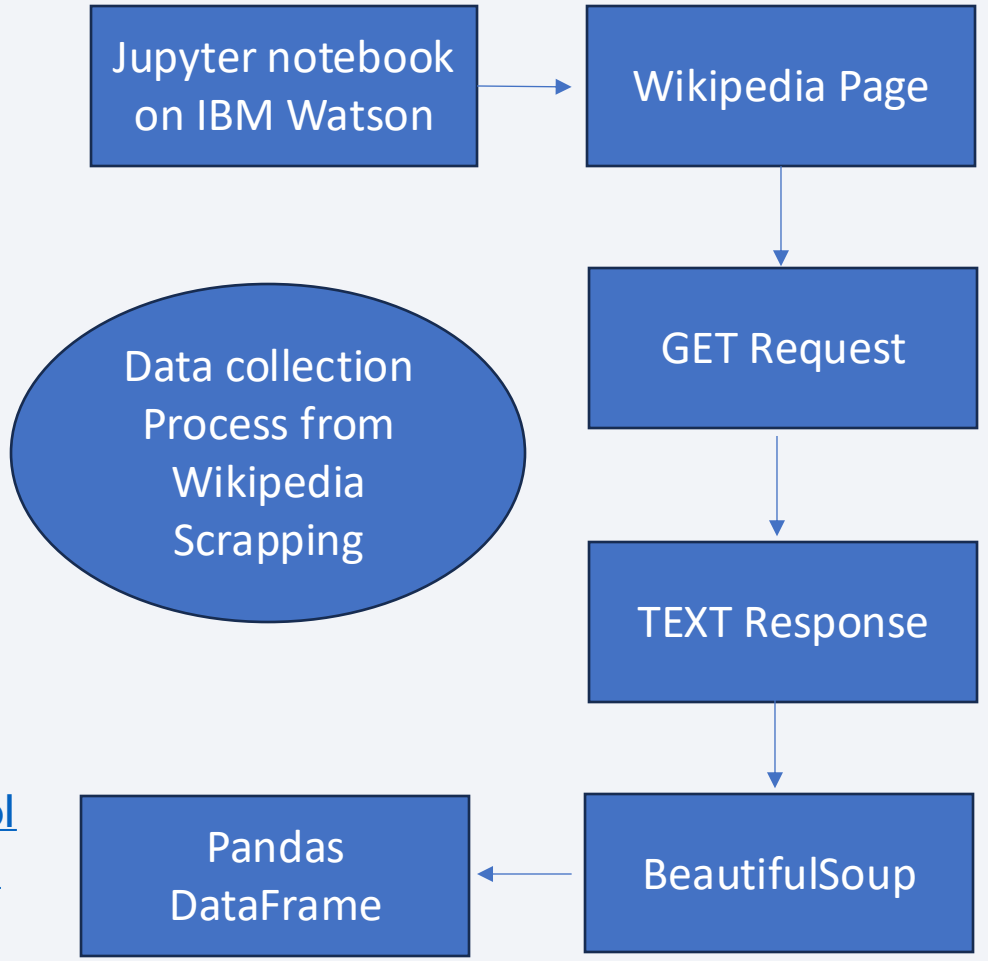
Data Collection – SpaceX API

- Accessed the official SpaceX REST API to retrieve historical launch data.
- Used Python's requests library to send HTTP GET requests and obtain structured JSON data.
- Parsed and normalized JSON responses using json_normalize and converted them into Pandas DataFrames.
- Focused on extracting key fields such as flight_number, launch_date, launch_site, rocket, payload, and landing outcome.
- Saved the cleaned API response data locally for integration with web-scraped sources.
- Ensured reproducibility by wrapping the data collection process in reusable Python functions.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/01_Complete_Data_Collection_API_Lab.ipynb



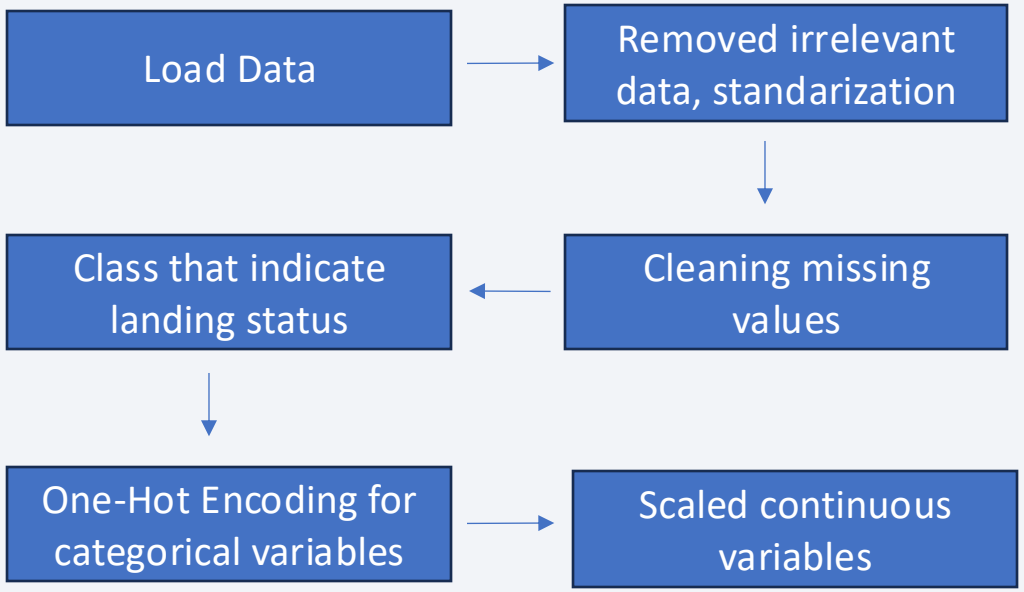
Data Collection - Scraping

- Performed web scraping on Wikipedia to extract supplementary data not provided by the SpaceX API.
- Targeted tables containing launch history, booster versions, payload mass, and landing outcomes.
- Used Python's requests library to fetch HTML content and BeautifulSoup for parsing.
- Extracted and cleaned table data, handling inconsistencies in formatting and missing values.
- Merged the scraped dataset with the API dataset on shared fields such as flight number and launch date.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/02_Complete_Data_Collection_Web_Scraping%20lab.ipynb



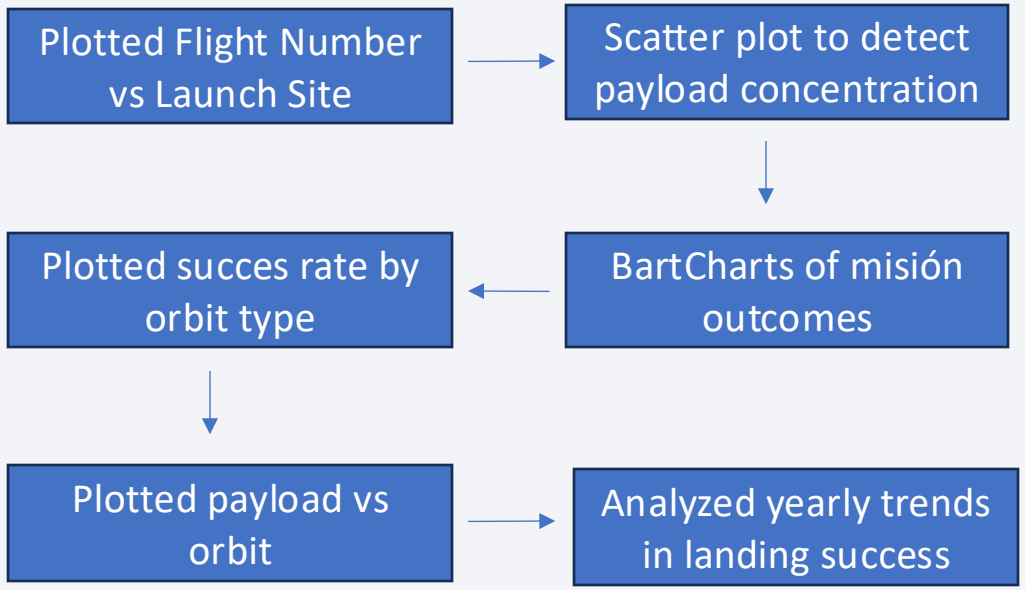
Data Wrangling

- Merged API and web-scraped datasets into a unified Pandas DataFrame using flight_number as the key.
- Removed irrelevant columns and standardized feature names for consistency.
- Handled missing values using imputation and conditional filtering.
- Created the binary target variable Class to indicate landing success (1) or failure (0).
- Applied One-Hot Encoding to convert categorical features (e.g., Orbit, Launch Site, Booster Version) into numerical format.
- Scaled continuous variables such as Payload Mass using StandardScaler to prepare for model training.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/03_Data_Wrangling.ipynb



EDA with Data Visualization

- Plotted Flight Number vs Launch Site to identify the distribution of launches across locations.
- Created Payload Mass vs Launch Site scatter plots to detect payload concentration patterns.
- Built bar charts of mission outcomes to examine the frequency of successful and failed landings.
- Visualized success rate by orbit type using grouped bar charts.
- Generated scatter plots for payload vs. orbit to observe how orbit selection may influence payload mass and outcomes.
- Analyzed yearly trends in landing success using line plots to reveal temporal performance improvements.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/04_EDA_with_Visualization.ipynb



EDA with SQL

- Queried all unique launch site names from the dataset to identify total distinct launch locations.
- Filtered launch sites starting with 'CCA' to focus on Cape Canaveral-based missions.
- Calculated total payload mass carried by boosters affiliated with NASA.
- Computed the average payload mass for missions using the F9 v1.1 booster version.
- Retrieved the date of the first successful ground pad landing.
- Listed boosters with successful drone ship landings carrying payloads between 4000 and 6000 kg.
- Counted total successes and failures in mission outcomes.
- Identified boosters that carried the maximum payload mass ever launched.
- Found failed drone ship landings in 2015, showing associated boosters and launch sites.
- Ranked all landing outcomes between June 2010 and March 2017 in descending frequency.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/05_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Added circle markers to indicate launch site locations, with size and color representing launch outcome success rate.
- Included popup tooltips with site name and coordinates to enable user-friendly exploration.
- Used colored markers (green for success, red for failure) to represent individual mission outcomes by site.
- Plotted lines from each launch site to nearby infrastructure (e.g., highways, railways, coastline) to visualize proximity.
- Calculated and displayed distances from launch sites to key infrastructure points using great-circle distance calculations.
- Purpose: These map objects enhance spatial understanding of launch site distribution, landing outcome patterns, and the influence of infrastructure proximity on mission planning and success.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/06_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

Plots and Interactions Added:

- Dropdown menu to filter data by launch site (e.g., all sites vs. individual sites).
- Pie chart displaying total success counts or success vs. failure distribution per site.
- Range slider to filter missions by payload mass interval.
- Scatter plot showing the correlation between payload mass and landing success, color-coded by booster version.

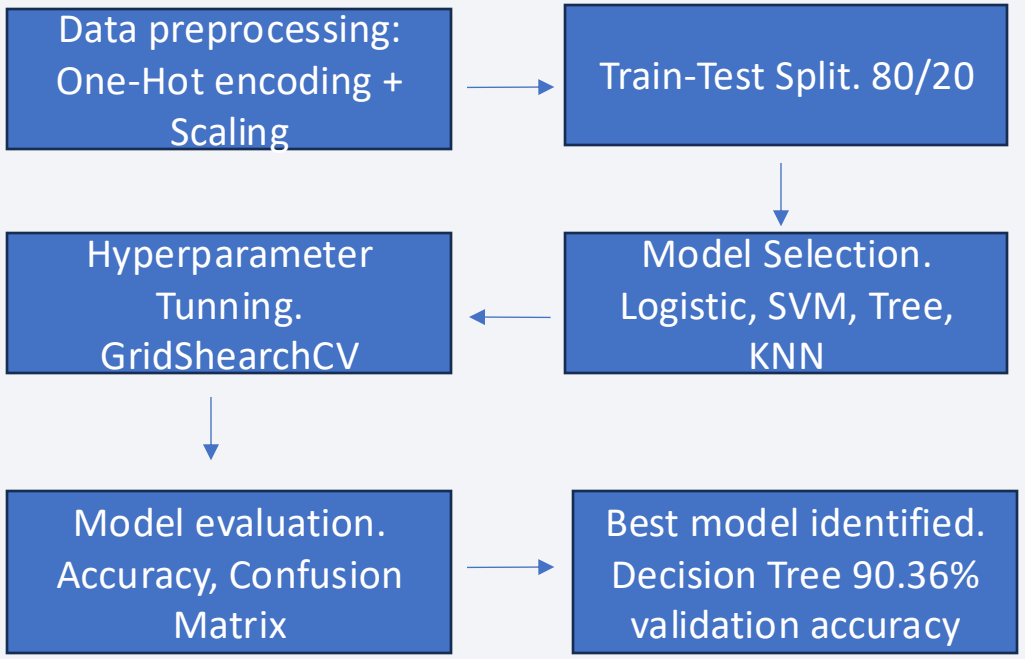
Why These Elements Were Included:

- The dropdown enables focused analysis per launch site and reveals site-specific trends.
- The pie chart provides a quick and intuitive overview of mission success distribution.
- The range slider allows dynamic payload filtering to explore how payload mass influences outcomes.
- The scatter plot illustrates multivariate relationships, helping identify optimal payload ranges and booster performance.
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/07_spacex-dash-app.py

Predictive Analysis (Classification)

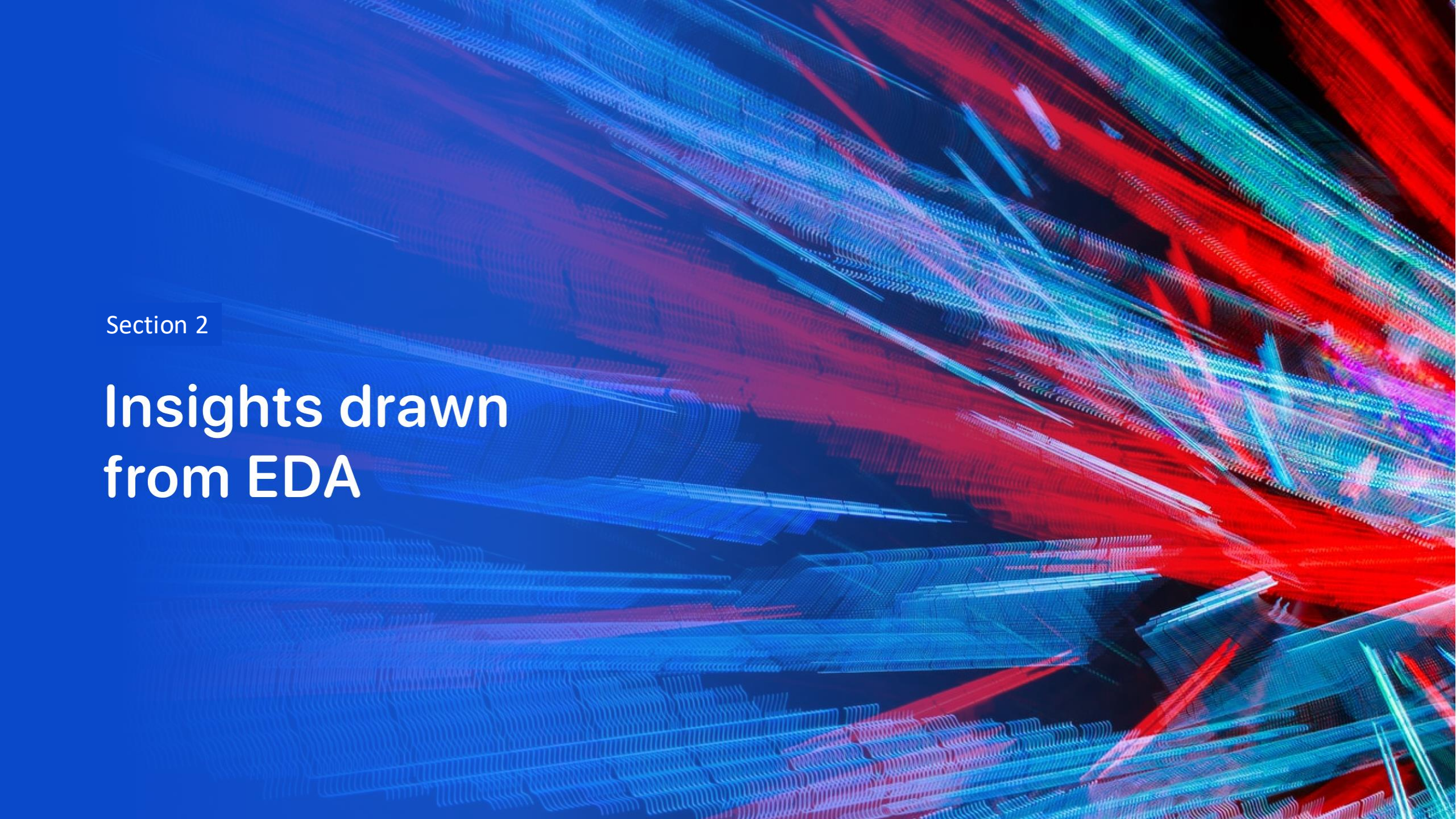
Model Development Process:

- Cleaned and standardized features; applied one-hot encoding to categorical variables.
- Scaled numeric features using StandardScaler to ensure uniform input for all models.
- Split the data into training and test sets using train_test_split with 80/20 ratio.
- Implemented multiple classifiers: Logistic Regression, SVM, Decision Tree, and KNN.
- Tuned hyperparameters for each model using GridSearchCV with 10-fold cross-validation.
- Selected the best model (Decision Tree Classifier) based on validation accuracy (90.36%).
- https://github.com/MatthiasAlmonacid/proyecto_10/blob/main/08_Machine_Learning_Prediction.ipynb

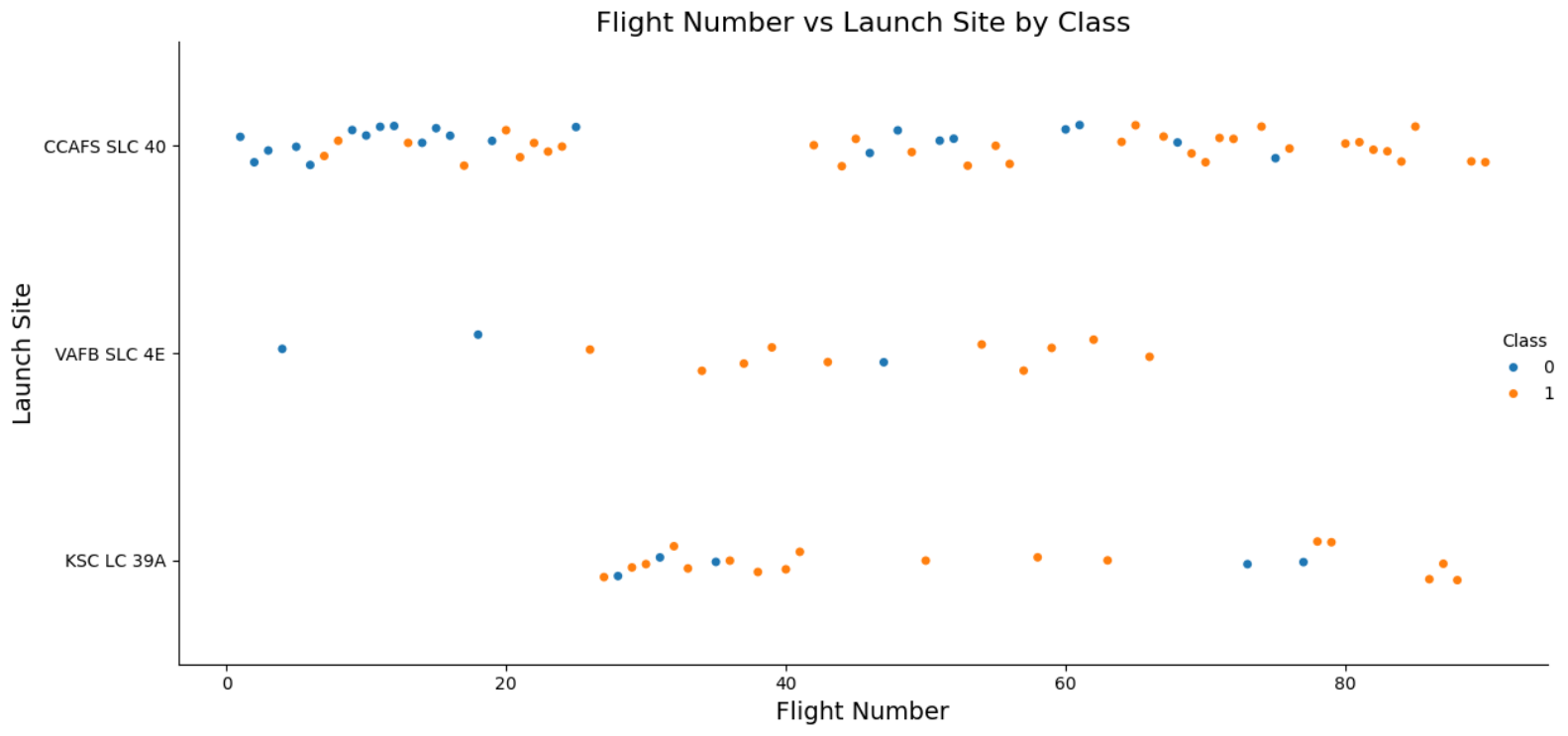


Results

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

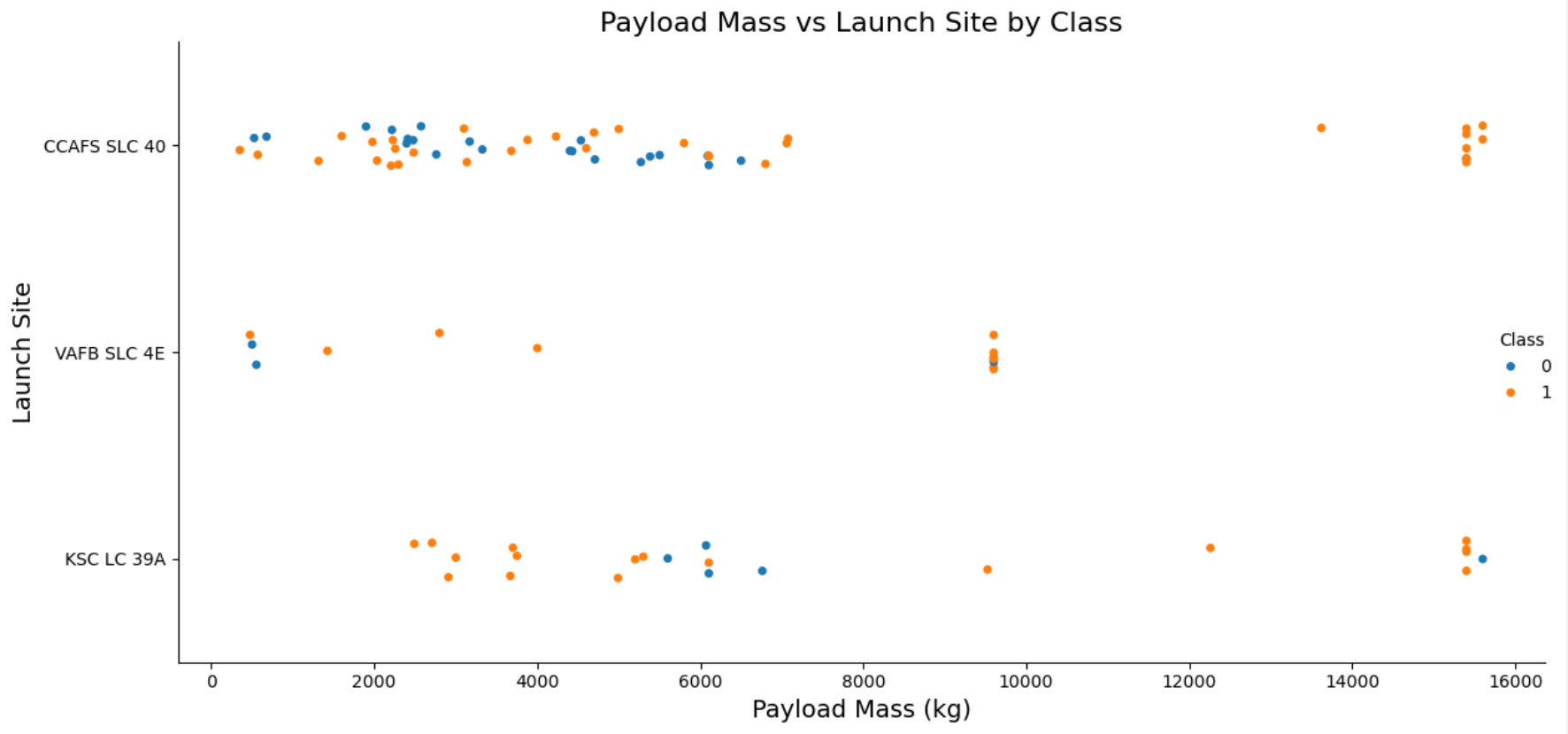

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



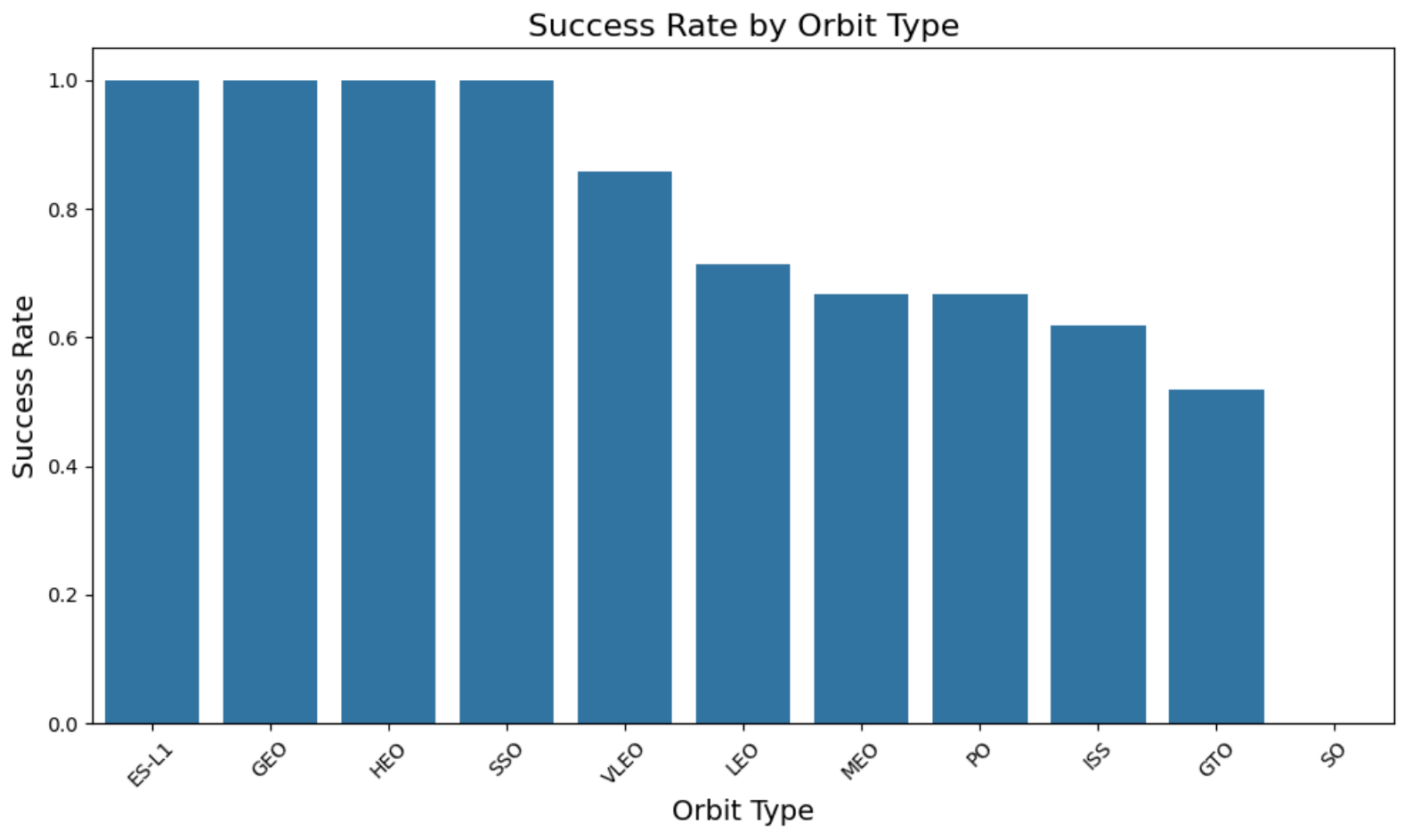
We observe that CCAFS SLC 40 has the highest number of launches, with increasing landing success over time. KSC LC 39A also shows a strong pattern of success in later missions, while VAFB SLC 4E has more variability. The trend suggests that experience over time and site-specific conditions may influence landing outcomes.

Payload vs. Launch Site



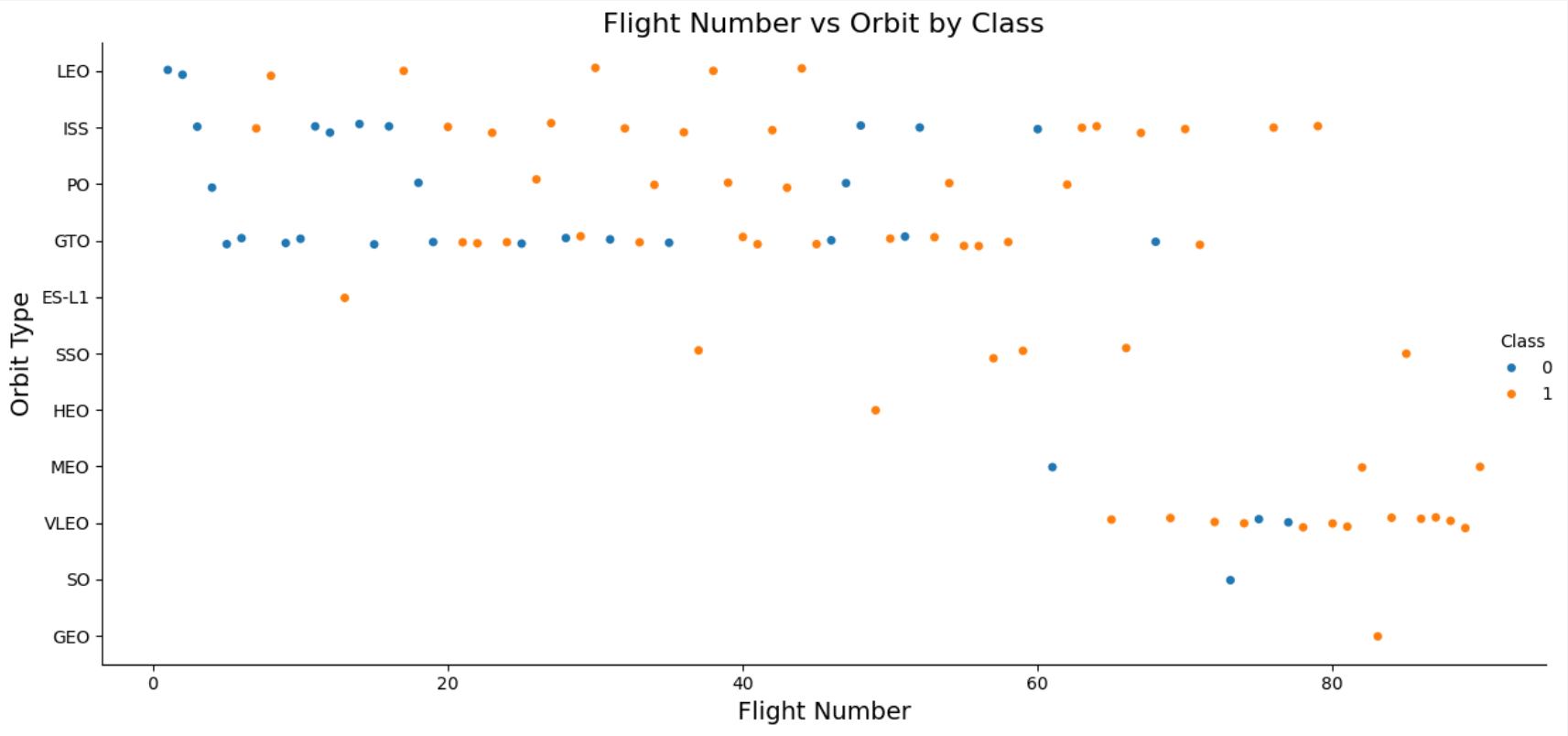
The plot reveals that CCAFS SLC 40 has the widest range of payload masses and a balanced distribution of success and failure. KSC LC 39A shows a high success rate, even with heavier payloads, while VAFB SLC 4E appears less consistent. This visualization suggests that site-specific factors and payload weight may influence landing outcomes.

Success Rate vs. Orbit Type



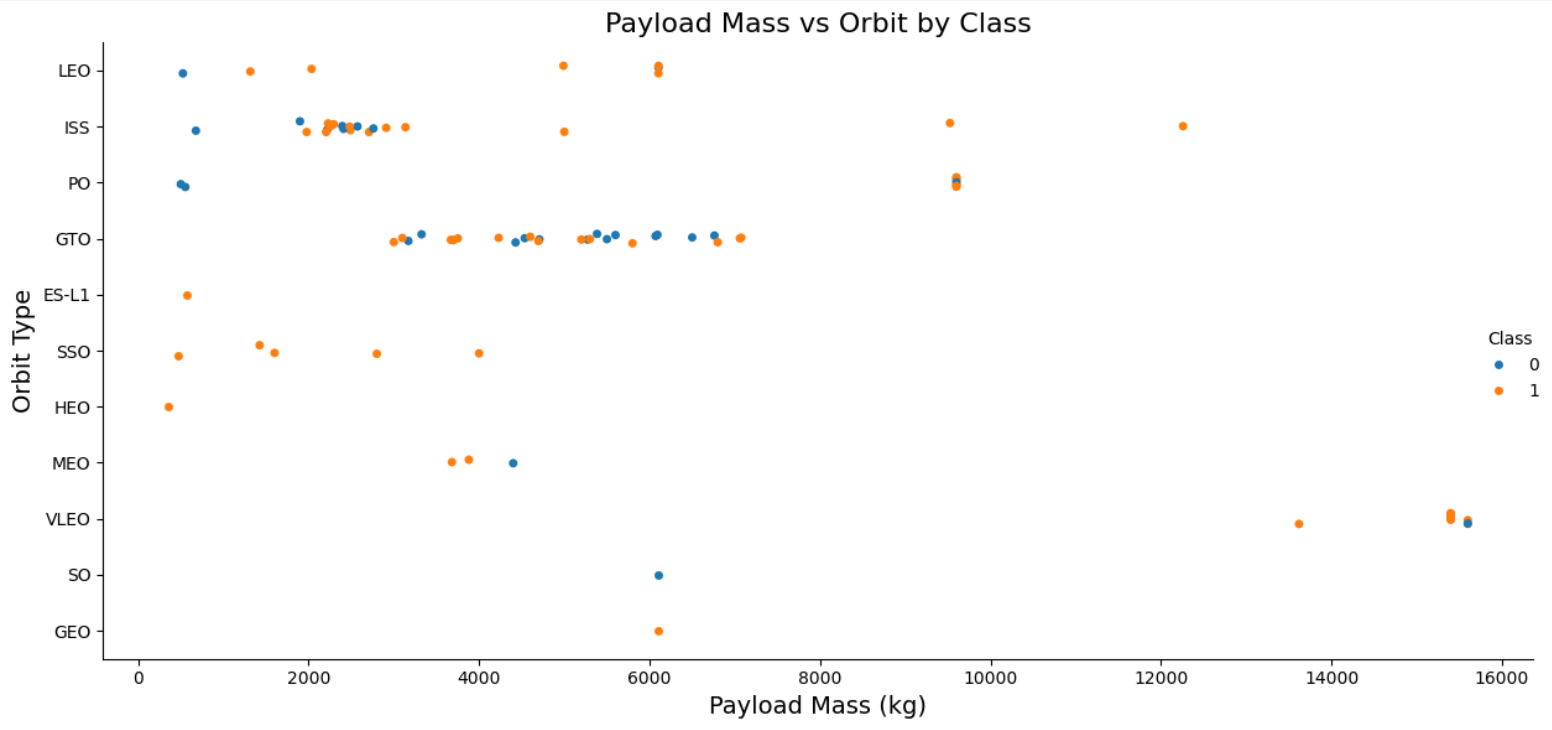
This bar chart shows the average success rate of Falcon 9 landings for different orbit types.ES-L1, GEO, HEO, and SSO achieved perfect landing records (100%), while GTO and SO show lower performance. The chart suggests that missions to higher or more stable orbits tend to have higher landing success, possibly due to optimized mission profiles or lower re-entry complexity. In contrast, GTO missions—often with heavier payloads and longer burns—present more landing challenges.

Flight Number vs. Orbit Type



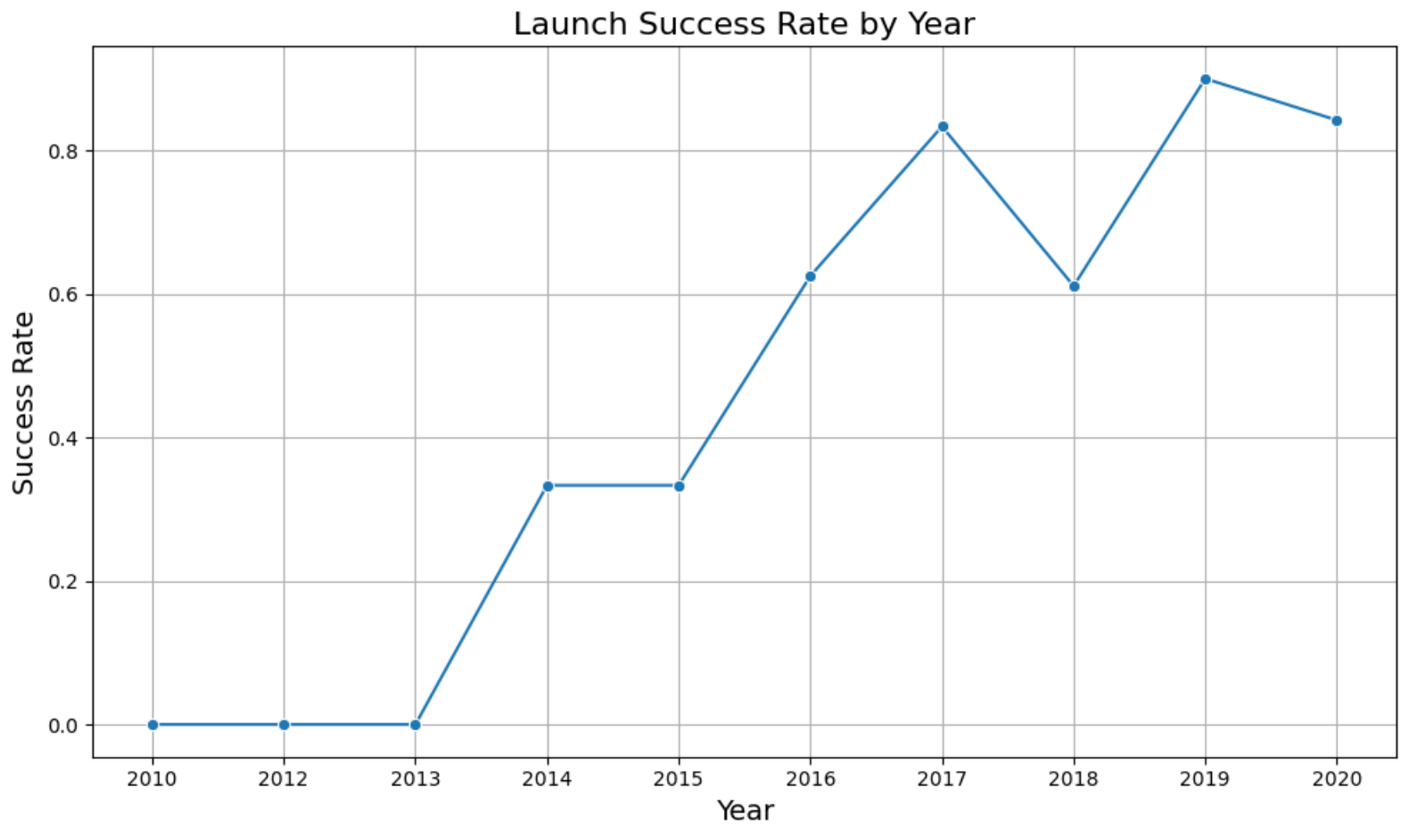
We observe that early missions had more failures across all orbit types. Over time, especially in later flight numbers, success becomes more consistent—particularly for orbits like ISS, LEO, and VLEO. This indicates a learning curve and technical improvements by SpaceX in achieving reliable landings across various orbital missions.

Payload vs. Orbit Type



We observe that successful landings (Class 1) occurred across various orbit types, particularly LEO, GTO, and ISS, regardless of payload mass. However, heavier payloads beyond 10,000 kg are less frequent and more prone to failure. This suggests that both orbit type and payload weight play important roles in landing success.

Launch Success Yearly Trend



This line chart shows the annual success rate of Falcon 9 first stage landings from 2010 to 2020.A clear upward trend is visible, reflecting significant improvements in SpaceX's landing technology and reliability. After 2015, success rates increased steadily, peaking at over 90% in 2019. The trend highlights continuous progress in mission control, engineering, and landing precision.

All Launch Site Names

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

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

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

This query retrieves the distinct names of all launch sites from the dataset.

It allows us to identify which sites SpaceX has used for Falcon 9 missions.

Launch Site Names Begin with 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

This query filters the dataset to retrieve records where the launch site name starts with 'CCA', which corresponds to Cape Canaveral Air Force Station (CCAFS).

Limiting the output to 5 records allows a quick look at typical missions launched from this location.

Total Payload Mass

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

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

Total_Payload
48213

This query calculates the total payload mass (in kilograms) for all launches associated with NASA Commercial Resupply Services (CRS) missions. Filtering by the customer field ensures that only missions contracted by NASA are included in the sum. The total payload carried for these missions was 48,213 kg, demonstrating NASA's significant contribution to Falcon 9 launch activity.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Avg_Payload
2928.4

This query calculates the average payload mass carried by Falcon 9 v1.1 booster versions.

The result of 2,928.4 kg provides insight into the typical payload capacity handled by this specific variant of the Falcon 9 rocket.

First Successful Ground Landing Date

```
%%sql SELECT MIN(Date) AS First_Successful_Ground_Pad_Landing
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Ground_Pad_Landing
2015-12-22

This query retrieves the earliest date on which a Falcon 9 first stage successfully landed on a ground pad.

The result shows that the first ground landing occurred on December 22, 2015, marking a major milestone in SpaceX's reusability efforts.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 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 all booster versions that successfully landed on a drone ship while carrying payloads between 4000 and 6000 kg.

These missions demonstrate the reliability of Falcon 9 Full Thrust boosters in handling mid-weight payloads with precise offshore landings.

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

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

This query counts the total number of missions by outcome type. The vast majority of missions were successful (98 out of 101), highlighting SpaceX's strong performance record.

A small number of exceptions include in-flight failure and missions with unclear payload status. The duplicate "Success" label may indicate a data labeling issue.

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

This query identifies all Falcon 9 Block 5 booster versions that transported the heaviest payloads, providing insight into which boosters were deployed for high-capacity missions.

2015 Launch Records

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

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

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

This query lists all drone ship landing failures that occurred in the year 2015.

Both failures involved F9 v1.1 boosters and took place at the Cape Canaveral SLC-40 launch site.

This highlights early challenges in SpaceX's offshore landing operations.

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

```
%%sql 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;
```

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

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

This query summarizes all landing outcomes between June 2010 and March 2017.

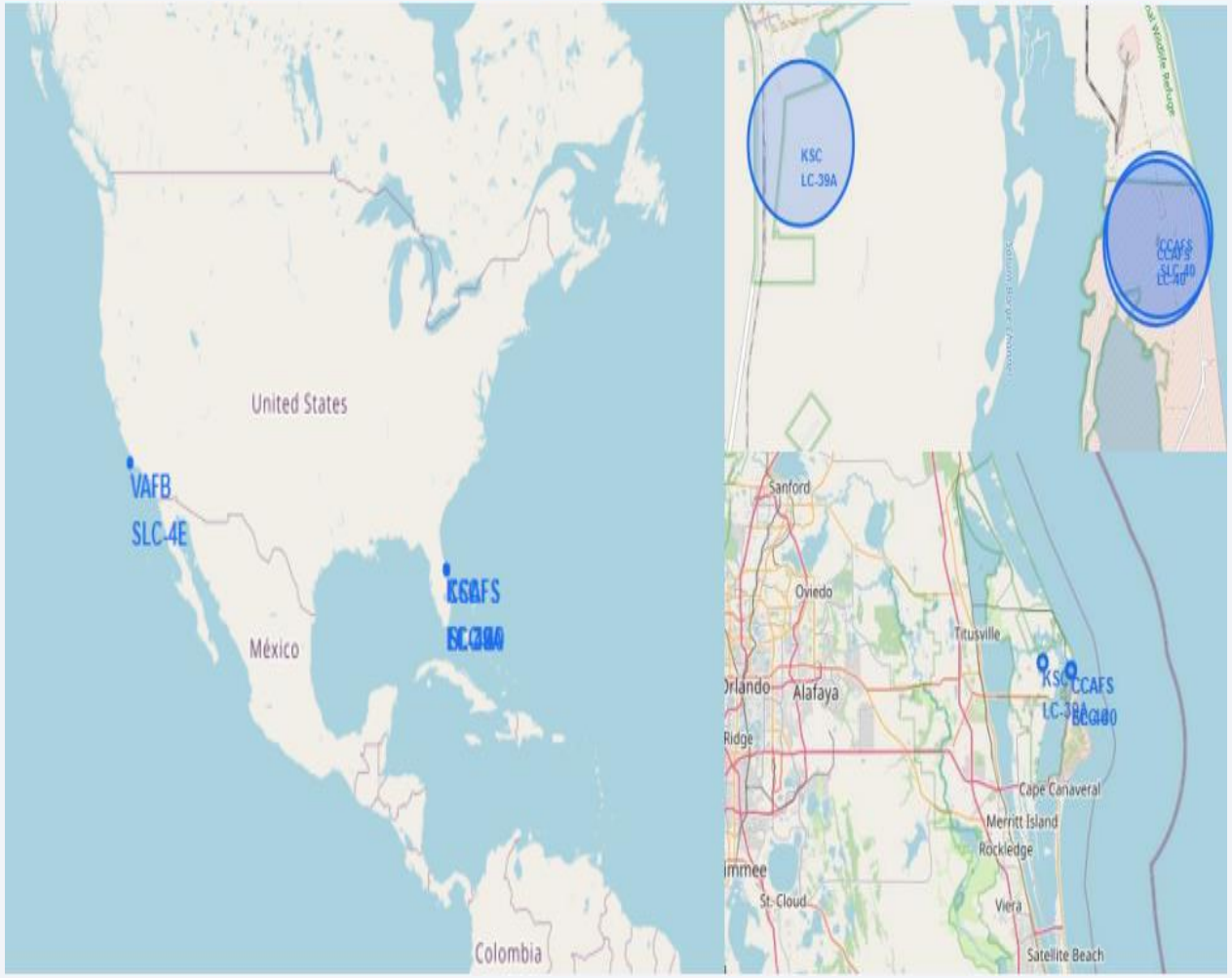
Most early missions had no landing attempt, followed by mixed drone ship results.

The data reveals SpaceX's gradual transition from experimental recovery phases to more consistent landing attempts, particularly using drone ships and ground pads.

Section 3

Launch Sites Proximities Analysis

Launch Sites



This map presents the three primary launch sites used by SpaceX for Falcon 9 missions.

Key Elements Highlighted:
Launch locations are visualized with circle markers and site labels.
Inset maps on the right zoom into Cape Canaveral, showing the proximity of KSC and CCAFS.

This geographic representation supports visual comparison of site density, proximity to coastlines, and infrastructure.This map aids in understanding logistical factors such as trajectory needs, recovery operations, and launch frequency per site.

Success rate for launch locations

This Folium-based visualization displays clustered markers of launch events across all sites:

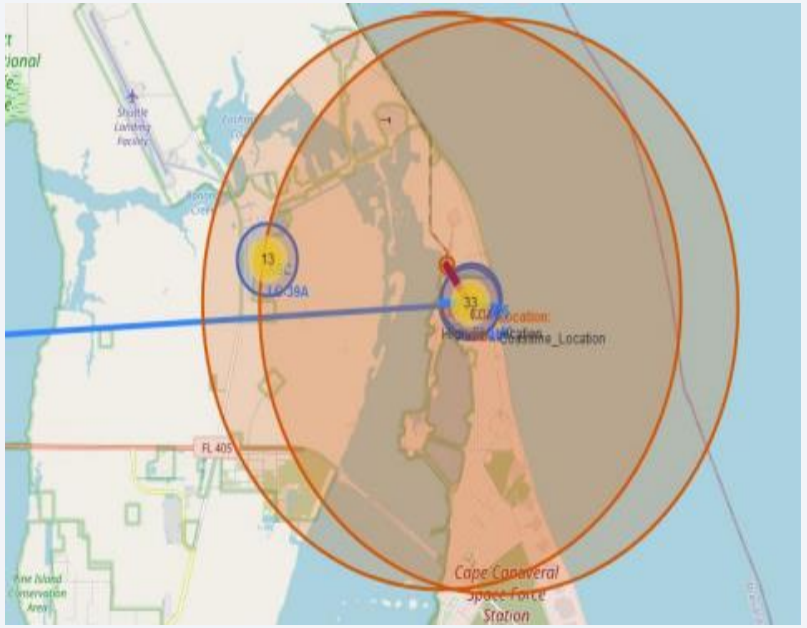
Marker Clusters show total launches per site (e.g., 46 at CCAFS, 10 at VAFB).

Color-coded icons: success, failure, grouped counts.

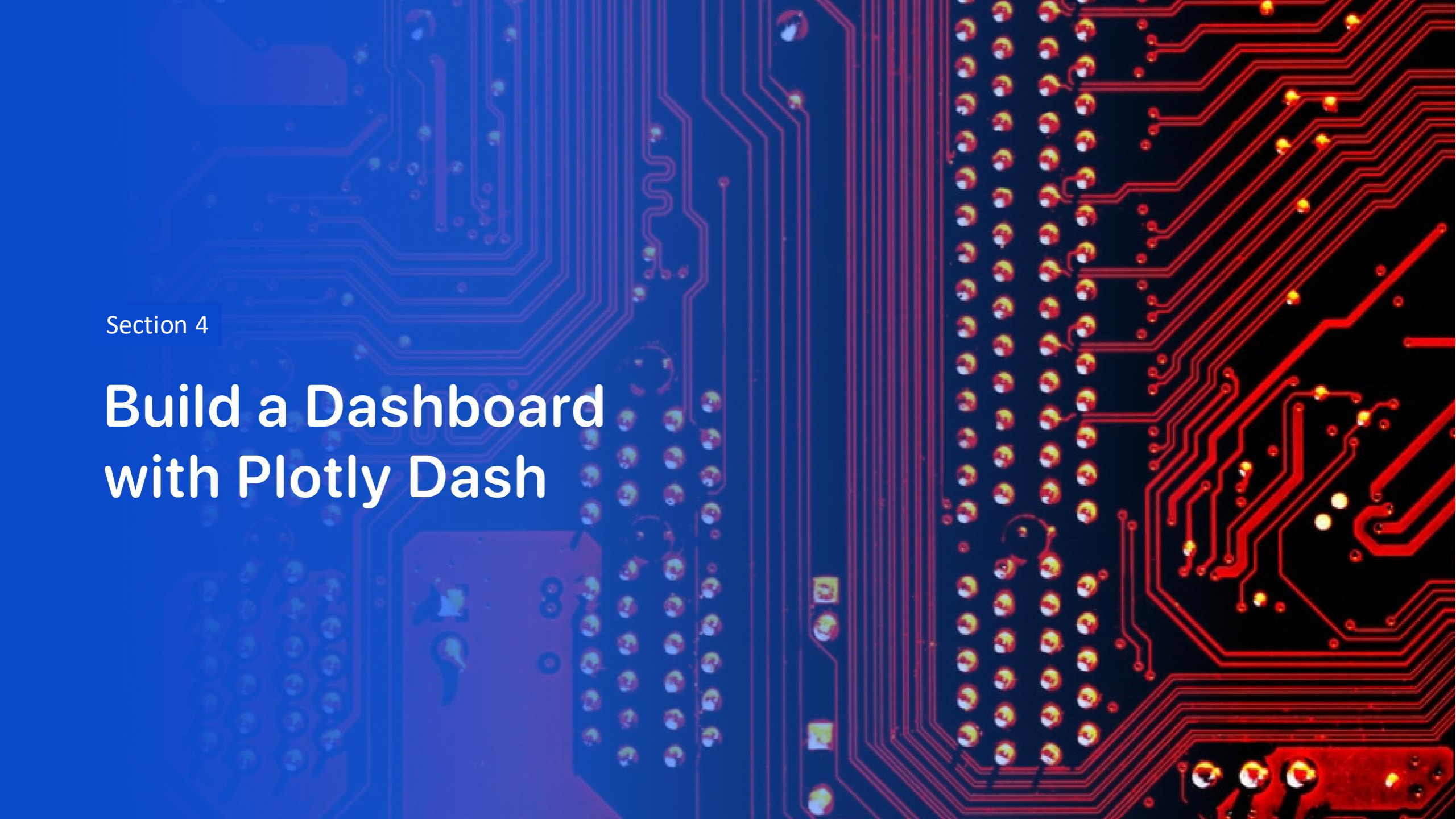
Zoomed insets reveal detailed launch distributions near Cape Canaveral and Vandenberg.

These patterns highlight site activity volume and success/failure dispersion.

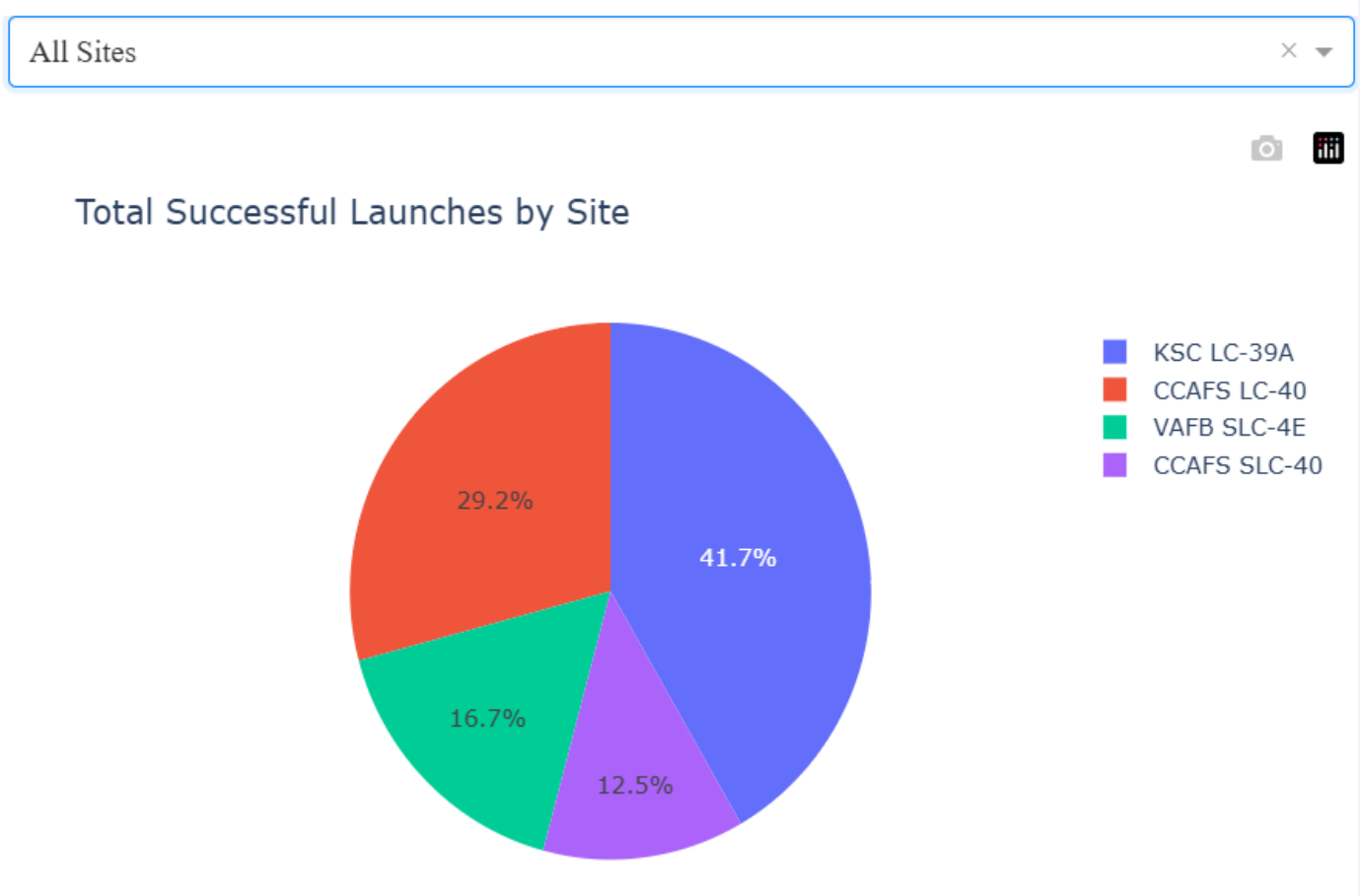
<Folium Map Screenshot 3>



These Folium-based maps display the distance and spatial relation between CCAFS SLC-40, KSC LC-39A, and nearby infrastructure: Concentric circles represent distance buffers (e.g., 10 km, 20 km).
Lines indicate proximity to reference points such as the coastline and highways.
Zoomed view highlights exact launch coordinates at CCAFS with infrastructure context.
Insight: Cape Canaveral's strategic positioning offers excellent access to coastlines and infrastructure—an important factor for launch logistics, payload recovery, and safety zones.


Section 4
Build a Dashboard with Plotly Dash

Launch success of all siter

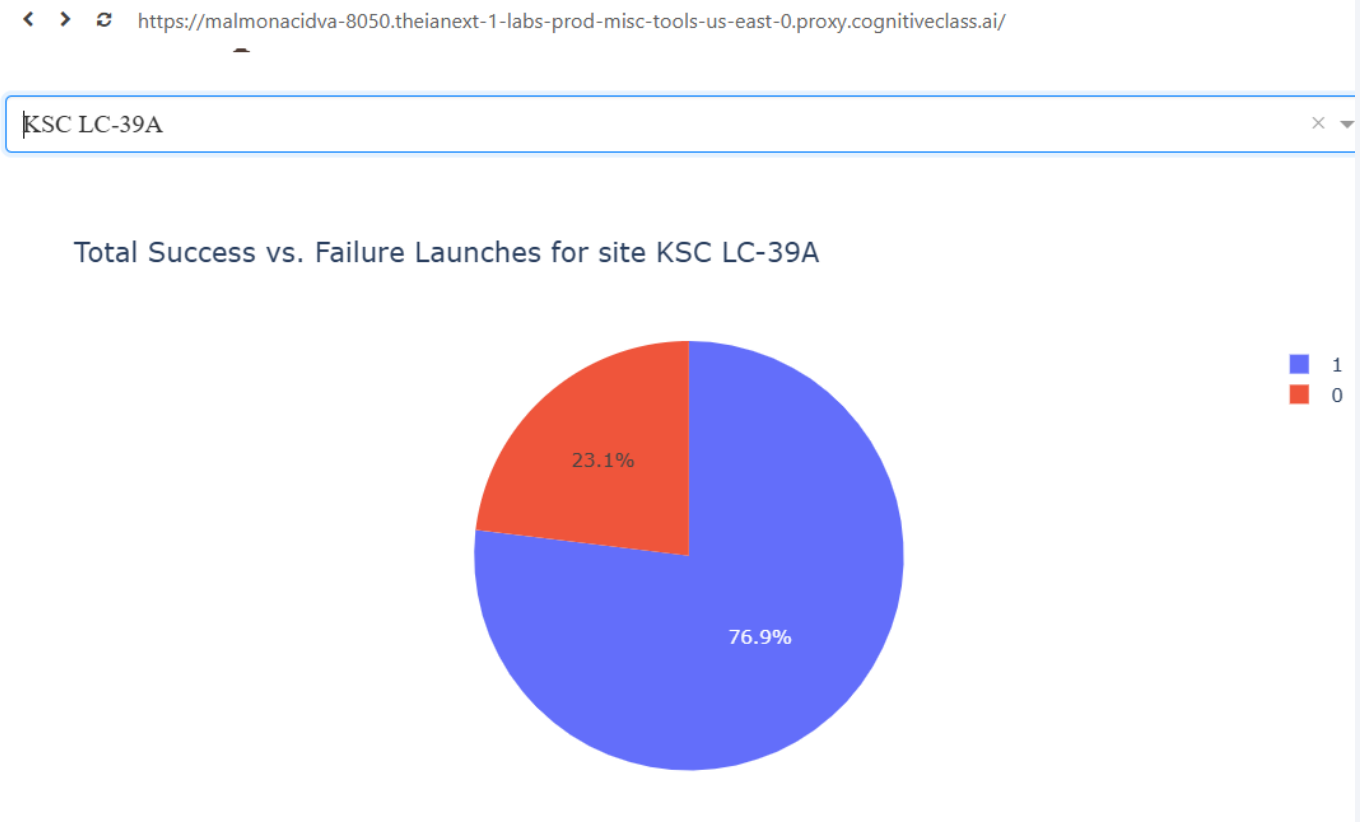


This pie chart from the Plotly Dash dashboard shows the distribution of successful launches across SpaceX's major launch sites:

- KSC LC-39A leads with 41.7% of total successful launches.
- CCAFS LC-40 follows with 29.2%.VAFB SLC-4E and other site entries represent a smaller share.
- The dropdown at the top allows filtering by specific launch site, dynamically updating the chart.

Insight: KSC LC-39A has emerged as the most reliable and frequently used launch site, reflecting its strategic importance for successful missions

Launch site with highest launch success

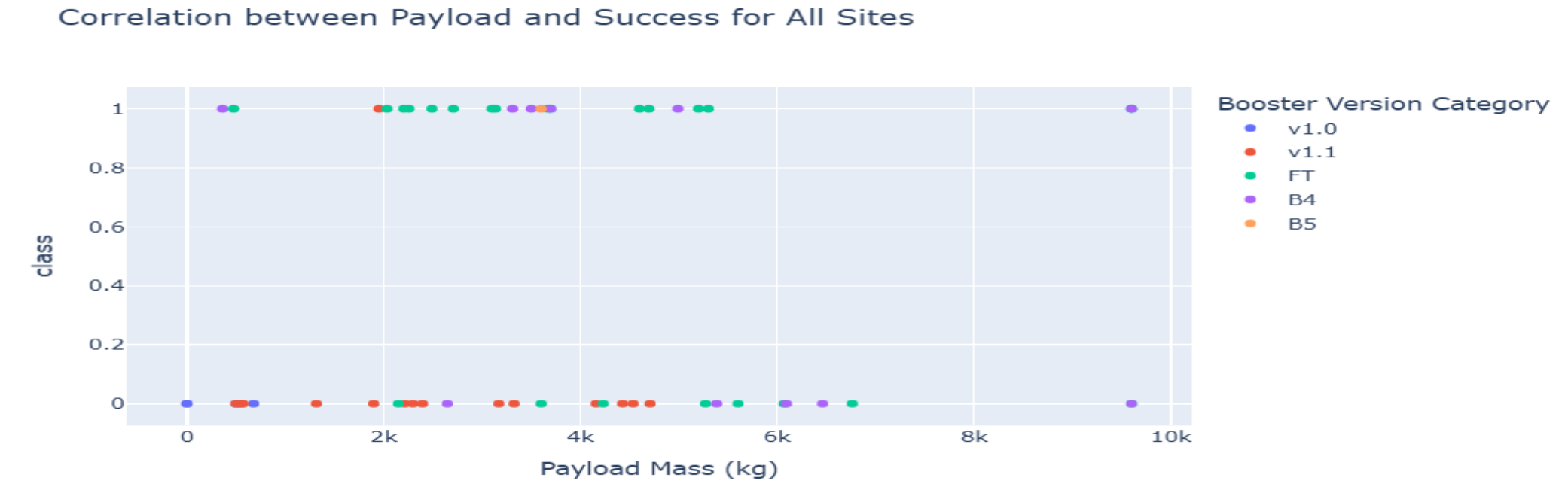


This pie chart shows the success rate for launches from Kennedy Space Center LC-39A:

- 76.9% of launches were successful (blue)
- 23.1% resulted in failure (red)

Insight:
KSC LC-39A maintains a high success rate, reinforcing its role as SpaceX's most reliable and strategic launch site.

Payload vs. Launch Outcome scatter plot for all sites



This scatter plot visualizes the correlation between payload mass and launch success across all sites and booster versions:

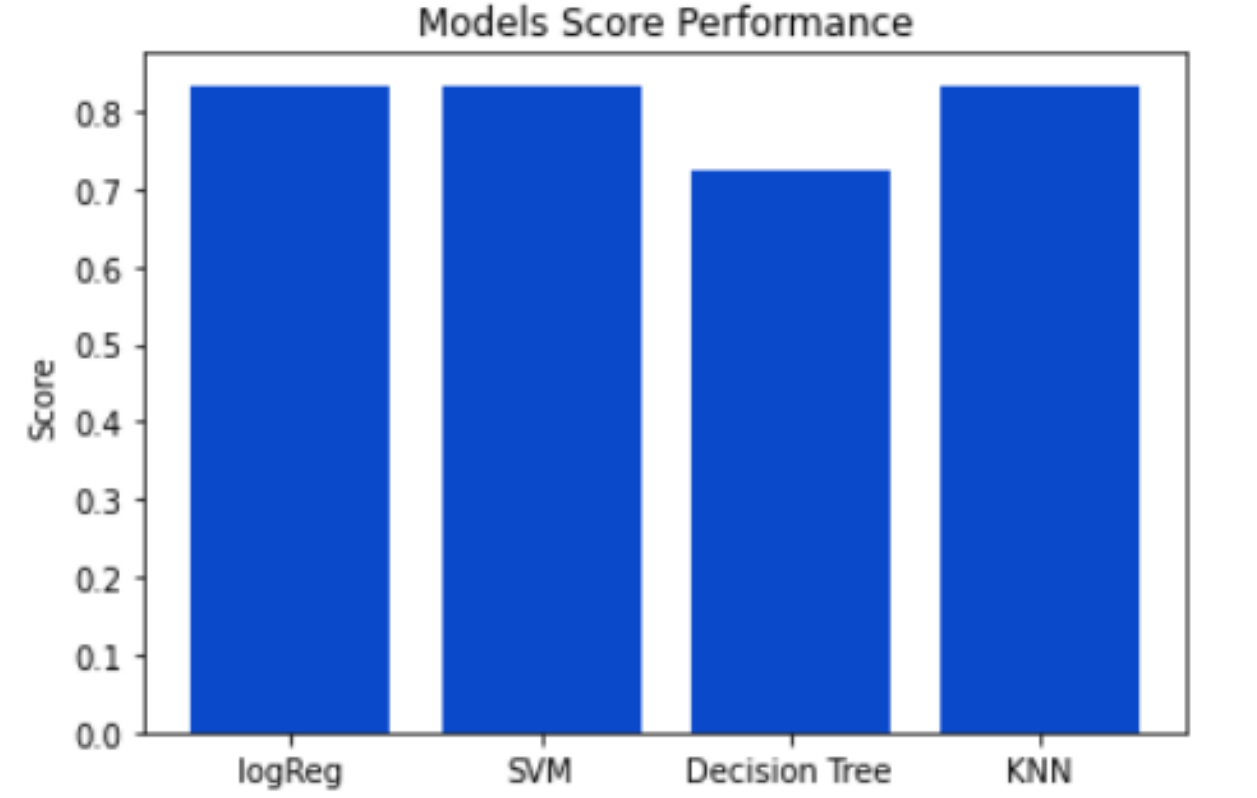
Y-axis shows success (1) or failure (0) X-axis represents payload mass in kg. Colors indicate different Booster Version Categories

Insight: There is no clear linear correlation between payload size and mission outcome, suggesting that booster version and operational factors may play a more critical role in success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



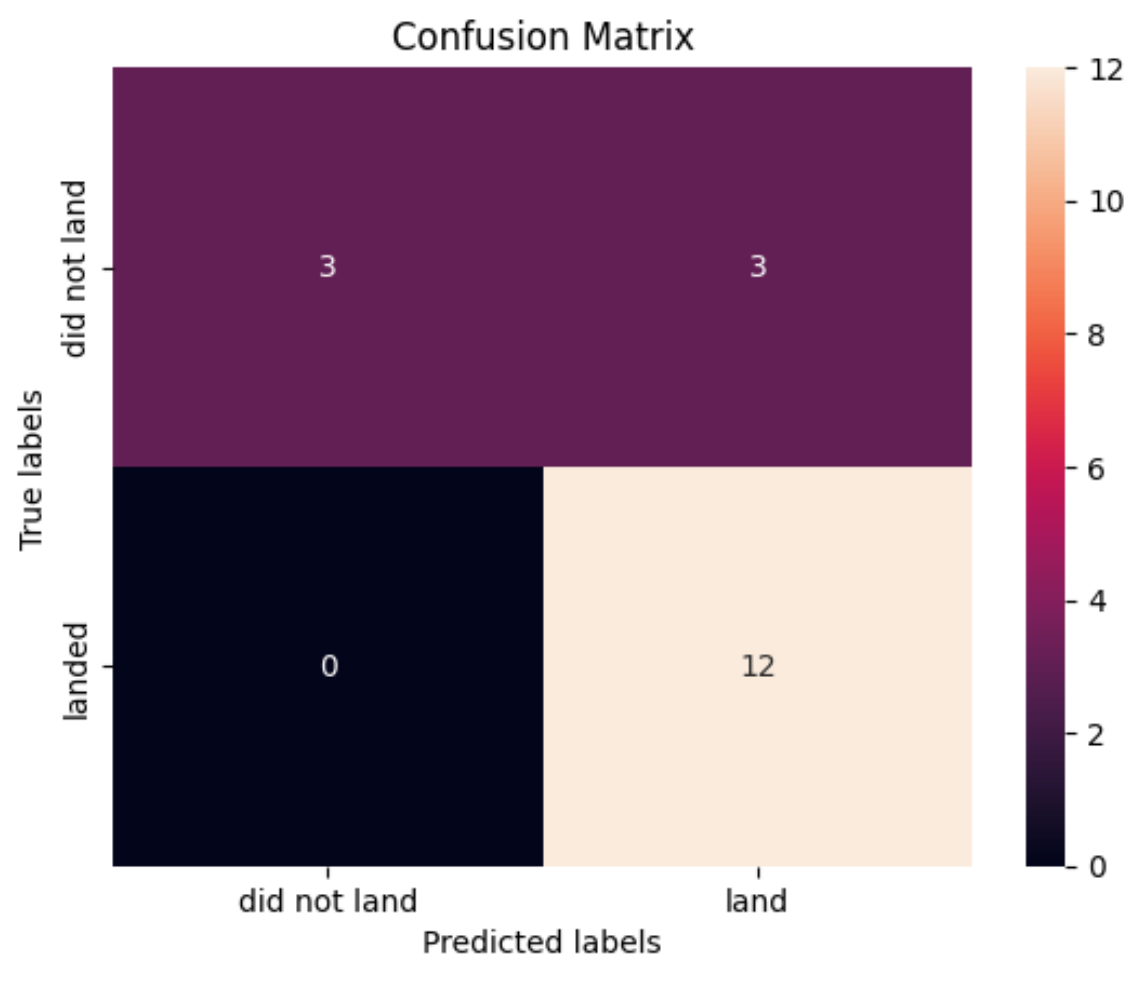
Logistic Regression -> Validation Accuracy: 0.8464, Test Accuracy: 0.8333

SVM -> Validation Accuracy: 0.8482, Test Accuracy: 0.8333

Decision Tree -> Validation Accuracy: 0.9036, Test Accuracy: 0.8333

KNN -> Validation Accuracy: 0.8482, Test Accuracy: 0.8333

Confusion Matrix




```
Decision Tree -> Validation Accuracy: 0.9036, Test Accuracy: 0.8333
```

This confusion matrix visualizes the prediction results of the best-performing model:
Decision Tree.

- True Positives (landed): 12
- False Positives (predicted land but didn’t): 3
- True Negatives: 3
- False Negatives: 0
- Validation Accuracy: 90.36%
- Test Accuracy: 83.33%

Insight: The model is effective at predicting successful landings, but still yields some false positives, meaning it occasionally overestimates success.

Conclusions

- The exploratory and interactive visualizations revealed key trends in SpaceX launch performance, including strong success rates at KSC LC-39A.
- Launch success is not strongly correlated with payload mass, but booster version and site selection play a relevant role.
- Interactive dashboards and maps allowed for a comprehensive spatial and categorical exploration of the dataset.
- A Decision Tree classifier achieved the highest validation accuracy (90.36%), indicating solid predictive performance for mission outcomes.
- This project showcases the power of combining API data collection, EDA, geospatial tools, dashboards, and ML into a unified data science workflow.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!