IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fernando
21 Feb 2025

Outline

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

Executive Summary

- **Summary of methodologies:**
  - Data Collection API
  - Data Collection Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with Folium
  - Machine Learning Prediction
- **Summary of all results:**
  - Exploratory Data Analysis with SQL
  - Interactive Analytics Screenshots
  - Predictive Analytics Result

Introduction

- **Project background and context:**

SpaceX offers Falcon 9 rocket launches at a cost of $62 million, significantly lower than other providers, whose costs exceed $165 million per launch. A major factor in these cost savings is SpaceX's ability to reuse the first stage of the rocket.

By accurately predicting whether the first stage will land successfully, we can estimate the overall cost-efficiency of a launch. This insight is valuable for other companies aiming to compete with SpaceX in the commercial space launch industry.

The objective of this project is to develop a machine learning pipeline that can predict the likelihood of a successful first-stage landing for Falcon 9.

- **Problems you want to find answers:**
  - What factors influence the successful landing of a rocket's first stage?

    Identifying key variables that contribute to a safe and stable landing.
  - How do different conditions and features interact to affect landing success rates?

    Understanding the relationships between various technical and environmental parameters.
  - What operational conditions are necessary to maximize landing success?

    Determining optimal launch and landing conditions for future missions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Fetch data using SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Extract relevant features and resolve missing values
- 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, evaluate classification models

Data Collection

Data Source:

1. SpaceX API:

Data was retrieved through HTTP requests to various SpaceX API endpoints, including:

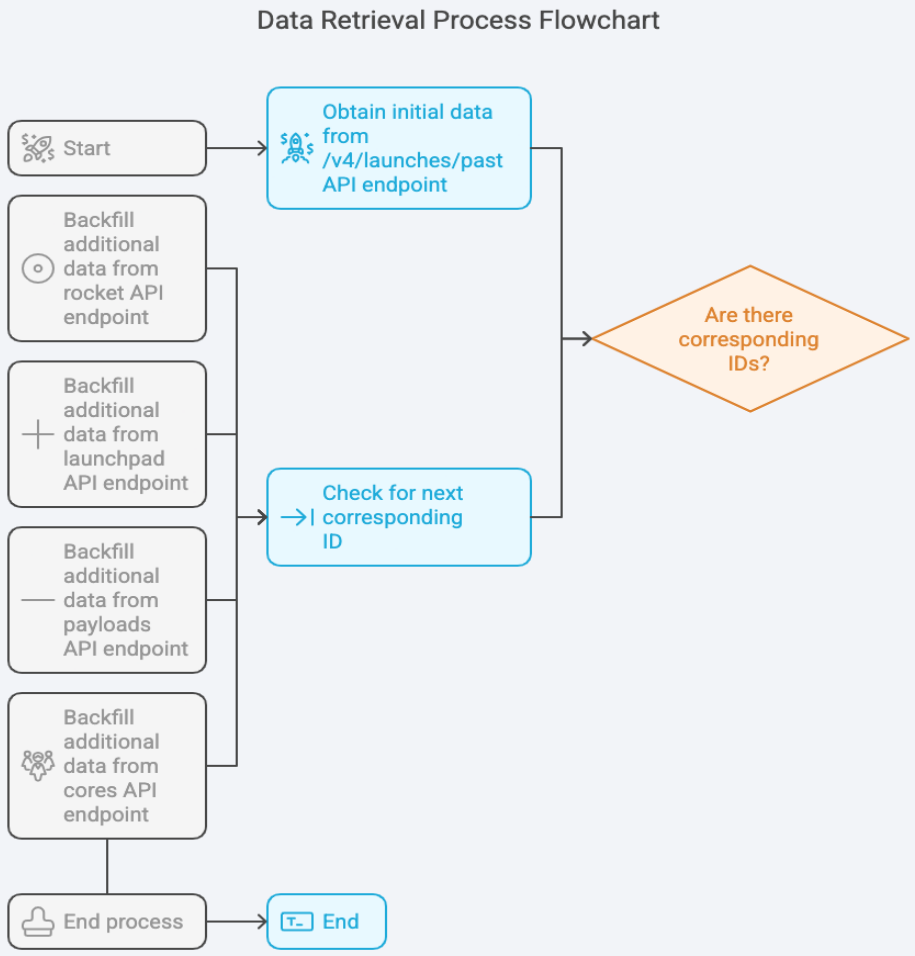
- /v4/launches/past – Initial launch data.
- /v4/rockets, /v4/launchpads, /v4/payloads, /v4/cores – Additional details such as rocket specifications, launchpad locations, payload information, and core booster details.

2. Falcon 9 Wikipedia Page:

- Additional historical launch records were extracted through web scraping using BeautifulSoup.
- The launch data was retrieved from HTML tables, parsed, and transformed into a structured pandas DataFrame for further analysis.

Data Collection – SpaceX API

- **Process:**
  - The primary dataset was retrieved from the /v4/launches/past API endpoint.
  - Supplementary details were sourced from the rockets, launchpads, payloads, and cores endpoints to enrich records with corresponding IDs.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone

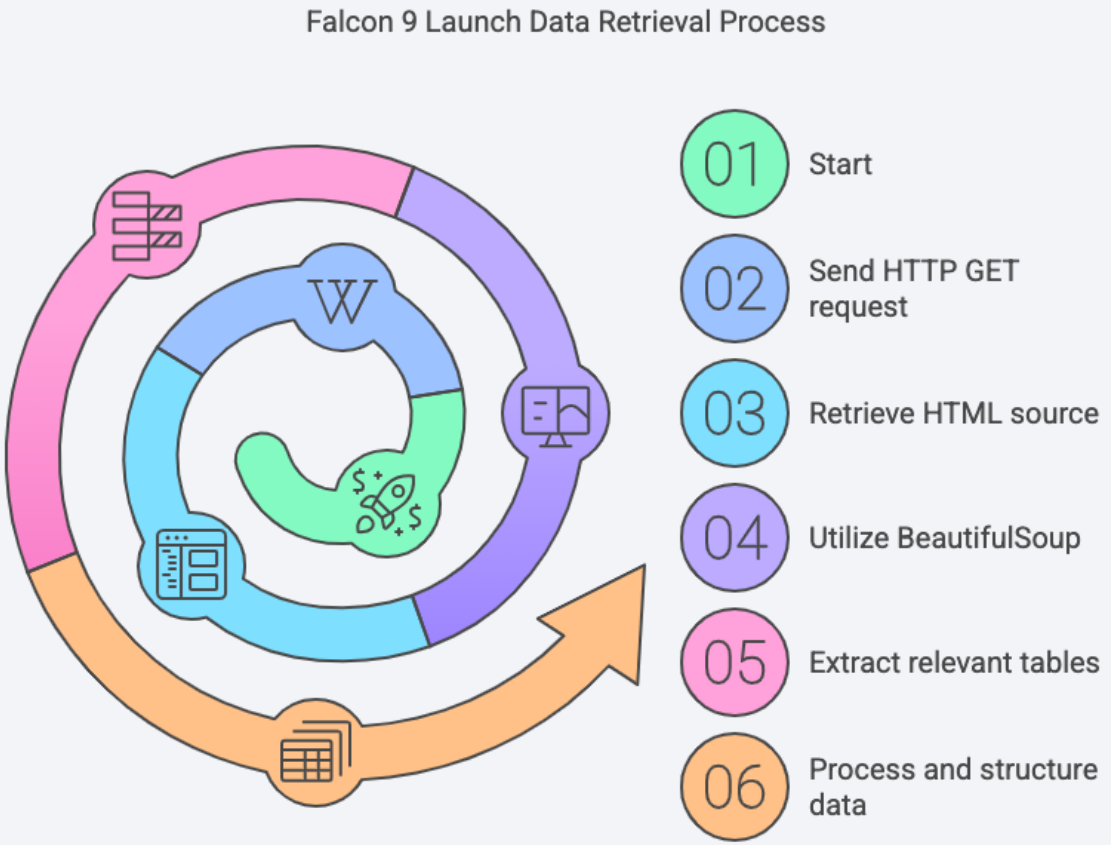


Data Collection - Scraping

Process:

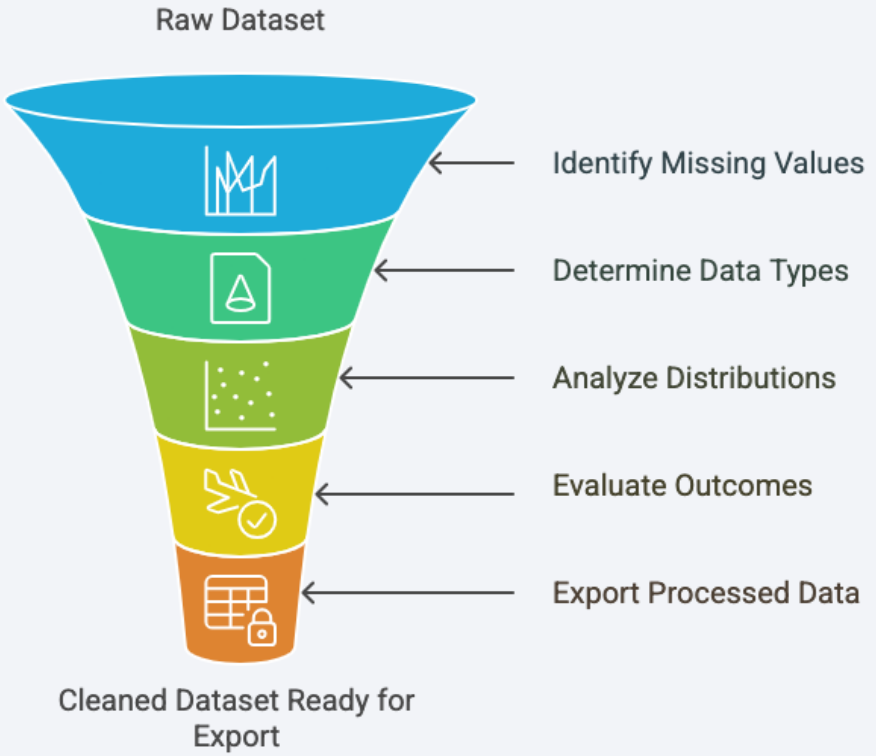
- Sent an HTTP GET request to the Falcon 9 launch page on Wikipedia to retrieve the HTML source.
- Utilized **BeautifulSoup** to parse the HTML and extract relevant tables containing launch data.
- Processed and structured the extracted data into a **Pandas DataFrame** for further analysis.

- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone



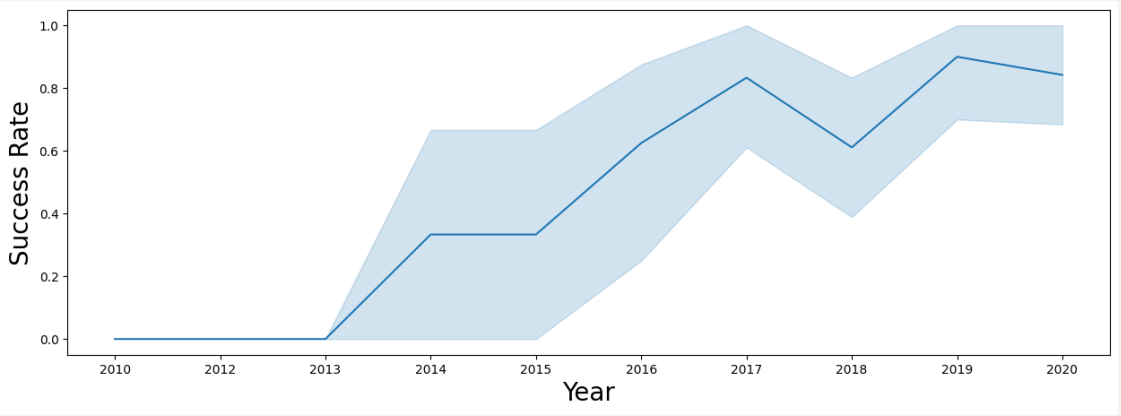
Data Wrangling

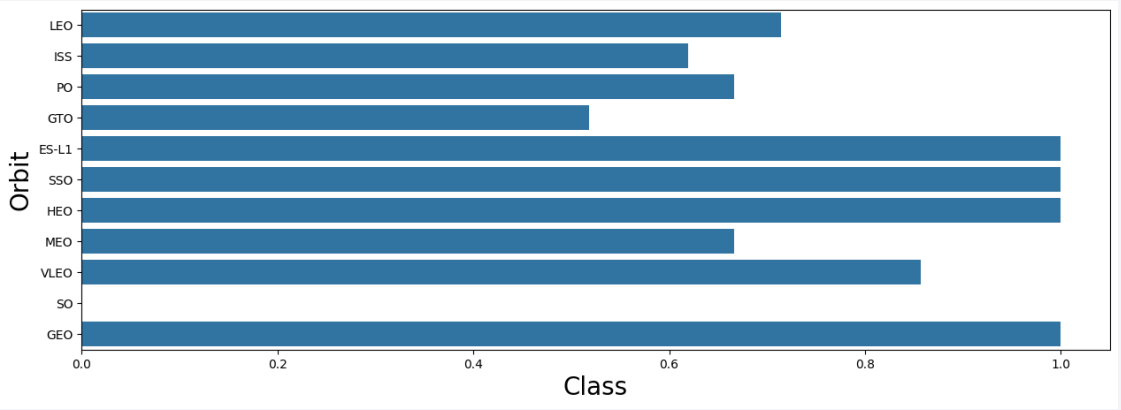
- **Process:**
  - **Performed Exploratory Data Analysis (EDA)** to uncover patterns and insights within the dataset.
  - **Identified and handled missing values** by calculating their percentage for each attribute.
  - **Determined data types** by categorizing columns as numerical or categorical.
  - **Analyzed launch distributions**, including the number of launches per site and the frequency of different orbit types.
  - **Examined landing outcomes**, grouping them into binary categories (success or failure).
  - **Created a "Class" label** based on landing outcomes to serve as the target variable for training machine learning models.
  - **Exported the cleaned and processed data** to a CSV file for further analysis and modeling.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone



EDA with Data Visualization

- **Exploratory Data Analysis (EDA) Process:**
  - **Visualized key relationships** to understand the impact of various factors on launch success.
  - **Examined correlations** between flight number and launch site, as well as payload and launch site.
  - **Analyzed success rates** across different orbit types.
  - **Explored trends** between flight number and orbit type to identify patterns.
  - **Tracked yearly success rates** to observe improvements over time.
  - **Converted numerical columns to float64** for consistency in data processing.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone





EDA with SQL

- **Loaded SpaceX Dataset into SQLite Database**
- **Queries Performed:**
  - Identify unique launch sites.
  - Retrieve 5 records where the launch site starts with 'CCA'.
  - Calculate total payload mass for boosters launched by 'NASA (CRS)'.
  - Find the average payload mass of Falcon 9 v1.1 boosters.
  - Get the date of the first successful ground landing.
  - List booster versions with successful drone ship landings for payloads between 4000-6000kg.
  - Count successful vs. failed mission outcomes.
  - Identify booster versions that carried the maximum payload.
  - Extract mission details (month, outcome, booster, launch site) for failed drone ship landings in 2015.
  - Show outcome distribution between June 4, 2010, and March 20, 2017.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone

Build an Interactive Map with Folium

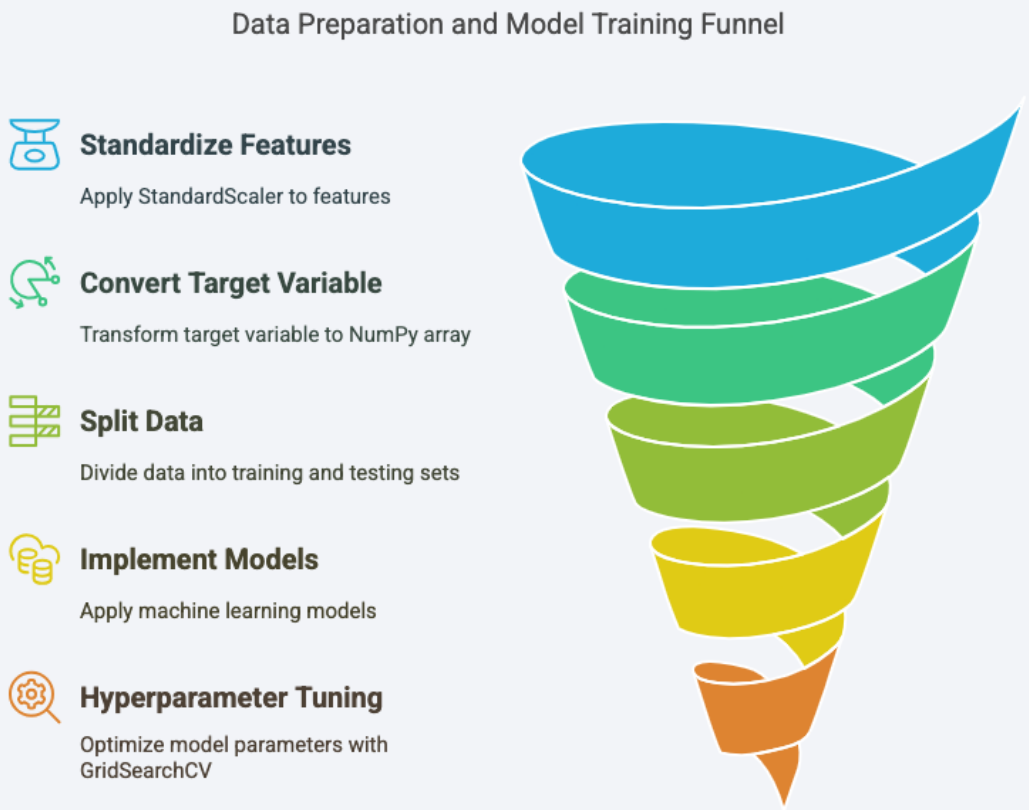
- **Various map objects to analyze geographical patterns in the launch data:**
  - **Markers**: Plotted all launch sites to visualize their locations.
  - **Color-Coded Markers**: Indicated successful (green) and failed (red) launches to assess site performance.
  - **Circles**: Highlighted launch sites and their surrounding areas.
  - **Lines**: Measured distances between launch sites and key landmarks like railways, highways, and coastlines.
- **Reason for Adding These Objects:**
  - **To identify launch success trends** by visually distinguishing successful vs. failed launches.
  - **To explore geographic influences** on launch sites, such as proximity to infrastructure and safety considerations.
  - **To enhance data-driven decision-making** by analyzing how location impacts launch outcomes.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone

Build a Dashboard with Plotly Dash

- **Interactive Plotly Dash dashboard to enable data exploration, featuring:**
  - **Pie Chart**:
    - When all sites are selected: Displays the distribution of successful launches across all sites.
    - When a specific site is selected: Shows the success vs. failure ratio for that site.
  - **Scatter Plot**:
    - When all sites are selected: Visualizes the relationship between payload mass, booster version, and launch outcomes across all sites.
    - When a site is selected: Filters the scatter plot to show data for only that site.
- **Payload Mass Range Selector**: Allows users to filter the scatter plot based on payload mass.
- **Reason for Adding These Plots and Interactions:**
  - **Pie Chart** helps compare overall launch success rates across different sites.
  - **Scatter Plot** provides insights into how payload mass and booster versions impact launch success.
  - **Interactive Filters** allow for dynamic data exploration, making it easier to identify trends and correlations.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone

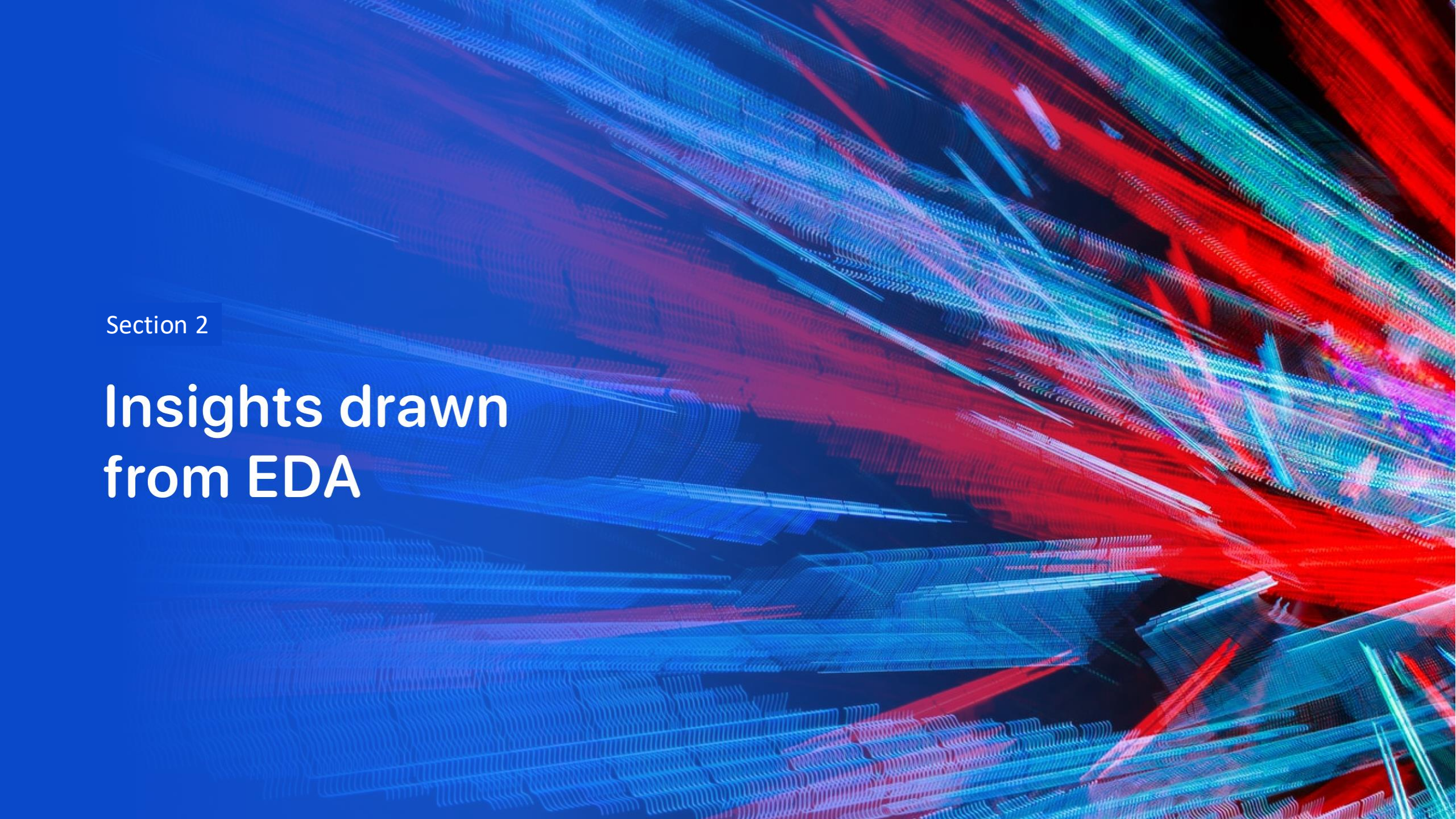
Predictive Analysis (Classification)

- **Data Preparation**:
  - Loaded the dataset and applied **StandardScaler** to standardize features.
  - Converted target variable **Y** into a NumPy array.
  - Split the data into **training and testing sets** for model evaluation.
- **Model Training & Evaluation**:
  - Implemented multiple machine learning models:
    - Logistic Regression
    - Support Vector Classifier (SVC)
    - Decision Tree Classifier
    - K-Nearest Neighbors (KNN)
  - Used **GridSearchCV** to test different hyperparameter combinations and identify the best-performing model.
- **Notebook:**
  - https://github.com/Fernando-Lim/data-science-capstone

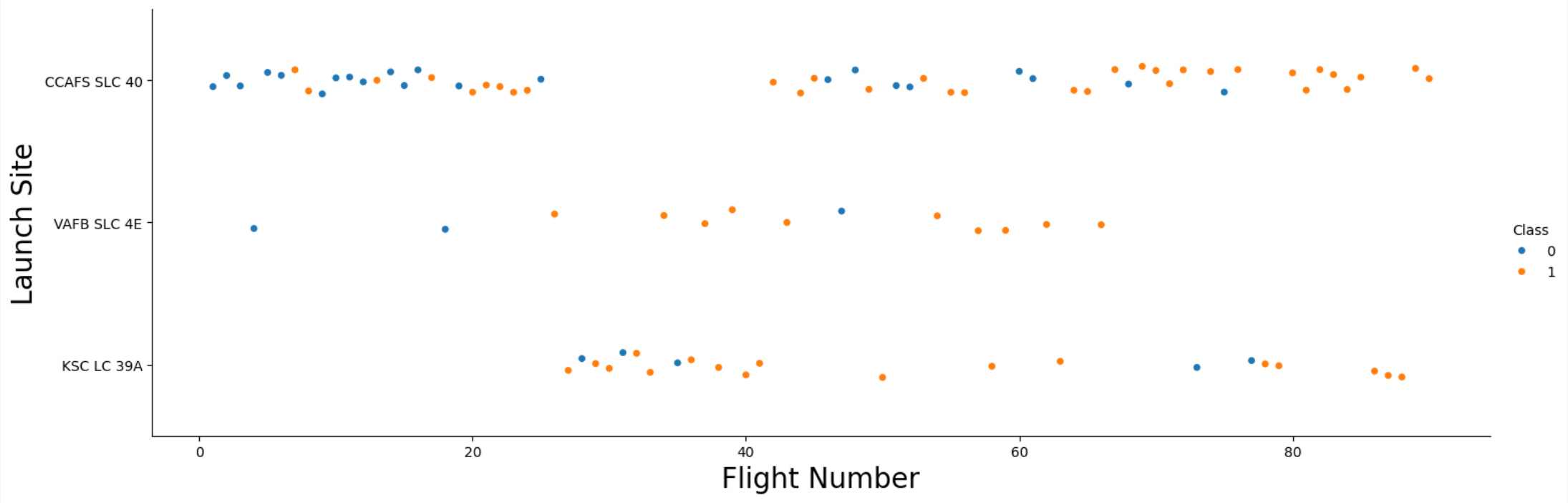


Results

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


Section 2
Insights drawn from EDA

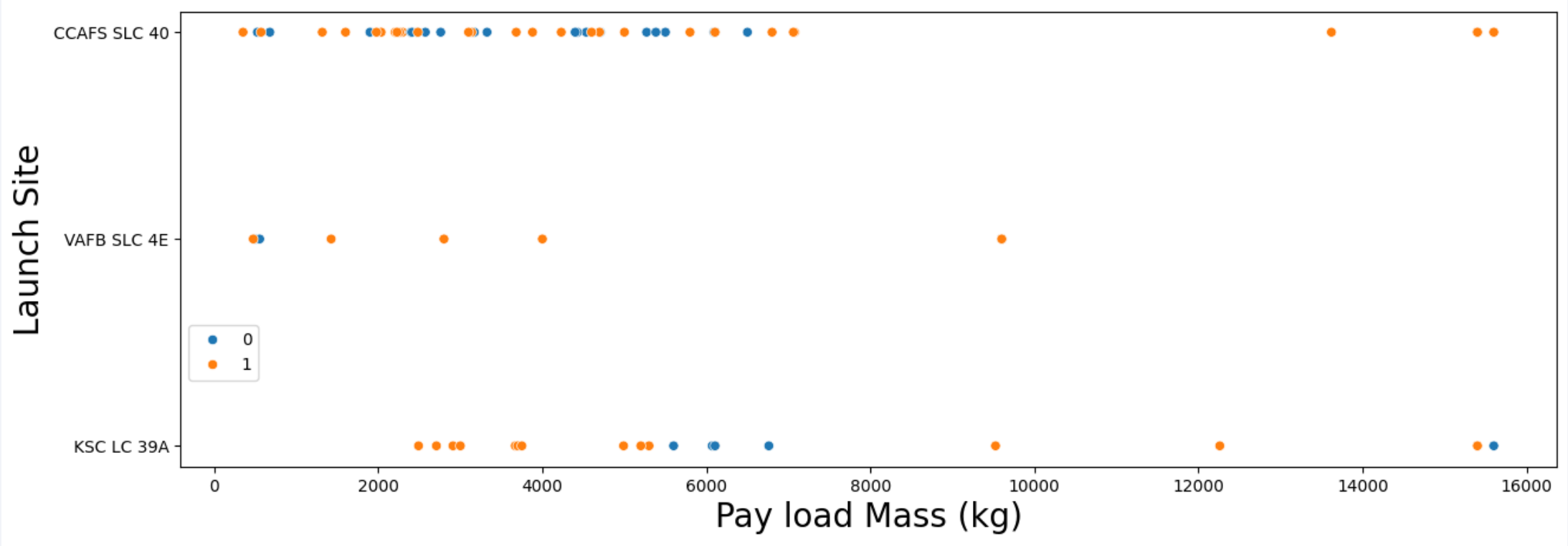
Flight Number vs. Launch Site



Explanation:

Over time, all launch sites have experienced both successful and failed first-stage landings, with a noticeable improvement in success rates. In the early stages, most launches resulted in failure, indicating that technological advancements and process improvements played a significant role in increasing reliability. Among the sites, CCAFS SLC 40 has conducted the highest number of launches overall. However, when evaluating relative success rates, VAFB SLC 4E appears to have a comparatively higher proportion of successful landings.

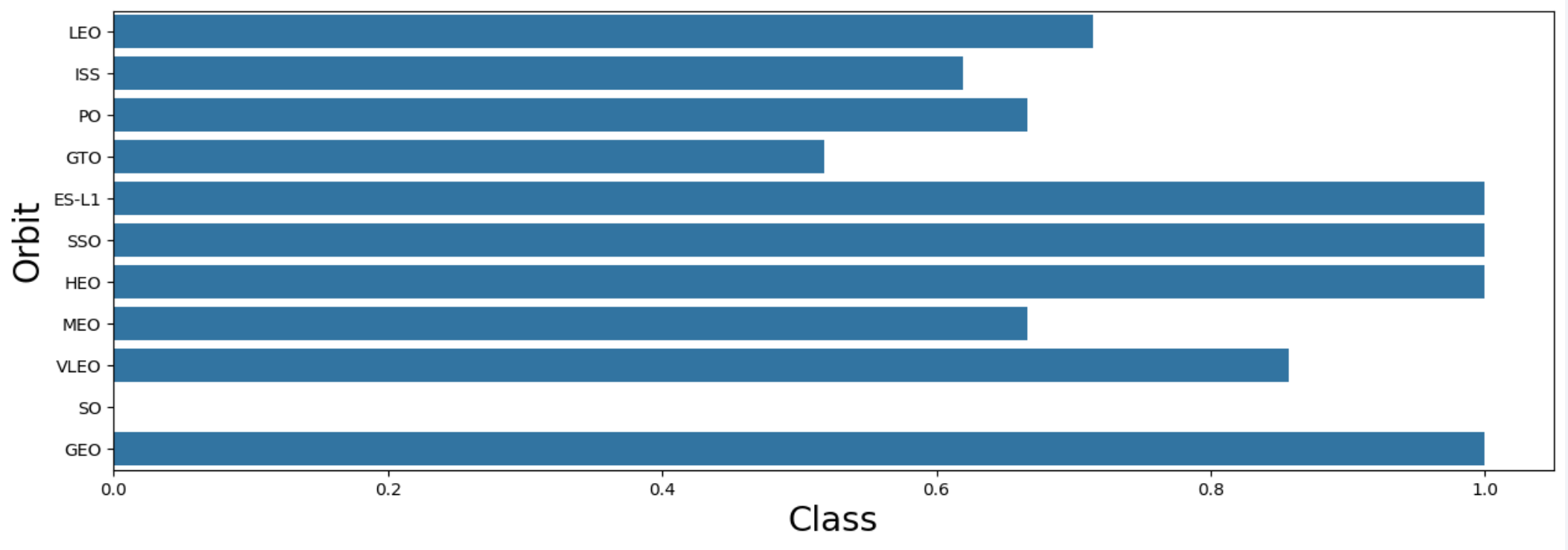
Payload vs. Launch Site



Explanation:

All launch sites have handled a wide range of payload weights, from light to heavy payloads. In the early stages, most flights carried lighter payloads, which also corresponded with a higher failure rate in landings. This trend suggests that advancements in technology and operational strategies have contributed to an improved success rate, even for missions carrying heavier payloads.

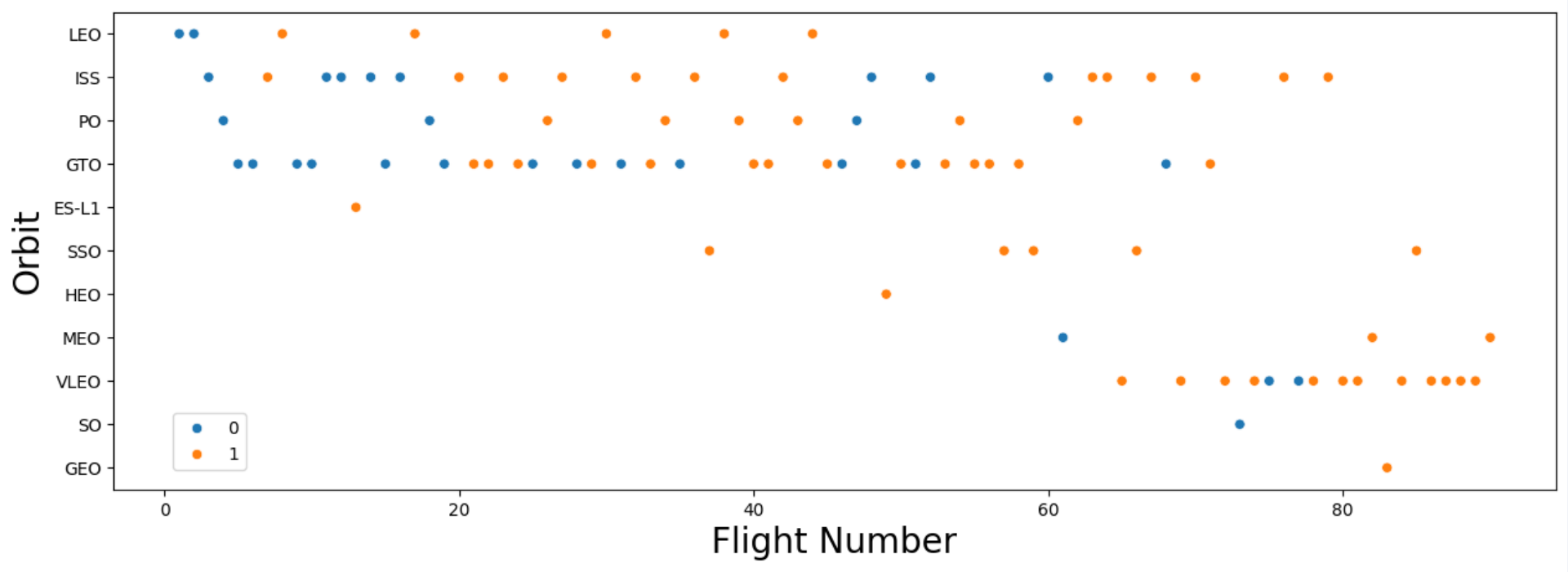
Success Rate vs. Orbit Type



Explanation:

Certain orbit types, such as ES-L1, SSO, HEO, and GEO, have consistently demonstrated high success rates in first-stage landings. In contrast, orbits like GTO show more varied outcomes, indicating that some orbit types may present additional operational or technological challenges. Additionally, the SO orbit type has only one recorded launch, making it difficult to draw meaningful conclusions from the limited data available.

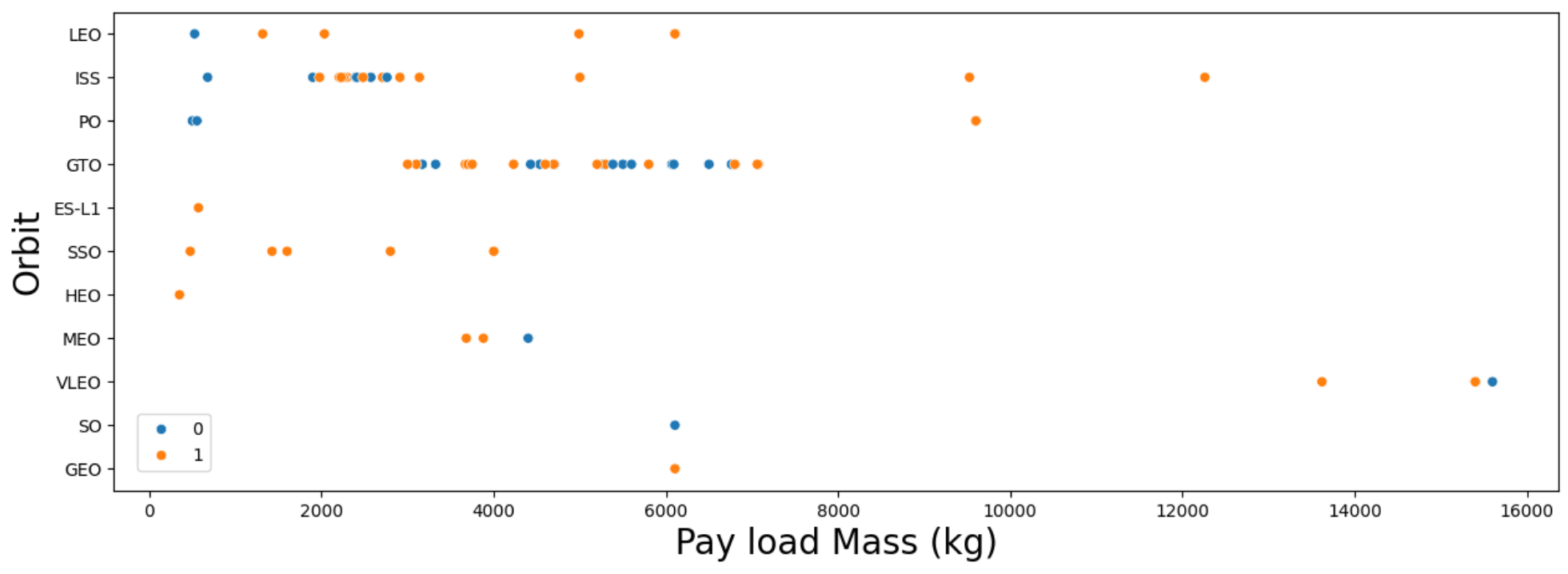
Flight Number vs. Orbit Type



Explanation:

Various orbit types are represented across different flight numbers, though some orbits were only attempted in later missions. As flight numbers increase, there is a clear trend of improved landing success, suggesting that experience and continuous advancements in technology have contributed to better outcomes over time.

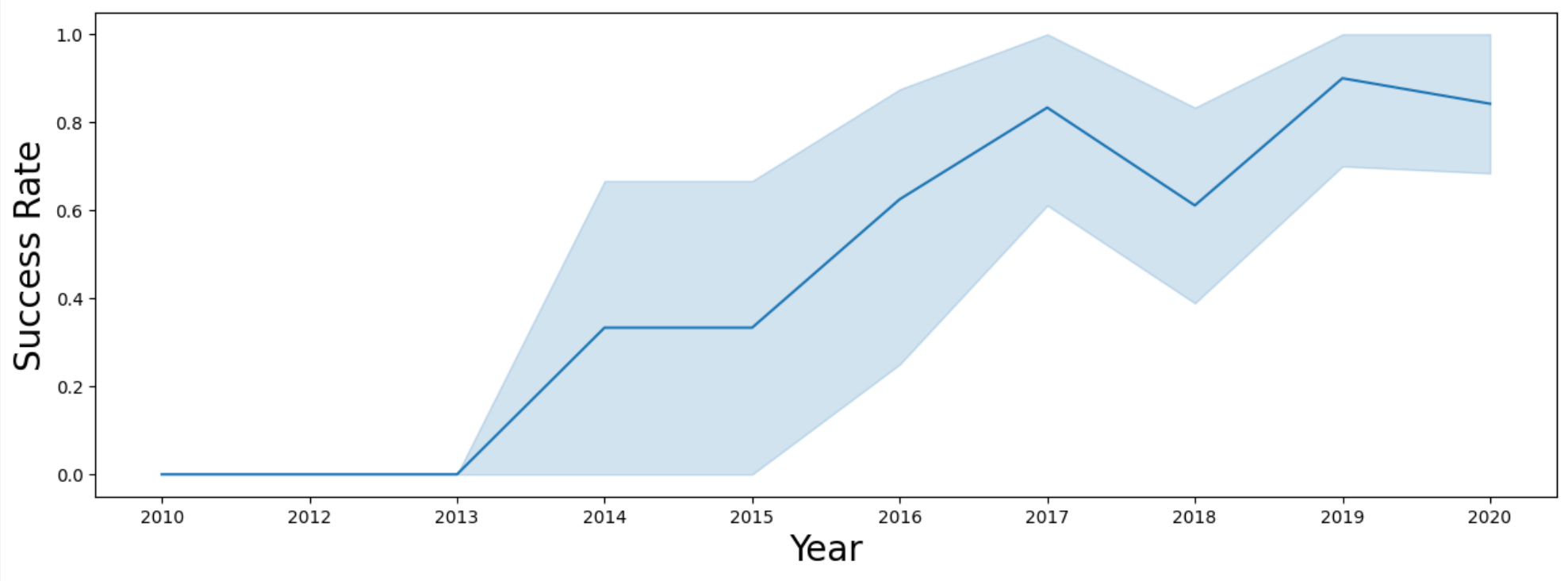
Payload vs. Orbit Type



Explanation:

Different orbits accommodate a wide range of payload masses, though some, such as SSO, MEO, HEO, and GEO, generally fall within a lower range. Orbits with more constrained payload ranges tend to have higher landing success rates. While payload mass alone does not directly determine mission success, its relationship with orbit type suggests a meaningful correlation.

Launch Success Yearly Trend



Explanation:

The yearly trend highlights a steady improvement in first-stage landing reliability, evolving from early challenges to higher success rates over time. Since 2016, SpaceX has achieved continuous year-over-year progress, with a slight setback in 2018 before continuing its upward trajectory.

All Launch Site Names

```
%%sql

SELECT DISTINCT Launch_Site from SPACEXTABLE;
```

✓ 0.0s

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

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

Explanation:

Used Distinct key word to show only unique launch site name from the SpaceX Data

Launch Site Names Begin with 'CCA'

```
%%sql

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

✓ 0.0s Python

```
 * 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

Explanation:

Used query above to display five record where launch sites begin with ‘CCA’

Total Payload Mass

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

✓ 0.0s

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

TOTAL_PAYLOAD
45596

Explanation:

Used query above to calculate the total payload carried by boosters from NASA (CRS) is 45,596 kg

Average Payload Mass by F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
 ✓  0.0s

 * sqlite:///my_data1.db
Done.

AVG_PAYLOAD_MASS
2534.6666666666665
```

Explanation:

Used query above to calculated the average payload mass caried by booster version F9 v1.1 is 2,534.6 kg

First Successful Ground Landing Date

```
%%sql

SELECT MIN(Date) as LaunchDate FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

✓ 0.0s

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

LaunchDate
2015-12-22

Explanation:

Used query above to observe that the first successful landing outcome on ground pad occurred on 22 December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql

SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

✓ 0.0s Python

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

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

Explanation:

Used query above to filter which boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Explanation:

Used wildcard and where clause on query to observe total number of successful and failure mission outcomes

```
%%sql

SELECT CASE
    WHEN Mission_Outcome LIKE 'Success%' THEN 'Success'
    WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
END as Mission_Status,
COUNT(*)
FROM SPACEXTABLE
GROUP BY Mission_Status;
```

✓ 0.0s

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

Mission_Status	COUNT(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

Explanation:

Determined the boosters that carried the maximum payload using subquery in where clause and max function

```
%%sql

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

✓ 0.0s

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

Explanation:

Used combination of WHERE, AND, CASE, and Wild Card to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
%%sql

SELECT
    CASE strftime('%m', Date)
        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,
    Landing_Outcome, Booster_Version, Launch_Site, Date
FROM SPACEXTABLE
WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

✓ 0.0s

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

Month	Landing_Outcome	Booster_Version	Launch_Site	Date
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

Explanation:

Used WHERE to filter between 2010-06-04 to 2010-03-20, GROUP BY outcome to group landing outcome and ORDER BY clause to order the group landing outcome in DESC order

```
%%sql

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

✓ 0.0s

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

Landing_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

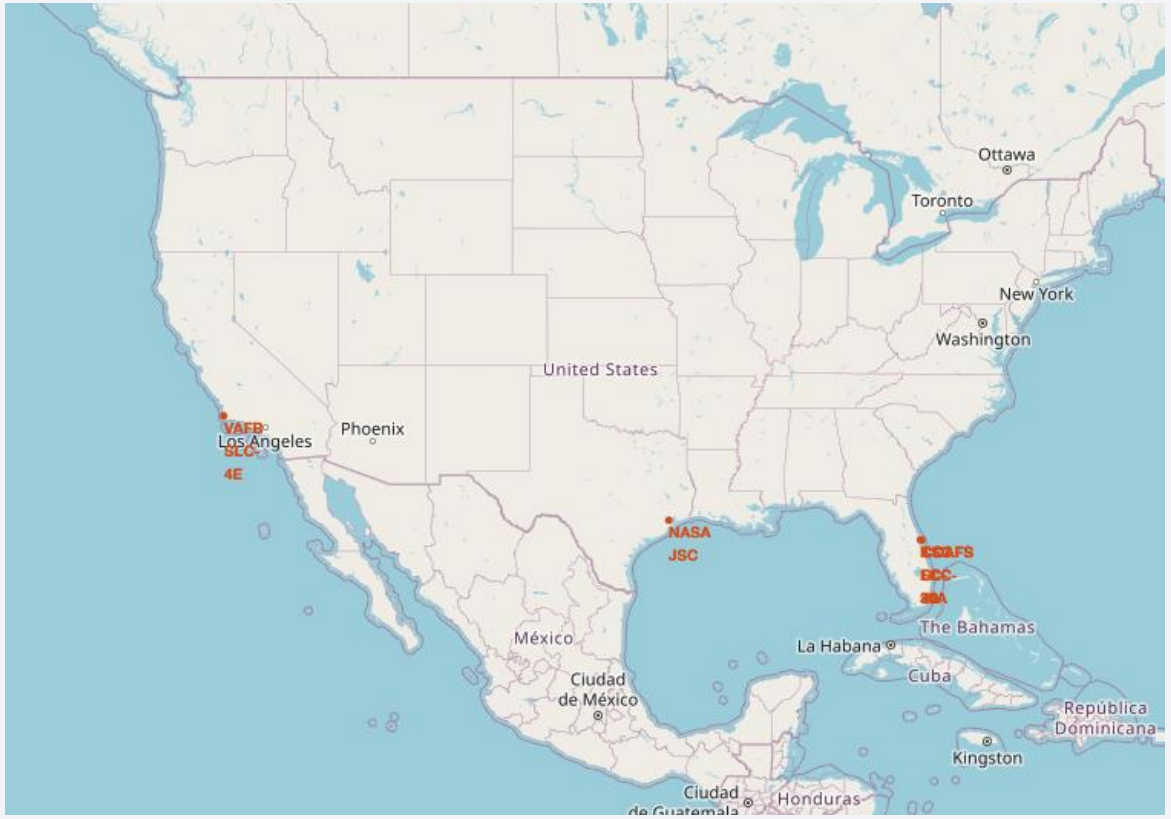
Section 3

Launch Sites Proximities Analysis

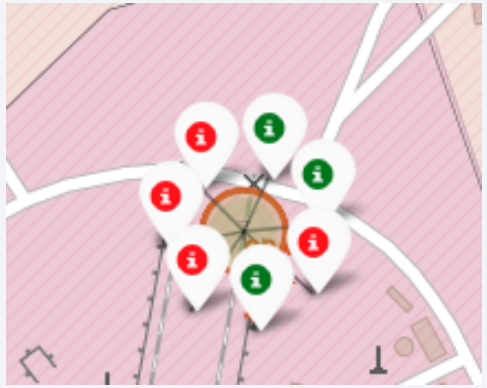
Launch Site Locations

Explanation:

SpaceX launch sites located in the United States of America coasts that near coastal regions in Floria and California to reduce catastrophic failures affecting human activities

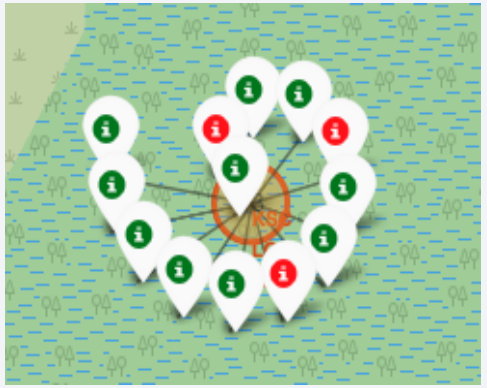


Launch Sites Outcomes



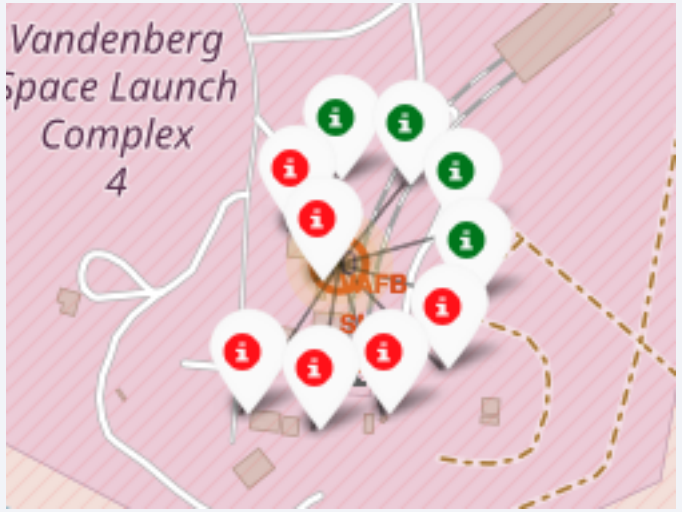
CCAFS SLC-40

CCAFS LC-40



KSC LC-39A

Florida



VAFB SLC-4E

California

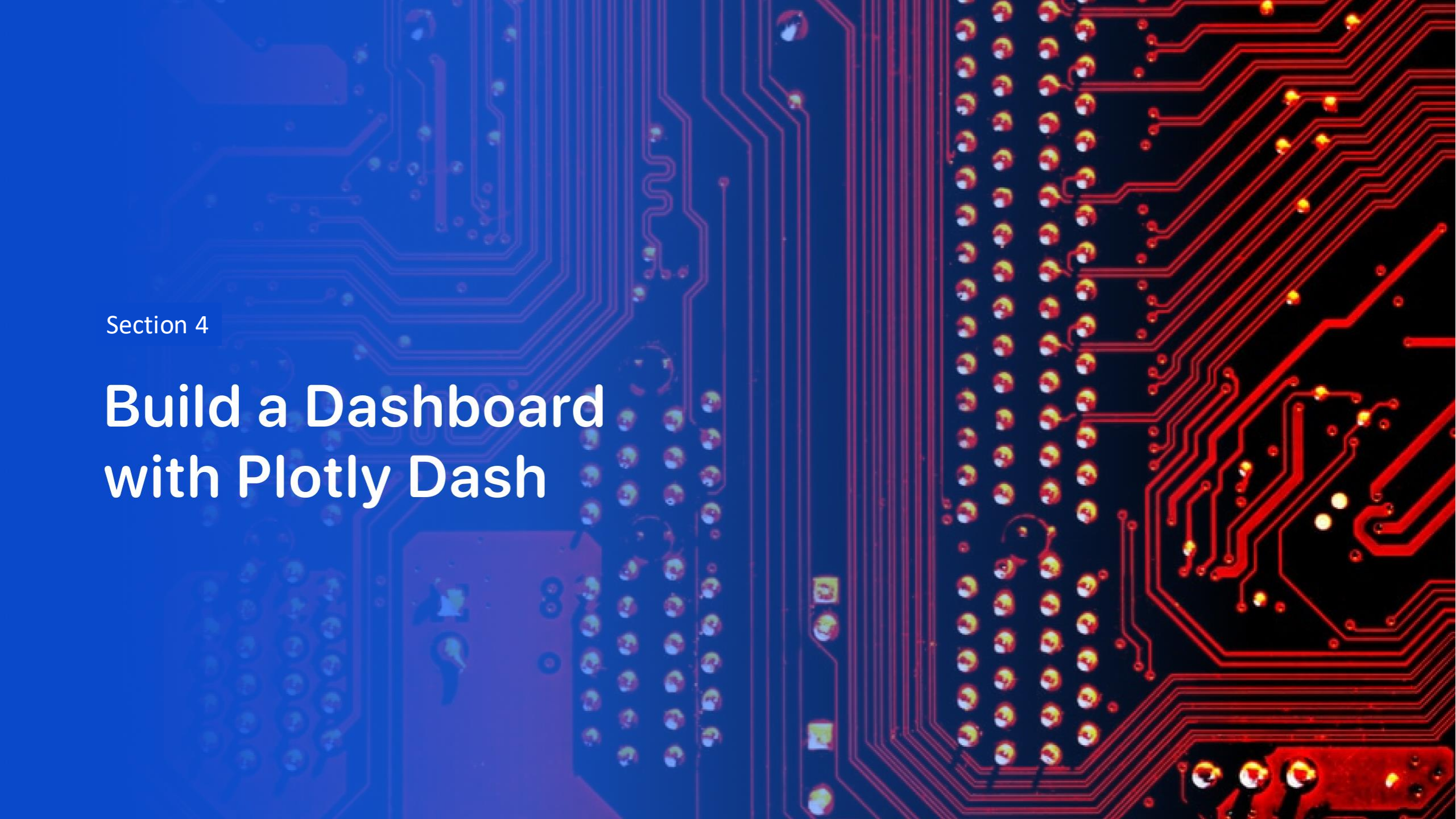
Legend:

- Successful Launches
- Failures Launches

Proximate Launch Sites to Notable Locations

Distance:

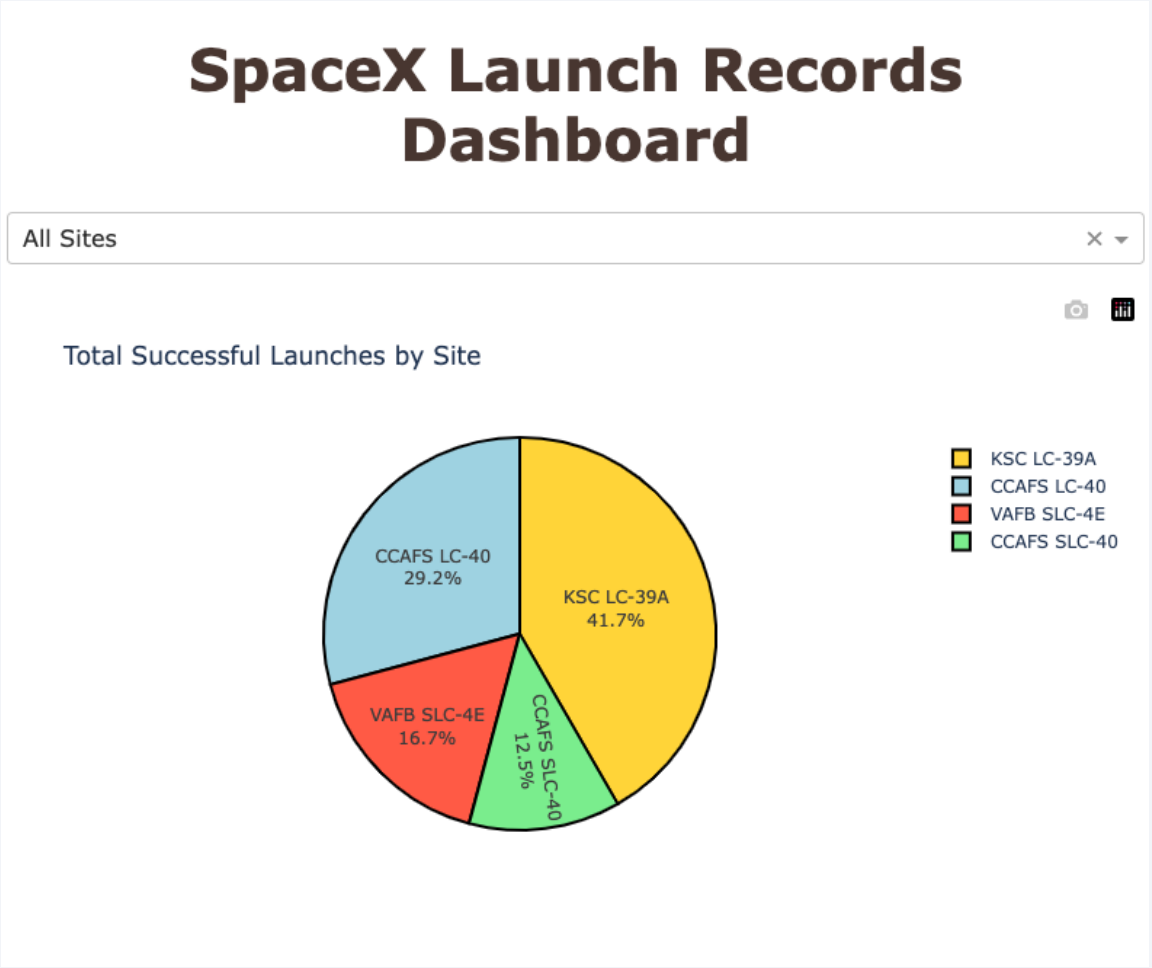
1. Railway = 1.02 KM
2. Roadway = 0.59 KM
3. Coast = 0.86 KM


Section 4
Build a Dashboard with Plotly Dash

Success Percentage by each launch site

Explanation:

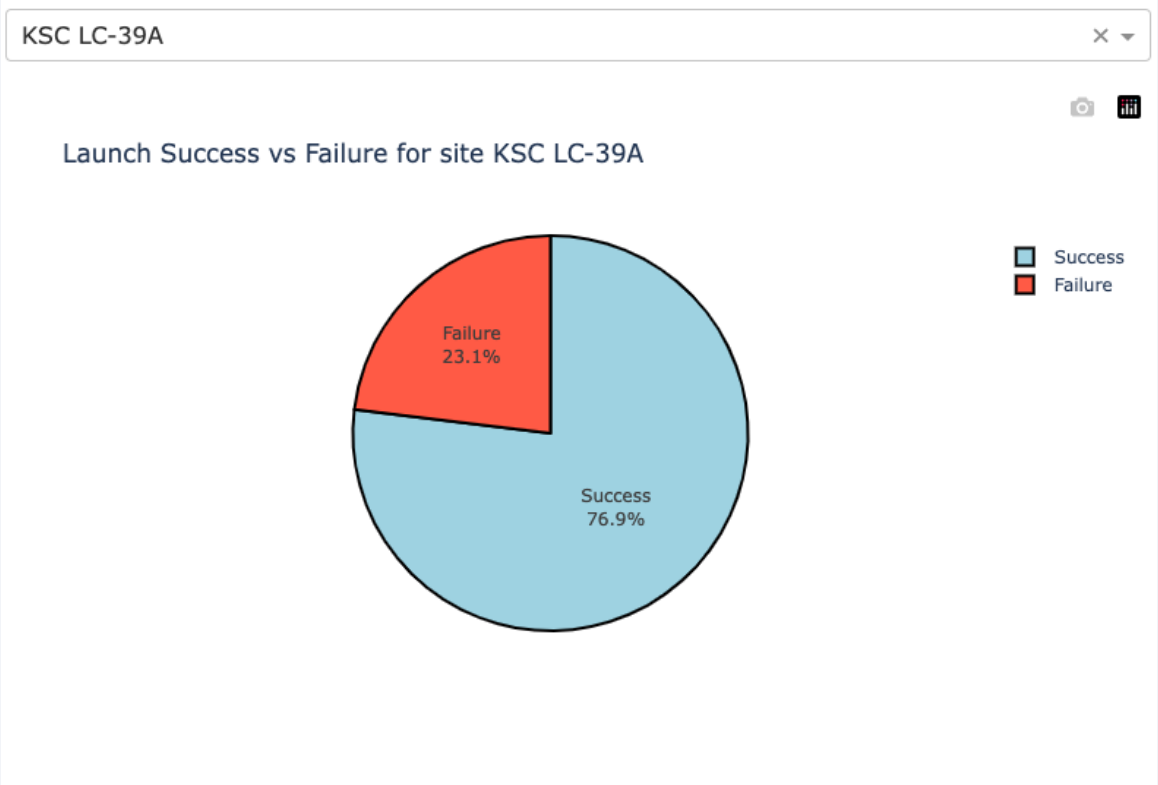
- KSC LC-39A had the most successful launches, followed by CCAFS LC-40.
- VAFB SLC-4E and CCAFS SLC-40 is the lowest launch success.



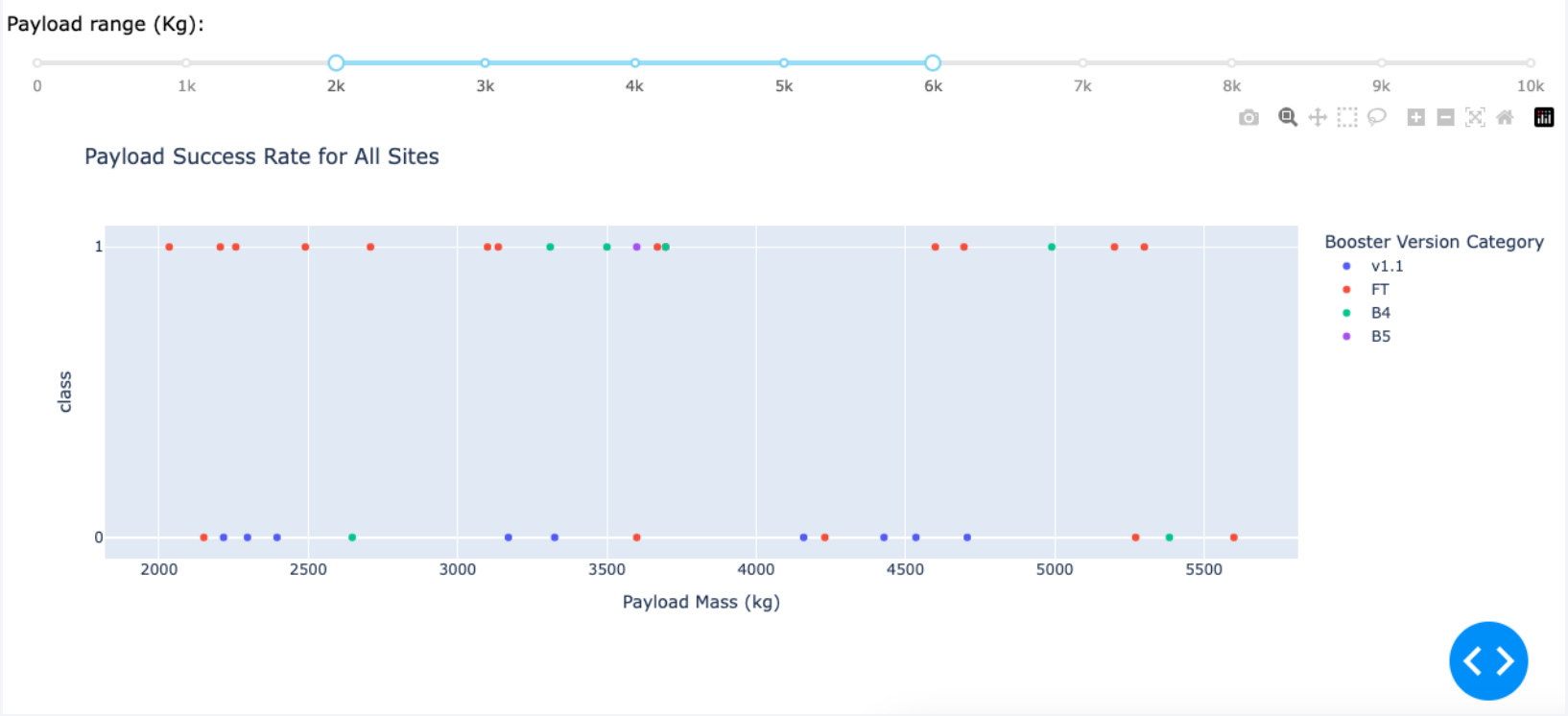
Highest Launch Success Ratio

Explanation:

KSC LC-39A achieved a 76.9% success rate which is the highest ratio of successful landings compare to other site



Payload Mass Range



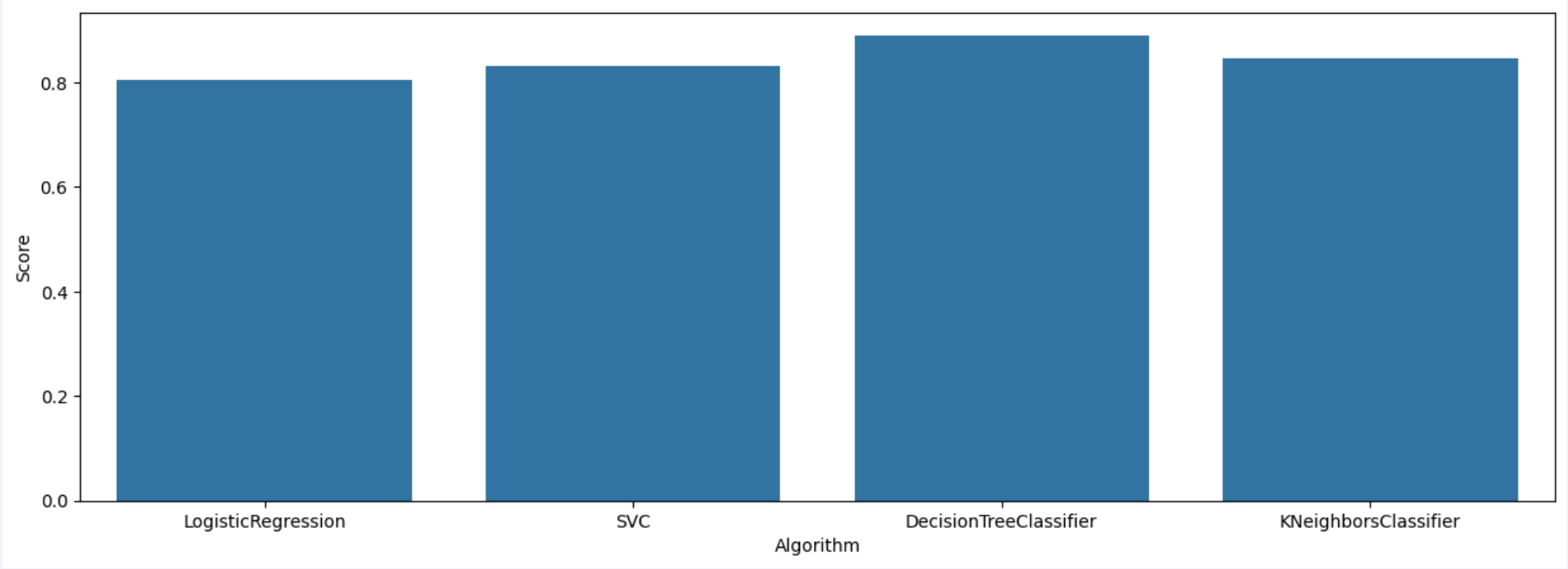
Explanation:

Payload range between 2,000 and 6,000 kg shows that v1.1 boosters performed the worst but we can see that FT boosters had the best success rate, followed by B4.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



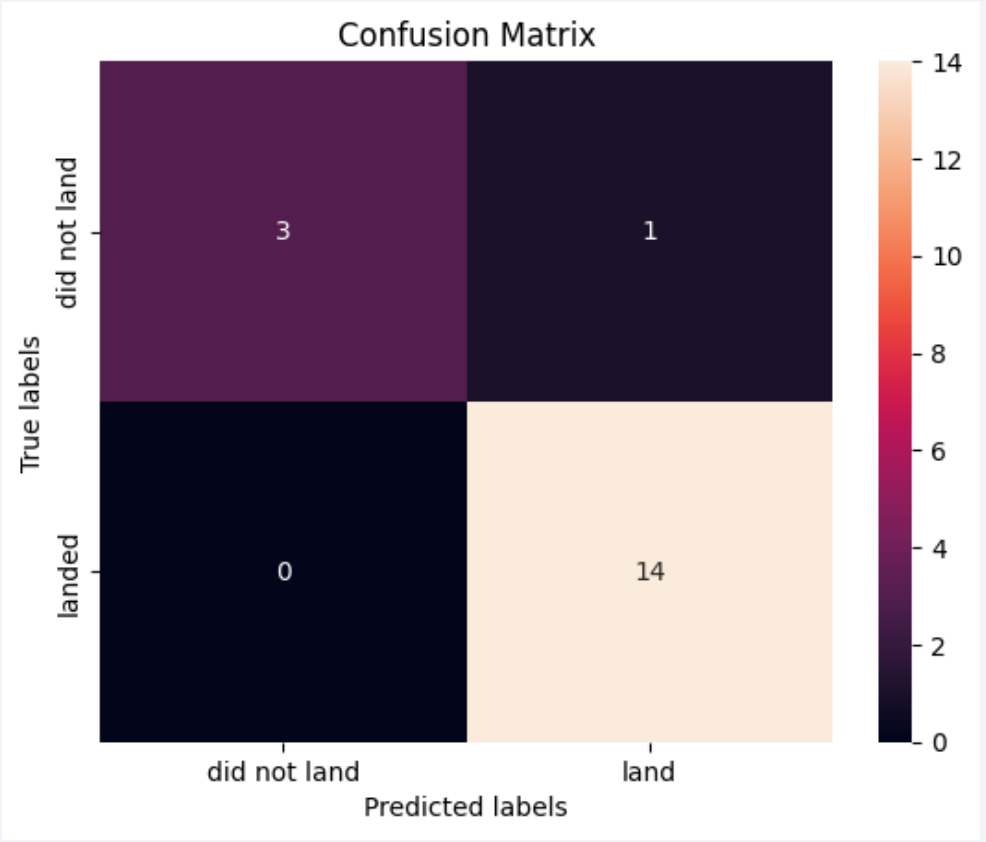
Explanation:

Based on algorithms comparison, Decision tree classifier is the model with the highest classification accuracy

Confusion Matrix

Explanation:

The confusion matrix for the decision tree classifier indicates its ability to differentiate between successful and failed landings. It correctly predicted 14 successful landings (true positives) and 3 failed landings (true negatives). However, one failed landing was misclassified as a success (false positive), while no successful landings were incorrectly predicted as failures (false negatives). The main issue with the model is the occurrence of false positives, where unsuccessful landings are mistakenly classified as successful.



Conclusions

- Launch success rates have steadily improved over time, indicating continuous technological and operational advancements.
- Higher launch frequency at a site correlates with greater overall success rates.
- The most successful orbits include ES-L1, GEO, HEO, SSO, and VLEO.KSC LC-39A recorded the highest number of successful launches, followed closely by CCAFS LC-40.
- Among the predictive models tested, the Decision Tree Classifier demonstrated the best performance with high accuracy, precision, and recall, making it the most suitable model for this task.

Appendix

- Notebook:
  - https://github.com/Fernando-Lim/data-science-capstone
- Charts:
  - https://github.com/Fernando-Lim/data-science-capstone/tree/main/flowcharts
- SpaceX API:
  - https://github.com/r-spacex/SpaceX-API/tree/master/docs#rspacex-api-docs
- Wikipedia:
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Thank you!