IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jaganathan A P
03/02/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of Methodologies:

- This project aims to develop a predictive model for the success of SpaceX Falcon 9's first-stage landing to help bid competitively against SpaceX for future launches.
- By analyzing historical data, the model forecasts landing outcomes based on key factors like rocket specifications, launch site, and launch parameters.
- With SpaceX's ability to reuse their first stage, which significantly lowers their launch costs, this model enables us to assess and optimize our launch strategies, offering competitive pricing and highlighting potential cost-saving opportunities by improving our own reusable rocket capabilities.

Executive Summary

Summary of the Results:

- The model predicts the success of SpaceX Falcon 9's first-stage landings using key factors like Booster version, Orbit, Payload mass and Launch site.
- After testing multiple algorithms, all the models showed similar performance, achieving around 83% accuracy. Key insights include the importance of launch site, payload mass and booster versions having higher impact on success rate.
- The models offer valuable insights for providing more competitive bids, potentially helping to decide on the quotes.

Introduction

Project background and context

- SpaceX has revolutionized the space industry with its Falcon 9 rocket, particularly by reusing the first stage of the rocket, significantly lowering the cost of space launches. This reusability has given SpaceX a competitive edge, allowing them to offer lower launch prices compared to traditional space companies.
- In this context, the goal of this project is to develop a predictive model that forecasts the success of Falcon 9's first-stage landing. By accurately predicting landing outcomes, we aim to better understand the key factors influencing landing success and improve our ability to compete with SpaceX in future bids for space launches.

Introduction

…continued

- By leveraging historical data from Falcon 9 launches, this project seeks to find the key features which contribute to the success or failure of the successful first stage landing. Ultimately, the model will support more competitive pricing and enhance our ability to compete with SpaceX’s industry-leading cost structure.

Problems to Address

1. What factors most influence the success or failure of SpaceX Falcon 9's first-stage landings? Can we predict landing outcomes with high accuracy?
2. Which features (e.g., launch site, payload mass and booster versions) are most critical in determining landing success?
3. How can we leverage insights from landing success predictions to offer more competitive bids against SpaceX?
4. What is the impact of SpaceX's ability to reuse its first stage on pricing, and how can we close the gap to offer comparable cost-effective launch services?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

- The data was collected from publicly available sources, including SpaceX launch records. It includes information from historical Falcon 9 missions, such as rocket specifications, launch sites, landing outcomes (success or failure), and other relevant factors like booster version and payload mass. This dataset provides the foundation for training and testing the predictive model.

Perform data wrangling

- The data was cleaned by handling missing values and outliers. Key features like booster version, payload mass, and launch site were extracted and encoded. Numerical features were scaled for model consistency. The dataset was split into training and testing sets, and techniques like oversampling were used to balance the class distribution for better model performance.

Methodology

Executive Summary

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

First, various classification models (e.g., Logistic Regression, Random Forest, SVM and KNN) were trained on the processed data. Hyperparameters were tuned using techniques like Grid Search to optimize model performance. The models were then evaluated using metrics such as accuracy, precision, recall, and F1 score to assess their effectiveness, ensuring reliable predictions of landing success. Cross-validation was used to validate performance across different datasets.

Data Collection

Identify Data Sources

- Collected historical data from public sources such as SpaceX's official website, space-related APIs, and websites like Wikipedia.
- Gathered information on Falcon 9 launches, including rocket specifications, landing success/failure, launch sites, location, and other relevant features.

Data Extraction

- Used web scraping techniques to collect raw data from SpaceX's API and relevant databases.
- Manually retrieved missing or supplementary data from official launch reports when possible.

Data Structuring

- Organized the data into structured formats (CSV, JSON, or databases) for ease of processing.
- Ensured that all data entries had consistent formats (e.g., date and time, categorical variables).

Data Integration

- Combined data from different sources to create a comprehensive dataset.
- Matched each rocket launch to its respective landing success/failure and external factors (e.g., rocket type, launch site).

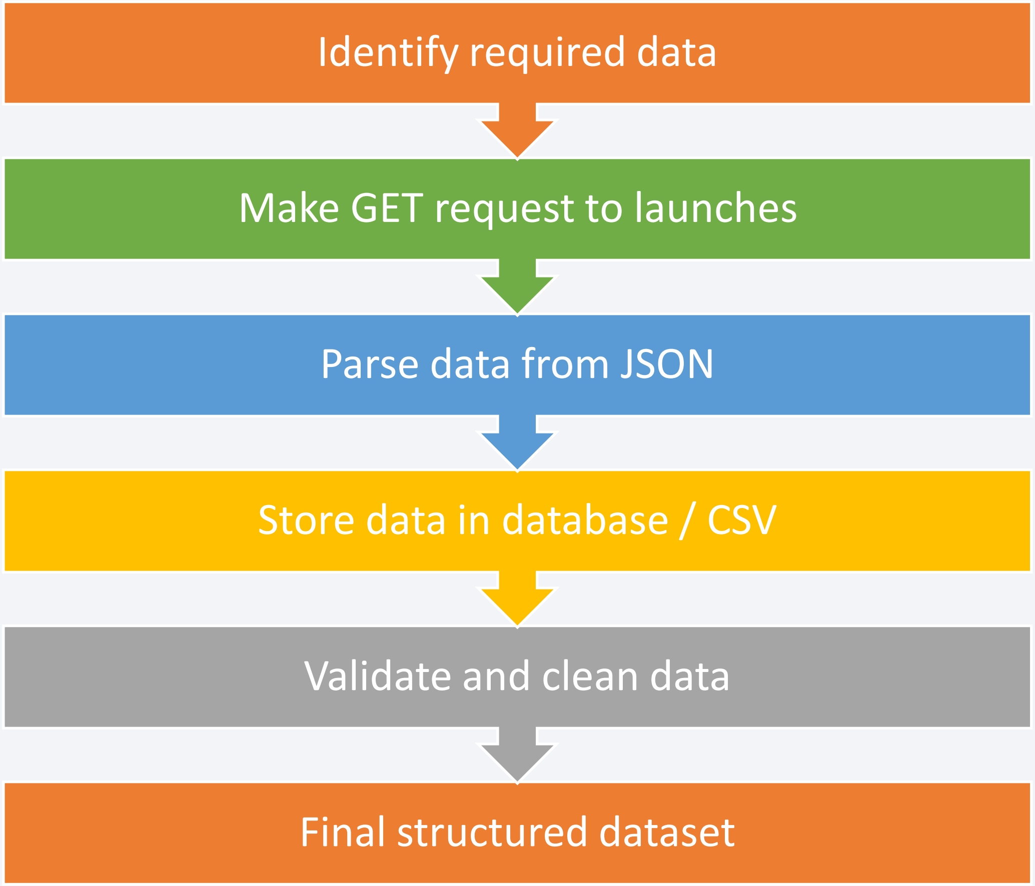
Data Quality Check

- Ensured accuracy and completeness by cross-referencing with trusted sources.
- Removed or corrected any discrepancies, duplicates, or erroneous data entries.

Data Collection – SpaceX API

Key phrases:

- Key data points—launch details, rocket specification, landing success/failure.
- API calls: GET / launches



Identify required data

Make GET request to launches

Parse data from JSON

Store data in database / CSV

Validate and clean data

Final structured dataset

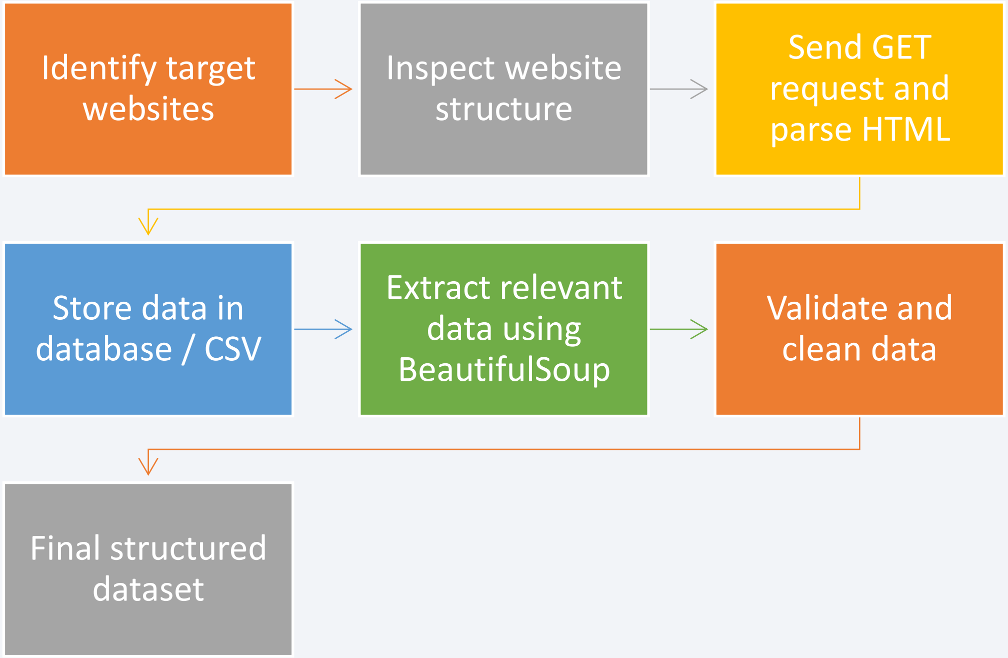
GitHub URL of the completed SpaceX API calls notebook:

[Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · apjaganathan/Capstone]

Data Collection – Scraping

Key phrases:

- Select reliable sources (e.g., SpaceX website, Wikipedia) for launch data.
- Analyze the website's HTML to locate the required data.
- Use **BeautifulSoup** and **Requests** to extract key information (e.g., launch sites, landing outcomes).


Identify target websites
Inspect website structure
Send GET request and parse HTML
Store data in database / CSV
Extract relevant data using BeautifulSoup
Validate and clean data
Final structured dataset

GitHub URL of the completed Webscraping notebook:

Capstone/jupyter-labs-webscraping.ipynb at main · apjaganathan/Capstone

Data Wrangling

Data Cleaning

Handle Missing Values: Impute or remove missing data using techniques like mean imputation or deletion of incomplete records.

Remove Duplicates: Identify and remove duplicate entries to ensure data accuracy.

Data Transformation

Feature Engineering: Extract and create new features.

Data Normalization/Scaling: Scale numerical features for consistency across models.

Data Encoding

Convert Categorical Data: Encode categorical variables like launch site and booster version using techniques like one-hot encoding.

Data Integration

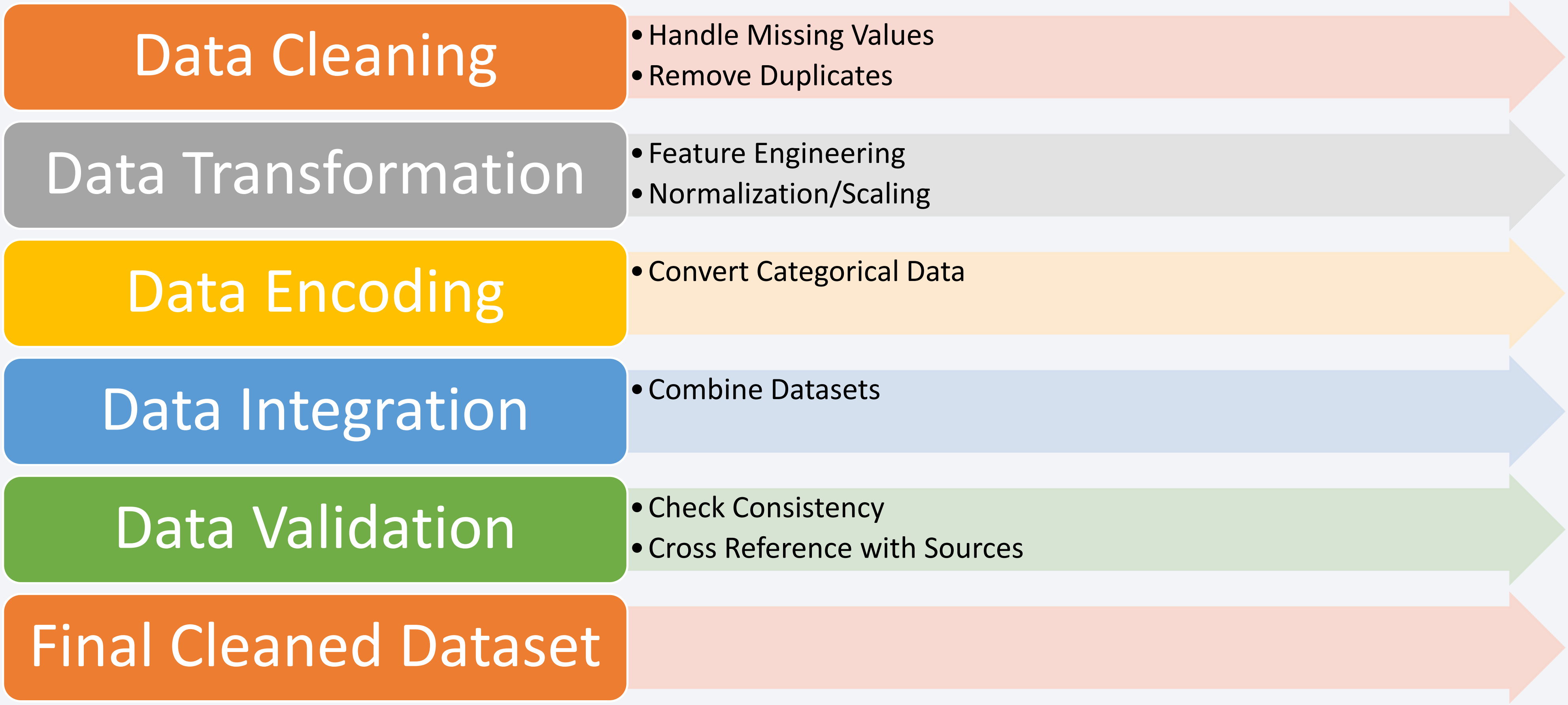
Combine Datasets: Merge data from multiple sources into a single comprehensive dataset.

Data Validation

Check Consistency: Ensure all data entries are consistent and accurately formatted.

Cross-Reference: Validate data points with trusted sources to ensure quality and correctness.

Data Wrangling



Data Cleaning

- Handle Missing Values
- Remove Duplicates

Data Transformation

- Feature Engineering
- Normalization/Scaling

Data Encoding

- Convert Categorical Data

Data Integration

- Combine Datasets

Data Validation

- Check Consistency
- Cross Reference with Sources

Final Cleaned Dataset

GitHub URL of the completed Data Wrangling notebook:

[Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · apjaganathan/Capstone]()

EDA with Data Visualization

Summary of Charts Used:

1.Scatter Plots:

- **Flight Number vs. Payload Mass & Launch Sites:** To visualize relationships and trends between flight number, payload mass, and launch sites.
- **Flight Number vs. Orbit & Payload Mass:** To explore how flight numbers correlate with different orbit types and payload mass.

2.Bar Chart:

- **Success Rate by Orbit:** To compare success rates across different orbits, making categorical comparisons easy.

3.Line Plot:

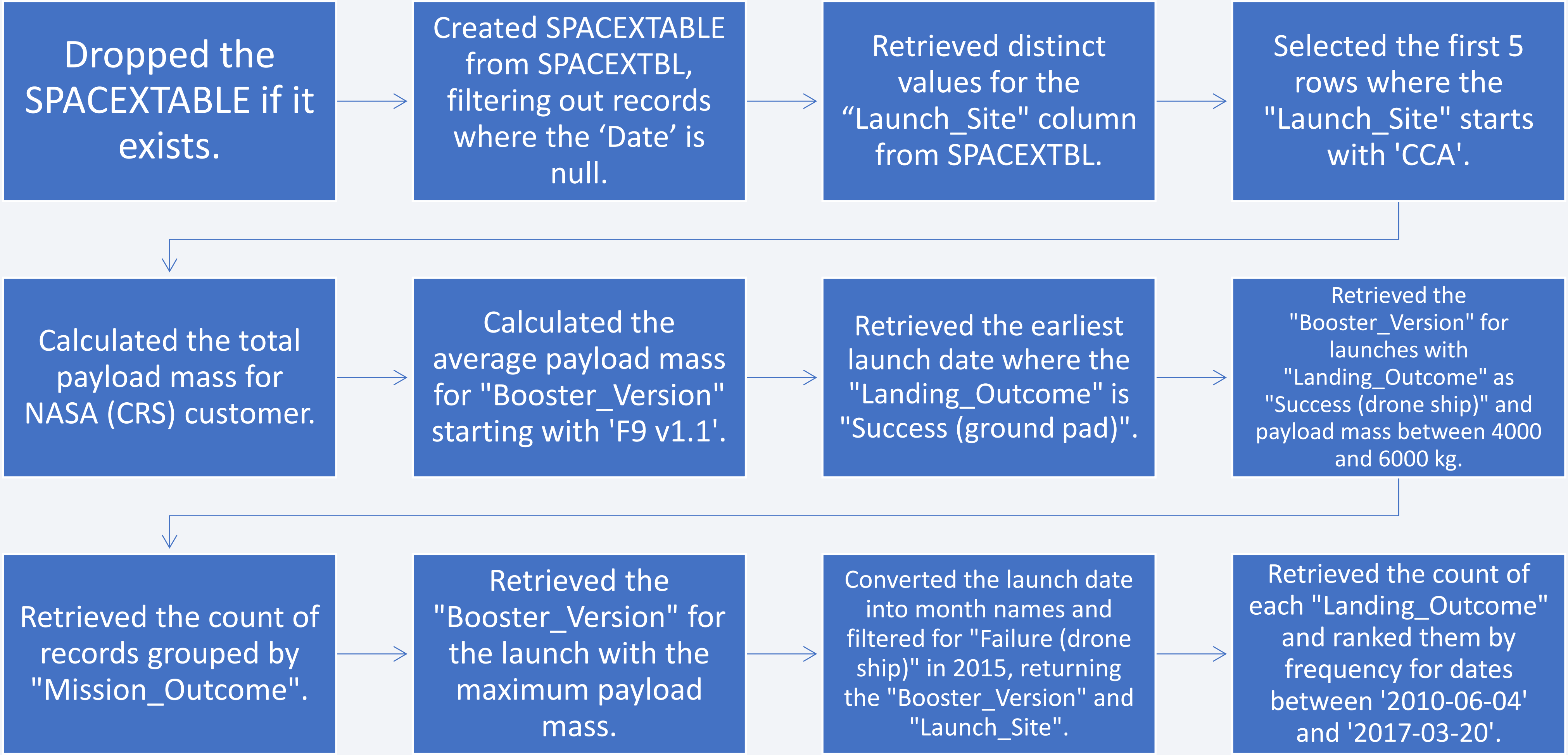
- **Average Launch Success by Year:** To track changes in launch success rates over time.

These charts help identify patterns and trends, providing valuable insights for launch analysis.

GitHub URL of the completed EDA with data visualization notebook:

[Capstone/edadataviz.ipynb at main · apjaganathan/Capstone]()

EDA with SQL



Dropped the SPACEXTABLE if it exists.

Created SPACEXTABLE from SPACEXTBL, filtering out records where the ‘Date’ is null.

Retrieved distinct values for the “Launch_Site" column from SPACEXTBL.

Selected the first 5 rows where the "Launch_Site" starts with 'CCA'.

Calculated the total payload mass for NASA (CRS) customer.

Calculated the average payload mass for "Booster_Version" starting with 'F9 v1.1'.

Retrieved the earliest launch date where the "Landing_Outcome" is "Success (ground pad)".

Retrieved the "Booster_Version" for launches with "Landing_Outcome" as "Success (drone ship)" and payload mass between 4000 and 6000 kg.

Retrieved the count of records grouped by "Mission_Outcome".

Retrieved the "Booster_Version" for the launch with the maximum payload mass.

Converted the launch date into month names and filtered for "Failure (drone ship)" in 2015, returning the "Booster_Version" and "Launch_Site".

Retrieved the count of each "Landing_Outcome" and ranked them by frequency for dates between '2010-06-04' and '2017-03-20'.

GitHub URL of the completed EDA with SQL notebook:

Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · apjaganathan/Capstone

Build an Interactive Map with Folium

Circle and Marker for Launch Sites: Created a circle and marker for each launch site.

Marker Cluster for Success/Failed Launches: Used marker clusters to group success and failed launches at each launch site.

Mouse Position: Added a mouse position feature to display the current coordinates (latitude and longitude) when the user moves their mouse over the map.

Polyline for Nearest Coastline, Railway and Highway Distance: Calculated and visualized the distance to the nearest coastline, railway and highway using a polyline.

These map objects were added to provide clear, interactive, and informative visualization of launch site data, launch outcomes, and geographical relationships, enhancing the user's ability to analyze and explore the data spatially.

GitHub URL of the completed Interactive map with Folium map notebook:

Capstone/lab_jupyter_launch_site_location.ipynb at main · apjaganathan/Capstone

Build a Dashboard with Plotly Dash

Summary of Plots/Graphs and Interactions:

Dropdown for Launch Site Selection:

Added a dropdown to select launch sites, making the dashboard interactive and customizable.

Pie Chart for Success / Failure by Launch Site:

Displayed a pie chart of success and failure rates for the selected launch site, with 'All-site' showing success percentages across sites for easy comparison.

Range Slider for Payload Mass:

Added a range slider to filter payload mass between 0 and 10,000 kg, enabling exploration of its impact on success/failure.

Scatter Plot for Success / Failure vs. Payload Mass:

Displayed a scatter plot to show the relationship between payload mass and launch success/failure, highlighting how different payload ranges impact landing outcomes

GitHub URL of the completed Plotly Dash lab:

Capstone/Plotly_Dash code at main · apjaganathan/Capstone

Build a Dashboard with Plotly Dash

Explanation:

These plots and interactions allow for a detailed, dynamic exploration of the data.

The dropdown and range slider enable filtering, making it easy for users to select a specific launch site and payload mass range.

The pie chart provides a quick overview of launch success/failure at the selected site, while the scatter plot offers deeper insights into how payload mass and booster version impact success rates.

These interactive features improve user engagement and make the analysis more customizable.

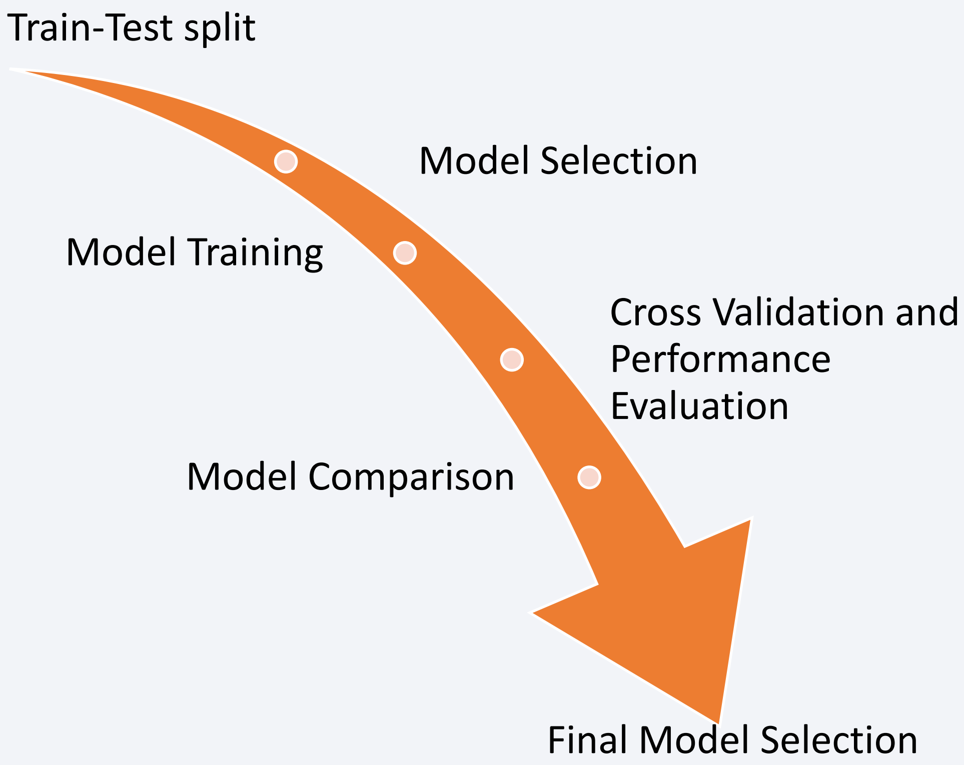
GitHub URL of the completed Plotly Dash lab:

[Capstone/Plotly_Dash code at main · apjaganathan/Capstone]()

Predictive Analysis (Classification)

Key phrases:

- **Train-Test Split**: Dividing data into training and test sets for evaluation
- **Model Evaluation**: Using performance metrics like accuracy, confusion matrix
- **Hyperparameter Tuning**: Fine-tuning parameters using GridSearchCV
- **Model Comparison**: Evaluating and comparing the models' performance
- **Final Model Selection**: Deciding on the best model based on overall performance and business needs


Train-Test split

Model Selection

Model Training

Cross Validation and Performance Evaluation

Model Comparison

Final Model Selection

GitHub URL of the completed predictive analysis lab:

Capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · apjaganathan/Capstone

Exploratory data analysis results

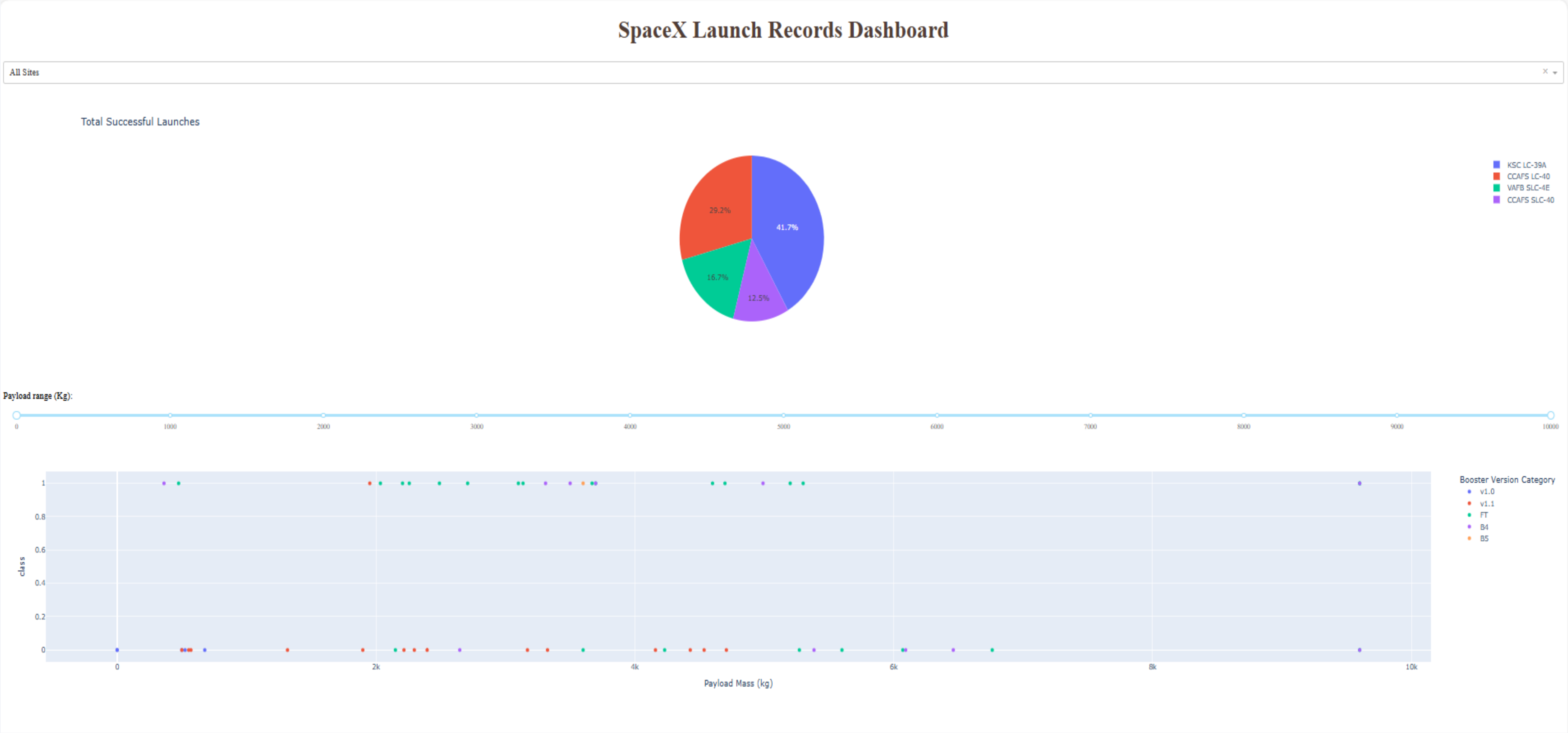
As the flight number increases, the likelihood of a successful landing improves, particularly in the LEO orbit.

For heavier payloads (greater than 10,000 kg), there are no launches from the VAFB-SLC site. Success rates for heavy payloads are higher for Polar, LEO, and ISS orbits, while GTO shows no clear pattern between success and failure.

Landing success rates have steadily increased from 2013 to 2020.

Results

Screenshot of the Interactive analytics demo



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Predictive Analysis Results

Four models : Logistic Regression, KNN, SVM, and Random Forest were evaluated, all achieving an **accuracy of 83.33%** on the test data.

The models showed similar performance in terms of **accuracy**, **precision**, **recall**, and **F1-score**.

Despite the close performance, **Random Forest** was preferred due to its ability to capture complex relationships and handle non-linear data.

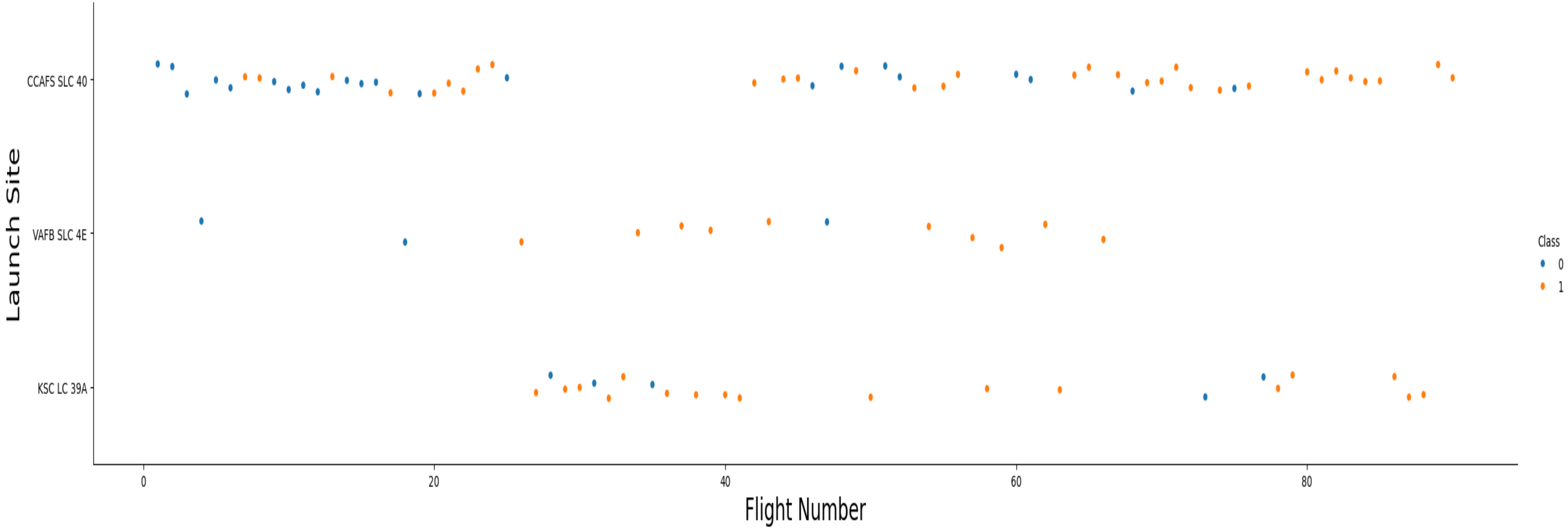
Key features influencing landing success included **flight number** and **payload mass**, with successful landings more likely as the flight number increases.

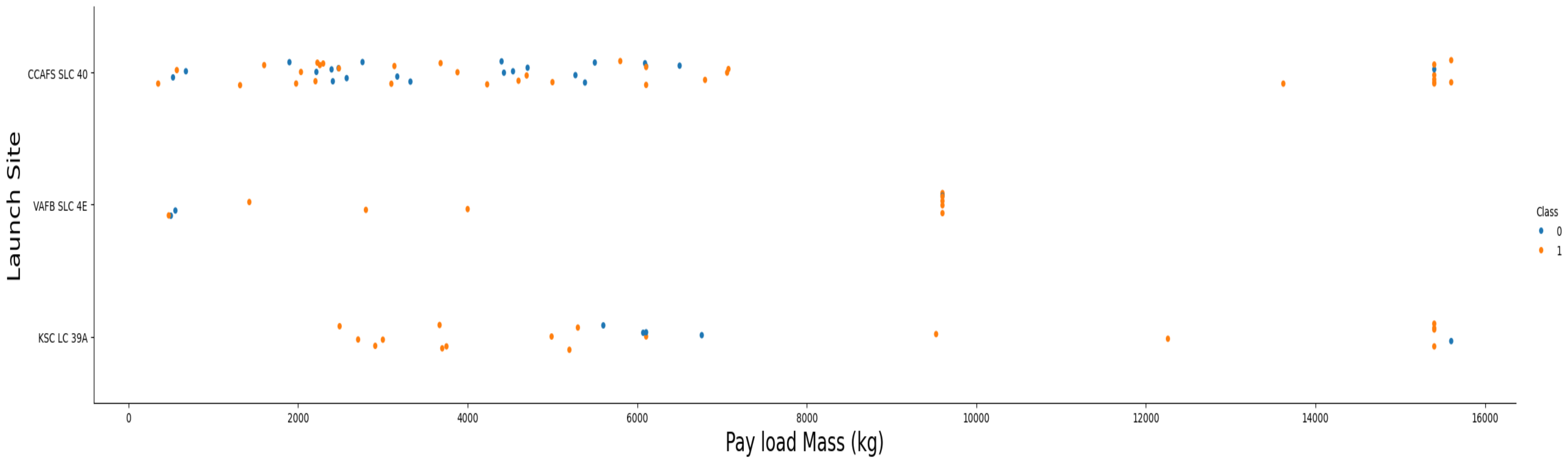
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

We observed that as the Flight number increases, the success rate for each of the launch sites were higher.



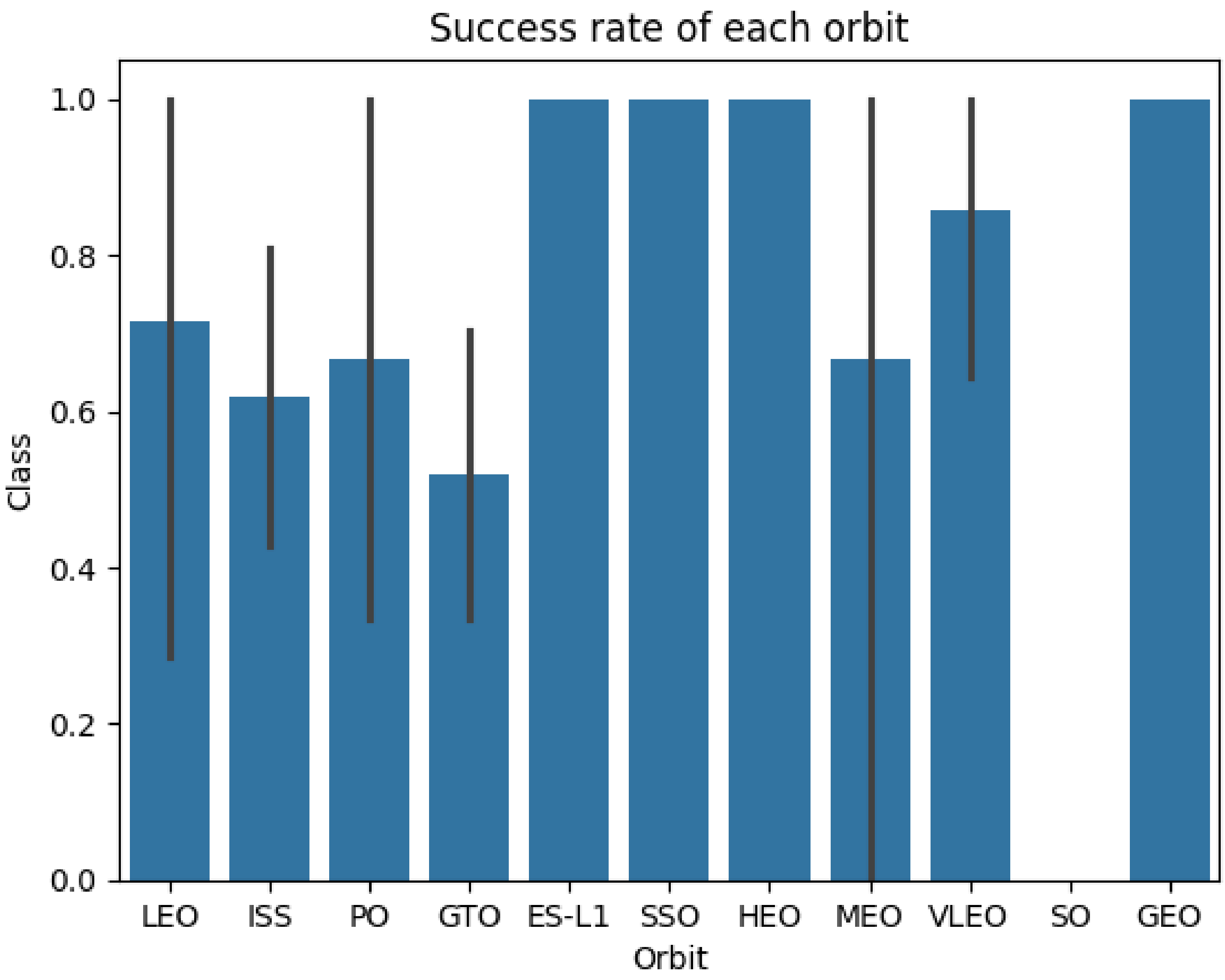


Payload vs. Launch Site

When we observe the Payload Mass Vs. Launch Site scatter point chart, we find that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

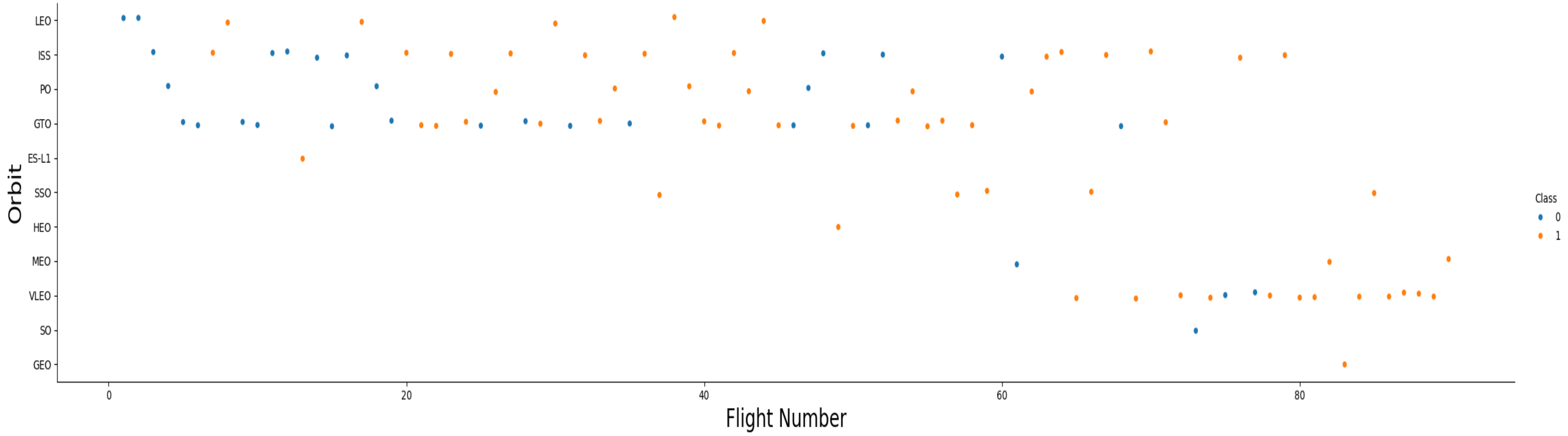
Success Rate vs. Orbit Type

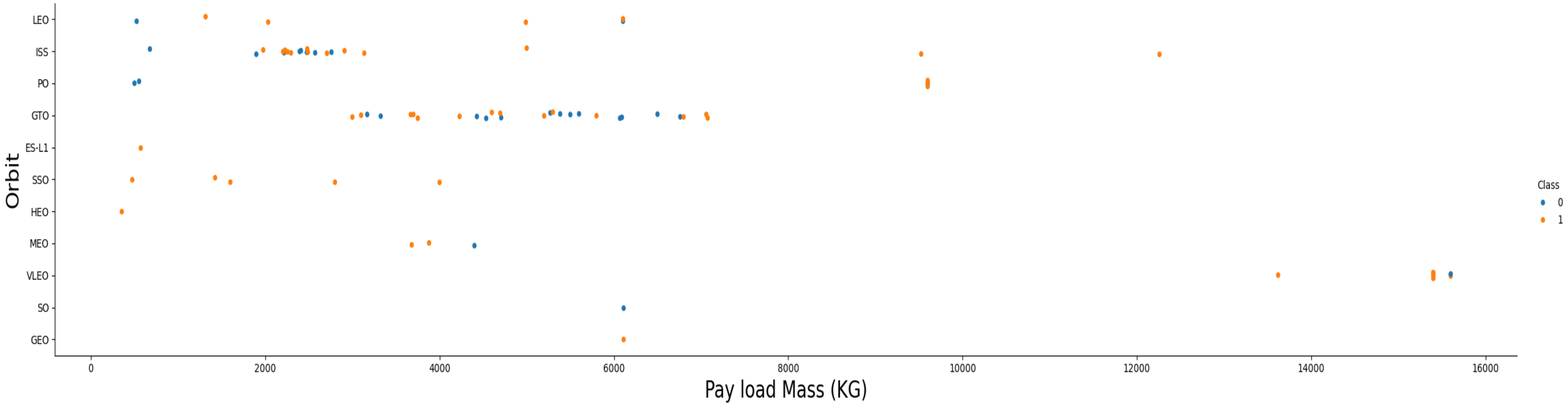
From the bar chart we can observe that the success rate for ES-L1, SSO, HEO and GEO orbits were the highest.



Flight Number vs. Orbit Type

- As the flight number increases, there is likely a trend indicating a shift in the types of orbits used over time, reflecting SpaceX's evolving launch capabilities.
- The plot also reveals how certain orbit types (like LEO or GTO) are more frequent at different stages of the flight program.



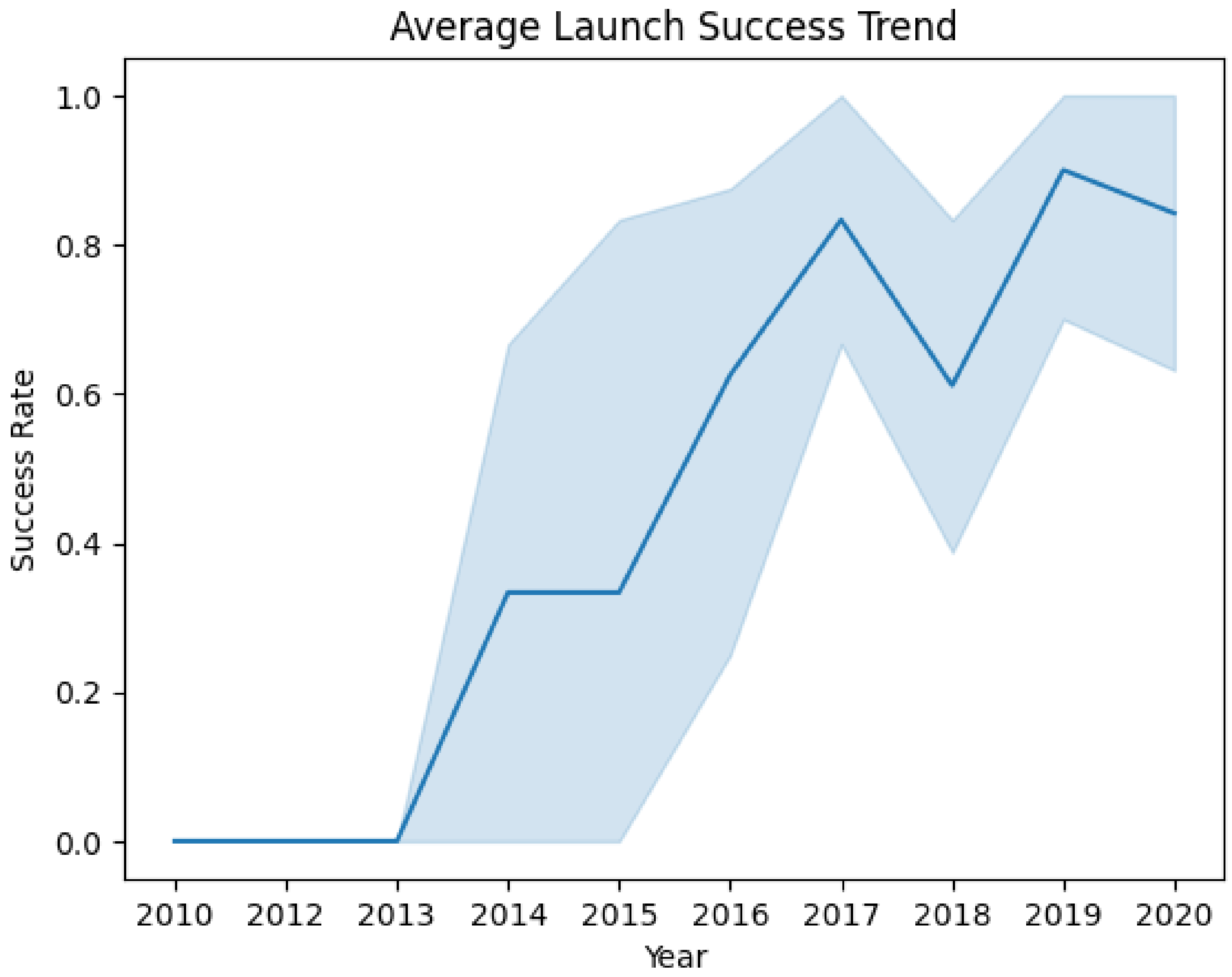


Payload vs. Orbit Type

Heavy payloads tend to have higher landing success rates in Polar, LEO, and ISS orbits. However, for GTO, distinguishing between successful and unsuccessful landings is challenging, as both outcomes are observed.

Launch Success Yearly Trend

The success rate has shown a steady increase from 2013 to 2020.



All Launch Site Names

- %sql select distinct "Launch_Site" from SPACEXTBL;
- This query pulls distinct values from the Launch_Site column, providing a list of all the unique launch sites used for SpaceX launches recorded in the dataset. This helps in understanding the different locations from which Falcon 9 rockets are launched.

Launch Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

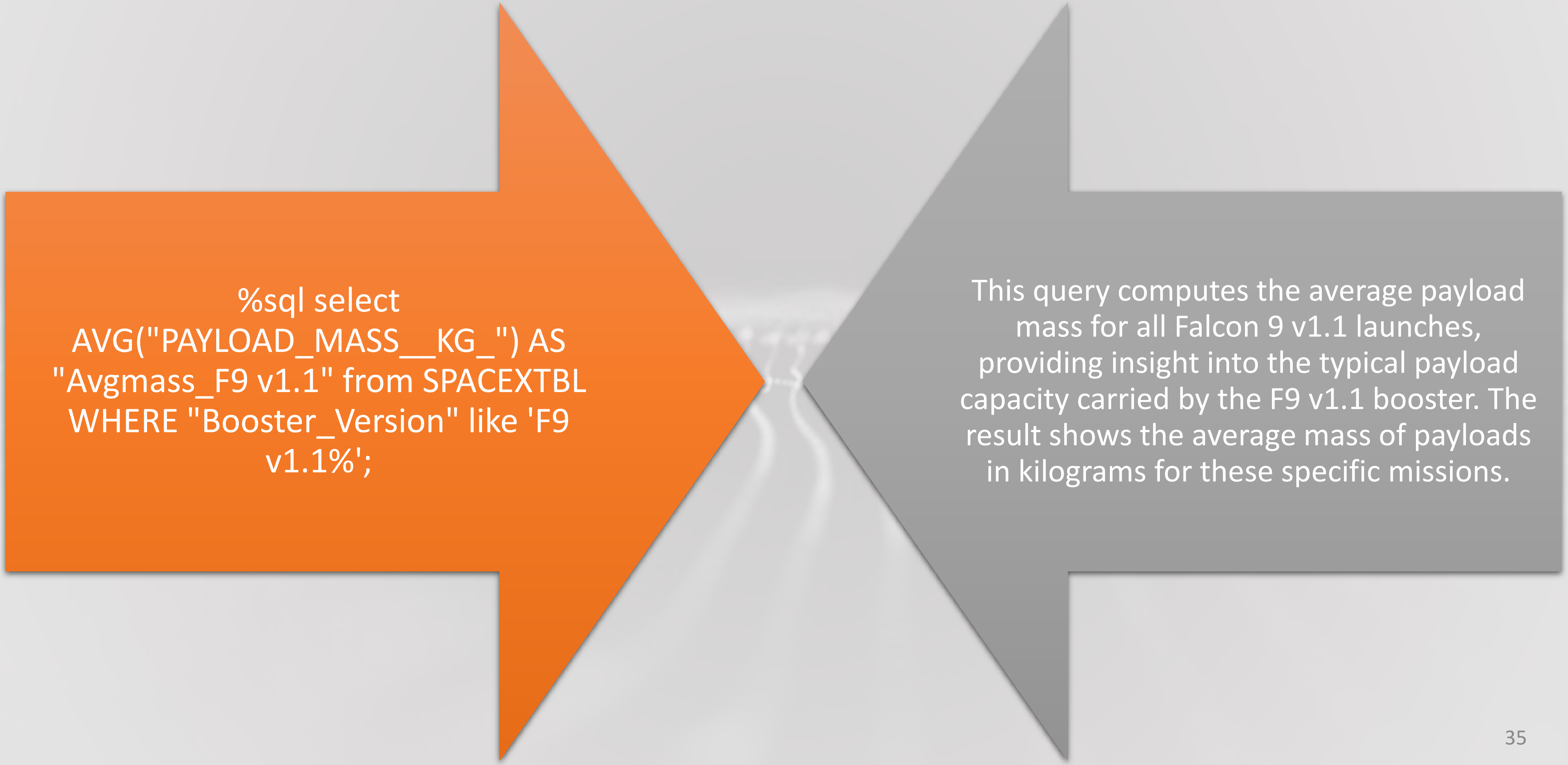
Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5;
- This query filters the data to only include launches from launch sites starting with "CCA", which in this case would include sites like CCAFS-LC. The result shows the first 5 launch records.
- This helps in analyzing specific launch sites, like those starting with "CCA", and their corresponding mission outcomes.

Total Payload Mass

- %sql select SUM("PAYLOAD_MASS__KG_") as "Totalmass_NASA(CRS)" FROM SPACEXTBL WHERE "Customer" = "NASA (CRS)";
- This query sums up the payload mass for all SpaceX missions where NASA (CRS) is listed as the customer. The result shows the total mass of payloads launched for NASA (CRS) in kilograms, helping to understand the scale of NASA's payloads handled by SpaceX.

Average Payload Mass by F9 v1.1



```
%sql select
AVG("PAYLOAD_MASS__KG_") AS
"Avgmass_F9 v1.1" from SPACEXTBL
WHERE "Booster_Version" like 'F9
v1.1%';
```

This query computes the average payload mass for all Falcon 9 v1.1 launches, providing insight into the typical payload capacity carried by the F9 v1.1 booster. The result shows the average mass of payloads in kilograms for these specific missions.

First Successful Ground Landing Date

- %sql select min("Date") from SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)";
- This query finds the earliest launch date from the SPACEXTBL table where the landing outcome was "Success (ground pad)". The result shows the first recorded successful ground pad landing by SpaceX, helping to identify the initial success in this type of landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select "Booster_Version" from SPACEXTBL where "Landing_Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000;
- This query lists the different booster versions that successfully landed on a drone ship, with payloads between 4000 and 6000 kg. The result highlights which booster versions were used for these specific missions, showcasing the ability of different versions to achieve successful drone ship landings with medium-weight payloads.

Total Number of Successful and Failure Mission Outcomes

- %sql select "Mission_Outcome", count(*) as "outcome_no" from SPACEXTBL group by "Mission_Outcome";
- This query provides a summary of the different mission outcomes and the number of missions corresponding to each outcome. It helps in understanding how many successful, failed, and partially successful missions SpaceX has had, based on the Mission_Outcome field in the dataset.

Boosters Carried Maximum Payload

```
%sql select "Booster_Version"
from SPACEXTBL WHERE
"PAYLOAD_MASS__KG_" =
(SELECT
MAX("PAYLOAD_MASS__KG_")
FROM SPACEXTBL);
```

This query finds the booster version used for the mission that carried the heaviest payload, by identifying the launch with the maximum payload mass.

The result shows which booster version was used for the heaviest payload, helping to identify the most powerful Falcon 9 variant in terms of payload capacity.

2015 Launch Records

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

- This query converts the month (from the Date column) into a readable month name and retrieves the booster version and launch site for missions in 2015 that had a failed landing on a drone ship. The result helps in identifying which months had failed drone ship landings, along with the associated booster versions and launch sites.

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

- %sql select "Landing_Outcome", count(*) as "Outcome_count", ROW_NUMBER() OVER (ORDER BY COUNT(*) DESC) AS "RANK" from SPACEXTBL WHERE "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by "RANK";
- This query groups data by Landing Outcome, counts occurrences within a date range, and ranks them by frequency to highlight the most common outcomes.

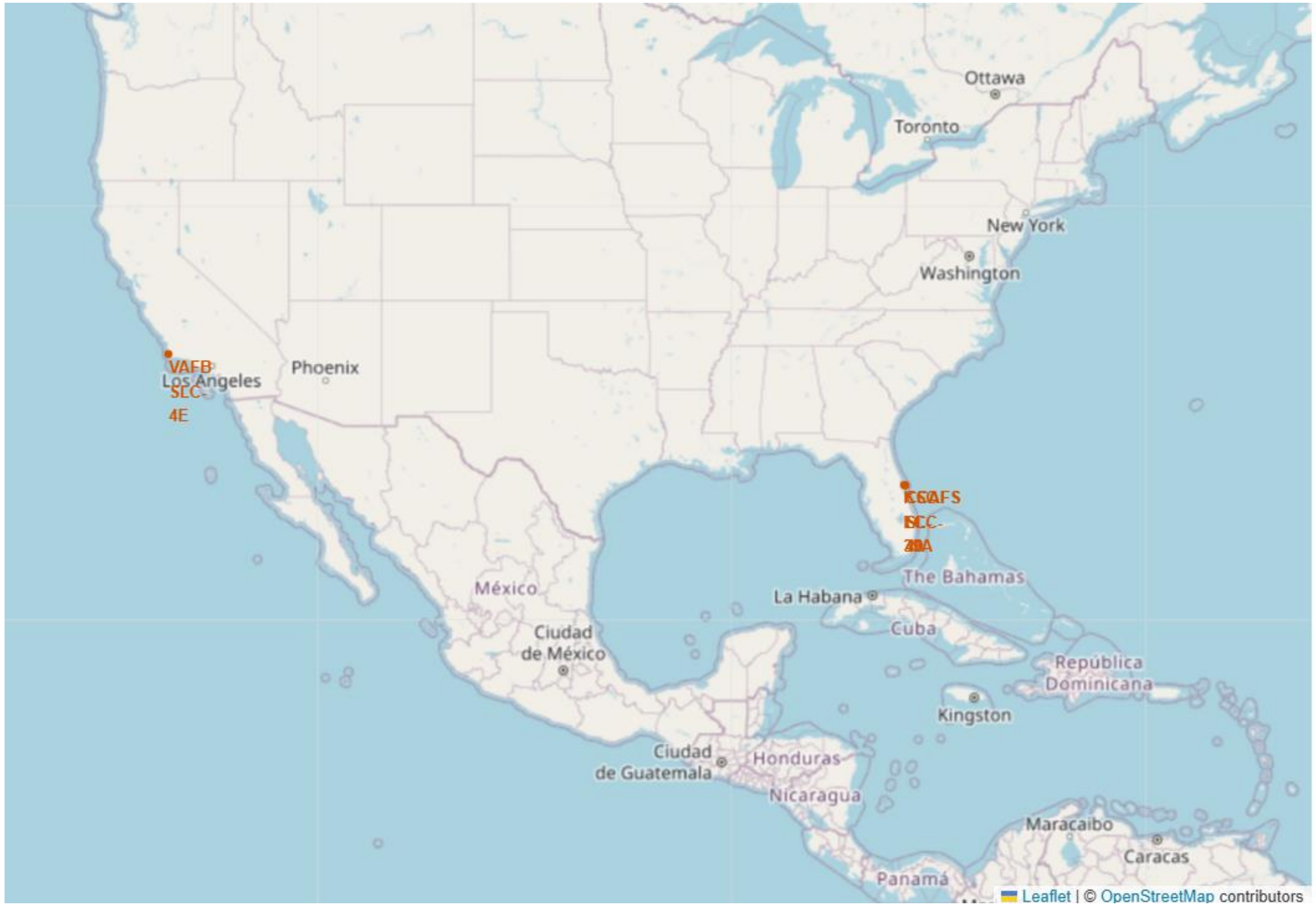
Section 3

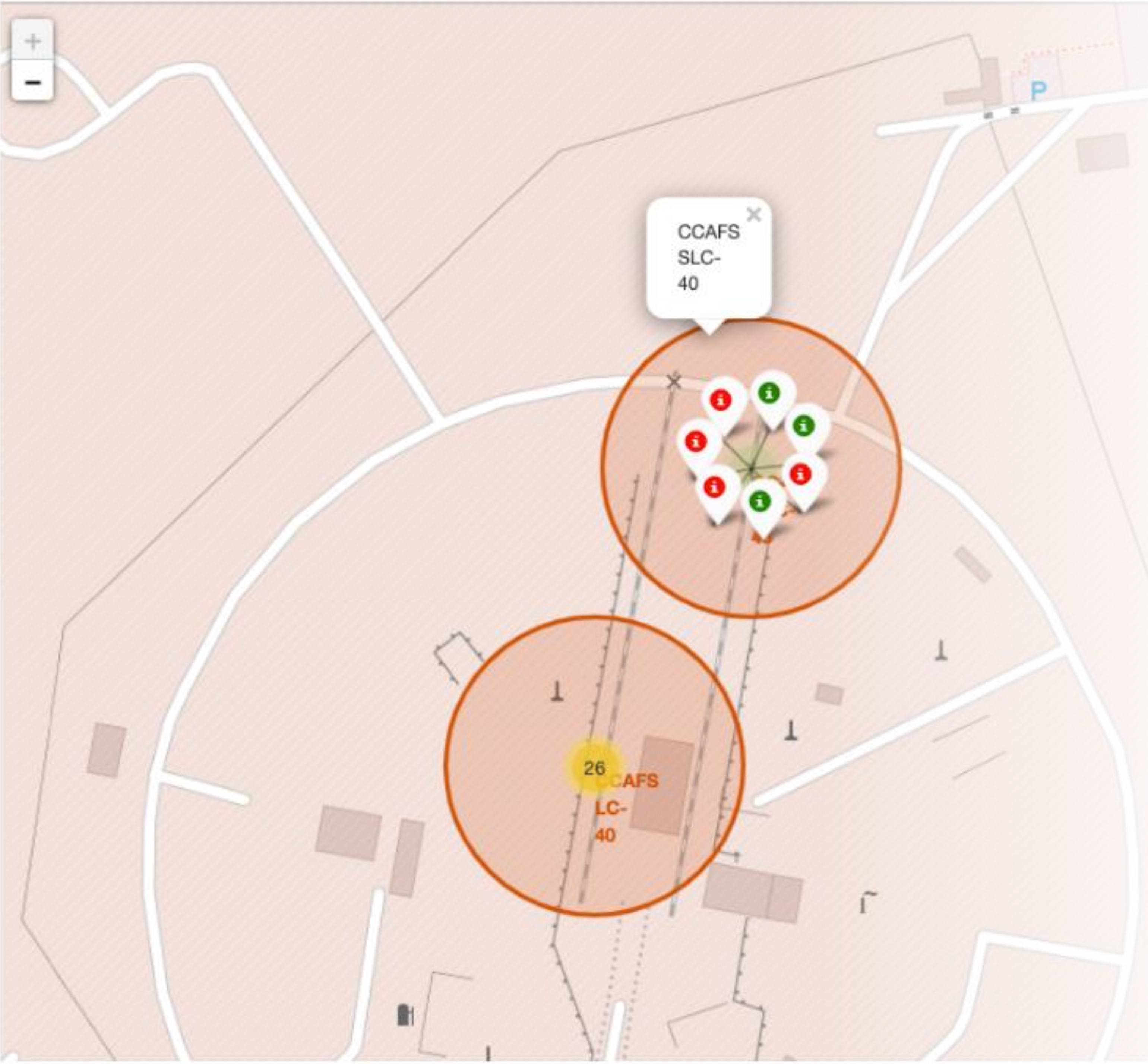
Launch Sites Proximities Analysis

Map of SpaceX Launch Sites’ Location Markers

- SpaceX’s launch sites are spread across different regions of the U.S., with a concentration in Florida, supporting both commercial and government launches.

- The sites are strategically placed for different types of missions, including missions to low Earth orbit (LEO), geostationary orbit (GEO), and polar orbits.





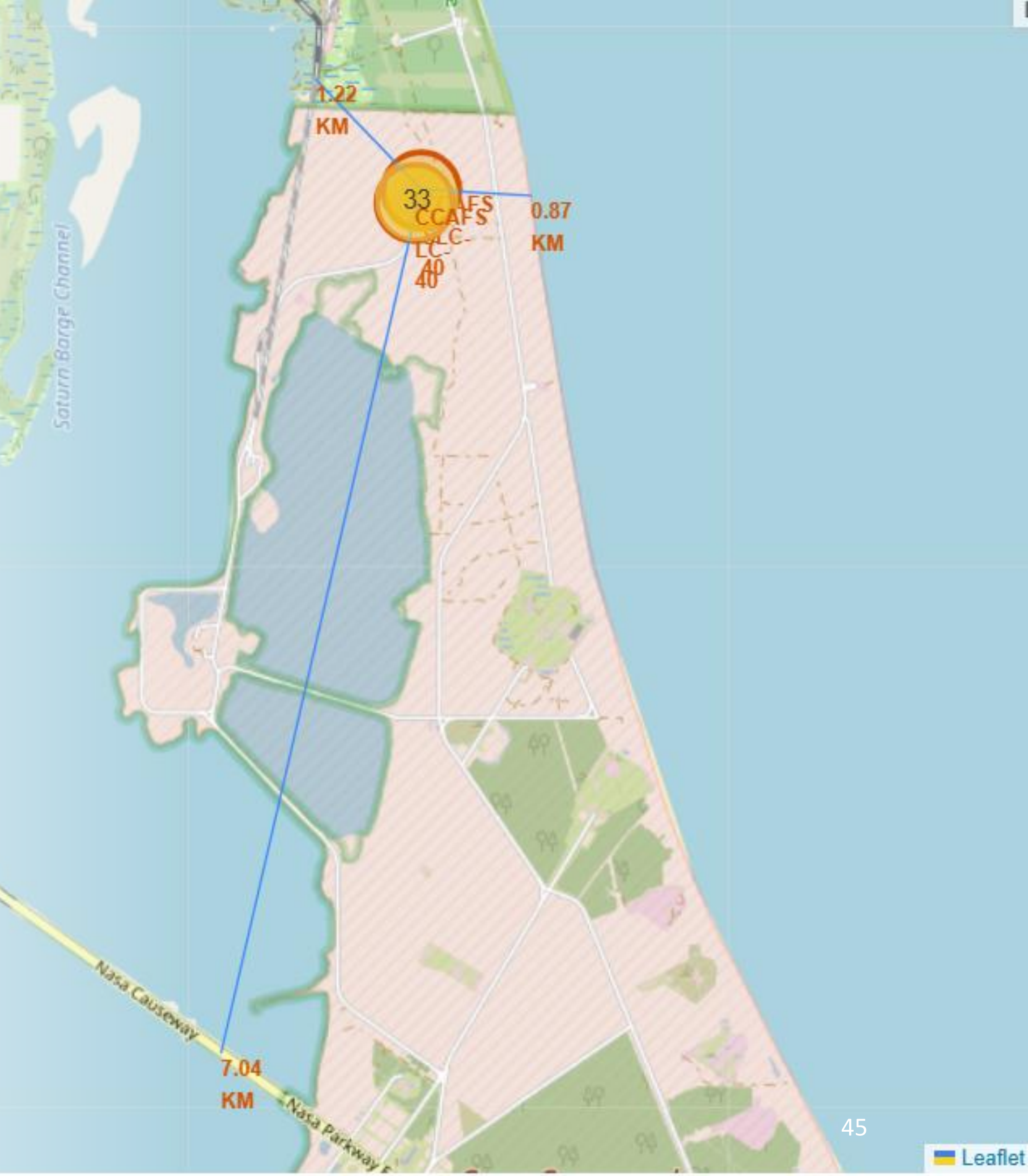
Map of SpaceX Launch Outcomes with Color-Labeled Markers

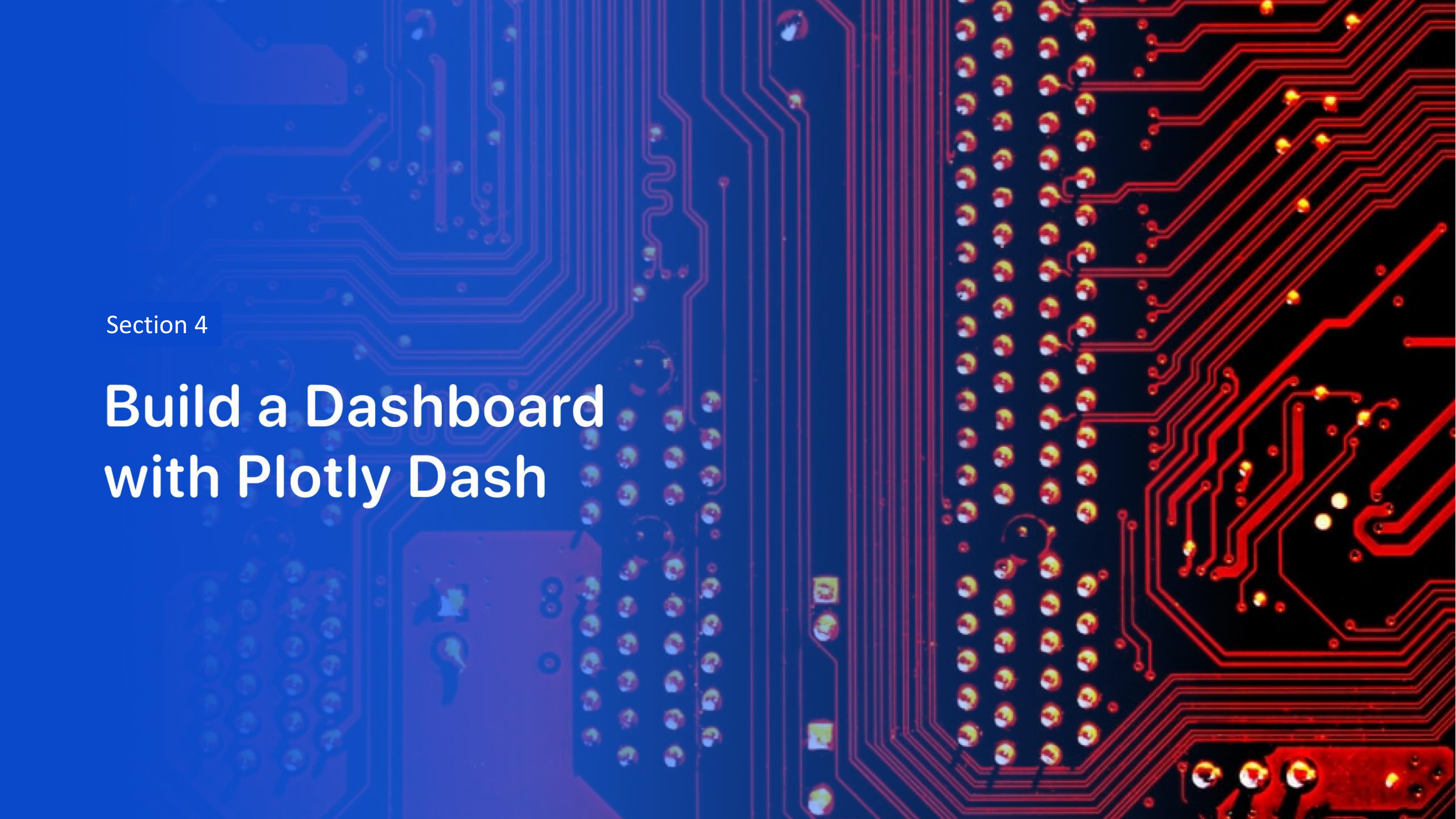
The color-coded markers clearly show the distribution of successful and failed launches, with green for success and red for failure, helping to quickly assess performance at different sites.

Proximity Map of CCAFS SLC-40

Distances to Nearest Coastline, Railway, and Highway

- Launch sites are generally not directly adjacent to railways but are in moderate proximity for logistics purposes.
- Launch sites like CCAFS SLC-40 are usually close to highways for easier transportation of rockets and equipment.
- They are very close to the coastline to allow launches over the ocean and reduce risk to populated areas.
- Launch sites are typically located a safe distance away from cities to minimize risk in case of an accident or failure.

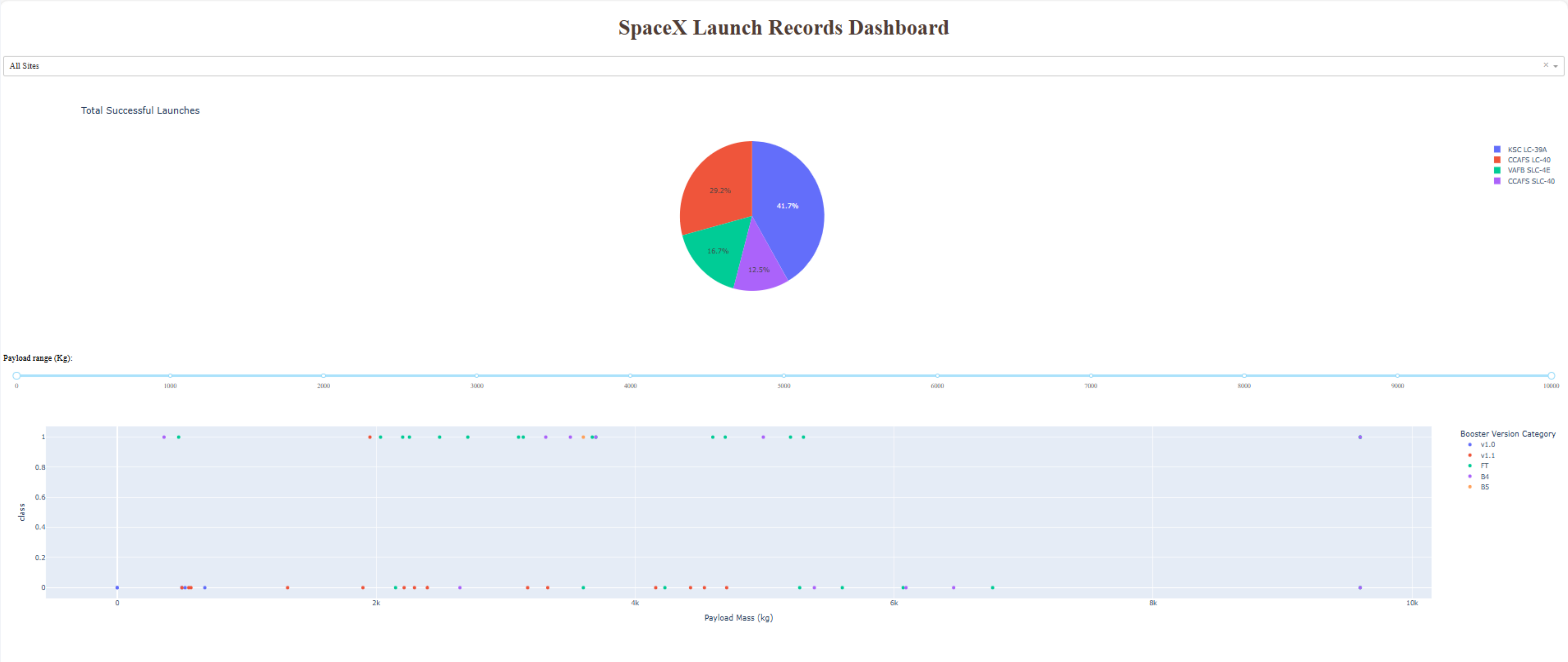




Section 4

Build a Dashboard with Plotly Dash

Pie Chart of Launch Success Distribution Across SpaceX Sites

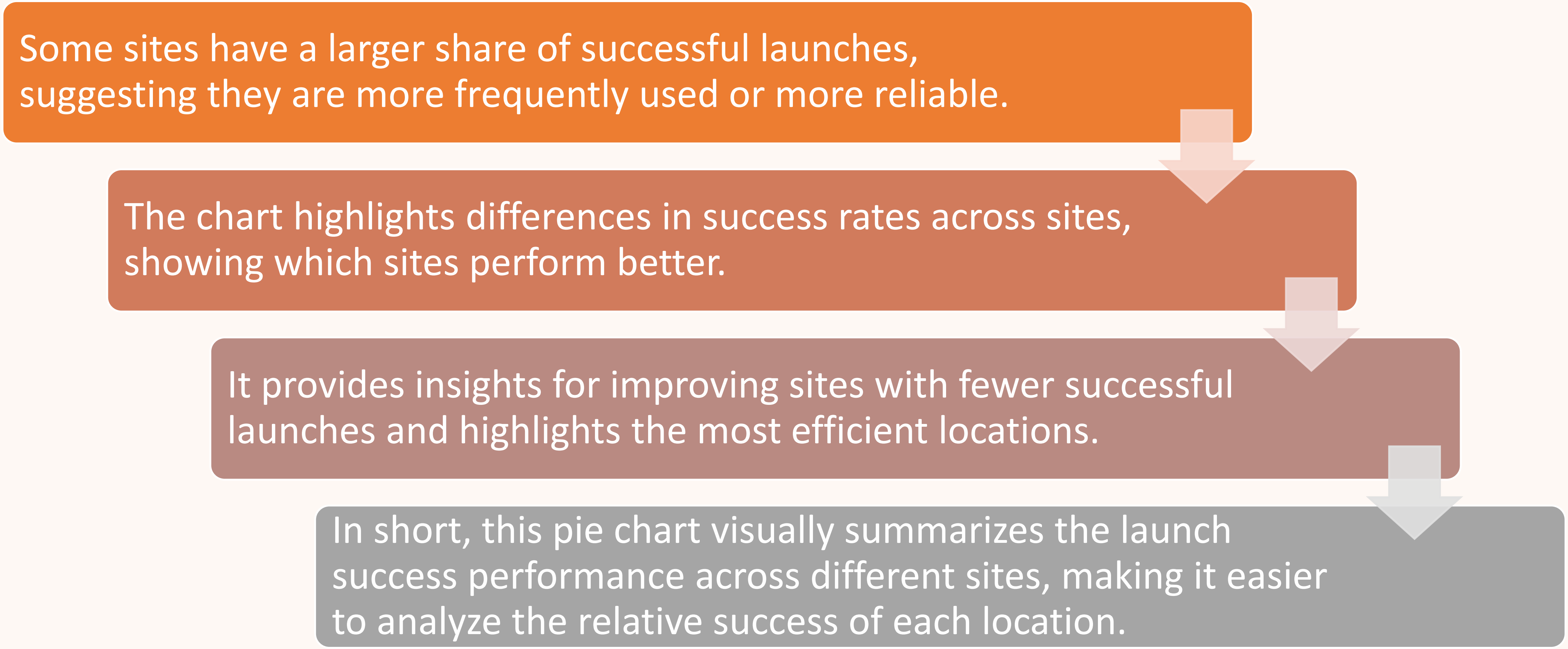


SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Pie Chart of Launch Success Distribution Across SpaceX Sites



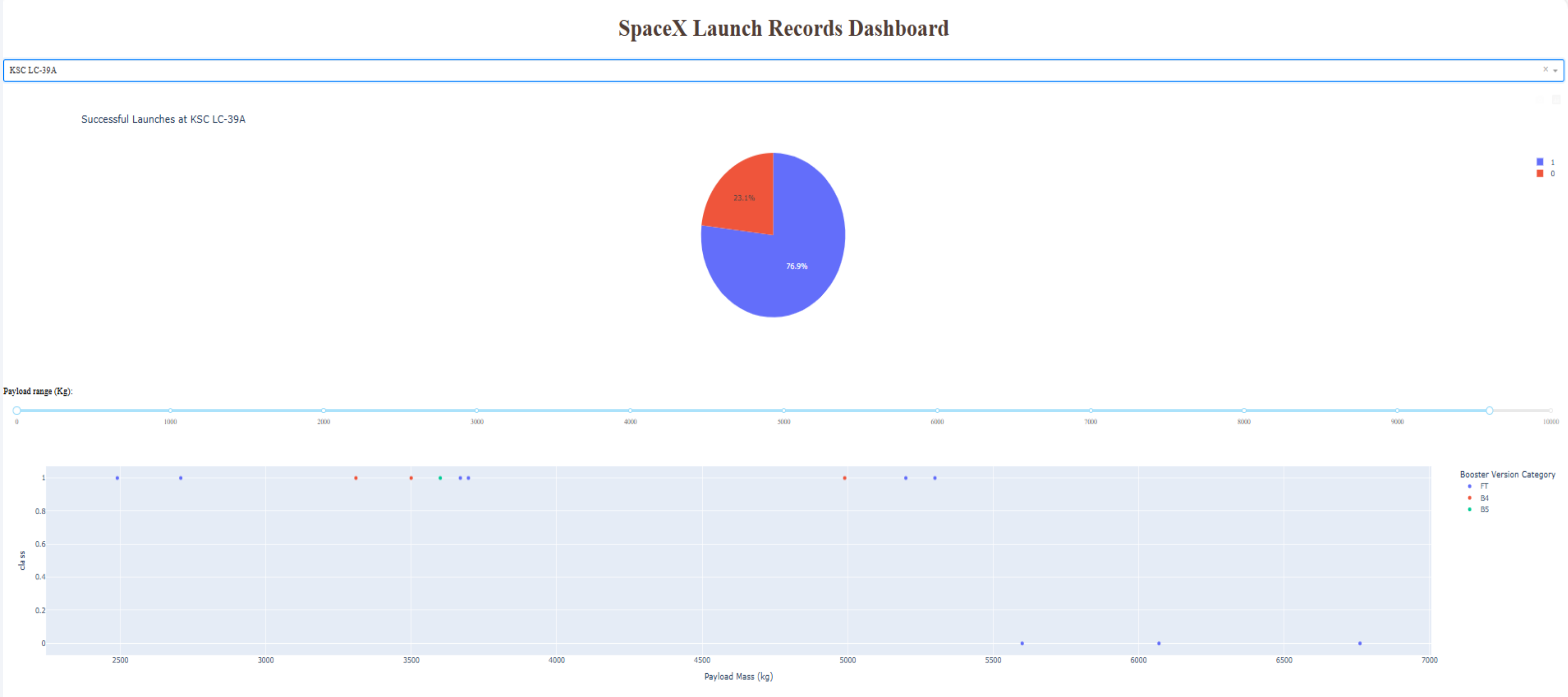
Some sites have a larger share of successful launches, suggesting they are more frequently used or more reliable.

The chart highlights differences in success rates across sites, showing which sites perform better.

It provides insights for improving sites with fewer successful launches and highlights the most efficient locations.

In short, this pie chart visually summarizes the launch success performance across different sites, making it easier to analyze the relative success of each location.

Launch Success Ratio at KSC LC-39A



SpaceX Launch Records Dashboard

KSC LC-39A

Payload range (Kg):

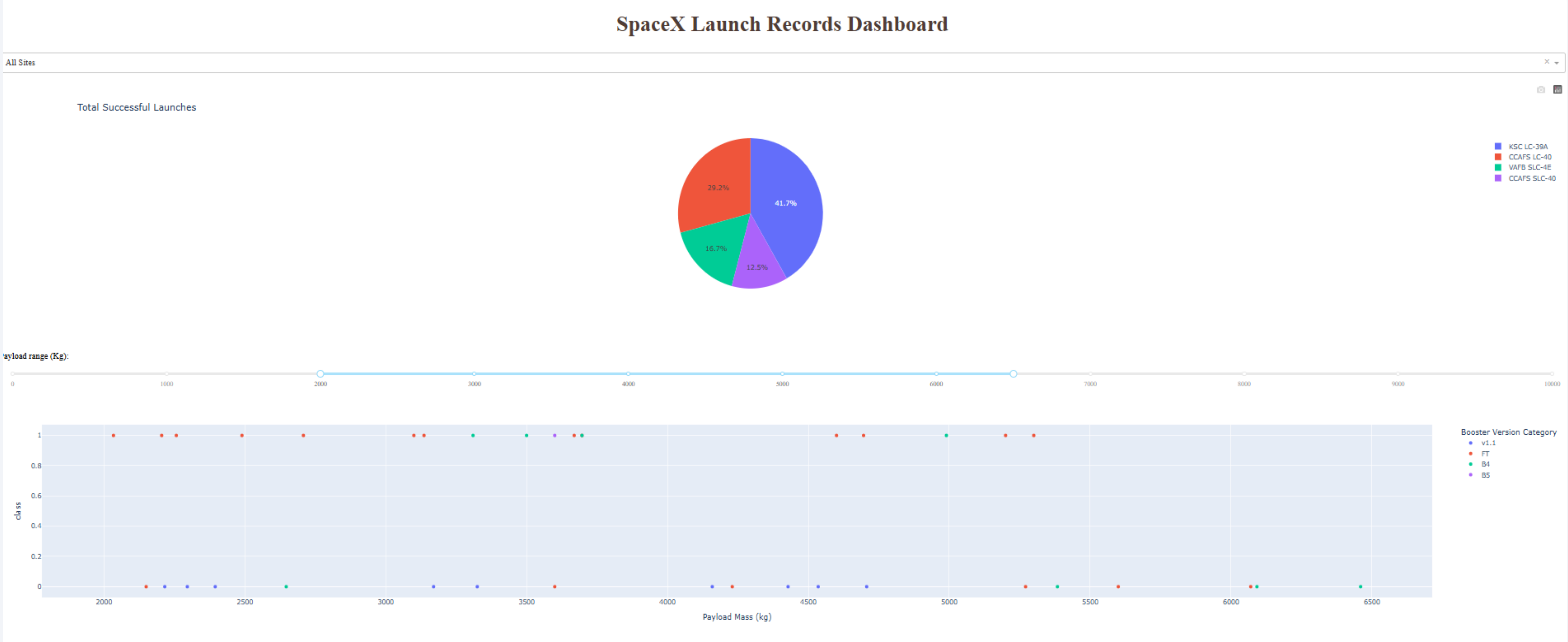
Launch Success Ratio at KSC LC-39A

The dominant segment in the pie chart represents successful launches, indicating that KSC LC-39A has a very high launch success ratio, reinforcing its reliability as a key launch site for SpaceX.

This high success ratio reflects well on KSC LC-39A's performance over time, demonstrating its capability for consistent and successful rocket launches.

The smaller segment representing failures or partial successes shows that KSC LC-39A experiences relatively few issues, reinforcing its role as one of SpaceX's most dependable launch locations.

Payload vs. Launch Outcome:
Success Rates Across Payload Ranges and Booster Versions



SpaceX Launch Records Dashboard

All Sites

ayload range (Kg):

Payload vs. Launch Outcome:
Success Rates Across Payload Ranges and Booster Versions

The scatter plot with the range slider provides a detailed view of how payload mass influences launch success, with heavier payloads generally having a lower success rate.

The booster version also plays a significant role, with newer versions showing better performance, especially for larger payloads.

The findings help to understand how payload size and booster version affect SpaceX's launch outcomes across different sites.

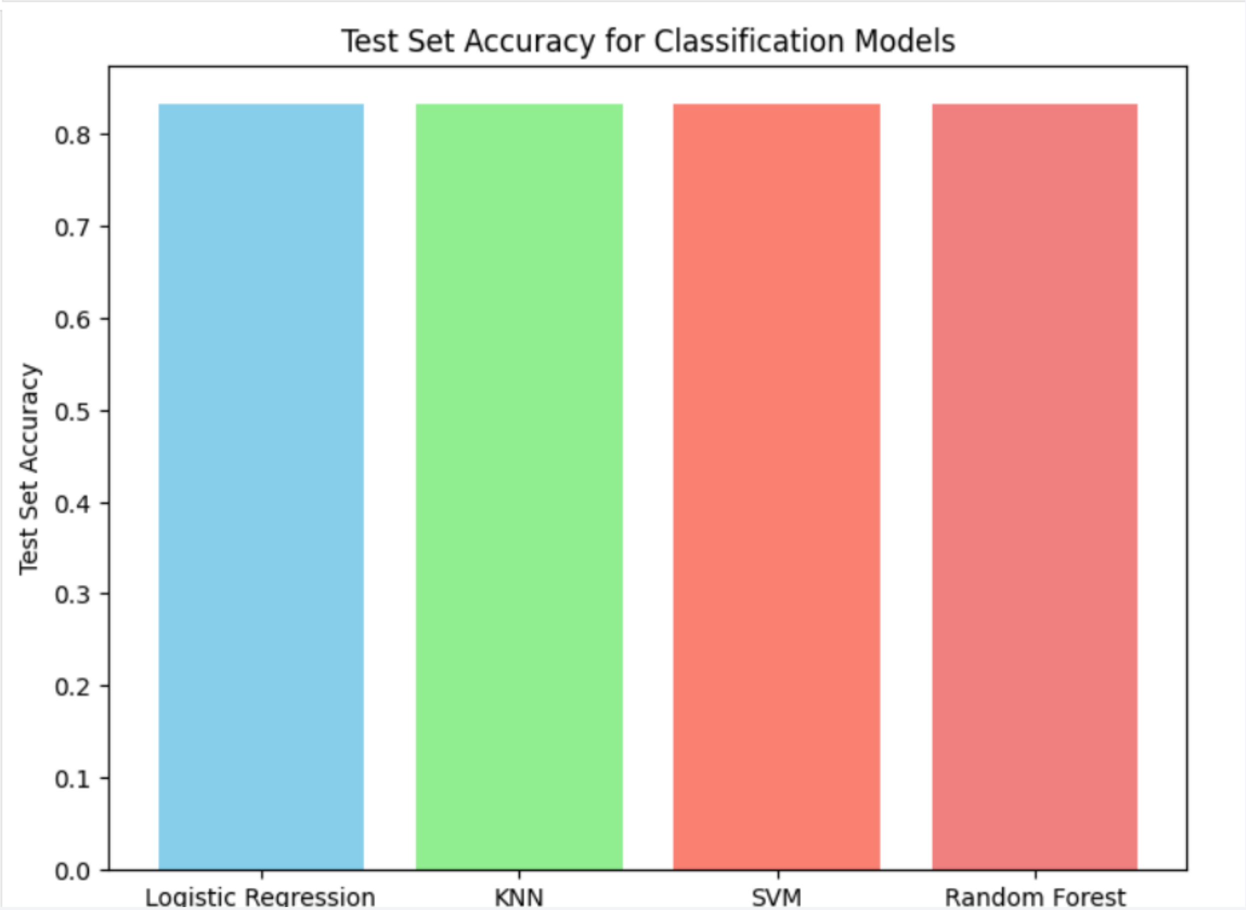
Section 5

Predictive Analysis (Classification)

Classification Accuracy

All four models : Logistic Regression, KNN, SVM, and Random Forest demonstrate similar test set accuracies of approximately 83.33%

This indicates that, for this particular dataset, each model performs almost equally well in predicting the success of the Falcon 9 first stage landing, suggesting that none of the models significantly outperforms the others in terms of accuracy.

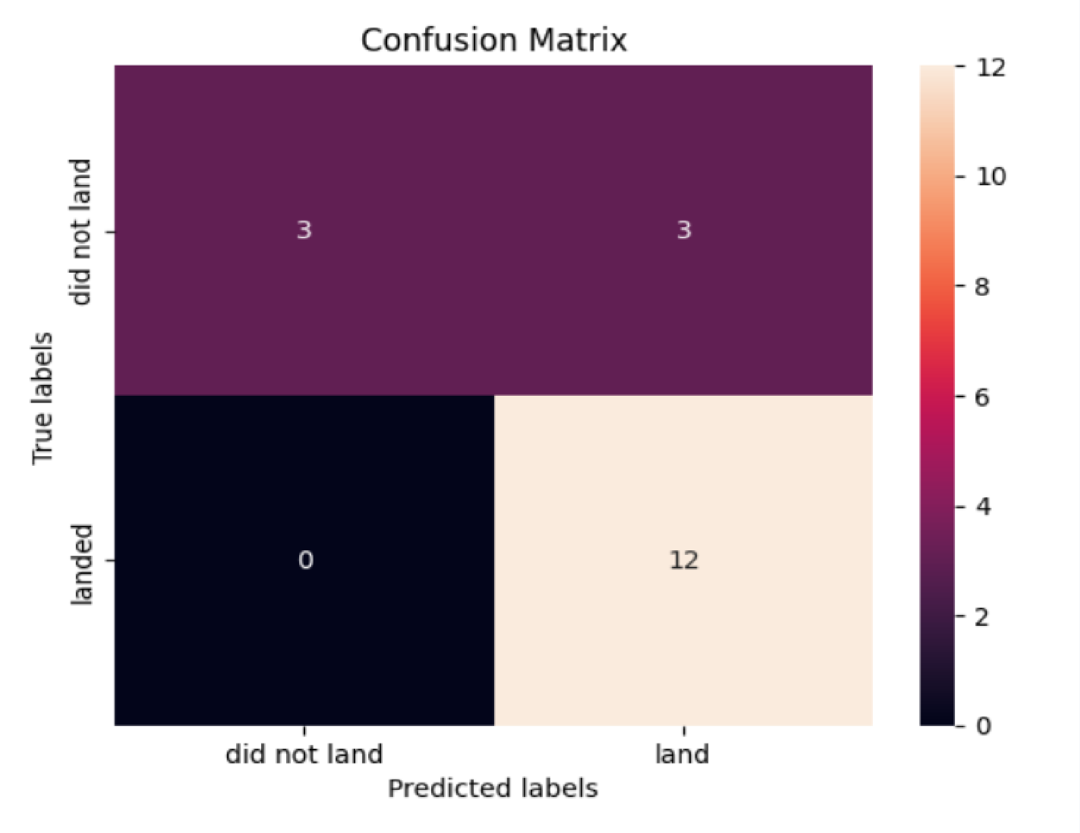


Confusion Matrix

Perfect Recall (100%): The models correctly identifies all successful landings with no false negatives.

Moderate Precision (80%): The models occasionally misclassifies unsuccessful landings as successful (false positives).

Good Accuracy (83.33%): Overall, the models perform well, but reducing false positives would improve its precision.



Conclusions

- **Model Performance:** All models (Logistic Regression, KNN, SVM, Random Forest) achieved similar accuracy (~83.33%), showing no significant performance differences.
- **EDA Insights:**
  - Flight Number and Payload Mass increase the likelihood of successful landings.
  - KSC LC-39A has the highest success rate among sites.
- **Confusion Matrix:** 100% recall and 80% precision, suggesting good performance but room to reduce false positives.
- **Launch Site Proximity:** Sites are near railways, highways, and coastlines, balancing safety and logistics.
- **Visualization Insights:** Scatter plots and pie charts reveal trends in payload and success rates, with certain sites having higher success.
- **Future Recommendations:** Improve precision by reducing false positives and optimize launch site and payload strategies for better success.

Thank you!