IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Héctor W. Pérez Vilcapaza
24 de Agosto 2025

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- **Summary of all results**
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result from Machine Learning Lab

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

This project employs a comprehensive approach to predict the successful landing of the Falcon 9 first stage, incorporating data collection, processing, exploratory analysis, interactive visualizations, and predictive modeling.

Data Collection Methodology:

Data was sourced from the SpaceX API, which provided detailed records of Falcon 9 launches, including launch dates, sites, payloads, and outcomes.

Perform Data Wrangling:

Data cleaning involved handling missing values, standardizing formats, and ensuring consistency. Key features were extracted and new features engineered to enrich the dataset.

Perform Exploratory Data Analysis (EDA) Using Visualization and SQL:

- Visualized launch success rates, payloads, and launch sites using Matplotlib and Seaborn.
- Executed SQL queries to derive insights and answer specific questions regarding the dataset

Methodology

Perform Interactive Visual Analytics Using Folium and Plotly Dash:

- Used Folium to create interactive maps displaying launch sites and outcomes.
- Developed a Plotly Dash application with interactive components like dropdowns and sliders to analyze launch success rates and payload ranges.

Perform Predictive Analysis Using Classification Models:

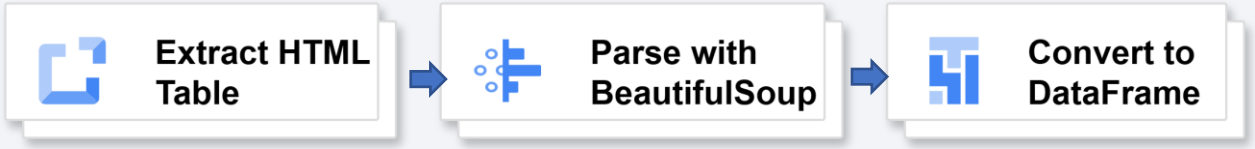
- Built and evaluated various classification models including Logistic Regression, SVM, KNN, and Decision Trees.
- Employed GridSearchCV for hyperparameter tuning.
- Evaluated models based on accuracy, and identified the best performing model for predicting landing success.

Github URL: https://github.com/hperezv/IBM_Data_Science_Professional_Certificate.git

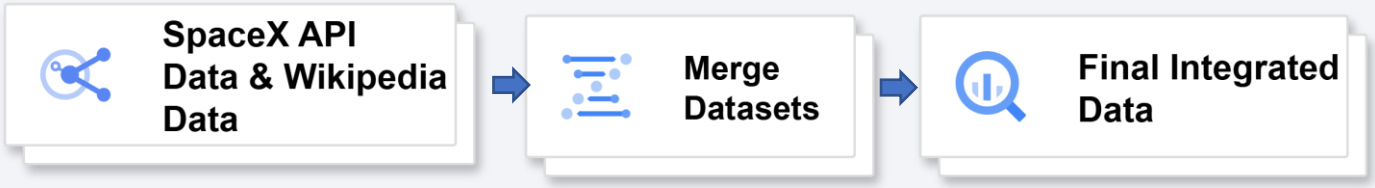
Data Collection

Step 1: SpaceX API Request

Step 2: Web Scraping Wikpedia



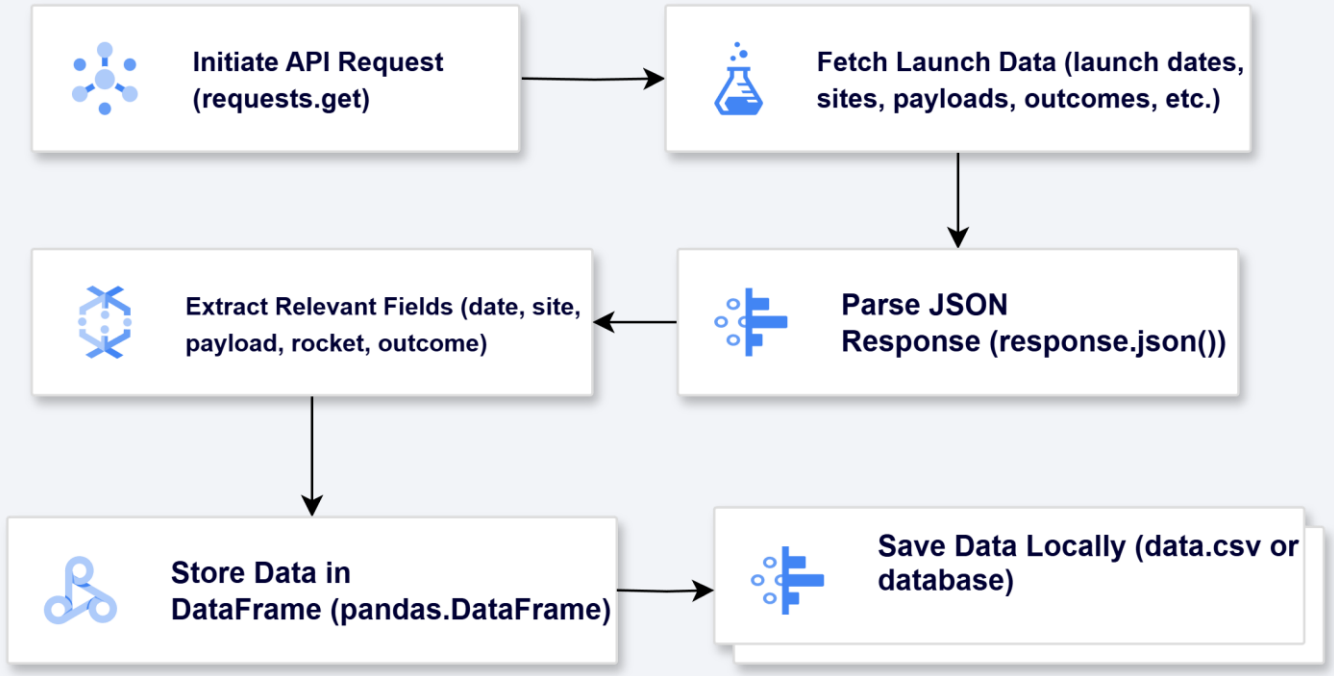
Step 3: Data integration



Data Collection – SpaceX API

- **Step 1: Initiate API Request**
  - Use Python's `requests` library to connect to the SpaceX API.
  - Endpoint: `https://api.spacexdata.com/v4/launches`
- **Step 2: Parse API Response**
  - Convert API response from JSON to a Python dictionary.
  - Extract relevant fields: launch date, launch site, payload mass, rocket type, outcome.
- **Step 3: Store Data Locally**
  - Save extracted data into a pandas DataFrame.
  - Store the DataFrame locally for further processing.

- GitHub URL:

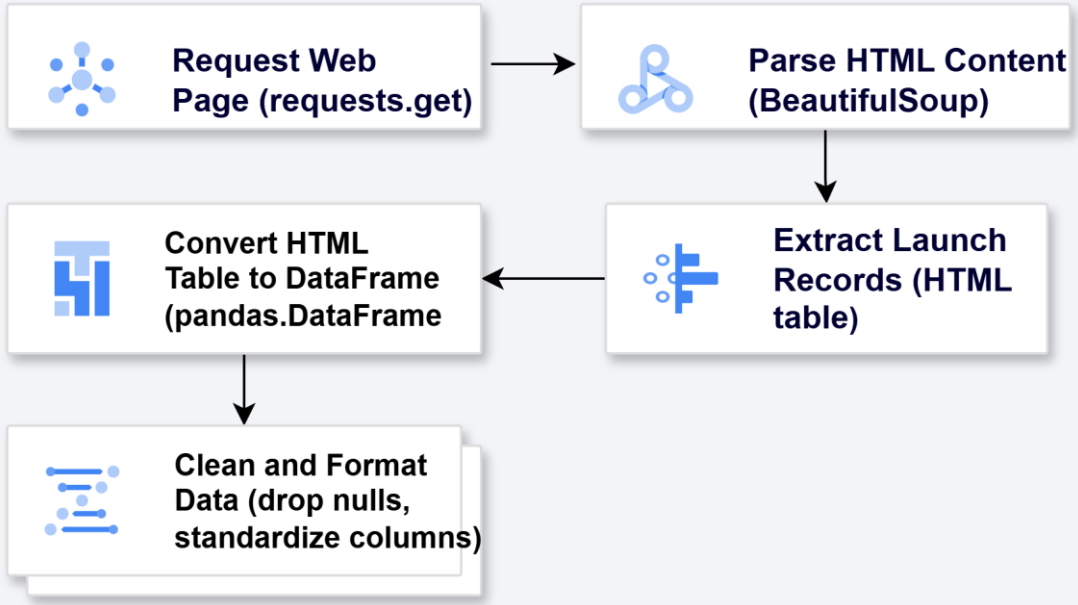


Data Collection - Scraping

- **Step 1: Initiate Web Scraping**
  - Use Python's `requests` library to fetch the HTML content of the Wikipedia page.
  - Target URL: `https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches`
- **Step 2: Parse HTML Content**
  - Use `BeautifulSoup` to parse the HTML content.

Extract the HTML table containing Falcon 9 launch records.

- **Step 3: Convert to DataFrame**
  - Convert the extracted HTML table into a pandas DataFrame.
  - Clean and format the DataFrame, ensuring data consistency.



Data Wrangling

Overview: Data wrangling involves cleaning, transforming, and organizing raw data into a structured format suitable for analysis.

- **Step 1: Data Cleaning**
- Identify and fill or remove missing values in the dataset.
- Use appropriate imputation techniques or drop rows/columns with excessive missing data.
- **Step 2: Data Transformation**
- Convert data types to appropriate formats (e.g., date-time, numerical).
- Standardize text (e.g., lowercase, remove whitespace).
- Create new features from existing data (e.g., extract year from date).
- Normalize/scale numerical features to ensure consistency

Data Wrangling

Step 3: Data Integration

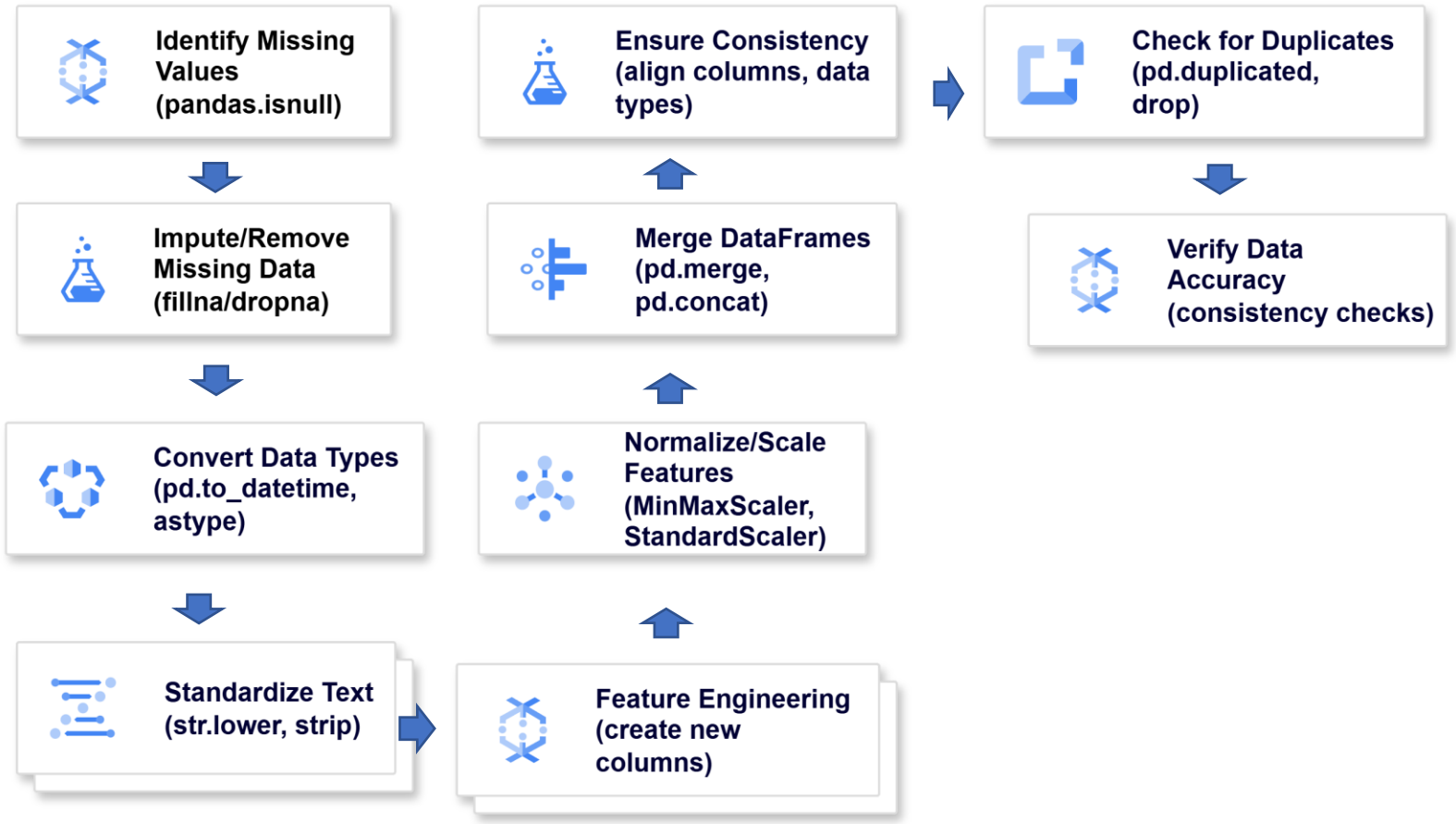
- Merge datasets collected from different sources (API, web scraping) into a single cohesive dataset.
- Ensure consistent column names and data formats across datasets.

Step 4: Data Validation

- Check for duplicate records and remove them.
- Verify the accuracy and consistency of data entries.

GitHub URL: 3. labs-jupyter-spacex-Data wrangling.ipynb

Data Wrangling flow Chart



EDA with Data Visualization

Overview:

- Exploratory Data Analysis (EDA) involves visually exploring and summarizing the main characteristics of a dataset. The goal is to understand the data's distribution, identify patterns, and uncover relationships between variables.

Charts Plotted:

1. Histograms:

- Purpose: Used to visualize the distribution of numerical variables such as launch success rates, payload mass, and flight number.
- Why: Helps in understanding the spread and central tendency of the data, identifying outliers, and assessing data skewness.

2. Bar Charts:

- Purpose: Used to compare categorical variables such as launch outcomes (success/failure) across different categories like launch sites or rocket types.
- Why: Provides a clear comparison of frequencies or proportions within categorical data, highlighting patterns or trends.

EDA with Data Visualization

3. Line Charts:

- Purpose: Used to track trends over time, such as the success rate of Falcon 9 launches across different years.
- Why: Reveals temporal patterns and helps in understanding performance trends or changes over specific periods.

4. Scatter Plots:

- Purpose: Used to explore relationships between two numerical variables, such as payload mass vs. launch success.
- Why: Identifies correlations or dependencies between variables, visualizing how one variable changes concerning another.

5. Heatmaps:

- Purpose: Used to visualize correlation matrices between multiple numerical variables.
- Why: Helps in identifying strong correlations (positive or negative) between variables, aiding feature selection or understanding multicollinearity.

6.Box Plots:

- Purpose: Used to display the distribution of numerical data through their quartiles.
- Why: Visualizes the spread and skewness of data, highlighting outliers and comparing distributions across different categories.
- Github URL:5. jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Aggregate Queries:

- Calculated total number of launches.
- Counted successful and failed launches.
- Calculated success rates by launch site and rocket type.

Join Queries:

- Joined tables to link launch records with additional data (e.g., rocket details).
- Combined datasets for comprehensive analysis.

Filtering Queries:

- Filtered data to focus on specific launch outcomes (success/failure).
- Applied conditions to extract launches based on criteria like launch date or rocket configuration.

Sorting Queries:

- Sorted data to identify trends or outliers.
- Ordered launches by date or success rate for analysis.
- **Subqueries:**
- Nested queries to calculate derived metrics (e.g., average payload mass per launch site).
- Subqueries used to perform detailed analysis within larger datasets.
- GitHub URL:4. jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Map Objects Created

Markers:

- Placed markers to indicate launch sites on the map.
- Each marker represents a specific geographical location where SpaceX launches have occurred.
- **Circles:**
- Added circles around launch sites to visually represent proximity zones.
- Circles help visualize the areas around launch sites that might influence operational decisions.

Lines:

- Drew lines to connect launch sites with their proximities or other relevant locations.
- Lines provide spatial context and connections between different points of interest related to launches.
- Github URL: 6. lab_jupyter_launch_site_location.ipynb

Reasons for Adding Objects

Markers:

- To pinpoint exact launch locations for spatial reference.
- Helps users identify where SpaceX has conducted launches geographically.

Circles:

- Illustrates the potential impact zones around launch sites.
- Provides a visual representation of safety perimeters or operational boundaries.

Lines:

- Shows connections or relationships between launch sites and relevant features.
- Enhances understanding of spatial relationships and dependencies.

Build a Dashboard with Plotly Dash

Plots/Graphs Added

Success Pie Chart:

- Displays the distribution of successful and failed launches.
- Helps visualize the overall success rate and performance trends.

Success-Payload Scatter Plot:

- Shows the relationship between payload mass and launch success.
- Allows users to explore how payload mass influences mission outcomes.

- Github URL:spacex_dash_app.py

Interactions Added

Launch Site Dropdown:

- Enables users to select specific launch sites for analysis.
- Facilitates filtering and focused exploration based on geographical locations.

Range Slider for Payload:

- Allows users to adjust payload mass ranges dynamically.
- Offers flexibility in examining launch success concerning payload mass variations.

Predictive Analysis (Classification)

1. Data Preprocessing:

- Standardized features to ensure all variables contribute equally.
- Split data into training and test sets for model validation.

2. Model Selection:

- Explored multiple classification algorithms: SVM, Decision Trees, and K-Nearest Neighbors (KNN).
- Chose algorithms suitable for binary classification tasks based on project requirements.

3. Hyperparameter Tuning:

- Used GridSearchCV to systematically search for optimal hyperparameters.
- Tuned parameters such as C (SVM), max_depth (Decision Trees), and n_neighbors (KNN).
- Github URL:7. SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

4. Model Evaluation:

- Evaluated models using cross-validation techniques to ensure robustness and generalizability.
- Utilized metrics like accuracy, precision, recall, and F1-score to assess model performance.

5. Improvement Iterations:

- Iteratively adjusted models based on insights from validation results.
- Fine-tuned hyperparameters to maximize predictive accuracy and reliability.

6. **Selection of Best Performing Model:**

- Identified the model with the highest accuracy on the test set as the best performer.
- Considered both training and test set performance to avoid overfitting and ensure real-world applicability

Reasons for Adding Plots and Interactions

Success Pie Chart:

- Provides a quick overview of mission success rates.
- Essential for stakeholders to understand overall performance metrics at a glance.

Success-Payload Scatter Plot:

- Helps identify correlations between payload characteristics and launch outcomes.
- Supports decision-making processes related to payload planning and operational strategies.

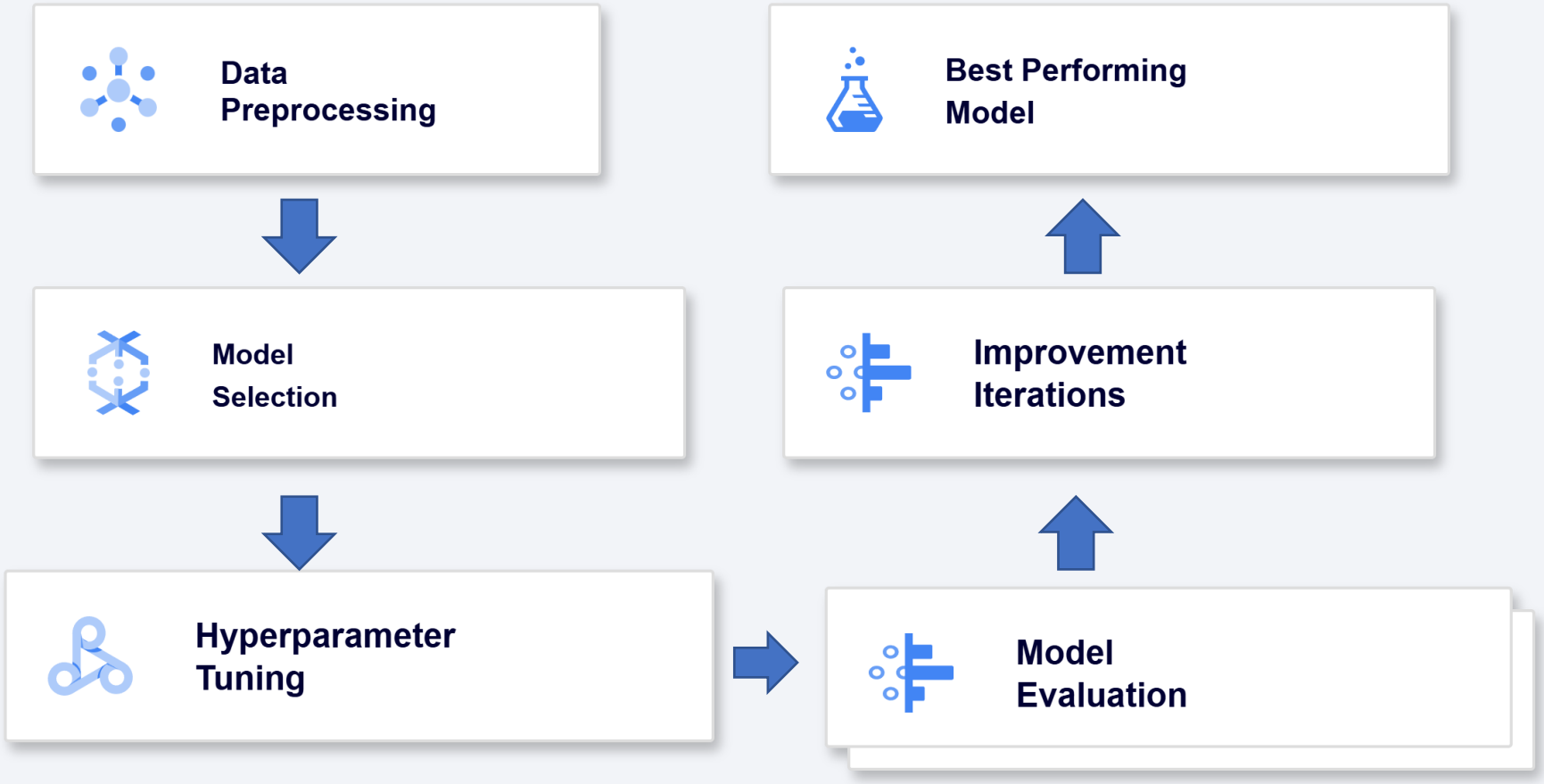
Launch Site Dropdown:

- Enhances user experience by focusing analysis on specific launch locations.
- Allows for regional insights and comparisons across different launch sites.

Range Slider for Payload:

- Offers interactive exploration of how payload mass affects mission success.
- Enables detailed analysis and insights into payload-related performance factors

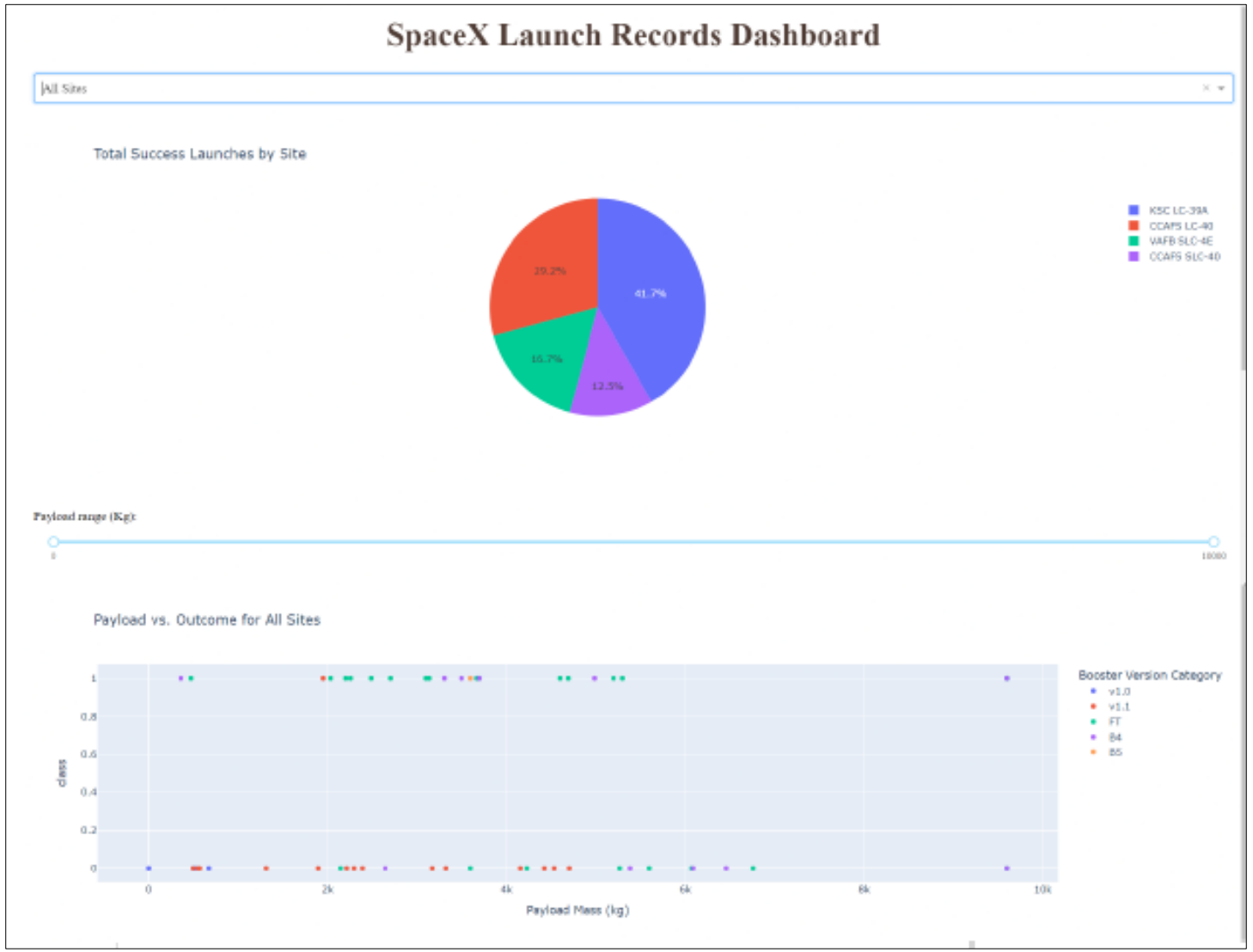
Predictive Analysis (Flowchart)

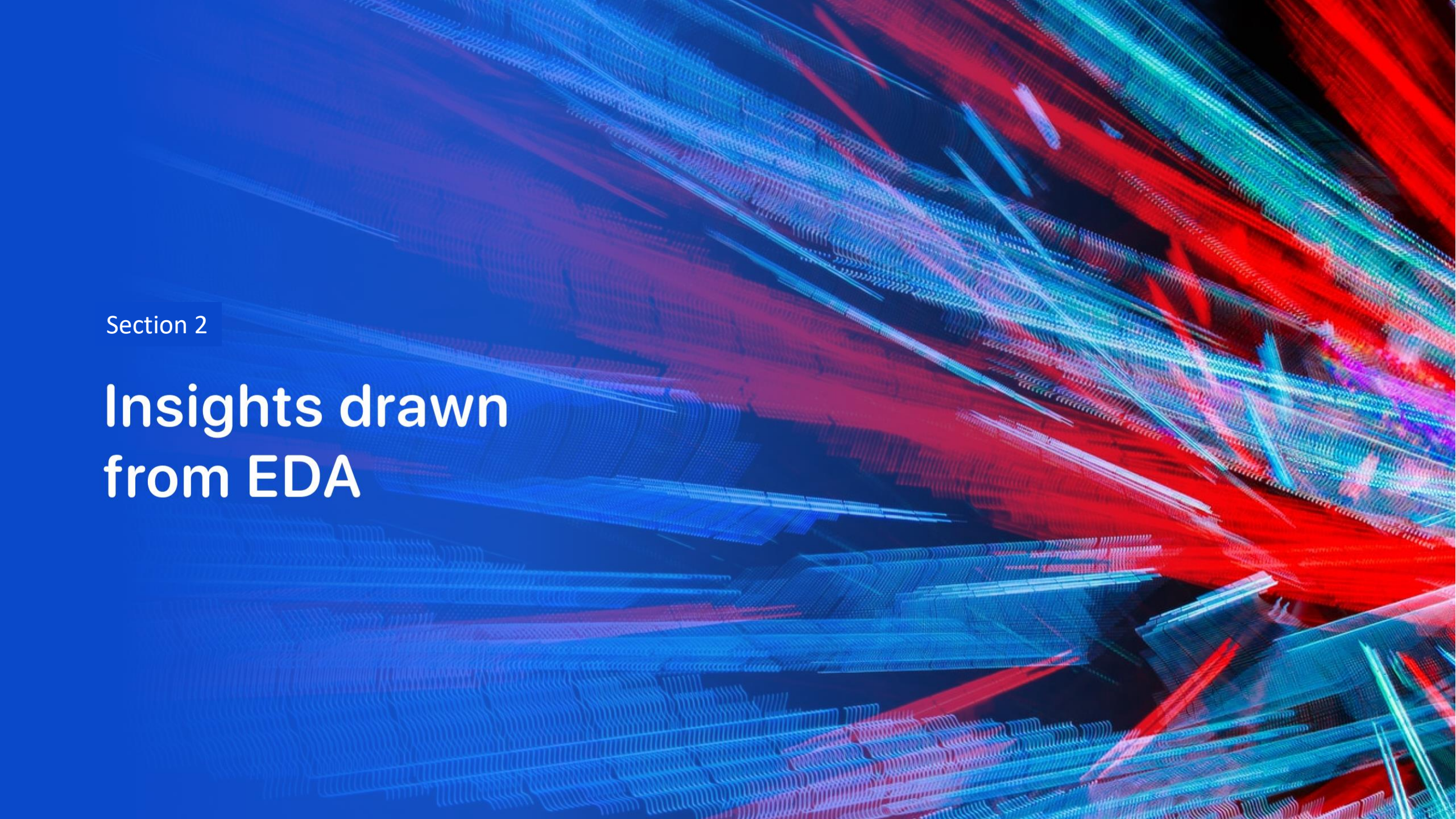


Results

This dashboard provides a clear snapshot of SpaceX's launch operations, highlighting their strategic use of multiple launch sites to support their ambitious launch manifest while maintaining high success rates across all locations

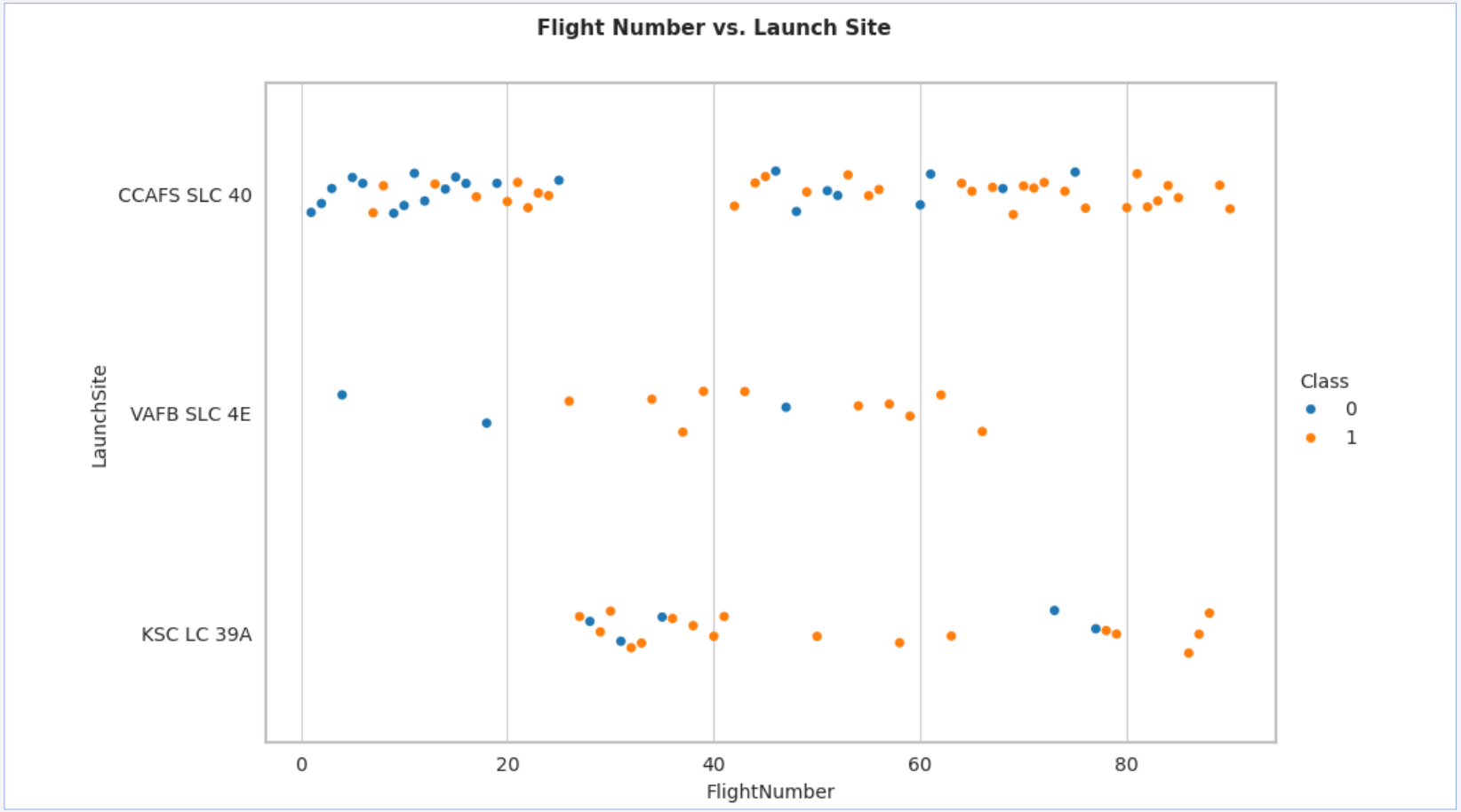
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results




Section 2
Insights drawn from EDA

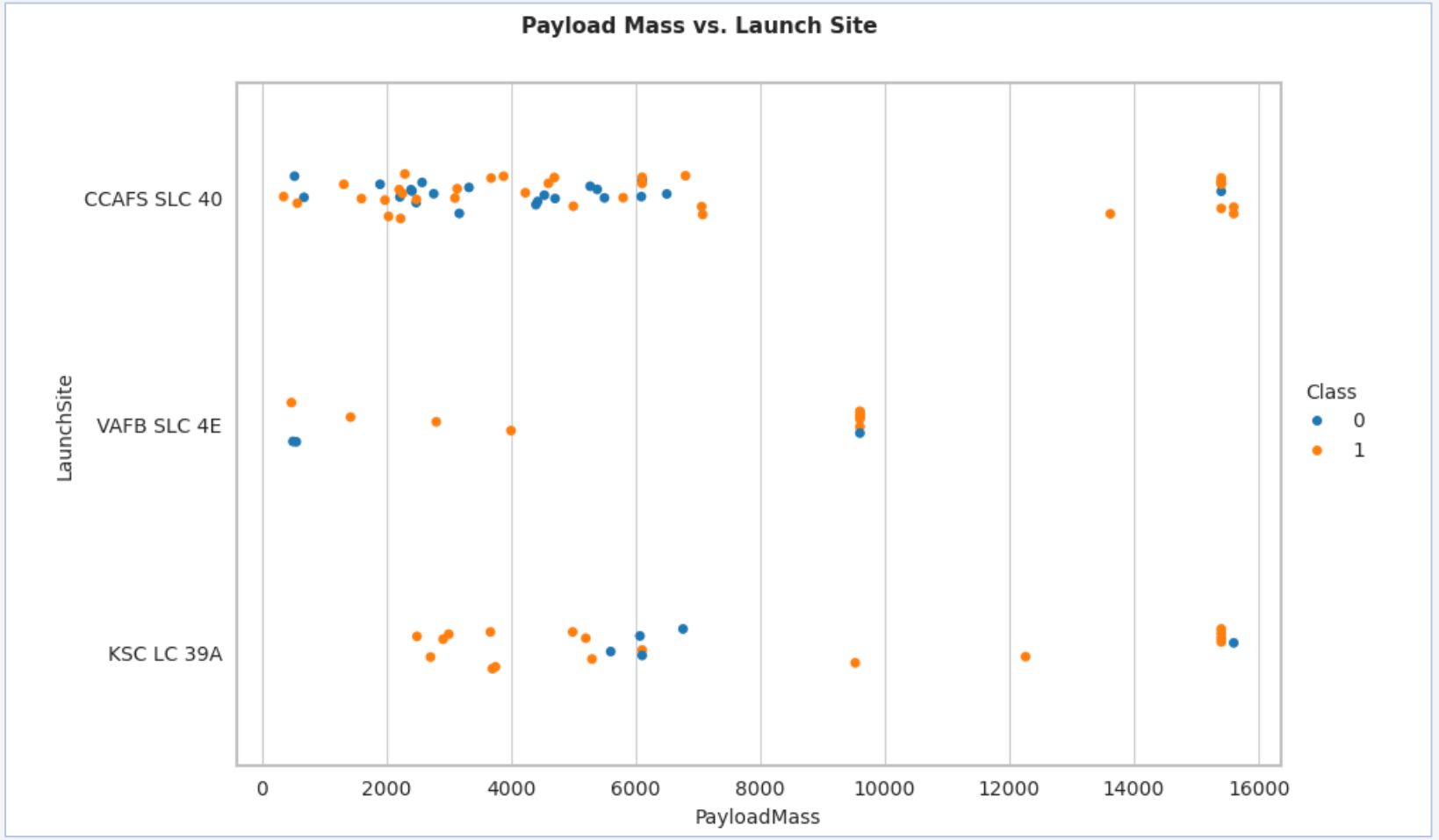
Flight Number vs. Launch Site

- **Mixed Outcomes at Major Launch Sites**: Both CCAFS SLC 40 and KSC LC 39A have a mix of successful (orange) and unsuccessful (blue) landings, indicating that factors other than the launch site itself may influence the landing success.
- **Consistent Activity Across Flight Numbers**: Launches are spread across a wide range of flight numbers at all sites, suggesting consistent activity over time without a clear trend of increasing or decreasing landing success.



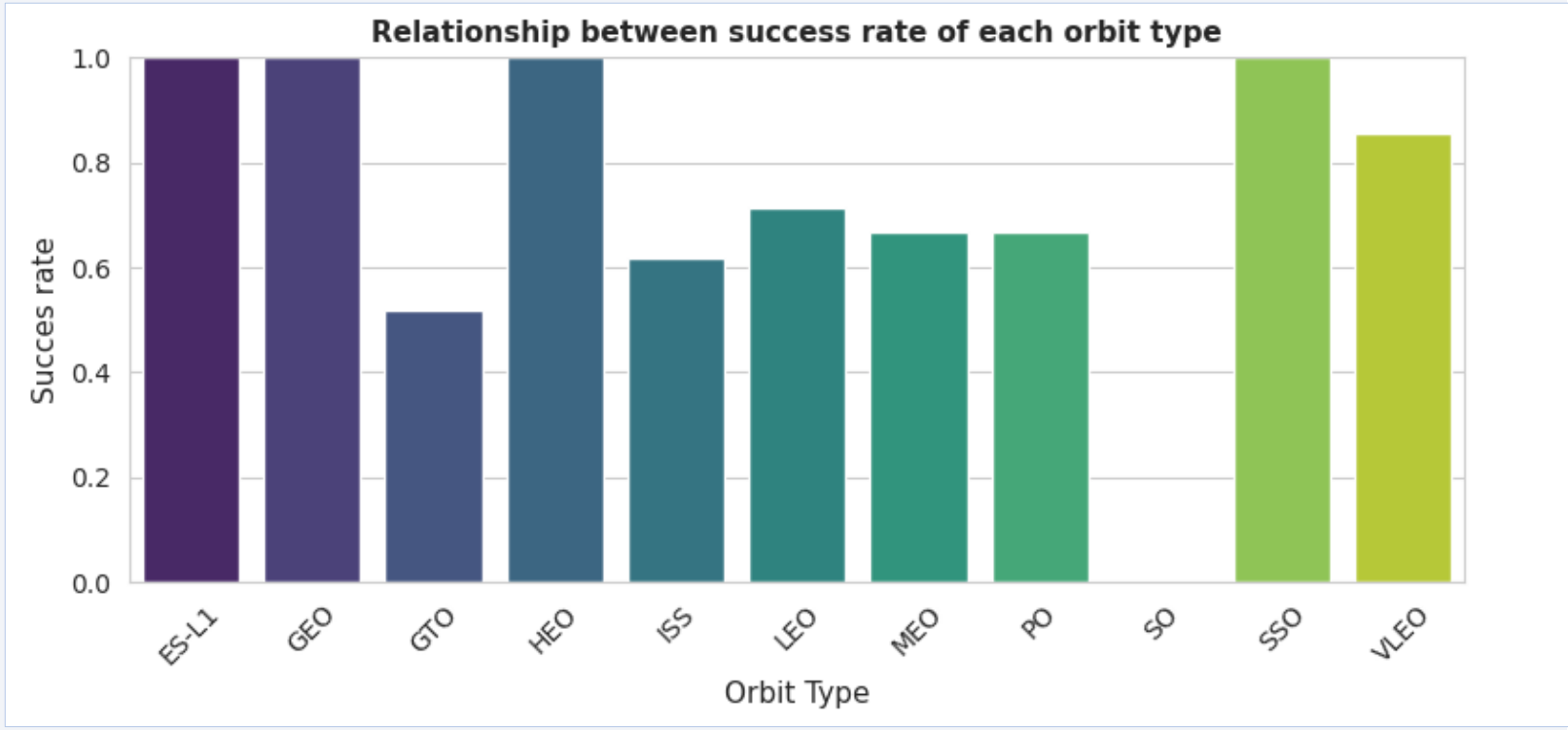
Payload vs. Launch Site

- Payload Distribution: Most launches from the CCAFS SLC 40 site handle payloads below 10,000 kg, while the VAFB SLC 4E and KSC LC 39A sites have a wider range of payload masses, indicating varied mission profiles.
- High-Capacity Launches: The KSC LC 39A site is frequently used for launching heavier payloads, with multiple launches carrying over 15,000 kg, suggesting its suitability for high-capacity missions.



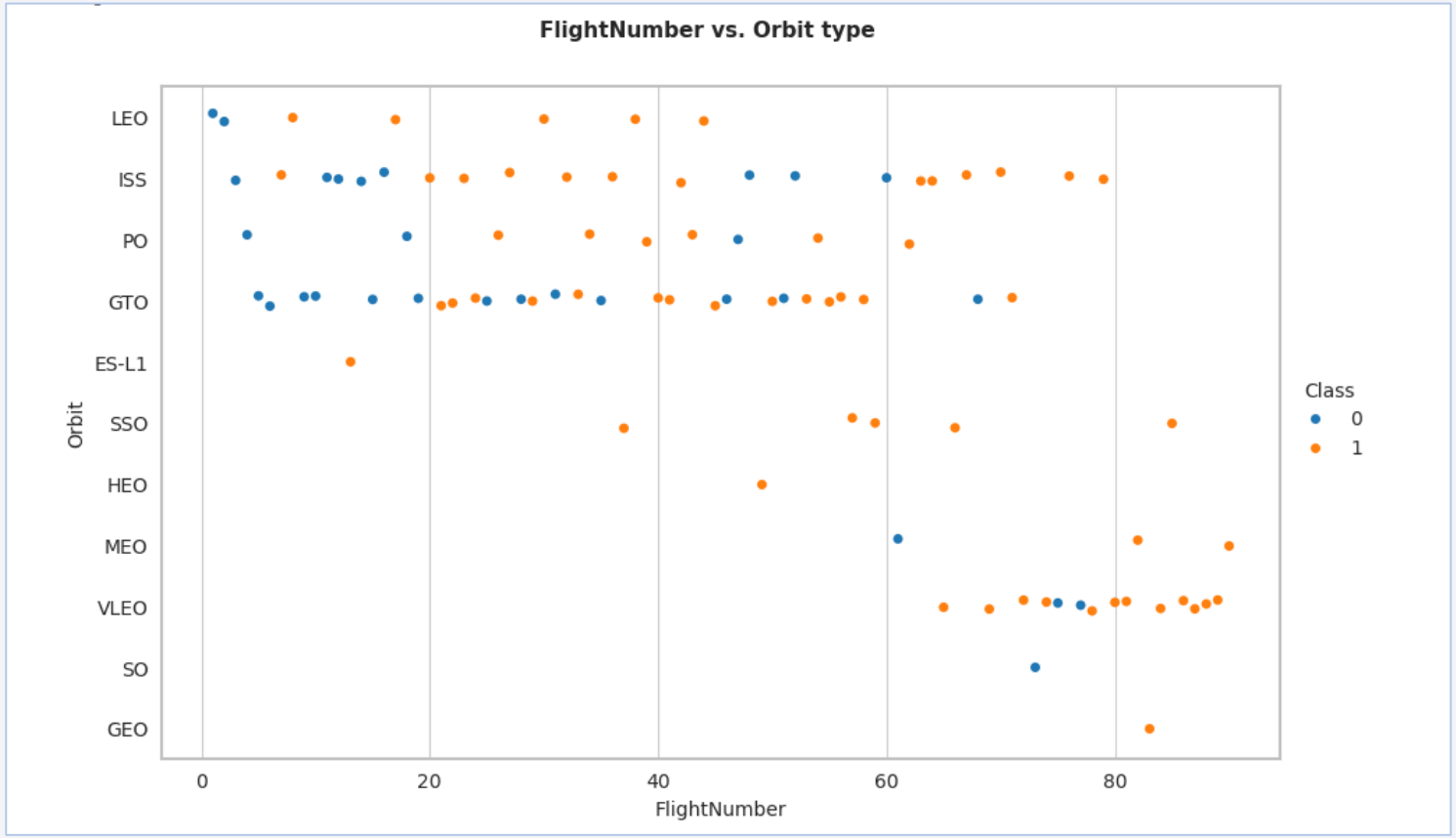
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
- ES L1;
- GEO;
- HEO; and
- SSO.
- Followed by:
- VLEO (above 80%); and
- LFO (above 70%).



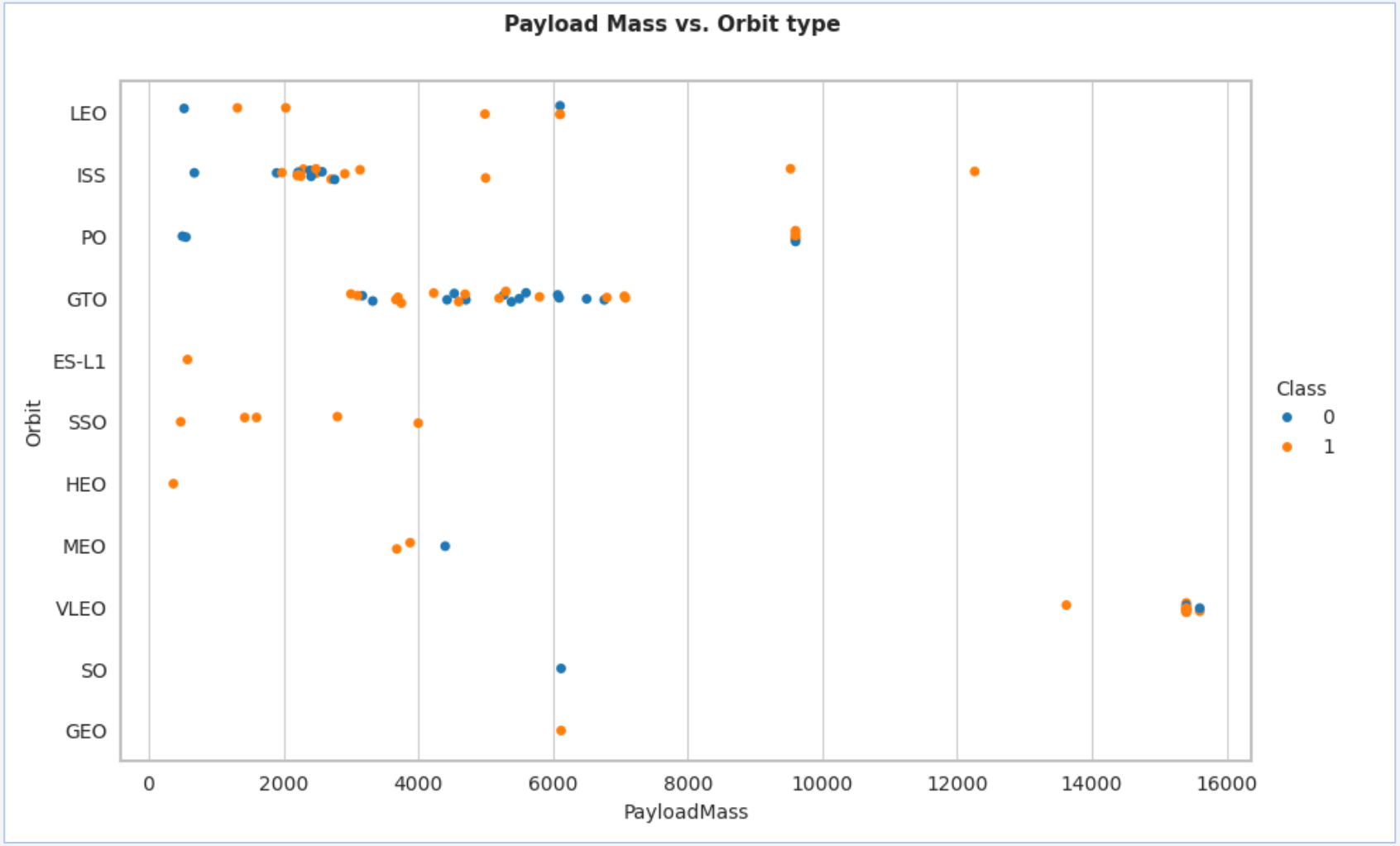
Flight Number vs. Orbit Type

- Increased Success Over Time: The success rate of Falcon 9 launches improves significantly with higher flight numbers, indicating that experience and iterative improvements contribute to better outcomes.
- Orbit-Specific Performance: Early flights to GTO and ISS orbits had mixed outcomes, but recent missions to these orbits show a higher success rate, reflecting advancements in mission planning and execution.



Payload vs. Orbit Type

- Successful landings are more frequent across all orbit types, especially for payloads less than 6000 kg.
- Higher payload masses (above 10,000 kg) show a mix of successes and failures, indicating increased difficulty with heavier payloads



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020
- It seems that the first three years were a period of adjusts and improvement of technology.

All Launch Site Names

- According to data, there are four launch sites:
- They are obtained by selecting unique occurrences of “ launch_site ” values from the dataset.

```
%%sql
select distinct Launch_Site from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- The list show 5 records where launch sites begin with `CCA`
- The query returned the first 5 launches from Cape Canaveral Air Force Station (CCAFS) Launch Complex 40, showing SpaceX's early operational history at this site

```
%%sql
select * from SPACEXTBL
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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- This SQL query calculates the sum of all payload mass (in kilograms) that SpaceX has launched to date, as recorded in their database. The result of 619,967 kg represents:

```
[10]: %%sql
      select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL
       * sqlite:///my_data1.db
      Done.
[10]: payloadmass
      619967
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- This SQL query calculates the average payload mass carried by SpaceX's Falcon 9 v1.1 rocket variant. The result of approximately 2,535 kg reveals:

```
%%sql
select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version like 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- This query identifies the historic date when SpaceX successfully landed a Falcon 9 rocket on a ground pad for the first time.

```
%%sql
select min(SPACEXTBL.Date) from SPACEXTBL
where Landing_Outcome="Success (ground pad)"
```

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

min(SPACEXTBL.Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- This query identifies specific Falcon 9 booster versions that successfully landed on drone ships while carrying medium-to-heavy payloads between 4,000-6,000 kg.

```
%%sql
select Booster_Version from SPACEXTBL
where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- This query provides a comprehensive overview of SpaceX's mission success rate by counting all mission outcomes in their database.

```
%%sql
select Mission_Outcome, COUNT(*) from SPACEXTBL
group by 1
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- This query identifies all Falcon 9 Block 5 booster versions that carried the maximum payload mass recorded in SpaceX's launch history.

```
%%sql
select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- This query identifies Falcon 9 boosters that failed to land successfully on drone ships during the year 2015

```
%%sql
select substr(Date, 4, 2) as "month", Booster_Version, Launch_Site from SPACEXTBL
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

month	Booster_Version	Launch_Site
5-	F9 v1.1 B1012	CCAFS LC-40
5-	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- This query shows the successful landing outcomes during SpaceX's early development phase of reusable rocket technology (June 2010 - March 2017).

```
%%sql
select Landing_Outcome, COUNT(Landing_Outcome) from SPACEXTBL
where Landing_Outcome like 'Success%' and SPACEXTBL.Date between '2010/06/04' AND '2017/03/20'
group by 1
order by 2 DESC
```

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

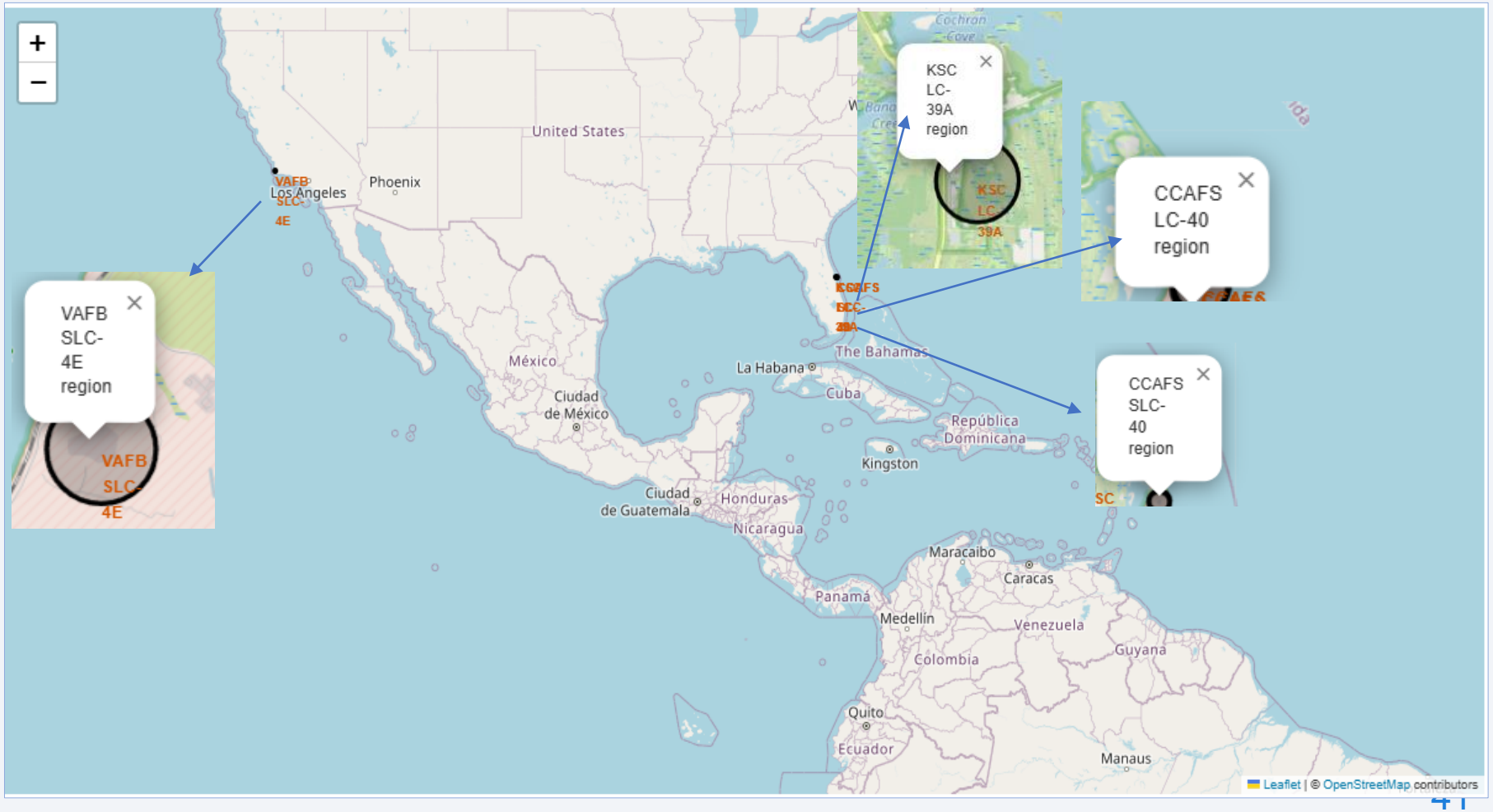
Landing_Outcome	COUNT(Landing_Outcome)
Success (drone ship)	12
Success (ground pad)	8

Section 3

Launch Sites Proximities Analysis

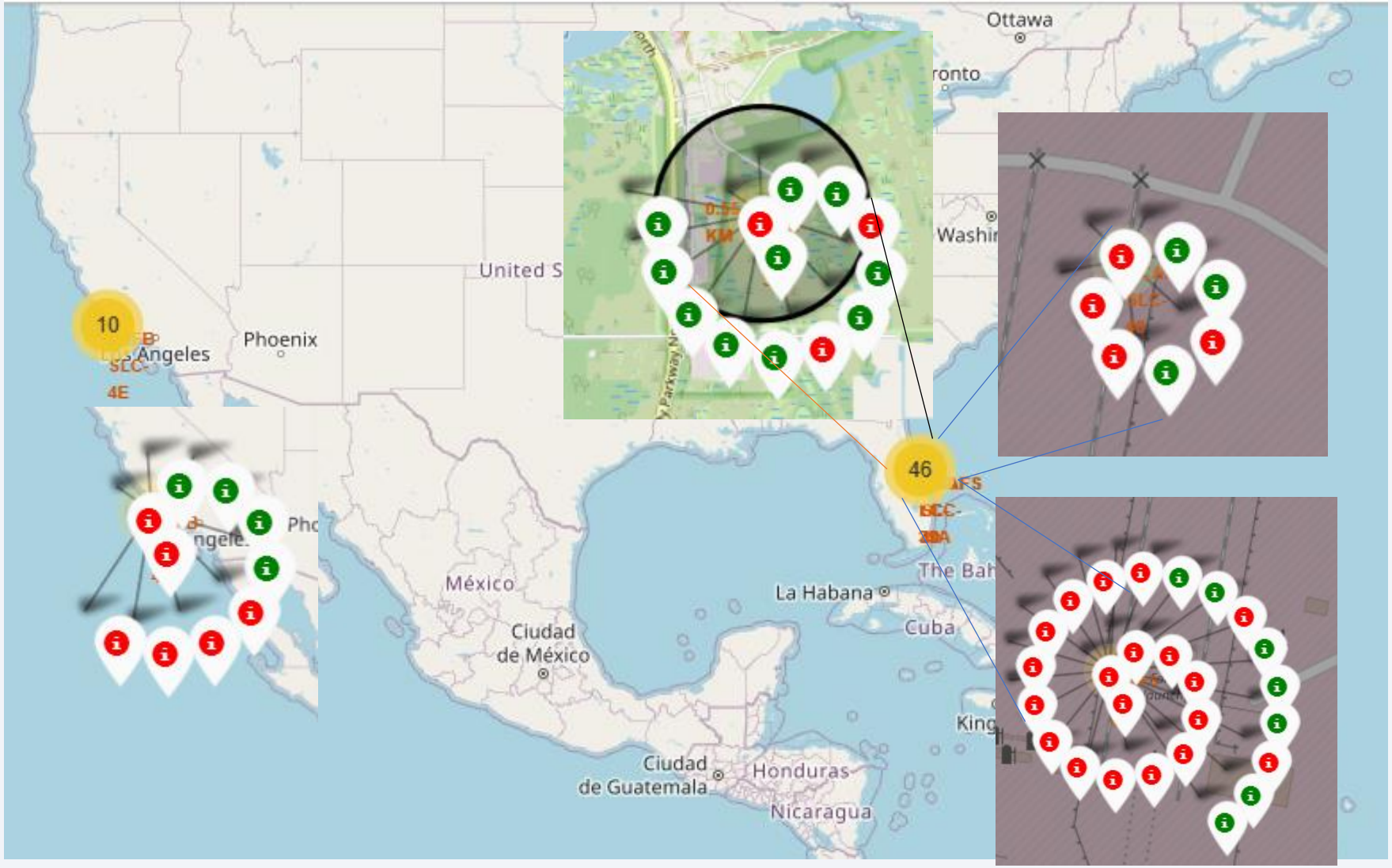
All launch sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads. The Cape Canaveral sites (CCAFS LC-40 and CCAFS SLC-40) and Kennedy Space Center (KSC LC-39A) are near the coast in Florida.
- Vandenberg Air Force Base (VAFB SLC-4E) is also near the coast in California.

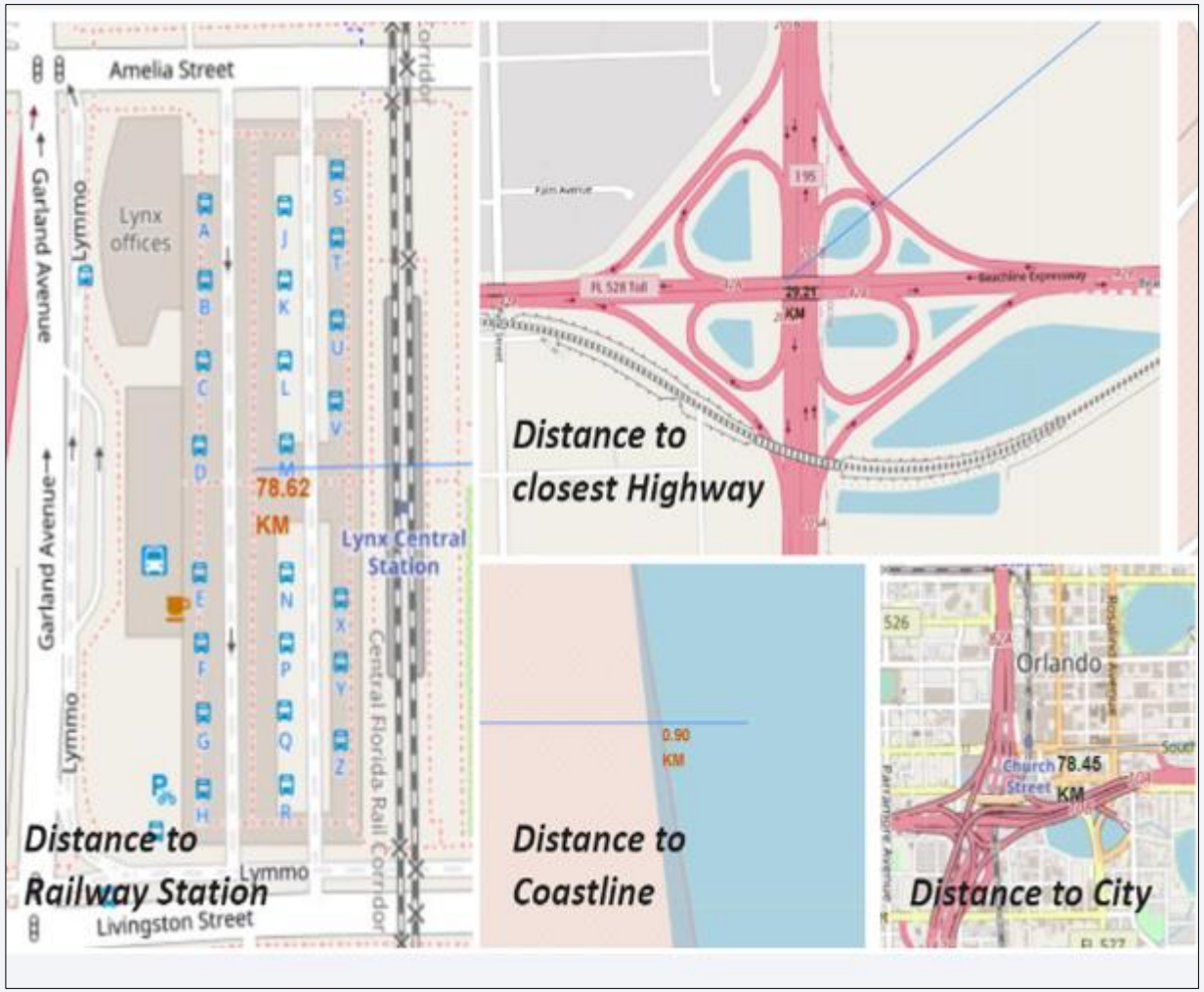


Launch Outcomes by Site

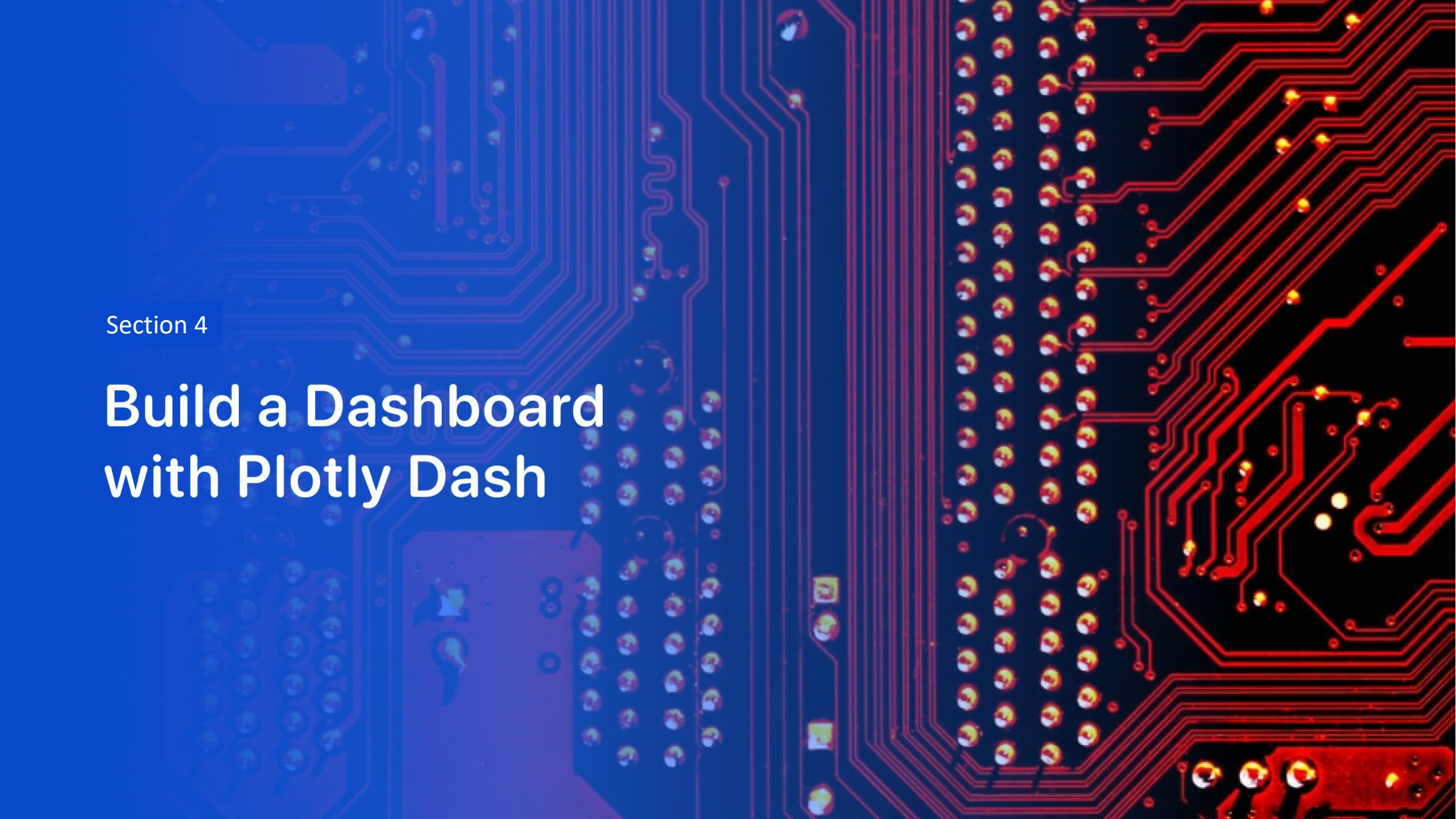
- Example of KSC LC 39A launch site launch outcomes
- Green markers indicate successful and red ones indicate failure.



Launch Site distance to landmarks

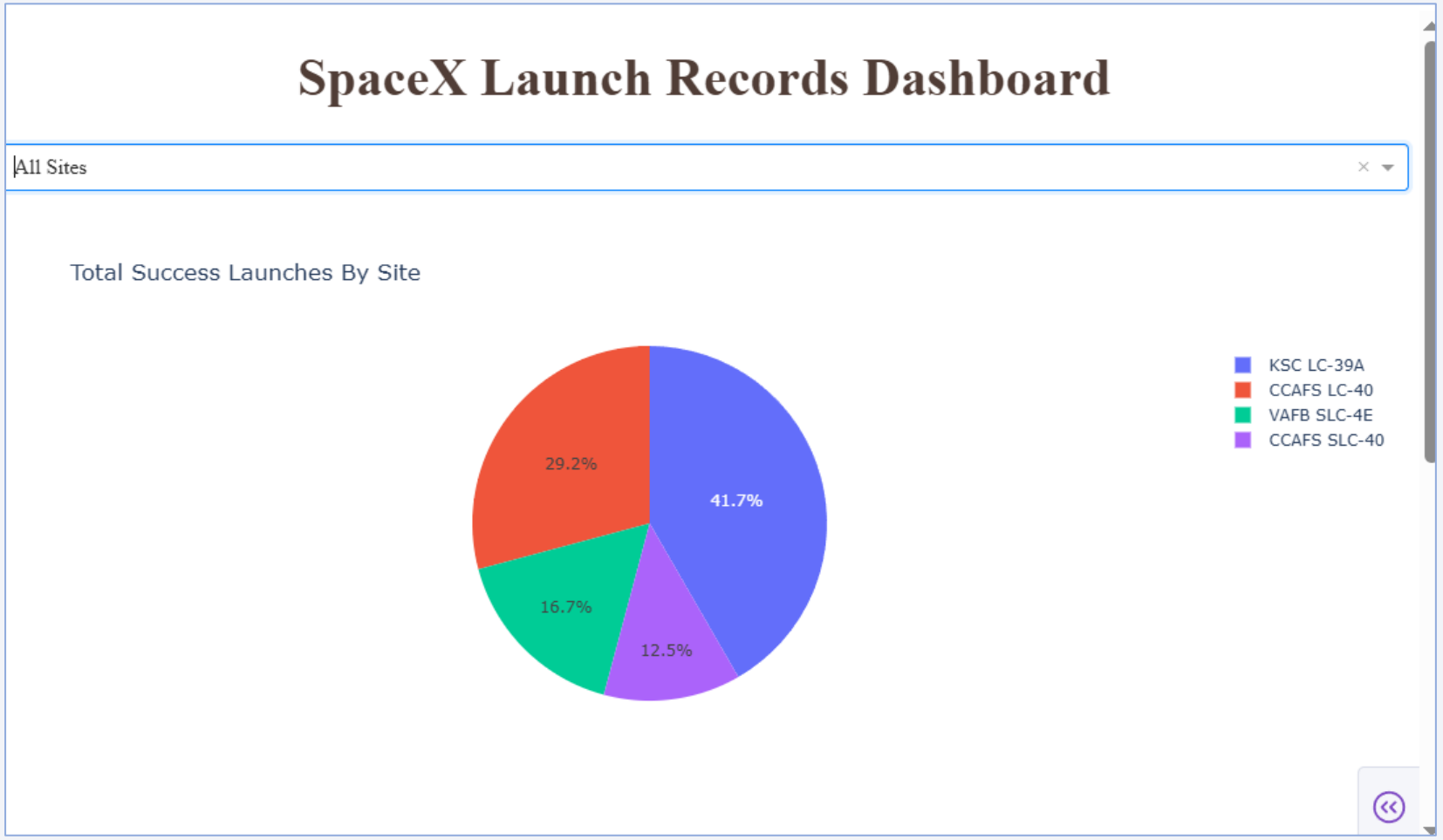


- This plot provides a visual representation of the distance between the CCAFS SLC-40 launch site and the closest coastline. The calculated distance is approximately 0.87 kilometers, as indicated by the marker. The added PolyLine clearly shows the straight-line distance, highlighting the proximity of the launch site to the coast. This close proximity to the coastline is typical for launch sites to facilitate over-water flight paths and safe recovery operations, ensuring minimal risk to populated areas.


Section 4
Build a Dashboard with Plotly Dash

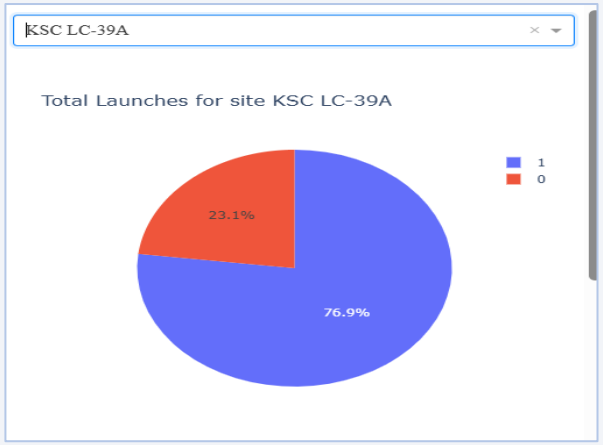
Successful Launches by Site

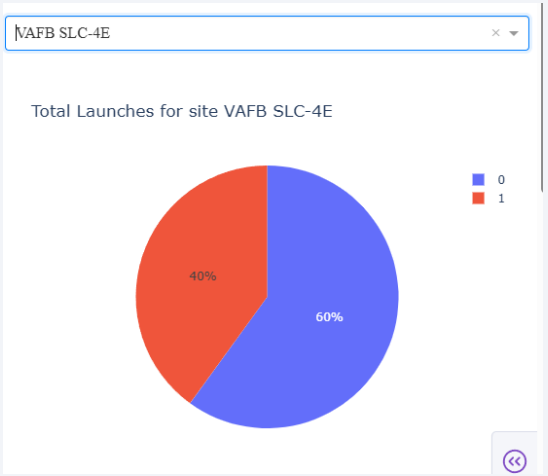
- He dashboard shows the distribution of successful launches across SpaceX's four main launch sites:
- KSC LC-39A: 29.2% of successful launches
- CCAFS LC-40: 41.7% of successful launches
- VAFB SLC-4E: 16.7% of successful launches
- CCAFS SLC-40: 12.5% of successful launches

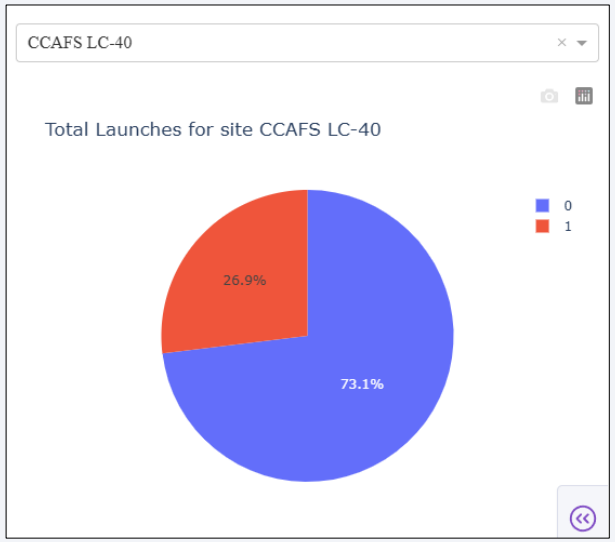


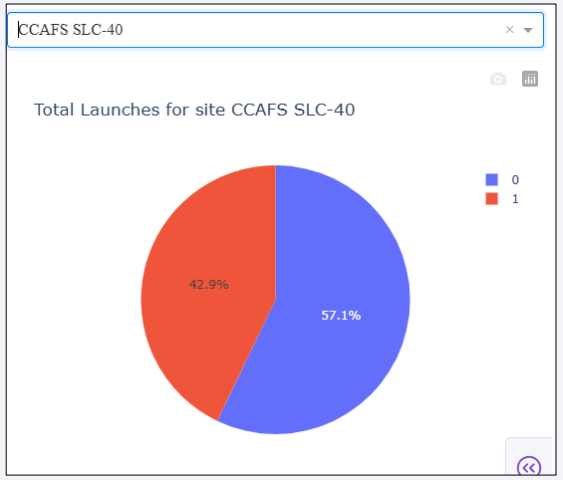
Launch Success Ratio Performance Analysis

- This dashboard highlights SpaceX's impressive launch reliability across multiple launch sites, with all facilities maintaining success rates well above industry averages, demonstrating their operational excellence and consistent performance regardless of location



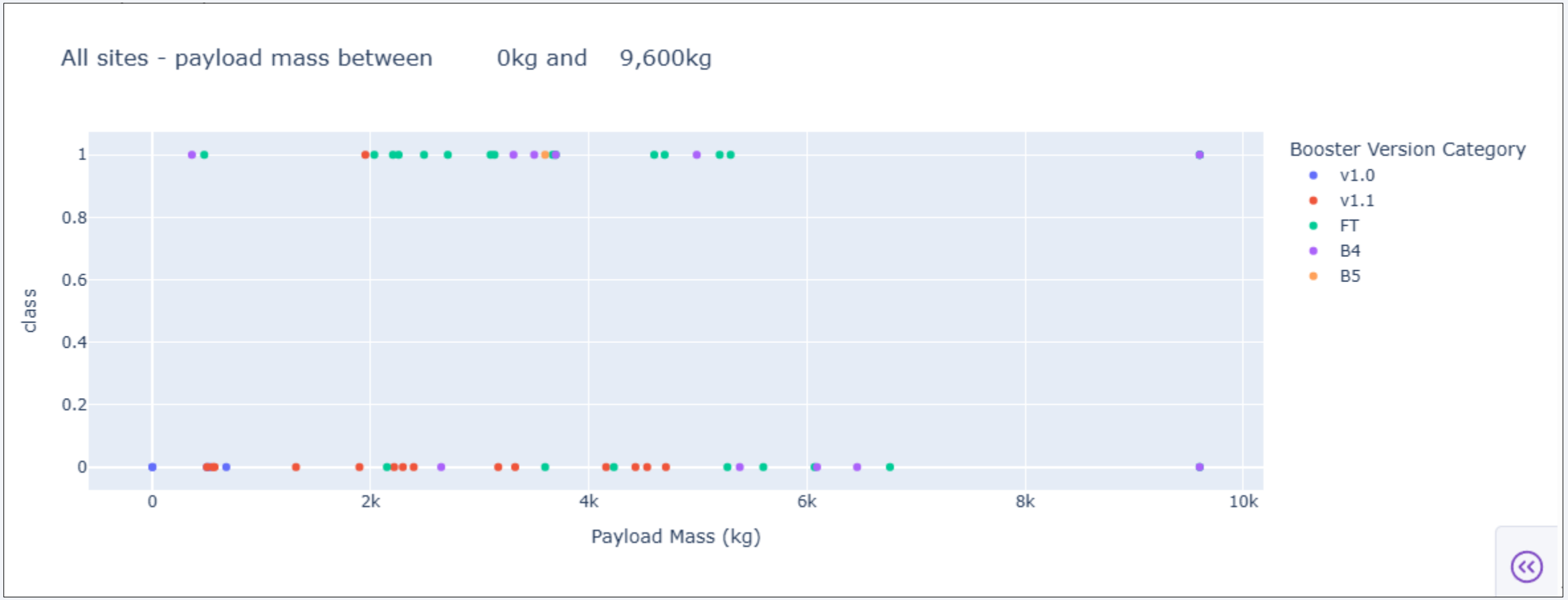






Payload vs. Launch Outcome

The data shows that while early versions had mass-dependent success rates, current Block 5 rockets deliver consistent >90% success regardless of payload mass within the operational range, representing a remarkable achievement in aerospace

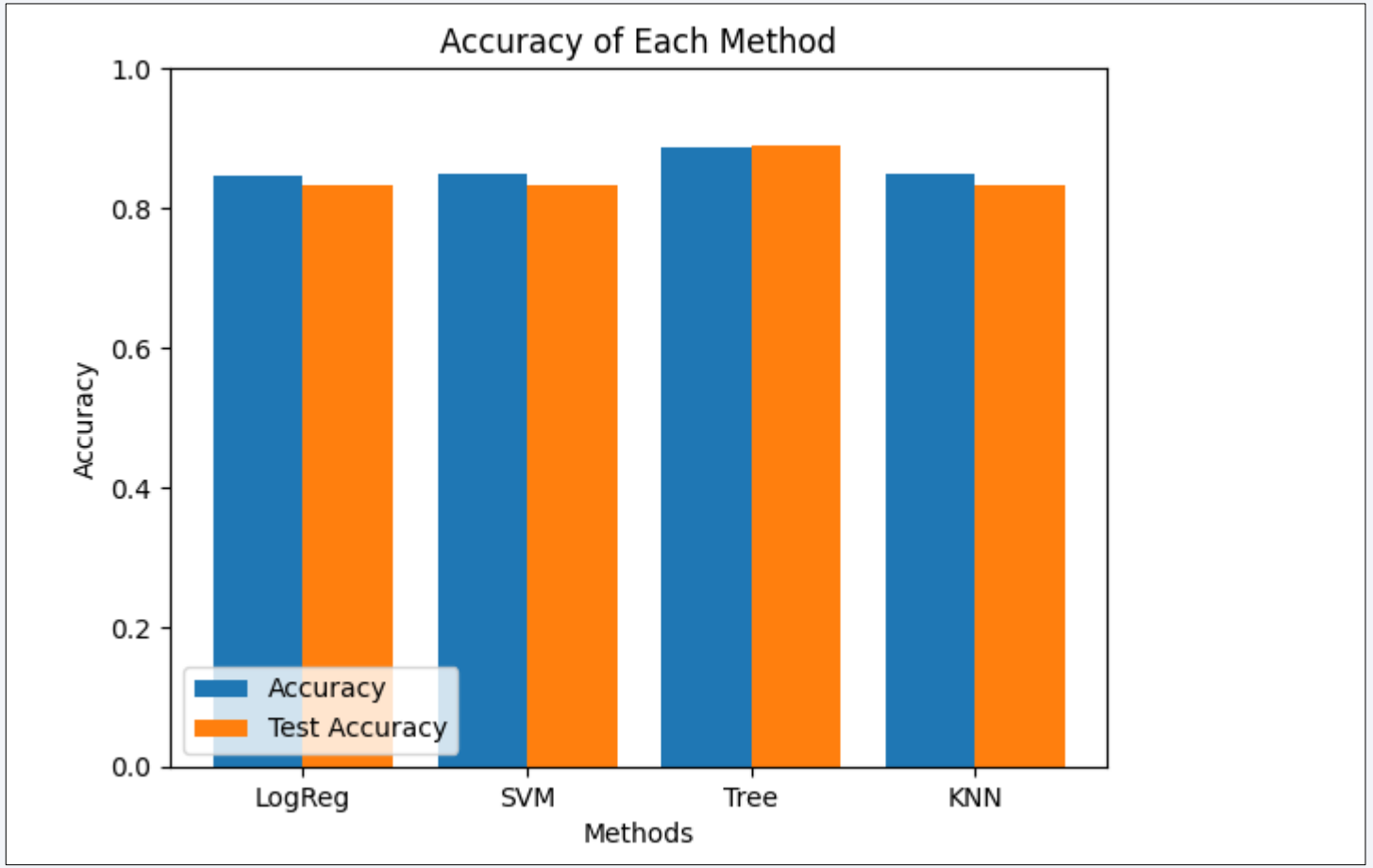


Section 5

Predictive Analysis (Classification)

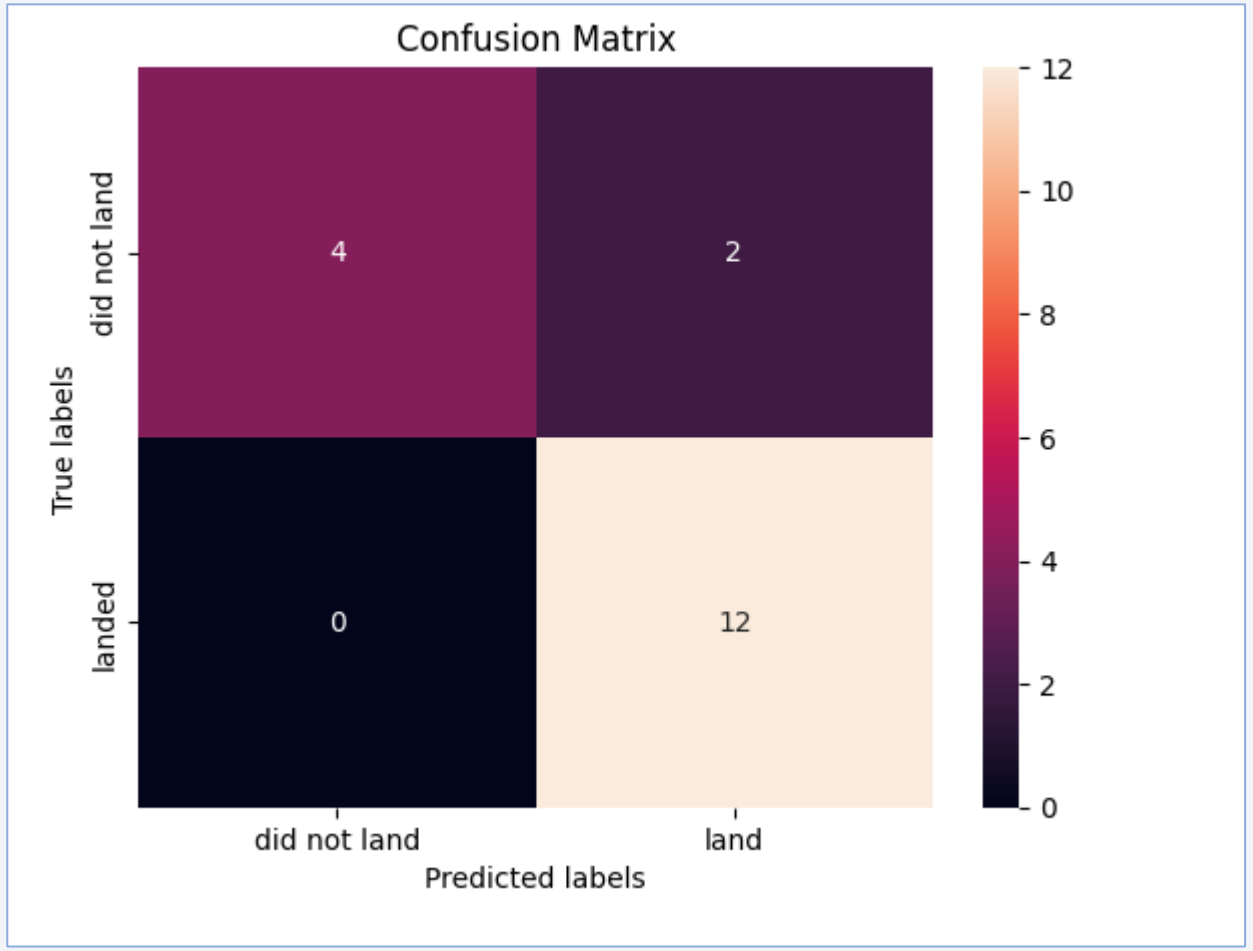
Classification Accuracy

- The Decision Tree (Tree) model has the highest test accuracy. The test accuracy bar for Tree is visually the tallest among all methods, indicating it performs best on unseen data. This aligns with the cross-validation scores from your code, where Tree achieved scores up to approximately 0.87, suggesting strong performance.



Confusion Matrix

- Confusion Matrix Analysis for Decision Tree Model
- Based on the image you provided, I can see the confusion matrix for the Decision Tree model (which was identified as the best performing model). Let me explain what this confusion matrix represents.



Conclusions

- Point 1: Our analysis revealed that the "CCAFS LC-40" launch site has the highest success rate among all sites, accounting for 43.7% of successful launches. This indicates that this site might have optimal conditions or processes that contribute to a higher success rate.
- Point 2: The scatter plot analysis showed that the "FT" booster version has a high success rate across various payload masses, demonstrating its reliability and robustness compared to other booster versions. This suggests that future missions might benefit from utilizing this booster version for improved success rates.
- Point 3: No clear pattern was observed linking higher payload masses to lower success rates, indicating that factors other than payload mass, such as launch site conditions and booster versions, play a more significant role in determining the outcome of a launch
- Point 4: Interactive data visualizations using Folium and Plotly Dash provided valuable insights into the geographical and operational patterns of SpaceX launches. These tools allowed for a deeper understanding of the data, enabling stakeholders to make informed decisions based on comprehensive visual analytics.

Conclusions

- In conclusion, our predictive analysis and interactive visualizations have not only shed light on key factors influencing SpaceX's launch success but also provided a robust framework for future assessments and decision-making in the aerospace industry. The insights gathered can help improve launch strategies and contribute to the ongoing success of reusable rocket technology.

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!