IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fahimeh Feshki
2/1/2026

Outline

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

Executive Summary

- **Summary of methodologies**

Public SpaceX launch data was collected using APIs and web scraping techniques.

The data was cleaned, processed, and analyzed using exploratory data analysis, interactive visualizations, and machine learning classification models.

- **Summary of all results**

Multiple classification models were trained to predict first-stage landing success.

The Decision Tree model achieved the highest test accuracy, showing strong potential for predicting launch outcomes and supporting cost estimation.

Introduction

• This project analyzes SpaceX Falcon 9 launch data to understand the factors that influence the successful landing and reuse of the first-stage booster. Since reusability significantly reduces launch costs, predicting landing success is critical for competitive pricing in the commercial space industry.

• The main objective is to predict whether the Falcon 9 first stage will land successfully based on mission parameters such as payload mass, launch site, orbit type, and flight history, using data analysis and machine learning techniques.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Falcon 9 launch data was collected using the SpaceX REST API, which provides detailed information about launches, payloads, rockets, and landing outcomes. To supplement and validate the API data, additional historical launch records were obtained by web scraping Falcon 9 launch tables from Wikipedia.

Perform data wrangling

Data wrangling involved cleaning and preprocessing the collected data by handling missing values, removing inconsistent records, standardizing column formats, and encoding categorical variables. A binary target variable was also created to indicate whether the Falcon 9 first stage landed successfully.

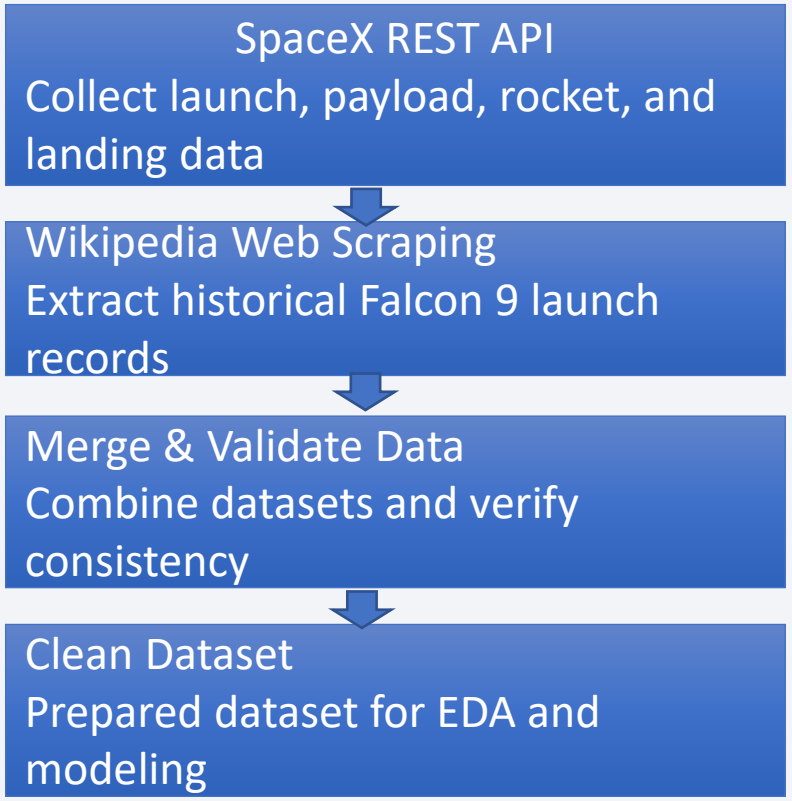
Methodology

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
  - Classification models were built using Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). Hyperparameters were tuned using cross-validation techniques, and model performance was evaluated based on test accuracy to identify the best-performing model for predicting first-stage landing success.

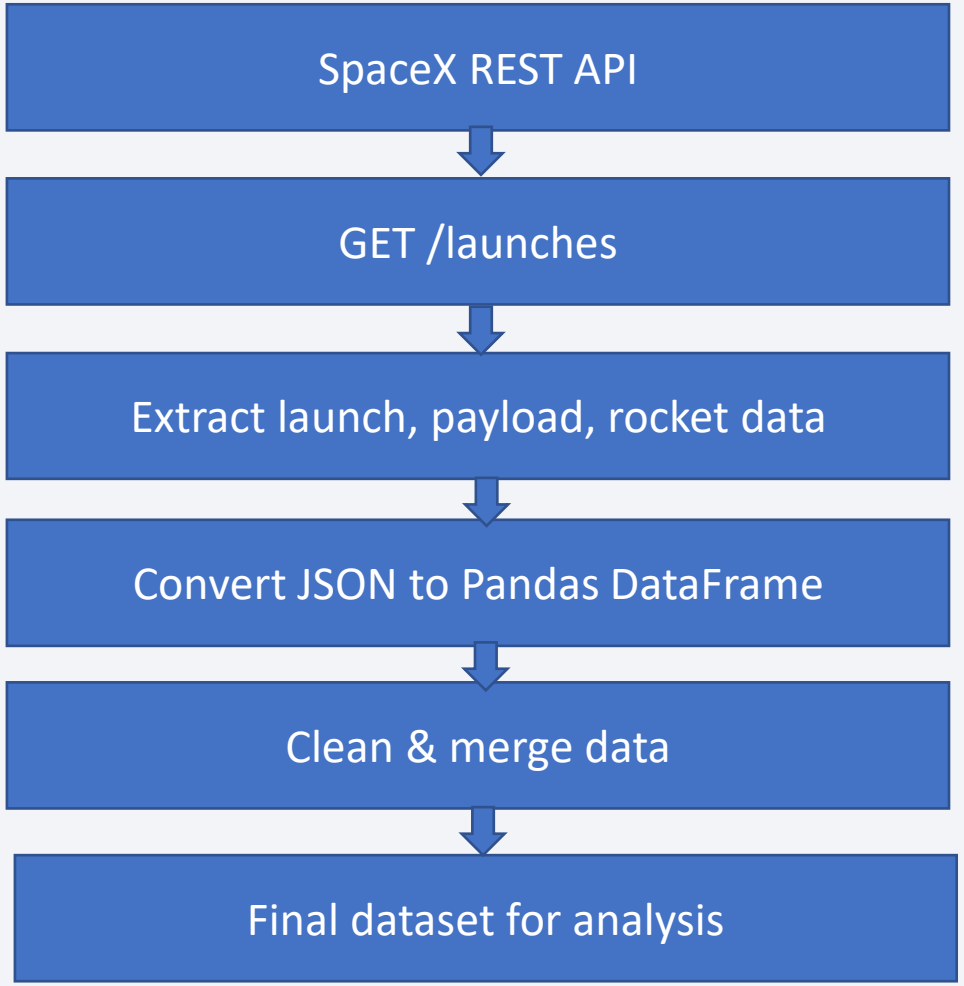
Data Collection Process

- Launch data was collected from multiple sources.
- The SpaceX REST API was used to retrieve structured information on launches, payloads, rockets, and landing outcomes.
- Historical Falcon 9 launch records were additionally extracted from Wikipedia using web scraping.
- All datasets were then merged, validated for consistency, and cleaned to prepare a final dataset for exploratory analysis and machine learning.



Data Collection – SpaceX API

The flowchart summarizes the SpaceX API data collection pipeline, from API requests to the final cleaned dataset used for analysis.



GitHub Reference (Completed SpaceX API Data Collection Notebook):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/1_the%20Data%20Collection%20API.ipynb

Data Collection - Scraping

This flowchart summarizes the web scraping process used to extract historical Falcon 9 launch records from Wikipedia, clean and standardize the data, and merge it with the SpaceX API dataset for analysis.

GitHub Reference (Completed Web Scraping Notebook):

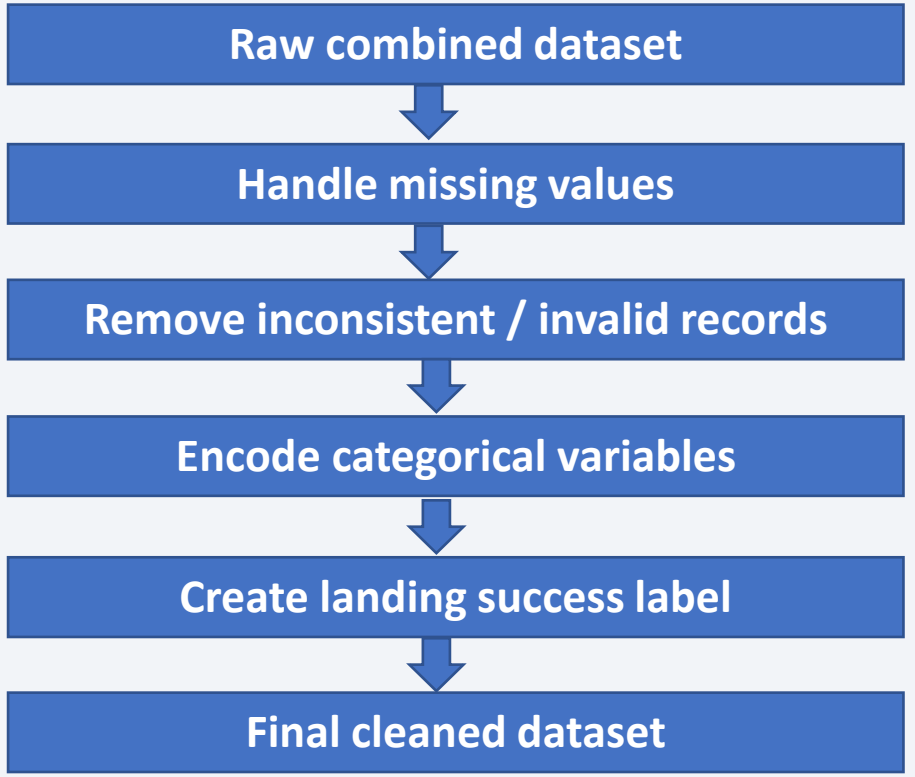
https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone Project/blob/main/2_the%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

This slide summarizes the data wrangling steps applied to prepare the dataset for analysis and modeling. Missing values and inconsistent records were handled, categorical variables were encoded into numerical features, and a binary target variable was created to represent first-stage landing success.

GitHub Reference (Completed Data Wrangling Notebook):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/3_Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter plots were plotted to explore relationships between flight number, payload mass, launch site, orbit type, and landing success.
- Bar charts were used to visualize and compare success rates across different orbit types.
- A line chart was used to observe the yearly trend in launch success rate.
- These charts were selected to effectively identify patterns, correlations, and trends that influence Falcon 9 landing outcomes.

GitHub Reference (Completed EDA with Data Visualization Notebook):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/5_the%20EDA%20with%20Visualization.ipynb

EDA with SQL

- Executed SQL queries to identify unique launch sites.
- Retrieved launch records based on specific launch site patterns.
- Calculated total and average payload mass for selected customers and booster versions.
- Analyzed landing outcomes to identify success and failure counts.
- Used subqueries to find boosters with maximum payload mass.
- Extracted time-based insights such as successful landings by year and month.
- Ranked landing outcomes within a specific date range.

GitHub Reference (Completed EDA with SQL Notebook):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/4_the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Map Objects Used

- Markers for each SpaceX launch site (with popups).
- Circle markers to visualize launches and success/failure class.
- Polylines showing distances from launch sites to the nearest coastline points.
- Markers for nearby features (coastline, highway, railway).
- Distance labels (DivIcon) to display KM values on the map.

Purpose & Insights

- • To clearly locate and compare launch sites on an interactive map.
- To visually separate successful vs unsuccessful landings using color-coded points.
- To measure and illustrate proximity to coastlines (useful for safety and recovery logistics).
- To check whether launch sites are close to key infrastructure such as highways and railways.
- To make distance insights easy to read directly on the map without manual calculation.

GitHub Reference (Completed Interactive Map with Folium Notebook):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/6_Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Dashboard Components

- Created an interactive **pie chart** to show the success vs. failure rate of Falcon 9 launches.
- Added a **dropdown menu** to allow users to filter results by launch site.
- Built a **scatter plot** showing the relationship between payload mass and launch success.
- Included a **range slider** to filter launches based on payload mass.
- Enabled dynamic updates of charts based on user selections using Dash callbacks.

Purpose & Insights

- The pie chart provides a clear overview of launch success rates and allows quick comparison across launch sites.
- The dropdown interaction helps users focus on a specific launch site instead of viewing aggregated data.
- The scatter plot is useful for analyzing how payload mass impacts launch success.
- The range slider enables more detailed exploration of payload effects by narrowing the data range.
- Interactive components improve user experience and allow exploratory data analysis without modifying cod

GitHub Reference (Completed Plotly Dash Lab):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/7_spacex-dash-app.py

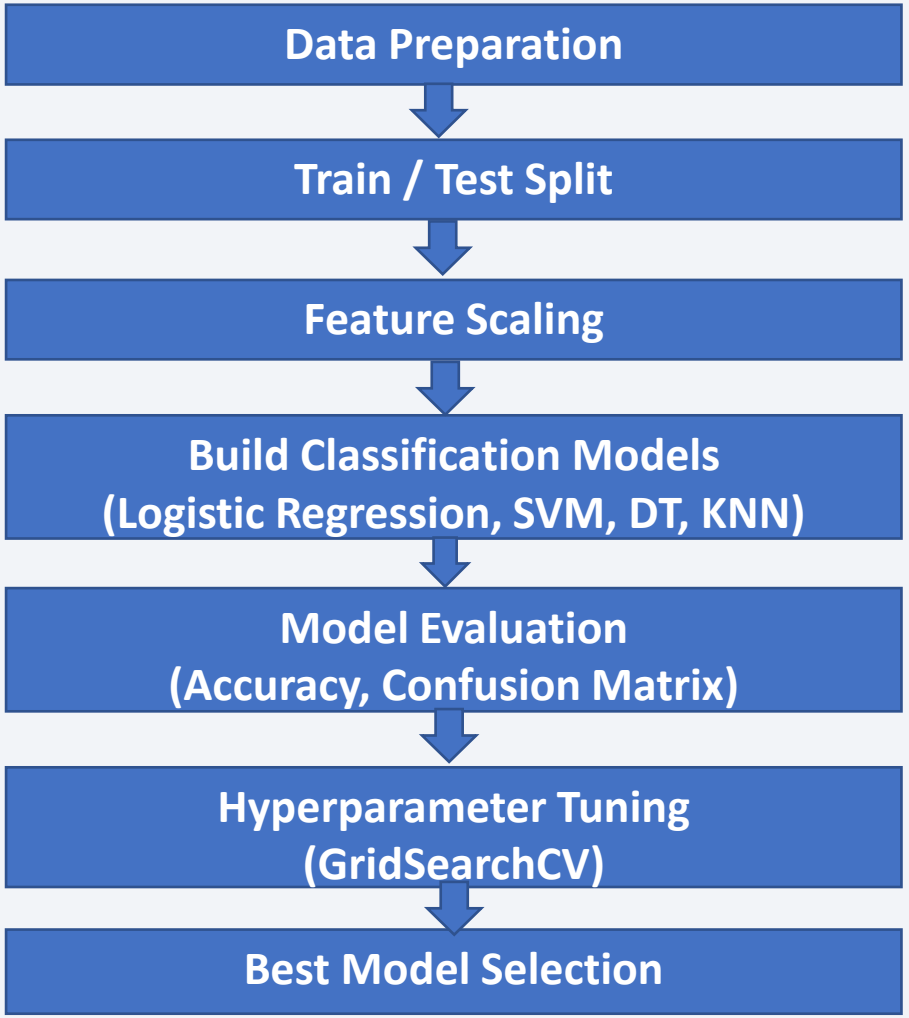
Predictive Analysis (Classification)

Model Development Workflow

- Prepared the final dataset and defined a binary target variable representing first-stage landing success.
- Split the data into training and test sets to ensure unbiased model evaluation.
- Applied feature scaling where required to improve model performance.
- Built multiple classification models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- Evaluated model performance using accuracy scores and confusion matrices to compare classification results.
- Improved model performance through hyperparameter tuning using Grid SearchCV.
- Selected the best-performing model based on evaluation metrics and overall classification performance.

GitHub Reference (Predictive Analysis Lab):

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/8_Complete%20the%20Machine%20Learning%20Prediction%20lab%20(2).ipynb



Results

Exploratory data analysis results

- The launch success rate increases significantly with higher Flight Number, indicating learning and reuse effects over time.

- Payload Mass alone is not a decisive factor; successful landings also occur at higher payload weights.

- Launch site significantly impacts landing success:
  - KSC LC-39A and CCAFS SLC-40 show the highest success rates.
  - VAFB SLC-4E has fewer launches but strong recent performance.

- Orbit type influences outcomes:
  - LEO, ISS, and SSO missions demonstrate higher landing success.
  - GTO missions show comparatively lower success rates.

- Yearly analysis reveals a clear upward trend in landing success after 2013, with peak performance during 2019–2020.

Results

Interactive Analytics

- Tools:
  - Folium Interactive Map
  - Plotly Dash Dashboard

- Folium Interactive Map:
  - Markers represent launch sites.
  - Circles visualize landing success rates.
  - Polylines show distances to the coastline.
  - Polylines show distances to the nearest railway, highway, and city.

- Purpose:
  - To analyze geographical factors affecting launch and landing success.

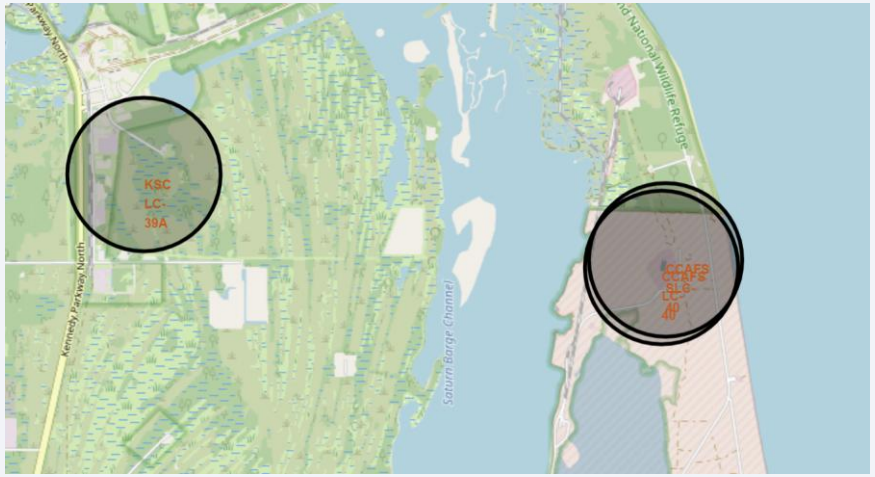
.

Results

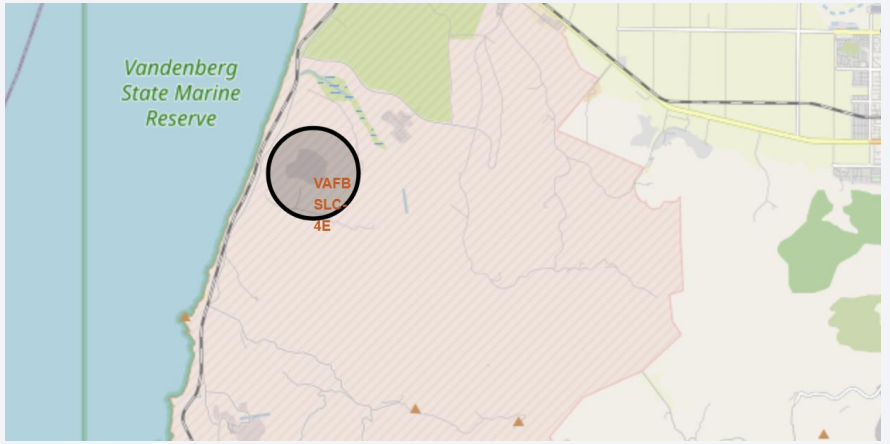
Folium Interactive Map:

- Markers represent launch sites and allow interactive exploration of geographic distribution.

Overview map showing the geographic distribution of all SpaceX launch sites across the United States.



KSC LC-39A, CCAFS SLC-40, and CCAFS LC-40 (Florida launch sites)

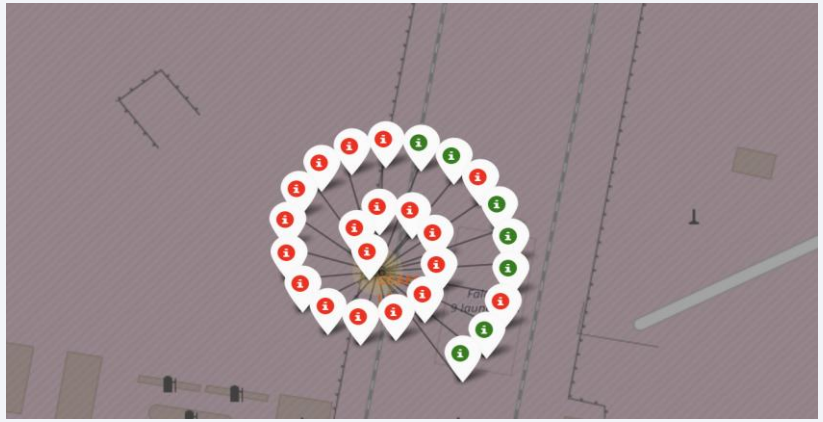


VAFB SLC-4E (California launch site)

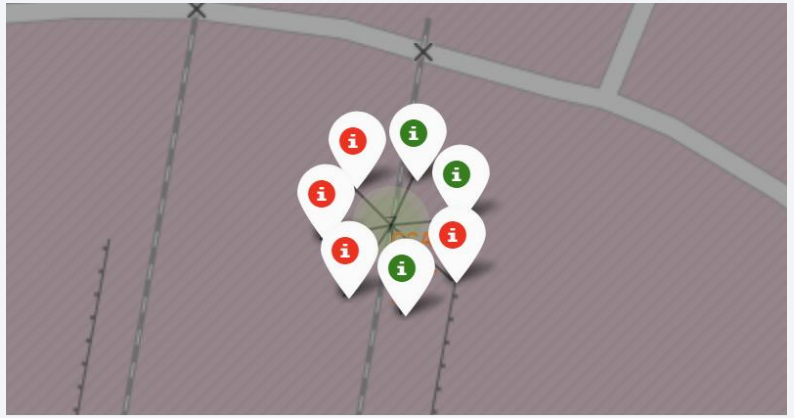
Results

Folium Interactive Map:

Circles visualize landing success rates across different launch sites.

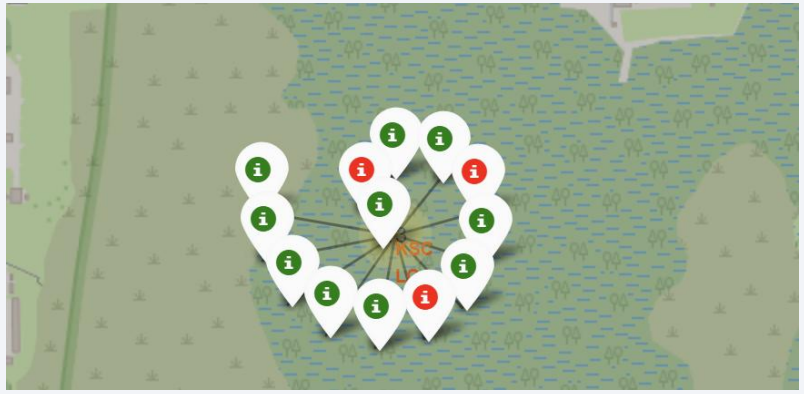


CCAFS LC-40



CCAFS SLC-40

VAFB SLC-4E



KSC LC-39A

Results

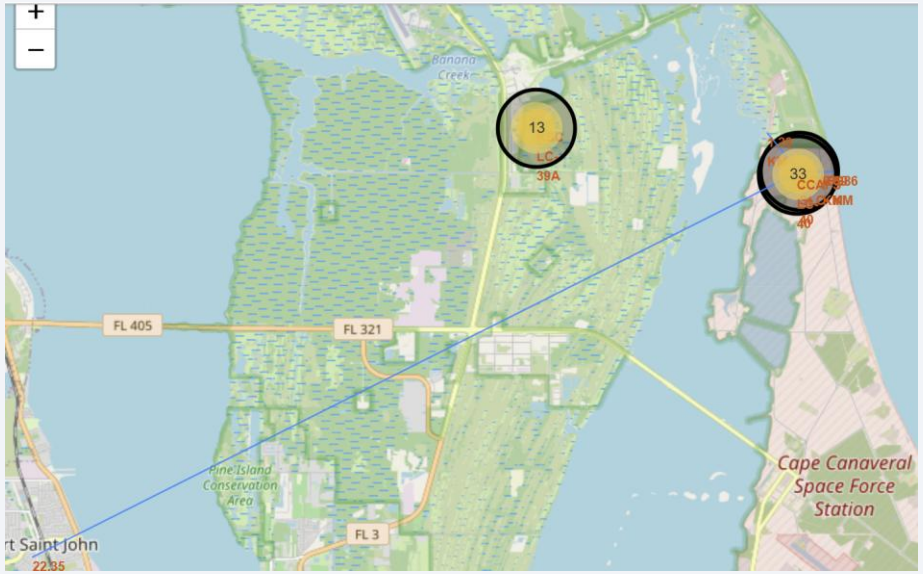
Folium Interactive Map:

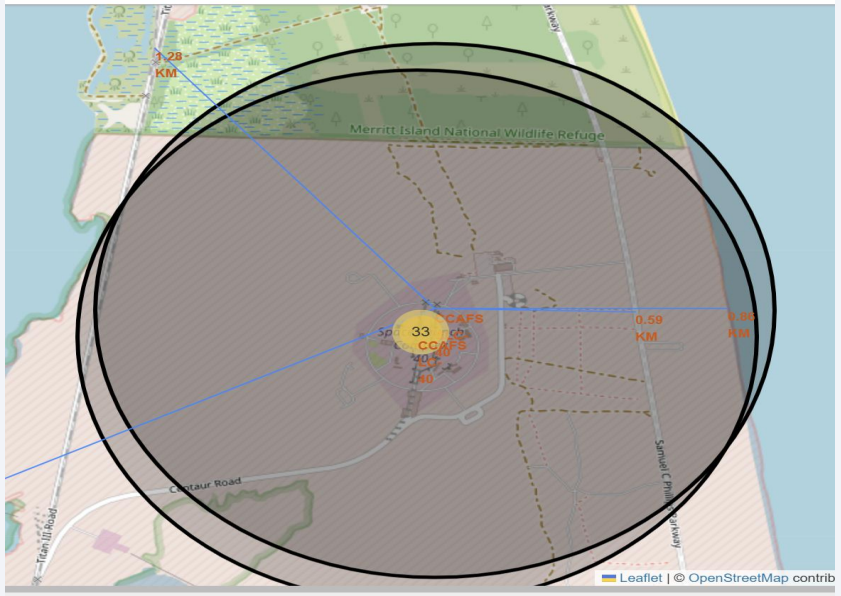
- Polyline shows distance from launch site to coastline.

Results

Folium Interactive Map:

- The polylines represent the distances between the launch site and the nearest railway, highway, and city.





Results

Interactive Analytics

Plotly Dash Dashboard:

- Pie chart displays success vs. failure distribution.
- Scatter plot shows payload mass versus landing outcome.
- Dropdown enables launch site selection.
- Range slider filters payload mass interactively.

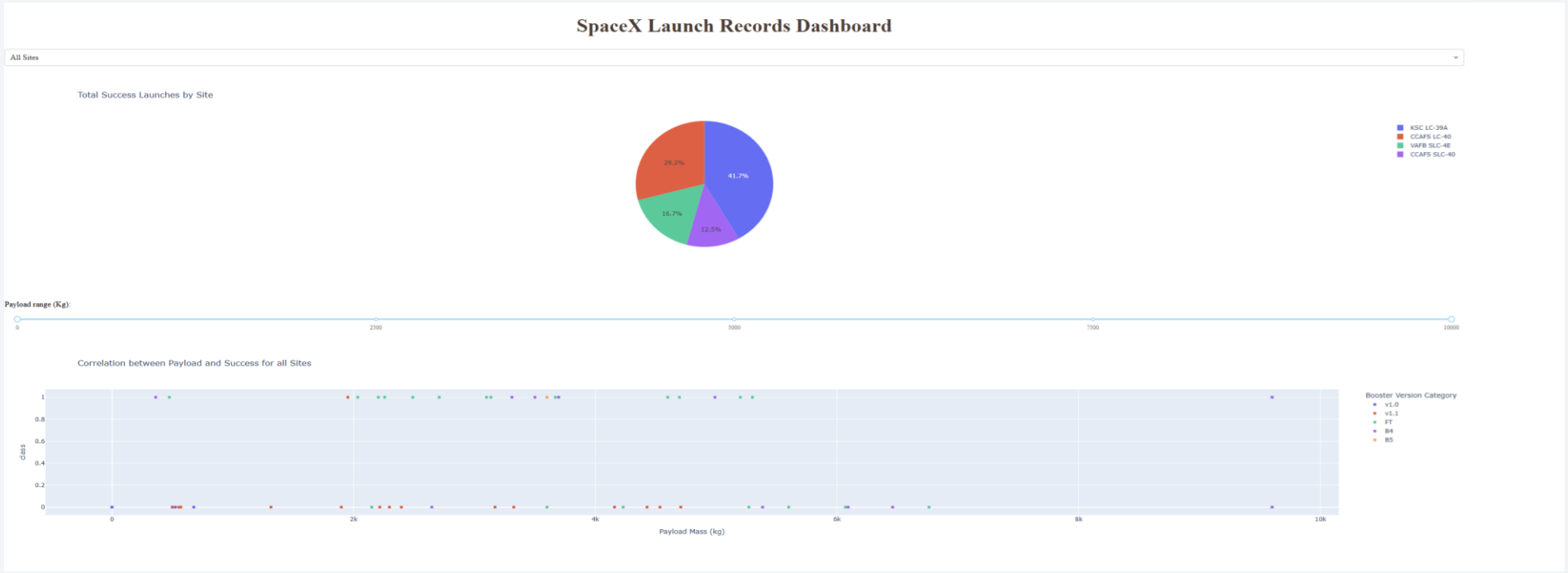
• Purpose:

- To allow dynamic exploration and comparison of launch conditions.

Results

Plotly Dash Dashboard:

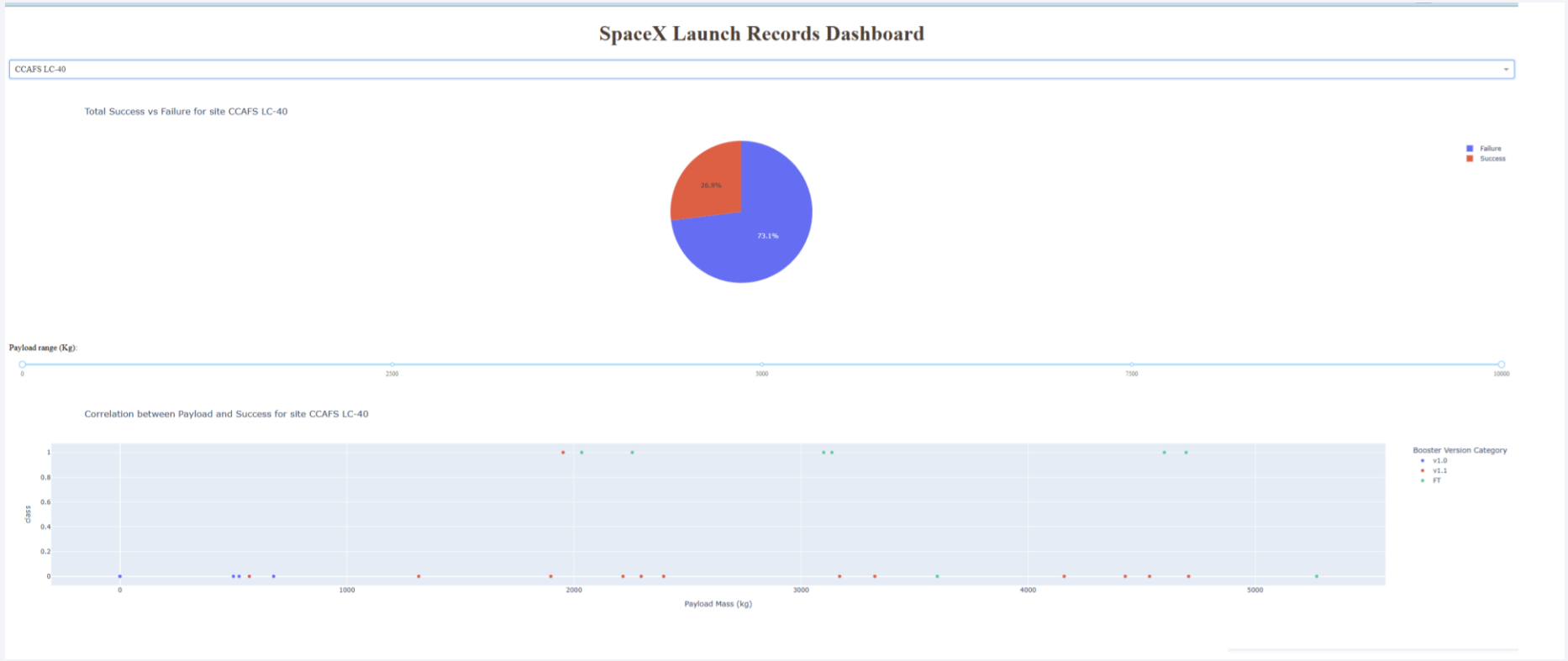
- An interactive Plotly Dash dashboard was built to explore launch success by launch site and payload mass.
- Dropdown filters allow users to dynamically select launch sites and explore payload ranges.



All Sites — Overall launch success and payload–success correlation.

Results

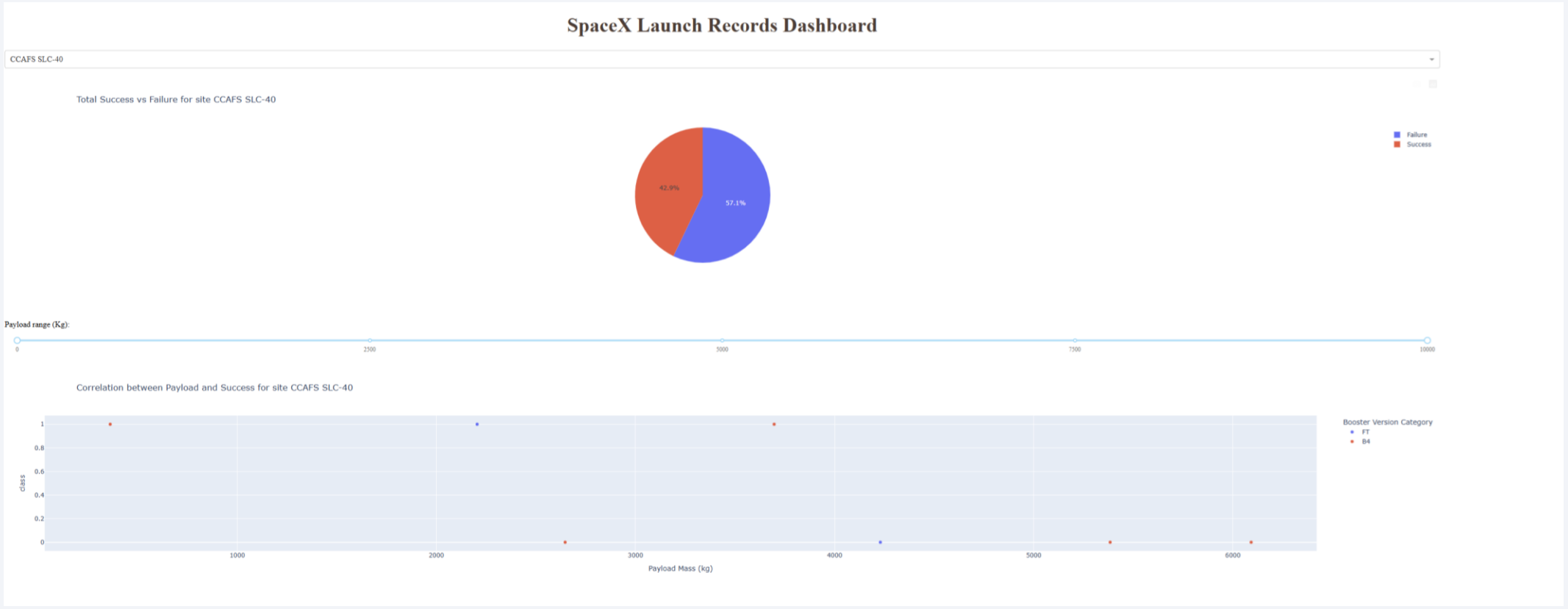
Plotly Dash Dashboard:



CCAFS LC-40 — Launch performance overview

Results

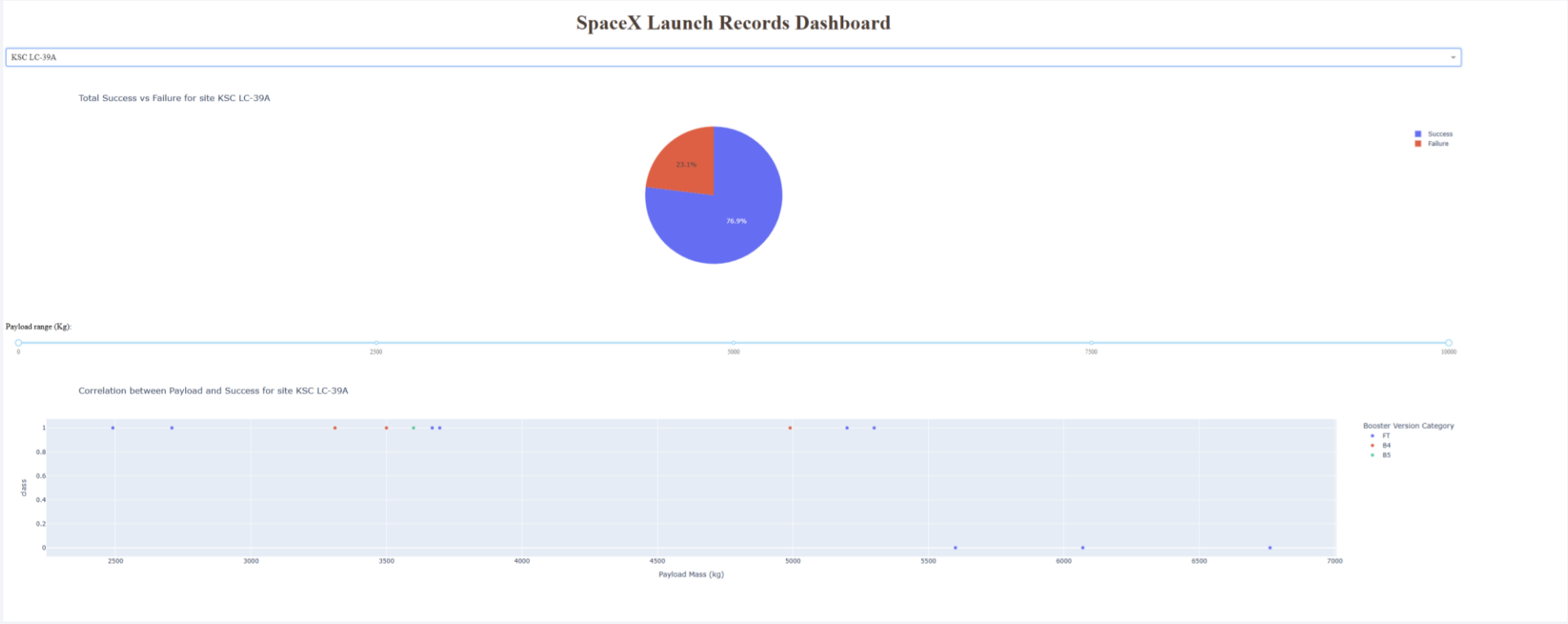
Plotly Dash Dashboard:



CCAFS SLC-40 — Launch performance overview

Results

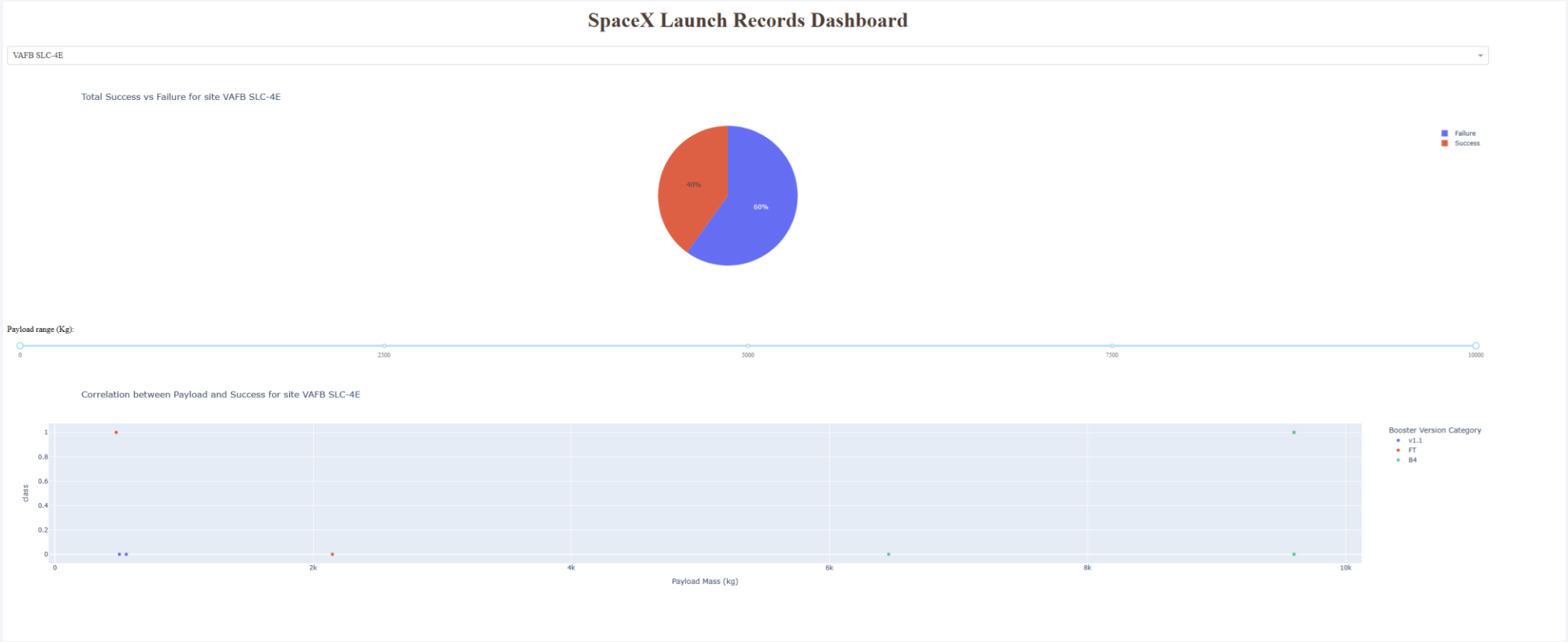
Plotly Dash Dashboard:



KSC LC-39A — Launch performance overview

Results

Plotly Dash Dashboard:



VAFB SLC-4E — Performance

Results

Interactive Analytics Results

Geographical Insights from Launch Sites :

- **Railway proximity:**

  Launch sites are located close to railway lines, enabling efficient transportation of equipment and supplies.

- **Highway proximity:**

  Major highways are nearby, providing convenient ground access and logistical support.

- **Coastline proximity:**

  Launch sites are situated close to the coastline, allowing safer launch trajectories over the ocean and reducing risks to populated areas.

- **Distance from cities:**

  Launch sites maintain a buffer distance from dense urban areas, enhancing safety while remaining accessible to workforce and infrastructure.

Results

Predictive Analysis Results

- Four classification models were trained and evaluated:

  Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

- Model performance was optimized using GridSearchCV with cross-validation on the training data.
- Cross-validation (training) accuracy results:
  - Logistic Regression: 0.846
  - SVM: 0.848
  - Decision Tree: 0.875
  - KNN: 0.848
- Performance was further analyzed using test data.

Results

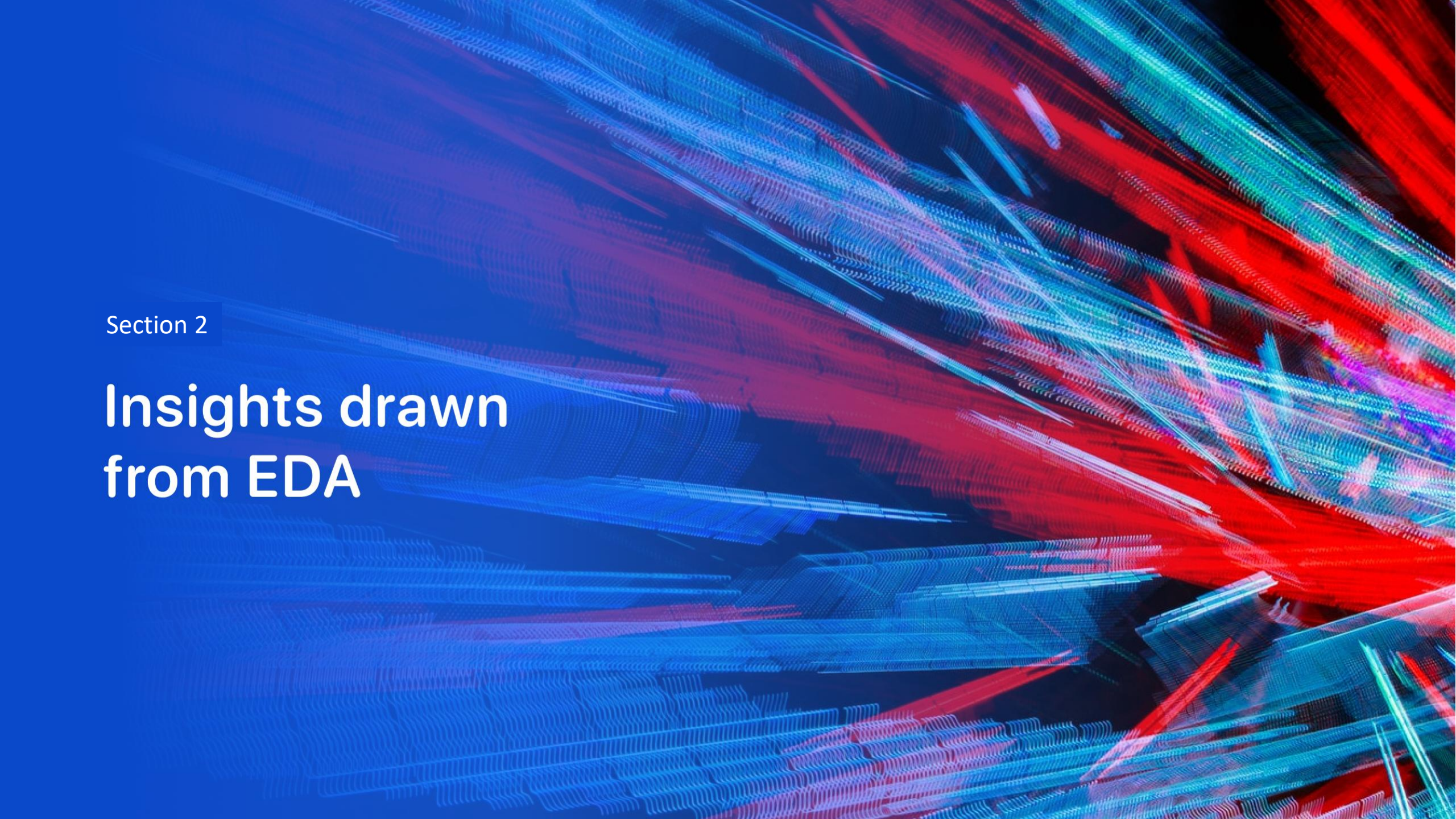
Predictive Analysis Results

- The Decision Tree model achieved the highest cross-validation accuracy on the training data.

- Test accuracy evaluation showed that all models achieved the same performance (0.833) on the held-out test set.

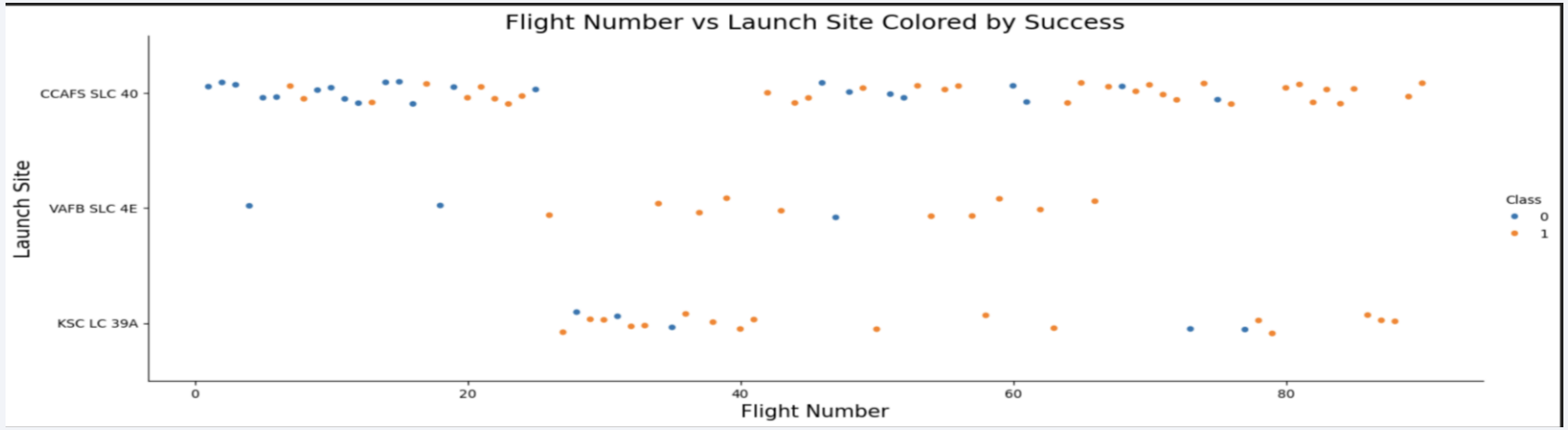
- Based on cross-validation performance and overall evaluation, the Decision Tree model was selected as the best-performing model.

- These results indicate that multiple machine learning models can effectively predict Falcon 9 first-stage landing success, with similar generalization performance on unseen data.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Flight Number vs. Launch Site

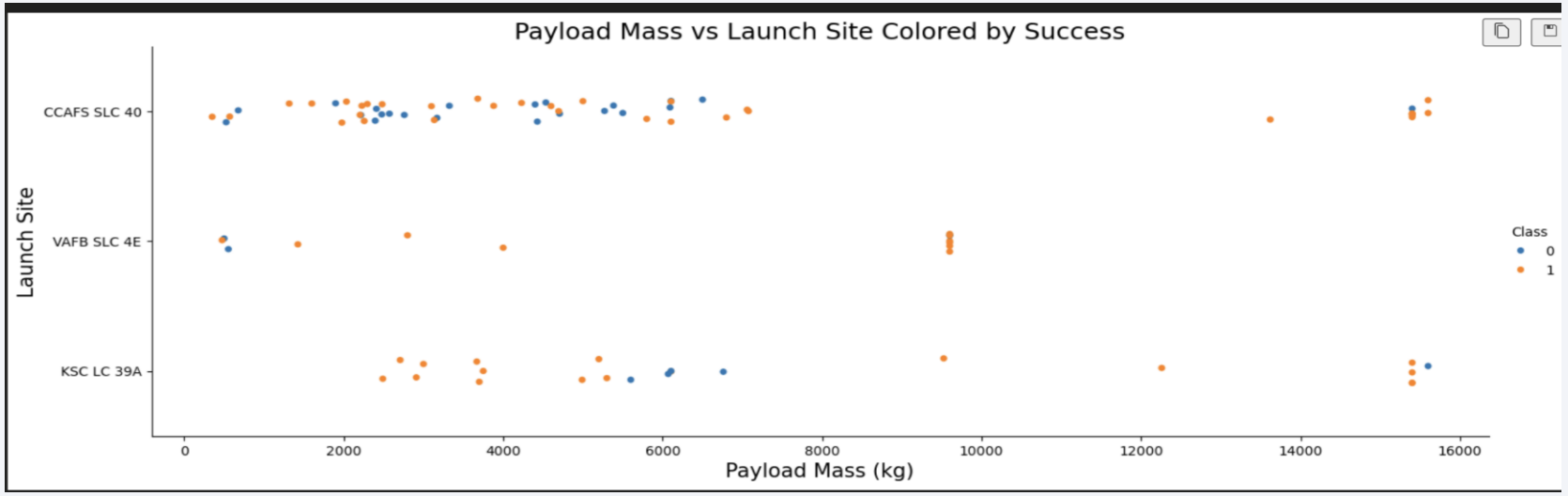


- A scatter plot of **Flight Number vs. Launch Site** was created.
- The scatter plot is **colored by landing outcome** (success vs. failure).
- Each point represents a Falcon 9 launch at a specific launch site.

- The screenshot shows the scatter plot of Flight Number versus Launch Site.
- Early flight numbers include more unsuccessful landings.
- As flight numbers increase, successful landings become more frequent.
- This indicates that launch success improves over time with operational experience across all launch sites.

Payload vs. Launch Site

Payload vs. Launch Site

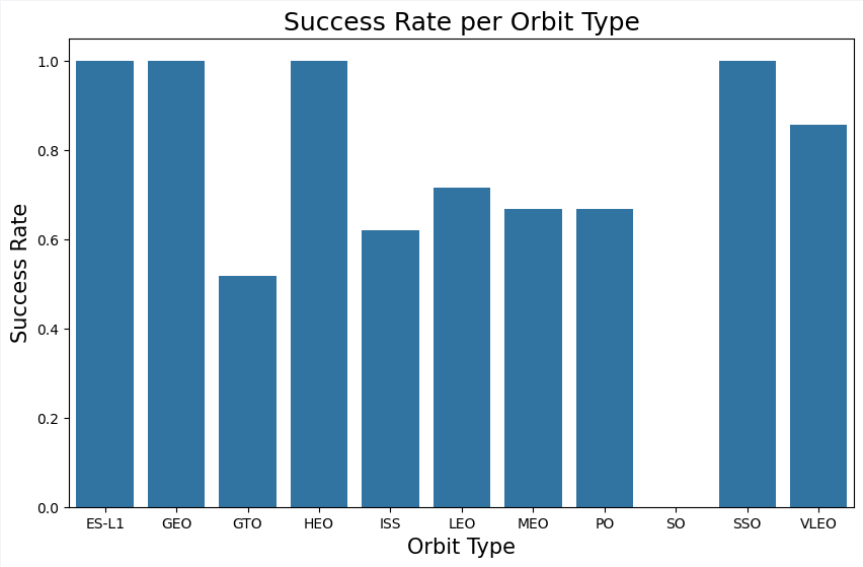


- The scatter plot illustrates the relationship between payload mass and launch site, colored by landing outcome.
- VAFB SLC-4E is primarily used for lighter payload missions and has no launches with payloads above 10,000 kg.
- Heavy payload missions are mainly launched from CCAFS SLC-40 and KSC LC-39A.
- CCAFS SLC-40 supports a wide range of payload masses with successful landings.
- KSC LC-39A handles medium to heavy payloads and shows consistently high success rates.
- Overall, launch site selection is strongly influenced by payload mass and mission requirements.

Success Rate vs. Orbit Type

Success Rate vs. Orbit Type

- The bar chart shows the success rate for each orbit type.
- ES-L1, GEO, HEO, and SSO achieve the highest success rates, close to 100%.
- These orbit types have very few or no failed launches in the dataset.
- VLEO also demonstrates a high success rate, slightly below 100%.
- In contrast, GTO, ISS, LEO, MEO, and PO show moderate success rates, indicating a mix of successful and unsuccessful launches.
- Overall, mission profile and orbit type have a significant impact on launch success, with more specialized or less frequently used orbits showing higher reliability.



Flight Number vs. Orbit Type

Flight number vs. Orbit type

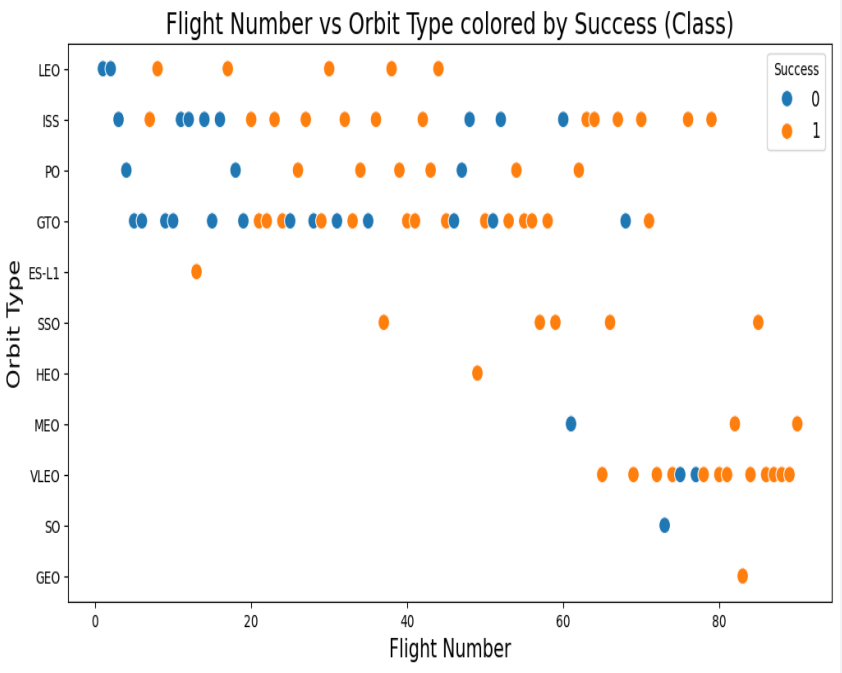
LEO (Low Earth Orbit):
- A clear trend is observed where landing success increases with flight number.
- Early missions show more failures, while later launches are predominantly successful, indicating learning effects, improved reliability, and increasing operational maturity over time.

GTO (Geostationary Transfer Orbit):
- No clear relationship is observed between flight number and landing success.
- Successful and unsuccessful launches are distributed across both early and later flights, suggesting that mission complexity, rather than accumulated experience alone, plays a major role.

ISS and PO orbits:
- Mixed outcomes are observed across different flight numbers, with a general improvement in success rates in later missions.

VLEO:
- Missions mostly appear in later flight numbers and show high success rates, indicating mature operations at the time these missions were introduced.

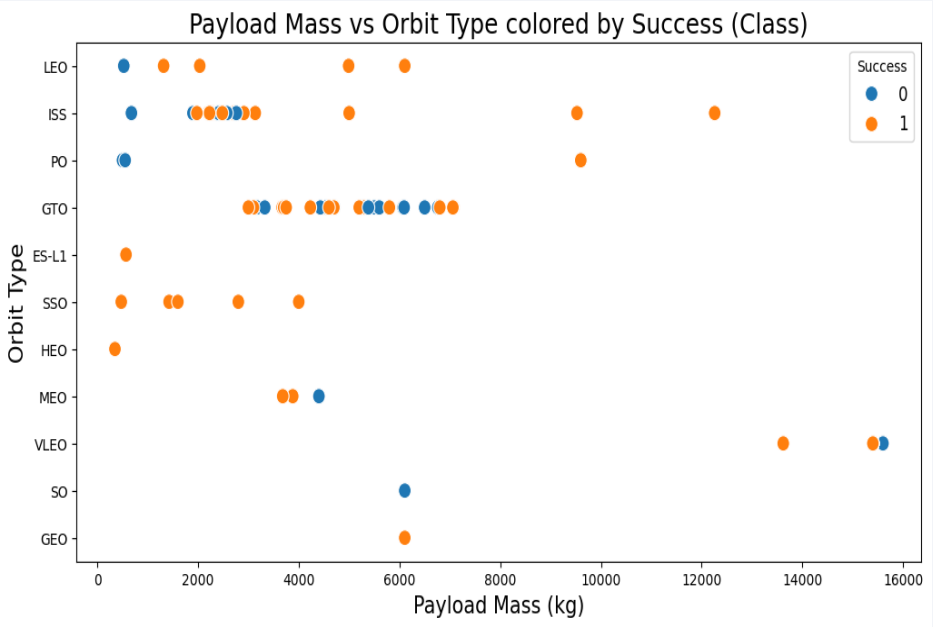
Specialized orbits (ES-L1, SSO, GEO, HEO):
- These orbits have a limited number of launches but are mostly successful, making it difficult to infer strong trends with respect to flight number.



Payload vs. Orbit Type

Payload vs. Orbit type

- **LEO (Low Earth Orbit):**
  Launches cover a wide range of payload masses.
  Success rate is high across different payload values, indicating reliable and mature operations.
- **ISS:**
  Payload masses are moderate and tightly clustered.
  Most launches are successful due to standardized mission profiles.
- **GTO (Geostationary Transfer Orbit):**
  Payload masses are moderate to high.
  Both successful and failed launches occur, showing higher mission complexity and risk.
- **VLEO:**
  Appears only at very high payload masses.
  Mostly successful launches, likely representing later missions with improved technology.
- **SSO and ES-L1:**
  Generally lower payload masses.
  Mostly successful launches, but with limited data points.
- **GEO, SO, HEO, and MEO:**
  Few launches overall.
  Mixed outcomes suggest orbit complexity impacts success more than payload mass.



Key Insight:
Payload mass alone does not determine launch success; orbit type and mission complexity play a more significant role.

Launch Success Yearly Trend

Temporal Trend of Launch Success Rate

- From 2010 to 2013:
  Launch success rate is very low or near zero, indicating early development and testing phases.

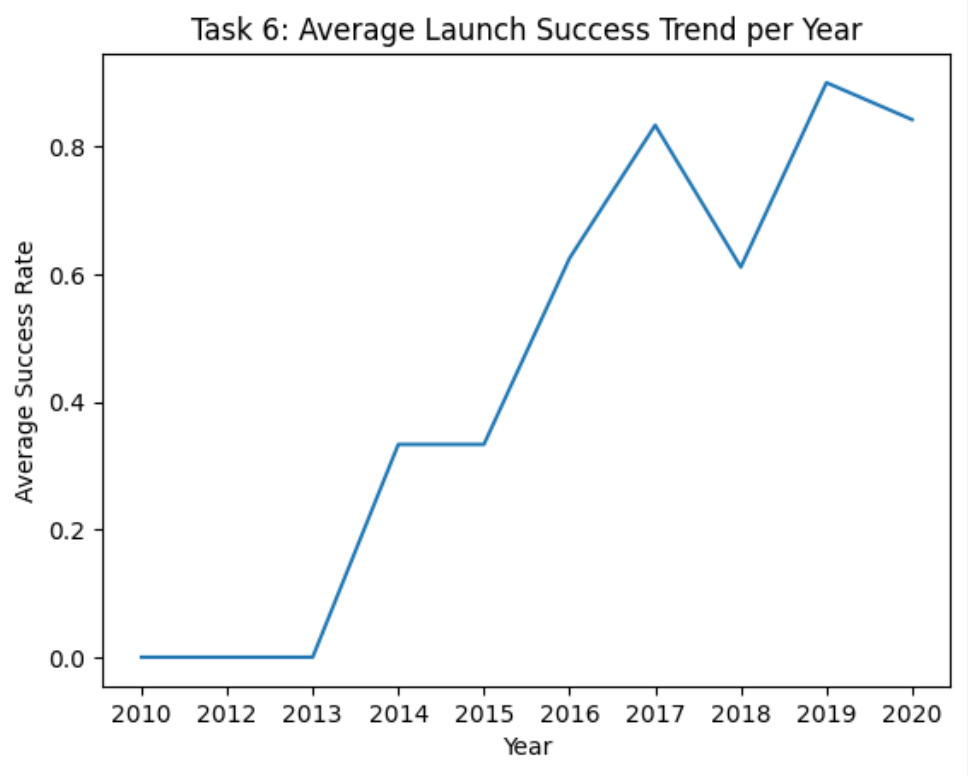
- 2014–2015:
  A noticeable improvement appears, suggesting technological learning and operational stabilization.

- 2016–2017:
  Success rate increases sharply, reaching a high level.
  This reflects significant improvements in launch reliability and experience.

- 2018:
  A temporary decline in success rate occurs, possibly associated with increased mission complexity or experimental launches.

- 2019–2020:
  Success rate recovers and remains consistently high (above 80%),
  indicating a mature and reliable launch system.

Key Insight:
Overall launch success improves significantly over time, showing strong learning effects, technological advancement, and increased operational efficiency. The success rate shows a clear upward trend from 2013 to 2020.



All Launch Site Names

A query was executed to identify all unique Falcon 9 launch sites in the dataset.

```
# Task 1: Display the names of all unique launch sites from the SPACEXTABLE
# This query retrieves distinct values from the 'Launch_Site' column to show all launch locations
query1 = 'SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;'
df_task1 = pd.read_sql_query(query1, con)
df_task1
```

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

- **CCAFS LC-40**: Major site for Falcon missions in Florida.
- **CCAFS SLC-40**: Another site in Florida, frequently used for Falcon launches.
- **KSC LC-39A**: Kennedy Space Center, used for high-profile launches.
- **VAFB SLC-4E**: Vandenberg site, often used for polar orbit missions.

Launch Site Names Begin with 'KSC'

Launch Site Names Begin with 'KSC'

```
# Task 2: Display 5 records where launch sites start with 'KSC'
# The LIKE operator is used to match patterns in the 'Launch_Site' column

query2 = 'SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "KSC%" LIMIT 5;'
df_task2 = pd.read_sql_query(query2, con)
df_task2
```

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

The query retrieves 5 launch records where the launch site name starts with KSC, corresponding to launches from Kennedy Space Center (KSC LC-39A).

These records show a variety of payload masses, mission types, and landing outcomes. All missions were successfully launched, with different landing results including ground pad landings, drone ship landings, and no landing attempts.

Total Payload Mass

Total Payload Mass Carried by NASA (CRS) Launches

```
# Task 3: Calculate the total payload mass for boosters launched by NASA (CRS)
# SUM function is used to get the total payload mass
query3 = 'SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer"="NASA (CRS)";'
df_task3 = pd.read_sql_query(query3, con)
df_task3
```

	Total_Payload_Mass
0	45596

- In this SQL query, the SUM function is used to calculate the total payload mass for boosters launched by the customer NASA (CRS).
- The result of this query shows that the total payload mass for launches by NASA (CRS) is 45,596 kg.

Average Payload Mass by F9 v1.1

Average Payload Mass Carried by Falcon 9 v1.1 Boosters

```
# Task 4: Calculate the average payload mass for boosters with version F9 v1.1
# AVG function is used to calculate the mean payload mass
query4 = 'SELECT AVG("Payload_Mass__kg_") AS Avg_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version"="F9 v1.1";'
df_task4 = pd.read_sql_query(query4, con)
df_task4
```

	Avg_Payload_Mass
0	2928.4

- In this SQL query, the AVG function is used to calculate the average (mean) payload mass for boosters with version Falcon 9 v1.1.
- The result shows that the average payload mass carried by this booster version is 2,928.4 kg.

First Successful Ground Landing Date

Date of First Successful Drone Ship Landing

```
# Task 5: Find the earliest date when a successful landing on drone ship occurred
# MIN function retrieves the minimum date for the condition
query5 = 'SELECT MIN("Date") AS First_Successful_Landing_DS FROM SPACEXTABLE WHERE "Landing_Outcome"="Success (drone ship)";'
df_task5 = pd.read_sql_query(query5, con)
df_task5
```

	First_Successful_Landing_DS
0	2016-04-08

- This SQL query uses the MIN function to retrieve the earliest date on which a successful landing occurred on a drone ship.The result indicates that the first successful drone ship landing took place on April 8, 2016.

Successful Ground Pad Landings with Payload between 4000 and 6000

Boosters with successful ground pad landings (4000–6000 kg payload)

```
# Task 6: List booster names that had a successful landing on ground pad and payload mass between 4000 and 6000 kg

query6 = 'SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome"="Success (ground pad)" AND "Payload_Mass__kg_" BETWEEN 4000 AND 6000;'
df_task6 = pd.read_sql_query(query6, con)
df_task6
```

	Booster_Version
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

- This query identifies booster versions that successfully landed on a ground pad with payload masses between 4000 and 6000 kg.The result shows three boosters that meet these criteria: F9 FT B1032.1, F9 B4 B1040.1, and F9 B4 B1043.1.

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes:

```
# Task 7: Count total successful and failed mission outcomes
# GROUP BY is used to aggregate counts for each outcome type
query7 = 'SELECT "Mission_Outcome", COUNT(*) AS Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";'
df_task7 = pd.read_sql_query(query7, con)
df_task7
```

	Mission_Outcome	Count
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

- The query counts the total number of mission outcomes in the SpaceX dataset, including successful and failed missions.The results show that the vast majority of missions were successful, with 98 missions classified as successful.In addition, there was 1 failed mission during flight and 1 mission with a successful launch but unclear payload status, which is grouped separately in the dataset.

Boosters Carried Maximum Payload

The names of the boosters that have carried the maximum payload mass:

```
# Task 8: List all booster versions that carried the maximum payload mass
# We first find the maximum payload mass using a subquery and then select boosters that carried it
query8 = '''
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);
'''
df_task8 = pd.read_sql_query(query8, con)
df_task8
```

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

- The query identifies the maximum payload mass in the dataset and returns all booster versions that carried this maximum payload.

2017 Launch Records

This query displays the month, successful ground pad landing outcomes, booster versions, and launch sites for launches in the year 2017.

```
# Task 9: List month, successful ground pad landings, booster version, and launch site for 2017
# SQLite does not support month names, so we extract the month from the Date column using substr()

query9 = '''
SELECT substr("Date",6,2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE
WHERE substr("Date",1,4)='2017' AND "Landing_Outcome"="Success (ground pad)";
'''
df_task9 = pd.read_sql_query(query9, con)
df_task9
```

[32]

	Month	Landing_Outcome	Booster_Version	Launch_Site
0	02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
1	05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
2	06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
3	08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
4	09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
5	12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- In 2017, successful ground pad landings occurred in multiple months, primarily at the KSC LC-39A launch site. The only exception was in December, when a landing took place at CCAFS SLC-40. These landings involved different versions of the Falcon 9 booster, ranging from F9 FT B1031.x to F9 B4/B5 variants.

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

Counts and ranks landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

The query counts and ranks different landing outcomes within the specified date range.

The results show that "No attempt" was the most frequent outcome, indicating that many early launches did not include landing attempts.

This is followed by both successful and failed drone ship landings, reflecting the gradual development and testing of recovery technologies during this period.

```
# Task 10: Count and rank landing outcomes between 2010-06-04 and 2017-03-20
# ORDER BY Count descending to rank the outcomes

query10 = '''
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;
'''
df_task10 = pd.read_sql_query(query10, con)
df_task10
```

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

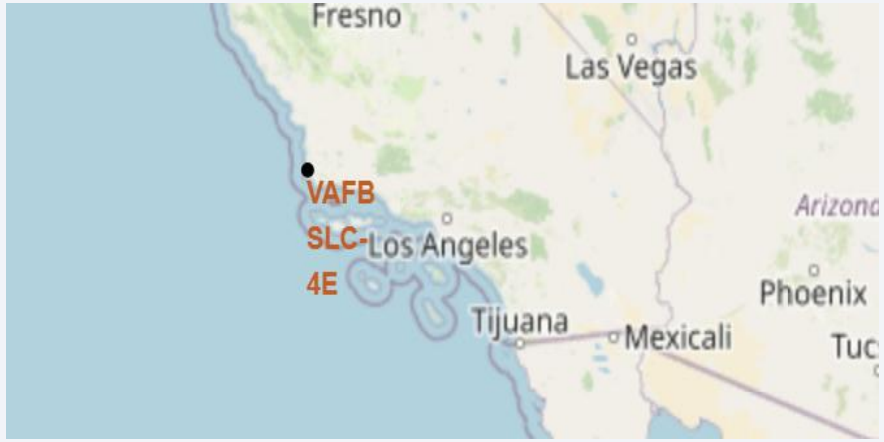

Section 3
Launch Sites
Proximities Analysis

Global Launch Sites Map with Markers

Global Launch Sites Map with Markers

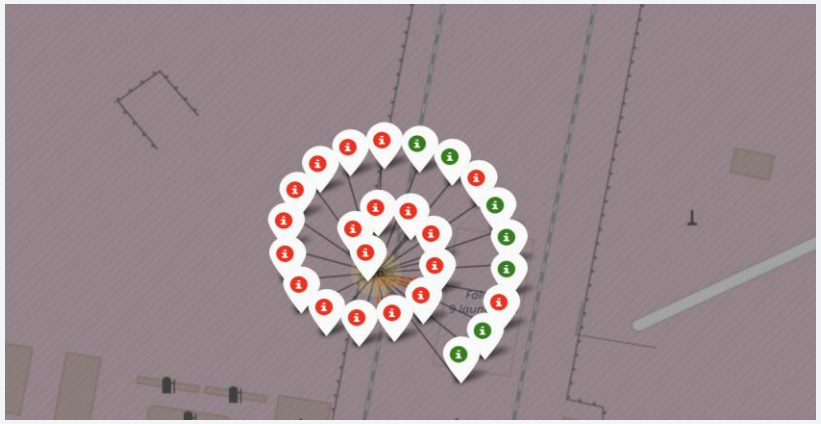
Global Launch Sites Map with Markers

Global Launch Sites Map with Markers

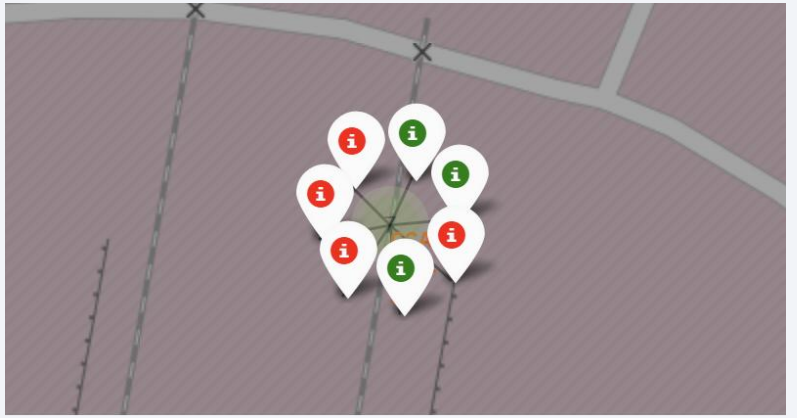


Screenshots represent a global map with the locations of all launch sites marked by location markers.
The markers show the geographical coordinates (latitude and longitude) for each launch site, such as CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.You can clearly see how the launch sites are distributed across the United States, primarily near coastlines, providing optimal safety for rocket launches.
The markers are interactive, allowing for zooming and panning to explore each site.
Key findings include the clustering of launch sites in coastal areas to minimize risk in case of rocket failure, as well as the diversity of launch sites spread across both East and West coasts of the U.S. for different mission needs (e.g., cargo missions, satellite launches, etc.).

Launch Outcomes on the Global Map

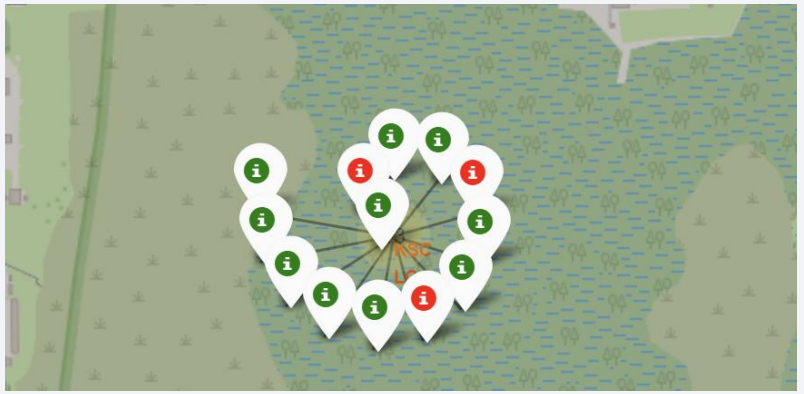


CCAFS LC-40



CCAFS SLC-40

VAFB SLC-4E

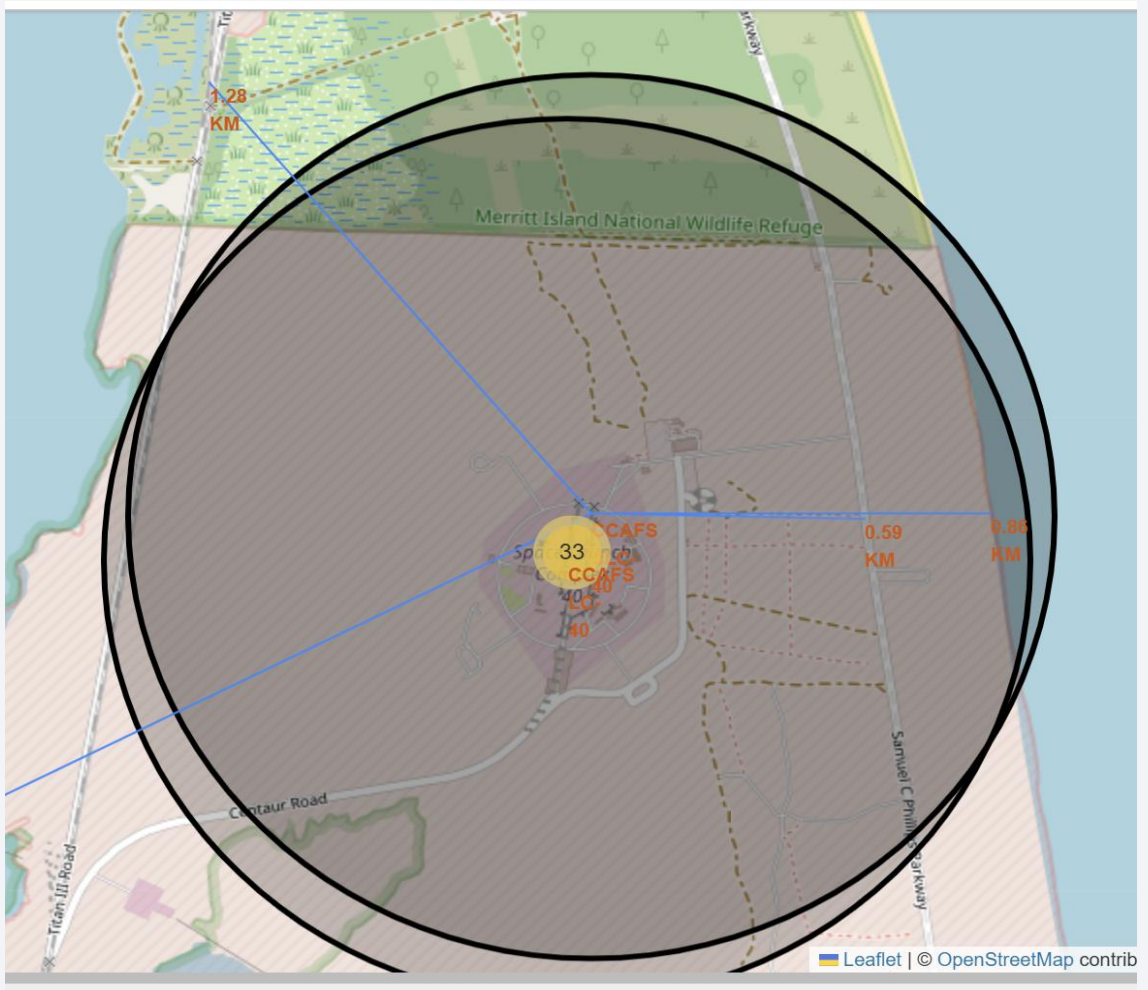


KSC LC-39A

Launch Outcomes on the Global Map

- Based on the color-labeled markers in the marker clusters, you can clearly identify the launch sites with high success rates.
- The green markers indicate successful launches, while red markers represent failed launches.
- Sites with a higher number of green markers, such as KSC LC-39A, show relatively higher success rates. Conversely, sites with more red markers, such as CCAFS LC-40 and VAFB SLC-4E, have lower success rates.
- This visual representation highlights the correlation between the number of successful landings and the launch site's overall performance.

Launch Site Proximities and Distance Analysis



Launch Site Proximities and Distance Analysis

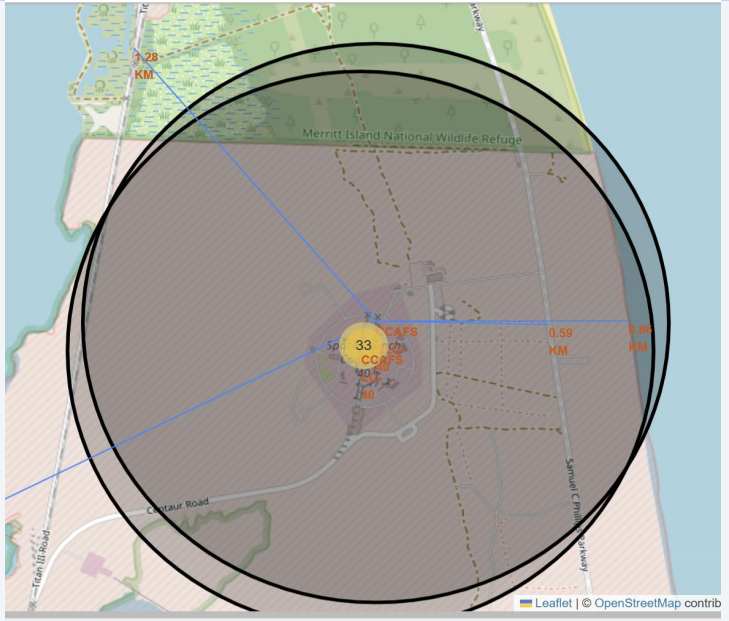
The selected launch site is **CCAFS SLC-40** (Cape Canaveral Air Force Station Space Launch Complex 40), located on the east coast of Florida, USA. The distances to nearby infrastructure and geographical features were measured directly from the folium map using cursor-selected coordinates.

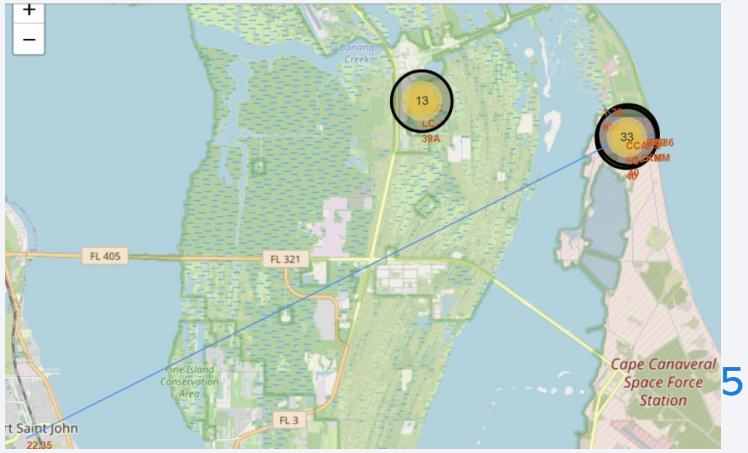
Railways: The launch site is approximately 1.28 km away from the nearest railway line, indicating convenient access for transporting heavy equipment and materials.

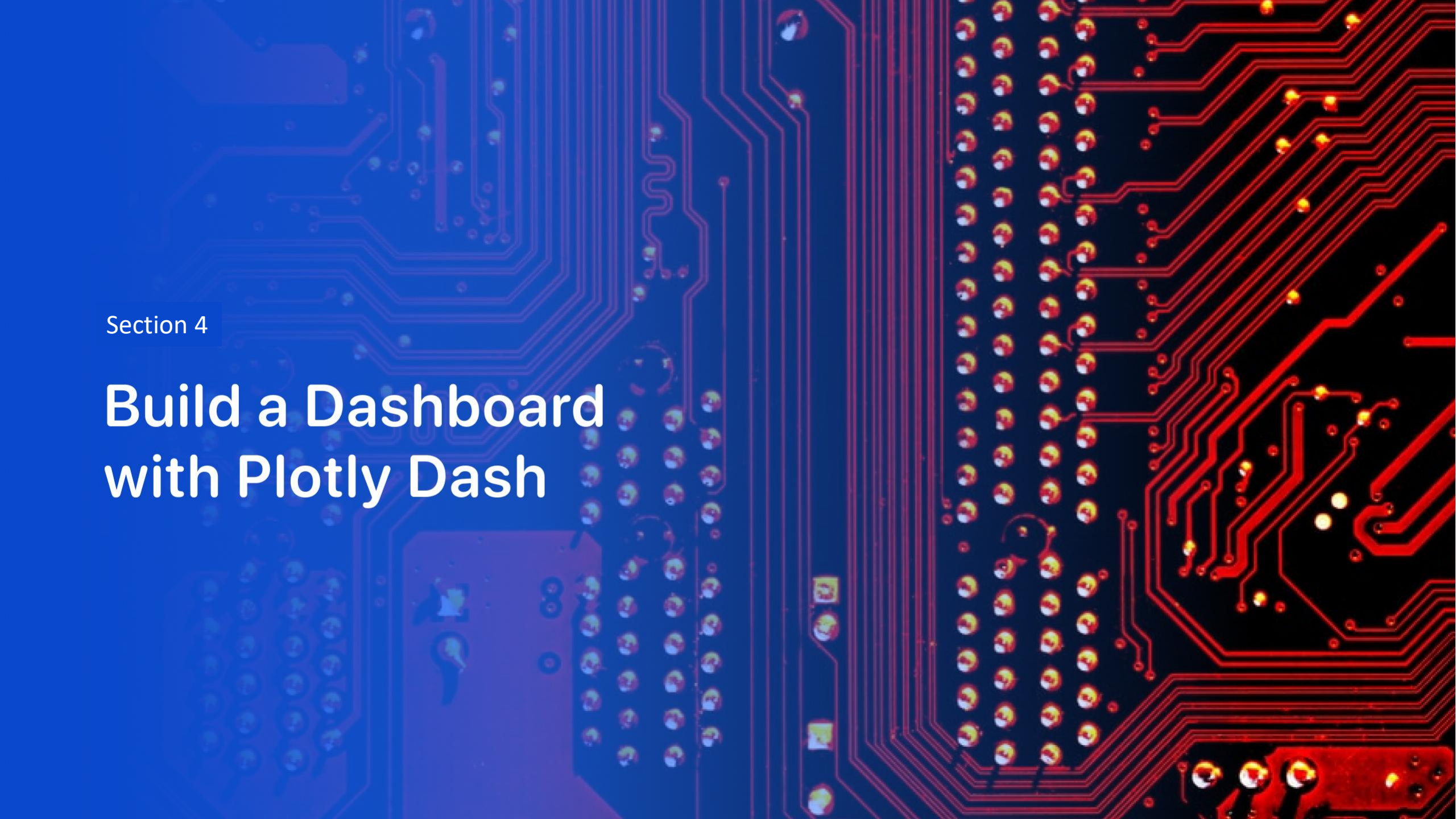
Highways: The site is about 0.59 km from a nearby highway, providing strong ground transportation and logistical support.

Coastline: The launch site is located roughly 0.86 km from the coastline, which is ideal for launch safety as rockets can fly over the ocean, minimizing risks to populated areas.

Urban areas: The nearest city, Port St. John, is approximately 22.35 km away, ensuring a sufficient safety buffer from urban regions while still allowing access to workforce and infrastructure.

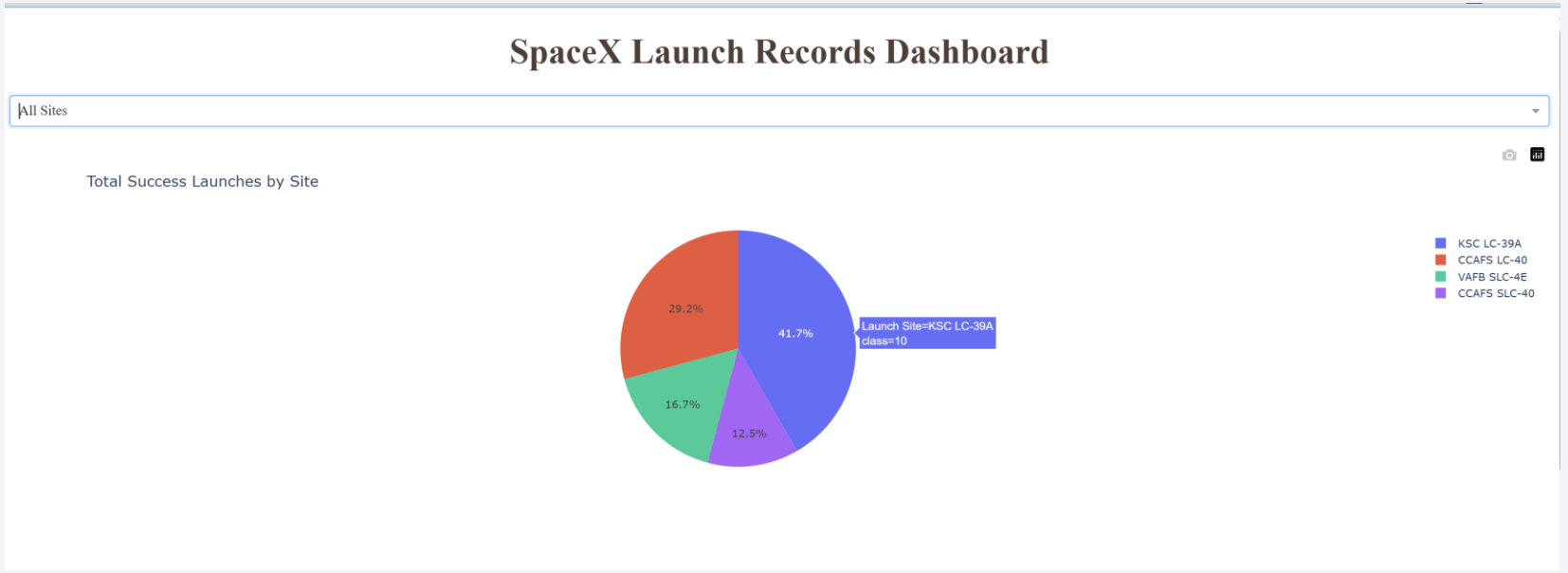





Section 4
Build a Dashboard with Plotly Dash

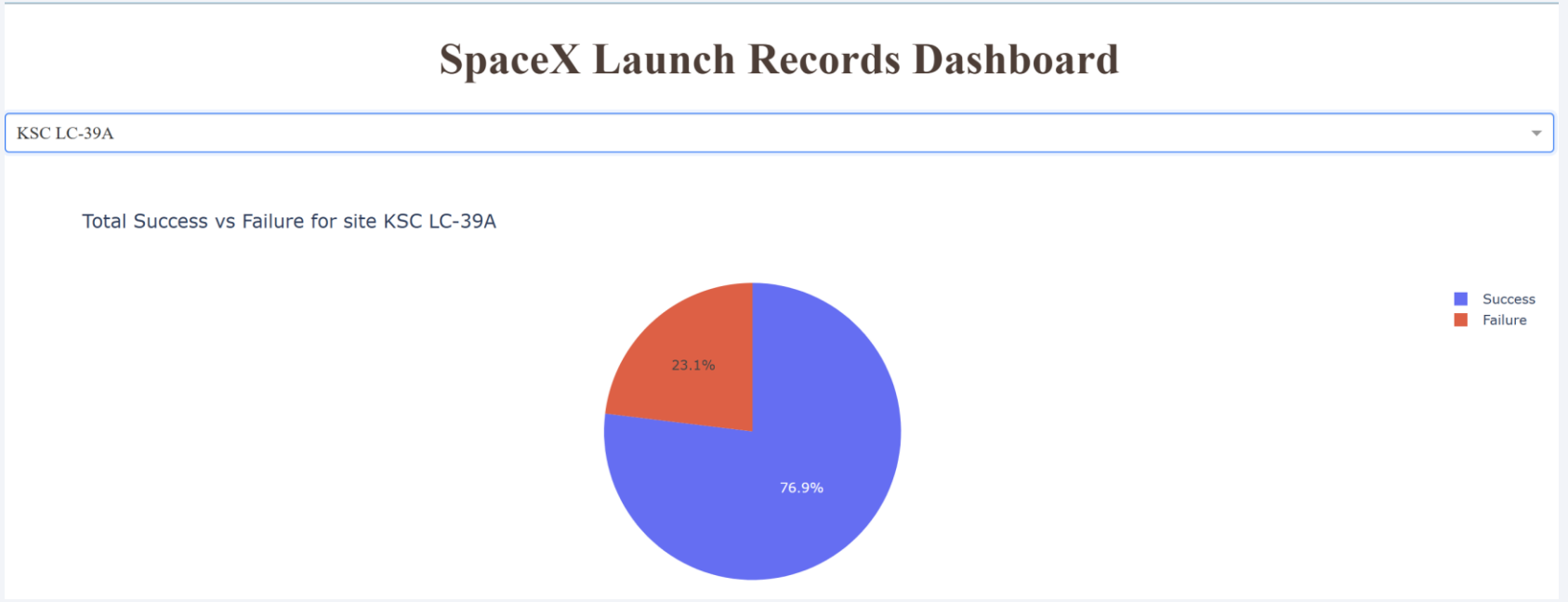
Launch Success Count by SpaceX Launch Site

The pie chart displays the count of successful launches for each SpaceX launch site. Each segment represents a launch site's share of total successful missions. The chart clearly shows that **CCAFS LC-40 and KSC LC-39A dominate the success counts**, reflecting their higher usage and more frequent launch operations, while other sites contribute a smaller portion of successful launches. This visualization provides a quick comparison of launch performance across sites.



Highest Launch Success Ratio – KSC LC-39A

- The pie chart shows the launch outcome distribution for **KSC LC-39A**, which has the highest launch success ratio among all sites. The blue segment represents successful launches (about **76.9%**), while the red segment shows failures (about **23.1%**). This indicates that the majority of missions launched from KSC LC-39A were successful, highlighting the site's strong performance and operational reliability compared to other launch sites.



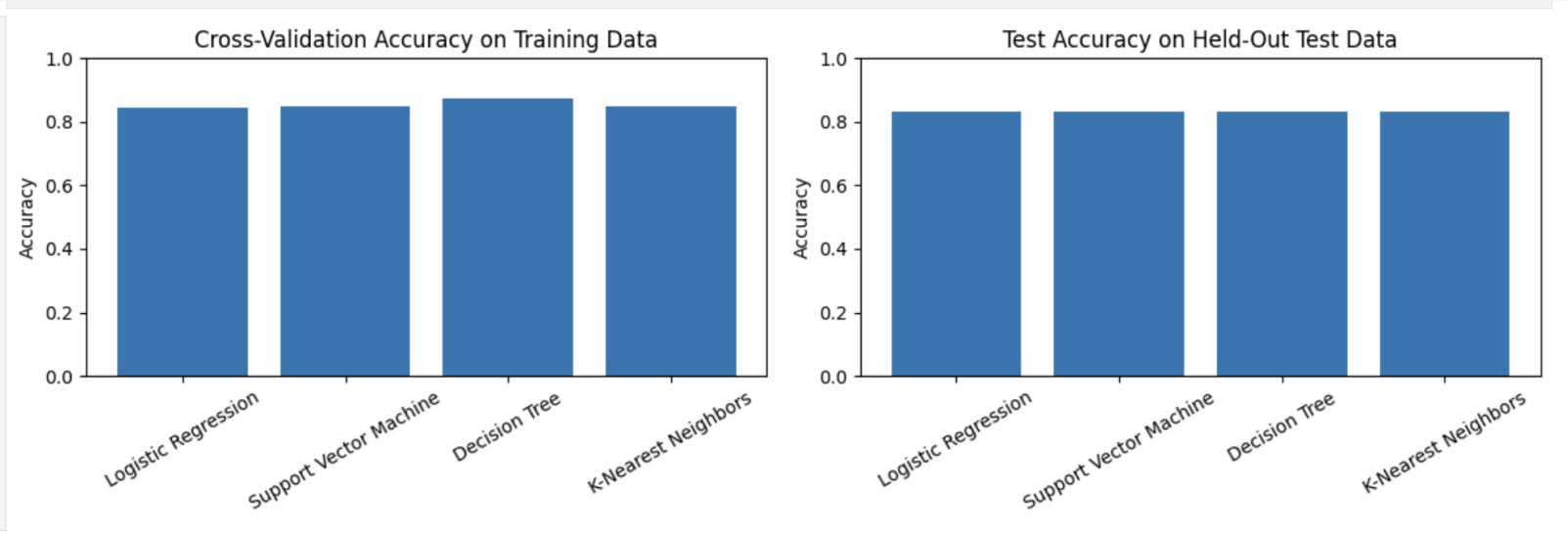
Payload vs. Launch Outcome Scatter Plot (All Launch Sites)

- This scatter plot shows the relationship between **payload mass (kg)** and **launch outcome (Class)** for all launch sites. The x-axis represents payload mass, while the y-axis indicates launch success (1 = success, 0 = failure). Different colors correspond to **booster version categories** (v1.0, v1.1, FT, B4, B5), and the **range slider** at the top allows filtering launches by payload range.
- From the plot, higher success rates are more frequently observed in the **mid to high payload ranges (roughly 3,000–6,000 kg)**, especially for **newer booster versions such as FT, B4, and B5**. Earlier versions (v1.0 and v1.1) show more failures, particularly at lower payload masses. Overall, the visualization suggests that **newer booster versions and moderate-to-high payload ranges are associated with higher launch success**

Section 5

Predictive Analysis (Classification)

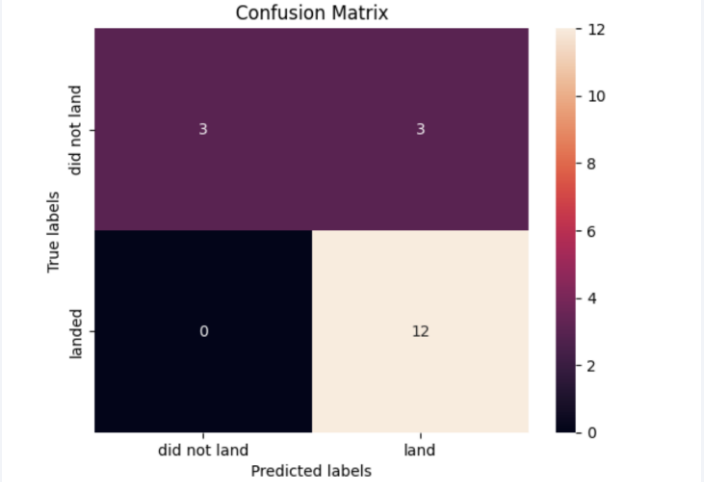
Classification Accuracy



- The bar charts visualize the classification accuracy of all built models.
- The left chart shows cross-validation accuracy on the training data, where the Decision Tree achieved the highest validation accuracy.
- The right chart shows accuracy on the held-out test data, where all models achieved very similar performance (approximately 83%).
- Although the Decision Tree performed slightly better during training, no single model significantly outperformed the others on the test dataset.

Confusion Matrix

```
26]: yhat = tree_cv.predict(X_test)
     plot_confusion_matrix(Y_test,yhat)
```



The confusion matrix of the Decision Tree model, selected as the best performing model during training, shows that the model correctly predicts most successful landings with no false negatives and only a small number of false positives. This indicates strong classification performance, particularly in identifying successful landings.

It is important to note that the confusion matrices of all models were nearly identical. This is mainly due to the small size of the test dataset, where even a few misclassifications can lead to similar accuracy values and confusion matrices across different models.

Overall, while minor differences exist during training, all models demonstrate comparable performance on the test data.

Conclusions

- Launch data was successfully collected, cleaned, and standardized. Landing outcomes were transformed into a binary success label to support both exploratory analysis and machine learning tasks.
- Exploratory Data Analysis (EDA) revealed clear relationships between launch success, payload mass, orbit type, and launch site, highlighting the importance of mission context beyond payload alone.
- Interactive Folium maps showed that launch sites are strategically located near coastlines and key infrastructure while maintaining safe distances from populated areas.
- An interactive dashboard demonstrated that mid-range payloads (≈ 2000–6000 kg) achieve the highest success rates and that KSC LC-39A consistently performs as the most successful launch site.
- Multiple classification models (Logistic Regression, SVM, Decision Tree, and KNN) were trained and tuned using cross-validation to predict landing success.
- The Decision Tree model achieved the highest cross-validation accuracy during training, indicating strong learning capacity on the available features.
- On the held-out test set, all models achieved similar accuracy (≈ 83%), with nearly identical confusion matrices due to the limited test sample size—suggesting no single model clearly outperformed the others in generalization.

Appendix

- Python notebooks for data collection, data wrangling, EDA, feature engineering, and machine learning
- API requests and web scraping scripts used to collect SpaceX launch data
- SQL queries used for filtering, aggregation, grouping, and ranking analysis
- Charts and visual outputs, including EDA plots, Folium maps, and a Plotly Dash dashboard
- Model evaluation results, including accuracy scores and confusion matrices
- Cleaned, labeled, and feature-engineered CSV datasets
- **GitHub Repository (Complete Capstone Project):**

https://github.com/fahimeh-feshki/Data-Science-and-Machine-Learning-Capstone-Project

Thank you!