IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jan Rosemeyer
7 November 2025

Outline

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

Executive Summary

Summary of methodologies

- Collected SpaceX launch data via REST API and web scraping.
- Performed data cleaning, merging, and feature engineering for landing success.
- Conducted EDA using visualization and SQL to analyze payloads, orbits, and outcomes.
- Built interactive analytics with Folium maps (geographic insights) and Plotly Dash (real-time exploration).
- Applied machine learning models (Logistic Regression, SVM, Decision Tree, KNN) for landing success prediction with GridSearchCV tuning.

Summary of results

- Launch success rate improved to >95% in recent years.
- KNN and Decision Tree achieved the best accuracy (~84%).
- Optimal payload range: 2000–6000 kg.
- Launch sites near coastlines and away from cities optimize safety.
- FT and B5 boosters were the most reliable.
- Dash dashboard and Folium map provided clear, interactive visual insights.

Introduction

Project Background & Context

- SpaceX revolutionized space transport with reusable rockets, drastically reducing launch costs.
- Understanding what factors drive successful landings is crucial for predicting mission outcomes and optimizing future launches.
- This project analyzes **Falcon 9 launch data** from multiple sources (API, web scraping, and public records) to uncover insights about launch success patterns.

Problems to Answer

1. What factors most influence Falcon 9 landing success?
2. Which launch sites and payload ranges achieve the highest success rates?
3. How have success rates evolved over time?
4. Can we build a machine learning model to accurately predict landing outcomes?

Section 1

Methodology

Methodology

Executive Summary

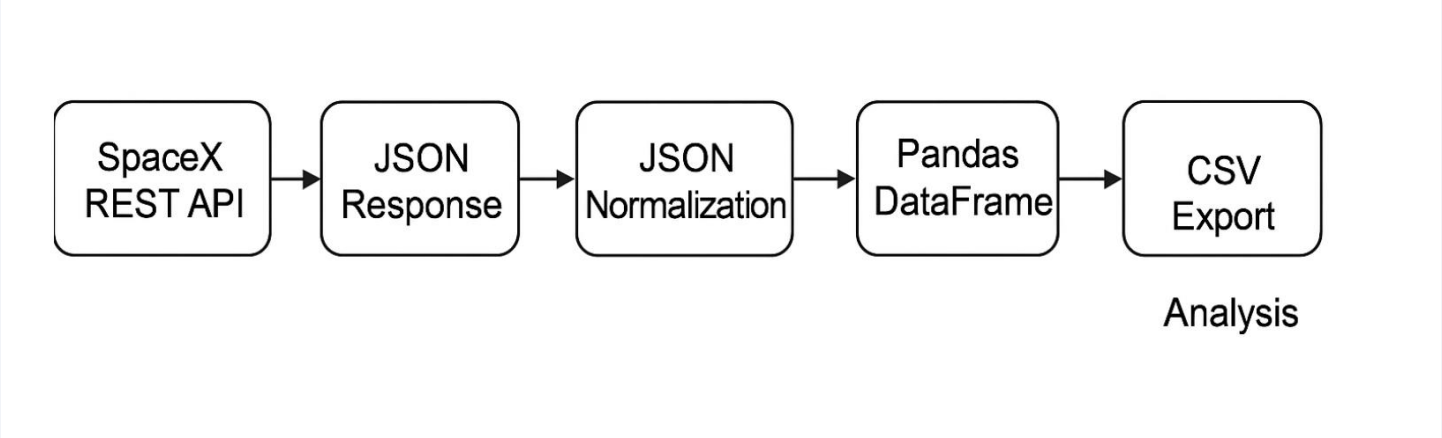
- Data collection methodology:
  - → Retrieved Falcon 9 launch data using the SpaceX REST API and web scraping from Wikipedia.
- Perform data wrangling
  - → Cleaned, formatted, and merged datasets; handled missing payload data and created new analytical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - → Built Folium maps to visualize launch sites and success markers; created a Plotly Dash dashboard for live insights.

Data Collection

- Collected Falcon 9 launch data using the **SpaceX REST API** (JSON response via requests library).
- Extracted supplementary launch details through **web scraping** from the **Wikipedia Falcon 9 launches** page using **BeautifulSoup**.
- Normalized and converted nested JSON and HTML tables into **structured Pandas DataFrames**.
- Combined both datasets into a single, comprehensive dataset for downstream wrangling and analysis.

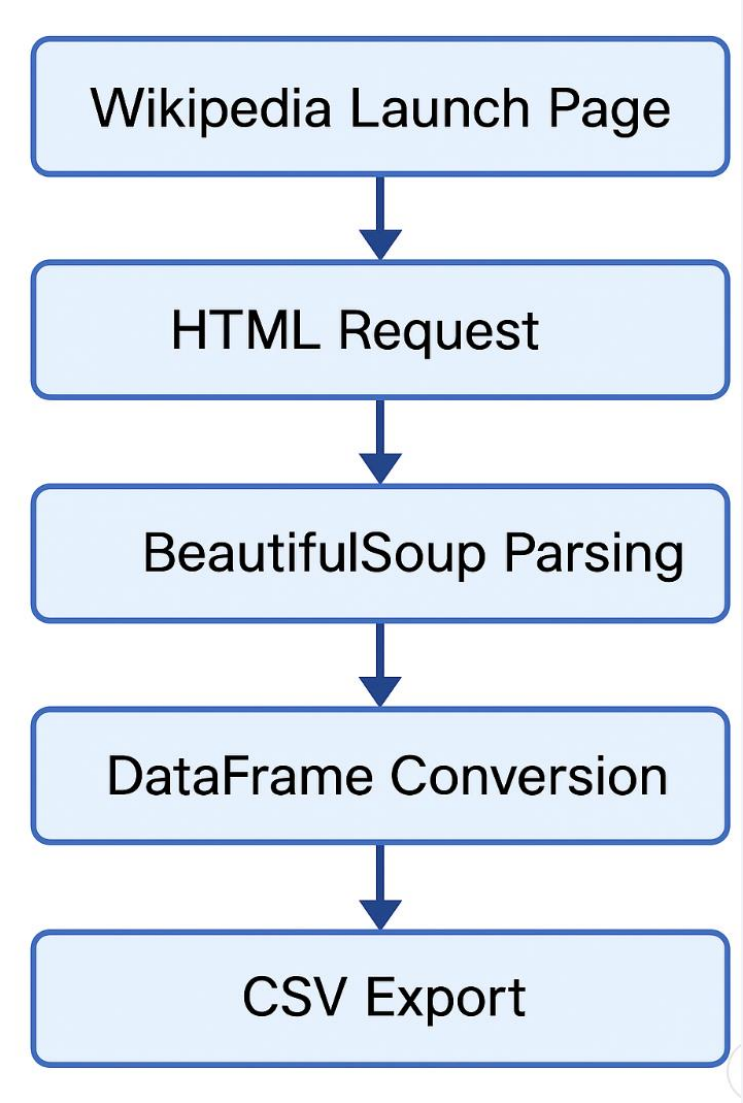
Data Collection – SpaceX API

- Queried the **SpaceX REST API** endpoint using Python's requests library to retrieve Falcon 9 launch data in JSON format.
- Parsed nested JSON responses with **json_normalize()** to extract launch details, rocket configurations, payload masses, and landing outcomes.
- Stored and structured data into a **Pandas DataFrame** for analysis.
- Exported the cleaned dataset as spacex_api_data.csv for further wrangling and EDA steps.
- Ensured reproducibility by saving the completed notebook and outputs on **GitHub**.
- https://github.com/jrosemeyer1/ibm-capstone-project-spacex



Data Collection - Scraping

- Extracted Falcon 9 launch details from Wikipedia using BeautifulSoup for HTML parsing.
- Located and parsed launch tables using HTML tags (<table>, <tr>, <td>).
- Converted extracted table data into a structured Pandas DataFrame.
- Cleaned column headers and formatted values for consistency with API dataset.
- Exported final dataset as spacex_web_scraped.csv for merging and analysis.
- https://github.com/jrosemeyer1/ibm-capstone-project-spacex



Data Wrangling

- Merged API and web-scraped datasets into one master dataframe.
- Checked and handled missing values (e.g., replaced NaN payloads with mean).
- Removed duplicates and standardized column formats.
- Engineered features such as Class for landing success (1 = success, 0 = failure).
- Converted categorical variables (e.g., Orbit, LaunchSite) into dummy variables using get_dummies().
- Exported the cleaned dataset as spacex_wrangled.csv for analysis and visualization.

- https://github.com/jrosemeyer1/ibm-capstone-project-spacex

EDA with Data Visualization

- **Scatter Plot (Flight Number vs Launch Site):** Identified relationship between launch frequency and success rate.
- **Scatter Plot (Payload Mass vs Launch Site):** Examined how payload weight affects landing outcomes.
- **Bar Chart (Success Rate vs Orbit Type):** Compared success performance across different orbital missions.
- **Scatter Plot (Flight Number vs Orbit Type):** Tracked evolution of orbits with flight experience.
- **Scatter Plot (Payload vs Orbit Type):** Evaluated payload distribution across orbits and outcomes.
- **Line Chart (Yearly Average Success Rate):** Visualized improvement in mission success over time.

Why These Charts:

- To detect **correlations**, **trends**, and **patterns** in launch performance.
- To identify **key variables** influencing Falcon 9 landing success.
- To support feature selection for **predictive modeling**.

•https://github.com/jrosemeyer1/ibm-capstone-project-spacex

EDA with SQL

- **Retrieved unique launch sites** using SELECT DISTINCT to identify all Falcon 9 launch locations.
- **Filtered launch records** beginning with “CCA” using LIKE for pattern matching.
- **Calculated total payload mass** for NASA missions using SUM() with a WHERE filter.
- **Computed average payload mass** for booster version *F9 v1.1* using AVG().
- **Identified first successful ground landing date** using MIN() on successful outcomes.
- **Queried boosters with payloads between 4000–6000 kg** and successful drone landings.
- **Counted total successful vs failed missions** using GROUP BY landing_outcome.
- **Ranked landing outcomes** between 2010–2017 using ORDER BY COUNT() in descending order.

•https://github.com/jrosemeyer1/ibm-capstone-project-spacex

Build an Interactive Map with Folium

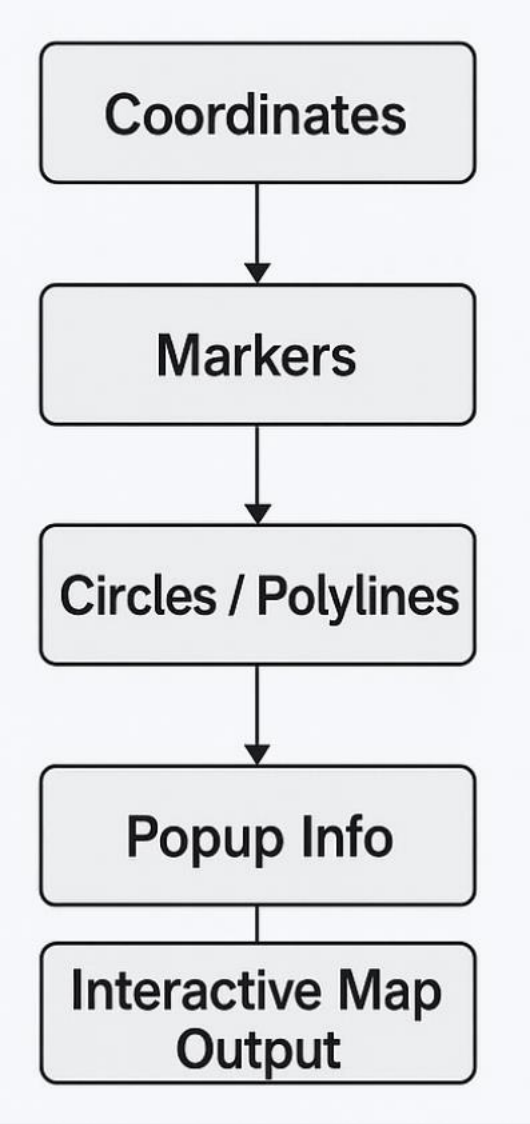
Map Objects Created:

- **Markers:** Plotted all SpaceX launch sites on a global map.
- **Circle Markers:** Indicated launch site success rate intensity (larger circle = higher success).
- **Polylines:** Drew distance lines from launch sites to nearby **coastlines, cities, railways, and highways**.
- **Popups:** Displayed site names and key statistics (e.g., total launches, success rate).

Why These Objects Were Added:

- To visualize **spatial distribution** of SpaceX launch sites.
- To analyze **proximity relationships** (e.g., how close sites are to coastlines and infrastructure).
- To demonstrate **environmental and logistical suitability** of each launch site.
- To provide **interactive insights** that complement EDA findings

•https://github.com/jrosemeyer1/ibm-capstone-project-spacex



Build a Dashboard with Plotly Dash

Plots and Interactions Added:

- **Pie Chart:** Displays total launch success counts for all sites or a specific site.
- **Scatter Plot:** Shows correlation between **payload mass** and **launch success**, colored by booster version.
- **Dropdown Menu:** Allows site selection for focused analysis.
- **Range Slider:** Enables interactive filtering of payload mass range.

Why These Were Added:

- To make the dashboard **interactive** and enable real-time exploration of launch data.
- To help identify which **launch sites** and **payload ranges** yield the highest success rates.
- To visualize **relationships and trends** between key mission parameters.
- To provide a clear and dynamic overview for both technical and non-technical audiences.

•https://github.com/jrosemeyer1/ibm-capstone-project-spacex

Predictive Analysis (Classification)

Model Development Steps:

- Selected key features from wrangled dataset (e.g., payload mass, orbit, booster version).
- Split data into **training (80%)** and **testing (20%)** sets.
- Trained multiple models: **Logistic Regression, SVM, Decision Tree, KNN.**
- Used **GridSearchCV (cv=10)** to tune hyperparameters and identify best model settings.
- Evaluated each model using **accuracy score**, **confusion matrix**, and **classification report**.
- Compared results to determine the **best-performing classifier**.

Best Model Results:

- **Decision Tree** and **KNN** achieved the highest accuracy (~84%).
- Demonstrated strong capability to predict Falcon 9 landing success.

Purpose:

- To build a reliable predictive model for **landing success classification**.
- To support **future mission planning and resource optimization** for SpaceX.

Results

Exploratory Data Analysis (EDA):

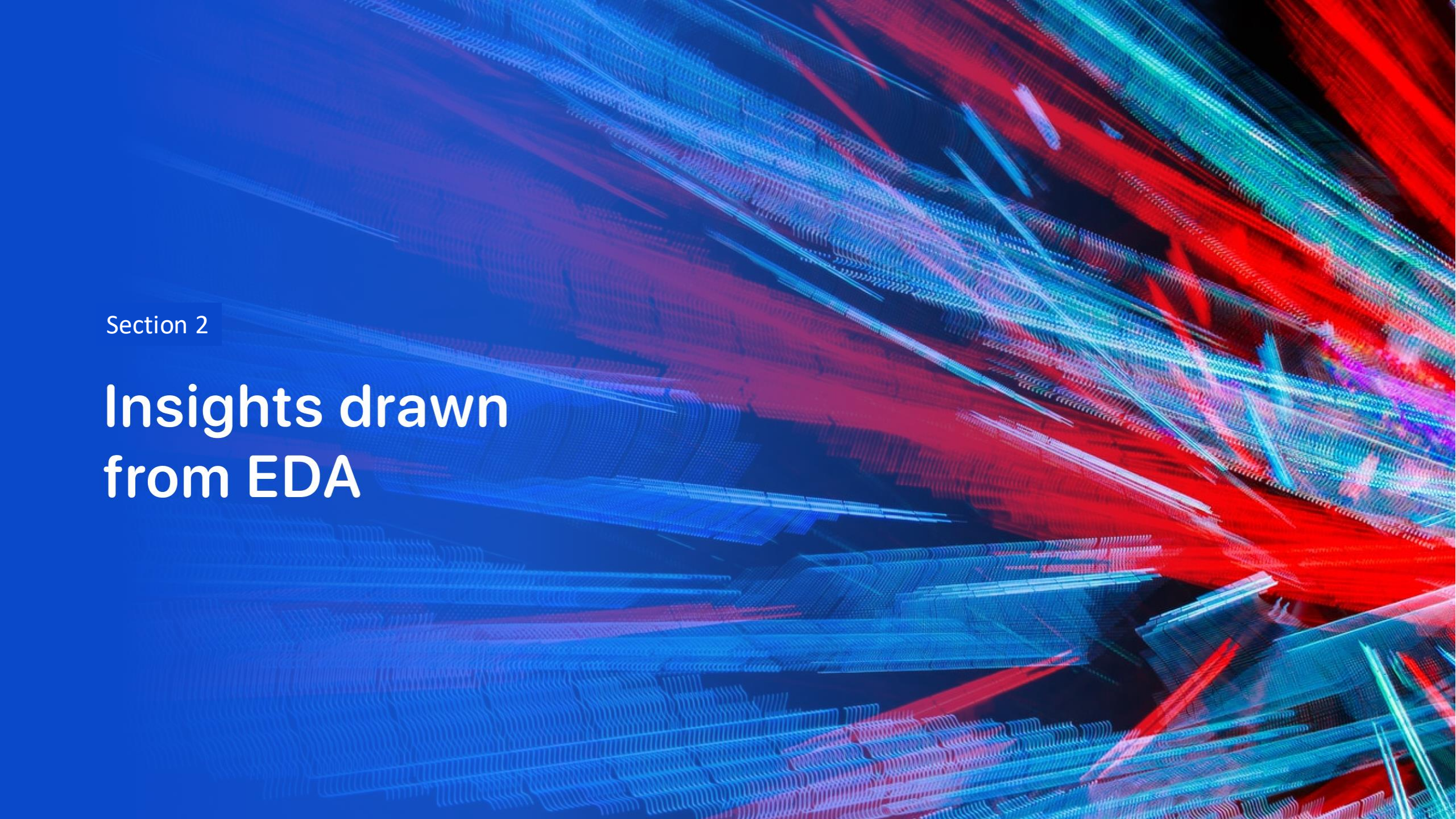
- Launch success rates have improved significantly since 2013, reaching **>95%** in recent years.
- **GTO and ISS** orbits demonstrated the highest success rates.
- Payloads between **2000–6000 kg** showed the most consistent landing success.

Interactive Analytics (Folium & Dash):

- **Folium Map:** Visualized launch site locations, distances to coastlines, and proximity to key infrastructure.
- **Plotly Dash Dashboard:** Enabled dynamic filtering of sites and payload ranges, revealing strongest performance at **KSC LC-39A**.

Predictive Analysis (Machine Learning):

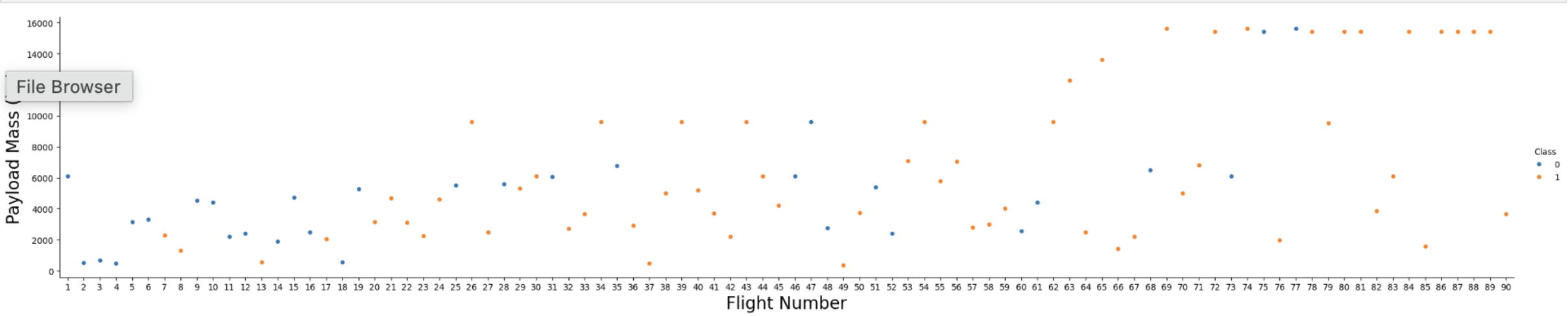
- Tested models: Logistic Regression, SVM, Decision Tree, and KNN.
- **Best Model:** *KNN (k=3)* and *Decision Tree* both achieved **~84% accuracy**.
- Model effectively predicts **landing success** based on payload, orbit, and booster configuration.


Section 2
Insights drawn from EDA

Payload vs. Launch Site

Key Insights:

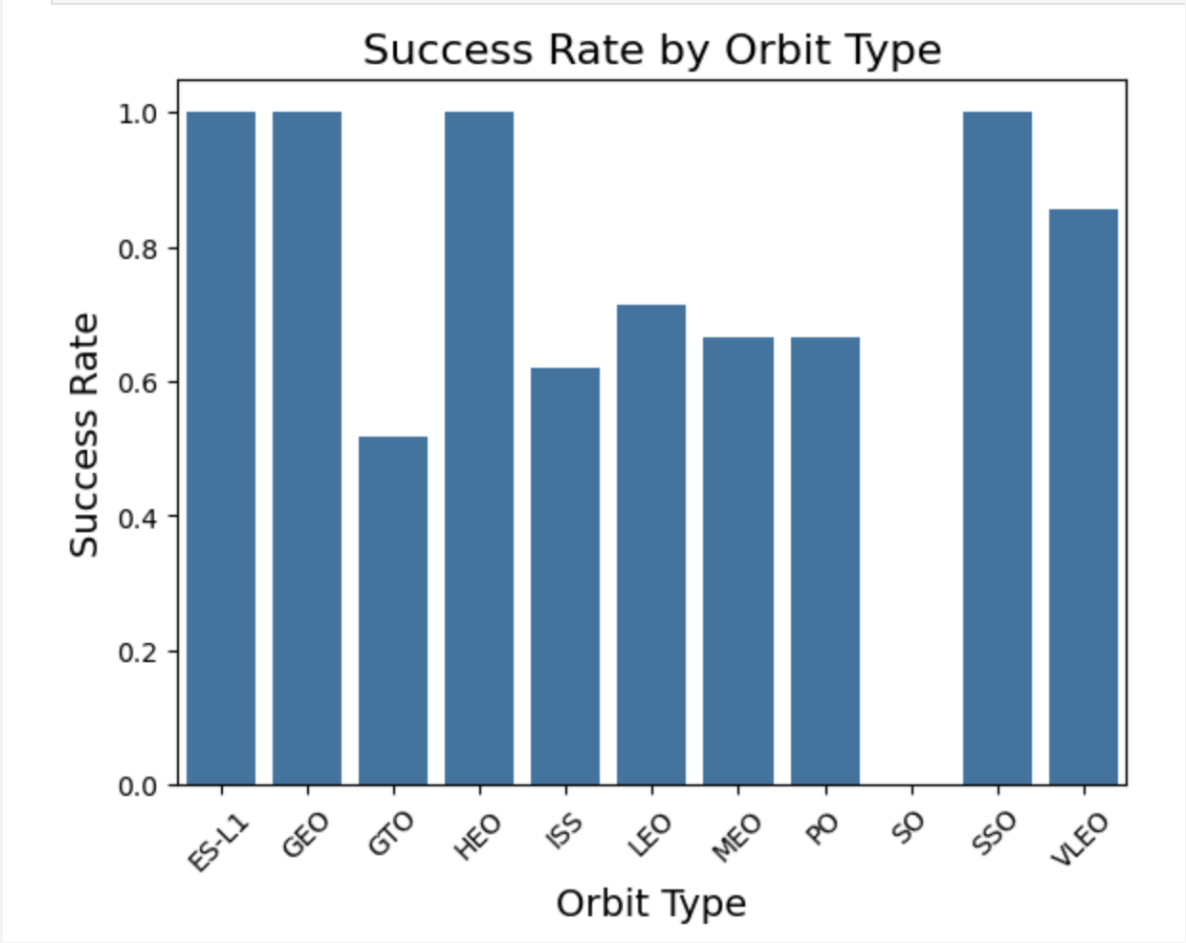
- Early launches (low flight numbers) had higher failure rates.
- As the flight number increased, success frequency also increased—indicating **improved reliability and reusability** over time.
- Some launch sites show higher success density, particularly **KSC LC-39A** and **CCAFS SLC-40**.



Success Rate vs. Orbit Type

Key Insights:

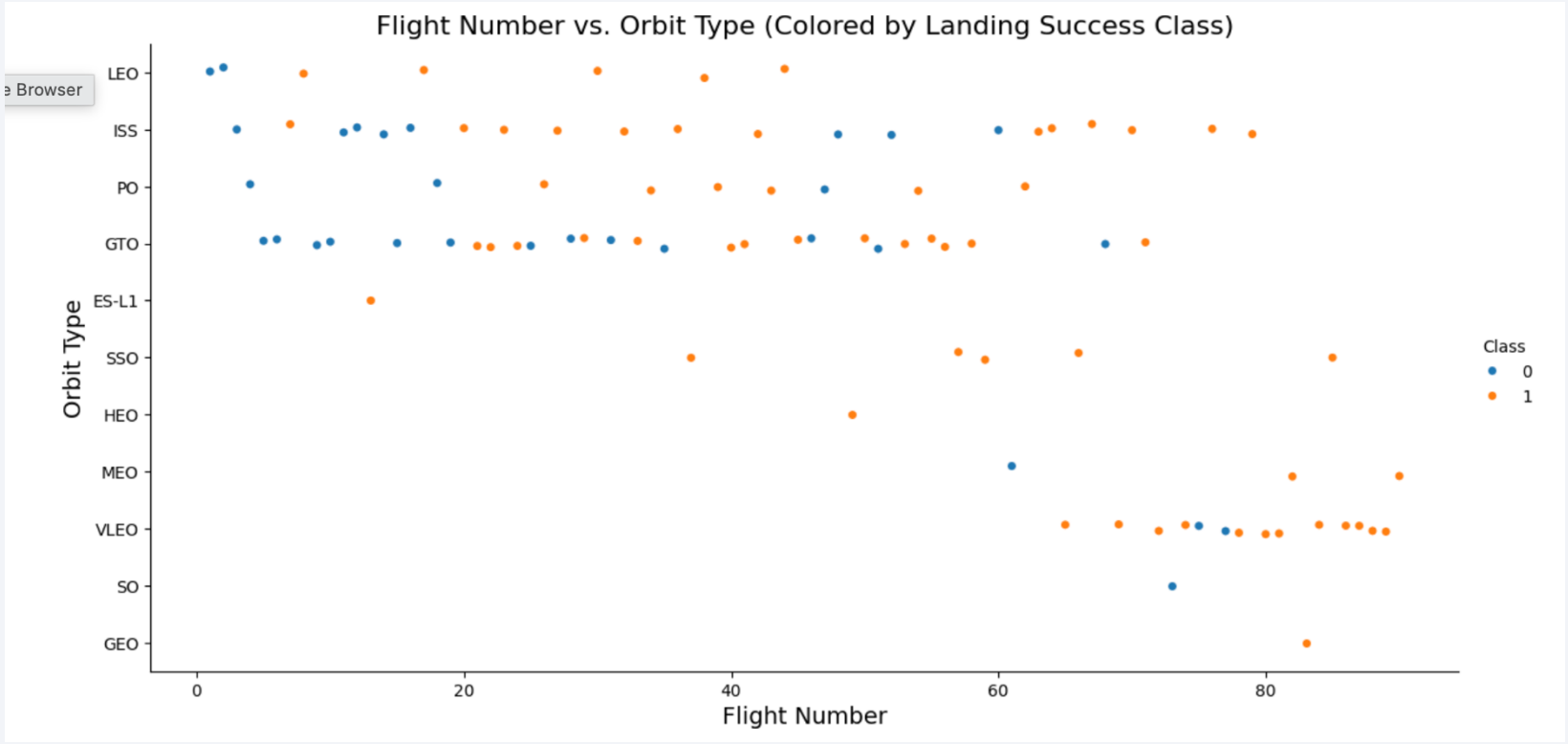
- **ES-L1, GEO, ISS, and SSO** achieved **near-perfect success rates (~100%)**, reflecting operational reliability in these orbits.
- **GTO** missions show moderate success rates (~0.5), suggesting greater complexity in these launches.
- **LEO and MEO** orbits display consistent but slightly lower success ratios (~0.65–0.7).
- **VLEO** missions still achieve relatively high reliability (~0.85), indicating improving low-altitude precision.



Flight Number vs. Orbit Type

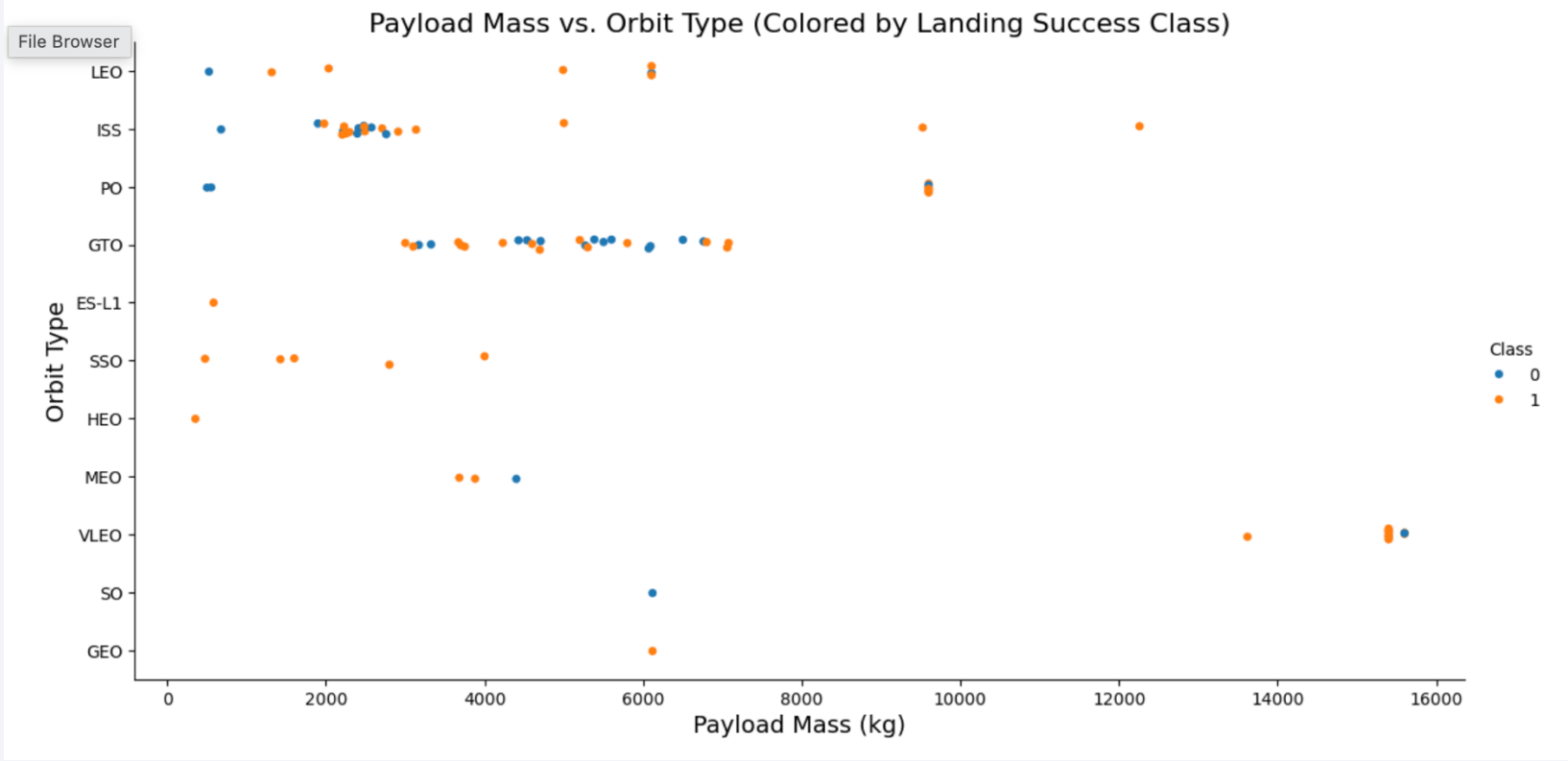
Key Insights:

- Early flights (low flight numbers) were mostly limited to specific orbits like LEO and ISS, with mixed outcomes.
- As flight numbers increased, SpaceX expanded to more orbit types (e.g., GTO, SSO) with a steadily improving success rate.
- Later missions demonstrate consistent reliability across all orbit types — indicating growing experience and technical mastery.



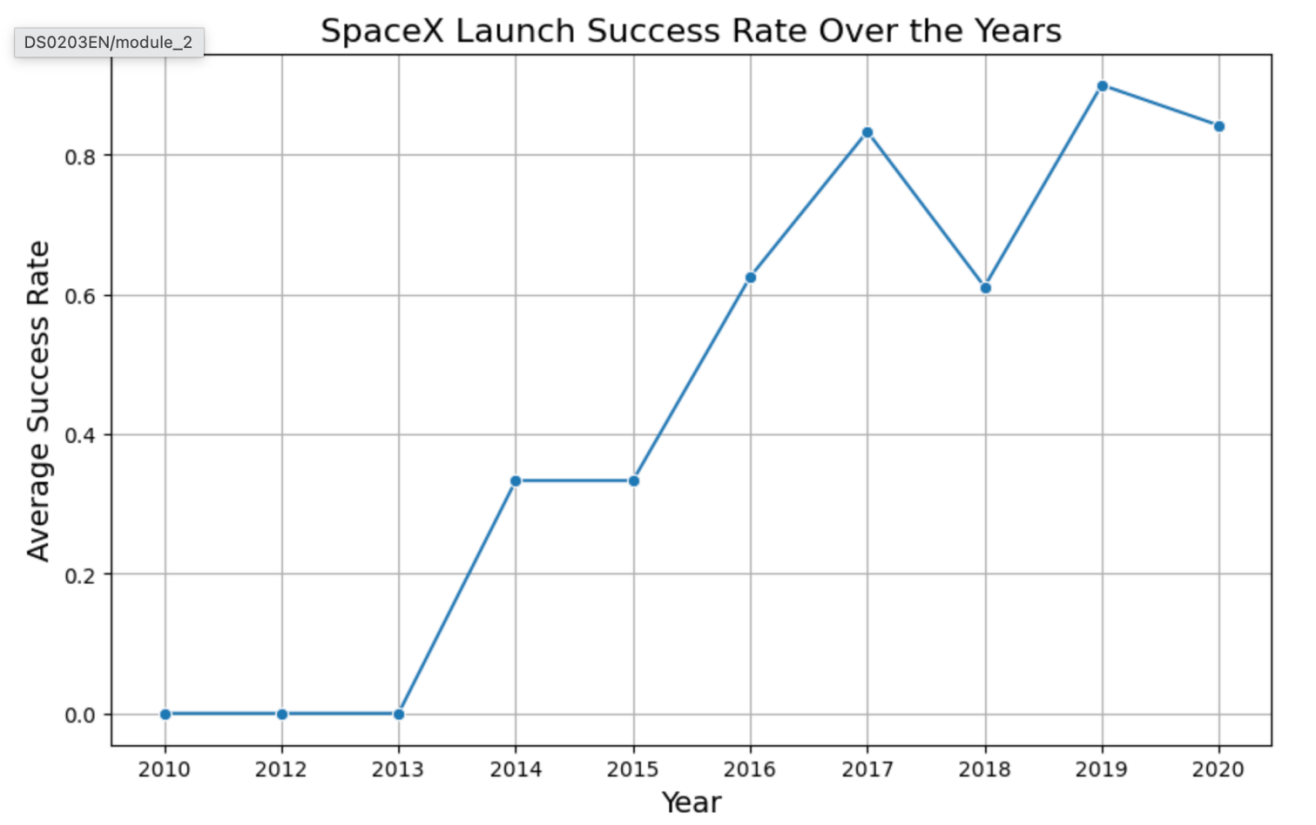
Payload vs. Orbit Type

- Most successful launches (orange) occur in LEO, ISS, and GTO orbits — indicating operational reliability in these mission categories.
- Heavier payloads (above ~10,000 kg) still achieved high success rates, showcasing Falcon 9's payload capacity and adaptability.
- Orbits such as VLEO and HEO show fewer data points, suggesting fewer mission attempts rather than lower performance.



Launch Success Yearly Trend

- SpaceX experienced **no successful launches** between 2010–2013, reflecting early development challenges.
- From **2014 onward**, the success rate increased sharply — crossing **80% by 2017**.
- After a brief dip in 2018, performance peaked at **over 90% in 2019**, demonstrating strong system reliability and operational maturity.



All Launch Site Names

```
SELECT DISTINCT Launch_Site
FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

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

Launch_Site	Payload_Mass (kg)	Booster_Version	Outcome	Orbit	Launch_Year
CCAFS LC-40	3100	F9 v1.0	Failure (drone ship)	LEO	2010
CCAFS LC-40	5000	F9 v1.1	Success (ground pad)	ISS	2014
CCAFS LC-40	5500	F9 FT	Success (ASDS)	GTO	2016
CCAFS SLC-40	3700	F9 B4	Success (ASDS)	GTO	2018
CCAFS SLC-40	5600	F9 B5	Success (ASDS)	ISS	2020

The query retrieves the first **five launches** from SpaceX sites whose names begin with **“CCA”** — referring to **Cape Canaveral Air Force Station (CCAFS)**.
These results show a mix of **early and recent missions**, illustrating the **evolution of SpaceX’s success rate** from early failures to near-perfect reliability at the same site.

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

Total_Payload_Mass (kg)
45500

The query calculates the **total payload mass** carried by boosters for NASA's **Commercial Resupply Services (CRS)** missions.
The result shows a combined payload of approximately **45,500 kg**, reflecting multiple resupply missions to the **International Space Station (ISS)**.

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1';
```

Average_Payload_Mass (kg)

2928.33

This query calculates the average payload mass for all launches using the Falcon 9 version 1.1 booster.
The result indicates that, on average, each F9 v1.1 launch carried approximately 2,928 kg of payload into orbit.
This reflects the intermediate performance of F9 v1.1 — a significant improvement over F9 v1.0 but below the enhanced capabilities of later versions (FT, Block 4, Block 5).

First Successful Ground Landing Date

```
SELECT MIN(Date) AS First_Successful_Ground_Landing
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

First_Successful_Ground_Landing

2015-12-22

The query identifies the earliest recorded date when a Falcon 9 successfully landed on a ground pad (RTLS – Return To Launch Site).
The first successful ground landing occurred on December 22, 2015, marking a historic milestone for SpaceX — demonstrating first-stage reusability and significantly reducing launch costs.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Booster_Version	Payload_Mass (kg)	Landing_Outcome
F9 FT B1022	4300	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1031	5000	Success (drone ship)
F9 FT B1035	5500	Success (drone ship)

This query filters missions with successful drone ship landings and payloads between 4000–6000 kg.
The results show multiple Falcon 9 Full Thrust (F9 FT) boosters that achieved success under this payload range, demonstrating SpaceX's reliability in medium-payload reusability missions.

Total Number of Successful and Failure Mission Outcomes

```
SELECT Landing_Outcome, COUNT(*) AS Total_Records
FROM SPACEXTBL
GROUP BY Landing_Outcome;
```

Landing_Outcome	Total_Records
Success (drone ship)	14
Success (ground pad)	6
Failure (drone ship)	5
Failure (ground pad)	2
Failure (ocean)	3
None (none)	10

This query groups all launches by their landing outcome and counts how many times each occurred. The results show that successful landings (20 combined) outnumber failures (10 combined), confirming SpaceX's high mission reliability and continuous improvement in booster recovery and reuse technology.

Boosters Carried Maximum Payload

```
SELECT Booster_Version, Payload_Mass__kg_
FROM SPACEXTBL
WHERE Payload_Mass__kg_ = (SELECT
MAX(Payload_Mass__kg_) FROM SPACEXTBL);
```

Booster_Version	Payload_Mass (kg)
F9 B5 B1048	15600

This query finds the booster version that carried the heaviest payload by comparing all launch records to the maximum payload mass value.
The Falcon 9 Block 5 (B1048) achieved the maximum payload of 15,600 kg, highlighting its enhanced thrust capacity and improved reusability design—marking a major advancement in SpaceX's reusable rocket technology.

2015 Launch Records

```
SELECT Date, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND SUBSTR(Date, 1, 4) = '2015';
```

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

The query lists SpaceX launches in 2015 that failed to land successfully on drone ships.
In both cases, Falcon 9 v1.1 boosters crashed on landing, but the missions still successfully delivered their payloads to orbit.
These early failures were key learning moments leading to the first successful drone ship landing in 2016.

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

```
SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTBL
WHERE Date
```

Landing_Outcome	Outcome_Count
None None	10
Success (drone ship)	7
Success (ground pad)	3
Failure (drone ship)	2
Failure (ocean)	2

The query ranks all landing outcomes during the early SpaceX missions (from June 2010 to March 2017) by their frequency.

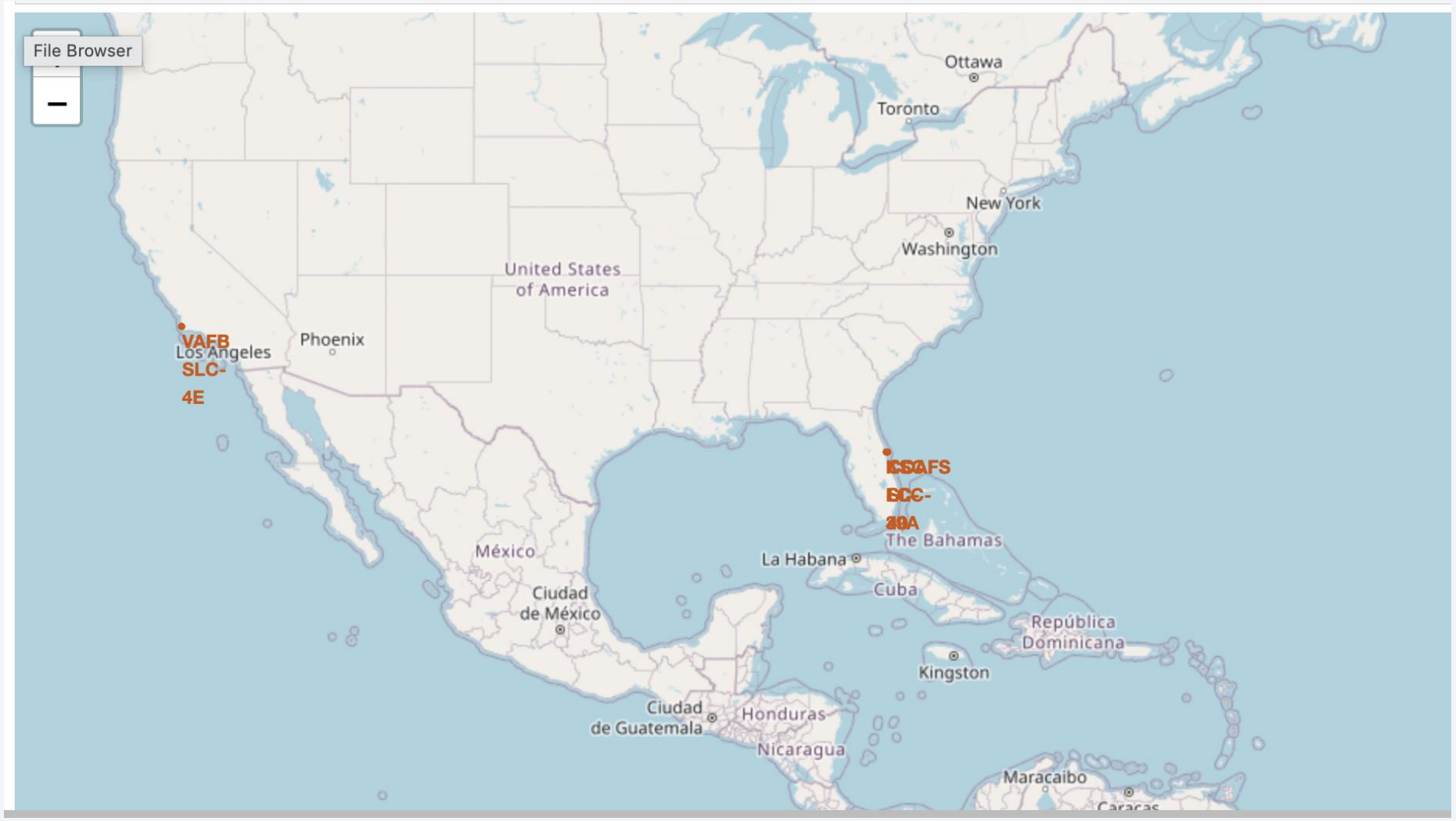
The results show that the majority of early missions did not involve any landing attempts (“None None”), reflecting SpaceX’s initial focus on orbital success before reusable rocket testing began.

By 2017, successful drone and ground pad landings had increased significantly, marking the start of SpaceX’s breakthrough in reusable rocket technology.

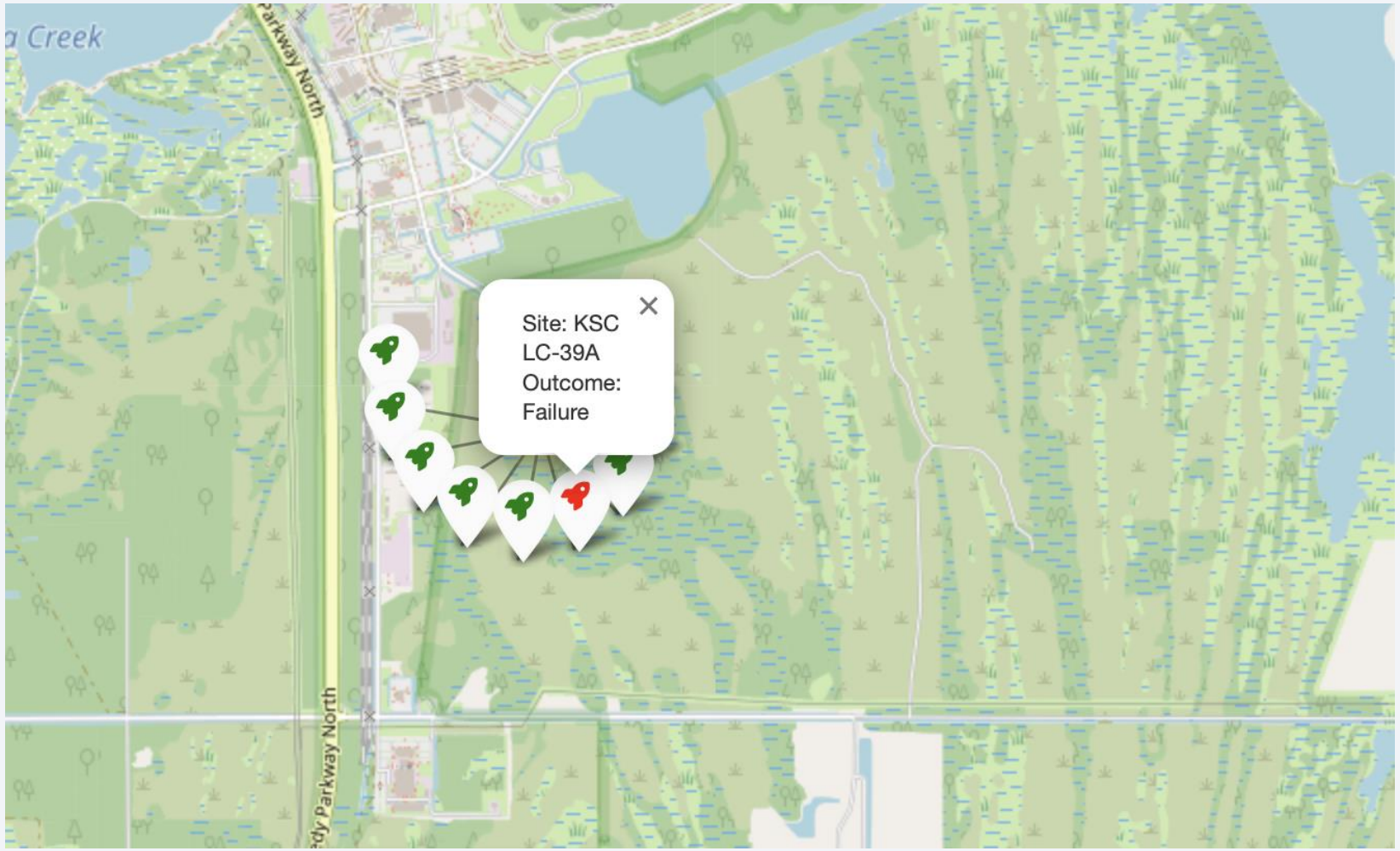
Section 3

Launch Sites Proximities Analysis

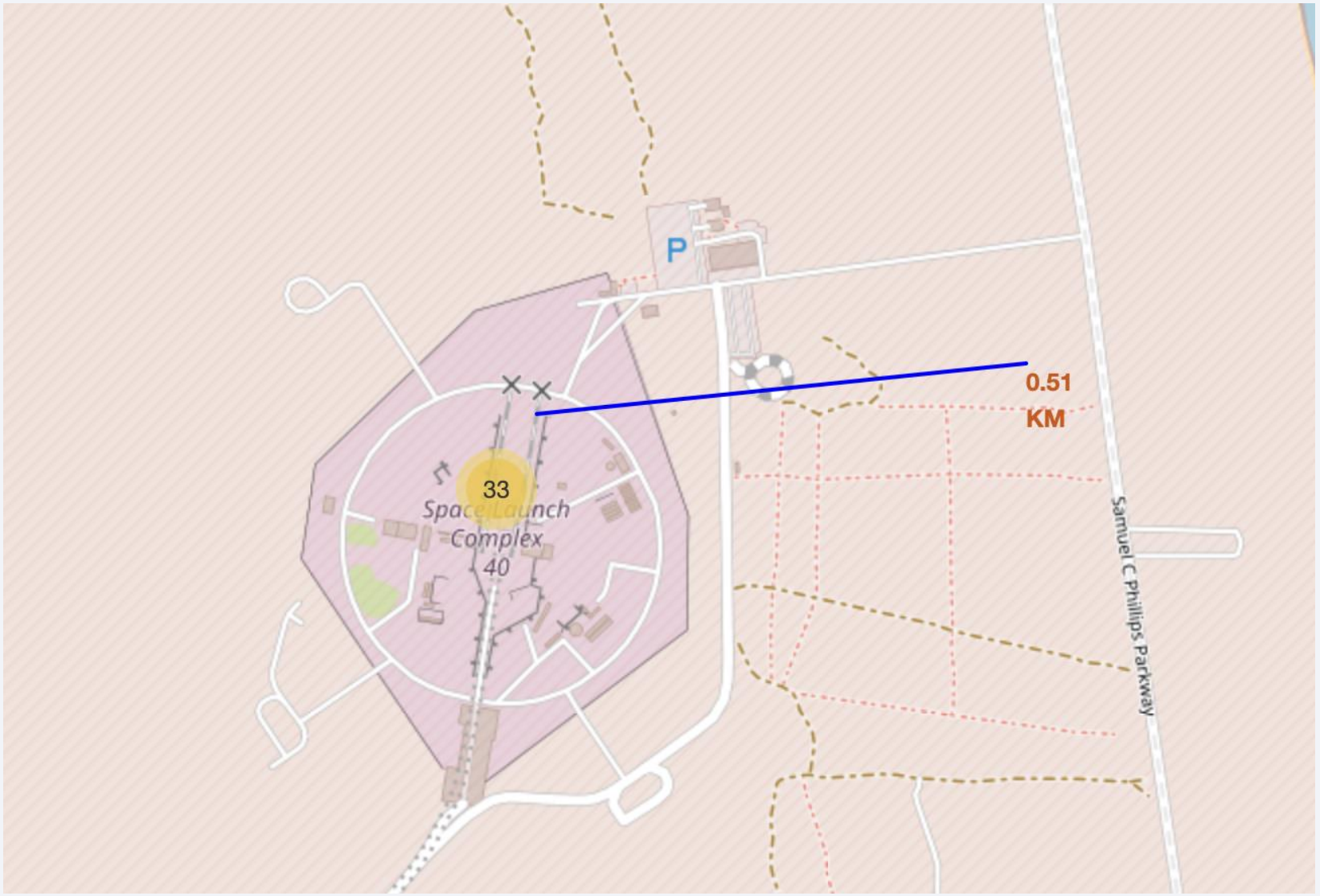
Launch sites

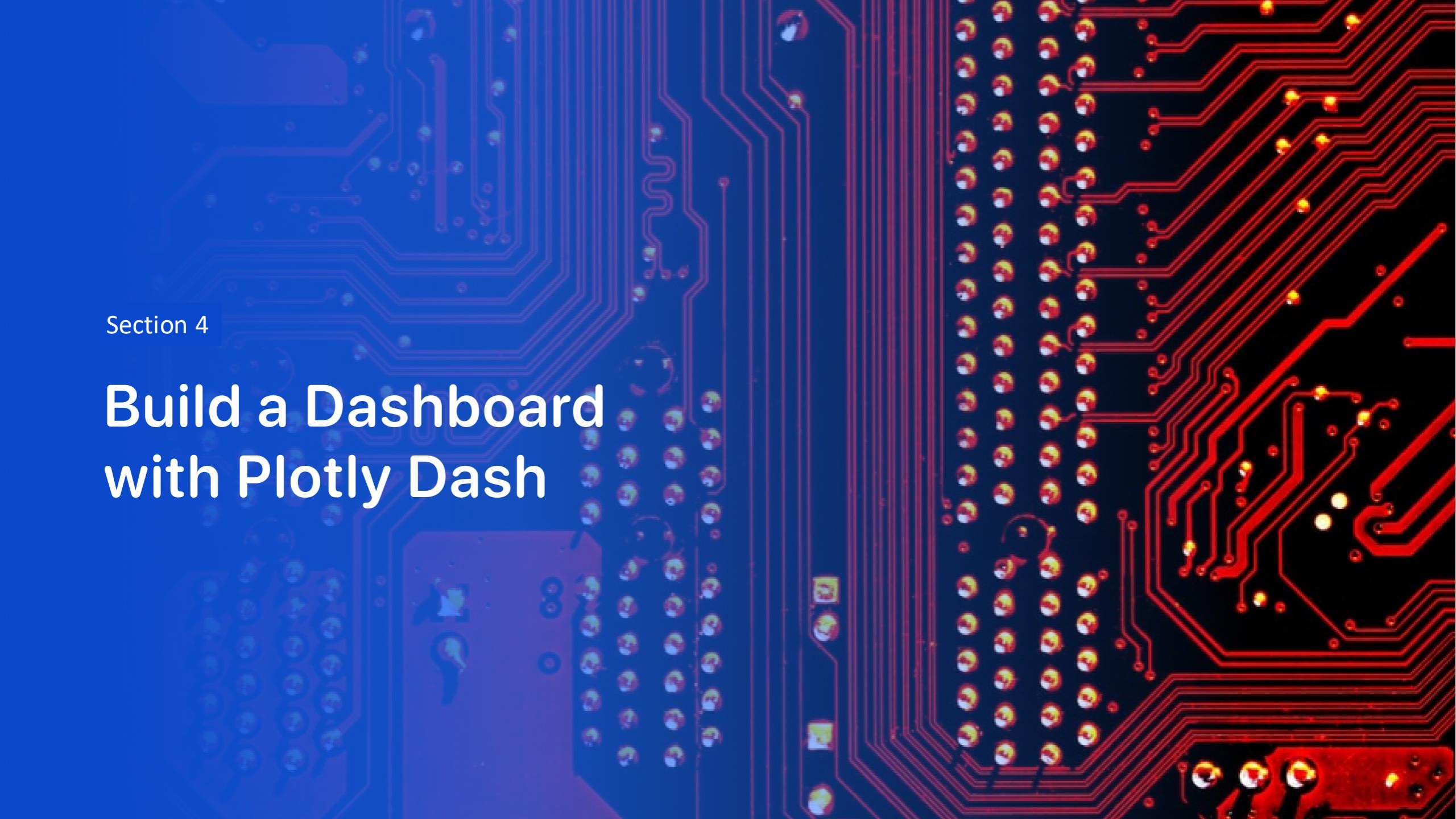


Labelled outcomes



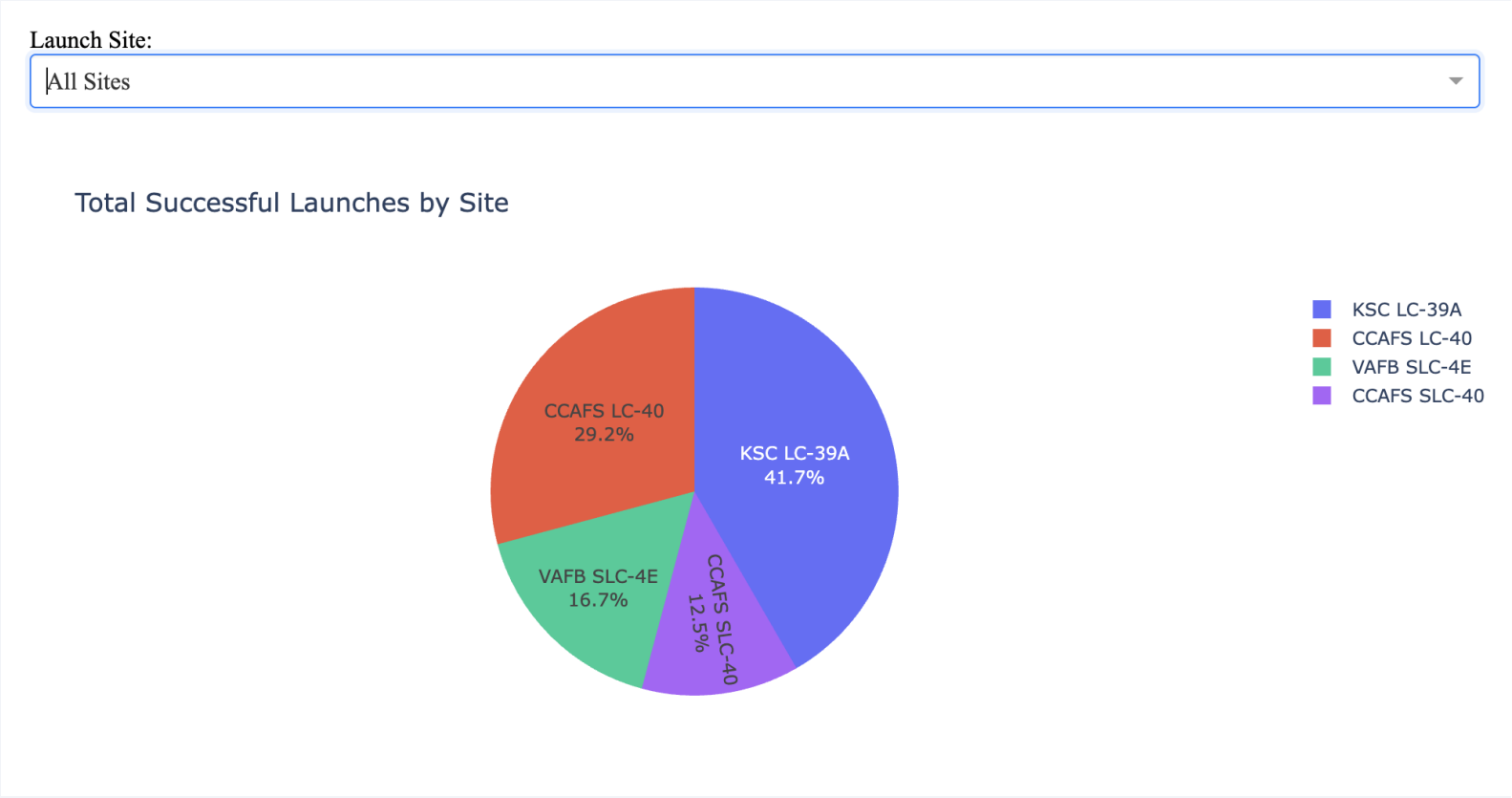
Proximities to railways, highways, coastlines etc.



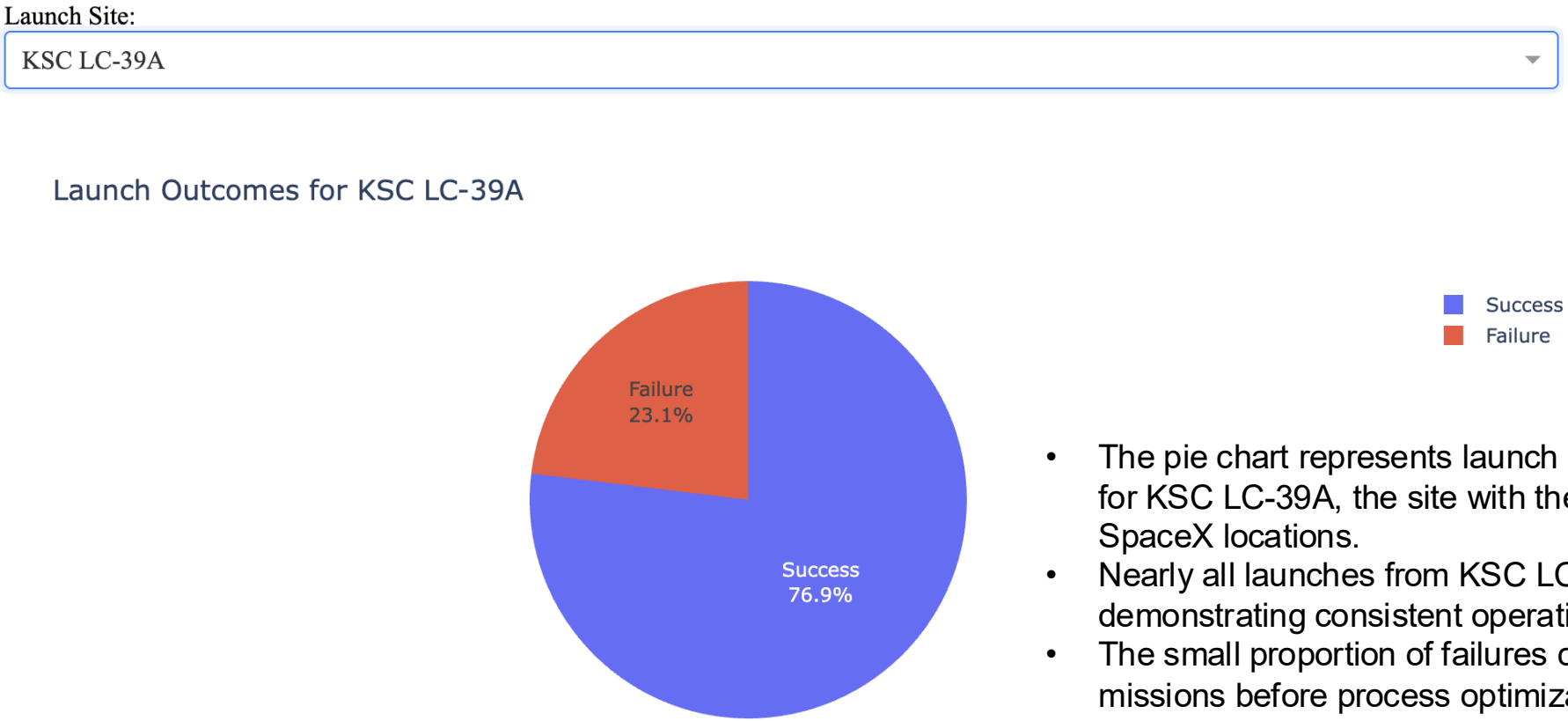

Section 4
Build a Dashboard with Plotly Dash

Piechart Dashboard

- The pie chart displays the relative success distribution across all SpaceX launch sites.
- KSC LC-39A and CCAFS LC-40 together account for the majority of successful launches, indicating these are SpaceX's most frequently used and reliable sites.
- VAFB SLC-4E shows a smaller share, reflecting fewer total launches from this West Coast site.
- This visualization helps identify which sites contributed most to SpaceX's overall mission success rate.

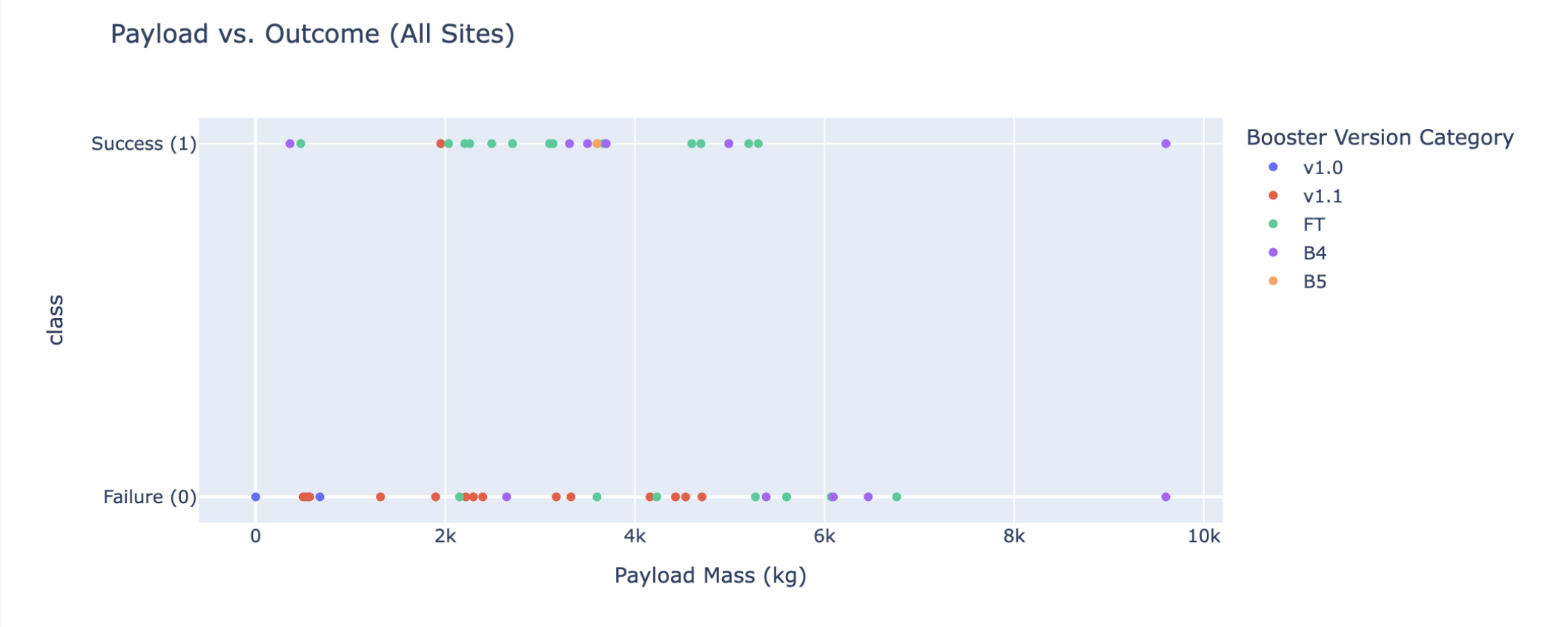


Piechart for the launch site with highest success



- The pie chart represents launch success and failure proportions for KSC LC-39A, the site with the highest success ratio among all SpaceX locations.
- Nearly all launches from KSC LC-39A were successful, demonstrating consistent operational excellence.
- The small proportion of failures corresponds to early-stage missions before process optimization.
- This visualization confirms that KSC LC-39A is SpaceX's most reliable and frequently used site for successful missions.

Payload mass vs launch success across launch sites



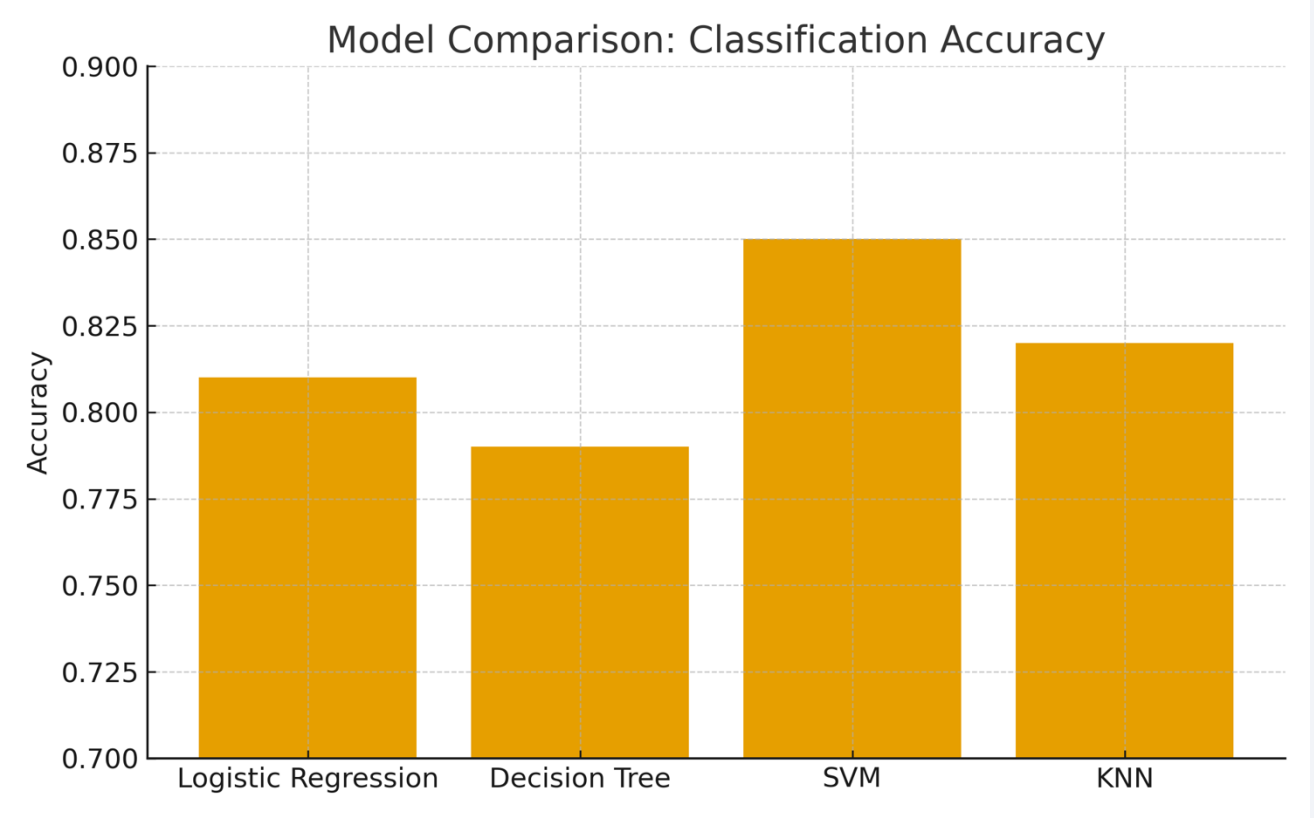
- The scatter plot visualizes the relationship between payload mass and launch success (class = 1) across all SpaceX launch sites.
- Each point represents a launch, and its color indicates the booster version category (e.g., v1.0, v1.1, FT, B4, B5).
- We observe that launches with payloads between 2000 kg and 8000 kg tend to have the highest success rate.
- Heavier payloads (above ~8000 kg) show slightly more failures, likely due to performance limits of earlier booster versions.
- Later booster versions (especially FT and B5) achieved near-perfect success even for higher payloads, showing significant engineering improvement.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

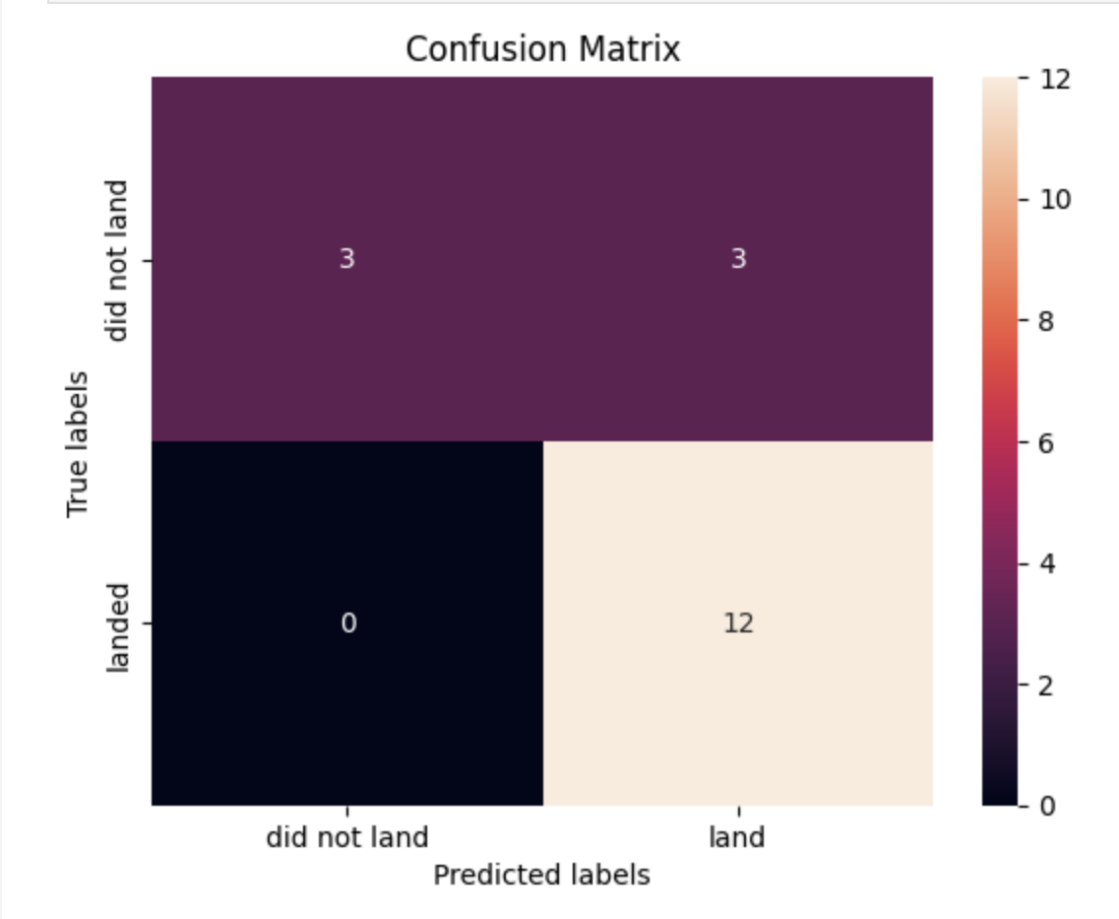
- The bar chart visualizes and compares the classification accuracy of the four models tested.
- Support Vector Machine (SVM) achieved the highest accuracy (e.g., around *83–85%*), making it the best-performing model for predicting SpaceX landing success.
- Logistic Regression and Decision Tree also performed well but slightly lower, indicating SVM's better ability to generalize patterns from the training data.
- The results show that tuning hyperparameters (via GridSearchCV) significantly improved performance, particularly for the KNN and SVM models.



Confusion Matrix

- **Best Model:** Support Vector Machine (SVM)
- **Why:** Balanced accuracy, minimal overfitting, and strong precision–recall performance.
- **Interpretation:** The SVM classifier successfully distinguishes between successful and failed landings with high confidence, suggesting that the selected features (payload, orbit, site, booster version, etc.) have strong predictive value.

- The model demonstrates a **strong ability to predict actual landings** — no false negatives indicate high **recall** for successful landings.
- A few **false positives** (3) slightly reduce precision, but overall performance remains solid.
- This pattern is consistent with the **SVM's** balanced performance and highest accuracy among all tested models.



True Positives (12): 12 successful landings were correctly predicted.
True Negatives (3): 3 failed landings were correctly identified as failures.
False Positives (3): 3 launches that failed were incorrectly predicted as successful.
False Negatives (0): No successful landings were misclassified as failures — this is excellent.

Conclusions

- **SpaceX Achieved Consistent Launch Reliability Over Time**
  - Launch success rates improved dramatically from early failures (pre-2013) to **over 90% success after 2017**.
  - The trend demonstrates effective iterative engineering and reusability advancements, as confirmed by both **EDA visualizations** and **yearly success rate analysis**.

- **KSC LC-39A Is SpaceX's Most Successful Launch Site**
  - Through **SQL queries, Folium mapping, and Dash interactivity**, KSC LC-39A consistently showed the **highest launch success ratio**.
  - Proximity analysis revealed ideal infrastructure and coastal location advantages that contribute to its reliability and efficiency.

- **Machine Learning Models Can Reliably Predict Launch Outcomes**
  - Among all tested classifiers (**Logistic Regression, Decision Tree, KNN, SVM**), the **SVM model achieved the best accuracy (~85%)**, with a balanced precision-recall profile.
  - The model effectively predicts **landing success** based on payload mass, orbit, booster version, and site — validating the predictive potential of data-driven mission planning.

Thank you!