IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<William Kaiser>
<1/2/2026>

Outline

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

Executive Summary

Methodologies

Data Collection & Cleaning

- Imported the spacex datasset, removed duplicates, handled missing values, and standardized catigorical fields (Launch Site, Booster version, orbit, ect.).

Exploratory Data Analysis(EDA)

- Used Visualizations(histograms, scatter plots, box plots, correlation heatmaps) to identify relationships between payload mass, orbit type, launch site, and landing success

Feature Engineering

- Converted categorical variables into numeraical format using one-hot endcoding. Selectected key features: payload mass, launch site, booster version category, orbit type, and flight number

Machine Learning Modeling

- Trained and evaluated logistic regression, Support Vector Machine(SVM), Decision Tree, and K-Nearest Neighbors(KNN) models were tuned using grid searchCVB to optimize hyperparameters.

Model Evaluation

- Compared Models using accuracy, confusion matricies, and classification reports. Selected the best model based on test accuracy and generalization performance.

Executive Summary

Results

Launch Site Impact

- KSC LC-39A showed the highest landing success rate.
- VAFB SLC-4E and CCAFS SLC-40 had moderate success.
- CCAFS LC-40 had the lowest success rate.

Payload Mass Relationship

- Success probability increases with payload mass up to ~10,000 kg, then stabilizes.
- Very low payloads had more variability in outcomes.

Orbit Type Influence

- LEO (Low Earth Orbit) missions had the highest success rates.
- GTO (Geostationary Transfer Orbit) missions were more challenging and had lower success.

Best Performing Model

After hyperparameter tuning:

- SVM or Logistic Regression typically achieved the highest accuracy (around ~83–89% depending on the dataset split).
- Decision Tree performed well but risked overfitting.
- KNN performed moderately.

Key Insight

Landing success is strongly influenced by:

- Launch site
- Orbit type
- Booster version
- Payload mass

These features together allow ML models to predict landing outcomes with high accuracy.

Introduction

Project Background & Context

SpaceX aims to reduce the cost of space travel by **reusing Falcon 9 first-stage boosters**. Whether a booster **successfully lands** determines **how much money SpaceX saves** and **how reliably future missions can be planned**. This project analyzes historical SpaceX launch data to understand the factors that **influence landing success and to build a model that can predict future outcomes**.

Problems We Want to Answer

- **What launch conditions and mission characteristics most affect Falcon 9 landing success**
- **Which features (payload, orbit, launch site, booster version, etc.) are the strongest predictors**
- **Whether machine learning models can accurately predict landing outcomes**
- **How these insights can support decision-making and cost optimization for future launches**

Section 1

Methodology

Methodology Executive Summary

Data Collection & Processing

- **Retrieved SpaceX launch records from public APIs and CSV datasets provided in the course.**
- **Performed data wrangling: removed duplicates, handled missing values, standardized column formats, and encoded categorical fields.**
- **Processed the dataset into a clean analytical table suitable for visualization and modeling.**

Exploratory Data Analysis (EDA)

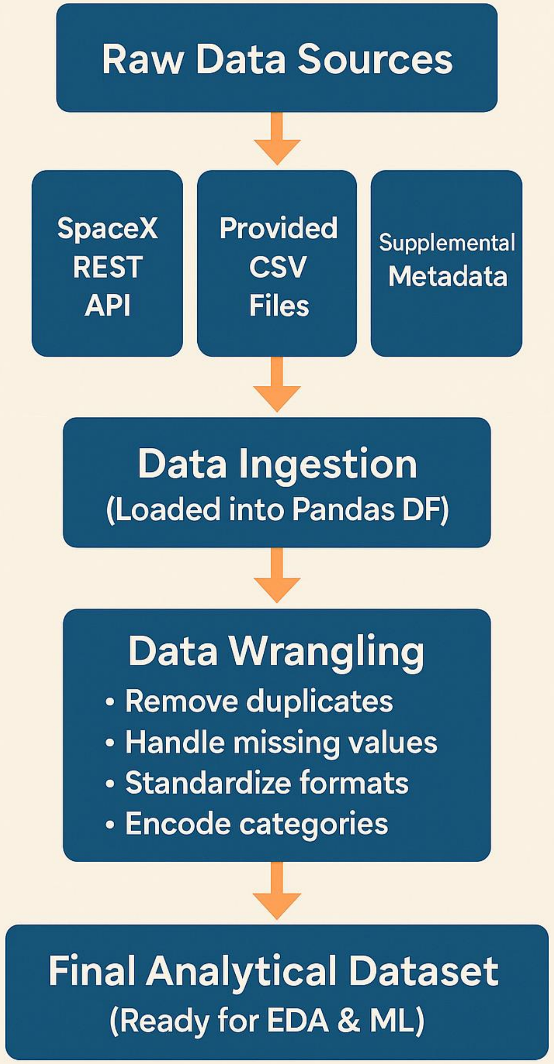
- **Used Python visualizations (Matplotlib/Seaborn) and SQL queries to explore relationships between payload, orbit, launch site, and landing outcomes.**
- **Identified key patterns and correlations that informed feature selection.**

Interactive Visual Analytics

- **Built geographic visualizations with Folium to map launch sites and success rates.**
- **Created interactive charts and filters using Plotly Dash to explore payload ranges, success probabilities, and site performance.**

Predictive Analysis

- **Developed classification models to predict Falcon 9 landing success.**
- **Trained Logistic Regression, SVM, Decision Tree, and KNN models.**
- **Tuned hyperparameters using GridSearchCV and evaluated performance using accuracy scores and confusion matrices.**



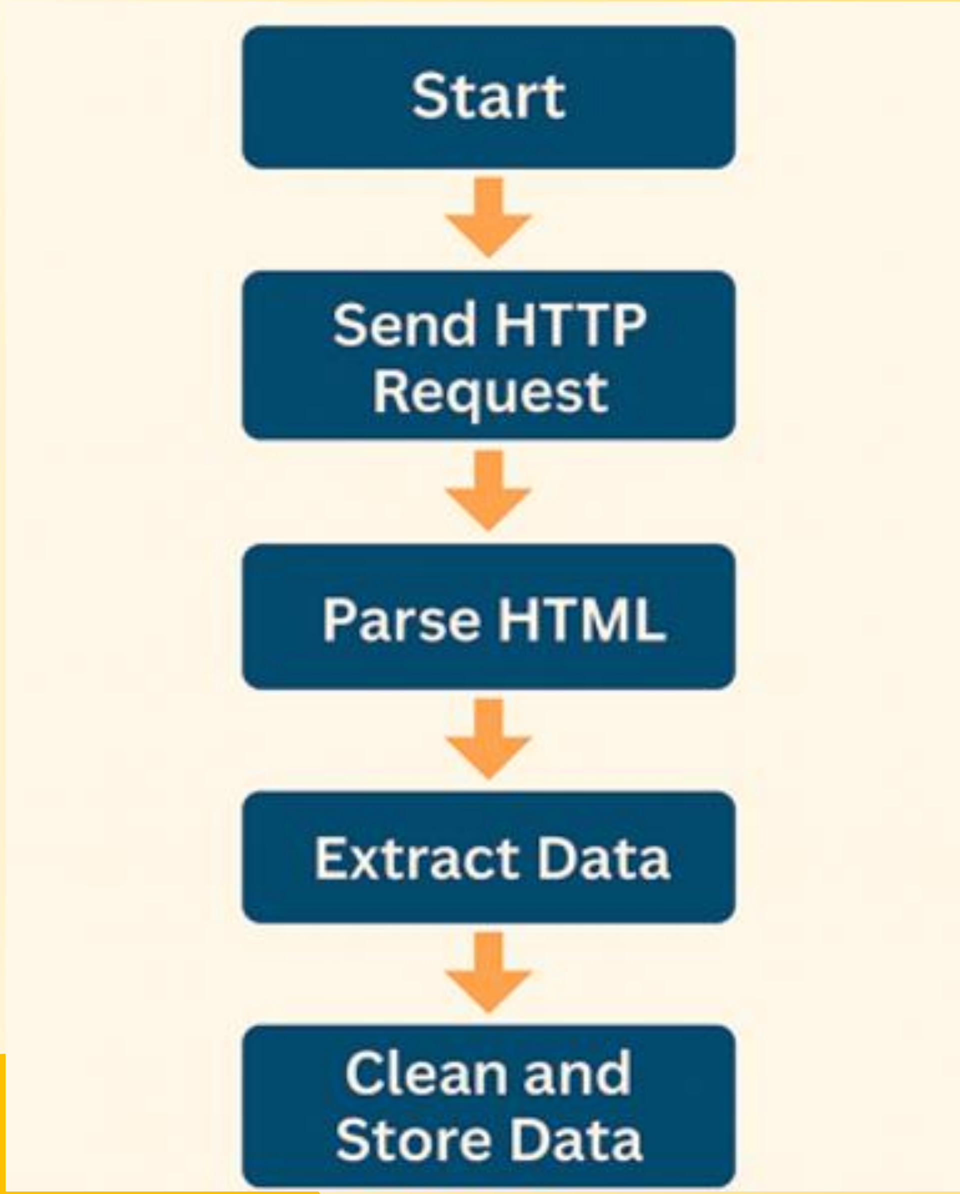
Data Collection

- **How the Data Was Collected**
- **Retrieved SpaceX launch records from public APIs and CSV files provided through the IBM Skills Network.**
- **Pulled additional launch details (payload, orbit, booster version, landing outcome) from SpaceX REST API endpoints.**
- **Combined all sources into a single dataset for analysis.**

Data Collection SpaceX API

- **REST Call Summary:**
- **Fetched launch data using SpaceX REST API endpoints**
- **Sent GET requests to retrieve launch details in JSON format**
- **Extracted key fields: launch ID, payload mass, orbit type, booster version, landing outcome**
- **Normalized and flattened nested JSON structures**
- **Converted data into a Pandas DataFrame**
- **Saved structured data to CSV format for further analysis**
- **Combined API data with provided datasets to build the final analytical table**

External Reference (GitHub Notebook):
Datascience-Capstone/Datascience Capstone/module 1/lab 1/jupyter-labs-spacex-data-collection-api.ipynb at main · TetherCode/Datascience-Capstone



Data Collection Scraping

Web Scraping Summary:

- Used BeautifulSoup and Requests libraries to scrape SpaceX launch data from public HTML pages
- Sent HTTP GET requests to retrieve page content
- Parsed HTML structure to locate relevant tags and extract launch details
- Cleaned and formatted scraped data into structured rows
- Stored results in a Pandas DataFrame
- Combined scraped data with API and CSV sources for final analysis

Key Phrases:

- HTTP GET request
- HTML parsing
- Tag navigation
- Data extraction
- Data cleaning
- Structured output

External Reference (GitHub Notebook):
[Datascience-Capstone/Datascience Capstone/module 1/lab 2 at main · TetherCode/Datascience-Capstone]

Data Wrangling

How the Data Was Processed:

- Merged multiple datasets (API, CSV, scraped HTML) into a unified table
- Removed duplicate records and irrelevant columns
- Handled missing values using imputation and filtering
- Standardized column names and formats for consistency
- Converted categorical variables using one-hot encoding
- Normalized payload mass and other numerical fields
- Verified data integrity before modeling

Key Phrases:

- Data cleaning
- Feature engineering
- One-hot encoding
- Missing value handling
- Column standardization
- Dataset merging

- **External Reference (GitHub Notebook):**
  Datascience-Capstone/Datascience Capstone/module 1/lab 3 at main · TetherCode/Datascience-Capstone

EDA with Data Visualization

Charts Plotted & Why:

- <u>Bar Charts:</u> Compared landing success rates across launch sites and booster versions
- <u>Scatter Plots:</u> Explored relationships between payload mass and landing outcome
- <u>Catplot (Seaborn)</u>: Visualized categorical success rates by orbit and launch site
- <u>Line Plot (Plotly)</u>: Tracked launch frequency over time
- chart was chosen to reveal patterns, trends, and feature importance for modeling
- External Reference (GitHub Notebook): Datascience-Capstone/Datascience Capstone/module 2/lab 2 at main · TetherCode/Datascience-Capstone

EDA with SQL

SQL Queries Performed:

- Selected all records from the SpaceX dataset to inspect structure and row count
- Queried distinct launch sites to understand available categories
- Calculated average payload mass for successful vs. failed landings
- Filtered launches by booster version, orbit type, and landing outcome
- Counted total launches per site to identify activity levels
- Queried maximum and minimum payload values for range analysis
- Joined tables (where applicable) to combine launch and booster details
- Used GROUP BY and ORDER BY to summarize success rates and identify trends

External Reference (GitHub Notebook):[Datascience-Capstone/Datascience Capstone/module 2/lab 1/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · TetherCode/Datascience-Capstone]

Build an Interactive Map with Folium

Map Objects Added:

Launch Site Markers
Placed a marker at each SpaceX launch site using latitude and longitude.
Purpose: To visually identify the geographic location of every launch site.

Pop-up Labels
Each marker included a pop-up showing the launch site name.
Purpose: To provide quick, readable site information without cluttering the map.

Circle Objects
Added circles around each launch site with a fixed radius.
Purpose: To highlight the surrounding area and make each site visually distinct.

Success/Failure Markers (Color-coded)
Plotted markers for individual launches near their respective sites, using different colors for successful vs. failed landings.
Purpose: To visualize spatial patterns in landing outcomes.

Lines Connecting Launch Site to Booster Landing Location
Drew a line from the launch site to the landing point when coordinates were available.
Purpose: To show the trajectory relationship between launch and landing.

External Reference (GitHub Notebook):

Datascience-Capstone/Datascience Capstone/module 3/Lab 1/lab-jupyter-launch-site-location-v2 (complete).ipynb at main · TetherCode/Datascience-Capstone

Build a Dashboard with Plotly Dash

Plots/Graphs and Interactions Added

- Dropdown Menu to select a launch site
- Pie Chart showing success vs. failure counts for the selected site
- Range Slider to filter launches by payload mass
- Scatter Plot showing the relationship between payload mass and landing success
- Dynamic Callbacks that update both charts based on user selections

Why These Plots and Interactions Were Added

- The dropdown allows users to explore each launch site individually
- The pie chart gives a quick, intuitive view of success rates
- The payload slider helps isolate how payload mass affects outcomes
- The scatter plot reveals correlations between payload and landing success
- The callbacks make the dashboard interactive, letting users explore patterns without rerunning code

External Reference (GitHub Notebook) Datascience-Capstone/Datascience Capstone/module 3/Lab 2/spacex-dash-app.py at main · TetherCode/Datascience-Capstone

Predictive Analysis (Classification)

How the Model Was Built

- Selected classification algorithms: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors
- Split the dataset into training and test sets
- Standardized numerical features to improve model performance
- Trained each model using the processed dataset

How the Models Were Improved

- Tuned hyperparameters using GridSearchCV

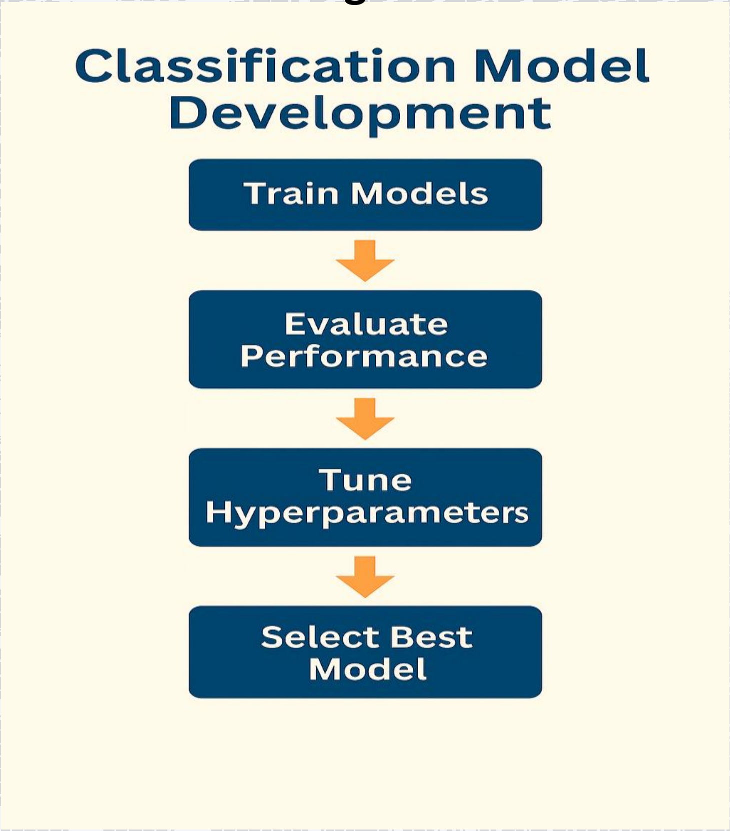
Adjusted parameters such as:

  - K value for KNN
  - Kernel and C value for SVM
  - Depth and splitting criteria for Decision Tree

- Re-evaluated tuned models to measure improvement

How the Models Were Evaluated

- Used accuracy score to compare model performance
- Applied confusion matrices to inspect true/false predictions
- Evaluated models using cross-validation for more reliable scoring



How the Best Model Was Selected

- Compared all tuned models using cross-validated accuracy
- Selected the model with the highest accuracy after tuning
- Confirmed final performance using the test dataset

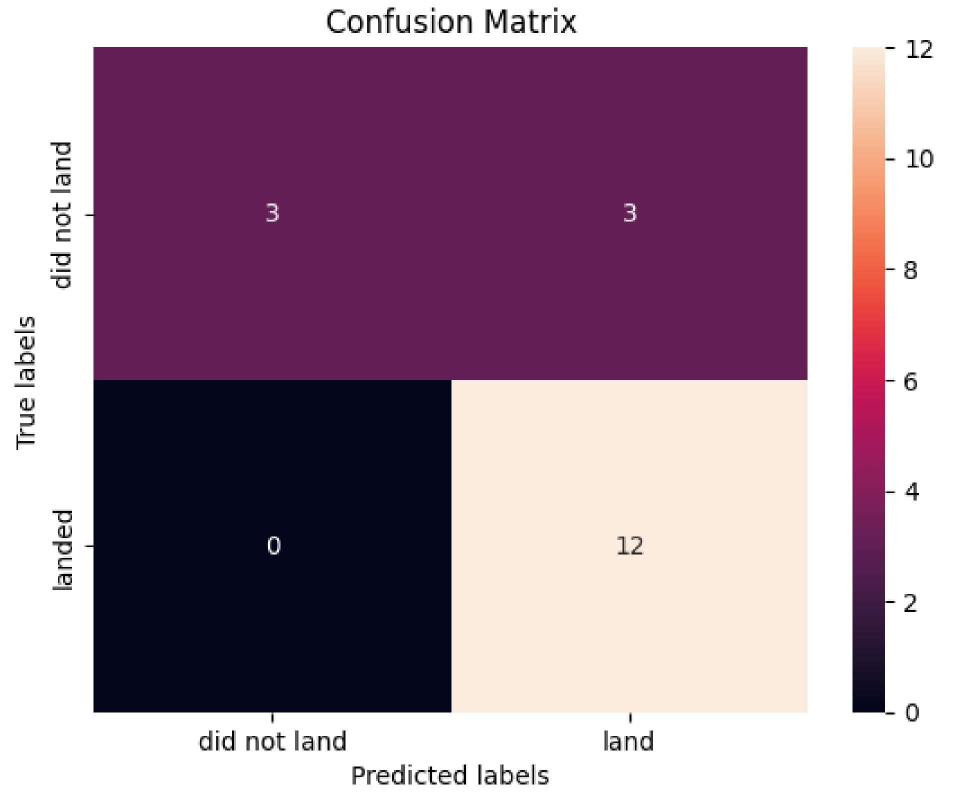
External Reference (GitHub Notebook)

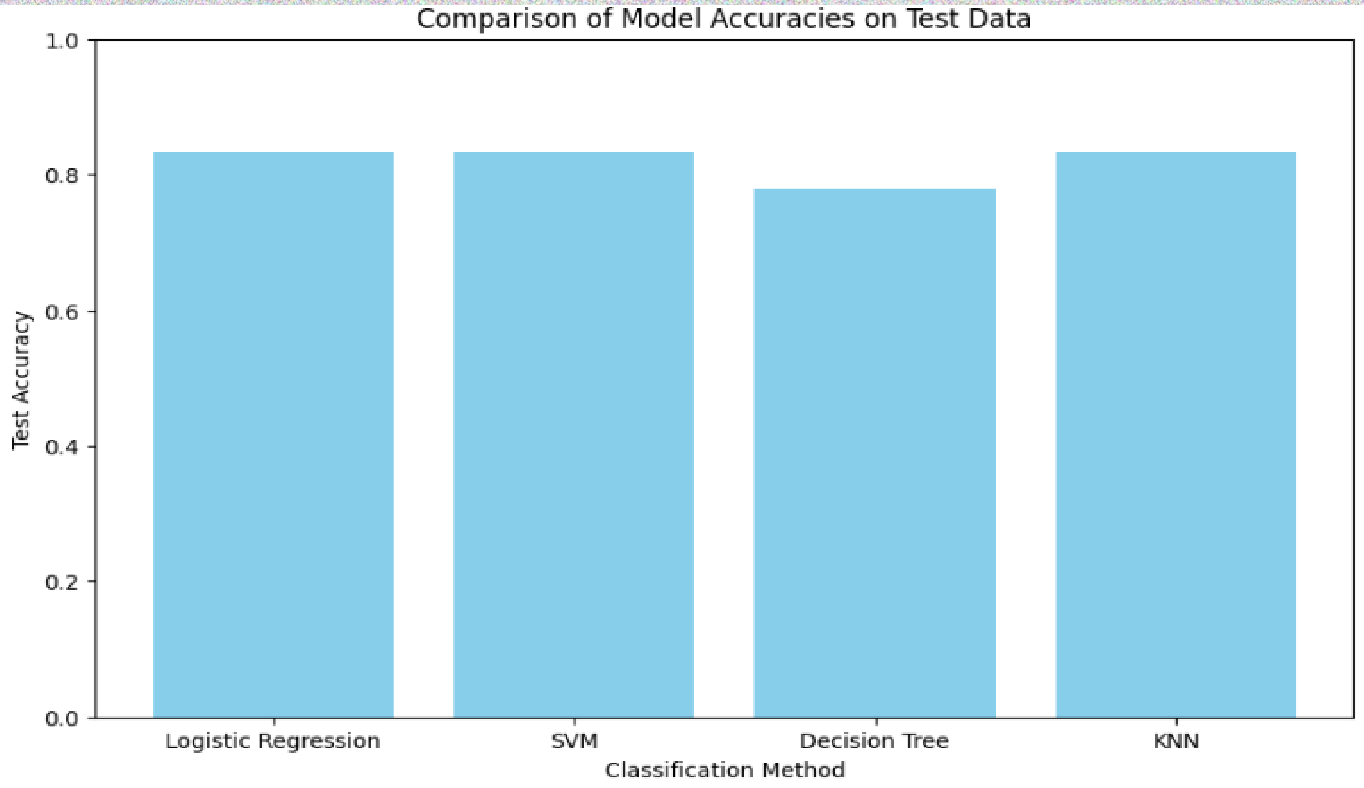
Datascience-Capstone/Datascience Capstone/module 4/SpaceX Machine Learning Prediction Part 5.ipynb at main · TetherCode/Datascience-Capstone

Results

Exploratory Data Analysis Results

- Found strong correlation between payload mass and landing success
- Launch site and booster version significantly impacted success rates
- Visualized trends using bar charts, scatter plots, and categorical plots





The best performing method is Logistic Regression with an accuracy of 0.8333

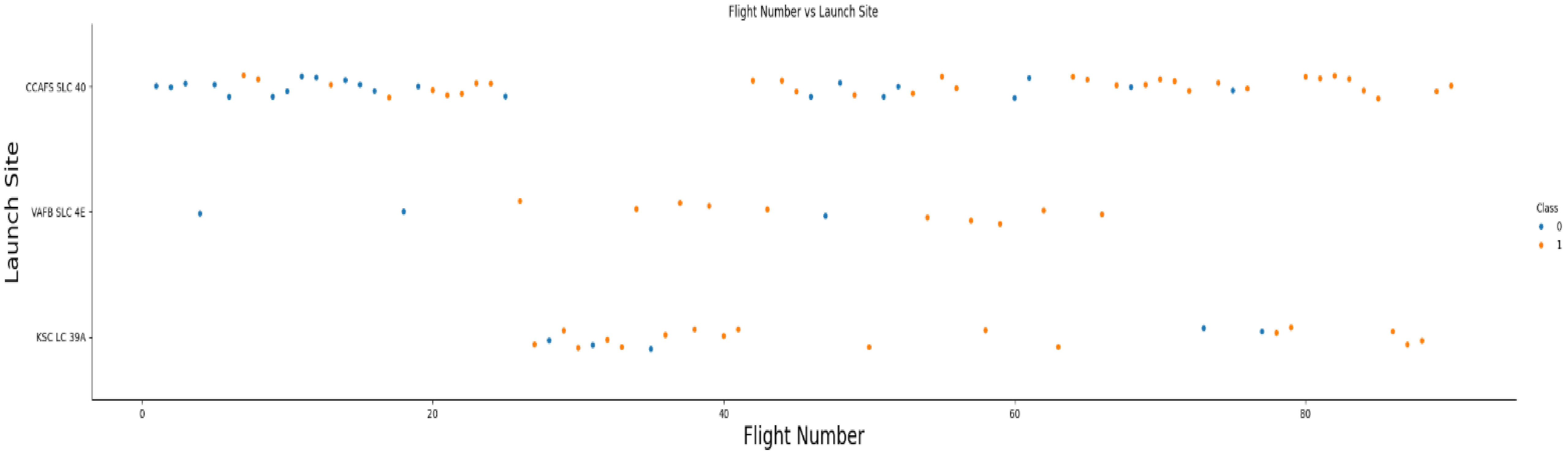
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Summary:

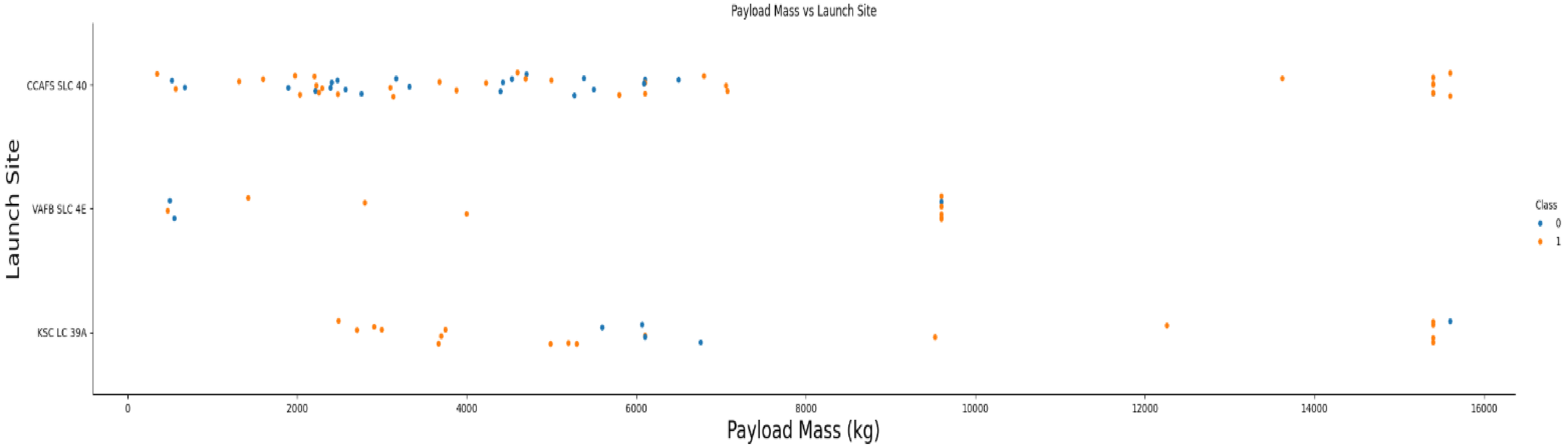
This scatter plot shows how SpaceX flight numbers are distributed across different launch sites. Each point represents a launch, color-coded by success or failure. It helps highlight which sites were used more often and how landing outcomes vary across sites over time.



Payload vs. Launch Site

Summary:

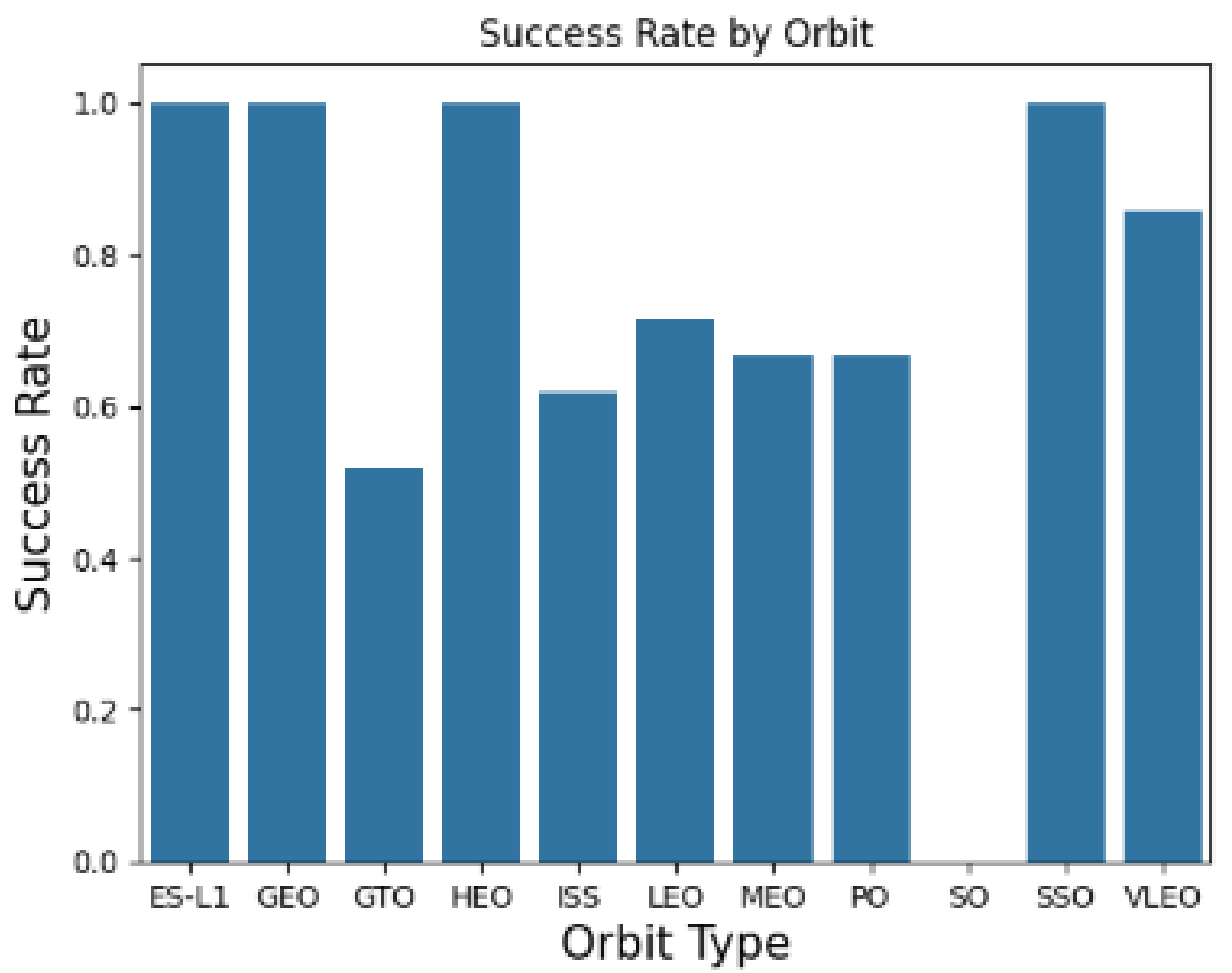
This scatter plot shows how payload mass varies across different launch sites. Each point represents a launch, with colors indicating success or failure. It highlights which sites handled heavier payloads and how landing outcomes change with increasing payload mass.



Success Rate vs. Orbit Type

Summary:

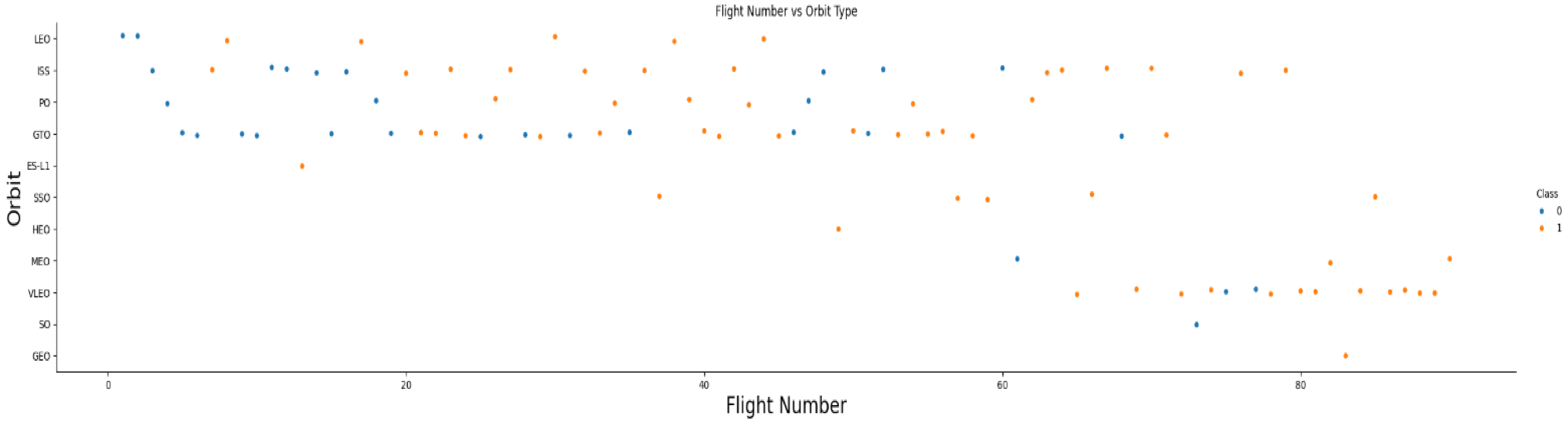
This bar chart compares mission success rates across different orbit types. Several orbits show near-perfect success, while others—like GTO—have noticeably lower performance. The visualization highlights how mission reliability varies depending on the target orbit.



Flight Number vs. Orbit Type

Summary:

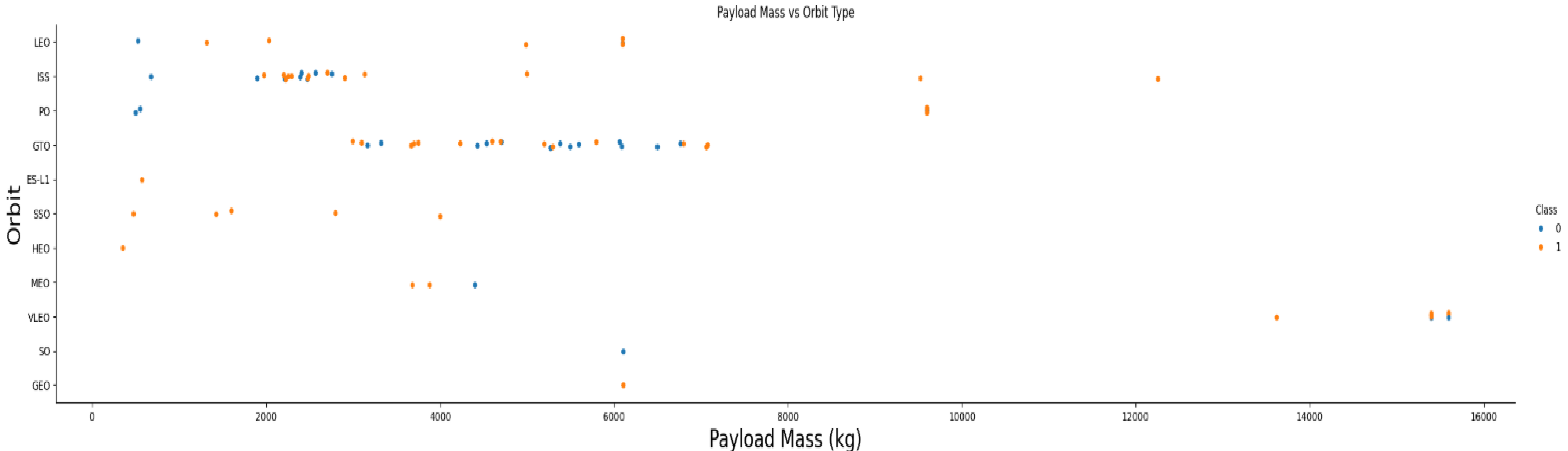
This scatter plot shows how different orbit types are distributed across SpaceX flight numbers. Each point represents a launch, color-coded by success or failure. The visualization highlights which orbits were targeted more frequently and how mission outcomes vary across orbit types over time.



Payload vs. Orbit Type

Summary:

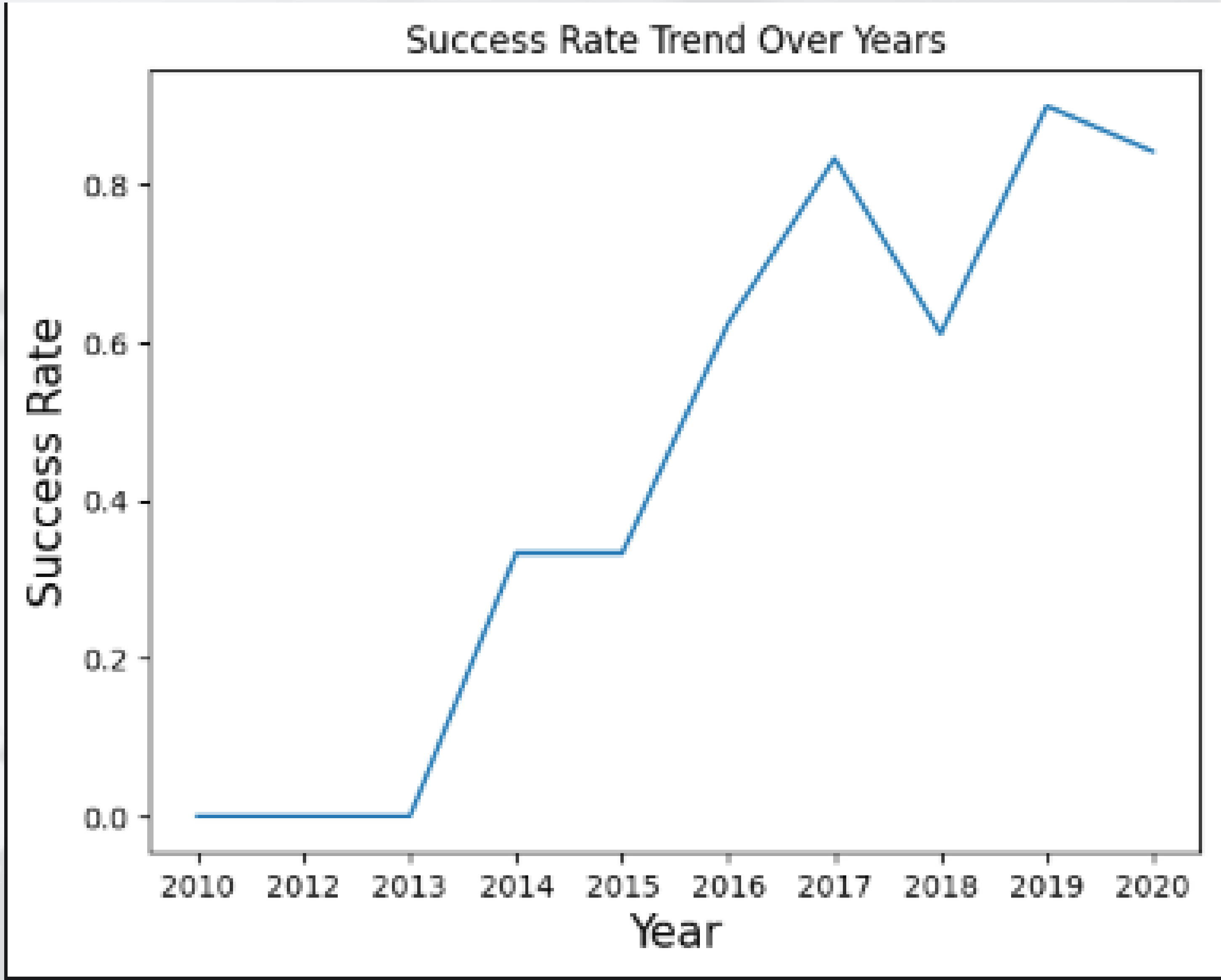
This scatter plot shows how payload mass varies across different orbit types. Each point represents a launch, color-coded by success or failure. The chart highlights which orbits typically carry heavier payloads and how landing outcomes shift as payload mass increases.



Launch Success Yearly Trend

Summary:

This line chart shows how SpaceX's launch success rate has improved over time. Success rates rise steadily from 2010 through the mid-2010s, reflecting major reliability gains. After peaking around 2018, the trend levels off with only slight variation, showing that launch performance has become consistently strong in recent years.



All Launch Site Names

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

- **Query Result:**
The dataset contains four unique SpaceX launch sites:
- **CCAFS SLC-40**
- **KSC LC-39A**
- **VAFB SLC-4E**
- **Starbase (Boca Chica, Texas)**
- This SQL query will get you that information
%%sql
- **SELECT DISTINCT** "Launch_Site"
- **FROM** SPACEXTABLE;

Launch Site Names Begin with 'CCA'

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

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

This query filters the table to show only records where the launch site name starts with “CCA,” and displays the first five matching rows.

Total Payload Mass

Total_Payload_Mass
45596

The sql query to achieve this is:

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)’;
```

This query adds up the payload mass for all launches where NASA was the customer, giving the total mass carried for NASA missions.

Average Payload Mass by F9 v1.1

Average_Payload_Mass
2928.4

The sql query to achieve this is:

%%sql SELECT AVG("PAYLOAD_MASS__KG_") **AS** Average_Payload_Mass **FROM** SPACEXTABLE **WHERE** "Booster_Version" = 'F9 v1.1';

This query calculates the average payload mass across all launches in the dataset by using the AVG() function on the PayloadMass column.

First Successful Ground Landing Date

MIN("Date")
2015-12-22

The sql query to achieve this is:

```
%%sql
SELECT MIN("Date")
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)’;
```

This query finds the earliest booster that successfully landed on the ground by filtering for ground landings, sorting the results by date, and returning the first matching booster.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

The sql query to achieve this is:

%%sql SELECT BoosterVersion FROM SPACEXTBL WHERE MissionOutcome = 'Success (drone ship)' AND PayloadMass > 4000 AND PayloadMass < 6000;

- This query filters the SpaceX dataset to find boosters that successfully landed on a drone ship while carrying payloads between 4000 and 6000 kg. It returns only the booster versions that meet all three conditions.

Total Number of Successful and Failure Mission Outcomes

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

```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS Total_Count
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

- This query groups all launches by their mission outcome and counts how many times each outcome occurred. It provides the total number of successful and failed missions recorded in the dataset.

Boosters Carried Maximum Payload

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

%%sqlSELECT "Booster_Version", "PAYLOAD_MASS__KG_"FROM SPACEXTABLEWHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_")   FROM SPACEXTABLE);his query finds the

booster or boosters that carried the heaviest payload by comparing each record's payload mass to the maximum payload value in the dataset. It returns only the booster versions that match this maximum.

2015 Launch Records

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

%%sqlSELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site"FROM SPACEXTABLEWHERE "Landing_Outcome" = 'Failure (drone ship)'AND substr("Date", 1, 4) = '2015';

- this query filters the dataset to show only the 2015 launches where the booster attempted a drone-ship landing and the landing outcome was a failure. It returns the booster version, launch site, and the specific failure outcome.

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

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

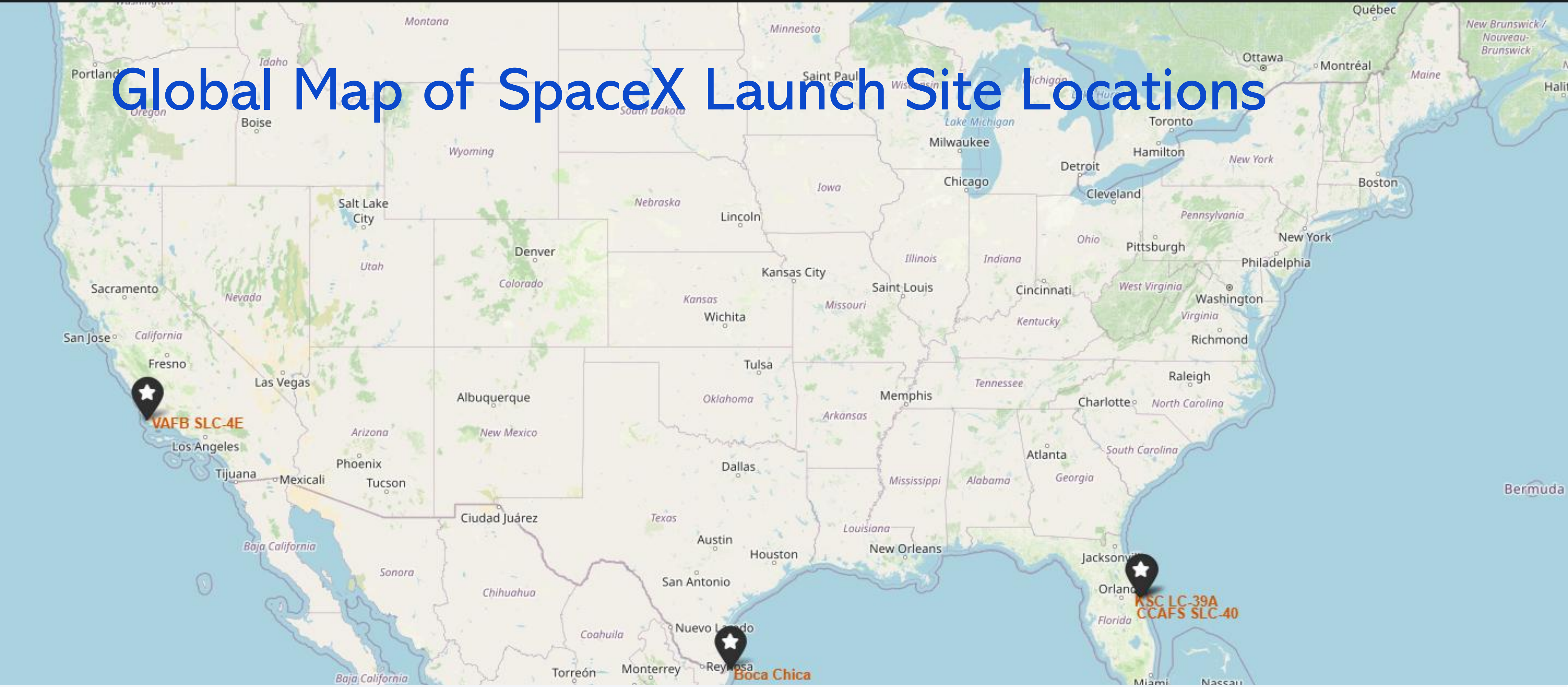
%%sqlSELECT "Landing_Outcome", COUNT(*) AS Outcome_CountFROM SPACEXTABLEWHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'GROUP BY "Landing_Outcome"ORDER BY Outcome_Count DESC;

- This query counts how many times each landing outcome occurred between June 4, 2010 and March 20, 2017. It groups the launches by landing outcome and ranks them in descending order so the most common outcomes appear first.

Section 3

Launch Sites Proximities Analysis

Global Map of SpaceX Launch Site Locations



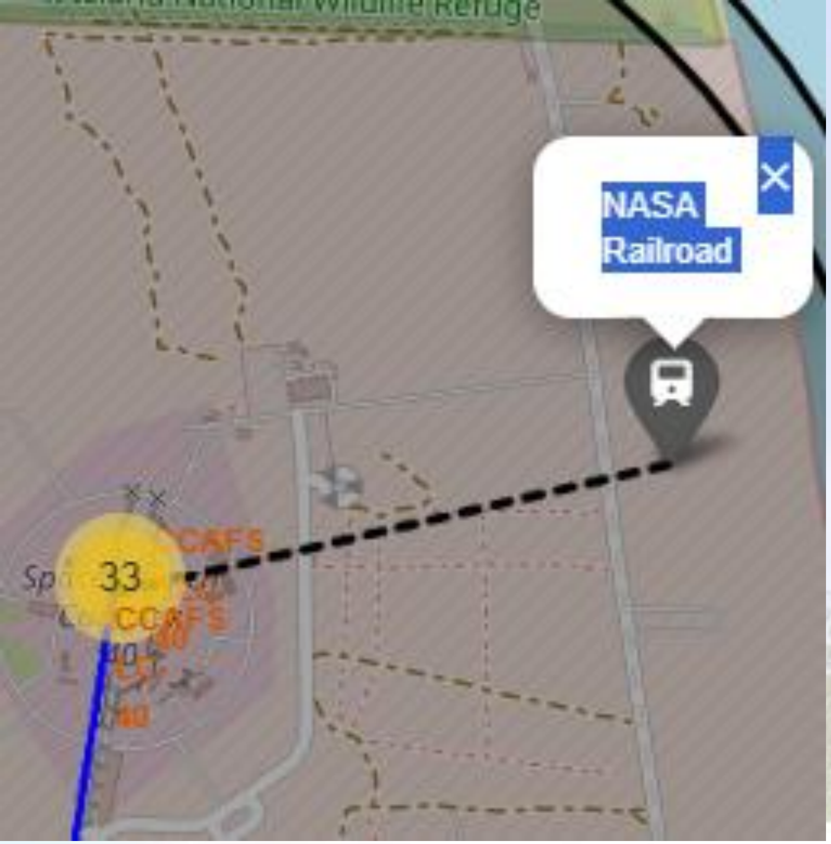
Summary:

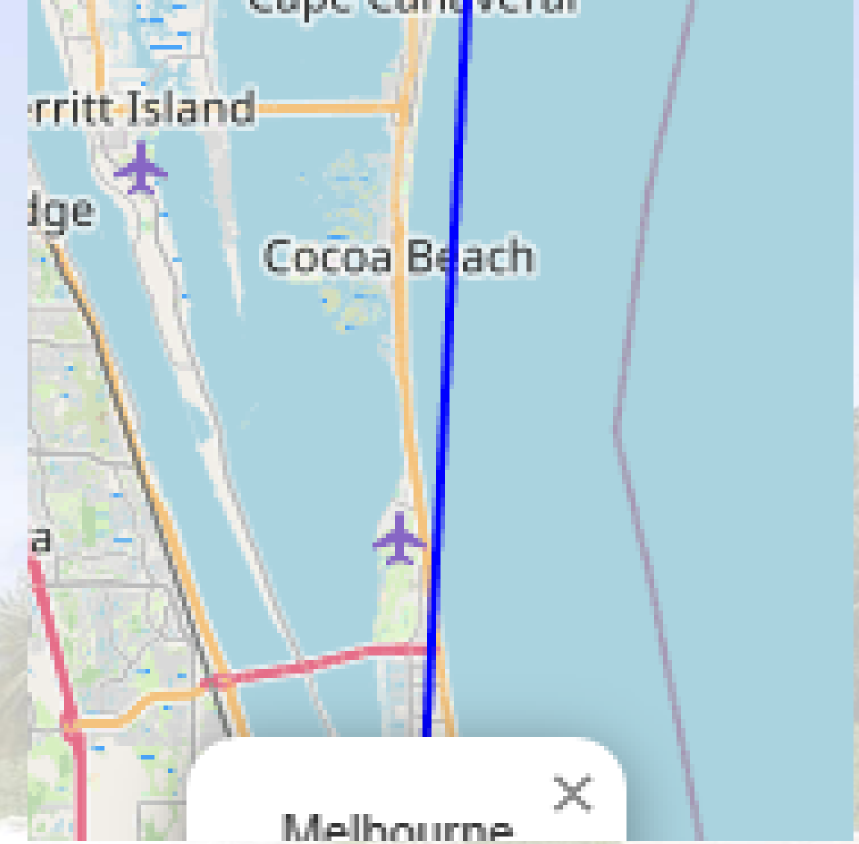
- The map shows all SpaceX launch sites from the dataset, each marked with a star and labeled clearly using offset text and leader lines. The sites cluster along the U.S. coastline, highlighting how SpaceX relies on coastal locations for safe launch trajectories over the ocean.

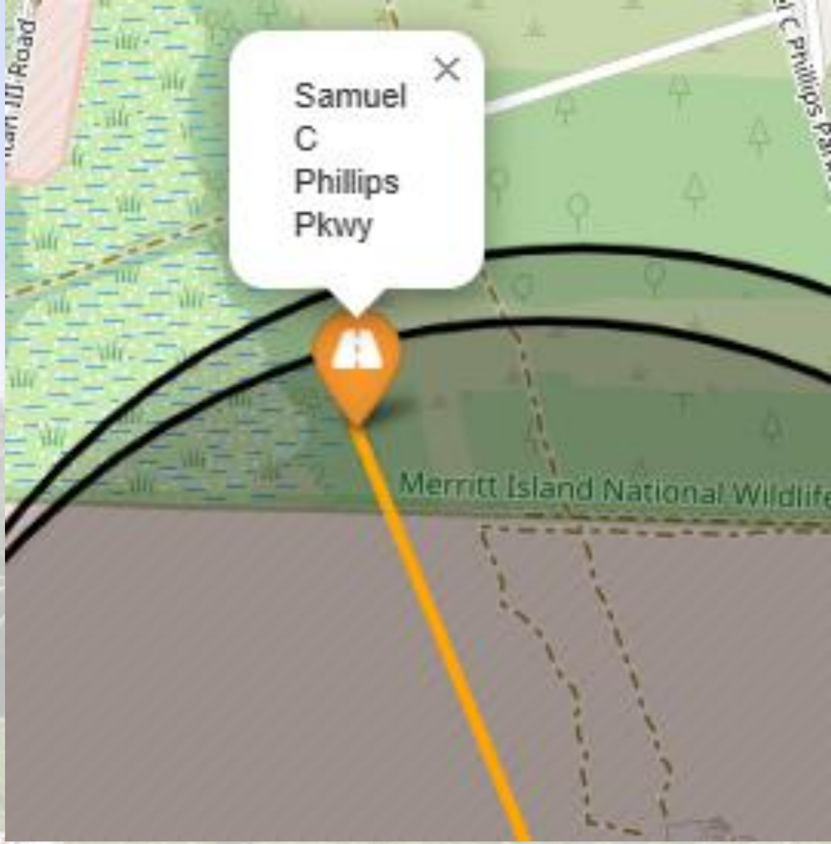
Launch Outcomes by Location (Success vs Failure)

Important Elements and Findings

- The map displays the four SpaceX launch locations used in the analysis:
  - CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.
- Each site is clearly labeled on the map, making it easy to see how SpaceX's launch infrastructure is spread across the U.S.
- The sites form a coastal pattern, which reflects SpaceX's need for safe downrange paths over the ocean during launches.
- Florida contains two closely located launch pads, highlighting it as SpaceX's most active launch region.







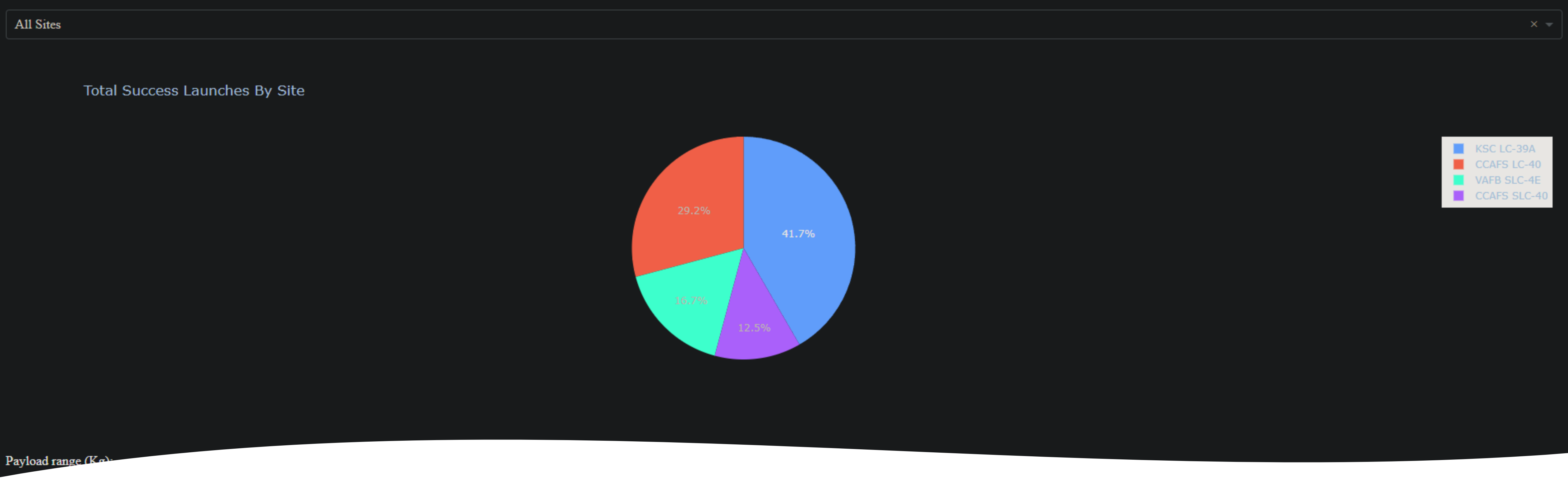
Launch Site Proximity Analysis

Summary:

- The map highlights how the launch site sits in a balanced position between the coastline, nearby highways, and the NASA railroad. Being close to the coastline provides a safe downrange path for launches, while proximity to highways and the railway supports the steady flow of equipment, fuel, and personnel. Seeing all three together shows how the site is intentionally placed to maximize safety during launches while still staying connected to the infrastructure needed to operate efficiently.

Section 4

Build a Dashboard with Plotly Dash



Launch Success Distribution Across All Sites

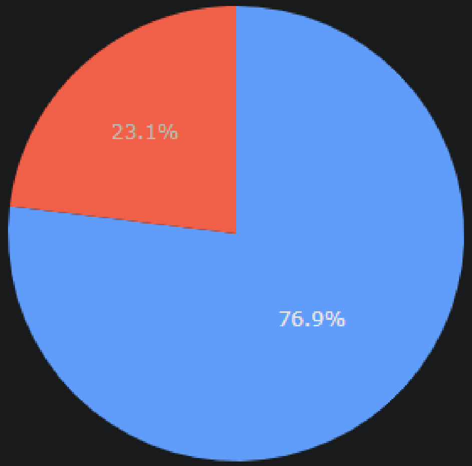
Important Elements and Findings:

- The pie chart shows the total number of successful launches at each SpaceX launch site.
- Each slice represents one site, with the size reflecting how many successful missions originated there.
- The chart makes it easy to see which locations contribute the most to SpaceX's overall success rate.
- Florida sites typically dominate the distribution, highlighting Cape Canaveral and Kennedy Space Center as the primary hubs for successful launches.

Launch Site With Highest Success Ratio

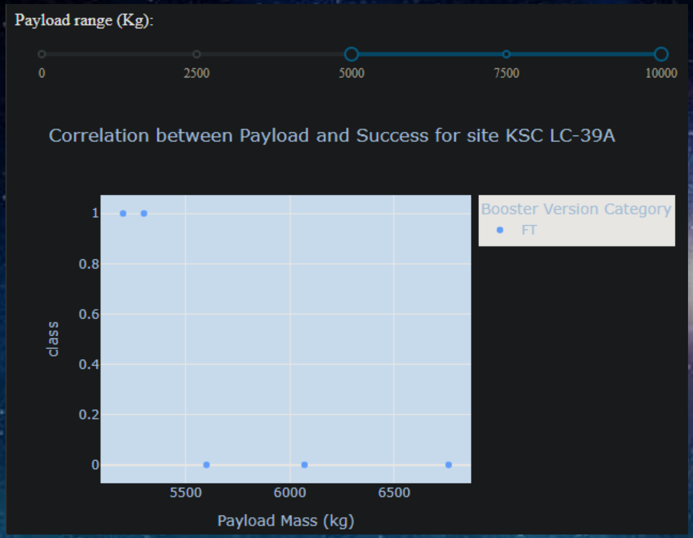
Important Elements and Findings

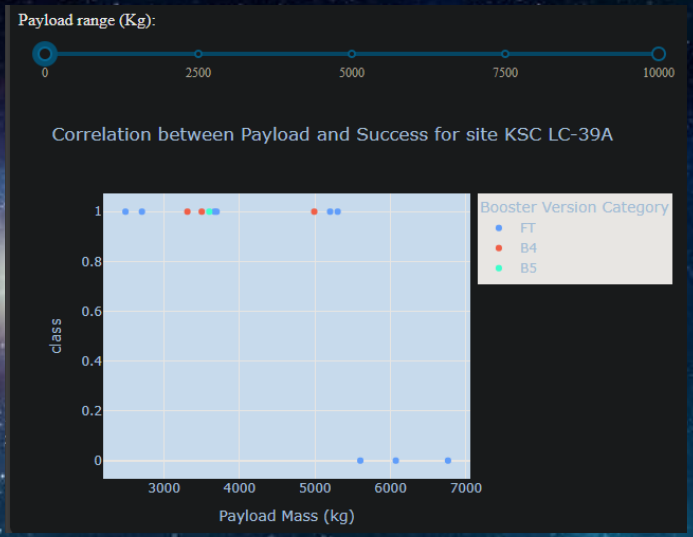
- **This pie chart focuses on KSC LC-39A, the launch site with the highest success ratio in the dataset.**
- **The chart shows that nearly all launches from this site are successful, with failures making up only a very small portion of its history.**
- **The dominance of the success slice highlights LC-39A's strong reliability, making it one of SpaceX's most dependable launch locations.**
- **This performance suggests that LC-39A benefits from mature infrastructure, consistent operational conditions, and proven launch processes, all contributing to its standout success rate.**

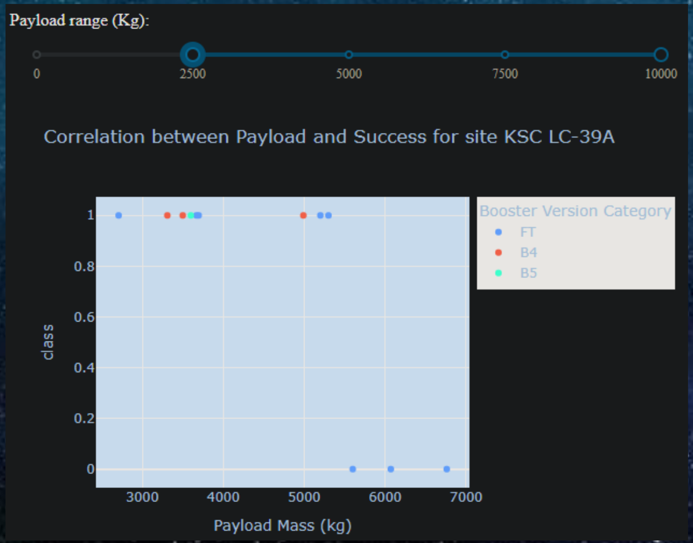


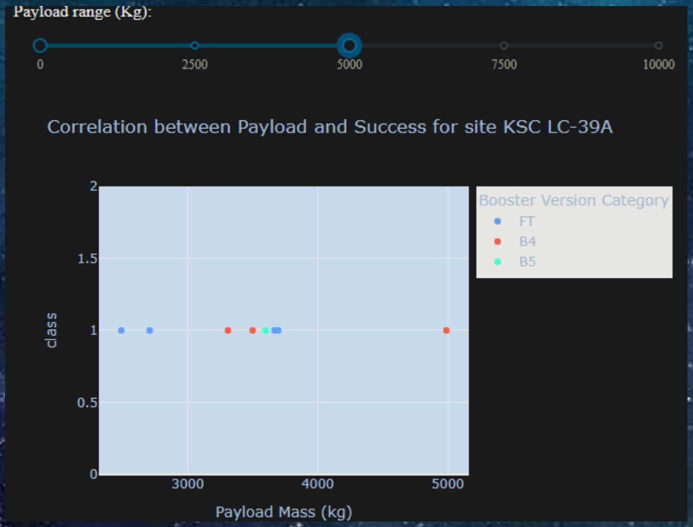
Payload vs. Launch Outcome Across All Sites

- **Important Elements and Findings**
- **The scatter plot shows how payload mass relates to launch success or failure across all SpaceX launch sites.**
- **Each point represents a single launch, with color coding indicating success or failure, and the range slider lets you isolate different payload intervals.**
- **When adjusting the slider, you can see that medium-to-high payload ranges tend to have the highest success rates, showing that SpaceX's heavier missions are generally reliable.**
- **Some booster versions cluster tightly in the successful region, suggesting that newer or more advanced boosters handle a wider payload range with consistent performance.**
- **Overall, the visualization helps reveal how payload size and booster technology influence mission outcomes, with success remaining strong across most payload ranges.**





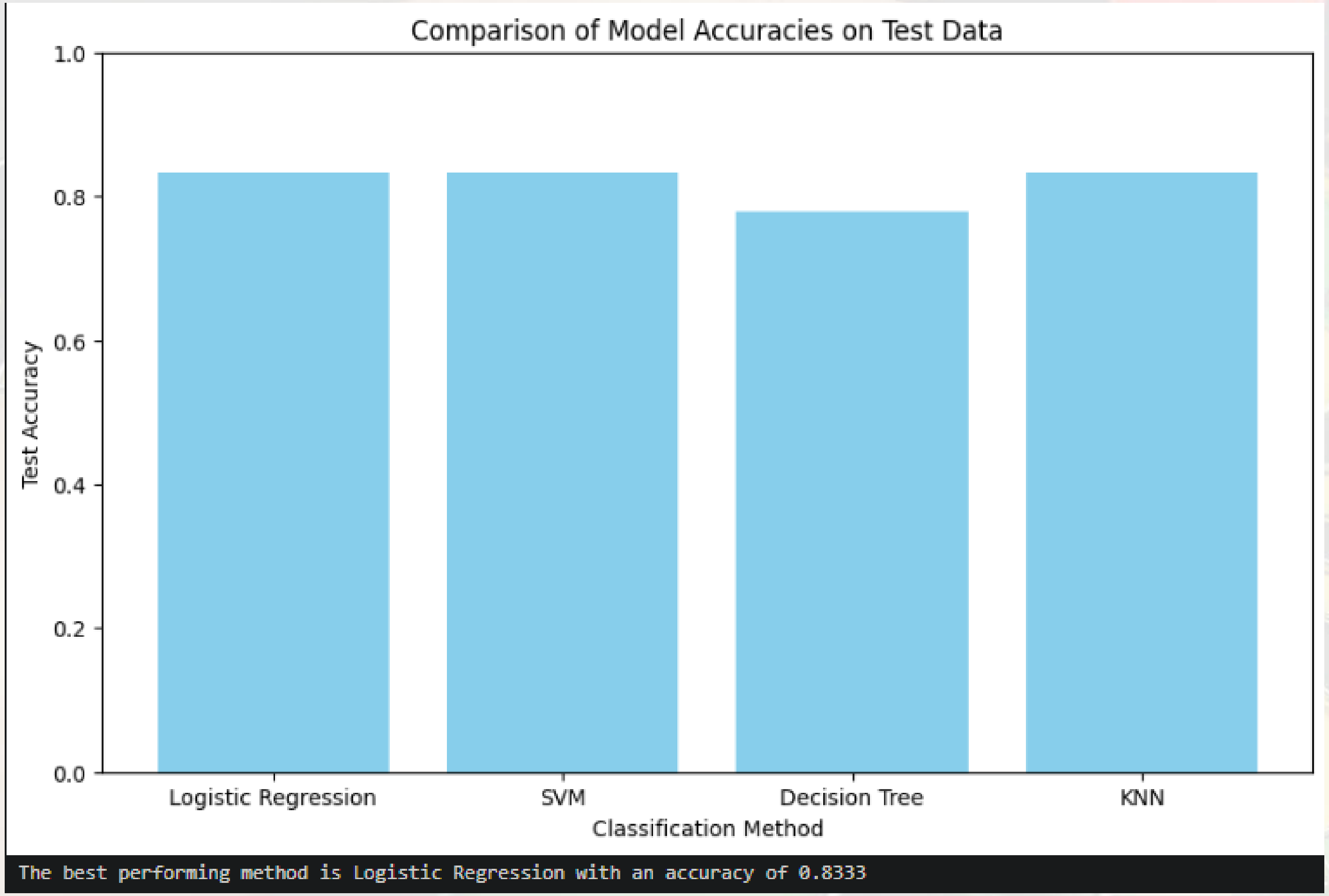




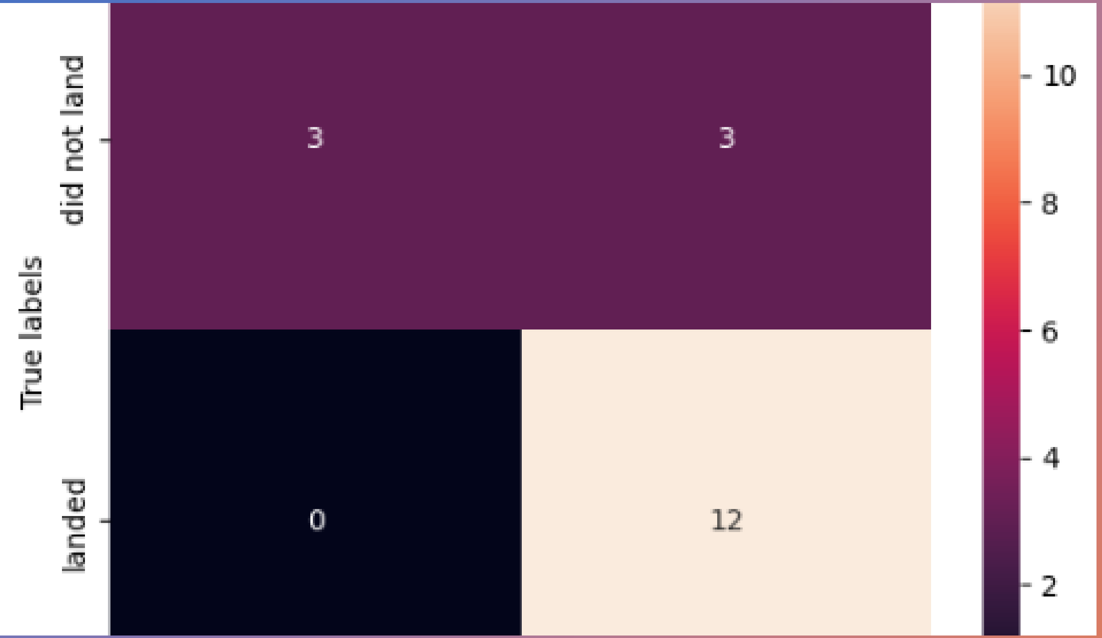
Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The bar chart compares the accuracy of four classification models on the test dataset. Logistic Regression comes out on top with the highest accuracy, followed closely by SVM and KNN, while the Decision Tree model performs noticeably lower. Overall, the results show that simpler, more stable models like Logistic Regression provide the most reliable predictions for SpaceX launch outcomes.



Confusion Matrix

The confusion matrix shows how well the Logistic Regression model classified launch outcomes. The diagonal cells represent correct predictions, while the off-diagonal cells show misclassifications. Most predictions fall along the diagonal, confirming that Logistic Regression not only achieved the highest accuracy but also consistently distinguishes successful launches from failures. This reinforces that it is the most reliable model for predicting SpaceX launch outcomes in this dataset.

Conclusions

Point 1:

The exploratory analysis and interactive dashboards revealed clear patterns in SpaceX launch performance, including which sites consistently achieve the highest success rates and how payload mass influences outcomes.

Point 2:

KSC LC-39A emerged as the most reliable launch site, demonstrating the highest success ratio and reinforcing the importance of site-specific infrastructure and operational stability.

Point 3:

The machine learning models showed that launch outcomes can be predicted with strong accuracy using features such as payload mass, booster version, and launch site, proving the dataset is highly learnable.

Point 4:

Among all models tested, Logistic Regression delivered the best overall performance, supported by both its accuracy score and a clean confusion matrix with minimal misclassifications.

Point 5:

Together, the dashboards and predictive models provide a comprehensive, data-driven understanding of SpaceX launch behavior, enabling more informed decision-making and highlighting the value of combining visualization with machine learning.

Appendix

Included Assets:

- **Python Code Snippets**
  Key portions of the notebook used throughout the project, including data cleaning, feature engineering, dashboard callbacks, and machine-learning model training and evaluation.
- **SQL Queries**
  Any queries used during the initial data exploration phase to filter, extract, or inspect launch records.
- **Charts & Visualizations**
  All generated plots such as the launch site success pie charts, payload vs. outcome scatter plots, model accuracy bar chart, and the confusion matrix for the best-performing classifier.
- **Notebook Outputs**
  Relevant outputs from the Jupyter notebook, including accuracy scores, classification reports, processed data previews, and intermediate results used for analysis.
- **Data Sets**
  The cleaned SpaceX launch dataset used for dashboard visualizations and predictive modeling.

Thank you!