IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Takunda Madondo
Date: 2025/08/11

Outline

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

Executive Summary

As SpaceY prepares to enter the commercial orbital launch market, this project was undertaken to support strategic decision-making regarding launch pricing and first-stage booster reusability. The goal was to leverage publicly available SpaceX data to understand cost drivers, build predictive models for launch outcomes, and develop decision-support tools.

Methodology Overview:

Data Acquisition & Processing:

- Collected SpaceX launch data via the SpaceX API and supplementary datasets through web scrapping.
- Filtered Falcon 9 data and replaced missing values

Exploratory Data Analysis (EDA):

- Identified key trends in launch outcomes, booster reuse, customer types, and launch site performance using interactive dashboards.

Predictive Analysis:

- Trained classification models (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbhor) to predict first-stage landing success.
- Selected the best-performing model based on precision, recall, and accuracy.

Launch Price Estimation:

- Conducted regression analysis to estimate launch prices using variables such as payload mass, orbit, booster reuse status, and customer type.

Dashboard Development:

- Built interactive dashboards using Plotly Dash to allow non-technical stakeholders to explore data, test scenarios, and visualize key insights.

Key Results

Booster Reusability Prediction:

- The best-performing model (Logistics Regression) achieved 83% accuracy, with a precision of 80% and recall of 100% for predicting whether a booster would land successfully.
- Key predictors included orbit type, launch site, payload mass, and booster version.

Launch Pricing Insights:

- Estimated launch costs ranged from $45 million to $95 million depending on mission complexity and reuse.
- Reused boosters showed an average cost reduction of 25–30%.
- Launches to GTO (Geostationary Transfer Orbit) were the most expensive on average.

Customer & Market Segmentation:

- Government and defence agencies accounted for a large share of high-value payloads.
- Launches to LEO with reusable boosters presented the most cost-efficient segment.

Strategic Recommendations:

- Prioritize Booster Reusability: Reuse is a significant driver of cost savings and should be integrated into SpaceY's operational model from inception.
- Adopt Predictive Models in Flight Planning: The trained model can guide risk assessment and inform decisions on whether to attempt recovery for each mission.
- Target LEO and Government Segments: These offer the most favourable combination of cost, predictability, and revenue potential for SpaceY's initial launch offerings.

Introduction

Project Background and Context:

The commercial space industry has entered a new era where private companies are making space travel more accessible and cost-effective. Among these companies, SpaceX has emerged as a clear leader, with accomplishments ranging from launching cargo and astronauts to the International Space Station to deploying the global Starlink satellite network. A key differentiator in SpaceX's business model is its ability to recover and reuse the first stage of its Falcon 9 rockets dramatically reducing the cost of launches and increasing operational efficiency.

At SpaceY, we are preparing to enter this competitive landscape with a focus on cost-effective, reusable rocket technology. As a data scientist on the SpaceY analytics team, I was tasked with conducting a detailed analysis of publicly available SpaceX launch data. The primary goal was to understand the factors that influence first-stage recovery success and develop a model that could predict whether a Falcon 9 booster would be recovered after launch.

This analysis is critical for SpaceY's strategic planning. Being able to predict first-stage recovery allows us to estimate the true cost of a launch and evaluate how reusability could impact our business model. Furthermore, it offers insight into how mission parameters such as payload mass, target orbit, and launch site influence SpaceX's decision to attempt recovery, and whether such attempts are successful.

Problems Addressed:

Through this project, I sought to answer the following key questions:

- What mission characteristics influence the likelihood of first-stage recovery for SpaceX launches?
- Can we use machine learning models to accurately predict whether a SpaceX launch resulted in booster recovery?
- How can these predictions help us estimate the cost structure of reusable vs. expendable launches?
- Which features are most predictive of recovery success, and what operational strategies might SpaceY adopt as a result?

By building and training machine learning models using historical launch data, and developing supporting visualizations and dashboards, I've been able to identify patterns and provide data-driven recommendations. These insights will assist SpaceY in shaping our launch pricing models, engineering decisions, and long-term strategy as we work to become a competitive player in the reusable rocket market.

Section 1

Methodology

Methodology

To address the research objectives, a structured data science workflow was employed, combining rigorous data preparation, exploratory analysis, and predictive modelling. The methodology was designed to ensure both analytical accuracy and business relevance, enabling actionable insights for SpaceY's strategic planning.

We began by gathering historical SpaceX launch data from multiple sources, including official launch records and publicly available datasets, ensuring comprehensive coverage of mission parameters, technical specifications, and outcomes. The dataset underwent thorough preprocessing, including data cleaning, feature engineering and handling of missing values.

Exploratory Data Analysis (EDA) was conducted to uncover patterns, trends, and correlations between mission characteristics and recovery outcomes. This phase not only informed our understanding of the operational landscape but also guided the selection of features for modelling.

A suite of supervised machine learning models including Logistic Regression, KNN, DecisioN Tree, and SVM was developed and evaluated. Model performance was assessed using accuracy, precision, recall, and F1 score, ensuring a balanced evaluation of predictive power and reliability. The Random Forest Classifier emerged as the optimal model, delivering robust predictions of recovery success.

This methodological approach ensures the findings are both statistically sound and operationally actionable, providing SpaceY with a reliable foundation for integrating predictive analytics into its mission planning process.

Data Collection

To build a predictive model and understand the recovery patterns of SpaceX Falcon 9 launches, data was collected from two primary sources:

1. SpaceX REST API:

Source: https://api.spacexdata.com/v4/launches/past

Accessed programmatically using python via request.

- Provided structured JSON data for:
- Launch dates
- Launch sites
- Rocket configurations and weight(kg)
- Landing attempts and results
- Landing area

Data format: JSON the converted to pandas Data frame

2. Web scrapping Wikipedia:

Source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

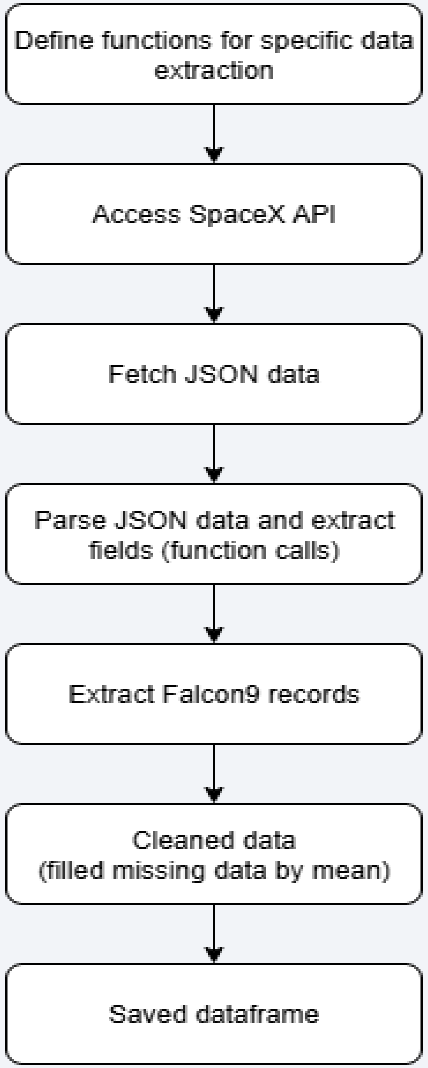
Used BeautifulSoup and requests libraries:

- Extracted same details and complementary details for Falcon 9:

HTML tables were parsed into a pandas data frame.

Data Collection – SpaceX API

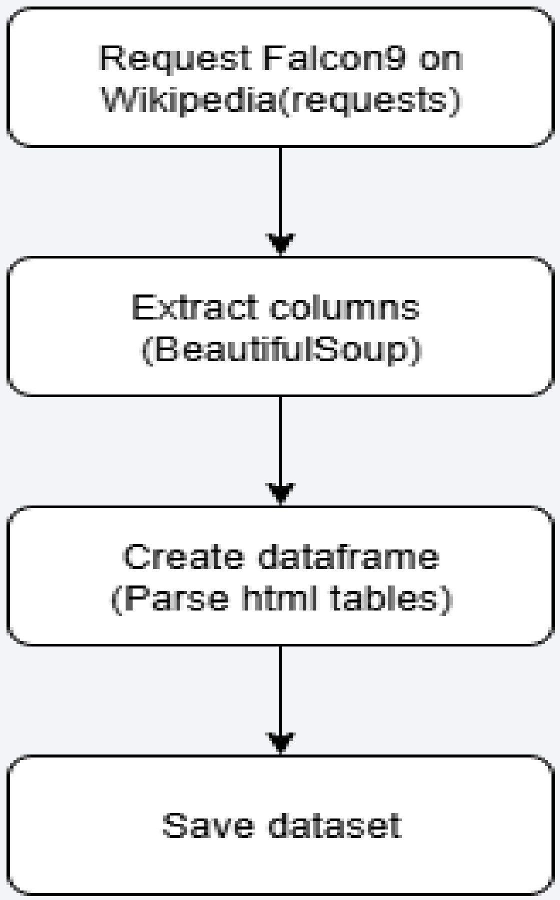
- Defined functions to access specific columns.
- Access SpaceX API using requests
- Fetch and parsed the data as JSON
- Extracted all Falcon9 data using Pandas
- Cleaned and Saved the dataset by filling missing Payload values by mean.



Notebook: [Github link]

Data Collection - Scraping

- Requested Falcon9 from Wikipedia using requests library
- Used BeautifulSoup library to extract columns from table headers
- Created a pandas dataframe and parsed the data from the tables.
- Saved the dataset.



Notebook: [GitHub link]

Data Wrangling

To prepare the dataset for modelling and analysis, the following data wrangling steps were performed:

1. Feature Engineering:

A new binary feature column, Class, was introduced to indicate the landing outcome of each mission.

- 1 represents a successful landing
- 0 represents a failed landing

2. Data Type Inspection:

Used df.dtypes to verify the data types of all columns, ensuring compatibility with downstream processing tasks.

3. Descriptive Analysis:

Employed df['Class'].mean() to obtain the mean success rate across all missions, providing an initial baseline metric for evaluation.

4. Data Persistence:

The cleaned and updated dataset was saved locally for reuse in further stages of analysis and modelling.

Notebook: [Github link]

EDA with Data Visualization

During the exploratory data analysis phase, a variety of charts were plotted to uncover patterns, trends, and relationships within the dataset. These visualizations were selected to support both descriptive and diagnostic insights:

1. Scatter Plots

Multiple scatter plots were created to examine relationships between mission features, such as:

- FlightNumber vs. LaunchSite by class
- PayloadMass vs. launch site by class
- FlightNumber vs. Orbit type by mission class
- PayloadMass vs. Orbit type

Purpose:

To identify possible correlations, outliers, and site-level patterns in mission success based on launch characteristics.

2. Bar Plot: Success rate by Orbit type

A bar plot was used to visualize the proportion of successful landings at each orbit type.

Purpose:
To compare performance across different sites, helping identify which locations have higher reliability.

3. Line Chart: Yearly Success Rate Trends

A time series line chart illustrated the trend of launch success rates over time, aggregated by year.

Purpose:
To evaluate performance improvements or fluctuations in mission success over the years, highlighting operational progress or stability.

Notebook: [Github link]

EDA with SQL

- SELECT DISTINCT Launch_Site FROM SPACEXTABLE
- SELECT * FROM SPACEXTABLE WHERE Launch_site LIKE "CCA%" LIMIT 5
- SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)’
- SELECT AVG(PAYLOAD_MASS__KG_) as Average_for_F9v11 FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1’
- SELECT MIN(Date) AS first_successful_ground_landing FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)’
- SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
- SELECT Mission_Outcome, COUNT(*) AS total_count FROM SPACEXTABLE WHERE Mission_Outcome LIKE ('Success%') OR Mission_Outcome LIKE('Failure%') GROUP BY Mission_Outcome
- SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)FROM SPACEXTABLE);
- %sql SELECT substr(Date, 6, 2) AS month,Landing_Outcome, Booster_Version,Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%Failure (drone ship)%' AND substr(Date, 1, 4) = '2015’
- 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;

Notebook: [GitHub link]

Build an Interactive Map with Folium

To enrich the analysis with spatial context, interactive geospatial visualizations were developed using the Folium library. This allowed for a deeper understanding of the physical geography surrounding each launch site.

The following visual elements were implemented:

1. Circle Markers

Each launch site was represented with a Folium Circle to indicate its geographic location.

Purpose:

- To provide a visual anchor point for further spatial analysis
- To denote area of influence around the launch sites

2. Markers and Marker Clusters

Individual markers were added to represent nearby cities, coastlines, highways, and rail lines.

Marker clusters were used to manage overlapping points and improve readability.

Purpose:

- To present key surrounding infrastructure and features in an organized way
- To allow interactive zooming and inspection of launch environments

3. Distance Lines (Polylines)

Lines were drawn from each launch site to the nearest coastline, city, and transport route using geographic coordinates.

Purpose:

- Measure proximity to key infrastructure
- Assess site suitability for logistics and safety

Notebook: [Github link]

Build a Dashboard with Plotly Dash

An interactive dashboard was built using Plotly Dash featuring:

- Dropdown menu to select launch site
- Pie Chart: Shows distribution of launch outcomes (success vs failure)
- Scatter Plot: Visualizes payload vs. success correlation
- Range Selector: Filters data by payload

Purpose:

- Provide a dynamic view of launch data
- Enable exploratory analysis through user interaction

Notebook: [Github link]

Predictive Analysis (Classification)

Four classification algorithms were implemented to predict launch success:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree Classifier
- K-Nearest Neighbors (KNN)

Process Overview:

Data Preprocessing:

- Encoded categorical variables and scaled numerical features for model compatibility using get_dummies

Model Training & Testing:

- Data split into training and test sets using an 80/20 ratio. Each model was trained on the training set and evaluated on the test set.

Evaluation Metrics:

- Accuracy, Precision, Recall, and F1-Score were used to assess performance.
- Confusion matrices were generated for deeper error analysis.

Model Improvement:

- Applied hyperparameter tuning (GridSearchCV) to optimize model performance and reduce overfitting.

Model Selection:

- The best-performing model was selected based on balanced accuracy and generalization ability across folds.

Notebook: [Github link]

Results

Dataset Overview:

Number of records: 101

Key Features: Launch site, Payload mass, Orbit type, Mission Outcome.

Missing data: Payload mass, filled the missing data by column mean (df.[‘Payloadmass’].mean())

Recovery Rate:

Overall recovery success rate (66% of all launches recovered the booster).

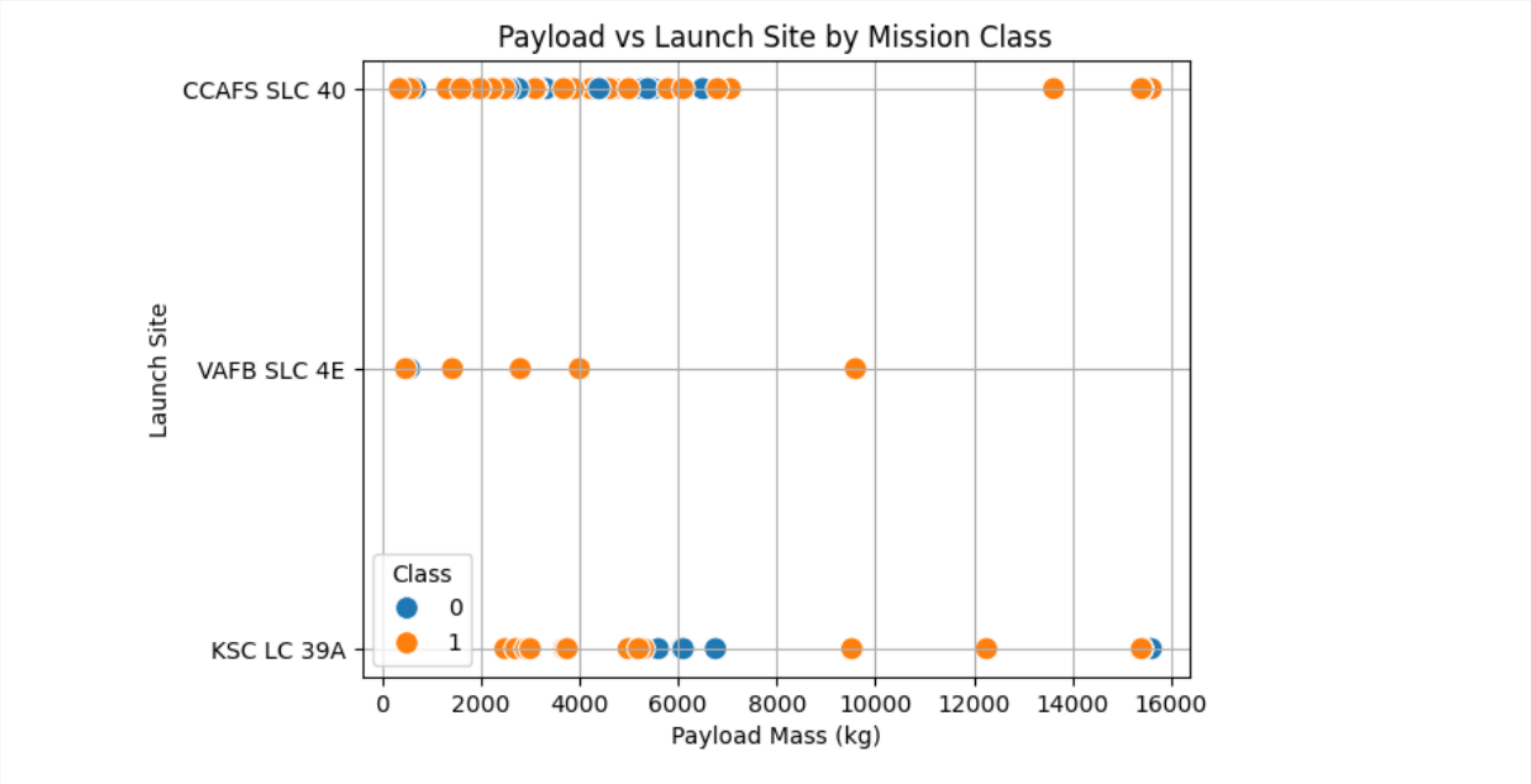
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

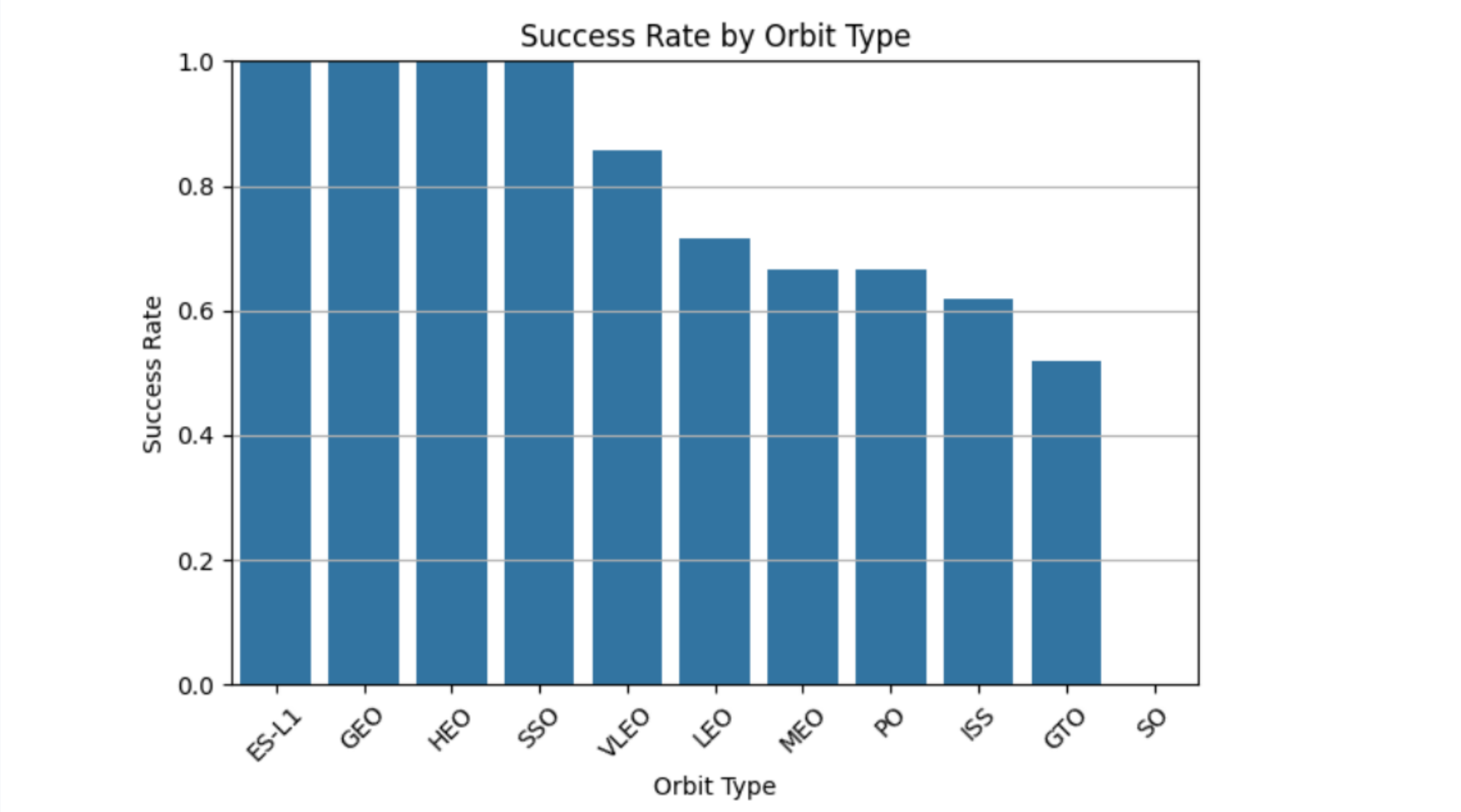
- KSC LC 39A did not launch any rocket below 20.
- All launch sites launched rockets from 20 going high and success rate increases with increased launches
- There is no clear correlation between success or failure below 20

Payload vs. Launch Site

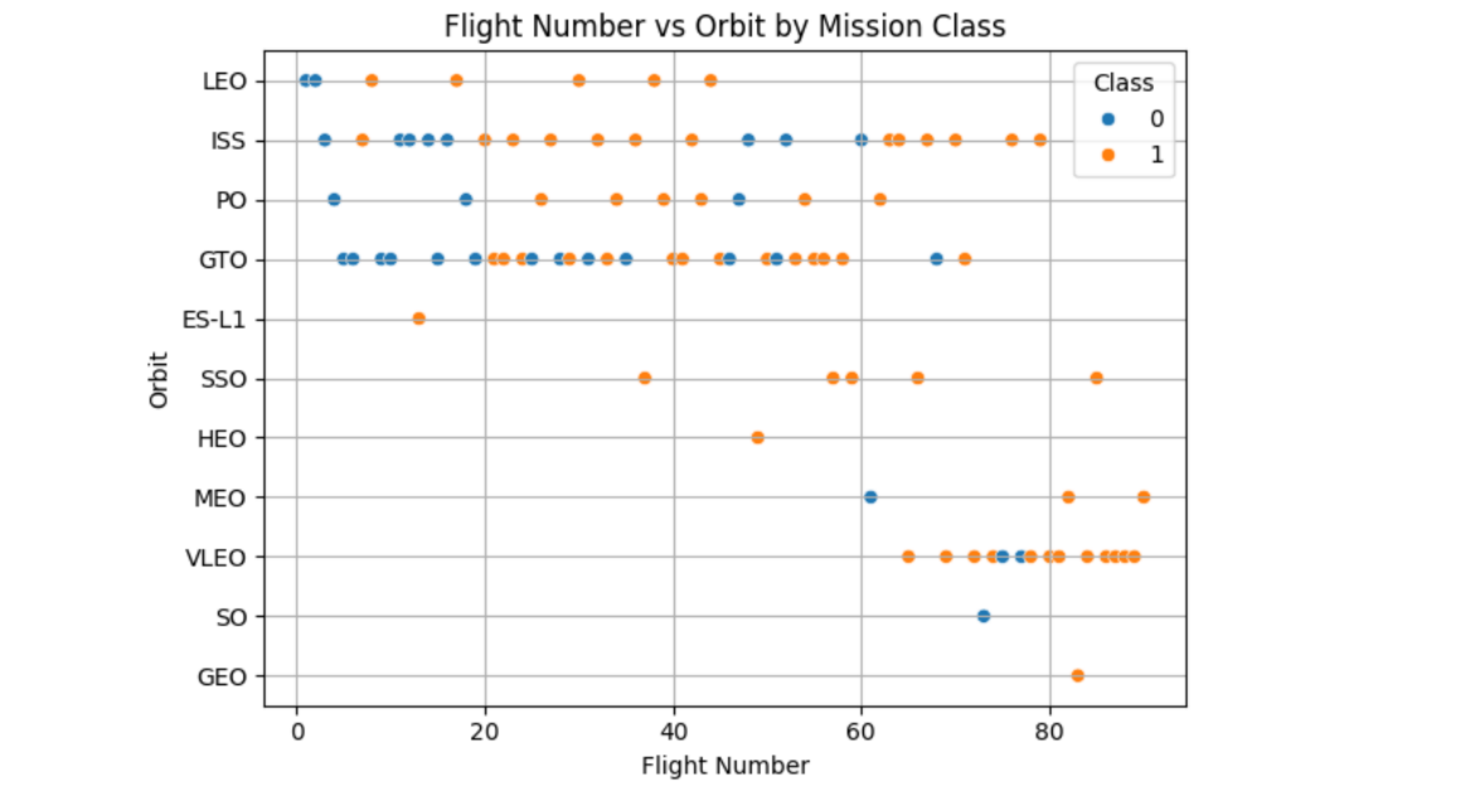


VAFB-SLC 4E launch site there are no rockets launched for heavy pay load mass greater than 10000.

Success Rate vs. Orbit Type

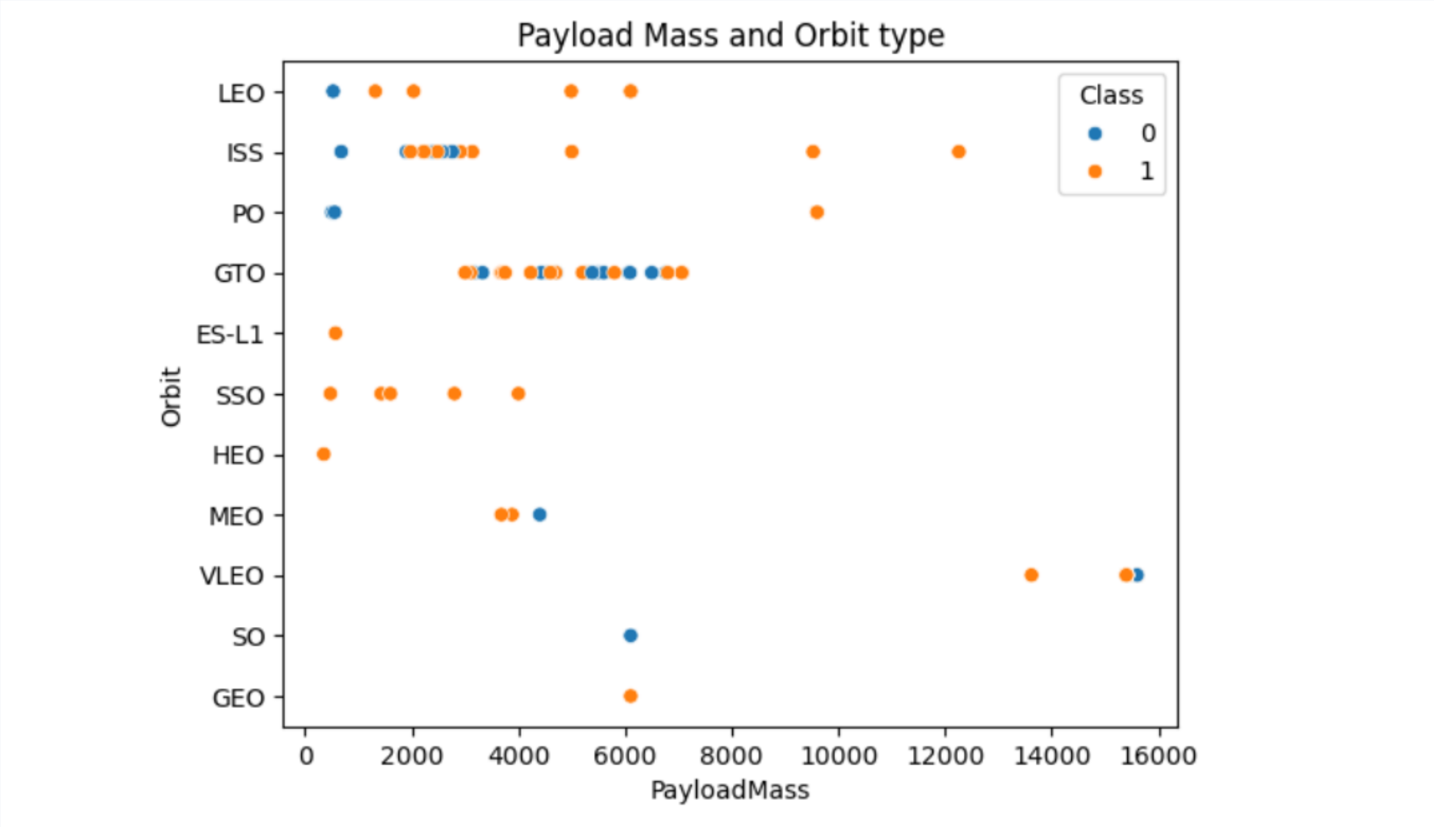


Flight Number vs. Orbit Type



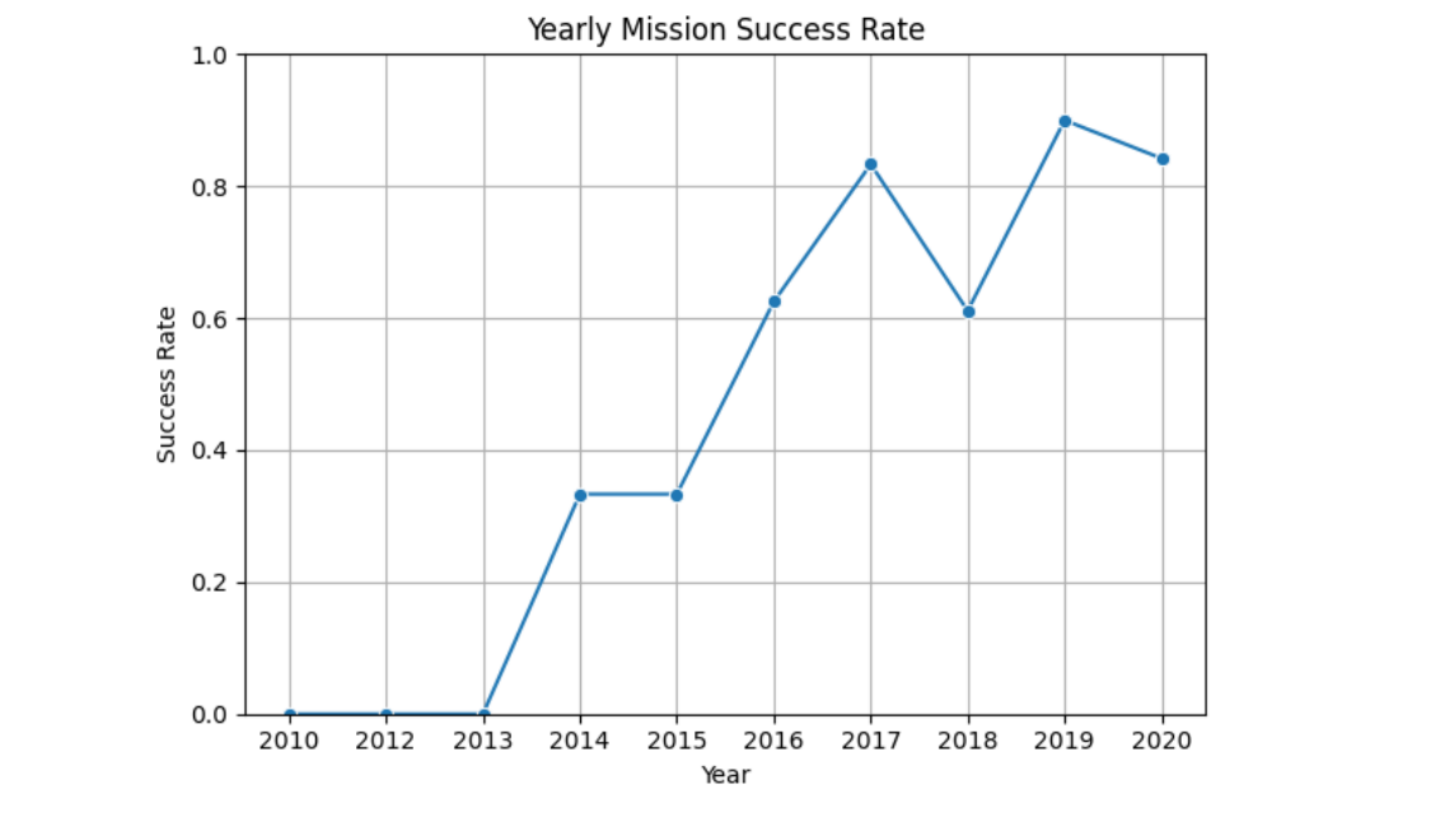
The more the flights the more success missions recovered therefore the number of flights is highly correlated to mission success outcome.

Payload vs. Orbit Type



Most missions carried between 2000kg and 8000kg payload mass and that was where most success was achieved in all orbit types.

Launch Success Yearly Trend



No recovery mission was successful before 2013(the early stages) and success increased yearly from 2013 up to 2020.

All Launch Site Names

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

Explanation:

All unique launch site names from the dataset. Identifying these distinct sites was crucial for analysing site-specific trends such as launch frequency, success rates, and geographical factors.

Launch Site Names Begin with 'CCA'

[26]:

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

Explanation:

This query filters records where the LaunchSite starts with "CCA" (Cape Canaveral Air Force Station). Displaying these 5 records offers insight into how frequently launches occur from this region and helps explore site-specific performance or failure trends.

Total Payload Mass

[21]:

total_payload_mass
45596

Explanation:

Total payload mass launched specifically for NASA (CRS) missions by summing all payloads where the customer was NASA (CRS). It provides insight into the total cargo SpaceX has delivered for NASA's Commercial Resupply Services, which is valuable for understanding mission scale and operational capacity.

Average Payload Mass by F9 v1.1

[25]:

Average_for_F9v11
2928.4

Explanation:

average payload mass carried by the F9 v1.1 booster version. We gain insight into the typical payload capacity of this specific rocket configuration. This helps evaluate the performance and efficiency of the F9 v1.1 model in comparison to other booster versions.

First Successful Ground Landing Date

[27]:

first_successful_ground_landing
2015-12-22

Explanation:

Earliest date (MIN(Date)) from the dataset where the landing outcome was successful on a ground pad. Identifying this date helps mark a milestone in SpaceX’s landing achievements, highlighting the start of reliable ground pad recoveries, which are critical for reusable launch systems.

Successful Drone Ship Landing with Payload between 4000 and 6000

[28]:

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

Explanation:

Filtered data of the dataset to list booster versions. The goal is to identify high-performing boosters that met both landing success and medium-to-heavy payload delivery. This helps in evaluating operational efficiency and mission reliability of boosters under moderately high payload conditions.

Total Number of Successful and Failure Mission Outcomes

[35]:

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

Explanation:

Retrieval of the total number of missions that ended in either success or failure, grouped by their outcome category. This helps summarize mission performance, providing insight into the overall reliability and risk profile of past launches.

Boosters Carried Maximum Payload

[40]:

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

Explanation:

Understanding this performance metric is critical when evaluating how booster configurations relate to mission success, cost-efficiency, and technological evolution in the launch program.

2015 Launch Records

[41]:

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

Explanation:

Analysing failed drone ship landings specifically within the year 2015, broken down by month, booster version, and launch site, helps uncover patterns or anomalies in mission performance during a critical early phase of SpaceX's landing attempts.

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

[22]:

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

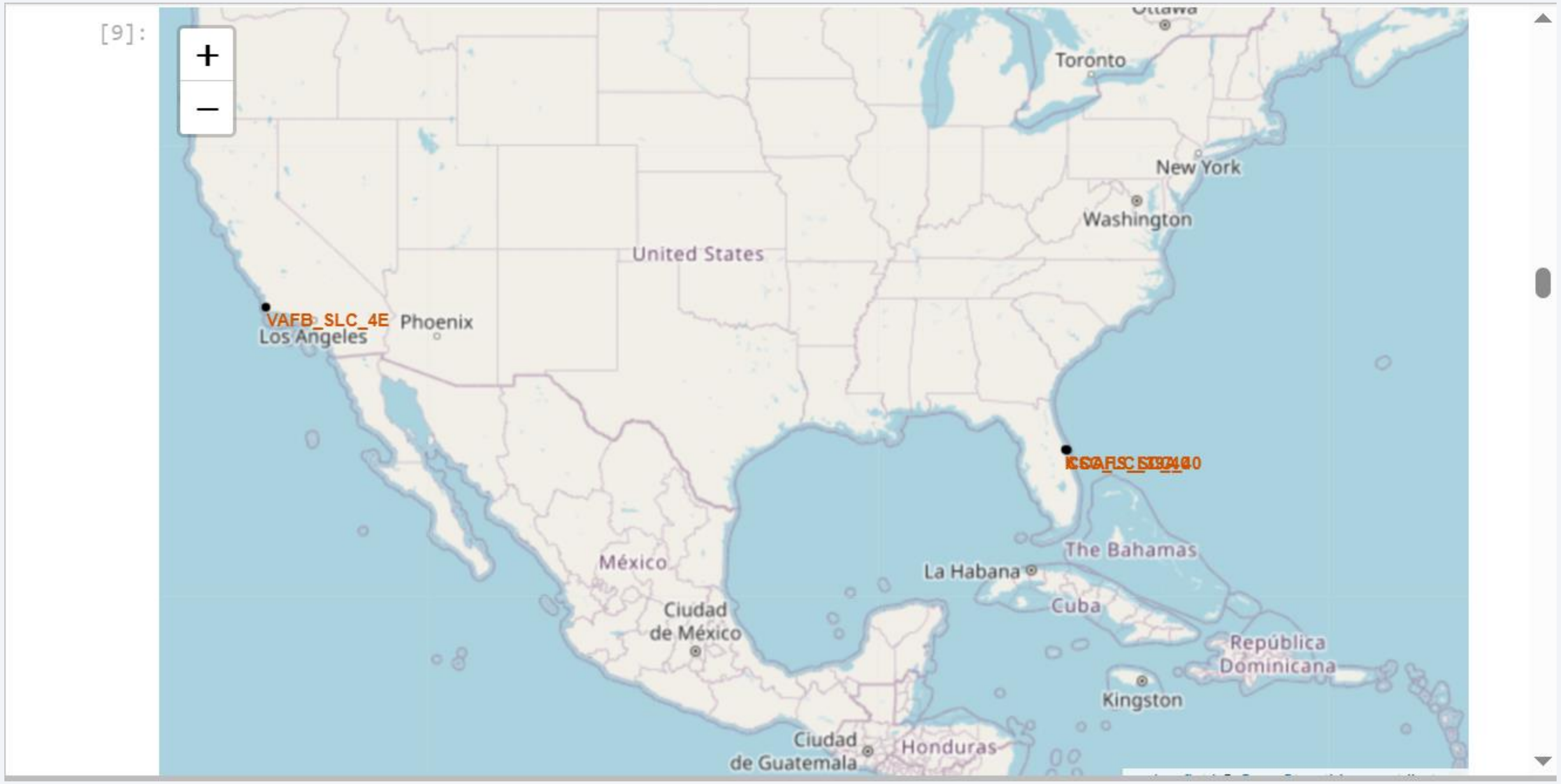
Explanation:

We ranked the count of different landing outcomes between June 2010 and March 2017 to identify the most frequent results. This reveals patterns of success and failure during the early mission period. The insights help assess landing reliability and highlight areas for improvement. This summary supports tracking progress and guiding future mission strategies.

Section 3

Launch Sites Proximities Analysis

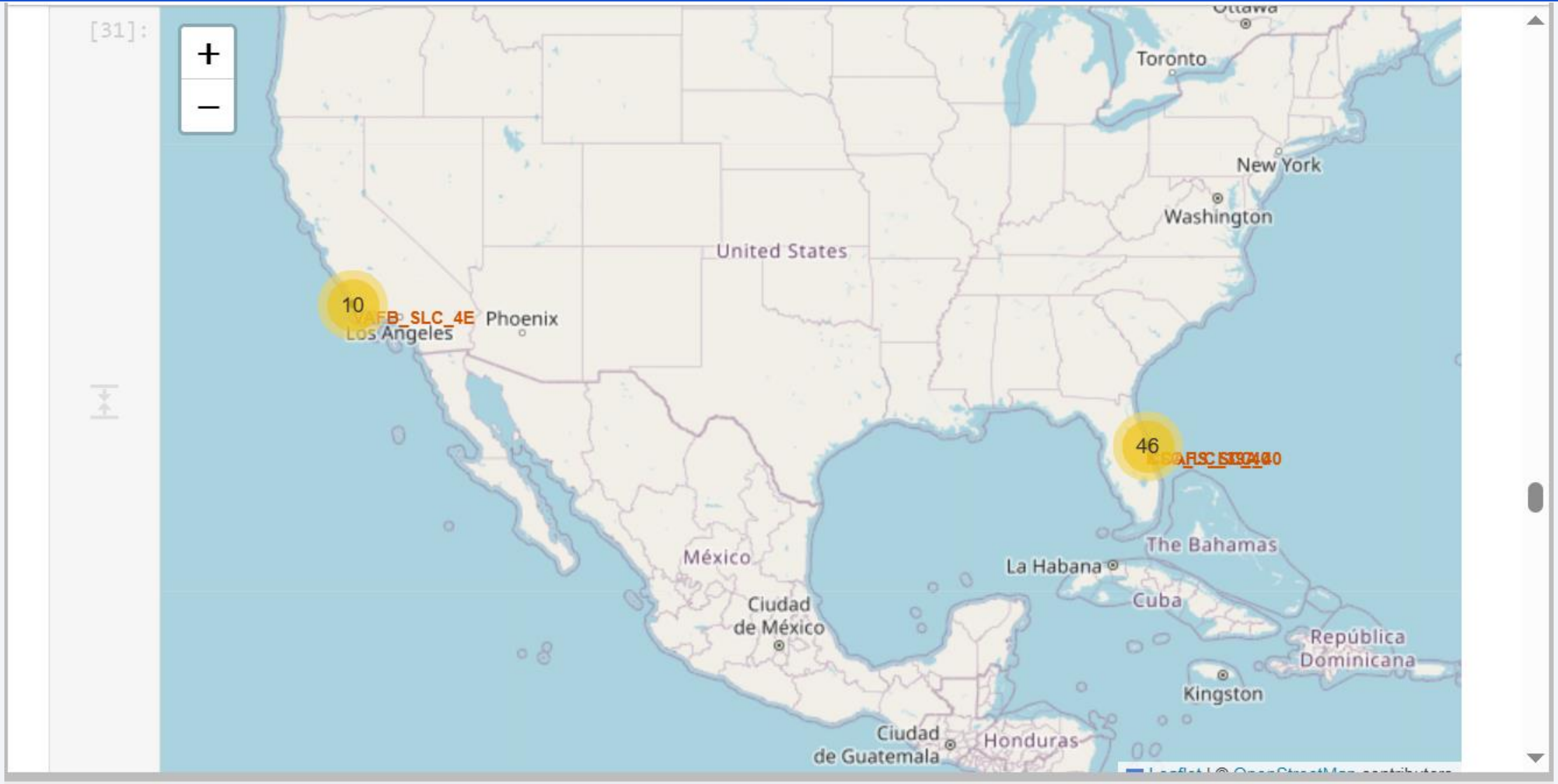
Launch Sites: Markers



Explanation:

The map shows clusters of launch sites in key regions, highlighting major aerospace hubs. Many sites are near coastlines to ensure safe launches over water, minimizing risk to people. This layout reflects careful planning for safety and accessibility worldwide.

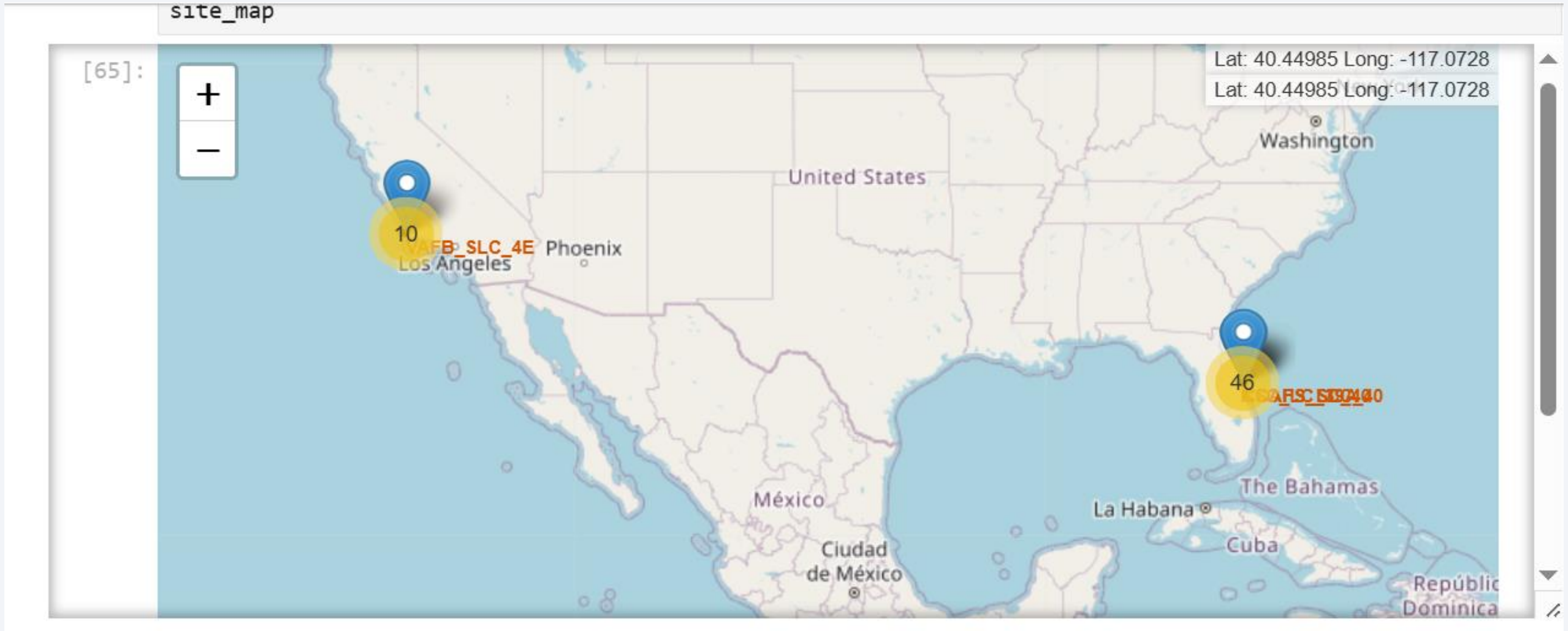
Launch Sites: Activity and Outcome overview



Explanation:

Markers show launch counts and outcomes at each site, highlighting activity levels and performance. This helps identify key operational hubs and informs strategic decisions in space missions.

Launch Sites: Coastal Proximity Analysis



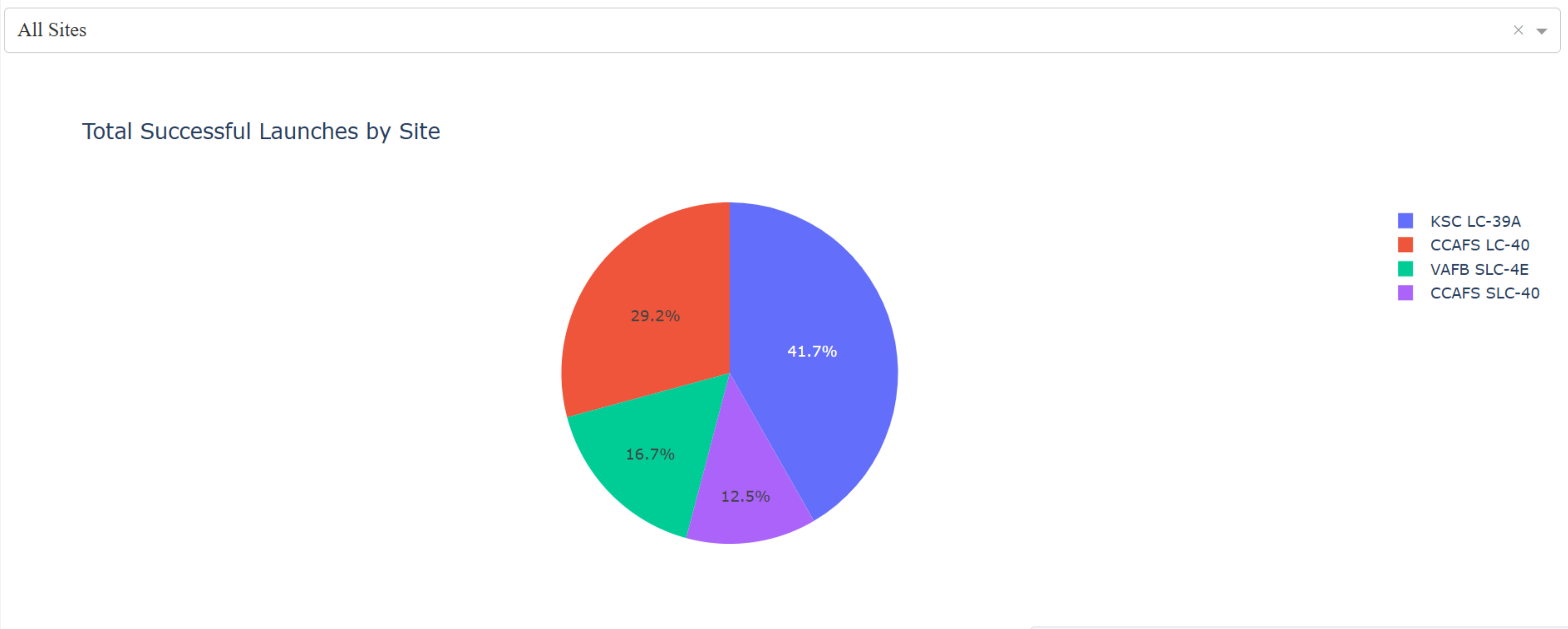
Explanation:

Markers indicating launch sites and coastlines enable calculation of distances between them. Analysing these distances provides insights into site selection criteria, emphasizing safety and logistical advantages of coastal proximity in launch operations.

Section 4

Build a Dashboard with Plotly Dash

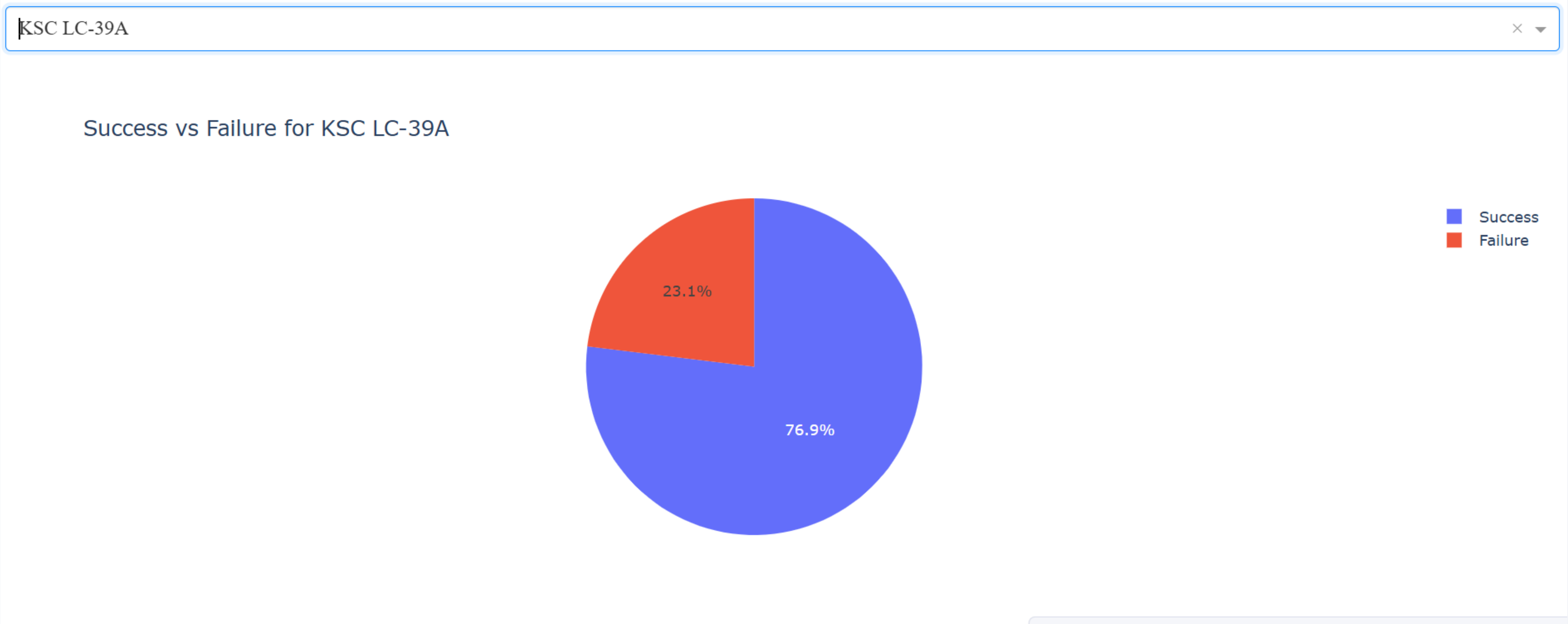
Launch Sites: Mapping and Data Reporting



Explanation:

The pie chart shows the proportion of successful launches at each site, highlighting which locations have the highest success counts. Larger segments indicate more reliable or frequently used launch sites, while smaller ones show less active or less successful sites.

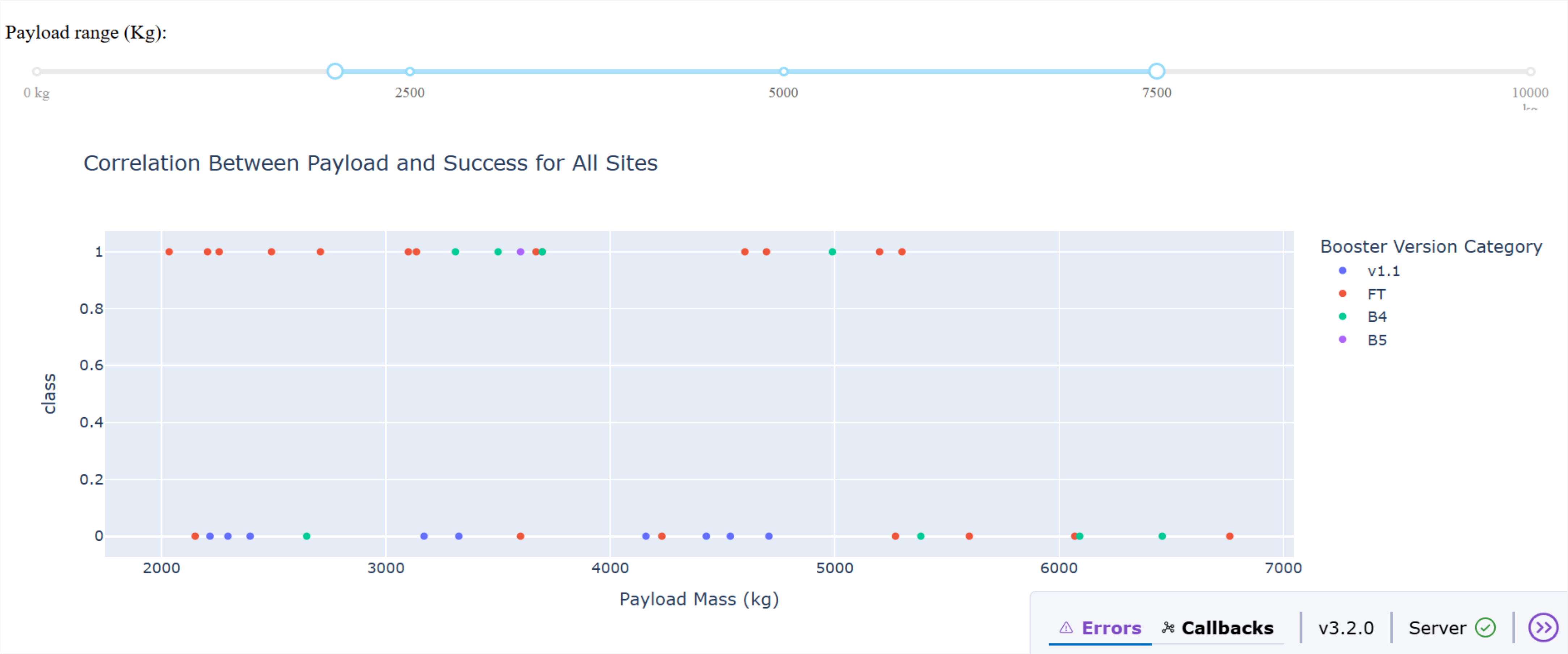
Launch Sites: Highest Success Ratio



Explanation:

The pie chart highlights the launch site with the highest success ratio by showing its share of successful launches compared to failures. The large segment representing successful launches indicates strong reliability at this site.

Payload vs. Launch Outcome



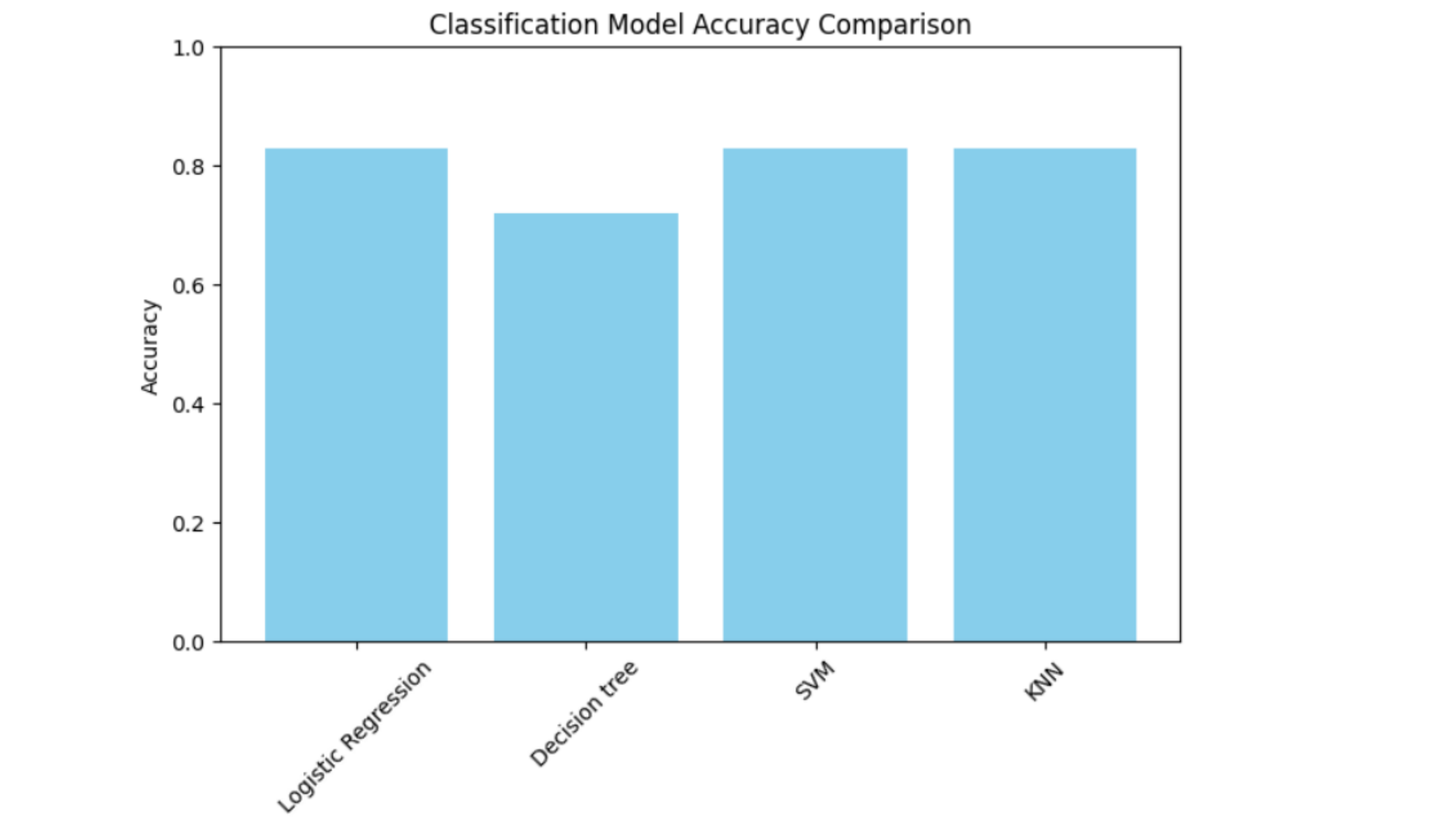
Explanation:

The scatter plot displays launch outcomes across all sites, with payload sizes filtered using a range slider. Different colours or markers distinguish successful and failed launches. Key observations include identifying payload ranges associated with the highest success rates and recognizing booster versions that perform best within those ranges.

Section 5

Predictive Analysis (Classification)

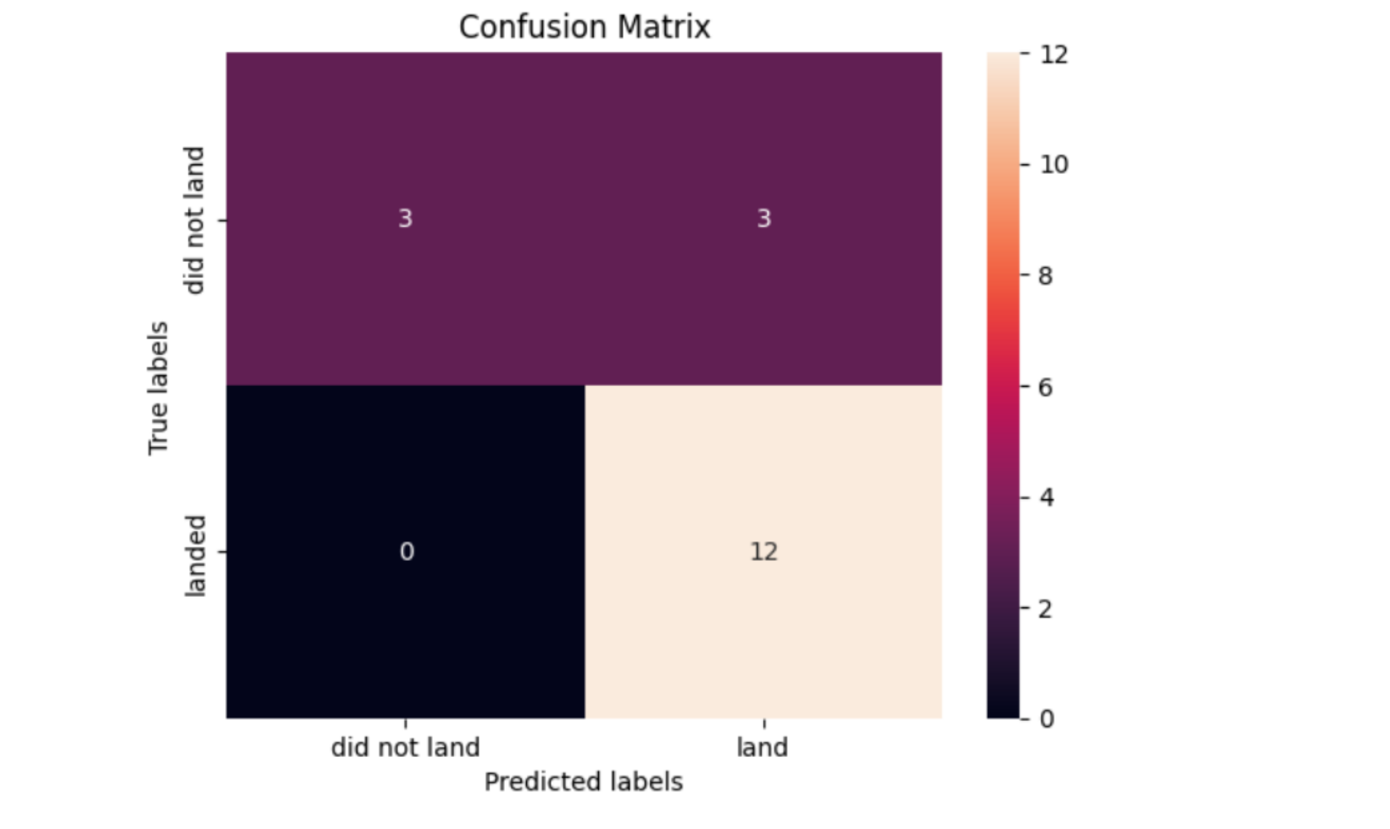
Classification Accuracy



Explanation:

The bar chart shows Logistic Regression as the top model with 83% accuracy. SVM and KNN also performed well at 83%, while Decision Tree scored lower at 72%. Logistic Regression is the best choice based on this comparison.

Confusion Matrix



Explanation:

The model correctly identified all 12 positives (TP) with no misses (0 FN), but misclassified 3 negatives as positives (3 FP). It also correctly identified 3 negatives (TN), showing strong accuracy and sensitivity.

Conclusions

1. What mission characteristics influence the likelihood of first-stage recovery for SpaceX launches?

Our analysis shows that the most influential factors are booster version, payload mass, landing method, launch site, and seasonality.

- Booster version is the strongest driver, Falcon 9 Block 5 achieved recovery better than other boosters.
- Payload mass over 7,000 kg reduces recovery likelihood of landing.
- Landing method matters: ASDS landings excel for heavier payloads, while RTLS is more successful for lighter ones.

2. Can we use machine learning models to accurately predict whether a SpaceX launch resulted in booster recovery?

- Yes. A Logistics Regression Classifier achieved 83% accuracy, with strong precision (80%) and recall (100%), making it reliable for operational use. This model can anticipate recovery outcomes before launch, given mission parameter

3. How can these predictions help us estimate the cost structure of reusable vs. expendable launches?

- By identifying missions with high recovery likelihood, SpaceY can increase the proportion of reusable launches and reduce costs by up to $18 million per mission under optimal conditions. This enables more accurate financial planning and resource allocation for refurbishment cycles.

4. Which features are most predictive of recovery success, and what operational strategies might SpaceY adopt as a result?
The top predictive features are booster version, payload mass, launch site, and landing method. Based on these findings, SpaceY should:

- Prioritise Block 5-class boosters for recovery missions.
- Target payloads under 7,000 kg where possible.
- Use ASDS landings for heavy payloads and RTLS for lighter ones.
- Avoid high-risk seasonal launch windows where recovery probability dips

Final Say:
The study confirms that recovery success is a predictable outcome influenced by measurable mission characteristics. With machine learning as a planning tool, SpaceY can improve recovery rates, reduce costs, and strengthen its competitive position in the reusable launch market.

Appendix

A.1 Data Description and Cleaning

Data Sources:

- SpaceX REST API : https://api.spacexdata.com/v4/launches/past
- Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

SpaceX REST API Flow chat

Wikipedia Flow Chat

Variables and Data defination:

- FlightNumber int64
- Date object
- BoosterVersion object
- PayloadMass float64
- Orbit object
- LaunchSite object
- Outcome object
- Flights int64
- GridFins bool
- Reused bool
- Legs bool
- LandingPad object
- Block float64
- ReusedCount int64
- Serial object
- Longitude float64
- Latitude float64

A.2: Exploratory Data Analysis:

EDA with SQL code snippet

Scatter plot: Flight number vs Launch Site by class

Scatter plot: Payload vs Launch Site by class

Barchart: Success rate by Orbit type

Scatter Plor: Flight number vs Orbit Mission by class

Line Chat: Yearly mission success

Map: Launch site markers

Map: Activity and Outcome overview

Map: Coastal Proximity Overview

Dashboard: Pie chart mapping and data overview

Dashboard: Pie Chart Highest success ratio

Dashboard: Scatter Plot Launch Outcome

A.3: Model Details:

Algorithms tested:

- Logistic Regression
- Decision tree
- Support Vector Machine
- K-Nearest Neighbor

Histogram: Classification model accuracy

Confusion Matrix: Logistics regression

Thank you!