IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nguyen Viet Dung
6th September 2024

Outline

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

Executive Summary

- **Summary of Methodologies**

- Data Collection & Preparation:

+ Collecting data on SpaceX launches, focusing on variables like launch site, payload mass, booster version, orbit type, and launch success outcomes.

+ Adding a new column for the landing class (success or failure) and standardized the data for modeling purposes.

+ Splitting data into training and test sets for model building and evaluation.

- Exploratory Data Analysis (EDA):

+ Performing EDA to understand key trends, such as how payload mass, launch site, and orbit type influence launch outcomes. Insights included the relationship between success rates and flight numbers, as well as site-specific performances.

- Machine Learning Model Development:

+ Training several models, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees. We applied GridSearchCV to fine-tune hyperparameters for each model.

+ Each model was evaluated using accuracy, precision, and recall metrics based on the test data.

- Model Evaluation & Selection:

+ The Decision Tree model showed the best performance with high accuracy in predicting whether the first stage would successfully land. The confusion matrix revealed minimal misclassification, making it the most reliable model.

Executive Summary

- **Summary of All Results**

- Launch Site Success Rates: Launch sites such as KSC LC-39A demonstrated the highest success rate (76.9%), while CCAFS areas showed a higher failure rate, providing key insights into site-specific performance.

- Payload Mass vs. Success: Lighter payloads (<4,000 kg) had a higher success rate across most booster versions, particularly with the FT and B4 versions, which showed consistent landing success.

- Model Performance: The Decision Tree model was the most effective in predicting the outcome of launches, with a high accuracy rate and minimal false predictions, as shown by the confusion matrix.

- Impact of First-Stage Reuse: First-stage reuse plays a critical role in reducing launch costs. The ability to predict successful landings allows Space Y to plan and optimize launch strategies effectively, mirroring SpaceX's cost-saving measures.

Introduction

- **Project Background and Context**

The commercial space age has made space travel more accessible, with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX leading the way. SpaceX stands out due to its cost-efficient rocket launches, largely thanks to its ability to reuse the first stage of the Falcon 9 rocket. The Falcon 9 is relatively inexpensive to launch at $62 million, compared to other providers that charge upwards of $165 million. Much of this savings comes from SpaceX's reuse of the first stage, a key factor in lowering launch costs.

In this project, a new rocket company called Space Y, founded by Billionaire Allon Musk, seeks to compete with SpaceX. As a data scientist working for Space Y, the goal is to gather data about SpaceX and use machine learning models to predict whether the first stage of the Falcon 9 will be reused.

- **Problems I Want to Find Answers To**

1. How does the reuse of the first stage affect the cost of SpaceX launches ?
2. Can a machine learning model be trained to predict whether SpaceX will successfully reuse the first stage, based on publicly available data ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API
- Perform data wrangling
  - Collected data then was processed through some technical methods to be available and then converted into the data frame
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was split into training and testing sets and trained using various classification models like Logistic Regression and Support Vector Machine (SVM). GridSearchCV was used to select the best model for each algorithm. Finally, the top models were compared based on their accuracy.

Data Collection

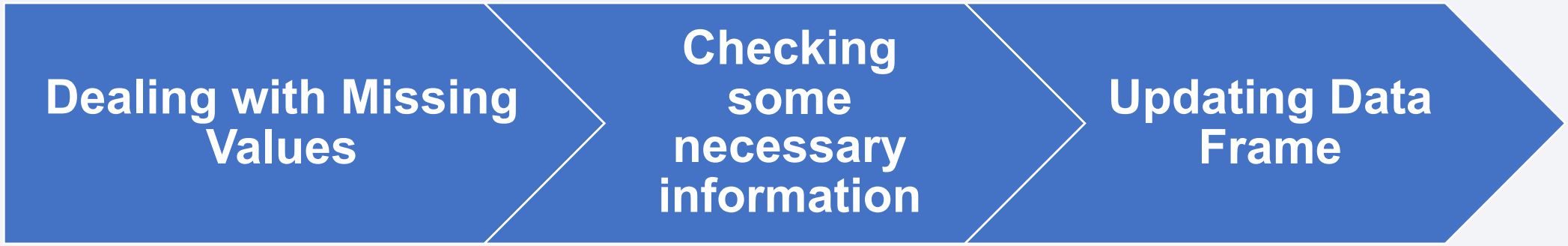
Data Collection – SpaceX API

GitHub URL : Data_Collection_Link

Data Collection - Scraping

GitHub URL : Web_Scraping_Link

Data Wrangling



GitHub URL : Data_Wrangling_Link

EDA with Data Visualization

Selected Plots :

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass vs. Launch Site
4. Launch Success Rate vs. Orbit Type
5. Flight Number vs. Orbit Type
6. Payload Mass vs. Orbit Type
7. Launch Success Rate vs. Year

Reason for Choosing These Plots

These plots were chosen to uncover patterns and relationships that directly influence the success of the Falcon 9's first-stage reuse. They focus on critical variables like payload mass, orbit type, launch sites, and time, which are essential for understanding the factors that lead to successful landings. These insights are key to training machine learning models that will predict whether a first stage can be reused, impacting the cost of each launch and helping Space Y compete with SpaceX.

GitHub URL : EDA_Link

EDA with SQL

- Retrieved the unique names of launch sites in the space mission data.
- Displayed 5 records where the launch sites start with "CCA."
- Calculated the total payload mass carried by boosters launched by NASA (CRS) missions.
- Computed the average payload mass carried by the booster version "F9 v1.1."
- Listed the date of the first successful landing outcome on a ground pad.
- Identified the names of boosters that succeeded on a drone ship and carried a payload mass between 4000 and 6000 kg.
- Displayed the total number of successful and failed mission outcomes.
- Listed the booster versions that carried the maximum payload mass.
- Retrieved records showing the month names, failed drone ship landings, booster versions, and launch sites for the year 2015.
- Ranked the count of landing outcomes (e.g., failure on drone ship, success on ground pad) between 2010-06-04 and 2017-03-20 in descending order.

These queries help analyze key details about launch sites, mission outcomes, and payloads, aiding the evaluation of booster performance and landing outcomes.

GitHub URL : EDA_SQL_Link

Build an Interactive Map with Folium

Features that I used to add into my map :

- **Circles**: Highlight each launch site with a 5000-meter radius for easy identification.
- **Markers**: Label launch sites with custom markers for clear visibility.
- **Mouse Position**: Added to track coordinates of nearby infrastructure for proximity analysis.
- **Distance Markers/Lines**: Show distances between launch sites and nearby cities, railways, or highways.

Purpose :

These objects help analyze geographic factors that may affect launch success and logistics, supporting the project's goal of understanding site influence on outcomes.

GitHub URL : Folium_Map_Link

Build a Dashboard with Plotly Dash

Summary of Dashboard Features

+ Dropdown for Launch Site Selection: A dropdown allows users to filter data by specific launch sites or view all sites.

+ Pie Chart for Launch Success: Displays total successful launches across all sites or by a selected site.

+ Range Slider for Payload Mass: Enables users to filter data by payload mass range.

+ Scatter Plot for Payload vs Launch Success: Visualizes the relationship between payload mass and launch success, with additional color coding for booster version categories.

Purpose of Plots and Interactions

+ The **Dropdown** helps users focus on individual launch sites, providing deeper insights into specific site performance.

+ The **Pie Chart** offers a clear overview of success vs. failure rates, useful for comparing launch outcomes across different sites.

+ The **Range Slider** allows for flexible exploration of how payload mass affects success rates, aiding in performance analysis.

+ The **Scatter Plot** helps identify correlations between payload size and launch success, supporting decision-making and predictive analysis.

GitHub URL : DashBoard_Link

Predictive Analysis (Classification)

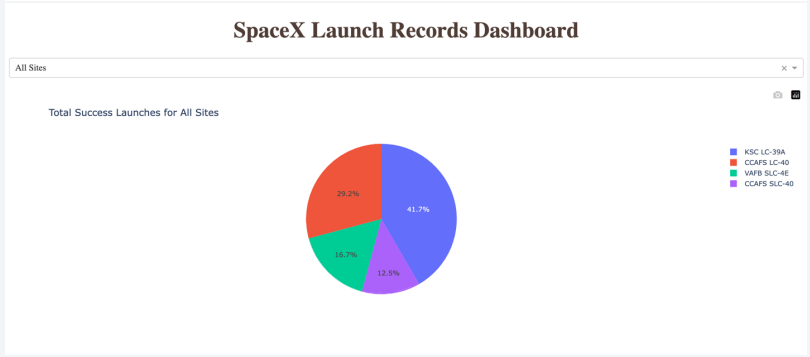
GitHub URL : Machine_Learning_Prediction_Link

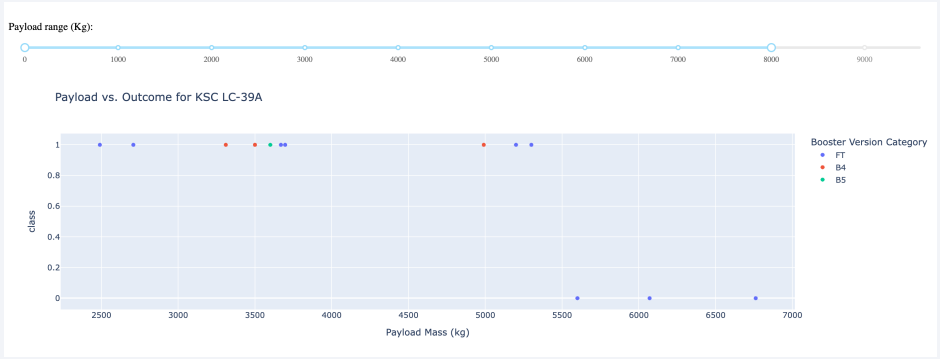
Results

- **Exploratory data analysis results**

As flight numbers increase, the first stage is more likely to land successfully, regardless of payload mass. The VAFB-SLC launch site did not handle heavy payloads (>10,000 kg). Orbits such as ES-L1, GEO, HEO, and SSO have the highest success rates. In LEO, success is tied to the number of flights, while GTO shows no clear relationship. Heavier payloads tend to have higher success rates in Polar, LEO, and ISS orbits, but GTO shows mixed outcomes. Overall, success rates have consistently improved since 2013, peaking in 2020.

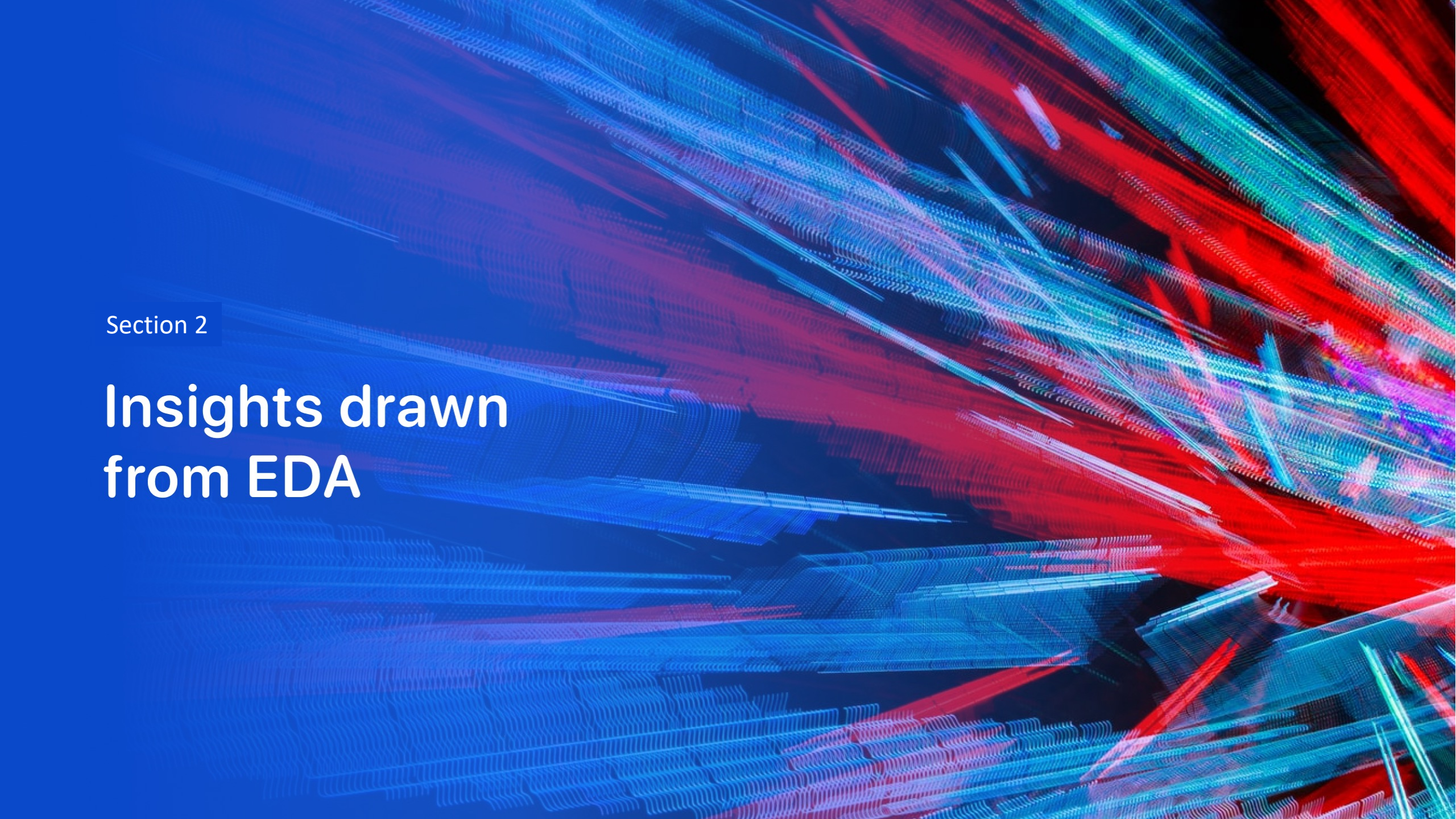
- **Interactive analytics demo in screenshots**



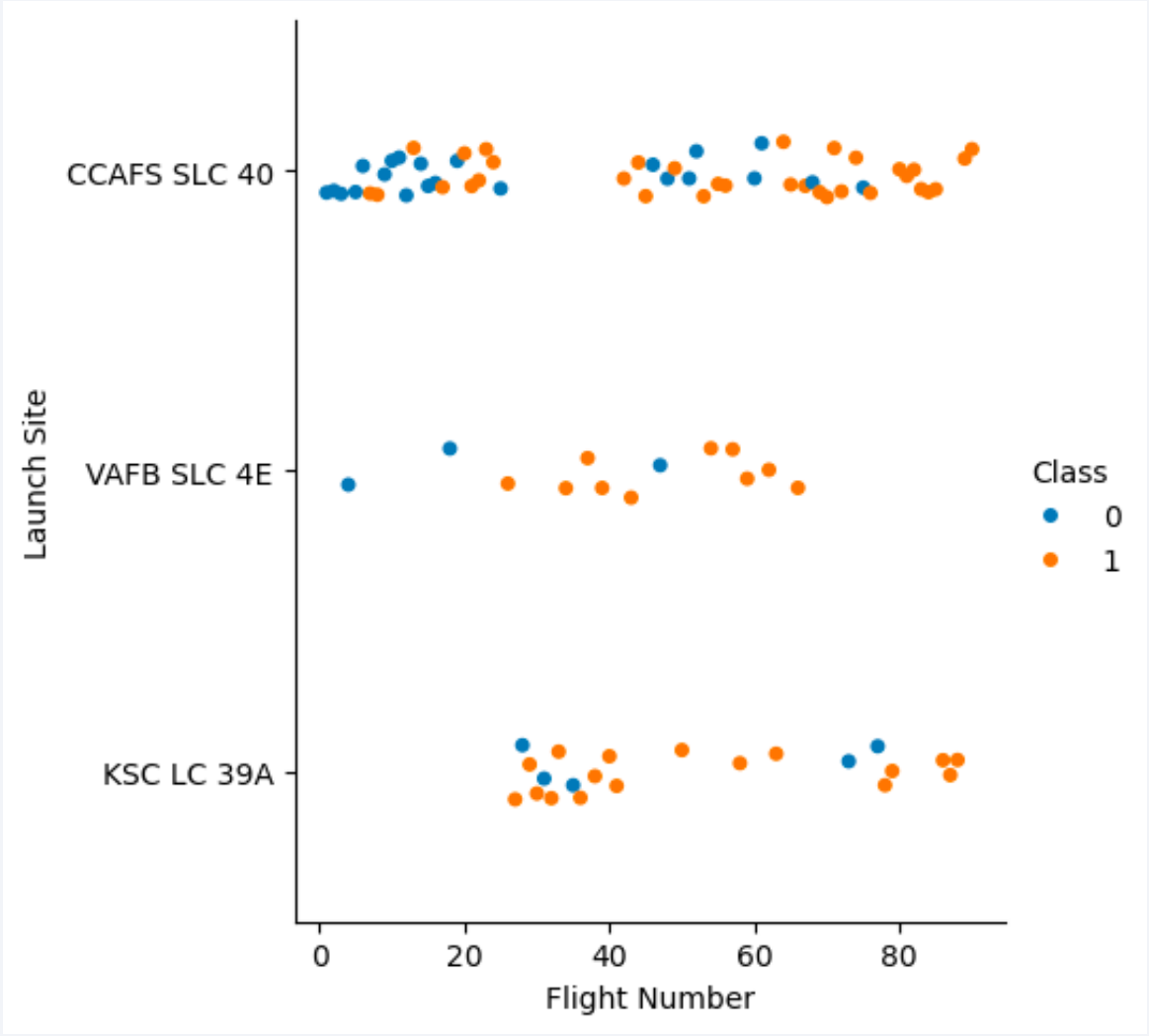


- **Predictive analysis results**

The model Decision Tree was applied and provided highest output accuracy rates.

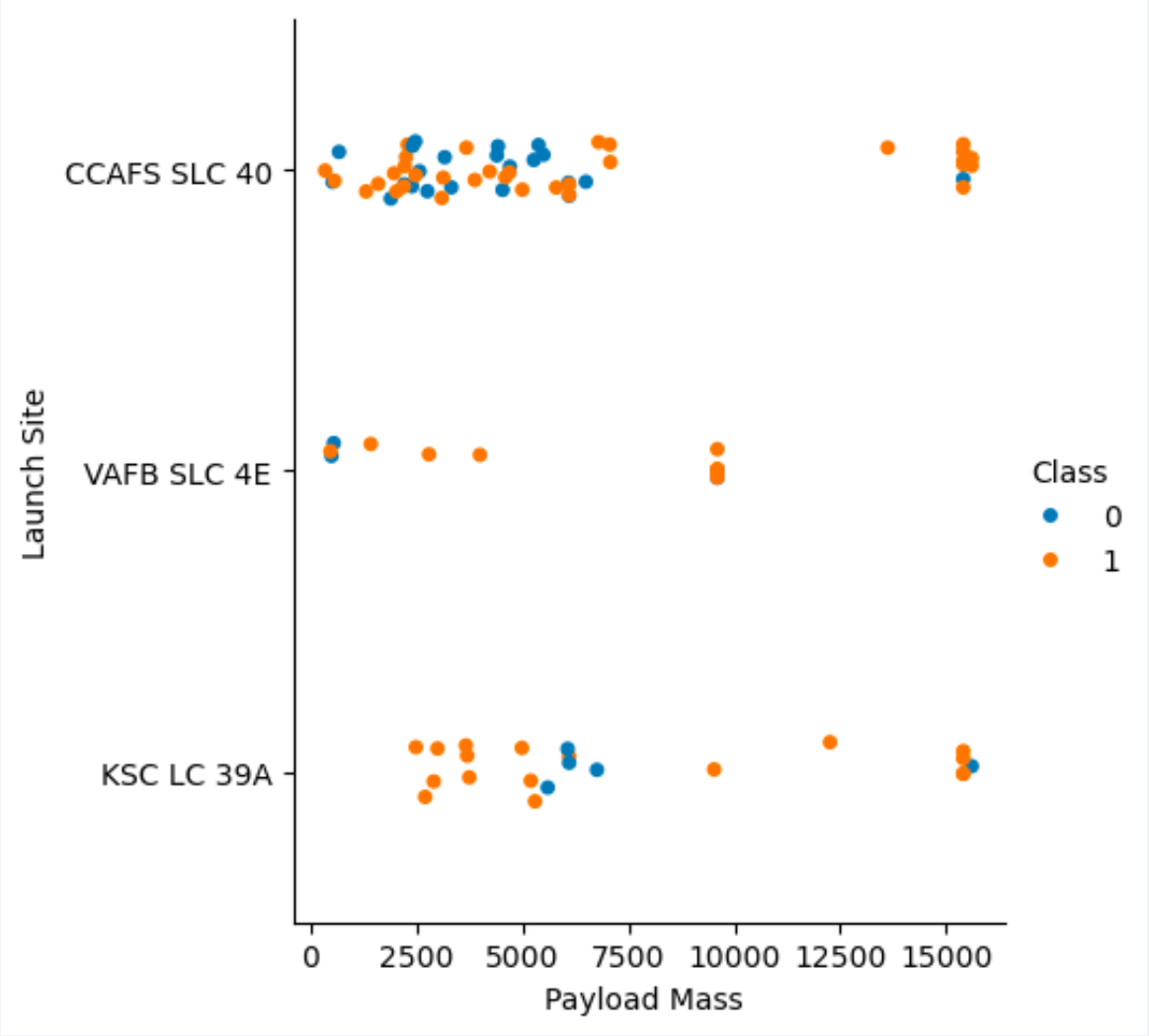

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



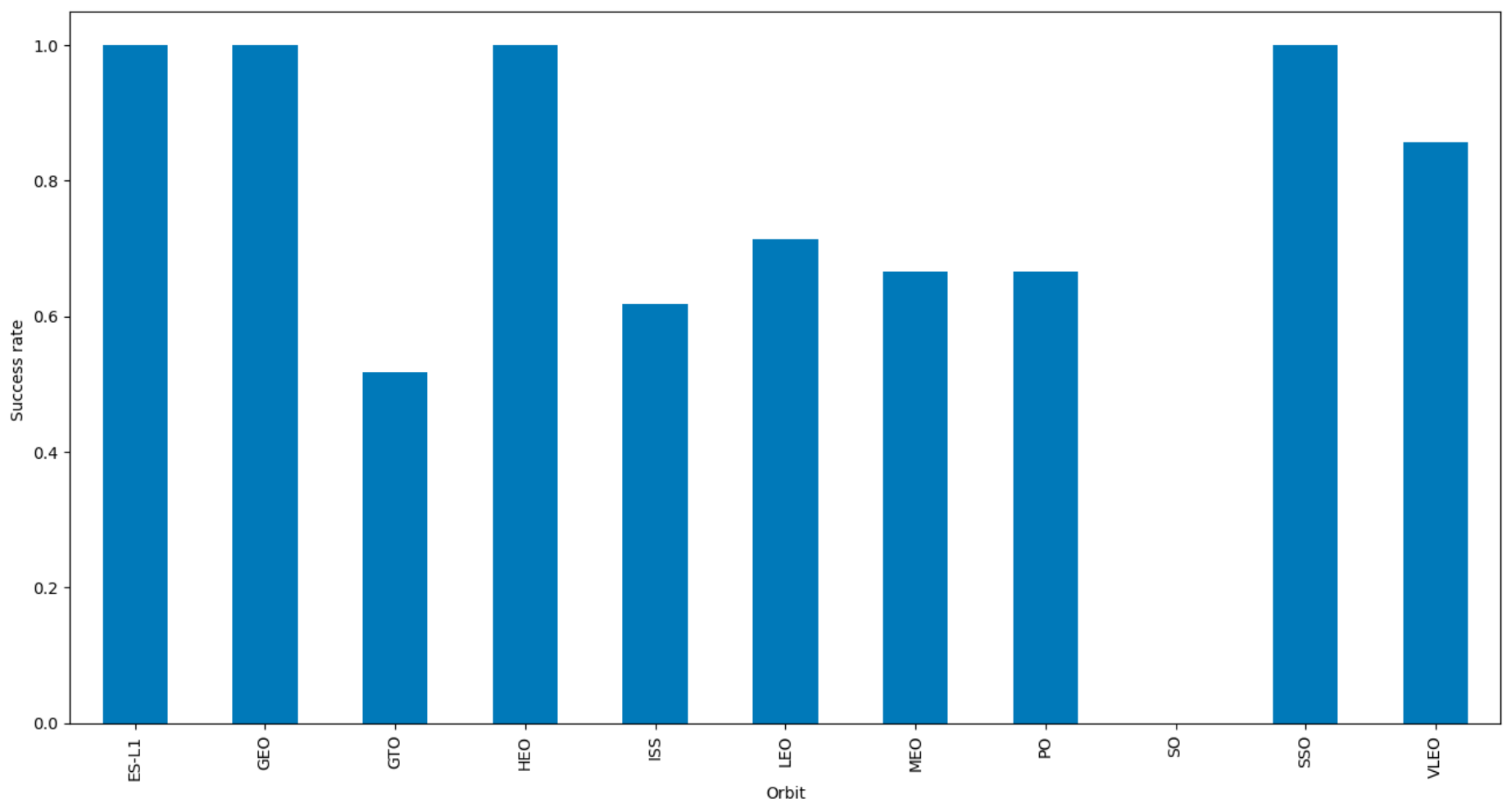
A scatter plot of Flight Number vs Launch Site

Payload vs. Launch Site



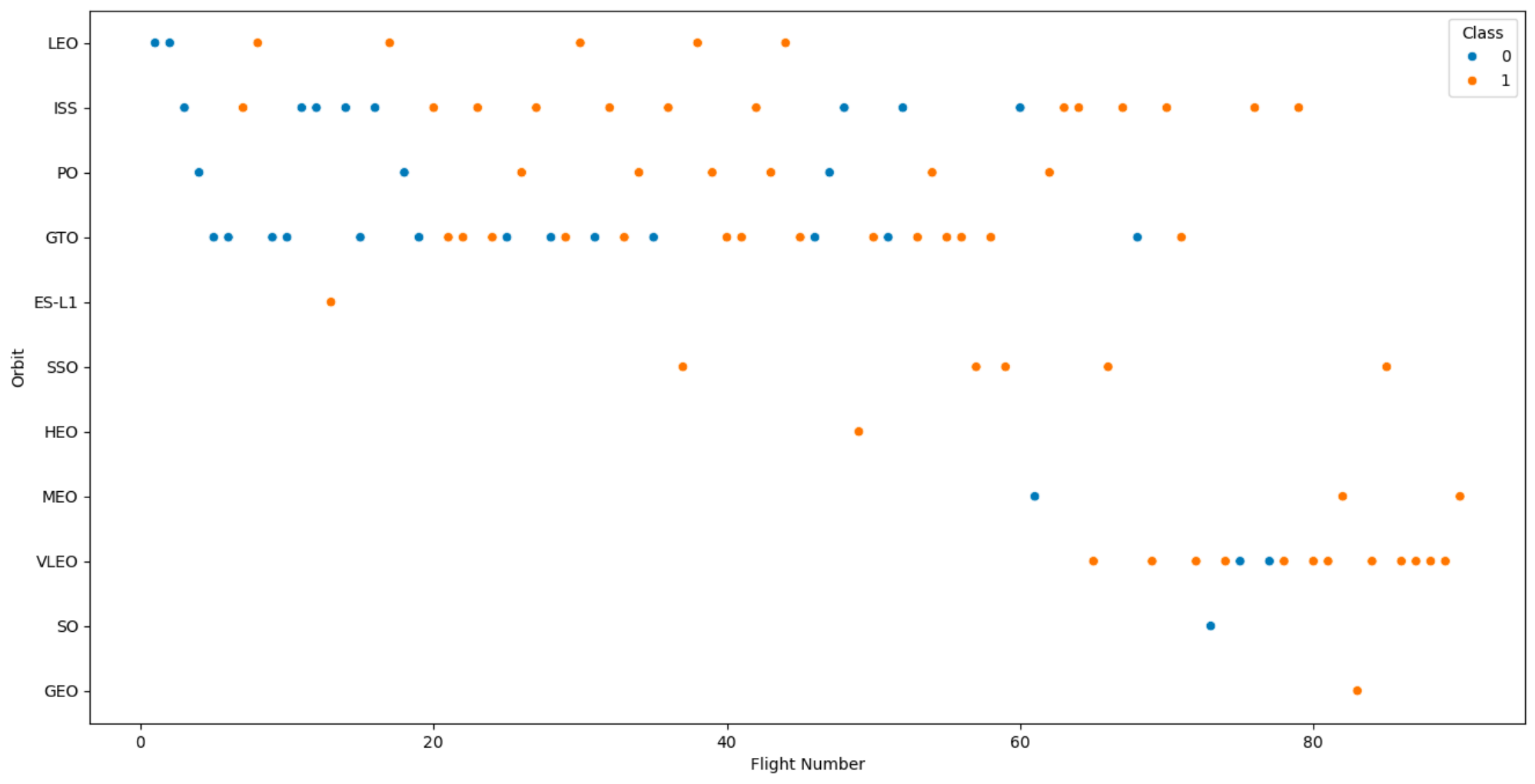
A scatter plot of Payload vs Launch Site

Success Rate vs. Orbit Type



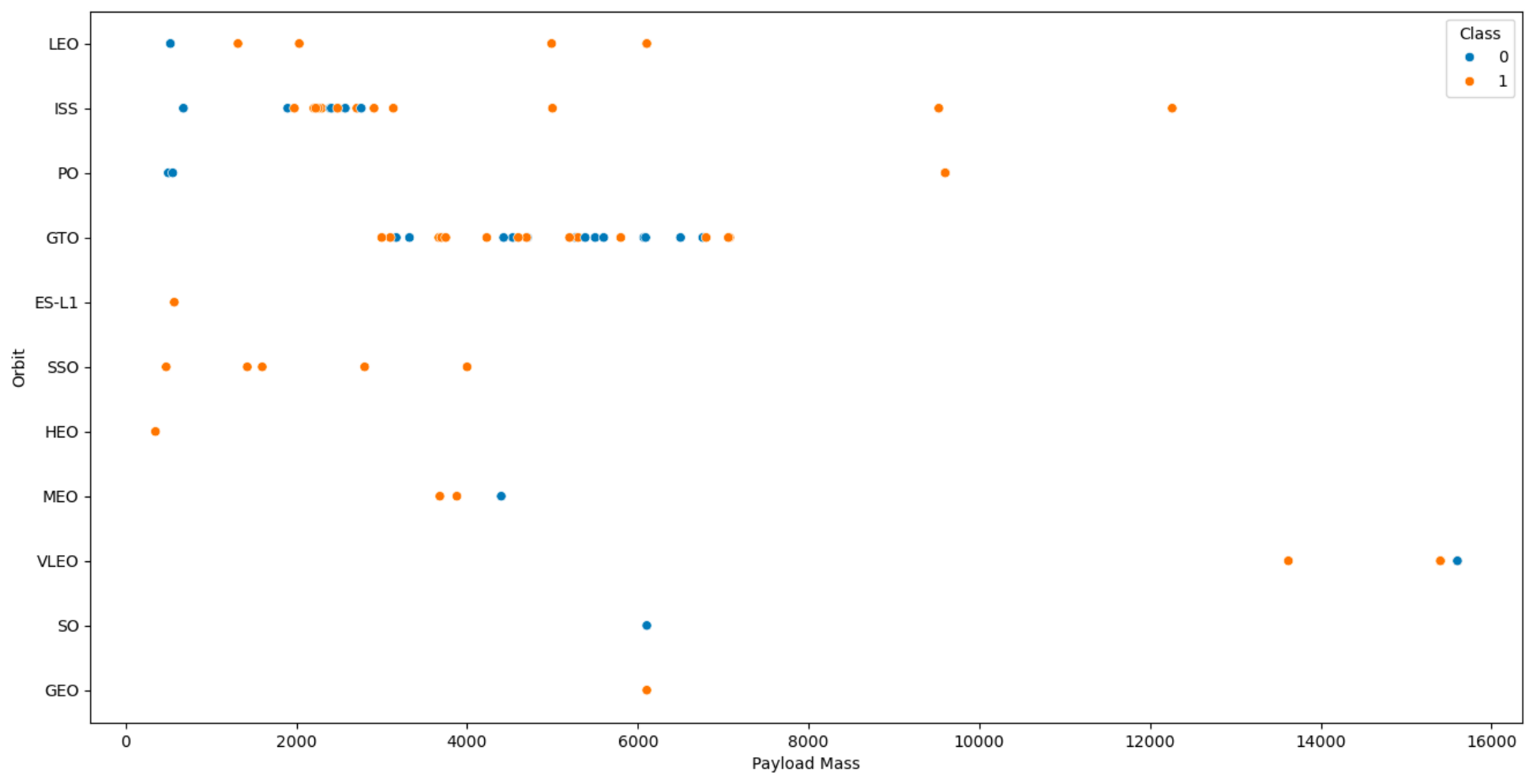
A bar chart for the success rate of each orbit type

Flight Number vs. Orbit Type



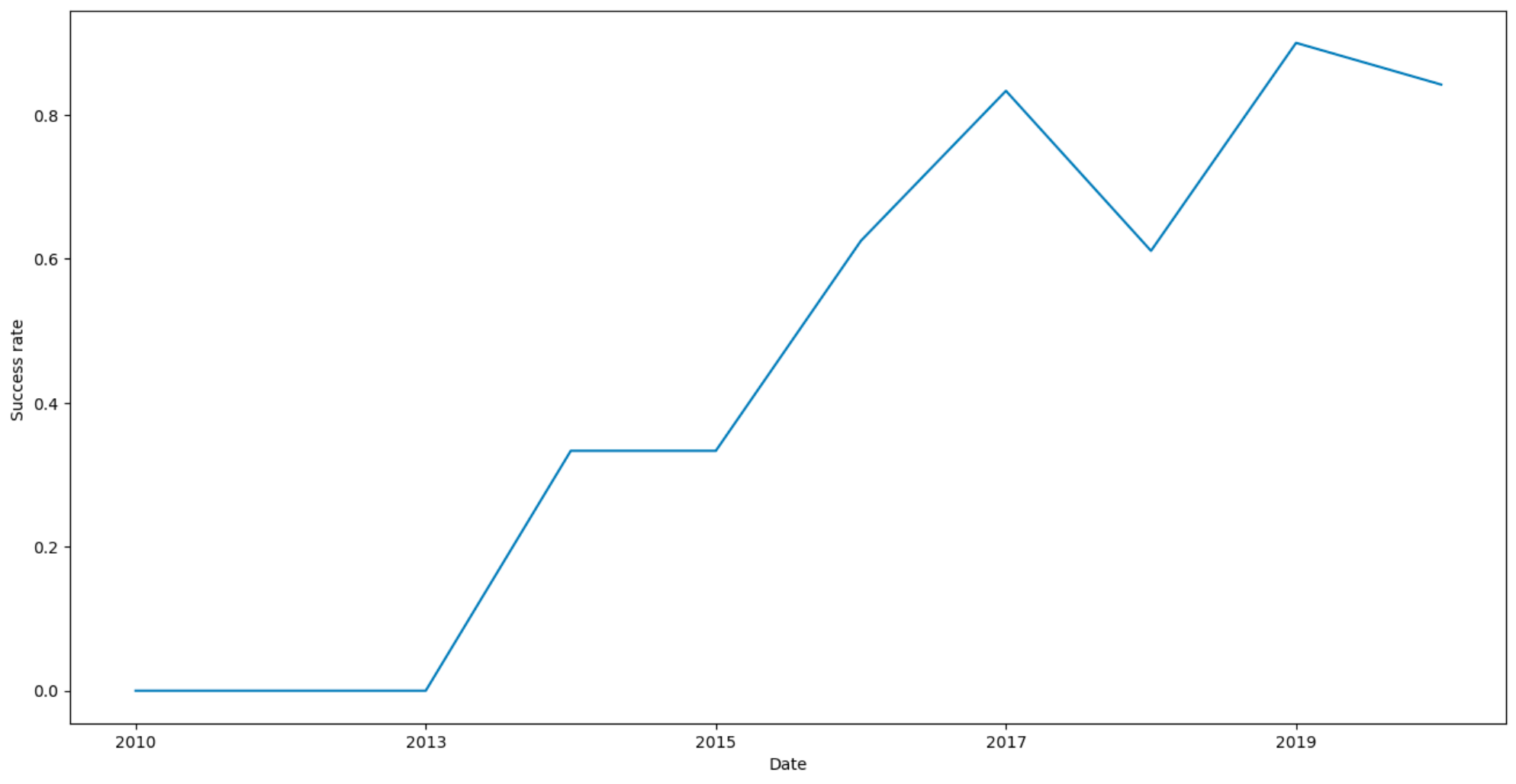
A scatter point of Flight number vs Orbit type

Payload vs. Orbit Type



A scatter point of payload vs orbit type

Launch Success Yearly Trend



A line chart of yearly average success rate

All Launch Site Names

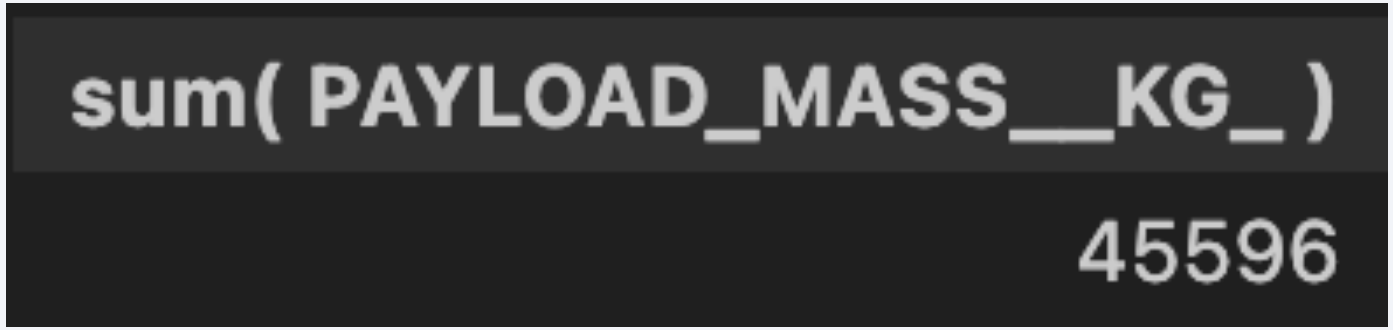
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

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

It’s just 5 records where launch sites begin with the string 'CCA’, not all

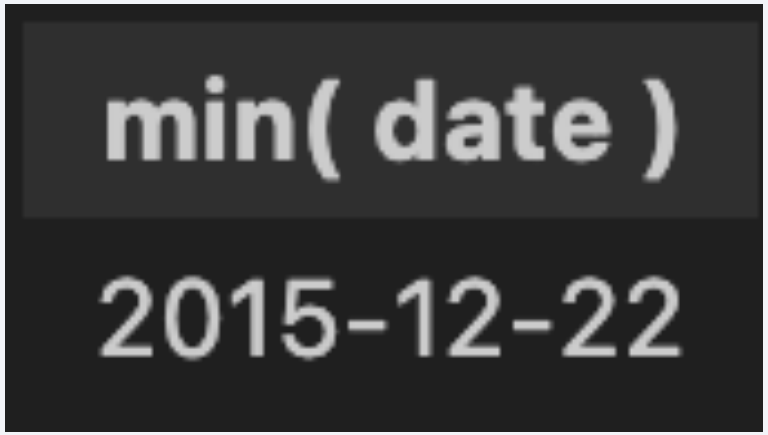
Total Payload Mass



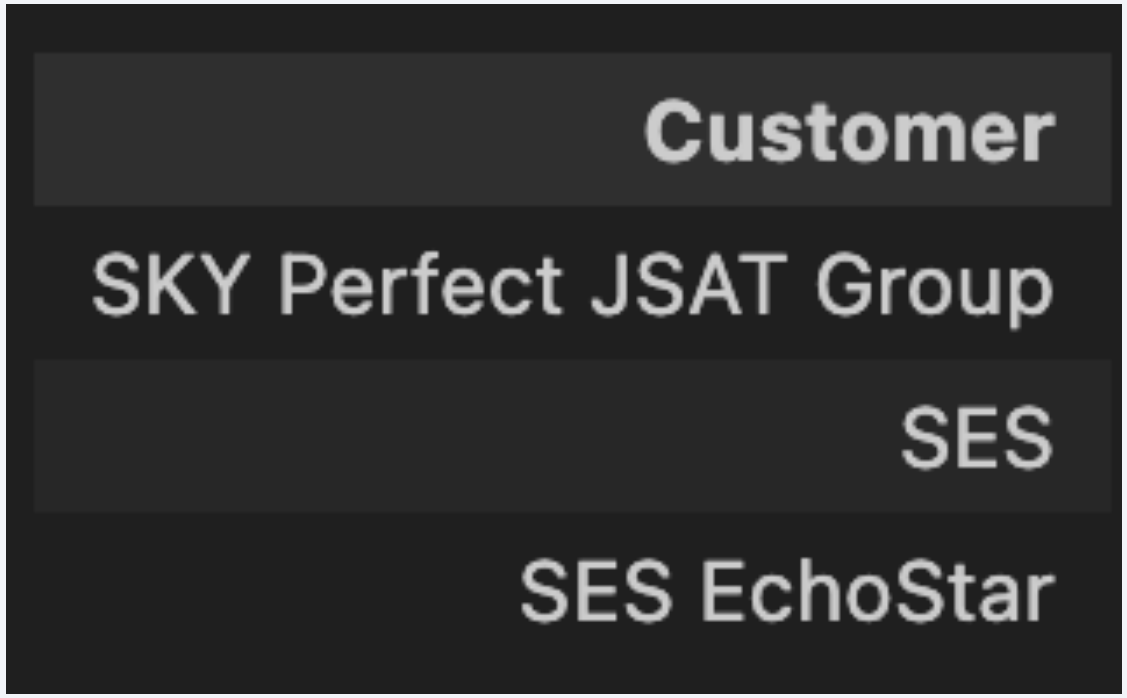
Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date



Successful Drone Ship Landing with Payload between 4000 and 6000



Total Number of Successful and Failure Mission Outcomes

Successful Mission Outcomes	Unsuccessful Mission Outcomes
100	1

Boosters Carried Maximum Payload

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

2015 Launch Records

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

Records of Failed Landing Outcomes on Drone Ships, Booster Versions, and Launch Sites for the Months in 2015

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

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

No attempt was the most frequent outcome, indicating many missions didn't attempt a landing, likely due to mission profiles.

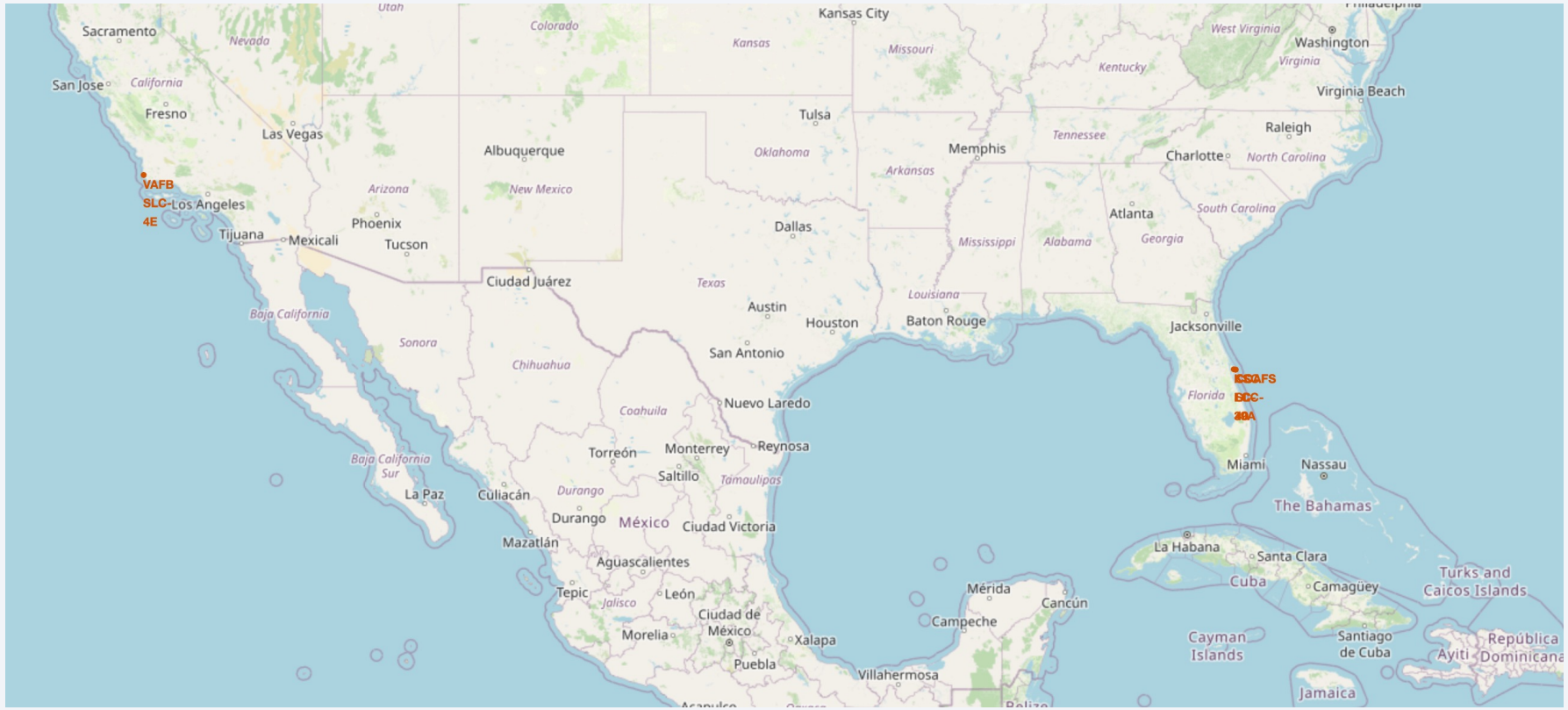
Drone ship landings showed balanced success and failure rates, suggesting they are more challenging compared to other methods.

Ground pad landings had a high success rate, highlighting their reliability compared to drone ships.

Ocean landings (controlled or uncontrolled) and parachute failures occurred infrequently, indicating these were less common or emergency scenarios.

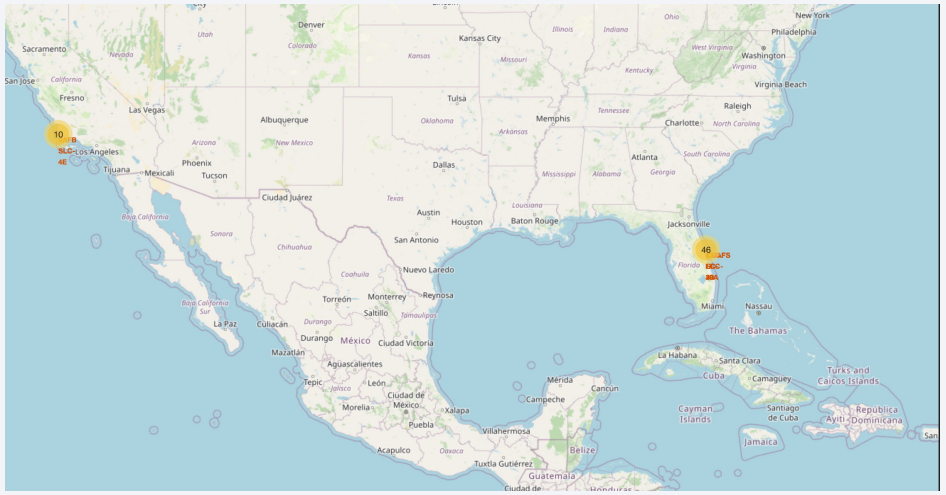

Section 3
Launch Sites
Proximities Analysis

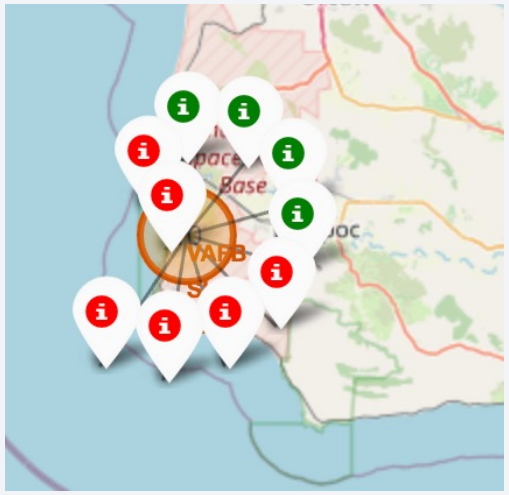
Locations of launch sites

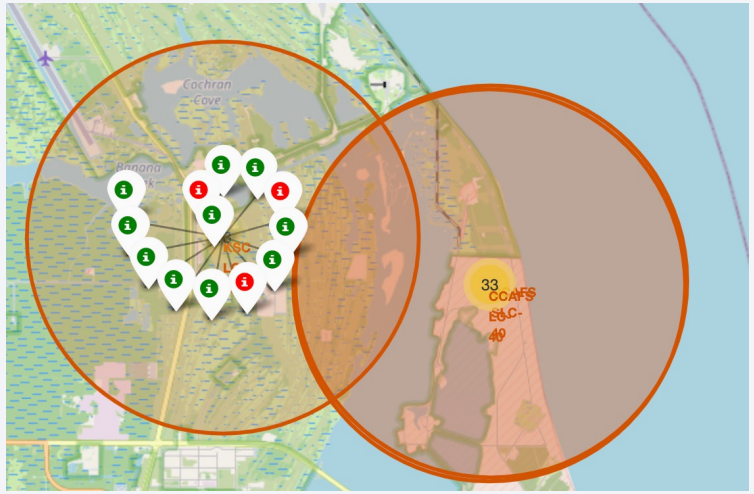


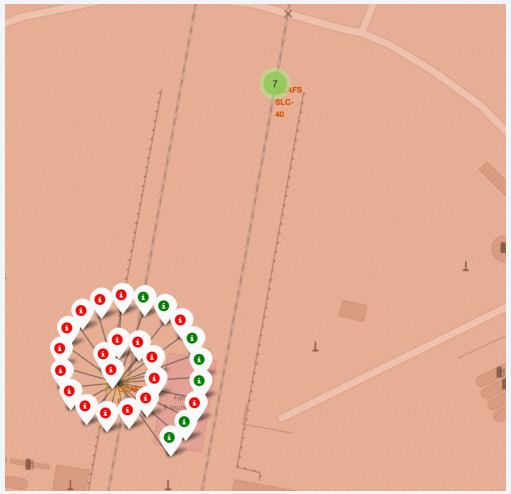
Both VAFB and CCAFS areas share geographic similarities as coastal, flat-terrain launch sites near large bodies of water. Their proximity to oceans supports safe launch trajectories and recovery operations, making them ideal locations for space missions.

Success and Failure Rates







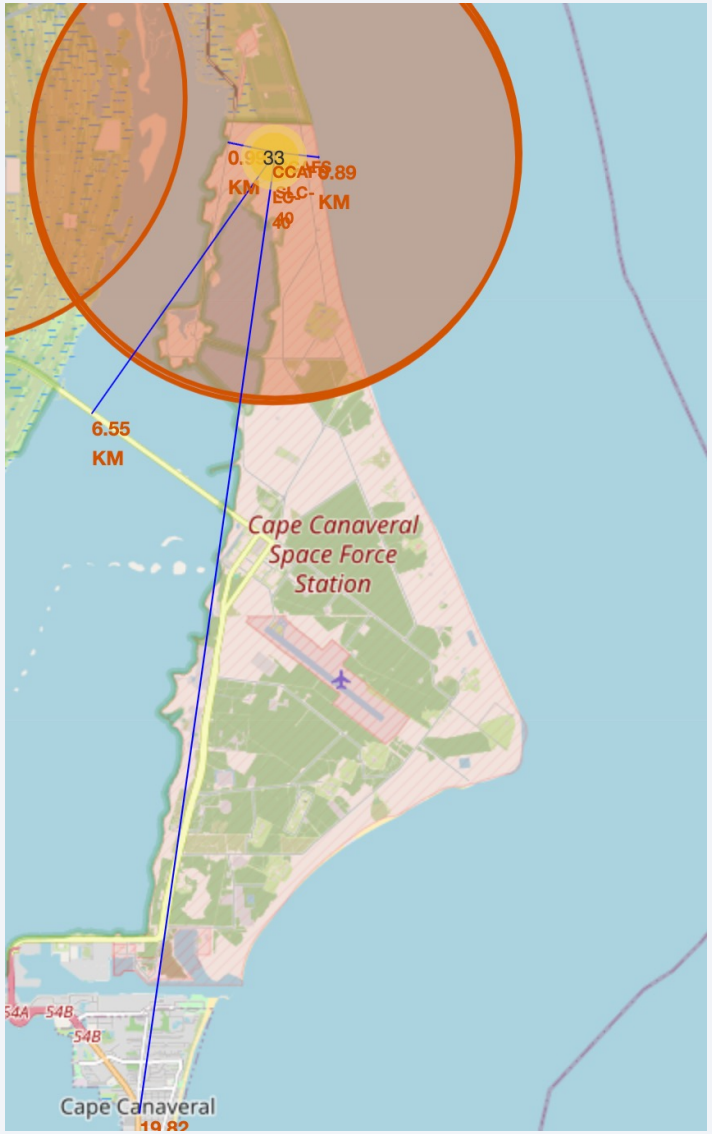


1.High Failure Rate at CCAFS: The CCAFS areas have a much higher failure rate, suggesting more complex or challenging missions are launched from this site.

2.KSC-LC 39A's High Success Rate: KSC-LC 39A shows strong reliability with a higher success rate, likely due to better infrastructure or less challenging launches.

3.Balanced Outcomes at VAFB-SLC 4E: VAFB-SLC 4E has a balanced success/failure record, indicating either improvements or more complex missions.

Geospatial Insights



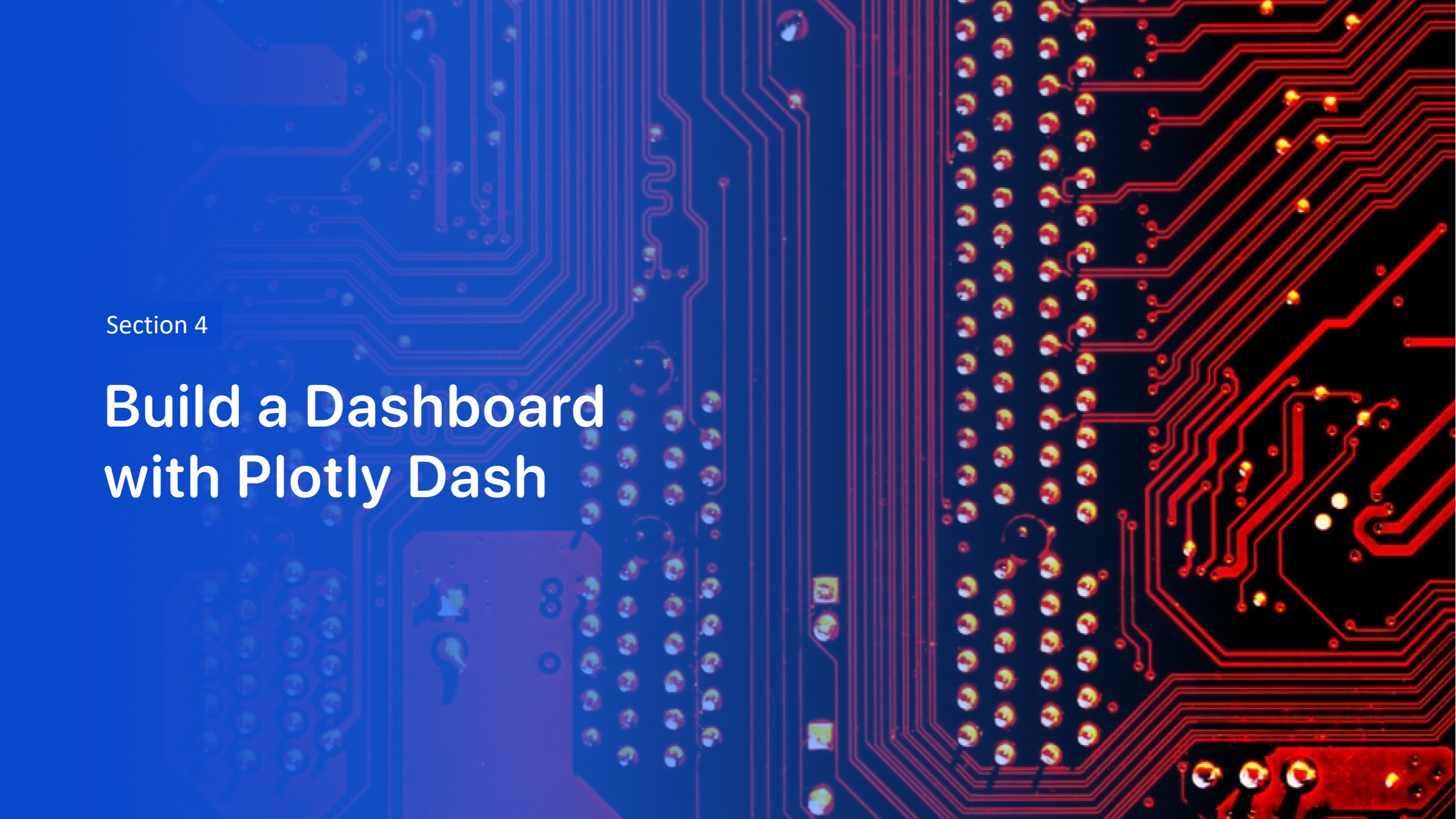
Insights:

1.Proximity to Railways: Launch sites are close to railways, likely for easy transportation of heavy rocket components and equipment.

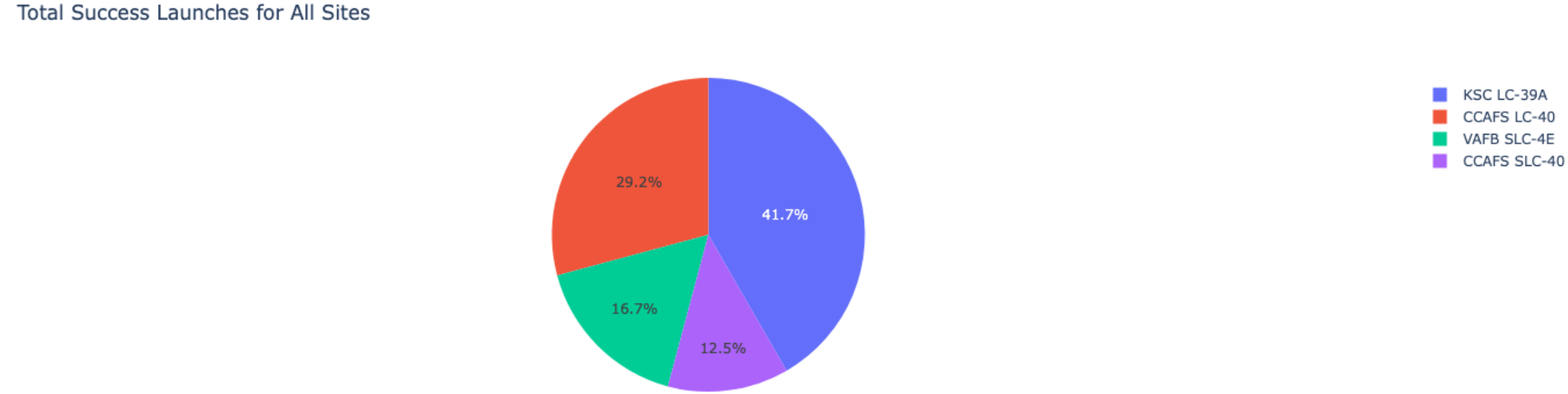
2.Distance from Highways: Launch sites are not near highways, possibly to reduce risks in case of accidents or launch failures.

3.Proximity to Coastlines: All launch sites are near coastlines, providing safe over-water flight paths and easier recovery options for rockets.

4.Distance from Cities: Launch sites maintain a safe distance from cities to ensure public safety and minimize the risk of damage in case of failures.

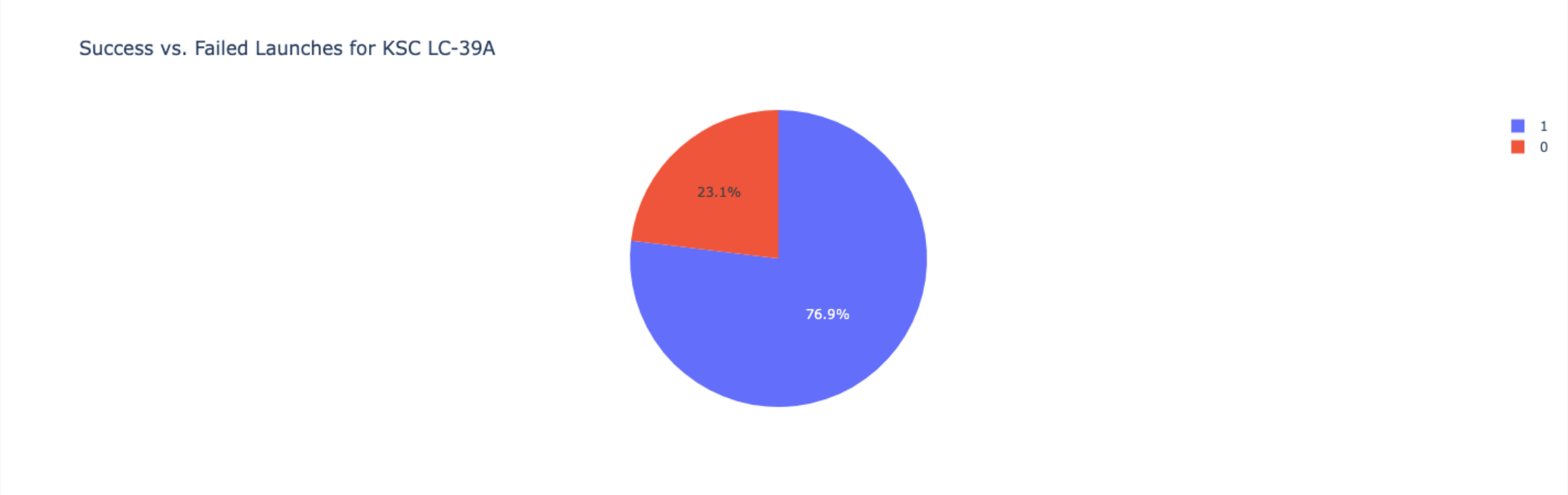

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites



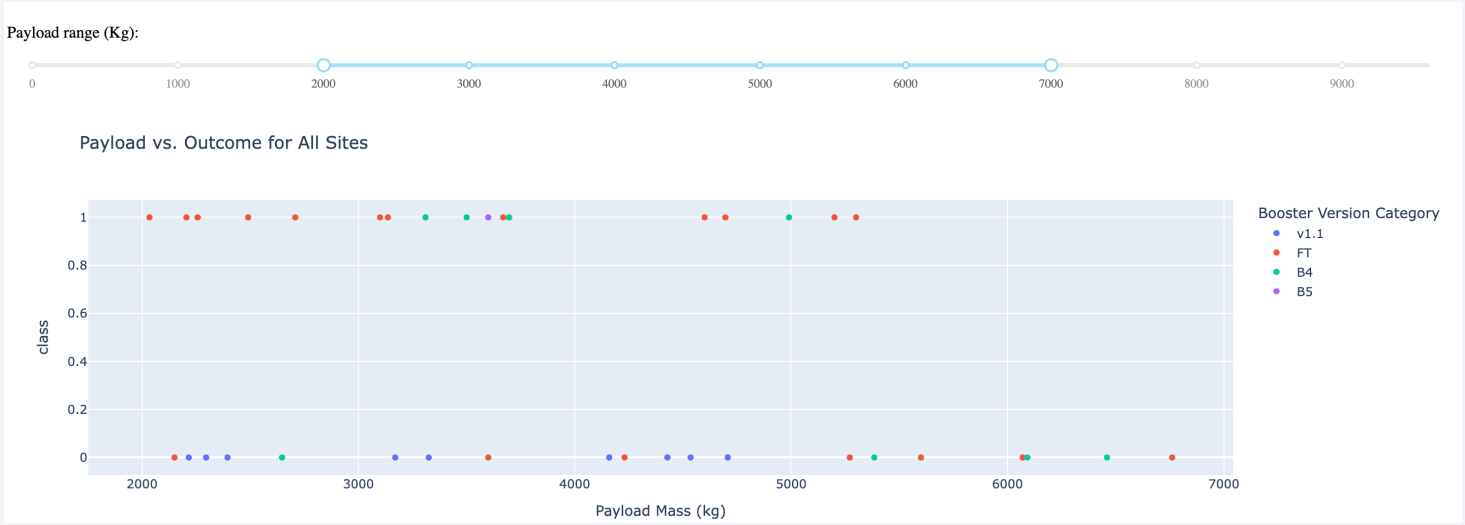
+ This figure helps us understand the relative success rates of different launch sites, with **KSC LC-39A** standing out as the most successful.
+ Besides, it shows that **VAFB** and **CCAFS** both contribute significantly, though some sites perform better than others, which is crucial for improving predictive models based on location.

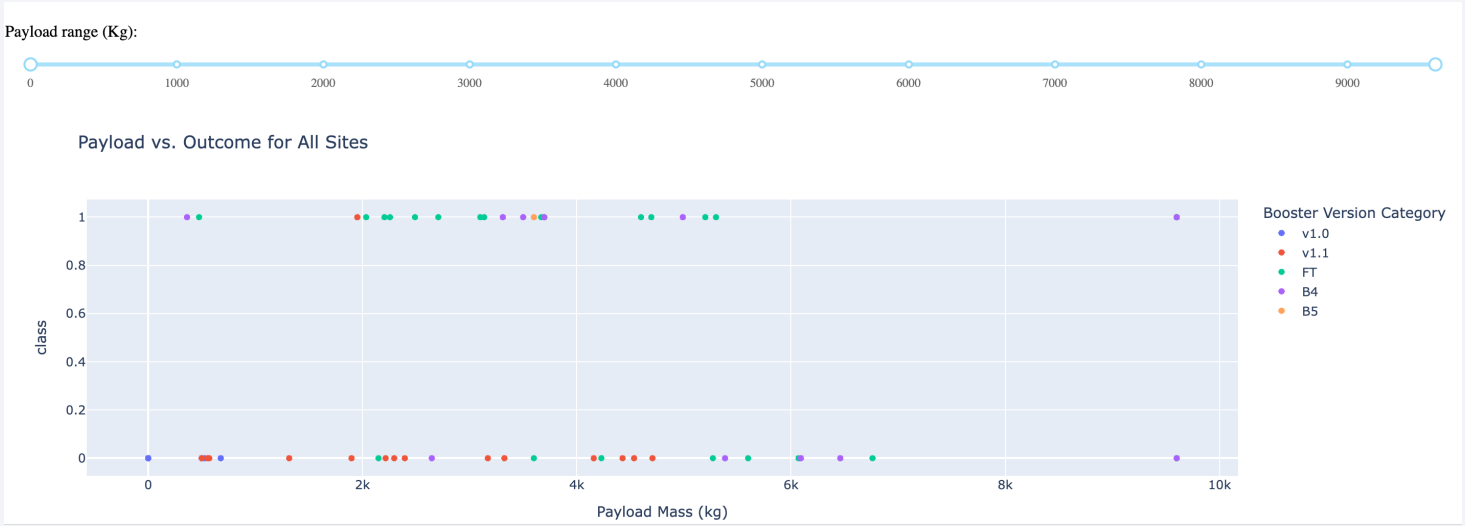
KSC LC-39A – The highest launch success rate



+ This chart emphasizes **KSC LC-39A's** role as a high-performing launch site with a strong success rate, making it crucial for predictive analysis and future mission planning.
+ The high success ratio reinforces **KSC LC-39A** as a preferred site for missions aiming for successful outcomes. This can be factored into models to predict the likelihood of success based on location.

Payload ~ Outcome for All Sites





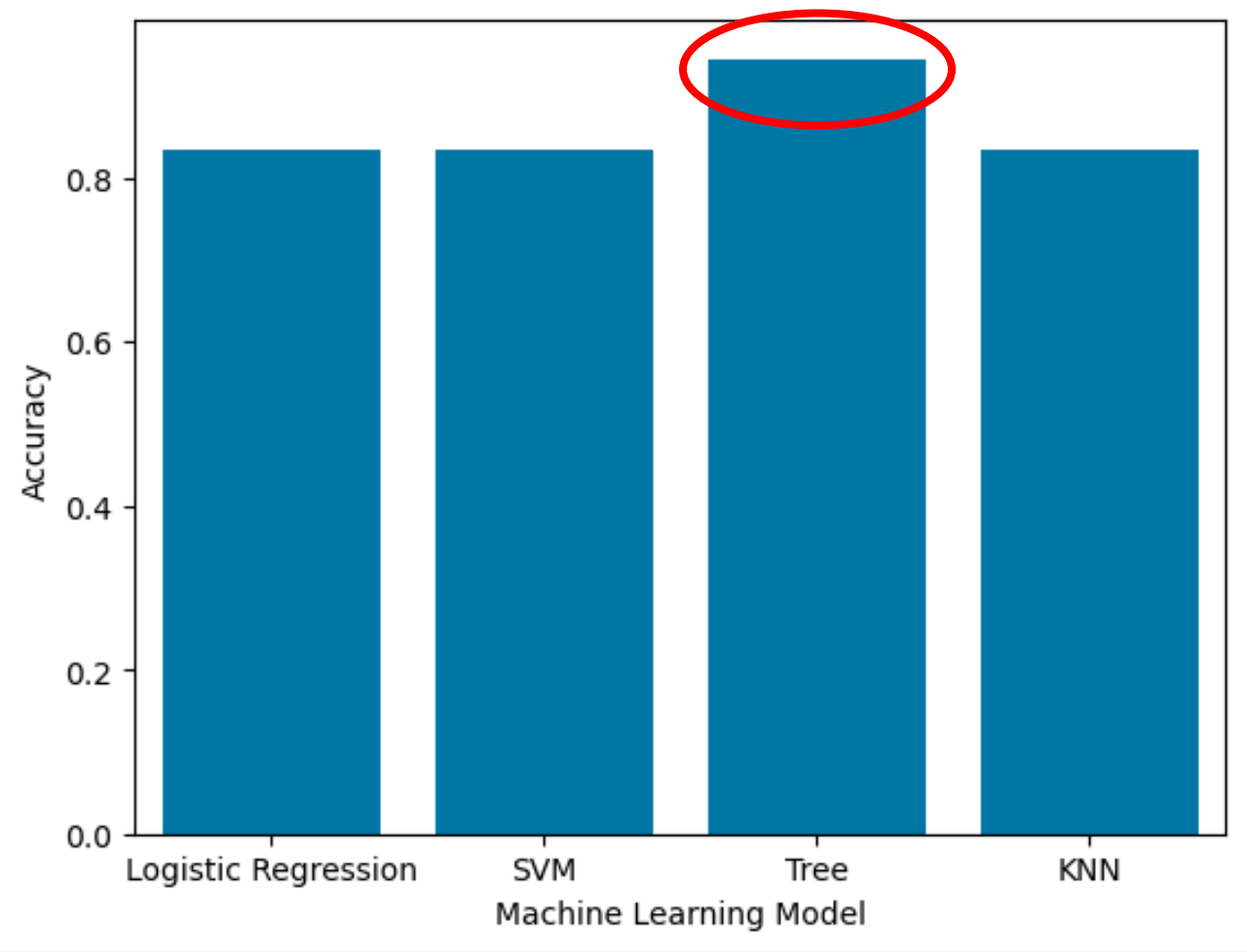
Booster Version Insight:
The strong performance of newer booster versions like **FT** and **B4** for a wide range of payloads can be used to improve predictions on launch outcomes.

Payload Mass and Success Rate:
The clear relationship between lighter payloads and higher success rates can help inform future mission planning and model training for predicting successful landings based on payload size.

Section 5

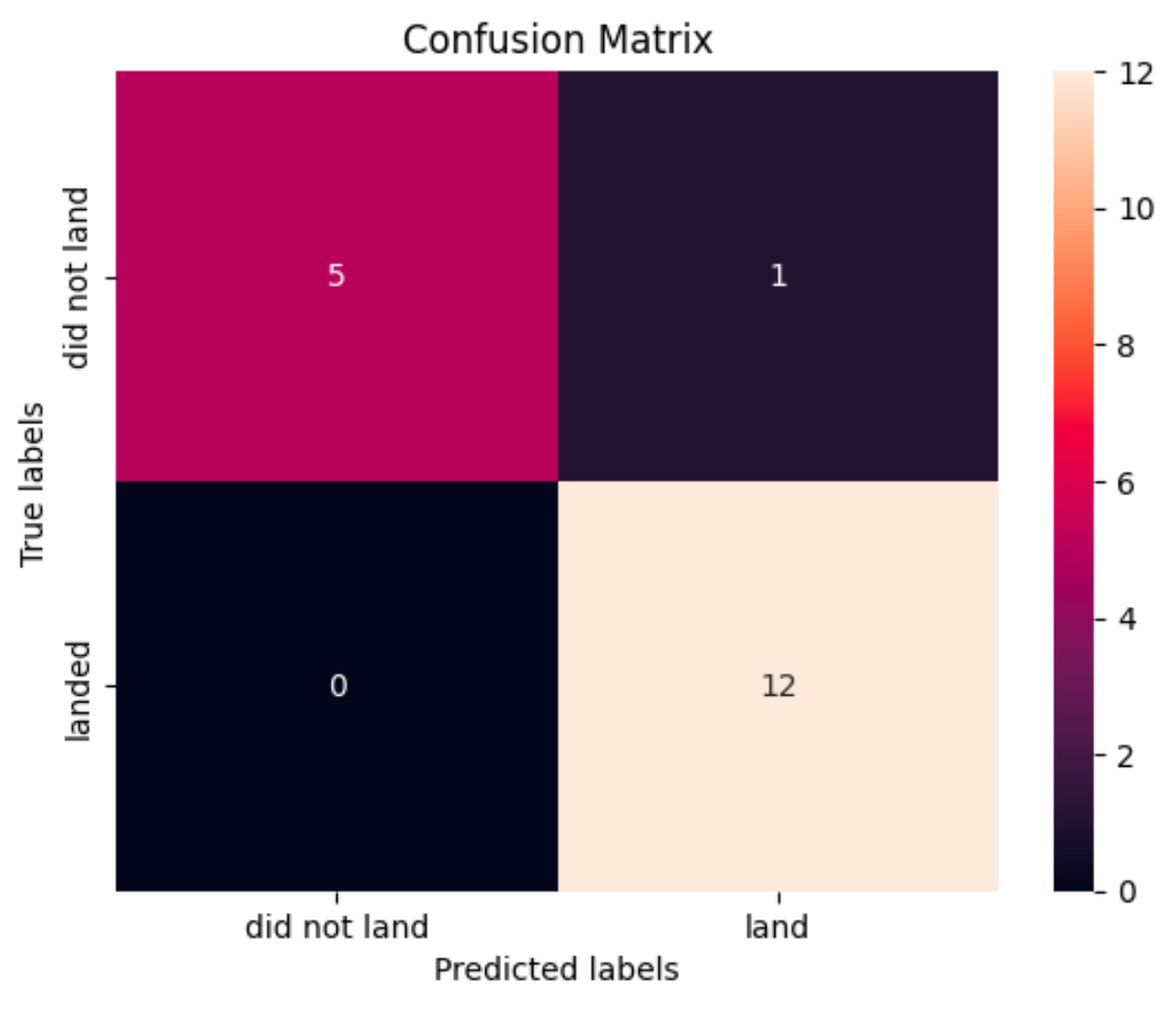
Predictive Analysis (Classification)

Classification Accuracy



The best model was Decision Tree, with the highest accuracy when applied with test dataset, around 0.94

Confusion Matrix



This Decision Tree model performs well, with high accuracy: minimal false positives and no false negatives. This is crucial for the project because the goal is to predict whether the Falcon 9 first stage will land successfully. Accurate predictions enable Space Y to estimate the likelihood of reusability, which directly affects launch costs and operational efficiency.

By using the Decision Tree model, we can confidently make cost-saving decisions regarding rocket reusability, improving Space Y's competitiveness against SpaceX in the commercial space industry.

Conclusions

Insights:

+ The reuse of the first stage has a significant impact on reducing launch costs. As seen from the landing success rates (as reflected in pie charts and scatter plots), certain factors like launch site, payload mass, and orbit type influence the success of the first stage landing. Successful landings, especially with newer booster versions (e.g., FT, B4), occur more frequently, which allows SpaceX to reuse the stage, dramatically cutting costs.

+ From the confusion matrix and model performance evaluation (like the Decision Tree), we see that the machine learning models, particularly the Decision Tree, perform well in predicting the outcomes of landings. The model has a high accuracy rate, with minimal misclassifications, as reflected by the confusion matrix.

Conclusion:

+ The data strongly indicates that the more successful landings SpaceX achieves, the lower the launch costs due to reusability. This has been reflected in launch site performance, where sites like KSC LC-39A have consistently higher success rates, contributing to more frequent first-stage reuse.

+ The Decision Tree Model shows high accuracy and reliability in predicting whether the first stage will land successfully. This indicates that it is highly feasible to use machine learning to predict first-stage reusability, allowing Space Y to make informed decisions about launch costs and strategy.

Overall Conclusion:

By analyzing the launch data and using machine learning models like the Decision Tree, we can conclude that first-stage reuse plays a pivotal role in reducing costs, and a well-tuned model can accurately predict reusability, helping Space Y compete effectively with SpaceX.

Thank you!