IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ROHIT ADITYA RAMBHATLA

Outline

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

Executive Summary

Methodologies Used for Data Analysis

1. **Data Collection**:
    1. Information was gathered through **web scraping** and the **SpaceX API**, enabling access to relevant public data sources.
2. **Exploratory Data Analysis (EDA)**:
    1. The collected data was processed through **data wrangling**, ensuring it was clean and structured for analysis.
    2. Various **visualization techniques** were applied to identify trends and patterns.
    3. An **interactive EDA approach** was used to explore key insights dynamically.
3. **Machine Learning Prediction**:
    1. Predictive models were developed to determine the most significant factors influencing the success of SpaceX launches.
    2. The best-performing model was identified, highlighting the key characteristics that drive successful launches.

Summary of Findings

- **Valuable data** was successfully collected from publicly available sources.
- **EDA helped identify** the most critical features that influence launch success.
- **Machine learning models** provided insights into the most impactful factors, allowing for data-driven decision-making to optimize future launches.

Introduction

The goal of this analysis is to **assess the feasibility of the new company, Space Y, in competing with SpaceX** by analyzing key factors influencing rocket launches.

Key Considerations

1. Estimating the Total Cost of Launches

One of the primary cost drivers in space missions is the **successful landing of the first-stage rocket**, as reusability significantly reduces expenses. By leveraging historical data on SpaceX launches, a **predictive model** can estimate the likelihood of successful landings. This allows Space Y to optimize its strategy, **minimizing costs** and **maximizing reusability** for competitive pricing.

2. Determining the Best Launch Location

The choice of launch site plays a crucial role in **mission success, efficiency, and cost-effectiveness**. Factors such as **proximity to the equator**, **weather conditions**, and **air traffic control restrictions** influence the selection. Data analysis can identify the **most optimal locations** based on past successful launches, ensuring higher efficiency for Space Y.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API (link)
  - Web Scrapping from Wikipedia (link)
- Perform data wrangling

Cleaned and structured the raw data for analysis. Cleaned and structured the raw data for analysis.

- 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 collected data was normalized to ensure consistency and then split into training and test datasets for model evaluation. Four different classification models were tested, each using various parameter combinations. The accuracy of each model was measured to determine the best-performing one for predicting launch success

Data Collection

The data was gathered from two main sources:

1. **SpaceX API**

Retrieved structured launch data, including rocket specifications, mission details, and landing outcomes.

2. **Web Scraping**

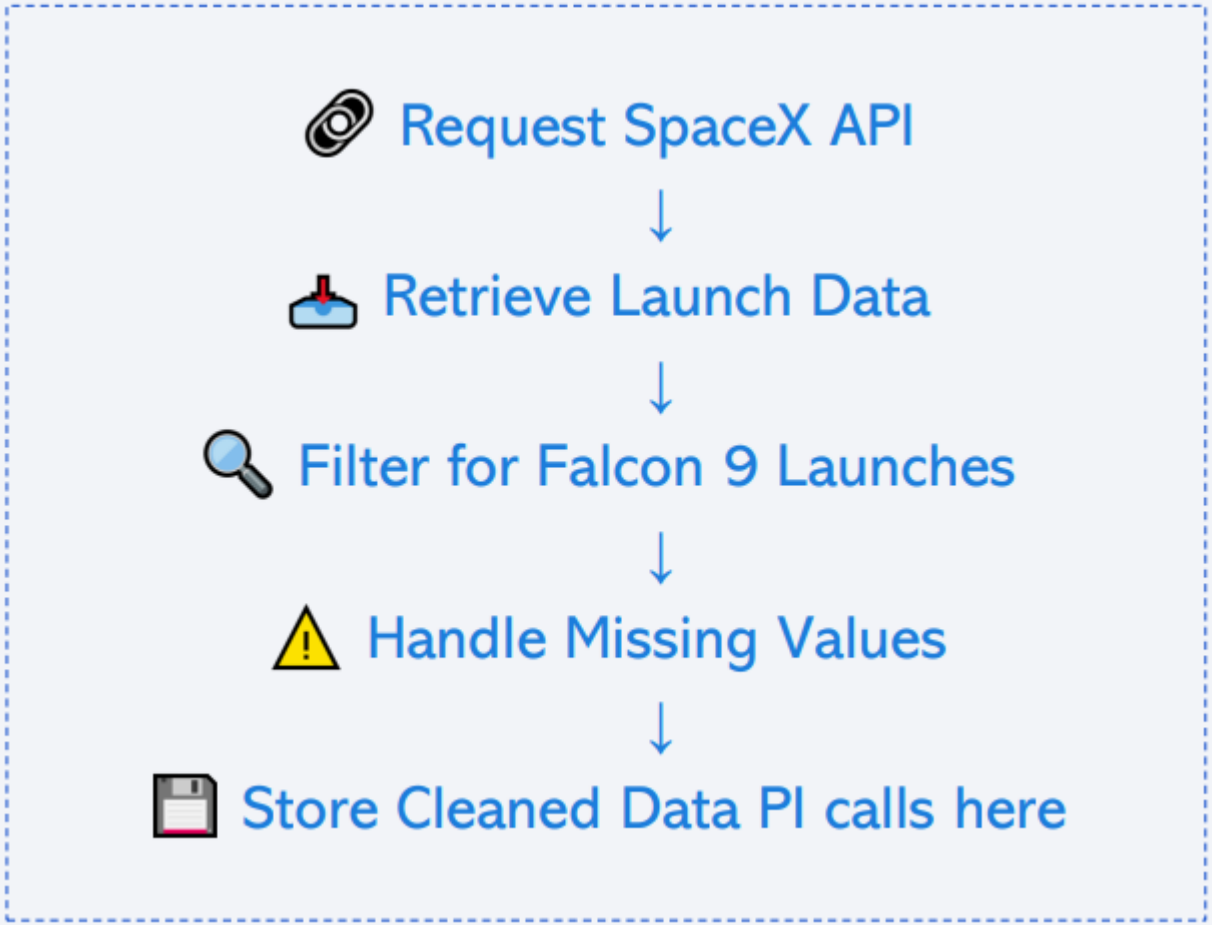
Extracted additional launch details to supplement API data. Here's a simple flowchart to illustrate the process:

Fetch Data → Clean & Merge → Analyze & Label → Store & Use for Modeling

Data Collection – SpaceX API

The API was used to **fetch important launch details**, following a structured process (as shown in the flowchart).

After retrieving the data, it was **cleaned and stored** for further analysis and modeling



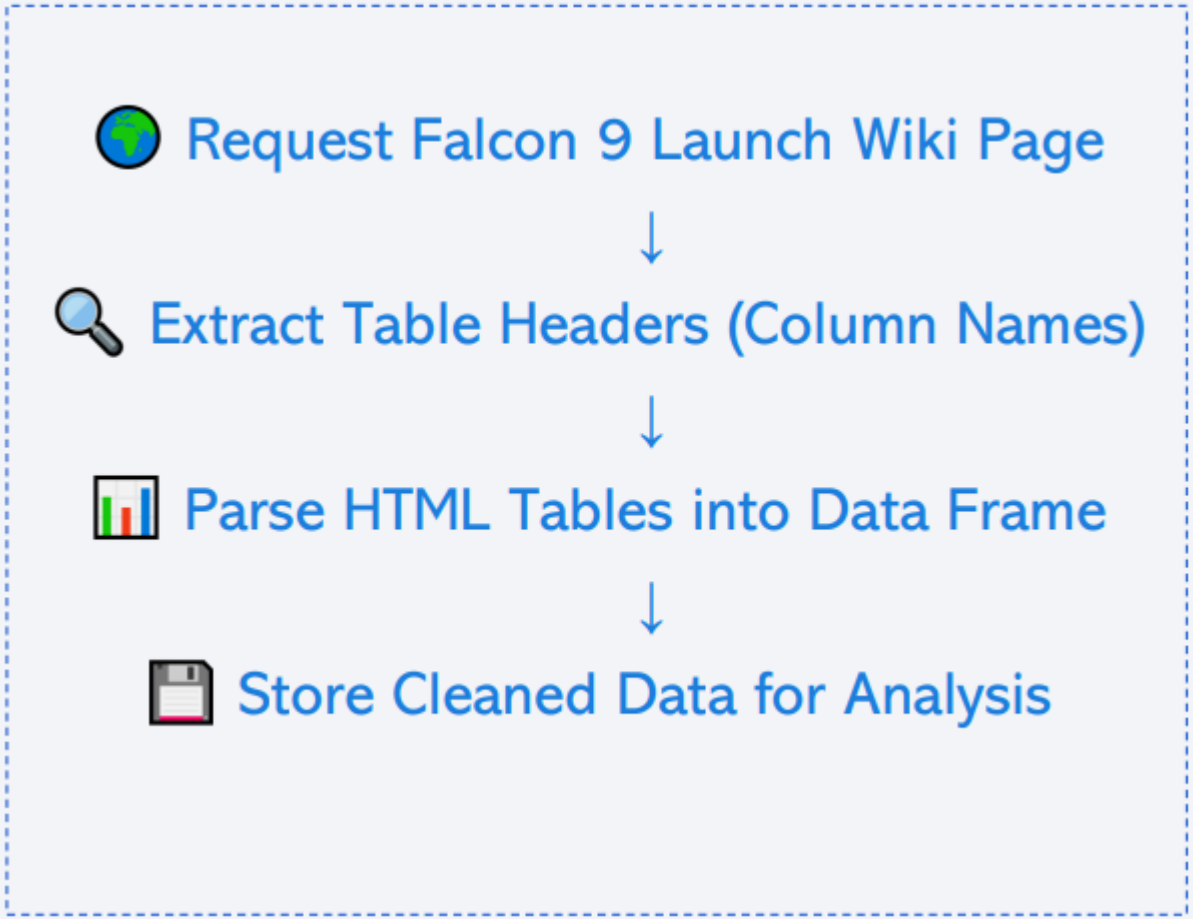
Source Code:
https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

The data was **extracted, cleaned, and structured** following the flowchart process.

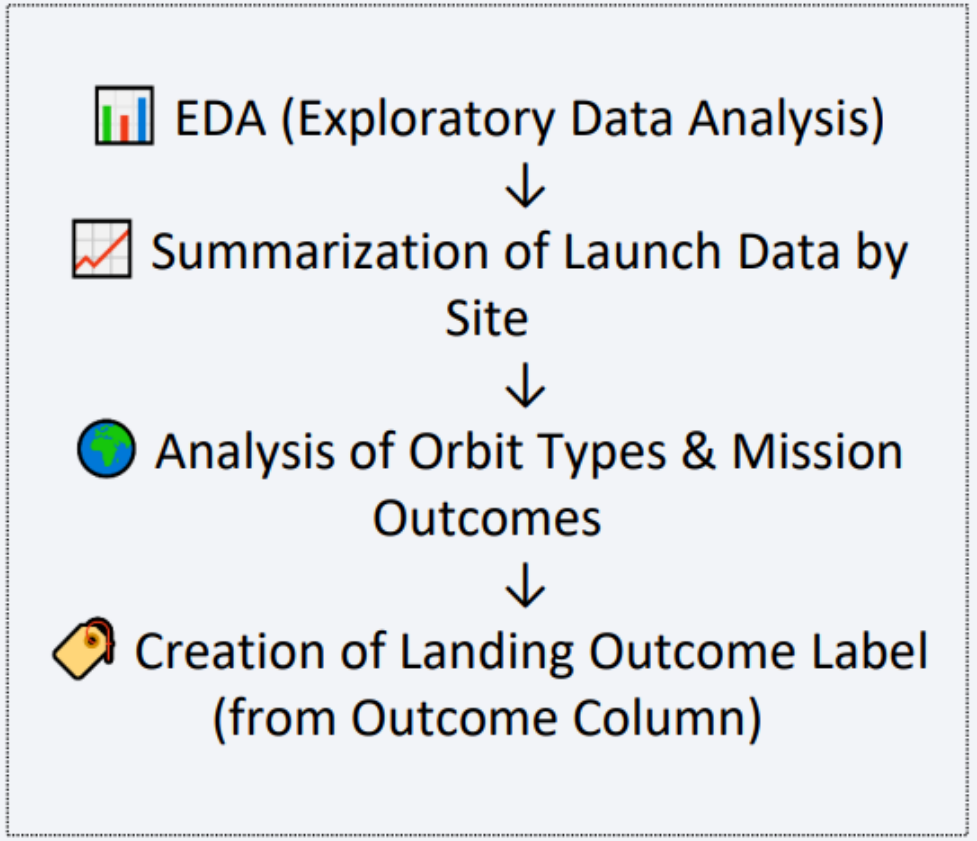
After processing, the cleaned data was **stored for further analysis**.

Source Code:
https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

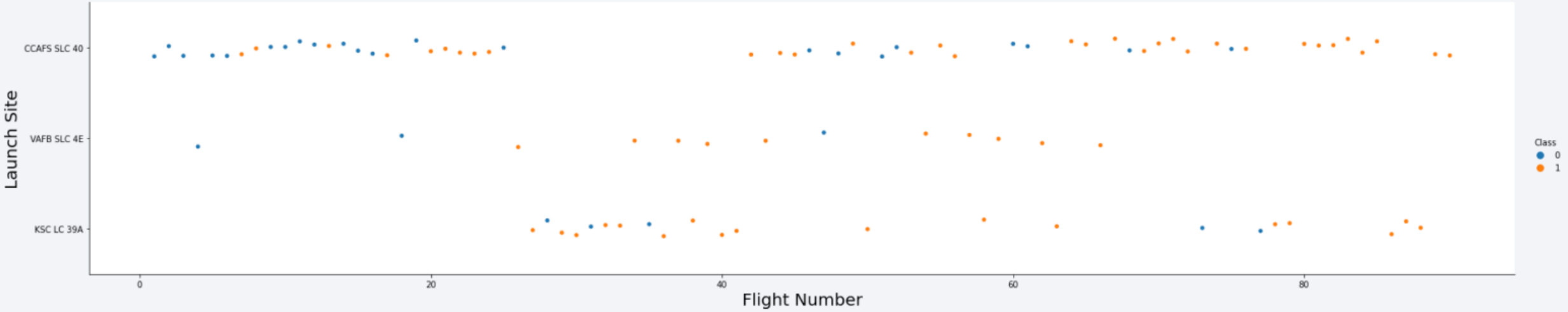
- The first step was to explore and understand the dataset, identifying its key features and checking for any patterns or irregularities.
- We calculated the **total number of launches** for each launch site, helping us determine which sites had the most frequent or successful launches.
- We analyzed the distribution of **orbit types** used across different missions and examined how **mission outcomes** (success/failure) varied by orbit type.
- Based on the information in the "Outcome" column, we created **landing outcome labels** to categorize launches as successful or failed. This would be used for future predictive modeling



Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

Scatter and bar plots were used to explore key relationships. **Payload Mass vs. Flight Number** showed payload trends over time, while **Launch Site vs. Flight Number** highlighted site activity. **Launch Site vs. Payload Mass** identified sites handling heavier payloads, and **Orbit Type vs. Flight Number** revealed orbit usage trends. Lastly, **Payload Mass vs. Orbit Type** showed how payload weight influenced orbit selection



Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

SQL Queries Performed

- Retrieved **unique launch site names** from the dataset.
- Identified the **top 5 launch sites** starting with "CCA".
- Calculated the **total payload mass** for boosters launched by NASA (CRS).
- Found the **average payload mass** carried by the **F9 v1.1** booster version.
- Determined the **date of the first successful landing** on a ground pad.
- Listed **boosters that successfully landed on a drone ship** with payloads between **4000-6000 kg**.
- Counted the **total number of successful and failed missions**.
- Identified booster versions that carried the **maximum payload mass**.
- Analyzed **failed landings on drone ships in 2015**, including booster versions and launch sites.
- Ranked the **frequency of landing outcomes** (e.g., "Failure (drone ship)" or "Success (ground pad)") between **2010-06-04 and 2017-03-20**.

Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

With **Folium Maps**, different markers and shapes were used to visualize various features:

- **Markers**: Used to represent specific points, such as **launch sites**.
- **Circles**: Highlighted areas around important locations, like the **NASA Johnson Space Center**, to show the surrounding region.
- **Marker Clusters**: Grouped multiple events (e.g., **launches** at a launch site) into clusters to represent multiple occurrences at the same location.
- **Lines**: Showed the **distances** between two coordinates, helping visualize connections or travel paths between locations.

Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Data Visualization Insights

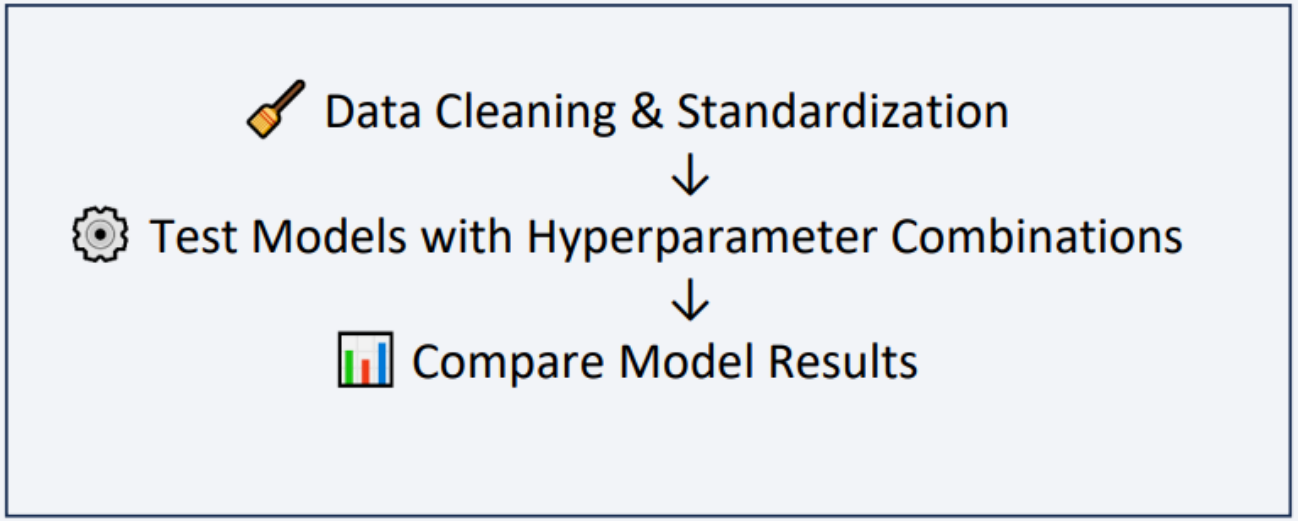
- **Percentage of launches by site**: Showed the distribution of launches across different sites.
- **Payload range**: Displayed the range of payloads for each launch.

This combination of visualizations enabled a quick analysis of the relationship between **payload sizes** and **launch sites**, helping identify the most optimal launch locations based on payload capacities.

Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

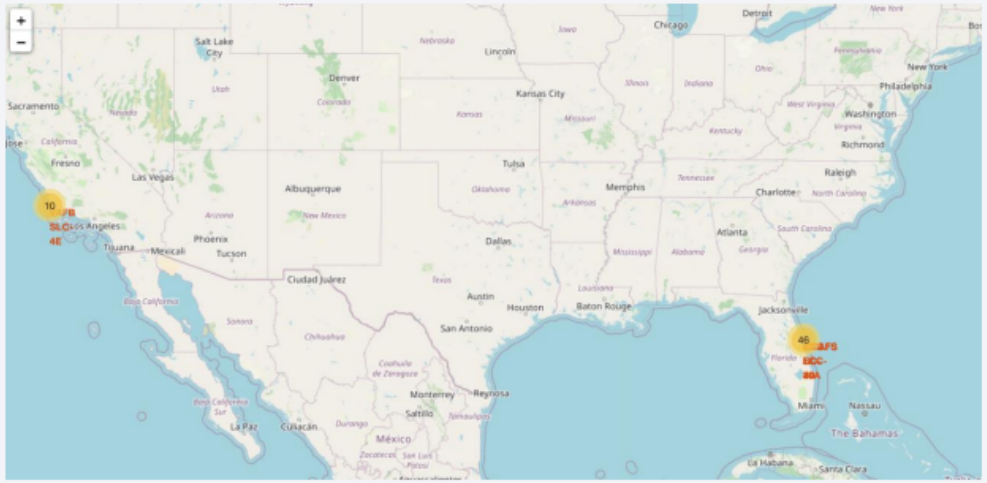
Four **classification models** were evaluated and compared to predict launch outcomes



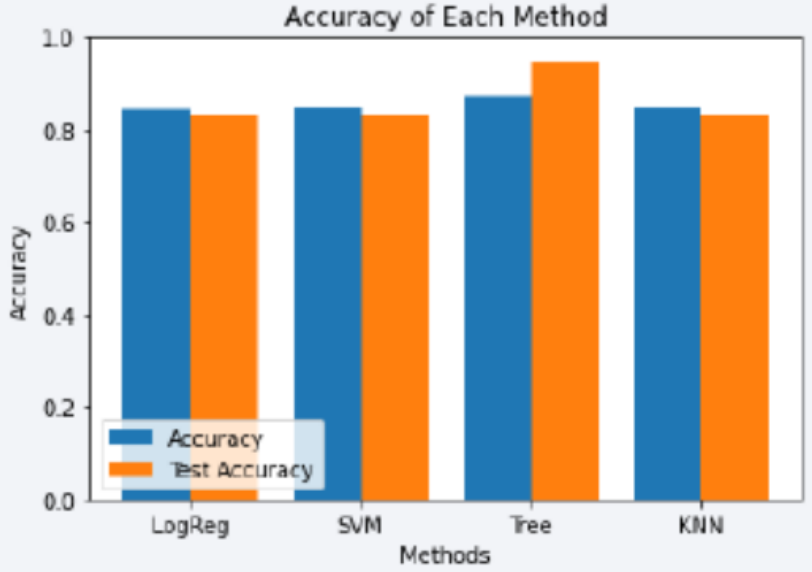
Source Code: https://github.com/Culossa/Applied-Data-Science-Captstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

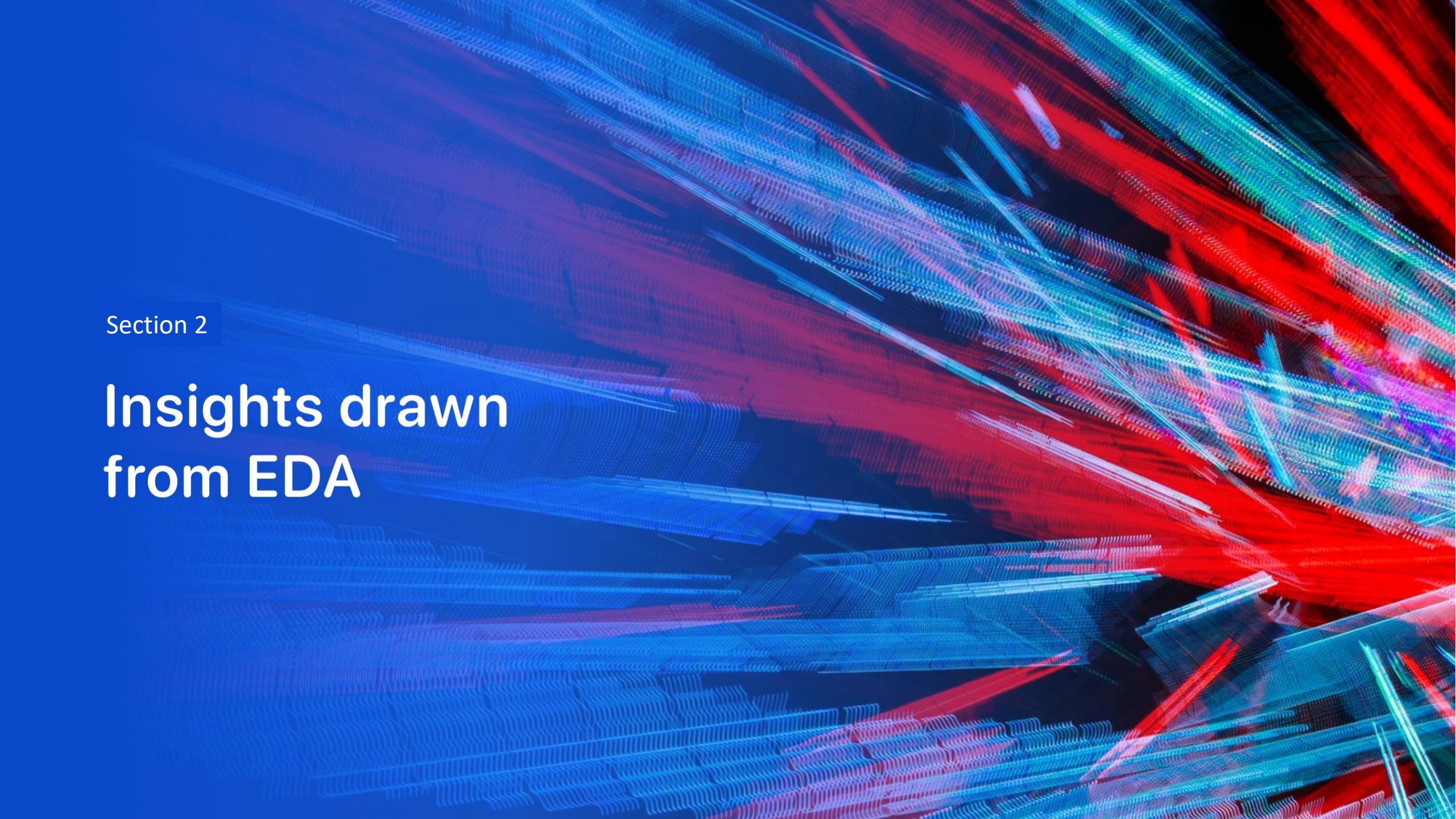
- **SpaceX Launch Sites**: SpaceX operates **4 different launch sites** for its missions.
- **Initial Launches**: The **first launches** were conducted for **SpaceX itself and NASA**.
- **Average Payload of F9 v1.1**: The **average payload** carried by the **F9 v1.1 booster** is **2,928 kg**.
- **First Successful Landing**: The **first successful landing** of a rocket occurred in **2015**, five years after the initial launch.
- **Successful Drone Ship Landings**: Many **Falcon 9 booster versions** successfully landed on **drone ships**, particularly those carrying **payloads above the average**.
- **Mission Success Rate**: Nearly **100% of mission outcomes** were successful, showcasing SpaceX's reliability.
- **Failed Landings in 2015**: In **2015**, two **F9 v1.1 booster versions** (**B1012** and **B1015**) failed to land on **drone ships**.



Interactive analytics helped reveal that **launch sites** are typically located in **safe areas**, often near the **sea**, and are equipped with **strong logistical infrastructure** to support operations. This ensures both safety and efficient handling of launches.



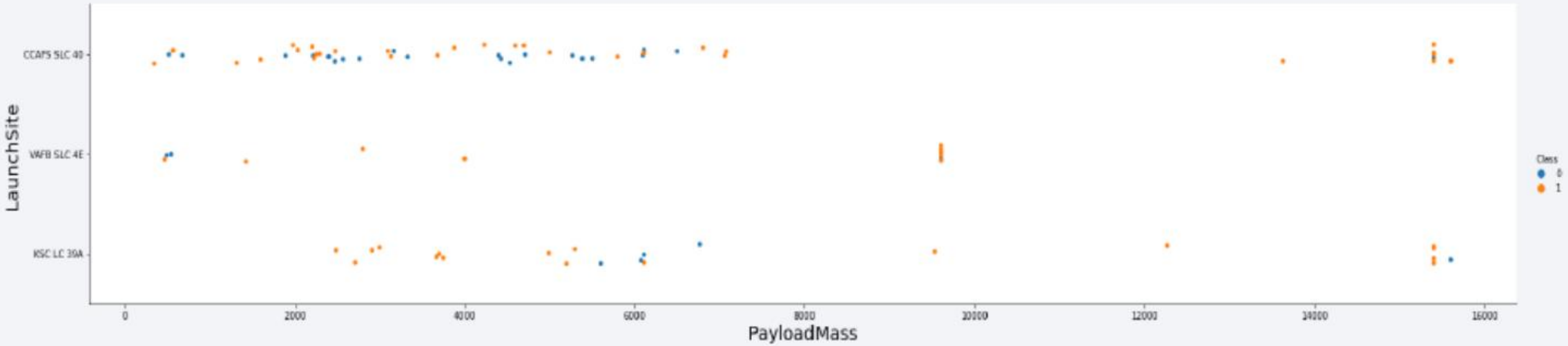
Predictive analysis revealed that the **Decision Tree Classifier** is the most effective model for predicting successful landings. It achieved an **accuracy of over 87%** overall, with even higher accuracy of **over 94%** on the test data


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

The plot indicates that **CCAFS SLC 40** is currently the most successful launch site, with the majority of recent launches achieving success. **VAFB SLC 4E** ranks second, followed by **KSC LC 39A** in third place. Additionally, the data shows a noticeable **improvement in the overall success rate** of launches over time, reflecting SpaceX's increasing reliability and expertise in space missions

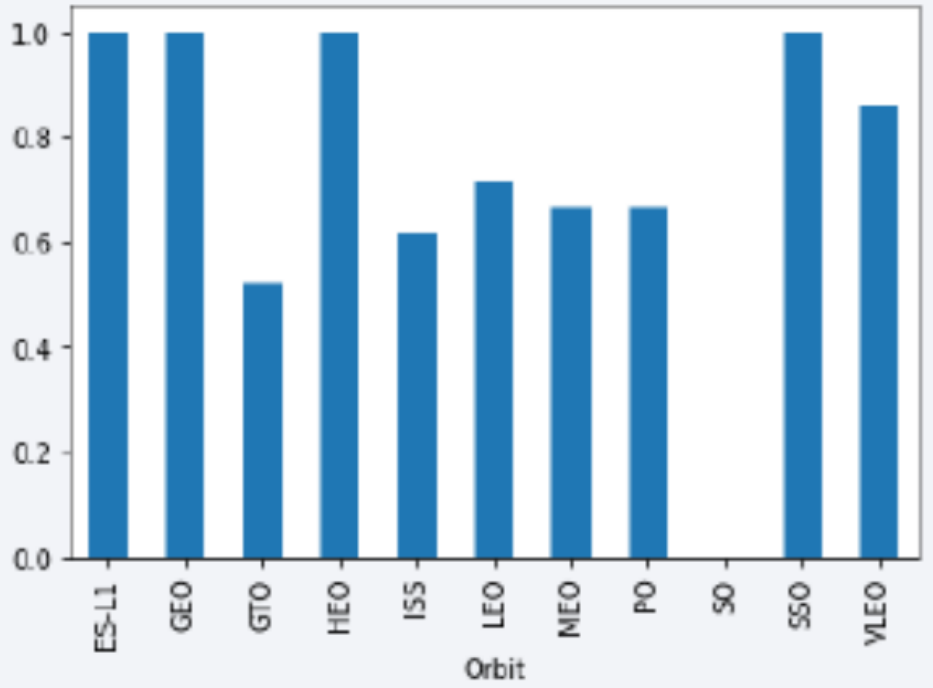
Payload vs. Launch Site



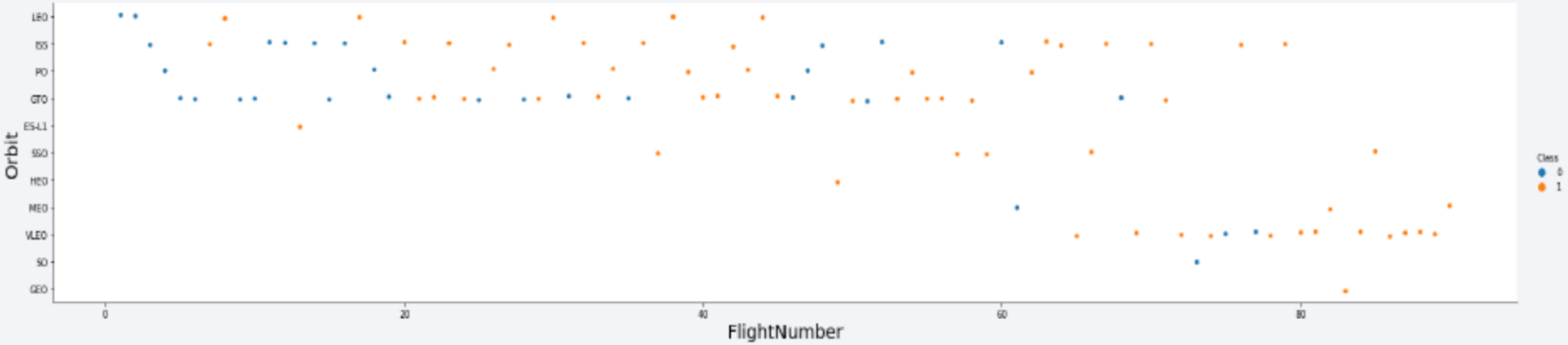
Payloads exceeding **9,000 kg** (approximately the weight of a school bus) show an **exceptionally high success rate** during launches. However, **payloads over 12,000 kg** appear to be feasible only from specific launch sites, namely **CCAFS SLC 40** and **KSC LC 39A**, suggesting that these sites are better equipped to handle such heavy payloads successfully

Success Rate vs. Orbit Type

The highest success rates are observed for missions targeting the **ES-L1**, **GEO**, **HEO**, and **SSO** orbits, indicating strong performance in these mission types. Following closely are the **VLEO** and **LFO** orbits, with success rates above **80%** and **70%**, respectively. These trends suggest that SpaceX has optimized its performance for these specific orbits, leading to a higher probability of successful missions.

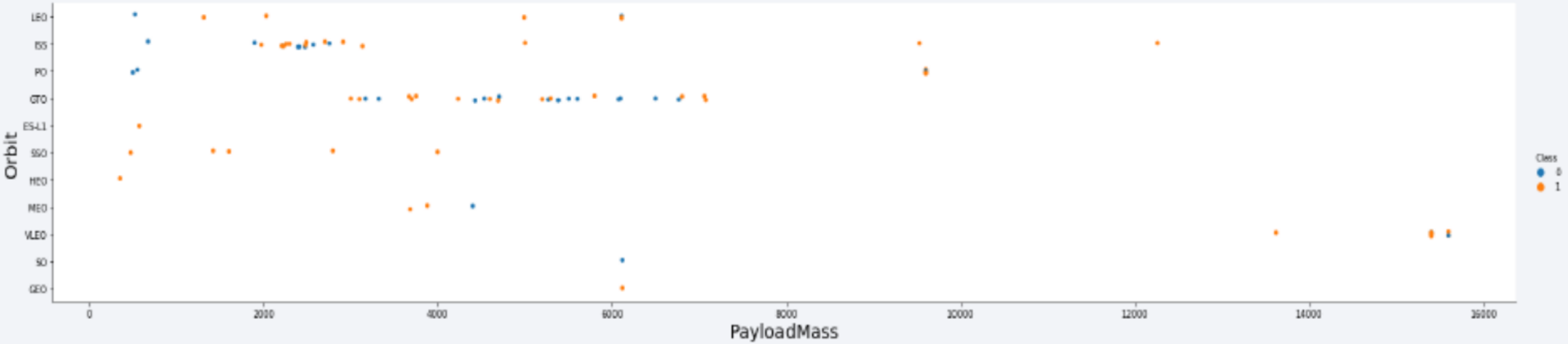


Flight Number vs. Orbit Type



It appears that the **success rate** has consistently improved over time across all orbits, reflecting SpaceX's growing expertise. Of particular interest is the **VLEO orbit**, which has seen a recent **increase in launch frequency**, indicating it may be a **new business opportunity** for SpaceX, as demand for missions in this orbit continues to rise.

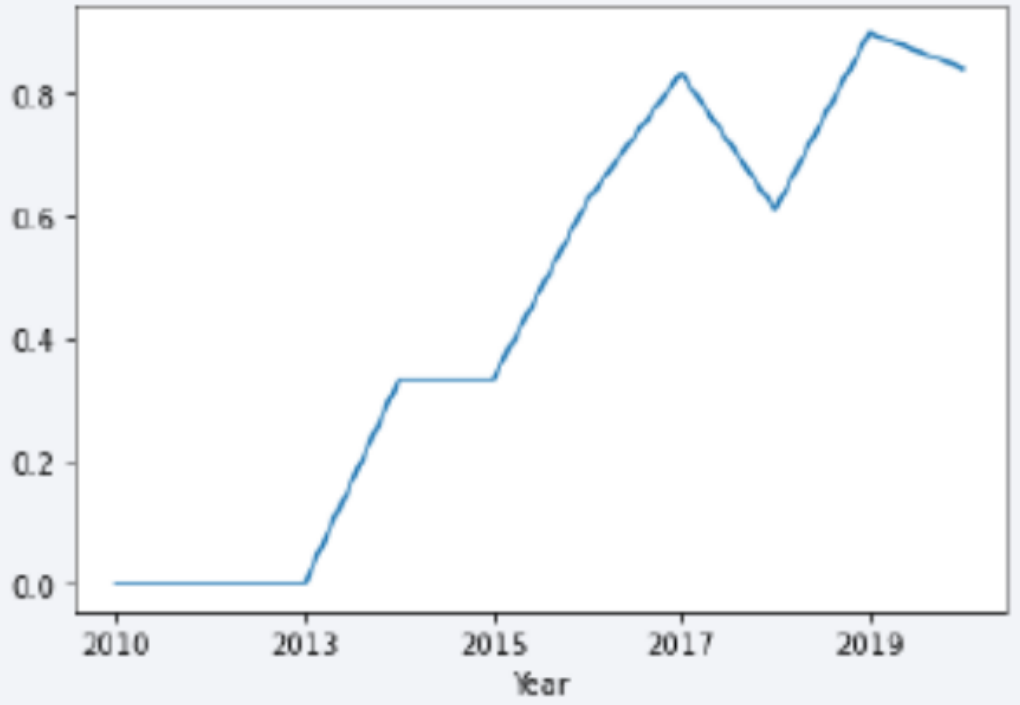
Payload vs. Orbit Type



It appears that for the **GTO (Geostationary Transfer Orbit)**, there is **no significant relationship** between **payload size** and the **success rate**. In contrast, the **ISS orbit** shows a **wide range of payloads** along with a **good success rate**, indicating versatility and reliability for missions to this orbit. Additionally, there have been **fewer launches** to the **SO (Sun-synchronous Orbit)** and **GEO (Geostationary Orbit)**, possibly due to the more specialized nature of these orbits.

Launch Success Yearly Trend

The **success rate** of SpaceX launches **began to improve significantly in 2013** and continued to rise steadily until 2020. The **first three years** of operations appear to have been a **phase of adjustments and technological advancements**, where early challenges were addressed, leading to greater reliability and higher mission success rates over time



All Launch Site Names

The dataset reveals a total of **four unique launch sites**, identified by extracting distinct values from the **"launch_site"** column. This method ensures that each launch site is counted only once, providing a clear overview of SpaceX's operational location

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

The query retrieves **five records** from the launches table where the **launch site name begins with "CCA"**. The **"LIKE 'CCA%'"** condition filters out only those launch sites that start with **"CCA"**, ensuring relevant results

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

Total Payload Mass

This query calculates the **total payload mass** carried by **boosters launched for NASA**.

Total Payload (kg)
111.268

The SUM(payload_mass) function adds up all payload weights from launches where the **customer is NASA**, providing the total mass transported for NASA missions.

Average Payload Mass by F9 v1.1

The query calculates the **average payload mass** carried by the **F9 v1.1 booster version**.

Avg Payload (kg)
2.928

The AVG(payload_mass) function computes the **mean payload weight** for all launches where the **booster version is 'F9 v1.1'**, providing insight into its typical payload capacity.

First Successful Ground Landing Date

This query retrieves the **earliest date** of a **successful landing on a ground pad**.

Min Date
2015-12-22

The MIN(date) function finds the **first recorded success**, while the WHERE landing_outcome = 'Success (ground pad)' condition ensures only **ground pad landings** are considered.

Successful Drone Ship Landing with Payload between 4000 and 6000

This query retrieves the **names of boosters** that **successfully landed on a drone ship** and carried a **payload mass between 4,000 kg and 6,000 kg**.

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

The condition landing_outcome = 'Success (drone ship)' ensures only successful drone ship landings are included, while payload_mass BETWEEN 4000 AND 6000 filters the payload range

Total Number of Successful and Failure Mission Outcomes

This query calculates the **total number of successful and failed mission outcomes** by **grouping** the records based on landing_outcome.

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

The COUNT(*) function counts the number of occurrences for each outcome, providing a breakdown of **successful and failed missions**.

Boosters Carried Maximum Payload

This query retrieves the **names of boosters** that have carried the **maximum payload mass**.

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

The subquery SELECT MAX(payload_mass) FROM launches identifies the **highest payload mass recorded**, and the outer query selects the **booster versions** that match this payload, ensuring only the heaviest-lifting boosters are listed.

2015 Launch Records

This query retrieves the **booster versions, launch site names, and dates** for **failed landings on a drone ship in 2015**.

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

The condition landing_outcome = 'Failure (drone ship)' filters for only **unsuccessful landings**, and YEAR(date) = 2015 ensures results are limited to the **year 2015**.

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

This query **counts** the number of each **landing outcome** (e.g., **Failure (drone ship)**, **Success (ground pad)**) within the date range **from June 4, 2010, to March 20, 2017**.

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

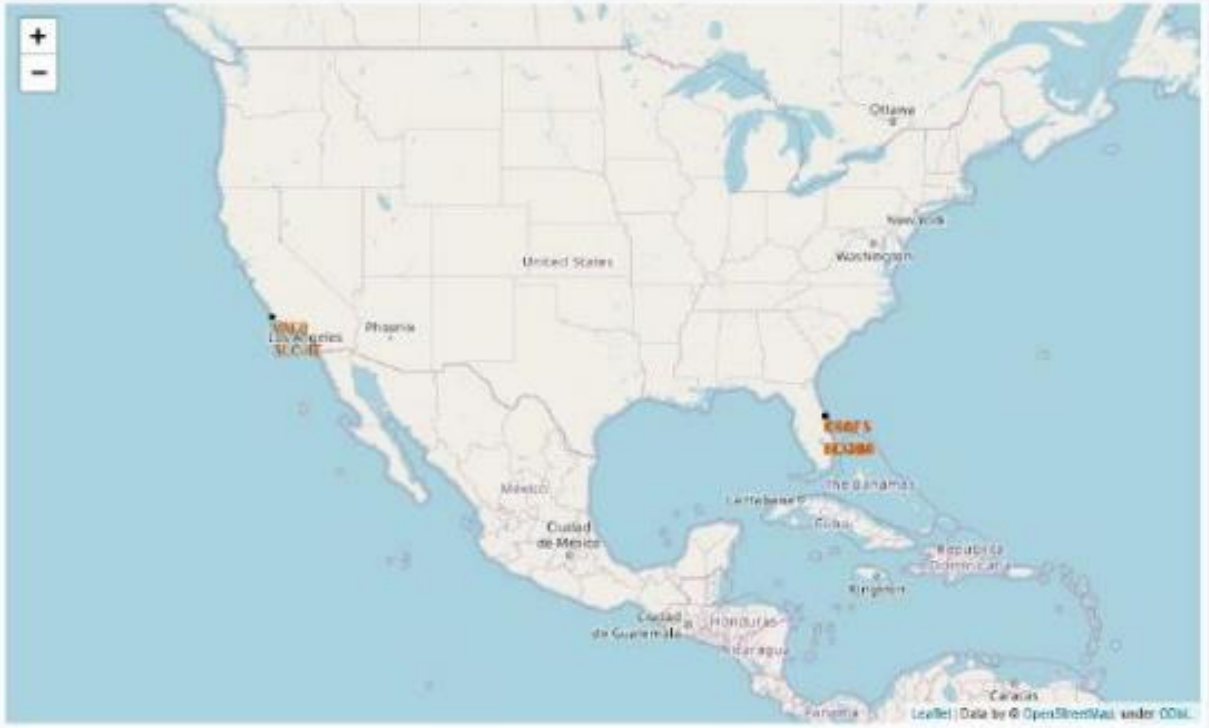
The GROUP BY landing_outcome groups the results by each outcome type, and ORDER BY count DESC ranks the results in **descending order** based on the count, showing the most frequent landing outcomes first.

Section 3

Launch Sites Proximities Analysis

Strategic Location of Launch Sites

Launch sites are typically located **near the sea** for **safety reasons**, as launching rockets over water reduces the potential risks to populated areas in case of a failure during takeoff or landing. However, these sites are **not too far from roads and railroads**, ensuring that they are easily **accessible for transportation** of materials, equipment, and personnel.

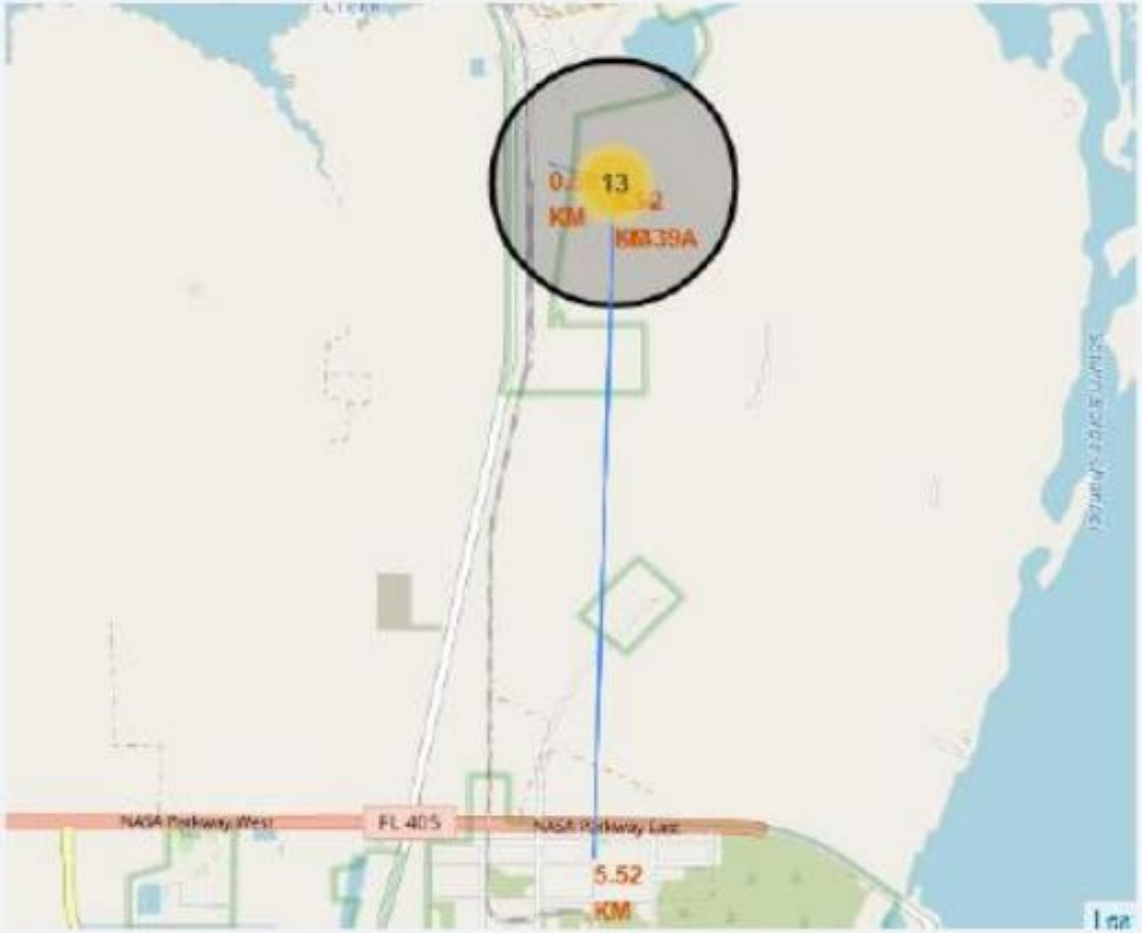


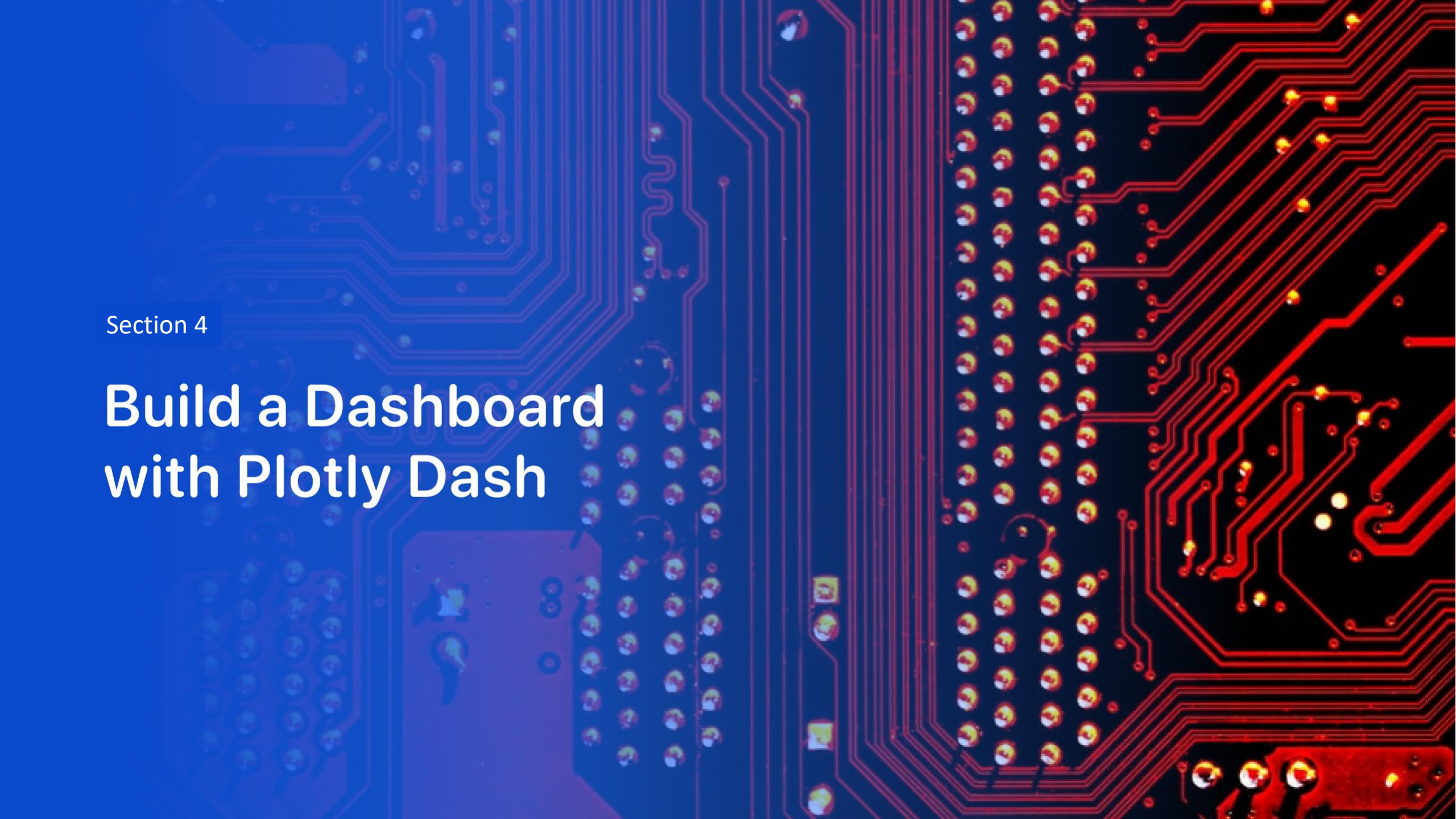
Launch Outcomes of Sites

At the **KSC LC-39A launch site, green markers** represent **successful launches**, while **red markers** indicate **failed launches**.

Logistics and Safety of Launch Site

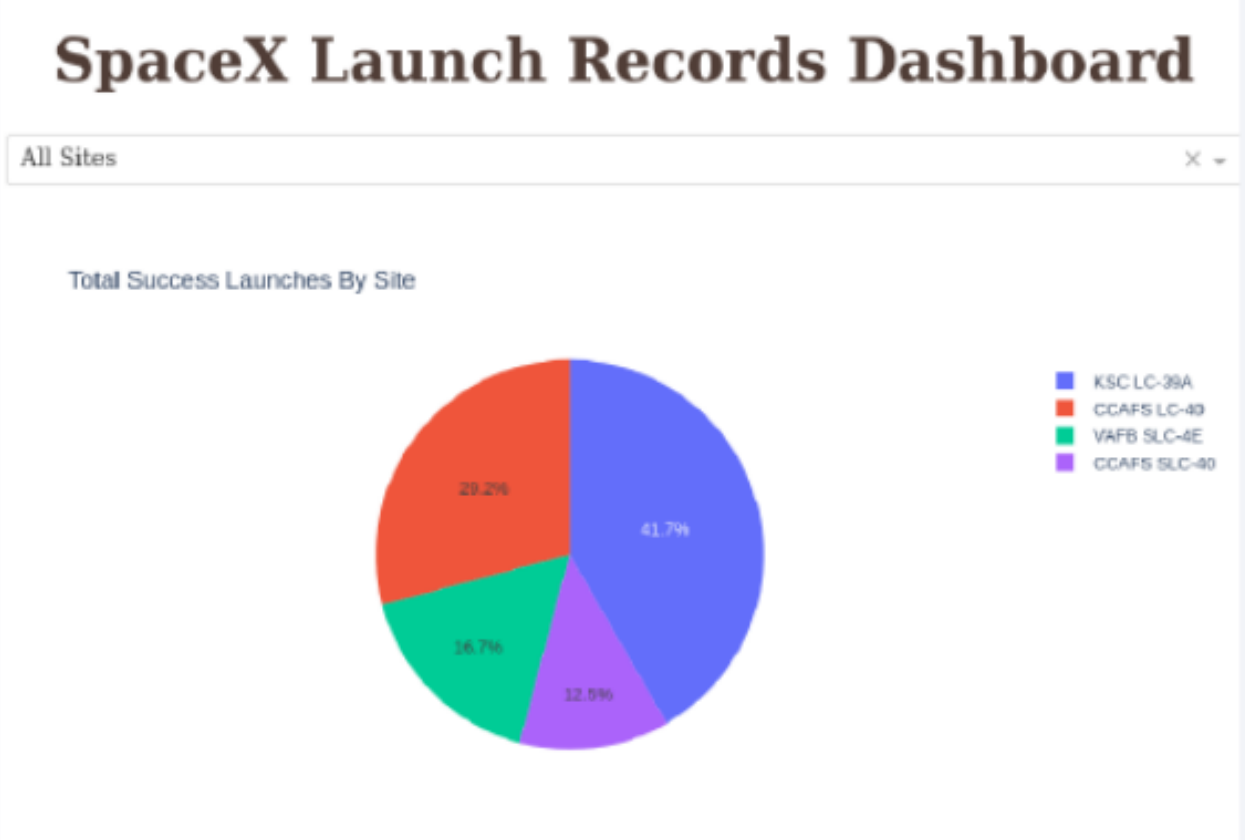
The **KSC LC-39A launch site** is strategically located with **good logistics** in mind, being close to **railroads** and **roads** for easy access and efficient transportation of materials, equipment, and personnel. Additionally, it is situated in an area that is **relatively far from inhabited regions**, prioritizing **safety** by reducing the risk to nearby populations in case of a launch failure




Section 4
Build a Dashboard with Plotly Dash

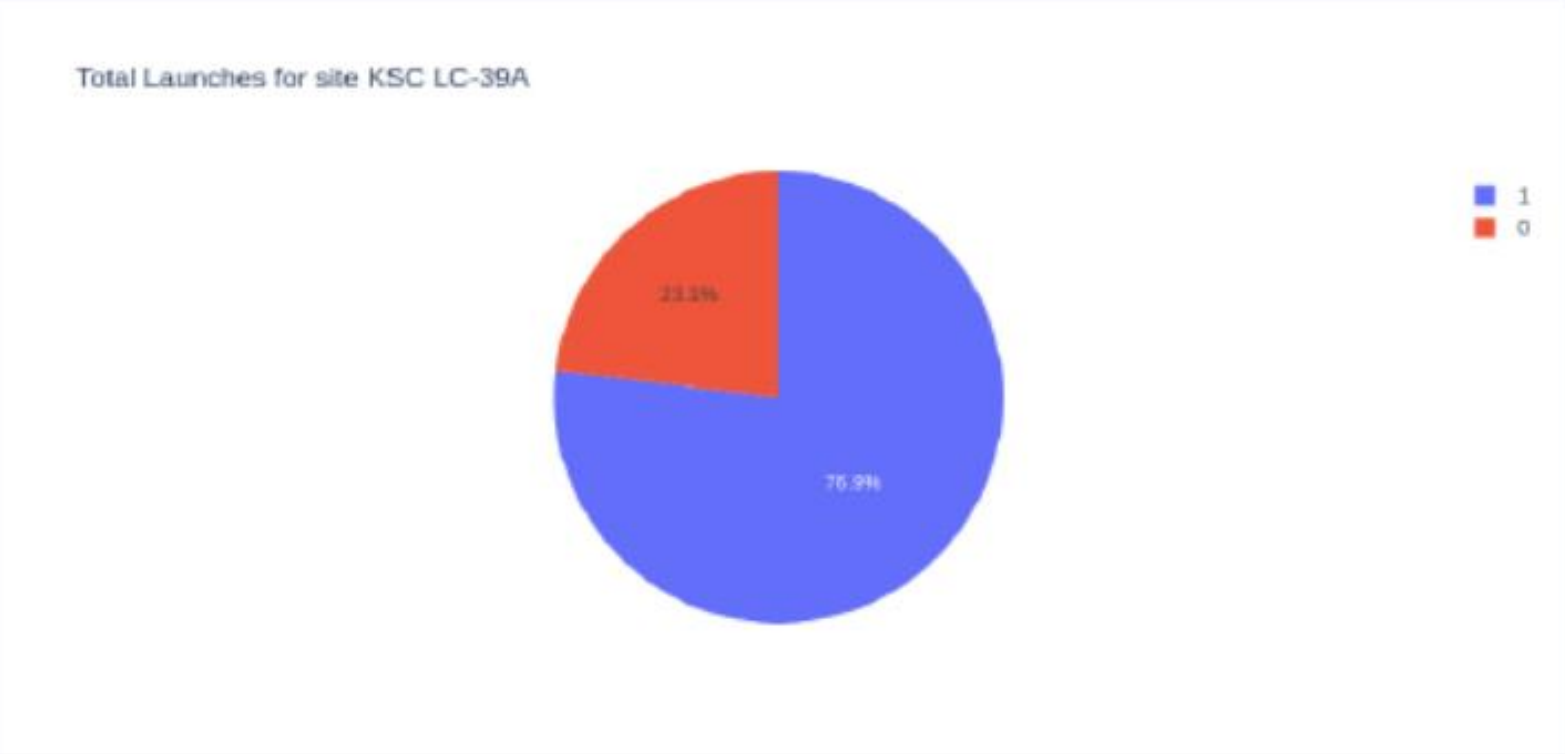
Success Rates by Site

The chart will show the percentage of successful missions for each launch site, offering insight into which sites have performed better in terms of mission outcomes



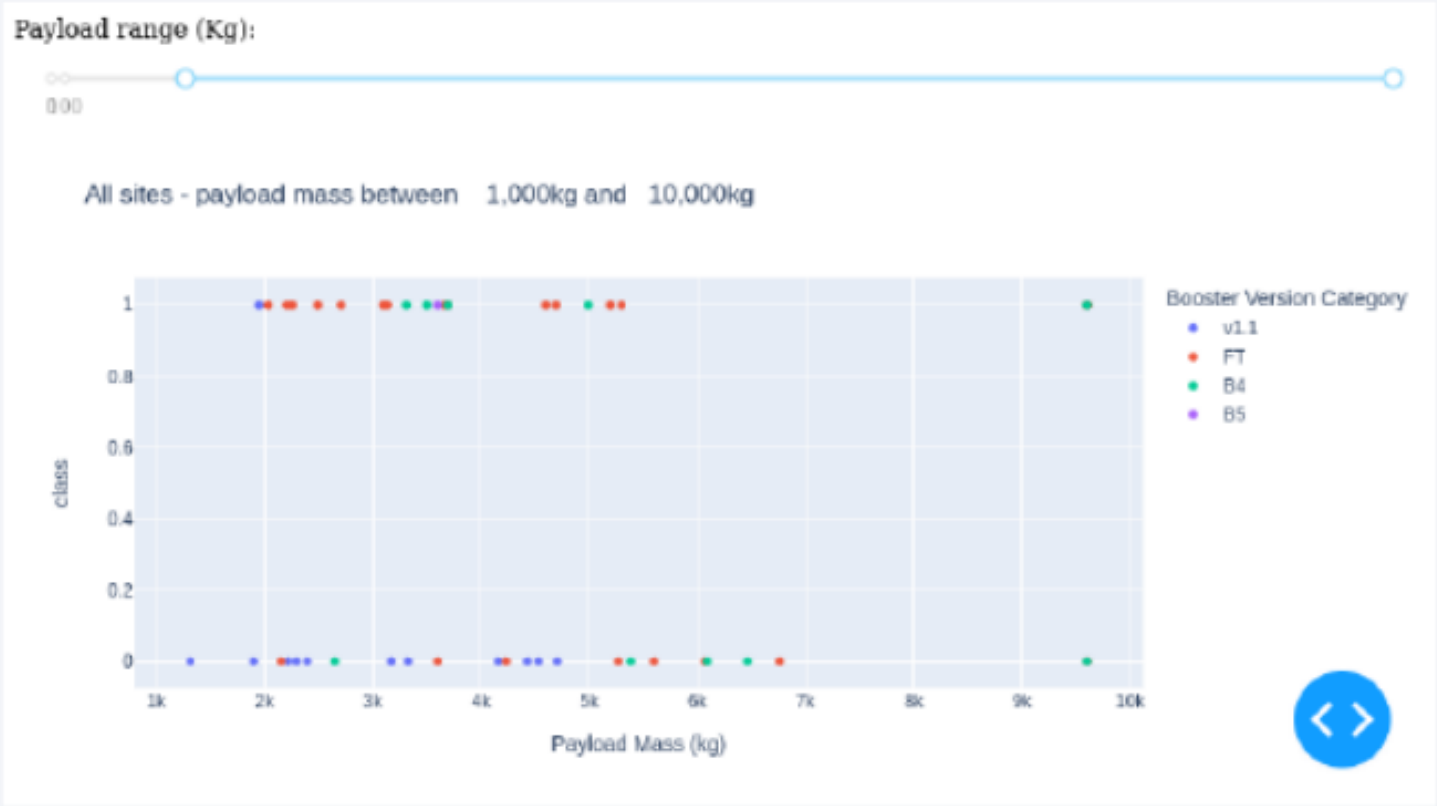
Launch Success Rate at KSC LC – 39A

This pie chart represents the **success rate** of launches at a specific site, showing that **76.9%** of missions were successful, while the remaining **23.1%** resulted in failure.



Analyzing Success Rates Across Different Payload Ranges

The **scatter plot** shows the relationship between **payload mass** and **launch outcomes** across all sites, with a range slider for selecting different payload values. **Payloads under 6,000 kg** and **FT boosters** are the most successful combination, demonstrating the highest success rates.

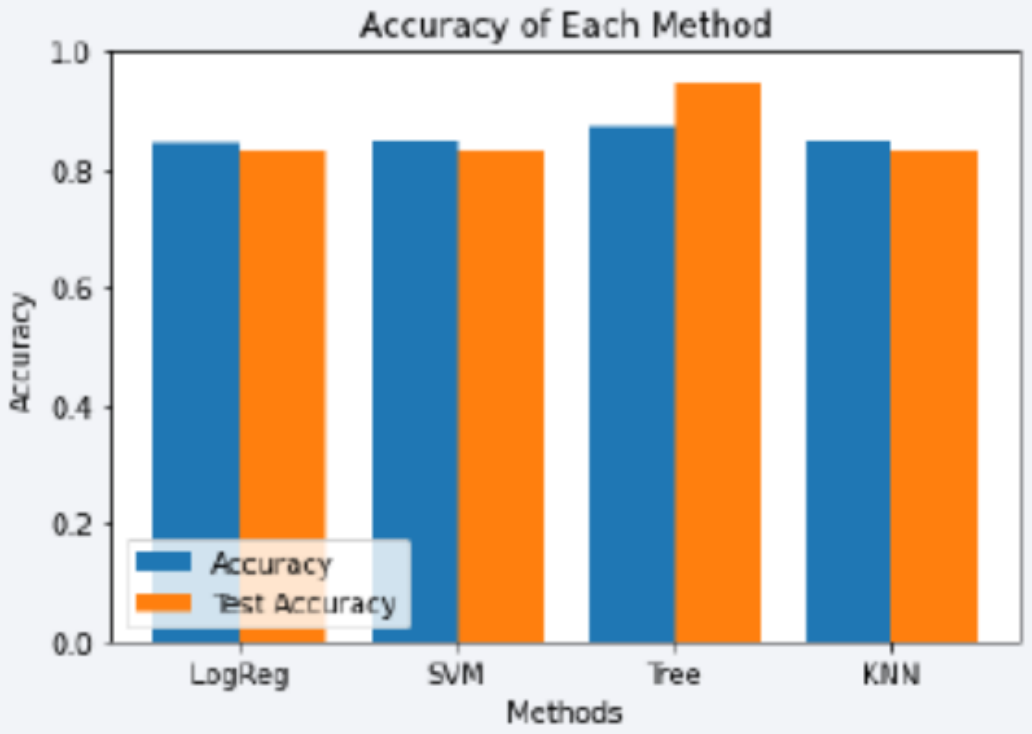


Section 5

Predictive Analysis (Classification)

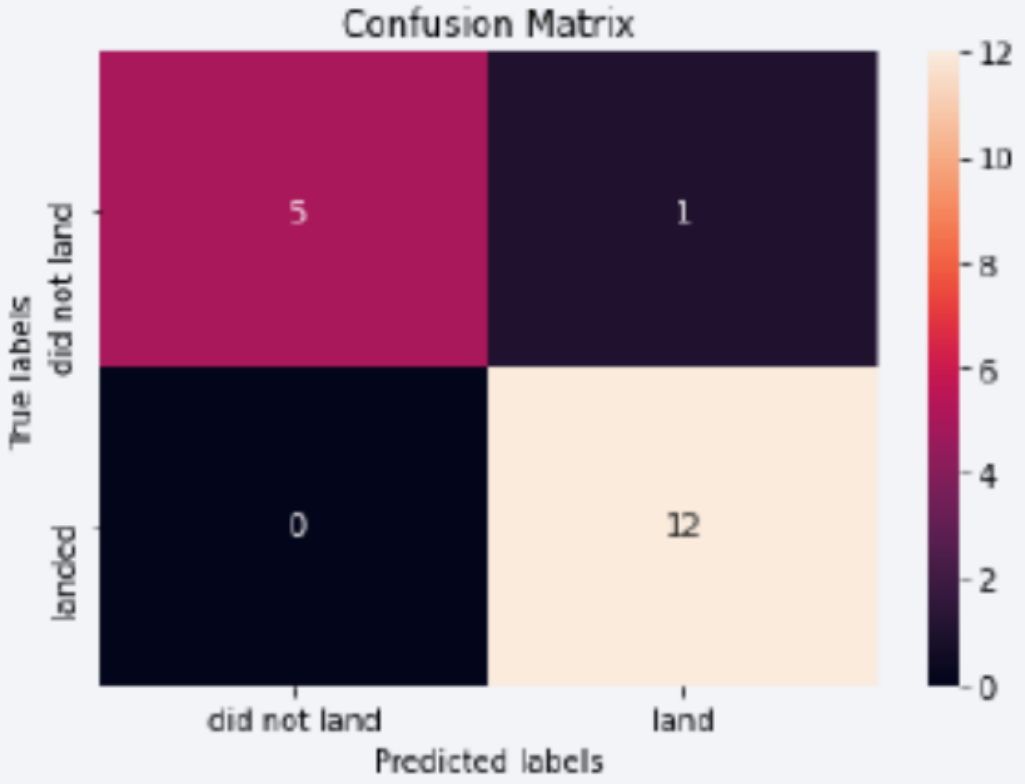
Classification Accuracy

Four different **classification models** were tested to predict mission outcomes, and their **accuracies** were plotted for comparison. The **Decision Tree Classifier** achieved the highest accuracy, with results exceeding **87%**, making it the most effective model for this task compared to the others.



Confusion Matrix

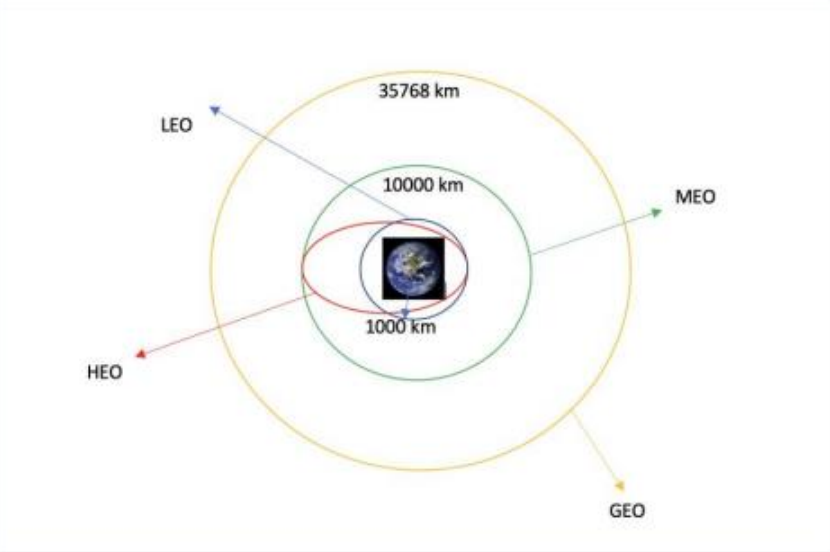
The **confusion matrix** for the **Decision Tree Classifier** demonstrates its effectiveness by showing a high number of **true positives** (correctly predicted successes) and **true negatives** (correctly predicted failures). This indicates that the model is accurately distinguishing between successful and failed missions, with relatively few **false positives** and **false negatives**, confirming its high accuracy



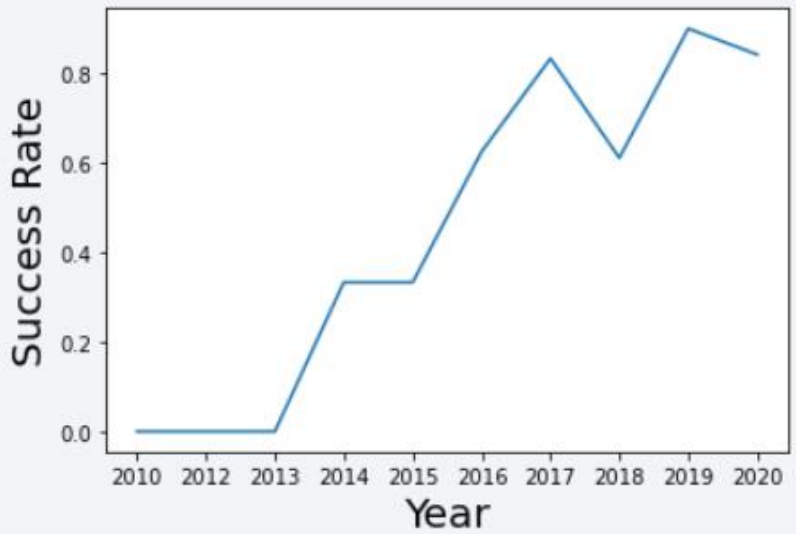
Conclusions

- **Data Sources**: Multiple data sources were analyzed, leading to refined conclusions throughout the process.
- **Best Launch Site**: KSC LC-39A is identified as the best launch site based on mission success.
- **Payload Impact**: Launches with payloads over 7,000kg tend to be less risky and more successful.
- **Improvement Over Time**: While most mission outcomes are successful, successful landing outcomes have increased over time due to improvements in rocket technology and processes.
- **Prediction Model**: The Decision Tree Classifier is an effective model for predicting successful landings, which can increase profits by optimizing launch success rates.

Appendix

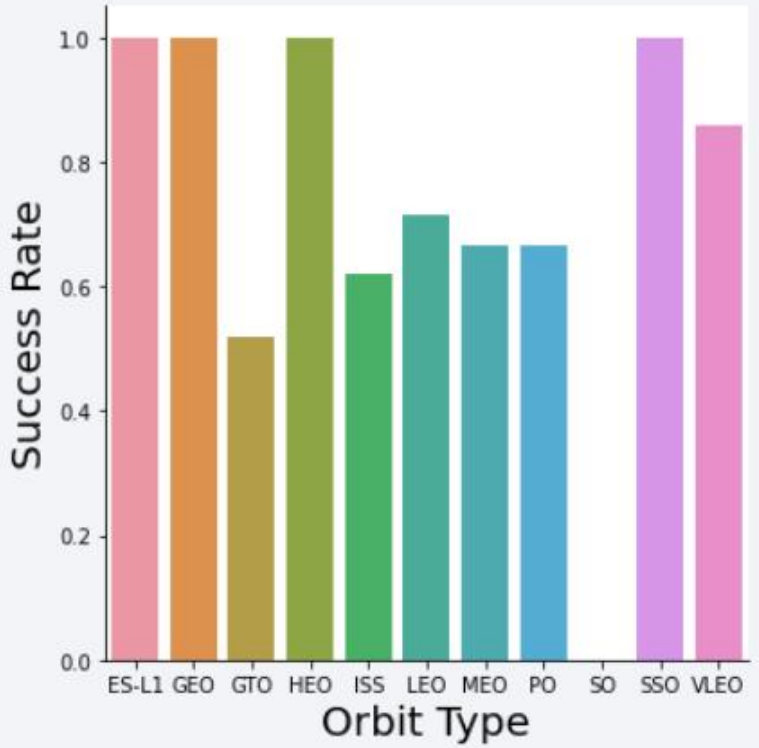


Sites by Orbit Type



Success Rate from 2013 till 2020

Success Rate by Orbit Type



Thank you!