IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nevanda F. P
12/12/2024

Outline

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

Executive Summary

Methodologies and Results :

- Data regarding Falcon 9 first-stage landings from 2010 to 2020 was gathered through a public API (https://api.spacexdata.com/) that is not associated with SpaceX and publicly available resources such as Wikipedia (https://en.wikipedia.org/wiki/SpaceX). Additional data sets were sourced as part of the course materials.
- The data preparation process involved cleaning and wrangling to extract the landing outcome as the dependent variable for use in machine learning models. Insights were derived from the dataset using SQL queries and various data visualization techniques, including static plots, interactive maps, and an interactive dashboard. For predictive analysis, machine learning models such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN) were employed.
- The dataset contained information such as flight number, launch date, payload mass, orbit type, launch site, and mission outcomes. Predictive modeling demonstrated that Logistic Regression, SVM, and KNN provided comparable performance in analyzing the data for predictive purposes.

Introduction

- Background and Objectives

SpaceX, a leader in the aerospace industry, has revolutionized space travel by significantly reducing launch costs through the reuse of the Falcon 9 rocket's first stage. With each launch priced at approximately $62 million—compared to over $165 million for non-reusable rocket providers—SpaceX has achieved groundbreaking milestones, including missions to the International Space Station, launching a satellite internet constellation, and sending astronauts into space.

To compete with SpaceX, a rival rocket launch company aims to predict the success of Falcon 9 first-stage landings. Accurate predictions could influence cost efficiency and launch strategies. Using publicly available data, machine learning models can help determine whether the first stage will successfully land and be reused, directly impacting the pricing of launches.

Key Questions to Explore :

- What insights can be derived from the available data on Falcon 9 first-stage landings?
- How do factors such as payload mass, launch site, flight frequency, and orbit type affect landing success?
- What is the success rate trend for Falcon 9 first-stage landings over time?
- Which machine learning model provides the highest accuracy for predicting landing outcomes, enabling effective binary classification for future launches?
- Will a future Falcon 9 first-stage landing succeed?

Section 1

Methodology

Methodology

- Data on SpaceX Falcon 9 first-stage landings was gathered through multiple sources, including a public SpaceX REST API, web scraping, and additional datasets provided in CSV format. The data preparation process involved filtering, handling missing values, and applying one-hot encoding to ensure the dataset was ready for analysis and modeling.
- Exploratory Data Analysis (EDA) was conducted using SQL and various data visualization techniques to uncover patterns and insights. Interactive visualizations were developed using tools like Folium and Plotly Dash to provide an intuitive understanding of the data.
- To predict landing outcomes, classification models were built, tuned, and evaluated. These models were optimized to identify the most accurate predictions and parameters, leveraging machine learning to analyze the success factors of first-stage landings.

Data Collection

- The datasets were sourced from multiple channels, including a response from an IBM-hosted replica of a publicly accessible SpaceX API providing launch data in JSON format. Additional data was retrieved from a revision of a Wikipedia page containing launch data in HTML tables (as of 9 June 2021), along with supplemental CSV files provided during the course.

Data Collection – SpaceX API

- To prepare the data for analysis, information was requested from the SpaceX API using custom functions, with the responses decoded into JSON and converted into dataframes using .json_normalize(). The data was organized into dictionaries before being further processed into dataframes. Specific filters were applied to isolate Falcon 9 launches, and missing payload mass values were replaced with the calculated mean. The cleaned and filtered data was then exported to a CSV file for further analysis.
- Github : https://github.com/nevandaa/Data-Science-Capstone/blob/main/1-SpaceX-Data-Collection.ipynb

Send GET request to API

Extract nested data

Convert date format

Use defined functions to generate specific columns of data

Combine separate columns into a DataFrame

Filter launches with rockets other than the Falcon 9

Handle missing values

Data Collection - Scraping

- SpaceX Data is scraped from HTML Tables on the Wikipedia Page of SpaceX through a permanently-linked version of the page
- The scraped data is loaded to a dataframe to be processed
- Github : https://github.com/nevandaa/Data-Science-Capstone/blob/main/2-SpaceX-Web-Scraping.ipynb

Web Scrape the page to get the entire HTML text

Create a BeautifulSoup object from the response text content

Select the tables

From the launch table, extract the column names from the tags

Create a Pandas Dataframe

Data Wrangling

- The data wrangling process involved cleaning and preparing the dataset for analysis. Key steps included determining labels and calculating the number of launches at each site, the frequency of different orbit types, and mission outcomes per orbit type. A binary classification column was added to the dataset to represent the landing outcomes, where successful landings were labeled as 1 and unsuccessful ones as 0.
- Landing outcomes were categorized based on their success or failure on different platforms: successful landings on ground pads or drone ships were labeled as "True RTLS" and "True ASDS," respectively, while failures in these categories were labeled as "False RTLS" and "False ASDS." The processed data was consolidated into a comprehensive DataFrame and exported to a CSV file for further use.
- Github : https://github.com/nevandaa/Data-Science-Capstone/blob/main/3-SpaceX-Data-Wrangling.ipynb

EDA with Data Visualization

- The Exploratory Data Analysis (EDA) revealed key trends and patterns in the SpaceX Falcon 9 launch dataset through various visualizations:
  - **Launch Site Success Trends**: Scatter plots showed the relationship between flight numbers and payloads across different launch sites. They highlighted that newer launches tend to have higher success rates, and KSC LC-39A emerged as the site with the highest success ratio.
  - **Orbit Success Trends**: Certain orbits, including ES-L1, GEO, HEO, and SSO, demonstrated a 100% success rate, while others like SO had no successful landings.
  - **Payload and Launch Success**: Scatter plots indicated that heavier payloads (especially between 2,000 kg and 5,000 kg) correlated with higher success rates. Success rates varied across orbit types, with lighter payloads showing consistent success in specific orbits such as LEO and ISS.
  - **Time-Based Success Trends**: Line plots revealed an overall increase in landing success rates over time, with noticeable improvements from 2013 onwards.
- These visualizations provided valuable insights into how factors such as launch site, payload, orbit type, and time influenced the outcomes of Falcon 9 launches.
- Github : https://github.com/nevandaa/Data-Science-Capstone/blob/main/5-SpaceX-Data-Visualization.ipynb

EDA with SQL

- The SQL-based Exploratory Data Analysis (EDA) provided detailed insights into the SpaceX Falcon 9 dataset:
  - **Launch Sites**: Identified the unique launch sites, with a query showing records of sites starting with "CCA."
  - **Payload Analysis**: Calculated the total payload mass carried by NASA (45,596 kg) and the average payload mass for the F9 v1.1 booster (2,928 kg).
  - **Mission Outcomes**: Queried the total count of successful (61) and failed (40) mission outcomes, along with records of failed drone ship landings in 2015.
  - **Landing Success Trends**: Found the date of the first successful ground landing (December 22, 2015) and identified boosters that successfully landed on drone ships with payloads between 4,000 kg and 6,000 kg.
  - **Performance Across Time**: Ranked the frequency of landing outcomes between 2010 and 2017 in descending order, highlighting "No attempt" as the most common outcome.
- This SQL analysis enriched the understanding of launch sites, payload characteristics, and success trends, offering precise numerical insights to complement other analysis methods.
- Github Link : https://github.com/nevandaa/Data-Science-Capstone/blob/main/4-SpaceX-SQLite.ipynb

Build an Interactive Map with Folium

- The interactive map created with **Folium** provided a detailed geographic view of the SpaceX launch sites and their respective success rates:
  - **Launch Site Markers**: The map displayed markers for each launch site. Green markers represented successful landings, while red markers indicated failures.
  - **Launch Site Success**: The map allowed users to visually compare success rates at different sites. Launch sites with higher success rates showed a higher density of green markers, indicating a greater number of successful Falcon 9 first-stage landings.
  - **Distance Analysis**: Colored lines were drawn to show the proximity of launch sites to critical geographical features, such as the coastline, rail lines, and highways. For example, CCAFS LC-40 was located 0.92 km from the coastline, 1.33 km from the rail line, and 0.19 km from the perimeter road.
  - **Geographical Insights**: The map illustrated how launch sites are strategically located near the equator for orbital advantages and are close to coastlines for safety, minimizing the risk of damage from failed launches.
- This interactive map enhanced the visual exploration of launch site data, providing geographical insights into SpaceX's operations.
- Github Link : https://github.com/nevandaa/Data-Science-Capstone/blob/main/6-SpaceX-Interactive-Visual-Analysis.ipynb

Build a Dashboard with Plotly Dash

- The **Plotly Dash** dashboard allowed for interactive exploration of the SpaceX Falcon 9 landing data, featuring key visual elements:
  - **Launch Site Selection**: A dropdown menu enabled users to choose between one or all launch sites, displaying corresponding data on the dashboard.
  - **Pie Chart**: The pie chart displayed the distribution of successful and unsuccessful Falcon 9 first-stage landings at a selected launch site. When multiple sites were selected, it showed the overall success rate across all launch sites.
  - **Payload Mass Slider**: A slider allowed users to filter payload masses, visualizing the relationship between payload mass and landing success.
  - **Scatter Plot**: The scatter plot provided a visual breakdown of Falcon 9 first-stage landings, showing the distribution of landings by payload mass, mission outcome, and booster version category.
  - **Insights**: The dashboard enabled users to easily explore how factors such as launch site, payload mass, and booster version influenced the success of Falcon 9 first-stage landings.
- This interactive dashboard offered an intuitive interface to explore the dataset and gain insights from the visualized data.

Predictive Analysis (Classification)

- The **Predictive Analysis** using classification models evaluated the success of Falcon 9 first-stage landings, and the following results were observed:
  - **Model Performance**: All classification models—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN)—performed similarly on the dataset, with the Decision Tree model showing slightly better performance than the others.
  - **Evaluation Metrics**: Models were assessed using accuracy scores, precision, recall, and F1 scores. The Decision Tree model slightly outperformed others in terms of accuracy.
  - **Confusion Matrix**: For the Logistic Regression model, the confusion matrix showed 12 true positives, 3 true negatives, 3 false positives, and 0 false negatives. This resulted in an accuracy of 83.3%, with a precision of 80% and a recall of 100%.
  - **Model Comparison**: Despite minor differences in performance, all models provided similar results, indicating that the dataset was well-suited for predicting landing outcomes. The Decision Tree model's slight edge in accuracy was noteworthy, but no model drastically outperformed the others.
- This predictive analysis demonstrated the utility of machine learning models for forecasting the success of SpaceX Falcon 9 first-stage landings.
- Github : https://github.com/nevandaa/Data-Science-Capstone/blob/main/7-SpaceX-Machine%20Learning%20Prediction.ipynb

Results

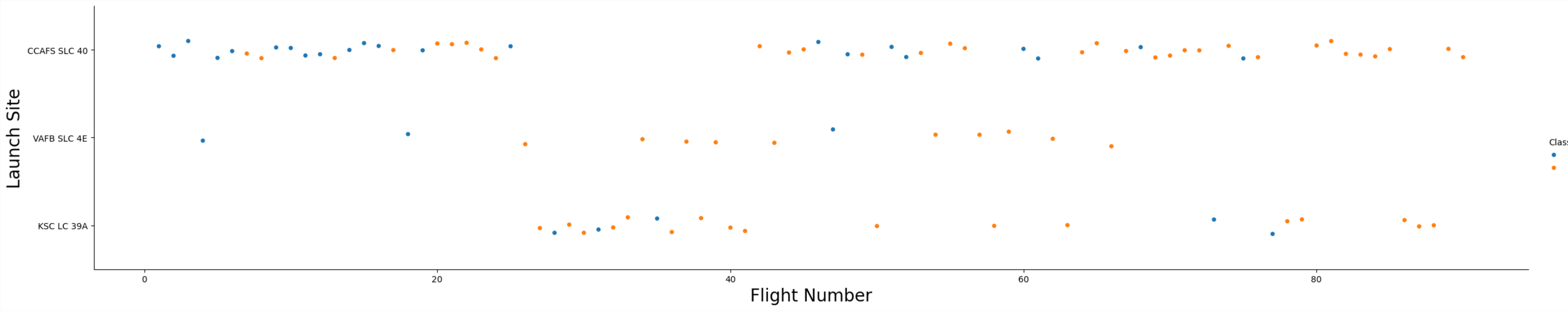
- The analysis yielded several key findings from the SpaceX Falcon 9 first-stage landing dataset:
  - **Launch Success Over Time**: The success rate of Falcon 9 first-stage landings has significantly improved over time, particularly as more flights were completed.
  - **Top Launch Site**: KSC LC-39A had the highest success rate among the launch sites, with Falcon 9 landings showing increasing success as flight numbers grew.
  - **Orbit Type Success**: Orbits such as ES-L1, GEO, HEO, and SSO achieved a 100% success rate for Falcon 9 first-stage landings, while others showed varying levels of success.
  - **Payload and Success**: The success rate of landings correlated with payload mass, with heavier payloads generally leading to higher success rates, especially for specific orbit types like LEO and ISS.
  - **Predictive Model Findings**: Machine learning models used for prediction showed that Logistic Regression, SVM, and KNN performed similarly, with Decision Trees slightly outperforming the others in accuracy.
- Overall, the analysis demonstrated a trend toward higher success rates for Falcon 9 first-stage landings and highlighted the key factors influencing these outcomes, including launch site, payload mass, and orbit type.

Section 2

Insights drawn from EDA

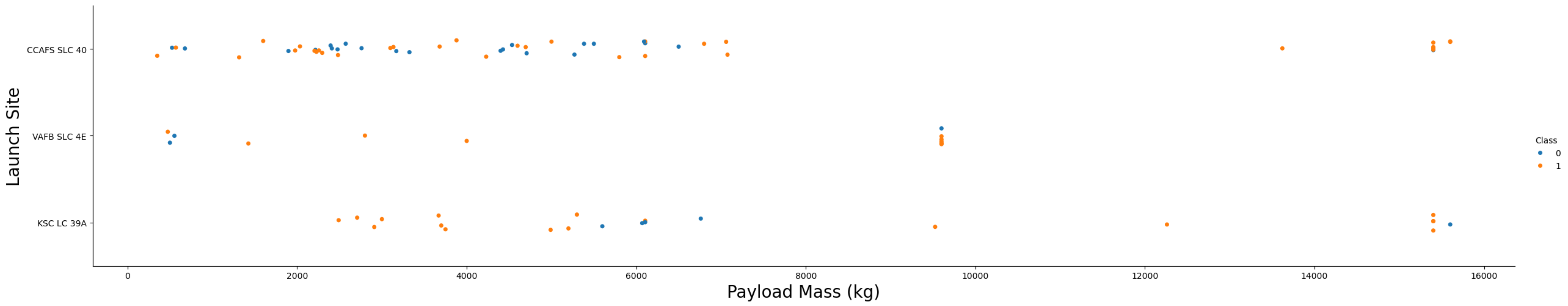
Flight Number vs. Launch Site

- Success rate is more seen at later flights
- Half of the launches is from CCAFS SLC 40 Launch Site



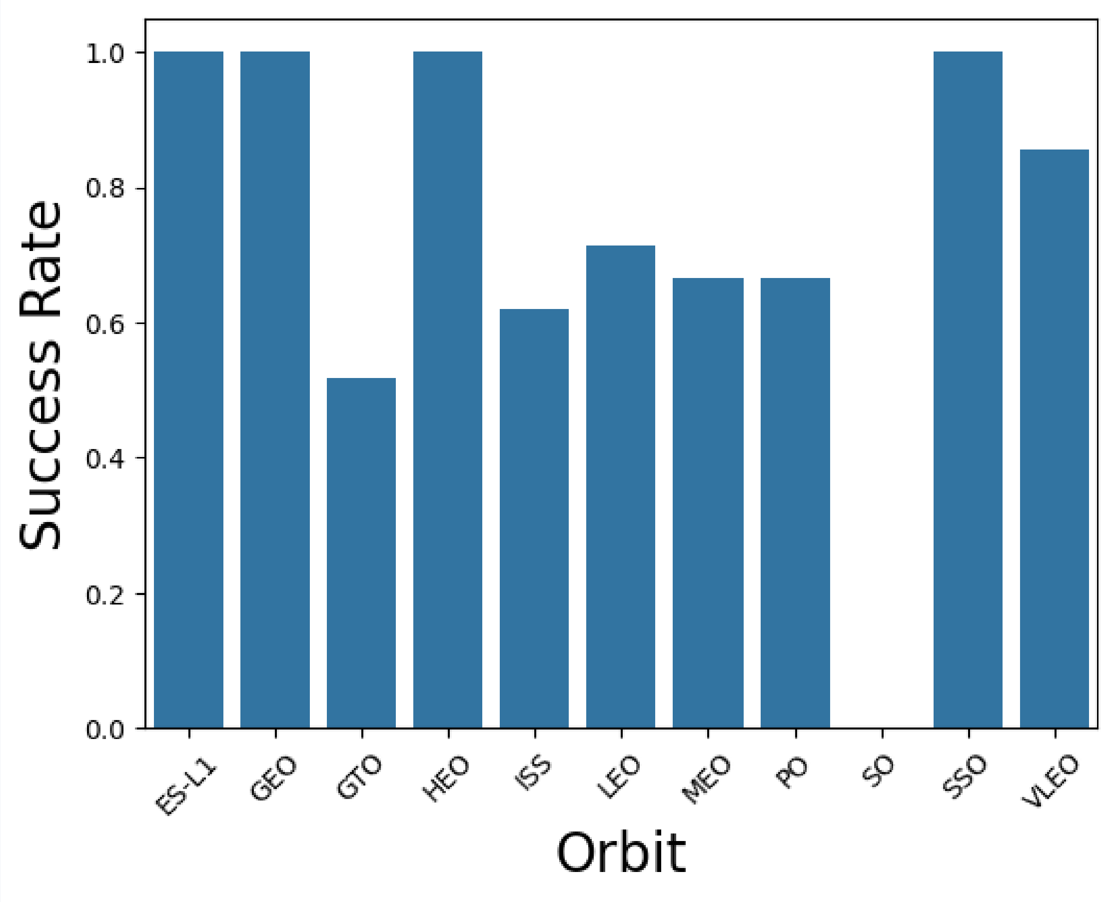
Payload vs. Launch Site

- The higher the payloas mass, the higher the success rate
- Most Launch with a payload mass higher than 7000 kg is successful



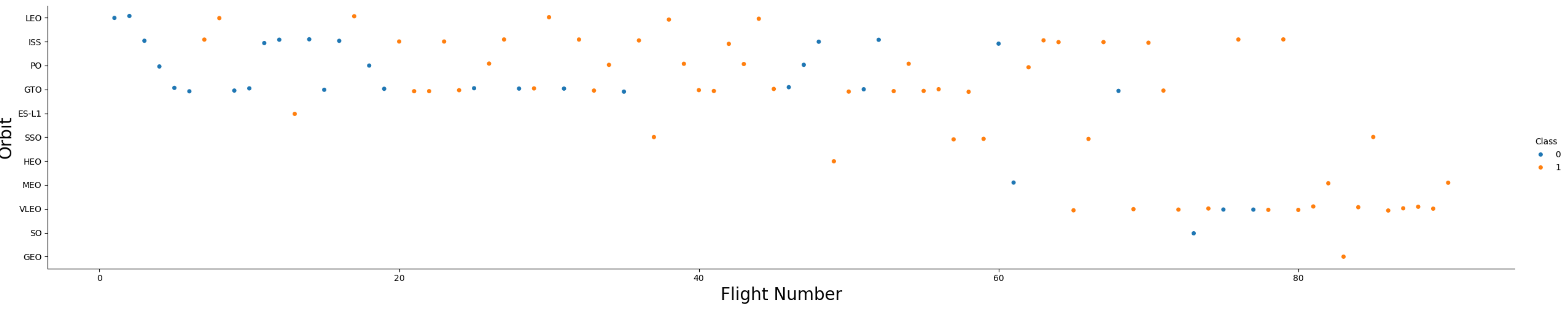
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a 100% Success Rate
- GTO, ISS, LEO, MEO, PO have a 50 - 80% Success Rate
- SO have a 0% Success Rate



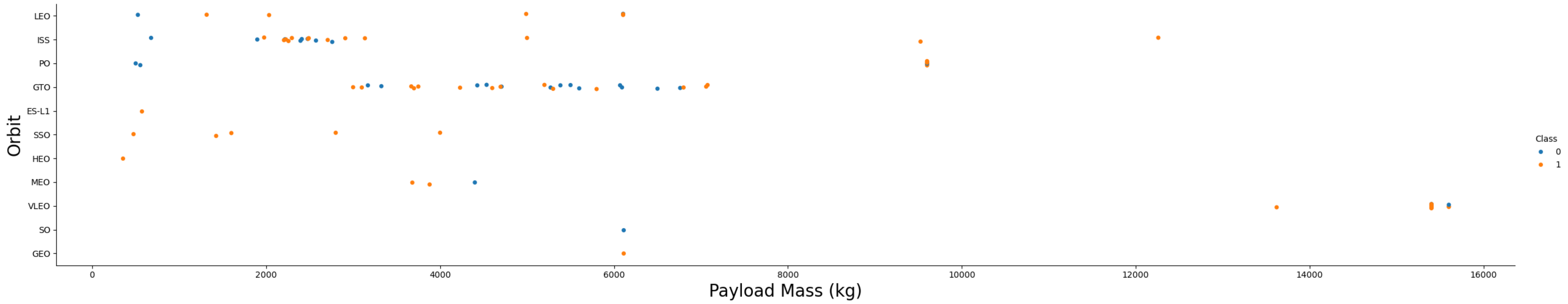
Flight Number vs. Orbit Type

- Larger Flight Numbers is associated with higher success rates



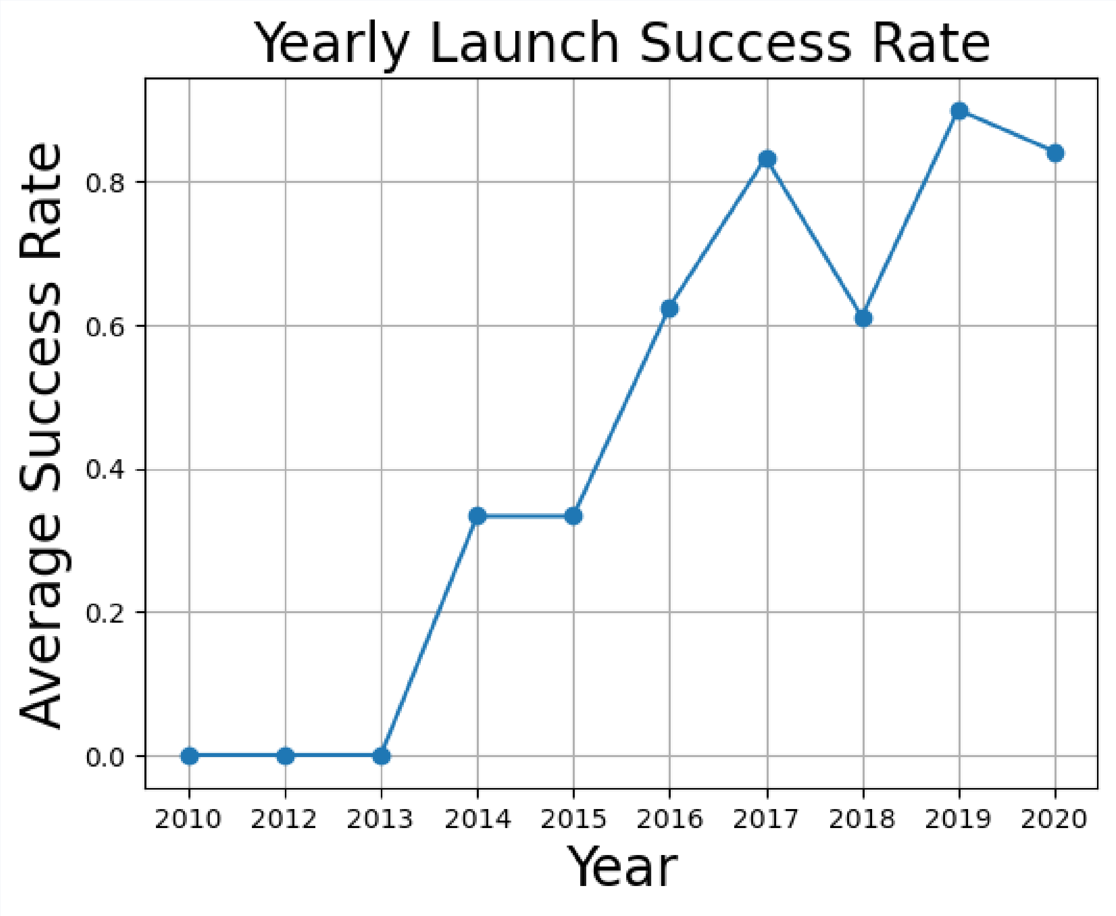
Payload vs. Orbit Type

- Some orbit types have better results with heavy payloads
- The GTO Orbit have mixed success rate with heavier payloads



Launch Success Yearly Trend

- The Success rate has increased significantly over the years, with a slight decline in 2018



All Launch Site Names

- There are 4 unique launch sites
- Query : %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

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

Launch Site Names Begin with 'CCA'

- This query is a to show 5 records of launch site that starts with CCA, all of the shown data is from CCAFS LC-40
- Query : %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA by using the SUM Method. Result : 48213 KG
- Query : %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 using the AVG method
- Query : %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad, using the minimum function of a success outcome
- Query : %sql SELECT MIN(DATE) \

FROM SPACEXTBL \

WHERE LANDING_OUTCOME = 'Success (ground pad)'

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The booster version of the successful attempt with 4000 – 6000 payload mass is shown below
- Query : %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes. There is 61 Successful mission outcomes and the remaining is failed mission
- Query : %sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE GROUP BY Landing_Outcome;

Landing_Outcome	Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query : %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query : %sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)';

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

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

- The most common landing outcome is no attempt.
- Query : %sql 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;

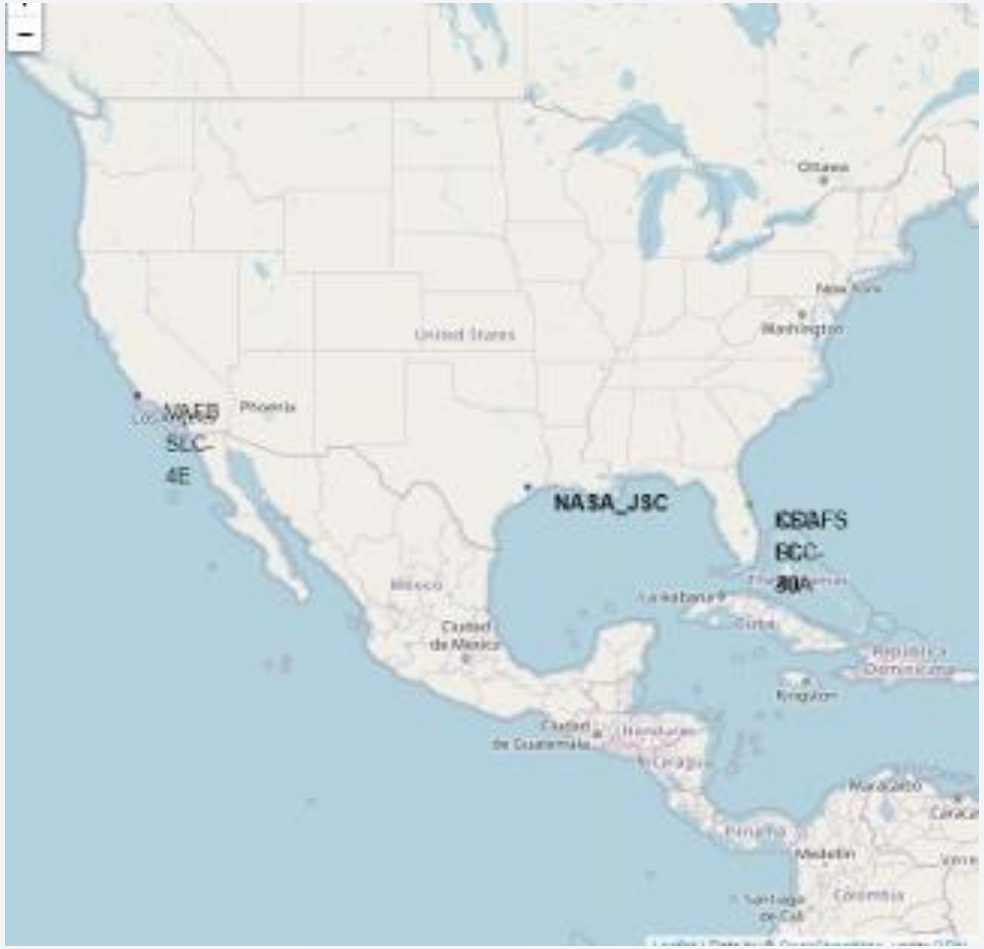
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

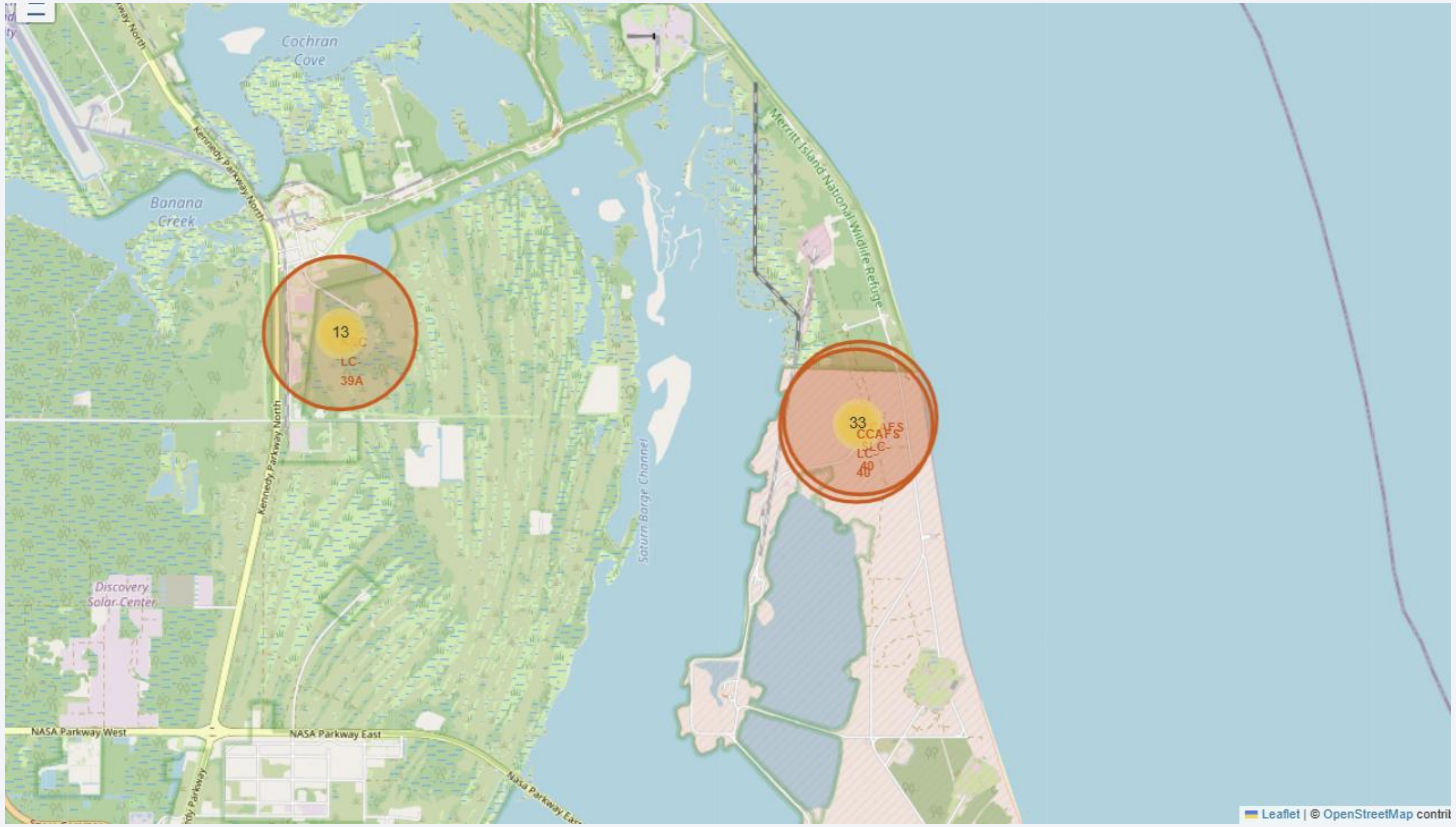
Launch Sites Proximities Analysis

Launch Site Locations

- VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A | Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 | Cape Canaveral Air Force Station Space Launch Complex 40



Launch Outcomes



Distance to Proximities

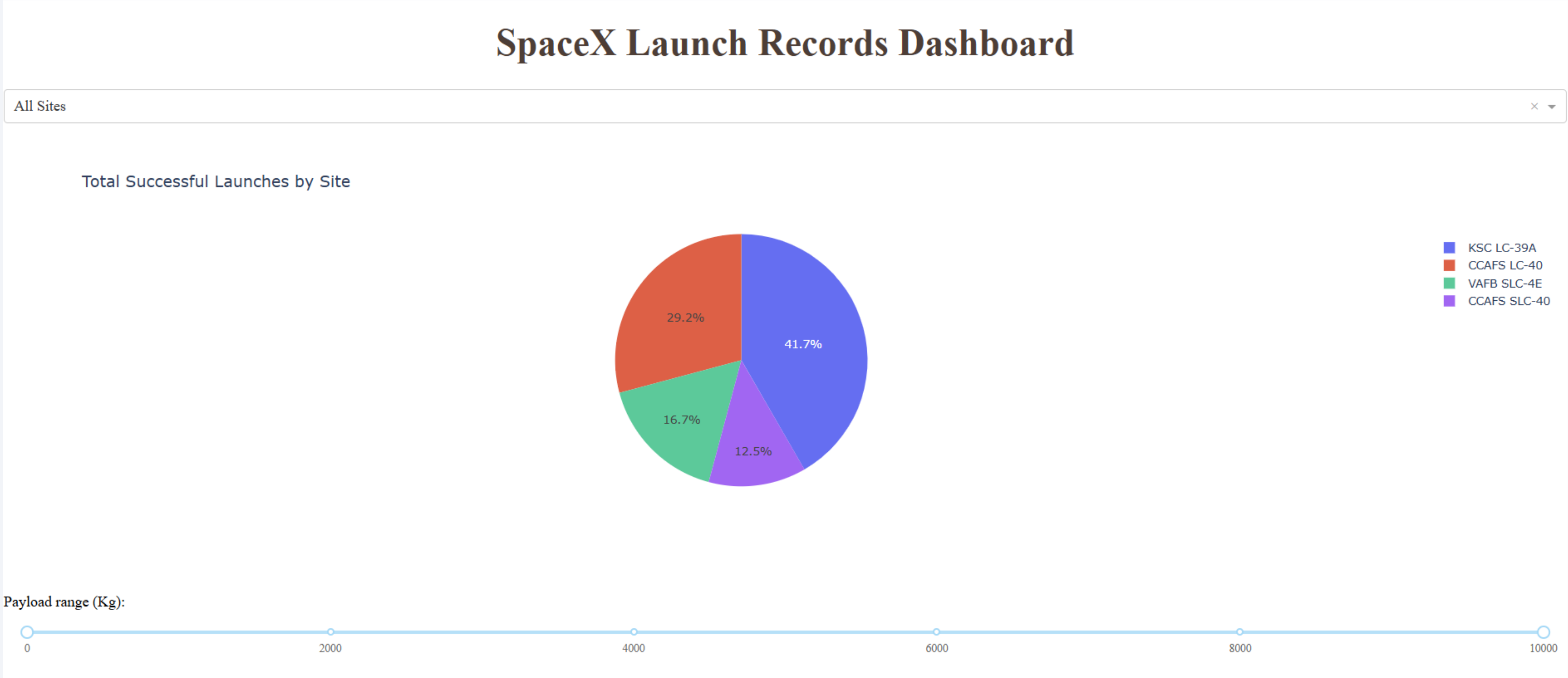
- CCAFS SLC-40
- 1 km from nearest coastline, 21 km from nearest railway, 23.23 km from nearest city, 26.88 km from nearest highway

Section 4

Build a Dashboard with Plotly Dash

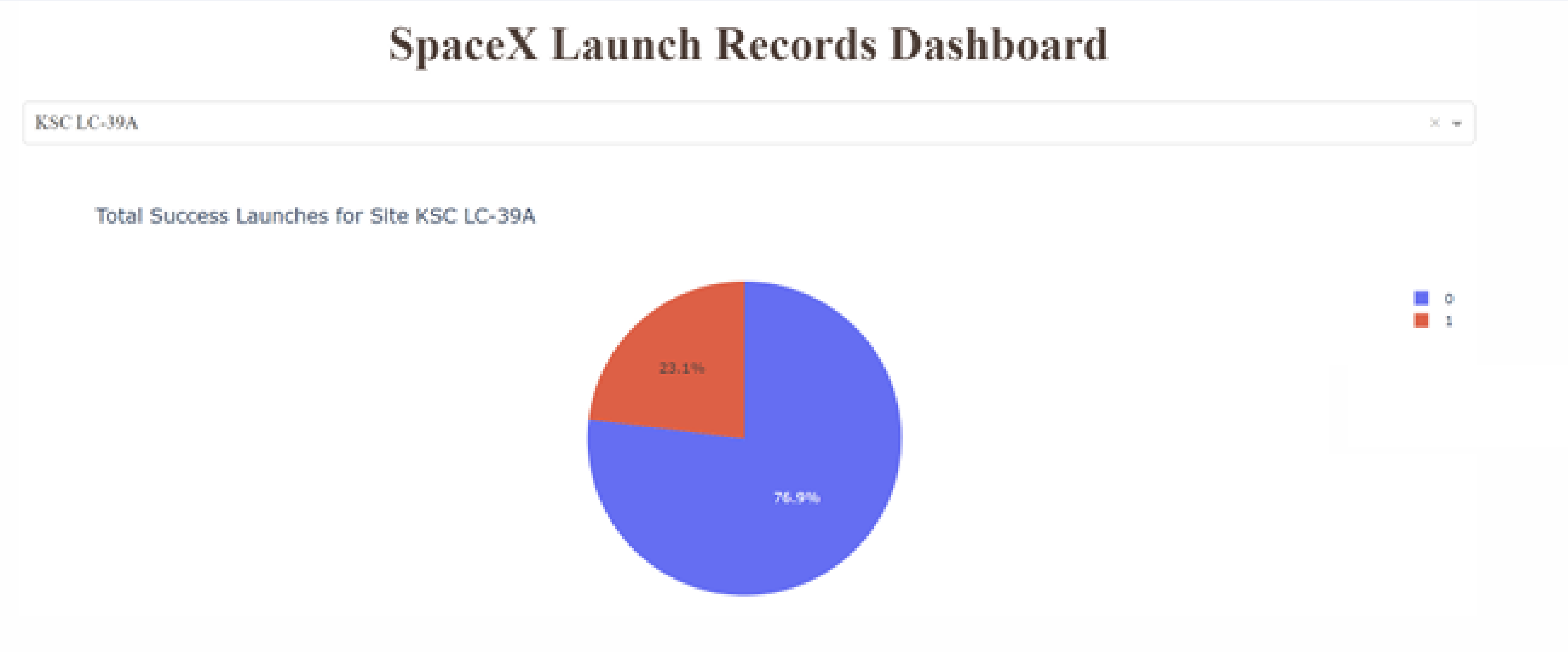
Launch Success All Sites

- The highest success rate is from launch site KSC LC-39A



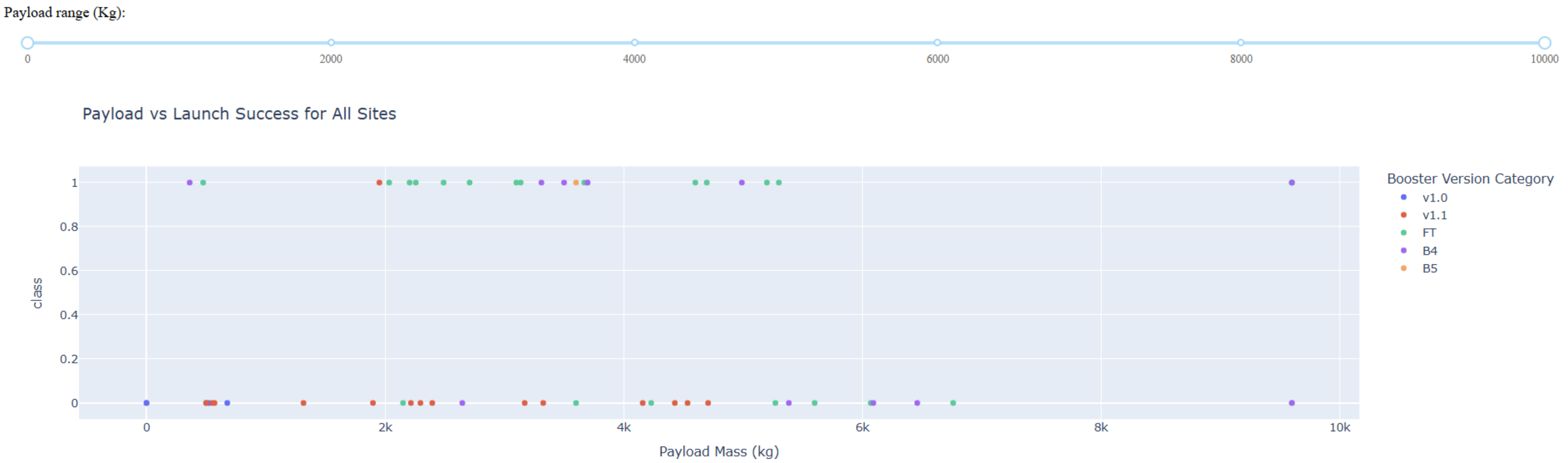
Launch Success in Highest Site

- KSC LC-39A has 10 successful launches and 3 failed launches



Payloads Mass and Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate

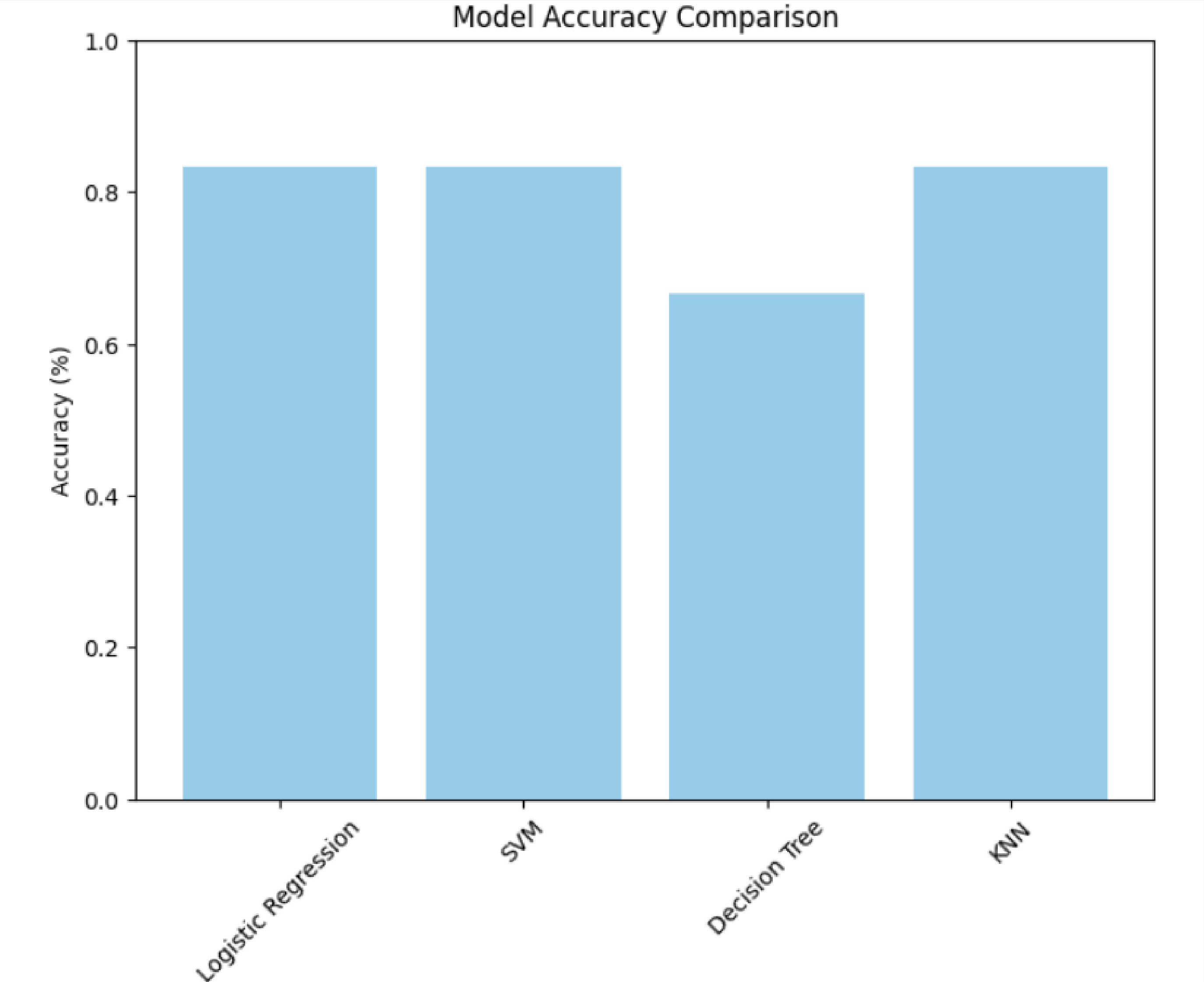


Section 5

Predictive Analysis (Classification)

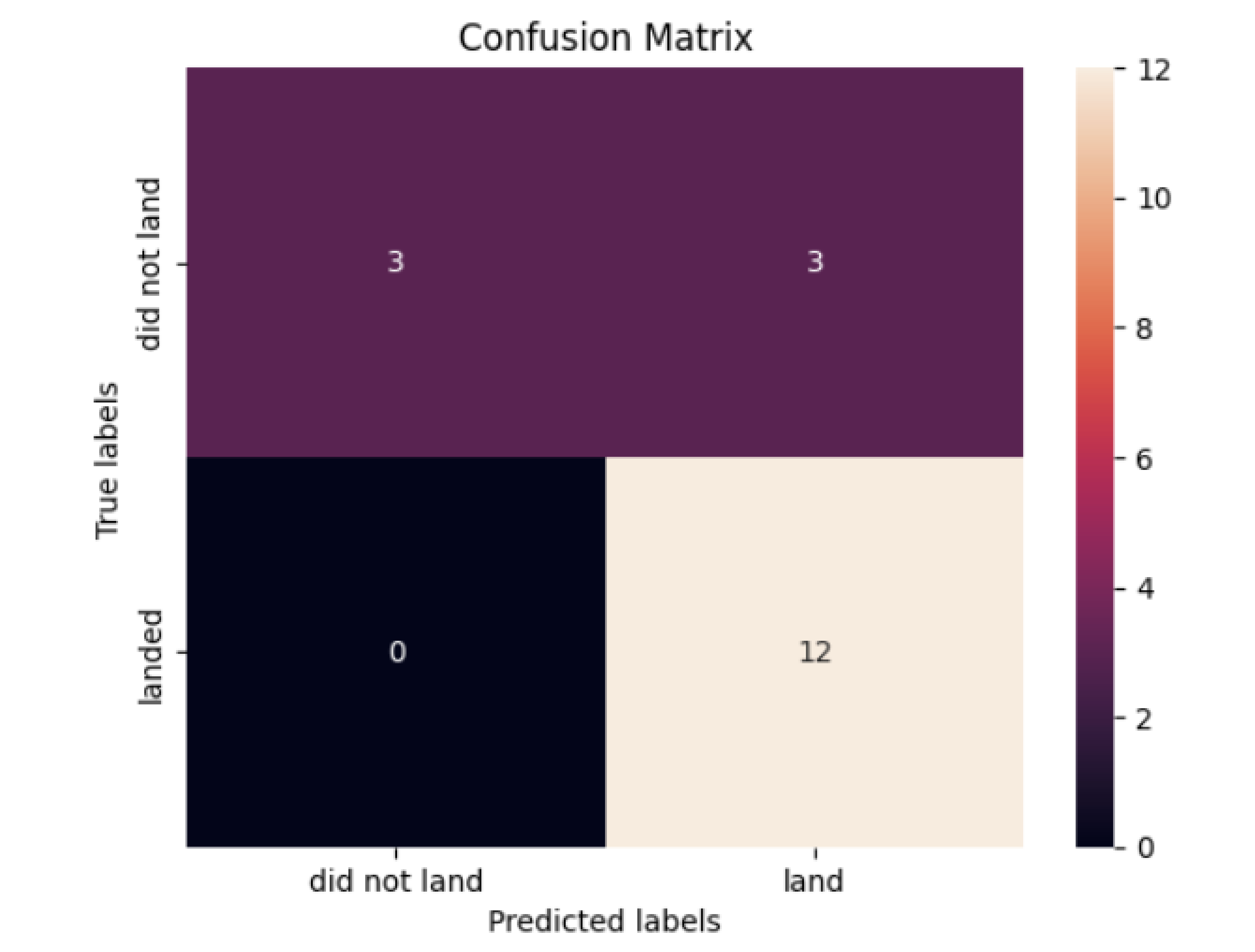
Classification Accuracy

- Logistic Regression is the model with the highest accuracy



Confusion Matrix

There are 12 True Positives, 3 True Negatives, 3 False Positives, and 0 False Negatives



Conclusions

- The analysis yielded several key findings from the SpaceX Falcon 9 first-stage landing dataset:
  - **Launch Success Over Time**: The success rate of Falcon 9 first-stage landings has significantly improved over time, particularly as more flights were completed.
  - **Top Launch Site**: KSC LC-39A had the highest success rate among the launch sites, with Falcon 9 landings showing increasing success as flight numbers grew.
  - **Orbit Type Success**: Orbits such as ES-L1, GEO, HEO, and SSO achieved a 100% success rate for Falcon 9 first-stage landings, while others showed varying levels of success.
  - **Payload and Success**: The success rate of landings correlated with payload mass, with heavier payloads generally leading to higher success rates, especially for specific orbit types like LEO and ISS.
  - **Predictive Model Findings**: Machine learning models used for prediction showed that Logistic Regression, SVM, and KNN performed similarly, with Decision Trees slightly outperforming the others in accuracy.
- Overall, the analysis demonstrated a trend toward higher success rates for Falcon 9 first-stage landings and highlighted the key factors influencing these outcomes, including launch site, payload mass, and orbit type.

Thank you!