IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daayum Mohsin
28 August 2024

Outline

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

Executive Summary

Methodology Overview:

- Data Collection from SpaceX API
- Web Scraping for Additional SpaceX Data
- Data Wrangling and Preparation
- Exploratory Data Analysis (EDA) with SQL
- Data Visualization Using Python, Pandas, and Matplotlib
- Launch Site Analysis with Interactive Visual Tools (Folium & Plotly Dash)
- Machine Learning for Landing Prediction

Key Results Summary:

- Findings from Exploratory Data Analysis (EDA)
- Interactive Dashboards and Visual Analytics
- Predictive Modeling (Classification Analysis)

Introduction

- SpaceX has fundamentally transformed the space industry through the development of a series of groundbreaking rockets, each designed with a focus on cost-efficiency, reusability, and high performance.
- Building on the success of Falcon 1, SpaceX developed the Falcon 9, a larger, more powerful rocket capable of delivering heavier payloads to orbit. First launched in 2010, the Falcon 9 is a two-stage rocket powered by nine Merlin engines in the first stage and a single Merlin Vacuum engine in the second stage.
- The cost of the launch was 62 million USD, other providers have costs upwards of 165 million USD. This is due to saving launchers and costs vary due to a successful first stage landing.

<u>Problems you want to find answers:</u>

- The primary problem is to determine whether the first stage of the Falcon 9 rocket will successfully land. This is crucial because SpaceX's ability to reuse the first stage of the rocket is a key factor in reducing launch costs. Predicting the success of these landings can help estimate the overall cost of a launch.
- Another problem addressed is how this prediction can be used by other companies. If an alternative company wants to bid against SpaceX for a rocket launch, understanding the likelihood of a successful first stage landing can influence the bid price and overall competitiveness in the market.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
    - The data collected was enhanced by generating a landing outcome label derived from the outcome data after summarizing and analyzing the features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - The data collected up to this point underwent a normalization processs to ensure consistency and comparability across all features.
  - After normalization, the data was then split into training and test datasets to facilitate model evaluation.
  - **Four distinct classification models** were applied to the data, each model designed to predict outcomes based on the features provided.
  - The performance of these models was rigorously assessed, with the accuracy of each model being evaluated across a variety of parameter combinations.

Data Collection

Data sets were collected:

Space X API (https://api.spacexdata.com/v4/rockets/)

and

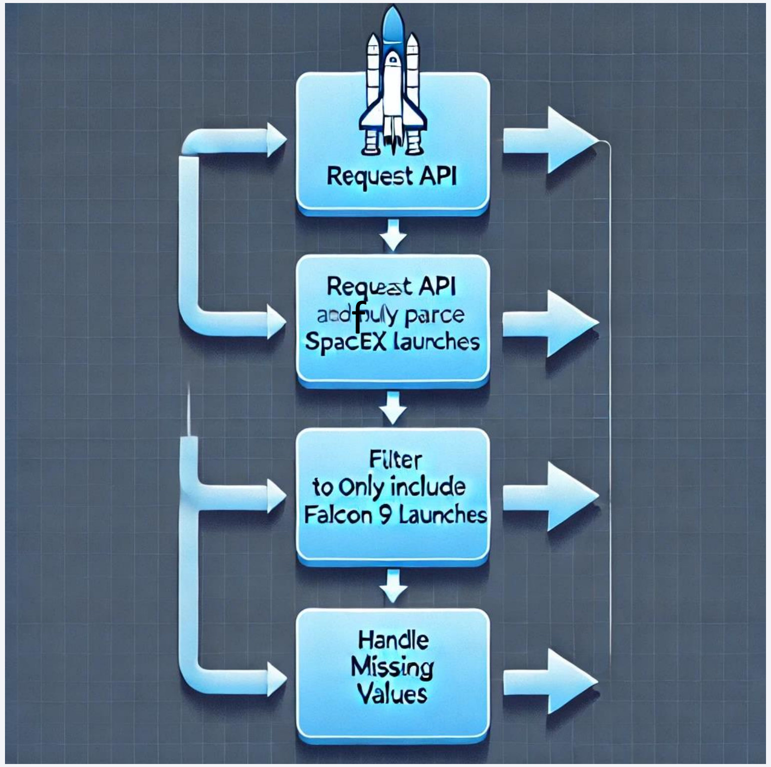
From Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping techniques.

Data Collection – SpaceX API

•SpaceX provides a public API that allows for the retrieval and subsequent utilization of data.

•This API was utilized following the steps outlined in the adjacent flowchart, with the processed data being stored for further analysis.

- Source Code: URL:https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(1)%20Space-X-%20first-landing-Data%20Collection-API.ipynb



Data Collection - Scraping

•Data from SpaceX launches can also be obtained from Wikipedia.
•The data is downloaded from Wikipedia following the steps outlined in the flowchart, after which it is stored for further use

- Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(2)%20Web%20scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20from%20Wikipedia.ipynb

Request the Falcon9 Launch Wiki page

⇩

Extract all column/ variable names from the HTML table header

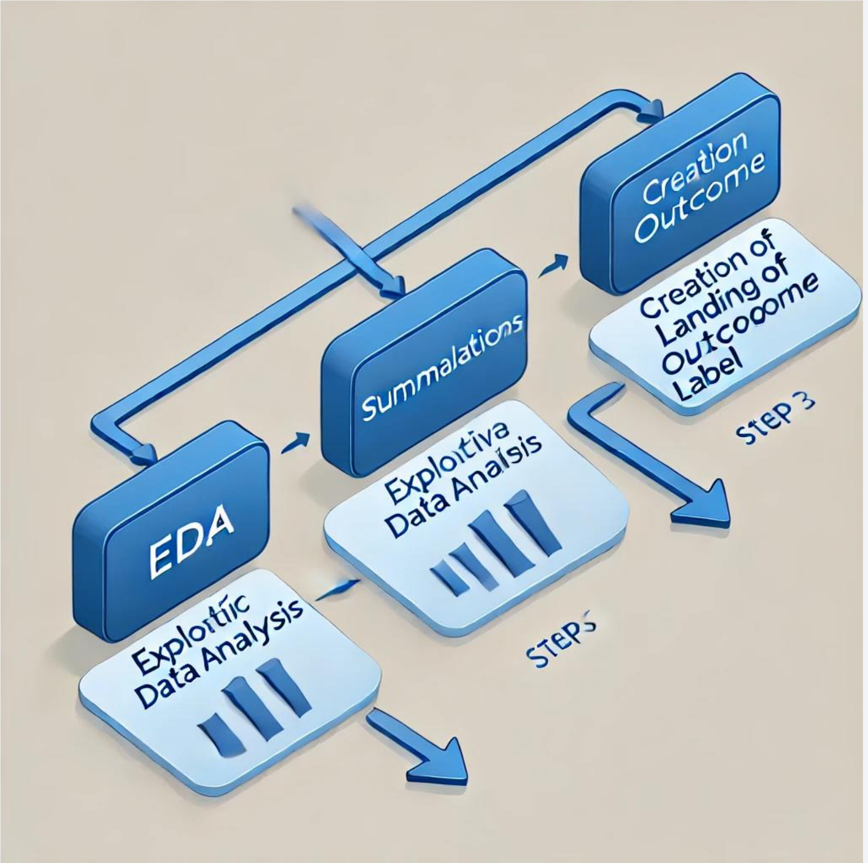
⇩

Create a data frame by parsing the launch HTML tables

Data Wrangling

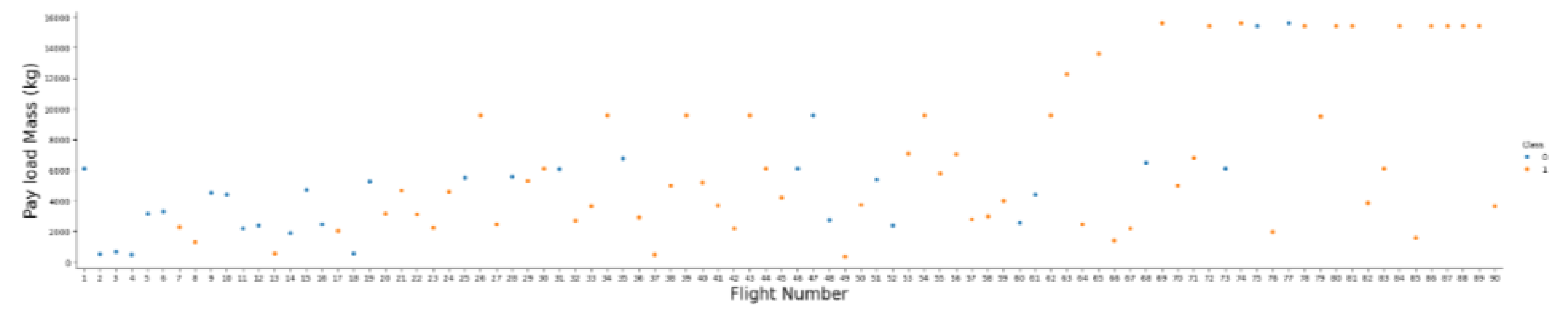
•The process began with performing Exploratory Data Analysis (EDA) on the dataset.
•Next, summaries were generated, including the number of launches per site, the frequency of different orbit types, and the mission outcomes associated with each orbit.
•Finally, a landing outcome label was created based on the data from the Outcome column.

•Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(3)%20Space-X-%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatterplots and Barplots were utilized to examine the relationships between pairs of features in the dataset:
- Key features:
  - -Payload Mass vs. Flight Number
  - -Launch Site vs. Flight Number
  - -Launch Site vs. Payload Mass
  - -Orbit vs. Flight Number
  - -Payload vs. Orbit



Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(5)%20Space-X-%20Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

- SQL queries that were executed:
- **List of Unique Launch Sites**: Retrieved the names of all distinct launch sites used in space missions.
- **Top 5 Launch Sites Starting with 'CCA'**: Identified the top five launch sites whose names begin with the prefix 'CCA.'
- **Total Payload Mass by NASA (CRS) Boosters**: Calculated the total payload mass carried by boosters launched by NASA (CRS).
- **Average Payload Mass for F9 v1.1 Boosters**: Determined the average payload mass carried by the booster version F9 v1.1.
- **Date of First Successful Ground Pad Landing**: Found the date when the first successful landing on a ground pad was achieved.
- **Boosters with Successful Drone Ship Landings and Specific Payload Mass**: Listed the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg.
- **Total Count of Successful and Failed Mission Outcomes**: Tallied the total number of both successful and failed mission outcomes.
- **Booster Versions with Maximum Payload Mass**: Identified the booster versions that have carried the maximum payload mass.
- **Failed Drone Ship Landings in 2015**: Extracted the failed landing outcomes on drone ships in 2015, along with their booster versions and launch site names.
- **Rank of Landing Outcomes Between Specific Dates**: Ranked the count of landing outcomes, such as failures on drone ships and successes on ground pads, within the date range from 2010-06-04 to 2017-03-20.
- Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(4)%20SQL%20Notebook%20for%20Peer%20Assignment.ipynb

Build an Interactive Map with Folium

•Folium Maps were utilized with various elements such as markers, circles, lines, and marker clusters to visualize data.
•Markers were used to represent specific points of interest, such as launch sites.
•Circles highlighted areas around particular coordinates, like the NASA Johnson Space Center.
•Marker clusters grouped events at each coordinate, for example, multiple launches at a single launch site, while lines indicated distances between two locations

Source Code:
https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(6)%20Space-X-%20Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

•Various graphs and plots were employed to effectively visualize the data on dash.

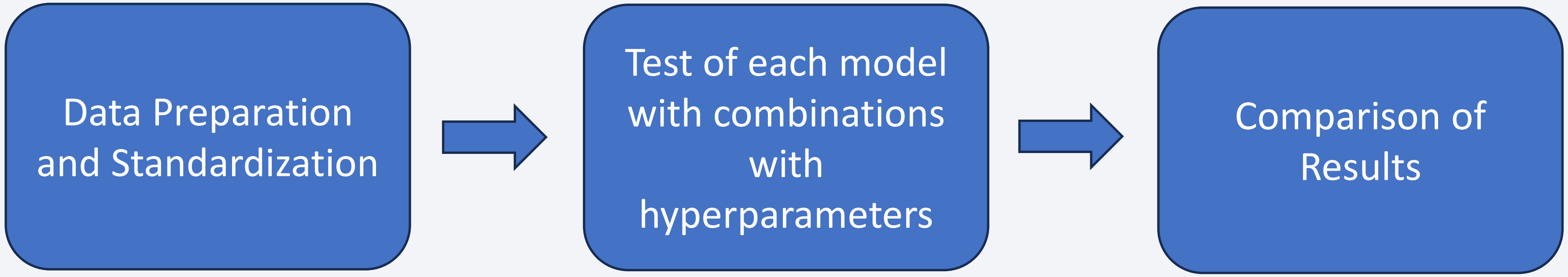
•A **pie chart** was used to depict the percentage distribution of launches by site.
•A **bar plot** illustrated the range of payloads across different launches.

•This combination of visualizations facilitated a quick analysis of the relationship between payloads and launch sites, aiding in the identification of the most suitable launch sites based on payload requirements.

Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(7)

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbours.



Source Code: https://github.com/Daayum03/Applied-Data-Science-Capstone/blob/main/(8)%20Space-X-%20Falcon%209%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results:
  - SpaceX operates from four distinct launch sites.
  - The initial launches were conducted for SpaceX and NASA.
  - The F9 v1.1 booster has an average payload capacity of 2,928 kg.
  - The first successful landing outcome occurred in 2015, five years after the inaugural launch.
  - Several Falcon 9 booster versions successfully landed on drone ships, often carrying payloads above the average weight.
  - Nearly all mission outcomes have been successful, with close to 100% success rate.
  - In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships.
  - The success rate of landing outcomes has improved progressively over the years.

Results

•Interactive analytics revealed that launch sites are typically located in secure areas, often near the sea, and are supported by robust logistical infrastructure.

•The analysis highlighted that most launches occur at East Coast launch sites.

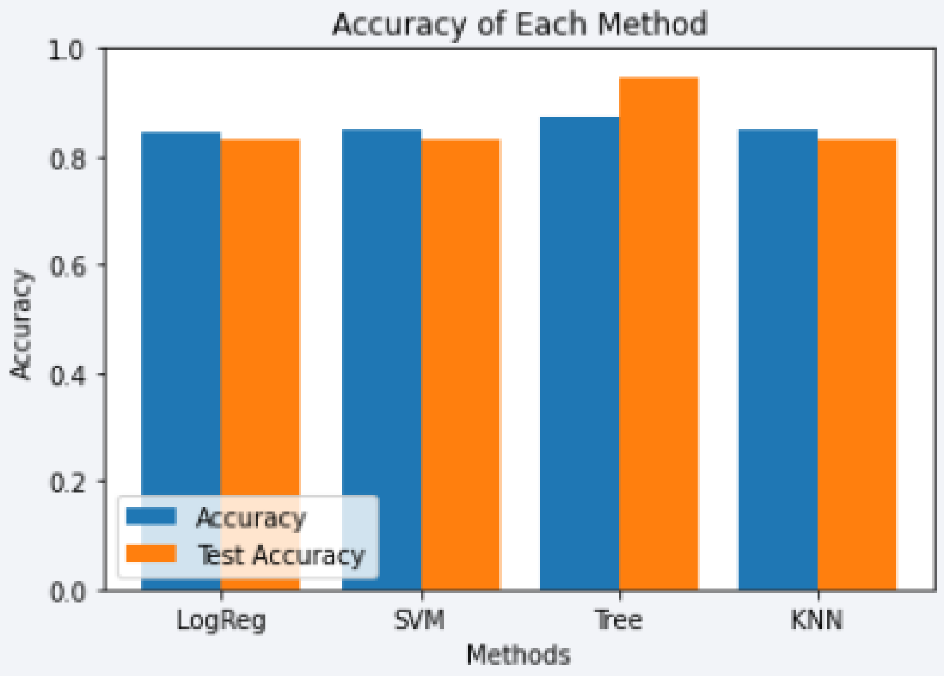
•These strategically chosen locations not only ensure safety but also optimize the efficiency of launch operations, as all launch sites are in proximity to the equator line.

Results

•The Logistic Regression, SVM, and KNN models all had similar performance, with accuracies slightly below the Decision Tree Classifier, suggesting they are also viable options depending on the specific application.

•The Decision Tree Classifier outperformed the other models, achieving the highest accuracy with over 87% in both training and test datasets had an accuracy of 94%.

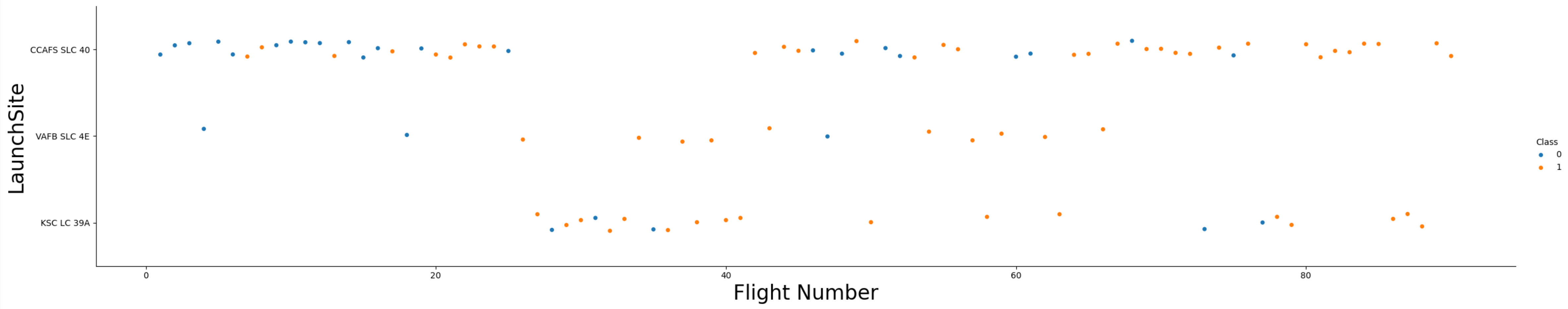
•All models demonstrated consistent performance, with accuracies generally above 80%, indicating reliable classification across different methods.



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

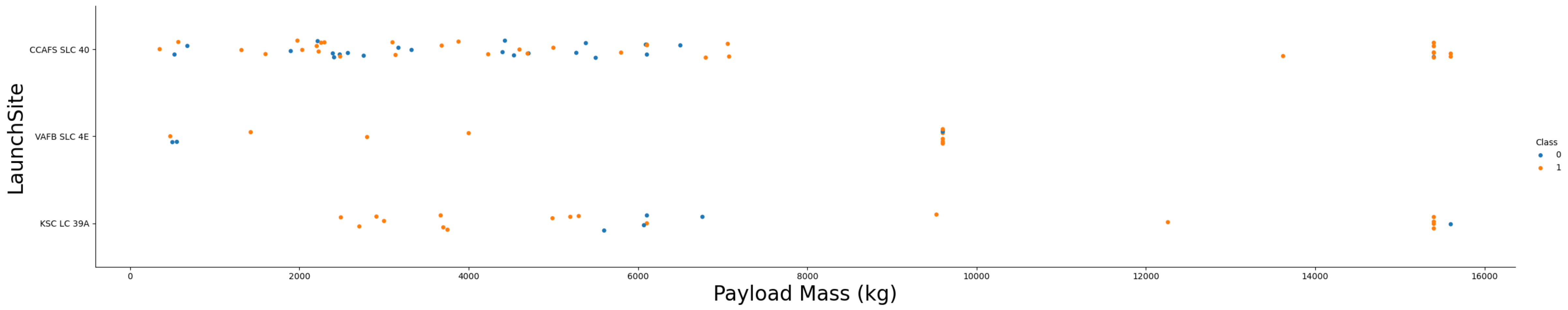


- The plot highlights CCAF5 SLC 40 as the top-performing launch site, with the highest success rate in recent launches.

- VAFB SLC 4E ranks second, followed by KSC LC 39A in third place.

- The data also shows a noticeable improvement in the overall success rate over time.

Payload vs. Launch Site



•Payloads above 9,000kg consistently achieve high success rates.

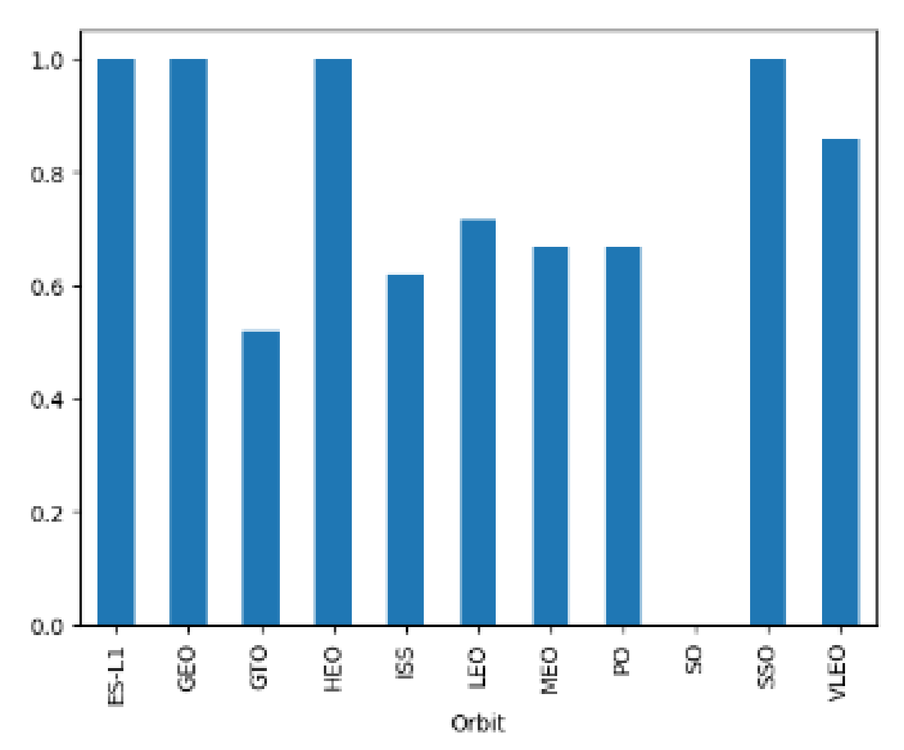
•Launch sites CCAFS SLC 40 and KSC LC 39A are the only locations capable of handling payloads exceeding 12,000kg.

•These sites are crucial for heavy payload missions, ensuring successful outcomes for the most demanding launches.

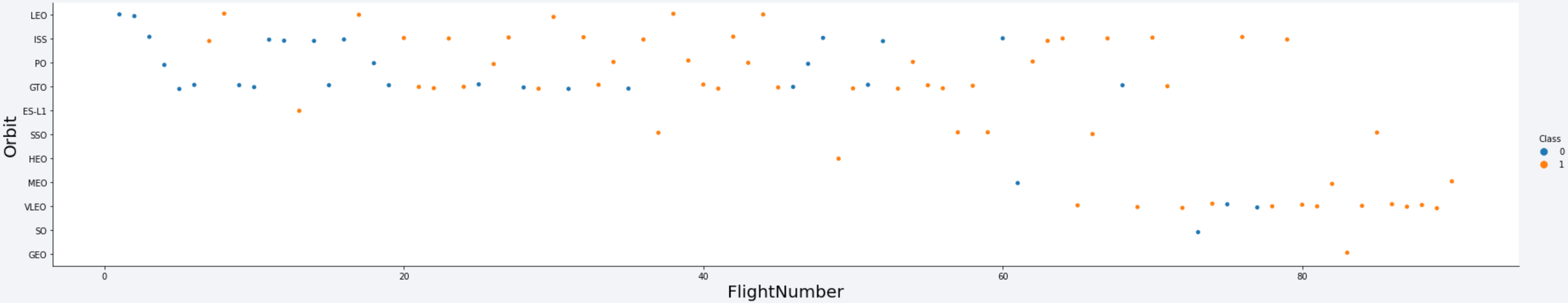
Success Rate vs. Orbit Type

- Highest Success Rates:
  - ES-L1 (Earth-Sun L1)
  - GEO (Geostationary Earth Orbit)
  - HEO (Highly Elliptical Orbit)
  - SSO (Sun-Synchronous Orbit)

- Next Highest Success Rates:
  - VLEO (Very Low Earth Orbit): Above 80%
  - LEO (Low Earth Orbit): Above 70%



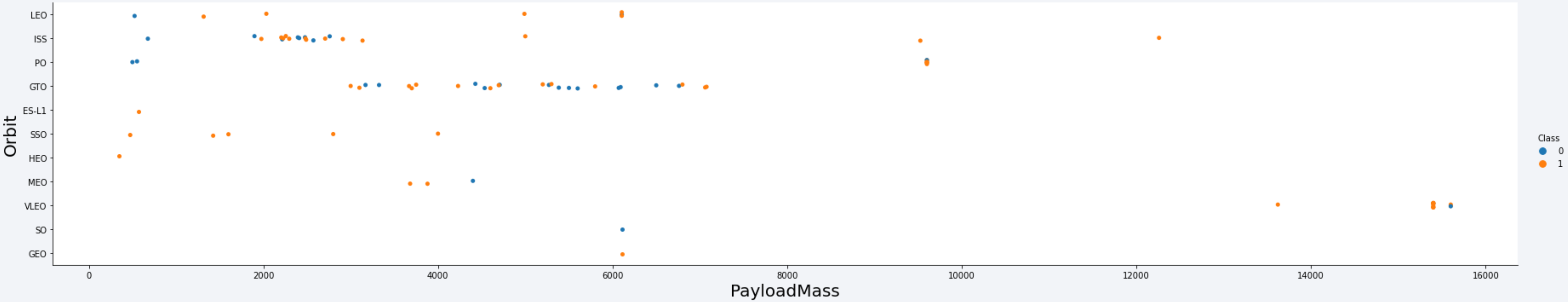
Flight Number vs. Orbit Type



•Improved Success Rate: The success rate of missions has improved over time across all orbital destinations.

•VLEO as a Business Opportunity: The recent increase in frequency of missions to Very Low Earth Orbit (VLEO) suggests it is emerging as a new business opportunity.

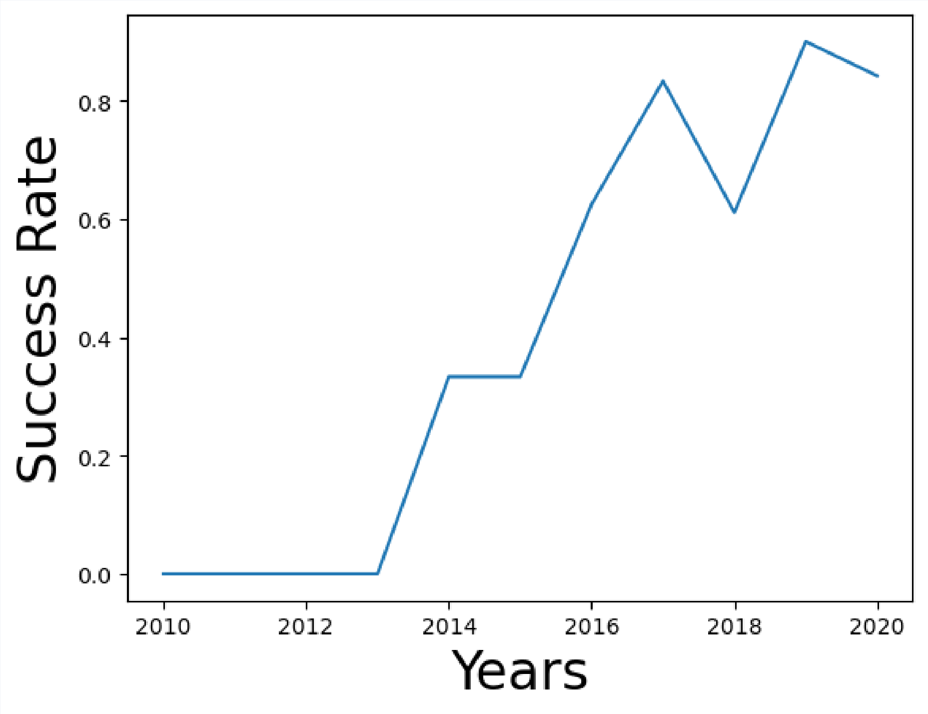
Payload vs. Orbit Type



•GTO Payload and Success Rate: No correlation exists between payload size and success rate for missions to Geostationary Transfer Orbit (GTO).
•ISS Orbit Success: The ISS orbit supports a wide range of payloads and maintains a strong success rate.
•SO and GEO Launches: Launches to Sun-synchronous Orbit (SO) and Geostationary Orbit (GEO) are relatively few in number.

Launch Success Yearly Trend

- From the period 2013-2019 there is a dramatic increase in the Success likelihood. Due to supply constraints from COVID-19, the success likelihood had decreased in 2020.
- The increase can be seen due to improvements in rocket technology, and through periods of adjustments to new technology, where the likelihood is stagnant i.e. 2014-2015.



All Launch Site Names

- According to the Data given, these are the list of the four launch sites:

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

- All obtained information from “Launch_site” values in the dataset

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Listed above are five samples of Cape Canaveral Launches

Total Payload Mass

- The total payload carried by boosters from NASA:

Total Payload Mass (kg)
111,268

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is:

Average Payload Mass (kg)
2,928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad:

First Success Date
2015-12-22

- This highlights the period of time between 2010-2015, where Space-X had a period of No successful ground landings from its first inaugural launch in 2010. Due to the advancement of rocket technology, new heights were reached to conserve Falcon 9 rockets and saved millions of dollars.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here are the list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Booster Version 4000-6000
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:

Mission Outcome	Occurences
Success	99
Success(Payload status unclear)	1
Failure(During Flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Here is a list of the names of the booster which 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

2015 Launch Records

- A List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

- The list of recorded failures in 2015 were 2 occurrences. Both being at Cape Canaveral Launch point, Florida.

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

- Ranking the number of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:
- This table gives us an understanding that the number of No attempts is the highest occurrence between the 2010-2017 period.

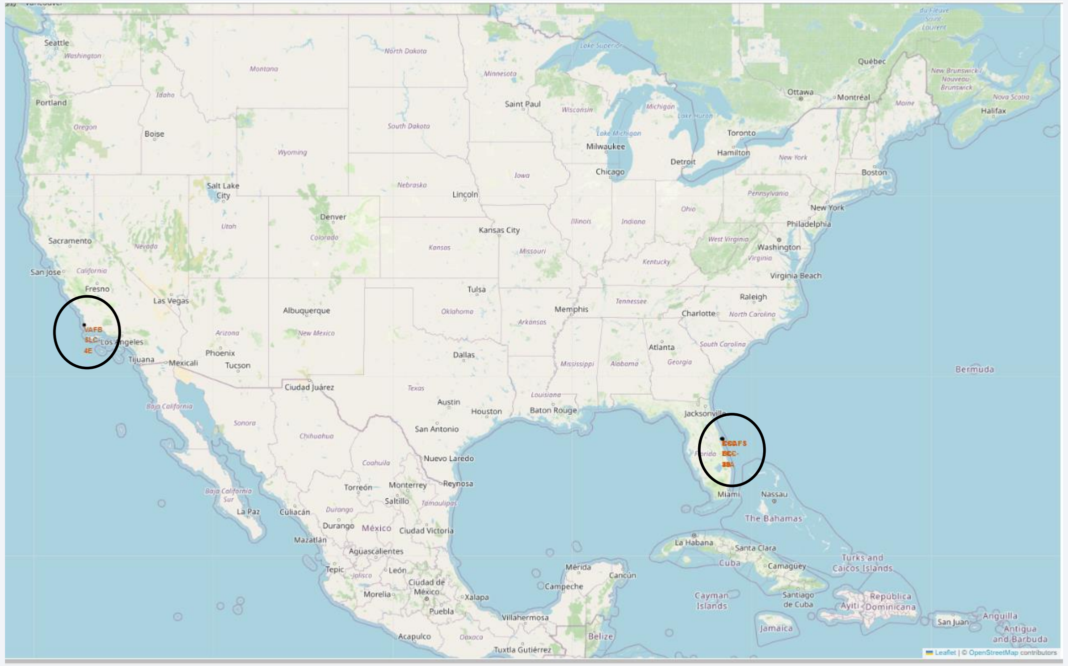
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success(ground pad)	3
Failure (parachute)	2
Uncontrolled(ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

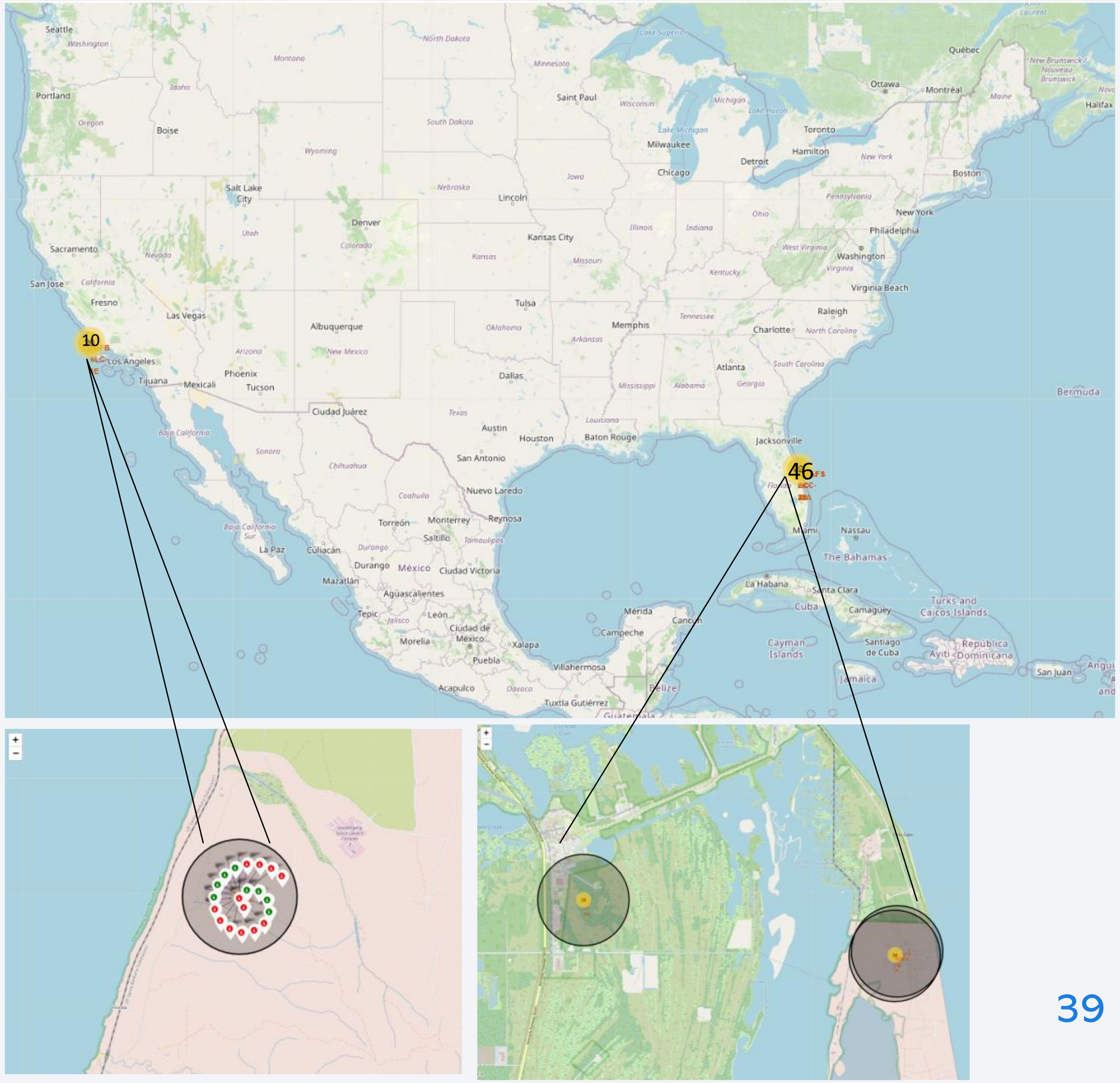
All Launch Sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.



Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes:
    - Green = Successful Launch
    - Red = Failed Launch

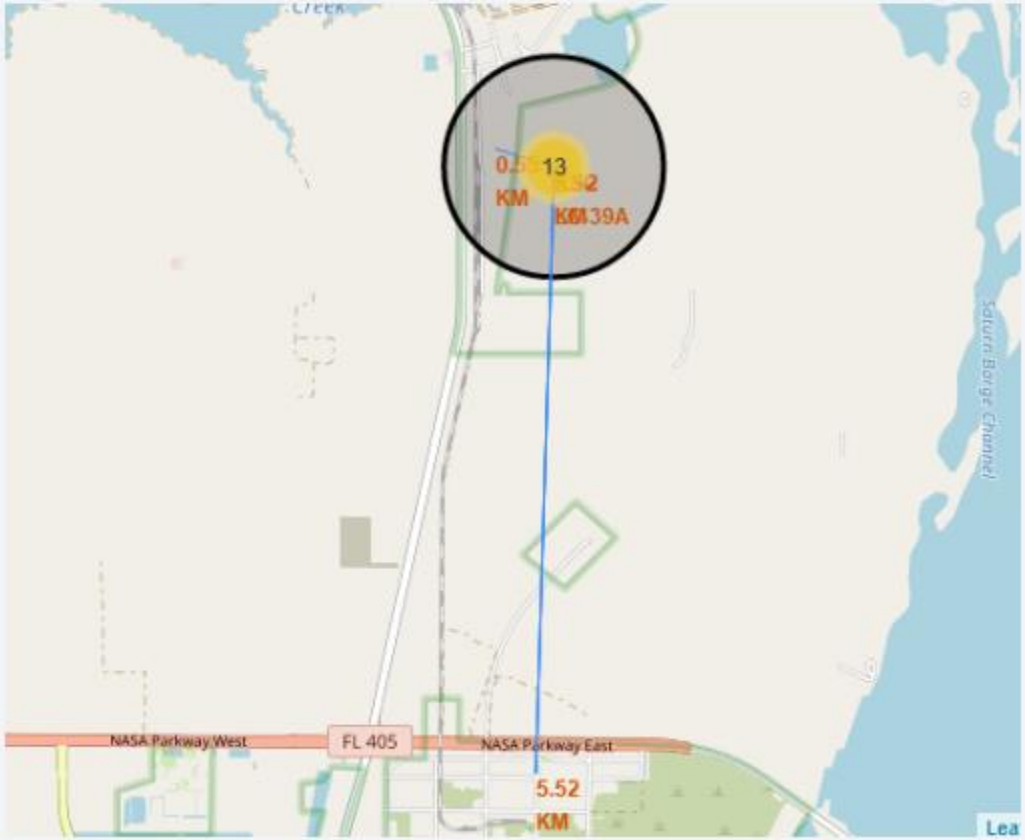


Proximity to Urban Infrastructure

•Strategic Location: Launch site KSC LC-39A is well-positioned near both railroad and major road infrastructure, facilitating efficient logistics and transportation.

•Safety Consideration: The site is relatively far from inhabited areas, reducing the risk to human populations in case of launch-related incidents.

•Accessibility: Proximity to key transportation networks ensures easy access for personnel, equipment, and materials necessary for launch operations

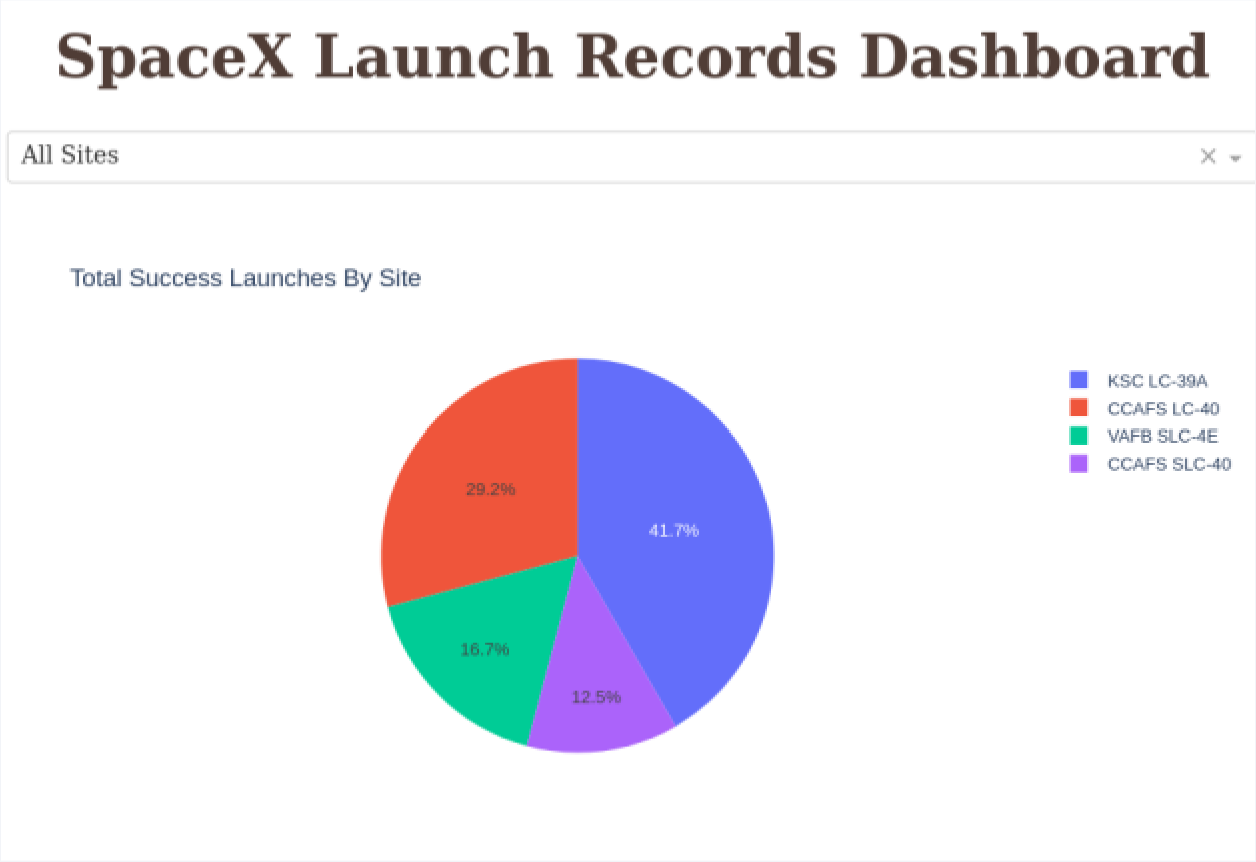


Section 4

Build a Dashboard with Plotly Dash

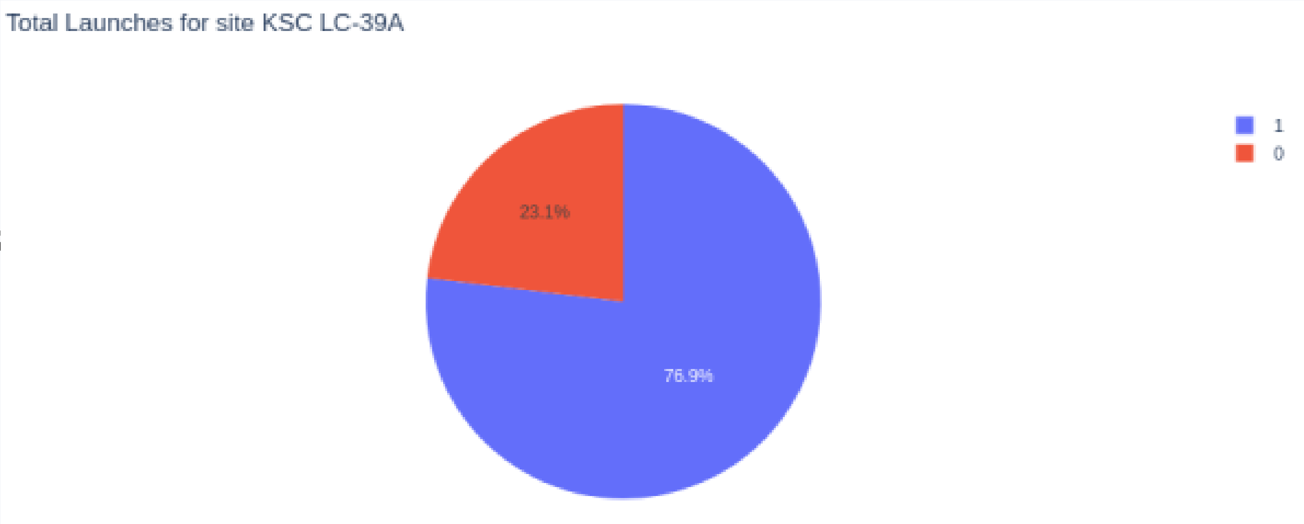
Space-X Successful Launches by Site

- The pie chart displayed shows the distribution of successful launches across various SpaceX launch sites, with KSC LC-39A having the highest success rate at 41.7%, followed by CCAFS LC-40 at 29.2%.
- This shows the dominance of KSC LC-39A in terms of successful launches, indicating its significance as a primary launch site.



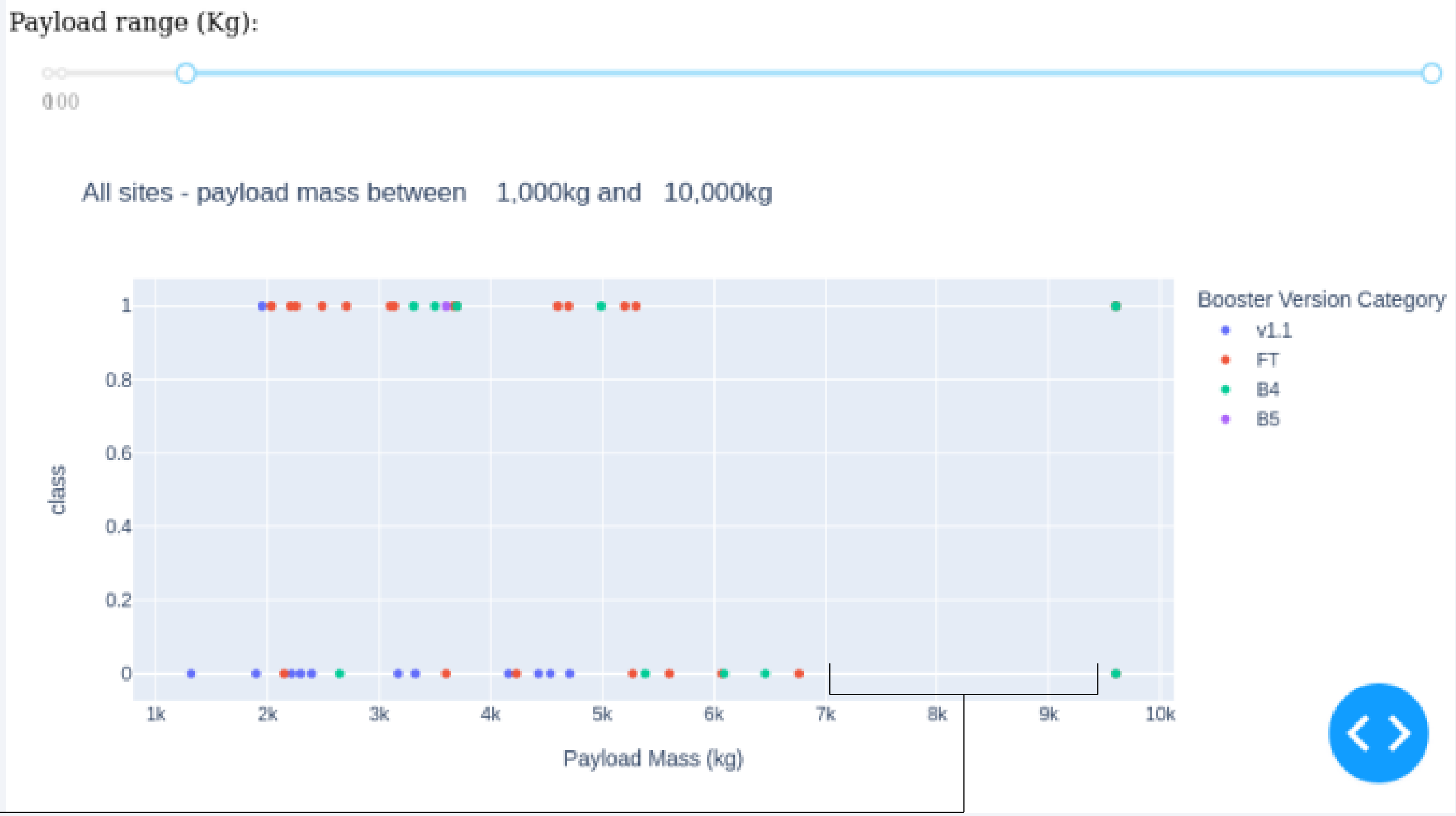
Successfulness of KSC LC-39A Launches

- The number of successful launches at the KSC LC-39A are 76.9%.
- Less than a quarter (23.1%) of launches were failed at the site NASA's Kennedy Space Center in Merritt Island Florida.



Dashboard Payload vs Outcome

- Payloads under 6,000kg and **FT** boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg

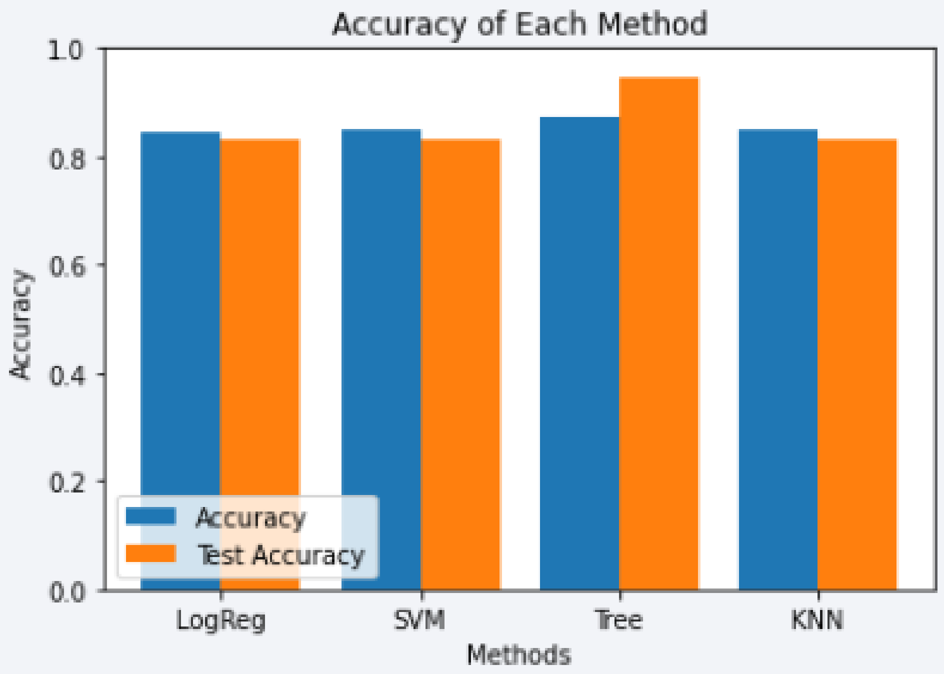


Section 5

Predictive Analysis (Classification)

Classification Accuracy

•Model Testing: Four classification models were tested, with their accuracies compared side by side.
•Top Performer: The Decision Tree Classifier emerged as the most accurate model, achieving accuracy rates exceeding 87%.



Confusion Matrix

- The **confusion matrix** provided shows the performance of a Decision Tree Classifier.
- Here's an analysis:
  - True Positives (Bottom-right, 12 cases): The classifier correctly predicted 12 instances where the event (landing) occurred, demonstrating strong performance in identifying successful landings.
  - True Negatives (Top-left, 5 cases): The model correctly identified 5 instances where the event did not occur (did not land), further proving its reliability.
  - False Positives (Top-right, 1 case): There was 1 instance where the model incorrectly predicted a landing when there was none, indicating a small error in predicting successful outcomes.
  - False Negatives (Bottom-left, 0 cases): The classifier did not incorrectly predict any non-landings when a landing occurred, which is a positive outcome showing zero false negatives.

Conclusions

•**Data Source Analysis and Refinement:**
Multiple data sources were analyzed, and conclusions were refined throughout the process to ensure accuracy and reliability.

•**Optimal Launch Site Identified:**
The analysis indicates that the best launch site is Kennedy Space Center Launch Complex 39A (KSC LC-39A), Merritt Island, Florida.

•**Weight-Based Risk Assessment:**
Launches carrying payloads above 7,000 kg are found to be less risky, suggesting that heavier missions have a higher success rate.

•**Improvement in Mission and Landing Success:**
Although most missions result in successful outcomes, the analysis reveals that successful landing outcomes have improved over time. This improvement is attributed to the evolution of processes and advancements in rocket technology.

•**Predictive Modeling for Landing Success:**
A Decision Tree Classifier has been identified as an effective tool for predicting successful landings. Implementing this model could potentially increase profitability by reducing the risk of failed missions.

Appendix

- Folium images were not produced on GitHub, so therefore I took screenshots and uploaded them.
- An improvement to make would be avoiding the duplication of datasets to have a double number of samples in the folium map charts.
- To understand the differences between accuracy and True accuracy

Thank you!