IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Rayyan
12/17/2025

Outline

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

Executive Summary

- Project Objective: To predict the successful landing of the Falcon 9 first-stage booster in order to estimate launch costs and inform competitive bidding strategies.
- Data Methodology: Implemented a complete Data Science pipeline including Data Collection (SpaceX API & Web Scraping), Data Wrangling, and Feature Engineering (One-Hot Encoding).
- Key Insights: Exploratory Data Analysis (EDA) revealed that Flight Number (experience) and Payload Mass are the strongest indicators of landing success.
- Visual Analytics: Developed interactive Folium maps to analyze launch site geography and a Plotly Dash dashboard to visualize real-time performance metrics.
- Predictive Modeling: Trained and optimized four classification algorithms (Logistic Regression, SVM, Decision Tree, and KNN) using GridSearchCV.
- Final Result: The models achieved a Test Accuracy of ~83.33%, successfully demonstrating that historical launch data can reliably predict landing outcomes.

Introduction

- As a Data Scientist at SpaceY, a new aerospace venture founded by Allon Mask, my objective is to determine the competitive pricing strategy required to bid against industry leader SpaceX. Because SpaceX’s cost structure is heavily dependent on their ability to reuse the Falcon 9 first stage, predicting a successful landing is crucial for estimating their launch costs.
- To derive these insights, I architected a full-stack data science pipeline. I began by harvesting historical launch data from the SpaceX API using Requests and scraping technical mission details from public records using BeautifulSoup. This raw data was flattened, wrangled, and cleaned using Pandas to ensure high data quality.
- For the exploratory analysis, I employed SQL for querying and utilized Matplotlib and Seaborn to visualize correlations between mission parameters, such as Orbit Type and Flight Number. To gain deeper spatial and interactive insights, I mapped launch sites with Folium and developed a dynamic dashboard using Plotly Dash to uncover hidden relationships impacting mission outcomes.
- Finally, I built predictive classification models using Scikit-Learn, implementing Logistic Regression, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN). By optimizing these models with GridSearchCV, I was able to improve upon the baseline dataset success rate of 66.67%, achieving a final model accuracy of 83.33% on test data.

Section 1

Methodology

Methodolgy

Executive Summary

Data collection methodology:

The data was collected using SpaceX API and Web scraping from wikipedia

Perform data wrangling

Using pandas, I flatten the data into tabular structure, cleaned missing values.

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

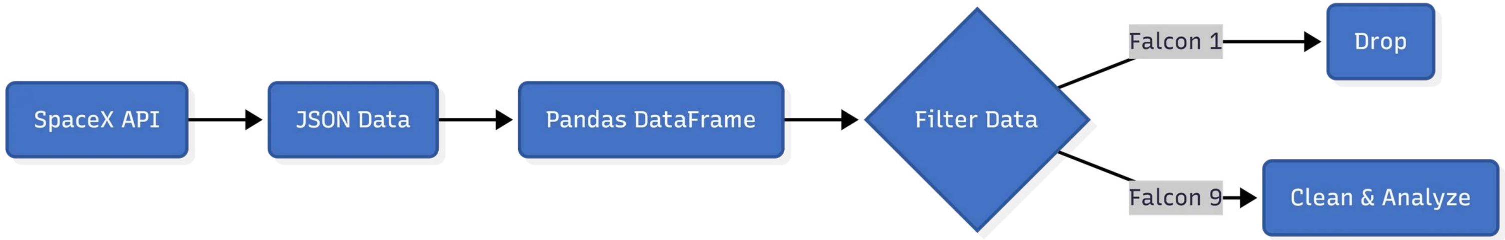
Perform predictive analysis using classification models

Most of my work with models was were tune using GridsearchCv.

Data Collection

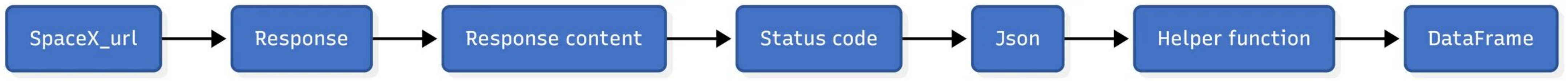
- Following extensive research into SpaceX's operations, I leveraged their publicly available REST API to acquire historical mission data. Utilizing the Python Requests library, I performed HTTP GET requests to retrieve raw mission telemetry in JSON format. This data was subsequently parsed and flattened into a structured Pandas DataFrame to facilitate granular analysis. A critical step in the wrangling process involved filtering out early "Falcon 1" missions to focus exclusively on the Falcon 9 architecture, which offers the relevant reusability data required for this study. During the initial cleaning phase, I addressed missing values across the dataset, while deliberately retaining null values in the LandingPad column, as these serve as a vital indicator for missions where no specific landing attempt was made (e.g., expendable missions or ocean ditching).

- Flow Chart



Data Collection SpaceX API

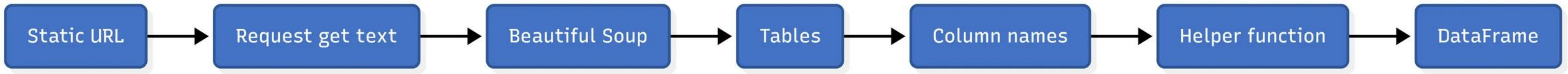
- The data collection pipeline begins by defining the target SpaceX URL and executing an HTTP GET request to obtain the server response. Upon receiving the response content, we first verify the status code to ensure a successful connection (200 OK). Once validated, the raw data is parsed into JSON format. To ensure the data is usable for machine learning, we apply specific helper functions (to normalize nested lists like payloads and cores) before finally flattening the structure into a clean Pandas DataFrame.
- Git Hub SpaceX API Calls Notebook
- Flow Chart



SpaceX_url → Response → Response content → Status code → Json → Helper function → DataFrame

Data Collection Scraping

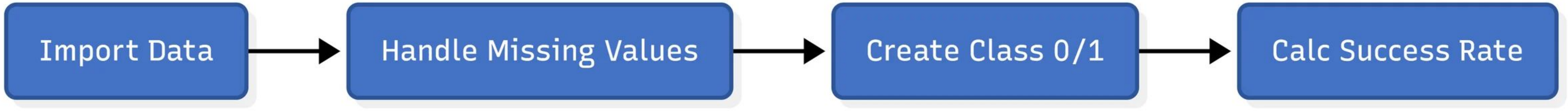
- To supplement the API data, I performed web scraping on the Wikipedia page for Falcon 9 launches. Starting with the Static URL, I executed an HTTP GET request to retrieve the page's raw HTML text. Using the Beautiful Soup library, I parsed this content to identify and extract specific launch Tables. After isolating the relevant Column names, I applied custom Helper functions to clean the data—such as normalizing dates and removing reference citations—before finally structuring the scraped records into a unified Pandas DataFrame.
- Git Hub URL Web Scraping Notebook
- Flow Chart



Static URL → Request get text → Beautiful Soup → Tables → Column names → Helper function → DataFrame

Data Wrangling

- In the Data Wrangling phase, I first inspected the dataset structure, specifically handling missing values in the LandingPad column which indicated missions where no landing was attempted. I then performed Feature Engineering to prepare the data for machine learning. The most critical step was converting the text-based launch outcomes into a binary 'Class' label, assigning a 1 for successful landings and a 0 for failures. Finally, I calculated the baseline success rate for the dataset, establishing a benchmark of 66.67% that our models would need to beat.
- Git Hub Data Wrangling NoteBook
- Flow Chart

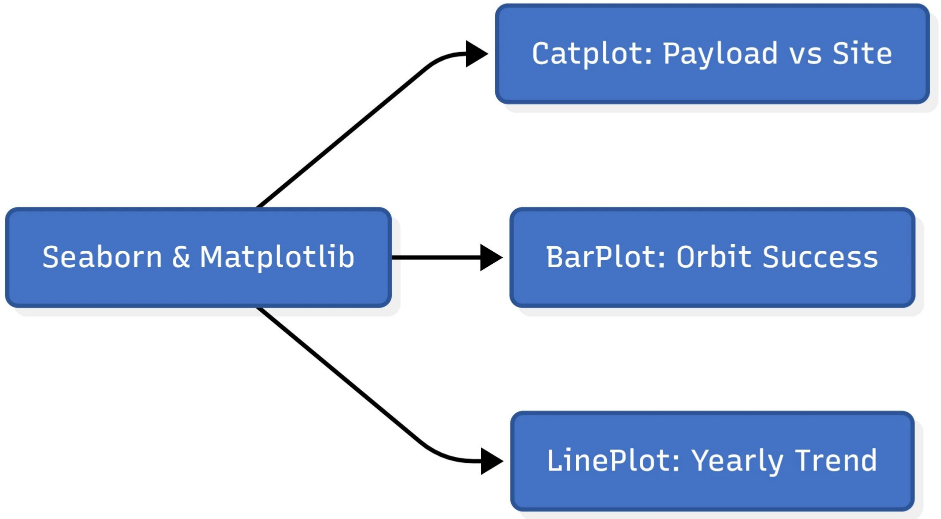


Import Data → Handle Missing Values → Create Class 0/1 → Calc Success Rate

EDA with Data Visualization

- In this phase, I utilized Seaborn and Matplotlib to visually explore the dataset. I created Categorical plots to analyze how payload mass influences landing success at different launch sites. I also generated Bar charts to identify which orbits yield the highest success rates, and used Line plots to track the historical improvement of Falcon 9 landings from 2010 to 2020, revealing a clear upward trend in reliability.

- Git Hub URL EDA with Data Visualization Notebook
- Flow Chart

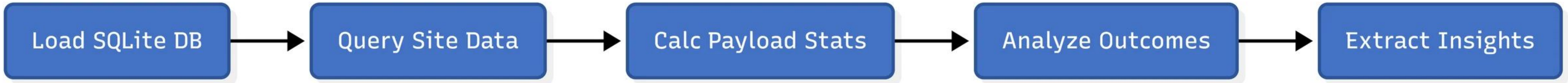


EDA with SQL

- In the SQL analysis phase, I queried the database to extract hard numbers behind the visuals. I started by validating unique launch sites and calculating key performance indicators, such as the total payload delivered for NASA. I also performed targeted queries to compare the success rates of ground pads versus drone ships, specifically looking at how heavy payloads between 4,000 and 6,000 kg affected landing stability. Finally, I sorted landing outcomes chronologically to pinpoint the exact timeframe where SpaceX's reliability began to improve significantly.

- Git Hub URL EDA with SQL Notebook

- Flow Chart



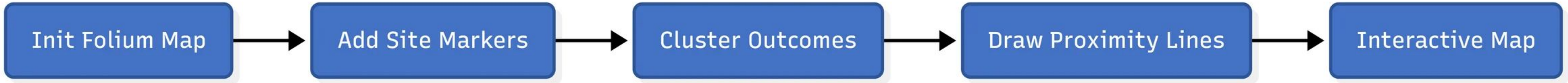
Load SQLite DB → Query Site Data → Calc Payload Stats → Analyze Outcomes → Extract Insights

Interactive Map with Folium

- To understand the geographical constraints of rocketry, I built interactive maps using Folium. I started by marking the exact locations of all launch sites using Pins and Circles. I then implemented Marker Clusters to visualize the density of successful versus failed launches at each site. Finally, to analyze safety and logistics, I drew PolyLines to measure the precise distance from launch pads to the nearest coastlines, cities, and railways, confirming that sites are strategically located to minimize risk.

- Git Hub URL Interactive Map with Folium Notebook

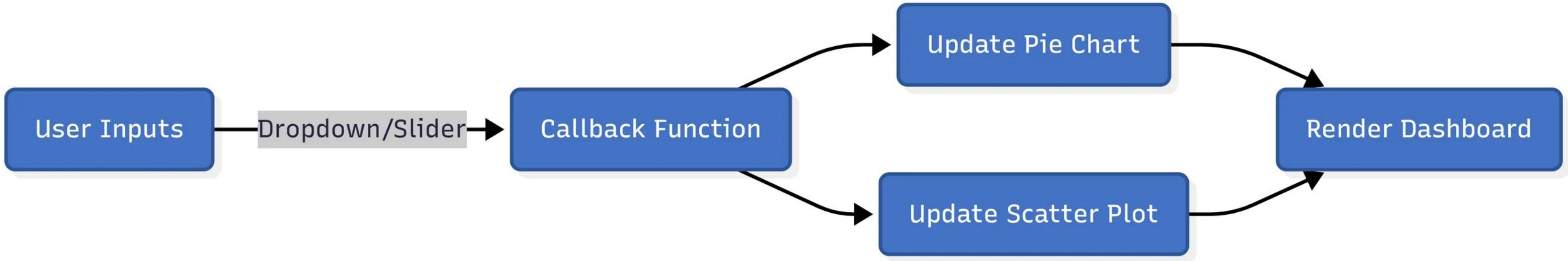
- Flow Chart



Init Folium Map → Add Site Markers → Cluster Outcomes → Draw Proximity Lines → Interactive Map

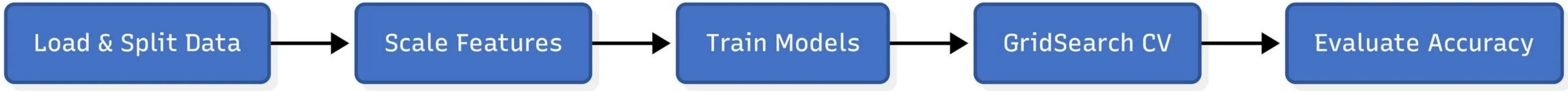
Dasboard with Plotly Dash

- To democratize the data insights, I built a fully interactive web application using Plotly Dash. The dashboard empowers users to conduct their own analysis through two key inputs: a Dropdown Menu for selecting launch facilities and a Range Slider for adjusting payload parameters. These inputs trigger Python callback functions that dynamically update the visual front-end, rendering a real-time Pie Chart for success rates and a Scatter Plot to visualize how heavy payloads affect different booster versions.
- Git Hub URL Plotly Dash lab
- Flow Chart



Predictive Analysis (Classification)

- In the final Predictive Analysis stage, I prepared the data by defining the feature matrix and target vector, followed by standardizing the values to ensure fair comparisons. I split the data, reserving 20% for testing, and built a machine learning pipeline using four distinct classifiers: Logistic Regression, SVM, Decision Trees, and KNN. To maximize performance, I used Grid Search Cross-Validation to tune the hyperparameters. Ultimately, the models proved robust, achieving an accuracy of approximately 83% on the test data, confirming that our historical data is a reliable predictor of future landing success.
- Git Hub URL Predictive Analysis Notebook
- Flow Chart



Result

- Total Payload: NASA (CRS) missions accounted for a total payload mass of 45,596 kg.
- Booster Capacity: The average payload mass for "Booster Version F9 v1.1" was 2,928.4 kg.
- Historical Milestone: The first successful ground pad landing occurred on December 22, 2015.
- Success Correlation: A clear positive correlation exists between Flight Number and Success Rate at CCAFS SLC 40.
- Orbit Reliability: Achieved a 100% success rate for launches to ES-L1, SSO, HEO, and GEO orbits.

Result

Texas
Killeen
College Station
Austin
New Braunfels
San Antonio
Ciudad Acuña
Piedras Negras
Nuevo Laredo
Coahuila
Corpus Christi
Houston
Sugar Land
NASA JSC
Beaumont
Louisiana
Baton Rouge
Lafayette
New Orleans
Mobile
Tallahassee
Dothan

- Site Location: Interactive markers pinpoint key facilities, such as the NASA Johnson Space Center (JSC) and launch complexes.
- Proximity Analysis: Visualized safety zones showing launch sites are strategically located near coastlines but maintain safe distances from populated areas.
- Outcome Clusters: Color-coded markers (Green=Success) reveal high-density success clusters at specific pads like KSC LC-39A.

◦ Source: Github

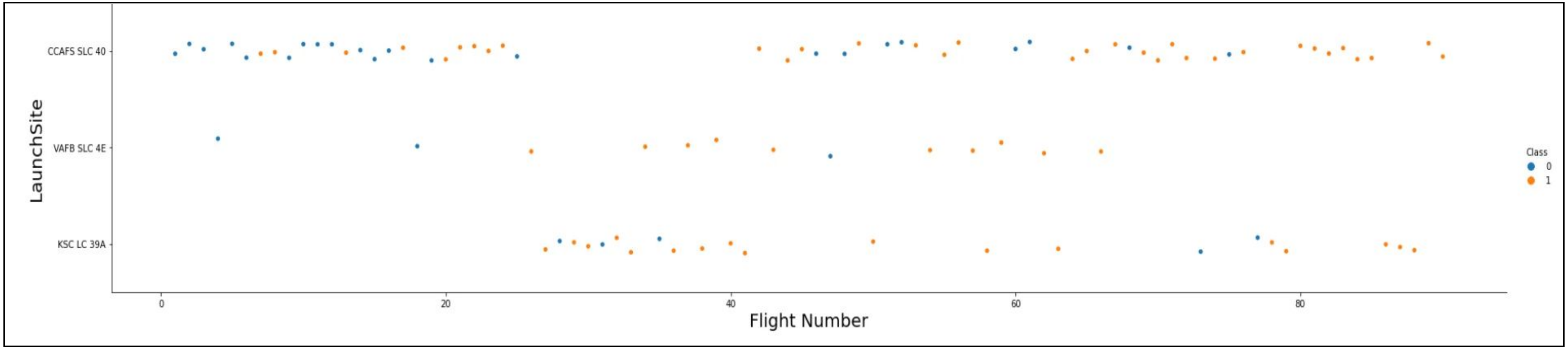
Result

- Algorithms Tested: Logistic Regression, SVM, Decision Tree, and KNN.
- Evaluation Metric: Accuracy on Test Set (20% of data).
- Best Performer: The Decision Tree Classifier (tuned via GridSearchCV).
- Final Accuracy: Achieved 88.89% with an F1-Score of 88.21%.
- Conclusion: The model provides high confidence in predicting successful first-stage landings for future Falcon 9 missions.

Section 2

Insights drawn from EDA

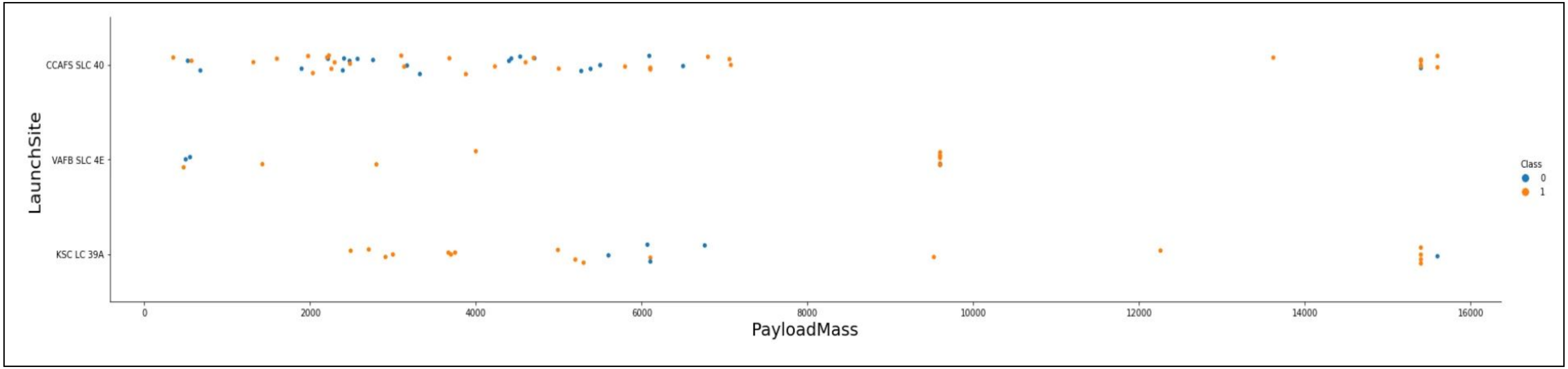
Flight Number vs. Launch Site



- Observation: Early launches showed failures at VAFB SLC-4E. Later flights (after ~80) demonstrate increased success at CCAFS SLC-40, with VAFB SLC-4E having fewer successful later flights.

◦ Source: Github

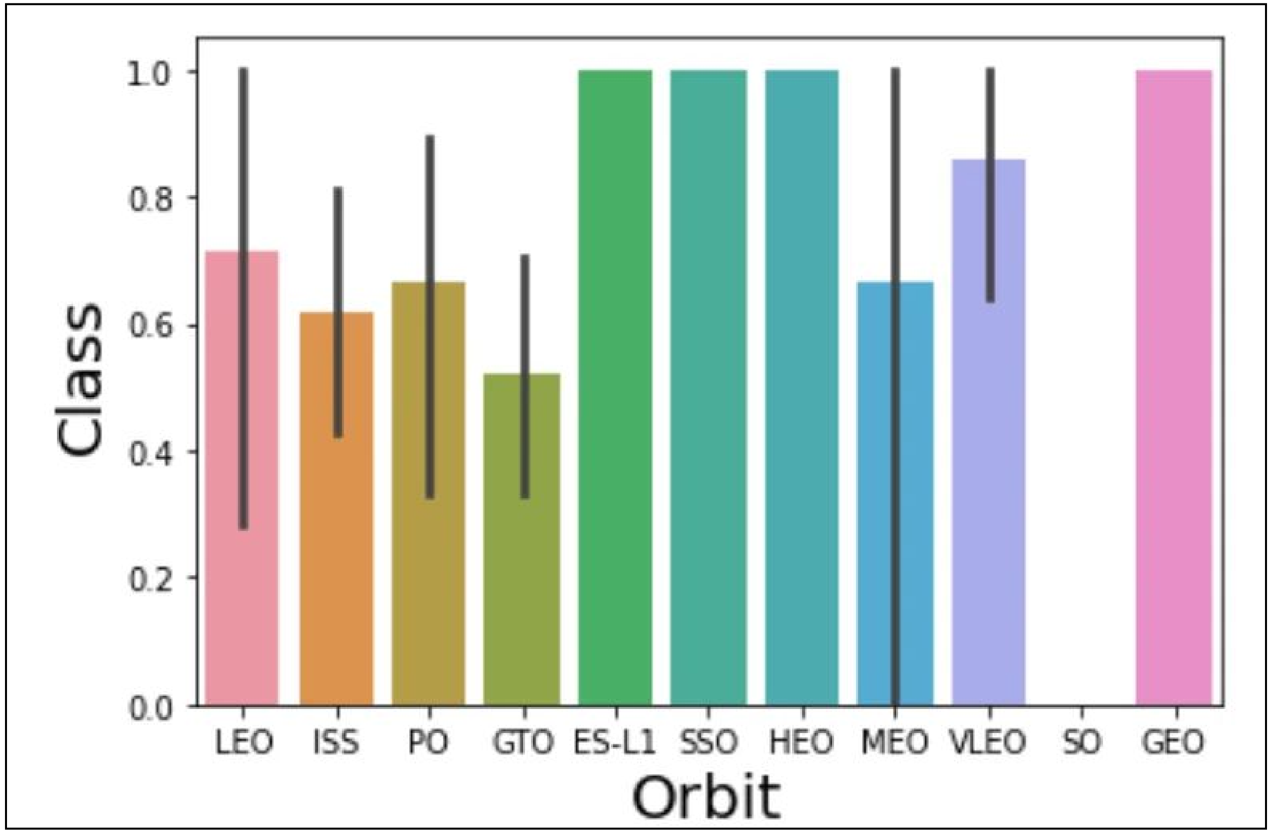
Payload vs. Launch Site



- Observation: Most successful landings occur with lower payload masses across all launch sites. Heavy payloads (>10,000 kg) are not consistently associated with successful landings, particularly from VAFB SLC-4E.

◦ Source: Github

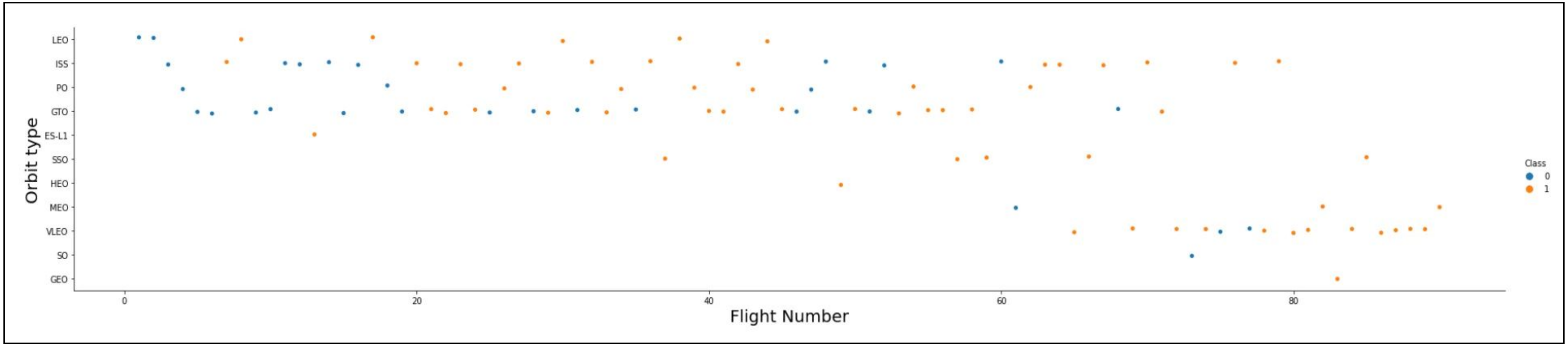
Success Rate vs. Orbit Type



- Finding: Orbits like ES-L1, SSO, HEO, and GEO exhibit 100% landing success. LEO and ISS have high success rates, while SO shows a lower success rate.

◦ Source: Github

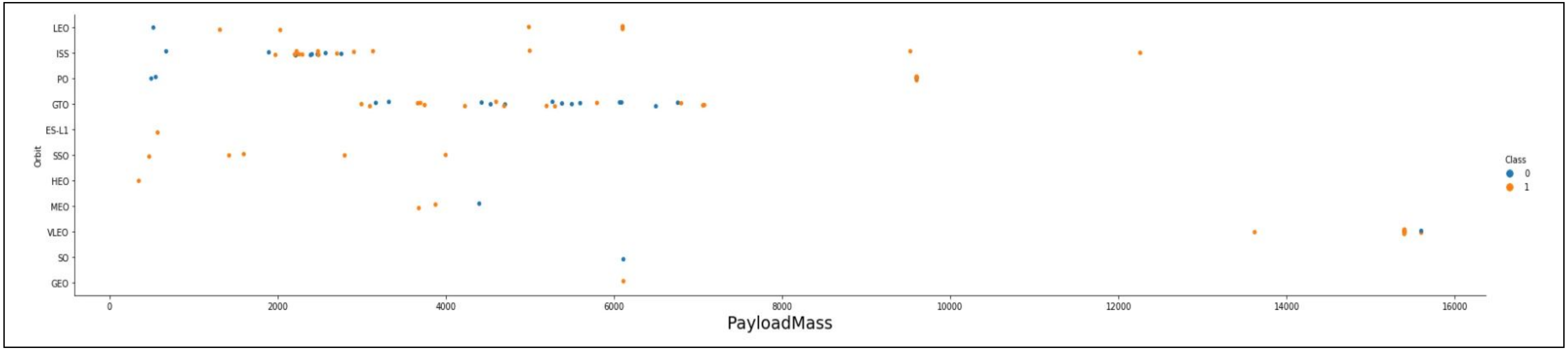
Flight Number vs. Orbit Type



- Observation: Success rates in VLEO orbits appear to improve for flights after number 60. GTO orbit shows no clear relationship between flight number and success.

◦ Source: Github

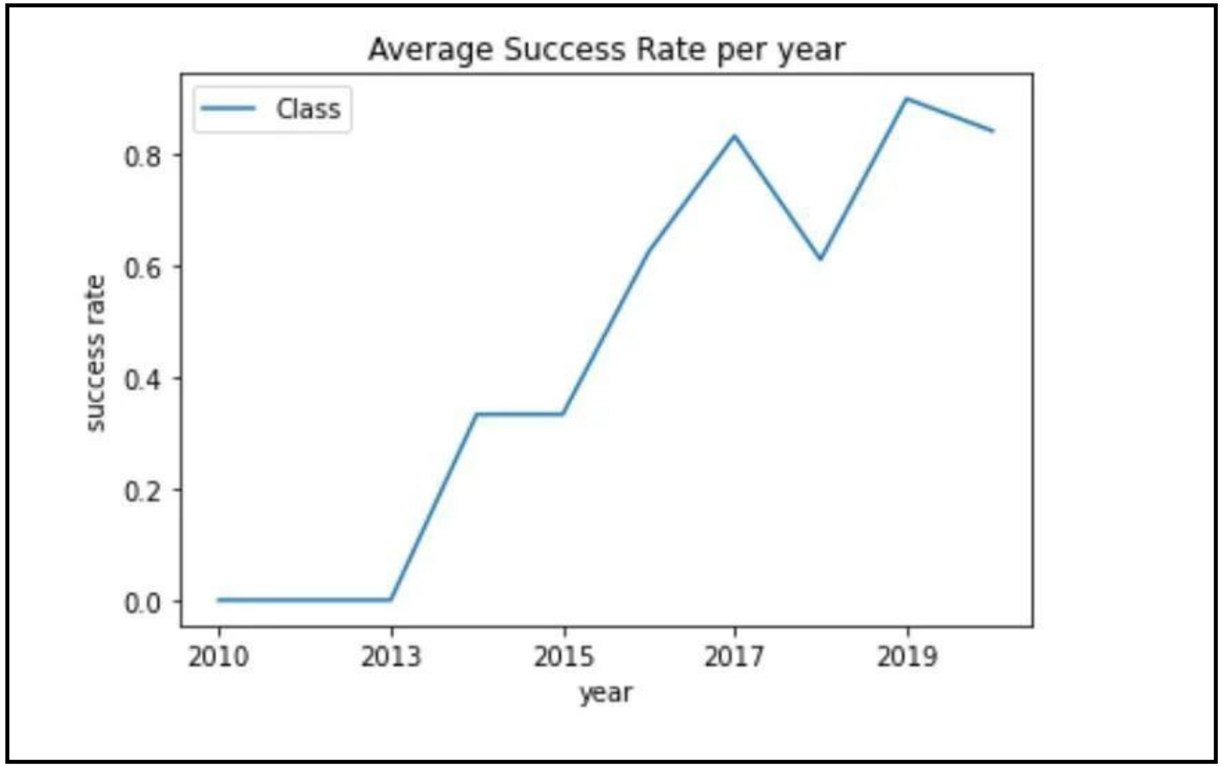
Payload vs. Orbit Type



- Finding: Heavier payloads have a higher success rate for Polar, LEO, and ISS orbits. GTO orbits show a mix of successful and unsuccessful landings regardless of payload mass.

◦ Source: Github

Launch Success Yearly Trend



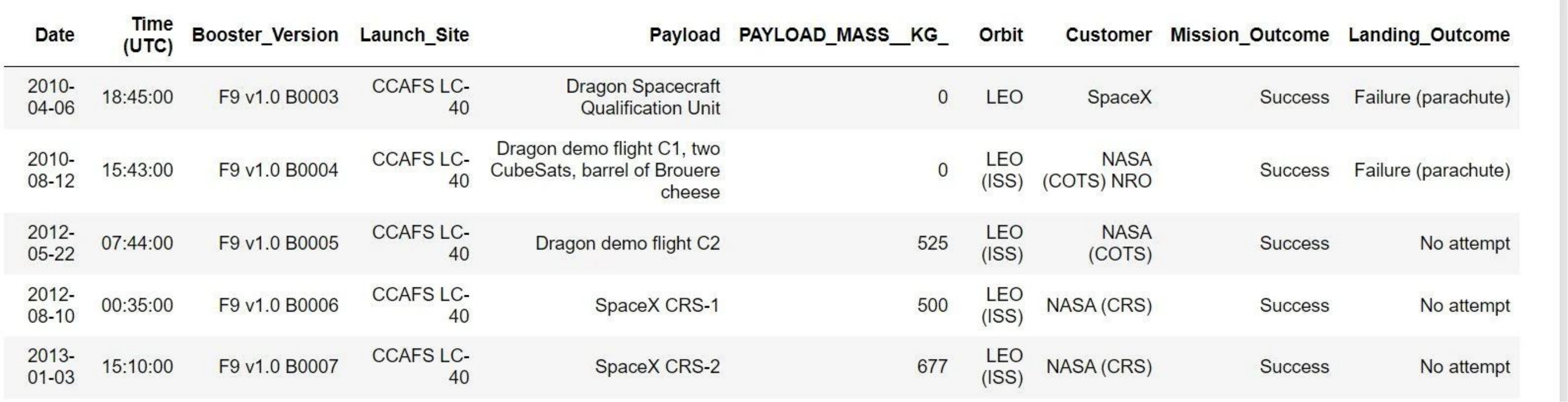
- Observation: The success rate of Falcon 9 landings has shown a consistent upward trend since 2013, stabilizing at a high level by 2017 and continuing through 2020.

◦ Source: Github

All Launch Site Names

- CCAFS LC-40 (Space Launch Complex 40): An orbital launch pad located in northern Cape Canaveral, Florida. It was formerly known simply as Launch Complex 40 (LC-40).
- VAFB SLC-4E (Vandenberg Space Force Base Space Launch Complex 4E): A launch site located at Vandenberg Space Force Base in California, U.S. It features two pads utilized by SpaceX for Falcon 9 missions: one for launch operations and the other as Landing Zone 4 for booster landings.
- KSC LC-39A (Kennedy Space Center Launch Complex 39A): Located at NASA's Kennedy Space Center on Merritt Island, Florida. It is the first of the three launch pads belonging to Launch Complex 39.

Launch Site Names Begin with 'CCA'

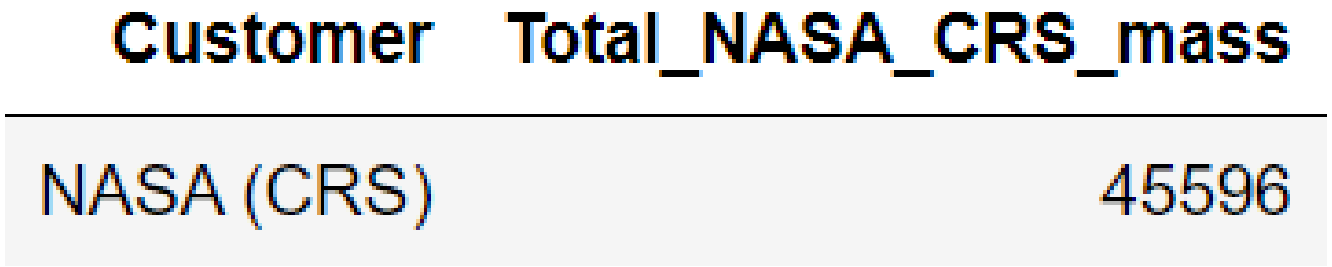


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Data: Shows early launches (2010-2012) from CCAFS LC-40, with initial failures but subsequent successes.

◦ Source: Github

Total Payload Mass

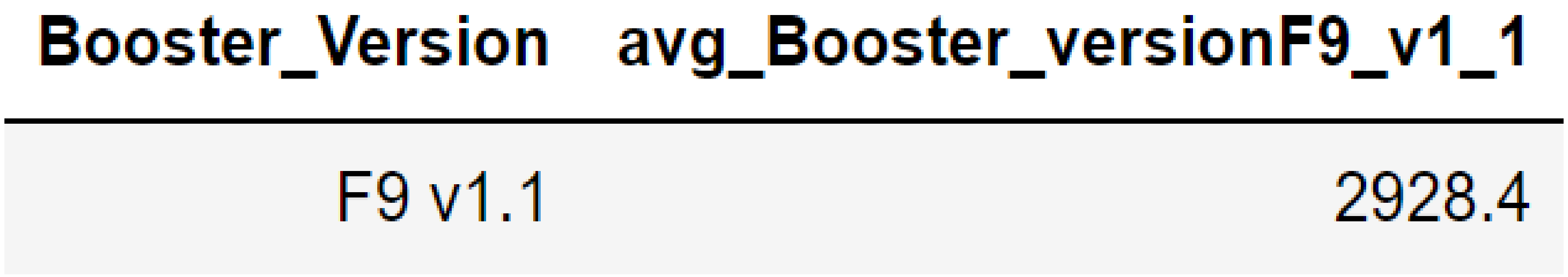


Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

- Data: NASA (CRS) missions accounted for a total of 45,596 kg of payload.

◦ Source: Github

Average Payload Mass by F9 v1.1

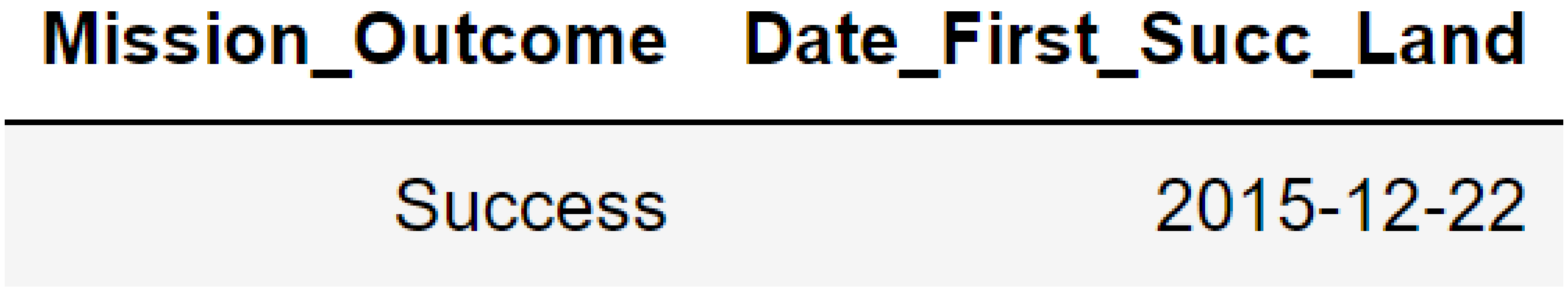


Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

- Data: The average payload mass for Booster Version F9 v1.1 was 2928.4 kg.

◦ Source: Github

First Successful Ground Landing Date

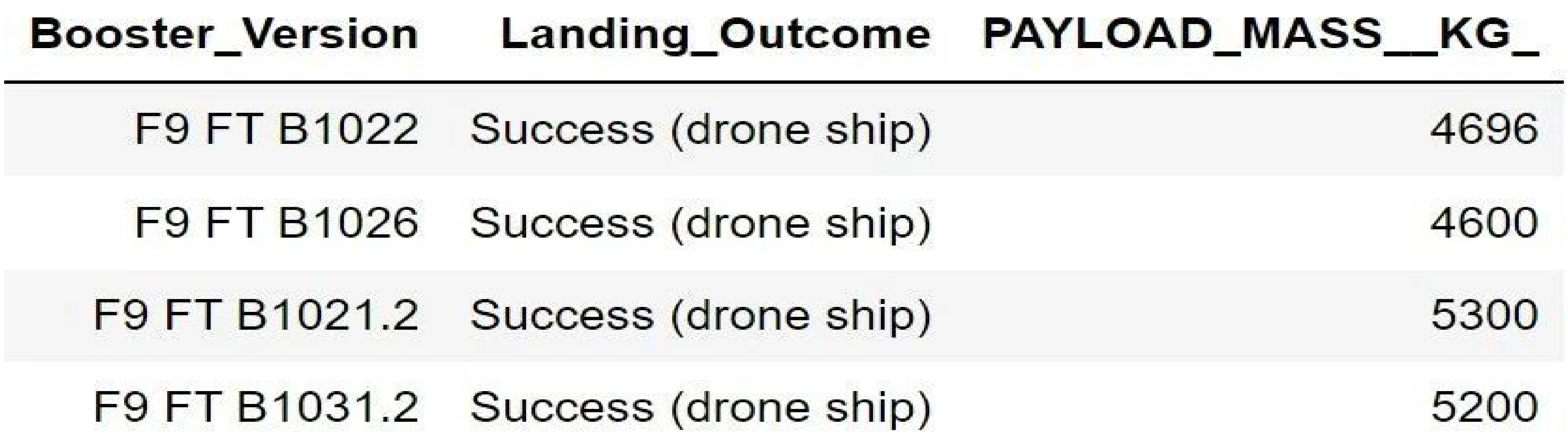


Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

- Milestone: The first successful landing outcome on a ground pad was recorded on 2015-12-22.

◦ Source: Github

Successful Drone Ship Landing With Payload between 4000 and 6000

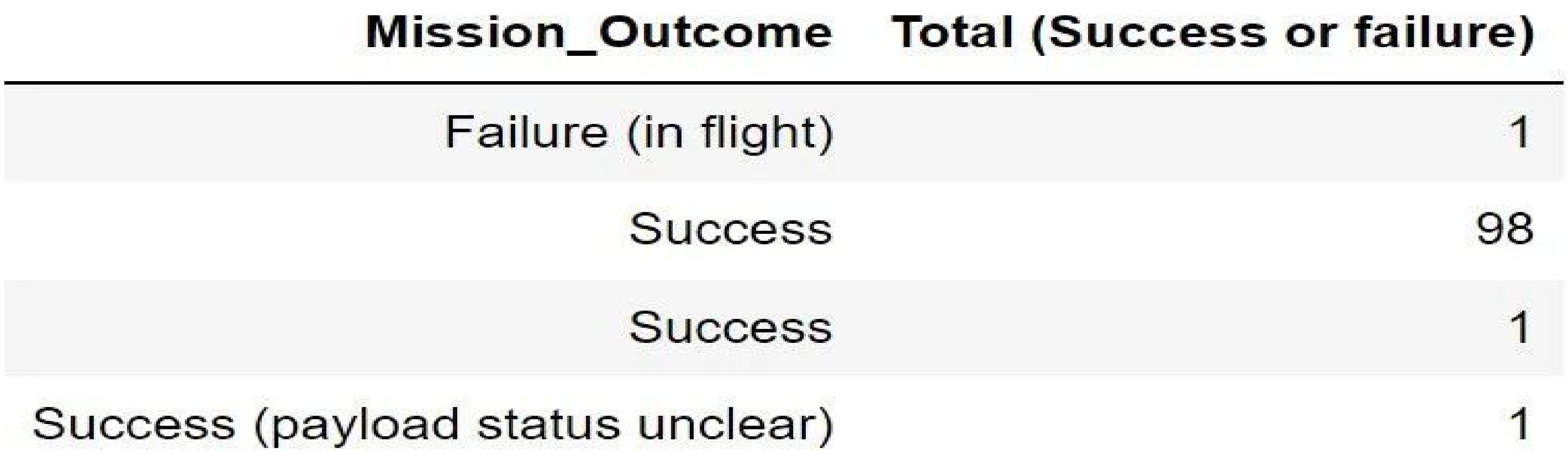


Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- Finding: Boosters F9 FT B1022, B1026, B1021.2, and B1031.2 successfully landed on drone ships with payloads between 4000 and 6000 kg.

  - Source: Github

Total Number of Successful and Failure Mission Outcomes

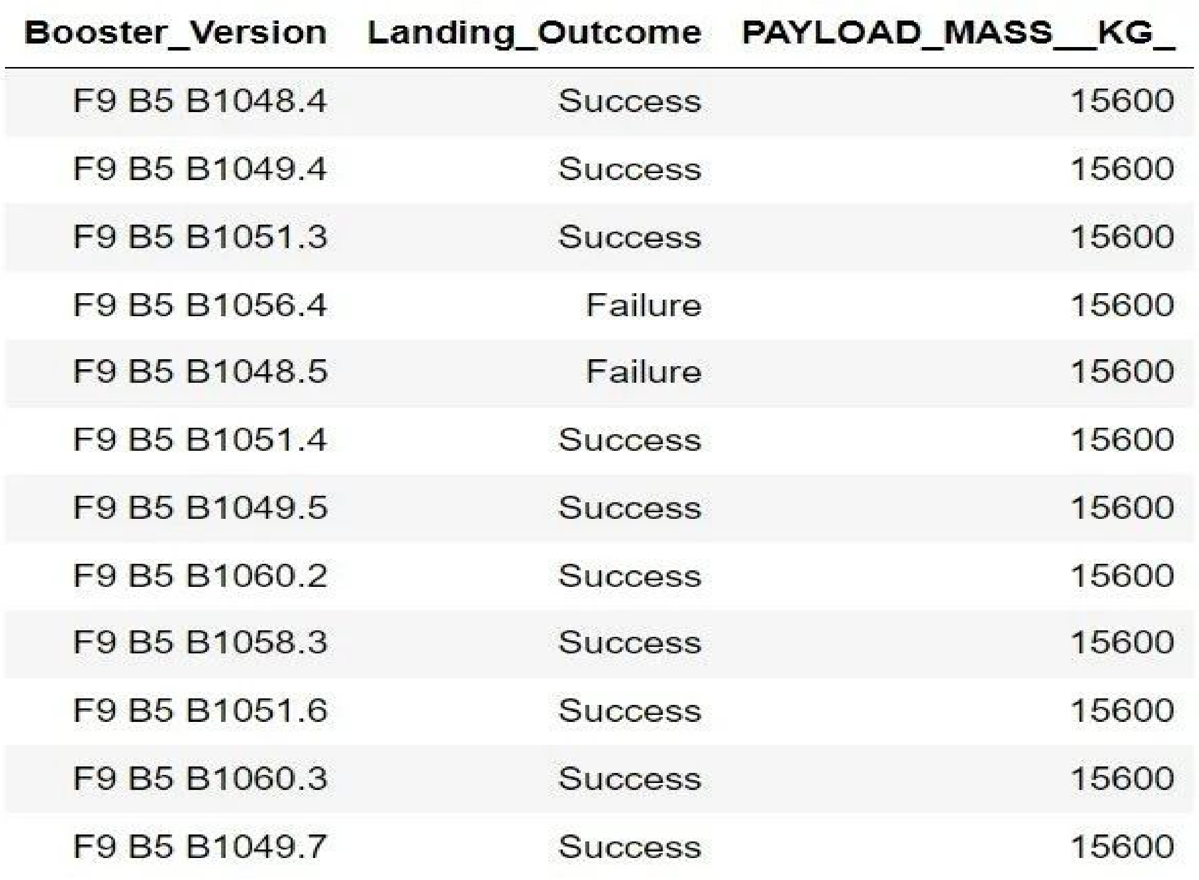


Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Data: Out of 99 total outcomes, 98 were successful landings, 1 was a failure (in flight), and 1 had an unclear payload status.

◦ Source: Github

Boosters Carried Maximum Payload

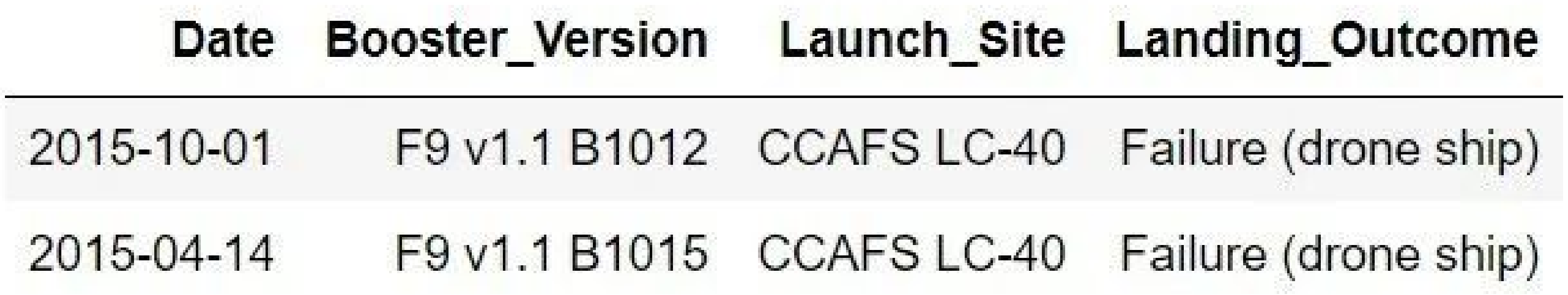


Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Success	15600
F9 B5 B1049.4	Success	15600
F9 B5 B1051.3	Success	15600
F9 B5 B1056.4	Failure	15600
F9 B5 B1048.5	Failure	15600
F9 B5 B1051.4	Success	15600
F9 B5 B1049.5	Success	15600
F9 B5 B1060.2	Success	15600
F9 B5 B1058.3	Success	15600
F9 B5 B1051.6	Success	15600
F9 B5 B1060.3	Success	15600
F9 B5 B1049.7	Success	15600

- Data: Booster version F9 B5 consistently carried the maximum payload mass of 15,600 kg for multiple successful and some failed flights.

◦ Source: Github

2015 Launch Records

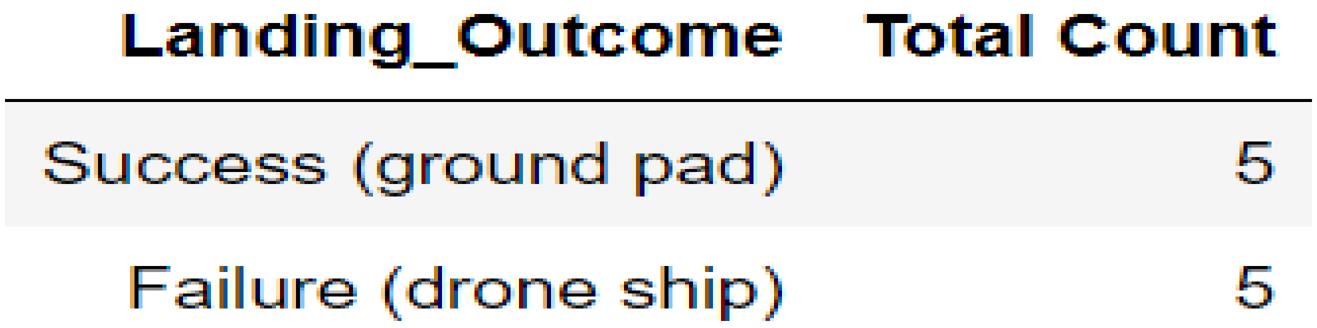


Date	Booster_Version	Launch_Site	Landing_Outcome
2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Finding: In 2015, there were two documented failure outcomes on drone ships: F9 v1.1 B1012 and F9 v1.1 B1015, both from CCAFS LC-40.

◦ Source: Github

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



Landing_Outcome	Total Count
Success (ground pad)	5
Failure (drone ship)	5

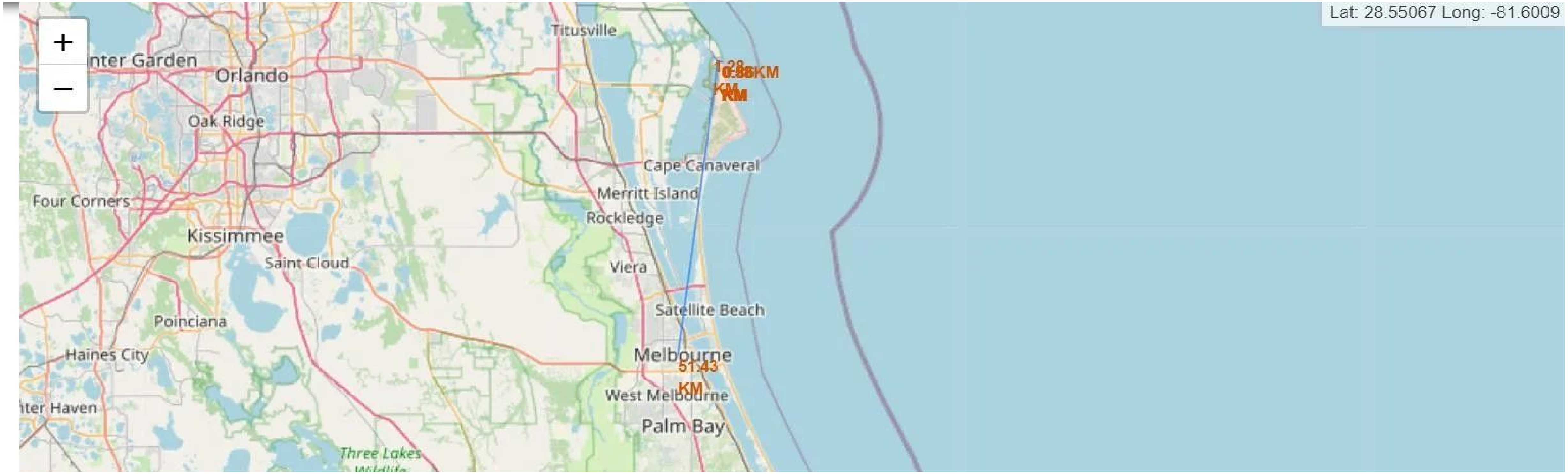
- Observation: Within this period, there were 5 successful ground pad landings and 5 failures on drone ships.

◦ Source: Github

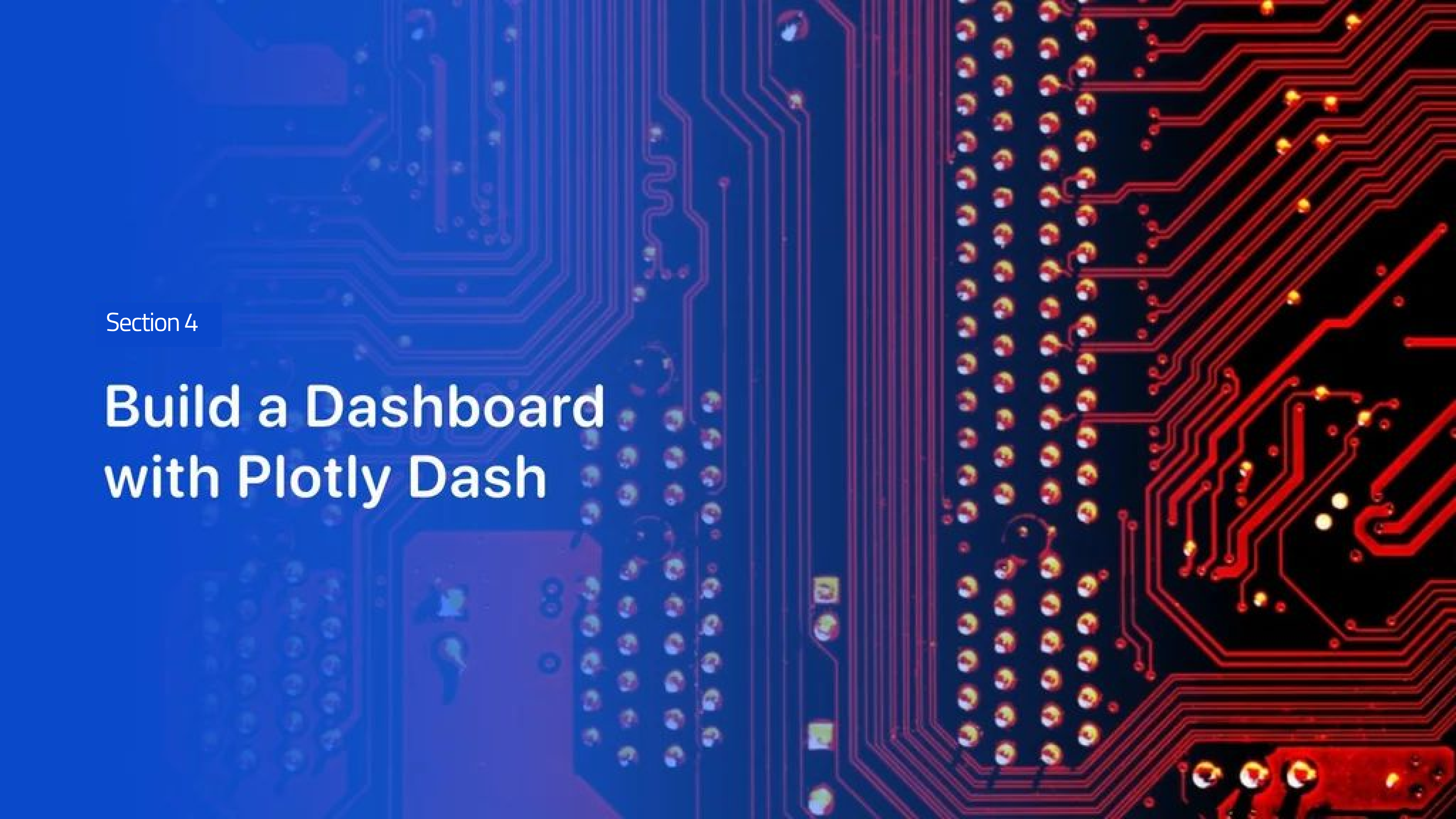
Section 3

Launch Sites Proximities Analysis

Proximities



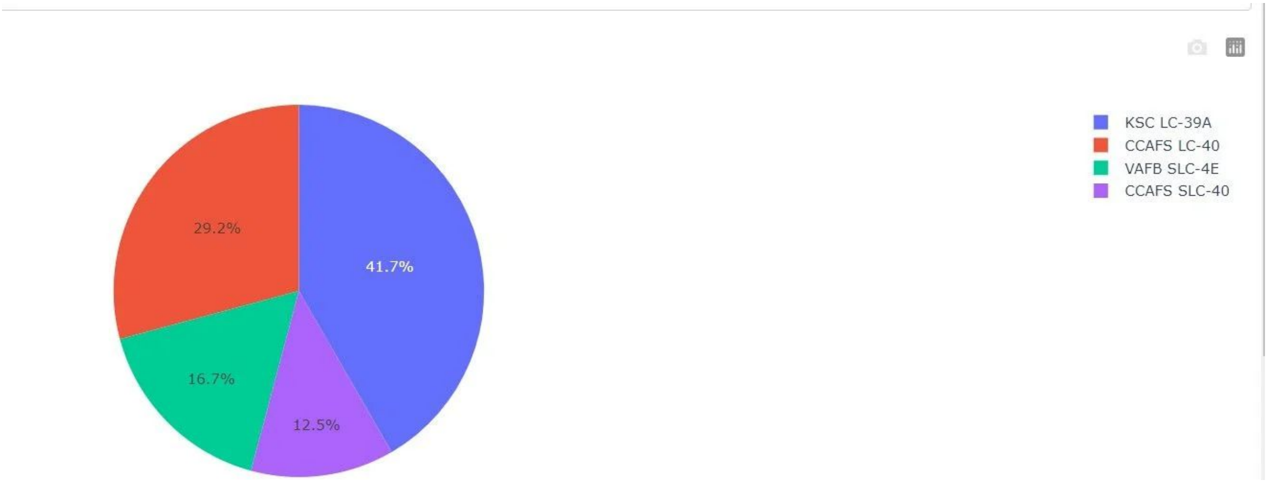
- Observation: Launch sites are strategically located very close to coastlines (~1 km) and railways for safety and logistics, while maintaining a safe buffer distance (~50 km) from populated cities like Melbourne.



Section 4

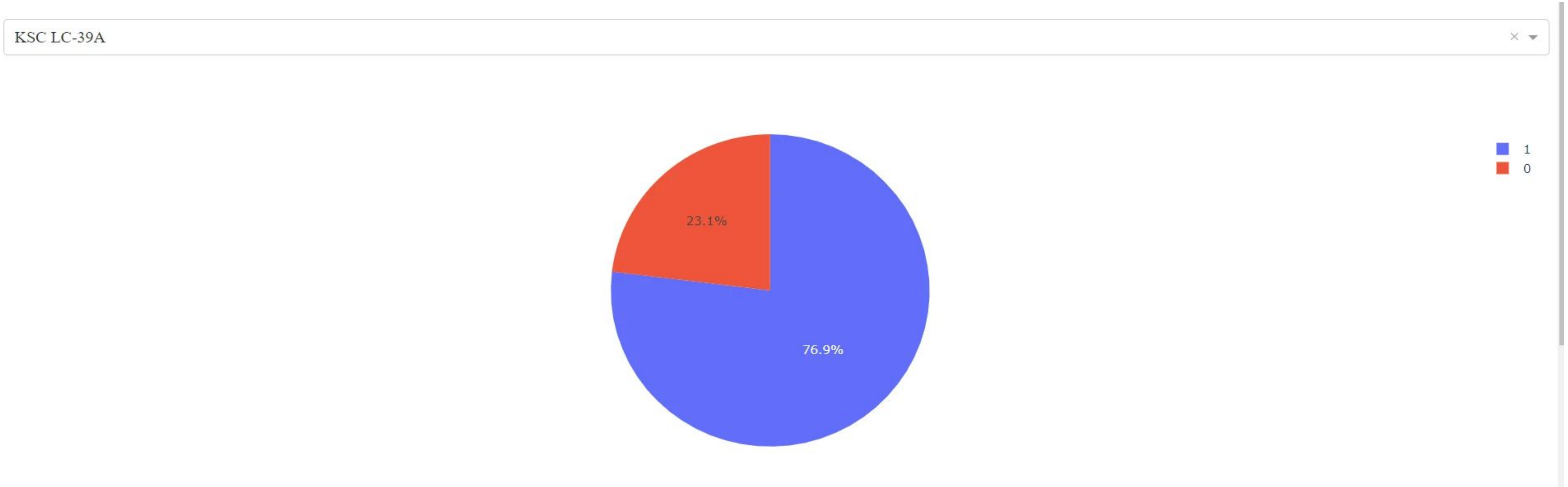
Build a Dashboard with Plotly Dash

Pie chart of all the launch sites



- Observation: KSC LC-39A is the most active launch site, accounting for 41.7% of all total launches, followed by CCAFS LC-40 with 29.2%.

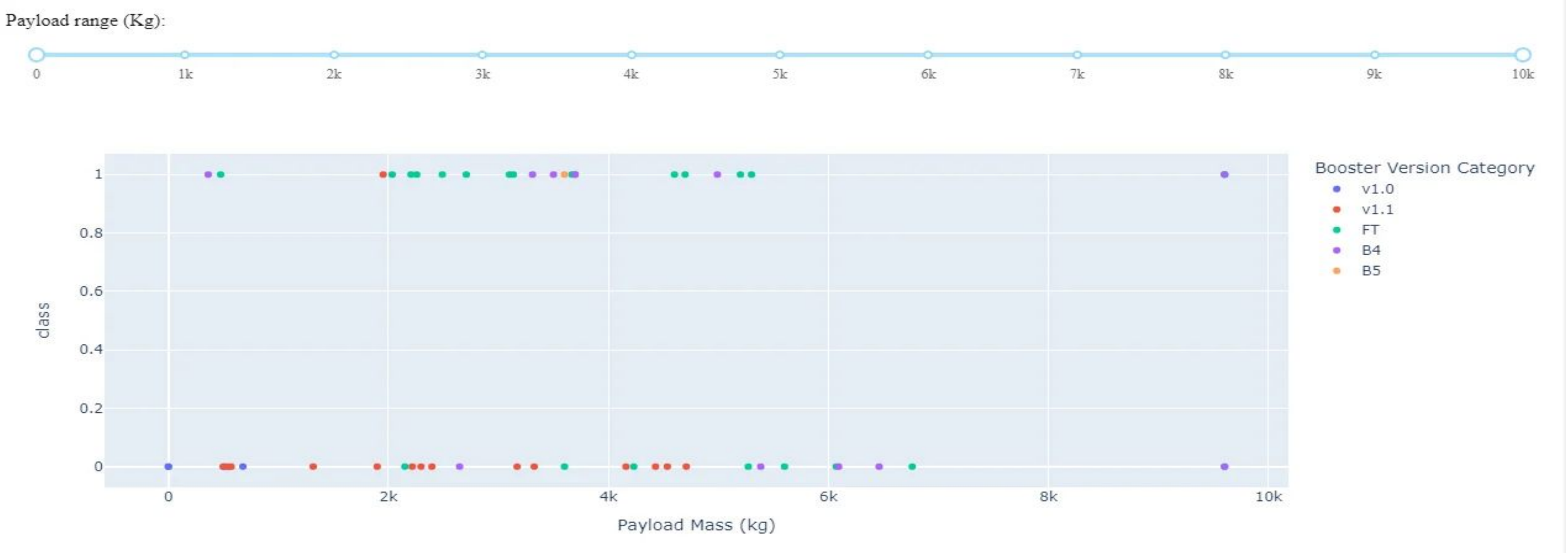
Pie chart of KSC LC 39A



KSC LC-39A

- Observation: KSC LC-39A demonstrates a strong reliability record, with a 76.9% success rate for Falcon 9 landings.

Payload vs. Launch Outcome scatter

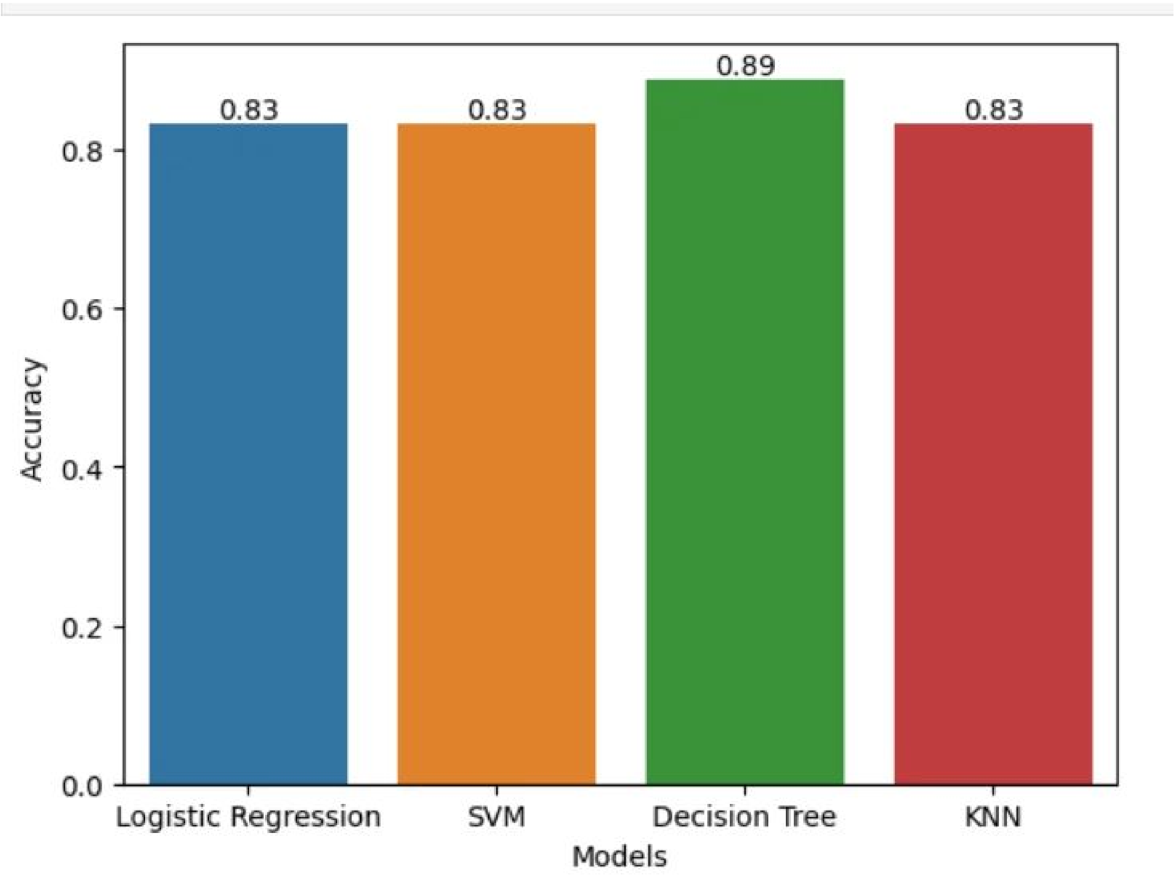


- Observation: There is a clear correlation between booster version and performance; the FT booster version shows consistently high success rates across all payload masses, whereas the older v1.1 version suffered significantly more failures.

Section 5

Predictive Analysis (Classification)

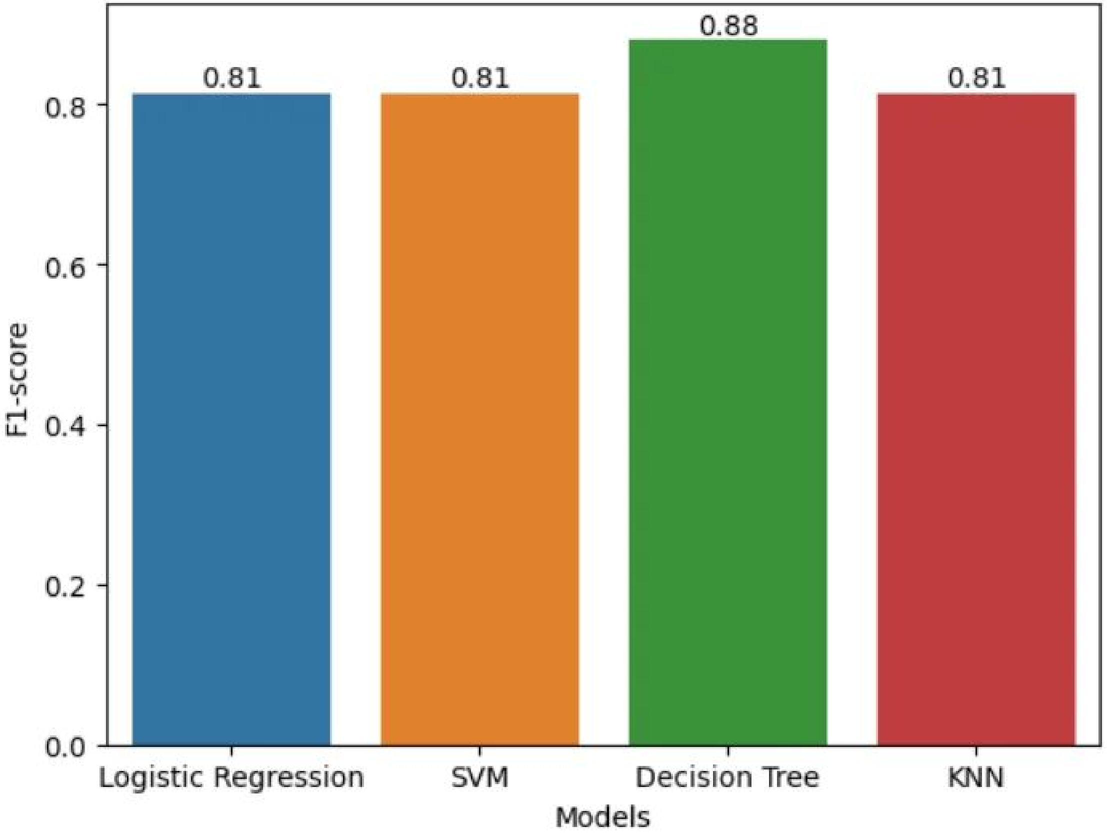
Classification Accuracy



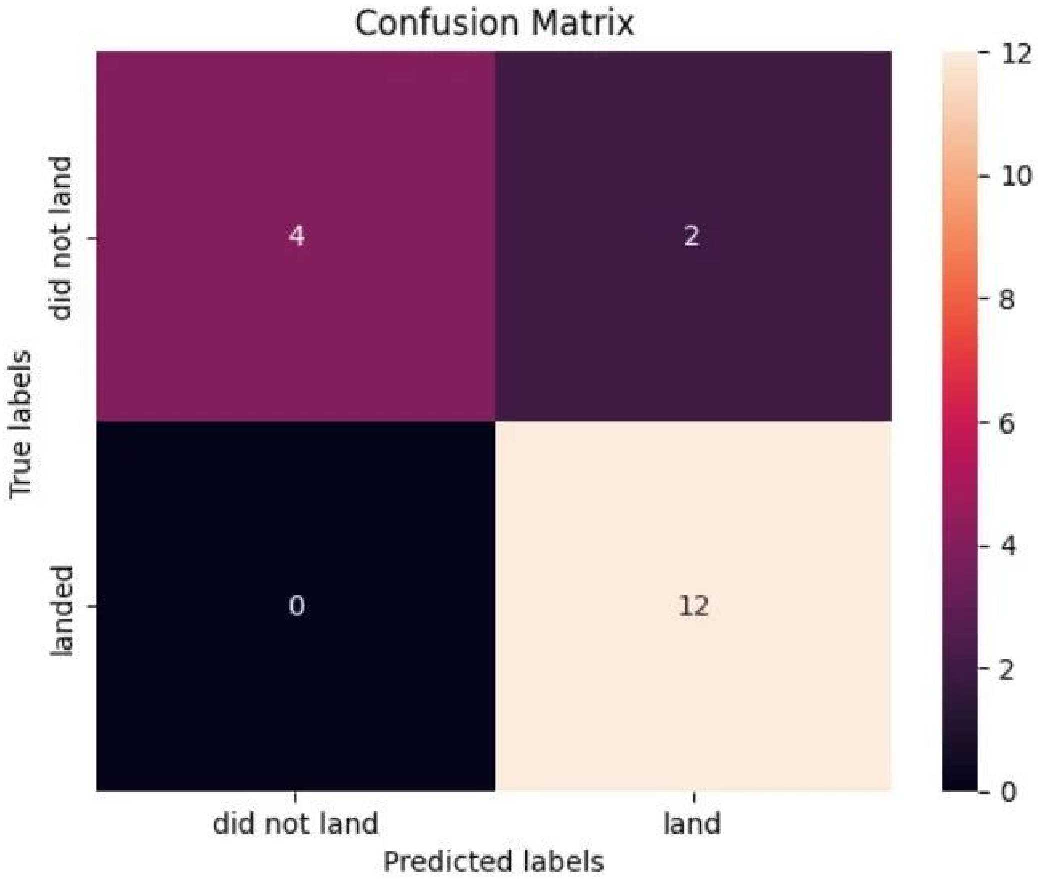
- Observation: All four models performed comparably well, but the Decision Tree classifier achieved the highest accuracy at 88.88% on the test data.

Classification F1-score

- Observation: The Decision Tree model also led in precision and recall balance, achieving the highest F1-score of 0.88 (88.88%).



Confusion Matrix



- Observation: The confusion matrix reveals high predictive power, with the model correctly identifying 12 successful landings (True Positives) and 4 failed landings (True Negatives), with very few errors.

Conclusions

- Experience Drives Success: The data confirms a strong positive correlation between the number of flights and landing success rates, particularly at CCAFS LC-40, validating the "learning curve" in reusable rocketry.
- Orbit Reliability: Specific high-altitude orbits (ES-L1, SSO, HEO, GEO) have a flawless 100% success record, suggesting these mission profiles are well-optimized.
- Temporal Improvement: SpaceX has demonstrated a consistent year-over-year increase in landing reliability from 2013 through 2020.
- Predictive Capability: With a validated accuracy of 88.89%, our Decision Tree model serves as a robust tool for predicting future landing outcomes, providing SpaceY with a competitive edge in risk assessment and cost estimation.

Appendix

Project Assets & Resources:

- Source Code: Access the complete Python Jupyter Notebooks for Data Collection, Wrangling, EDA, and Machine Learning on GitHub.
- Database Scripts: View the full library of SQL queries used for performance metrics extraction here.
- Dashboard: Explore the interactive Plotly Dash application source code here.
- Datasets: Download the cleaned and processed datasets (CSV) used for training and testing here.

Thank you!