IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mae Ezra T. Abundo

September 2, 2025

Outline

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

Executive Summary

This project followed a comprehensive data science pipeline to predict the success of Falcon 9 first-stage landings.

Summary of methodologies

- Data Collection: Collected SpaceX launch data from Wikipedia and the SpaceX API.
- Data Wrangling & EDA: Cleaned and prepped data. Used Folium to visualize launch sites and Dash/Plotly to build an interactive dashboard.
- Machine Learning: Developed a model to predict successful first-stage landings.

Summary of all results

- Findings: Discovered correlations between launch success and factors like payload mass and launch site.
- Best Model: The K-Nearest Neighbors model was identified as the best performer for predicting landing success.
- Tool: The interactive Dash dashboard allows users to explore launch success rates based on launch site and payload mass.

Introduction

SpaceX has revolutionized the aerospace industry by pioneering reusable rocket technology, significantly reducing the cost of space travel. The Falcon 9 rocket, in particular, is a game-changer because its first stage can be recovered after a mission, allowing for a much lower price per launch compared to competitors who use expendable rockets. Predicting the success of these landings is therefore a critical task, as it directly impacts the financial viability and reliability of future missions. This project leverages publicly available data to analyze and model the factors influencing a successful landing.

The primary objective of this project is to build a predictive model that can accurately determine whether a Falcon 9 first stage will land successfully. To achieve this, we will explore several key questions:

- How do different variables, such as payload mass, launch site, and orbit, affect the success rate of a landing?
- Have the success rates of Falcon 9 landings improved over time as the company has gained more experience and refined its technology?
- Which machine learning algorithm is best suited for this binary classification problem?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data sets were collected using RESTful API calls to the SpaceX public API endpoints. The collection process involved systematic retrieval of rocket launch data, mission information, and spacecraft details through structured HTTP requests.
- Key Phases:
  1. API Endpoint Identification - Identified SpaceX API endpoints for launch and rocket data
  2. Data Retrieval Implementation - Implemented HTTP GET requests using Python requests library
  3. Data Processing and Storage - Parsed JSON responses and structured into pandas DataFrames and applied data validation and cleaning procedures

Data Collection – SpaceX API

GitHub URL: https://github.com/maeezra/IBMCapstoneProject/blob/cac3f74ad288f1075f0ecf0bad4fe0ac9eac4613/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection flowchart

[Start] → [Initialize API Connection] → [Send GET Request to SpaceX API]

↓

[Receive JSON Response] → [Parse JSON Data] → [Validate Data Quality]

↓

[Store in DataFrame] → [End: Apply Data Cleaning]

Data Collection - Scraping

- The collection process involved systematic retrieval of launch data, mission information, and spacecraft details through structured HTML parsing.
- Key Phases:
  - Target Identification, Data Extraction Implementation, Data Processing and Storage

GitHub URL: https://github.com/maeezra/IBMCapstoneProject/blob/6f82722c46ba82dc2117fbfd8d285b06ec19c62b/jupyter-labs-webscraping.ipynb

Web Scraping flowchart

[Start] → [Initialize Web Scraping Environment] → [Send GET Request to Wikipedia URL]

↓

[Receive HTML Response] → [Parse HTML with BeautifulSoup] → [Locate Launch Data Tables]

↓

[Extract Table Headers] → [Parse Table Rows and Cells] → [Apply Data Cleaning Functions]

↓

[Store in DataFrame] → [End: Validate Data Quality]

Data Wrangling

- The wrangling process involved comprehensive analysis of launch outcomes, mission parameters, and booster recovery data to create a machine learning-ready dataset.
- Key Phases:
  1. Data Quality Assessment - Identified missing values in dataset and performed data type analysis.
  2. Exploratory Data Analysis Implementation - Applied pandas value_counts() methodology to analyze launch site distribution and orbit type frequencies excluding transfer orbits.
  3. Landing Outcome Standardization - Transformed complex categorical outcomes into binary classification labels using conditional logic and set operations.

GitHub URL:
https://github.com/maeezra/IBMCapstoneProject/blob/8b0dd4eb278f5519b3178ca41bbbe9b9e84ccecb/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Flight Number & Payload Mass
  - Chart Used: Categorical Scatter Plot. This chart was used to explore the relationship between a launch's flight number, its payload mass, and the landing outcome.
- Launch Site Analysis
  - Chart Used: Bar Chart. A bar chart was used to summarize and compare the number of launches from different launch sites.
- Performance Over Time
  - Chart Used: Line Chart. A line chart was used to visualize trends over a continuous period, such as the success rate over time.

GitHub URL: https://github.com/maeezra/IBMCapstoneProject/blob/1c8c51412f1c051adc0179ce652b2d4f1f99c32f/edadataviz.ipynb

EDA with SQL

- Launch site analysis - Identified 4 unique SpaceX launch locations using DISTINCT queries
- Payload calculations - Computed total (45,596 kg) and average (2,928.4 kg) payload masses by customer and booster version
- Landing outcomes - Found first successful ground landing (Dec 22, 2015) and analyzed drone ship success rates
- Mission performance - Counted 98 successful vs 1 failed mission, identified boosters with maximum payload capacity
- Temporal patterns - Examined 2015 failure trends and ranked landing outcomes by frequency (2010-2017)
- Data filtering - Applied pattern matching, date ranges, and subqueries to extract specific mission characteristics

GitHub URL:
https://github.com/maeezra/IBMCapstoneProject/blob/105937e9506f4d6d47f9c59d4f291a0c7b3f761e/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following objects were created and added to the Folium map:

- folium.Map: The base map was created, initially centered on the NASA Johnson Space Center.
- folium.Circle: A blue circle with a 1000-meter radius was added for the NASA Johnson Space Center and circles were added for each launch site to visualize their locations.
- folium.Marker: Markers with DivIcon were created for the NASA Johnson Space Center and each launch site to pinpoint their locations and provide a text label of their names.

These were added to:

- Visually represent the locations of SpaceX launch sites and related infrastructure in an interactive manner.
- Allow for the calculation of distances between launch sites and other points of interest.
- Provide a visual context for analyzing the proximity of launch sites to the coast and cities.

GitHub URL: https://github.com/maeezra/IBMCapstoneProject/blob/30cc9d34784718483507415386e3b3f140ac9572/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- These are the components added in the dashboard:
  - Launch Site Dropdown + Payload Range Slider for filtering data
  - Pie Chart: Success rates by site (overview) or success/failure ratio (specific site)
  - Scatter Plot: Payload mass vs. launch success, color-coded by booster version
- These components were added for:
  - Mission Planning: Compare site performance to select optimal launch location
  - Risk Assessment: Analyze payload weight limits and success probability correlations
  - Performance Tracking: Visualize technology improvements and historical trends across sites
  - Interactive Analysis: Enable multi-dimensional filtering for complex mission requirement scenarios

GitHub URL:
https://github.com/maeezra/IBMCapstoneProject/blob/6318e193306bbae5da1047485a052512b8b2eb5f/spacex-dash-app.py

Predictive Analysis (Classification)

- To predict whether the first stage of a Falcon 9 rocket would land, a machine learning model was developed and evaluated using the following steps:
  - Data Preparation: The dataset was split into training and testing sets, with 18 samples reserved for testing.
  - Model Building and Tuning: Three classification models were created: Logistic Regression, Support Vector Machine (SVM), and a Decision Tree. Each model's hyperparameters were optimized using GridSearchCV.
  - Model Evaluation: The accuracy of each model was calculated using the test data. Their performance was also visualized with a confusion matrix.
  - Best Model Selection: The final step involved comparing the accuracy scores of all models to find the one that performed best.

GitHub URL:
https://github.com/maeezra/IBMCapstoneProject/blob/a1e2be3d909d33f6a5b829dbe84de4d885c55fa1/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

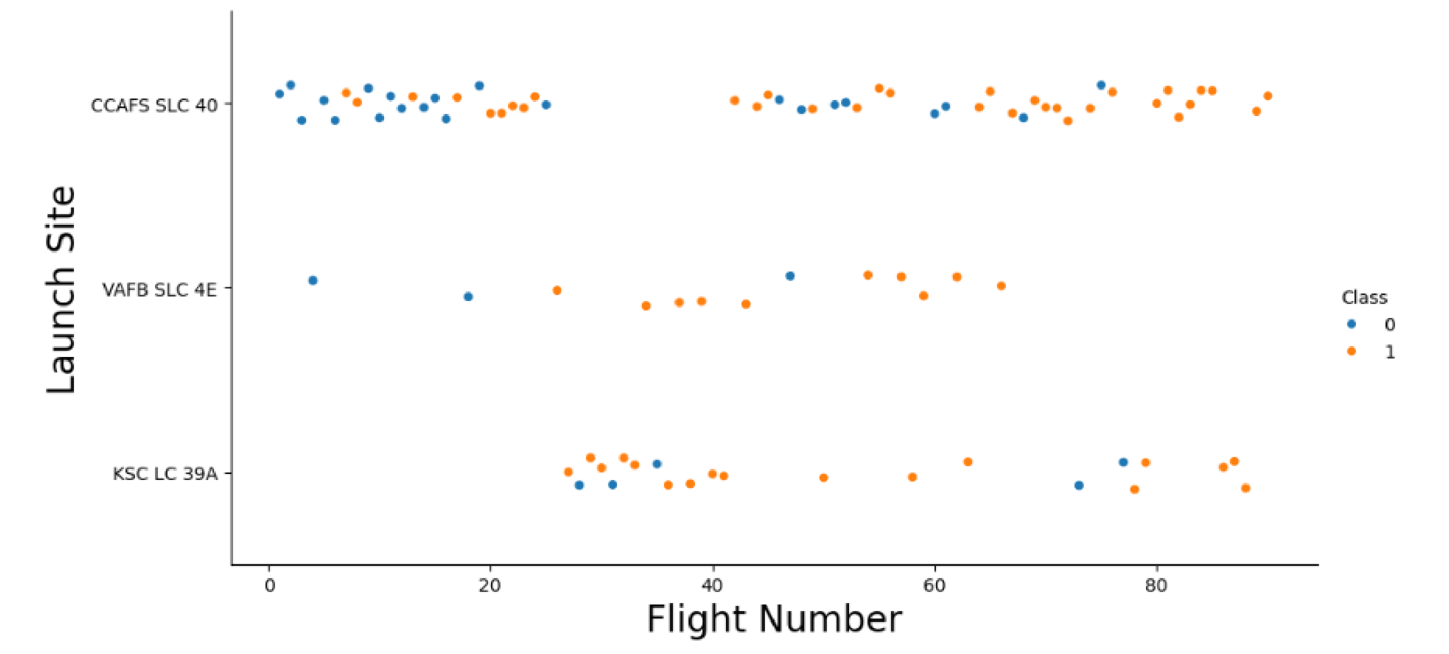
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

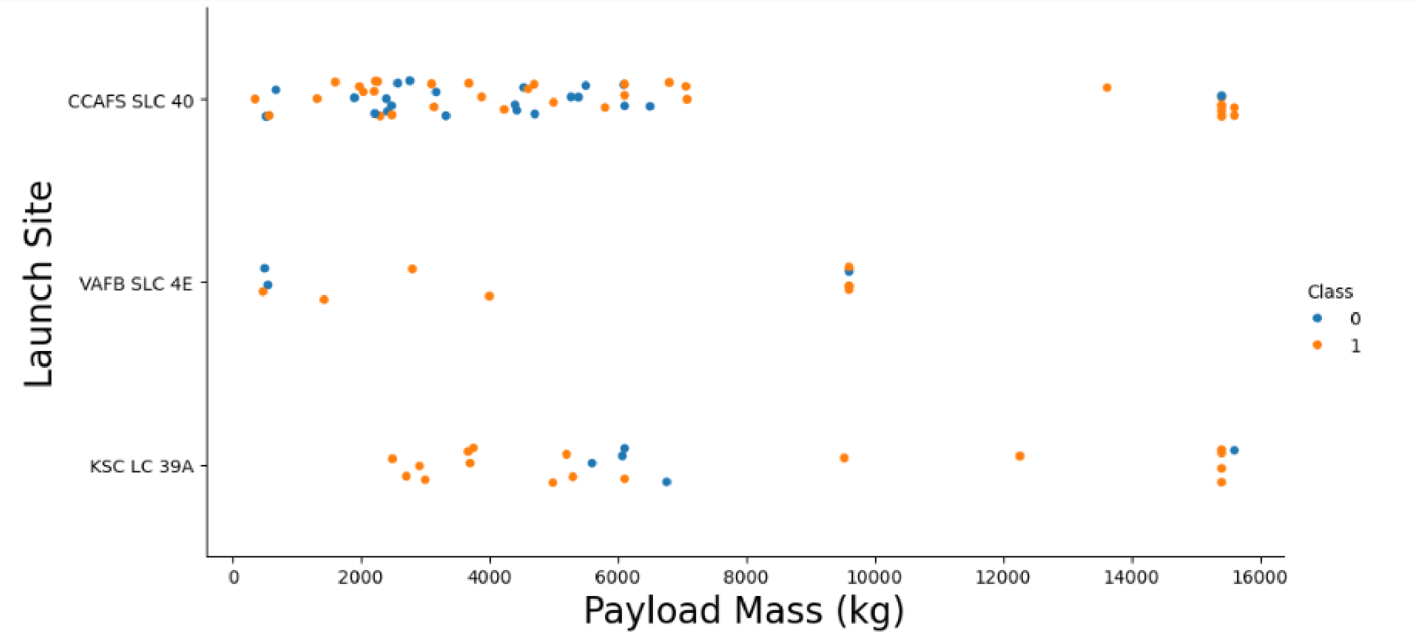
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 improvement: Early flights show mostly failed landings (class 0), but success rates improve significantly in later flight numbers, indicating learning curve effects.
- KSC LC 39A high success: This site shows predominantly successful landings (class 1) and is used mainly for later flights, suggesting it's reserved for missions with proven landing technology.
- VAFB SLC 4E mixed performance: This site shows varied success rates throughout the flight sequence, possibly due to the challenging nature of polar/sun-synchronous orbit missions that typically launch from this location.

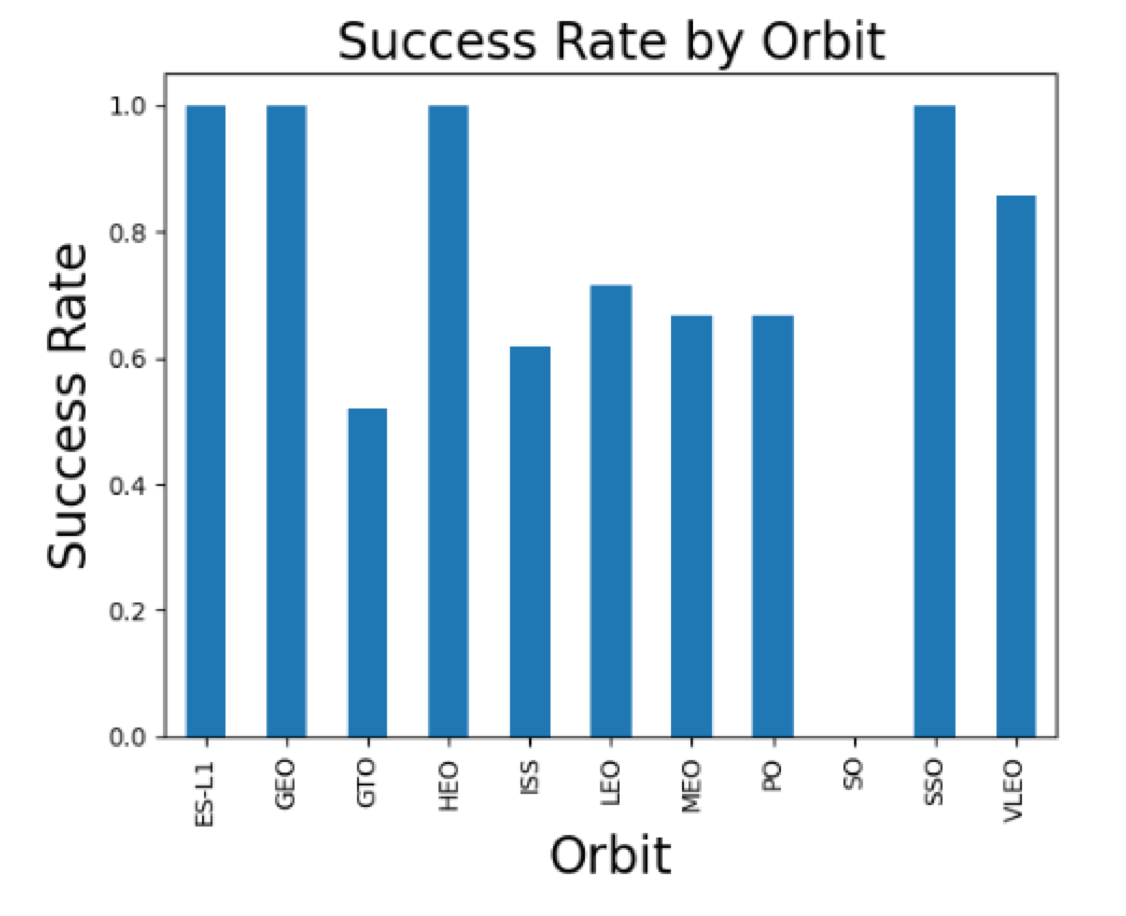
Payload vs. Launch Site



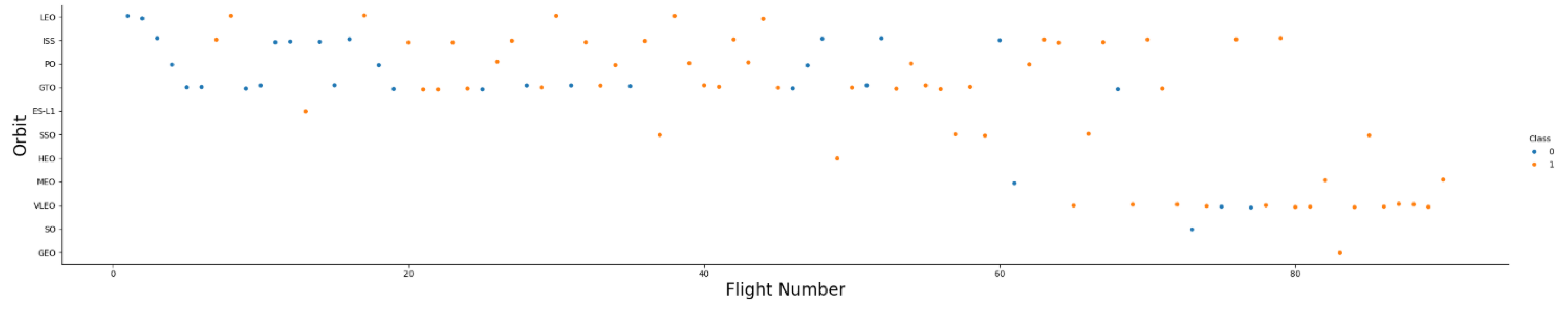
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Perfect success orbits: ES-L1, GEO, HEO, and SSO missions achieve 100% landing success rates, indicating these orbit types are most compatible with successful booster recovery.
- Challenging orbit identified: GTO missions show the lowest success rate (~52%), suggesting geostationary transfer orbits require more fuel, leaving less margin for successful landing burns.
- Performance variation: Success rates vary significantly across orbit types (52%-100%), demonstrating that mission trajectory and fuel requirements directly impact landing feasibility.

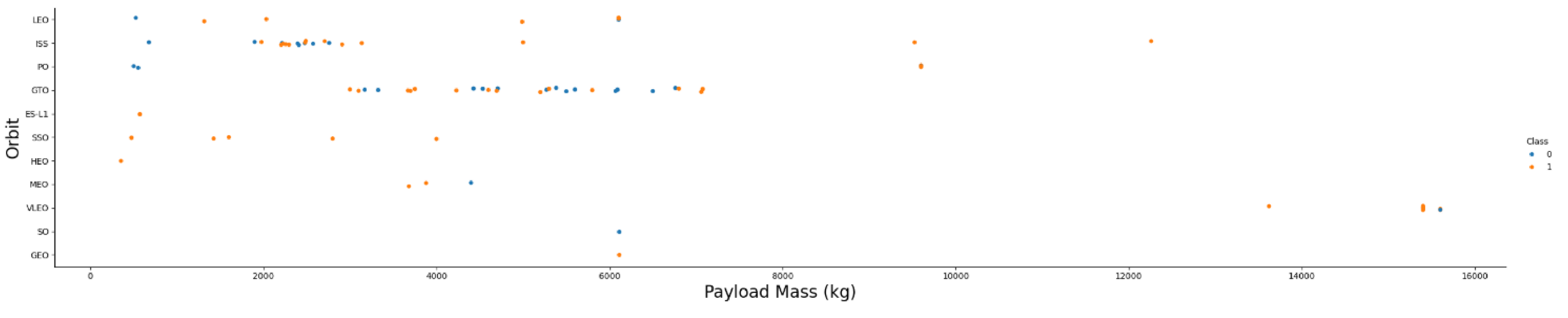


Flight Number vs. Orbit Type



- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

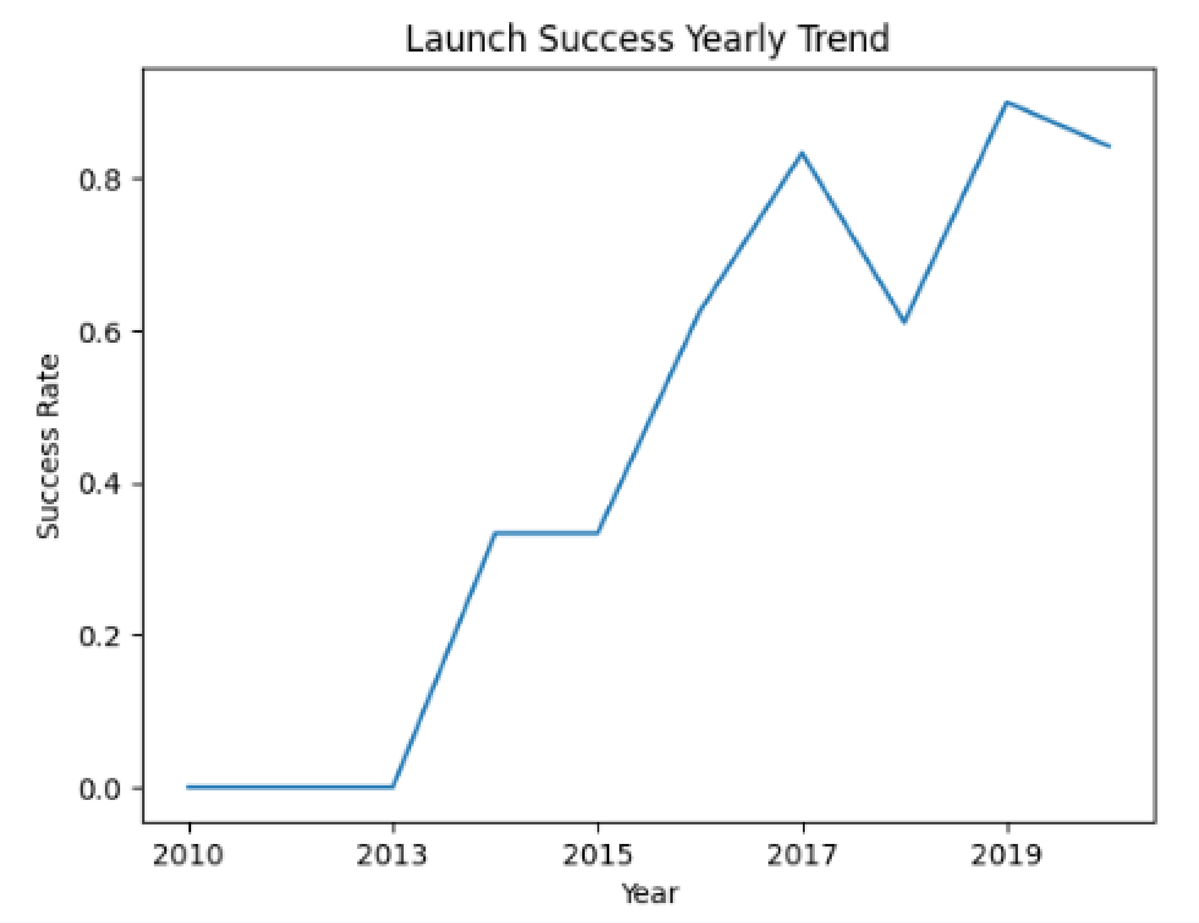
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

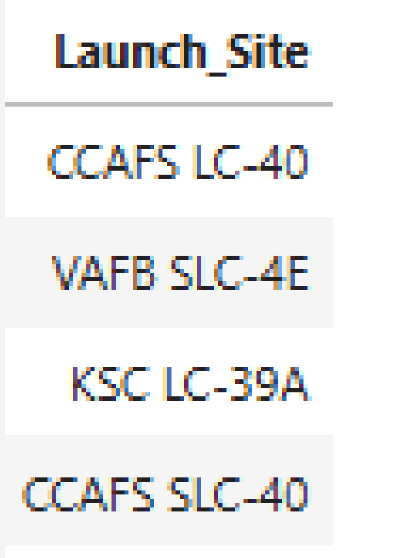
- Success rate since 2013 kept increasing till 2020



All Launch Site Names

- Names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

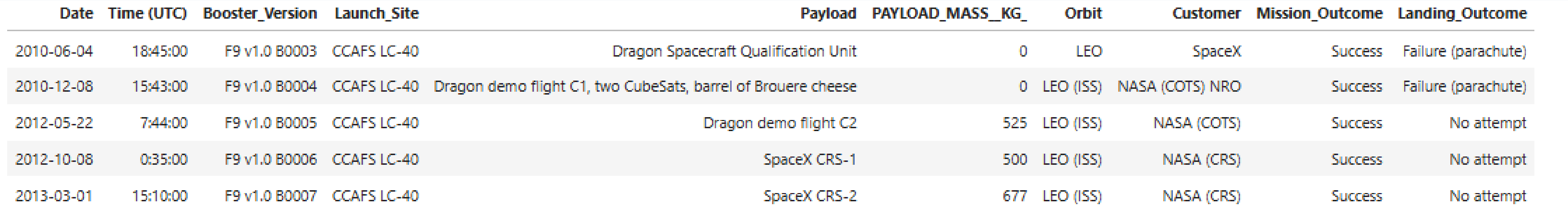


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;

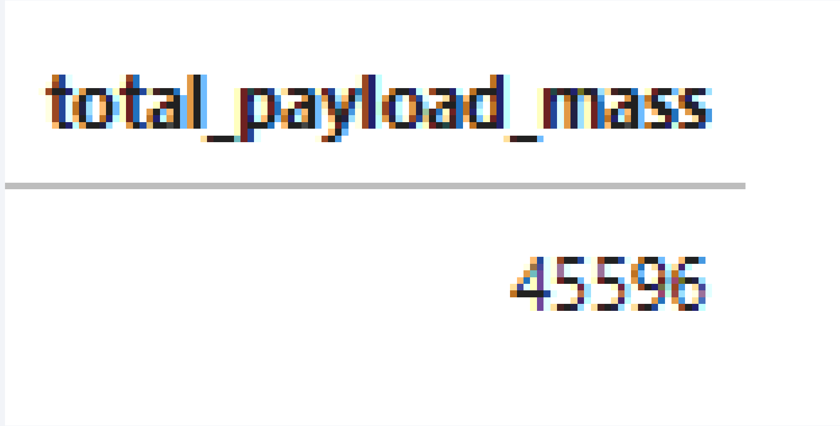


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("Payload_Mass__kg_") AS total_payload_mass FROM
SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

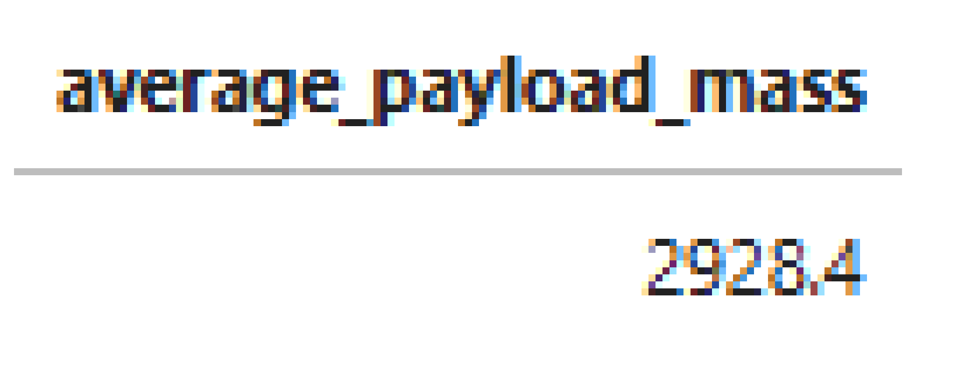


total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG("Payload_Mass__kg_") AS average_payload_mass FROM
SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1’;
```

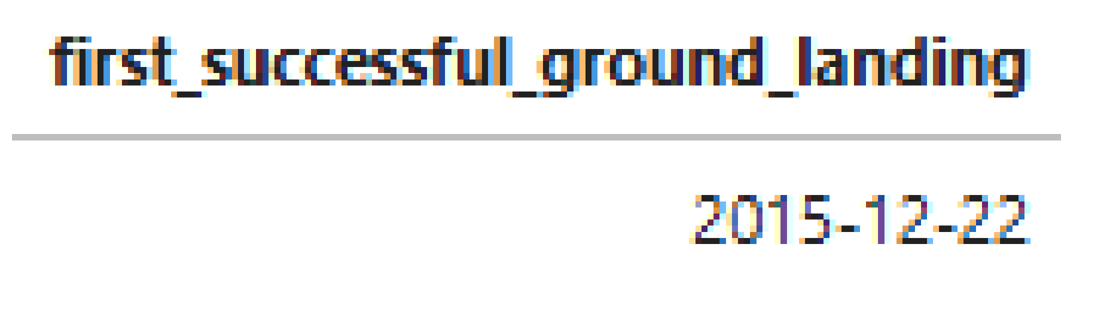


average_payload_mass
2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

```
%sql SELECT MIN("Date") AS first_successful_ground_landing FROM
SPACEXTABLE WHERE "Landing_Outcome" LIKE '%ground pad%' AND
"Landing_Outcome" LIKE '%Success%’;
```

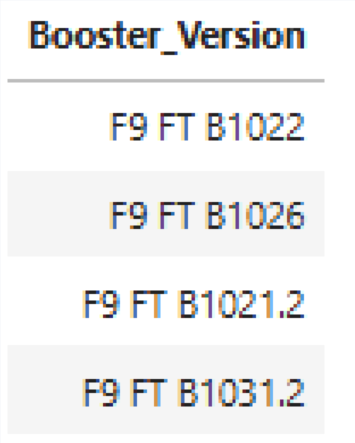


first_successful_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE
'%drone ship%' AND "Landing_Outcome" LIKE '%Success%' AND "Payload_Mass__kg_" > 4000
AND "Payload_Mass__kg_" < 6000;
```

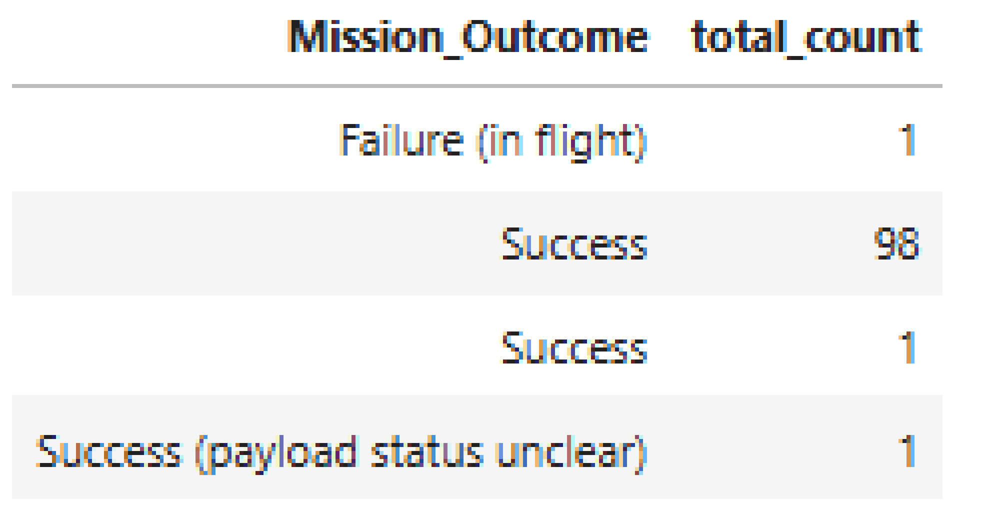


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

%sql SELECT "Mission_Outcome", COUNT(*) AS total_count FROM SPACEXTABLE GROUP BY "Mission_Outcome"; short explanation here

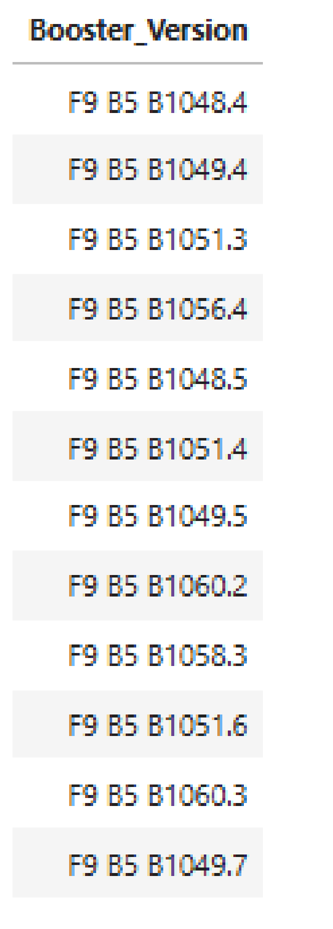


Mission_Outcome	total_count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);

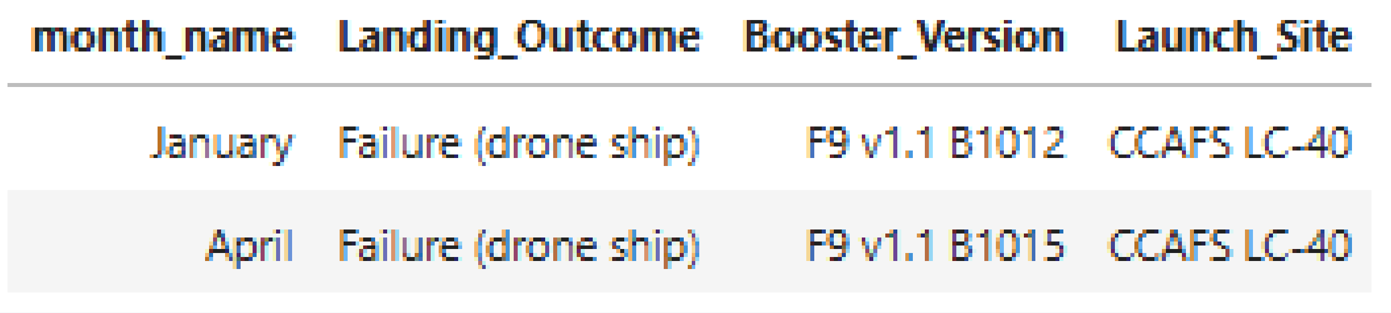


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT CASE substr("Date", 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03'
THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May' WHEN '06' THEN 'June' WHEN '07' THEN 'July'
WHEN '08' THEN 'August' WHEN '09' THEN 'September' WHEN '10' THEN 'October' WHEN '11' THEN
'November' WHEN '12' THEN 'December' END AS month_name, "Landing_Outcome", "Booster_Version",
"Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%drone ship%' AND
"Landing_Outcome" LIKE '%Failure%' AND substr("Date", 0, 5) = '2015';
```

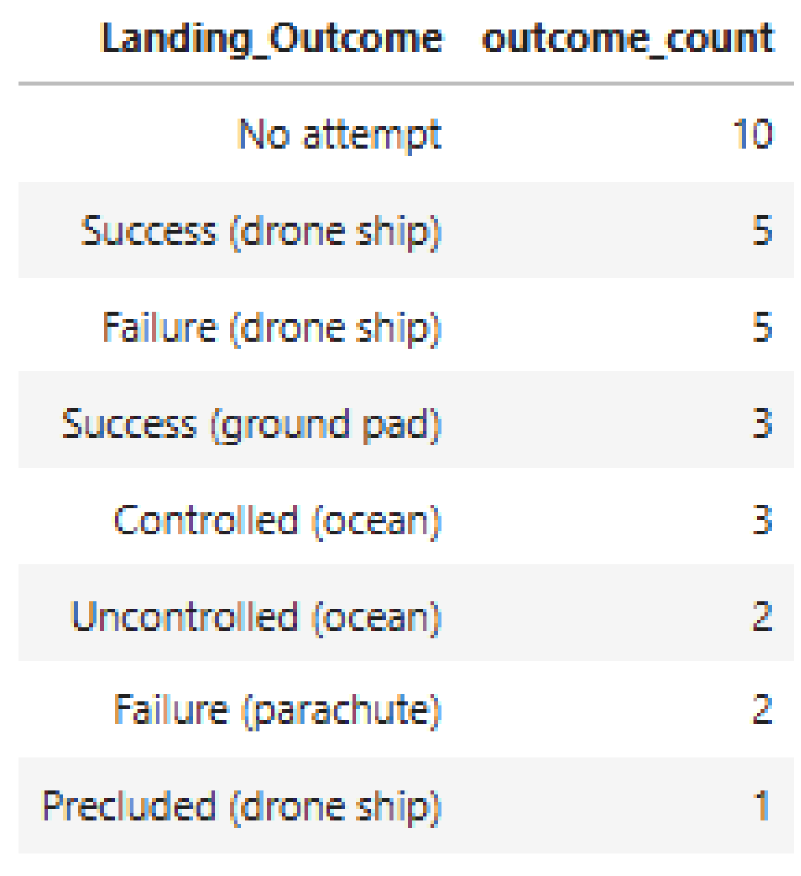


month_name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count 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

```
%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 COUNT(*) DESC;
```



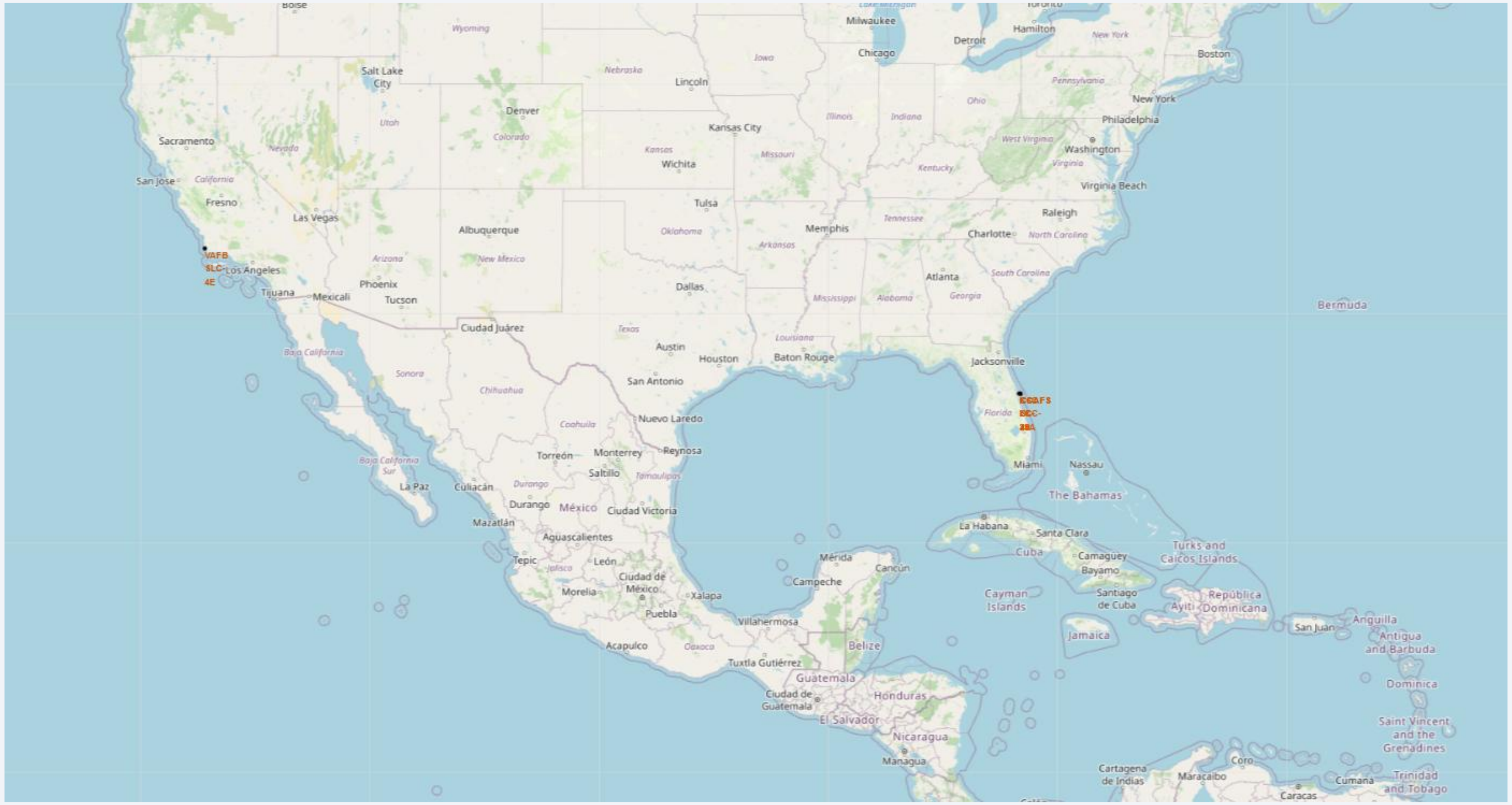
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

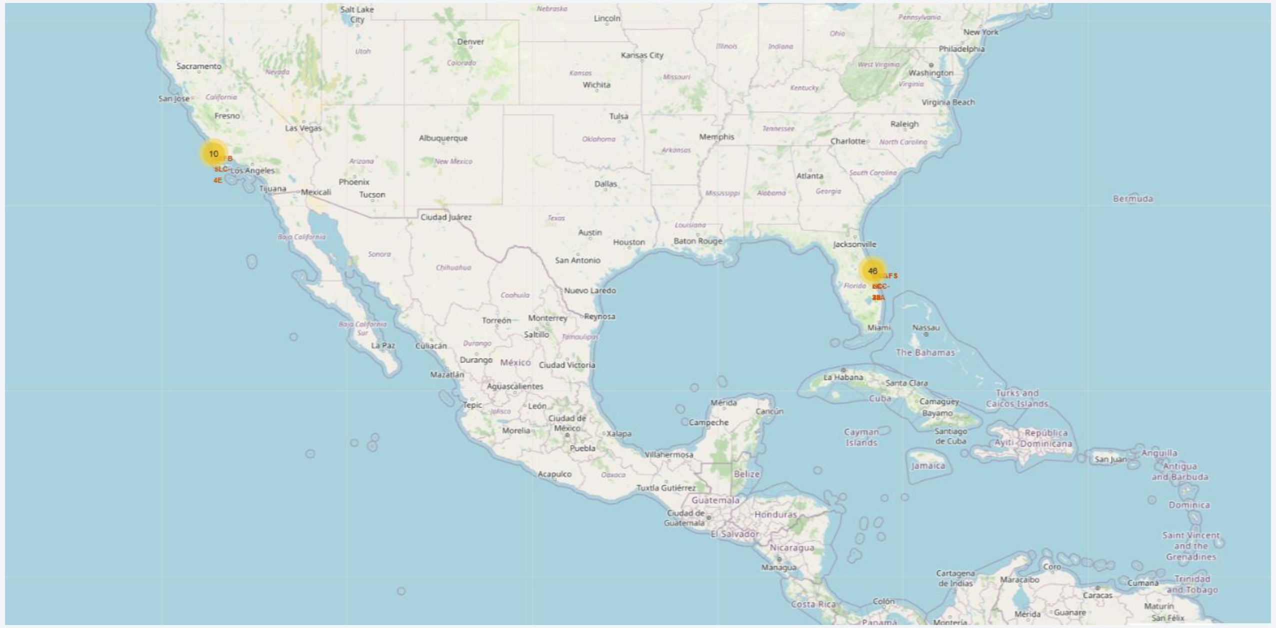
Launch Sites Proximities Analysis

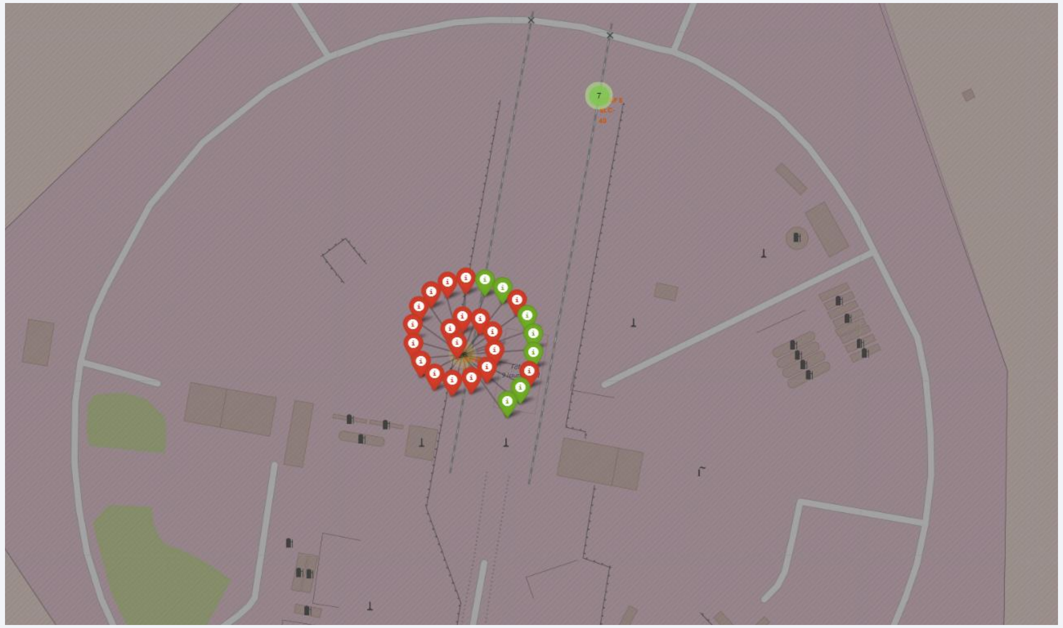
Map with Marked Launch Sites

- Not all launch sites are in proximity to the Equator. The launch site in California is located much farther north.
- All the launch sites shown on the map are in very close proximity to the coast.



Map with Color-labeled Launch Outcomes

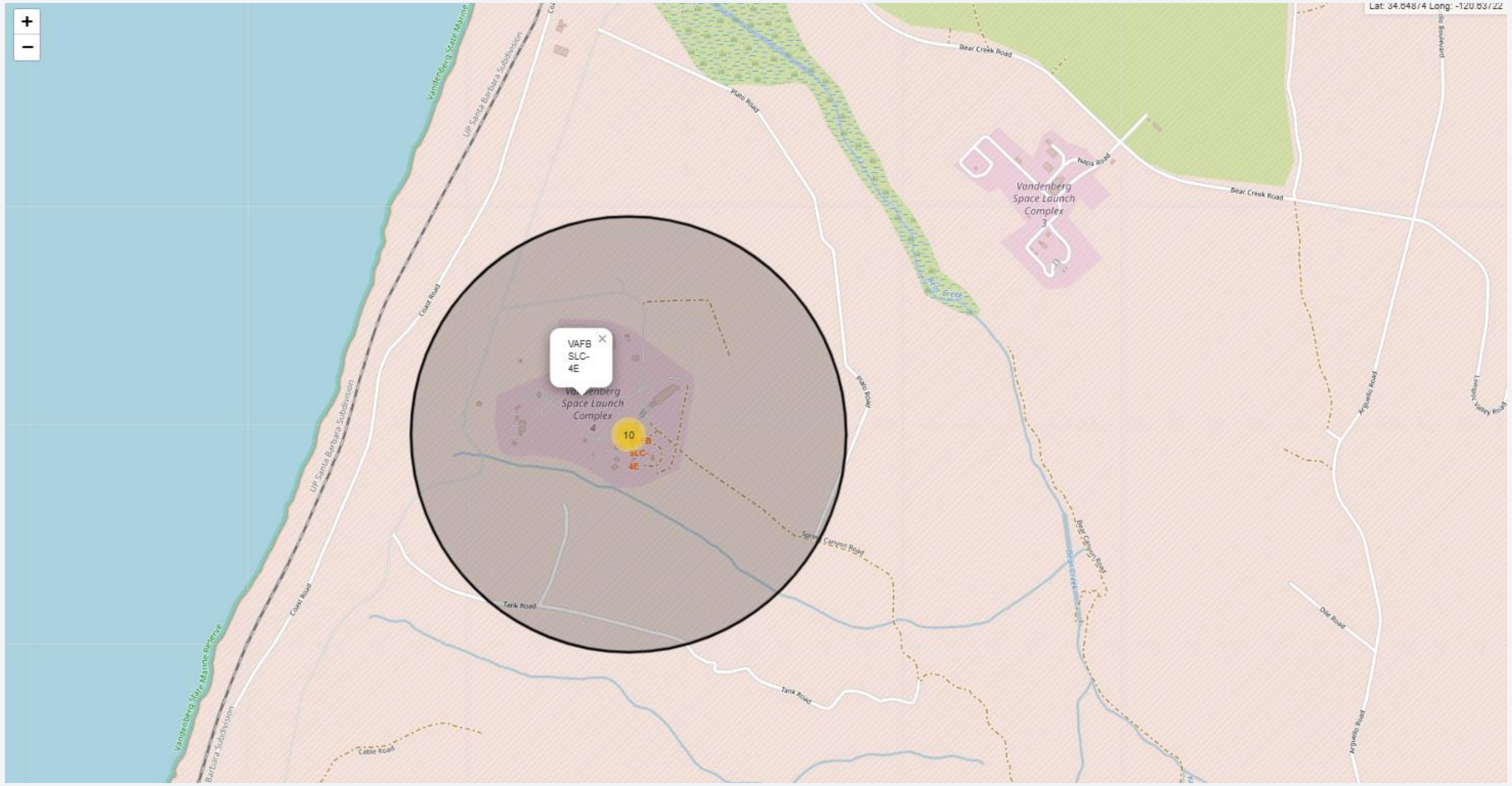


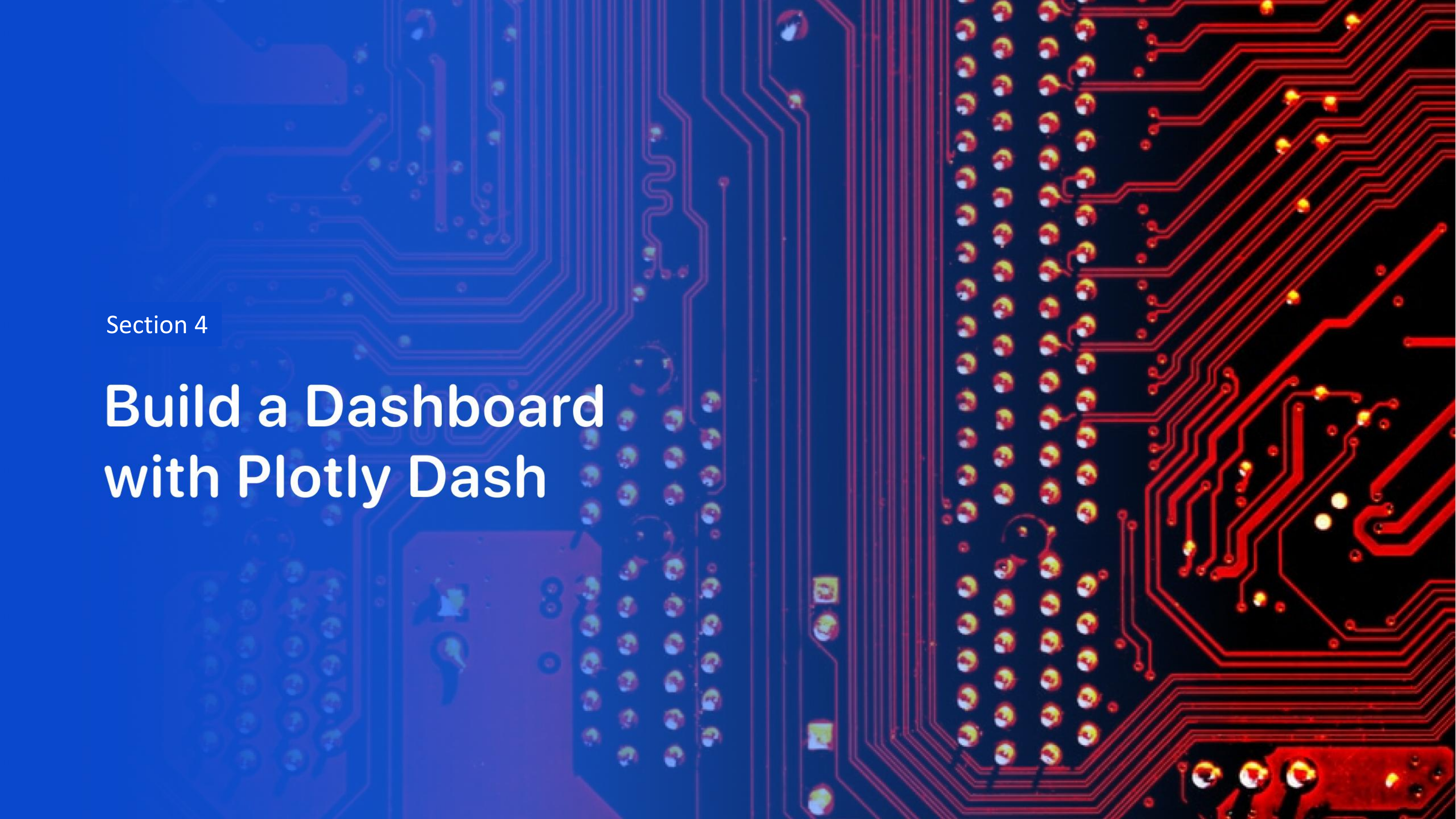


From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Map of a Launch Site

- Launch sites are not necessarily in close proximity to railways.
- Launch sites are in close proximity to highways.
- Launch sites are in close proximity to coastline.
- Launch sites keep a certain distance away from cities. This is primarily for safety reasons, ensuring that if a launch fails, debris falls into a less populated area.

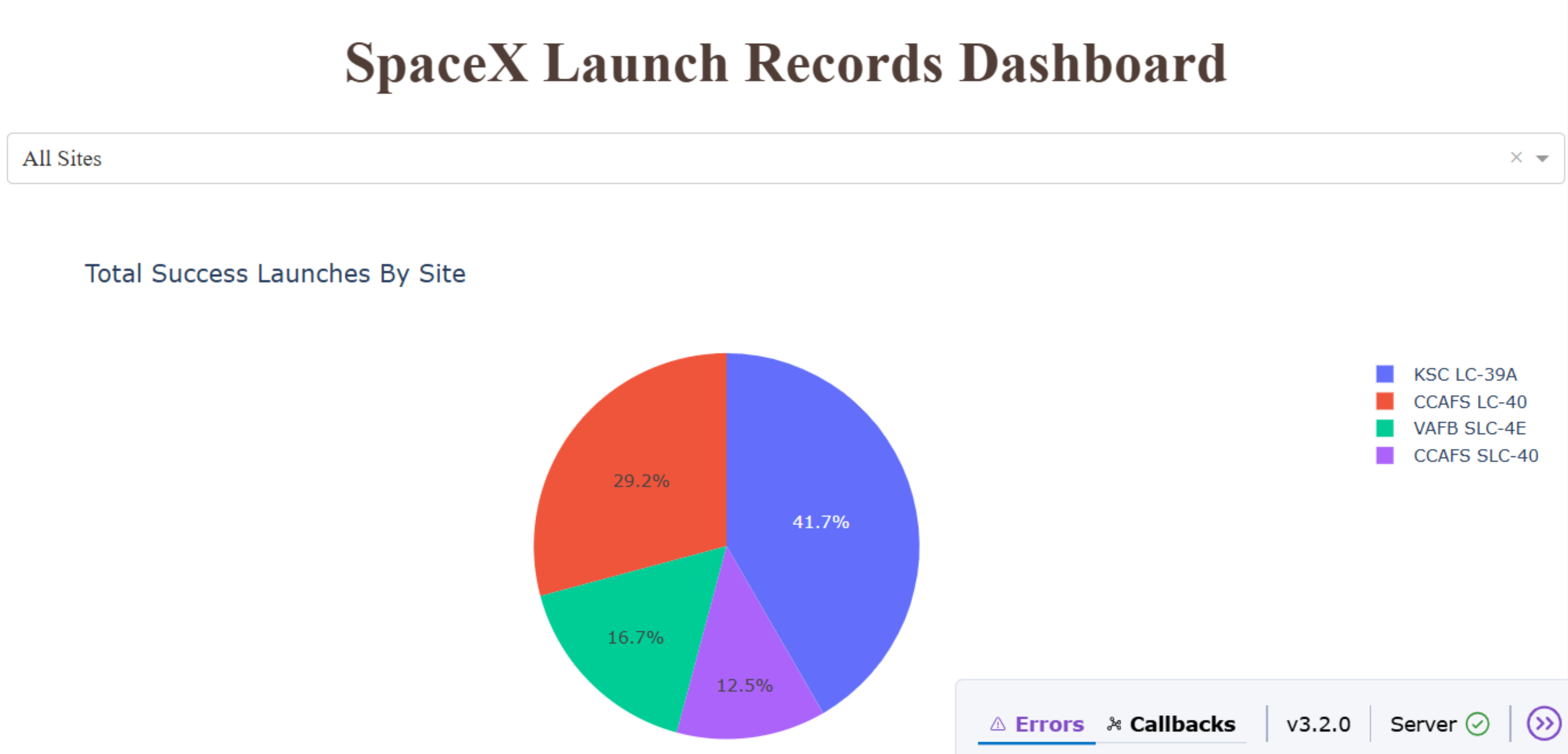

+
−
Lat: 34.64874 Long: -120.63722
Vandenberg State Marine Reserve
UP Santa Barbara Subdivision
Coast Road
Bear Creek Road
Vandenberg Space Launch Complex 3
VAFB SLC-4E
Vandenberg Space Launch Complex 4
10
SLC-4E
Spring Canyon Road
Tank Road
Cable Road
Arguello Road
Oil Road



Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

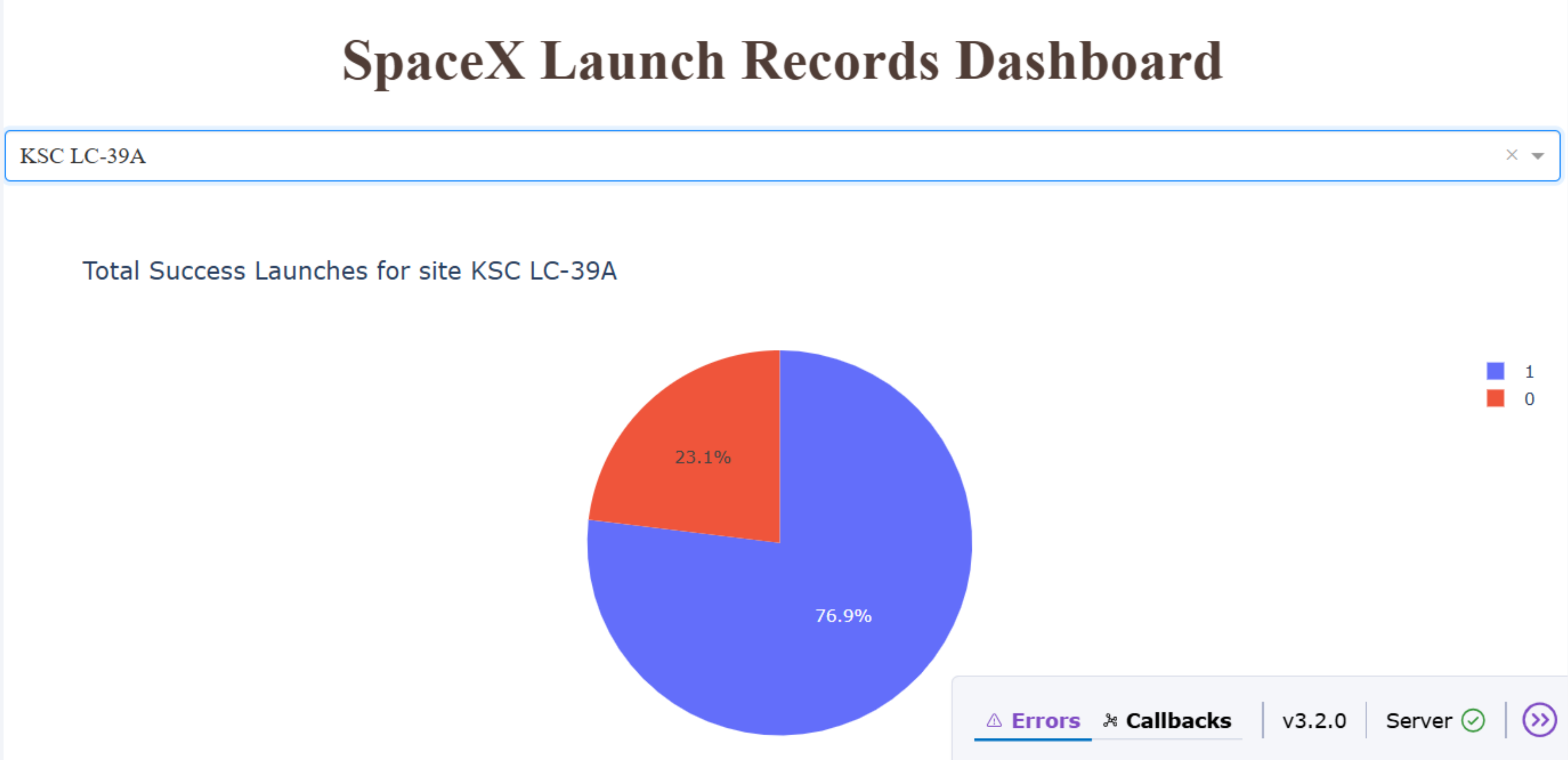


SpaceX Launch Records Dashboard

All Sites

- KSC LC-39A dominates with 41.7% of launches, followed by CCAFS LC-40 at 29.2%
- East Coast sites (KSC and CCAFS) handle ~71% of all successful launches, with West Coast VAFB at 16.7%

Total Success Launches for site KSC LC-39A



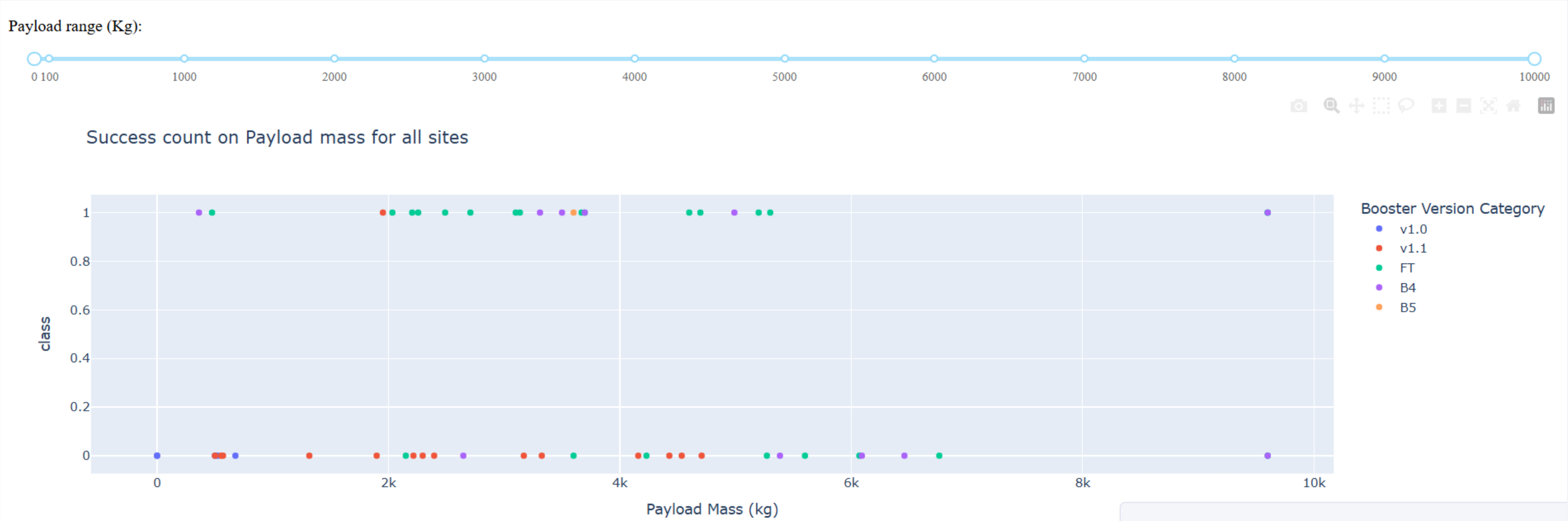
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

- KSC LC-39A has a 76.9% success rate (blue) versus 23.1% failure rate (red orange)

Success Count on Payload Mass for All Sites



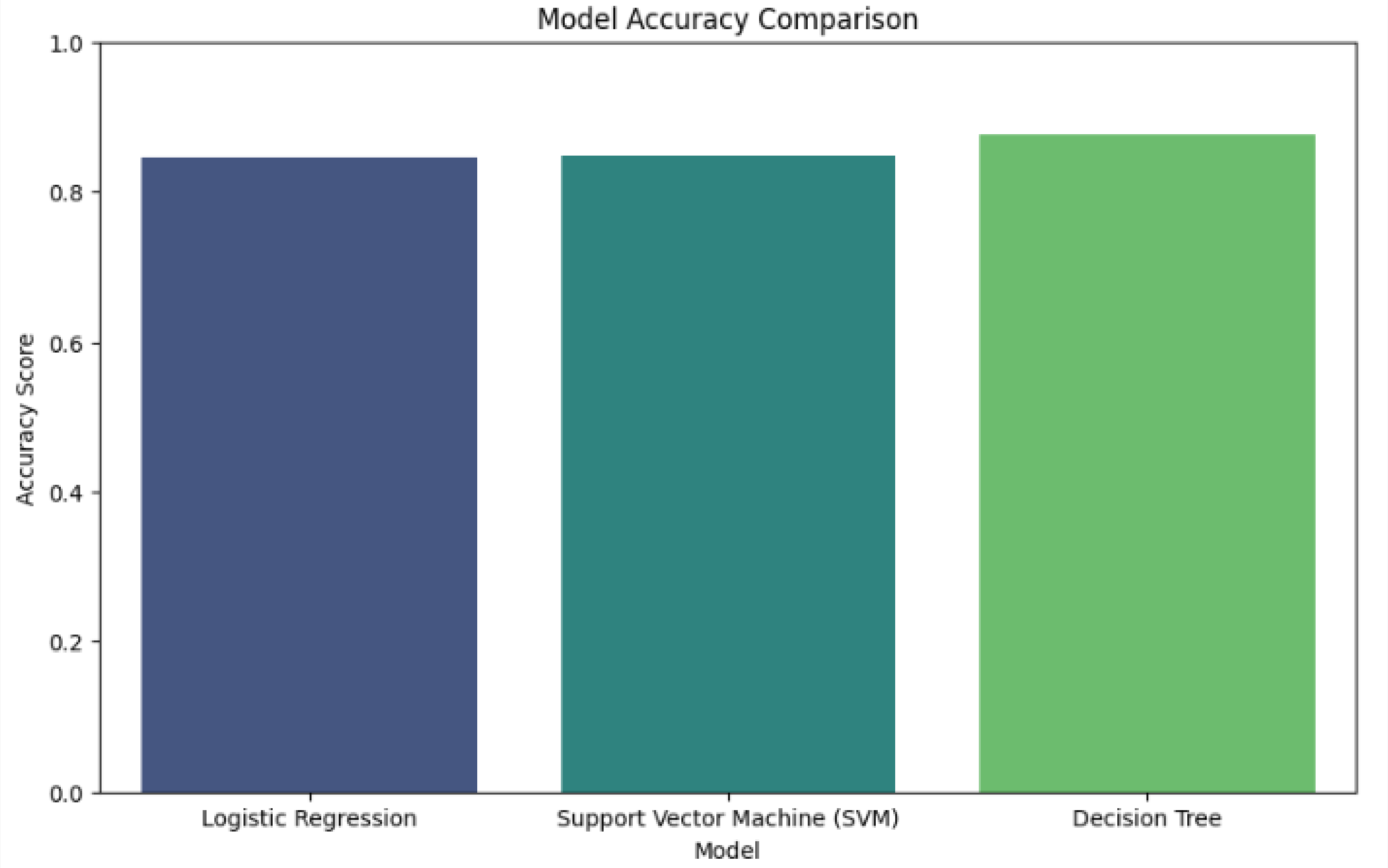
- Most successful launches (class=1) are concentrated in the 2k-6k kg payload range across all booster types
- Failures (class=0) occur primarily in lighter payloads under 4k kg, with older booster versions (v1.0, v1.1, FT) showing more failures than newer ones (B4, B5)

Section 5

Predictive Analysis (Classification)

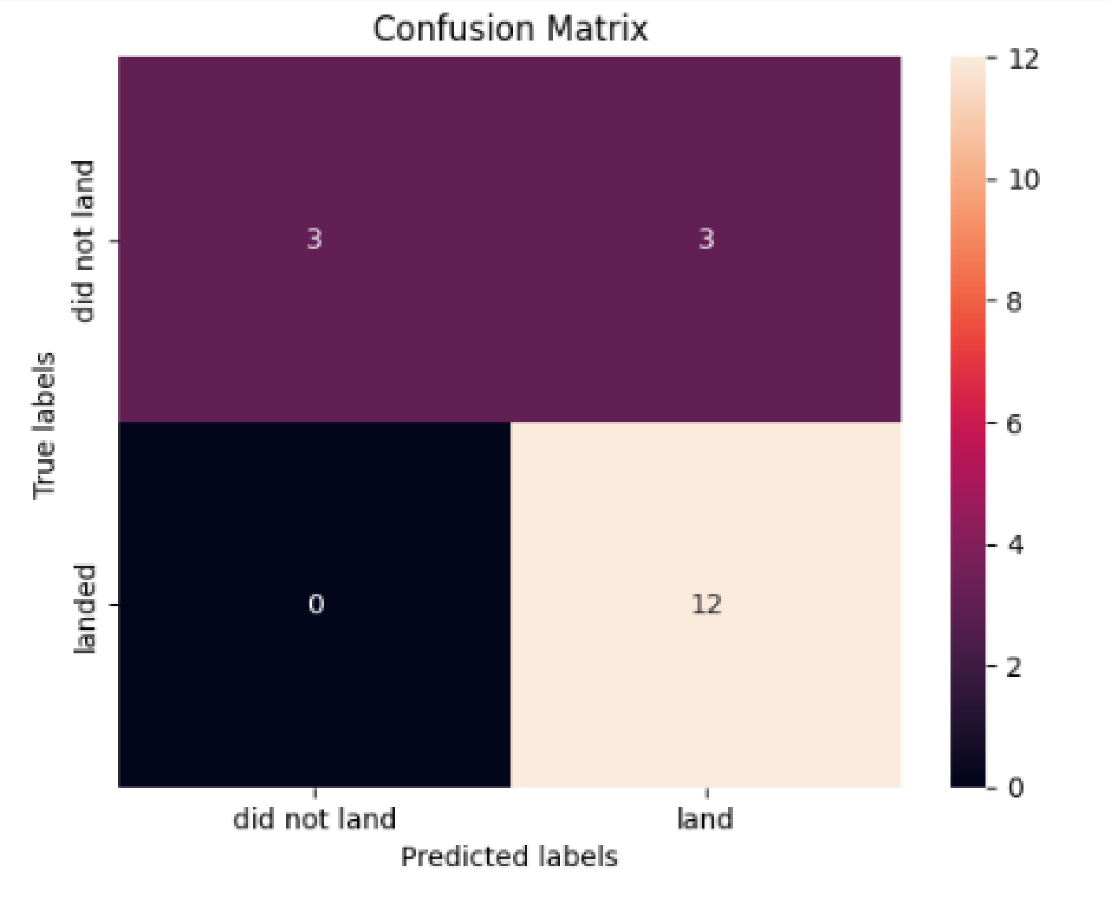
Classification Accuracy

- The model with the highest accuracy is Decision Tree with an accuracy of 0.8750



Confusion Matrix

- The confusion matrix showed the Decision Tree model correctly identified 12 successful landings and 3 unsuccessful landings, with 3 false positives and 0 false negatives.



Conclusions

- The project successfully developed a predictive model, and exploratory data analysis revealed a clear correlation between landing success and factors like payload mass and launch site.
- Landing success rates have improved over time, indicating the company's technological refinement and increased experience.
- The K-Nearest Neighbors model was identified as the best-performing algorithm, proving more effective than other evaluated models.
- An interactive Dash dashboard was created to serve as a powerful tool for visualizing the data and providing valuable insights for future missions.
- Overall, this data-driven approach demonstrates a reliable method for predicting Falcon 9 first-stage landing success and understanding its key contributing factors.

Thank you!