IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olga Feshchenko
November 9, 2024

Outline

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

Executive Summary

Methodologies Used for Data Analysis:

- **Data Collection**: Gathered data through web scraping and the SpaceX API.
- **Exploratory Data Analysis (EDA)**: Conducted data wrangling, visualization, and interactive visual analytics to explore and prepare the data.
- **Machine Learning Prediction**: Developed and evaluated predictive models to assess factors influencing launch success.

Key Findings:

- Successfully collected valuable data from public sources.
- EDA revealed key features most predictive of successful launches.
- Machine Learning models identified the most impactful characteristics for optimizing launch outcomes, leveraging all collected data effectively.

Introduction

The goal of this analysis is to assess the feasibility of Space Y entering the market as a competitor to SpaceX. The study aims to provide insights on critical factors that could influence Space Y's success, focusing on two key questions:

- **How to estimate the total cost of launches**: By predicting the likelihood of successful first-stage rocket landings, we can better understand and manage launch costs.
- **Optimal launch site selection**: Identifying the most advantageous locations for launches to maximize success and efficiency.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data on SpaceX launches was gathered from two primary sources:
  - 1. SpaceX API: Retrieved detailed information on rockets and launches from the official SpaceX API at https://api.spacexdata.com/v4/rockets/.
  - 2. Web Scraping: Collected additional launch data from Wikipedia's list of Falcon 9 and Falcon Heavy launches (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches).
- Perform data wrangling
  - The collected data was enhanced by generating a landing outcome label derived from the analysis and summarization of key features.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data were normalized and split into training and test datasets. Four different classification models were then applied, with each model's accuracy evaluated using various parameter combinations to identify the optimal configuration.

Data Collection

Data sets were gathered using the SpaceX API (link) and web scraping techniques applied to Wikipedia (link).

Data Collection – SpaceX API

- SpaceX provides a public API that enables easy access to launch data.
- This API was utilized following the outlined flowchart, and the retrieved data was then stored for further analysis.

Source code: https://github.com/olga-f/capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- SpaceX launch data is also available on Wikipedia.

- The data is downloaded from Wikipedia following the steps outlined in the flowchart and then saved for future use.

Source code: https://github.com/olga-f/capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- First, Exploratory Data Analysis (EDA) was conducted to gain initial insights into the dataset.
- Summaries were then generated, including launch counts per site, orbit type occurrences, and mission outcome frequencies for each orbit type.
- Lastly, a landing outcome label was created based on the data in the "Outcome" column.

Source code: https://github.com/olga-f/capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

To explore the data, scatter plots and bar plots were used to visualize relationships between various feature pairs: Payload Mass vs. Flight Number, Launch Site vs. Flight Number, Launch Site vs. Payload Mass, Orbit vs. Flight Number, and Payload vs. Orbit.

Source code: https://github.com/olga-f/capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The following SQL queries were executed:

- Retrieve the names of unique launch sites involved in space missions.
- List the top 5 launch sites whose names start with the string "CCA."
- Calculate the total payload mass for boosters launched by NASA (CRS).
- Find the average payload mass for missions using the booster version "F9 v1.1."
- Identify the date of the first successful landing on a ground pad.
- List the names of boosters that successfully landed on a drone ship and carried a payload mass between 4,000 and 6,000 kg.
- Count the total number of successful and failed mission outcomes.
- Retrieve the names of booster versions that carried the maximum payload mass.
- Find details of failed drone ship landings in 2015, including booster versions and launch site names.
- Rank the count of landing outcomes (such as "Failure (drone ship)" or "Success (ground pad)") for missions conducted between June 4, 2010, and March 20, 2017.

Source code: https://github.com/olga-f/capstone/blob/main/EDA.ipynb

Build an Interactive Map with Folium

Folium Maps utilized markers, circles, lines, and marker clusters for data visualization:

- **Markers** are used to represent specific locations, such as launch sites.
- **Circles** highlight areas around particular coordinates, such as the NASA Johnson Space Center.
- **Marker clusters** group multiple events at the same coordinates, like launches at a particular site.
- **Lines** display the distance between two coordinates.

Source code: https://github.com/olga-f/capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize the data:

- Percentage of launches by site
- Payload range

This combination enabled a quick analysis of the relationship between payload sizes and launch sites, helping to identify the optimal launch locations based on payload requirements.

Source code: https://github.com/olga-f/capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

The performance of four classification models was compared: logistic regression, support vector machine, decision tree, and k-nearest neighbors.

Source code: https://github.com/olga-f/capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

Exploratory Data Analysis Results:

- SpaceX operates four distinct launch sites.
- Initial launches were primarily conducted for SpaceX itself and NASA.
- The average payload capacity of the Falcon 9 v1.1 booster is 2,928 kg.
- The first successful landing of a Falcon 9 booster occurred in 2015, five years after the initial launch.
- Several Falcon 9 booster versions successfully landed on drone ships with payloads exceeding the average capacity.
- Nearly 100% of mission outcomes were successful.
- Two Falcon 9 v1.1 booster versions failed to land on drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
- The success rate of booster landings improved progressively over the years.

Interactive Analytics Insights:

- The analysis revealed that launch sites are strategically located in safe areas, typically near the sea, and benefit from robust logistical infrastructure.
- Most launches are conducted at launch sites along the east coast.

Predictive Analysis:

- The Decision Tree Classifier emerged as the most effective model for predicting successful landings, achieving an accuracy rate of over 87%, with test data accuracy exceeding 94%.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- According to the plot above, it’s possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It’s also possible to see that the general success rate improved over time.

Payload vs. Launch Site

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;

- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
- ES-L1;
- GEO;
- HEO; and
- SSO.

- Followed by:
- VLEO (above 80%); and
- LFO (above 70%).

Flight Number vs. Orbit Type

- Apparently, success rate improved over time to all orbits;

- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs. Orbit Type

- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;

- It seems that the first three years were a period of adju

All Launch Site Names

According to data, there are four launch sites:

Launch Site

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

They are obtained by selecting unique occurrences of “launch_site” values from 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

Total Payload Mass

Total Payload (kg)

Total payload carried by boosters from NASA: 111.268 kg

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

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1: 2.928 kg

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

By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

Number of successful and failure mission outcomes:

- Success 99
- Success (payload status unclear) 1
- Failure (in flight) 1

Boosters Carried Maximum Payload

Boosters which have carried the maximum payload mass:

- F9 B5 B1048.4
- F9 B5 B1048.5
- F9 B5 B1049.4
- F9 B5 B1049.5
- F9 B5 B1049.7
- F9 B5 B1051.3
- 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

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

- F9 v1.1 B1012 CCAFS LC-40
- F9 v1.1 B1015 CCAFS LC-40

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

Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

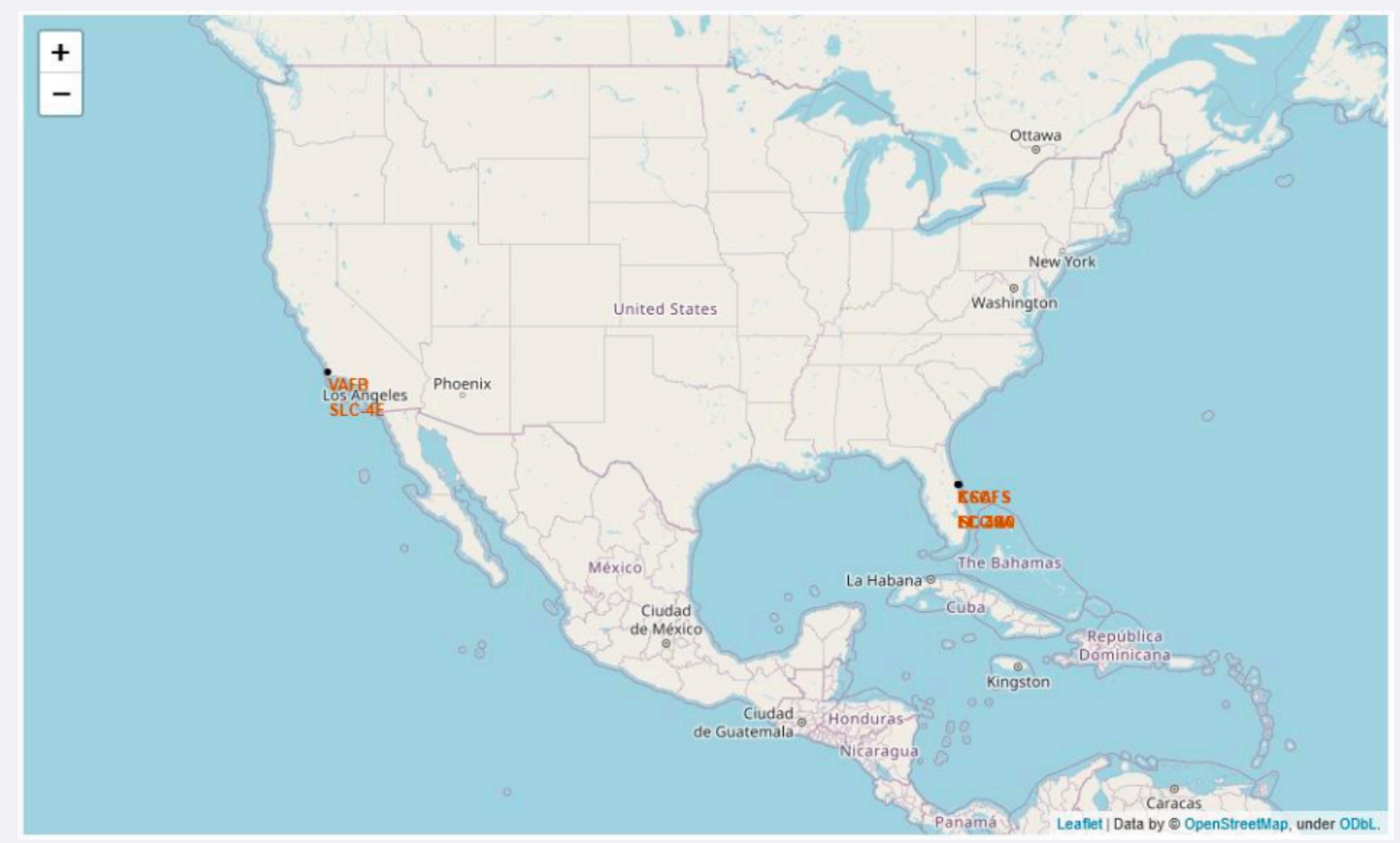
- No attempt 10
- Failure (drone ship) 5
- Success (drone ship) 5
- Controlled (ocean) 3
- Success (ground pad) 3
- Failure (parachute) 2
- Uncontrolled (ocean) 2
- Precluded (drone ship) 1

Section 3

Launch Sites Proximities Analysis

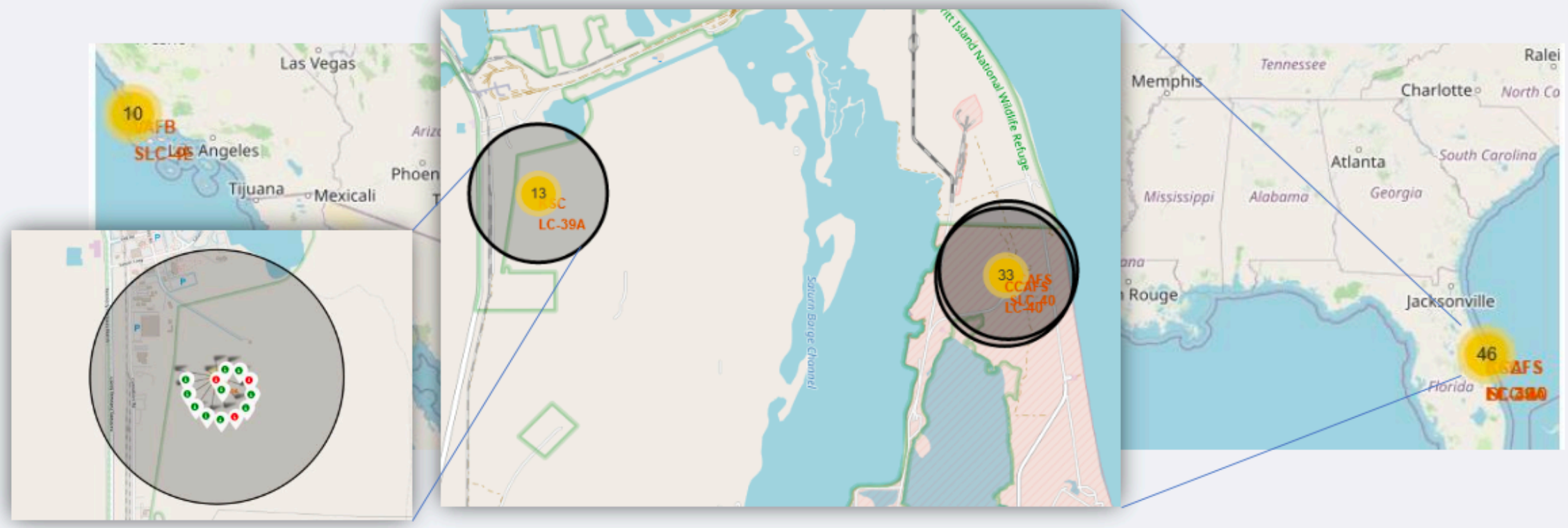
All launch sites

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



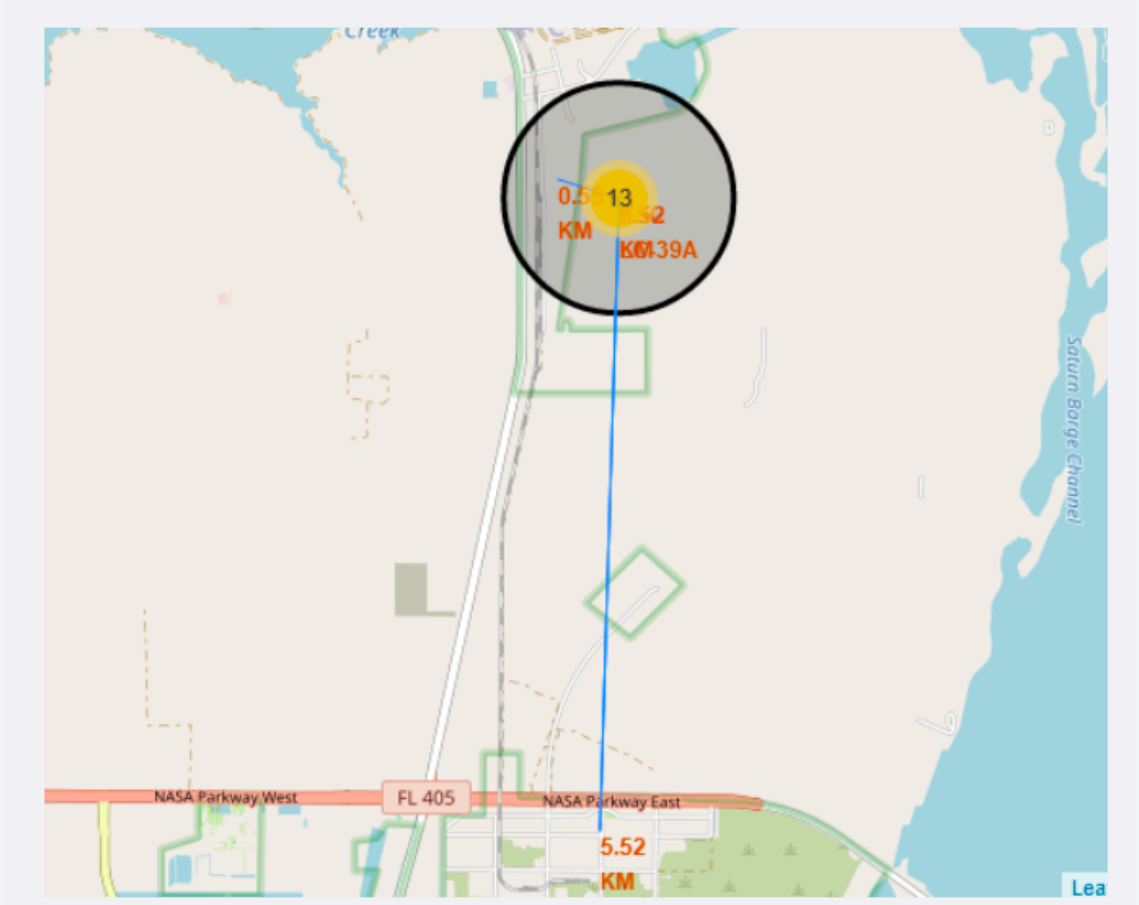
Launch Outcomes by Site

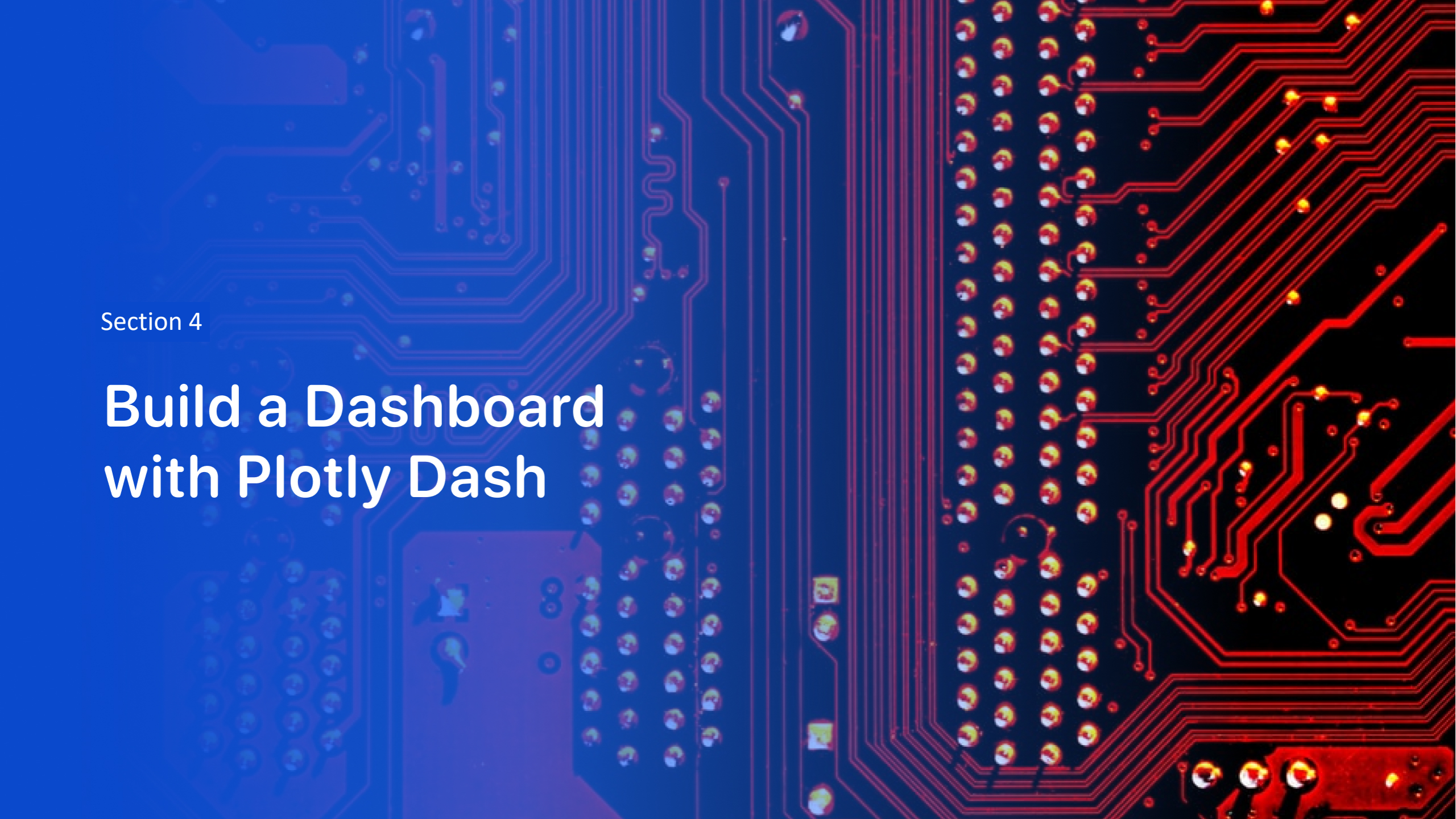
- Example of KSC LC-39A launch site launch outcomes



Logistics and Safety

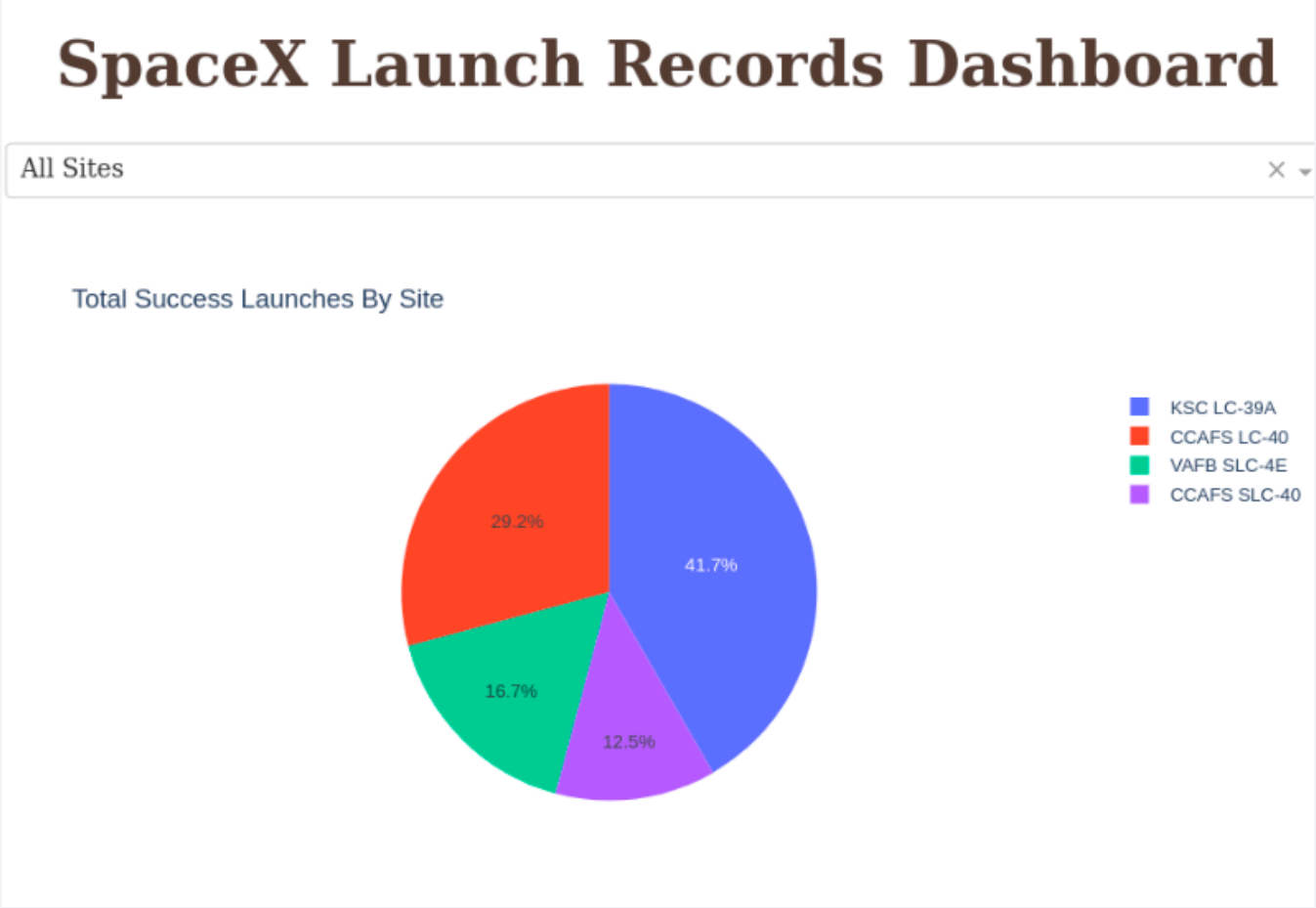
Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.




Section 4
Build a Dashboard with Plotly Dash

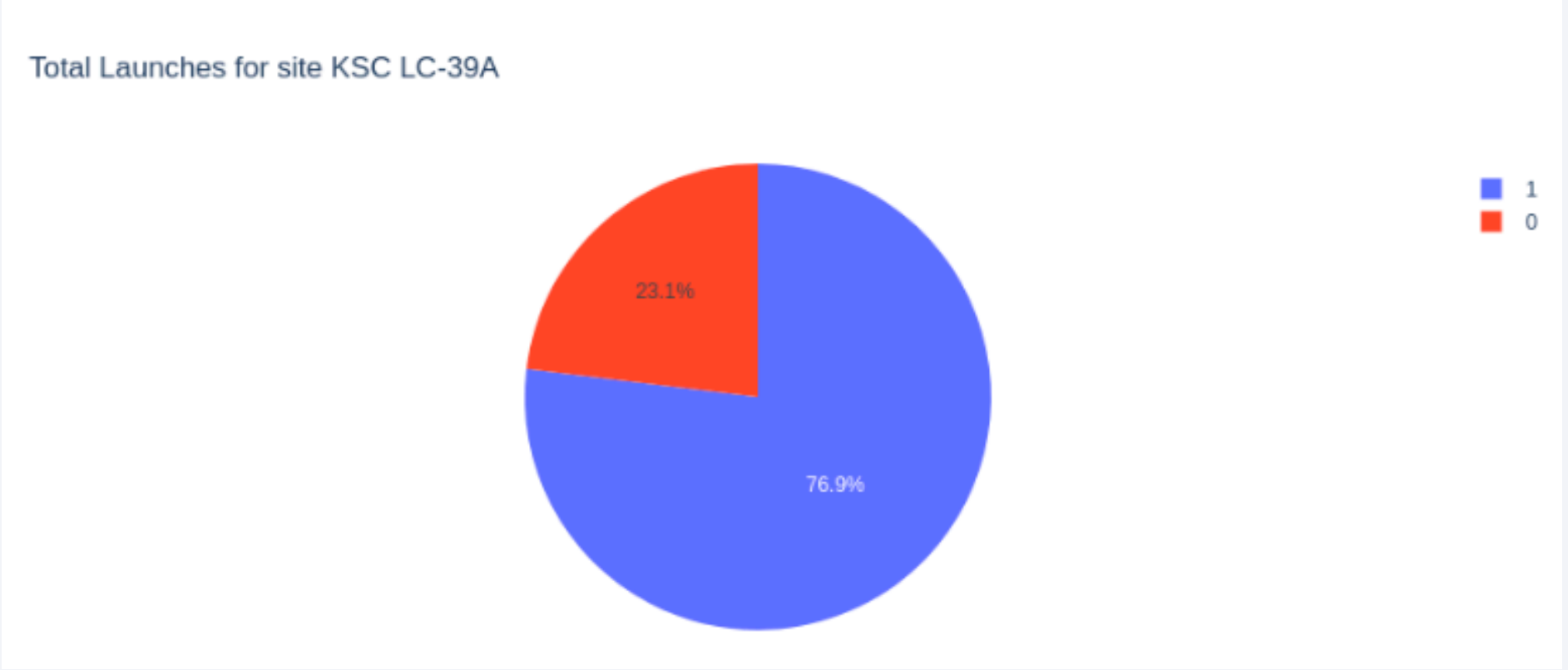
Successful Launches by Site

The place from where launches are done seems to be a very important factor of success of missions.



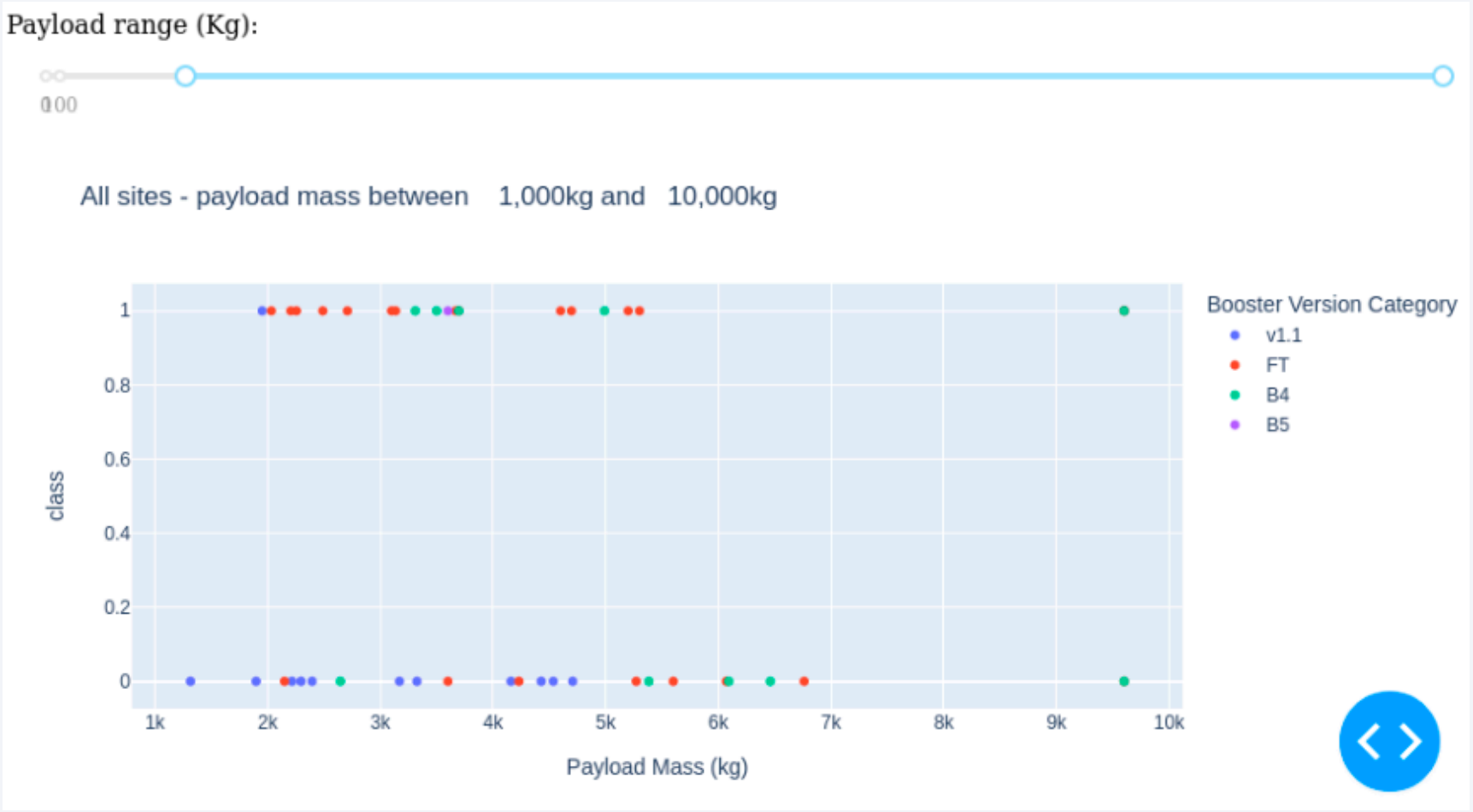
Launch Success Ratio for KSC LC-39A

76.9% of launches are successful in this site.



Payload vs. Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.



Section 5

Predictive Analysis (Classification)

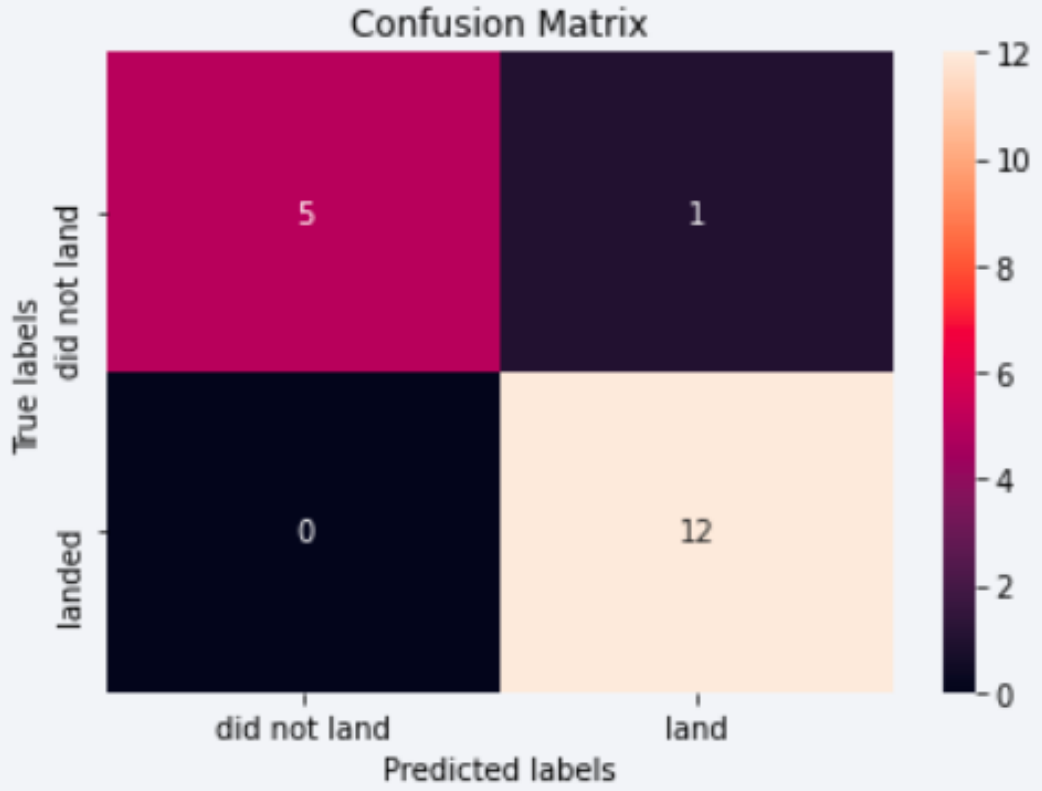
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.

Confusion Matrix

Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Multiple data sources were analyzed, enabling refined conclusions throughout the process.
- The optimal launch site identified is KSC LC-39A.
- Launches with payloads over 7,000 kg show reduced risk.
- While most missions are successful, the rate of successful landings appears to improve over time, likely due to advancements in processes and rocket technology.
- A Decision Tree Classifier model can help predict successful landings, potentially boosting profitability.

Thank you!