IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

IJAZ AHMED

01.09.2025

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Studied factors influencing Falcon 9 launch success by collecting and preparing data from APIs and web scraping, engineered features for predictive machine learning models, and developed an interactive dashboard for live analytics and visualization

Introduction

SpaceX has made space travel far more affordable — $62M per launch vs. $165M for competitors — mainly because the first stage is reusable.

Using Machine Learning models, our goal is to identify the best predictive model for determining when the first stage can be reused successfully.

Key Questions:

- How often is reuse successful?
- Which factors (payload, orbit, launch site, booster version, etc.) influence the outcome?

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

- API Calls
- Collected launch data using SpaceX REST API
- Retrieved structured JSON responses with launch details (rocket, payload, orbit, landing outcome)
- Converted JSON into Pandas DataFrame for analysis
- Web Scraping
- Used requests module to fetch HTML pages (e.g., Wikipedia Falcon 9 launches)
- Parsed HTML with BeautifulSoup (html.parser)
- Extracted tables containing launch history and outcomes
- Combined with API data for a richer dataset

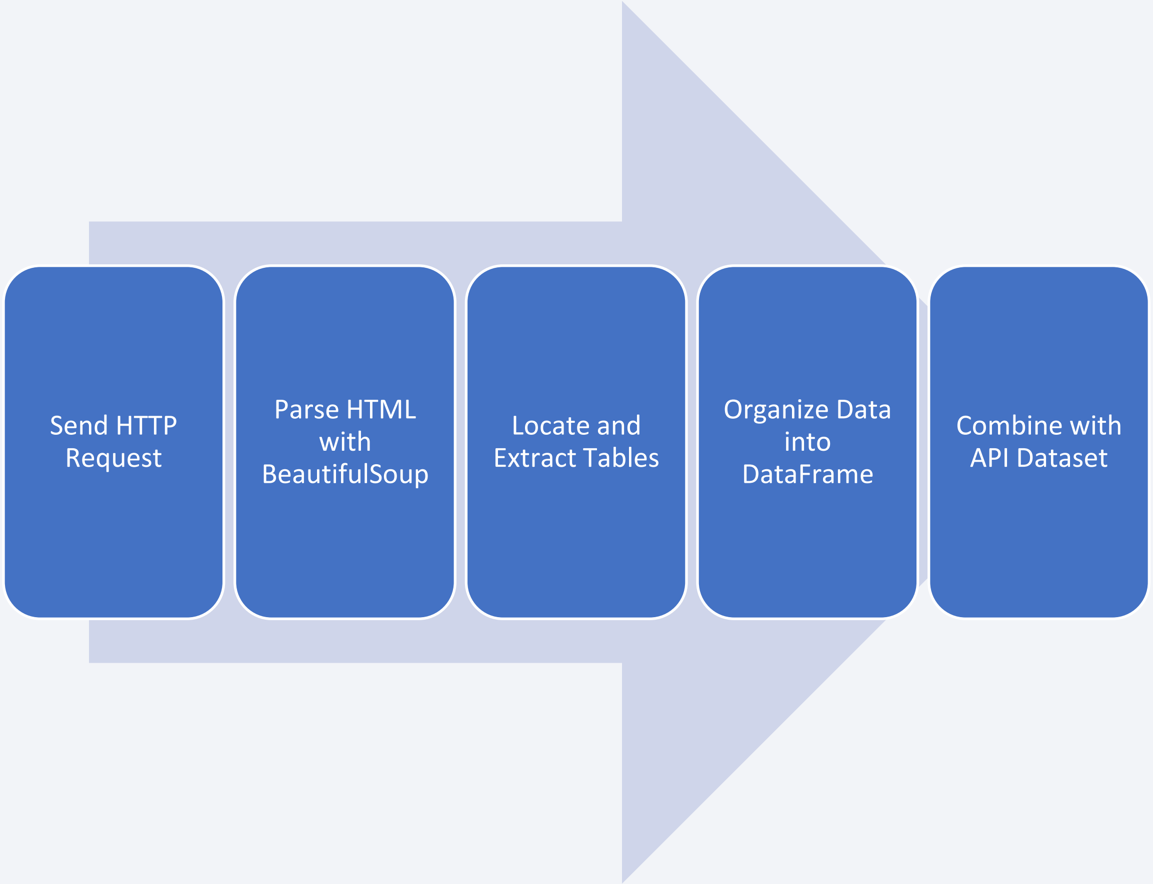
Data Collection – SpaceX API

- API Calls
- Collected launch data using SpaceX REST API
- Retrieved structured JSON responses with launch details (rocket, payload, orbit, landing outcome)
- Converted JSON into Pandas DataFrame for analysis
- [Notebook URL]()

Access SpaceX API	• Endpoint: /v4/launches/past
Fetch Launch Data	• Request JSON response for all past launches
Normalize JSON	• Convert raw JSON into a Pandas DataFrame
Extract Rocket Information	• Call /v4/rockets/{id} to get booster version
Extract Payload Information	• Call /v4/payloads/{id} to get payload mass & orbit
Extract Launch Site Information	• Call /v4/launchpads/{id} for site name & coordinates
Extract Core Information	• Call /v4/cores/{id} for flights, reuse, block, serial
Combine Data	• Merge booster, payload, site, and core details into one dataset
Filter Falcon 9 Launches	• Remove Falcon 1 and multi-core launches
Save Final Dataset	• Export as dataset_part_1.csv for later analysis

Data Collection - Scraping

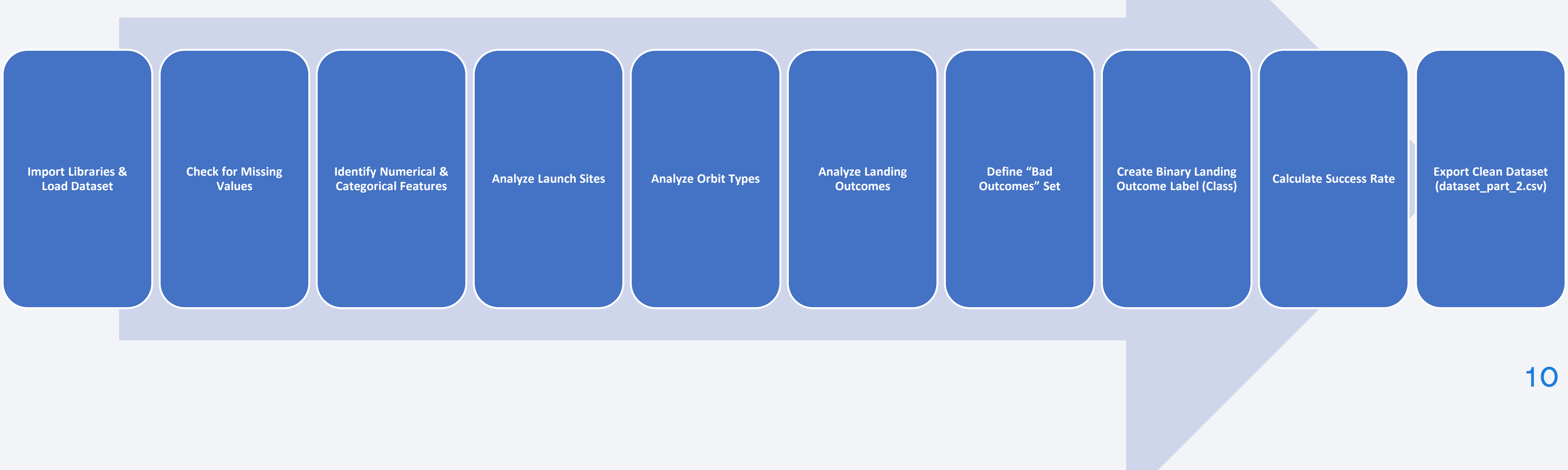
- Web Scraping
- Used requests module to fetch HTML pages (e.g., Wikipedia Falcon 9 launches)
- Parsed HTML with BeautifulSoup (html.parser)
- Extracted tables containing launch history and outcomes
- Combined with API data for a richer dataset



- URL

Data Wrangling

In the data wrangling step, we loaded the Falcon 9 dataset, checked for missing values, and identified numerical and categorical features. We analyzed the distribution of launches across sites, orbits, and landing outcomes, then grouped all unsuccessful outcomes into a single “bad outcomes” category. Using this, we created a new binary column Class, where 1 represents a successful landing and 0 represents a failure. This simplified label will be used as the training target for our machine learning models, and the cleaned dataset was saved as dataset_part_2.csv for subsequent analysis. NOTEBOOK: [GitHub URL]



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Inteactive Map with Folium

- In the Folium interactive map, I created and added the following objects:
- Markers with Labels (DivIcon): Placed at the nearest city, railway, and highway to the launch site. Each marker displayed the type of feature (City, Railway, Highway) along with its distance in kilometers.
- Colored Lines (PolyLine): Drew straight lines from the launch site to each proximity feature (city, railway, highway) for visual clarity.
- Custom Colors:
- City → Purple (#6f42c1)
- Railway → Brown (#8B4513)
- Highway → Orange (#fd7e14)
- Coastline -> Blue
- [Notebook URL]()

Build an Inteactive Map with Folium

- Markers make it easy to identify nearby infrastructure and visualize the exact location of cities, railways, and highways relative to launch sites.
- Lines help quantify and illustrate the distance from the launch site to these key features.
- Color-coding ensures each category (City, Railway, Highway) is distinct and easy to interpret on the map.
- These map objects were essential for exploring questions like:Are launch sites close to railways for easy transport of rocket components?
- Are they close to highways for logistical support?Do they maintain distance from cities for safety?How close are they to the coastline, which is important for safe rocket trajectories?

Build an Inteactive Map with Folium: Findings

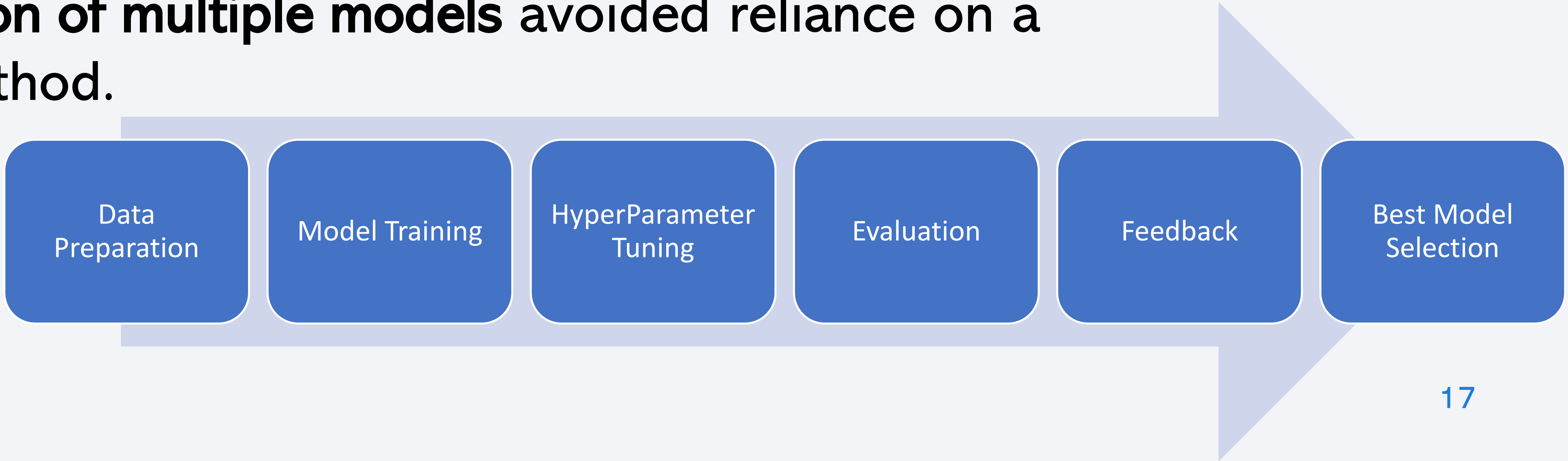
- From the distance calculations for one example site:
- City Distance: ~61 km (launch sites are kept far from cities, ensuring safety)
- Railway Distance: ~0.78 km (very close, useful for transporting heavy rocket parts)
- Highway Distance: ~25.6 km (within reasonable proximity for logistics)
- These results suggest that launch sites are strategically placed:
- Close to infrastructure like railways and highways for operational efficiency.
- Far from cities to minimize risks to populated areas.
- Near the coastline (seen in other parts of the lab) for safer rocket launches over the ocean.

Build a Dashboard with Plotly Dash

- Built an interactive Plotly Dash dashboard to analyze SpaceX Falcon 9 launch data. The dashboard includes:
- Dropdown Input (Launch Site Selection): Allows users to select All Sites or a specific launch site.
- Pie Chart (Success Distribution): Updates dynamically based on the dropdown. For All Sites, it shows the distribution of successful launches by site. For a specific site, it shows success vs. failure counts.
- Range Slider (Payload Selection): Enables filtering of launches based on payload mass (0–10,000 kg).
- Scatter Plot (Payload vs Success): Updates dynamically based on both dropdown and slider inputs. It shows the relationship between payload mass and launch outcome (success/failure), with booster version categories used as color labels.
- Here is the link for script

Predictive Analysis (Classification)

Systematic tuning avoided bias toward default parameters.
Cross-validation ensured stability across folds.
Confusion matrix gave clarity on types of errors (false positives).
Comparison of multiple models avoided reliance on a single method.



[Github Link]

Results

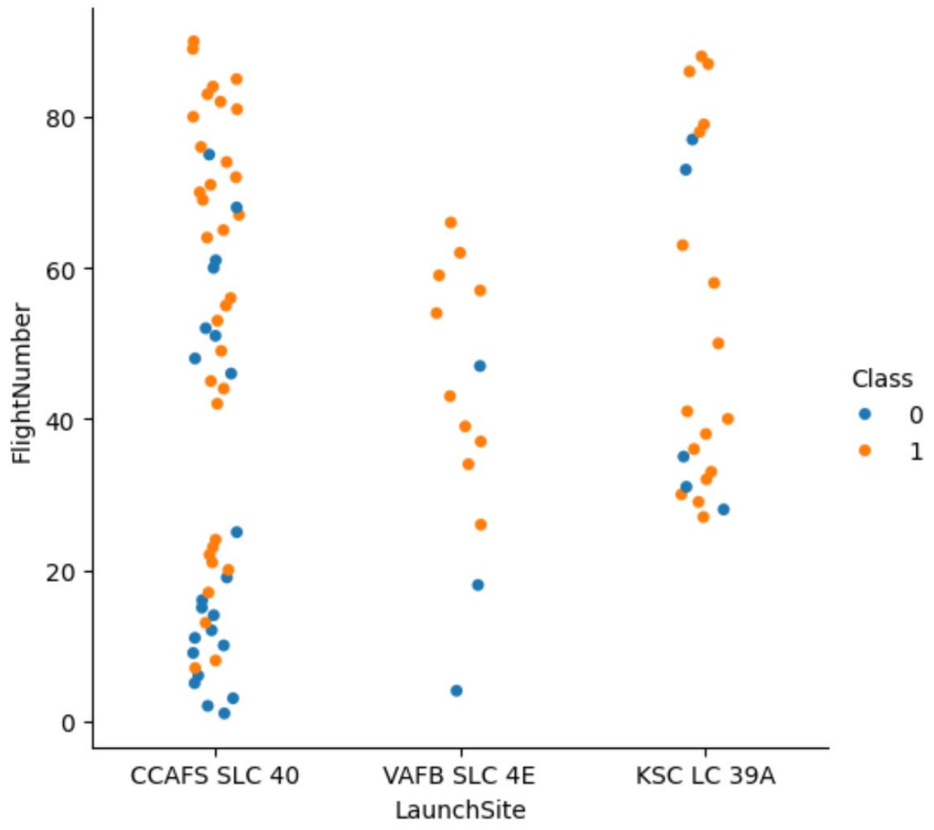
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

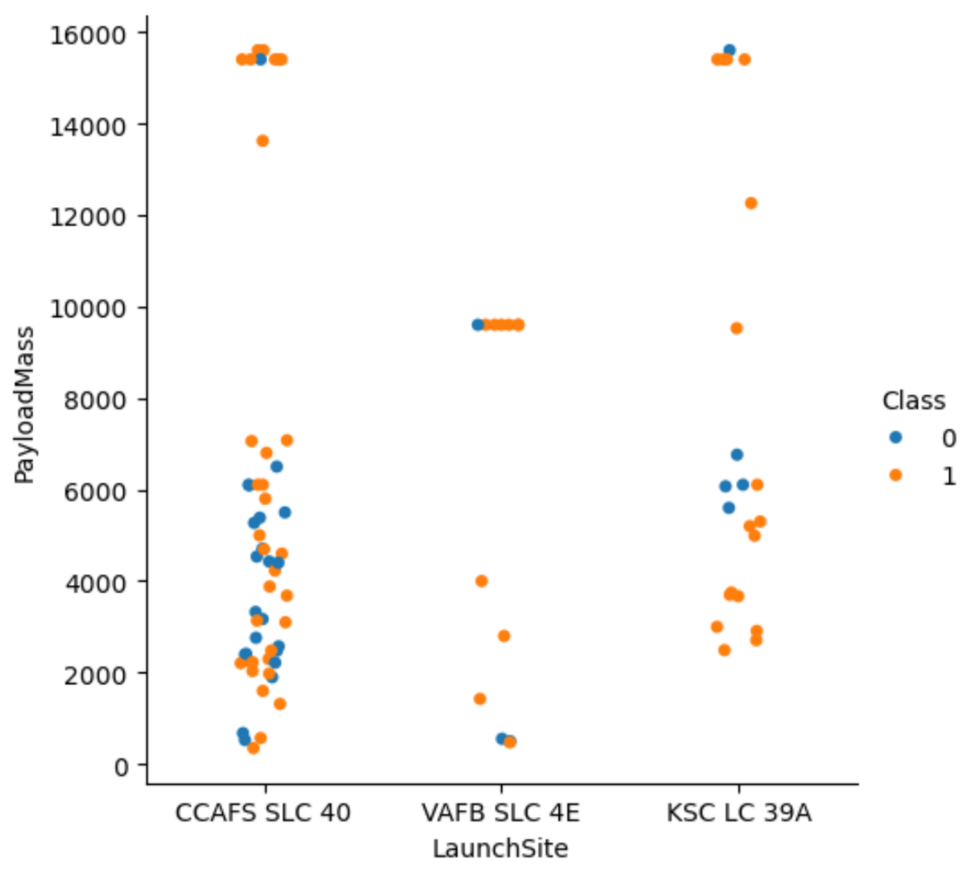
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations



Payload vs. Launch Site

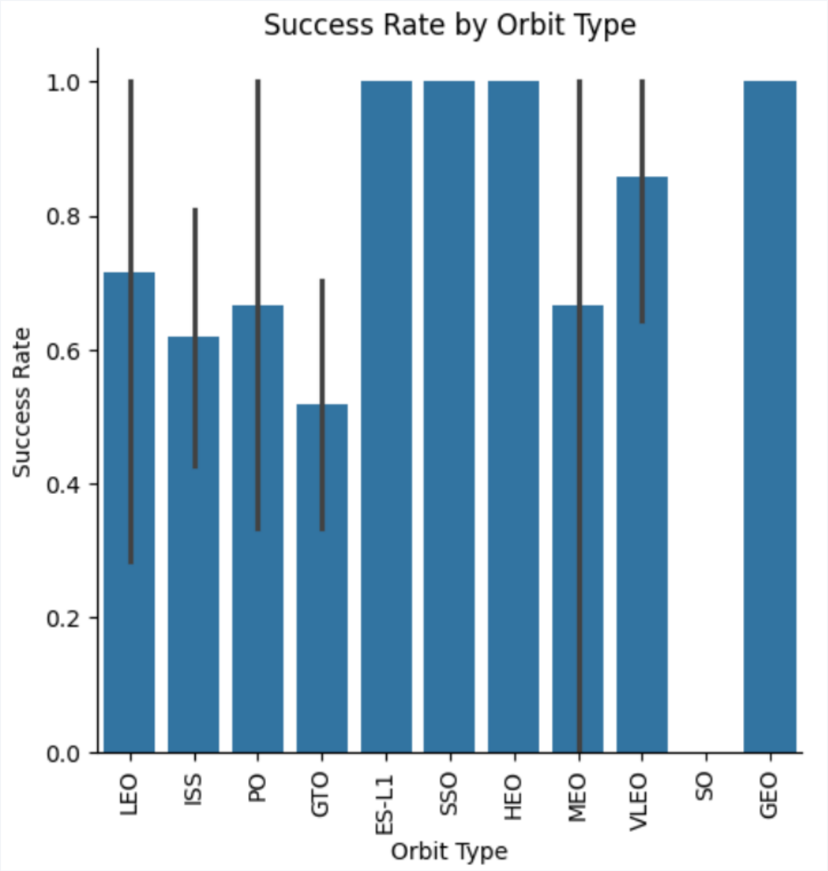
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



Success Rate vs. Orbit Type

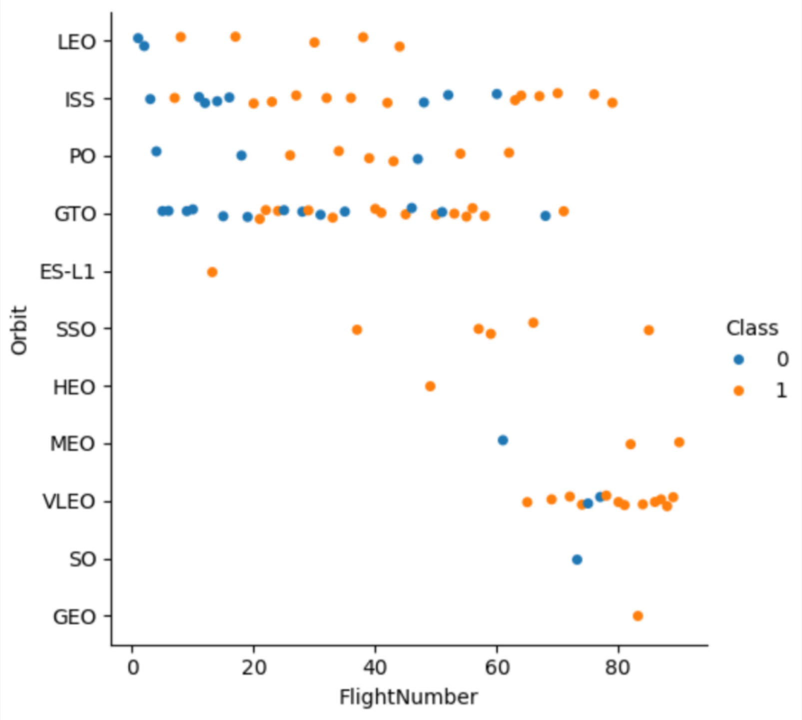
- Show a bar chart for the success rate of each orbit type

- Show the screenshot of the scatter plot with explanations



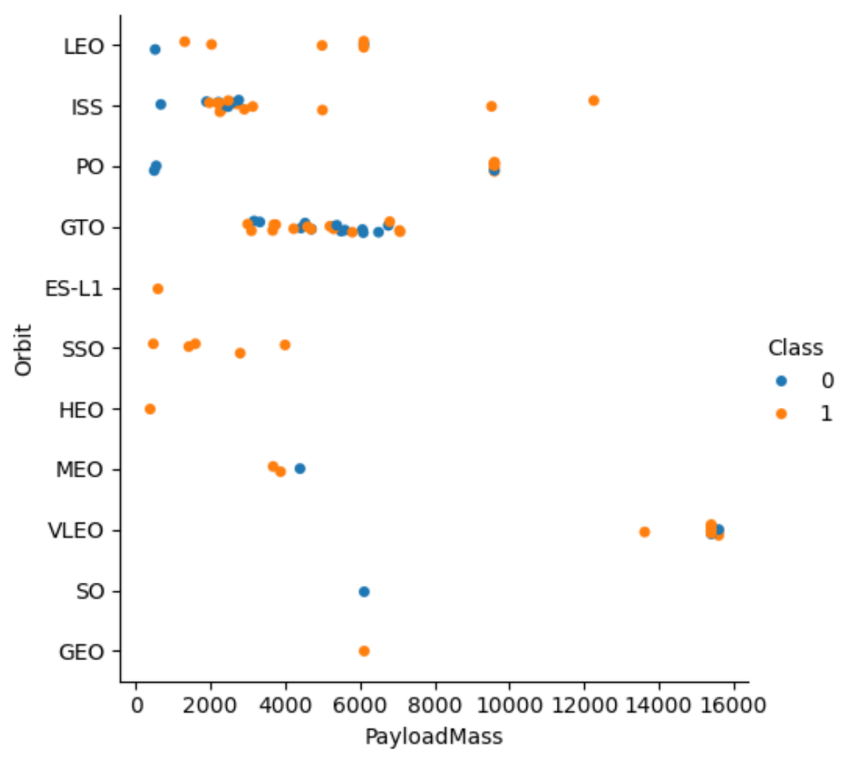
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations



Payload vs. Orbit Type

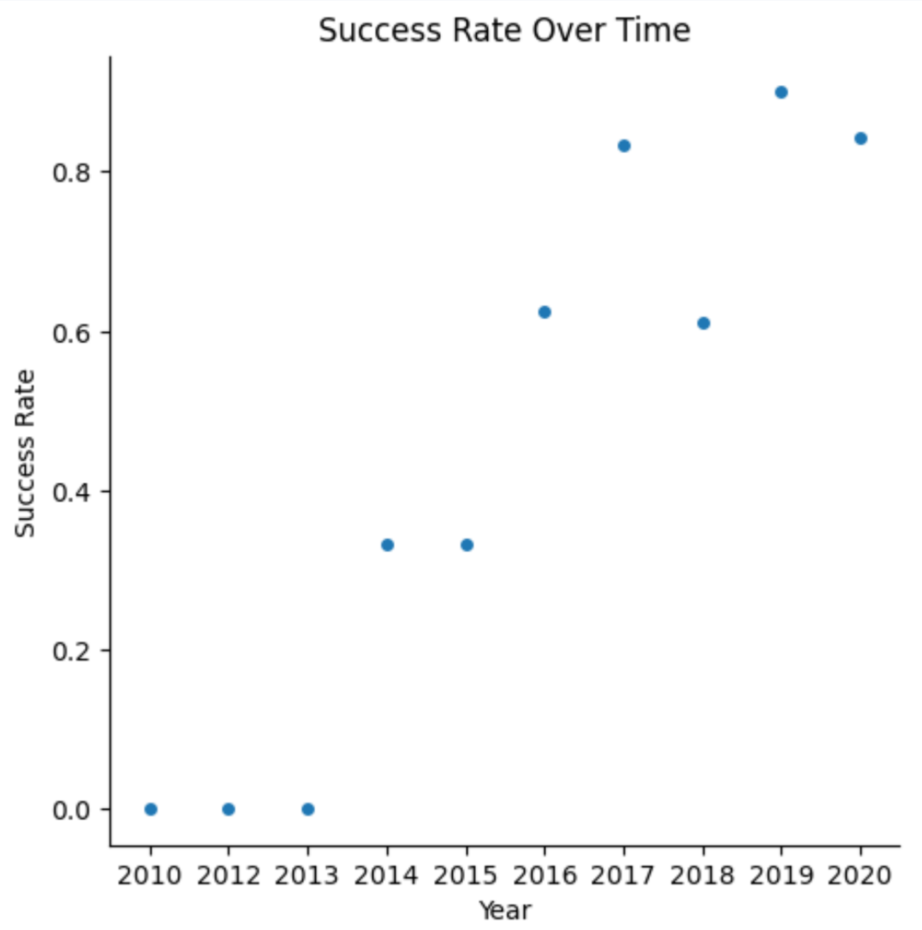
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



Launch Success Yearly Trend

- Shows a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- We have the same one, since the Task didn't ask for Distinct

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The total payload carried by boosters from NASA

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- For the F9 v1.1, we have the following 2928.4 kgs as the average payload mass.

- For reference, total payload for NASA
- Was 45596 kgs

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Here is the date for the first successful landing outcome on ground pad

Around Christmas Eve for the year 2015

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Out of 61 Successful Landings, only 4 fall within this range

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

Total Number of Successful and Failure Mission Outcomes

- 61 Success Outcomes, whereas only 10 resulted in Failure
- Not counting the no-attempt made here as a failure or a success.

```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success%'
```

 * sqlite:///my_data1.db
Done.

COUNT(*)
61

```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Failure%'
```

 * sqlite:///my_data1.db
Done.

COUNT(*)
10

Boosters Carried Maximum Payload

- Listed are the names of the booster which have carried the maximum payload mass
- A lot of Booster Versions have carried the Maximum Payload

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
- Both were on the same Launch Site and drone ship.

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

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

- For drone ships, the success rates in the given date range is almost 50%, and for 10 we have no attempt at all.
- Total count of 31, but only 21 attempts.

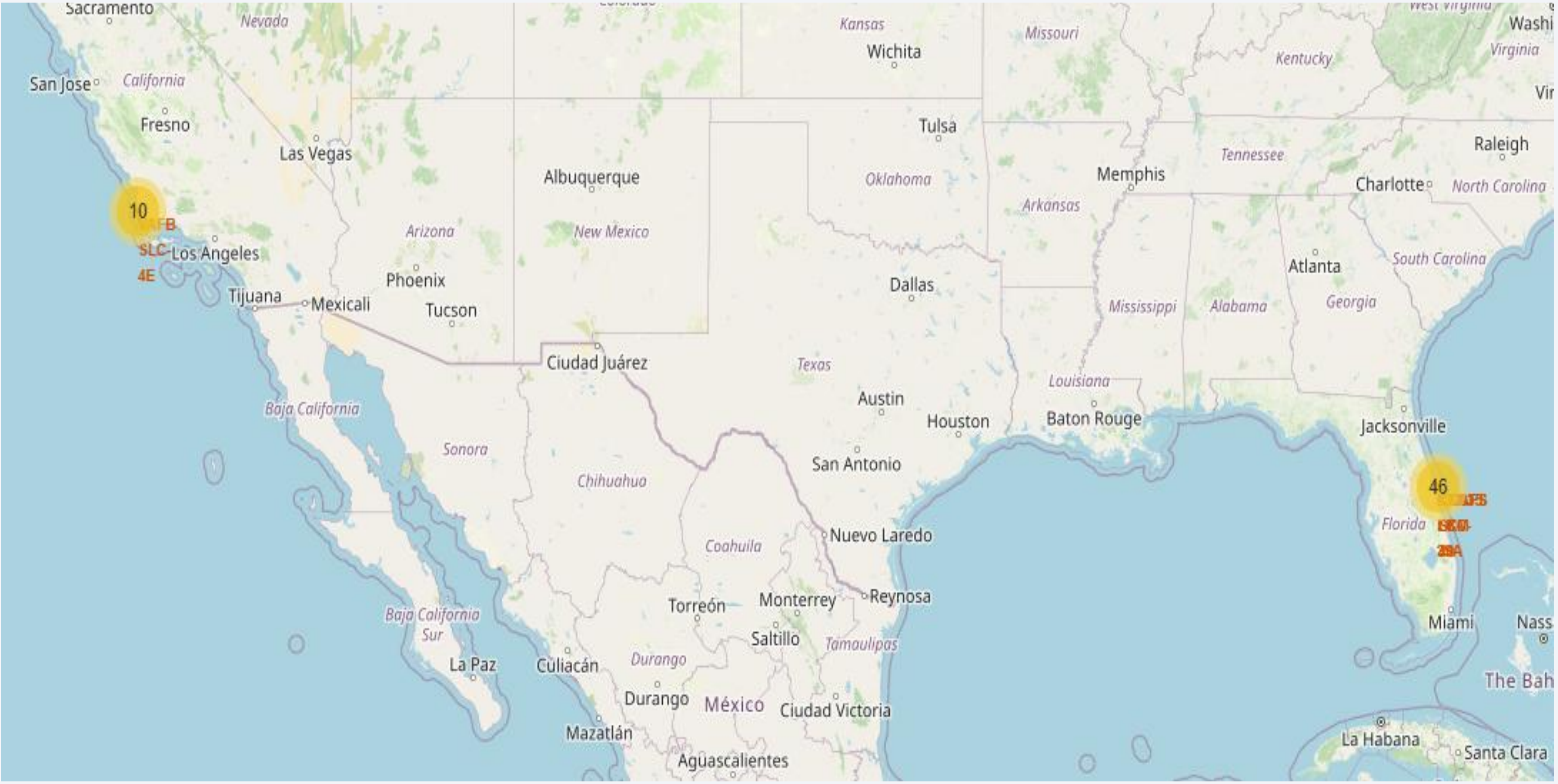
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Interactive Folium Map (Global Map)

Launch Sites



Distance Markers

On the interactive Folium map, we added **markers and lines** to show distances from the launch site to key proximities:

City Distance ~61 km away (purple marker)

Railway Distance ~0.78 km away (brown marker, very close)

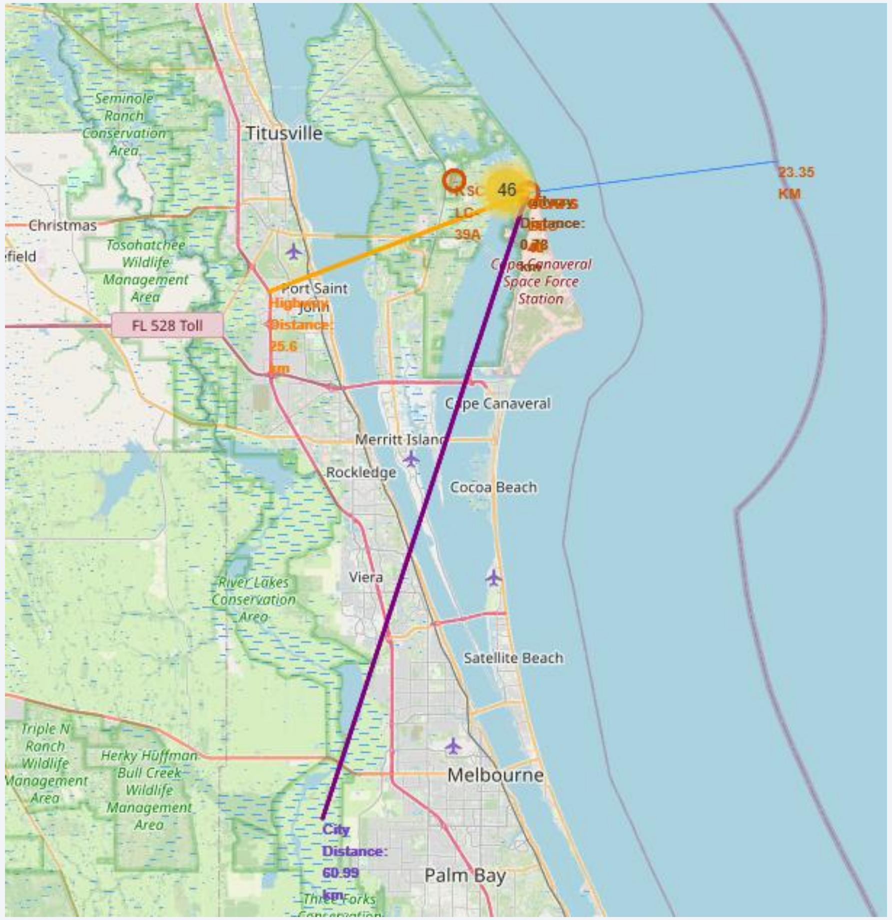
Highway Distance ~25.6 km away (orange marker)

Insights:

Launch sites are strategically close to **railways and highways** for logistics and transport.

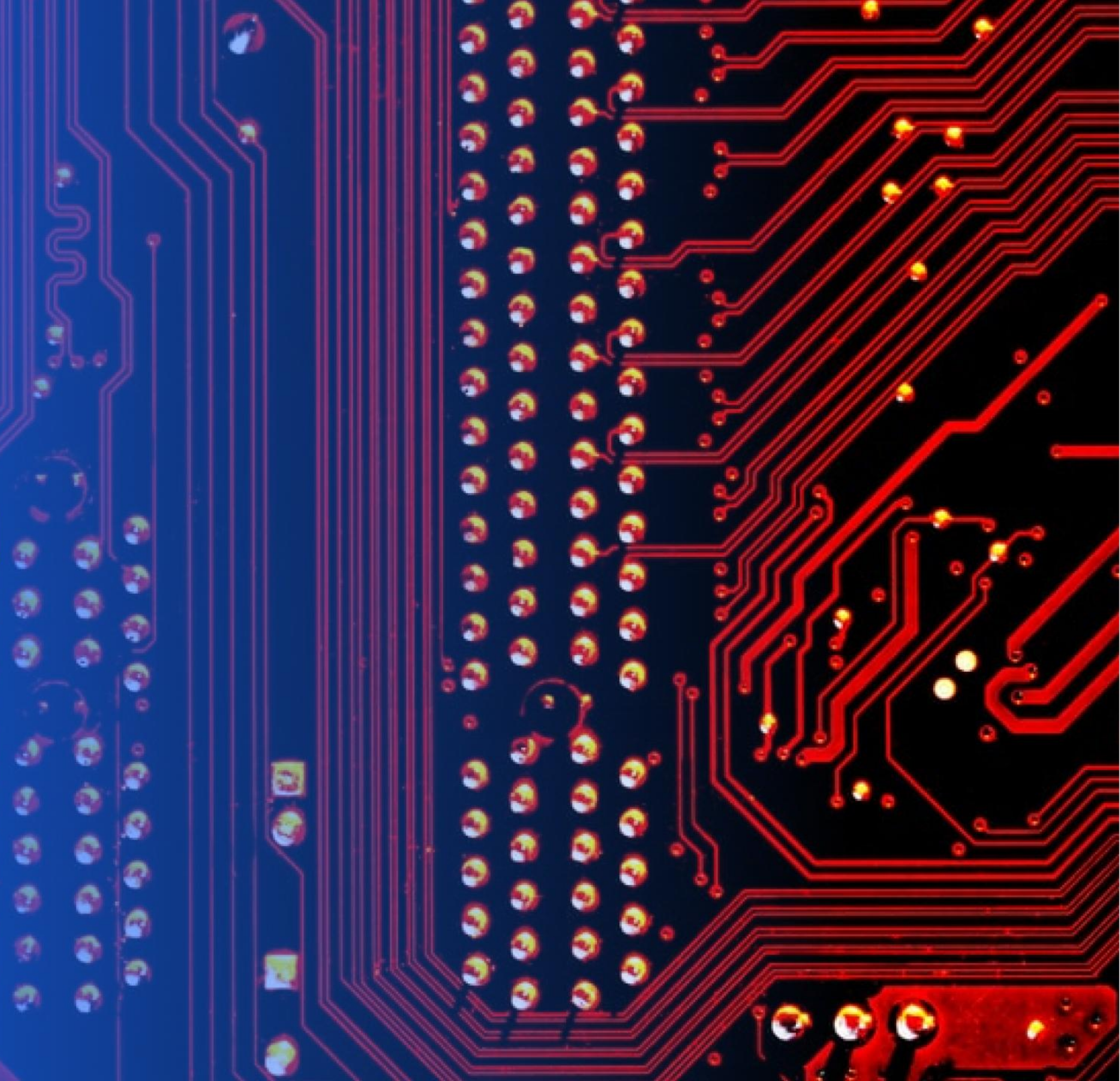
Sites maintain a **safe buffer from cities** for security and safety.

Coastline proximity supports **safe landings and ocean recovery missions**.

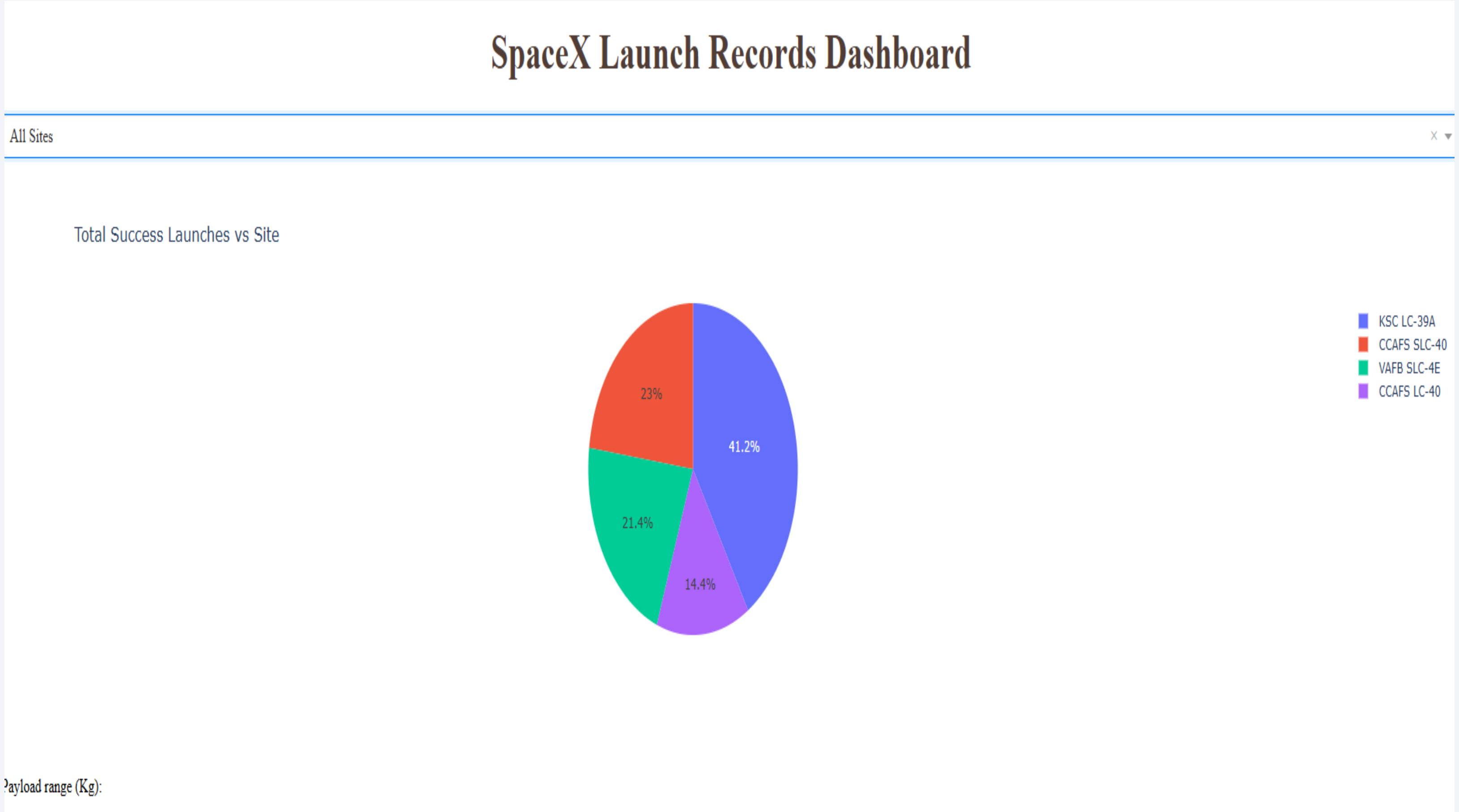


Section 4

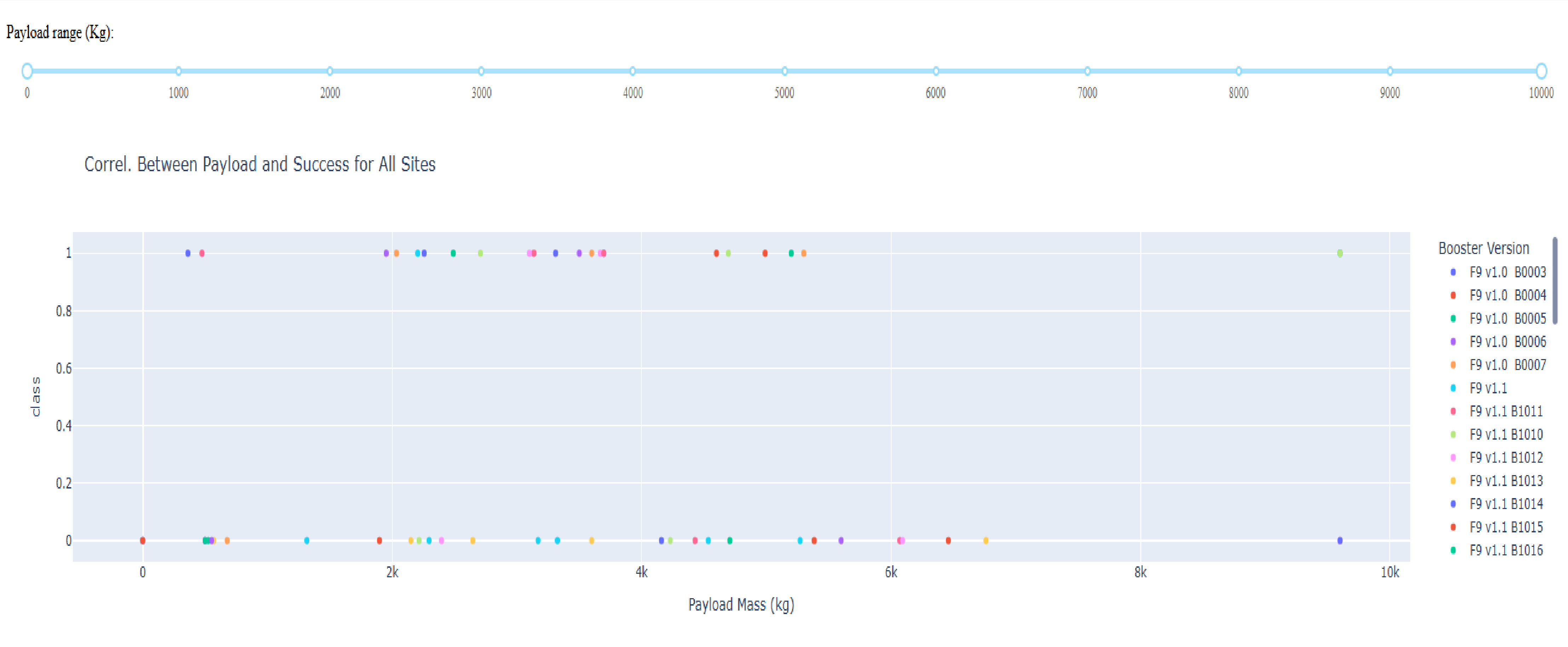
Build a Dashboard with Plotly Dash



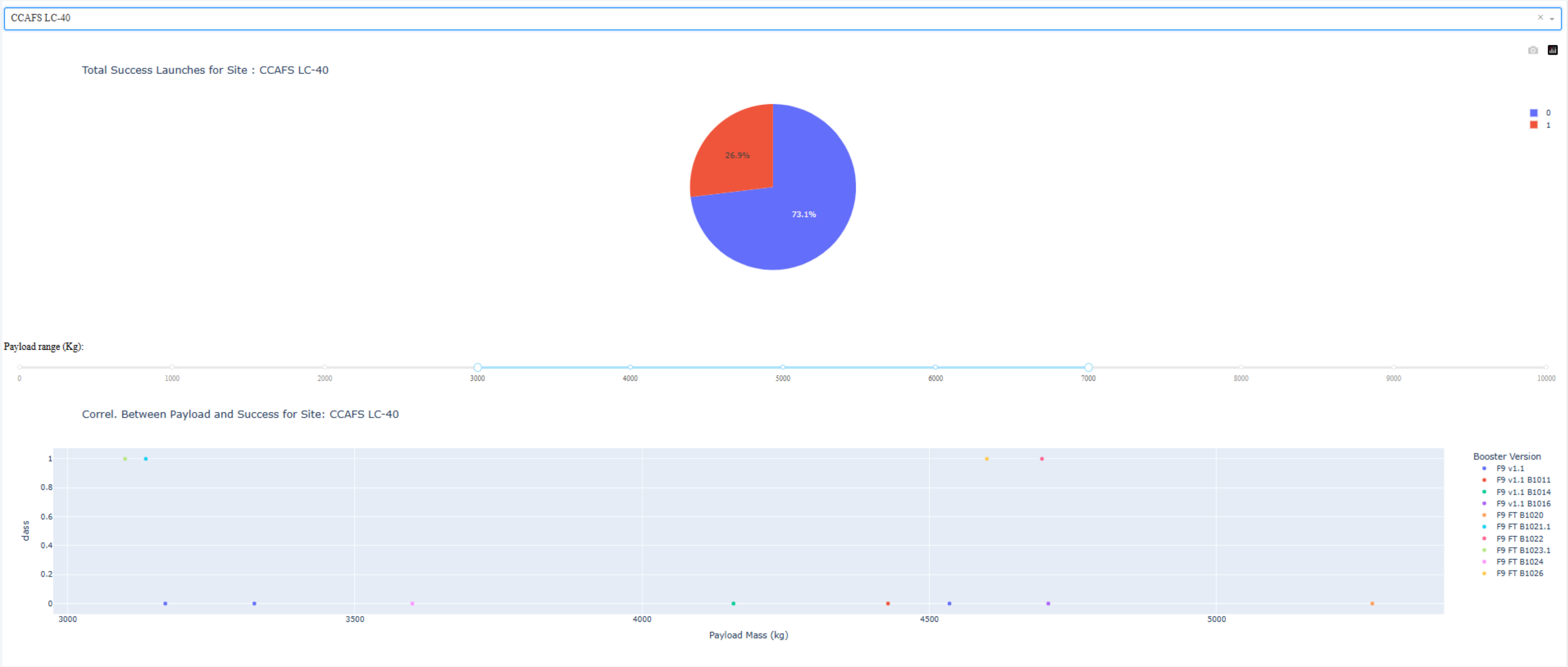
Pie Chart for All Sites



Payload for All Sites



One Site Selected with Limited Payload Range

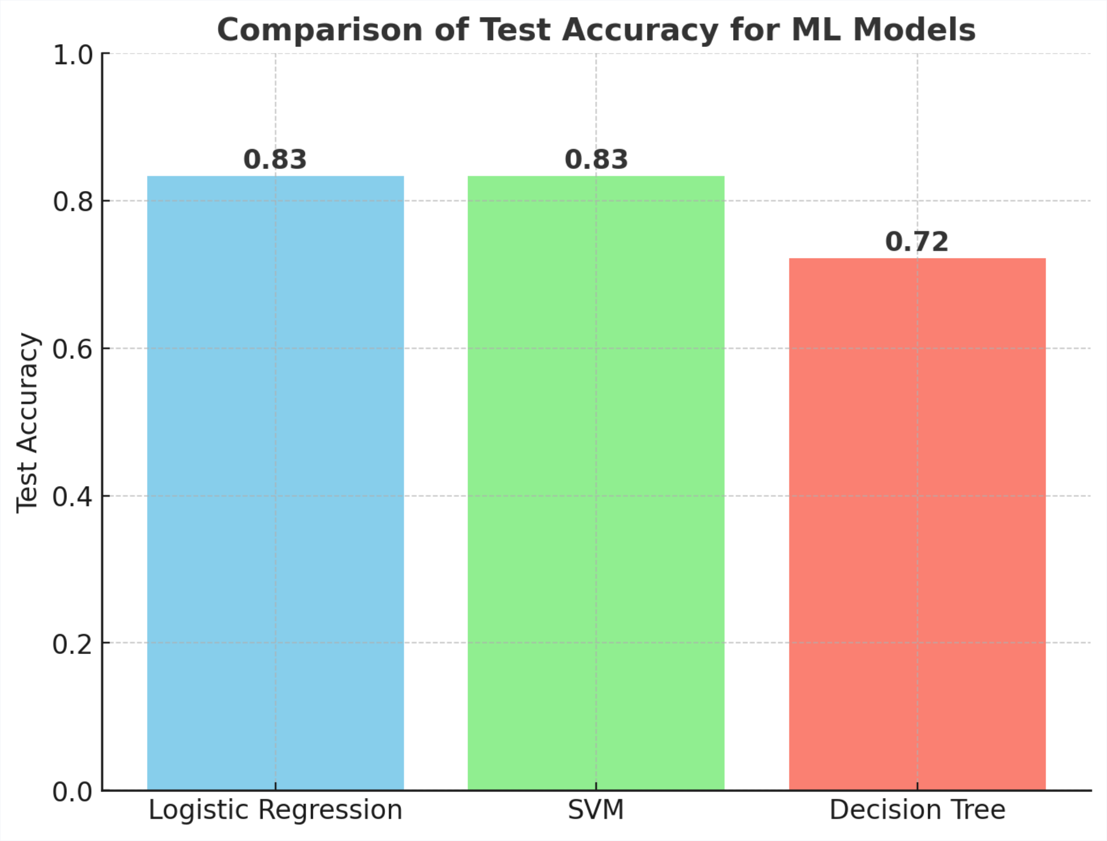


Section 5

Predictive Analysis (Classification)

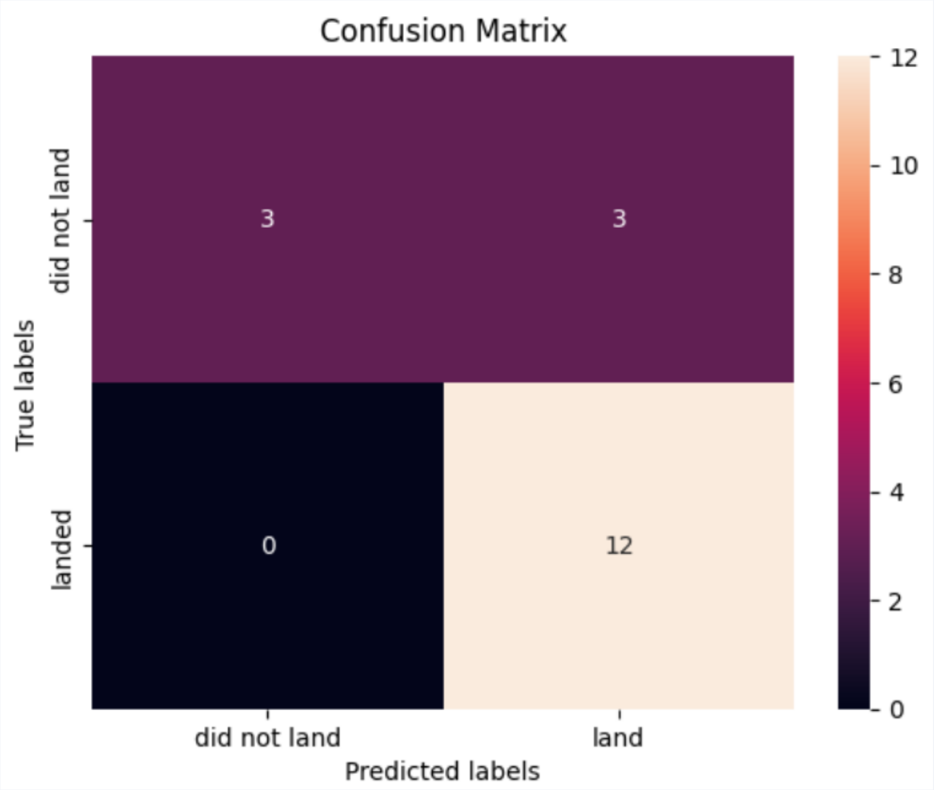
Classification Accuracy

Considering accuracy, stability, and interpretability, SVM emerged as the most reliable model. It provided balanced performance, strong generalization, and the advantage of being straightforward to explain to stakeholders, making it the preferred choice for predicting Falcon 9 first-stage landing success.



Confusion Matrix(SVM)

- High Recall for Successful Landings: SVM correctly predicted all 12 landings (no false negatives). This means the model is highly effective at identifying when a booster lands successfully.
- Moderate Performance for Failures: Out of 6 failed landings, the model correctly predicted 3 and misclassified 3 as landings (false positives). This shows that failures are harder for the model to distinguish.
- Balanced Accuracy (~83%): With 15/18 predictions correct, the accuracy is about 83%, matching the earlier test set accuracy.



Conclusions

- Best Models were KNN and SVM
- Reused Boosters perform better
- Orbits like LEO, ISS have higher success rates than GTO
- Heavier payloads reduce success probability

Appendix

- Special thanks for Peers and Instructers

Thank you!