Winning Space Race with Data Science

<Zahra Vahidi Ferdousi>
<December 2025>

Outline

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

Executive Summary

Summary of methodologies

Data Collection: utilized SpaceX Rest API and Web Scraping (BeautifulSoup) to gather Falcon 9 launch data.

Data Wrangling: Processed using Pandas to handle missing values, filter for Falcon 9, and convert categorical variables (One-Hot Encoding).

EDA: Conducted using SQL queries for statistical analysis and Matplotlib/Seaborn for data visualization.

Interactive Analytics: Built geospatial maps using Folium and an interactive dashboard using Plotly Dash.

Predictive Analysis: Trained and tuned classification models (Logistic Regression, SVM, KNN, Decision Tree) to predict landing success.

Summary of all results

- Exploratory analysis revealed a strong correlation between success rates and time (improving yearly).
- Heavier payloads generally have a high success rate, though specific orbits (like GTO) pose different challenges.
- Launch sites are strategically located near coastlines for safety.
- Predictive models achieved an accuracy score of approximately **83.33%** on the test data, with Decision Trees, KNN, SVM, and Logistic Regression performing similarly after hyperparameter tuning.

Introduction

Project background and context

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars.
- The savings are due to SpaceX's ability to reuse the first stage of the rocket.

Problems you want to find answers

- Can we predict if the first stage will land successfully before the launch?
- How do variables like Payload Mass, Orbit Type, and Launch Site affect the landing success?
- This prediction determines the cost of a launch: if the first stage lands, the cost is significantly lower.

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

We collect data from two different sources:

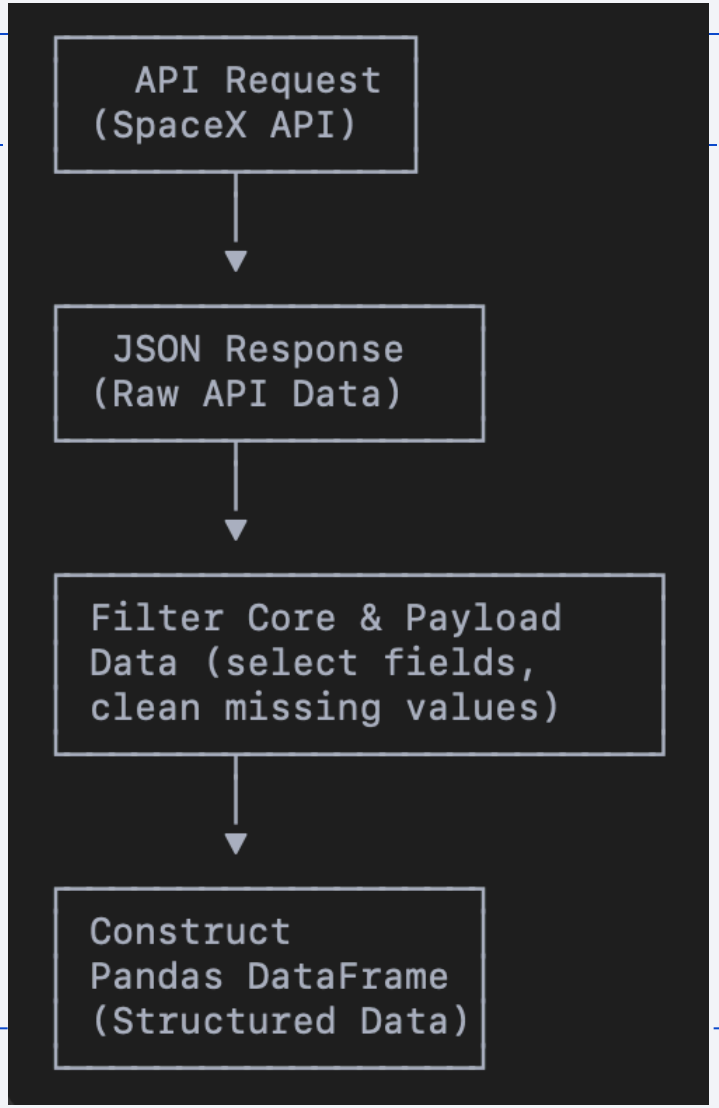
- **Source 1:** Request data from SpaceX REST API (https://api.spacexdata.com/v4/launches/past)
- **Source 2:** Scrape launch data from the Wikipedia page "List of Falcon 9 and Falcon Heavy launches" using ***BeautifulSoup***

Data Collection – SpaceX API

Key Phrases:

request.get()
Json_normalize()
Filtering for Falcon9

GitHub URL: (ihttps://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb).



flowchart of data collection request

Data Collection - Scraping

- **Key Phrases:**
  - BeautifulSoup
  - find_all(‘table’)
  - Iterating table rows (tr) and cells (td)
- GitHub URL : (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb)

```
Start
  ↓
Define Target URL (Wikipedia SpaceX Launches page)
  ↓
Send HTTP Request (requests.get)
  ↓
Receive HTML Response
  ↓
Parse HTML Content (BeautifulSoup)
  ↓
Locate Tables (find_all('table'))
  ↓
Select Relevant Launch Table
  ↓
Extract Table Headers (th elements)
  ↓
Clean & Store Column Names
  ↓
Iterate Over Table Rows (tr elements)
  ↓
Extract Cell Data (td elements)
  ↓
Clean & Process Data
  ├─ Date & Time parsing
  ├─ Booster version extraction
  ├─ Payload & mass extraction
  ├─ Landing outcome parsing
  ↓
Store Each Row in Dictionary
  ↓
Append Dictionaries to List
  ↓
Create Pandas DataFrame
  ↓
Handle Missing / Inconsistent Values
  ↓
Final Clean DataFrame Ready for Analysis
  ↓
End
```

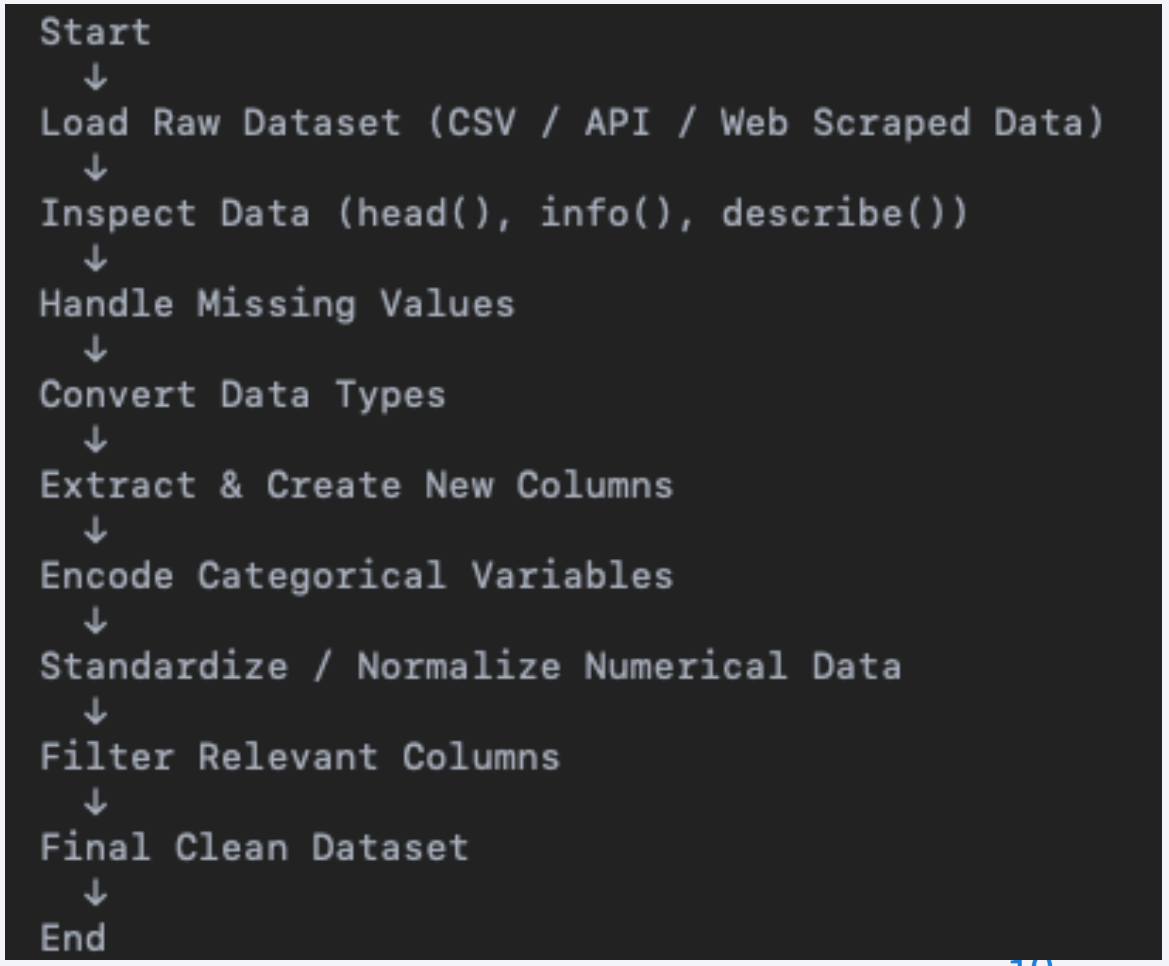
flowchart of web scraping

Data Wrangling

-Process:

- Filtered data to include only Falcon 9 launches.
- Replaced missing Payload Mass values with the mean mass.
- Created a ***Class*** column: 1 for Success, 0 for Failure.

- GitHub URL (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb)



Flowchart of data wrangling

EDA with Data Visualization

- Created scatter plots to visualize relationships between Payload, Flight Number, and Launch Site. Used Bar charts to compare success rates across different Orbit types.
- GitHub URL : (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/edadataviz%20(1).ipynb)

EDA with SQL

- Summary of SQL queries:
  - Unique launch sites.
  - Total payload mass carried by NASA.
  - Success rates per booster version.
  - Ranking landing outcomes by count.
- GitHub URL: (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(2).ipynb)

Build an Interactive Map with Folium

- Added markers for launch sites, color-coded markers for launch outcomes (Green=Success, Red=Failure), and ***MousePosition*** plugins to calculate distances to coastlines and railways.
- GitHub URL : (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location%20(1).ipynb)

Build a Dashboard with Plotly Dash

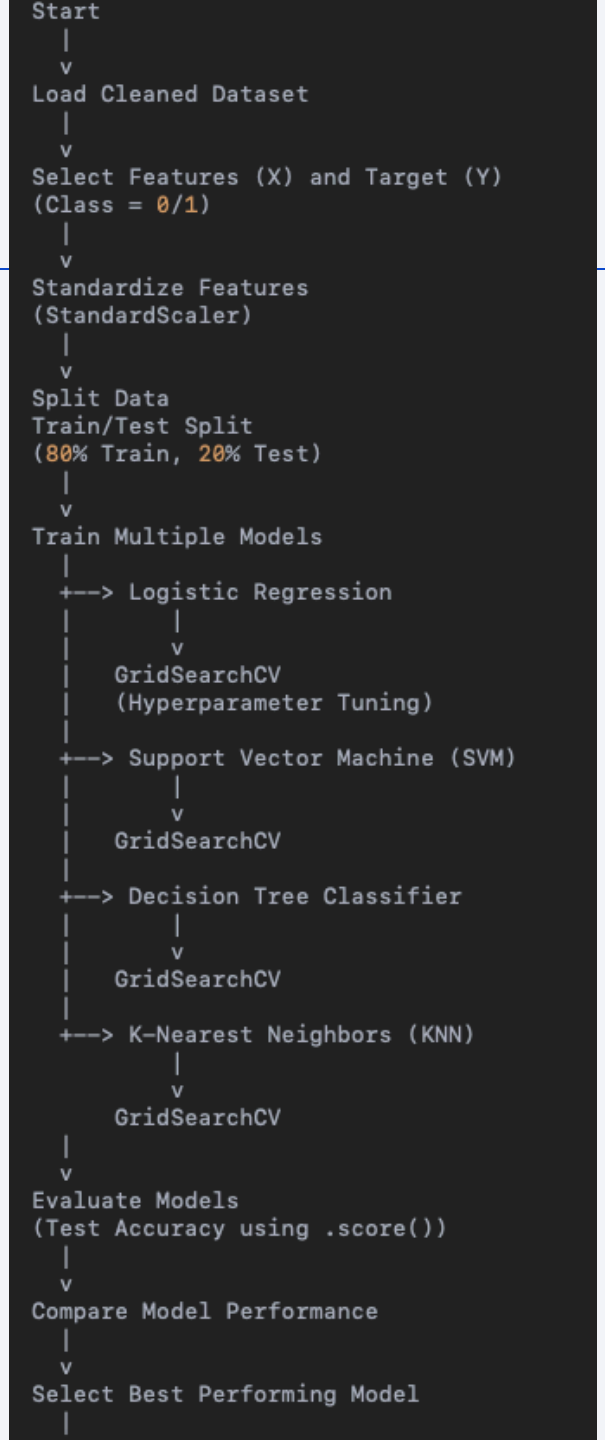
Created a dashboard with:
- A Dropdown to select specific Launch Sites.
- A Range Slider to filter by Payload Mass.
- A Pie Chart showing total success vs. failure.
- A Scatter plot showing correlation between Payload Mass and Success.

GitHub URL : (https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/spacex-dash-app%20(1).py)

Predictive Analysis (Classification)

- Standardized the data (StandardScaler).
- Split data into training and testing sets (train_test_split).
- Trained Logistic Regression, SVM, Decision Tree, and KNN.
- Used ***GridSearchCV*** to find the best hyperparameters.

GitHub URL:(https://github.com/h4y4h0o/IBM_Applied_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb)



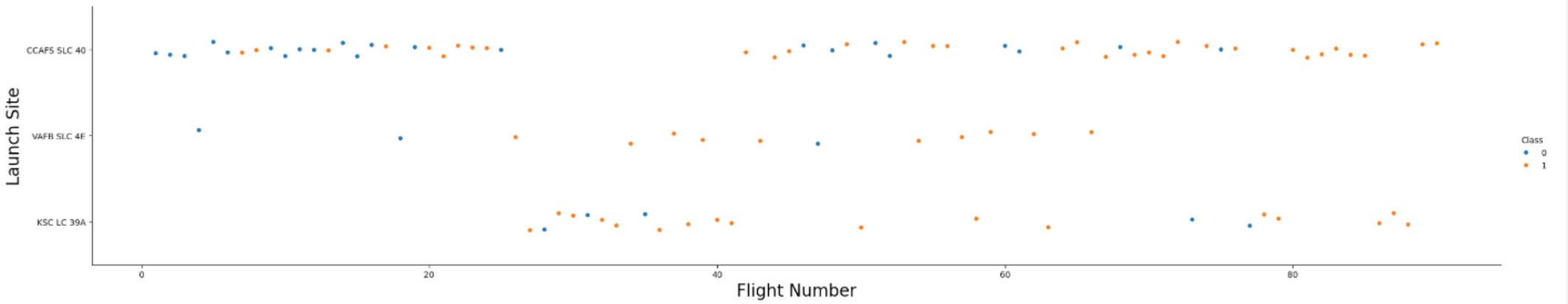
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

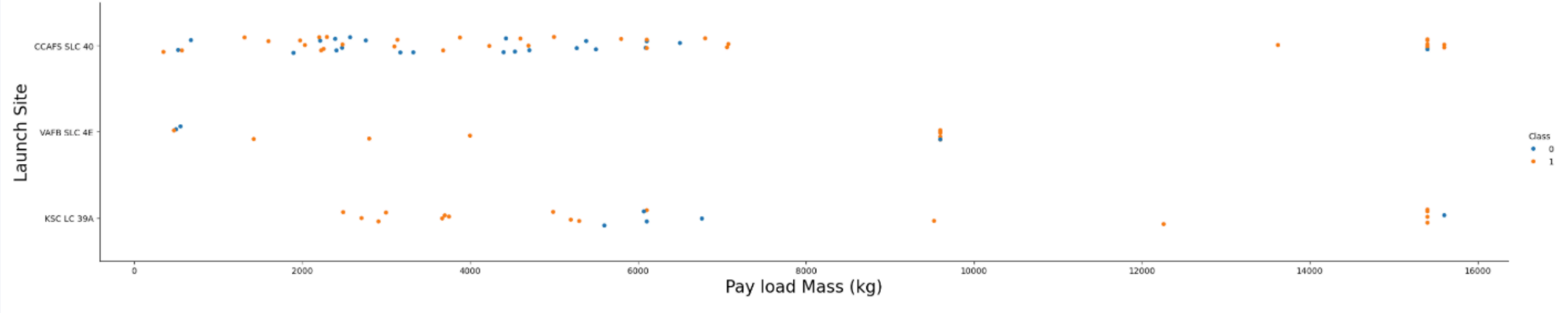


- scatter plot of Flight Number vs. Launch Site

Observation: As flight numbers increase (indicating time progression), the success rate generally increases.

Explanation: CCAFS SLC-40 has the highest volume of launches. VAFB SLC-4E is used less frequently but has a high success rate in later flights.

Payload vs. Launch Site



Scatter plot of Payload vs. Launch Site

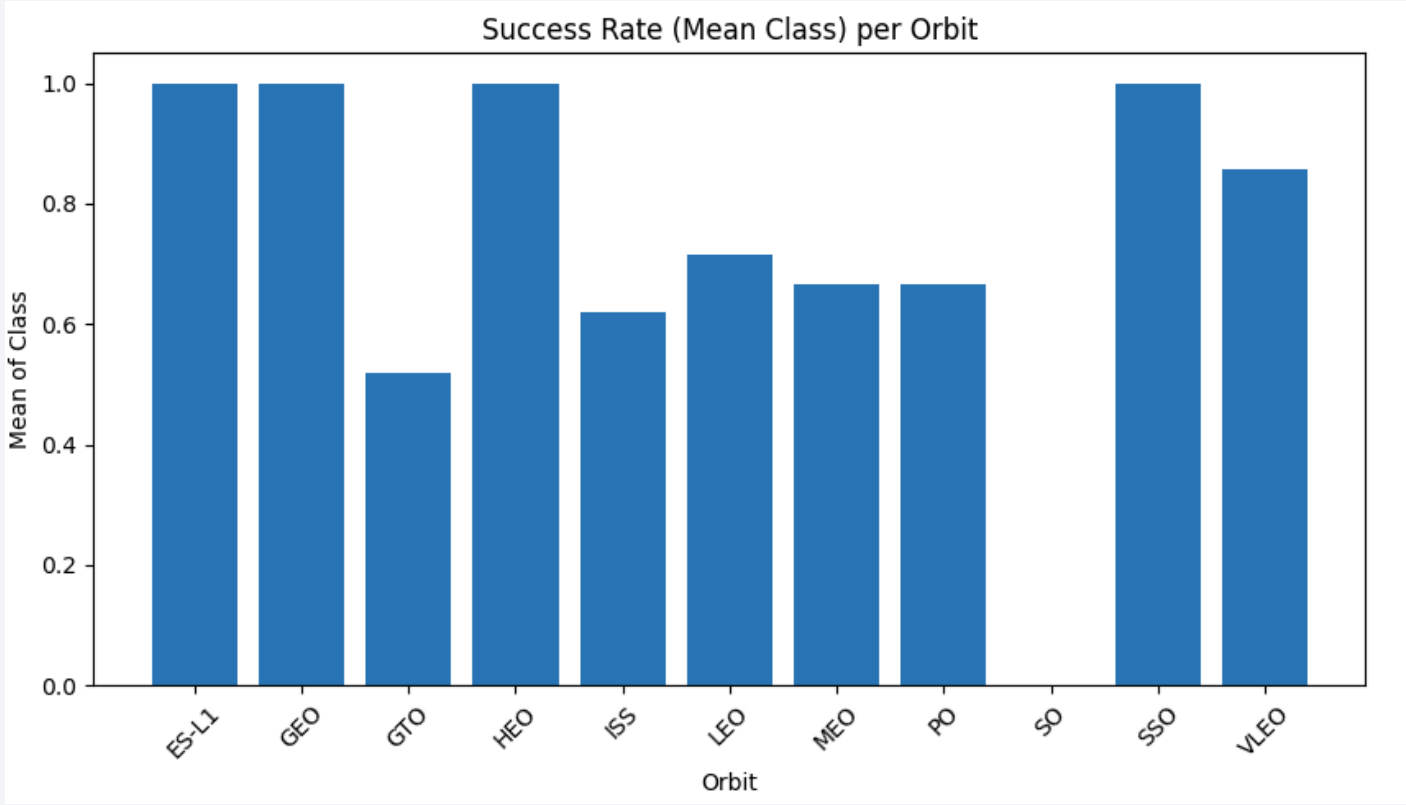
Observation: The VAFB-SLC launch site appears to handle lower payload masses compared to CCAFS SLC-40 and KSC LC-39A.

Explanation: There is no distinct cluster indicating that higher payload mass ensures failure; success is distributed across payload ranges at the major sites.

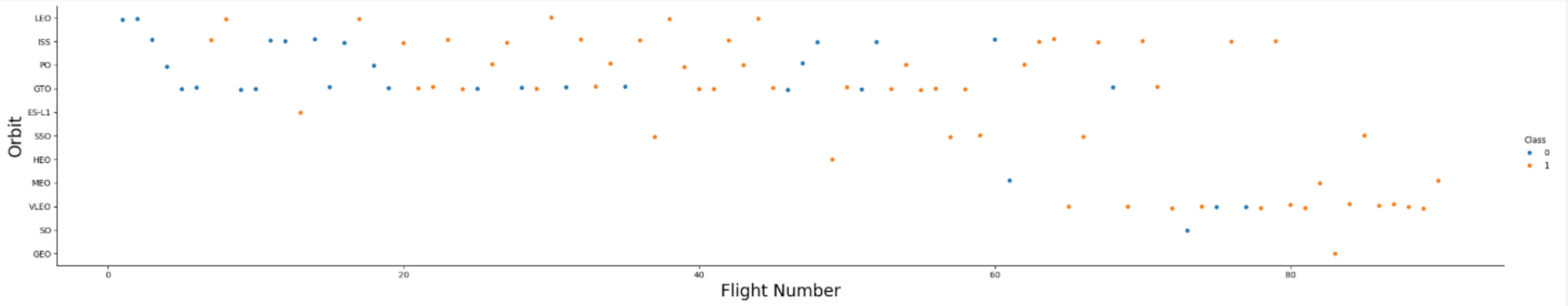
Success Rate vs. Orbit Type

Observation: ES-L1, GEO, HEO, and SSO orbits have the highest success rates (near 100%).

Explanation: Orbits like SO (Sun-Synchronous) showed lower success rates in this dataset, while the GTO (Geostationary Transfer Orbit) is the most popular orbit type with a mixed success rate.



Flight Number vs. Orbit Type

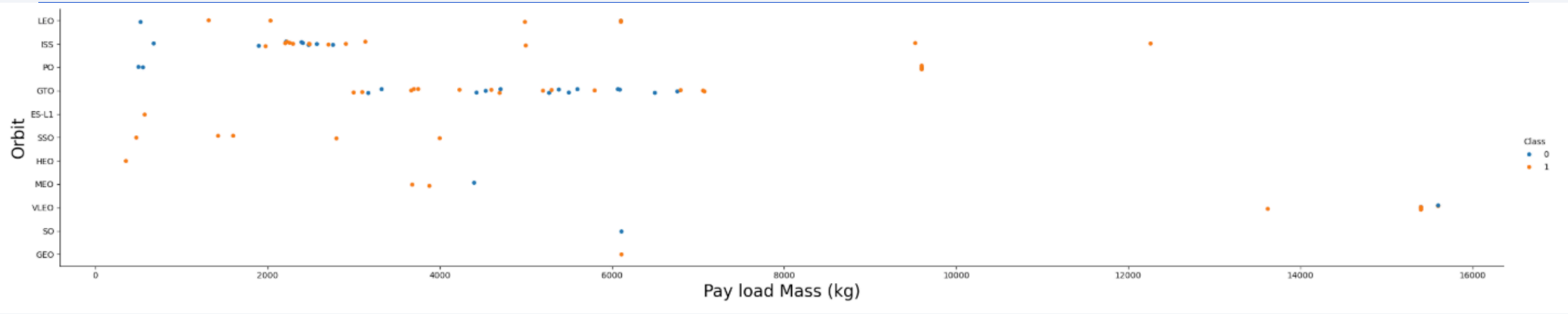


Scatter plot of flight numbers vs. orbit type

Observation: In the early years (low flight numbers), launches were primarily to LEO.

Explanation: As flight numbers increased, SpaceX expanded to more complex orbits like GTO and ISS (VLEO).

Payload vs. Orbit Type



Scatter plot of payload vs. Orbit Type

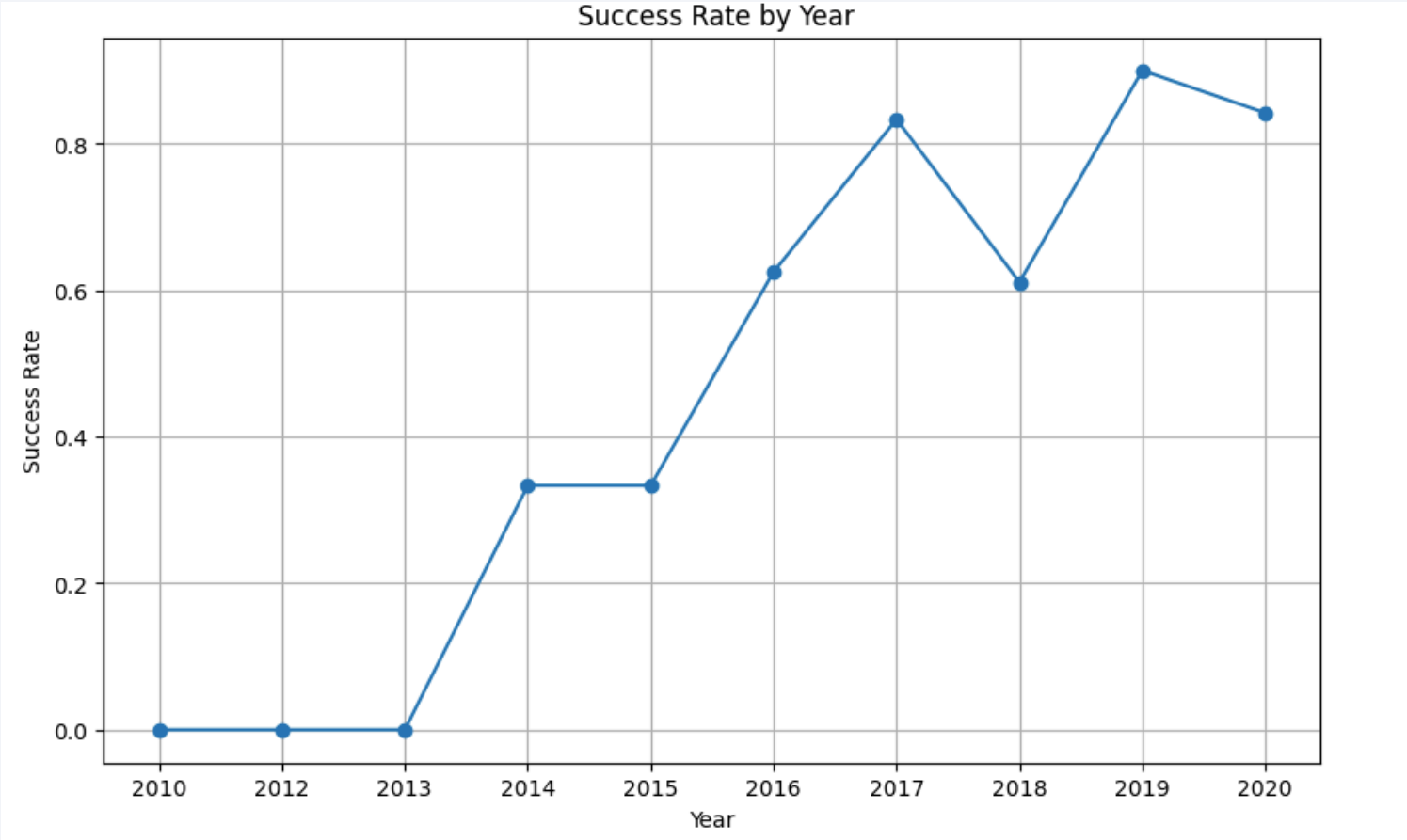
Observation: Heavy payloads are typically destined for GTO and ISS (LEO).

Explanation: Polar, SSO, and HEO orbits typically carry lighter payloads.

Launch Success Yearly Trend

Observation: The success rate has steadily increased from 2013 to 2020.

Explanation: 2013 had a 0% success rate (for landing), but by 2019/2020, the success rate stabilized near 90%, proving the reliability of the reusable technology.



Line plot of launch success yearly trend

All Launch Site Names

- 4 Sites found: **CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E**
- Query : select distinct "Launch_Site" from SPACEXTABLE

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Query : 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

5 records where launch sites begin with `CCA`

Total Payload Mass

- Calculate the total payload carried by boosters from NASA : **45596 KG**
- Query: select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer"=="NASA (CRS)"

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:
  2928.4

- Query: select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like "F9 v1.1 »

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad : **2015-12-22**
- Query : select MIN("Date") from SPACEXTABLE where "Landing_Outcome"="Success (ground pad) »

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Query: select "Booster_Version" from SPACEXTABLE where "Landing_Outcome"="Success (drone ship)" 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

- Calculate the total number of successful and failure mission outcomes: ***100***
- Query: select COUNT("Mission_Outcome") as "Success" from SPACEXTABLE where "Mission_Outcome" like ‘Success%'

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:
- Query: select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("

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

- Present your query result with a short explanation here

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

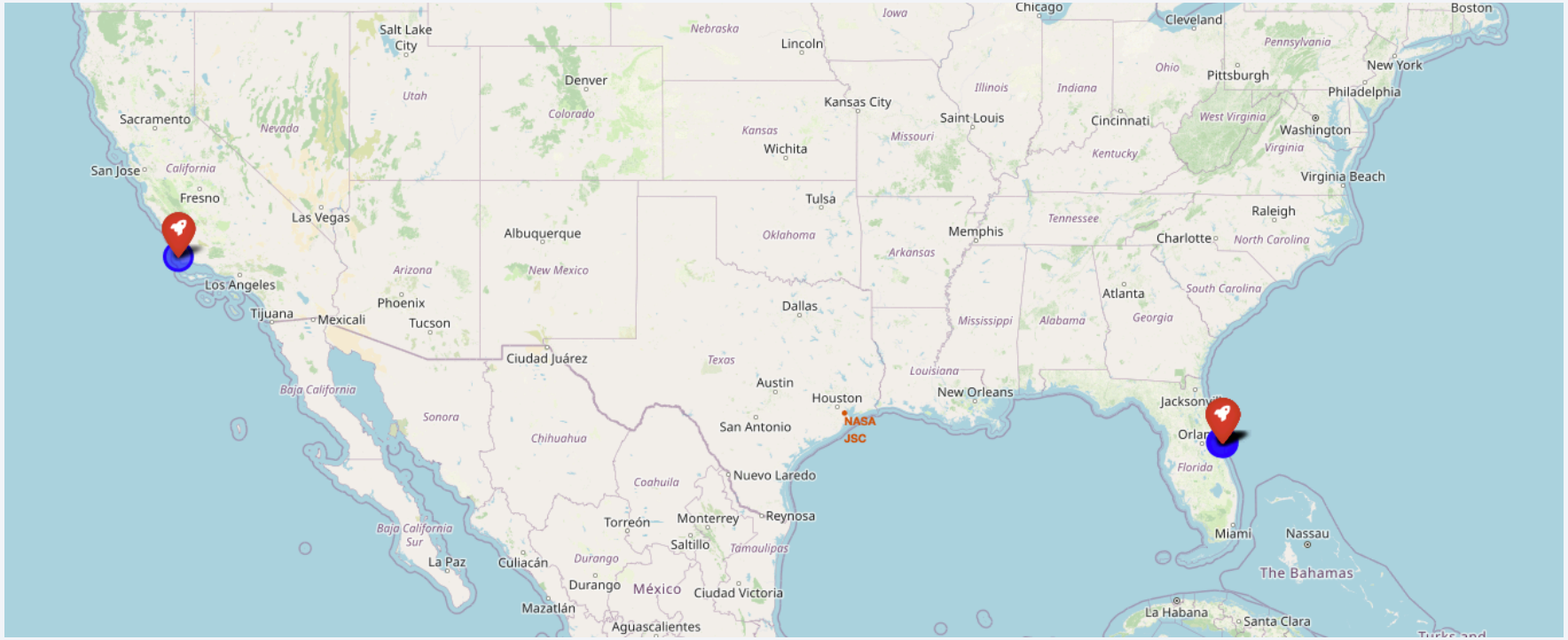
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

Launch site locations



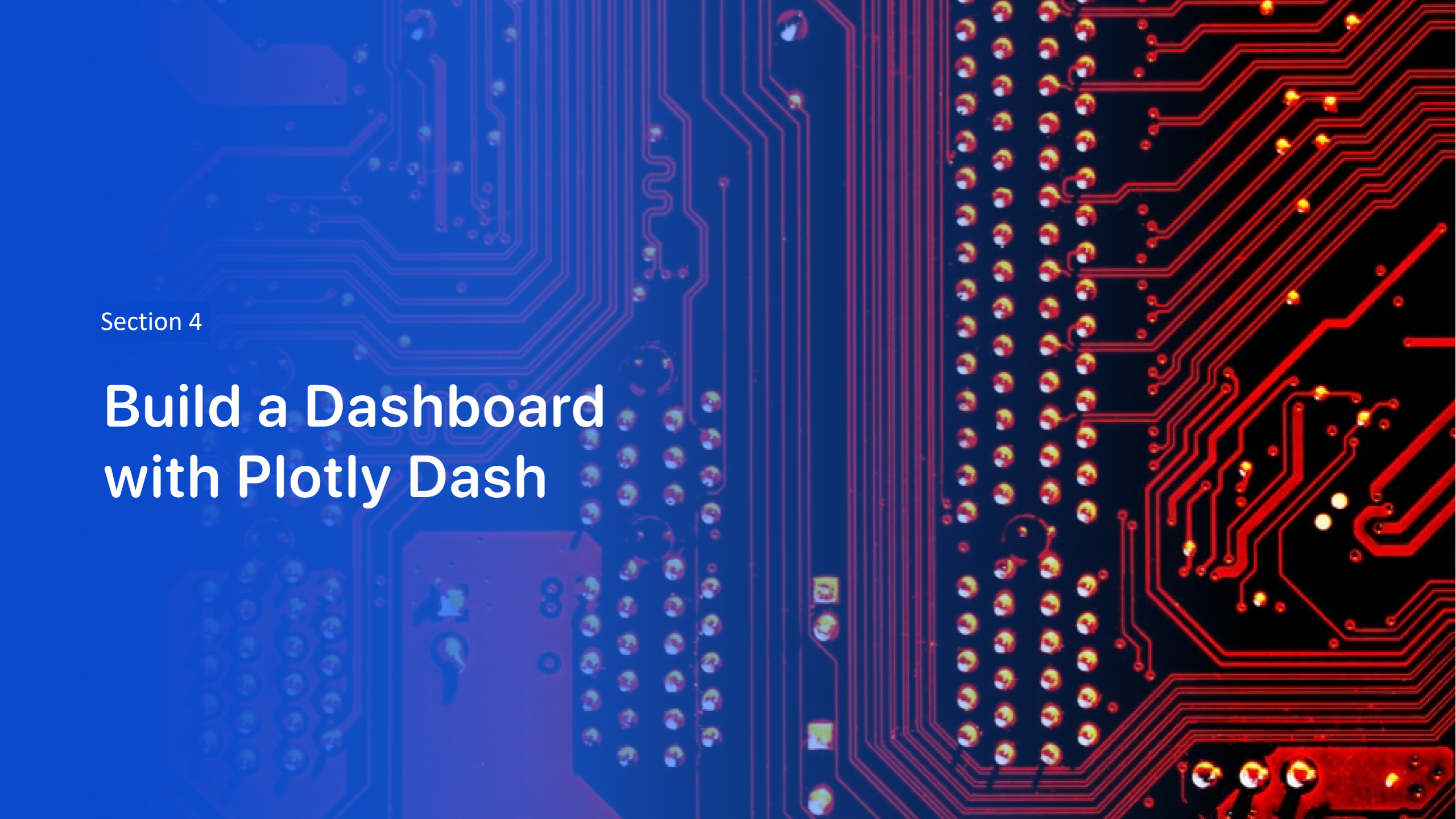
Explanation: All launch sites are located close to the equator and on coastlines to take advantage of the Earth's rotation and for safety (dropping debris into the ocean).

Launch Success/Failure Clusters

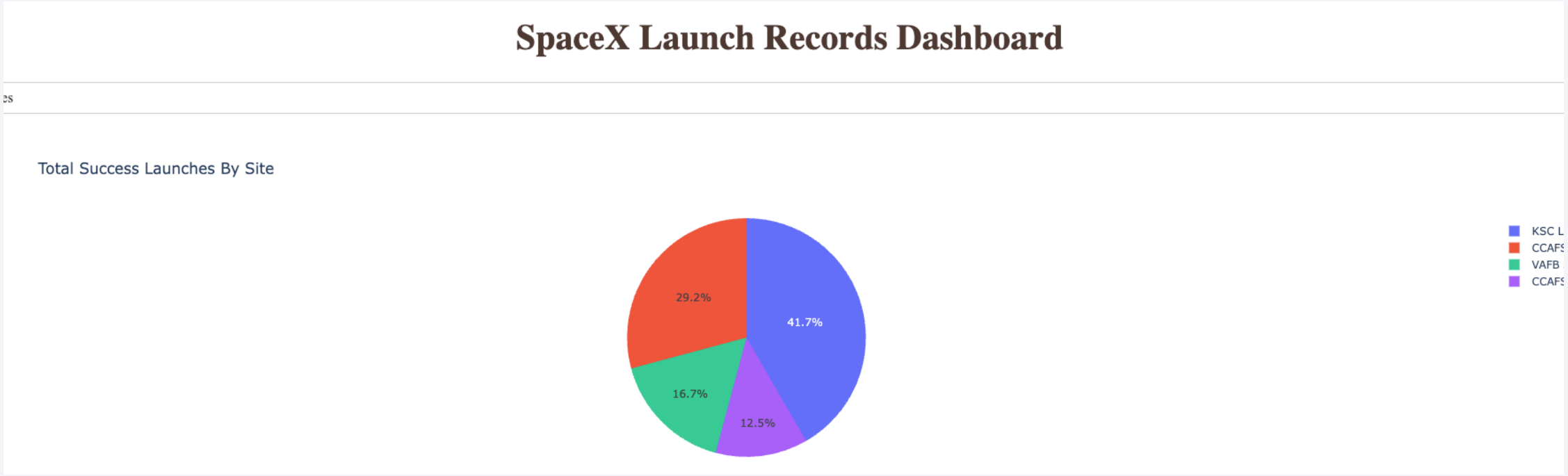
Explanation: The map shows clusters of Green markers (Success) and Red markers (Failure). The KSC LC-39A site shows a very dense cluster of green markers, indicating high reliability.

Proximity to Coastlines and Railways

Explanation: Measurement lines show that launch sites are very close to the coastline (<1 km) but maintain a safe distance from city centers. They are also located near railways for heavy equipment transport.

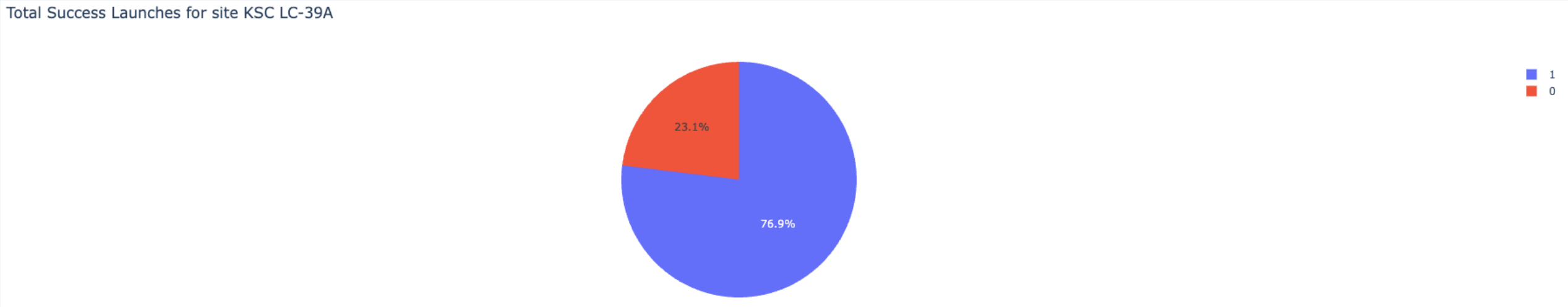

Section 4
Build a Dashboard with Plotly Dash

Success Count Pie Chart



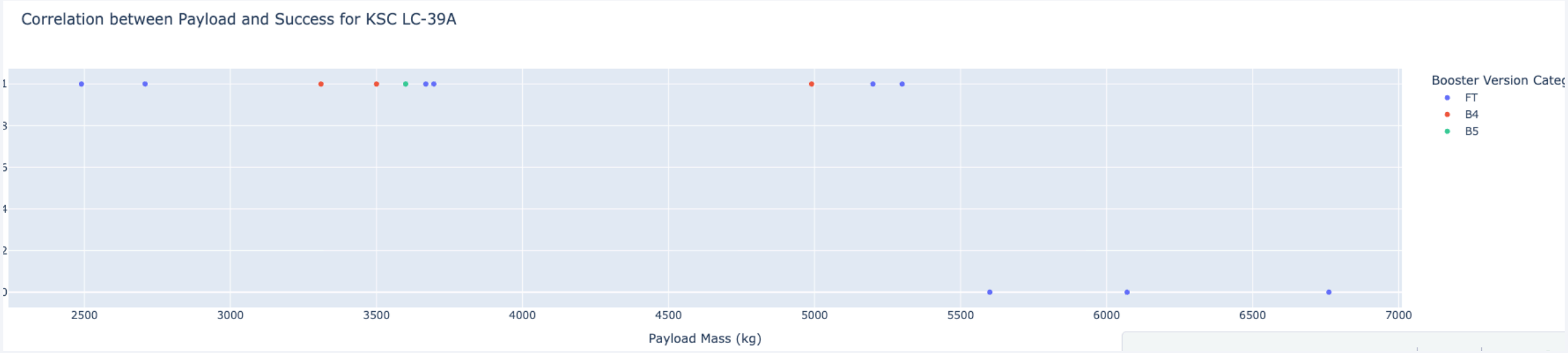
Explanation: When "All Sites" is selected, the pie chart shows the contribution of each site to total successes. KSC LC-39A is usually the largest contributor.

Success Rate for KSC LC-39A



Explanation: When a specific site (e.g., KSC LC-39A) is selected, the chart shows the split between Success (1) and Failure (0). KSC typically shows a success rate of over 75%.

Correlation between Payload and success



Explanation: The scatter plot shows that the Booster Version "FT" (Full Thrust) has the highest success rate across various payload masses. Payloads between 2000kg and 5000kg have a particularly high density of successful landings.

Section 5

Predictive Analysis (Classification)

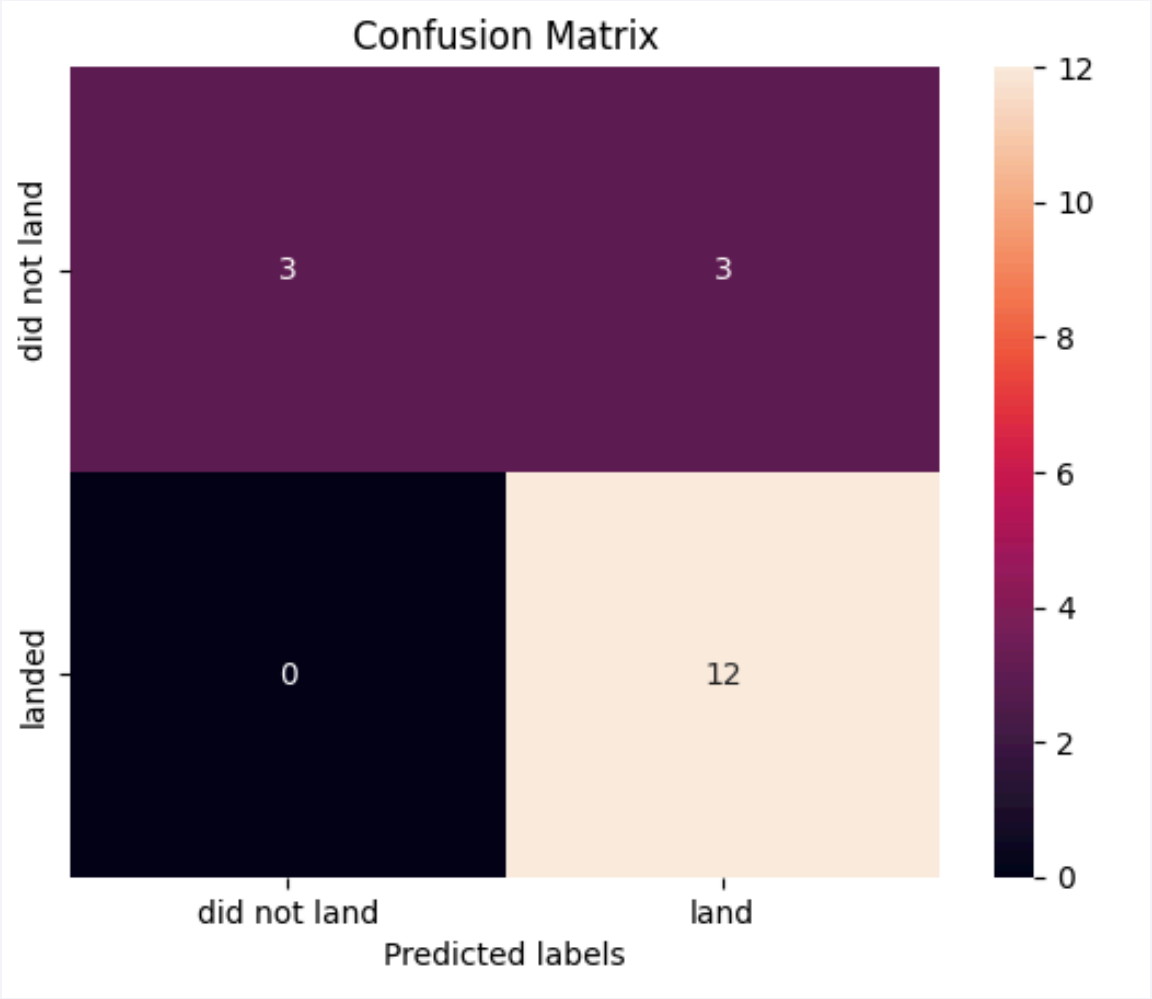
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

True Positives: The model correctly predicted the rocket would land.
False Positives: The model predicted a landing, but it crashed (Type I error).
False Negatives: The model predicted a crash, but it landed (Type II error).
The model generally has a high True Positive rate but may struggle with False Positives due to the imbalance of successful flights in the data.



Conclusions

- **Model Performance:** We can predict the success of a rocket landing with roughly **83% accuracy** using classification models like SVM or Logistic Regression.
- **Temporal Trends:** SpaceX's success rate has improved drastically over time, stabilizing near 90% in recent years.
- **Orbit Factors:** Launches to LEO, ISS, and GTO have high success rates; however, specific conditions (like high payload mass to GTO) still present risks.
- **Strategic Geography:** Launch sites are optimized for physics (equator proximity) and logistics (coastlines/railways), which is confirmed by geospatial analysis.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!