IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tsuzumi Sato
Aug 13th, 2025

Outline

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

Executive Summary

In this project, Falcon 9 launch data was collected from the SpaceX REST API and Wikipedia tables using requests, BeautifulSoup, and pandas.read_html(). The dataset was cleaned by handling missing values, normalizing site names, converting dates, and creating a binary landing success variable. Exploratory data analysis with matplotlib, seaborn, and Folium identified trends in payload, launch site, and orbit. Features were engineered through date extraction, one-hot encoding, and standardization of numeric variables. Multiple classification models—Logistic Regression, SVM, Decision Tree, and KNN—were trained and tuned via 10-fold cross-validation (GridSearchCV). The tuned Decision Tree Classifier achieved the highest test accuracy (~0.83–0.85), with launch site, payload mass, and orbit emerging as key predictors. The model provides reliable landing success probabilities, enabling more accurate cost estimates and supporting strategic mission planning.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. We create a machine learning pipeline to predict if the first stage will land.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API(JSON Data)
  - Wikipedia tables using webscraping
- Perform data wrangling
  - Handle missing values, normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

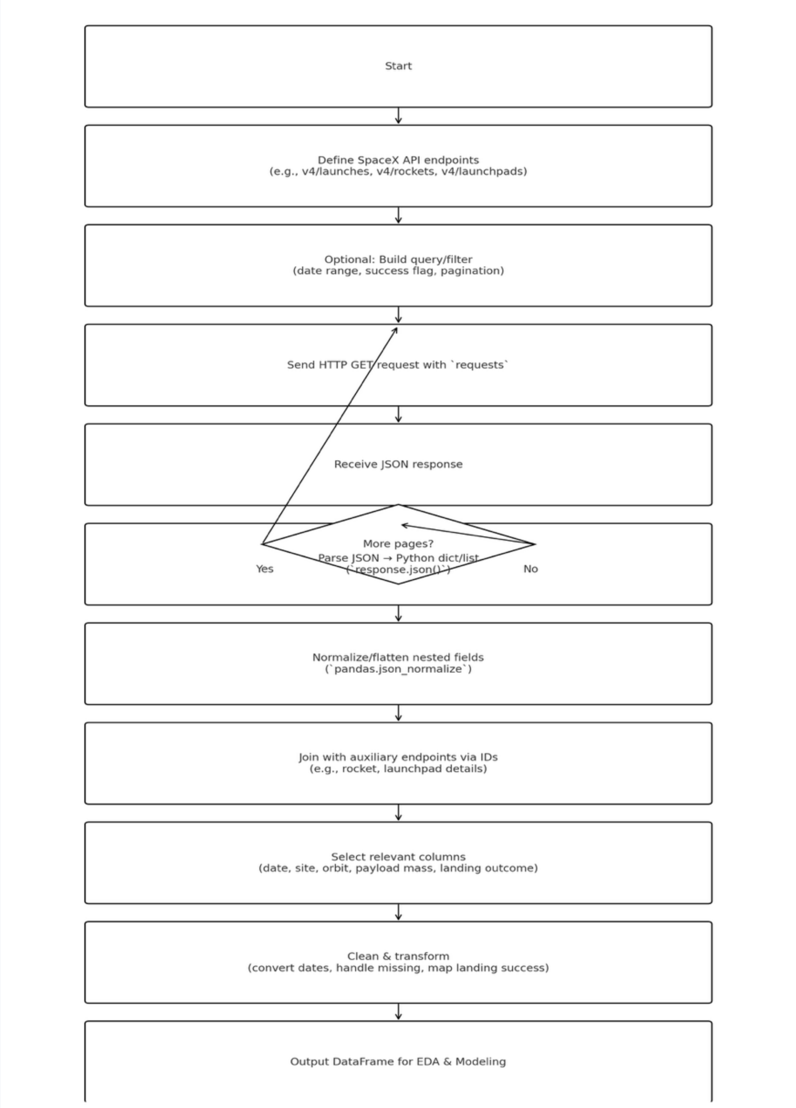
Data Collection

- Data Sets were collected using the following methods:
  - SpaceX REST APIs for launch data
  - Wikipedia mission tables (historical launches)

- Tools used:
  - Requests for API calls
  - BeautifulSoup for HTML parsing
  - Pandas.readhtml( ) for table extraction

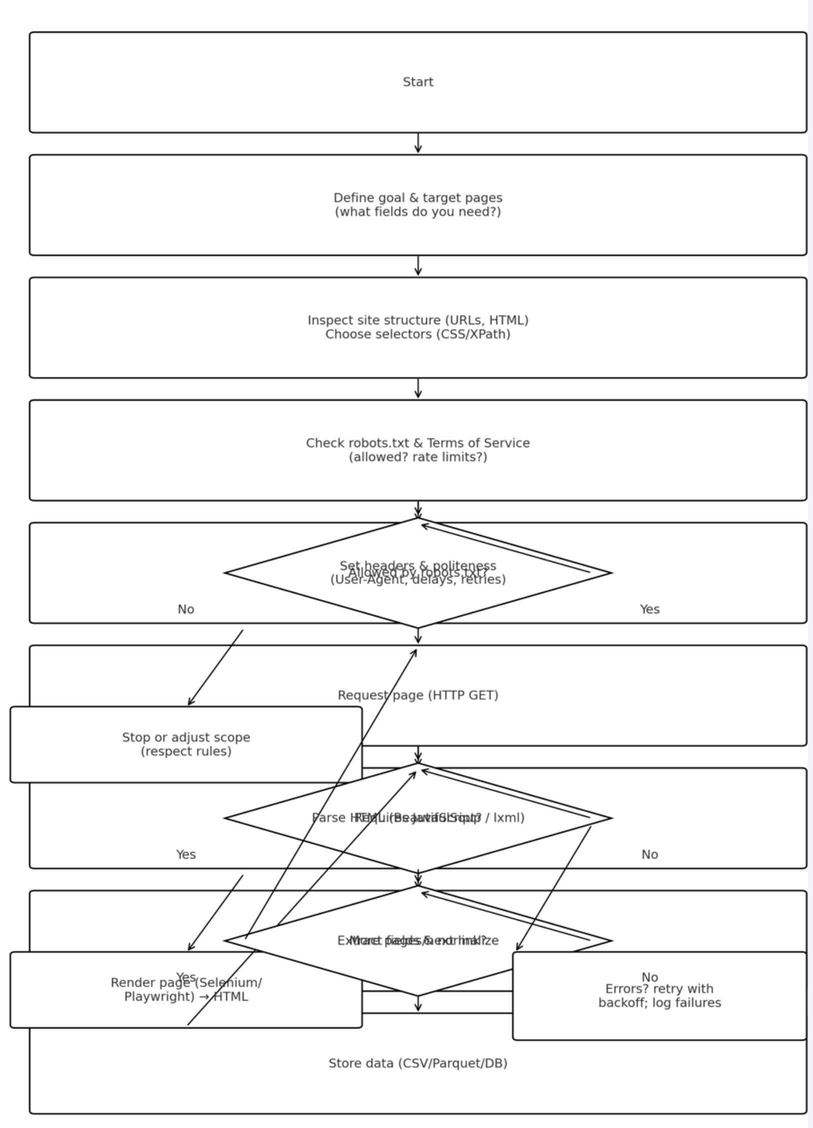
Data Collection – SpaceX API

- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/data-collection-api.ipynb



Data Collection - Scraping

- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/Webscraping.ipynb



Data Wrangling

- **Handling Missing Data:** Drop rows or impute where possible.
- **Standardization:** Normalize site names and mission outcomes.
- **Date Conversion:** Convert launch dates to datetime format.
- **Target Variable:** Create binary LandingOutcome column (1 = success, 0 = failure).
- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Used matplotlib and seaborn for trend analysis
- Folium maps for geospatial analysis to display launch sites and success rates
- Heatmaps for correlation analysis to identify relationships between features
- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/EDA%20Visualization%20Lab.ipynb

EDA with SQL

- Some of the SQL used to understand the data set
  - Finding unique launch sites
  - Finding average payloads for each booster
  - Ranking the count of each boosters' landings.
- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/EDA%20-%20SQL.ipynb

Build an Interactive Map with Folium

- Added markers for each launch site to find geographical patterns.
    - Are launch sites closer to railways? Highways?
    - Are they in close proximity to the city?
- Discover which geographical factor influenced landing successes.
- Notebook URL:
    - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/Launch%20Site%20Location%20with%20Folium.ipynb

Predictive Analysis (Classification)

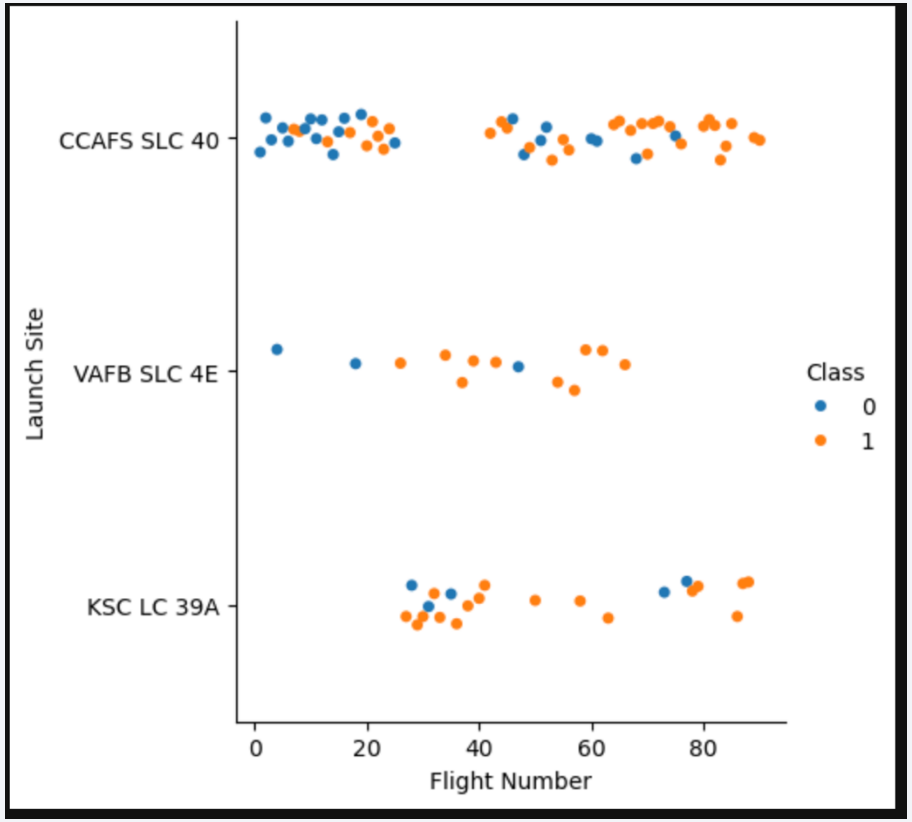
- Tested Logistic Regression, Support Vector Machine, Decision Tree, K-nearest neighbors
- Hyperparameter Tuning: GridSearchCV with CV = 10 for each model
- Notebook URL:
  - https://github.com/TsuzumiTTD/ibm-DataScience-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Section 2

Insights drawn from EDA

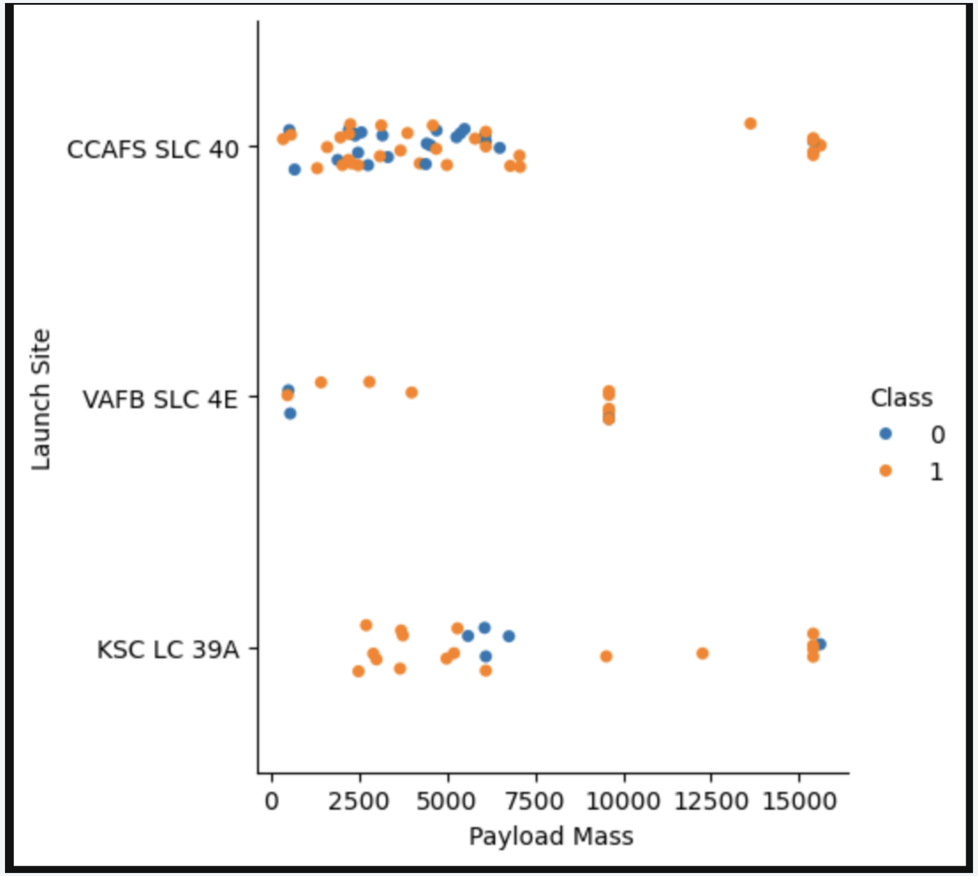
Flight Number vs. Launch Site

- We can observe that for launch site CCAFS SLC 40 and VAFB SLC 4E, the success rate increased with more flight number.



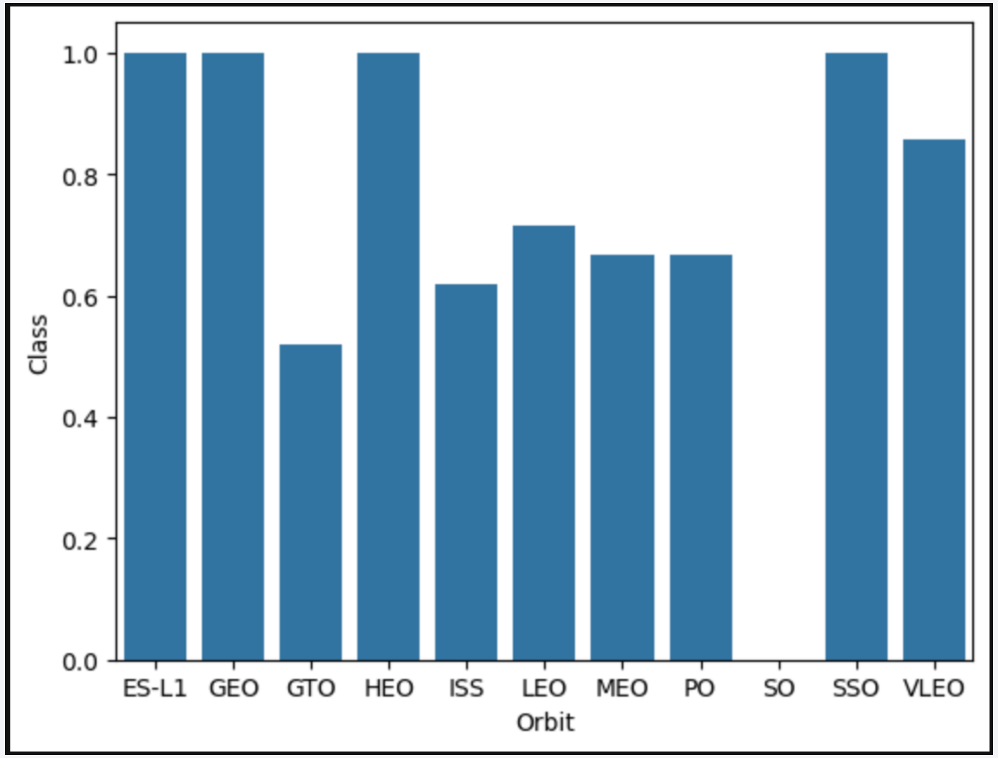
Payload vs. Launch Site

- For VAFB SLC 4E, there were no launches exceeding 10000 payload mass. Also success rate tend to be higher with heavier payloads.
- No correlation observed for the other launch sites.



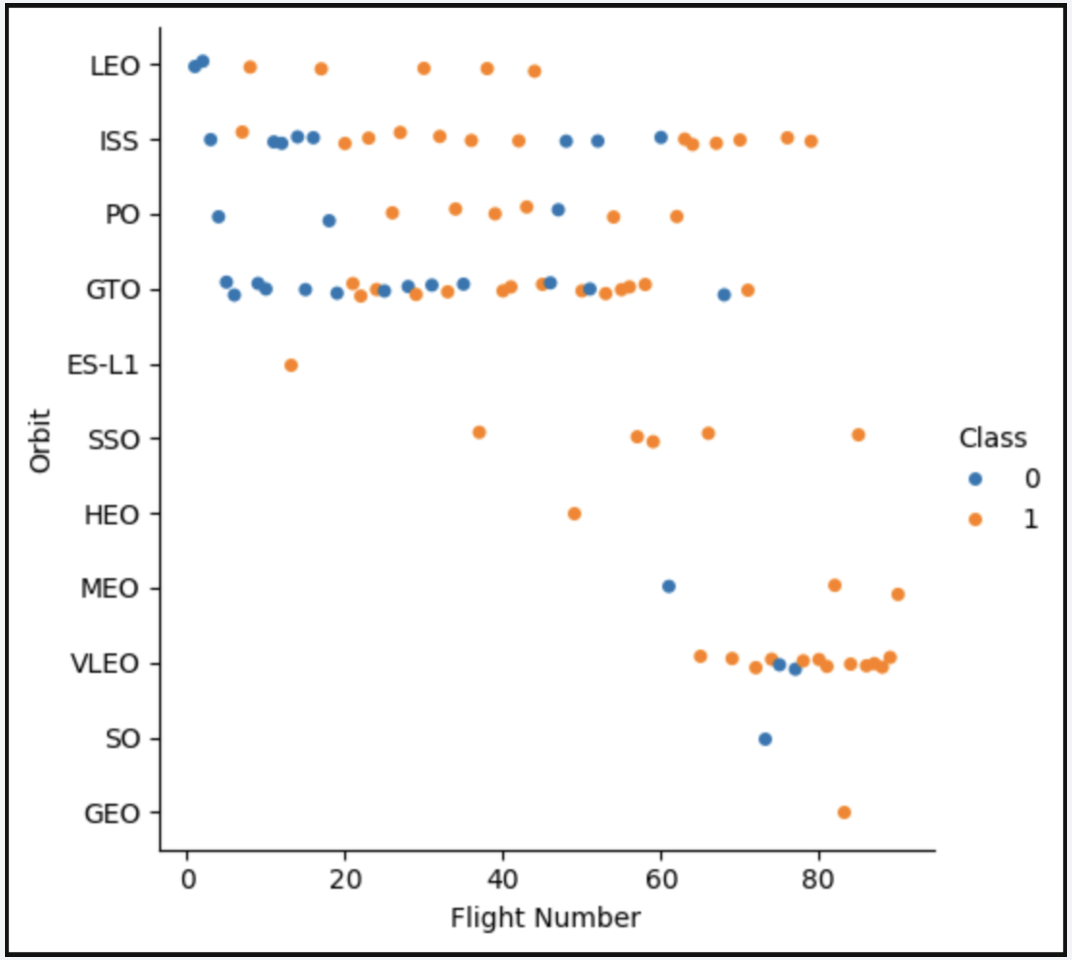
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO tend to have the highest success rates



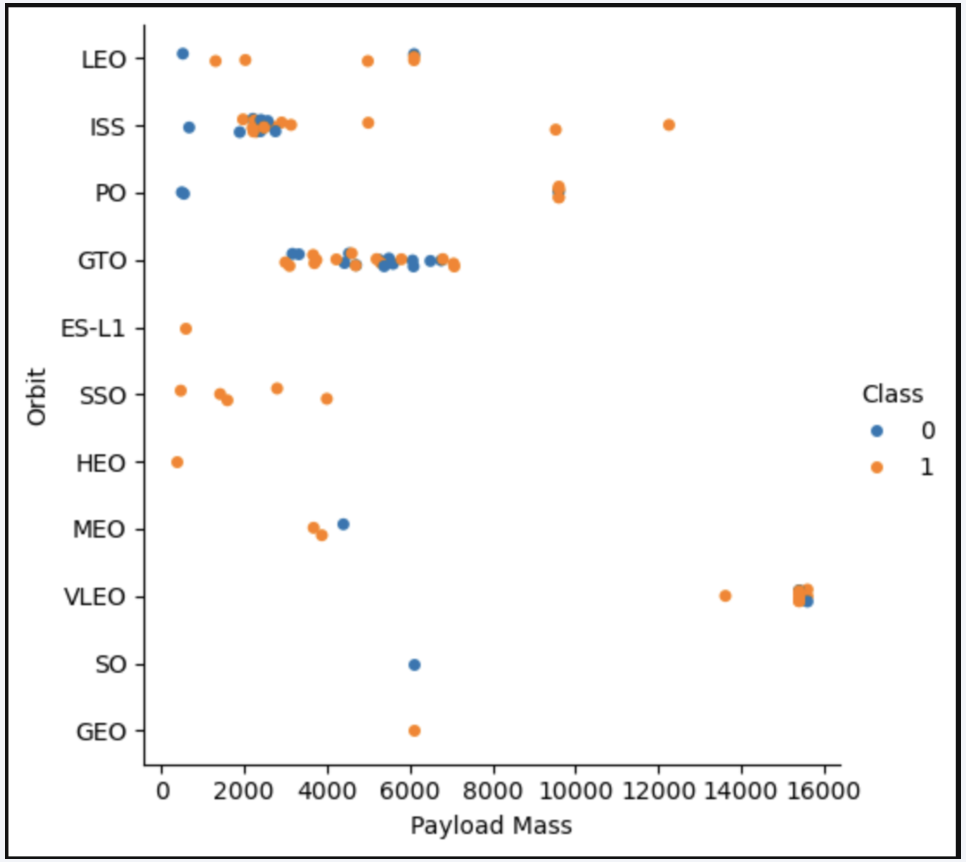
Flight Number vs. Orbit Type

- For LEO orbit, the success rate increases with flight number. Whereas for GTO and ISS, there is no correlation.



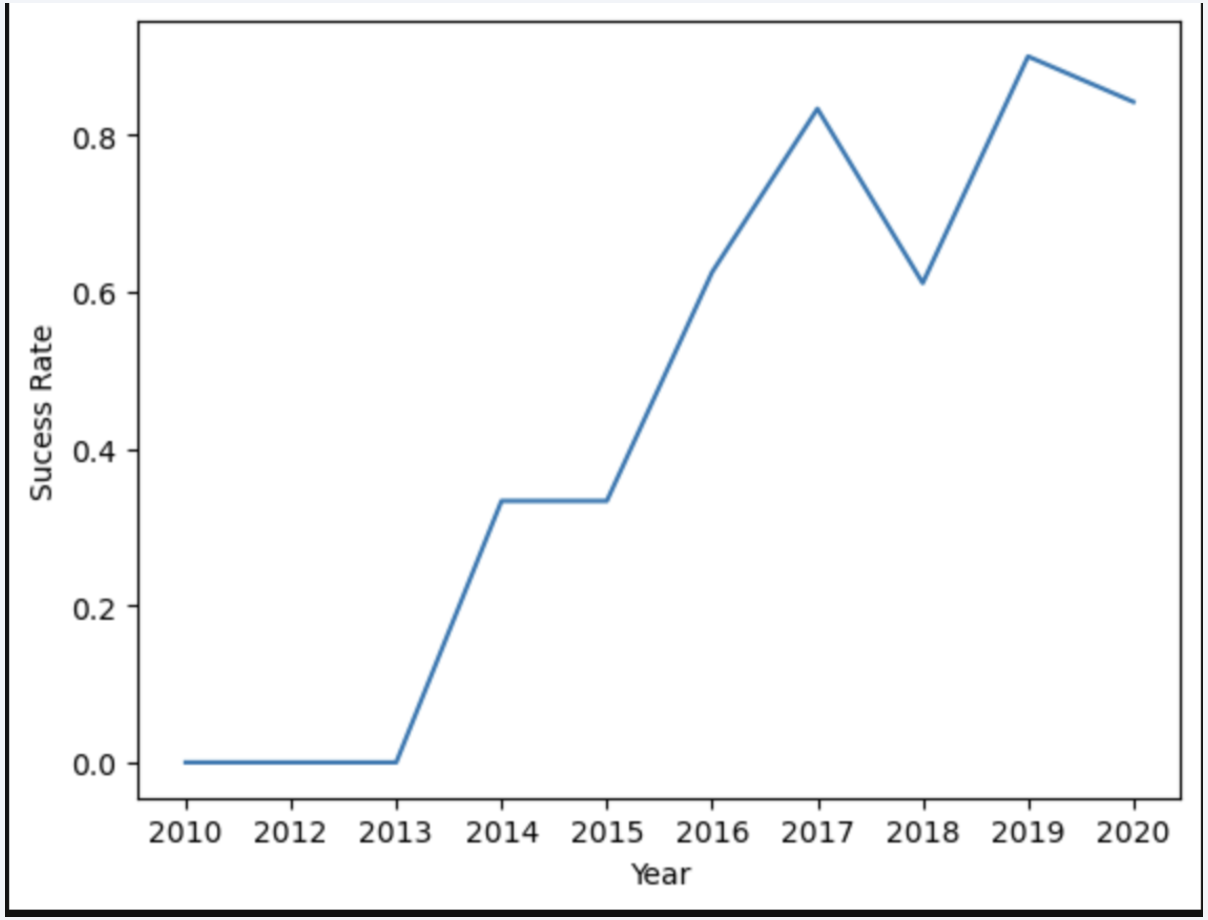
Payload vs. Orbit Type

- Heavier payloads tend to be more successful for LEO, ISS, and PO orbits. While GTO shows minimal correlation.



Launch Success Yearly Trend

- Success rates have been on the rise since 2013 with a slight drop in 2018.



All Launch Site Names

- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-4
- Query:
  - %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL

Launch Site Names Begin with 'CCA'

- Query:
  - %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%'

Date	Time	Booster Version	Launch Site	Payload Description	Payload Mass (kg)	Orbit	Customer(s)	Launch 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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:
  - 45596
- Query:
  - %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:
  - 2928.4
- Query:
  - %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AvgPayloadMass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad:
  - 2015-12-22
- Query:
  - %sql SELECT MIN(Date) AS FirstSuccesfulLanding FROM SPACEXTBL WHERE Mission_Outcome = "Success" AND Landing_Outcome LIKE "%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:
  - %sql SELECT Booster_Version FROM SPACEXTBL WHERE Mission_Outcome = "Success" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

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

- Query:
  - %sql SELECT Mission_Outcome, COUNT(*) AS MissionCount FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%' OR Mission_Outcome LIKE 'Success%' GROUP BY Mission_Outcome

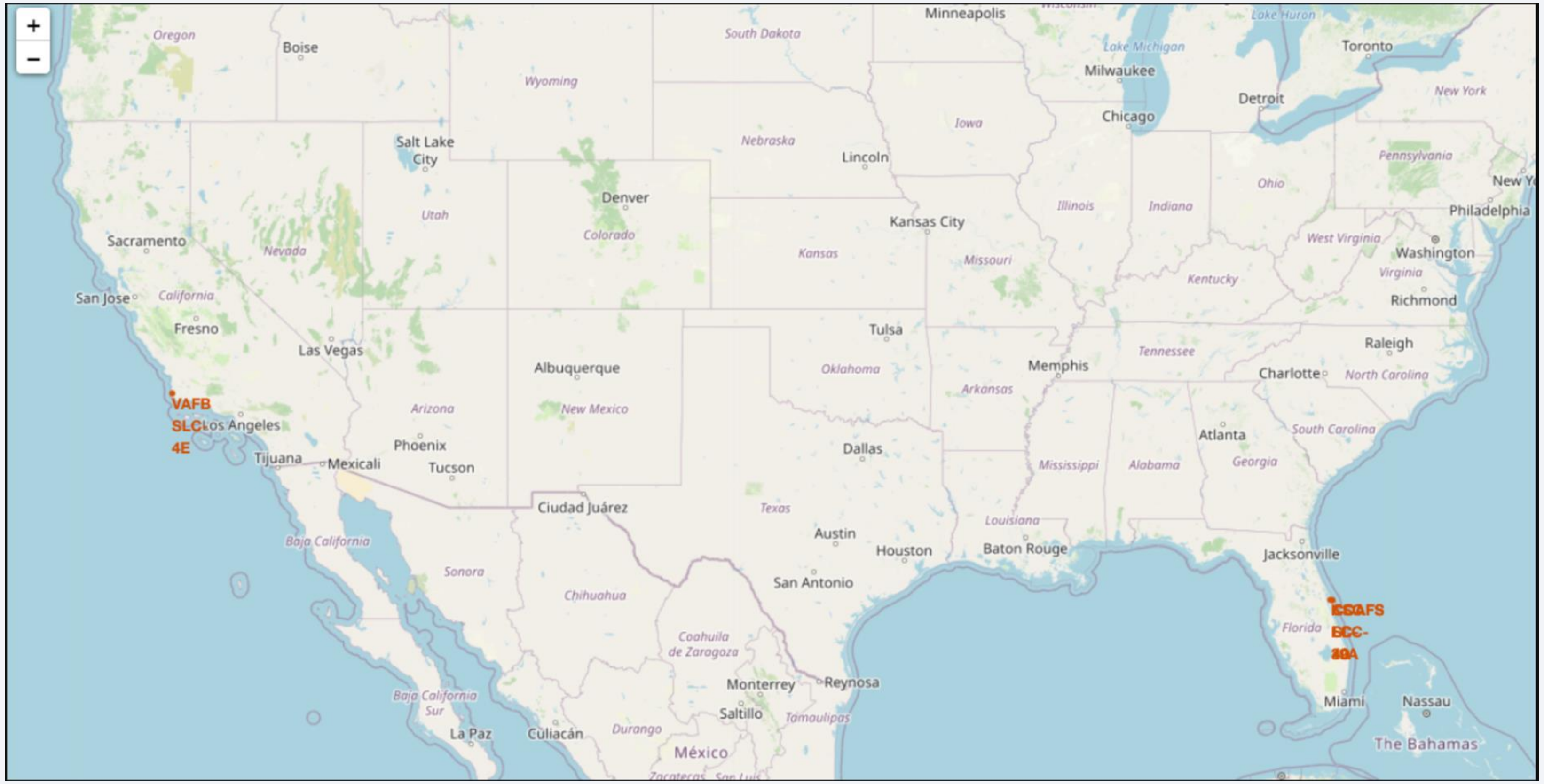
Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:
  - 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
- Query:
  - %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

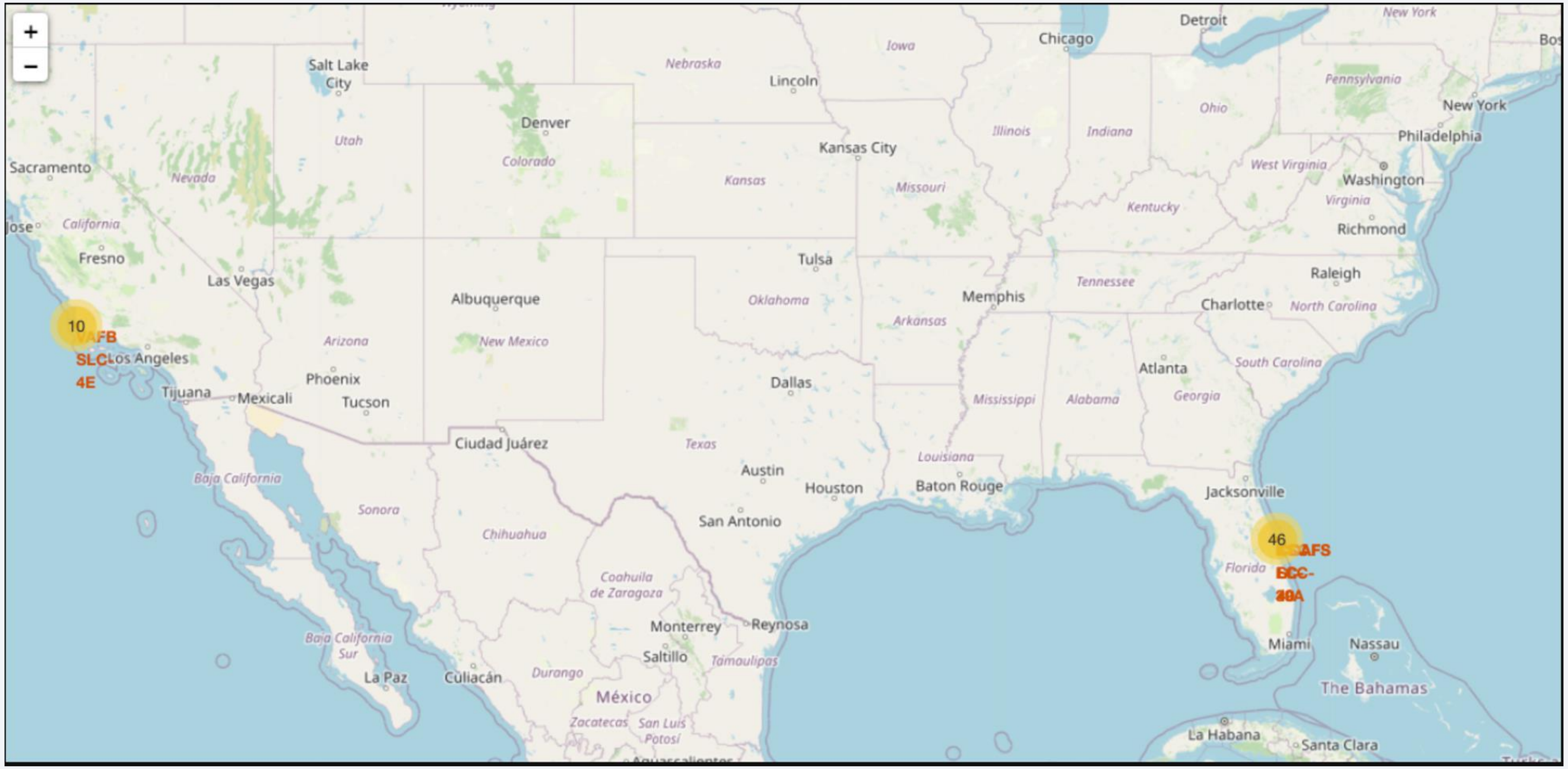
Section 3

Launch Sites Proximities Analysis

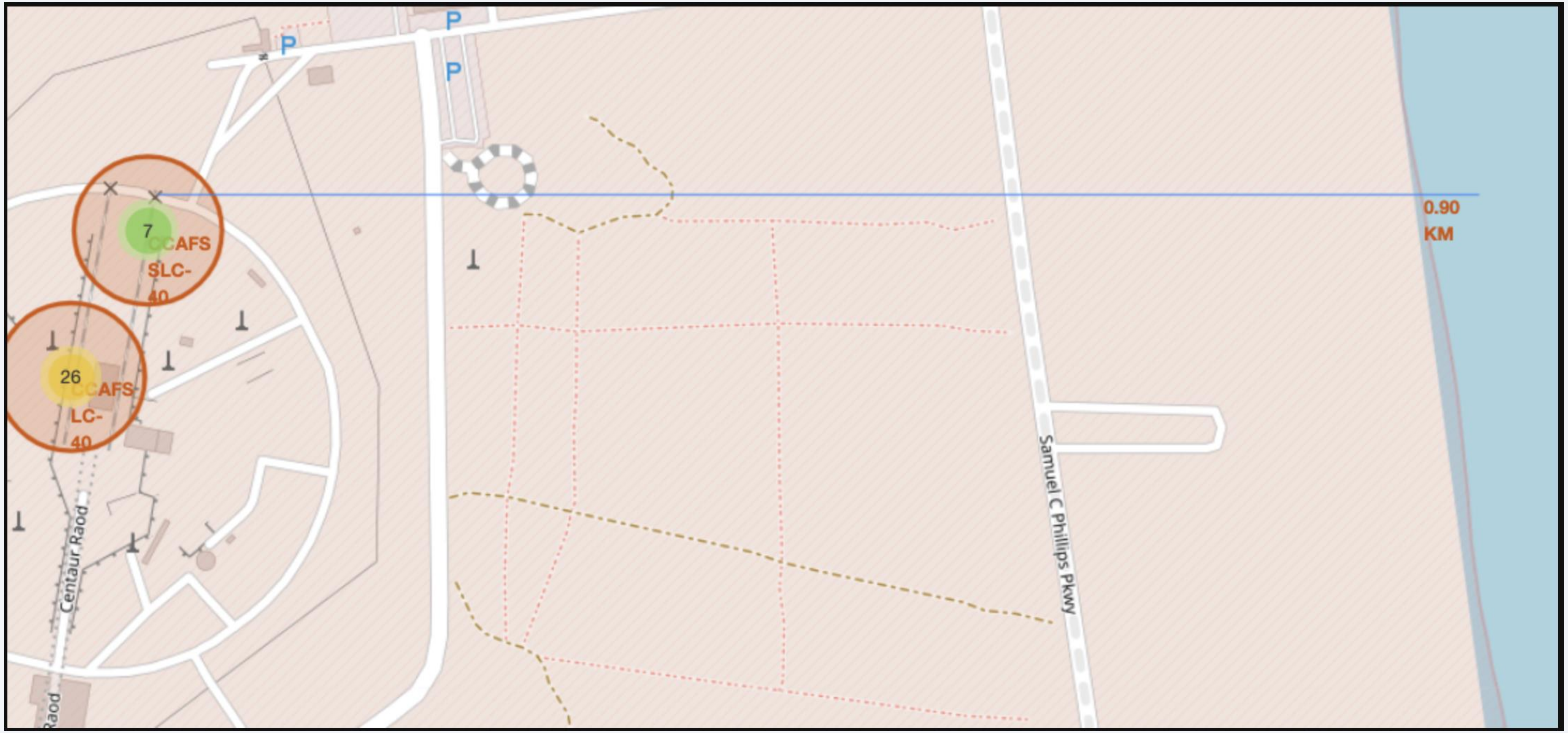
Map of All Launch Sites



All Launch Outcomes



Distance from Coast Line



Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

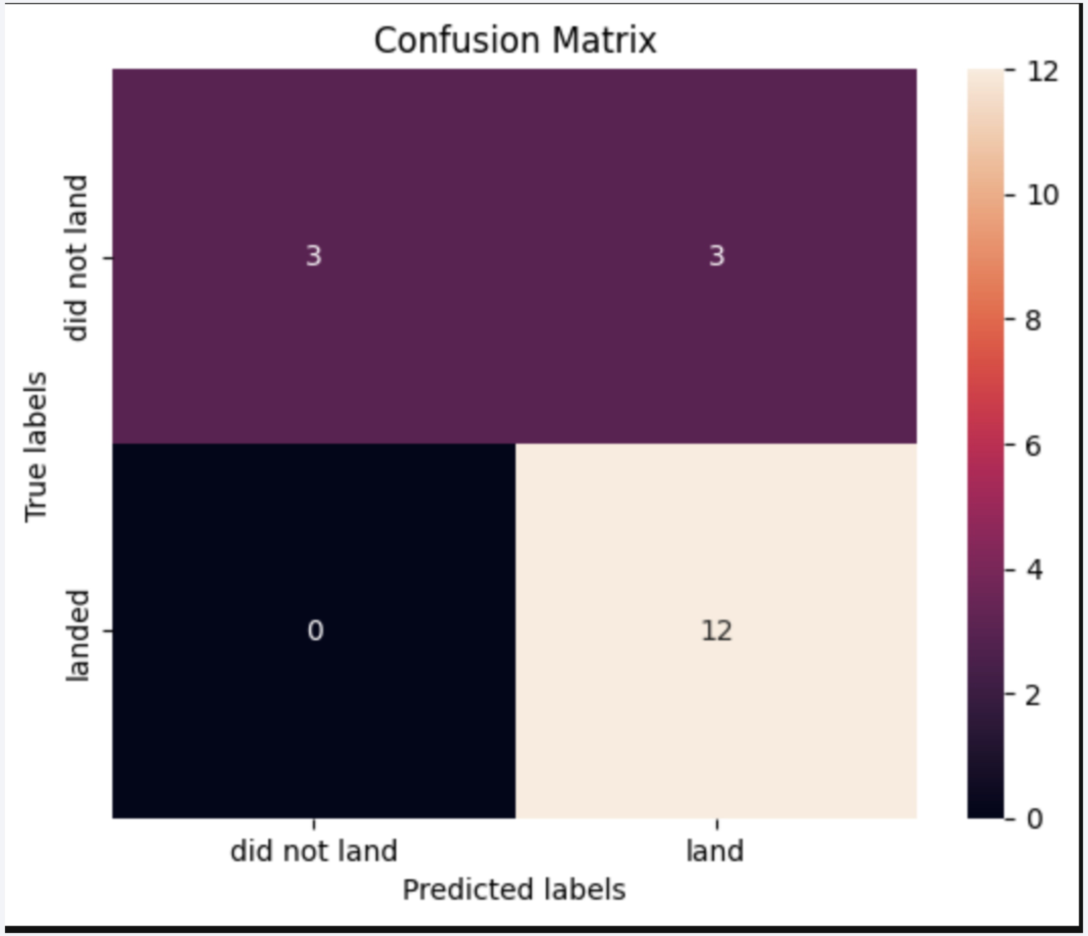
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree had the highest accuracy of 0.83%

Confusion Matrix



Conclusions

- ✅ **Accurate Prediction:** The tuned Decision Tree Classifier achieved ~83–85% test accuracy in predicting Falcon 9 first stage landing success.
- 📊 **Key Drivers Identified:** Launch site, payload mass, and orbit type were the most influential features affecting landing outcomes.
- 🚀 **Operational Insights:** Certain sites (e.g., KSC LC-39A) consistently show higher success rates, while very heavy or very light payloads lower the probability of a successful landing.
- 💡 **Business Value:** The model can be integrated into pre-launch planning to estimate landing success, supporting cost forecasting and resource allocation.
- 🔄 **Scalability:** The approach can be adapted for new launch data, ensuring the prediction system improves as more missions occur.
- 🔍 **Data-Driven Decision Making:** Empowers stakeholders to make informed choices on mission logistics and pricing strategies based on predicted success probabilities.

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!