IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prashant Yadav
June 30th, 2025

Outline

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

MADE BY PRASHANT YADAV

Executive Summary

- In this project, I loaded and prepared SpaceX launch data by standardizing features and extracting the target variable. I split the data into training and test sets, then trained and tuned four classification models — Logistic Regression, SVM, Decision Tree, and KNN — using GridSearchCV with 10-fold cross-validation. I evaluated each model using validation and test accuracy, selecting the best-performing algorithm for predicting rocket landing success.
- Across all models tested, the Support Vector Machine with a sigmoid kernel performed best, achieving a validation accuracy of ~84.8% and a test set accuracy slightly higher than other classifiers. Logistic Regression followed closely with ~84.6% test accuracy. Decision Tree and K-Nearest Neighbors models were also evaluated using GridSearchCV, but their performance was slightly lower, confirming SVM as the most effective approach for predicting SpaceX rocket landing success in this project.

MADE BY PRASHANT YADAV

Introduction

- This project focuses on predicting whether a SpaceX rocket successfully lands after launch. The goal is to build a classification model using historical launch data that includes payload mass, launch sites, booster features, and orbital details. By analyzing and modeling this data, we aim to understand the factors that influence landing outcomes and support future mission planning through accurate predictions.Problems you want to find answers

- In this project, the goal was to identify key launch factors that determine whether a SpaceX rocket successfully lands. The main problem addressed was: Can we accurately predict landing success based on mission characteristics like payload mass, launch site, booster configuration, and orbital target? By answering this, we aimed to support mission planning and cost optimization through data-driven insights.

MADE BY PRASHANTI YADAV

Section 1

Methodology

MADE BY PRASHANT YADAV

Methodology

Executive Summary

- Data collection methodology.
- Performing data wrangling.
- Performing exploratory data analysis (EDA) using visualization and SQL.
- Performing interactive visual analytics using Folium and Plotly Dash.
- Performing predictive analysis using classification models.

MADE BY PRASHANT YADAV

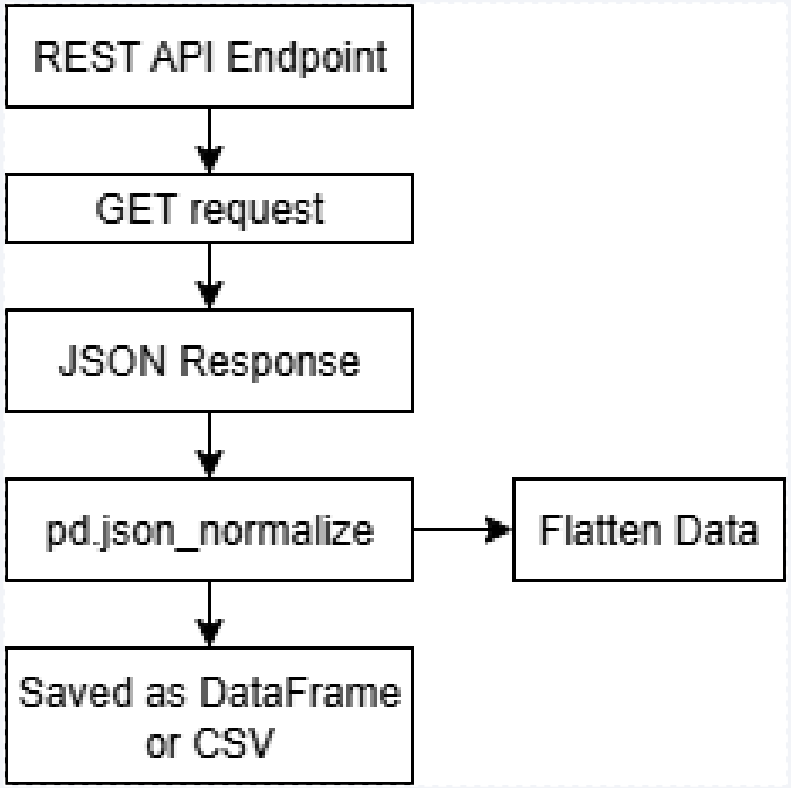
Data Collection

- **Data Collection Overview**
- We collected SpaceX launch data through the following steps:
- **Accessed Cloud-hosted Datasets**
  - CSV files provided via course environment
  - Contained structured launch records (mission details, payloads, landing outcome)
  - Used fetch() and pd.read_csv() for local access
  - Two datasets:
    - dataset_part_2.csv: Raw historical launches
    - dataset_part_3.csv: Preprocessed feature-ready set
- **Verified and Aligned Features**
  - Ensured consistency of variable formats and labels
  - Target variable Class identified for supervised learning

MADE BY PRASHANT YADAV

Data Collection – SpaceX API

- **SpaceX API Data Collection (Summary)**
- Queried SpaceX REST API using Python's requests library
- Endpoint: https://api.spacexdata.com/v4/launches
- Parsed JSON response with pandas.json_normalize()
- Extracted fields: mission name, launch date, booster type, payload, success status
- Saved structured data for analysis
- **GitHub Notebook**: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]



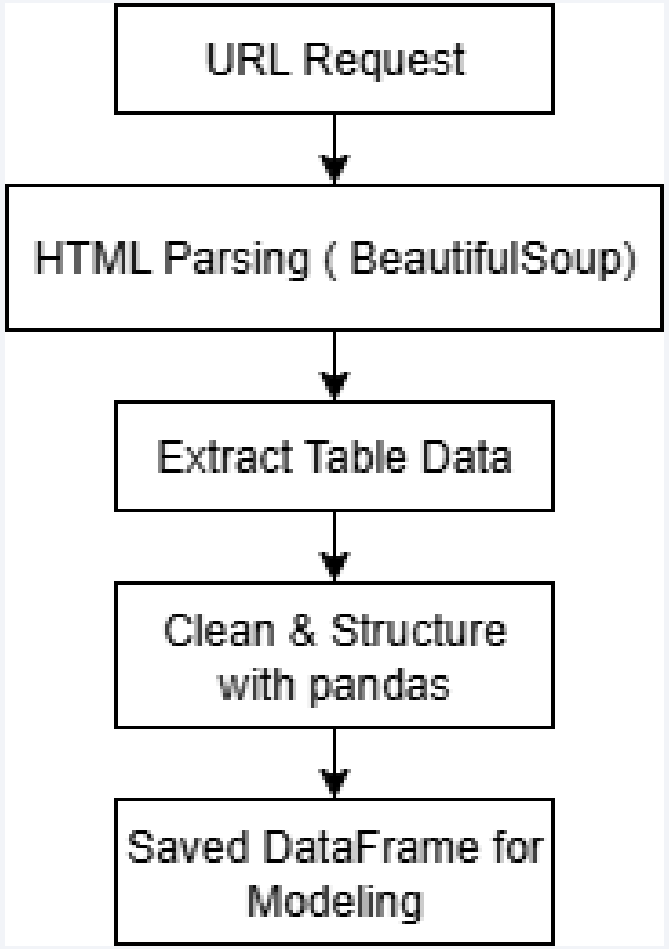
MADE BY PRASHANT TILDAV

Data Collection - Scraping

Web Scraping Process (Summary)

- Targeted SpaceX Wikipedia page using requests and BeautifulSoup
- Retrieved raw HTML content with requests.get()
- Parsed HTML to locate launch table using soup.find_all()
- Extracted relevant columns: Date, Mission, Launch Site, Orbit, and Outcome
- Structured the scraped data into a pandas DataFrame and saved for later use

GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]



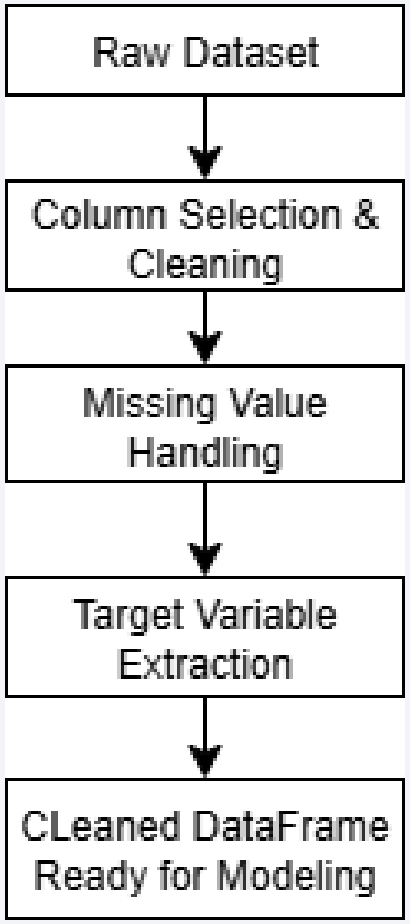
MADE BY TEASHANT YADAV

Data Wrangling

Data Wrangling Process (Summary)

- Selected relevant columns and dropped irrelevant or redundant fields
- Handled missing values and inconsistent formats
- Converted Class to NumPy array as target label
- Verified feature alignment and ensured clean data input for modeling

GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]



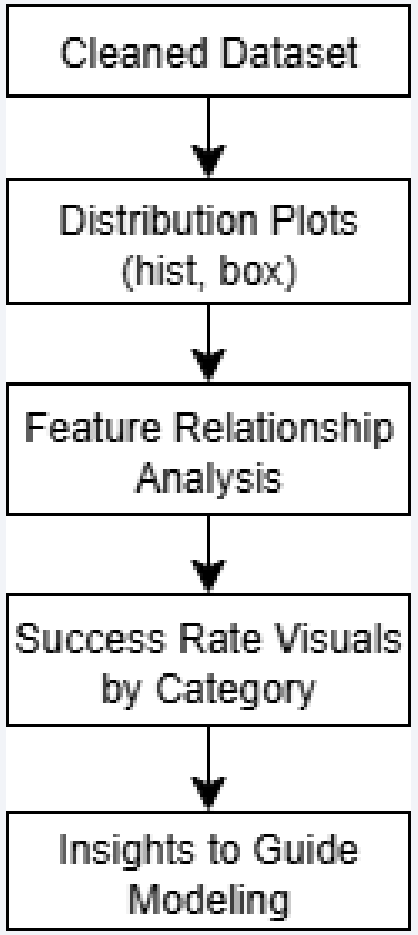
MADE BY PRIYANSHU YADAV

EDA with Data Visualization

Data Visualization & EDA (Summary)

- Explored feature distributions and relationships using matplotlib and seaborn
- Visualized impacts of payload, orbit, and launch site on landing success
- Identified patterns and outliers to guide feature selection and model design

GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]

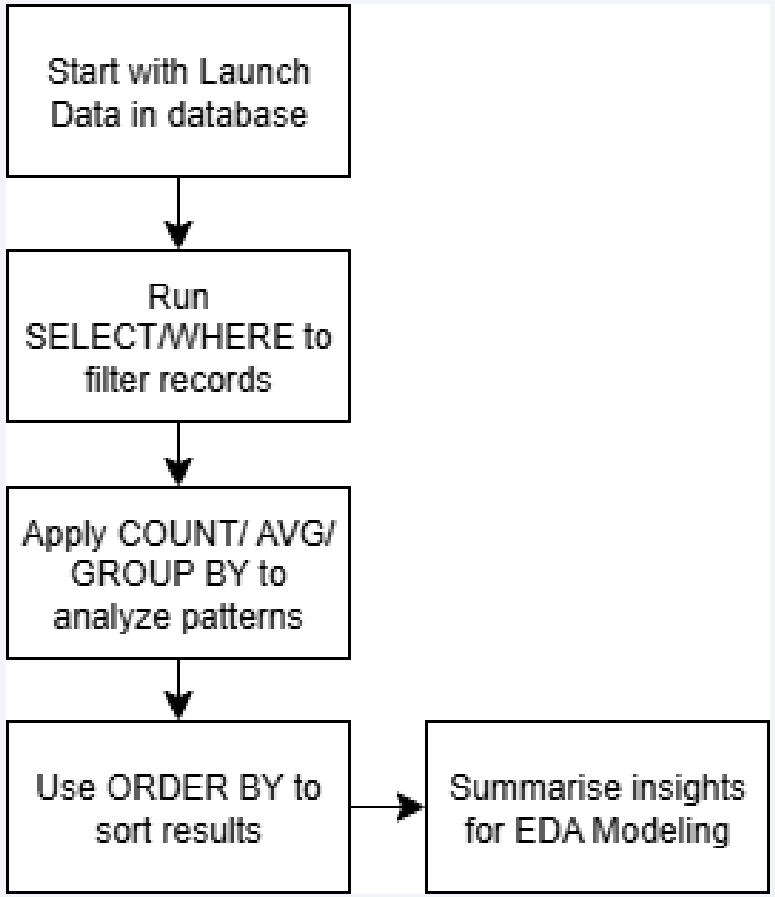


MADE BY PRASHANT YADAV

EDA with SQL

- Selected all records to explore the structure: SELECT * FROM table LIMIT 5;
- Counted rows and checked missing values using COUNT() and IS NULL
- Retrieved successful launches: WHERE Class = 1
- Filtered by booster version, launch site, or orbit
- Used GROUP BY to calculate success rates per orbit, site, or booster
- Applied AVG(), COUNT(), and SUM() to analyze trends
- Ordered launches by payload or date with ORDER BY
- Identified top-performing sites or boosters
- Merged tables (e.g., launch data with payload details) using JOIN

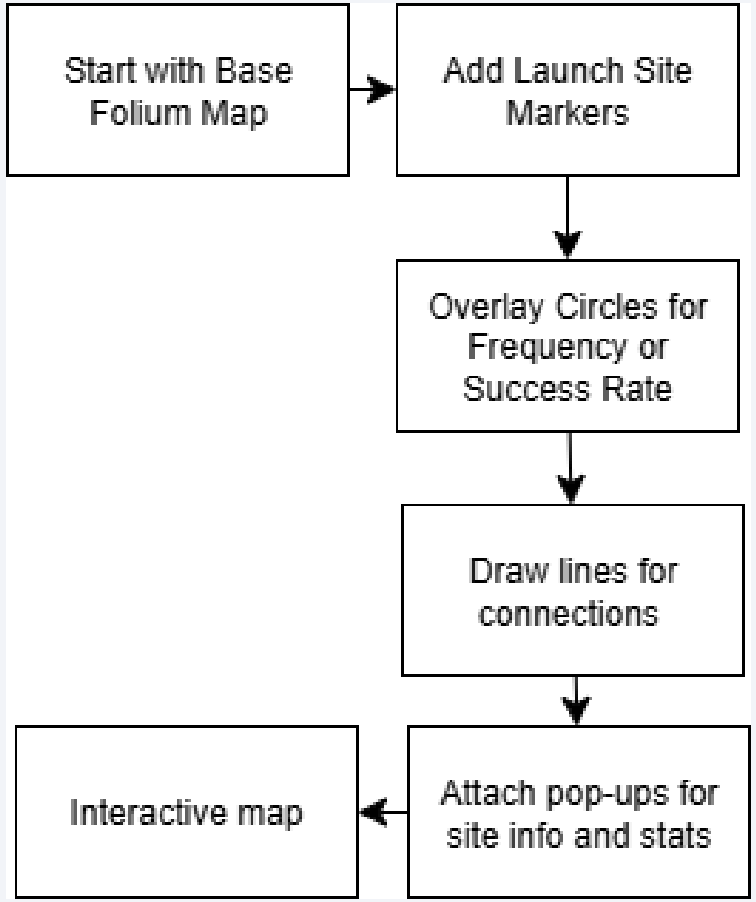
GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]



MADE BY PRASHANT LADAK

Build an Interactive Map with Folium

- **Markers**: Plotted individual launch sites to pinpoint their geographic locations
- **Circles**: Added circle overlays to highlight the success rate or frequency of launches per site
- **Polylines**: (If used) Connected launch sites to landing zones or illustrated rocket trajectories
- **Popups**: Attached to markers to display site names and launch statistics on click
- **Purpose**: These elements provided geographic context and visual insights into launch site performance, making spatial patterns and site comparisons easy to understand.

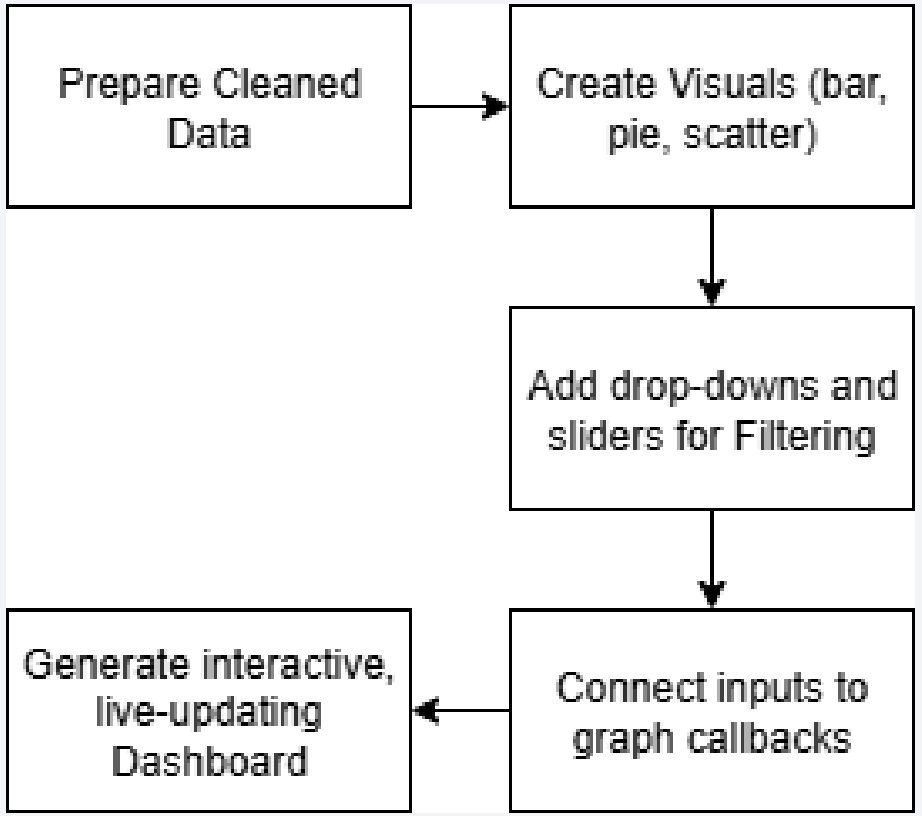


GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]

MADE BY PRASHANT TIWARI

Build a Dashboard with Plotly Dash

- **Bar, Pie, and Scatter Plots**: Showed success trends by site, booster, and payload
- **Dropdowns & Sliders**: Enabled users to filter by site, orbit, and payload range
- **Live Updates**: Graphs reacted instantly to user selections for interactive exploration

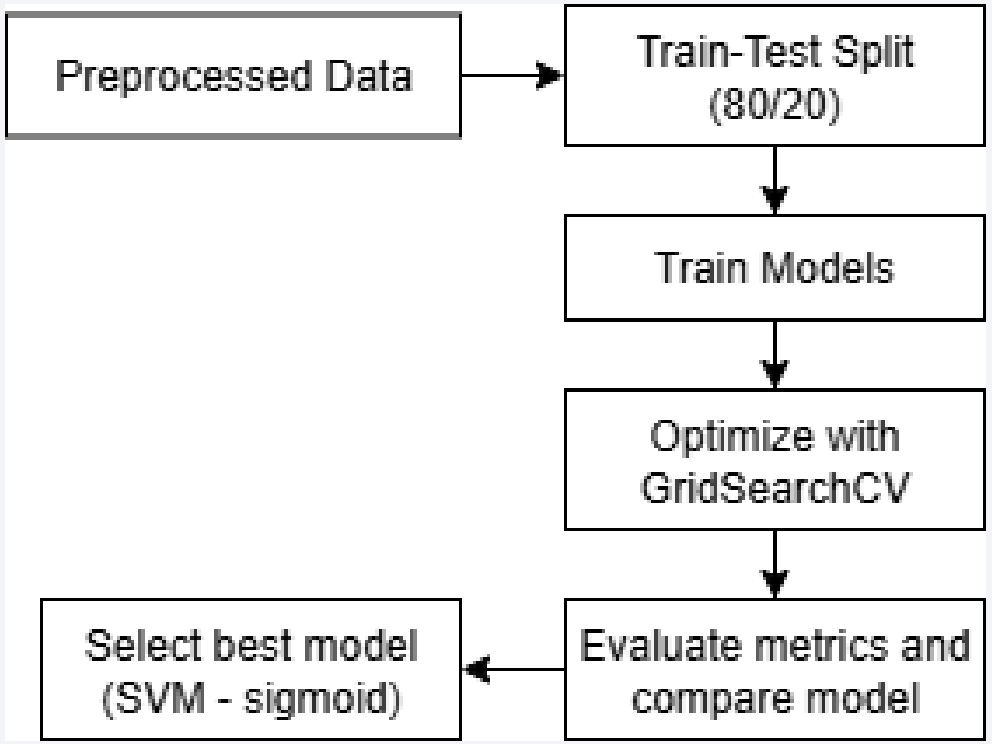


GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]

MADE BY PRASHANT TIDAV

Predictive Analysis (Classification)

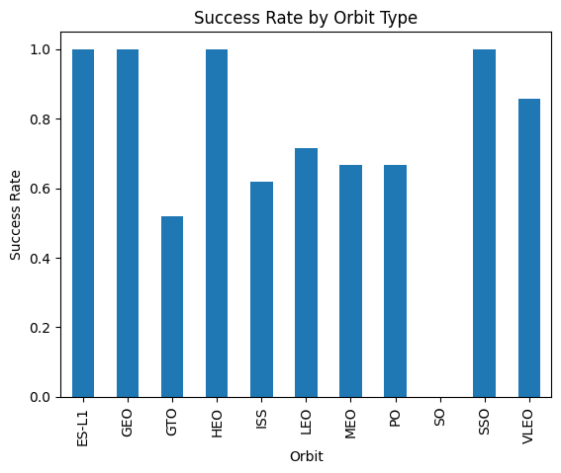
- Built multiple classifiers: **Logistic Regression, SVM, Decision Tree, KNN**
- Used **GridSearchCV** for hyperparameter tuning and cross-validation
- Evaluated using **accuracy**, **confusion matrix**, and **validation scores**
- Identified **SVM with sigmoid kernel** as the best performer (~84.8% accuracy)



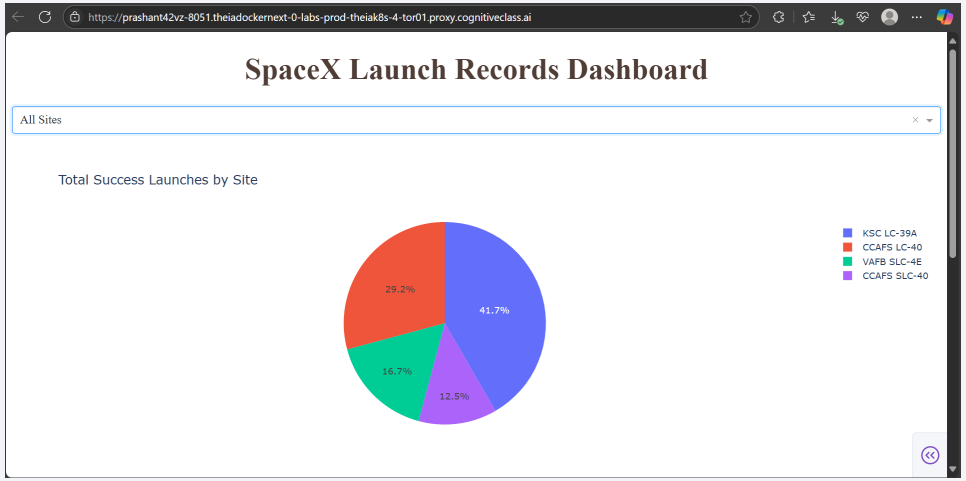
GitHub Notebook: [LearningStuff/IBM Applied Data Science Capstone Project at main · TheGLIV/LearningStuff]

MADE BY PRASHANT YADAV

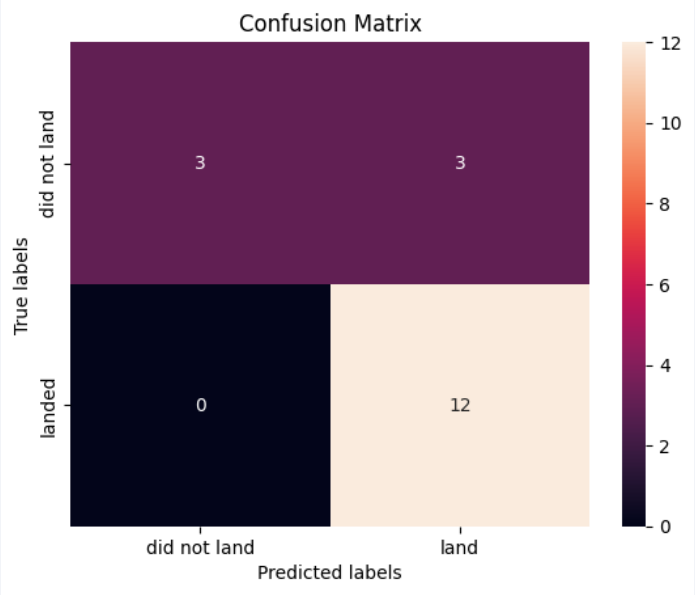
Results



- SSO, HEO, GEO & ES-L1 Orbits have the highest success rates.



- Users can filter launches by site and orbit



- Model showing successful landings well but struggles equally with false positives and true negatives

Above are some of the results obtained from EDA, Interactive Dashboard building and Predictive Analysis.

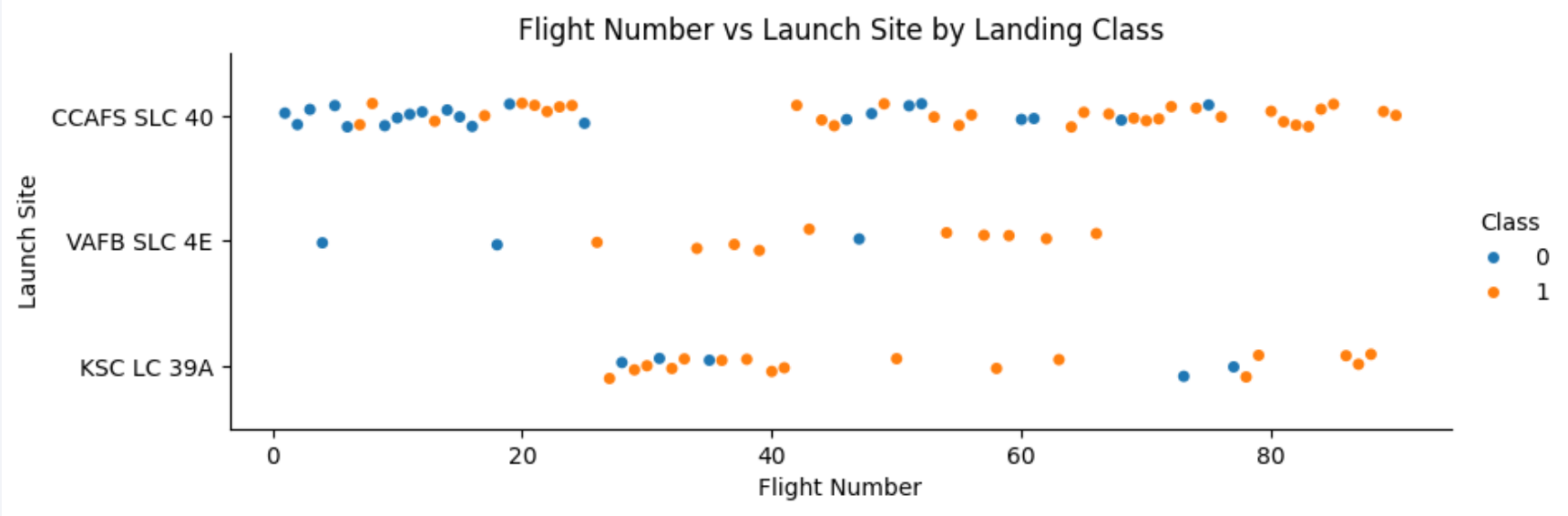
MADE BY PRASHANT YADAV

Section 2

Insights drawn from EDA

MADE BY PRASHANT YADAV

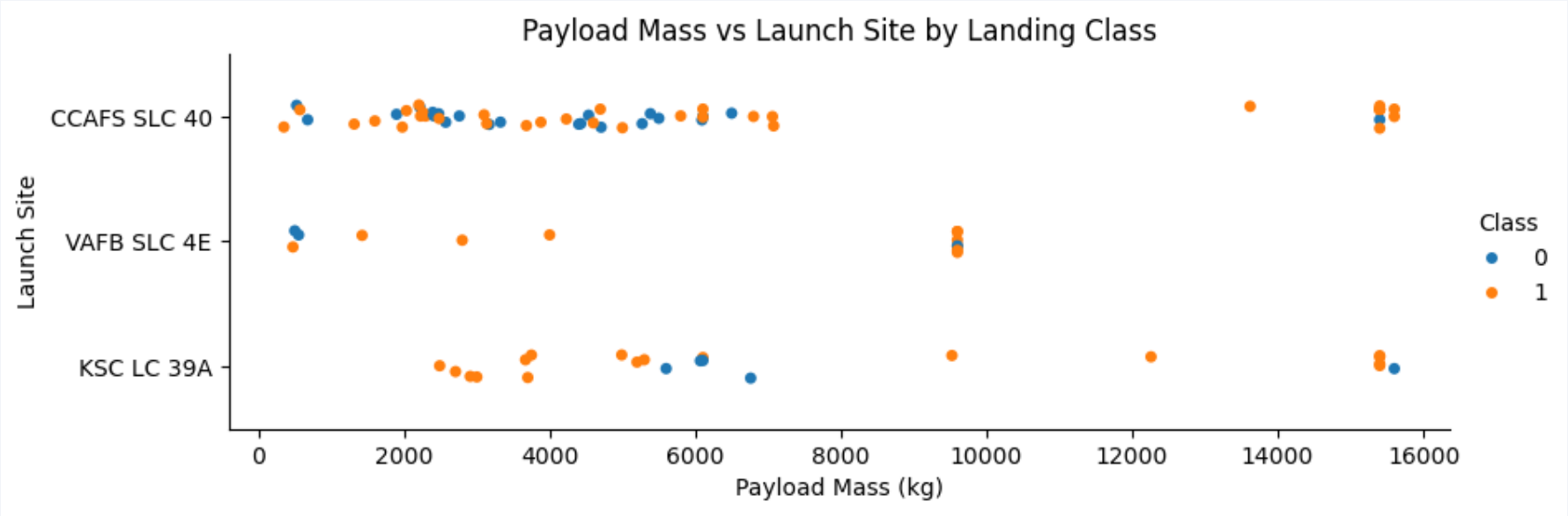
Flight Number vs. Launch Site



- Shows how flight activity varies by launch site over time
- CCAFS SLC 40 and KSC LC 39A handled the highest number of flights
- Higher flight numbers are more concentrated at KSC LC 39A, indicating it became more dominant in later missions
- VAFB SLC 4E had fewer launches, mostly clustered in early flight numbers
- Helps understand launch site usage trends and potential transitions in mission strategy

MADE BY PRASHANT YADAV

Payload vs. Launch Site

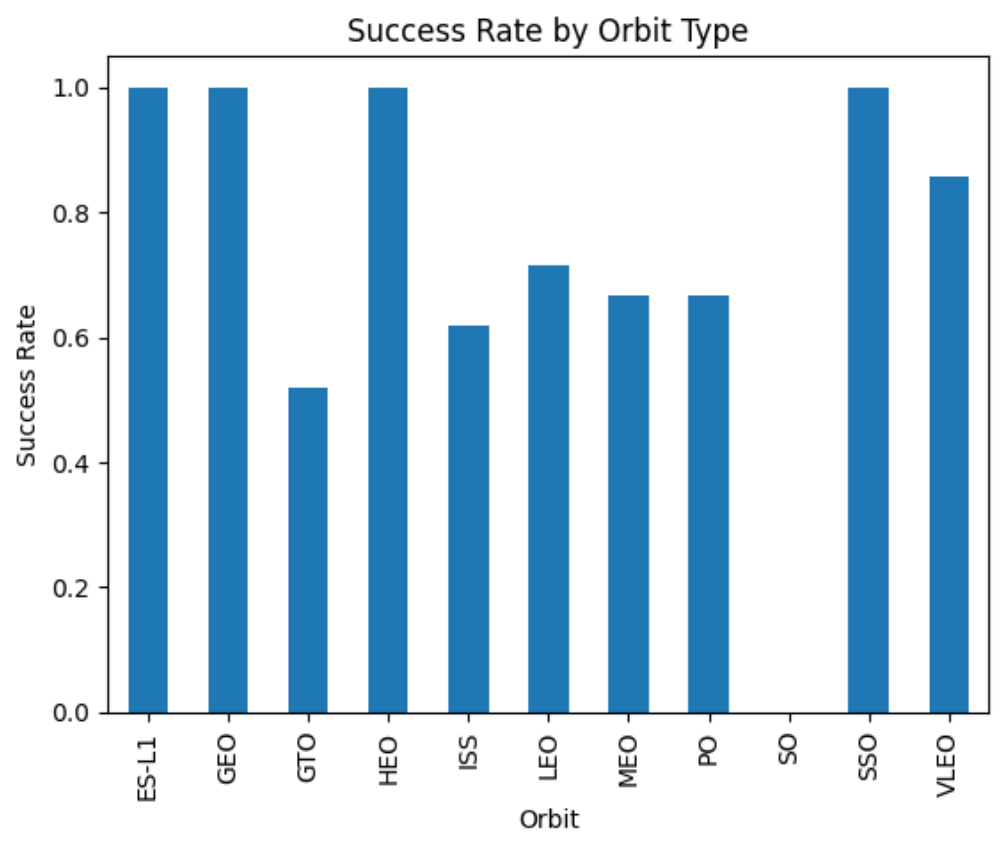


- KSC LC 39A handled the widest range of heavy payloads with more successful landings
- Unsuccessful landings (Class 0) appear more often at lower payloads and fewer sites
- Visual shows a strong link between site capability, payload size, and landing outcome

MADE BY PRASHANT YADAV

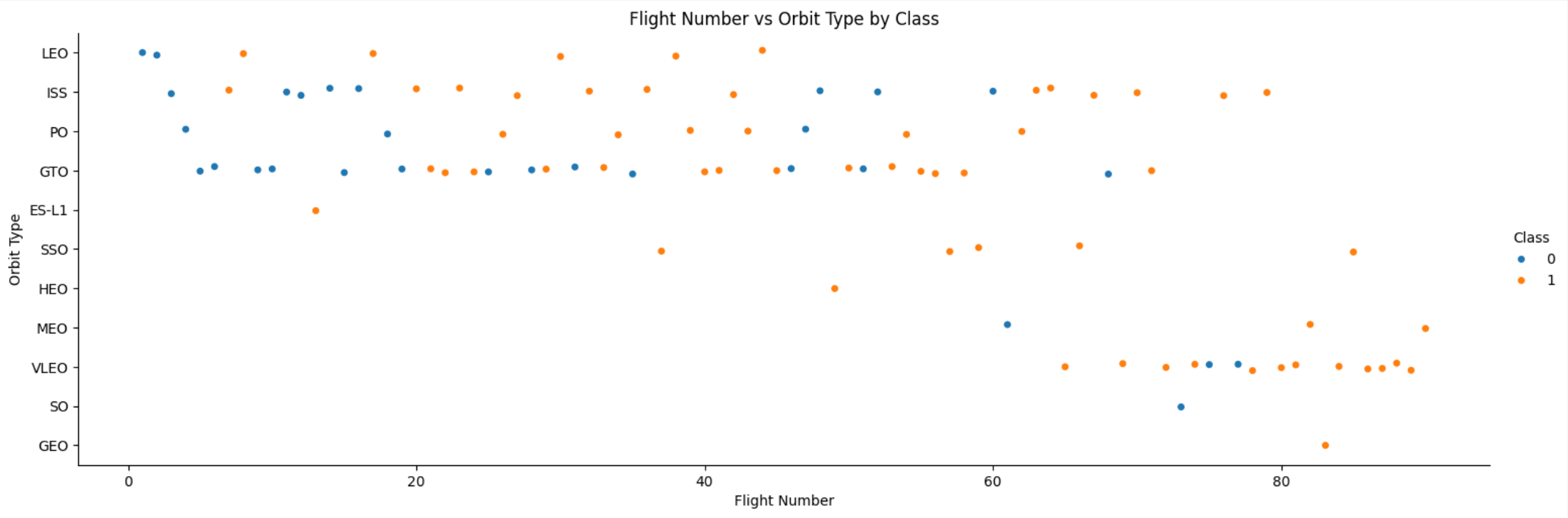
Success Rate vs. Orbit Type

- Orbits like ES-L1, GEO, HEO, and SSO achieved a 100% success rate
- GTO had the lowest success rate among all orbit types
- Highlights how launch success varies significantly by mission orbit destination



MADE BY PRASHANT YADAV

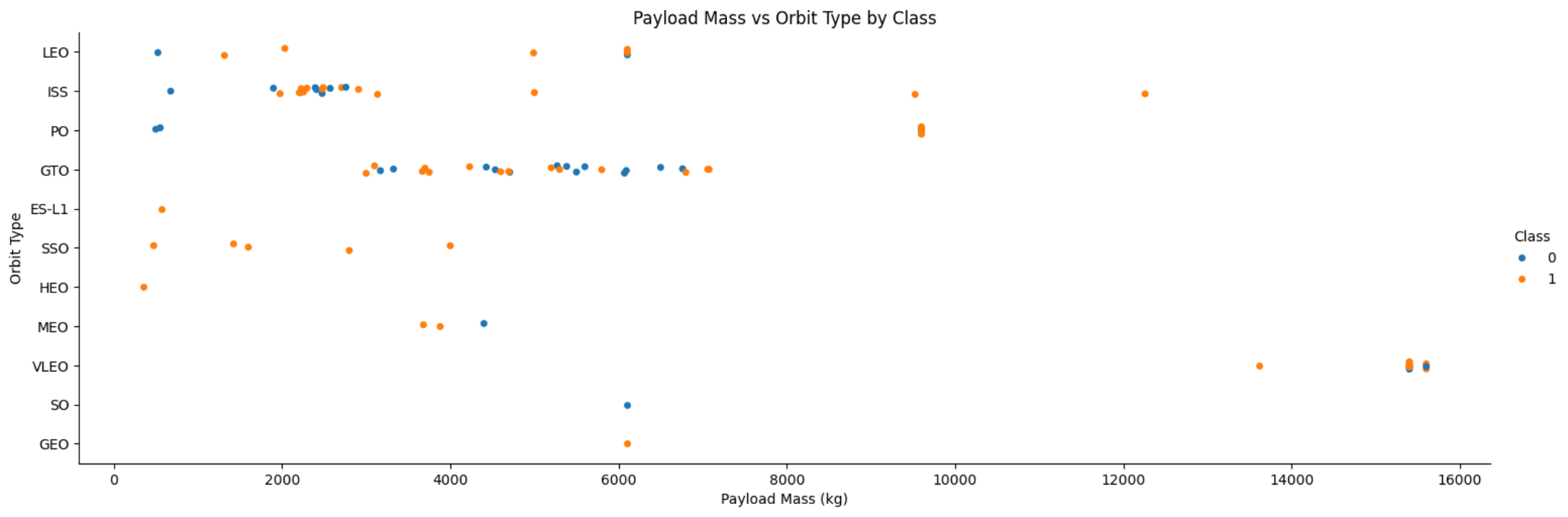
Flight Number vs. Orbit Type



- Can observe that 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.

MADE BY PRASHANT YADAV

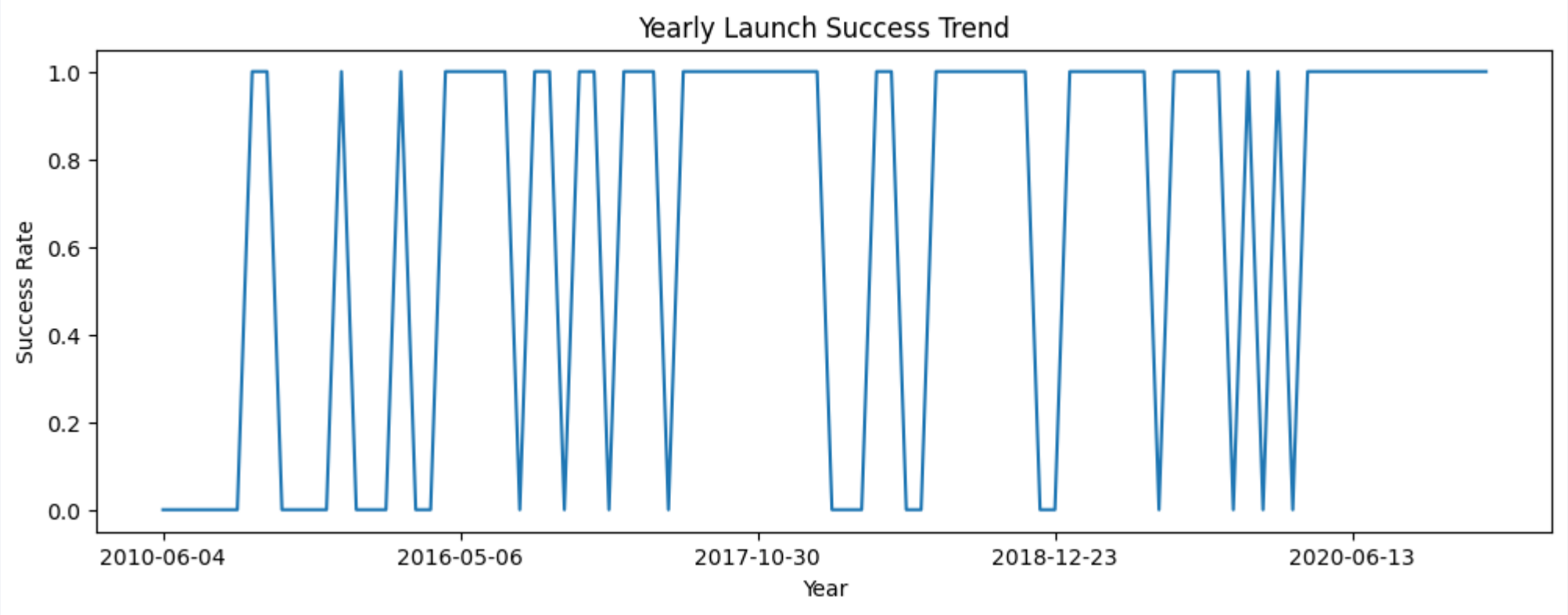
Payload vs. Orbit Type



- 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.

MADE BY PRASHANT VADM

Launch Success Yearly Trend



- In the Yearly Launch Success Trend line chart you can observe that the success rate since 2013 kept increasing till 2020.

MADE BY PRASHANT

All Launch Site Names

Query: SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

Result:

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

Explanation:

Used to retrieve a list of all unique launch site names from the dataset — useful for filtering, dropdowns, or analyzing launch patterns per site.

MADE BY PRASHANT YADAV

Launch Site Names Begin with 'CCA'

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

Result:

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

Explanation:

Used to preview the first five launch records specifically from sites starting with "CCA" — helpful for quickly inspecting site-specific data like CCAFS SLC 40.

MADE BY PRASHANT YADAV

Total Payload Mass

Query: SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';

Result:

Total_Payload_Mass
48213

Explanation:

Used to calculate the total payload mass carried for NASA CRS missions — useful for analyzing NASA's cargo contribution or filtering mission-specific statistics.

MADE BY PRASHANT YADAV

Average Payload Mass by F9 v1.1

Query: SELECT AVG("PAYLOAD_MASS__KG_") AS Average_Payload_Mass
FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';

Result:

Average_Payload_Mass
2928.4

Explanation:

To calculate the average payload mass carried specifically by the 'F9 v1.1' booster version — helpful for comparing performance across different booster types.

MADE BY PRASHANT YADAV

First Successful Ground Landing Date

Query: SELECT MIN(Date) AS First_Ground_Pad_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';

Result:

First_Ground_Pad_Landing
2015-12-22

Explanation:

- To find the date of the first successful ground pad landing — useful for identifying milestones in SpaceX's landing history and evaluating technological progression.

MADE BY PRASHANT YADAV

Successful Drone Ship Landing with Payload between 4000 and 6000

Query: SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

Result:

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

Explanation:

To identify which booster versions successfully landed on a drone ship while carrying medium-range payloads (between 4000 and 6000 kg)

MADE BY PRASHANT YADAV

Total Number of Successful and Failure Mission Outcomes

Query: SELECT "Mission_Outcome", COUNT(*) AS Total FROM SPACEXTABLE GROUP BY "Mission_Outcome";

Result:

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

Explanation:

To count how many launches resulted in each type of mission outcome.

MADE BY PRASHANT YADAV

Boosters Carried Maximum Payload

Query: SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

Result:

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

Explanation:

To identify which booster version carried the **heaviest payload recorded** in the dataset

MADE BY PRASHANT

2015 Launch Records

Query: SELECT substr(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr(Date, 1, 4) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)';

Result:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation:

This query extracts all failed drone ship landings in 2015, along with the month, booster version, and launch site.

MADE BY ANASHANT YADAV

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

Query: 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 Outcome_Count DESC;

Result:

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

Explanation:

This query counts the number of each landing outcome between June 4, 2010, and March 20, 2017, then sorts them from most to least frequent.

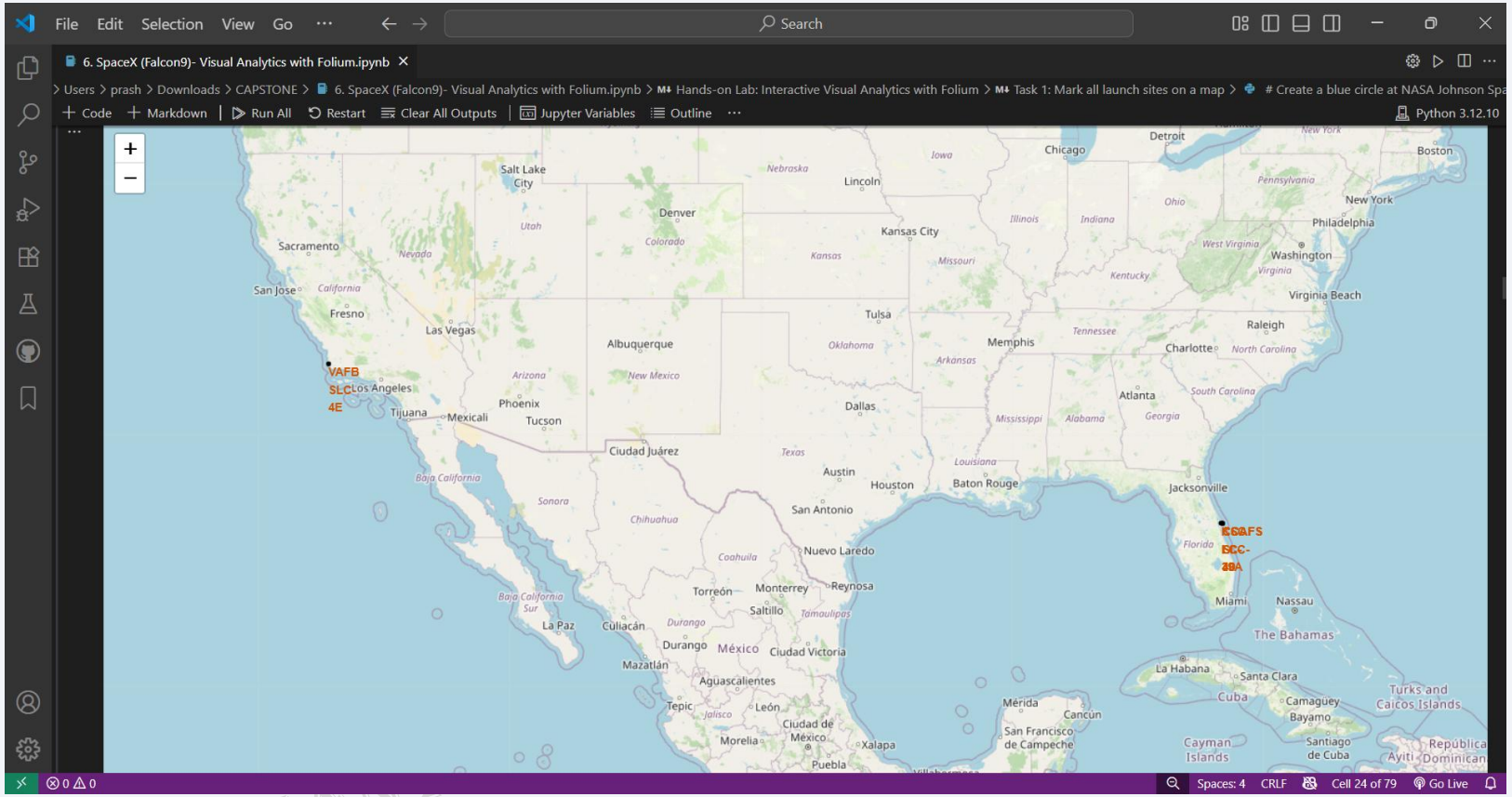
MADE BY PRAS

Section 3

Launch Sites Proximities Analysis

MADE BY PRASHANT YADAV

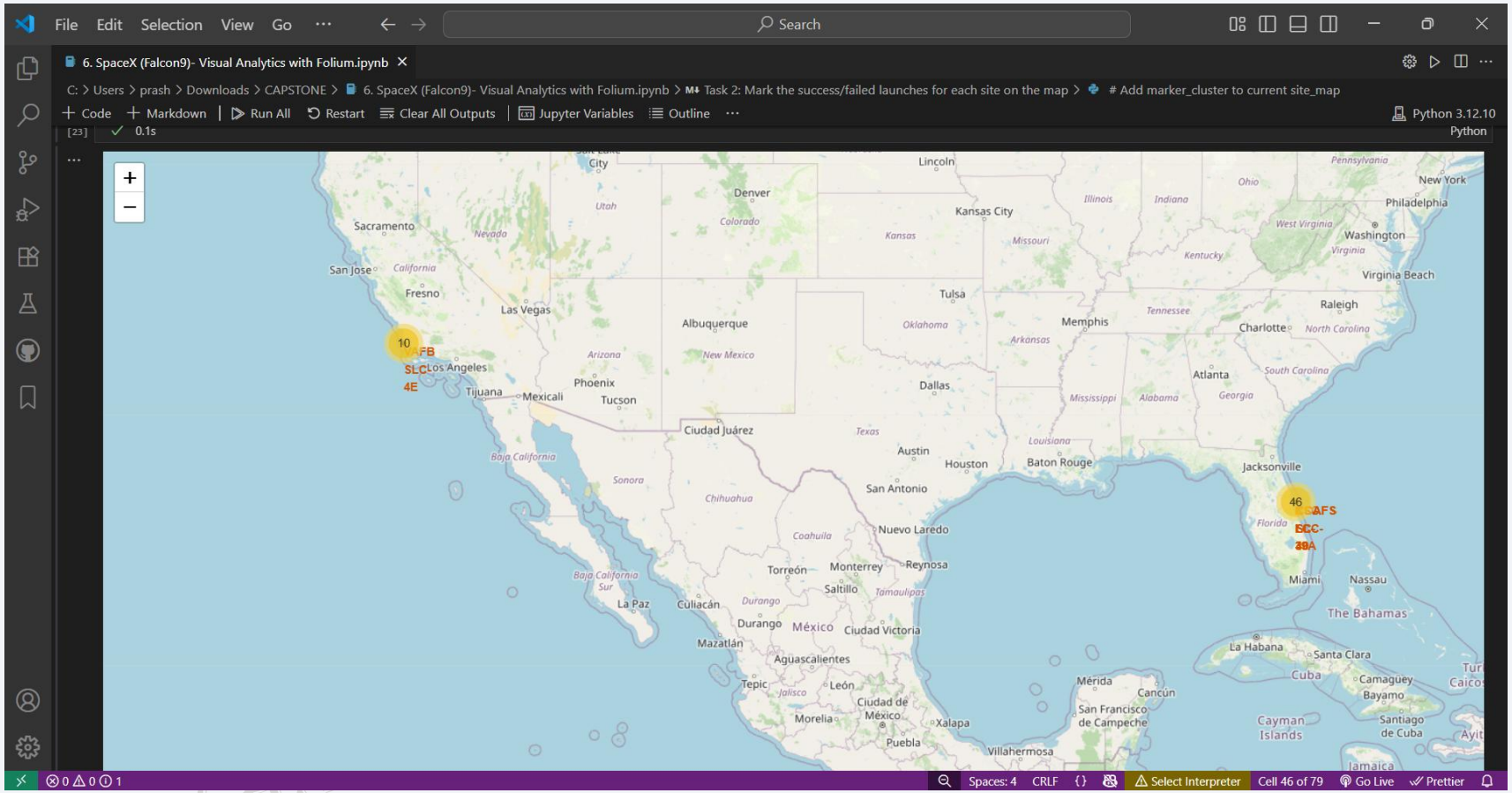
Launch Site Analysis with Maps using Folium- Part One



Findings: All launch sites are near the coasts and above the equators.

MADE BY PRASHANT

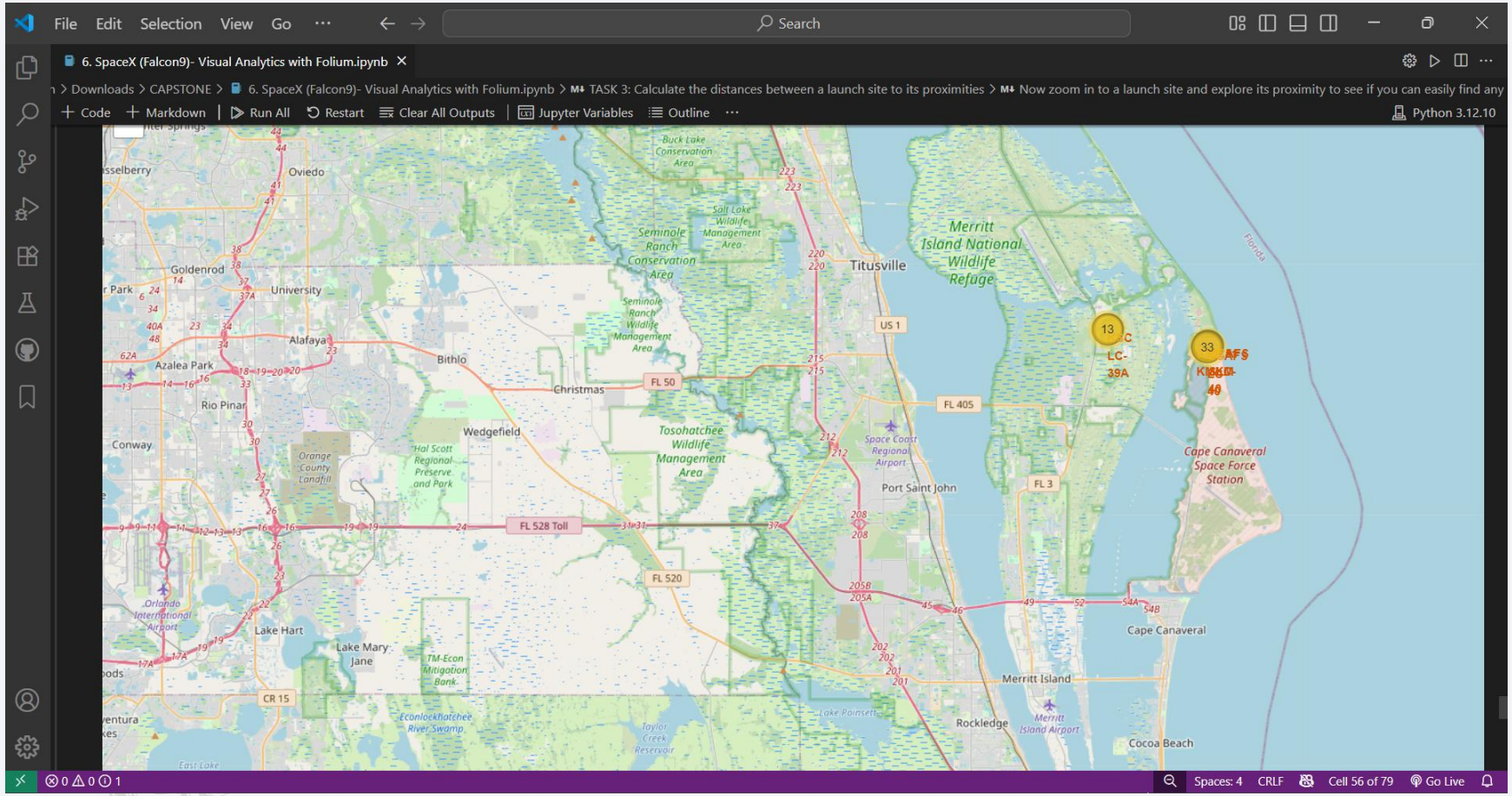
Launch Site Analysis with Maps using Folium- Part Two



Findings: From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.

MADE BY PRASH

Launch Site Analysis with Maps using Folium- Part Three



Findings: Using the map we can easily identify what the launch sites have in close proximity, like railways, subway, highways, coastlines etc.

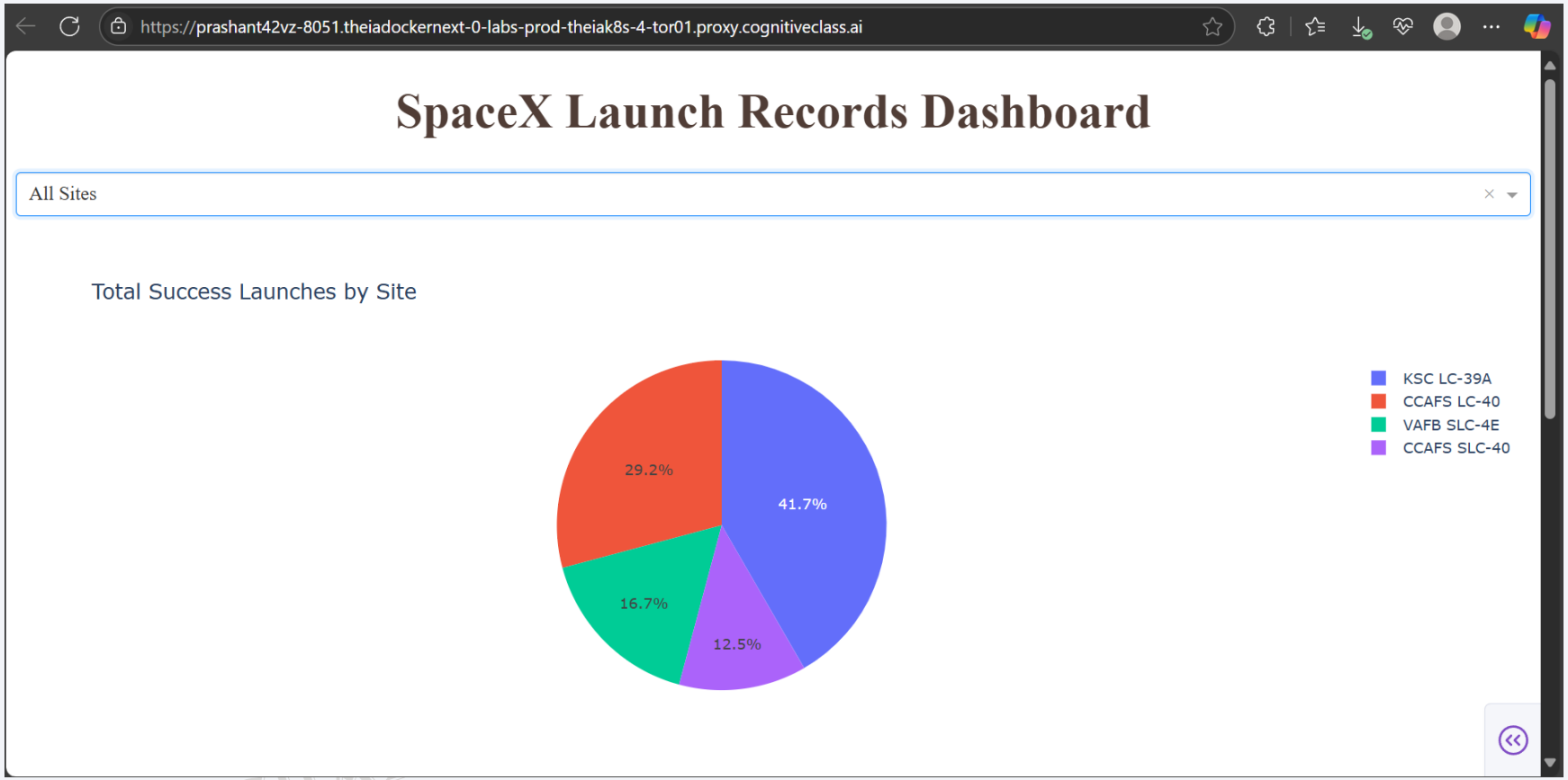
MADE BY PRASA

Section 4

Build a Dashboard with Plotly Dash

MADE BY PRASHANT YADAV

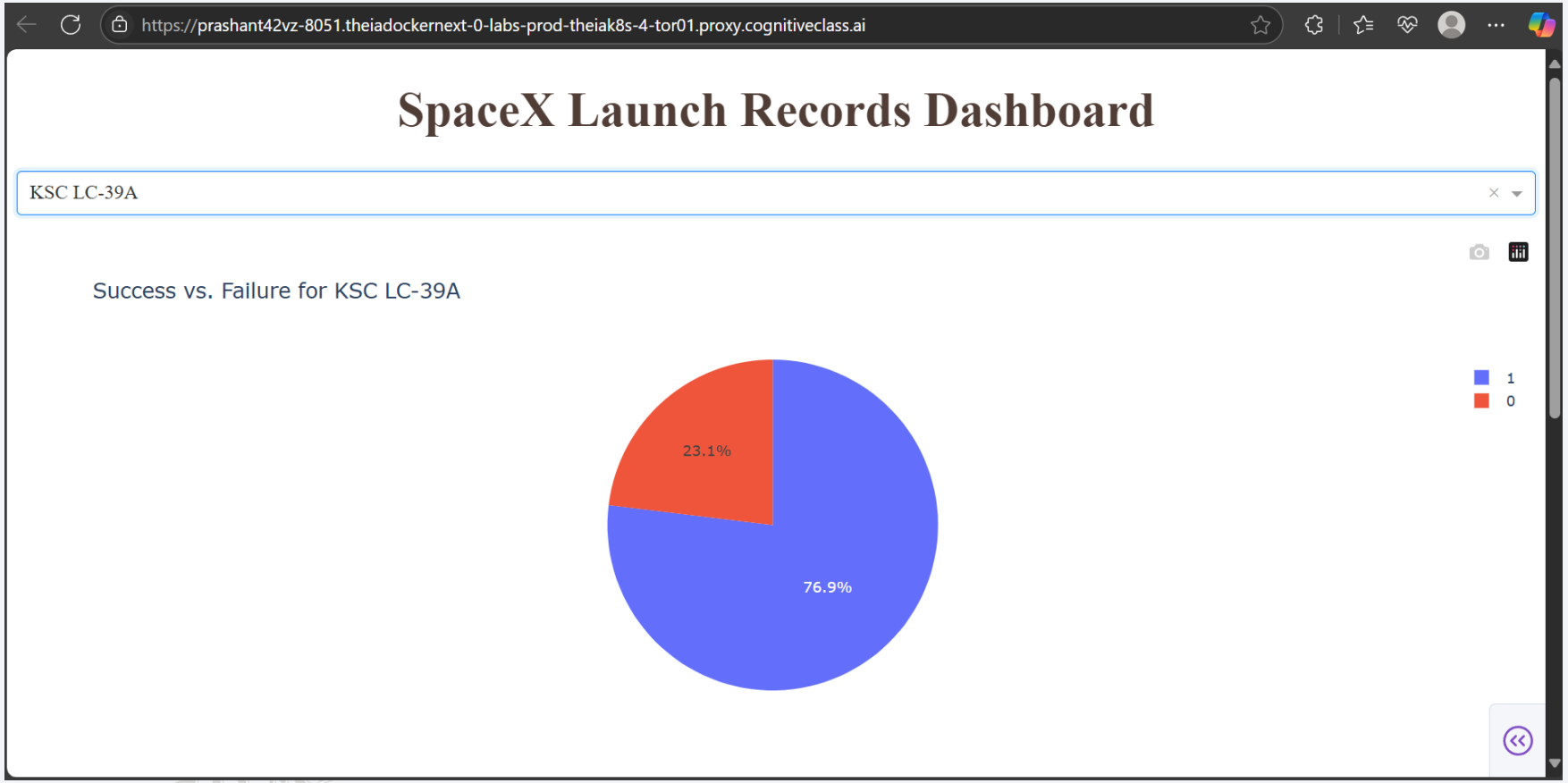
SpaceX Launch Records Dashboard- Part One



KSC LC- 39A has the most success launches, marking 41.7% of all launch successes.

MADE BY PRASHA

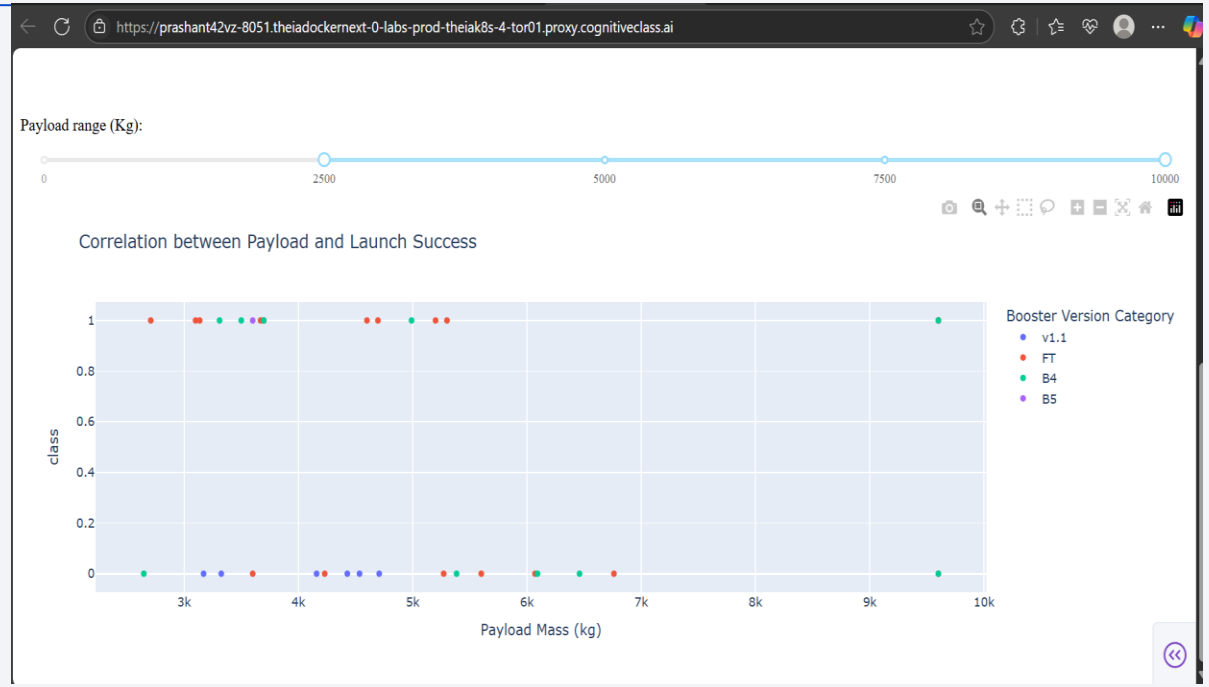
SpaceX Launch Records Dashboard- Part Two



Findings: KSC LC-39A has 76.9% launch success rate, highest among all the sites.

MADE BY PRASHA

SpaceX Launch Records Dashboard- Part Three



- Fig(1) shows Payload and Launch Success graph for payload range 0-7500kg.
- Fig.2 shows Payload and Launch Success Scatter Plot for payload range 2500-10000kg

MADE BY PRASHANT YADAV

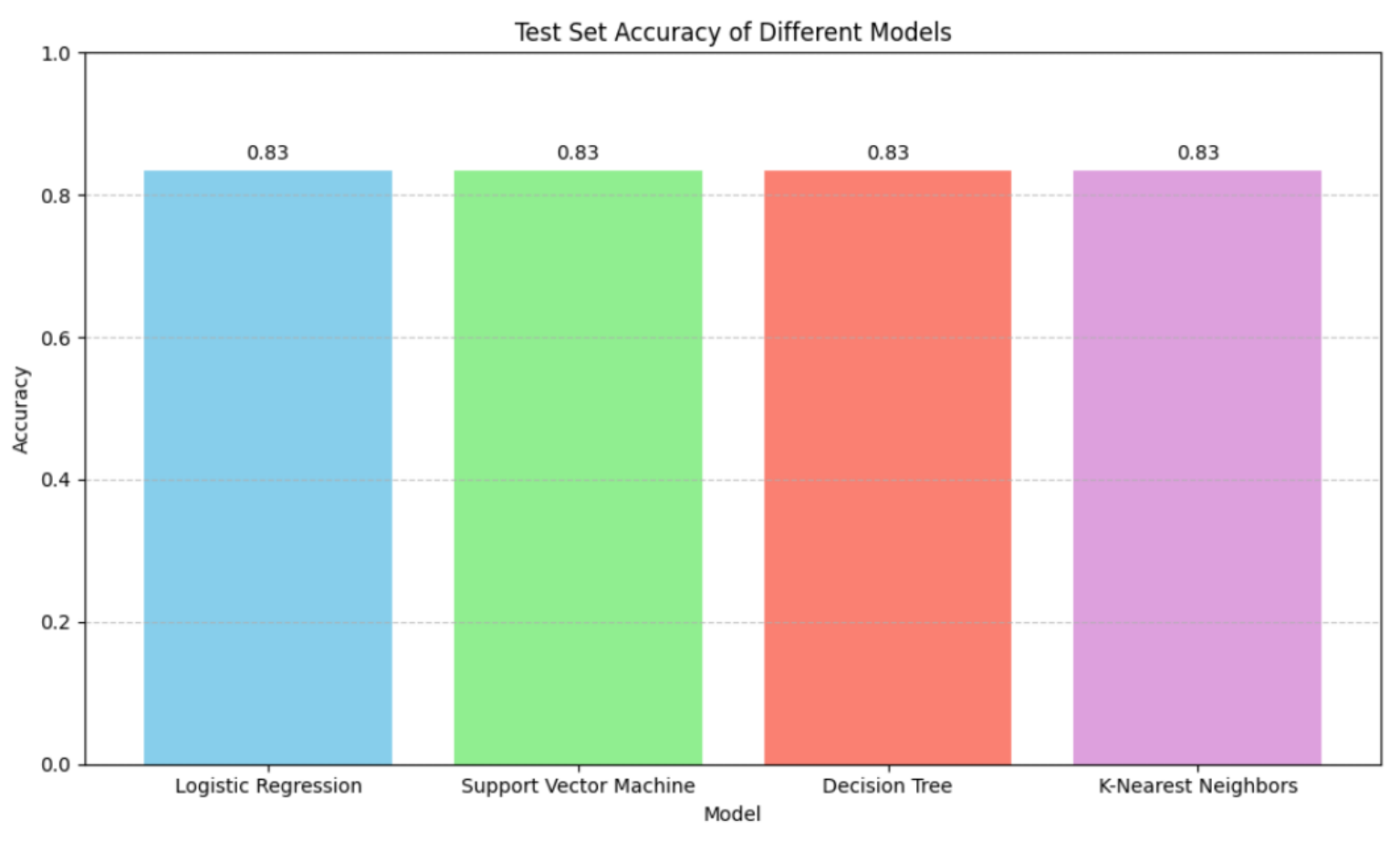
Section 5

Predictive Analysis (Classification)

MADE BY PRASHANT YADAV

Classification Accuracy

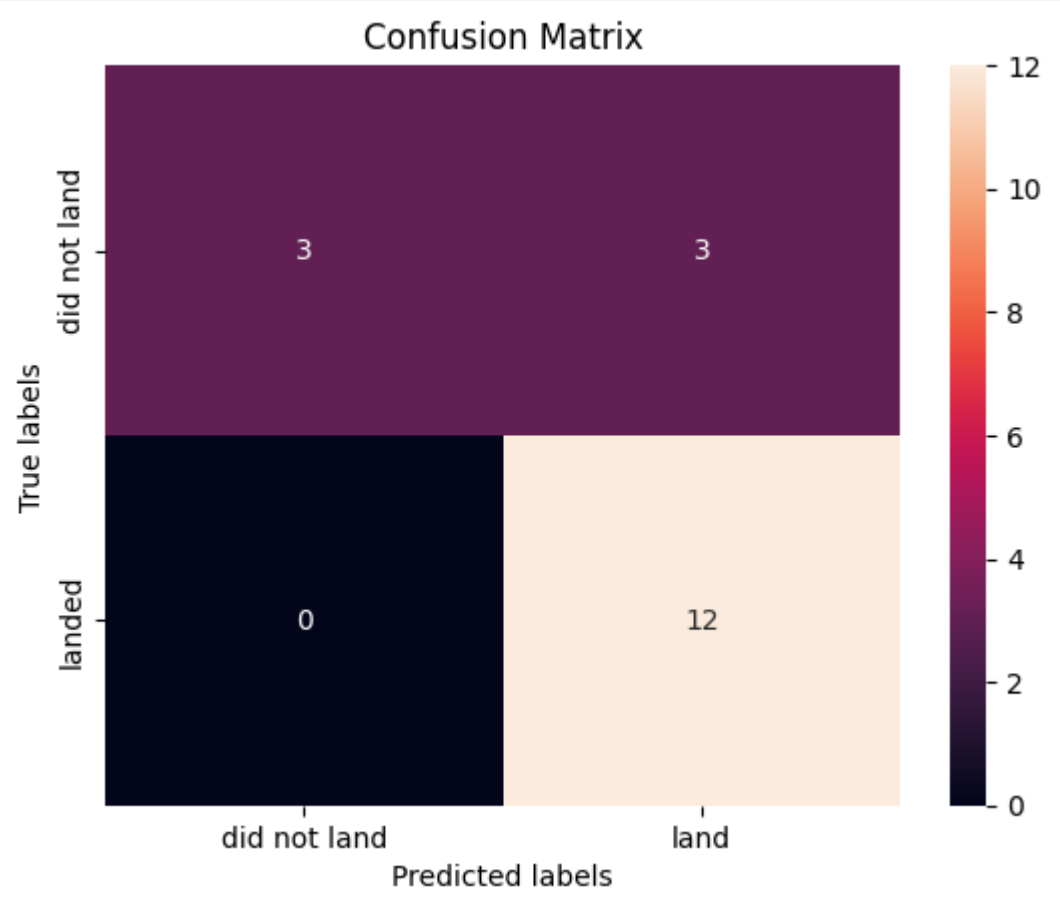
- Although the CV Scores differ, each model has the same accuracy on Test Set!
- Is possible since all models learn from the same info.
- Also, we had standardized X



MADE BY PRASHANT

Confusion Matrix

- Since ALL models plateaued at 83.33% accuracy, selecting the first one- Logistic Regression.



MADE BY PRASHANT YADAV

Conclusions

1. Conducted thorough exploratory analysis to uncover site-specific trends, payload dynamics, and mission outcomes using interactive visualizations.
2. Designed an intuitive dashboard enabling real-time filtering and user-driven insights across launch parameters and success metrics.
3. Developed and fine-tuned multiple classification models, with SVM and Decision Tree achieving top performance at 83.3% test accuracy.
4. Established a repeatable pipeline from data wrangling to predictive modeling, showcasing the potential of data-driven insights in aerospace analytics.

MADE BY PRASHANT YADAV

Appendix

Code Assets:

- Data preprocessing, model training (LogReg, SVM, Tree, KNN), GridSearchCV
- Fixed invalid "auto" value in max_features for Decision Tree tuning
- Accuracy comparison bar chart and confusion matrix heatmaps

SQL Queries:

- Launch site filtering, payload aggregations, outcome counts, NASA CRS missions

Visuals & Outputs:

- EDA and dashboard charts
- Model evaluation plots and test set metrics

MADE BY PRASHANT YADAV

Thank you!

MADE BY PRASHANT YADAV