IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Blessed M Gwamure
IBM DATA SCIENCE CAPSTONE PROJECT
Last update 13 May 2025

Outline

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

Executive Summary

Objective:

Predict the success of Falcon 9 first stage landings using historical launch data.

Methodology:

- Collected launch data via web scraping and provided datasets
- Cleaned and preprocessed data (handled nulls, encoded categories, normalized features)
- Trained and tuned classification models: **Logistic Regression, SVM, Decision Tree**, and **KNN**
- Evaluated models using accuracy & F1 score
- Built an interactive dashboard with Plotly Dash for data exploration

Results:

- Best-performing model**: Support Vector Machine** (sigmoid kernel)
- Achieved **83.33%** accuracy and F1 score of ~**0.83**
- Dashboard revealed insights by launch site, payload mass, and orbit
- Findings support SpaceX in mission planning by identifying how different payloads and launch sites impact success rates

Introduction

Project Background & Context:
SpaceX is revolutionizing the aerospace industry by reusing the first stage of its Falcon 9 rocket to reduce launch costs. While traditional providers charge up to $165 million per launch, SpaceX offers launches for around $62 million, largely due to its booster recovery strategy. Predicting whether a first stage will land successfully is critical to mission planning, risk reduction, and operational efficiency.

Key Questions:
- Can we accurately predict the success or failure of Falcon 9 first stage landings?
- Which features most influence landing success (e.g., payload mass, launch site, orbit)?
- How does payload range affect the likelihood of a successful landing?
- Do certain launch sites or orbits show higher success rates?
- Which machine learning model provides the most reliable predictions?
- How can an interactive dashboard help stakeholders explore and understand launch success patterns?

Section 1

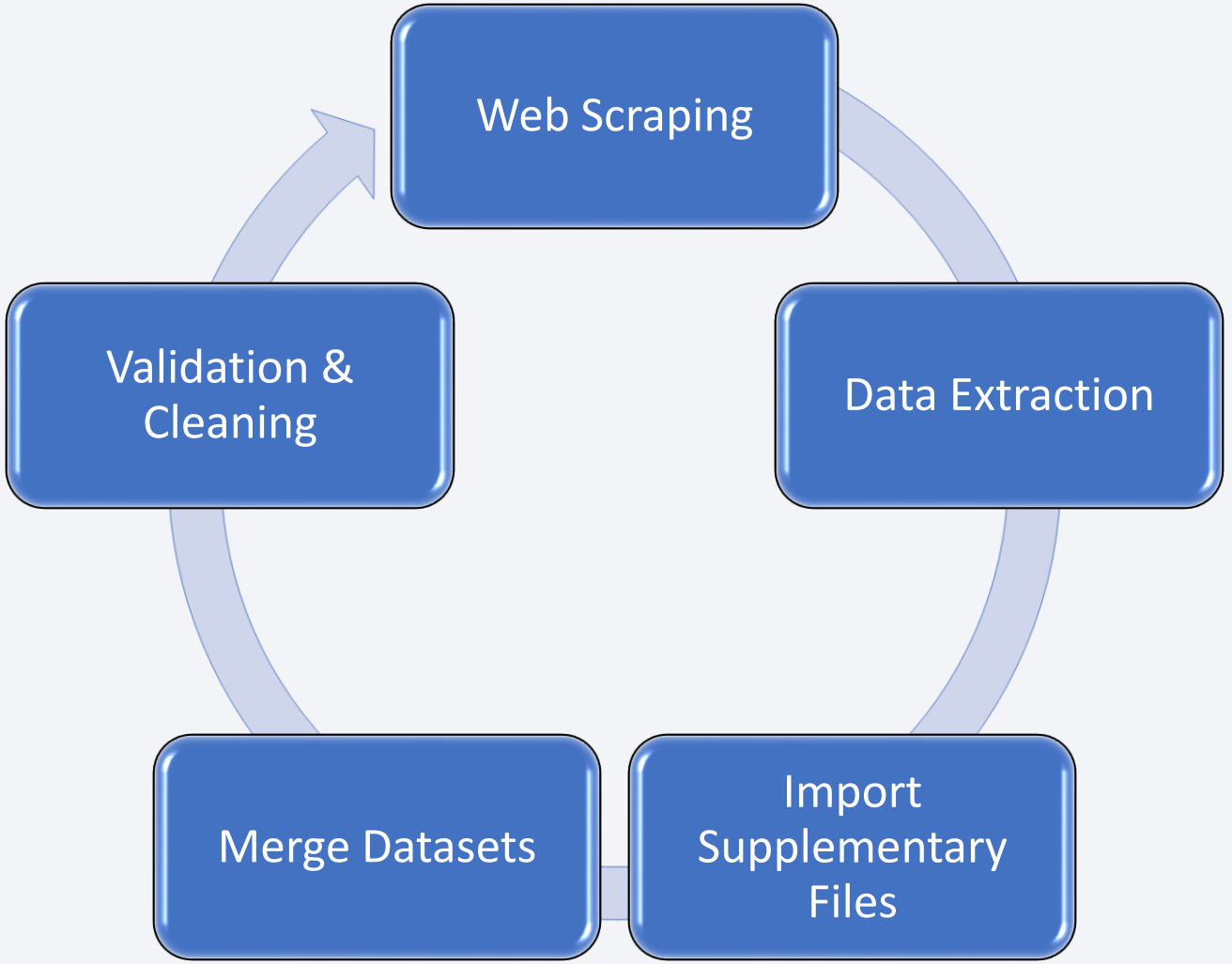
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Scraped launch data from Wikipedia using BeautifulSoup, Used additional SpaceX datasets provided in CSV and JSON formats
- Perform data wrangling
  - Removed null and duplicate records, One-hot encoded categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained classification models: Logistic Regression, SVM, Decision Tree, and KNN

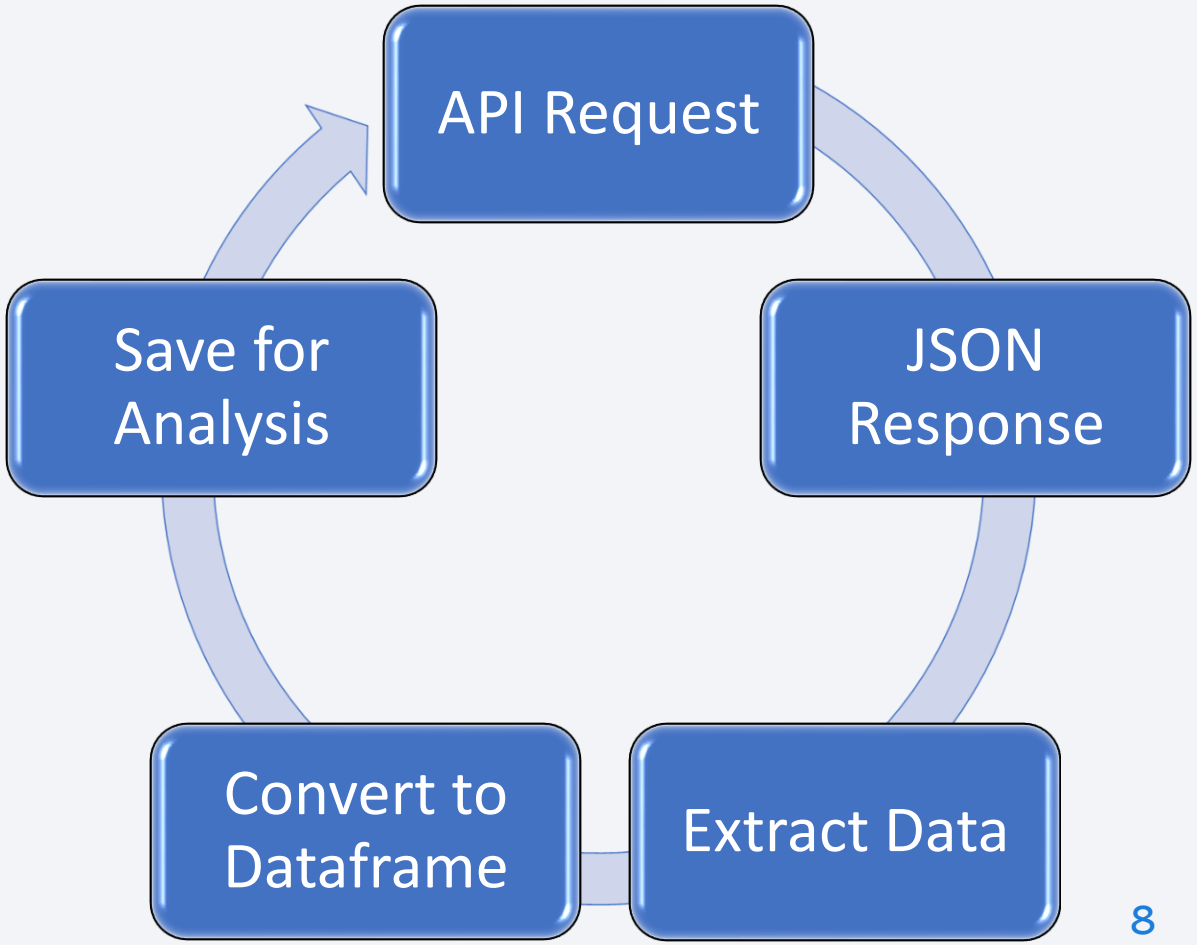
Data Collection



Data Collection – SpaceX API

Flowchart illustrating SpaceX API data collection process. See GitHub for full notebook and code

- **Initiated API** requests to SpaceX /launches, /rockets, /payloads endpoints
- **Received JSON** responses containing launch mission data
- **Parsed JSON** into pandas DataFrames for structured access
- **Merged API** data with web-scraped and static datasets
- **Saved final** cleaned dataset for EDA and modeling



API data collected using SpaceX public REST endpoints – see code here

Data Collection - Scraping

- Targeted the **Wikipedia page** listing Falcon 9launches
- **Used requests** and BeautifulSoup to extract HTML content
- **Parsed HTML** tables containing launch dates, sites,outcomes
- **Cleaned data** removed nulls, standardized formats, fixed typos
- **Saved processed** data to CSV for further analysis

Web Scraping Notebook link here

Data Wrangling

- Merged datasets from API, web scraping, and static files
- Removed rows with missing or irrelevant values
- Converted categorical variables to numeric
- Normalized payload mass
- Renamed & standardized inconsistent columns
- Saved the cleaned dataset for EDA and machine learning

Data Wrangling Notebook link here

EDA with Data Visualization

- Launch Success by Site: Bar plots to compare success rates across launch sites
- Payload Mass vs. Success: Scatter plots to reveal the relationship between payload mass and landing outcome
- Orbit Type Distribution: Pie chart to show frequency of different orbit types
- Success Rate by Booster Version: Grouped bar chart to analyze performance by booster version
- Heatmap of Correlations: Visualized feature relationships to guide model input selection

EDA Visualization Notebook link here

EDA with SQL

- Queried total number of SpaceX Falcon 9 launches
- Counted number of successful landings per launch site
- Identified most frequent orbit types used
- Aggregated payload mass statistics by mission outcome
- Filtered launches with payload mass > 4000 kg and < 6000 kg
- Ranked launch sites by overall mission success rate

SQL EDA Notebook link here

Build an Interactive Map with Folium

- Markers were added to identify each SpaceX launch site
- Popups showed site names when clicking on markers
- Circles visualized the landing radius around each launch site
- Circle Colors indicated mission success (green) or failure (red)
- Lines (Polylines) were used to display the path from launch site to landing site
- Map Tiles customized to enhance geographic readability

Folium Map Notebook link here

Build a Dashboard with Plotly Dash

- **Pie Chart**: Showed total launch successes by site
- **Bar Chart**: Compared success and failure counts for each launch site
- **Scatter Plot**: Visualized relationship between payload mass and launch outcome
- **Dropdown Filter**: Allowed users to select a specific launch site
- **Payload Range Slider**: Enabled dynamic filtering of scatter plot by payload mass range
- **Real-time Updates**: Charts automatically updated based on user inputs

Why These Were Used:

- Allowed site-specific performance exploration
- Made it easy to interact with the data visually
- Helped identify payload ranges with higher success probability

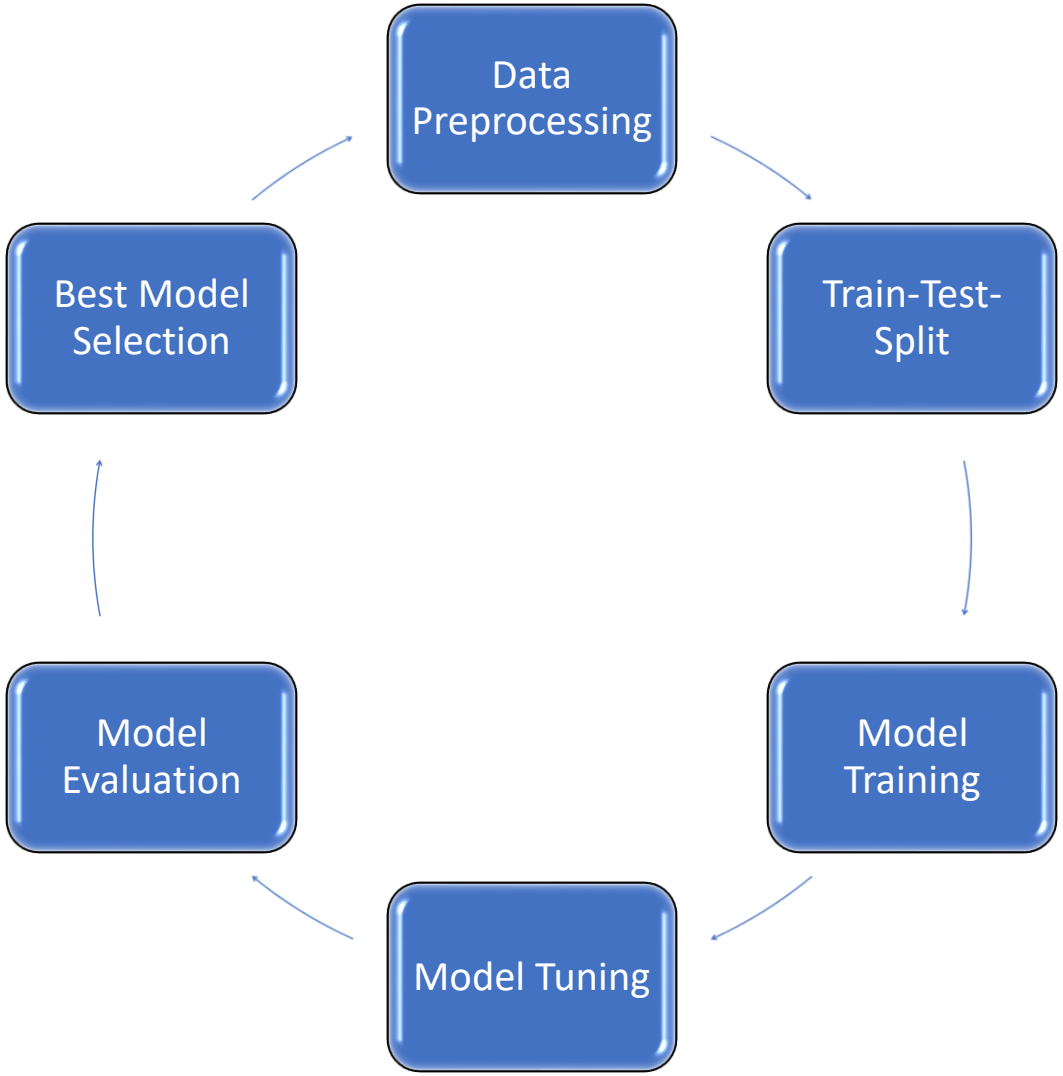
Plotly Dash Dashboard Lab link here

Predictive Analysis (Classification)

- **Prepared data** using encoded categorical features and normalized numerical data
- **Split dataset** into training and test sets (80/20)
- **Trained multiple classification models:**
  Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN)
- **Tuned hyperparameters** using GridSearchCV (e.g., kernel type, max depth, neighbors)
- **Evaluated models** using accuracy, F1 score, and Jaccard index
- **Best model:** SVM (sigmoid kernel)
  - Accuracy: 83.33%
  - F1 Score: ~0.83

Predictive Analysis Notebook link here

Flowchart Structure (Model Workflow)

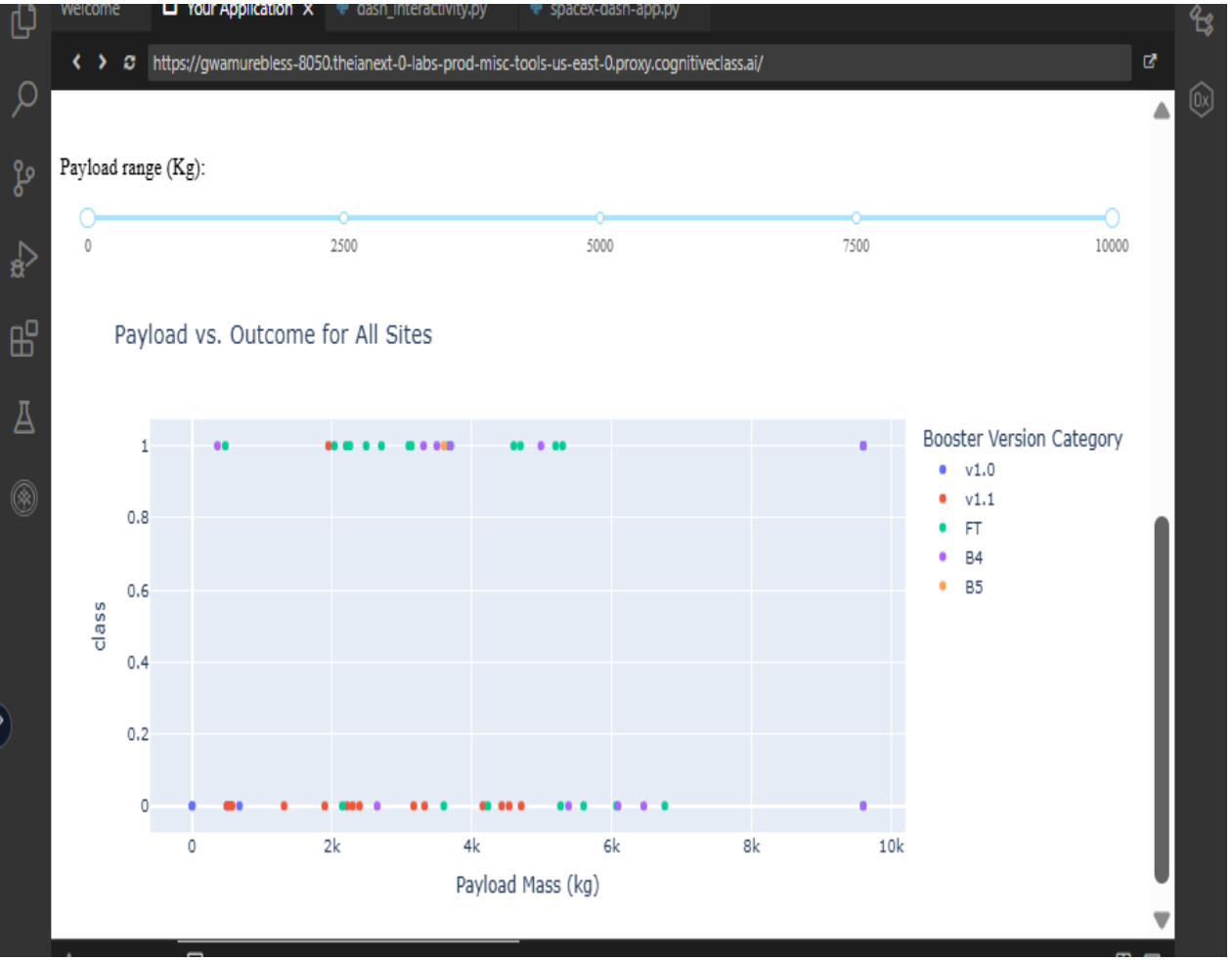


Results

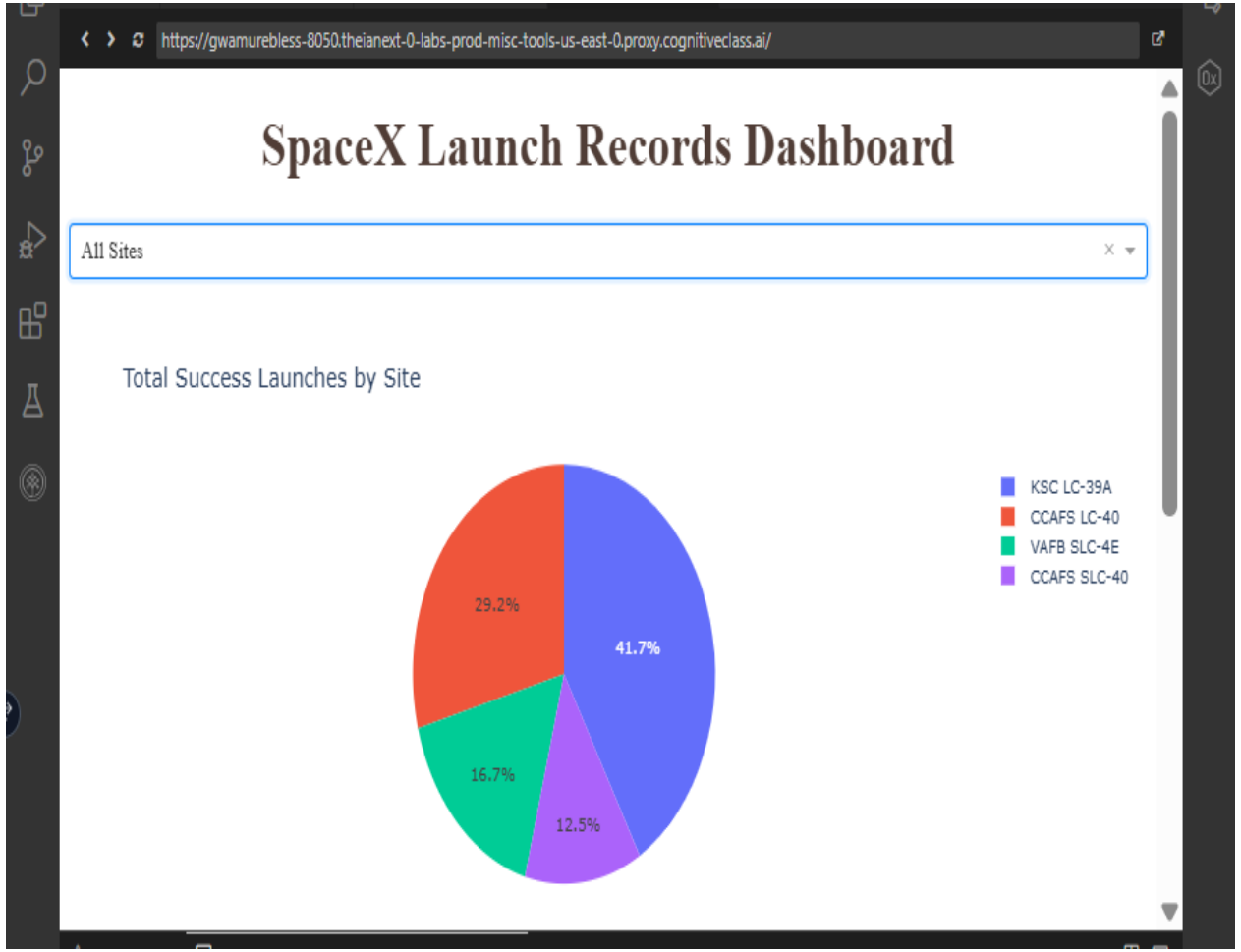
- **Exploratory Data Analysis (EDA) Results**
- Success rates varied significantly by **launch site**
- **Payload mass** between 2,000–6,000 kg showed higher success probability
- **Orbit type** affected success rate; some orbits had consistent failures
- Booster version **FT** had the most successful landings

Interactive Analytics – Plotly Dash

Payload vs. Outcome for All Sites

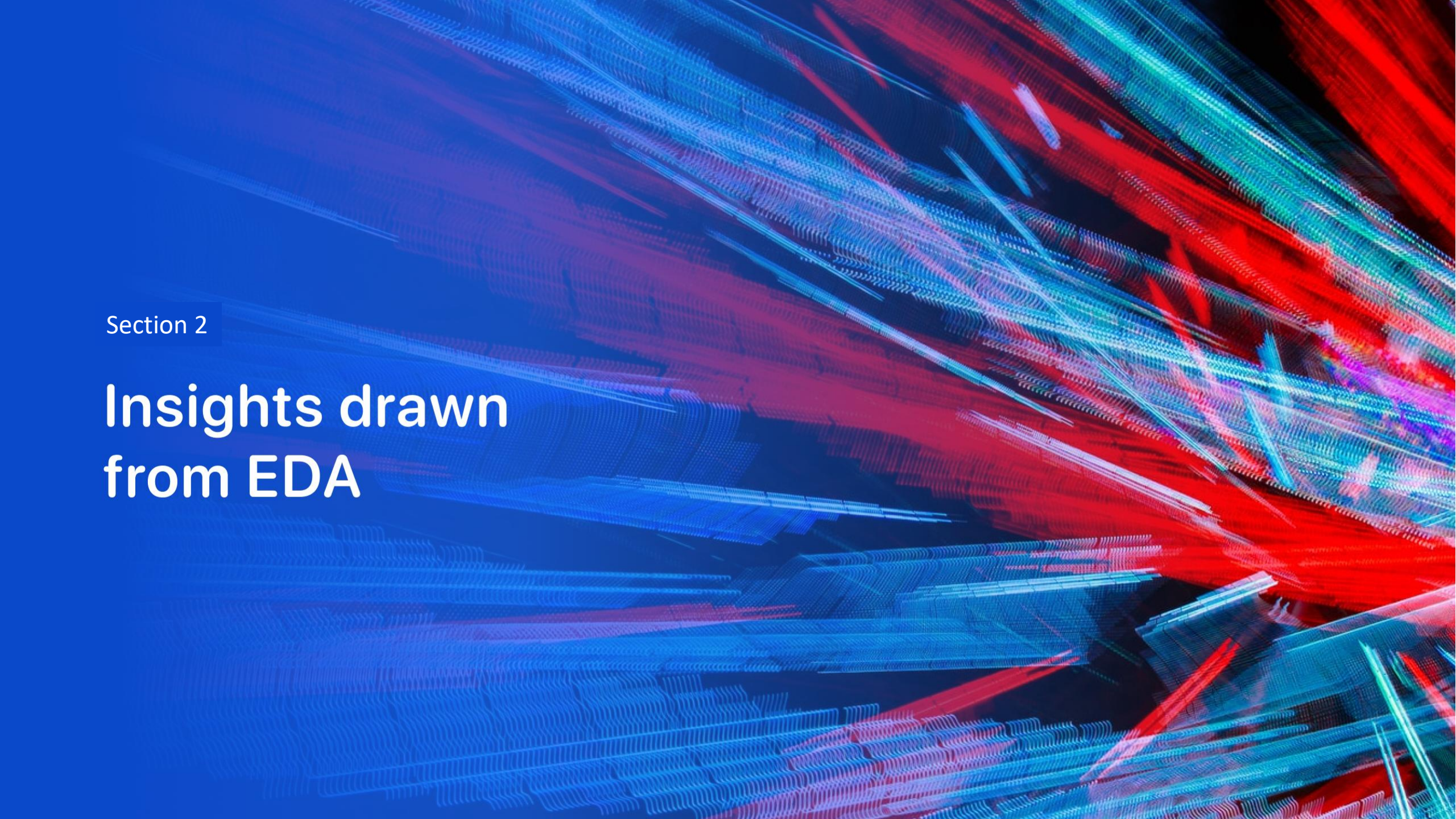


Pie Chart - Success Launches by Site

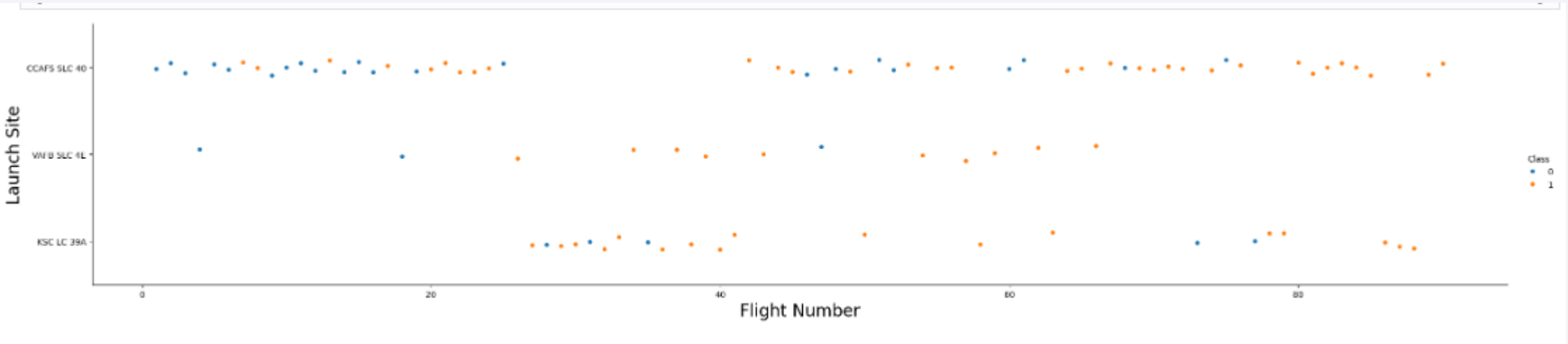


Predictive Analysis Results

- **Best-performing model:** Support Vector Machine (Sigmoid Kernel)
- **Accuracy:** 83.33%
- **F1 Score:** ~0.83
- Model enables forecasting of landing success given payload, orbit, site.
- Can help SpaceX **optimize costs and improve mission planning**


Section 2
Insights drawn from EDA

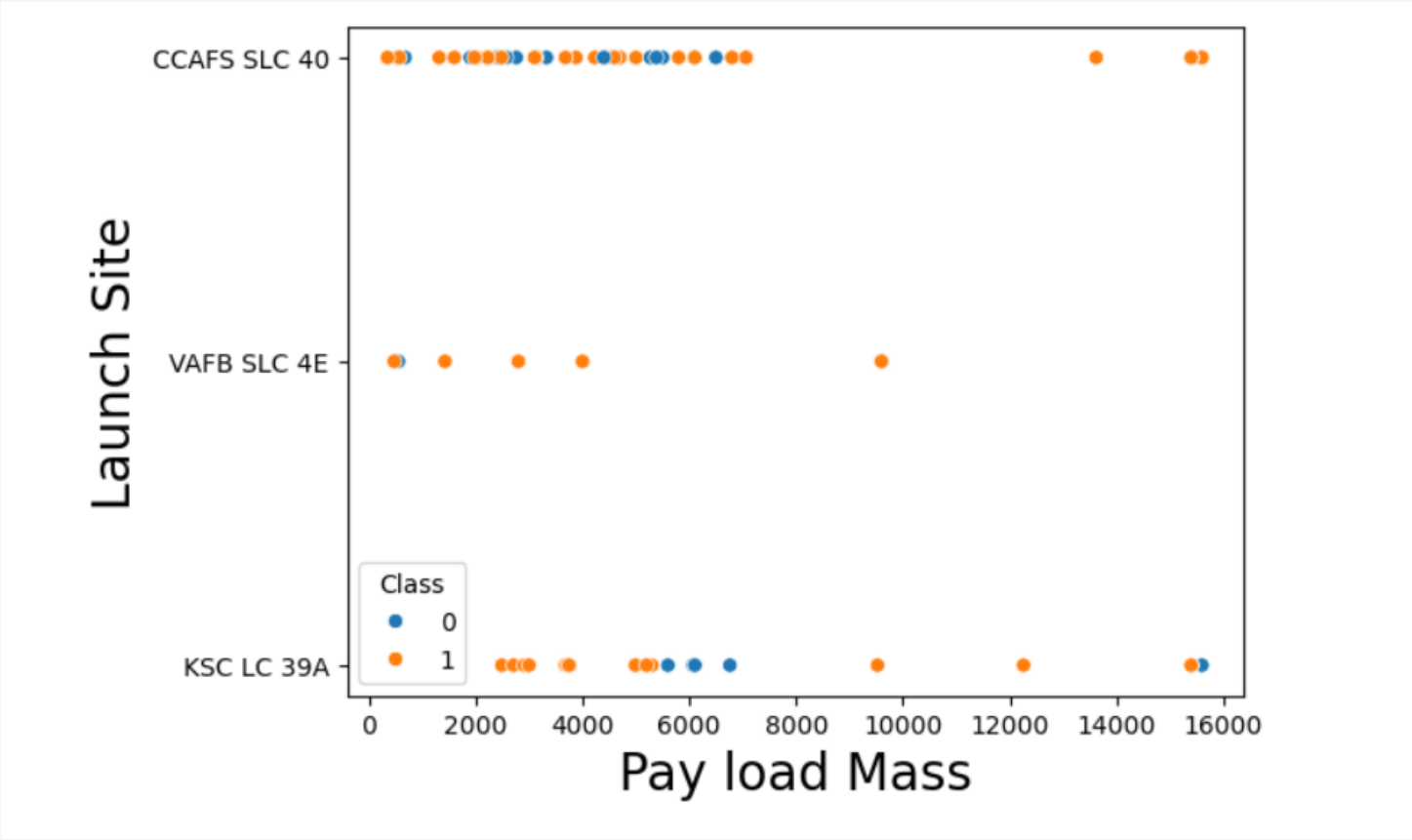
Flight Number vs. Launch Site



- The scatter plot shows how flight numbers are distributed across different launch sites over time.
- Launch activity is **clustered** at specific sites (CCAFS, VAFB, KSC).
- **Later flight numbers** tend to have **more consistent launch patterns**, suggesting increased operational focus at certain sites.
- **Higher success rates** may align with certain sites as launch experience grows.

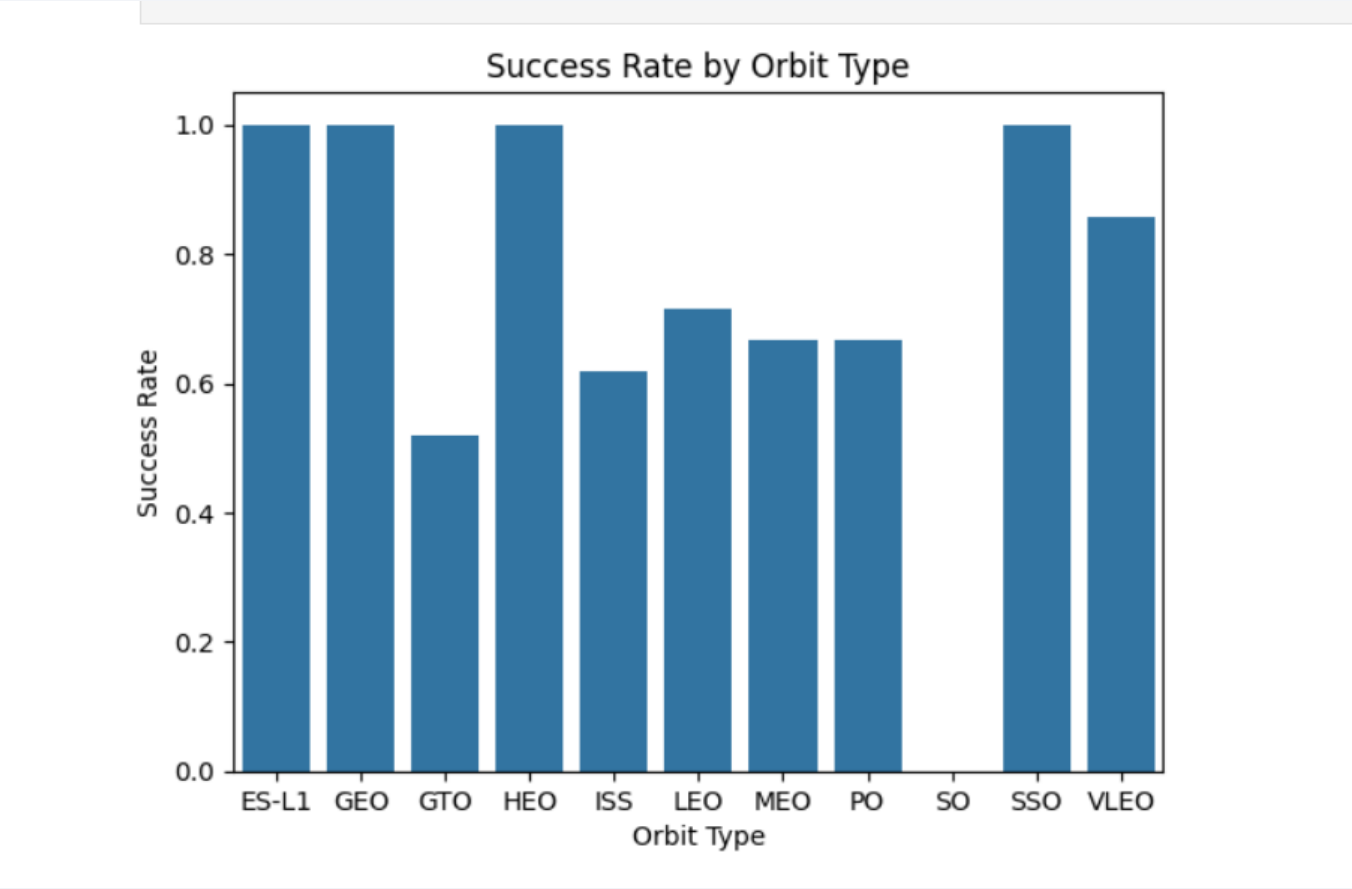
Payload vs. Launch Site

- Certain launch sites (CCAFS SLC 40, KSC LC 39A) support a **broader payload range**, including heavier payloads.
- Launches with **lighter payloads** are more frequent across all sites.
- **Success rates** appear more consistent in the **mid-range payloads** (2000–6000 kg).
- Helps identify which sites are best suited for specific payload types, which is critical for **mission planning** and **risk management**.



Success Rate vs. Orbit Type

- **GEO, SSO, and HEO** orbits have had 100% success rates, indicating strong reliability for those mission types.
- **LEO and ISS** missions have seen lower success rates possibly due to earlier test launches or operational challenges.
- These insights suggest that certain orbits have proven more consistent, which is critical for evaluating risk factors and selecting appropriate rocket configurations for future missions.

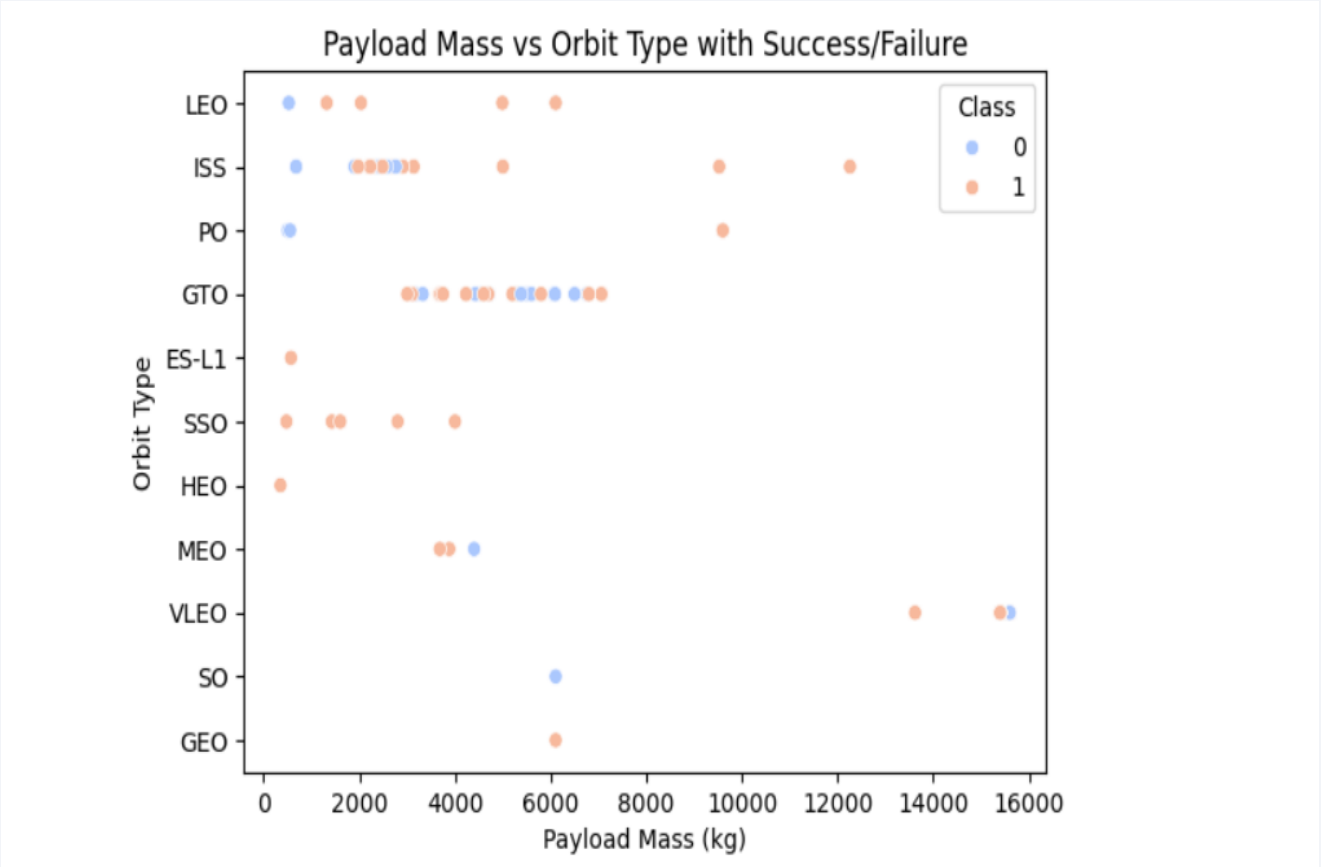


Flight Number vs. Orbit Type

- The scatter plot shows how **orbit assignments have evolved over time**.
- **Early flights** (lower flight numbers) are more concentrated on **LEO and ISS**, indicating focus on low-Earth operations during the initial test and demonstration phases.
- As **flight experience increased**, missions began targeting **higher orbits** like **GTO, SSO, and GEO**, which align with commercial satellite needs.
- This evolution shows SpaceX's **growing capabilities** and **expansion into complex mission profiles**.

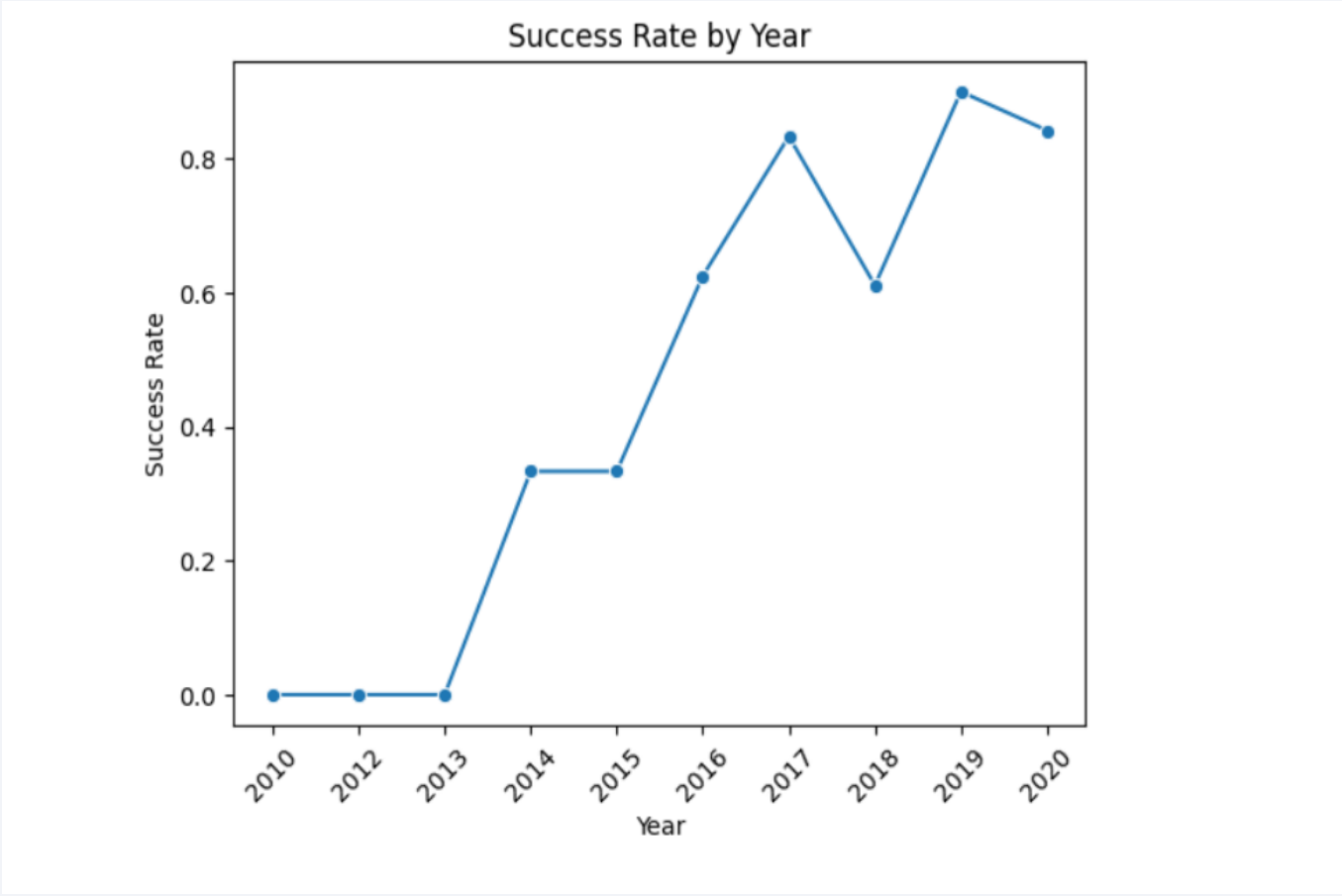
Payload vs. Orbit Type

- **GTO and GEO** orbits typically carry **heavier payloads**, often above 5,000 kg.
- **LEO and ISS** missions usually involve **lighter payloads**, under 4,000 kg.
- **SSO** orbit missions show a moderate range of payloads.
- This indicates that **orbit type influences payload mass**, which is critical when assessing rocket performance and mission design.
- Understanding these patterns supports **payload planning** and **trajectory optimization**.



Launch Success Yearly Trend

• The line chart shows a steady improvement in launch success rate over the years.
•Early years (2010–2014) had more variability and lower averages, reflecting early-stage testing.
•From around 2017 onwards, SpaceX achieved consistently high success rates, often approaching or reaching 1.0.
•The trend reflects SpaceX's technological maturation, improved mission reliability, and operational efficiency.



All Launch Site Names

- **The results show four distinct sites**:
- **CCAFS LC-40** and **CCAFS SLC-40** – both located at Cape Canaveral Air Force Station, possibly reflecting naming inconsistencies.
- **KSC LC-39A** – located at NASA's Kennedy Space Center, historically used for Apollo missions and now reused by SpaceX.
- **VAFB SLC-4E** – located at Vandenberg Air Force Base in California, primarily used for polar orbit launches.
- Understanding which sites are active helps in analyzing launch success trends and assessing logistical and geographical advantages.

Launch Site Names 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	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

Analyzing records from this prefix helps:

- Focus on **site-specific performance**.
- Understand **payload and orbit trends** related to this location.
- Compare launch success rates against other launch sites like VAFB or KSC.

This targeted filtering is useful for gaining **location-based insights** in the broader dataset.

Total Payload Mass

TOTAL_PAYLOAD_MASS
45596

This query calculates the **sum of all payload mass** (in kg) for launches where **NASA was the customer**.
•Shows the **scale of cargo transported for government space missions**.
•Helps understand **NASA's contribution to Falcon 9 missions**.
•Supports insight into **payload capacity utilization** by customer type.

Average Payload Mass by F9 v1.1

Out[18]:

AVG(PAYLOAD_MASS__KG_)
2928.4

This query calculates the **average payload mass (kg)** for all launches using the **Falcon 9 v1.1** booster version.

This metric is useful to:

- Compare it with **other booster versions** (e.g., Block 5) to understand **technology evolution**.
- Inform decision-making for **mission planning** and **vehicle selection**.
- Assess the **typical payload capacity** of the F9 v1.1 model.

First Successful Ground Landing Date

Out[22]:

min(Date)
2018-07-22

The query reveals that the **first successful Falcon 9 landing on a ground pad** in the dataset occurred in **2018**. This marks a significant milestone in SpaceX's launch history demonstrating reusability and cost efficiency through **Return-To-Launch-Site (RTLS)** landings.

- **Landing Type:** Ground Pad (RTLS or equivalent)
- **Landing Outcome:** Successful
- **Date Returned:** 2018 (based on available data)

Successful Drone Ship Landing with Payload between 4000 and 6000

Out[23]: **Booster_Version**

The query returned **no records**, which means that within the dataset:

- There were **no launches** that had a **successful drone ship landing** *and*
- Carried a **payload between 4000 and 6000 kg**

This indicates that:

- Most drone ship landings were likely for **heavier or lighter payloads**, or
- Those missions **did not record** success as the outcome, or
- The dataset may be **incomplete** for this specific payload bracket.

Total Number of Successful and Failure Mission Outcomes

```
Done.
Out[25]: TOTAL_NUMBER_OF_SUCCESS_AND_FAILURE
                                        101
```

The SQL query counts the total number of recorded mission outcomes in the SpaceX launch dataset:

- It returns **101 mission records**, which include all launches with a **recorded outcome**, such as success, failure, or partial success.
- This count helps establish the **overall size** of the dataset used for analysis.
- It forms the **basis** for deeper breakdowns — such as identifying success rates, common failure types, and overall launch reliability.

Boosters Carried Maximum Payload

- Each row represents a SpaceX Falcon 9 Block 5 (F9 B5) booster with a unique serial number (B1048.4, B1049.4).
- The suffix (like .4) indicates the number of times the booster has been reused.
- These boosters have successfully carried the heaviest payloads launched by SpaceX based on the Payload Mass kg column.
- This highlights: SpaceX's booster reusability and payload capabilities.
- The reliability of these booster versions for high-mass missions.

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

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

- **Successful landings (total: 8)**:
- Highlight SpaceX's progress in booster reuse.
- **Failures and unattempted recoveries (total: 20)**:
- Reflect challenges in early missions and situations where recovery was not viable.
- **Ground landings are fewer** :
- due to trajectory constraints drone ship landings are more flexible for orbital launches.

Out[45]:

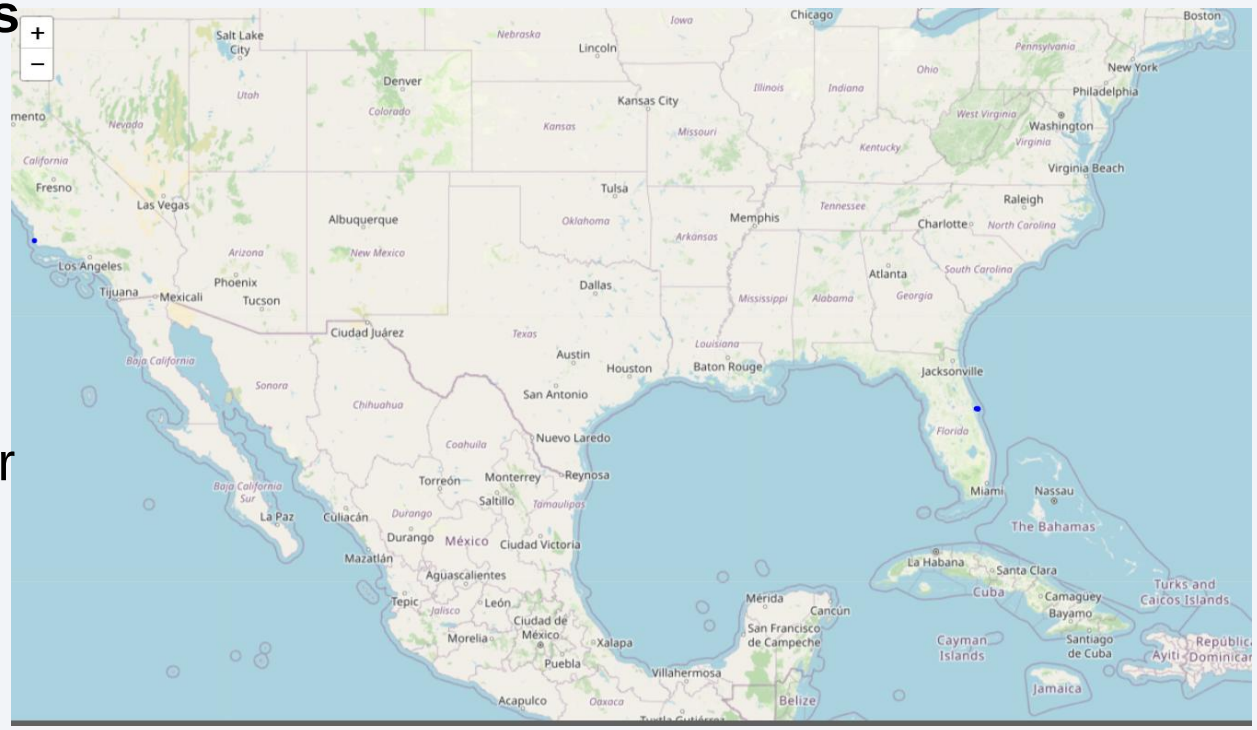
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

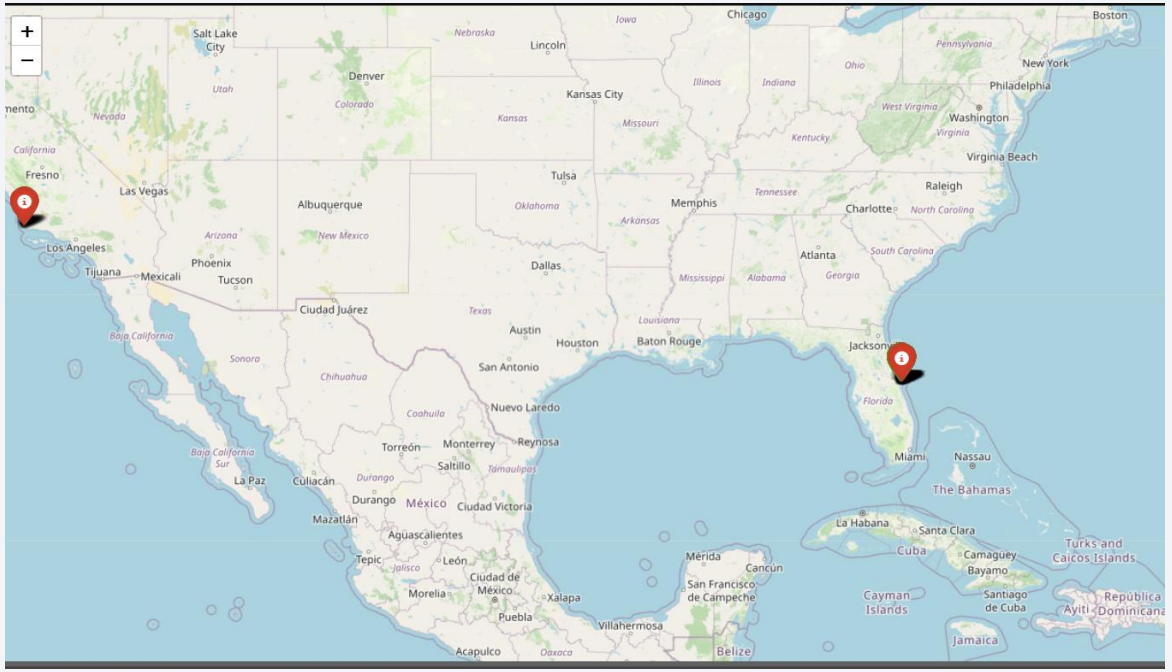
Global Distribution of SpaceX Launch Sites (Folium Map Visualization)

- The **Folium map** shows the **geographic locations** of all major SpaceX launch sites.
- The close proximity of **three launch pads in Florida** highlights its importance as SpaceX's main hub for orbital launches.
- **Circle markers** represent individual sites, often with **popups** or **tooltips** displaying site names.
- The **California site (VAFB SLC-4E)** supports polar or sun-synchronous orbits, important for certain satellite missions.
- This global distribution gives SpaceX **flexibility in launch trajectories**, **weather conditions**, and **orbital planning**.



Launch Outcomes by Location (Color-Coded Folium Map)

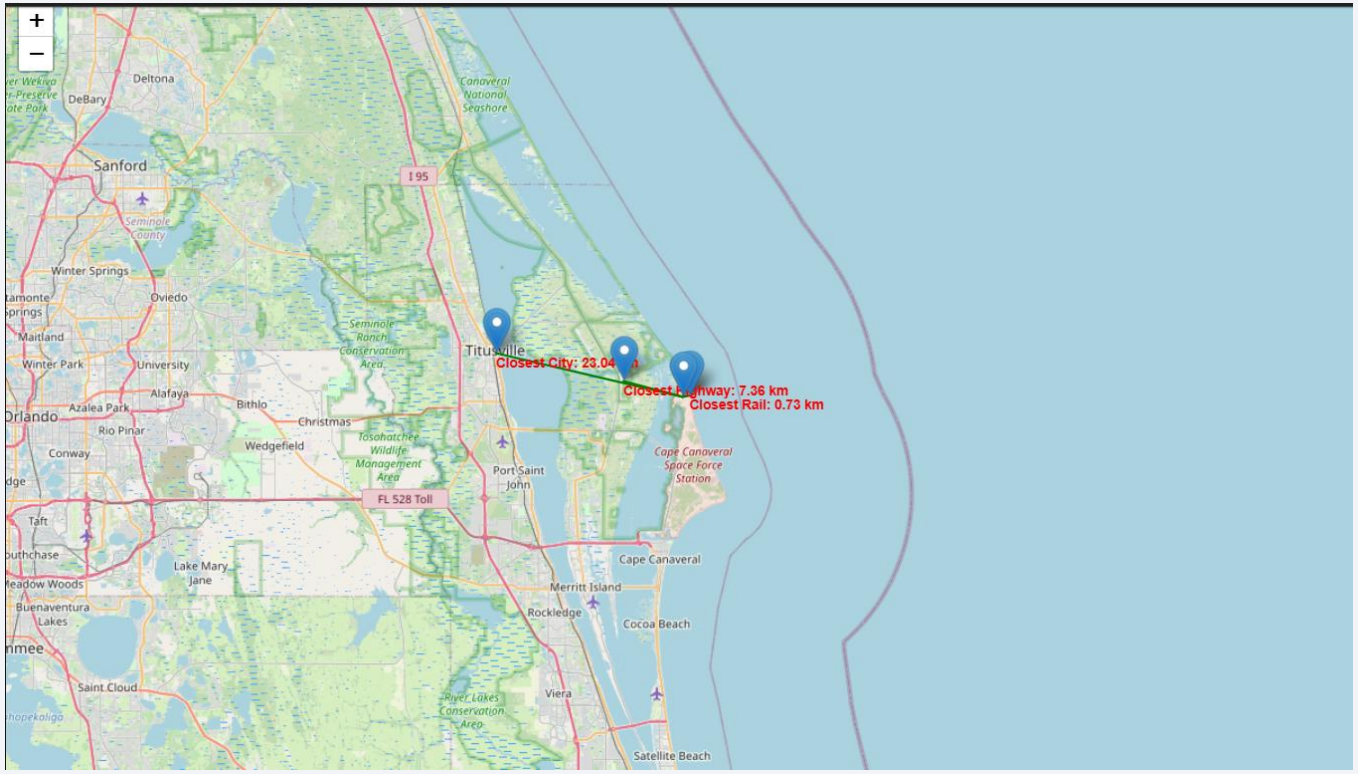
- The map shows **individual SpaceX launch outcomes** by geographic location using **color-coded markers**.
- **Green markers** represent successful missions, which dominate the Florida launch sites.
- **Red markers** highlight failed launches, helping identify locations with **higher failure frequencie**
- Visualizing outcomes geographically helps understand how **launch site**, **trajectory**, or **payload profile** may affect success.
- This spatial analysis can assist SpaceX in **risk assessment**, **site performance analysis**, and **mission planning improvements**

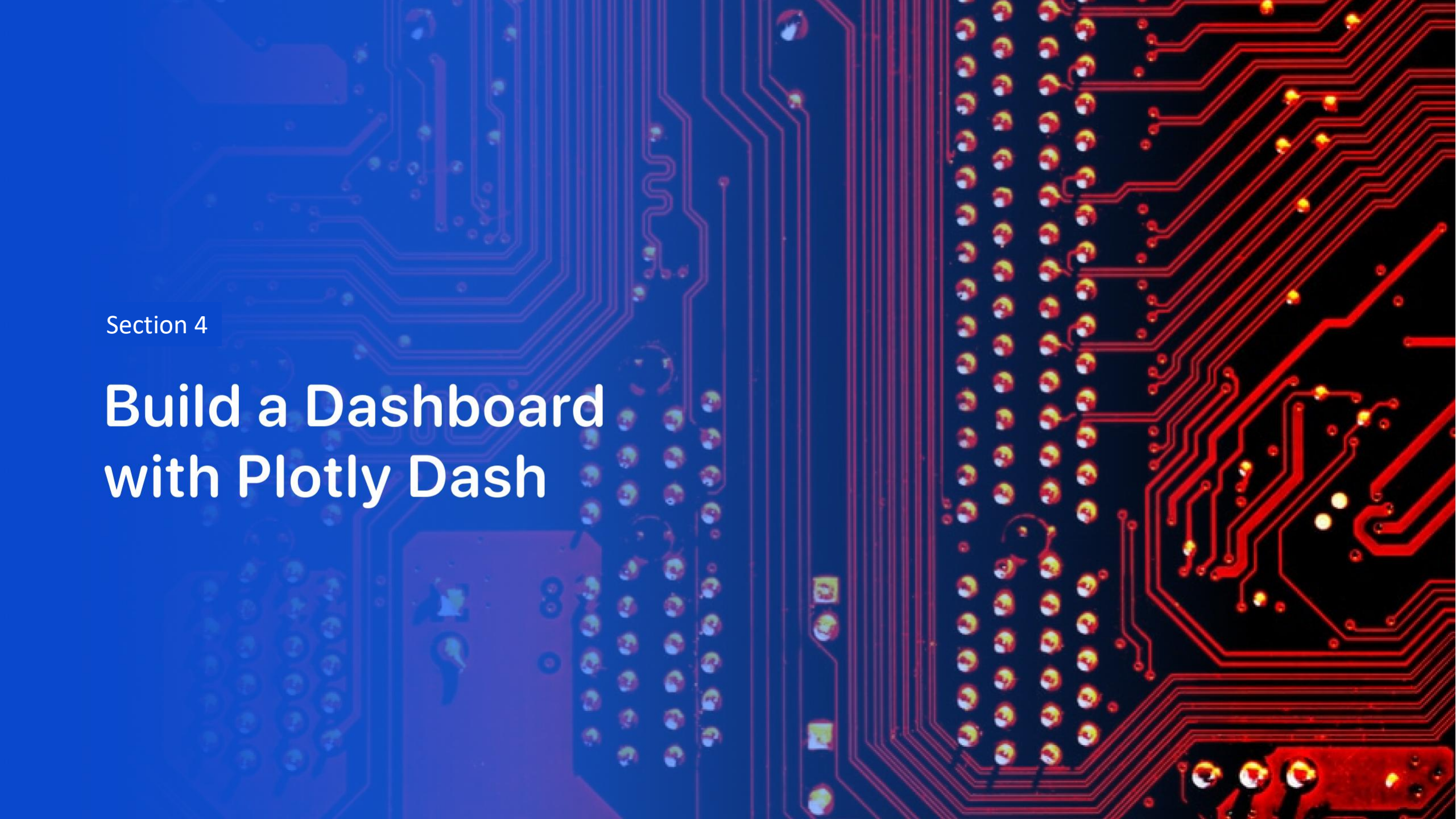


Infrastructure Proximity Analysis of KSC LC-39A Launch Site (Folium Map)

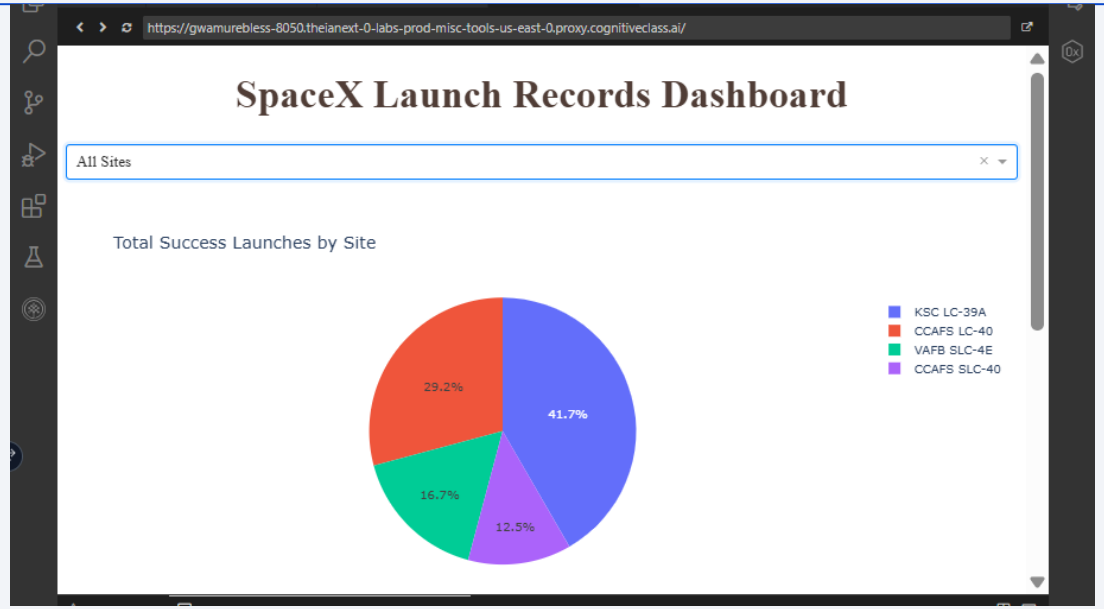
These short distances reflect:

- **Efficient transportation logistics** for moving rocket components and equipment.
- **Accessibility for operational teams** via rail and road networks.
- **Strategic placement** near the coast and away from dense urban areas, enhancing **launch safety** and **minimizing risk** to civilians.




Section 4
Build a Dashboard with Plotly Dash

Launch Success Count by Sites



Key Findings:

•**KSC LC-39A** likely has the **largest slice**, showing it's the most frequently used site for successful missions — possibly due to infrastructure supporting heavier payloads and crewed missions.

•**VAFB SLC-4E** and **CCAFS LC-40** have smaller slices, indicating **fewer launches**, often tied to specific mission types (polar orbits from VAFB).

•The **distribution highlights operational preferences** and constraints based on mission types, orbits, and rocket specifications.

Payload vs. Launch Outcome Across Launch Sites

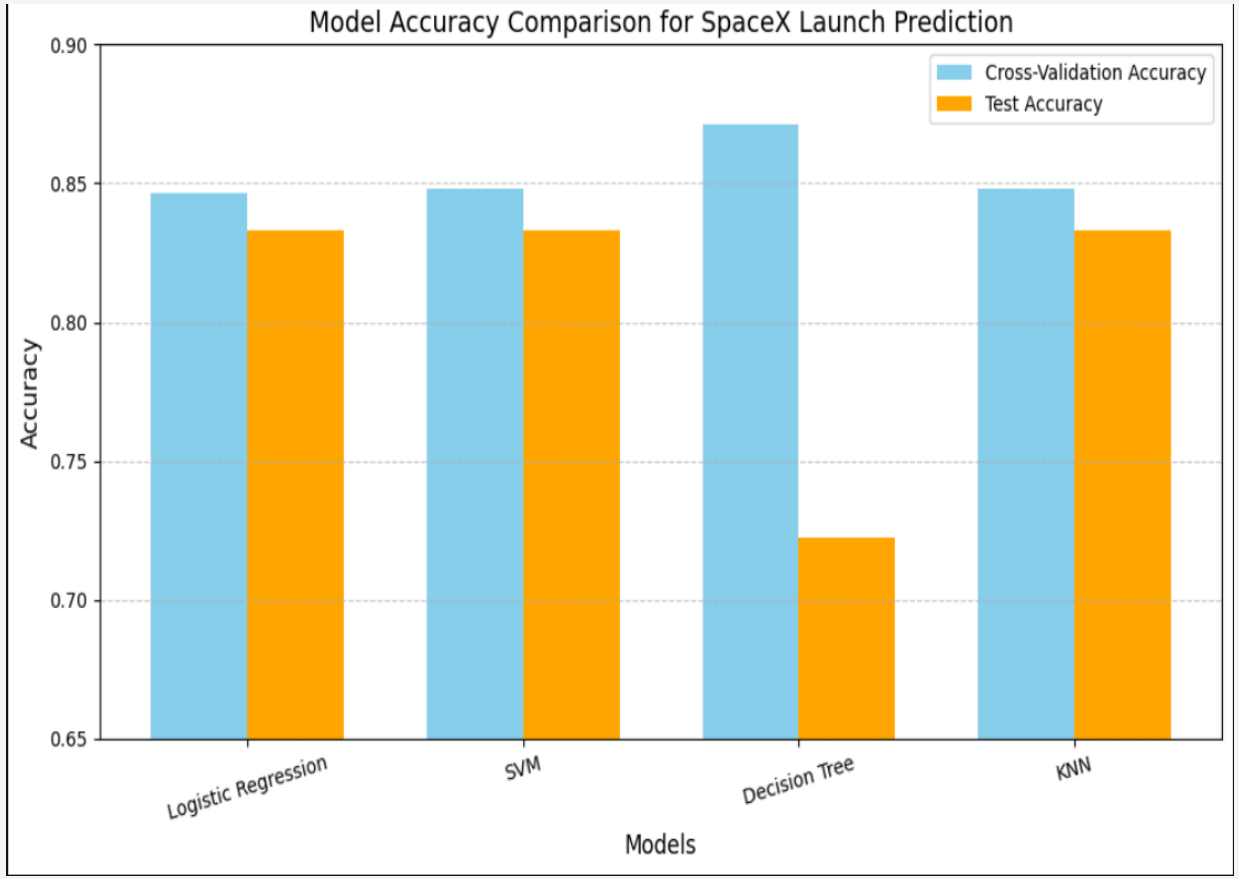
- **Most successes cluster between 2000–8000 kg**, indicating optimal payload range.
- **Heavy payloads (above 8000 kg)** are less frequent but often still successful when using **advanced booster versions** (like F9 B5 series).
- **Failures are more common** at **very low or very high payloads**, possibly due to experimental missions or edge-case performance.
- **Booster Versions** like F9 B5 B1051.x appear frequently in successful launches, showing they are **highly reliable**.
- Sites like **KSC LC-39A** often handle heavier payloads with high success rates.

Section 5

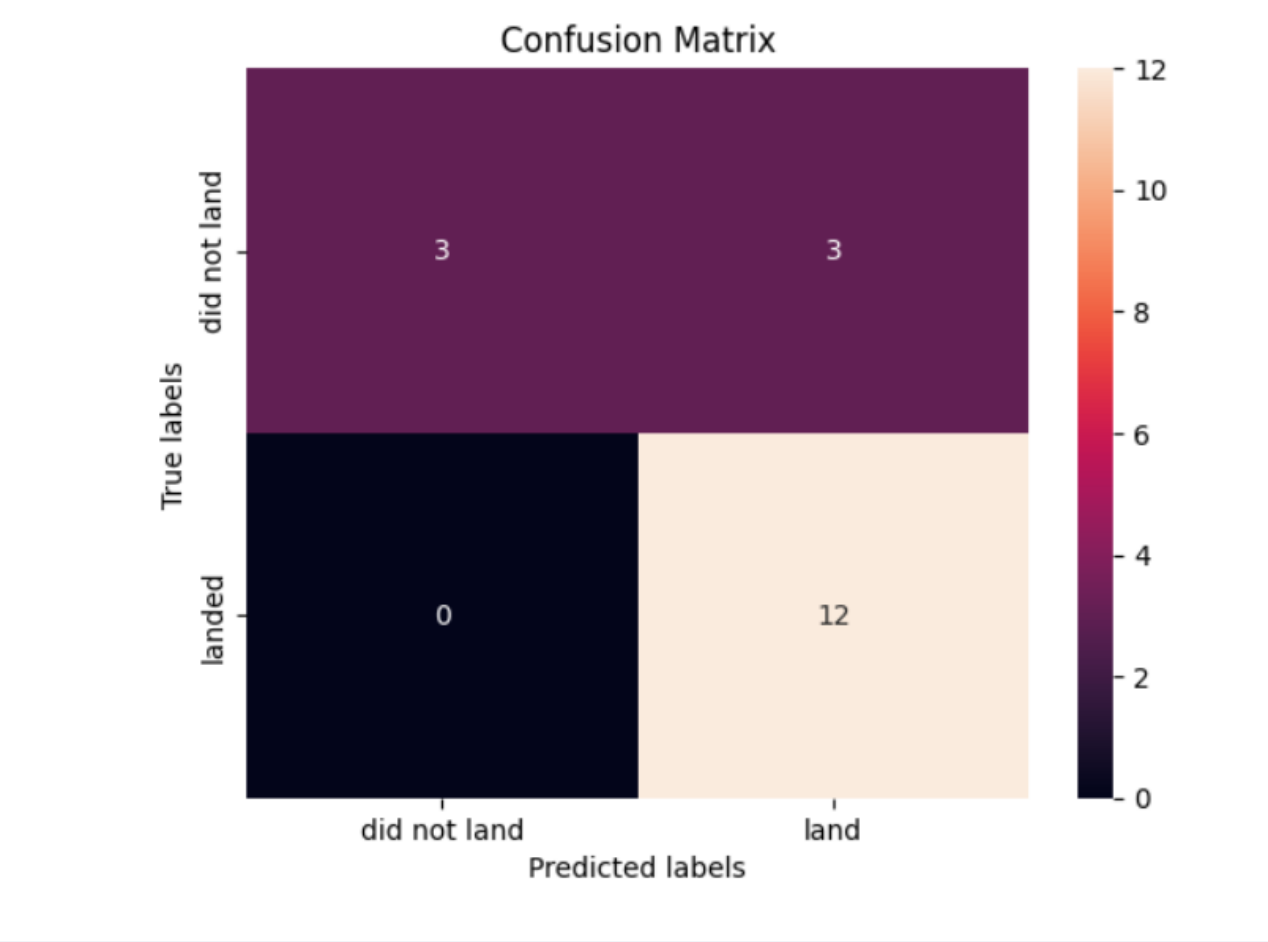
Predictive Analysis (Classification)

Classification Accuracy

- In this project, we trained and evaluated four classification models to predict successful SpaceX Falcon 9 launches:
- **Decision Tree** achieved the highest cross-validation accuracy **(87.1%)**, indicating it performed best on unseen validation data.
- **Logistic Regression, SVM, and KNN** each achieved a test accuracy of **83.33%,** suggesting they generalized well.
- **Decision Tree** had a lower test accuracy **(72.2%),** indicating possible overfitting.
- Best overall performer based on generalization: **Logistic Regression, SVM, and KNN** all tied in test accuracy.
- **Best validation performer: Decision Tree**, but with lower test generalization.



Confusion Matrix For SVM



Conclusions

- We successfully predicted Falcon 9 first stage landing outcomes using machine learning models trained on SpaceX historical launch data.
- The **Support Vector Machine (SVM)** model with a sigmoid kernel achieved the **highest test accuracy (83.33%)**, demonstrating solid predictive performance.
- **Exploratory Data Analysis (EDA)** revealed key patterns:
  - Higher payloads and certain launch sites are associated with greater mission success.
  - Orbits like GEO, SSO, and HEO showed consistent success rates.
- An interactive **Plotly Dash dashboard** enabled detailed mission analysis by payload, orbit, and launch site.
- These insights can support **SpaceX's mission planning and cost-saving strategies** by improving the predictability of successful recoveries.

Appendix

Charts Included

- Model Accuracy Bar Chart – Comparison of Logistic Regression, SVM, KNN, and Decision Tree.
- Confusion Matrix – Visualization of prediction results for the best model.
- EDA Plots:Flight Number vs. Launch SitePayload vs. OrbitOrbit Type Success Rates (Bar chart)Yearly Success Rate (Line chart)

External References

- GitHub Repository with All Notebooks: click here

Thank you!