IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noah George
12/24/2025

Outline

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

Executive Summary

- **Methodology**
- Collected SpaceX launch data via **REST API** and public datasets
- Cleaned and engineered features using **Python (Pandas, NumPy)**
- Performed **Exploratory Data Analysis (EDA)** with SQL and visualizations
- Built and evaluated **machine learning classifiers** (KNN, Logistic Regression, SVM)
- Developed an **interactive dashboard** to explore launch success patterns
- **Key Results**
- Launch success is strongly influenced by **payload mass** and **launch site**
- Certain booster versions show **higher success consistency**
- All three models achieved **similar predictive performance**
- **Logistic Regression** was selected for its **simplicity, interpretability, and robustness**
- Interactive dashboard enables **real-time filtering and insight discovery**

Introduction

- **Background & Context**
- SpaceX has significantly reduced launch costs through **reusable rocket technology**
- Understanding factors that influence **launch success** is critical for reliability and efficiency
- Historical launch data provides an opportunity to **analyze patterns and predict outcomes** using data science

- **Key Questions**
- Which factors most strongly influence **launch success or failure**?
- How do **payload mass**, **launch site**, and **booster version** impact outcomes?
- Can machine learning models **accurately predict launch success**?
- Which model provides the **best balance of performance and interpretability**?

Section 1

Methodology

Methodology

- **Executive Summary**
- **Data Collection Methodology**
- Collected historical SpaceX launch data using a **REST API** and supplemental public datasets
- **Data Wrangling**
- Cleaned, filtered, and transformed raw data
- Handled missing values and engineered relevant features for analysis
- **Exploratory Data Analysis (EDA)**
- Analyzed launch outcomes using **SQL queries** and **Python visualizations**
- Identified relationships between launch success, payload mass, launch site, and booster version
- **Interactive Visual Analytics**
- Built **Folium maps** to analyze launch site locations
- Developed an interactive **Plotly Dash dashboard** for dynamic exploration of launch success patterns
- **Predictive Analysis**
- Trained **classification models** (KNN, Logistic Regression, SVM) to predict launch success
- Tuned hyperparameters and evaluated models using **accuracy and confusion matrices**
- Selected **Logistic Regression** for its interpretability and comparable performance

Data Collection

- **How the Data Was Collected**
- Retrieved historical SpaceX launch data using a **REST API**
- Extracted structured launch attributes (date, site, payload, booster, outcome)
- Supplemented API data with **public datasets** for launch site and geographic context
- Scrapped SpaceX web pages for launch data from tables, parsing HTML content
- Stored raw data in **Pandas DataFrames** for further processing

Data Collection – SpaceX API

- **Data Collection Process (Key Phrases)**
- REST API requests
- JSON response parsing
- Data extraction & normalization
- DataFrame creation
- Initial data validation

- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

```
SpaceX API
    ↓
API Requests (GET)
    ↓
JSON Responses
    ↓
Data Parsing & Extraction
    ↓
Pandas DataFrames
    ↓
Clean Dataset for Analysis
```

Data Collection - Scraping

- **Web Scraping Process (Key Phrases)**
- Targeted public SpaceX launch pages
- HTTP requests to retrieve HTML content
- HTML parsing using **BeautifulSoup**
- Extraction of launch attributes (date, site, booster, outcome)
- Structured data storage in **Pandas DataFrames**
- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

```
Public Web Pages
        ↓
HTTP Request
        ↓
HTML Response
        ↓
Parse HTML (BeautifulSoup)
        ↓
Extract Relevant Fields
        ↓
Structured Dataset (DataFrame)
```

Data Wrangling

- **How the Data Were Processed (Key Phrases)**
- Removed irrelevant and duplicate records
- Handled missing and inconsistent values
- Converted data types and standardized formats
- Encoded categorical variables for analysis
- Created clean, analysis-ready datasets

```
Raw Collected Data
        ↓
Data Cleaning
        ↓
Missing Value Handling
        ↓
Feature Engineering
        ↓
Data Transformation
        ↓
Clean Dataset for EDA & Modeling
```

EDA with Data Visualization

- **Exploratory Data Analysis (EDA)**
- **Charts Used & Purpose**
- **Bar Charts** – compared launch success rates across launch sites and booster versions
- **Scatter Plots** – examined the relationship between payload mass and launch outcome
- **Line / Trend Charts** – analyzed changes in launch success over time
- **Categorical Comparisons** – identified differences in success patterns across key variables

- **Why These Charts Were Used**
- To identify **patterns, trends, and outliers** in launch data
- To compare **categorical variables** against launch outcomes
- To support **feature selection** for predictive modeling

- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/edadataviz.ipynb

EDA with SQL

- **Summary of SQL Queries Performed**
- Queried launch records by **launch site** to compare success and failure counts
- Aggregated launch outcomes using **GROUP BY** and **COUNT** functions
- Filtered data based on **payload mass ranges** and launch conditions
- Analyzed relationships between **booster version**, launch site, and success rate
- Extracted summary statistics to support visualization and modeling

- **Purpose of SQL Analysis**
- Efficiently explore large datasets
- Validate findings from Python-based EDA
- Generate structured summaries for visualization

- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Map Objects Created**
- **Markers** – identified SpaceX launch site locations
- **Circle Markers** – represented launch outcomes and relative success frequency
- **Polylines** – visualized distances between launch sites and key geographic features
- **Pop-up Labels** – displayed launch site details and contextual information
-

- **Why These Objects Were Used**
- To visualize the **geographic distribution** of launch sites
- To assess proximity to **coastlines, cities, and infrastructure**
- To identify spatial patterns related to **launch success**
- To enhance interactivity and intuitive data exploration
- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- **Plots & Interactions Added**
- **Pie Chart** – displayed launch success distribution by launch site
- **Scatter Plot** – visualized payload mass versus launch outcome
- **Dropdown Menu** – allowed users to filter results by launch site
- **Range Slider** – enabled dynamic filtering by payload mass range
- **Color Encoding** – differentiated booster version categories
-

- **Why These Plots & Interactions Were Used**
- To allow **interactive exploration** of launch success factors
- To compare performance across **launch sites and payload sizes**
- To identify **patterns and thresholds** affecting launch outcomes
- To support **user-driven analysis** rather than static results
- https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/dash_interactive.py

Predictive Analysis (Classification)

- **Model Development Summary (Key Phrases)**
- Selected relevant features from EDA results
- Split data into **training and testing sets**
- Trained multiple classification models (KNN, Logistic Regression, SVM)
- Tuned hyperparameters to improve model performance
- Evaluated models using **accuracy and confusion matrices**
- Compared results and selected the **best-performing model**
- **Best Model Selection**
- Models achieved **similar accuracy and confusion matrices**
- **Logistic Regression** was selected for:
  - Comparable predictive performance
  - Simplicity and interpretability
  - Lower risk of overfitting

  - https://github.com/Noahgeorge21/spacex-falcon-9-landing-prediction-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

```
Feature Selection
        ↓
Train / Test Split
        ↓
Model Training (KNN, LR, SVM)
        ↓
Hyperparameter Tuning
        ↓
Model Evaluation
        ↓
Model Comparison
        ↓
Final Model Selection
```

Results

- **Exploratory Data Analysis Results**
- Launch success rates vary significantly by **launch site**
- **Higher payload mass** shows a stronger relationship with launch outcome
- Certain **booster versions** demonstrate more consistent success
- Trends over time indicate **improving launch reliability**

- **Interactive Analytics (Screenshots)**
- Dashboard screenshots demonstrate:
    - Launch site filtering via dropdown
    - Payload mass range adjustment via slider
    - Dynamic updates to success distributions and scatter plots
- Interactive maps show geographic context and spatial relationships

- **Predictive Analysis Results**
- Classification models achieved **comparable accuracy levels**
- Confusion matrices showed similar prediction performance across models
- **Logistic Regression** selected as the final model due to:
    - Strong predictive performance
    - Interpretability
    - Model simplicity and robustness

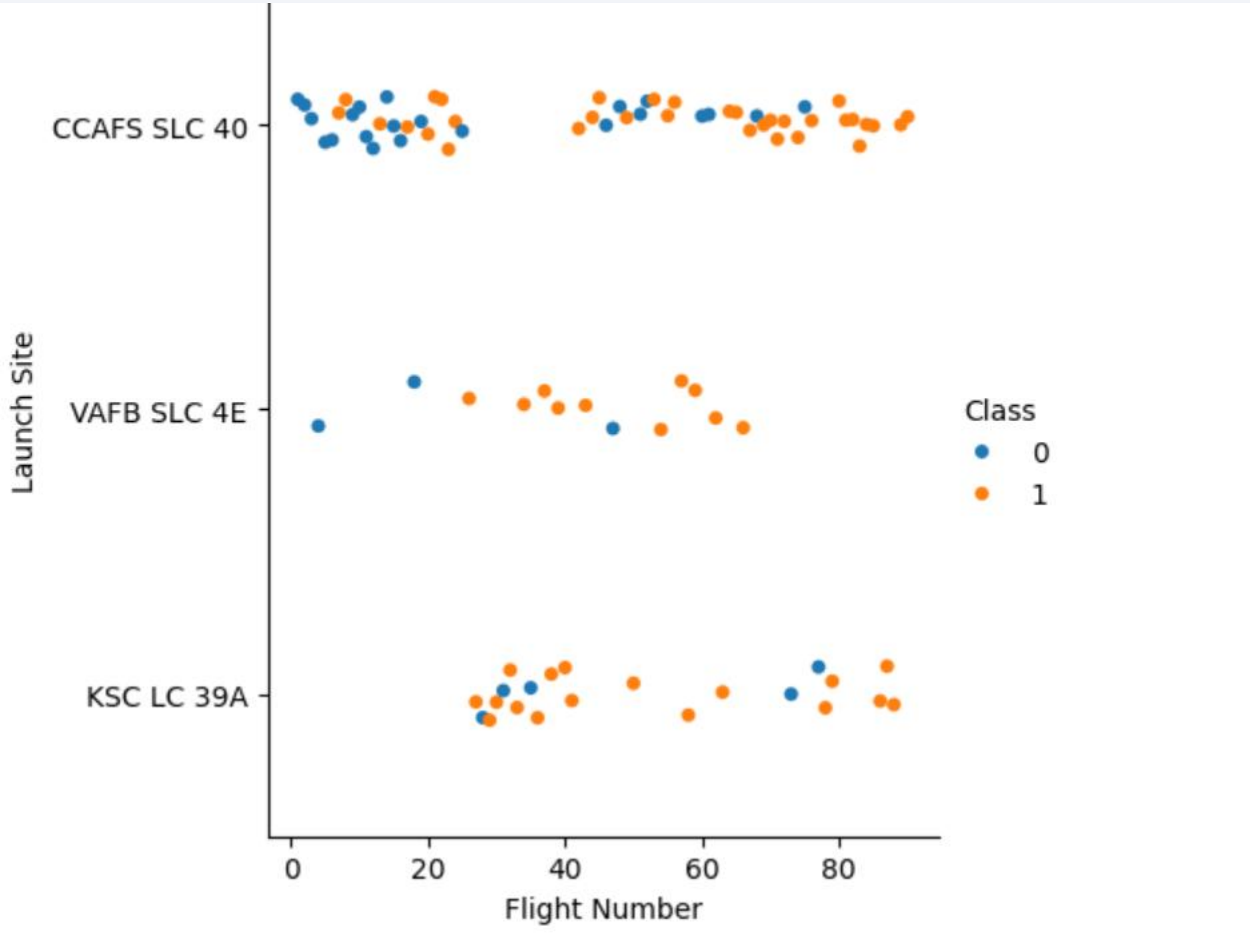
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

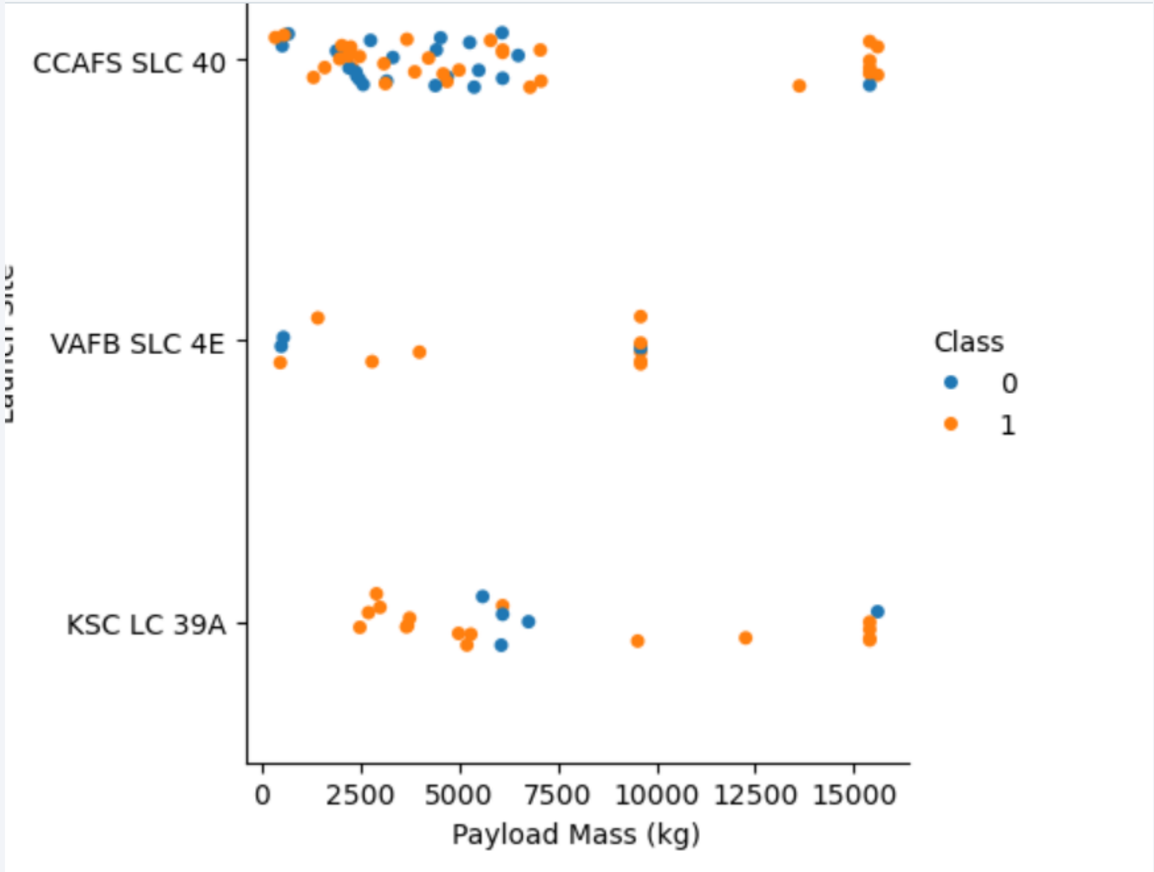
- **What the Plot Shows**
- Each point represents a **SpaceX launch**
- **X-axis:** Flight Number (chronological order of launches)
- **Y-axis:** Launch Site (categorical)
- Points are distributed across multiple launch sites over time
-

- **Key Observations**
- Launch frequency **increases over time**, especially at **CCAFS LC-40** and **KSC LC-39A**
- Newer launch sites appear more frequently in **later flight numbers**
- Indicates **operational expansion and improved launch capacity** over time
- No single launch site dominates early flights, but patterns emerge as SpaceX scales



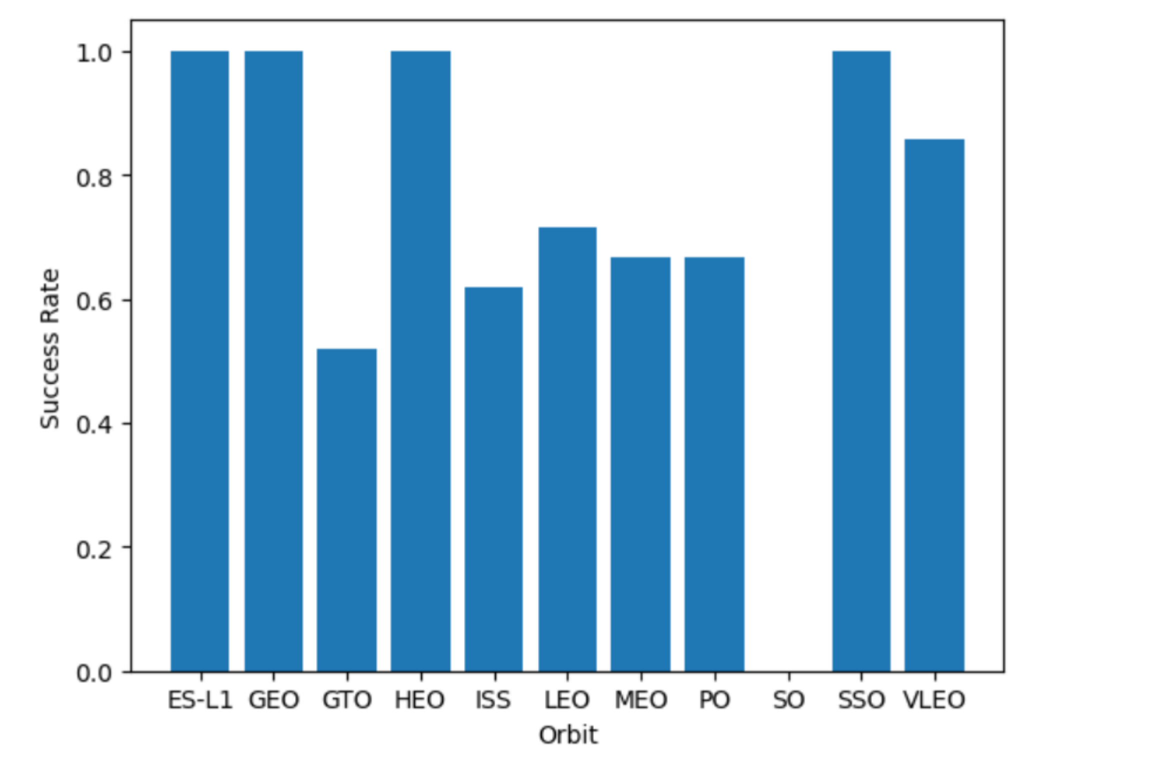
Payload vs. Launch Site

- **What the Plot Shows**
- Each point represents a **SpaceX launch**
- **X-axis:** Payload Mass (kg)
- **Y-axis:** Launch Site (categorical)
- Distribution of payload sizes across different launch sites
-
- **Key Observations**
- **Heavier payloads** are more frequently launched from **KSC LC-39A**
- **CCAFS LC-40** supports a wide range of payload masses
- **VAFB SLC-4E** primarily handles **lower to mid-range payloads**
- Payload distribution suggests **specialization by launch site**



Success Rate vs. Orbit Type

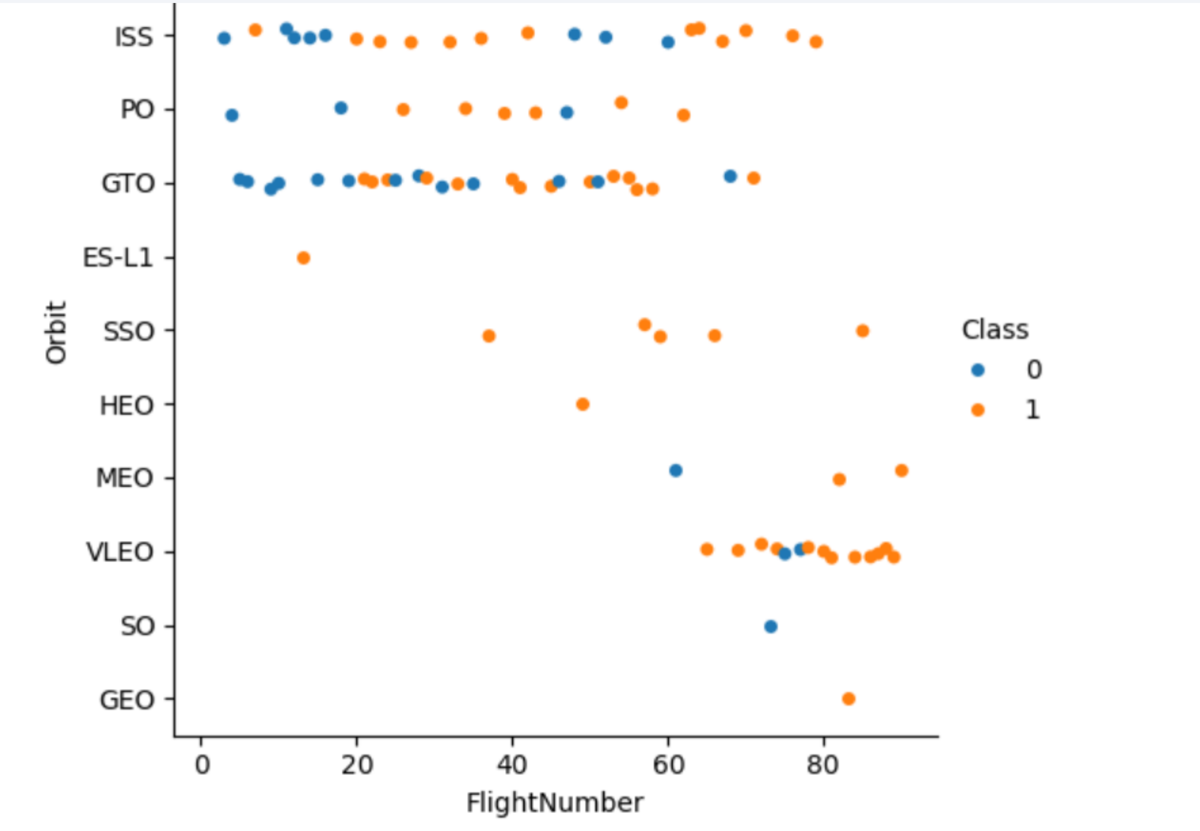
- **What the Chart Shows**
- **X-axis:** Orbit Type (LEO, GTO, ISS, PO, etc.)
- **Y-axis:** Launch Success Rate
- Each bar represents the **proportion of successful launches** for a given orbit
-
- **Key Observations**
- **LEO missions** show the **highest success rates**, reflecting operational maturity
- **GTO missions** have slightly lower success rates due to higher mission complexity
- Less frequent orbits show **greater variability** in success outcomes
- Orbit type is a meaningful factor influencing **launch reliability**



Flight Number vs. Orbit Type

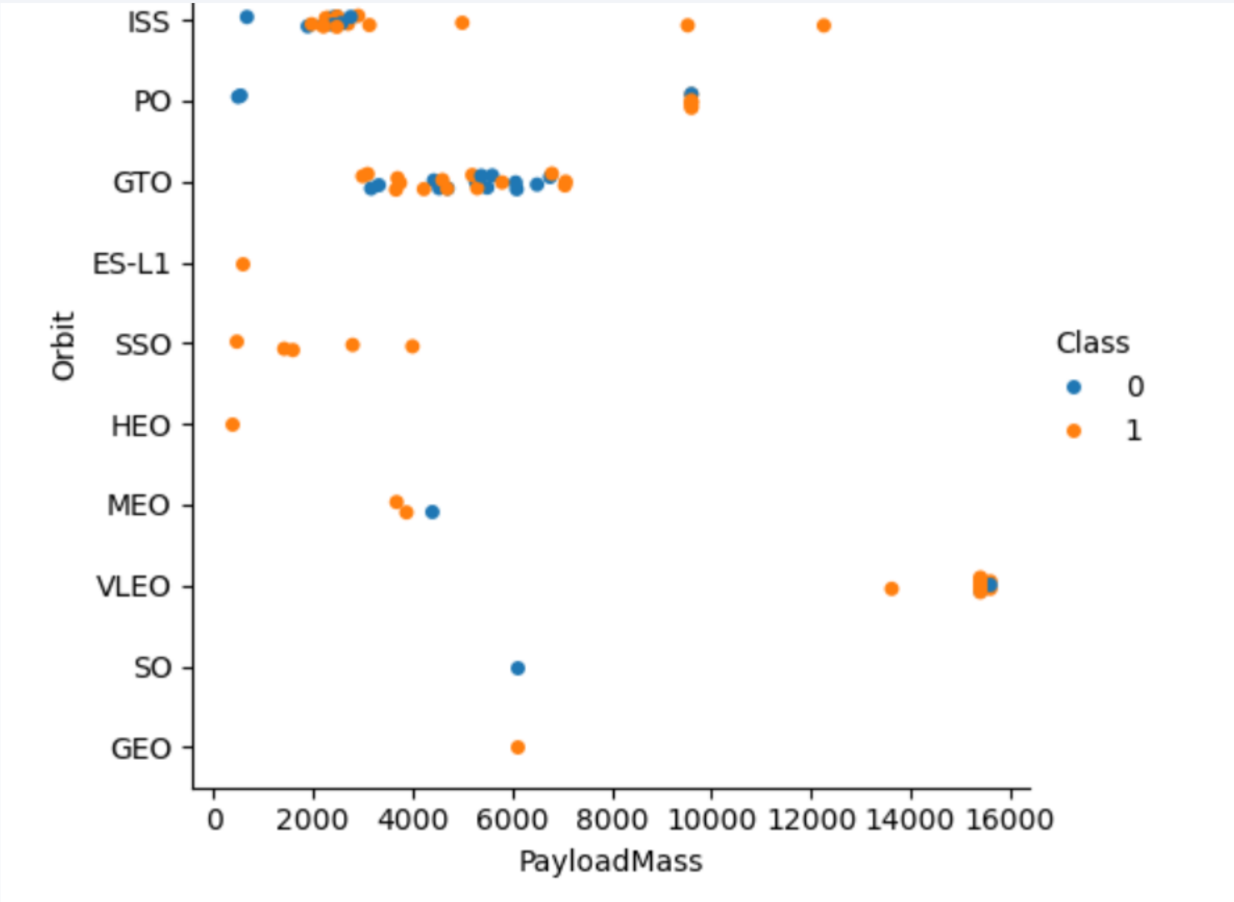
- **What the Plot Shows**
- Each point represents a **SpaceX launch**
- **X-axis:** Flight Number (chronological sequence)
- **Y-axis:** Orbit Type (categorical)
- Distribution of mission orbits over time
-

- **Key Observations**
- Early flights are concentrated in **LEO missions**, indicating initial operational focus
- **GTO and ISS missions** become more frequent as flight numbers increase
- Expansion into more complex orbits reflects **growing launch capability and confidence**
- Orbit diversity increases over time, suggesting **program maturity**



Payload vs. Orbit Type

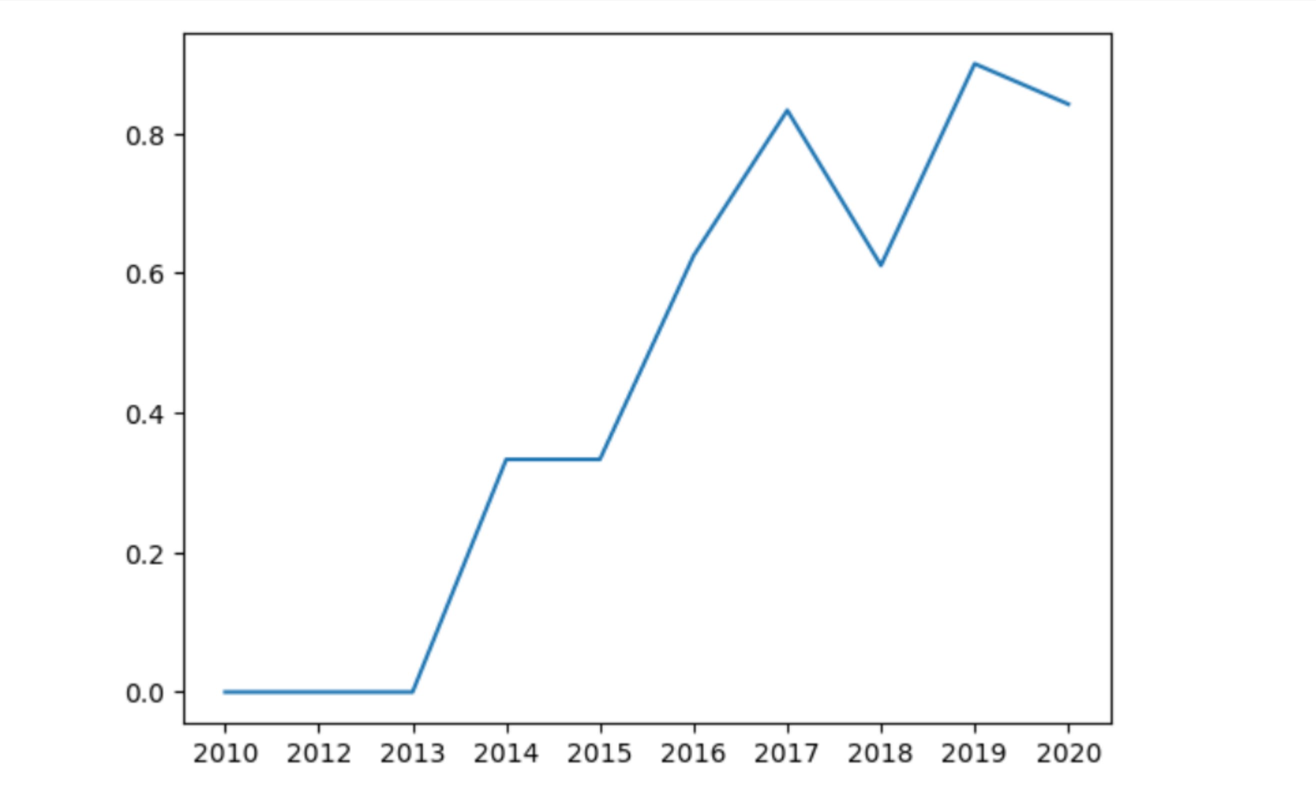
- **What the Plot Shows**
- Each point represents a **SpaceX launch**
- **X-axis:** Payload Mass (kg)
- **Y-axis:** Orbit Type (LEO, GTO, ISS, PO, etc.)
- Distribution of payload sizes across different mission orbits
-
- **Key Observations**
- **LEO missions** span a wide range of payload masses, reflecting operational flexibility
- **GTO missions** generally involve **higher payload masses**, indicating greater mission complexity
- **ISS missions** cluster within a consistent payload range, reflecting standardized mission requirements
- Less frequent orbit types show **more variability** due to limited launch samples



Launch Success Yearly Trend

- **What the Chart Shows**
- **X-axis:** Year
- **Y-axis:** Average Launch Success Rate
- Each point represents the **average success rate for that year**, connected to show trends over time
-

- **Key Observations**
- Launch success rate **improves steadily over time**
- Early years show **greater variability**, reflecting early-stage development
- Later years demonstrate **consistently high success rates**
- Trend indicates increased **operational maturity and reliability**



All Launch Site Names

- **Query Result**
- The dataset contains the following **unique SpaceX launch sites**:
- **CCAFS LC-40**
- **CCAFS SLC-40**
- **KSC LC-39A**
- **VAFB SLC-4E**

- **Short Explanation**
- These launch sites represent **distinct geographic locations** used for different mission profiles
- Each site supports launches with **specific orbital and payload requirements**
- Identifying unique launch sites is essential for **comparative analysis** and **feature selection** in modeling

Launch Site Names Begin with 'CCA'

- **Query Result**
- Retrieved **5 launch records** where the launch site name begins with **“CCA”**
- Results include launches from **Cape Canaveral Air Force Station (CCAFS)** sites

- **Short Explanation**
- This query filters launch records using a **string pattern match** (LIKE 'CCA%')
- It demonstrates how SQL can be used to **search and subset categorical text data**
- Filtering by launch site prefix supports **targeted analysis** of specific launch facilities

Total Payload Mass

- **Query Result**
- Calculated the **total payload mass** carried by boosters on **NASA-sponsored launches**
- Result represents the **sum of payload mass (kg)** across all NASA missions in the dataset

- **Short Explanation**
- This query filters records where the **customer/agency is NASA**
- Uses **aggregation (SUM)** to compute total payload mass
- Helps quantify **NASA’s contribution to overall launch payload volume**
- Demonstrates use of SQL for **conditional aggregation and metric extraction**

Average Payload Mass by F9 v1.1

- **Query Result**
- Calculated the **average payload mass (kg)** for launches using **booster version F9 v1.1**
- Result reflects the **mean payload capacity** delivered by this booster variant

- **Short Explanation**
- This query filters launch records by **booster version = F9 v1.1**
- Uses the **AVG** aggregation function on payload mass
- Helps compare **performance characteristics across booster versions**
- Demonstrates SQL-based **filtering and aggregation**

First Successful Ground Landing Date

- **Query Result**
- Identified the **earliest date** when a **successful landing on a ground pad** occurred
- Result marks the **first confirmed ground landing achievement** in the dataset

- **Short Explanation**
- The query filters records for **successful landing outcomes** on **ground pads**
- Uses **date ordering** to find the earliest occurrence
- Highlights a **key milestone in SpaceX's reusable rocket development**
- Demonstrates SQL use for **conditional filtering and chronological analysis**

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Query Result**
- Retrieved the **names of boosters** that:
  - Successfully **landed on a drone ship**
  - Carried payloads **greater than 4,000 kg and less than 6,000 kg**
- The result lists booster versions that met **both landing success and payload constraints**

- **Short Explanation**
- The query applies **multiple conditions**:
  - Landing outcome = successful
  - Landing type = drone ship
  - Payload mass within the specified range
- Demonstrates SQL use of **conditional filtering (WHERE)** and **range constraints**
- Helps identify boosters capable of **handling mid-range payloads with successful recovery**

Total Number of Successful and Failure Mission Outcomes

- **Query Result**
- Calculated the **total number of successful missions**
- Calculated the **total number of failed missions**
- Results provide an overall view of **launch reliability** in the dataset

- **Short Explanation**
- The query groups missions by **outcome status (success vs. failure)**
- Uses **COUNT** aggregation to summarize mission outcomes
- Establishes a **baseline success rate** for further analysis and modeling
- Confirms class distribution for **classification model evaluation**

Boosters Carried Maximum Payload

- **Query Result**
- Identified the **booster name(s)** associated with the **maximum payload mass** in the dataset
- These booster(s) represent the **highest payload capacity missions** recorded

- **Short Explanation**
- The query first determines the **maximum payload mass** using an aggregation function
- It then filters the dataset to retrieve the **booster(s) that carried this payload**
- Highlights the **upper performance limits** of SpaceX boosters
- Demonstrates SQL usage of **subqueries and conditional filtering**

2015 Launch Records

- **Query Result**
- Listed **failed landing outcomes** that occurred on **drone ships** in **2015**
- Retrieved associated **booster versions** and **launch site names** for each failure
- Results capture early challenges during SpaceX's **booster recovery development phase**

- **Short Explanation**
- The query filters records by:
  - **Year = 2015**
  - **Landing type = drone ship**
  - **Landing outcome = failure**
- Displays relevant operational details (booster version and launch site)
- Helps analyze **early-stage landing failures** and their operational context
- Demonstrates SQL use of **multi-condition filtering and temporal analysis**

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

- **Query Result**
- Ranked **landing outcomes** by their **total occurrence count**
- Included outcomes such as:
  - **Failure (drone ship)**
  - **Success (ground pad)**
  - Other landing result categories
- Results are ordered in **descending frequency** between **2010-06-04 and 2017-03-20**

- **Short Explanation**
- The query filters launches within the specified **date range**
- Groups records by **landing outcome type**
- Uses **COUNT** aggregation and **ORDER BY DESC** to rank outcomes
- Highlights how landing success and failure frequencies evolved during SpaceX's **early reusability period**
- Demonstrates SQL usage for **temporal filtering, aggregation, and ranking**

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- **What the Map Shows**
- A **global map** displaying all SpaceX **launch site locations**
- **Markers** indicate individual launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Each marker represents a **distinct geographic launch facility**
-
- **Important Elements on the Screenshot**
- **Location markers** pinpoint launch sites across different regions
- Launch sites are primarily located along **coastal areas**, supporting safe launch trajectories
- The map provides **geographic context** for later spatial analysis
- Interactive elements (zoom, pan, popups) enhance exploration of launch locations

<Folium Map Screenshot 2>

- **What the Map Shows**
- A geographic map of SpaceX launch sites with **color-coded markers**
- Marker colors represent **launch outcomes**:
  - **Green:** Successful landing
  - **Red:** Failed landing
- Each marker corresponds to a **specific launch event**
-

- **Important Elements on the Screenshot**
- **Color-coded markers** visually distinguish success vs. failure outcomes
- **Pop-up labels** provide launch site and outcome details
- Map interactions (zoom and pan) allow closer inspection of clustered launches
- Multiple markers at the same site highlight **historical performance trends**

<Folium Map Screenshot 3>

- **What the Map Shows**
- A focused map view of a **selected SpaceX launch site**
- Nearby **infrastructure and geographic features**, including:
  - **Coastline**
  - **Highways**
  - **Railways**
- **Distance measurements** displayed between the launch site and nearby features
-

- **Important Elements on the Screenshot**
- **Markers** identify the launch site and surrounding points of interest
- **Lines (polylines)** show the measured distance between the launch site and each feature
- **Distance labels** provide quantitative proximity information
- Interactive zoom enables precise spatial inspection

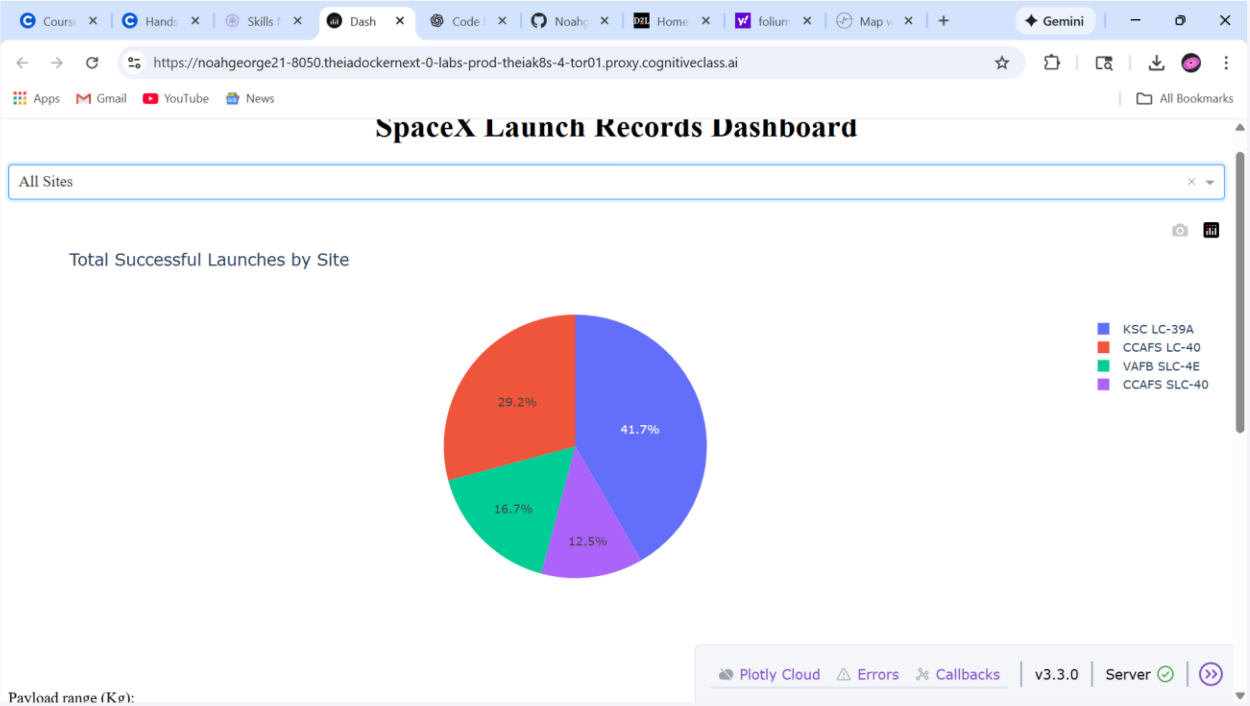
Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- **What the Dashboard Shows**
- A **pie chart** displaying the **total number of successful launches** for **all launch sites**
- Each slice represents a **launch site**, sized by its success count
- Interactive hover reveals **site name and success totals**
-

- **Important Elements on the Screenshot**
- **Pie chart segments** compare launch success across sites
- **Color-coded slices** differentiate each launch site
- **Interactive tooltips** provide exact success counts
- The chart updates dynamically when filters are applied in the dashboard



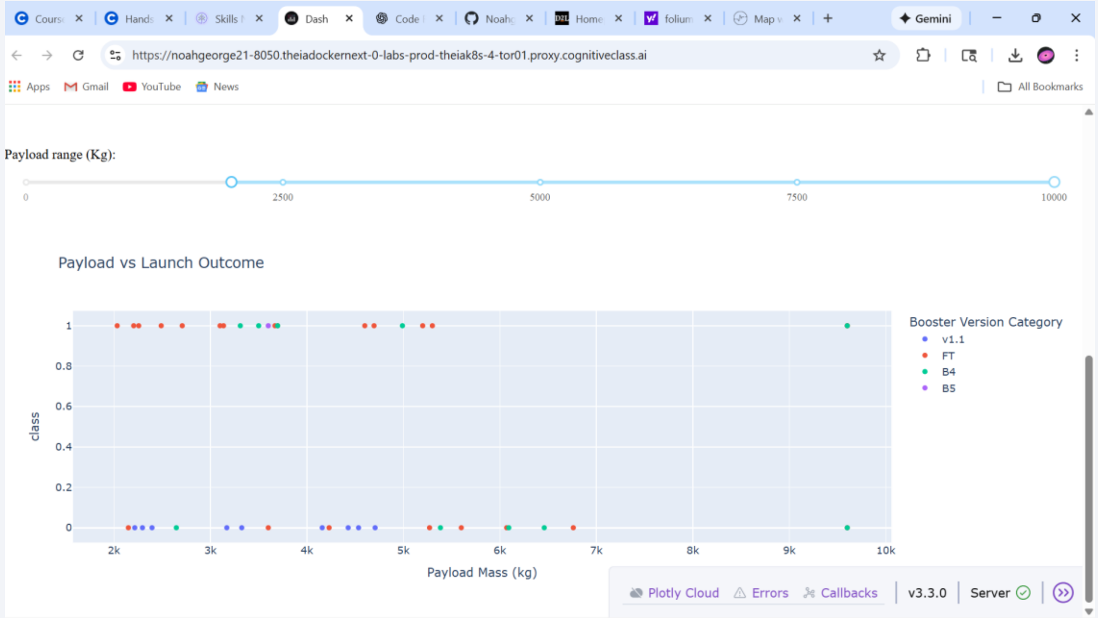
<Dashboard Screenshot 2>

- **What the Dashboard Shows**
- A **pie chart** displaying **success vs. failure outcomes** for the **launch site with the highest success ratio**
- Each slice represents the **proportion of successful and failed launches**
- Chart updates dynamically based on the selected launch site

- **Important Elements on the Screenshot**
- **Binary outcome slices** (Success vs. Failure)
- **Color contrast** clearly distinguishing outcomes
- **Interactive tooltips** showing exact counts or percentages
- Title reflects the **selected top-performing launch site**

<Dashboard Screenshot 3>

- **What the Dashboard Shows**
- An interactive **scatter plot** of **Payload Mass (kg) vs. Launch Outcome**
- **X-axis:** Payload Mass
- **Y-axis:** Launch Outcome (Success = 1, Failure = 0)
- **Color-coded points** represent different **booster versions**
- A **range slider** allows dynamic filtering of payload mass
-



- **Important Elements on the Screenshot**
- **Range slider** demonstrating different payload intervals
- **Filtered scatter points** updating in real time
- **Color legend** identifying booster versions
- Clear separation of success and failure outcomes

Section 5

Predictive Analysis (Classification)

Classification Accuracy

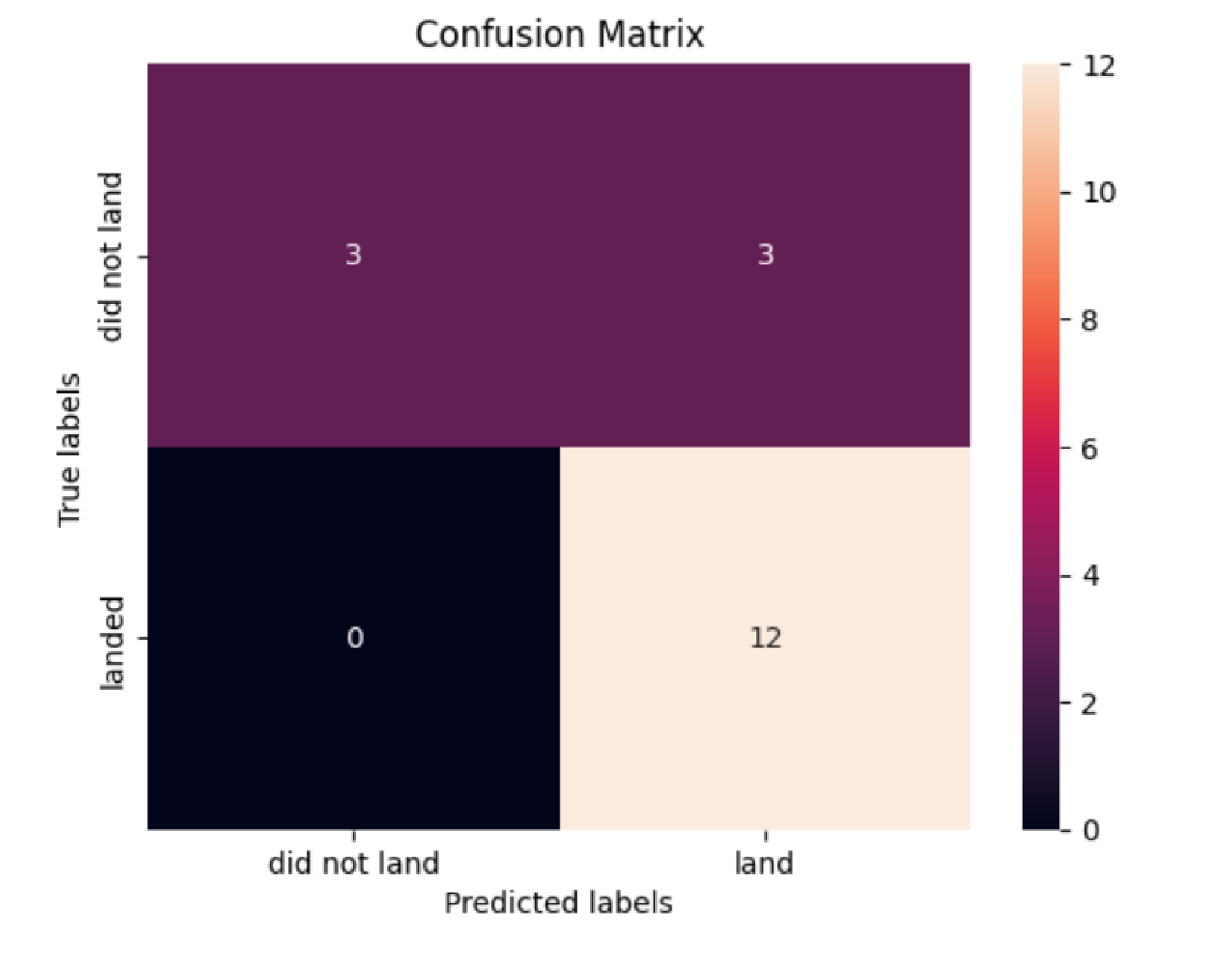
- **What the Chart Shows**
- A **bar chart** comparing the **classification accuracy** of all trained models:
  - **K-Nearest Neighbors (KNN)**
  - **Logistic Regression**
  - **Support Vector Machine (SVM)**
- **Y-axis:** Model accuracy
- **X-axis:** Classification model type
-

- **Important Elements on the Chart**
- Each bar represents the **test accuracy** of a trained model
- Heights allow **direct visual comparison** across models
- Accuracy values are derived from **held-out test data**

Confusion Matrix

- **What the Confusion Matrix Shows**
- A summary of **predicted vs. actual launch outcomes**
- Rows represent **actual outcomes** (Success / Failure)
- Columns represent **model predictions** (Success / Failure)
- Each cell shows the **count of predictions** in that category

- **How to Interpret the Results**
- **True Positives:** Successful launches correctly predicted as successful
- **True Negatives:** Failed launches correctly predicted as failures
- **False Positives:** Failures incorrectly predicted as successes
- **False Negatives:** Successes incorrectly predicted as failures

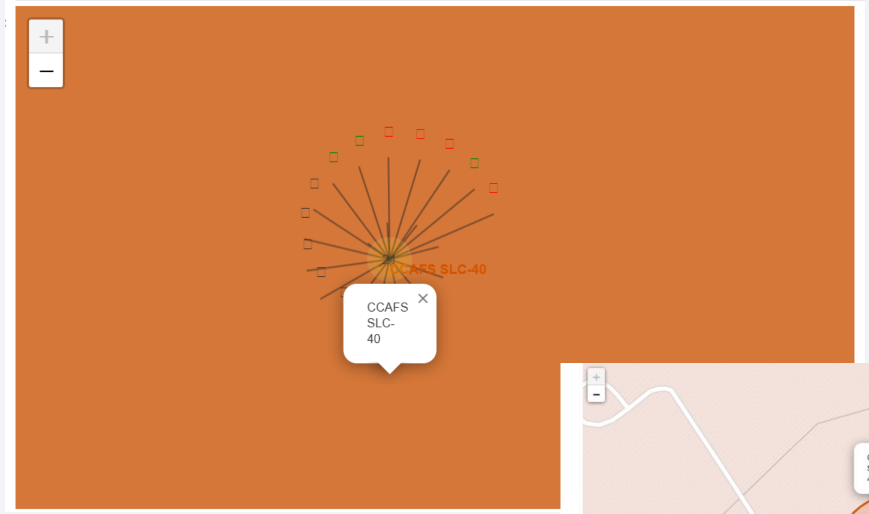


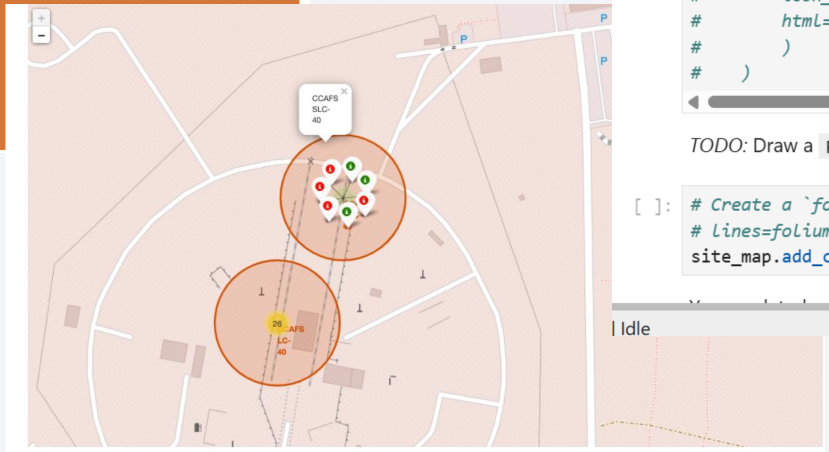
Conclusions

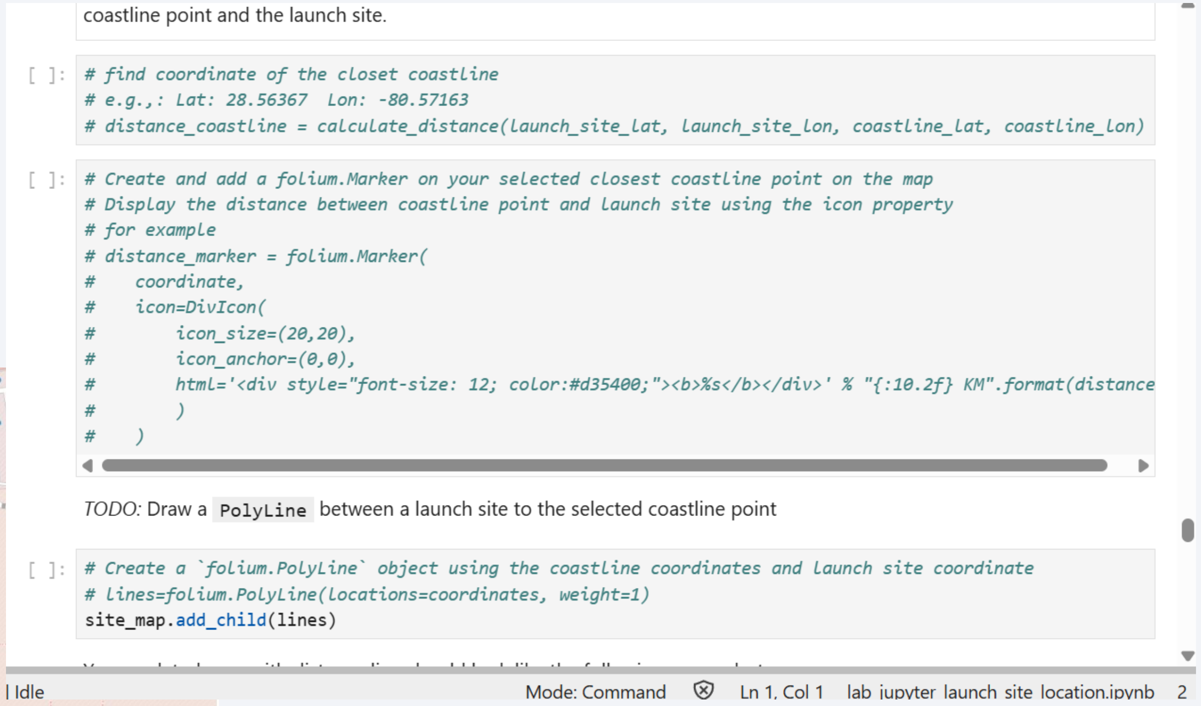
- **Conclusions & Key Takeaways**
- **Launch success is influenced by multiple factors**, including launch site, payload mass, orbit type, and booster version
- **Exploratory and interactive analyses** revealed clear operational patterns and geographic trends
- **Classification models (KNN, Logistic Regression, SVM)** achieved comparable performance in predicting launch success
- **Logistic Regression** was selected as the final model due to its strong accuracy, interpretability, and robustness
- The end-to-end workflow demonstrates how **data collection, analysis, visualization, and modeling** can support data-driven decision making

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







coastline point and the launch site.

```
[ ]: # find coordinate of the closet coastline
     # e.g.,: Lat: 28.56367  Lon: -80.57163
     # distance_coastline = calculate_distance(launch_site_lat, launch_site_lon, coastline_lat, coastline_lon)
```

```
[ ]: # Create and add a folium.Marker on your selected closest coastline point on the map
     # Display the distance between coastline point and launch site using the icon property
     # for example
     # distance_marker = folium.Marker(
     #    coordinate,
     #    icon=DivIcon(
     #        icon_size=(20,20),
     #        icon_anchor=(0,0),
     #        html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>' % "{:10.2f} KM".format(distance
     #        )
     #    )
```

TODO: Draw a `PolyLine` between a launch site to the selected coastline point

```
[ ]: # Create a `folium.PolyLine` object using the coastline coordinates and launch site coordinate
     # lines=folium.PolyLine(locations=coordinates, weight=1)
     site_map.add_child(lines)
```

Idle   Mode: Command   Ln 1, Col 1   lab_jupyter_launch_site_location.ipynb   2

Thank you!