IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Apoorv Srivastava
10th March 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis
  - Interactive Maps with Folium
  - Dashboard with Plotly & Dash
  - Predictive Data analysis
- Summary of all results
  - Results of Exploratory data analysis with Charts
  - Analytics data in screenshots
  - Comparision of Predictive analysis methods

Introduction

- Space Y, which aims to compete with SpaceX, was founded by billionaire industrialist Allon Musk
- Our task is to set the price for each launch.
- We will accomplish this by gathering information about Space X and building dashboards for our team.
- We'll also decide if SpaceX will reuse the first stage by using a machine learning model to forecast whether SpaceX would reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

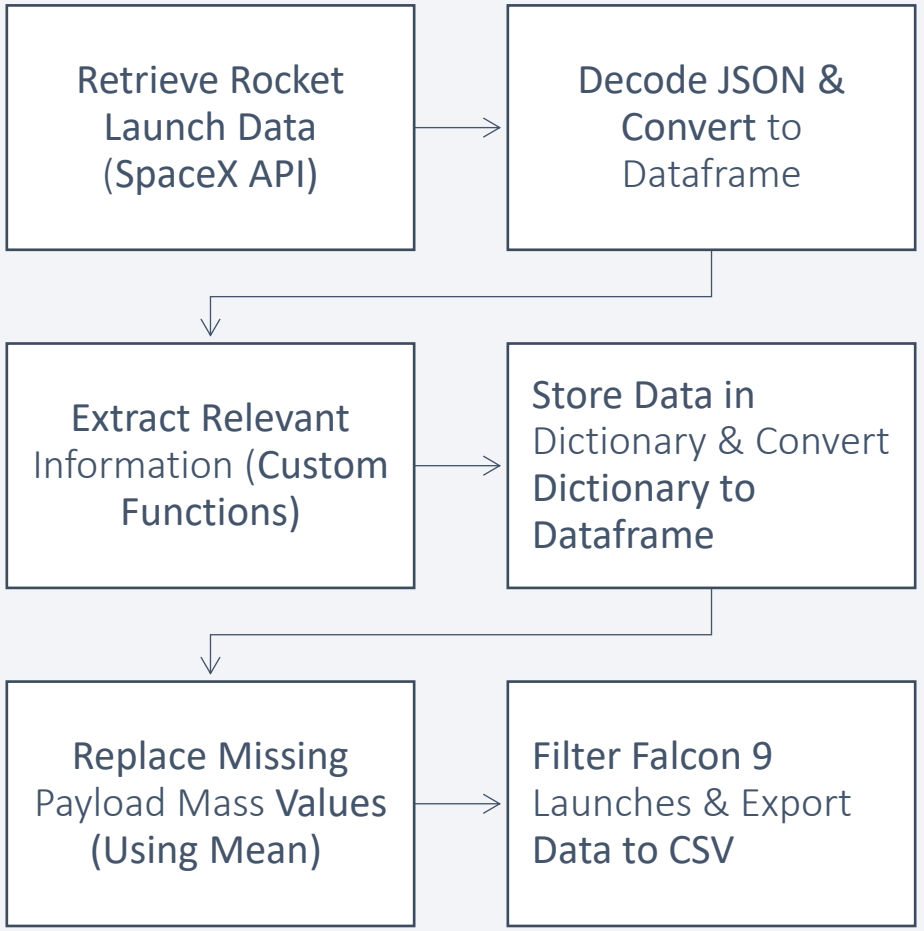
- Data collection methodology:
  - Data were sourced from the SpaceX REST API and Wikipedia launch table
- Perform data wrangling
  - Filtering the Data, and Handling missing Values
  - One hot encoding the labels to prepare it for classification in the next step
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Folium and Plotly Dash were utilized for interactive visual analytics, and model accuracy was assessed using Scikit-learn.

Data Collection

- Data collection involved API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia page.
- Both methods were necessary to gather complete launch information for a comprehensive analysis.
- The SpaceX REST API provided details such as flight number, date, booster version, payload mass, orbit, launch site, and landing outcome.
- Web scraping from Wikipedia extracted additional details like launch site, payload, customer, booster version, and launch outcome.
- Combining these sources ensured a richer dataset for analysis and model development.

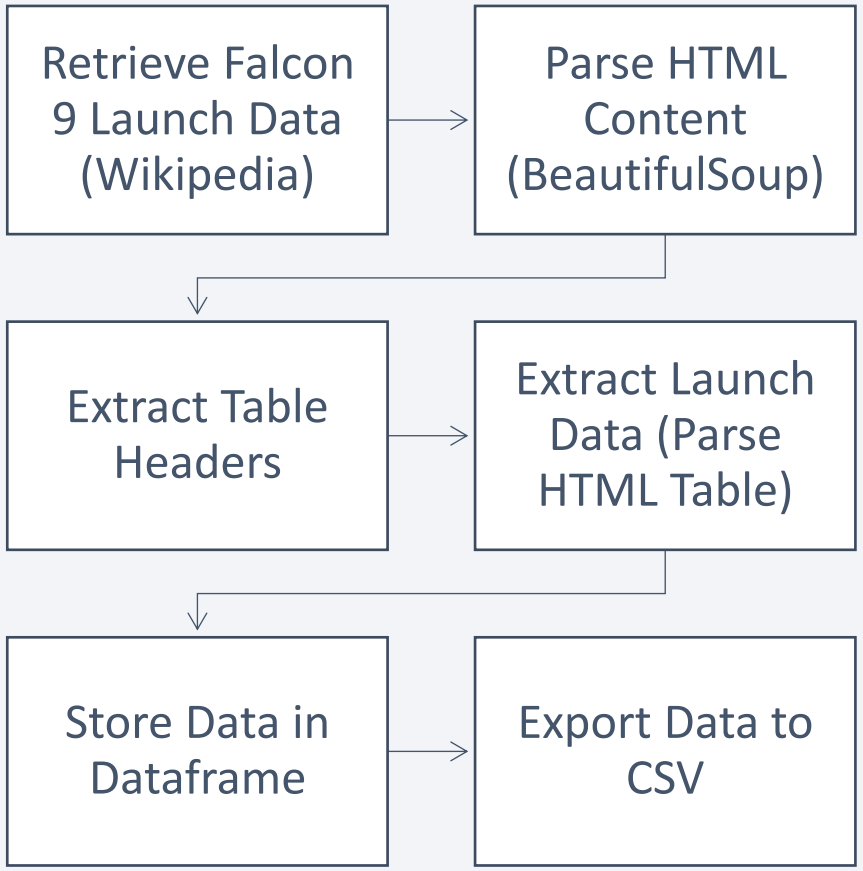
Data Collection – SpaceX API

- Collect launch data from the SpaceX API.
- Process API Response & Extract Relevant Information.
- Organize Data in a structured dataframe.
- Handle Missing Data – Replace missing values in the Payload Mass column with the calculated mean.
- Filter Falcon 9 Launches – Refine the dataset to include only Falcon 9 launch records.
- Export Data – Save the final cleaned dataset as a CSV file.
- Github Link



Data Collection - Scraping

- Retrieve Falcon 9 Launch Data from Wikipedia.
- Parse HTML Content – Create a BeautifulSoup object from the HTML response.
- Extract Table Headers – Identify and extract column names from the HTML table.
- Extract Launch Data – Parse the table to collect relevant launch information.
- Export Data – Save the cleaned data as a CSV file.
- GitHub link



Data Wrangling

- The dataset includes various cases where the booster did not land successfully, with outcomes categorized based on landing attempts.
- “True Ocean” indicates a successful ocean landing, while “False Ocean” signifies an unsuccessful attempt in the ocean.
- “True RTLS” represents a successful landing on a ground pad, whereas “False RTLS” indicates a failed attempt. Similarly, “True ASDS” and “False ASDS” refer to successful and unsuccessful drone ship landings, respectively.
- These outcomes are converted into training labels: 1 for a successful landing and O for an unsuccessful one.
- Exploratory Data Analysis (EDA) is performed to analyze training labels, launch site frequencies, orbit occurrences, and mission outcomes per orbit type.
- A landing outcome label is created from the “Outcome” column, and the cleaned data is exported as a CSV file.
- GitHub link

EDA with Data Visualization

- EDA involves visually exploring and summarizing a dataset to identify distributions, patterns, and relationships between variables.
- Histograms: Used to visualize the distribution of numerical variables like launch success rates and payload mass, helping to identify central tendencies, outliers, and skewness.
- Bar Charts: Compare categorical variables such as launch outcomes across different launch sites or rocket types, making it easier to spot patterns in categorical data.
- Line Charts: Track trends over time, such as Falcon 9's success rate across different years, helping to reveal performance changes over time.

EDA with Data Visualization

- Scatter Plots: Explore relationships between two numerical variables (e.g., payload mass vs. launch success), helping to identify correlations useful for machine learning models.
- Heatmaps & Box Plots: Heatmaps visualize correlations between multiple numerical variables, while box plots display data distributions, highlighting outliers and skewness.
- Key Plots Created: Charts such as Flight Number vs. Payload Mass, Payload Mass vs. Launch Site, and Success Rate Yearly Trend were used to analyze relationships and trends in the dataset.
- GitHub URL

EDA with SQL

- Launch Site Analysis: Retrieved unique launch sites and filtered records where site names start with “CCA”.
- Payload Mass Queries: Calculated total payload mass for NASA (CRS) launches and the average payload mass for the F9 V1.1 booster version.
- Landing Success Analysis: Identified the date of the first successful ground pad landing and boosters with successful drone ship landings carrying payloads between 4000-6000 kg.
- Mission Outcome Queries: Counted total successful and failed missions and identified booster versions that carried the maximum payload mass.

EDA with SQL

- Failure Analysis: Listed failed drone ship landings, their booster versions, and launch sites for 2015.
- Landing Outcome Ranking: Ranked landing outcomes (success/failure) between 2010-2017 in descending order.
- SQL Query Techniques: Used aggregate queries (total launches, success rates), sorting (trends/outliers), joins (linking launch records with rocket details), subqueries (e.g., average payload per site), and filtering (specific outcomes and criteria).
- GitHub URL

Build an Interactive Map with Folium

- Launch Site Markers: Added markers for NASA Johnson Space Center and all SpaceX launch sites with popups and text labels to display their locations and proximity to the equator and coast.
- Outcome-Based Markers: Used color-coded markers (green for success, red for failure) within a Marker Cluster to highlight launch success rates at each site.
- Distance Visualization: Added lines to show distances from the CCAFS LC-40 launch site to nearby features such as railways, highways, coastlines, and the closest city.
- Markers for Spatial Reference: Placed markers to pinpoint exact launch locations and provide a clear visual representation of SpaceX's launch sites.

Build an Interactive Map with Folium

- Circles for Proximity Zones: Added circles around launch sites to represent safety perimeters and operational impact areas.
- Lines for Spatial Context: Connected launch sites to nearby infrastructure, enhancing understanding of geographic relationships and dependencies.
- GitHub URL

Build a Dashboard with Plotly Dash

- Launch Site Selection & Success Visualization: Added a dropdown which updates a pie chart showing total successful launches for all sites or success vs. failure rates for a selected site.
- Payload Mass Interaction: Included a range slider to adjust and filter payload mass, allowing users to explore how payload mass influences launch success.
- Success-Payload Relationship: Integrated a scatter plot to visualize the correlation between payload mass and launch success across different booster versions.
- Interactive Exploration: Enabled dynamic filtering and analysis, helping users examine launch performance based on site selection, payload mass, and success rates.
- GitHub URL

Predictive Analysis (Classification)

- Data Preprocessing: Standardized features to ensure equal contribution, then split the dataset into training and test sets for validation.
- Model Selection & Hyperparameter Tuning: Tested classification models (SVM, Decision Trees, KNN) and used GridSearchCV to optimize hyperparameters for better performance.
- Model Evaluation & Refinement: Assessed models using accuracy, precision, recall, and F1-score, iteratively adjusting them based on validation results.
- Final Model Selection: Chose the best-performing model based on test accuracy while ensuring it generalizes well without overfitting.
- GitHub URL

Predictive Analysis (Classification)

Data Preprocessing

Standardize features

- Split data into training and test sets

Model Selection

Test classification models (SVM, Decision Trees, KNN)

Hyperparameter Tuning

Use GridSearchCV to optimize parameters (C for SVM, max_depth for Decision Trees, n_neighbors for KNN)

Model Evaluation

Assess performance using accuracy, precision, recall, and F1-score

Improvement Iterations

Adjust models based on validation results

Fine-tune hyperparameters for better accuracy

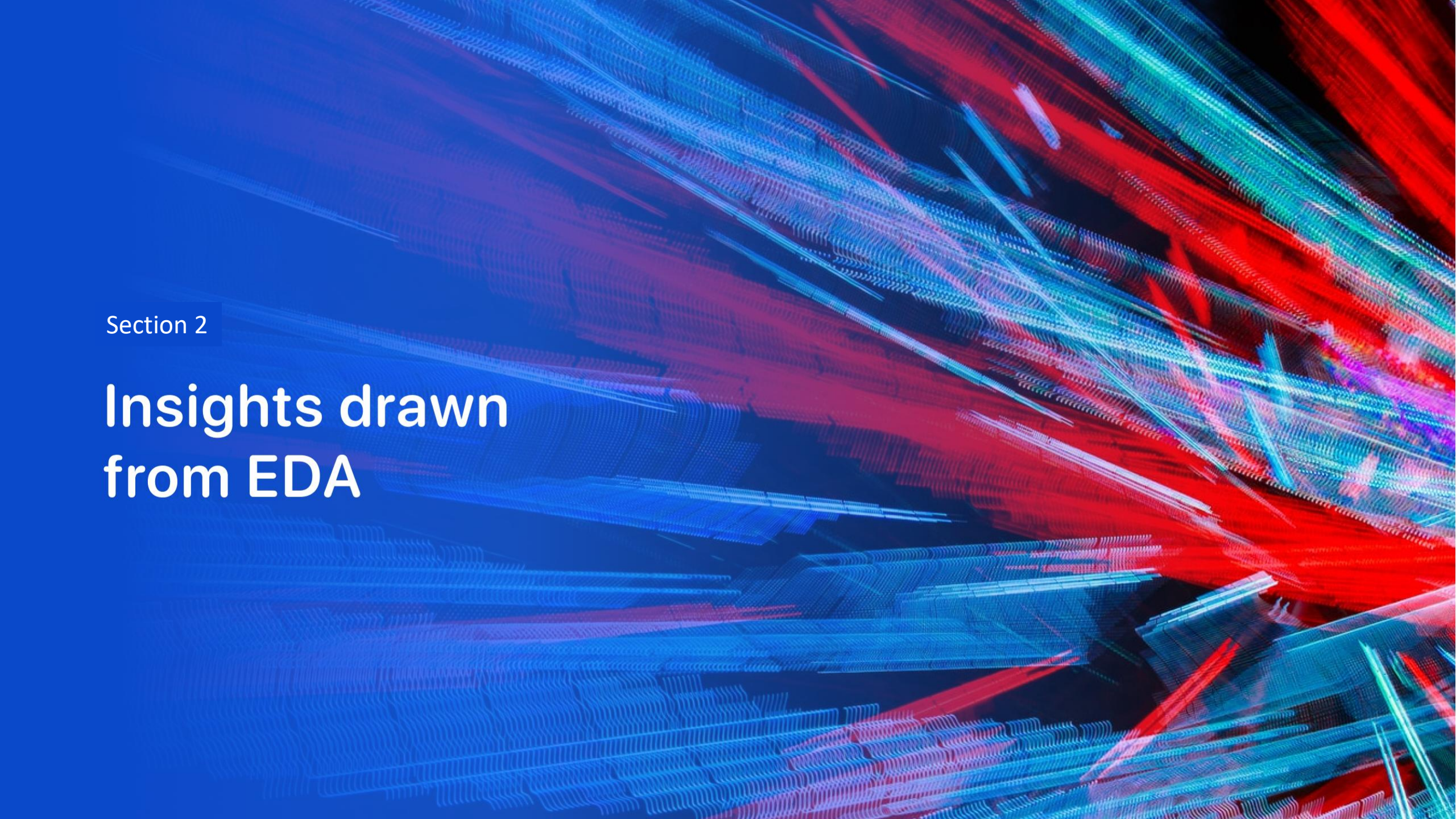
Best Model Selection

Choose the model with the highest test accuracy

Ensure generalizability to prevent overfitting

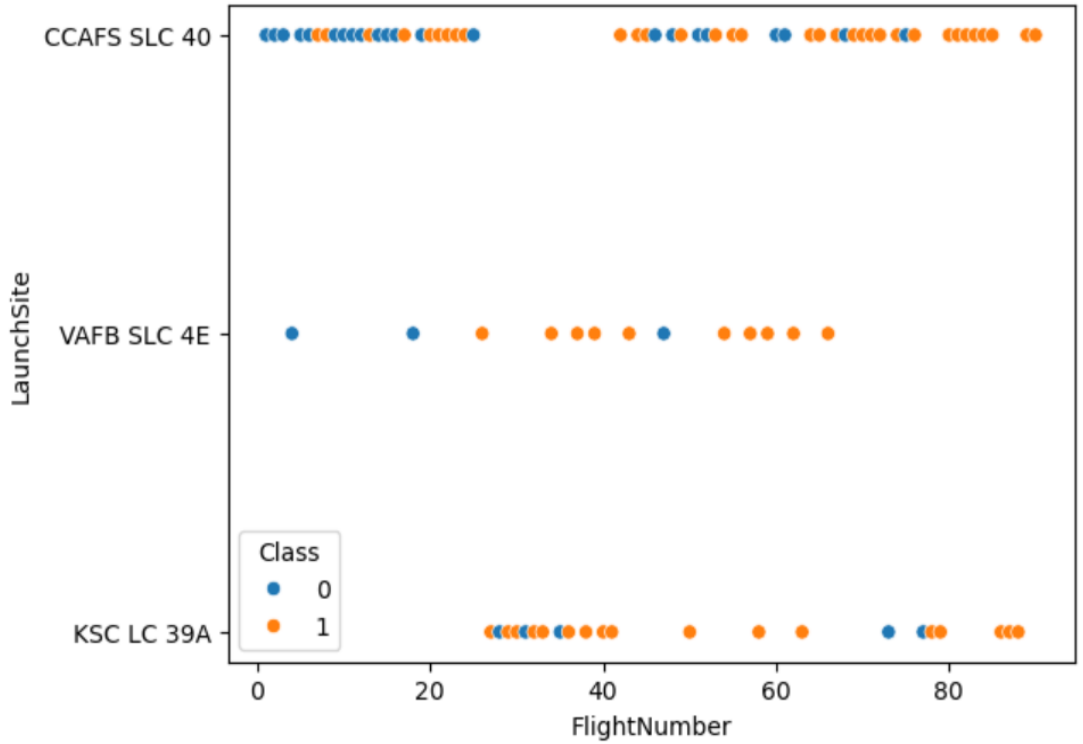
Results

- The Results are Split into multiple sections:
  - Insights drawn from EDA:
    - Exploratory data analysis with seaborn
    - Exploratory data analysis with SQL
  - Launch Sites and Proximity Analysis: Maps with markers for sites
  - Build a Dashboard with Plotly Dash: Dashboard to view successful launches
  - Predective Analysis for successful 1st stage performance

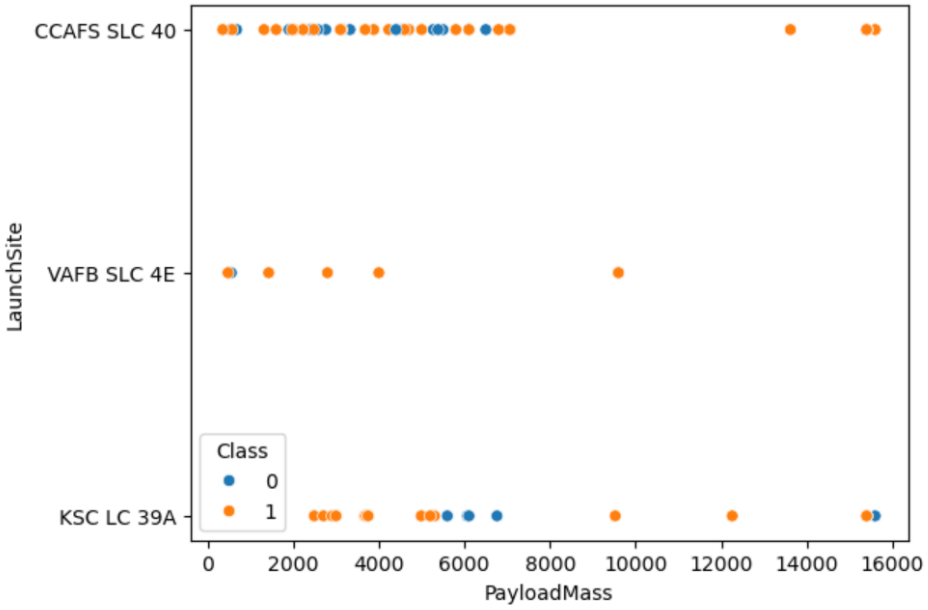

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Launch Success Trends: Early flights failed while recent flights succeeded, indicating improved success rates over time. VAFB SLC 4E and KSC LC 39A have higher success rates, while CCAFS SLC 40 accounts for nearly half of all launches.
- Landing Outcomes & Activity Patterns: Mixed success at major launch sites suggests other influencing factors beyond location. Flight numbers are evenly distributed, showing consistent launch activity over time.



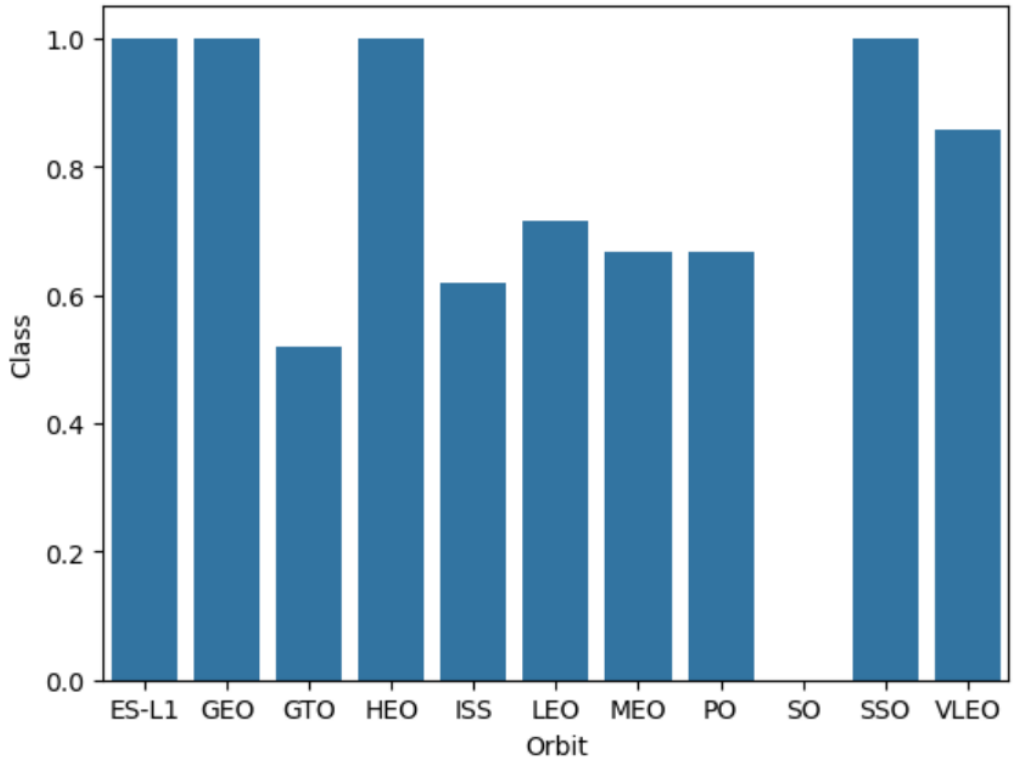
Payload vs. Launch Site



- Payload Mass & Success Rate: Higher payload mass is linked to higher success rates, with most launches over 7,000 kg succeeding. KSC LC 39A has a 100% success rate for payloads under 5,500 kg.
- Launch Site Payload Distribution: CCAFS SLC 40 primarily handles payloads under 10,000 kg, while VAFB SLC 4E and KSC LC 39A support a wider range, with KSC LC 39A frequently launching payloads over 15,000 kg, indicating its role in high-capacity missions.

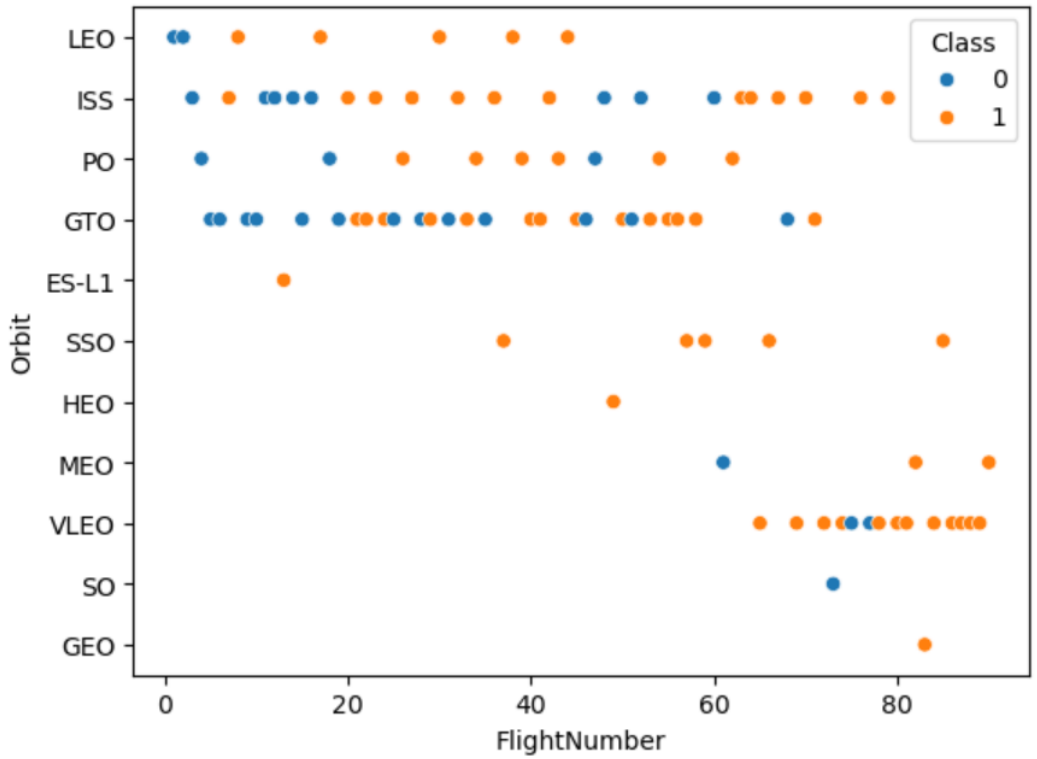
Success Rate vs. Orbit Type

- Orbit Success Rates: ES-L1, GEO, HEO, and SSO have a 100% success rate, while ISO has 0% success. GTO, ISS, LEO, MEO, and PO have success rates between 50% and 85%.
- Challenges in GTO Missions: GTO missions have a lower success rate, indicating additional complexities or challenges in achieving successful landings.



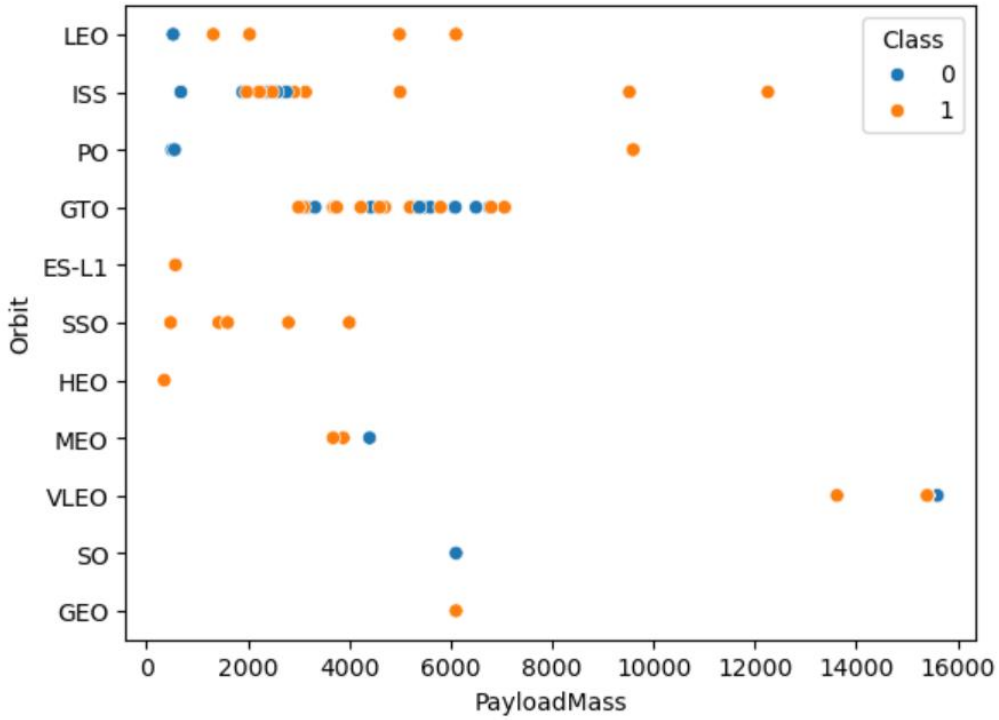
Flight Number vs. Orbit Type

- Flight Experience & Success: Falcon 9's success rate improves with more flights, highlighting the impact of experience and iterative improvements.
- Orbit-Specific Trends: In LEO, success correlates with flight numbers, while in GTO, no clear pattern exists. Recent missions to GTO and ISS show higher success rates, indicating advancements in mission execution.



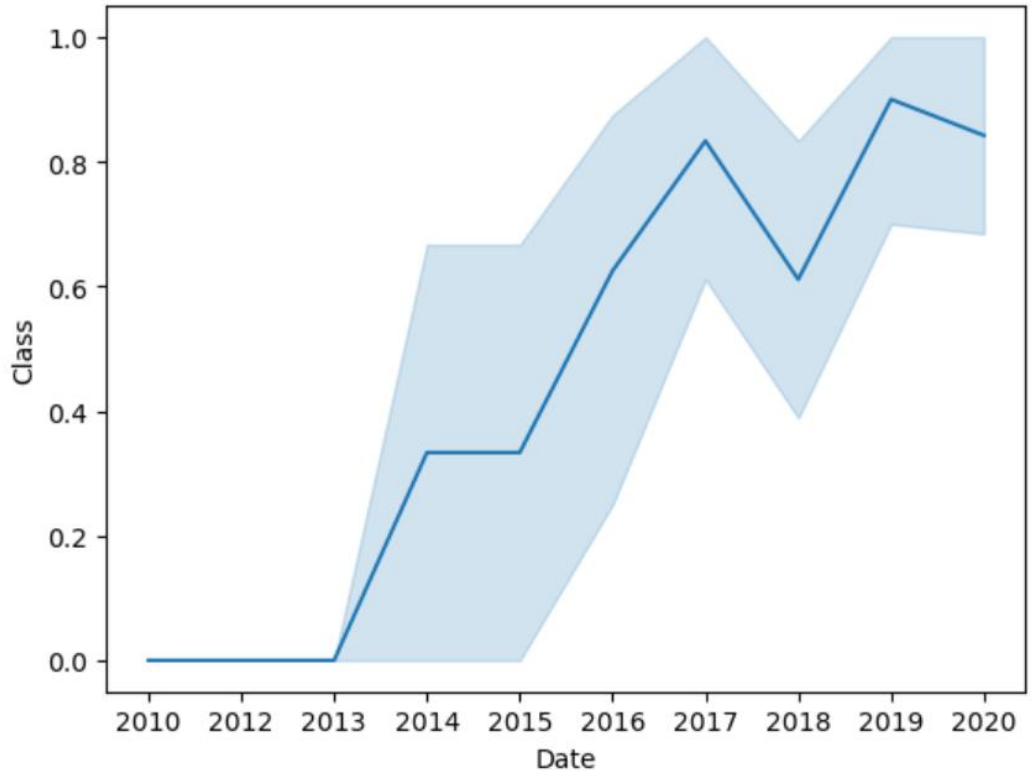
Payload vs. Orbit Type

- Payload Impact on Orbit Success: Heavy payloads negatively affect GTO landings but positively influence GTO and Polar LEO (ISS) orbits.
- Success & Payload Mass: Landings are more successful for payloads under 6,000 kg, while payloads over 10,000 kg show mixed success, highlighting increased challenges with heavier loads.



Launch Success Yearly Trend

- Increasing Success Over Time: Falcon 9's success rate has steadily improved since 2013, surpassing 80% by 2020.
- Overall Trend & Setback: Despite a dip in 2018, the long-term trajectory shows increasing reliability in launches.



All Launch Site Names

- Identifying Launch Sites: Listed all unique launch sites used in the space mission.

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

Launch Site Names Begin with 'CCA'

- Filtering Launch Sites: Retrieved five records where the launch site name starts with ‘CCA’.

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total Payload by NASA (CRS): Calculated the total payload mass carried by boosters for NASA (CRS) missions.

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average Payload for F9 v1.1:
  Computed the average payload mass carried by the F9 v1.1 booster version.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First Ground Pad Success: Identified the date of the first successful landing on a ground pad.

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed boosters that successfully landed on a drone ship with payloads between 4,000 and 6,000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- Mission Outcome Summary: Counted the total number of successful and failed mission outcomes.

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

Boosters Carried Maximum Payload

- Boosters with Maximum Payload: Identified booster versions that carried the highest payload mass.

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

- Failed Drone Ship Landings (2015): Listed unsuccessful drone ship landings, including booster versions and launch sites, for 2015.

month_names	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

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

- Landing Outcome Rankings: Ranked the number of landing outcomes (e.g., Failure on drone ship, Success on ground pad) from June 4, 2010, to March 20, 2017, in descending order.

RANK() OVER (ORDER BY L_COUNT DESC)	LANDING_OUTCOME
1	No attempt
2	Success (drone ship)
2	Failure (drone ship)
4	Success (ground pad)
4	Controlled (ocean)
6	Uncontrolled (ocean)
6	Failure (parachute)
8	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

Global Map of Launch Site Locations

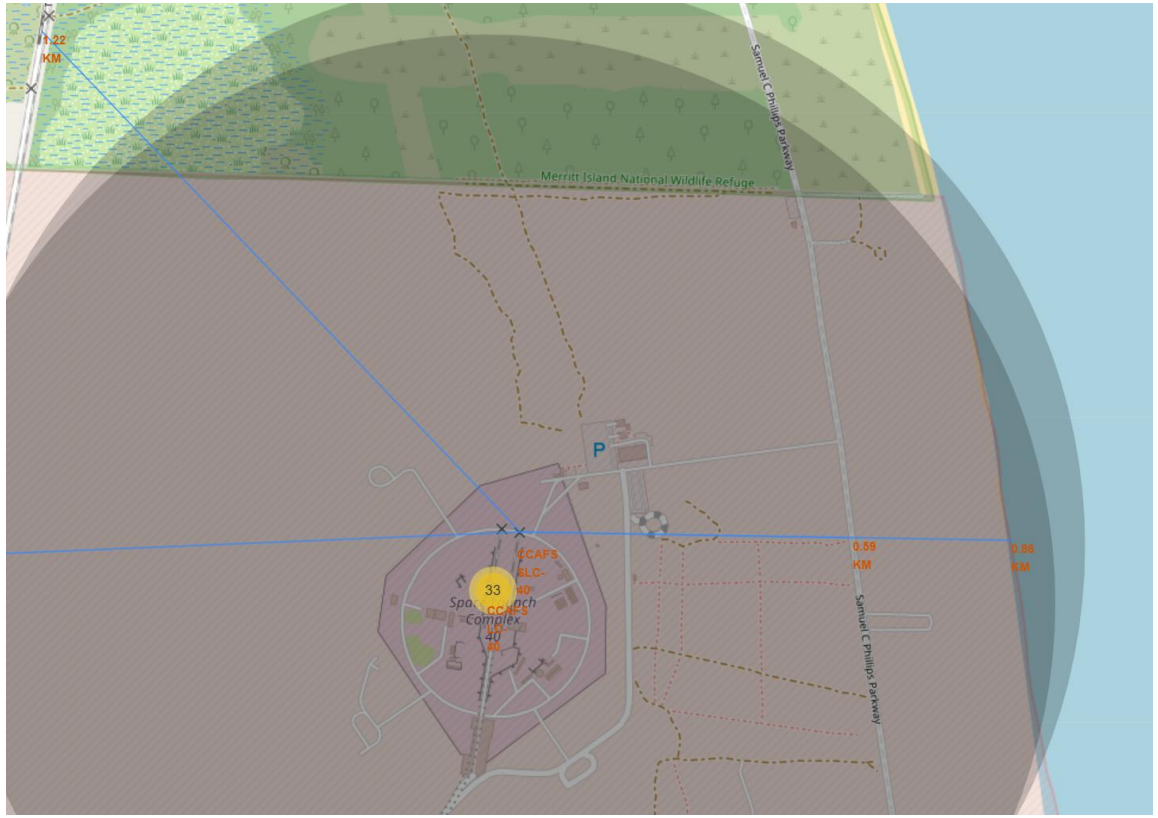
- Equator Advantage: Many launch sites are near the equator to take advantage of Earth's rotational speed (1670 km/h), which helps rockets achieve orbital velocity more efficiently due to inertia.
- Coastal Locations for Safety: All launch sites are near the coast to minimize the risk of debris falling on populated areas by launching rockets over the ocean.
- Not All Sites Near the Equator: While Florida's launch sites are relatively close, Vandenberg Air Force Base (VAFB SLC-4E) is farther from the equator at a latitude of 34.63.
- All Sites Near Water: Every major launch site, including those in Florida and California, is located near a coastline to ensure safer launches and landings.

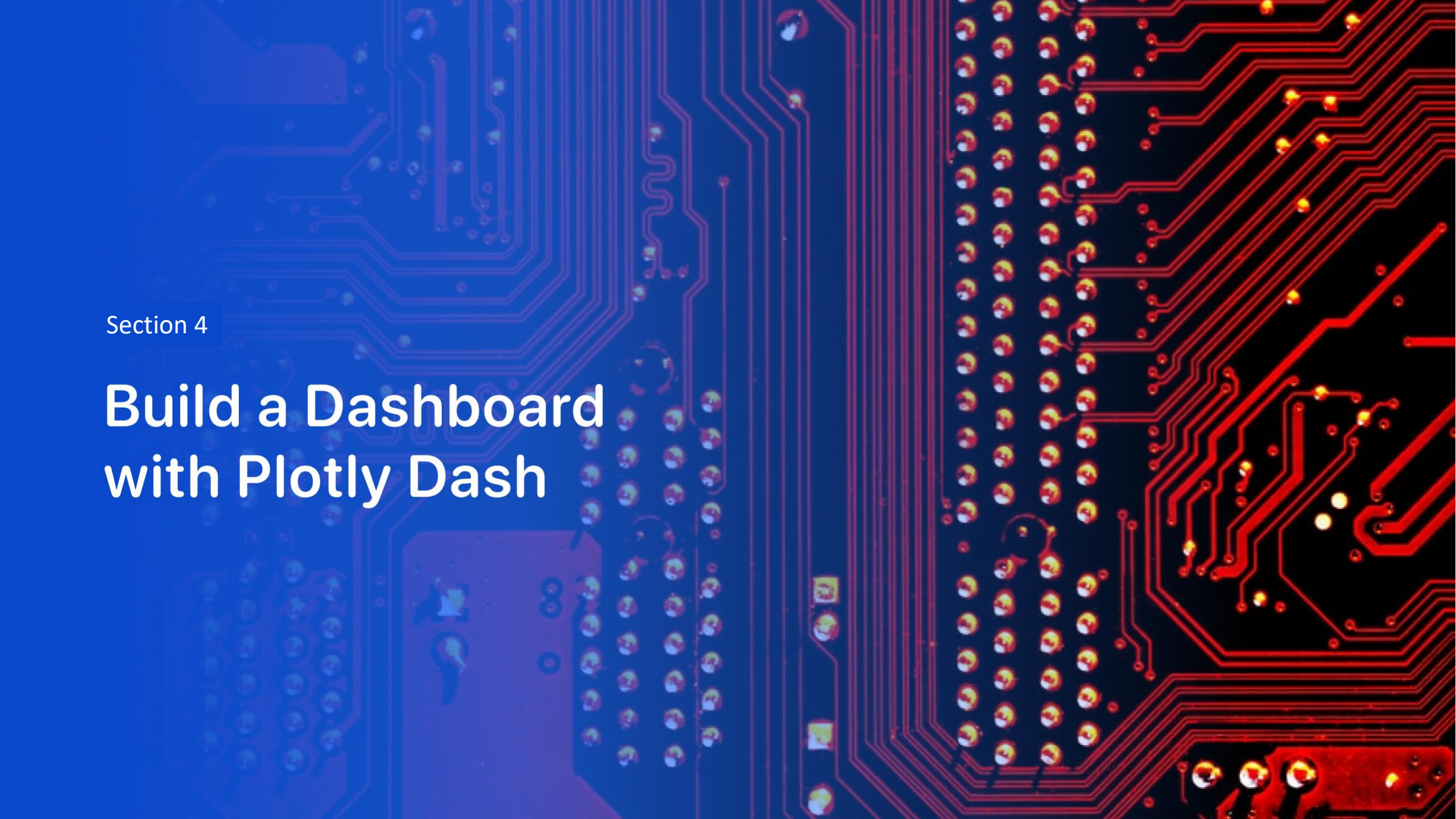
Color-Coded Launch Records

- Green markers represent successful launches, while red markers indicate failures.
- KSC LC-39A has a notably high success rate compared to other sites.
- Clustered markers improve visualization, making it easier to identify trends and compare launch site performance.
- Example: At CCAFS LC-40, 19 out of 26 launches were unsuccessful, highlighting a lower success rate.

Proximity and Safety Considerations for Launch Sites

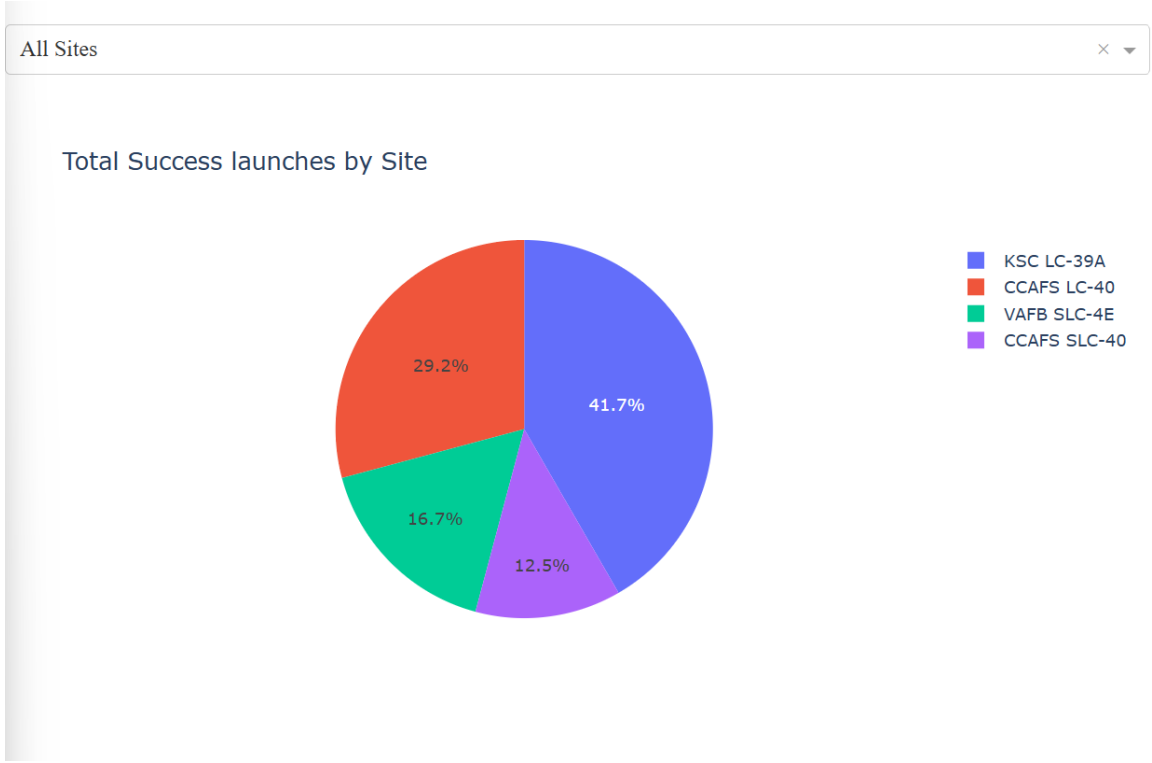
- CCAFS LC-40 is relatively close to key infrastructures: railway (1.22 km), highway (0.59 km), coastline (0.86 km), and Titusville (77.32 km).
- A failed rocket traveling at high speeds could pose risks to populated areas within 15-20 km.
- CCAFS LC-40 is just 0.51 km from the coastline, ensuring safe over-water flight paths and minimizing risks to land-based populations.
- The visual representation with PolyLines highlights the importance of launch site placement for safety and recovery operations.




Section 4
Build a Dashboard with Plotly Dash

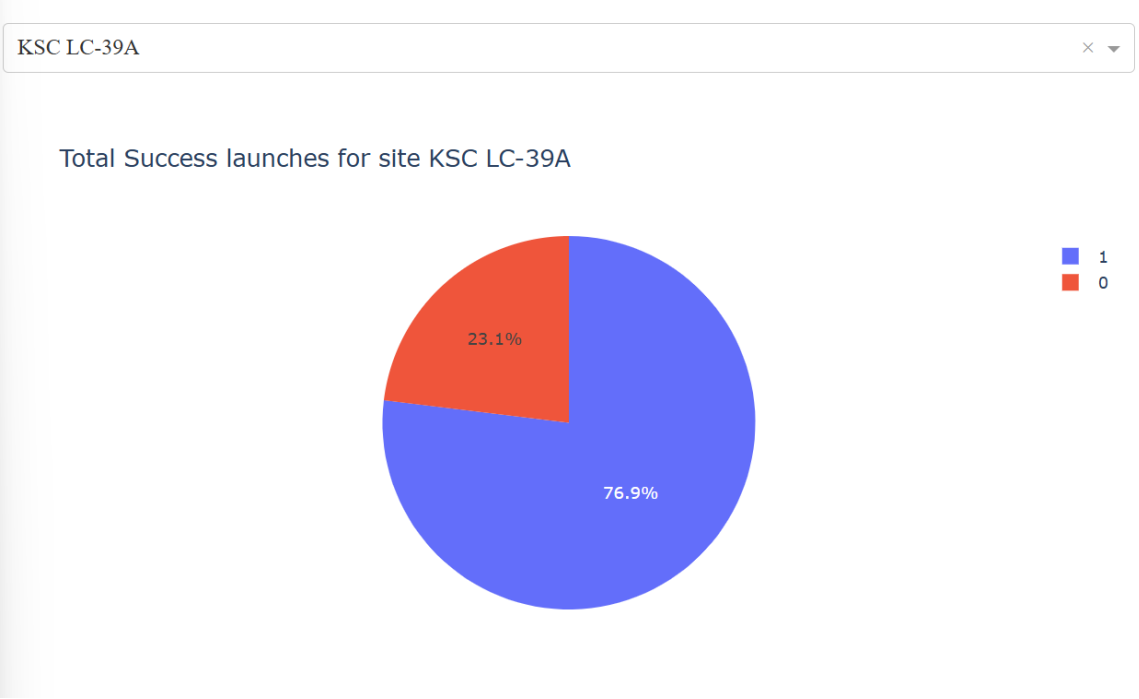
Successful launches for all sites

- KSC LC-39A: The Most Successful Launch Site it accounts for 41.7% of all successful launches, making it the most reliable launch site.
- Other success rates:
  - CCAFS LC-40: 29.2%
  - CCAFS SLC-40: 12.5%
  - VAFB SLC-4E: 16.7%



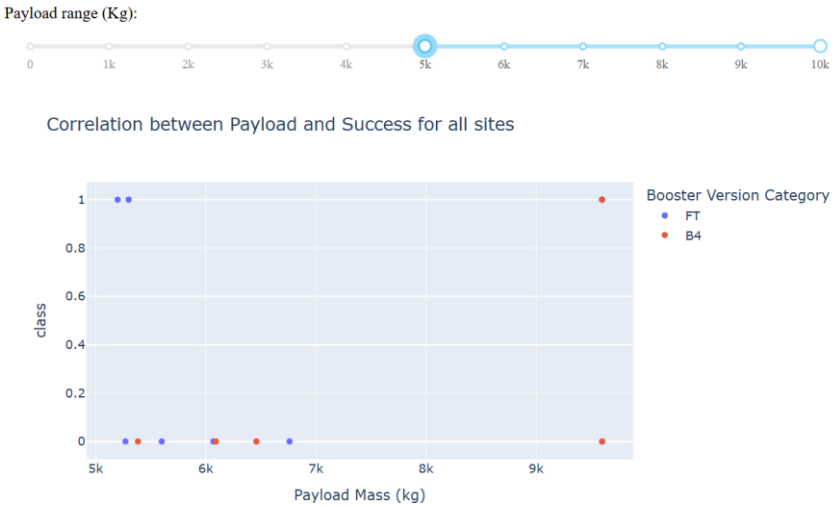
Best Launch Site performance

- KSC LC-39A: Highest Launch Success Rate
- Success Rate: 76.9% (10 successful, 3 failed landings).
- Failure Rate: 23.1%.
- The data highlights KSC LC-39A's strong performance in SpaceX's launch history.



Comparing Payload Masses

- Payload and Booster Version Performance
- Payloads between 2000 and 5500 kg have the highest success rate.
- Booster version “FT” is the most frequently used and performs well across different payload masses.
- Booster version “v1.0” has fewer launches, requiring further analysis.

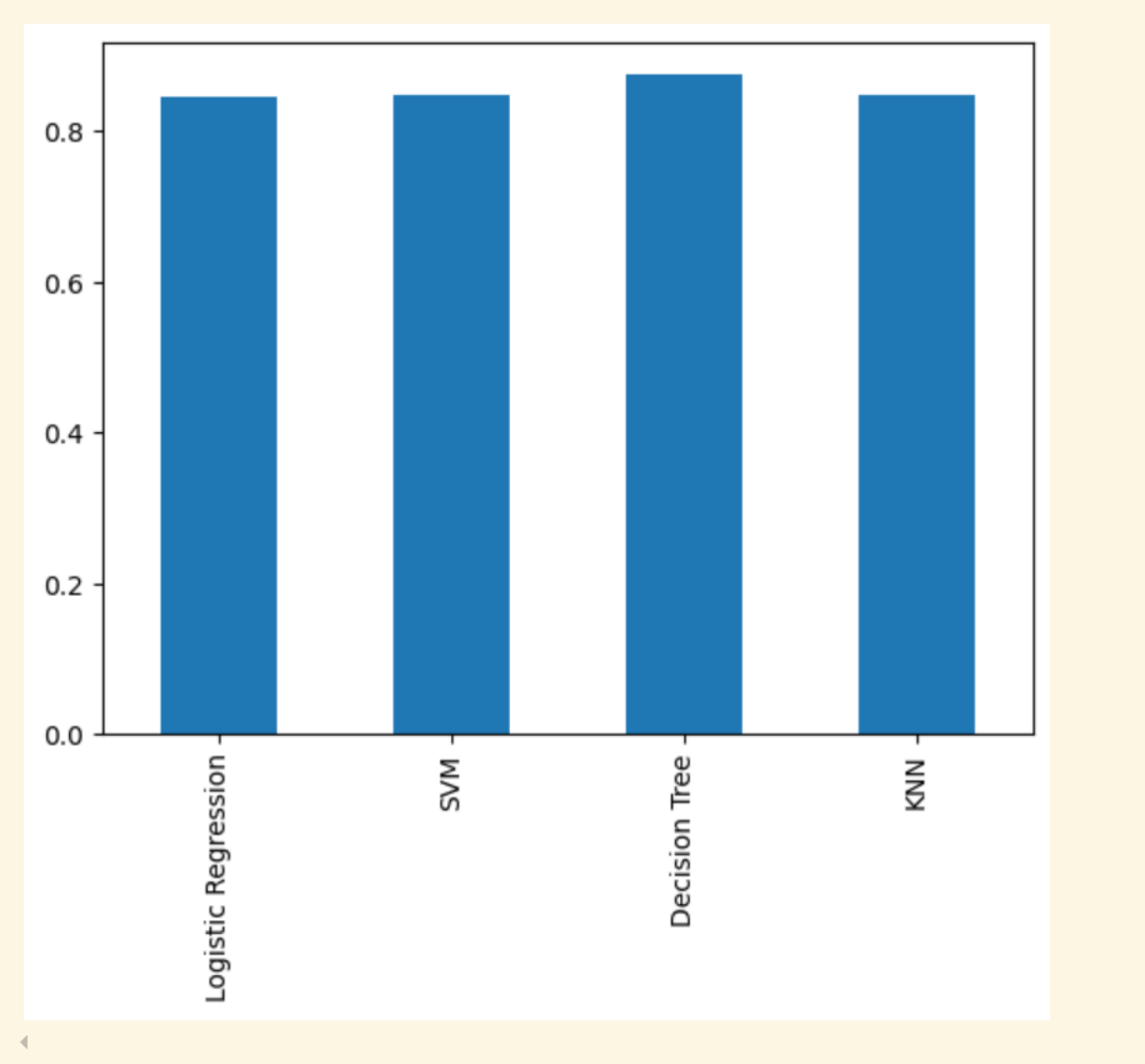


Section 5

Predictive Analysis (Classification)

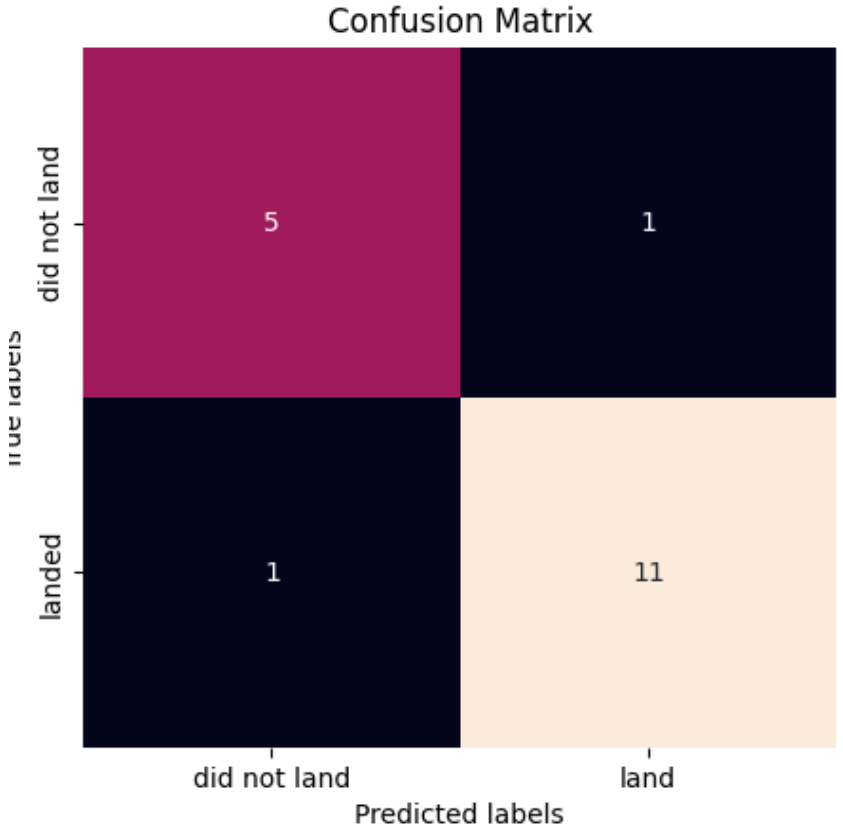
Classification Accuracy

- Due to the small test sample size (18 samples), test set scores alone cannot confirm the best model.
- Evaluating all models on the whole dataset confirms that the Decision Tree model performs best, with the highest accuracy (0.87).
- Other models, including Logistic Regression, SVM, and KNN, achieved an accuracy of 0.8333, making Decision Tree the most suitable choice for this dataset.



Confusion Matrix

- High Accuracy (87.4%): The model effectively predicts Falcon 9 first-stage landings with a strong balance of true positives and true negatives.
- Low False Negatives: Every actual successful landing was correctly identified, ensuring high reliability for aerospace operations.
- Low False Positives: A single false positive is less critical than false negatives, making the model suitable for practical use.
- Balanced Performance: The model slightly favors predicting success, aligning well with industry needs for safety, cost estimation, and mission planning.



Conclusions

- Decision Tree Model is the Best: It achieved the highest accuracy, making it the most effective algorithm for this dataset.
- Payload Mass and Success: Launches with lower payload mass tend to have better outcomes.
- Geographical Influence: Most launch sites are near the Equator, and all are close to the coast, optimizing launch conditions.
- Increasing Success Over Time: The success rate of launches has steadily improved over the years.
- KSC LC-39A Leads in Success: This launch site has the highest success rate among all sites.
- Orbit Performance: ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Takeaways

- Launch Site Impact: KSC LC-39A has the highest success rate (41.7%), suggesting optimal conditions or processes.
- Booster Performance: The “FT” booster version is highly reliable across various payload masses, making it a strong candidate for future missions.
- Payload Mass Not a Sole Factor: Success rates do not show a clear pattern based on payload mass, highlighting the importance of site conditions and booster versions.
- Visual Insights: Interactive data tools (Folium, Plotly Dash) provided valuable geographical and operational patterns, aiding in 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

Thank you!