IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

TarunS
28 September 2025

Outline

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

Executive Summary

- Performed SpaceX Launch Data Collection through API and Webscraping
- Conducted EDA on SpaceX launch data to uncover patterns in launch success.
- Performed predictive modeling using Logistic Regression, SVM, Decision Trees, and KNN to determine the best model for predicting Stage 1 launch success
- Identified key factors influencing launch success: payload range, booster version, and launch site
- Conclusion: Block 5 booster and payloads between 2000–6000 kg had the highest success rates.

Introduction

- Aim of the Project: SpaceX aims to reduce launch costs through reusability. Understanding launch success factors is critical to determine the highly successful method to launch and optimize costs.
- Problem Statement: What factors influence launch success? Can we predict outcomes based on payload, booster, and site?

Section 1

Methodology

Methodology

Executive Summary

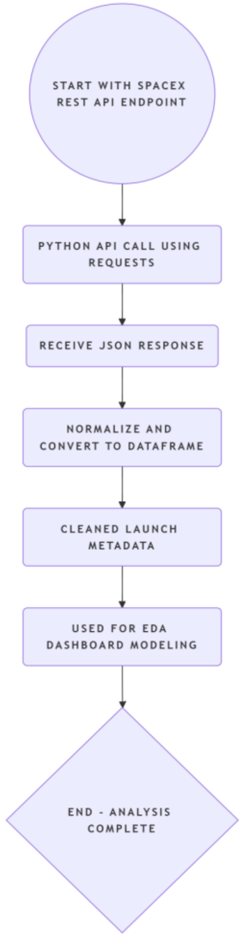
- Data collection methodology:
  - Data was collected by using the REST API calls to SpaceX and web scraping for booster details.
- Perform data wrangling:
  - Cleaned payloads, standardized booster names, merged datasets.
  - Created derived features such as year of launch, binary success label (class), and booster category.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized features, split data, tuned hyperparameters using GridSearchCV
  - Compared all models based on test accuracy and validation performance

Data Collection

- Data Sources: Combined data from REST APIs, web scraping, and cloud-hosted CSVs.
- SpaceX REST API: Used to retrieve launch metadata including flight number, payload mass, orbit, and launch site.
- Web scraping: Extracted booster version details and landing outcomes from SpaceX's official launch archive
- Cloud datasets: Supplemented with curated CSVs from IBM Cloud Object Storage for modeling and dashboard development.
- Automated pipelines: Used Python scripts to fetch, clean, and merge data into a unified DataFrame.
- Schema alignment: Ensured consistent column naming and data types across sources.

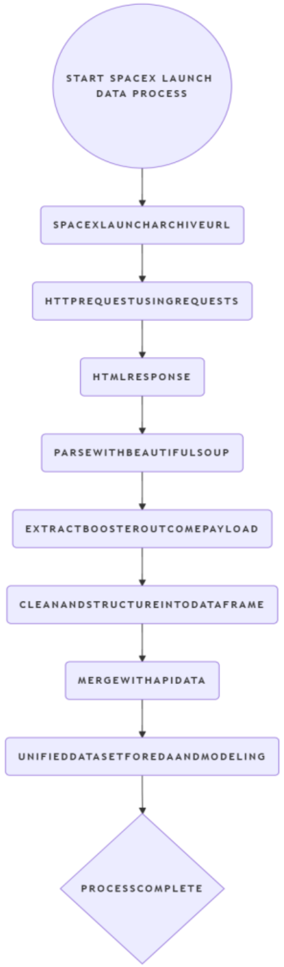
Data Collection – SpaceX API

- Retrieved launch metadata using SpaceX REST API via structured GET requests.
- Extracted fields such as flight number, launch site, payload mass, orbit type, booster version, and landing outcome.
- Automated the API calls using Python and stored results in a Pandas DataFrame.
- Ensured schema consistency and handled missing values during ingestion.
- Used the API data as the foundation for EDA, dashboard visualizations, and predictive modeling.
- GitHub URL - Link



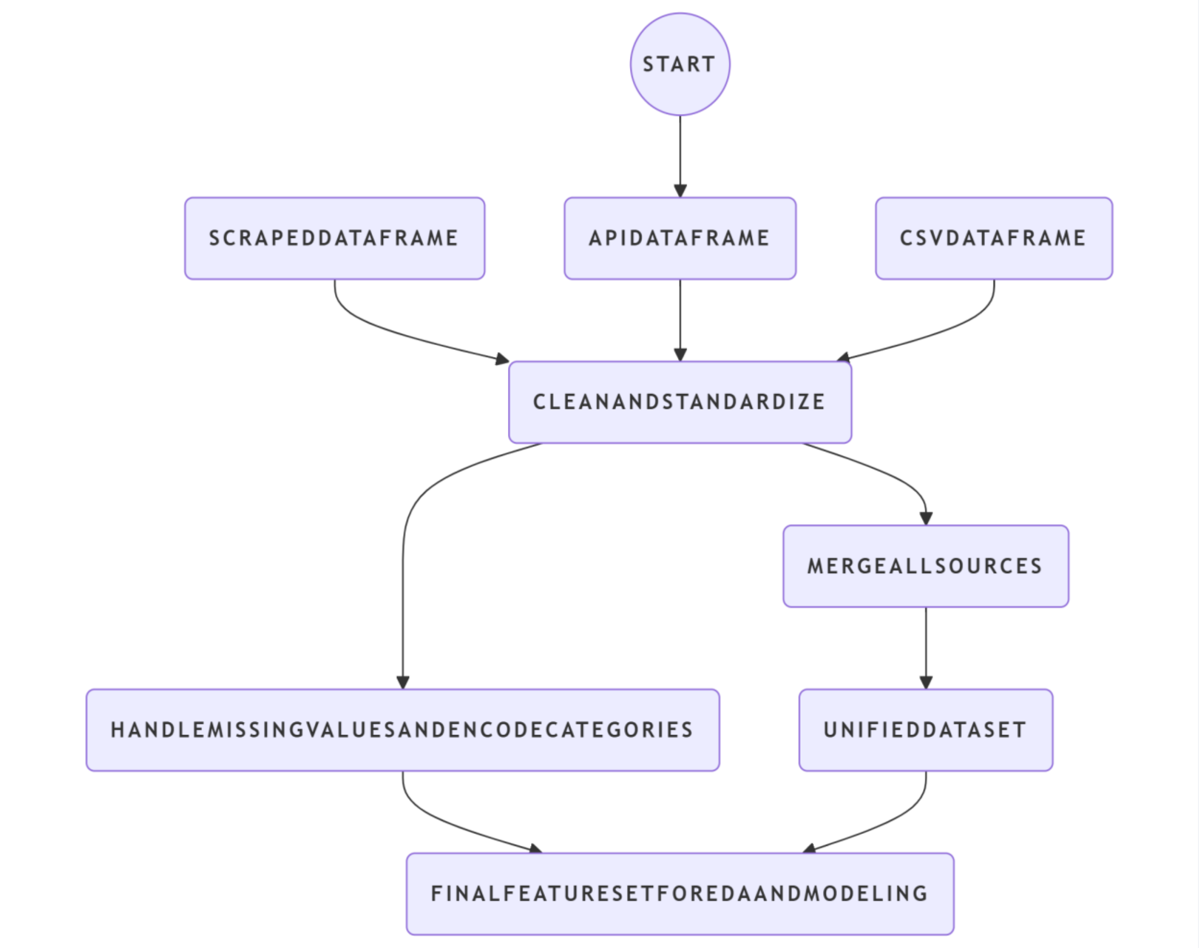
Data Collection - Scraping

- Scraped SpaceX launch archive to extract booster version, landing outcome, and payload details not available via API.
- Used Python with BeautifulSoup to parse HTML tables from SpaceX's mission logs.
- Cleaned and structured scraped data into a Pandas DataFrame for merging with API results.
- Validated scraped data against API data to ensure consistency and completeness.
- Stored final output as a CSV for reproducibility and modeling in future.
- GitHub URL: Link



Data Wrangling

- Unified multiple datasets from API, scraping, and cloud sources into a single DataFrame.
- Handled missing values using imputation and filtering.
- Standardized column names and ensured consistent data types across features.
- Created derived features such as class (binary success label), year, and booster categories.
- Encoded categorical variables using one-hot encoding for modeling compatibility.
- Validated schema alignment to ensure smooth integration with dashboards and ML pipelines.
- GitHub URL: Link



EDA with Data Visualization

- **Pie Charts:**

  - Launch Success by Site: Compared success ratios across launch sites. Site with Highest Success Rate: Identified VAFB SLC-4E as the most reliable site.

- **Scatter Plots:**

  - Payload vs. Launch Outcome: Revealed optimal payload range (2000–6000 kg) for success. Booster Version vs. Success: Highlighted Block 5 boosters as most effective.

- **Bar Charts:**

  - Success Rate by Orbit Type: Compared performance across different orbit categories. Booster Performance: Ranked boosters by success count.

- **Line Chart:** Yearly Success Trend: Showed improvement in launch success over time.
- **Interactive Filters:** Used sliders and dropdowns to dynamically explore payload ranges, booster types, and launch sites.
- **Folium Map:** Mapped launch sites with success/failure overlays and also proximity to key infrastructure
- **Plotly Dash Dashboard:** Dynamic filters updating live charts to highlight top payload and booster success
- **GitHub URL:** Link 1 - Folium | Link 2 - Plotly Dash

EDA with SQL

- Identified all unique launch sites using SELECT DISTINCT and filtered launch records for CCAFS LC-40 Payload using LIKE 'CCA%'
- Calculated total payload mass for NASA missions using SUM() and WHERE customer = 'NASA' and Computed average payload for booster version F9 v1.1 using AVG() and WHERE booster_version = 'F9 v1.1'
- Retrieved first successful ground landing date using MIN() and WHERE landing_outcome = 'Success (ground pad)' and listed boosters with successful drone ship landings and payload between 4000–6000 using BETWEEN and multiple WHERE conditions
- Counted total success and failure outcomes using GROUP BY landing_outcome and found booster with maximum payload using MAX() and ORDER BY payload_mass DESC
- Filtered failed drone ship landings in 2015 using WHERE landing_outcome = 'Failure (drone ship)' AND YEAR(date) = 2015 and ranked landing outcomes between 2010–06–04 and 2017–03–20 using GROUP BY, COUNT(), and ORDER BY DESC
- GitHub URL: Link

Build an Interactive Map with Folium

Map Objects Added:

- Markers: Plotted all SpaceX launch sites globally for spatial reference.
- Color-coded Circles: Indicated launch outcomes — green for success, red for failure.
- Lines: Connected launch sites to nearby infrastructure (coastlines, highways, railways).
- Popups: Displayed site name, payload mass, and booster version on click.

Why These Objects Were Added:

- To visualize launch site distribution and outcome patterns.
- To perform proximity analysis for infrastructure impact on launch logistics.
- To enhance interactivity and insight through clickable areas such as popups.

- **GitHub URL:** Link

Build a Dashboard with Plotly Dash

Plots and Interactions Added:

- Pie Charts: Launch success count by site and to find site with highest success ratio
- Scatter Plot: Payload vs. launch outcome with dynamic payload range slider
- Dropdown Filters: Select launch site and booster version to update visuals
- Range Slider: Adjust payload mass range to explore success patterns

Why These Were Added:

- To enable interactive visualization of payload, site, and booster impact
- To identify optimal payload ranges and high-performing boosters
- To make insights measurable and interactive for various launch metrics

GitHub URL: Link 1 | Link 2

Predictive Analysis (Classification)

Summary of Model Development:

- Selected key features: payload mass, booster version, launch site, orbit type
- Encoded categorical variables and standardized numerical features
- Split data into training and test sets (80/20)
- Applied four classifiers: Logistic Regression, SVM, Decision Tree, KNN
- Tuned hyperparameters using GridSearchCV with 10-fold cross-validation
- Evaluated models using test accuracy and confusion matrix
- Identified best-performing model based on validation score and generalization
- Outcome: SVM with RBF kernel achieved the highest validation accuracy, followed closely by Logistic Regression.
- **GitHub URL:** Link

Results

Exploratory Data Analysis (EDA) Results:

- Uncovered key patterns in launch success across payload ranges, booster versions, and launch sites using pie charts, scatter plots, bar charts, and SQL queries.

Interactive Analytics Demo:

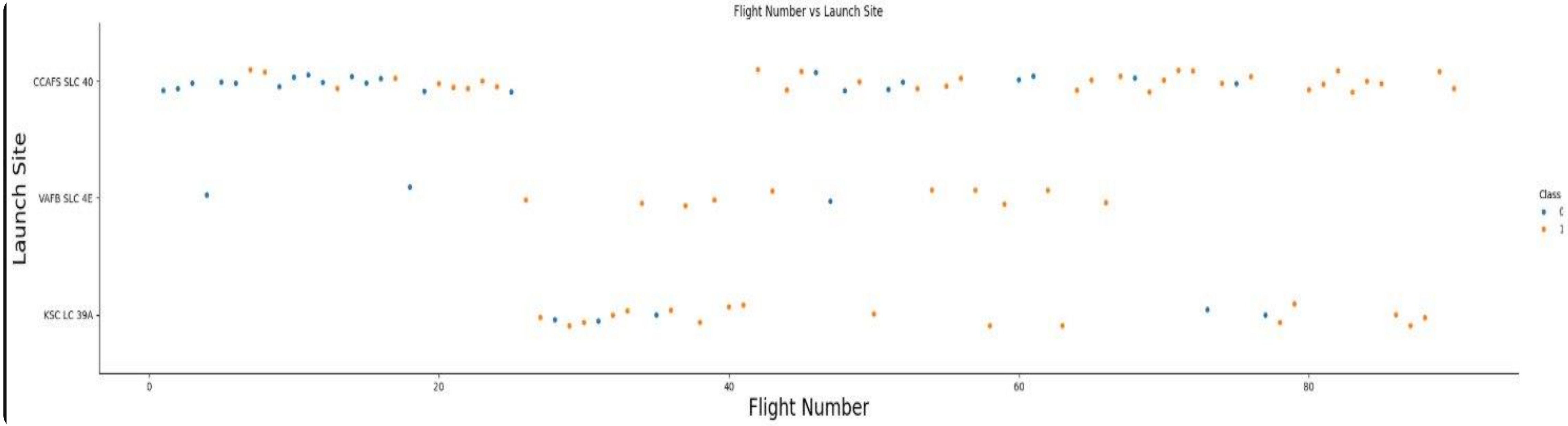
- Built dynamic dashboards with Plotly Dash and interactive maps with Folium to explore payload success, booster performance, and site proximity in real time.

Predictive Analysis Outcomes:

- Trained and tuned four classification models; SVM with RBF kernel achieved highest validation accuracy. Confusion matrix confirmed strong predictive power on unseen data.

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

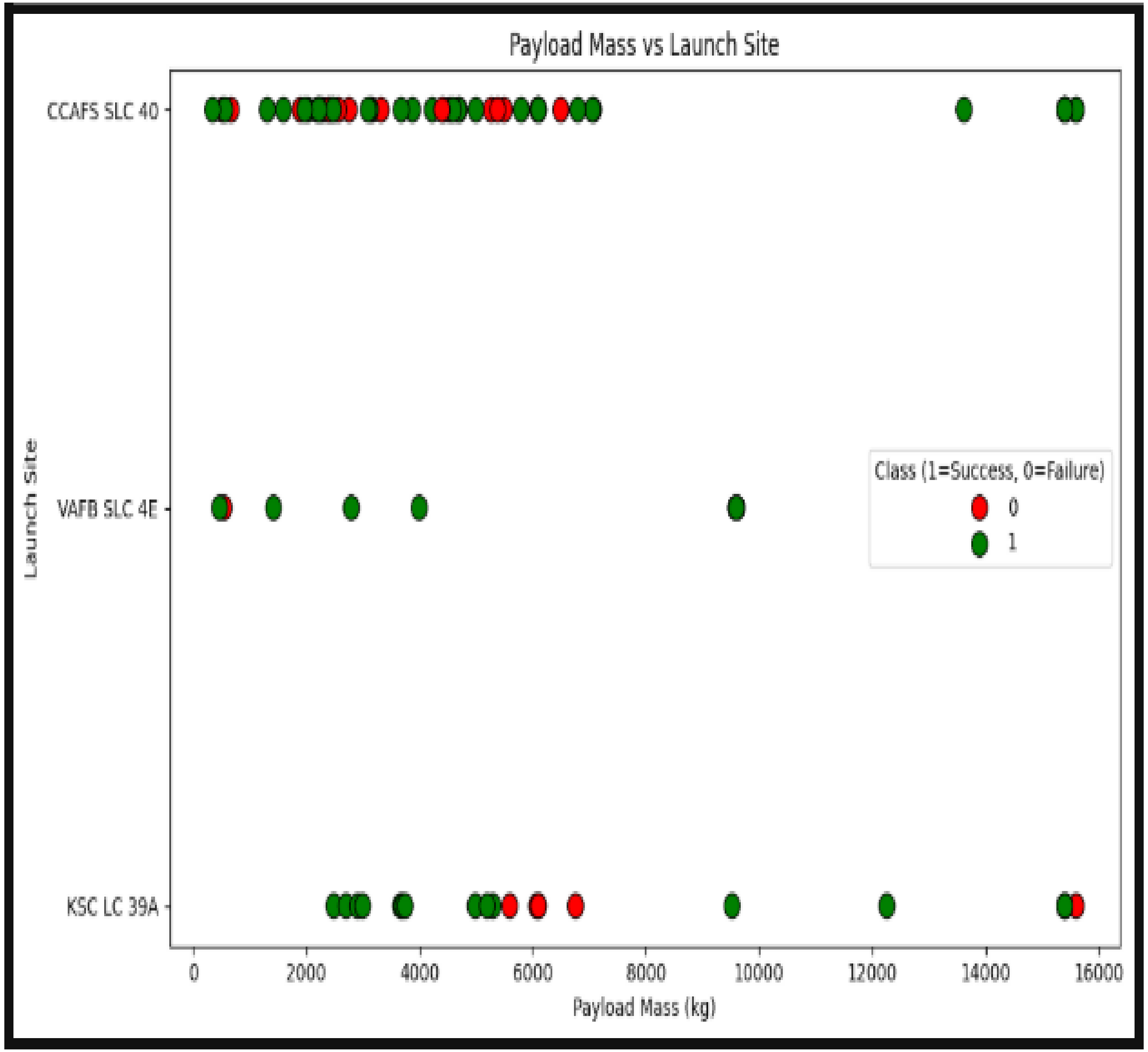
Scatter Plot of Flight Number vs. Launch Site:

- Plotted Flight Number on the x-axis and Launch Site on the y-axis.
- Each point represents a unique launch event.
- Color-coded by launch outcome (success/failure) for added context.
- **Orange (Class 1):** Represents successful launches and where launch outcome was classified as a success.
- **Blue (Class 0):** Represents failed launches and where launch outcome was classified as a failure.

Payload vs. Launch Site

Scatter Plot of Payload vs. Launch Site:

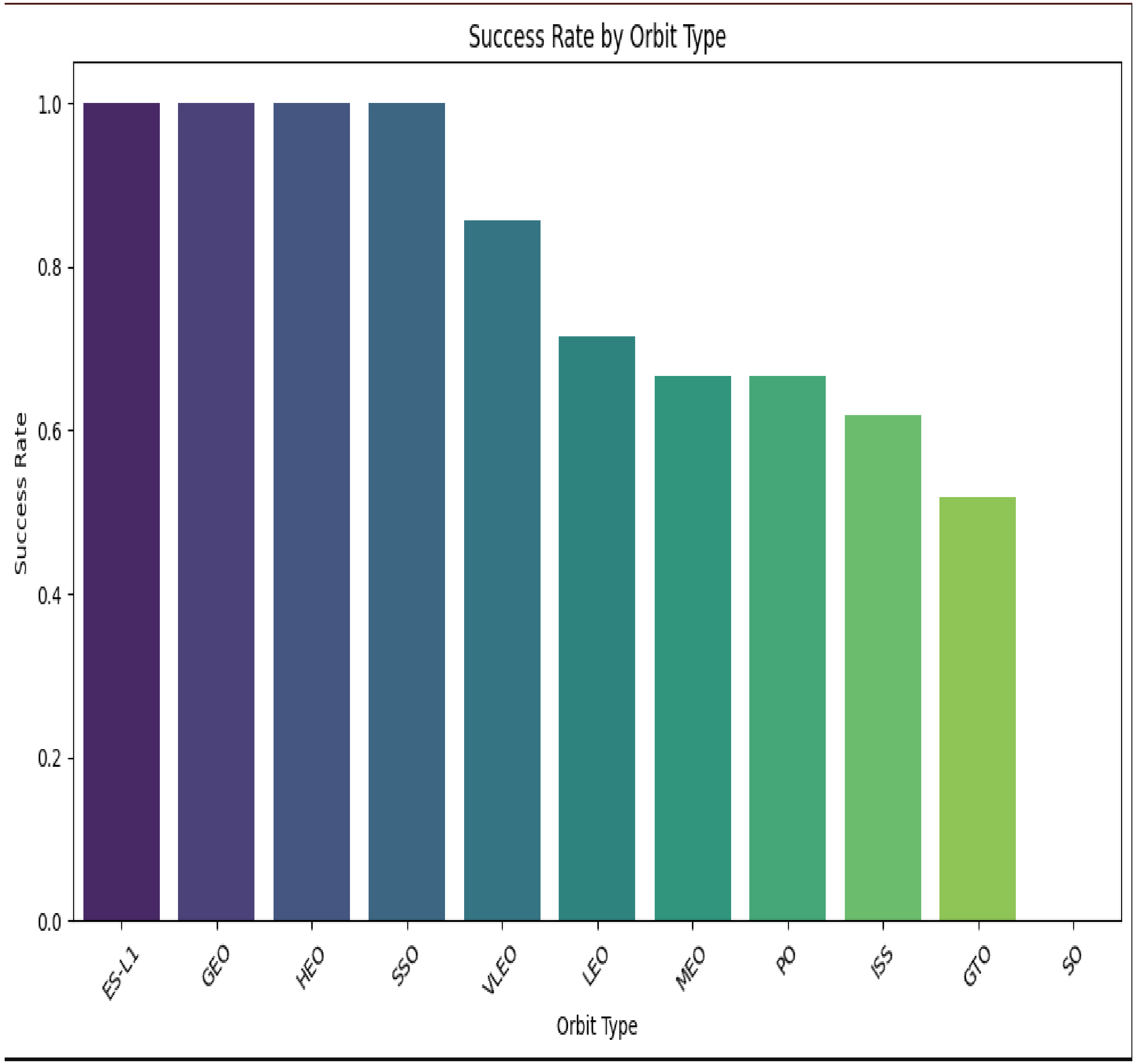
- Plotted Payload Mass on the x-axis and Launch Site on the y-axis.
- Each point represents a unique launch event.
- Color-coded by launch outcome (success/failure) for added context.
- Green (Class 1): Represents successful launches and where launch outcome was classified as a success.
- Red (Class 0): Represents failed launches and where launch outcome was classified as a failure.



Success Rate vs. Orbit Type

Bar Chart of Success Rate vs. Orbit Type:

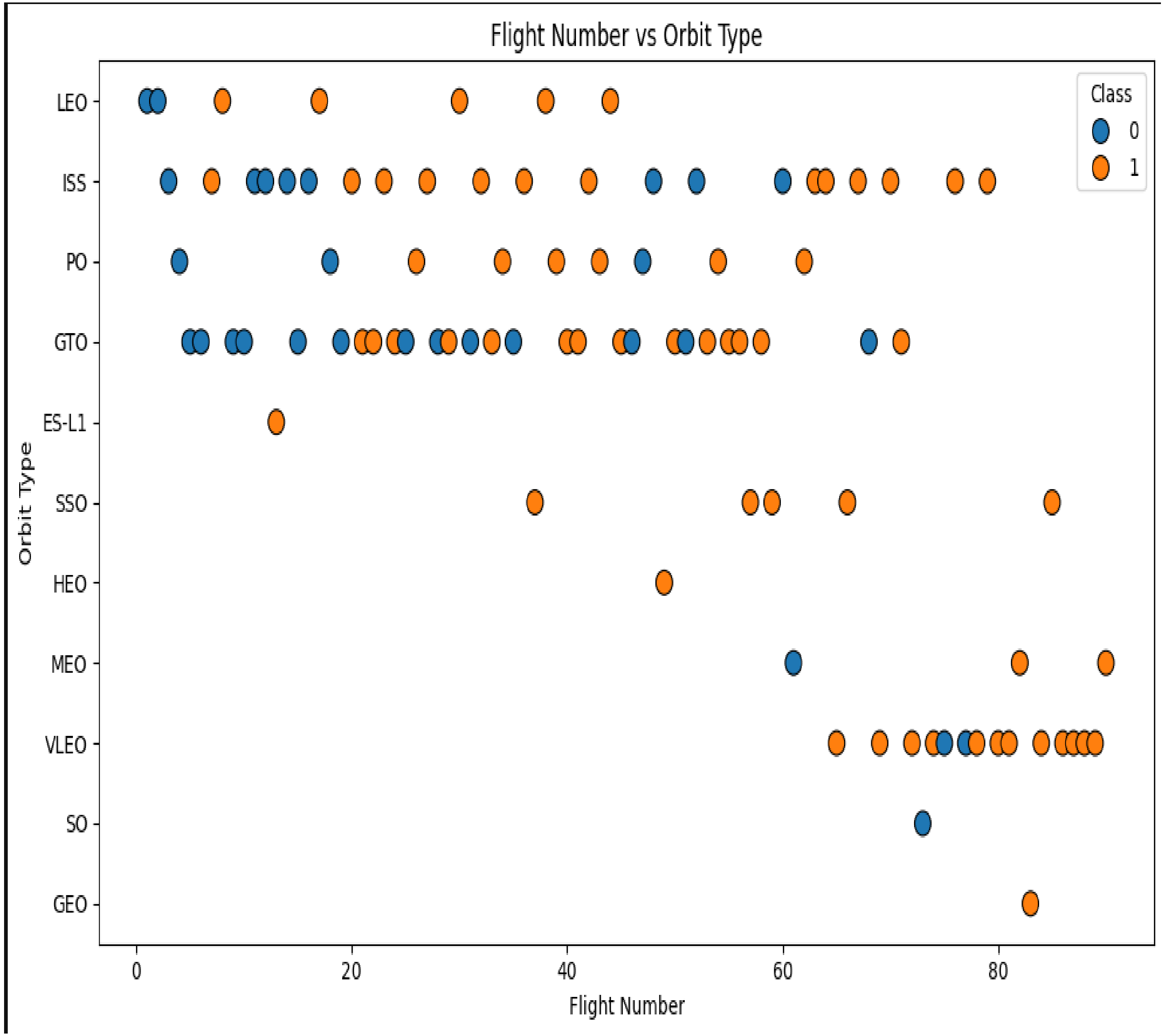
- Plotted Orbit Type on the x-axis and Success Rate on the y-axis.
- Each bar represents the proportion of successful launches for a given orbit type.
- Height of the bar indicates reliability of missions targeting that orbit.
- ES-L1, GEO, and HEO showed highest success rates, indicating strong mission reliability.
- SO had the lowest success rate, suggesting higher risk or complexity for that orbit.



Flight Number vs. Orbit Type

Scatter Plot of Flight Number vs. Orbit Type:

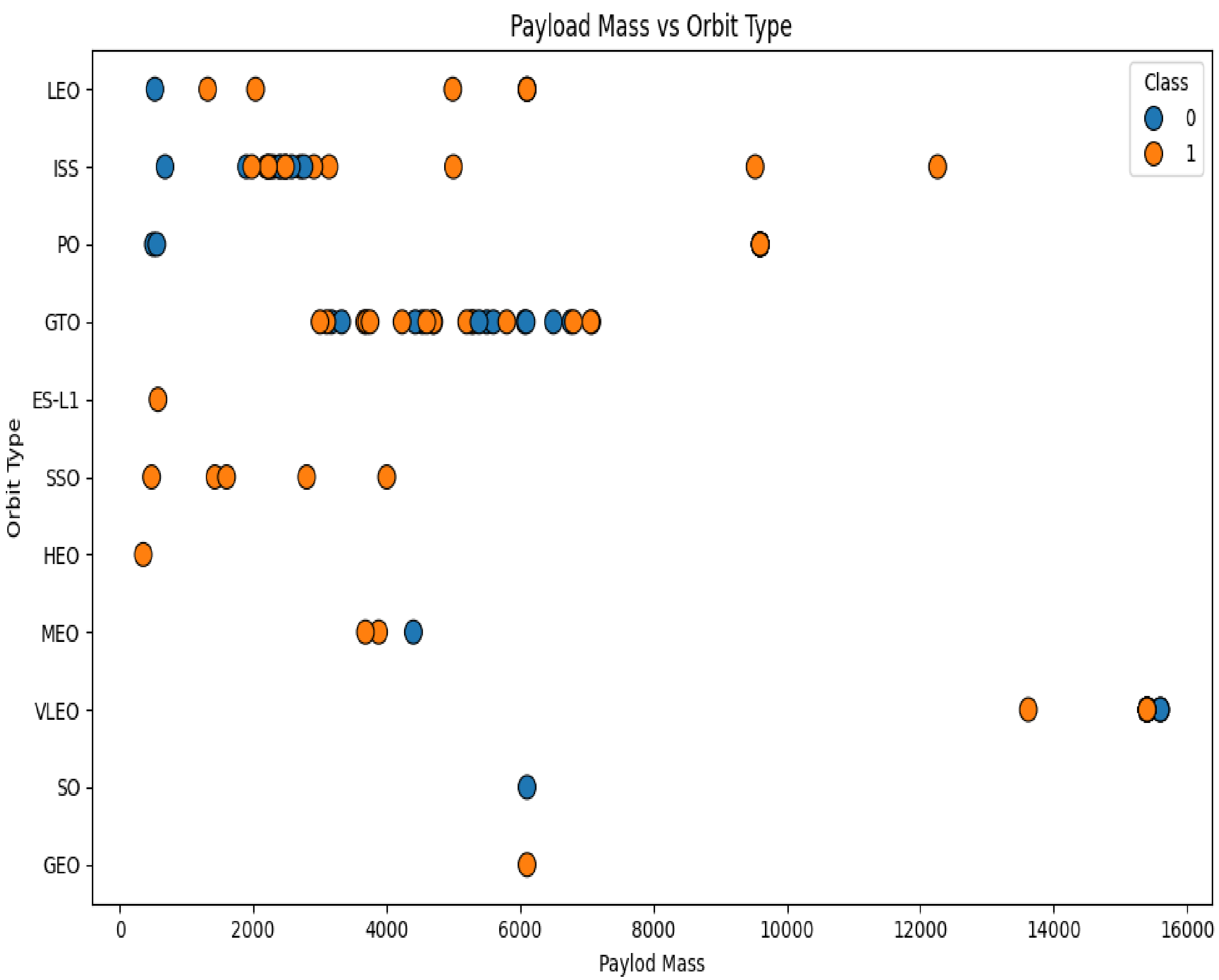
- Plotted Flight Number on the x-axis and Orbit Type on the y-axis.
- Each point represents a unique launch event.
- Color-coded by launch outcome (success/failure) for added context.
- Orange (Class 1): Represents successful launches and where launch outcome was classified as a success.
- Blue (Class 0): Represents failed launches and where launch outcome was classified as a failure.
- In LEO orbit, success improves with higher flight numbers.
- In GTO orbit, success appears unrelated to flight number — outcomes remain mixed.



Payload vs. Orbit Type

Scatter Plot of Payload vs. Orbit Type:

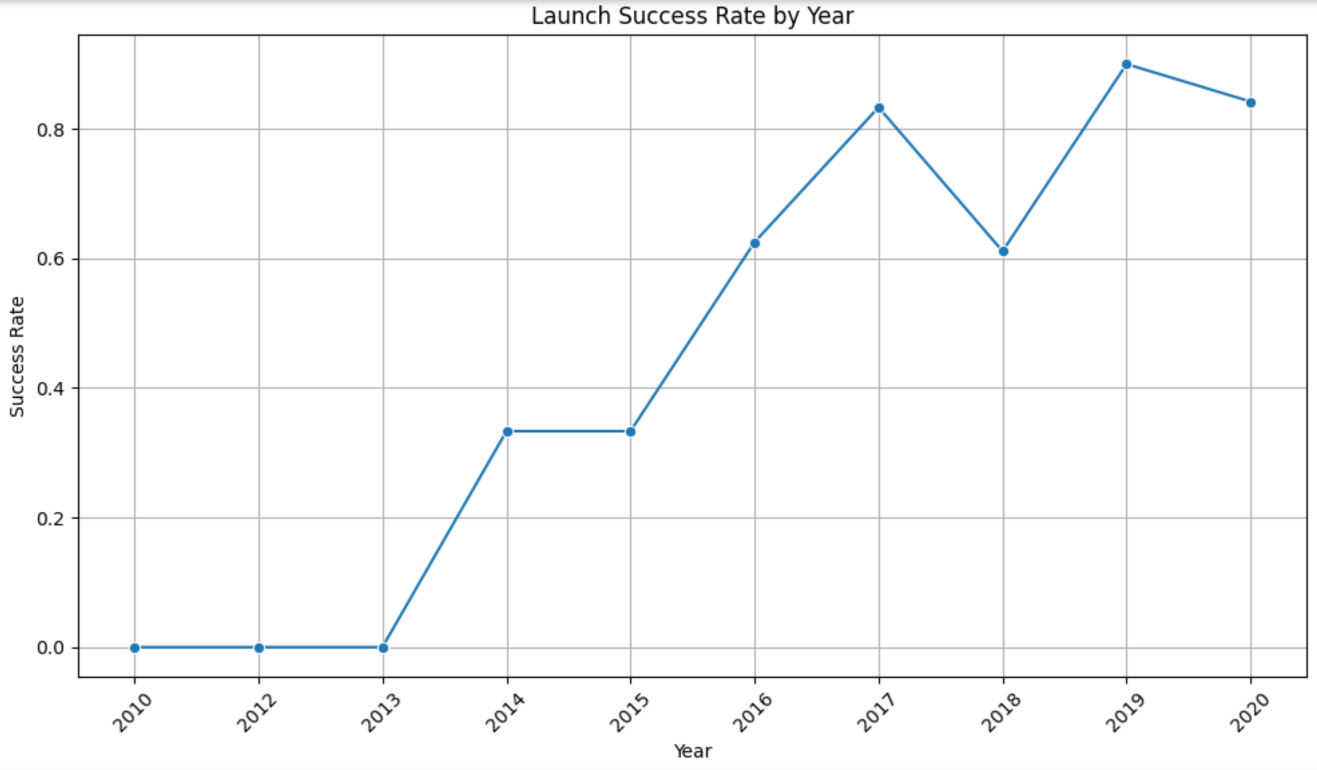
- Plotted Payload Mass on the x-axis and Orbit Type on the y-axis.
- Each point represents a unique launch event.
- Color-coded by launch outcome (success/failure) for added context.
- Orange (Class 1): Represents successful launches and where launch outcome was classified as a success.
- Blue (Class 0): Represents failed launches and where launch outcome was classified as a failure.
- In LEO orbit, higher payloads tend to correlate with successful outcomes.
- In GTO orbit, success appears inconsistent across payload masses.



Launch Success Yearly Trend

Line Chart of Launch Success Rate by Year:

- Plotted Year on the x-axis and Average Launch Success Rate on the y-axis.
- Each point represents the average success rate for launches in that year.
- Line chart shows trend of increasing reliability over time.
- Success rate remained at 0.0 from 2010 to 2013, then steadily improved.
- Peaked near 0.9 in 2017 and 2019, indicating strong operational success rate.
- Slight dips in 2018 and 2020 suggest occasional setbacks.



All Launch Site Names

SQL Query to Retrieve Unique Launch Sites:

- Used the query: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
- Returned the following unique launch site names:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

- These sites represent the primary launch locations used by SpaceX across different missions. Identifying them helped in analyzing site-specific performance and trends.

Launch Site Names Begin with 'CCA'

SQL Query to Retrieve Launch Sites Beginning with 'CCA':

- Used the query: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- Returned 5 records where the launch site name starts with “CCA”:
- All records are from CCAFS LC-40, showing early SpaceX missions.
- Payloads range from 0 to 677 kg, mostly targeting LEO orbits.
- This query helped filter site-specific launch history and also showed that Mission outcomes were successful.
- However, landing outcomes varied, including parachute failures and no attempts.

Total Payload Mass

SQL Query to Calculate Total Payload Mass for NASA (CRS):

- Used the query: %sql SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
- Returned payload mass values for each NASA (CRS) mission.
- Summing these values gives the total payload mass carried by NASA boosters.
- This helped quantify NASA's payload mass range to launch operations – with the lowest payload mass being 500 KG and the highest being 3310 KG

Average Payload Mass by F9 v1.1

SQL Query to Calculate Average Payload Mass for Booster Version F9 v1.1:

- Used the query: %sql SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
- Returned payload mass values for missions using F9 v1.1:- 3170, 3325, 2296, 1316, 4535 kg
- These values can be analyzed to assess the typical payload capacity of the F9 v1.1 booster.
- This analysis may help evaluate booster performance and improve mission outcome success.

First Successful Ground Landing Date

SQL Query to Find First Successful Ground Pad Landing Date:

- Used the query: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
- Returned the earliest or the first date of successful ground landing: 2015-12-22
- This query highlights a major milestone in SpaceX's reusability efforts, showcasing the first confirmed success of landing a booster on solid ground.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query to List Boosters with Successful Drone Ship Landings and Payload Mass Between 4000 and 6000 kg:

- Used the query: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
- Returned the following booster versions:

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1029
4. F9 FT B1031.2

- These boosters successfully landed on drone ships while carrying payloads between 4000 and 6000, demonstrating reliable performance.

Total Number of Successful and Failure Mission Outcomes

SQL Query to Count Successful and Failed Mission Outcomes:

- Used the query: %sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome
- Returned the following outcome counts:
- Success: 99
- Success (payload status unclear): 1
- Failure (in flight): 1
- This breakdown highlights SpaceX's high mission success rate, with only one confirmed in-flight failure and one unclear payload status.

Boosters Carried Maximum Payload

SQL Query to List Boosters That Carried the Maximum Payload Mass:

- Used the below query: SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- Returned the following booster versions that carried the heaviest payloads:
- F9 B5 B1048.4, F9 B5 B1046.4, F9 B5 B1051.3, F9 B5 B1056.2, F9 B5 B1049.4, F9 B5 B1051.4, F9 B5 B1058.2, F9 B5 B1056.3, F9 B5 B1059.2, F9 B5 B1051.5, F9 B5 B1058.3, F9 B5 B1051.6, F9 B5 B1058.4 and F9 B5 B1049.7
- These boosters represent the peak payload capacity achieved in SpaceX missions, showcasing the robustness of the Falcon 9 Block 5 series.

2015 Launch Records

SQL Query to List Failed Drone Ship Landings in 2015:

- Used the query: SELECT substr(Date,0,5), Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 0, 5) = '2015'
- Returned the following records:
- Booster Versions: F9 v1.1 B1022, F9 v1.1 B1015
- Launch Site: CCAFS LC-40
- Landing Outcome: Failure (drone ship)
- These records highlighted early challenges in drone ship recovery during 2015 as part of SpaceX's reusability efforts to lower costs.

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

SQL Query to Rank Landing Outcomes Between 2010-06-04 and 2017-03-20:

- Used the 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
- Returned the following ranked outcomes:
- No attempt: 10 | Success (drone ship): 5 | Failure (drone ship): 5 | Success (ground pad): 4 | Controlled (ocean): 2 | Uncontrolled (ocean): 2 | Precluded (drone ship): 1
- This ranking highlighted the evolution of landing strategies, with many early missions opting out of recovery and gradual improvements in drone ship and ground pad landings.

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites on Folium Map

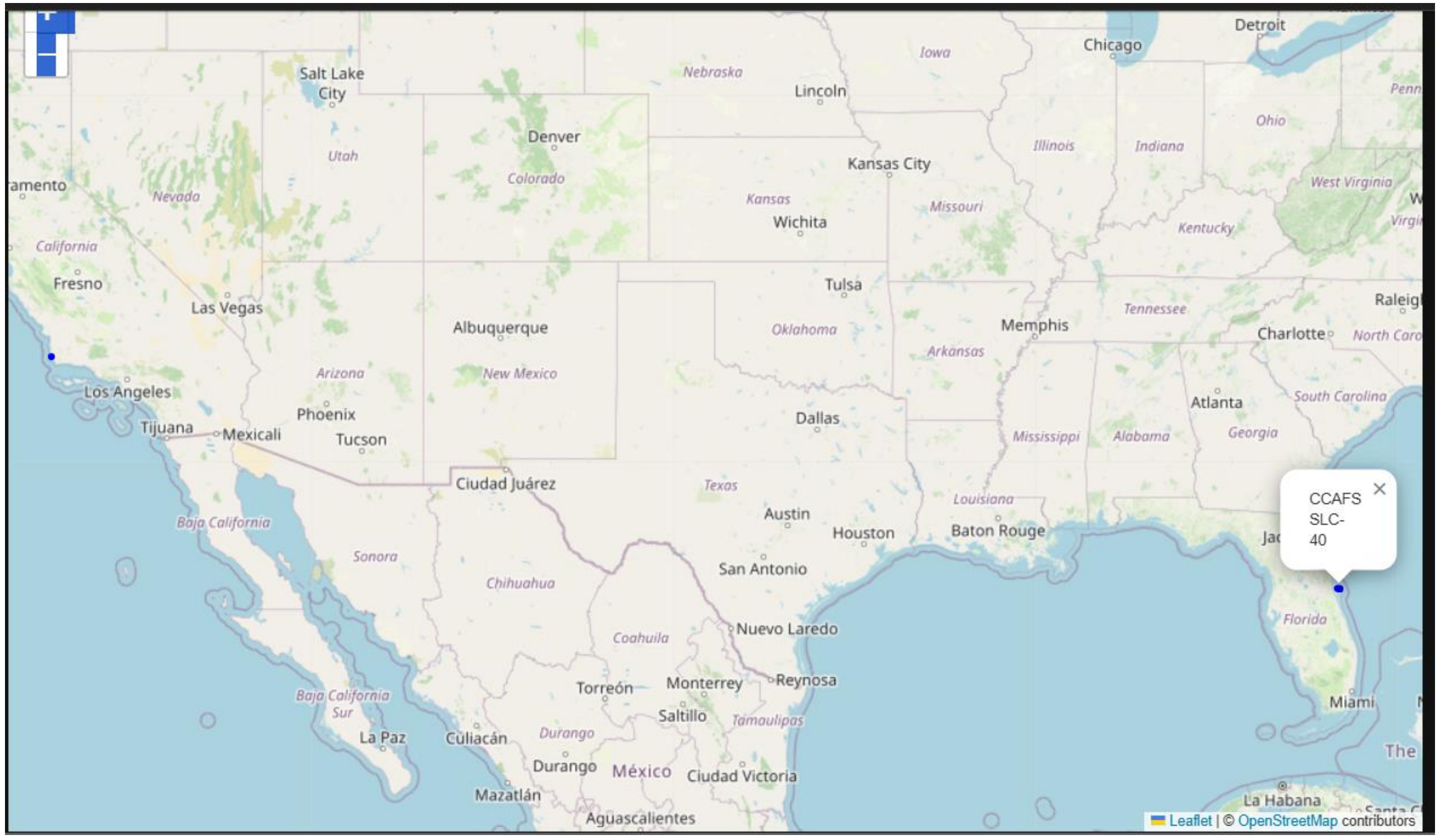
- Explored the Folium map and captured a screenshot showing all launch site markers.

Key elements visible on the map:

- Markers for each launch site, including CCAFS LC-40, VAFB SLC-4E, and KSC LC-39A
- Interactive features, such as zoom and pan, for deeper exploration of site locations

Findings:

- Launch sites are strategically located near coastlines for safe booster recovery.
- Most sites are concentrated in the southeastern and western United States.
- The map provides a clear visual of SpaceX's operational footprint and logistical planning.



SpaceX Launch Outcomes Distribution Across Sites

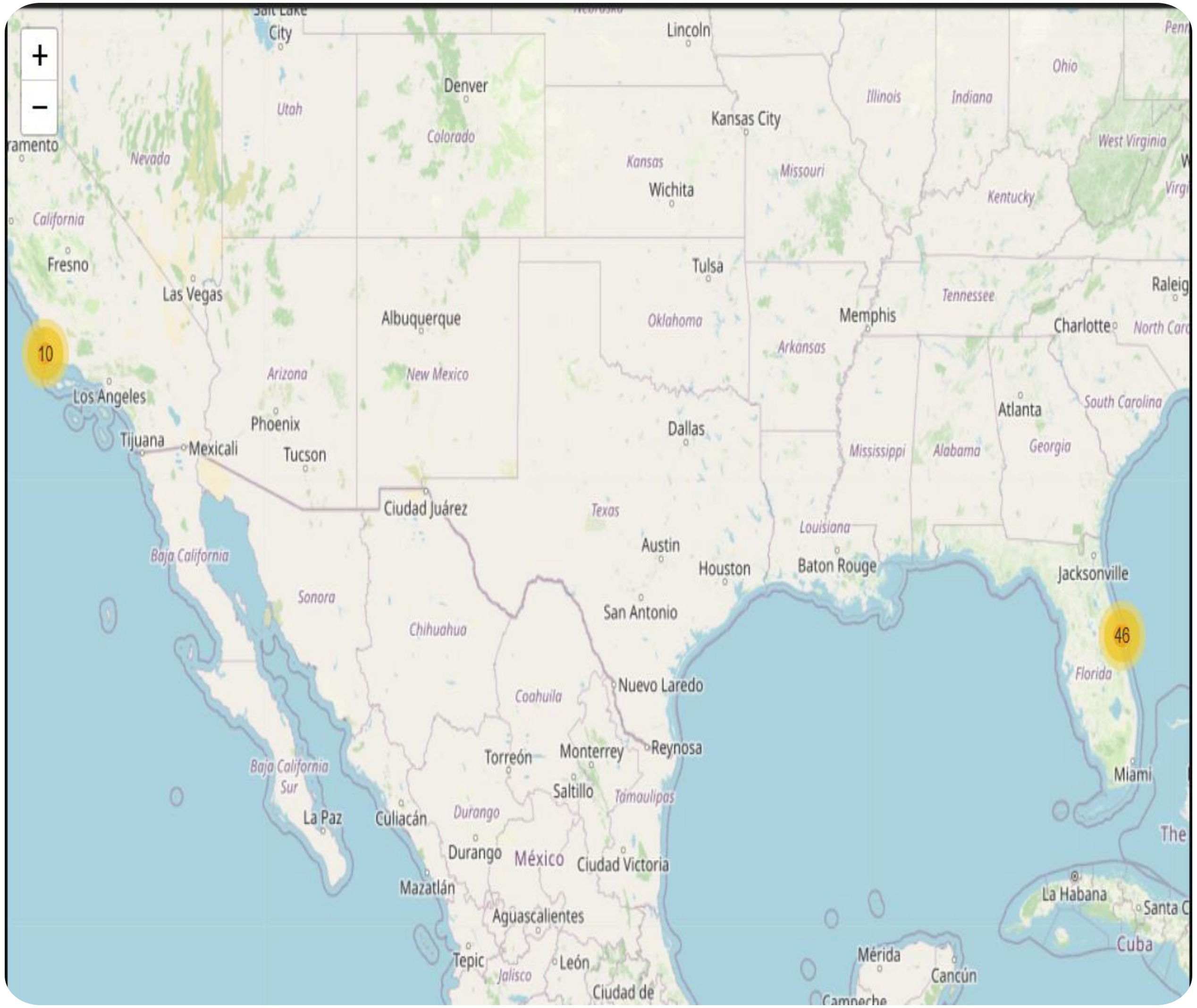
- Explored the Folium map and captured a screenshot showing color-coded launch outcome markers.

Key elements visible on the map:

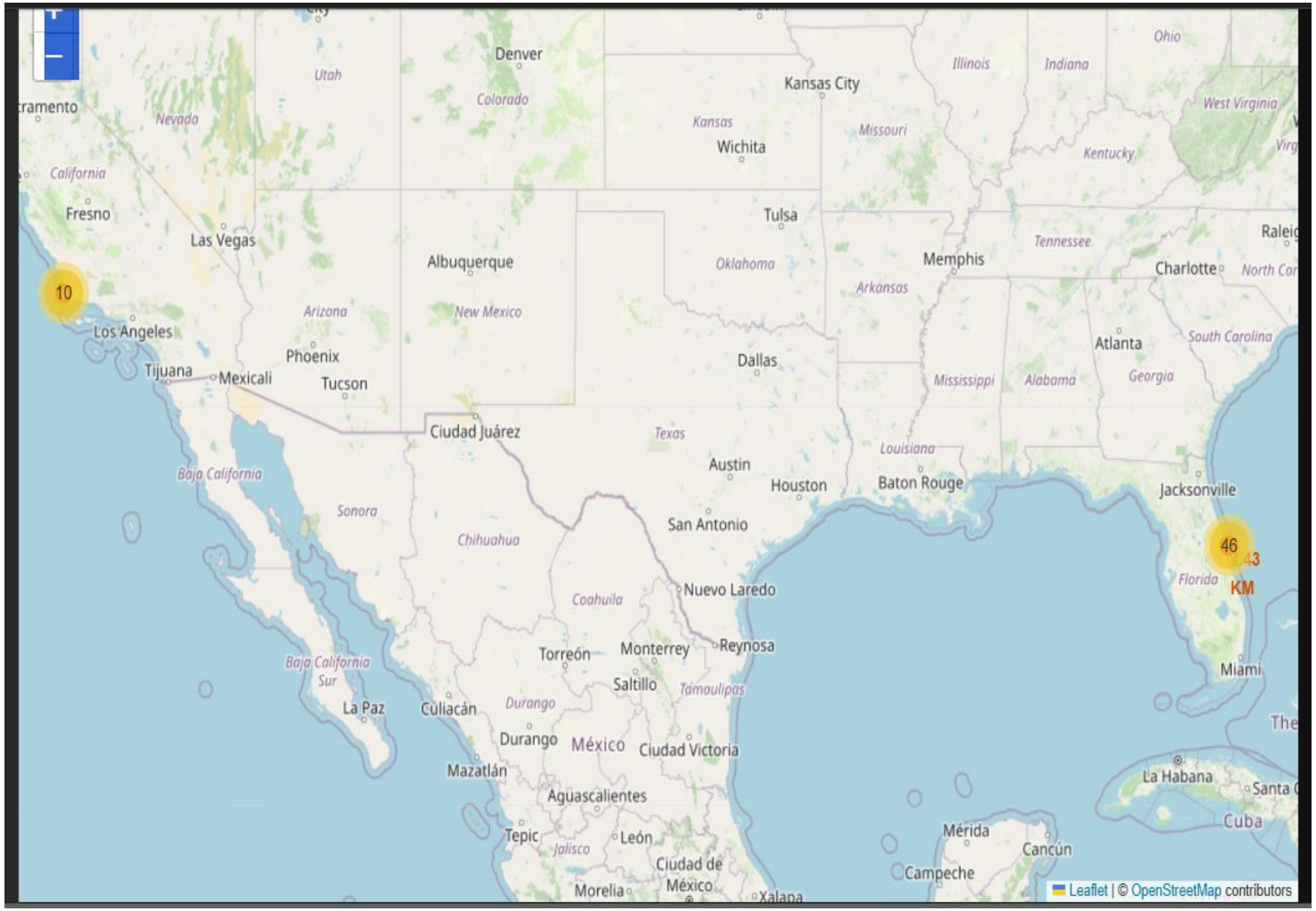
- Colored markers represent launch outcomes: Green for successful landings, Red for failures and Yellow for no attempts or ambiguous outcomes.
- Clustered markers indicate high launch activity at key sites like CCAFS LC-40 and VAFB SLC-4E.
- Folium map shows proximity to coastlines, aiding recovery of boosters.

Findings:

- Most successful landings are concentrated near Florida and California coastlines.
- Drone ship landings are typically offshore, while ground pad successes are near launch sites.
- The map provides a clear visual summary of operational reliability by location.



Folium Map of Launch Site Proximity to Infrastructure



- Explored the Folium map and captured a screenshot showing a selected launch site with proximity overlays.

Key elements visible on the map:

- Launch site marker (e.g., CCAFS SLC-40)
- Distance lines connecting the site to nearby features:
- Coastline - around 7.43 km
- Railway and highway access points: clearly marked for logistical relevance
- Color-coded zones indicating operational areas and safety buffers

Findings:

- Launch sites are strategically positioned near coastlines for booster recovery.
- Proximity to highways and railways supports efficient transport of payloads and hardware.
- Visual overlays help assess logistical feasibility and emergency planning.

Section 4

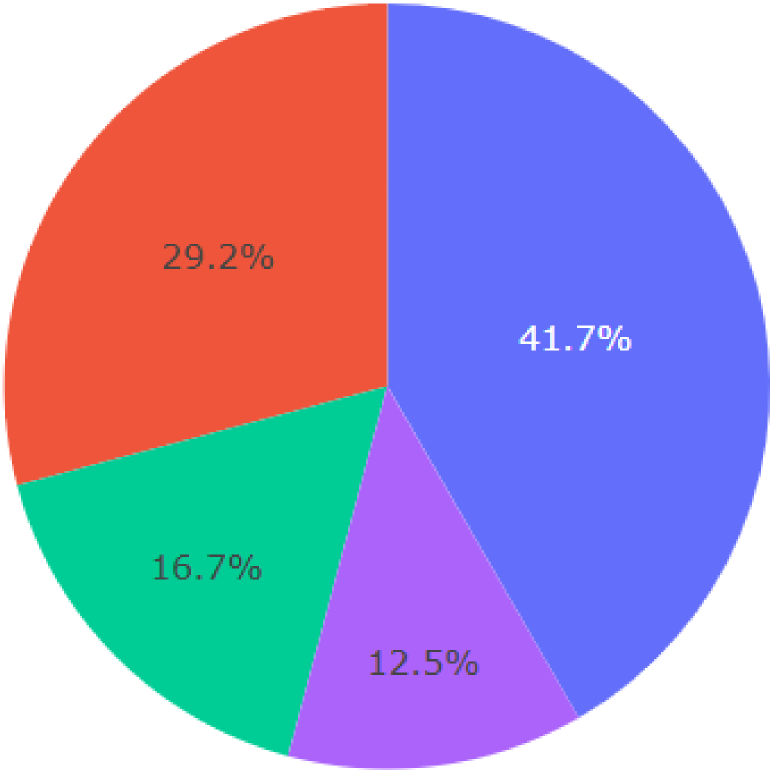
Build a Dashboard with Plotly Dash

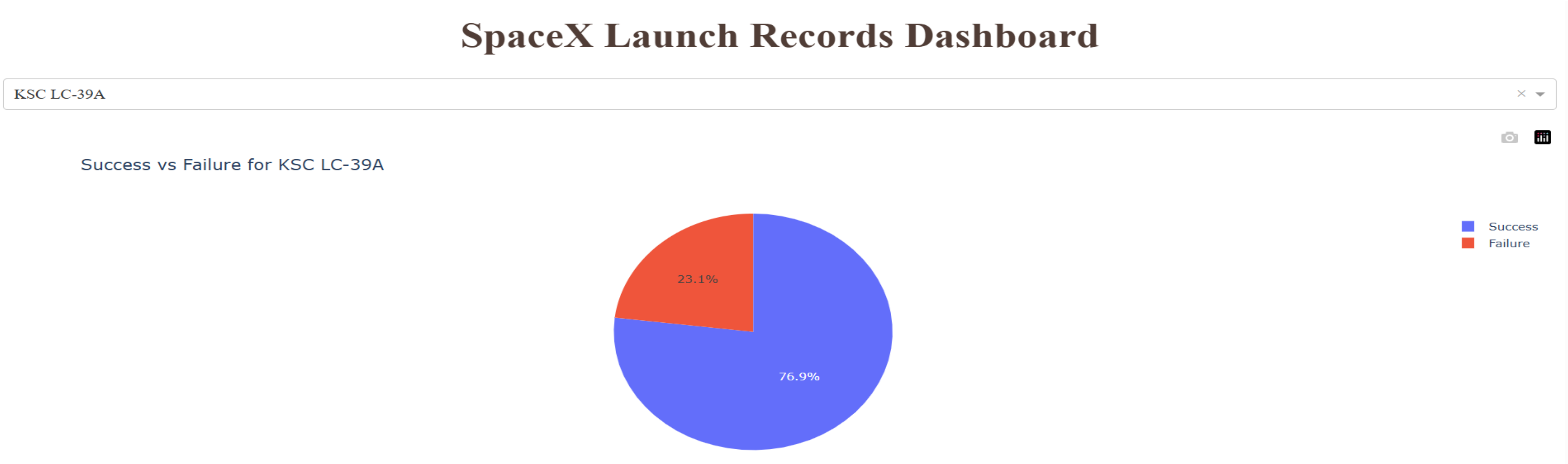
Launch Success Distribution Across Sites

- Displayed a pie chart showing the proportion of successful launches from each SpaceX site:
- KSC LC-39A: 41.7%
- CCAFS LC-40: 20.8%
- VAFB SLC-4E: 16.7%
- CCAFS SLC-40: 12.5%

Findings:

- KSC LC-39A leads in successful launches, reflecting its role in high-profile missions.
- CCAFS sites collectively contribute over 33%, highlighting Florida’s strategic importance.
- VAFB SLC-4E supports West Coast operations, especially polar orbit missions.





Launch Outcome Ratio at KSC LC-39A

- Displayed a pie chart showing the success vs failure ratio for Kennedy Space Center Launch Complex 39A with Success: 76.9% | Failure: 23.1%

Findings:

- KSC LC-39A has the highest launch success ratio among all SpaceX sites.
- The high success rate reflects its use for high-priority missions and advanced booster versions.
- The chart provides a clear visual of operational reliability at this flagship launch site.

Payload Mass vs Launch Success by Booster Version

- Displayed a scatter plot with:
- X-axis: Payload Mass (kg), filtered using an interactive range slider (0–9600 kg)
- Y-axis: Launch Outcome Class (0 = Failure, 1 = Success)
- Color-coded points: Represent different Booster Version Categories (v1.0, v1.1, FT, B4, B5)

Findings:

- Booster Version B5 shows the highest success rate across all payload ranges.
- Launches with payloads between 4000–6000 kg tend to have more consistent success, especially with FT and B5 boosters.

Section 5

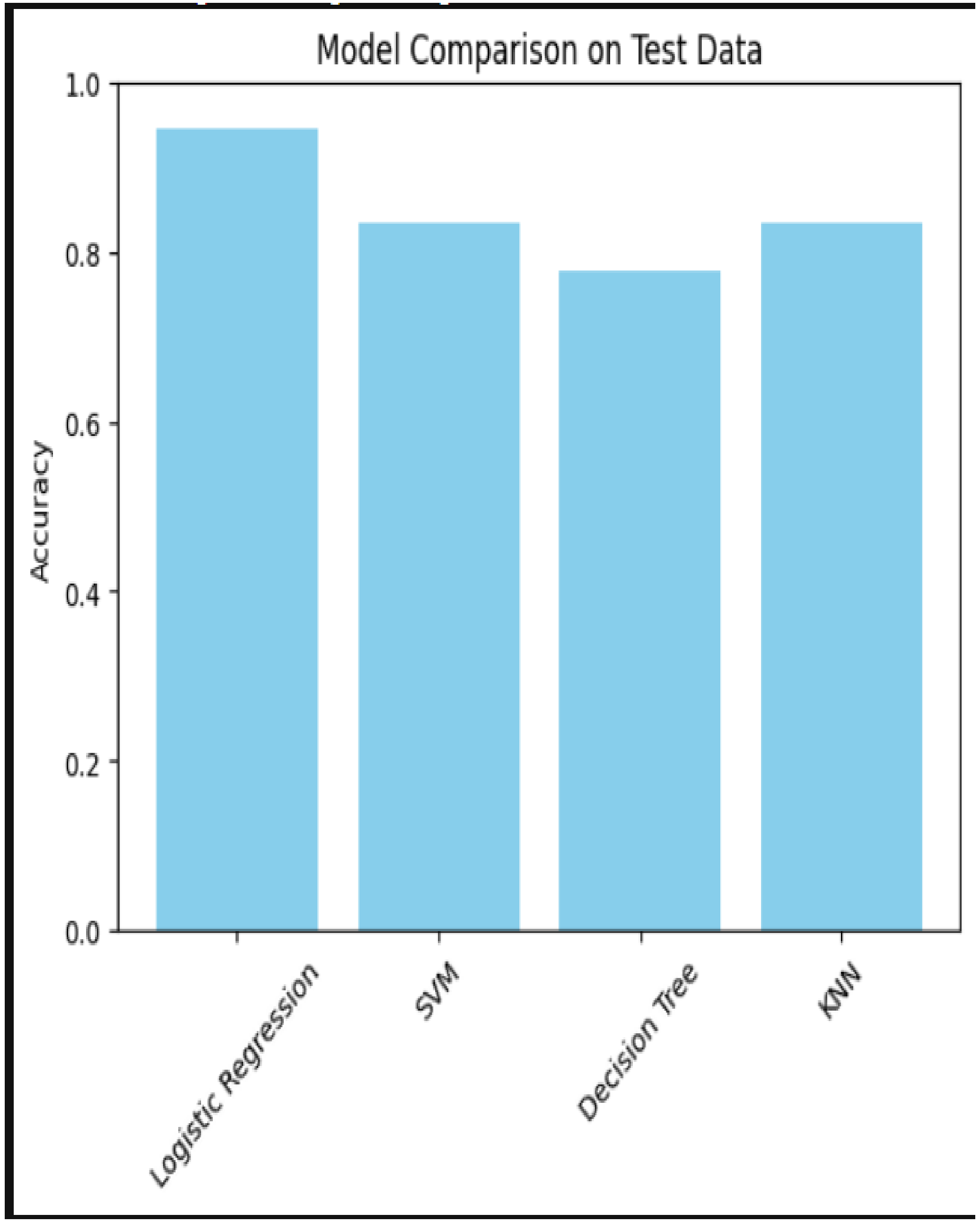
Predictive Analysis (Classification)

Classification Accuracy

- Displayed a bar chart comparing test accuracy of four classification models:
- Logistic Regression: 94.4%
- SVM: 83.3%
- KNN: 83.3%
- Decision Tree: 77.8%

Findings:

- Logistic Regression is the best-performing model, achieving the highest accuracy.
- SVM and KNN perform similarly, but below Logistic Regression.
- Decision Tree shows the lowest accuracy, suggesting potential overfitting or poor generalization.
- This comparison helps justify model selection for deployment and highlights the importance of evaluating multiple classifiers.



Confusion Matrix

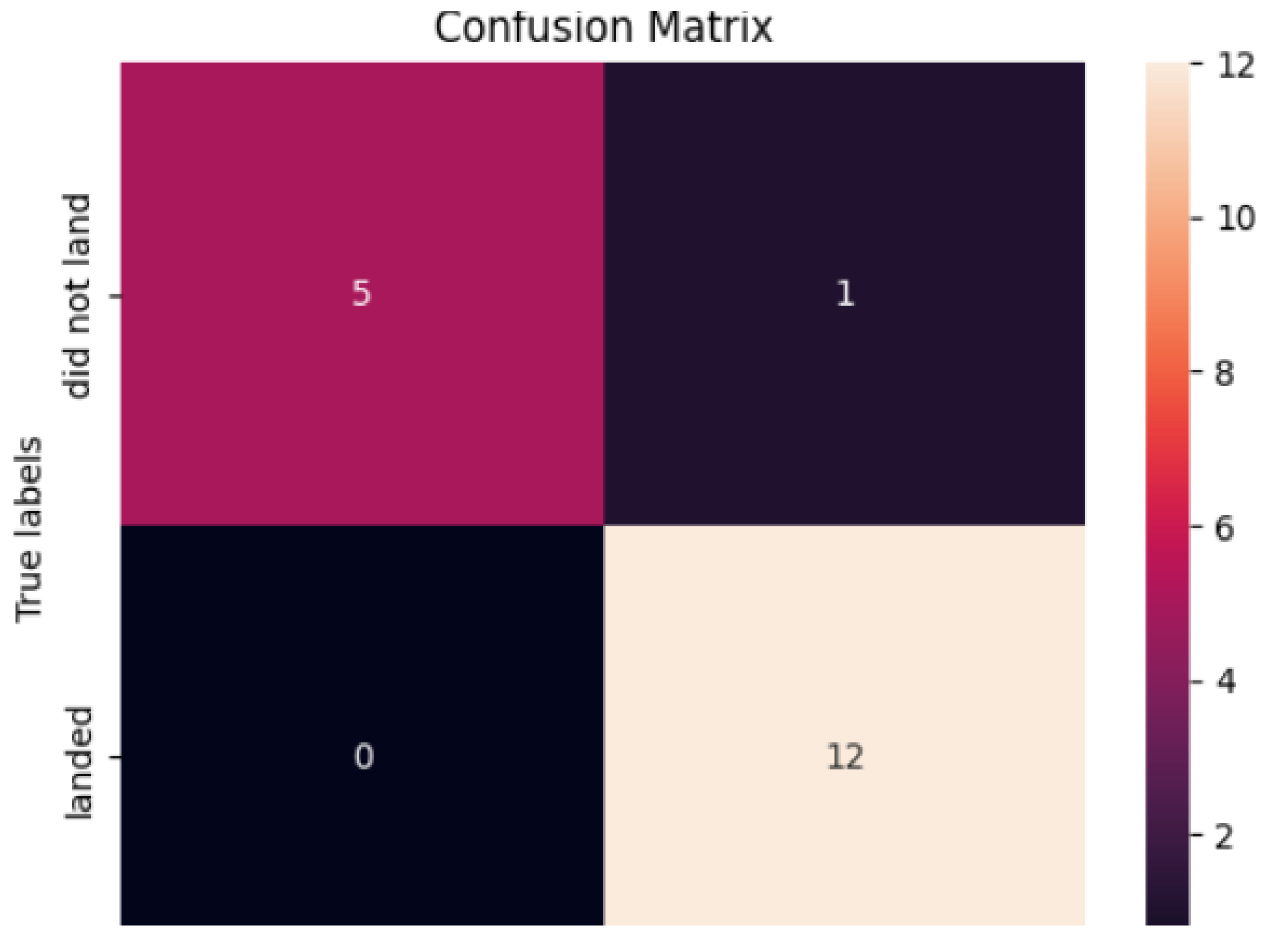
- Confusion Matrix of Best Performing Model (Logistic Regression)

Confusion matrix values:

- True Positives (land predicted as land): 12
- True Negatives (did not land predicted as did not land): 5
- False Positives (did not land predicted as land): 1
- False Negatives (land predicted as did not land): 0

Findings:

- The model shows high precision and recall, especially for successful landings.
- Only one misclassification occurred, indicating strong generalization.
- This validates Logistic Regression as the most reliable classifier in your pipeline.



Conclusions

- This project aimed to predict SpaceX launch success using historical mission data, combining geospatial insights, payload dynamics, and booster performance.
- Through Data Wrangling, EDA with SQL queries, and Visualization with Folium maps, and interactive dashboards, we uncovered patterns in payload mass, landing outcomes, and booster reliability.
- Logistic Regression emerged as the most accurate classification model, achieving a test accuracy of 94.4% with minimal misclassifications.
- Geospatial analysis revealed strategic launch site placements and proximity to infrastructure such as near coastlines, supporting operational efficiency.
- The dashboard visualizations provided clear, audience-ready insights into launch trends, site performance, and model predictions.

Next Steps:

- Integrate real-time launch data via APIs for dynamic dashboard updates.
- Explore ensemble models or hyperparameter tuning to further improve prediction accuracy.

Appendix

- This section includes sources that supported the main analysis.

Raw Data Sources:

- SpaceX launch records from public datasets and mission logs - Link
- Payload and booster metadata extracted via SQL queries from my_data1.db

Resources & Tools Used:

- Python (pandas, scikit-learn, Folium, Plotly Dash)
- SQLite for structured querying
- Jupyter Notebook for interactive development

Acknowledgements:

- SpaceX for publicly available launch data
- OpenStreetMap contributors for geospatial mapping

References:

- Documentation from scikit-learn, Folium, and Plotly
- SQL syntax guides and geospatial visualization tutorials

Thank you!