IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andres Melendez
24 December 2025

Outline

Executive Summary

- This project analyzes SpaceX Falcon 9 launch data to identify factors that influence successful first-stage booster landings.
- Data was collected using the SpaceX REST API and web scraping, then cleaned and processed through data wrangling techniques.
- Exploratory data analysis using visualizations and SQL revealed strong relationships between launch site, orbit type, payload mass, and landing success.
- Interactive analytics were built using Folium maps and a Plotly Dash dashboard to visualize launch locations, success rates, and payload trends.
- Multiple classification models were trained and evaluated, with the best-performing model achieving strong predictive accuracy for landing outcomes.
- Results provide data-driven insights that can help predict mission success and support cost-efficient rocket reusability decisions.

Introduction

- SpaceX has significantly reduced the cost of space travel by developing reusable Falcon 9 rocket boosters, making landing success a critical factor in mission efficiency.
- Understanding the conditions that lead to successful first-stage landings can help improve mission planning, reduce launch costs, and increase overall reliability.
- This project uses historical SpaceX launch data to explore how variables such as launch site, orbit type, payload mass, and booster configuration impact landing outcomes.
- The primary problem addressed in this analysis is determining which factors most strongly influence whether a Falcon 9 first stage successfully lands.
- Additionally, this project aims to build a predictive classification model capable of estimating landing success prior to launch.

Section 1

Methodology

Methodology

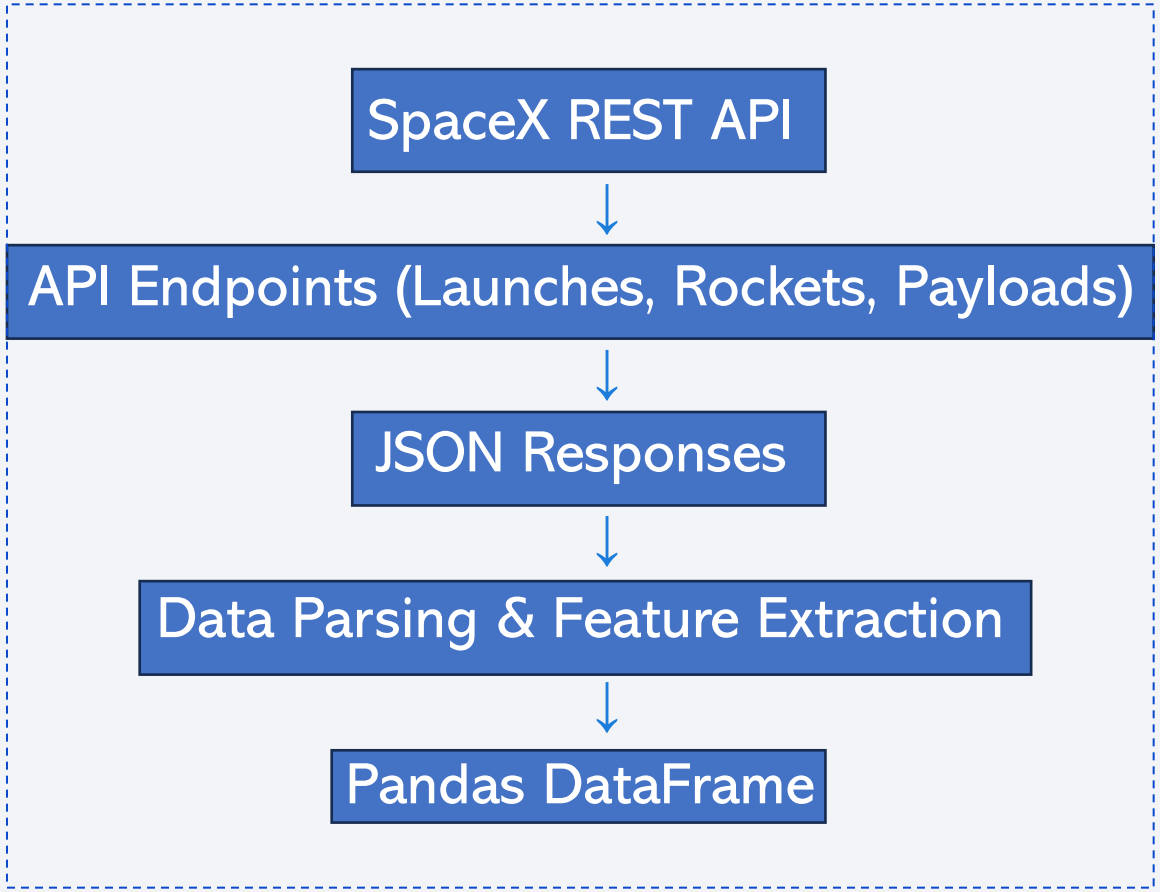
- **Executive Summary**
- **Data Collection Methodology**
  - Launch data was collected using the SpaceX REST API and supplemented with web-scraped historical launch records.
- **Data Wrangling**
  - Raw datasets were cleaned, merged, and transformed by handling missing values, encoding categorical variables, and selecting relevant features.
- **Exploratory Data Analysis (EDA)**
  - Data was analyzed using statistical summaries, visualizations, and SQL queries to identify patterns and relationships affecting landing success.
- **Interactive Visual Analytics**
  - Folium maps were used to visualize launch locations and outcomes, while a Plotly Dash dashboard enabled interactive exploration of payload and success trends.
- **Predictive Analysis**
  - Multiple classification models were built, tuned, and evaluated to predict first-stage landing success based on mission characteristics.

Data Collection

- Multiple datasets were collected to support the analysis of SpaceX Falcon 9 launch and landing outcomes.
- Primary launch data was retrieved programmatically using the SpaceX REST API, providing structured information on missions, rockets, payloads, launch sites, and landing results.
- Additional historical launch details were gathered through web scraping from publicly available SpaceX-related sources.
- The data collection process followed a structured pipeline to ensure accuracy, completeness, and consistency across sources.
- Collected datasets were stored in tabular format and prepared for downstream data wrangling and analysis.

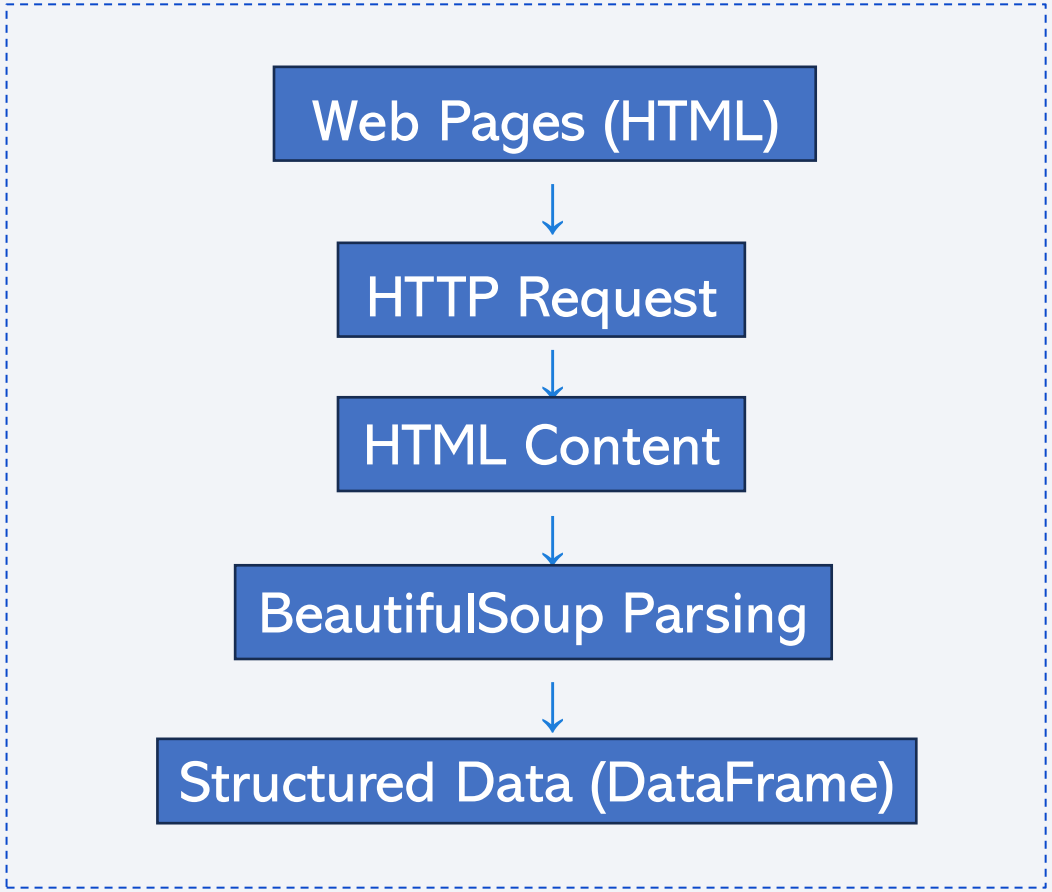
Data Collection – SpaceX API

- Launch data was collected programmatically using the **SpaceX REST API**, which provides structured mission and rocket information.
- API endpoints were used to retrieve data related to launches, rockets, payloads, launch sites, and landing outcomes.
- JSON responses were parsed and transformed into structured Pandas DataFrames for analysis.
- Relevant features such as flight number, payload mass, orbit type, launch site, and landing success were extracted.
- The finalized dataset was saved for further data wrangling, exploratory analysis, and modeling.
- **GitHub Repository (SpaceX API Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-data-collection-api.ipynb



Data Collection - Scraping

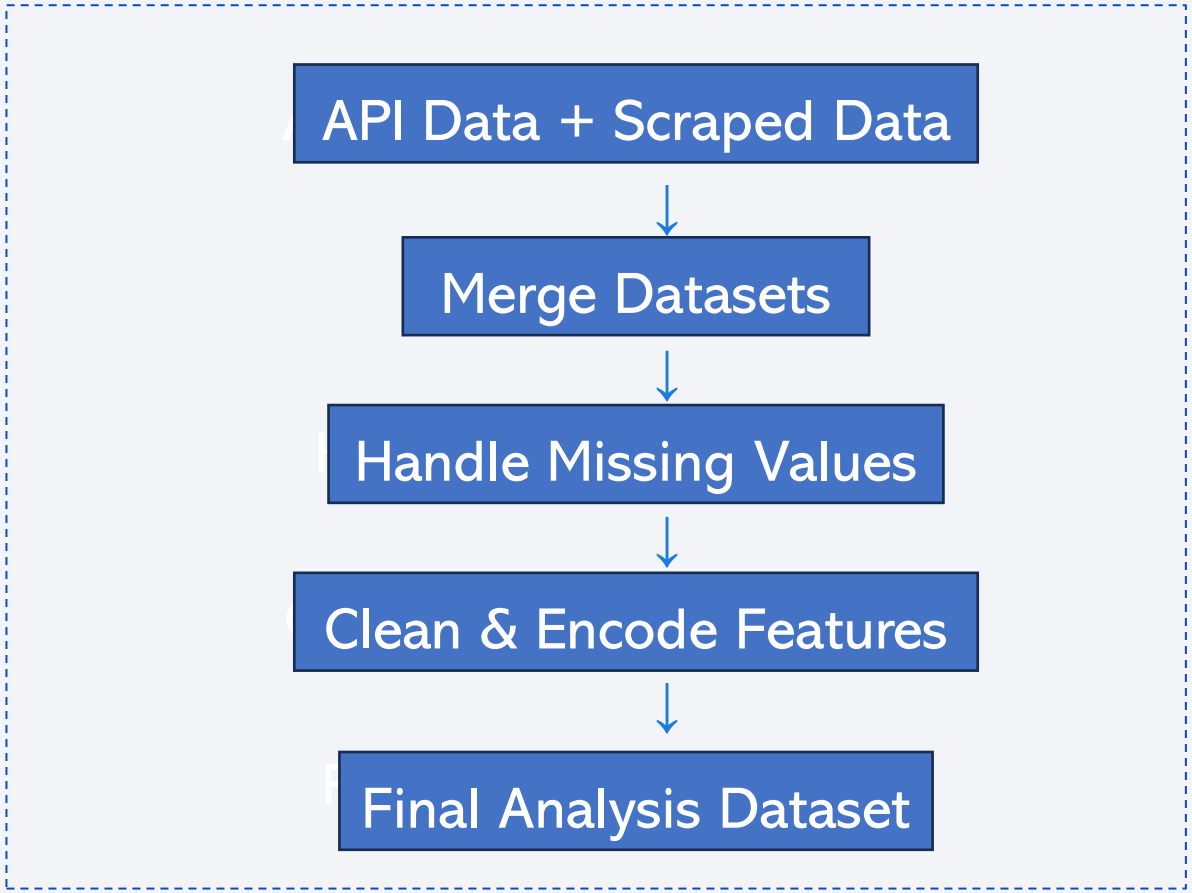
- Web scraping was used to collect additional historical SpaceX launch data not available through the SpaceX REST API.
- HTML content was retrieved from publicly available web pages containing launch records.
- The BeautifulSoup library was used to parse HTML tables and extract relevant launch information.
- Extracted data was cleaned, structured, and converted into Pandas DataFrames.
- The resulting dataset was merged with API-collected data to support comprehensive analysis.
- **GitHub Repository (Web Scraping Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-webscraping.ipynb



Data Wrangling

- Describe Data collected from the SpaceX REST API and web scraping sources was combined into a single unified dataset.
- Missing values were identified and handled appropriately to ensure data completeness and reliability.
- Categorical variables such as orbit type, launch site, and landing outcome were cleaned and encoded for analysis.
- Numerical features including payload mass and flight number were standardized and validated.
- The final dataset was structured to support exploratory data analysis, interactive visualization, and predictive modeling.
- **GitHub Repository (Data Wrangling Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-Data%20wrangling.ipynb

Data Wrangling



EDA with Data Visualization

- Scatter plots were used to examine relationships between numerical and categorical variables, such as flight number, payload mass, launch site, orbit type, and landing outcome.
- Bar charts were created to compare launch success rates across different orbit types and launch sites.
- Line charts were used to analyze trends in launch success over time and observe performance improvements across years.
- These visualizations helped identify patterns, correlations, and potential predictors of successful first-stage landings.
- The insights gained from visual exploration informed feature selection for predictive modeling.
- **GitHub Repository (EDA with Data Visualization Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-eda.ipynb

EDA with SQL

- Queried the dataset to identify all unique SpaceX launch sites and analyze their distribution.
- Filtered launch records to find specific launch sites based on name patterns (e.g., sites beginning with “CCA”).
- Calculated total and average payload masses for different booster versions and mission types.
- Identified the dates and conditions of the first successful booster landings.
- Analyzed successful and failed landing outcomes across different launch sites and years.
- Retrieved records for missions with specific payload mass ranges and landing outcomes.
- Ranked landing outcomes over a defined time period to compare mission success rates.
- **GitHub Repository (EDA with SQL Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-eda-sql.ipynb

Build an Interactive Map with Folium

- Location markers were added to display the geographic positions of SpaceX launch sites across the globe.
- Color-coded markers were used to distinguish between successful and failed first-stage landing outcomes.
- Circles were added to visualize the relative proximity of launch sites to important infrastructure such as coastlines, railways, and highways.
- Distance lines and radius indicators were used to calculate and display distances between launch sites and nearby features.
- These map objects help visually assess how location and surrounding infrastructure may influence launch and landing success.
- **GitHub Repository (Folium Map Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-Folium.ipynb

Build a Dashboard with Plotly Dash

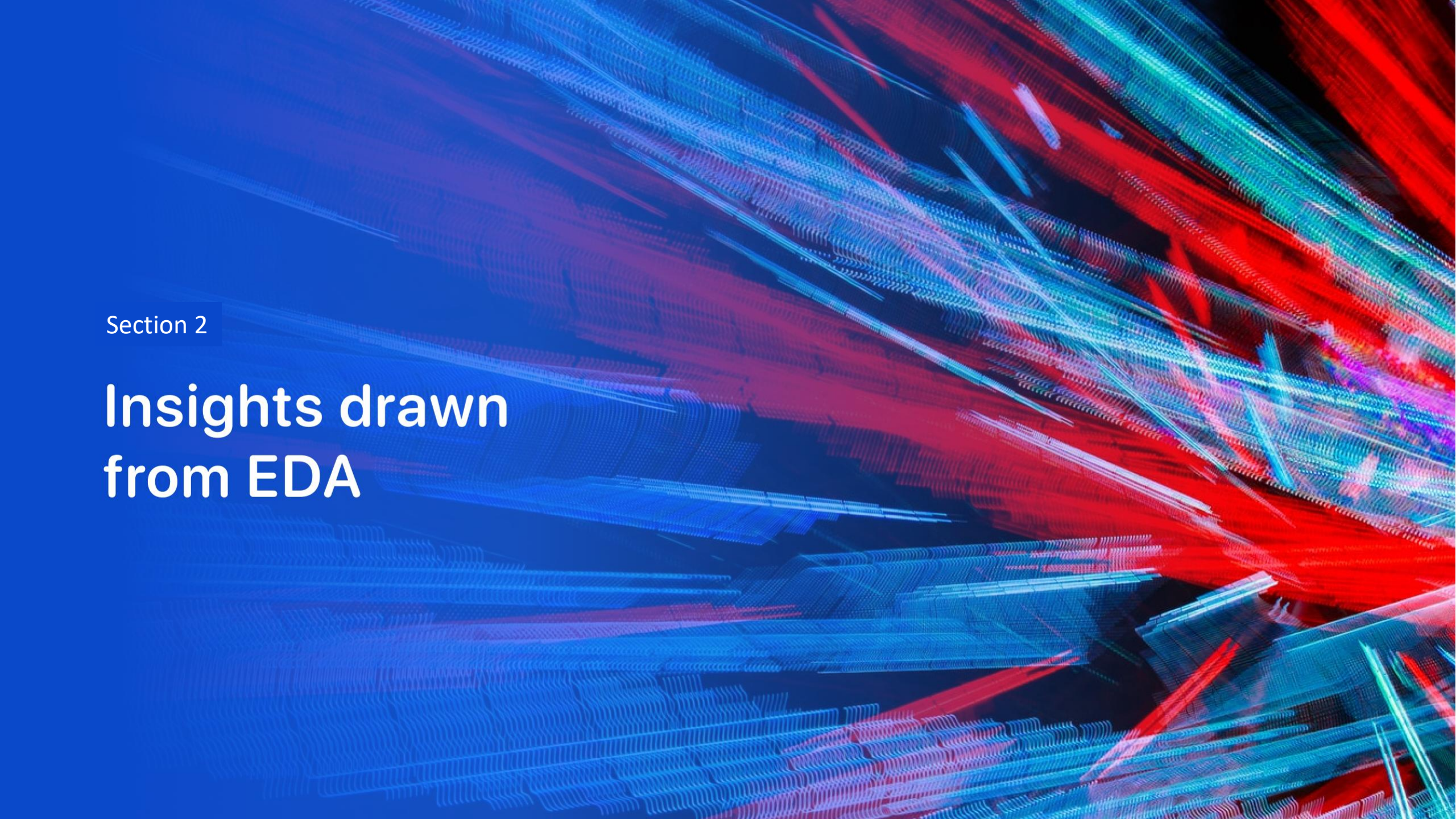
- Interactive pie charts were created to display launch success counts and success ratios across different launch sites.
- A scatter plot was added to visualize the relationship between payload mass and launch outcome.
- User controls such as dropdown menus and range sliders were implemented to allow dynamic filtering by launch site and payload range.
- These interactive elements enable users to explore how payload mass and launch location impact landing success.
- The dashboard provides an intuitive way to analyze complex relationships that are not easily observed in static charts.
- **GitHub Repository (Plotly Dash Lab):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-Plotly.ipynb

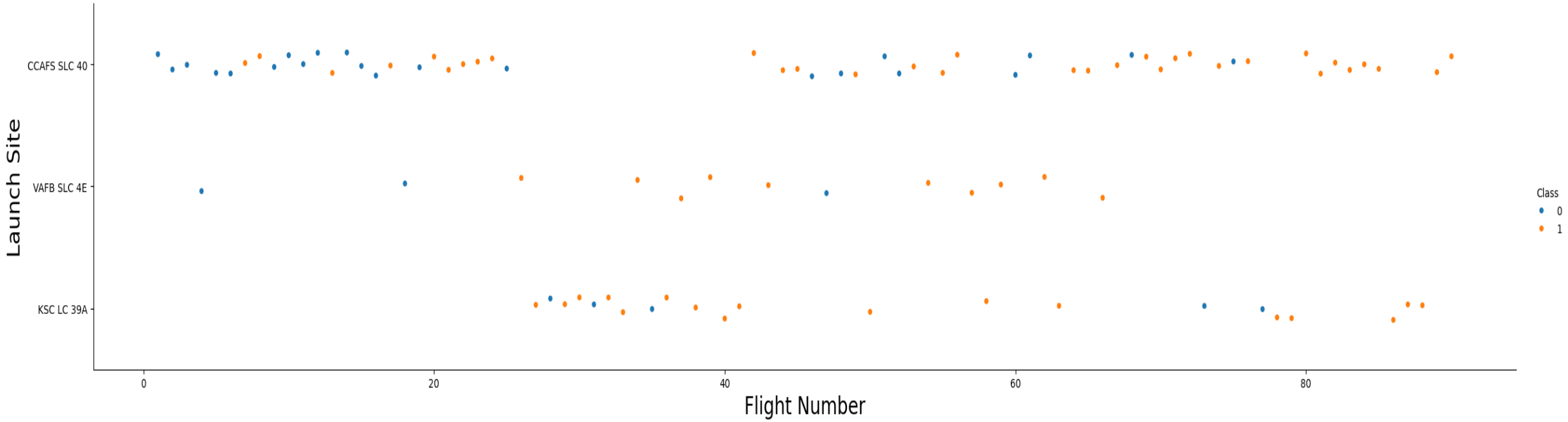
Predictive Analysis (Classification)

- A supervised classification approach was used to predict first-stage booster landing success.
- Relevant features such as payload mass, orbit type, launch site, flight number, and booster configuration were selected from the processed dataset.
- The dataset was split into training and testing sets to evaluate model generalization.
- Multiple classification models were built and compared, including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors.
- Hyperparameter tuning and cross-validation were applied to improve model performance.
- Models were evaluated using accuracy and confusion matrices to identify the best-performing classifier.
- **GitHub Repository (Predictive Analysis Notebook):** https://github.com/Dre2322/Coursera/blob/main/IBM%20Data%20Science/Course10/Course10-ML-Prediction.ipynb

Results

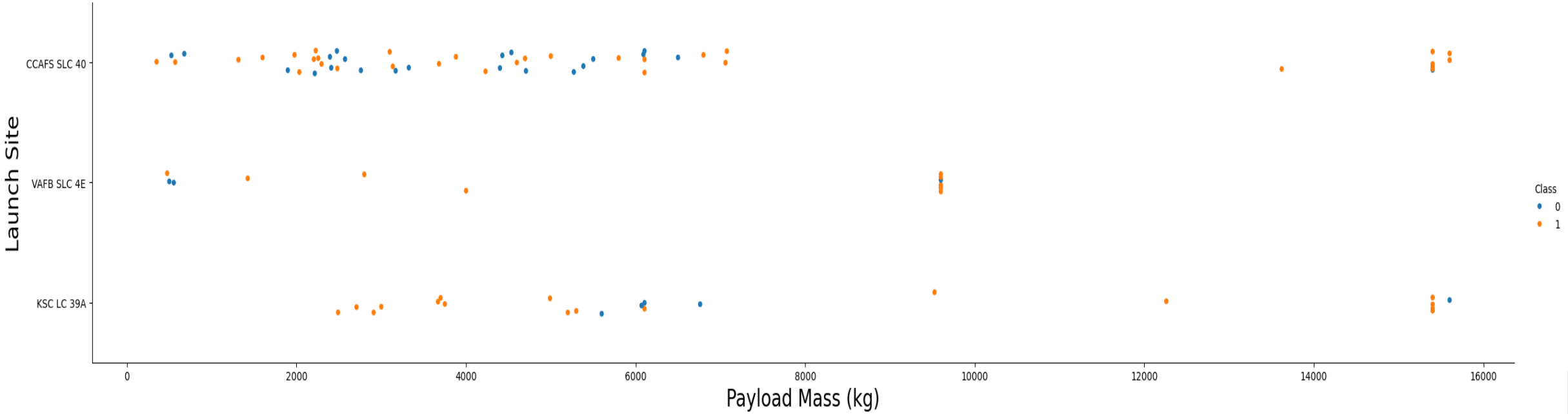
- Exploratory data analysis results highlighting relationships between launch characteristics and landing success.
- Interactive analytics results demonstrated through screenshots of Folium maps and the Plotly Dash dashboard.
- Predictive analysis results comparing classification model performance and identifying the best-performing model.


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- The scatter plot shows how launch frequency and landing success vary across different SpaceX launch sites as flight numbers increase.
- Early missions show more failures across all sites, while later flights demonstrate a higher concentration of successful landings.
- This trend suggests that operational experience and launch site maturity contribute positively to first-stage landing success.

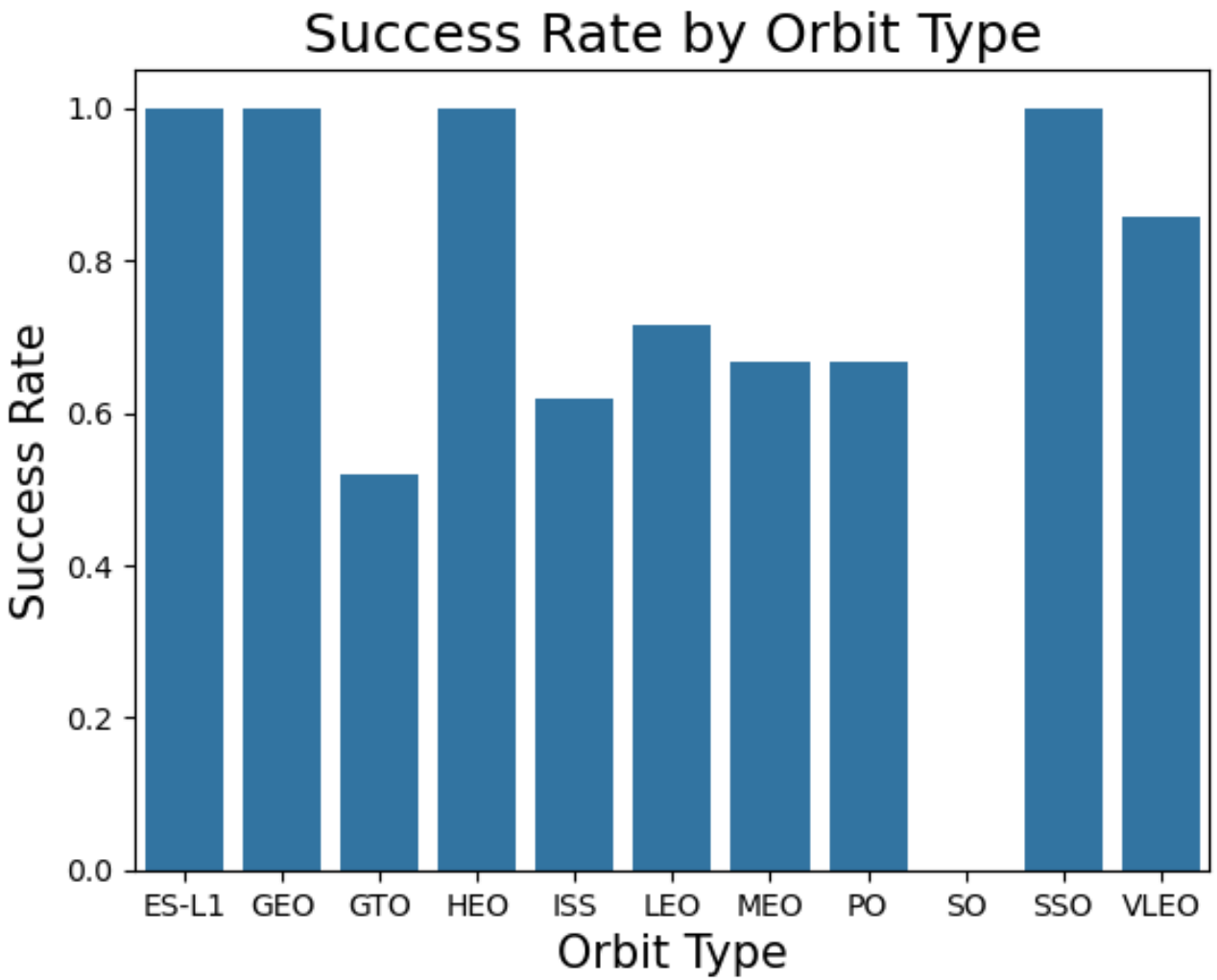


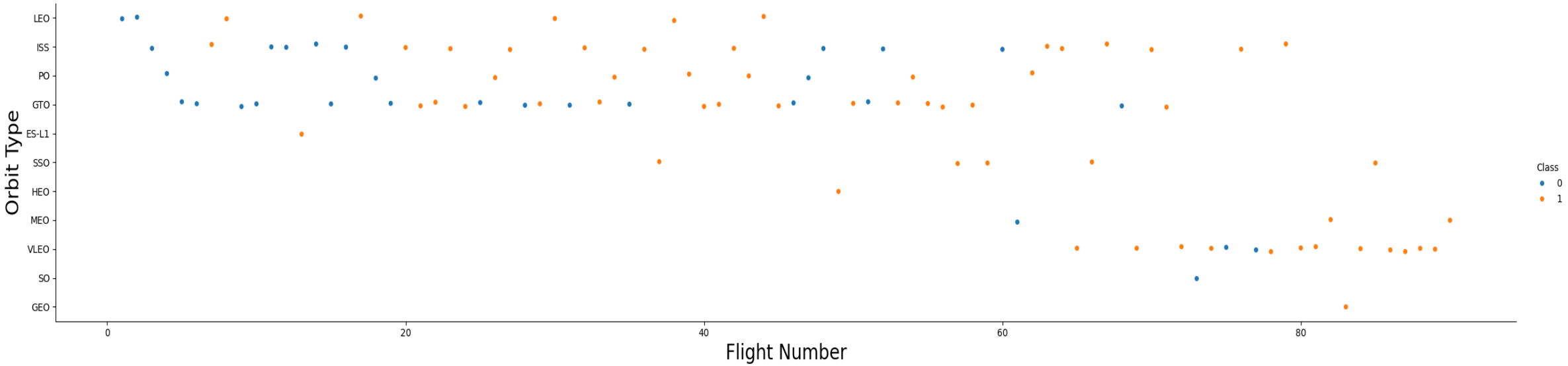
Payload vs. Launch Site

- The scatter plot illustrates the relationship between payload mass and launch site, with landing outcomes shown for each mission.
- Higher payload masses tend to be associated with fewer successful landings, particularly at certain launch sites.
- This suggests that payload mass is an important factor influencing first-stage landing success and should be considered in predictive modeling.

Success Rate vs. Orbit Type

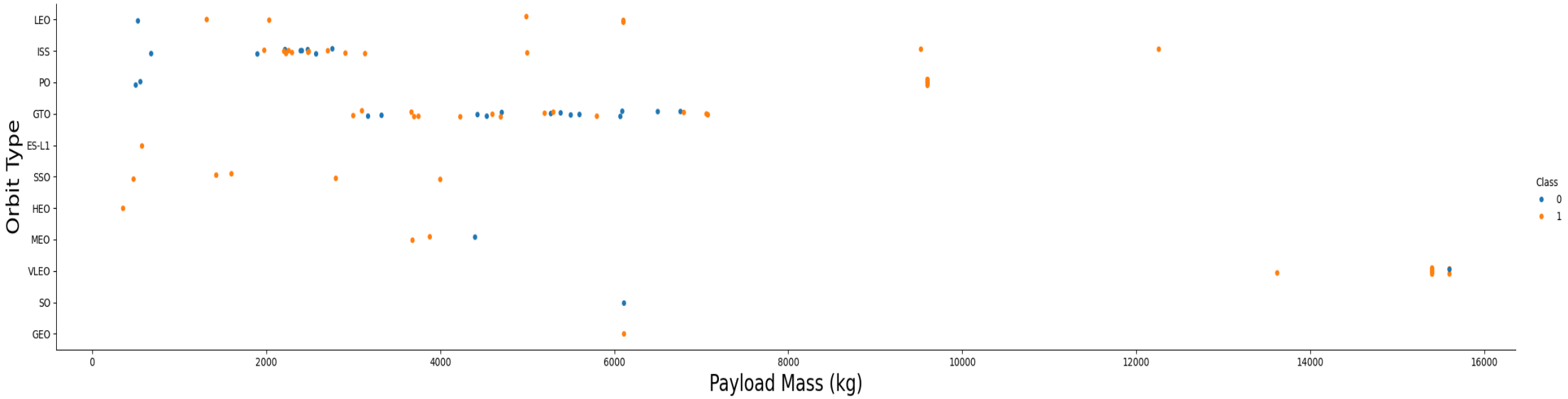
- The bar chart compares first-stage landing success rates across different orbit types.
- Certain orbit types show significantly higher success rates, indicating that mission profile and orbital requirements influence landing outcomes.
- This suggests that orbit type is an important categorical feature and should be included in predictive modeling.





Flight Number vs. Orbit Type

- The scatter plot shows how launch missions are distributed across different orbit types as flight numbers increase.
- Later flights exhibit a higher concentration of successful landings across multiple orbit types compared to early missions.
- This pattern suggests that increasing operational experience improves landing success regardless of orbit complexity.

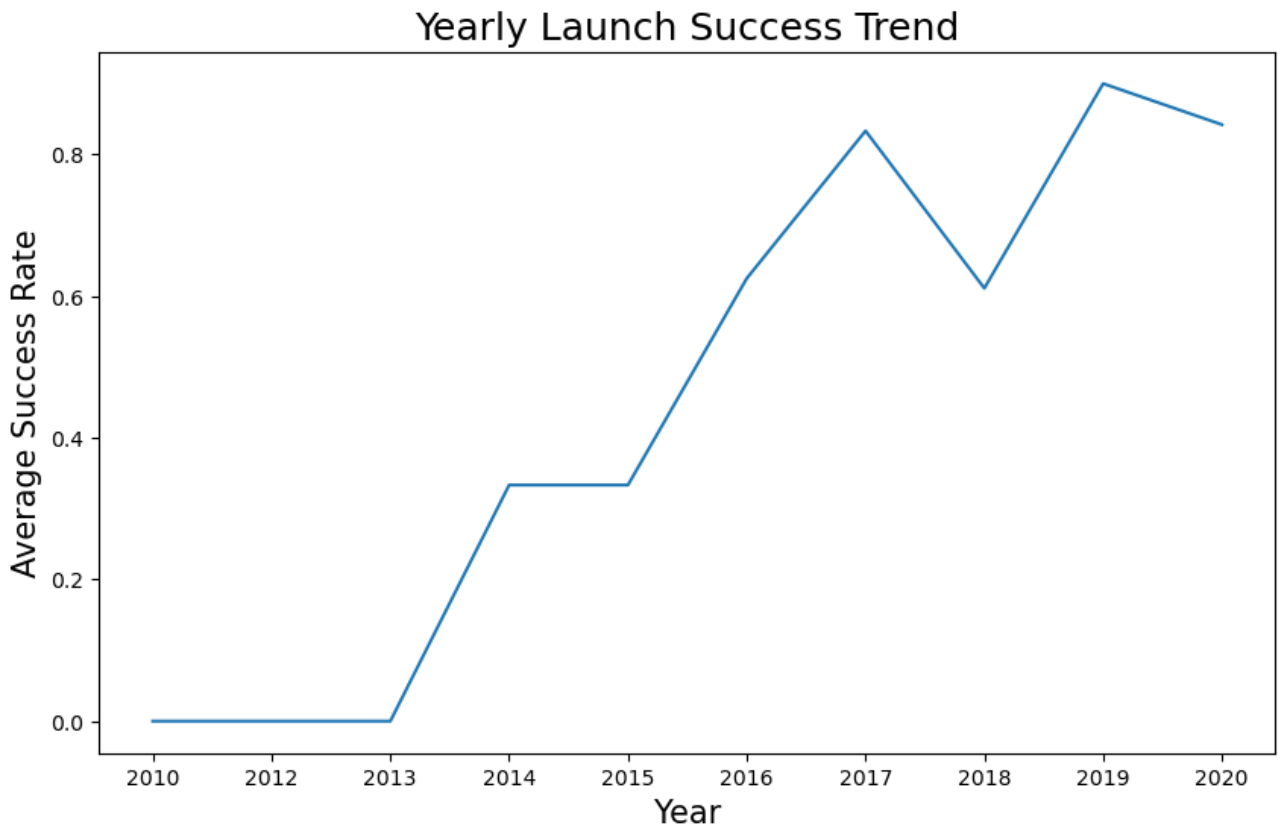


Payload vs. Orbit Type

- The scatter plot shows how payload mass varies across different orbit types and how it relates to first-stage landing outcomes.
- Higher payload masses are more frequently associated with lower landing success, particularly for more demanding orbit types.
- This indicates that both payload mass and orbit type jointly influence landing difficulty and are important features for predictive modeling.

Launch Success Yearly Trend

- The line chart shows the yearly average first-stage landing success rate over time.
- Landing success rates increase steadily in later years, reflecting improvements in rocket design, recovery techniques, and operational experience.
- This upward trend highlights SpaceX’s progress toward reliable booster reusability and supports the feasibility of predicting mission success.



All Launch Site Names

- This query retrieves the distinct launch site names used in the dataset.
  The results show that SpaceX primarily launched missions from three main locations: CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E, with minor naming variations present in the raw data.

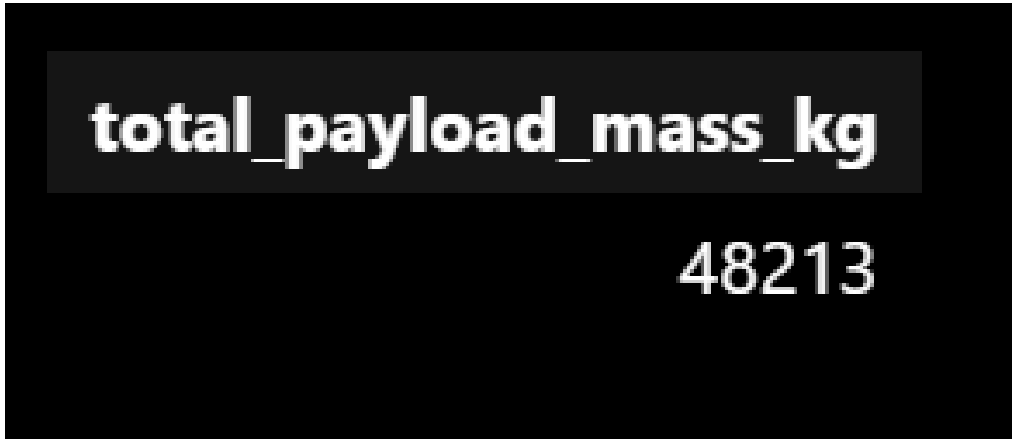
Launch Site Names Begin with 'CCA'

- This query retrieves five launch records where the launch site name begins with “CCA.”
  The results confirm that Cape Canaveral Air Force Station (CCAFS) is a frequently used launch site for early SpaceX missions.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

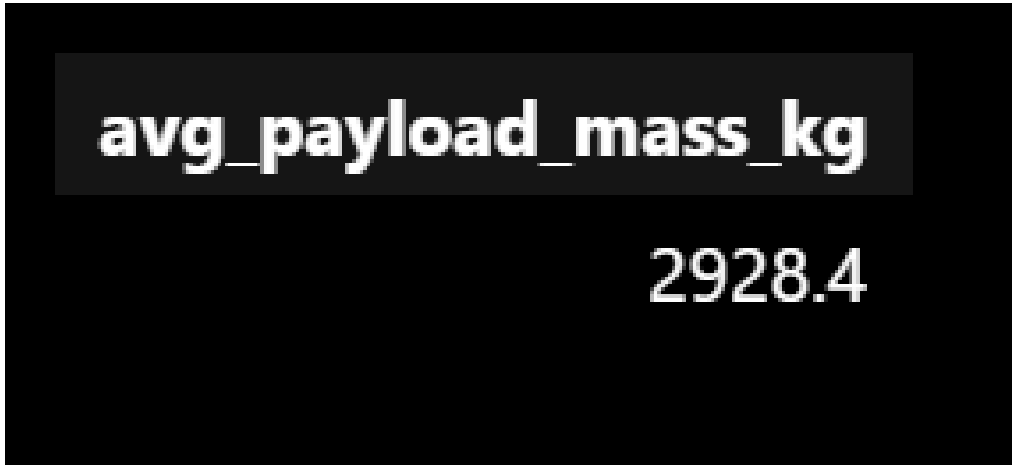
Total Payload Mass

- This query calculates the total payload mass carried by SpaceX boosters for NASA missions.
  The result shows that SpaceX has delivered a cumulative payload mass of 48,213 kg for NASA-contracted launches.



Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried by the Falcon 9 v1.1 booster version.
  The result shows that F9 v1.1 missions carried an average payload of approximately 2,928 kg, reflecting the typical capacity of this booster configuration.



First Successful Landing Outcome Date

- This query identifies the date of the first successful first-stage landing achieved by SpaceX.
  The result shows that the first successful landing occurred on December 22, 2015, marking a major milestone in SpaceX's reusable launch vehicle development.

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query identifies booster versions that successfully landed on a drone ship while carrying payloads between 4,000 and 6,000 kg.
  The results demonstrate that multiple Falcon 9 Full Thrust boosters achieved reliable recovery performance under moderate payload conditions.

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

Total Number of Successful and Failure Mission Outcomes

- This query summarizes the total number of mission outcomes recorded in the dataset. The results show that SpaceX missions achieved a high success rate, with 100 successful missions and only one in-flight failure, demonstrating strong overall mission reliability.

Mission_Result	Total_Count
Failure (in flight)	1
Success	100

Boosters Carried Maximum Payload

- This query identifies the booster versions that carried the maximum payload mass recorded in the dataset.
  The results show that multiple Falcon 9 Block 5 boosters each delivered the maximum payload of 15,600 kg, demonstrating SpaceX's ability to consistently support heavy payload missions using its most advanced booster configuration.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- This query lists missions from 2015 that experienced failed drone-ship landing outcomes, along with the corresponding booster versions and launch sites.
  The results reflect early recovery attempts during SpaceX's booster reusability development phase, which contributed to improved landing success in subsequent years.

Date	Booster_Version	Launch_Site	Landing_Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- This query ranks landing outcomes between June 4, 2010, and March 20, 2017, based on their frequency.
  The results show that early SpaceX missions primarily involved no landing attempts or unsuccessful recovery efforts. In contrast, successful drone ship and ground pad landings became more common as landing technology matured.

Landing_Outcome	Total_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

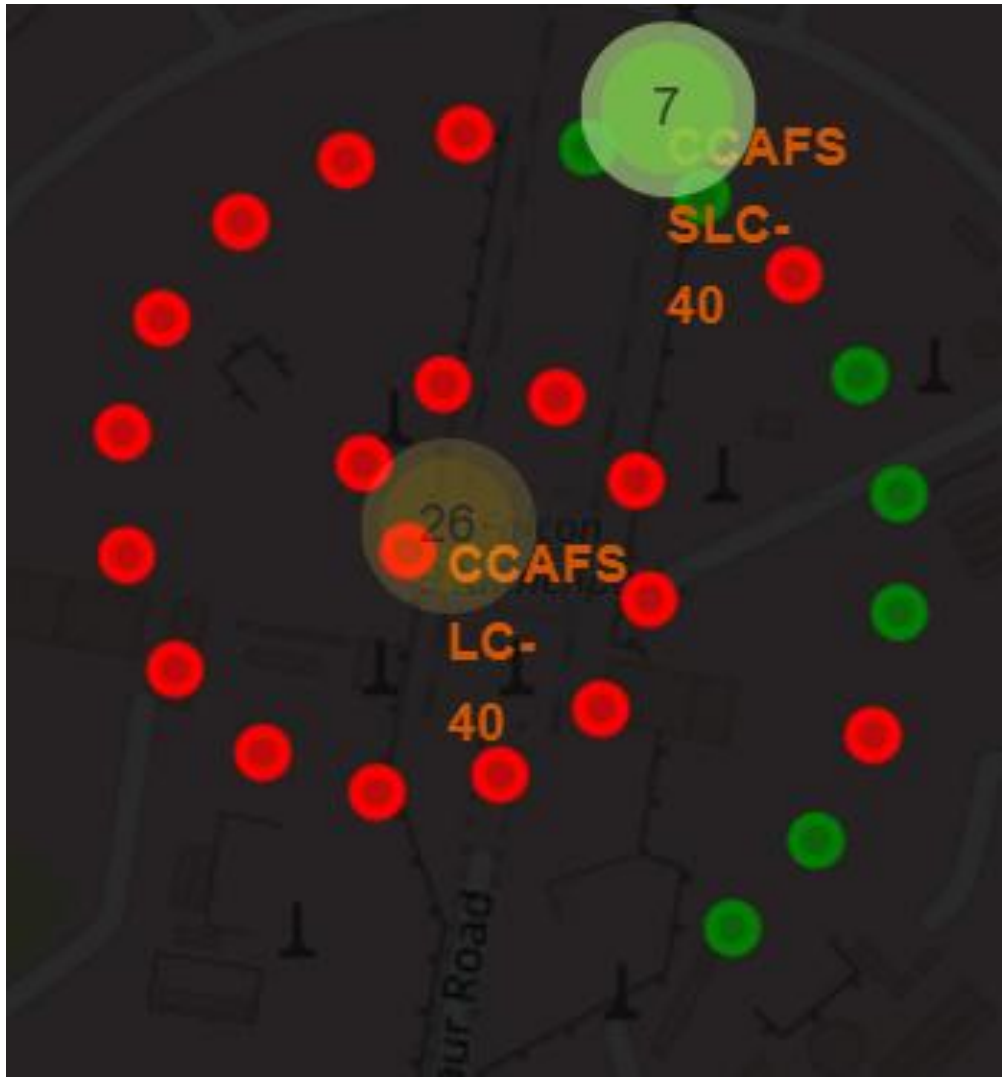
Launch Sites Proximities Analysis

Global Launch Site Locations (Folium Map)

- This Folium map displays the global geographic locations of all SpaceX launch sites using interactive markers. Although SpaceX launch operations are currently concentrated in the United States, the global map view provides geographic context relative to other continents and oceans.
  The locations in California and Florida support different orbital inclinations and mission profiles, highlighting how launch site geography influences orbital access and mission planning.

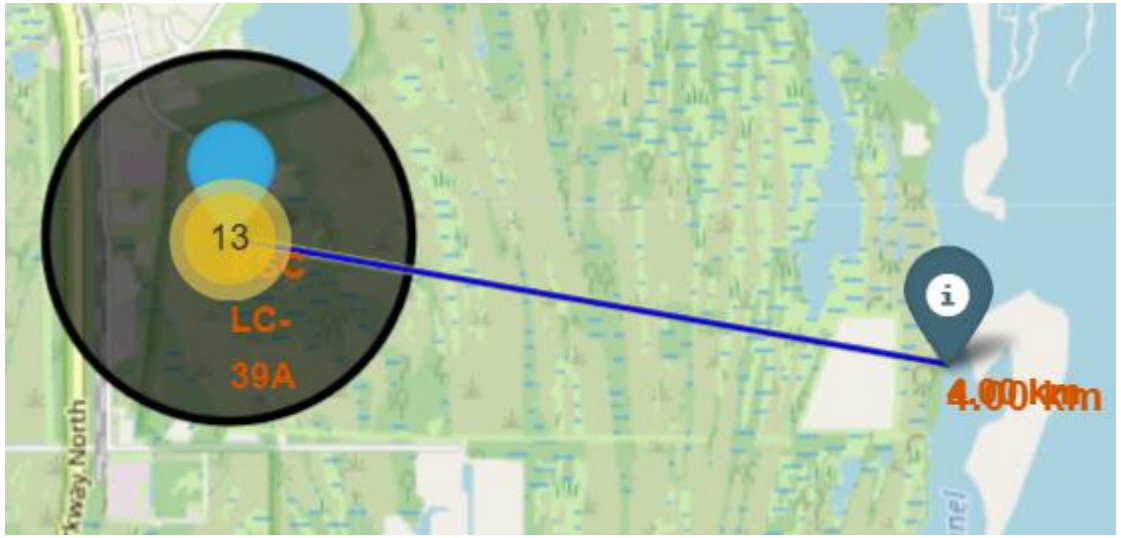
Color-Coded Launch Landing Outcomes by Site

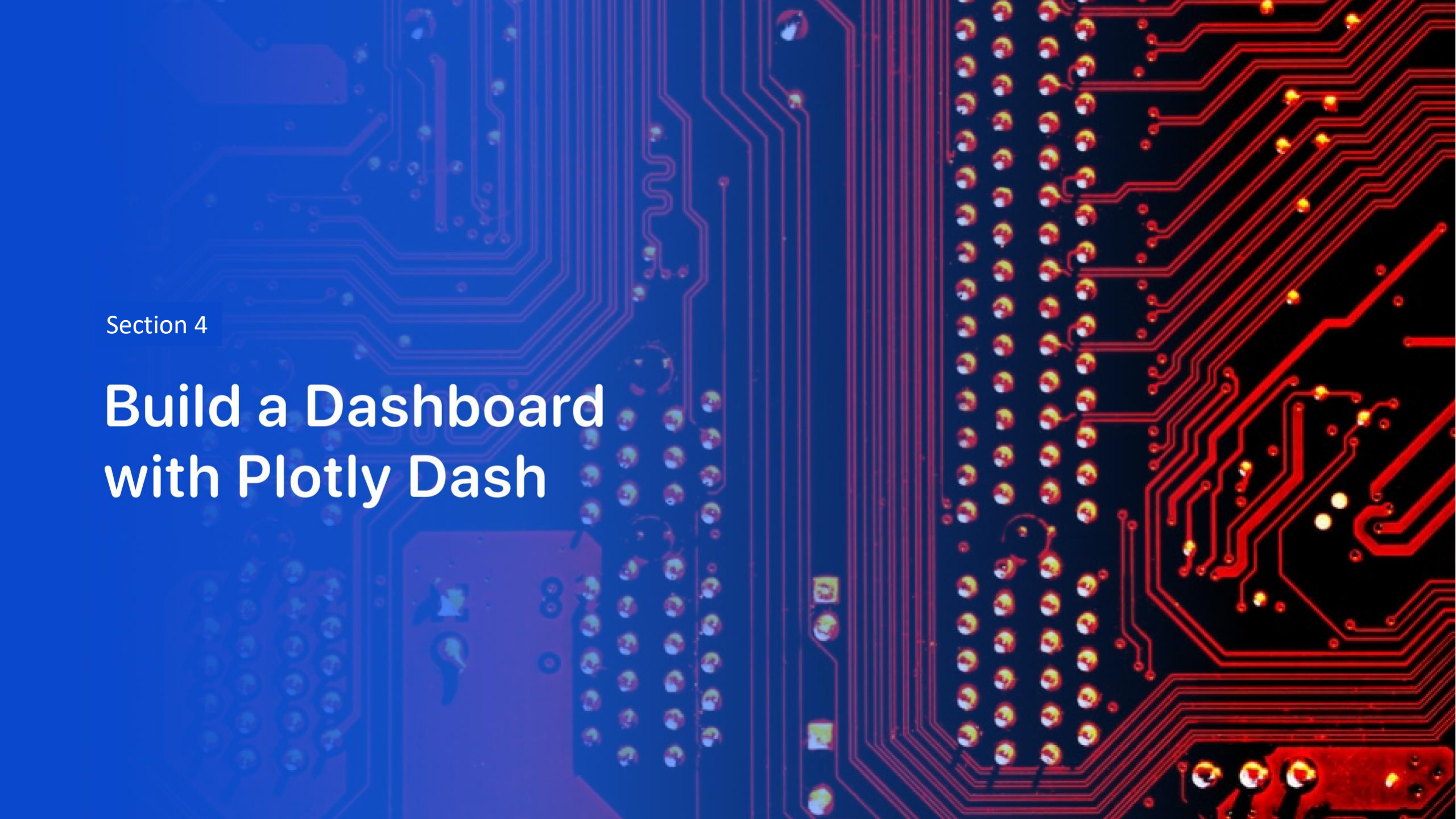
- This map displays individual SpaceX launches with color-coded markers representing landing outcomes.
  Green markers indicate successful landings, while red markers represent failed attempts.
  The visualization reveals clusters of early failures near launch sites and a growing concentration of successful landings over time, particularly at Cape Canaveral (CCAFS), highlighting operational improvement and increased landing reliability.



Proximity Analysis: KSC LC-39A to Nearest Coastline

- This map illustrates a proximity analysis between the KSC LC-39A launch site and the nearest coastline. Using latitude and longitude coordinates, the great-circle distance was calculated with the Haversine formula and visualized with a connecting line and distance label. The short distance (approximately 4 km) highlights the strategic advantage of coastal launch sites, particularly in terms of safety considerations, recovery operations, and launch logistics.




Section 4
Build a Dashboard with Plotly Dash

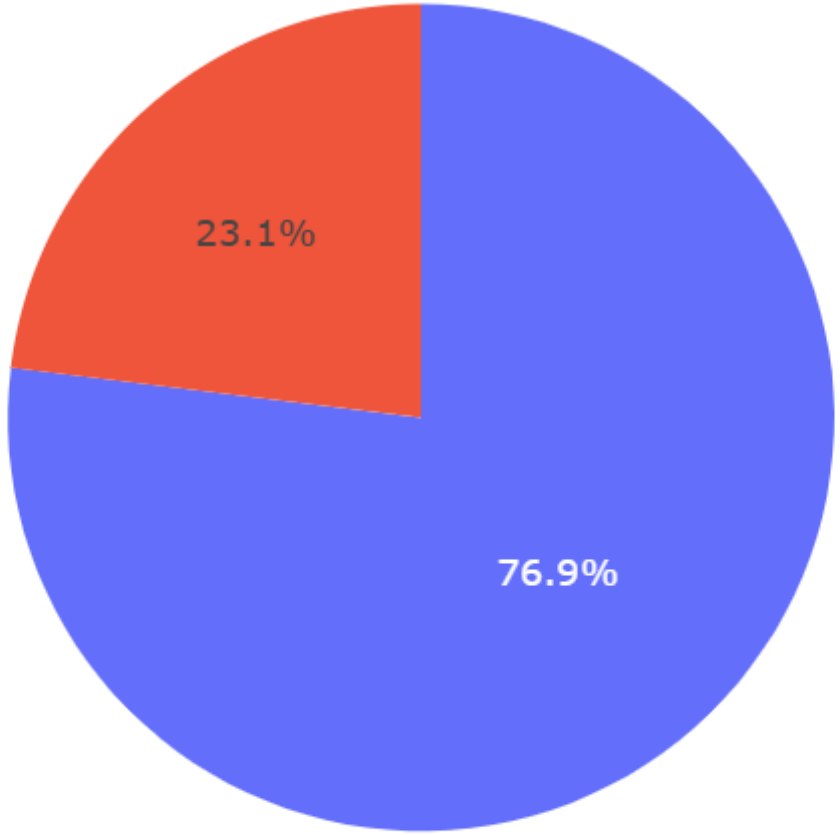
Total Successful Launches by Site

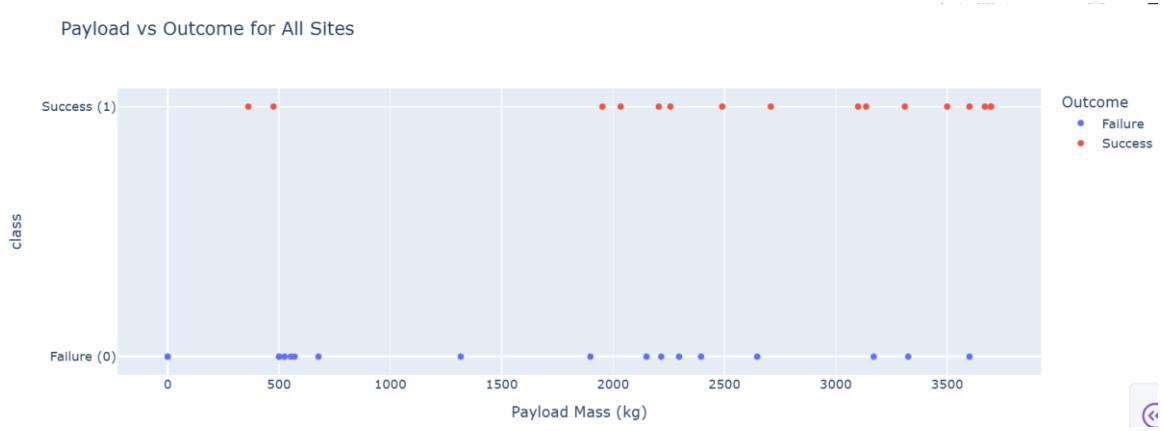
Launch Success Distribution by Launch Site

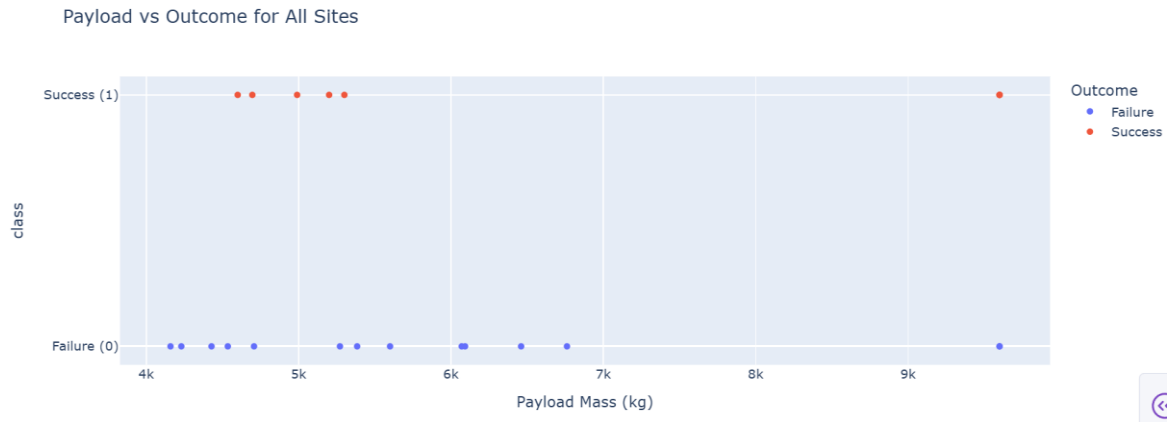
- The pie chart illustrates the distribution of successful SpaceX launches across all launch sites.
- Kennedy Space Center (LC-39A) accounts for the largest share of successful launches, indicating its central role in later missions.
- Cape Canaveral (CCAFS LC-40) and Vandenberg (SLC-4E) contribute smaller but significant portions, reflecting differences in mission frequency and orbital targets.

Launch Outcome Distribution for KSC LC-39A

- This pie chart illustrates the proportion of successful and failed launches at KSC LC-39A, the launch site with the highest success ratio.
- Approximately **76.9%** of launches from this site were successful, indicating strong operational reliability.
- The high success rate suggests that KSC LC-39A benefits from improved launch infrastructure and later-stage mission maturity.







Payload Mass vs Launch Outcome Across All Sites

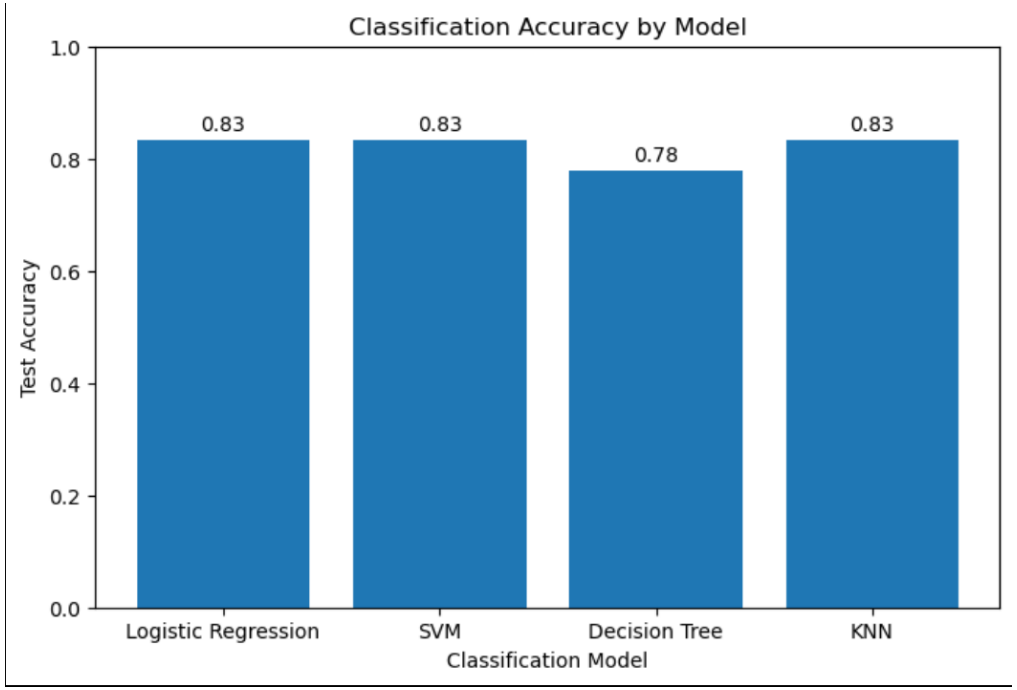
- The scatter plots show the relationship between **payload mass** and **launch outcome**.
- Lower payload masses generally show a higher concentration of successful launches.
- As payload mass increases, failures become more frequent, indicating increased mission complexity.
- This suggests payload mass is an important factor in predicting launch success.

Section 5

Predictive Analysis (Classification)

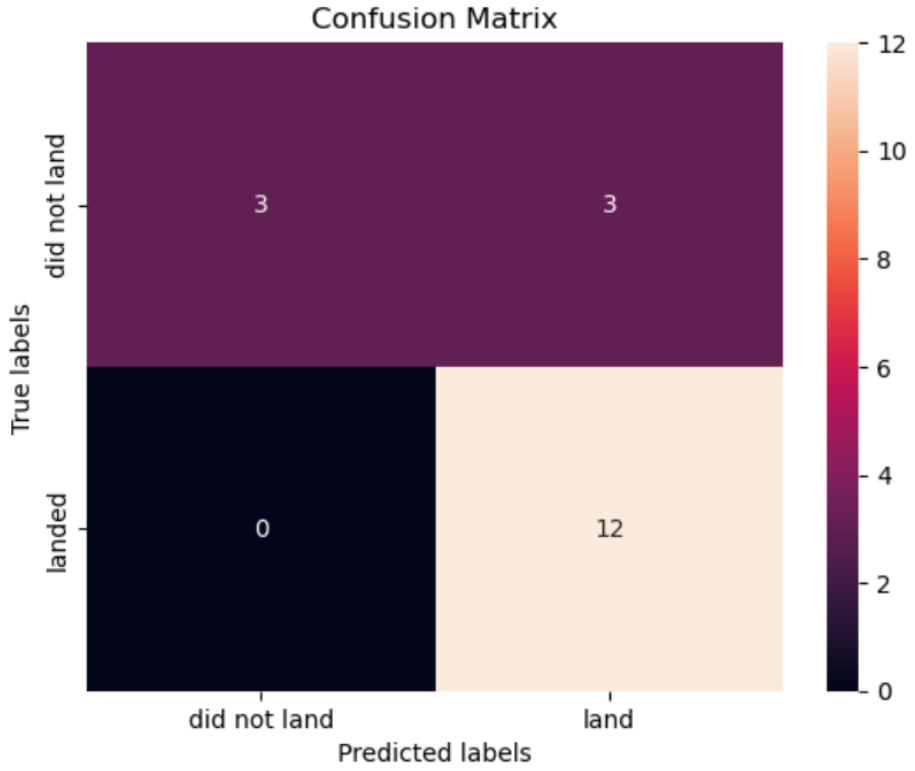
Classification Accuracy

- The bar chart shows the test accuracy of four classification models.
  Logistic Regression, SVM, and KNN achieved the highest accuracy at approximately 83.3%, while the Decision Tree model performed slightly worse at 77.8%.
  Therefore, Logistic Regression, SVM, and KNN are tied as the best-performing models for this dataset.



Confusion Matrix

- The confusion matrix shows that the Logistic Regression model correctly predicts most landing outcomes.
- The model achieved 12 true positives and 3 true negatives, with no false negatives, indicating it successfully identifies all actual successful landings. Although there are a small number of false positives, the absence of false negatives highlights the model's strong ability to detect successful landings, supporting its high overall classification accuracy.
- The model prioritizes recall for successful landings, ensuring no successful launches are misclassified as failures.



Conclusions

- Launch success rates increased over time, reflecting operational learning and process optimization.
- Payload mass and orbit type are key factors affecting landing success.
- Launch sites are geographically optimized for recovery operations.
- Machine learning models achieved high classification accuracy, particularly in identifying successful landings.
- An end-to-end data science workflow successfully explains historical outcomes and predicts landing success.

Appendix

- **Python notebooks** for data preprocessing, EDA, and machine learning models
- **SQL queries** used for payload, launch site, and mission outcome analysis
- **Model evaluation outputs**, including accuracy scores and confusion matrix
- **Folium map visualizations** showing launch sites and spatial analysis
- **Supporting tables and figures** referenced throughout the presentation

Thank you!