IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alberto Correa Peña
15-16-2025

Outline

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

Executive Summary

- **Methodologies Used**
  - **Data Collection**: Acquired public data through web scraping and the SpaceX API.
  - **Exploratory Data Analysis (EDA)**: Conducted data wrangling, visualization, and interactive visual analytics to understand the data structure and distributions.
  - **Machine Learning Modeling**: Built and evaluated predictive models to identify factors influencing launch success.

- **Key Findings**
  - Successfully gathered valuable information from public sources.
  - EDA enabled the identification of the most relevant features for predicting launch outcomes.
  - Machine Learning models were trained and compared using cross-validation. The best-performing model highlighted the key features that most influence launch success, providing insights for data-driven decision-making.

Introduction

- Objective
  - Evaluate the feasibility of the new company Space Y to compete in the aerospace industry against Space X, by leveraging data analytics and machine learning.
- Key Questions to Address
  - What is the most effective way to estimate total launch costs, based on predictions of successful first-stage rocket landings?
  - Which launch sites offer the best conditions for achieving successful missions?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from two main sources:
    - SpaceX API: https://api.spacexdata.com/v4/rockets/
    - Web Scraping: Data retrieved from publicly available launch records (Wikipedia)
- Perform data wrangling
  - The dataset was enhanced by creating a new label for landing outcomes.
  - Additional feature engineering was performed after analyzing and summarizing relevant variables.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Conducted using SQL and visualizations.
  - Included feature engineering to prepare data for model training.
  - Helped to understand which variables influence mission outcomes.
- Interactive visual analytics
  - Implemented through:
    - Folium: for geographic mapping of launch sites and success/failure markers.
      - Plotly Dash: to create dynamic dashboards and explore patterns interactively.

Methodology

Executive Summary

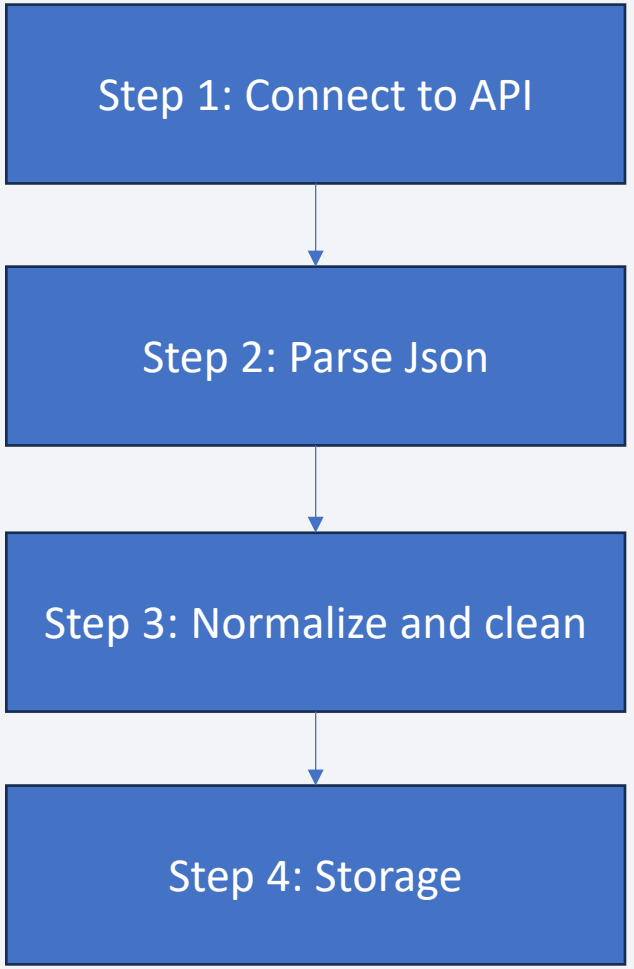
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated multiple classification models.
  - Models were trained on data obtained from:
    - SpaceX API: https://api.spacexdata.com/v4/rockets/
    - Web Scraping: Publicly available launch records (e.g., Wikipedia)

Data Collection

- Primary data sources:
  - SpaceX API – Structured launch data retrieved programmatically.
  - Web Scraping (Wikipedia) – Supplementary data collected from public records of launch outcomes.
- Process Summary:
  1. Access API endpoint to obtain JSON-formatted launch records.
  2. Scrape and parse launch details from Wikipedia using BeautifulSoup.
  3. Merge both datasets using common launch identifiers.
  4. Clean and normalize data to ensure consistency across features.
- Key Output:
  - Combined dataset with location, date, payload, orbit, and landing outcome.
  - Ready for analysis and machine learning tasks.

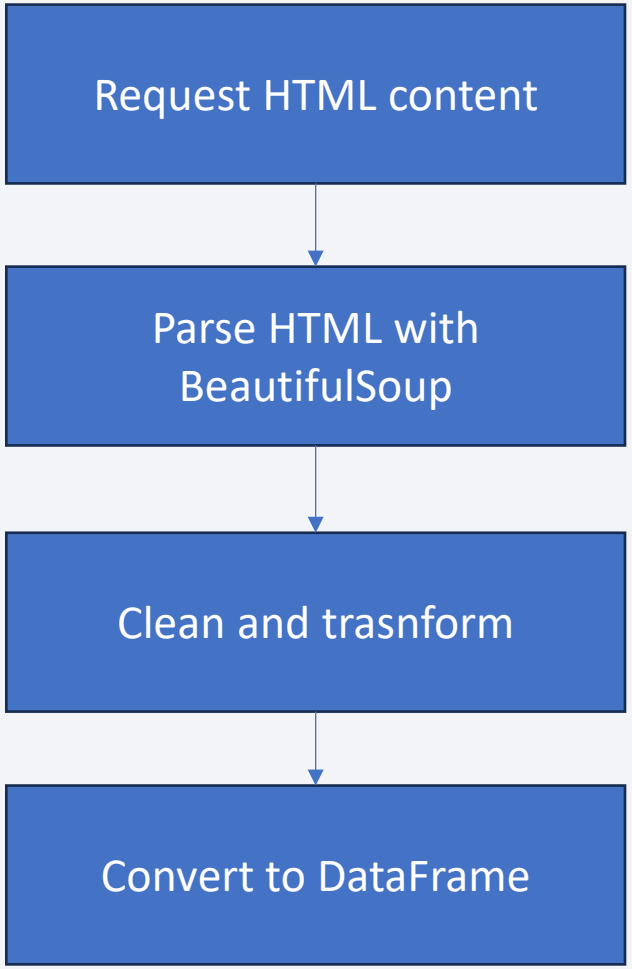
Data Collection – SpaceX API

- Step 1: Connect to the SpaceX REST APIAccess the endpoint:
  - https://api.spacexdata.com/v4/rockets
  - Perform a GET request using Python (requests.get())
- Step 2: Parse JSON response
  - Extract structured data on rocket launches, boosters, payloads, and landing outcomes
- Step 3: Normalize and clean
  - Transform nested fields into tabular format using pandas.json_normalize()
  - Handle missing values and data types
- Step 4: Store locally
  - Export as CSV or use a DataFrame for direct analysis



Data Collection - Scraping

- Target URL:
  - Wikipedia page with Falcon 9 and Falcon Heavy launch records (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Request HTML Content:
  - Used requests to fetch raw HTML content of the page.
- Parse HTML with BeautifulSoup:
  - Extracted the specific HTML <table> containing launch data.
- Clean & Transform:
  - Removed unwanted columns, renamed headers, and handled NaN values.
- Convert to DataFrame:
  - Structured the cleaned data into a Pandas DataFrame for analysis.

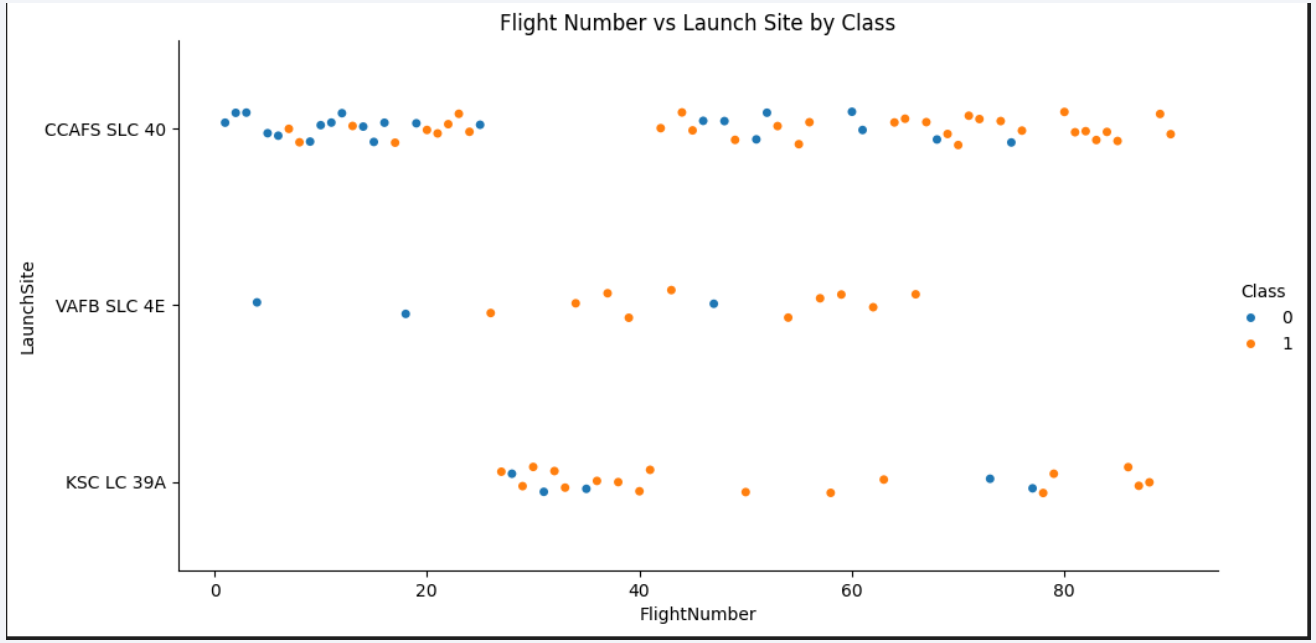


Data Wrangling

- Loaded Raw Data:
  - Loaded datasets from both SpaceX API and web scraping results.
- Created New Feature:
  - Generated a binary variable Landing Outcome to indicate success (1) or failure (0) of first-stage landings.
- Filtered and Cleaned:
  - Removed duplicate entries, standardized column names, handled missing values, and dropped irrelevant features.
- Merged Sources:
  - Joined API data and web-scraped tables into a unified DataFrame for analysis
- .Final Dataset:
  - Output stored in a clean Pandas DataFrame, ready for EDA and Machine Learning.

EDA with Data Visualization

- To explore the dataset, we used scatterplots and barplots to visualize relationships between key variables.
- These visual tools allowed us to uncover correlations, patterns, and outliers among flight features.



EDA with SQL

- The following SQL queries were executed to explore and summarize key insights from the dataset:
  - Retrieve unique launch sites from the mission records
  - Identify the Top 5 launch sites starting with 'CCA'
  - Calculate total payload mass for boosters launched by NASA (CRS)
  - Find the average payload mass for booster version F9 v1.1
  - Determine the first successful ground pad landing date
  - List boosters that:
    - Landed successfully on drone ships, andCarried a payload between 4000 and 6000 kg
    - Count the total number of successful and failed launches
  - Identify booster versions with the highest payload capacity
  - Extract failed drone ship landings in 2015 with booster and launch site details
  - Rank landing outcomes (e.g., success/failure) between 2010 to 2020

Build an Interactive Map with Folium

We used Folium to create dynamic and interactive maps that represent key insights spatially:

- Markers pinpoint specific locations, such as SpaceX launch sitesCircles highlight areas around coordinates (e.g., NASA Johnson Space Center)
- Marker Clusters group multiple markers at a single coordinate to avoid visual clutter (e.g., several launches at one site)
- Lines illustrate distances between coordinates (e.g., launch site to landing location)

Build a Dashboard with Plotly Dash

To provide a more interactive and user-friendly exploration of our analysis, we created a dashboard using Plotly Dash:

- Included dynamic plots and graphs to display mission outcomes by launch site and payload mass.
- Used interactive dropdowns and filters to allow users to select launch sites, payload ranges, or booster versions.
- Visualized success vs failure rates, payload trends, and correlation between orbit types and mission results.

Predictive Analysis (Classification)

To identify the best performing classification model for predicting mission success:

- We split the dataset into training and testing sets using an 80/20 ratio.
- Applied GridSearchCV with cross-validation (cv=10) to tune hyperparameters and evaluate models consistently.
- Tested multiple classification algorithms:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K-Nearest Neighbors (KNN)
- The models were compared based on their cross-validated accuracy scores, and the best model was selected based on performance on the test set.

Results

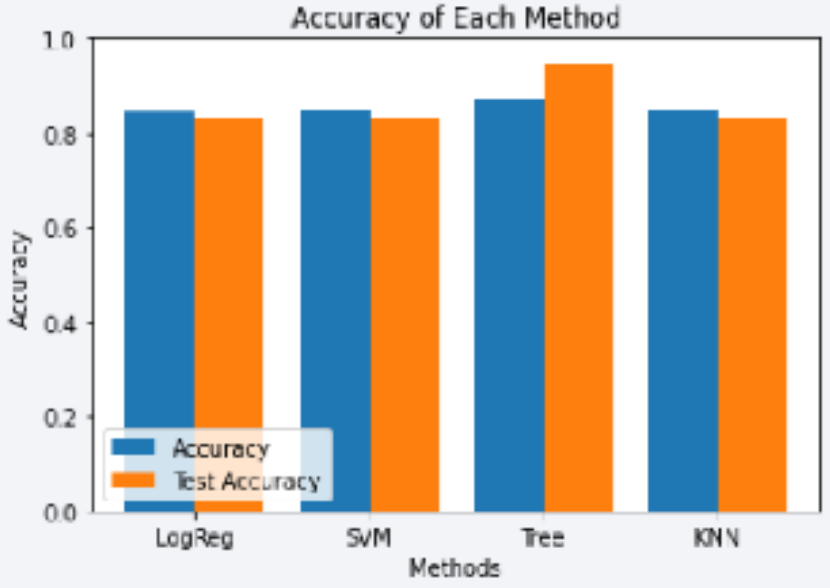
- Exploratory Data Analysis (EDA) – Key Findings
- SpaceX operates from 4 different launch sites.Initial launches were directed to SpaceX and NASA locations.
- The average payload for F9 v1.1 boosters is approximately 2,928 kg.
- The first successful landing occurred in 2015, five years after the initial launch.Several Falcon 9 booster versions succeeded in landing on drone ships when carrying payloads above the average.
- Nearly 100% of missions resulted in successful outcomes in recent years.
- Two specific booster versions (F9 v1.1 B1012 and B1015) failed landing attempts on drone ships in 2015.
- Landing success rates have improved over time, showing a clear upward trend in performance.

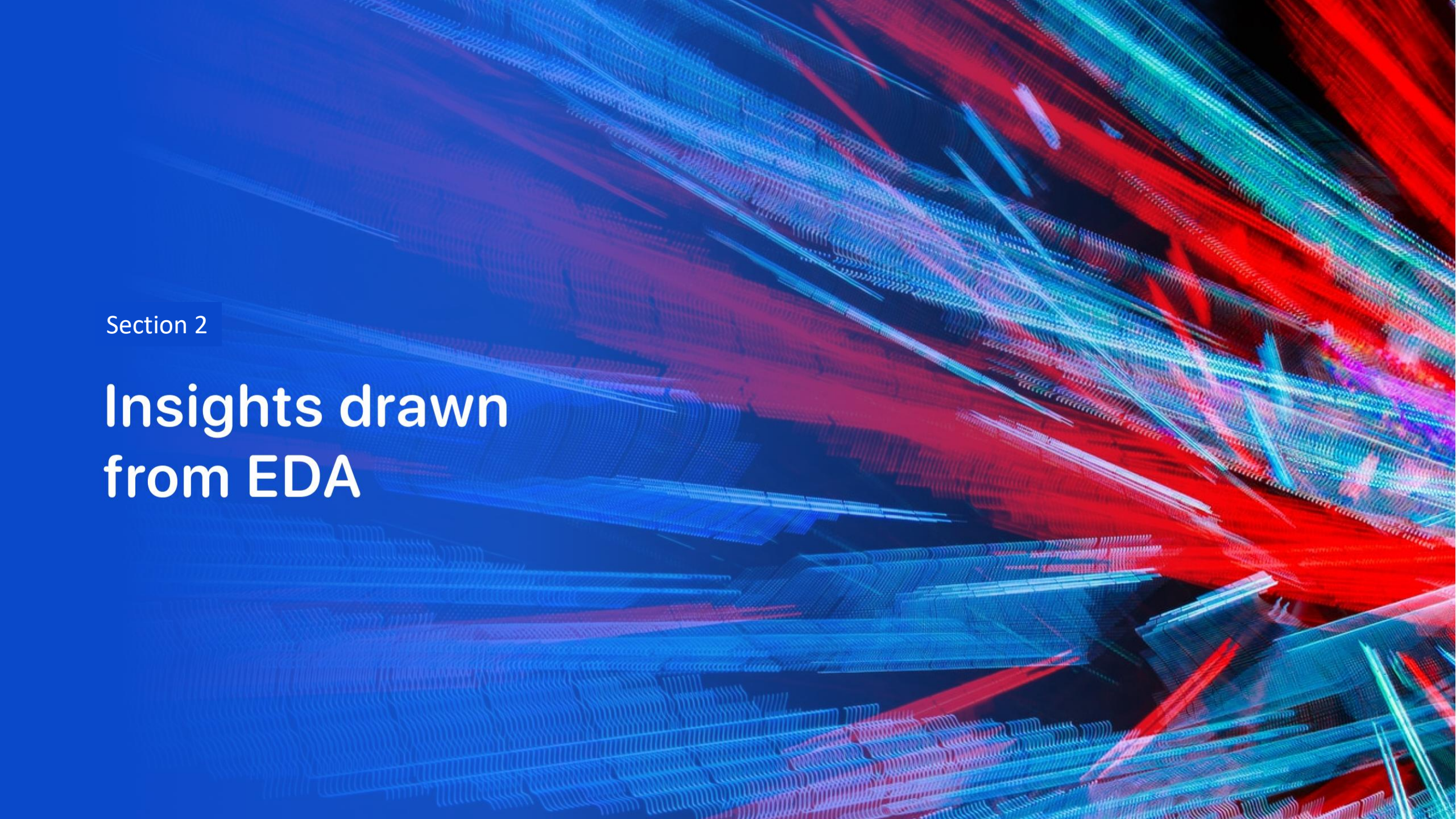
Results

- Exploratory Data Analysis (EDA) revealed relationships between features such as payload mass, orbit type, launch sites, and mission success. Key patterns and outliers were identified through scatterplots and bar charts.
- Interactive Maps were built using Folium to visualize launch sites, success rates, and distances to NASA centers. Marker clusters and circles helped highlight patterns geographically.

Results

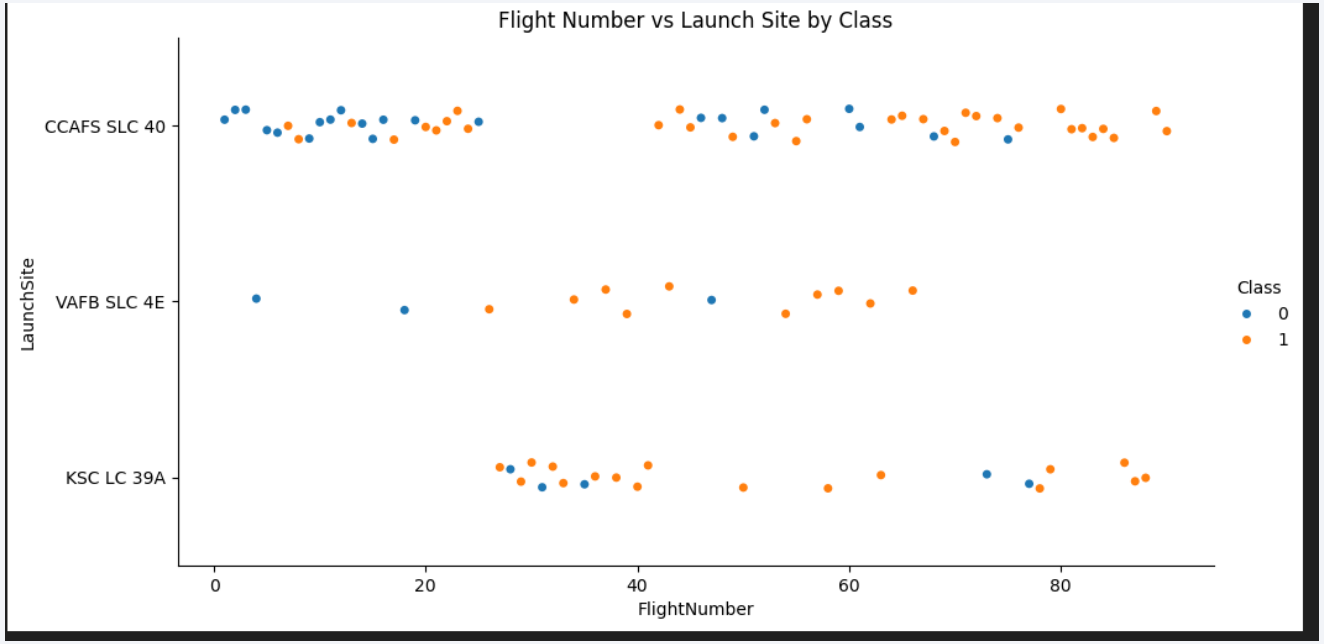
- Predictive Modeling was conducted using Logistic Regression, SVM, Decision Trees, and KNN.
- After hyperparameter tuning with GridSearchCV and cross-validation, the best-performing model achieved an accuracy of ~89%, successfully predicting the likelihood of mission success.




Section 2
Insights drawn from EDA

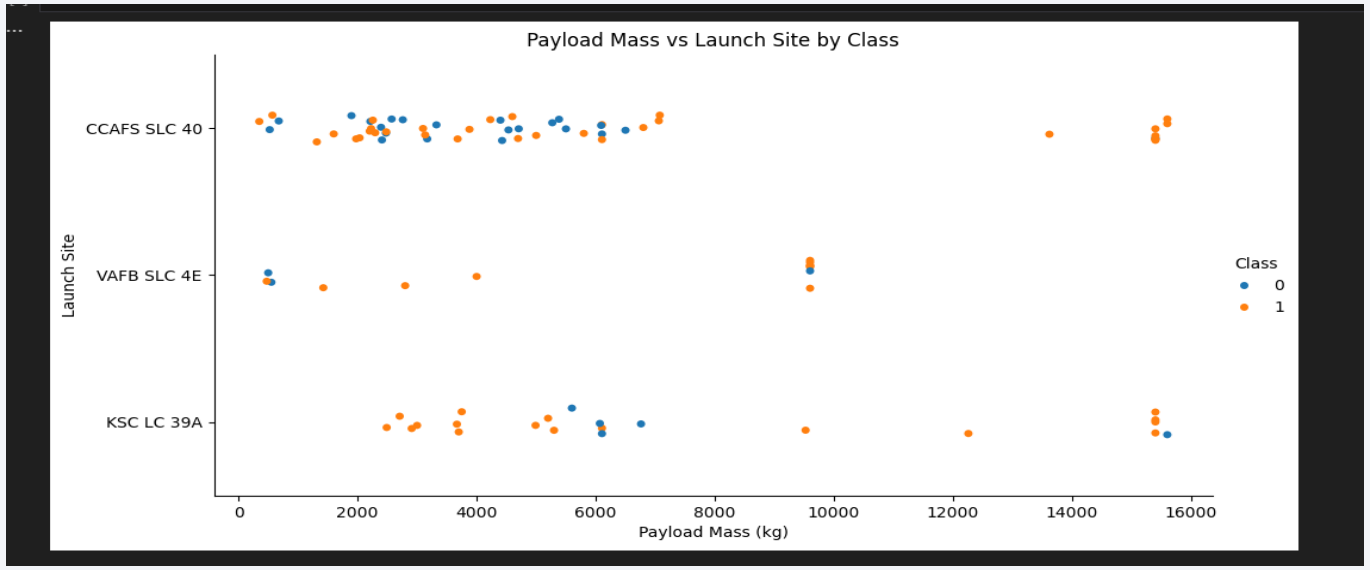
Flight Number vs. Launch Site

- This scatter plot illustrates the relationship between the flight number and the launch site, categorized by the launch outcome (class):
  - Class 1 (orange) represents successful landings, while Class 0 (blue) represents failed landings.
  - Most launches from CCAFS SLC 40 span the entire range of flight numbers and show a higher concentration of successful launches as the number of flights increases.
  - KSC LC 39A appears later in the timeline (higher flight numbers) and shows mostly successful outcomes, indicating technological and operational improvements.
  - VAFB SLC 4E shows fewer launches overall, but with a reasonable number of successful missions.
  - As the number of flights increases, success rate improves across all sites, especially in CCAFS and KSC, indicating growing launch reliability over time.



Payload vs. Launch Site

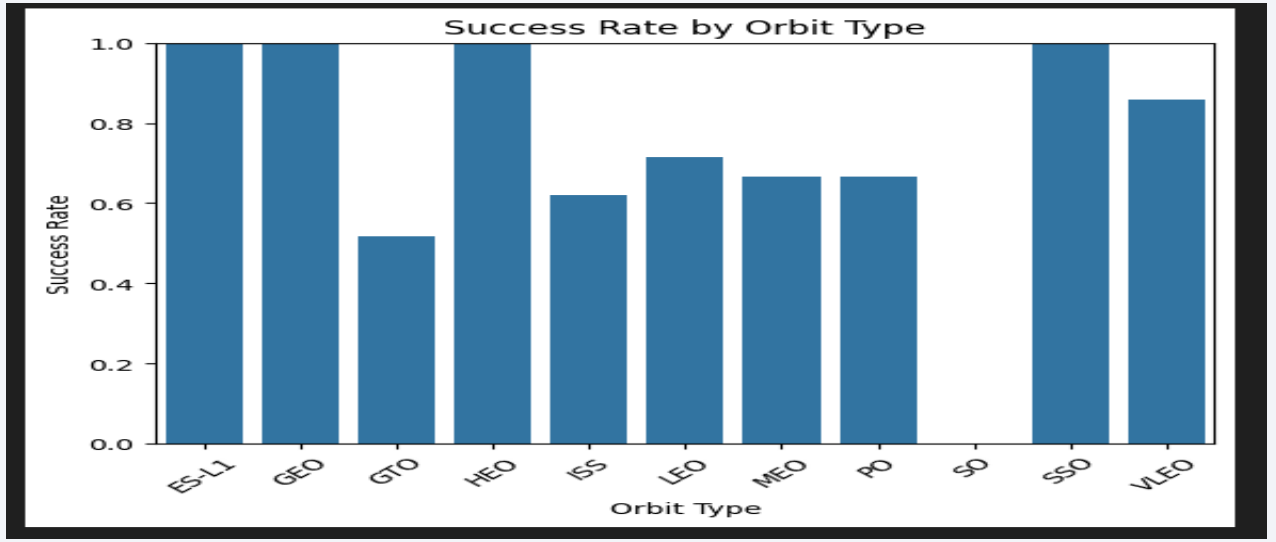
- This scatter plot displays the relationship between the payload mass (kg) and the launch site, categorized by mission outcome
  - :Class 1 (orange) indicates successful landings, and Class 0 (blue) indicates failed landings.
  - Most launches from CCAFS SLC 40 involve payloads between 2,000 and 6,000 kg, with a majority being successful. Some very high payload missions (>15,000 kg) were also mostly successful.
  - KSC LC 39A handled both mid-range and very heavy payloads, with a balanced mix of successes and failures. However, most high-mass payloads launched here succeeded, suggesting good infrastructure and experience.
  - VAFB SLC 4E conducted fewer launches, generally with lower and mid-range payloads, and showed mixed outcomes.
  - In general, mission success is not strictly dependent on payload mass, although heavier payloads tend to succeed more often at major launch sites like KSC LC 39A and CCAFS SLC 40.



Success Rate vs. Orbit Type

This bar chart illustrates the success rate of SpaceX missions based on the target orbit type:

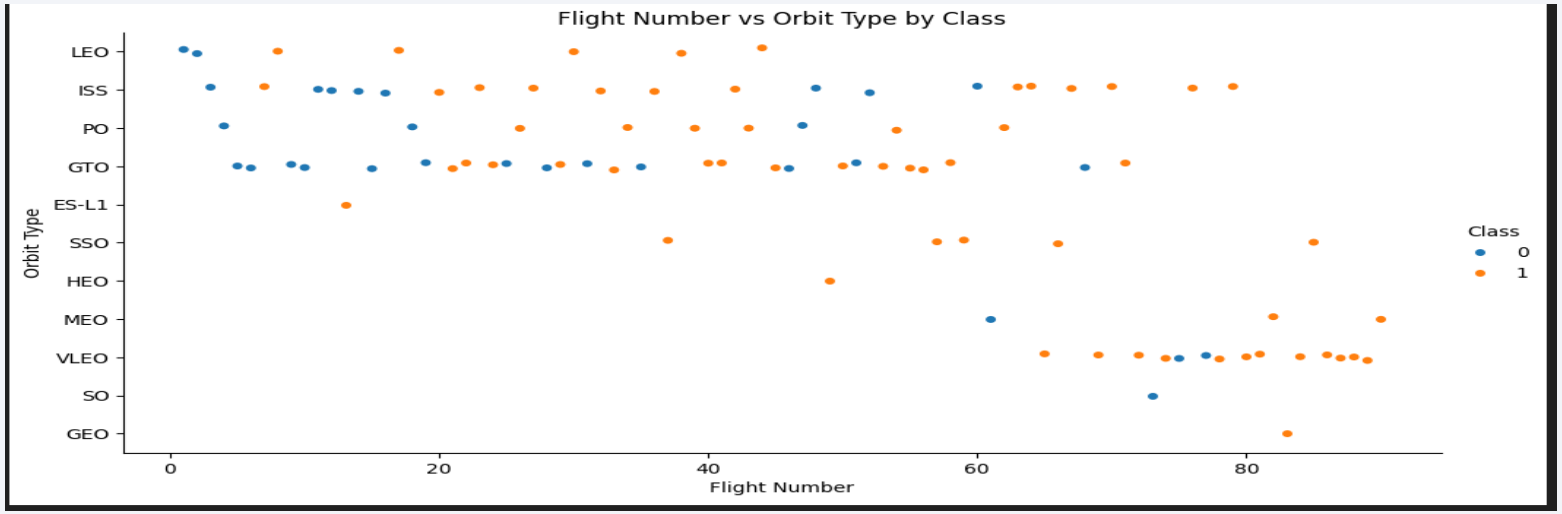
- ES-L1, GEO, HEO, and SSO orbits exhibit a 100% success rate, indicating high reliability for missions targeting these orbital paths.
- VLEO (Very Low Earth Orbit) also shows a strong success rate (≈ 87%), suggesting it's a relatively stable target.
- In contrast, GTO (Geostationary Transfer Orbit) has the lowest success rate, below 55%, indicating higher risks or operational complexity.
- Orbits like LEO, MEO, and ISS show moderate success rates between 63% and 73%.
- The data implies that missions targeting higher or specialized orbits tend to be more consistent, while missions to transitional or lower-stability orbits (like GTO or ISS) present more challenges.



Flight Number vs. Orbit Type

This scatterplot shows the distribution of SpaceX flight numbers across various orbit types, differentiated by mission success (Class 1) and failure (Class 0):

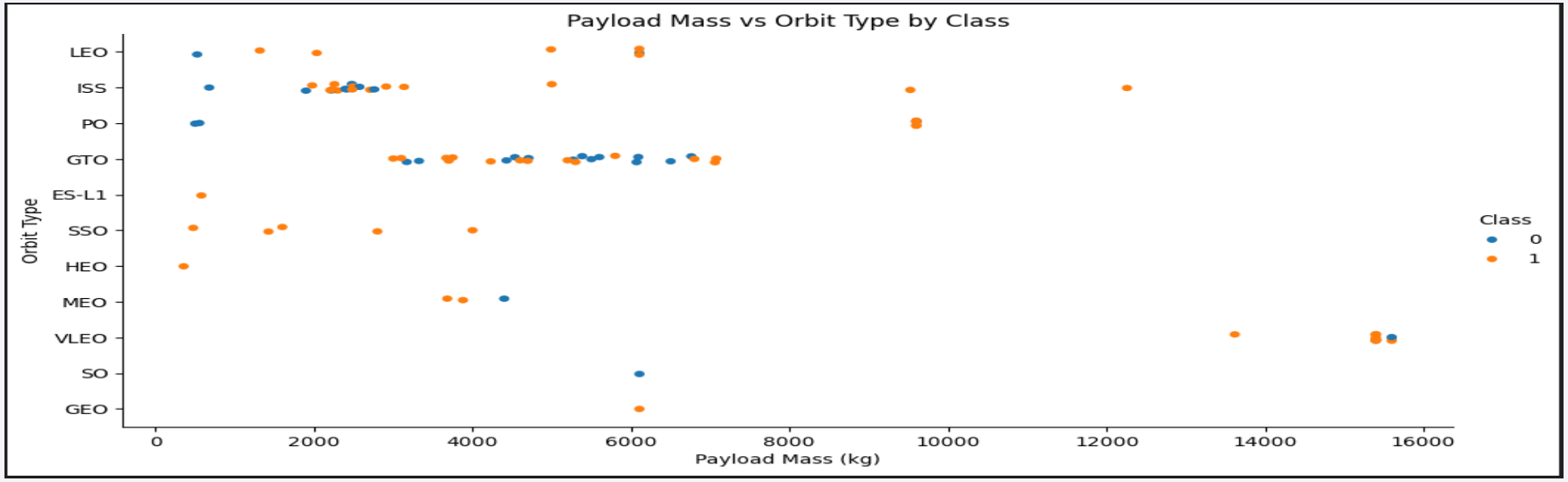
- Higher flight numbers (i.e., more recent launches) are increasingly associated with successful missions (orange points) across almost all orbit types.
- Orbit types such as LEO, GTO, and ISS had early launches with mixed outcomes, but over time success rates improved.
- VLEO and SSO appear more frequently in later flights, showing consistent success (most orange dots), suggesting they were introduced later and benefited from technical improvements.
- Some orbit types, like HEO and GEO, are less common but had only successful launches, possibly due to cautious planning or fewer missions.



Payload vs. Orbit Type

This scatterplot analyzes the relationship between payload mass (kg) and orbit type, differentiating between mission outcomes (Class 1 for success, Class 0 for failure):

- Most successful launches (Class 1) are concentrated across payloads below 7000 kg, especially in GTO, ISS, and VLEO orbits.
- Some orbits like VLEO, HEO, and SSO had consistently successful missions even with very high payloads (above 15000 kg), showing strong engineering performance.
- Failures (blue dots) tend to occur in low to mid payload ranges, particularly in LEO, GTO, and ME orbits, suggesting early-stage limitations or risk-prone mission profiles.
- GTO and VLEO had the widest range of payload masses, reflecting their adaptability across mission types.



Launch Success Yearly Trend

This line chart illustrates the evolution of SpaceX mission success rates from 2010 to 2020:

- From 2010 to 2013, the success rate remained at 0%, indicating early development and testing phases.
- A noticeable improvement began in 2014, with a steady increase in successful launches over the following years.
- The peak success rate occurred in 2019, reaching over 90%, reflecting enhanced technology, reliability, and mission planning.
- The dip in 2018 suggests possible anomalies or mission complexities that briefly affected reliability.
- Overall, the trend reveals significant progress in SpaceX's launch capabilities, confirming a strong learning curve and operational maturity over the decade.

All Launch Site Names

According to the dataset, there are four unique launch sites used by SpaceX:

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

These values were identified by extracting the distinct entries from the launch_site column in the dataset. Recognizing these unique launch locations is essential for mapping and analyzing launch success based on geographical distribution.

Launch Site Names Begin with 'CCA'

This table shows the first five launches from the dataset. Key insights include:

- All launches occurred at CCAFS LC-40, confirming its importance in early missions.
- The payloads were related to demonstration or resupply missions for NASA (COTS and CRS) and SpaceX.
- The initial payload masses were relatively small, reflecting early test missions (e.g., 0 to 677 kg).
- Although all missions were marked as successful, the landing outcomes were either failures due to parachute issues or no attempt, showing that reusable booster recovery was not implemented yet.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- This value reflects the total mass of payloads delivered by SpaceX rockets for NASA CRS missions.
- It provides insight into NASA's role in resupply operations and its dependency on commercial partners.
- The use of SQL allowed efficient filtering based on customer criteria.

Total_Mass
48213

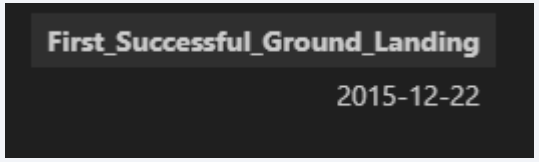
Average Payload Mass by F9 v1.1

- The result shows the average payload mass carried per mission by the F9 v1.1 booster version.
- This insight helps assess the performance and reliability of this booster in lifting moderate payloads.
- It provides a baseline to compare with other booster versions (e.g., F9 Full Thrust or F9 Block 5).

Average_Mass
2534.6666666666665

First Successful Ground Landing Date

- This date marks a major milestone in reusable rocket technology.
- The Falcon 9 successfully landed on a ground pad for the first time, reducing launch costs and enabling rapid reuse.
- It reflects SpaceX's transition toward sustainable and cost-effective orbital transportation.



Successful Drone Ship Landing with Payload between 4000 and 6000

- These boosters achieved successful landings on drone ships despite carrying moderate to heavy payloads.
- This highlights their reliability and performance in recovering expensive components under demanding 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

- The vast majority of missions were successful, indicating a high reliability of SpaceX launches.
- The single in-flight failure was a rare occurrence, suggesting strong operational consistency.
- The presence of unclear or duplicate success entries suggests a need for data cleaning or metadata clarification.

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

Boosters Carried Maximum Payload

- All booster versions capable of lifting the **maximum payload** belong to the **F9 Block 5 series**, indicating it is the most **powerful and reliable version** developed for heavy missions.

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

- Two drone ship landings failed in 2015, both launched from **CCAFS LC-40** using **F9 v1.1** boosters, occurring in **January and April**. This suggests early challenges in drone ship recovery attempts during that year.

Month	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

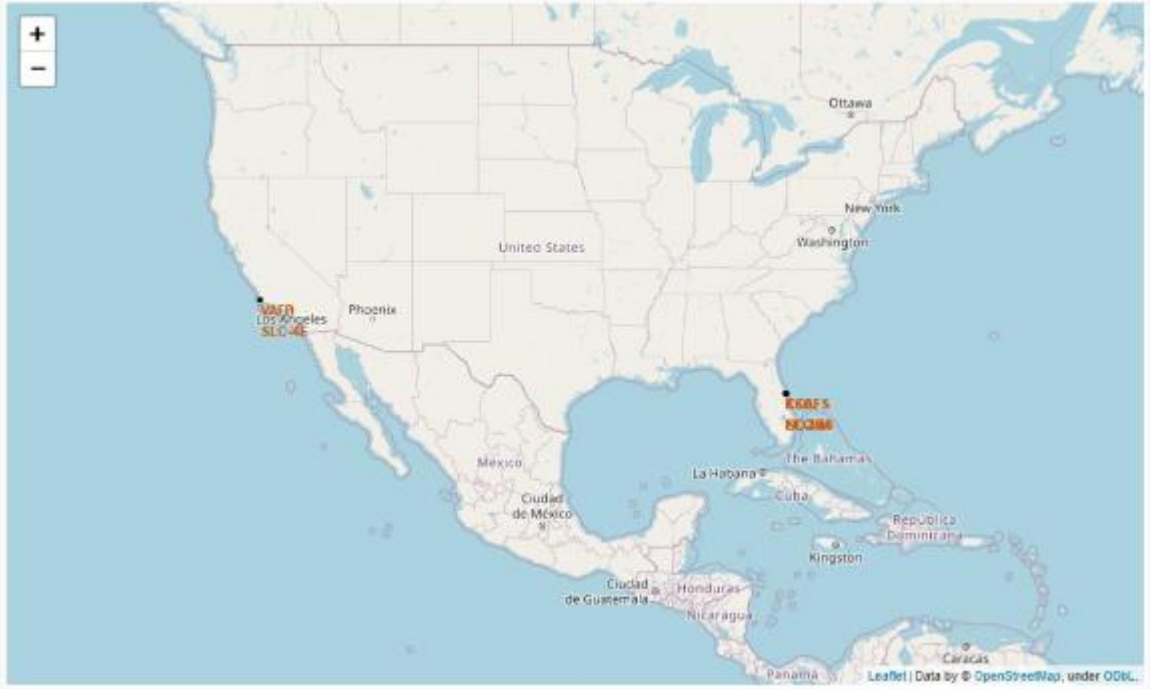
- The most frequent outcome was "No attempt", indicating earlier missions didn't plan or attempt recovery.
- Drone ship landings had both notable successes and failures (5 each), reflecting technological learning.
- Ground pad landings had only 3 successful attempts in this period, suggesting they became more viable later.

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

Section 3

Launch Sites Proximities Analysis

All launch Sites



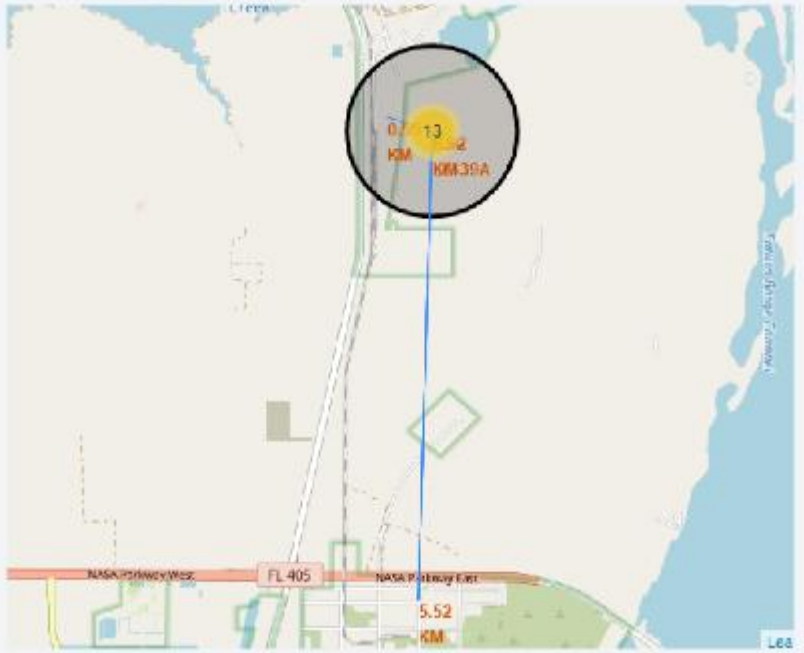
- The map shows all launch sites used by SpaceX missions, including:
  - VAFB SLC-4E (California)
  - CCAFS LC-40 / SLC-40 and KSC LC-39A (Florida)
- These launch sites are strategically located close to the sea, likely due to safety considerations, such as minimizing risks during takeoff or recovery operations.
- They are also positioned not far from major roads and rail infrastructure, allowing easy transport of rockets and payloads.

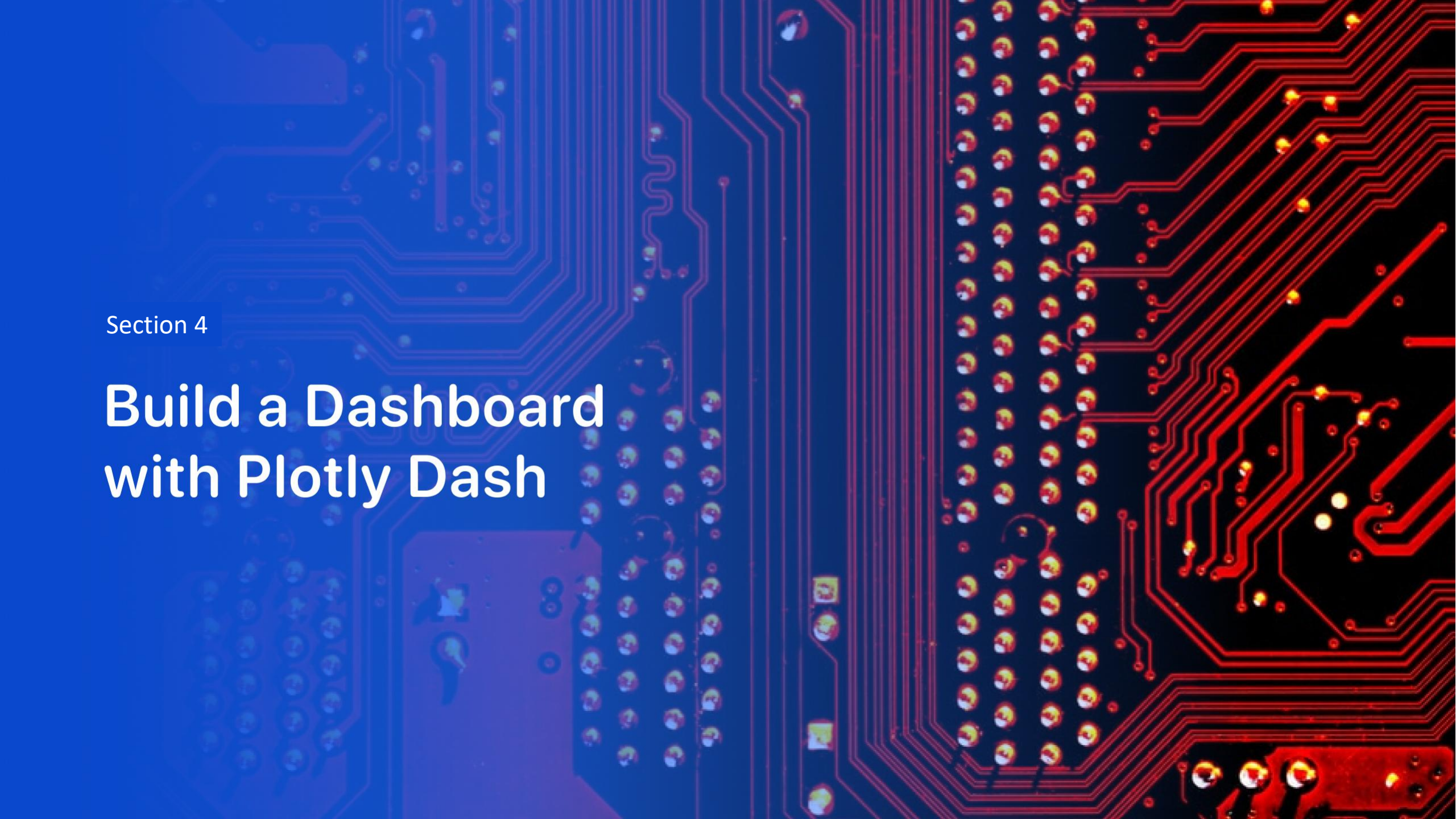
Lauch Outcomes by Sites

- The map displays launch outcomes by location, focusing on KSC LC-39A as an example.
- Each marker cluster shows the number of launches and their outcomes at a specific site.Green markers represent successful launches, while red markers indicate failures.
- This visualization helps identify which launch sites have higher reliability based on historical success rates.

Logistics and Safety

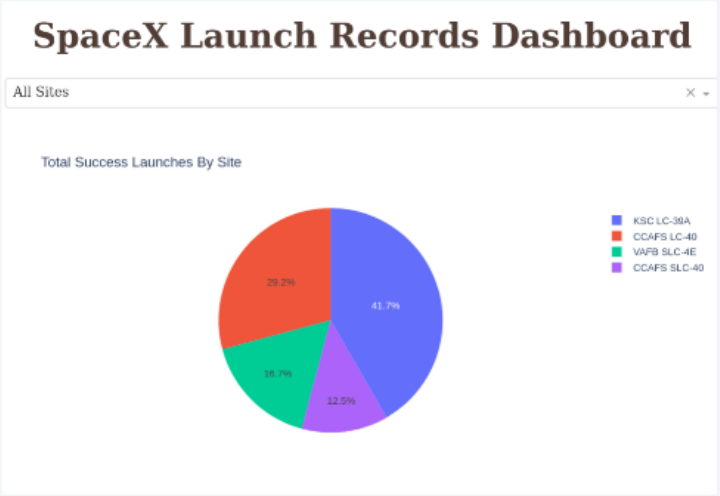
- The launch site KSC LC-39A offers strategic logistic advantages.
- It is located close to major roads and railroads, which facilitates transportation of equipment and personnel.
- Additionally, the site is relatively distant from densely populated areas, which enhances safety in case of launch anomalies or emergencies.




Section 4
Build a Dashboard with Plotly Dash

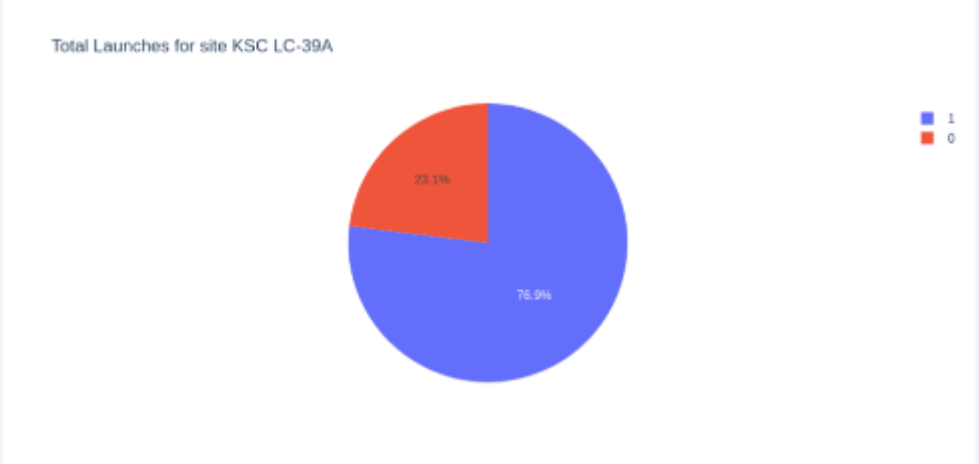
Successful Launches by Site

- The location of the launch site appears to be a significant factor influencing mission success rates.
- According to the dashboard, KSC LC-39A has the highest proportion of successful launches, followed by CCAFS LC-40 and VAFB SLC-4E.
- This may reflect differences in infrastructure, weather conditions, or mission types assigned to each site.



Launch Success Ratio for KSC LC-39A

- A total of 76.9% of launches conducted from KSC LC-39A were successful, indicating a high reliability for this launch site.
- This success rate highlights the site's strong performance and may be linked to factors such as infrastructure, technical readiness, and logistical advantages.



Payload vs. Launch Outcome

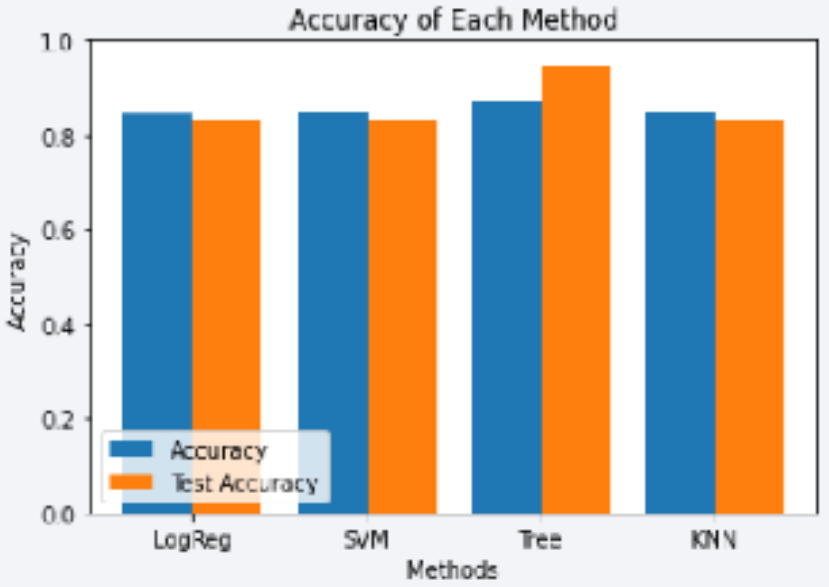
- Payloads under 6,000 kg combined with FT (Full Thrust) boosters show the highest success rate among all combinations.
- This suggests that lighter payloads and advanced booster versions are optimal for mission reliability.

Section 5

Predictive Analysis (Classification)

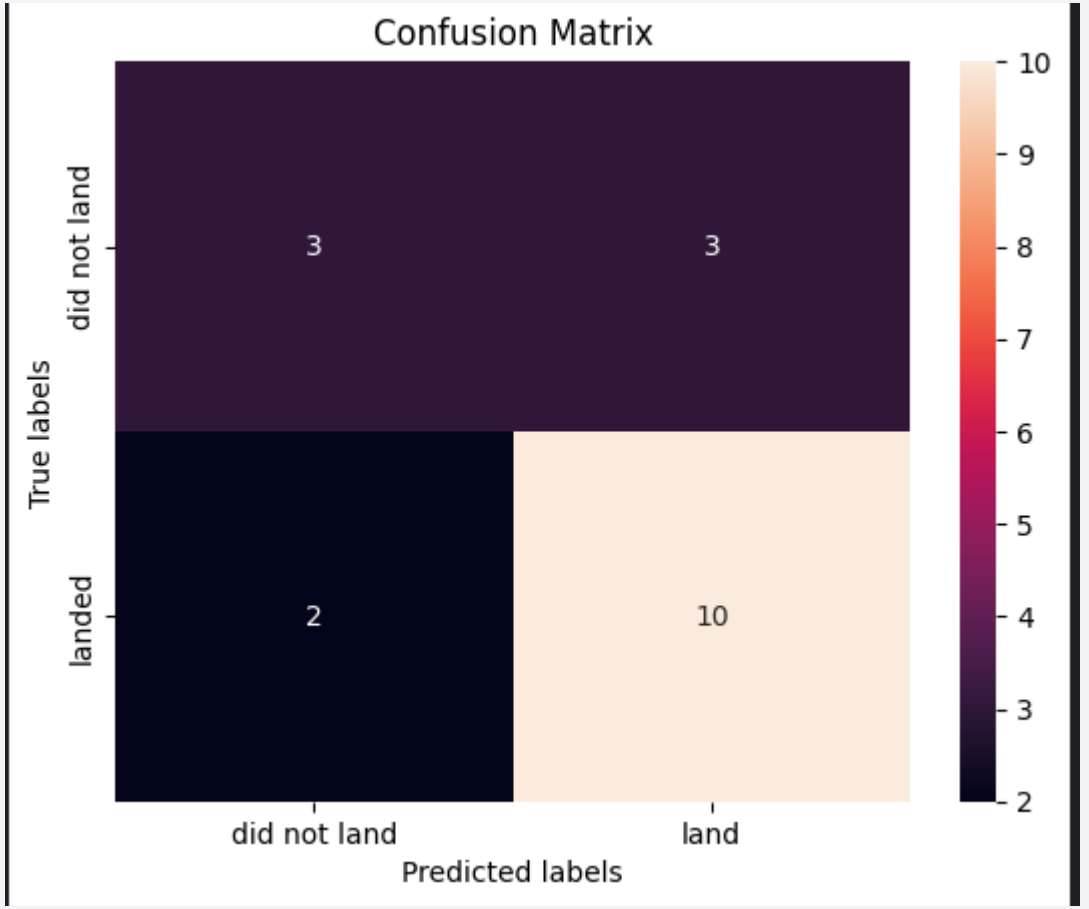
Classification Accuracy

- Four classification models were evaluated:
  - Logistic Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbors.
- Their respective training and test accuracies are displayed in the adjacent bar chart.
- Among them, the Decision Tree Classifier achieved the highest test accuracy, exceeding 87%, making it the best-performing model in this analysis.



Confusion Matrix

- This confusion matrix corresponds to the best-performing model in the classification task.
- The matrix shows:
  - True Positives (TP): 10 launches correctly predicted to land.
  - True Negatives (TN): 3 launches correctly predicted to not land.
  - False Positives (FP): 3 launches predicted to land but did not.
  - False Negatives (FN): 2 launches predicted to not land but actually landed.
- Despite a few misclassifications, the model performs well overall, especially in predicting successful landings.



Conclusions

- Exploratory Data Analysis (EDA):
  - Four main launch sites were identified, with KSC LC-39A standing out for its high number of successful launches.
  - The success rate has shown an upward trend over the years, exceeding 80% after 2017.
- Geospatial Visualization (Folium):
  - Launch sites are strategically located near the sea, enhancing safety and logistics.
  - Success clusters were particularly concentrated around KSC LC-39A, with fewer failure records.
- SQL Analysis:
  - The highest average payload mass was observed in booster version F9 v1.1, with total payloads exceeding 48,000 kg for NASA (CRS) missions.
  - Most boosters with payloads between 4000–6000 kg and successful drone ship landings belong to the F9 FT series.

Conclusions

- Interactive Dashboard with Plotly Dash:
  - The interactive interface made it easy to observe patterns, such as the relationship between payload mass and launch outcome.
  - Payloads under 6000 kg combined with FT or B5 boosters resulted in the highest success probabilities.
- Predictive Modeling and Evaluation:
  - Classification models tested included Logistic Regression, SVM, KNN, and Decision Tree.
  - The Decision Tree Classifier achieved the highest accuracy, over 87%.
  - The confusion matrix confirmed strong model performance, showing high true positive and true negative rates.

Thank you!