IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adisorn Promkaewngarm
7 January 2026

Outline

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

Executive Summary

- **Objective:** Predict Falcon 9 first-stage landing success to optimize launch cost estimations.
- **Methodology:** Collected data via SpaceX API and Wikipedia; performed EDA with SQL and visualization; built Folium maps and Plotly Dash dashboards; and tuned classification models.
- **Key Results:** While baseline models achieved 83.33%, the Decision Tree model reached a superior test accuracy of 94.44%.

Introduction

- **Project Background:** SpaceX has significantly reduced space access costs by making rocket stages reusable. Predicting successful first-stage landings is essential for determining the actual cost of a launch.
- **Context:** While competitors charge up to **$165 million** per launch, SpaceX advertises the Falcon 9 at **$62 million**. Much of these savings depend on whether the first stage can be recovered and reused.
- **Problems to Answer:**
  - Can we accurately predict if the Falcon 9 first stage will land successfully based on public launch data?
  - What factors (e.g., payload mass, orbit type, launch site) have the highest impact on landing success?
  - Which machine learning algorithm provides the most reliable prediction for landing outcomes?

Section 1

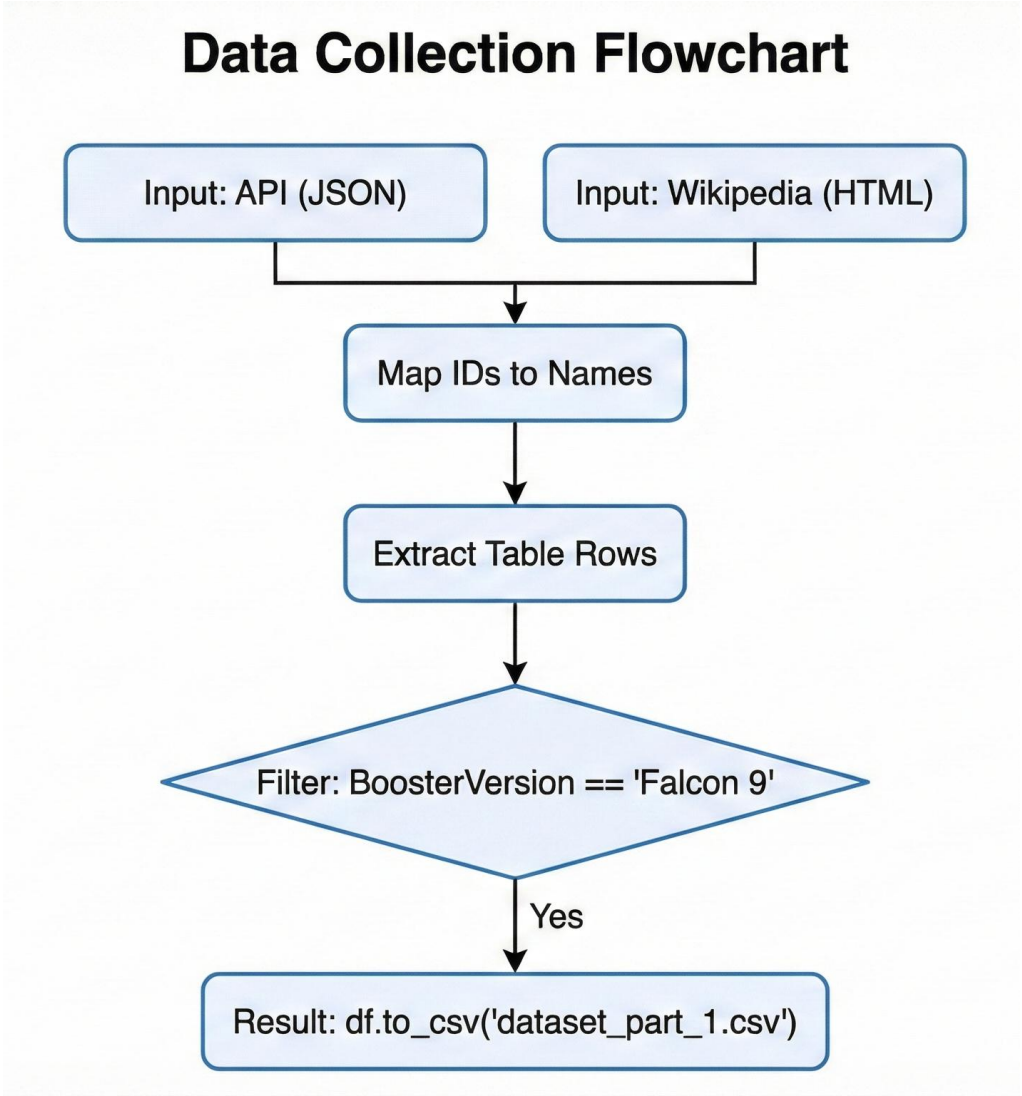
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Extracted data via SpaceX REST API and Wikipedia web scraping with BeautifulSoup
- Perform data wrangling
  - Cleaned data, handled missing values, and created a binary landing outcome variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained and evaluated Logistic Regression, SVM, Decision Tree, and KNN models using GridSearchCV.

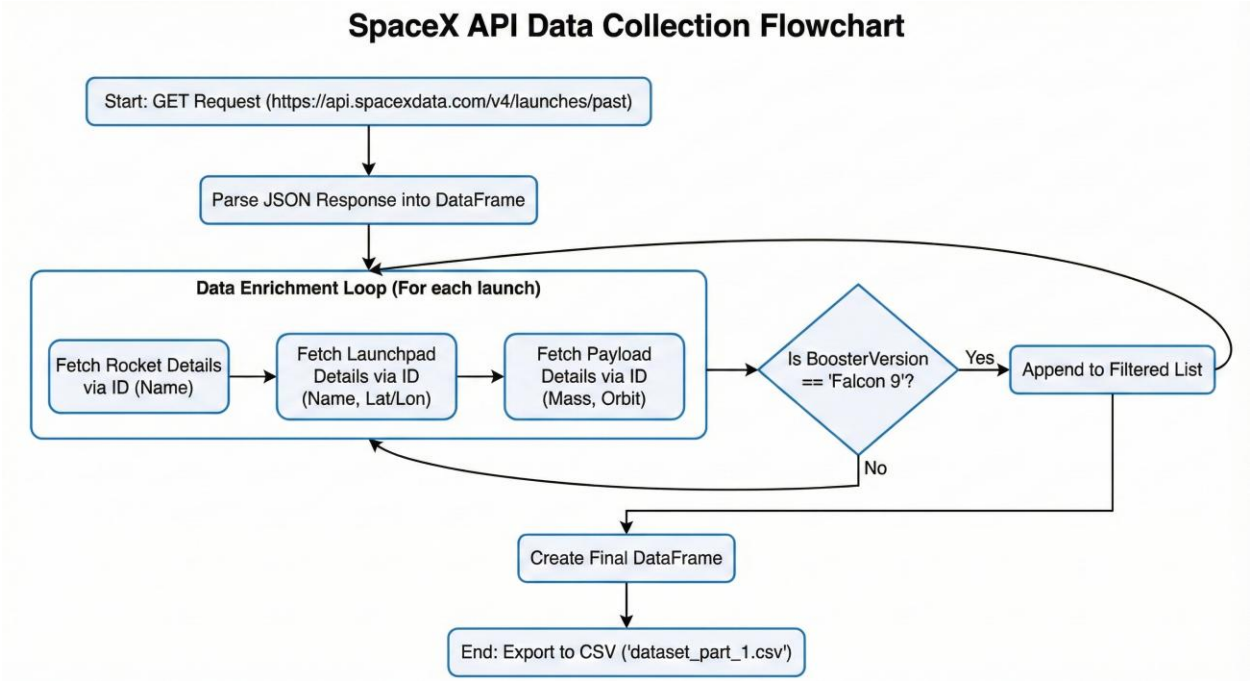
Data Collection



- **API Data Acquisition:** Requests data from the SpaceX API endpoint and uses helper functions to resolve IDs into human-readable names.
- **Web Scraping:** Employs BeautifulSoup to parse HTML tables from Wikipedia for historical launch data.
- **Dataset Consolidation:** Filters for Falcon 9 launches and merges data into a structured Pandas DataFrame.

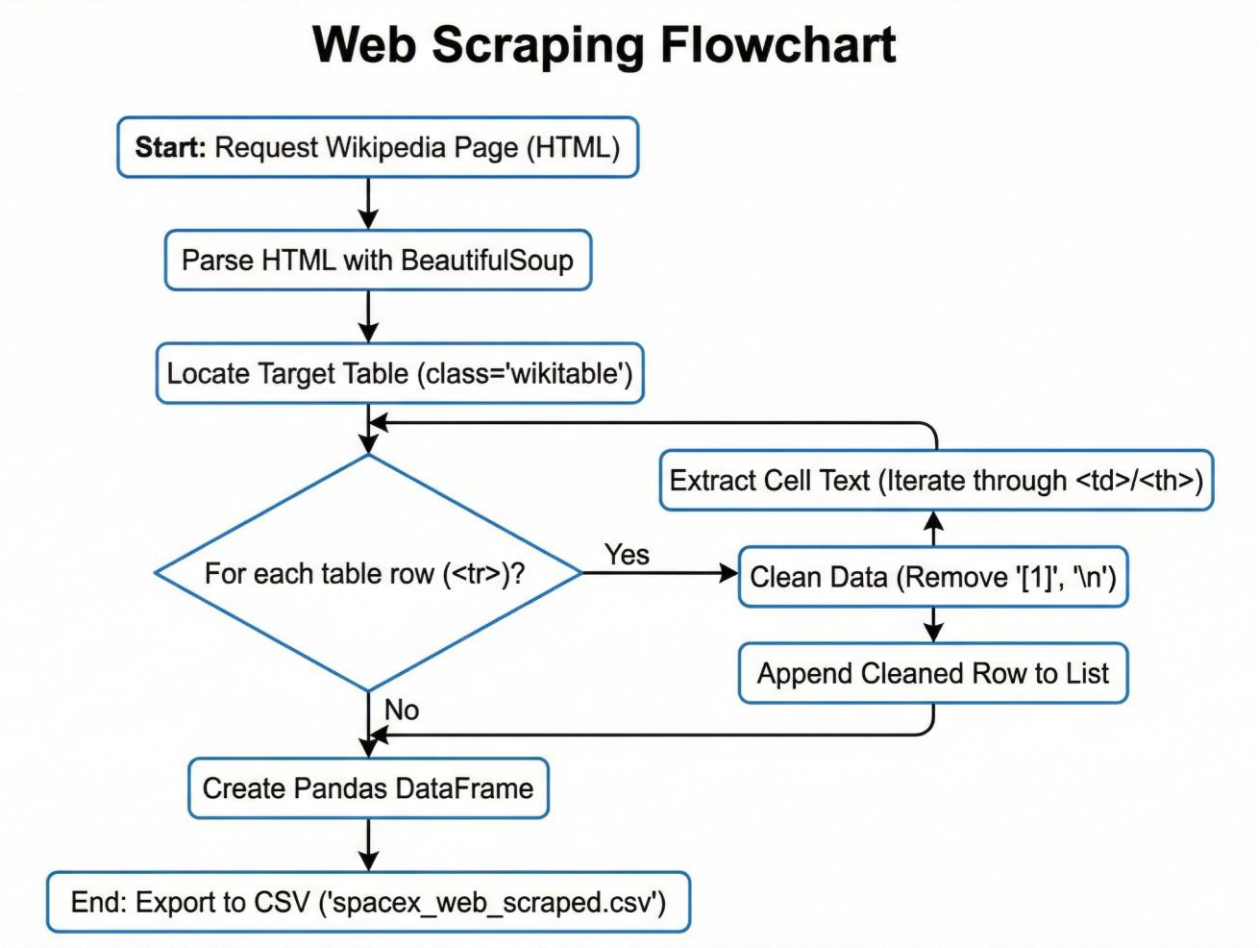
Data Collection – SpaceX API

- **REST API Methodology:**
  - Request: Called the SpaceX API endpoint to retrieve historical launch data in JSON format.
  - Parsing: Mapped technical IDs (Rocket, Payload, Launchpad) to human-readable names via secondary API calls.
  - Filtering: Focused specifically on Falcon 9 launches by filtering out "Falcon 1" data.
- **Output:**
  - Generated a structured dataframe exported as dataset_part_1.csv.
- **GitHub Repository:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/ca62457cc8b12ecce2e4b6de1bc74b4de05af6d7/05-01-01%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb



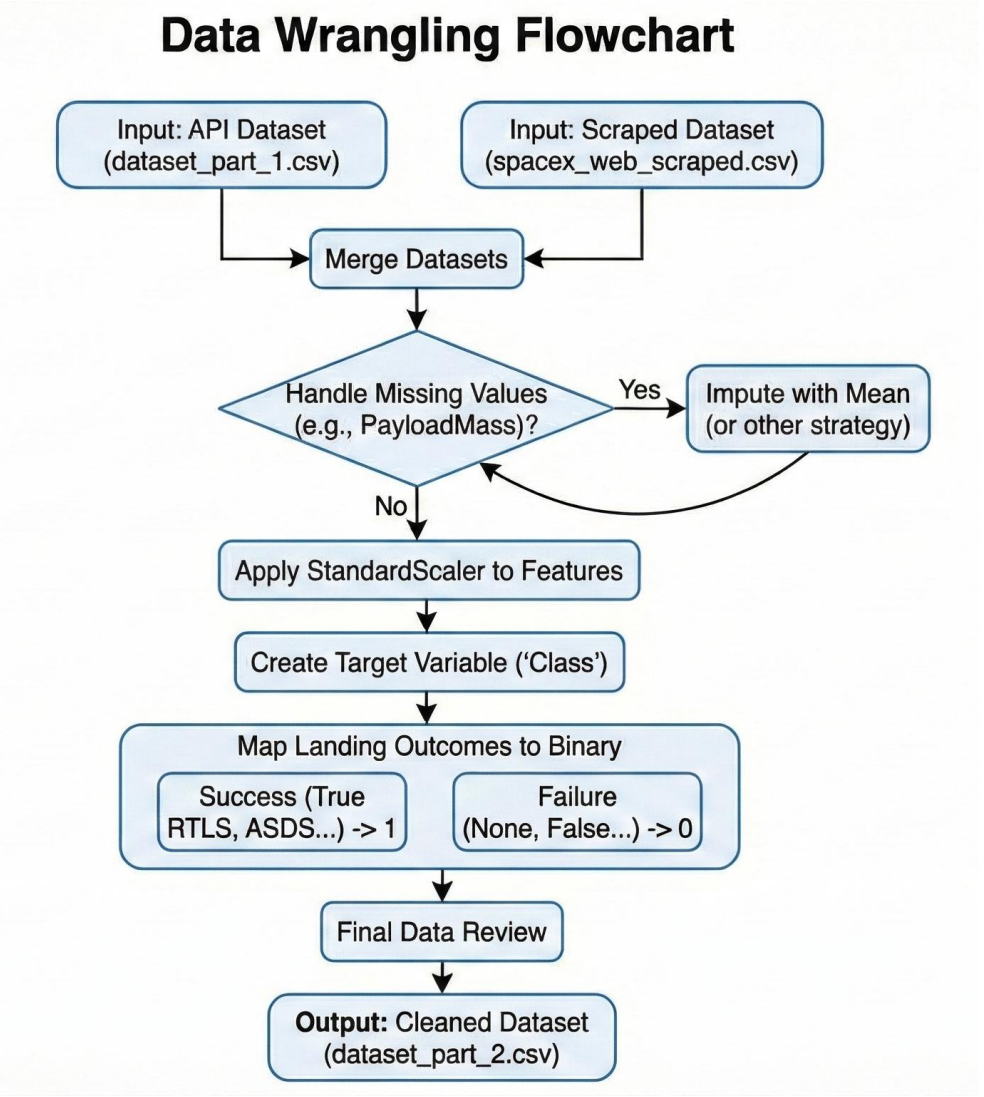
Data Collection - Scraping

- **Data Collection – Scraping**
  - Process: Used BeautifulSoup to parse HTML tables from Wikipedia.
  - Methodology: Iterated through wikitable rows to extract launch details and cleaned HTML tags/citations.
  - Output: Filtered for Falcon 9 and exported to spacex_web_scraped.csv.
- **GitHub URL:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-01-02%20Webscraping.ipynb



Data Wrangling

- **Data Wrangling**
  - Data Processing: Cleaned and transformed raw datasets from API and web scraping to prepare for exploratory analysis and machine learning.
  - Missing Value Imputation: Handled null entries in PayloadMass by replacing them with the column mean.
  - Success Labeling: Created a binary Class column where 1 represents a successful landing and 0 represents a failure.
  - Standardization: Applied StandardScaler to the feature matrix X to normalize data, ensuring all features contribute equally to the model's objective function.
- **GitHub URL:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-01-03%20Data%20wrangling-v2.ipynb



EDA with Data Visualization

- **EDA with SQL**
  - Objective: Query dataset to identify launch trends and success patterns.
  - Database: Loaded cleaned data into SQLite/DB2 for relational analysis.
  - Key Findings: Identified unique launch sites, calculated total payload masses, and determined the highest success rates per booster version.
  - Result: Verified that success rates improved with mission frequency and specific orbit types.
- GitHub URL:
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-02-01%20eda-sql-coursera_sqllite.ipynb
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-02-02%20eda-dataviz-v2.ipynb

EDA with SQL

- **Objective:** Query the dataset to identify patterns and success metrics.
- **Database:** Loaded dataset_part_2.csv into a SQL environment for relational analysis.
- **Query Summaries:**
  - **Launch Sites:** Identified unique sites and determined total launches per location.
  - **Payload Analysis:** Calculated total/average payload mass and identified boosters carrying specific mission types (e.g., NASA CRS).
  - **Success Trends:** Filtered records to find the most successful landing outcomes and booster versions.
  - **Records Analysis:** Identified the first successful ground landing and ranked launch sites by success rate.
- **GitHub URL:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-02-01%20eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

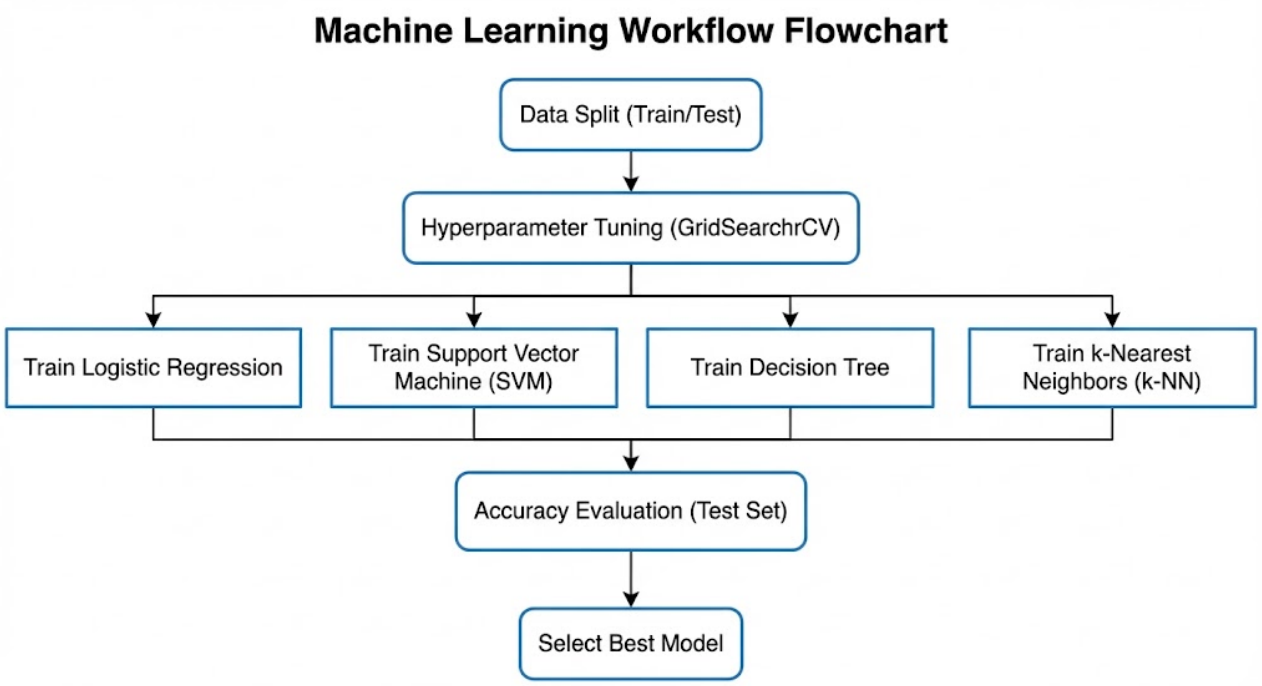
- **Map Objects**: Created a global base map with Markers for each launch site, Circles to visualize safety perimeters, and Lines (PolyLines) to calculate distances to the nearest coastlines and highways.
- **Purpose:** To visually analyze the geographical distribution of launch sites and their proximity to critical infrastructure and landing zones.
- GitHub URL:
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-03-01%20launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- **Interactive Components:** Developed a web application using Plotly Dash featuring a Dropdown for launch site selection and a RangeSlider to filter by payload mass.
- **Visualizations:**
  - Pie Chart: Displays the total success launches for all sites or a specific selected site.
  - Scatter Plot: Shows the correlation between payload mass and landing success, color-coded by booster version.
- **Purpose:** To provide an interactive tool for real-time data exploration and to identify the optimal launch conditions for success.
- **GitHub URL:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-03-02%20Plotly.ipynbb

Predictive Analysis (Classification)

- **Model Building:** Developed and trained four classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (k-NN).
- **Evaluation & Improvement:** Utilized GridSearchCV for hyperparameter tuning to optimize model performance across all algorithms.
- **Validation Strategy:** Implemented a Train/Test split on the standardized feature matrix to rigorously evaluate the models on unseen data and prevent overfitting.
- **Performance Metrics:** Defined Accuracy Scores and Confusion Matrices as the primary evaluation criteria to measure the precision and recall of each model in predicting successful first-stage landings.
- **Optimization Goal:** The primary methodological objective was to identify the specific model that best captured non-linear patterns in payload and orbit data to minimize cost-estimation risks.
- **GitHub URL:**
  - https://github.com/adisorn242/SpaceX-Presentation/blob/main/05-04-01%20SpaceX-Machine-Learning-Prediction-Part-5.ipynb

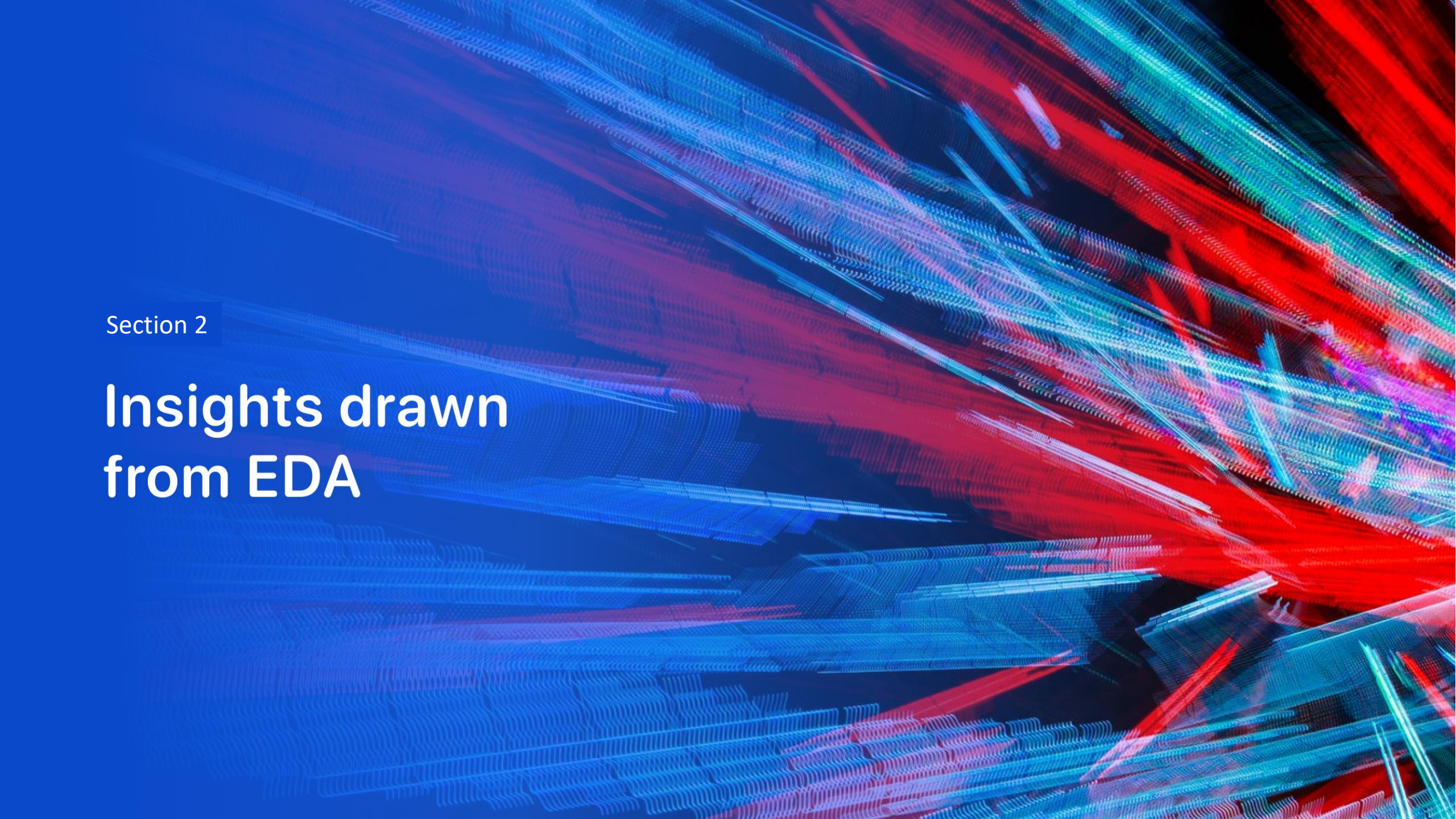


Results

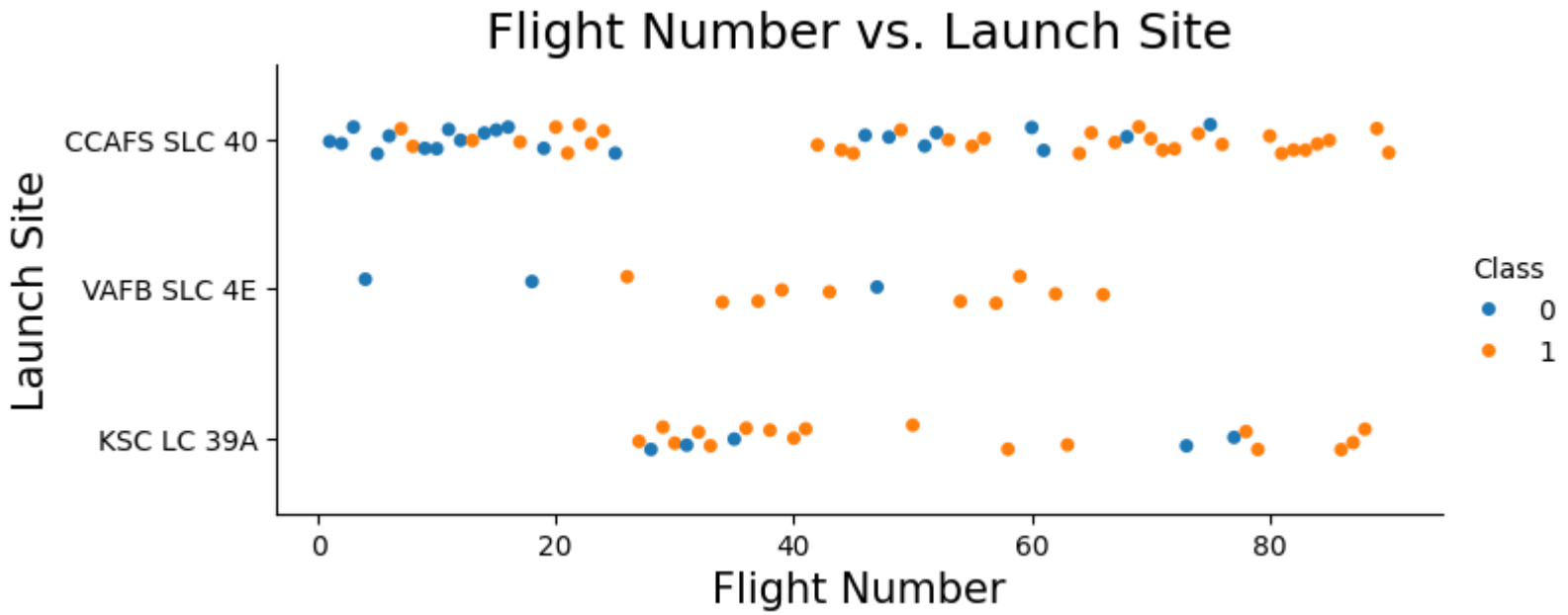
- **Exploratory Data Analysis (EDA)**
  - Launch Site Success: KSC LC-39A demonstrated the highest success rate at 76.9%, while CCAFS SLC-40 handled the most frequent and diverse mission profiles.
  - Learning Curve: Success rates improved significantly as Flight Numbers increased, reflecting the maturation of SpaceX's technology over time.
  - Payload Trends: Heavier payloads (over 8,000 kg) showed higher landing success rates, primarily at VAFB SLC-4E and KSC LC-39A
- **Interactive Analytics Demo**
  - Geospatial Insights: Folium maps visualized that all launch sites are strategically located near coastlines and at lower latitudes for safety and efficiency.
  - Dynamic Exploration: The dashboard identified that certain high-energy orbits, such as ES-L1, GEO, HEO, and SSO, achieved a perfect 100% success rate.
  - Model Accuracy: Logistic Regression, SVM, and k-NN models achieved an identical test accuracy of 83.33% while Decision Trees achieved the highest accuracy at 94.44%.

Results

- **Predictive Analysis Results**
  - Model Optimization: All classification algorithms (Logistic Regression, SVM, Decision Tree, and k-NN) were trained and refined using GridSearchCV to ensure optimal hyperparameter tuning.
  - Performance Baseline: The Logistic Regression, SVM, and k-NN models achieved a consistent and identical test accuracy of 83.33%.
  - Superior Model: The Decision Tree emerged as the best-performing algorithm, outperforming the other models with a test accuracy of 94.44%.
  - Predictive Precision: The confusion matrix for the Decision Tree demonstrated high reliability, correctly identifying 12 successful landings in the test set.
  - High Recall: The top-performing model achieved zero false negatives, meaning it never failed to predict a successful landing when one actually occurred.

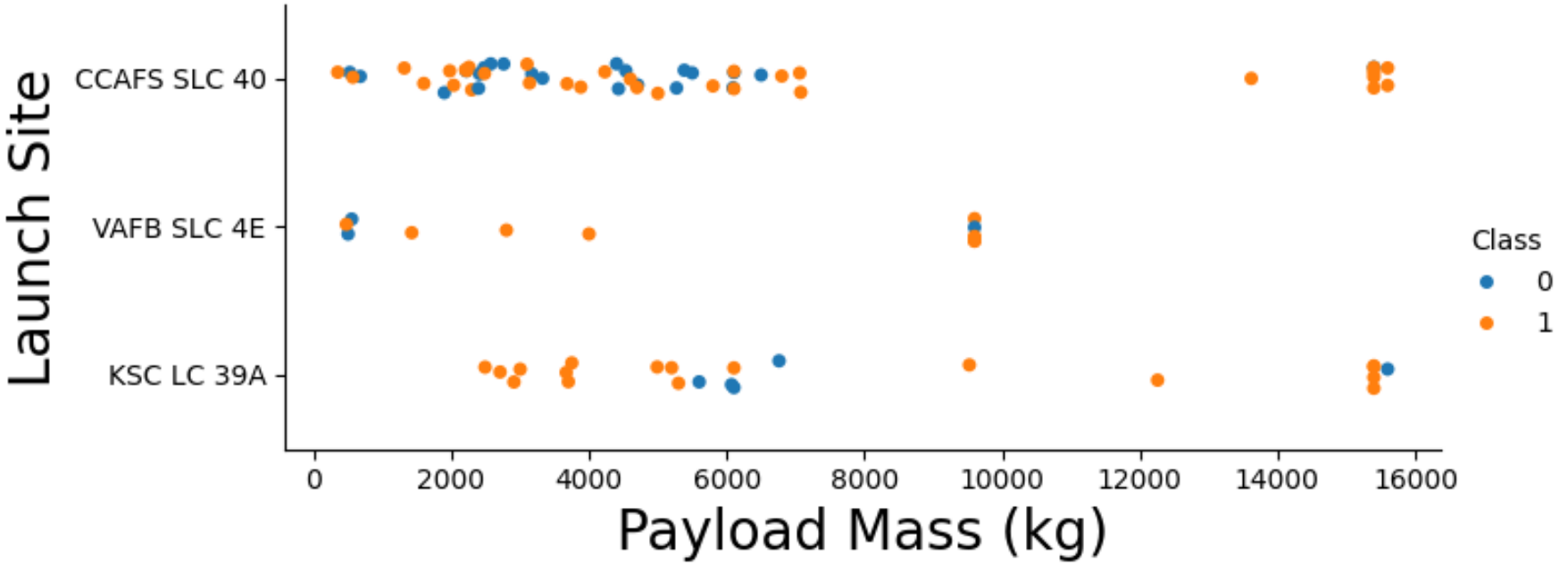

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- As flight numbers increase, landing success rates significantly improve across all sites, particularly at KSC LC-39A.
- Site Trends: CCAFS SLC-40 handled early experimental phases, while later missions shifted toward KSC LC-39A with higher consistency.

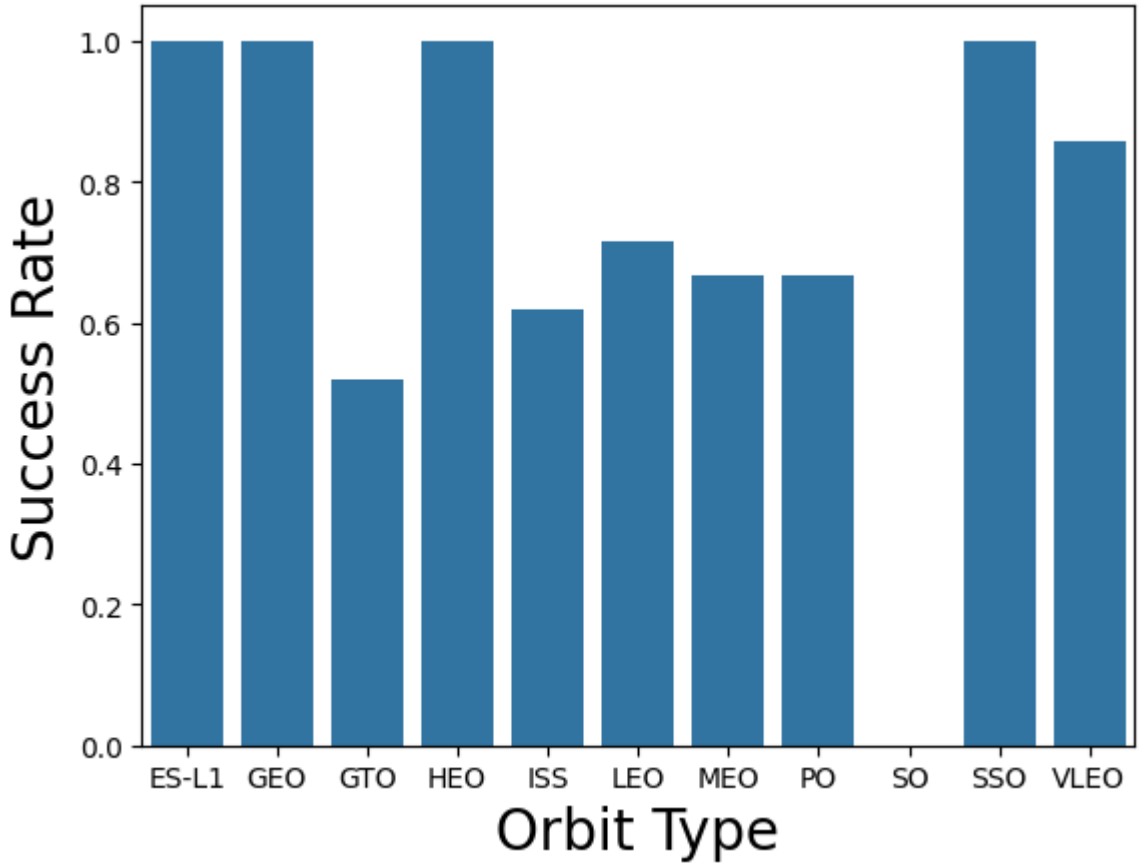
Payload vs. Launch Site



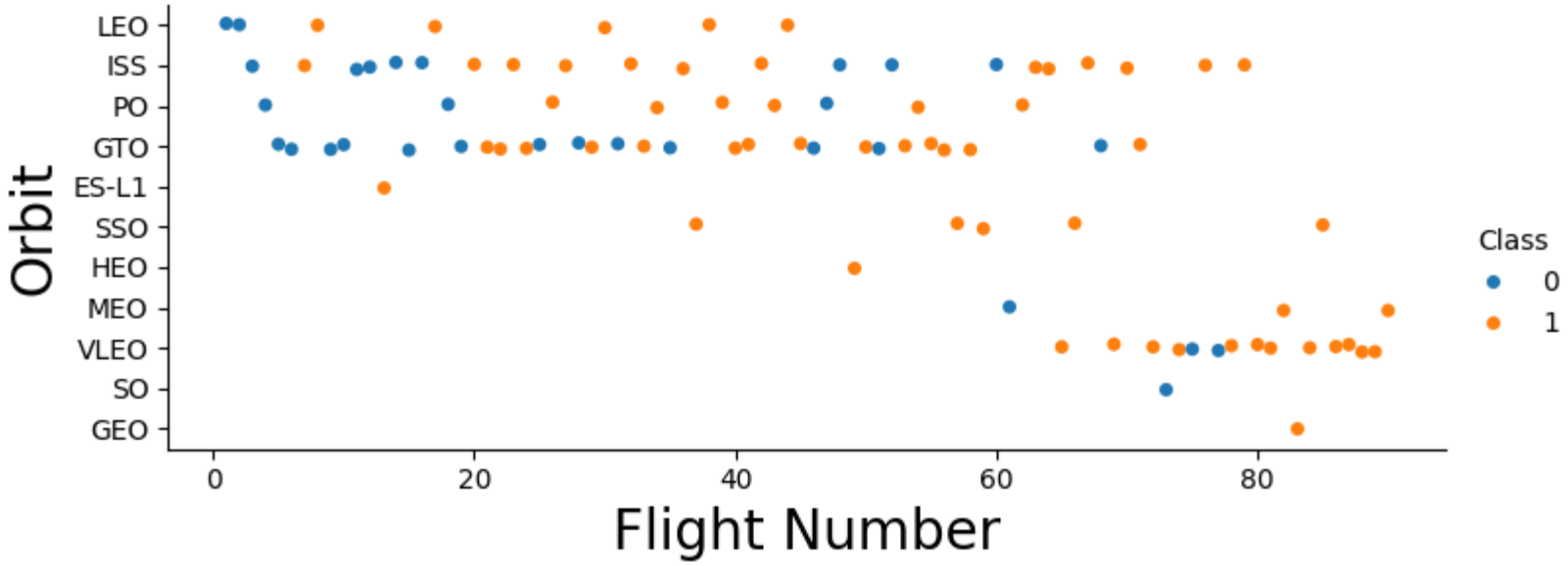
- **Observation:** Higher payload masses (over 8,000 kg) show a significantly higher landing success rate, particularly at **KSC LC-39A** and **VAFB SLC-4E**.
- **Site Trends: CCAFS SLC-40** manages a diverse range of payloads, while heavier missions are more concentrated at **KSC LC-39A** with high consistency.

Success Rate vs. Orbit Type

- **Observation:** The bar chart shows that orbits ES-L1, GEO, HEO, and SSO achieved a perfect 100% success rate.
- **Performance Analysis:** Orbits like GTO and ISS show lower landing success rates (approx. 50-60%), while VLEO demonstrates high reliability above 80%.

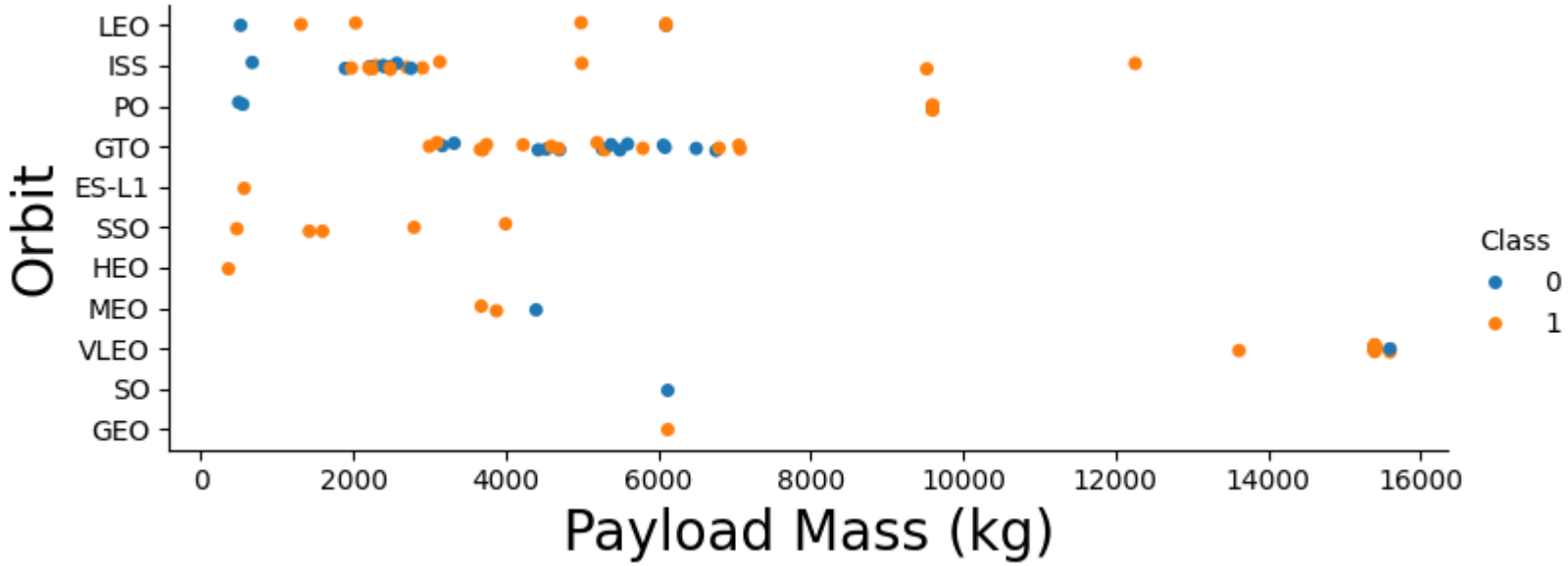


Flight Number vs. Orbit Type



- **Observation:** Early missions were primarily focused on LEO, ISS, and GTO orbits with inconsistent landing results.
- **Evolution:** Later missions show a clear shift towards VLEO orbits, coinciding with a significant increase in successful landings (Class 1).

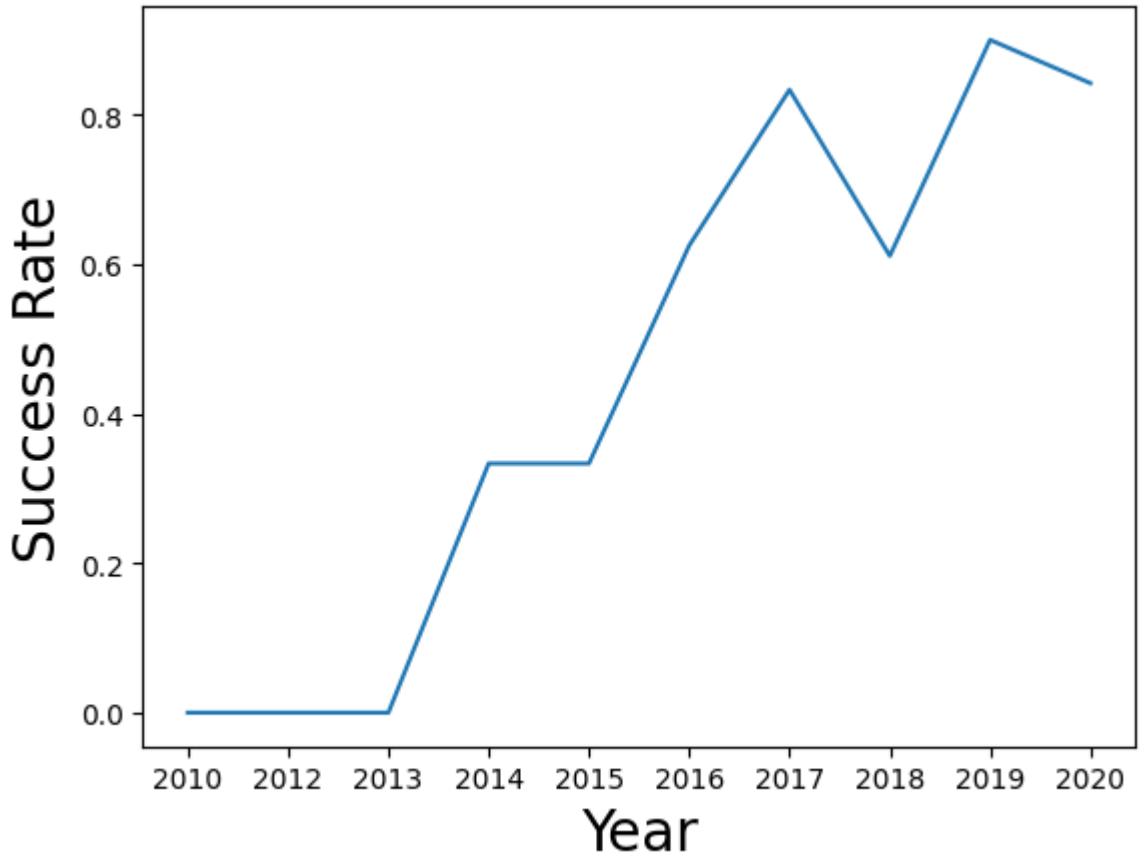
Payload vs. Orbit Type



- **Observation:** Payloads destined for **VLEO** and **ISS** orbits show high landing success rates across varying masses.
- **Correlation:** Missions to **GTO** orbits are concentrated between 3,000 kg and 6,000 kg, with a mix of successful and failed landings.
- **Performance:** Success is less dependent on payload mass for low-altitude orbits (LEO, SSO) compared to high-energy transfer orbits.

Launch Success Yearly Trend

- **Observation:** The line chart shows a clear upward trend in landing success rates, starting from 0% in 2013 and reaching approximately 85% by 2020.
- **Key Milestones:** Significant improvements are visible after 2015, with the success rate surpassing 60% in 2016 and peaking near 90% in 2019.
- **Analysis:** This positive trend reflects the rapid maturation of SpaceX's reusable rocket technology and mission reliability over the decade.



All Launch Site Names

- SQL Query:
  - Executed SELECT DISTINCT Launch_Site FROM SPACEXTBL to identify all unique mission departure points.
- Query Results: The analysis identified
  - CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.
- Explanation:
  - The results show the primary pads used across Florida and California, noting a slight naming variation for the Cape Canaveral site in the raw dataset.

Launch Site Names Begin with 'CCA'

]:

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

- **SQL Query:** Executed SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5 to retrieve the first five records from Cape Canaveral sites.
- **Observation:** All five initial records originated from CCAFS LC-40, carrying various Dragon spacecraft versions to LEO orbits between 2010 and 2013.
- **Outcome:** These early missions primarily resulted in "Failure (parachute)" or "No attempt" for landing outcomes, reflecting the experimental nature of SpaceX's early recovery program.

Total Payload Mass

- **Objective:** Calculate the total weight of all cargo successfully delivered to space specifically for the NASA (CRS) contract.
- **SQL Implementation:** Applied the SUM() aggregate function to the PAYLOAD_MASS__KG_ column, filtering by the customer name using a WHERE clause.
- **Finding:** The cumulative payload delivered for these missions totals 45,596 kg.
- **Significance:** This metric highlights the substantial logistics capacity SpaceX has provided to NASA, representing a core segment of their early operational success and revenue stream.

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- **Analysis Goal:** Calculated the mean payload weight specifically for missions utilizing the **F9 v1.1** booster version to establish a performance baseline.
- **Methodology:** Used the SQL AVG() aggregate function on the payload mass column while filtering for the specific booster version string in the database.
- **Key Finding:** The average payload mass successfully transported by this booster variant is 2,928.4 kg.
- **Significance:** This result identifies the typical operational load for this iteration, allowing for direct capacity comparisons against newer Falcon 9 upgrades.

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- **Analysis Goal:** Identified the historical milestone of the first successful rocket landing on a ground pad within the dataset.
- **Methodology:** Filtered the mission records by the landing outcome 'Success (ground pad)' and sorted the results chronologically to isolate the earliest occurrence.
- **Key Finding:** The first successful landing on a ground pad was achieved on **2015-12-22**.
- **Significance:** This date represents a major breakthrough in SpaceX's reusability program, proving that orbital-class boosters could be successfully returned to a land-based facility for refurbishment.

First_Success_Ground_Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Analysis Goal:** Identified booster versions that achieved successful drone ship landings while carrying mid-range payloads between 4,000 and 6,000 kg.
- **Methodology:** Filtered the database for the 'Success (drone ship)' landing outcome and restricted the results to the specified payload mass range.
- **Key Finding:** Four specific boosters met these criteria: **F9 FT B1022**, **F9 FT B1026**, **F9 FT B1021.2**, and **F9 FT B1031.2**.
- **Significance:** This identifies the "Full Thrust" (FT) booster generation as highly reliable for recovering heavier orbital payloads on autonomous spaceport drone ships.

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

Total Number of Successful and Failure Mission Outcomes

- **Analysis Goal:** Aggregated the total counts for all mission outcomes recorded in the dataset to assess overall reliability.
- **Methodology:** Utilized the SQL GROUP BY clause on the Mission_Outcome column combined with the COUNT(*) aggregate function to categorize every flight entry.
- **Key Finding:** The query identified 98 clear "Success" outcomes, with minor variations including 1 "Failure (in flight)" and 1 "Success (payload status unclear)".
- **Significance:** This high success-to-failure ratio highlights SpaceX's operational maturity, with the vast majority of missions meeting their primary orbital delivery objectives.

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

Boosters Carried Maximum Payload

- **Analysis Goal:** Identified the specific booster versions that have demonstrated the highest heavy-lift capability by carrying the maximum recorded payload mass in the dataset.
- **Methodology:** Executed a SQL query to filter the Booster_Version records where the PAYLOAD_MASS__KG_ matched the absolute maximum value found across the entire table.
- **Key Finding:** Several boosters, primarily from the **F9 B5** (Block 5) generation, successfully carried the maximum payload, including **B1048.4**, **B1049.4**, **B1051.3**, **B1056.4**, and others in the series.
- **Significance:** These results highlight the maturity of the Falcon 9 Block 5 architecture, showcasing its consistent ability to launch heavy mission profiles while maintaining high operational frequency.

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

- **Analysis Goal:** Identified specific mission failures during the early experimental phase of drone ship landings in **2015**.
- **Methodology:** Filtered the database for the year **2015** and isolated records with a **"Failure (drone ship)"** landing outcome, extracting the month, booster version, and launch site.
- **Key Findings:** Two notable failures occurred at **CCAFS LC-40**:
  - **January (Month 01):** Booster **F9 v1.1 B1012**.
  - **April (Month 04):** Booster **F9 v1.1 B1015**.
- **Significance:** These records document the iterative "trial and error" process SpaceX underwent to perfect autonomous drone ship recovery technology.

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

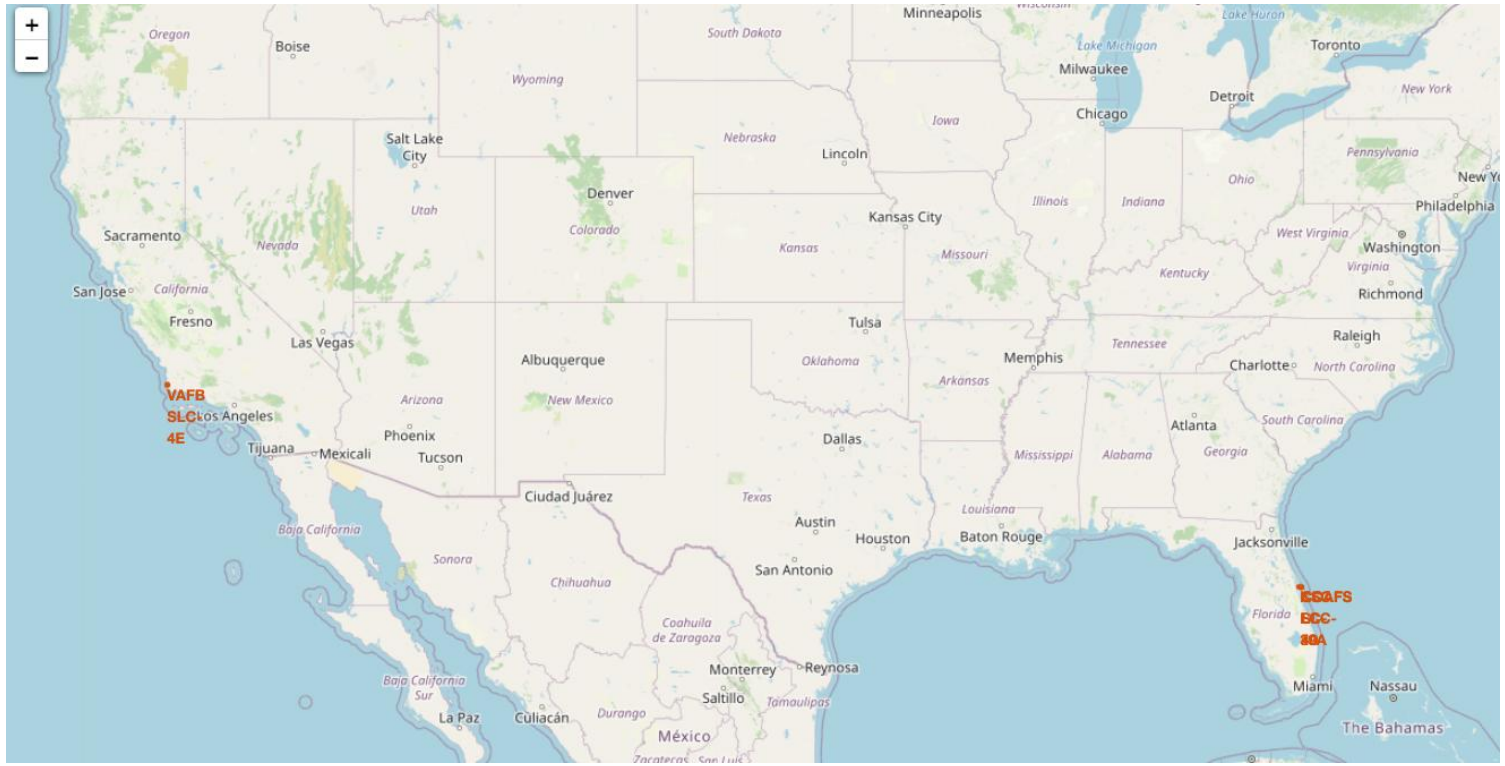
- **Analysis Goal:** Ranked the frequency of different landing outcomes within a specific historical timeframe to identify the most common mission results.
- **Methodology:** Aggregated landing data between **2010-06-04** and **2017-03-20**, sorting the counts in descending order to highlight dominant trends.
- **Key Finding:** "No attempt" was the most frequent outcome (10), followed by an equal number of "Success (drone ship)" and "Failure (drone ship)" at 5 each.
- **Significance:** These results reflect the transition from early non-recovery missions to the high-stakes testing phase of the autonomous drone ship landing system.

	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

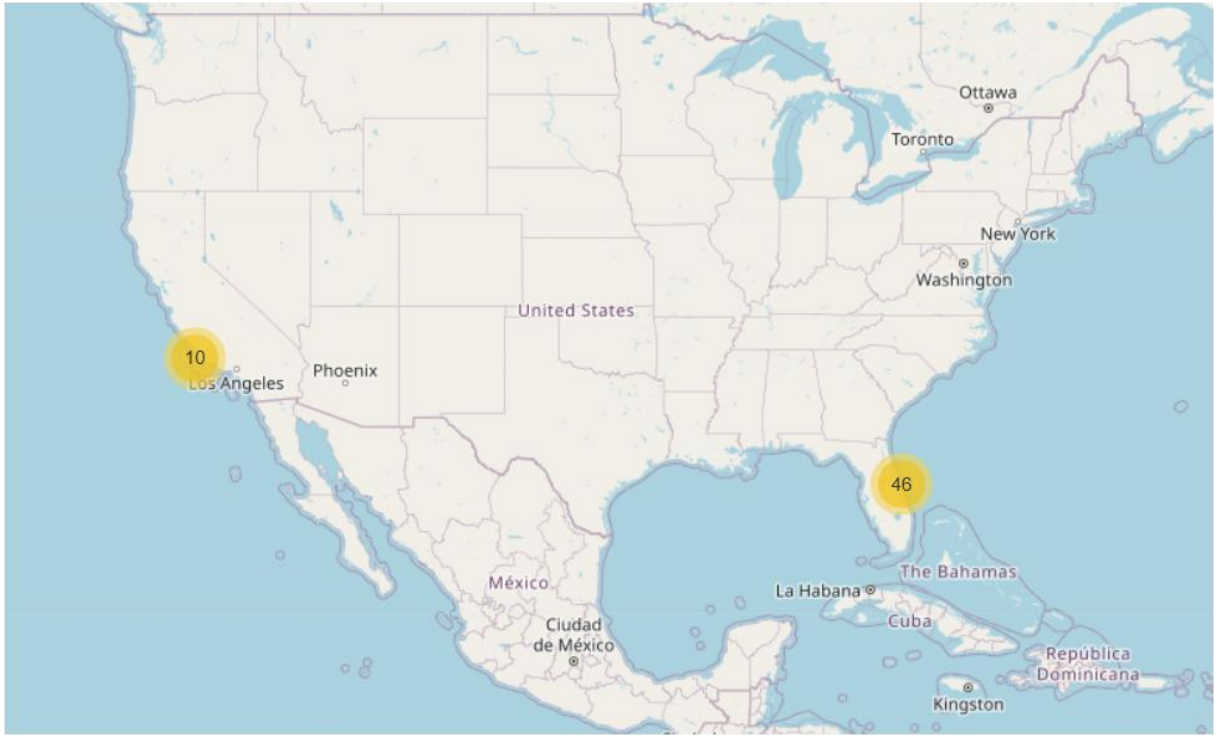
Launch Sites Proximities Analysis

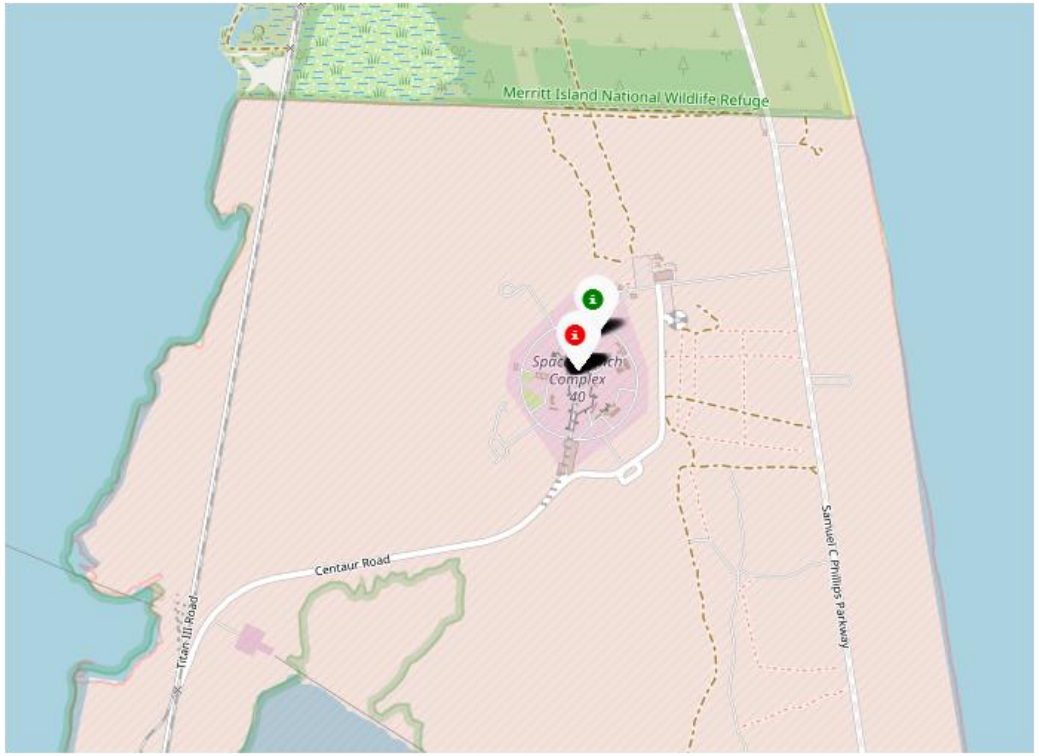
SpaceX Launch Sites



- **Observation:** Interactive mapping reveals all SpaceX launch sites are strategically located along coastlines to ensure safe flight paths over open water.
- **Analysis:** The proximity to the equator for Florida sites leverages Earth's rotation to maximize payload efficiency during orbital injection.

The success/failed launches for each site

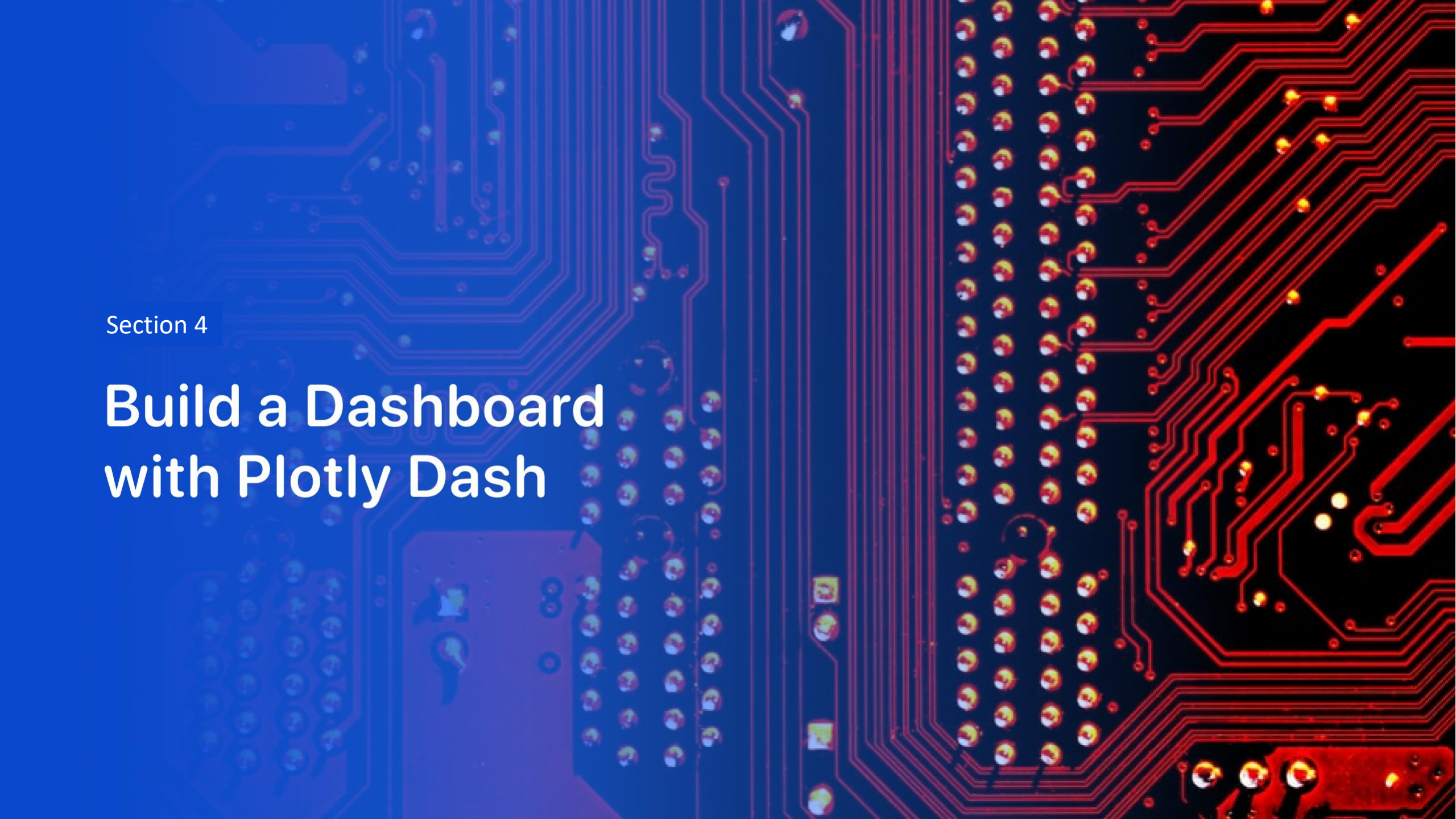




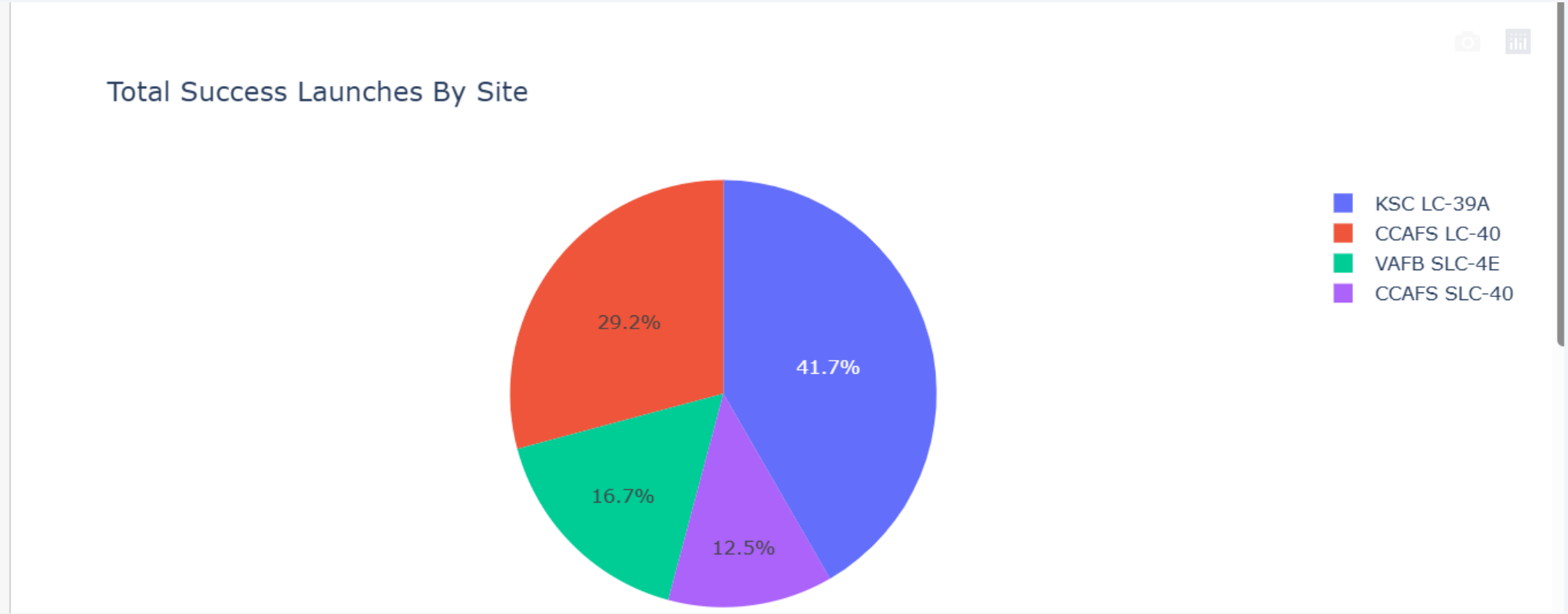
- The interactive map shows that major sites like **KSC LC-39A** and **CCAFS SLC-40** achieved higher landing success rates as operational experience increased.

Calculate the distances between a launch site to its proximities

- Proximity analysis shows that launch sites are strategically located near coastlines to minimize risk and optimize logistics for rocket stage recovery.

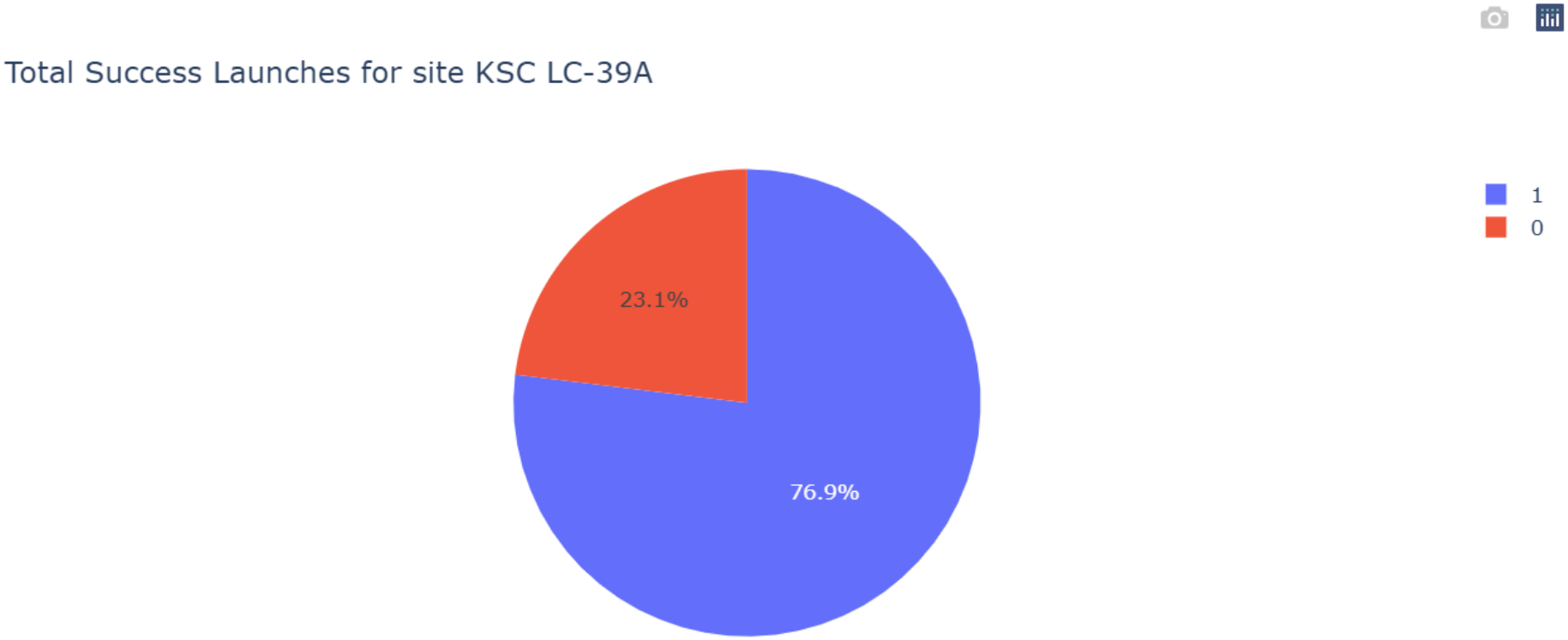

Section 4
Build a Dashboard with Plotly Dash

Total Launch Success Count by Site



- KSC LC-39A leads with 41.7% of total successful launches, while CCAFS LC-40 and VAFB SLC-4E contribute 29.2% and 16.7% respectively.

Launch with Highest Success Ratio: KSC LC-39A



- The filtered dashboard shows that **KSC LC-39A** achieved a high success rate of **76.9%**, significantly outperforming failed attempts at this primary hub.

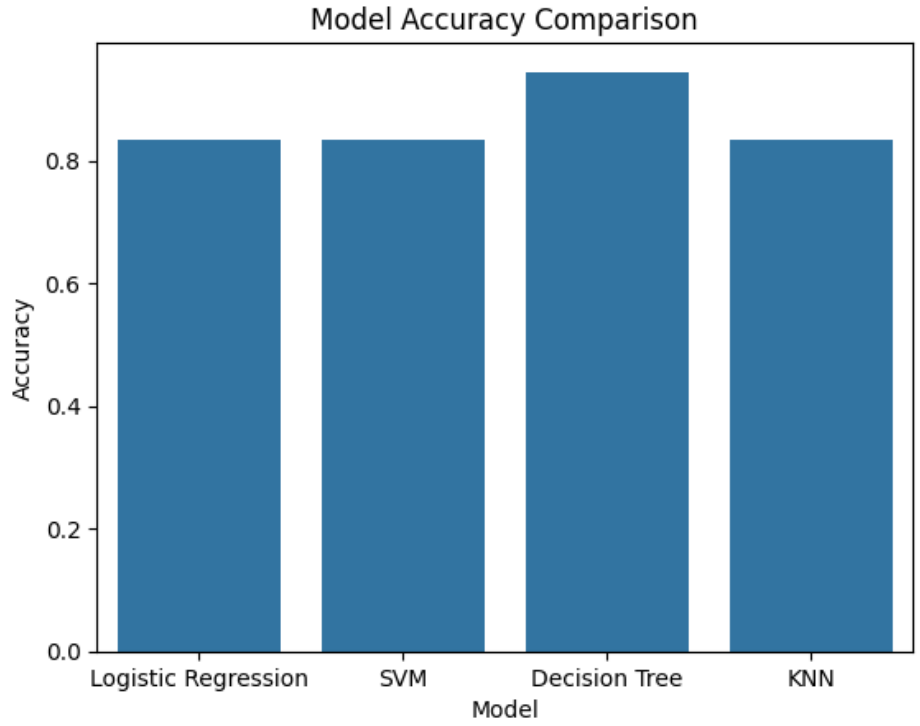
Correlation Between Payload and Success for All Sites

- The scatter plot for all launch sites shows that successful missions are most densely concentrated in the 2,000 kg to 6,000 kg payload range, with the FT booster version achieving the highest volume of successes.

Section 5

Predictive Analysis (Classification)

Classification Accuracy: Model Comparison on Test set

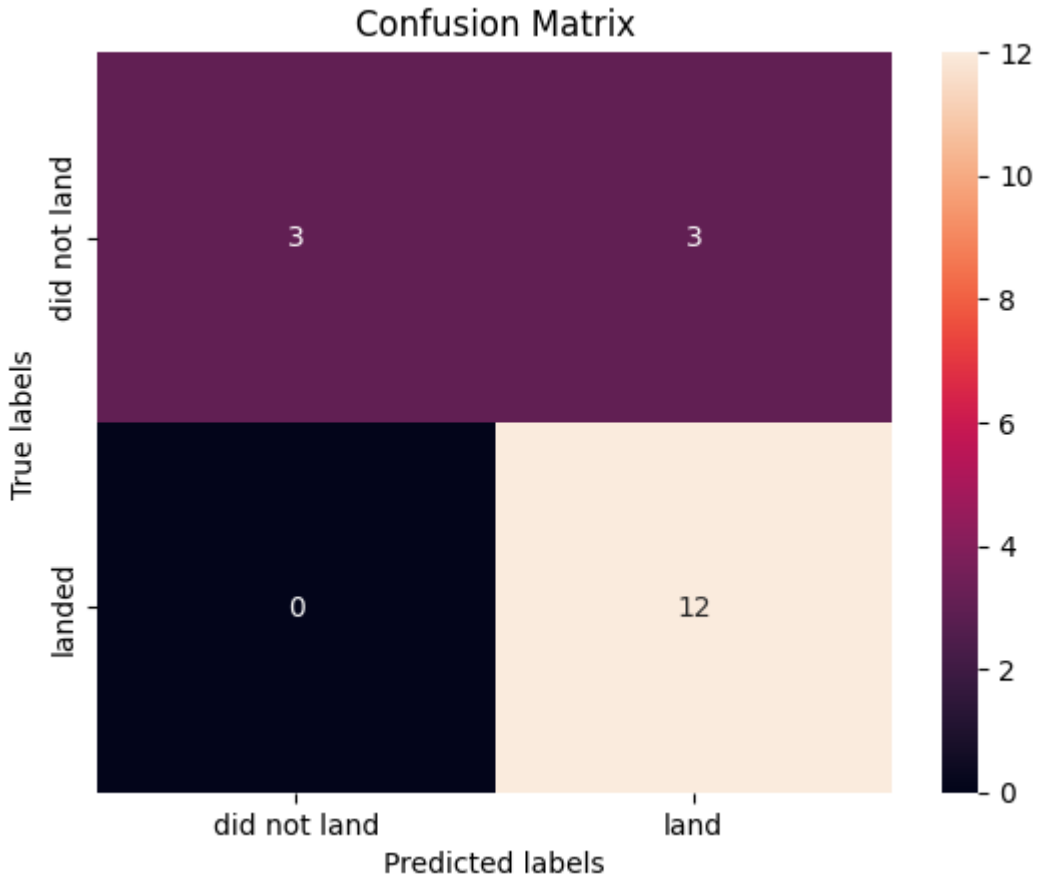


	Model	Accuracy
0	Logistic Regression	0.8333
1	SVM	0.8333
2	Decision Tree	0.9444
3	KNN	0.8333

- The Decision Tree is the most accurate model, outperforming the others by 11%. While linear and distance-based models reached a performance ceiling at 83.33%, the tree-based approach better captured the complex patterns in the data.

Confusion Matrix: Decision Trees (Best)

- **Observation:** The confusion matrix for the Decision Tree model demonstrates high reliability, correctly identifying **12** successful landings and **3** failures.
- **Insight:** The model achieved zero false negatives, meaning it never failed to predict a successful landing, though it produced **3** false positives.
- **Performance:** The Decision Tree was the best-performing model among those tested, achieving an overall accuracy of **0.9444**.



Key Analytical Insights

- **Site Dominance:** KSC LC-39A is the most successful individual launch site, accounting for 41.7% of all successful landings.
- **Performance Baseline:** When filtered specifically, KSC LC-39A demonstrates a high success-to-failure ratio of 76.9%.
- **Success Corridor:** Successful missions are most densely concentrated in the 2,000 kg to 6,000 kg payload range.
- **Technology Maturation:** Landing success rates showed a clear upward trend over time, rising from 0% in 2013 to approximately 85% by 2020.
- **Booster Reliability:** The FT (Full Thrust) booster version achieved the highest volume of successful recoveries, particularly for mid-range payloads.

Final Conclusions

- **Operational Scale:** SpaceX has demonstrated high-capacity logistics, delivering a cumulative payload of 45,596 kg for NASA (CRS) missions alone.
- **Strategic Geography:** Mapping confirms launch sites are optimized for safety and efficiency, positioned near coastlines to leverage Earth's rotation.
- **Predictive Superiority:** Among all evaluated models, the Decision Tree emerged as the best predictor for landing success.
- **Model Accuracy:** The Decision Tree achieved a superior test accuracy of 94.44%, outperforming linear and distance-based models by 11%.
- **High Precision:** The best-performing model's confusion matrix highlighted zero false negatives, correctly identifying all 12 successful landings in the test set.

Appendix

- GitHub : https://github.com/adisorn242/SpaceX-Presentation/tree/main

Thank you!