IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Miguel Domingos Marques
22/05/2025

Outline

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

Executive Summary

Summary of methodologies:

Data collection
Data wrangling
Exploratory Data Analysis with Data Visualization
Exploratory Data Analysis with SQL
Building an interactive map with Folium
Building a Dashboard with Plotly Dash
Predictive analysis (Classification)

Summary of all results:

Exploratory Data Analysis results
Interactive analytics demo in screenshots
Predictive analysis results

Introduction

Project Background and Context

SpaceX has revolutionized commercial spaceflight by significantly lowering launch costs. Their Falcon 9 rocket is advertised at around $62 million per launch, compared to competitors charging upwards of $165 million. A key factor behind this cost reduction is the company's ability to recover and reuse the first stage of the rocket. Accurately predicting whether the first stage will successfully land can directly impact launch cost estimations.

Using publicly available data combined with machine learning techniques, this project aims to forecast the probability of a successful first stage landing for Falcon 9 rockets.

Key Questions to Explore

- What impact do variables such as payload mass, launch site, flight count of the booster, and orbit type have on the likelihood of a successful first stage landing?
- Has the frequency of successful landings improved over time, indicating enhanced reliability?
- Among various machine learning models, which method provides the most effective binary classification performance for predicting first stage recovery?

Section 1

Methodology

Methodology

Executive Summary

Data Collection and Processing Approach

- Acquired data through the SpaceX RESTful API to obtain detailed launch and rocket information.
- Supplemented data by scraping relevant launch and mission details from Wikipedia pages.

Data Preparation

- Cleaned and refined the dataset by filtering out irrelevant or inconsistent records.
- Addressed missing or incomplete data through appropriate imputation and handling techniques.
- Transformed categorical variables into a machine-readable format using One-Hot Encoding, facilitating binary classification.

Exploratory Analysis

- Conducted exploratory data analysis (EDA) combining visualizations and SQL queries to uncover underlying patterns and relationships.
- Developed interactive visual dashboards leveraging tools such as Folium for geospatial mapping and Plotly Dash for dynamic data exploration.

Predictive Modeling

- Constructed, optimized, and validated several classification algorithms to predict the success of the first stage landing.
- Applied hyperparameter tuning and rigorous evaluation metrics to identify the most accurate and reliable predictive model.

Data Collection

- The dataset was compiled through a combination of API calls to the SpaceX REST API and web scraping techniques targeting SpaceX's Wikipedia page. Utilizing both sources ensured a comprehensive and enriched dataset for a more thorough analysis of launch missions.

Data Fields from SpaceX REST API:

- Flight Number, Launch Date, Booster Version, Payload Mass (kg), Orbit Type, Launch Site Identifier, Mission Outcome, Number of Flights for Booster, Grid Fins Presence, Reusability Status, Number of Landing Legs, Landing Pad Designation, Block Number, Count of Reuses, Booster Serial Number, Launch Longitude, Launch Latitude

Data Fields Extracted from Wikipedia Scraping:

- Flight Number,Launch Site Name, Payload Description, Payload Mass (kg), Orbit Category, Customer Name, Launch Result, Booster Version Details, Landing Outcome of Booster, Launch Date, Launch Time

Data Collection – SpaceX API

- **Data Retrieval and Processing Steps:**
- **1º** Sent requests to the SpaceX API to fetch rocket launch data.
- **2º** Parsed the API response by decoding it with `.json()` and converted the nested JSON data into a flat table using `json_normalize()`.
- **3º** Extracted specific launch details by applying custom functions tailored to retrieve only relevant fields from the API response.
- **4º** Organized the extracted information into a structured dictionary format.
- **5º** Created a pandas DataFrame from this dictionary for easier data manipulation.
- **6º** Filtered the DataFrame to retain only Falcon 9 launches for focused analysis.
- **7º** Addressed missing values in the `Payload Mass` column by imputing them with the mean value of that column.
- **8º** Finally, exported the cleaned and filtered dataset to a CSV file for subsequent analysis.
- GitHub URL: Data Collection API

Data Collection - Scraping

- **1º** Sent a request to the Wikipedia page containing Falcon 9 launch information.
- **2º** Parsed the retrieved HTML content using BeautifulSoup to create a parse tree.
- **3º** Extracted column headers by locating and reading the table header elements (<th>) within the HTML.
- **4º** Parsed the relevant HTML table rows to collect launch data from each cell.
- **5º** Compiled the extracted data into a structured dictionary format.
- **6º**Converted the dictionary into a pandas DataFrame for easy data handling.
- **7º** Saved the finalized DataFrame as a CSV file for further use.
- GitHub URL: WebScraping

Data Wrangling

The dataset contains various scenarios where the booster's landing attempt either succeeded or failed. For instance:

Ocean landings:

- *True Ocean* indicates a successful splashdown in a designated ocean region.
- *False Ocean* indicates a failed ocean landing attempt.

Return To Launch Site (RTLS) landings:

- *True RTLS* means the booster successfully landed back on a ground pad near the launch site.
- *False RTLS* means the booster failed to land at the launch site.

Autonomous Spaceport Drone Ship (ASDS) landings:

- *True ASDS* means a successful landing on the drone ship at sea.
- *False ASDS* means the landing attempt on the drone ship was unsuccessful.
- To prepare the data for machine learning, these detailed outcomes are consolidated into a binary training label:

Label "1" — booster landed successfully (any True outcome)

Label "0" — booster landing was unsuccessful (any False outcome)

Exploratory Data Analysis (EDA) and Labeling

- **1º** Calculate the total number of launches at each launch site.
- **2º** Determine the frequency and distribution of different orbit types.
- **3º** Analyze mission outcomes grouped by orbit category to identify patterns.
- **4º** Generate a binary landing outcome label derived from the `Outcome` column to be used for predictive modeling.
- **5º** Export the processed dataset with labels to a CSV file for further analysis.
- GitHub URL:Data Wrangling

EDA with Data Visualization

Types of Charts and Their Purposes:

Flight Number vs. Payload Mass

Flight Number vs. Launch Site

Payload Mass vs. Launch Site

Flight Number vs. Orbit Type

Payload Mass vs. Orbit Type and Success Rate

Scatter Plots:

These plots help reveal relationships between continuous variables and identify trends or groupings that might influence landing success. Such relationships can inform feature selection for machine learning models.

- **Bar Charts:**

  Used to compare the success rates of different orbit types and launch sites. Helpful for visualizing categorical variables like launch site or orbit and their impact on landing outcomes.

- **Line Charts:**

These plots were used to illustrate how landing success rates have evolved over the years, providing a time-series view of performance improvements. Each of these visualizations played a critical role in understanding the structure and dynamics of the dataset prior to model development.

GitHub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

- **Identifying Unique Launch Locations**
  Queried the dataset to display the distinct names of all launch sites used in the missions.
- **Filtering Launch Sites**
  Retrieved the first five records of launch sites that begin with the prefix `'CCA'`.
- **Payload Analysis for NASA Missions**
  Calculated the total payload mass carried specifically by missions launched under the *NASA (CRS)* program.
- **Booster Version-Specific Payload Metrics**
  Computed the average payload mass carried by boosters of the version *F9 v1.1*.
- **First Ground Pad Success**
  Identified the date of the earliest successful landing on a ground pad (RTLS), marking a significant milestone in reusability.
- **Drone Ship Landings with Specific Payload Range**
  Listed the boosters that landed successfully on drone ships and carried payloads between 4000 and 6000 kg.
- **Overall Mission Outcomes Summary**
  Aggregated the total count of both successful and failed mission outcomes for a high-level overview.
- **Maximum Payload Booster Versions**
  Identified which booster versions carried the highest payload masses across all launches.
- **Drone Ship Failures in 2015**
  Listed failed drone ship landings during 2015 along with their respective booster versions and launch sites.
- **Landing Outcome Ranking**
  Ranked different types of landing outcomes (e.g., *Failure (drone ship)*, *Success (ground pad)*) in descending order, limited to missions between *June 4, 2010* and *March 20, 2017*.

- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Map Objects Added:

Markers:
Placed on all SpaceX launch sites and the NASA Johnson Space Center, including popup and text labels to identify each location.

Circles:
Added around each marker to emphasize the launch site areas and enhance visibility on zoomed-out views.

Colored Markers:
Green for successful landings and red for failures, added using `MarkerCluster` to group them by location and simplify analysis.

Lines (Polylines):
Drawn from the KSC LC-39A launch site to nearby infrastructure: railway, highway, coastline, and closest city to visualize proximity and logistic accessibility.

Purpose of These Objects:

- To **highlight the geographical distribution** of SpaceX launch sites.
- To **distinguish mission outcomes visually** using color-coded markers.
- To **analyze success rates by site** via clustering.
- To **assess infrastructure proximity** for logistic and strategic insight.

- GitHub URL: VisualAnalytics with Folium

Build a Dashboard with Plotly Dash

- Summary of Plots and Interactions in the Dashboard
- **1. Launch Sites Dropdown List (Interaction)**
- Allows users to filter the dashboard data by selecting a specific launch site or viewing data for all sites.
- Launch performance can vary significantly by site. This control enables focused analysis on individual launch sites or comparison across all sites.
- **2. Pie Chart of Launch Successes (Plot)**
- Shows the distribution of successful launches. When "All Sites" is selected, it displays total successful launches per site; for a specific site, it shows the ratio of successes vs. failures.
- Pie charts provide an immediate visual summary of success rates, making it easy to compare site performance or understand launch reliability at a specific site.
- **3. Payload Mass Range Slider (Interaction)**
- Lets users select a range of payload masses to filter the data.
- Payload mass is an important factor that may impact launch success. This slider allows investigation of how different payload ranges correlate with success rates.
- **4. Scatter Chart of Payload Mass vs. Launch Success (Plot)**
- Visualizes the relationship between payload mass and launch success, with points colored by booster version category.
- This scatter plot helps identify patterns or trends—such as whether heavier payloads or certain booster versions influence launch outcomes—providing deeper insight into launch dynamics.
- GitHub URL: spacex dash app

Predictive Analysis (Classification)

- **1º Extract Target Variable**
  - Create a NumPy array from the `"Class"` column in the dataset.
  - This array serves as the labels for classification.
- **2º Data Standardization**
  - Use `StandardScaler` to standardize feature data (zero mean, unit variance).
  - Fit the scaler to the data and transform it to normalize the feature scales.
- **3º Train-Test Split**
  - Split the dataset into training and testing sets using `train_test_split`.
  - This allows for model training and unbiased evaluation on unseen data.
- **4º Model Selection with GridSearchCV**
  - Apply `GridSearchCV` with 10-fold cross-validation (`cv=10`) to tune hyperparameters.
  - Evaluate and optimize multiple classifiers: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors.
- **5º Model Evaluation**
  - Calculate accuracy scores on the test set using the `.score()` method for each model.
  - Generate confusion matrices to assess classification performance and error types.
- **6º Performance Metrics Comparison**
  - Compare models using metrics such as Jaccard similarity score and F1-score.
  - Identify the best performing model based on these comprehensive evaluation criteria.
- GitHub URL: Machine Learning Prediction

Results

Exploratory Data Analysis (EDA) Results

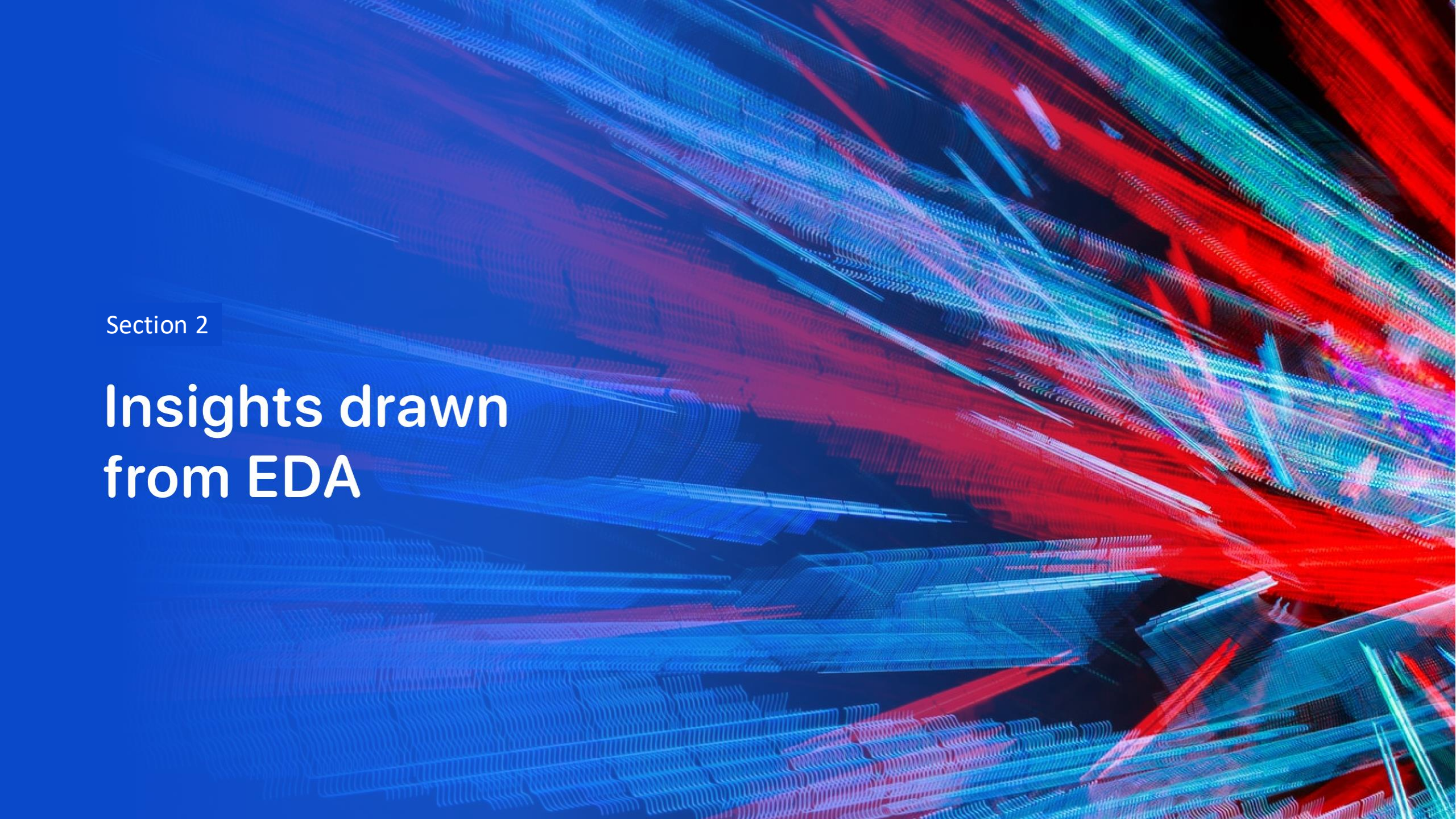
- **Summary Statistics:**
  Provided descriptive stats (mean, median, std) for key variables like payload mass, launch outcome, and booster versions.
- **Launch Site Distribution:**
  Visualized the number of launches per site, showing which launch sites are most active.
- **Success Rates:**
  Analyzed launch success rates overall and by site, highlighting sites with higher or lower success.
- **Payload Impact:**
  Explored how payload mass relates to launch success, revealing any notable correlations or patterns.
- **Booster Version Insights:**
  Examined how different booster versions performed in terms of success rates.

Interactive Analytics Demo (Screenshots)

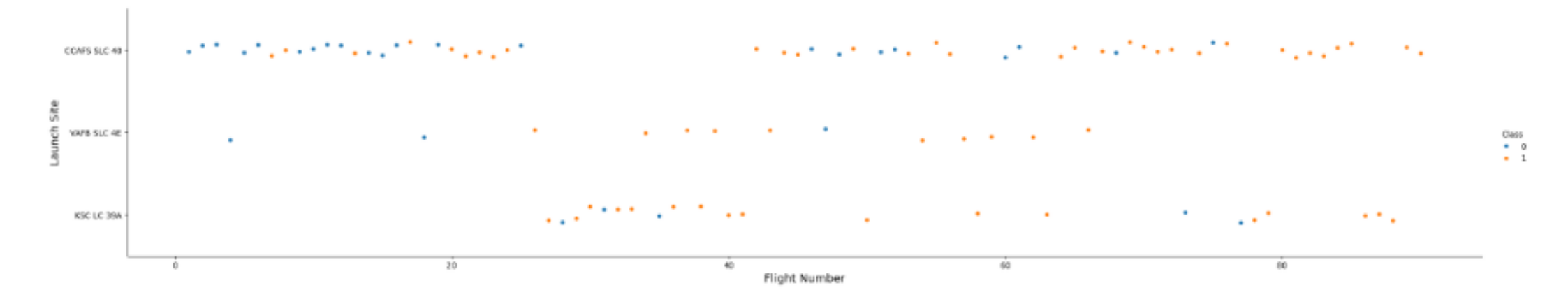
- **Launch Site Dropdown:**
  showing how selecting different launch sites updates the dashboard visuals dynamically.
- **Success Pie Chart:**
  Displays success distribution changes when switching between “All Sites” and specific launch sites.
- **Payload Range Slider:**
  Demonstrates filtering of launches by payload mass and its immediate effect on scatter plots.
- **Scatter Plot of Payload vs. Success:**
  Visualizes the correlation between payload and launch success, color-coded by booster version category.

Predictive Analysis Results

- **Models Tested:**
  Logistic Regression, Support Vector Machines, Decision Tree, and K-Nearest Neighbors with hyperparameter tuning using GridSearchCV.
- **Model Performance:**
  Accuracy scores on test data and confusion matrices were generated for each model.
- **Best Model Selection:**
  Compared using Jaccard similarity score and F1-score metrics. The model with the highest scores was identified as best performing.
- **Insights:**
  Predictive models can reliably classify launch success based on features such as payload mass and launch site, enabling data-driven launch planning.

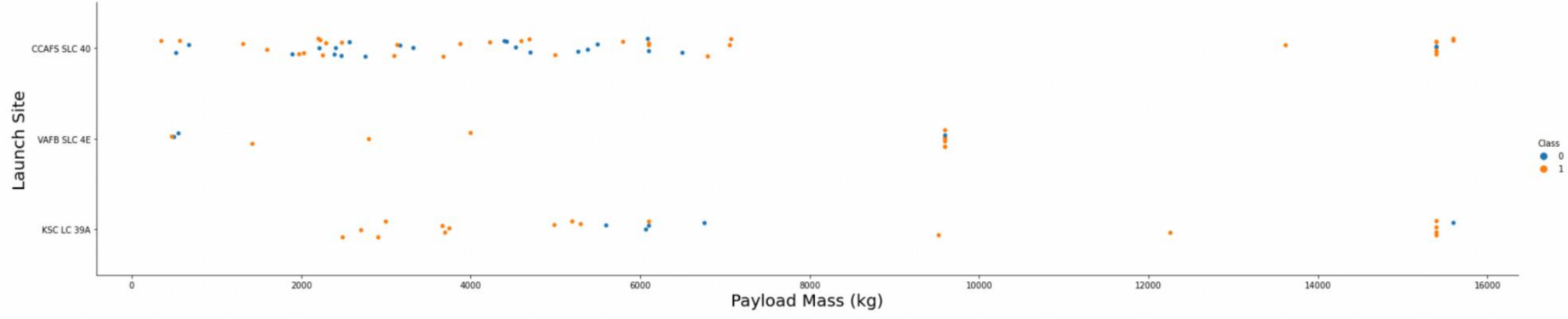

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



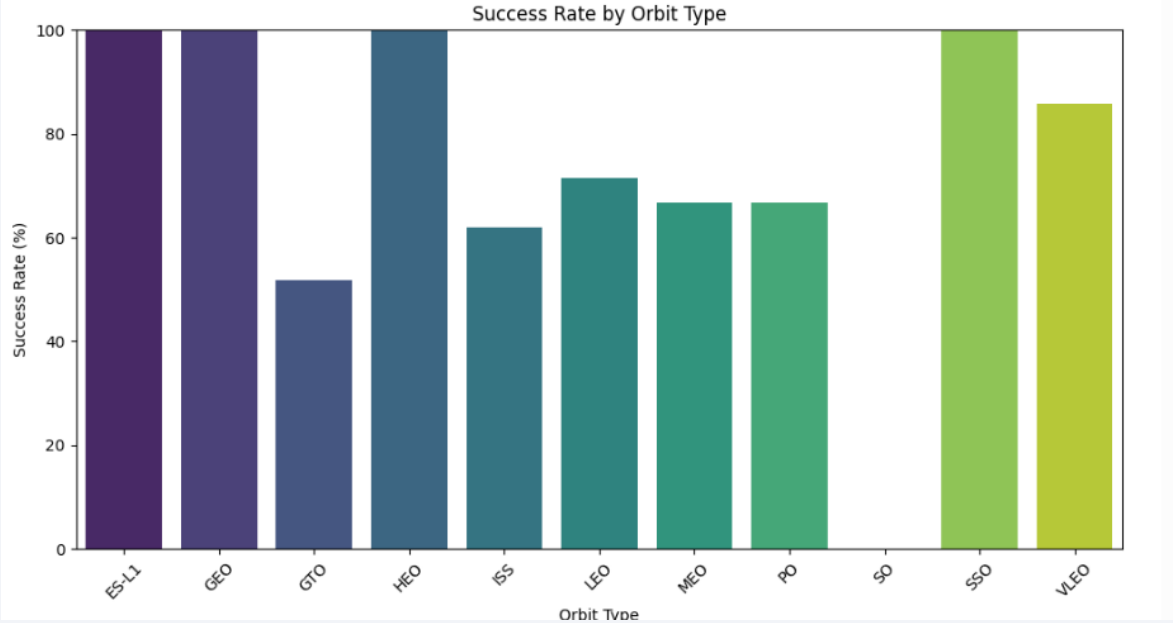
- **Early flights had failures, but recent flights are all successful.** This suggests improvements over time, like technology, experience, or procedures.
- **About half of all launches happen at CCAFS SLC 40.** This makes it a major launch site in the dataset.
- **VAFB SLC 4E and KSC LC 39A show higher success rates.** These sites appear more reliable or have better conditions for successful launches.
- **It's reasonable to assume newer launches are more likely to succeed.** Success rates improve as experience and tech advance.

Payload vs. Launch Site



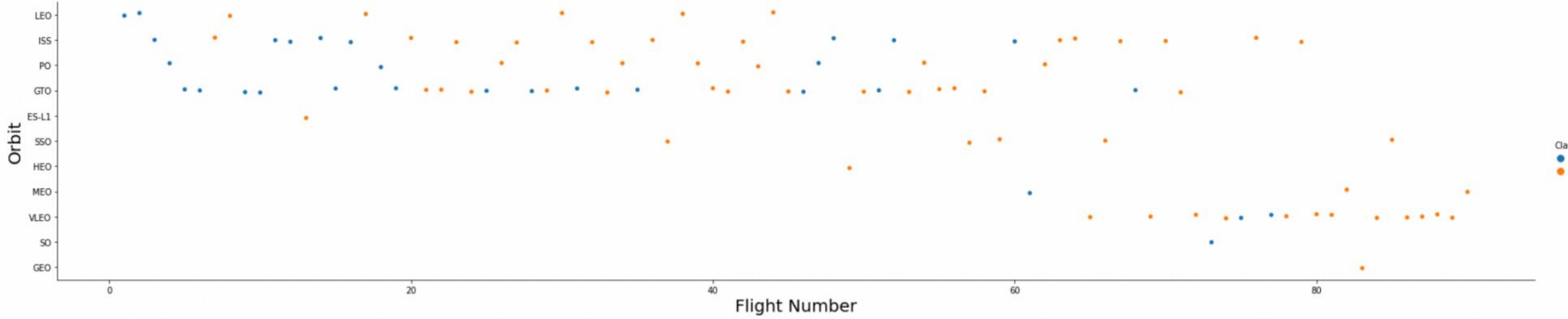
- **At every launch site, launches with higher payload mass tend to have higher success rates.** This indicates that heavier payloads might be prioritized or prepared more carefully.
- **Most launches carrying payloads over 7,000 kg were successful.** This suggests that very heavy payloads are reliably delivered.
- **KSC LC 39A has a perfect (100%) success rate even for payloads under 5,500 kg.** This shows exceptional performance at this site across different payload sizes.

Success Rate vs. Orbit Type



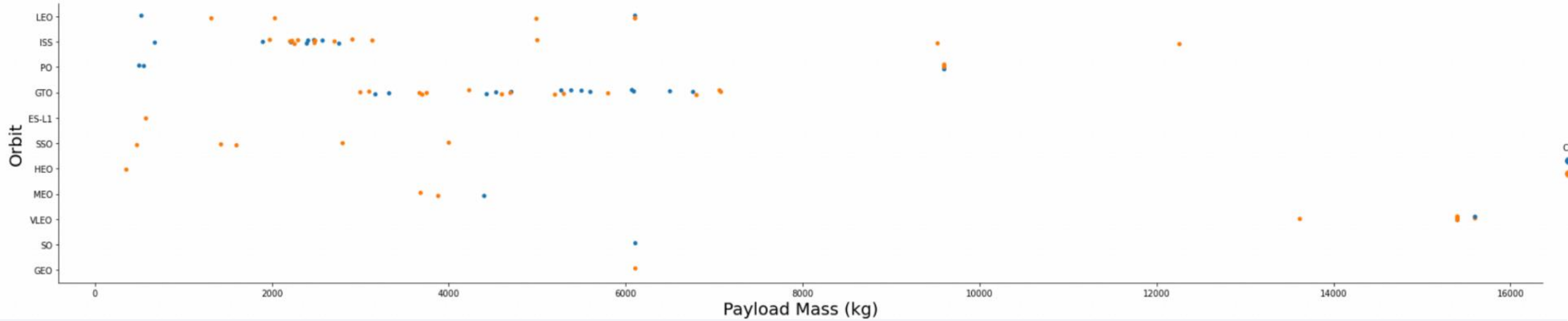
- **Orbits with a 100% success rate include ES-L1, GEO, HEO, and SSO.** Launches targeting these orbits have never failed in the data.
- **The SO orbit has a 0% success rate, meaning all attempts have failed.** This orbit seems particularly challenging or has very limited data.
- **Orbits with success rates between 50% and 85% include GTO, ISS, LEO, MEO, and PO.** These orbits show moderate to high reliability, but some failures still occur.

Flight Number vs. Orbit Type



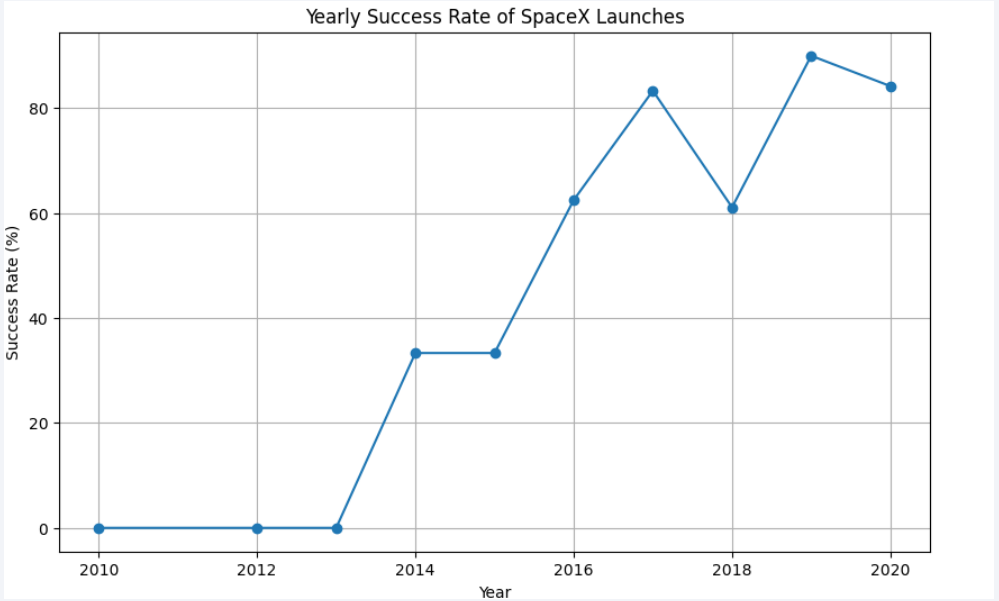
- **For launches targeting the LEO orbit, success seems to improve with the number of flights.** This suggests that as more missions occur, experience and learning lead to higher success rates in LEO.
- **In contrast, for the GTO orbit, there appears to be no clear link between the flight number and success rate.** This could mean that factors other than experience or flight count—like technical challenges specific to GTO—play a bigger role in success.

Payload vs. Orbit Type



- **Heavy payloads negatively affect success rates in GTO orbits.** This suggests that launching heavier payloads into GTO is more challenging, potentially lowering success rates.
- **In contrast, heavy payloads have a positive influence on success rates in ISS orbits.** For these orbits, heavier payloads may be associated with better-prepared missions or more reliable launches.

Launch Success Yearly Trend



- **The success rate of launches has been steadily increasing from 2013 through 2020.** This indicates continuous improvements in technology, processes, and experience during this period, leading to more reliable missions over time.

All Launch Site Names

]:

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

- Displaying the names of the unique launch sites in the space mission

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

Displaying 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

total_payload_mass
45596

Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

avg_payload_mass
2928.4

- Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

min(date)
2015-12-22

- Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

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

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

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

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

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

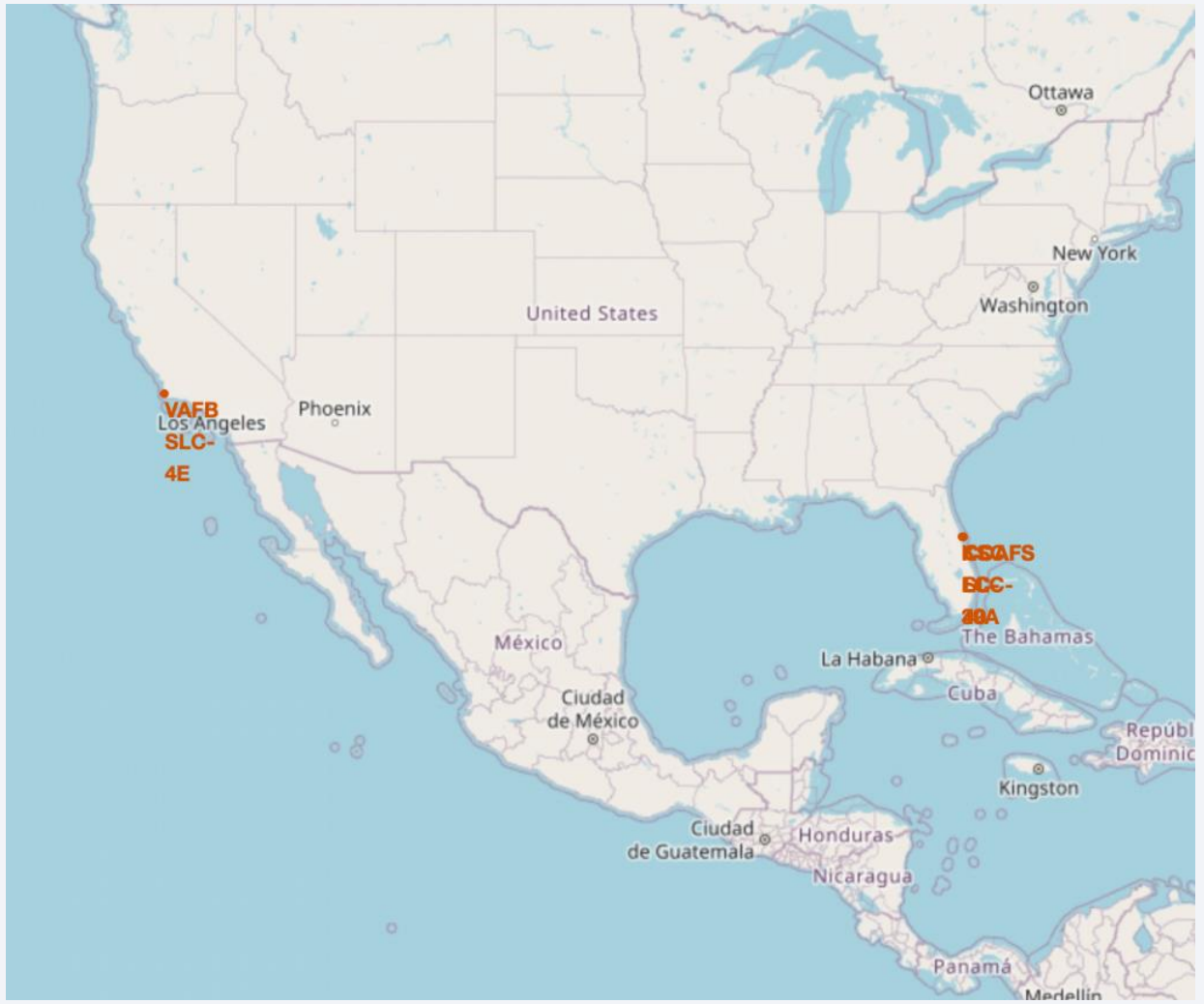
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

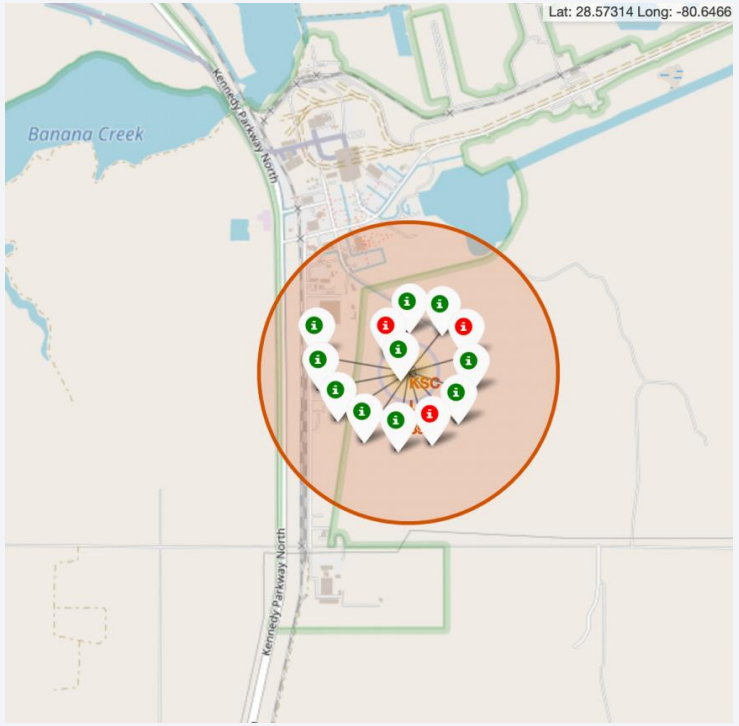
All launch sites' location markers on a global map

- **Most launch sites are near the equator** because the Earth's rotational speed is highest there—about 1670 km/h. This speed, due to inertia, gives rockets an extra velocity boost when launching eastward, making it easier and more efficient to reach orbit.
- **Launch sites are also located near coasts** to ensure rockets can be launched over open ocean. This greatly reduces the risk to human life and property in case of launch failures or falling debris.



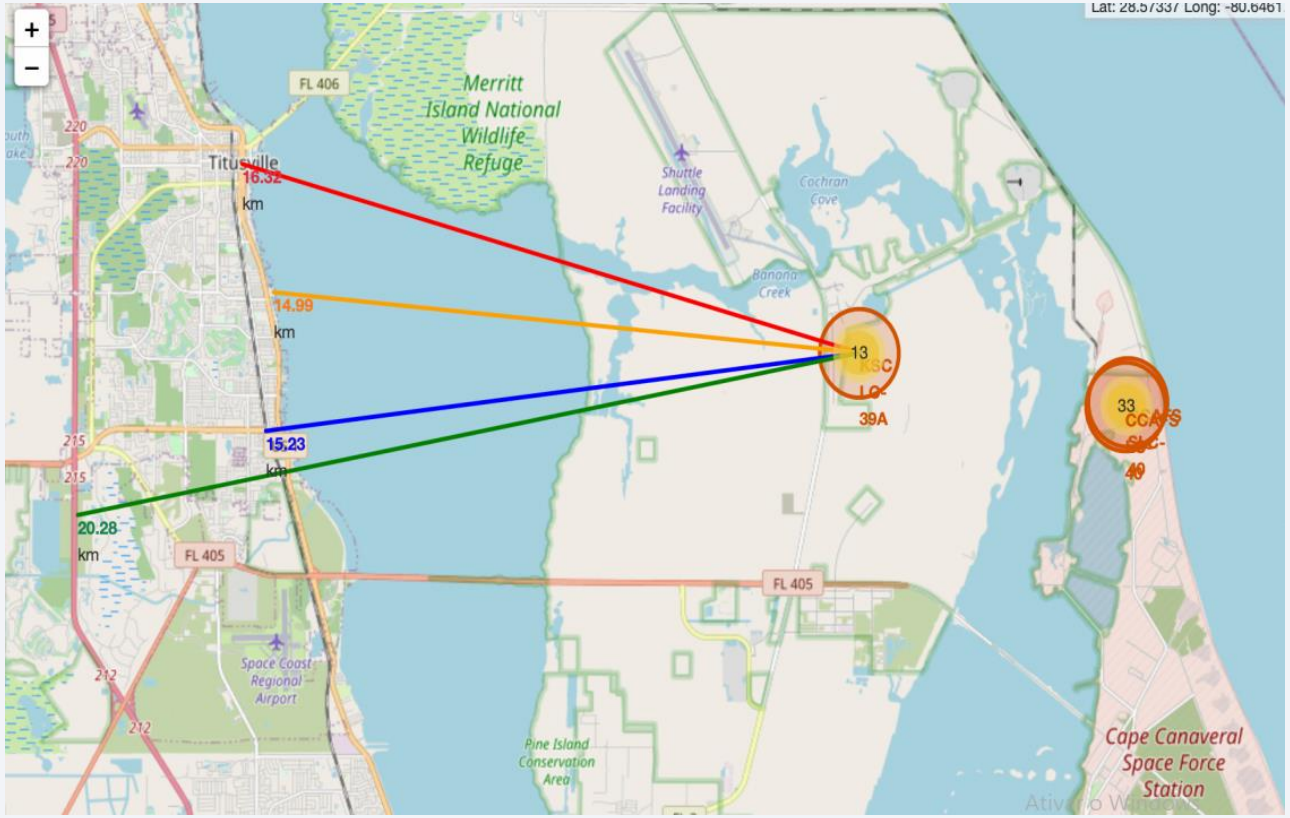
Colour-labeled launch records on the map

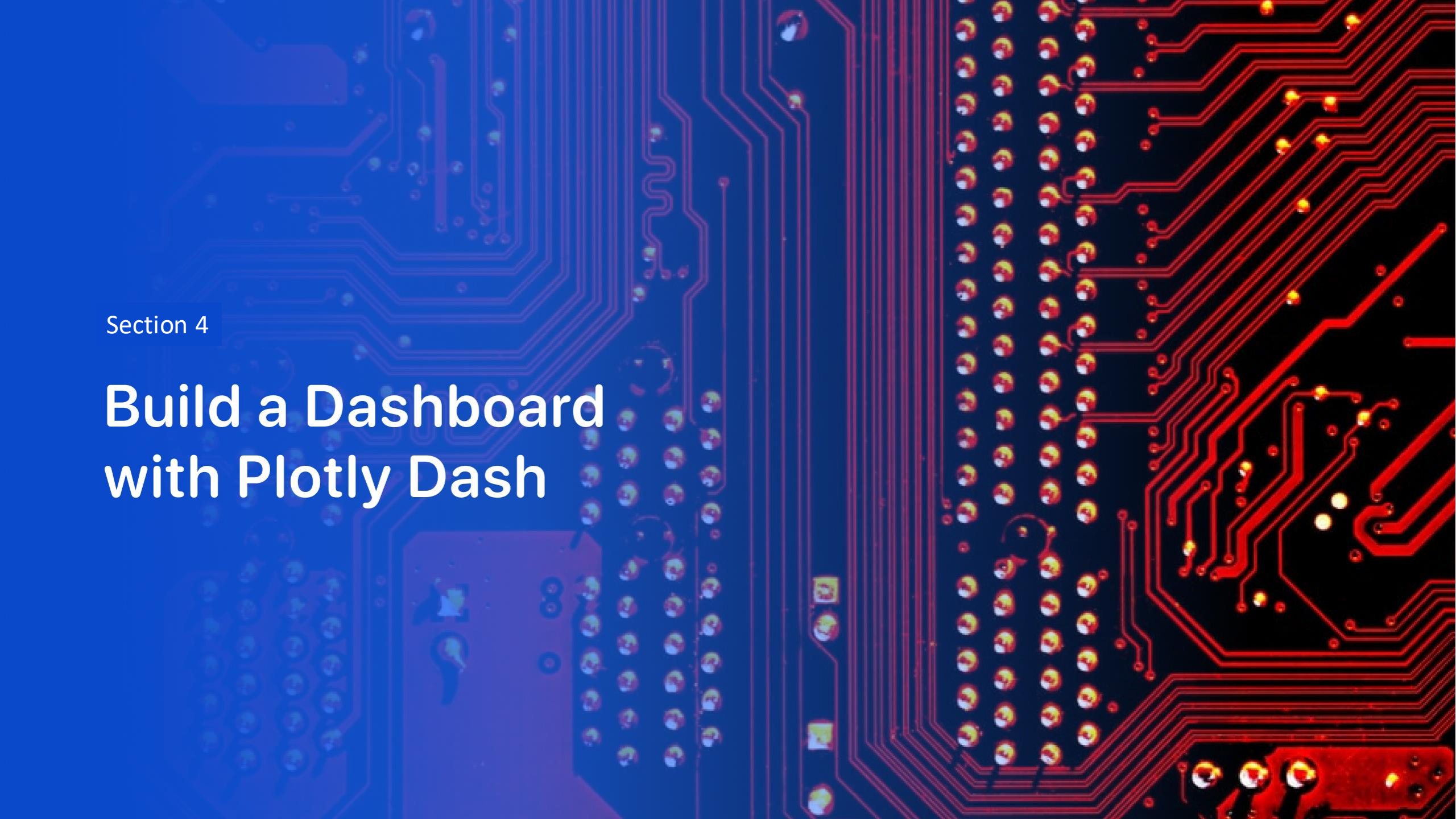
- **Color-coded markers make it easy to visually assess launch success.** Green markers represent successful launches, while red markers indicate failures.
- **By looking at the distribution of green and red markers, we can quickly identify which launch sites perform better.**
- **KSC LC-39A stands out with mostly (or all) green markers, indicating a very high success rate.** This makes it one of the most reliable launch sites in the dataset.



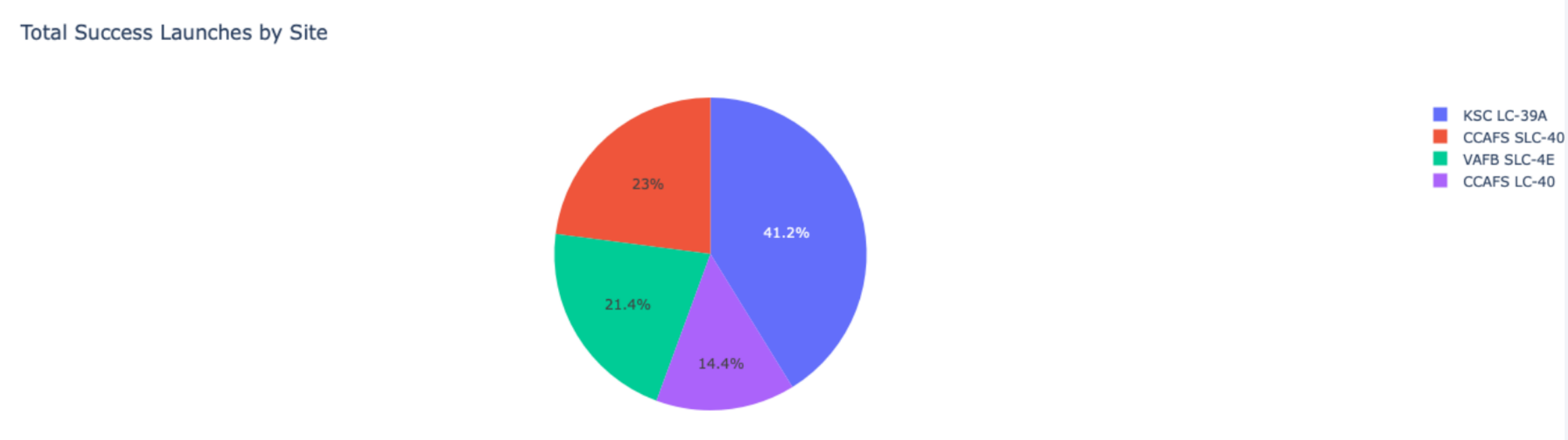
Distance from the launch site KSC LC-39A to its proximities

- **The launch site KSC LC-39A is located relatively close to key infrastructure and population centers:**
  - ~15.23 km from a railway
  - ~20.28 km from a highway
  - ~14.99 km from the coastline
  - ~16.32 km from the city of Titusville
- **This proximity provides logistical advantages** for transporting equipment and personnel but also highlights **potential risks.**
- **In the event of a launch failure, a rocket moving at high speed can travel 15–20 km within seconds**, which could pose a significant danger to nearby populated or infrastructural areas.



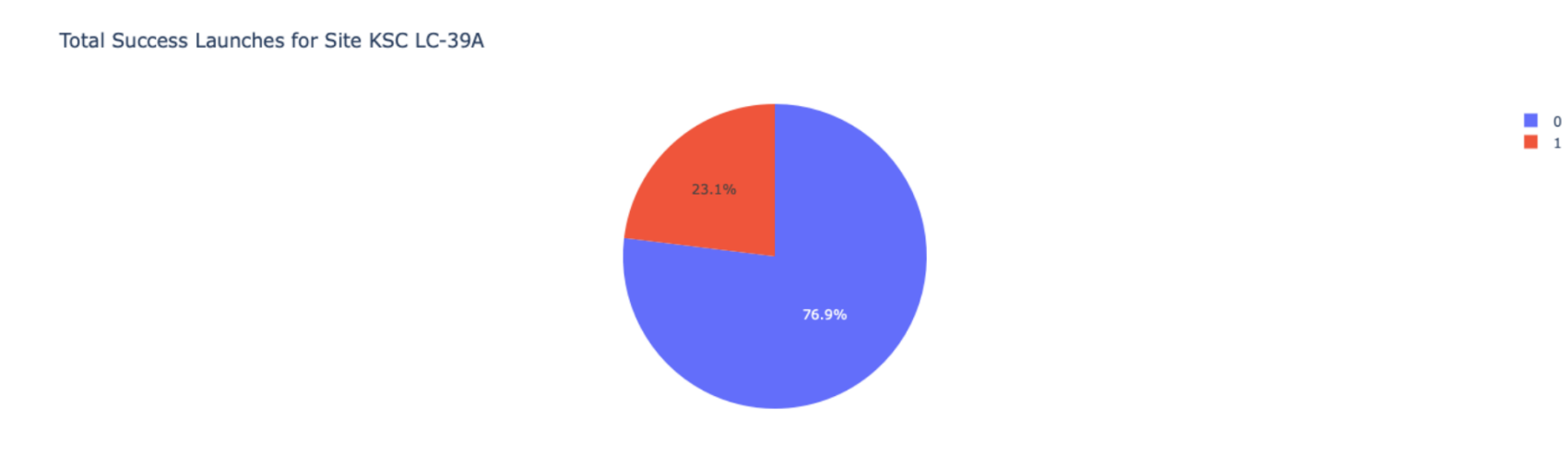

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

<Dashboard Screenshot 2>



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

<Dashboard Screenshot 3>

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Model Evaluation Summary**
- Based on the Test Set scores, we cannot definitively determine which method performs best.
- The identical or similar scores may be due to the **small test sample size** (18 samples), which limits the reliability of the evaluation.
- To address this limitation, we evaluated all models using the **entire dataset**.
- The results from the full dataset indicate that the **Decision Tree model** outperforms the others. It achieved **higher overall scores and the highest accuracy**, confirming it as the best-performing model in this analysis.

Scores and Accuracy of the Test Set

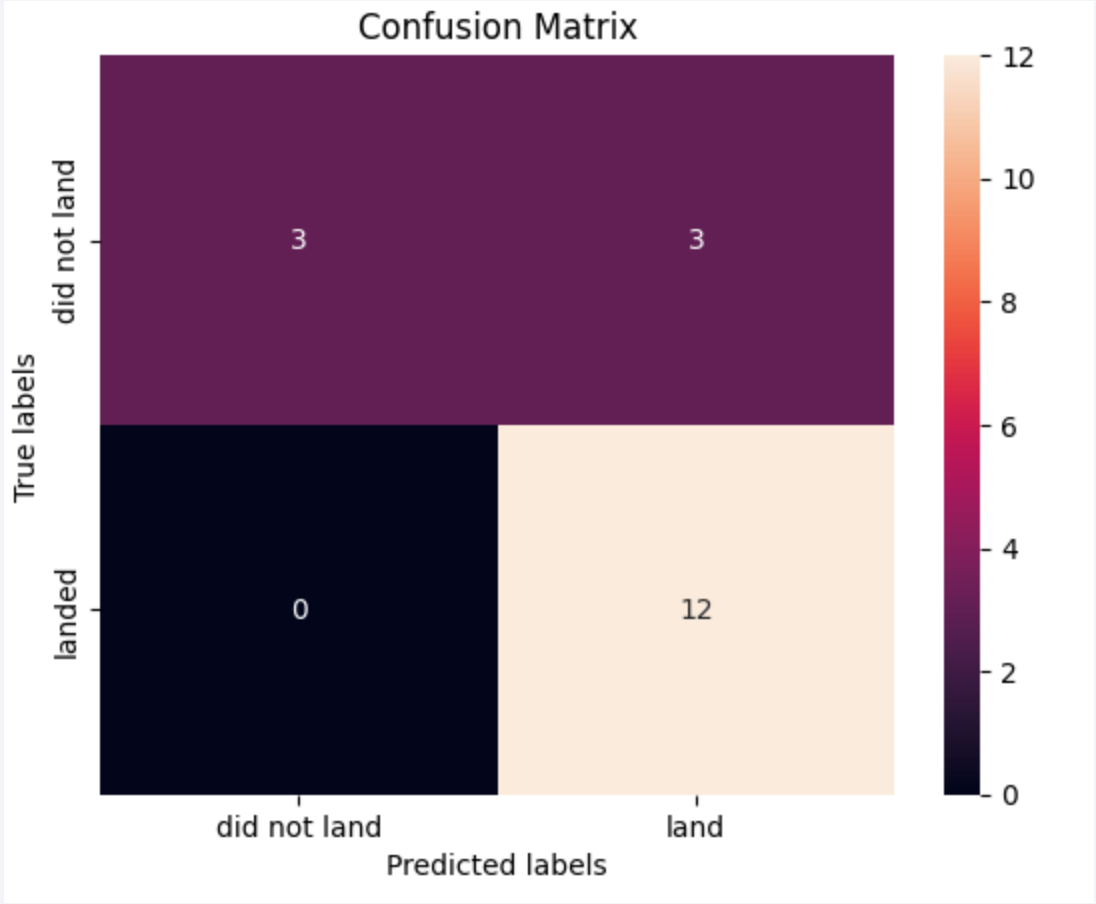
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- The confusion matrix indicates that the **Logistic Regression model is able to distinguish between the different classes**.
- However, the primary issue lies in the number of **false positives**, suggesting that the model tends to incorrectly predict the positive class more often than desired.
- This may impact the precision of the model and should be considered when selecting the appropriate model for deployment.



Conclusions

- The **Decision Tree model** outperforms other algorithms and is the most suitable for this dataset.
- **Launches with lower payload masses** tend to have higher success rates compared to those with heavier payloads.
- Most **launch sites are located near the Equator** and are situated close to **coastal regions**, likely to facilitate easier launch logistics and safety.
- The **success rate of launches has shown a positive trend over the years**, indicating improvements in technology and processes.
- The **KSC LC-39A site** has the **highest success rate** among all the launch sites analyzed.
- Missions targeting **ES-L1, GEO, HEO, and SSO orbits** have achieved a **100% success rate**, demonstrating reliability for these orbital destinations.

Thank you!