IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yufei Peter Cheng
Nov 26, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Collected and processed SpaceX Falcon 9 launch data to identify key factors influencing landing success.
  - Performed end-to-end data science workflow: data cleaning, feature engineering, exploratory analysis, visualization, and machine learning modeling.
  - Applied and compared four supervised learning algorithms—**Logistic Regression**, **Support Vector Machine**, **Decision Tree**, and **K-Nearest Neighbors**—with hyperparameter tuning using **GridSearchCV** for optimal performance.
  - Evaluated models using test accuracy and confusion matrices to assess reliability and error types.
- Summary of all results
  - All four machine learning models achieved **consistent test accuracy of ~83%**, indicating stable and predictable performance across methods.
  - Models successfully identified patterns between payload mass, launch site, booster category, and landing outcome.
  - Consistent confusion matrices show that the primary classification challenge is **false positives** (predicting "landed" when the booster did not land).
  - Findings demonstrate that **machine learning can reliably predict booster landing success**, supporting SpaceX's mission of improving reusability and lowering launch costs.

Introduction

- Project background and context
  - SpaceX has transformed the aerospace industry by reusing Falcon 9 boosters, significantly reducing launch costs.
  - Predicting whether a booster can successfully land is crucial for improving mission planning, cost efficiency, and future launch reliability.
  - Falcon 9 landing outcomes depend on multiple factors such as payload mass, launch site, booster version, and flight conditions.
  - This project applies data science and machine learning techniques to analyze historical launch data and identify the key drivers of landing success.
- Problems we want to find answers
  - What factors most strongly influence whether a Falcon 9 booster lands successfully?
  - Can we build a machine learning model that accurately predicts landing outcomes?
  - Which machine learning method performs best for this classification task?
  - How consistent are predictions across different algorithms?

Section 1

Methodology

Methodology

Executive Summary

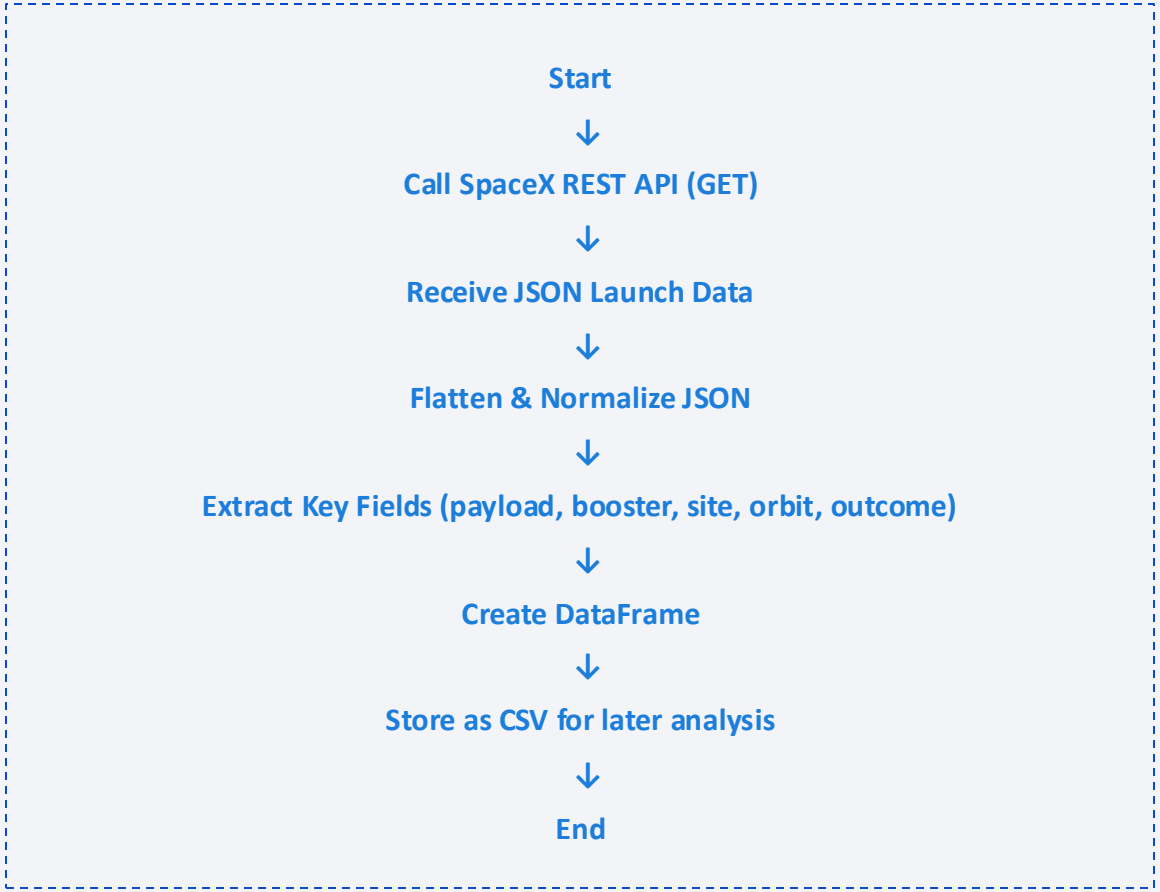
- Data collection methodology:
  - Collected SpaceX Falcon 9 launch records from publicly available APIs and datasets.
  - Combined launch metadata, payload specifications, booster information, and landing outcomes into a unified dataset.
- Perform data wrangling:
  - Cleaned missing and inconsistent values, standardized categorical variables, and engineered features such as landing success labels.
  - Filtered and transformed the dataset to prepare it for analysis and modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - Built logistic regression, SVM, decision tree, and KNN models to predict landing success.
  - Tuned model hyperparameters using GridSearchCV with cross-validation.
  - Evaluated and compared model performance to determine the best-performing classifier.

Data Collection

- Primary launch data was retrieved using the SpaceX public REST API.
- Payload mass and additional details were collected from Wikipedia using web scraping.
- The two data sources were cleaned, merged, and stored as a unified dataset for EDA and modeling.

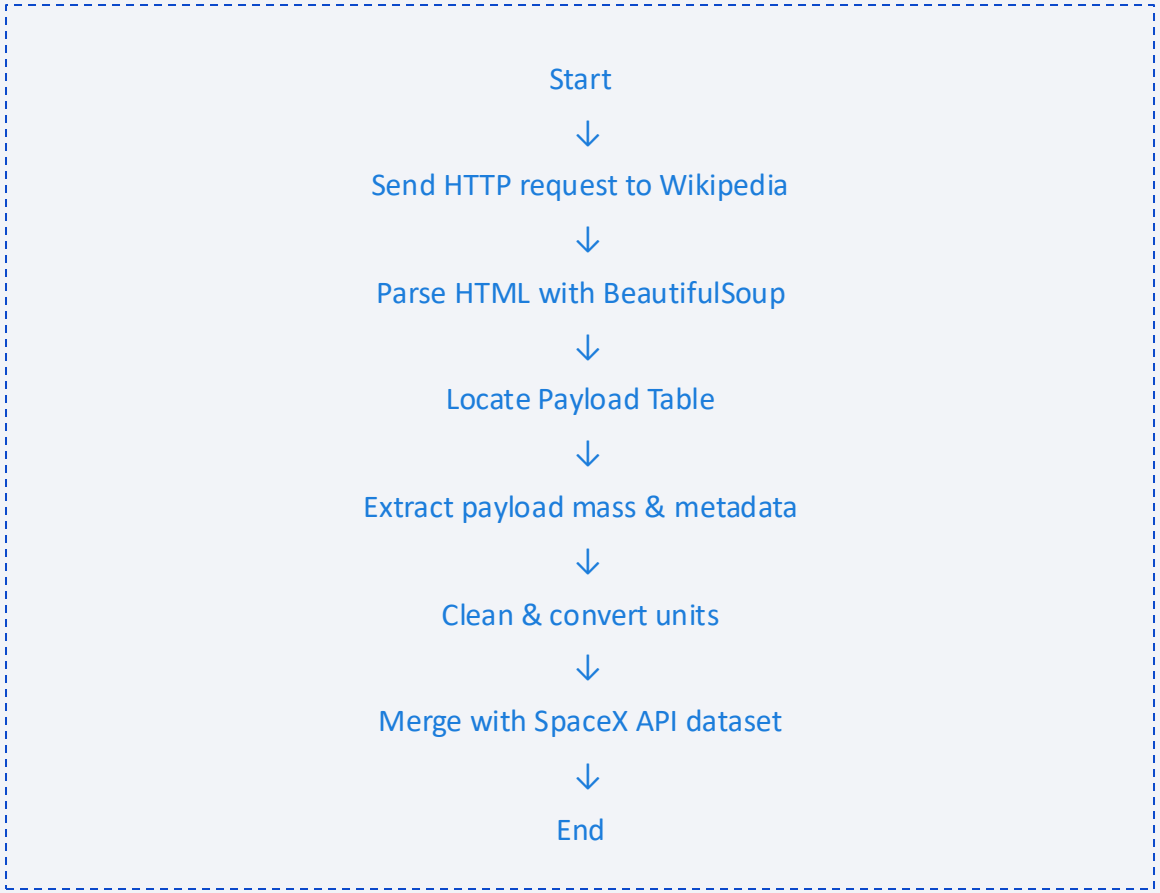
Data Collection – SpaceX API

- **Key steps of the API data collection process:**
  - Send GET requests to the SpaceX REST API
  - Receive JSON responses
  - Parse and normalize JSON into structured tables
  - Extract key features (payload, booster version, launch site, orbit, landing outcome)
  - Save the processed dataset for further analysis
- **GitHub Notebook Reference (API Collection)**
  - https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- **Key steps of the scraping process**
  - Request HTML content from Wikipedia
  - Parse page using BeautifulSoup
  - Locate payload tables and extract rows
  - Clean and standardize payload mass values
  - Merge the scraped payload data with the API dataset
- **GitHub Notebook Reference (Scraping)**
  - https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/jupyter-labs-webscraping.ipynb

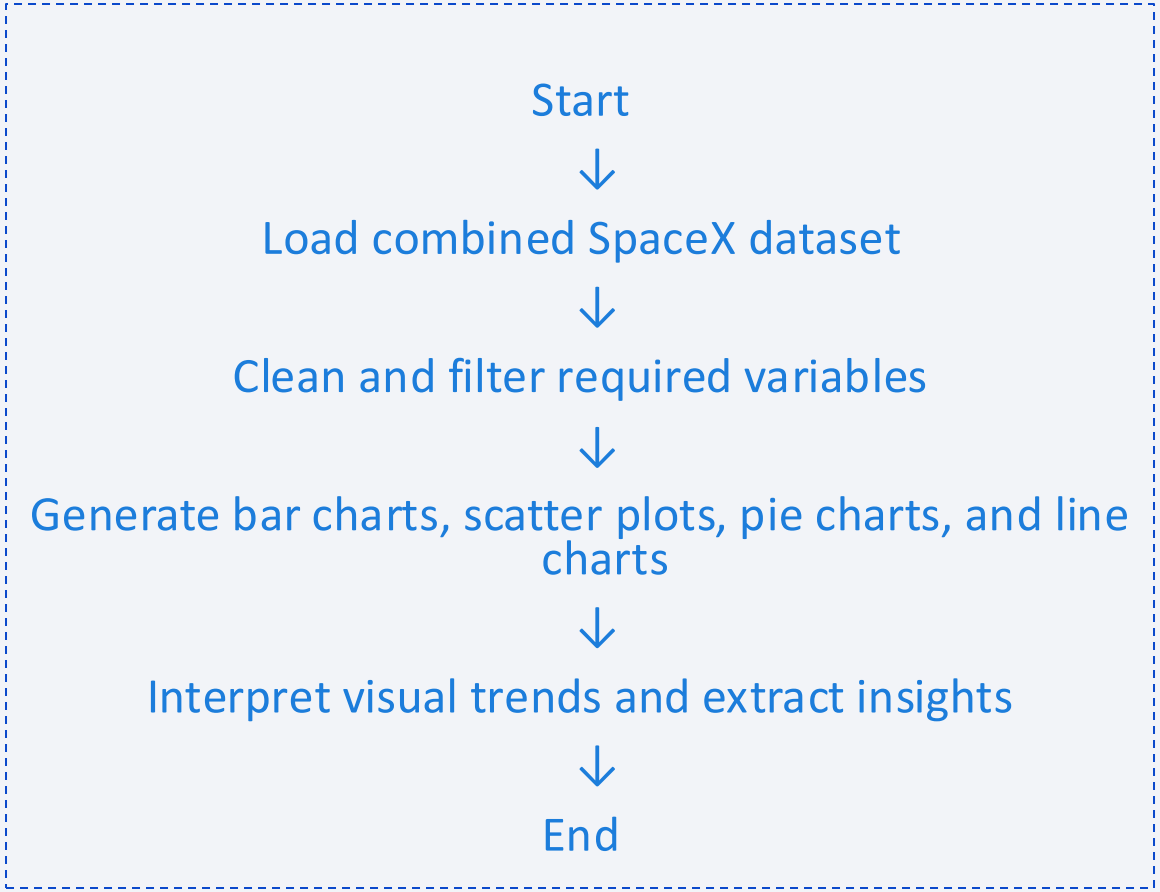


Data Wrangling

- **How the data were processed**
  - Removed duplicates and handled missing values (e.g., payload mass).
  - Engineered new features such as Booster Version Category and Orbit Encodings.
  - Converted categorical variables using One-Hot Encoding.
  - Removed irrelevant columns and formatted the final model-ready dataset.
- **GitHub Notebook Reference (Data Wrangling)**
  - https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We produced several charts to explore the merged SpaceX dataset, including:
  - **Bar charts** to compare launch success rates across launch sites
  - **Scatter plots** showing the correlation between payload mass and landing outcome.
  - **Pie charts** summarizing overall success vs. failure distribution.
  - **Line charts** illustrating changes in launch success rate over time.
- **GitHub Notebook Reference** (EDA with Data Visualization):
  - https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/edadataviz.ipynb

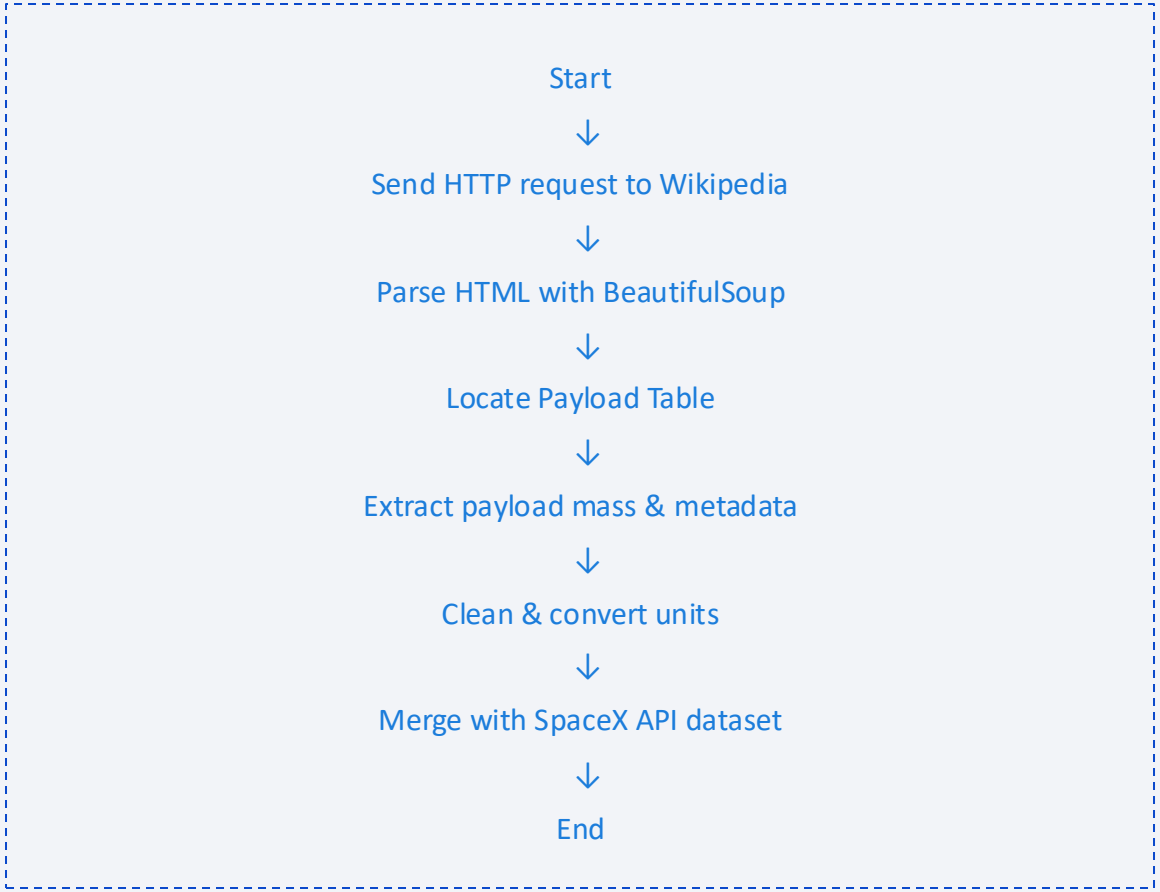


EDA with SQL

- Using SQL queries, we explored the dataset stored in SQLite, including:
  - Selecting launch records and filtering by launch site
  - Calculating the **number of successful vs. failed landings**
  - Aggregating **mean payload mass** per booster version
  - Identifying **launch outcome distributions per orbit type**
  - Joining tables to merge payload, launch, and booster data
- Full notebook (EDA with SQL):
  - **https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/jupyter-labs-eda-sql-coursera_sqllite.ipynb**

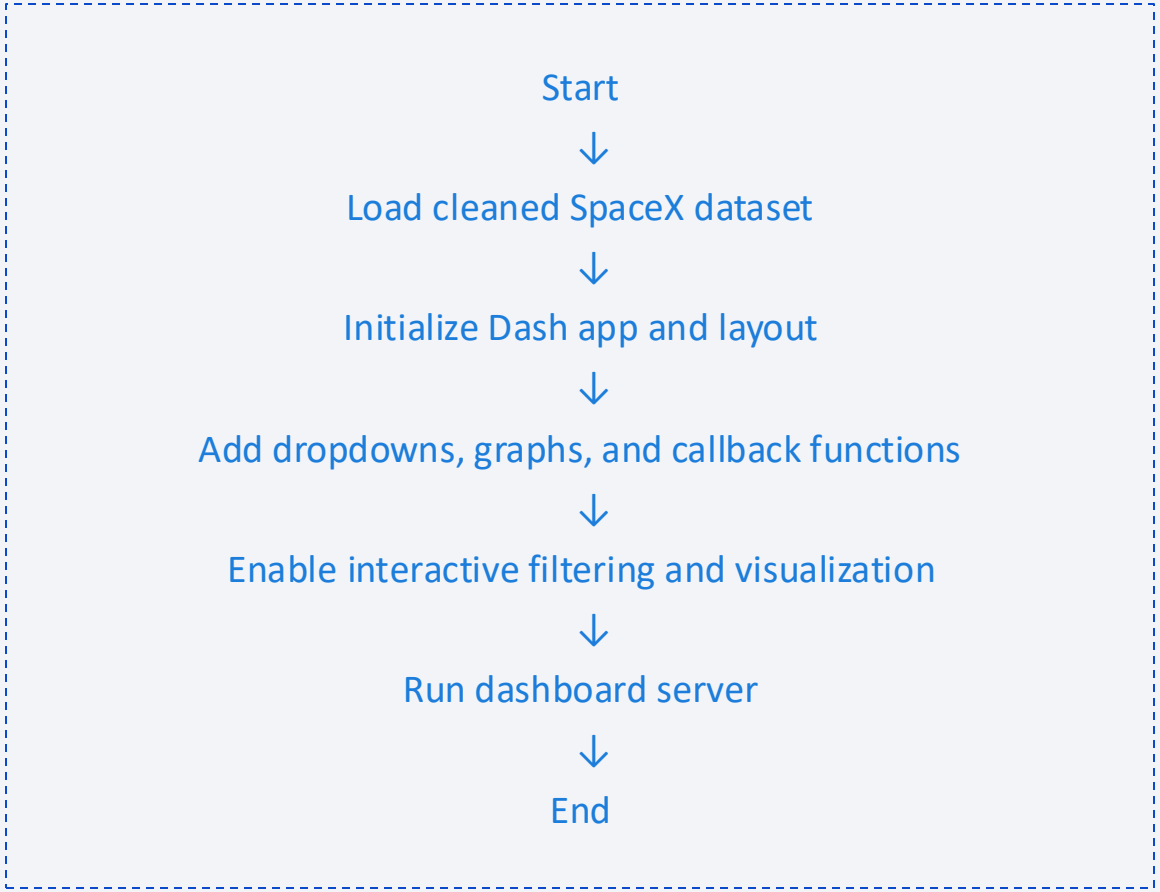
Build an Interactive Map with Folium

- We created an interactive map to visualize SpaceX launch sites using Folium.
- Objects added:
  - **Markers** for each launch site location
    Why: Displays exact geographic positioning.
  - **Circles** showing surrounding radius
    Why: Highlights safety zones and regional context.
  - **Polylines** connecting launch sites to booster landing locations
    Why: Visualizes mission trajectories and landing distances.
  - **Pop-ups** with site information
    Why: Allows users to interact and learn details on-click.
- Full Folium notebook:
  - **https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/lab_jupyter_launch_site_location.ipynb**



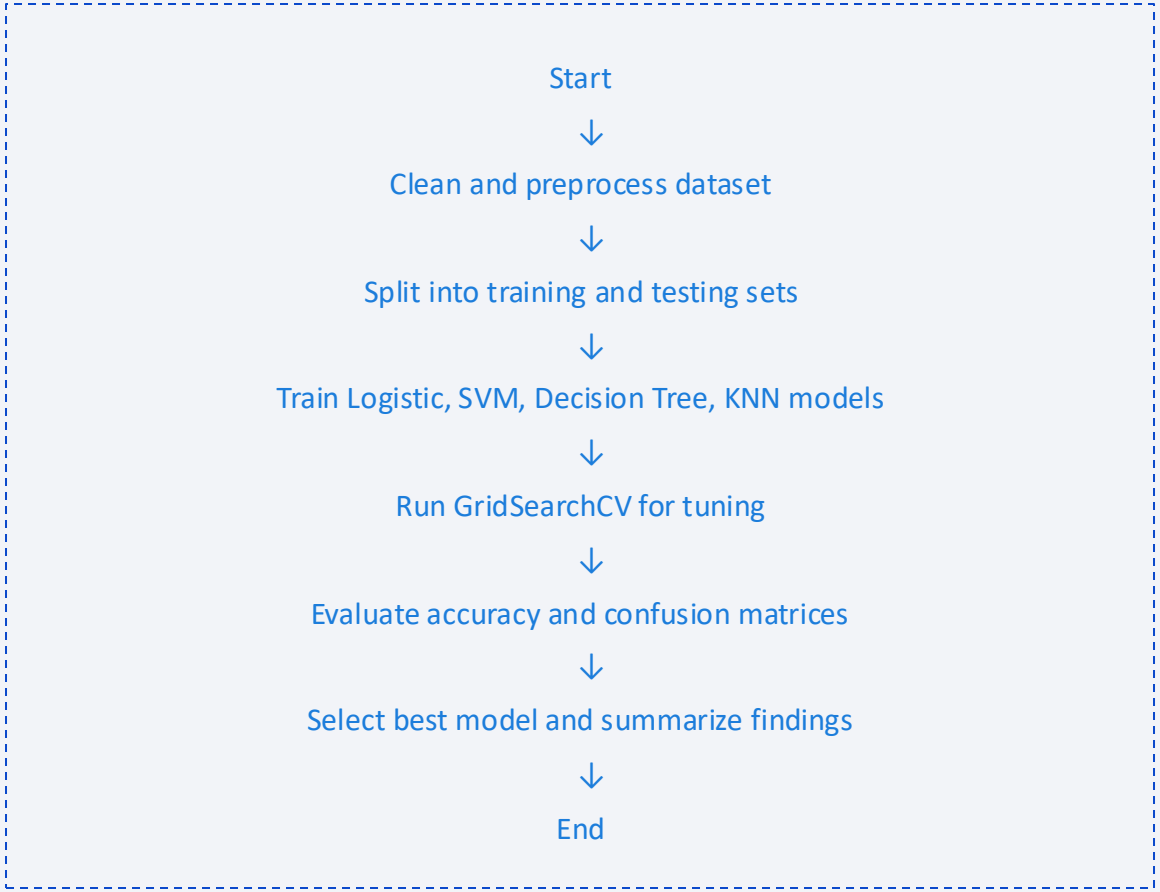
Build a Dashboard with Plotly Dash

- The interactive dashboard includes:
  - **Dropdown filters** for launch site selection
    Why: Enables exploratory analysis by location.
  - **Pie charts** summarizing landing success per site
    Why: Allows quick comparison of performance.
  - **Scatter plots** linking payload mass to launch outcome
    Why: Helps visualize payload influence on success.
- Full Dash Application Code:
  - https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/spacex-dash-app.py



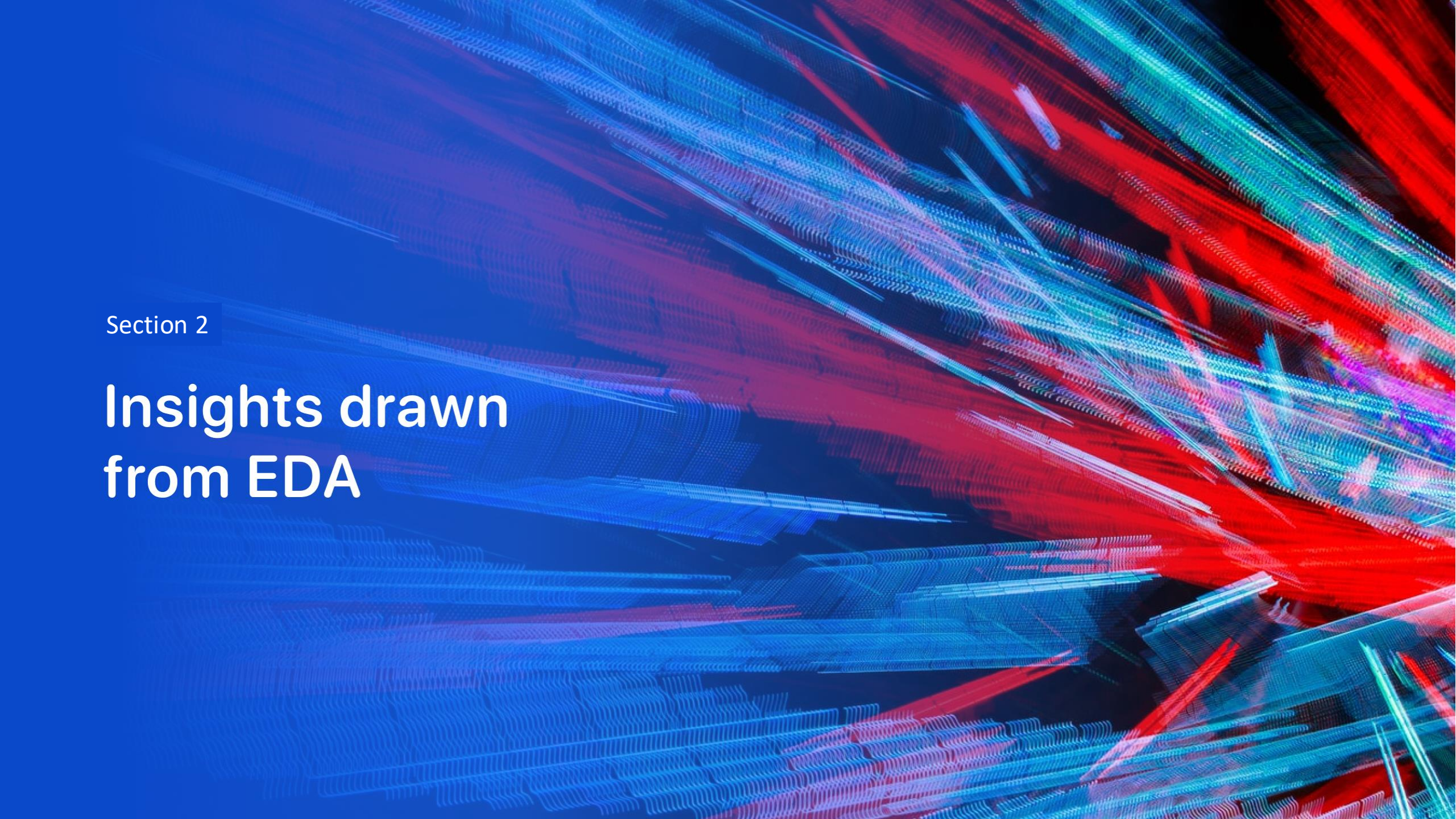
Predictive Analysis (Classification)

- Classification models tested:
  - **Logistic Regression**
  - **Support Vector Machine**
  - **Decision Tree Classifier**
  - **K-Nearest Neighbors**
- Process steps:
  - Train/test splitting
  - Hyperparameter tuning using **GridSearchCV**
  - Model evaluation using **accuracy score and confusion matrix**
  - Comparison of all models to select the best performing one
- Best model based on your results:
  All models performed similarly with ~83.3% accuracy, but further tuning could differentiate them.
- Full predictive analysis notebook:
  - **https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science/blob/2a957fdd2331c8a0bf5cb202ae5abb889ea96315/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb**

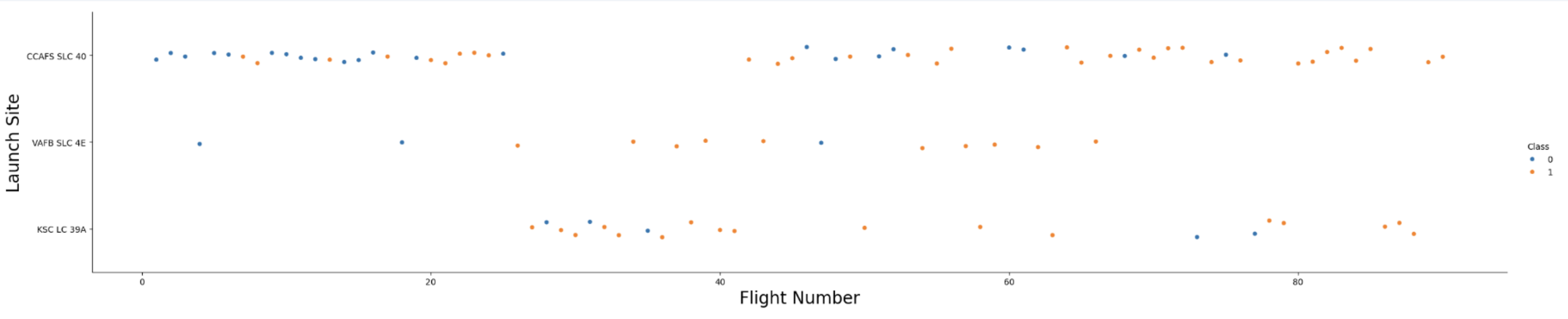


Upcoming Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

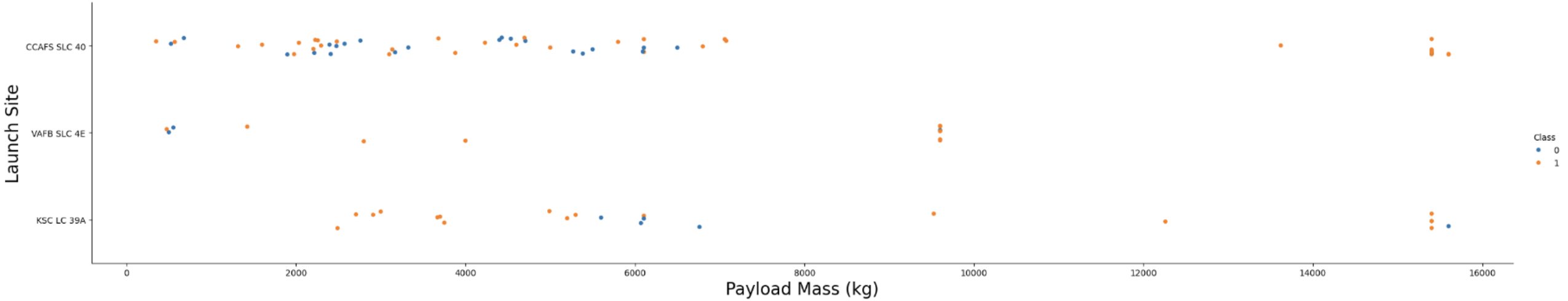

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- **Key Insights**
  - **Higher flight numbers generally show more successful landings**, indicating improvement over time.
  - **Launch sites differ in performance** — some sites have more successful outcomes than others.
  - The plot shows **both temporal learning (more flights → better success)** and **site-specific patterns**.

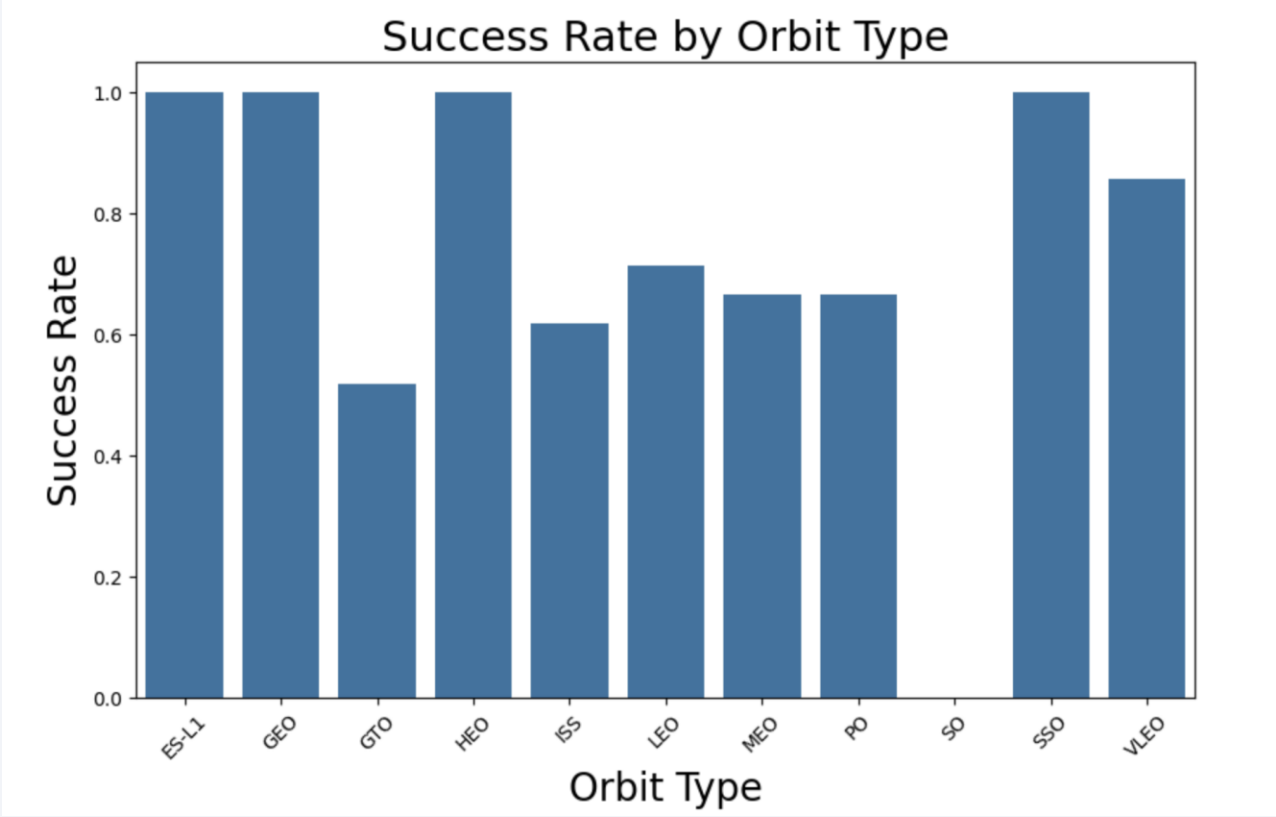
Payload vs. Launch Site



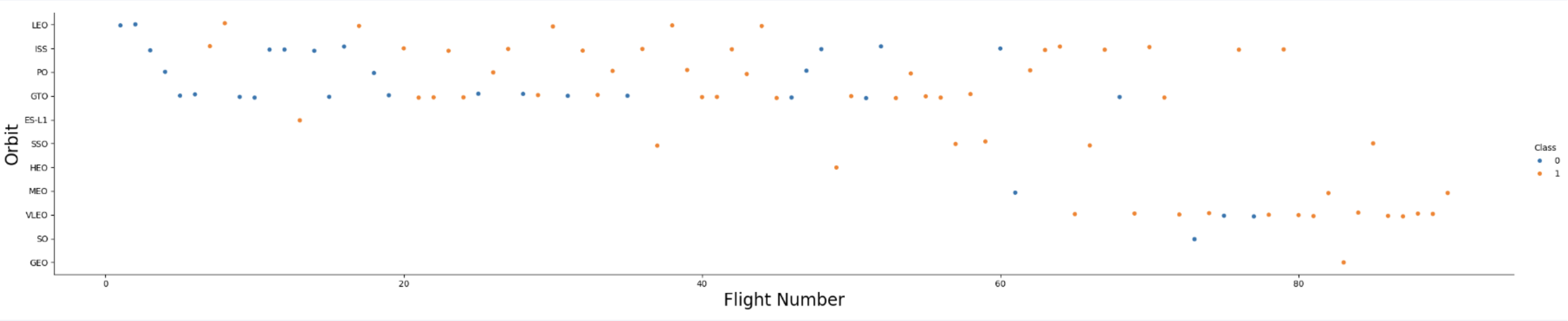
- **Key Insights**
  - **Higher payloads do not always guarantee success** — outcomes vary across sites even at similar payload levels.
  - **Each launch site handles different payload ranges**, showing variation in mission profiles.
  - Success patterns appear **more dependent on site and mission type** than on payload alone.

Success Rate vs. Orbit Type

- **Key Insights**
  - **Success rates vary widely across orbit types**, showing that some missions are more reliable than others.
  - **GEO, ES-L1, and SSO orbits show the highest reliability**, while **FTS-required or high-energy orbits** tend to have lower success rates.
  - This suggests that **mission complexity and orbit requirements** play a major role in launch outcomes.

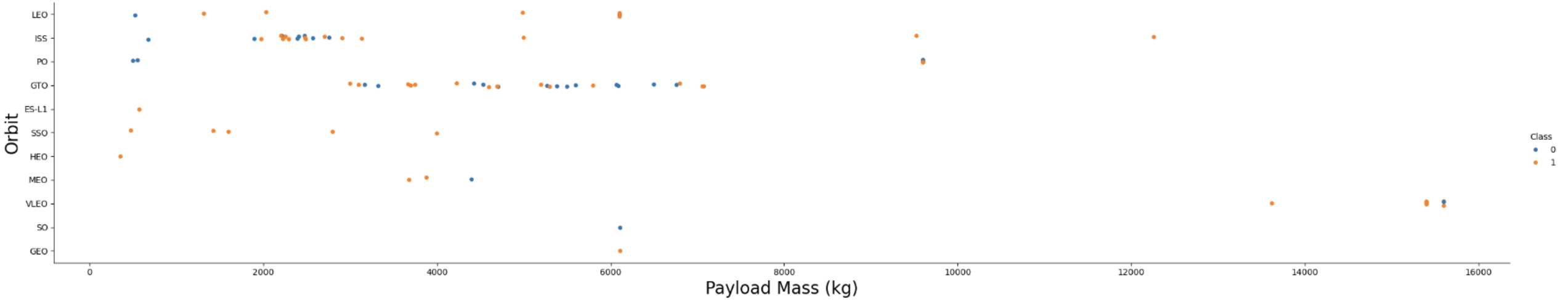


Flight Number vs. Orbit Type



- **Key Insights:**
  - The plot shows **how different orbits correspond to early vs. later Falcon 9 missions**.
  - **Early flights** targeted mostly common orbits (LEO, ISS, GTO), while **later flights** expanded to more diverse and higher-complexity orbits.
  - The distribution also suggests that **success rates improved over time**, as later flights (higher flight numbers) have more successful outcomes.

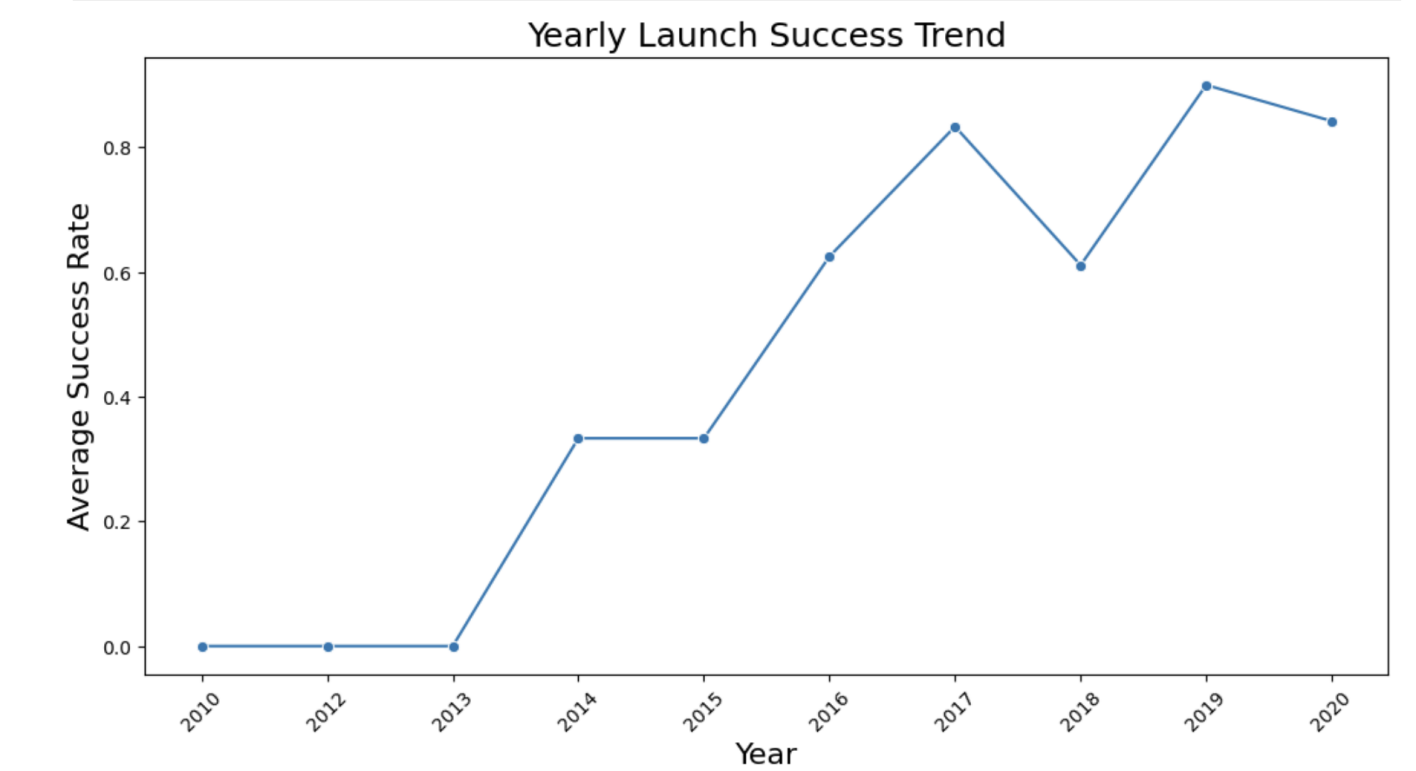
Payload vs. Orbit Type



- **Key Insights:**
  - Different orbit types require different payload ranges, and the plot shows which orbits typically carry heavier or lighter payloads.
  - High-mass payloads appear mostly on a few specific orbits (e.g., GTO, GEO), while others consistently carry lighter payloads.
  - The outcome colors also show that **success rates remain high across most payload levels**, with failures scattered but not concentrated at heavy masses.

Launch Success Yearly Trend

- **Key Insights:**
  - The yearly success rate shows a clear upward trend, indicating consistent improvements in launch reliability over time.
  - Early years had low or unstable success, but after 2014 the success rate rose sharply and remained high, suggesting maturation of technology and operations.



All Launch Site Names

- **Key Insights:**
- There are **four unique launch sites** in the dataset:
  - CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.
  - These sites represent the main locations used for SpaceX launches during the period covered in the data.

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

Launch Site Names Begin with 'CCA'

- **Key Insights:**
- These five records show launches from sites whose names begin with **“CCA”**, which correspond to the Cape Canaveral launch complexes (e.g., CCAFS LC-40). This confirms that multiple early SpaceX missions were conducted from Cape Canaveral.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- **Key Insights:**
- The total payload mass carried by NASA-associated boosters is **45,596 kg**, indicating the cumulative cargo launched across all NASA-related missions in the dataset.

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- **Key Insights:**
- The average payload carried by the F9 v1.1 booster is **2,928.4 kg**, showing the typical cargo capacity for missions using this version.

Avg_Payload
2928.4

First Successful Ground Landing Date

- **Key Insights:**
- SpaceX achieved its first successful ground landing on **2015-12-22**.

First_Ground_Success_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Key Insights:**
- **The query found four boosters meeting the conditions: B1022, B1026, B1021.2, and B1031.2.**

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

Total Number of Successful and Failure Mission Outcomes

- **Key Insights:**
- The query shows that SpaceX had **100 successful missions** and only **1 failure**, indicating a very high mission reliability.

Outcome_Class	Total_Missions
Failure	1
Success	100

Boosters Carried Maximum Payload

- **Key Insights:**
- These boosters all appear at the maximum recorded payload mass, meaning they shared the heaviest payload missions in the dataset.

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

- **Key Insights:**
- In 2015, two drone-ship landings failed, both using F9 v1.1 boosters and both launched from CCAFS LC-40.

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

- **Key Insights:**
- Between 2010-06-04 and 2017-03-20, “No attempt” was the most common landing outcome, followed by both successful and failed drone-ship landings.

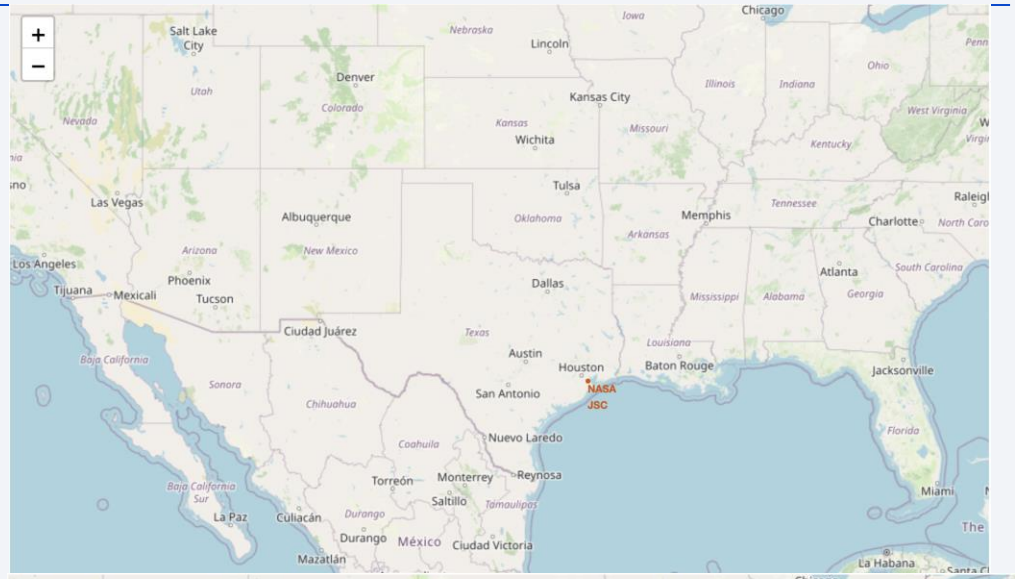
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

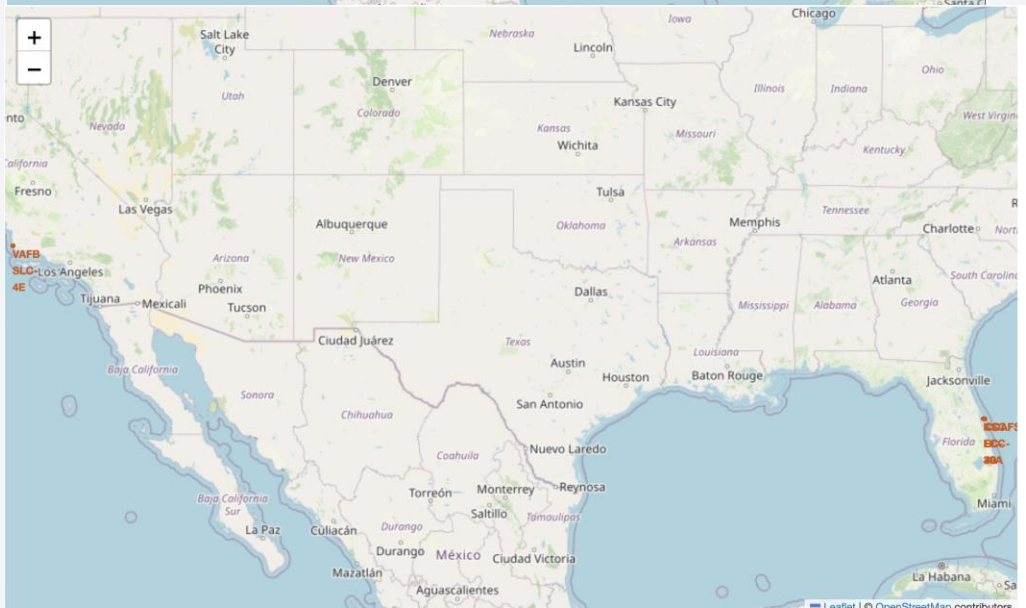
Section 3

Launch Sites Proximities Analysis

Mark All Launch Site on Map

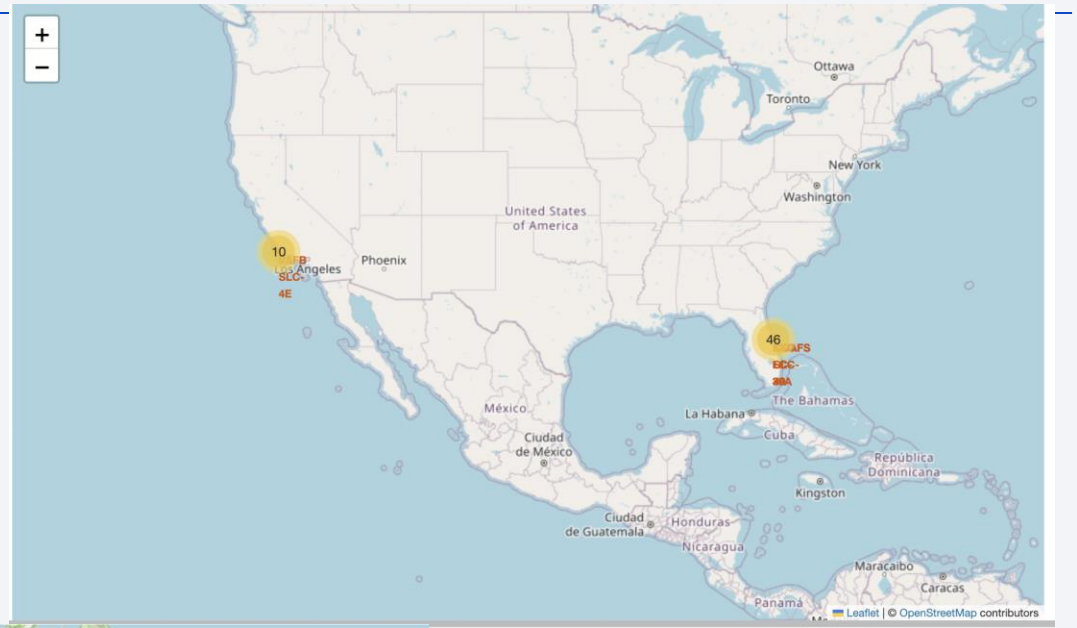
- **Key Insights:**
  - This map displays all SpaceX Falcon 9 launch sites, each marked by a location pin on the U.S. map.
  - All launch sites are positioned **very close to the coastline**, which reduces risk during launch and allows safe booster landings over water.
  - All sites lie in the **southern United States**, but **none** are close to the Equator line. Launching closer to the equator provides additional velocity from Earth's rotation, but SpaceX's U.S.-based operations rely instead on coastal access and existing NASA/DoD infrastructure.

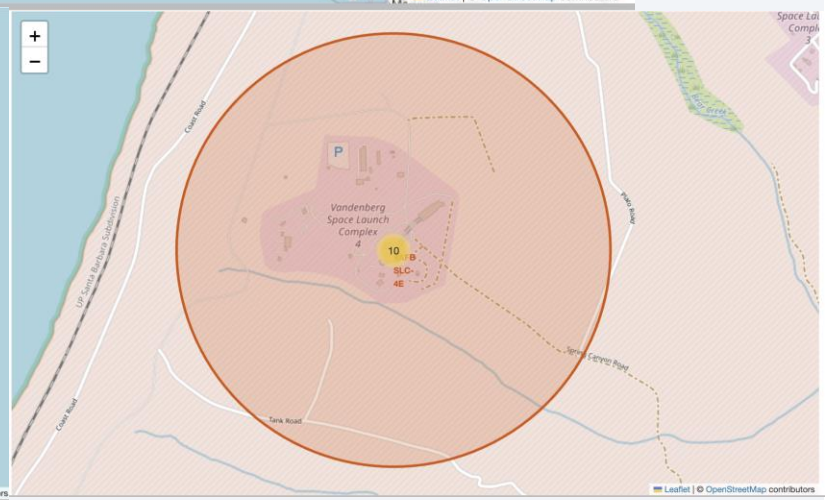




Success/Failed Launches for Each Site on Map

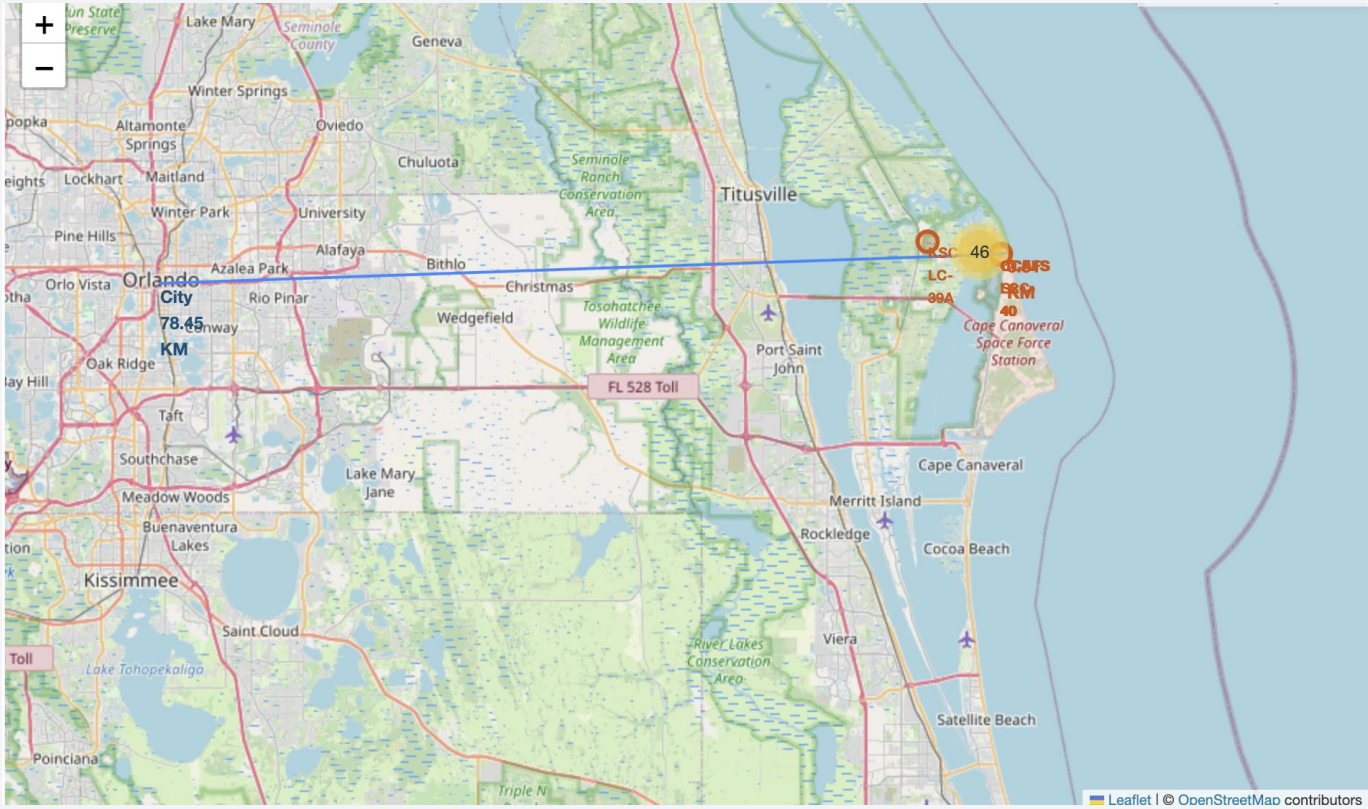
- **Key Insights**
  - Each marker cluster shows the number of Falcon 9 launches at that site, with **color-coded outcomes** (green for success, red for failure).
  - From the clusters, we clearly see that **Cape Canaveral (Florida)** has the **highest launch count and a very high success rate**, indicated by mostly green markers.
  - **Vandenberg (California)** also shows a strong success pattern but with fewer total launches.
  - **Boca Chica, Texas** has fewer missions overall, but the cluster still shows **predominantly successful outcomes**.

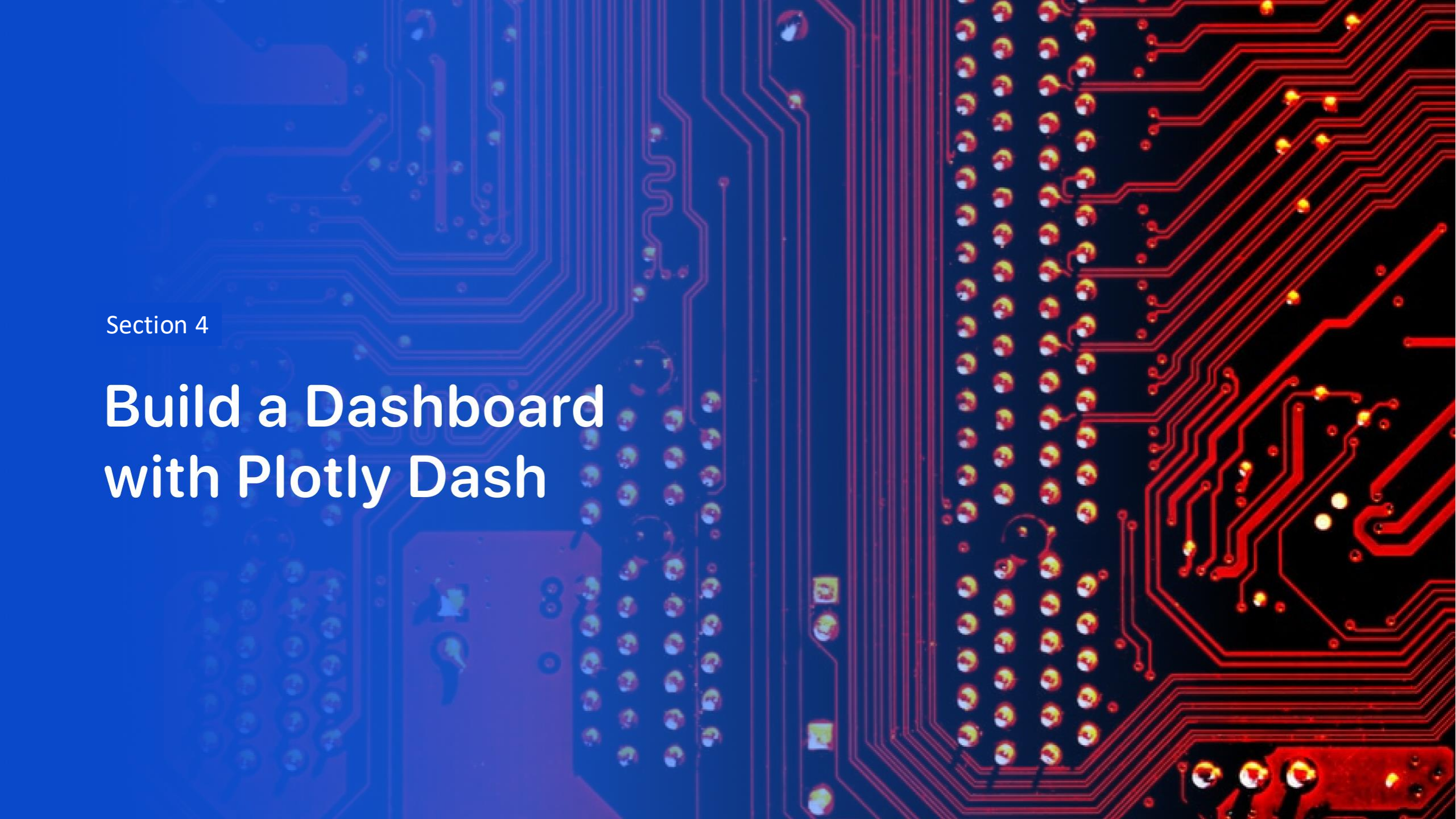




Distance From City to Launch Site

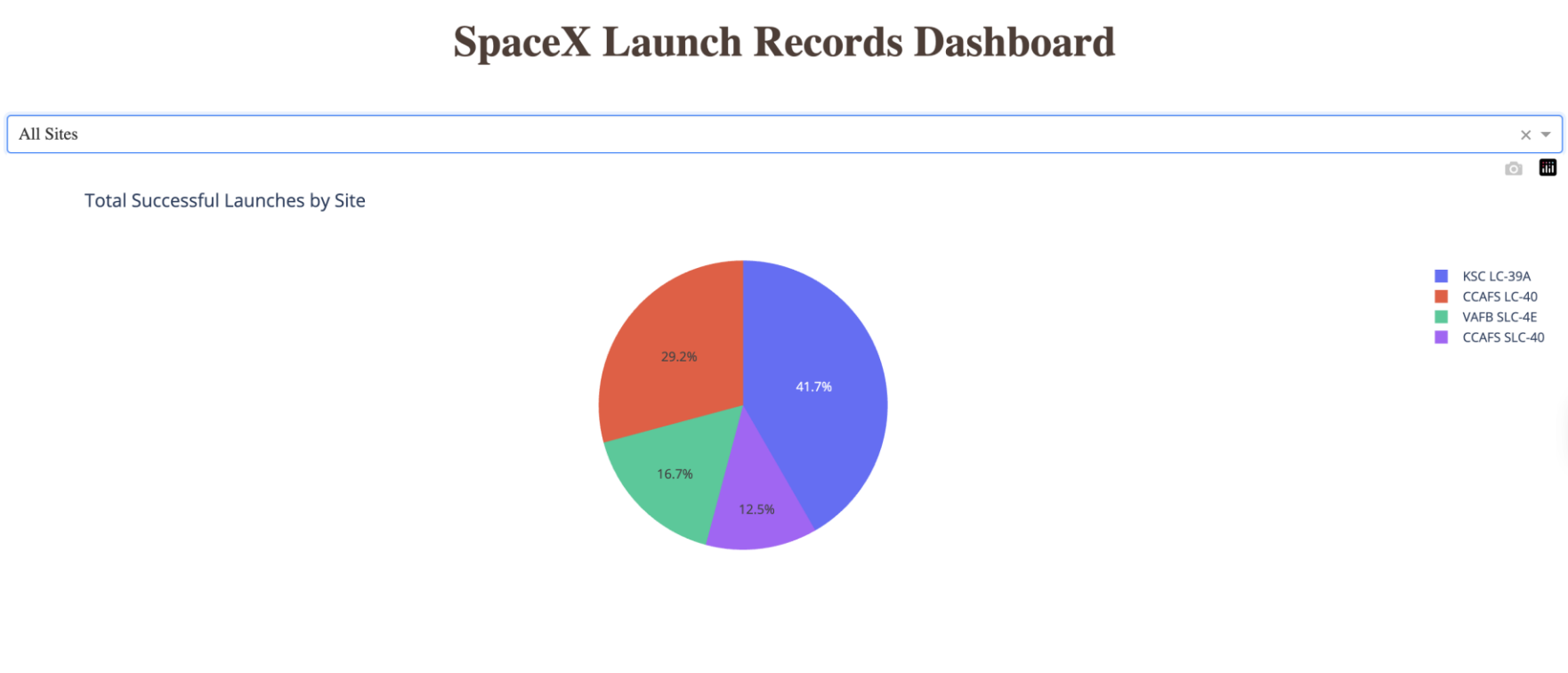
- **Key Insights**
  - The distance lines show that **all launch sites are located very close to the coastline**, which helps rockets safely drop boosters over the ocean and reduces risk to populated areas.
  - Launch sites are also **near major highways**, enabling transportation of large rocket components and access for engineering teams.
  - Most sites sit **close to railways**, which SpaceX and NASA traditionally use for transporting heavy equipment.
  - Importantly, each launch site maintains a **significant buffer distance from major cities** (e.g., ~78 km from Orlando to Cape Canaveral). This separation ensures safety during launches and minimizes risk in the event of debris or anomalies.




Section 4
Build a Dashboard with Plotly Dash

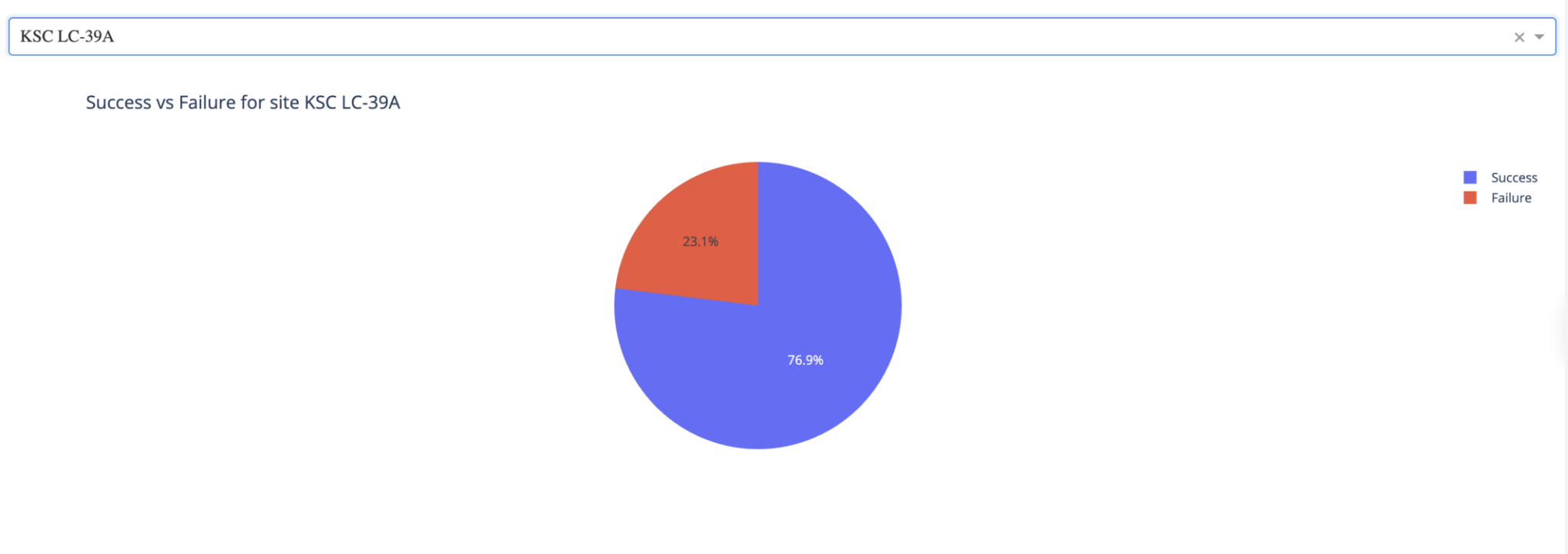
Total Successful Launches at All Sites

- **Key Insights**
- The pie chart shows the **distribution of all successful Falcon 9 launches across SpaceX's launch sites**.
- **KSC LC-39A** accounts for the **largest share of successful missions**, highlighting its role as SpaceX's most active and capable launch facility.
- **CCAFS LC-40** also contributes a significant portion of successes, reflecting its long operational history with SpaceX.
- **VAFB SLC-4E** handles fewer missions since it is primarily used for polar-orbit and sun-synchronous launches.
- **CCAFS SLC-40** shows the smallest share, consistent with it being used less frequently relative to the major launch pads.



The Site with Highest Launch Success Ratio

- **Key Insights**
- The pie chart shows the **Success vs. Failure distribution for KSC LC-39A**, the launch site with the **highest success ratio**.
- **KSC LC-39A achieves roughly 77% successful launches**, making it SpaceX's most reliable site in terms of success rate.
- The relatively small failure portion (about **23%**) indicates strong operational performance, consistent procedures, and high mission reliability at this site.
- Overall, the chart highlights **LC-39A's role as SpaceX's most dependable launch pad** among all sites.



Payload vs. LaunchOutcome for All Sites

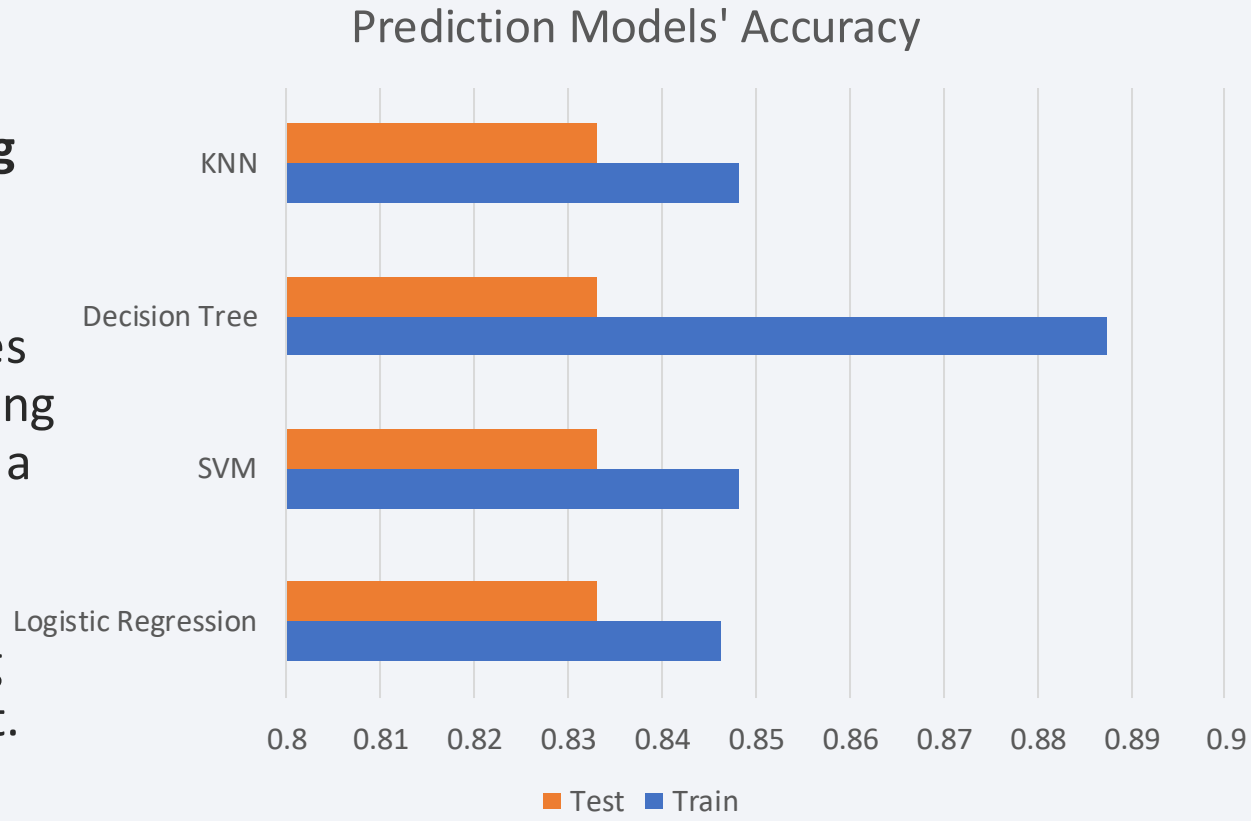
- **Key Insights**
- The scatter plots show how **payload mass relates to launch outcomes** across different SpaceX sites.
- Overall, **payload is not strongly correlated with launch success** — missions succeed across both low and high payload ranges.
- **Booster versions matter more than payload**: newer boosters (especially **B5**) consistently appear in the successful outcomes across all payload ranges.
- As the payload slider increases, the plot shows that **heavy-payload missions are still reliably successful**, indicating strong booster performance and robust launch capability.

Section 5

Predictive Analysis (Classification)

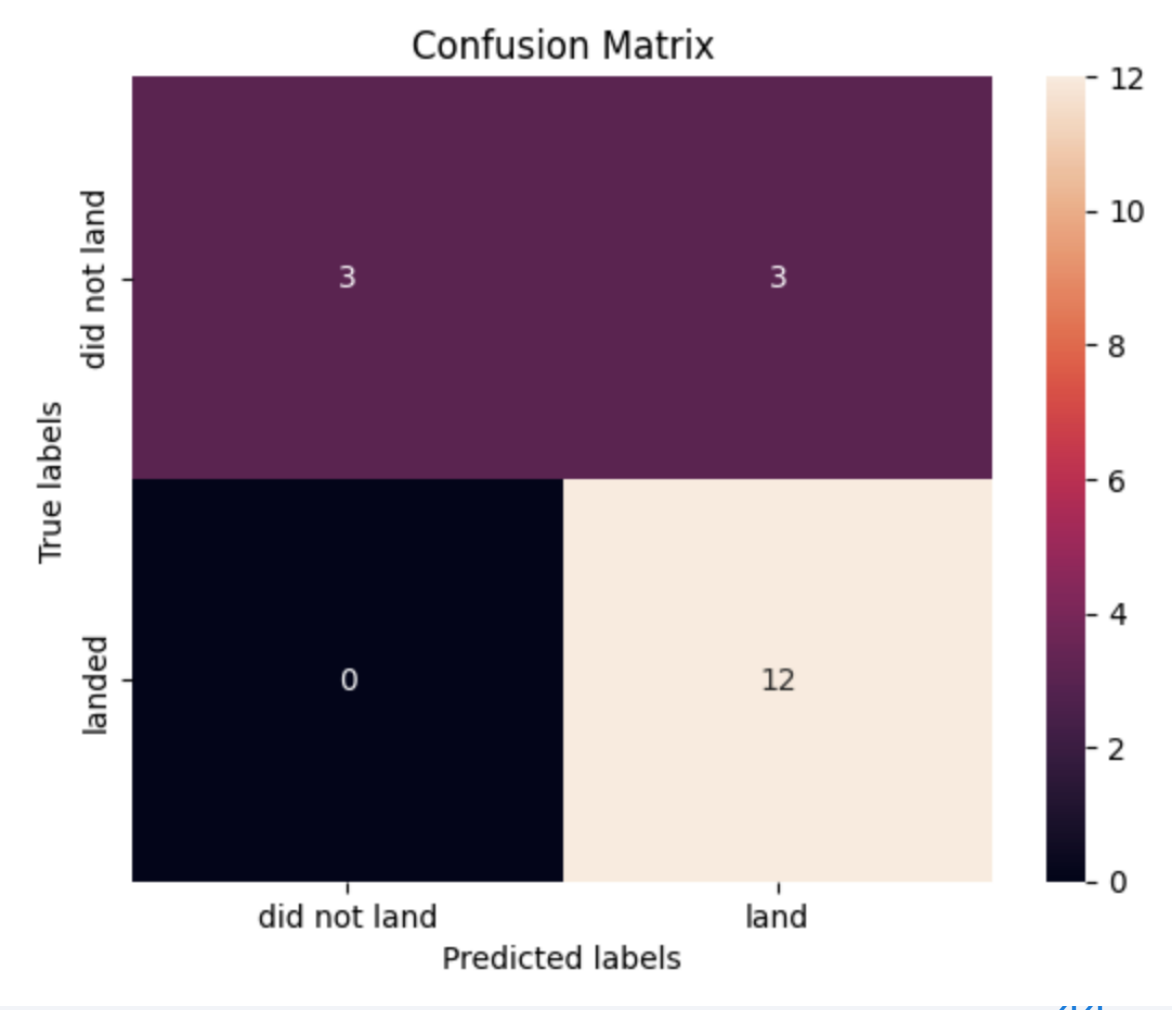
Classification Accuracy

- **Key Insights**
- This bar chart compares the **test and training accuracy** for four classification models: KNN, Decision Tree, SVM, and Logistic Regression.
- Among all models, the **Decision Tree** achieves the **highest accuracy**, especially on the training data, indicating strong performance but also a possibility of overfitting.
- KNN, SVM, and Logistic Regression show similar accuracy levels, with SVM performing slightly better than the others on the test set.



Confusion Matrix of Decision Tree

- **Explanation**
- This confusion matrix shows how well the **Decision Tree** model classifies landing outcomes.
- The model correctly predicted **12 successful landings** and **3 failed landings**.
- It incorrectly classified **3 failed landings as successful**, and made **0 false negatives** (no successful landing was predicted as failure).
- Overall, the model is strong at identifying successful landings, but slightly less accurate at detecting failures.



Conclusions

- Falcon 9 launch missions show **high overall reliability**, with only a small number of failures across all years.
- Ground landings began in **2015**, and landing success rates increased significantly afterward.
- All launch sites are located **near coastlines** and **away from major cities**, supporting safety and optimal launch trajectories.
- **KSC LC-39A** has the **highest launch success ratio** among all SpaceX sites.
- Payload mass shows **no strong correlation** with launch outcome across sites.
- Among the classification models built, the **Decision Tree** achieved the highest test accuracy.
- The Decision Tree's confusion matrix shows it **correctly identifies successful landings well**, though it slightly misclassifies some failed attempts.
- Overall, the prediction models demonstrate that **landing outcome can be reasonably predicted** using mission features like payload, site, and booster version.

Appendix

- Github Code Reference:
- https://github.com/PeterCheng0906/Winning-Space-Race-with-Data-Science.git

Thank you!