IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason Echevarria
3/23/2024

Outline

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

Executive Summary

Methodology Summary:

- Data Gathering:
  - Utilize SpaceX API to collect relevant data.
- Data Wrangling:
  - Perform data cleaning and preprocessing using Pandas.
- Exploratory Data Analysis (EDA):
  - Conduct EDA using SQL to understand dataset structure.
  - Visualize data with Matplotlib and Seaborn for insights.
- Dashboard Creation:
  - Develop an interactive dashboard with Plotly Dash for data visualization.
- Predictive Analysis:
  - Implement machine learning algorithms (e.g., decision trees, logistic regression, SVM, KNN) using scikit-learn packages in Python.

Executive Summary Cont.

Summary of the results:

Through meticulous data collection, wrangling, and exploratory analysis, coupled with the implementation of diverse machine learning algorithms, we achieved a remarkable 90% accuracy in predicting the success of Falcon 9 first stage landings. This predictive capability enables us to provide precise estimations of launch costs, essential for informed decision-making and competitive bidding within the aerospace industry.

Introduction

The goal:

This project aims to predict the successful landing of the Falcon 9 first stage. SpaceX offers Falcon 9 rocket launches at a significantly lower cost compared to other providers, mainly due to the ability to reuse the first stage. Predicting successful landings can help estimate launch costs, crucial for competing bids in the rocket launch market.

The problem:

Through this project, we seek to address the challenges of predicting landing success, estimating launch costs, and providing valuable insights for competitors in the aerospace industry.

Section 1

Methodology

Methodology

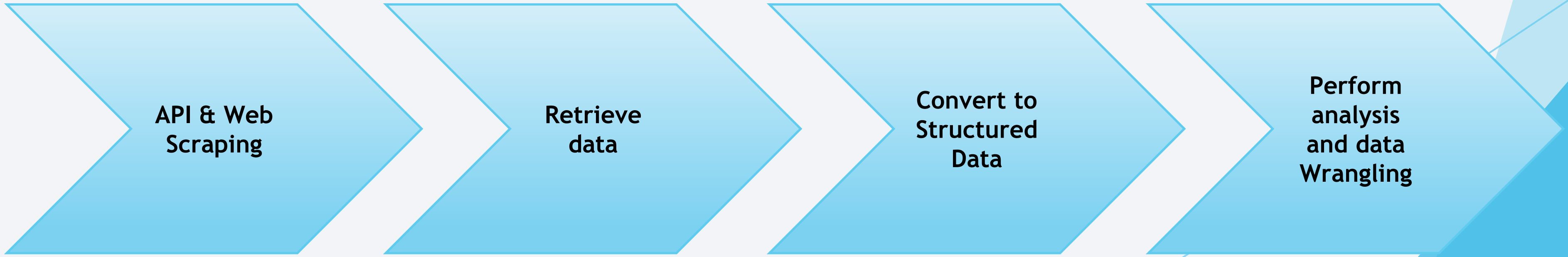
- Data collection methodology:
  - Utilizing a combination of web scraping and PAI access techniques, data related to Falcon 9 rock launches were obtained including information such as launch dates, success/failure status, mission details, payload information, and landing outcomes.
- Perform data wrangling
  - The data was processed by cleaning inconsistencies, transforming unstructured data into structured formats and engineering features to enhance predictive modeling such as one hot encoding.

Methodology Cont.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models such as SVM, K-Nearest Neighbors, Logistic Regresion, etc.
  - Utilize scikit-learn's functionalities to construct each model, perform hyperparameter tuning via techniques like GridSearchCV, and assess their performance using appropriate evaluation metrics such as accuracy and confusion matrices.

Data Collection

- Implore the use of web scraping techniques & beautiful soup python package
  - Gathered info from SpaceX public API
  - Gathered info from SpaceX’s Wikipedia page
- Retrieval of diverse information such as launch sites, mission details, payload information, landing outcomes, etc.
- Load unstructured information to structured Pandas data frame
- Subsequently perform analysis and data wrangling



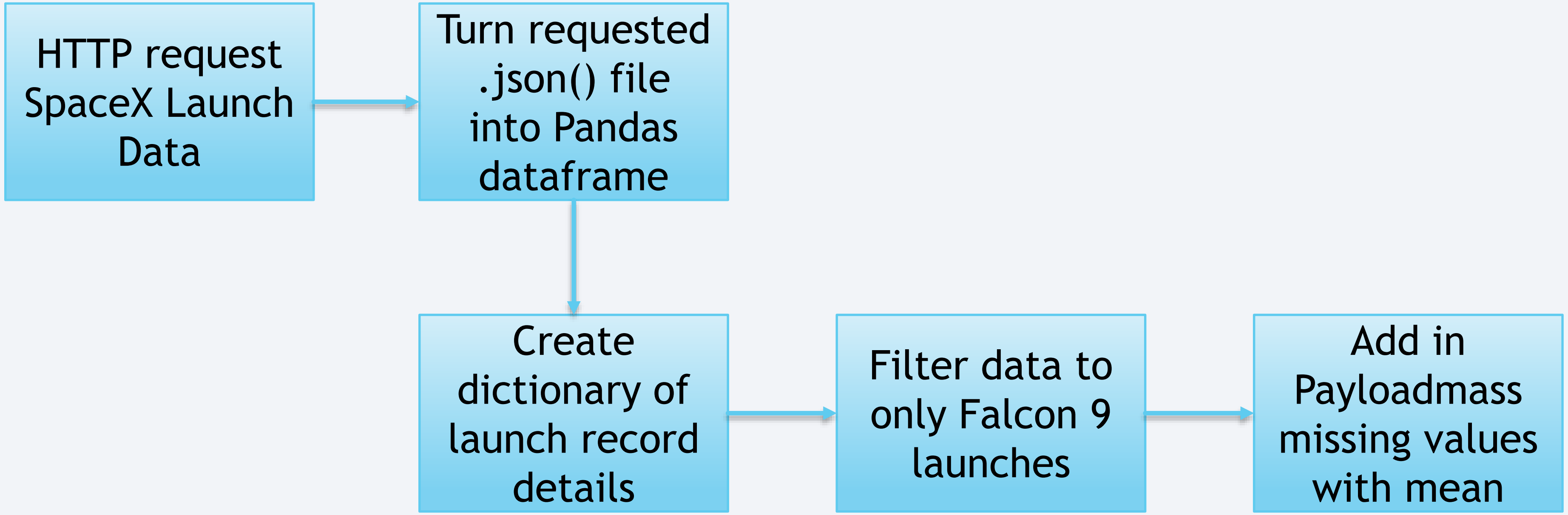
API & Web Scraping

Retrieve data

Convert to Structured Data

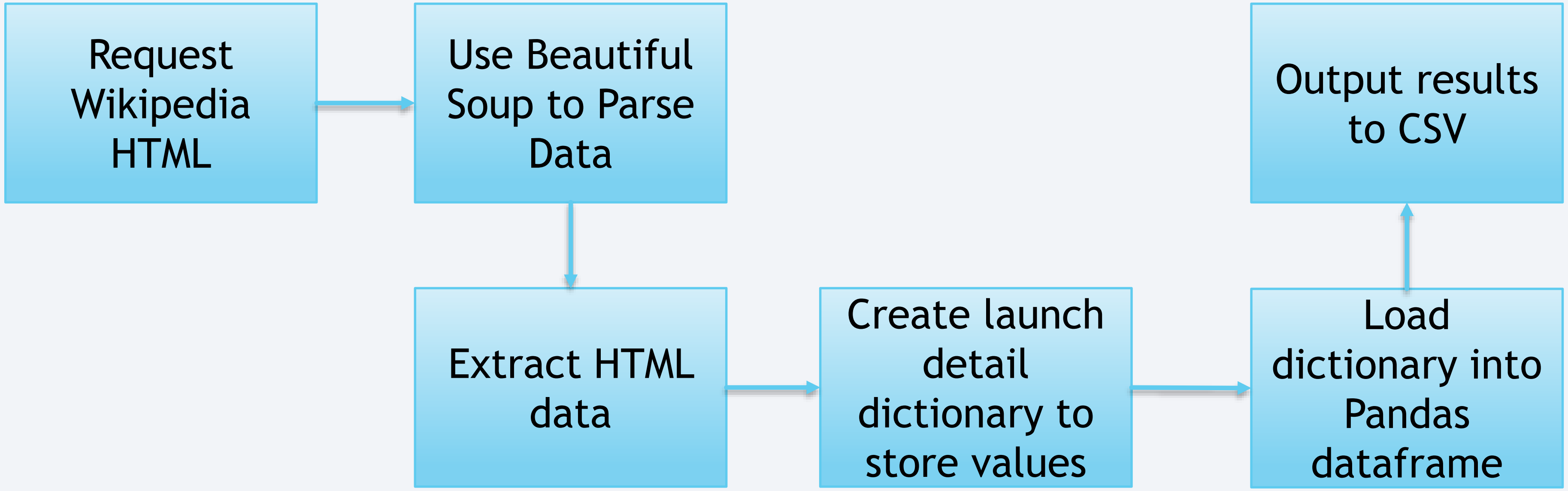
Perform analysis and data Wrangling

Data Collection – SpaceX API


HTTP request SpaceX Launch Data
Turn requested .json() file into Pandas dataframe
Create dictionary of launch record details
Filter data to only Falcon 9 launches
Add in Payloadmass missing values with mean

- SpaceX data collection API

Data Collection - Scraping



Request Wikipedia HTML

Use Beautiful Soup to Parse Data

Extract HTML data

Create launch detail dictionary to store values

Load dictionary into Pandas dataframe

Output results to CSV

- Web Scraping Falcon 9 from Wikipedia

Data Wrangling

- Data was read from CSV to Pandas data frame
- Identify key columns such as orbit, landing_outcome, payloadmass, etc.
- Decode values such as "True ASDS" and "False Ocean" to identify successful and unsuccessful missions
- Created a landing outcome column called "class" where:
  - If landing_class = 0 then it is a failed outcome
  - If landing_class = 1 then it is a successful outcome
- Export file as CSV for future analysis
- SpaceX Falcon 9 Data Wrangling

EDA with Data Visualization

- **Flight Number vs. Payload Mass -** Analyzed payload mass variation across flight numbers.
- **Flight Number vs Launch Site -** Evaluated launch site distribution over flight numbers.
- **Payload Mass vs Launch Site -** Compared payload masses at different launch sites.
- **Orbit vs. Success Rate -** Assessed mission success rates by orbit type.
- **Flight Number vs Orbit -** Investigated orbit distribution across flight numbers.
- **Payload vs Orbit -** Examined payload distribution across orbits.
- **Success Yearly Trend -** Reviewed yearly trends in mission success.

- SpaceX Falcon 9 EDA data visualizations

EDA with SQL

- Loaded SpaceX CSV data into Pandas data frame
- Calculate the total payload mass launched by NASA (CRS) missions.
- Compute the average payload mass for missions with booster version 'F9 v1.1'.
- Find the earliest date of a successful landing on a ground pad.
- Identify distinct booster versions for missions with successful landings on a drone ship, with payload mass between 4000 and 6000 kg.
- Count the total number of successful and failed mission outcomes.
- Find the booster version associated with the maximum payload mass.
- Convert the date format to month-year and filter data for failed missions on a drone ship in 2015.

- SpaceX Falcon 9 EDA SQL

Build an Interactive Map with Folium

- **Launch Sites Markers -** Created markers to indicate the locations of launch sites on the map, providing visual reference points for the geographical distribution of launches.
- **Successful and Unsuccessful Landings Circles -** Utilized circles to represent successful and unsuccessful landing locations, offering a visual depiction of the outcomes of Falcon 9 landings at various sites.
- **Proximity Examples to Key Locations -** Incorporated lines and markers to illustrate the proximity of launch sites to key locations such as railways, highways, coasts, and cities. This visualization aids in understanding the accessibility and strategic positioning of launch sites relative to important infrastructure and population centers.
- SpaceX Falcon 9 Folium Map

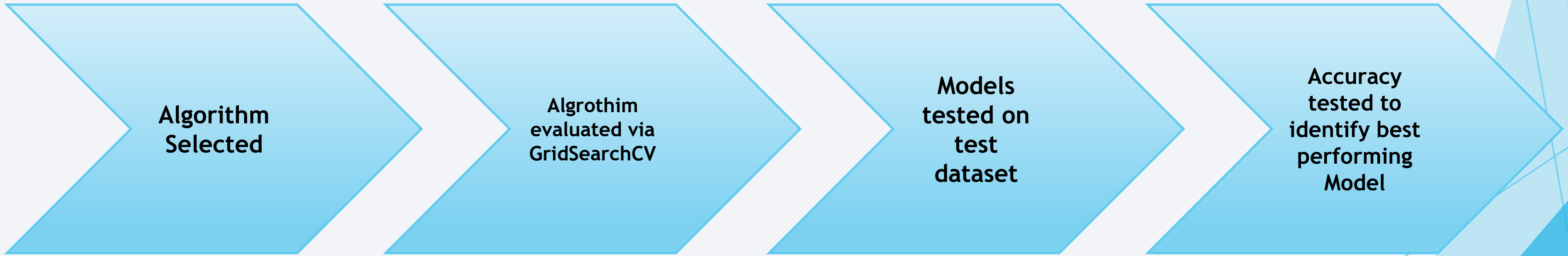
Build a Dashboard with Plotly Dash

- **Launch Site Dropdown -** Enabled selection of launch sites to filter data, providing users with the ability to focus on specific locations of interest.
- **Pie Chart for Success Launches -** Displayed the total count of successful launches for all locations or the success vs. failed launches count for a selected site, allowing users to see the corresponding success rate.
- **Payload Range Slider -** Allowed users to specify a payload mass range using a slider, facilitating exploration of the relationship between payload mass and launch success.
- **Scatter Chart for Payload vs. Launch Success -** Visualized the correlation between payload mass and launch success for all locations or a selected site within the specified payload range, aiding in understanding the impact of payload mass on mission outcomes.

- SpaceX Falcon 9 Plotly Interactive Dashboard

Predictive Analysis (Classification)

- The predictive analysis performed used various classification algorithms (Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbors) and the python package scikit-learn.
  - Each algorithm is evaluated using cross-validation and hyperparameter tuning via GridSearchCV to find the best performing model configuration.
  - The models are then tested on the test dataset, and their performance is assessed using accuracy scores and confusion matrices.
  - The classification model with the highest accuracy on the test dataset is identified as the best performing model.



Algorithm Selected

Algrothim evaluated via GridSearchCV

Models tested on test dataset

Accuracy tested to identify best performing Model

  - [SpaceX Falcon 9 Machine Learning Predictive Analysis]

Results

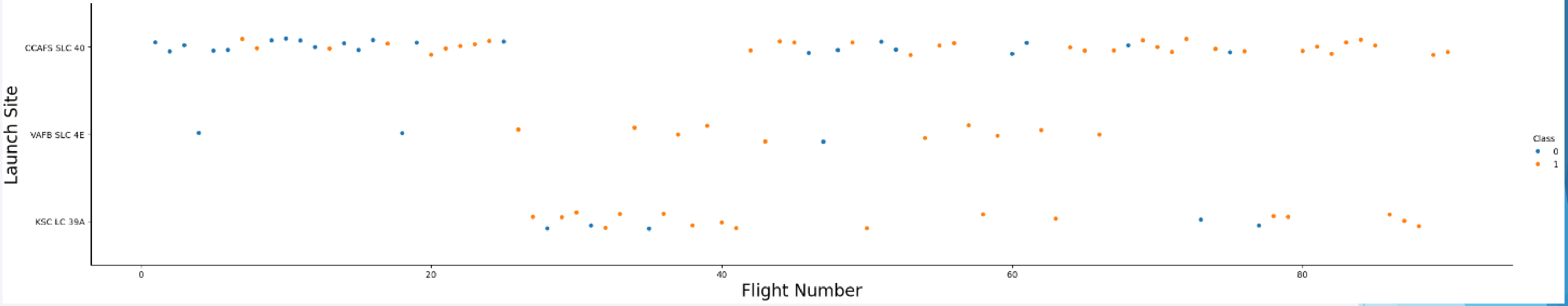
- Project Goal: Predict successful landing of Falcon 9 first stage.
- Importance: Estimating launch costs critical for competitive bidding in rocket launch market.
- Challenges Addressed: Predicting landing success, estimating launch costs, providing insights for competitors.
- Methodology: Employed various classification algorithms (Logistic Regression, SVM, Decision Tree, KNN).
- Performance: Decision Tree model achieved highest accuracy of 90% in predicting successful landings.
- Implications: Reliable prediction of landing success facilitates accurate launch cost estimation and enhances competitiveness in the aerospace industry.

The analysis aimed to predict the successful landing of the Falcon 9 first stage, crucial for estimating launch costs and informing competitive bids in the aerospace industry. By employing various classification algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors, the study evaluated their performance in predicting landing outcomes. Among these models, the Decision Tree algorithm demonstrated the highest accuracy of 90% in predicting successful landings, thus providing valuable insights for estimating launch costs and facilitating competitive bidding processes in the rocket launch market.

Section 2

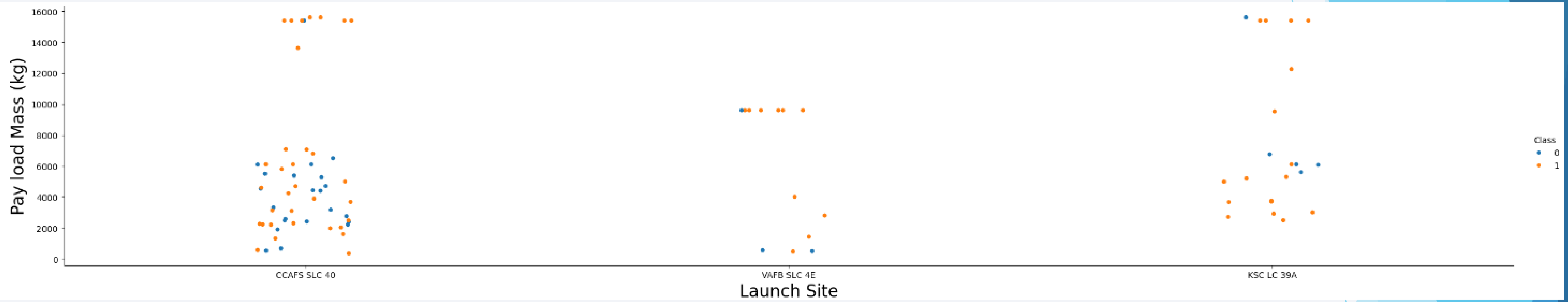
Insights drawn from EDA

Flight Number vs. Launch Site



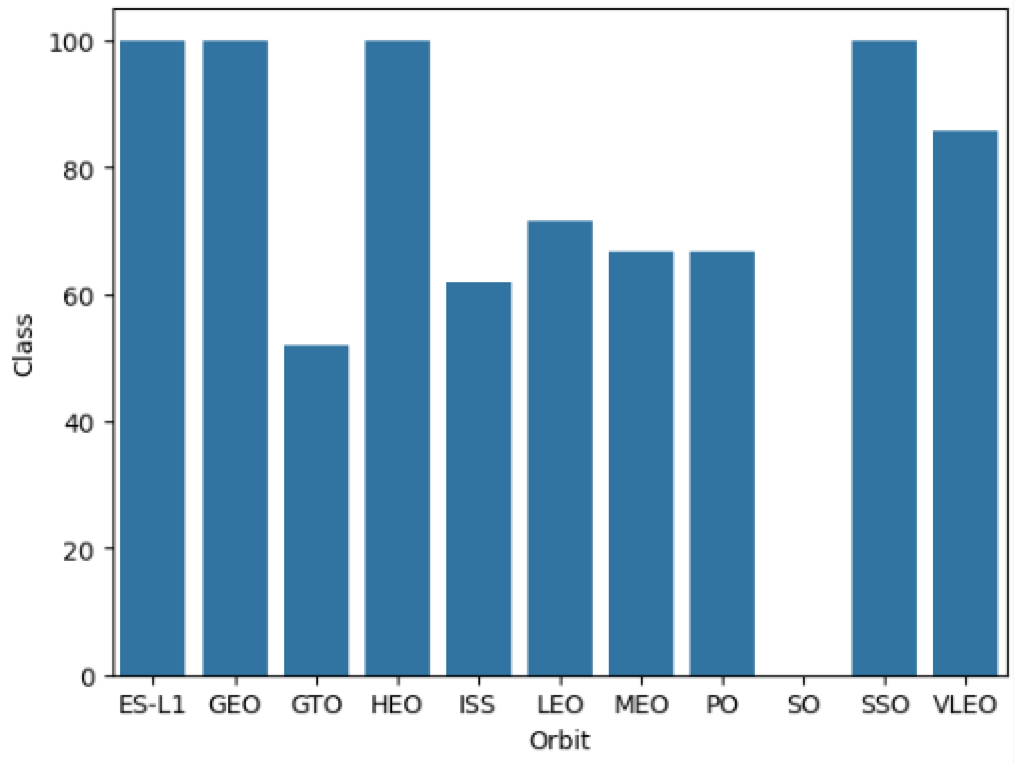
- In this scenario, each dot represents a launch, with the x-axis denoting the flight number and the y-axis representing the launch site. By coloring the dots blue for failed attempts (where class = 0) and orange for successful attempts (where class = 1), the plot visually distinguishes between successful and unsuccessful launches. This visualization helps to identify that the earlier flight numbers had significantly more failed outcomes compared to later flight numbers.

Payload vs. Launch Site



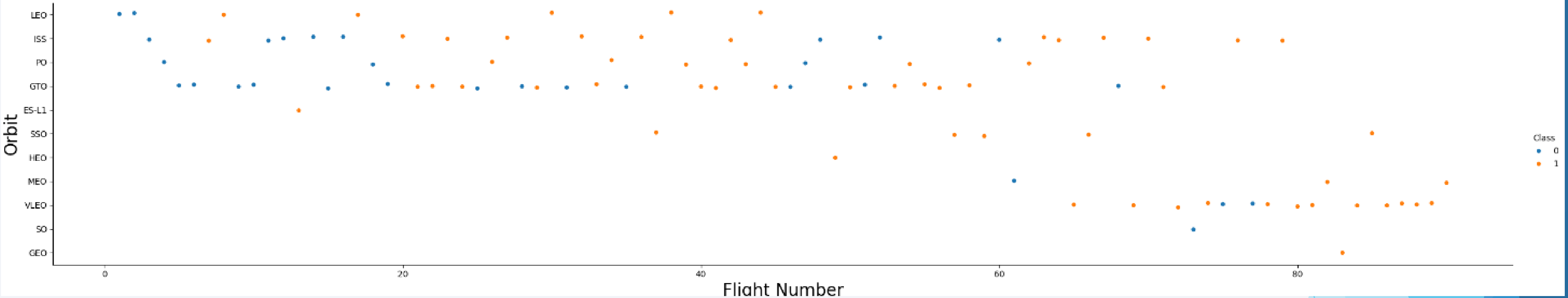
- In this scenario, each dot represents a launch, with the x-axis denoting the launch site and the y-axis representing the Payload Mass (KG). By coloring the dots blue for failed attempts (where class = 0) and orange for successful attempts (where class = 1), the plot visually distinguishes between successful and unsuccessful launches. This plot helps to visualize which launch sites have the high concentrations of successes and failures based on Payload Mass. CCAFS SLC 40 has both the highest number of launches and greatest number of failures with lower payload mass.

Success Rate vs. Orbit Type



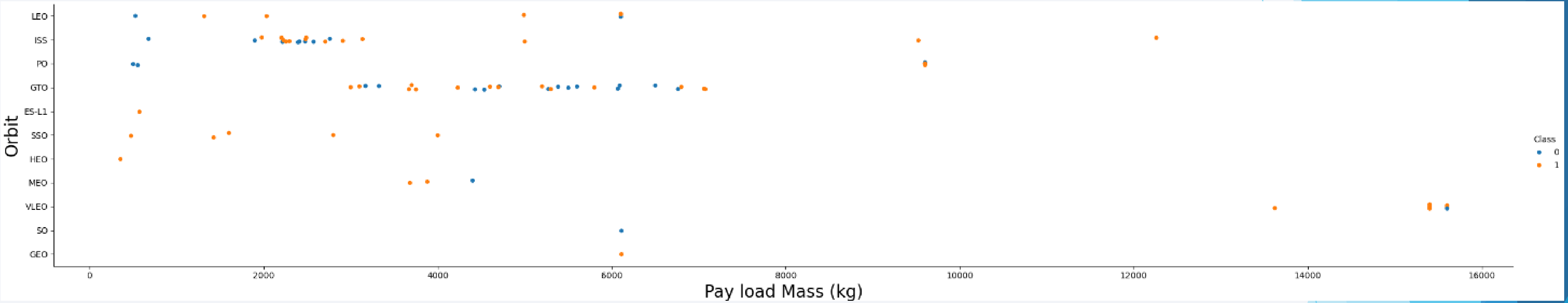
- This bar chart of class vs orbit serves the purpose of visually comparing the distribution of successful and failed launch outcomes across different orbital paths. Each bar represents a specific orbit, with the height of the bar indicating the frequency of successful and failed launches in that orbit category. This visualization makes it clear that the orbital path has a great impact on the success rate.

Flight Number vs. Orbit Type



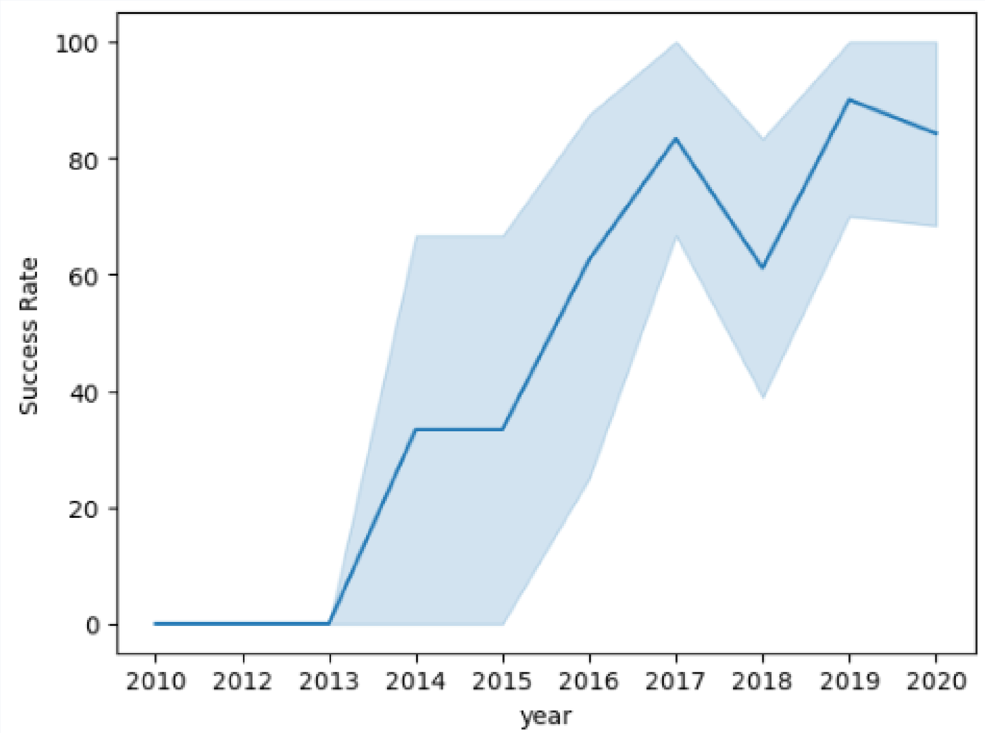
- In this scenario, each dot represents a launch, with the x-axis denoting the flight number and the y-axis representing the orbit. By coloring the dots blue for failed attempts (where class = 0) and orange for successful attempts (where class = 1), the plot visually distinguishes between successful and unsuccessful launches. It is clear from this visualization that the orbit paths GTO has the greatest number of failures with earlier flight numbers and improves with more flights.

Payload vs. Orbit Type



- In this scenario, each dot represents a launch, with the x-axis denoting the flight number and the y-axis representing the launch site. By coloring the dots blue for failed attempts (where class = 0) and orange for successful attempts (where class = 1), the plot visually distinguishes between successful and unsuccessful launches. This scatter plot aids in identifying how payload mass varies across different orbital paths and provides valuable information for mission planning/payload selection.

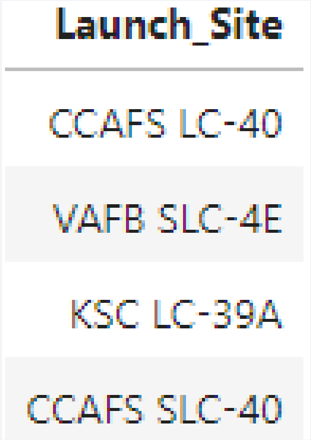
Launch Success Yearly Trend



- By plotting success rate against year and including a trend line, this graph provides insights into the performance of Falcon 9 launches over the years. It allows the observation of any long-term patterns or changes in launch success rates and assess the effectiveness of improvements or changes implemented by SpaceX over time.

All Launch Site Names

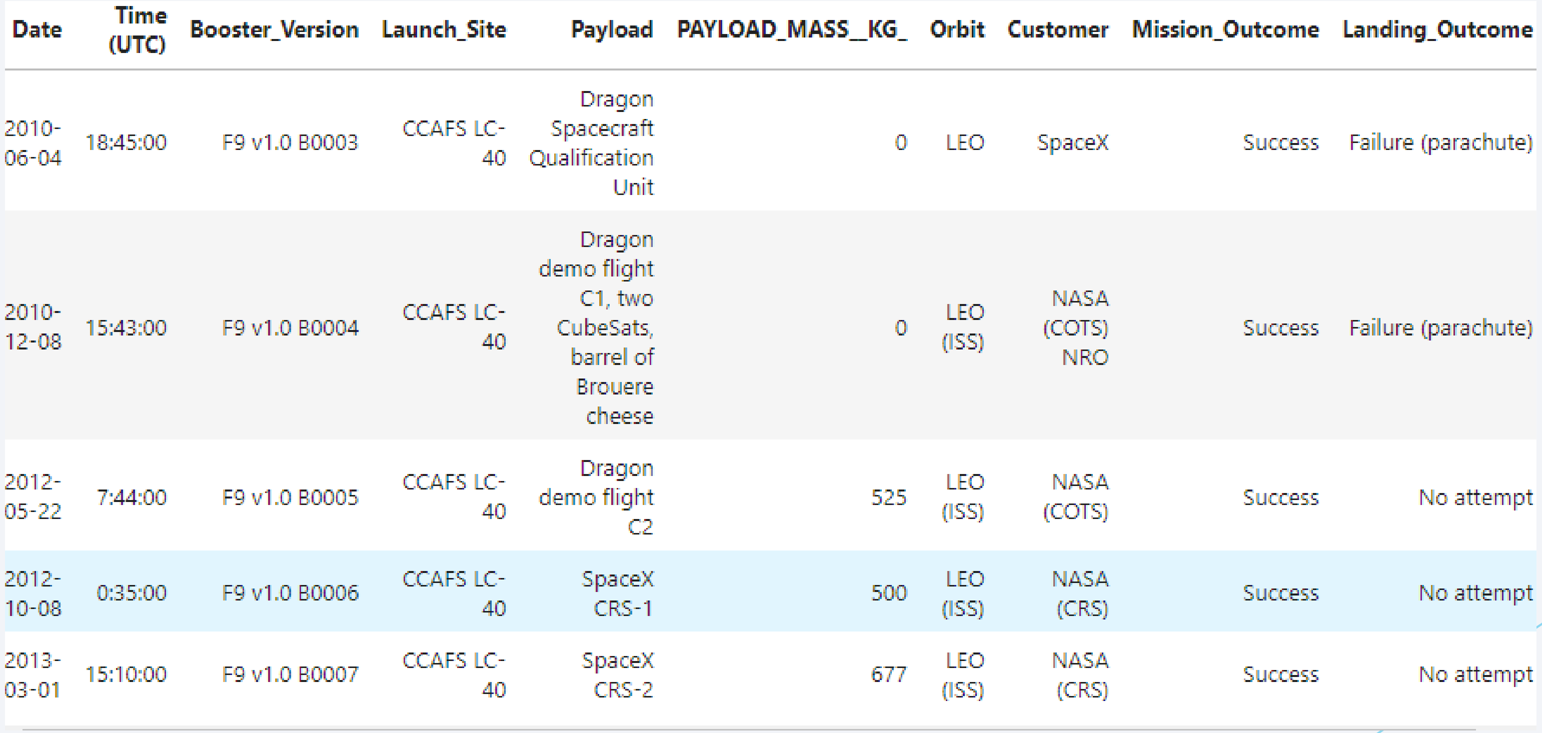
- The following query retrieves the unique launch sites from the SPACEXTABLE dataset, providing a list of distinct launch locations where Falcon 9 launches have occurred:
    - %sql SELECT DISTINCT Launch_Site from SPACEXTABLE



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

Launch Site Names Begin with 'CCA'

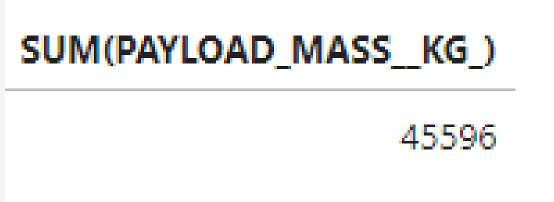
- This SQL query selects the first five records from the SPACEXTABLE dataset where the launch site begins with 'CCA', providing a glimpse of the details for Falcon 9 launches specifically from sites with names starting with 'CCA’:
  - %sql SELECT * from SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5



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

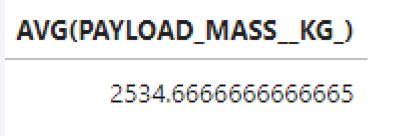
- The following query calculates the total payload mass (in kilograms) for launches conducted by the customer "NASA (CRS)" from the SPACEXTABLE dataset, grouping the results by the customer name.Present your query result with a short explanation here
  - %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)' GROUP BY CUSTOMER



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

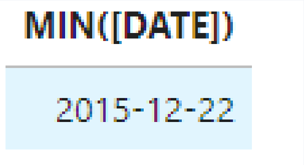
- The following query calculates the average payload mass (in kilograms) for launches where the booster version contains the term "F9 v1.1" from the SPACEXTABLE dataset.
    - %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE '%F9 v1.1%



```
AVG(PAYLOAD_MASS__KG_)
        2534.6666666666665
```

First Successful Ground Landing Date

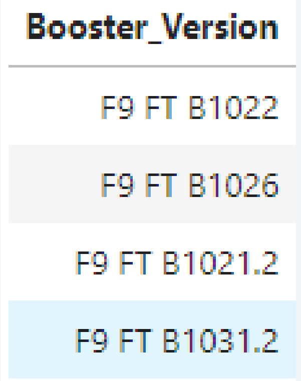
- The following query retrieves the earliest date of successful landings on ground pads from the SPACEXTABLE dataset.
  - %sql SELECT MIN([DATE]) FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)'



MIN([DATE])
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

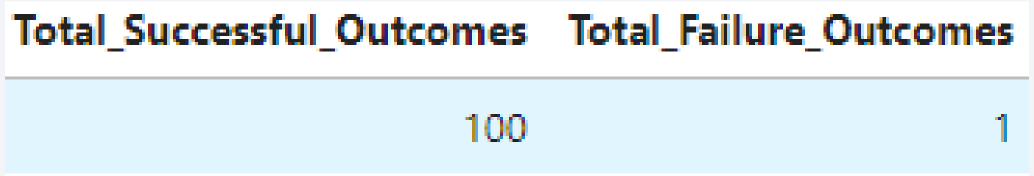
- The following query selects unique booster versions from the SPACEXTABLE dataset for launches with successful landings on drone ships and payload masses between 4000 and 6000 kilograms.
  - %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)



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

Total Number of Successful and Failure Mission Outcomes

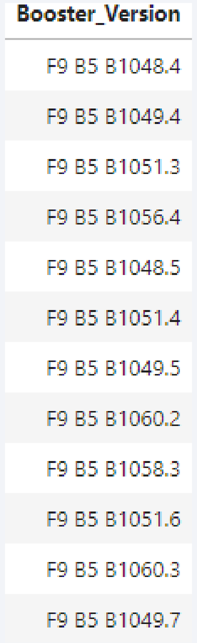
- The following query calculates the total number of successful and failed mission outcomes from the SPACEXTABLE dataset by counting occurrences where the mission outcome contains the term 'Success' or 'failure', respectively.
  - %sql SELECT sum(CASE WHEN Mission_Outcome like '%Success%' THEN 1 ELSE 0 END) AS Total_Successful_Outcomes, sum(CASE WHEN Mission_Outcome like '%failure%' THEN 1 ELSE 0 END) AS Total_Failure_Outcomes from SPACEXTABLE



Total_Successful_Outcomes	Total_Failure_Outcomes
100	1

Boosters Carried Maximum Payload

- The following query selects unique booster versions from the SPACEXTABLE dataset where the payload mass is equal to the maximum payload mass recorded in the dataset.
  - %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

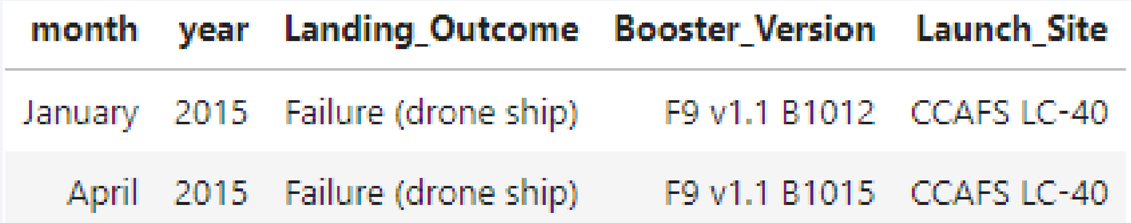


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- The following query extracts the month and year from the Date column, converts the numeric month to its corresponding name (e.g., '01' becomes 'January'), and selects specific columns such as month, year, landing outcome, booster version, and launch site from the SPACEXTABLE dataset. It filters the results to include only records where the landing outcome is 'Failure (drone ship)' and the year is '2015'.

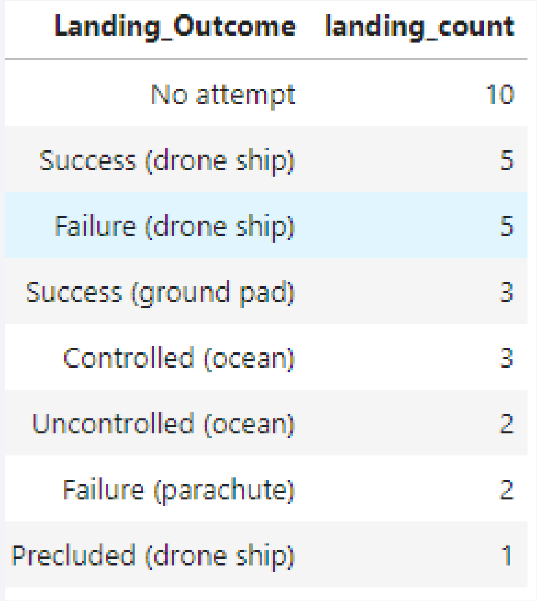
  - %sql select distinct case when substr(Date, 6,2) = '01' then 'January' when substr(Date, 6,2) = '02' then 'February' when substr(Date, 6,2) = '03' then 'March' when substr(Date, 6,2) = '04' then 'April' when substr(Date, 6,2) = '05' then 'May' when substr(Date, 6,2) = '06' then 'June' when substr(Date, 6,2) = '07' then 'July' when substr(Date, 6,2) = '08' then 'August' when substr(Date, 6,2) = '09' then 'September' when substr(Date, 6,2) = '10' then 'October' when substr(Date, 6,2) = '11' then 'November' when substr(Date, 6,2) = '12' then 'December' end as month, substr(Date,0,5) as year, landing_outcome, booster_version, launch_site from SPACEXTABLE where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5) ='2015'



month	year	Landing_Outcome	Booster_Version	Launch_Site
January	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The following query counts the occurrences of each landing outcome (success or failure) within the specified date range from June 4, 2010, to March 20, 2017, in the SPACEXTABLE dataset. It groups the results by landing outcome and orders them by the count of occurrences in descending order.
  - %sql SELECT landing_outcome, COUNT(*) AS landing_count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY landing_count DESC;



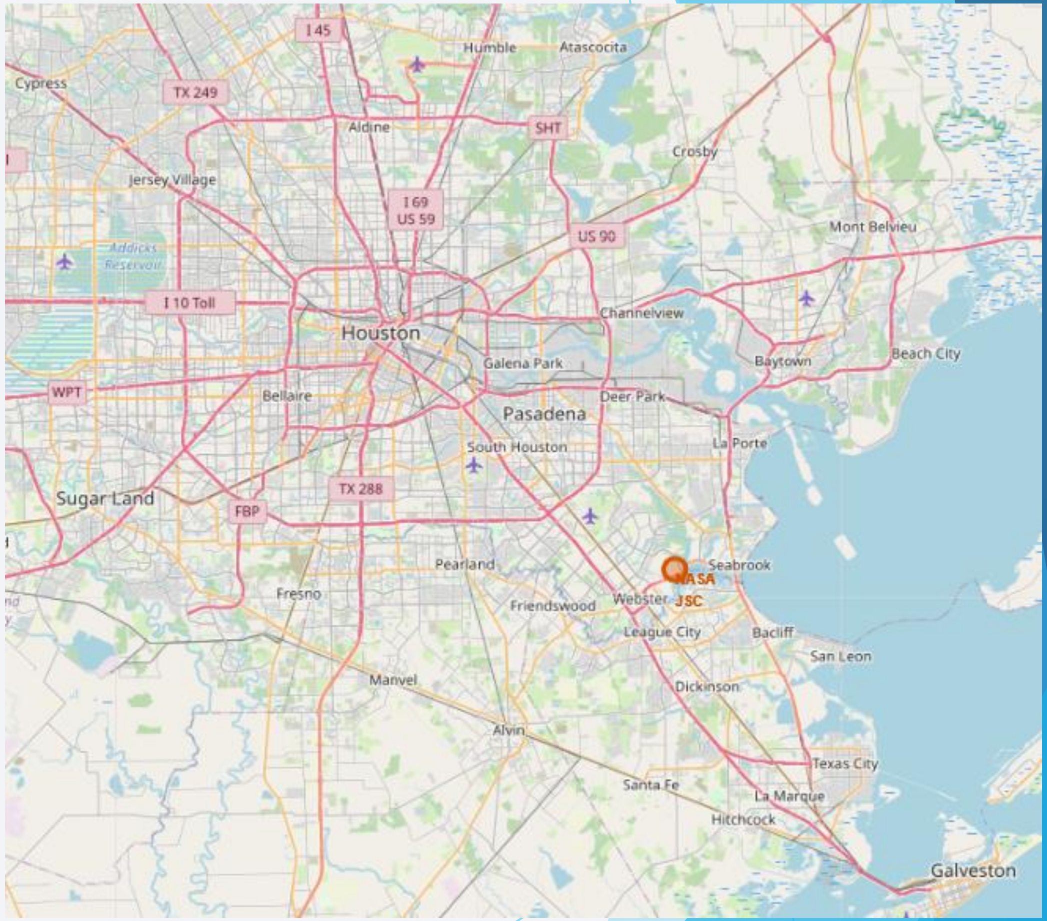
Landing_Outcome	landing_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

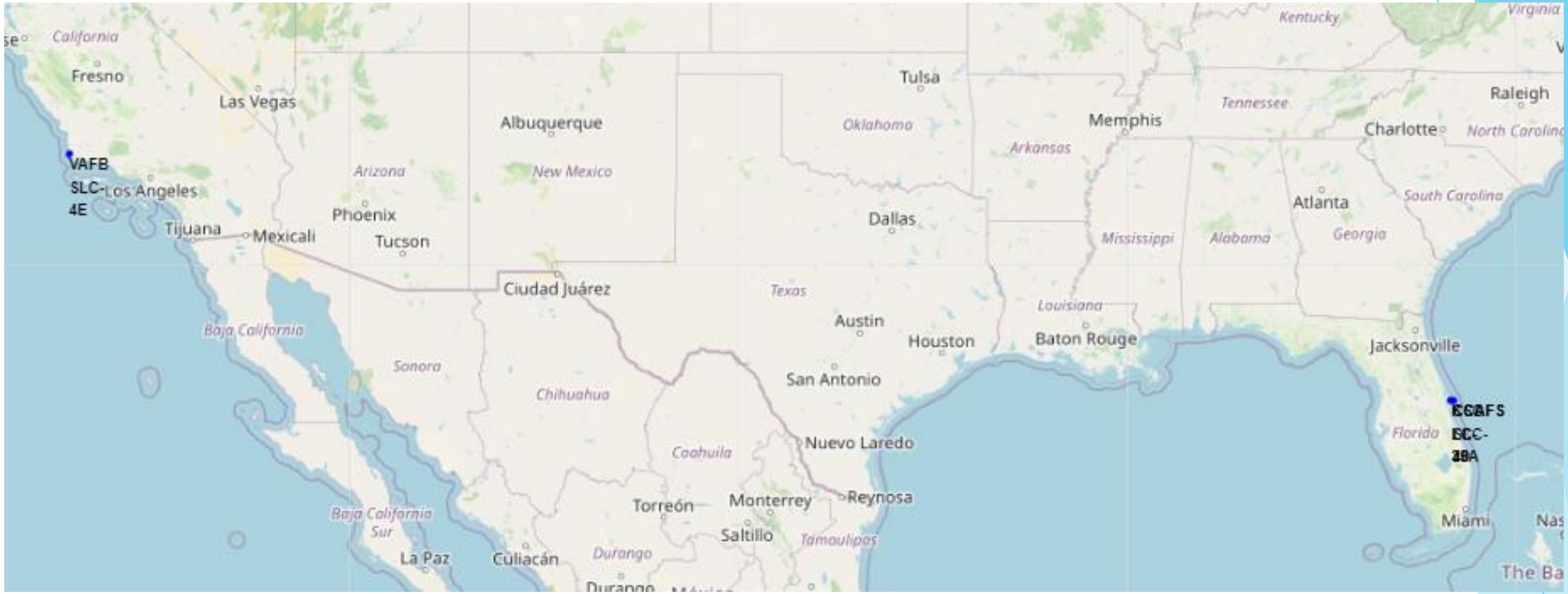
Launch Sites Proximities Analysis

Folium Map NASA JSC Marker

- The screenshot in Folium marks the location of NASA JSC, providing a clear visual reference on the map interface. It includes essential elements such as a distinct marker, satellite imagery, and navigation controls, aiding users in identifying and exploring the site's geographic context.

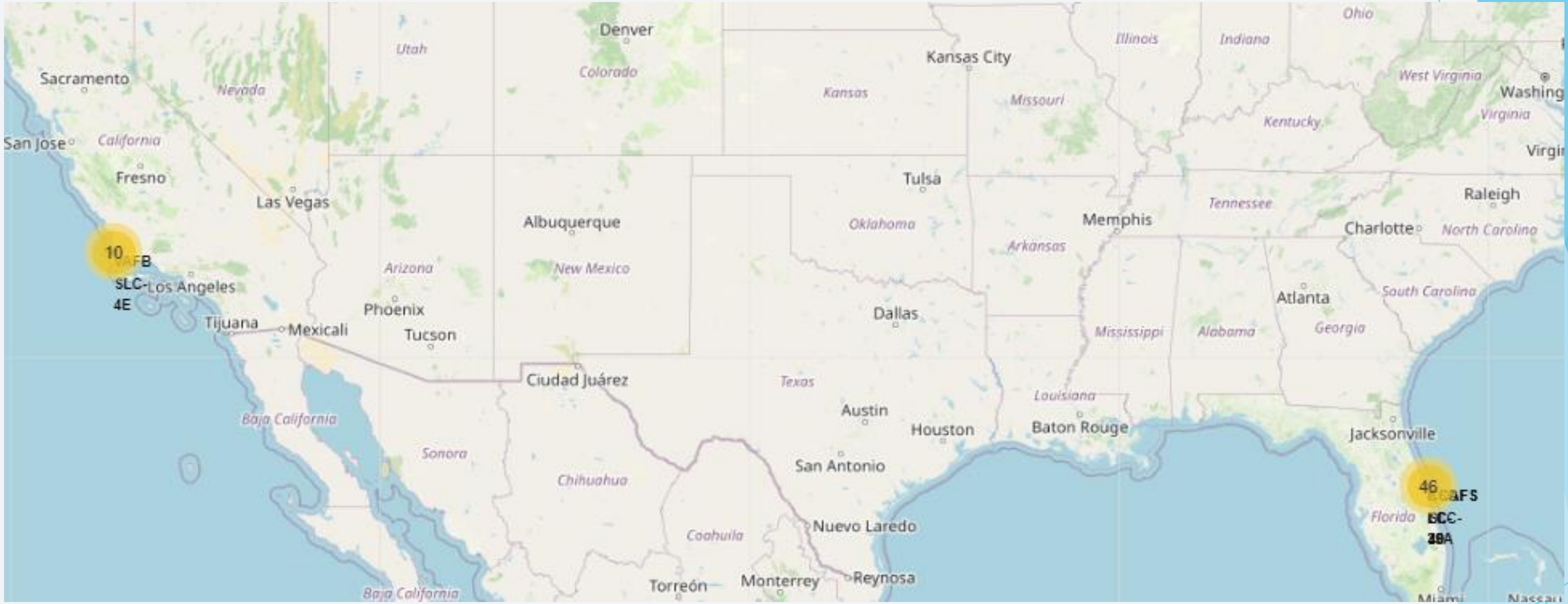


Folium Map All Launch Site Markers



- The US Folium map of launch sites displays distinct markers representing various launch locations across the country, providing users with a clear visual reference. It offers essential navigation controls and may include additional information about each site, facilitating exploration and understanding of the geographical context of space launch activities in the United States.

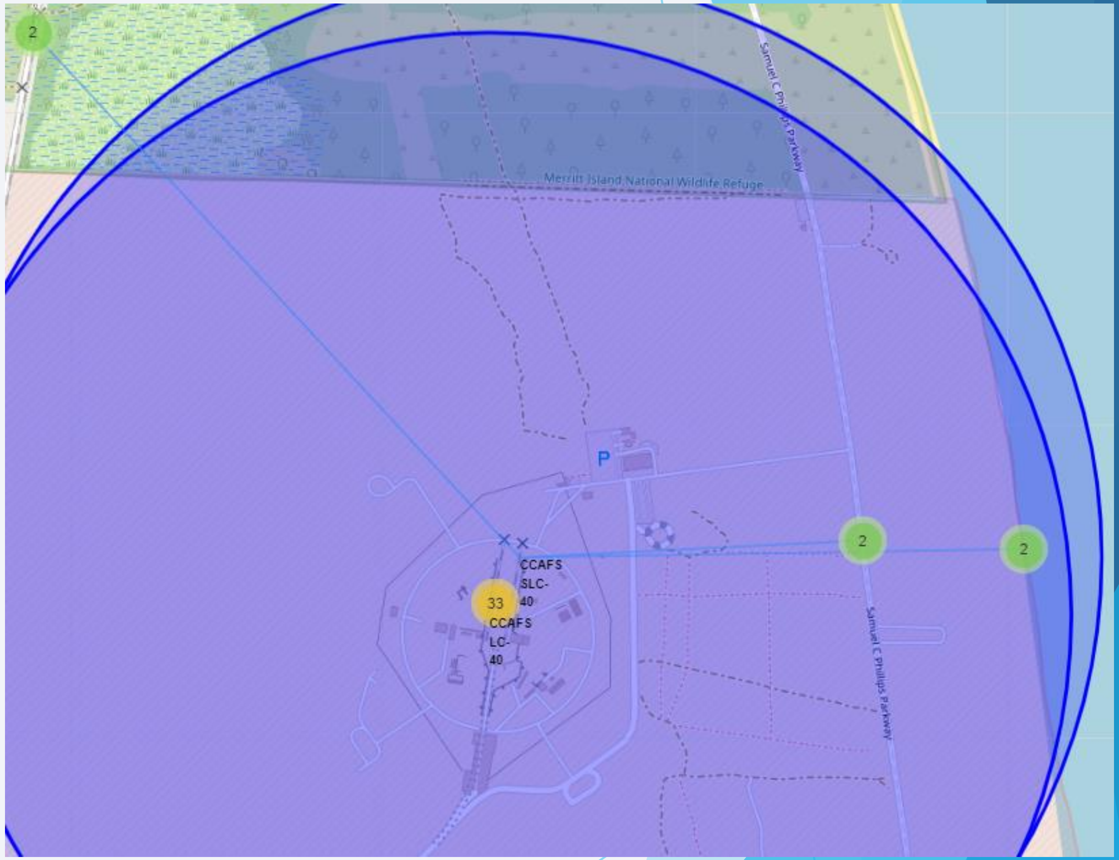
Folium Map Color Labeled Launch Outcomes

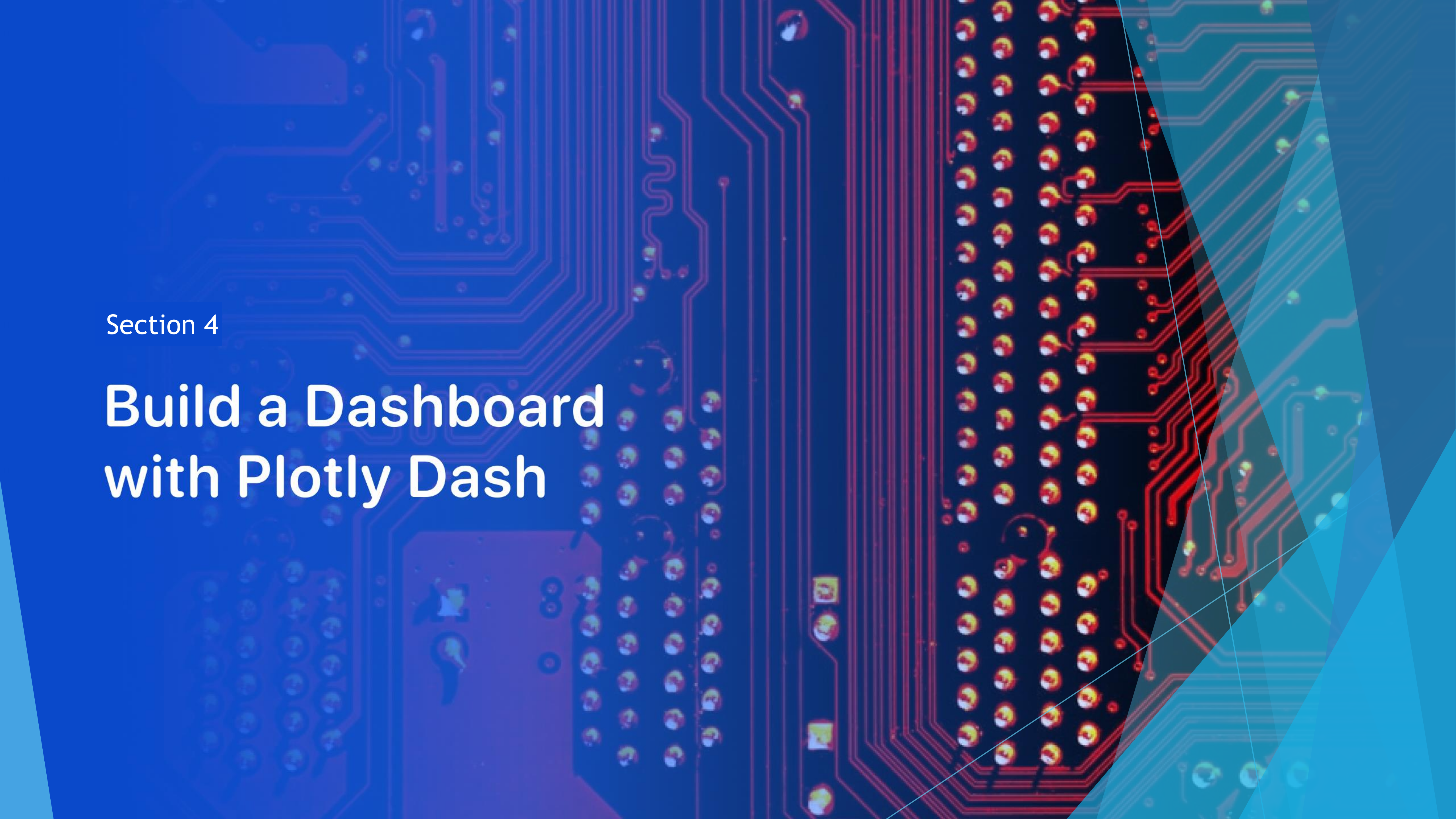


The US Folium map of launch outcomes displays distinct colored circles representing various launch outcomes across the country, providing users with a clear visual reference. It offers essential navigation controls and may include additional information about each site, facilitating exploration and understanding of the geographical context of space launch activities in the United States.

CCAFS SLC 40 Launch Site Proximities

- The screenshot in Folium marks the location of CCAFS SLC 40, providing a clear visual reference on the map interface. It also includes markers indicating the location of the nearest coastline, highway and railway. This map also shows a proximity line that indicates the distances to the launch site when clicked on.

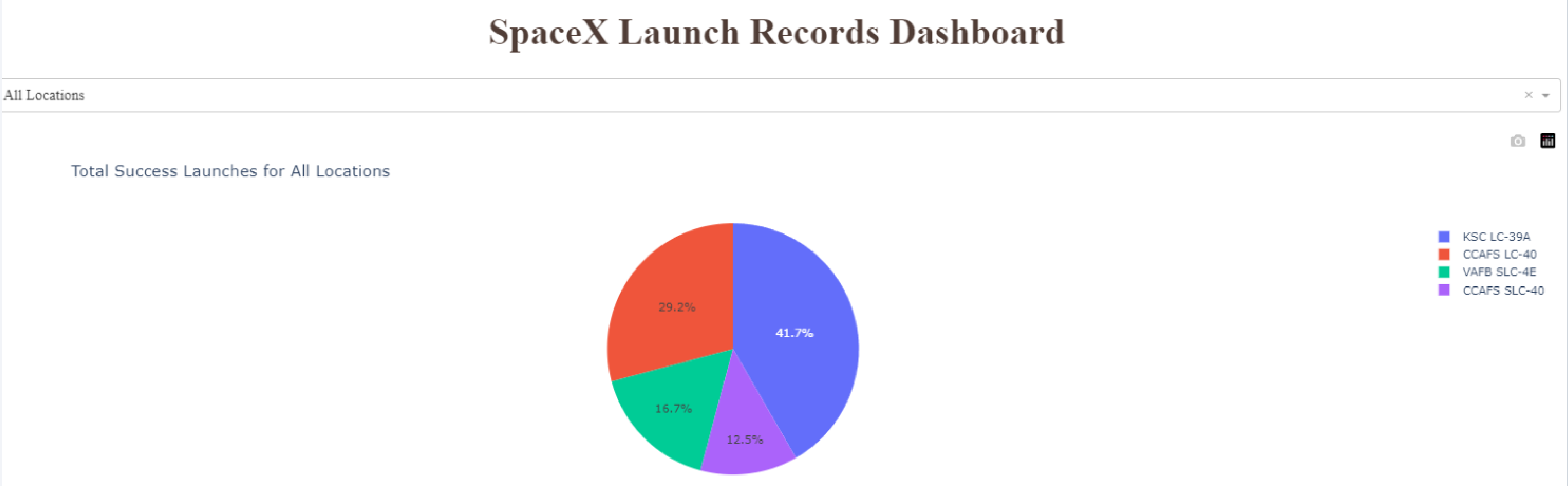

2
Samuel C Phillips Parkway
Merritt Island National Wildlife Refuge
P
2
2
CCAFS SLC-40
33
CCAFS LC-40
Samuel C Phillips Parkway



Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



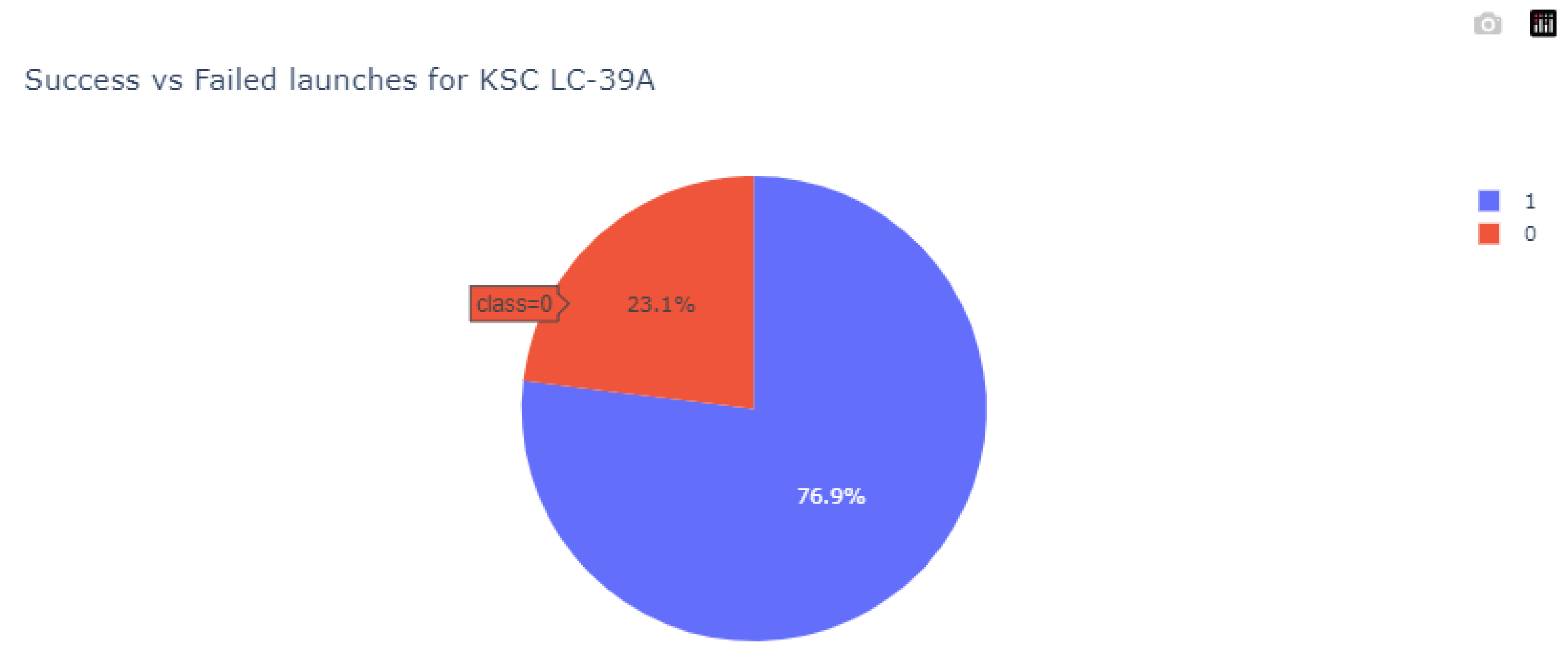
SpaceX Launch Records Dashboard

All Locations

Total Success Launches for All Locations

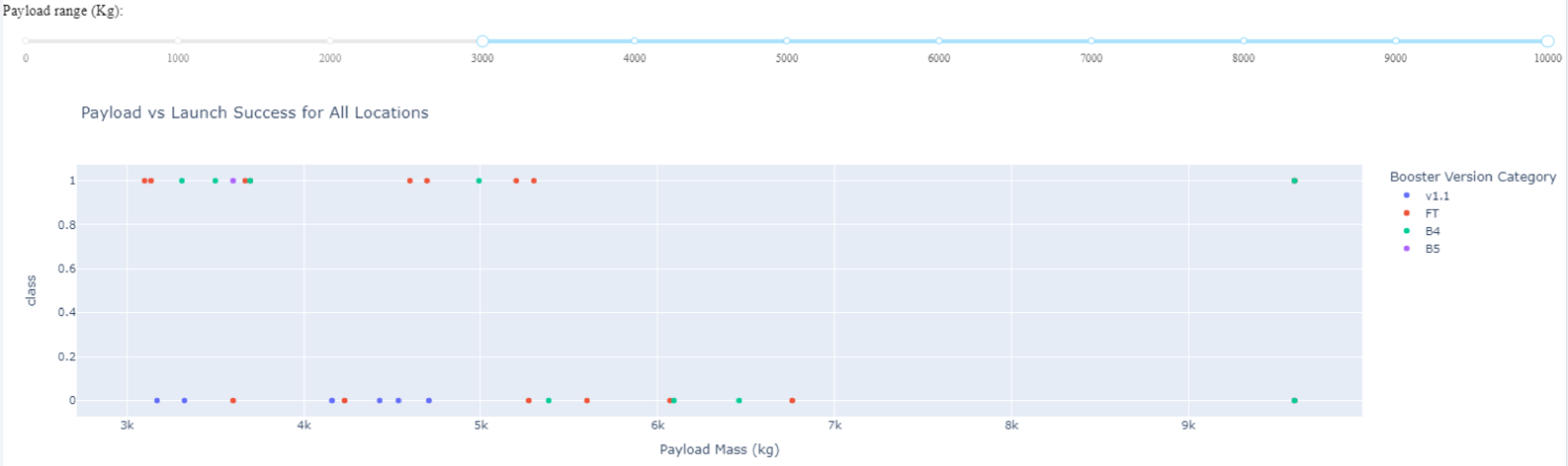
The screenshot of this Plotly Dashboard indicates the success rate based on the launch locations listed in the legends on the right. The dashboard is interactive and when the drop down is selected, a launch site can be selected for additional information on its success rate.

SpaceX Launch Records Dashboard – KSC LC-39A



When the drop down is selected for a launch site, additional information can be gathered regarding its success rate. In this example, the highest success rate launch site KSC LC-39A launch was selected. This pie chart indicates that the launches for KSC LC-39A were successful 76.9% of the time and 23.1% were unsuccessful.

SpaceX Launch Records Dashboard – Payload vs Launch Outcome Scatter Plot



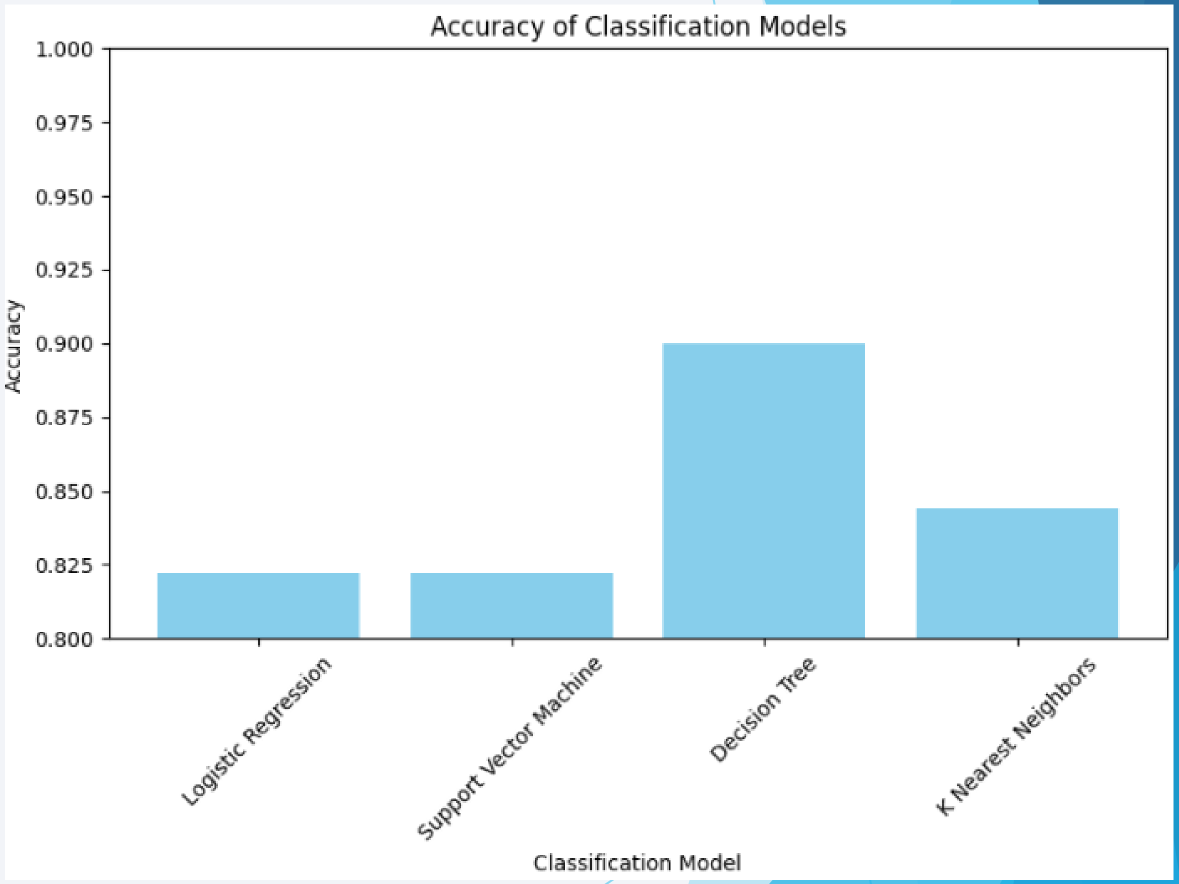
The Plotly scatter plot aims to illustrate the relationship between successful launches and payload mass, with a slider enabling users to adjust the range of payload mass displayed. Additionally, the use of different colors represent booster versions. This provides insights into potential correlations between successful launches, payload mass, and the specific booster versions utilized.

Section 5

Predictive Analysis (Classification)

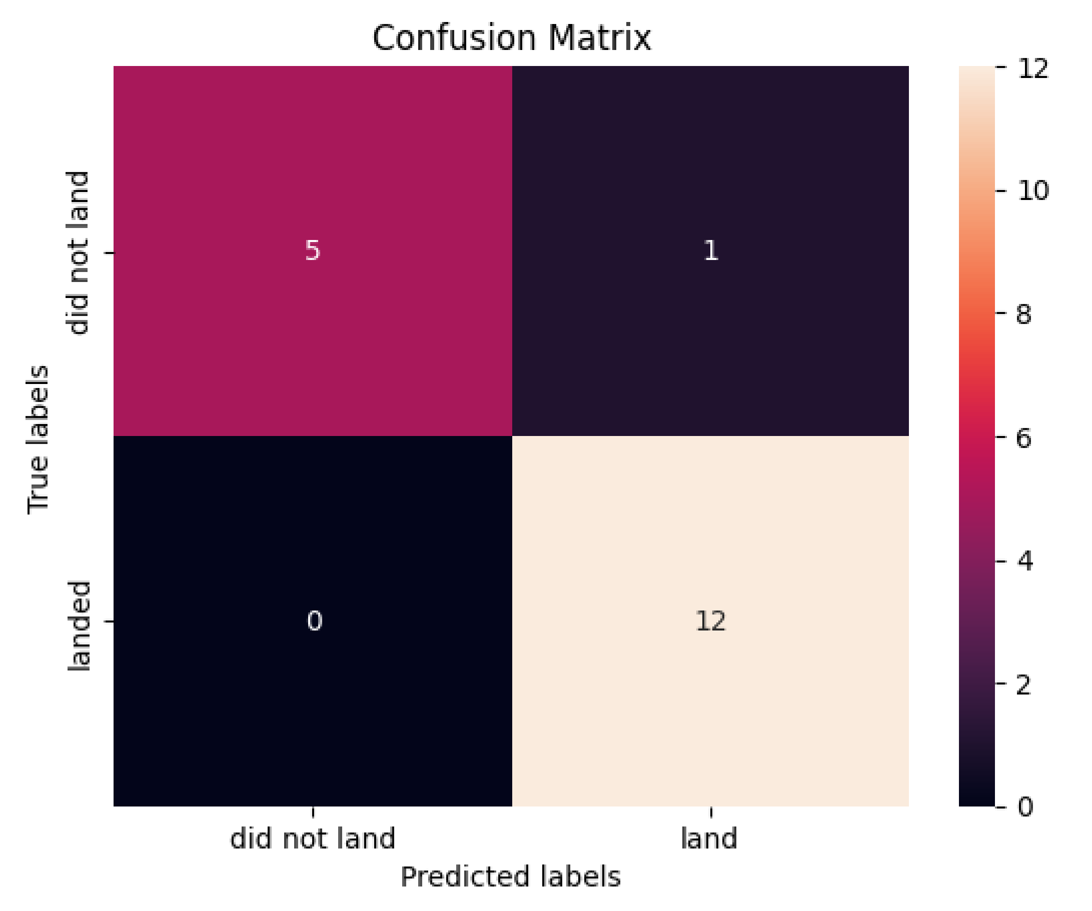
Classification Accuracy

- The following screenshot of a bar chart visualizes the accuracy of different classification models used in the project. Each bar represents a model, with the height of the bar indicating its accuracy score. From the chart, it's evident that the Decision Tree model achieved the highest accuracy, followed by K Nearest Neighbors while Logistic Regression and Support Vector Machine performed slightly lower in accuracy.



Confusion Matrix

- The following confusion matrix summarizes the performance of the classification model by showing the correct and incorrect predictions. In this model:
  - The model correctly predicted "landed" 5 times.
  - The model incorrectly predicted "not landed" 1 time when the actual label was "landed".
  - The model incorrectly predicted "landed" 0 times when the actual label was "not landed".
  - The model correctly predicted "not landed" 12 times.



Conclusions

- Classification models like Decision Trees, Logistic Regression, SVM, and KNN were employed to predict successful Falcon 9 first stage landings.
- Decision Tree model exhibited the highest accuracy, 90%, among the tested models, followed closely by K-Nearest Neighbors.
- Understanding launch site, payload mass, booster version, and orbit are crucial factors in predicting successful landings.
- Exploratory Data Analysis highlighted trends such as success rates varying by launch site and the correlation between payload mass and launch success.
- Time-series analysis revealed a positive trend in launch success rates over the years.
- SQL queries were utilized to extract specific data subsets for in-depth analysis, such as payload mass statistics and launch outcome counts.
- This project showcased the significance of data preprocessing, model tuning, and evaluation for accurate predictive modeling.

Appendix

- Github Repository with all scripts, data extracts, and screenshots can be found here:
  - IBM Course 10 Capstone Project - Jason Echevarria

Thank you!