IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luis Requejo

Jan 02, 2025

Outline

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

Executive Summary

- Summary of methodologies

  1. Data Collection: Utilized web scraping and data wrangling techniques to compile a comprehensive dataset.
  2. Data Understanding: Gained insights into the dataset through:
     - SQL queries
     - Exploratory data analysis (EDA)
     - Visual analytics and map generation using Folium
  3. Model Development: Built predictive models for landing success using machine learning techniques.

- Summary of all results

  - We applied various machine learning algorithms, including Logistic Regression, SVM, and KNN, to predict the likelihood of SpaceX reusing the first stage of its Falcon 9 rocket. By aggregating data from multiple sources, we developed a robust database for analysis and forecasting.
  - Our models achieved an average prediction accuracy of 83.3%, offering a competitive edge in the commercial space sector by enhancing the reliability of reusability forecasts.

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches at a cost of $62 million per launch, significantly lower than the $165 million charged by other providers. This cost efficiency is largely attributed to SpaceX's ability to reuse the rocket's first stage. Predicting whether the first stage will successfully land is crucial, as it directly impacts the overall cost of a launch. Such predictions could be invaluable for alternative companies aiming to compete with SpaceX for rocket launch contracts.

- Problems you want to find answers
  - This project leverages various data science methodologies and techniques to predict the outcome of Falcon 9 rocket landings. Accurate predictions can help save both time and money, providing insights that could optimize decision-making in the commercial space industry.

Section 1

Methodology

Methodology

Executive Summary

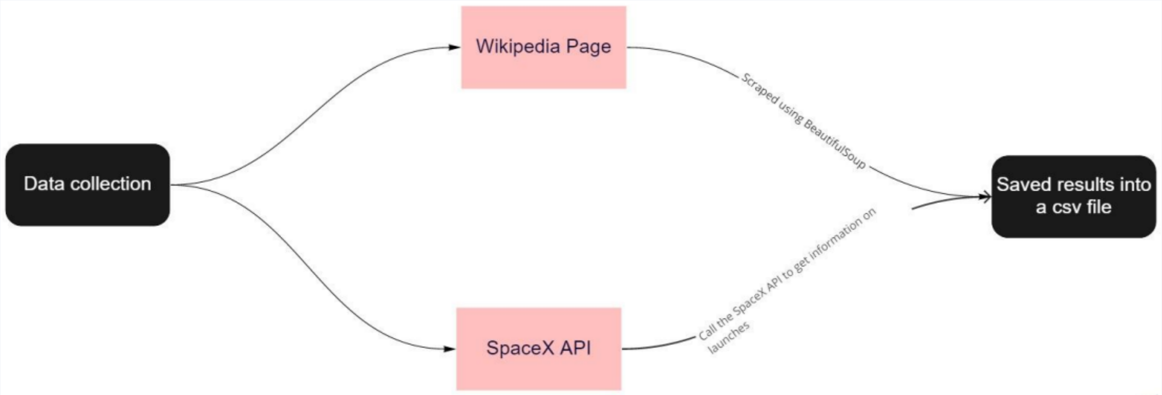
- Data collection methodology:
  - Data was collected from a SpaceX API and scraped from a Wikipedia page containing information about all past SpaceX rocket launches
- Perform data wrangling
  - The launch outcomes were split into "good" outcomes and "bad" outcomes and assigned a numerical value of 0 or 1, missing values were replaced with averages for that column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were built using the scikit learn python package, fine tuned using GridSearchCV and evaluated using the model's accuracy score and confusion matrix

Data Collection

- API
  - Request and parse the SpaceX launch data using the GET request
  - Filter the data frame to only include Falcon 9 launches
  - Dealing with Missing Values by replacing it with mean values for PayloadMass

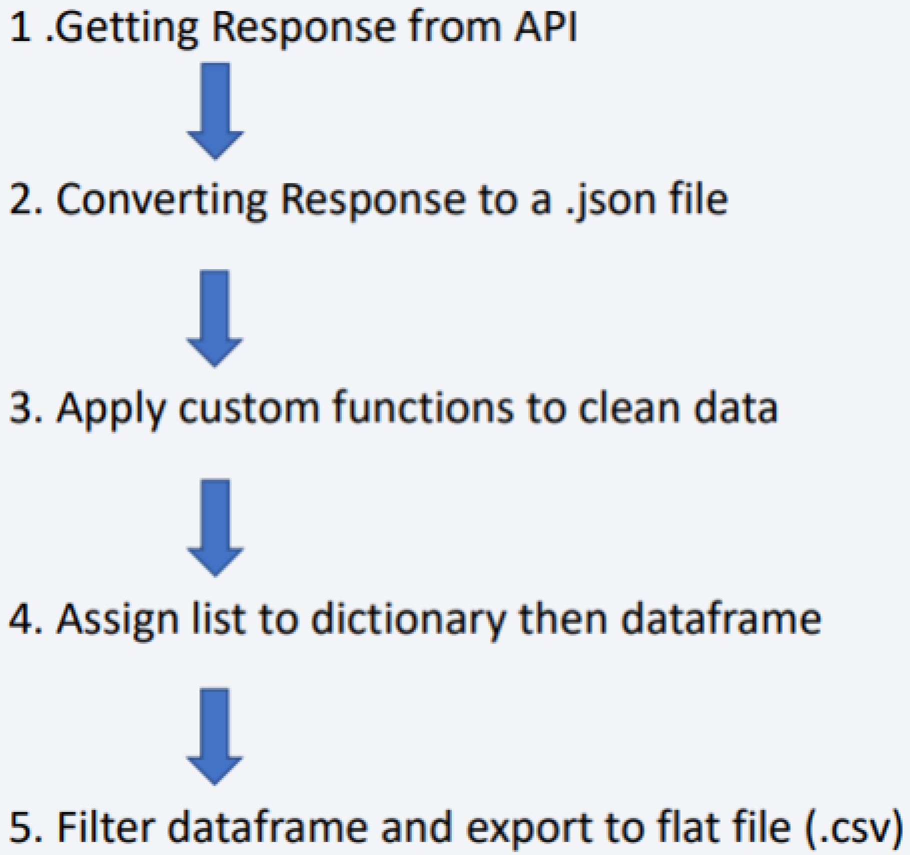
Web Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from HTML table header
- Create a data frame by parsing the launch HTML tables


Data collection
Wikipedia Page
SpaceX API
Scraped using BeautifulSoup
Call the SpaceX API to get information on launches
Saved results into a csv file

Data Collection – SpaceX API

- GitHub URL of the completed REST API lab: https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/01%20Cleaning%20the%20Data/jupyter-labs-spacex-data-collection-api.ipynb



1 .Getting Response from API

2. Converting Response to a .json file

3. Apply custom functions to clean data

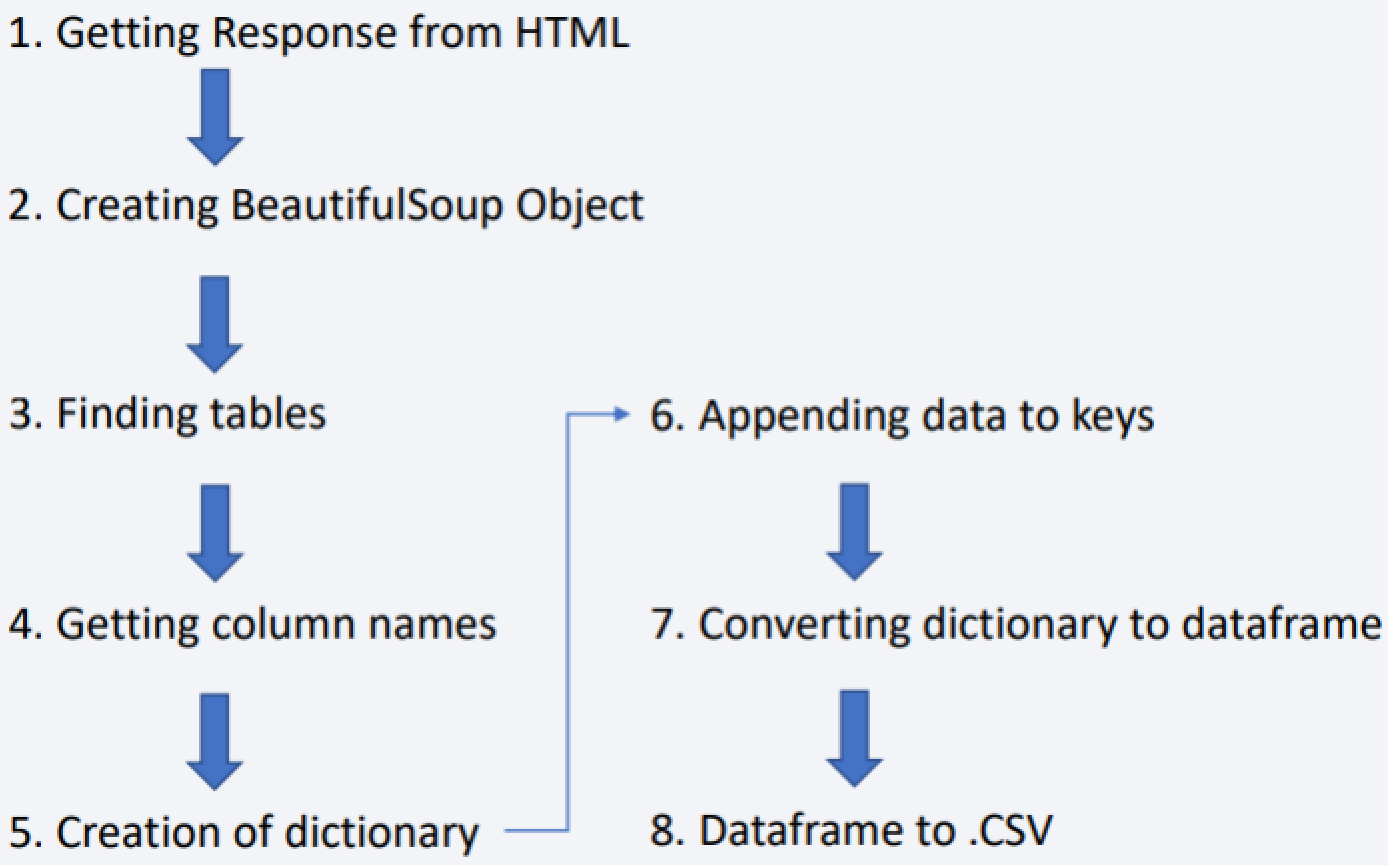
4. Assign list to dictionary then dataframe

5. Filter dataframe and export to flat file (.csv)

Data Collection - Scraping

- GitHub URL of the completed lab:

https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/01%20Cleaning%20the%20Data/jupyter-labs-webscraping%20(1).ipynb

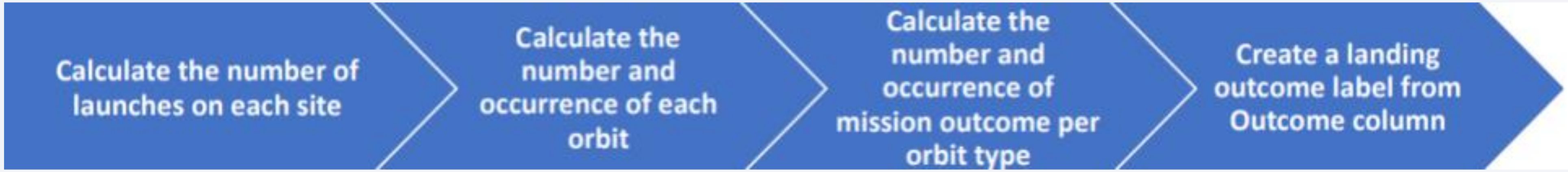

1. Getting Response from HTML
2. Creating BeautifulSoup Object
3. Finding tables
4. Getting column names
5. Creation of dictionary
6. Appending data to keys
7. Converting dictionary to dataframe
8. Dataframe to .CSV

Data Wrangling

- GitHub URL of the completed lab:

https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/02%20Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

- The dataset contains various scenarios where the booster either successfully landed or failed. For instance:
  - **True Ocean** indicates a successful landing in a designated ocean area, while **False Ocean** represents an unsuccessful landing in the ocean.
  - **True RTLS (Return to Launch Site)** signifies a successful landing on a ground pad, whereas **False RTLS** denotes a failed attempt on a ground pad.
  - **True ASDS (Autonomous Spaceport Drone Ship)** reflects a successful landing on a drone ship, while **False ASDS** represents a failed landing on a drone ship.
- To streamline the analysis, these outcomes were transformed into binary training labels:
  - **1** for successful landings.
  - **0** for unsuccessful landings.


Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate the number and occurrence of mission outcome per orbit type
Create a landing outcome label from Outcome column

EDA with Data Visualization

- In the EDA process, we utilized three distinct charts:
  - Scatter charts were utilized to illustrate relationships between various factors such as:
    - Flight Number vs Payload Mass
    - Payload vs Launch Site
  - Bar charts are excellent for displaying categorical data such as:
    - Success rate vs. orbit
  - Line charts are excellent for displaying patterns such as:
    - Each year's average success year trend
- GitHub URL of the completed lab:

https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/04%20EDA%20with%20Visualization%20Lab/edadataviz.ipynb

EDA with SQL

- Various SQL queries were executed to analyze the SpaceX dataset. Examples include:

  1. Retrieve unique launch site names from the dataset.
  2. Display 5 records where launch site names begin with the string "CCA".
  3. Calculate the total payload mass carried by boosters launched for NASA's CRS missions.
  4. Determine the average payload mass carried by the booster version F9 v1.1.
  5. Identify the date of the first successful ground pad landing.
  6. List booster names with successful drone ship landings and payload mass between 4000 and 6000 kg.
  7. Count the total number of successful and failed mission outcomes.
  8. Find booster versions that carried the maximum payload mass.
  9. List failed drone ship landings along with booster versions and launch sites for the year 2015.
  10. Rank landing outcomes (e.g., Failure on drone ship or Success on ground pad) by count, within the date range 2010-06-04 to 2017-03-20, in descending order.

- GitHub URL of the completed lab: https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/03%20Exploratory%20Data%20Analysis%20SQL/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- We utilized Folium to create additional visualizations, enabling deeper exploration of our data. Key Folium map elements used include:
  - MarkerCluster():
    - Used to group markers with the same or nearby coordinates, reducing visual clutter.
    - Green markers represent successful launches (class=1), while red markers indicate unsuccessful launches (class=0).
  - Folium.PolyLine:
    - Creates lines connecting two locations on the map, based on latitude and longitude coordinates.
  - Folium.Circle:
    - Highlights specific areas on the map, such as launch site locations, with circles centered at given coordinates.
  - Folium.Marker:
    - Adds text labels to the map, specifying the names of launch sites, often in conjunction with Folium.Circle elements.

- GitHub URL of the completed lab:

  https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/05%20Interactive%20Visual%20Analytics%20with%20Folium%20lab/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- As part of our data analysis, we developed an interactive dashboard that enables users to explore and manipulate data in real time.
- The dashboard was implemented in Python using the Plotly Dash library and features input components like a dropdown menu and a range slider. These components dynamically interact with the visualizations, which include:
  - Pie Chart:
    - Displays the distribution of launch sites.
    - Highlights the success and failure rates for each site.
  - Scatter Plot:
    - Visualizes the relationship between payload mass, launch success, and the booster version used.
- This interactive design allows for a seamless and engaging exploration of SpaceX launch data.
- GitHub URL of the completed lab:
- https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/06%20Dashboard%20with%20Plotly%20Dash/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- Building the Model
  - Data Preparation: Loaded the dataset using NumPy and Pandas. Transformed the data as needed for analysis. Split the dataset into training and testing subsets. Verified the number of test samples.
  - Algorithm Selection: Chose suitable machine learning algorithms for classification tasks. Configured hyperparameters and integrated algorithms with GridSearchCV for optimization. Trained the models using the prepared datasets.
- Evaluating the Model
  - Measured accuracy for each algorithm.
  - Tuned hyperparameters to achieve optimal performance.
  - Plotted Confusion Matrices to visualize prediction results.
- Improving the Model
  - Applied Feature Engineering to enhance input data quality.
  - Performed Algorithm Tuning to refine model performance.
- Finding the Best Performing Model
  - The model with the highest accuracy score was selected as the best-performing classifier.
  - A summary of algorithm performance and scores is available in the notebook's results dictionary.
- GitHub URL of the completed lab:
- https://github.com/LuisRequejoM/IBM-DATA-SCIENCE/blob/main/Applied%20Data%20Science%20Capstone/07%20Machine%20Learning%20Predictive%20Analysis/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

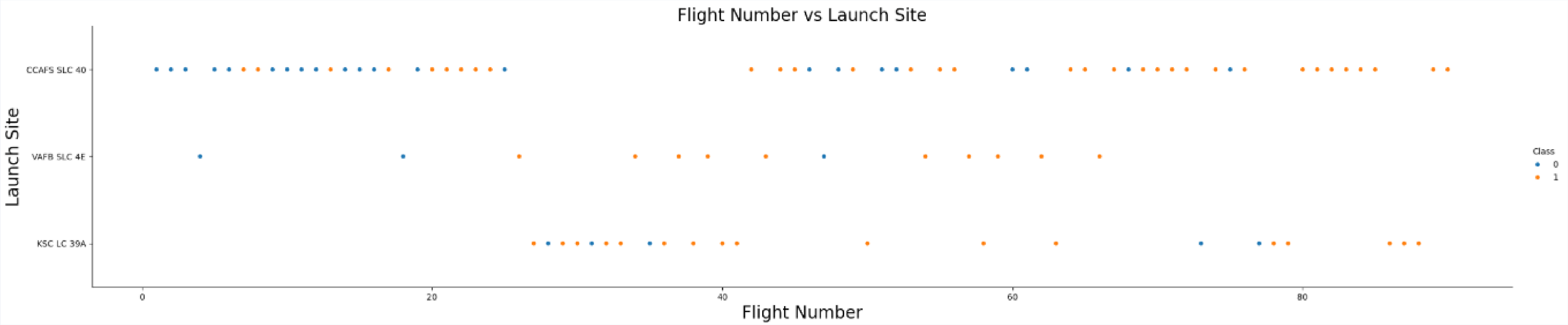
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

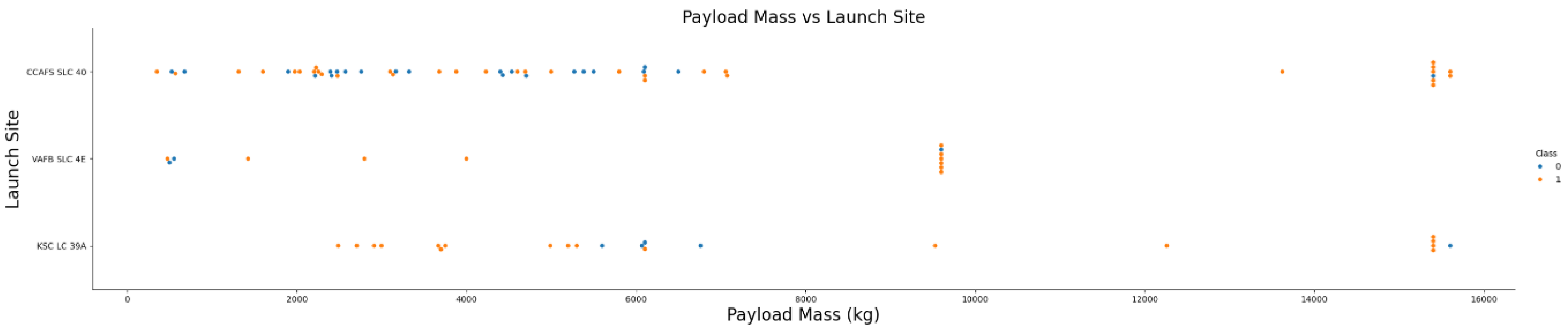
Flight Number vs. Launch Site

- The graph illustrates the relationship between launch sites and success rates across the number of flights. It is evident that the instances of success, represented by class 1, increase proportionally with the number of flights. This trend may indicate improvements in launch site performance or advancements in other contributing factors.



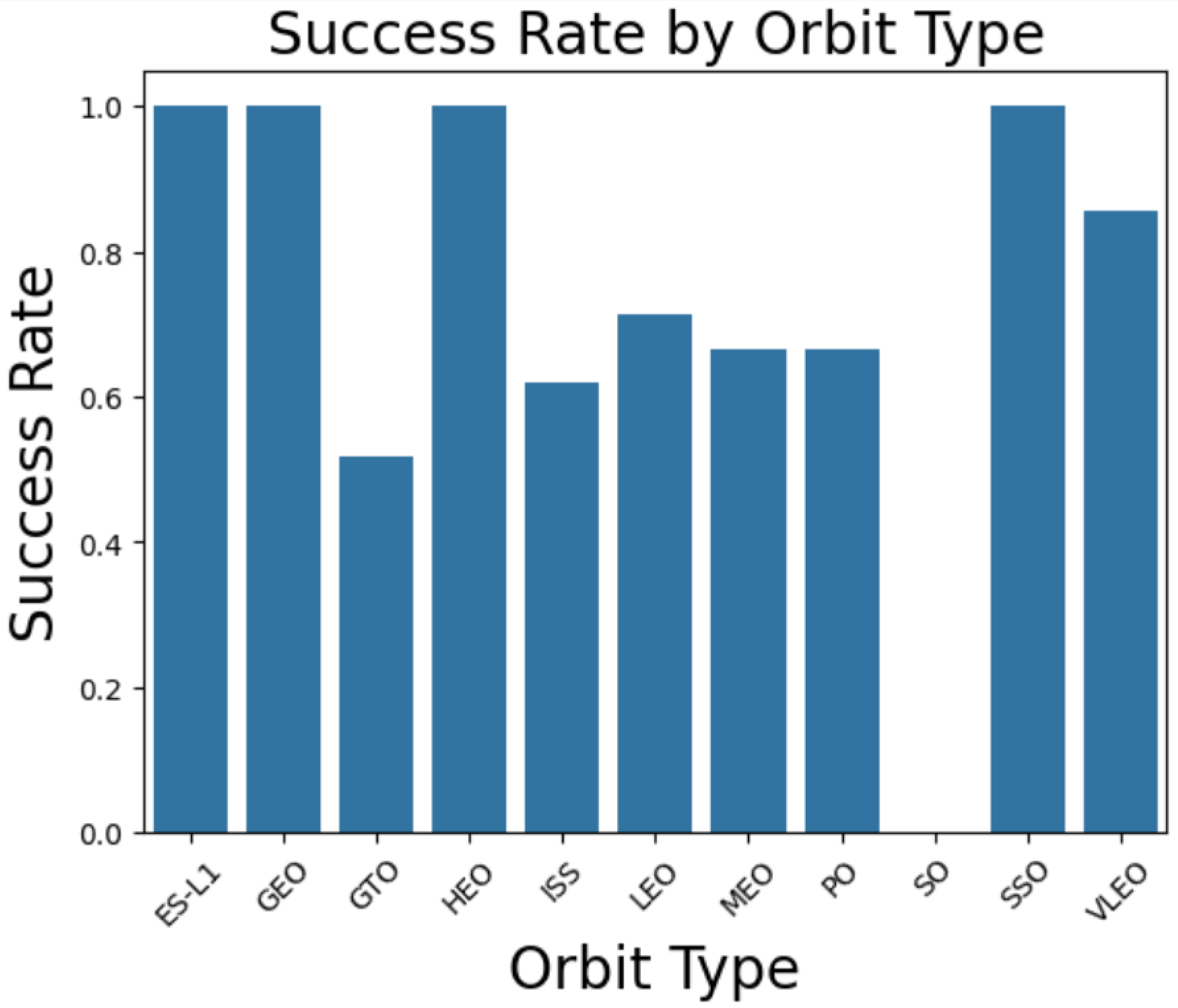
Payload vs. Launch Site

- The distribution of successful and unsuccessful landings at CCAFS SLC 40 is evenly spread across payload masses ranging from 0 to 8000 kg but shows greater success with larger payloads. At the VAFB SLC 4E launch site, most landings were successful regardless of payload mass. Similarly, KSC LC 39A recorded a high rate of successful landings. However, an anomaly was observed near the 6000 kg payload range, where landings were generally unsuccessful.



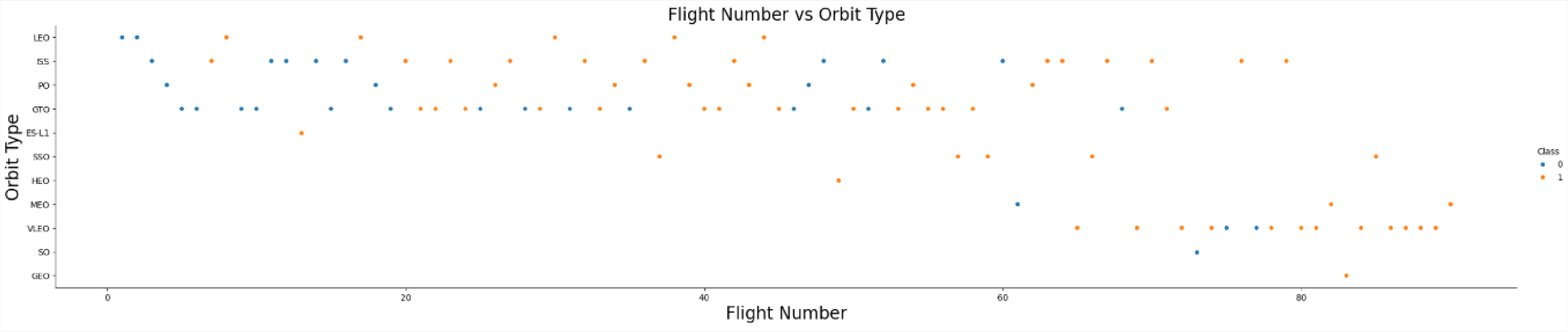
Success Rate vs. Orbit Type

- This graph highlights the most and least successful orbits. Orbits such as ES-L1, GEO, HEO, and SSO exhibited the highest success rates, while GTO had the lowest success rate. The SO orbit showed no significant outcomes.



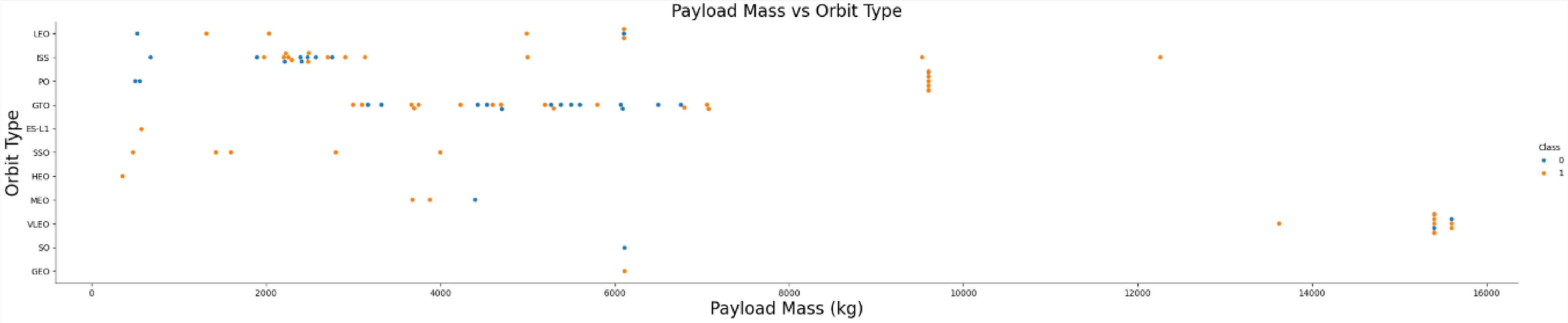
Flight Number vs. Orbit Type

- It is evident that in the LEO orbit, the success rate seems to increase with the number of flights. However, in the GTO orbit, no clear relationship between the number of flights and success rate is observed.



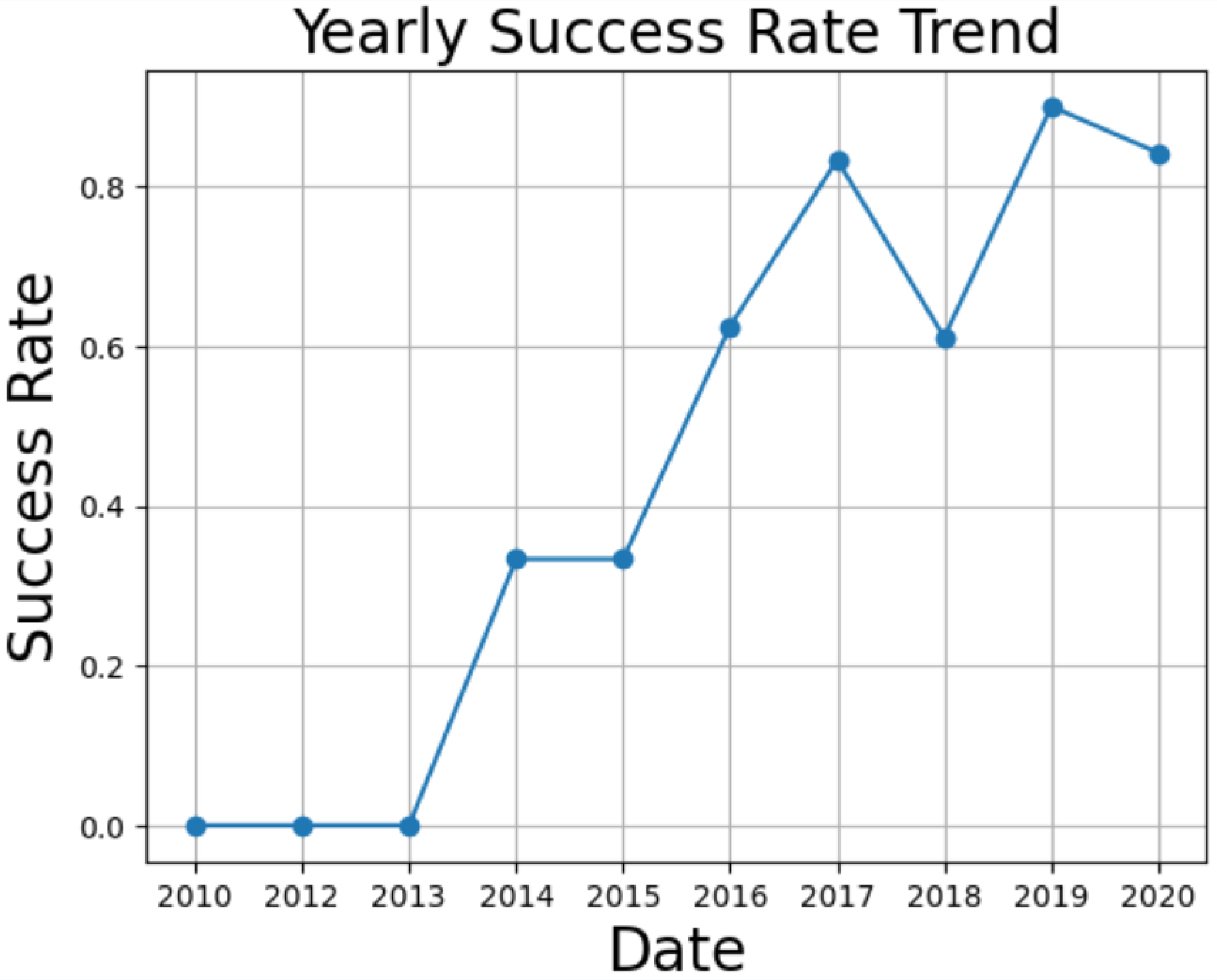
Payload vs. Orbit Type

- It is evident that heavier payloads negatively affect GTO orbits but have a positive impact on PO and LEO orbits. However, this pattern does not apply to ESL1 and SSO orbits.



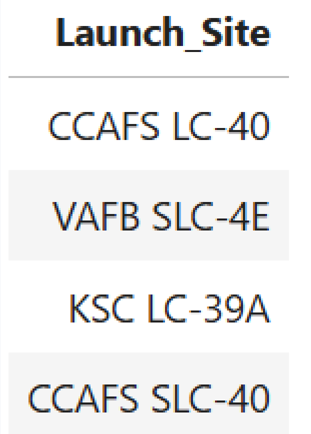
Launch Success Yearly Trend

- The success rate has shown consistent improvement since 2013 and is expected to continue rising until 2020. However, there was a noticeable decline between 2017 and 2018, followed by a recovery from 2018 to 2019.



All Launch Site Names

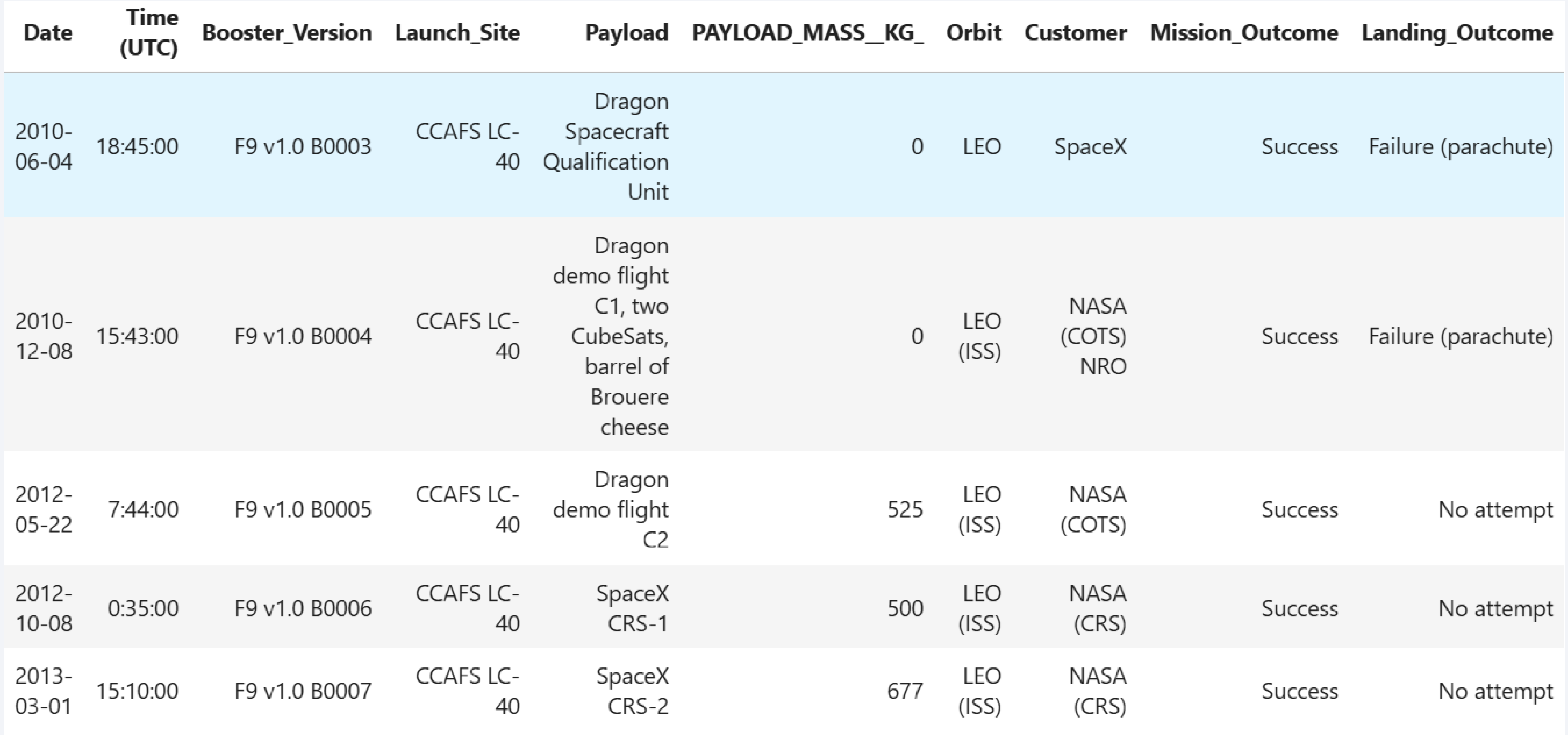
- Find the names of the unique launch sites



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

Launch Site Names Begin with 'CCA'

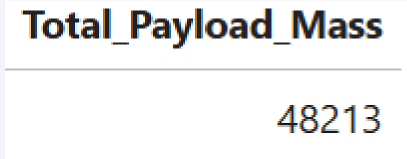
- Find 5 records where launch sites begin with `CCA`



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

Total Payload Mass

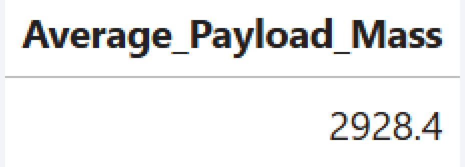
- Calculate the total payload carried by boosters from NASA



Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

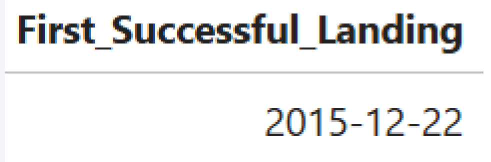
- Calculate the average payload mass carried by booster version F9 v1.1



Average_Payload_Mass
2928.4

First Successful Ground Landing Date

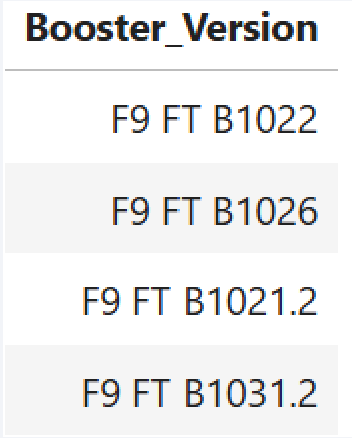
- Find the dates of the first successful landing outcome on ground pad



First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

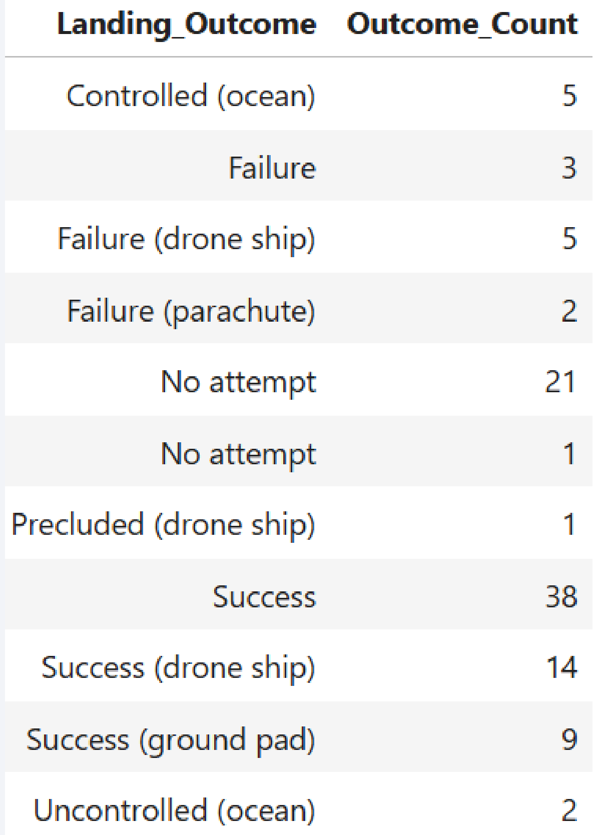
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



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

Total Number of Successful and Failure Mission Outcomes

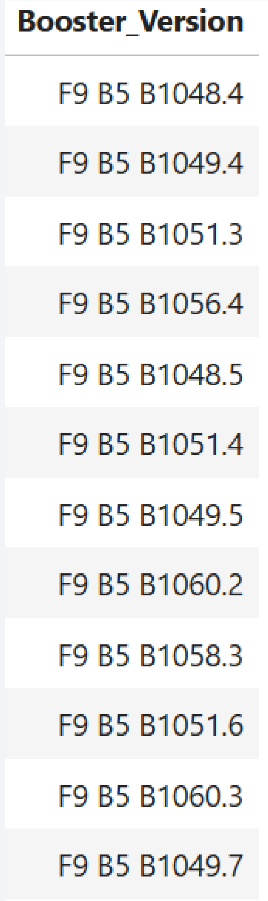
- Calculate the total number of successful and failure mission outcomes



Landing_Outcome	Outcome_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

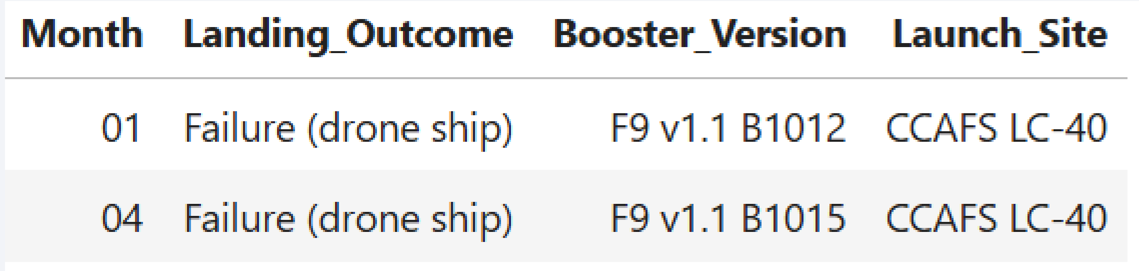
- List the names of the booster which have carried the maximum payload mass



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

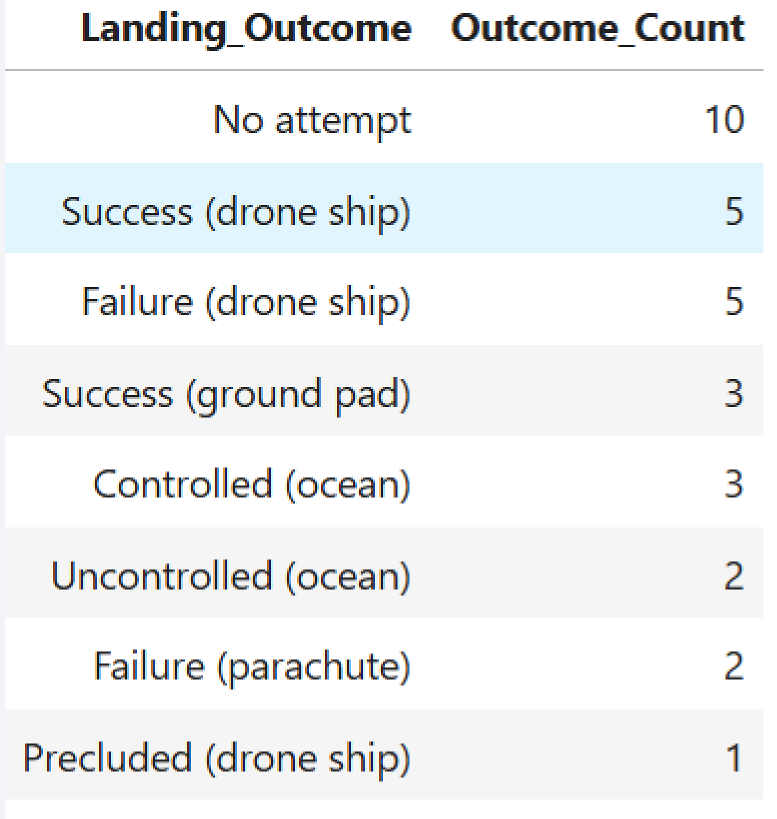
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

- Rank 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



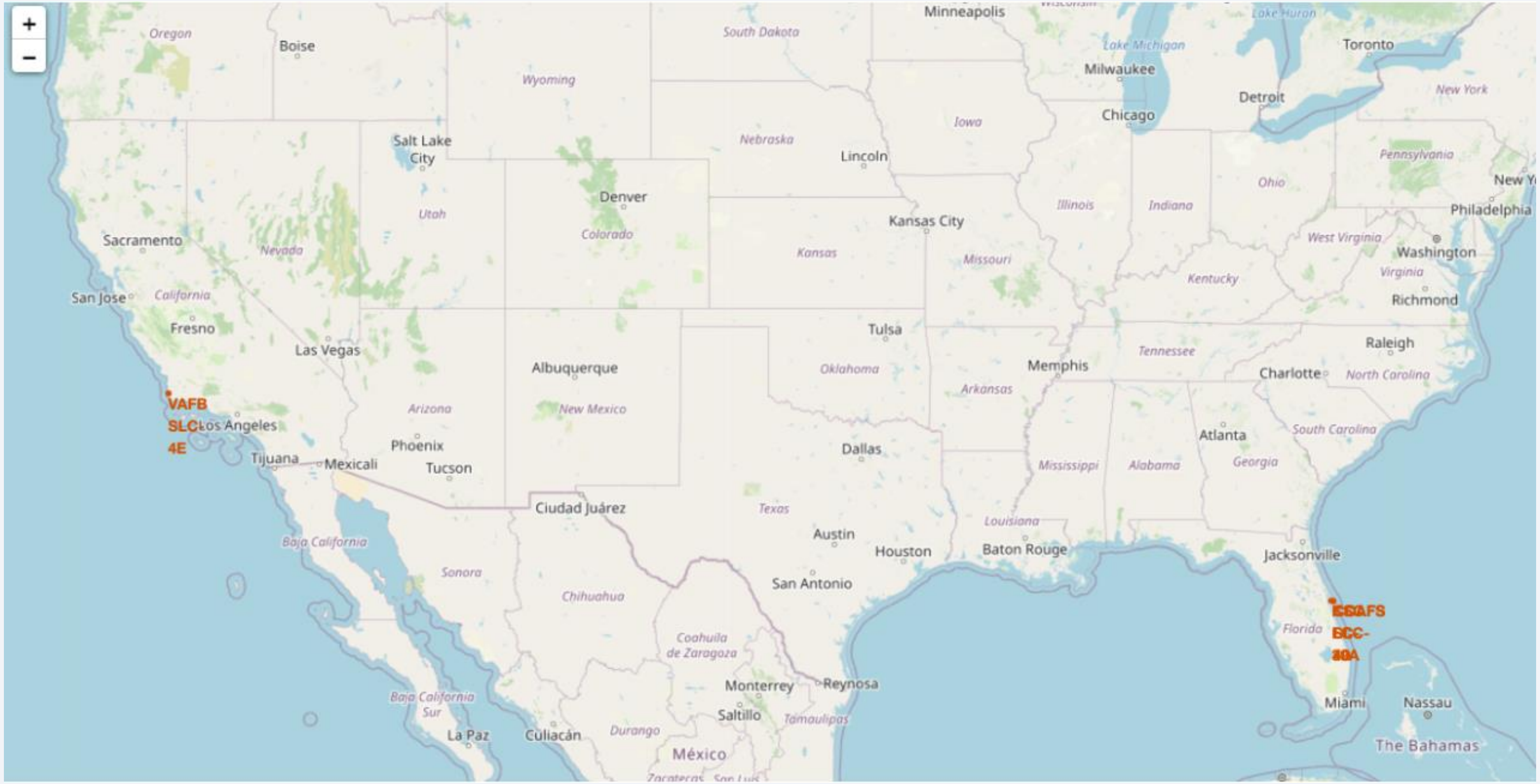
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

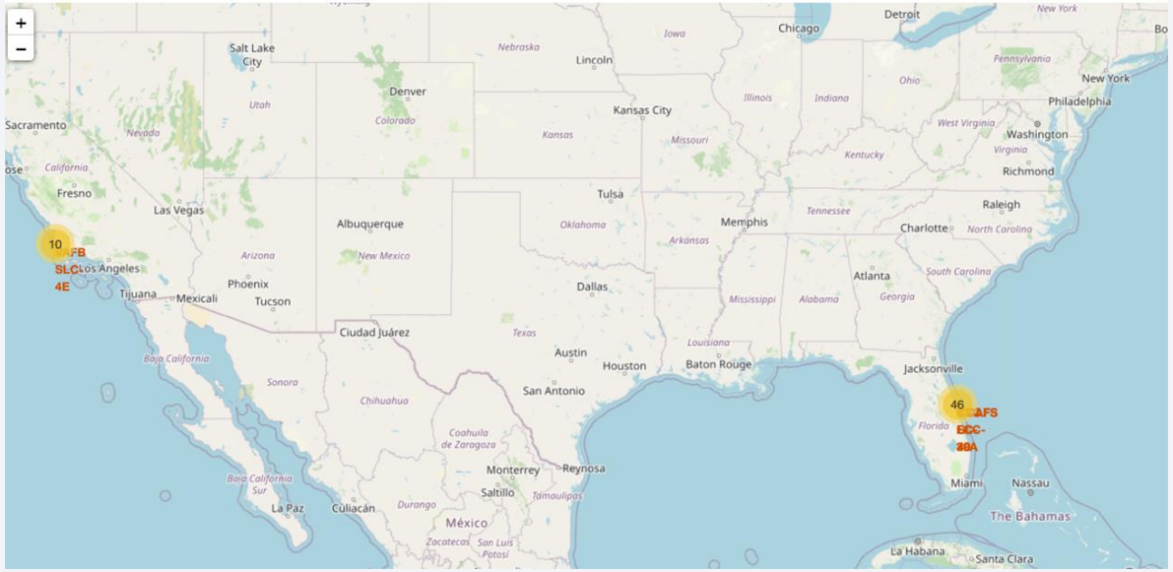
Location of launch sites

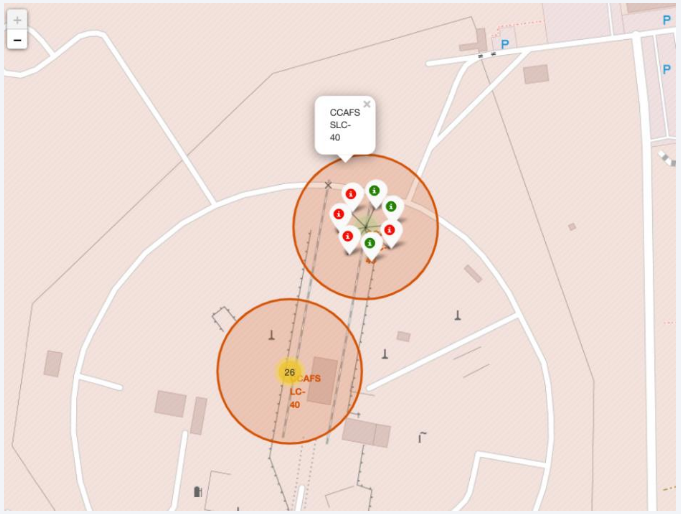
- There are 4 launch sites in total, one in California, USA, three in Florida, USA



Launch clusters

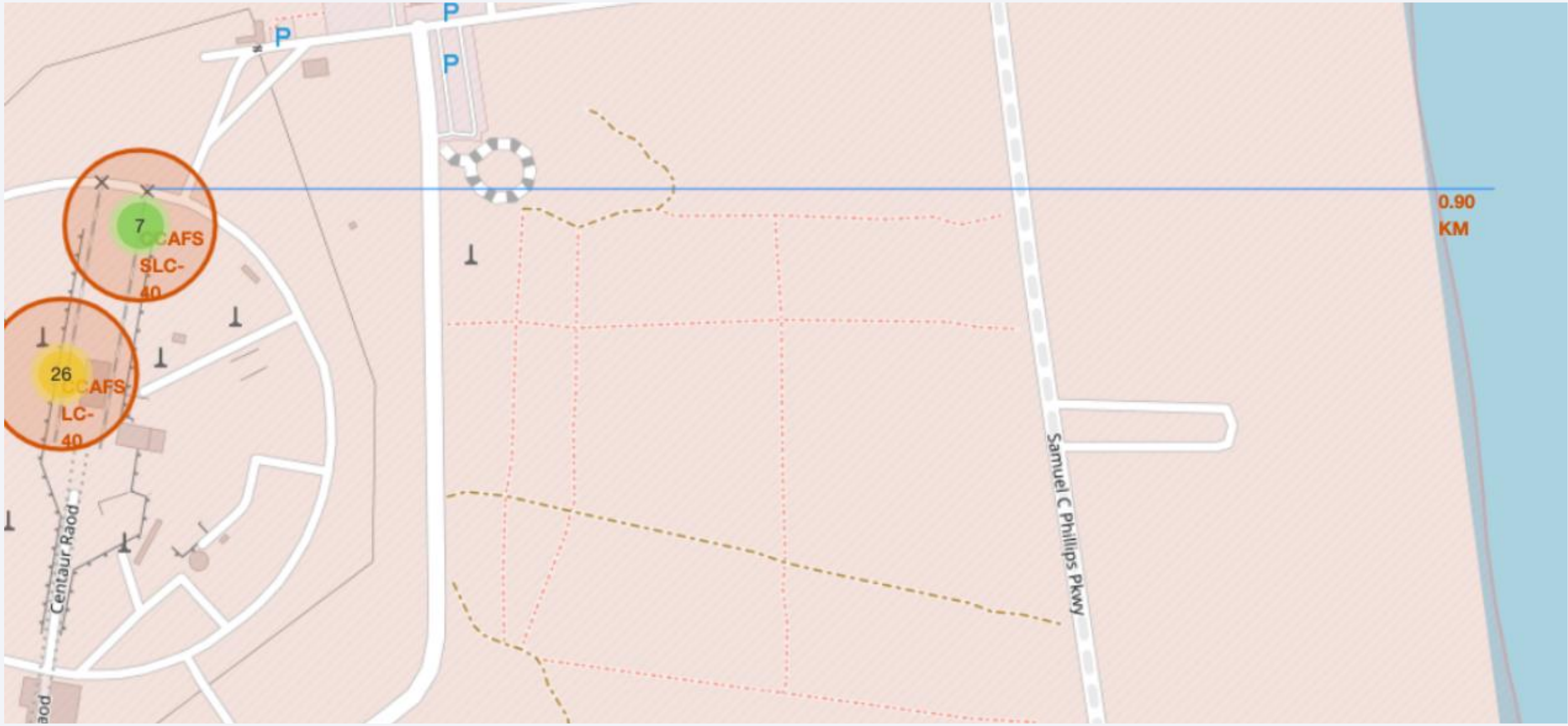
- On the right, there are two screenshots. The upper picture show the total numbers of success launches on east and west costs.
- The lower picture show a closer View of number of success at specific launch sites. Among all the sites, The green circle indicates KSC LC-39A has the highest number of success launches.





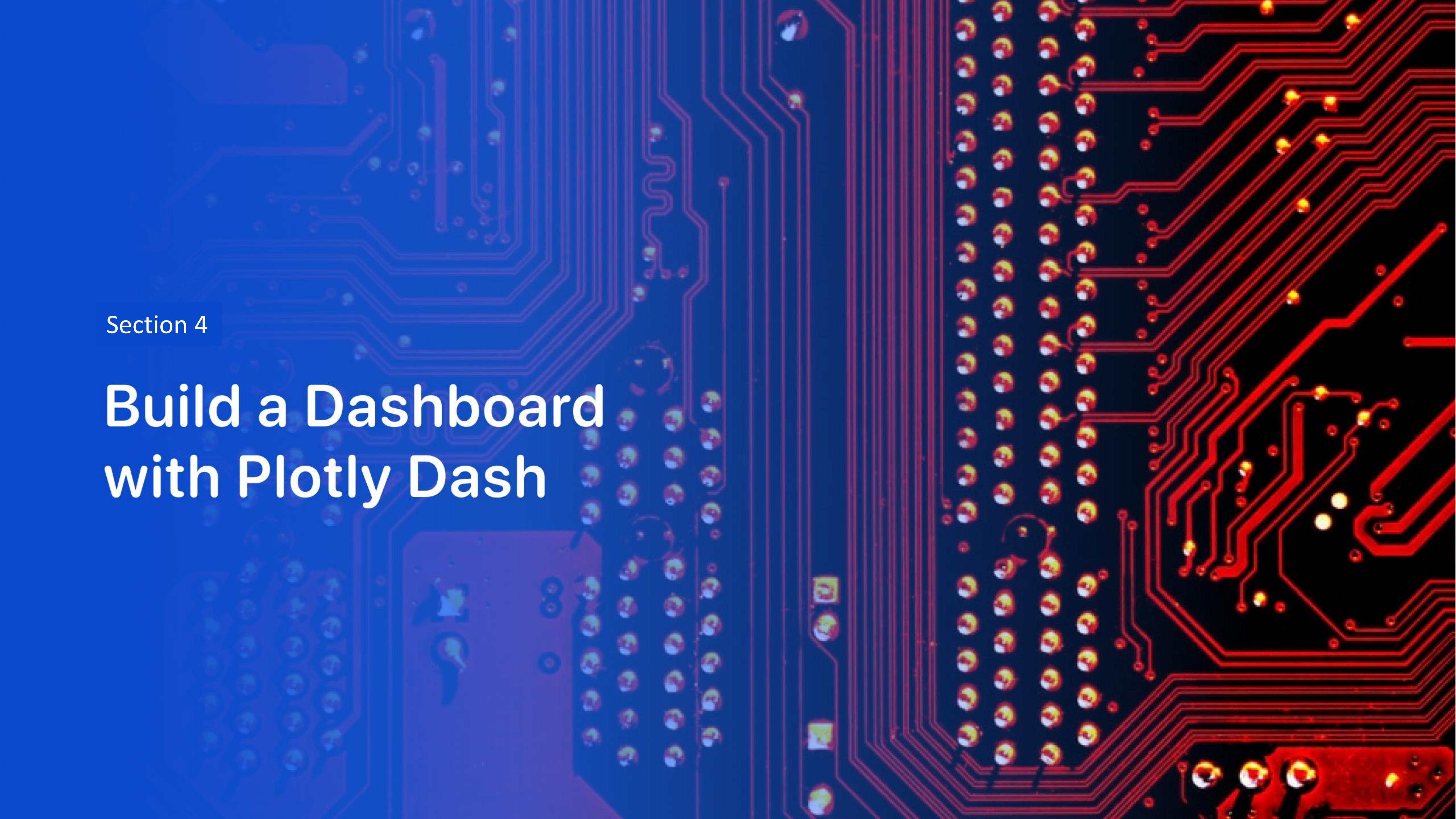
Distance between Sites and Closest Locations

- The picture on the right side depicts the distance between the launch Site CCAFS SLC-40 and the coastline.



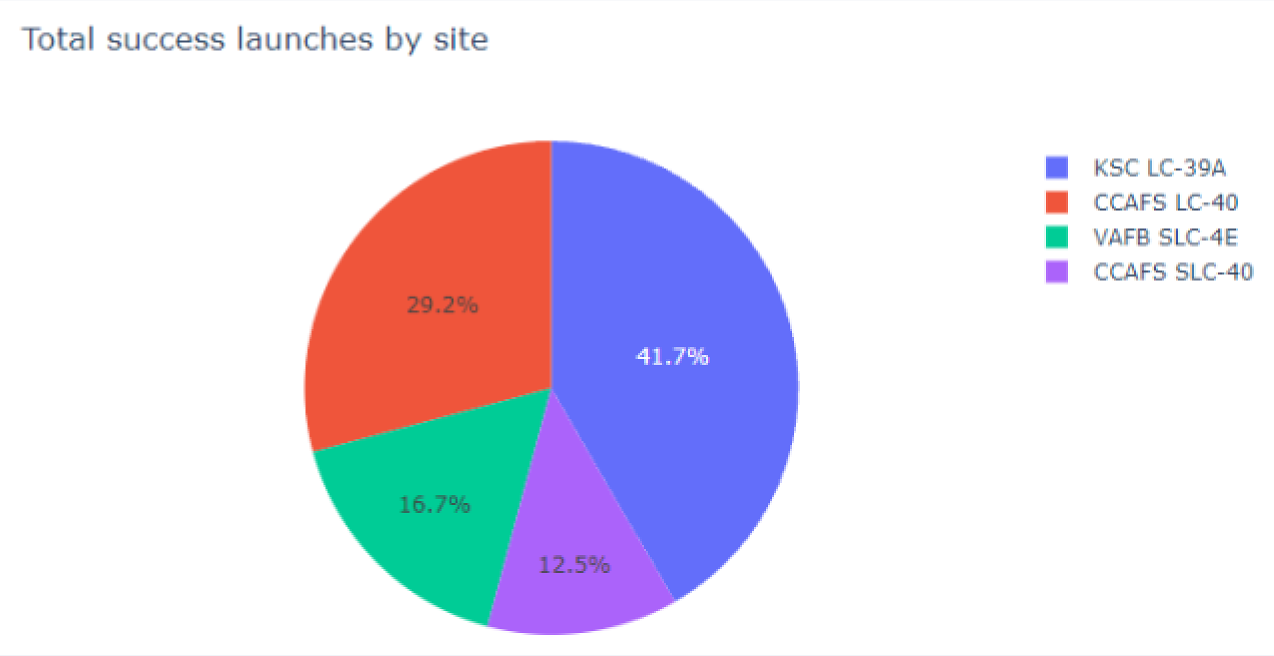
Section 4

Build a Dashboard with Plotly Dash



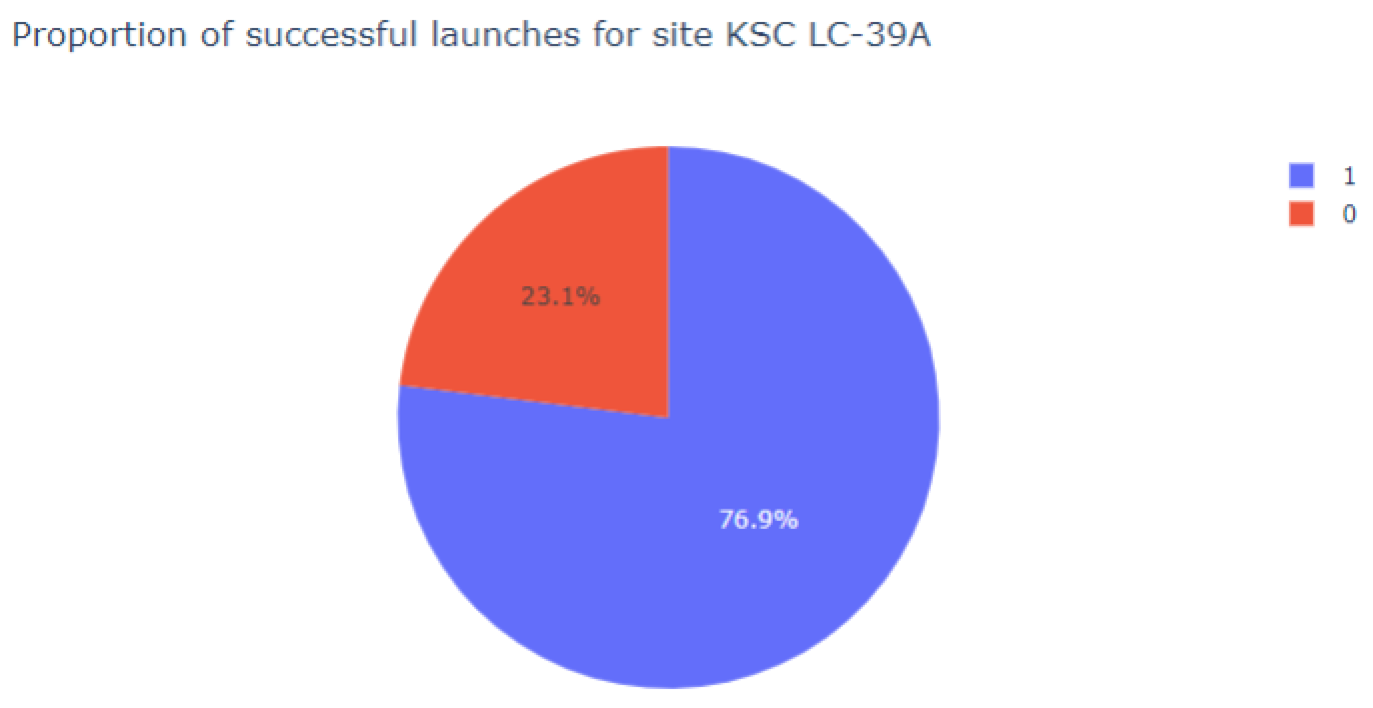
Total success launches by Site

- Almost half of all successful launches were done in KSC LC-39A



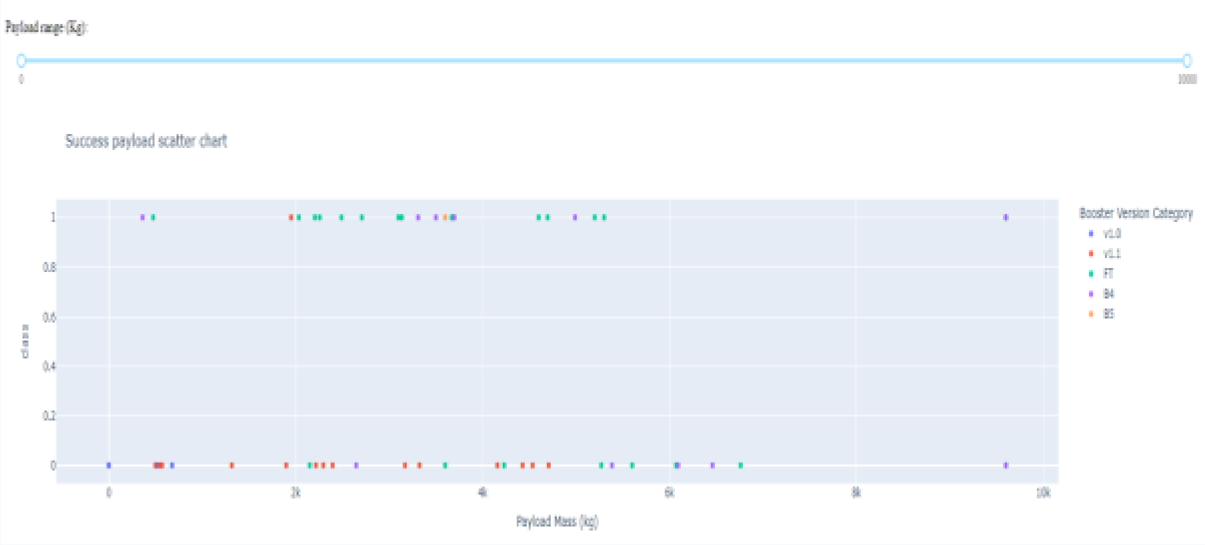
Proportion of successful launches for KSC LC-39A

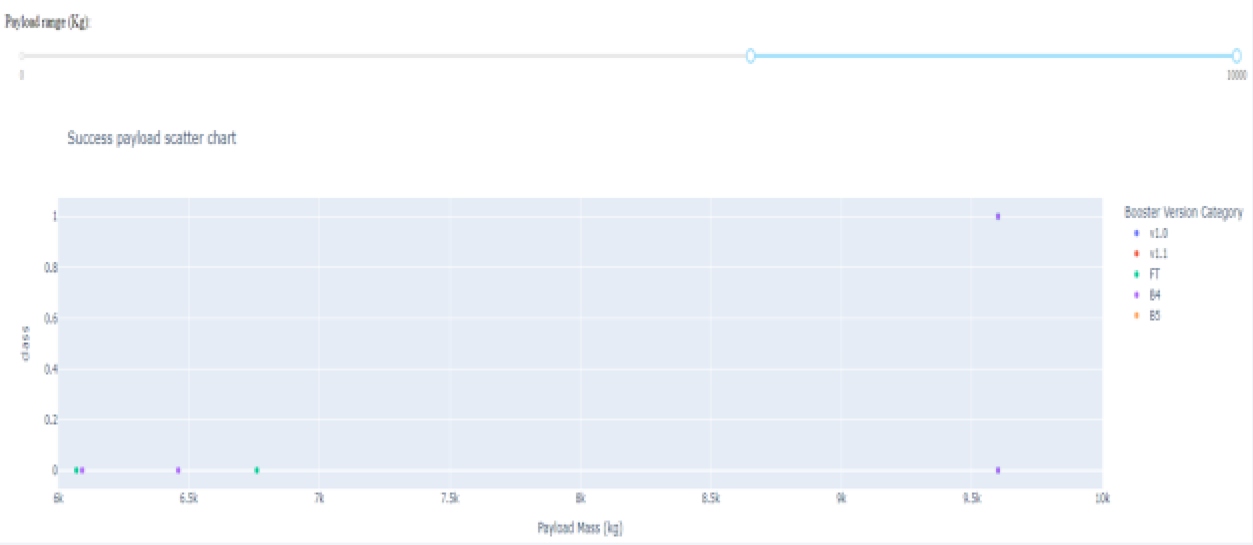
- ~3/4 of launches for KSC LC-39A were successful



Payload mass and launch outcome

- Almost all launches with a payload mass above 6000kg failed
- Most of the successful launches were in the FT booster version category



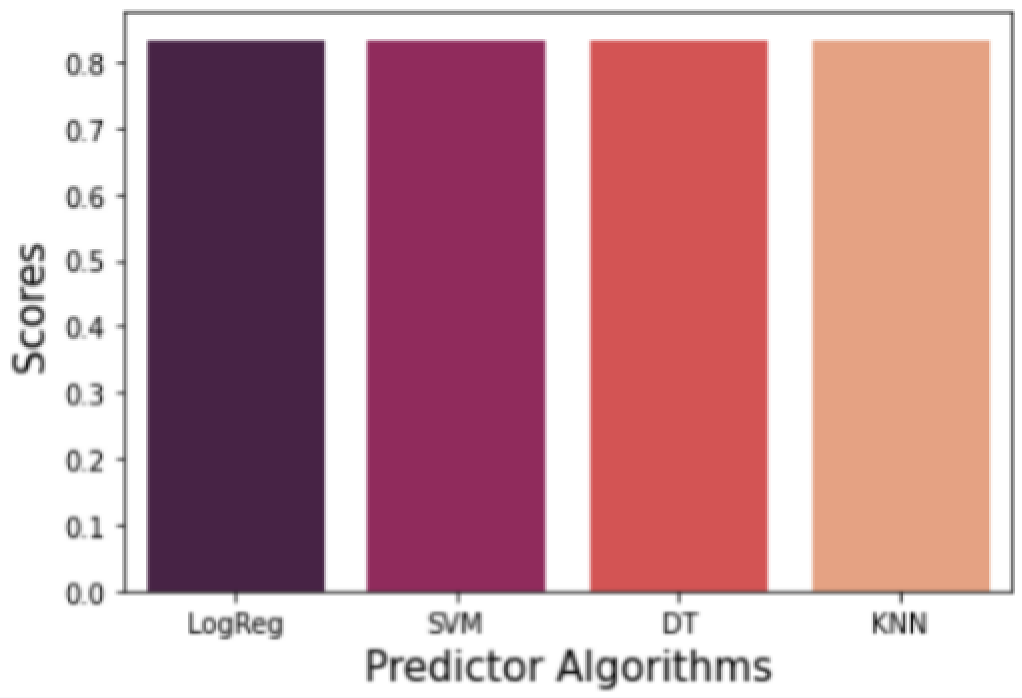


Section 5

Predictive Analysis (Classification)

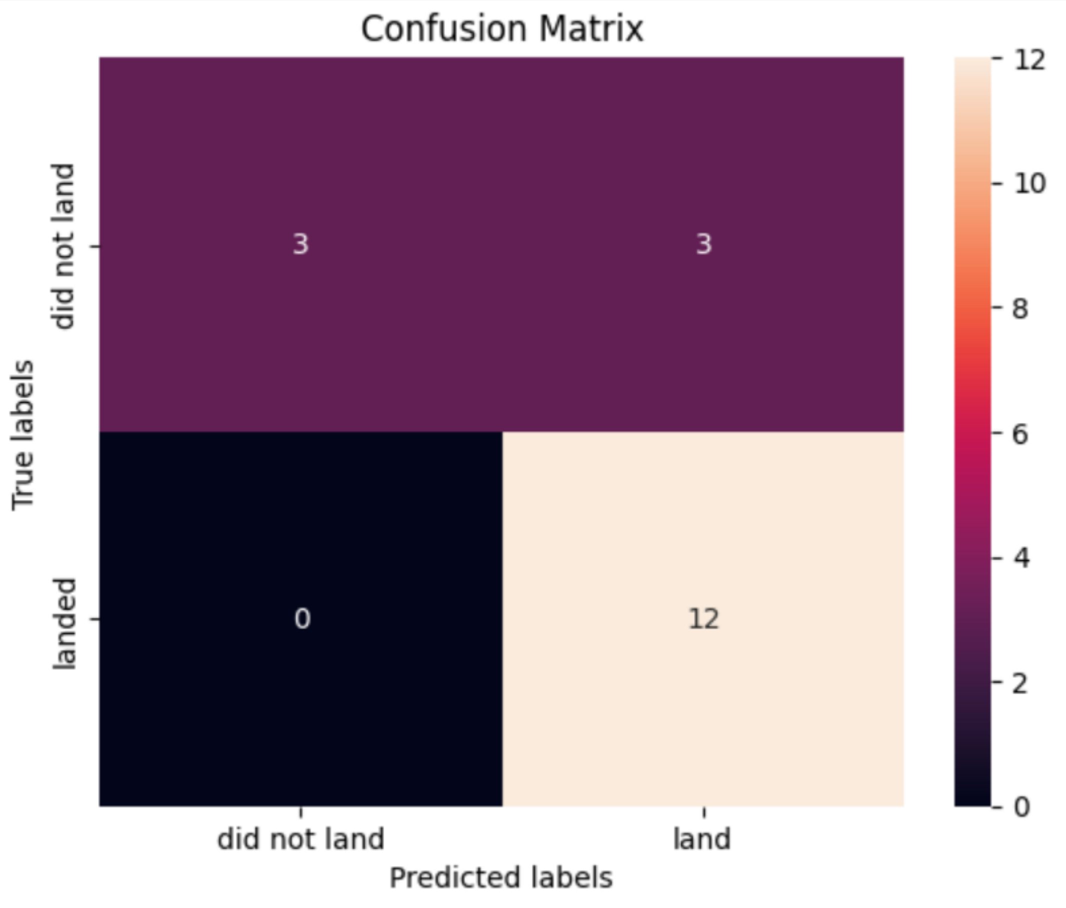
Classification Accuracy

- To validate a machine learning algorithm predicting the success or failure of a launch, the dataset is divided into training and testing sets. The models are then trained, and the hyperparameters are optimized using the GridSearchCV function. The best-performing models were identified using Logistic Regression, SVM, Decision Trees, and the KNN method. The bar graph on the right displays the performance scores for each algorithm. In this case, all models achieved the same score.



Confusion Matrix

- Based on the confusion matrix, out of the eighteen actual launch attempts, six were failures and twelve were successful. The algorithms predicted three failures and fifteen successes. While there was a small discrepancy, the prediction accuracy for all models was strong.



Conclusions

- To achieve the main objective of the project, it is essential to follow the stages in the correct order, including ETL, EDA, modeling, and implementation. Acquiring a dataset can be done through various methods, such as web scraping, downloading data from the internet, querying a database, or even creating one from scratch.
- To better understand the data, proper data wrangling and exploratory data analysis techniques must be applied. Visualizations, such as charts, maps, or interactive dashboards, help create a more comprehensive representation of the data.
- In this project, the four algorithms employed to predict the successful launch of Falcon 9 all demonstrated strong performance.

Thank you!