IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Durotoye Abdullahi Ololade

May 16, 2025

Outline

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

Executive Summary

- This project explores the commercial spaceflight industry, focusing on SpaceX's cost-effective Falcon 9 launches enabled by reusable first-stage rockets. Acting as data scientists for a competitor company, "Space Y," our goal was to predict whether a Falcon 9's first stage would successfully land and key to estimating launch costs.
- We collected data using the SpaceX REST API and web scraping, then cleaned and structured it using Pandas. Key features such as launch site, payload mass, and booster configuration were analyzed. Exploratory Data Analysis revealed that launch site and payload mass strongly influence landing success.
- An interactive dashboard was built using Folium and Plotly Dash, enabling real-time data exploration by site, payload, and success outcome. Finally, we built and tested machine learning models Logistic Regression, SVM, KNN, and Decision Trees, with all achieving 83% accuracy.
- This project demonstrates how data-driven insights and predictive models can support strategic decisions in the space industry, particularly around cost-effective mission planning.

Introduction

- The era of commercial spaceflight has arrived, transforming space access from a government-led endeavor into a competitive private industry. Companies such as Virgin Galactic, Rocket Lab, Blue Origin, and most notably, SpaceX, are pioneering this transformation. Among them, SpaceX stands out due to its technological achievements and cost-efficiency primarily due to its reusable Falcon 9 rocket stages. Recognizing this advantage, our project simulates the role of a data scientist at "Space Y," a new space launch company aiming to compete with SpaceX.
- The goal: To analyze public data and predict whether a Falcon 9's first stage will successfully land, impacting launch cost projections.

Section 1

Methodology

Created by Durotoye Abdullahi Ololade

Methodology

Executive Summary

- Data Collection Methodology:
  - Data was collected from the official SpaceX REST API (api.spacexdata.com/v4/launches/past) and supplemented with web scraping from Wikipedia.
- Perform Data Wrangling
  - Key attributes were extracted: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, and features like Grid Fins, Reused, Legs, etc. The dataset was filtered to include only Falcon 9 launches, missing values were handled, and categorical data was converted via one-hot encoding.
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform Interactive Visual Analytics using Folium and Plotly Dash
- Perform Predictive Analysis using Classification Models
  - A machine learning pipeline was built to predict landing success using preprocessing, model training, hyperparameter tuning, and evaluation with accuracy and confusion matrices.

Data Collection

We collected Falcon 9 launch data from two main sources:

- SpaceX REST API
- Web Scraping from Wikipedia

These sources provided structured and semi-structured data including launch details, rocket specs, and landing outcomes.

Data Collection – SpaceX API

API Endpoint

- AccessBase URL: https://api.spacexdata.com/v4/launches/past
- Accessed using Python requests library

Data Format

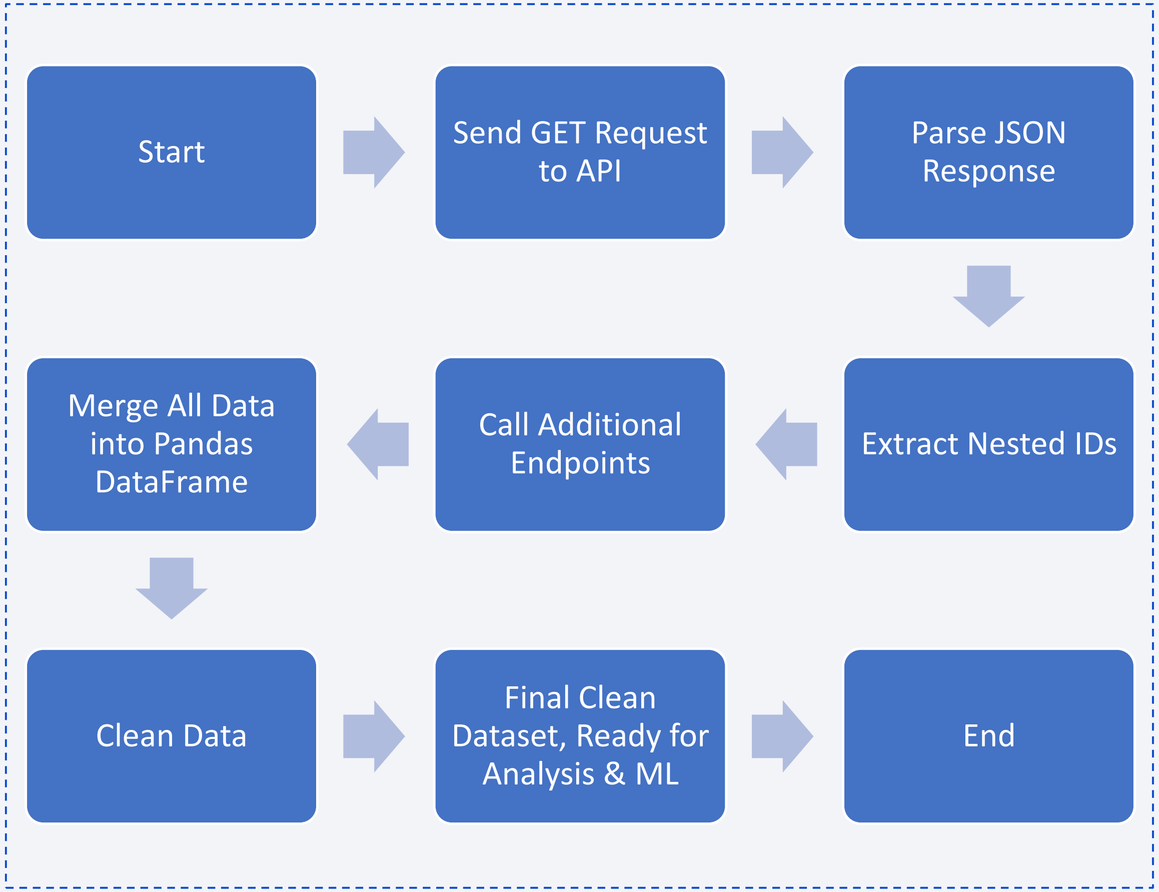
- Response: List of JSON objects, each representing a launch event. Used json_normalize() in pandas to flatten nested data

Data Cleaning & Transformation

- Removed Falcon 1 data
- Replaced null Payload Mass with mean value
- Prepared final dataset as a Pandas DataFrame for analysis

GitHub Repository Link:

- Data Collection API Notebook: Link


Start
Send GET Request to API
Parse JSON Response
Extract Nested IDs
Call Additional Endpoints
Merge All Data into Pandas DataFrame
Clean Data
Final Clean Dataset, Ready for Analysis & ML
End

Created by Durotoye Abdullahi Ololade

Data Collection - Scraping

Target source: Wikipedia Falcon 9 Launches Page

Tool used: BeautifulSoup for parsing HTML

Library used: Requests to get page content

Data and Processing:

- HTML tables of Falcon 9 launches, Convert HTML tables to Pandas DataFrame

Cleanup:

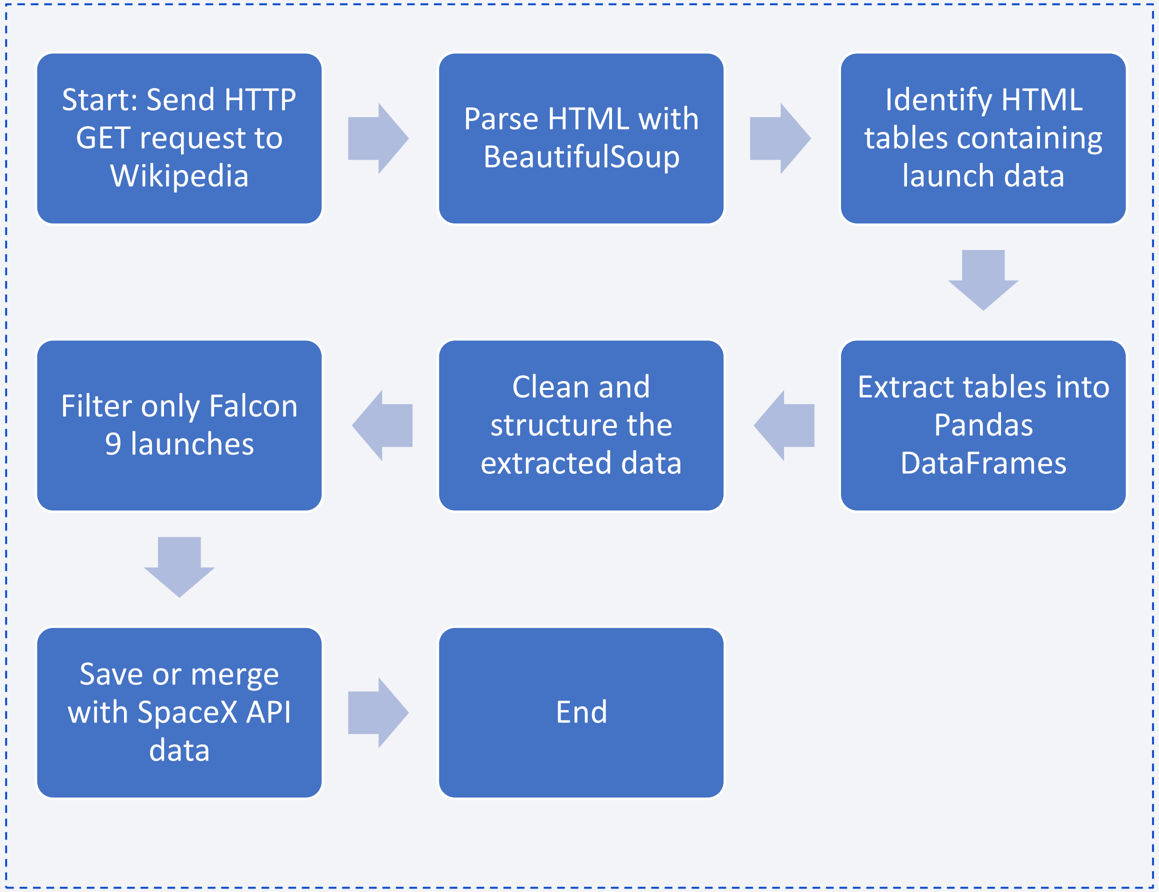
- Remove irrelevant rows/columns, handle missing values

Output:

- Structured dataset ready for merging with API data

GitHub Repository Link:

- Web Scraping Notebook: Link



Start: Send HTTP GET request to Wikipedia

Parse HTML with BeautifulSoup

Identify HTML tables containing launch data

Extract tables into Pandas DataFrames

Clean and structure the extracted data

Filter only Falcon 9 launches

Save or merge with SpaceX API data

End

Created by Durotoye Abdullahi Ololade

Data Wrangling

Class simplification:

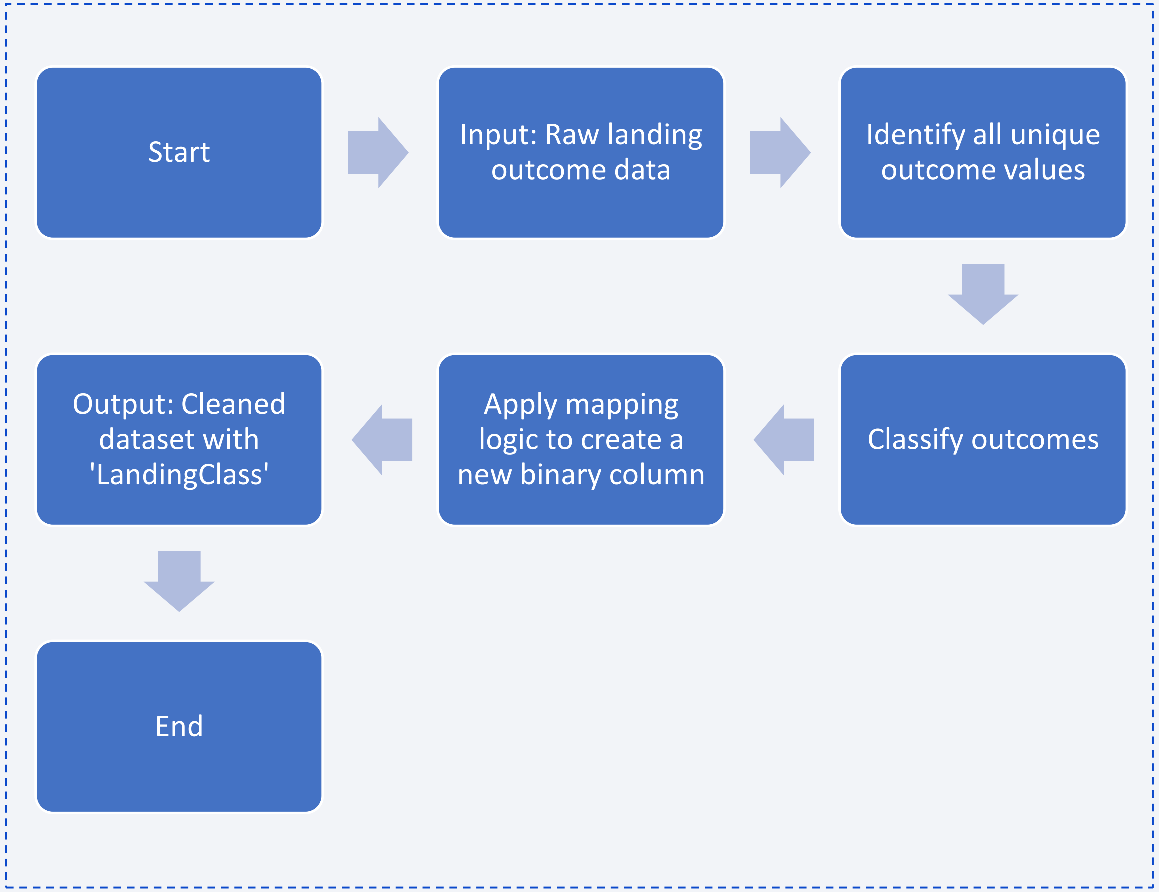
- Reduced multiple nuanced outcomes (e.g., "True ASDS", "False Ocean") into a single classification target

Label transformation: Converted textual landing outcomes to binary labels

- If outcome contains "True“, Label = 1 (Successful landing)
- if outcome contains "False“, Label = 0 (Unsuccessful landing)

GitHub Repository Link:

- Data Wrangling Notebook: [Link]



Start

Input: Raw landing outcome data

Identify all unique outcome values

Classify outcomes

Apply mapping logic to create a new binary column

Output: Cleaned dataset with 'LandingClass'

End

Created by Durotoye Abdullahi Ololade

EDA with Data Visualization

We plotted various charts to analyze the impact of different features on launch success:

- Scatter plots to visualize relationships between FlightNumber, PayloadMass, LaunchSite, and launch outcome (Class).
- Bar charts to compare success rates across different Orbit types.
- Line chart to observe yearly trends in launch success.
- OneHotEncoding was used to prepare categorical features for machine learning

GitHub Repository Link:

- EDA + Visualizations Notebook: Link

EDA with SQL

SQL Queries Summary

- Loaded SQL extension and created a working table from the original SpaceX dataset with non-null dates.
- Retrieved unique launch site names.
- Selected 5 records with launch sites starting with 'CCA'.
- Calculated total payload mass for boosters launched by NASA (CRS).
- Found average payload mass for booster version F9 v1.1.
- Identified the date of the first successful landing on a ground pad.
- Listed boosters that succeeded on drone ships with payload mass between 4000 and 6000.
- Counted total successful and failed mission outcomes.
- Queried booster versions that carried the maximum payload mass using subqueries.
- Displayed month names, failure drone ship landings, booster versions, and launch sites for 2015.
- Ranked landing outcomes between June 2010 and March 2017 by their count, in descending order.

GitHub Repository Link:

- SQL EDA Notebook: Link

Build an Interactive Map with Folium

Folium Map Summary

- Markers:
  Added markers at launch site coordinates to indicate the location of each SpaceX launch site. Each marker includes a popup with the site name for easy identification.
- Circles:
  Placed circle objects around each launch site to highlight their areas visually. The radius of the circles helps convey proximity and emphasize launch zones.
- Lines (Polylines):
  Drew lines from launch sites to their respective payload landing points (if available), helping to visualize flight paths and launch dynamics.
- Custom Icons:
  Used icons for specific launch sites to distinguish between them on the map, enhancing map readability.

Why These Objects Were Added

- To visually represent spatial information of the SpaceX launch and landing sites.
- To help users explore the geographic distribution and relationships between launch points and landing outcomes.
- To create an interactive, informative visualization that supports data interpretation beyond static charts or tables.

GitHub Repository Link:

- Folium Notebook: Link

Build a Dashboard with Plotly Dash

Plotly Dash Summary

Plots/Graphs Added:

- Pie Chart:
  Displays the proportion of successful vs. failed launches. Gives a quick overview of SpaceX mission success rate.
- Scatter Plot:
  Plots payload mass against mission outcome, colored by booster version. Helps identify how payload and booster types relate to launch success.

Interactions Added:

- Dropdown Menu:
  Lets users filter data by launch site. Enables site-specific analysis of launch performance.
- Range Slider:
  Allows users to select a payload mass range. Filters scatter plot to observe the effect of different payload weights on mission outcome.

Why These Were Added

- To create an interactive, user-friendly dashboard for data exploration.
- To allow users to explore patterns in launch success based on payload, site, and booster type.
- To support visual data analysis and enhance understanding of the SpaceX launch dataset.

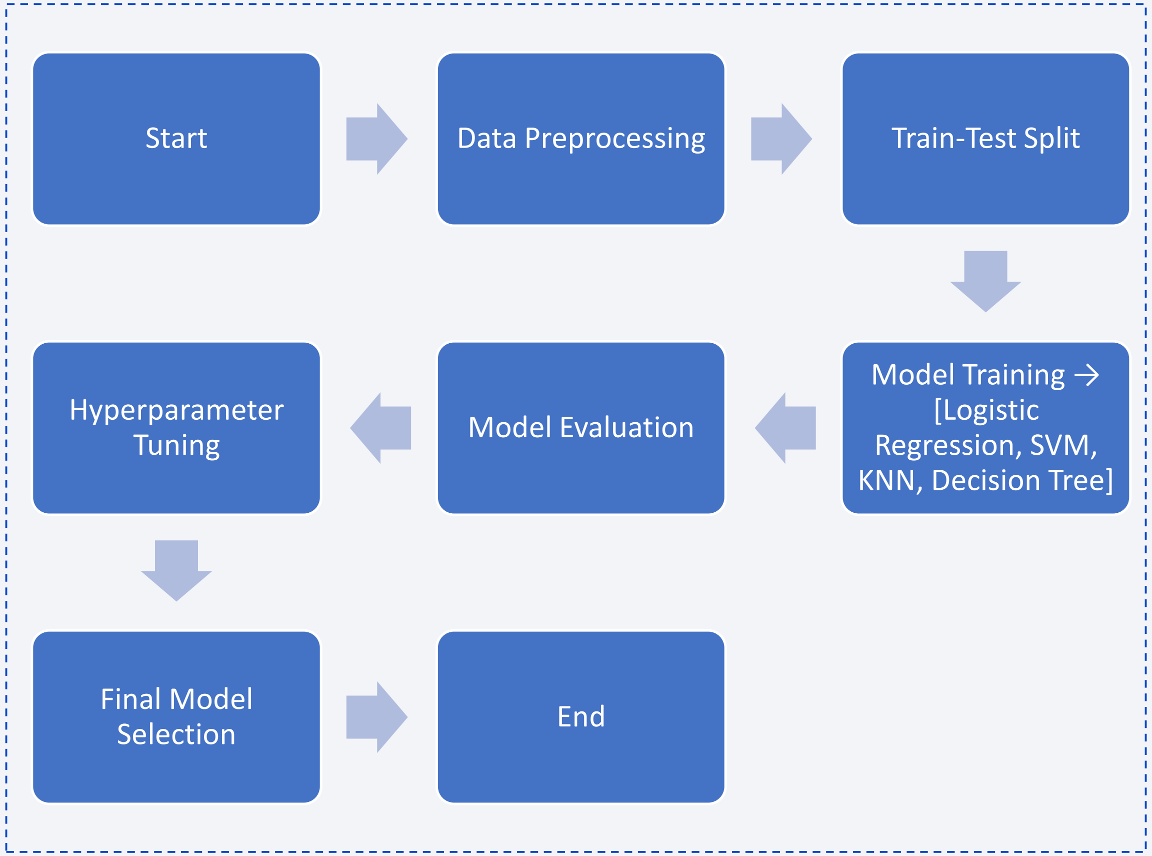
GitHub Repository Link:

- Dashboard Python File: Link

Predictive Analysis (Classification)

Predictive Model Development Summary

- Data Preparation: Cleaned and preprocessed the dataset, and then Split the dataset into training and test sets
- Model Training: Built and trained the following classification models: Logistic Regression, SVM, KNN, and Decision Tree
- Evaluation: Evaluated each model using Accuracy Score and Confusion Matrix
- Hyperparameter Tuning: Performed GridSearchCV on top models to find the best parameters, Improved performance by optimizing key hyperparameters
- Final Model Selection: Compared performance across models, Selected the best-performing model based on test accuracy and balanced metrics



GitHub Repository Link:

- Machine Learning Prediction Notebook: Link

Created by Durotoye Abdullahi Ololade

Results

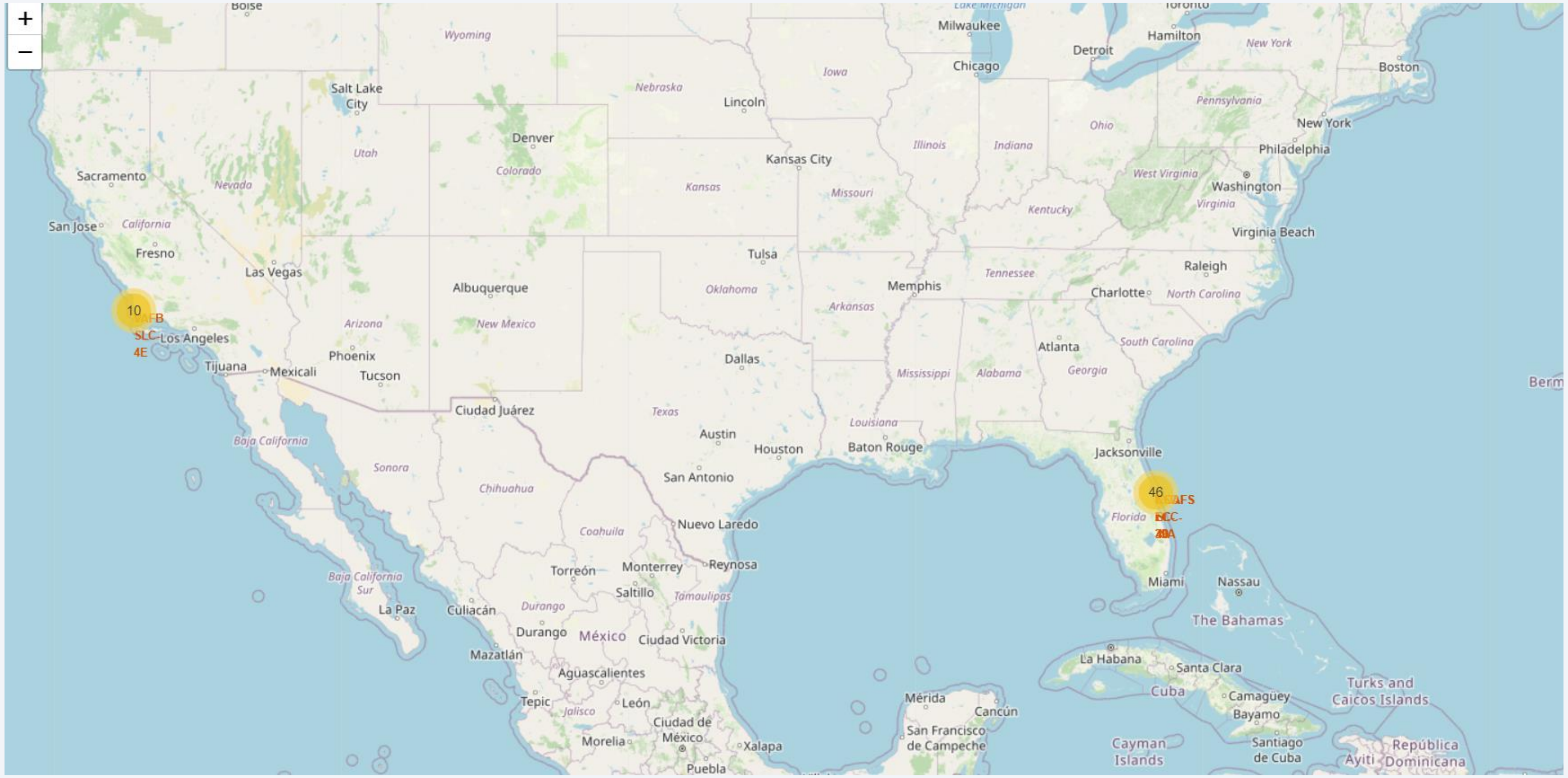
Exploratory Data Analysis (EDA) Results

- Flight Number vs. Launch Success:
  Success rates improved with higher flight numbers, indicating learning and process improvements over time.
- Payload Mass vs. Success:
  Rockets carrying heavier payloads still often landed successfully, especially in LEO, Polar, and ISS orbits.
- Launch Site vs. Outcome:
  Most launches occurred at CCAFS SLC 40, but KSC LC 39A had a higher success rate. VAFB SLC 4E had fewer launches with mixed outcomes.
- Orbit vs. Success Rate:
  Highest success rates (100%) were observed for ES-L1, GEO, HEO, and SSO orbits. GTO had the lowest success rate.
- Yearly Success Trend:
  Success rates have steadily increased from 2013 to 2020, showing overall improvement in mission outcomes.

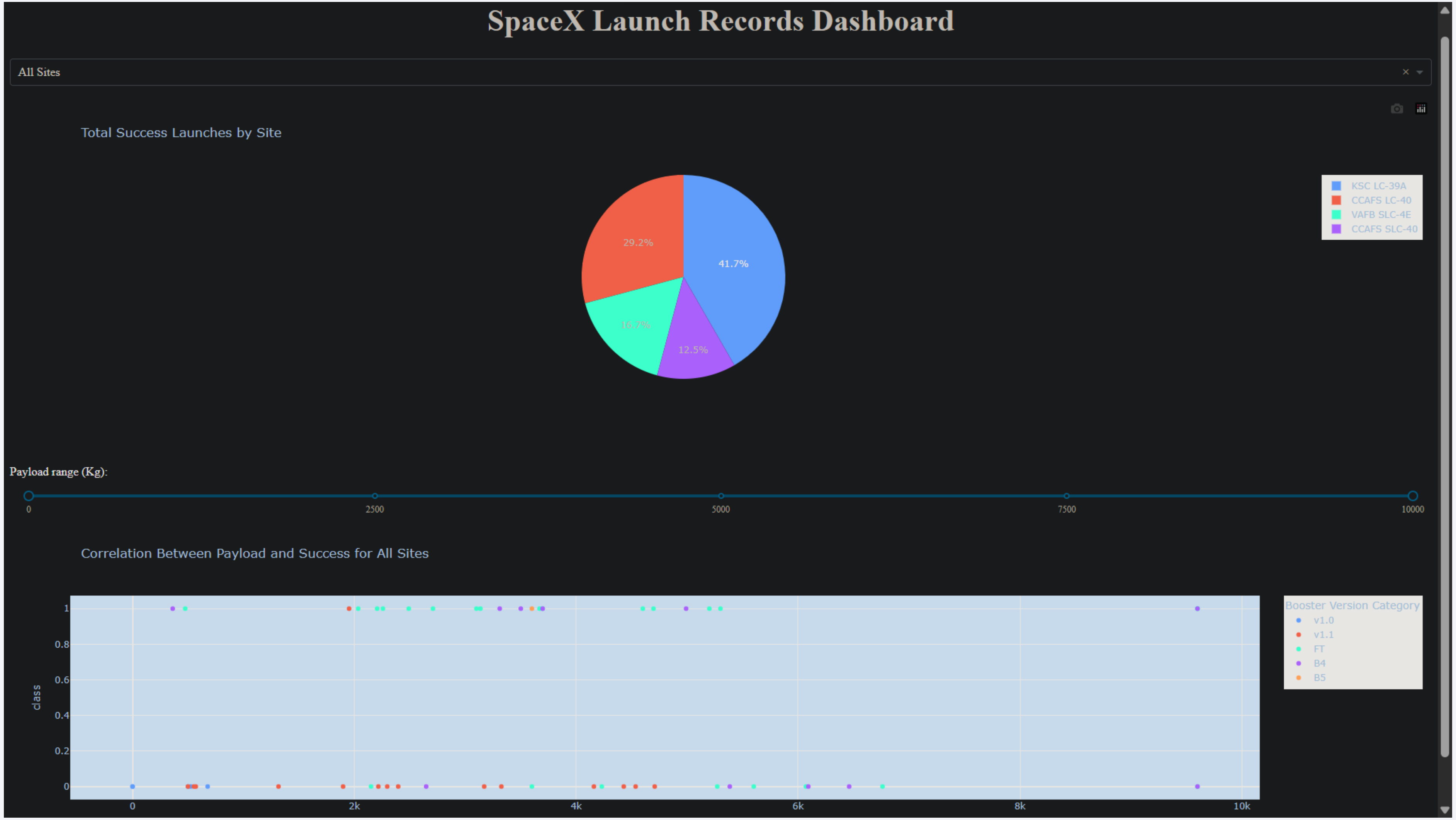
Results

Interactive analytics

- These objects and interactions enable users to explore the SpaceX data visually, providing deeper insight into success patterns by site, payload, and orbit.



Results



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Created by Durotoye Abdullahi Ololade

Results

Simple Question for the Interactive analytics Location Map

Are launch sites in close proximity to railways?

- Yes, based on the plotted distances, most launch sites appear to be within 1–5 KM from a railway.

Are launch sites in close proximity to highways?

- Yes, many launch sites are less than 1 KM from major roads or highways, aiding easy transportation of equipment.

Are launch sites in close proximity to coastline?

- Yes. All launch sites are within 1 KM or so from the coast, which is strategic for launch safety (rockets can fly over the ocean).

Do launch sites keep certain distance away from cities?

- Yes. Typically, launch sites are located at safe distances (10–50 KM) from urban areas, likely for safety and noise reasons.

Results

- Predictive analysis results

  Models Evaluated:

  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
  - Logistic Regression
  - Decision Tree

  Performance Outcome:

  - All four models achieved the same test set accuracy of 0.83.
  - Each model misclassified 3 instances where boosters did not land but were predicted as landed.

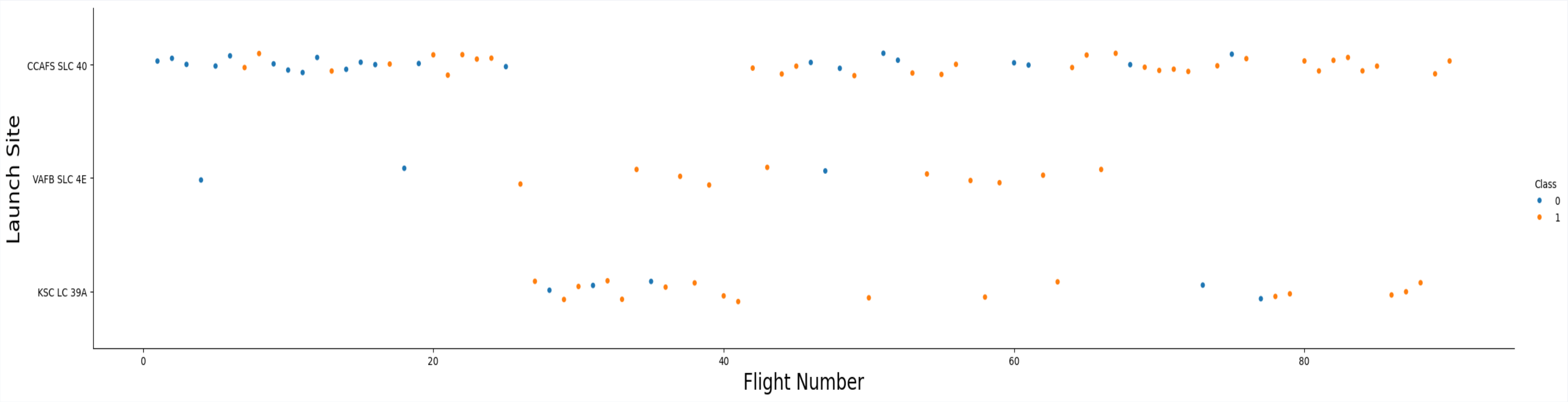
  Insight: The identical performance and consistent misclassification suggest the presence of external factors (e.g., weather, mechanical failure) not captured in the dataset, affecting model predictions.

Section 2

Insights drawn from EDA

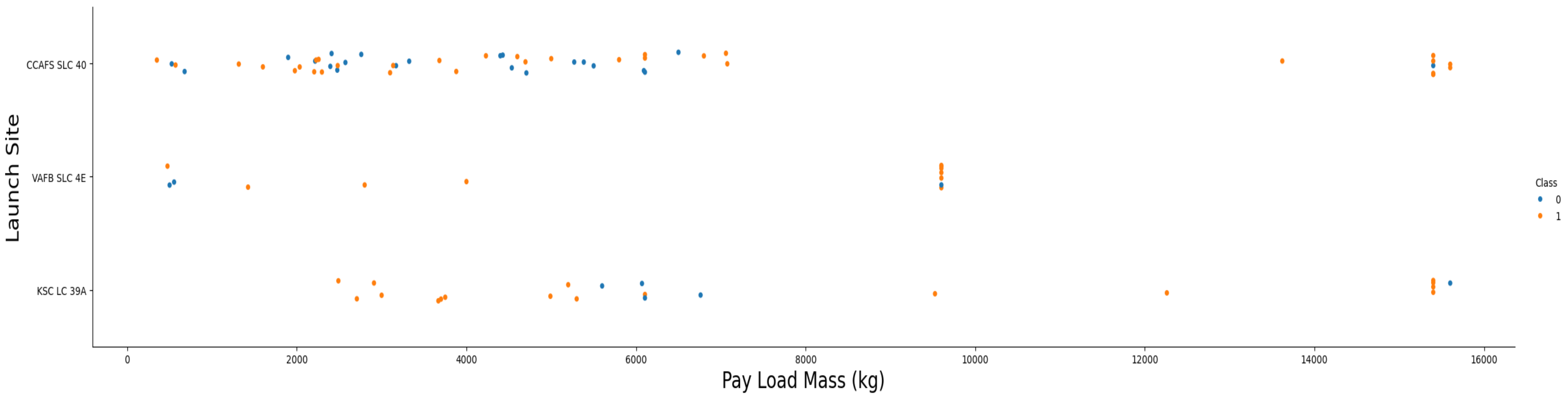
Created by Durotoye Abdullahi Ololade

Flight Number vs. Launch Site



This plot shows the success and failure of launches over time across different launch sites. Each dot represents a launch, color-coded by class (success or failure). CCAFS SLC 40 has the most launches, followed by KSC LC 39A and VAFB SLC 4E. Over time, the number of successful launches increases across all sites, indicating improved performance.

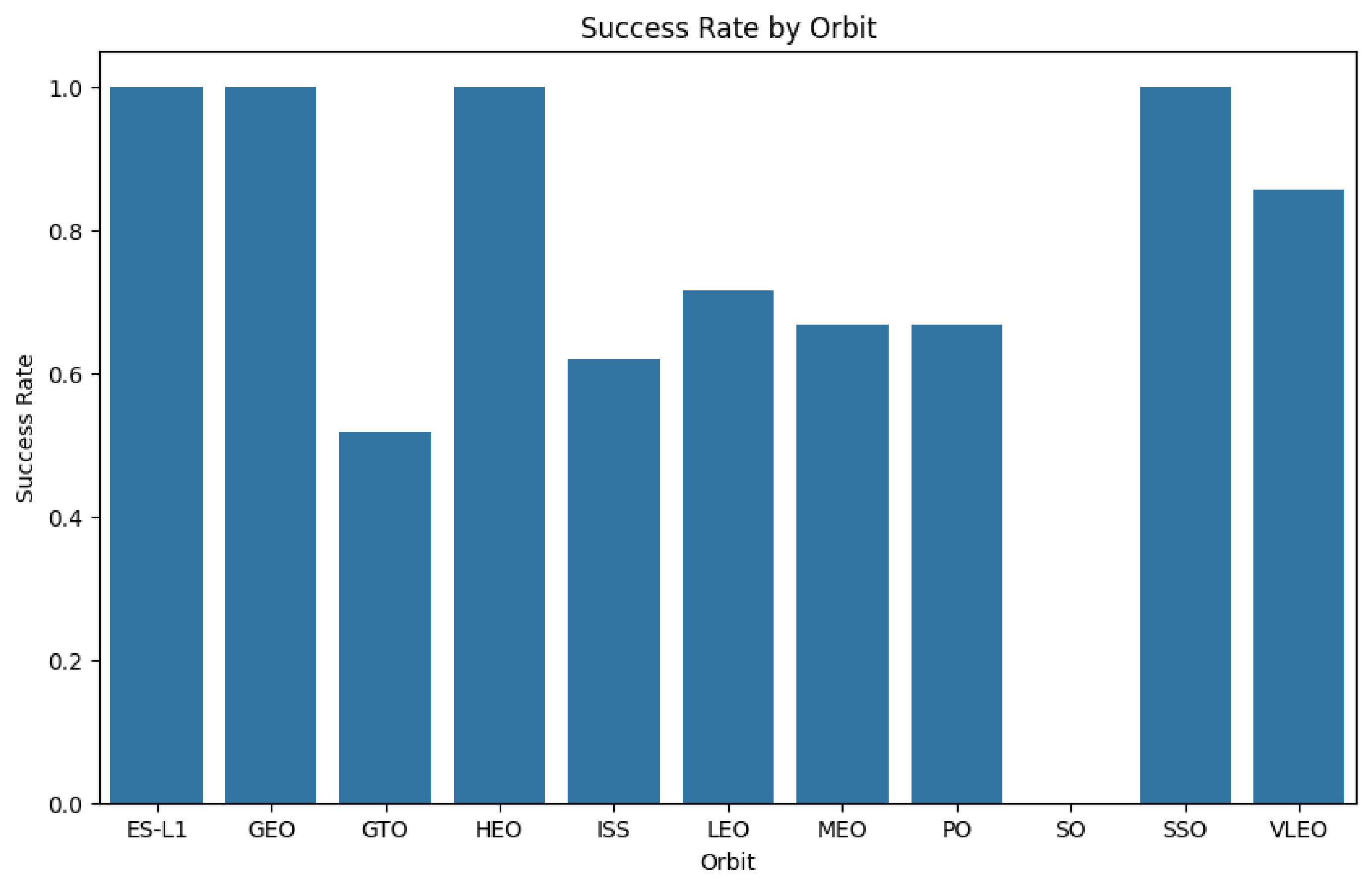
Payload vs. Launch Site



This scatter plot shows the relationship between payload mass and launch success at different sites. Successful missions (in orange) are observed across various payload sizes, especially in higher mass ranges. Launches above 10,000 kg are predominantly successful, reflecting reliability with heavier payloads. VAFB SLC 4E is frequently used for high-mass payloads with consistent success

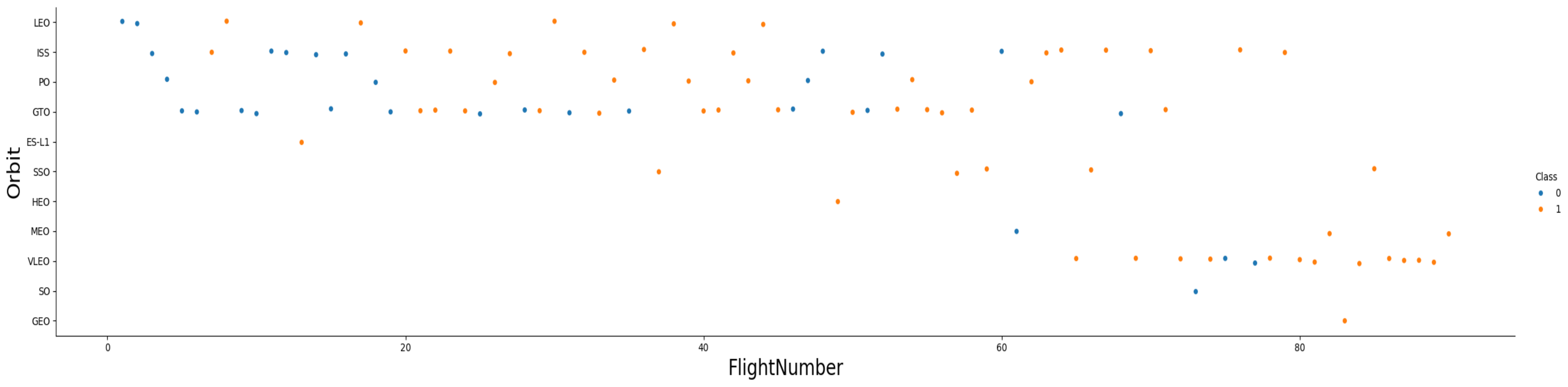
Success Rate vs. Orbit Type

The bar chart illustrates the success rate of launches based on the target orbit. ES-L1, GEO, HEO, and SSO orbits have achieved a 100% success rate. On the other hand, GTO shows the lowest success rate, indicating higher risk. Orbits like ISS, LEO, and MEO have moderate success rates, suggesting variability in mission outcomes.



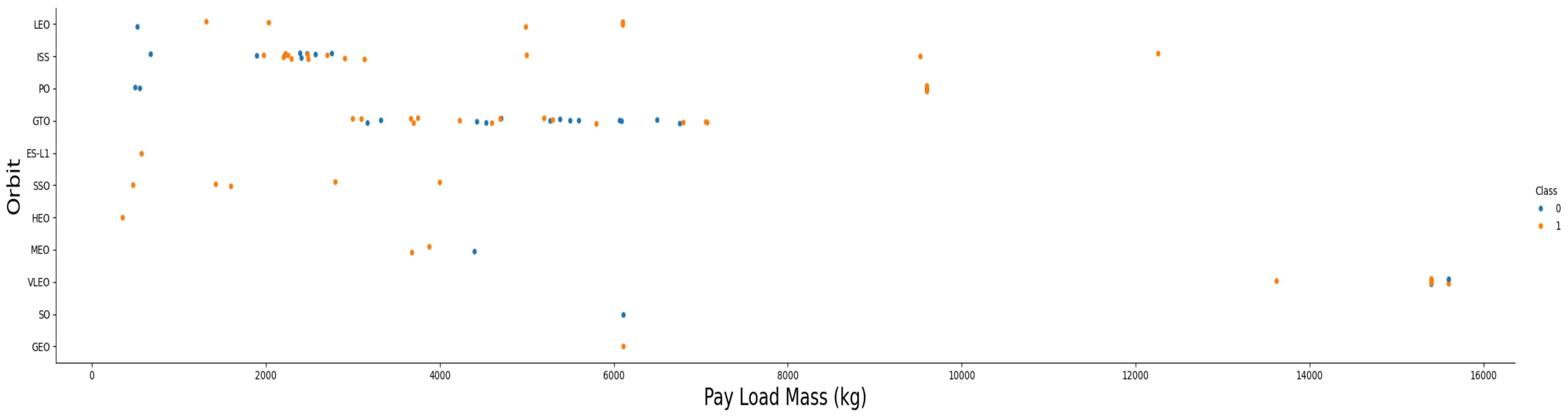
Created by Durotoye Abdullahi Ololade

Flight Number vs. Orbit Type



This scatter plot shows how launch success (class) varies with flight number and orbit type. As flight numbers increase (indicating more experience), the frequency of successful missions (class 1) becomes more dominant. This implies a learning curve, where increased experience leads to improved outcomes across different orbits.

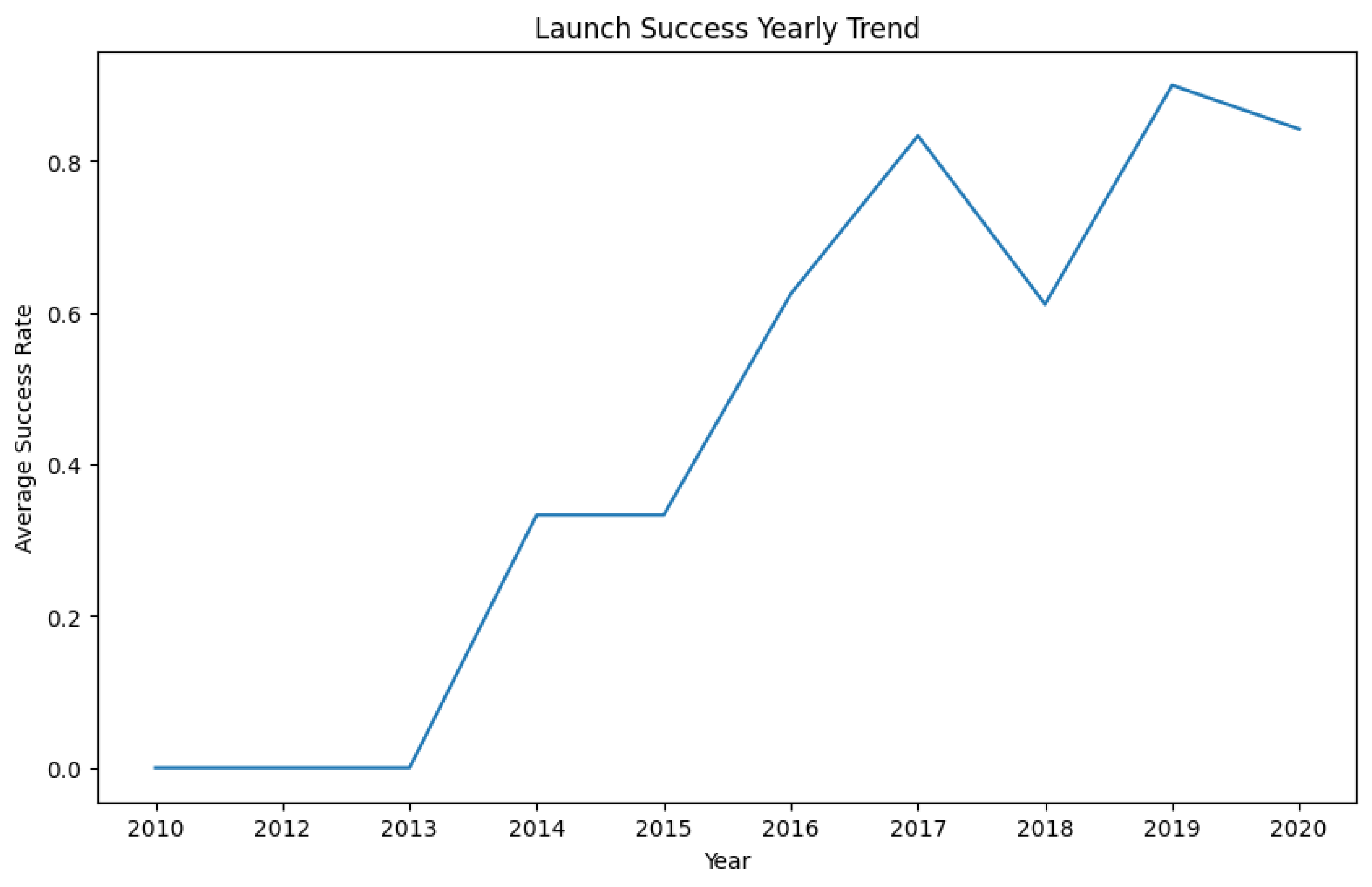
Payload vs. Orbit Type



This scatter plot compares payload mass (in kg) with the type of orbit, categorized by class (0 = failure, 1 = success). It suggests that while lighter payloads are sent to various orbits, successful launches (class 1) are more frequent regardless of mass. Heavier payloads tend to be associated with higher orbits like GEO or MEO, with a high success rate observed.

Launch Success Yearly Trend

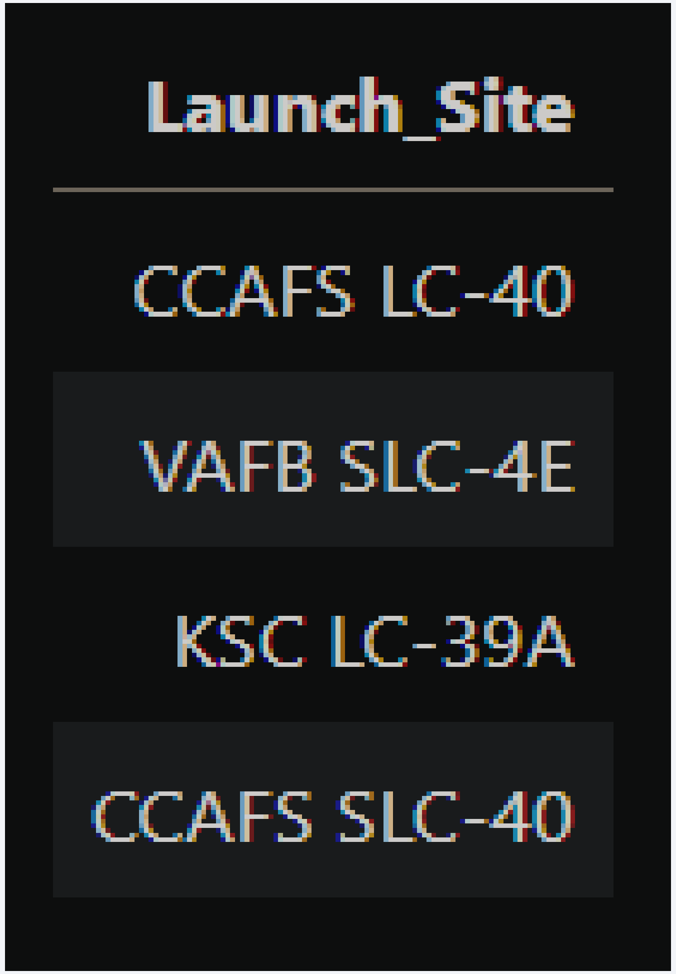
This line plot shows the average launch success rate per year from 2010 to 2020. It highlights a significant improvement in success rates, particularly from 2014 onward. The trend indicates that launch reliability increased over time, peaking in 2019 before slightly dropping in 2020. This could reflect technological improvements and better mission planning.



Created by Durotoye Abdullahi Ololade

All Launch Site Names

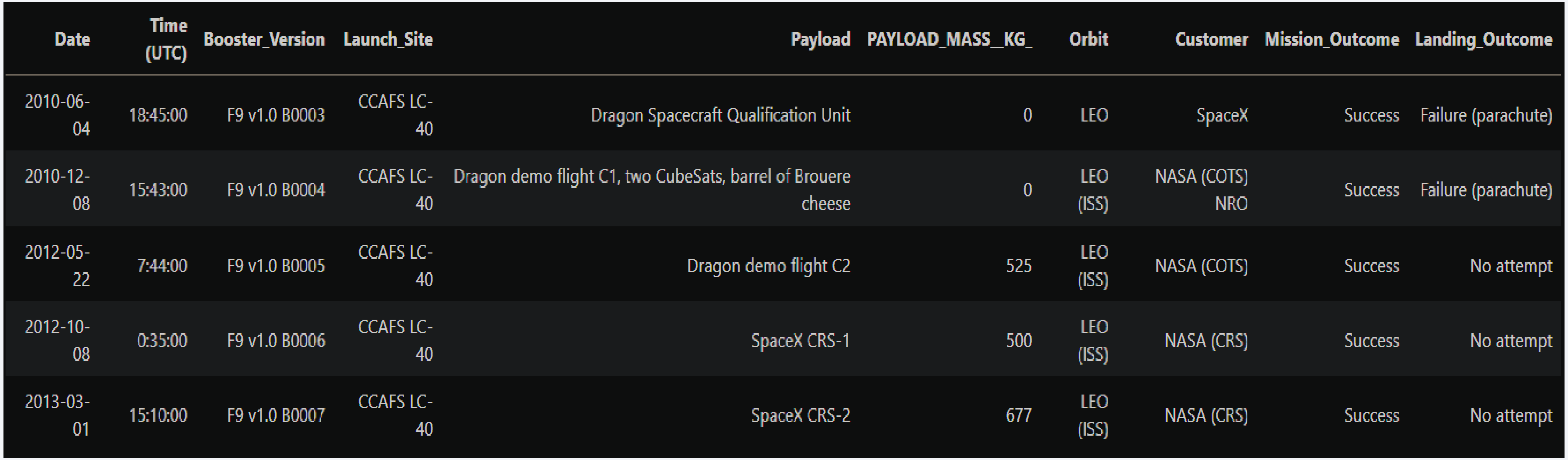
The Query Result shows the names of the unique launch sites in the SpaceX dataset



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

Created by Durotoye Abdullahi Ololade

Launch Site Names Begin with 'CCA'



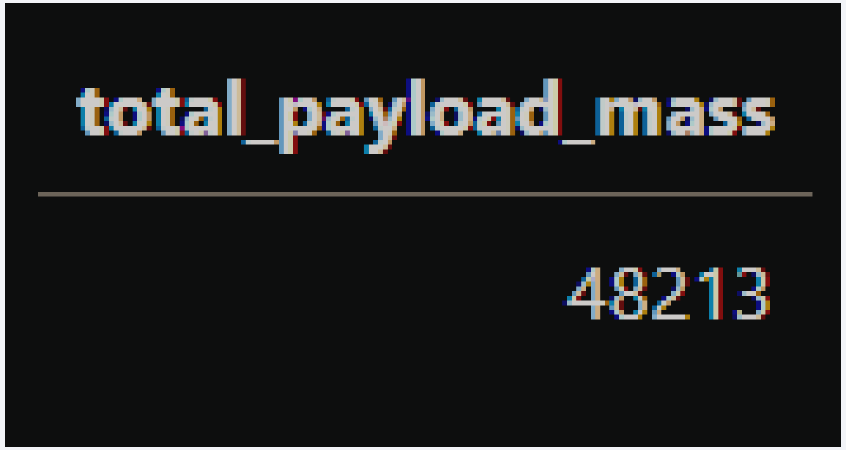
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

The Query Result shows the first 5 records where launch sites begin with `CCA`

Created by Durotoye Abdullahi Ololade

Total Payload Mass

The Query Result shows the Calculated total payload mass carried by boosters from NASA

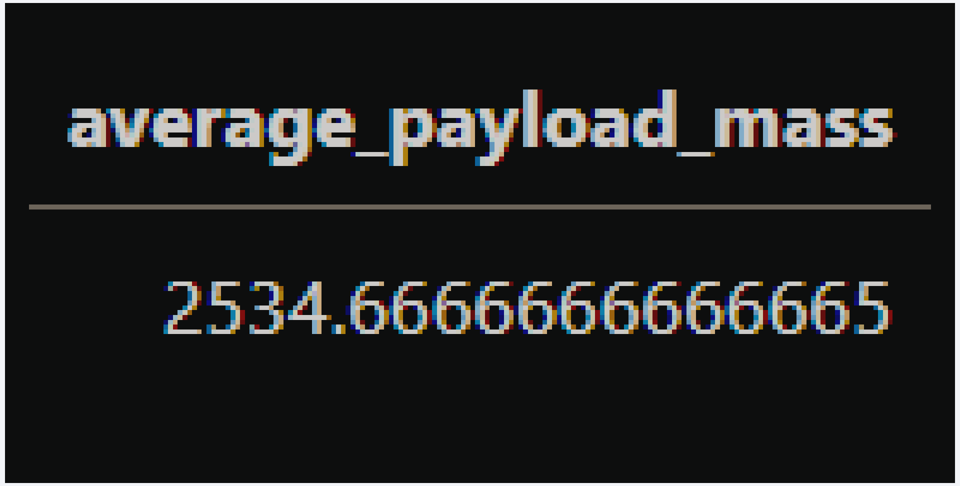


total_payload_mass
48213

Created by Durotoye Abdullahi Ololade

Average Payload Mass by F9 v1.1

The Query Result shows the Calculated average payload mass carried by booster version F9 v1.1

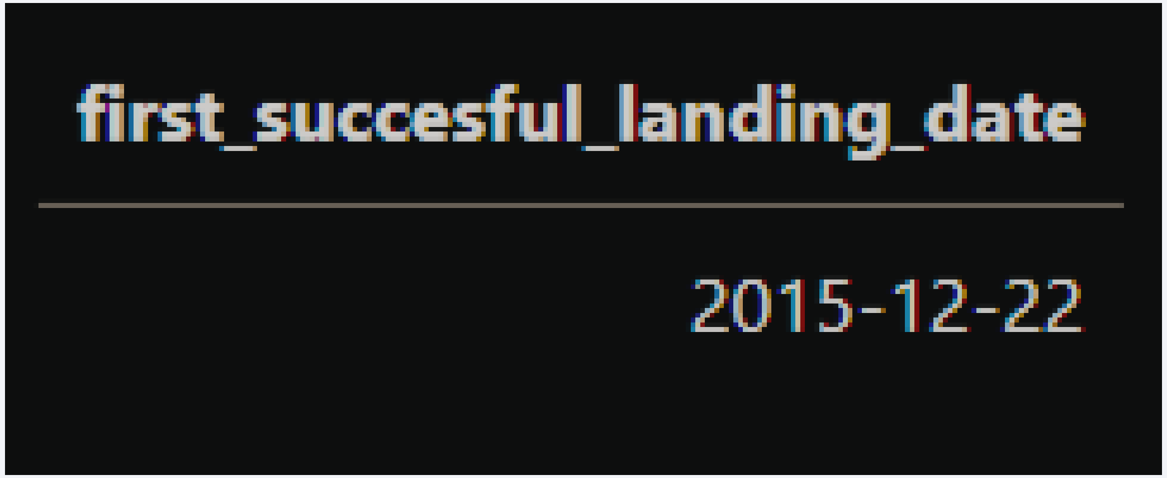


average_payload_mass
2534.6666666666665

Created by Durotoye Abdullahi Ololade

First Successful Ground Landing Date

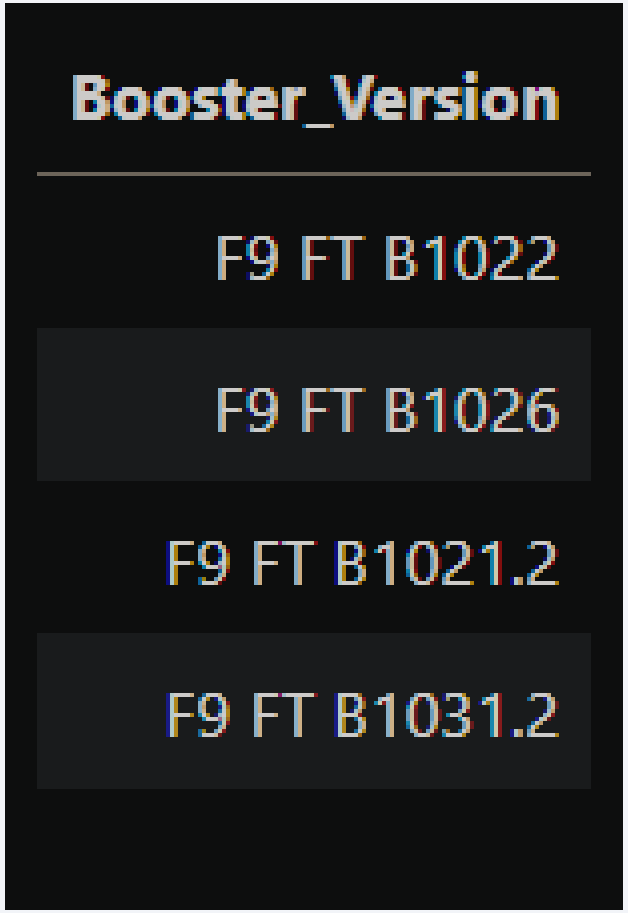
The Query Result shows the dates of the first successful landing outcome on ground pad



first_succesful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

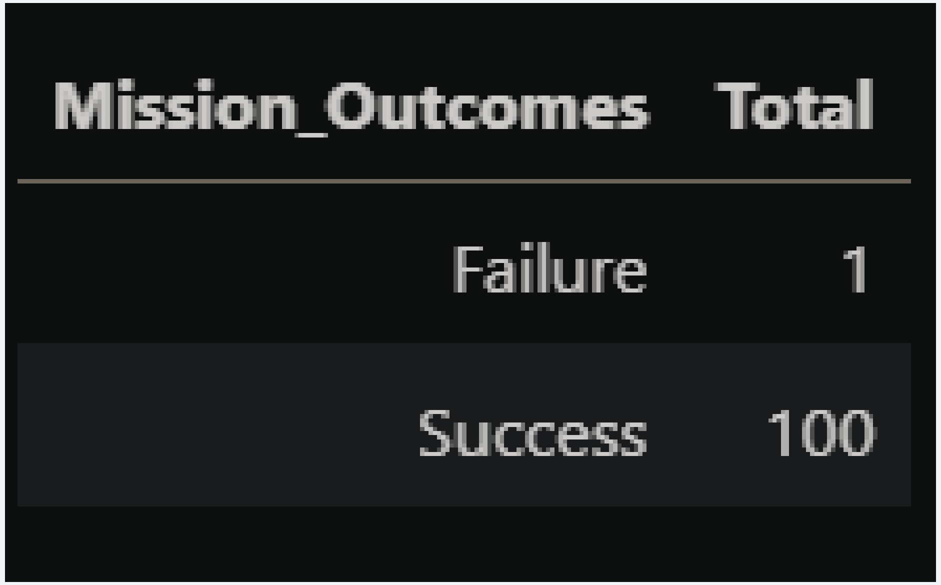


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Created by Durotoye Abdullahi Ololade

Total Number of Successful and Failure Mission Outcomes

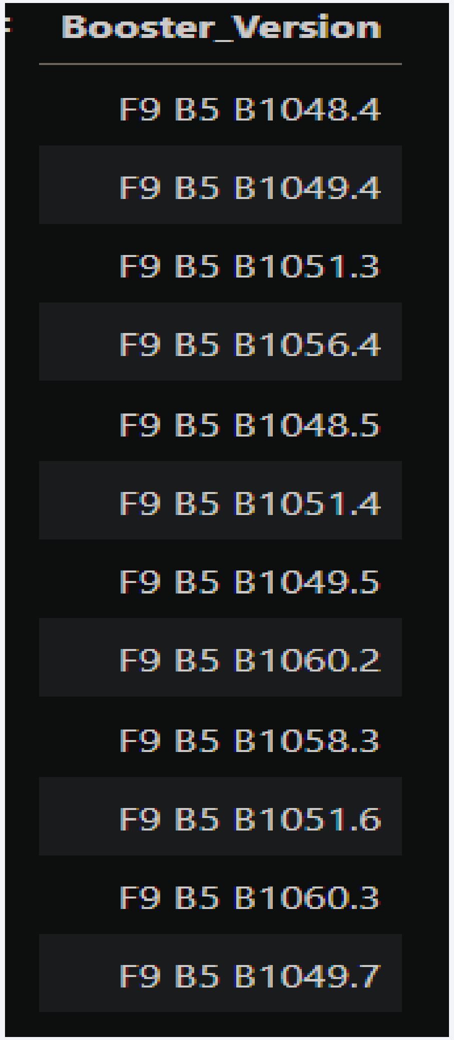
The Query Result shows the total number of successful and failure mission outcomes



Mission_Outcomes	Total
Failure	1
Success	100

Boosters Carried Maximum Payload

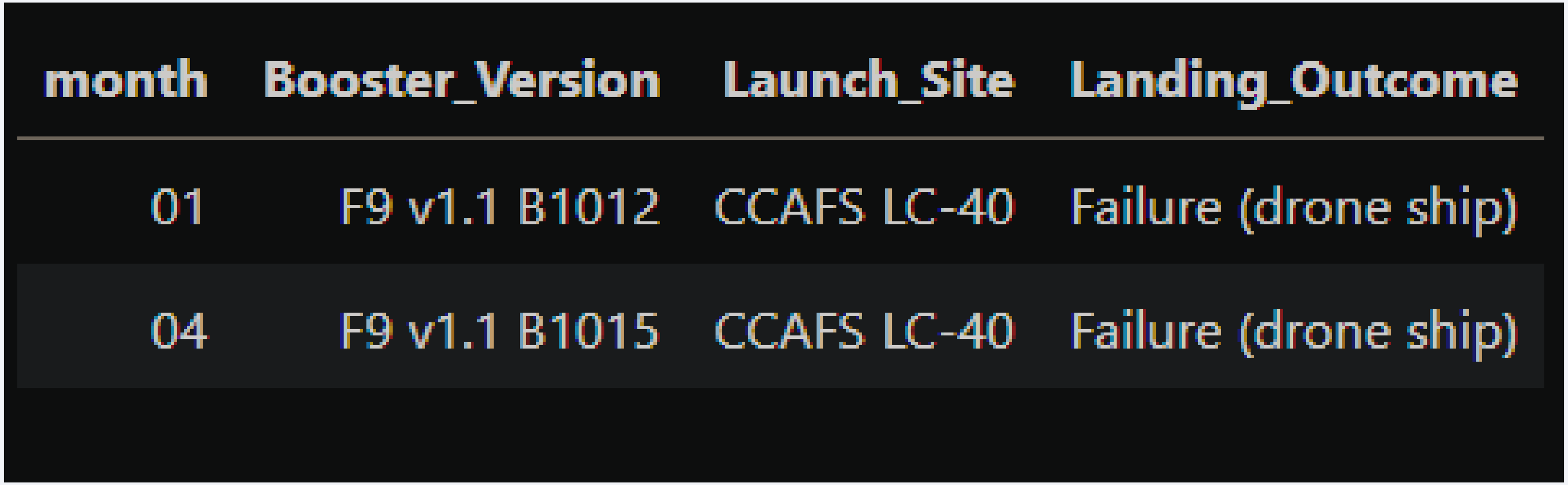
The Query Result shows the list of the names of the booster which have carried the maximum payload mass



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

Created by Durotoye Abdullahi Ololade

2015 Launch Records

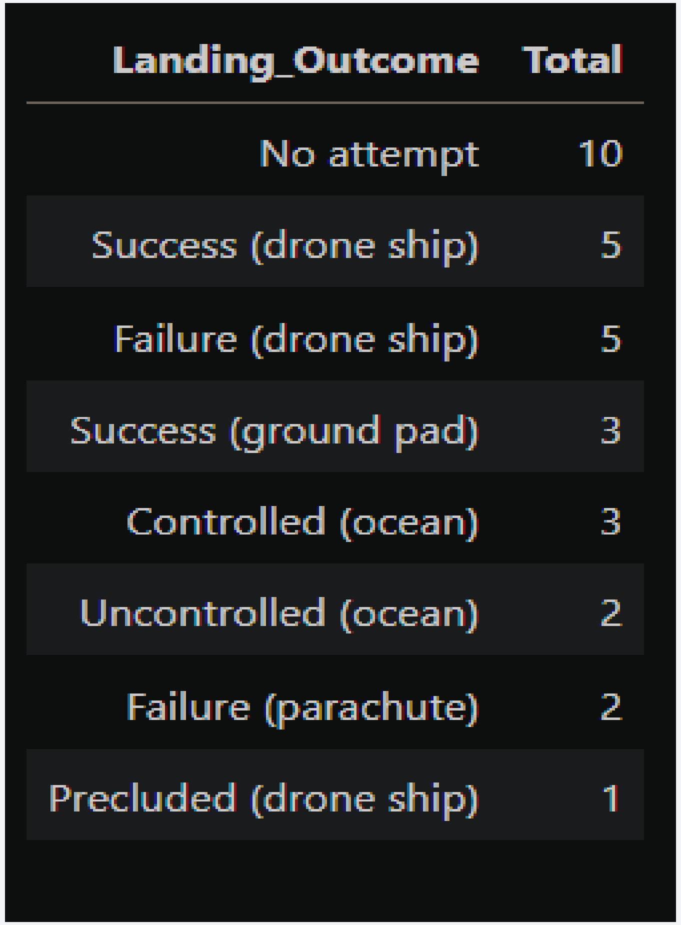


month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

This Screenshot explains the query result which list the records displaying the Month Names, Failure Landing Outcomes in drone ship ,Booster Versions and Launch Sites for the months in year 2015.

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

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



Landing_Outcome	Total
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

Created by Durotoye Abdullahi Ololade

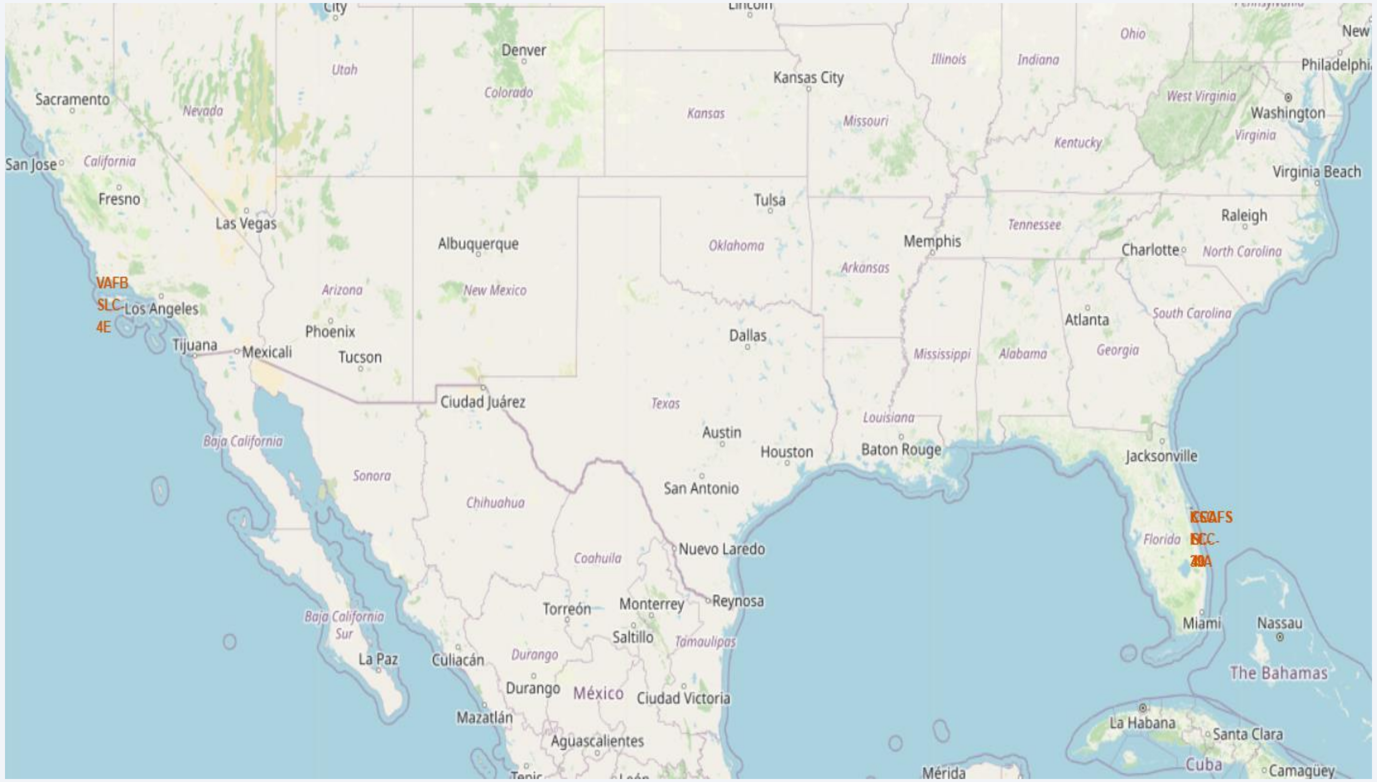
Section 3

Launch Sites Proximities Analysis

Created by Durotoye Abdullahi Ololade

All Launch Sites on a Global Map

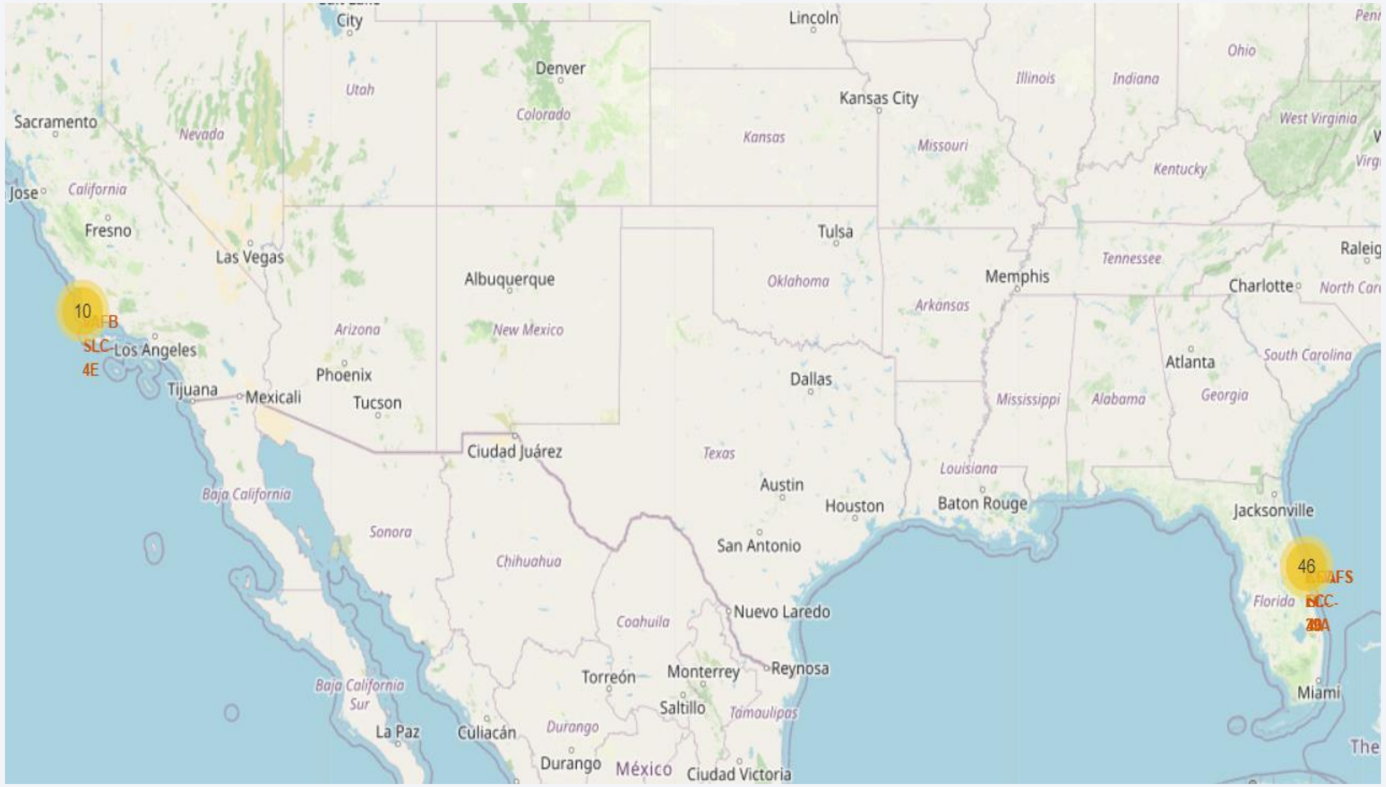
The map reveals that all major U.S. launch sites are strategically located along the coastlines, primarily in Florida and California. This placement minimizes risk by allowing rockets to launch over open water, avoiding populated areas. The eastern sites support low-inclination orbits, while western sites are ideal for polar orbits. The absence of launch facilities in the central U.S. underscores the importance of geographic safety and orbital efficiency in site selection.



Created by Durotoye Abdullahi Ololade

Launch Outcomes on the Global Map

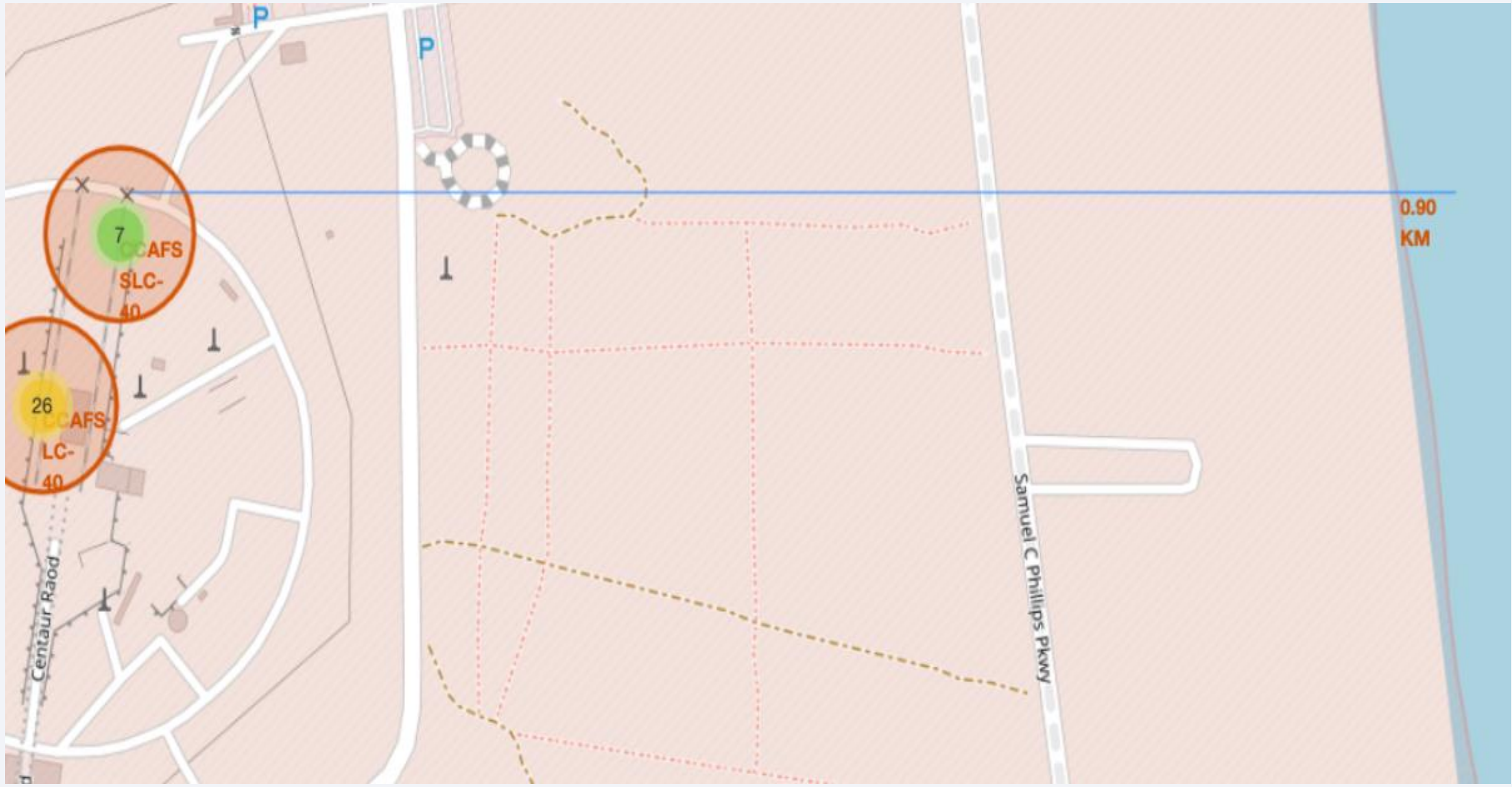
This shows a Folium map with U.S. rocket launch sites clustered in California (10 sites) and Florida (46 sites). Key locations include Vandenberg AFB in California and Cape Canaveral in Florida. These coastal sites support safe over-ocean launches and access to different orbital paths, reflecting their strategic importance for space missions. From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

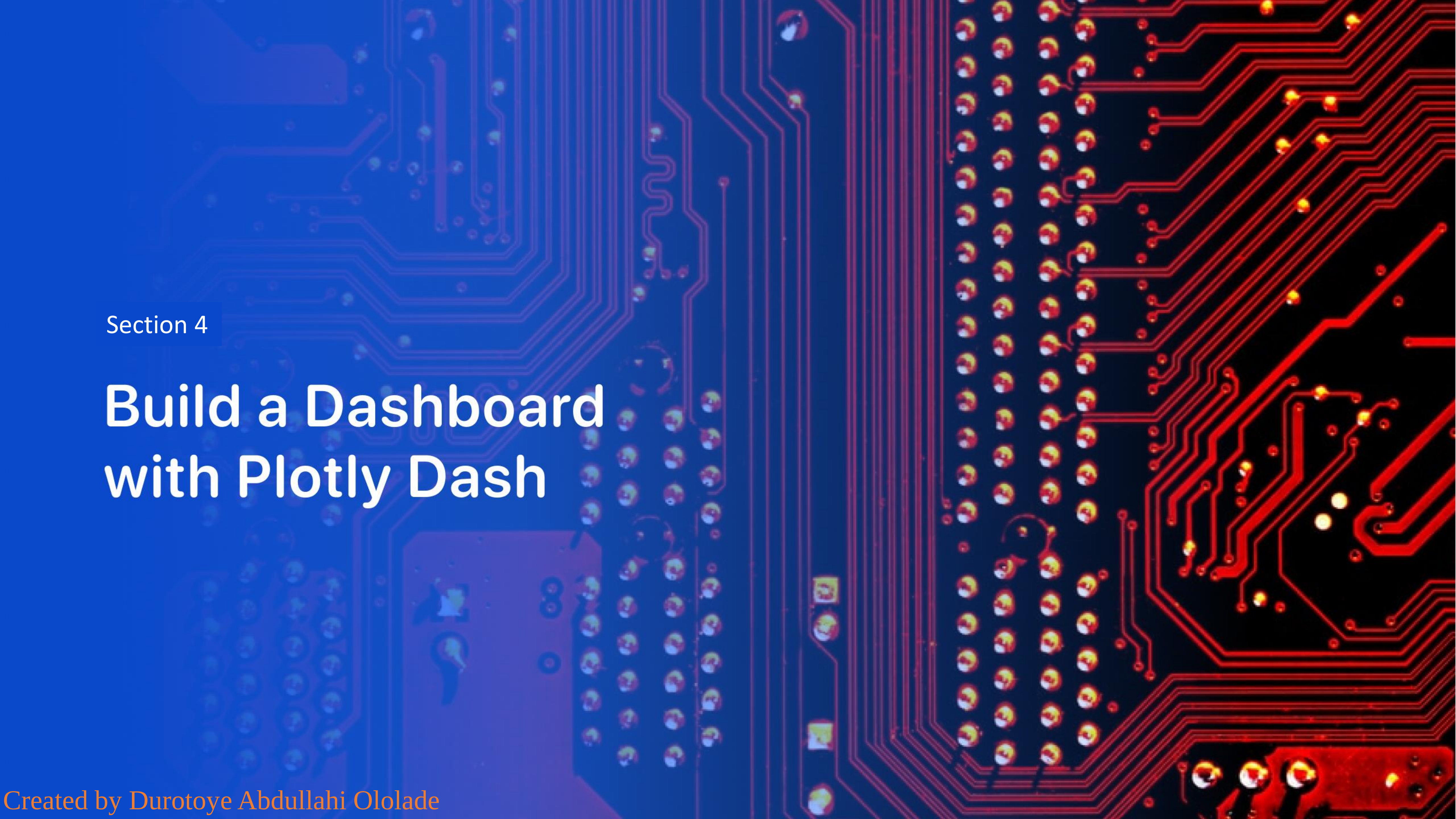


Created by Durotoye Abdullahi Ololade

CCAFS LC-40 Distance to Coastline

The screenshot shows CCAFS LC-40 located approximately 0.9 km from the coastline. This close proximity supports safe rocket launches over the Atlantic Ocean. The site is easily accessible via Samuel C. Phillips Parkway and nearby roads, ensuring smooth transport of equipment and personnel. The dense clustering of launch infrastructure highlights the area's role as a major hub for space operations.



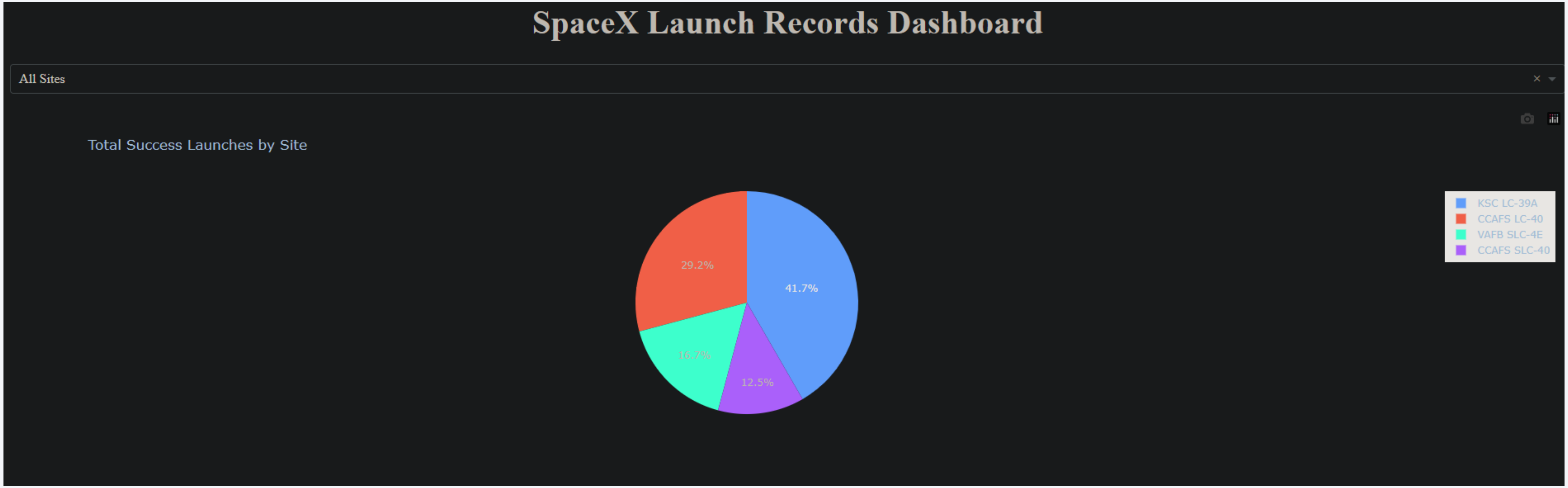


Section 4

Build a Dashboard with Plotly Dash

Created by Durotoye Abdullahi Ololade

Total Success Launches By Site

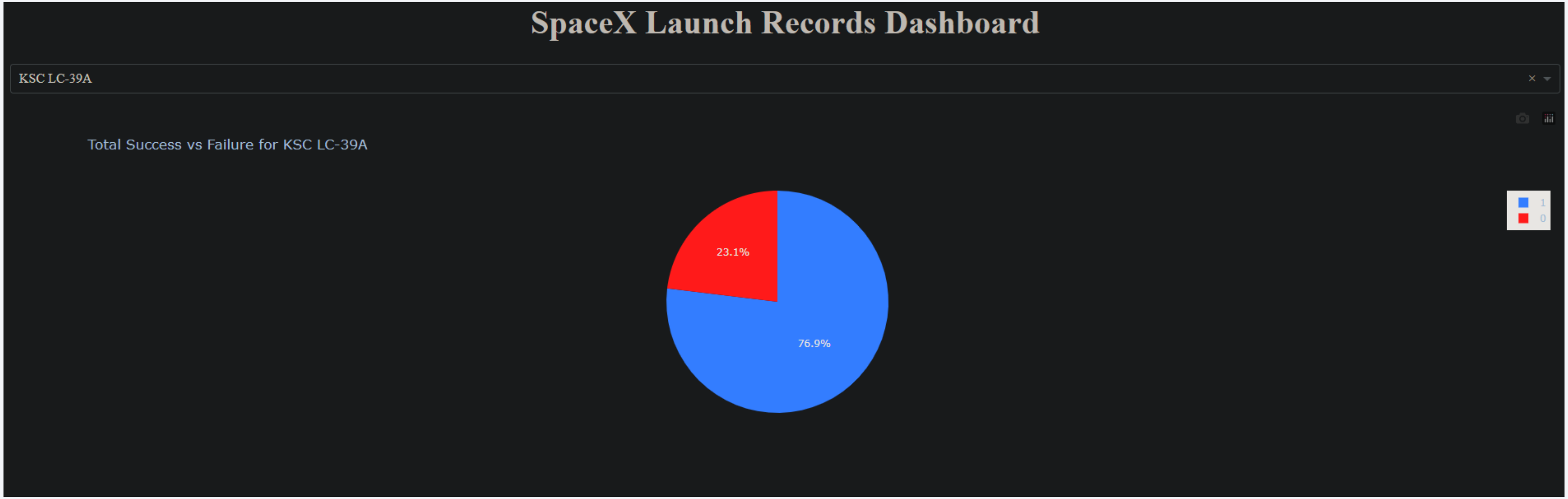


SpaceX Launch Records Dashboard

All Sites

From the Pie Chart, We can see that KSC LC-39A has the highest total success launches while CCAFS SLC-40 has the lowest total success launches.

Launch Site with Highest Launch Success Ratio



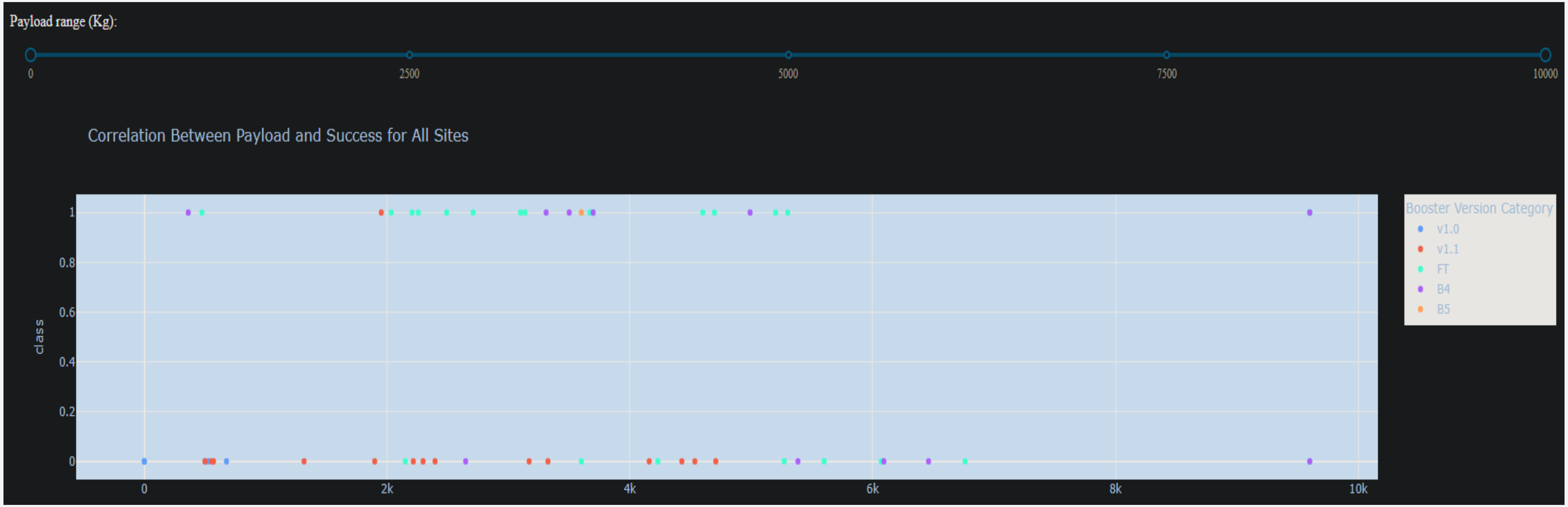
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success vs Failure for KSC LC-39A

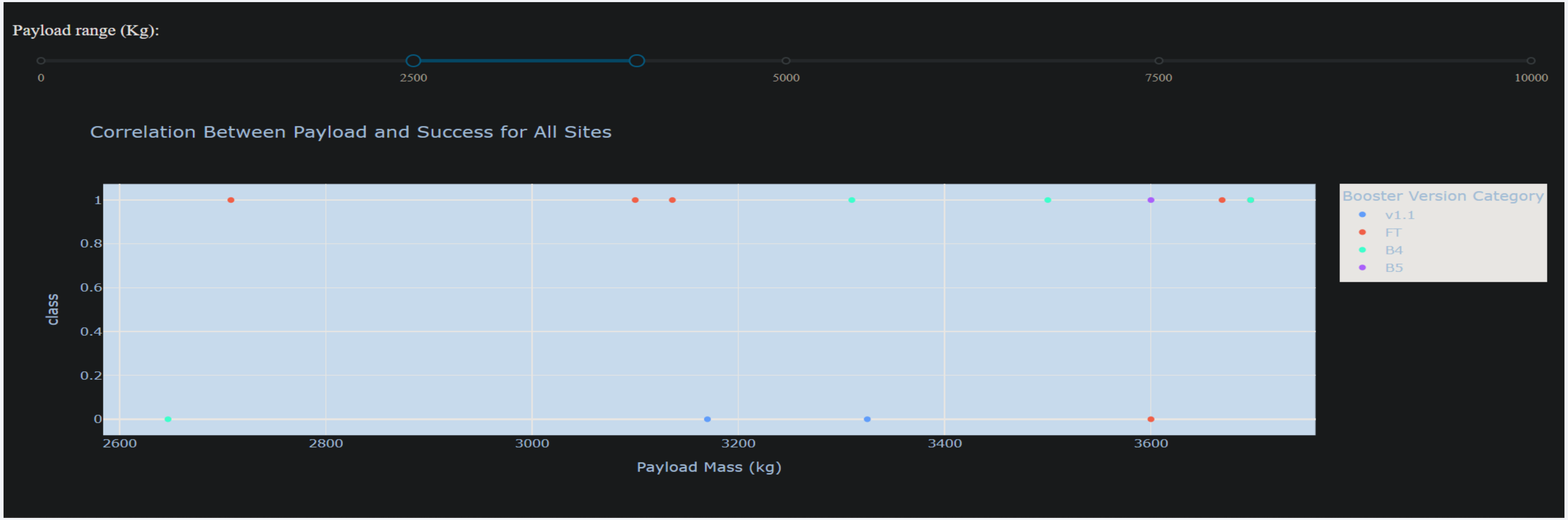
From the Pie Chart, We can see that KSC LC-39A has the highest Launch Success Ratio with 76.9% successes and 23.1% Failures.

Payload vs. Launch Outcome scatter plot for all sites



From the screenshot of the graph, we can see that the booster version with the largest success rate was the FT booster version with about 65% successes.

Payload vs. Launch Outcome scatter plot for all sites



From the screenshot of the graph, we can see that the payload range with the largest success rate was between with 2500-4000.

Created by Durotoye Abdullahi Ololade

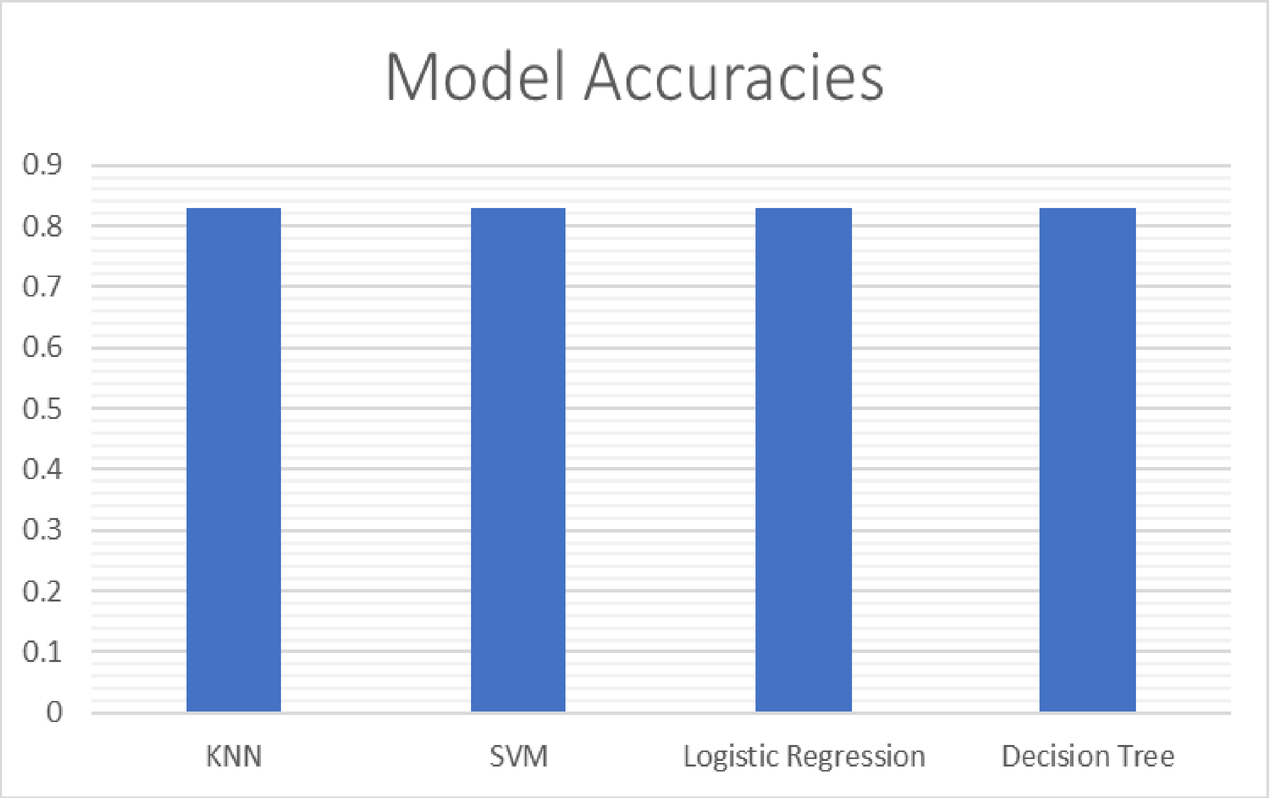
Section 5

Predictive Analysis (Classification)

Created by Durotoye Abdullahi Ololade

Classification Accuracy

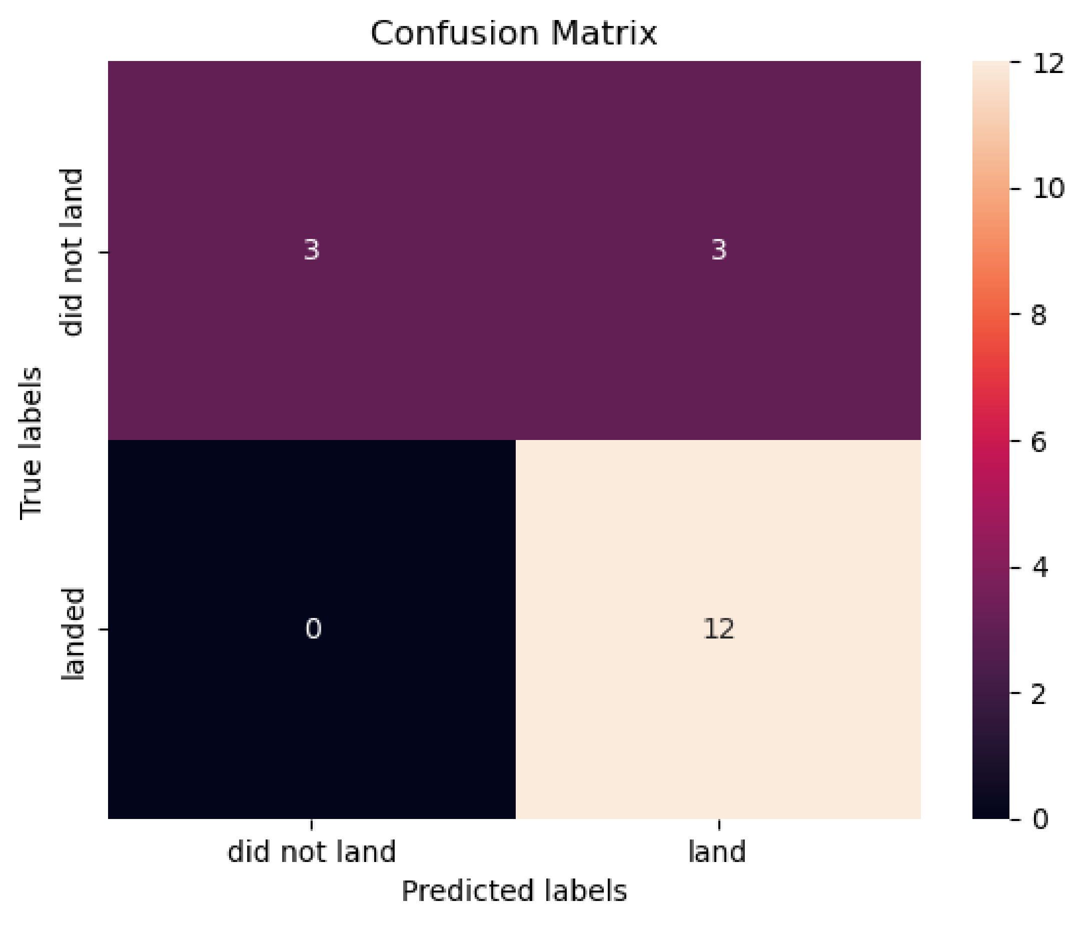
The bar chart displays the classification accuracy of four models: KNN, SVM, Logistic Regression, and Decision Tree. Each bar reaches a value of 0.83, indicating that all models performed equally well on the SpaceX Dataset. Since the accuracy values are identical, no single model outperforms the others in this metric. This could suggest that the dataset is either relatively simple to classify or that the models are well-tuned and perform similarly.



Created by Durotoye Abdullahi Ololade

Confusion Matrix

The confusion matrix shows that out of 18 samples, the model correctly classified all 12 instances of class 1, resulting in a perfect recall of 1.00 for that class. However, for class 0, it only correctly predicted 3 out of 6 cases, leading to a 50% accuracy for class 0. This indicates that the model leans toward predicting class 1, possibly due to class imbalance or stronger patterns in that class.



Created by Durotoye Abdullahi Ololade

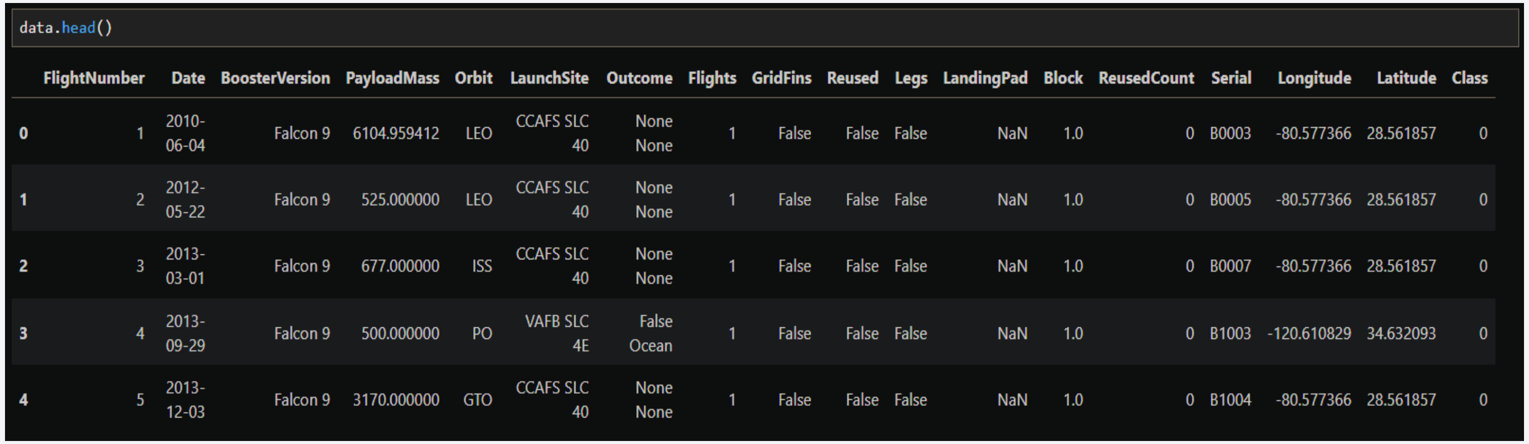
Conclusions

- This project effectively demonstrates how data science techniques spanning data acquisition, wrangling, visualization, and predictive modeling can be applied to the aerospace industry.
- The ability to predict Falcon 9's first stage landing success provides a significant strategic advantage, as it directly correlates with cost efficiency and mission planning.
- Through this capstone, we not only replicated real-world data analysis workflows but also gained actionable insights into spaceflight logistics that a company like "Space Y" could leverage in a competitive market.

Appendix

- Final Dataset for the Modeling

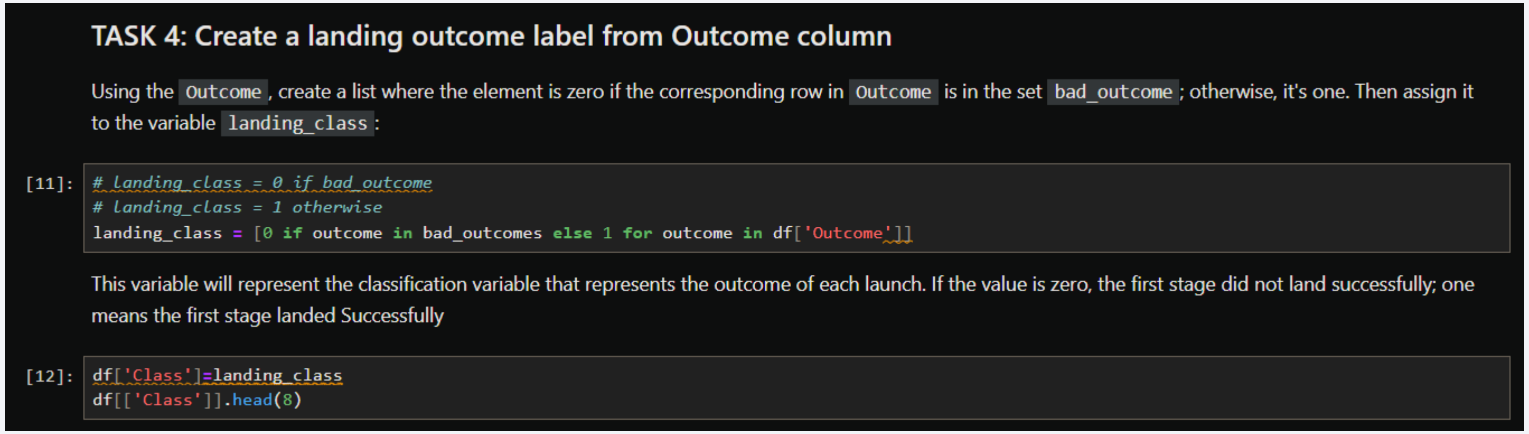
  Dataset: Link



```
data.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

- Python code for creating a landing outcome label from Outcome column



TASK 4: Create a landing outcome label from Outcome column

Using the `Outcome`, create a list where the element is zero if the corresponding row in `Outcome` is in the set `bad_outcome`; otherwise, it's one. Then assign it to the variable `landing_class`:

```
[11]: # landing_class = 0 if bad_outcome
      # landing_class = 1 otherwise
      landing_class = [0 if outcome in bad_outcomes else 1 for outcome in df['Outcome']]
```

This variable will represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

```
[12]: df['Class']=landing_class
      df[['Class']].head(8)
```

Created by Durotoye Abdullahi Ololade

Thank you!

Created by Durotoye Abdullahi Ololade