



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Keith Mitchell  
June 11, 2023



# Executive Summary

In this presentation, I explore the commercial space industry with a focus on SpaceX and its cost advantage in rocket launches due to first stage reuse. I highlight SpaceX's achievements, such as sending spacecraft to the International Space Station, launching the Starlink satellite internet constellation, and conducting manned missions to space. By employing methodologies including data gathering, data wrangling, data visualization, EDA using SQL and Pandas, and machine learning, I aim to predict the likelihood of first stage reuse. These predictions have implications for determining launch pricing and advancing reusable rocket technology.





# Outline

---

Executive Summary

---

Introduction

---

Methodology

---

EDA Insights

---

Proximities Analysis

---

Dashboards

---

Predictive Analysis (Classification)

---

Appendix

---

Conclusion



# Introduction

In this project, my role as a data scientist is to analyze the commercial space industry, with a focus on companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX. The main objective is to determine launch pricing for Space Y, a new rocket company aiming to compete with SpaceX. I will gather information about SpaceX, create informative dashboards, and use machine learning to predict whether SpaceX will reuse the first stage of their rockets.

SpaceX has achieved significant milestones, including sending spacecraft to the International Space Station, launching the Starlink satellite internet constellation, and conducting manned missions to space. Their cost advantage stems from the ability to reuse the first stage, with Falcon 9 rocket launches priced at \$62 million compared to other providers' costs exceeding \$165 million. Visual diagrams will help illustrate the size difference between the first and second stages, and the fairings that enclose the payload.

The project's focus is on predicting the successful reuse of the first stage. Not all landings are successful, and sometimes the first stage is intentionally sacrificed based on mission parameters. By completing this project, I will determine the pricing of each launch for Space Y, contribute to the advancement of reusable rocket technology, and enhance competitiveness in the commercial space industry.

Section 1

# Methodology



# Methodology



## Data collection methodology used:

API integration and web scraping to collect relevant information on SpaceX, rocket launches, and first stage landings

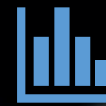


## Performed data wrangling:

Calculations were made to identify the number of launches at each site, the number and occurrence of each orbit, the number and occurrence of mission outcome per orbit type, and created a landing outcome label from the outcome data



## Performed exploratory data analysis (EDA) using visualization and SQL



## Performed interactive visual analytics using Folium and Plotly Dash



## Performed predictive analysis using classification models:

Machine learning used to determine the first stage of Falcon 9 landing outcome. Split data into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression.

# Data Collection

In this capstone project, I collected data for SpaceX launch analysis by utilizing two methods: API integration and web scraping.

Firstly, I accessed the SpaceX REST API to gather information about past launches, including rocket details, payload information, launch specifications, landing specifics, and outcomes. Through GET requests to specific API endpoints, such as /capsules, /cores, and /launches/past, I retrieved the data in JSON format.

Secondly, to complement the API data, I employed web scraping techniques using Python's BeautifulSoup package. By extracting valuable Falcon 9 launch records from HTML tables, I expanded the dataset with additional insights. This combination of API integration and web scraping enabled a comprehensive collection of data for further analysis and prediction in the SpaceX launch analysis.



# Data Collection – SpaceX API

- GitHub URL - [here](#)



Imported libraries



Defined helper functions



Extracted valuable information from the "cores" endpoint



Requested rocket launch data from SpaceX API



Requested and parsed the SpaceX launch data using the GET request



Created a Pandas dataframe with filter to only show 'Falcon 9' launches



Used the mean and the .replace() function to replace np.nan values



Exported to CSV



# Data Collection – Web Scraping

---

- GitHub URL - [here](#)



Requested the Falcon9  
Launch Wiki page from its  
URL



Created a BeautifulSoup  
object from the HTML  
response



Extracted all  
column/variable names  
from the HTML table  
header



Created a data frame by  
parsing the launch HTML  
tables



Exported to CSV

# Data Wrangling

---

In the dataset, there are different scenarios regarding the successful landing of boosters. Some landing attempts were successful, while others were not. The outcomes are categorized as follows:

- True Ocean indicates a successful landing in a specific region of the ocean
- False Ocean signifies an unsuccessful landing in the ocean
- True RTLS represents a successful landing on a ground pad
- False RTLS indicates an unsuccessful landing on a ground pad
- True ASDS denotes a successful landing on a drone ship
- False ASDS signifies an unsuccessful landing on a drone ship

I converted these outcomes into Training Labels where '1' means the booster successfully landed and '0' means the landing was unsuccessful.

GitHub URL - [here](#)



Calculated the number of launches on each site



Calculated the number and occurrence of each orbit



Calculated the number and occurrence of mission outcome per orbit type



Created a landing outcome label from 'Outcome' column

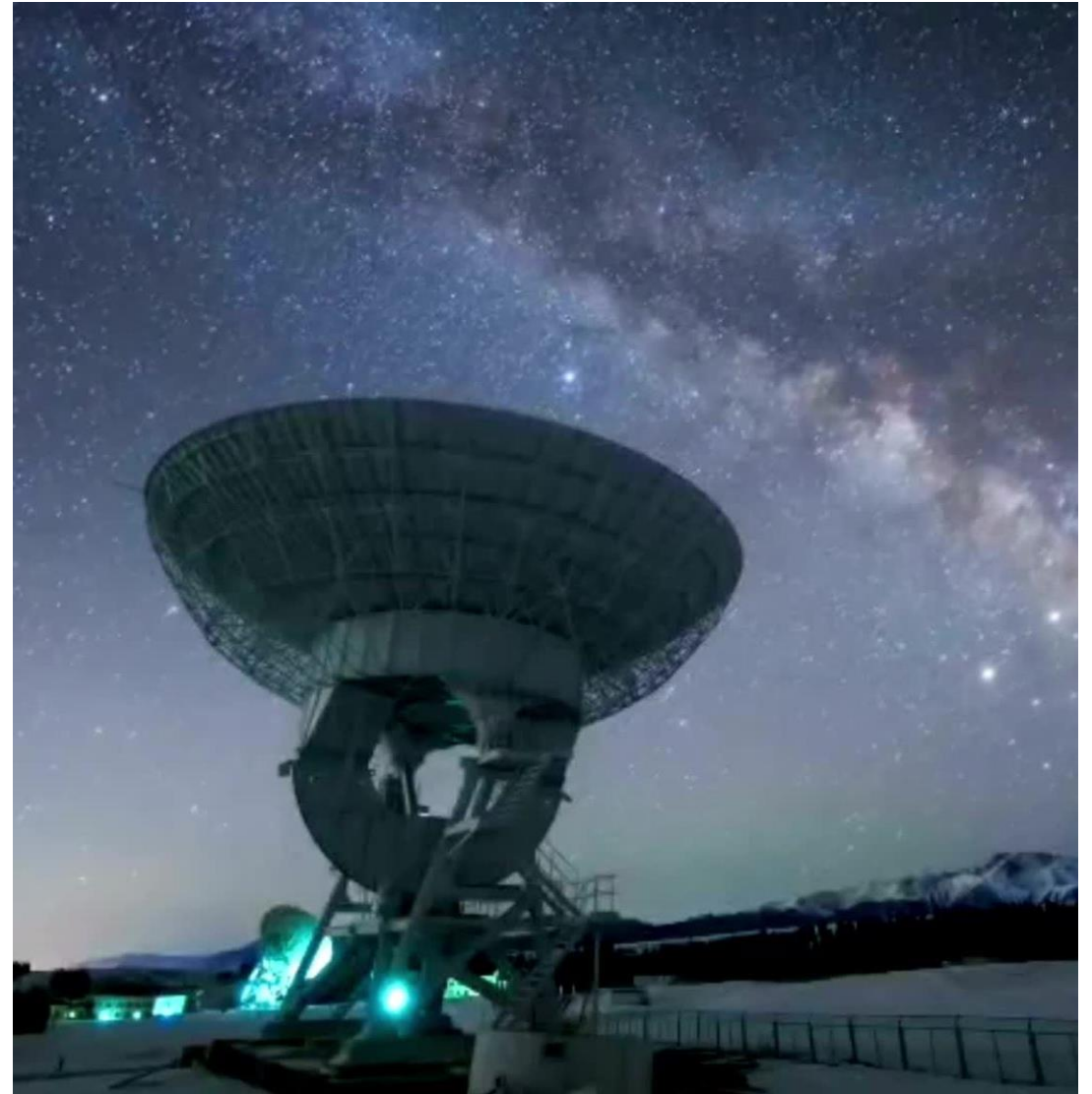


Exported to CSV

# Exploring Data with SQL

Summary of SQL queries performed:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved.
- Listed the names of the boosters that had success in drone ship and payload mass greater than 4000 but less than 6000

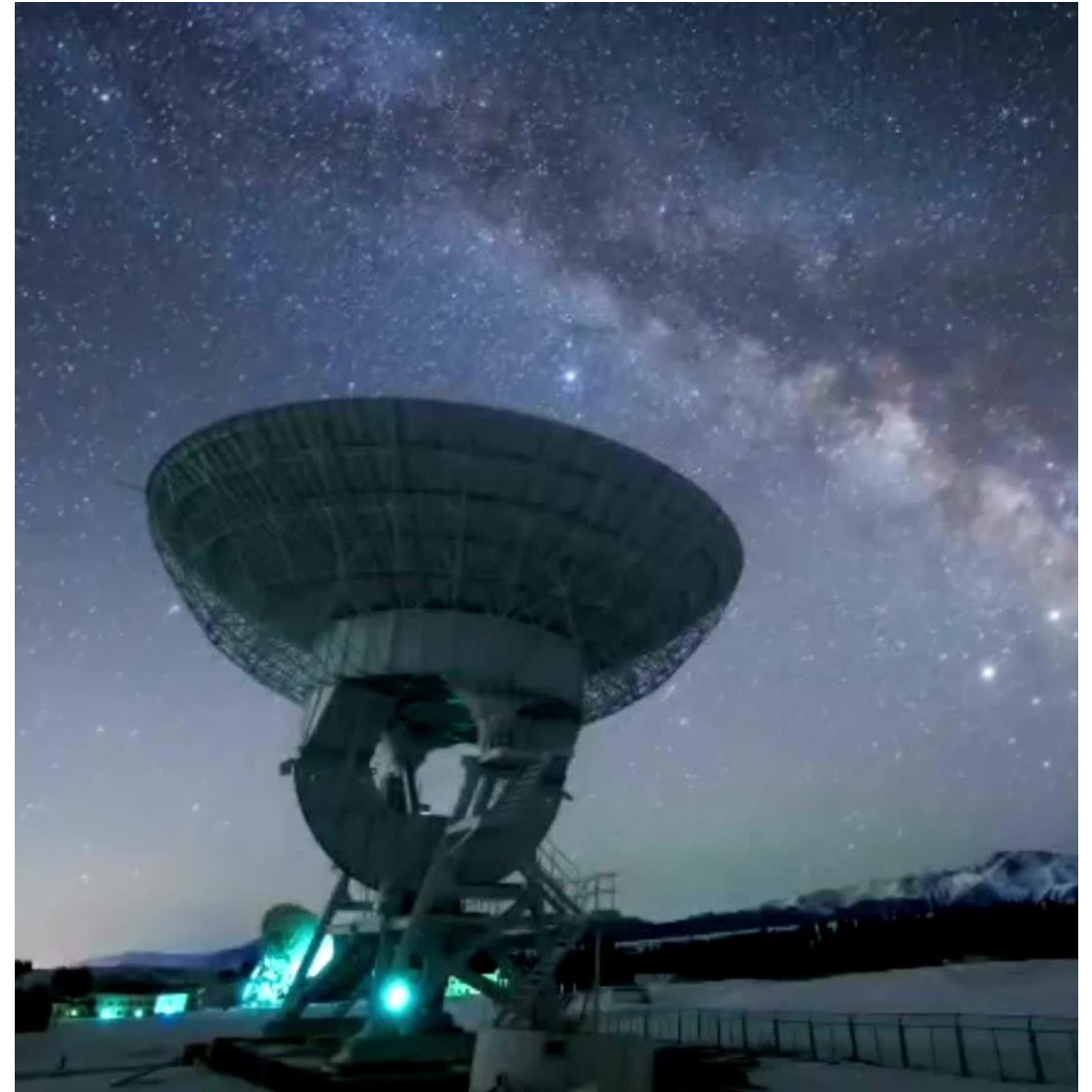




# Exploring Data with SQL (cont.)

- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass. Use a subquery
- Listed the records which will display the month names, failure landing outcomes in drone ship ,booster versions, and launch site for the months in year 2015.
- Ranked the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL - [here](#)



# Creating Engaging Maps with Folium

- folium.Circle was used to add a highlighted circle area for all launch sites
- Red markers were used for failed launches
- Green markers were used for successful launches
- Markers were added to clusters
  - From the color-labeled markers in marker clusters, it is easy to identify which launch sites have relatively high success rates.
- MousePosition was added on the map to get coordinates for a mouse over point on the map. This makes it so you can easily find the coordinates of any points of interests
- I used a polyline to display the distance between the coastline point and the launch site

GitHub link - [here](#)

# Building Dynamic Dashboards with Plotly Dash

A dashboard was created with the following;

- A dropdown list to enable Launch Site selection
- A pie chart to show the total successful launches count for all sites
  - If a specific launch site is selected, the Success vs. Failed counts for the site is shown
- A scatter chart to show the correlation between payload and launch success
- A callback function to render success-pie-chart based on selected site dropdown
- A callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output

GitHub URL - [here](#)






# Predictive Analysis (Classification)

I built, evaluated, improved, and found the best performing classification model by using the below processes.

- Data Preparation:
  - Cleaned and prepared the data by standardizing
  - Handled missing values and encoded categorical variables
  - Split the data into training and testing sets
- Model Selection:
  - Chose an appropriate classification algorithm
  - Considered logistic regression, decision trees, random forests, SVM, and neural networks
- Model Training:
  - Trained the initial model using the training data
  - Adjusted hyperparameters to optimize the model's performance



# Predictive Analysis (Classification)

- Model Evaluation:
  - Evaluated the model using appropriate metrics
  - Considered accuracy using the score method
  - Plotted confusion matrix
- Iterative Process:
  - Iterated through model training, evaluation, and improvement steps
- Grid Search and Cross-Validation:
  - Utilized grid search and cross-validation
  - Systematically searched parameter space and estimated model performance on unseen data
- Model Selection and Final Evaluation:
  - Selected the best performing model based on evaluation metrics
  - Evaluated the model on test data for final performance estimation

GitHub URL - [here](#)

## Predictive Analysis (Classification)

Created a NumPy array from the column Class

Standardized the data in X then reassigned it to the variable X using transform

Used the function `train_test_split` to split the data X and Y into training and test data

Created a logistic regression object then created a `GridSearchCV` object `logreg_cv` with `cv = 10`

Calculated the accuracy on the test data using the method `score`

Repeated the process for decision trees, random forests, SVM, and neural networks



The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

Section 2

# Insights drawn from EDA



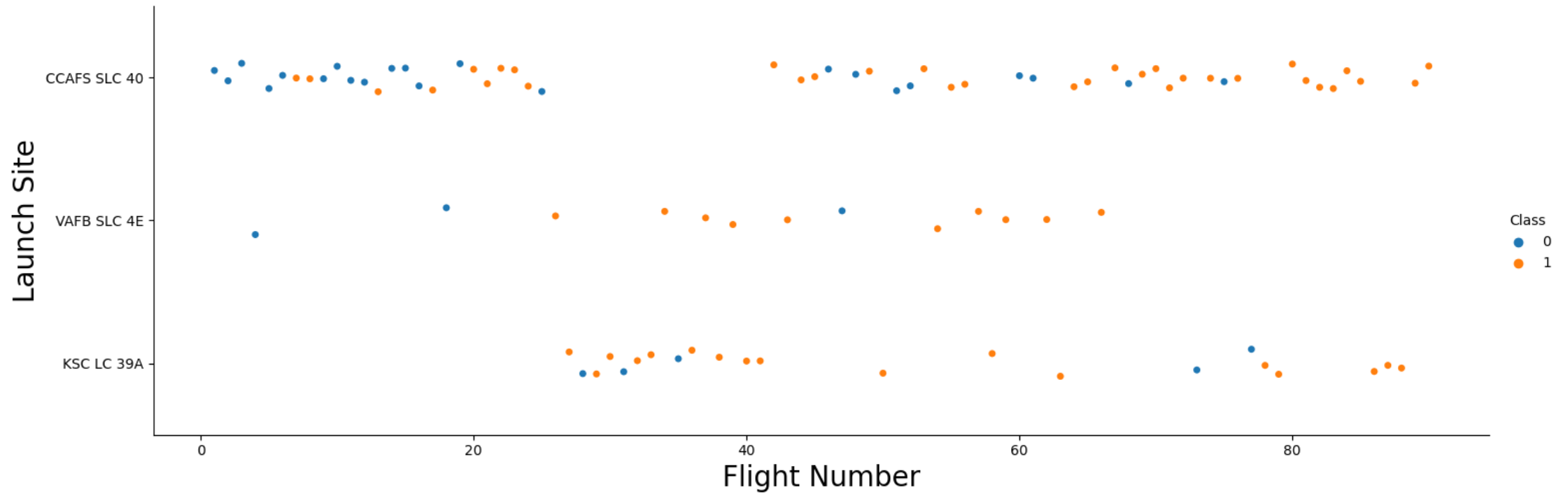


# Exploring Data through Visualizations

In the context of this project, a set of insightful visualizations have been crafted to enhance the understanding of complex data. By leveraging visual representation techniques, these charts bring clarity and depth to my analysis, contributing to a more comprehensive exploration of SpaceX's potential for successful landings.

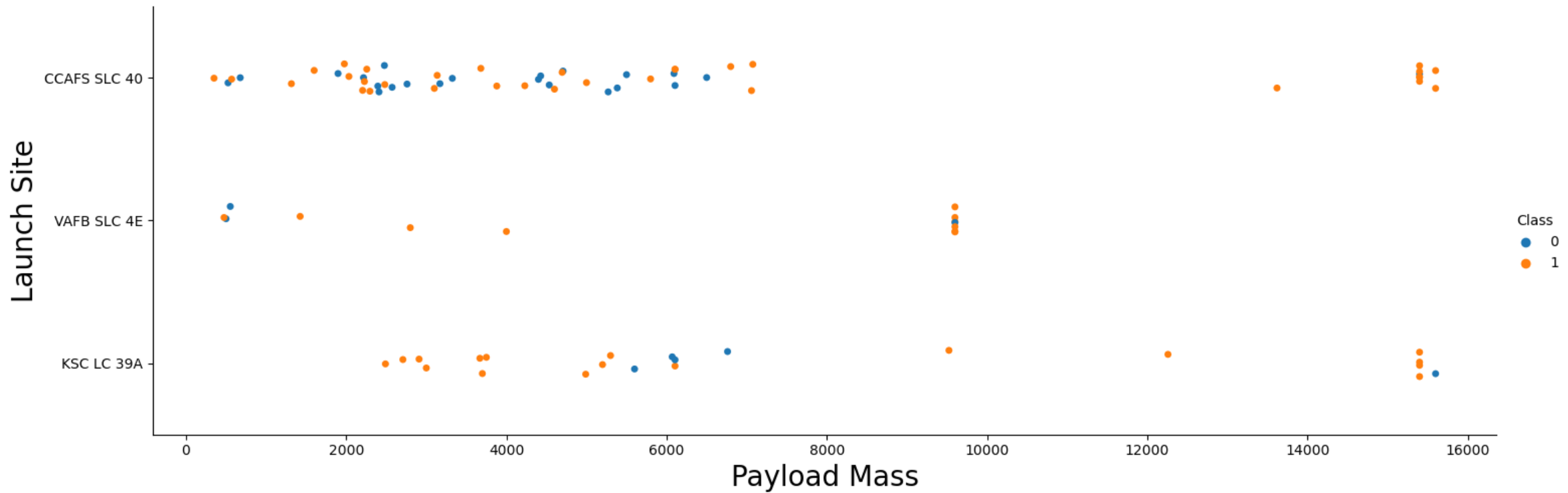
- Scatterplot charts
  - Used to visualize the relationship between;
    - Flight Number and Payload
    - Flight Number and Launch Site
    - Payload and Launch Site
    - Flight Number and Orbit type
    - Payload and Orbit type
- Bar chart
  - Used to visualize the relationship between success rate of each orbit type
- Line chart
  - Used to visualize launch success yearly trend

GitHub URL - [here](#)



## Flight Number vs. Launch Site

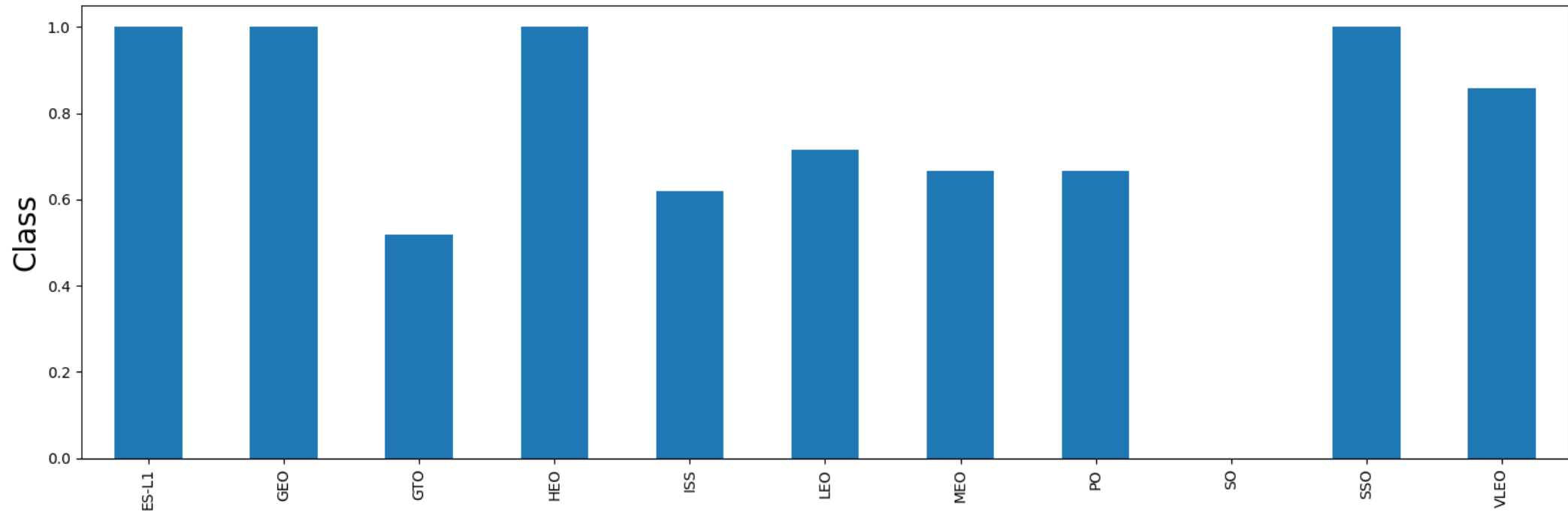
The scatterplot provides a visual representation of the relationship between flight numbers and launch sites. It allows for a comprehensive overview of how different launch sites are associated with specific flight numbers.



## Payload Mass vs. Launch Site

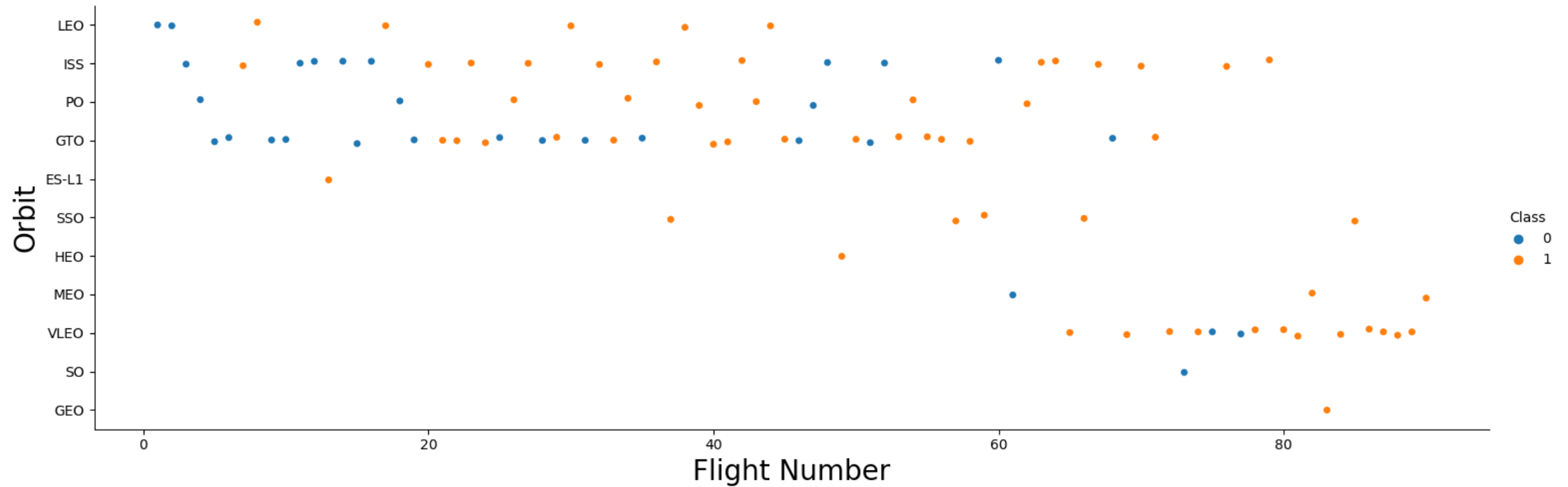
The scatterplot illustrates the relationship between payload mass and launch sites. It visually depicts how different launch sites handle payloads of varying masses. By examining the positioning of data points on the scatterplot, one can gain insights into the distribution and capacity of launch sites based on their ability to accommodate payloads of different sizes. The scatter plot indicates that CCAFS SLC 40 is associated with the highest number of successful launches, particularly in the vicinity of a payload mass of 15,000.





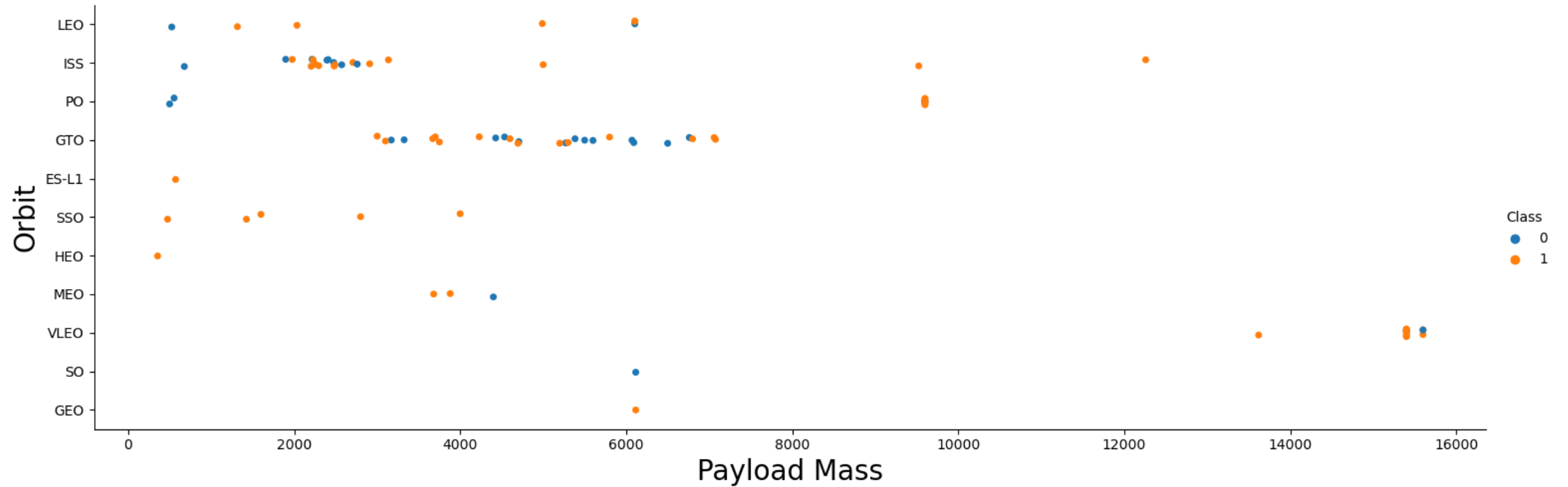
## Success Rate vs. Orbit Type

The bar chart presents a comparison of success rates across different orbit types. Each bar represents a specific orbit type, while the height of the bar corresponds to the corresponding success rate.



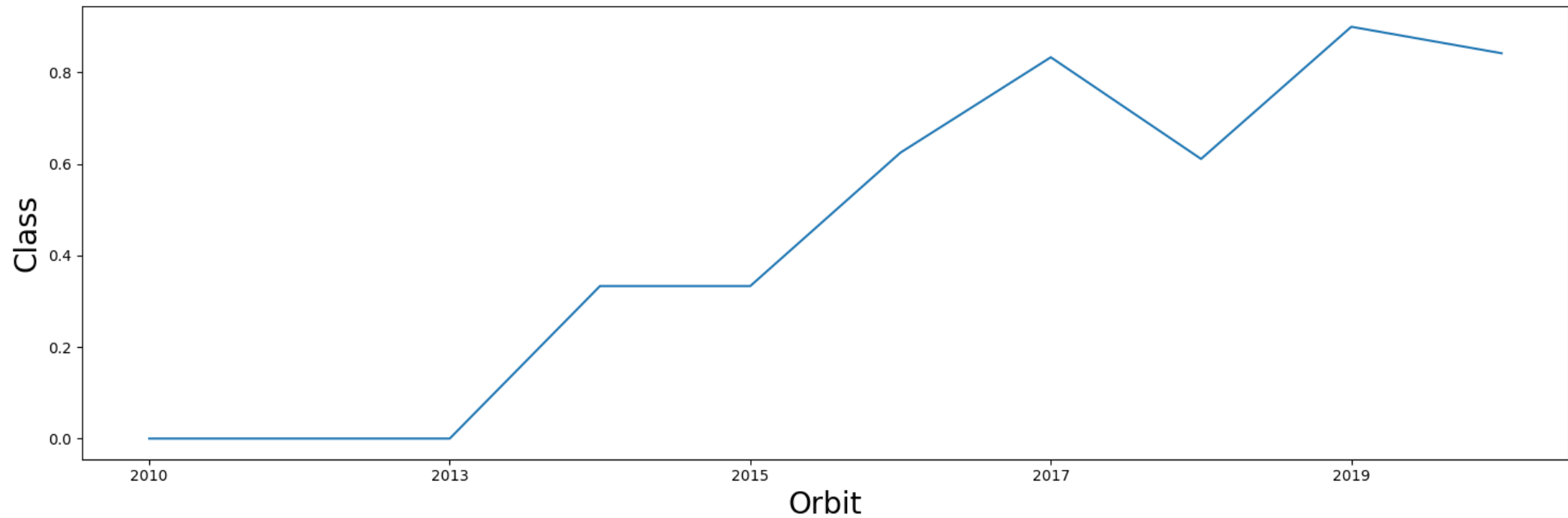
## Flight Number vs. Orbit Type

The scatterplot showcases the relationship between flight numbers and orbit types. By plotting the flight numbers along the x-axis and the corresponding orbit types on the y-axis, it allows for a comprehensive overview of the distribution and clustering of orbits based on flight numbers. The plot clearly demonstrates that the GTO orbit type has achieved the highest success rate among all orbit types.



## Payload Mass vs. Orbit Type

The scatterplot visually depicts the relationship between payload mass and orbit type. It provides valuable insights into how different types of orbits are associated with varying payload masses. Between a mass range of 2000 and 4000, the ISS orbit type emerges as the most successful.



## Launch Success Yearly Trend

The success rate has shown a consistent upward trend since 2013, demonstrating significant growth until 2020. However, there was a minor decline observed in 2018, which was followed by a resurgence in subsequent years. This pattern highlights the overall positive trajectory of success over the years, despite occasional fluctuations.

## Launch Sites

These are all of SpaceX's launch sites located in the United States. These strategically positioned sites enable SpaceX to conduct launches across different regions of the country. By leveraging these launch facilities, SpaceX can efficiently deploy their rockets and payloads, demonstrating their widespread operational capabilities within the United States.

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

## Launch Site Names Beginning with 'CCA'

The presented data consists of five records of launch sites, all of which share a common prefix of "CCA." Surprisingly, within this specific group, none of the launches resulted in successful landings, despite all missions being marked as successful. This observation highlights the significance of analyzing and understanding the intricacies of mission outcomes, as it reveals a potential discrepancy between overall mission success and the specific aspect of landing success in this particular subset of launch sites.

# Total Payload Mass

The collective payload transported by NASA boosters amounted to an impressive 45,596 kilograms.

This figure represents the cumulative mass of payloads delivered through NASA's booster missions, demonstrating the significant capacity and capability of these launch systems. The substantial payload weight underscores the importance of NASA's contribution to various space exploration and scientific endeavors.



# Average Payload Mass by F9 v1.1

---

Booster version F9 v1.1 has a noteworthy average payload mass of 2928.4 kilograms.

This indicates the capacity and capability of this specific booster version to reliably transport payloads of considerable weight. The average payload mass serves as a valuable metric for assessing the performance and efficiency of F9 v1.1 in delivering payloads to their intended destinations.



# First Successful Ground Landing Date

On July 22, 2018, a significant milestone was achieved as the first successful landing on a ground pad took place. This achievement marked a pivotal moment in the history of landing capabilities for the respective launch system. The successful landing outcome on the ground pad demonstrated the technological prowess and advancement in vertical landing capabilities.

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

## Successful Drone Ship Landing with Payload Mass between 4000 and 6000

The following are the names of boosters that have achieved successful landings on a drone ship, while also carrying a payload mass between 4000 and 6000. These boosters stand as prime examples of successful and precise landing operations, showcasing the ability to safely return to a drone ship platform while transporting payloads within a specific mass range.



MISSION_OUTCOME	TOT_NUM
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

### Total Number of Successful and Failure Mission Outcomes

The table presents the total count of both successful and failed mission outcomes, with a notable occurrence of 100 successful missions. This significant number underscores the effectiveness and accomplishments of these missions, reflecting a high success rate in achieving their objectives.

By examining the provided data, it becomes evident that the majority of the recorded mission outcomes resulted in successful achievements.

# Boosters Carrying Maximum Payload

The following boosters are notable for carrying the maximum payload mass. These boosters demonstrate the remarkable capacity and capability to transport payloads of significant weight, showcasing their exceptional performance in accommodating heavy payloads. Their ability to handle maximum payload masses reflects the effectiveness and efficiency of these boosters in meeting the demands of space missions requiring large payload capacities.

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

Month	Landing Outcome	Booster Version	Launch Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The displayed query result provides information on failed landing outcomes on a drone ship in the year 2015. The data includes details such as the respective booster versions and launch site names associated with these unsuccessful landing attempts.

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

Landing_Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	7

The presented data showcases the count of landing outcomes, categorized as either "Failure" on a drone ship or "Success" on a ground pad, within the time frame of June 4, 2010, to March 20, 2017. The results are arranged in descending order, providing a comprehensive overview of the frequency of these landing outcomes during the specified period.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible as a thin, curved line separating the dark surface from the deep blue of space.

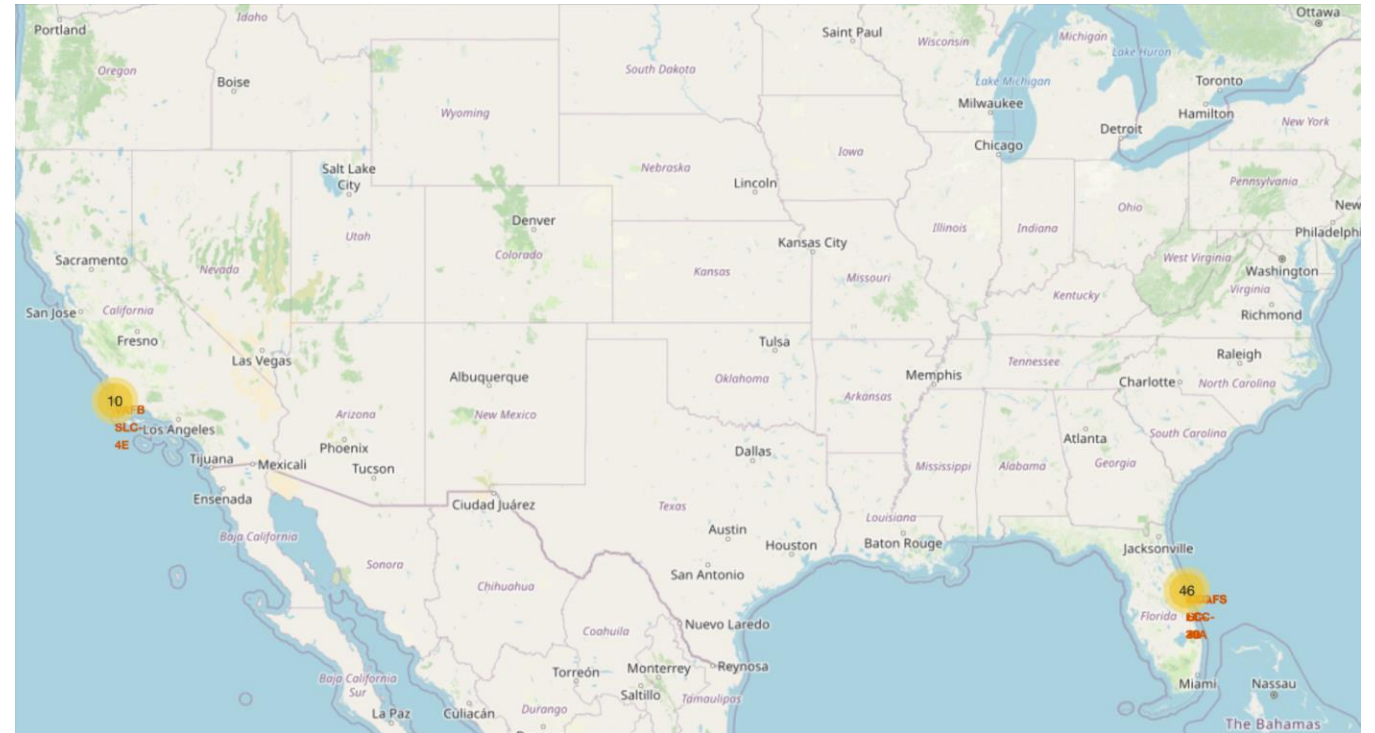
Section 3

# Launch Sites Proximities Analysis



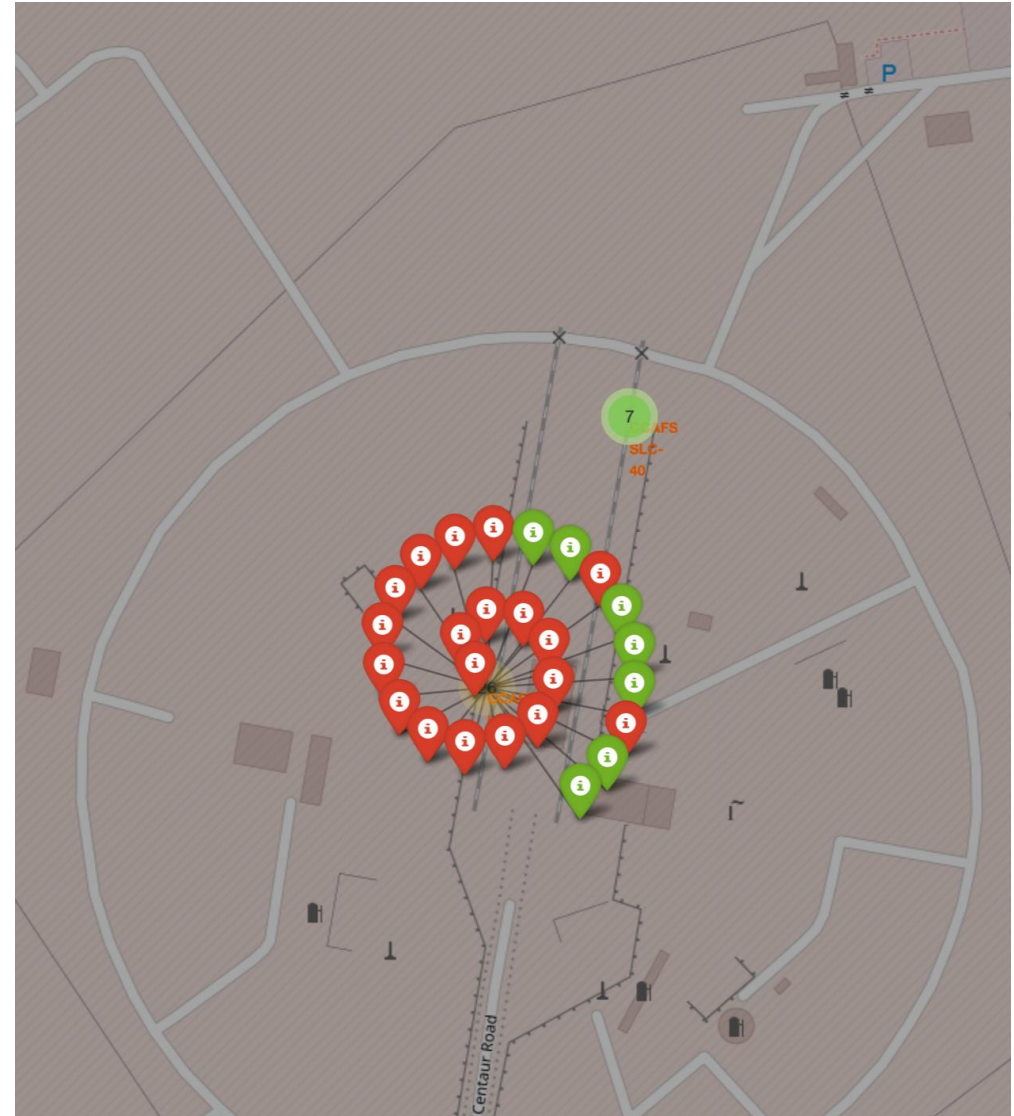
# Launch Sites

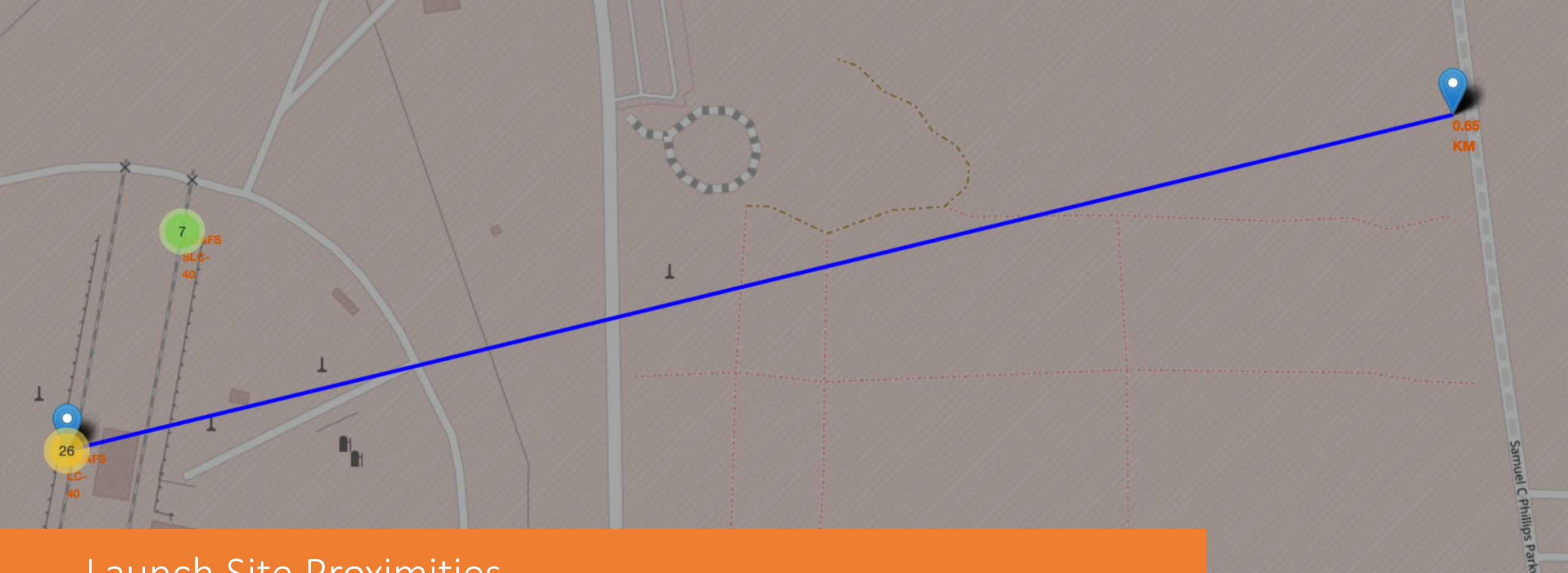
All of SpaceX's launch sites are strategically positioned in proximity to the east and west coastlines of the United States. This deliberate choice allows for efficient access to desired launch trajectories and facilitates launches over open bodies of water, minimizing risks to populated areas. The coastal locations of these launch sites provide logistical advantages for transport and deployment, supporting SpaceX's mission to deliver payloads to space with enhanced safety and efficiency.



# Launch Outcomes

An illustrative example is presented with the launch site "CCAFS LC-40" situated on the east coast of the United States in Florida. The map showcases red markers representing failed launches and green markers representing successful launches. Notably, for this specific launch site, there is a higher count of failed launches compared to successful ones, indicating a potential area of improvement and focus for future launch operations.





## Launch Site Proximities

The CCAFS LC-40 launch site is located in close proximity to the nearest highway, with a calculated distance of approximately 0.65 kilometers. This information highlights the advantageous accessibility of the launch site to a major transportation route, facilitating logistical operations and transportation of equipment and personnel. The close proximity to the highway ensures efficient connectivity and potential time-saving benefits for the launch site activities.





Section 4

# Build a Dashboard with Plotly Dash

# SpaceX Launch Records Dashboard

All Sites

×

▼

Total Launches for All Sites



Launch  
Success for All  
Sites

Among all the launch sites, KSC LC-39A stands out with the highest number of launches recorded, highlighting its significance and extensive usage. In contrast, CCAFS has had the least number of launches among the various sites, indicating relatively lower activity in terms of launches conducted from that specific location.

# SpaceX Launch Records Dashboard

KSC LC-39A

×

▼

Total Success Launches for site KSC LC-39A



Launch Site with  
Highest Launch  
Success Ratio

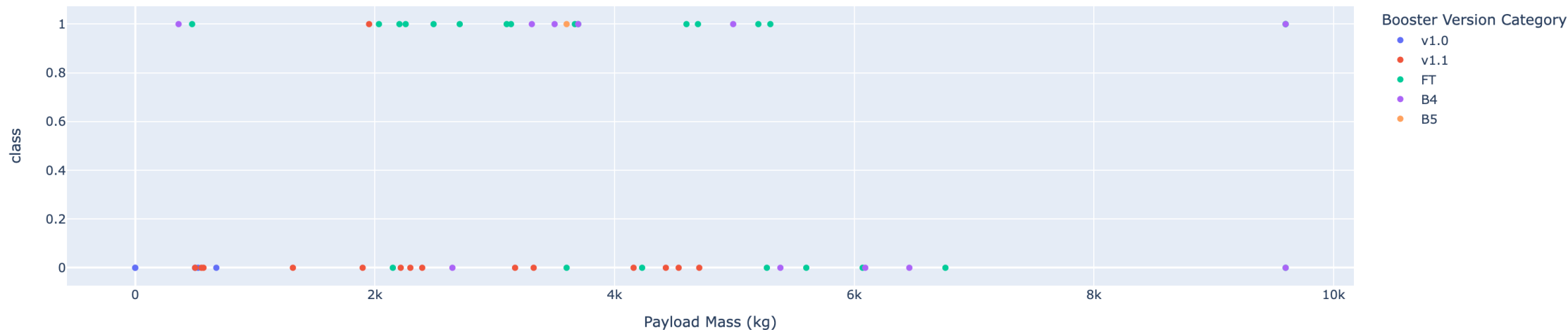
In addition to having the most launches, KSC LC-39A boasts the highest launch success ratio among all the launch sites. This noteworthy achievement underscores the reliability and effectiveness of launch operations conducted from this site. The high launch success ratio highlights KSC LC-39A as a premier location for achieving mission objectives with a greater degree of success compared to other launch sites.



Payload range (Kg):



Success count on Payload mass for all sites



# Payload vs. Launch Outcome for All Sites

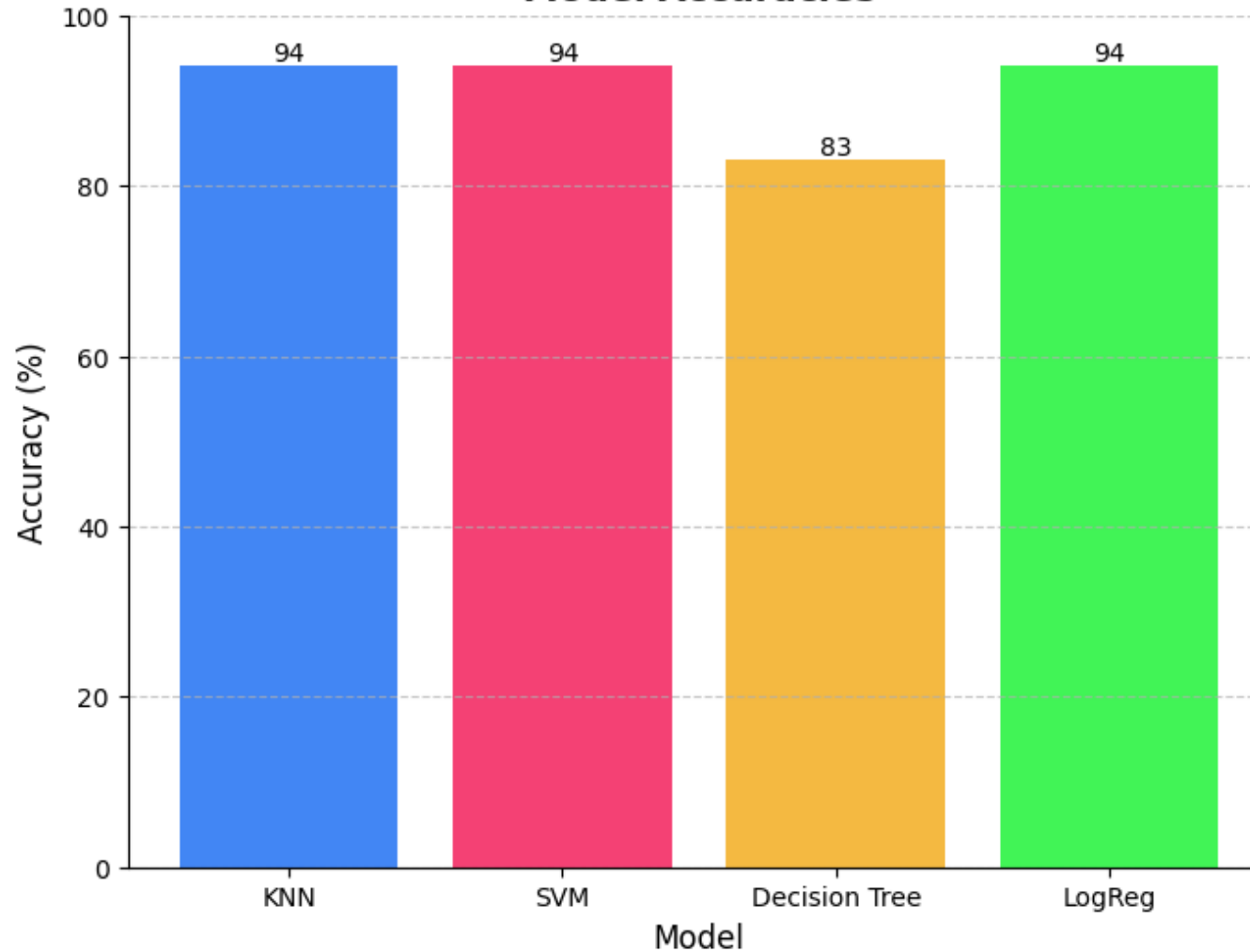
The scatterplot visually presents the correlation between payload mass and mission outcomes, offering a holistic representation. With payload mass on the x-axis and mission outcomes on the y-axis, the scatterplot enables the identification of patterns and trends. Notably, the data reveals that the range of payload masses between 2000 and 4000 kg exhibits the highest rate of mission success, emphasizing the significance of this payload mass range in achieving favorable mission outcomes.



Section 5

# Predictive Analysis (Classification)

**Model Accuracies**

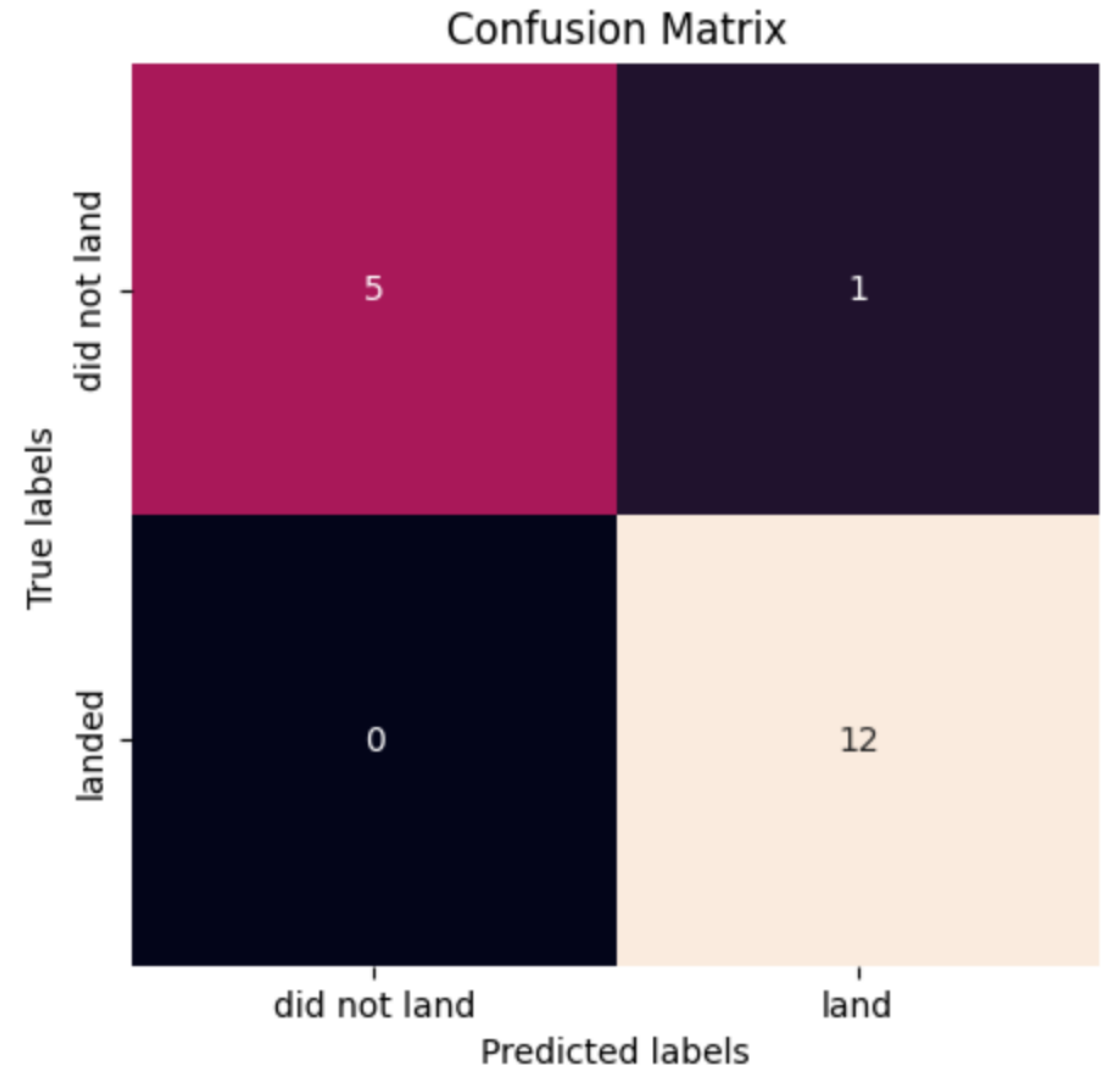


## Classification Accuracy

Among the models employed, namely K Nearest Neighbors, SVM, Logistic Regression, and Decision Tree, it is worth noting that all models, except the Decision Tree, achieved an impressive accuracy score of 94%. However, the Decision Tree model performed slightly lower with an accuracy rate of 83%. These results emphasize the strong performance of the K Nearest Neighbors, SVM, and Logistic Regression models in accurately predicting outcomes compared to the Decision Tree model.

# Confusion Matrix

Among the models utilized, the K Nearest Neighbors, SVM, and logistic regression models stand out as the top performers with an impressive score of 94%. In comparison to the other model employed, namely decision tree, the mentioned models demonstrate superior performance in accurately predicting outcomes.



# Conclusions



Based on the analysis conducted using K Nearest Neighbors, SVM, Logistic Regression, and Decision Tree models, it is evident that these models performed exceptionally well in accurately predicting the likelihood of first stage reuse. The high accuracy rates achieved by these models instill confidence in their predictive capabilities, enabling more informed decision-making in determining the viability of first stage reuse.



Furthermore, the data reveals that payload masses between 2000 and 4000 kg have the highest rate of mission success. This finding highlights the significance of considering payload mass within this range when assessing the likelihood of successful missions and potential first stage reuse.



KSC LC-39A, with its remarkable launch success ratio and the most launches among all sites, emerges as a key location for achieving successful launches and potential first stage reuse. The strategic positioning of all SpaceX launch sites along the east and west coastlines of the United States further enhances accessibility and operational efficiency, facilitating the implementation of first stage reuse strategies.



Moreover, the average payload mass carried by booster version F9 v1.1 is approximately 2928.4 kg. This data point provides valuable insights into the capacity and capabilities of this booster version, further informing the assessment of its potential for supporting first stage reuse.



Lastly, within the specified mass range of 2000 to 4000 kg, the ISS orbit type demonstrates the highest success rate. This observation emphasizes the reliability and effectiveness of the ISS orbit type for successful missions and increases the likelihood of achieving first stage reuse for payloads falling within this mass range.



In conclusion, considering the accuracy rates of the models, the relationship between payload mass and mission success, launch site performance, strategic positioning, booster capabilities, and successful outcomes of the ISS orbit type, a comprehensive evaluation can be made to predict the likelihood of first stage reuse, enabling informed decision-making in space mission planning and operations.



Thank you!

