Winning Space Race with Data Science

Kaleem U Allah
05/16/2024

Outline

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

Executive Summary

- The process consisted of many critical components. First, we completely grasped the difficulty of forecasting the successful landing of the Falcon 9 first stage to estimate launch costs. We then compiled a thorough dataset utilizing SpaceX's API and web scraping tools. The obtained data proceeded through an understanding step (exploratory data analysis), in which visualization and descriptive statistics helped us comprehend the correlations between various aspects, and feature engineering assisted us in picking significant features for our research. Moving on to the preparation step, the data has been cleansed and preprocessed to ensure its quality and usefulness. We created and evaluated a variety of machine learning models, including logistic regression, decision trees, and support vector machines, with cross-validation approaches used to avoid overfitting. The models were tested using several measures.
- The decision tree model emerged as the best performer, providing reliable predictions of landing success.

Introduction

- In the competitive landscape of the aerospace industry, the ability to reduce launch costs is a significant advantage. SpaceX has revolutionized this domain by successfully reusing the first stage of its Falcon 9 rockets, bringing down the cost of launches to 62 million dollars compared to the industry standard of 165 million dollars. This cost-effectiveness is primarily due to the successful landing and reuse of the first stage, a critical component of their operational model.
- The focus of this project is to predict the likelihood of a Falcon 9 first-stage landing successfully. Accurate predictions can provide essential insights into launch costs, aiding alternative companies in crafting competitive bids against SpaceX. By leveraging data science techniques, this project aims to understand the factors influencing landing success and develop predictive models to forecast this outcome reliably.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping techniques with Python's beautiful soup, requests modules, and pandas library
- Perform data wrangling
  - Data has been filtered out to only keep those related to Falcon 9. Then several information from different sources were aggregated to form a unique data frame where NaN values were dealt with. Feature engineering also helped drop irrelevant fields for further analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We used the machine learning library scikit-learn to build several classification models. Those models have been evaluated to keep the best performing.

Data Collection

The data collection process primarily involved extracting comprehensive datasets related to all SpaceX launches from the SpaceX API. Using Python's requests library, we made requests to gather detailed information on launch dates, payload mass, launch sites, booster versions, landing outcomes... Further, we utilize a Wikipedia web page with web scraping techniques to extract more useful data on Falcon 9 launches. The collected data was then processed to ensure consistency and stored both in CSV files and SQLite 3 databases.

Data Collection – SpaceX API

- Initial data collected from https://api.spacexdata.com/v4/rockets/
- Data has been parsed into DataFrame with pandas (Data content type: Json)
- Based on the IDs obtained from the first API call, call: https://api.spacexdata.com/v4/launchpads/ to obtain launchpads names and coordinates
- https://api.spacexdata.com/v4/payloads/ to obtain information about the payload (orbit, mass)
- https://api.spacexdata.com/v4/cores/ to obtain some useful information about the launch including its outcome
- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%201/spacex-data-collection-api.ipynb

Data Collection - Scraping

- Get request to`https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922`related to SpaceX's Falcon 9 launches
- •Response's text content from the website parsed with BeautifulSoup
- •All tables on the webpage has been extracted with query on the parsed html Response
- •The table of interest has been selected
- •A DataFrame has been created based on the table content

- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%201/webscraping.ipynb

Data Wrangling

- After being collected, a DataFrame has been built upon the collected Data
- Only relevant columns for our analysis have been kept
- All rows on non-Falcon 9 launches have been filtered out
- Missing values have been replaced by the mean of their columns
- Categorical fields are encoded using One-hot encoding
- An extra column called ‘Class’ is added. It equals 0 if a given launch was a failure and 1 if it was successful.
- The final dataframe was cast float type
- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%201/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Several plots were made to get how each variables are related to each other and to the target variable :
- Scatter plot: it helped us visualize the correlation between different variables and the target variable (the launch outcome); say otherwise, it helped identify how different features influence or impact the target variable and how they are correlated to each other.
- Bar chart: helped us gain quantitative information on categorical variables; for instance, It helped us visualize the success rate of each orbit
- Line plot: it helped us to visualize how a specific variable fluctuates or trends over time; we used it for instance to visualize how the success rate trends over years

- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%202/eda-dataviz.ipynb

EDA with SQL

SQL queries are used to perform the following EDA tasks :

- Display the names of unique launch sites in the space missions
- Display launch site records where the names begin with CCA
- Display the total payload mass carried by boosters launched by NASA (CRS) : 45 596 Kg
- Display average payload mass carried by booster version F9 v1.1: 2 928.4 Kg
- Get the date of the first successful landing in the ground pad: 2015-12-22
- List the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failed mission outcomes
- List the total number of successful and failed mission outcomes
- 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

https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%202/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

With Folium, different map objects object have been created :

- Circle: to highlight a specific area
- Markers: to mark a specific site and display its name when printed on it
- Markers Cluster: to group several related markers with the ability to quickly identify which sites have the highest rate of success outcomes
- Mouse Position: to quickly get coordinates of a place of interest; for instance, it helped us get the coordinates of some coastline points, highways, railways...
- Poly Line: to visualize the distance between a site and a coastline point or some place of interest to state how close the launch site is to those places.

- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%203/launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The following elements have been added to the dashboard :

- A pie chart to visualize the success and the failure rate of a selected site and all sites
- A dropdown input component to make the site selection
- A scatter point chart to visualize the correlation between the payload mass and the success for all sites and a specific site
- A range slider to select a payload range

- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%203/spacex_dash_app.py

Predictive Analysis (Classification)

Given the classification problem, we have built and evaluated several classification models following the steps :

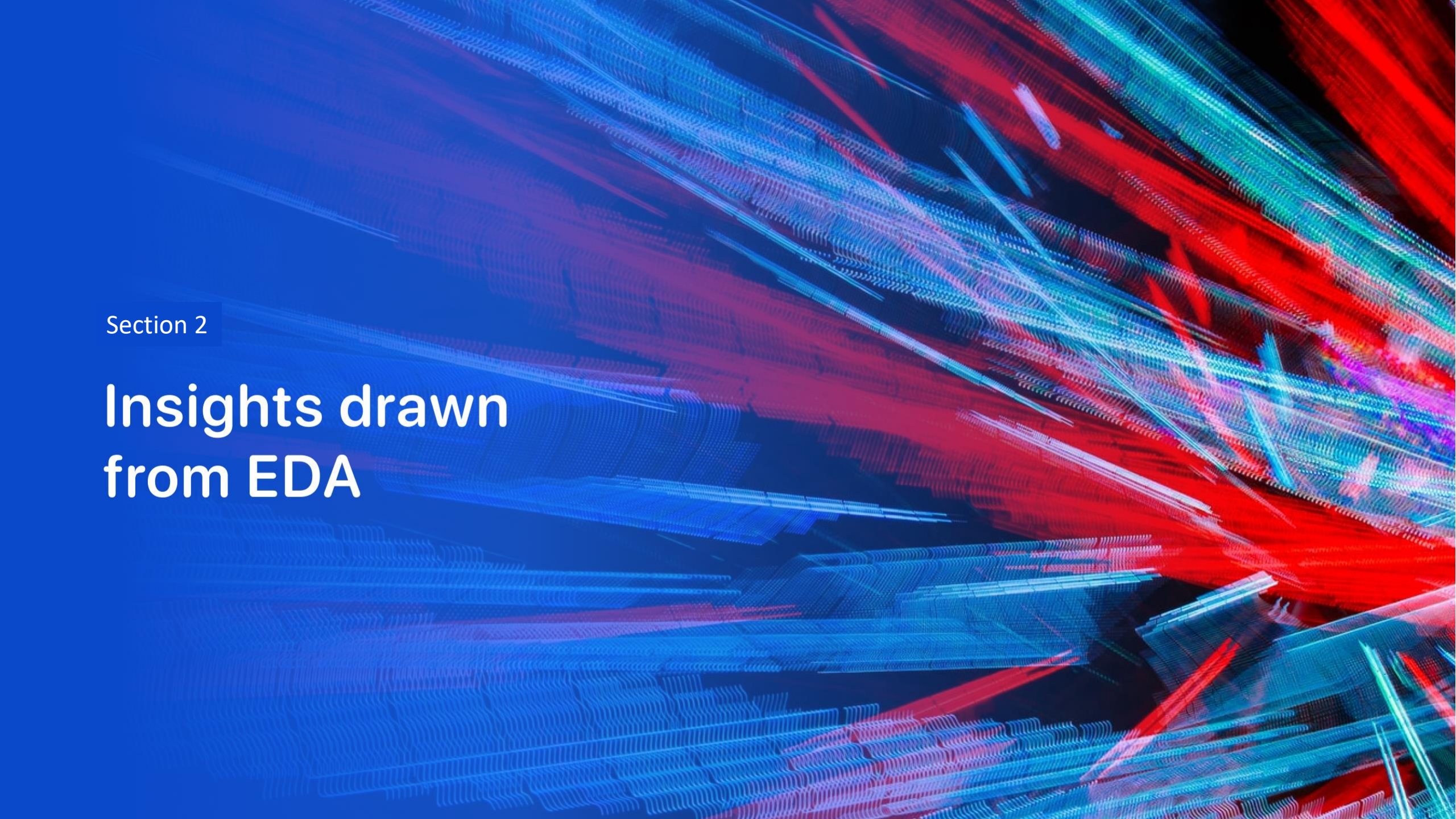
- We first standardize the dataset using a standard scaler
- Split the dataset into train and test sets to prevent overfitting
- Select a given model and a set of hyperparameters
- Build a model using gridsearchcv and cross-validation to select the best parameters
- Train the model on the train test
- Evaluate the model on the test set

- https://github.com/KaleemuAllah/IBM-Applied-DS-Capstone/blob/main/Week%204/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

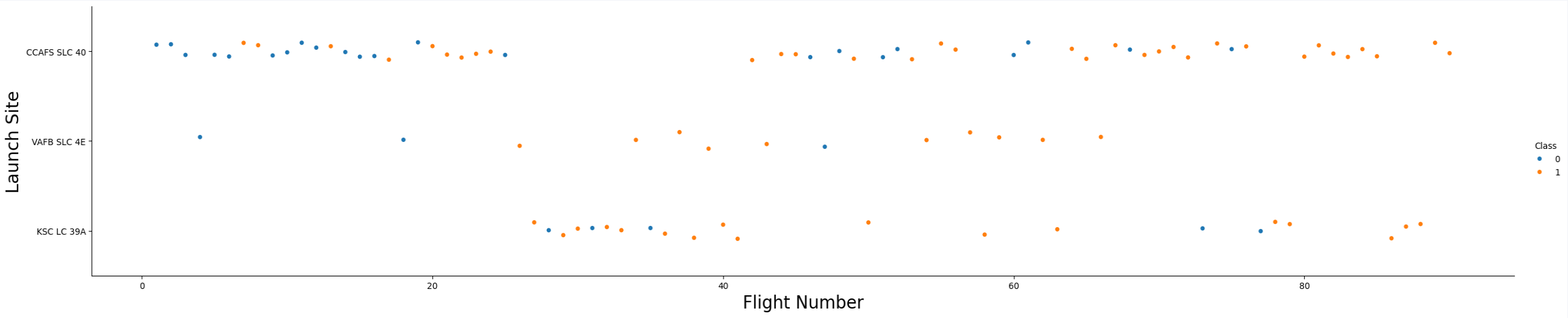
Results

The results section will fall into three sub-sections :

- Exploratory data analysis insights with SQL and static visualizations
- Interactive analytics
- Predictive analysis results

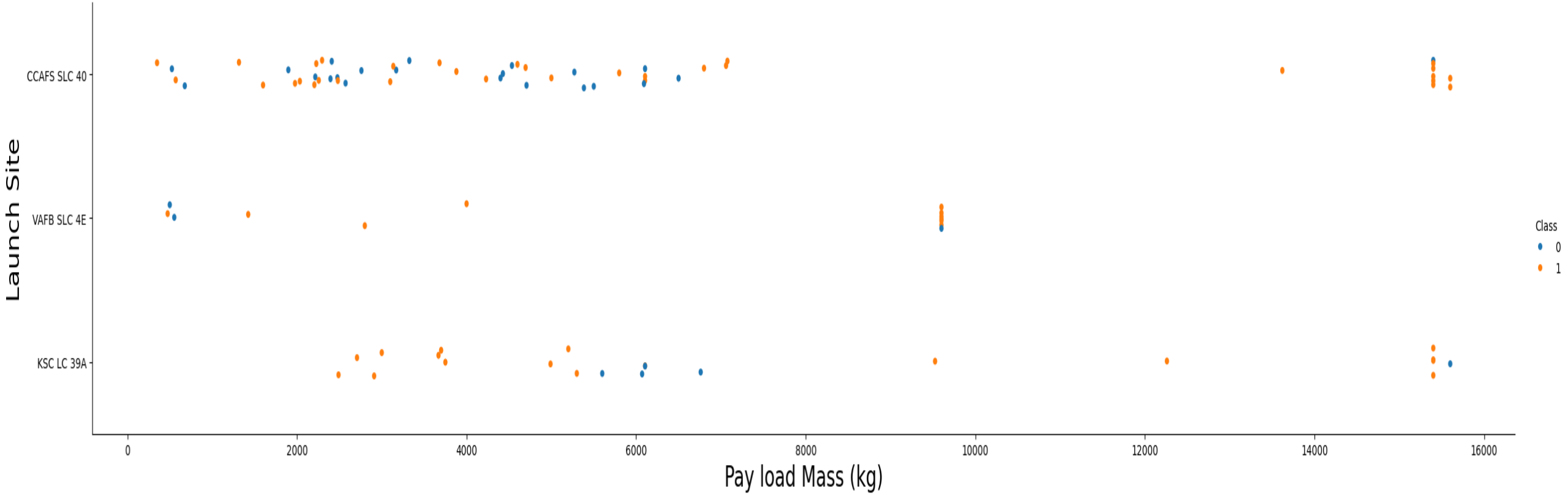

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

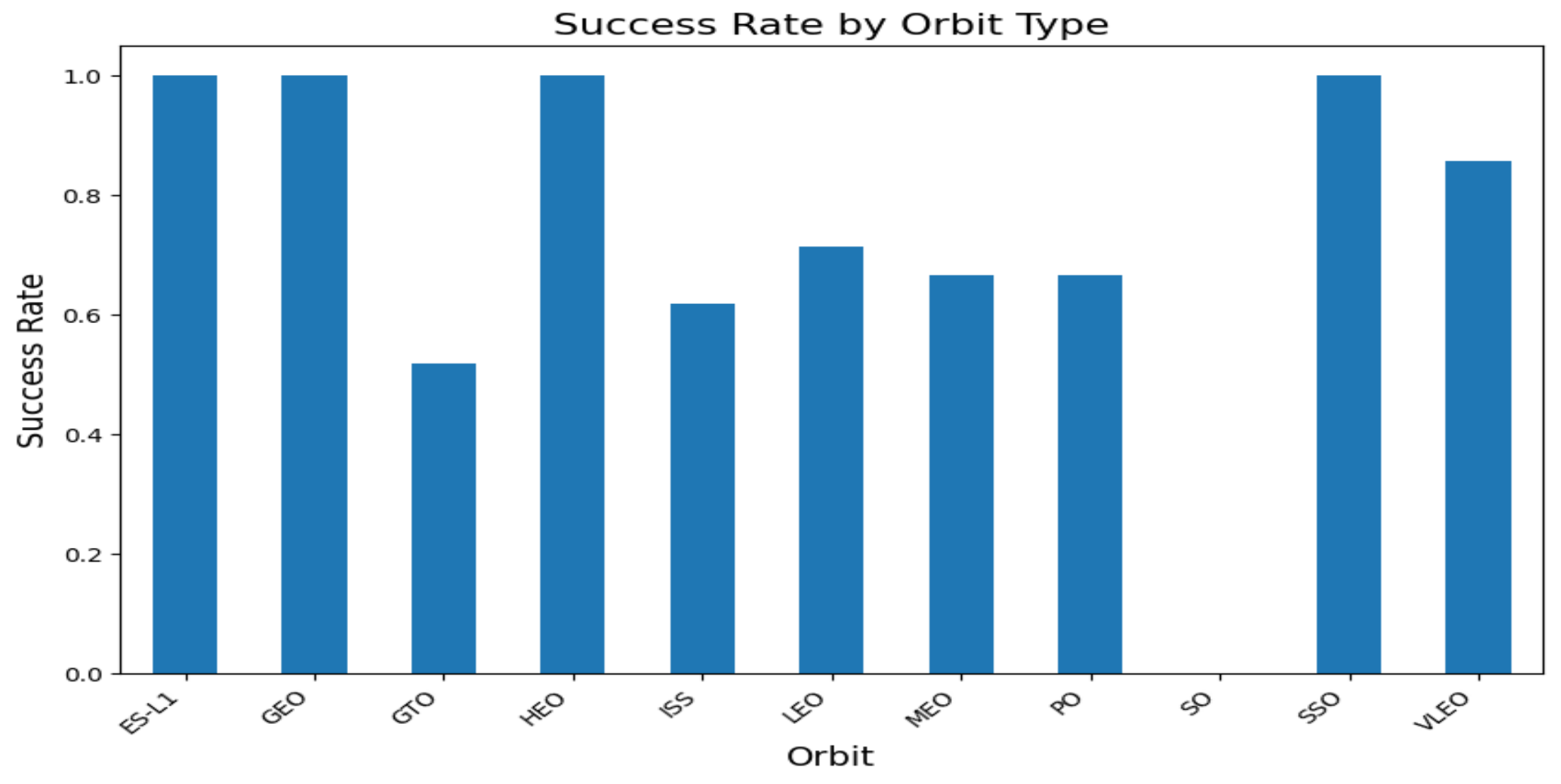


We see that as the flight number increases, the first stage is more likely to land successfully. The launch site is also important; it seems the more massive, the less likely the first stage will return.

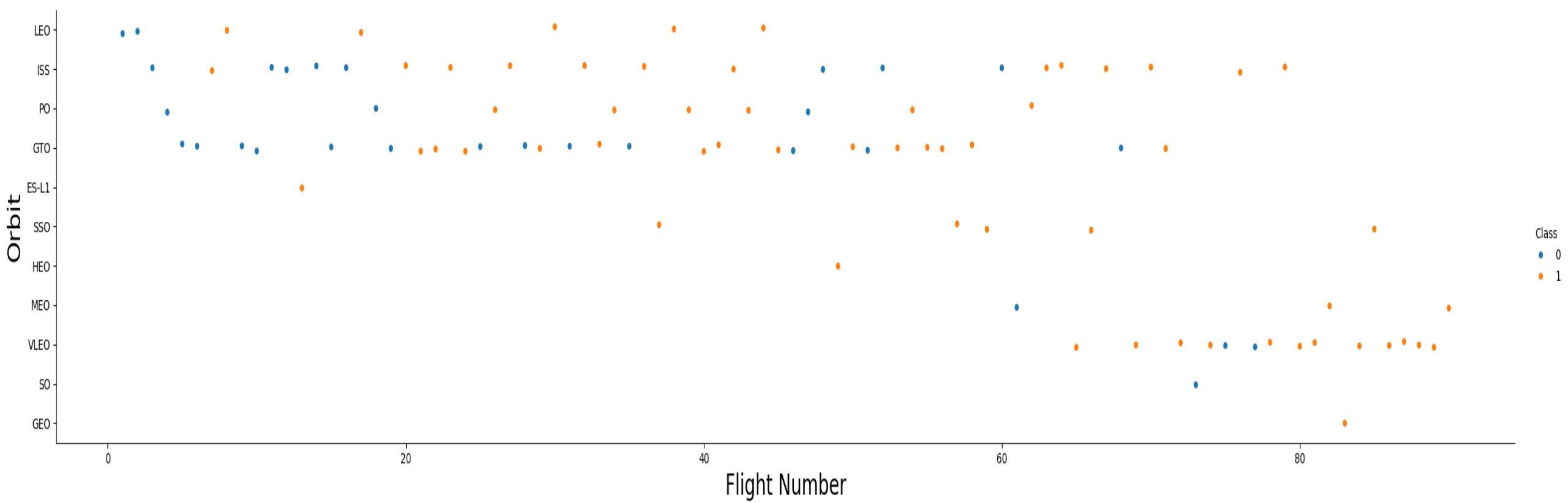
Payload vs. Launch Site



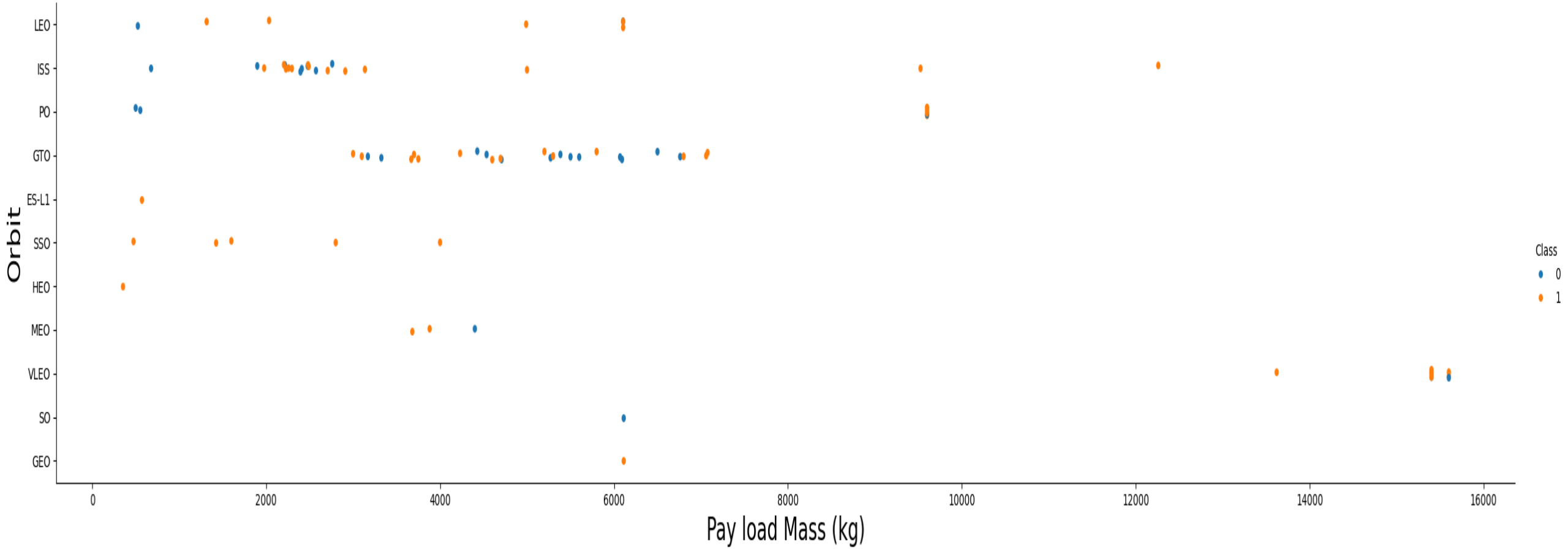
Success Rate vs. Orbit Type



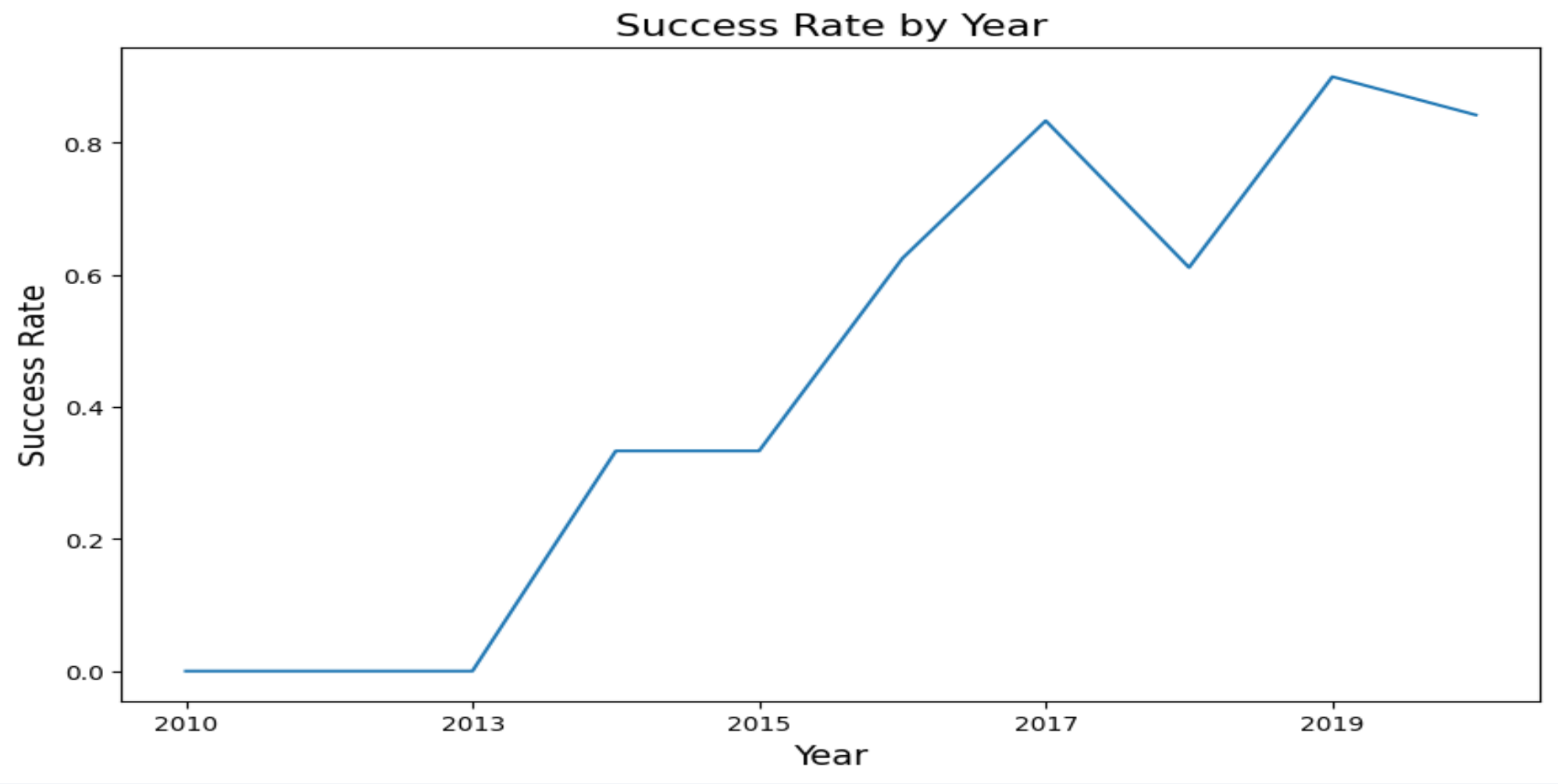
Flight Number vs. Orbit Type



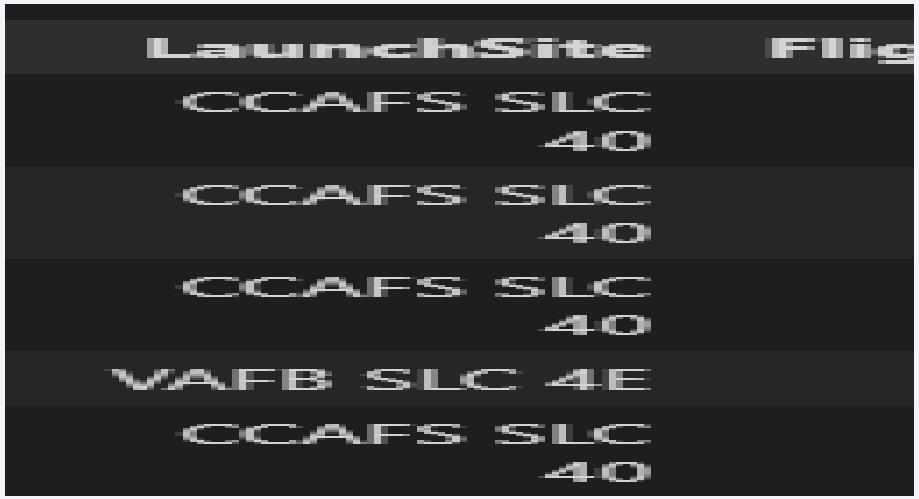
Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

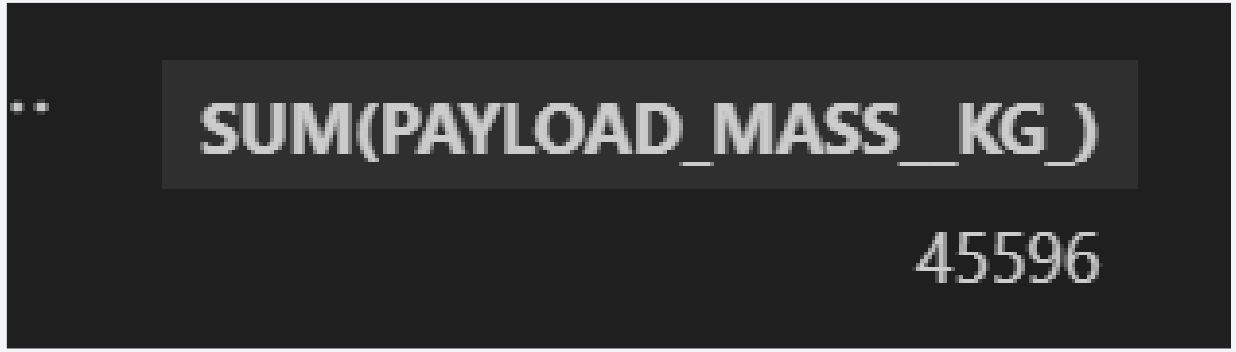


LaunchSite	Flig
CCAFS SLC 40	
CCAFS SLC 40	
CCAFS SLC 40	
VAFB SLC 4E	
CCAFS SLC 40	

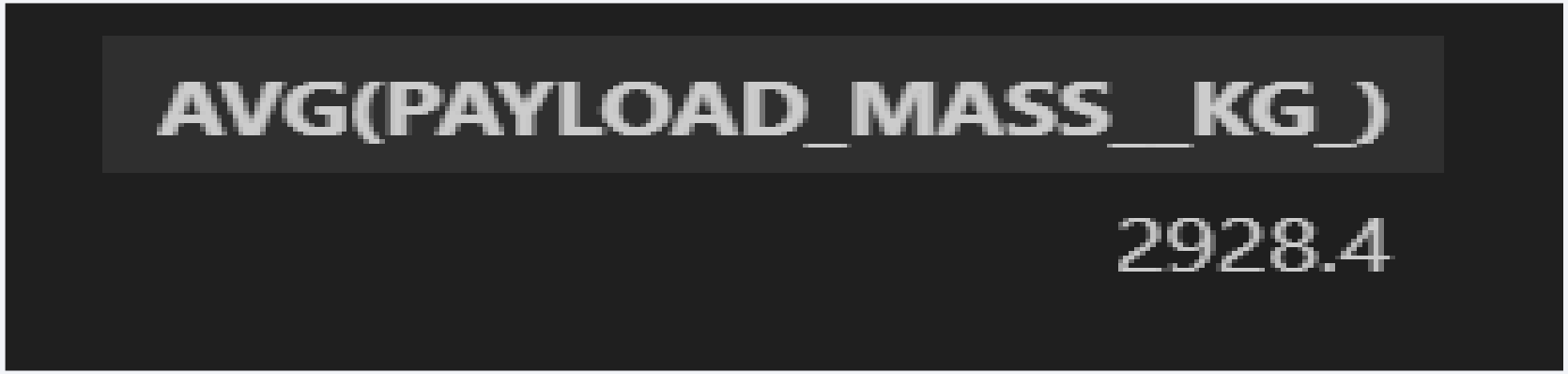
Launch Site Names Begin with 'CCA'

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

Total Payload Mass



Average Payload Mass by F9 v1.1



First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	total_outcome
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Booster Version With max payload mass
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
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

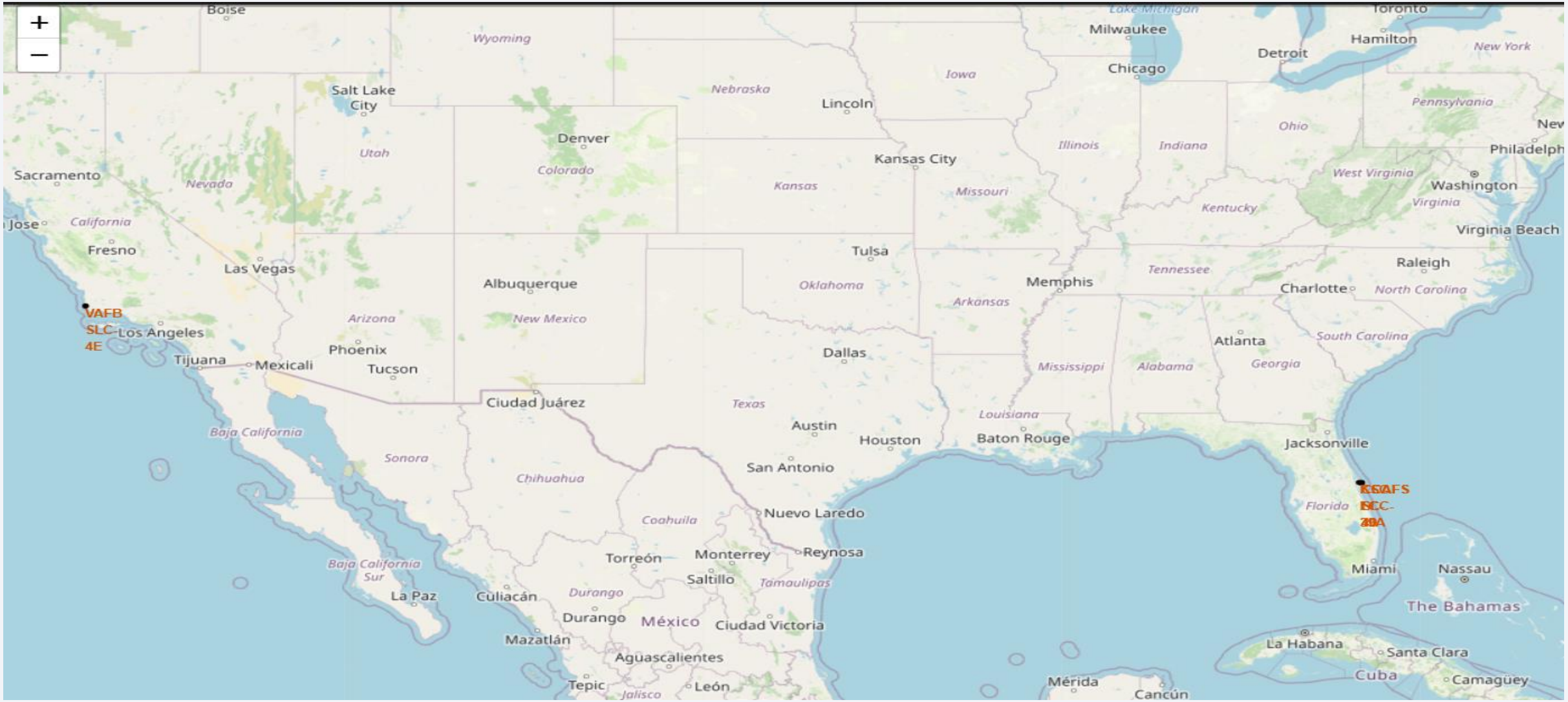
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

Section 3

Launch Sites Proximities Analysis

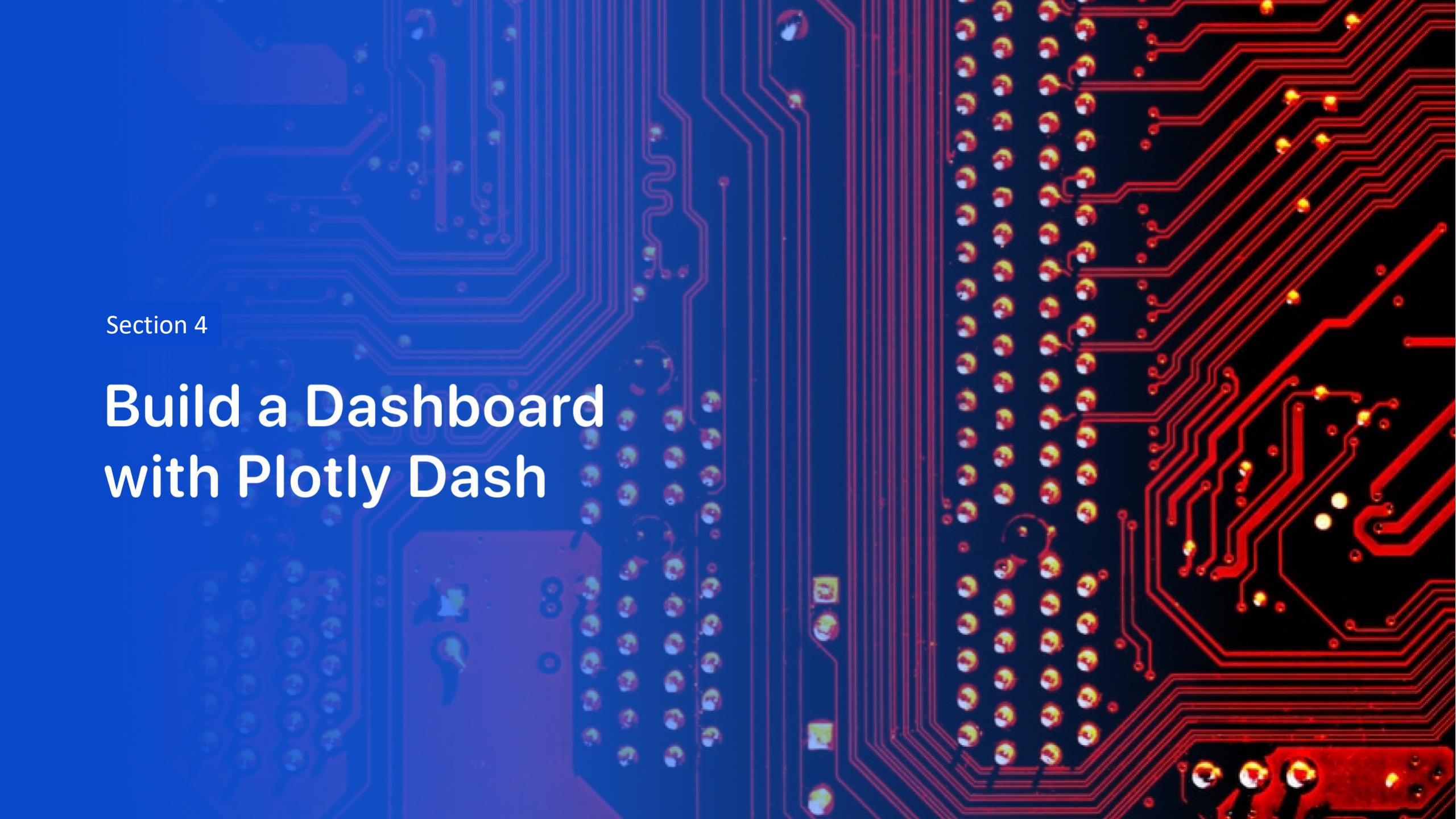
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

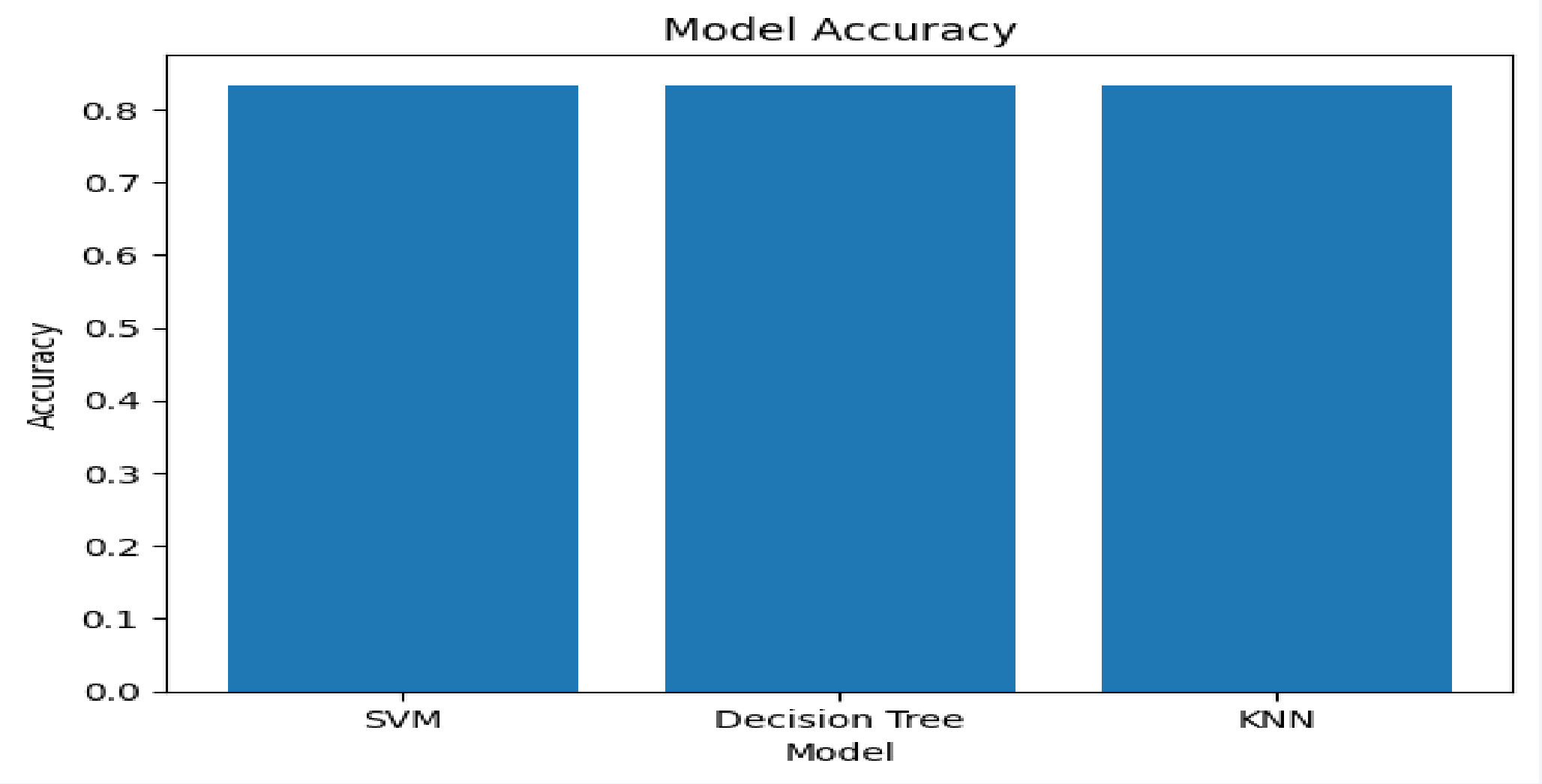
<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

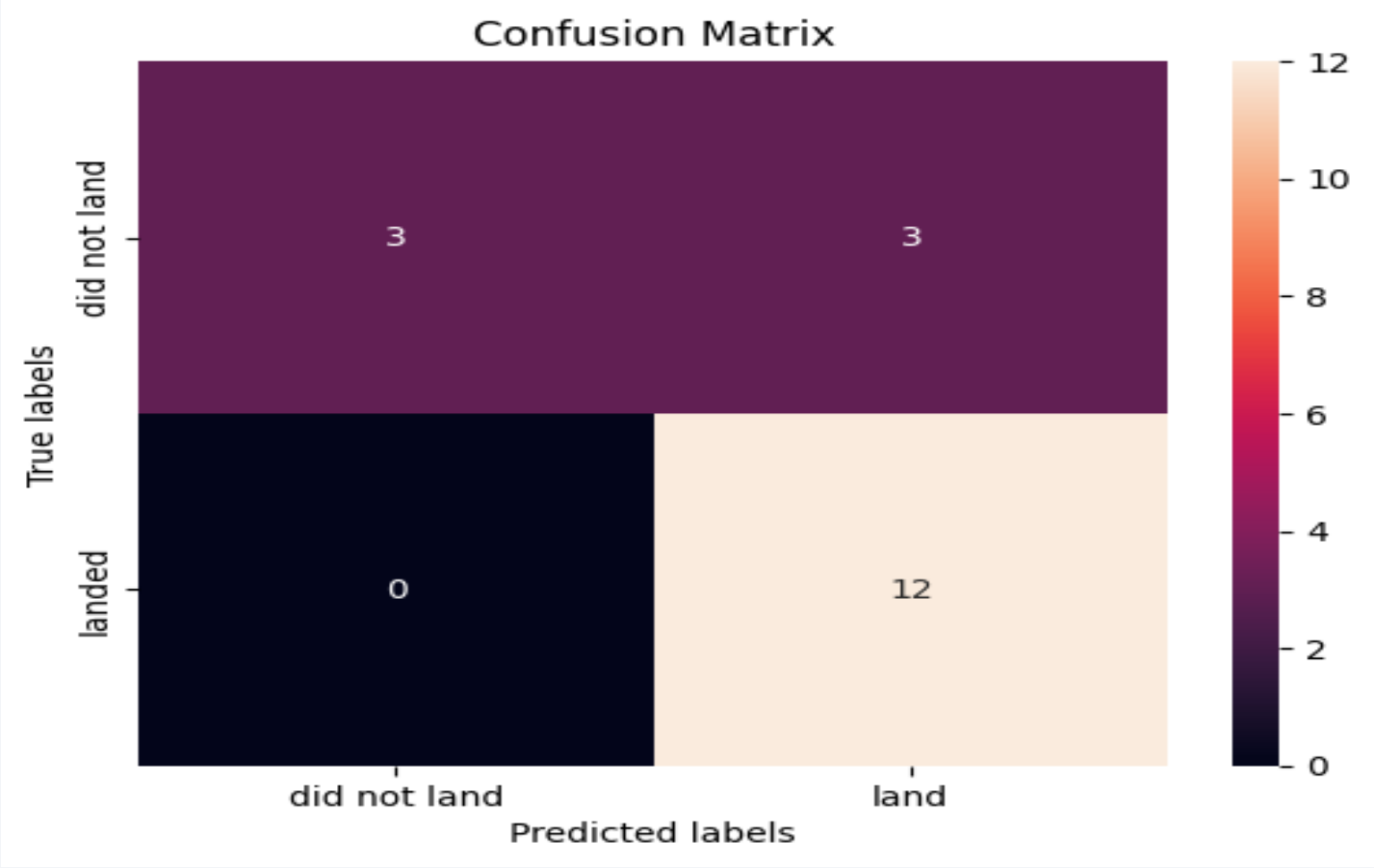
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- The objective of this project was to predict the successful landing of the Falcon 9 first stage, a critical factor in SpaceX's cost-effective launch operations.
- •Each feature of the Falcon 9 launch, such as the payload mass, the launch site, or the orbit type, may affect the mission outcome.
- •We collected comprehensive datasets exclusively from the SpaceX API and Wikipedia, which provided detailed information on launch parameters such as launch dates, payload mass, launch sites, booster versions, and landing outcomes...
- •During the exploratory data analysis (EDA) phase, we examined the collected data to uncover patterns, trends, and relationships among the variables. We identified significant features influencing landing success, such as payload mass, and launch site ...
- •We developed and tested several machine learning models to predict the successful landing of the Falcon 9 first stage. The models included logistic regression, support vector machine, decision tree, and K nearest neighbor. They all came out with the same score and the same confusion matrix during the evaluation phase. We finally keep the best model as the one with the highest accuracy during the cross-validation: the decision tree.

Appendix

Thank you!