IBM Developer
SKILLS NETWORK

Applied data science capstone

Winning Space Race with Data Science

Project Scanerio: SpaceX Falcon 9 First Stage Landing Prediction

OLUM ABEL
NOVEMBER 2024

Outline

- Executive Summary
- Table of Contents
- Introduction
- Methodology
- Result
- Conclusion
- Appendix

Executive Summary

This research aims to identify the key factors that contribute to the success of rocket landings. To achieve this, several methodologies were employed:

- **Data Collection**: Data was gathered through SpaceX's REST API and web scraping techniques.
- **Data Wrangling**: The dataset was processed to create a binary outcome variable indicating success or failure of landings.
- **Exploratory Data Analysis (EDA)**: Data visualization techniques were used to explore variables such as payload, launch site, flight number, and yearly trends.
- **Statistical Analysis**: SQL was used to calculate key statistics, including total payload, payload range for successful launches, and the overall count of successful and failed outcomes.
- **Launch Site Analysis**: The success rates of different launch sites were examined, along with their proximity to significant geographical markers.
- **Visualization**: The most successful launch sites and their corresponding payload ranges were visualized to identify patterns.
- **Predictive Modeling**: Several machine learning models—including logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbors (KNN)—were developed to predict the likelihood of a successful landing.
- This comprehensive approach enables a deeper understanding of the factors influencing rocket landing success and helps to predict future outcomes more accurately.

Introduction

Project Background

- SpaceX, a trailblazer in the space industry, aims to make space travel more accessible and affordable. The company has achieved significant milestones, including sending spacecraft to the International Space Station, launching a satellite constellation that provides global internet coverage, and conducting crewed space missions. A key factor in SpaceX's ability to reduce the cost of space missions is the innovative reuse of the Falcon 9 rocket's first stage, which keeps the cost of each launch at around $62 million. In contrast, competitors that cannot reuse the first stage face launch costs exceeding $165 million.
- By predicting whether the first stage of a rocket will successfully land, we can estimate the overall cost of a launch. This can be done by leveraging publicly available data and applying machine learning models to forecast the likelihood of first-stage reusability, providing valuable insights not only for SpaceX but also for other companies in the industry.

Key Exploration Questions

- How do factors such as payload mass, launch site, number of flights, and orbits influence the success of first-stage landings?
- How has the rate of successful landings evolved over time?
- What is the most effective predictive model for determining successful landings (binary classification)?

Section 1

Methodology

Methodology

- Data collection methodology:

  Data was collected using the following

  SpaceX Rest API

  Web Scrapping from Wikipedia

- Perform data wrangling

  Filtering the data

  Dealing with missing values

  Using One Hot Encoding to prepare the data to a binary classification

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  ➢ Building, tuning and evaluation of classification models to ensure the best results

Data Collection

The data was collected using various methods

- Data Collection - API
- We Request data from SpaceX API (rocket launch data)
- We decode response using .json() and convert to a dataframe using .json_normalize()
- We use the API again to get information about the launches using the IDs given for each launch
- We Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean() and with replace() function to replace np.nan values in the data
- Export data to csv file

Data Collection ע SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/01%20jupyter-labs-spacex-data-collection-API.ipynb

Data Collection - Scraping

- We performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page
- We perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
- we collected all relevant column names from the HTML table header
- We applied web scrapping to webscrap Falcon 9 launchrecords with BeautifulSoup.
- We parsed the table and converted it into a pandas dataframe.
- WeCreate a data frame by parsing the launch HTML tables

https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/02%20jupyter-labs-webscraping%20.ipynb

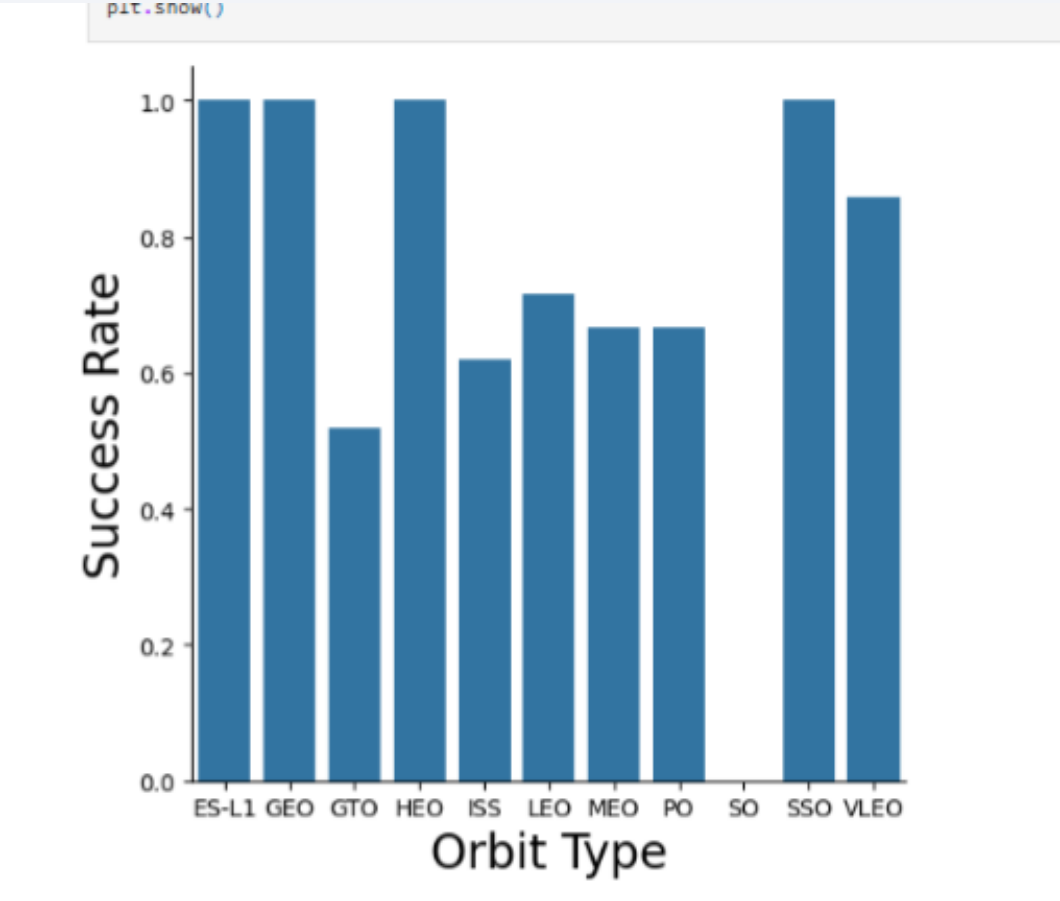
Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.

- https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/03%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.
- We created dummy variables to categorical columns
- We Use the function get_dummies and features dataframe to apply OneHotEncoder to the column Orbits, LaunchSite, LandingPad, and Serial.



https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/05%20EDA%20data%20visualization.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
- We display the names of unique launch sites in the space mission.
- We Display 5 records where launch sites begin with the string 'CCA'

- We displayed the total payload mass carried by boosters launched by NASA (CRS)
- We displayed the average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/04%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- •We assigned the feature launch outcomes (failure or success) to class 0 and1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities.
- We answered some question for instance:-Are launch sites near railways, highways and coastlines.-Do launch sites keep certain distance away from cities . Build an Interactive Map with Folium

https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/06%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

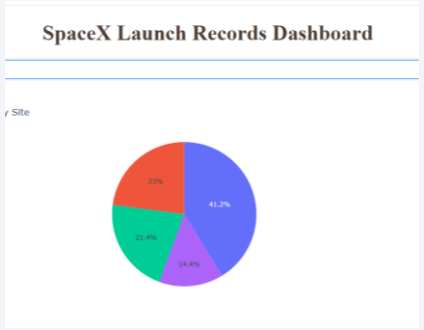
- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- Explain why you added those plots and interactions

Pie Charts Showing Total Launches by Site: The pie charts were added to visualize the distribution of launches across different sites.

Scatter Plot Showing the Relationship Between Outcome and Payload Mass (Kg) by Booster Version: This scatter plot helps in understanding how payload mass influences the outcome of the launch for different booster versions

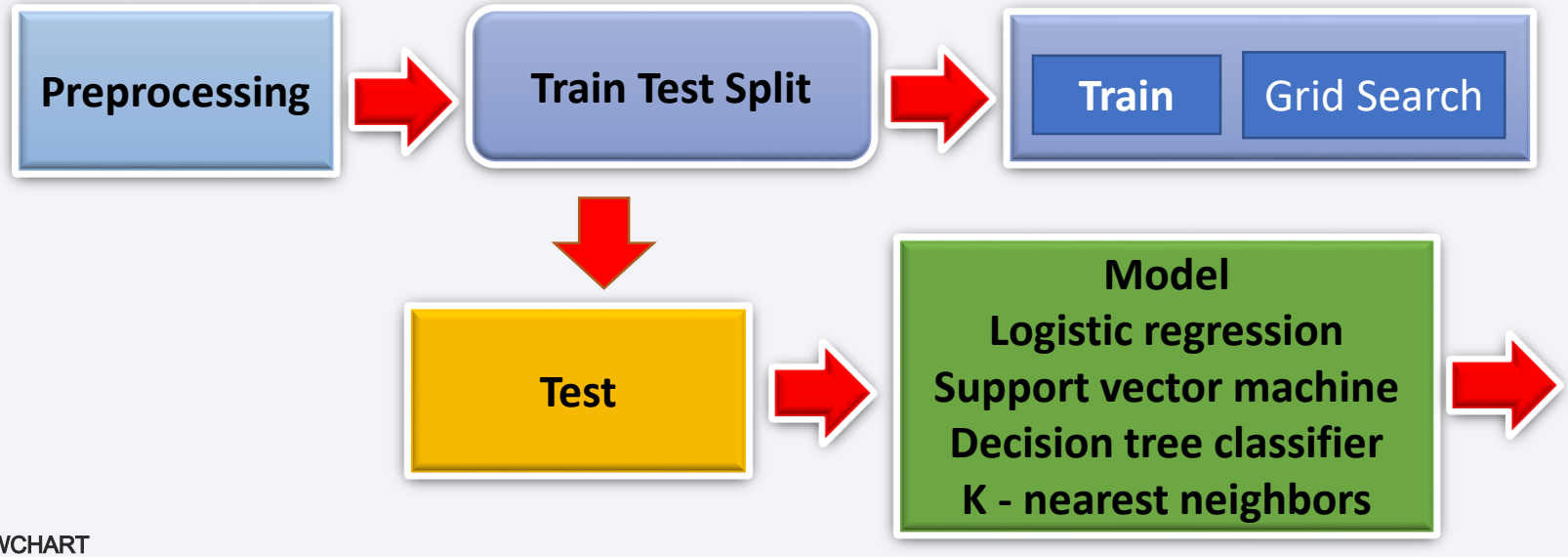
Interactive Dashboard: The interactive nature of the dashboard allows users to filter, zoom, and dynamically adjust the views, facilitating real-time exploration of the data. This enhances the user experience by allowing them to focus on particular subsets of the data, helping them draw conclusions quickly or dive deeper into specific aspects of the dataset.

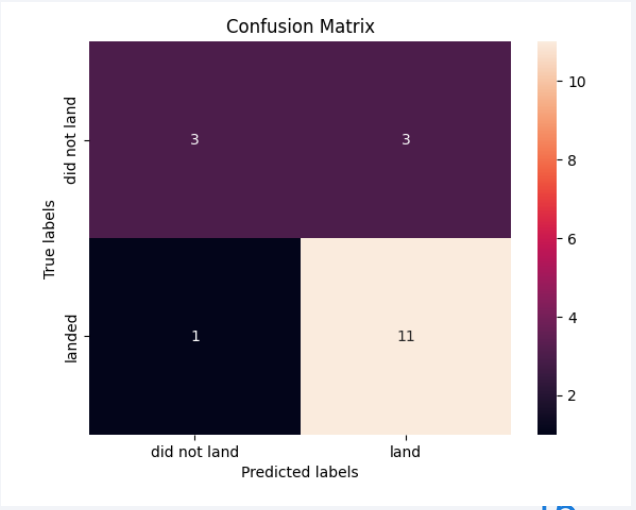
https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/07%20spacex_dash_app.py



Predictive Analysis (Classification)

- We create a NumPy array from the 'Class' column in the data by applying the to_numpy() method and assign it to the variable Y.
- we standardize the data in X through preprocessing and reassign the transformed data back to X.
- we perform data transformation and split the dataset into training and testing subsets.
- We then develop multiple machine learning models and optimize their hyperparameters using GridSearchCV.
- The accuracy was calculated on the test data using the score method as the evaluation metric. The model was then enhanced through feature engineering and algorithm tuning. After several iterations, the best-performing classification model was identified.



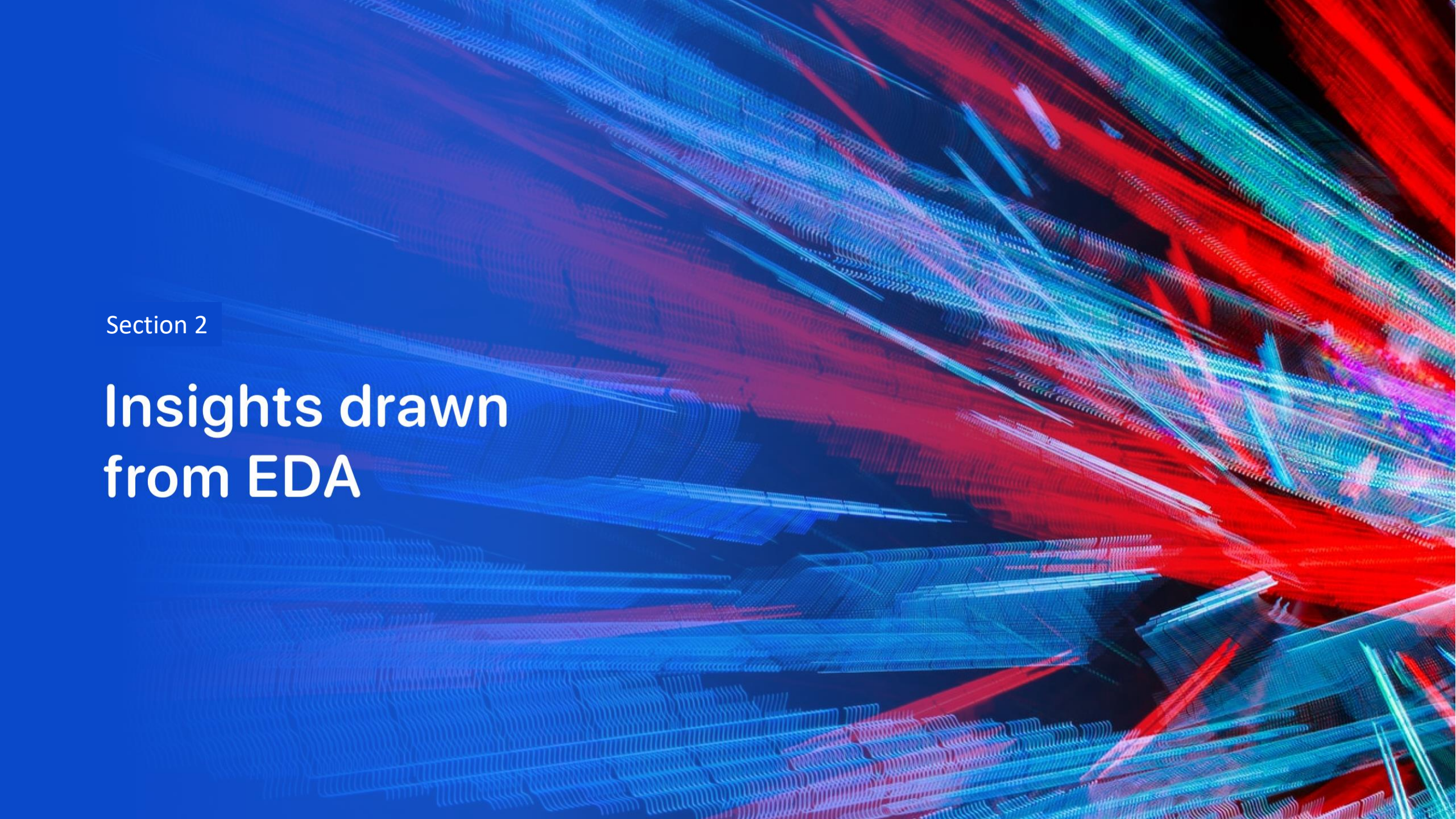


FLOWCHART

- https://github.com/abelronick/Applied_Data_Science_Capston/blob/main/08%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

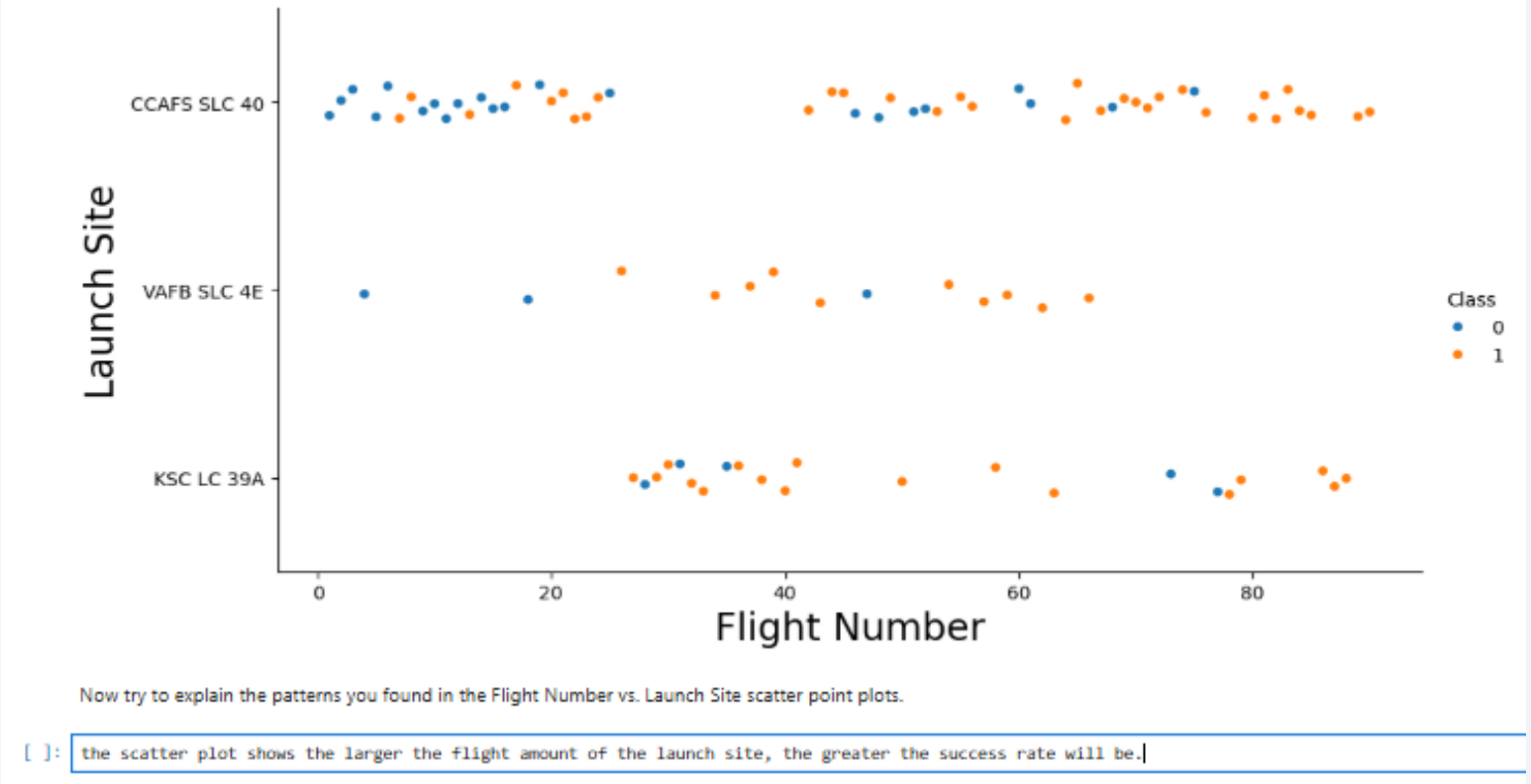
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results
- **Exploratory Data Analysis:**
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Visualization / Analytics:**
- Most launch sites are near the equator, and all are close to the coast
- **Predictive Analytics**
- All models performed similarly on the test set. The decision tree model slightly outperformed when looking at .best_score_

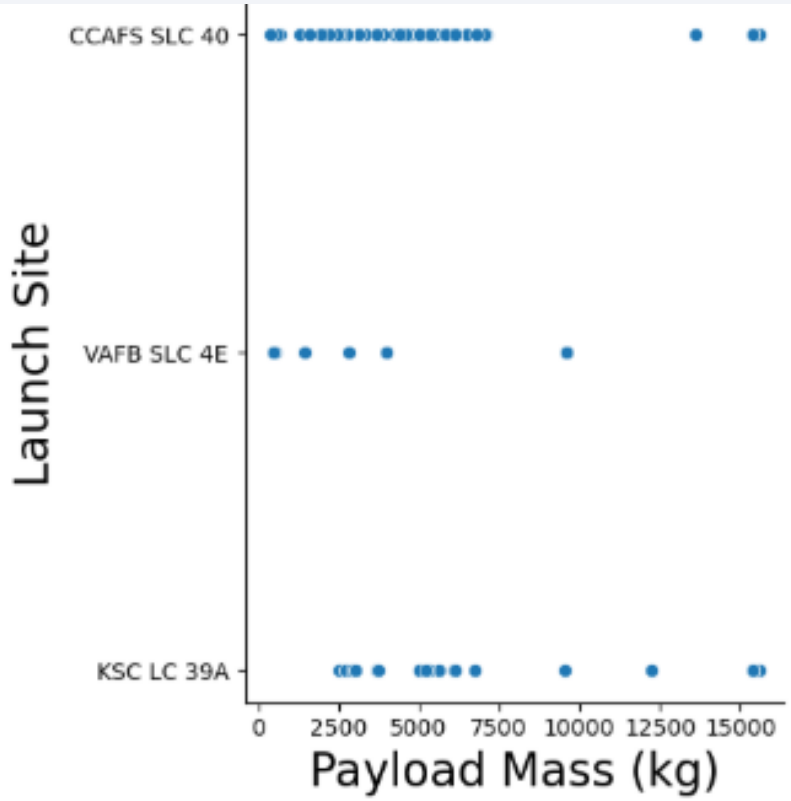

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- the scatter plot shows the larger the flight amount of the launch site, the greater the success rate will be.



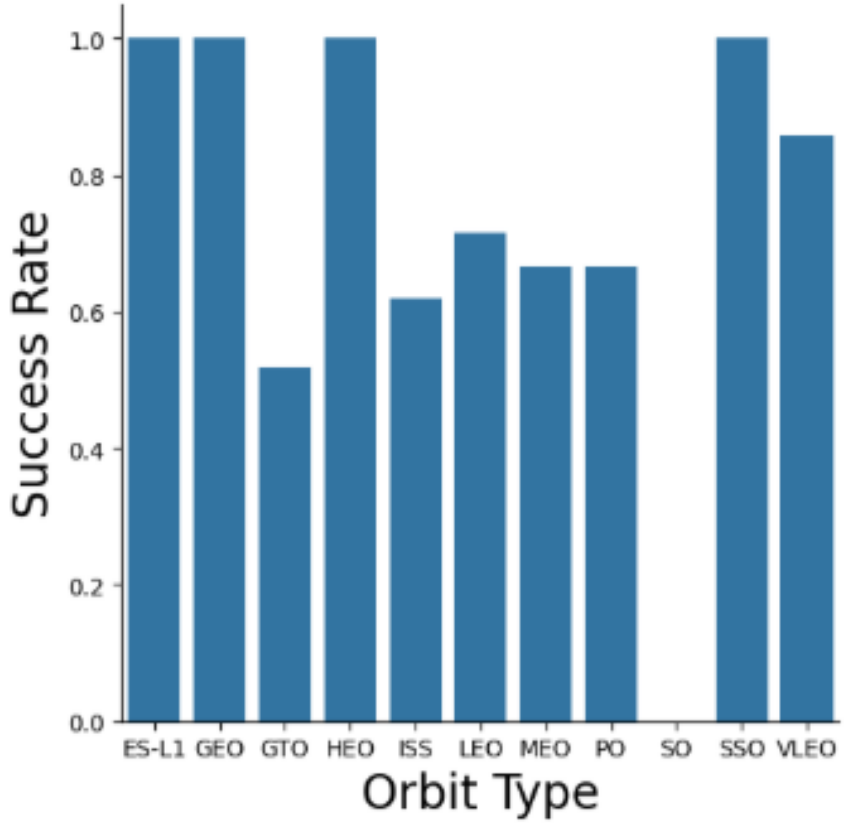
Payload vs. Launch Site



Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

From the bar chart plot the orbit type that has the highest success rate are ES-11, GEO, HEO, SSSO,

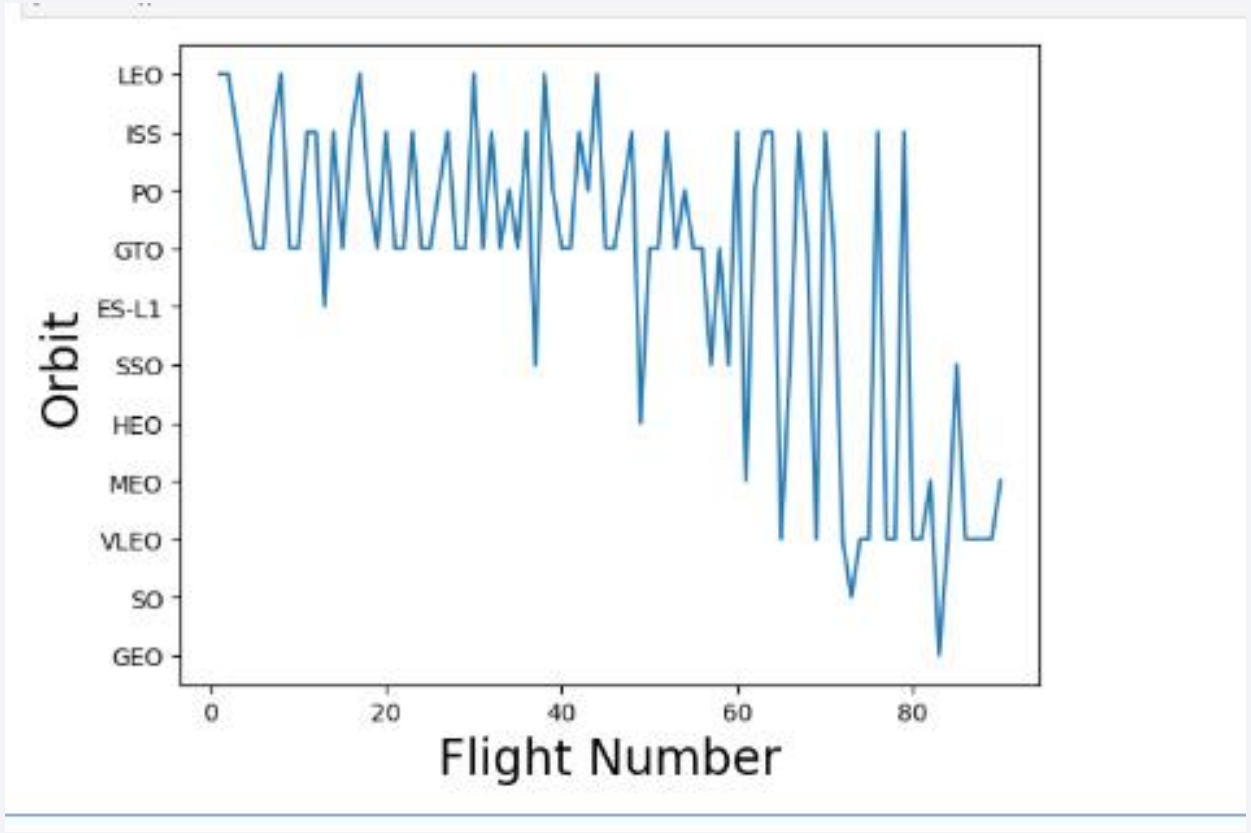


Analyze the plotted bar chart to identify which orbits have the highest success rates.

[ ]: from the bar chart plot the orbit type that has the highest success rate are ES-11, GEO, HEO, SSSO,

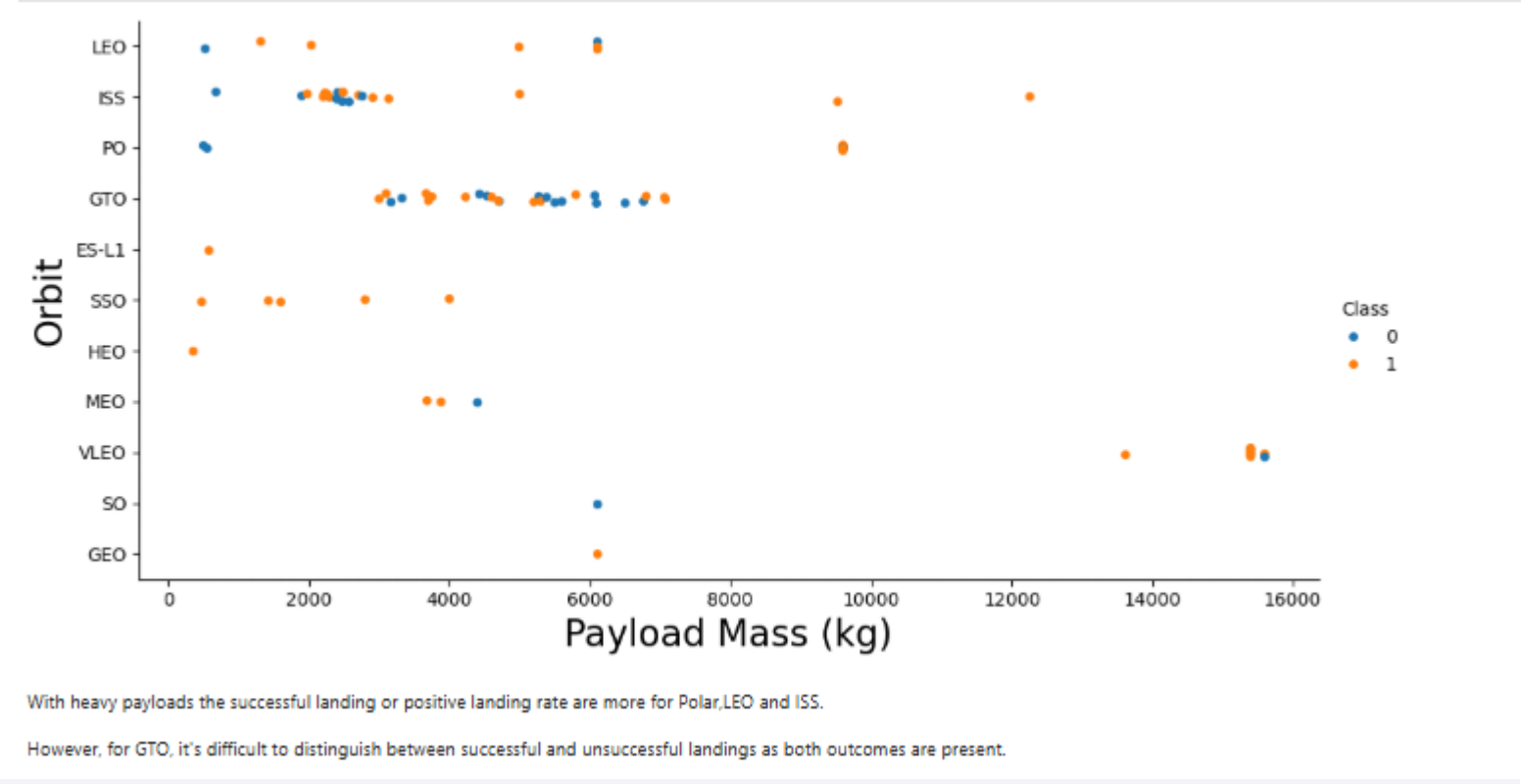
Flight Number vs. Orbit Type

From the line plot the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



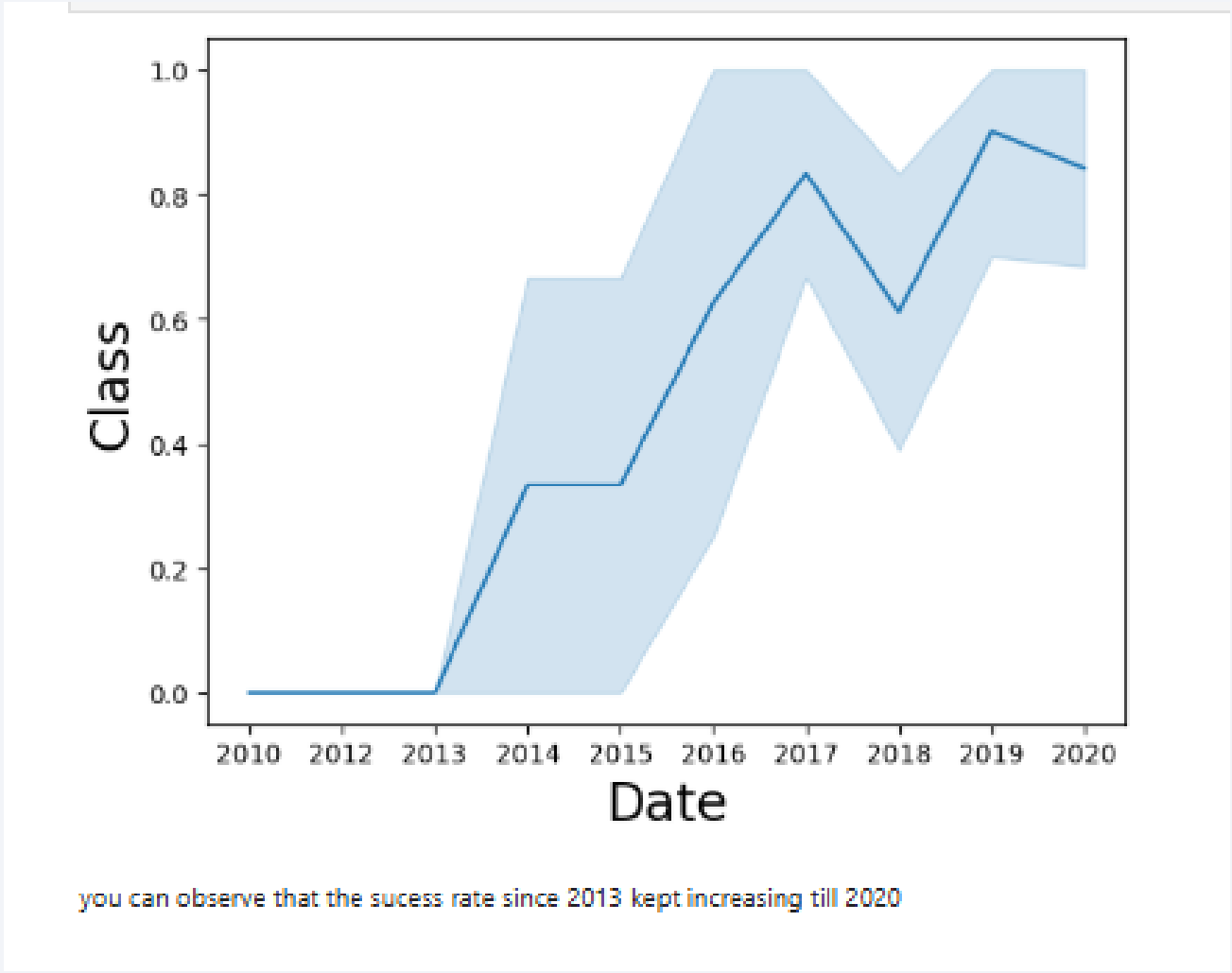
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

The line plot shows that the success rate keep increasing since 2013 till 2020



All Launch Site Names

the names of the unique launch sites

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

- **CCAFS LC-40**: This is one of the launch complexes located at Cape Canaveral Air Force Station (CCAFS) in Florida. It has been used for various space missions, including those by SpaceX.
- **VAFB SLC-4E**: Vandenberg Air Force Base Space Launch Complex 4 East (VAFB SLC-4E) is located in California and is primarily used for launching satellites into polar orbits.
- **KSC LC-39A**: Kennedy Space Center Launch Complex 39A (KSC LC-39A) is a historic launch site in Florida, known for its use in NASA's Apollo and Space Shuttle programs. It is now used by SpaceX for Falcon 9 and Falcon Heavy launches.
- **CCAFS SLC-40**: Similar to CCAFS LC-40, this is another launch complex at Cape Canaveral used by SpaceX for launching their Falcon rockets.
- These launch sites are integral to various space missions, providing the infrastructure needed to send payloads into space

Launch Site Names Begin with 'CCA'

[11]:

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

- CCAFS LC-40 and CCAFS SLC-40: These are launch complexes at Cape Canaveral Air Force Station, used for various space missions.
- Other records may include similar launch sites that start with CCA, indicating their location at Cape Canaveral.
- These records provide insight into the specific launch sites used for space missions, particularly those associated with Cape Canaveral. If you have any further questions or need additional assistance, feel free to ask!

Total Payload Mass

Total Payload Mass is 45596.

This value represents the sum of all payloads carried by NASA's boosters, providing insight into the total weight of materials launched into space by NASA.

Average Payload Mass by F9 v1.1

Average Payload Mass is 29284,

This value indicates the mean weight of payloads that the F9 v1.1 booster has carried across its missions. It provides insight into the typical payload capacity of this specific booster version.

First Successful Ground Landing Date

First Successful Landing Date was 2015- 12- 22

This date indicates when the first successful landing on a ground pad occurred, providing insight into the timeline of successful landings in the dataset.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Payload
- JCSAT-14
- JCSAT-16
- SES-10
- SES-11 / EchoStar 105
- Booster Names: The list of booster names indicates which boosters have successfully landed on a drone ship while carrying a payload mass within the specified range. This information can be useful for analyzing the performance and capabilities of different boosters.

Total Number of Successful and Failure Mission Outcomes

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

The result shows the number of missions for each type of landing outcome, such as "Success" or "Failure". This information provides insight into the overall success rate and challenges faced during the missions.

Boosters Carried Maximum Payload

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.

Booster Names: The list of booster names indicates which boosters have achieved the highest payload capacity in your dataset. This information can be useful for identifying the most capable boosters in terms of payload capacity.

2015 Launch Records

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Landing Outcome: This column will show the failed landing outcomes on a drone ship.
Booster Version: This column will list the versions of the boosters involved in these failed landings.
Launch Site: This column will provide the names of the launch sites where these missions took place.
This information can be useful for analyzing the challenges faced during drone ship landings in 2015.

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

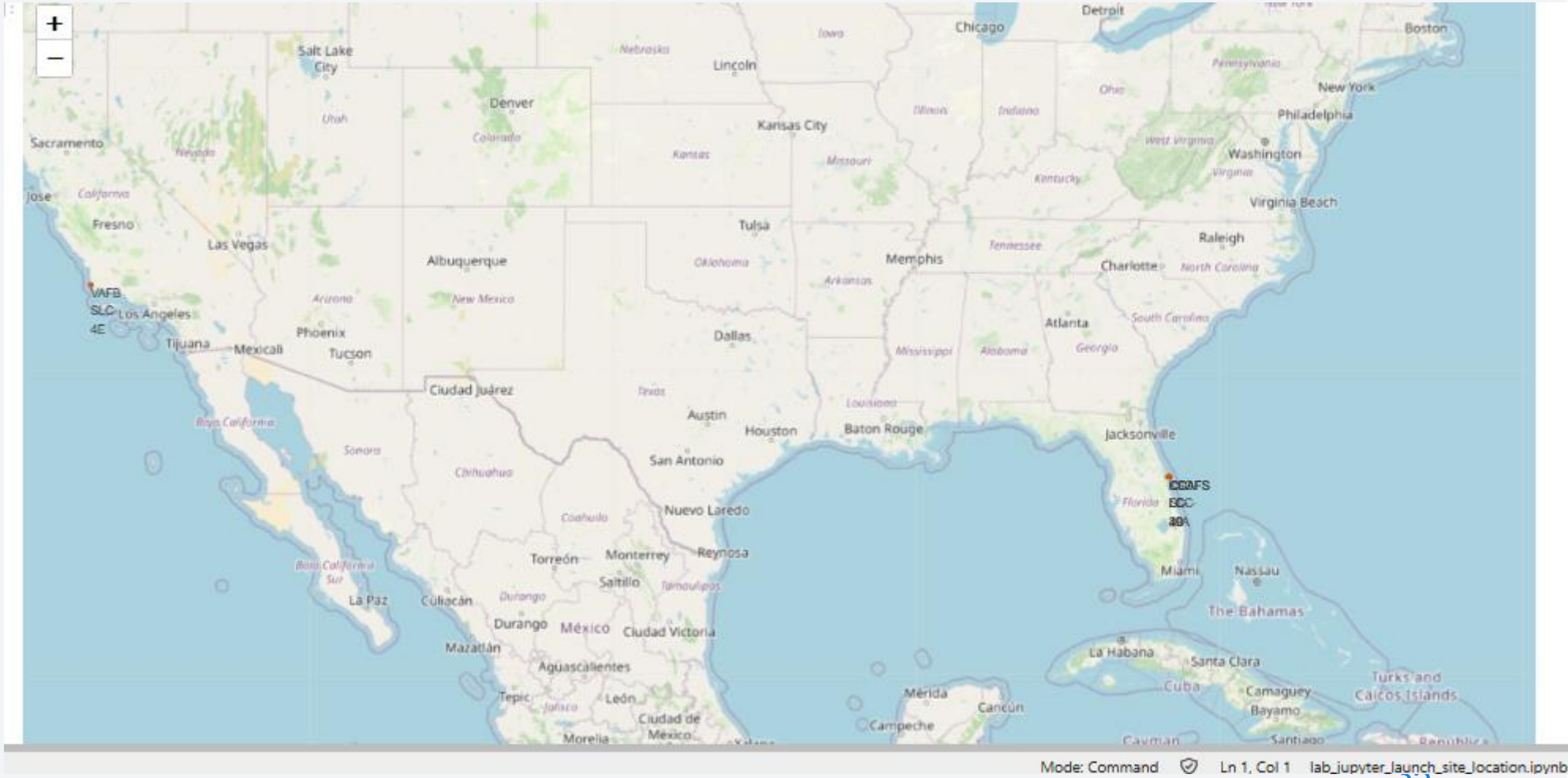
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	

Section 3

Launch Sites Proximities Analysis

All launch site location makers

Launch site location markers are strategically positioned to leverage geographical advantages. They are often located near the Equator to benefit from the Earth's rotational speed, which provides an additional velocity boost for rockets, particularly useful for launching satellites into geostationary orbits. Additionally, proximity to coastlines is common, allowing rockets to travel over water, minimizing risk to populated areas in case of launch failures. This strategic placement ensures both safety and efficiency in launch operations, balancing the need for optimal launch conditions with safety considerations.

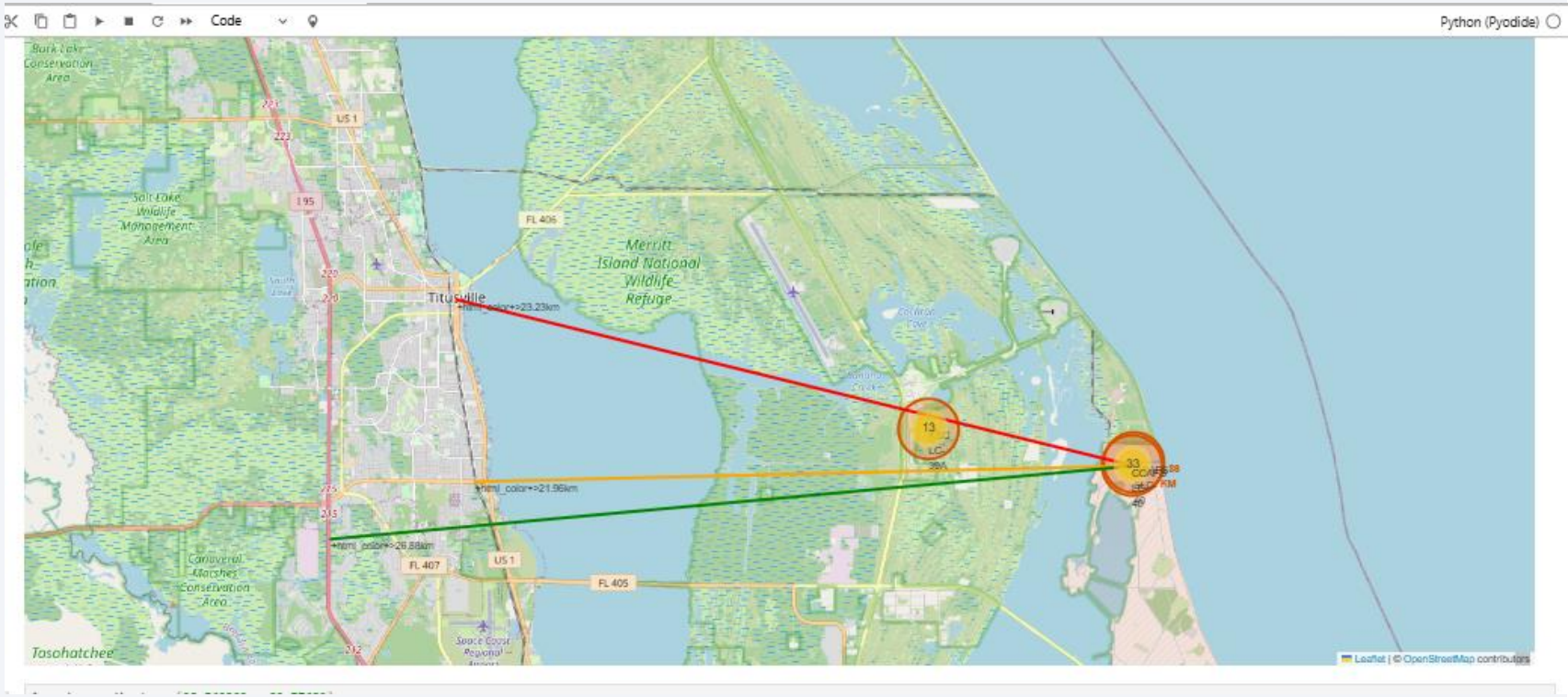


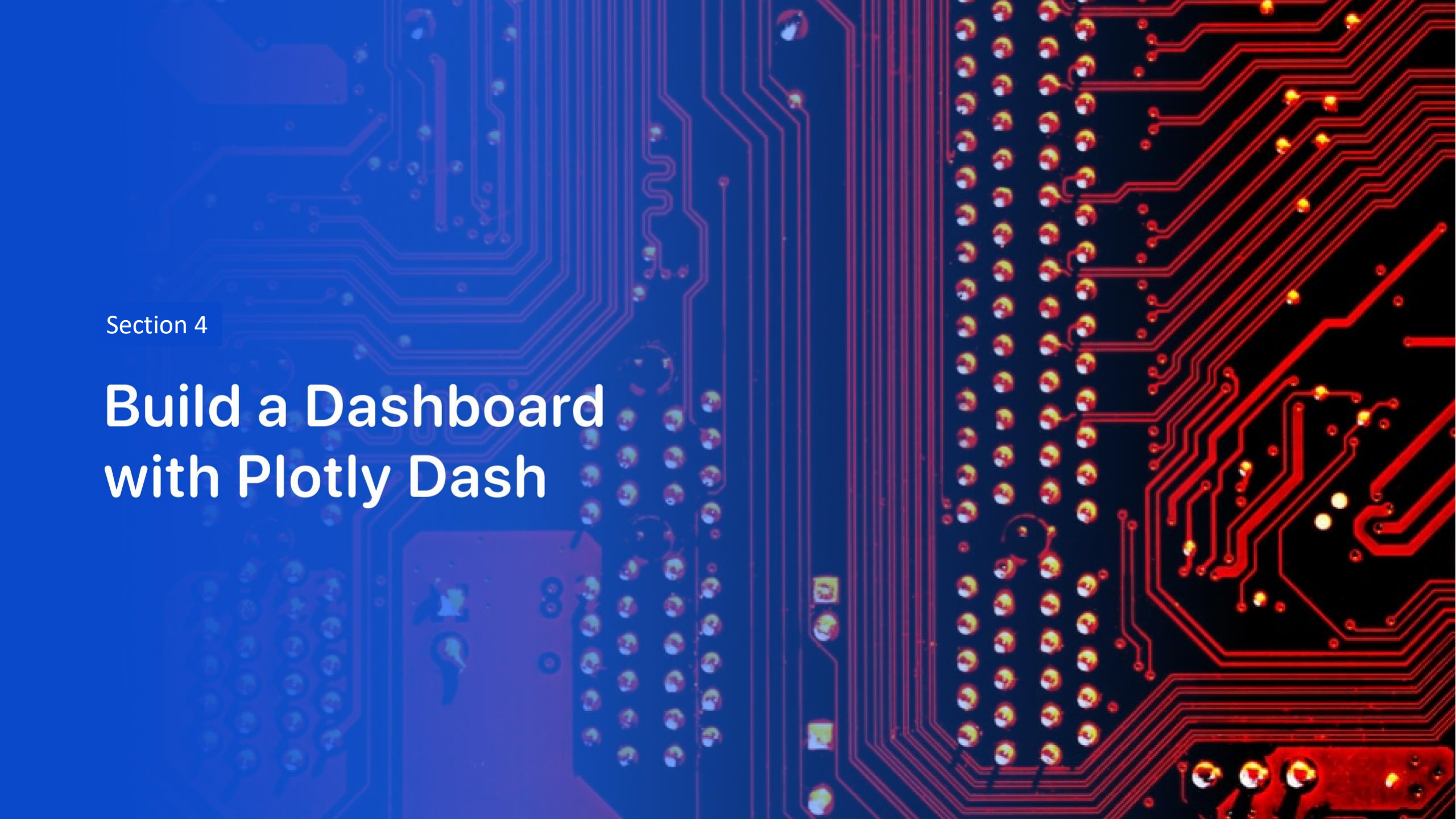
Color-labeled launch outcome

Color-labeled launch outcomes on the map visually represent the success and failure rates at each launch site, using distinct colors—typically green for successful launches and red for failures. This color coding allows for a quick assessment of each site's performance, with marker clusters helping to manage data density on the map. By analyzing the ratio of successful to failed launches, one can gauge the reliability of each site and potentially identify geographical or environmental factors affecting launch outcomes. This visualization aids in understanding historical performance and identifying areas for improvement.

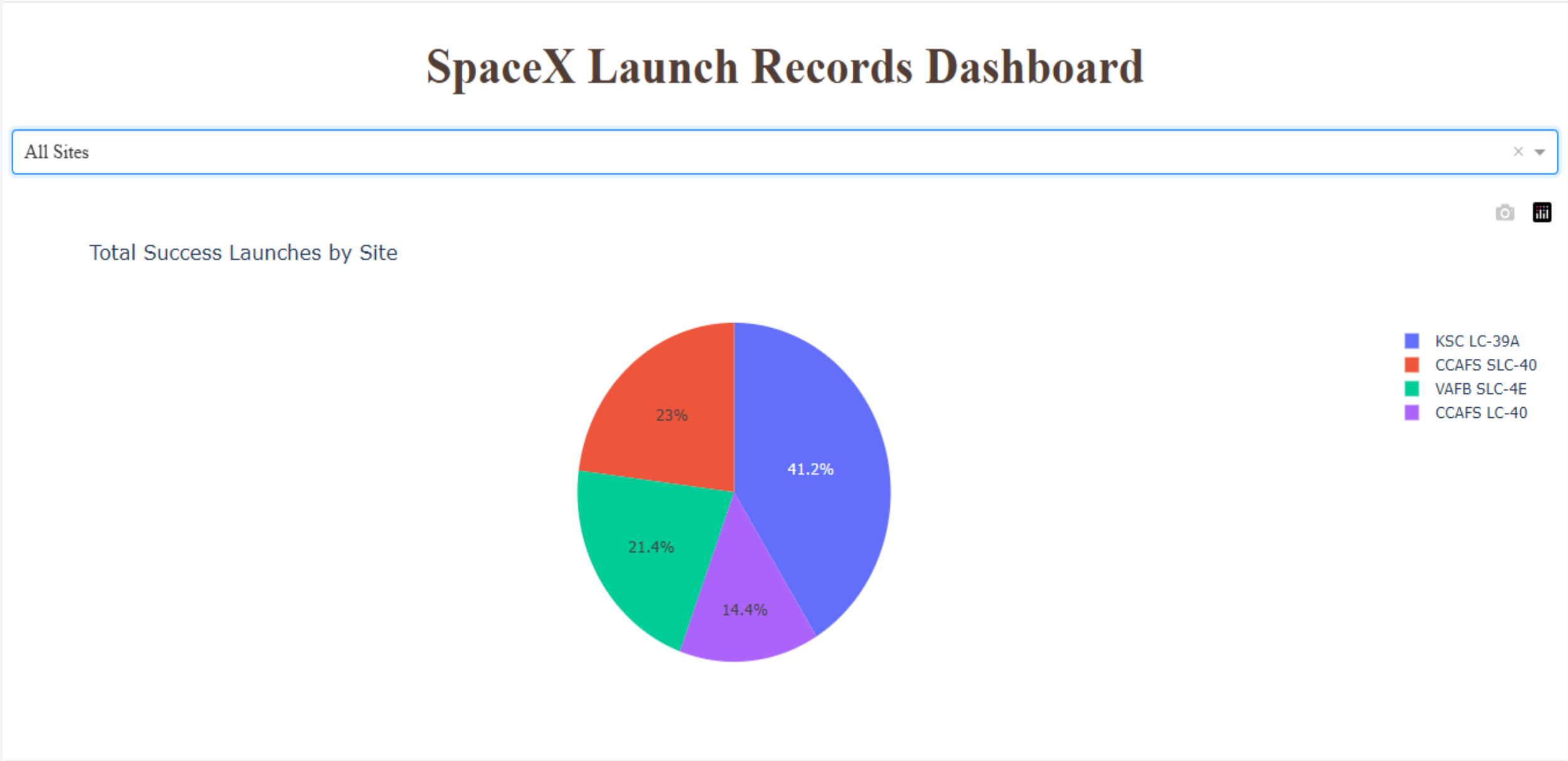
Distance between a launch site to its proximities

Analyzing the distances between a launch site and its proximities reveals strategic considerations for site placement, balancing logistical efficiency and safety. Proximity to railways and highways facilitates efficient transportation of materials and personnel, enhancing logistical operations. Being near coastlines allows for safe over-water flight paths, reducing risk to populated areas in case of launch anomalies. Additionally, maintaining a safe distance from cities minimizes risk to human populations and infrastructure, ensuring safety in the event of a launch failure. This analysis underscores the importance of these factors in the strategic planning and selection of launch sites.



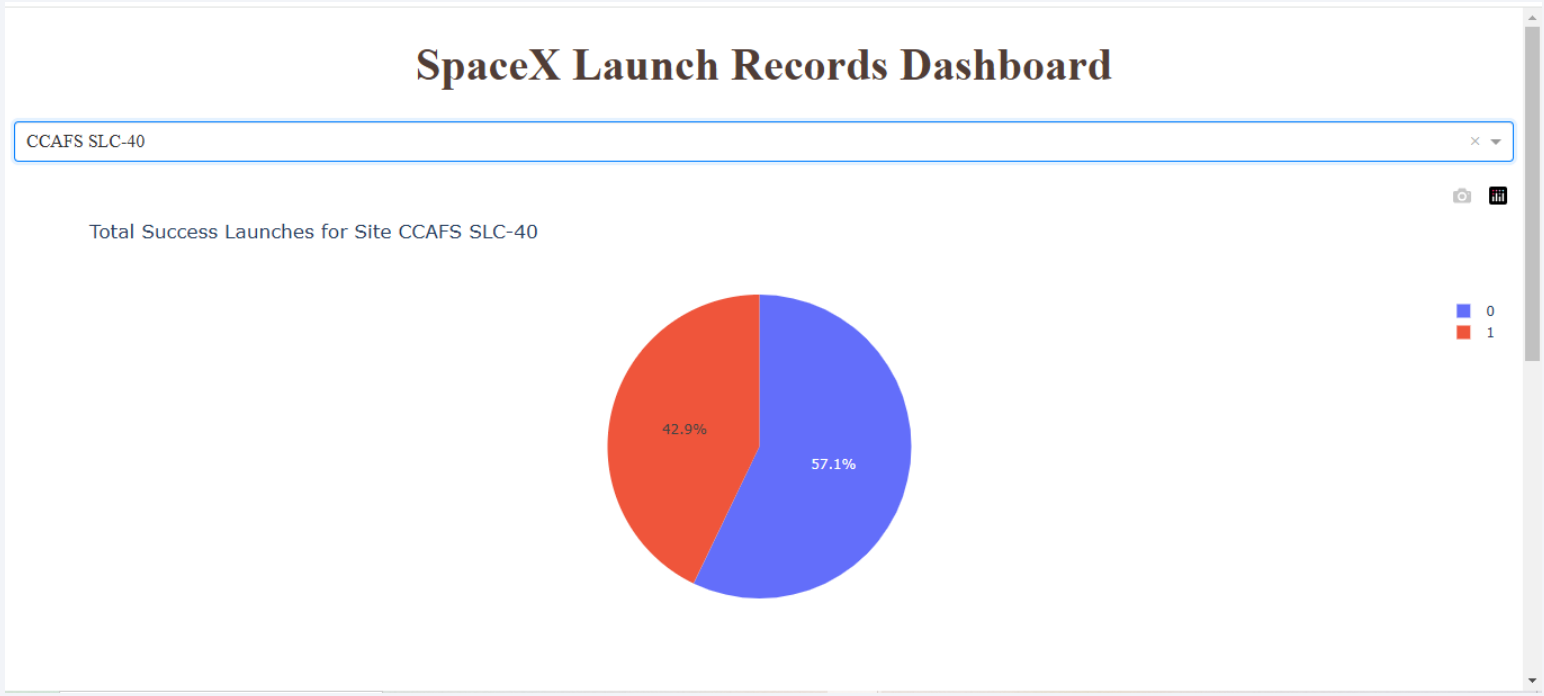

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all site



Highest success ratio

The pie chart shows the launch site with highest launch success ratio



Payload vs Launch outcome scatter plot

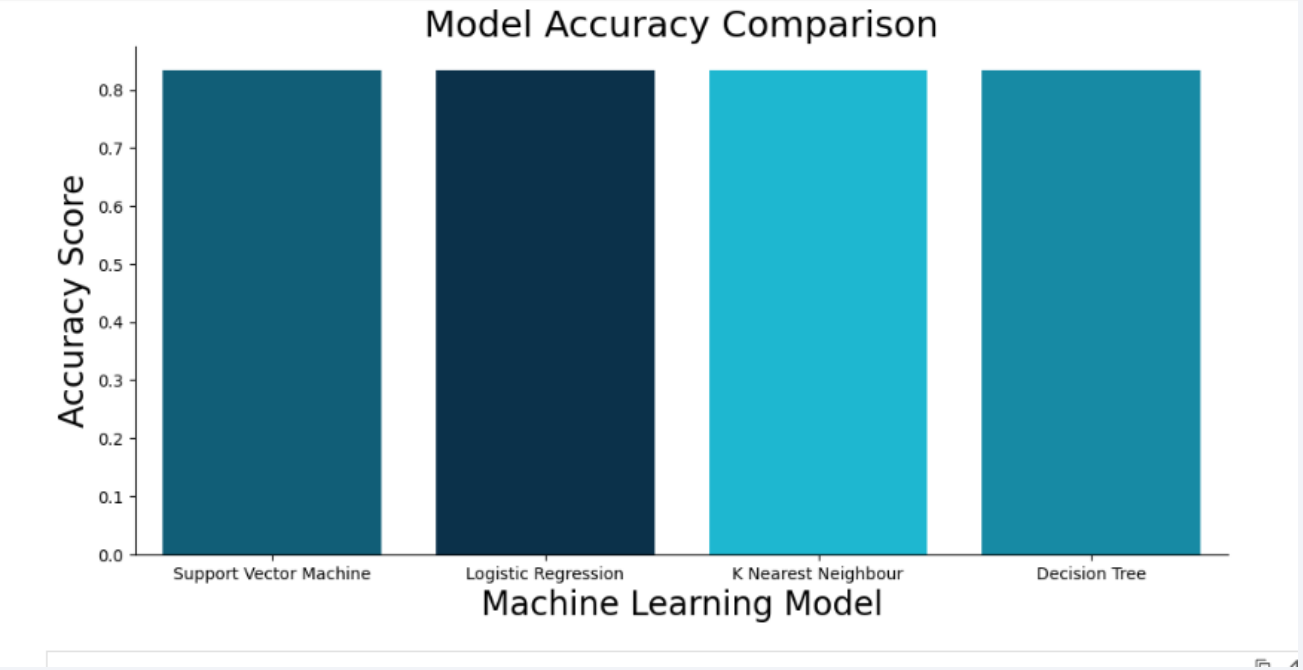
The scatter plot provides valuable insights into the success rates of different payload ranges and booster versions. The X-axis represents the payload mass, while the Y-axis indicates the launch outcome, with successful launches positioned higher on the plot. Color coding differentiates booster versions, allowing for easy identification of which versions are more successful. By examining clusters of points at the top of the plot, the payload range that has the highest success launch is between 2000 and 4000 kg , followed by payload range of 4000 to 6000kg ,

Booster version FT(green spots) has the highest success launches, followed by B4(purple spots) with the second highest launches .

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Find the method performs best:

[39]:
```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

[39]:

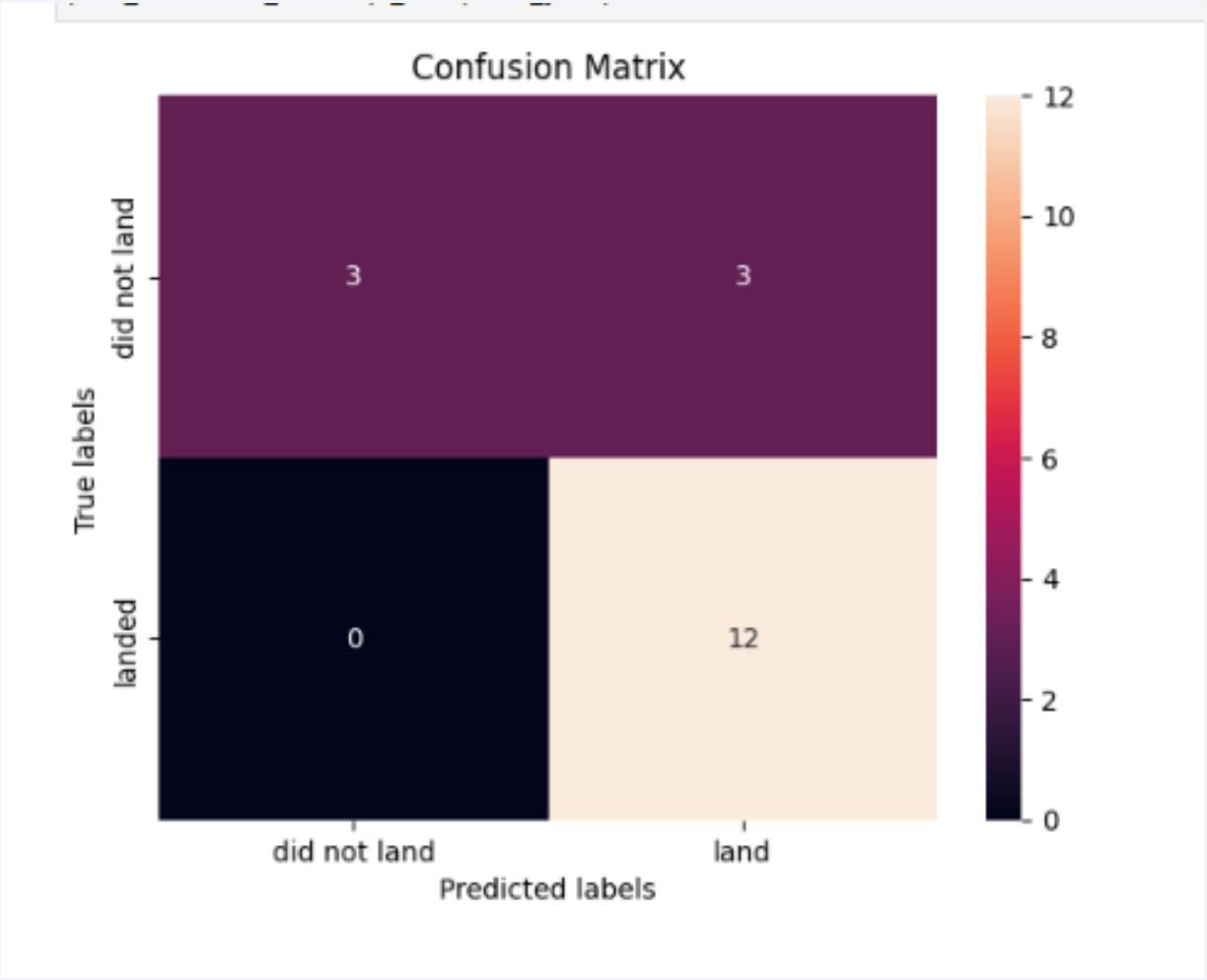
	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

The bar chart shows that the models have the same accuracy performance

.

Confusion Matrix

The confusion matrix for the Decision Tree classifier indicates that it struggles to differentiate between the various classes. A significant issue with the classifier is the presence of false positives, where an unsuccessful landing is incorrectly labeled as a successful landing.



Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!