IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zain Naboulsi
06/27/2023

Outline

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

Executive Summary - Methodologies

This report synthesizes an extensive SpaceX launch study, deploying numerous data streams, notably the SpaceX REST API and Wikipedia page scraping. Data wrangling hurdles like null values and irrelevant data got resolved, resulting in a well-rounded dataset. With the aid of Python tools, the data underwent scrutiny, manifesting trends and relationships between variables such as payload mass, launch site, and orbit type. An interactive Plotly dashboard got constructed, offering enhanced data interpretation with elements like a payload range slider and site-focused success rate visual.

Multiple classification models underwent training and evaluation for the prediction of rocket landing success. A robust method involving hyperparameter tuning and cross-validation got adopted, selecting the highest accuracy model as the final predictive instrument. This aids future SpaceX missions with data-driven decisions.

Executive Summary - Results

Insights of importance emerge from the launch data analysis:

Payload weight and launch failure rates display an inverse relationship. Beyond the 8000 mass mark, payload weight increases coincide with decreases in failure rates. This suggests more reliable operations with heavier payloads.

Four orbit types show the highest success rates: Lagrange Point 1 (ES-L1), Geostationary orbit (GEO), Highly Elliptical Orbit (HEO), and Sun-Synchronous Orbit (SSO). This information can guide mission planning to optimize for success.

Over time, the data reveals a positive trend, with a clear linear path towards increased mission success. This reflects our continuous improvement efforts and technological advancements.

However, the analysis also reveals a significant amount of missing data on launches. This may affect the accuracy of our findings and predictions. Efforts should be made to improve data collection and management processes.

Regarding launch sites, all are strategically located near the coast and away from major cities, minimizing potential risks to populated areas. They are also conveniently situated near heavy load-carrying resources like roads and railroads, facilitating logistics.

Despite these strategic locations, all launch sites had a success rate of less than 50%, with the highest being 40%. This indicates room for improvement in our launch operations.

Introduction

Background

Commercial space travel, once exclusive to government agencies, is now a reality with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX. SpaceX stands out with its Falcon 9 rocket. This rocket has cut space travel cost significantly, primarily due to its reusable first stage.

Problem Statement

A crucial question is predicting the successful landing of the first stage of a launch. This prediction can provide insights into the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data for this analysis was primarily sourced from the SpaceX REST API
  - Web scraping was used to obtain additional data from Wikipedia
- Performed data wrangling
  - Determining labels for our supervised models
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Hyperparameter tuning for multiple types of classifiers

Data Collection

The data for this analysis was primarily sourced from the SpaceX REST API, an open interface providing detailed information about SpaceX launches. Data, such as details about the rockets used, payload specifications, and landing outcomes, are all accessible via specific API endpoints.

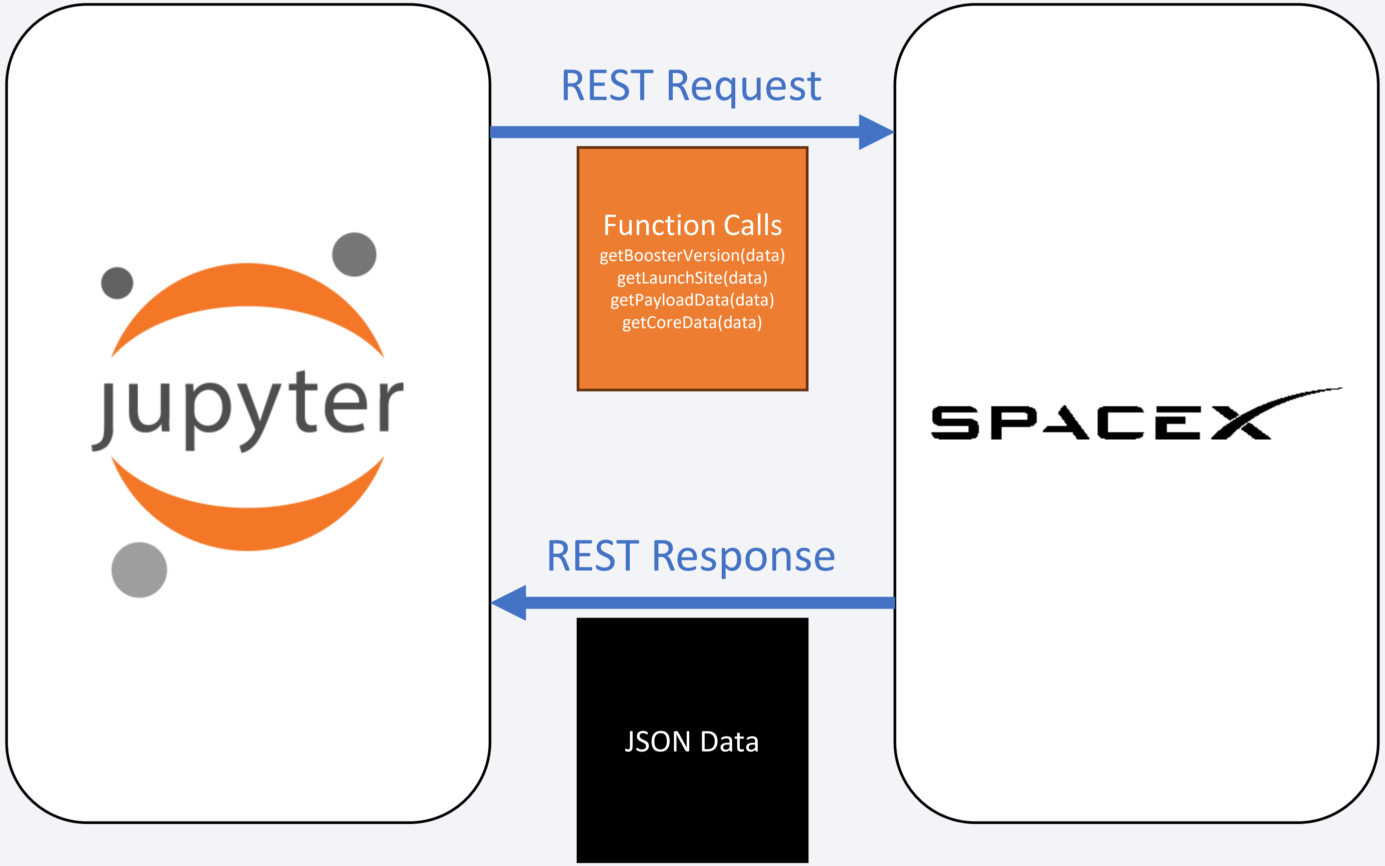
In addition, web scraping techniques were employed to extract Falcon 9 launch records from relevant Wiki pages. This additional data source enriched the insights from the API data, creating a more holistic picture of the Falcon 9 launches.

Data Collection – SpaceX API

We use the SpaceX REST API to collect detailed data about past rocket launches. Functions access specific API endpoints, retrieve JSON data, and process this data.

For example, the `getBoosterVersion` function interacts with the 'rocket' endpoint, fetches the JSON data, and extracts the name of the rocket. In the same way, `getLaunchSite`, `getPayloadData`, and `getCoreData` functions access and pull critical details from 'launchpad', 'payloads', and 'cores' endpoints, respectively.

These steps help us create a detailed dataset from various related sources, offering useful insights into the factors that impact the success of rocket landings.



REST Request

Function Calls
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)

REST Response

JSON Data

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

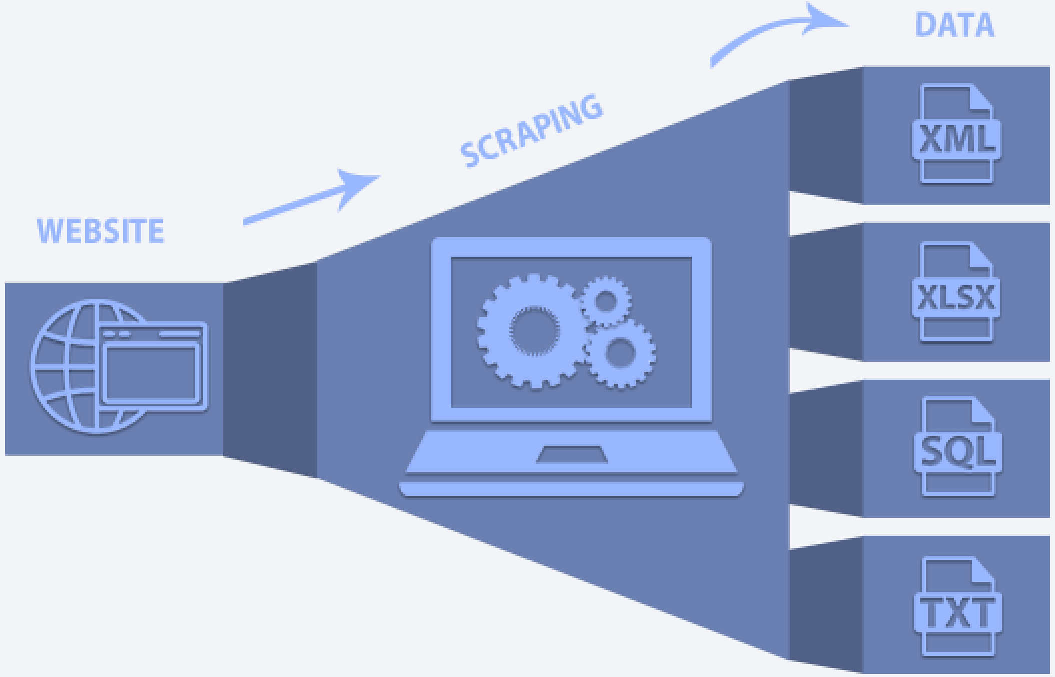
Data Collection Scraping

Web scraping is a technique we use to extract information from websites.

In this case, we extract Falcon 9 historical launch records from a Wikipedia page.

This particular page contains information within an HTML table, and we utilize a tool named BeautifulSoup to parse this data. Once the table is parsed, we convert this data into a DataFrame. The DataFrame structure allows us to further analyze and visualize the data, paving the way for insightful discoveries about Falcon 9 launch records.

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Data wrangling involves refining raw data into a more usable format.

As part of our analysis, we conduct Exploratory Data Analysis (EDA) to detect patterns, spot anomalies, or test hypotheses.

A critical aspect of this process involves determining labels for our supervised models.

Our dataset has instances where the Falcon 9 booster did not land as planned. These outcomes fall into various categories, such as successful or unsuccessful landing on ocean, ground pad, or drone ship, denoted by terms like 'True Ocean', 'False Ocean', 'True RTLS', 'False RTLS', 'True ASDS', and 'False ASDS'.

For simplicity and ease of modeling, we transform these outcomes into training labels. '1' indicates a successful landing, while '0' signifies an unsuccessful one. This transformation will aid in the development of predictive models in later stages.

Discovering
Structuring
Cleaning
Enriching
Validating
Publishing
Data Wrangling

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

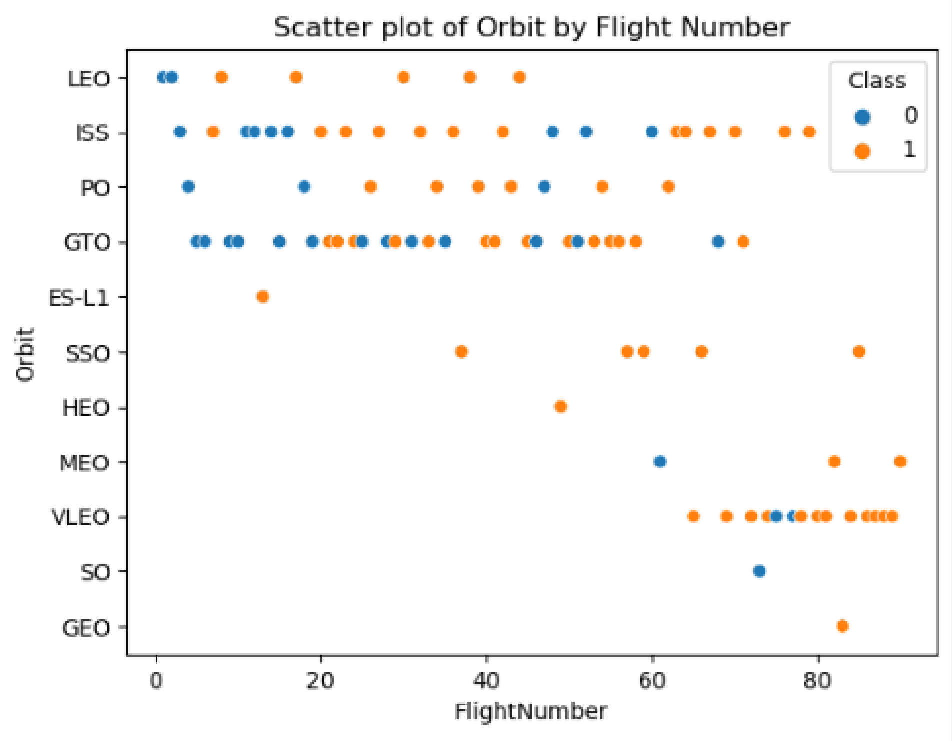
We charted various parameters against each other to understand their impact on launch success. The 'FlightNumber' against 'PayloadMass' plot showed an increased likelihood of first stage landing with increasing flight numbers, but a decreasing success rate with heavier payloads.

Examining success rates by launch sites, we found variances 'CCAFS LC-40' had a 60% success rate, while 'KSC LC-39A' and 'VAFB SLC 4E' had 77%. This was visualized using a 'catplot' with 'FlightNumber' and 'LaunchSite' on the axes, and the success class as the hue.

Investigating the relationship between launch sites and payload mass, we discovered that the 'VAFB-SLC' site hadn't launched any rockets for payloads over 10000.

A relationship was also sought between success rate and orbit type through a bar chart. It indicated that certain orbits had a higher success rate. The scatterplot of 'FlightNumber' against 'Orbit' type showed a relationship between success and flight numbers in 'LEO' orbit but not in 'GTO' orbit.

Finally, examining payload mass against orbit type, we found that successful landings were more frequent for 'Polar', 'LEO', and 'ISS' with heavier payloads, while 'GTO' had both successful and unsuccessful missions.



https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Identified unique launch sites involved in the space mission to understand the diversity of launch locations, essential for assessing geographical dependencies and operational flexibility.
- Showcased five records of launch sites starting with 'CCA' to illustrate the data consistency and validate naming conventions in the dataset, crucial for maintaining data integrity.
- Calculated the total payload mass carried by boosters launched by NASA (CRS) to assess the partnership's contribution to SpaceX's total payload volume, important for strategic partnership evaluations.
- Determined the average payload mass carried by the 'F9 v1.1' booster version to quantify its payload capacity, fundamental in performance comparison against other booster versions.
- Identified the date when the first successful ground pad landing was achieved to mark the milestone in SpaceX's reusability program, critical for tracking technological advancements and operational efficiency.
- Listed the names of boosters that had successful landings on a drone ship and carried a payload mass between 4000 and 6000 kg to highlight specific booster performances under defined conditions, relevant for fine-tuning future missions.
- Counted the total number of successful and failed mission outcomes to assess overall mission success rate, imperative for evaluating company performance and guiding future mission planning.
- Listed booster versions carrying the maximum payload mass to identify the high-performance variants, essential for optimizing mission design considering payload requirements.
- Showcased records displaying the month names, failure landing outcomes in drone ship, booster versions, and launch sites for the year 2015 to provide a detailed performance report for that year, useful for year-over-year performance comparisons.
- Ranked the count of various landing outcomes from June 4, 2010, to March 20, 2017, to analyze the trend and success rate over a specific period, necessary for trend analysis and predicting future success rates.

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Several graphics (shown later) were used to provide insight about the launch locations. We began by marking all launch sites on a map, which gave a clear view of their geographical distribution, providing us with immediate context of the environmental factors they might have been subjected to, like weather and population density. This visual representation allowed a straightforward comparison of location attributes among the sites.

Next, we added markers to indicate the success or failure of launches at each site. By utilizing color-coded markers (green for success, red for failure), we aimed to provide a clear, visual indicator of the success rate for each site. This provided a high-level overview of the performance of each site and potentially highlighted any correlation between location and success rate.

Finally, we calculated and displayed the distances between each launch site and two significant proximities: coastlines and major cities. By drawing lines between the launch site and these points of interest, we visually emphasized the spatial relationship. This aided in understanding logistical considerations, such as how easily materials and personnel could reach the site, the ease of shipping the rockets to the launch site, and safety factors, such as the remoteness from populated areas.

Overall, these additions to the map facilitated a comprehensive visual analysis that went beyond raw data, offering insights into factors that may have impacted the success of a launch.

https://nbviewer.org/github/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

A Plotly dashboard was developed to enhance understanding and interpretation of SpaceX launch data. The dashboard comprised of multiple components for data interaction and visualization.

The dashboard initially presented a dropdown list to allow selection of a specific Launch Site or an option to consider all sites. This interaction allowed executives to examine and compare data on a site-by-site basis or holistically, thus enabling analysis at varying levels of granularity.

To illustrate the success rate of launches, a pie chart was incorporated into the dashboard. When a specific site was selected from the dropdown, the pie chart showed the ratio of Successful to Failed launches for that site. When 'All Sites' option was chosen, it displayed the total successful launches for all sites. This interactive graphical representation provided a clear overview of launch outcomes, aiding in the identification of performance trends and potential issues at specific sites.

A payload range slider was another component added to the dashboard. This slider, which operated over a specified range of payload mass, allowed the user to filter the data based on the payload mass. This provided a valuable opportunity to observe and explore how the payload mass might influence launch success.

The final component was a scatter plot that depicted the correlation between payload mass and launch success. The plot was updated based on the site chosen from the dropdown and the payload range selected. This enabled executives to explore potential relationships between the payload mass and the success of the launches. The data points in the scatter plot were color-coded according to the Booster Version Category, offering further insight into the impact of different booster versions on launch success.

Overall, the interactive features and visualizations of the Plotly dashboard enabled an effective exploration and interpretation of the data, thereby aiding strategic decision-making and planning in SpaceX launch operations.

https://ibm-applied-data-science-capstone.onrender.com/

Predictive Analysis (Classification)

The process of developing a high-performing classification model commenced with exploratory data analysis. This entailed standardizing the data to ensure uniformity, thereby allowing effective comparisons between features. The data was subsequently partitioned into training and test sets, which facilitated both model development and subsequent evaluation.

Following this initial phase, we embarked on the task of hyperparameter tuning for multiple types of classifiers: Support Vector Machine (SVM), Classification Trees, Logistic Regression, and K Nearest Neighbors. Each classifier was explored using a GridSearchCV object, a technique that systematically works through multiple combinations of hyperparameters to determine the most effective ones. This approach was employed with a 10-fold cross-validation to ensure robustness, which involves splitting the data into 10 parts and iteratively training the model on nine parts while testing on the remaining part.

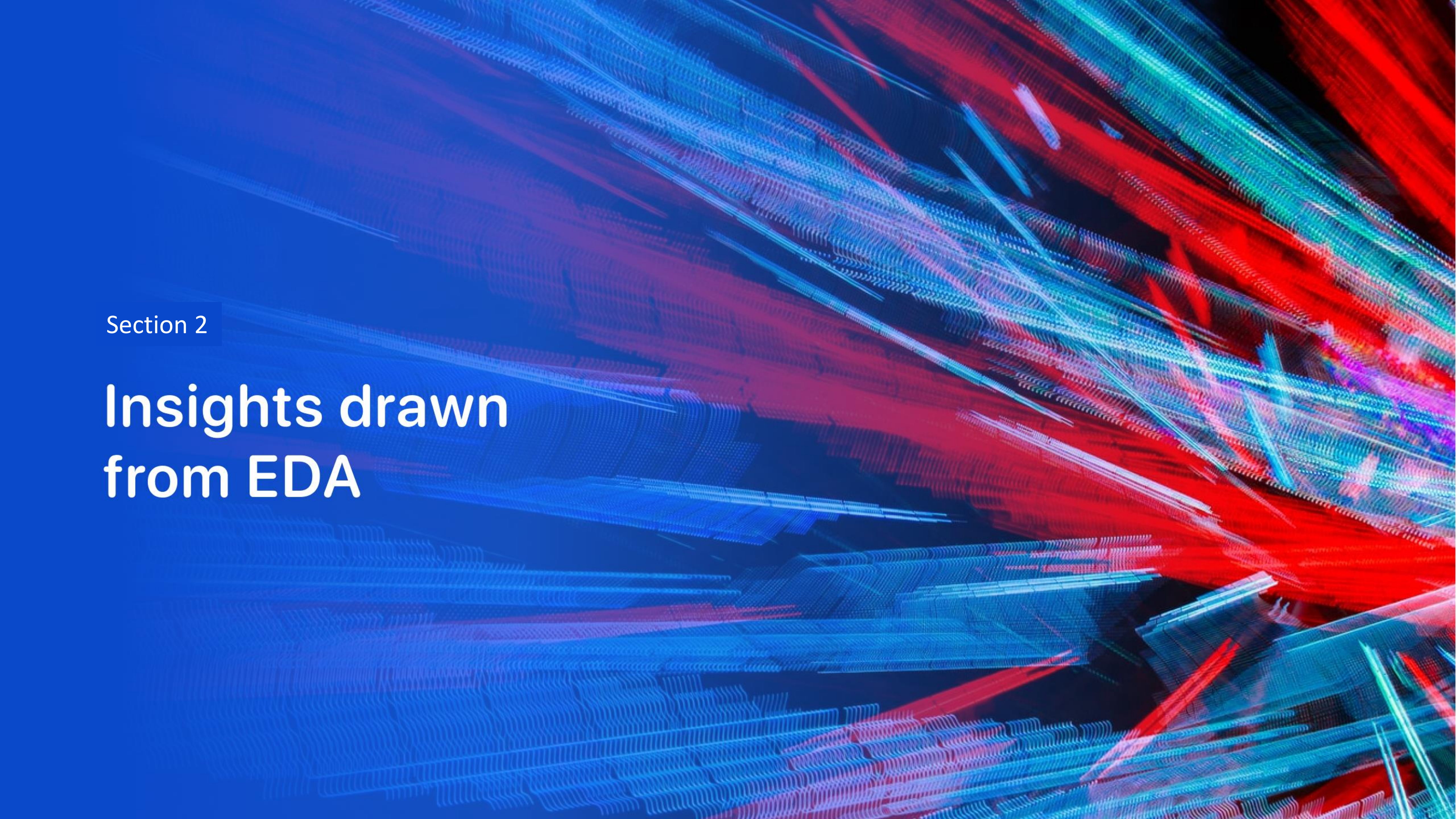
Once hyperparameters for each model were determined, they were fit to the training data, and their performance was evaluated on the test data. Accuracy was the primary metric used for this performance assessment. The confusion matrix was also plotted, which provides a visual representation of the model's performance by displaying the number of true positives, true negatives, false positives, and false negatives.

Finally, to identify the best performing model, each classifier's accuracy score was compared. The model with the highest accuracy was considered the most effective for this particular classification task. This thorough and systematic approach ensured the model's reliability and robustness, increasing confidence in its predictions and facilitating informed decision-making.

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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



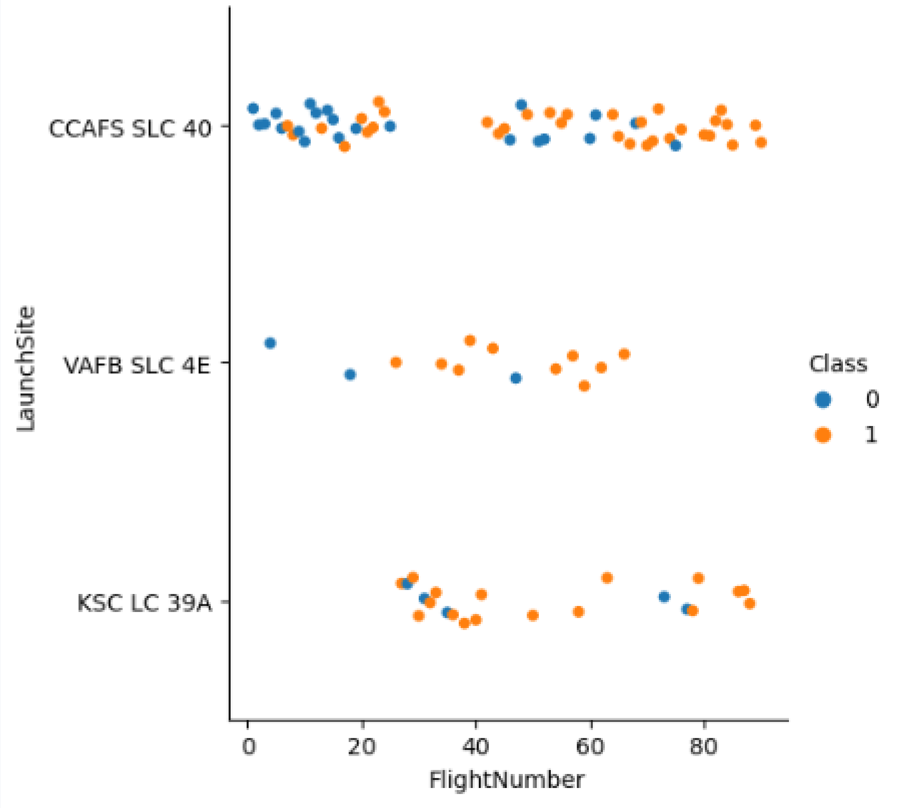
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

In our analysis of Flight Success as a function of Flight Numbers and Launch Sites we found nothing unusual.

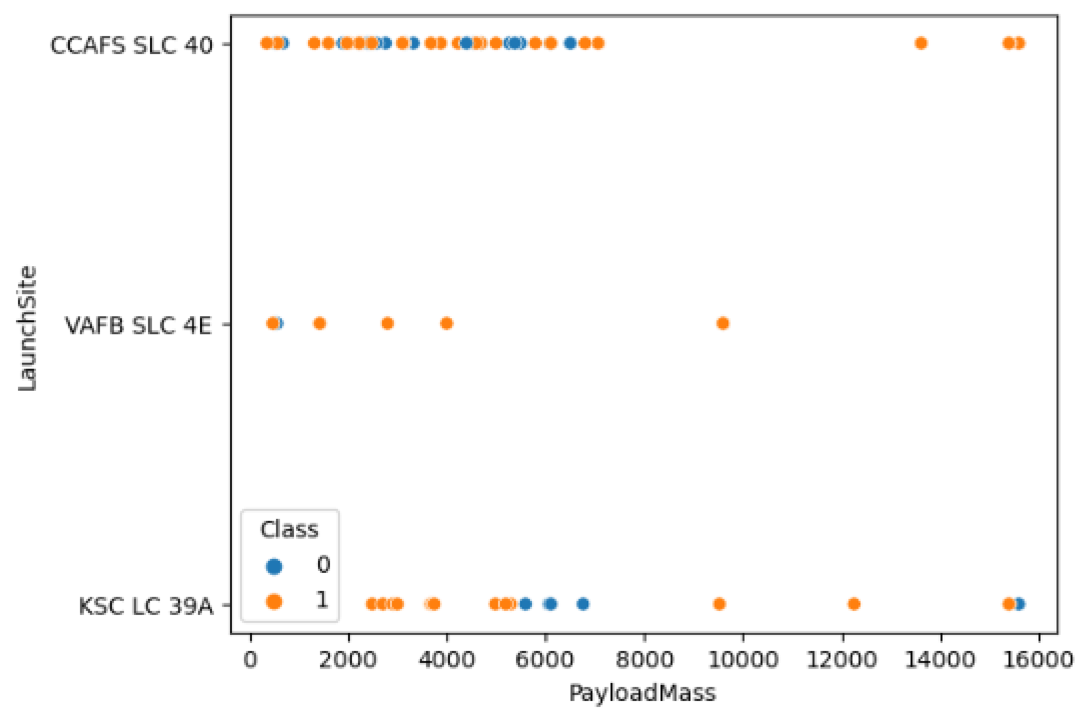
As would be expected, early flights were more failure prone than later ones since experience is gained over time.



Payload vs. Launch Site

A much more interesting pattern revealed itself when looking at Flight Success as a function of Payload Mass and Launch Site.

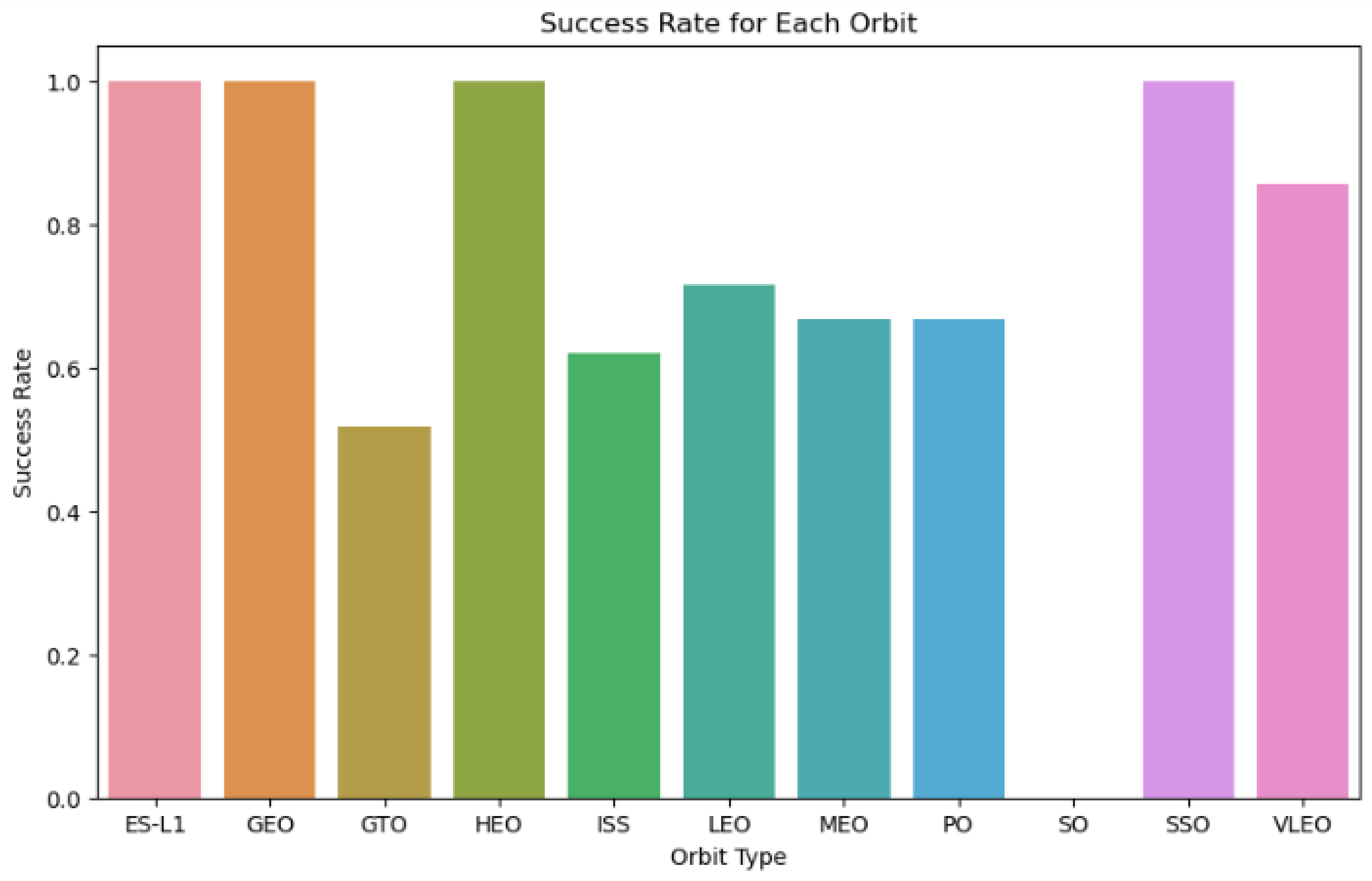
While there tended to be more failures at the lower half of payloads; as the weights went up the failures went down past the 8000 mass mark.



Success Rate vs. Orbit Type

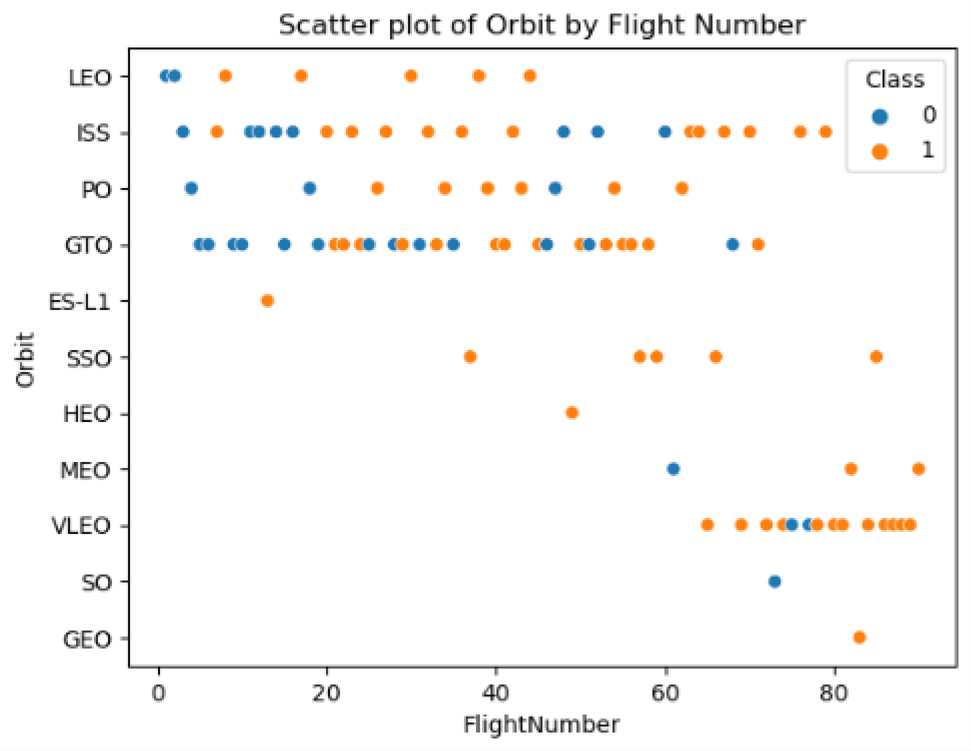
In our examination of Flight Success compared to Orbit Type, we found four types stood out above the rest:

- Lagrange Point 1 (ES-L1)
- Geostationary orbit (GEO)
- Highly Elliptical Orbit (HEO)
- Sun-Synchronous Orbit (SSO)



Flight Number vs. Orbit Type

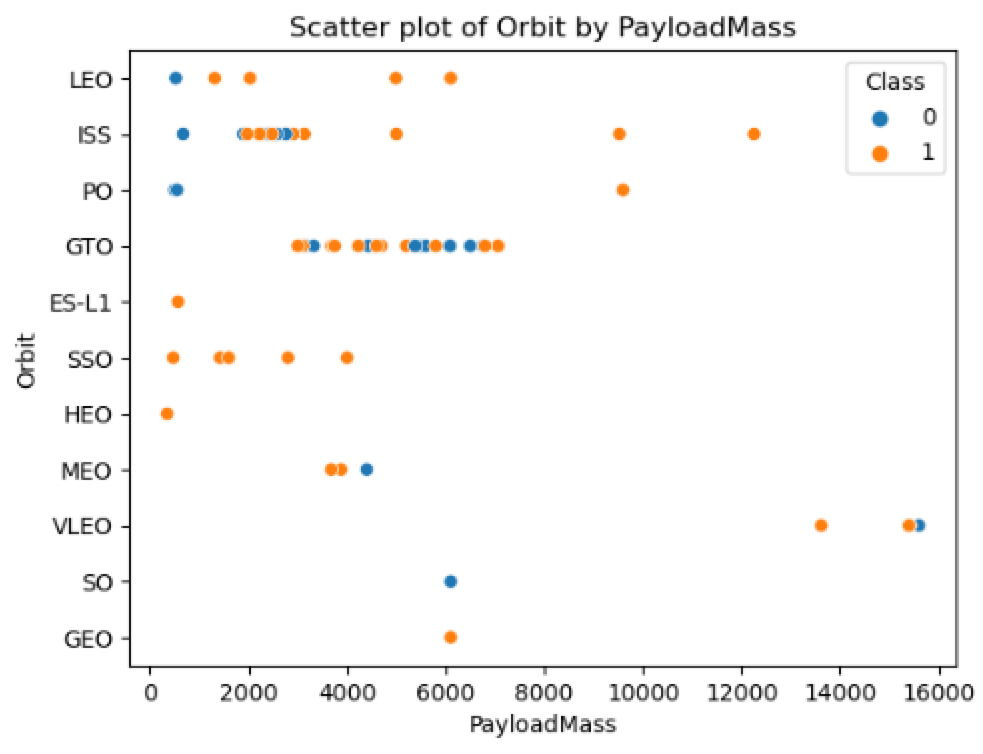
With regard to Flight Success as a function of Flight Number and Orbit Type, we found no discernable pattern other than the expected lower volume of failures in later flights due to experience gain.



Payload vs. Orbit Type

Payload Mass and Orbit Type have no discernable pattern with relation to Flight Success.

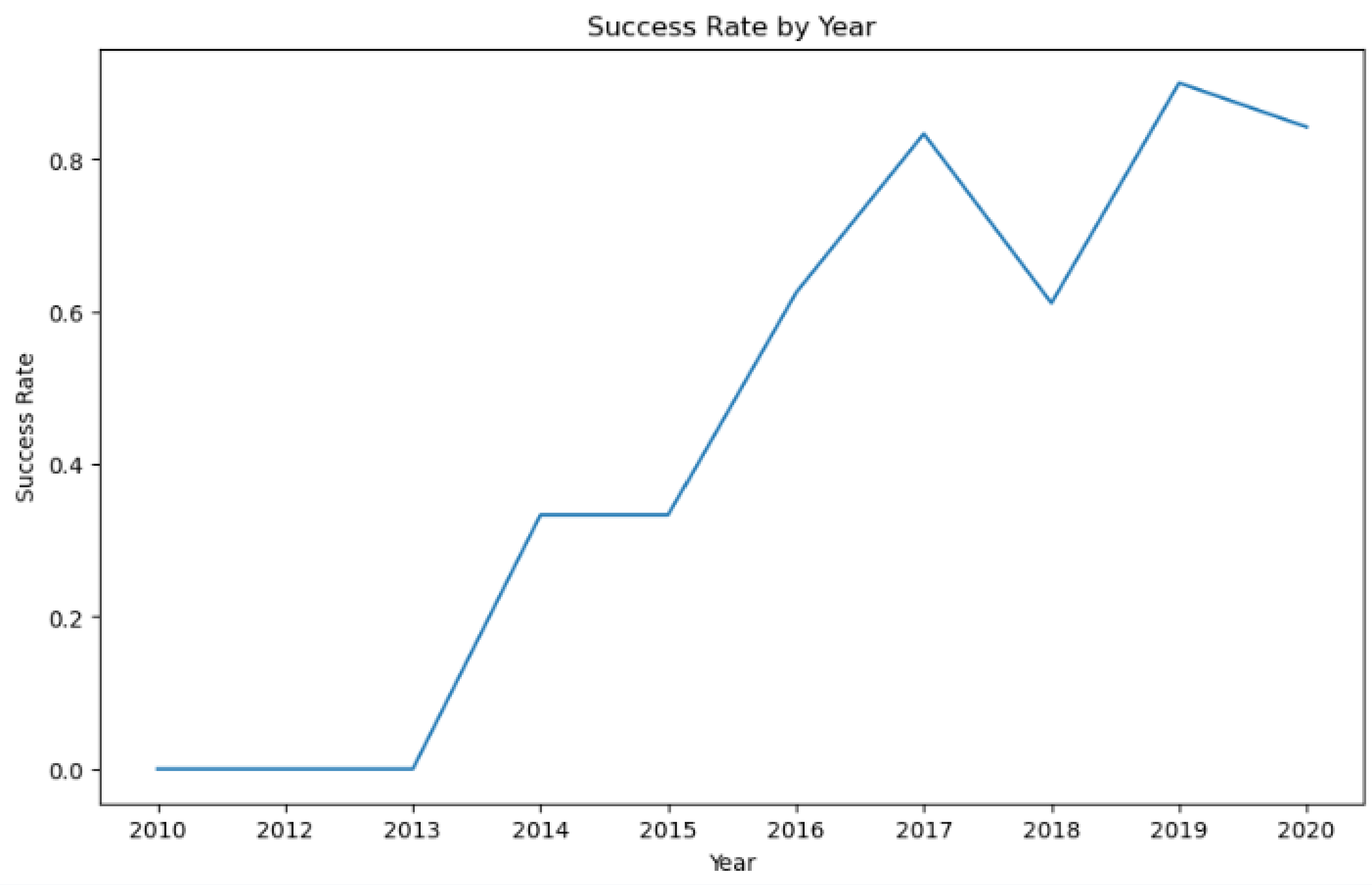
Interestingly, one of the smallest payloads is the only flight to the largest orbit (ES-L1).



Launch Success Yearly Trend

As expected, we can see a clear linear path toward more successful missions over time.

This, of course, is due to advancements in technology as well as experience gained launching earlier missions.



All Launch Site Names

In our analysis we looked at the following launch sites:

- CCAFS SLC-40, Cape Canaveral Space Launch Complex 40
- VAFB SLC-4E, Vandenberg Space Launch Complex 4
- KSC LC-39A, Kennedy Space Center Launch Complex 39A

Launch Site Names Begin with 'CCA'

For some random reason, we wanted to see the first five site entries that begin with ‘CCA’:

CCAFS SLC-40

CCAFS SLC-40

CCAFS SLC-40

CCAFS SLC-40

CCAFS SLC-40

Total Payload Mass

We calculated the total payload carried by boosters from NASA to be:

45,596 Kilograms

This equates to 100,521.97 Pounds or 50.26 Tons.

For perspective this is the approximate weight of 10 African Elephants.

Average Payload Mass by F9 v1.1

We also calculated the average payload mass carried by booster version F9 v1.1 to be 2,928.4 Kilograms. Which is a respectable payload relative to the total of 45,596 Kilograms for all payloads.

First Successful Ground Landing Date

We found the date of the first successful landing outcome on ground pad to be January the Eighth, in the Year of Our Lord, Two Thousand and Eighteen.

01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

We felt compelled to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

Total Number of Successful and Failure Mission Outcomes

- Naturally we would want to calculate the total number of successful and failure mission outcomes, which can be seen below:
  - Success, 99
  - Failure (in flight), 1
  - Success (payload status unclear), 1
  - No Data, 898

Boosters Carried Maximum Payload

- Following are the the names of the boosters which have carried the maximum payload mass for missions:
  - 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

- Following is a list of the failed landing outcomes on a drone ship, their booster versions, and launch site names for in year 2015:
  - 10, 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

Finally, we rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order:

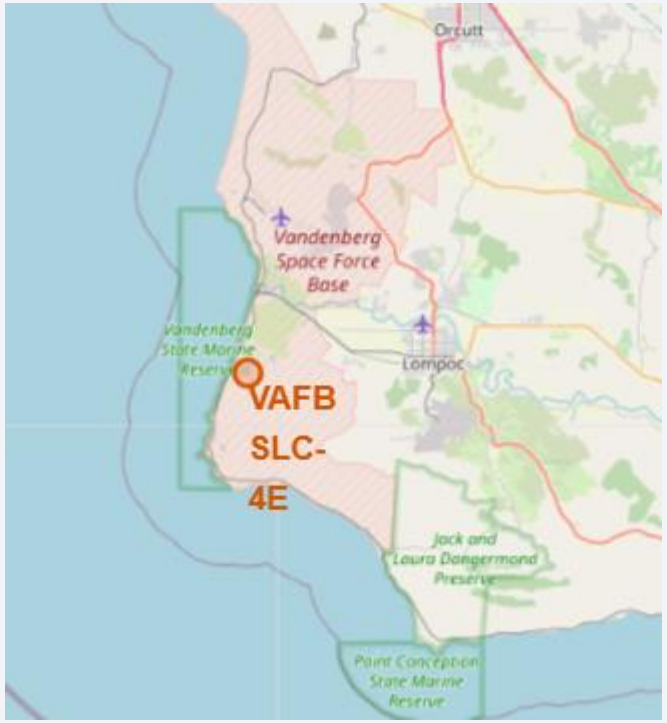
- No attempt, 10
- Success (ground pad), 5
- Success (drone ship), 5
- Failure (drone ship), 5
- Controlled (ocean), 3
- Uncontrolled (ocean), 2
- Precluded (drone ship), 1
- Failure (parachute), 1

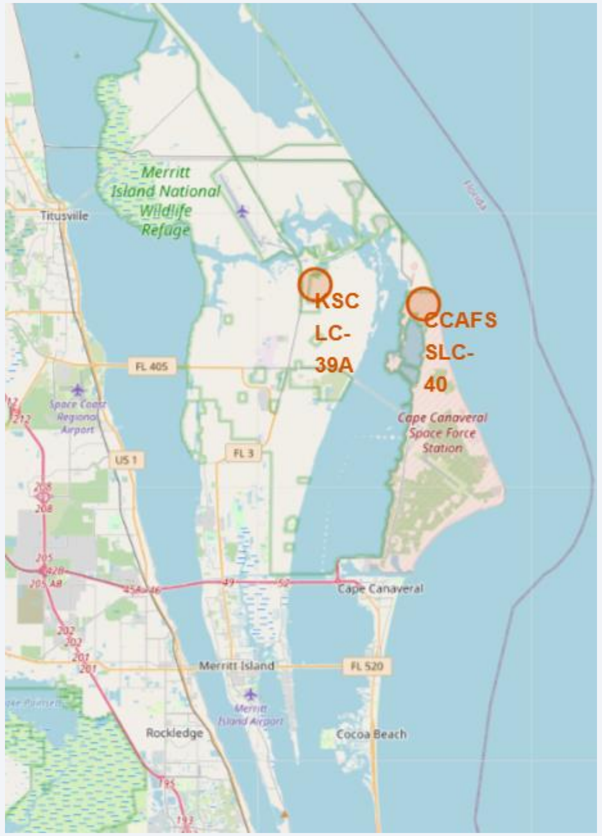
Section 3

Launch Sites Proximities Analysis

Launch Site Location Indicators

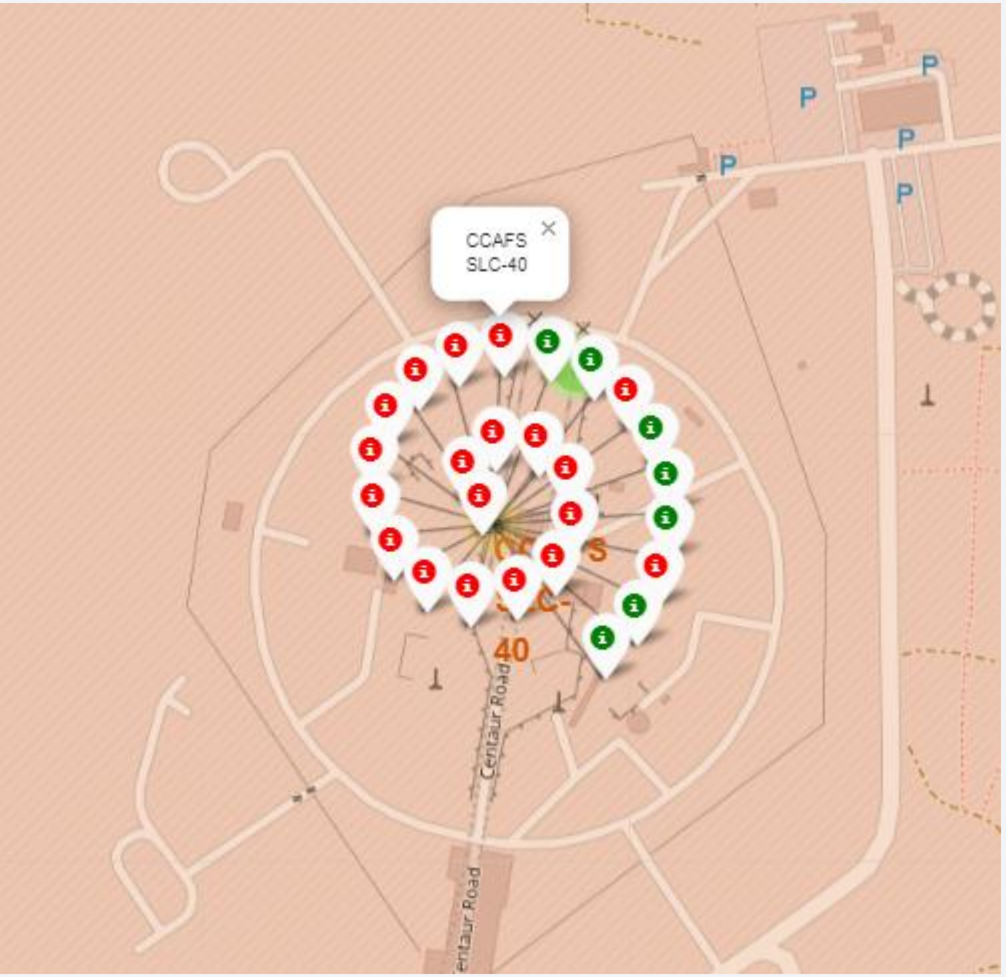
Here you can see all launch sites on a map, this gives a clear view of their geographical distribution, providing us with an immediate context of the environmental factors they might be subjected to, like weather and population density.


Orcutt
Vandenberg Space Force Base
Vandenberg State Marine Reserve
Lompoc
VAFB SLC-4E
Jack and Laura Dangermond Preserve
Point Conception State Marine Reserve

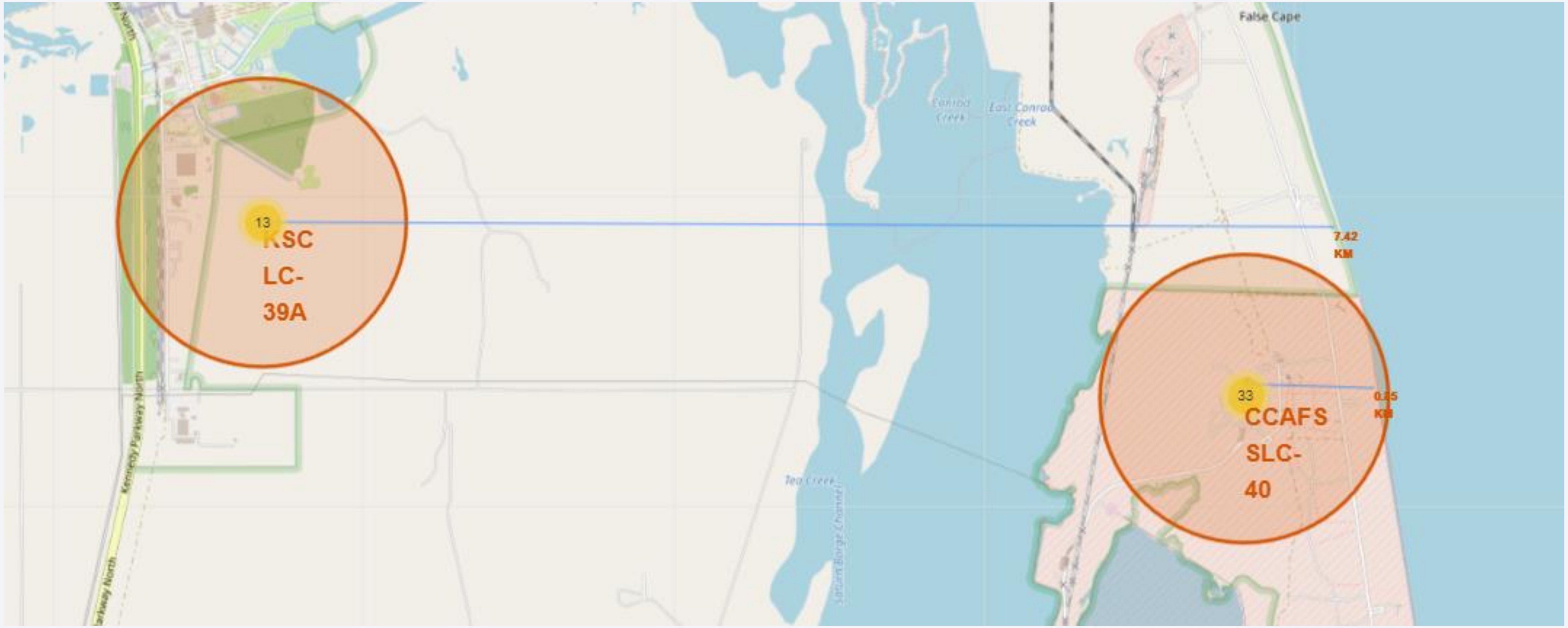


Launch Site Success / Fail Indicators

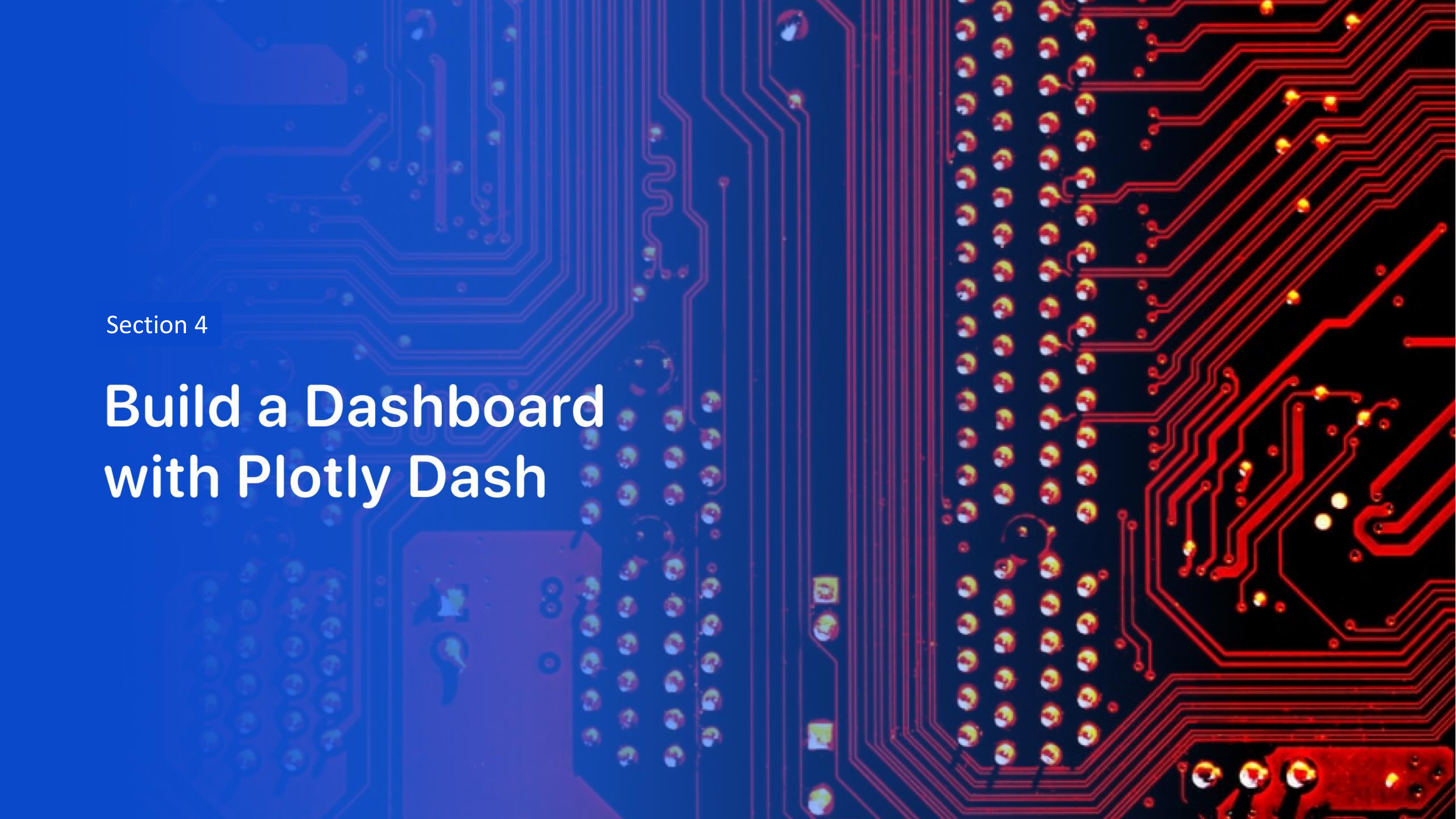
We added markers to indicate the success or failure of launches at each site. By utilizing color-coded markers (green for success, red for failure), we aimed to provide a clear, visual indicator of the success rate for each site.



Launch Site Proximity to Significant Landmarks



Finally, we calculated and displayed the distances between each launch site and its significant proximities, such as coastlines, and major cities. By drawing lines between the launch site and these points of interest, we visually emphasized the spatial relationship.



Section 4

Build a Dashboard with Plotly Dash

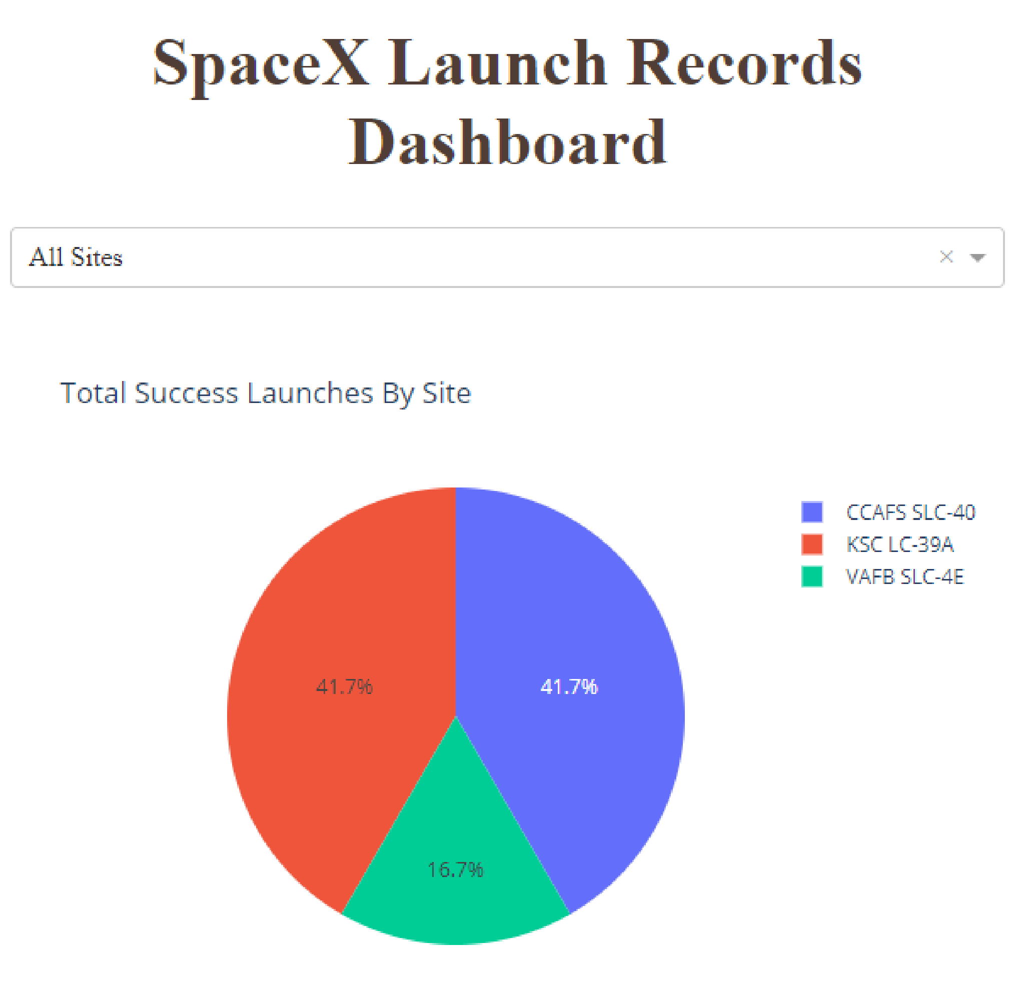
Dashboard: Total Success Launches by Site

The dashboard initially presented a dropdown list to allow selection of a specific Launch Site or an option to consider all sites.

Additionally, to illustrate the success rate of launches, a pie chart was incorporated into the dashboard.

When a specific site was selected from the dropdown, the pie chart showed the ratio of Successful to Failed launches for that site. When 'All Sites' option was chosen, it displayed the total successful launches for all sites.

This interactive graphical representation provided a clear overview of launch outcomes, aiding in the identification of performance trends and potential issues at specific sites.



SpaceX Launch Records Dashboard

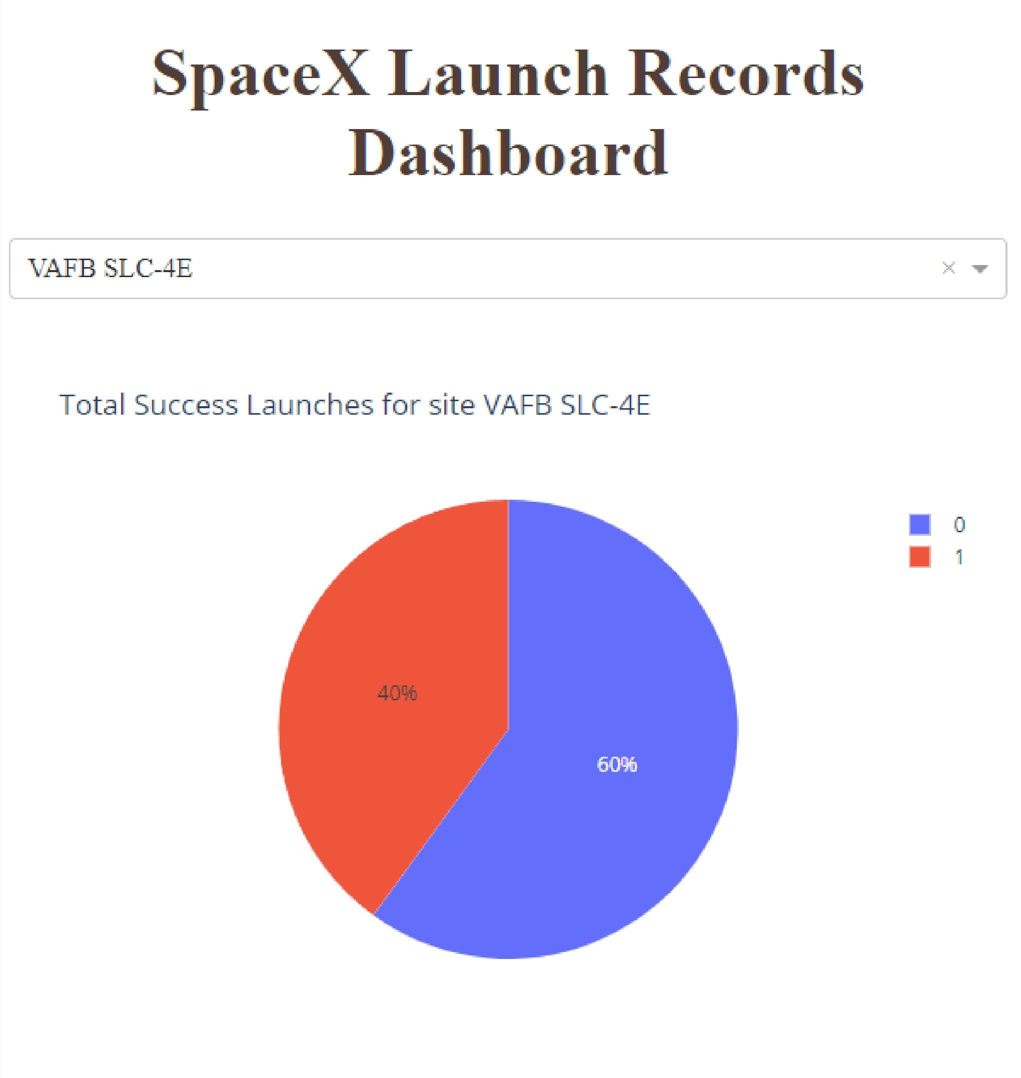
All Sites

Most Successful Launches

In this snapshot we can see the site with the most successful launches (40%): Vandenberg Space Launch Complex 4E.

Space Launch Complex 4 (SLC-4) is a launch and landing site at Vandenberg Space Force Base, California, U.S.

It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 (LZ-4) for SpaceX landings.



SpaceX Launch Records Dashboard

VAFB SLC-4E

Total Success Launches for site VAFB SLC-4E

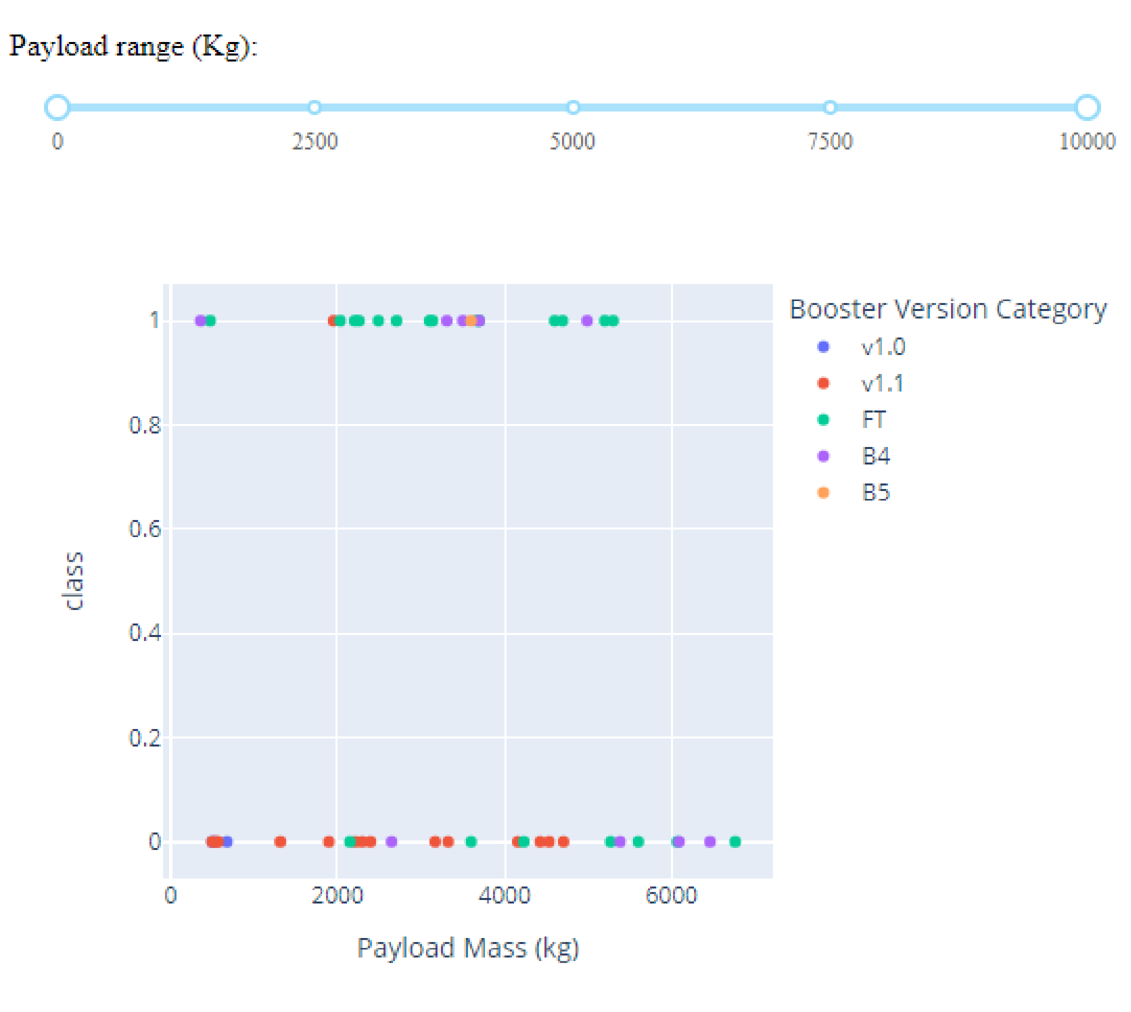
Outcome vs. Payload & Booster Version

This final component of the dashboard was a scatter plot that depicted the correlation between payload mass and launch success.

The plot was updated based on the site chosen from the dropdown and the payload range selected.

This enabled executives to explore potential relationships between the payload mass and the success of the launches.

The data points in the scatter plot were color-coded according to the Booster Version Category, offering further insight into the impact of different booster versions on launch success.



Section 5

Predictive Analysis (Classification)

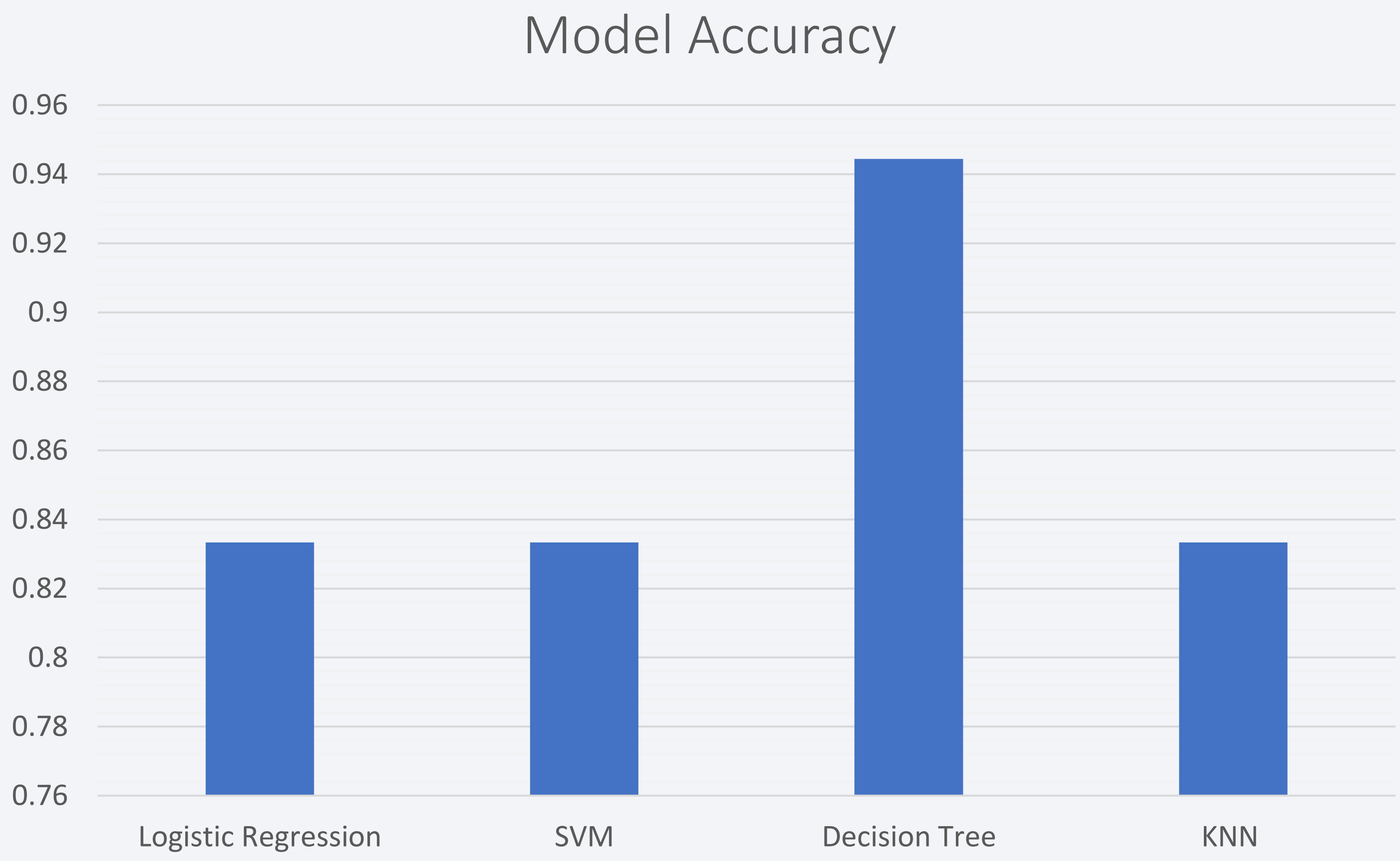
Classification Accuracy

Four distinct machine learning models, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN), underwent performance comparison.

Among all, the Decision Tree model led the pack, showcasing an accuracy of 94.44%.

In contrast, Logistic Regression, SVM, and KNN demonstrated matching accuracy levels of 83.33%.

Given these results, the Decision Tree model emerges as the top choice for future predictions, owing to its accuracy advantage.



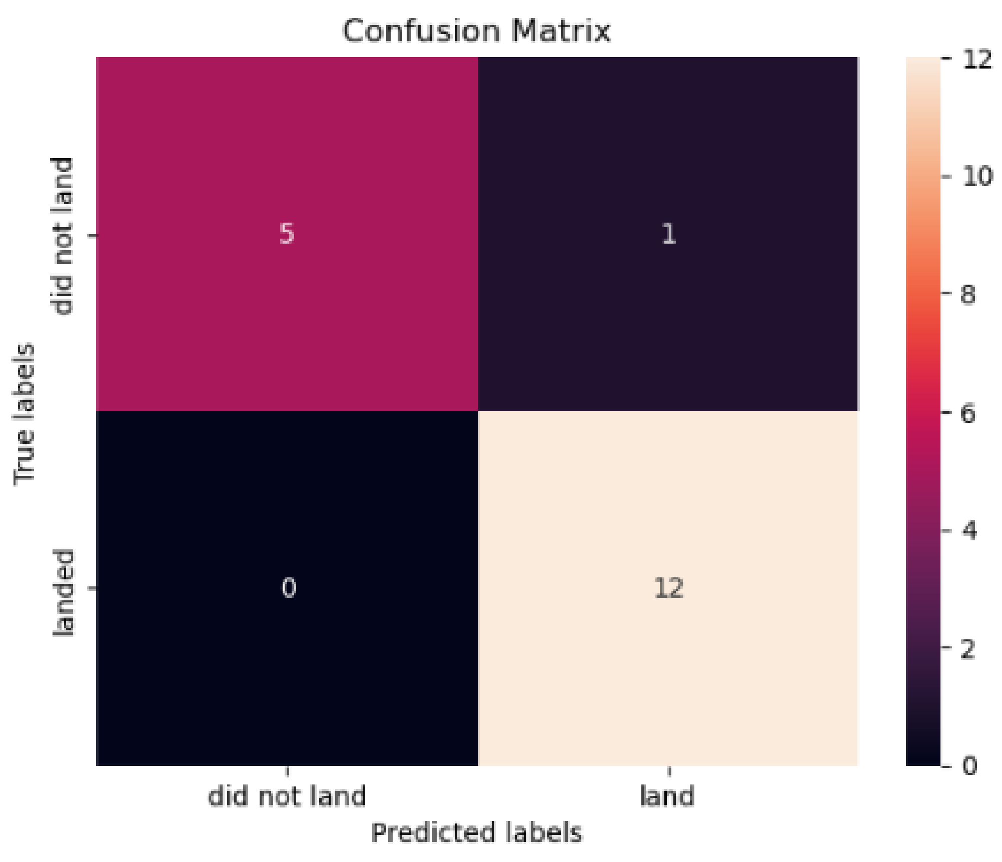
Confusion Matrix

The Decision Tree model's effectiveness was tested by evaluating its predictions. The results can be distilled into four numbers, arranged in a matrix. This matrix lets us understand the accuracy of our predictions.

First, we take a look at the top left corner of the matrix, which represents instances where the model correctly predicted a rocket would not land, totaling 5 such instances. On the other hand, the top right corner represents cases where the model incorrectly predicted that a rocket would land, but it did not, amounting to 1 such instance.

Focusing on the bottom row, the left corner accounts for scenarios where the model incorrectly predicted a rocket would not land, but it did. Fortunately, there were no instances of this type. Finally, the lower right corner represents instances where the model correctly predicted that a rocket would land, with a total of 12 such successful predictions.

The model demonstrated strong performance (94.44%) in its predictions, particularly in predicting successful rocket landings.



EDA Conclusions

- While there tended to be more failures at the lower half of payloads; as the weights went up the failures went down past the 8000 mass mark.
- In our examination of Flight Success compared to Orbit Type, we found four types stood out above the rest:
  - Lagrange Point 1 (ES-L1)
  - Geostationary orbit (GEO)
  - Highly Elliptical Orbit (HEO)
  - Sun-Synchronous Orbit (SSO)
- As expected, we can see a clear linear path toward more successful missions over time.
- There is significantly more missing data on launches than data on success or failures.

Launch Site Proximities Conclusions

- All launch sites are located relatively close to the coast.
- No launch site is close to any major city.
- All launch sites are close to resources that can be used to carry heavy loads on them (i.e. roads, railroads, etc…)

Dashboard Conclusions

- All launch sites had a less than 50% success rate.

- The best success rate for any launch site was 40%.

- B4 and FT booster types had a much higher success rate than all other boosters.

Predictive Analytics Conclusions

- All models exceeded 80% accuracy.

- The Decision Tree model is the clear choice at 94% accuracy.

Appendix

All artifacts for this project can be found on GitHub at:

https://github.com/suspicious-cow/IBM-Applied-Data-Science-Capstone

Thank you!