IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Huamantalla
2025–04–17

Outline

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

Executive Summary

- Summary of methodologies

This report uses Falcon 9 data obtained through the SpaceX API and web scraping techniques. The data was cleaned and transformed for analysis, including removal of null values and coding of categorical variables. Exploratory data analysis was applied using visualization tools and SQL queries. In addition, interactive visualizations were developed with Folium and Dash, and classification models were implemented to perform predictive analysis.

- Summary of all results

Exploratory analysis revealed that Falcon 9 first stage landing success has improved over time and with increasing numbers of launches. Patterns were identified that associate a higher success rate with certain orbits and specific payload masses, with orbits such as ES-L1, GEO, HEO and SSO showing 100% success. It was also observed that the launch site influences the payload mass, with some sites not handling heavy payloads. Through SQL analysis, the main launch sites were identified and the loads carried by different versions of the rocket were analyzed. The first successful on-platform recovery occurred in 2015. Finally, four classification models were applied in the predictive analysis. All achieved similar performance, being the decision tree model the one that presented the highest accuracy in cross-validation.

Introduction

- Project background and context

SpaceX has revolutionized the aerospace industry by introducing reusable rockets, significantly reducing launch costs and increasing operational efficiency. One of its most prominent launchers is the Falcon 9, whose performance and resilience have been the subject of great technical and commercial interest. This project aims to analyze the factors that influence the successful landing of the Falcon 9 first stage, considering variables such as orbit type, payload mass, launch site and flight history.

- Problems you want to find answers

From these data, we seek to answer key questions such as: what conditions favor landing success, what are the characteristics of the most successful missions, and is it possible to predict the outcome of a mission from these factors?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - SpaceX launch data was obtained from two sources:
    - SpaceX REST API → https://api.spacexdata.com/v4/rockets/
    - Web scraping related wiki pages → https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - A subset of the original data set was used, null values were removed, and coding was applied to categorical variables.
  - Subsequently, the data were divided into training and test sets, and feature engineering was applied along with balancing techniques when necessary.

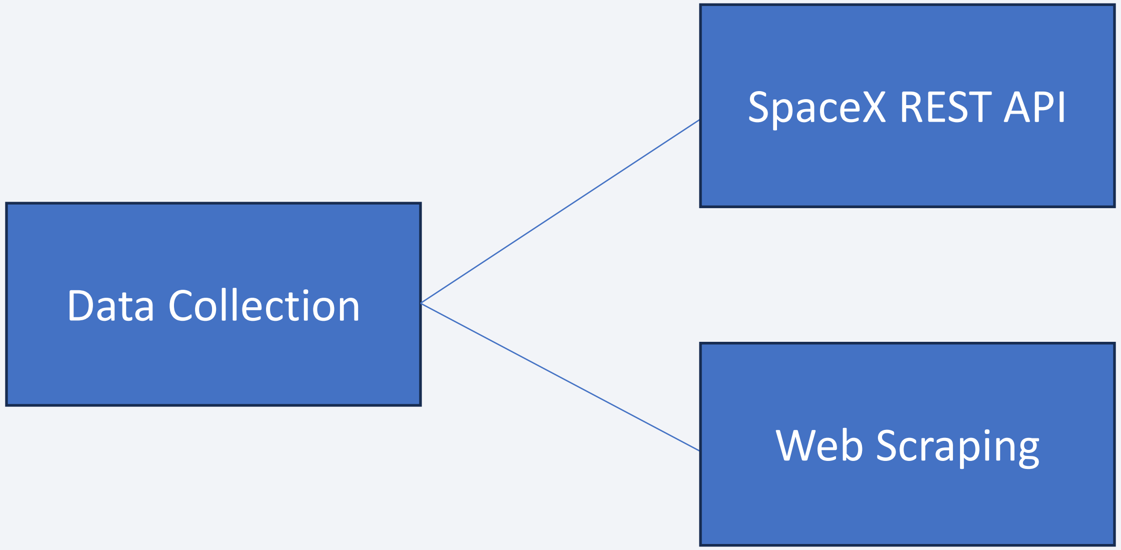
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models such as Logistic Regression, Decision Tree, among others, were built and adjusted, evaluating them using metrics such as F1-score, precision y recall to select the best performance.

Data Collection

- The dataset was collected using the SpaceX REST API y web scraping related wiki pages



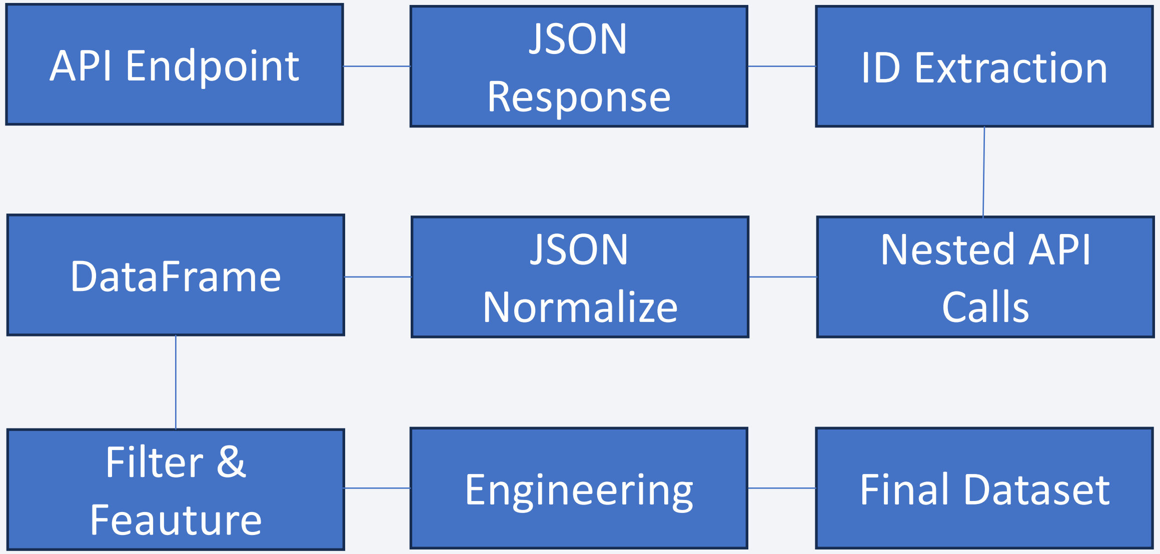
Data Collection

SpaceX REST API

Web Scraping

Data Collection – SpaceX API

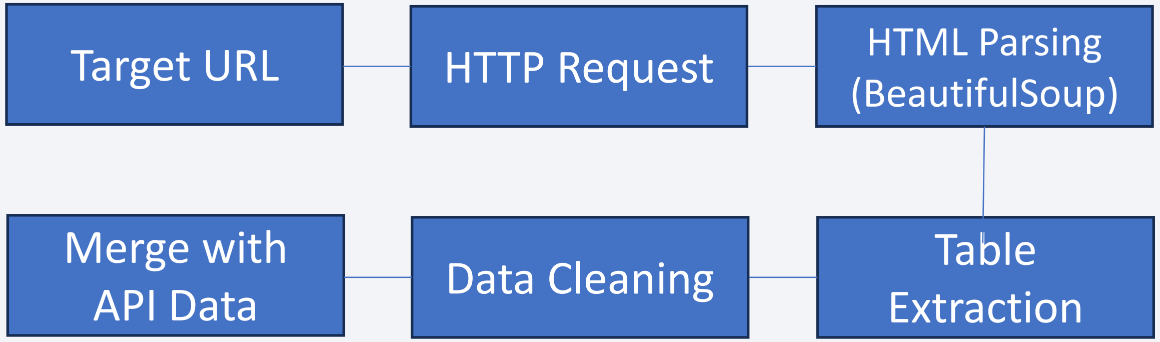
- The dataset was collected using the SpaceX REST API, which provides structured JSON responses.



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

Data Collection - Scraping

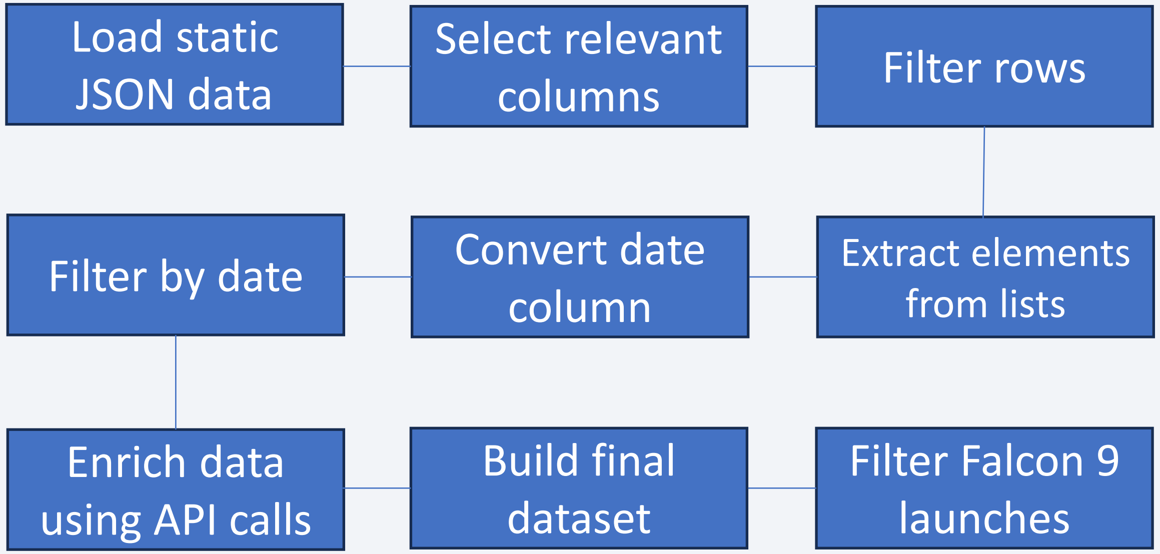
- To obtain additional information, web scraping was used.



https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- To prepare the dataset for analysis, we first loaded a subset of the SpaceX launch data using a static JSON file, without applying variable normalization.



Load static JSON data

Select relevant columns

Filter rows

Extract elements from lists

Convert date column

Filter by date

Enrich data using API calls

Build final dataset

Filter Falcon 9 launches

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

EDA with Data Visualization

The following shows which graphs and why they were used:

- Catplot → Used to visualize the distribution of categorical or discrete data: FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, Payload Mass vs Launch Site, FlightNumber vs Orbit type y Payload Mass vs Orbit type.
- Barplot → Used to display average Success Rate by Orbit Type.
- Lineplot → To show how Success Rate changes over time.

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

The following is a summary of the SQL queries performed:

- Successful launches were filtered to analyze the conditions that favor a successful landing.
- A date range filter was applied to observe trends or compare phases within a specific period.
- Data was grouped by launch site to identify the most frequently used locations.
- Results were sorted by count or date to highlight the most active sites or the temporal evolution of launches.
- Aggregate functions were used such as total launches and average payload to summarize key numerical information.
- Multiple conditions were combined to filter more specific data, like successful missions funded by a specific customer.

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following is a summary of the construction of the interactive map with Folium:

- A map was initialized centered at NASA Johnson Space Center using Folium.
- A circle and label were added at the NASA location to visually mark it and display its name.
- A loop added circles and labels for each launch site in the dataset, showing location and name using coordinates.
- A MarkerCluster was created to group all launch markers efficiently, reducing clutter on the map.
- Each launch was represented by a marker, and the color indicated whether it was successful or not.
- Mouse tracking functionality was added to display the latitude and longitude of the cursor on the map.
- Markers were added to show both the launch site and the nearest coastline.
- A label showed the distance from the launch site to the coastline directly on the map.
- A polyline was drawn between the launch site and the coastline to visualize their connection.
- Additional distance markers and lines were drawn from the launch site to other points like a city, a railway, and a highway, each labeled with the distance.

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

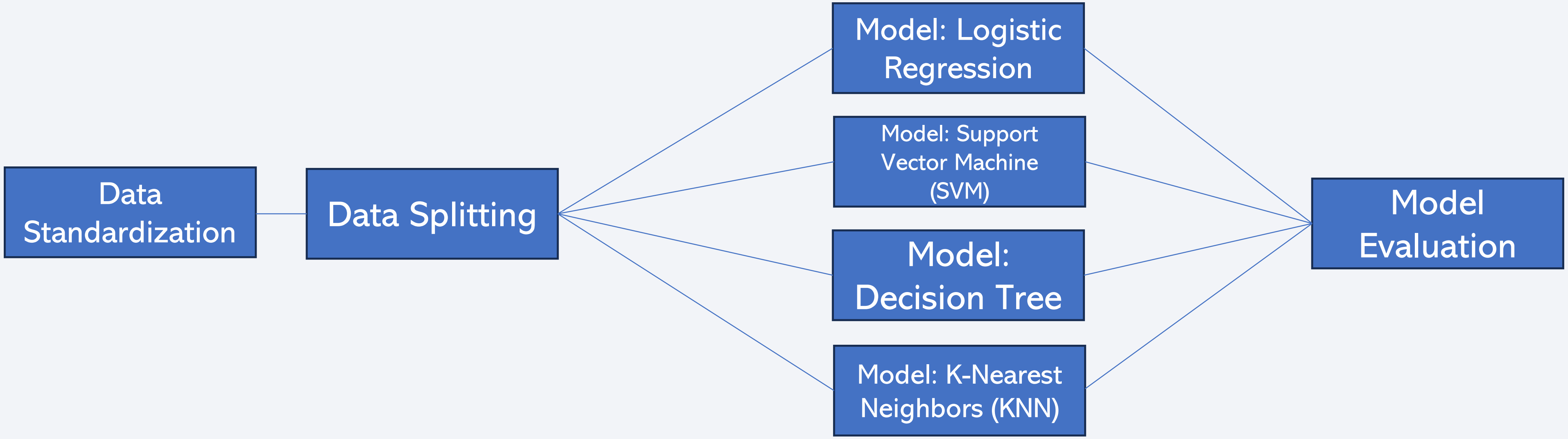
The following is a summary of the construction of the dashboard with Plotly Dash:

- An interactive web dashboard was developed to visualize the performance of SpaceX launches.
- A launch site filter allows users to analyze results either globally or by specific location.
- A dynamic pie chart is used to show:
    - Total successful launches per site (overview).
    - Success vs. failure rates at each site (detailed analysis).
- A payload range slider enables filtering by payload weight to examine its impact on launch success.
- A scatter plot visualizes the relationship between payload mass and success, segmented by booster version.
- Charts update automatically based on selected filters, making pattern identification and performance insights more accessible.

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/dash-app.py

Predictive Analysis (Classification)

- All four machine learning models (Logistic Regression, SVM, Decision Tree, KNN) showed similar and solid performance, with test accuracies around 83.3%, making them equally viable for classification under current conditions.


Data Standardization
Data Splitting
Model: Logistic Regression
Model: Support Vector Machine (SVM)
Model: Decision Tree
Model: K-Nearest Neighbors (KNN)
Model Evaluation

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

From the exploratory analysis of the data, with Pandas and Matplotlib, the following results were obtained on Falcon 9:

- It was found that as the number of flights increased, the payload mass increased as did the first stage landing success.
- It was found that as the number of flights increased for all three launch sites, the first stage landing success increased.
- The VAFB-SLC launch site was found to not launch rockets with high payload mass (greater than 10000).
- Launches that went to ES-L1, GEO, HEO and SSO orbits were found to have 100% first stage landing success. While for the GTO orbit, it provided the least success.
- It was found that for heavy payloads the PO, LEO and ISS orbits had higher first stage landing success than the other orbits.
- The success rate was found to be increasing from 2013 to 2020.

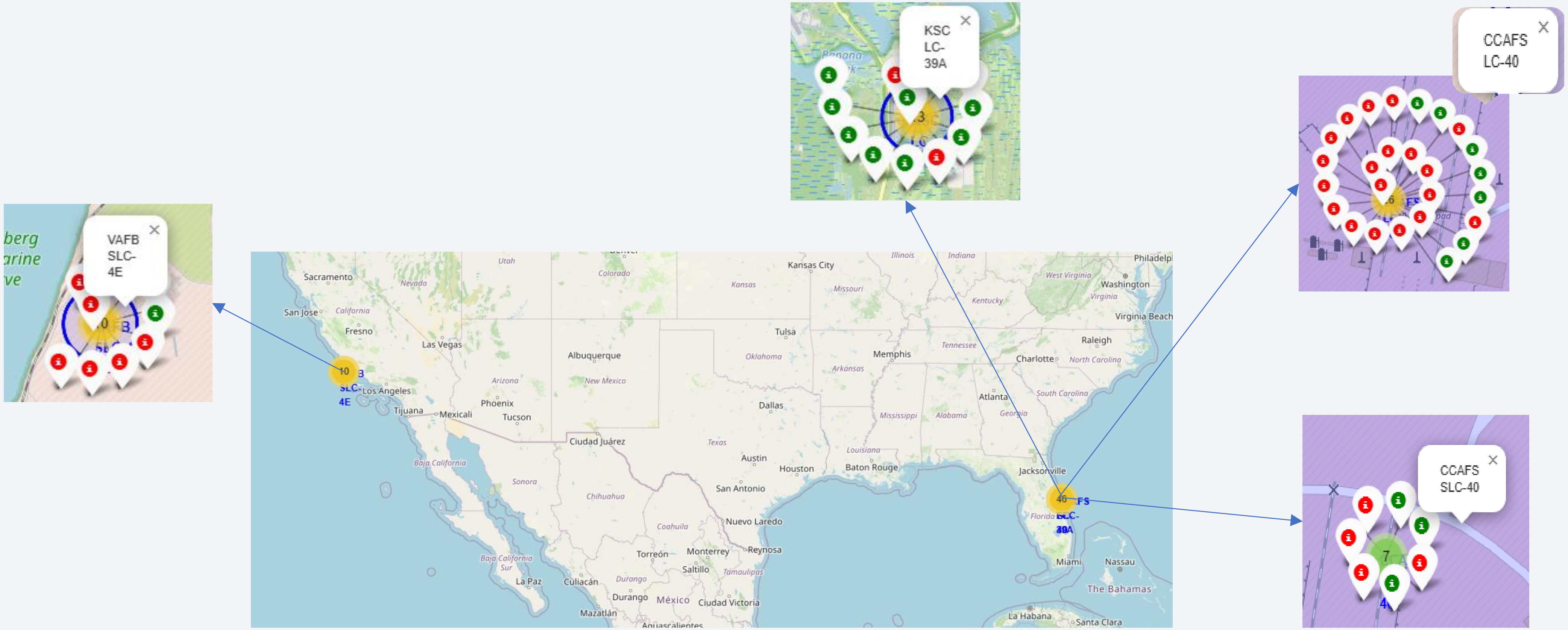
Results

From the exploratory analysis of the data, with SQL, the following results were obtained on Falcon 9:

- Four space mission launch sites were found: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.
- The total mass of payload carried by NASA launched rockets (CRS) was found to be 45,596 tons.
- The average payload mass carried by the F9 v1.1 booster version was found to be 2,535 tons.
- The date on which the first successful landing occurred on the landing pad was found to be 2015-12-22.
- The boosters that have been successful on the unmanned spacecraft and have a payload mass greater than 4000 but less than 6000 were found to be F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.
- The total number of missions successfully flown was found to be 100 and failed 1.
- Booster versions found to have carried the maximum payload mass were: 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, and F9 B5 B1049.7.

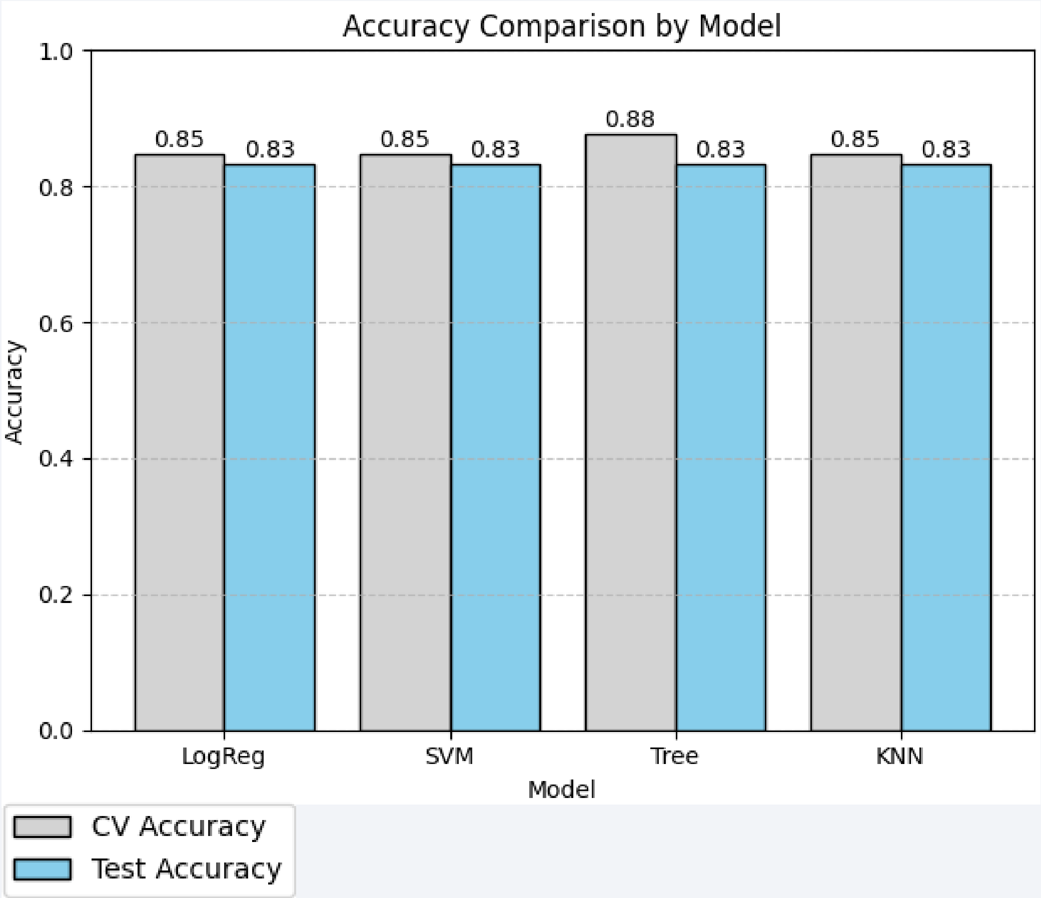
Results

- In the interactive visual analysis with Folium, we create markers on the map for all launch records. If a launch has been successful, we mark it with green and if it has failed with red.



Results

- In the predictive analysis we use 4 classification models: Logistic Regression, Support Vector Machines, Decision Tree and K-Nearest Neighbors. It was found that all models have the same level of accuracy in the test. The Decision Tree model has the highest accuracy in cross-validation.



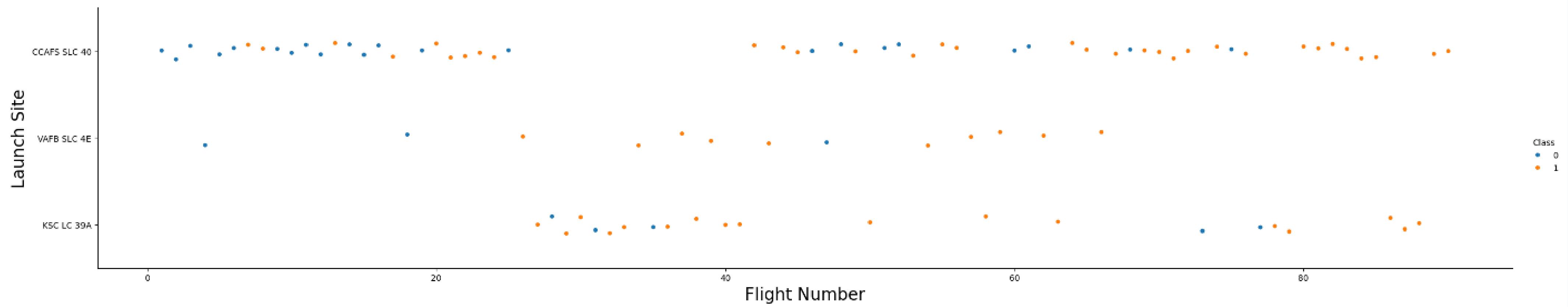
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Looking at the graph, two key points about SpaceX launches stand out:

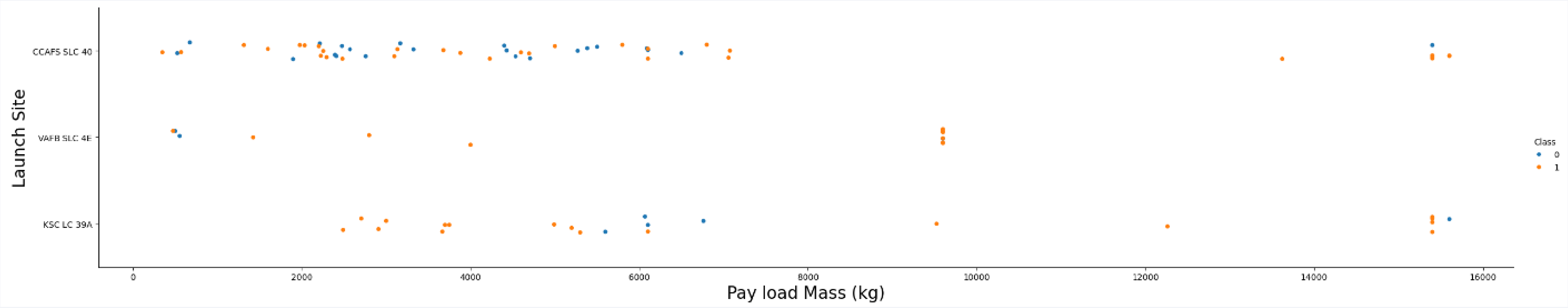
- First, landing success varies by site: CCAFS SLC-40 and KSC LC-39A show mostly successful landings (orange), while VAFB SLC-4E has a more even mix of successes and non-successes (blue).
- Second, the launch frequency differs: CCAFS SLC-40 is the site with the most launches represented, followed by KSC LC-39A , with VAFB SLC-4E having the least activity in this data set.



Payload vs. Launch Site

Observing this graph, which relates payload mass to launch site and landing outcome, we can draw two important conclusions:

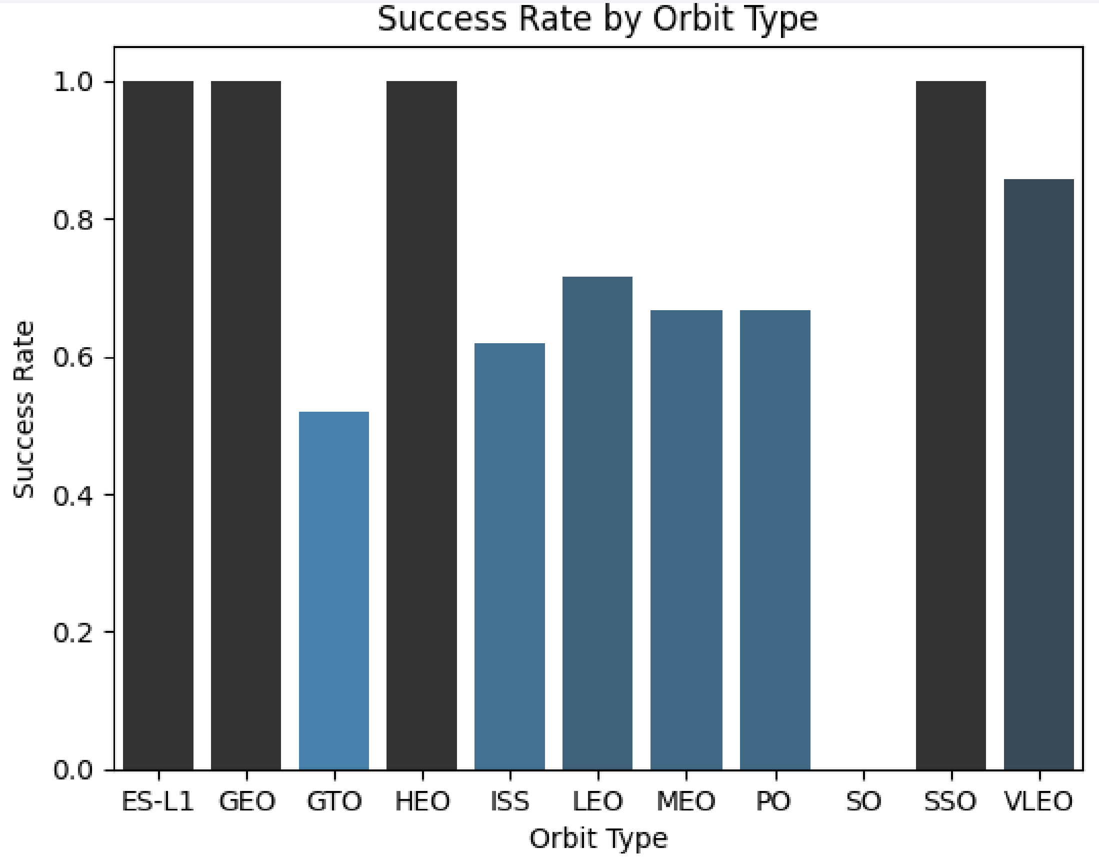
- Sites and Payloads: KSC LC-39A and CCAFS SLC-40 handle heavier payloads; VAFB SLC-4E focuses on light/medium ones.
- Success vs. Mass/Site: Successful heavy payload landings are possible (KSC/CCAFS) but not guaranteed. Across lighter/medium loads, all sites show mixed success/failure, with VAFB showing a balance.



Success Rate vs. Orbit Type

This bar chart shows us the Launch Success Rate, broken down by the Orbit Type the launches were targeting.

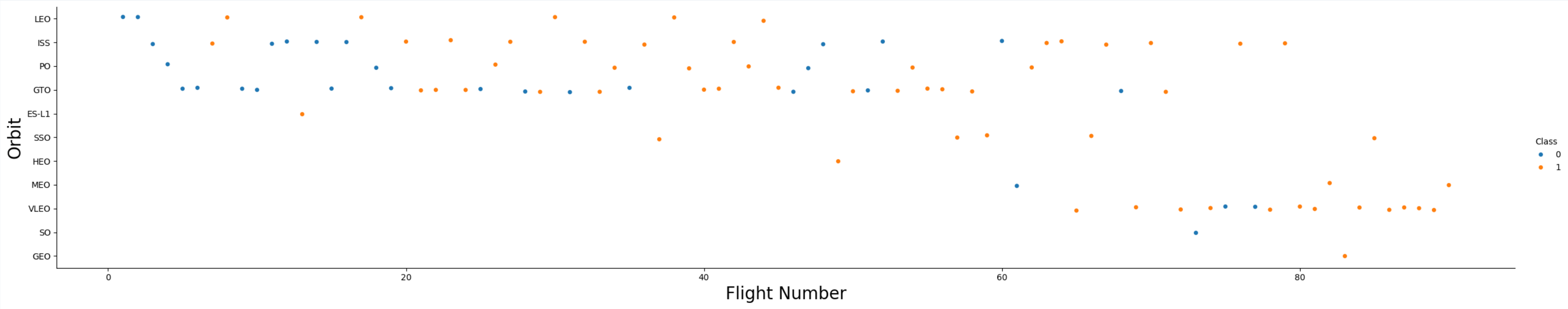
- High Success in Specific Orbits: We observe that several orbits, such as GEO, HEO, and SSO, show a 100% success rate (value of 1.0 on the chart) according to this data, indicating high reliability for reaching these destinations.
- Challenge in GTO: In contrast, the Geostationary Transfer Orbit (GTO) presents the lowest success rate of all, sitting slightly above 50%. This highlights it as the orbit with the greatest difficulties or history of failures in this dataset.



Flight Number vs. Orbit Type

This scatter plot shows us how launches to different orbits are distributed over time (represented by the Flight Number). Additionally, the color of each point indicates the mission or landing outcome: orange for success (Class 1) and blue for failure (Class 0).

- General Improvement of Success Over Time: A visual trend is observed: as we move along the flight number (towards the right), there appears to be a higher concentration of orange dots (successes) compared to blue ones (failures) across most orbits. This suggests a general improvement in the success rate as the launch program has progressed.
- Popularity and Evolution of Orbits: The ISS, GTO, and VLEO orbits are consistently popular destinations throughout the flight sequence. Interestingly, the SSO orbit shows a notable increase in its frequency as a destination in more recent missions (higher flight numbers).

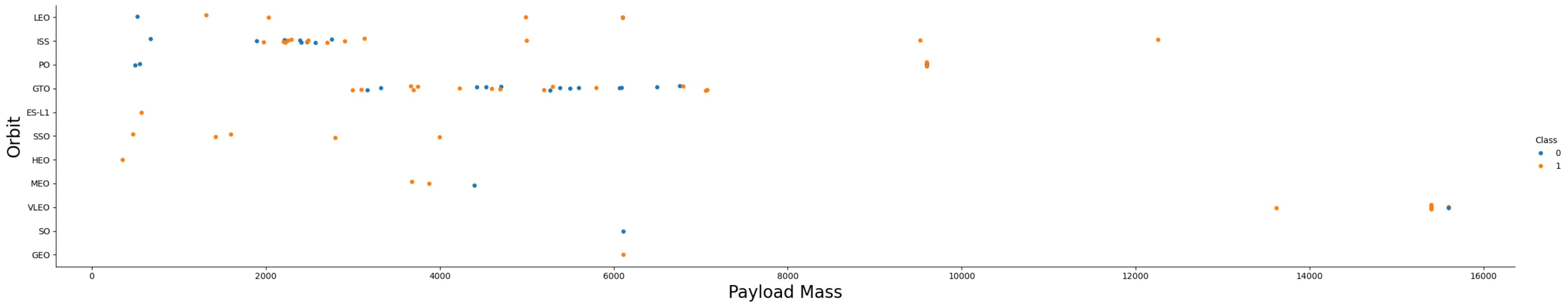


Payload vs. Orbit Type

This scatter plot illustrates the relationship between the Payload Mass being launched and the destination Orbit. The color of each point indicates whether the mission (or its landing phase) was successful (Class 1, orange) or a failure (Class 0, blue).

- Payload Ranges per Orbit: Different orbits tend to receive payloads of varying weights. The GTO and VLEO orbits are destinations for a wide range of masses, including the heaviest ones (over 10,000 kg). In contrast, orbits like ISS receive loads in a medium range, while LEO and SSO, in this visualization, are mainly associated with lighter or medium payloads.

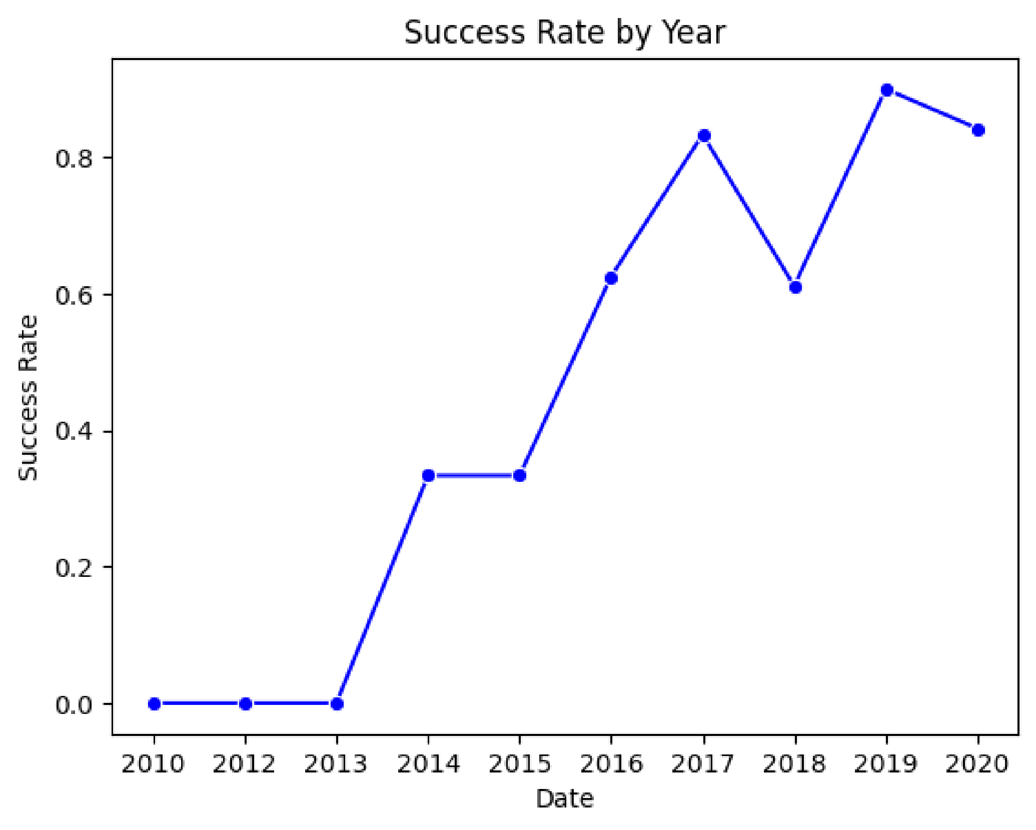
- Success and Challenges by Orbit/Mass: Success varies notably. The GTO orbit shows a considerable number of failures (blue dots) distributed across almost its entire mass range, standing out as a particularly challenging destination. On the other hand, ISS and SSO predominantly show successes (orange) for the payload masses associated with them here. It's important to note that even for the heaviest payloads (>14,000 kg), destined for VLEO and GTO, both successes and failures are recorded.



Launch Success Yearly Trend

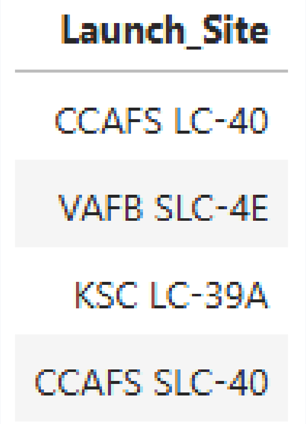
This line graph shows us the evolution of the Success Rate over the years, from 2010 to 2020. We can observe several distinct phases in this trend:

- 2010-2013: Low success (0%). Challenging start.
- 2013-2017: Strong success increase, reaching >80% in 2017. Notable improvement.
- 2017-2020: Fluctuations, but generally high success (~80-90%). Volatility after the peak.



All Launch Site Names

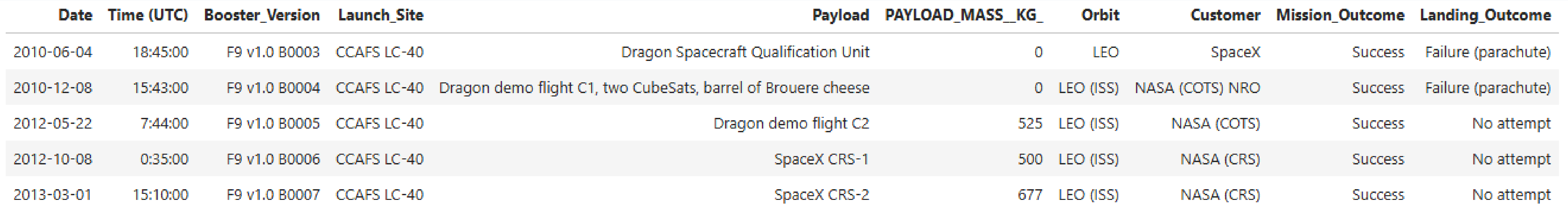
- Based on the data, there are 4 launch sites as observed in the following figure:



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

Launch Site Names Begin with 'CCA'

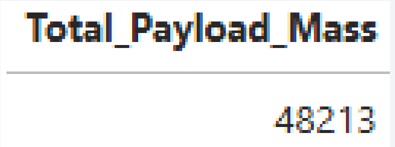
- The table presents the first 5 records whose launch sites begin with ‘CCA’.



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

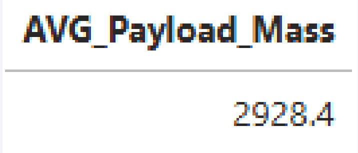
- The table presents the total payload mass carried by NASA launched rockets (CRS) in kg.



Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

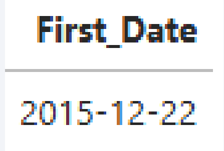
- The table presents the average payload mass carried by booster version F9 v1.1 in kg.



AVG_Payload_Mass
2928.4

First Successful Ground Landing Date

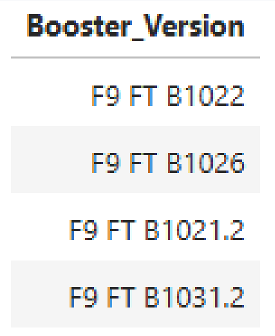
- The table presents the date when the first succesful landing outcome in ground pad was acheived.



First_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

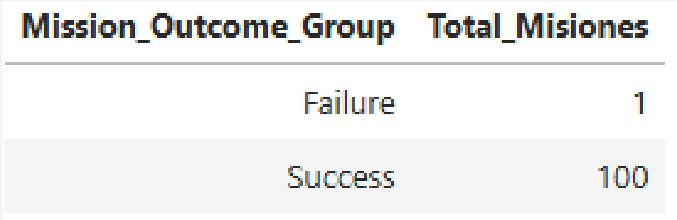
- The table presents the names of the boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 6000 kg.



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

Total Number of Successful and Failure Mission Outcomes

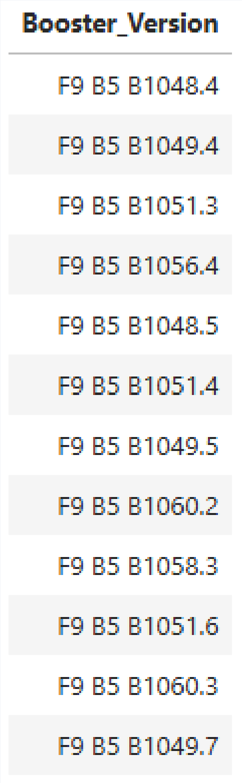
- The table presents the total number of successful and failure mission outcomes.



Mission_Outcome_Group	Total_Misiones
Failure	1
Success	100

Boosters Carried Maximum Payload

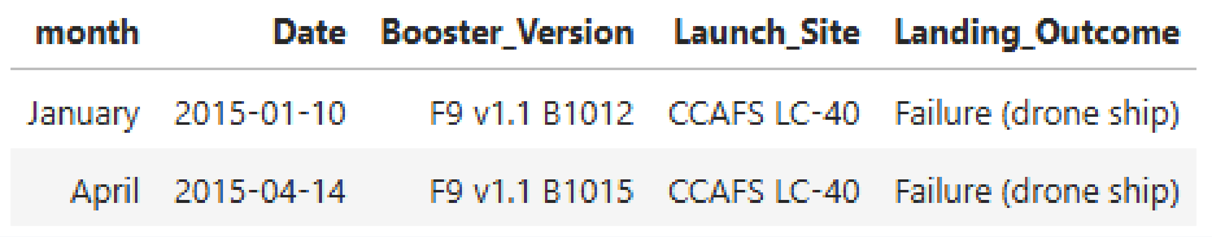
- The table presents all the booster versions that have carried the maximum payload mass.



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

2015 Launch Records

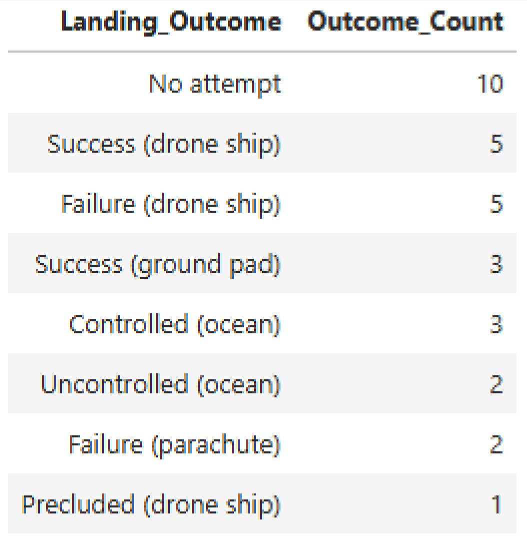
- The table presents the records that will show the names of the months, the results of the failed landing on the unmanned spacecraft, the propellant versions and the launch site for the months of 2015.



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

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

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



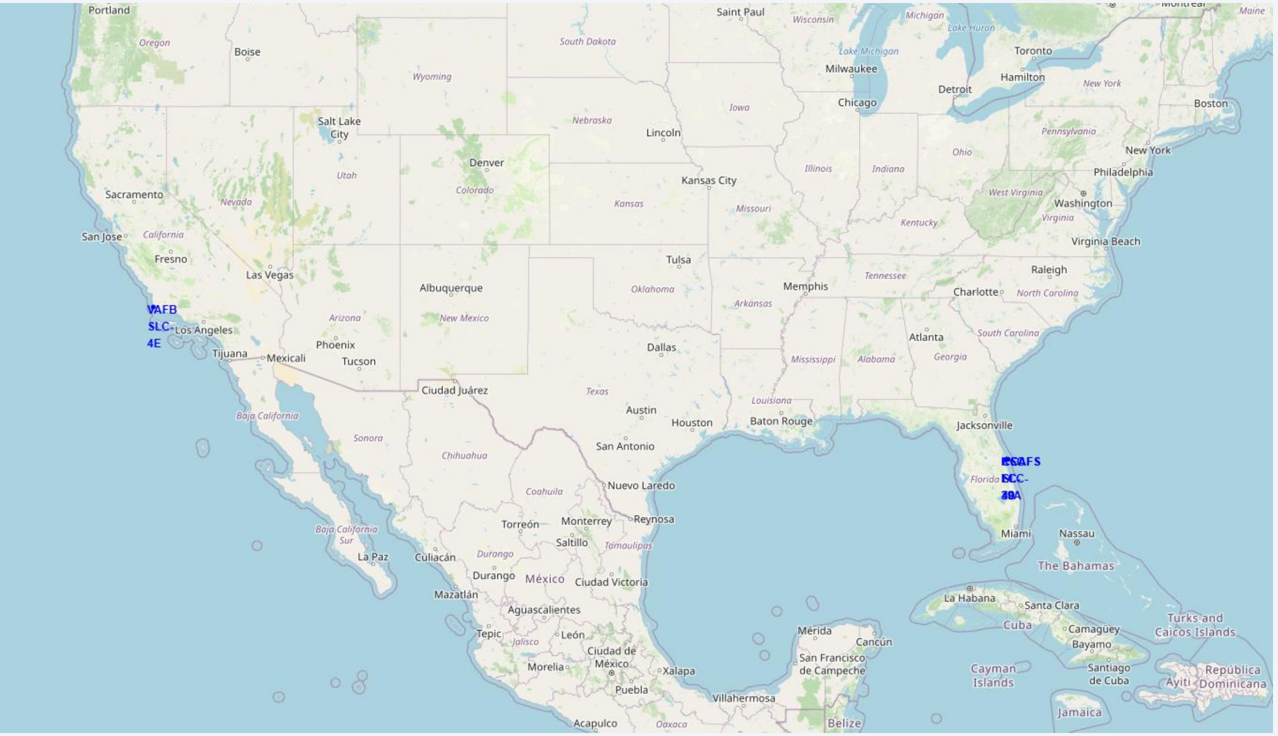
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

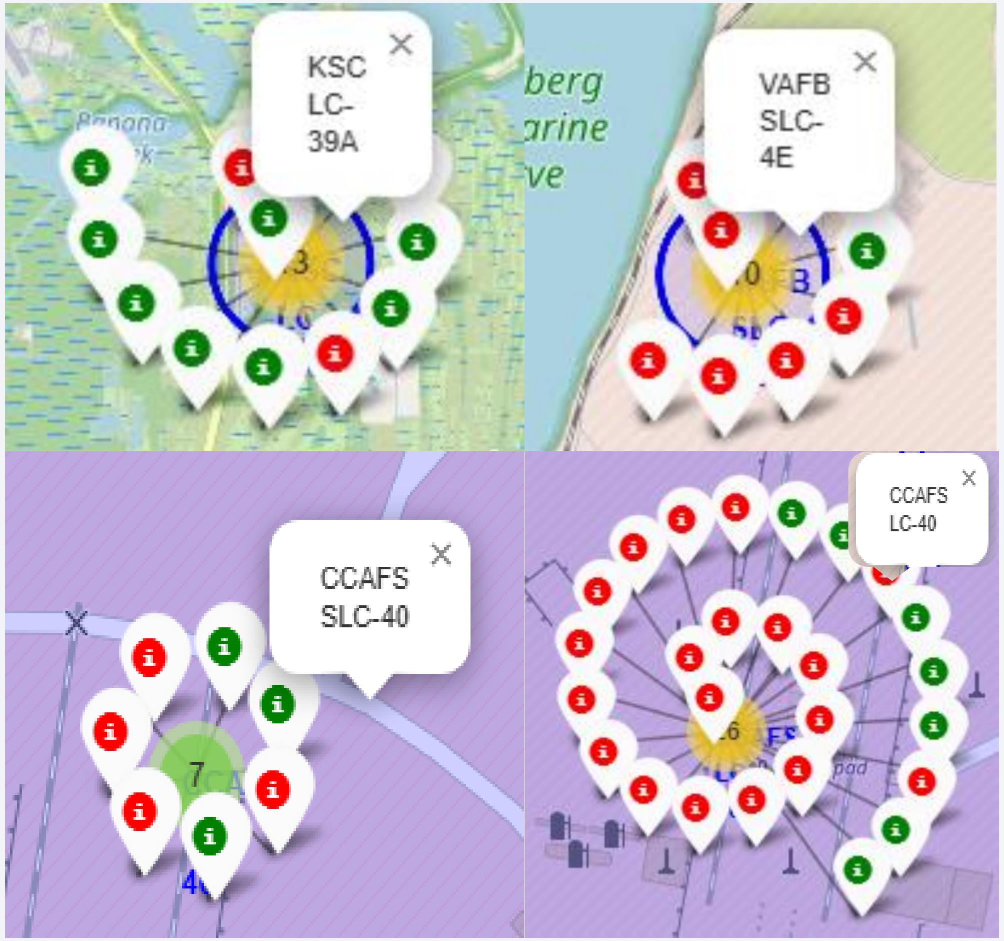
Launches Sites

- The map shows the locations on the map of SpaceX's launch sites.
- On the map, 4 launch points can be observed (although zooming in is necessary).
- The state of Florida presents 3 launch points, while the state of California has only 1.



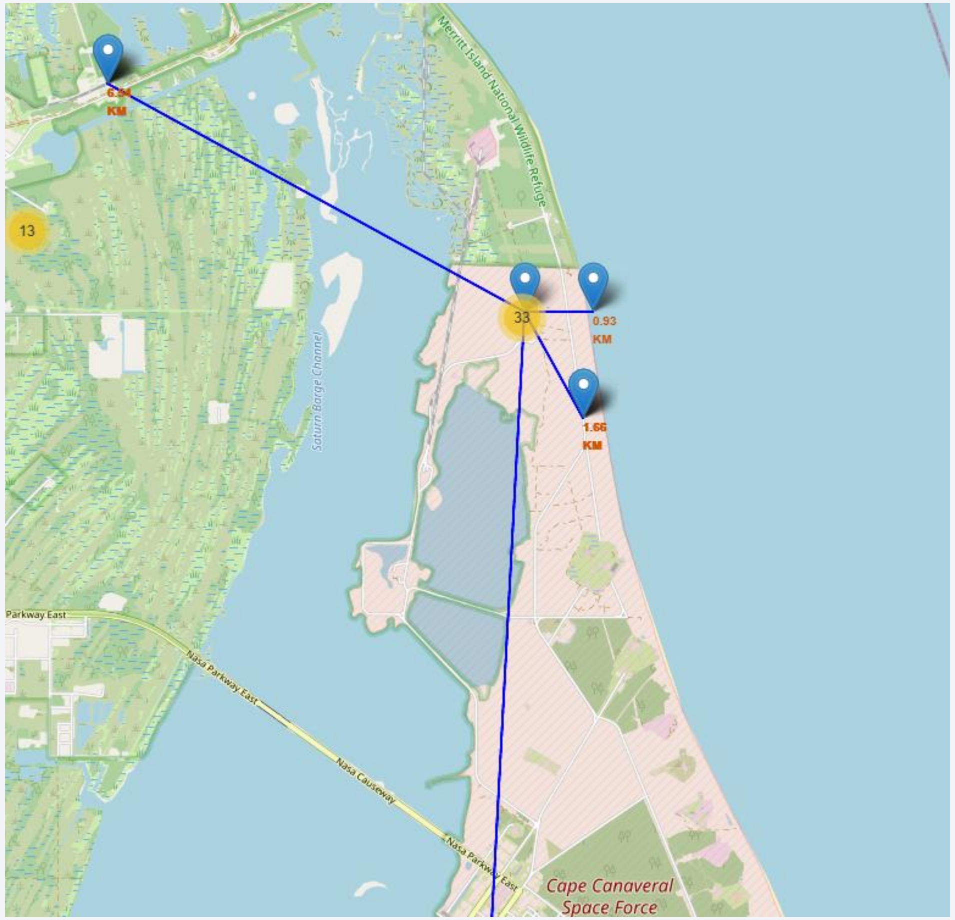
Launching labeled by color

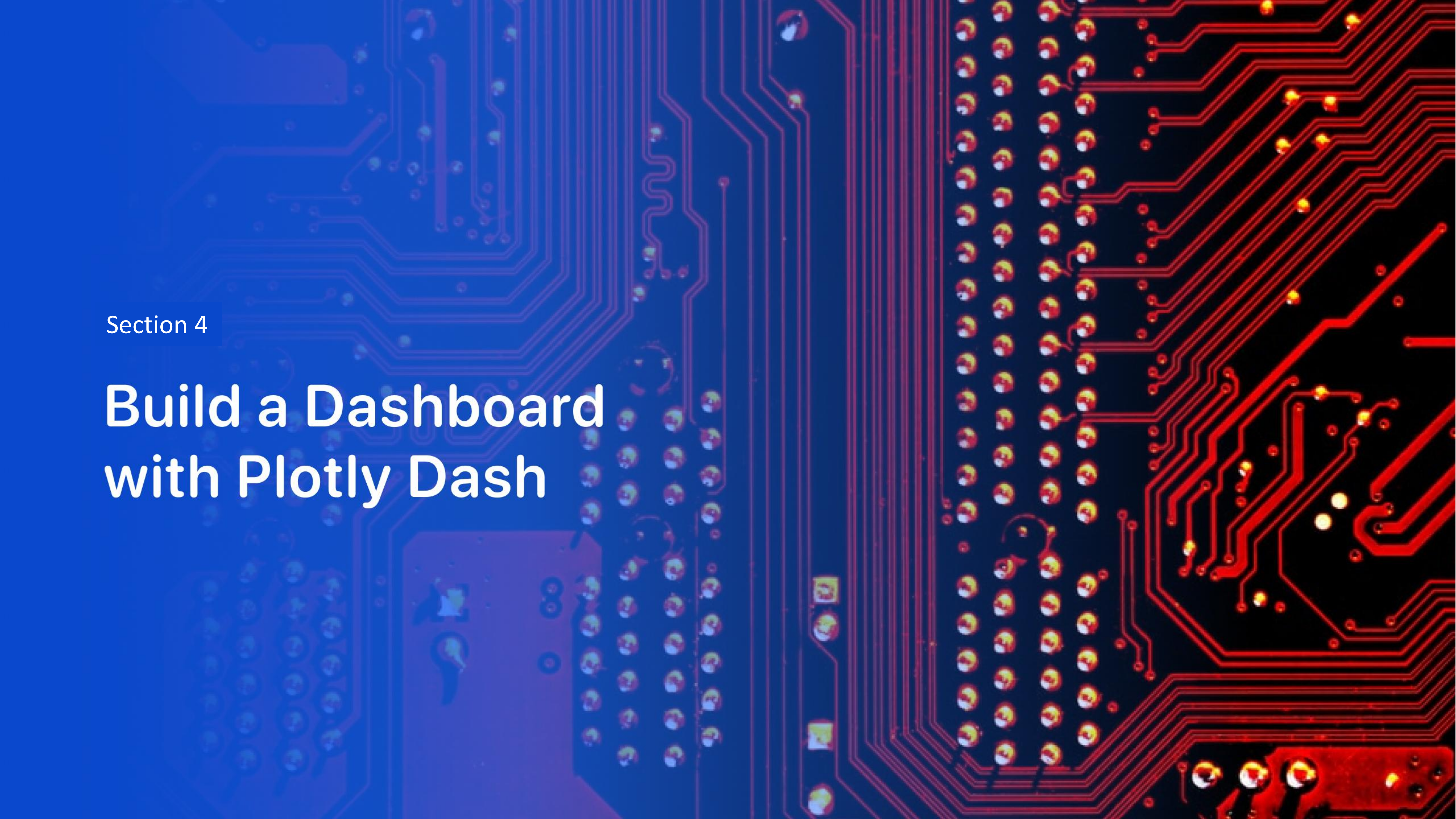
- The map shows the launches of each site labeled by color. Green color shows successful launches and red color shows unsuccessful launches.
- It is observed that the CCAFS LC-40 launch site has the highest number of launches and the worst ratio of successful launches.
- On the other hand, the KSC LC-39A site has the highest ratio of successful launches.



Proximities of the Launch Site CCAFS SLC-40

- The map shows the distance from 4 locations to the launching point CCAFS SLC-40.
- CCAFS SLC-40 is located at a distance of 6.54 km from a railway, 0.93 km from the coast, 1.66 km from a highway and 53.96 km from the city of Melbourne.



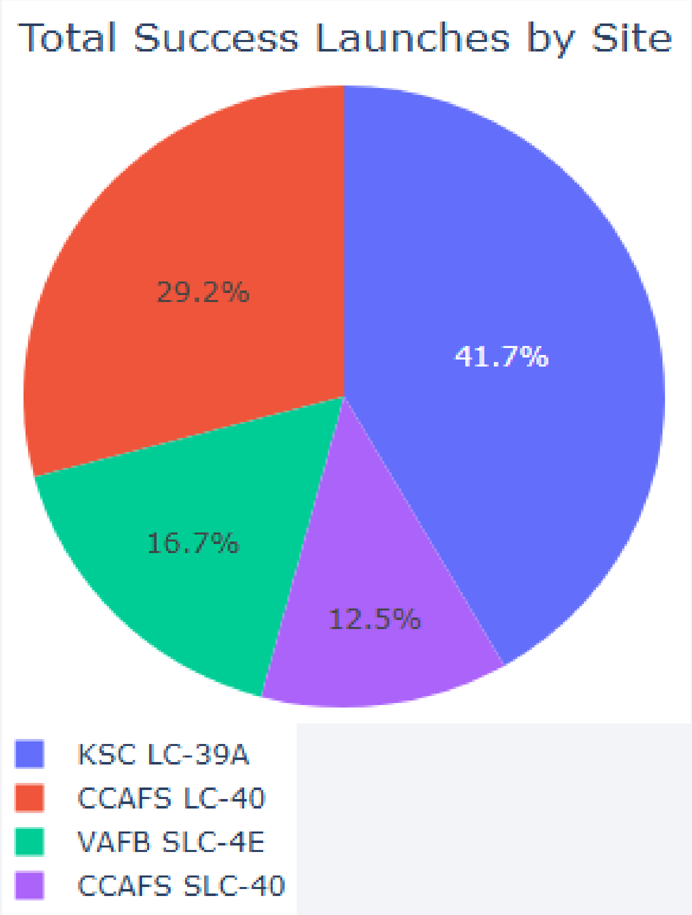


Section 4

Build a Dashboard with Plotly Dash

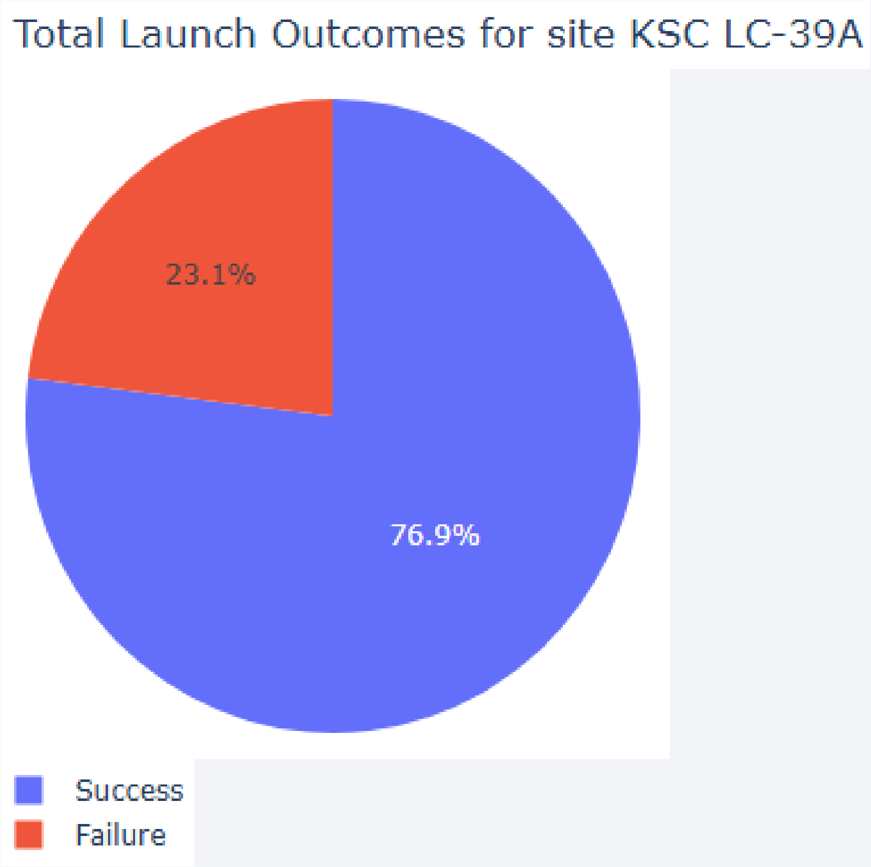
Success Launches by Site

- The chart shows the percentage distribution of successful launches from different launch sites.
- The KSC LC-39A site has the highest number of successful launches, accounting for 41.7% of the total. It is followed by CCAFS LC-40 with 29.2%, then VAFB SLC-4E with 16.7%, and finally CCAFS SLC-40 with 12.5%.
- This indicates that most of the successful launches have been carried out from the KSC LC-39A site, which may reflect its higher capacity, frequency of use, or reliability.



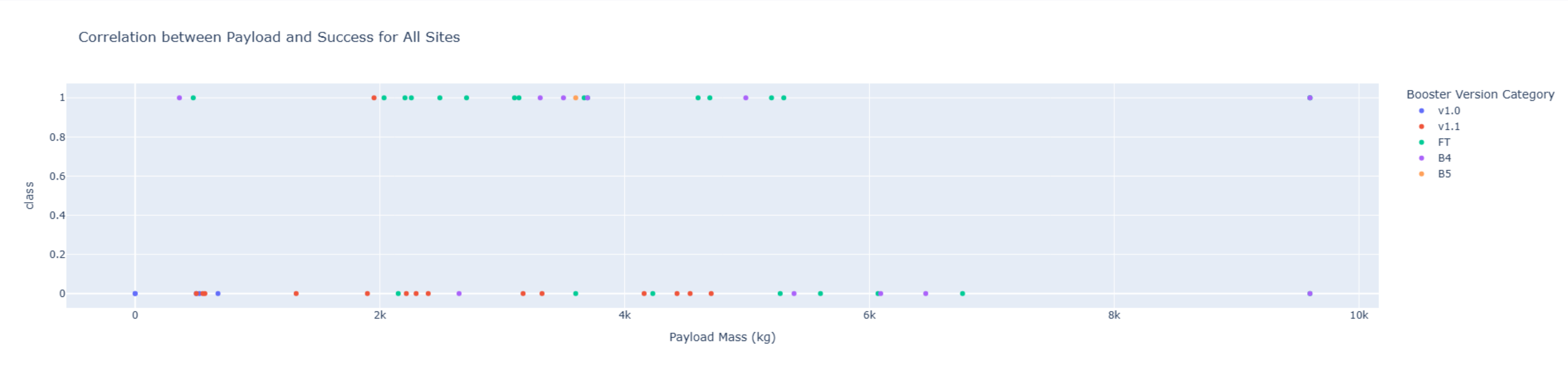
KSC LC-39A Site Launches

- The chart shows the percentage breakdown of successful and failed launches from the KSC LC-39A site.
- The KSC LC-39A site has a 76.9% success rate for launches, with 23.1% being unsuccessful.



Payload vs. Launch Outcome

- This scatter plot shows the relationship between the payload mass (X-axis, in kg) and the launch success (Y-axis, where 1 is success and 0 is failure).
- Each point represents an individual launch, and its color indicates the Booster version used (v1.0, v1.1, FT, B4, B5). The goal is to visualize if there is a correlation between payload mass and success, considering the different booster versions.
- The more modern versions (FT, B4, B5) have demonstrated the ability to launch significantly heavier payloads with high reliability, whereas the earlier versions (v1.0, v1.1) were used for lighter payloads with varying degrees of success.

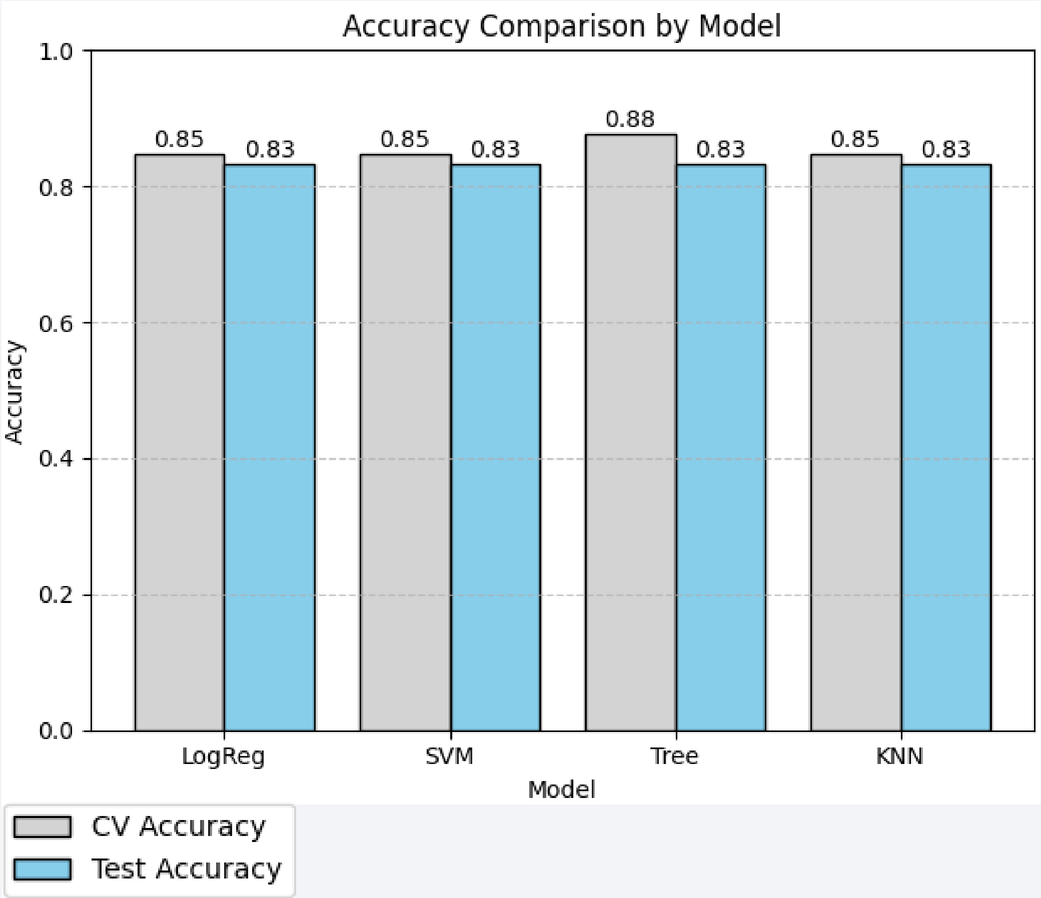


Section 5

Predictive Analysis (Classification)

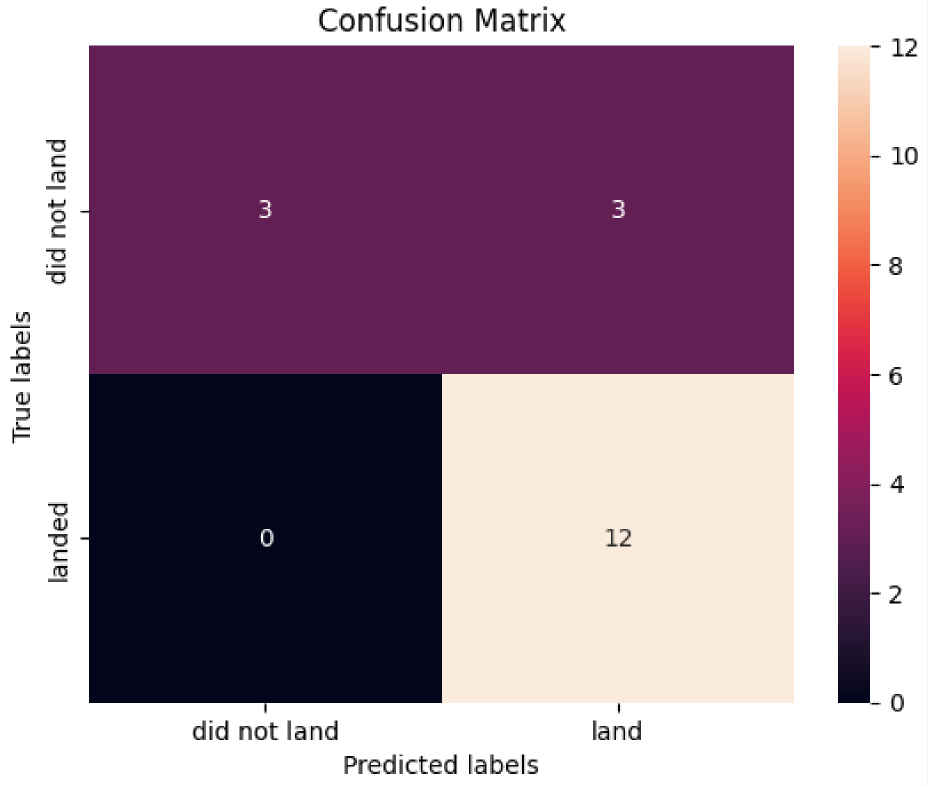
Classification Accuracy

- Comparison of classification model accuracy: Logistic Regression (LogReg), Support Vector Machines (SVM), Decision Tree (Tree), and K-Nearest Neighbors (KNN), divided into cross-validation accuracy (CV Accuracy) and test accuracy (Test Accuracy).
- The model with the highest accuracy in cross-validation is Decision Tree. For test accuracy all models have the same accuracy.



Confusion Matrix

- It was found that all models have the same test accuracy, so the confusion matrix of the Decision Tree classification model is shown, as it has the highest cross-validation accuracy. Although the explanation and the graph are based on this model, they are also valid for the other models.
- The confusion matrix helps to see how often a model gets each type of outcome right or wrong. In this case, the model correctly identified 12 real landings and predicted 3 landings that did not actually occur. It never failed to detect a real landing. This means it has a very good recall, although its precision could be improved.



Conclusions

- This project aimed to analyze the factors influencing the success of the Falcon 9 first stage landing, combining data science tools with an analytical perspective based on economics.
- Through data processing, visualization and predictive modeling, relevant patterns related to payload mass, orbit type, launch site and flight history were identified, all impacting the probability of mission success.
- The results obtained not only reflect relevant technical trends in Falcon 9 performance but also demonstrate the potential of data science as a tool for rigorous and systematic analysis of complex phenomena.
- This report represents my first approach to the practical use of methodologies such as exploratory analysis and classification models. While this involved challenges inherent in the learning process, it also marked a valuable starting point for further developing skills to complement my training as an economist.
- As future work, I would like to explore additional methodologies - beyond classification - that allow me to approach this type of analysis from different angles. I also plan to develop more reports of this type to deepen and refine the ability to communicate findings in a clear, effective and data-driven manner.

Appendix

- Project repository:

The complete code, visualizations, data and documentation of the project are available in the following GitHub repository:

https://github.com/andy-huamantalla/Applied-Data-Science-Capstone/tree/main/Appendix

Thank you!