IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ermesa Pepe
05/02/2024

Outline

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

Executive Summary

- The purpose of this analysis was to evaluate the performance of several classification models built for predicting the success of Falcon 9's first-stage landings. The accuracy of each model was assessed, and the results were visualised in a bar chart to facilitate comparison.
- A total of five classification models were evaluated: Support Vector Machine (SVM), Decision Trees, Logistic Regression, and K-Nearest Neighbors (KNN). Each model's accuracy was calculated based on its performance on the test dataset.
- The bar chart clearly illustrates the varying levels of accuracy achieved by each model. Among the models evaluated, the Decision Tree exhibited the highest accuracy at 94%, followed closely by SVM, Logistic Regression, and KNN, which demonstrated 83%.
- These findings provide valuable insights into the effectiveness of different classification algorithms for predicting the success of Falcon 9 first-stage landings. SpaceY can leverage this information to select the most suitable model for their needs and optimise their decision-making processes.
- Moving forward, further refinements and optimisations to the classification models can be explored to enhance their predictive performance and reliability. Additionally, ongoing monitoring and evaluation of model performance will be essential to ensure continued effectiveness in real-world applications.

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website at 62 million dollars; other providers cost upwards of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

- Problems you want to find answers

The problem we want to solve is How we can predict if the Falcon 9 first stage will land successfully maximizing the accuracy on which the first stage will land.

Section 1

Methodology

Methodology

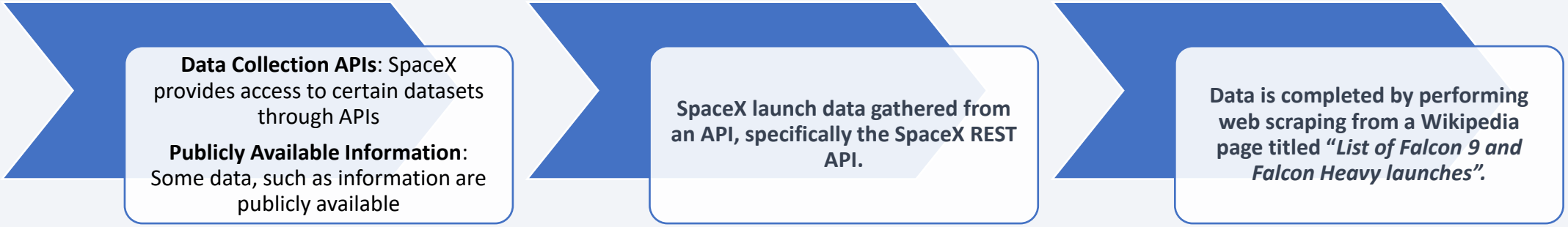
Executive Summary

- Data collection methodology
  - API and Web scraping
- Perform data wrangling
  - Transformation and Standardization
- Perform exploratory data analysis (EDA) using visualisation and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Decision Tree, KNN, and Logistic Regression evaluated on accuracy.

Data Collection

SpaceX launch data is gathered from an API, specifically the SpaceX REST API. Data is completed by performing web scraping from a Wikipedia page titled *"List of Falcon 9 and Falcon Heavy launches".*

Data is completed by performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches



Data Collection – SpaceX API

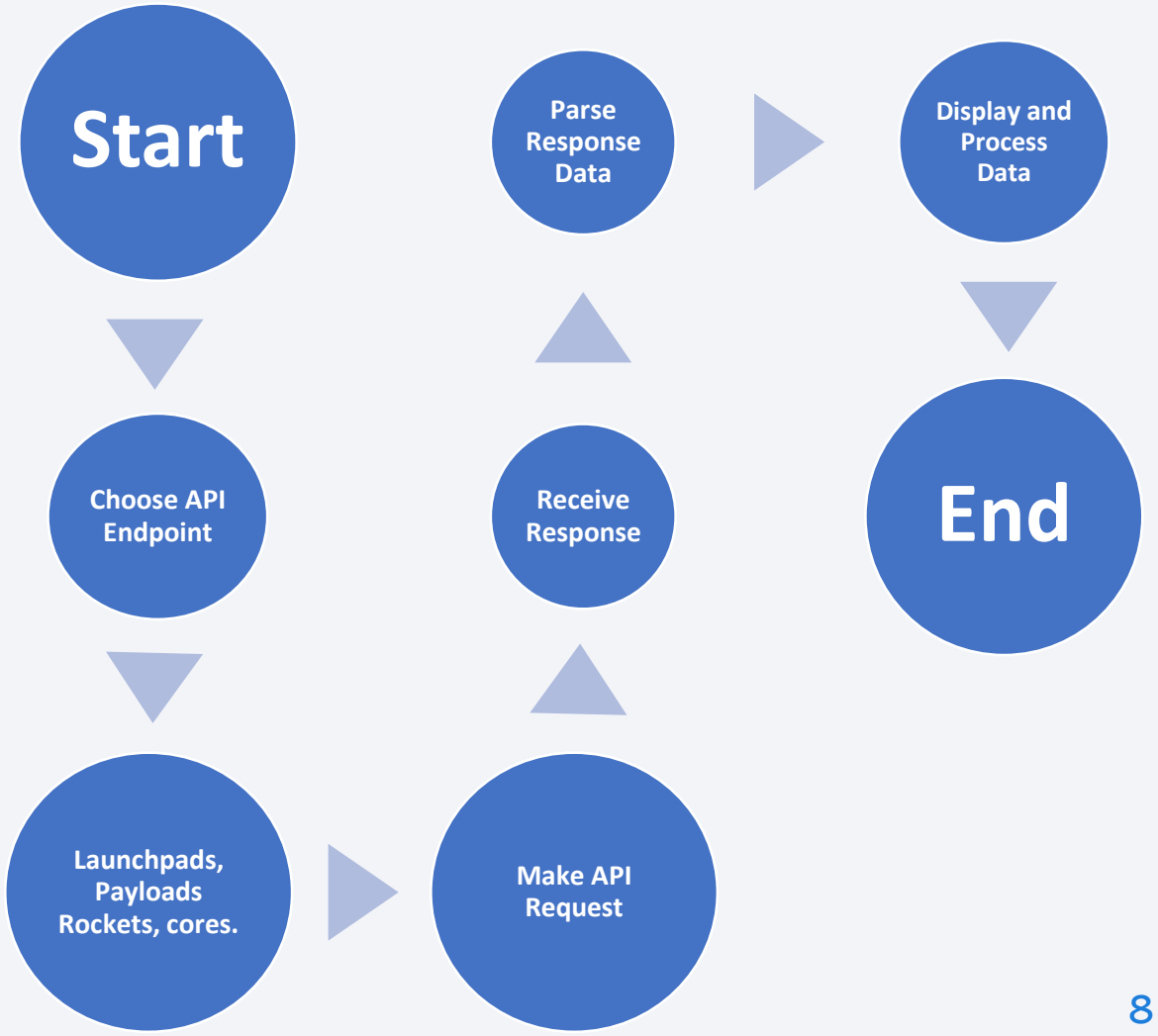
Data Collection consists mainly in:

- Request to the SpaceX API
- Clean the requested data

The flowchart shows the basic steps involved in a SpaceX API call

GitHub URL of the completed SpaceX API calls notebook:

https://github.com/clorofilla/SpaceX-Project/blob/34a063f339288d1ad34217a962c0057c545b62d4/jupyter-labs-spacex-data-collection-api.ipynb



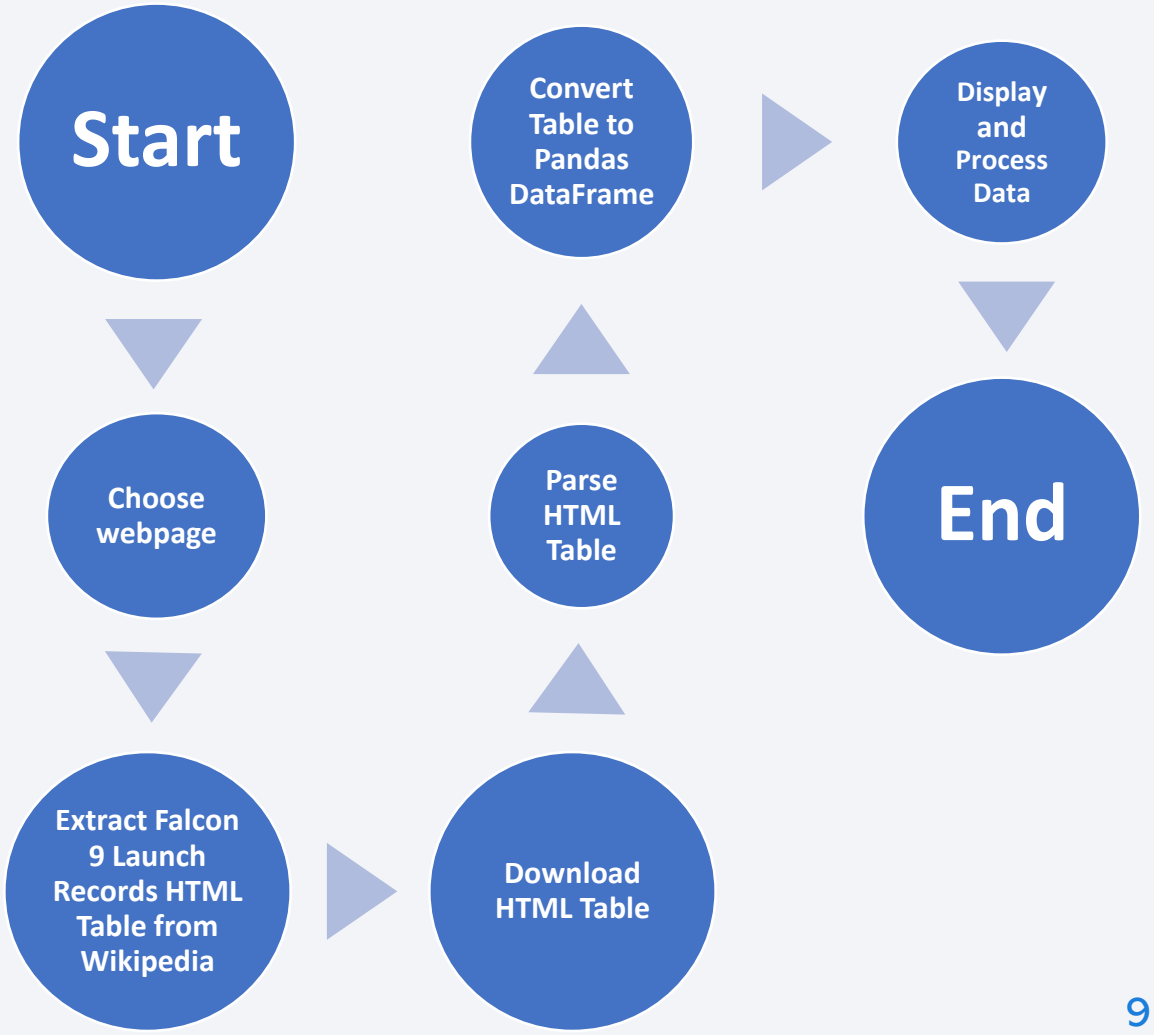
Data Collection - Scraping

The flow shows the steps to collect Falcon 9 historical launch records:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

GitHub URL of the completed web scraping notebook:

https://github.com/clorofilla/SpaceX-Project/blob/34a063f339288d1ad34217a962c0057c545b62d4/jupyter-labs-webscraping.ipynb



Data Wrangling

The data wrangling phase mainly consists of reviewing some dataset attributes and converting landing outcomes into Training Labels, with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful. All details are available on the GitHub repository:

https://github.com/clorofilla/SpaceX-Project/blob/297f081fadb7d34f5dcec934cc58249e80a0b780/labs-jupyter-spacex-Data%20wrangling.ipynb

Data Exploration

- Performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

Data Transformation

- Converted categorical landing outcomes into Training Labels with 1, meaning the booster successfully landed, and 0, meaning it was unsuccessful.

Feature Engineer

- Created a new attribute, “Outcome”, which is the label for training supervised models.

EDA with Data Visualization

- Python code was developed to conduct exploratory data analysis by manipulating data in a Pandas data frame. This stage mainly consisted of:
  - Exploratory Data Analysis
  - Preparing Data Feature Engineering
- We created and executed SQL queries to select and sort data
- Data visualisation skills were applied to visualise the data and extract meaningful patterns to guide the modelling process
- GitHub URL of your completed EDA with the data visualisation notebook:

https://github.com/clorofilla/SpaceX-Project/blob/34a063f339288d1ad34217a962c0057c545b62d4/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

This phase includes exploratory analysis using SQL queries to understand the *Spacex DataSet by*:

- Loading the dataset into the corresponding table in a Db2 database
- Executing SQL queries to answer the following questions:
  - ✓ Display the names of the unique launch sites in the space mission
  - ✓ Display the total payload mass carried by boosters launched by NASA (CRS)
  - ✓ List the date when the first successful landing outcome in the ground pad was achieved.
  - ✓ List the total number of successful and failed mission outcomes
  - ✓ List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - ✓ List the records that will display the month names, failure outcomes in drone ships, booster versions, and launch sites for the months in 2015.

- GitHub URL of your completed EDA with SQL notebook:

https://github.com/clorofilla/SpaceX-Project/blob/34a063f339288d1ad34217a962c0057c545b62d4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The map objects added to the folium map are markers, circles, lines, and polylines.

These were used mainly to:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

GitHub URL of your completed interactive map with Folium map:

https://github.com/clorofilla/SpaceX-Project/blob/34a063f339288d1ad34217a962c0057c545b62d4/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

The dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

The plots and interactions are created mainly to answer the following questions:

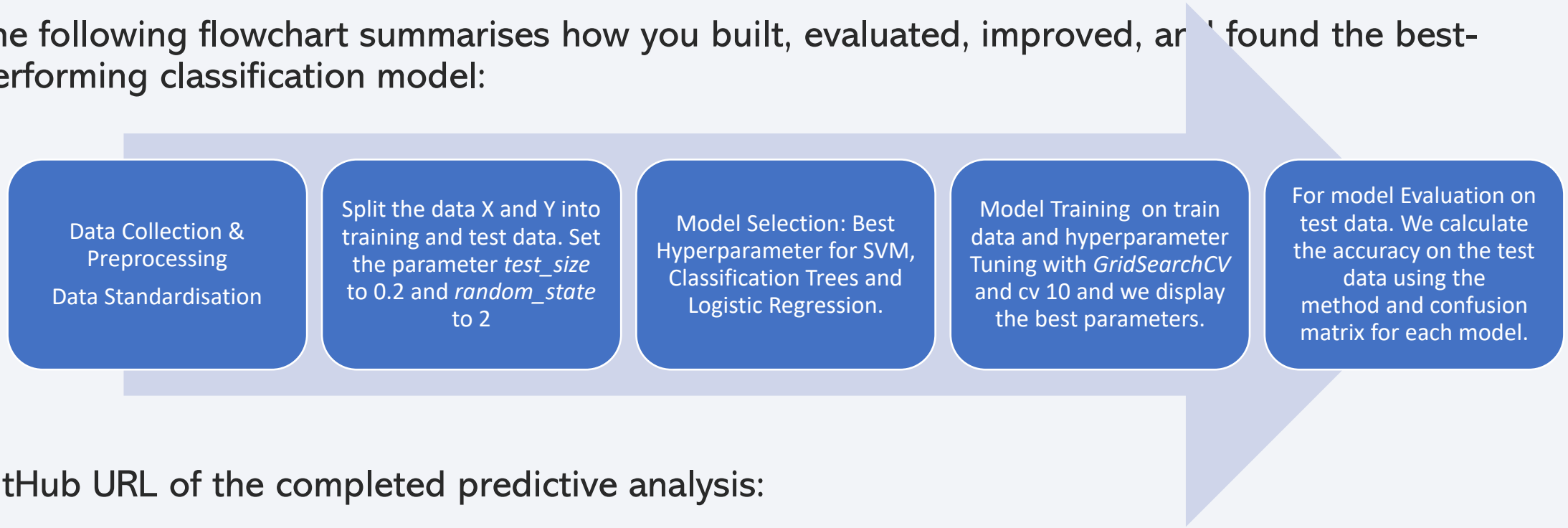
- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

- GitHub URL of your completed Plotly Dash: https://github.com/clorofilla/SpaceX-Project/blob/2b69ca8569afa3d5164eb57e1a35c9902c10e687/spacex_dash_app.py

Predictive Analysis (Classification)

We use machine learning to determine if the first stage of Falcon 9 will land successfully. We split the data into training and test data to find the best Hyperparameter for SVM, Decision Trees, KNN and Logistic Regression. Then, we see the method that performs best using test data.

The following flowchart summarises how you built, evaluated, improved, ar ' found the best-performing classification model:

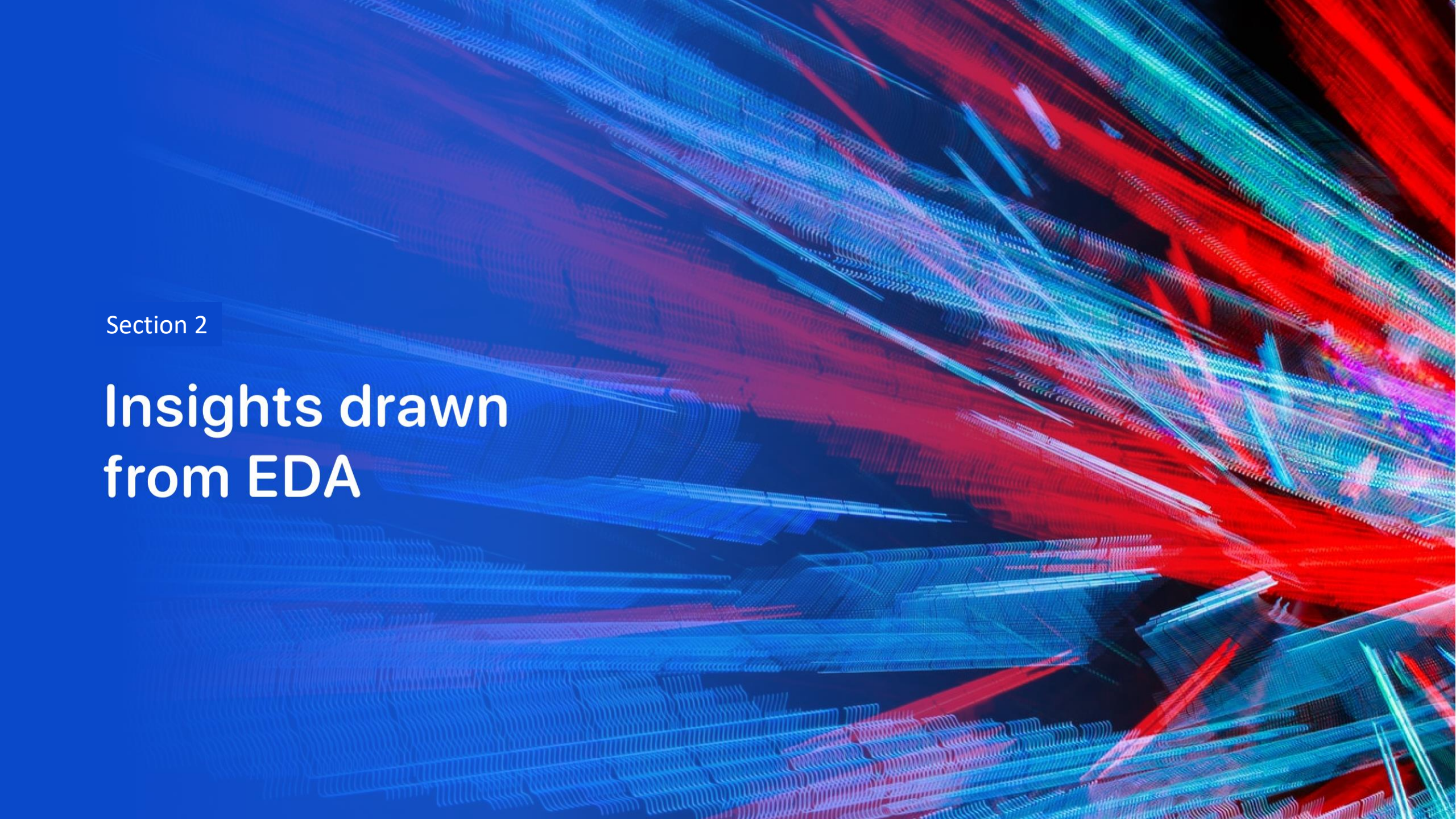


GitHub URL of the completed predictive analysis:

https://github.com/clorofilla/SpaceX-Project/blob/297f081fadb7d34f5dcec934cc58249e80a0b780/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

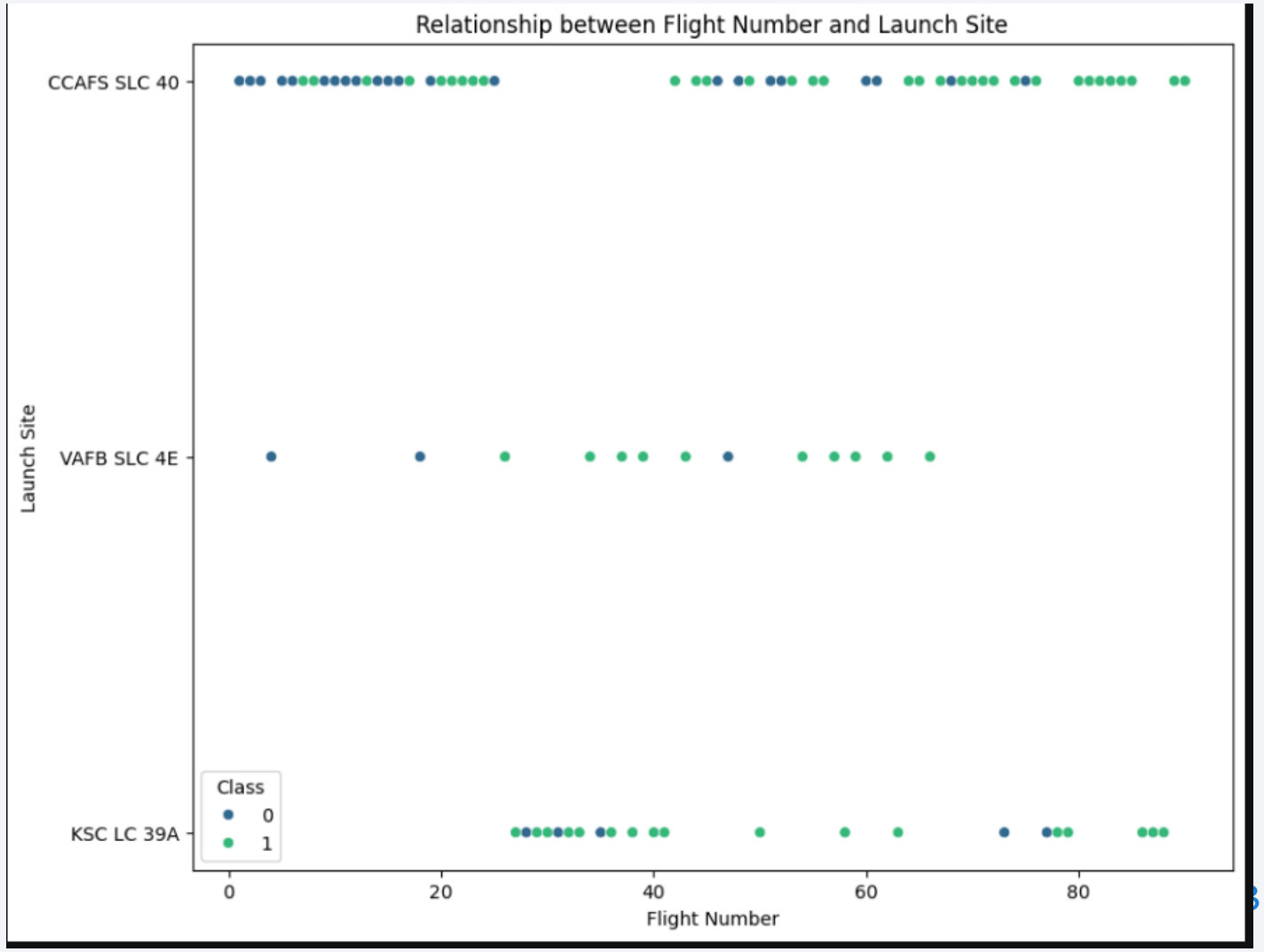
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%

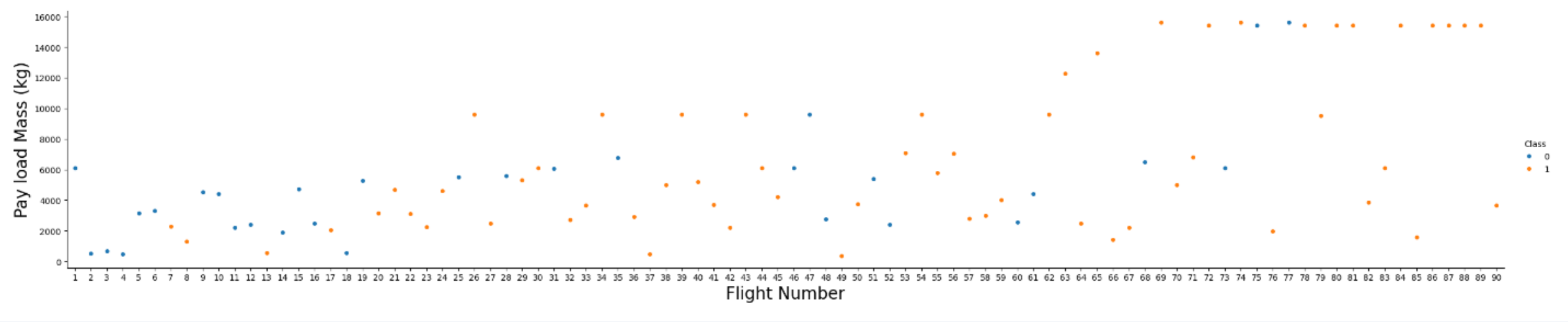
Flight Number vs. Launch Site

We can observe a positive correlation between the number of launches and the success of the launches. We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%



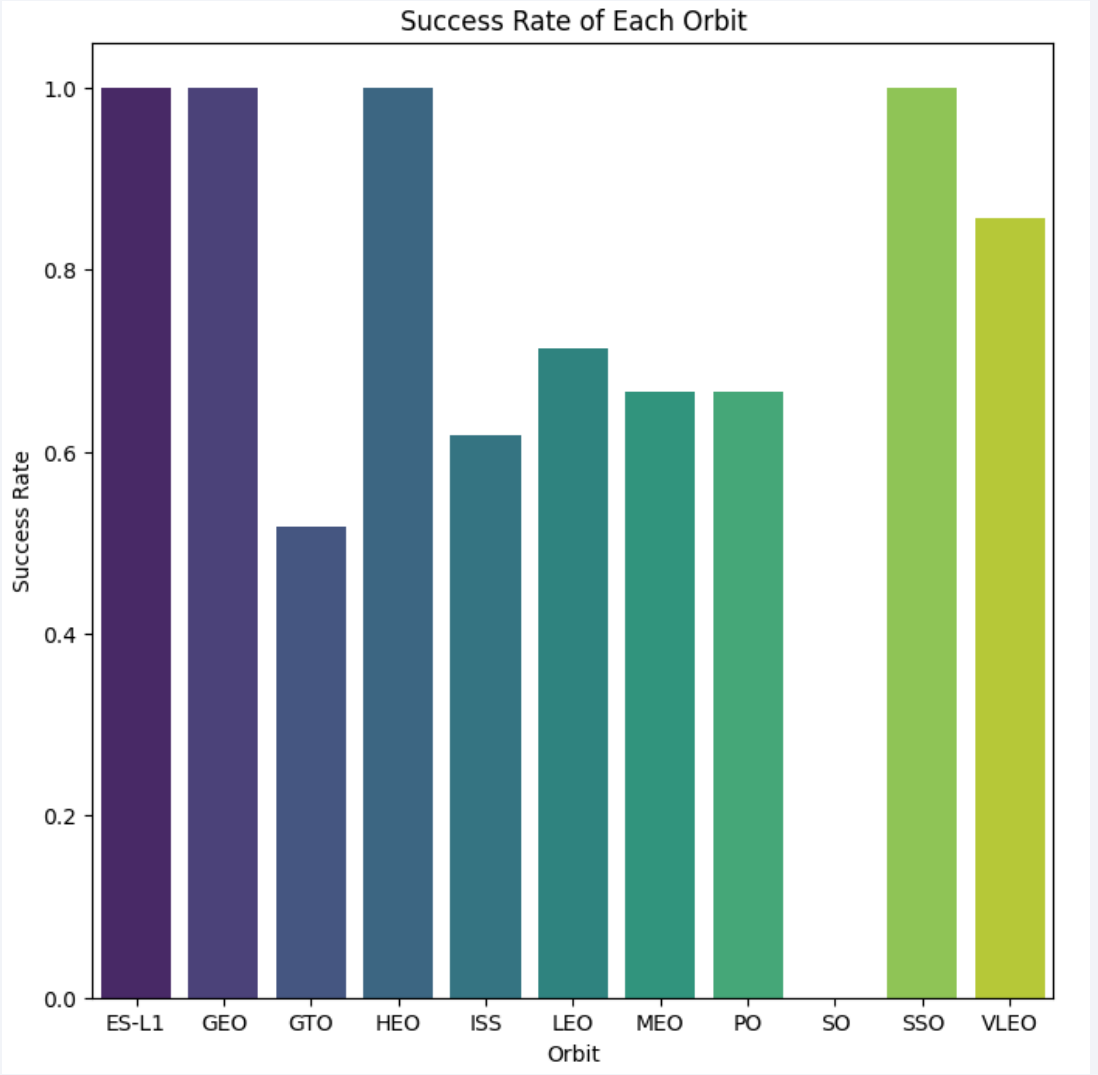
Payload vs. Launch Site

From the scatterplot, we can notice a positive correlation between Flight number and Payload Mass. As the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; the more massive the payload, the less likely the first stage will return.



Success Rate vs. Orbit Type

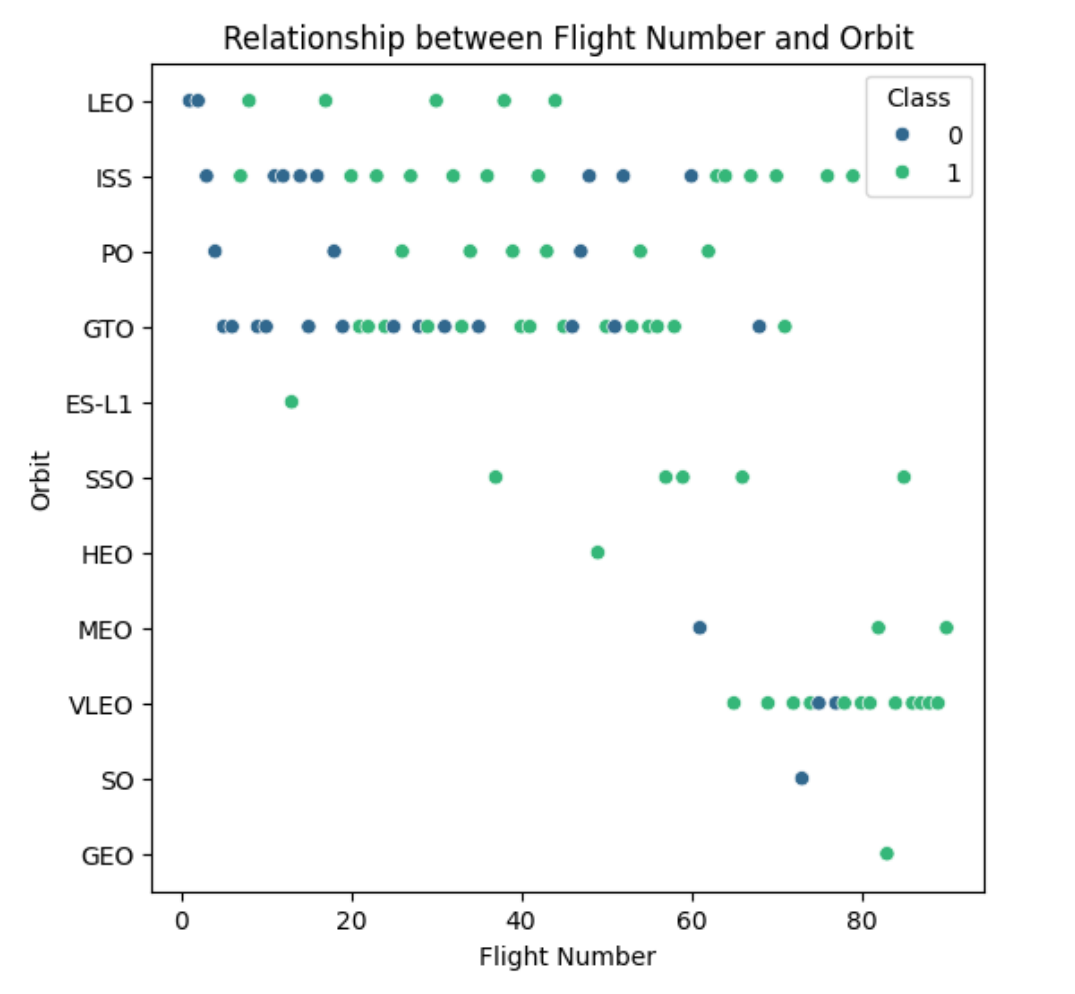
- This bar plot allows to visualise the success rate of each orbit, providing insight into which orbits have the highest and lowest success rates. It helps understand the performance of Falcon 9 launches in different orbital trajectories.
- We can observe that Orbits ES-L1, GEO, HEO, and SSO have a higher success rate.



Flight Number vs. Orbit Type

The scatter plot allows to visualise the relationship between Flight Number and Orbit, with different classes distinguished by color. It helps understand how the Flight Number and Orbit variables relate to each other and how they might correlate with different classes.

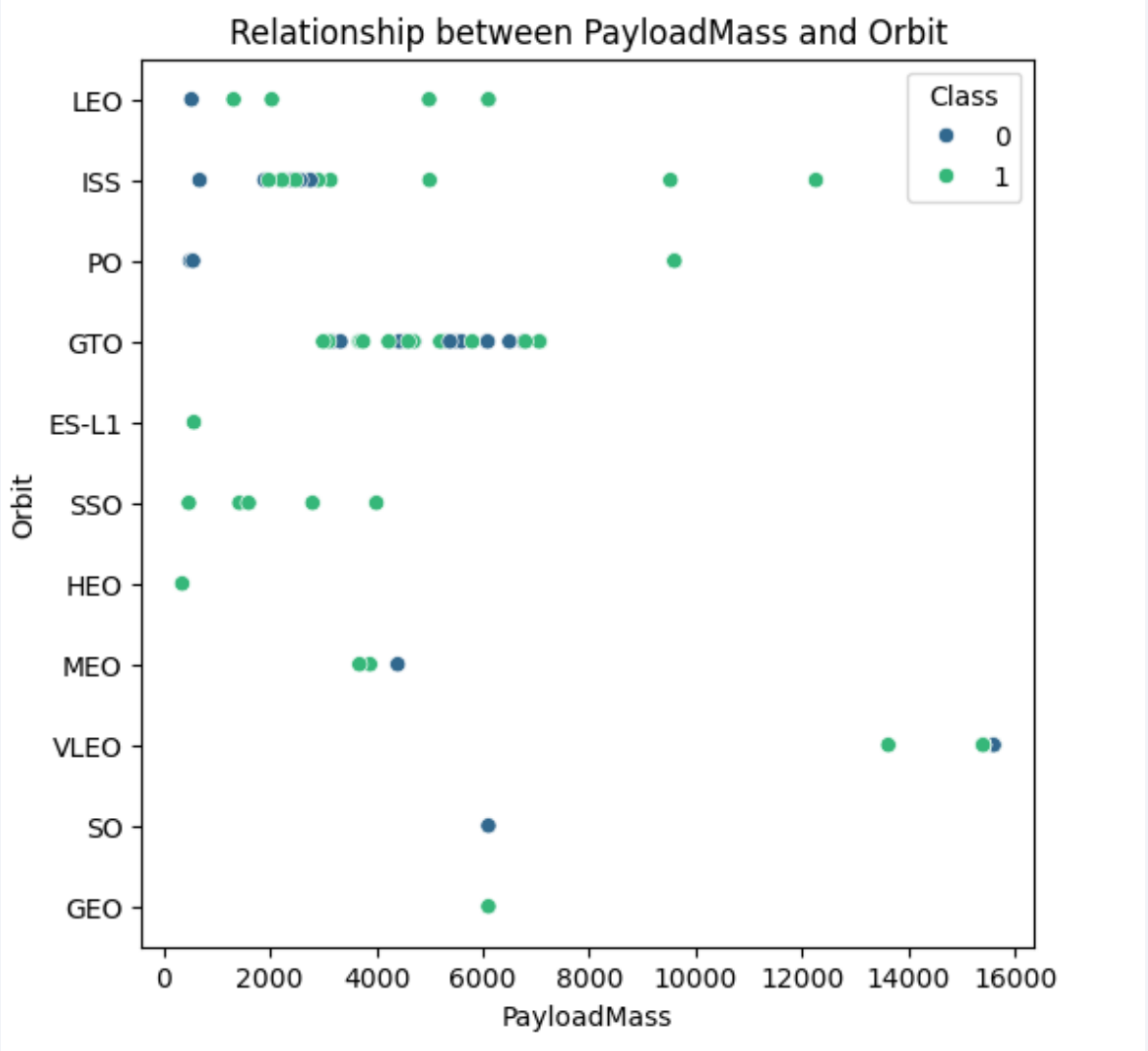
For example, in LEO orbit, Success appears related to the number of flights; conversely, there seems to be no relationship between flight numbers when in GTO orbit.



Payload vs. Orbit Type

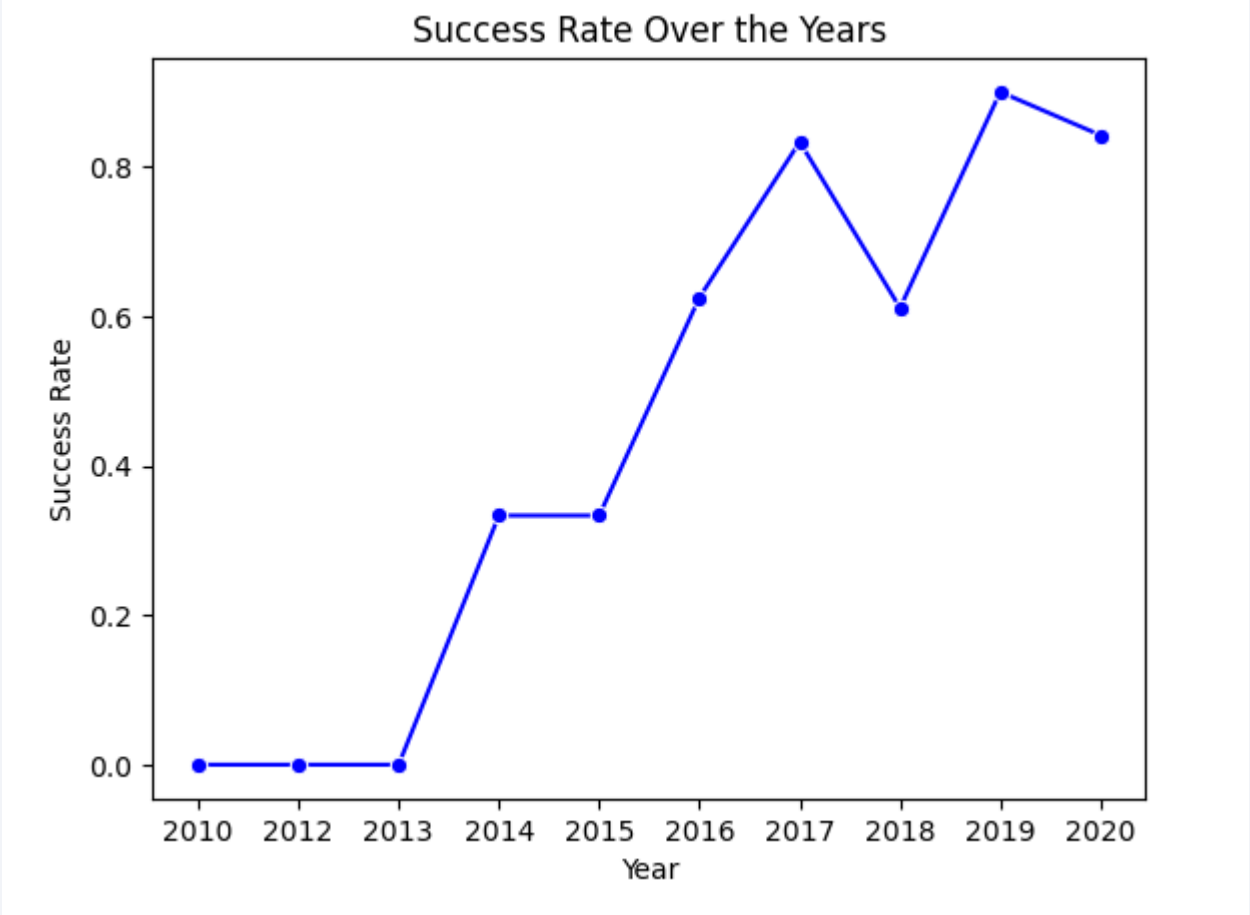
With heavy payloads, the successful landing or positive landing rate is higher for Polar, LEO and ISS.

However, we cannot distinguish this well for GTO as both positive landing rate and negative landing(unsuccessful mission) are there here.



Launch Success Yearly Trend

We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

The query result is the list of the names of the unique launch sites of distinct values from the *Launch_Site* column in the SPACEXTABLE table, ensuring that each launch site appears only once in the output.

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

Launch Site Names Begin with 'CCA'

The query returns the first 5 rows from the SPACEXTABLE where the *Launch_Site* column starts 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 query will return the total payload mass carried by boosters launched by NASA (CRS) in kilograms.

Tatal Mass carried by boosters launched by NASA(Kg)
45596

Average Payload Mass by F9 v1.1

The query returns the average payload mass carried by boosters with the version 'F9 v1.1'.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The query returns the date of the earliest successful mission recorded in the SPACEXTABLE.

Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query will return information about the booster version, landing outcome, and payload mass for successful landings on the drone ship with a payload mass between 4000 kg and 6000 kg

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

The query will return the total number of successful and failed mission outcomes from the SPACEXTABLE table, grouped by the respective outcomes, for a total of 99 successful missions.

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

Boosters Carried Maximum Payload

The query will return the names of the *booster_versions* that have carried the maximum payload mass recorded in the SPACEXTABLE table.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- The query will return the month, year, booster version, launch site, and landing outcome for missions in 2015 with a landing outcome of 'Failure (drone ship)'.

Month	Year	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The query returns the count of each landing outcome for missions that occurred between June 4, 2010, and March 20, 2017, sorted in ascending order of counts.

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

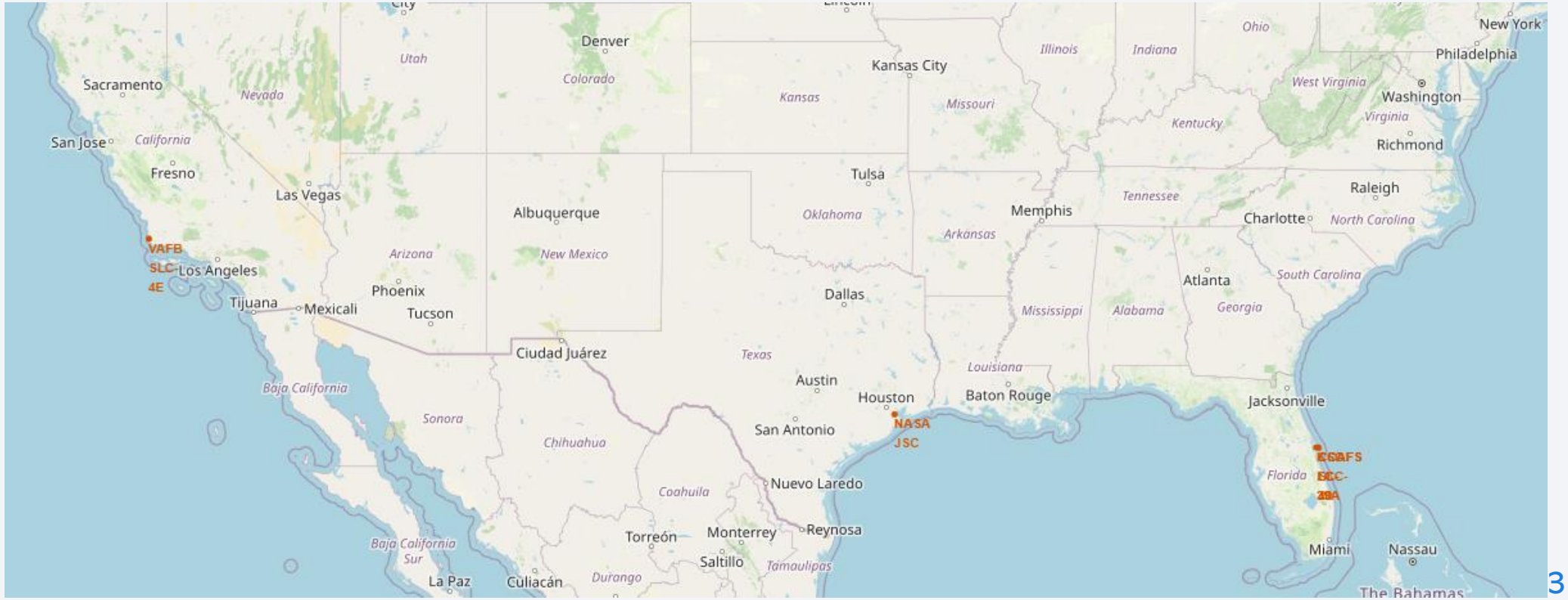
Section 3

Launch Sites Proximities Analysis

Launch sites on a map

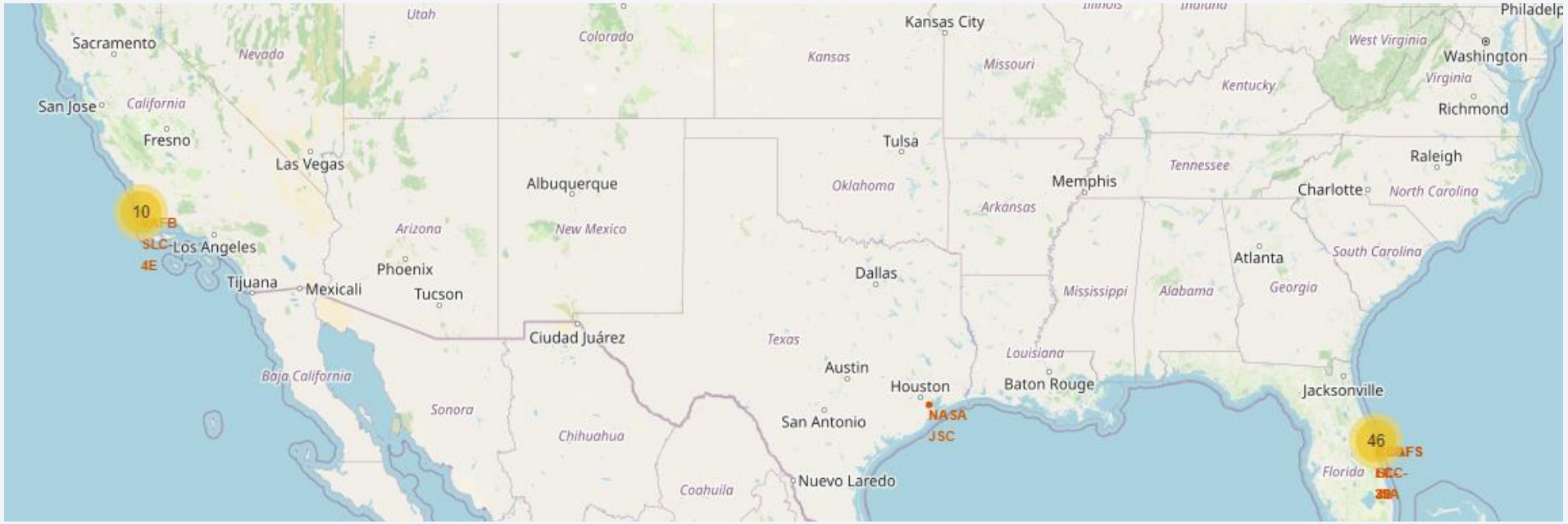
As we can notice, all launch sites are near the Equator line and very close to the coast. If a spacecraft is launched from a site near Earth's equator, it can take optimum advantage of the Earth's rotational speed.

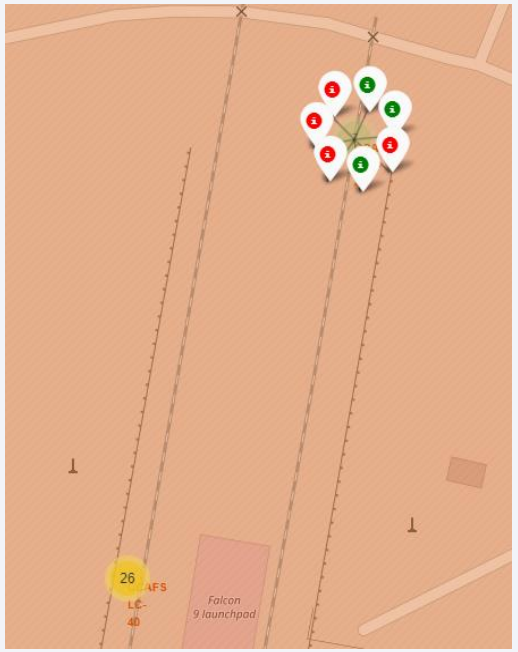
If and when a rocket fails, it has the potential to drop tons, of debris and unused propellant onto a potentially populated area; therefore, it makes sense that a rocket should be launched over an area with as little chance of collateral damage.



Success/failed launches for each site on the map

We provide an interactive map visualization of SpaceX launch sites, clustering them for better visualization and indicating the success or failure of each launch through marker colors.

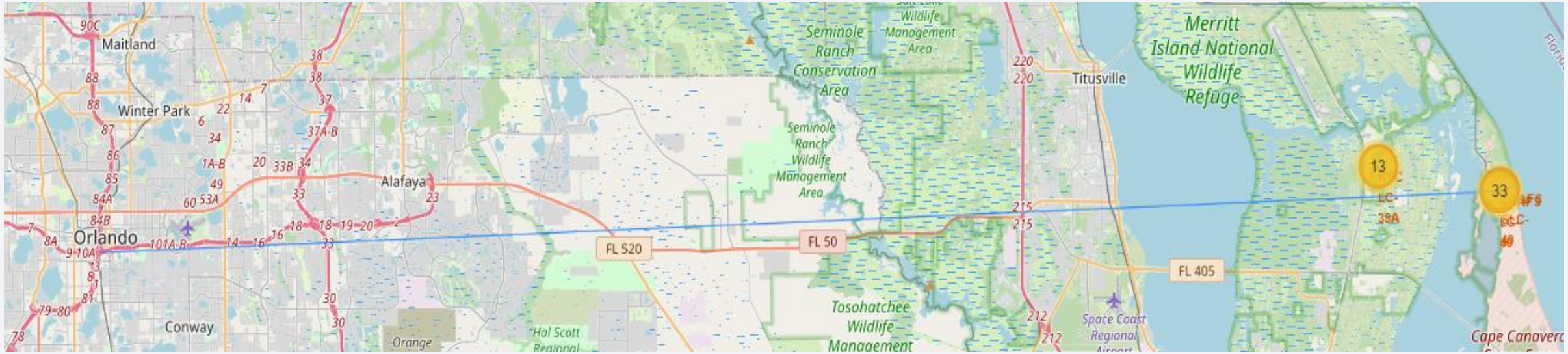


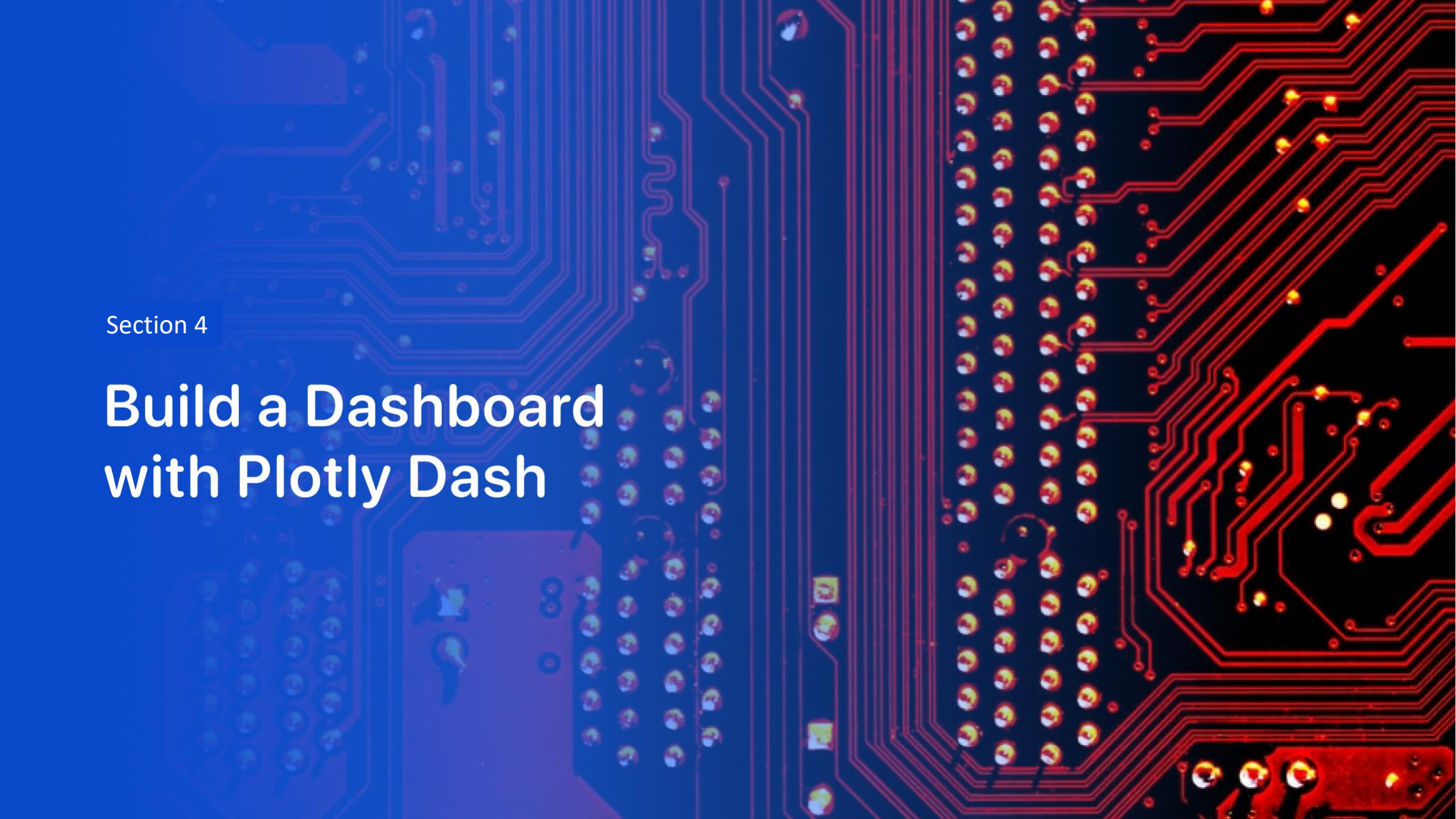


Distances between a launch site to its proximities 3

We can notice that launch sites are not near railways, highways or cities. Instead, they are in close proximity to the coastline.

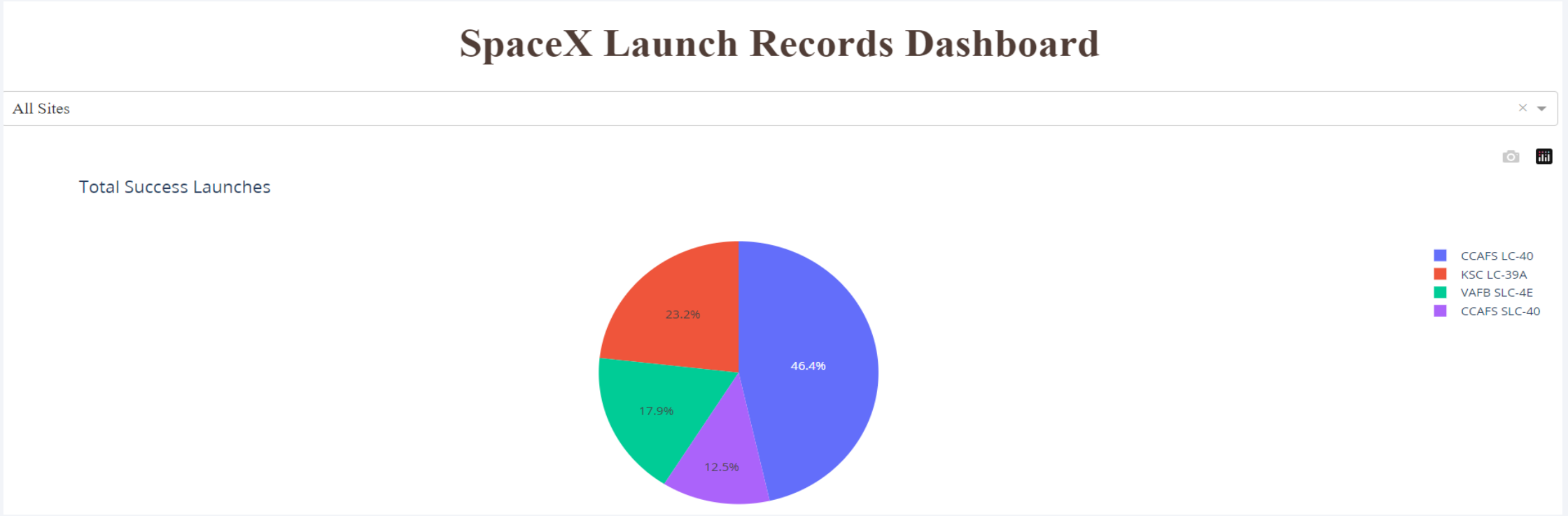
If and when a rocket fails, it has the potential to drop tons, of debris and unused propellant onto a potentially populated area; therefore, it makes sense that a rocket should be launched over an area with as little chance of collateral damage.




Section 4
Build a Dashboard with Plotly Dash

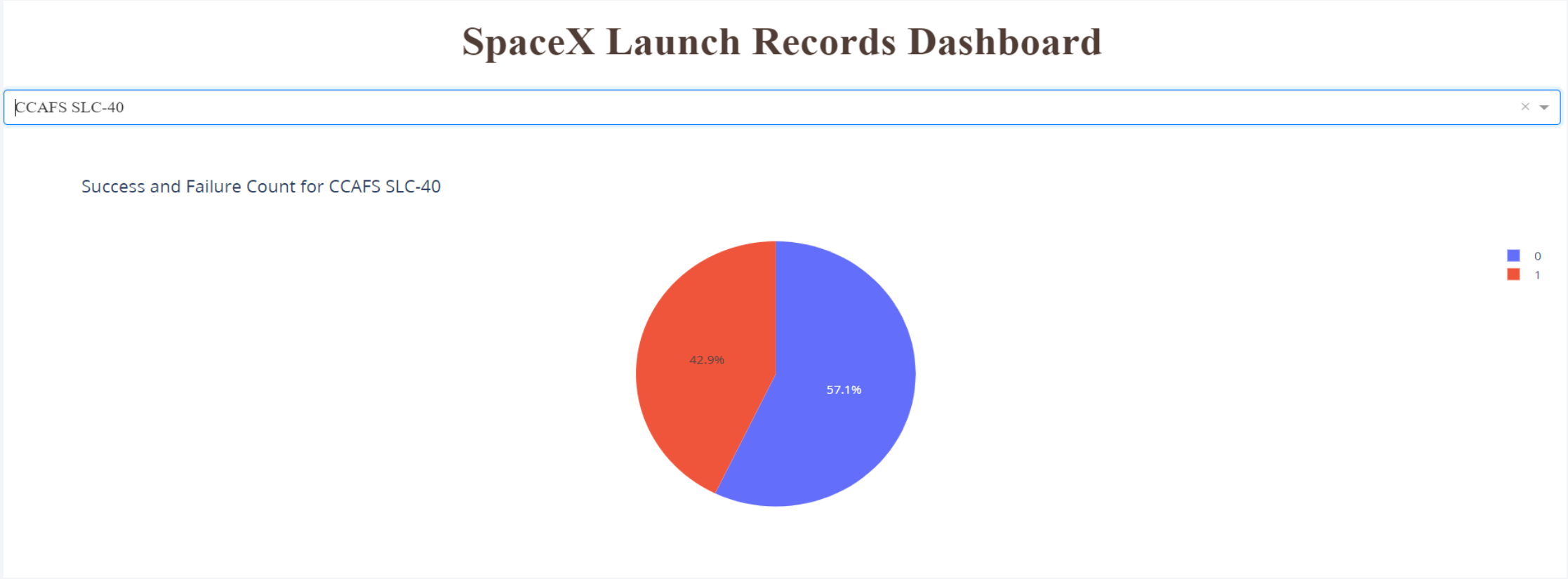
Total Launches by launch site

From the Dashboard, we can observe the total distribution of launches by launch site. We can notice that CCAPS LC-40 report the highest percentage of launches.



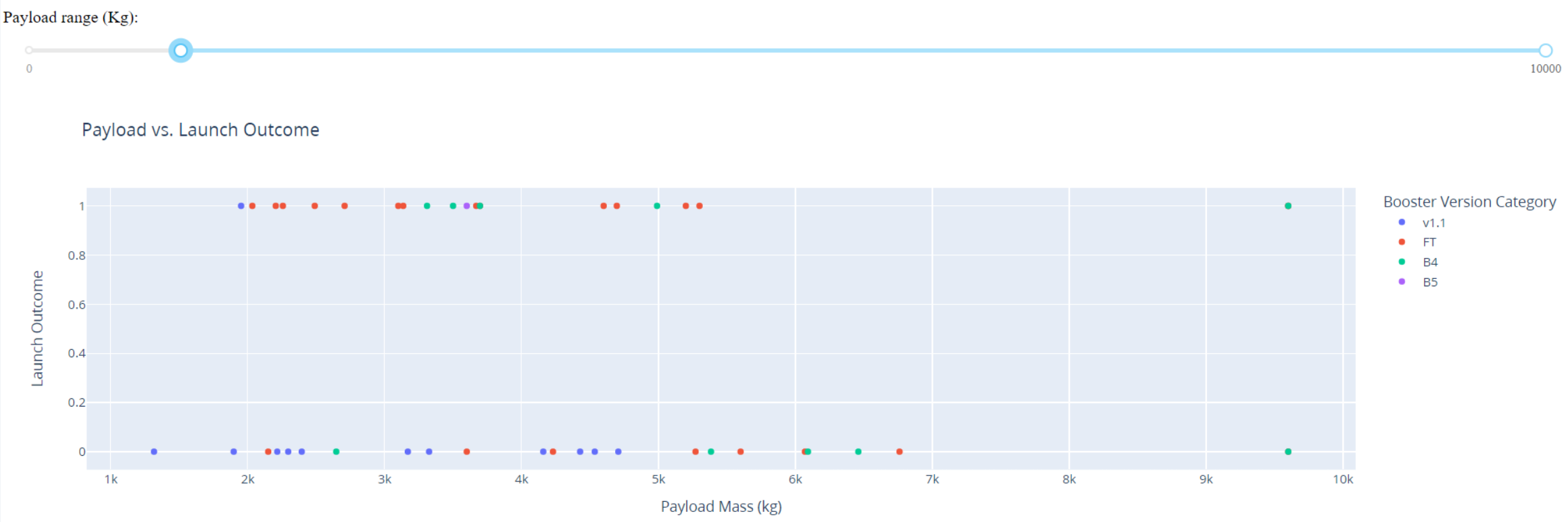
Launch site with highest launch success ratio

Below is the screenshot of the pie chart for the launch site with the highest launch success ratio. We can notice that the CCAFS SLC-40 launch site has the highest percentage of successful launches, 42,9%.



Payload vs. Launch Outcome

From the scatterplot for all sites, with different payloads selected and booster version, we can notice that we have the largest success rate with a payload between 2k and max 6k and boost version FT.

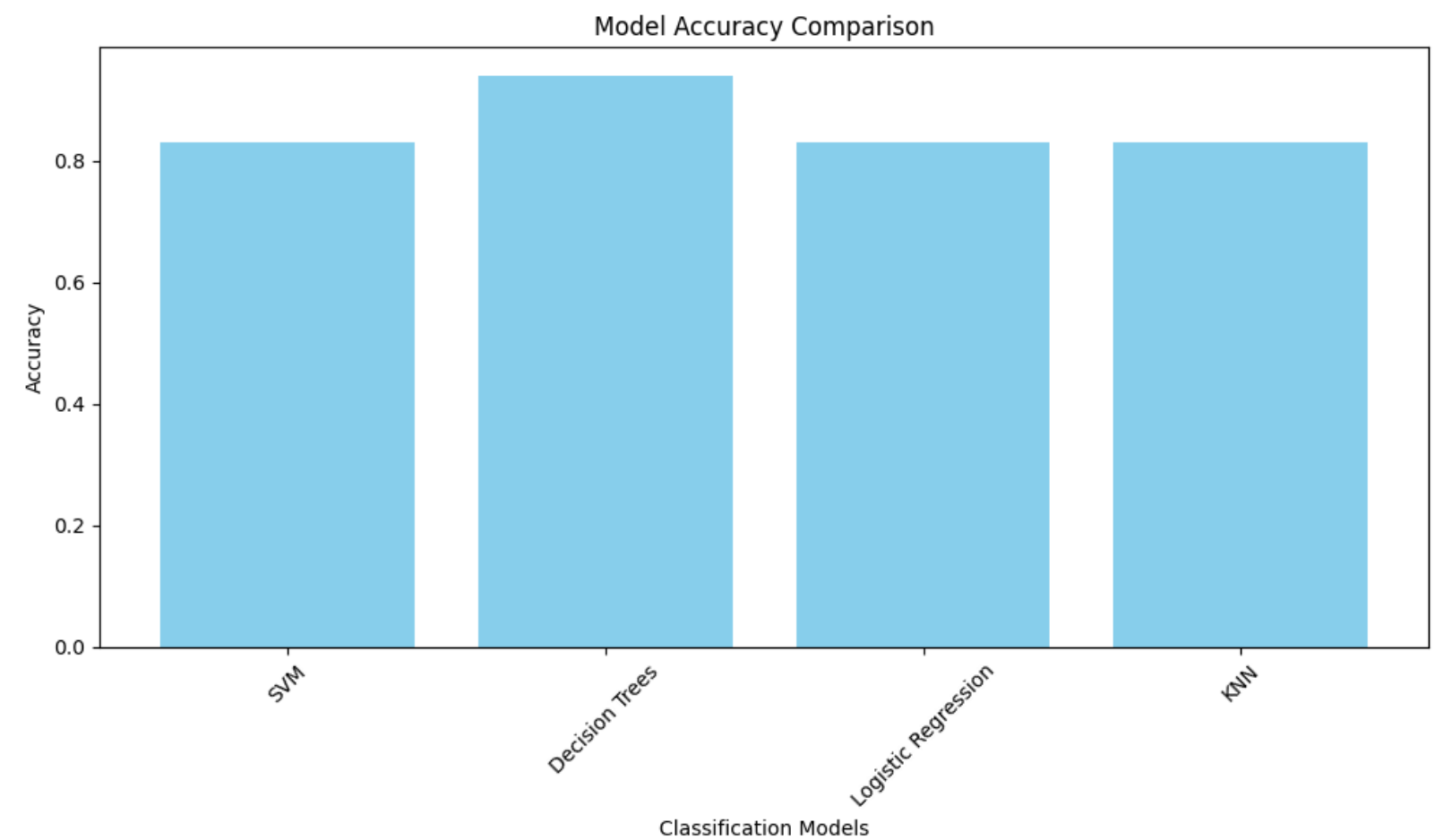


Section 5

Predictive Analysis (Classification)

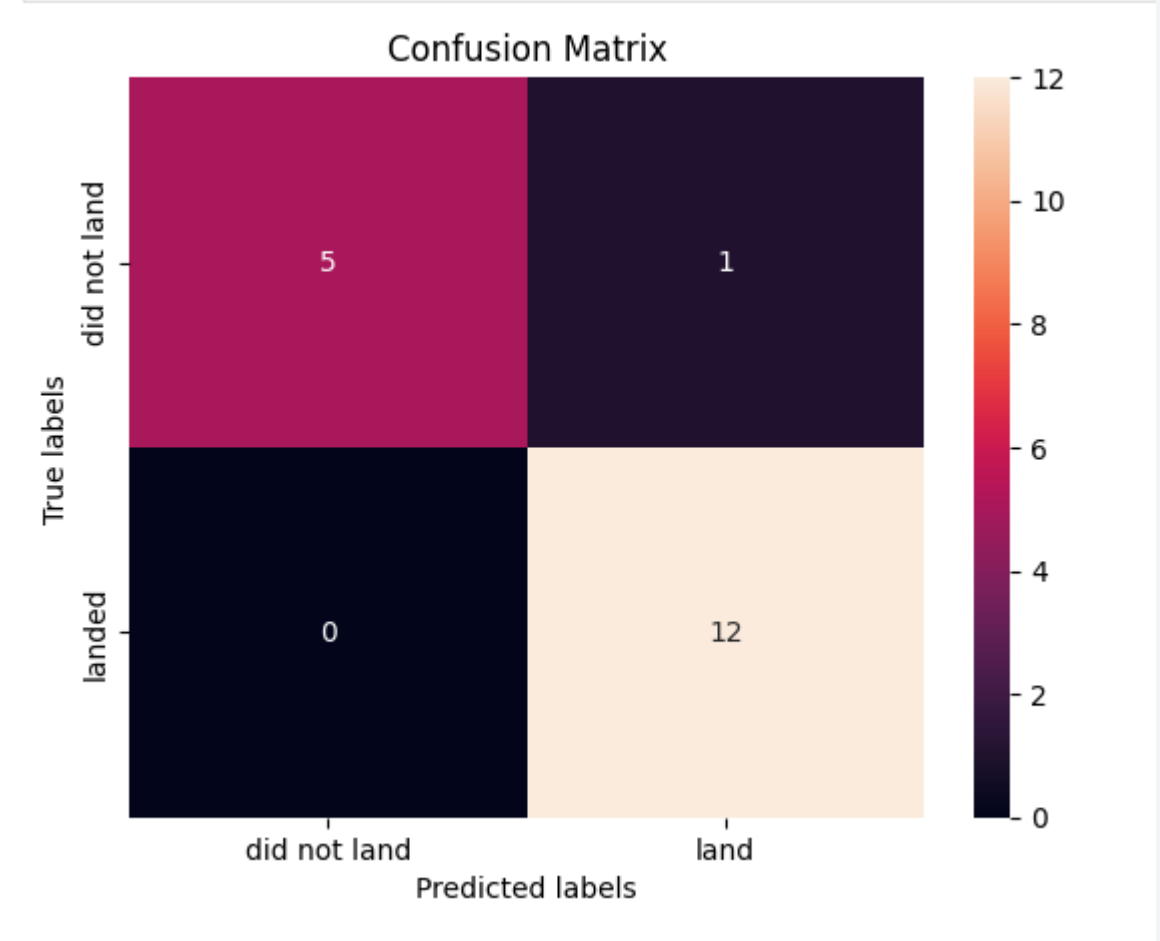
Classification Accuracy

The below bar chart visualizes the built model accuracy for all built classification models implemented. Decision Tree is the model that has the highest classification accuracy of 0.94



Confusion Matrix

The decision tree model is able to predict correctly a launch site's success and failure with a 94% accuracy. From the confusion matrix, we can notice that applied on the test data was able to correctly predict 17 cases with only one false positive.



Conclusions

- The dataset chosen offers very good results from applying exploratory data analyses and data visualization, such as the relationship between features like payload mass and launch outcome, the best location to conduct the launches and the best boost version. Thanks to the dashboard, it is easy to notice that the CCAFS SLC-40 launch site has the highest percentage of successful launches, 42,9%. And payload between 2k and max 6k and boost version FT.
- We can conclude that the Decision tree is the best model to adopt to predict launch outcomes with an accuracy of 0.94. This means that the model can predict on the test set with an accuracy of 94%.
- Because of the limited number of test sets, only 18 samples would be useful to gather more recent data to conduct tests further and compare the models again.

Appendix

GitHub link to complete project https://github.com/clorofilla/SpaceX-Project.git

Thank you!