IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olivia Mangwanda
28/01/2025

Outline

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

Executive Summary

Summary of methodologies:

1. Collect data using the SpaceX Api, perform data wrangling and formatting, then save the data to a data frame.
2. Collect the launch records of Falcon 9 from Wikipedia using web scrapping and save the data into a data frame.
3. Perform exploratory data analysis on the Space X and Falcon 9 data, then create landing outcome labels from the data.
4. Perform SQL queries to analyze the dataset
5. Perform data analysis and feature engineering on the dataset
6. Use folium to find the launch sites for the spaceships on the map
7. Use Machine learning algorithms to find the best hyperparameters for SVM, classification trees and logistic regression

Summary of all results

- Exploratory Data Analysis using descriptive statistics, graphs and SQL queries
- Dashboard of successful launches
- Predictive analysis scores

Introduction

- Project background and context

SpaceX has reengineered space travel, making it more resource efficient and cost effective. SpaceX has a Falcon 9 rocket with a cost of 62 million dollars as opposed to their competitors who offer 165 million dollars for a rocket. What makes SpaceX rockets more recourse efficient and cost effective is the ability to reuse the first stage of the rocket. If the first stage of the rocket lands perfectly, the cost of the launch will decrease. This purpose of this project is to predict whether the first stage of the rocket will land, so that it can be reused by SpaceX.

- Problems you want to find answers

- Which boosters have the highest success rate when it comes to landing the first stage?
- What is the success relationship between success rate and each orbit type?
- What is the yearly trend for the successful launches?
- What are the optimal parameters for each Machine Learning algorithm?
- Which machine learning model is the most optimal to predict whether the first stage will land or not?

Section 1

Methodology

Methodology

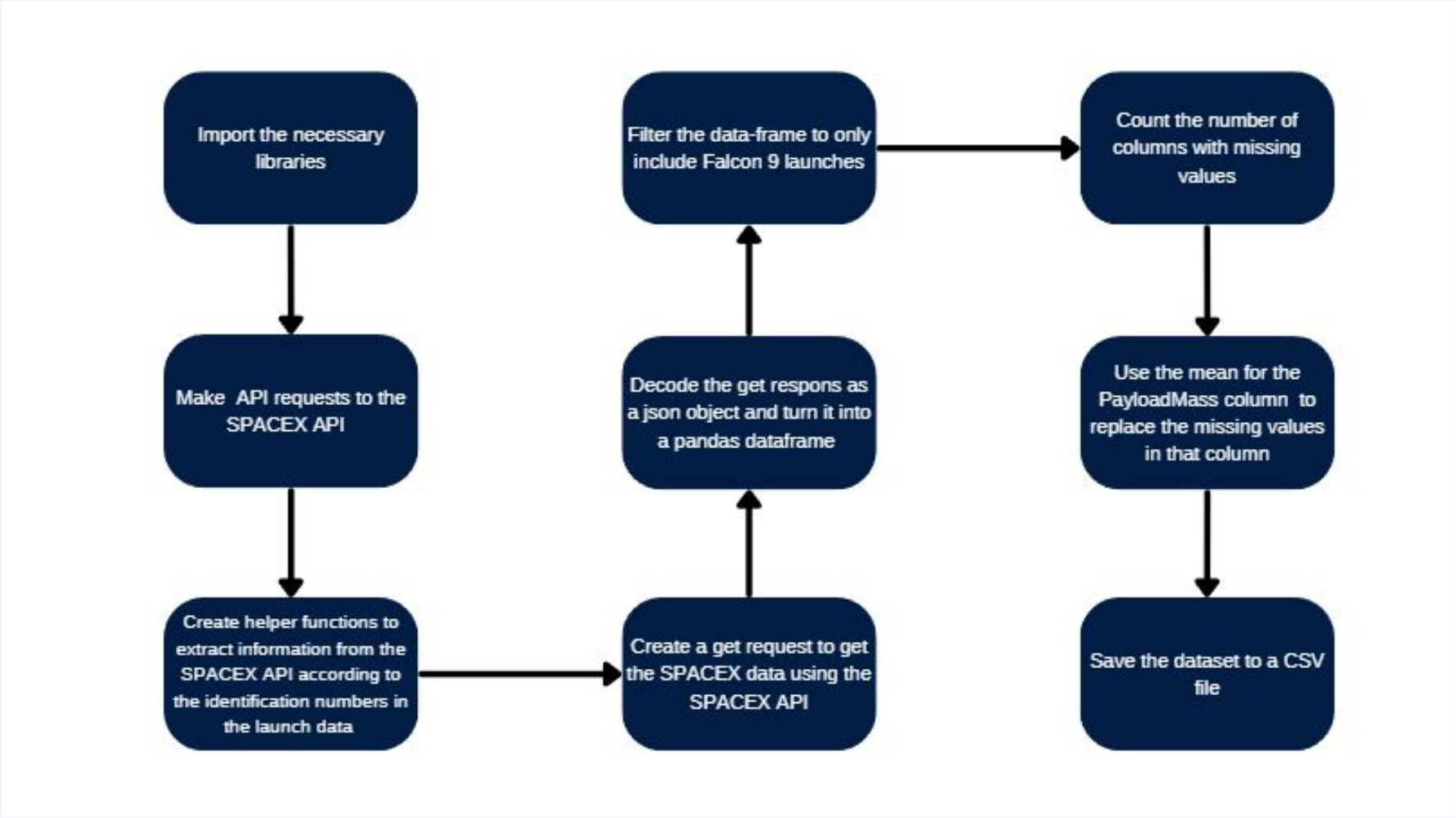
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

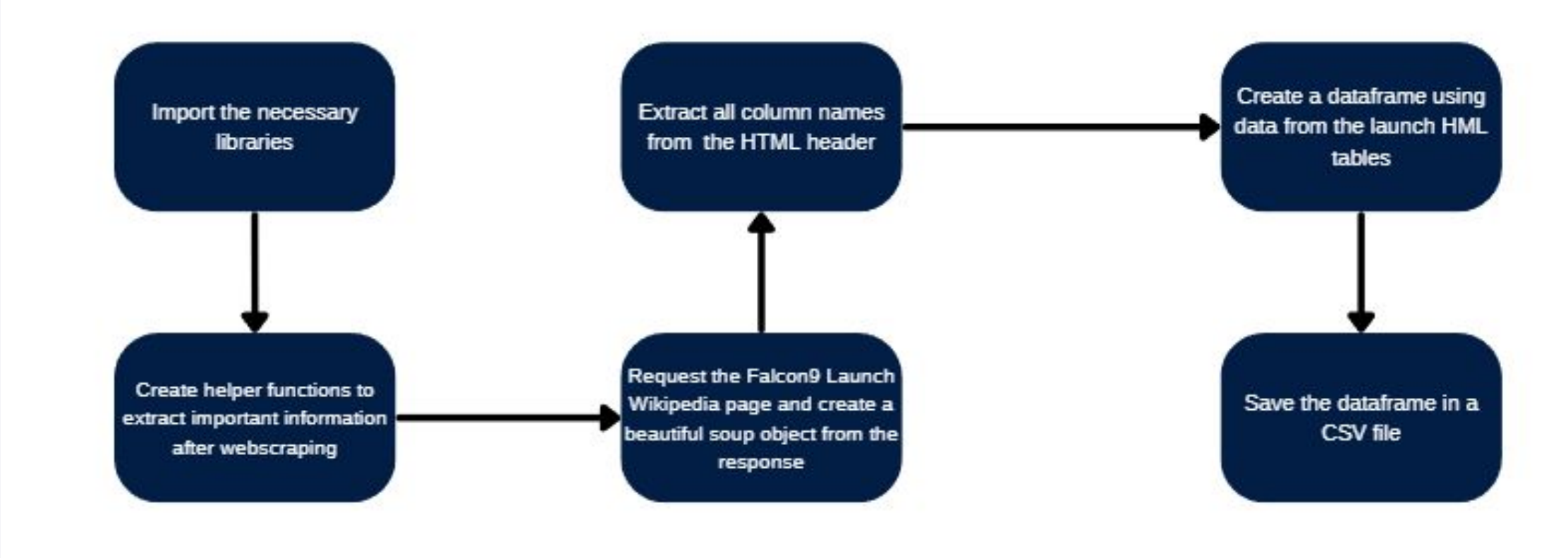
The data collection process involved collecting data from two different locations, namely the SpaceX REST API as well as web scrapping from the SpaceX Wikipedia page. The data from the Wikipedia page and API are then used together, for us to have a more in-depth analysis on the rocket launches. In turn providing us with better insight on the appropriate parameters to use for future launches so the most favorable outcome.

Data Collection – SpaceX API



Github link

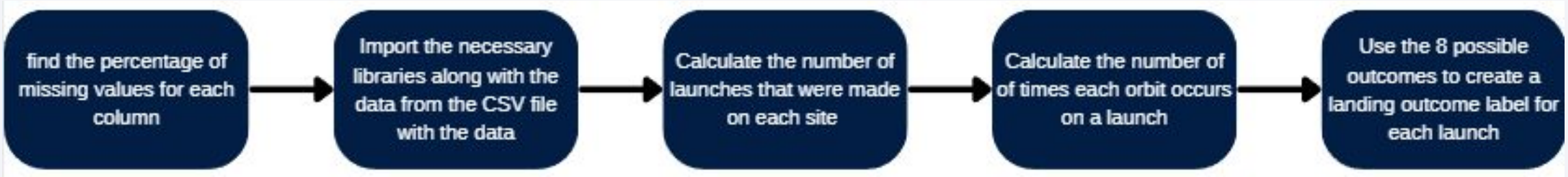
Data Collection - Scraping



Github link

Data Wrangling

We first perform statistical analysis to identify and calculate the number of missing values in each column, then we get the datatypes of each column. We then get the counts for the number of launches made at each of the 3 launch sites, the counts for the number of orbits made for each orbit, and then we can create landing outcome labels in numerical form for ease of use when we need to use the data later.



Github link

EDA with Data Visualization

- To visualize the relationship between flight number and launch site I used a scatter plot. This plot can show us which flights were successful at a specific launch site and which were not.
- To visualize the relationship between payload mass and launch site I used a scatter plot. This plot can show us that for the VAFB-SLC launch site there are no rockets launches for a payload greater than 100000
- To visualize the relationship between success rate of each orbit I used a bar chart. This chart can show us that the most successful orbits are ES-L1, GEO, HEO and SSO
- To visualize the relationship between flight number and orbit type I used a scatter plot. This plot can show us that LEO orbit flights became successful after more launches were made

Github link

EDA with Data Visualization (continued)

- To visualize the relationship between payload mass and orbit type I used a scatter plot. This plot can show us which orbit types are the most successful at a given payload and which orbits are best for larger payloads
- To visualize the relationship between launch success yearly trend I used a line plot. This plot can show us the improvement made in terms of successful launches made between 2010 and 2020

Github link

EDA with SQL

- Drop the SPACEX table if it already exists in the database. Then create the SPACEX table in the database.
- Use a select query to display the names of the unique sites in the space mission
- Use a select query to display 5 record where the launch sites begin with the string ‘CCA’
- Use a select query to display the total payload mass carried by the boosters launched by NASA
- Use a select query to display the average payload mass caried by booster version F9 v1.1

Github link

EDA with SQL (continued)

- Use a select query to list the date when the first successful landing on the ground pad occurred
- Use a select query to list the names of the boosters which have success in drone ship and also have a payload between 4000 and less than 6000
- Use a select query to list the total number of successful and failure missions
- Use a select query to list the names of the booster versions which have carried the maximum payload mass
- Use a select query to list the launch details for the months in 2015
- Use a select query to rank the count of landing outcomes between 2010/06/04 and 2017/03/20 in descending order.

Github link

Build an Interactive Map with Folium

- We extracted the launch sites with their respective longitude and latitude coordinates from the database. We then create a folium map object using NASA as our initial location, we use an orange circle with a marker for easy identification.
- After creating our initial map, and we can create a circle around the coordinates of a launch site to make the are for ease of identification.
- We then use markers with the name of the launch site based off the coordinates of the launch site so the area can be visible on the map.
- We can also use lines to show the distance between two specific coordinates, that way we know how far apart specific locations are from each other.

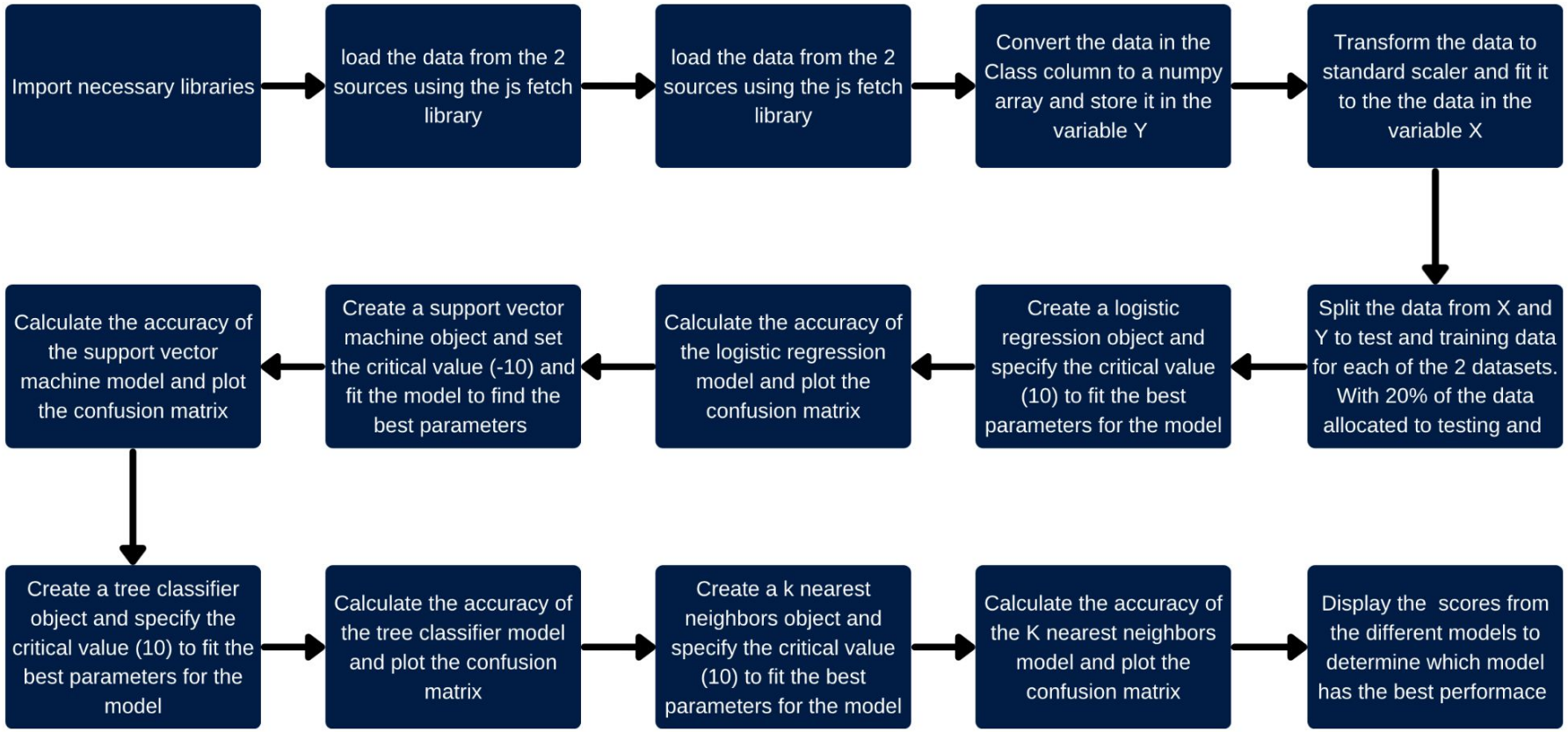
Github link

Build a Dashboard with Plotly Dash

- A dropdown list was added to allow ease of launch site selection, with all the launch sites as the default selection.
- Pie charts are used to show the total successful and unsuccessful launches when a specific launch site is selected. By default all the percentage of all launches per launch site are displayed. Pie charts are an intuitive way to easily view the impact of each launch site in relation to specific criteria.
- Scatter plots are a great way to represent the relationship between payload (independent variable) and launch success (dependant variable). The scatter plot shows the launch status of the launches at a specific launch site using specific booster versions and payloads, if all launch sites are selected, display all the launch sites.
- The slidebar with all the payloads is added to adjust the payload range to display on the scatter plot.

Github Link

Predictive Analysis (Classification)



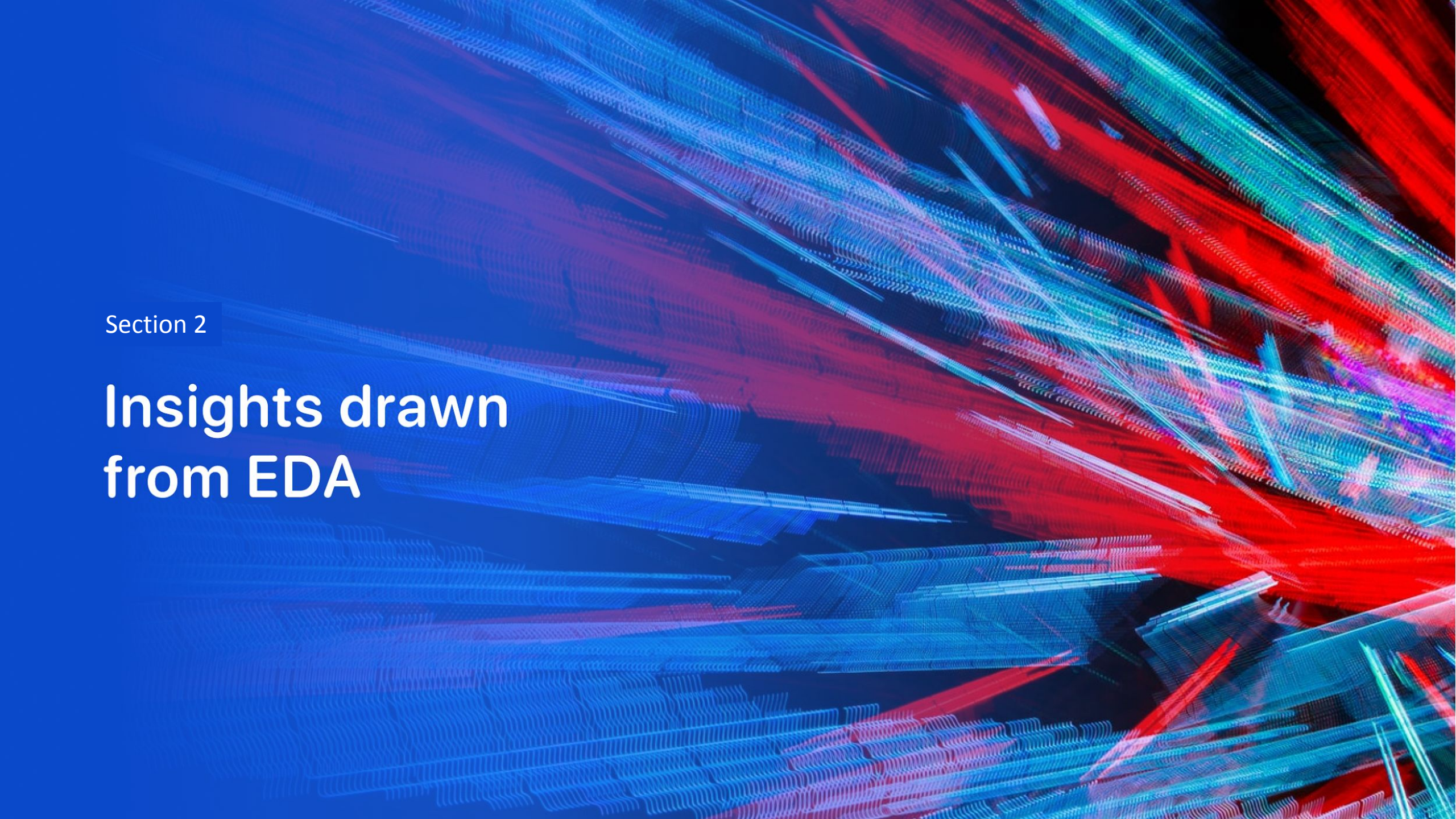
Github Link

Results

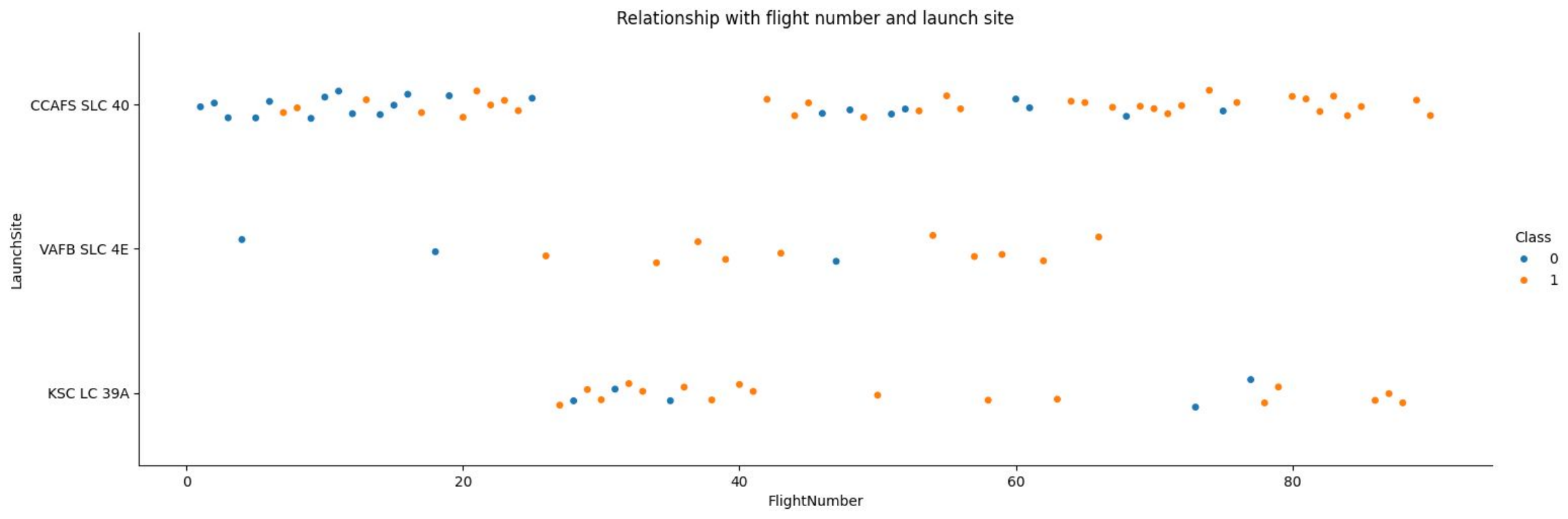
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

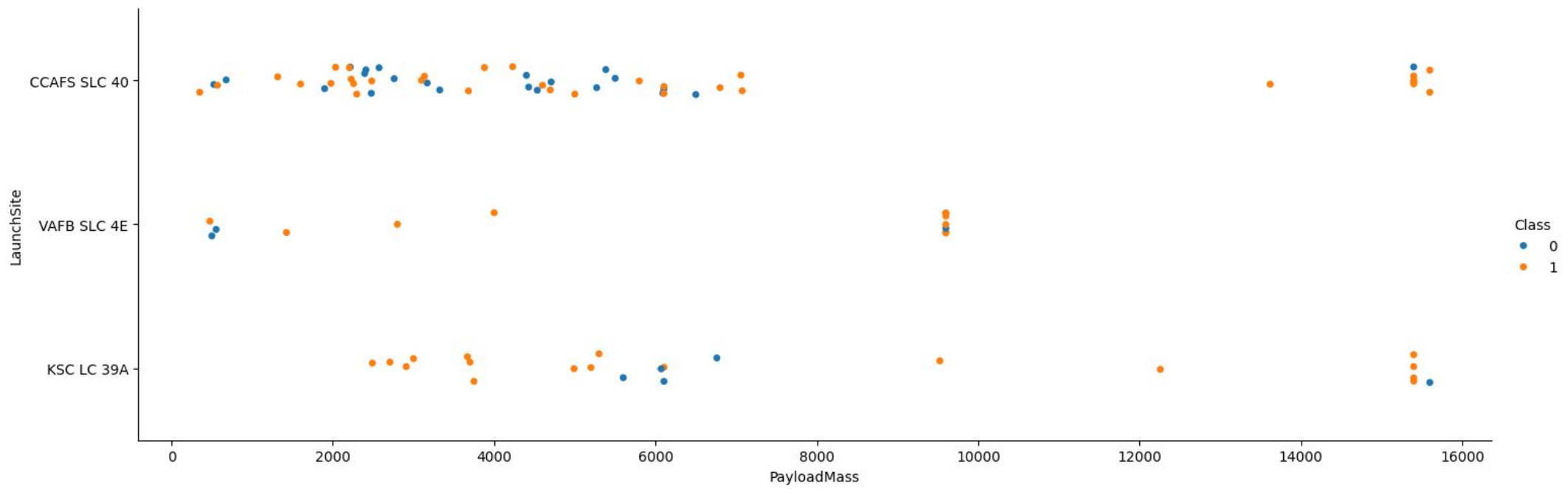


Flight Number vs. Launch Site



- The success rates for the launches at launch site CCAFS SLC 40 increased as the number of flights increased.
- As the number of flights for the launch Site VAFB SLC 4E increased, there is a lesser number of failed launches.
- Majority of the initial flights failed then over time the amount of failed launches decreased.

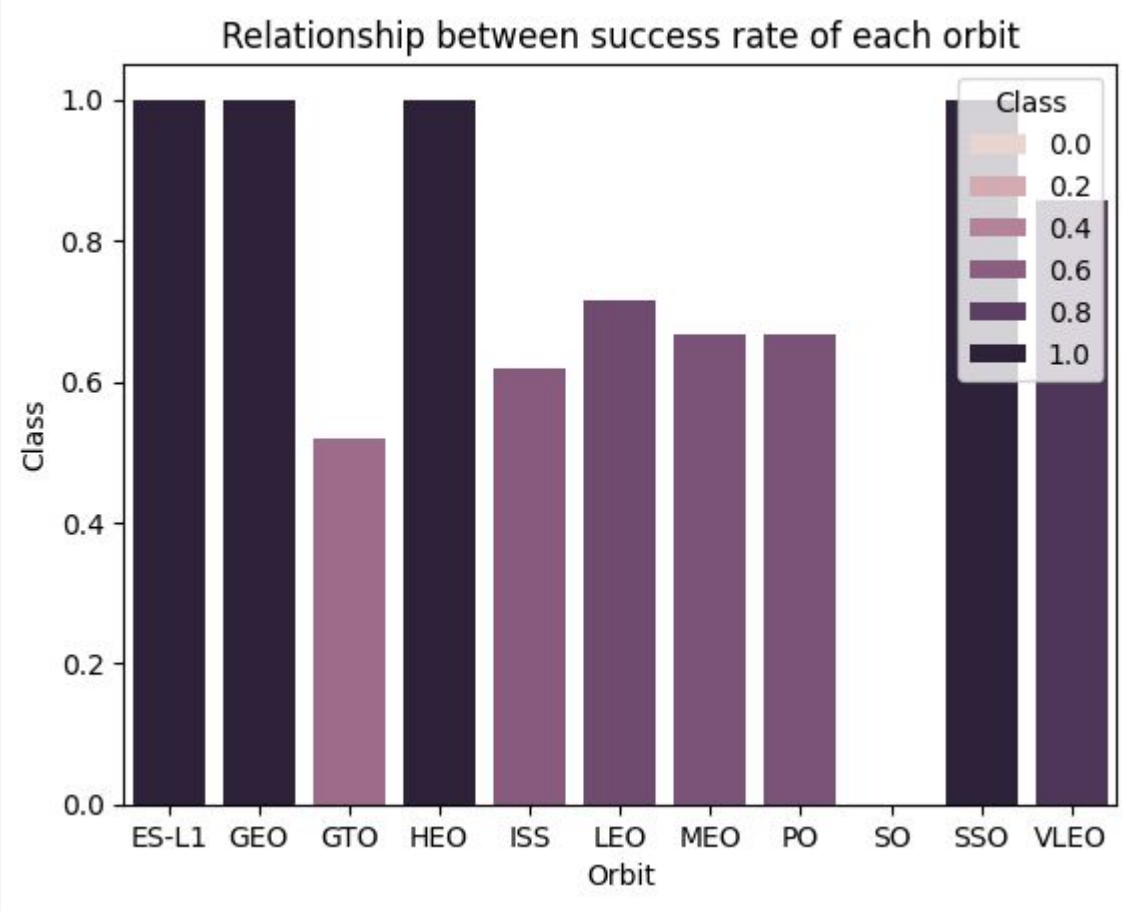
Payload vs. Launch Site



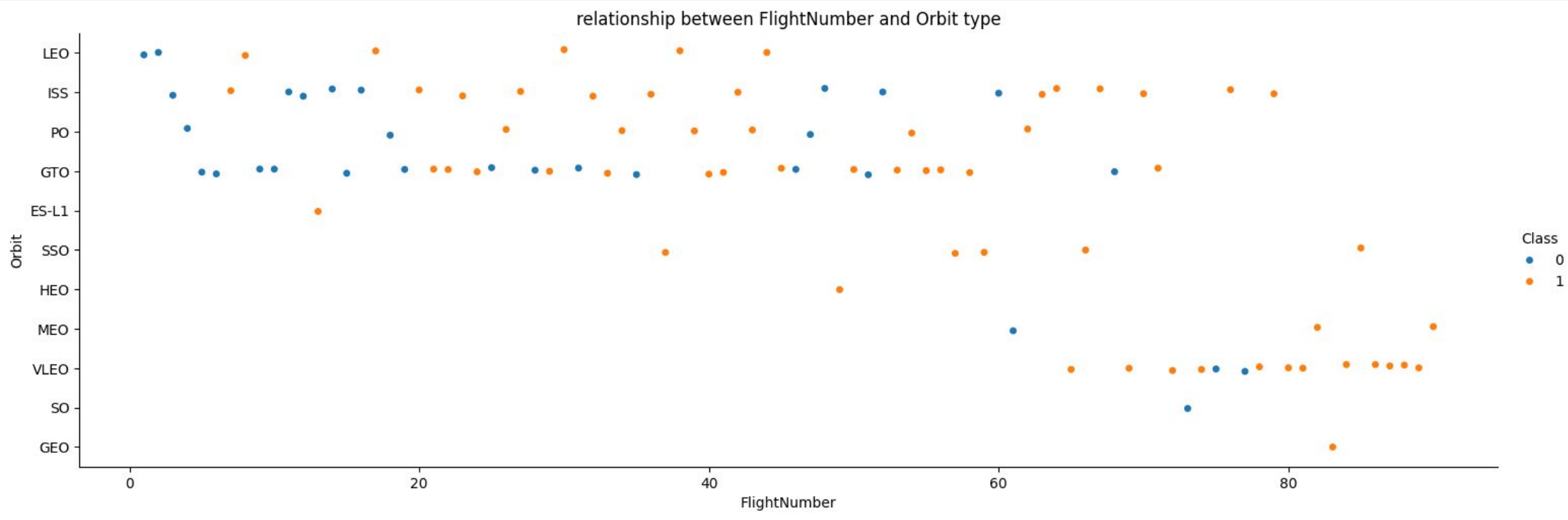
- The launches made at all 3 launch sites have a higher frequency of successful launches as the payload increases.
- The launches made at launch site KSC LC 39A have a highest frequency of successful launches for flights with a payload mass that is less than 6000 kg

Success Rate vs. Orbit Type

- The flights around the ES-L1, GEO, HEO and SSO orbits have a 100% success rate meaning that Space X can reuse the first stage of the rocket.
- The flights around the SO orbit completely failed with a 0% success rate.
- The flights around the GTO, ISS, LEO, MEO, PO and VLEO orbits have a success rate between 50% - 80%.

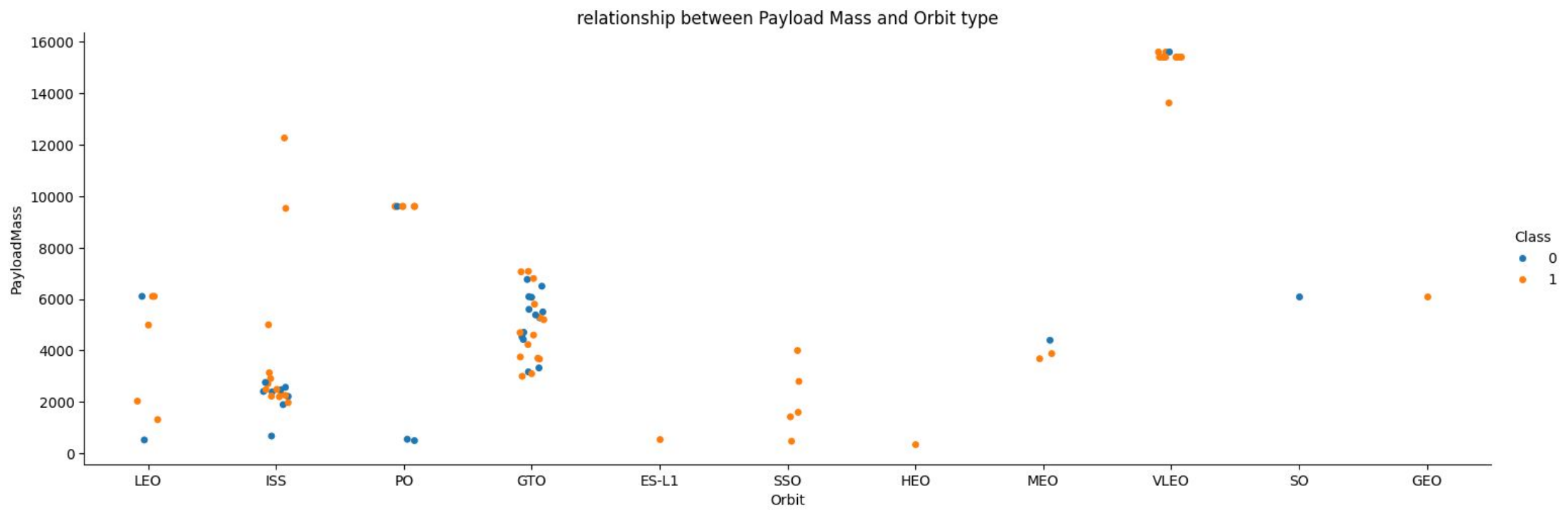


Flight Number vs. Orbit Type



- The flights around the LEO and MEO orbits only had failures in the earlier flights, but as the flight numbers increased, the success rate increased as well.
- The flights around the ES-L1, SSO, HEO and GEO all have continuously successful flights from their earliest flights to their latest ones.
- The flights around the GTO orbit do not seam to have any progress in terms of successful flights as the number of flights increase.
- The flight around the SO orbit was a failure, this orbit has no record of successful launches. This is the only orbit with a record of all failed flights.

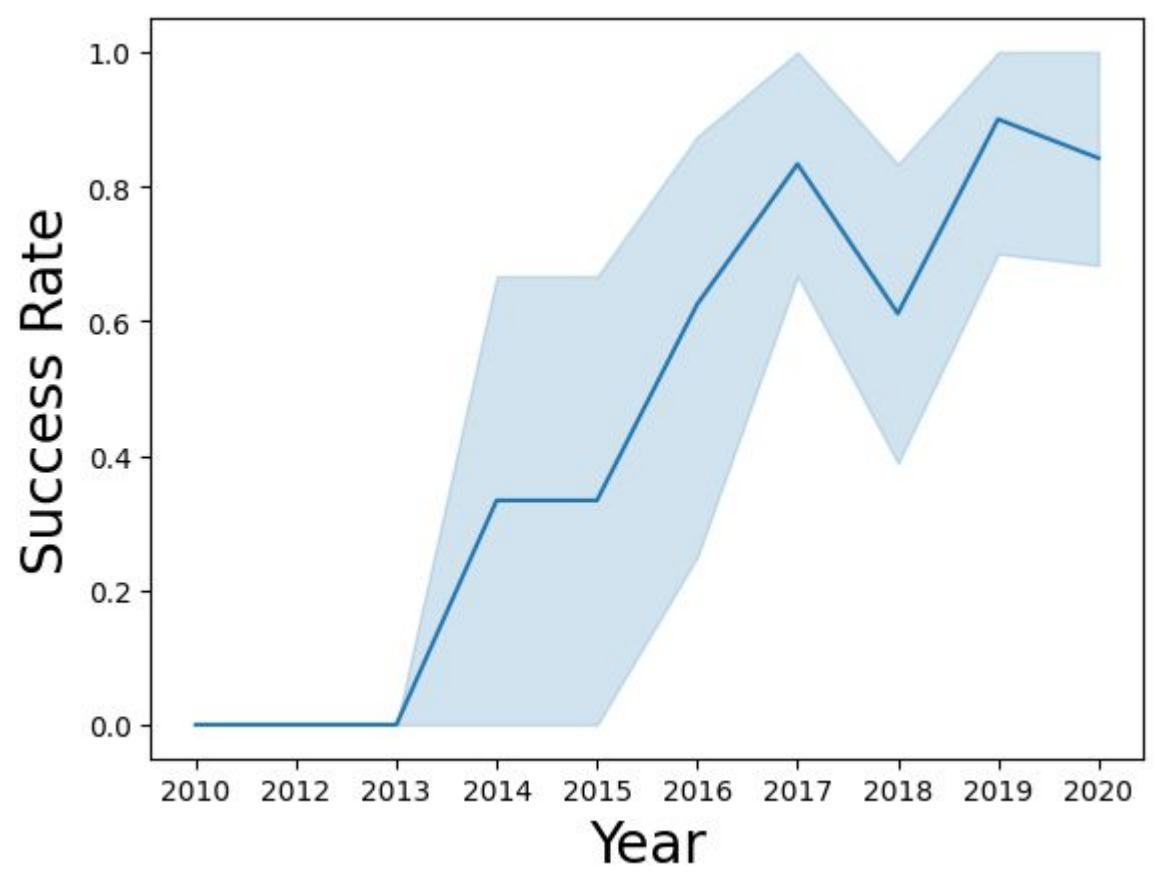
Payload vs. Orbit Type



- The flights around the ISS orbit have an increase in their success rate as the payload mass of each rocket increases.
- The flights around the ES-L1, HEO and GEO orbits have a 100% success rate on all fights even though the payload mass increases.
- None of the flights around the SO orbit were successful.
- Flights around the LEO, PO, GTO, MEO and VLEO orbits do not seem to have any significant improvements regardless of the payload mass.

Launch Success Yearly Trend

- The line chart shows an overall increase in the success rate of each launch over the years.
- There is a decline in the success rate in the year 2018.



All Launch Site Names

- This query displays the unique names of the launch sites used to launch a flight.
- To get the result, select the distinct launch sites from the SPACEX data

[10]:

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

Launch Site Names Begin with 'CCA'

[12]:

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

- This query displays the records of data where the Launch site name begins with ‘CCA’
- To get this result, display all the rows where the launchsite begins with CCA and limit the output to 5 records only.

Total Payload Mass

- This query displays the total payload carried by the NASA boosters
- To calculate the total payload carried by boosters from NASA, select the rows where NASA is the customer and find the sum of the payload mass column

[12]:

Total_payload
45596

Average Payload Mass by F9 v1.1

- This query is used to calculate the average payload mass carried by booster version F9 v 1.1
- To calculate the average payload mass carried by booster version F9v1.1, select the rows where the Booster_Version column contains F9 v 1.1 then calculate the average of the Payload Mass column

[15]:

Average_payload
2928.4

First Successful Ground Landing Date

- This query is used to find the dates of the first successful landing outcome on ground pad
- To display the date of the first successful landing, select the rows where the Landing Outcome column is Success (ground pad) then display the earliest date using the min function.

[16]:

First_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

[17]:

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

- This query is used to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- To get this result select the rows where the entries in the Payload mass column is between 4000 and 600 kgs and the entries in the Landing Outcome column entries are Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

[21]:

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

- This query is used to calculate the total number of successful and failure mission outcomes
- To get the result we select the rows that are grouped by the Mission outcome column, then count the total successful missions and display the value count.

Boosters Carried Maximum Payload

- This query is used to list the names of the booster which have carried the maximum payload mass
- To get these results, we select the Booster Version and Payload mass for the rows where the payload mass is the maximum payload in the table

[22]:

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

[27]:

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

- This query is used to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- To get this result selects the month from the Date column, Booster version, Launch Site and Landing Outcome, where the Landing Outcome is Failure (drone ship), and the year extracted from the Date column is 2015.

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

[28]:

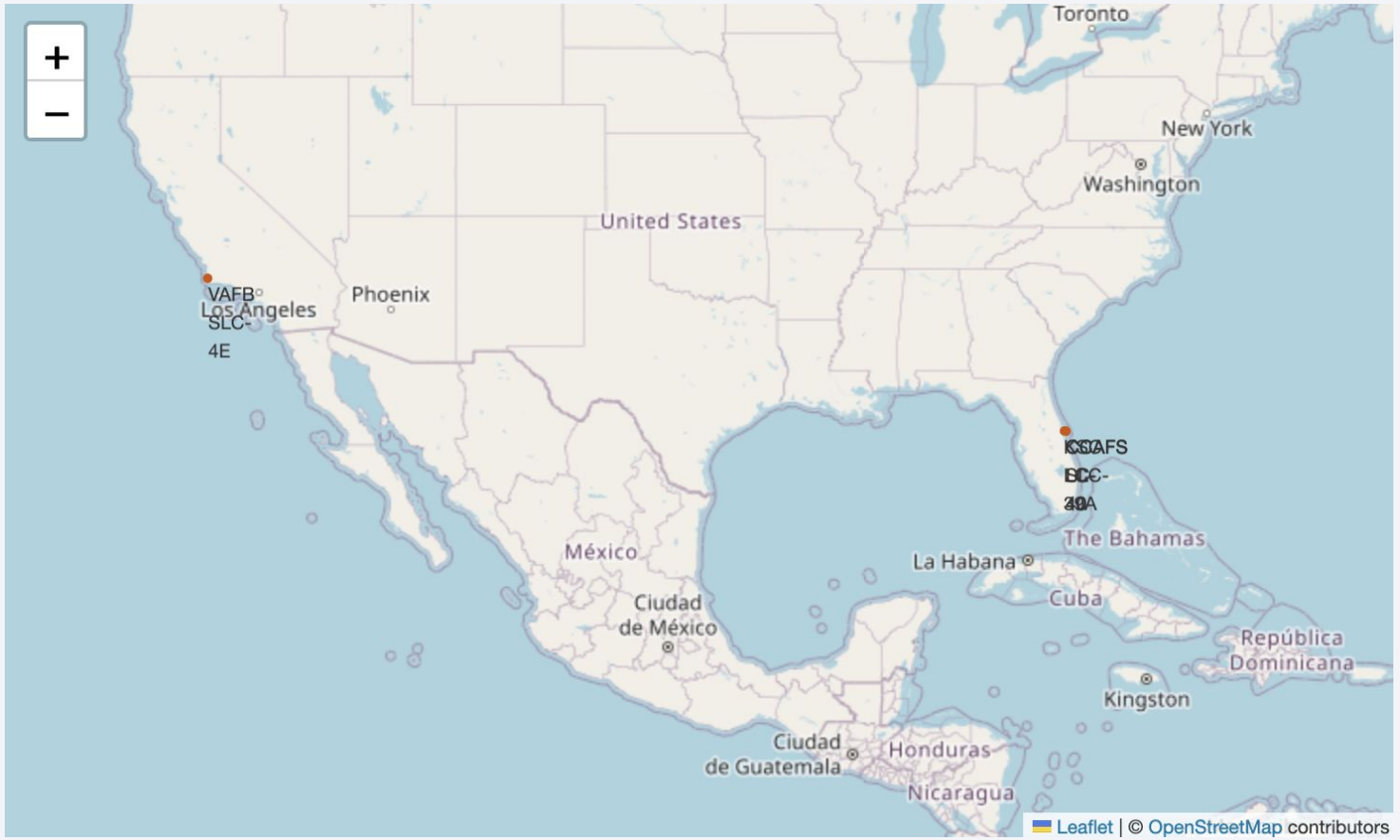
Landing_Outcome	Count_Outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query analyses the frequency of the different landing outcomes for SpaceX launches between 4 June 2020 and 20 March 2017.
- To get this result, the query selects the Landing Outcome and counts how many times each outcome appears, where the Date is between 2010-06-04 and 2017-03-20. The results are grouped by Landing Outcome and sorted in descending order based on the count

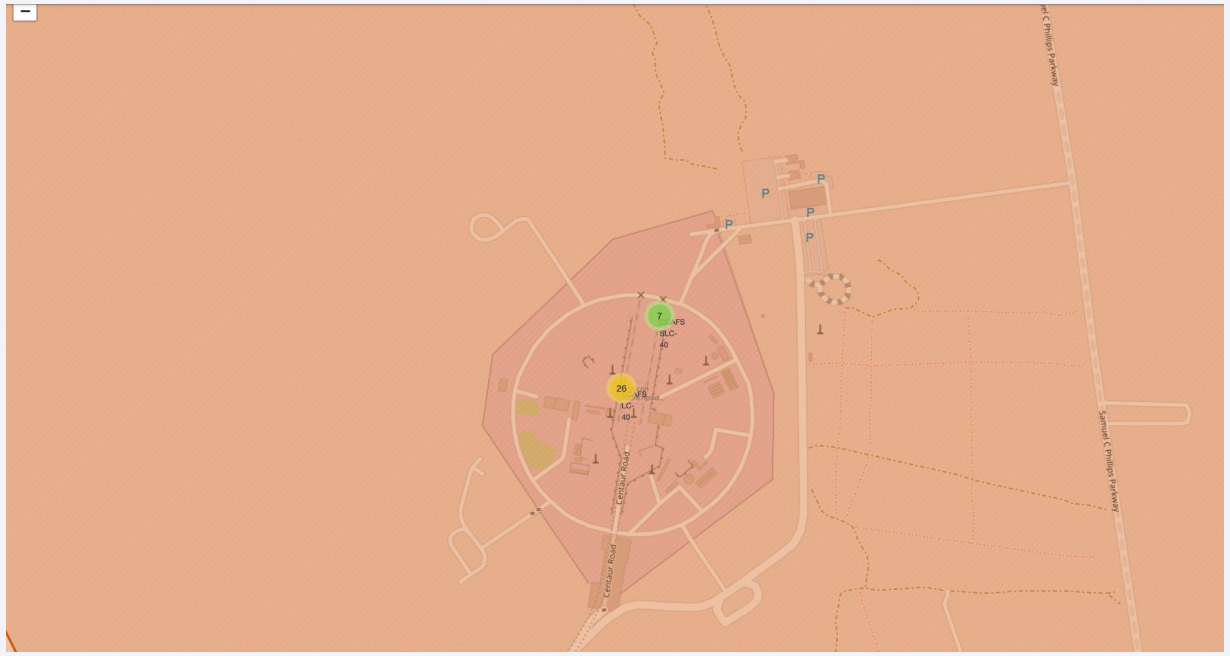

Section 3
Launch Sites
Proximities Analysis

Launch site markers on global map

- This map shows the launch sites used to laugh rockets. They are marked with red circles of ease of identification then labelled according to their name.

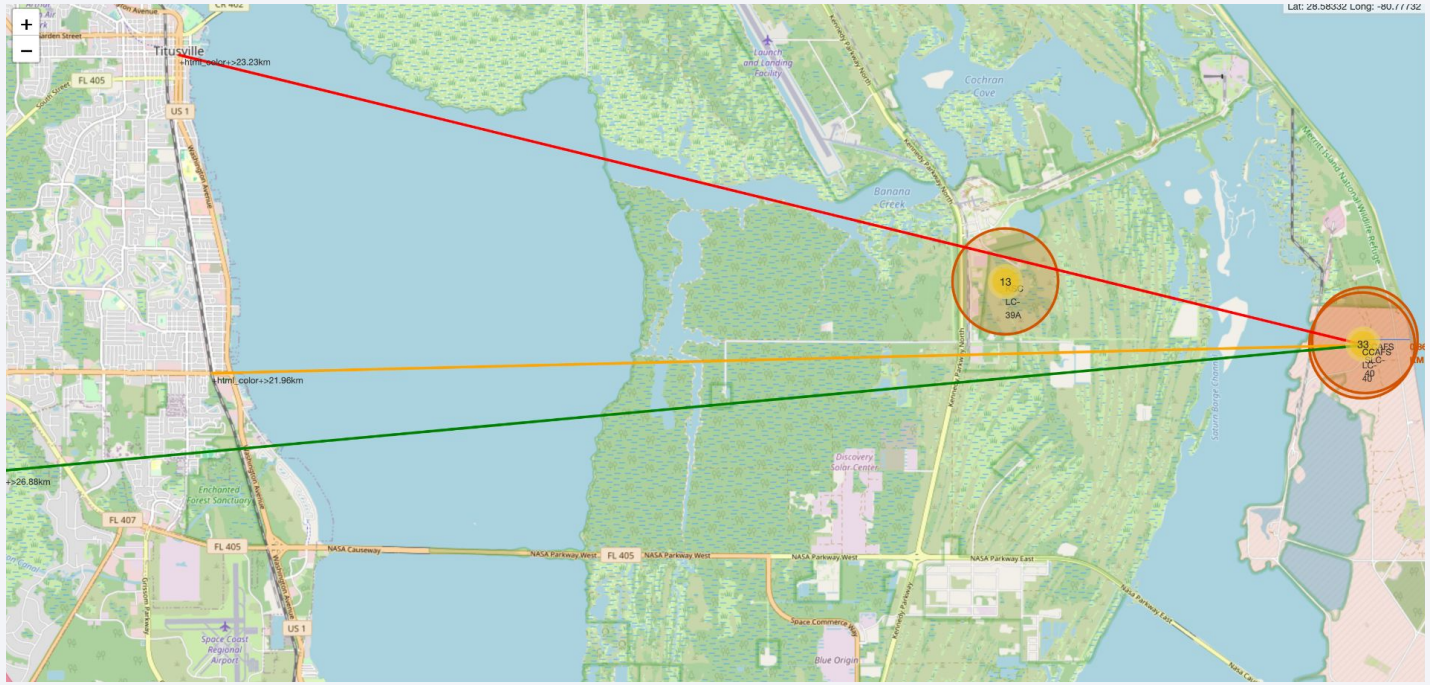


Launch outcomes

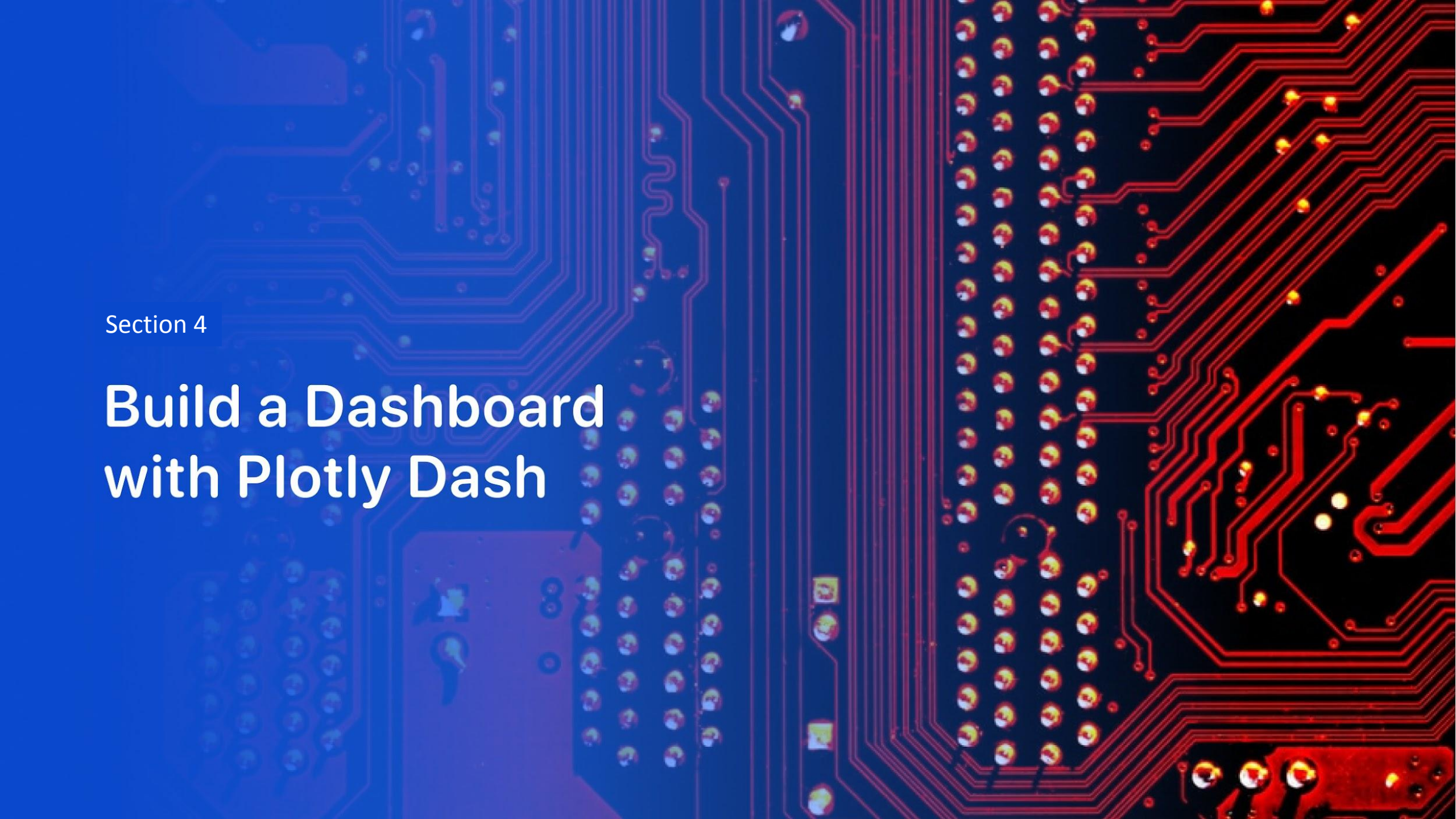


- The map displays the different launch outcomes for each launch site, with the green representing the successful launches and the red the failed launches

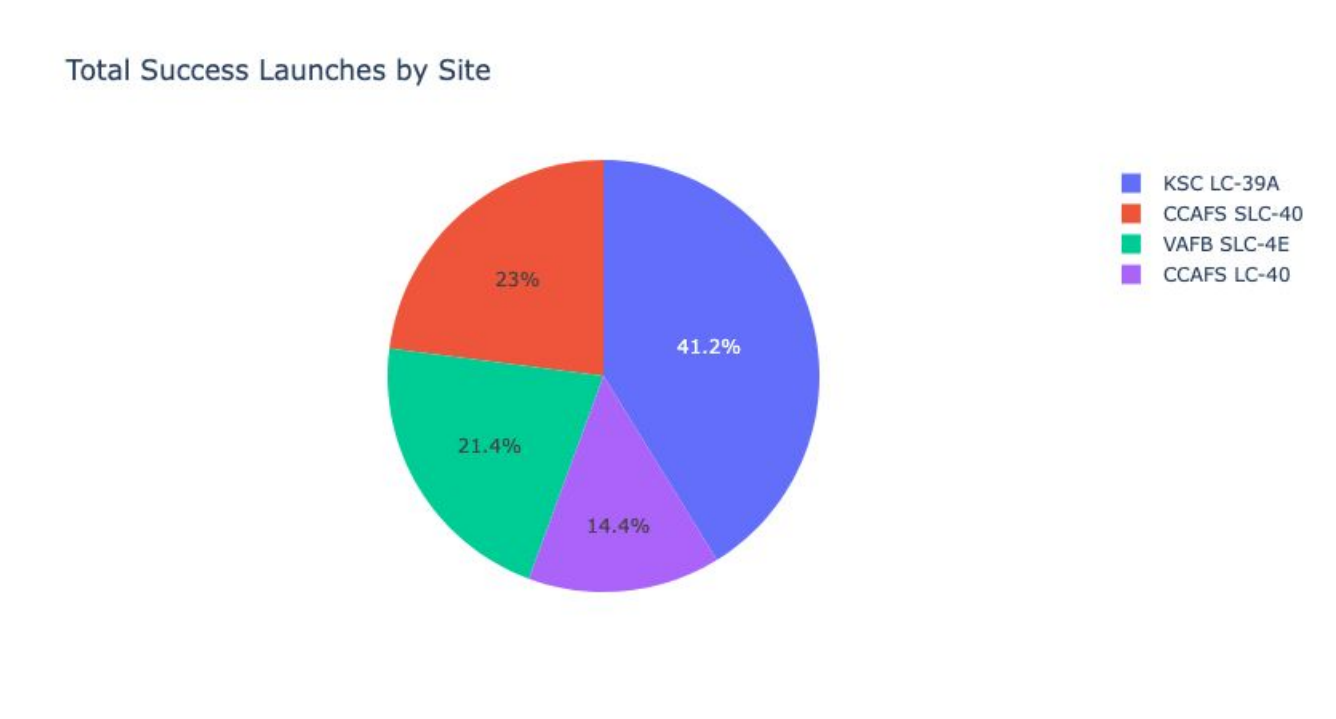
Launch site distance from various proximities



- The red line displays the distance between the launch sites and the city
- The orange line displays the distance between the launch sites and the railway
- The green line shows the distance from the launch sites and the highway
- The blue line shows the distance between the launch sites and the coast

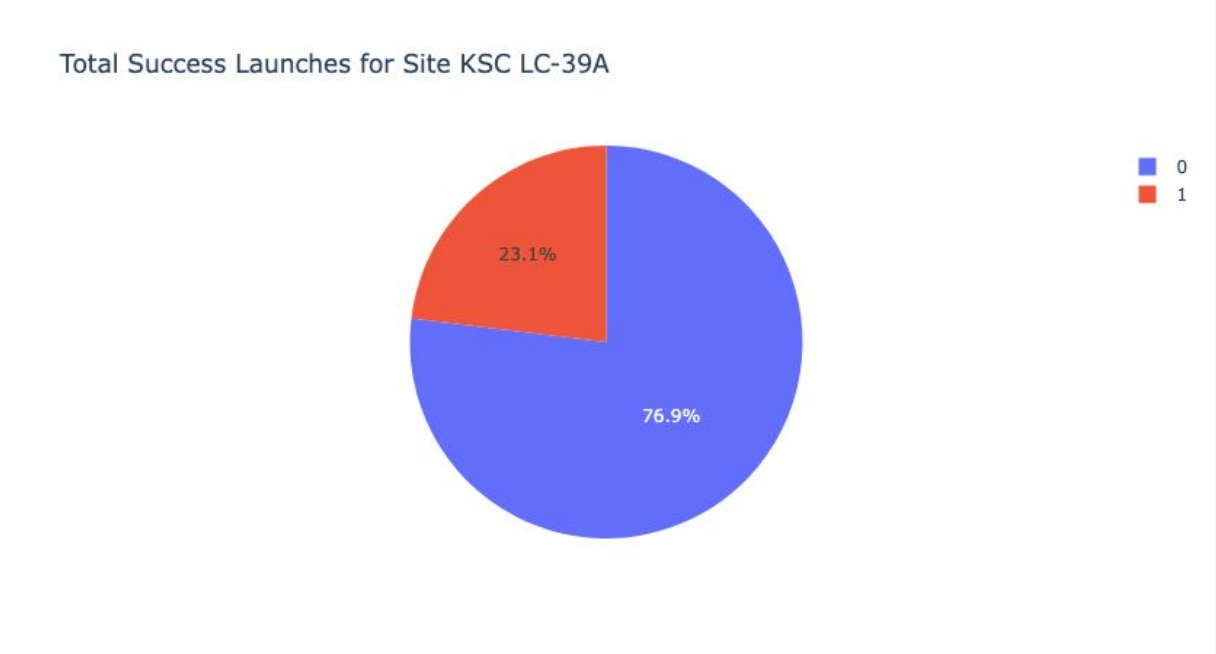

Section 4
Build a Dashboard with Plotly Dash

Total successful launches by site



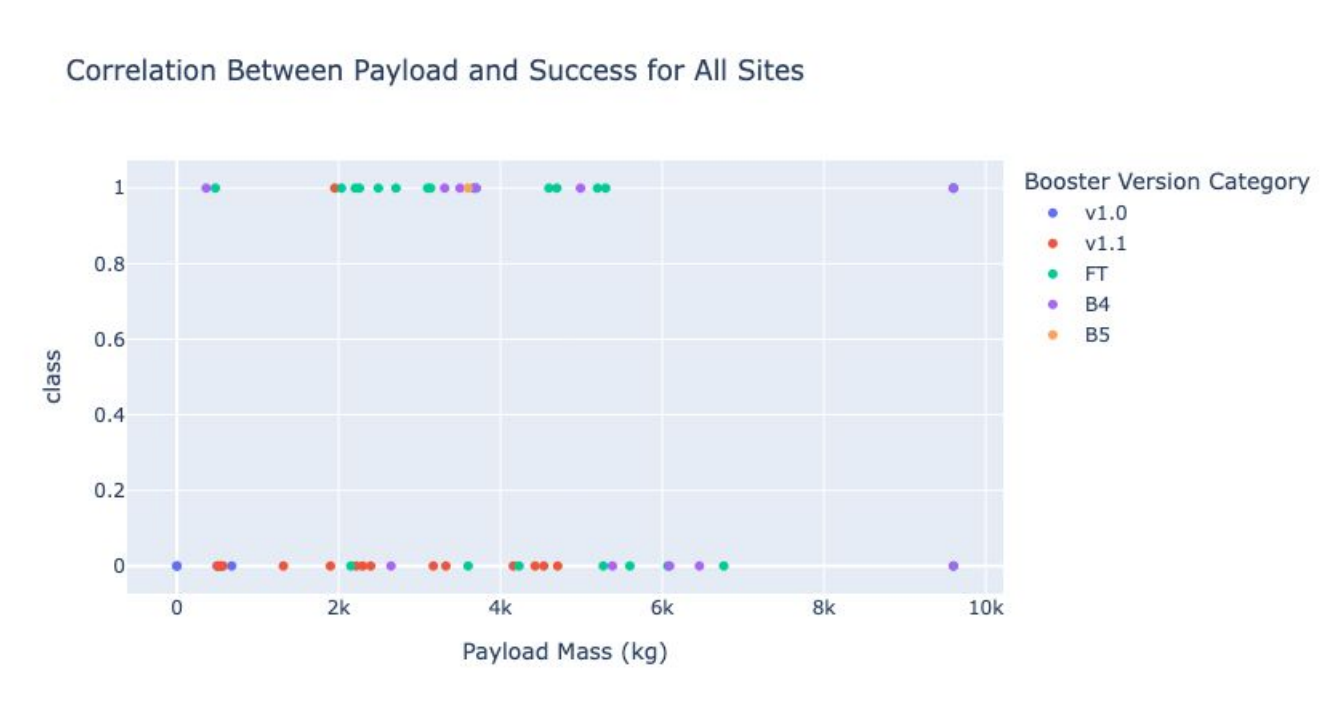
- The launch site with the most successful launches is KSC LC-39A
- The launch site with the least amount of successful launches is CCAFS LC-40

Highest launch success



- The launch site with the highest launch success ratio is KSC LC-39A, with a success rate of 76.9% and a 23.1% failure rate

Payload vs success for all sites



- The scatter plot displays the success and failure launches for each booster version. We can see that the FT and B4 boosters have more frequent success rates

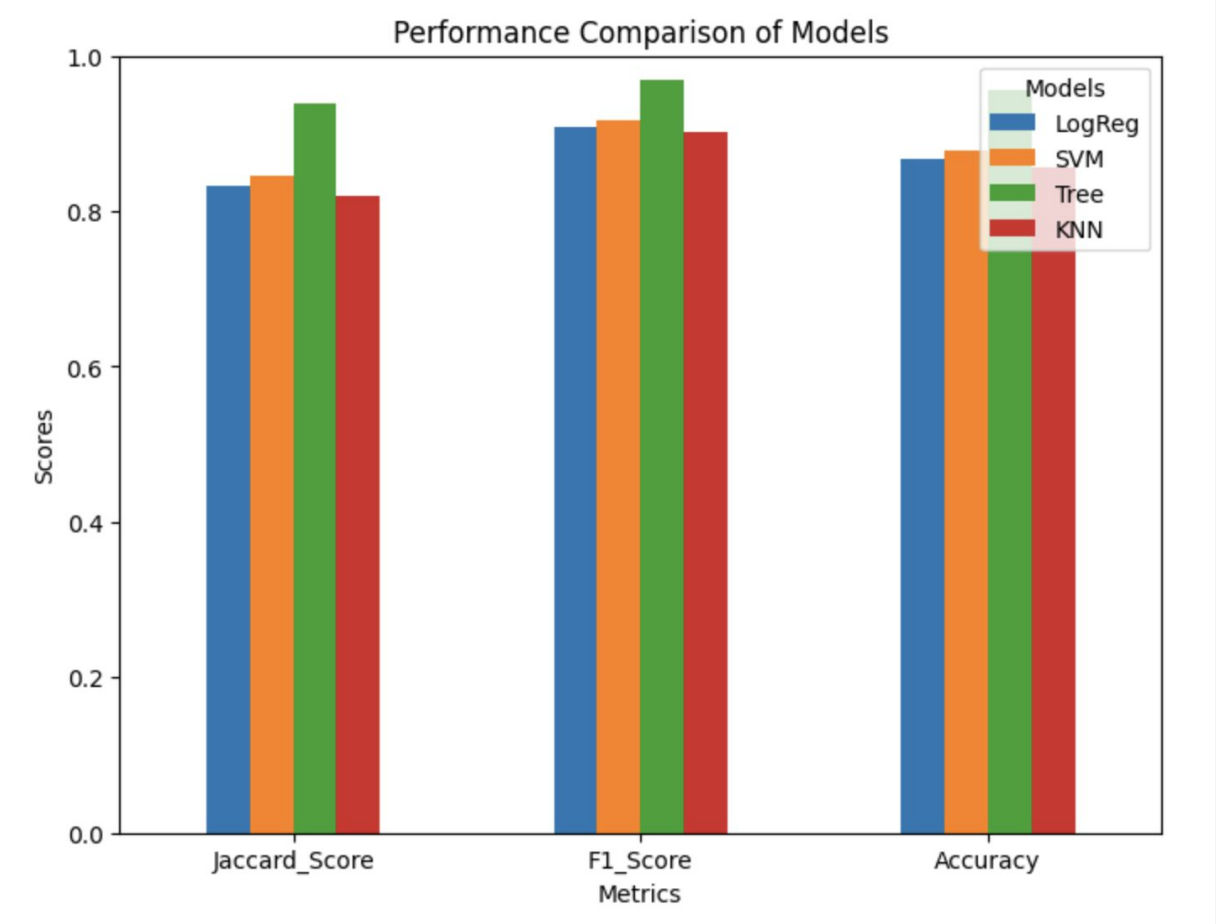
Section 5

Predictive Analysis (Classification)

Classification Accuracy

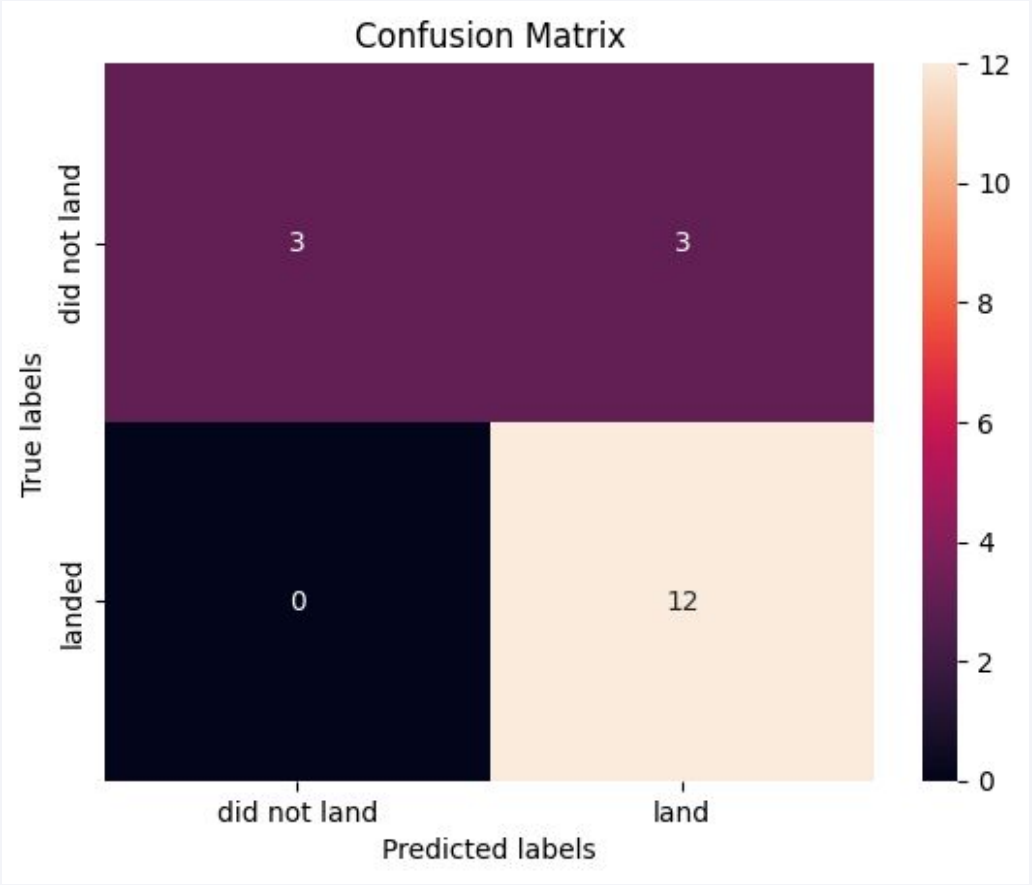
- The decision tree has the highest performance scores compared to the other classification models.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.840580	0.819444
F1_Score	0.909091	0.916031	0.913386	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556



Confusion Matrix

- Confusion matrix for the decision tree.
- There are 12 true positives, which are launches that were predicted to land that actually landed
- There are 3 true negatives, which are launches that were predicted not to land and did not land
- There are 0 false positives, there are no launches that were predicted to not land that did land.
- There are 3 false positives in the model, meaning these launches were predicted to land but they did not land



Conclusions

- The Decision Tree model is the most effective in making correct predictions overall.
- SVM and Logistic Regression show close performance, with SVM slightly outperforming LogReg in all three metrics. This indicates that both models are reliable, but SVM may have a slight advantage.
- KNN has the lowest scores in all metrics, with an Accuracy of 0.8556, indicating that it may not generalize as well as the other models. This suggests that KNN may be more sensitive to noise or less effective in capturing decision boundaries compared to the other models.
- Overall, If high accuracy and overall performance are the priority, the Tree model is the best choice. If a balance between performance and computational efficiency is needed, SVM or Logistic Regression might be suitable.

Appendix

- Github link: https://github.com/olivia9469/capstone_datascience.git

Thank you!