IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sairam Venkatachalam
29th May 2023

Outline

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

Executive Summary

Methods Used:

- Data collection: Combined SpaceX API and webscraping Wikipedia to obtain Falcon 9 launch data.
- Data wrangling: Used pandas dataframes to process the data and map landing outcomes to 'Class' 0 and 1.
- Exploratory data analysis (EDA): Employed visualization and SQL queries to explore the data and derive key metrics.
- Interactive visual analytics: Utilized Folium and Plotly Dash for interactive analysis, focusing on geo-spatial insights.
- Predictive analysis: Conducted classification modeling on cleaned data after scaling and categorical variable conversion.

Results showcased:

- EDA insights: Identified patterns, trends, and correlations through data exploration and visualization.
- Interactive analytics demo: Demonstrated interactive analysis capabilities with relevant screenshots.
- Predictive analysis outcomes: Trained machine learning models to predict landing outcomes successfully.

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

Project Overview

The Client

The client, **SpaceY**, is interested in competing with **SpaceX** in providing cost effective payload delivery to outer space, by successfully re-using the stage 1 part of these rockets. They are interested in being able to predict the likelihood of successfully landing the first stage of the rocket launches, in order to better estimate the cost for these launches.

Successfully landing of a the 1st stage on a drone ship

The Ask

The objective of this project was to perform data analysis on previous **SpaceX** launches and build a predictive model to effectively capture trends and patterns which will help predict the first stage landing outcomes of future launches.

Introduction

Key Questions to be answered

- What is the significance of various parameters such as launch site and orbit in determining the final landing outcome?
- How have launch success rates have changed over time?
- How are launch sites strategically located in proximity to key locations?
- Which parameter plays the greatest role in determining success?
- How do the top classification algorithms perform on test data?

Section 1

Methodology

Methodology (1/2)

Executive Summary

- Data collection methodology:
  - The data on Falcon 9 rocket launches was collected using a combination of the SpaceX API and by webscraping Wikipedia to obtain Falcon 9 launch data
- Perform data wrangling
  - The data was processes using pandas dataframes. The main task performed here was to map the different landing outcomes to 'Class' 0 and 1 to indicate success of the landing
- Perform exploratory data analysis (EDA) using visualization and SQL
  - The collected data was uploaded to sqlite in order to use SQL queries to explore the data and derive key metrics and outputs

Methodology (2/2)

Executive Summary

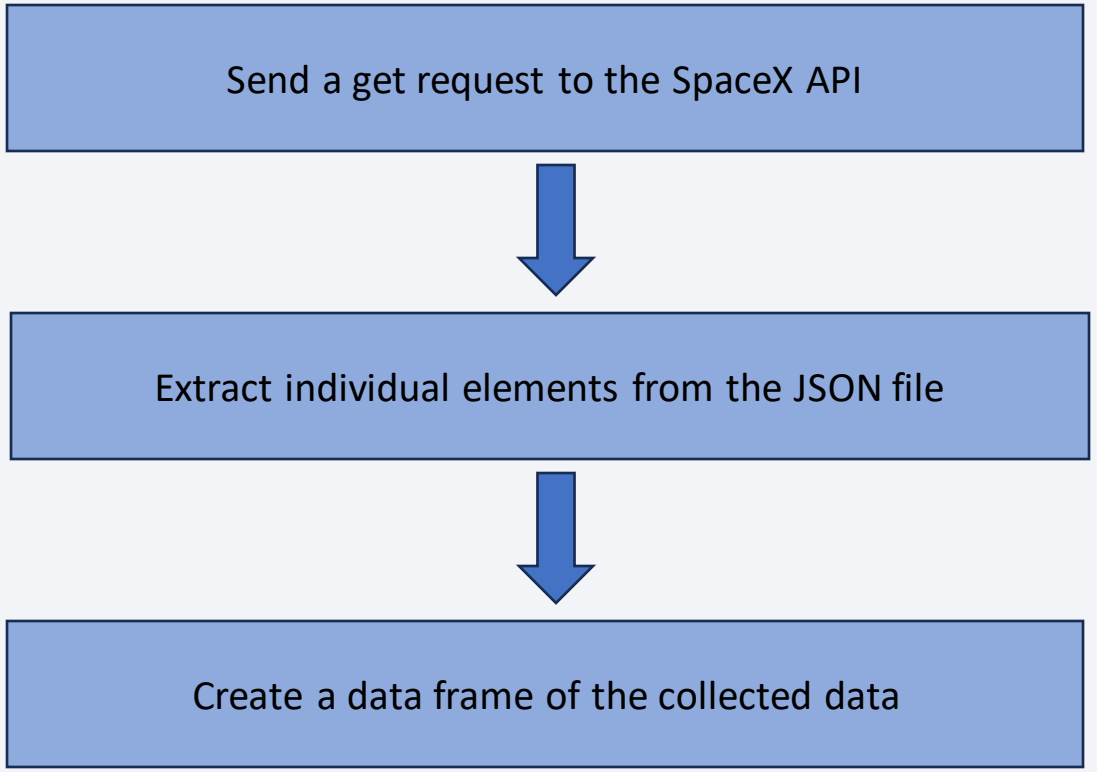
- Perform interactive visual analytics using Folium and Plotly Dash
  - Using the latitude and longitudes data points, geo-spatial analysis was performed on the launch data to identify success rates and proximities to key locations
- Perform predictive analysis using classification models
  - Once the data was cleaned, it had to be prepared for modelling by scaling it and converting categorical variables
  - The processed data was then split into a training and testing set
  - The training set was used to train various machine learning models to effectively be able to predict landing outcomes

Data Collection

- The data collection process was conducted utilizing two methods.
- In the initial approach, the SpaceX API was directly employed to acquire a JSON file encompassing launch records for the Falcon 9 rocket. Subsequently, the requisite attributes were extracted from the file.
- In the subsequent approach, webscraping techniques were employed on a Wikipedia page dedicated to Falcon 9 launches. The tables present on the page were parsed to retrieve the column names and corresponding data points.

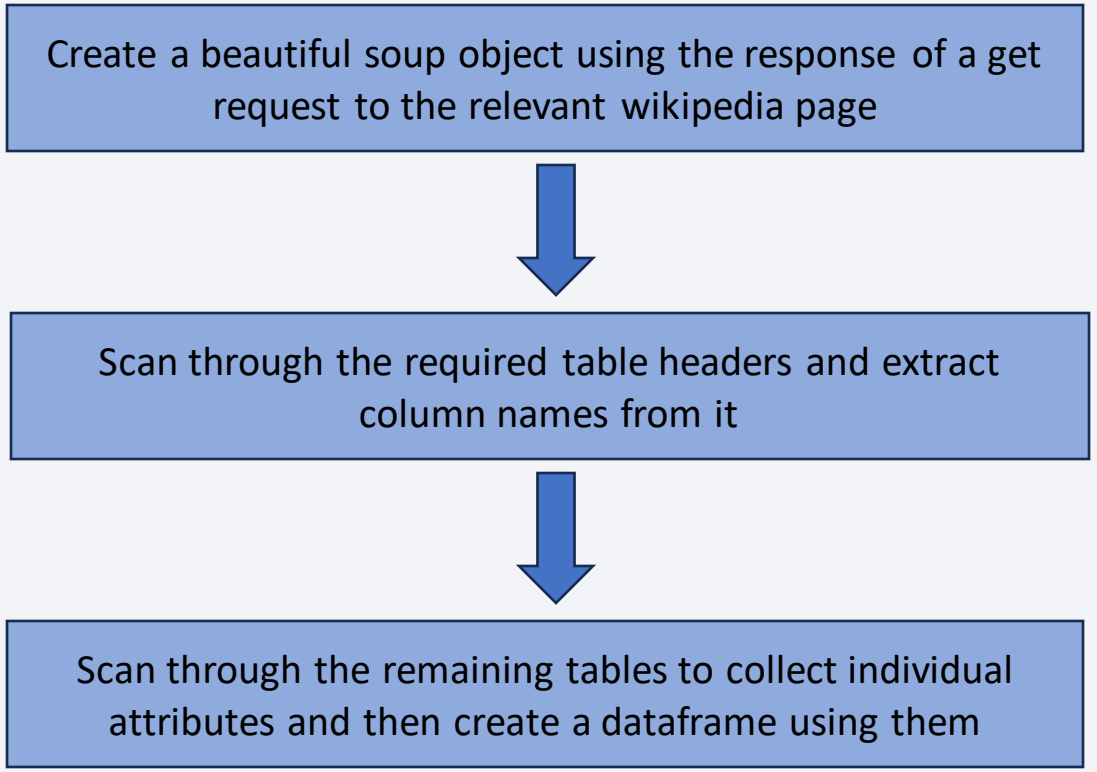
Data Collection – SpaceX API

- The SpaceX API was used to get data for Falcon 9 launches
- Using a get request we generate a json file and extract the required parameters from the individual attributes present
- Finally, we use these attributes to create a dataframe for further analysis
- Please refer to the link for the code



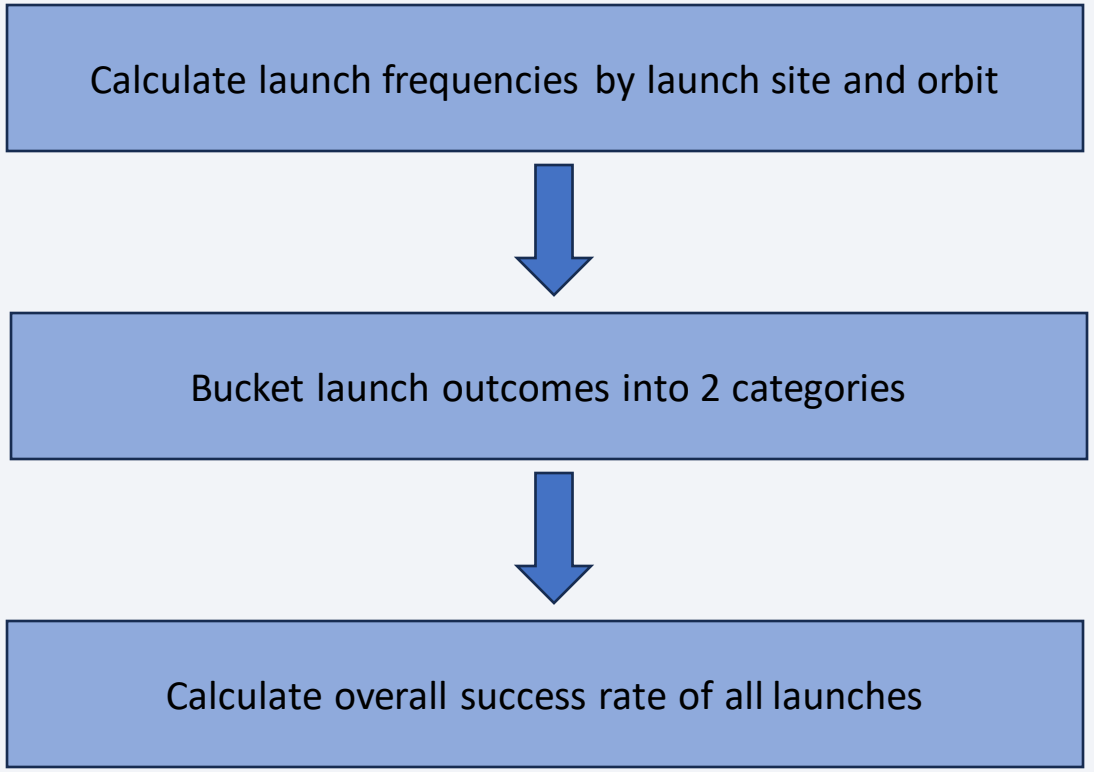
Data Collection – Webscraping

- Create a get request to the wikipedia page with Falcon9 launches
- Using Beautiful Soup, we create an object with the response from the get request
- We scan through the table containing the first launch in order to obtain the column names. Using these column names, we extract data from all tables on the page
- Please refer to the link for the code



Data Wrangling

- The data was processed using pandas dataframes
- Firstly, the launch counts were extracted by launch site and orbits in order to get an idea of data frequencies
- Further, the landing outcomes were mapped onto 2 classes '0' and '1' to represent success and failure
- Finally the mean of the class was calculated. This represents the overall success rate of all launches



EDA with Data Visualization

- A variety of scatter plots were generated to visualize the relationship between attributes such as flight number, payload mass, launch site and orbit. The objective was to determine the correlation between the combination of these attributes and the success rates for these flights
- Apart from this, the success rates by orbit was also visualized
- Please refer to the link for the code

EDA with SQL

- The data was examined by employing an extensive array of SQL queries. These queries encompassed various aspects, such as generating a list of unique values for booster names, landing outcomes, and launch sites.
- Additionally, an analysis was conducted of aggregated statistics, including payload mass categorized by booster and launch site. Furthermore dates were ranked based on the number of successful launches.
- Please refer to the link for the code

Building an Interactive Map with Folium

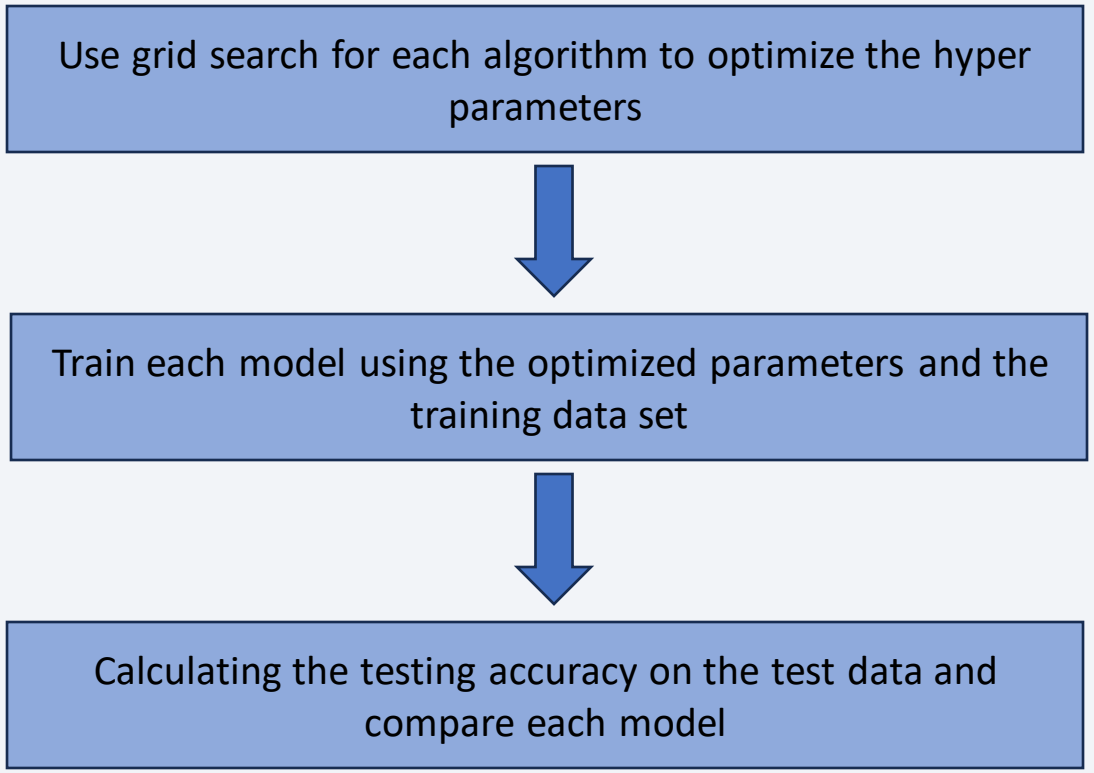
- The process began by marking the various launch sites on the US map using circles. This approach aimed to gain an understanding of the general geographical distribution of these sites.
- Next, each individual launch was marked on the map, and a color-coded scheme was employed to indicate their respective outcomes.
- Lastly, the distance between the launch sites and significant locations, such as coastlines and railways, was measured and marked on the map.
- Please refer to the link for the code

Build a Dashboard with Plotly Dash

- A dashboard was created using dash in python
- There are 2 interactive visuals present. The first one contains a pie chart to view the distribution of successful launches by launch site
- The second visual included a payload mass slider to view the success rates of landings based on the mass and booster version
- Please refer to the link for the code

Predictive Analysis (Classification)

- Initially, the data was divided into two subsets, namely the training set and the test set.
- Subsequently, a grid search was conducted to explore multiple algorithms, including logistic regression, support vector machines (SVM), decision trees, and k-nearest neighbors (KNN).
- Finally, the model was trained using the training data and the optimized parameters. The performance of the models was compared based on their testing accuracy
- Please refer to the link for the code



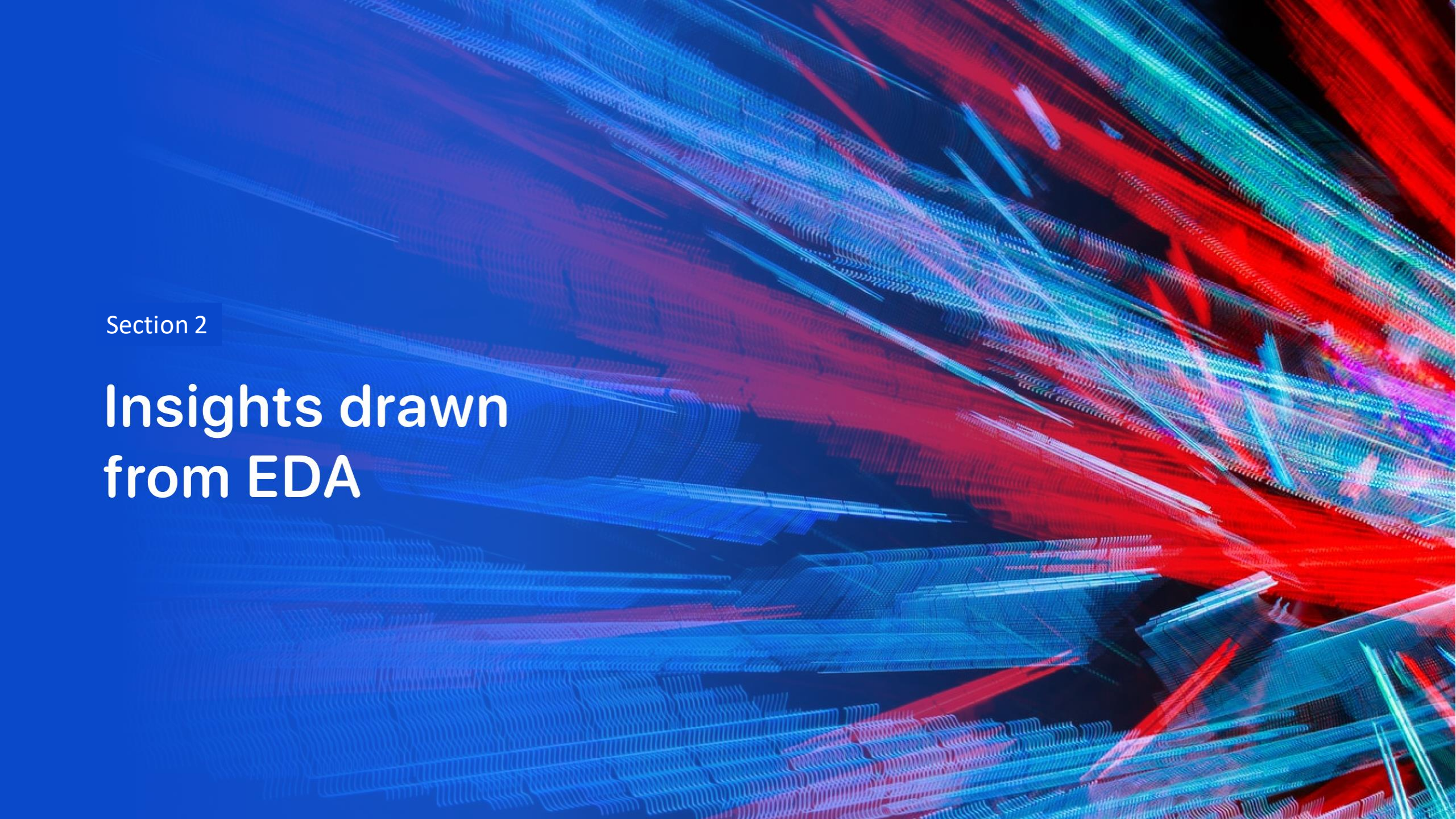
Results

The following results are presented in the next few sections

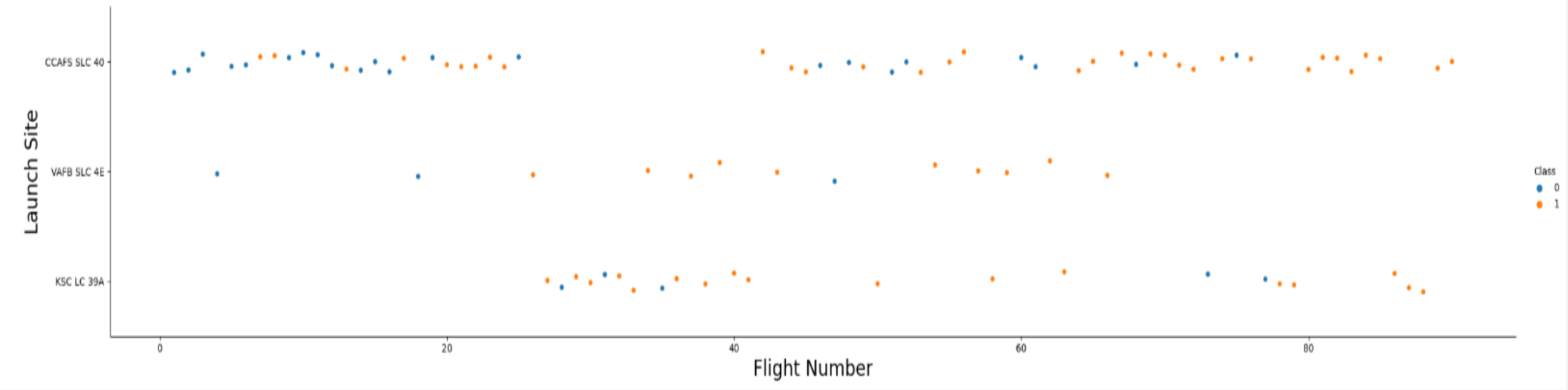
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

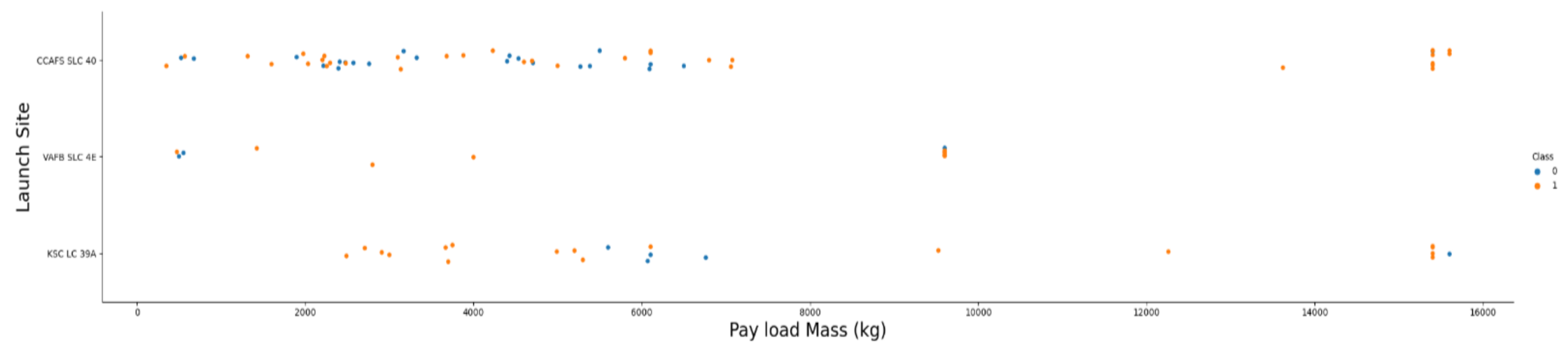


Flight Number vs. Launch Site



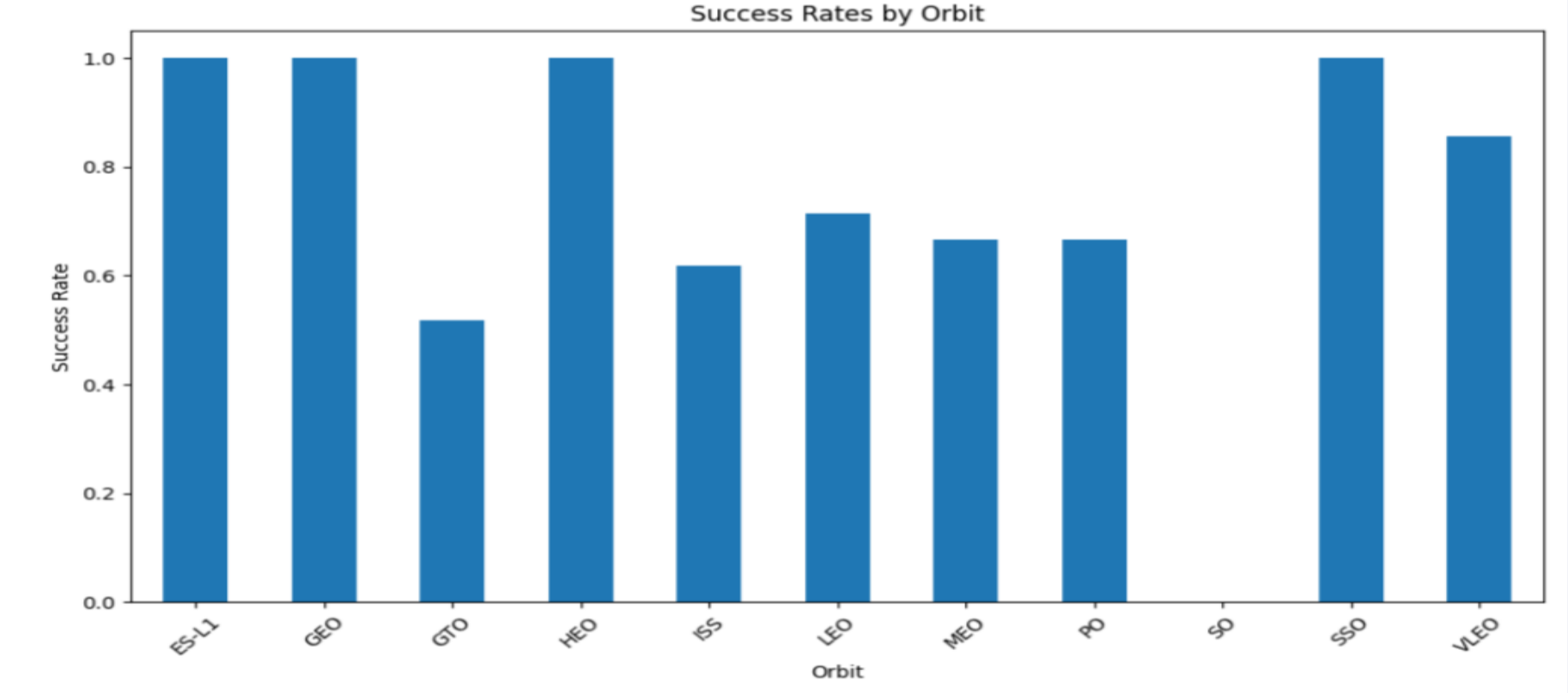
- As shown, the success rates seem to be getting higher as the flight number increases. This is an indication of that fact that success rates have improved over time
- The launch site KSC LC-39A has the highest success rate

Payload vs. Launch Site



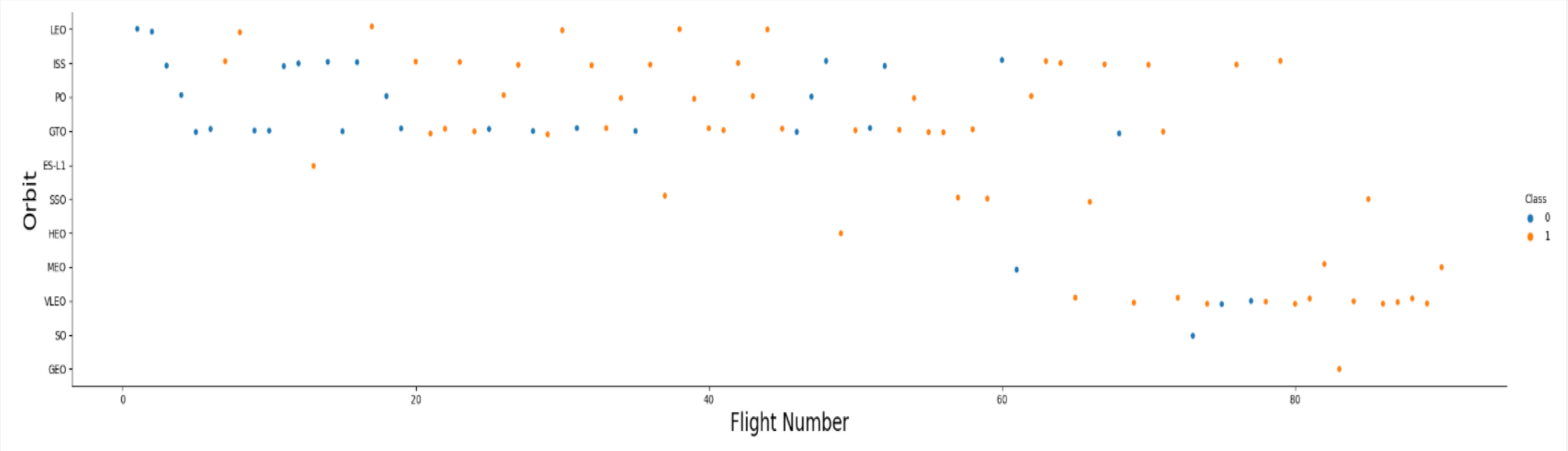
- The figure shows the scatter plot of the Payload mass vs the launch site
- The most noticeable cluster is when the mass is high and the launch site is CCAFS. We have a very high success rate in this case

Success Rate vs. Orbit Type



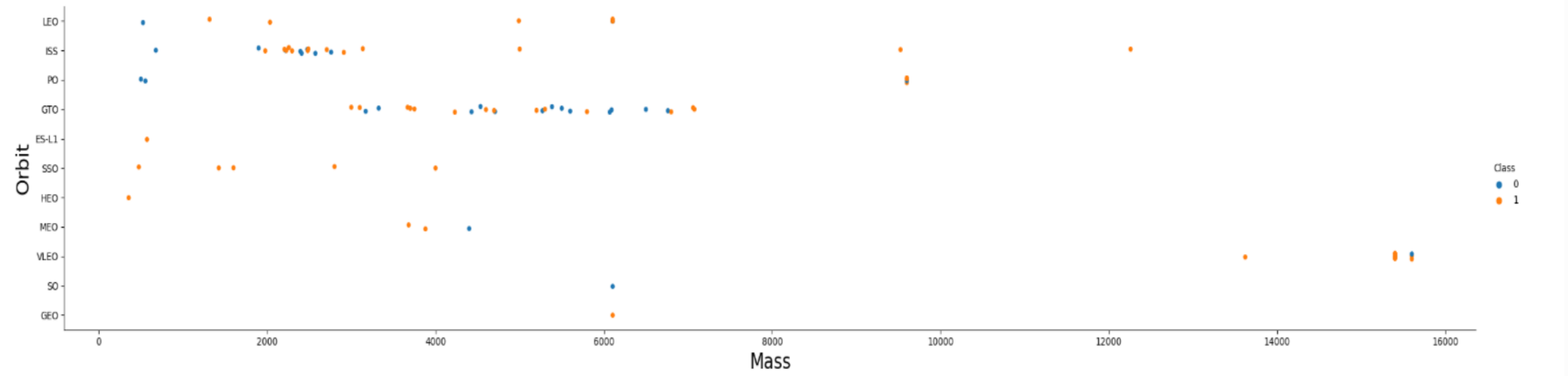
- The bar chart shows the success rates by orbit type for the launches
- While there are a few high success orbits such as GEO and ES-L1, there are plenty of orbits with a 50% or lower success rate

Flight Number vs. Orbit Type



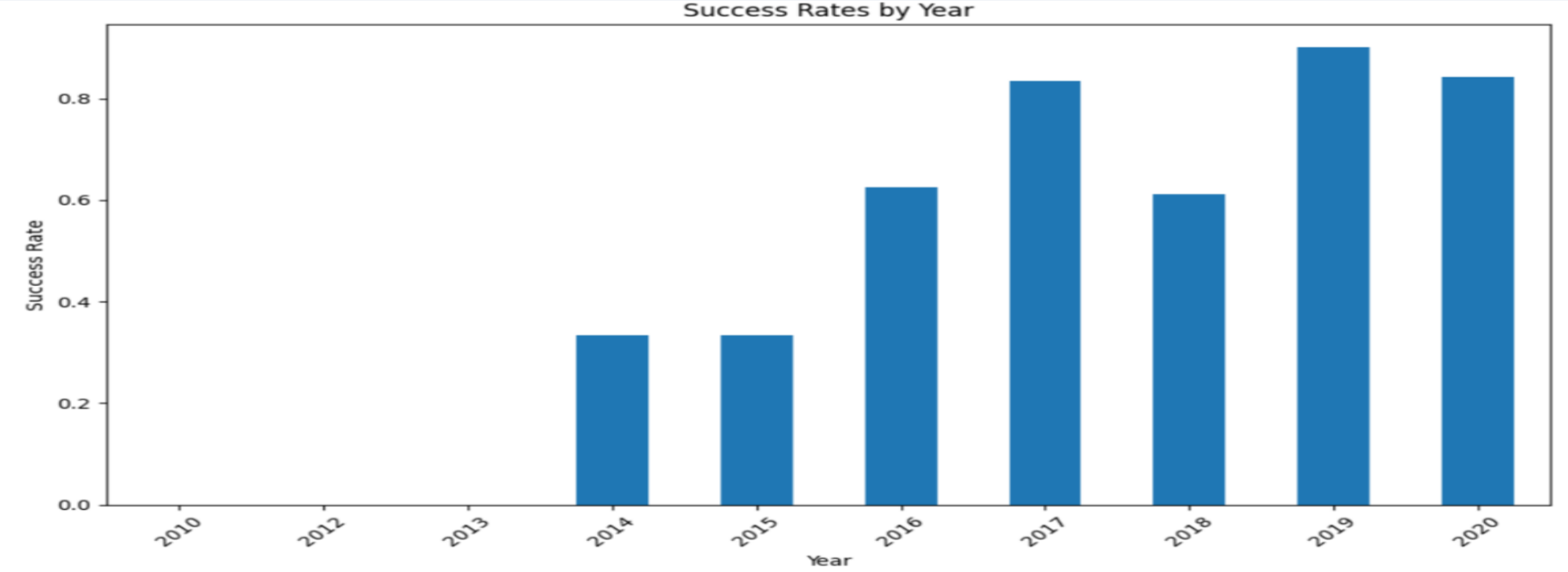
- The figure shows the scatter plot of the Flight number vs Orbit Type
- We can see a change in the trend of orbits used over time as the flight number changes. Initially, orbits such as LEO and ISS were more frequent, while recently, VLEO orbits are more prevalent

Payload vs. Orbit Type



- The figure shows the scatter plot of the Flight number vs Orbit Type
- We can observe a preference of orbit for specific mass ranges. For example, payloads with a higher mass generally use the VLEO low earth orbit

Launch Success Yearly Trend



- Since 2013, we can clearly see a sharp increase in the success rates for launches
- The highest success rate so far was achieved in 2019

All Launch Site Names

- The query result shows the 4 unique launch sites

[9]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- The query result shows us all the records of launches made from the launch site 'CCAFS LC-40'

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The query result shows us the total payload mass from all flights made for NASA

[20]:

Customer	Total_Payload_Mass
NASA (CRS)	45596.0

Average Payload Mass by F9 v1.1

- The query result shows the average mass of the payload across all flights made by the Falcon9 rocket with booster version F9 v1.1

[21]:

Booster_Version	Average_Payload_Mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- The query result shows that the earliest successful landing on a ground pad was made at the end of year 2015

[35]:

Landing_Outcome	Earliest_Date
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query result shows the Booster Versions which have succesfully managed to land on a drone ship and had a Payload mass between 4000 and 6000 kgs.
- There are 4 such booster versions

[37]:

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696.0	Success (drone ship)
F9 FT B1026	4600.0	Success (drone ship)
F9 FT B1021.2	5300.0	Success (drone ship)
F9 FT B1031.2	5200.0	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The query result shows the tally of mission outcomes. We can see that the vast majority of missions were successful. However, in a lot of these missions, the 1st stage of the rocket failed to land preventing its re-use

[41]:

Mission_Outcome	Total
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The query result shows the list of Booster Versions that have managed to carry the heaviest payloads

[16]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- The query result shows the failed landing outcomes on the drone ship in 2015
- We can see 2 results returned, both of which were launched from the CCAFS LC-40 launch site

[17]:

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

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

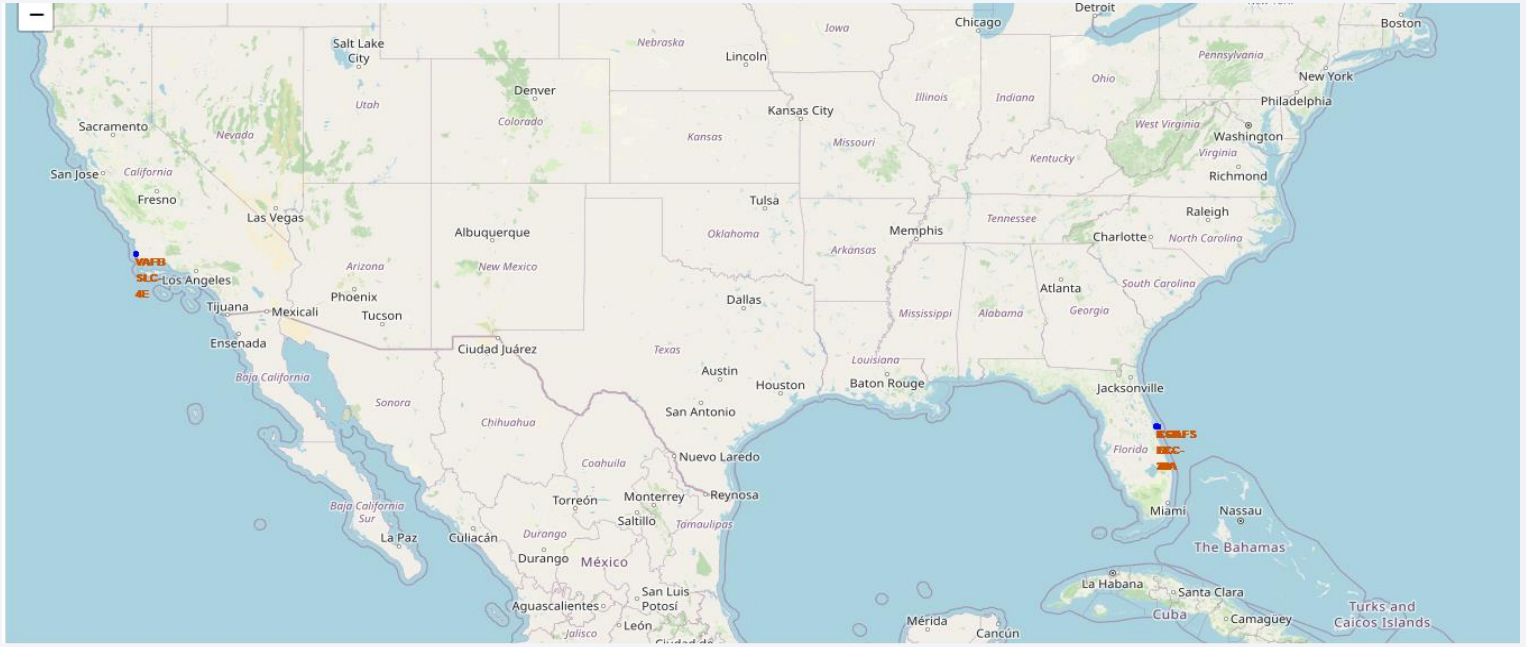
- The query shows the count of the different 1st stage landing outcomes between the date range
- Most of the landings were successful, however, a lot of times there was no attempt made to retrieve the 1st stage of the rocket

[26]:

Landing_Outcome	frequency
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

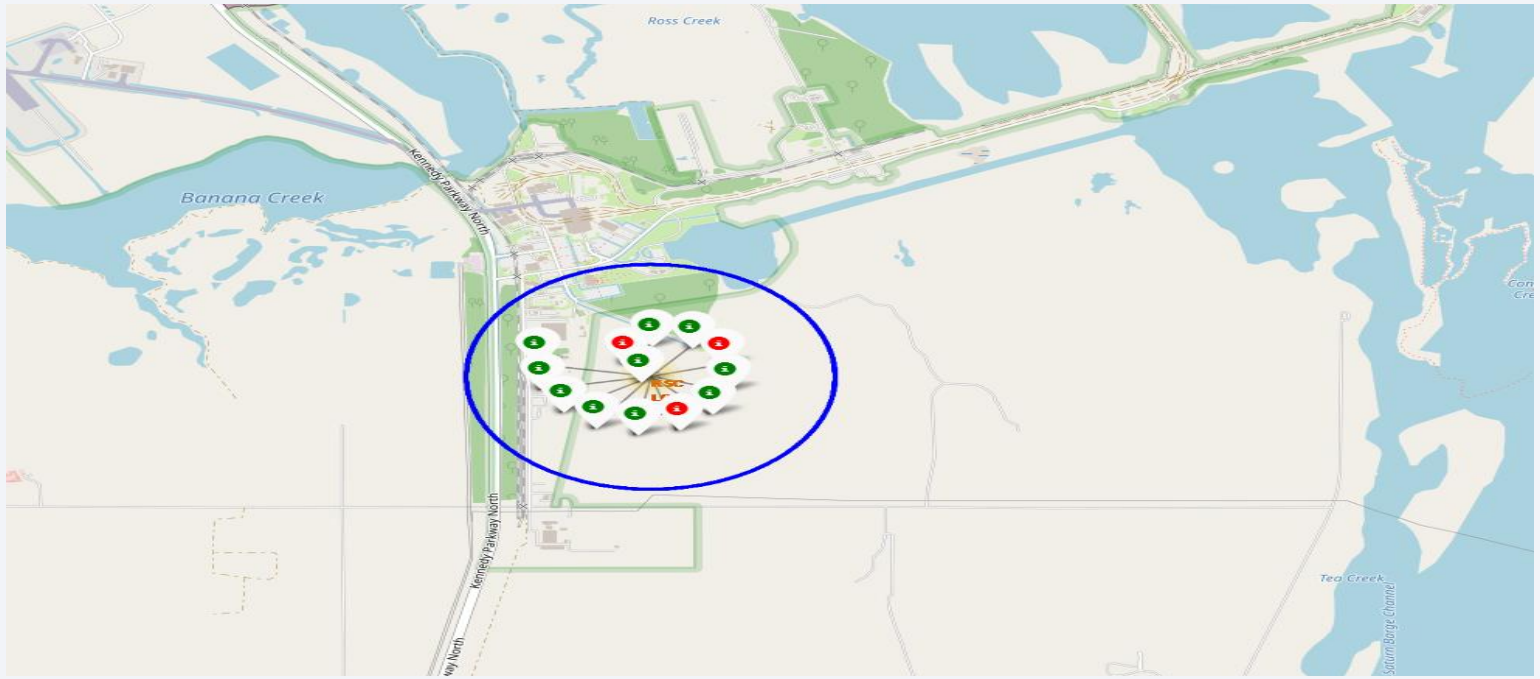

Section 3
Launch Sites
Proximities Analysis

Location of the launch sites



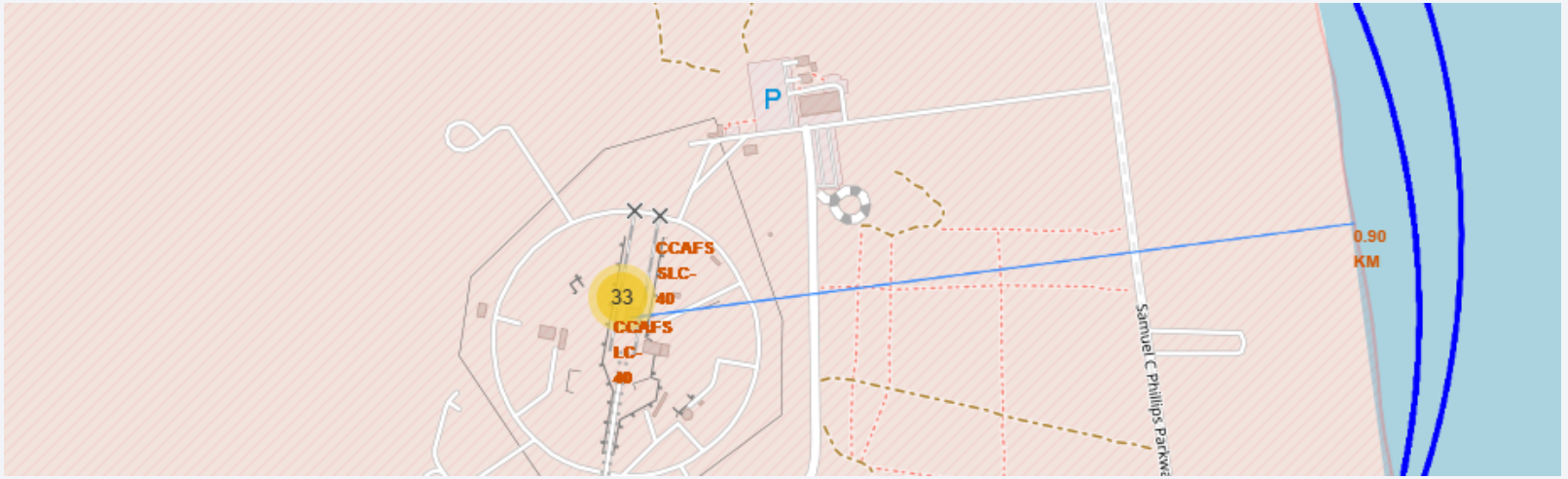
- As shown, the launch sites are located on either the west or east coast of the country with 'VAFB SLC-4E' being the only one in the west
- The proximity to the ocean allows for ocean landings and drone ship landings

Launch outcomes in 'KSC LC-39A'

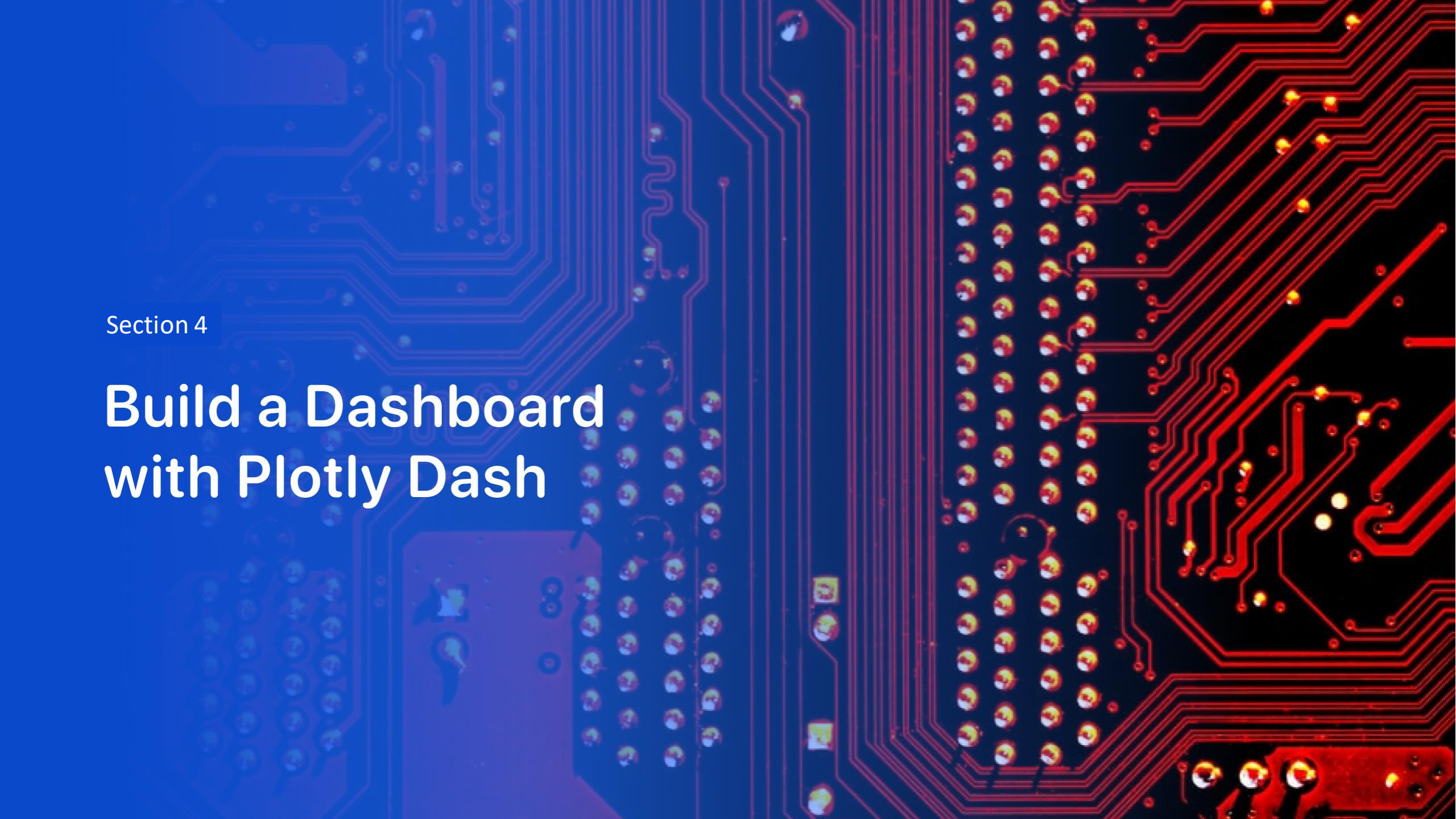


- The picture shows the outcomes of launches in the 'KSC LC-39A'
- Clearly this launch site has a very high success rate compared to others

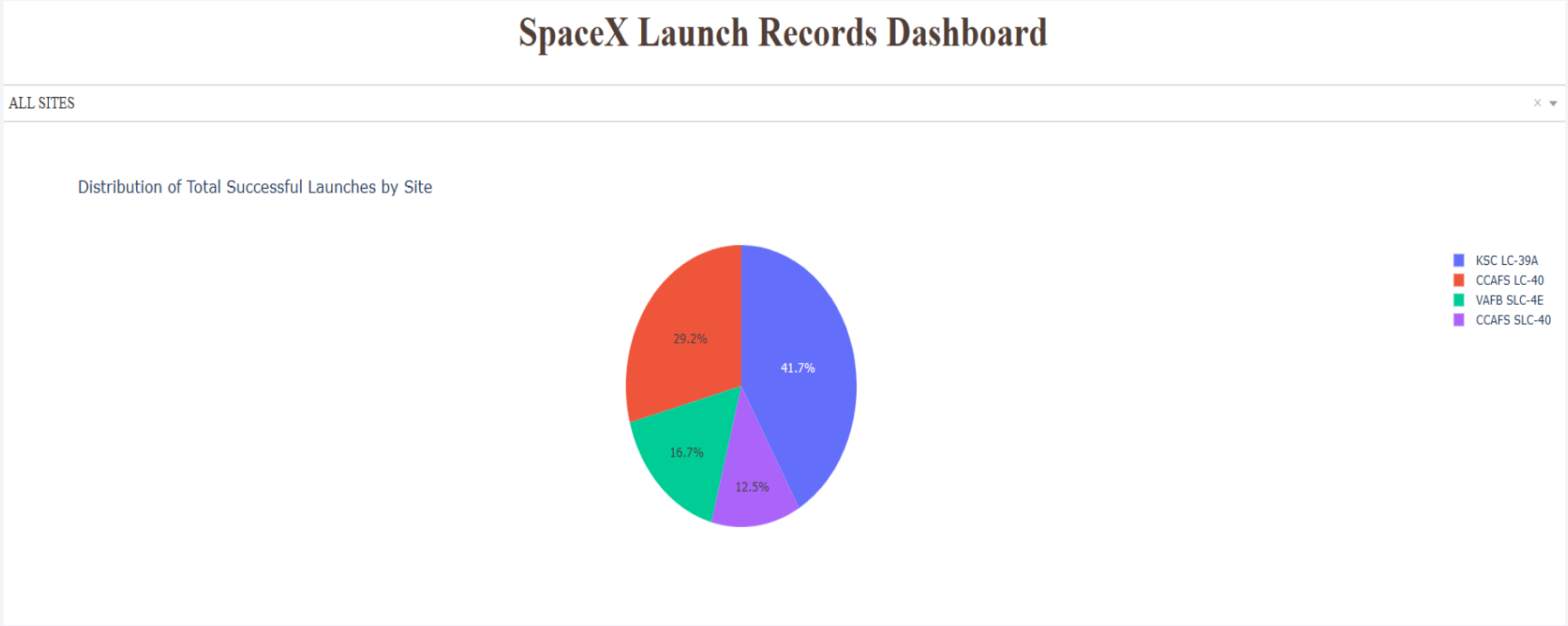
Proximity to key locations



- As we can see, the launch sites are located close to coastal regions to facilitate landing the 1st stage on drone ships
- Similarly, these launch sites are also located close to railways and away from major cities. The railways likely help in transporting heavy equipment and distance from cities ensures safety in case of launch failures

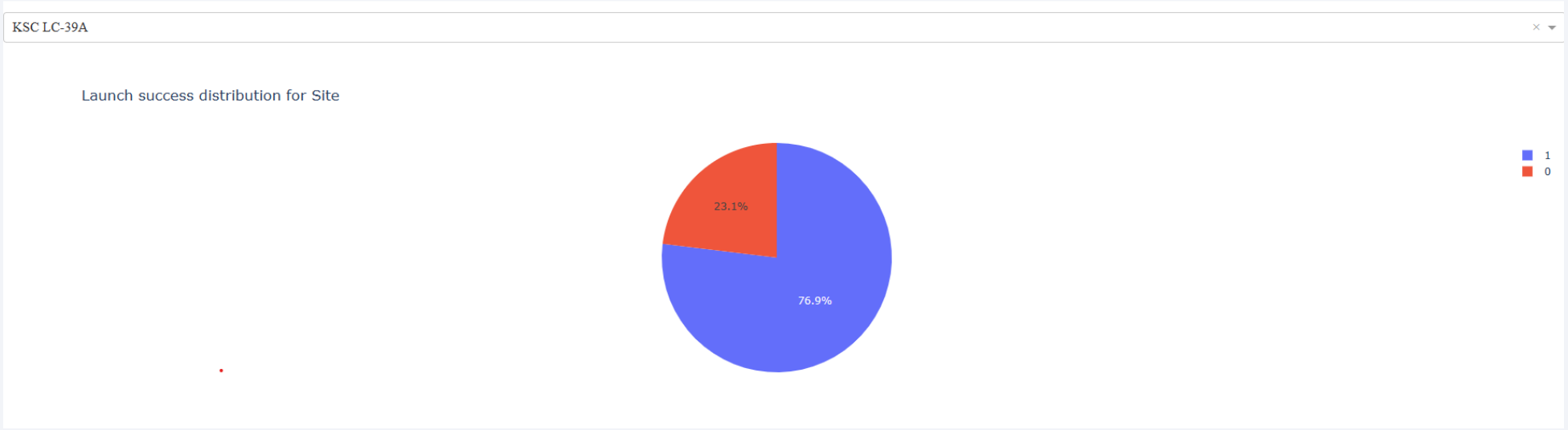

Section 4
Build a Dashboard with Plotly Dash

Distribution of successful launches by site



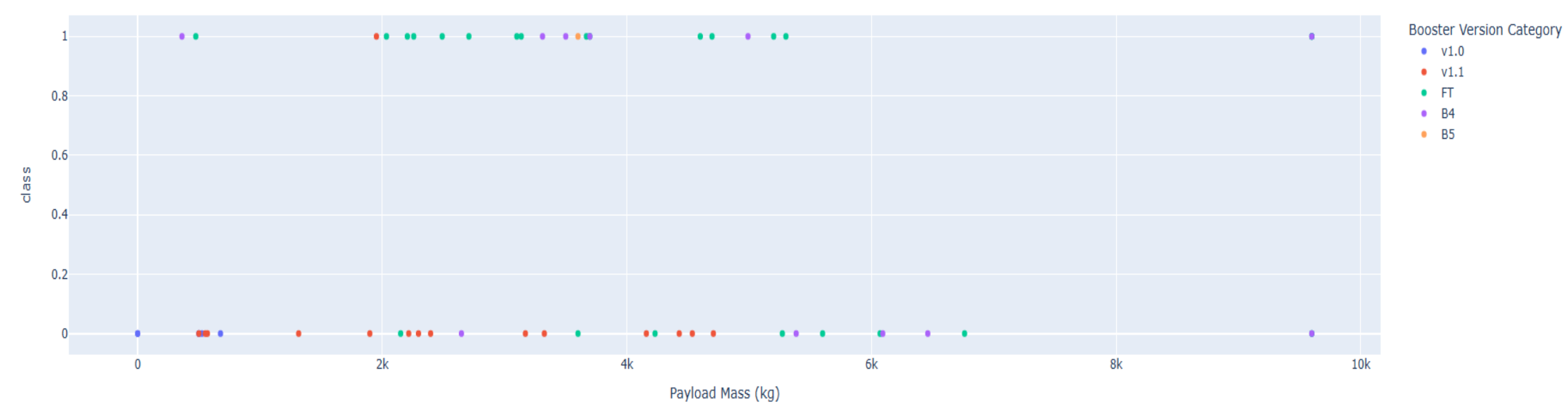
- The picture shows the distribution of successful launches by launch site
- As seen, the launch site 'KSC LC-39A' has historically yielded the most number of successful launches

Launch success distribution for KSC LC-39A



- The picture shows the success rate for site 'KSC LC-39A' which is the highest among all sites at nearly 77%
- Together with the most number of successful launches and highest success rate, this site is a key determiner in the success of future launches

Scatter plot for Payload mass and launch outcome



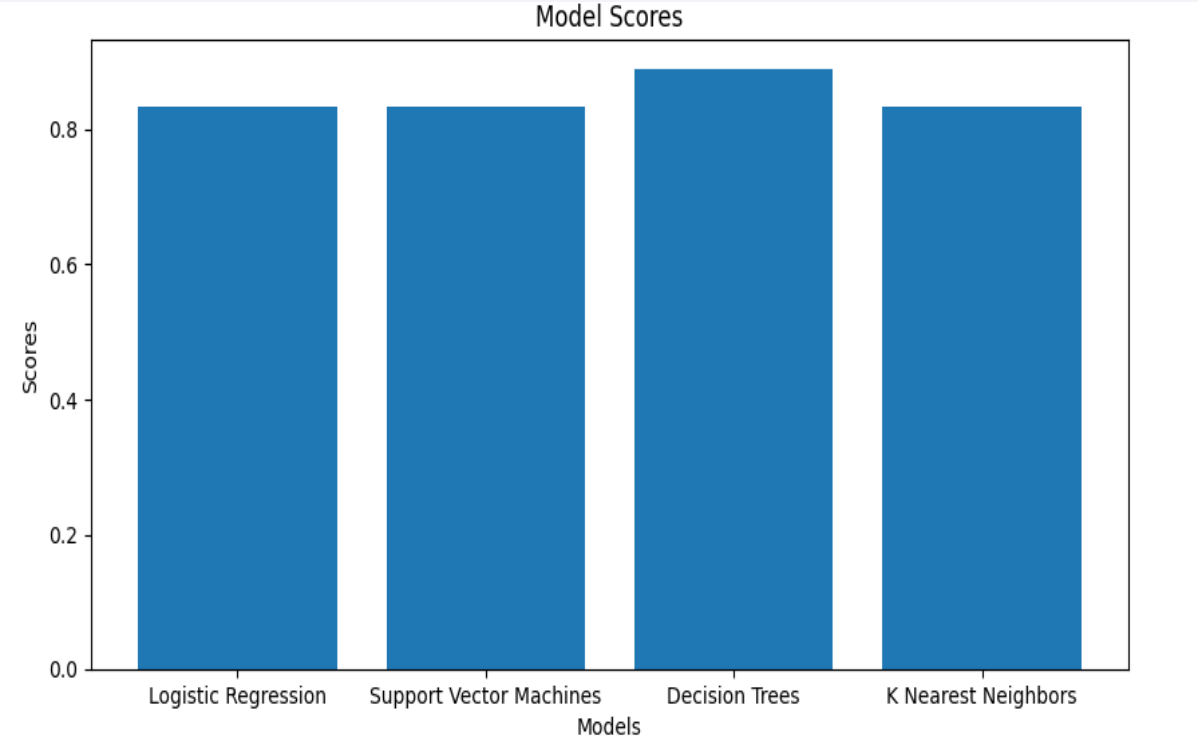
- As shown above, the booster version 'FT' has a very high success rate across multiple payload mass ranges
- In particular, the payload mass range between 2K and 5.5K sees the most number of successful launches where the 'FT' booster is the highest contributor

Section 5

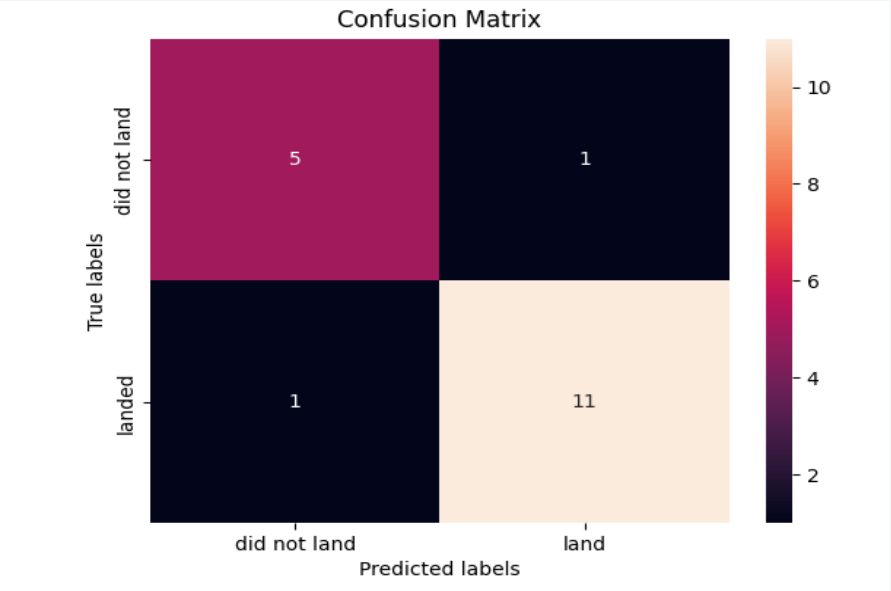
Predictive Analysis (Classification)

Classification Accuracy

- The figure shows the testing scores for the four models trained
- The Decision tree has the highest testing accuracy of almost 0.89
- However, this is achieved at a maximum depth of 18 which might make this model too complex



Confusion Matrix – Decision Tree



- Looking at the confusion matrix, it is clear to see that this model minimizes on both false positives and false negatives

Conclusions

The following are the key insights gained from this project:

- Success rates have generally increased over time
- Parameters such as flight number, booster version and launch site have the highest influence in determining success rates
- The launch site locations are strategically placed to ensure optimal logistical support as well as minimize danger to civilians
- The top machine learning classification models seem to be able to effectively predict the success rates of new flights with up to almost 90% accuracy

Thank you!