IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Amanullah Mohammed
07-Nov-2021

# Outline

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

# Executive Summary

- Summary of methodologies:
  - Data collected using the public Space X API.
  - Wrangle the data
  - Visualize the relationship between different attributes by using charts
  - Build an interactive map with Folium
  - Build a Plotly Dashboard
  - Built, evaluated, improved, and find the best performing classification model
- Summary of all results:
  - Found positive impact with increasing year, flight number and altitude of the orbit, and negative impact on the payload mass.
  - Launch sites are more closer to coasts, railway lines and highway and farther from the cities.
  - Launch Site ‘KSC LC-39A’ has the highest number of success launches while ‘CCAFS SLC-40’ has the highest success ratio.
  - Best performing classification model is Decision Tree Classification with accuracy of 88.57%

# Introduction

- We need is to determine the price of each launch by gathering information about the Space X launches and creating visual and reusable dashboards.
- We will determine if the Space X will reuse the first stage for which we will train a machine learning model and use public information to predict if Space X will reuse the first stage

Section 1

# Methodology

# Methodology

## Executive Summary

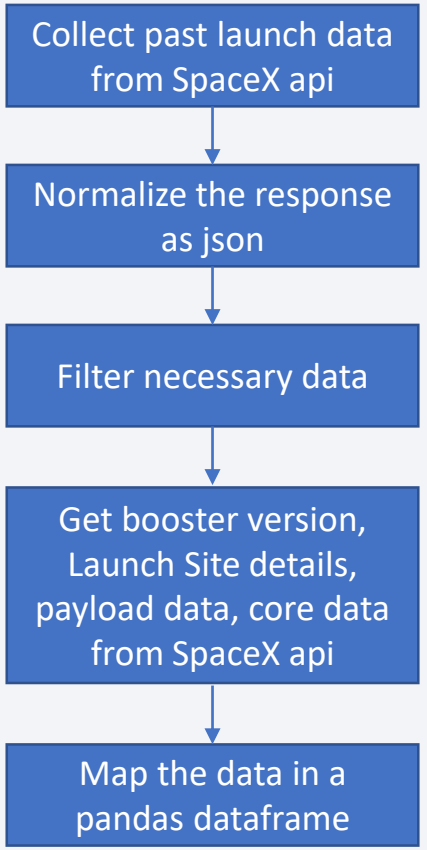
- Data collection methodology:
  - Data is gathered from the Space X REST API.
- Perform data wrangling
  - Wrangling Data using an API, Sampling Data, and Dealing with Nulls.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We have built a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.

# Data Collection

- Data sets were collected using the below two methods:
  - Using the SpaceX public REST API
  - Using Web Scraping from Wiki Pages

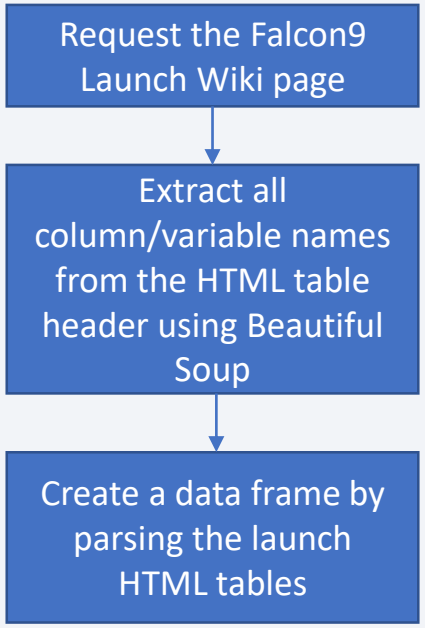
# Data Collection - SpaceX API

- The data is collected using the SpaceX launch data from SpaceX REST API. This API gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- The flowchart describes the data collection process.
- Notebook Link



# Data Collection - Scraping

- We used Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- Then we parsed the data from those tables and convert them into a Pandas data frame for further visualization and analysis.
- Notebook Link



# Data Wrangling

- We want to transform this raw data into a clean dataset which provides meaningful data.
- The data wrangling process includes:
  - Wrangling Data using an API – use only the required data for our analysis.
  - Dealing with Nulls - calculate the mean of the PayloadMass data and then replace the null values in PayloadMass with the mean
  - Use one hot encoding of categorical values.
- Notebook Link

# EDA with Data Visualization

- Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.
- Exploratory Data Analysis
  - We visualize the relationship between different attributes by using a scatter plot, bar graph or line chart and explain any patterns that we found.
- Features Engineering
  - Create dummy variables to categorical columns
- Notebook Link

# EDA with SQL

- To summarize, performed the below queries from the database:
    - Display the names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA‘
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1
    - List the date when the first successful landing outcome in ground pad was acheived.
    - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - List the total number of successful and failure mission outcomes
    - List the names of the booster_versions which have carried the maximum payload mass.
    - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - Rank 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
- Notebook Link

# Build an Interactive Map with Folium

- We build an interactive map with Folium in order 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
- Notebook Link

# Build a Dashboard with Plotly Dash

- We build a Plotly Dashboard to plot the below:
  - Add a dropdown list to enable Launch Site selection
  - Add a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site.
  - Add a slider to select payload range
  - Add a scatter chart to show the correlation between payload and launch success
- File Link

# Predictive Analysis (Classification)

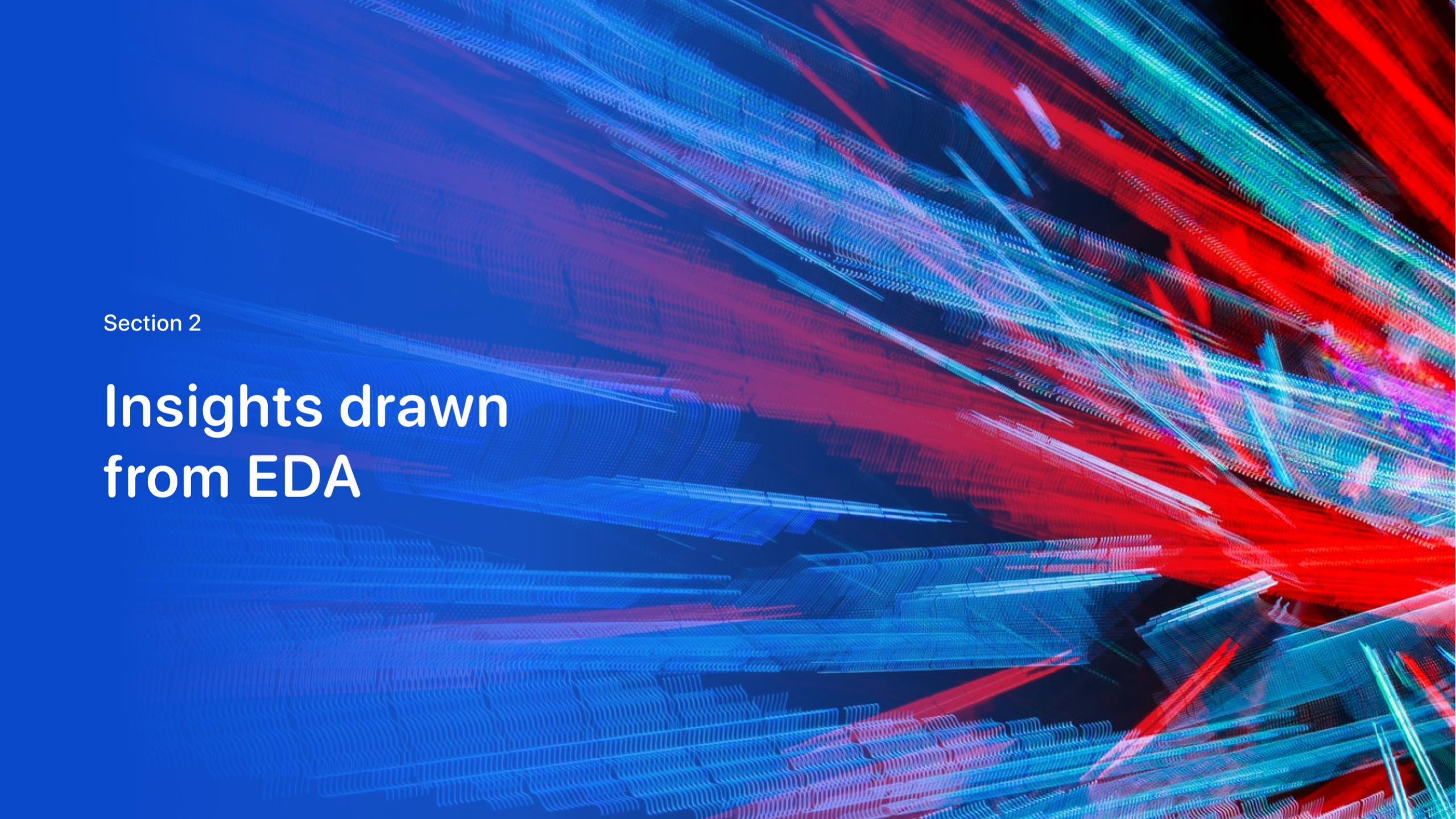
- We perform exploratory Data Analysis and determine Training Labels by:
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Then we find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Notebook Link

# Results

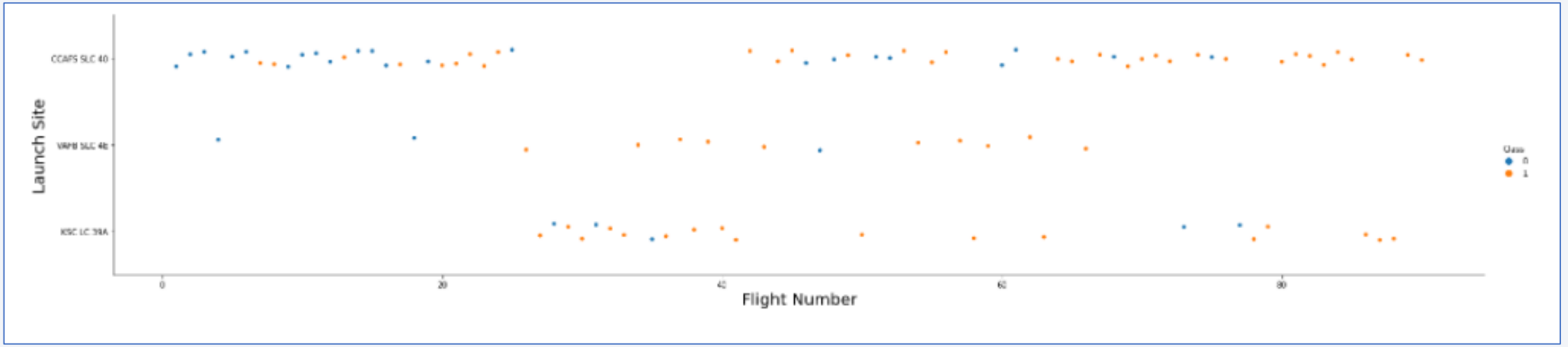
- Exploratory data analysis results:
  - The first stage is more likely to land successfully if case of the below scenarios:
    - The flight number is greater
    - The payload mass is between 2000 and 4000 kgs
    - The altitude of the orbit is higher
  - Further we also found that the chances of success keeps increasing as the year goes by.
- Interactive analytics demo in screenshots:
  - Launch Site ‘KSC LC-39A’ has the highest number of success launches while the launch site ‘CCAFS SLC-40’ has the least, but also ‘CCAFS SLC-40’ has the highest success ratio.
- Predictive analysis results
  - Our decision tree model predicts that 88.57% the first stage is likely to land successfully.

Section 2

# Insights drawn from EDA

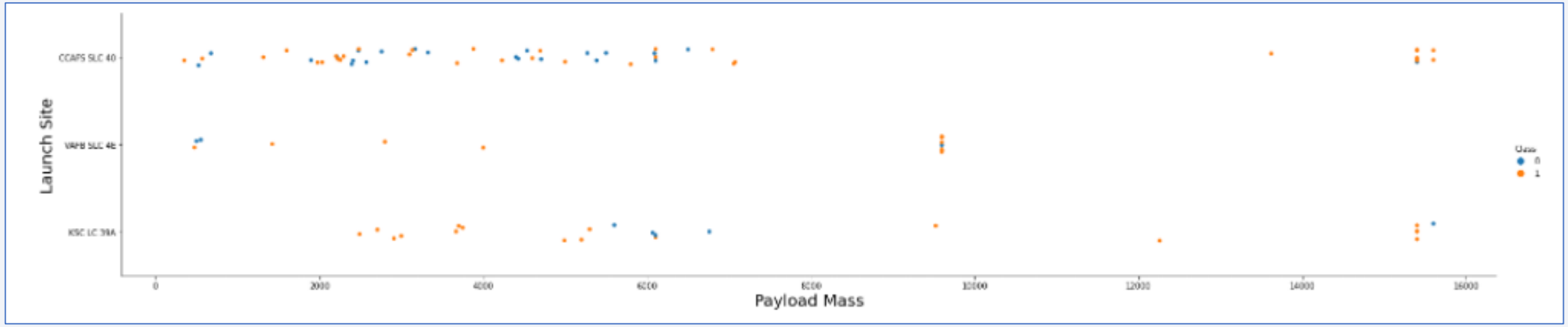


# Flight Number vs. Launch Site




- For the all launch sites, as the flight number increases, the first stage is more likely to land successfully.

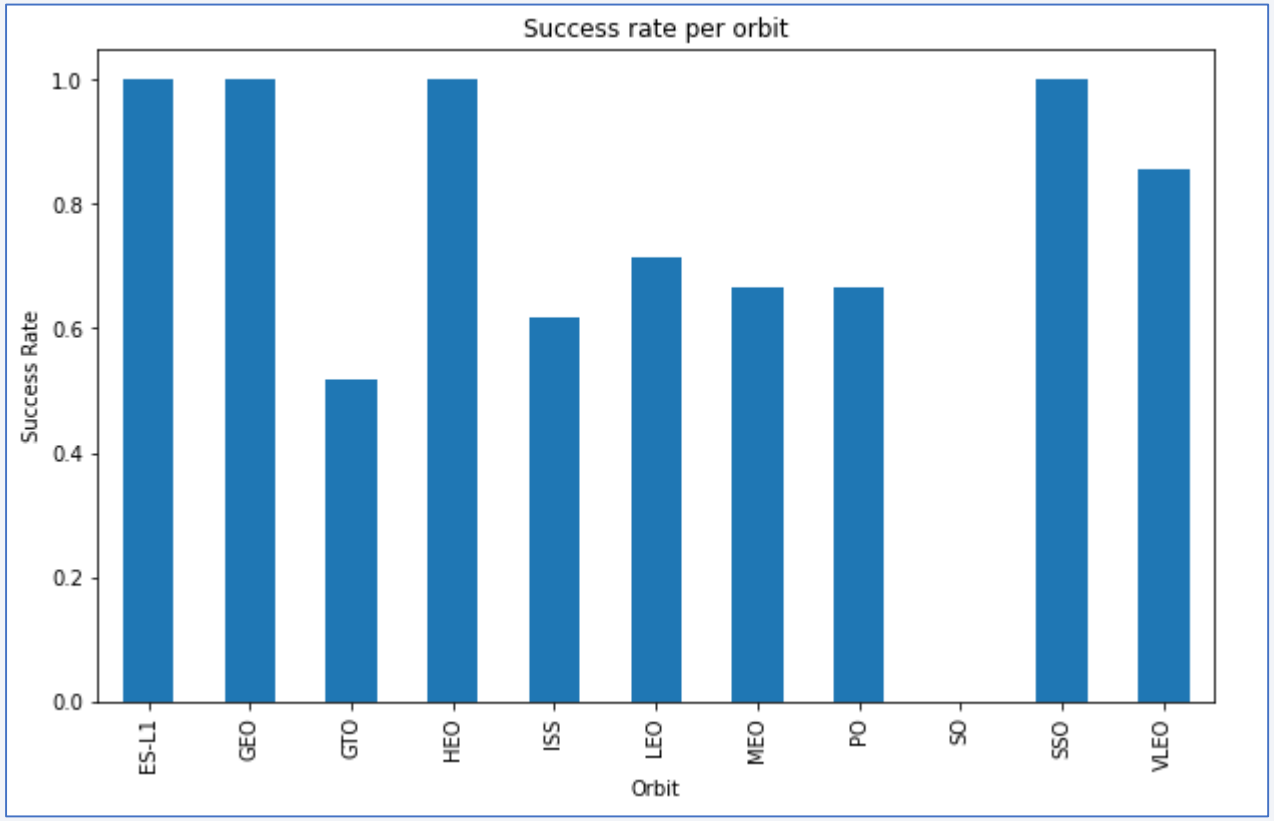
# Payload vs. Launch Site



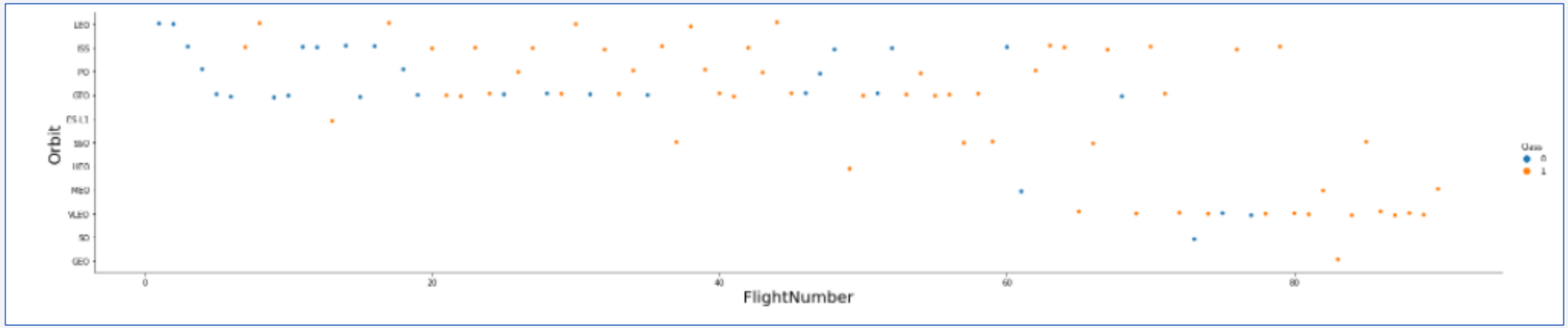

- The less the payload mass for each site, the better are the chances of successful outcome.

# Success Rate vs. Orbit Type

- The trend of Orbit vs Success Rate depicts that the success rate of the outcome is higher when the altitude of the orbit increases.

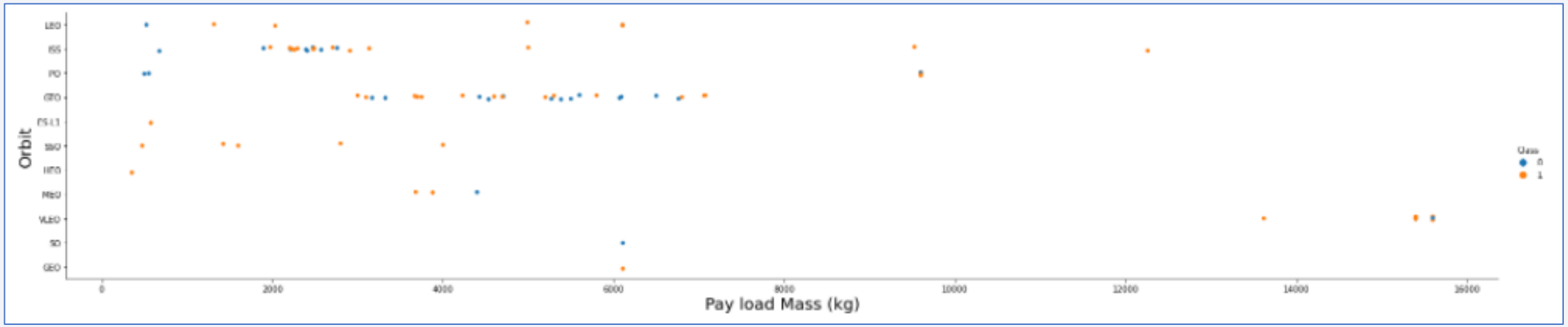


# Flight Number vs. Orbit Type




- The scatter point of Flight number vs. Orbit type depicts the below trends:
  - For LEO, ISS, PO, SSO and VLEO orbits, the trend is that as the flight number increases, the first stage is more likely to land successfully.
  - For the remaining orbits, there does not seems to be any relationship between the flight number and the orbit type.

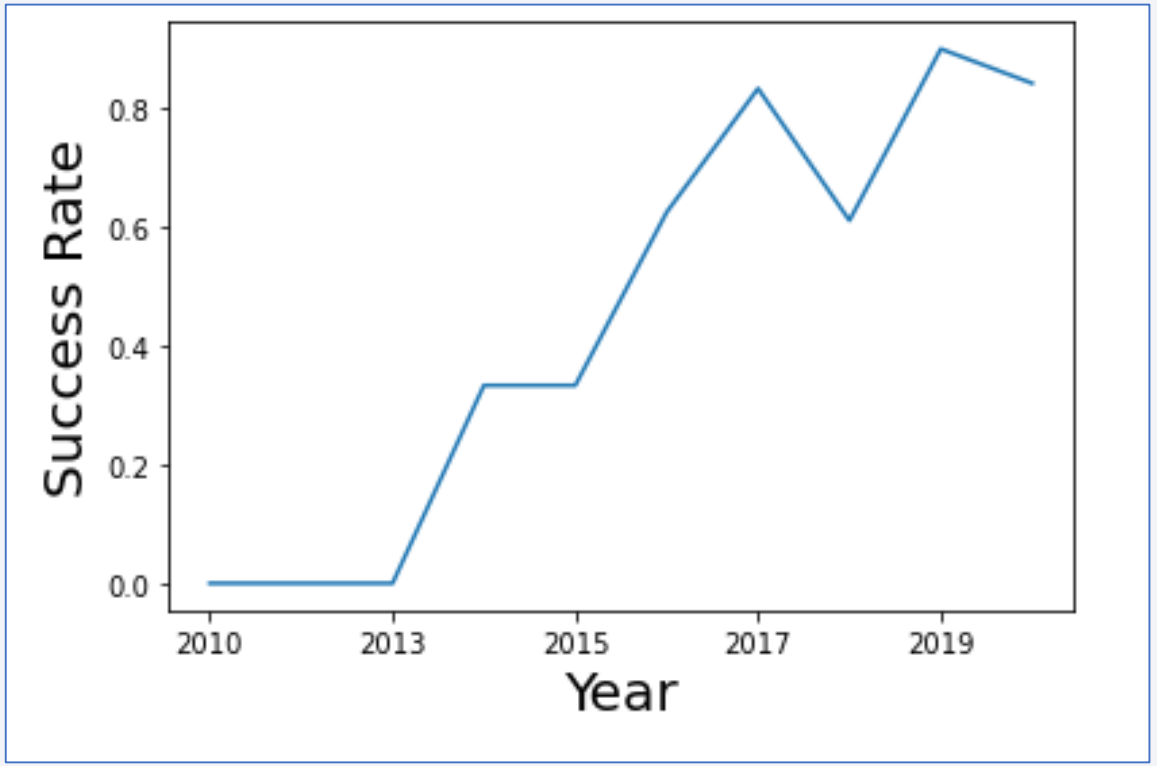
# Payload vs. Orbit Type




- Heavy payloads have a negative influence on GTO orbits and positive on SSO, LEO and ISS orbits

# Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



# All Launch Site Names

- The names of the unique launch sites:

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

- The SQL query used was to display the distinct LAUNCH_SITE from SPACEXTBL

# Launch Site Names Begin with 'CCA'

- To find 5 records where launch sites begin with `CCA`, we select the LAUNCH_SITE column from SPACEXTBL where LAUNCH_SITE is like ‘CCA%’ and we limit the outcome to 5.

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

# Total Payload Mass

- To calculate the total payload carried by boosters from NASA, we find the sum of all PAYLOAD_MASS__KG from SPACEXTBL where the CUSTOMER is ‘NASA (CRS)’, which comes out to be 45596 Kgs.

| total_payload_mass |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1, we use AVG() function of SQL to calculate the average of payload mass for F9 v1.1 Booster, which comes out to be 2928 Kgs.

| avg_payload_mass |
| --- |
| 2928 |

# First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad, we use the MIN() function of SQL.
- The dates of the first successful landing outcome on ground pad is:
  - 2015-12-22

| first_succ_outcome |
| --- |
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, we use the WHERE condition of SQL to filter the outcome and the payload mass.

| 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 total number of successful and failure mission outcomes are 100 and 1 respectively.

| total | failure | success |
|---|---|---|
| 101 | 1 | 100 |

# Boosters Carried Maximum Payload

- The list of the names of the booster which have carried the maximum payload mass are:
- It is noted that F9 B5 versions can carry maximum payloads

| 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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

| landing__outcome | booster_version | launch_site |
|---|---|---|
| Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

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

- Rank 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:

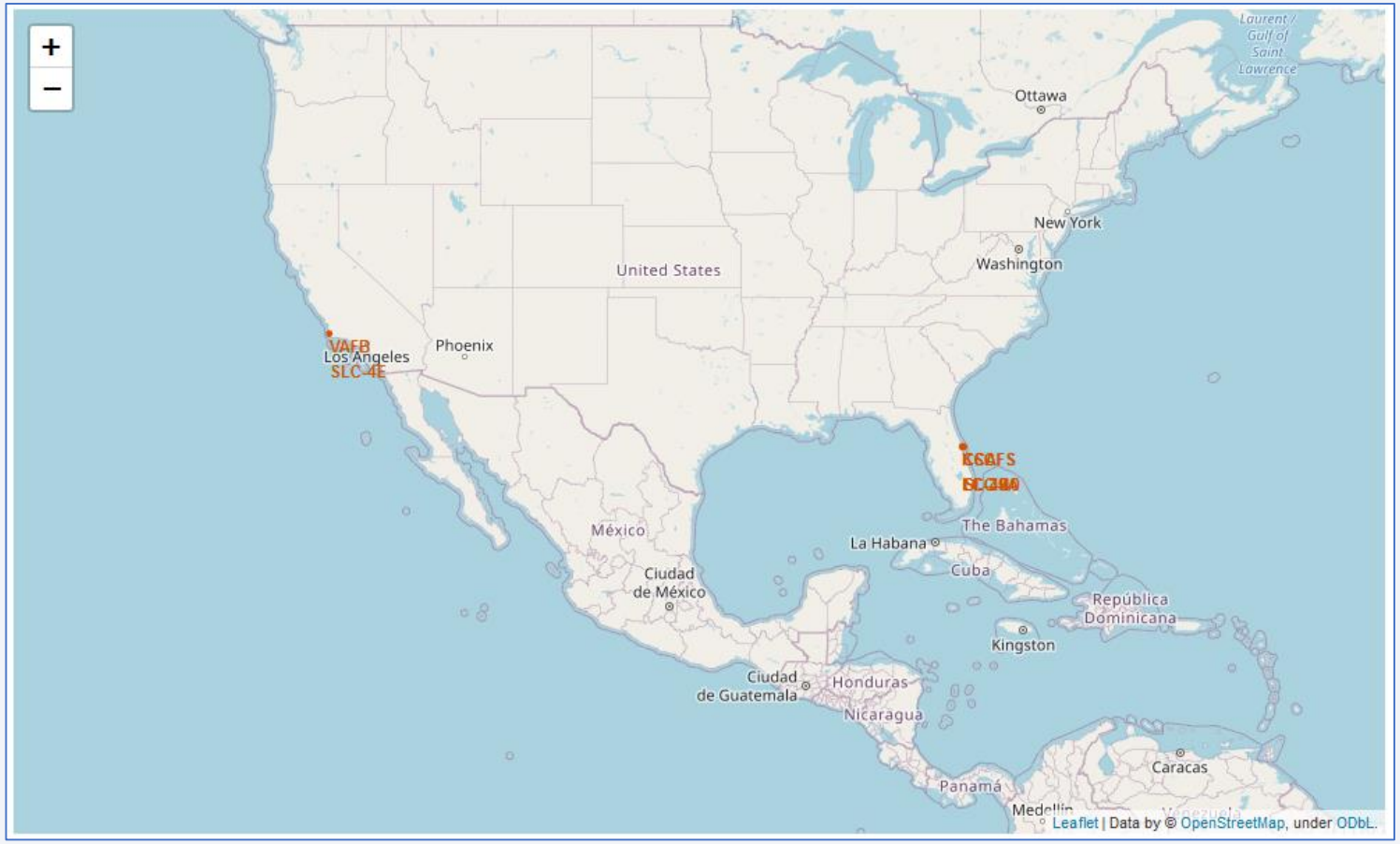
| landing__outcome | 2 |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Section 4

# Launch Sites Proximities Analysis

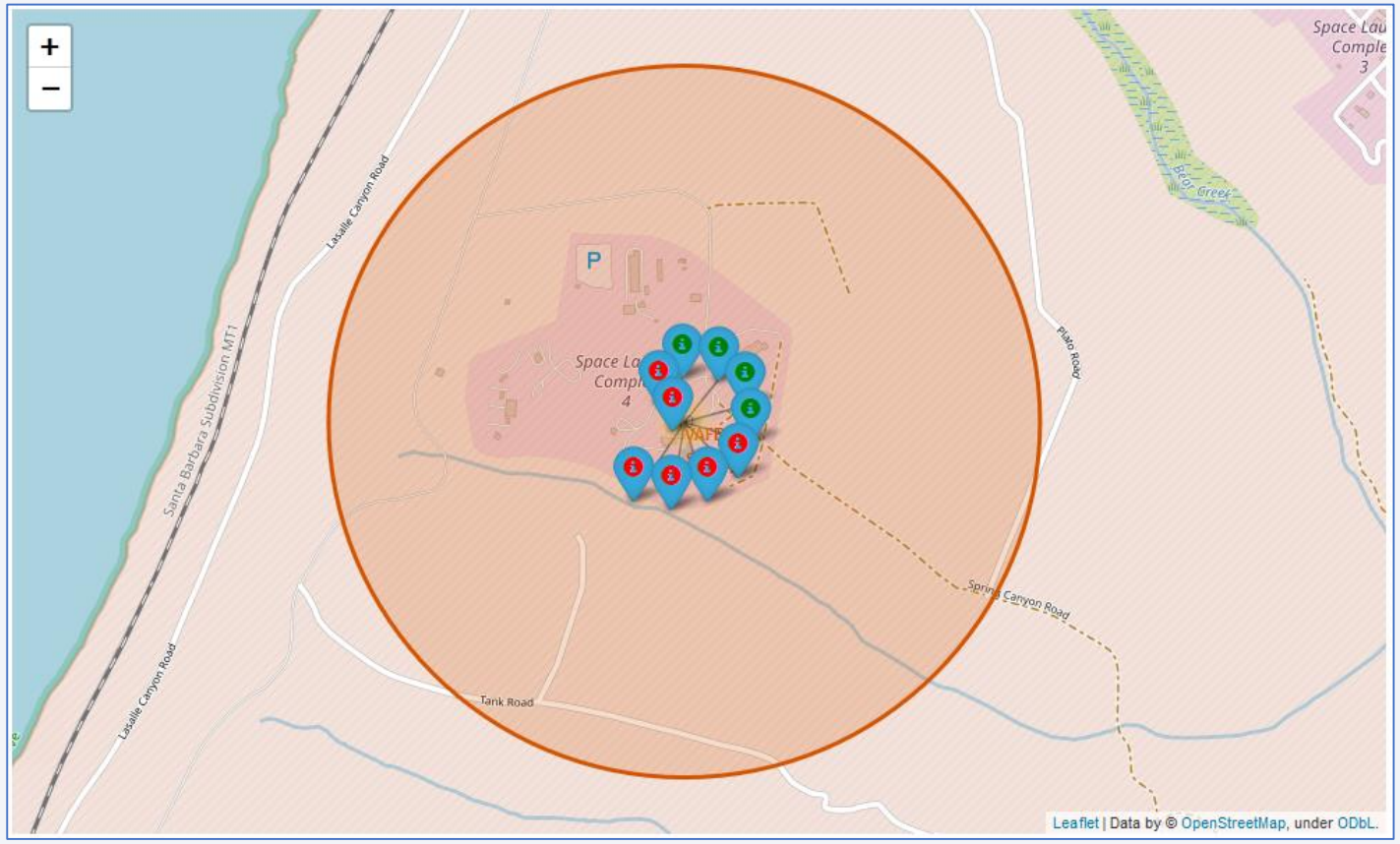
# Location of Launch Sites

- The map shows the location of all the launch site. It is noted that the launch sites are located on the east and west coasts of U.S.



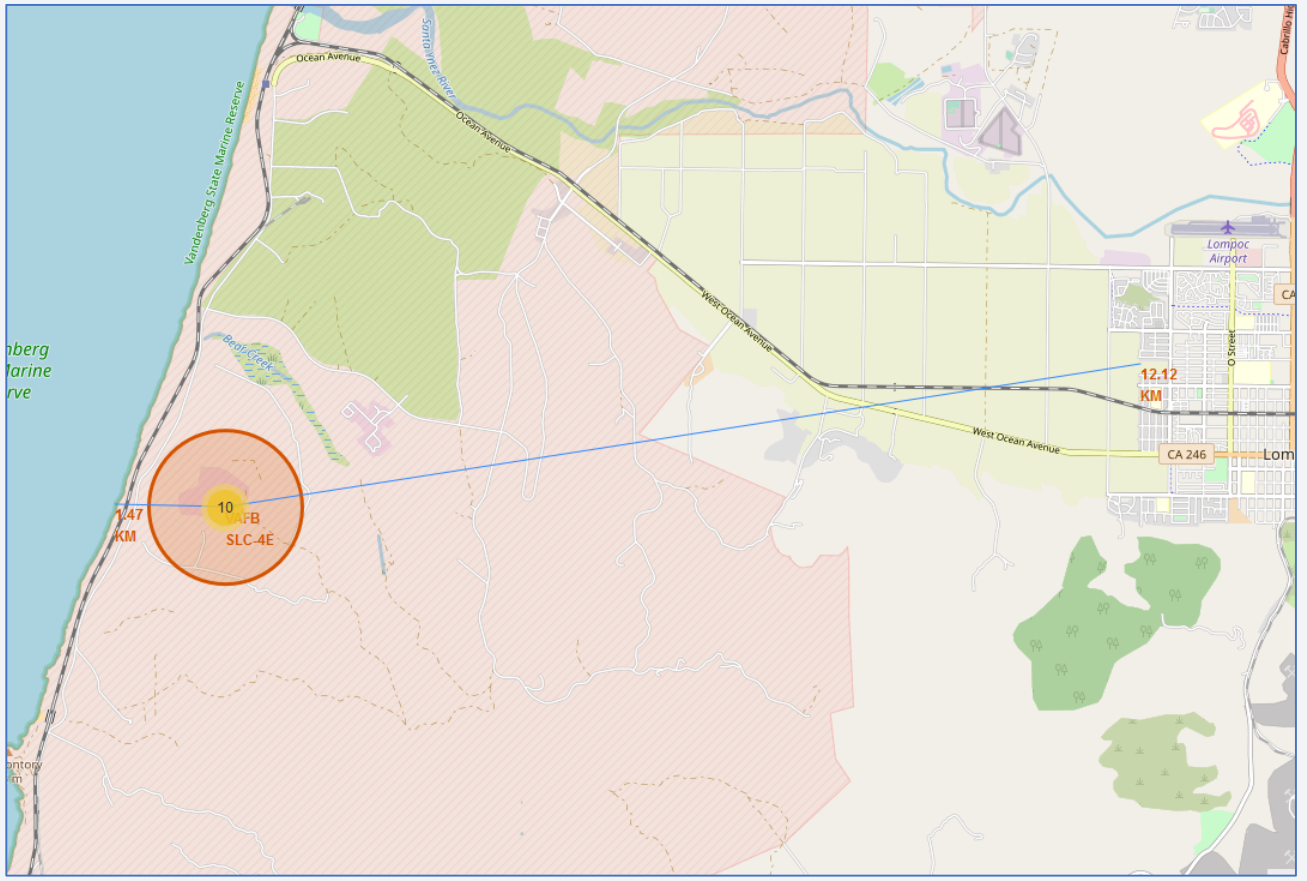
# Launch Outcomes of VAFB SLC-4E Site

- The launch outcome of the Launch Site VAFB SLC-4E is shown on the adjacent map.
- The colour coding of the markers are:
  - Red – failed outcome
  - Green -successful outcome



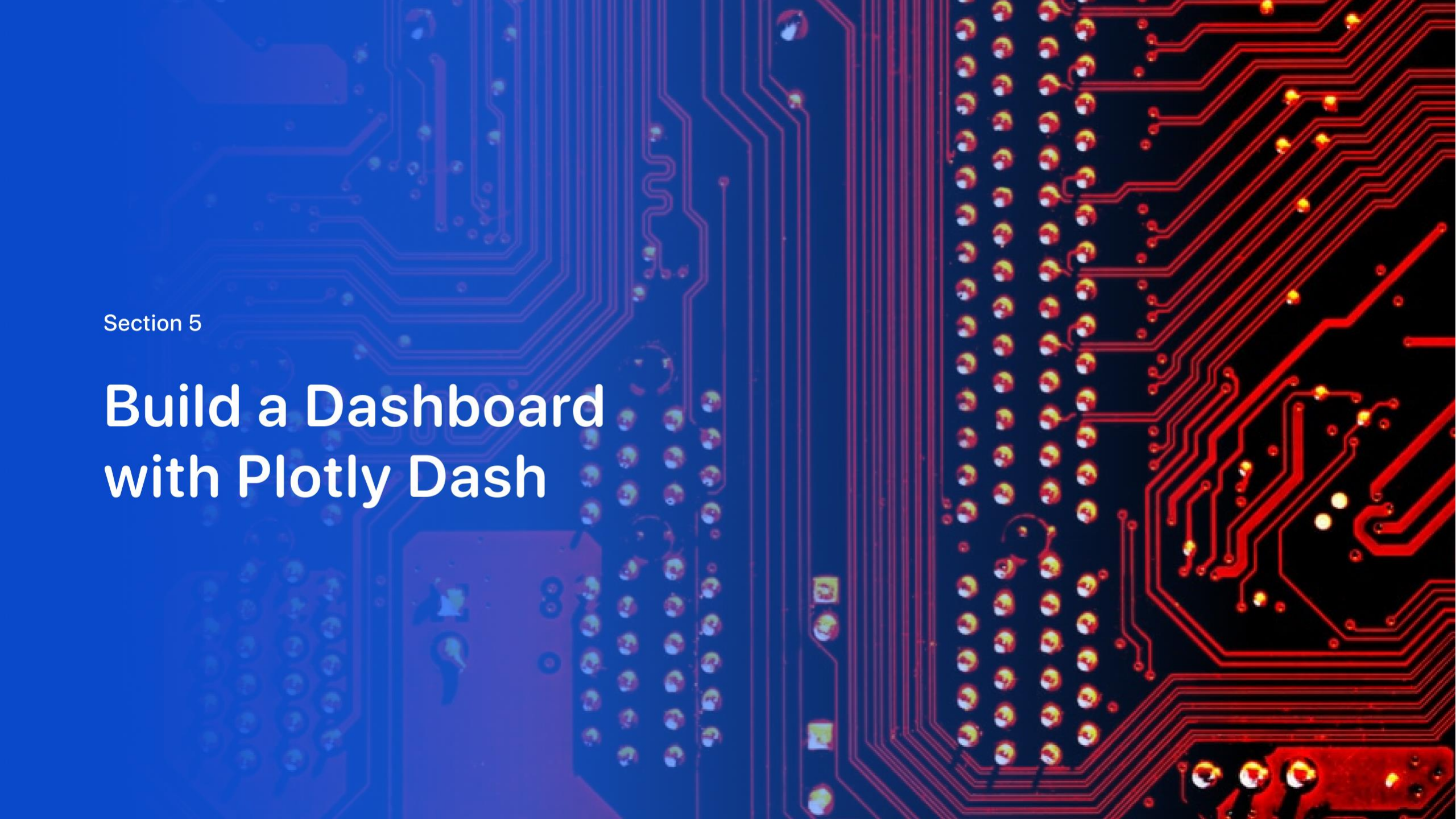
# Launch Site Proximities

- The adjacent map shows the distance of launch site from the coastline, railway lines, highway and cities.
- As noted from the map, the launch sites are more closer from coasts, railway lines and highway and farther from the cities.



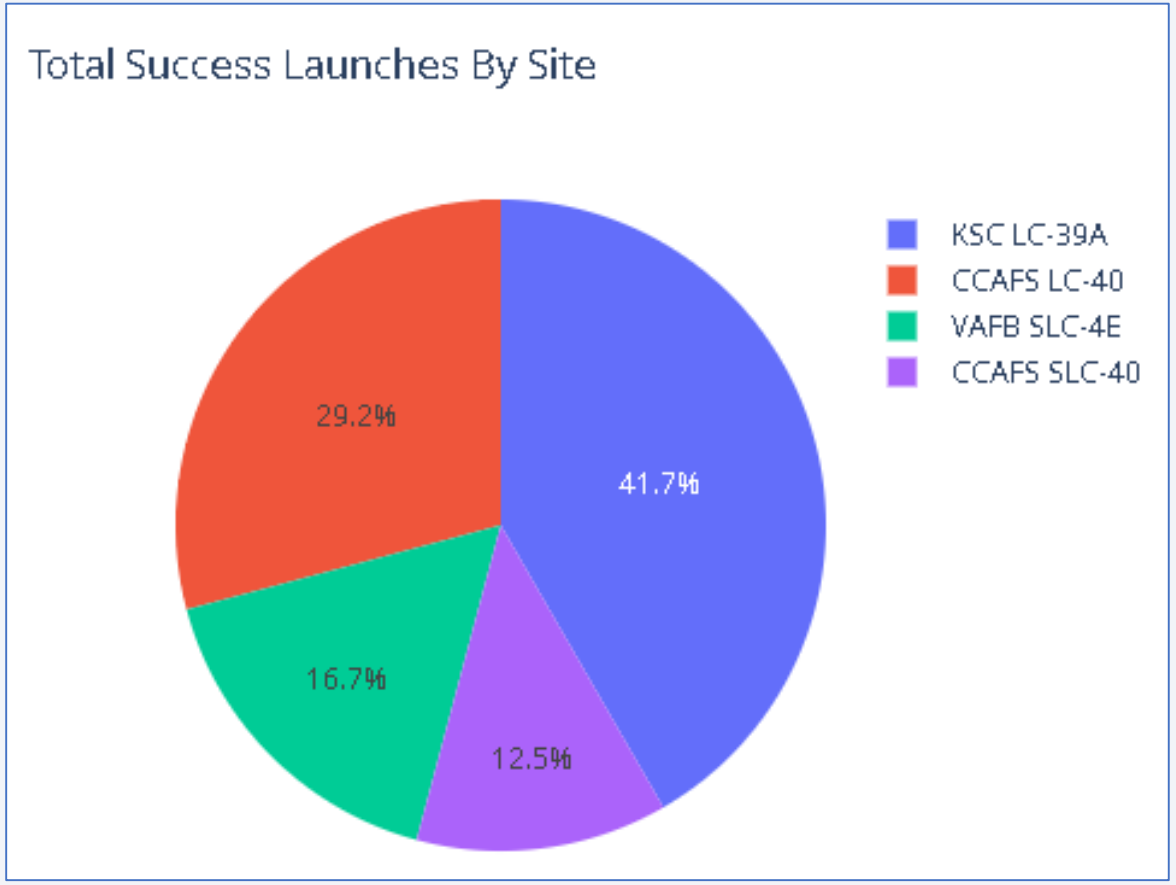
Section 5

# Build a Dashboard with Plotly Dash



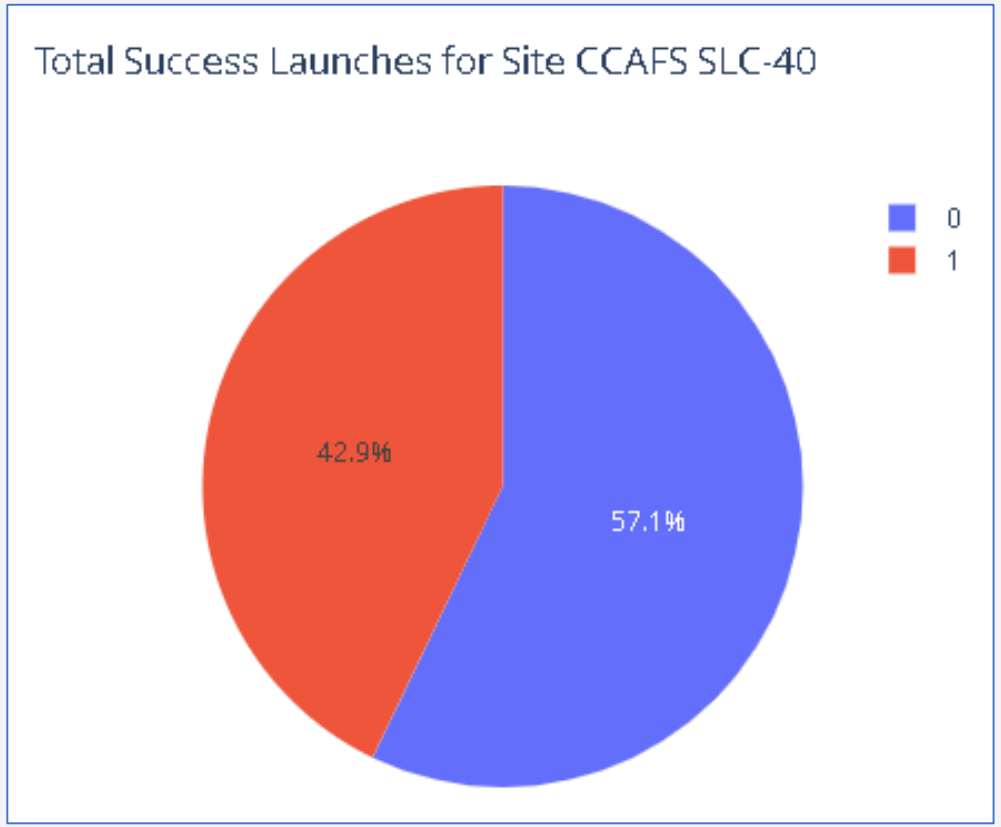
# Total Success Launches by Site

- The adjacent pie chart shows the total success launches by site.
- It is noted that Launch Site ‘KSC LC-39A’ has the highest number of success launches while the launch site ‘CCAFS SLC-40’ has the least.

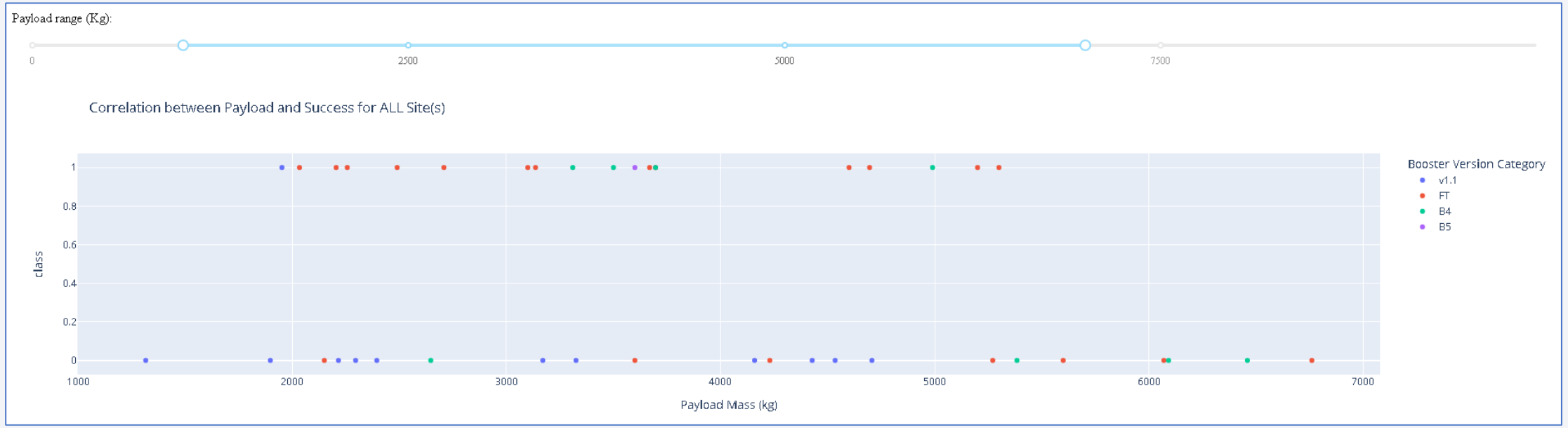


# The Launch Site With Highest Success Ratio

- The pie chart for the launch site with highest launch success ratio is shown on the right.
- The launch site ‘CCAFS SLC-40’ has 42.9% successful launches.



# Correlation between Payload and Success Outcome



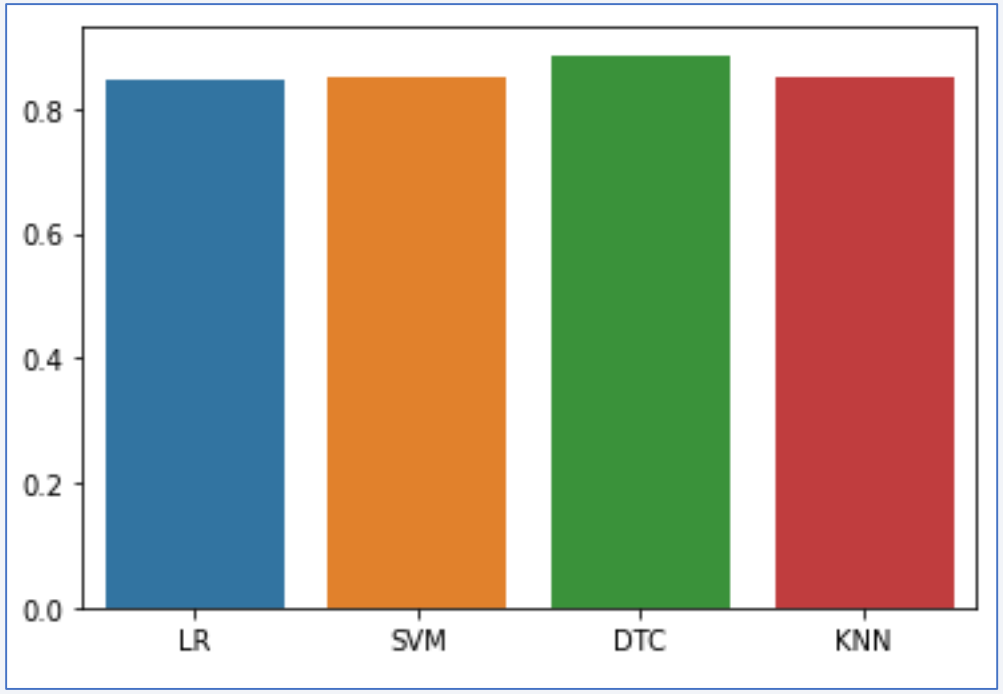

- The above screenshots shows Payload vs. Launch Outcome scatter plot for all sites.
- It is observed that the FT Booster version has the highest success rate while the payload range between 3000kgs and 4000kgs has the higher success rate.

Section 6

# Predictive Analysis (Classification)

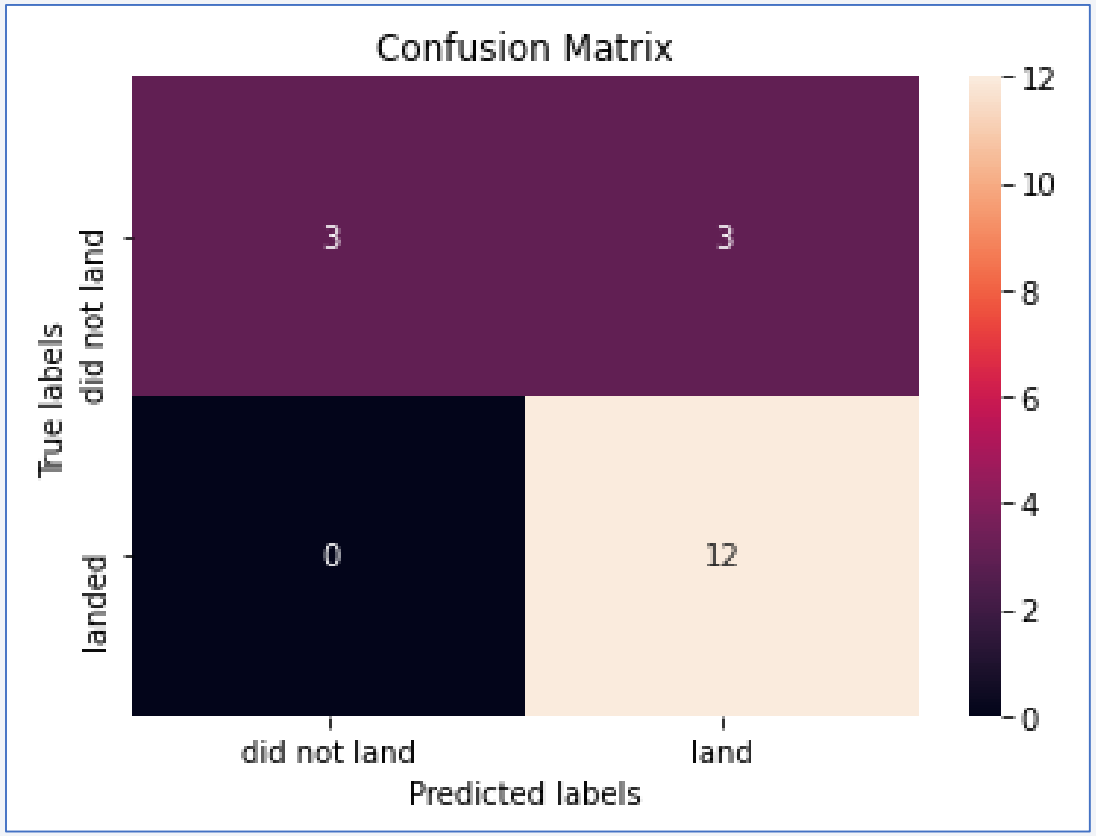
# Classification Accuracy

- The bar chart on the right shows the accuracy for all the built classification models.
- As shown in the graph, the Decision Tree Classifier Model has the highest accuracy of 88.57%.



# Confusion Matrix

- The confusion matrix of the best performing model – Decision Tree Classifier is shown in the figure.
- We note that the model predicted the outcome with a high degree of accuracy, having zero false positive and limited false negative.



# Conclusions

- We needed to determine the price of each launch by gathering information about the Space X launches, determine if the Space X will reuse the first stage for which we trained a machine learning model and used public information to predict if Space X will reuse the first stage or not.
- We used flight data from Space X public API and the Wiki pages to gather and collect information, performed data cleaning and wrangling, replaced necessary null values and discarded unnecessary information from the data.
- We studied the trend and found positive impact with increasing year, flight number and altitude of the orbit, and negative impact on the payload mass.
- By using an interactive map, we found out that the launch sites are more closer to coasts, railway lines and highway and farther from the cities.
- We also conclude that Launch Site ‘KSC LC-39A’ has the highest number of success launches while ‘CCAFS SLC-40’ has the highest success ratio.
- Our decision tree model predicts with accuracy that 88.57% the first stage is likely to land successfully.

# Appendix

- The raw data was collected from:
  - Past launch data from SpaceX API: https://api.spacexdata.com/v4/launches/past
  - Webscraping from the Wiki: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

# Thank you!