IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neha Adkar
01 Aug 2024

Outline

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

Executive Summary

- Methodologies:
  - We collected data through API and Webscraping
  - Performed data wrangling
  - Performed Exploratory Data Analysis(EDA) using visualization and SQL
  - Created Markers on Map using Folium
  - Created Dashboard to show the results
  - Trained and tested data using Machine learning and predicted results
- Summary of all results
  - EDA and Dashboard gave better insights in visualizing best performing launch sites, booster versions, orbits and success rates
  - Machine learning helped us predict the best model to determine the important characteristics of the datasets

Introduction

- The project was created to analyse data of SpaceX and use the result to evaluate viability for the company SpaceY to compete with Space X
- Problems needed to be answered:
    - Approximate cost of launches, their site locations
    - Which models can give accurate predictions
    - Which boosters, orbits, payload and other factors can give highest success rate of launches

Section 1

Methodology

Methodology

Executive Summary

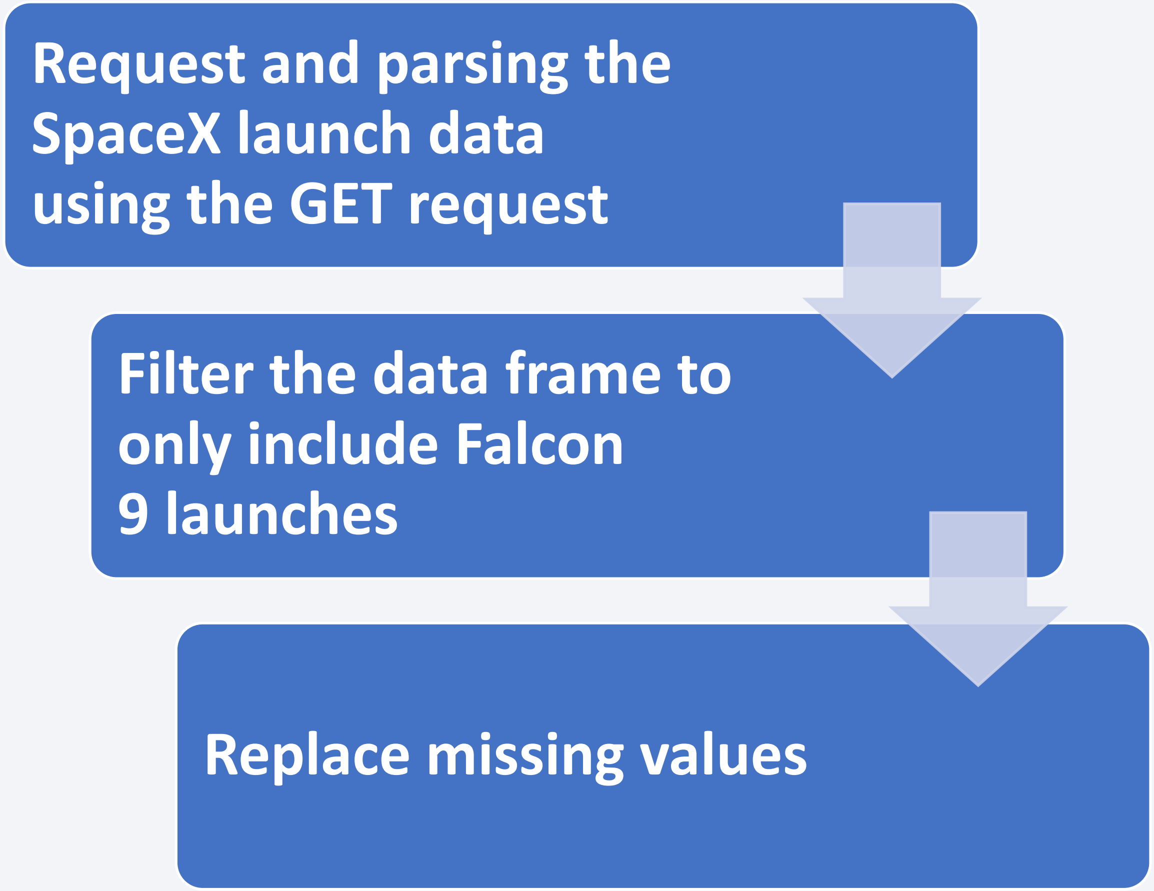
- Data collection methodology:
  - Using API
  - Using web scraping from Wiki pages
- Perform data wrangling
  - Summarizing and analyzing features from data collected by above methods
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data to be divided in training and testing data sets, classified by different models and accuracy score to be calculated to determine best performing model

Data Collection

- Data sets were collected by 2 methods
  - API
  - Webscraping from wiki page
- Used function like .json_normalise() to normalize data and converted data into pandas data frame in API
- Used Beautiful Soup library to extract data from web page, parse the data to html tables, later converted the tables into pandas data frame in webscraping

Data Collection – SpaceX API

- Here we fetched the data from SpaceX API. We converted the data into a data frame to perform basic data formatting to remove unwanted and missing values
- Collected values for: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Sites, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pads, Block , Reused Count , Serial, Longitude, Latitude
- SpaceX API calls notebook



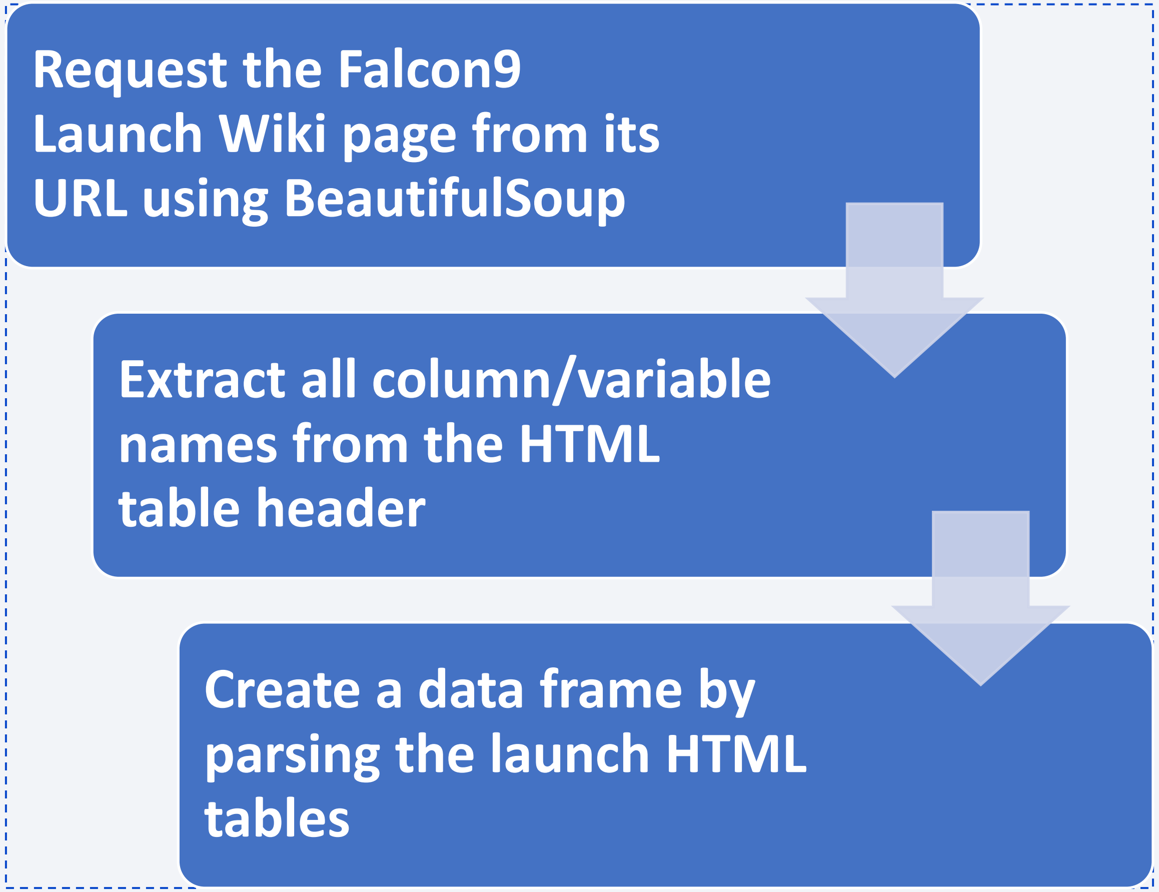
Request and parsing the SpaceX launch data using the GET request

Filter the data frame to only include Falcon 9 launches

Replace missing values

Data Collection - Scraping

- Here we scraped the data from a from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches to collect Falcon 9 historical launch records.
- Data extracted: Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- Web scraping notebook link



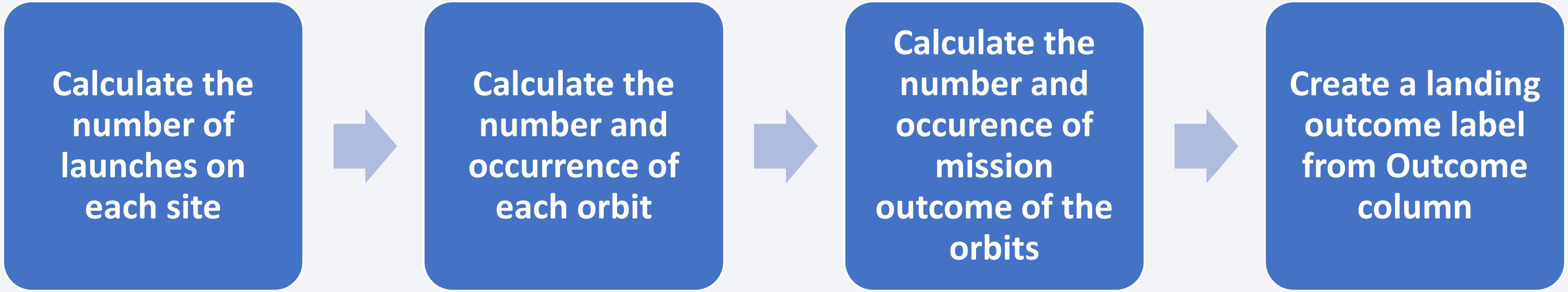
Request the Falcon9 Launch Wiki page from its URL using BeautifulSoup

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Here data like number of launches on each site, number and occurrence of each orbit, number and occurrence of mission outcomes were calculate to determine the landing outcome of the flight launches (Successful or Unsuccessful). These were determined based on launches performed in ocean, land and ground pad.
- Data wrangling notebook



Calculate the number of launches on each site

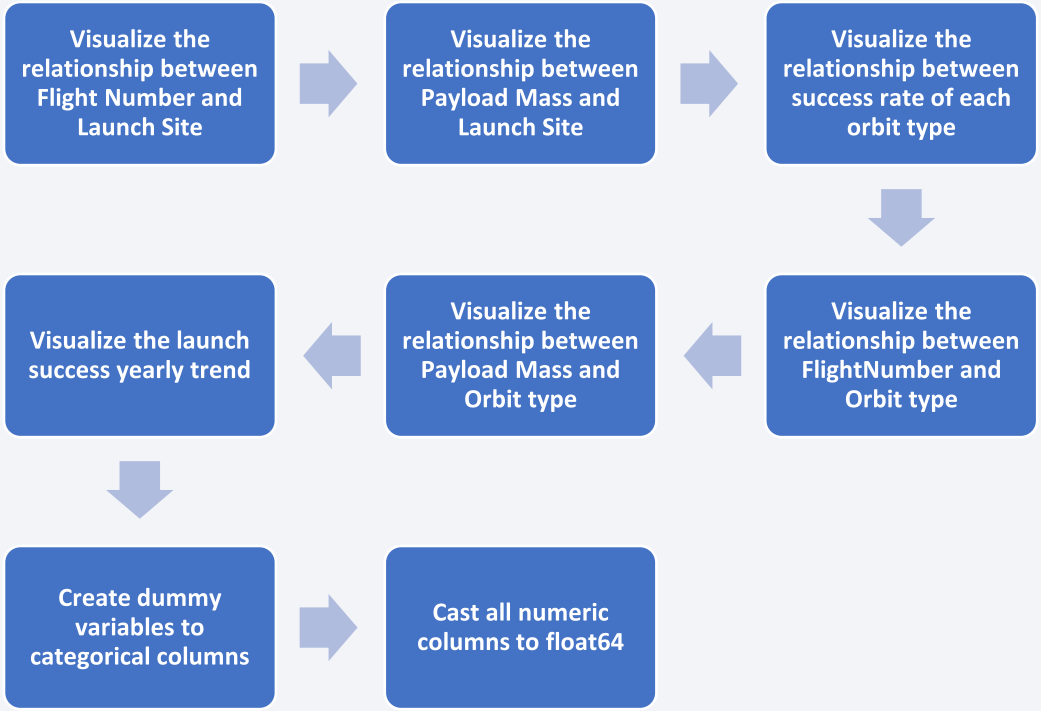
Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

EDA with Data Visualization

- Several charts were plotted to visualize the various categories with each other. Scatter plot was used to visualize relationship between Flight Number Launch Site, Payload Mass and Orbit type. To show success rate per Orbit, bar chart was used and to show the yearly trend of success rate, line chart was used.
- [EDA with data visualization notebook]



Visualize the relationship between Flight Number and Launch Site

Visualize the relationship between Payload Mass and Launch Site

Visualize the relationship between success rate of each orbit type

Visualize the relationship between FlightNumber and Orbit type

Visualize the relationship between Payload Mass and Orbit type

Visualize the launch success yearly trend

Create dummy variables to categorical columns

Cast all numeric columns to float64

EDA with SQL

- Here we used SQL to analyze the record for each payload carried during a SpaceX mission into outer space and determine the cost of first launch.
- Used “Distinct” to fetch unique Launch sites record
- Used “Like” to fetch records which have the given string present
- Used “min”, “max”, “count”, “avg”, “sum” to calculate minimum, maximum, count, average and total of records
- Used “group by” to perform calculations on the grouped data(success, failure)
- Used “order by” to sort data in ascending or descending order
- Created a subquery to list the names of the booster_versions which have carried the maximum payload mass.
- EDA with SQL notebook

Build an Interactive Map with Folium

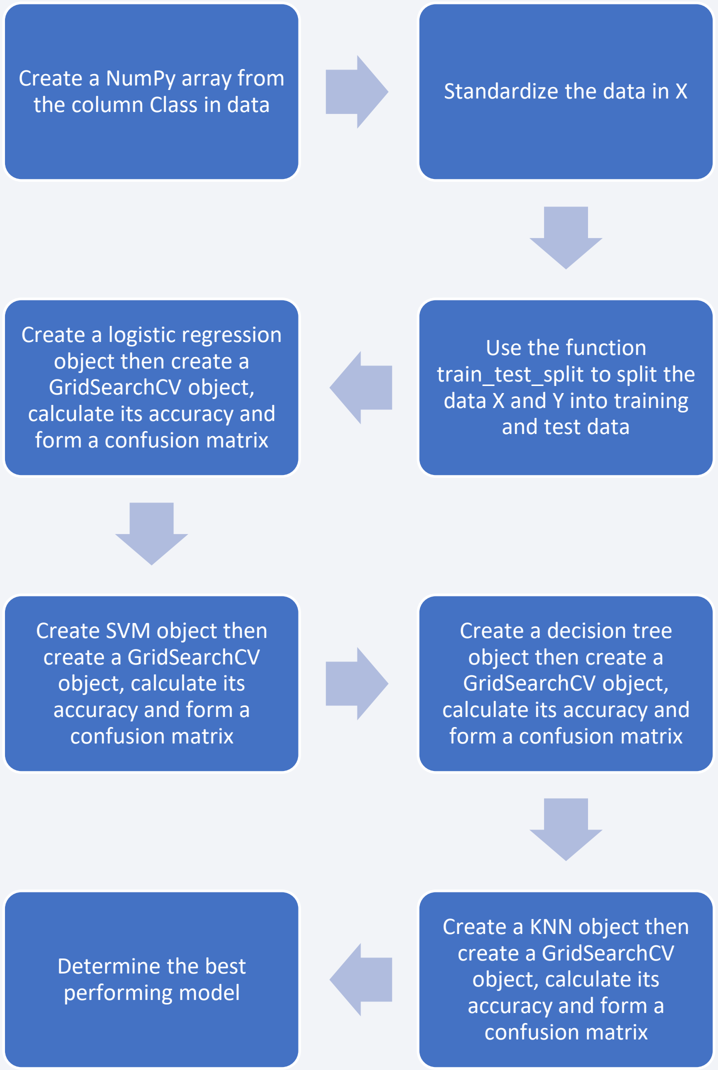
- Several markers and polylines were created on the map using Folium to show existing launch sites coordinates, their success rates and their proximities so that a better location can be chosen to build the launch site and increase its success rate
- Added circle and markers to indicate launch sites on the map using folium.circle and folium.map.Marker respectively
- Added markers for a cluster to show the successful and failed launches for the launch sites
- Created polylines to show the distances between a launch site and its proximities(nearby coastline, highway, railway or city)
- Interactive map with Folium map

Build a Dashboard with Plotly Dash

- The dashboard provides insights into the SpaceX launch data and enable users to answer questions about the data and help users understand the relationships between launch site, payload range, and launch success.
- In the dashboard, a dropdown list is created which allows users to select a specific launch site or all sites, and a pie chart shows the total successful launches count for all sites or the Success vs. Failed counts for the selected site.
- A slider is also created which enables users to filter the data by payload range, and a scatter chart shows the correlation between payload and launch success for the selected site(s).
- Plotly Dash lab

Predictive Analysis (Classification)

- Here we created a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- Used models like Logistic Regression, SVM, decision trees, K nearest objects to train the data.
- By training, testing and modeling the data, we got to know that Decision tree was best suitable for this dataset as it gave the highest accuracy score
- Predictive analysis lab



Create a NumPy array from the column Class in data

Standardize the data in X

Use the function train_test_split to split the data X and Y into training and test data

Create a logistic regression object then create a GridSearchCV object, calculate its accuracy and form a confusion matrix

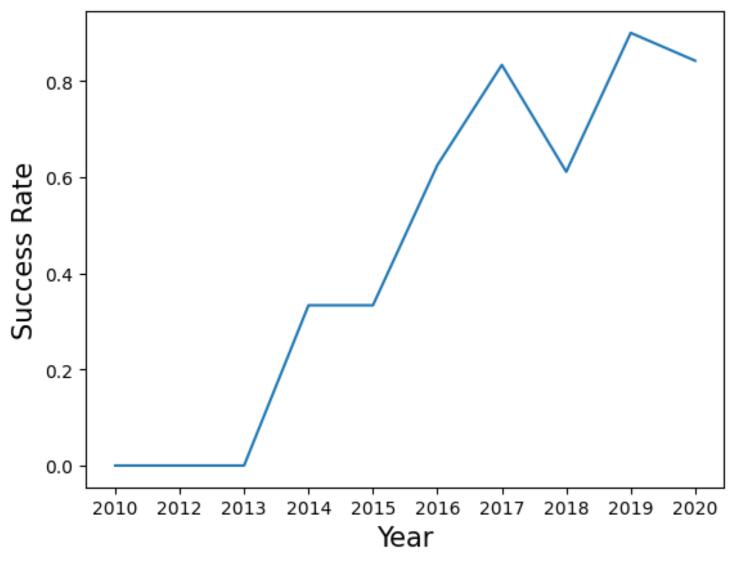
Create SVM object then create a GridSearchCV object, calculate its accuracy and form a confusion matrix

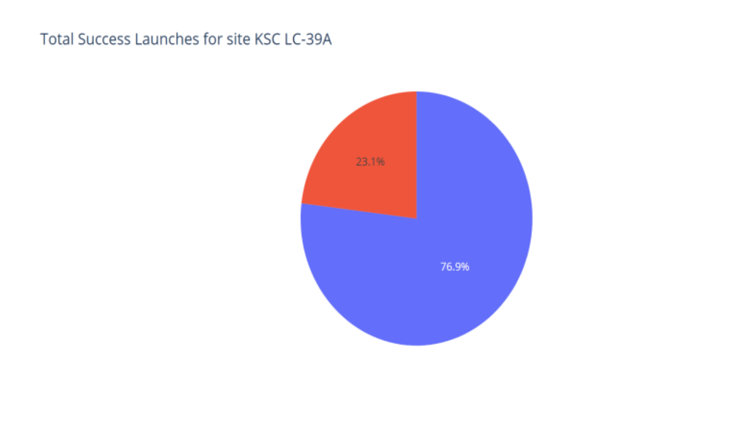
Create a decision tree object then create a GridSearchCV object, calculate its accuracy and form a confusion matrix

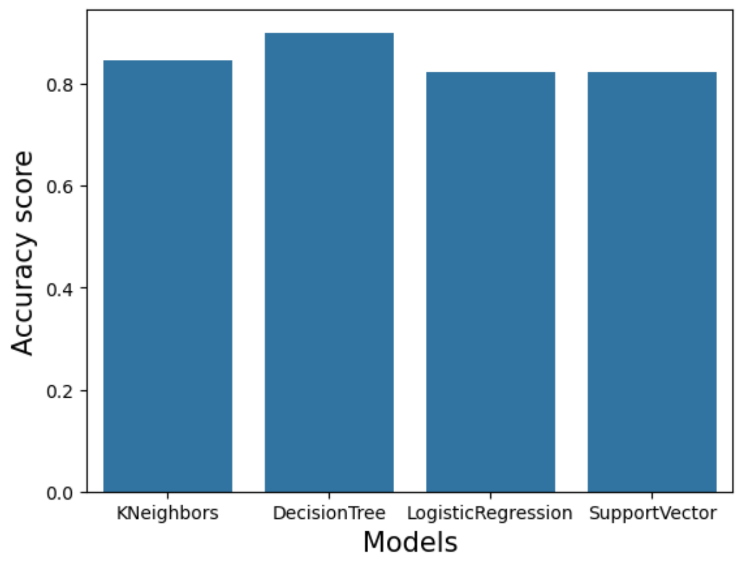
Create a KNN object then create a GridSearchCV object, calculate its accuracy and form a confusion matrix

Determine the best performing model

Results




Total Success Launches for site KSC LC-39A

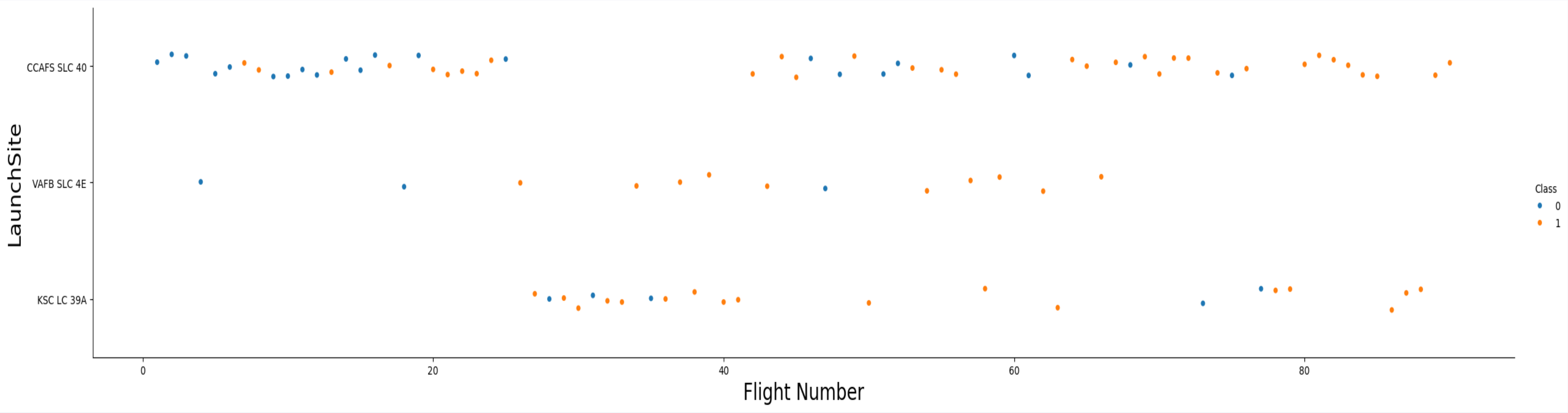


Section 2

Insights drawn from EDA

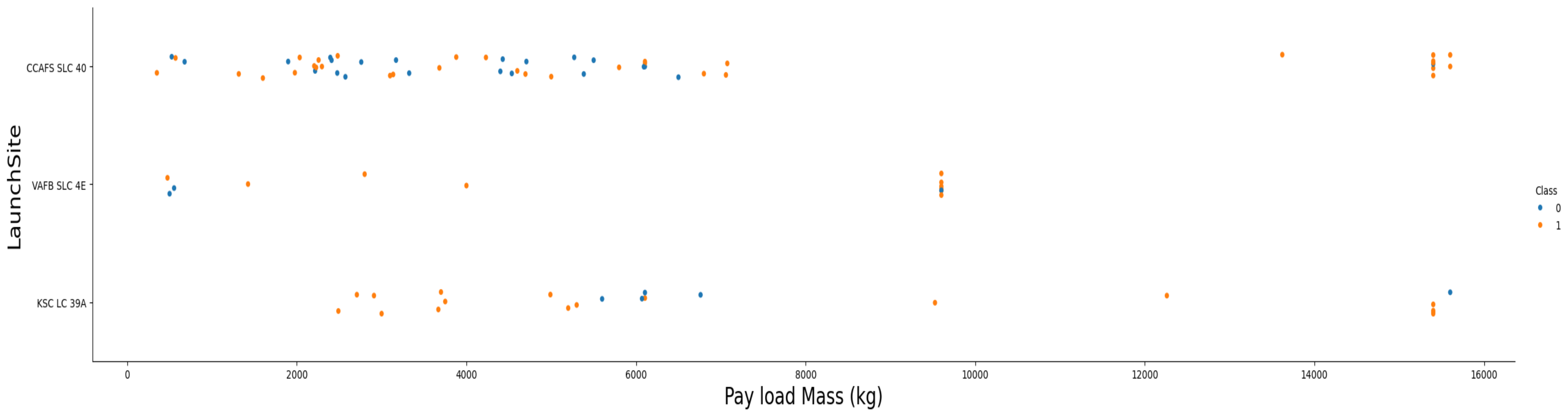
Flight Number vs. Launch Site

- Below is the scatter plot to show relation between Flight number and Launch Site
- VAFB SLC 4E had success despite low flight counts. CCAFS SLC 40 having highest flight count have approximately equal number of success and failures



Payload vs. Launch Site

- Below is a scatter plot showing relation between Payload mass and Launch Site
- We observe that VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000)

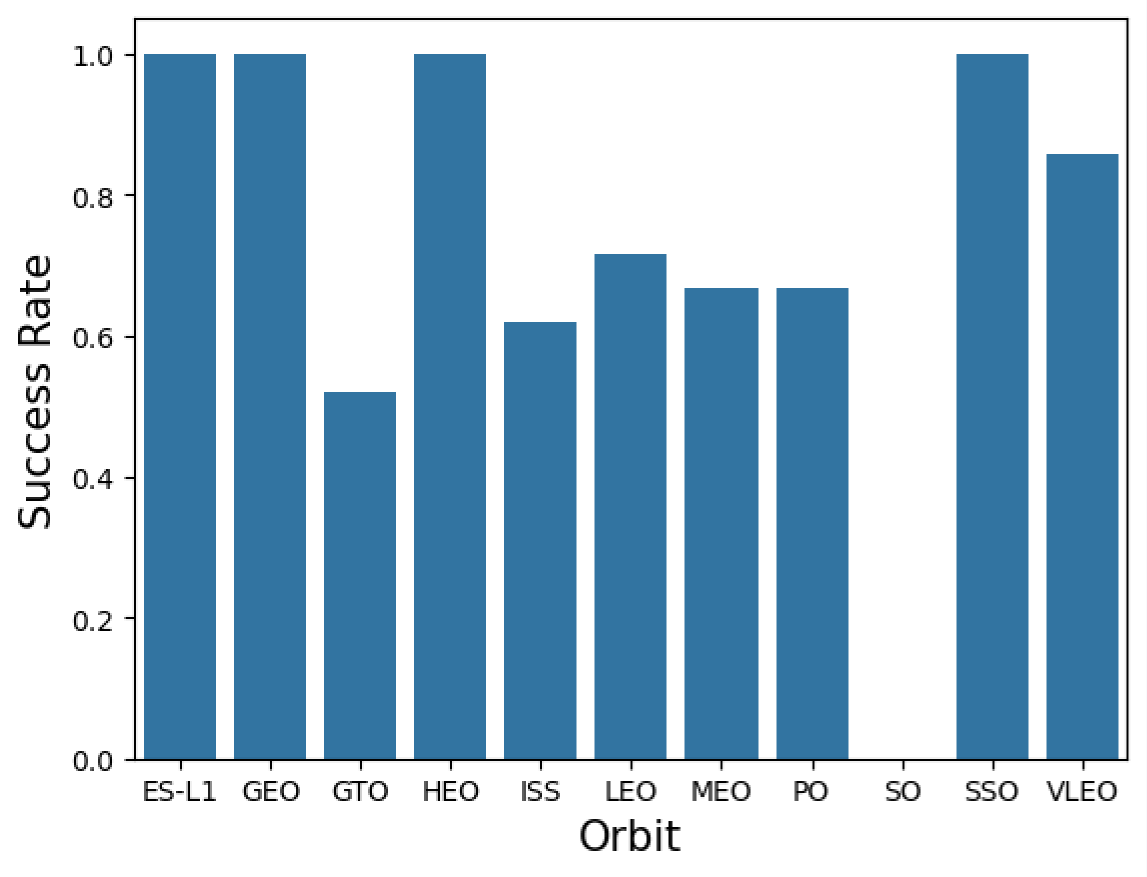


Success Rate vs. Orbit Type

- Here is the bar plot to check relationship between success rate and orbit type

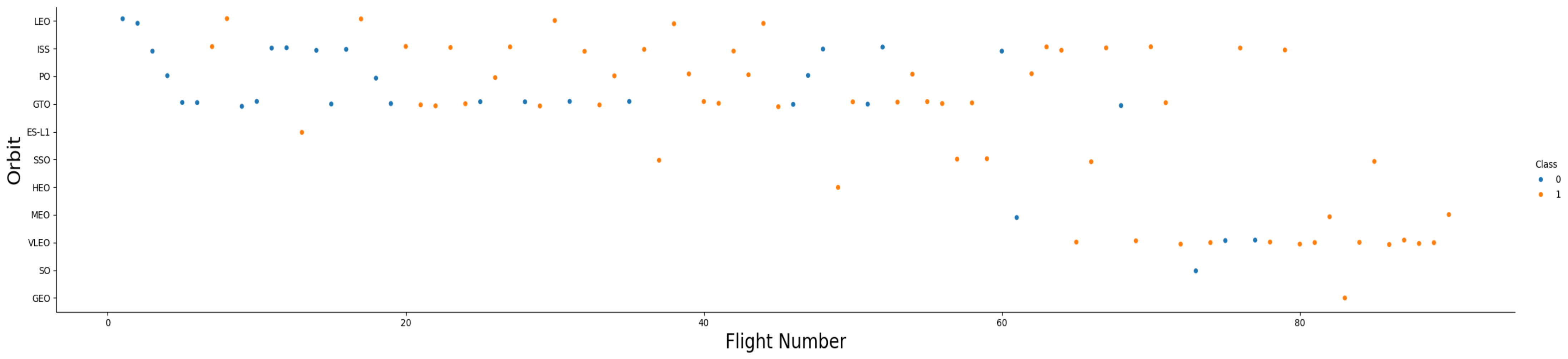
- ES-L1, GEO, HEO, SSO have highest success rate of 100% while SO has the lowest rate of 0%

- Rest got fair success rate above 50%



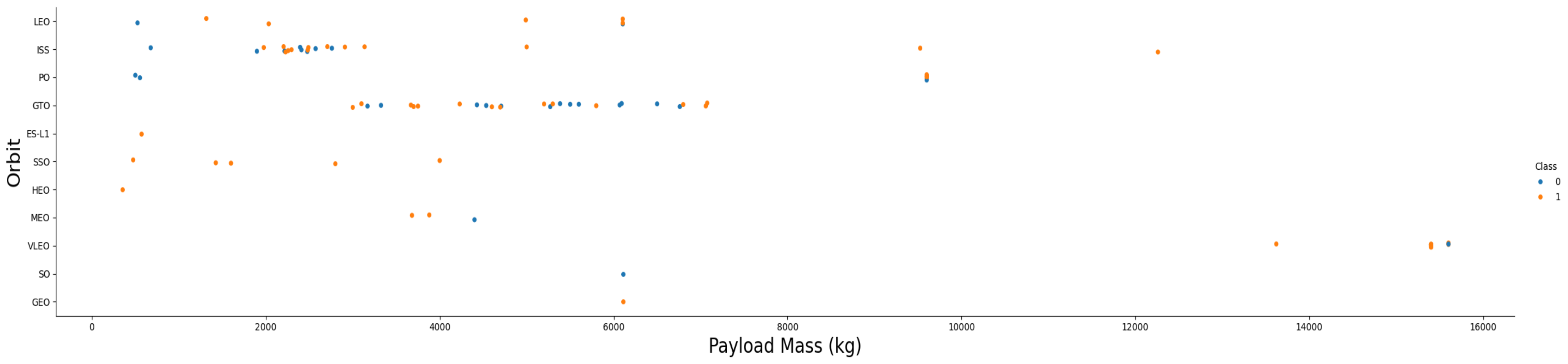
Flight Number vs. Orbit Type

- Below a scatter plot is created to see if there is any relationship between FlightNumber and Orbit type.
- We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



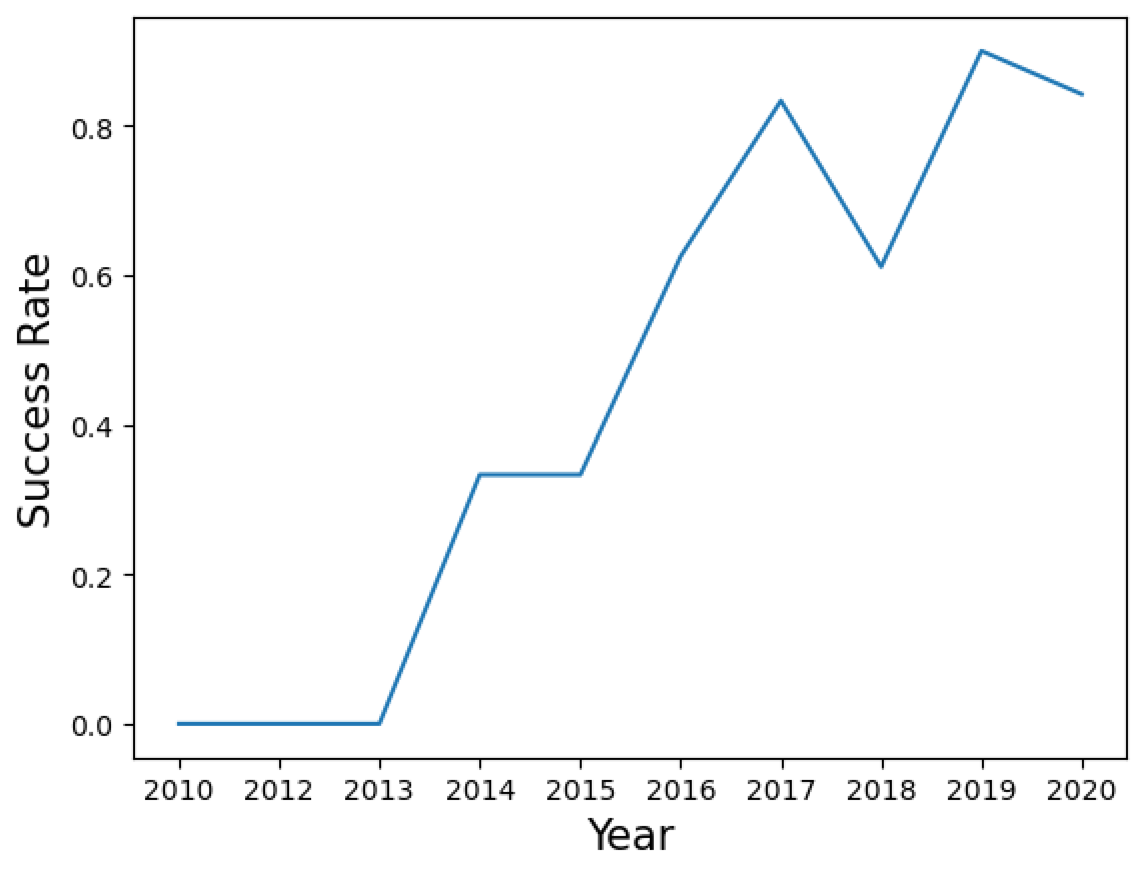
Payload vs. Orbit Type

- Below is the Payload Mass vs. Orbit scatter point chart to reveal the relationship between Payload Mass and Orbit type
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



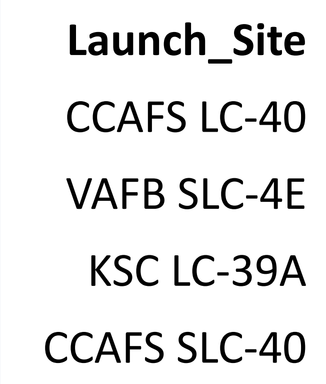
Launch Success Yearly Trend

- The chart shows average launch success trend from 2010 to 2020
- We can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Here is the list of Launch sites extracted through DISTINCT feature



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

Launch Site Names Begin with 'CCA'

- Below is the first 5 records which start with ‘CCA’
- These were extracted using “like ‘%CCA’” in select query

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

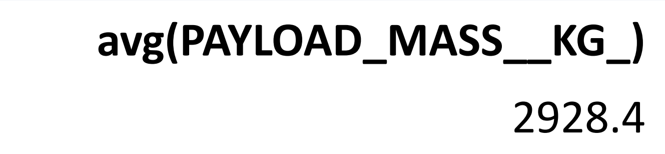
Total Payload Mass

- Below is the total payload carried by boosters from NASA
- Count was extracted using SUM

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

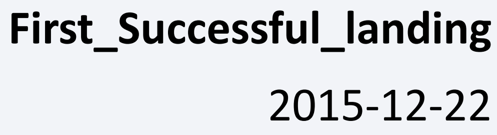
- Below is the average payload mass carried by booster version F9 v1.1
- Query was extracted using Avg



avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Below is the date of the first successful landing outcome on ground pad
- Extracted using Min feature



First_Successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below is the list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

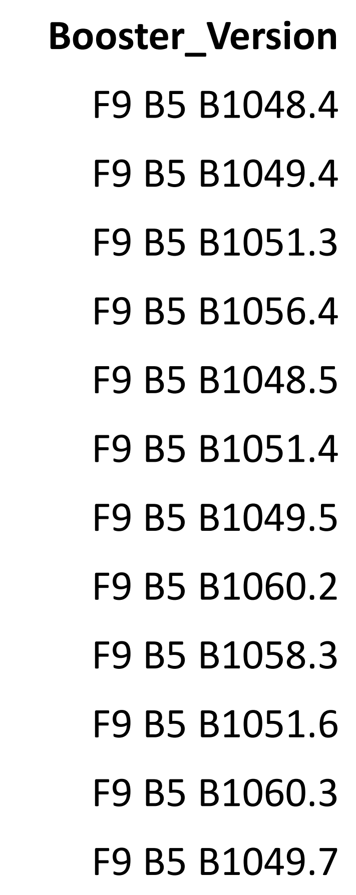
Total Number of Successful and Failure Mission Outcomes

- Below is the total number of successful and failure mission outcomes
- Extracted using count feature

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

Boosters Carried Maximum Payload

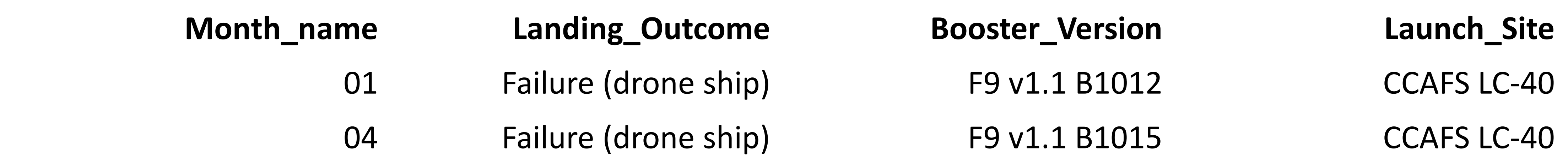
- Below is the list of names of the booster which have carried the maximum payload mass
- Extracted using subquery and max feature



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

- Below is the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Extracted using Substring feature as month had to be extracted for Date



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

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

- Below is Ranked list of 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

- Extracted the list using group by feature

Landing_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

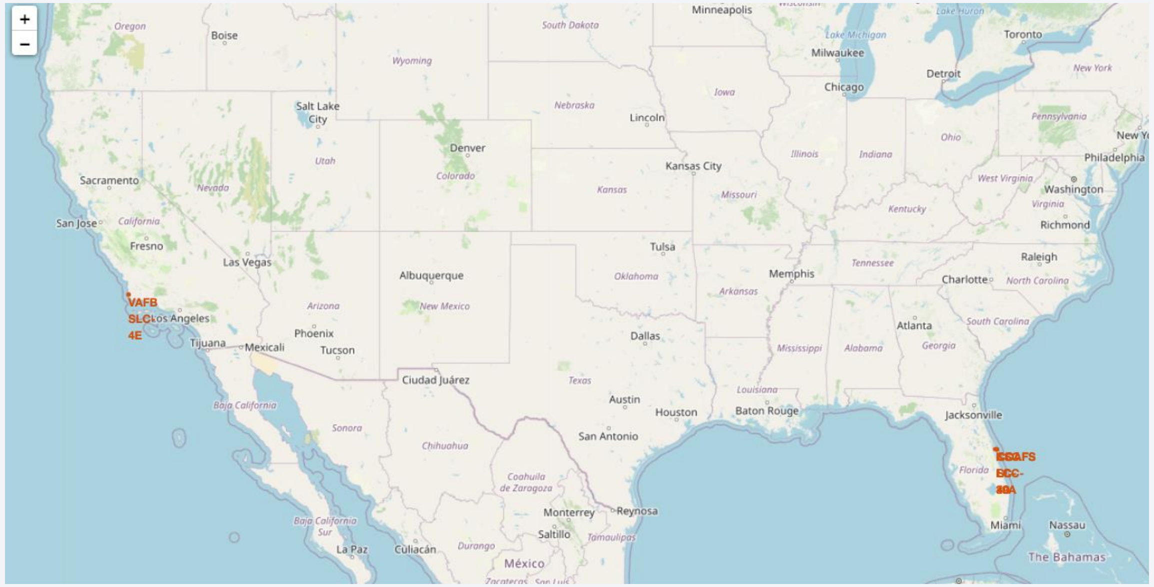
Section 3

Launch Sites Proximities Analysis

All launch sites’ location

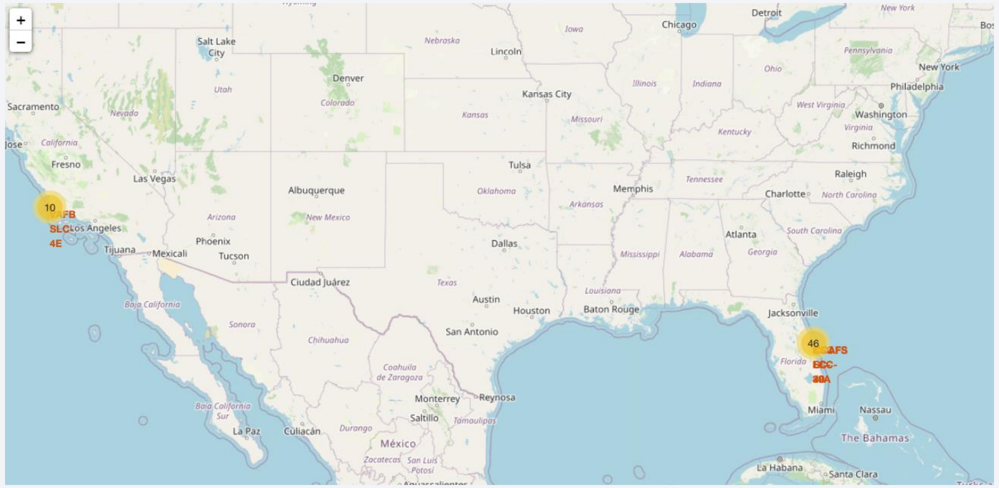
- Here is the generated folium map and make a proper screenshot to include all launch sites’ location markers on a global map

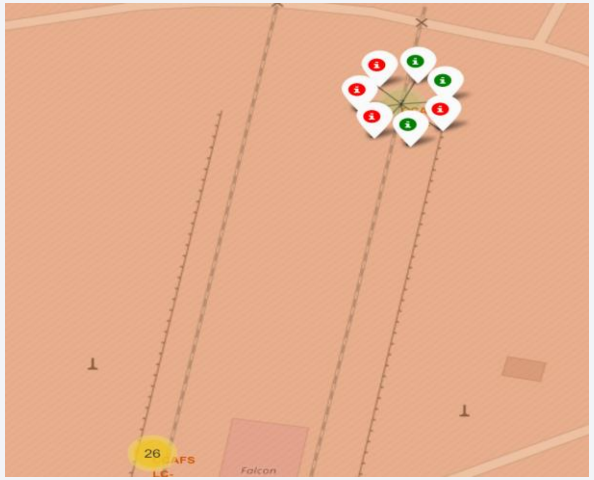
- Extracted using Folium.circle and Folium.Marker feature



Color-labeled launch outcomes

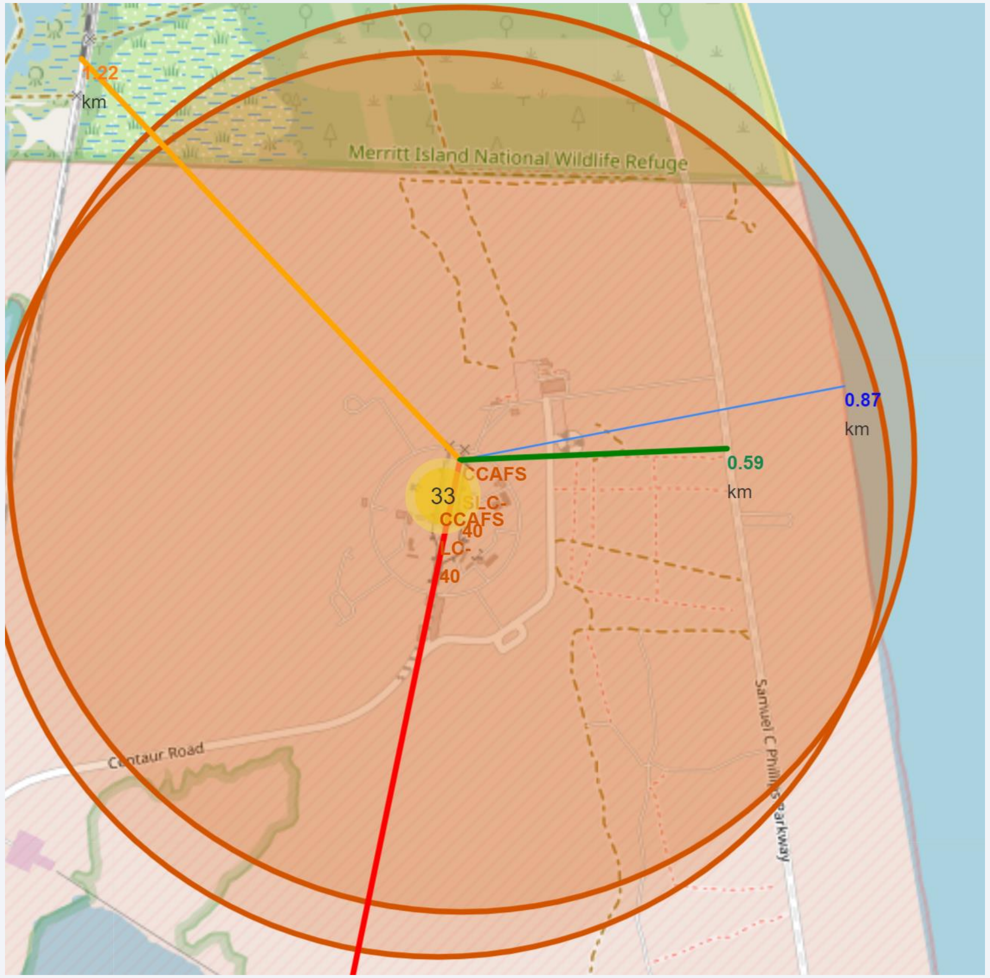
- The screenshot shows the color-labeled launch outcomes on the map which indicates number of launches performed on each site.
- Green marker shows the launch was a success whereas Red shows it was failure

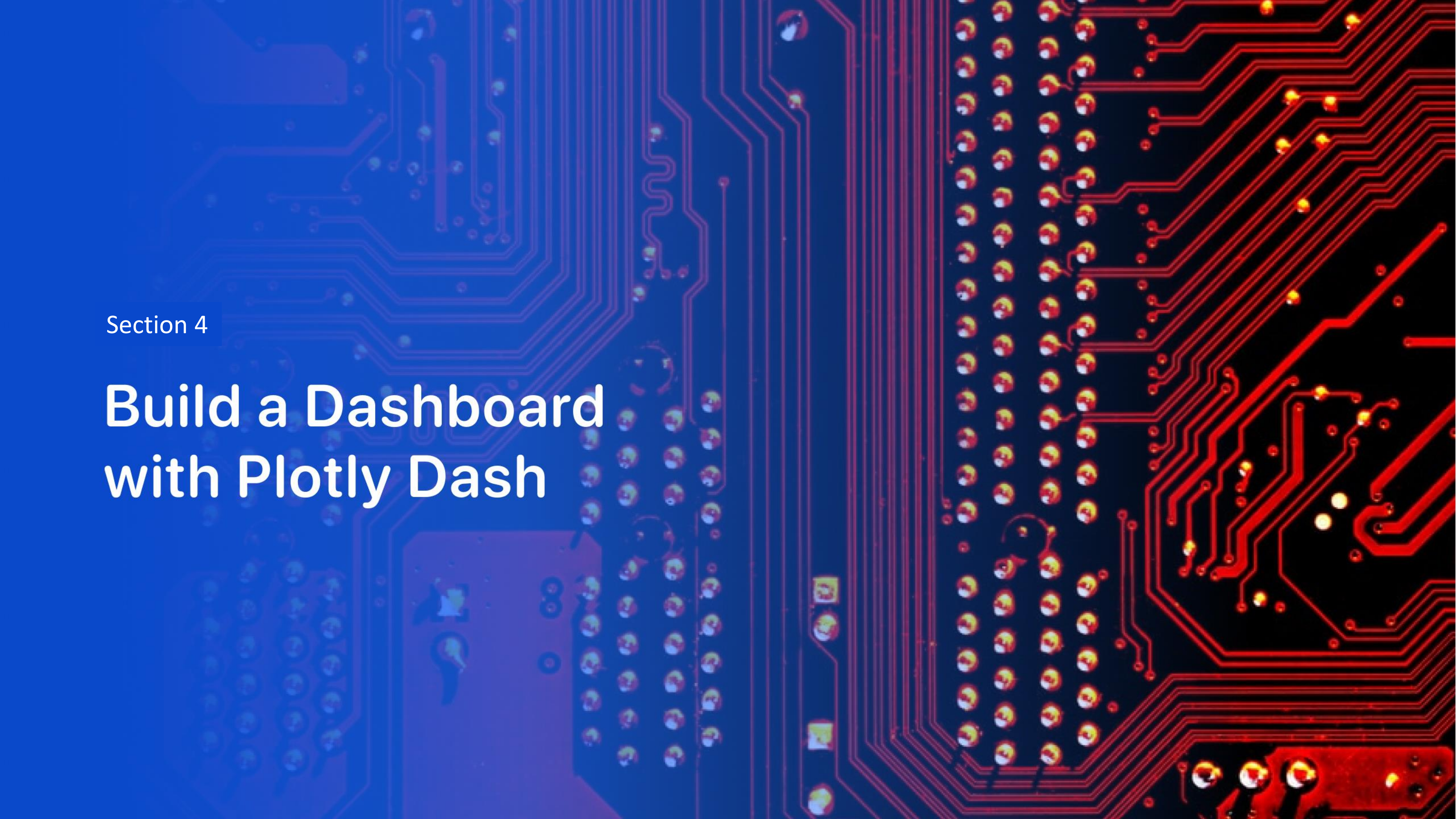




Launch site's proximities

- The screenshot shows a selected launch site to its proximities such as railway, highway, coastline, with distance calculated
- Highway is the closest to this launch site which can bring ease in transport of materials required. Having coastline nearby can be useful for the site to experiment on ocean landings.


1.22
km
Merritt Island National Wildlife Refuge
0.87
km
0.59
km
CCAFS
33
SLC-
CCAFS
40
LC-
40
Centaur Road
Samuel C Phillips Parkway

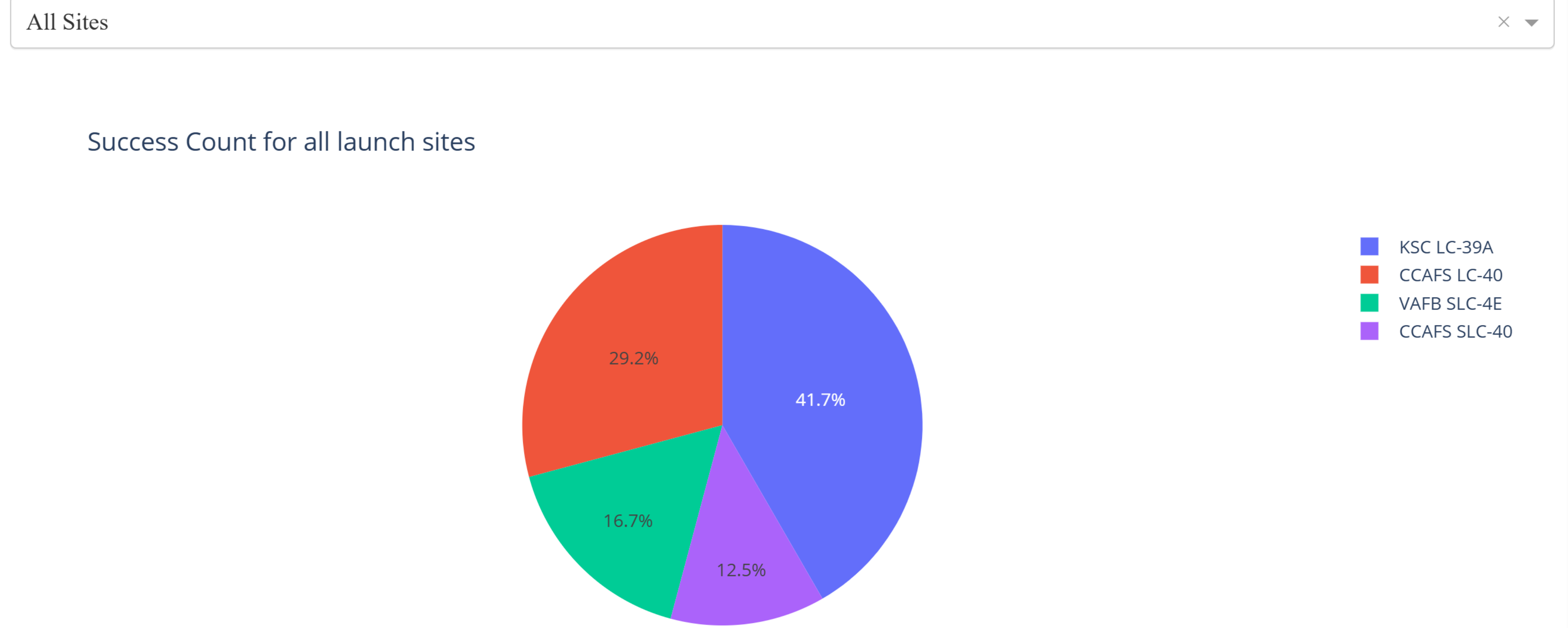


Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites

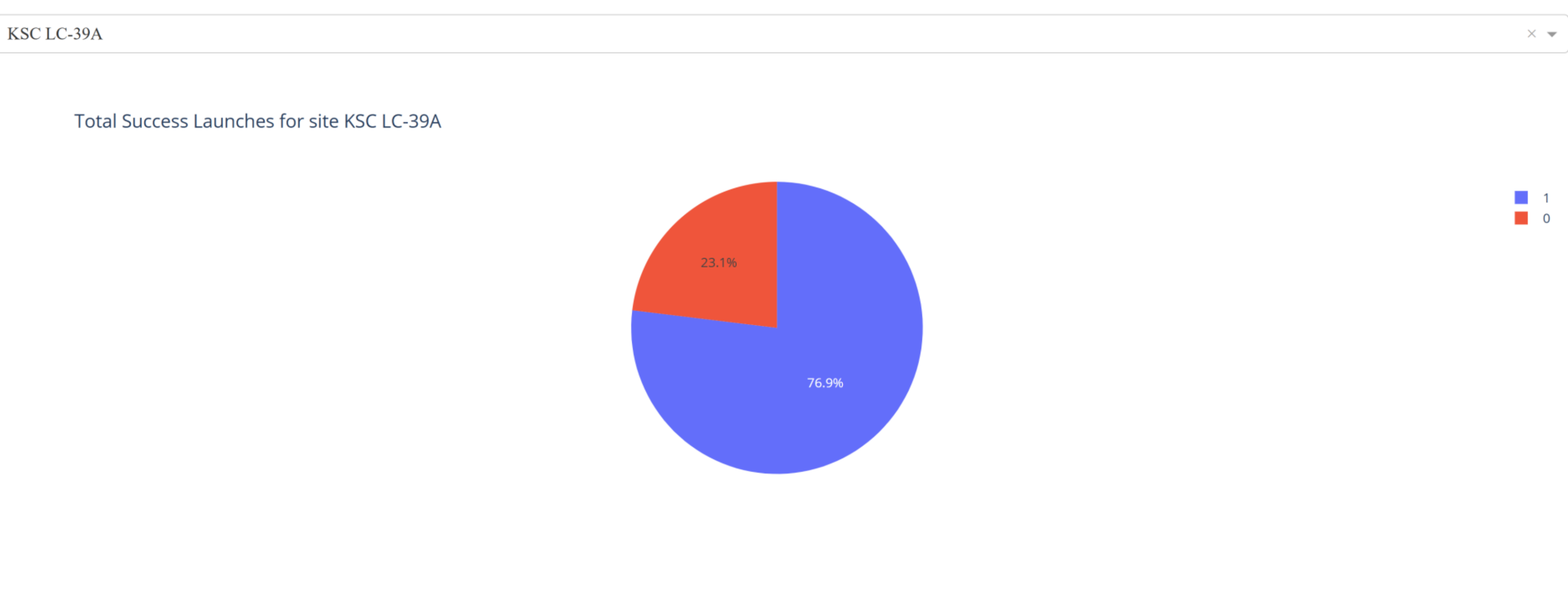
- Below pie chart shows the percent of success count for the launch sites which lead by KSC LC-39A



All Sites

Highest launch success ratio

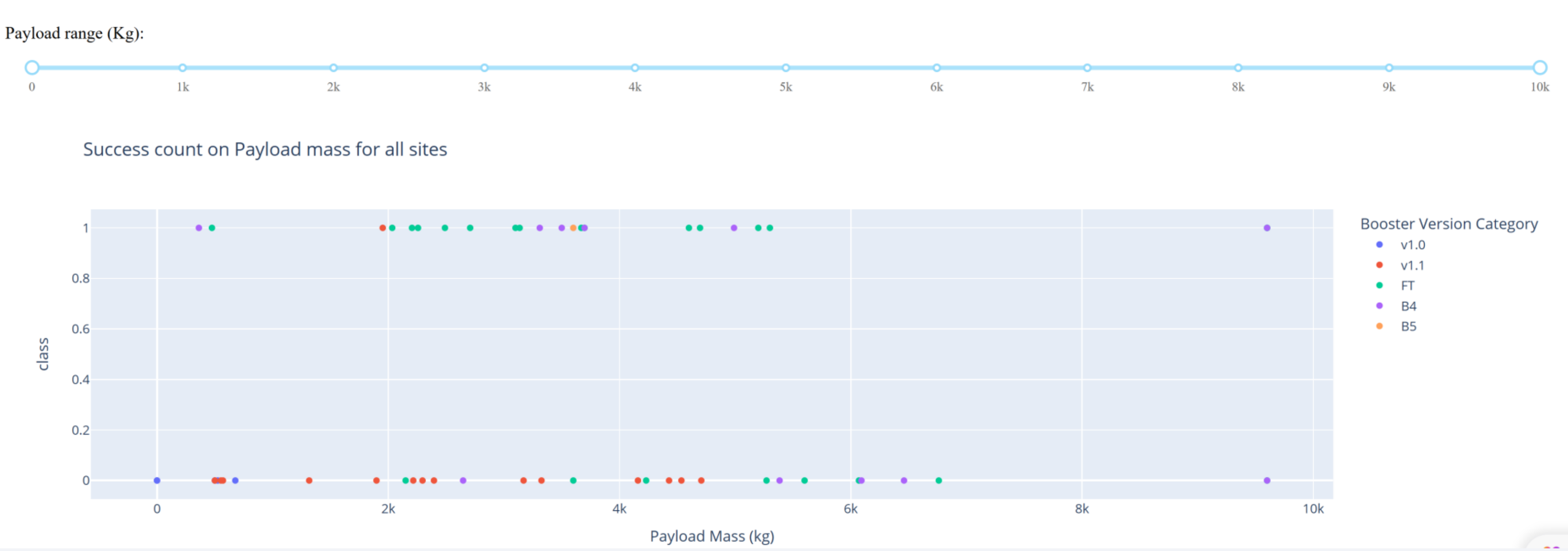
- Here we can observe KSC LC-39A having higher success rate compared to failure rate ratio



KSC LC-39A

Payload vs. Launch Outcome

- Here is the scatter plot depicting relation between payload and launch outcome.
- We can observe that FT and B4 booster versions have highest success count
- V1.0, v 1.1 and B5 have major failure count

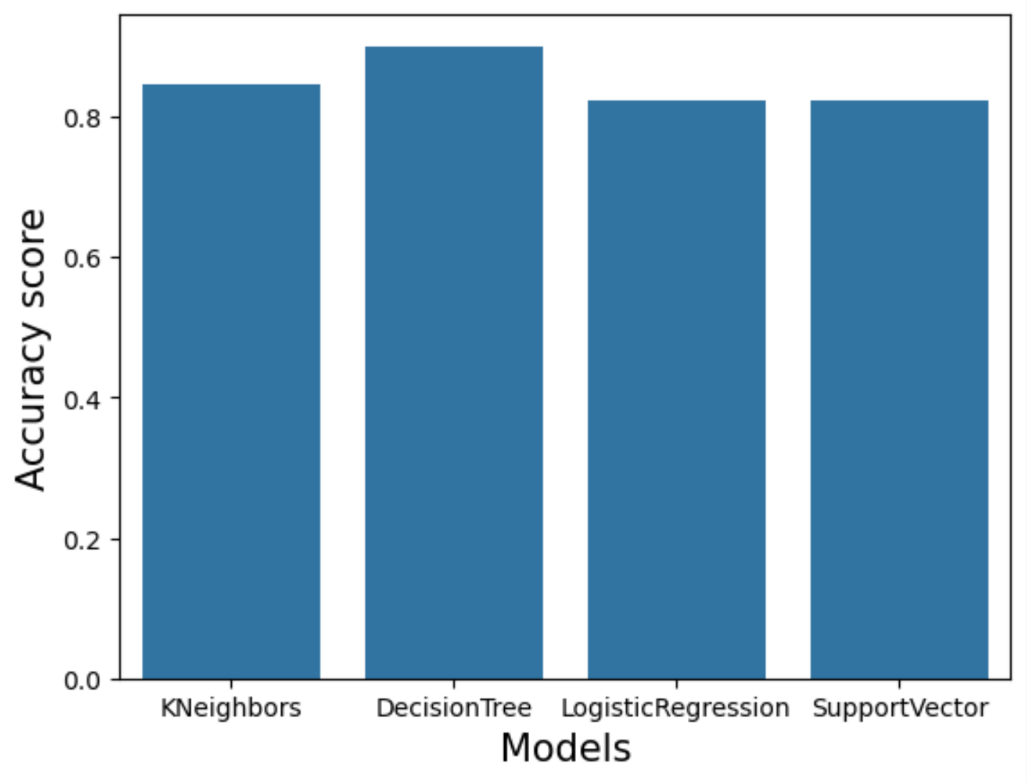


Section 5

Predictive Analysis (Classification)

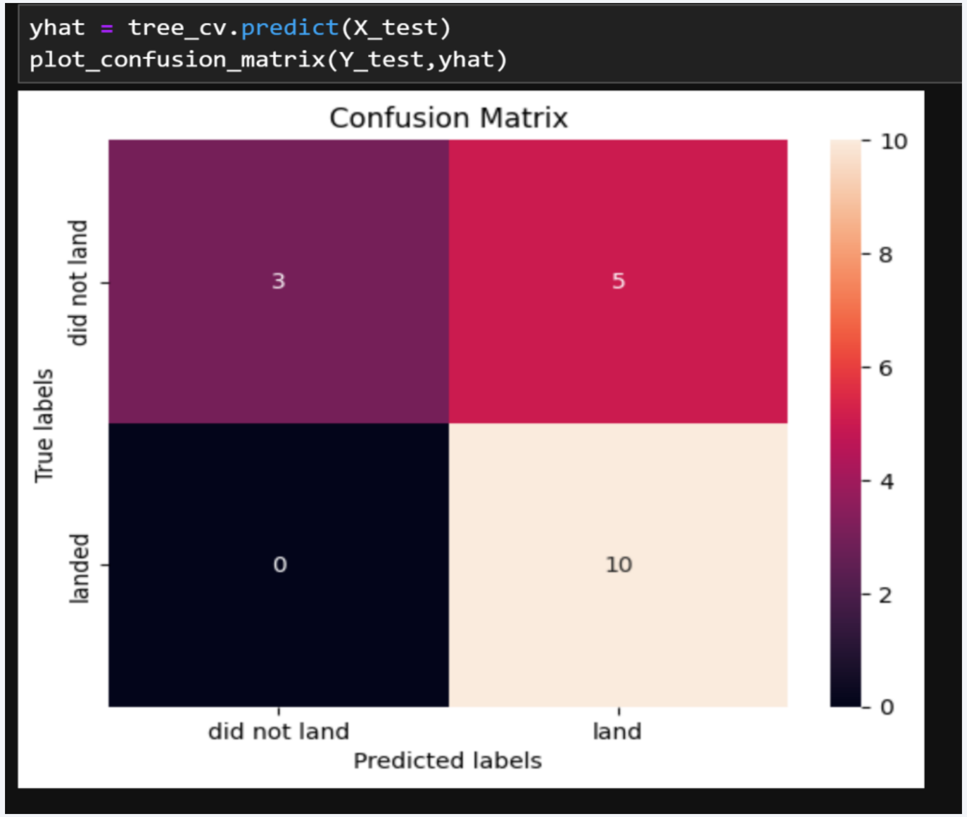
Classification Accuracy

- Here is a bar plot showing accuracy score for different classification models
- Decision tree showed the best accuracy score compared to other models



Confusion Matrix

- Below is the confusion matrix for optimal model determined earlier: Decision tree
- The model demonstrated a relatively high accuracy in predicting whether something has landed or not. It correctly identified 10 instances of landing (true positives) and 3 instances of not landing (true negatives). However, there were 5 instances where the model incorrectly predicted landing (false positives). No instances were incorrectly predicted as not landing (false negatives).



```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Conclusions

- KSC LC-39A had the highest success rate as a launch site
- FT and B4 booster versions had the highest success count
- ES-L1, GEO, HEO, SSO orbits had 100% success rate in launches
- Success rate has been increasing since 2013 till date
- Decision tree is the optimal model for the dataset as it gave the highest accuracy score

Appendix

- SpaceX API calls notebook
- Web scraping notebook link
- Data wrangling notebook
- EDA with data visualization notebook
- EDA with SQL notebook
- Interactive map with Folium map
- Plotly Dash lab
- Predictive analysis lab

Thank you!