IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ahmed Shahriar Sakib
09.07.2021

Outline

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

Executive Summary

Summary of methodologies

- In this project, the Falcon9 Launch data was collected via SpaceX API and Falcon 9 launch wiki page by web scraping using BeautifulSoup4, and later on stored in a csv for further analysis. Before performing any exploratory and visual analysis of the data, it was cleaned by getting rid of unwanted features, replacing the missing values, creating a label for the predictive analysis and finally, converting all the categorical features into numerical forms. After that, performed several visual analytics to get more insights about features and their relations with launching outcome. Finally, performed a predictive analysis to train and evaluate a best model that can give answer via prediction to our question **whether a landing would be successful or not**.

Summary of all results

- Exploratory and Visual Analysis:
  - Payload greater than 7500 kg had a higher chance of successful landing
  - Among 11 orbit types , ES-L1, GEO, HEO, SSO were 100% successful with less than 6000 kg payload

Executive Summary [Continued]

- Exploratory and Visual Analysis:
  - From 2010 to 2019, 61 missions had successful landing whereas 10 missions had failure landing, leading a success rate of 67.78%
  - Among the four Launch sites “KSC LC-39A” had the highest success rate (41.7%) whereas, “VAFB SLC-4E” had the lowest success rate (12.5%)
  - In 0-10000 kg payload range “FT” booster version has the largest success rate
- Launch Site Proximity Analysis:
  - SpaceX has 4 launch sites, one is near California, the other three is near Florida and South Texas.
  - All the sites are in near proximity to ocean
  - All the sites are bit far away from the city
- Predictive Analysis:
  - Logistic Regression and SVM performed well in comparison with other models, with the highest accuracy of 83.3%

Introduction

Project background and context

- SpaceX is a company that aims to make commercial space travel more affordable for everyone. This company can launch rockets in a relatively inexpensive manner because, it can reuse the first stage of the rocket *Falcon9*, which is the first orbital class rocket capable of reflight
- In this project, we are working for Space Y, which is going to compete with SpaceX. Our goal is to find the price of each launch. And this price mainly depends on the successful landing of the first stage

Problems we want to find answers

- Will SpaceX reuse the first stage?
- Will SpaceX attempt to land a rocket or not? i.e., will the first stage land successfully?

Section 1

Methodology

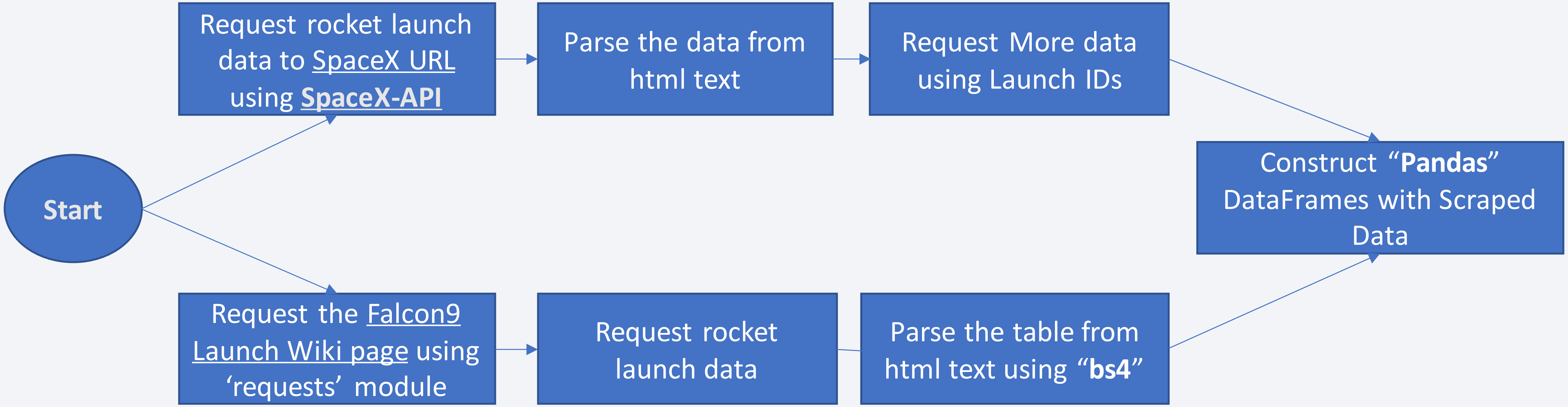
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via SpaceX API and web scraping using python “bs4” package.
- Performed data wrangling
  - Data was processed using python “pandas”, “numpy” packages.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Classification models were built, tuned, evaluated using python “sklearn”, “pandas”, and, “matplotlib” packages

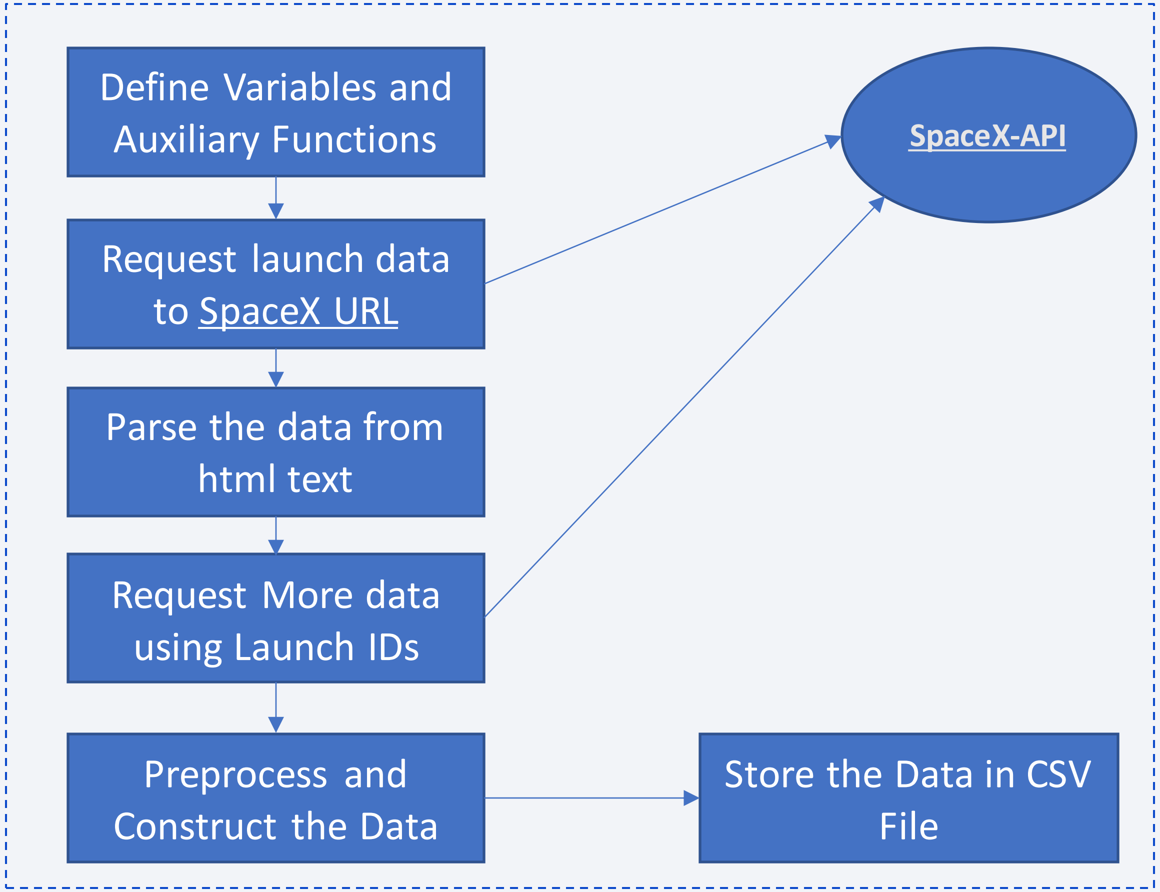
Data Collection

- Data were collected in two separate processes – using SpaceX API and web scraping the SpaceX wiki page. Scraped datasets were later preprocessed and stored in separate CSV files.


Start
Request rocket launch data to SpaceX URL using SpaceX-API
Parse the data from html text
Request More data using Launch IDs
Construct "Pandas" DataFrames with Scraped Data
Request the Falcon9 Launch Wiki page using 'requests' module
Request rocket launch data
Parse the table from html text using "bs4"

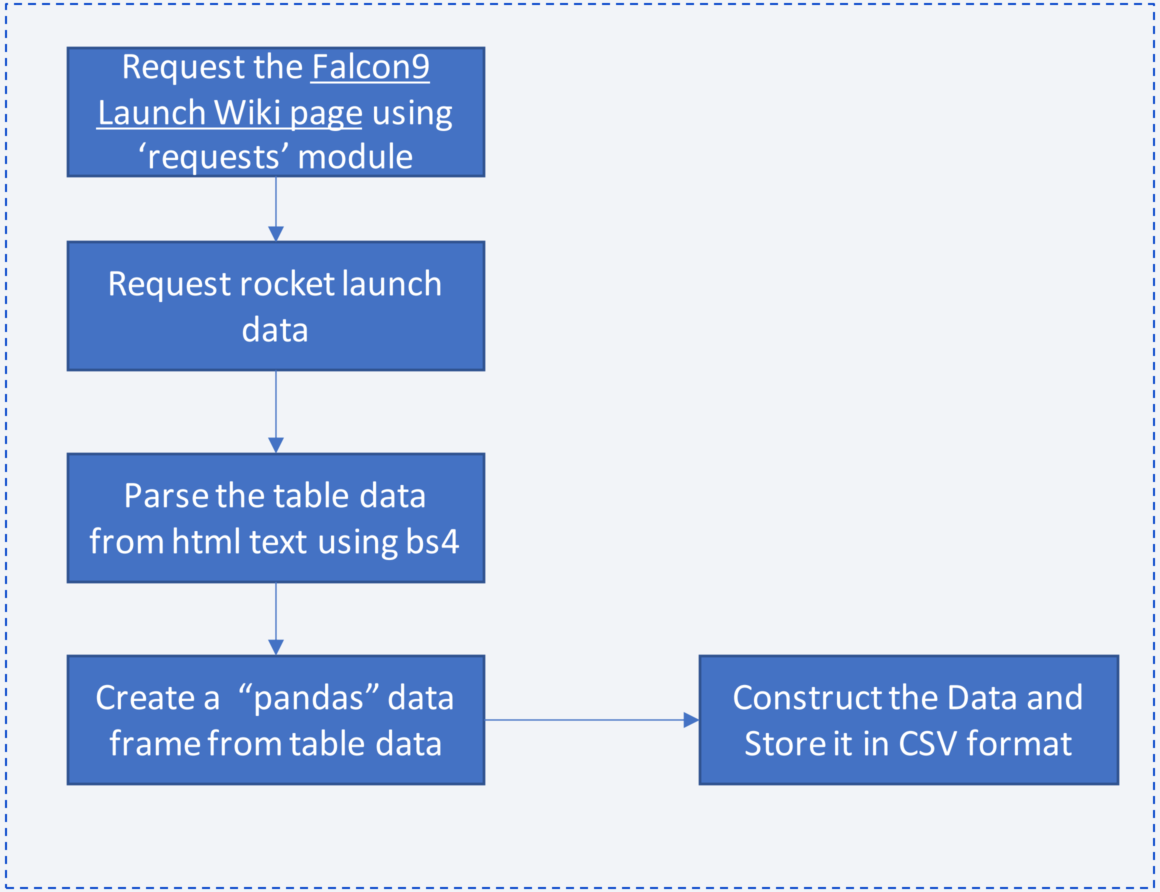
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook


Define Variables and Auxiliary Functions
Request launch data to SpaceX URL
Parse the data from html text
Request More data using Launch IDs
Preprocess and Construct the Data
Store the Data in CSV File
SpaceX-API

Data Collection - Scraping

- GitHub URL of the completed web scraping notebook



Request the Falcon9 Launch Wiki page using ‘requests’ module

Request rocket launch data

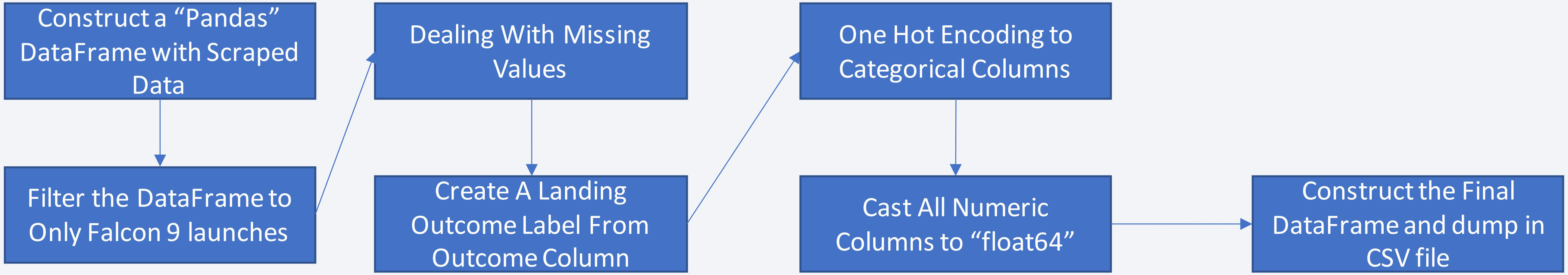
Parse the table data from html text using bs4

Create a “pandas” data frame from table data

Construct the Data and Store it in CSV format

Data Wrangling

- Summary - The **Falcon 1** launches data was filtered out the from scraped dataset. The null values inside of PayLoadMass were replaced by its mean value. A Landing Outcome label (“**landing_class**”) created from **Outcome** column. All categorical columns were converted using “One Hot Encode” and numerical columns were converted to float datatype.
- GitHub URL of completed data wrangling related notebook-1 , notebook-2, notebook-3


Construct a “Pandas” DataFrame with Scraped Data
Filter the DataFrame to Only Falcon 9 launches
Dealing With Missing Values
Create A Landing Outcome Label From Outcome Column
One Hot Encoding to Categorical Columns
Cast All Numeric Columns to “float64”
Construct the Final DataFrame and dump in CSV file

EDA with Data Visualization

For Explorative Data Analysis, following charts were used:

CatPlot :

- FlightNumber vs LaunchSite : To visualize the relationship between Flight Number and Launch Site
- Payload vs Launch Site : To visualize the relationship between Payload and Launch Site
- FlightNumber vs Orbit type : To visualize the relationship between Flight Number and Orbit type
- Payload vs Orbit type : To visualize the relationship between Payload and Orbit type

EDA with Data Visualization [Continued]

Bar Chart

- Orbit vs Success Rate : To visualize the relationship between success rate of each orbit type

Line chart

- Year vs Class : To visualize the launch success yearly trend

GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

SQL queries performed :

- To display the names of the unique launch sites in the space mission
- To display 5 records where launch sites begin with the string 'CCA’
- To display the total payload mass carried by boosters launched by NASA (CRS)
- To display average payload mass carried by booster version “F9 v1.1”
- To list the date when the first successful landing outcome in ground pad was achieved
- To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL [Continued]

SQL queries performed [Continued] :

- To list the total number of successful and failure mission outcomes
- To list the names of the booster versions which have carried the maximum payload mass. Use a subquery
- To list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- To 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

GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

Map objects which are created and added to the folium map are given below :

- **Markers** : Added to mark a specific area with a text label on a specific coordinate
- **Circles** : Added to highlight circle areas with a text label on a specific coordinate
- **MarkerCluster** : Marker clusters were used to simplify the containing many markers having the same coordinates.
- **MousePosition** : Used to get coordinate for a mouse over a point on the map (proximities). It helps to find the coordinates easily of any points of interests while exploring the map
- **PolyLine** : It draws polyline overlays on a map. It was used to denote the distance between a launch site and its proximities. (such as Railway station, city etc.)

Jupyter Notebook link to view the Lab Codes with Folium Map

GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

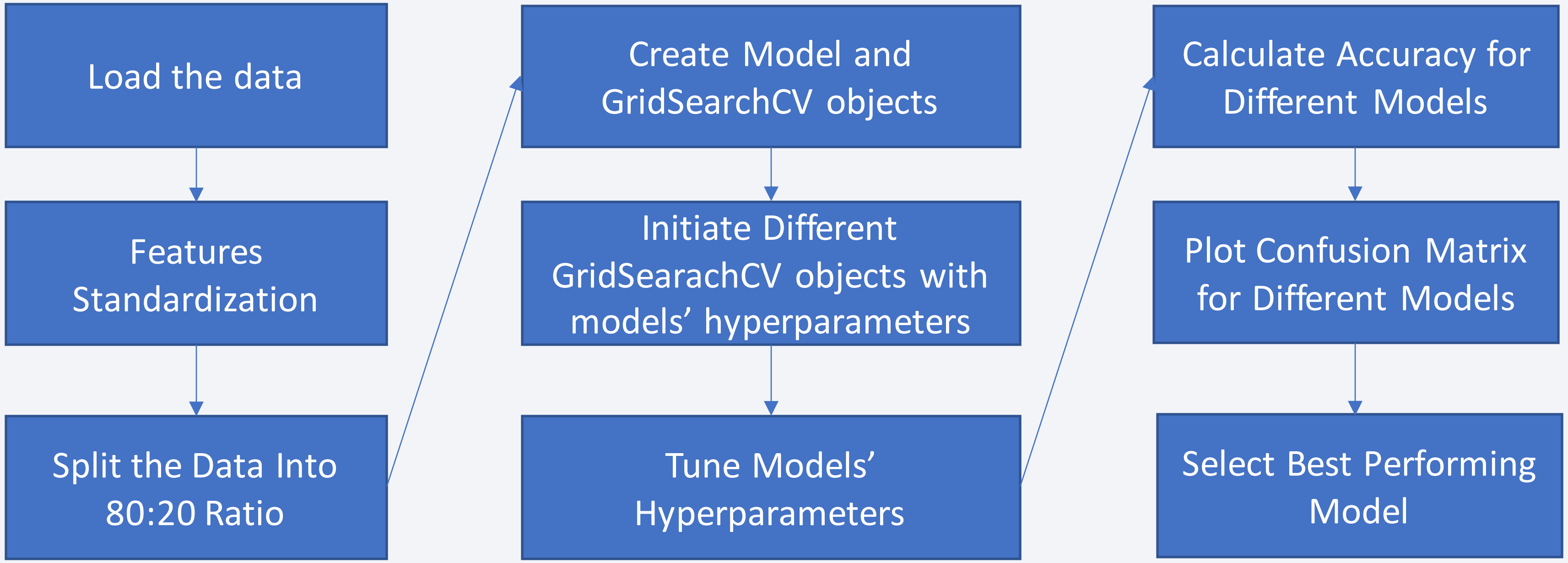
Plots/graphs and interactions which are added to the dashboard -

- **Drop-down Input Component** : To select different launch sites
- **Pie-chart** : To visualize total success launches ratio by site
- **Range Slider** : To select range of payload mass (kg)
- **Scatter Plot** : To visualize the correlation between payload and success launch for all sites

GitHub URL of the completed Plotly Dash lab

Predictive Analysis (Classification)

The Following chart illustrates the whole predictive analysis workflow: from loading the data to selecting best model for classification


Load the data
Features Standardization
Split the Data Into 80:20 Ratio
Create Model and GridSearchCV objects
Initiate Different GridSearachCV objects with models' hyperparameters
Tune Models' Hyperparameters
Calculate Accuracy for Different Models
Plot Confusion Matrix for Different Models
Select Best Performing Model

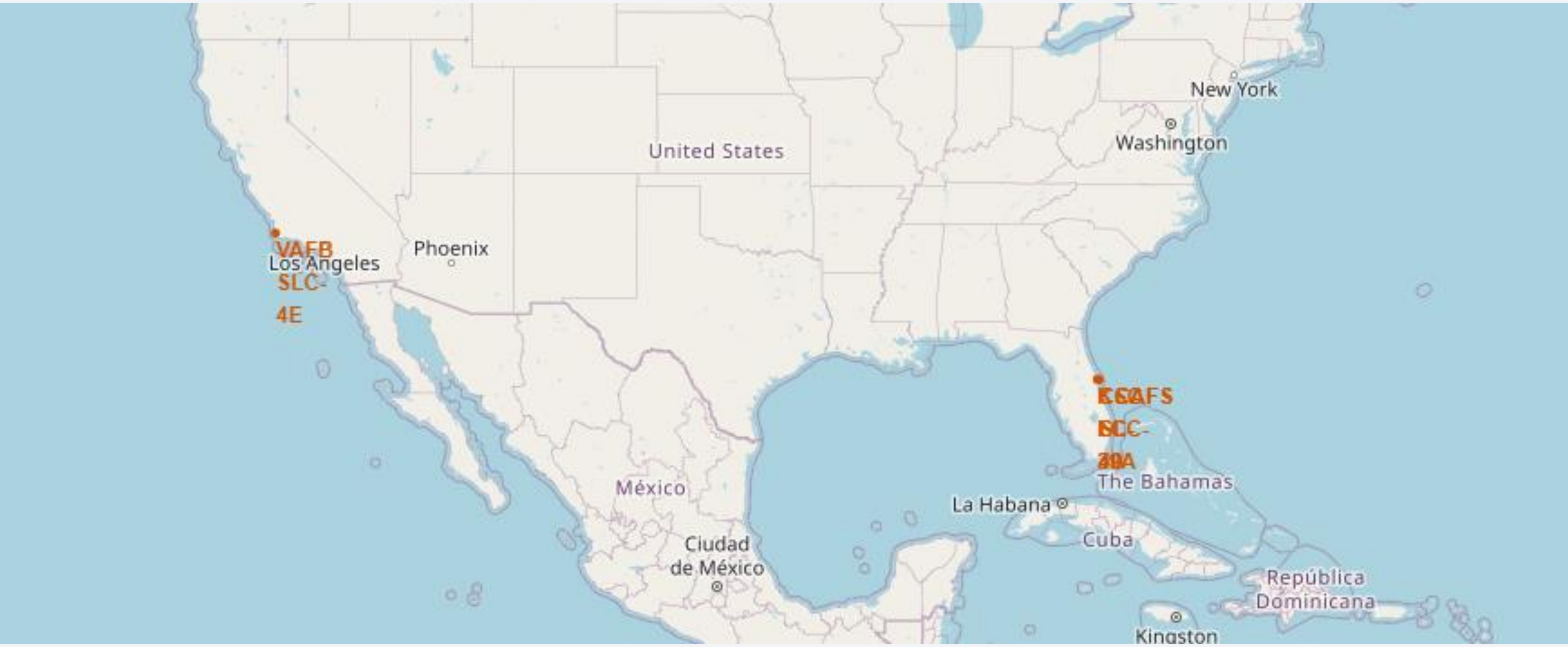
- GitHub URL of the completed predictive analysis lab

Results

- **Exploratory data analysis results -**
  - Payload greater than 7500 kg had a higher chance of successful landing
  - Among 11 orbit types , ES-L1, GEO, HEO, SSO were 100% successful with less than 6000 kg payload
  - From 2010 to 2019, 61 missions had successful landing whereas 10 missions had failure landing, leading a success rate of 67.78%
  - Among the four Launch sites “KSC LC-39A” had the highest success rate (41.7%) whereas, “VAFB SLC-4E” had the lowest success rate (12.5%)
  - In 0-10000 kg payload range “FT” booster version has the largest success rate

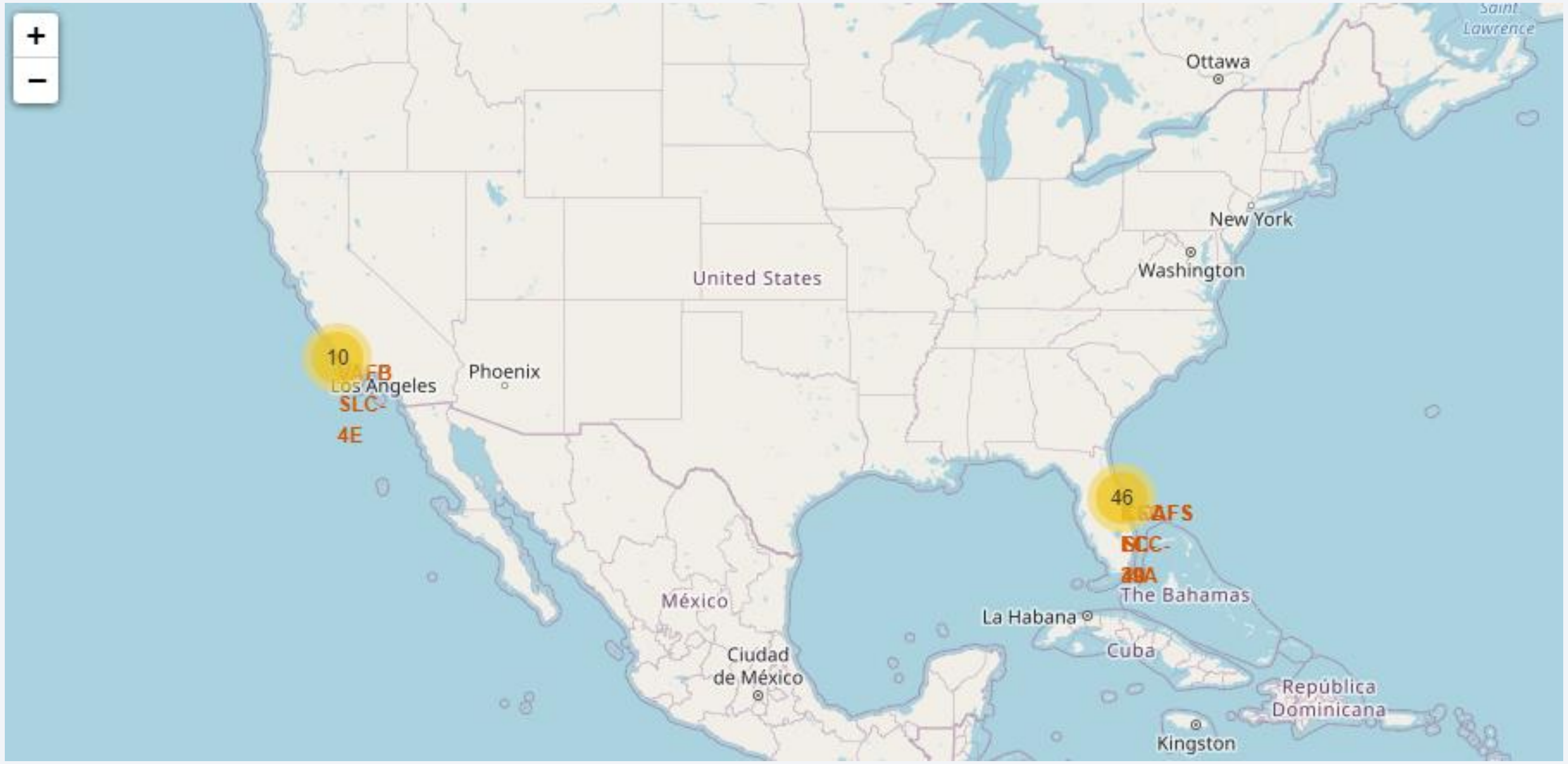
Results [Continued]

- Interactive analytics Folium Lab



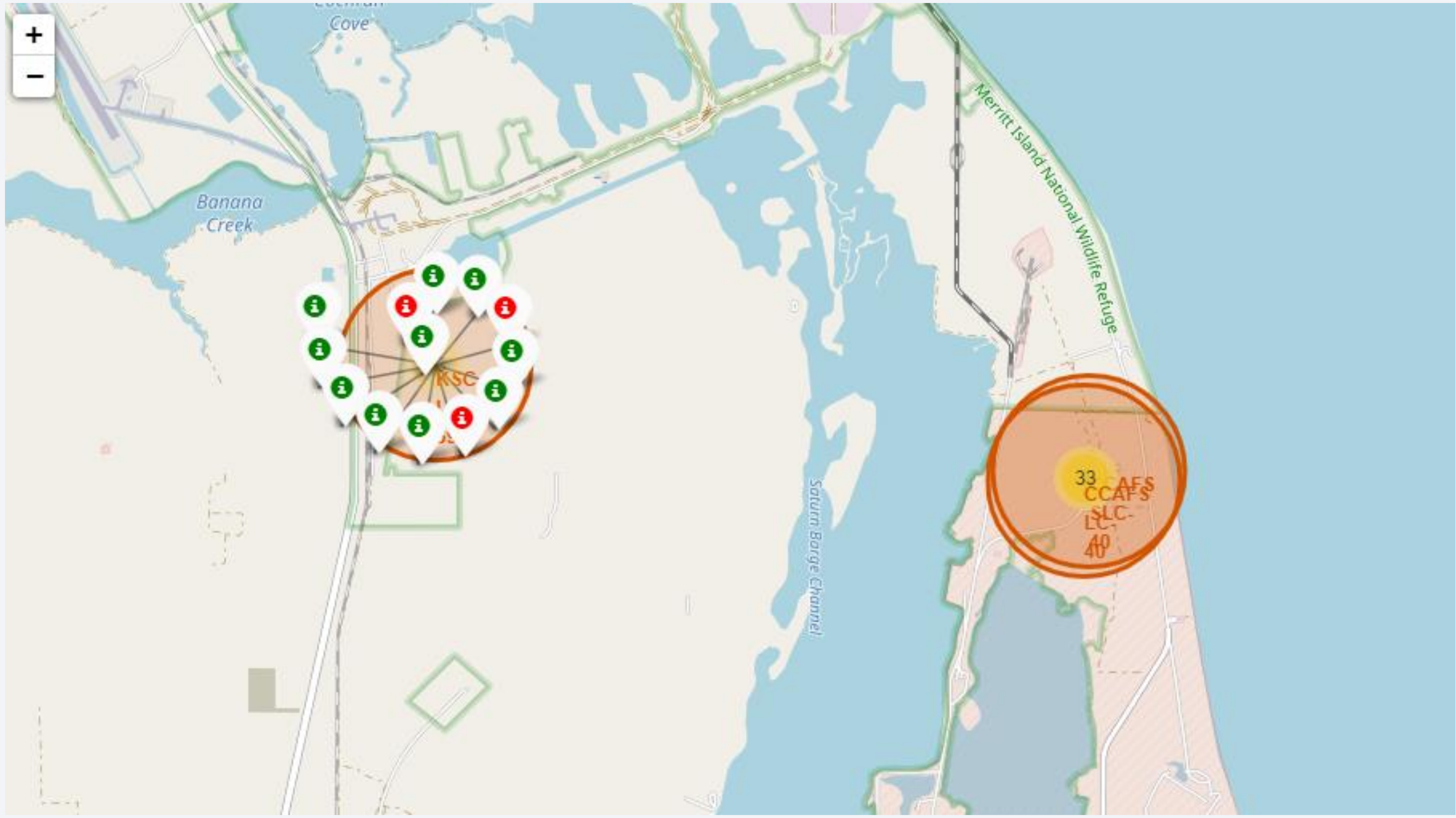
Results [Continued]

- Interactive analytics Folium Lab



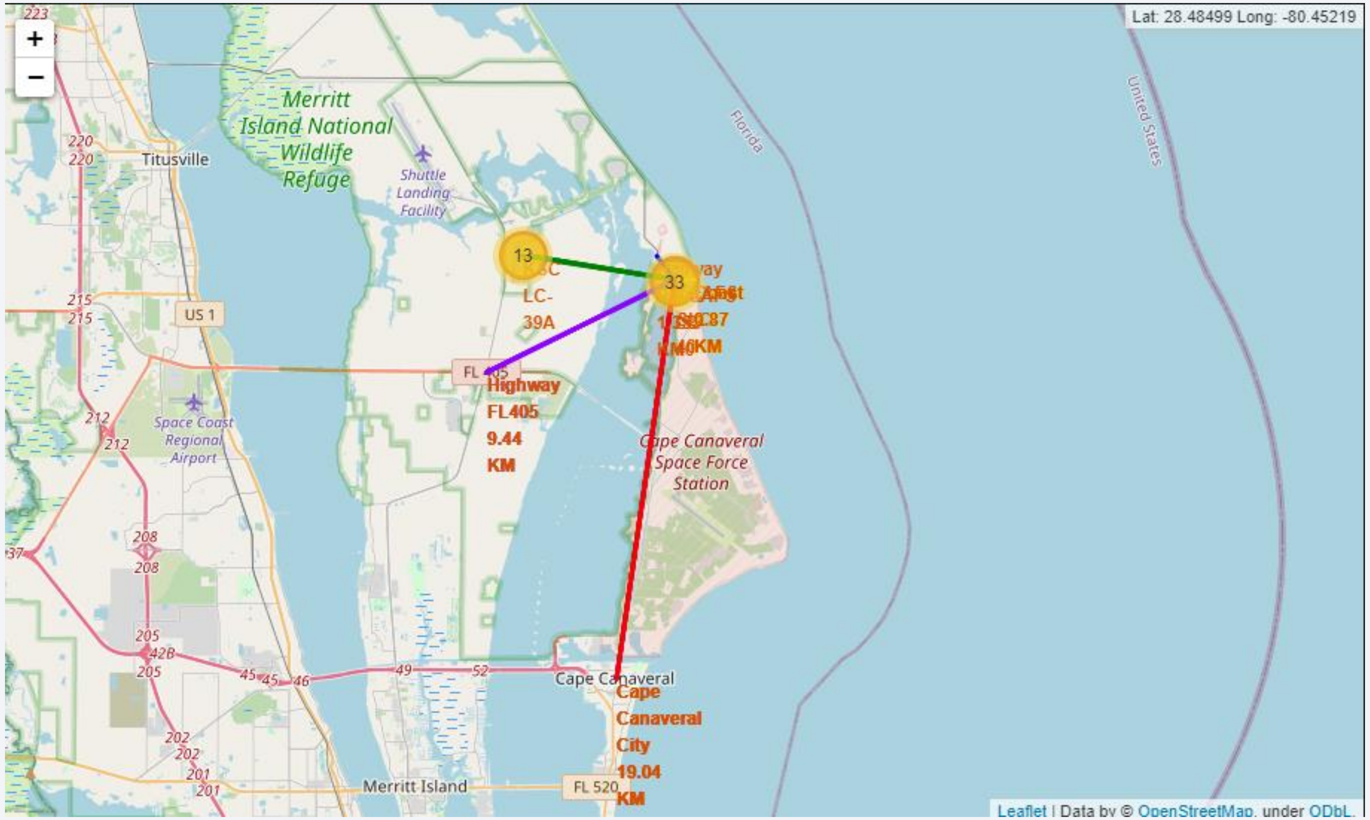
Results [Continued]

- Interactive analytics Folium Lab



Results [Continued]

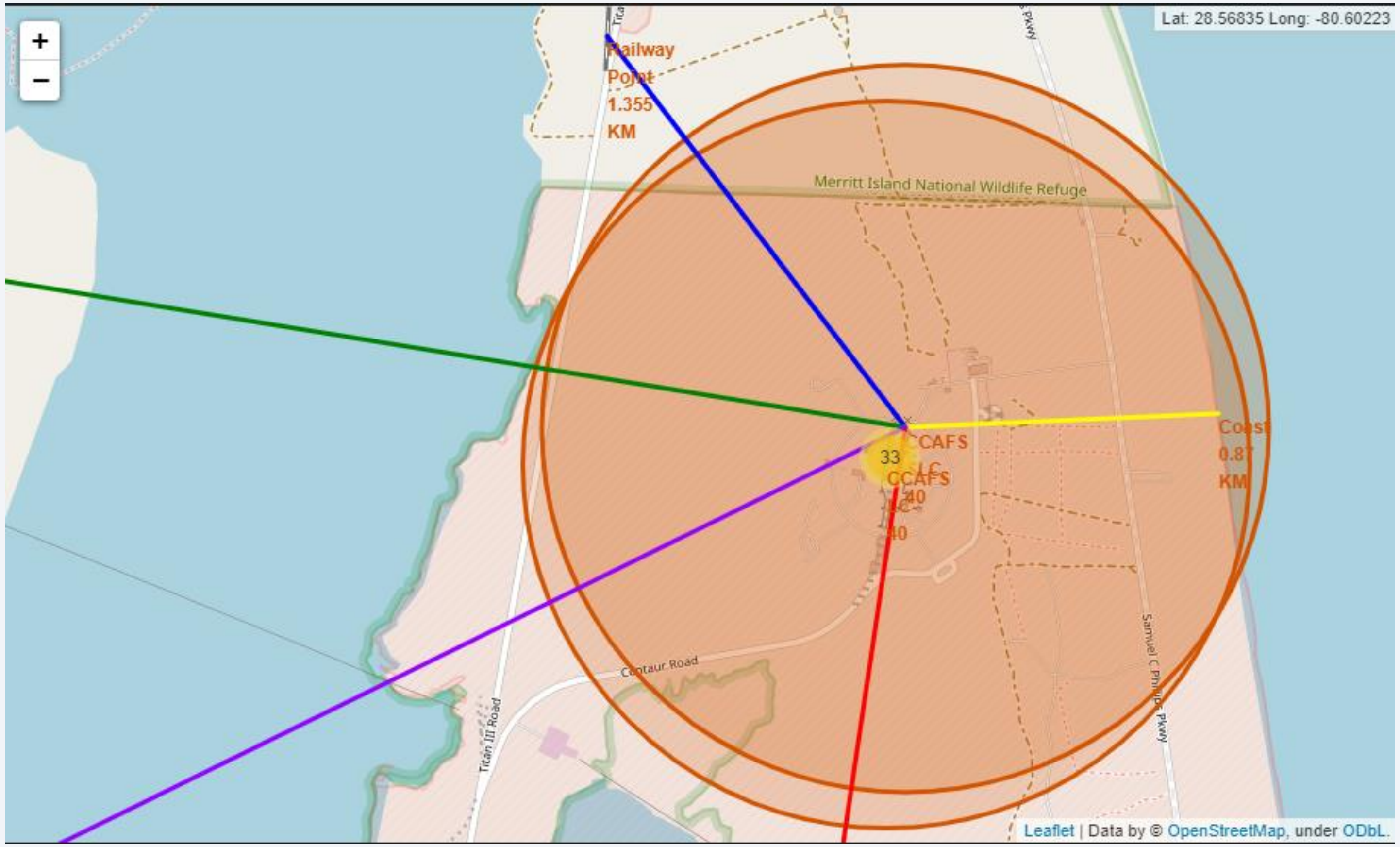
- Interactive analytics Folium Lab



Lat: 28.48499 Long: -80.45219
+
−
Merritt Island National Wildlife Refuge
Titusville
Shuttle Landing Facility
Florida
United States
13
33
LC-39A
US 1
FL 405
Highway FL405 9.44 KM
Space Coast Regional Airport
Cape Canaveral Space Force Station
Cape Canaveral
Cape Canaveral City 19.04 KM
Merritt Island
FL 520
Leaflet | Data by © OpenStreetMap, under ODbL.

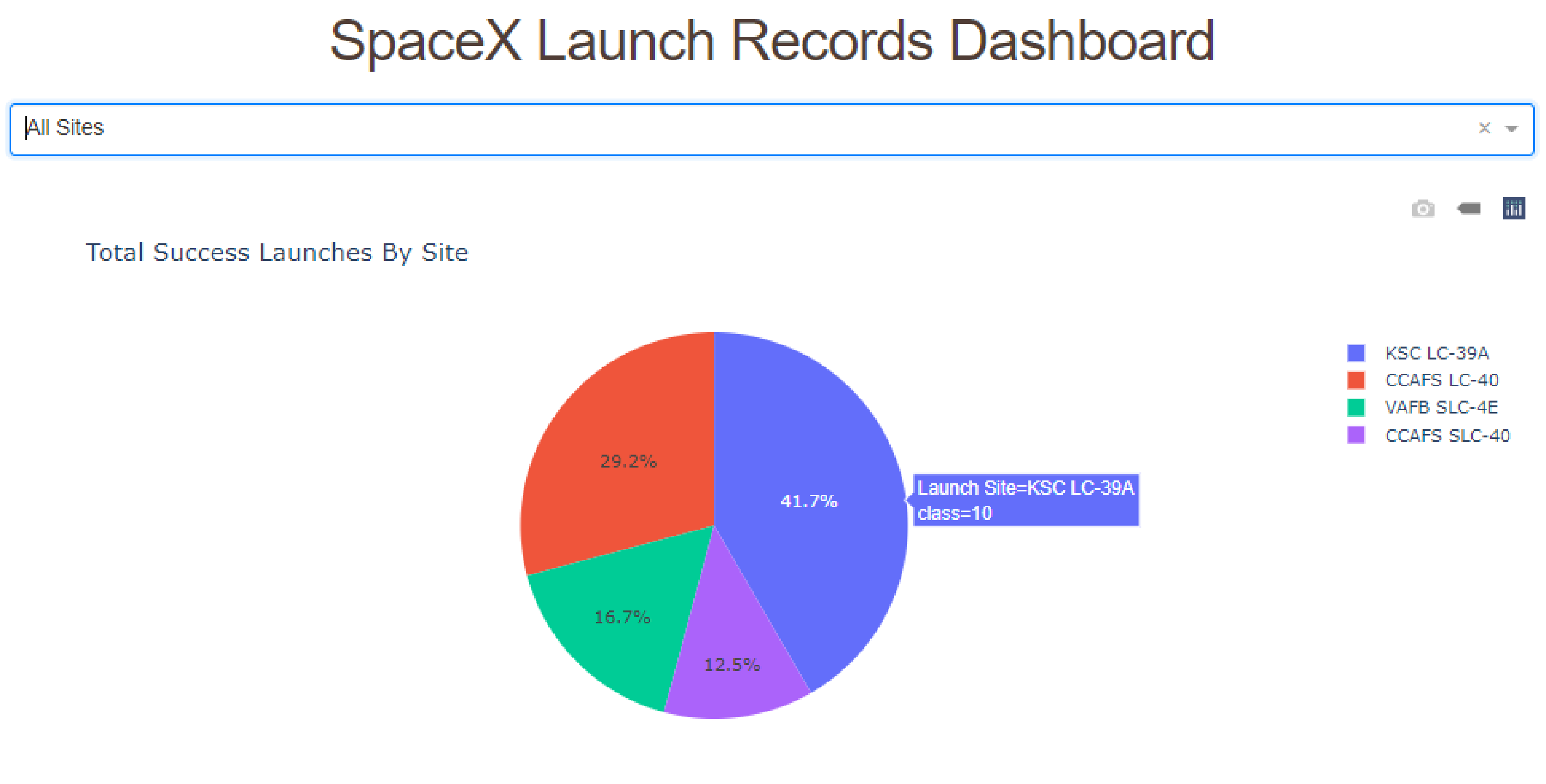
Results [Continued]

- Interactive analytics Folium Lab


+
−
Lat: 28.56835 Long: -80.60223
Railway Point 1.355 KM
Merritt Island National Wildlife Refuge
CCAFS LC-40
33
CCAFS LC-40
Coast 0.87 KM
Centaur Road
Titan III Road
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

Results [Continued]

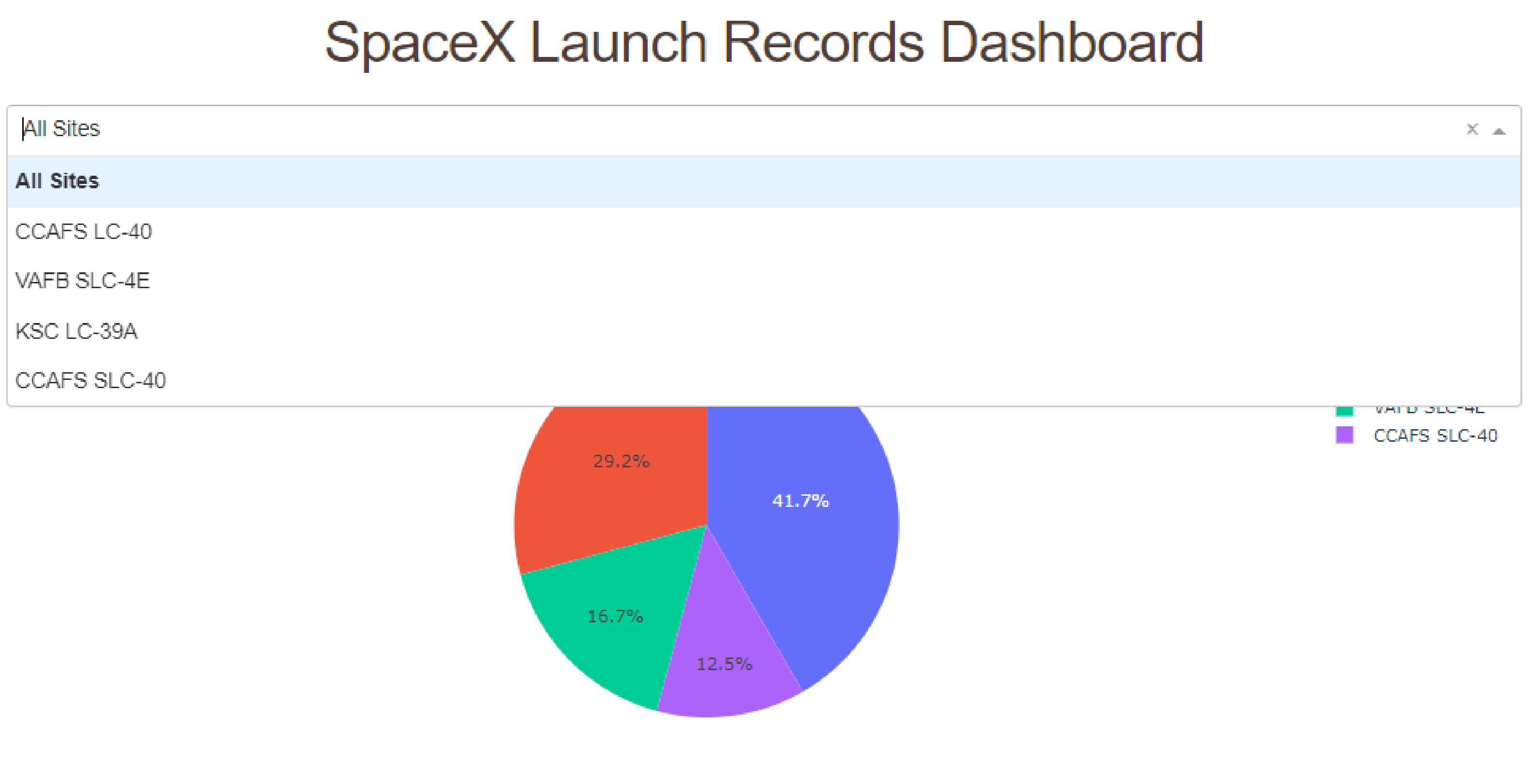
- Interactive analytics dashboard



SpaceX Launch Records Dashboard

Results [Continued]

- Interactive analytics dashboard

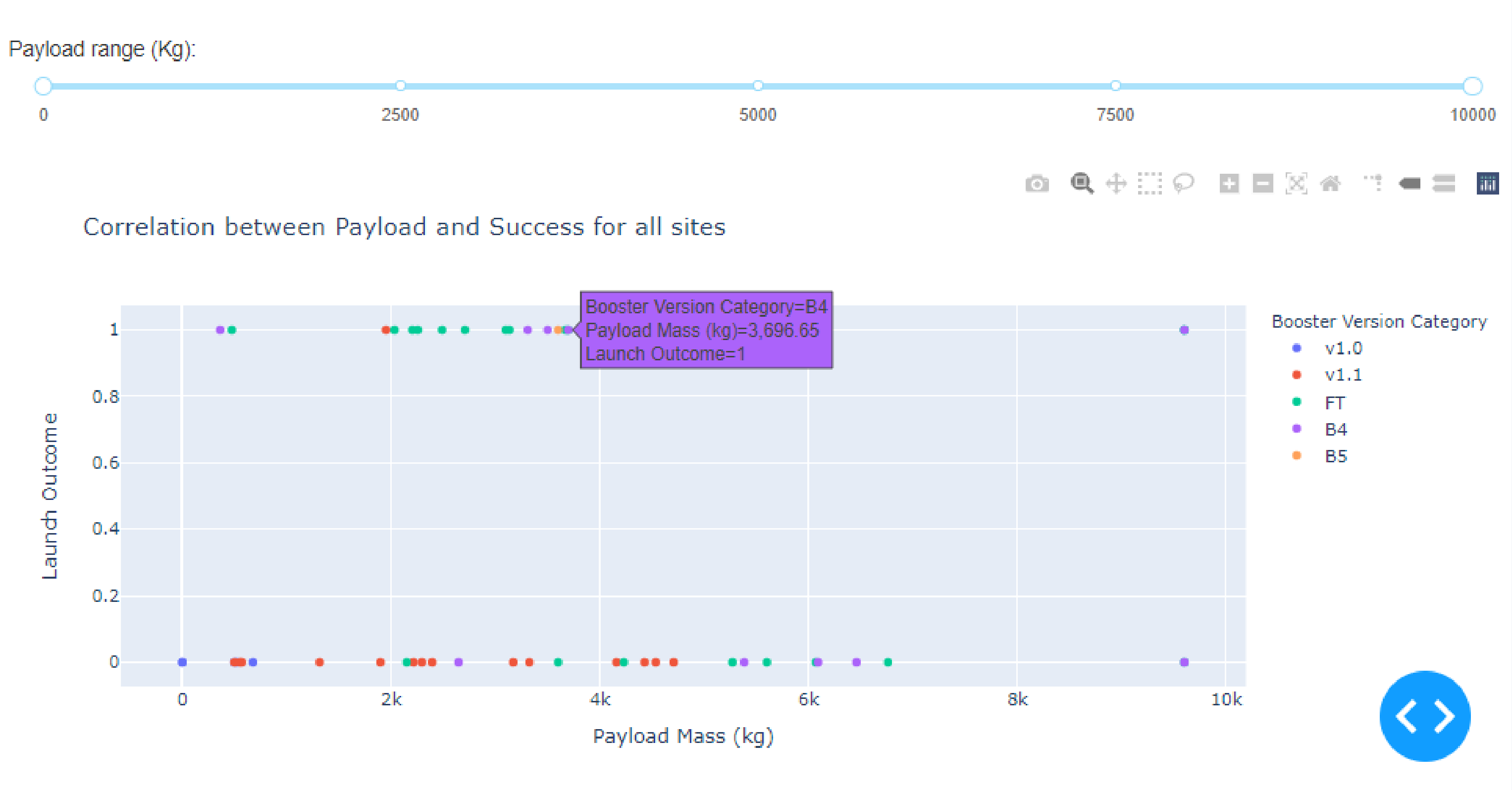


SpaceX Launch Records Dashboard

All Sites

Results [Continued]

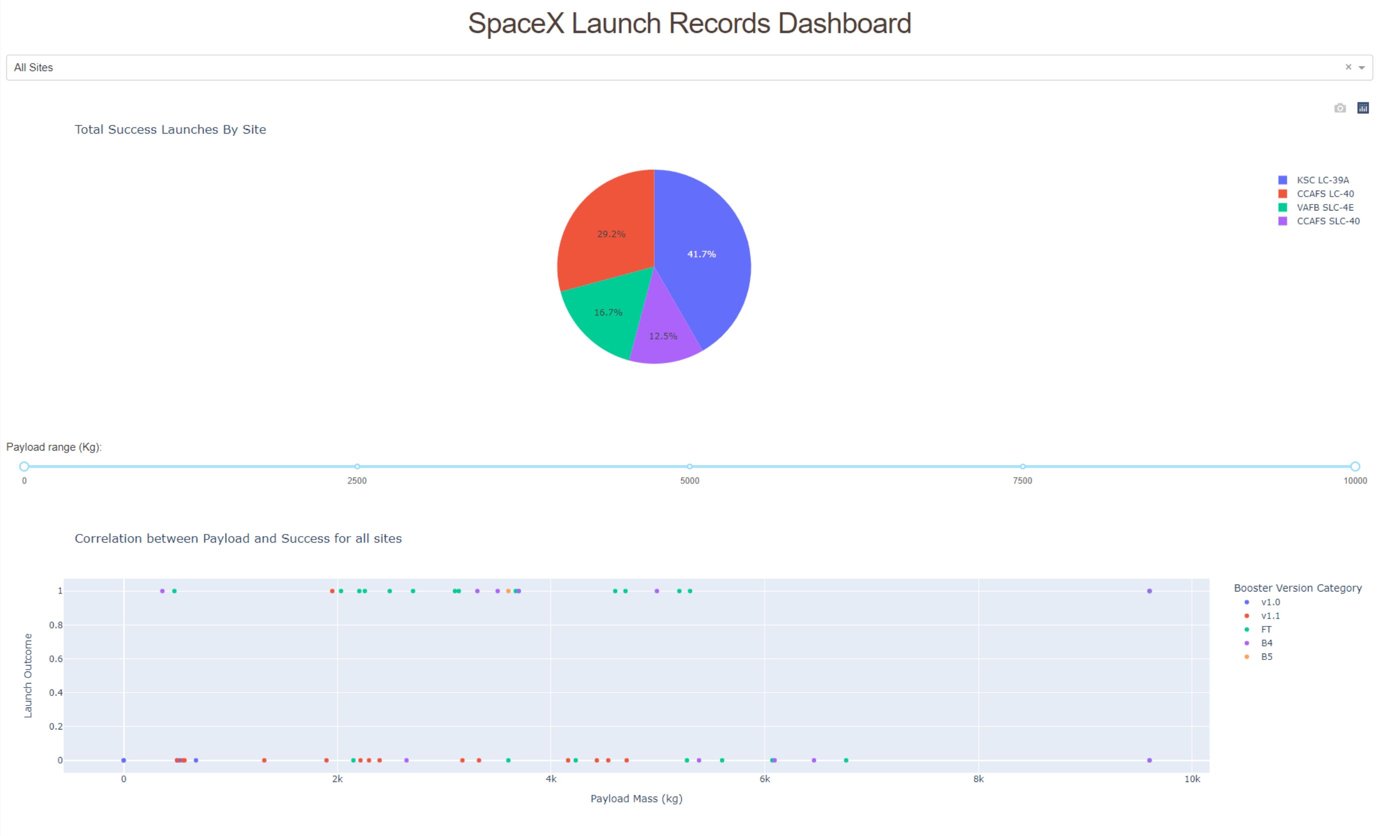
- Interactive analytics dashboard



Payload range (Kg):

Results [Continued]

- Interactive analytics dashboard



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Results [Continued]

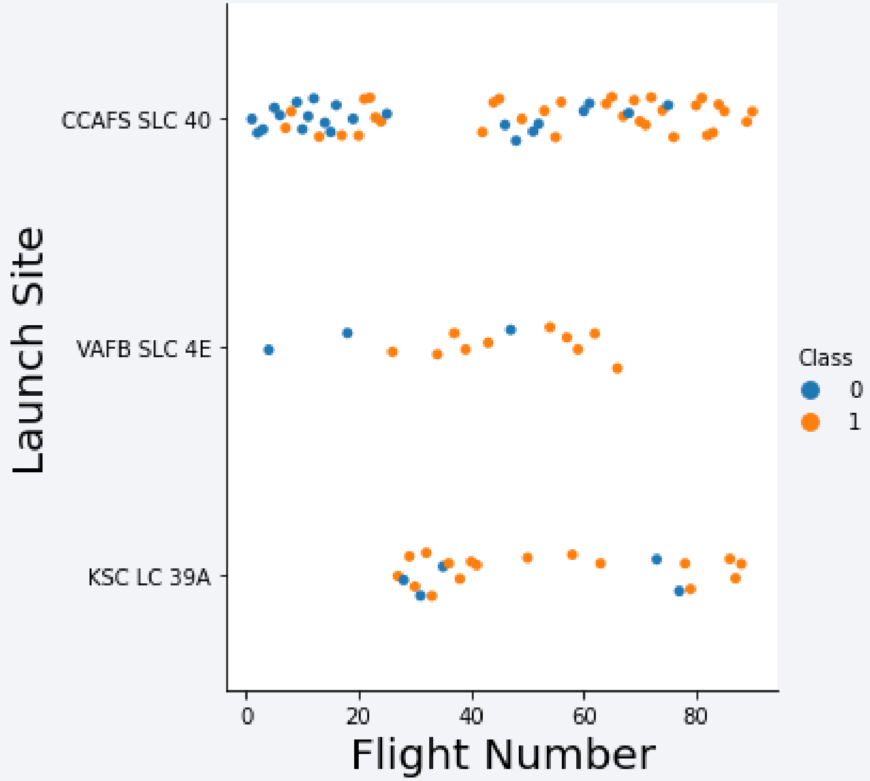
- Predictive analysis results
  - Performed a predictive analysis on the dataset using Logistic Regression, SVM, Decision Tree and KNN Classifier
  - Logistic Regression and SVM performed well in comparison with other models, with the highest accuracy of 83.3%

Section 2

Insights drawn from EDA

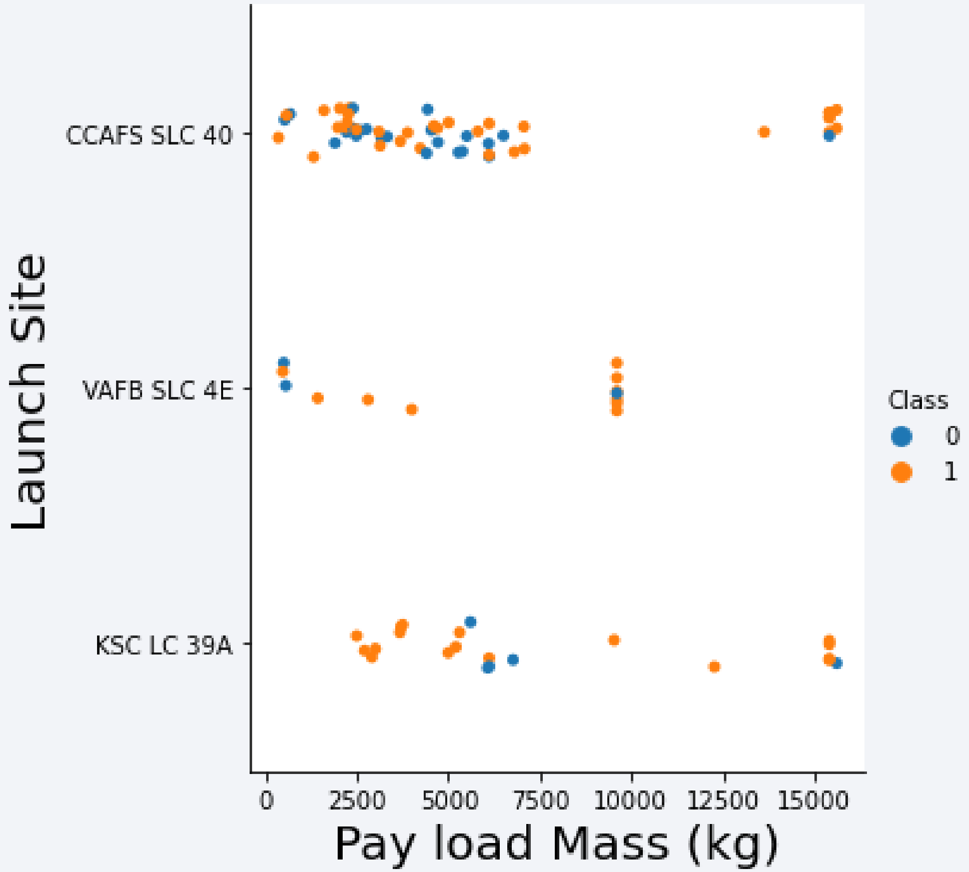
Flight Number vs. Launch Site

- Success rate was gradually increasing for every launch site
- After flight count reached 40, success rate increased dramatically



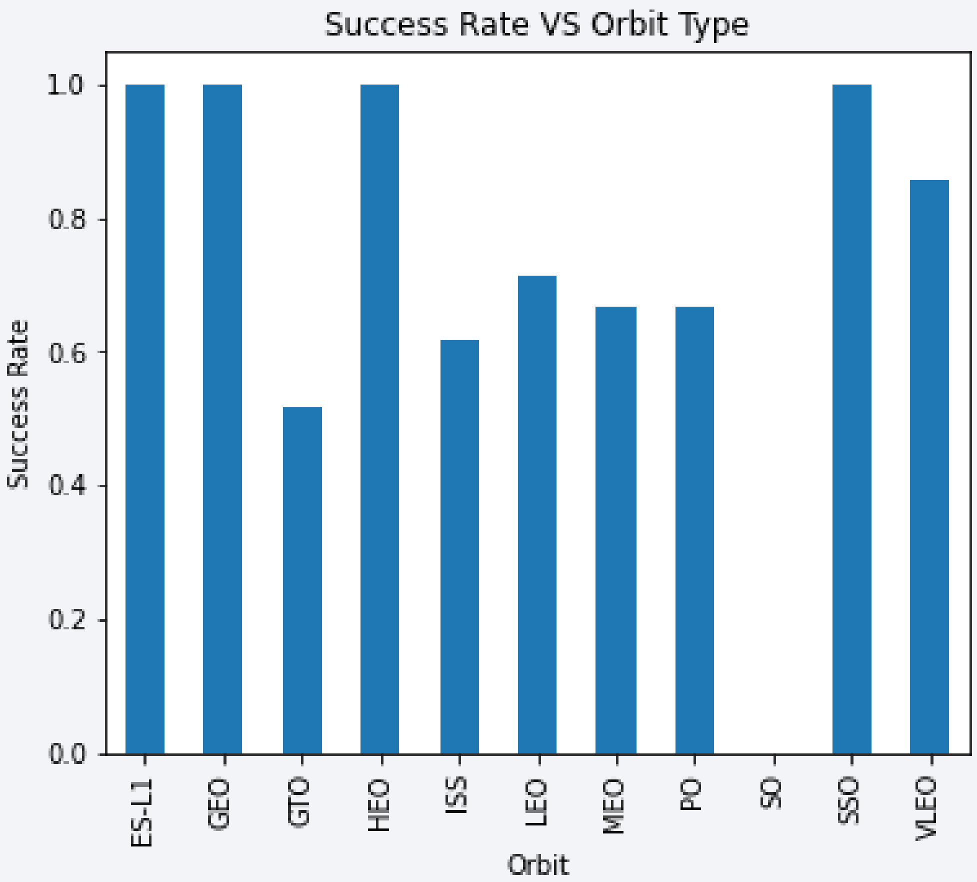
Payload vs. Launch Site

- Payload greater than 7500 kg had a higher chance of successful landing (in all cases)
- KSC LC 39A had the highest success rate among the 4 launch sites



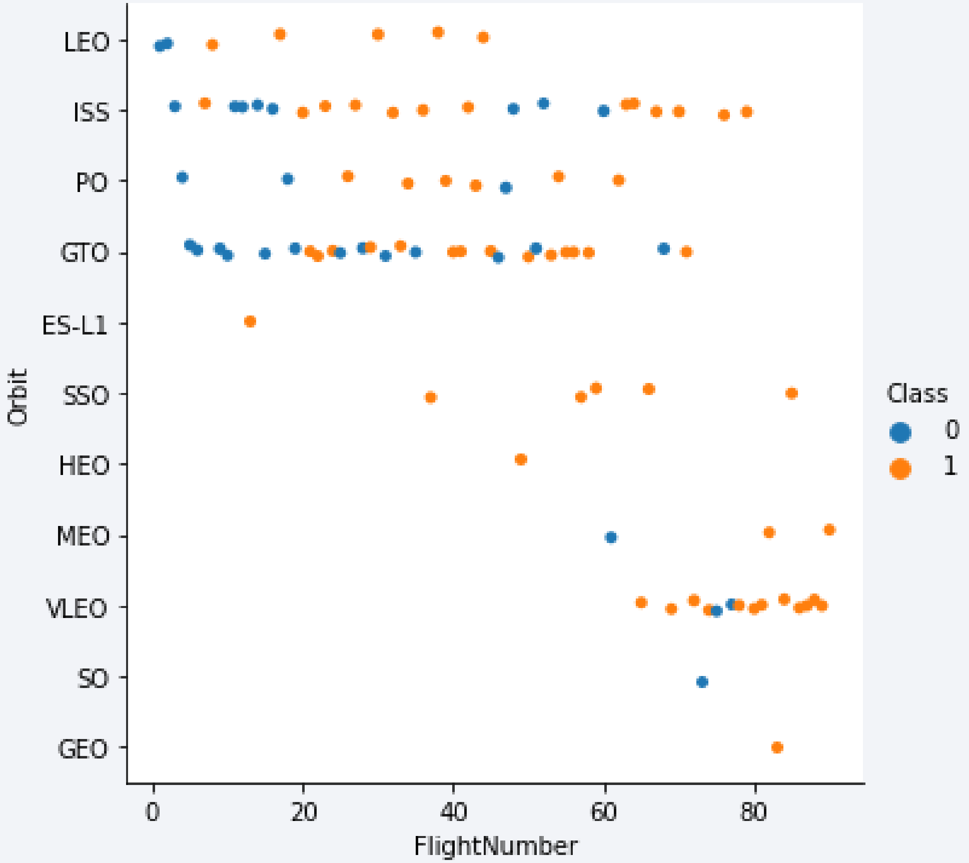
Success Rate vs. Orbit Type

- “SO” orbit had no success
- Among 11 orbit types , 4 types were 100% successful(ES-L1, GEO, HEO, SSO)
- “GTO” had approximately 50% success rate whereas the rest of the orbit types had more than 50% success rate



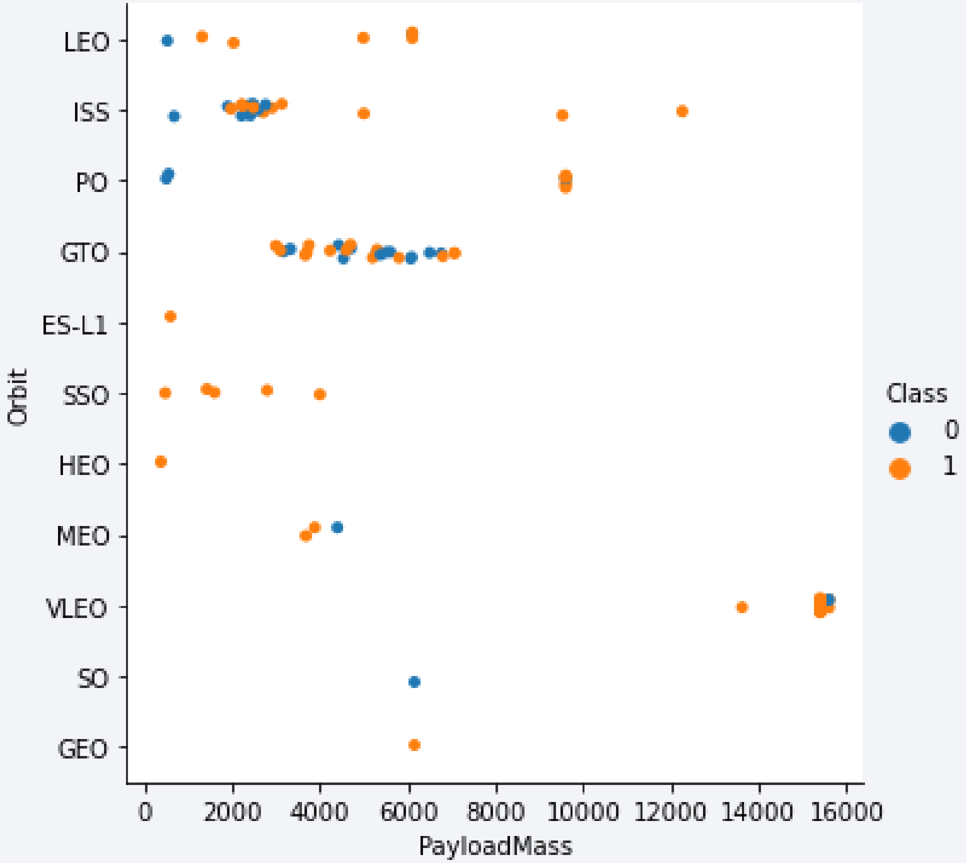
Flight Number vs. Orbit Type

- LEO , ISS, PO ,GTO had more flight counts than other orbit types
- SO,GEO,HEO orbit type had only 1 flight which was a failure, success and success respectively
- ES-L1, SSO, GEO, HEO had 100% success landing in terms of flights



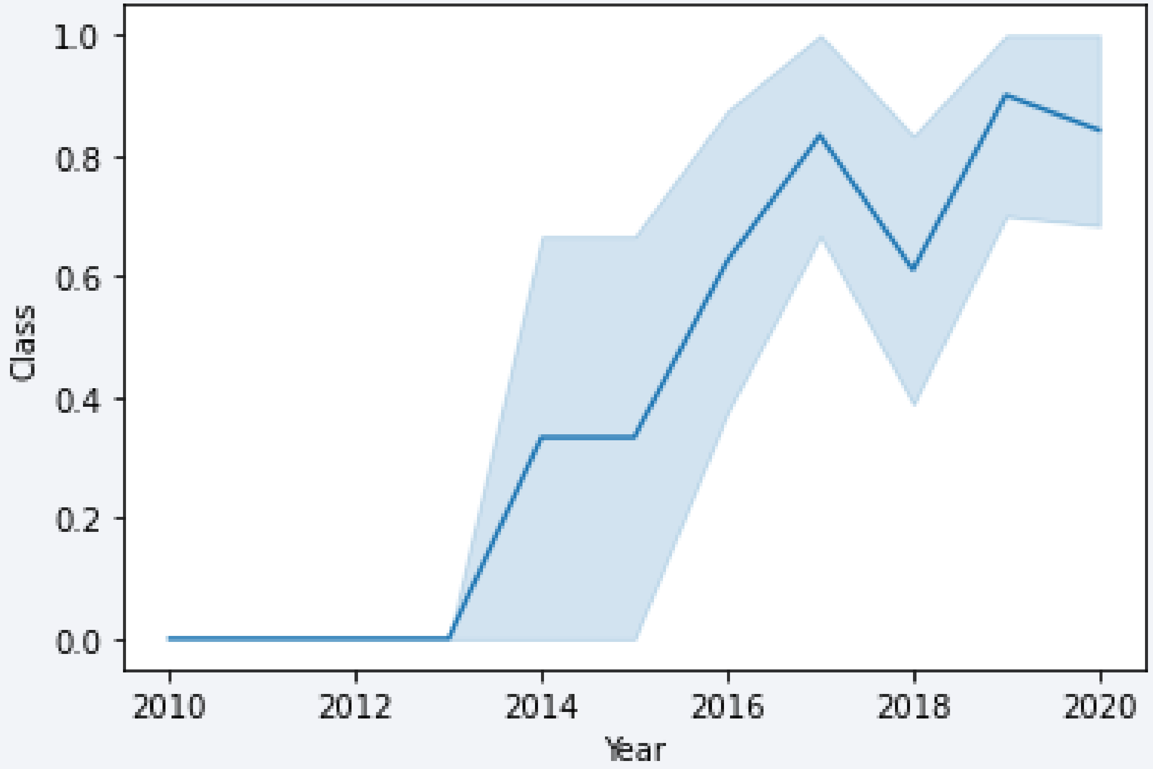
Payload vs. Orbit Type

- Only ISS, PO and VLEO orbits carried more than 8000 kg payload
- ES-L1, SSO, HEO,GEO had less than 6000 kg payload in their orbit with 100% success



Launch Success Yearly Trend

- Before 2013, there was no success
- After 2013 success rate increased dramatically
- Launch success peaked in 2019



All Launch Site Names

- There are total of 4 launch sites, the list is given below -

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- The below list represents 5 records where launch sites begin with `CCA`, all of them refers to the same launch site

Launch Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

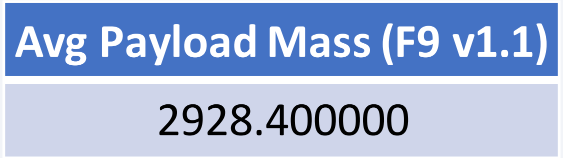
Total Payload Mass

- Total payload carried by boosters from NASA

Total Payload Mass
107010

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1. It is roughly 3000 kg



Avg Payload Mass (F9 v1.1)
2928.400000

First Successful Ground Landing Date

- The date of the first successful landing outcome on the ground pad. It is the end of 2015

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The below list represents the 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

- The total number of successful and failure mission outcome. The successful outcomes is 6 times of the failure outcomes

successful_outcomes	failure_outcomes
61	10

Boosters Carried Maximum Payload

- The below list represents the names of the booster which have carried the maximum payload mass. It’s a total of 12

Booster Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

- The below list represents the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015. It happened in the same launch site

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

- The below list represents the ranking of 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

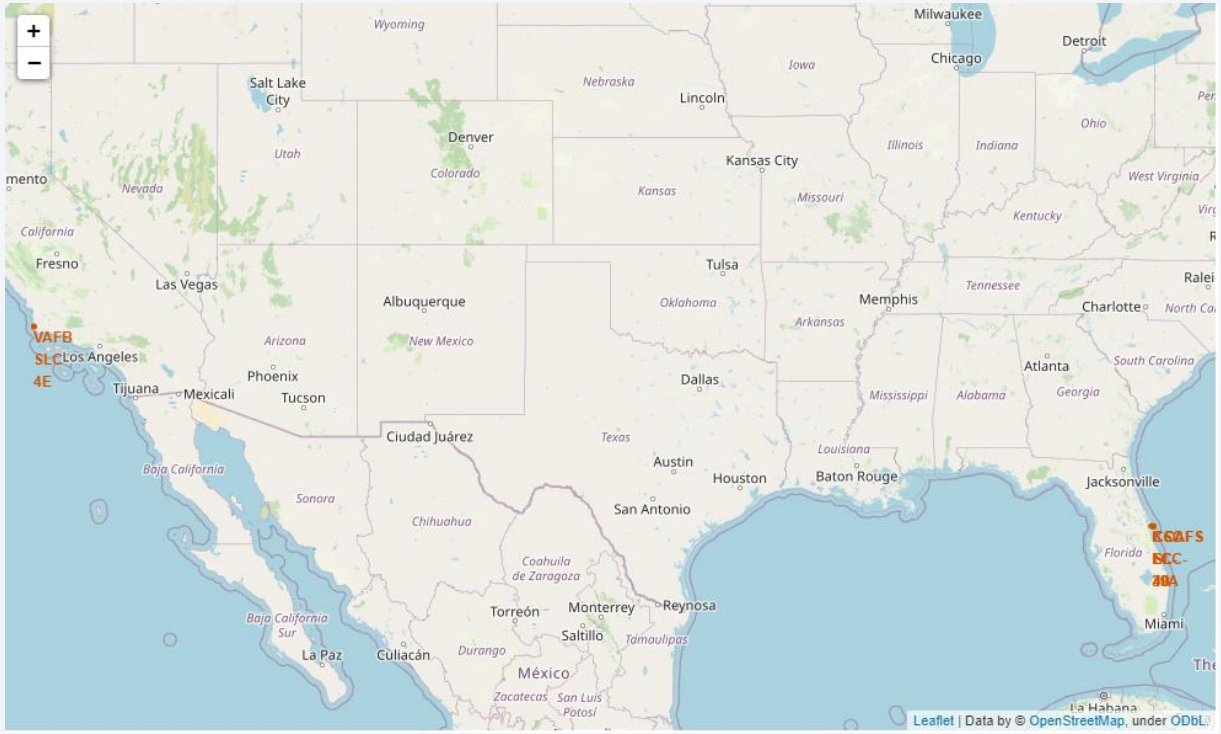
Landing Outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

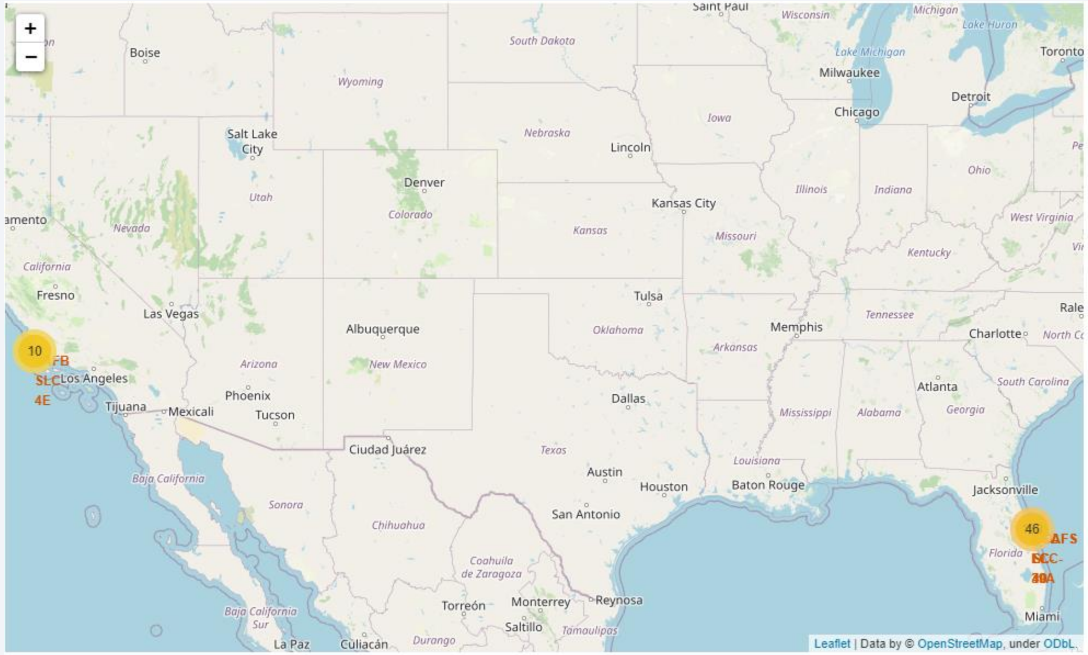
All Launch Sites Of SpaceX On The Map

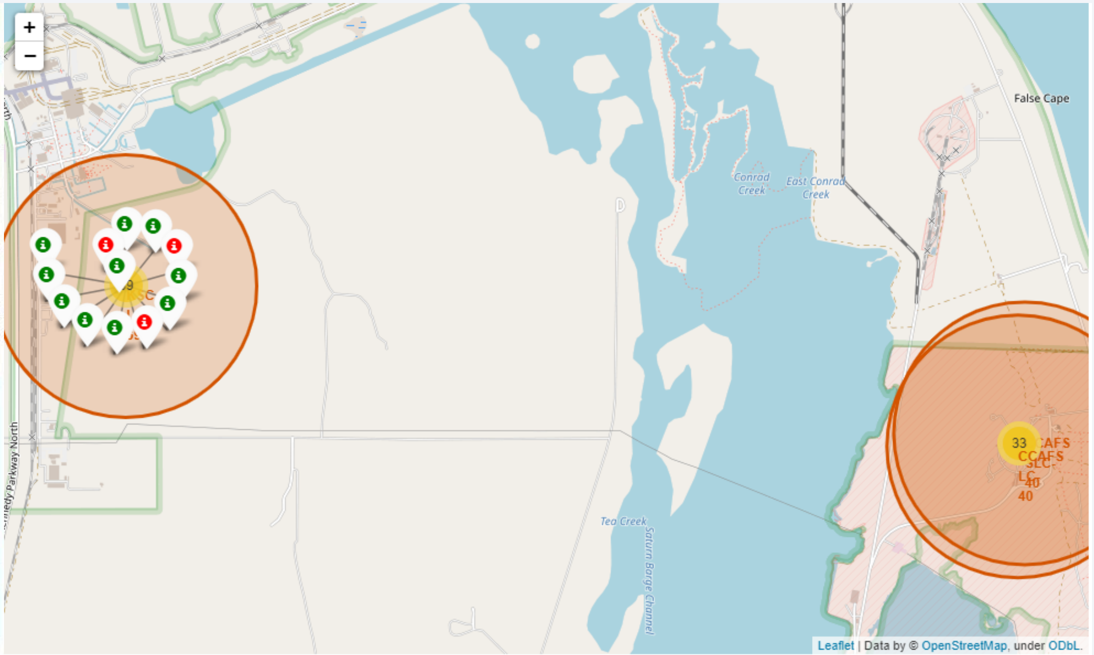
- SpaceX has 4 launch facilities, one is near California, the other three facilities are near Florida and South Texas. All facilities are near the ocean



Success/Failed Launches For Each Site On The Map

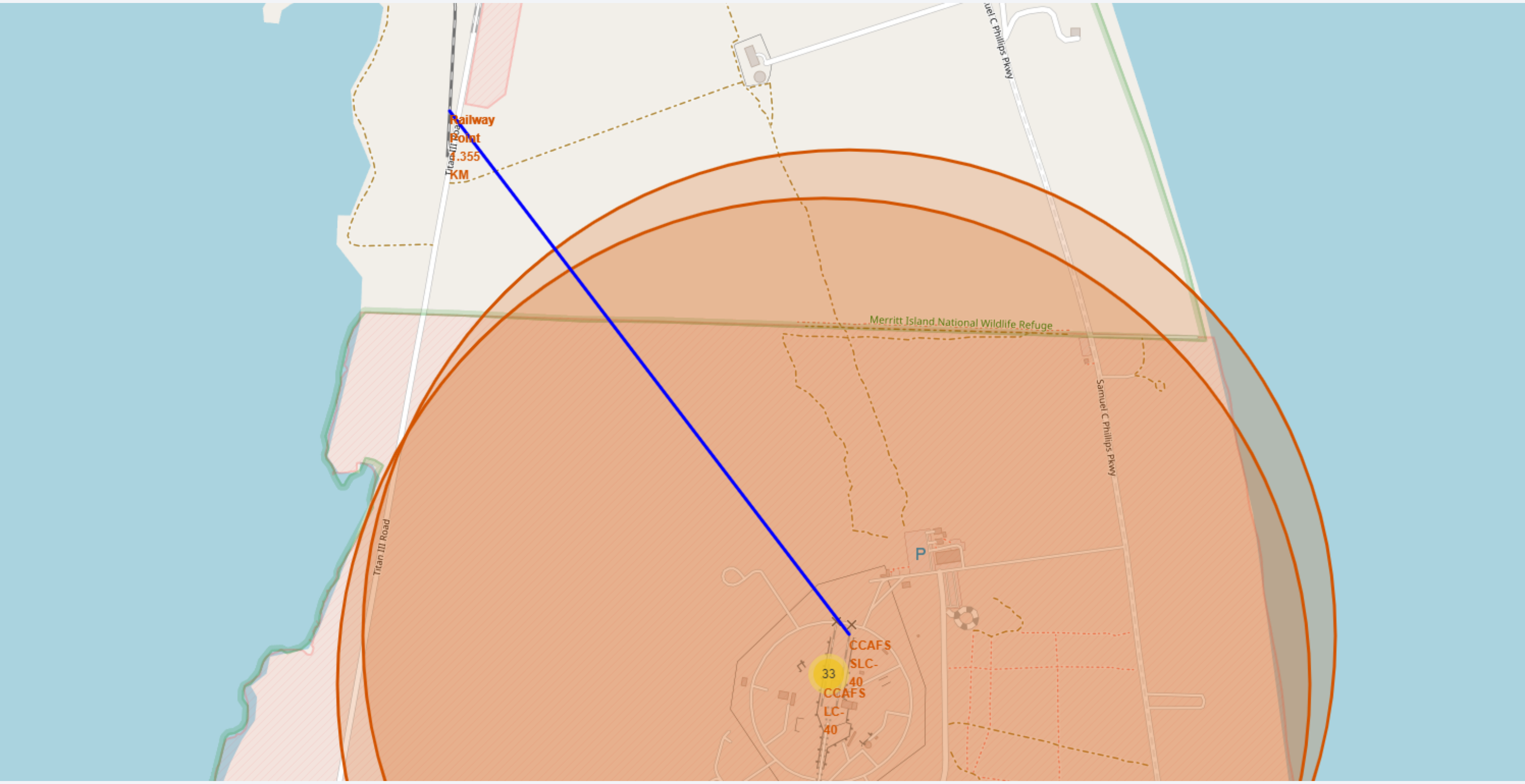
- Marker clusters is used to simplify the map containing many markers having the same coordinate (check 3 locations near Florida). Successful launches are marked using a green marker and failed launches are marked using a red marker





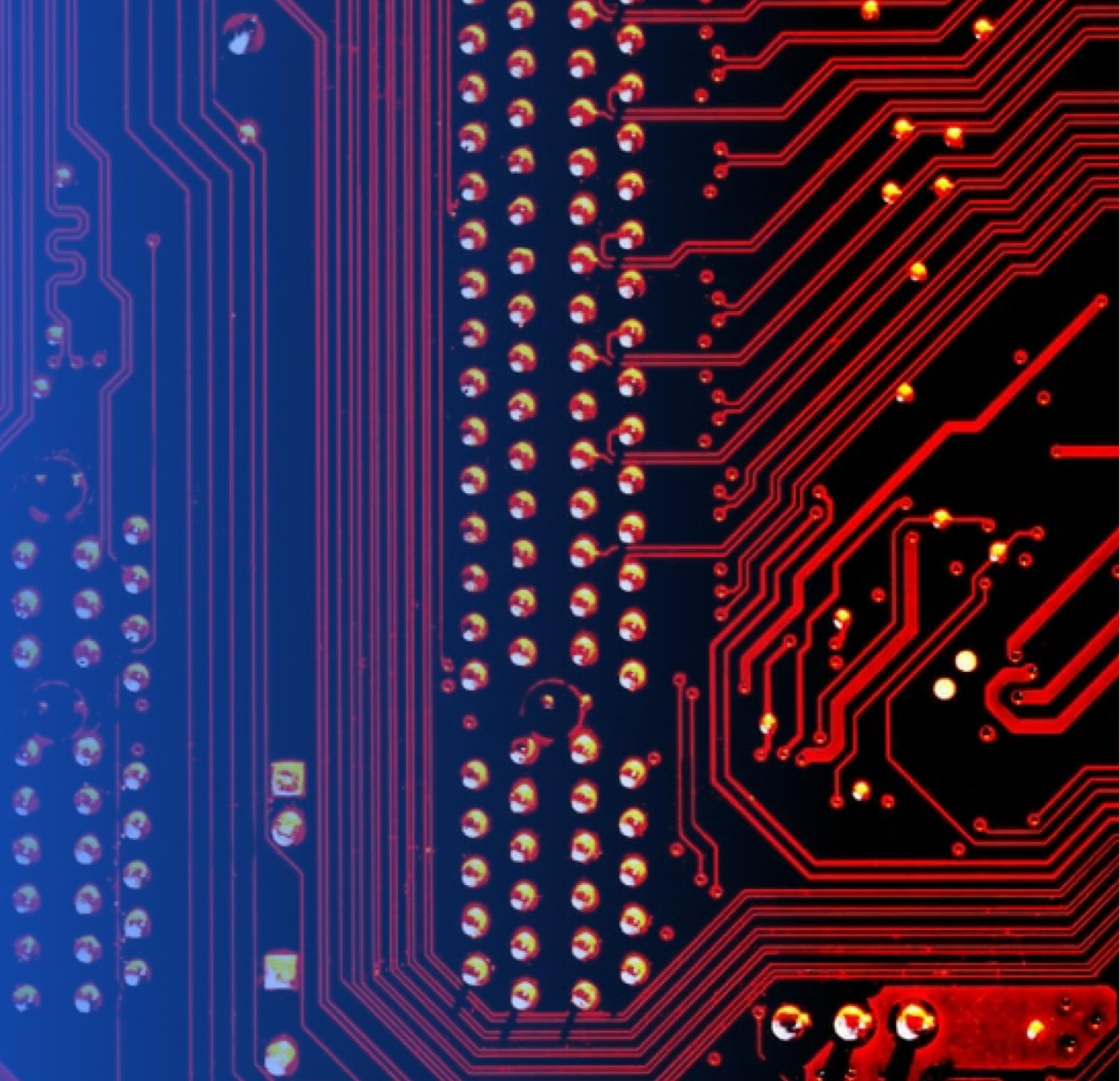
Distance Between A Launch Site To Its Proximities

- A blue distance line denotes the distance between a launch site (CCAFS SLC-40) to a nearby railway station, (denoted as “Railway Point 1.355 KM”)



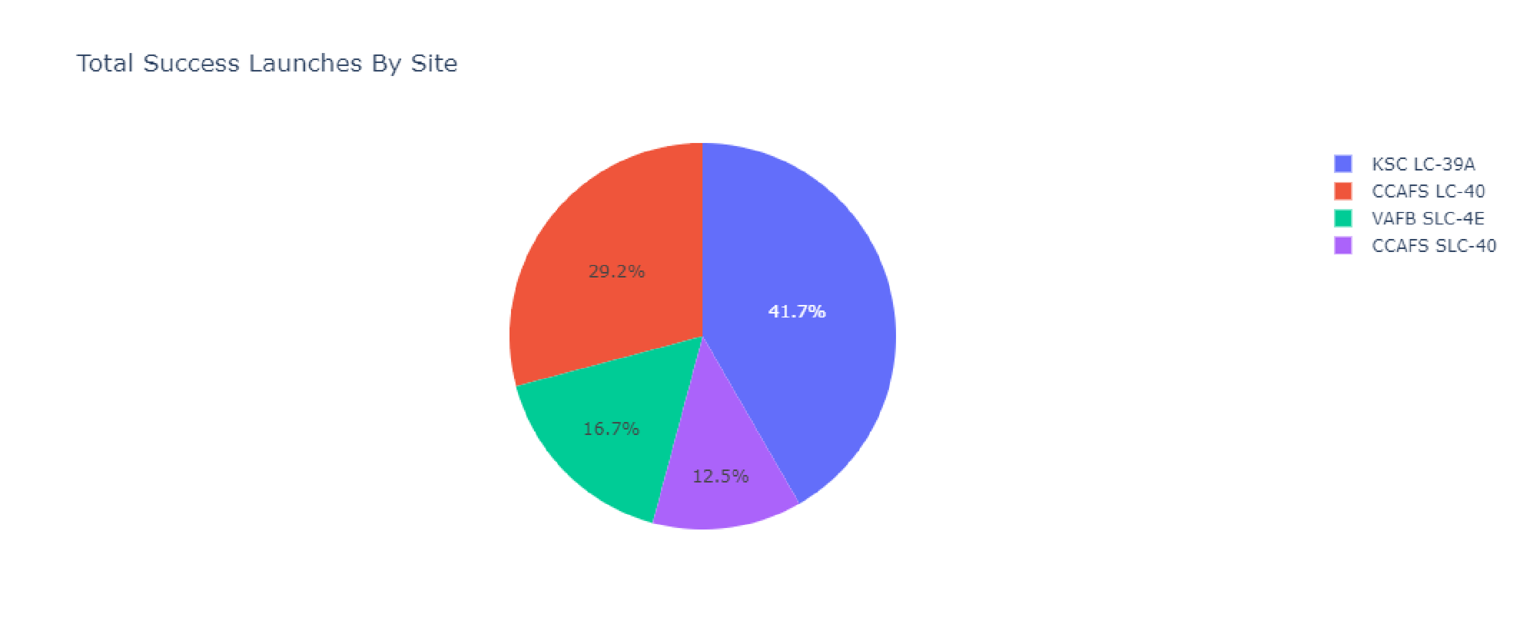
Section 5

Build a Dashboard with Plotly Dash



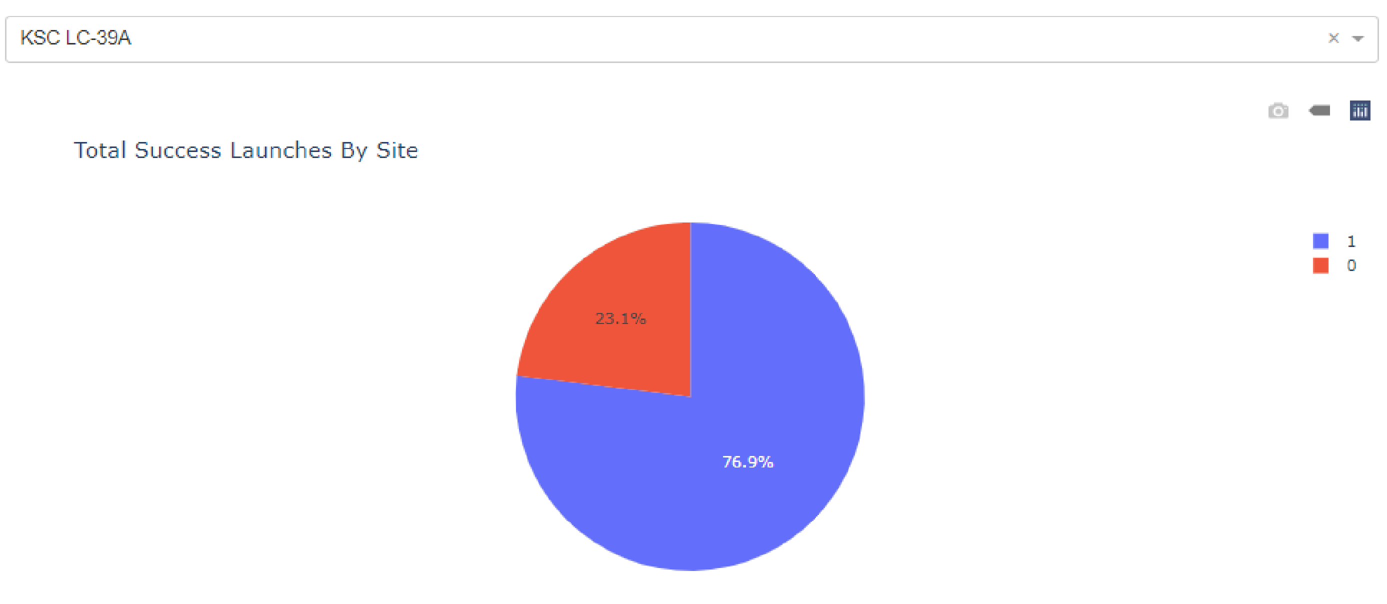
Total Success Launches By Site

- Below is the screenshot of launch success count for all sites, in a pie-chart
- “KSC LC-39A” has the highest success ratio whereas “VAFB SLC-4E” had the lowest success ratio among all four launch sites



Launch Site With Highest Launch Success Ratio

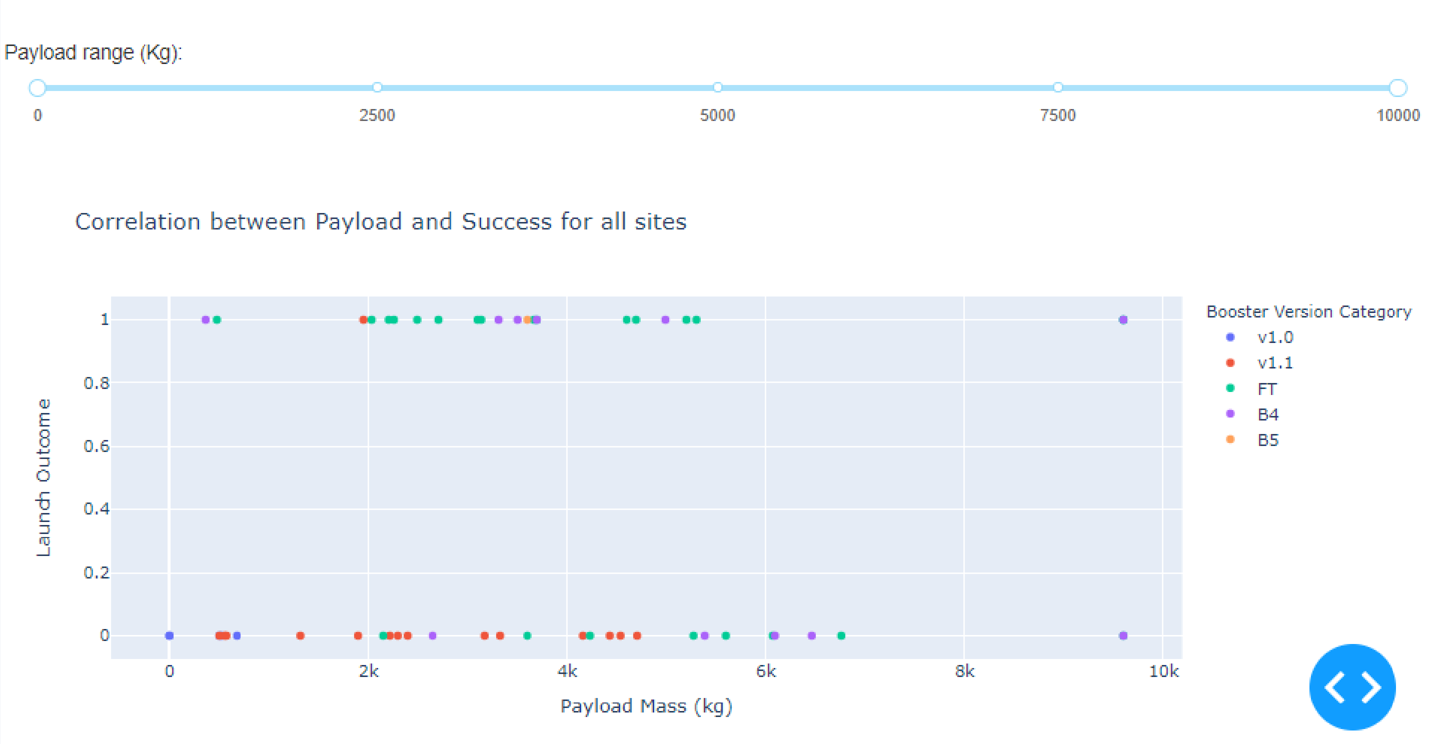
- Below is the screenshot of the pie-chart for the launch site with the highest launch success ratio – “KSC LC-39A”
- It has 77% of success ratio


KSC LC-39A

Total Success Launches By Site

Payload vs. Launch Outcome

- Below are the screenshot of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- In 0-10000 kg payload range "FT" booster version has the largest success rate



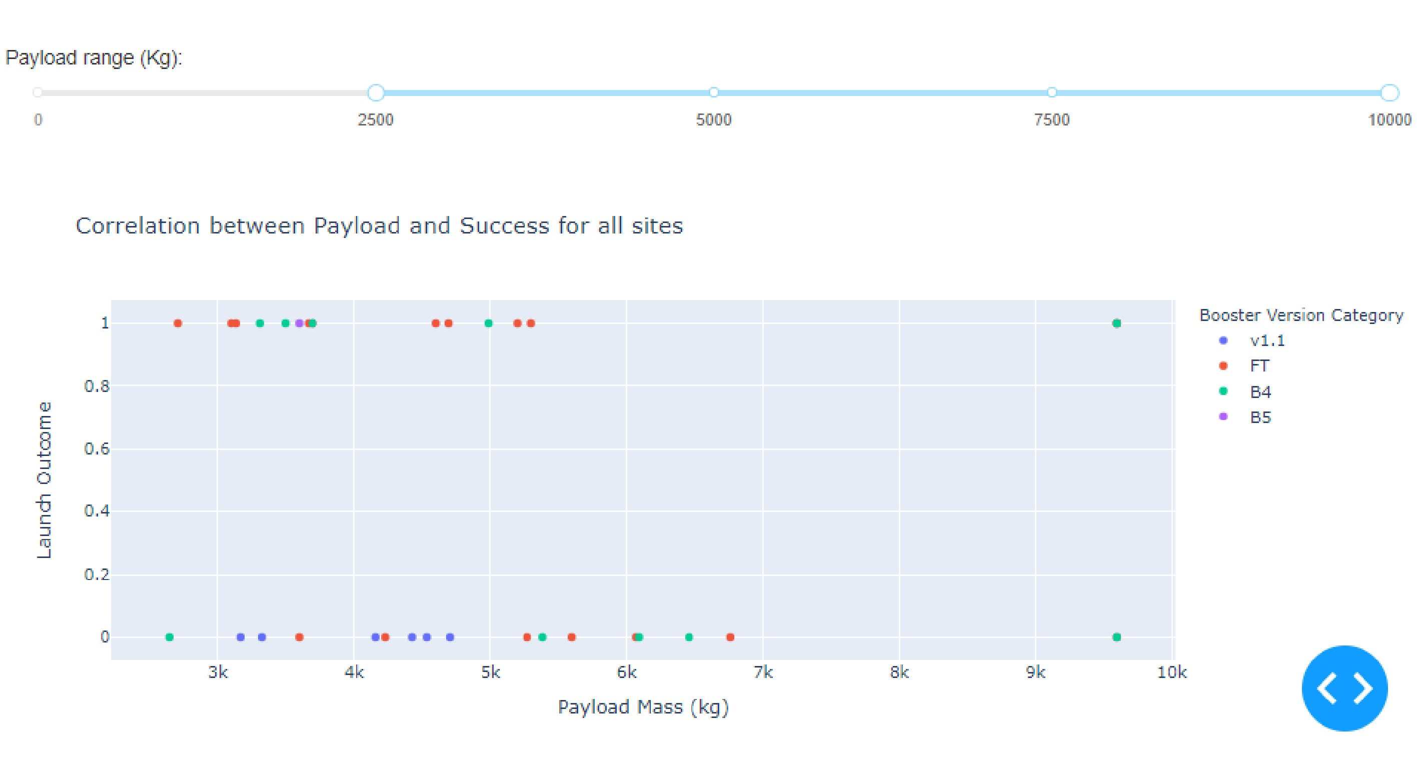
Payload range (Kg):

Correlation between Payload and Success for all sites

Payload vs. Launch Outcome [Continued]

Payload vs. Launch Outcome between 2500 kg and 10000 kg

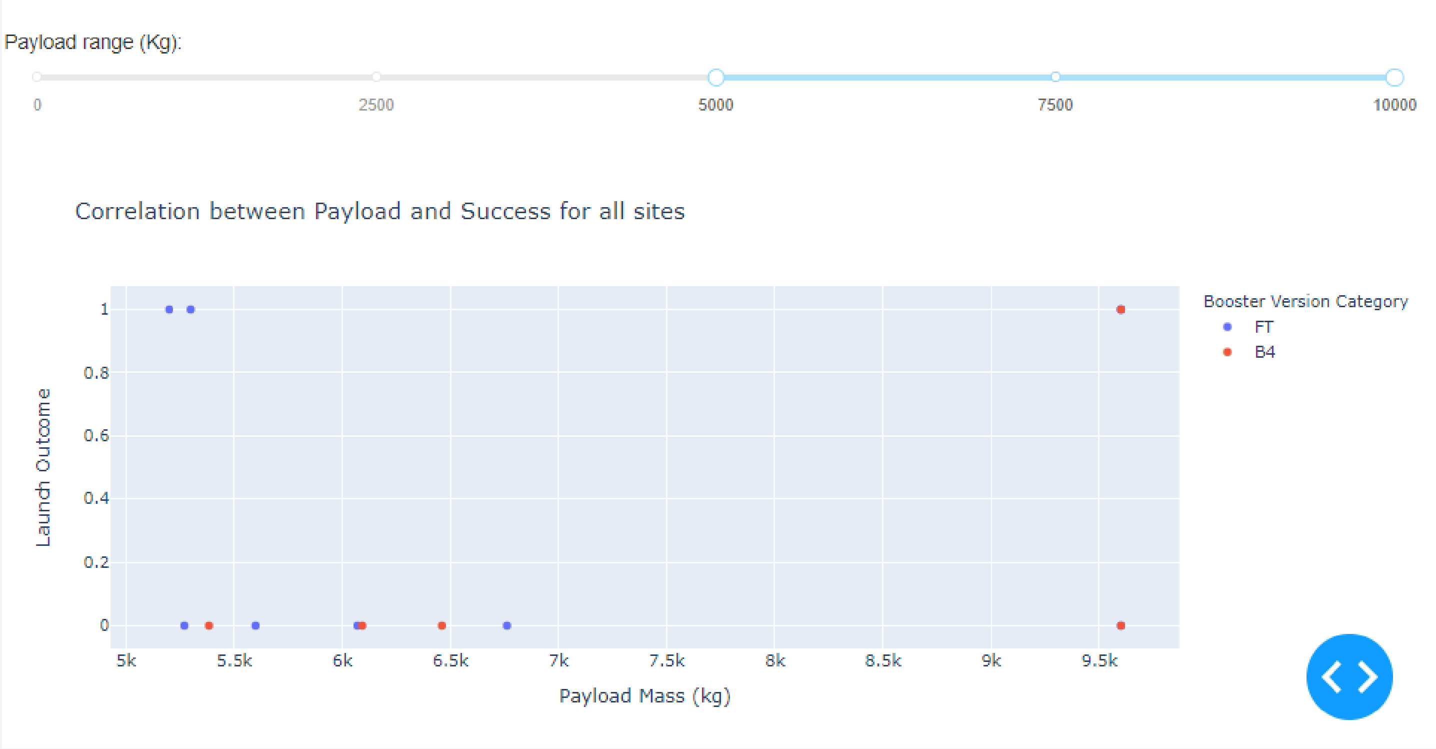
- In this range “FT” booster version has the largest success rate



Payload vs. Launch Outcome [Continued]

Payload vs. Launch Outcome between 5000 kg and 10000 kg

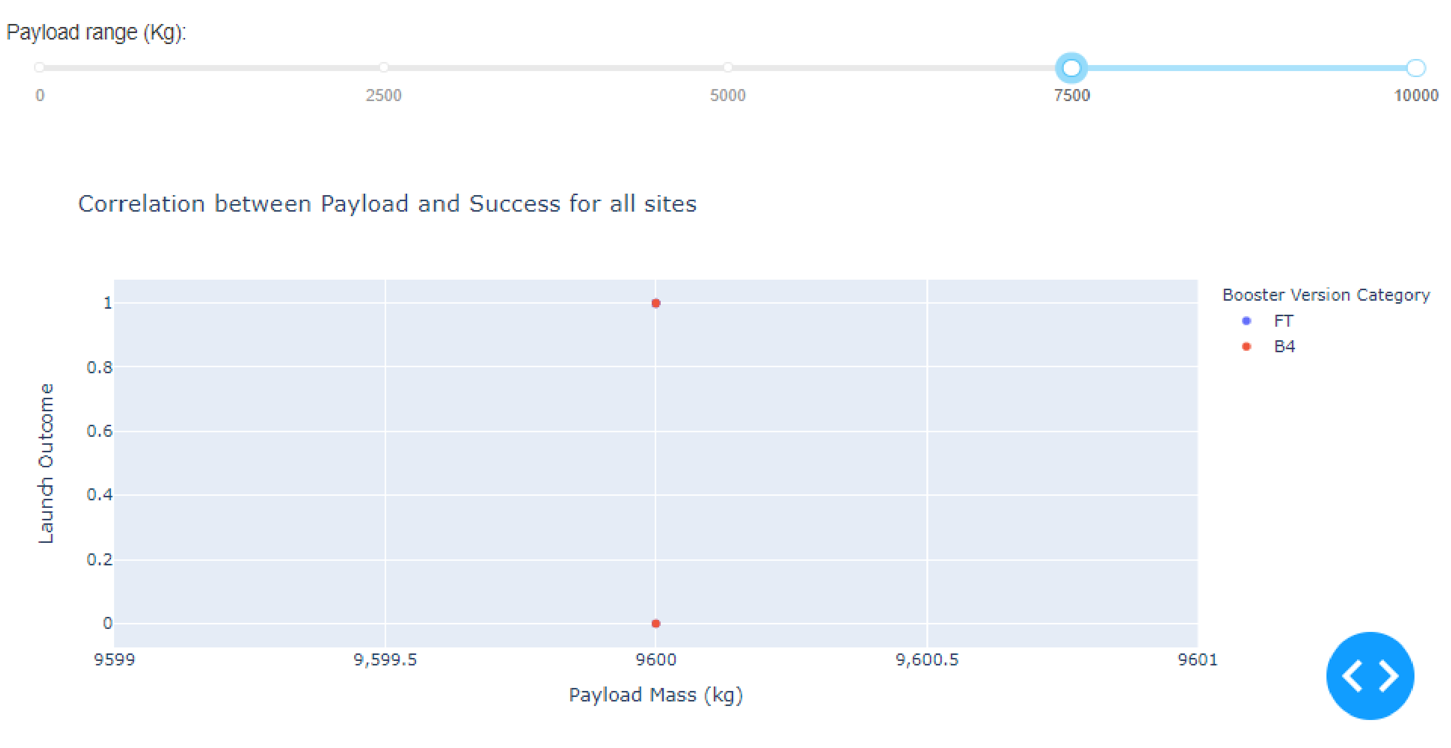
- In this range “FT” booster version has the largest success rate whereas “B4” booster version has no success



Payload vs. Launch Outcome [Continued]

Payload vs. Launch Outcome between 7500 kg and 10000 kg

- In this range only “B4” booster version is used and has equal success / failure rate

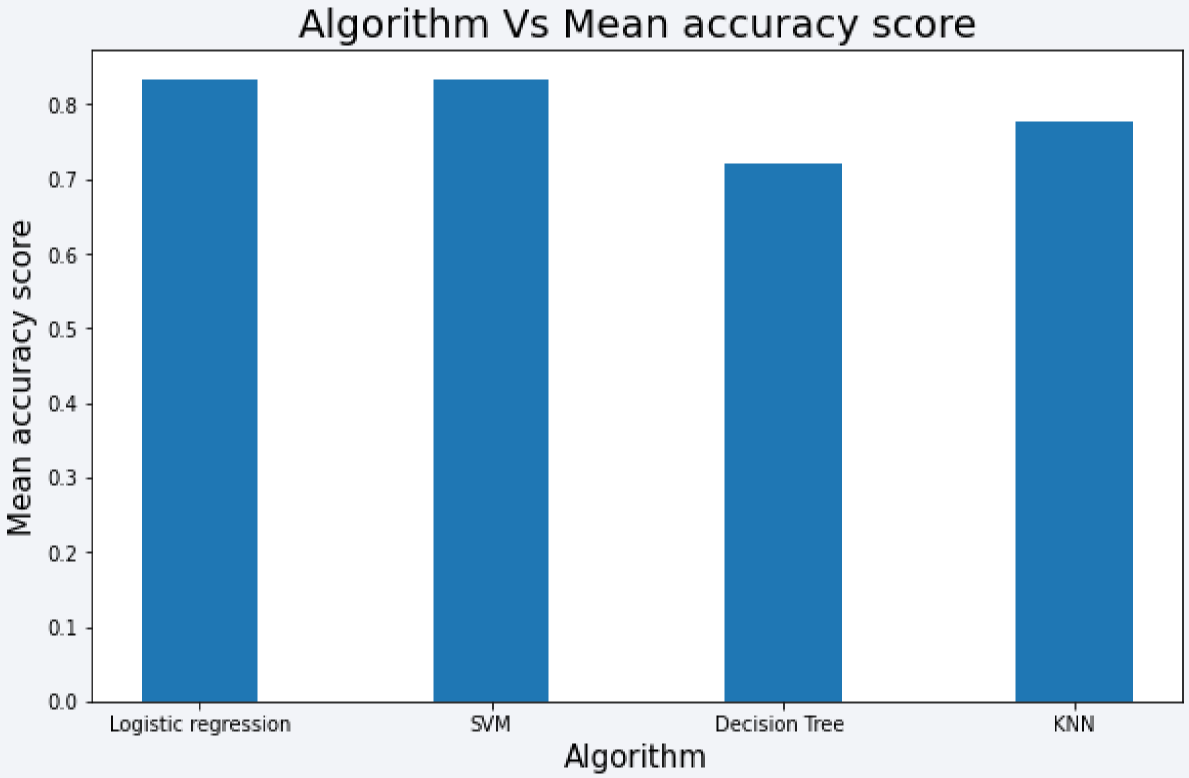


Section 6

Predictive Analysis (Classification)

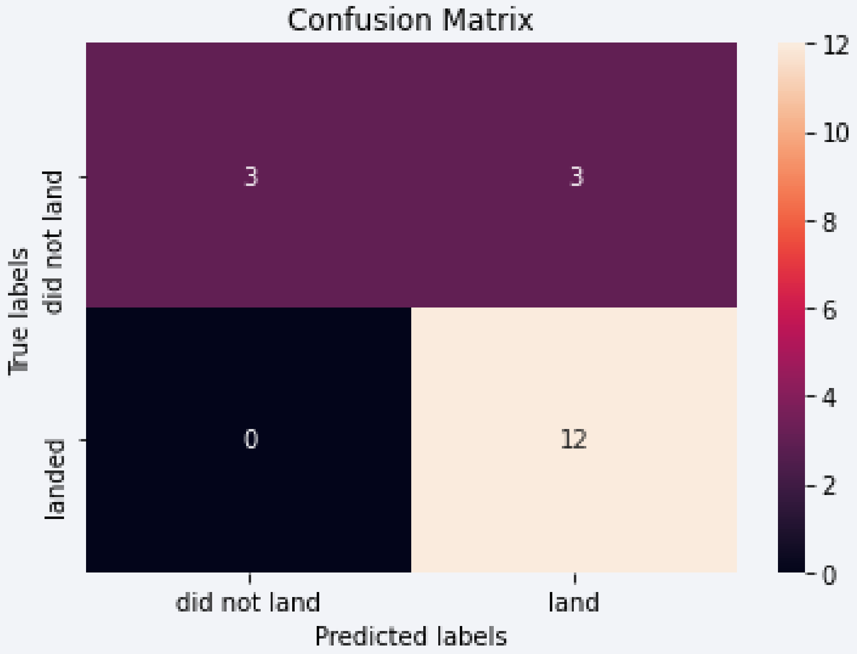
Classification Accuracy

- Logistic Regression and SVM has the highest accuracy which is approximately 0.83
- Decision Tree Classifier has the lowest accuracy (0.72) among all four models



Confusion Matrix

- Below is the confusion matrix of the best performing model (**Logistic Regression** and **SVM** – both generates the same result)
- It has True Positive value of 12 and False Negative of 0. It means this model correctly identified 12 instances which was originally belong to class 1(successful landing), conversely it predicted correctly only 50% (3 out of 6) of the class 0 (landing failure)



Conclusions

- The project explored the SpaceX data and built a classifier on **whether a rocket launch outcome (landing) would be successful or not**
- More data is needed for better prediction
- Different algorithms should be explored to generate a better accurate classifier (Gradient boosting, Neural Network etc.)
- An end-to-end pipeline should be built for classification model that automates most of its process.

Appendix

- Function to report about missing values from a pandas dataframe
- Code snippet on automated hyperparameter tuning
- Folium Map on SpaceX successful or failed launches for each launch site
- Folium Map on The distances between a launch site to its proximities

Thank you!