IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Monika Mittal

23rd Jan 2024

Outline

Executive Summary

- As a Data Scientist first we understand data and its working.
- Then we do data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- As a result we show Dashboard with useful insights

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problem we are going to solve is to determine cost of launch based on this other company can bid.

Section 1

Methodology

Methodology

Executive Summary

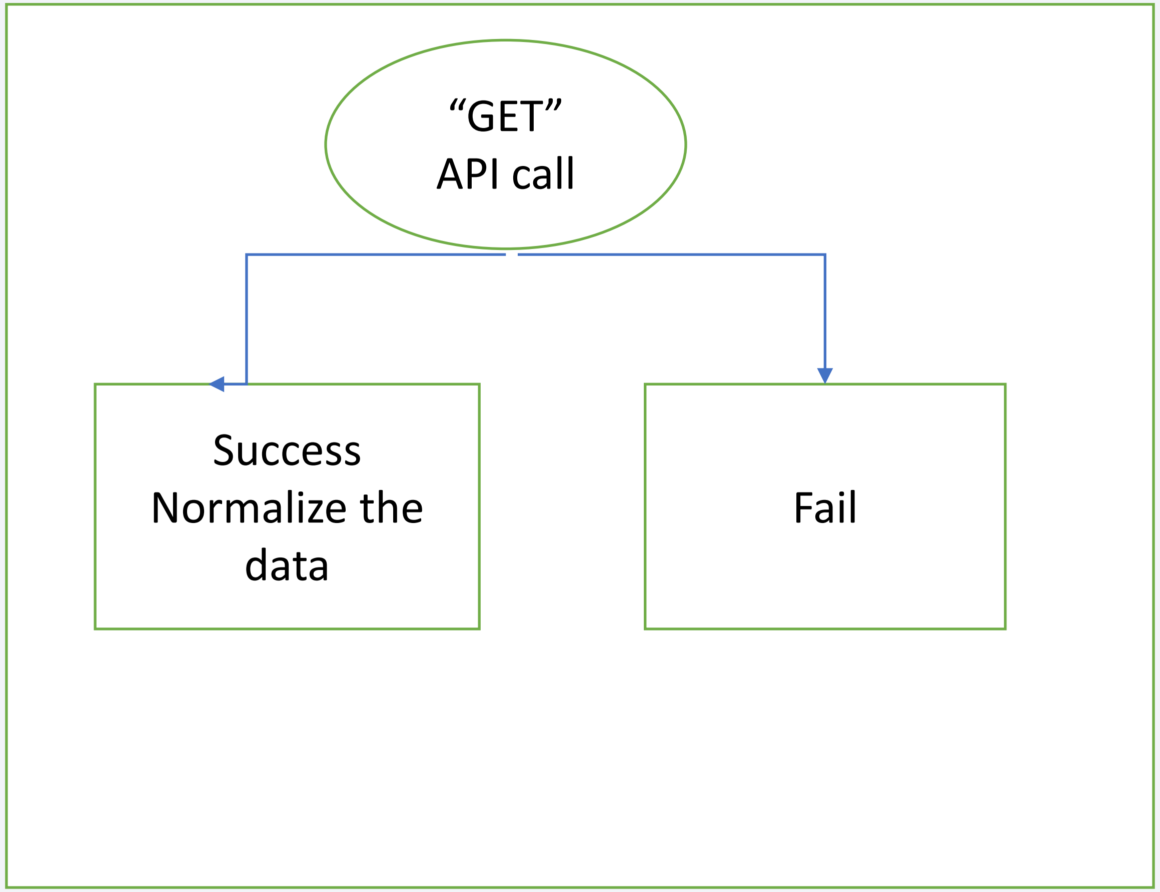
- Data collection methodology:
    - From CSV file hosted on SpaceX server with raw data we have collected data
- Perform data wrangling
    - Deal with null, missing or none value. Our data become nonempty.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - First standardize data, split it into train and test and fit in different models and check their accuracy.

Data Collection

- Data for SpaceX launch was collected via "GET" request by passing API URL to method.
- Then check for response status code if it's success then check for it's content by response.content
- Where the content of response is very massive, so we normalize the it using json normalizer method
- Now we have dataframe to ready in row and column format for Data wrangling process.

Data Collection – SpaceX API

- For SpaceX launch data we called https://api.spacexdata.com/v4/launches/past API and normalize it in json format
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/Spacex-data-collection-api.ipynb

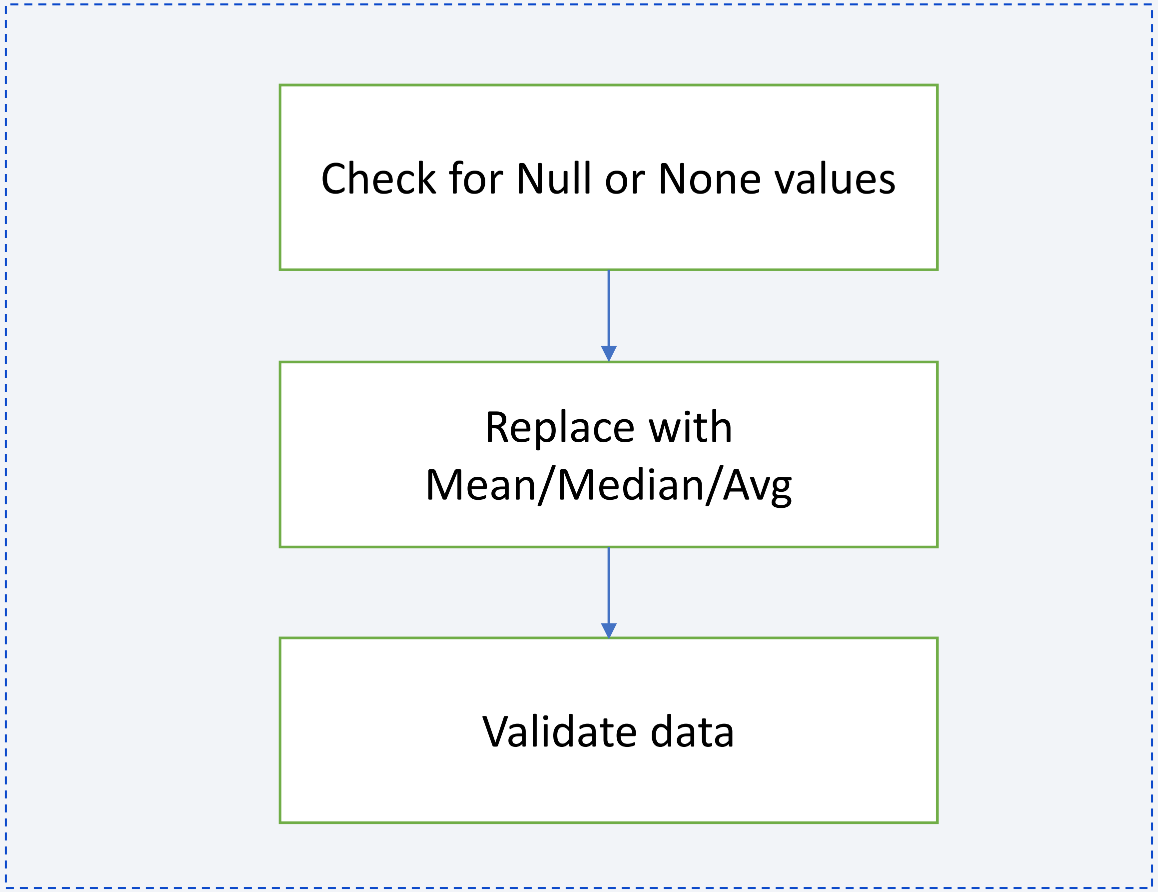


Data Collection - Scraping

- For Scraping we did Web scraping using Wikipedia page for SpaceX Falcon9 landing success.
- Using “BeautifulSoup” we extracted record from HTML tables of Wikipedia page
- Parse those tables to pandas dataframe for further exploratory analysis.
- Following is lab URL
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/Webscraping.ipynb

Data Wrangling

- Once Dataframe loaded check for null values and replace them with mean or average whichever is best.
- Validate the all cells
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/Spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Feature Engineering using `Pandas` and `Matplotlib`
- Exploratory Data Analysis
- Preparing Data Feature Engineering
- https://github.com/monikaa947/AppliedDataScience/blob/076c37fd8498f43b857cc2e888d25d7751ad63d9/EDA-DataVizualization_Pandas_Matplot.ipynb

EDA with SQL

- First, we understand the dataset of SpaceX
- Load the dataset into the corresponding table in a Db2 database
- Execute SQL queries
- Find unique Launch sites
- Display total and average payload mass carried by boosters launched by NASA (CRS)
- Used subqueries to filter data
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/EDA-SQL-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To highlight areas Circle object of Folium class used.
- Marker of Folium is used to mark location map like launch site.
- Lines were used to show distance between two marks on folium map.
- Mark all launch sites, success/failed launches for each site on a map
- Calculate the distances between a launch site to its proximities
- Here is Git link of lab
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/Falium_site_location.ipynb

Build a Dashboard with Plotly Dash

- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- Dashboard is developed get info about launch site success landing for all and specific with the help of pie and scatter plot of plotly express.
- Build Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real time
- https://github.com/monikaa947/AppliedDataScience/blob/236e5c7814c7fc5fa4adbc1a6f05d9f5cfa01017/Spacex_Dash_App.py

Predictive Analysis (Classification)

- For classification we used different models like logistic regression, support vector machine, decision tree classifier, K Neighbors Classifier with GridsearchCV and compared their scores to identify best model for classification.
- Process to develop different models is create a column for class, Standardize the data, Split into training data and test data and fit train dataset and check its score with test dataset.
- https://github.com/monikaa947/AppliedDataScience/blob/e48bc66c6d71ca43be7bced9f24c53dbd229cf0c/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

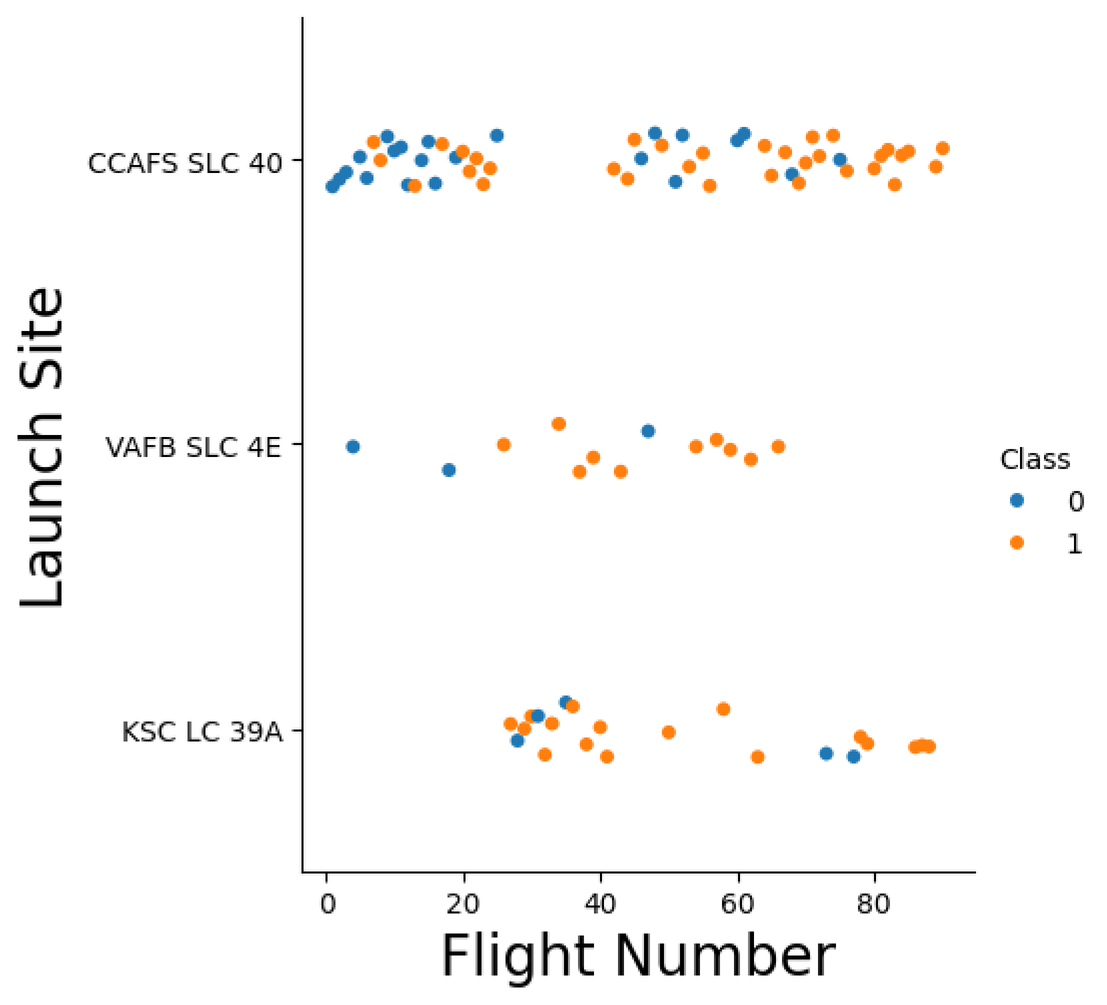
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

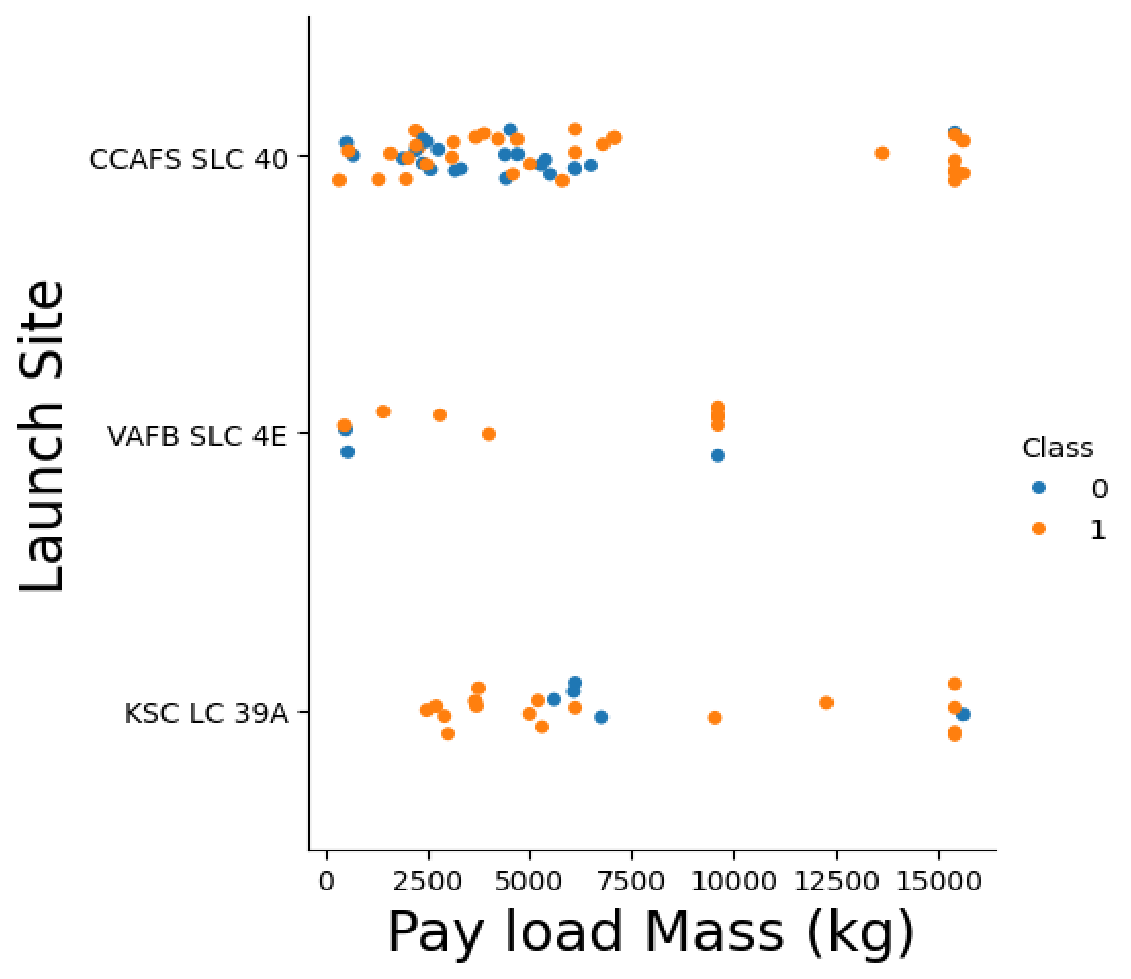
Flight Number vs. Launch Site

- Scatter plot depicts the landing of different launch site with class showing success and fail and it is related to flight number
- Based on lines on plot we can observe there is strong relation of KSC LC 39A launch with flights and CCSFS SLC 40 is less effected by flight numbers.



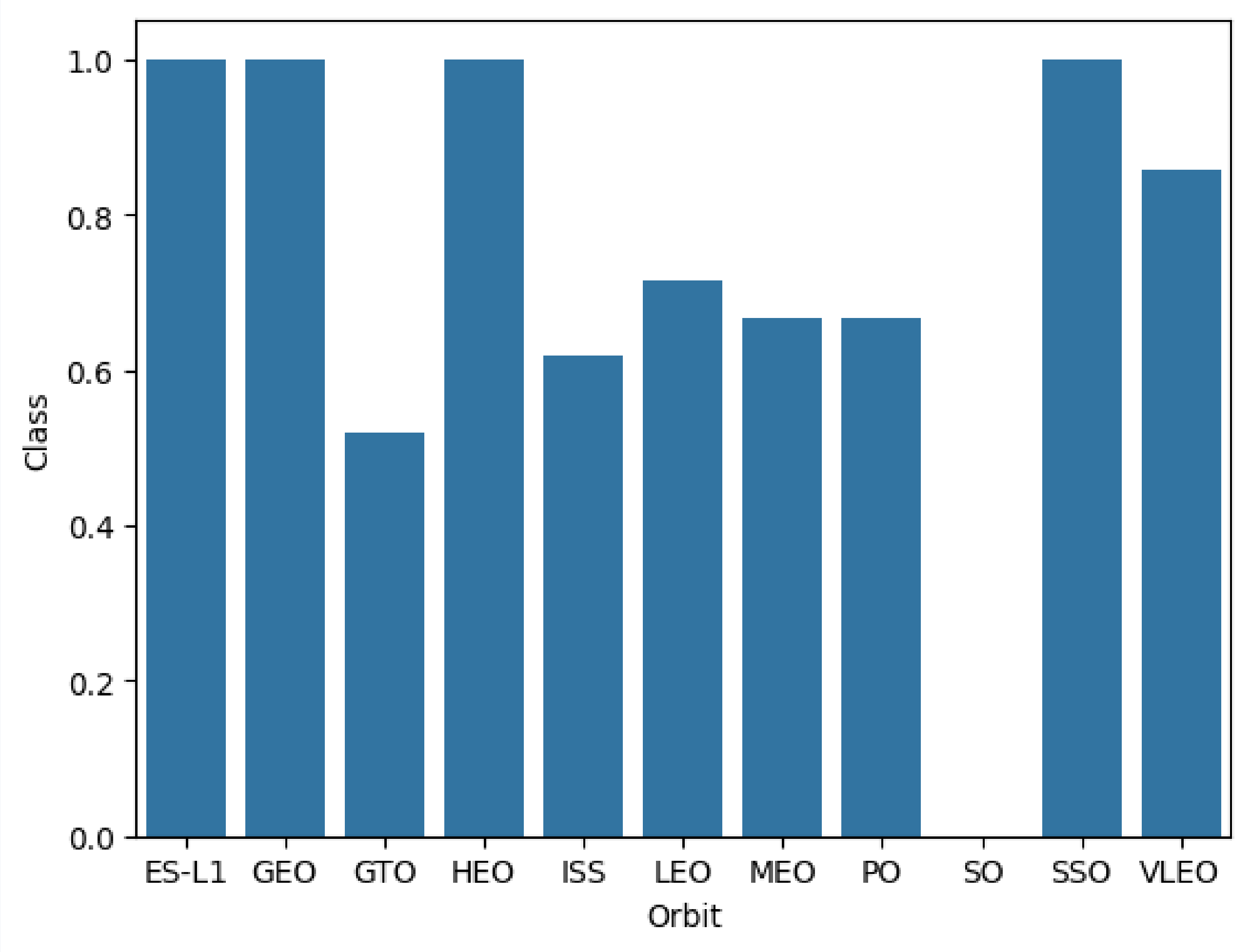
Payload vs. Launch Site

- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



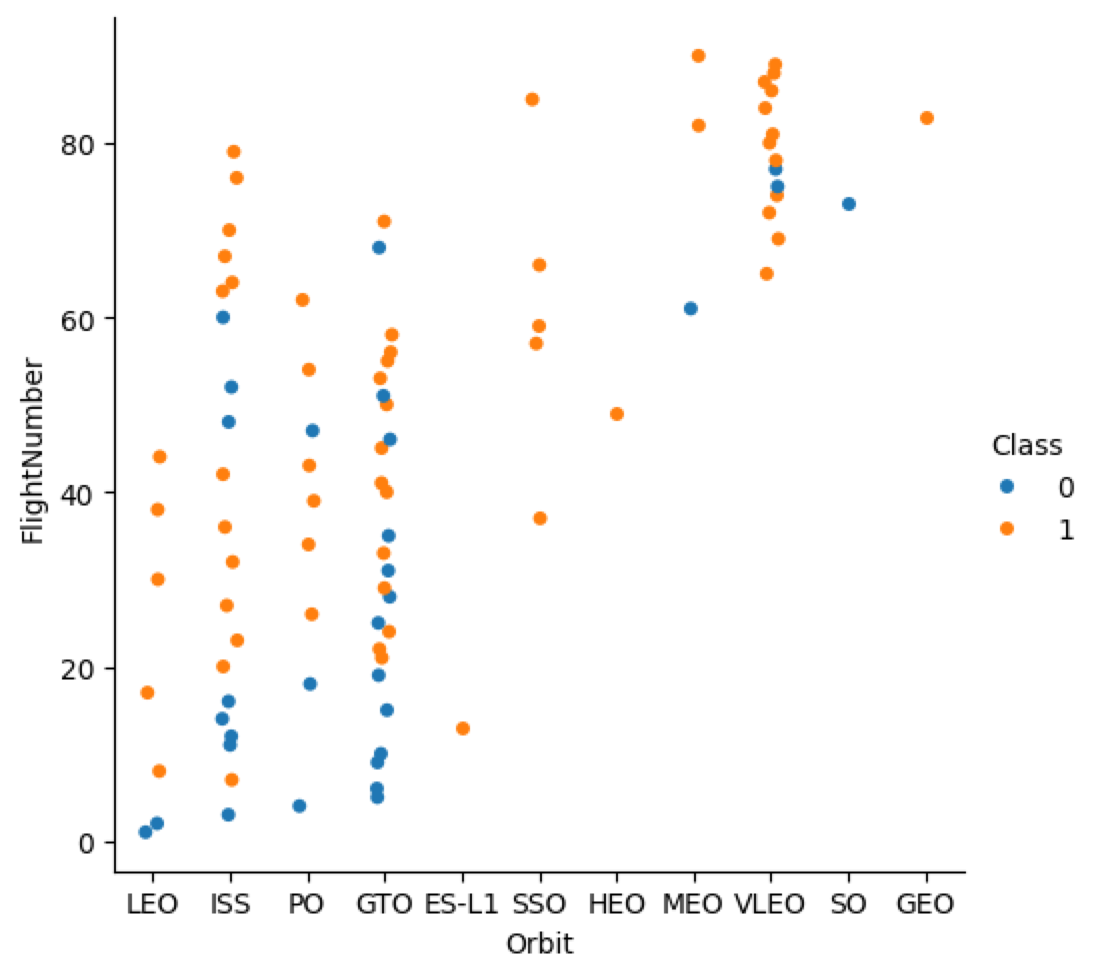
Success Rate vs. Orbit Type

- As per bar chart we can observe ES-L1, GEO, HEO, SSO orbits have high success rate which is 1.
- Where SO orbit doesn't have any or very low success rate.



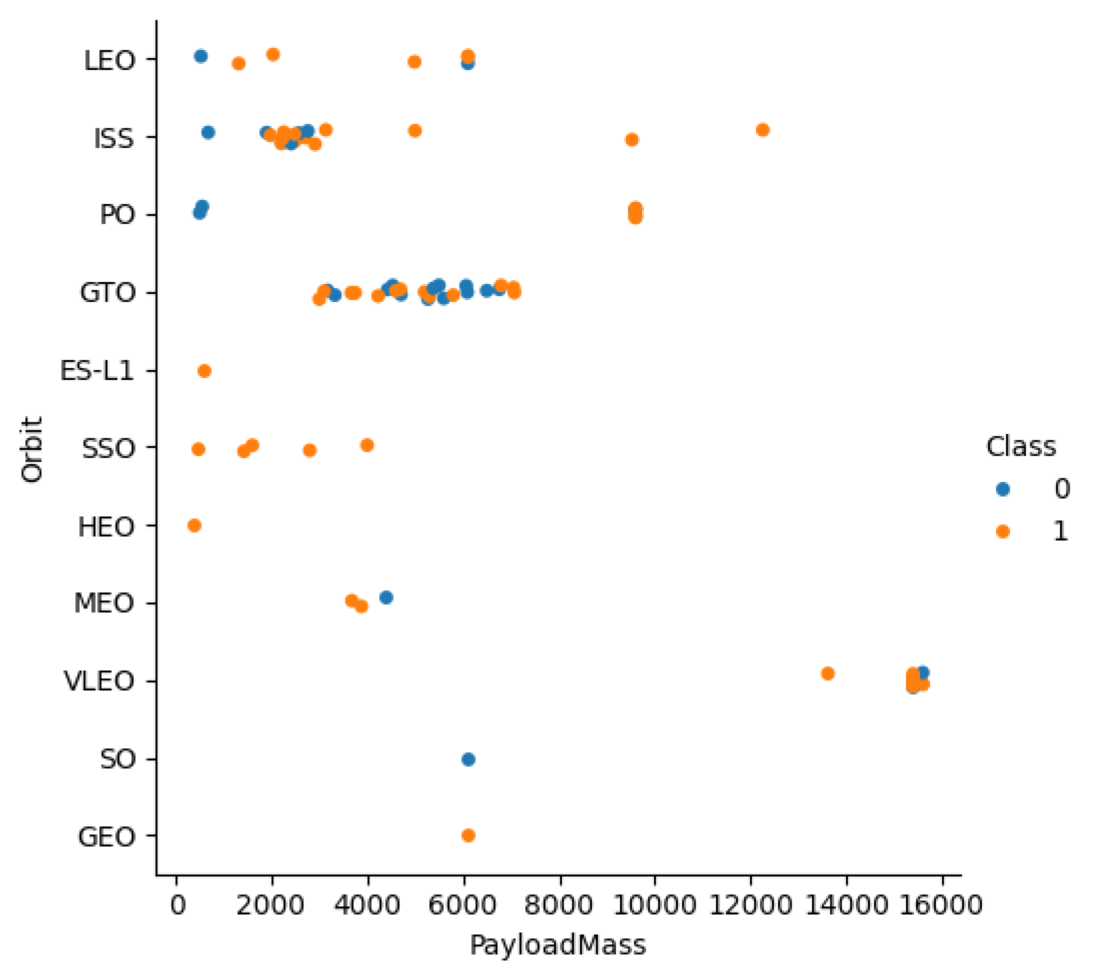
Flight Number vs. Orbit Type

- We can see that in the LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



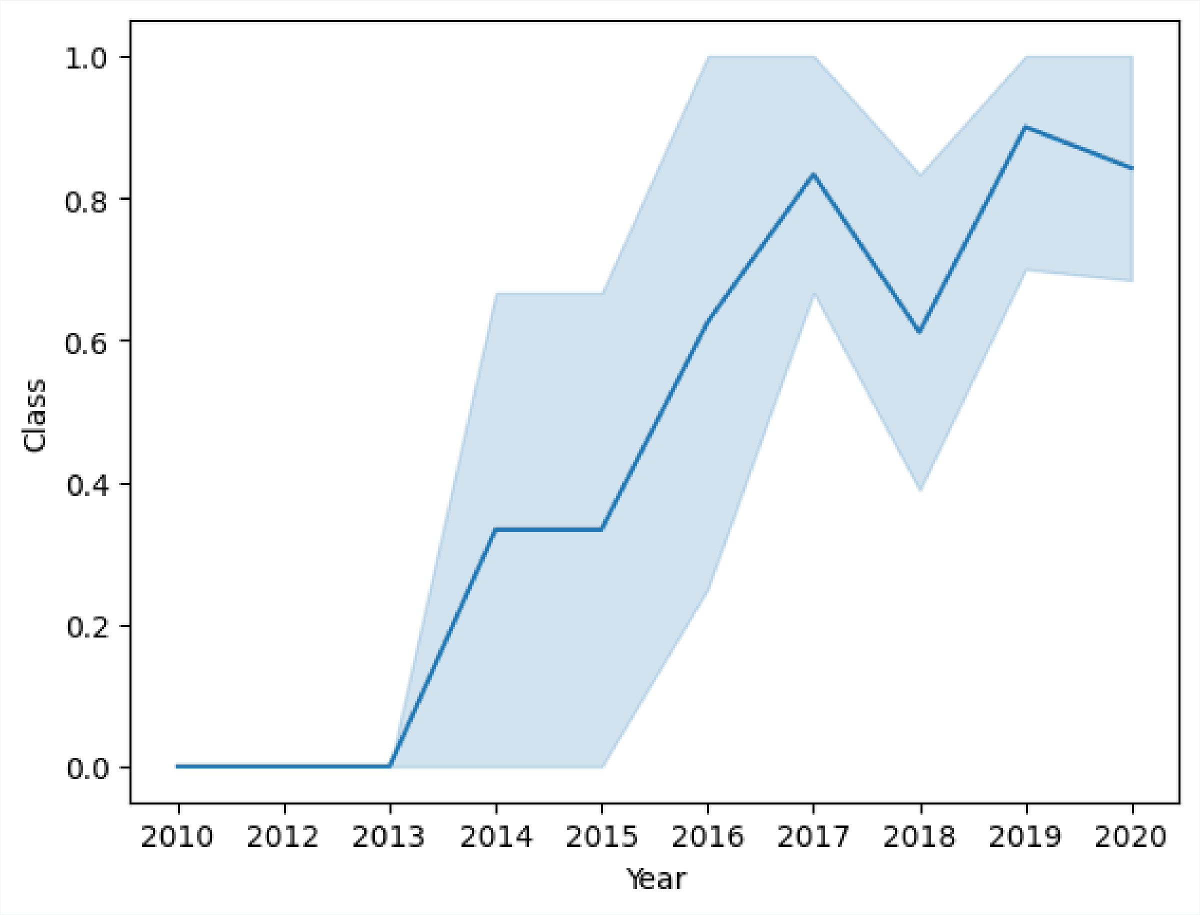
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Query to find unique launch site in database is "*%sql select DISTINCT Launch_Site from SPACEXTABLE*"
- Here is the result:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Used Like and limit operator to find launch site starting from 'CCA%' and top 5 records, here is the records. Following is sample response

Date	Time (UTC)	Booster_ Version	Launch_Si te	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)

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Used sum and where clause to find total of payload with specified customer 'NASA (CRS)'. Here is sample response

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Average function used to find payload mass average where booster version is 'F9 v1.1'. Here is sample response

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Min function used to find minimum date where landing outcome is “Success”. Here is sample response:

min(Date)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Where clause with multiple condition separated from AND operator Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000)

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- In where clause using Like operator, we found Success and Failure mission outcomes

Mission outcome Success

count(Mission_Outcome)
100

Mission outcome Failure

count(Mission_Outcome)
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Using subquery in where clause applied max on payload mass. Here is sample response

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2019-11-11	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We used substr fun to find month and year in where and selection query to perform filter. Here is sample response.

substr(Date, 6,2)	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	2015-02-11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	2015-03-02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	2015-04-14	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
- Applied multiple condition using AND and OR operator to filter data. Here is sample response

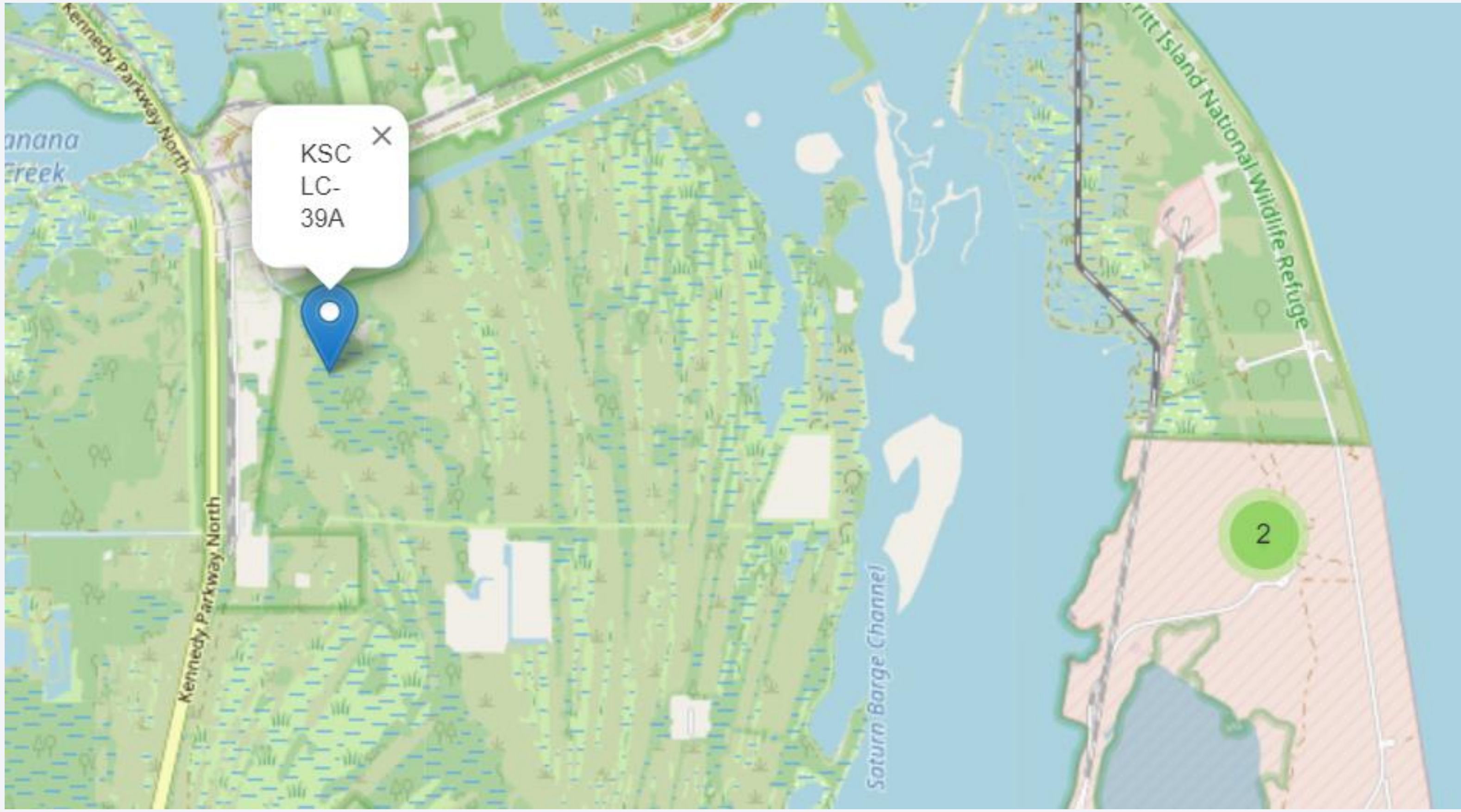
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

Section 3

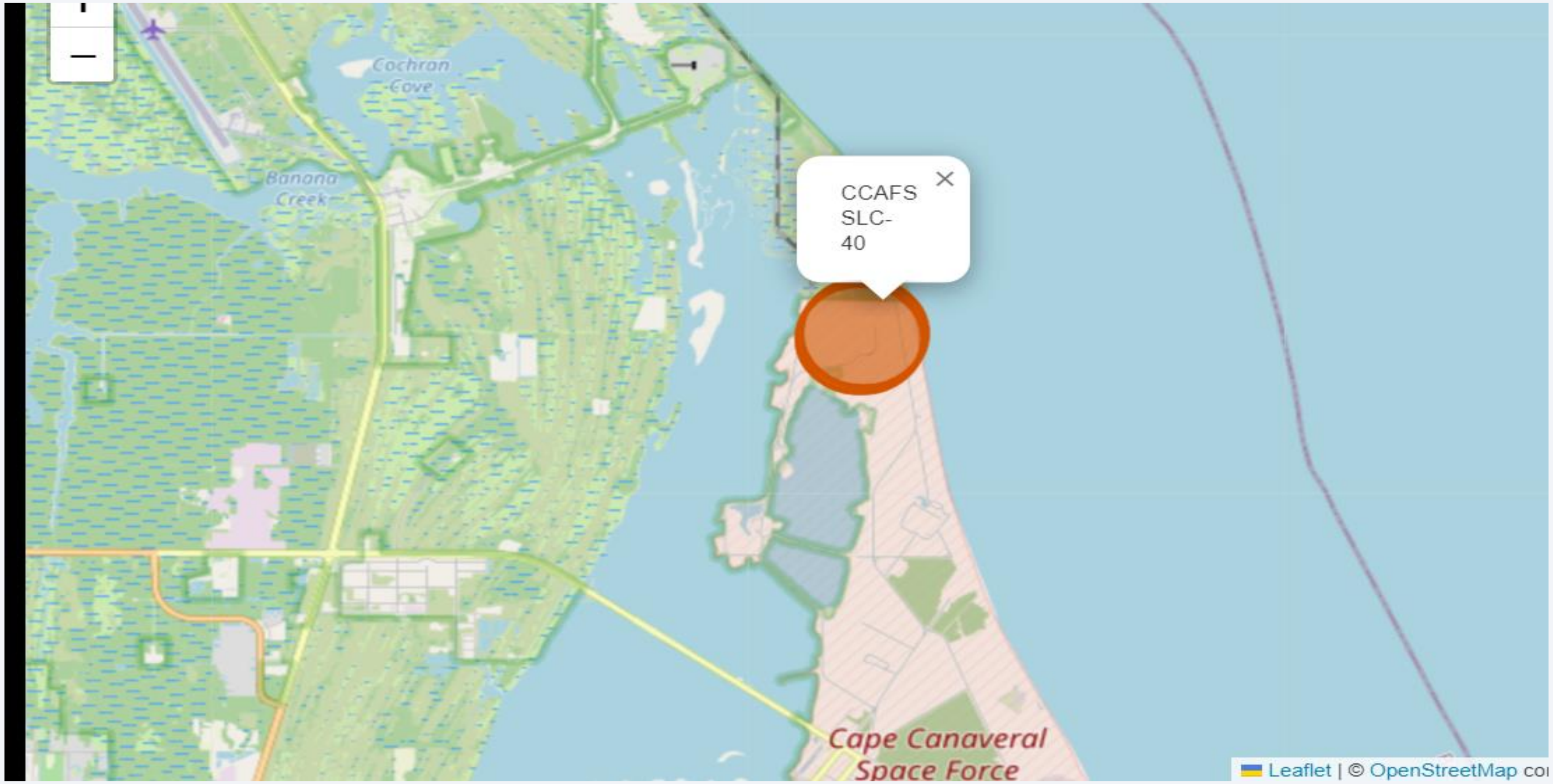
Launch Sites Proximities Analysis

All launch sites on Global map

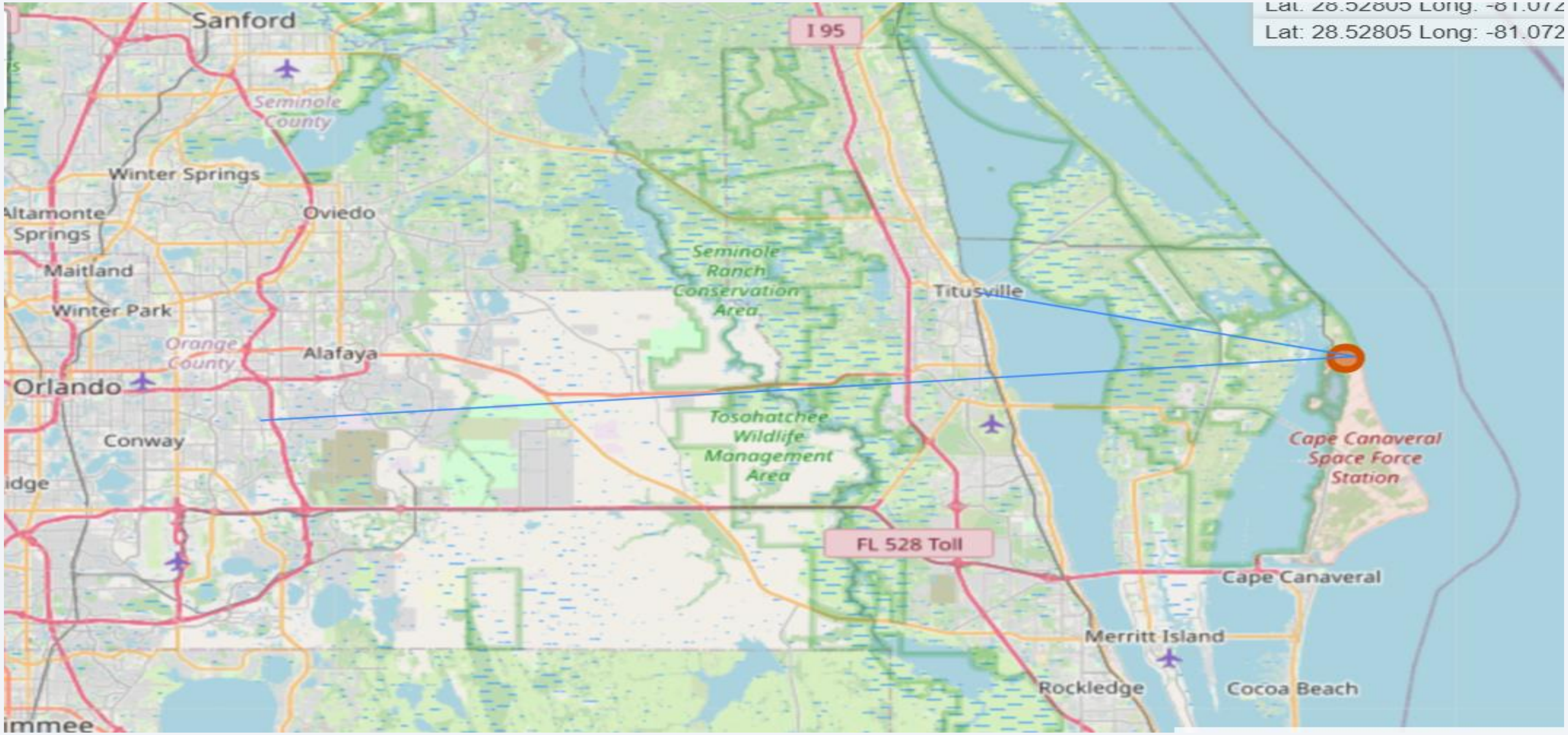
- We can see multiple markers on map.
- Green spot with 2 number shows 2 markers.



Success/Failed launches for each site



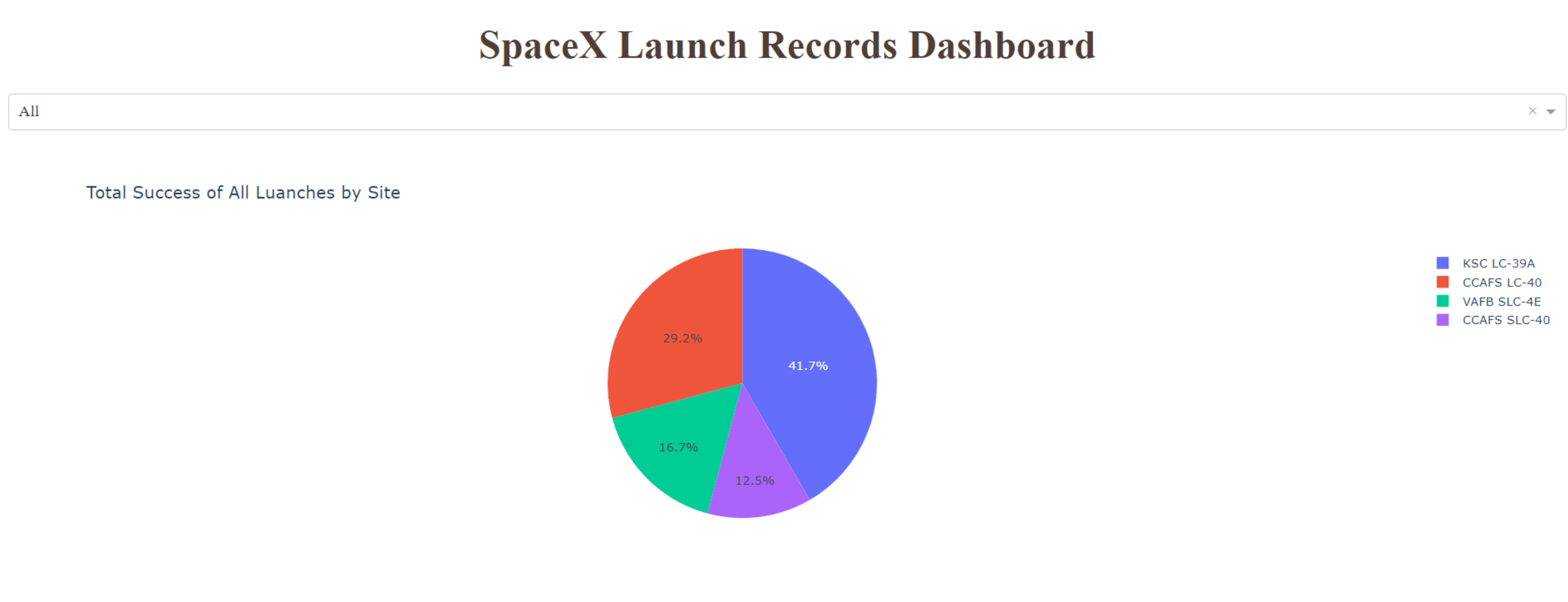
Distances between a Launch site to its proximities



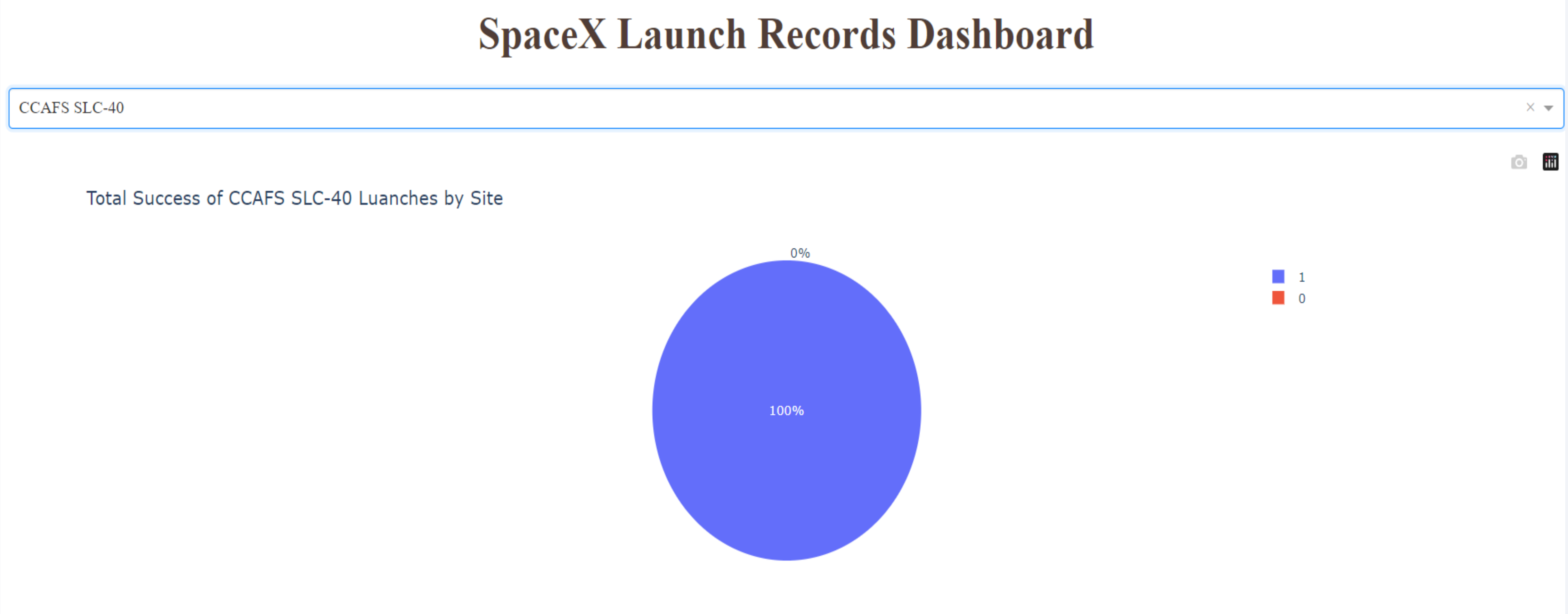
Section 4

Build a Dashboard with Plotly Dash

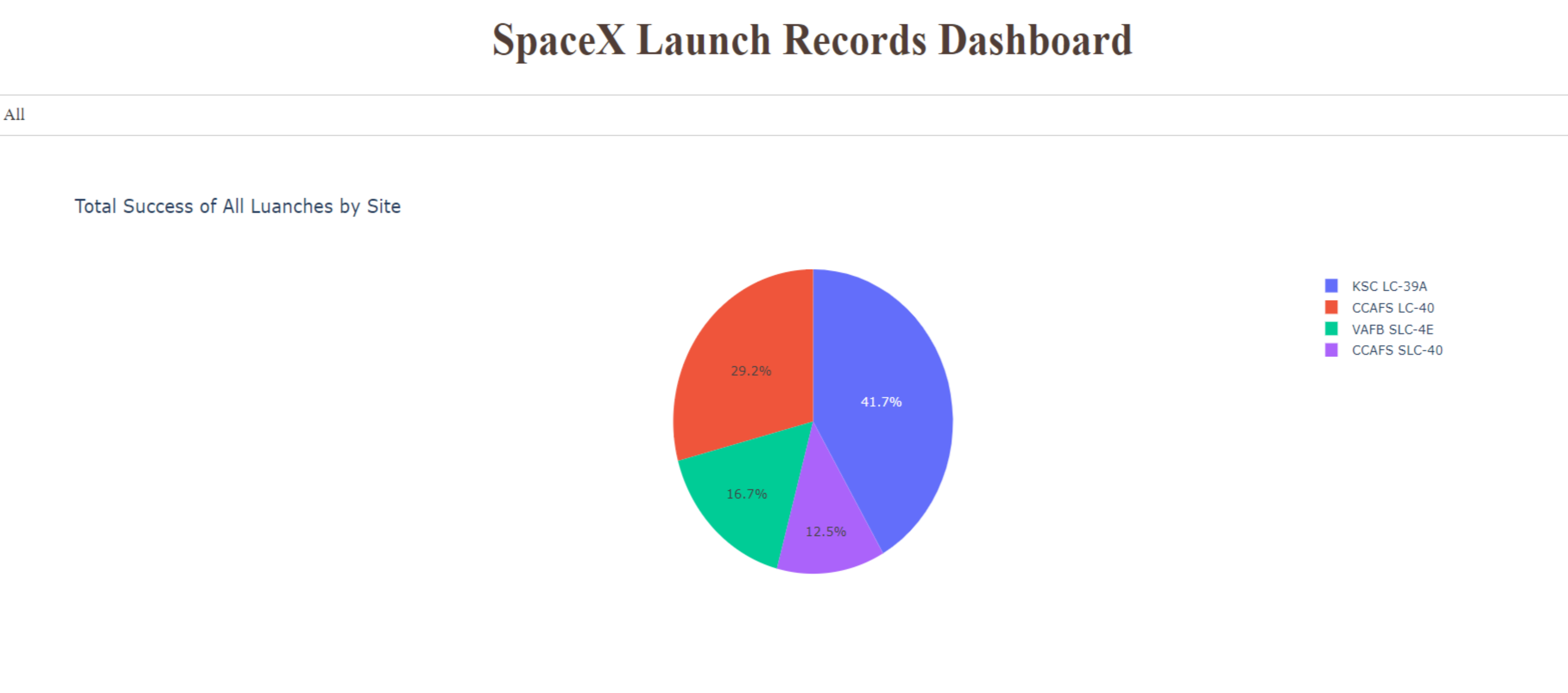
Dash for total success of all sites



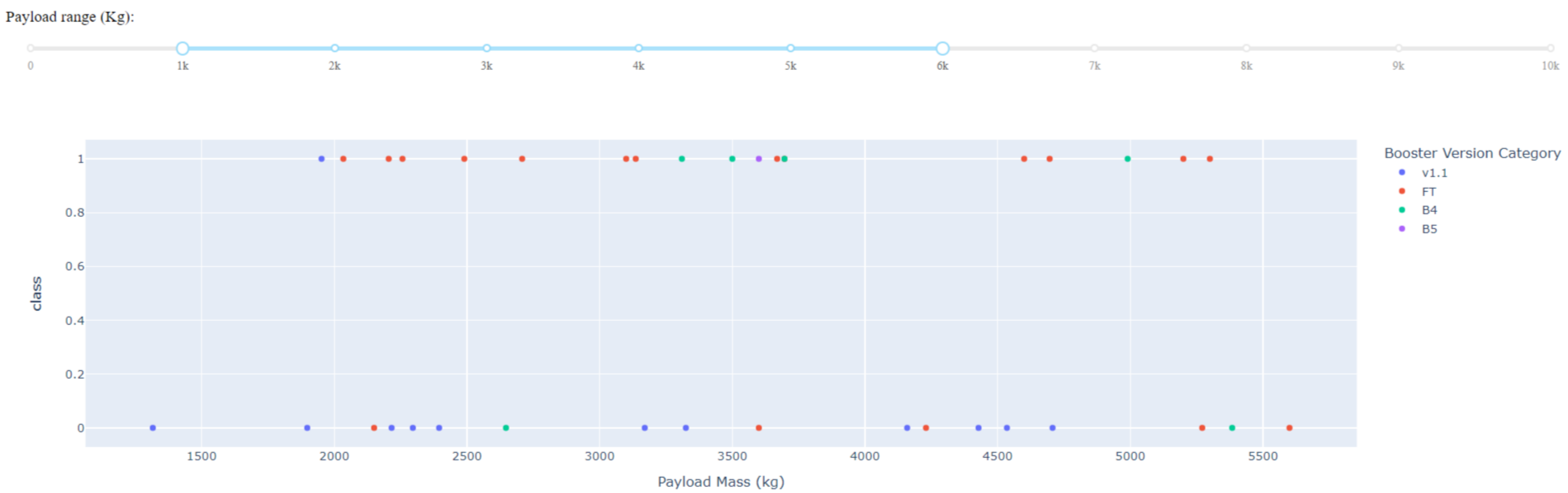
Success launch of CCAFS SLC-40



SpaceX Launch Records Dashboard



SpaceX Launch Records Dashboard

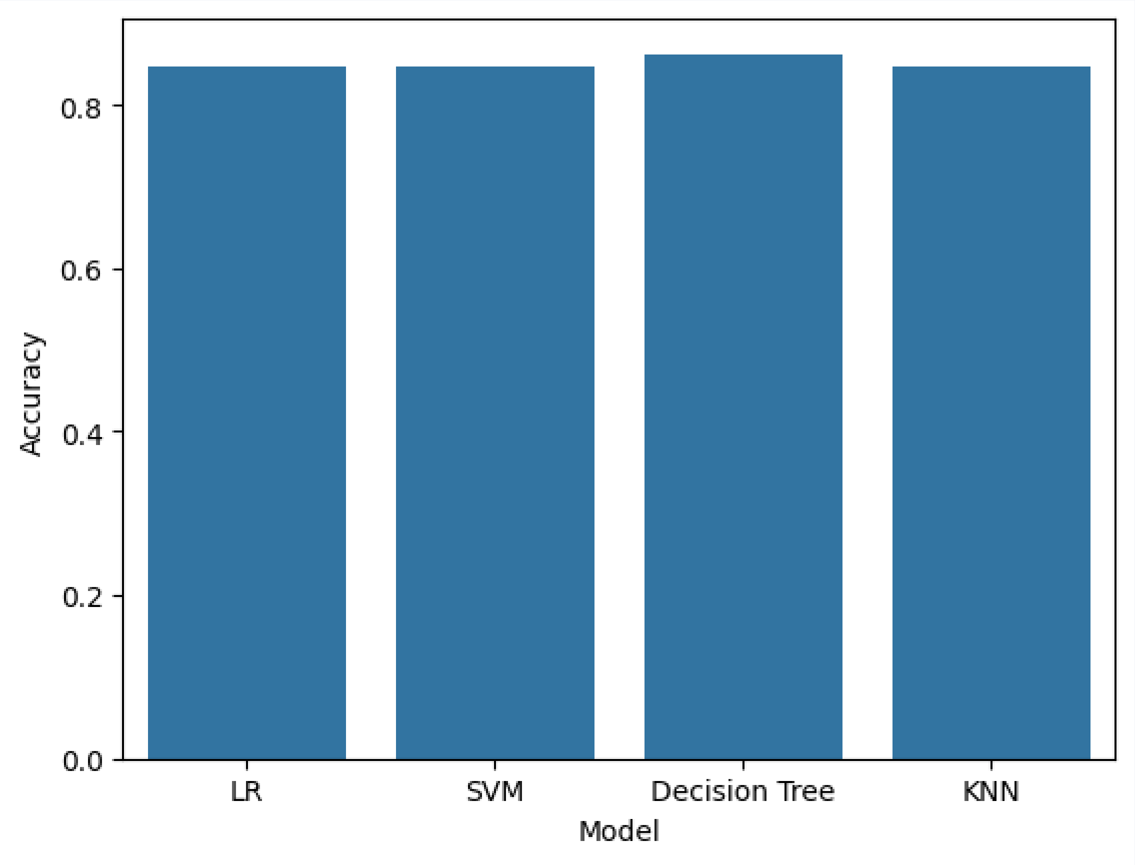


Section 5

Predictive Analysis (Classification)

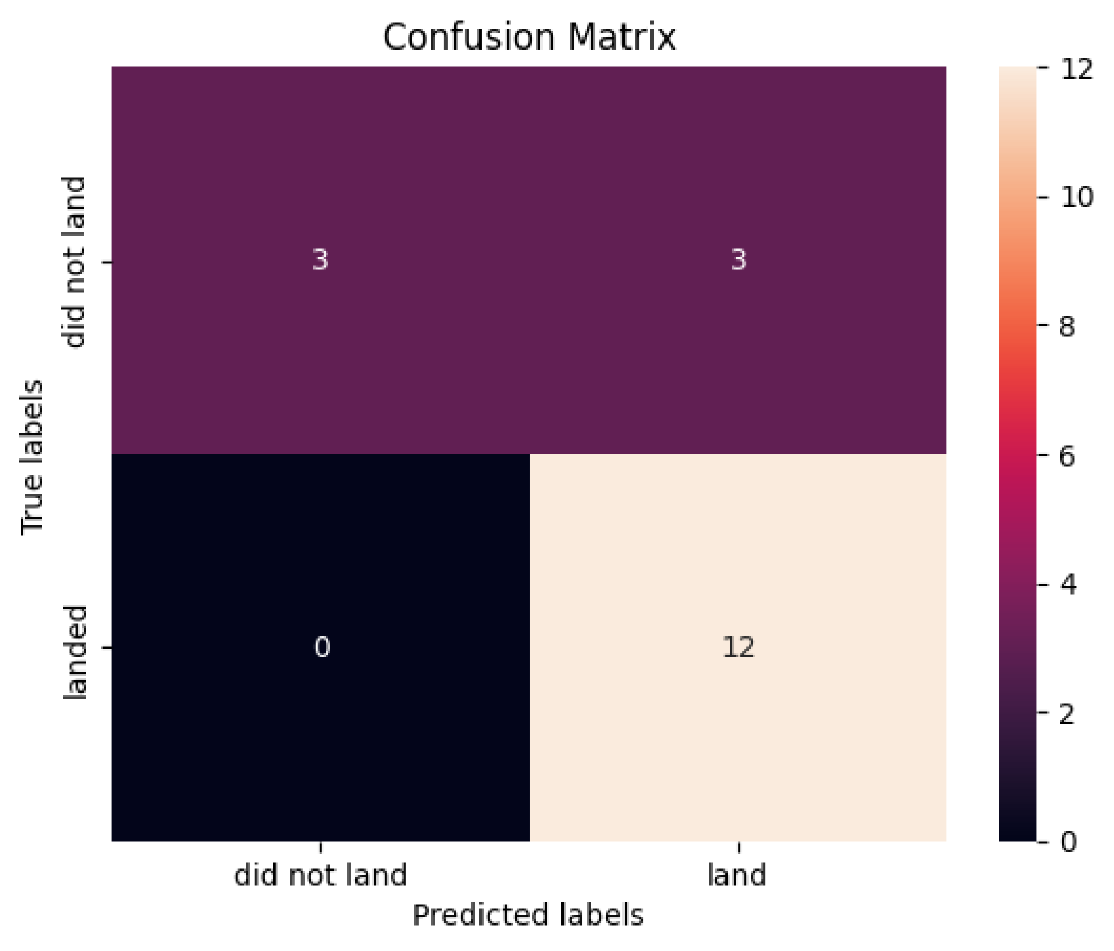
Classification Accuracy

- Based on Bar plot Decision tree have highest score
- May be decision tree has given more accuracy



Confusion Matrix

- Based on accuracy we can conclude that Decisions tree's confusion matrix gives more accurate answer

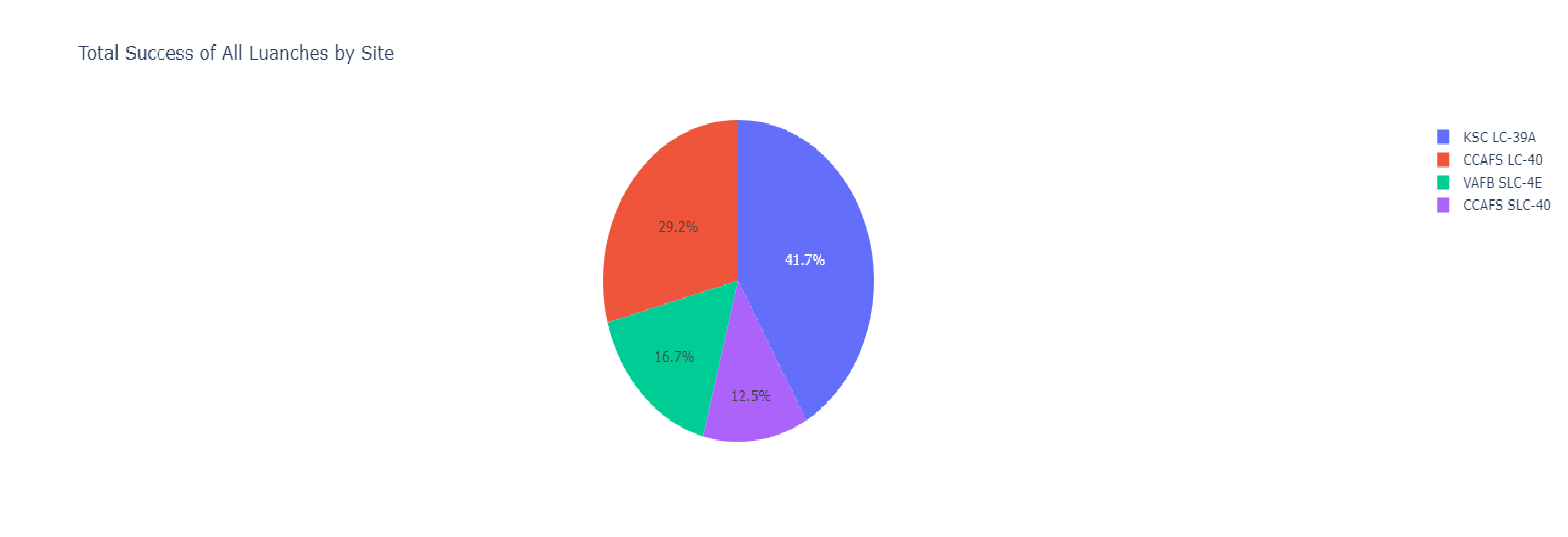


Conclusions

- After performing all task in labs, I have successfully implemented most of the concepts of Data Science using SpaceX Falcon 9 launching site.
- As a Data Scientist first we understand data and its working.
- Then we did data collection, Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- As a result we show Dashboard with useful insights

Appendix

- Success Rate of all launch site is as following



Thank you!