IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Reza Fahmi
20th October 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API
  - Data collection using web scrapping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Dashboard with Dash/Plotly
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis Result
  - Interactive Analytics
  - Predictive Analytics Result

Introduction

- Project Background
  - Rocket launching is an expensive project that needs to be managed carefully
  - The cost for a normal rocket launching project reaches 165 million dollars
  - SpaceX tries to reduce that cost to only 62 million dollars using different techniques with Falcon 9 launch
  - However, not all rocket launch will be successful
  - Finding probability of successful rocket launch will be beneficial for future launch
- Problem Statements
  - How is the correlations between parameters?
  - What are the most important factors in determining the success or failure of rocket launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scrapping from Wikipedia page
- Perform data wrangling
  - One-hot encoding was applied to categorical feature
  - Missing values were filled with the column mean value and newly needed column was added
  - Training labels were generated from available data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis using Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor
  - Use Grid Search for Model Optimization

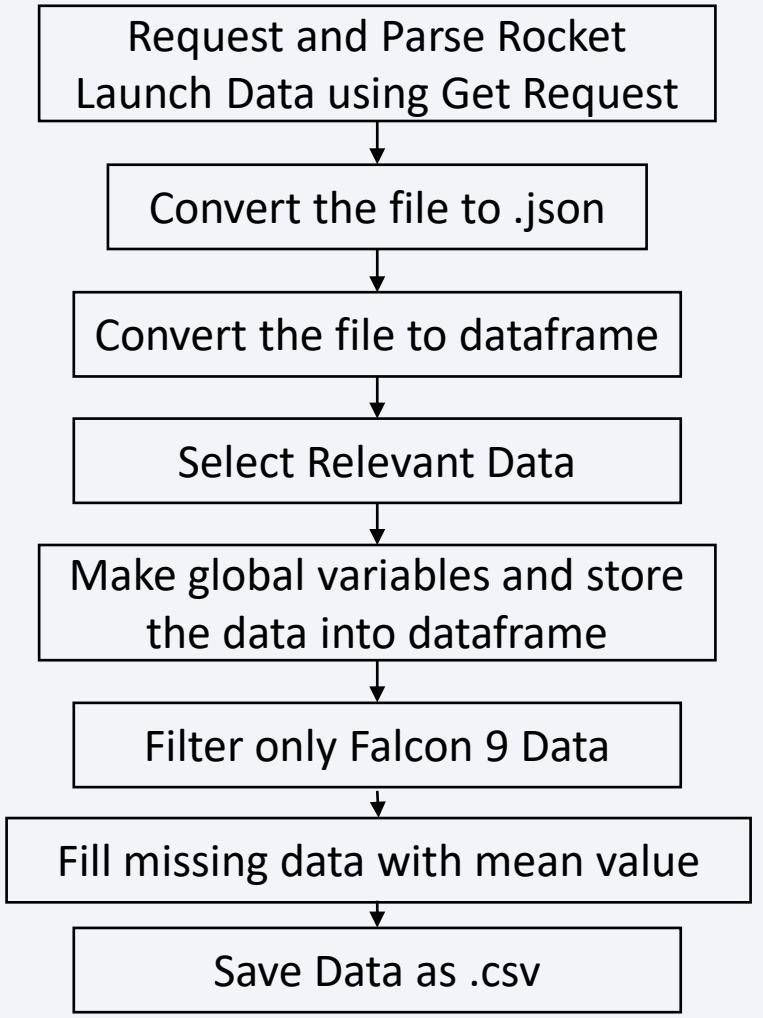
Data Collection

- The data were collected using two (2) main methods: using SpaceX API and Webscrapping method using BeutifulSoup.
  - For SpaceX API, the data was requested and parsed before converting it into .json format. After that, the data is converted into dataframe using .json_normalize(). The next step is to perform data wrangling by cleaning the data and filling the missing values
  - For webscrapping, the data was scraped from table data type for Falcon 9 launch record at Wikipedia URL before beautify it with BeautifulSoup. After the data was collected and parsed, the data was converted into pandas dataframe object.

Data Collection – SpaceX API

- The data was collected using get request and parsed before being converted into .json and later dataframe. The data is filtered as we only need Falcon 9 launch data. The data was cleaned and the missing values was filled with the mean value of the respective column

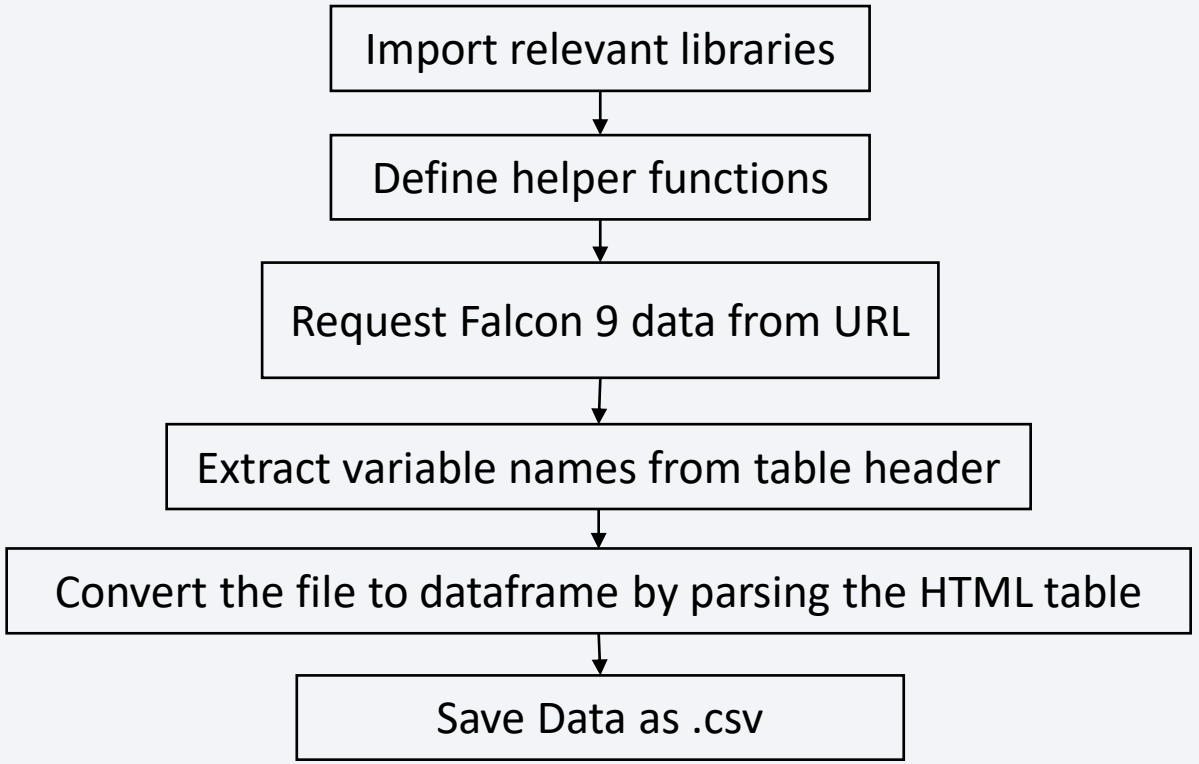
- https://github.com/rezafa8/ibmcapstone/blob/main/Data%20Collection-SpaceX%20API.ipynb



Data Collection - Scraping

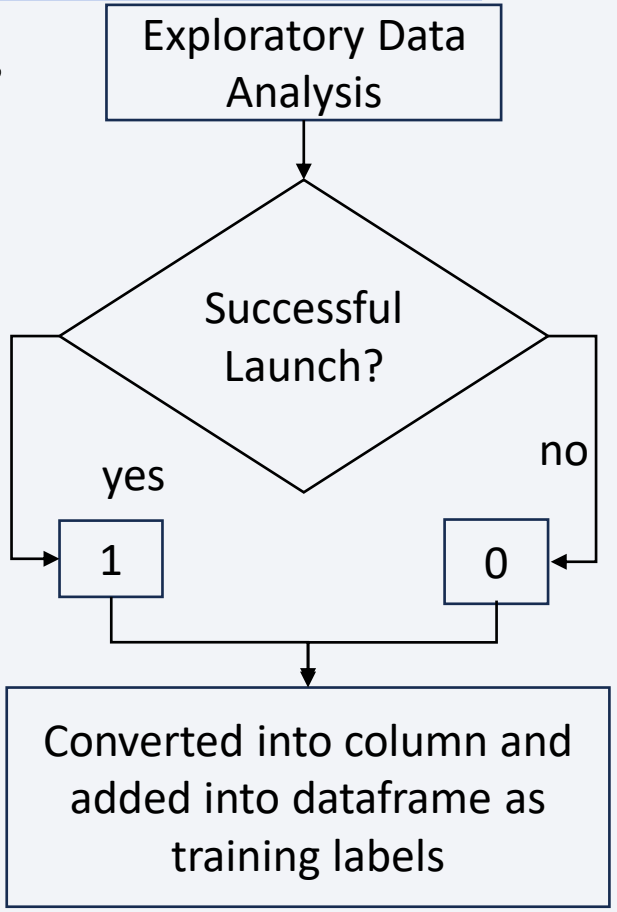
- Web scrapping using BeautifulSoup was performed at Wikipedia URL to extract Falcon 9 Launch Data
- The data was parsed and converted into dataframe

- https://github.com/rezafa8/ibmcapstone/blob/main/Web%20Scrapping.ipynb



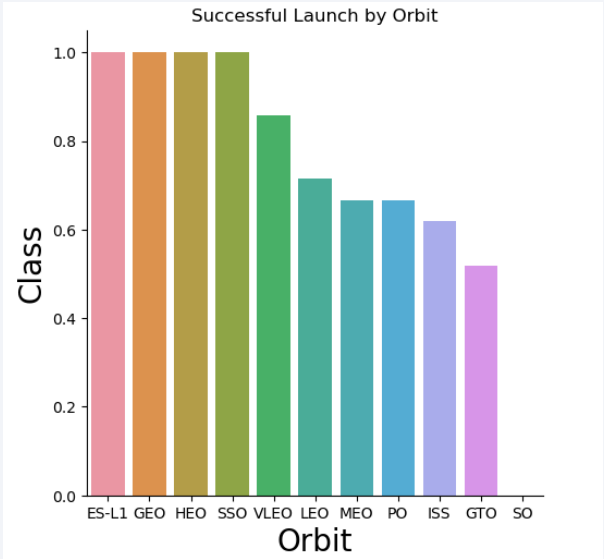
Data Wrangling

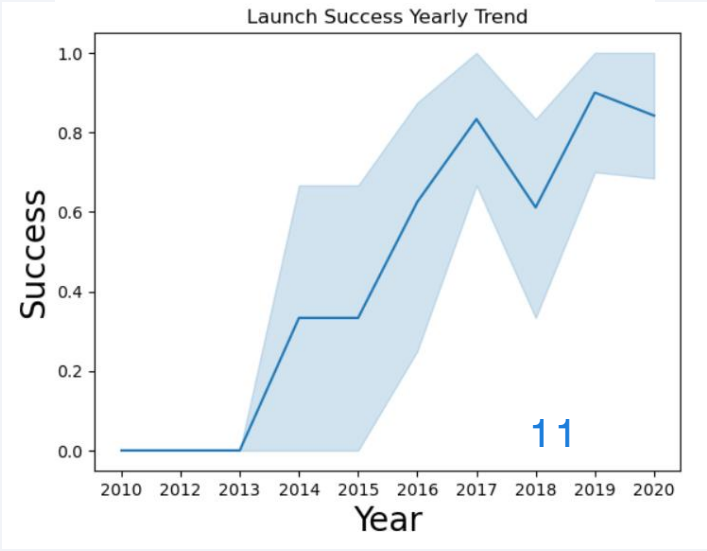
- Performing Exploratory Data Analysis and Determining Training Labels was the aim of data wrangling
- We check the percentage of missing values and datatypes
- Calculating the total launch from each launch site and orbit
- Determine and convert failed launch into a new binary class and convert it into a new column to be added into dataframe
- https://github.com/rezafa8/ibmcapstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- There are three main charts that are used for Exploratory Data Analysis: Scatter plot, line graph and bar graph.
- Scatter plot could show the relationship between two variables and is beneficial to determine the correlation between features in spaceX data frame. It is used to plot numerical variables
- Bar graph could be used to plot numerical variables and categorical variables. For example: the plot of percentage of successful launch based on the rocket's orbit
- It was found that the success rate increased throughout the years and the success rate is determined by several factors, including Orbit
- https://github.com/rezafa8/ibmcapstone/blob/main/EDA-Data%20Visualization.ipynb





EDA with SQL

- SQL is one of the most important database tools in industries
- We use SQL to do exploratory data analysis with the help of jupyter notebook
- From SQL, we could get for example:
  - Unique Launch Site
  - Average Payload Mass by certain booster version
  - First successful landing
  - Booster versions with maximum payload mass
  - Total successful and failure rocket launch
- https://github.com/rezafa8/ibmcapstone/blob/main/EDA-SQL.ipynb

Build an Interactive Map with Folium

- All launch site locations are mapped using markers and circles
- We differentiate the color for successful and failed rocket launch
- We could determine launch site with the most successful rocket launch rate
- We calculate the distance from the launch site with nearby infrastructures like highway, railway and cities
- https://github.com/rezafa8/ibmcapstone/blob/main/Folium.ipynb

Build a Dashboard with Plotly Dash

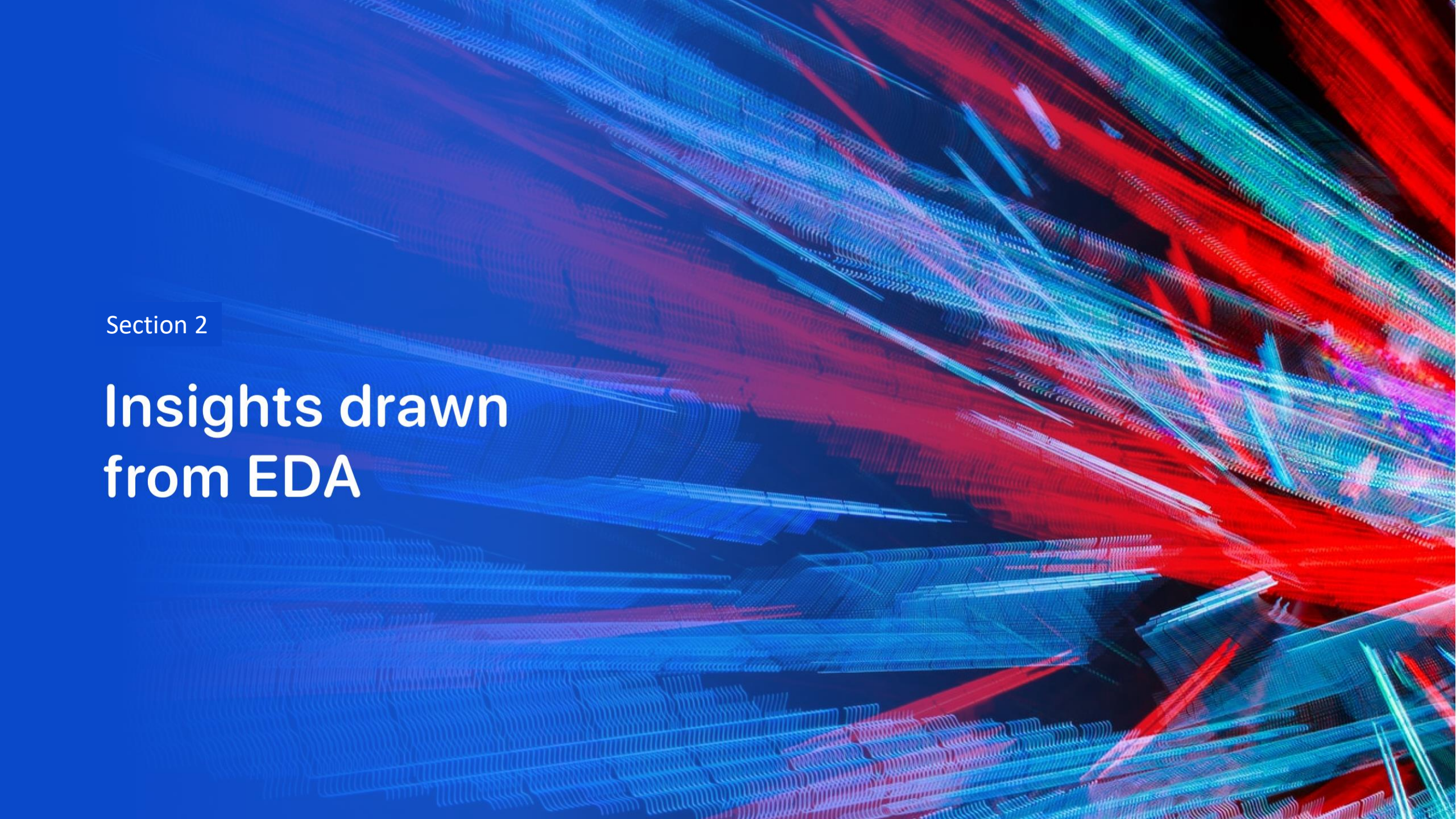
- Pie chart and scatter plot were added into dashboard. The data input could be chosen from 4 launch locations.
- The pie chart is important to show the proportion of total of successful rocket launch and successful rocket launch per launch site.
- Similarly, the scatter plot could highlight the importance of payload mass in kg to determine the success of rocket launch
- https://github.com/rezafa8/ibmcapstone/blob/main/dash-app.py

Predictive Analysis (Classification)

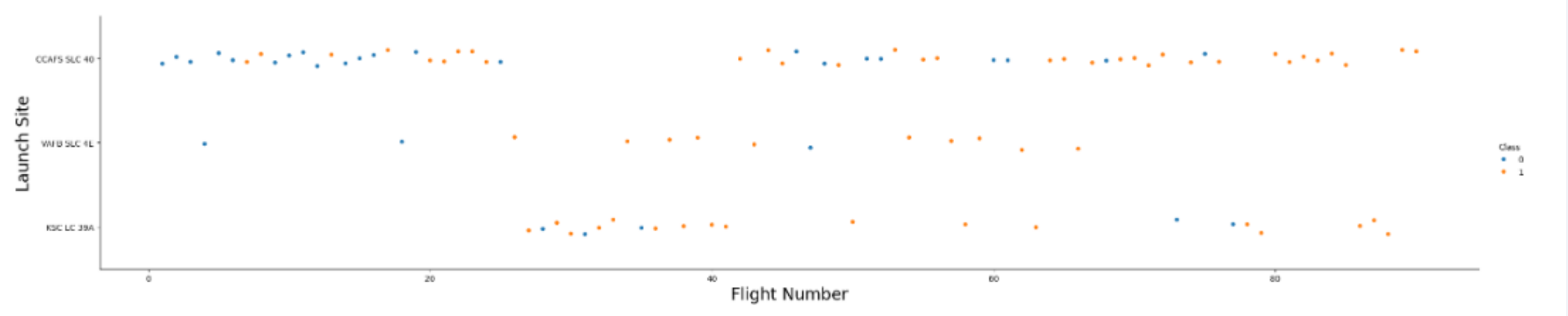
- The data were first imported into a dataframe using two different data. The training label (Y) is separated from the rest of the data and converted into numpy array.
- X parameters are then standardized to get the same weight for the datasets using StandardScaler() method.
- The data were then split into training and test dataset with 80% of the data were used for training.
- The training set were then analyzed using the following methods: linear regression, decision tree, SVM and K-Nearest Neighbor. The maximum accuracy was found by using Grid Search optimization method to calculate the optimum parameters for each method.
- Accuracy metrics were used to grade our model.
- All classification models perform similarly on the test set, achieving accuracy of 83.33% except Decision Tree with 72%.
- https://github.com/rezafa8/ibmcapstone/blob/main/Machine%20Learning.ipynb

Results

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

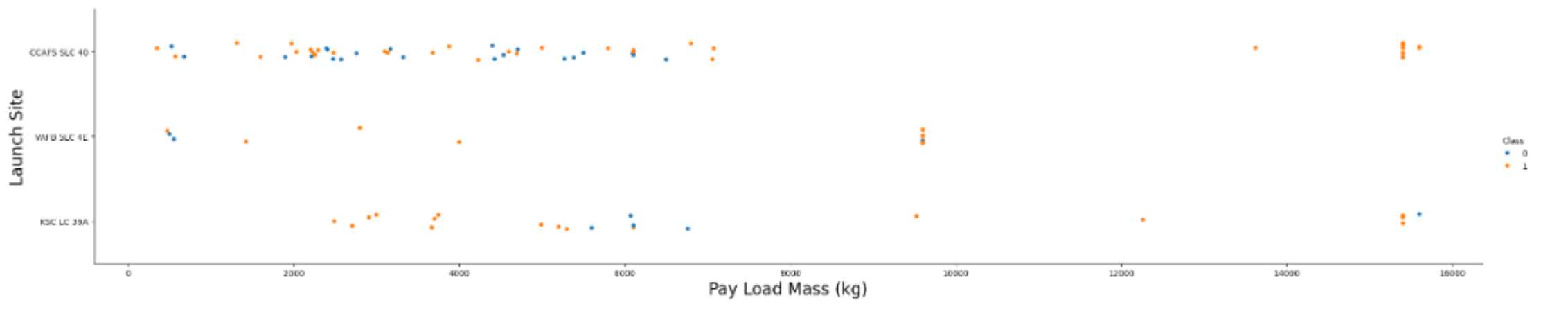

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The scatter plot shows the plot between Flight Number and Launch Site. The higher the flight number, the higher the success rate.

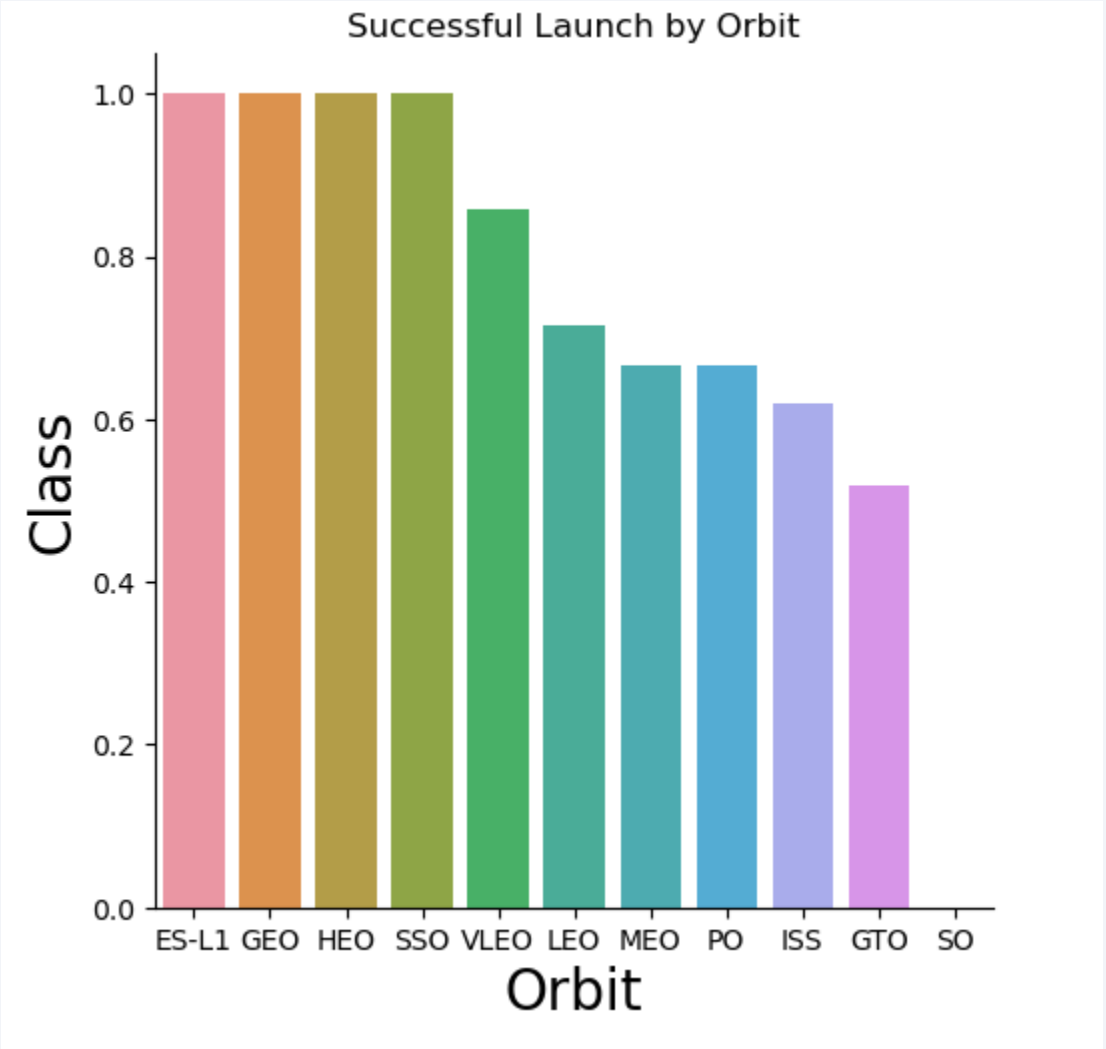
Payload vs. Launch Site



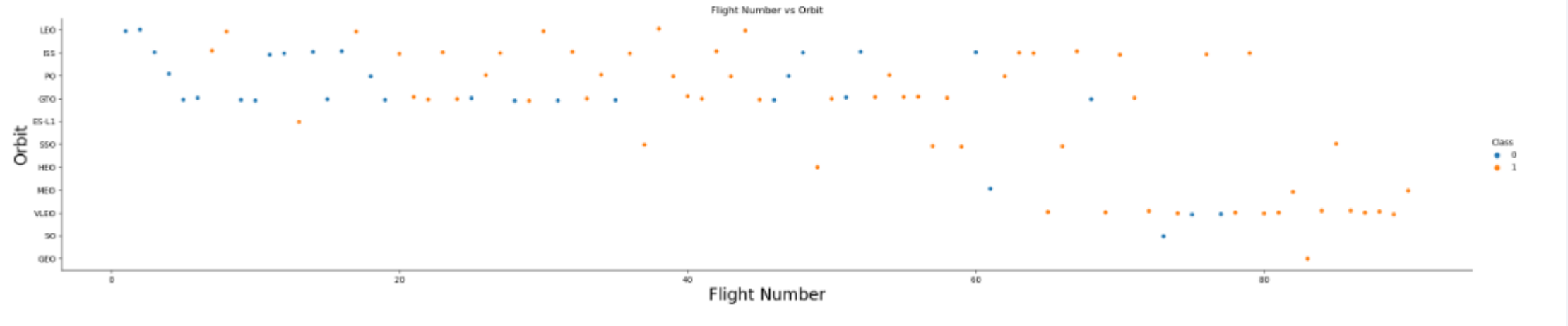
- The majority of Payload Mass for all Launch Site are less than 7000 kg.
- There are no rocket launch for VAFB-SLC 40 with payload mass of more than 10000 kg
- The higher the mass, the higher the success rate for CCAFS SLC-40

Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO and SSO have success launch rates of 100% compared to SO with 0%.
- The second highest mean success rate is at VLEO
- The mean success rate for all other orbit ranges between 50% to 70%.

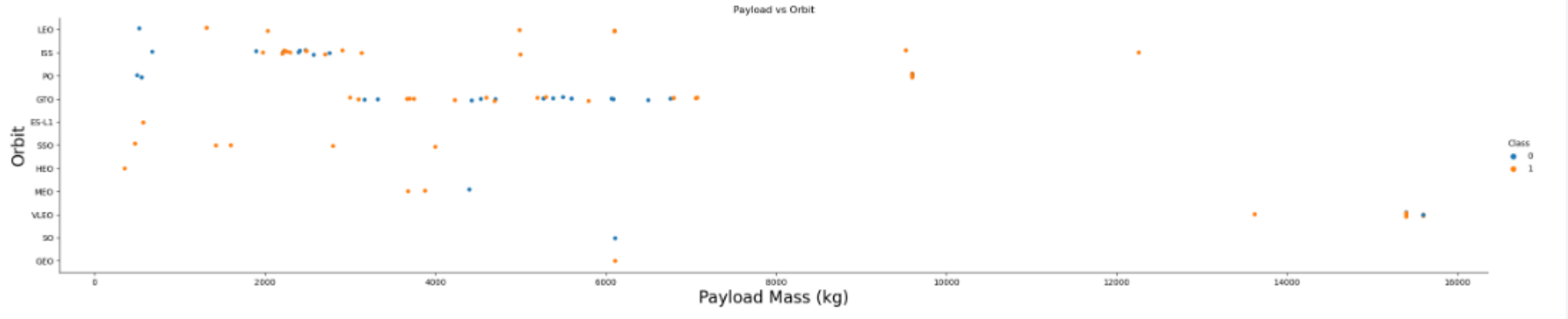


Flight Number vs. Orbit Type



- The success rate of Orbit LEO is correlated with the number of flights.
- However, there are nearly no correlation between the number of flights with orbit at GTO, shown by randomly successful launch

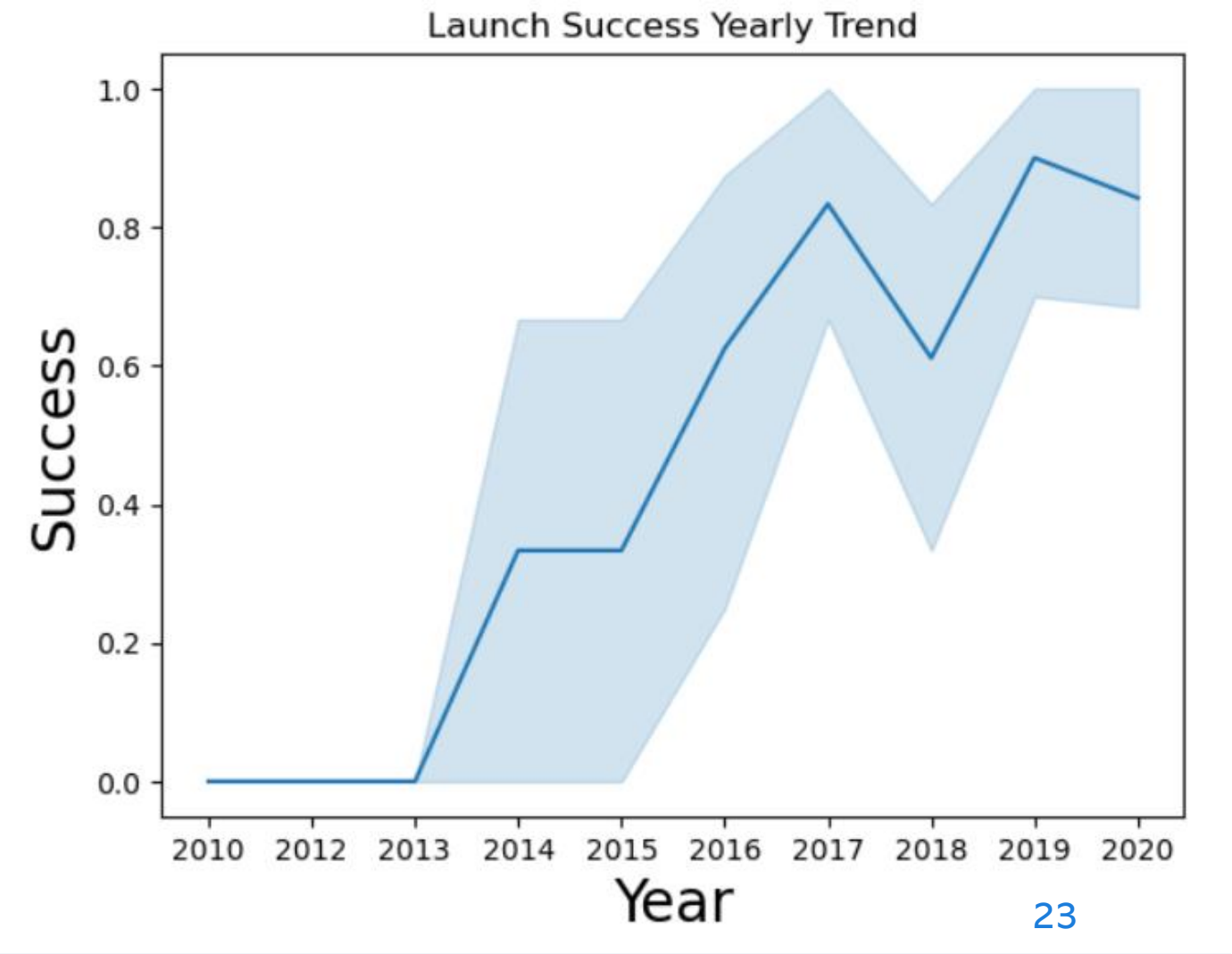
Payload vs. Orbit Type



- The heavier the payload, the more successful the launch in orbit LEO, PO and ISS
- On the other hand, the success and failure ratio for GTO is hard to determine as the success/failure ratio is randomly scattered

Launch Success Yearly Trend

- The launch success rate increased over the years from the year 2013 with anomaly in year 2018



All Launch Site Names

- %sql select distinct Launch_Site as Diff_Launch_Site from SPACEXTABLE was used to get the unique value for Launch Site
- There are four different launch sites: CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E and KSC LC-39A

Out[15]:

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

Launch Site Names Begin with 'CCA'

- We implement *like* query to find launch site names that begins with ‘CCA’ with the following query: %sql select * from SPACEXTABLE where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5

Out[16]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload value is 45596 kg for all booster launched by NASA (CRS)
- We use *sum* query for the column payload_mass_kg to get the desired result:
  %sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_kg from SPACEXTABLE where “Customer” like ‘NASA (CRS)’

Total_Payload_kg
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster F9 v1.1 is 2928.4 kg
- We use *avg* query for the column payload_mass_kg and *where* clause from Booster_Version to get the desired result: %sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTABLE where "Booster_Version" like 'F9 v1.1'

average_payload_mass
2928.4

First Successful Ground Landing Date

- The first successful landing happened in 2015-12-22
- We use *min* query on date column and *like* clause from Landing_Outcome to get the first successful landing date: %sql select min(Date) as first_successful_landing_date from SPACEXTABLE where "Landing_Outcome" like 'Success%'

Out[19]:

first_succssfull_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are four Booster Version that fulfill the criteria: F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2
- Distinct, where and like clause were used to get the desired results: %sql select distinct(Booster_Version) from SPACEXTABLE where "Landing_Outcome" like 'Success (drone ship)%' and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)

Out[35]:

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

Total Number of Successful and Failure Mission Outcomes

- There are total of 100 success launch as opposed to only 1 failure
- The following query is used: %sql select count(Mission_Outcome) as Total_Outcome, Mission_Outcome from spacextable group by Mission_Outcome

Out[17]:

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

Boosters Carried Maximum Payload

- There are 12 Boosters that have carried the highest payload value.
- Subqueries were used to extract boosters that carried maximum payload mass: %sql select booster_version from spacextable where PAYLOAD_MASS__KG_ == (select max(PAYLOAD_MASS__KG_) from spacextable)

Out[18]:

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

- There are two failed outcome in drone ship for the year 2015, happened in October and April 2015
- Queries were used to extract month and year from the table. Additionally, where clause was also used: %sql select substr(date,6,2) as Month_in_2015, landing_outcome, Booster_Version, Launch_Site from spacextable where substr(Date,0,5)='2015' and "Landing_Outcome"=='Failure (drone ship)'

Out[35]:

Month_in_2015	Landing_Outcome	Booster_Version	Launch_Site
10	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

- There are 8 different landing outcomes from 2010-06-04 to 2017-03-20.
- Count query, where and group by clause were used to get the data: %sql select count(Landing_outcome), landing_outcome from spacextable where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_outcome) desc

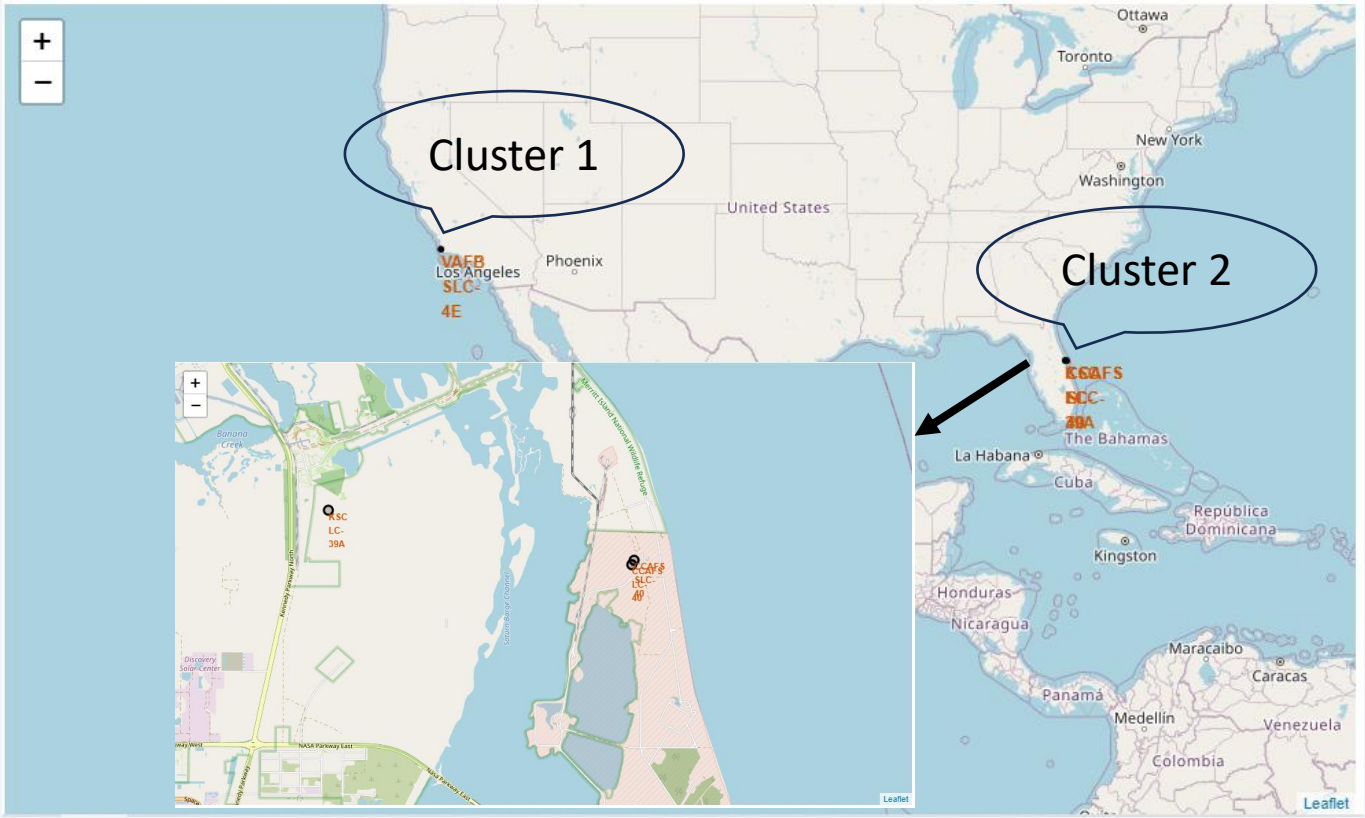
Out[20]:

count(Landing_outcome)	Landing_Outcome
10	No attempt
5	Success (ground pad)
5	Success (drone ship)
5	Failure (drone ship)
3	Controlled (ocean)
2	Uncontrolled (ocean)
1	Precluded (drone ship)
1	Failure (parachute)


Section 3
Launch Sites
Proximities Analysis

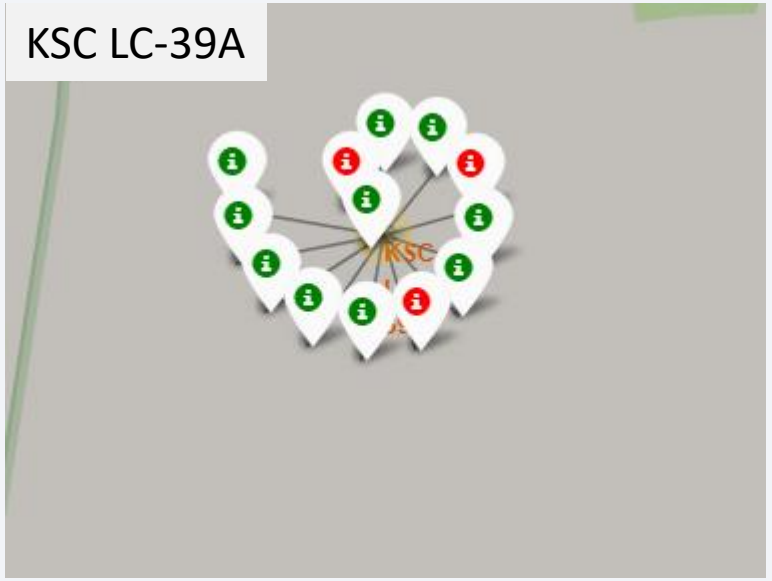
LAUNCH SITE LOCATIONS

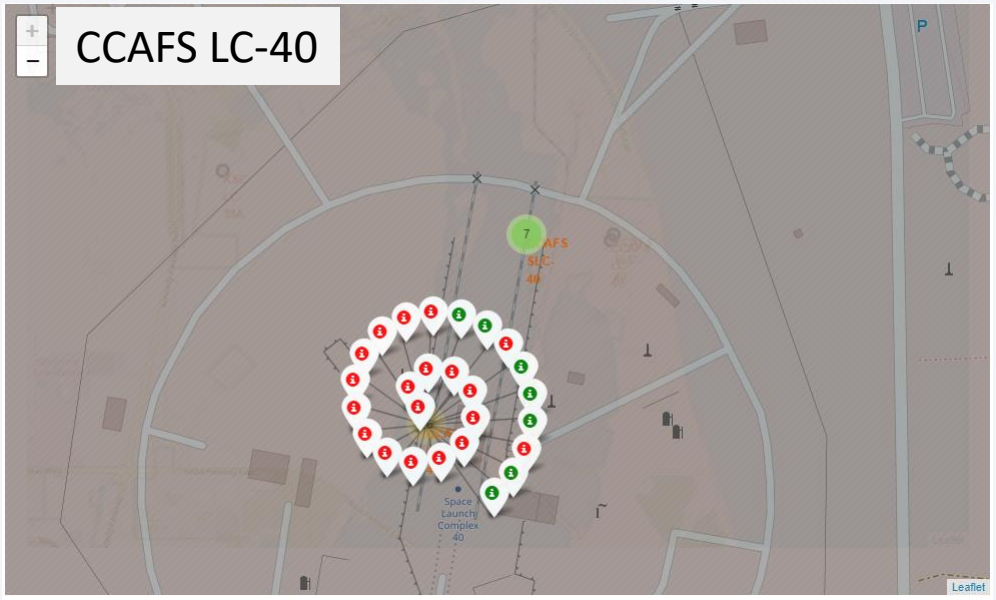
- There are two main clusters of launch site for SpaceX, located near Los Angeles (VAFB SLC-4E) and located to the east of Orlando (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A). Both are located near shore

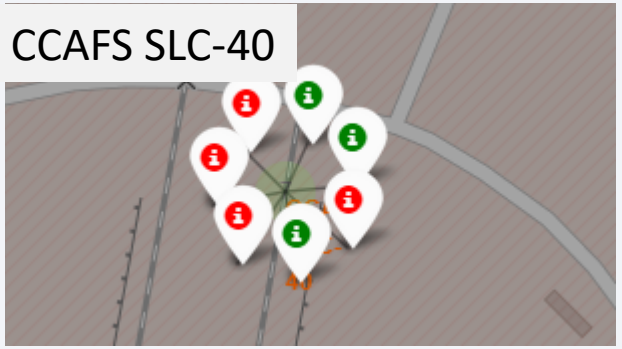


Investigating The Success Rate for Launch Site

- The success rate for each launch site could be investigated by making color-labeled marker. The green color shows successful launch as opposed to red color, which shows failed launch
- There are only 7 successful launches for Launch Site CCAFS LC-40 out of 26 launches

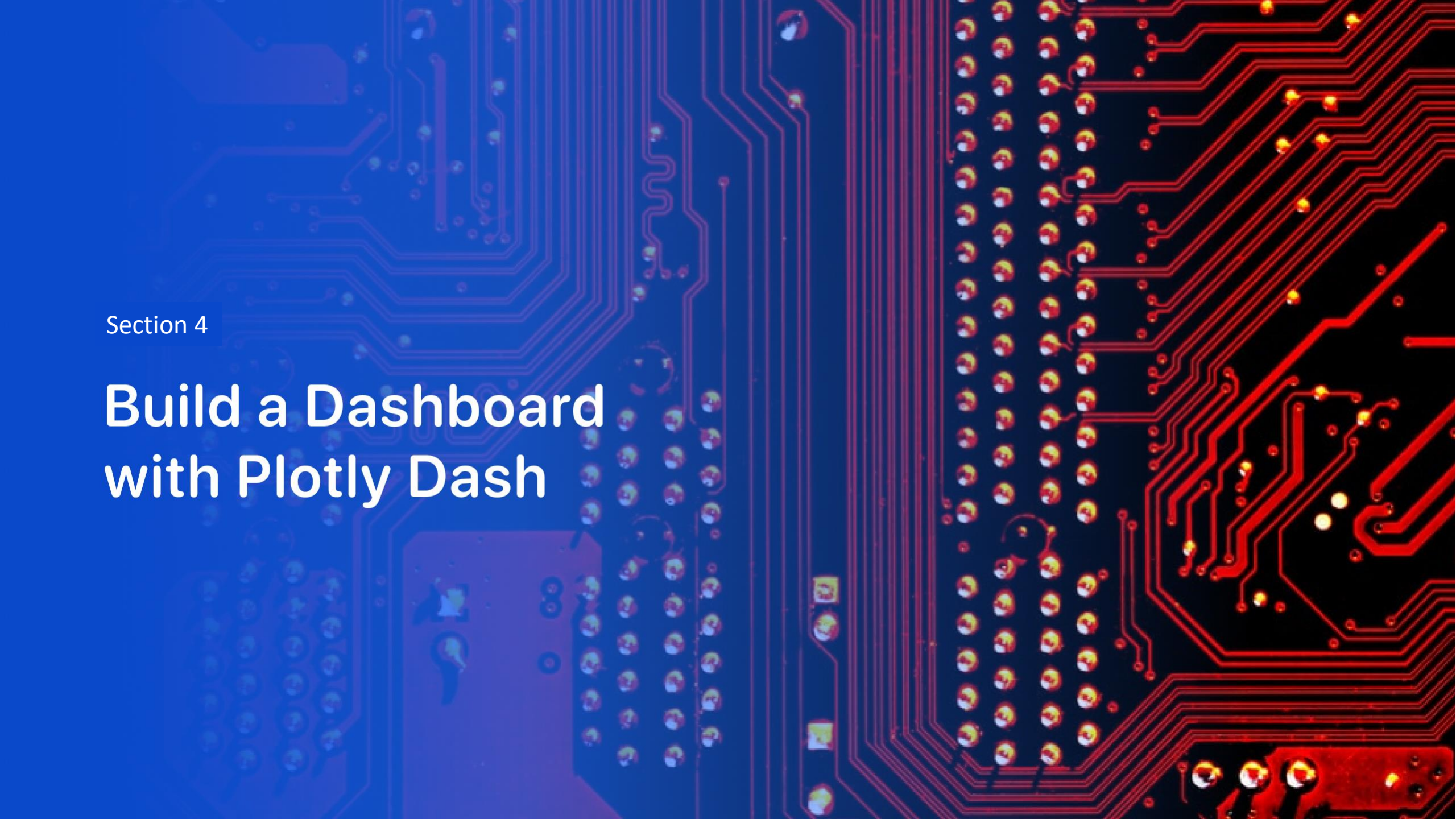






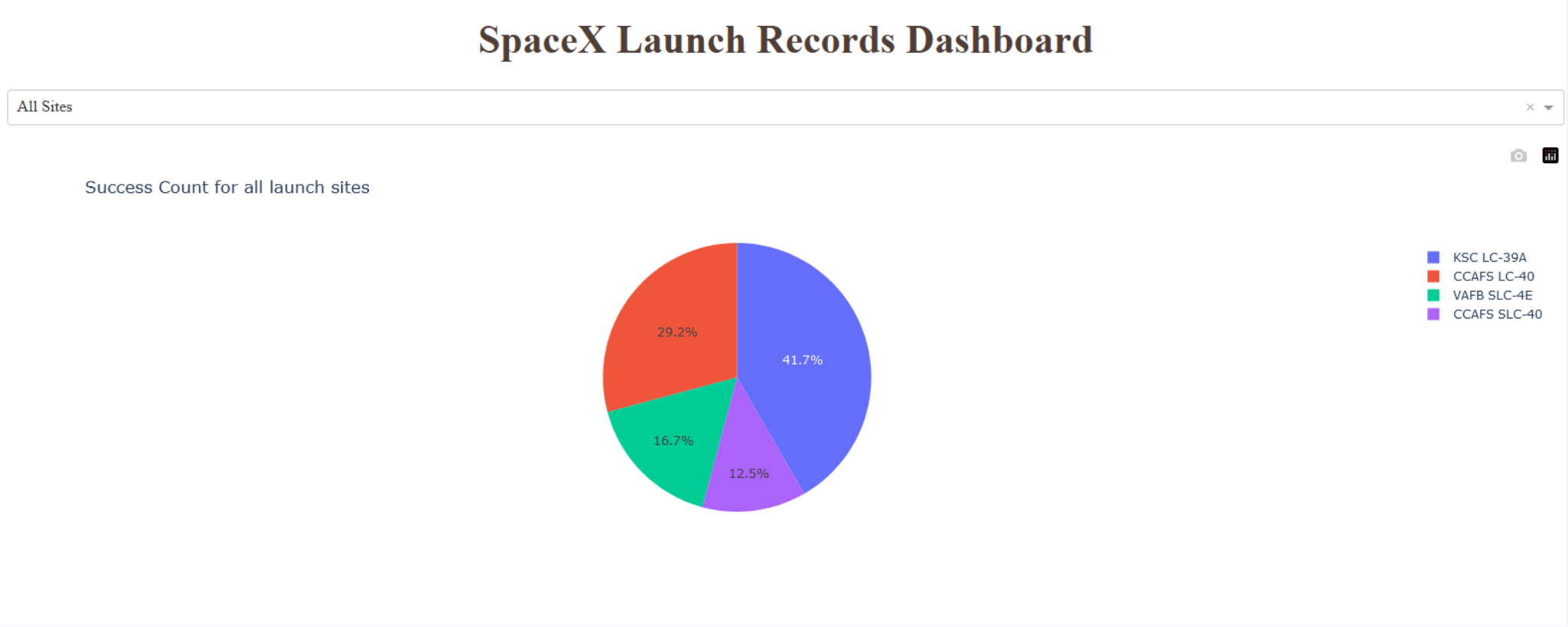
Distance from launch site to its proximities

- The distance between the launch site and its proximities could be calculated automatically using Folium. The proximities could include, but not limited to railways, highways, coastline, cities and airport.
- The distance to railways for selected launch site is 1.28 km, to highway 0.55 km and to coastline 0.83 km. From Melbourne city in the south, the distance is more than 50 km. The launch is situated far away from cities and near coastline, railways and highway


Section 4
Build a Dashboard with Plotly Dash

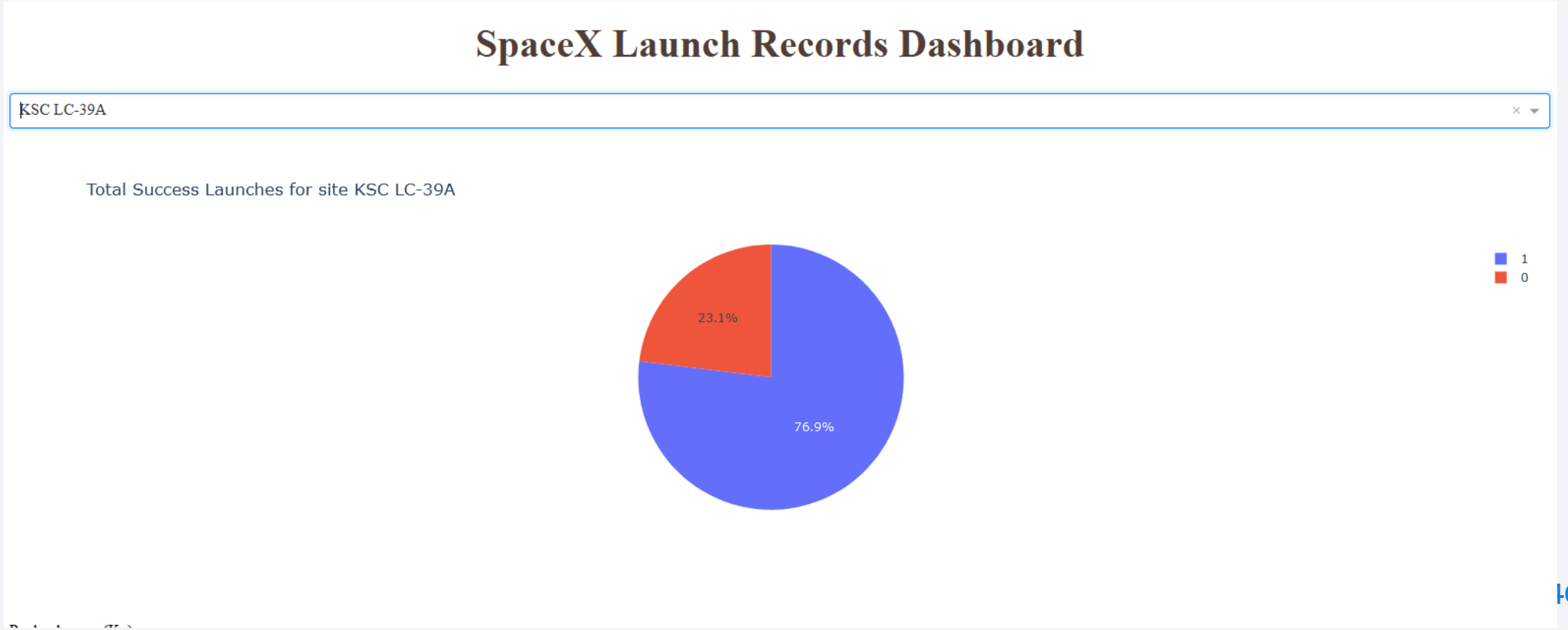
Success Percentage by Launch Site

- The highest number of successful launch comes from KSC LC-39A with 41.7%



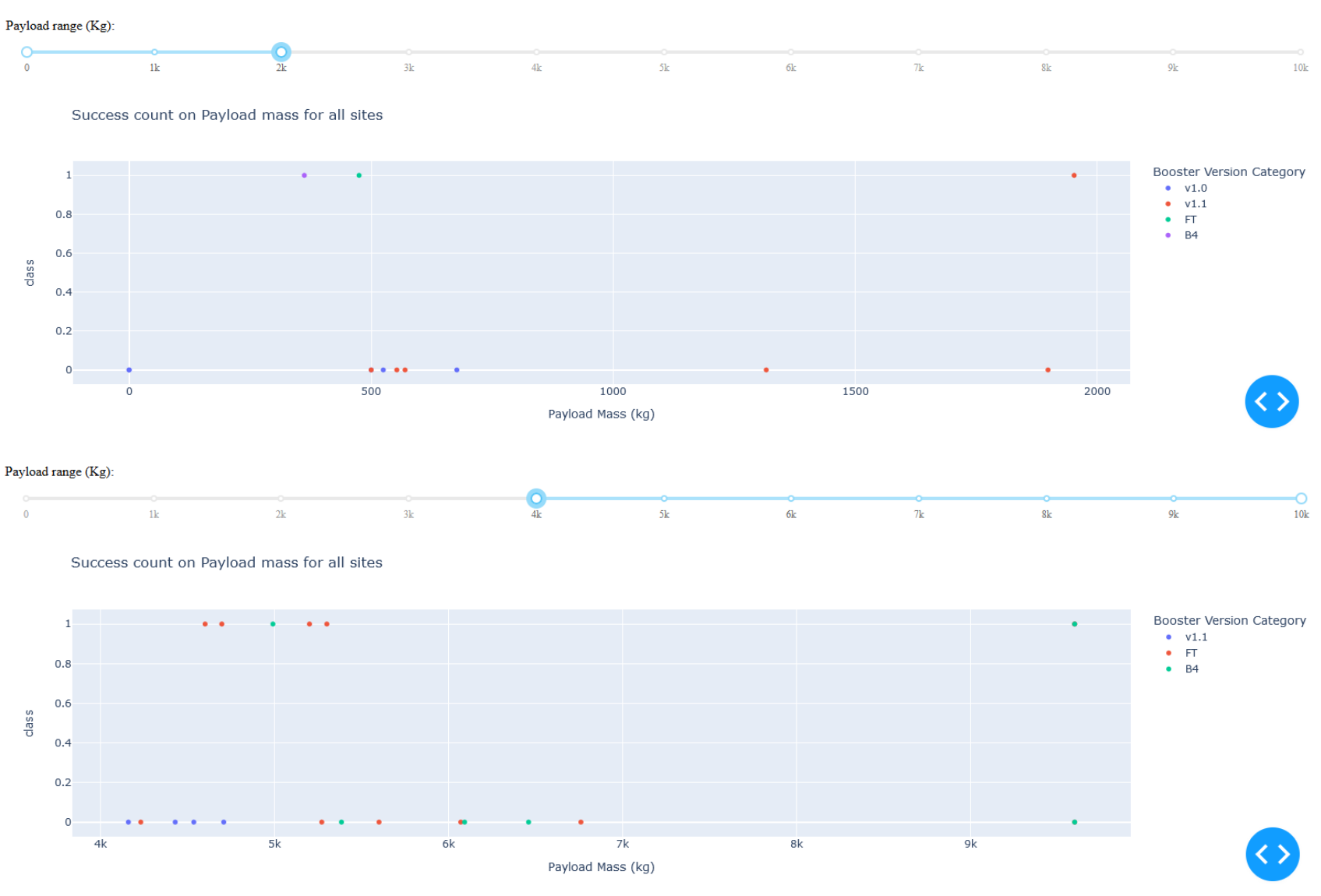
Launch Site with the highest successful percentage

- 76.9% launch from KSC LC-39A was successful with only 23.1% failure rate



Payload Mass and Success Rate Scatter Plot

- The heavier payload mass the lower the success rate

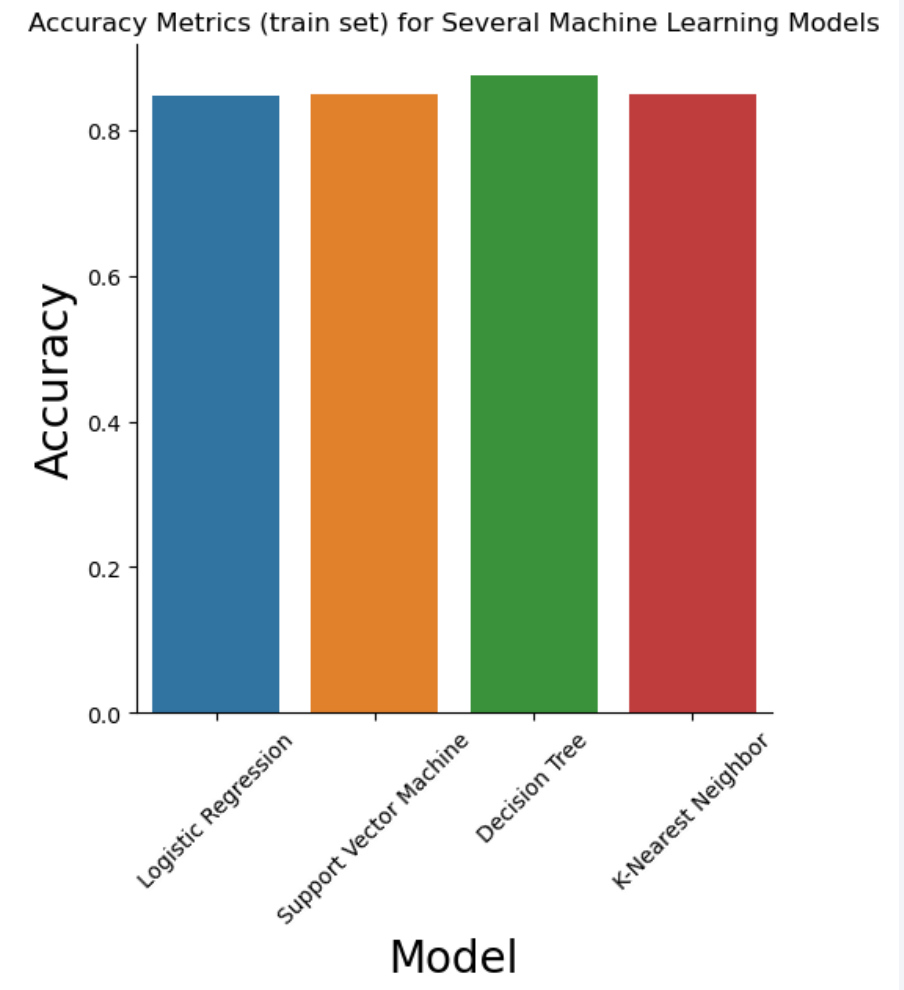


Section 5

Predictive Analysis (Classification)

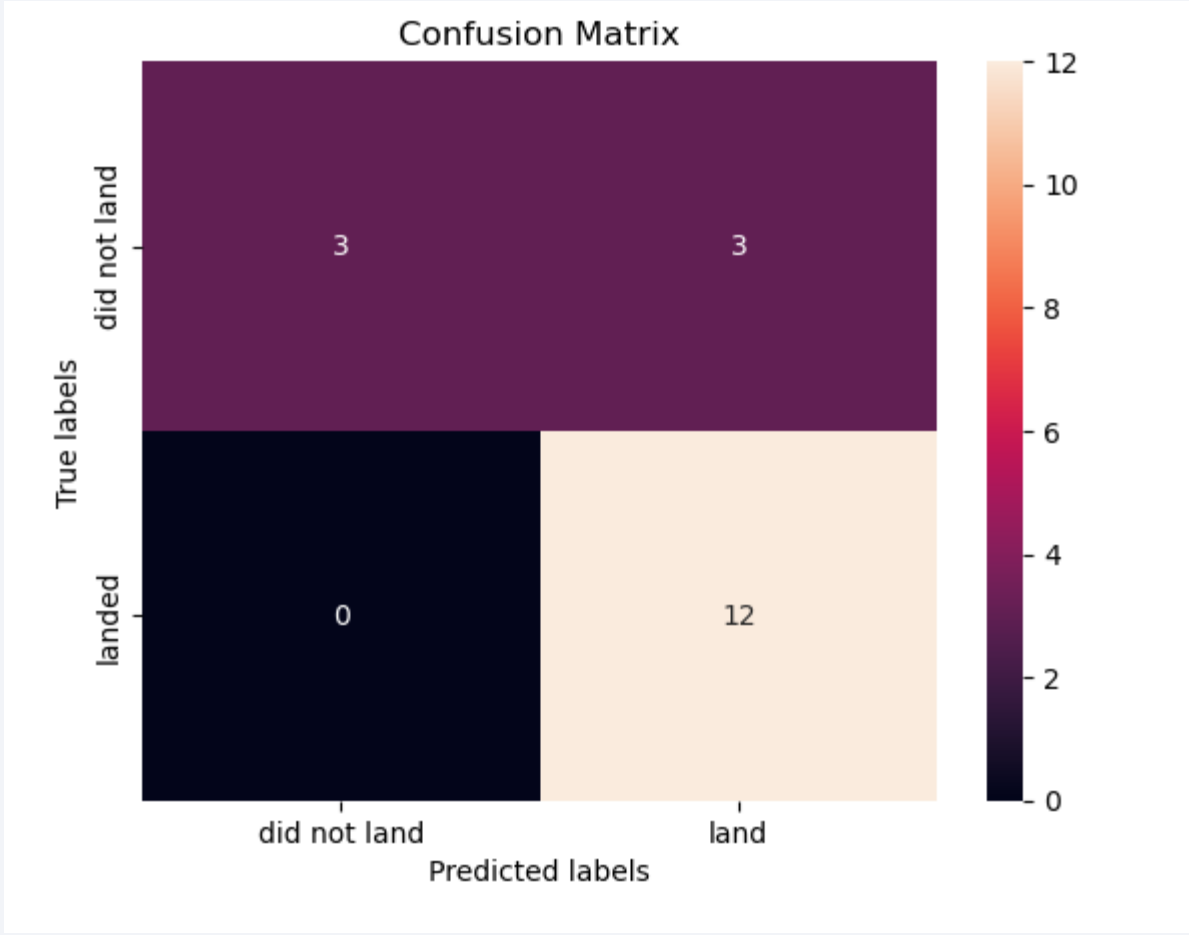
Classification Accuracy

- Decision Tree has the best accuracy for training data, reaching 87% accuracy. However, the accuracy decrease to 72 for the test set.
- Logistic Regression, SVM and K-Nearest Neighbor have the same accuracy for test dataset with 0.833.



Confusion Matrix

- Logistic Regression, SVM and K-Nearest Neighbor have the same confusion matrix (all score 83.33% accuracy).
- There are 3 false positive prediction (predicted landed, actual did not land)



Conclusions

- The success rate has a positive correlation with Flight Number
- Orbit ES-L1, GEO, HEO and SSO have success launch rates of 100% compared to SO with 0%.
- The success rate has been increasing since 2013
- Launch Sites are located near shore but far from cities
- KSC LC-39A has the highest success rate
- Decision Tree has the best accuracy for training data, reaching 87% accuracy. However, the accuracy decrease to 72 for the test set. Other models score 83.33% accuracy for the test set.

Thank you!