IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nwafor Patrick
10th March 2022

Outline

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

Executive Summary

- This research tries to predict whether the first stage of a falcon 9 space rocket will land or not. Data were gotten from the SpaceX web API which was cleaned and preprocessed to make ready to apply ML models on.
- Four classification ML models were used to predict the landing outcome of preprocessed test data and they all had the same accuracy score of 0.8334.

Introduction

- 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.
- if we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

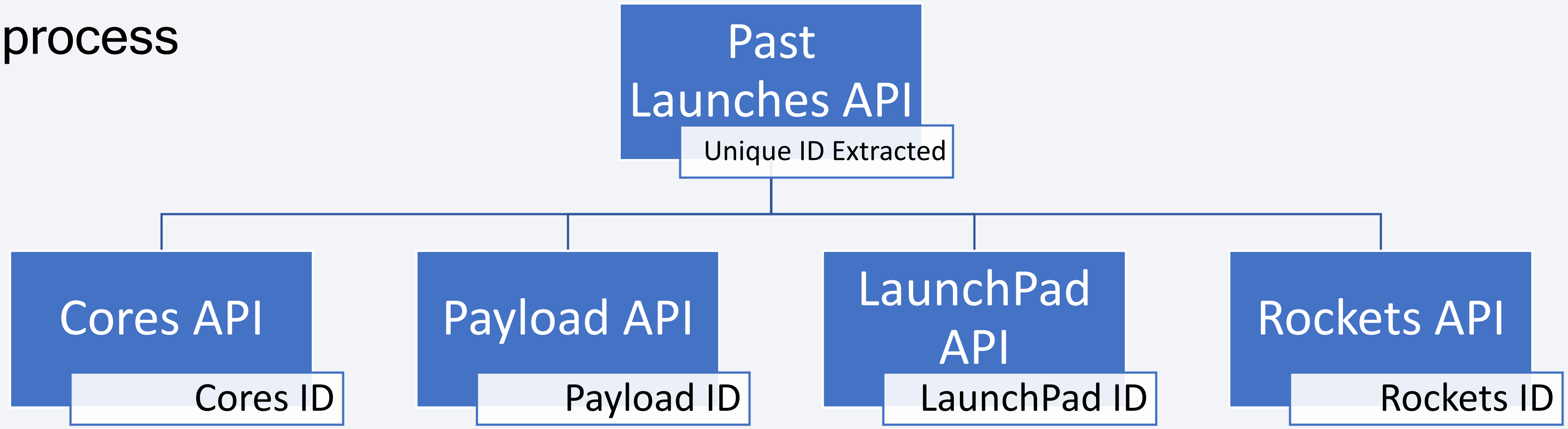
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected from the SpaceX web API ("https://api.spacexdata.com/v4/launches/past")
- First, an API call was made to the website above using python request.get(URL) method, which then returned a response.
- The response was parsed into a Pandas dataframe frame using the pd.json_normalize(response.json()) method.
- The unique ID for the Launchpad, Cores, Rockets and Payload were extracted from the Dataframe. These ID were used to further query the full details of the selected features
- Finally, the features were put together into a single Pandas Dataframe. This table formed the basis of our future manipulations and analysis.

Data Collection - SpaceX API

- GitHub Data Collection URL
- Click the link to view the data collection process



Data Collection - Scraping

- GitHub Data Collection with Web Scraping URL.

- Get the html text data using requests.get(url).text method
- Parse the html text to the beautiful soup object
- Extract all the columns from the table in the beautiful soup object
- Create an empty dictionary with keys as extracted column names
- Extract all the rows of the table and append to the dictionary
- Finally convert the dictionary to a Pandas DataFrame

Data Wrangling

Data Wrangling GitHub Link

- Read the data to a Pandas DataFrame using the pd.read_csv() method
- Then checked for null values in the df and data types of each column to ensure they are correct
- Got the unique landing outcomes of the df
- A new column called Class was created and all bad landing outcome were assigned a value 0 and good landing outcome 1

EDA with Data Visualization

- A Cartesian plot of flight number and Payload with a hue of Class to show the relationship between flight number and payload
- A Cartesian plot of Launch Site and Payload with a hue of Class to show the relationship between Launch Site and payload
- A Cartesian plot of Orbit type and Payload with a hue of Class to show the relationship between Orbit type and payload
- A bar plot to show the successful landing rate of each flight to a particular orbit.
- A line plot so show the trend of the successful landing rate per year.
- [GitHub URL of EDA with Data Visualization]

EDA with SQL

- Queried the unique launch site from db2 database
- Queried 5 records where launch sites begin with the string 'CCA'
- Queried the total payload mass carried by boosters launched by NASA (CRS)
- Queried average payload mass carried by booster version F9 v1.1
- Queried the date when the first successful landing outcome in ground pad was acheived.
- Queried the total number of successful and failure mission outcomes

EDA with SQL

- Queried the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Queried the names of the booster versions which have carried the maximum payload mass
- Queried the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranked 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 Link for EDA with SQL

Build an Interactive Map with Folium

- A folium map was created with the initial coordinates to NASA Johnson Space Center
- Various map objects were added to the map such as Circles, Markers, Mouse Position, Marker clusters and Polyline
- The Circles were used to mark the launch sites
- Polylines was used to draw a line from a launch site to another location such as an high way while also indicating the distance in kilometers
- Marker clusters were used to groups markers indicating the landing outcomes of each flights for each launch sites.
- Mouse position enabled allows us view the latitude and longitude at any point on the map where the mouse hovers.
- GitHub URL Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Added a pie chart and scatter plot in the dashboard
- The pie chart shows all the landing sites with their successful landing rate or it should a single launch site with their success and failure rate
- Scatter plot shows the relationship between the payload mass and the landing outcome with a hue of the booster version.
- Added a dropdown interactivity to enable you select a particular launch or all launch site for the Pie chat.
- Added a slider to select a range of payload mass you want to be displayed on the scatter plot against the landing outcome.
- GitHub URL for Dashboard with Plotly Dash

Predictive Analysis (Classification)

- Imported our preprocessed data set and split the dependent dataset(“Class”) from the independent features
- Next we normalized the independent features(X) dataset using the standard scaler method.
- Next we split the X and Y data set into train and test data.
- With the data ready to apply ML models, we choose classification models because we are trying to predict a categorical variable.
- GridSearchCV was used to run the following ML models: Decision tree, Logistics Regression, Support Vector Machine(SVM) and K Nearest Neighbor (KNN).
- Best parameter for each model was gotten from GridSearchCV.best_params_ attribute
- The model is then tested with the test data to determine its accuracy (Score)
- GitHub URL for Predictive Analysis

Predictive Analysis (Classification)

Imported data set and split to X and Y

Normalized the X data using standard scaler

Split X and Y data set into train and test data.

Fit the X and Y train data to the ML Model using GridSearchCV

Determine best parameter using GridSearchCV.best_params_ attribute

Predict the Y data from the X test data

Determine the accuracy(Score) of the predicted data

Model evaluation of each ML model by F1_SCORE

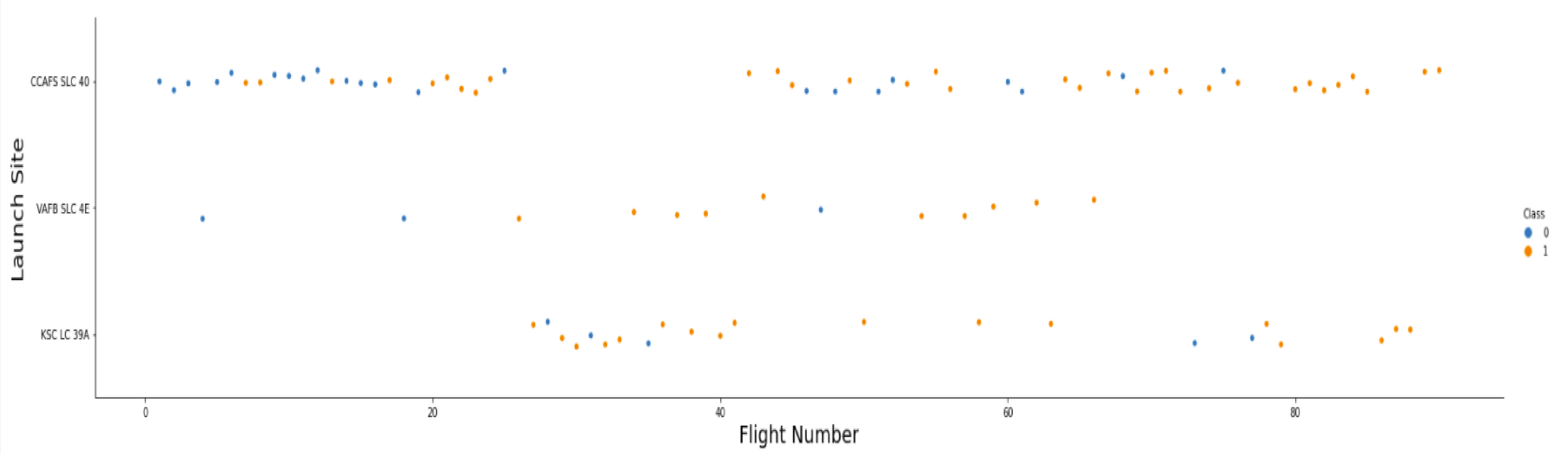
Results

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

Section 2

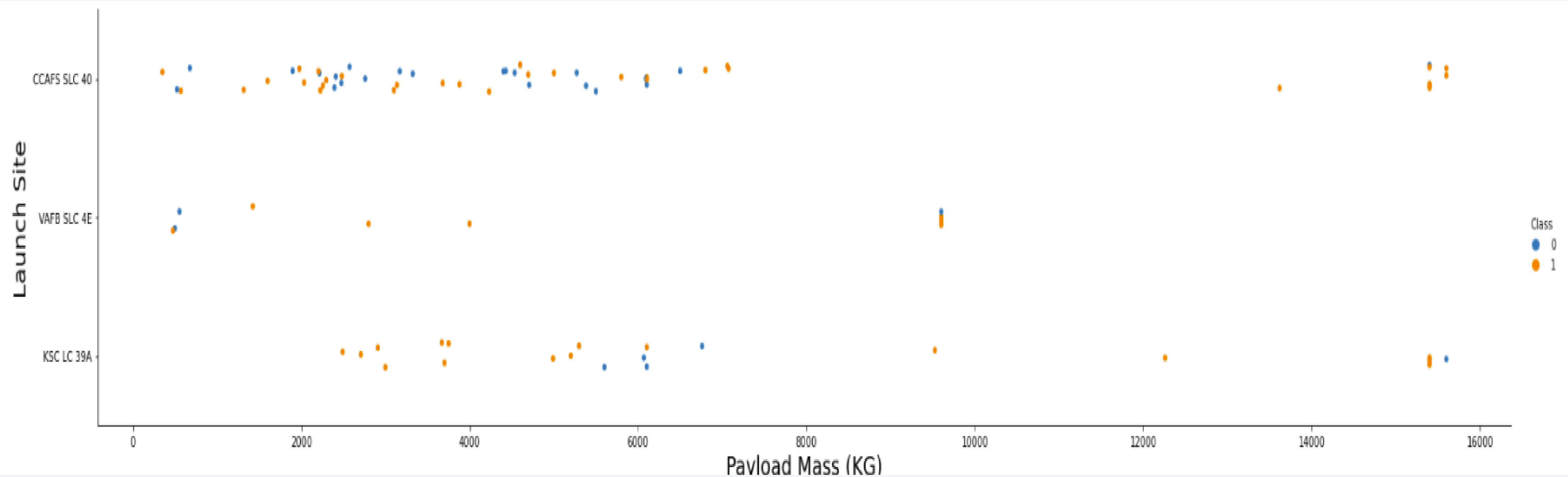
Insights drawn from EDA

Flight Number vs. Launch Site



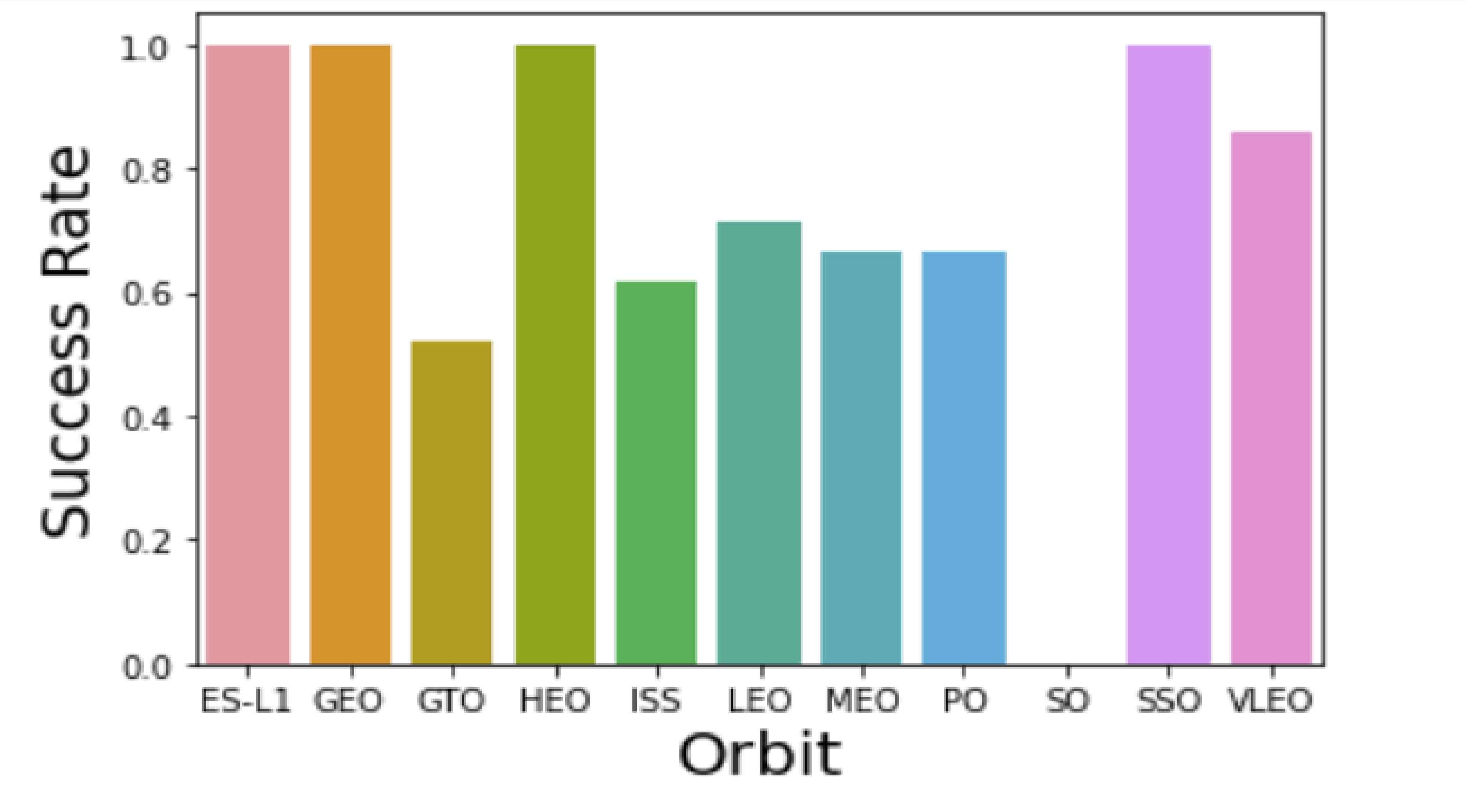
From the Cartesian plot above, it can be deduced that as the flight number increased, the Landing outcome success rate also increased for each launch sites

Payload vs. Launch Site



For each Launch Site aside CCAFS SLC 40, we can see a clear relationship between the payload mass and Launch Site. As the payload mass increased so did the success rate. No payload greater than 10000KG was used on the VAFB SLC 4E Launch Site.

Success Rate vs. Orbit Type

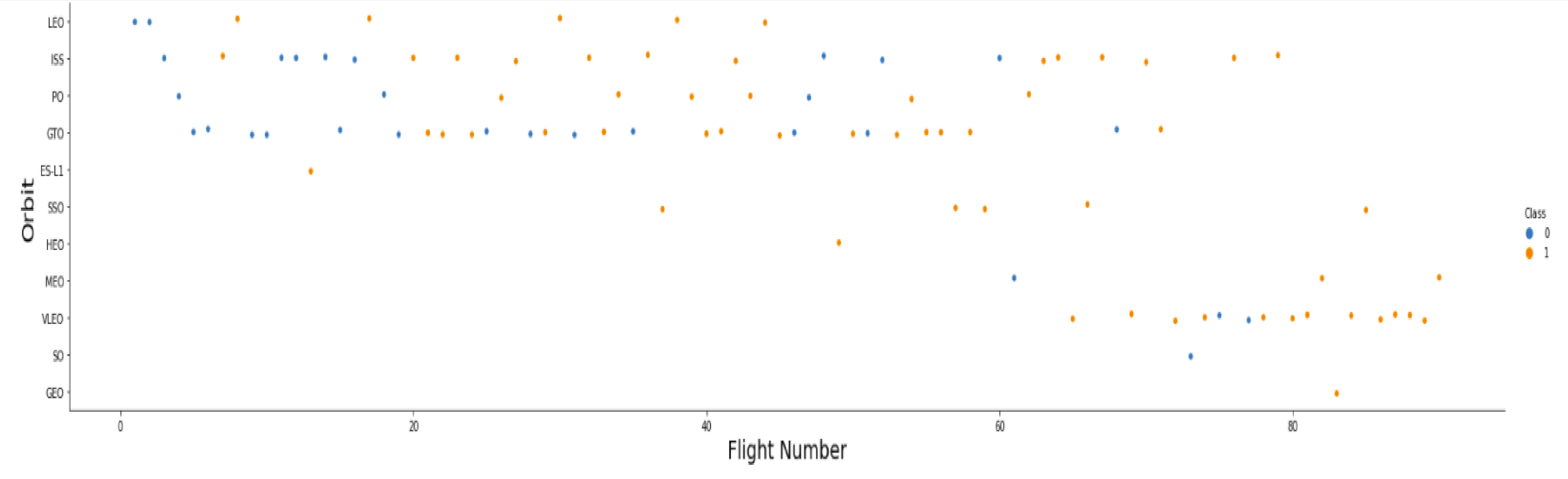


Orbit ES-L1, GEO,HEO and SSO
Had a successful
Landing rate of
100%

Orbit SO had a
Successful
Landing rate of 0%

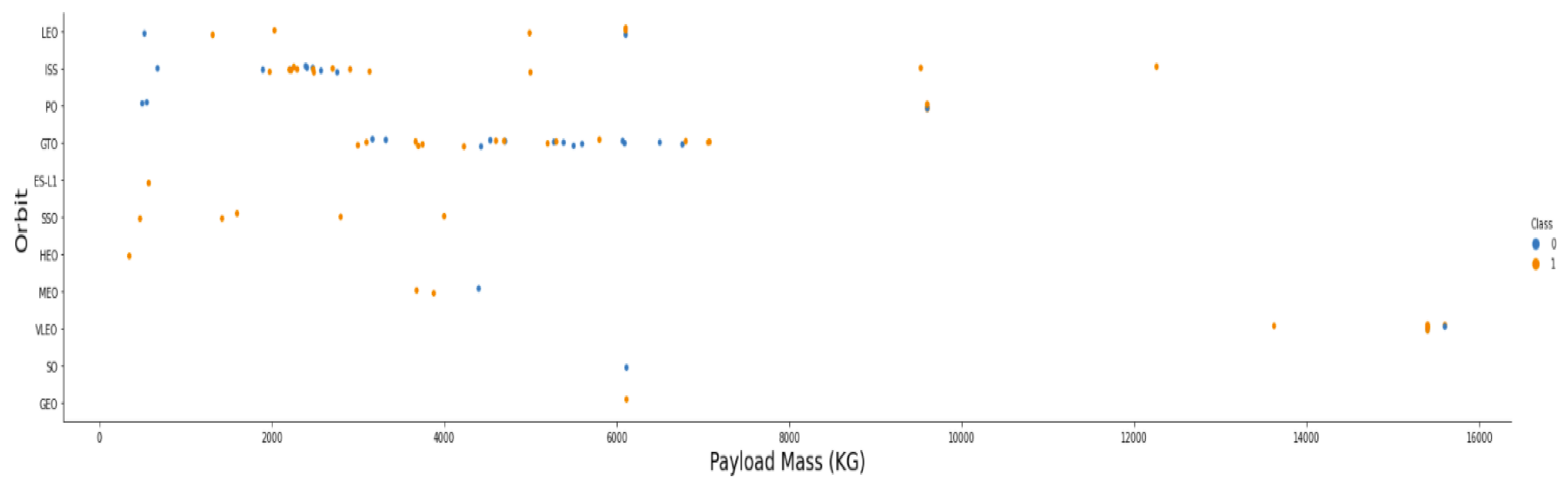
The remaining orbit
had a successful
Landing rate between
50 to 85%

Flight Number vs. Orbit Type



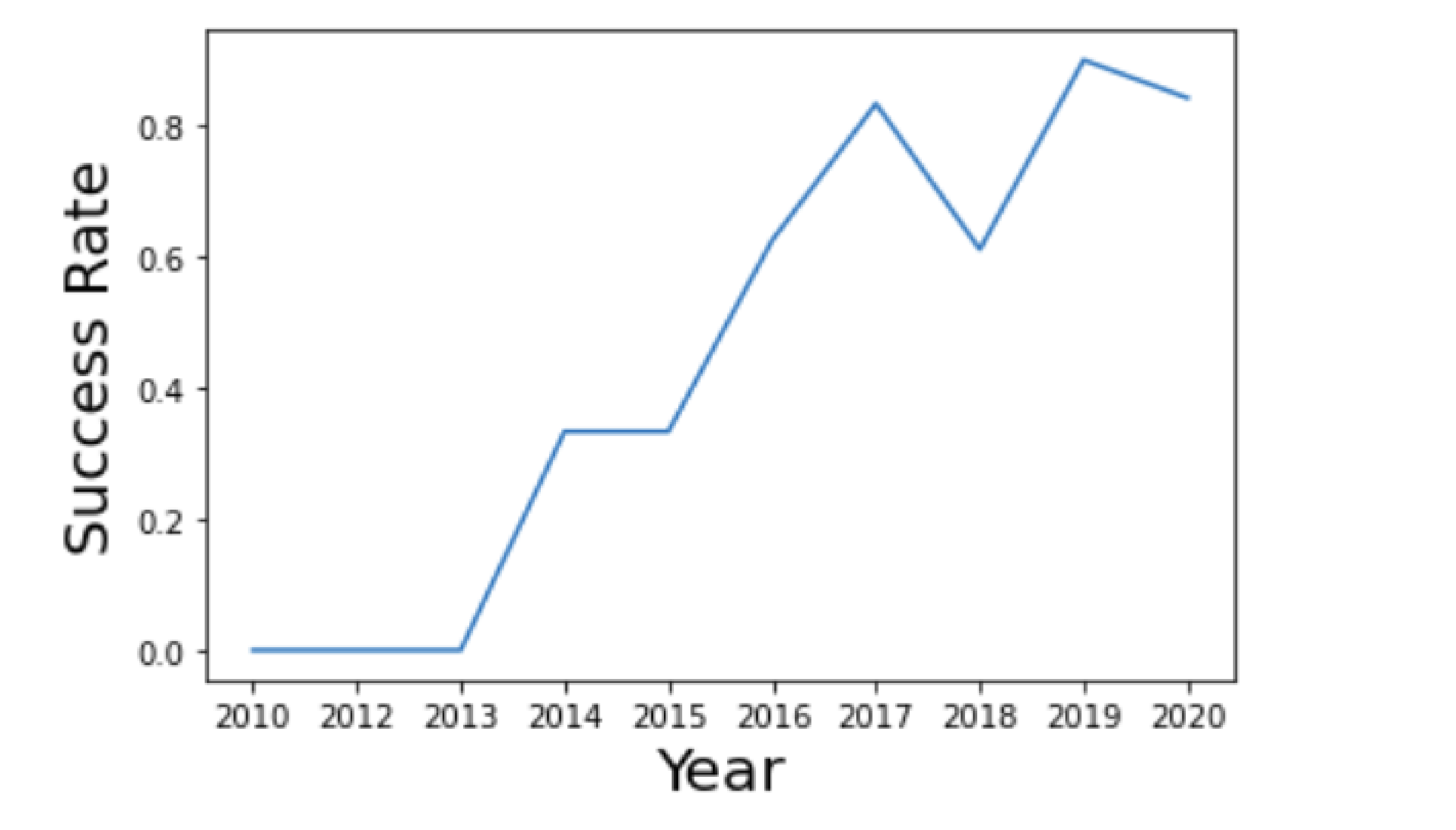
From the above plot, it can be seen that the LEO orbit Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for PO,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



As seen on the graph The landing success rate Kept increasing from 2013 To 2020.

There was also a dip in year 2018

All Launch Site Names

There are four Launch sites in the dataset which was gotten by using SQL Unique keyword on the Launch site column. The Launch sites includes:

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

Launch Site Names Begin with 'CCA'

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

The above result contains the first five records with launch sites that begins ‘CCA’. The query was done using SQL “LAUNCH_SITE LIKE 'CCA%‘” as the WHERE condition and setting the limit as 5

Total Payload Mass

- The total payload carried by boosters from NASA (CRS) is 45596 KG
- This is gotten using the SQL sum function on the payload mass column where customer is “NASA (CRS) ”

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928 KG
- This is gotten using the SQL avg function on the payload mass column where booster_version is “F9 v1.1”

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 22nd December 2015
- This is gotten using the SQL min function on the date column where landing outcome is “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- 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

S/N	Mission Outcome	Frequency
1	Failure (in flight)	1
2	Success	99
3	Success (Payload status unclear)	1

In total, there were 100 successful mission outcomes (including the mission with unclear payload status) and just one failure

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass includes:

- 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

The result was gotten using SQL subquery

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 includes:

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

- There are two failed landing outcomes on drone ship for the year 2020 both Launched from the same Launch Site

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

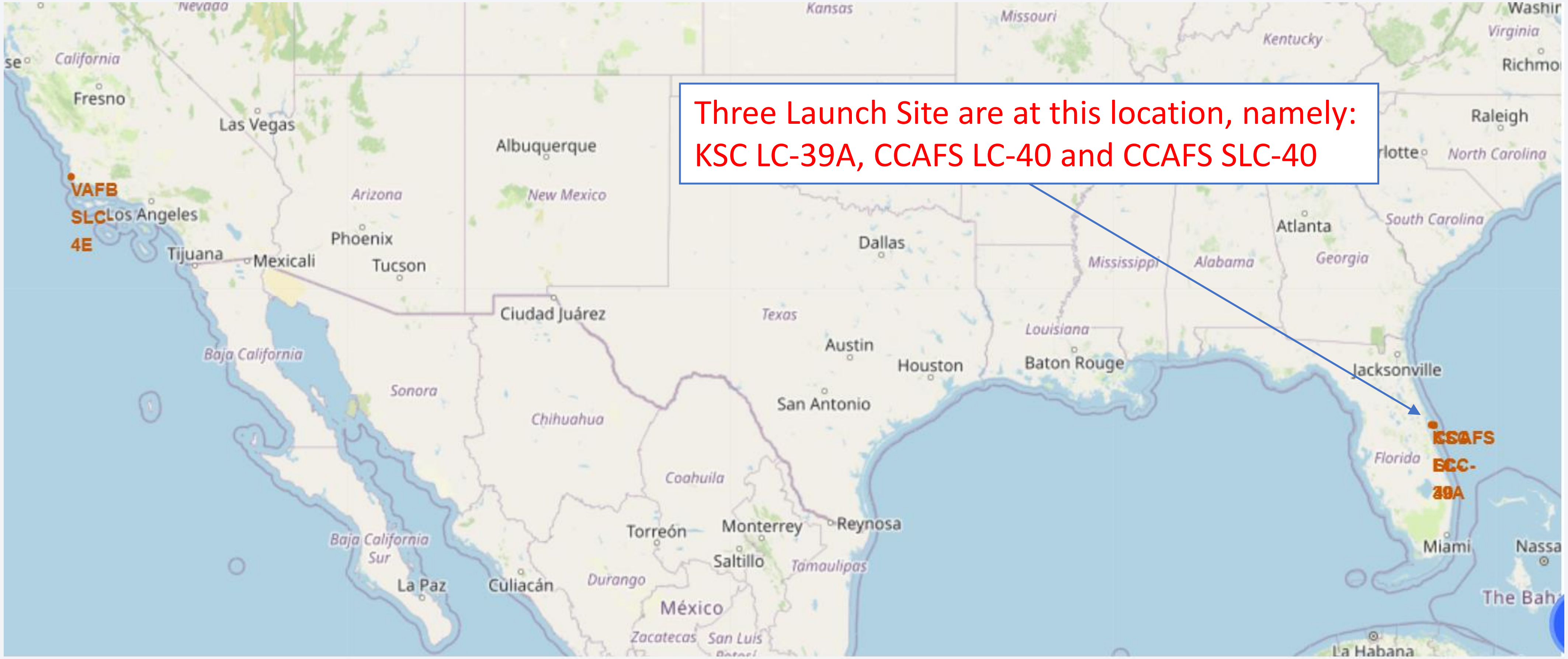
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing__outcome	frequency
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

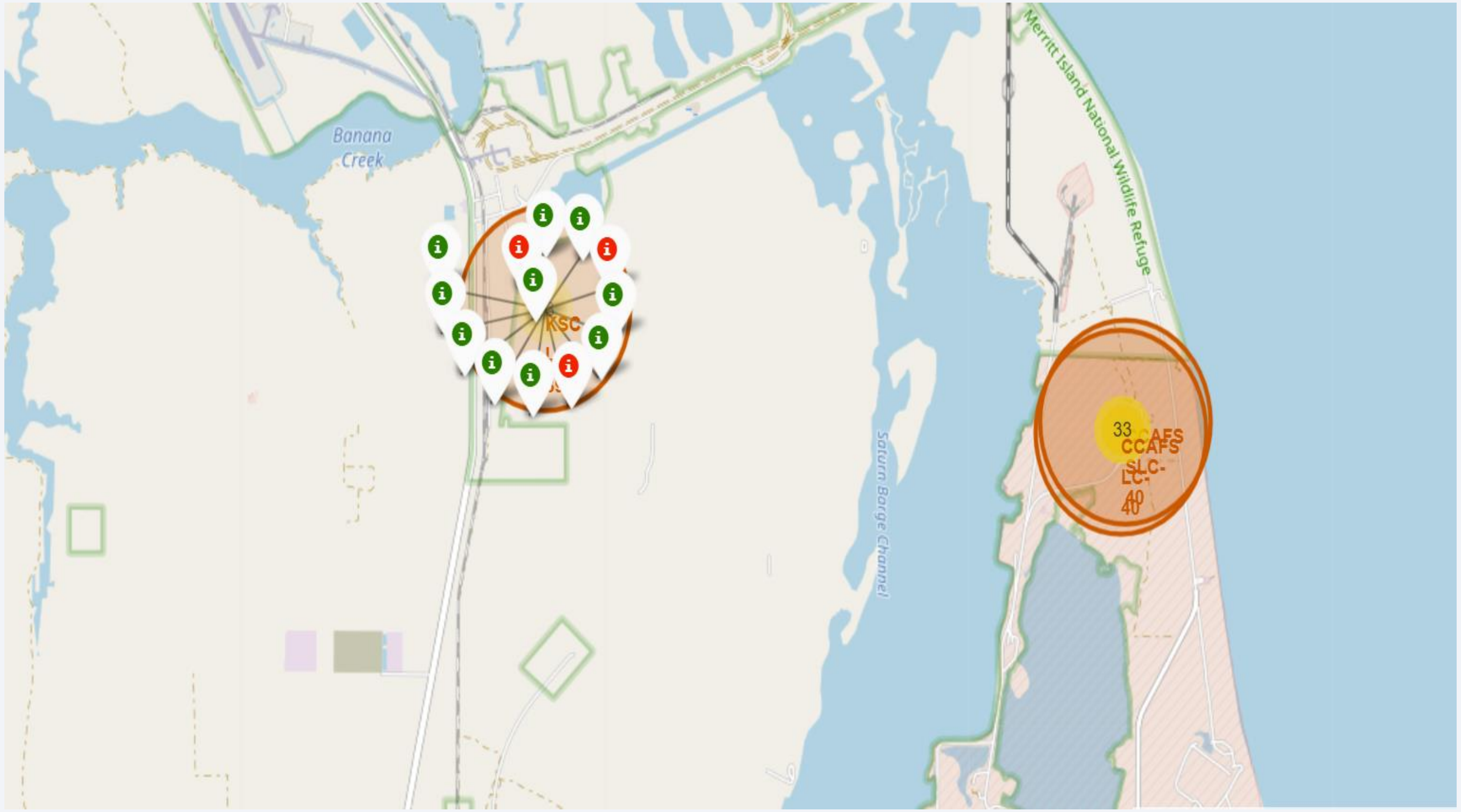
Section 3

Launch Sites Proximities Analysis

Folium Map With Launch Sites



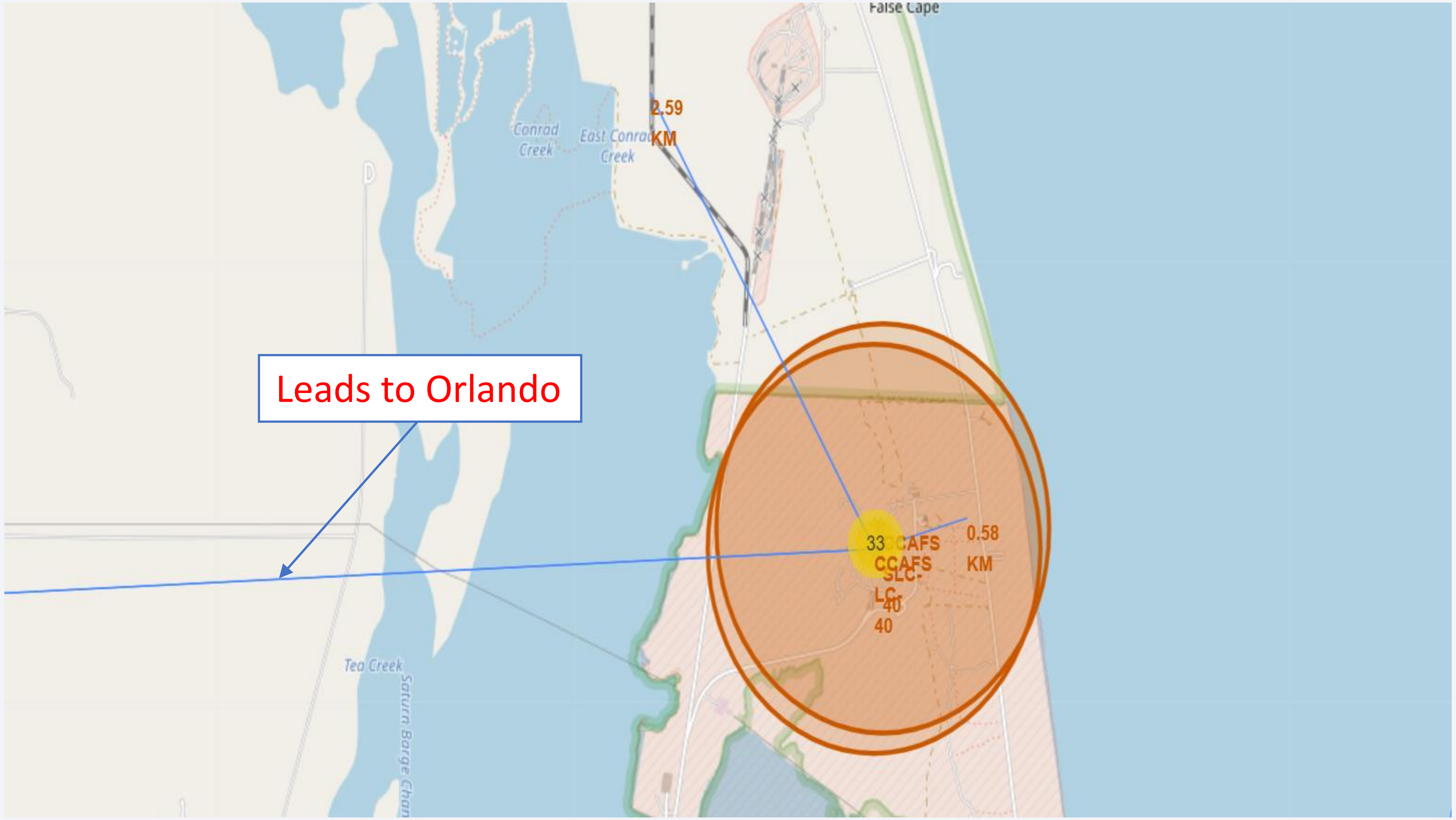
Folium Map with Labeled Outcomes



Folium Map Showing
Colored labeled
Landing outcome
From each Launch
Site

The Selected Launch
“KSC LC-39A” as seen
on the
Map had 10
successful and 3
Failure Landing
outcome

Distance from Launch Site



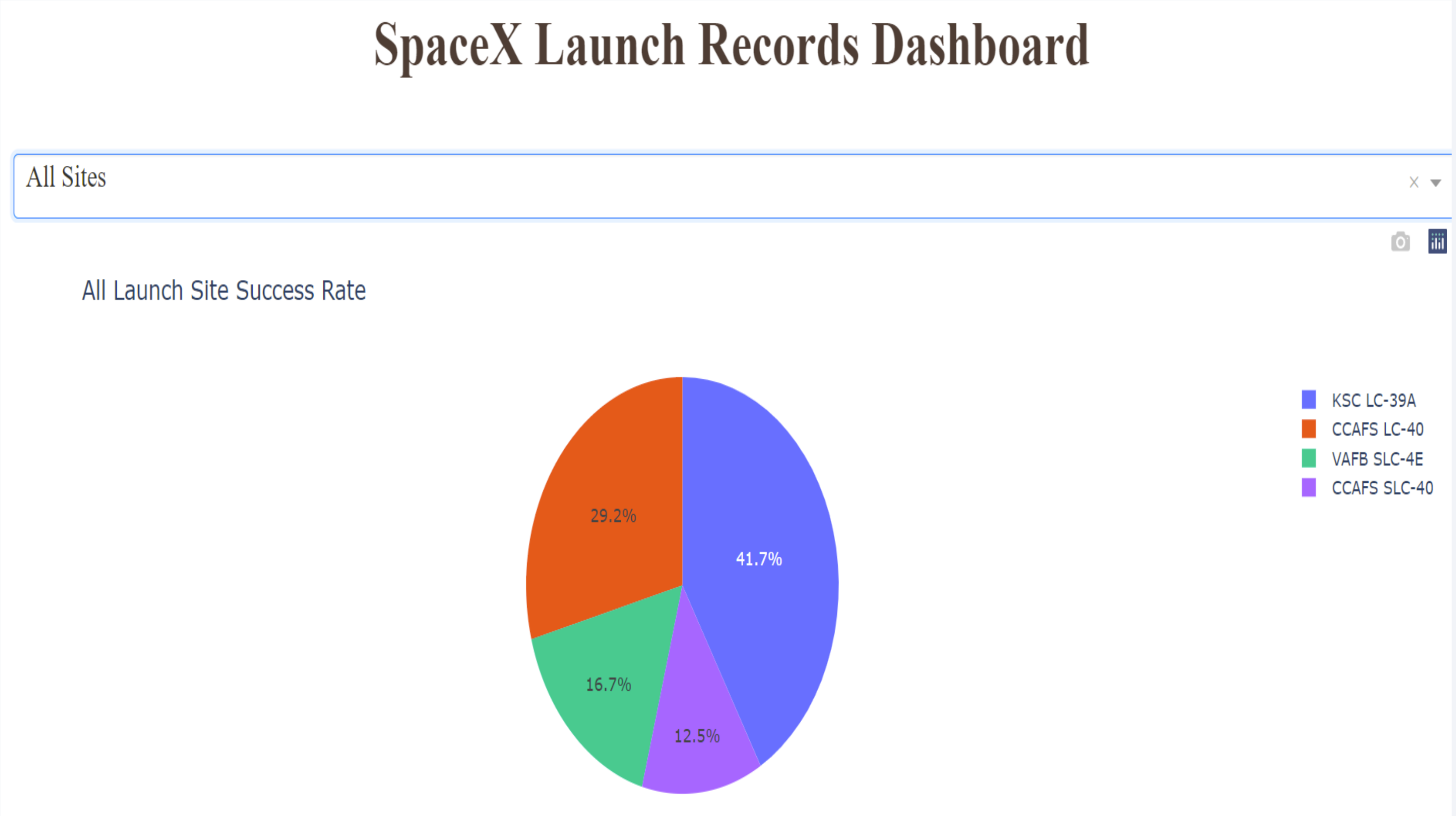
Folium Map showing Distance from Launch Site “SCAFS LC-40” to Nasa railroad, Highway and Orlando

A line is drawn from the launch site to the target location, also indicating the distance between both location on the map.

Section 4

Build a Dashboard with Plotly Dash

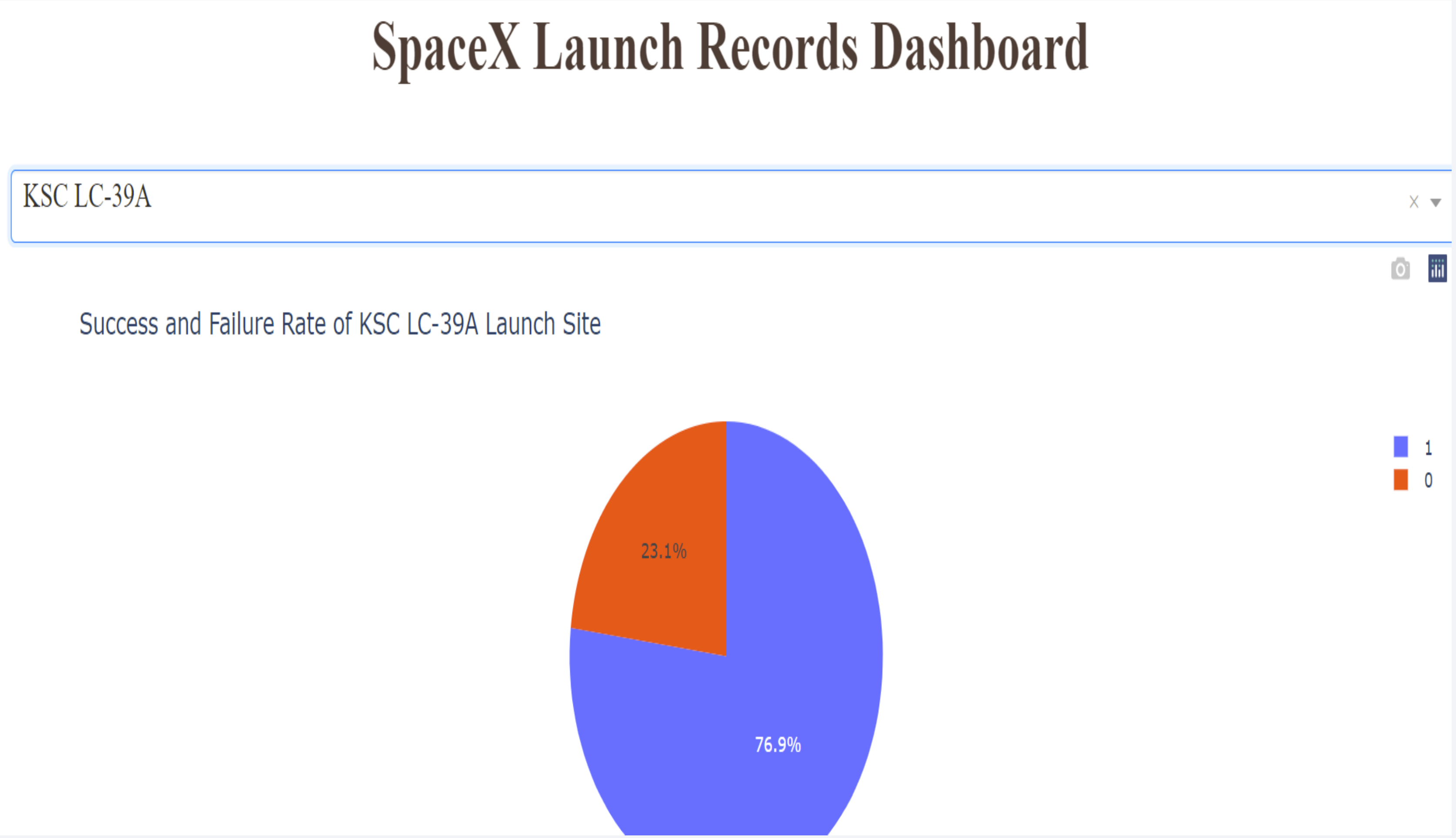
All Sites Pie Chart Dashboard



The Pie chart displays all Launch Sites and their successful landing rates when compared to other Launch Sites

It can be seen that KSC LC-39A had the highest landing success rate.

Best Performing Launch Site Pie Chart Dashboard



The best Performing Launch Site was KSC LC-39A with a success rate of 76.9% and failure rate of 23.1%

Dashboard Scatter Plot

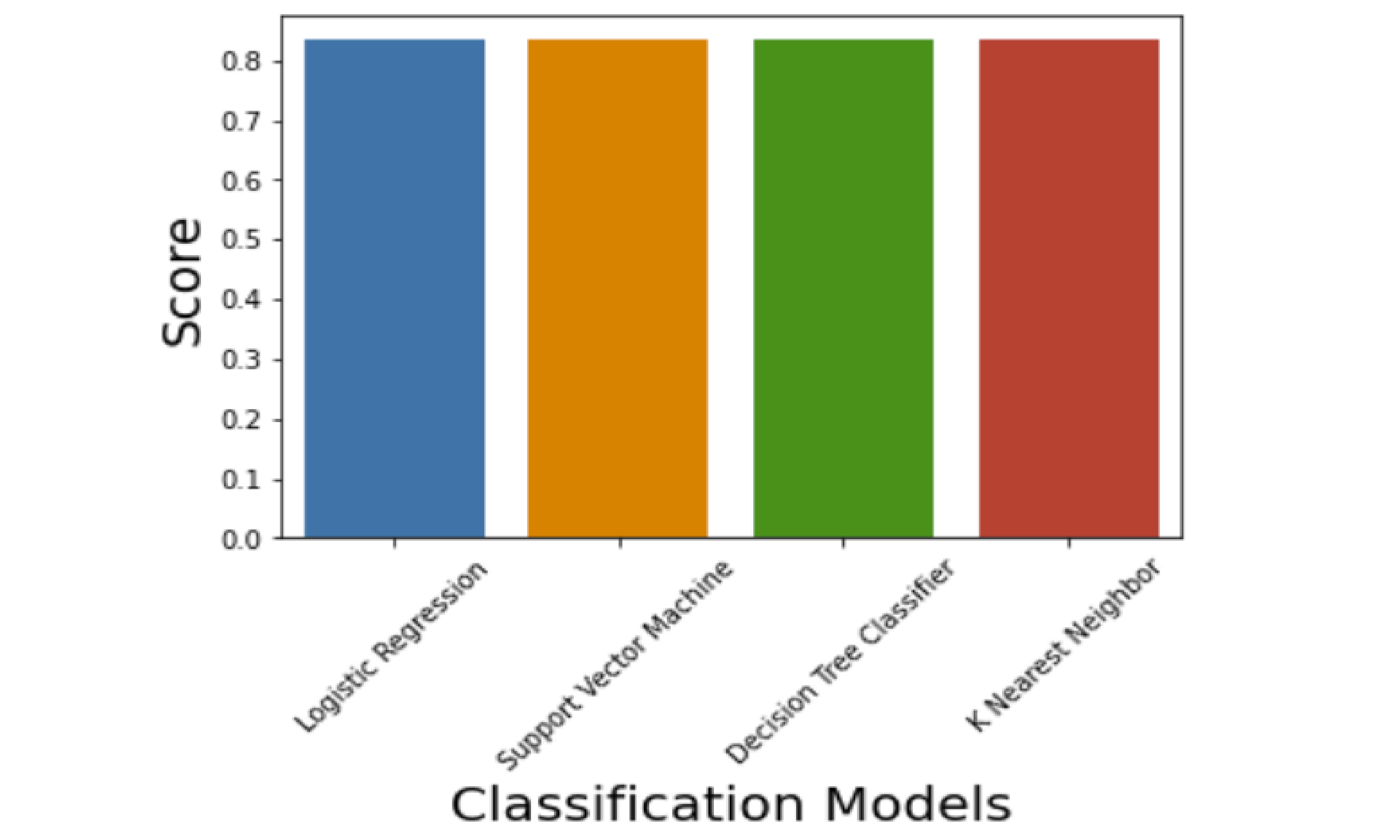
Booster Version
Category B5,
With a payload range of
3000
To 4000 performed
best.

This is because there
was only
One Launch of the B5
Booster
Version and it had a
successful
Landing outcome
(100%)

Section 5

Predictive Analysis (Classification)

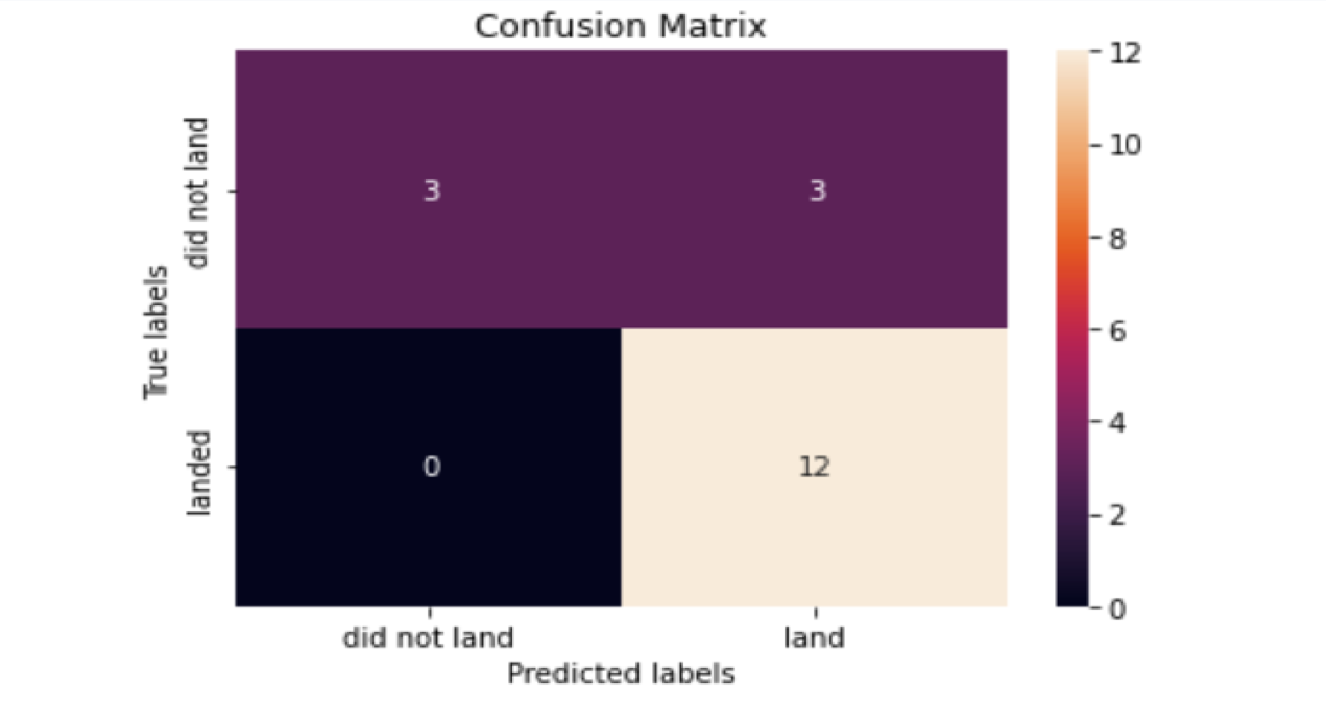
Classification Accuracy



All four models had a score of 0.83334 when tested with the test data as seen on the bar chart

Confusion Matrix

- Since all four models performed the same way, their confusion matrix were all the same as seen below.



Conclusions

- There is 100% chance of having a successful Landing outcome when going to orbits such as ES-L1, GEO,HEO and SSO
- All four classification models performed the same having a score of 0.8334 and the same confusion matrix
- The best Performing Launch Site was KSC LC-39A with a success rate of 76.9% and failure rate of 23.1%
- Finally, there is a yearly increase in successful landing outcome which can be linked to the advancement in rocket technology

Appendix

- GitHub link to full project

Thank you!