IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sooraj Sheregar
15 March 2023

Outline

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

Executive Summary

- Summary of methodologies:
  - We are trying to predict the SpaceX Falcon 9 first stage Landing by utilized the data collected from SpaceX API and data collected from Wikipedia with Web scraping. Some basic data wrangling and formatting is done to access the relevant data. Exploratory Data Analysis is performed and Training Labels are determined. Interactive visual analytics are developed using Folium and Plotly Dash. And predictive analysis is performed using classification models.
- Summary of all results
  - Based on the Accuracy score we have come to a conclusion that the models 'Decision tree' is best suited for predicting SpaceX Falcon 9 first stage Landing

Introduction

- SpaceX has gained worldwide attention for a series of historic milestones.
- It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The required data was collected by Requesting and parsing the SpaceX launch data form SpaceX API using the GET request
  - We have also collected data from Wikipedia using Web scraping through Beautiful Soup
- Perform data wrangling
  - First the data is filtered for Falcon 9 rockets only and then the missing data in the Payload Mass is replaced with the average payload mass.
  - The outcomes column is filled with different values of success and failure those data are transformed into binary values 1 for Success and 0 for Failure and placed in a new column 'Class'

Methodology

Executive Summary (continuation)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data is processed using One Hot Encoder to get the independent variable and the dependent variable, which is split into train and test set.
  - Find best Hyperparameter for SVM, Classification Trees, K-nearest neighbor and Logistic Regression

Data Collection

- The required data was collected by Requesting and parsing the SpaceX launch data form SpaceX API using the GET request

- We have also collected data from Wikipedia using Web scraping through Beautiful Soup.

Data Collection – SpaceX API

The API used is SpaceX API

It contains several Data we have filtered it to Falcon 9

The missing value in payload mass is replaced with the Average payload mass

GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL: Wikipedia
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	None	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	None	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	None	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44

Data Wrangling

- Payload mass missing values replaced with Average
- Exploratory data analysis to find patterns in data
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column for success and failure
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/EDA.ipynb

EDA with Data Visualization

- Scatterplot of **FlightNumber vs. PayloadMass**
  - Increase in FlightNumber : Increase in Success rate of landing
  - Increase in PayloadMass : Decrease in Success rate of landing
- Scatterplot of **FlightNumber vs LaunchSite**
  - CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - But CCAFS LC-40 also has higher number of launches compared to other two.
- Scatterplot of **Payload vs Launch Site**
  - VAFB SLC 4E used for launching with payload upto 10000Kg
  - CCAFS LC-40 & KSC LC-39A can be used for higher payload

EDA with Data Visualization

- Bar Chart **Orbit type vs Success rate**
  - SO has the lowest success rate
  - ES-L1, GEO, HEO, SSO has the highest Success rate
- Scatterplot of **FlightNumber vs Orbit**
  - SSO &LEO have better success rate
- Scatterplot of **Payload vs Orbit**
  - SSO payload mass is below 5000 Kg
  - VLEO payload mass is above13000 Kg
- Line Chart **Year vs Success rate**
  - With the increase in year the success rate is also increasing
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/EDA%20with%20SQL-coursera_sqllite.ipynb

Build an Interactive Map with Folium

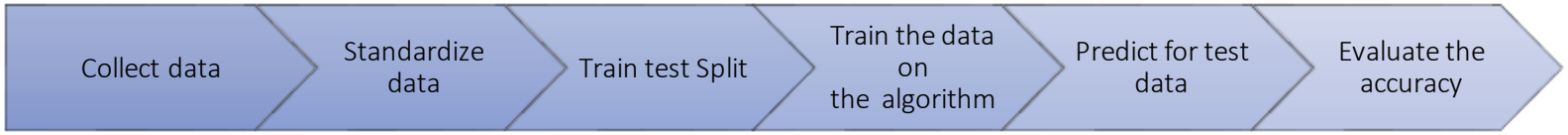
- Using Folium we have created markers, circles on the launch sites.
- We have created marker object of colour green for success and colour red for failure.
- We have created marker clusters for the launches so that its easily accesible.
- We have created lines to display distance from the launch site to nearest city, coast line, Highway, Railway and their distance. etc. you created and added to a folium map.
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- We have added Dropdown lists, Pie chart, slider, scatter plot, to the dashboard.
- Dropdown list is to enable Launch Site selection.
- Pie chart shows the total successful launches count for all sites & if a specific launch site was selected it shows the Success vs. Failed counts for the site.
- Slider is to select payload range.
- Scatter chart shows the correlation between payload and launch success
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/dashwithpython.py

Predictive Analysis (Classification)

- We have built 4 models they are Logistic regression, Support vector machine, Decision tree, K-nearest neighbors.
- GridSearchCV is used to evaluate the best hyperparameter suited for the model.
- Using the accuracy score and confussion matrix the best performing classification model is found to be Decision tree
- Model development process included



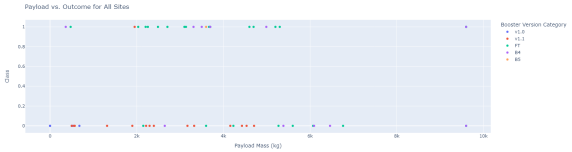
- GitHub URL: https://github.com/SoorajSheregar/Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

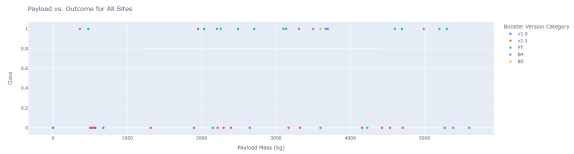
Results

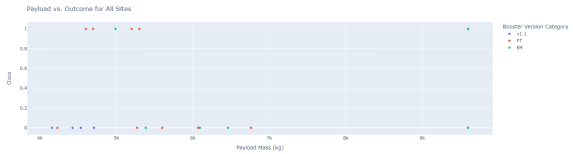
Exploratory data analysis results:

- Most of the flights are sent to GTO, ISS orbit as these are orbits where satellite are.
- Currently there are 4 Launch sites and 3 of them are located near Melbourne and 1 near Lompoc
- VAFB SLC 4E is used only to launch a max payload of 10000 kg
- Success rate is Increasing year by year with new Booster versions

- Interactive dash:





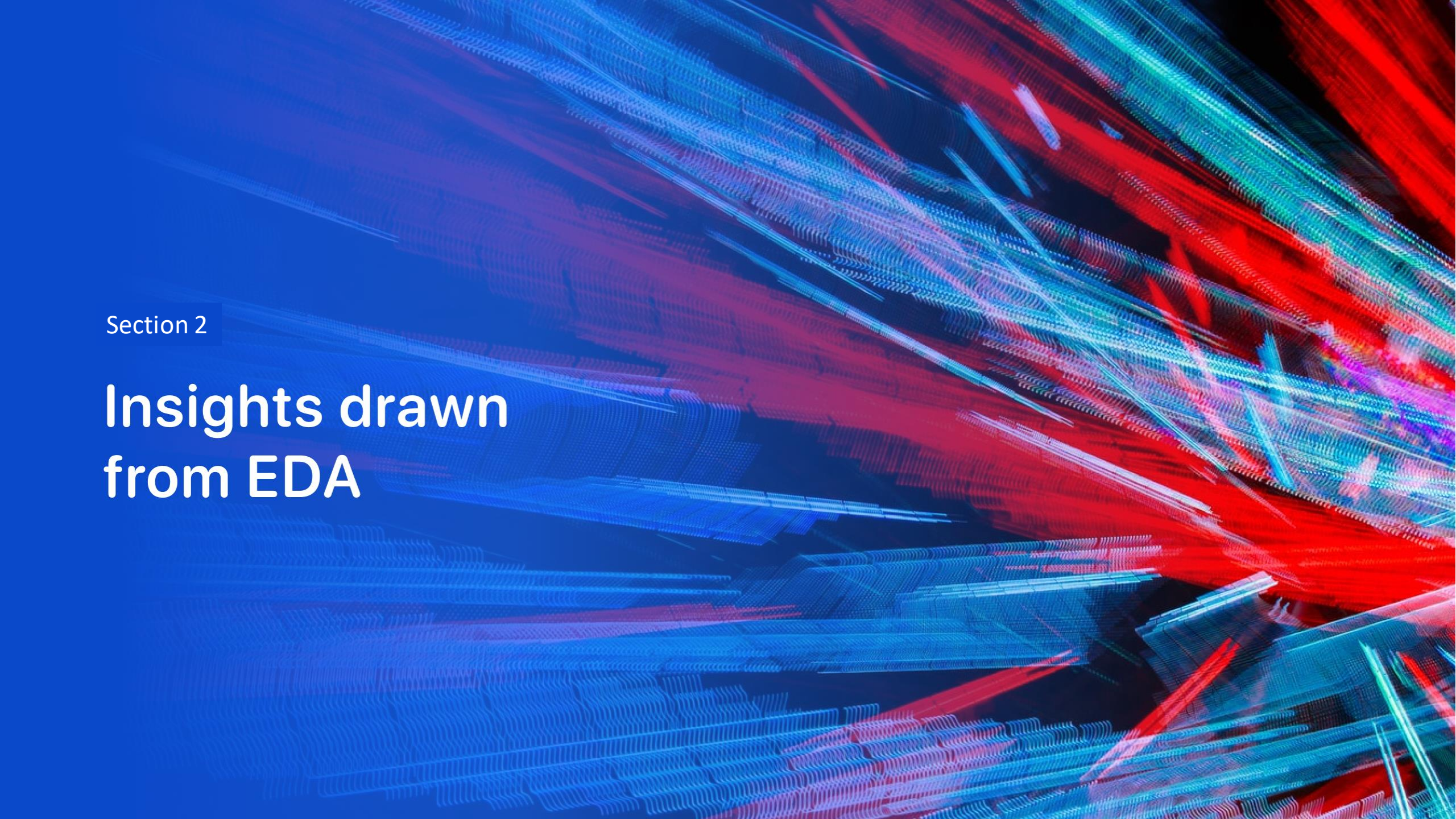


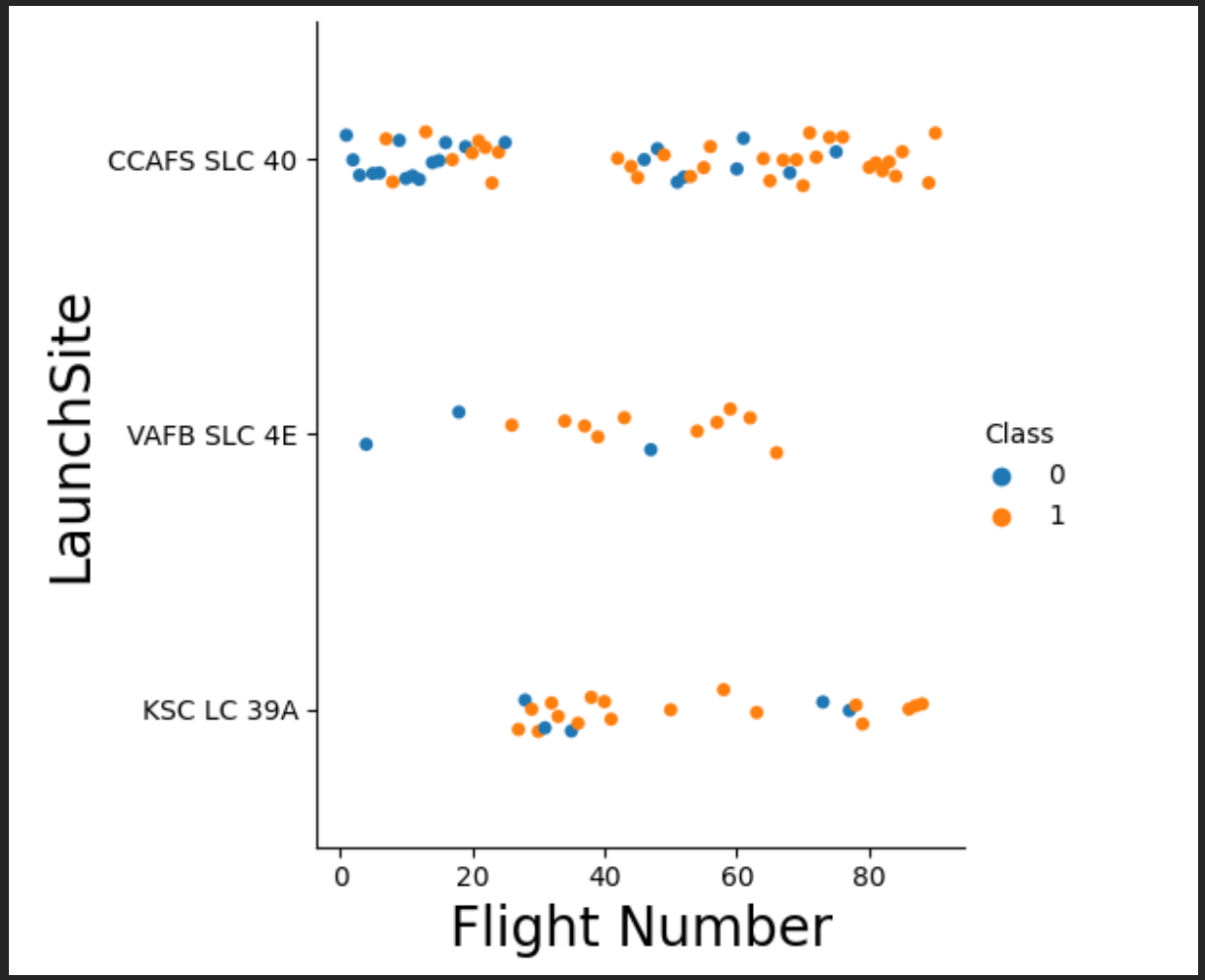
Predictive analysis results:

- Decision tree is the best model suited for the prediction

Section 2

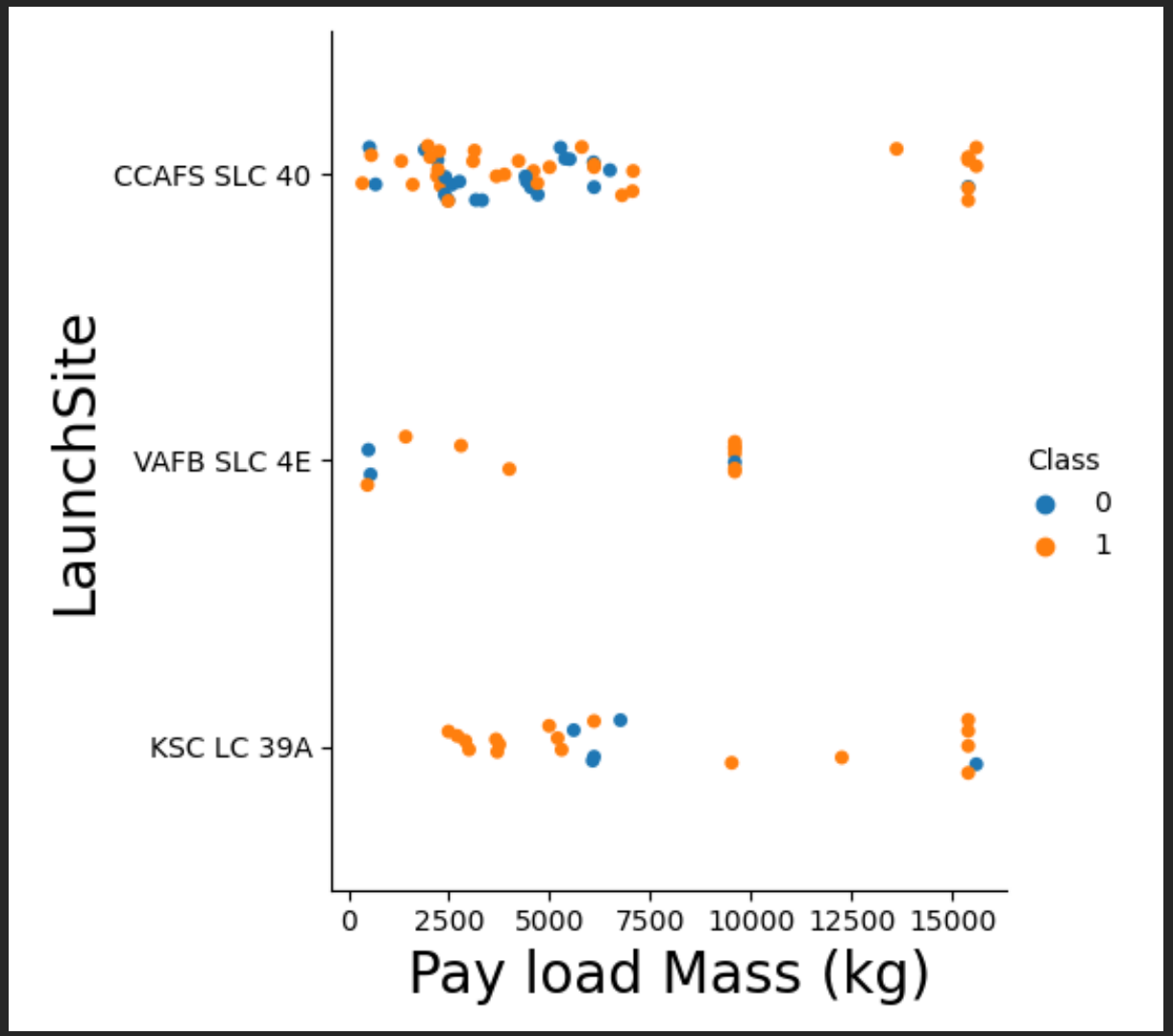
Insights drawn from EDA





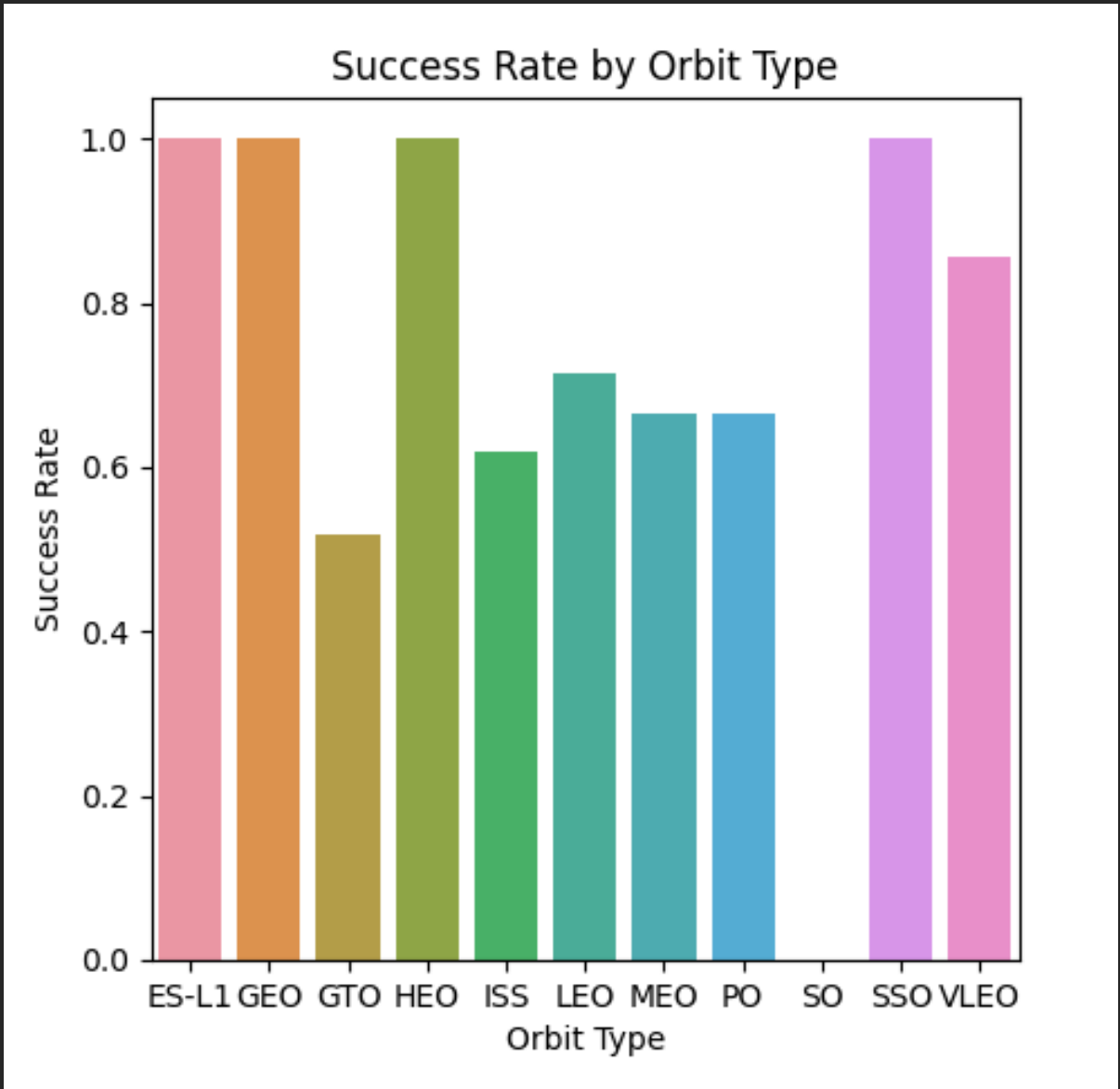
Flight Number vs. Launch Site

- We can see that CCAFS SLC 40 has the highest number of flights.
- Also if we check with the increase in flight numbers the success rate is more.



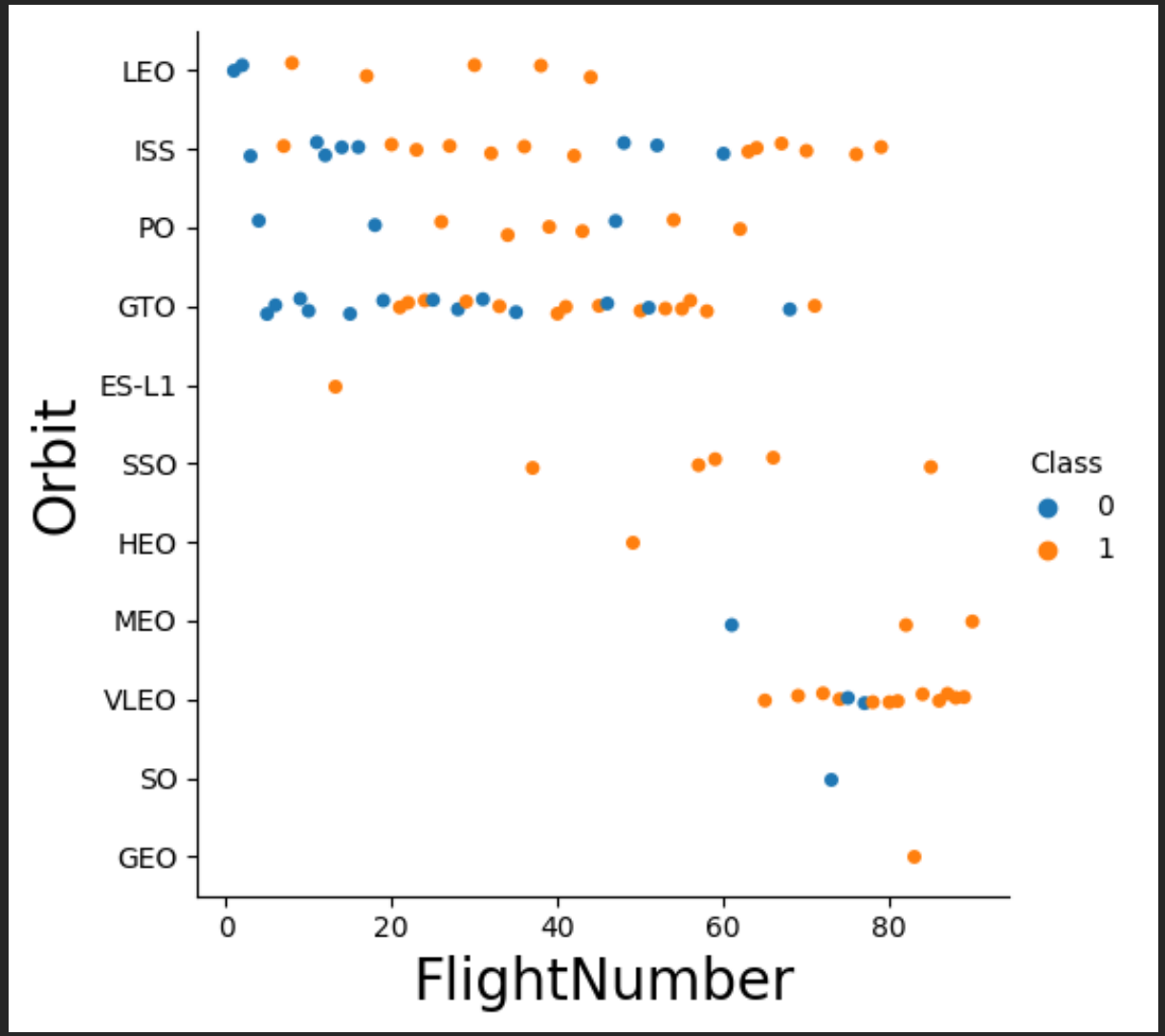
Payload vs. Launch Site

- CCAFS SLC 40 & KSC LC 39A can be used for payload mass greater than 10,000 Kg.
- Whereas VAFB SLC 4E is limited to 10,000 Kg.



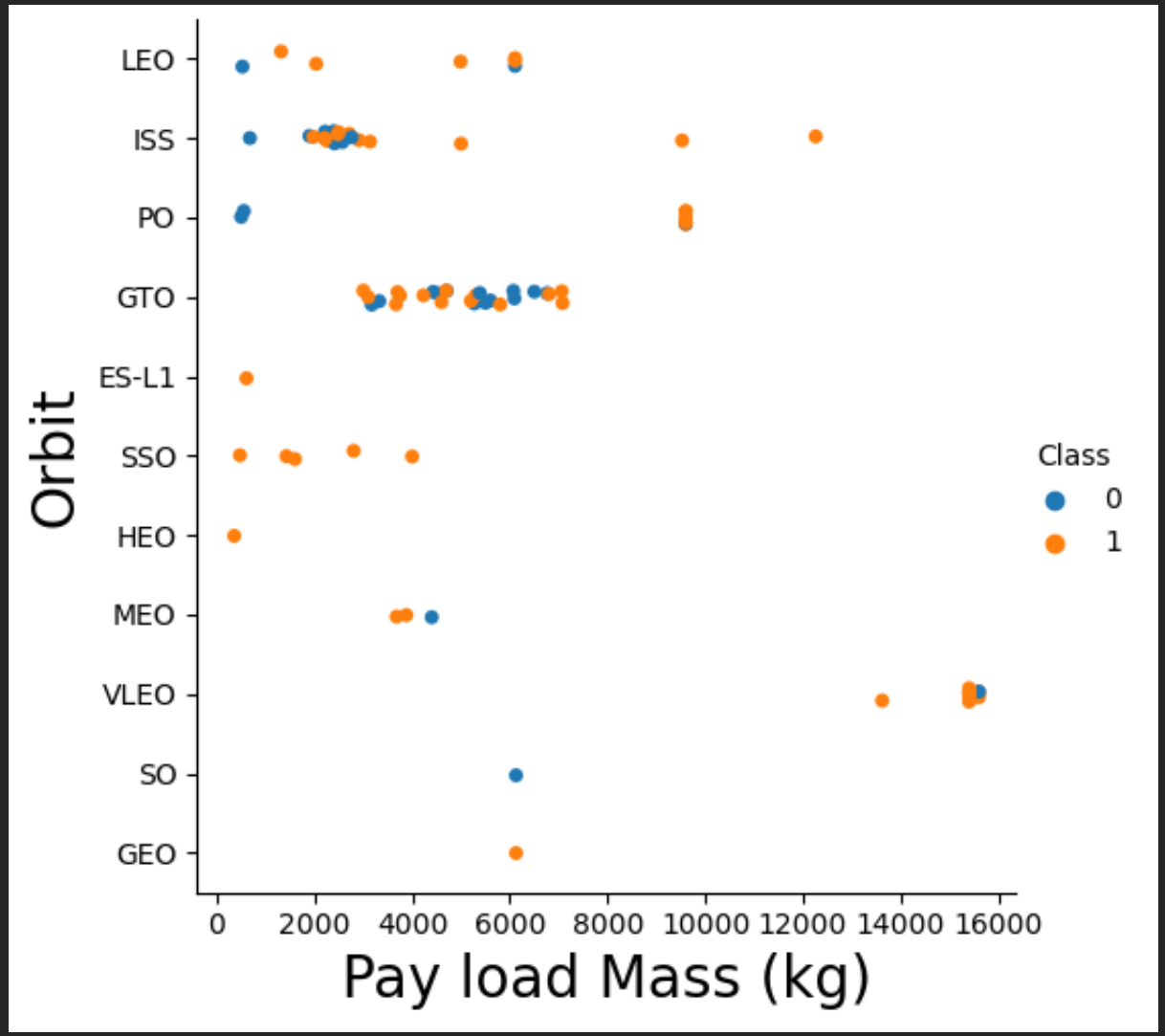
Success Rate vs. Orbit Type

- Success rate of ES-L1, GEO, HEO, SSO seems to be high but when accounted with total number of samples considered SSO can be considered as best performing.



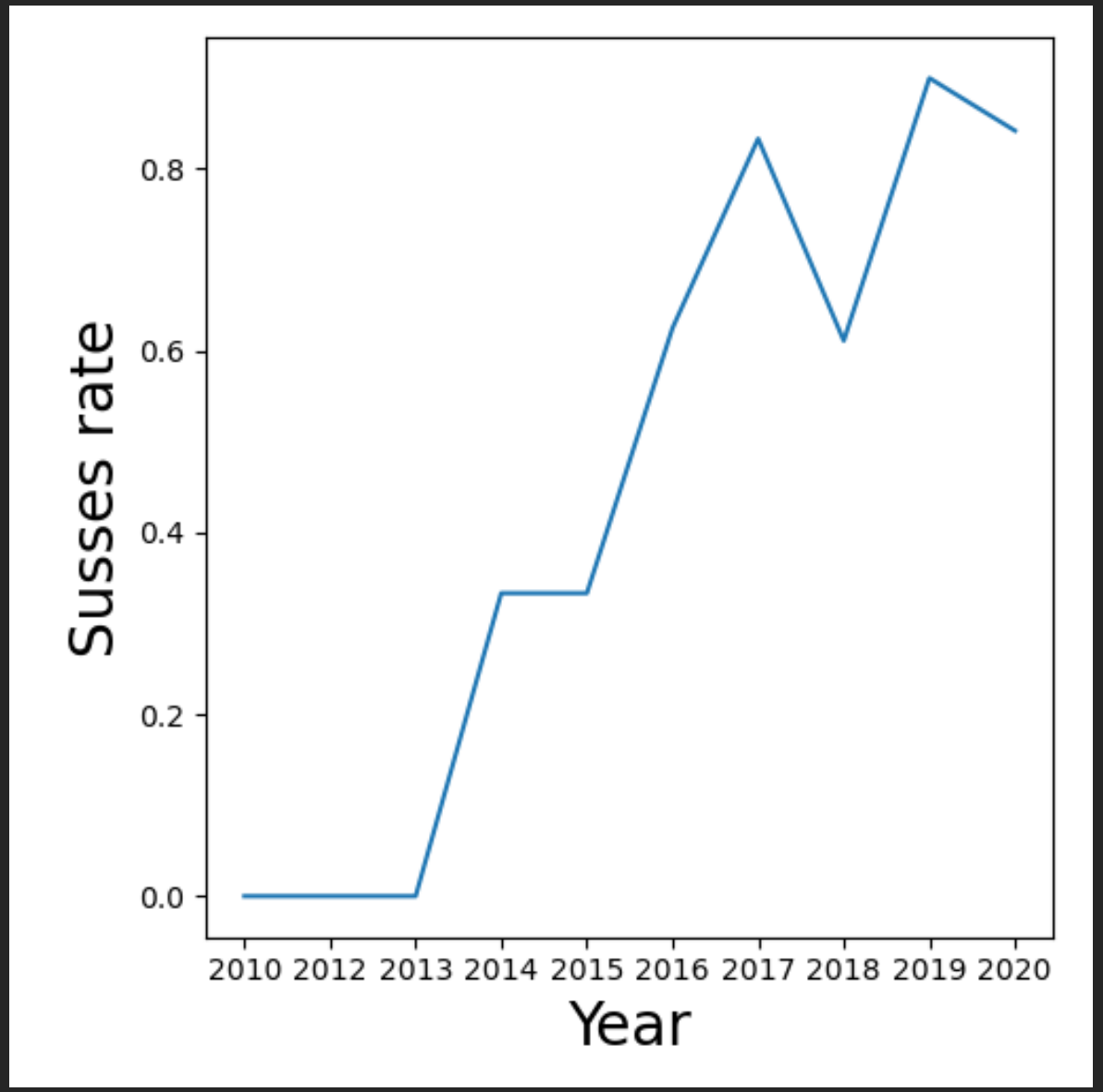
Flight Number vs. Orbit Type

- As we can see the highest flights sent are to GTO & ISS orbits.
- Least number of flights sent to ES-L1, GEO, SO & HEO, so it will be difficult to predict outcome due to insufficient data.



Payload vs. Orbit Type

- We can see that VLEO orbit requires highest payload approximately above 13,000 Kg.
- PO & ISS orbits come to the next range where payload mass is above 8,000 Kg.



Launch Success Yearly Trend

- We can Notice a significant increase in the success rate with Time

All Launch Site Names

- We have selected Distinct Launch Site names to display the launch site names without repetition.

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- We have applied a Where condition to meet the requirement to begin with `CCA` and a LIMIT to display only 5 rows

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Total Payload Mass

- We have calculated the total payload carried by boosters from NASA by applying the SUM() function and a Where condition to meet the requirement of Customer to be NASA.

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Vers
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass by F9 v1.1

- We have calculated the average payload mass carried by booster version F9 v1.1 by using the function AVG() and a WHERE condition to meet the requirement of Booster version to be F9 v1.1.

First Successful Ground Landing Date

```
%sql SELECT MIN(strftime('%Y
```

```
 * sqlite:///my_data1.db
Done.
```

first
2015-12-22

- To find the dates of the first successful landing outcome on ground pad we have used MIN() function and WHERE condition to meet the requirement of Successful ground landing.

```
%sql select Booster_Version from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' A
```

```
 * sqlite:///my_data1.db
Done.
```

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- We have listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 by using WHERE condition.

```
%sql select MISSION_OUTCOME , count(*) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- We have calculated the total number of successful and failure mission outcomes by using COUNT() feature.

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- We have listed the names of the booster which have carried the maximum payload mass by using the function MAX().

monthname	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

2015 Launch Records

- We have listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

- We have 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

count	Landing _Outcome
20	Success
8	Success (drone ship)
6	Success (ground pad)

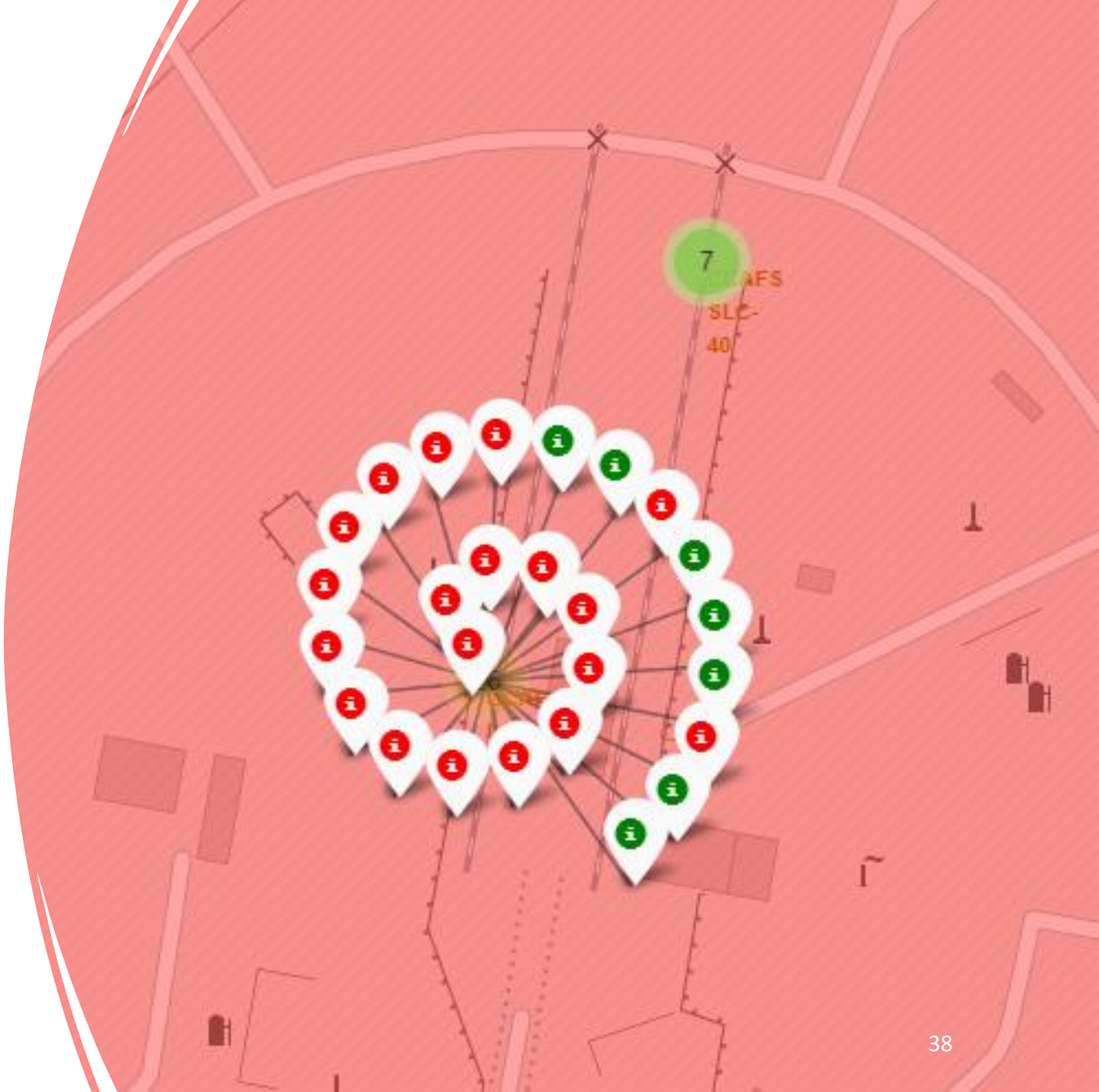

Section 3
Launch Sites
Proximities Analysis

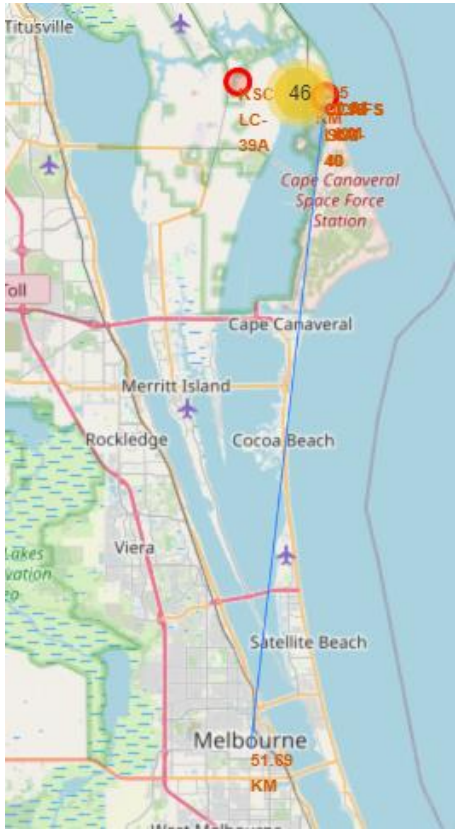
Launch Sites

- In the above Global map we can see all the launch sites location Marker
- We can see that launch sites are mainly in two areas, 3 of them are located near Melbourne and 1 near Lompoc.
- We can also notice that all the launch sites are near the coast line

Colour Labeled Launch Outcomes

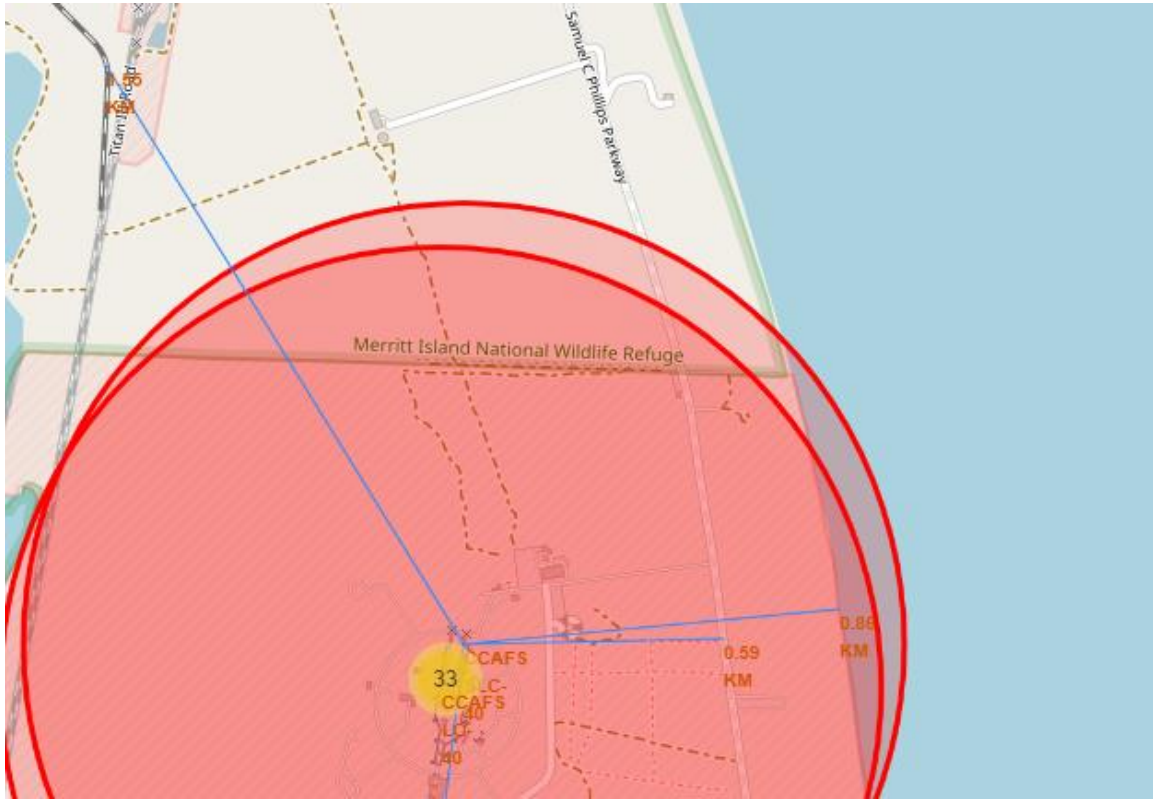
- Successful Outcomes are labeled Green on the map.
- Unsuccessful Outcomes are labeled red on the map.





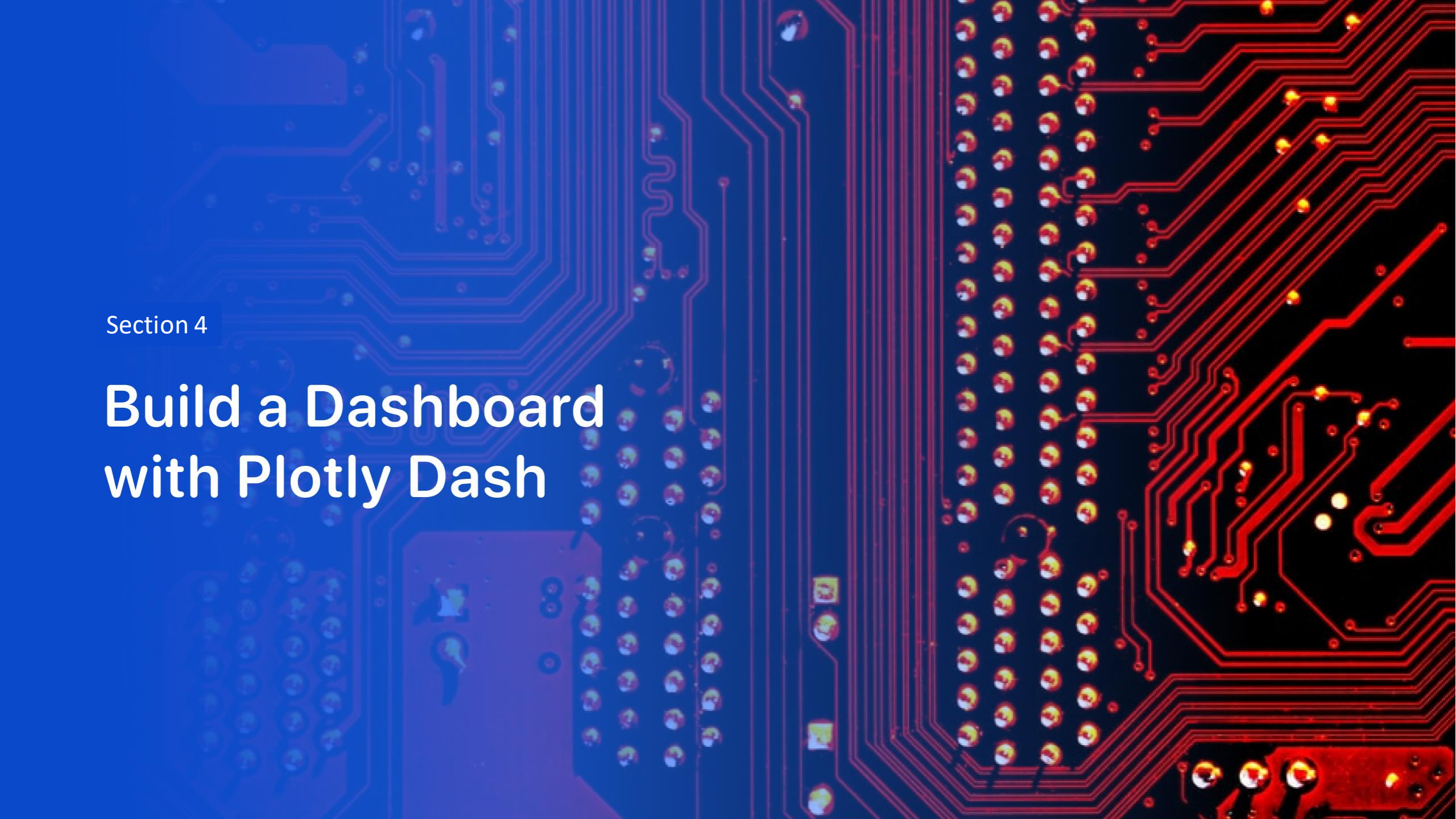
Distance of CCAFS SLC-40

1. Coastline- 0.85 km
2. Highway- 0.55 Km
3. Railway- 1.55 Km
4. Melbourne City- 51.59 Km



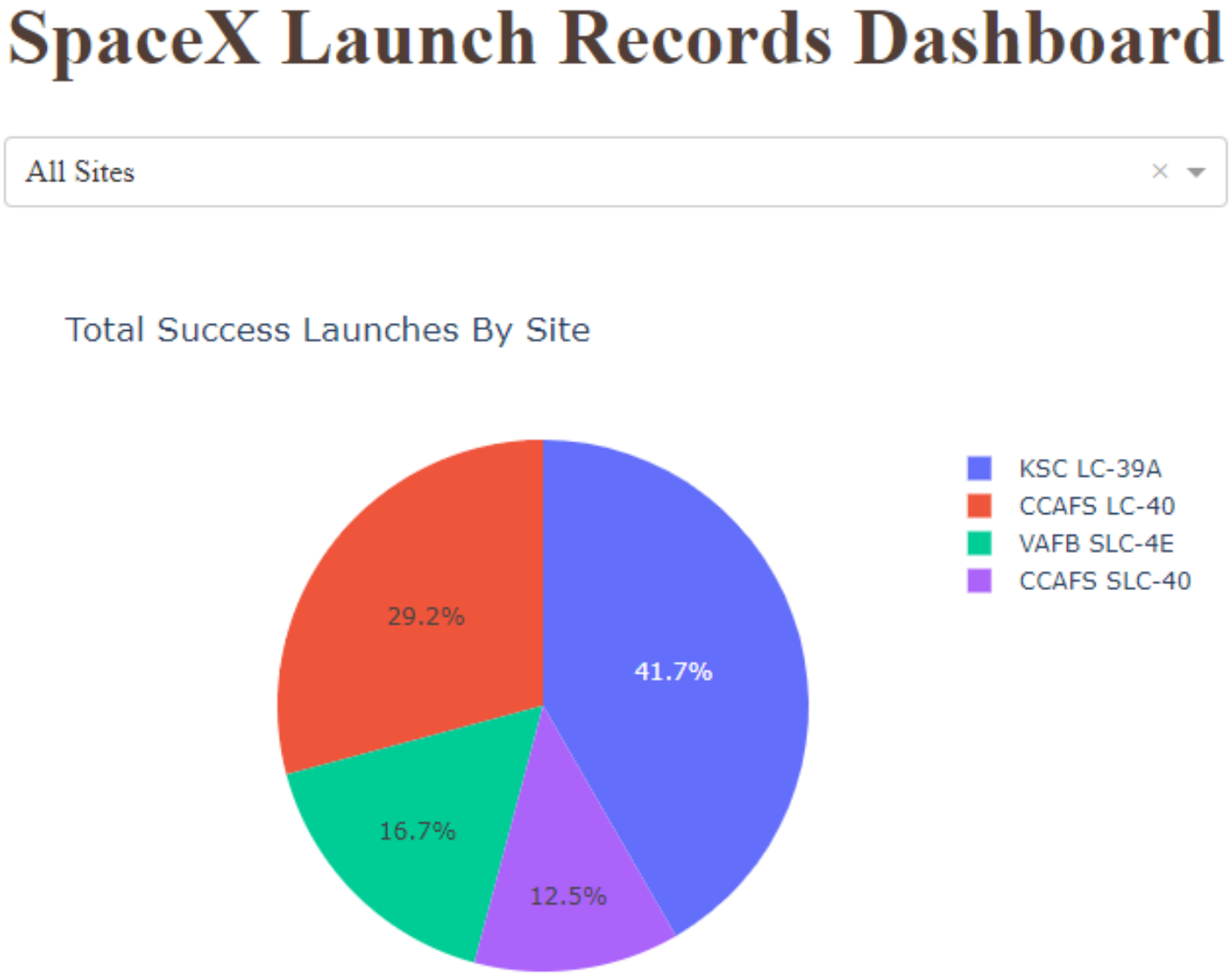
Launch site Distance

- Launch site distance from various areas are displayed on the map with a line and the Distance in Kilo-meters


Section 4
Build a Dashboard with Plotly Dash

All Sites Success Launches

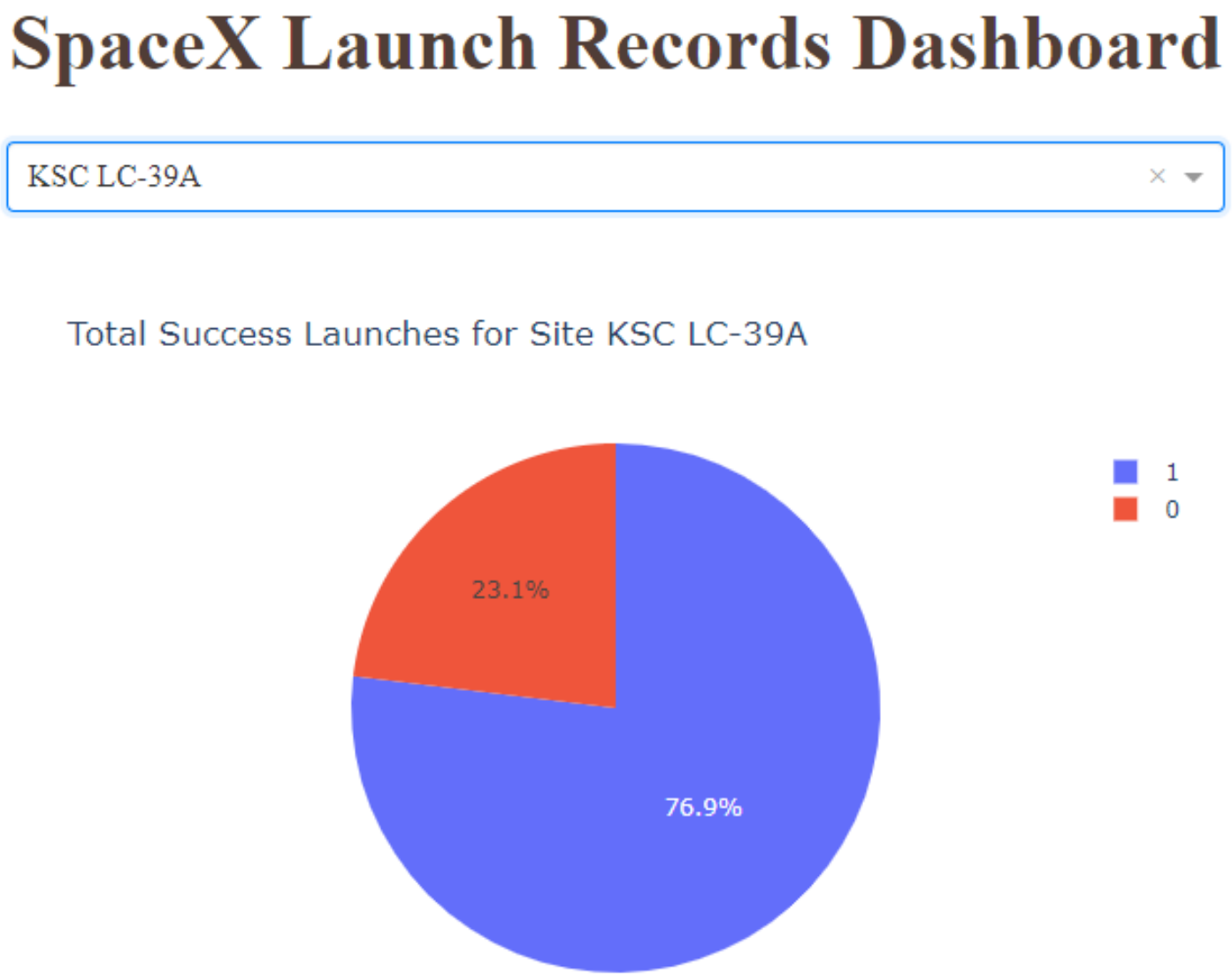
- We see that KSC LC-39A which comprises of 41.7% of all sites, has the highest Successful Launches
- And CCAFS SLC-40 which comprises of only 12.5% of all sites, is lowest when compared on the basis of success rate



Site With Highest Success

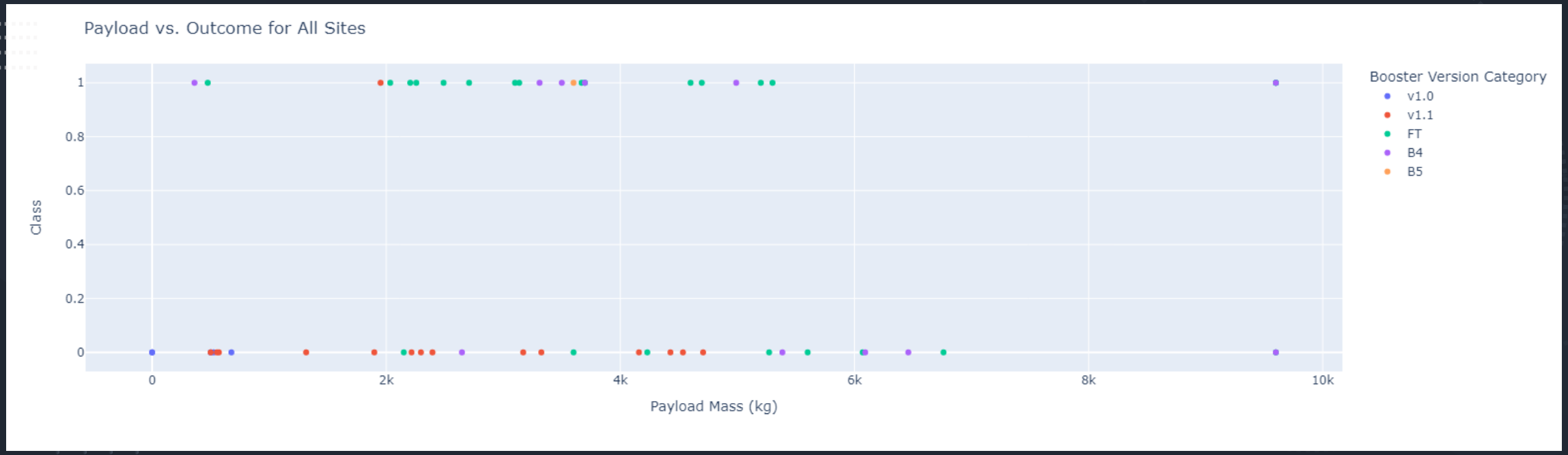
- As we know that KSC LC-39A has the highest successful launches when we dig into more data we see that Success rate at this site is 76.9% and the failure rate is 23.1%.

- So this implies that 3 out of 4 launches tend to be successful.



Interactive Scatterplot of Payload vs Outcome

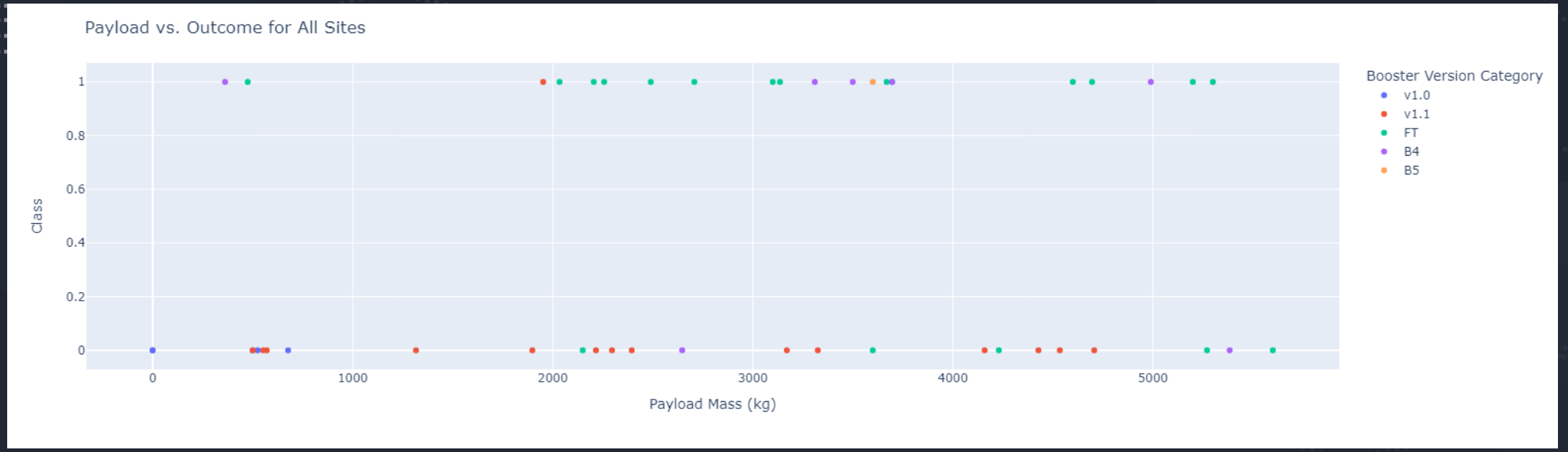
- When we choose the range between 0 to 10000 Kg, we notice that there are very fewer flight above 7000 Kg payload and those are of B4 Booster version.



Range 0 to 10,000 Kg

Interactive Scatterplot of Payload vs Outcome

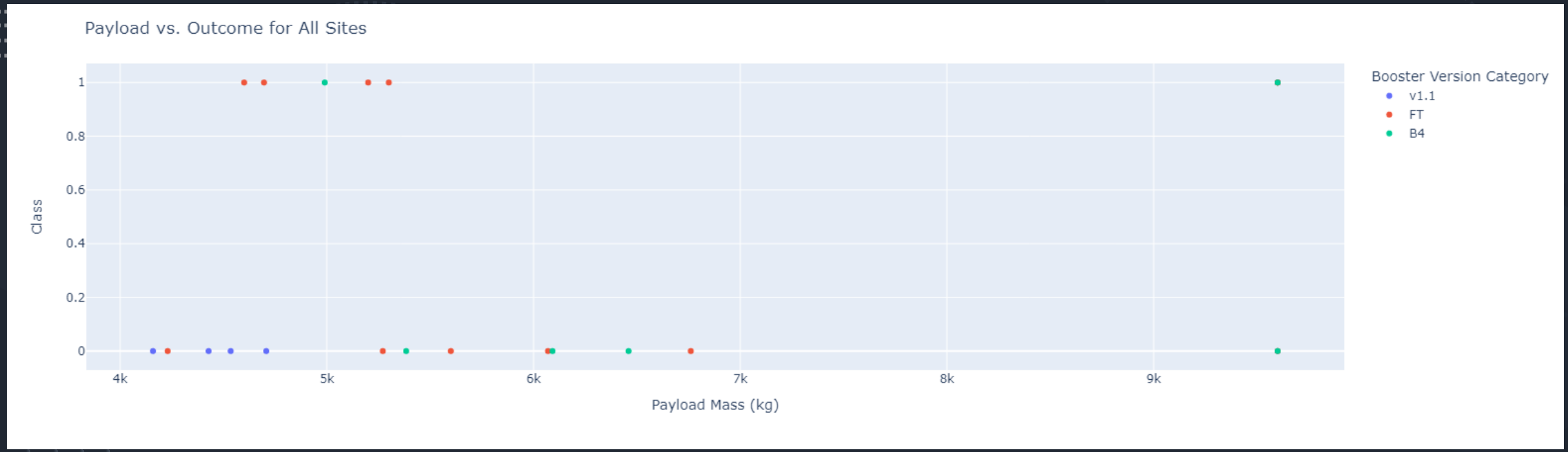
- In the Range between 0 to 6000 Kg we can see that Booster version FT have better success rate when compared to others and also notice that v1.0 and v1.1 have lowest success rate and have low payload mass.



Range 0 to 6,000 Kg

Interactive Scatterplot of Payload vs Outcome

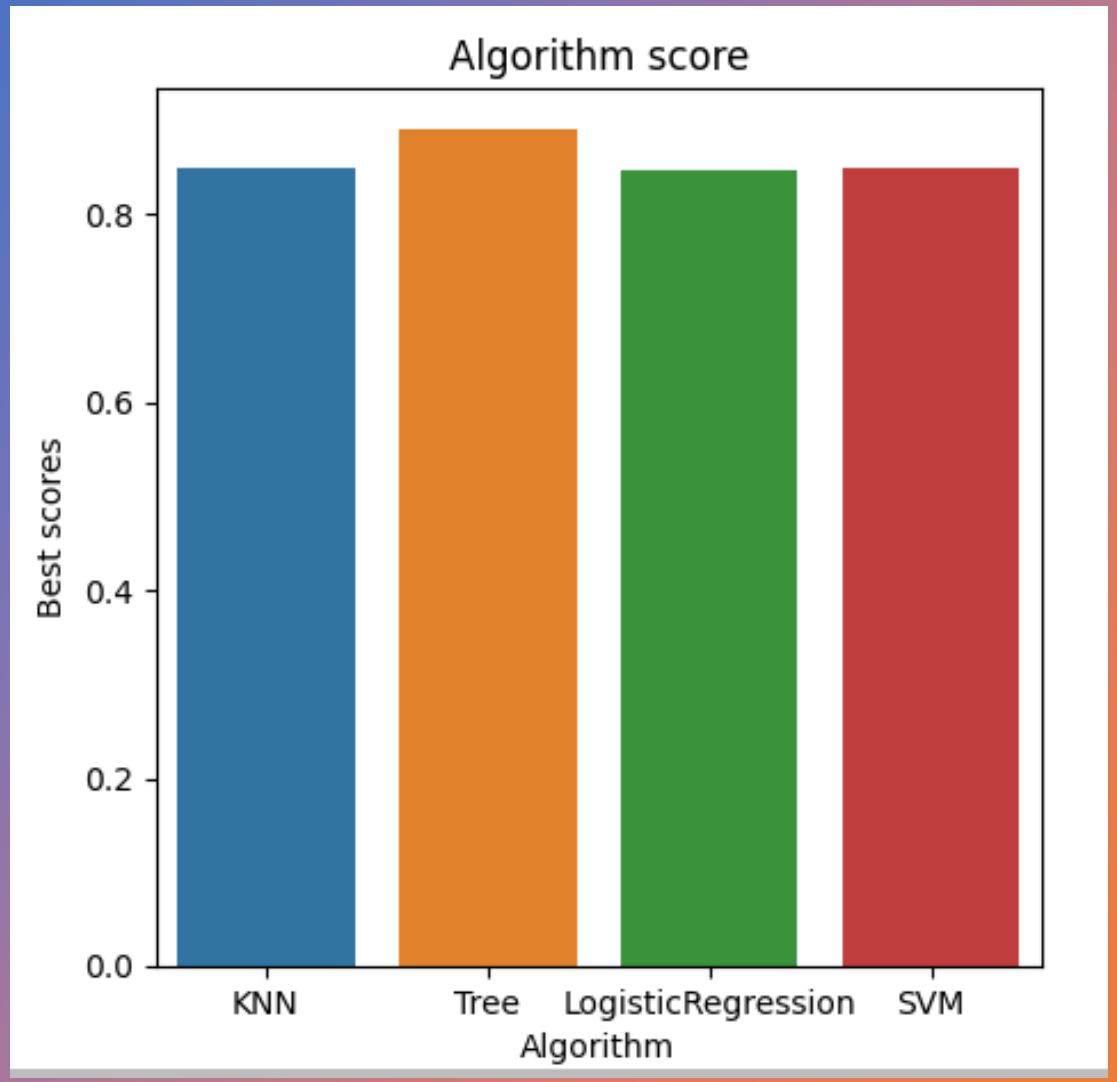
- If we check in the range 4000 to 10000 Kg, we notice that FT & B4 can be seen with success but mostly it is failure . So we can conclude that chances of failure in this payload range is more for v1.1, FT, B4.



Range 4,000 to 10,000 Kg

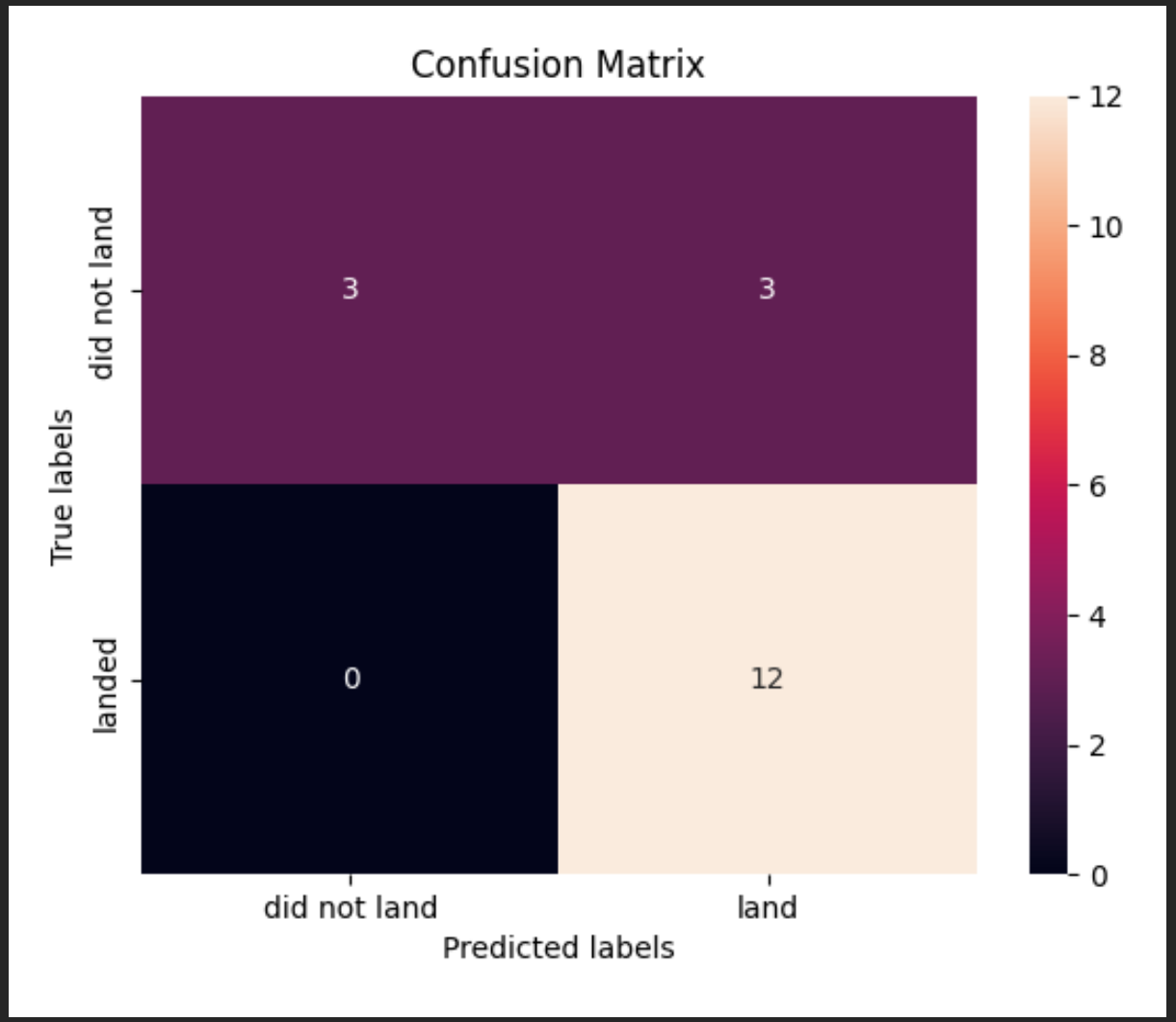
Section 5

Predictive Analysis (Classification)



Classification Accuracy

- Decision Tree model has the highest classification accuracy of around 0.89



Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

Conclusions

- The success rate of sending a flight to SEO orbit is comparatively higher than any other orbit
- With time the success rate is increasing.
- Launch sites are supposed to be far from human habitat.
- Out of all sites KSC LC-39A is preferable due to its high success rate.
- Booster version B4 & FT are preferable due to higher success rate.
- Decision Tree is the best model preferable to predict Landing Outcomes.

Appendix

- For more data related to this project please refer the following GitHub URL: https://github.com/SoorajSheregar/Capstone-Project

Thank you!