IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Samuel Konzi
February 2, 2022

Outline

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

Executive Summary

Summary of methodologies

In this report we are exploring the area of space travel, which has in the recent past become affordable and the opportunity for it made available by different companies. SpaceX is perhaps the most successful in the field because it can reuse the first stage. We thus use data from the SpaceX API and the SpaceX Wikipedia page to establish why. The data is collected by web scrapping using various python libraries such as requests, beautiful soup. Exploratory data analysis is performed using tools such as pandas, SQL. The cleaned data is then fed to a machine learning model to predict if SpaceX's rocket landing success rate.

Summary of all results

After exploring the data, we found that Spacex landing success rate is 83.33%. This percentage is only calculated for a successful landing as there are cases where the landing was attempted but failed due to an accident and of course cases where it didn't land at all. We observed several factor affecting the landing outcomes such as orbit type, Payload Mass , Launch site and so on. It was observed that generally the success rate has been on the rise since the period between 2013 and 2020.

Introduction

Project background and context

Space travel is now being made available for every one with SpaceX advertising on its website the best price of 65 million USD against the 165million USD price offered by other companies. Much of the price savings are because SpaceX can reuse the first stage of its Falcon 9 rocket.

Problems you want to find answers

Space Y a company owned by Allon Musk wants to compete with Space X.

For that it determines if SpaceX's rocket will land for reuse, and with that it can determine the cost of launch. Space Y asked to perform this prediction using a machine learning model.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- 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

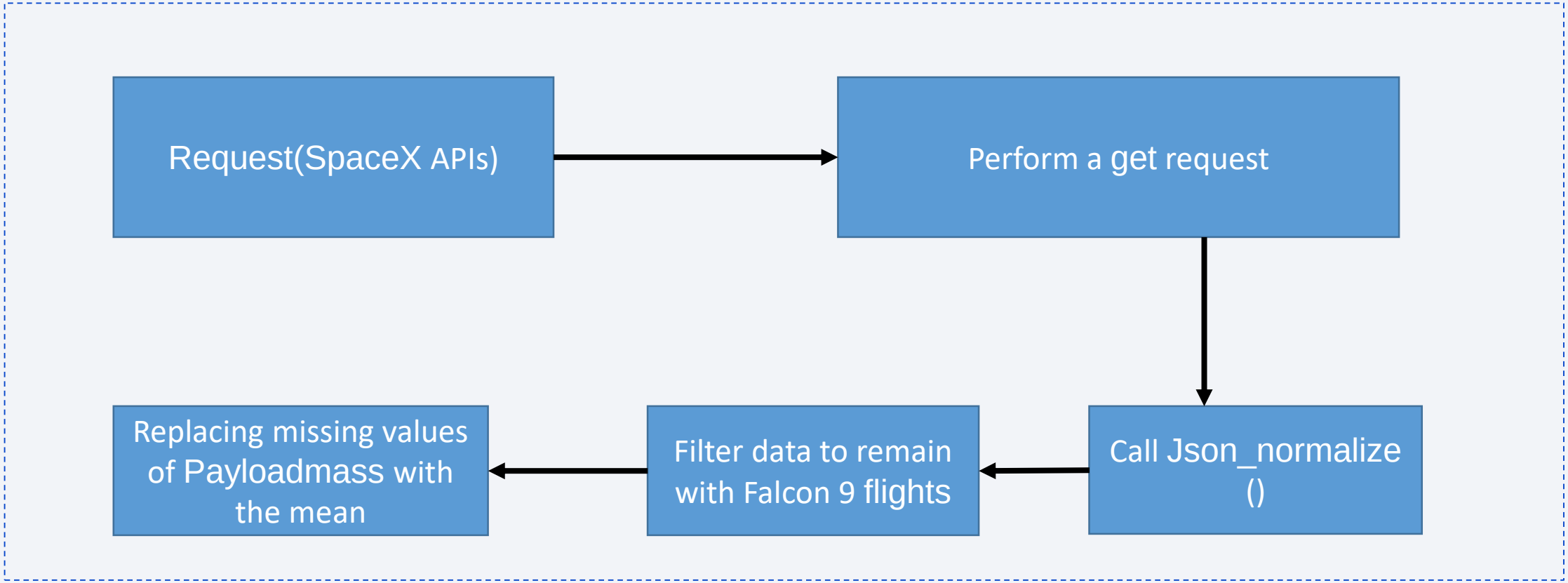
Data Collection

The data was collected in two ways.

- API Requests from SpaceX REST API – The response is a Json file which is turned into a Pandas Data frame using .json_normalize(). The following information is obtained(columns): Flight number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude and Latitude.
- Web scrapping the SpaceX's Wikipedia page –The scraping is done using the library requests and beautiful soup. We obtain the following columns: Flight Number, launch Site, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

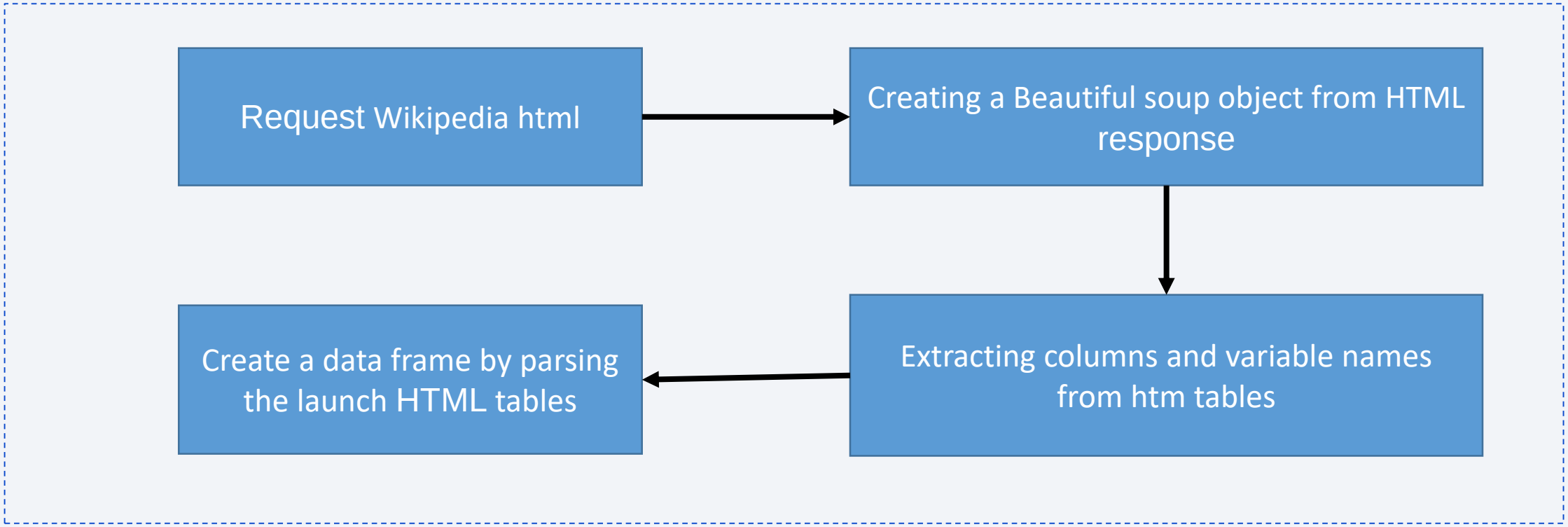
For each of the data collection methods there is a dedicated slide highlighting the various specific stages.

Data Collection – SpaceX API



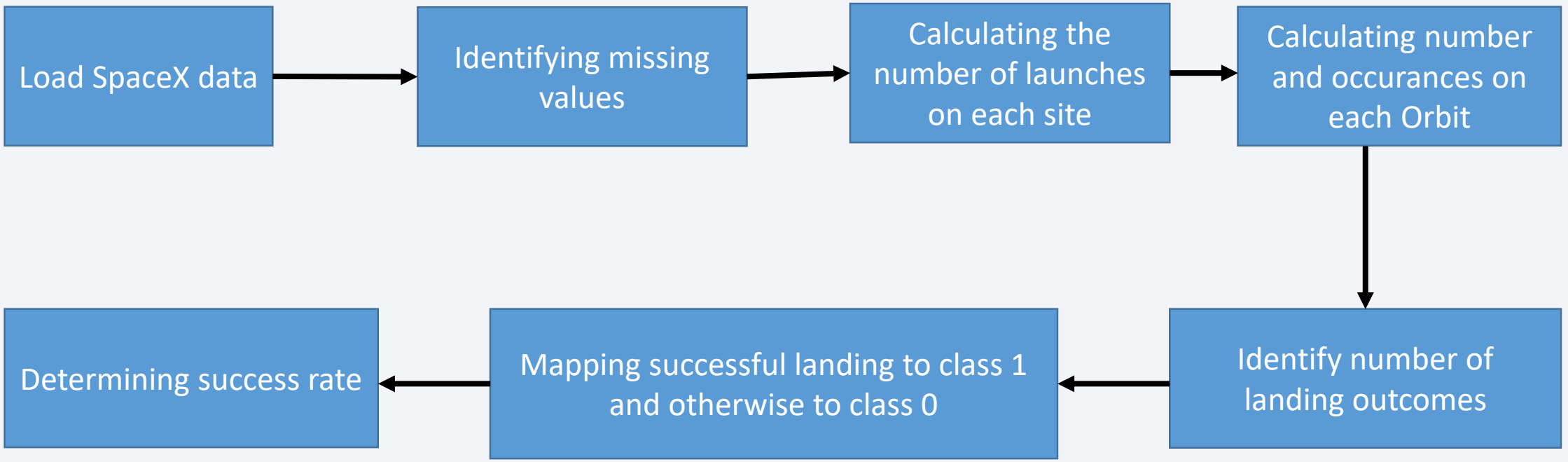
Github url: https://github.com/Konzisam/Applied_DataScience_capstone_project/blob/master/Data_collection_Spacex_RESTful_API.ipynb

Data Collection - Scraping



Githb url: https://github.com/Konzisam/Applied_DataScience_capstone_project/blob/master/Data_collection_web scraping_Wikipedia.ipynb

Data Wrangling



Github url: https://github.com/Konzisam/Applied_DataScience_capstone_project/blob/master/Data_wrangling_spacex.ipynb

EDA with Data Visualization

Exploratory data analysis was performed to determine the various relationships as well as feature engineering for compatibility with machine learning model.

Matplotlib and seaborn libraries were used to plot graphs of: Payload Mass vs. Flight Number, Flight Number vs. launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs. Orbit, Success rate vs year.

EDA with SQL

Further Exploratory analysis is done using SQL to better understand the dataset.

Some of the queried information include:

- Unique sites
- Total load mass in a Lauchsite
- Average Payload Mass carried by a specific booster
- First date of successful landing
- Successful landing outcomes with payload greater than 4000

Build an Interactive Map with Folium

In this section the Folium library was used to perform various tasks such as:

- Marking Launch sites
- Proximity examples for example: Railway, Highway, coast and city
- Adding circles with text labels
- Adding markers on a map

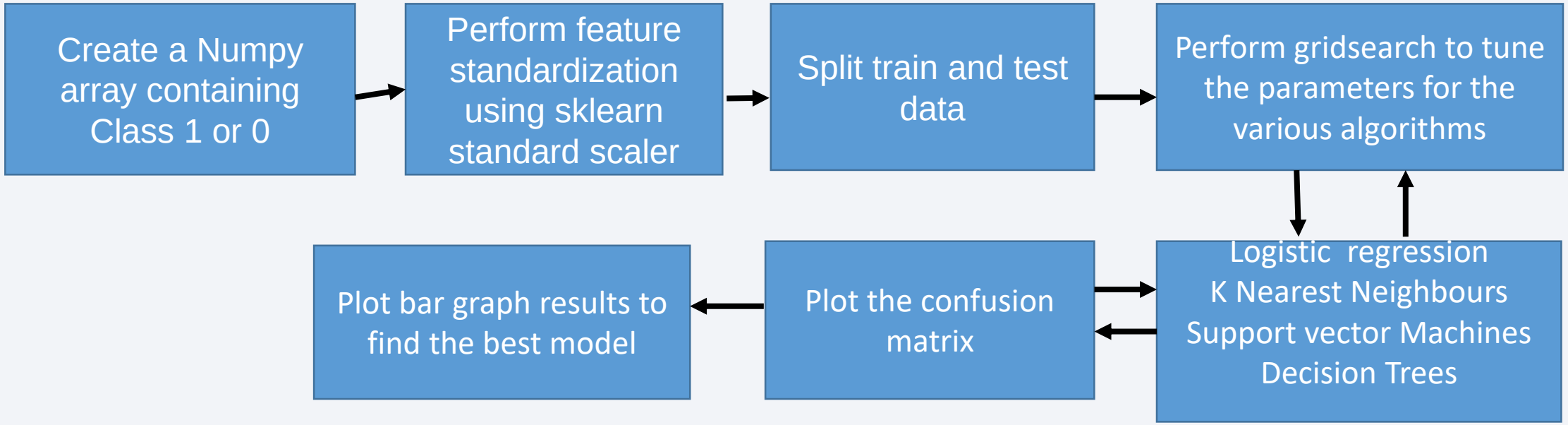
Build a Dashboard with Plotly Dash

A Dash board with a pie chart to visualize launch site success rate and scatter plot for the same was built using Plotly Dash.

The dashboard also has a drop down to select the site and the result is a pie chart highlighting the success rate of the specific site or all sites

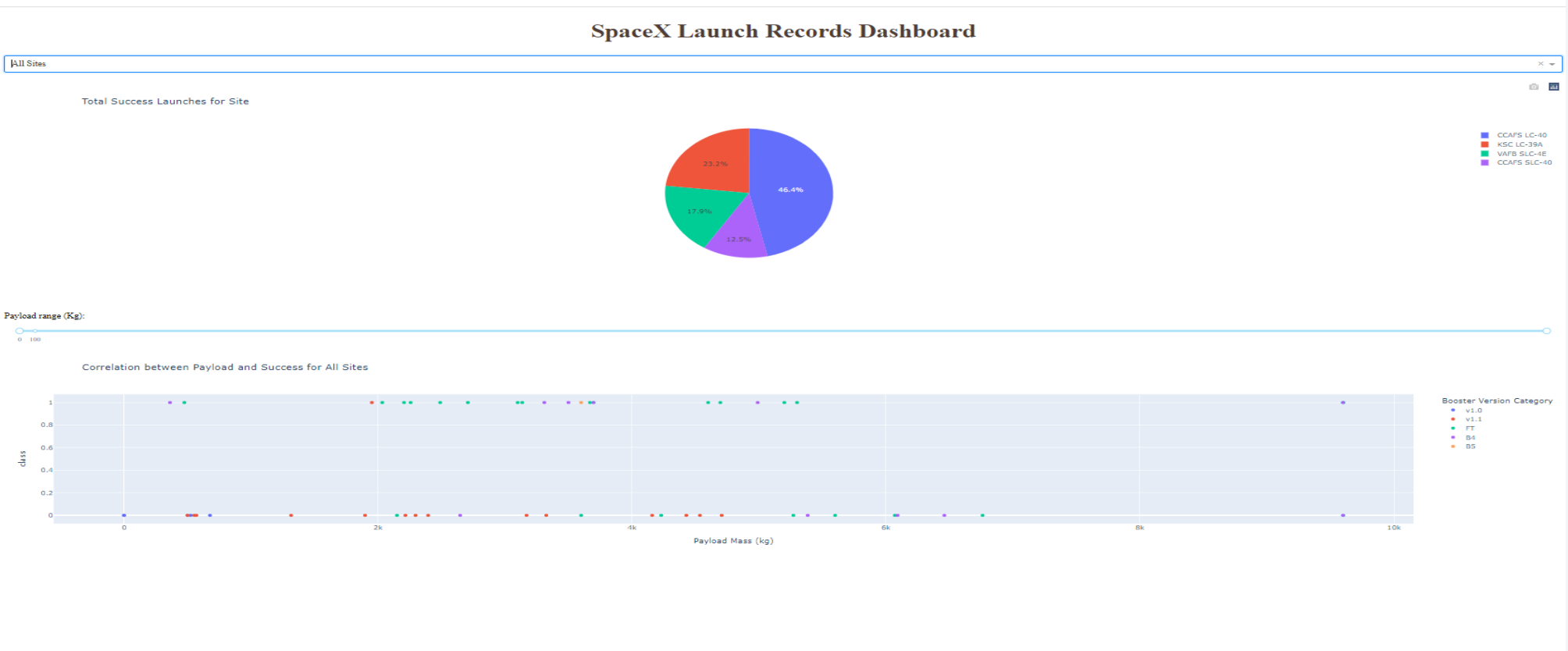
The scatter plot takes the site input and has a slider to select the Payload Mass over the range 0 – 10000kg. This is useful for visualizing the success across different launch Sites , Payload mass as well as booster version.

Predictive Analysis (Classification)



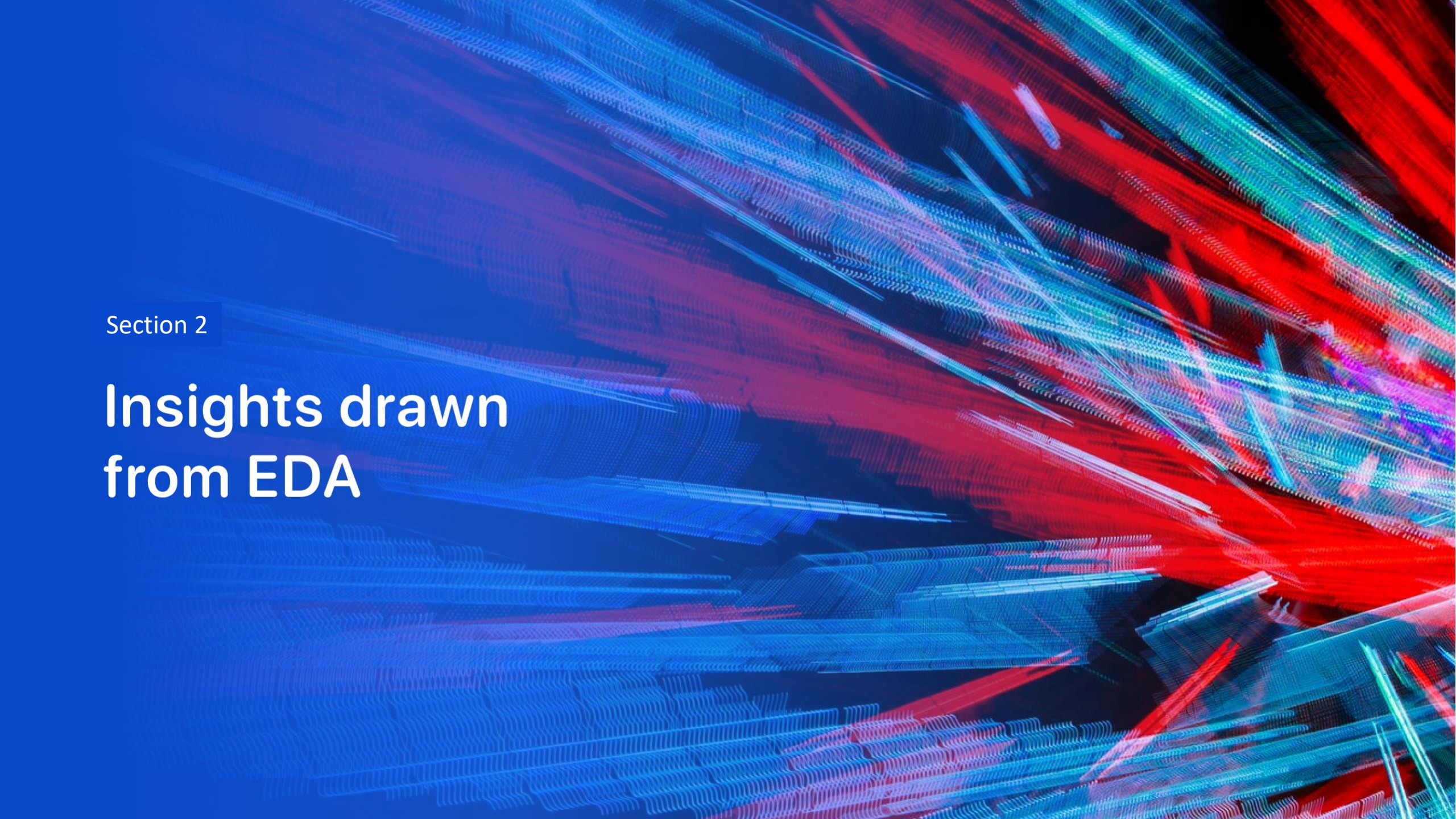
Github Url:
https://github.com/Konzisam/Applied_DataScience_capstone_project/blob/master/Machine_Learning_Prediction_SpaceX.ipynb

Results



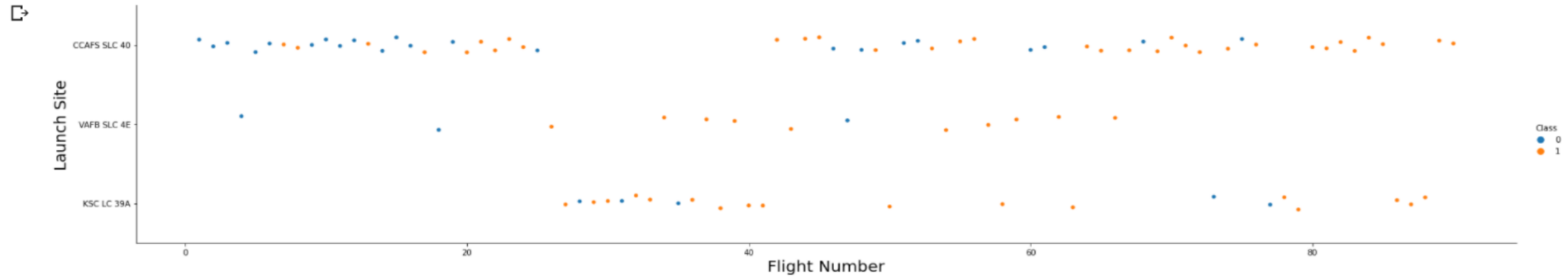
The success rate from Exploratory data analysis yielded a 66% success rate

The predictive analysis using machine learning results to model accuracy of 83.33%


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

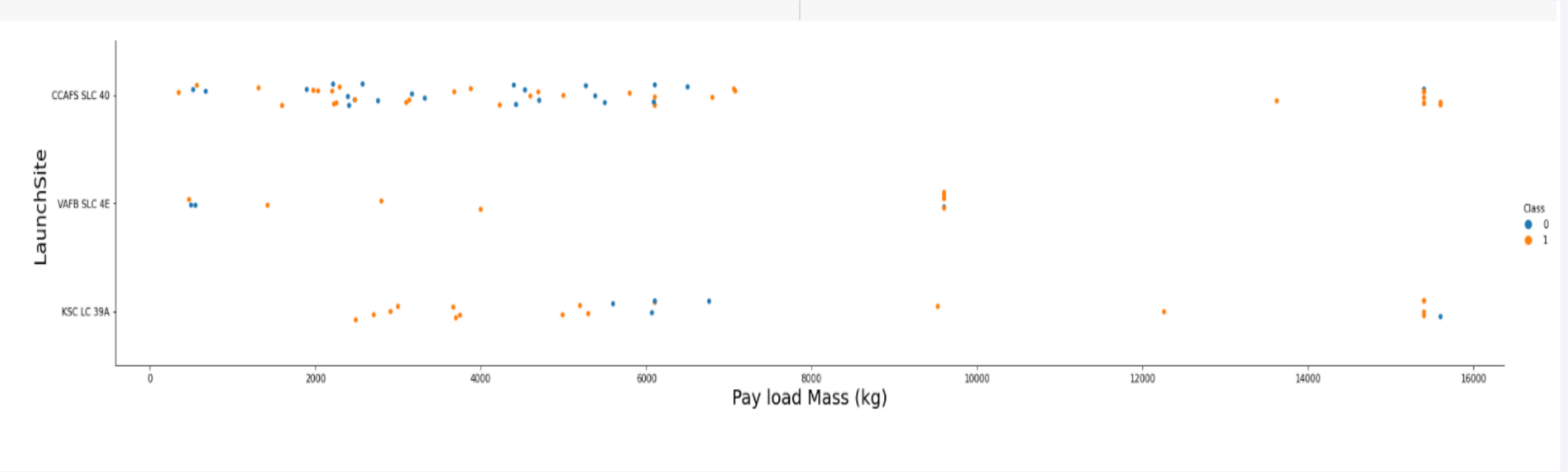


The orange dots represent successful launch, whereas the blue represent unsuccessful launch

The success rate of the site CCAFS SLC 40 is seen to be lower compared to the other sites. The site also has the highest launch volume. Al flights since around the 80th flight launched successfully.

It is also seen that as the number of flights increased(consequently time), the success rate significantly increased

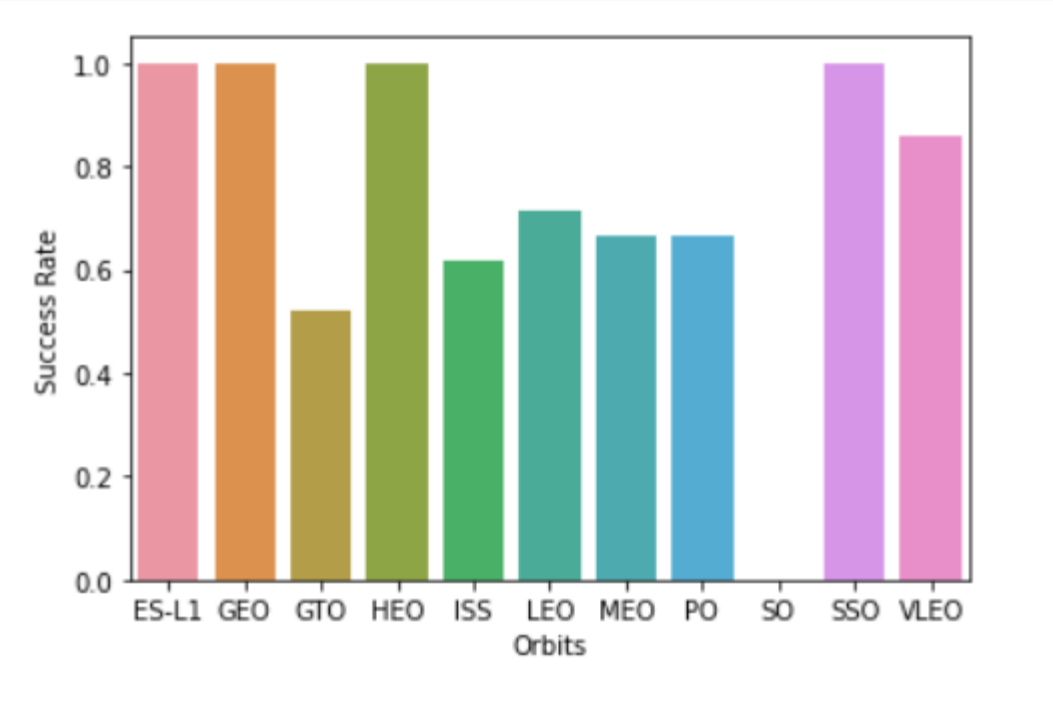
Payload vs. Launch Site



We observe that for the VAFB-SLC Launch site there are no heavy load mass(greater than 10000kg) rockets launched.

It is also observed that heavy Payload Mass rockets are more likely to land successfully

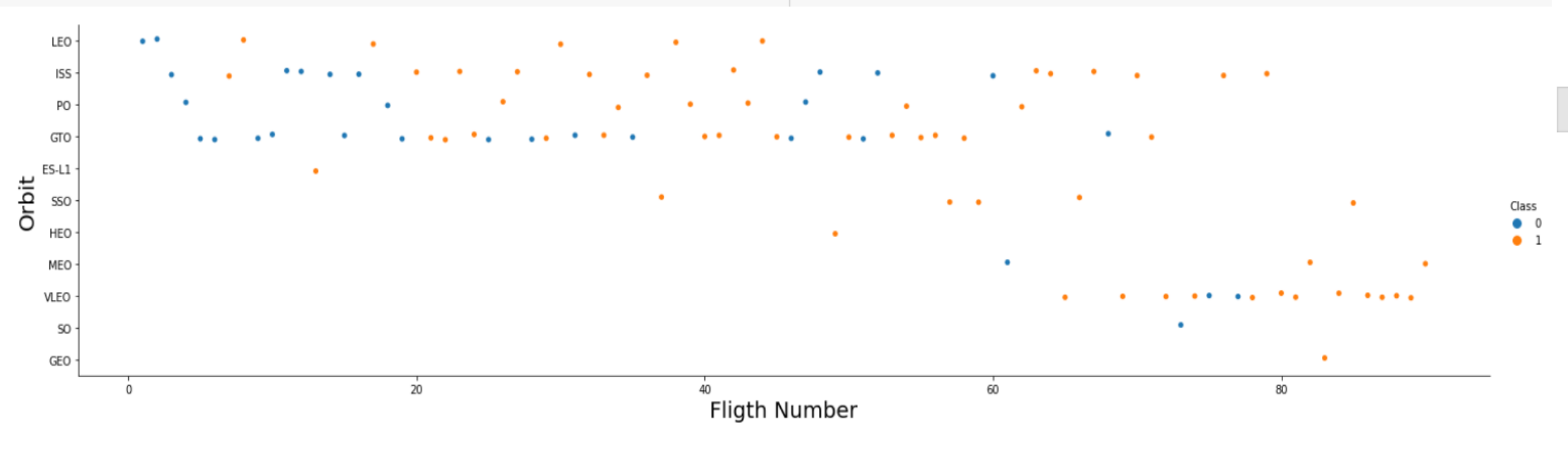
Success Rate vs. Orbit Type



The scale for the graph is 1 for 100% and 0 for 0%

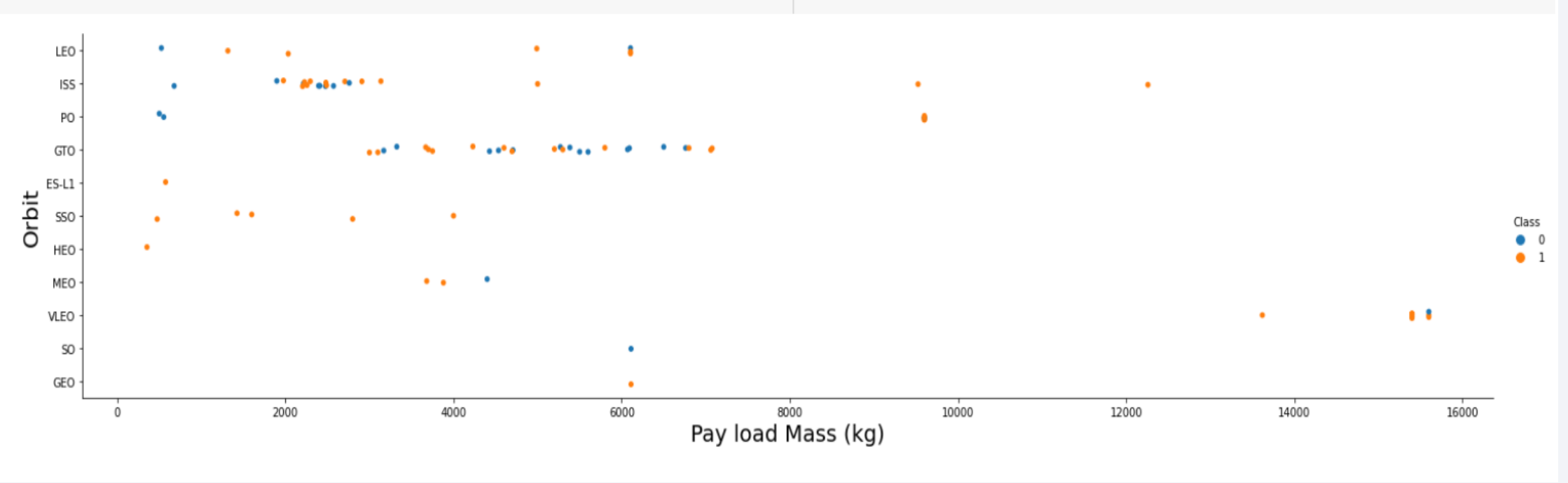
We can see that some of the Orbits exhibit 100% success rate but contain only a single launch and 5 launches for the SSO orbit. This might not mean much compared to the other orbits from which more flights have been launched.

Flight Number vs. Orbit Type



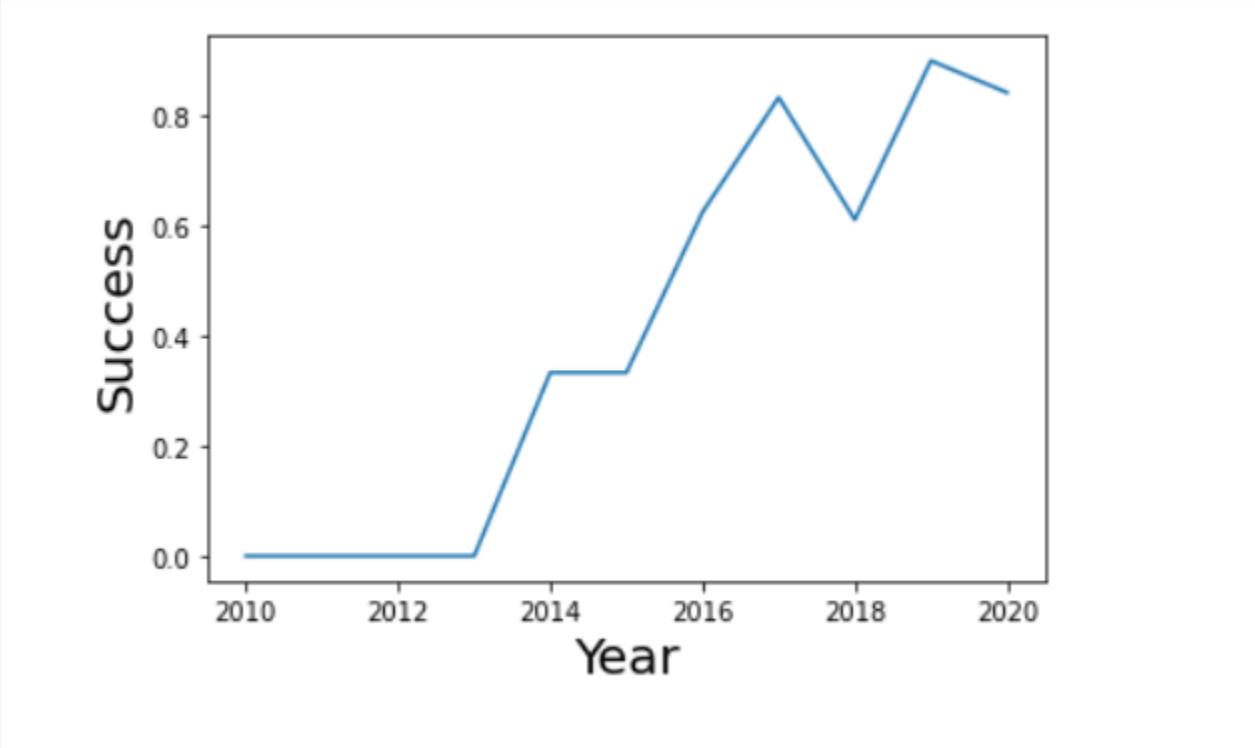
For the LEO orbit, the success rate seems to be related to the number of flight unlike in other orbits such as GTO Orbit.

Payload vs. Orbit Type



Its is seen that with heavy payloads the successful landing rate is more for Polar, LEO and ISS

Launch Success Yearly Trend



It can be observed that the success rate kept increasing since 2013 to 2020.

All Launch Site Names

```
run("select distinct(Launch_Site) from Spacex")
```

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

The distinct statement is used with the select statement to find the unique Launch Sites.

Launch Site Names Begin with 'CCA'

```
run('''select *
        from Spacex
          where Launch_Site like 'CCA%'
            limit 5
          ''')
```

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Finding the 5 records where launch sites begin with `CCA`

Total Payload Mass

```
[ ]  run('''select sum(PAYLOAD_MASS__KG_) as Total_Payload_mass
            from Spacex
              where Customer = 'NASA (CRS)'

              ''')
```

	Total_Payload_mass
0	45596

The total payload carried by boosters from NASA by calculating the sum and specifying customer using where clause

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
run('''select AVG(PAYLOAD_MASS__KG_)  as Average_Payload_Mass
        from Spacex
          where Booster_Version = 'F9 v1.1'
          ''')
```

	Average_Payload_Mass
0	2928.4

Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
run('''select min(Date) as first_date
        from Spacex
          where Landing_Outcome = 'Success (ground pad)'
          ''')
```

	first_date
0	22-12-2015

- 22-12-2015 is the first successful landing outcome on ground pad
- The min() function is used as well as specifying the where clause

Successful Drone Ship Landing with Payload between 4000 and 6000

```
run('''select Booster_Version
        from Spacex
          where Landing_Outcome ='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
          ''')
```

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

List of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. The ‘between, and’ clause is used in the where clause

Total Number of Successful and Failure Mission Outcomes

```
[ ]  run('''select Mission_outcome, count(*)
            from Spacex
              group by Mission_outcome
              ''')
```

	Mission_Outcome	count(*)
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

- The query outputs the total number of successful and failure mission outcomes
- The group by clause is used along with the SELECT statement

Boosters Carried Maximum Payload

```
run('''select Booster_Version
        from Spacex
          where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_ )
                                     from Spacex)
          ''')
```

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

- The query lists the names of the booster which have carried the maximum payload mass
- A sub query is used

2015 Launch Records

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
[ ]  run('''select Landing_Outcome, Booster_Version,Launch_Site
            from Spacex
              where Date like '%2015' and Landing_outcome ='Failure (drone ship)'
              ''')
```

	Landing_Outcome	Booster_Version	Launch_Site
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The query lists 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

Task 10

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

```
run('''select Landing_Outcome, count(*) as count_
        from Spacex
          where Landing_Outcome like 'Success%' and Date between '04-06-2010' and '20-03-2017'
           group by Landing_Outcome
            order by count_ desc
          ''')
```

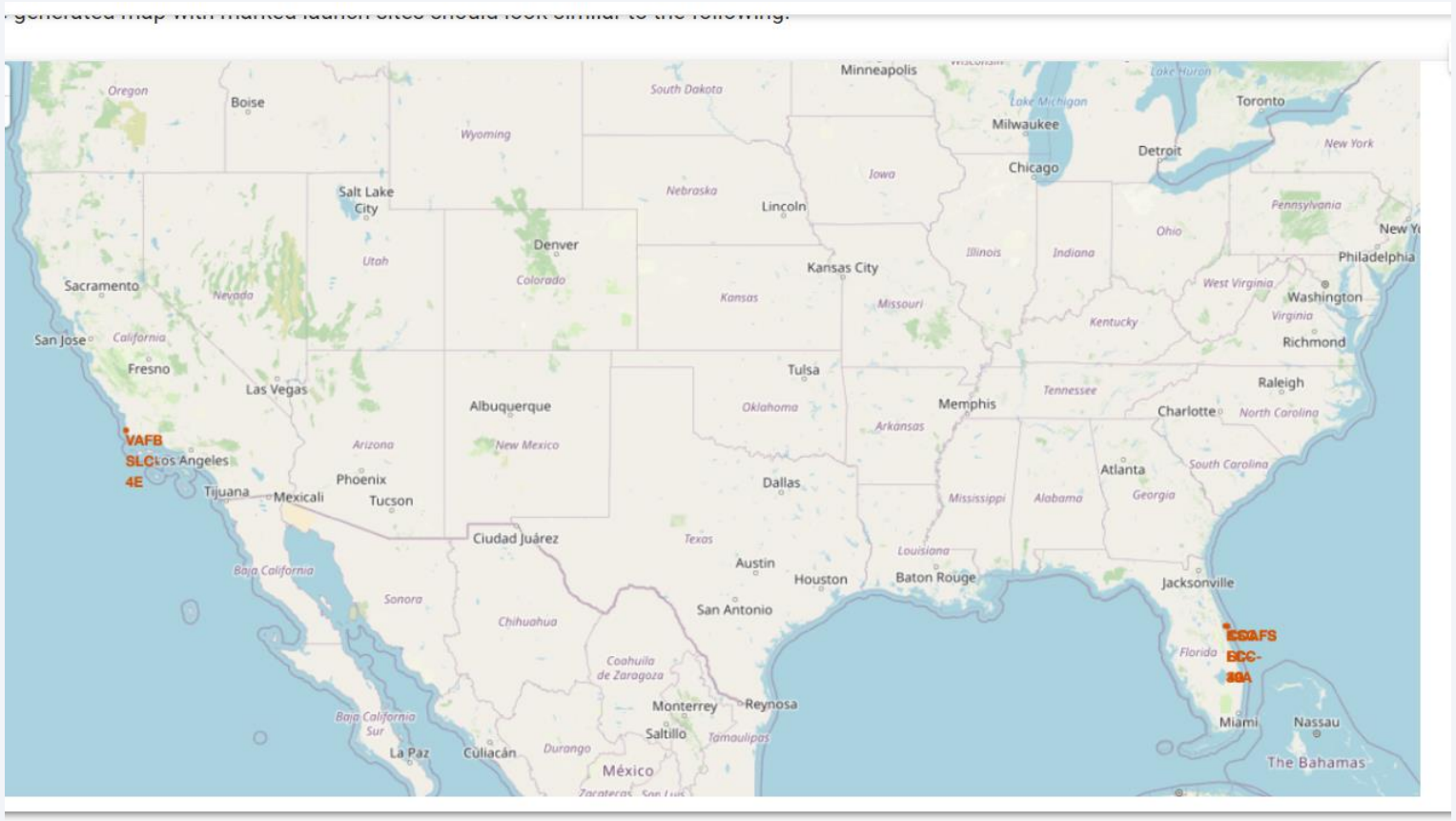
	Landing_Outcome	count_
0	Success	20
1	Success (drone ship)	8
2	Success (ground pad)	6

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

Section 3

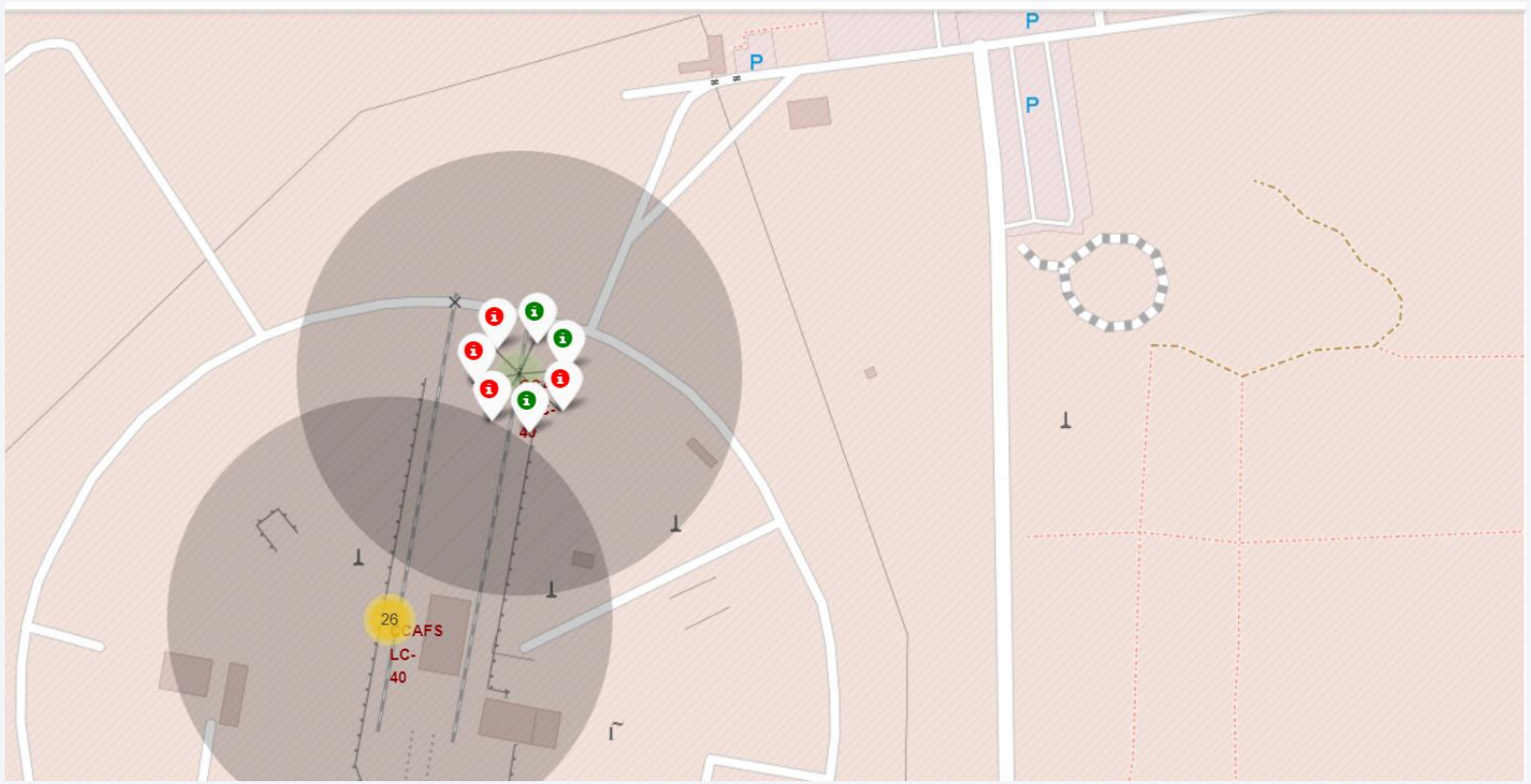
Launch Sites Proximities Analysis

Location of Launch sites with Folium



It is observed that the Launch Sites are located are located close to the ocean. Its also noted that the sites on the South East coast(Florida)

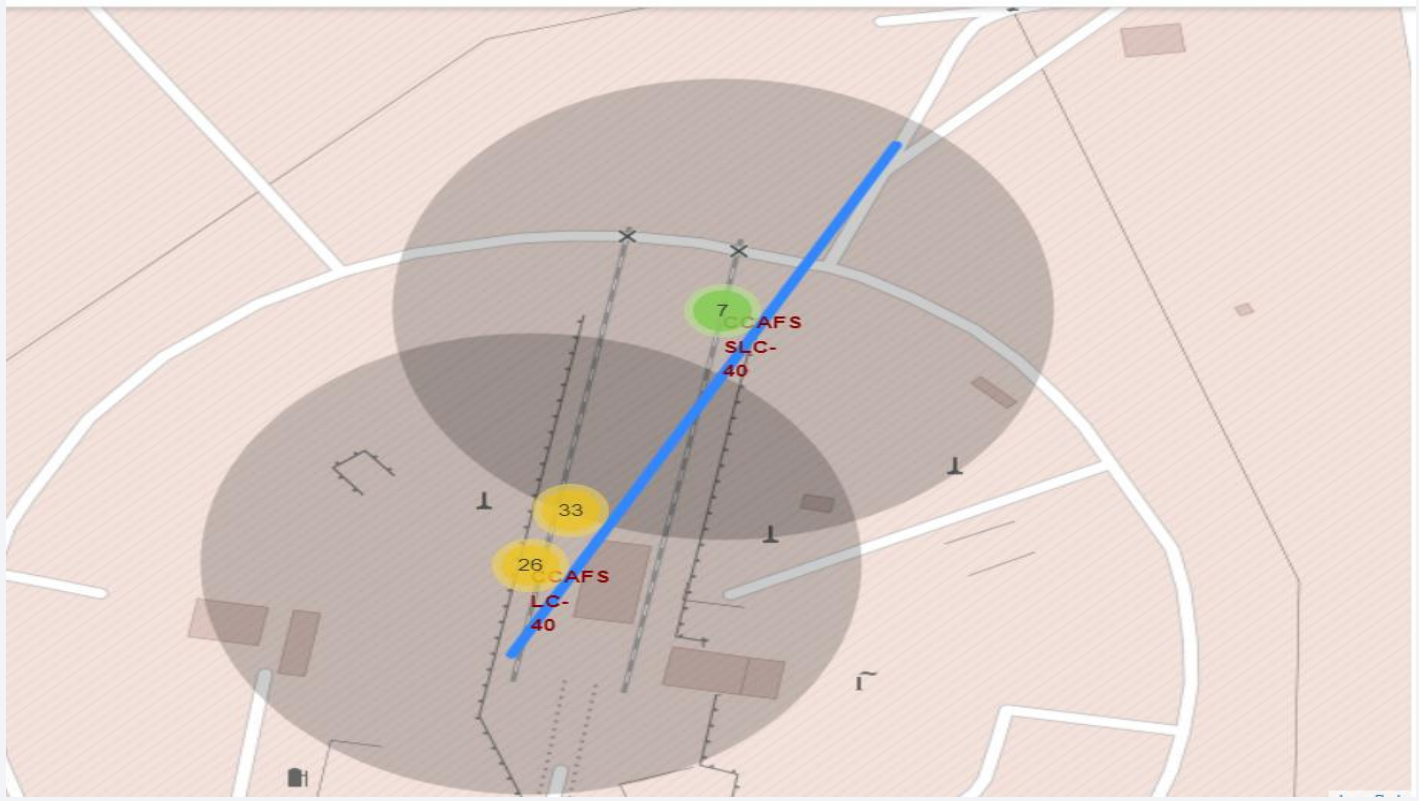
Color marked labels for Successful/Failed Launches



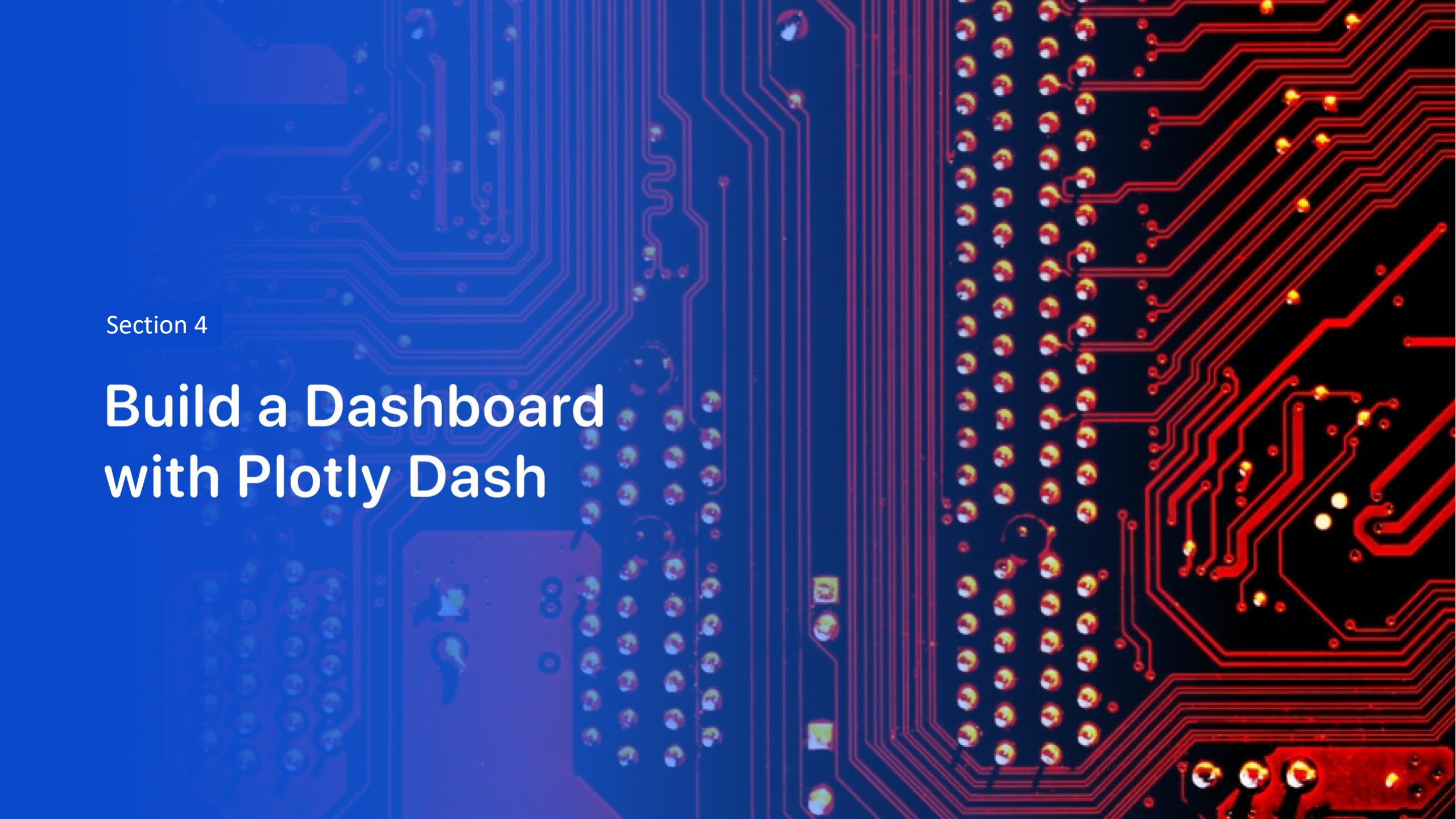
The green icons represent a successful landing and red otherwise.

For the Launch Site CCAFS SLC -40 in the screen shot, it is seen that there are 3 successful landings and 4 unsuccessful.

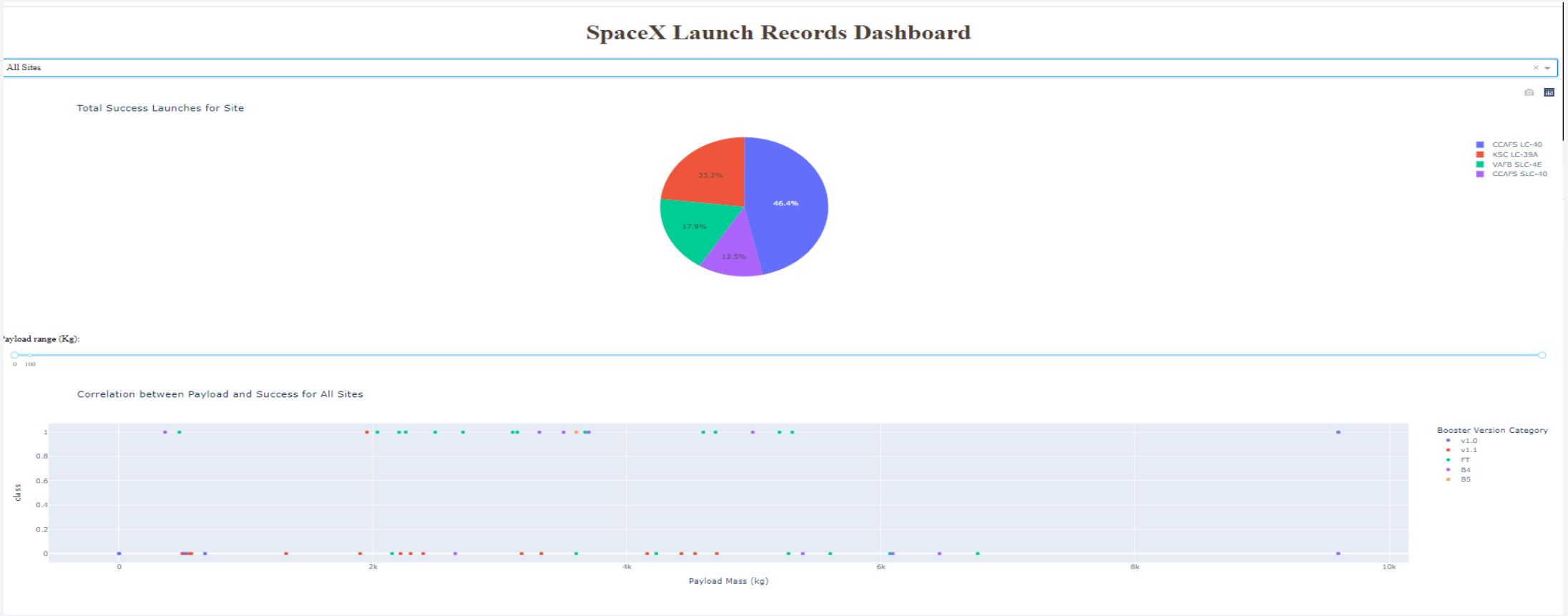
Launch Site to its proximities



It is observed that the launch sites are all located near railway lines, highways and far from cities. Which makes sense in terms of transport and consequently safety, since cities have high population.

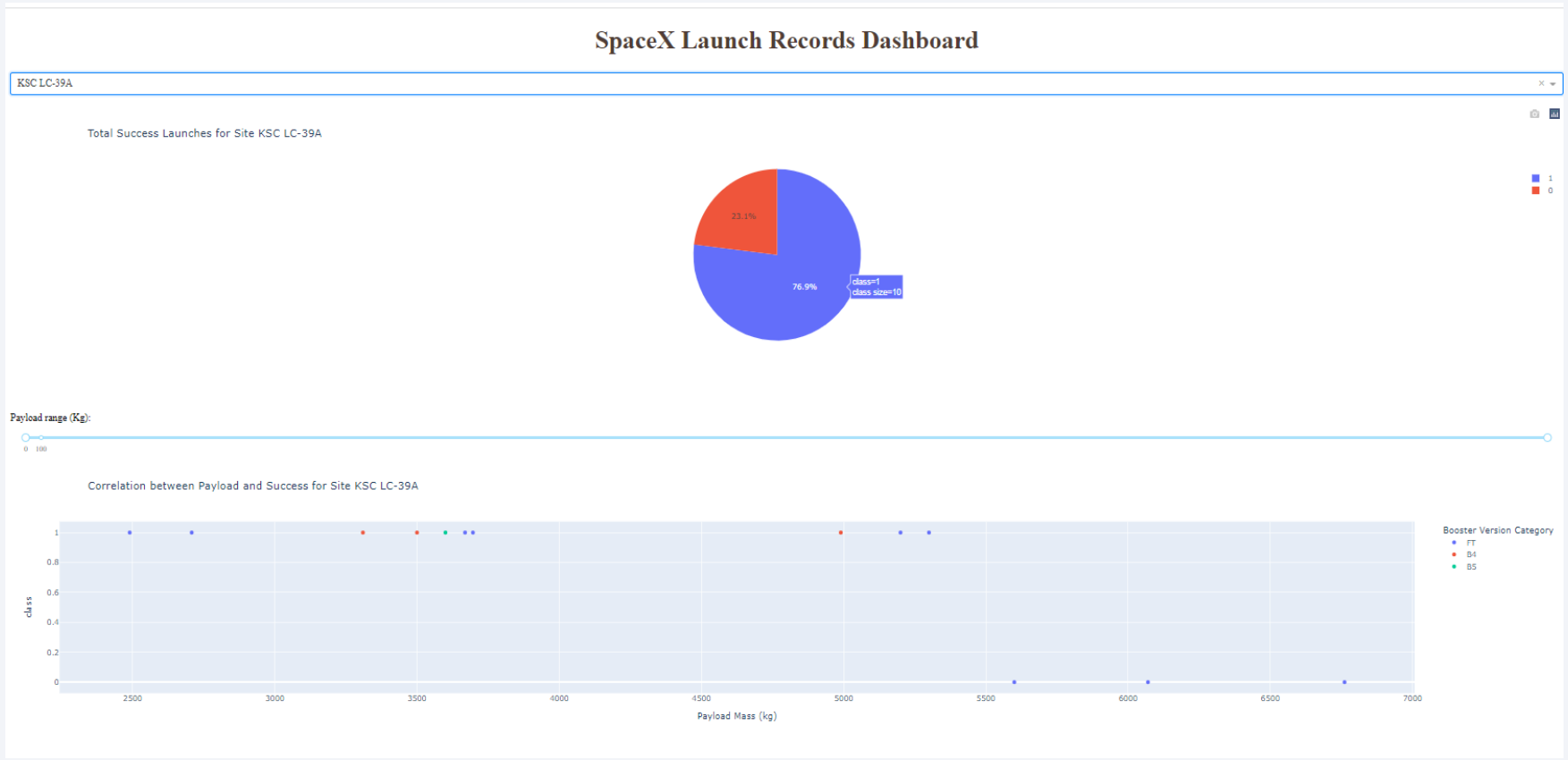

Section 4
Build a Dashboard with Plotly Dash

Visual display of Successful Launches in all Sites



The pie chart shows a comparison of the successful launches for all the sites. The site KSC LC -39 is seen to have the highest success count whereas CCAFS SLC -40 has the least. The low count of successful launches may be attributed to smaller sample.

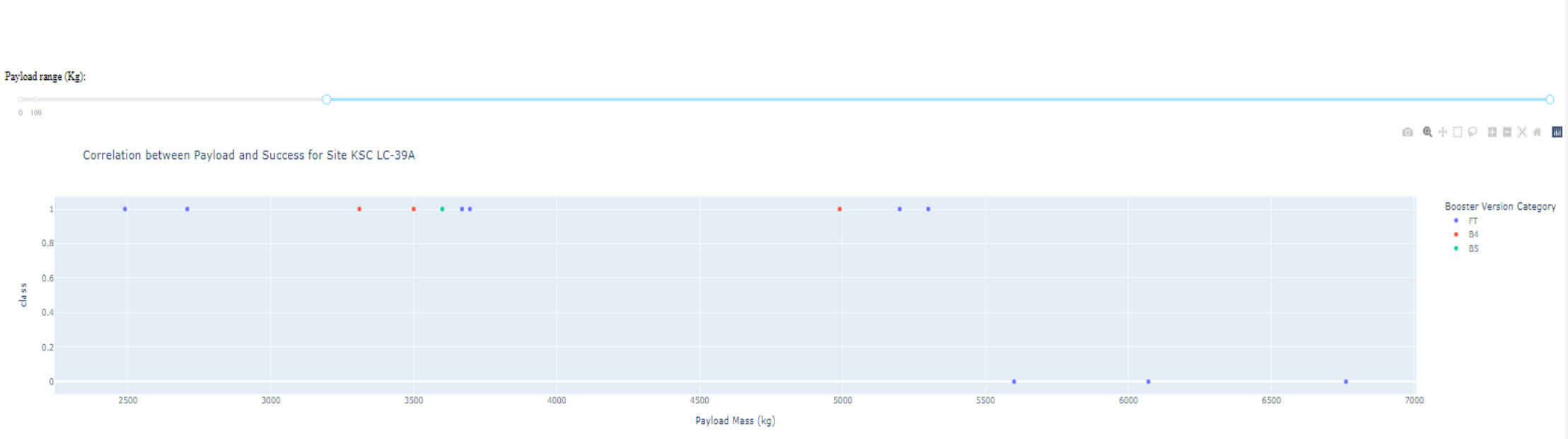
Site with Highest success rate



The Landing site KSC LC – 39A rate represented by color purple.

This site has the highest percentage of landings of 76.9% . I has only 3 failed landing and 10 successful landings

Payload vs Launch Outcome scatter plot

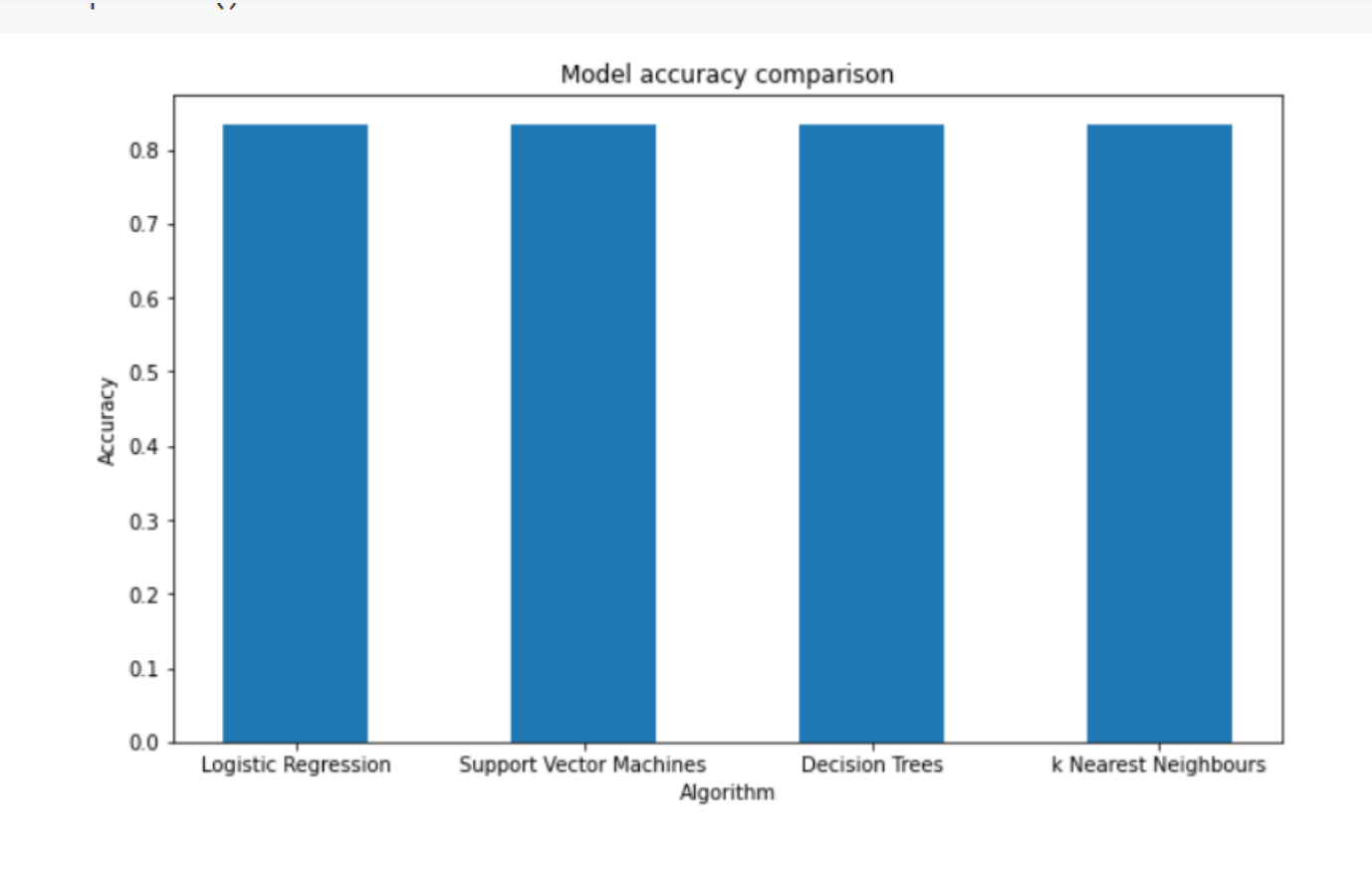


The different colors of the scatter plot show the different booster versions. The graphical view also make sit very clear to visualize the classes(success or failure).It can be seen that the booster version v1.1 has the most failures whereas the booster FT has the highest success rate

Section 5

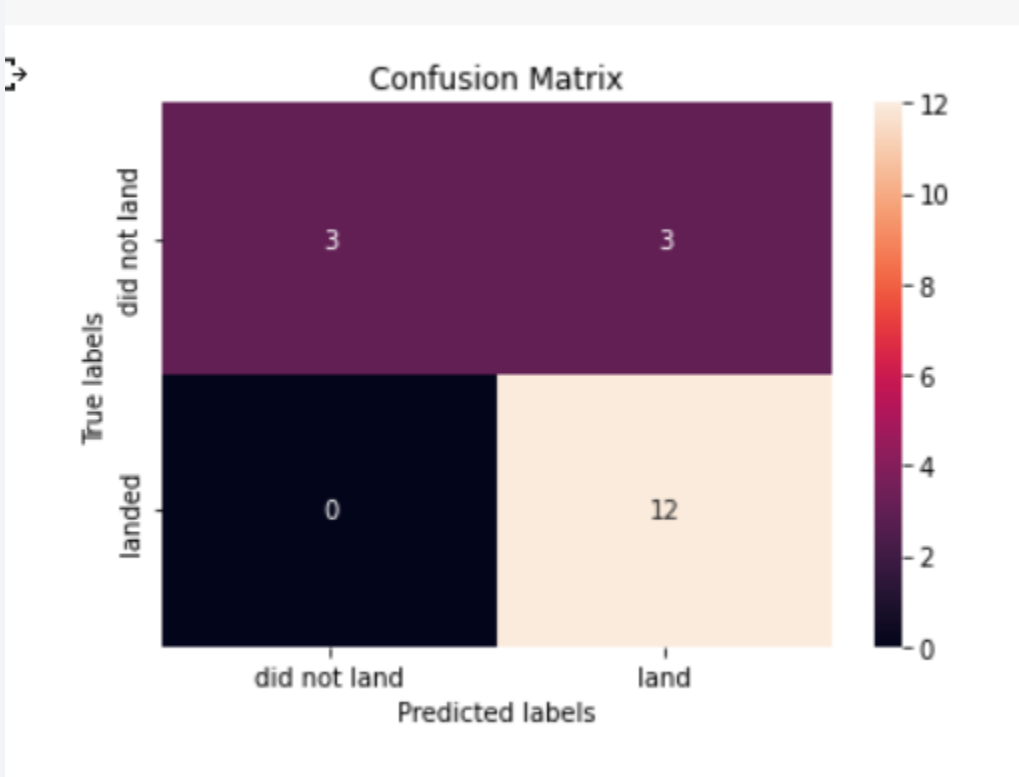
Predictive Analysis (Classification)

Classification Accuracy



As seen in the graph, all the models yielded the same classification score.

Confusion Matrix



The confusion matrix is the same for all models as well: It is noted that the model has a problem of over predicting false positives(predicting successful landing when the label was unsuccessful landing).

Conclusions

- A machine learning model for the company SpaceY with a success rate of 83.33 % was built.
- A dashboard to visualize the various factors affecting Launch success or failure was built along with various plots to support the arguments.
- With this information SpaceY can determine if SpaceX will re use their first stage and thus with this information predict their outcome and expense of a launch.
- There we setbacks such as the model predicting false positives which can be attributed to small data set used. The accuracy of the model can be increased by using a larger dataset.

Appendix

Github Url:

https://github.com/Konzisam/Applied_DataScience_capstone_project

Thank you to all the instructors:

Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Thank you!