IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abhishek Siwakoti

27th May 2023

Outline

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

Executive Summary

- Summary of methodologies used in this project:
  - Data Collection Using SpaceX API and web scrapping from Wikipedia
  - Data Wrangling
  - Exploratory Data Analysis (EDA) Using Python and SQL
  - Interactive Data Visualization with Folium and Plotly Dash
  - Predictive analysis using machine learning methods
- Summary of all results we'll be discussing in this project:
  - Results of EDA
  - Results of Interactive Data Visualizations
  - Results of predictive analysis

Introduction

- Project background and context
  - 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.
- Problems you want to find answers
  - What factors influence the successful landing of the rocket?
  - How does the interaction amongst these factors influence the rate of a successful or a failed landing?
  - Using these information, what conditions are to be established in order to have the highet possibe probability of a successful landing?

Section 1

Methodology

Methodology

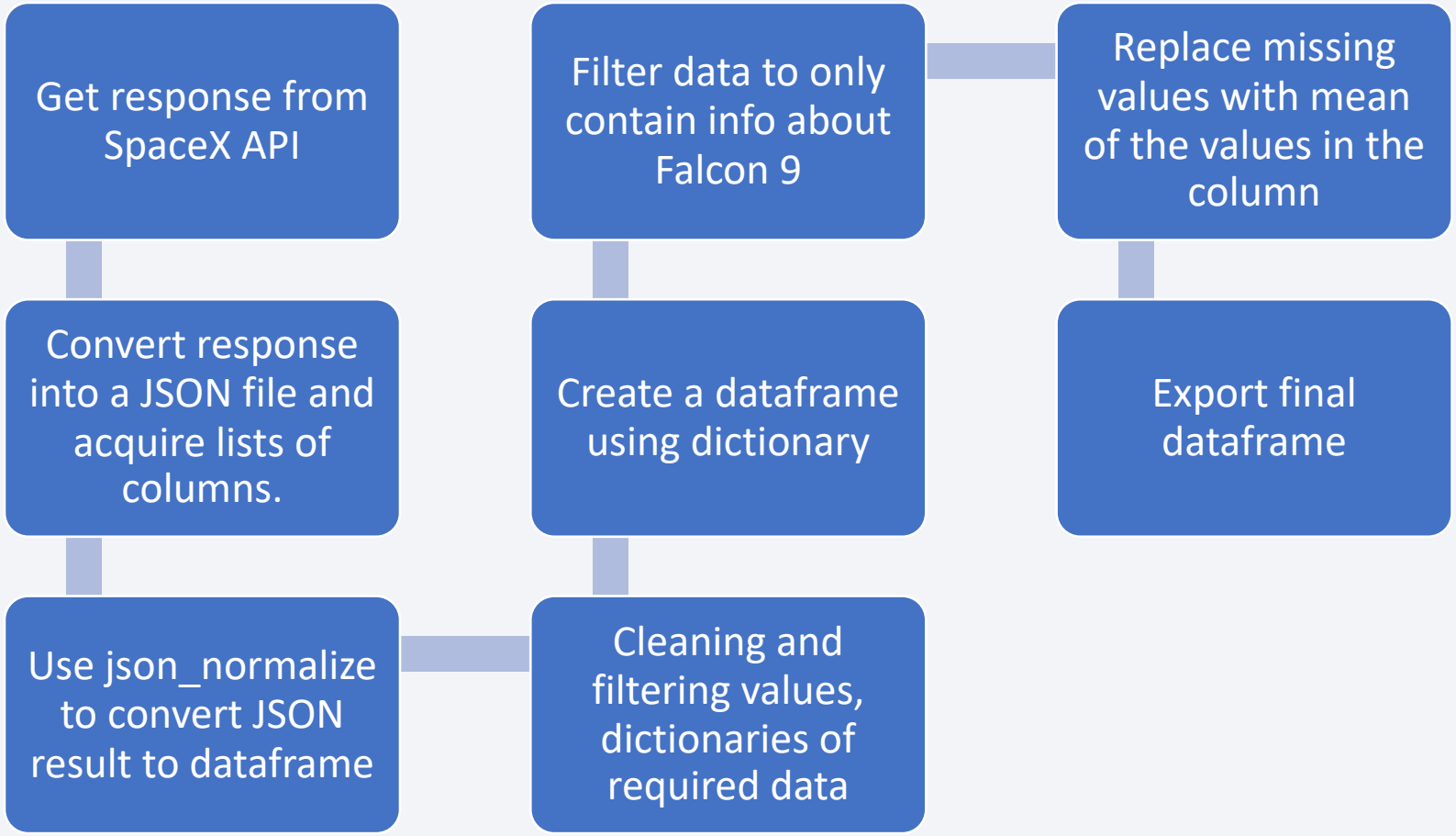
Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scrapping from Wikipedia
- Perform data wrangling
  - One-hot encoding was used in order to classify data and create various columns for labels of factors influencing landing outcome, as well as classify them as successful or unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV and machine learning, predictive analysis was performed.

Data Collection

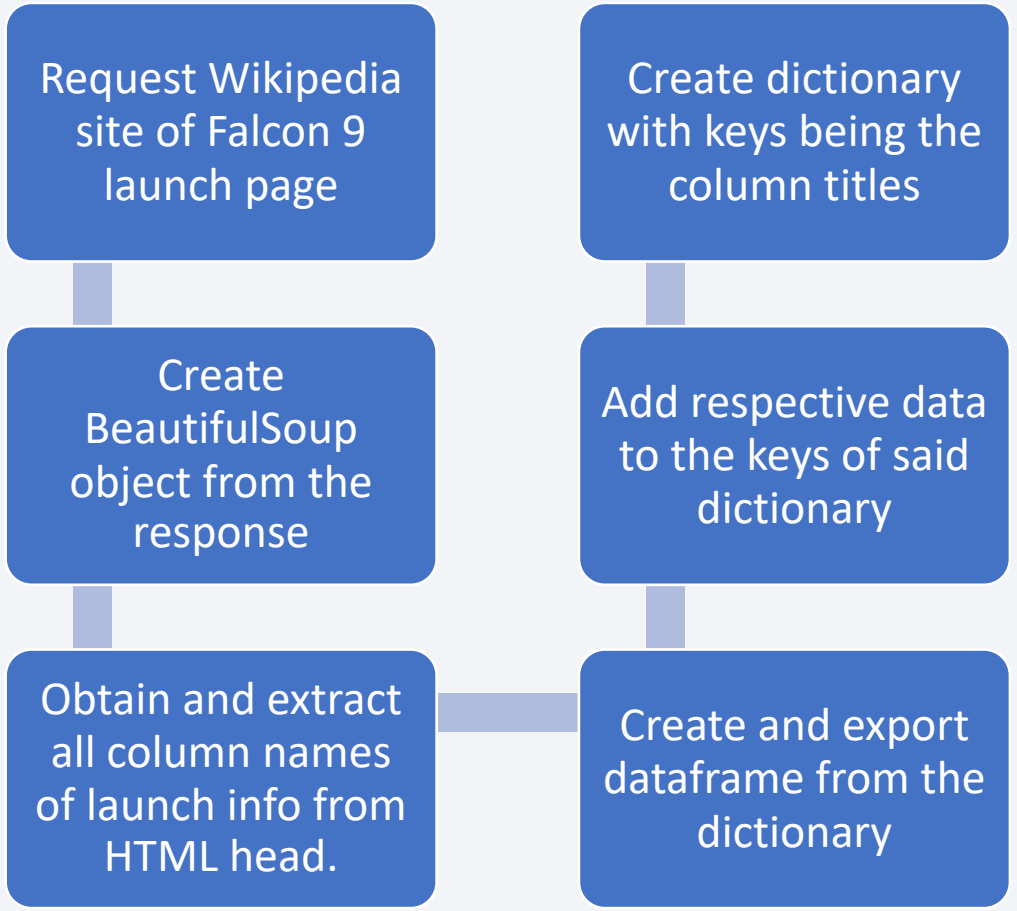
- Describe how data sets were collected.
  - Data sets were collected from two sources, one being SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - The other being web scrapping from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API



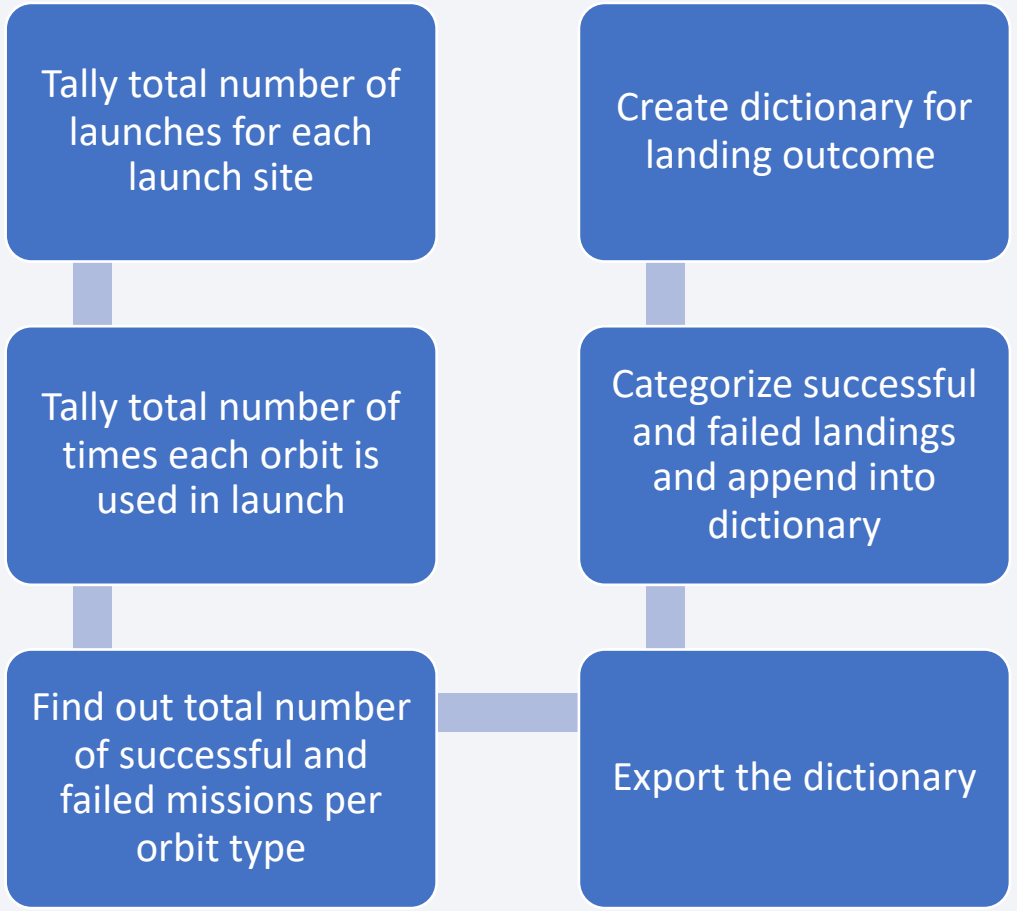
Github URL:
https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/1_Data_Collection_API.ipynb

Data Collection - Scraping



Github URL: https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/2_Data_Collection_Web_Scrapping.ipynbç

Data Wrangling



Github URL:
https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/3_Data_Wrangling.ipynb

EDA with Data Visualization

- The following charts were plotted:
  - Scatter plot between Flight Number and Launch Site, Payload Mass and Launch Site, Flight Number and Orbit, Payload Mass and Orbit
  - Bar Graph showing success rates of all orbit types
  - Line plot showing relationship between flight number and orbit
  - Line chart showing trend of success rate by year
- The purpose of these plots were to visually see if there is a possible relationship amongst these variables that need further analysis and to use them for predictive analysis.
- Github URL: https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/5_EDA_Data_Visualization.ipynb

EDA with SQL

- The following SQL queries were performed as a part of EDA:
  - Names of the unique launch sites in the space mission.
  - 5 records where the launch sites begin with the string 'KSC'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Listing date where the successful landing outcome in drone ship was achieved.
  - Listing names of boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing total number of successful and failure mission outcomes
  - Listing names of booster_versions which have carried the maximum payload mass.
  - Listing records which will display the month names, successful landing outcomes in ground pad booster versions, launch site for the months in year 2017
  - Ranking count of successful landing outcomes between 04-06-2010 and 20-03-2017 in descending order.

Git URL: https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/4_EDA_With_SQL.ipynb

Build an Interactive Map with Folium

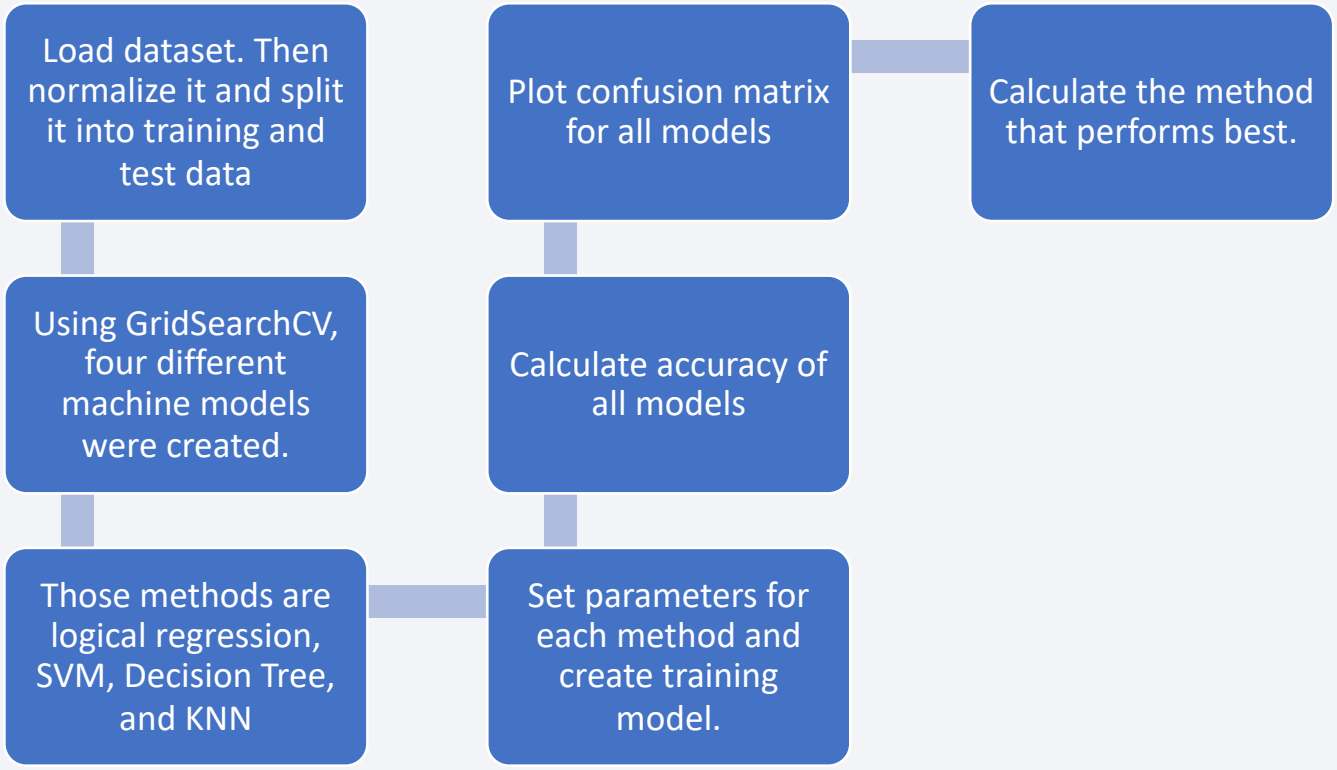
- Using Folium, markers, circles, lines and marker clusters were added in maps using Folium.
- Folium was used to mark things such as launch sites, areas of successful and failed landings (green for successful and red for failed). Things such as distance between launch sites to railways, cities, coastlines were also marked.
- This is done to get a better visual understanding of possible safe launch sites, while keeping distance with other key locations into consideration.
- Github URL: https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/6_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- The following plots and interactions were added:
  - Dropdown option to choose launch site/s.
  - Pie chart displaying success rate of launch site/s chosen.
  - Range slider which allows you to select payload mass.
  - plot with the x axis to be the payload mass and the y axis to be the launch outcome

- Github URL: https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/7_Interactive_Dashboard_Plotly.py

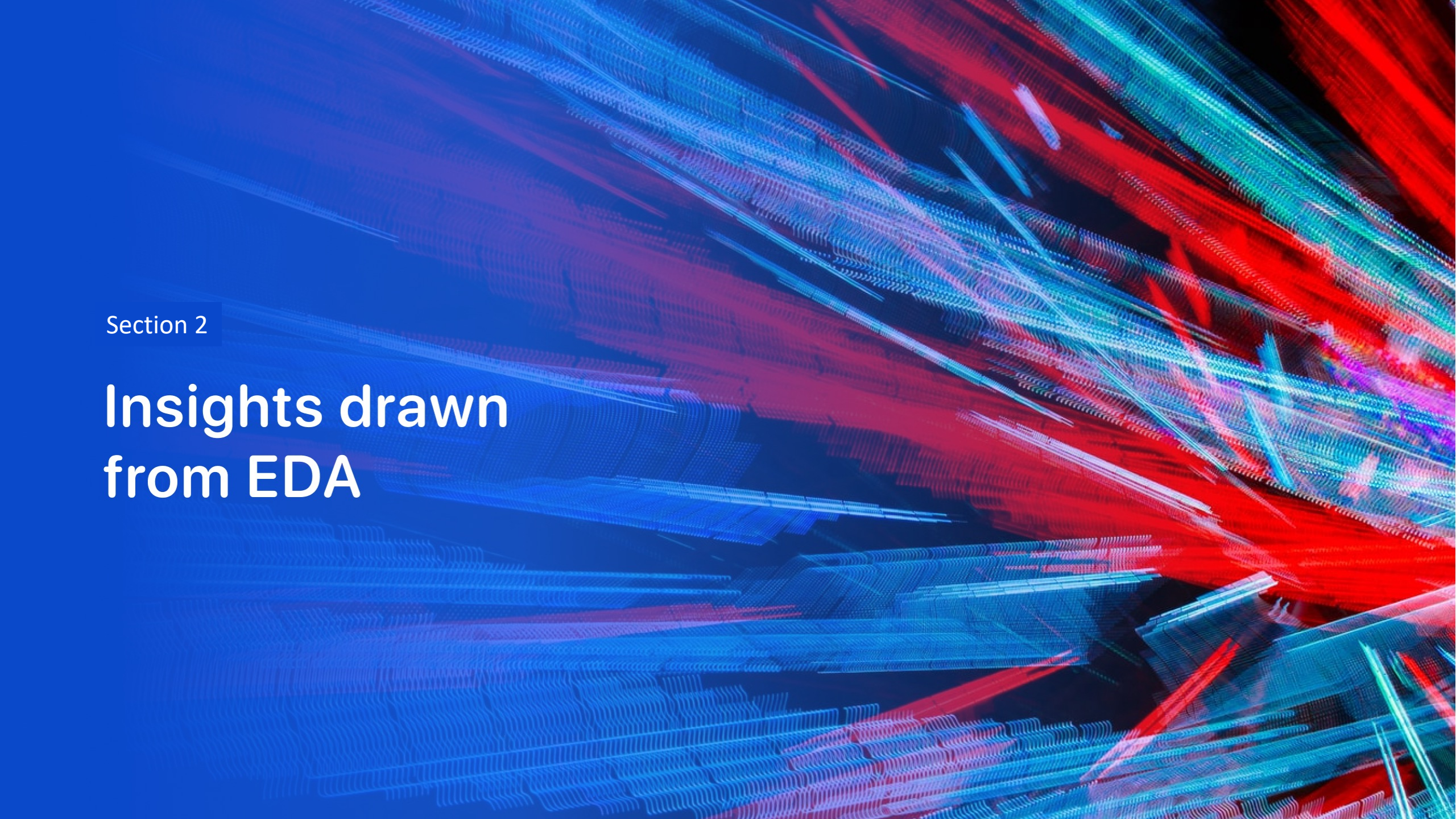
Predictive Analysis (Classification)



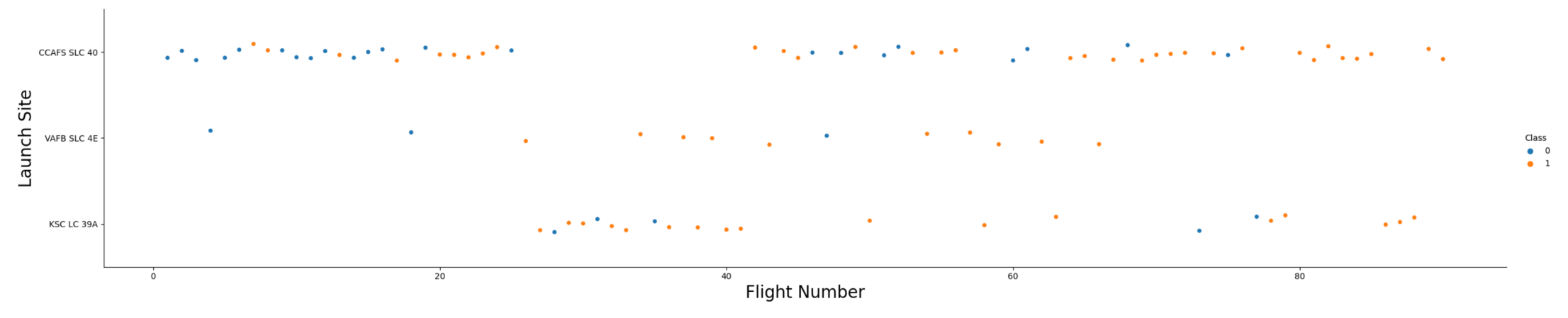
Github URL:
https://github.com/abhisheksiwakoti/IBM_Data_Science_Capstone_Project/blob/main/8_Machine_Learning_Prediction.ipynb

Results

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

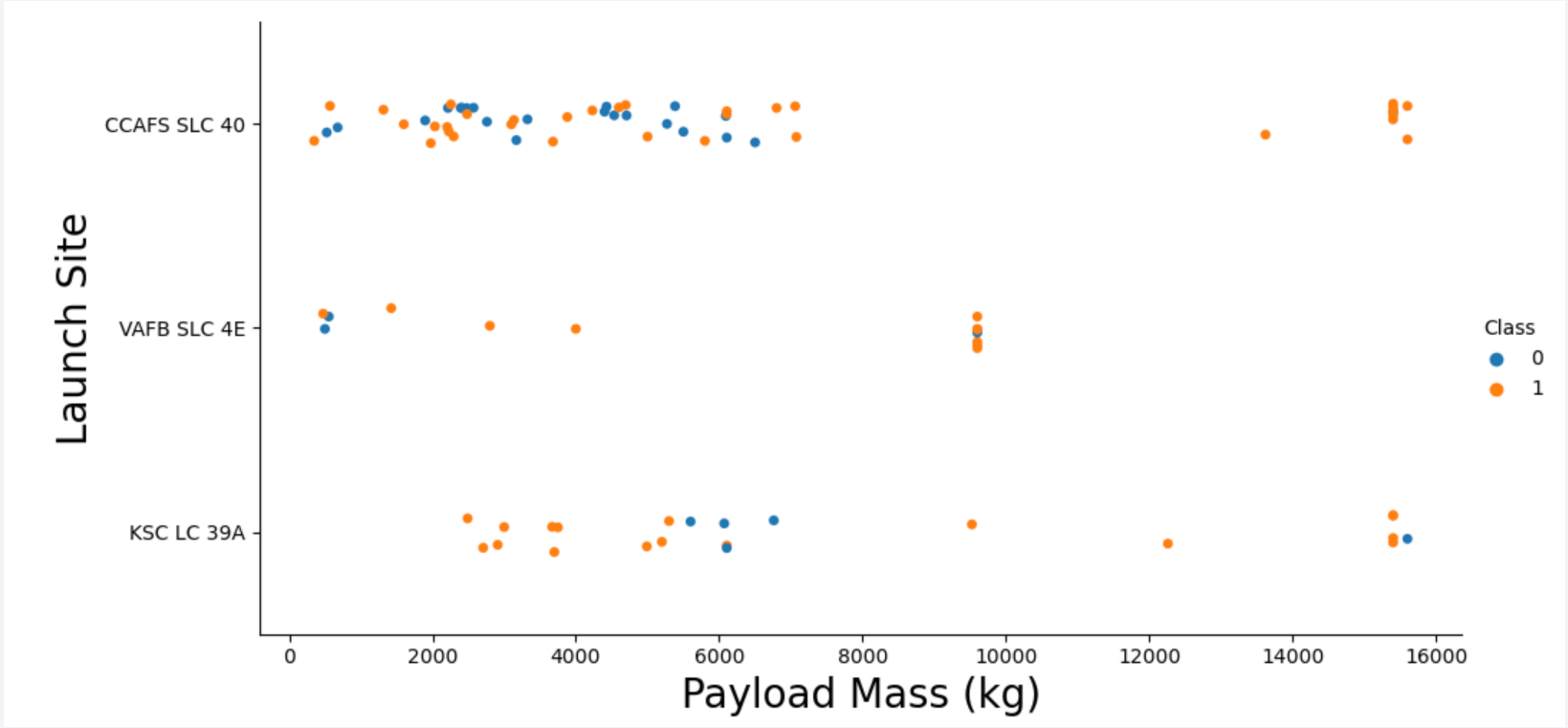

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



From this plot, we can see that CCAFS SLC 4 has the highest number of launches given high flight numbers, and the largest number of successful launches given the largest concentration of orange dots which indicate a successful launch. Therefore, we can conclude that CCAFS SLC 4 has the highest success rate. We can also conclude that all launch sites trend towards increasing success rate with more flight numbers.

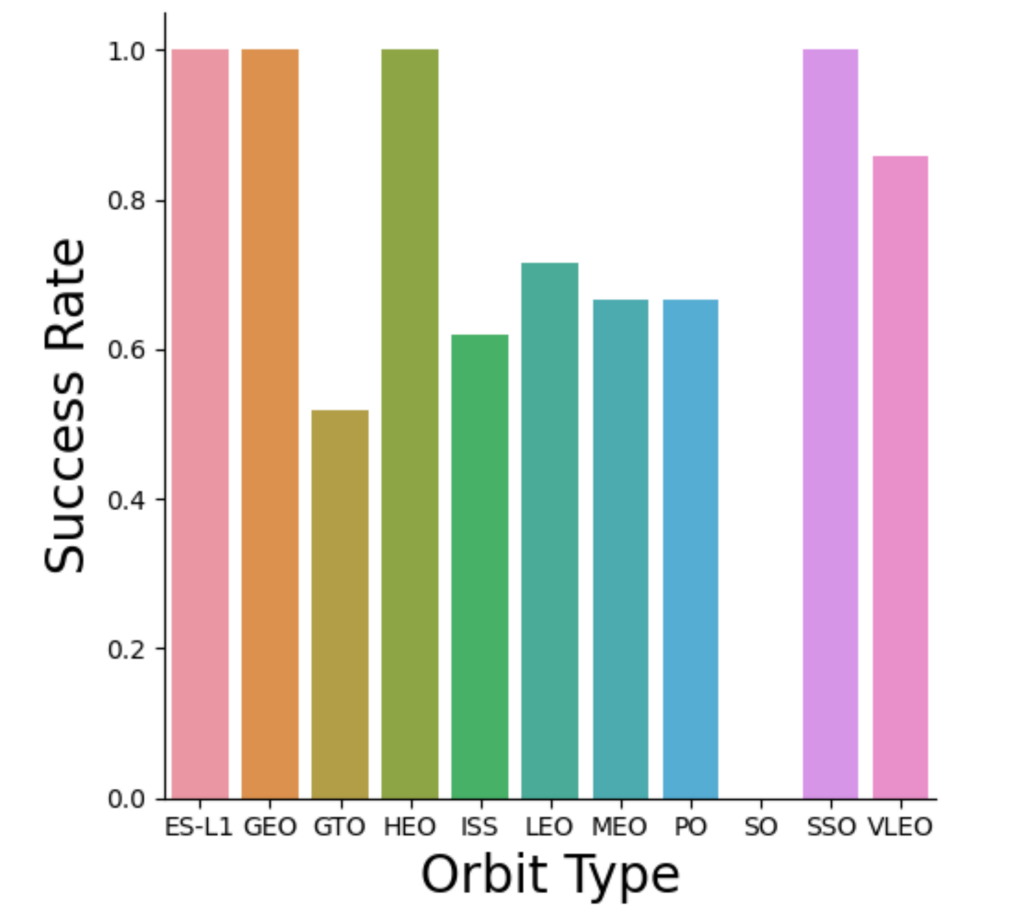
Payload vs. Launch Site



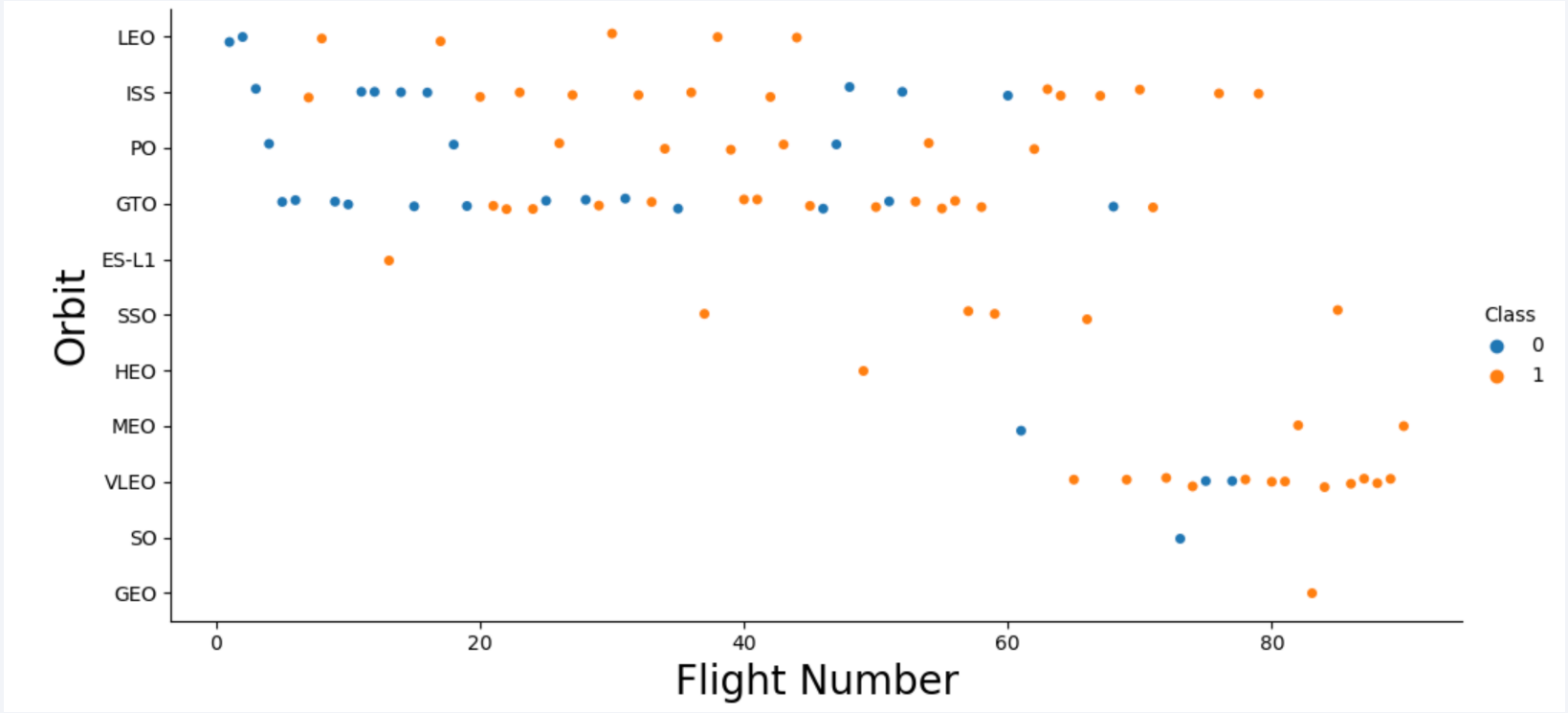
From the graph, we can see that the lower payload mass rockets have a higher success rate. Furthermore, we can see that CCAFS SLC 40 has the highest volume of launches and the highest success rate.

Success Rate vs. Orbit Type

The bar graph indicates that the orbit type ES-L1, GEO, HEO, and SSO have the highest success rate of 1.0 (100%). The orbit SO has the lowest success rate of 0.0 (0%).

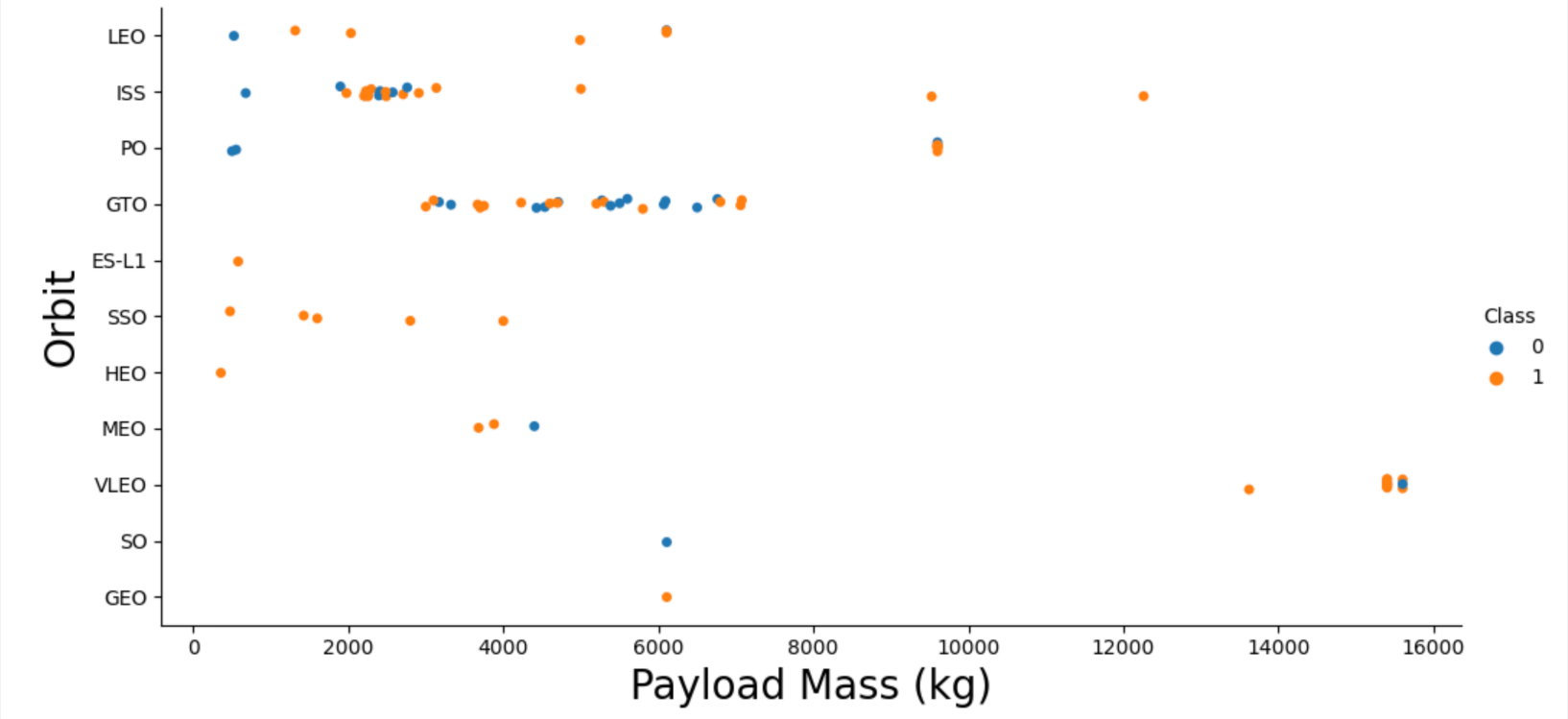


Flight Number vs. Orbit Type



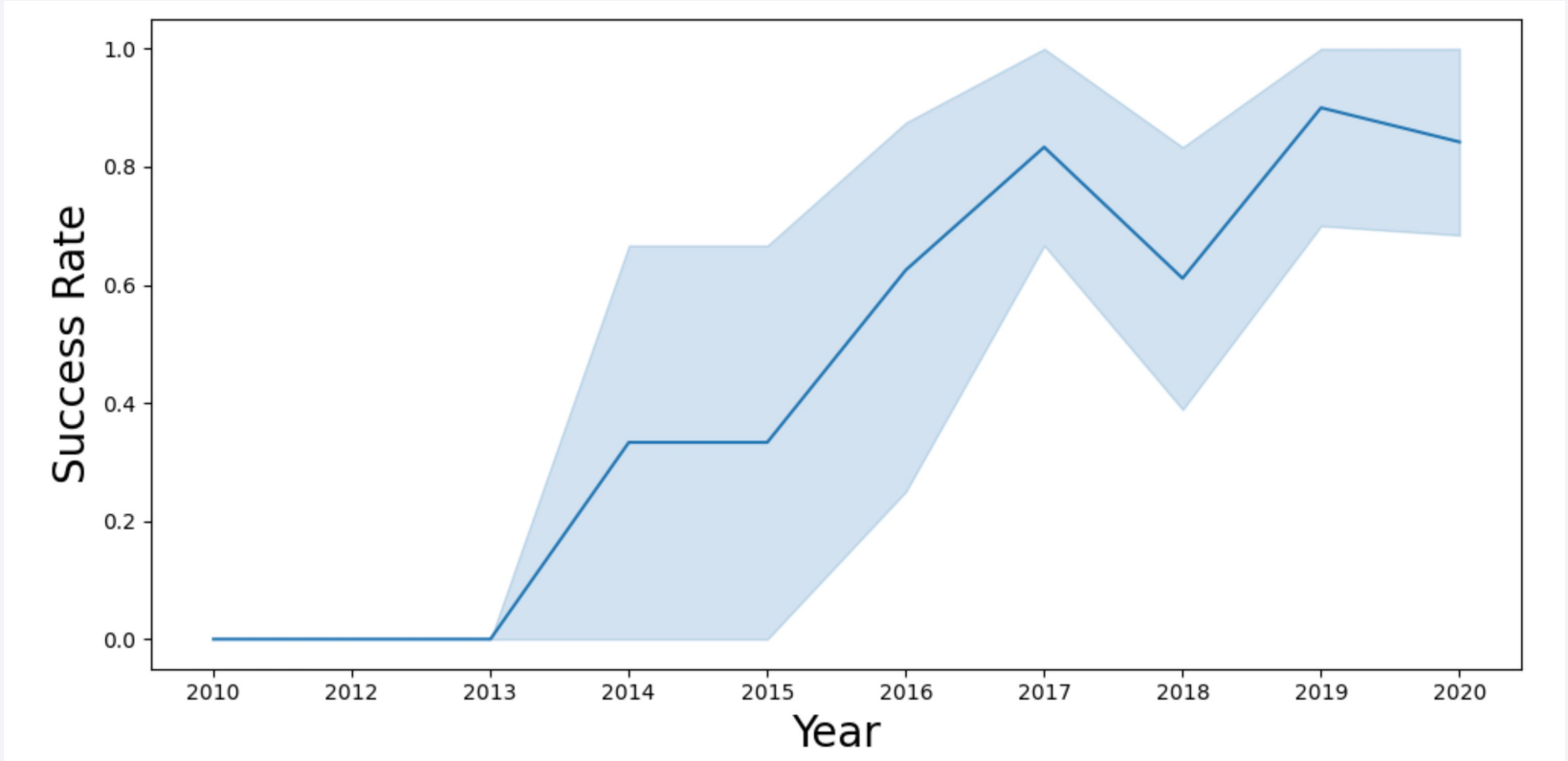
From this graph, it's clear that the orbits with the highest success rate have barely any flights, so the previous graph could have been misleading. Furthermore, visually speaking, GTO has the best balance of high number of flight numbers as well as a higher success rate.

Payload vs. Orbit Type



For all orbits lower payload mass look to positively correlate with a successful launch. The ideal payload mass for the best success rate looks to be between 2000 – 6000 kg.

Launch Success Yearly Trend



The graph indicates that as the years have gone on, the success rate of the launches have increased.

All Launch Site Names

- Using the command

```
%sql select distinct launch_site from SPACEXTBL;
```

  The following unique names for launch sites were obtained:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'KSC'

- Using the SQL command given below, we obtained launch sites names that begin with ‘KSC’

```
%sql select * from SPACEXTBL where launch_site like 'KSC%' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Land
19/02/2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490.0	LEO (ISS)	NASA (CRS)	Success	Su
16/03/2017	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600.0	GTO	EchoStar	Success	
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300.0	GTO	SES	Success	S
05/01/2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300.0	LEO	NRO	Success	Su
15/05/2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070.0	GTO	Inmarsat	Success	

Total Payload Mass

- Using the SQL command given below, the total payload mass carried by boosters launched by NASA was calculated.

```
%sql select sum(PAYLOAD_MASS__KG_) \
    from SPACEXTBL \
    where CUSTOMER = "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- Using the SQL command given below, the average payload mass carried by booster version F9 v1.1 was calculated.

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL \
    where Booster_Version like "%F9 v1.1%"
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Using the SQL command given below, the date of the first successful landing outcome on drone ship was found.

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = "Success (ground pad)"
```

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

min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Using the SQL command given below, the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were obtained. They are:

```
%sql select Payload from SPACEXTBL where Landing_Outcome = \
"Success (ground pad)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Payload
NROL-76
Boeing X-37B OTV-5
Zuma

Total Number of Successful and Failure Mission Outcomes

Using the SQL command given below, the total number of successful and failed mission outcomes were given.

```
%sql select Mission_Outcome, count(*) as Total_Number \
from SPACEXTBL group by Mission_Outcome order by Mission_Outcome
```

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

Mission_Outcome	Total_Number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using the SQL command given below, the names of the booster which have carried the maximum payload mass are listed.

```
%sql select Booster_Version from SPACEXTBL where \
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * 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
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Using the SQL command given below, the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017 are listed.

```
%sql select substr(Date,4,2) as Month, Date, Booster_Version,\
    Launch_Site, Landing_Outcome from SPACEXTBL \
    where Landing_Outcome = 'Success (ground pad)' \
    and substr(Date,7,4) = '2017';
```

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

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
02	19/02/2017	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
01	05/01/2017	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
03	06/03/2017	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
08	14/08/2017	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
07	09/07/2017	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
12	15/12/2017	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

- Using the SQL command given below, the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 are ranked in descending order

```
%sql select Landing_Outcome, count(*) as total_number from SPACEXTBL \
    where Date between '04-06-2010' and '20-03-2017' \
    group by Landing_Outcome order by total_number DESC;
```

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

Landing_Outcome	total_number
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

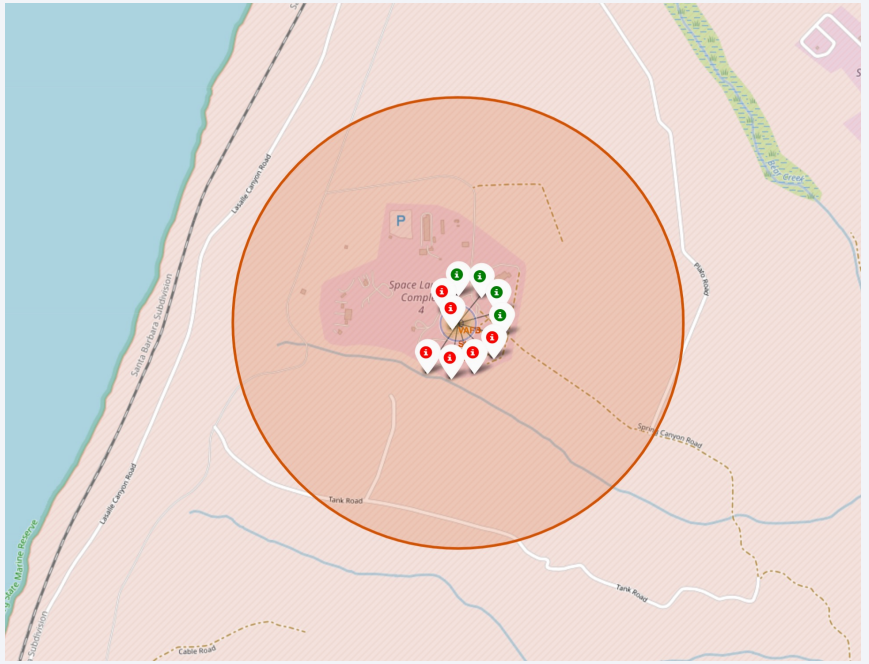

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

From the graph to the right, we can see that the launch sites are in USA, more specifically in Los Angeles and California. They are also by the coastline.

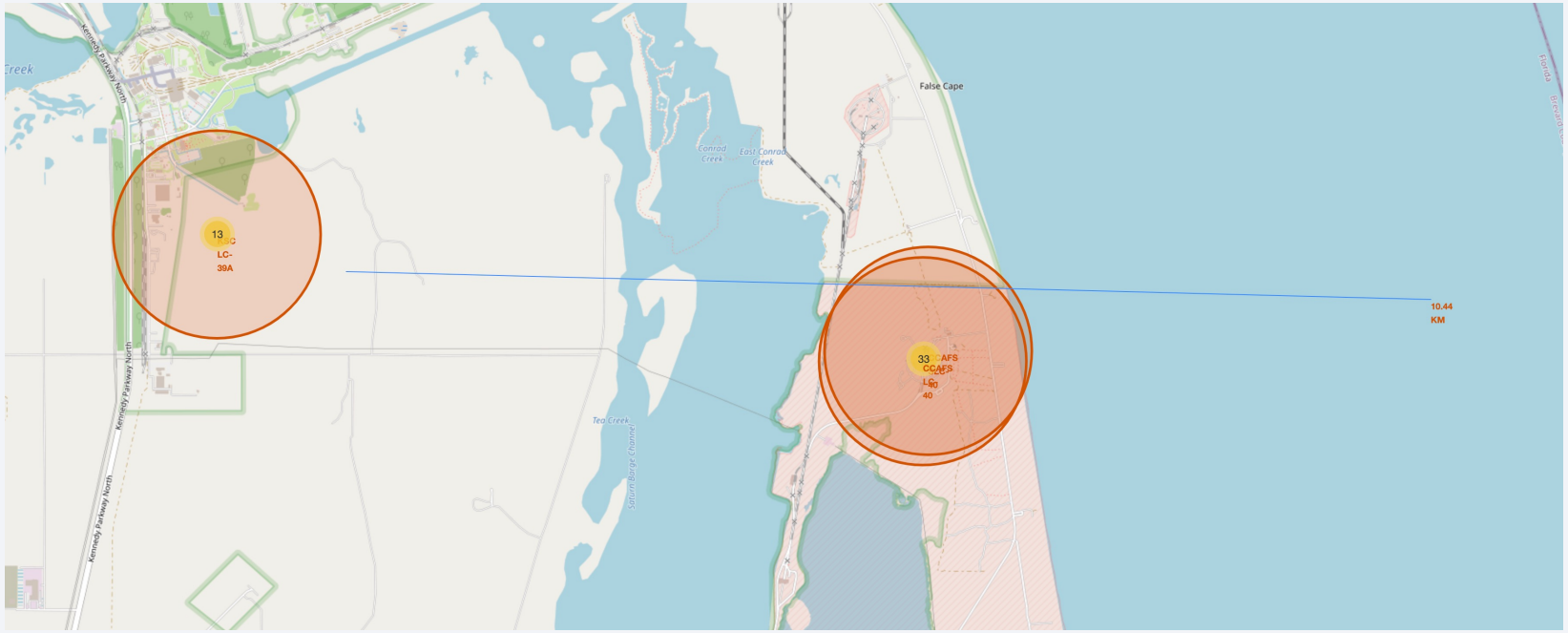
Color labelled launch sites markers

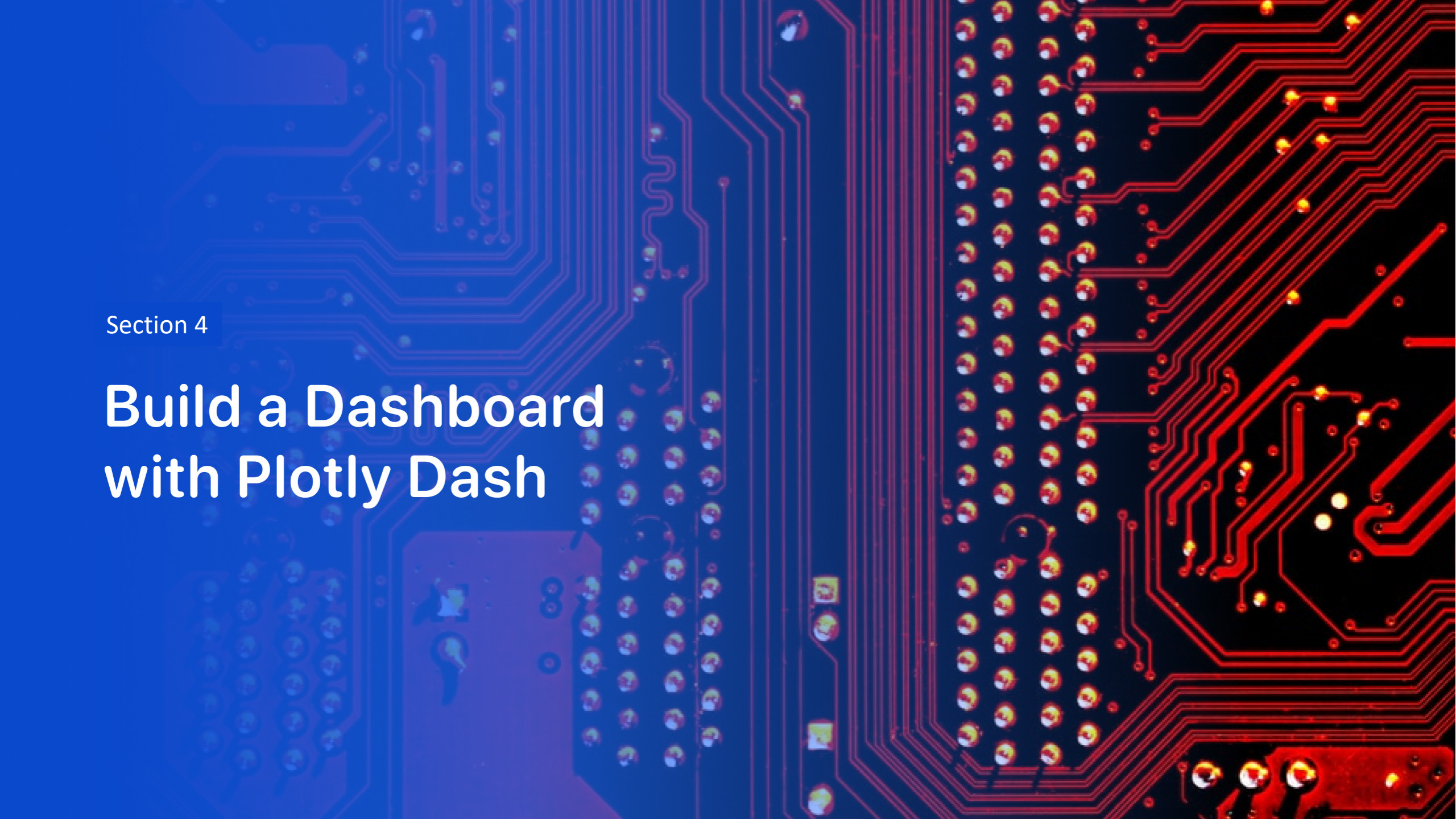
To the right, we have one example of cluster markers for Florida launch site. In folium, we can zoom in and click on clusters to show all the successful and failed landings in that location. The green labels indicate a successful launch, and the red labels indicate an unsuccessful launch.



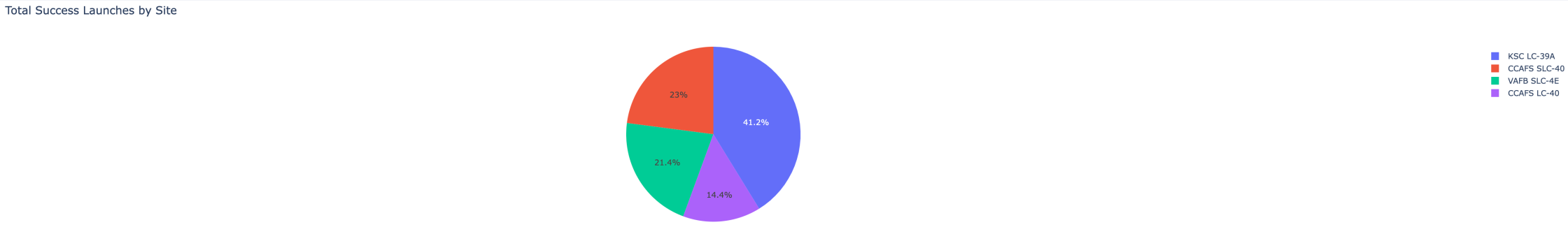
Analyzing Distance Of Launch Site From Important Infrastructure

While planning logistics of the location of a launch site, it’s important to make sure it’s close to important infrastructure like railways and highways, and close to the coastline for safety, as well as distant enough from residential areas.



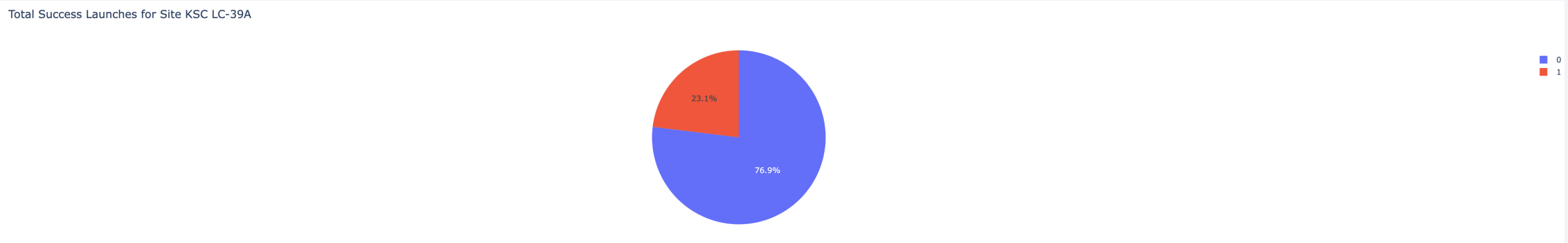

Section 4
Build a Dashboard with Plotly Dash

Distribution of success rate of each launch site



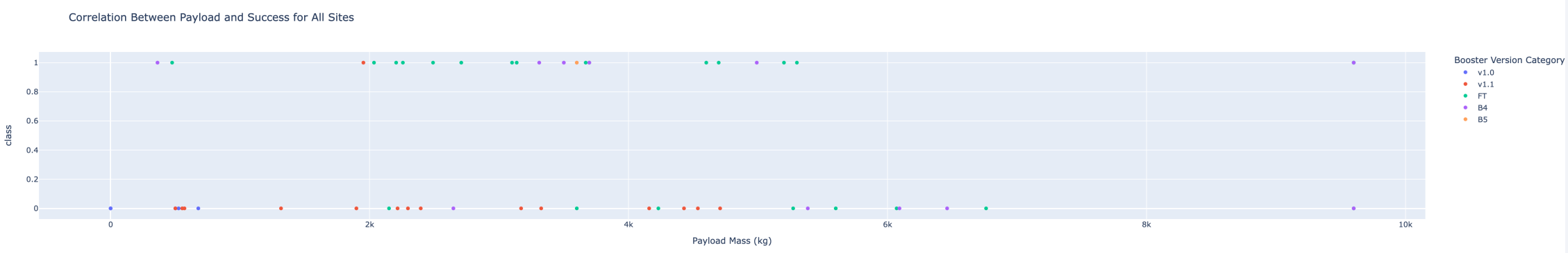
The pie chart above shows the proportion of success rates of each launch site. From the pie chart, we can tell that KSC LC-39A has the highest proportion of success launches with 41.2% of successful launches coming from this launch site.

Success Rate Of KSC LC-39A



KSC LC-39A, which has the highest proportion of successful launches has a 76.9% success rate and a 23.1% failure rate.

Payload Mass vs Launch Outcome



From the graph above, we can deduce that payloads between 2000– 6000 kg have the best chances of having a successful launch.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

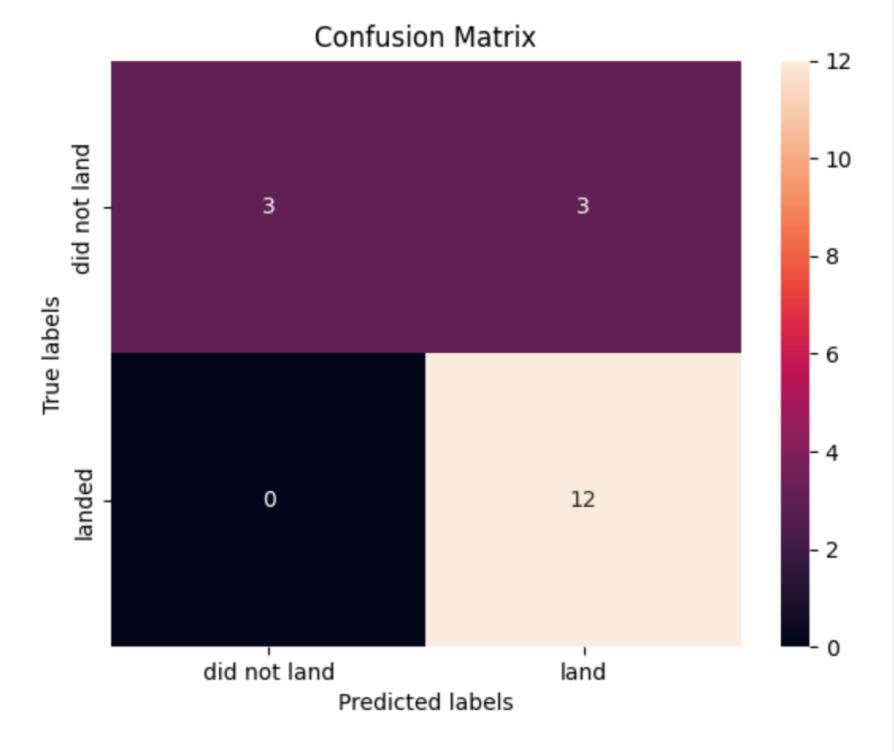
```
models = {'k_neighbours':knn_cv.best_score_,
              'decision_tree':tree_cv.best_score_,
              'logistic_regression':logreg_cv.best_score_,
              'support_vector': svm_cv.best_score_}

best_alg = max(models, key=models.get)
print('Best model is', best_alg,'with a score of', models[best_alg])
if best_alg == 'decision_tree':
    print('Best params is :', tree_cv.best_params_)
if best_alg == 'k_neighbours':
    print('Best params is :', knn_cv.best_params_)
if best_alg == 'logistic_regression':
    print('Best params is :', logreg_cv.best_params_)
if best_alg == 'support_vector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is decision_tree with a score of 0.875
Best params is : {'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}
```

- Running the code above, we have deduced that the Decision Tree method has the highest accuracy out of the 4 models examined.

Confusion Matrix

- An interesting thing observed was that despite the decision tree method having the highest classification accuracy, all methods had the same confusion matrix.



Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

Thank you!