Winning Space Race with Data Science

Nkemakolam Chinenye Onyemachi

15– Jul -2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

Space X Falcon 9 launches are advertised on its website on a cost of 65million dollars. This is far more cheaper than other providers. This is due to the fact that Space X can reuse its first stage. The aim of this project is to determine if the first stage will land successfully using machine learning algorithms. This is to enable us to use the information for an alternate company “Space Y” That wants to bid against space X in the future.

- Problems you want to find answers
  - The factors that will determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - Operating conditions that needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Space X API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- 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

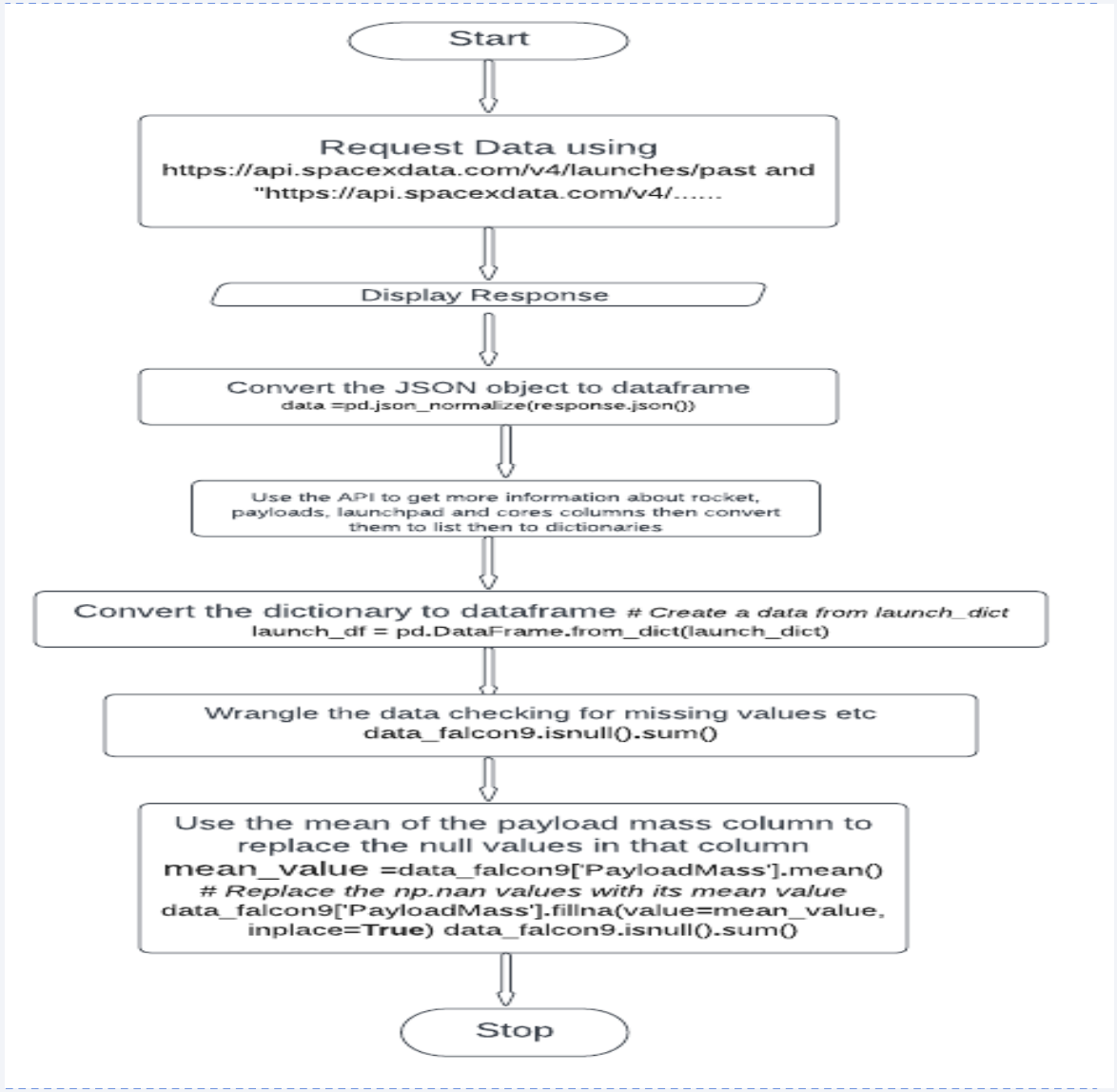
- Data collection was done using get request to the SpaceX API.
- Next, I decoded the response content as a JSON object using .json() function call and converted into a pandas dataframe using .json_normalize().
- I then cleaned the data, checked for missing values and fill in missing values where necessary.
- I also performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for further analysis.

Data Collection – SpaceX API

- The ‘Get’ request was used with the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

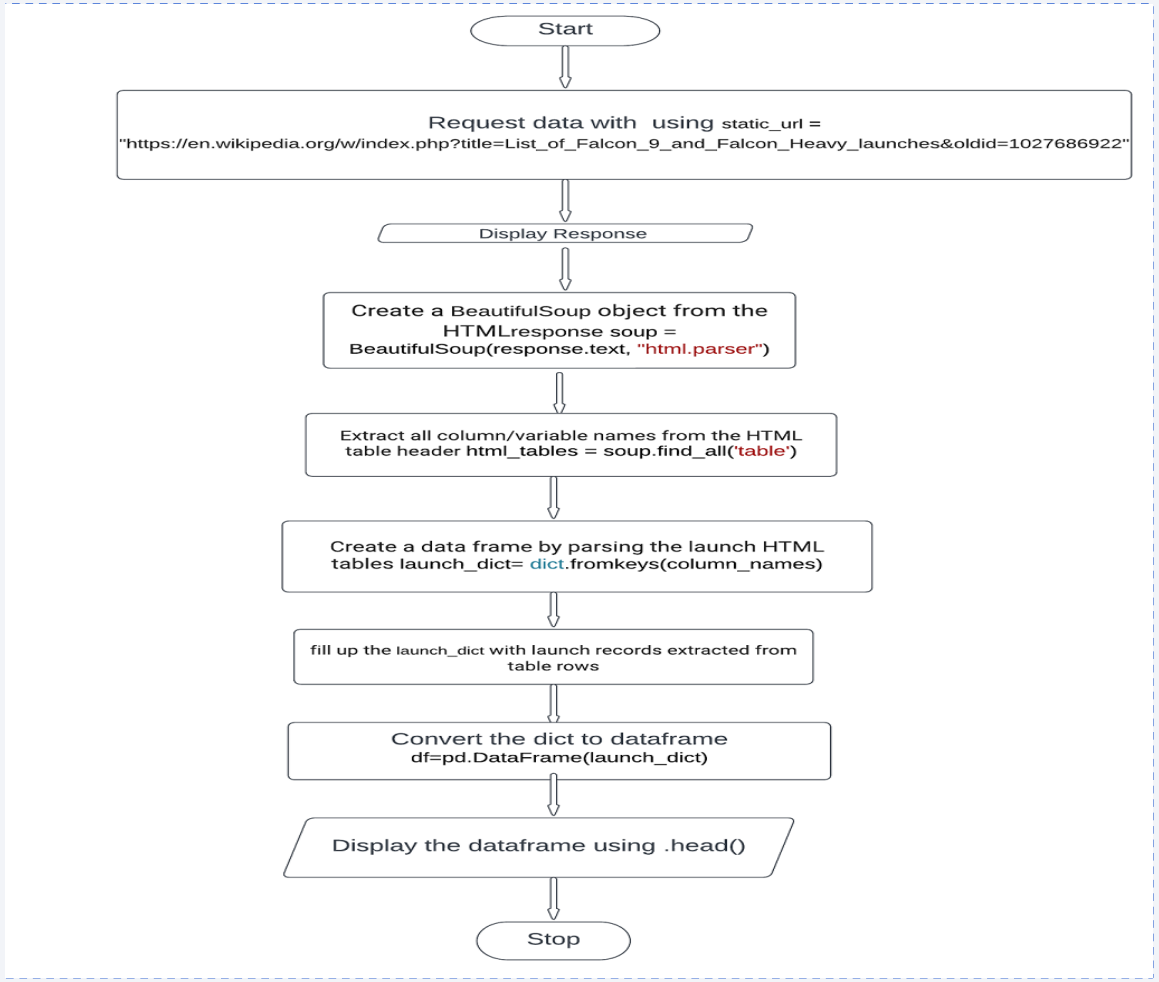
- The link to the notebook is

https://github.com/NkemDev/Applied-Capstone-Project/blob/main/jupyter_labs_spacex_data_collection_api.ipynb



Data Collection - Scraping

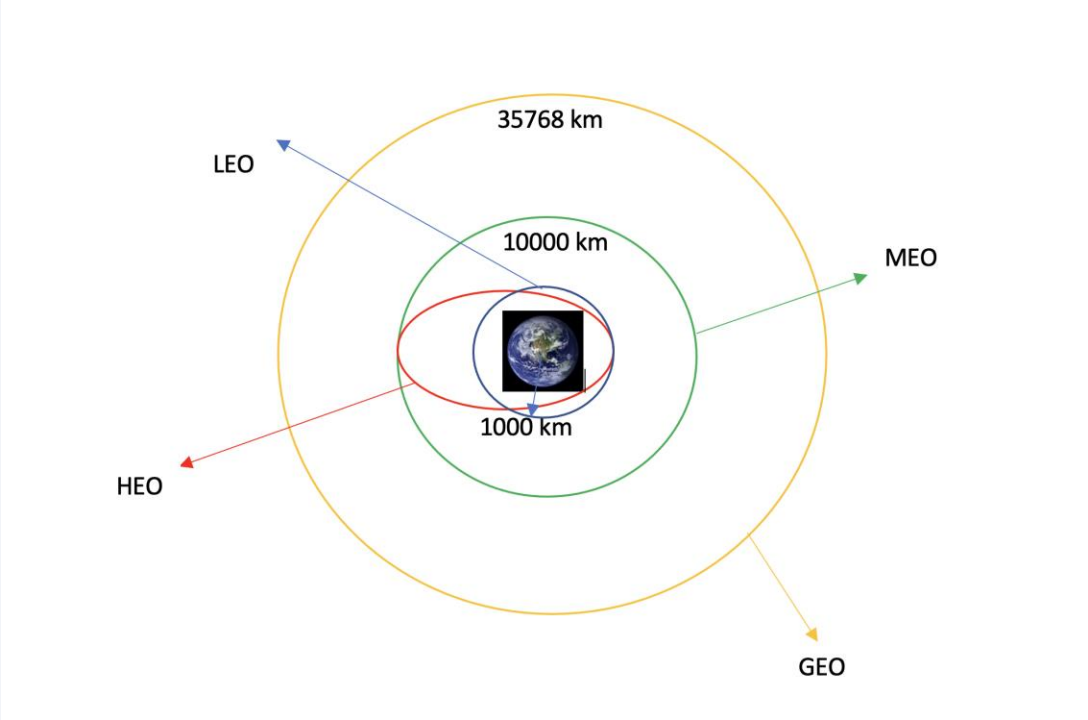
- We applied web scrapping to web scrape Falcon 9 launch records with BeautifulSoup. I parsed the table and converted it into a pandas dataframe.
- The link to the notebook that contains the web scraping with BeautifulSoup is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/jupyter_labs_webscraping.ipynb



Data Wrangling

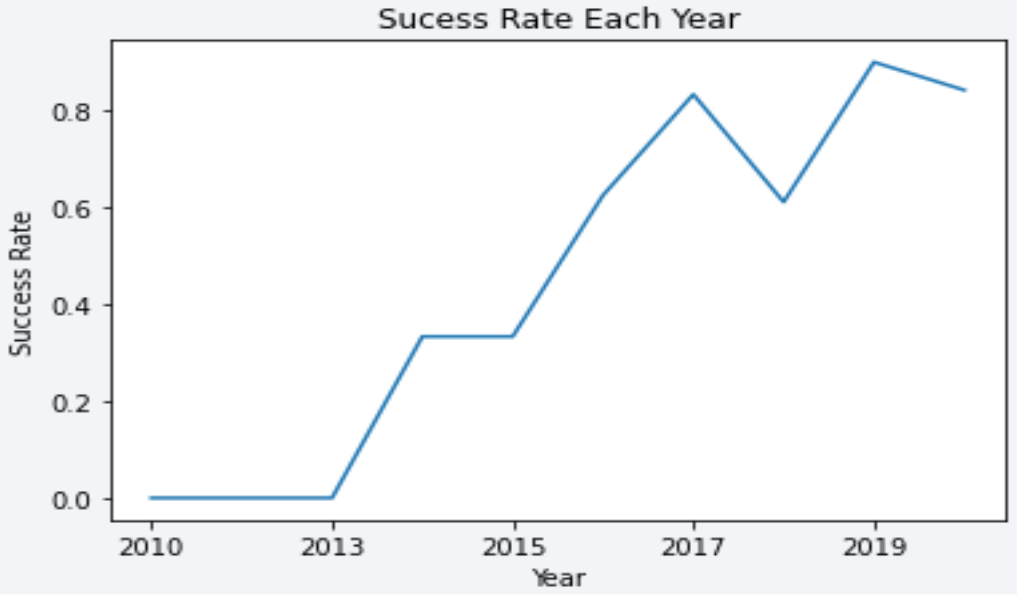
- Exploratory data analysis was performed and the training labels were determined
- The number of launches were calculated at each site and the number and occurrence of each orbits
- The landing outcome label was created from the outcome column the results exported to a csv file.
- The link of the notebook is

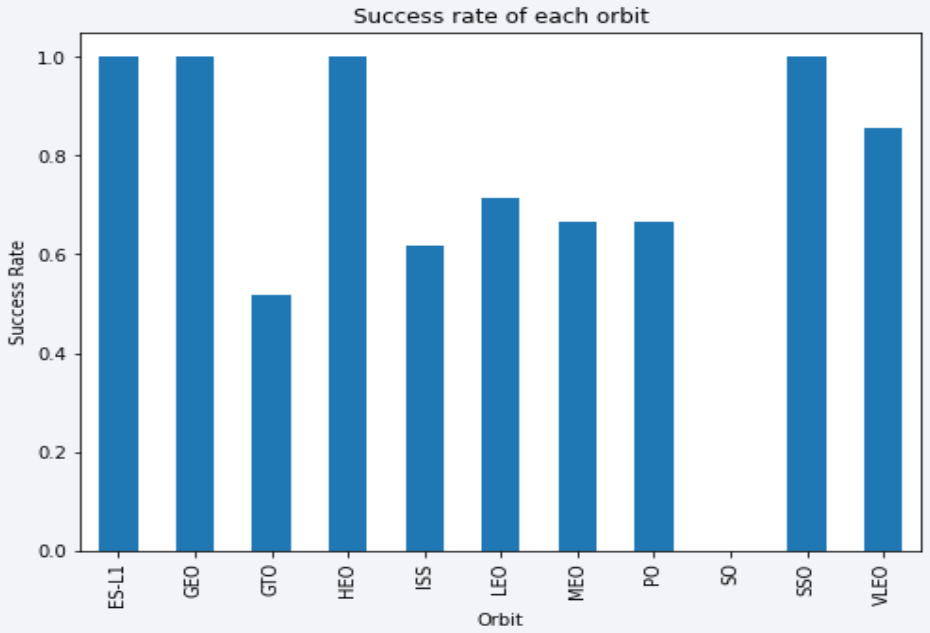
https://github.com/NkemDev/Applied-Capstone-Project/blob/main/spacex_Data_wrangling.ipynb



EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.





The link to the notebook in github is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/Eda_dataviz.ipynb

EDA with SQL

- The dataset was loaded into an sqlite database. I then applied EDA to gain insights from the data using SQL queries.
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- The link of the notebook on github is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/eda_sql_coursera_sqllite.ipynb

Build an Interactive Map with Folium

- I marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- I assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, I identified which launch sites have relatively high success rate.
- I calculated the distances between a launch site to its proximities. We answered some question for instance
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.
- The link to the notebook on GitHub is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

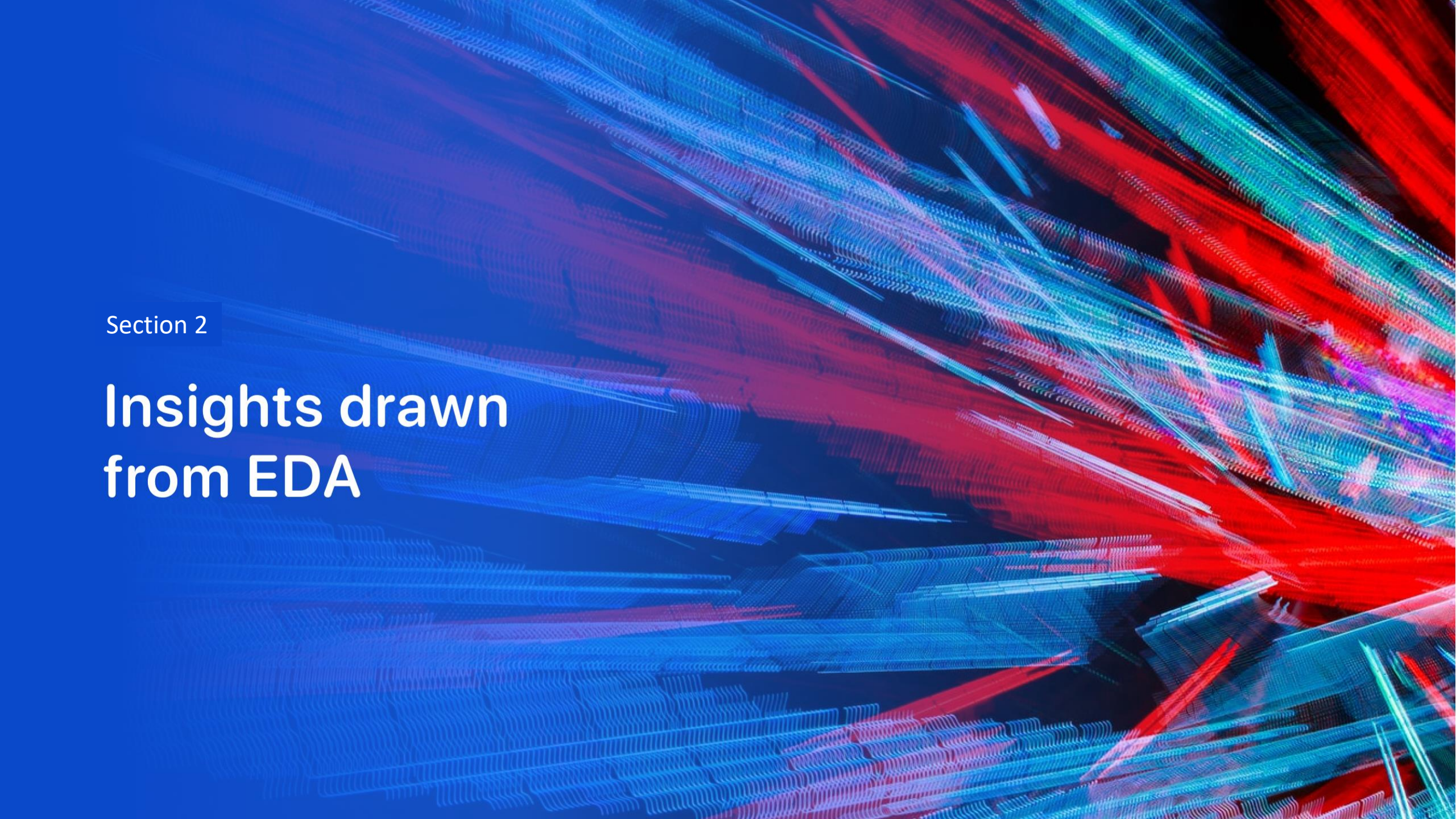
- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The link to the notebook on GitHub is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/plotydash.py

Predictive Analysis (Classification)

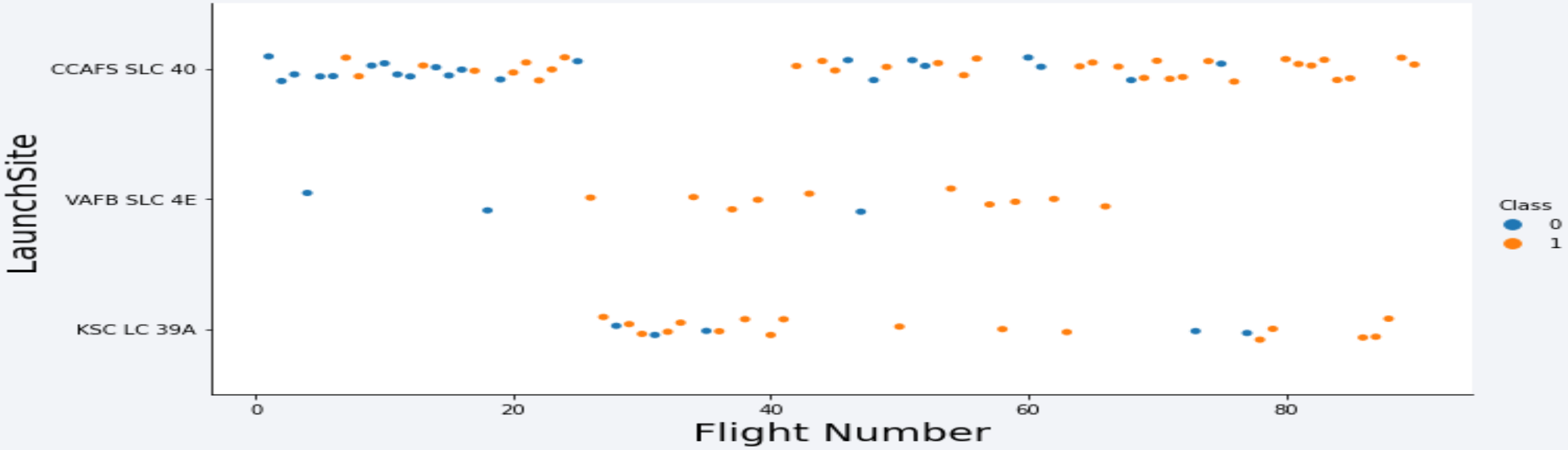
- Data was loaded and transformed using pandas and numpy, the data was then split into training and testing.
- Machine learning models were then built and tune different hyper parameters were tuned using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- The link to the Github notebook is https://github.com/NkemDev/Applied-Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

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

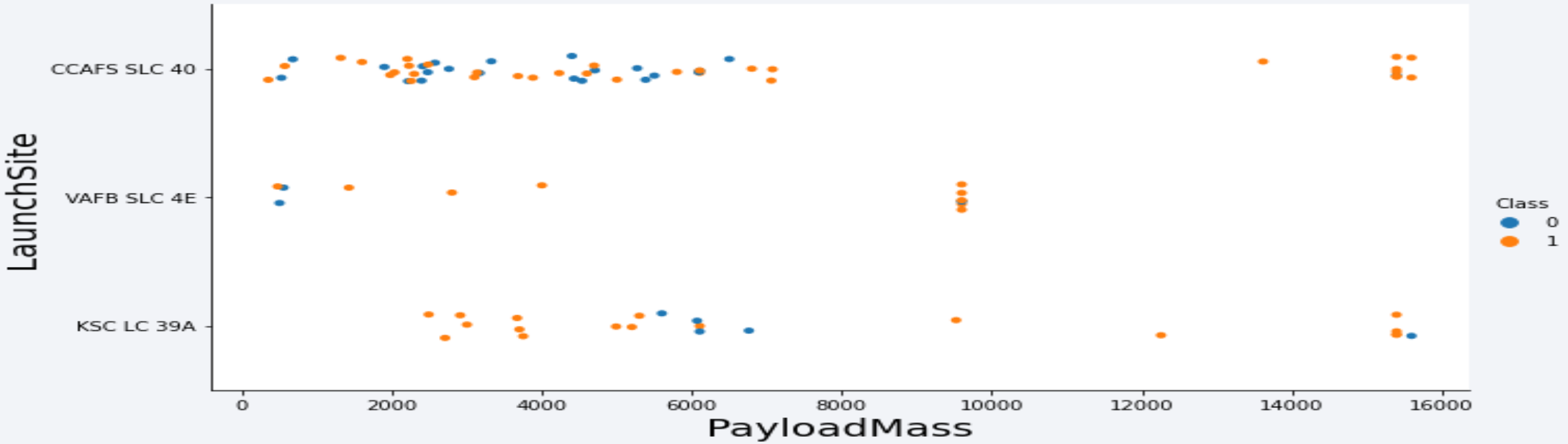

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



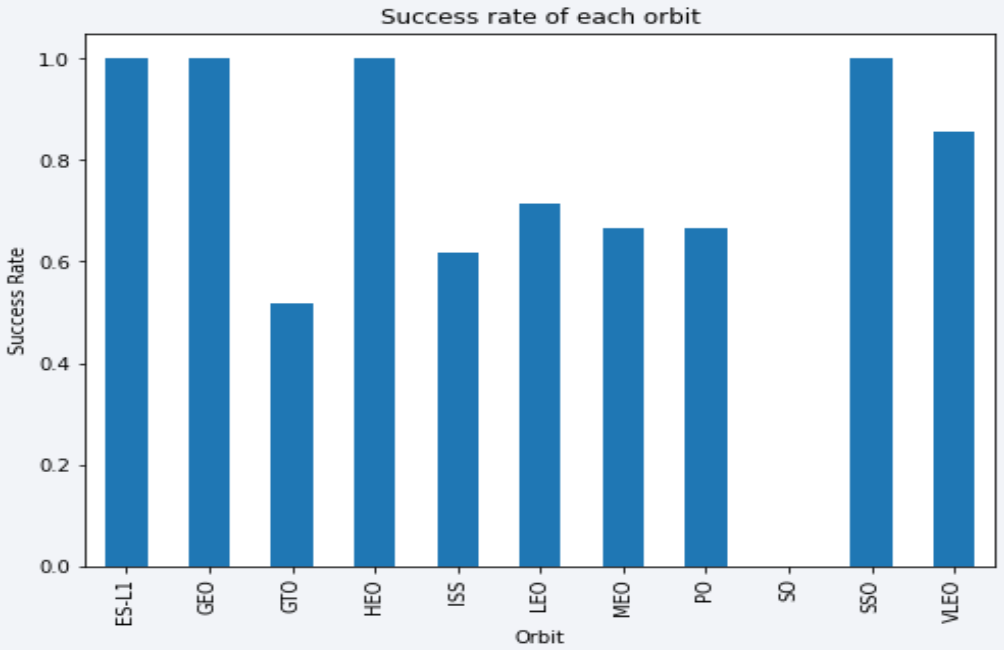
There are more launches with CCAFS SLC 40. The chance of success launch is 50%, VAFB SLC 4E has a 77% success rate in launching. KSCLC 39 also has a 77% success rate. Also for VAFB SLC 4E, There are no flights after number 60.

Payload vs. Launch Site



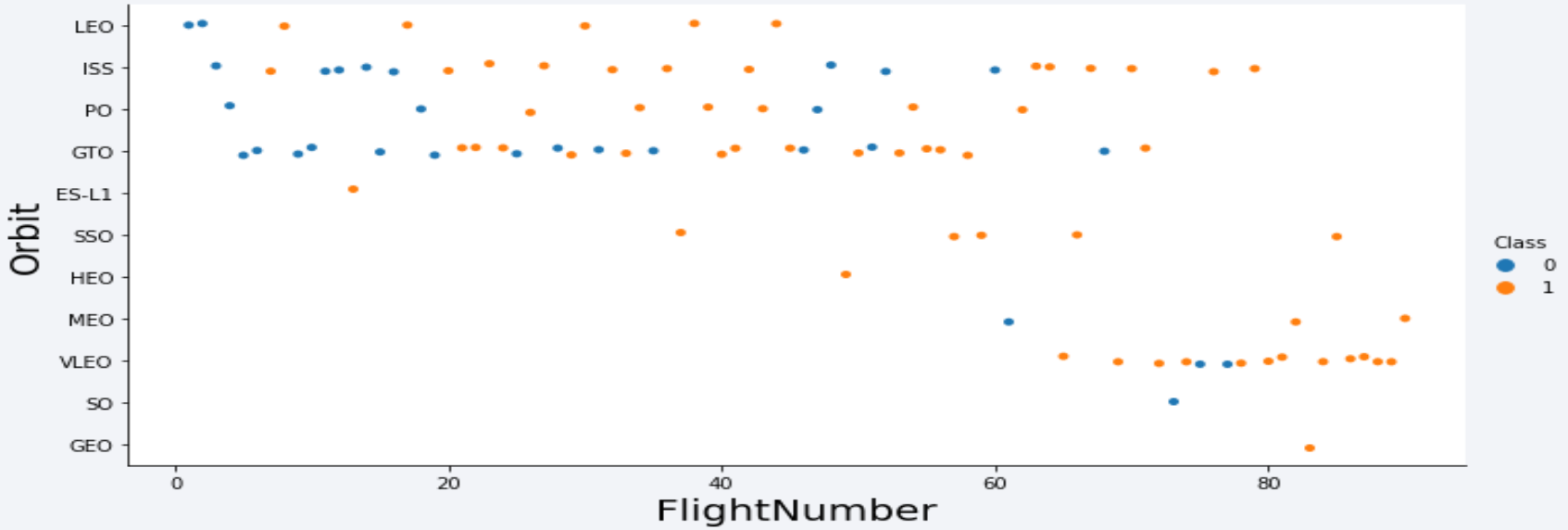
Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



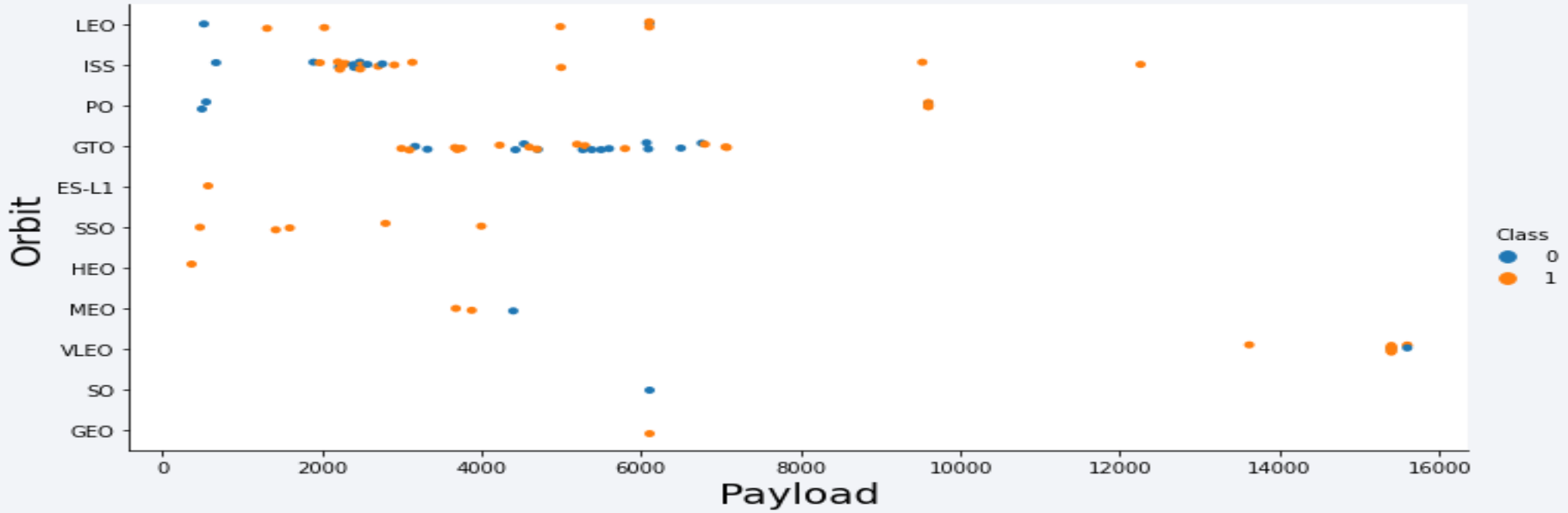
ES-L1, HEO, SSO have the same highest rates while GTO have the lowest success rate.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the 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 Polar, 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

The success rate since 2013 kept increasing till 2020

All Launch Site Names

DISTINCT was used to show launch sites from SPACETBL

```
%%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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%%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	Orbit	Customer	Mission_Outcome	Landing _Outcome
4/6/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
8/12/2010	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1/3/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

We used the query above to display five launch sites that begin with the string ‘CCA’

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[ ] %%sql
    SELECT SUM(Payload_Mass) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
 SUM(Payload_Mass)
 45596
```

The total payload mass was determined using the query above

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG(Payload_Mass)FROM SPACEXTBL WHERE Booster_Version ='F9 v1.1'
```

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

AVG(Payload_Mass)
2928.4

The query above was used to determine the average payload mass where booster version =F9 v1.1

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[16] %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome   LIKE 'Success%ground%pad%'
```

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

MIN(Date)
1/5/2017

The query above was used to determine the first successful landing outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[17] %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%drone%ship%' AND Payload_Mass BETWEEN 4000 AND 5999
```

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

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

The query above was used to determine the names of boosters which have success in drone ship and have payload mass between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[18] %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTBL;
```

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

COUNT(Mission_Outcome)
101

The query above was used to determine the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[19] %sql SELECT Booster_Version FROM SPACEXTBL WHERE Payload_Mass=(SELECT MAX(Payload_Mass) 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

The query to select the booster version which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2),Mission_Outcome,Booster_Version,Launch_Site FROM SPACEXTBL where substr(Date,7,4)='2015';
```

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

substr(Date, 4, 2)	Mission_Outcome	Booster_Version	Launch_Site
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

The query above was used to display the date launch records in 2015

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[ ]  %sql SELECT Landing_Outcome FROM SPACEXTBL WHERE Date BETWEEN '04-06-10' and '20-03-2017' ORDER BY Date DESC
```

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

Landing_Outcome
No attempt
Success
No attempt
Success (ground pad)
No attempt
Success
Success
Success (ground pad)
Success (drone ship)
Controlled (ocean)
Failure
Success
Failure
Failure (drone ship)

The query above was used to display successful landing outcomes between date 04-06-2010 and 20-03-2017
Using the **WHERE** clause where date between 04-06-2010 and 20-03-2017 **ORDER** BY date

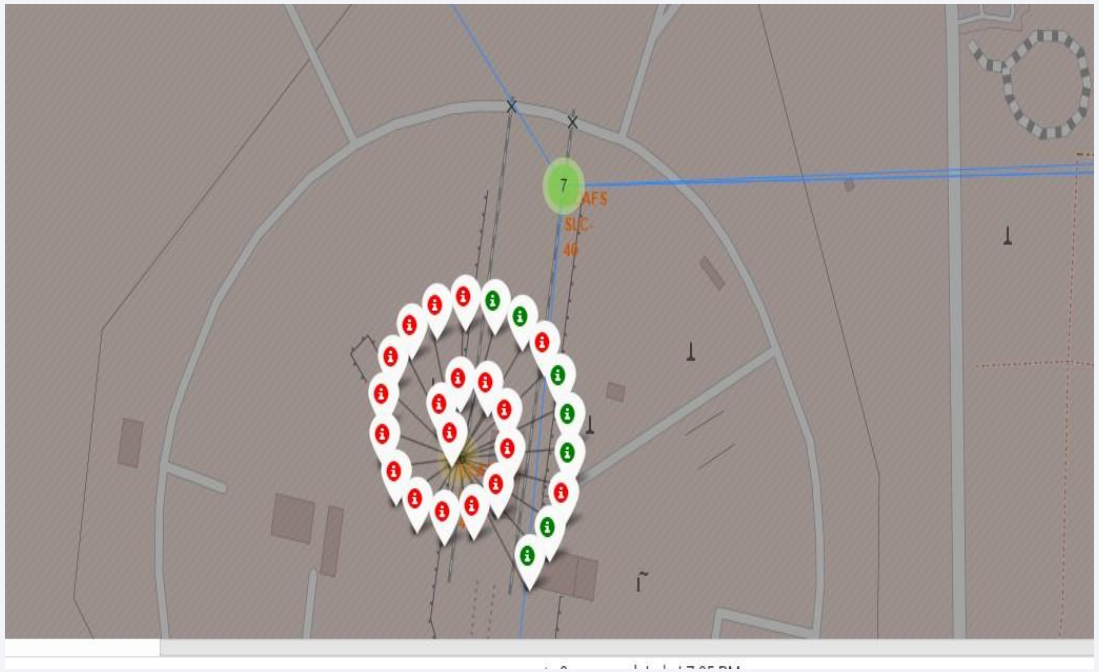
Section 3

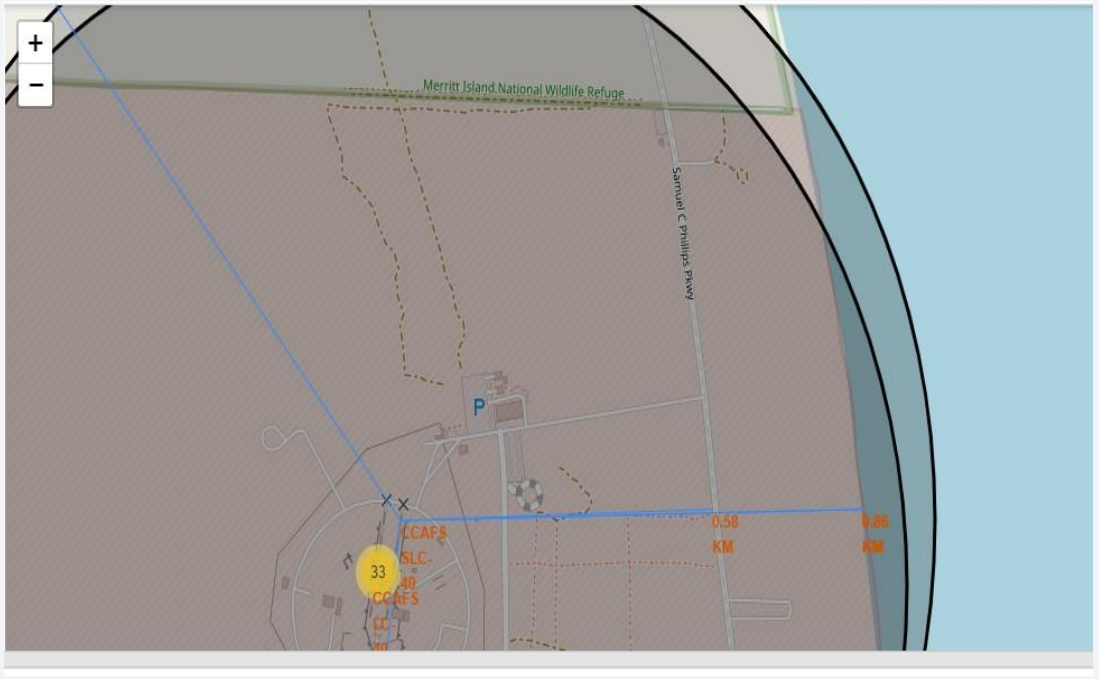
Launch Sites Proximities Analysis

All Launch sites global markers

The launch sites were located in the western and eastern coast of the United states.

Key locations on the location map

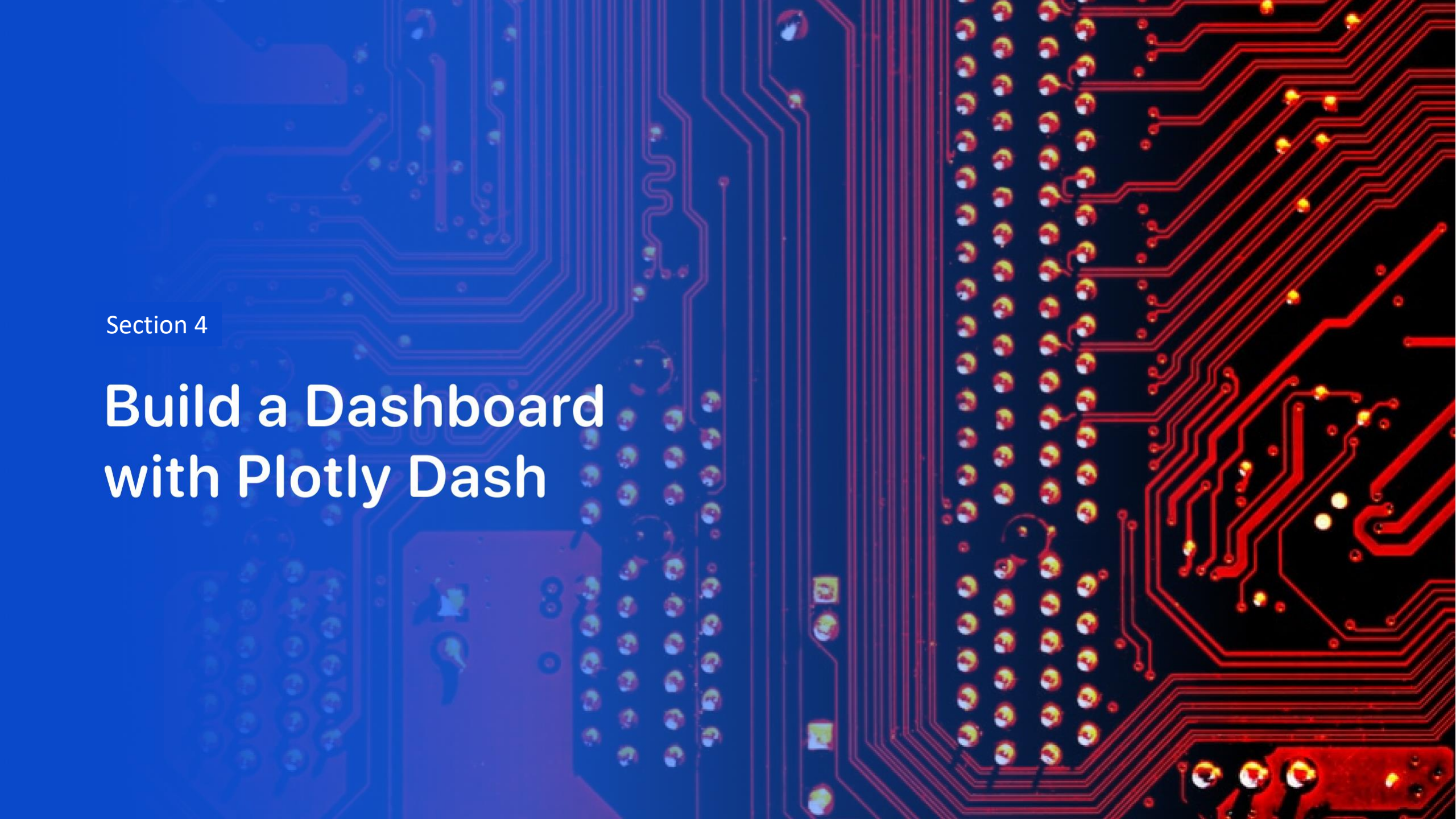




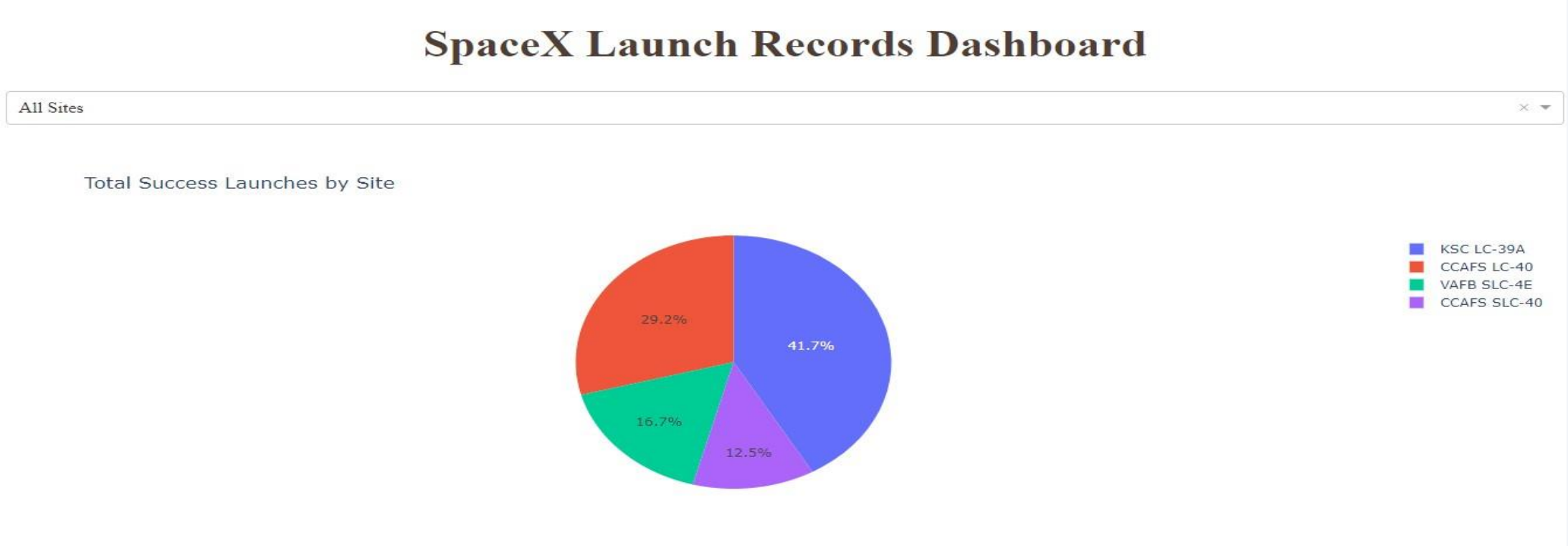
distance_highway = 0.5834695366934144 km
distance_railroad = 1.2845344718142522 km
distance_city = 51.43416999517233 km
The distance from the launchsite to the sea is 0.58km

Color-Coded Launch Markers

Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

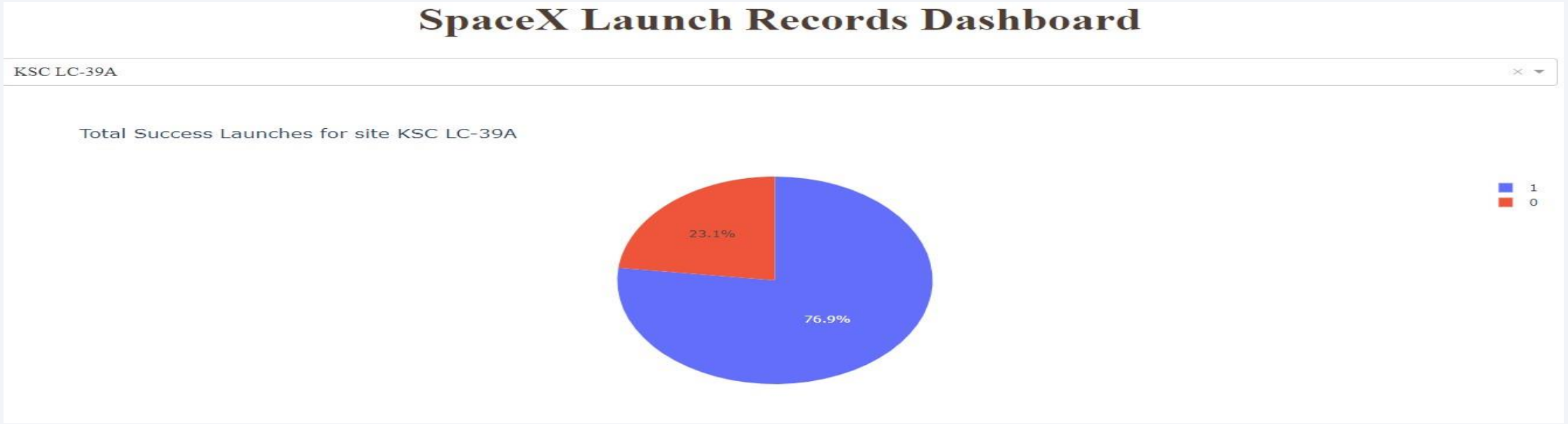

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



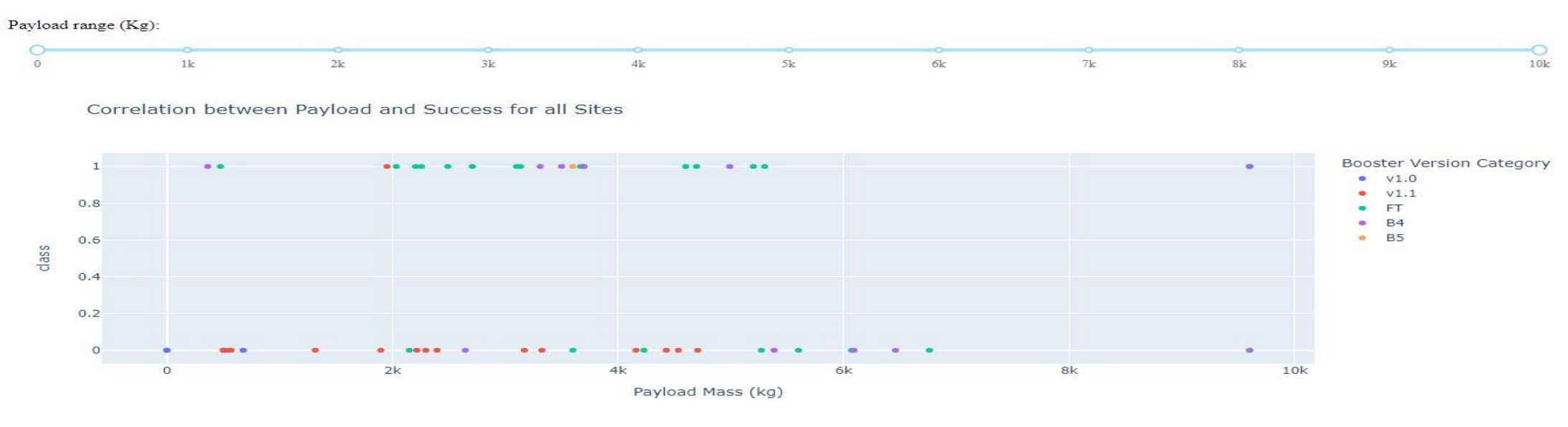
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings,
but a majority of the successful landings where performed before the name change.
VAFB has the smallest share of successful landings.

Highest Success Rate Launch Site



KSC LC-39A has the highest success rate in launches.

Payload Mass vs. Success vs. Booster Version Category



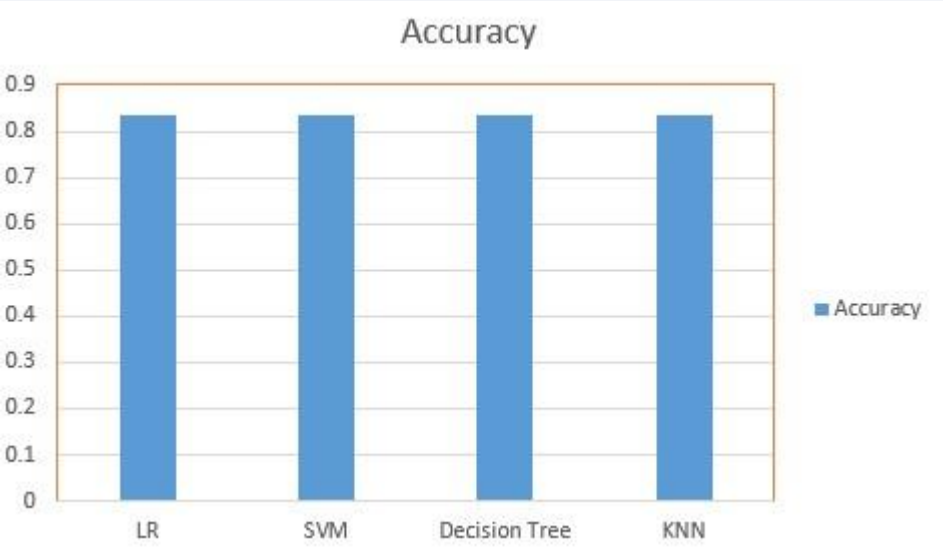
Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 10k. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Classification on test data



For the test data, the accuracy is the same for all the ML models .
In the training data, Decision tree had the highest Accuracy score.

Confusion Matrix

Since all models performed the same for the test set, the confusion matrix is the same across all models.
The models predicted 12 successful landings when the true label was successful landing.
The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
Our models over predict successful landings.

Conclusions

- The task was to develop a machine learning model for Space Y who wants to bid against SpaceX.
- The goal of model is to predict when Stage 1 will successfully land to save about $100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- I Created data labels and stored data into an sqlite database
- I Created a dashboard for visualization
- I created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Thank you!