IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Basel Al Shargabi
Date7/22/2024

Outline

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

Executive Summary

Introduction

This project aims to develop a machine learning pipeline to predict the success of SpaceX's Falcon 9 rocket landings. By achieving a high success rate, the project seeks to enable Space Y to offer competitive services and reduce launch costs significantly.

Methodology

The data collection involved two primary methods:

1.API Requests: Data was collected from SpaceX's API.

2.Web Scraping: Additional data was obtained from Wikipedia using BeautifulSoup.

The data underwent extensive wrangling, including handling missing values and applying one-hot encoding to categorical features. Exploratory Data Analysis (EDA) was performed using visualization techniques and SQL queries to uncover insights. Interactive visual analytics were created using Folium and Plotly Dash. Finally, predictive models were developed using various classification algorithms, and hyperparameters were tuned using GridSearchCV.

Executive Summary

Results:

Key findings from the analysis include:

1. **Launch Site Success Rates:** KSC LC-39A had the highest success rate among all launch sites.
2. **Payload and Orbit Analysis:** Heavy payloads tend to have higher success rates for specific orbits like PolarLEO and ISS, while GTO orbits showed mixed results.
3. **Yearly Trends:** The success rate of launches has steadily increased since 2013, reaching its peak in 2020.

The decision tree classifier emerged as the best-performing model for predicting landing success, with a high accuracy score. However, the model's confusion matrix revealed some issues with false positives, indicating areas for further improvement

Executive Summary

Conclusion

The project successfully developed a predictive model that can help Space Y optimize their launch processes and reduce costs. The insights gained from the EDA and the interactive visualizations provide valuable information for decision-making and strategic planning. The continuous improvement of the model and further analysis of additional features will enhance the accuracy and reliability of the predictions.

Introduction

- Project background and context
  - Space X 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 Space X can reuse the first stage. Therefore a Space Y company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully that will allow it to minimize the cost of their service to reach the level of Space X
- Problems you want to find answers
  - Build a model to determine if the rocket will land successfully to be used to build the rockets system.
  - The interaction amongst various features that determine the success rate of a successful landing
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was collected using SpaceX 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

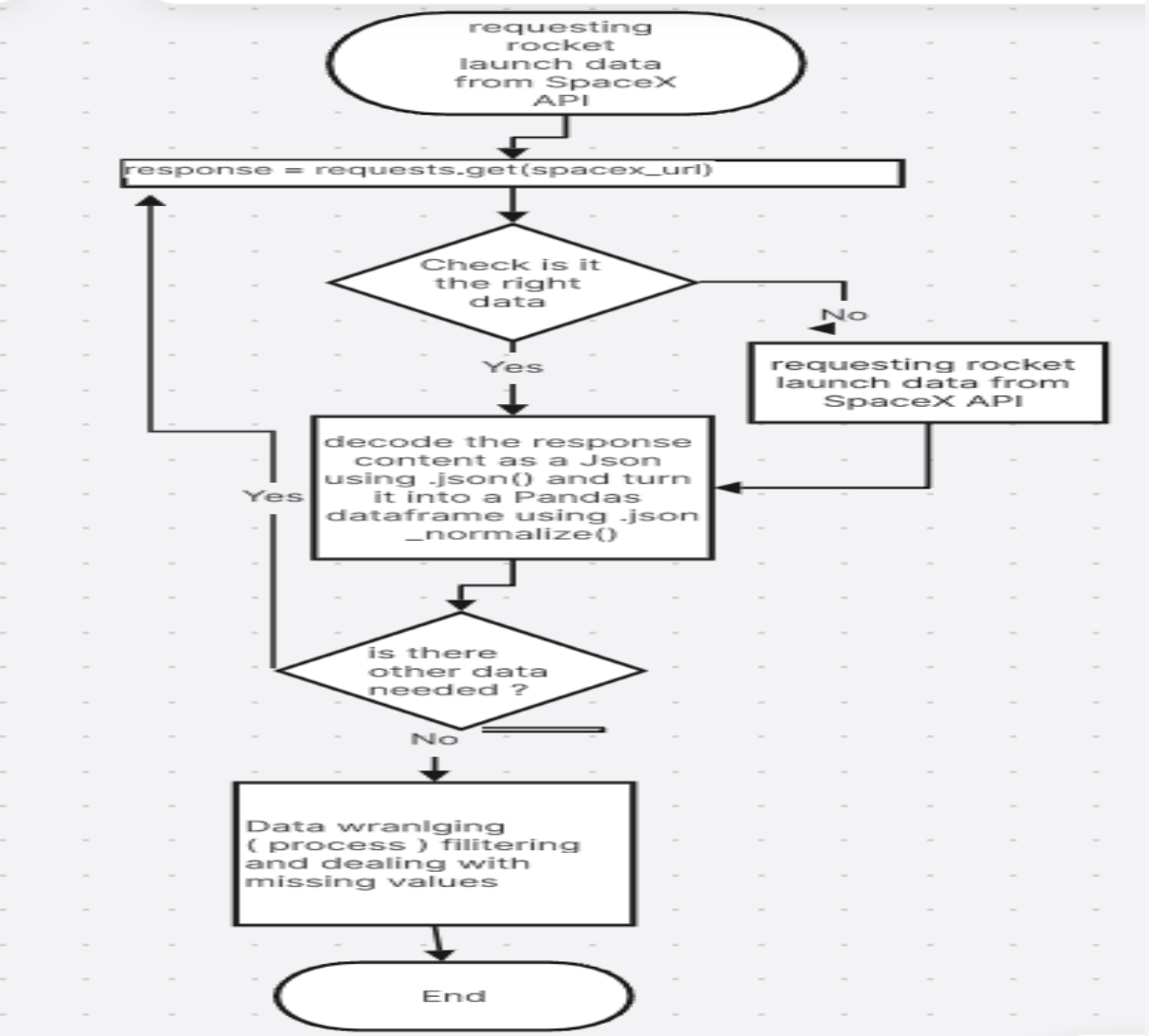
- Data collection was done using get request to the SpaceX API.- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas data frame using .json_normalize().
- We then cleaned the data, checked for missing values and fill in missing values where necessary. In addition, we 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 data frame for future analysis.

Data Collection – SpaceX API

- We used the get request to 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/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

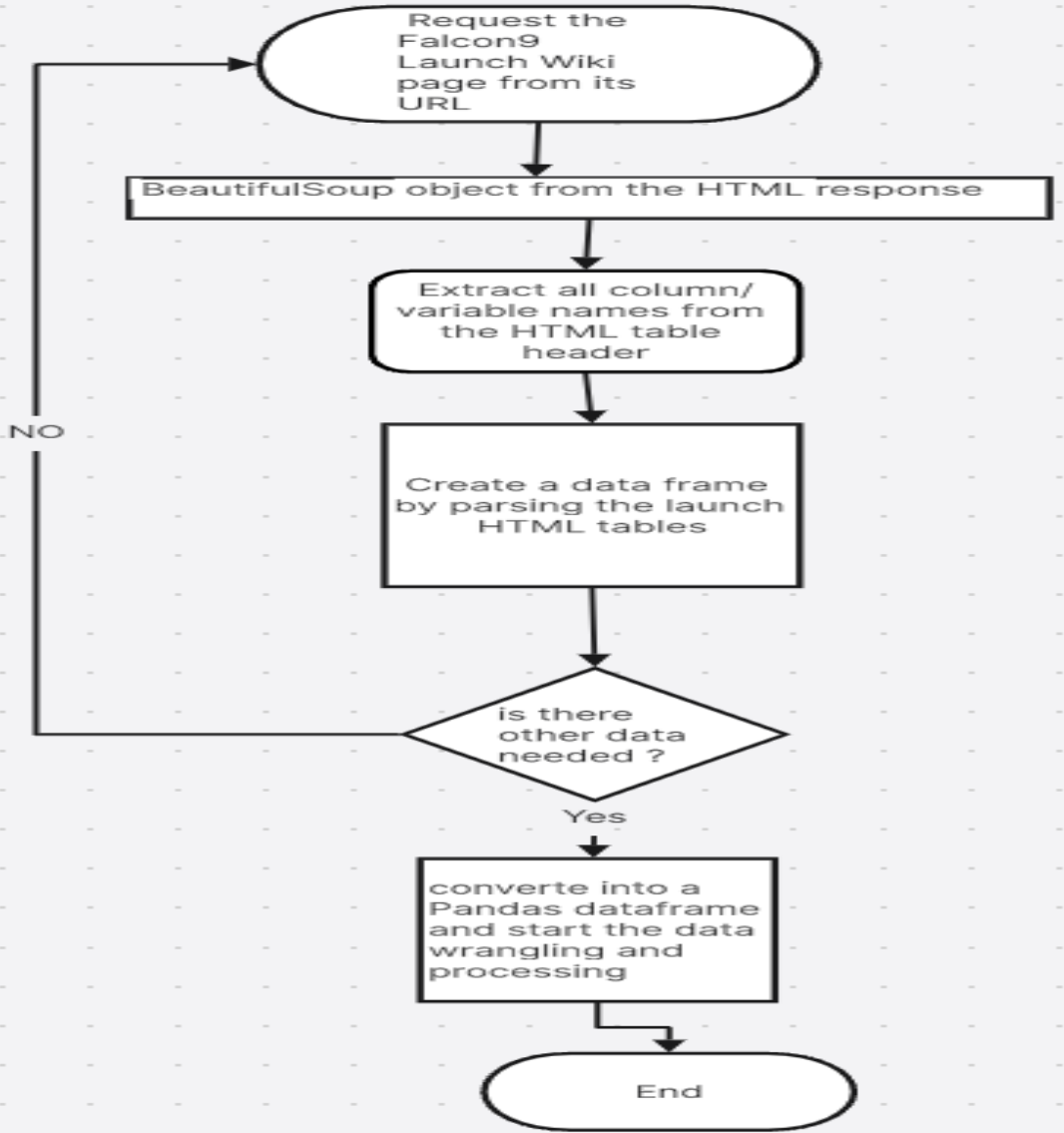


Data Collection - Scraping

- We applied web scrapping to web scrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas data frame.

- The link to the notebook is

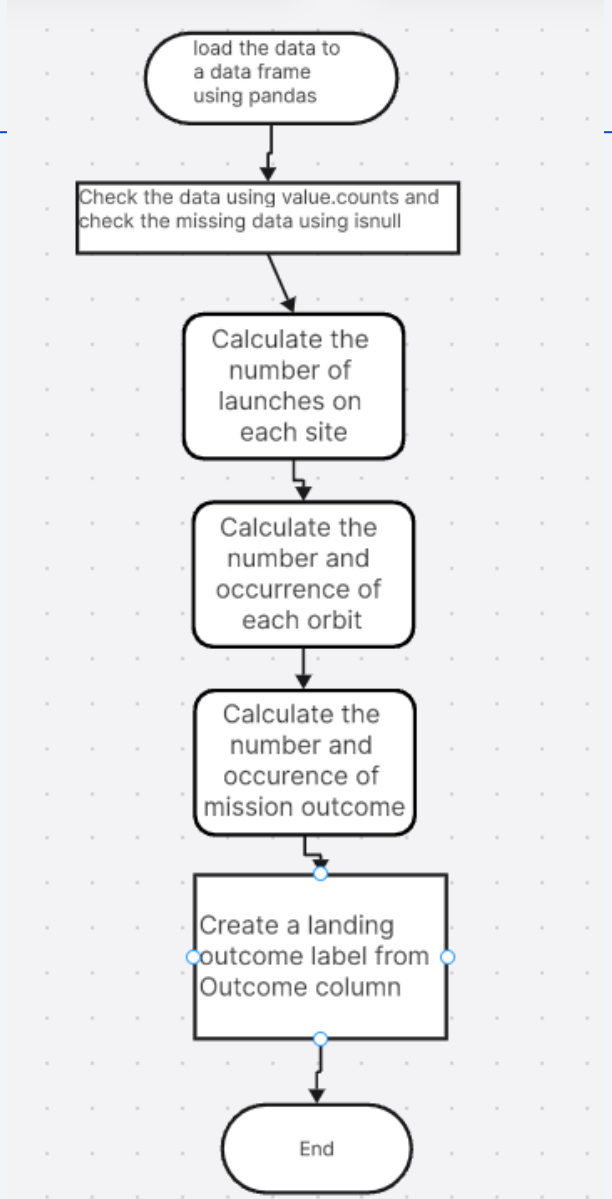
https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

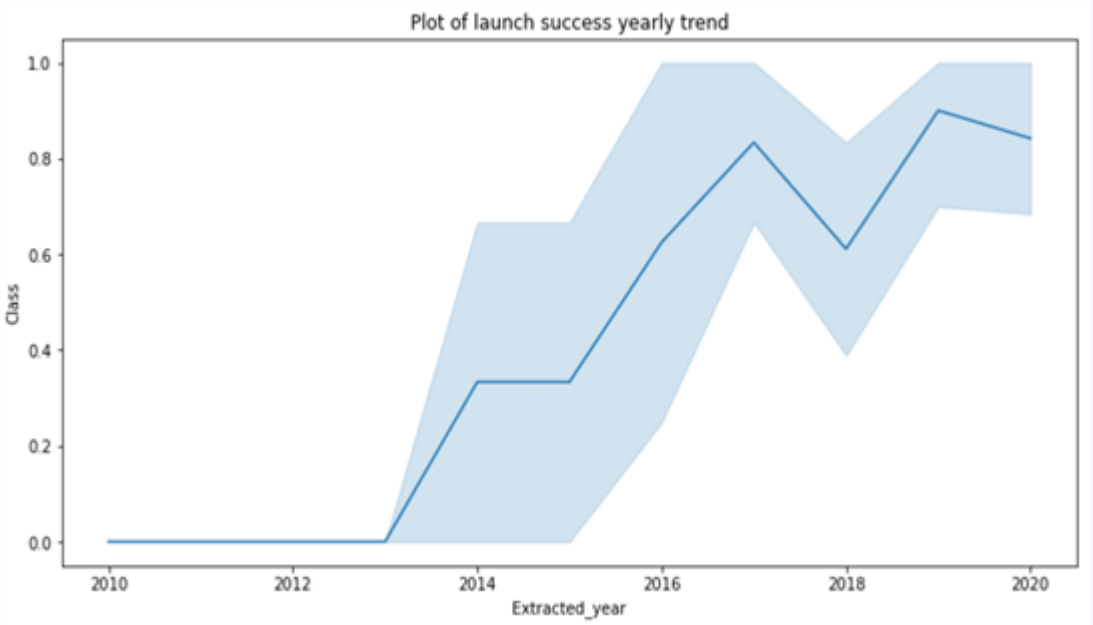
- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.
- The link to the notebook is

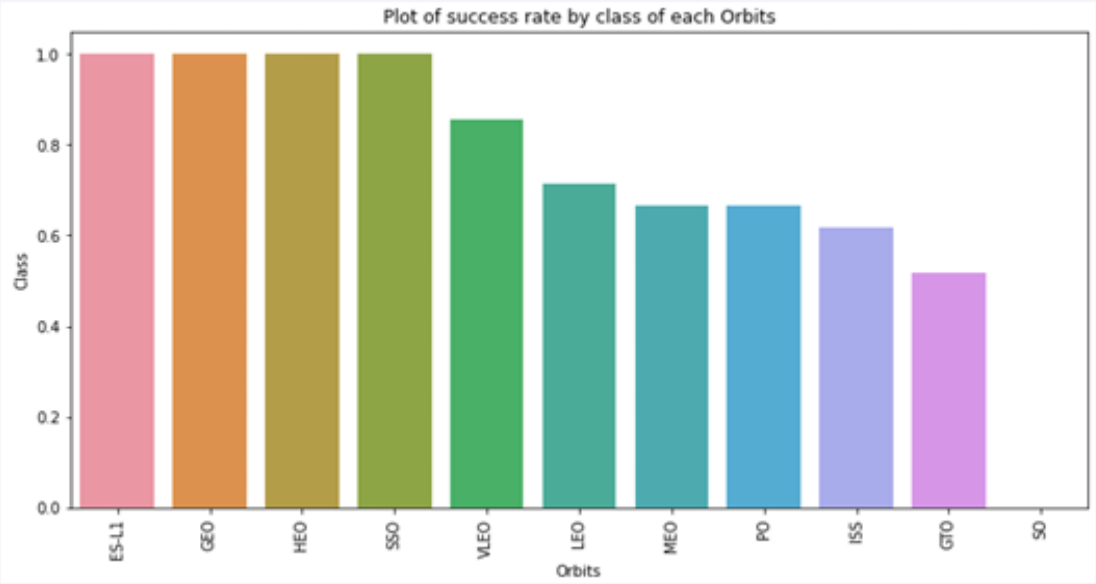
https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



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 is https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - ➢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 to the notebook is ：https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We 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.
- We 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, we identified which launch sites have relatively high success rate.
- We 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 for the note : https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

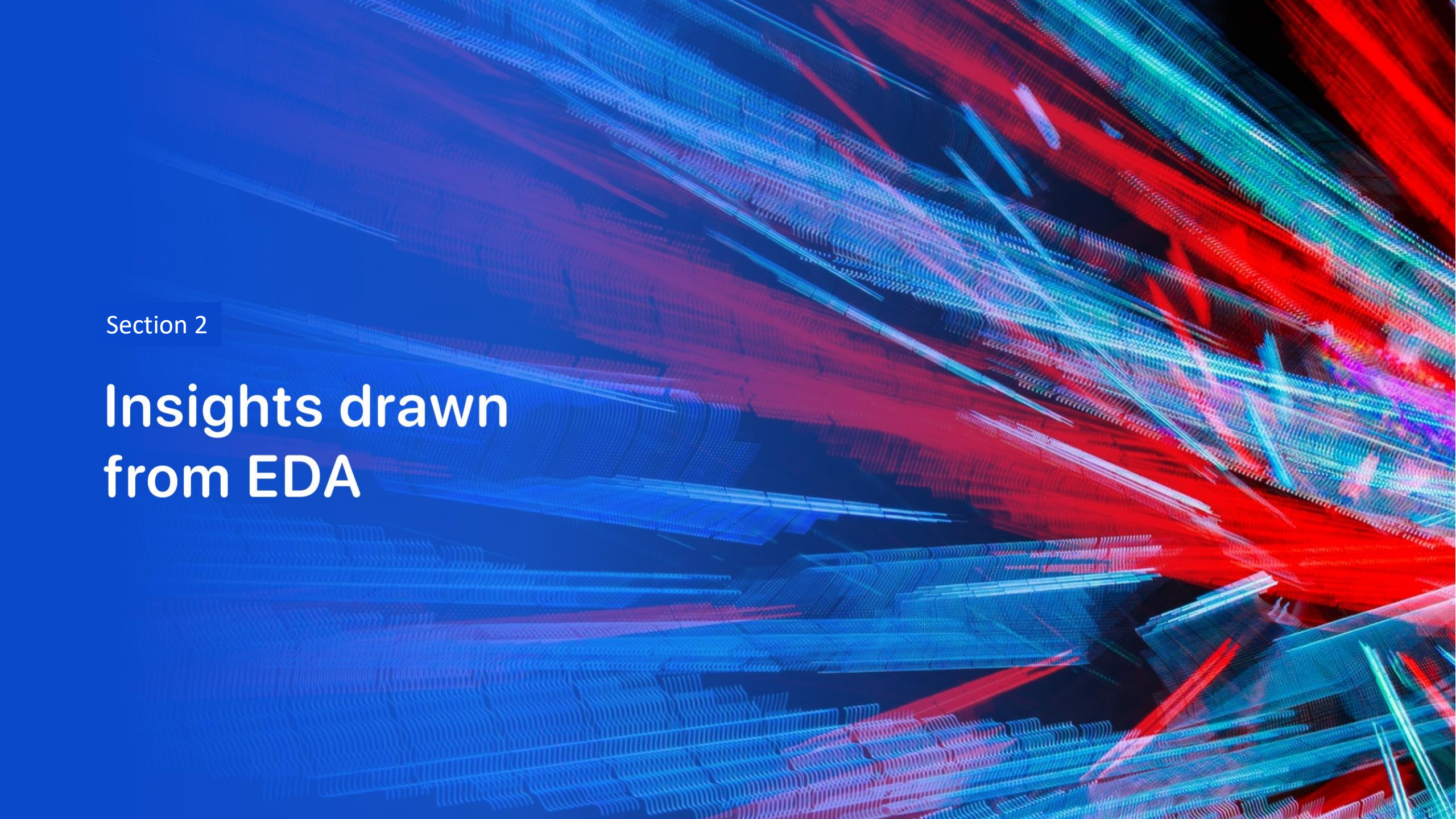
- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The link to the notebook is https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters 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 notebook is https://github.com/chuksoo/IBM-Data-ScienceCapstone-SpaceX/blob/main/Machine%20Learning%20Prediction.ipynb

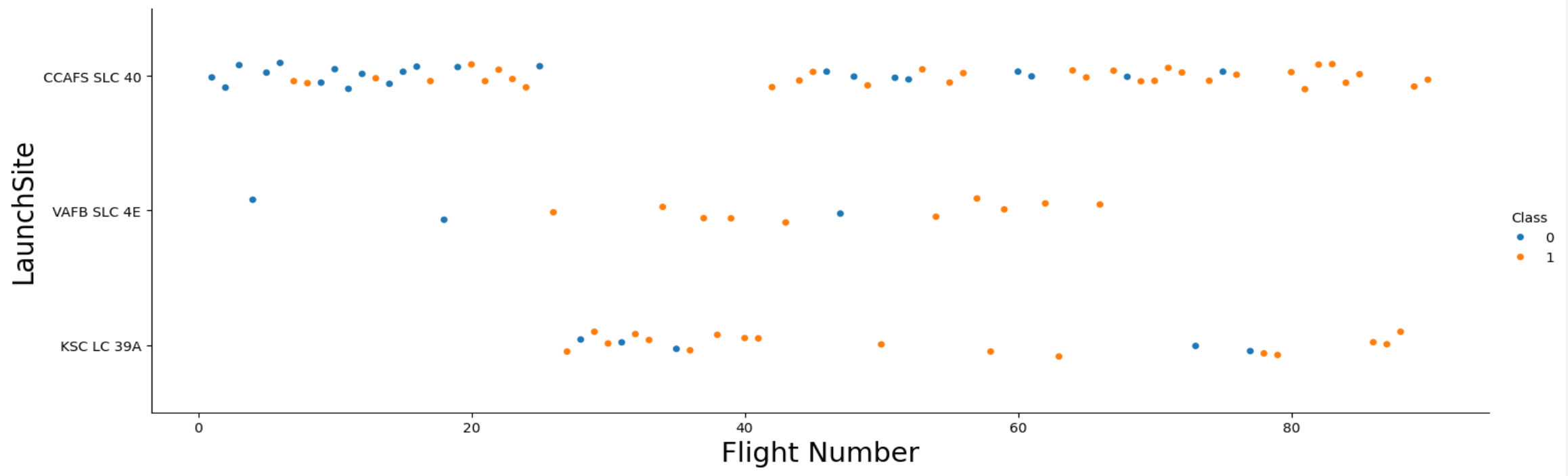
Results

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


Section 2
Insights drawn from EDA

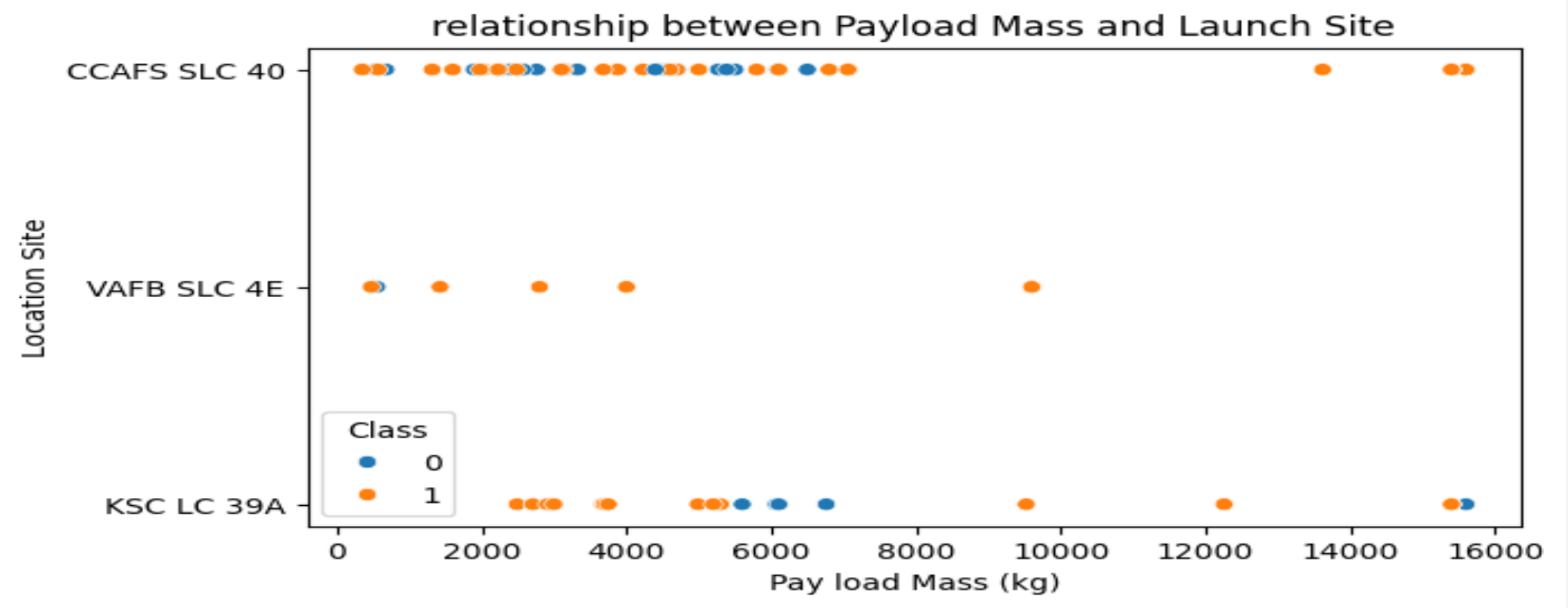
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



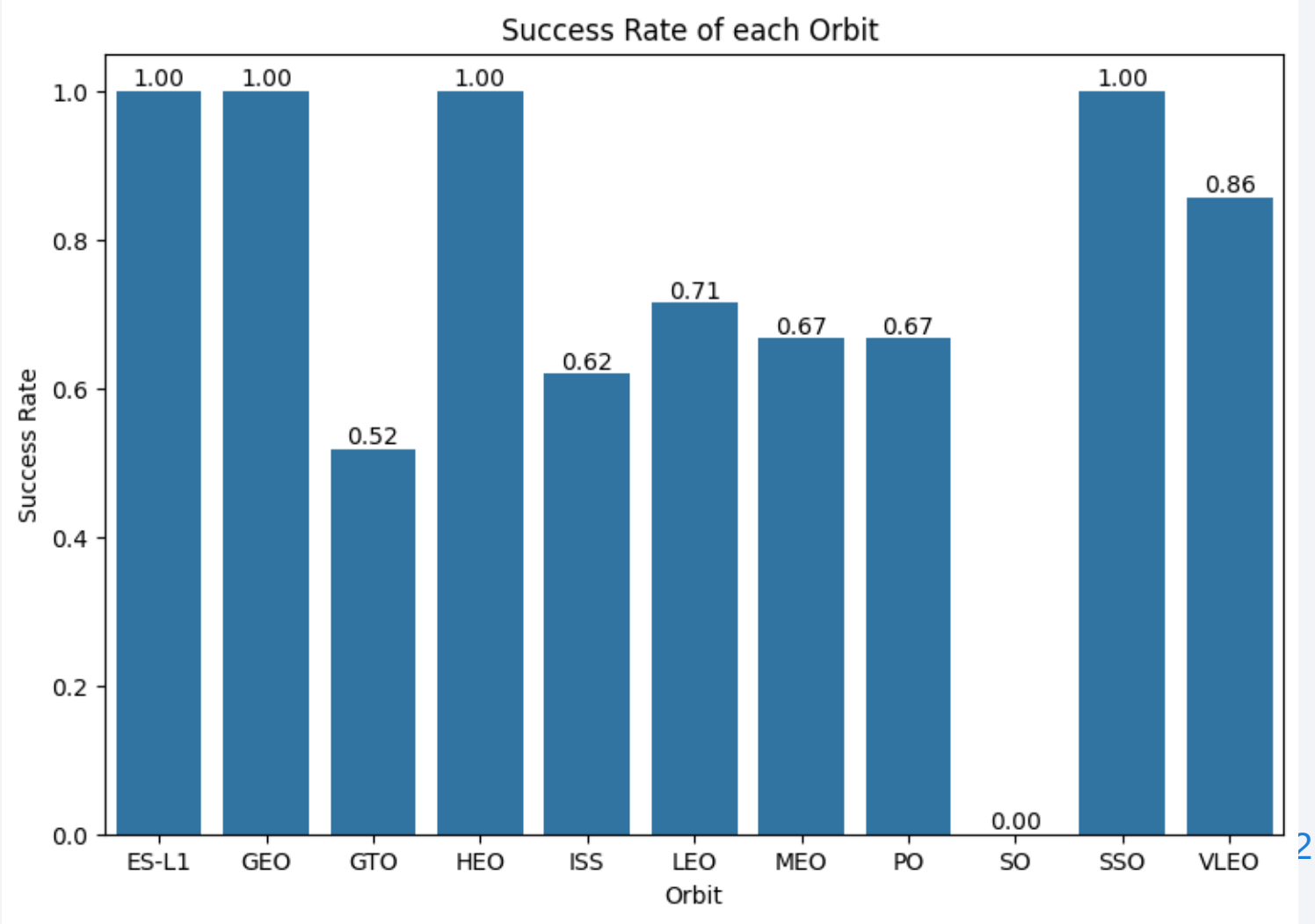
Payload vs. Launch Site

- We can observe Payload Mass 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)



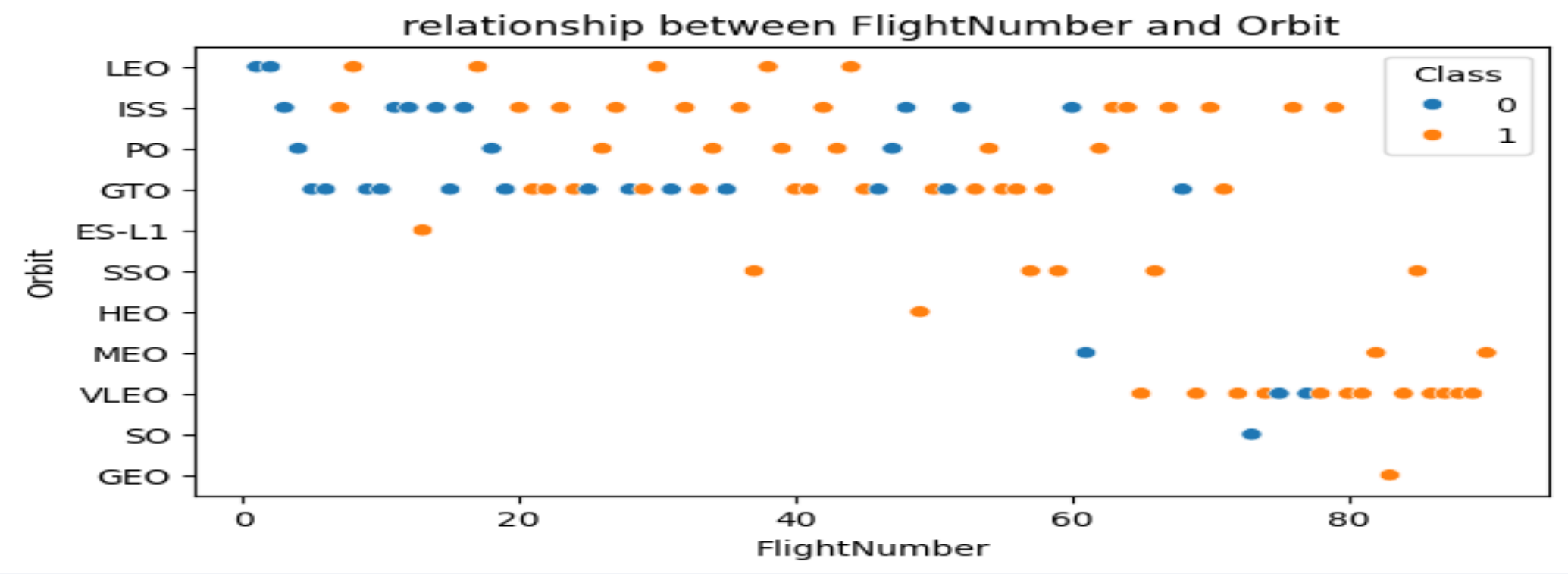
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate with the SO location having zero success rate .



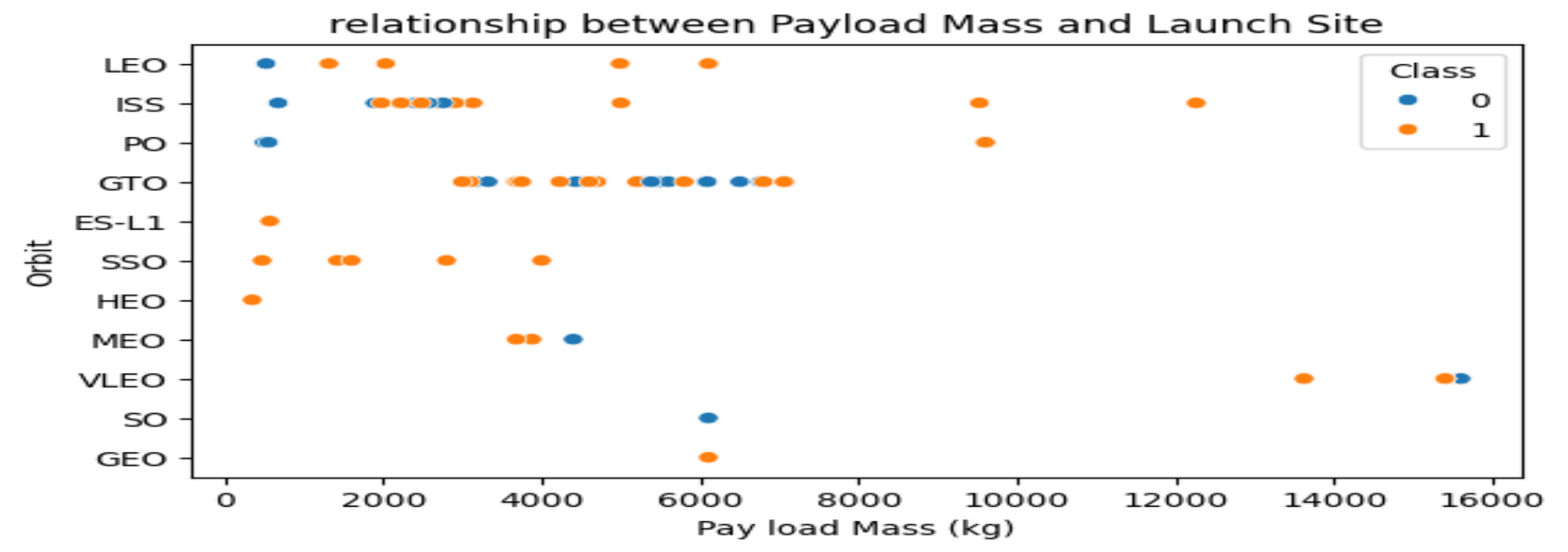
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



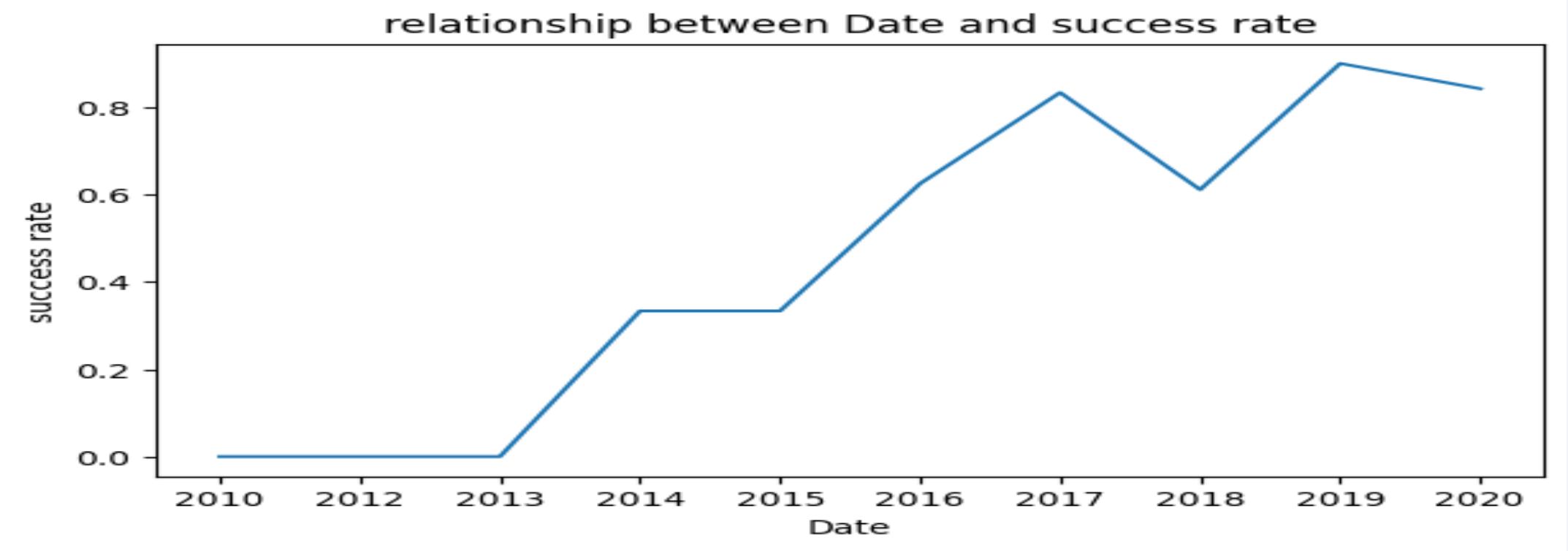
Payload vs. Orbit Type

- We can observe that With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020



All Launch Site Names

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

In [36]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Out[36]:

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

Task 2

Launch Site Names Begin with 'CCA'

- We used the query below to display 5 records where launch sites begin with `CCA`

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

```
[41]:
%%sql
SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

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

ut[41]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

Task 3

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

In [51]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer LIKE 'NASA (CRS)'
```

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

Out[51]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

Task 4

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

In [56]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1'
```

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

Out[56]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

In [60]:
```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

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

Out[60]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 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

In [67]:
```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
LIMIT 5;
```

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

Out[67]:

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like '%' to filter for WHERE MissionOutcome was a success or a failure

In [91]:
```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Count
FROM SPACEXTBL
WHERE Mission_Outcome LIKE 'Success'
   OR Mission_Outcome LIKE 'Failure (in flight)'
GROUP BY Mission_Outcome

UNION ALL

SELECT 'TOTAL' AS Mission_Outcome, COUNT(*) AS Count
FROM SPACEXTBL
WHERE Mission_Outcome LIKE 'Success'
   OR Mission_Outcome LIKE 'Failure (in flight)'
```

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

Out[91]:

Mission_Outcome	Count
Failure (in flight)	1
Success	98
TOTAL	99

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX**() function

Task 8

List the names of the booster_versions which hav

In [92]:
```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Payload_Mass__KG_ = (
    SELECT MAX(Payload_Mass__KG_)
    FROM SPACEXTBL
);
```

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

Out[92]:

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

- We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

In [103…

```
%%sql
SELECT Date
    FROM SPACEXTBL
WHERE substr(Date, 1, 4) = '2015'
  AND Landing_Outcome LIKE 'Failure (drone ship)';
```

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

Out[103…

Date
2015-01-10
2015-04-14

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

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the **GROUP BY** clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

[11]:
```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count;
```

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

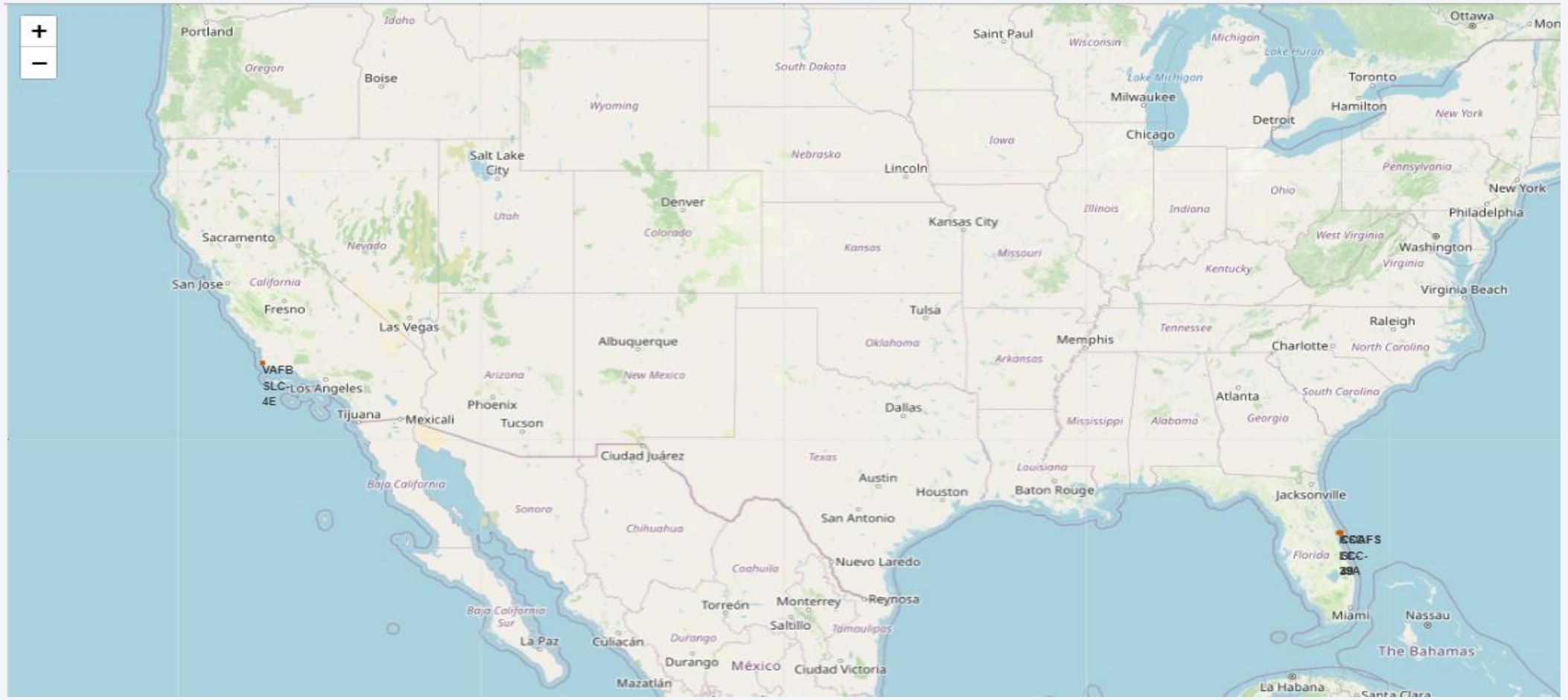
[11]:

Landing_Outcome	Count
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

Section 3

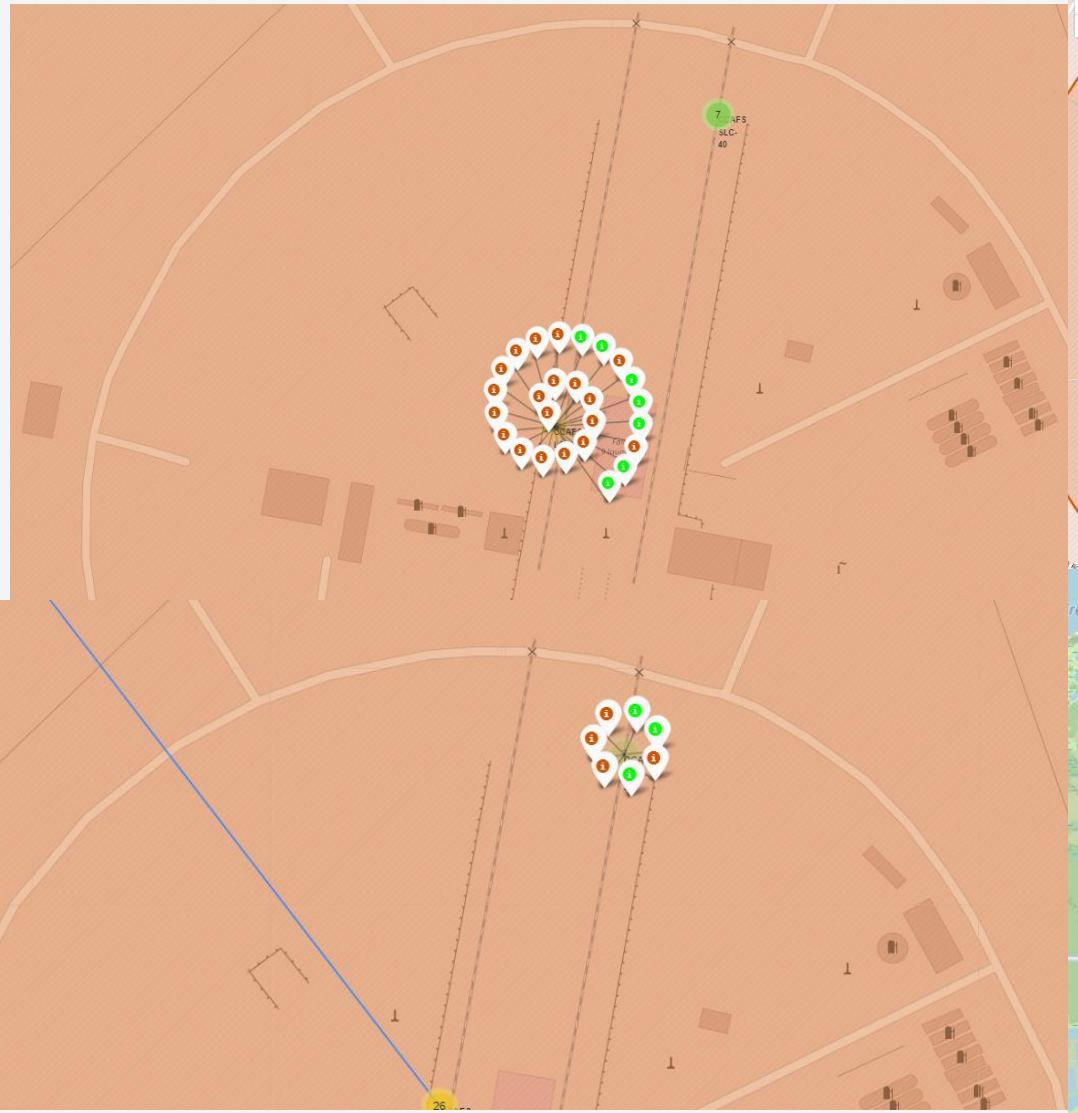
Launch Sites Proximities Analysis

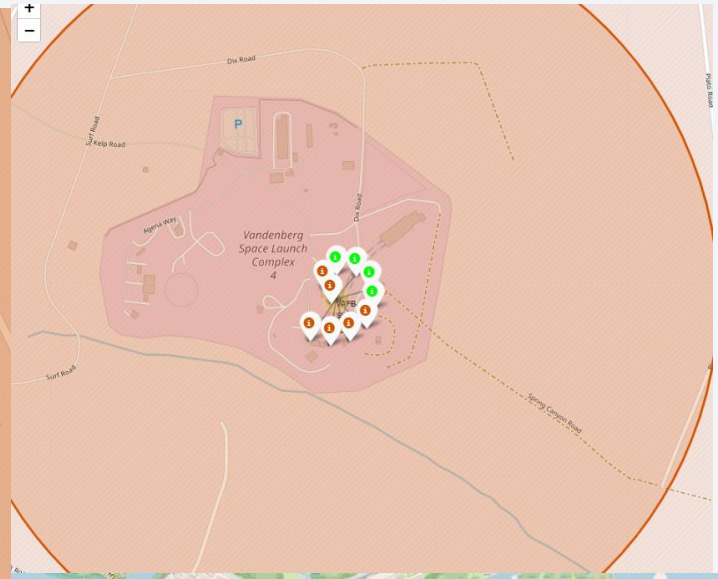
All launch sites global map markers

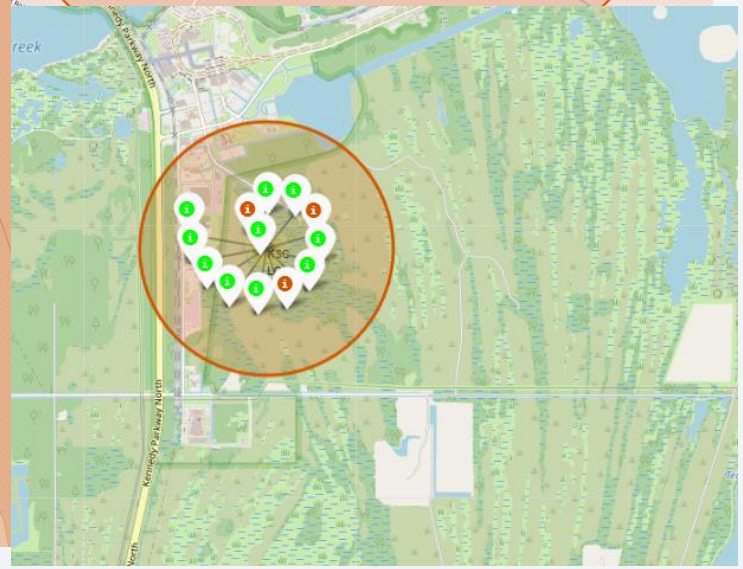


We can note that all Space lunch locations are in the US coasts areas (Florida and California)

Markers showing launch sites with color labels

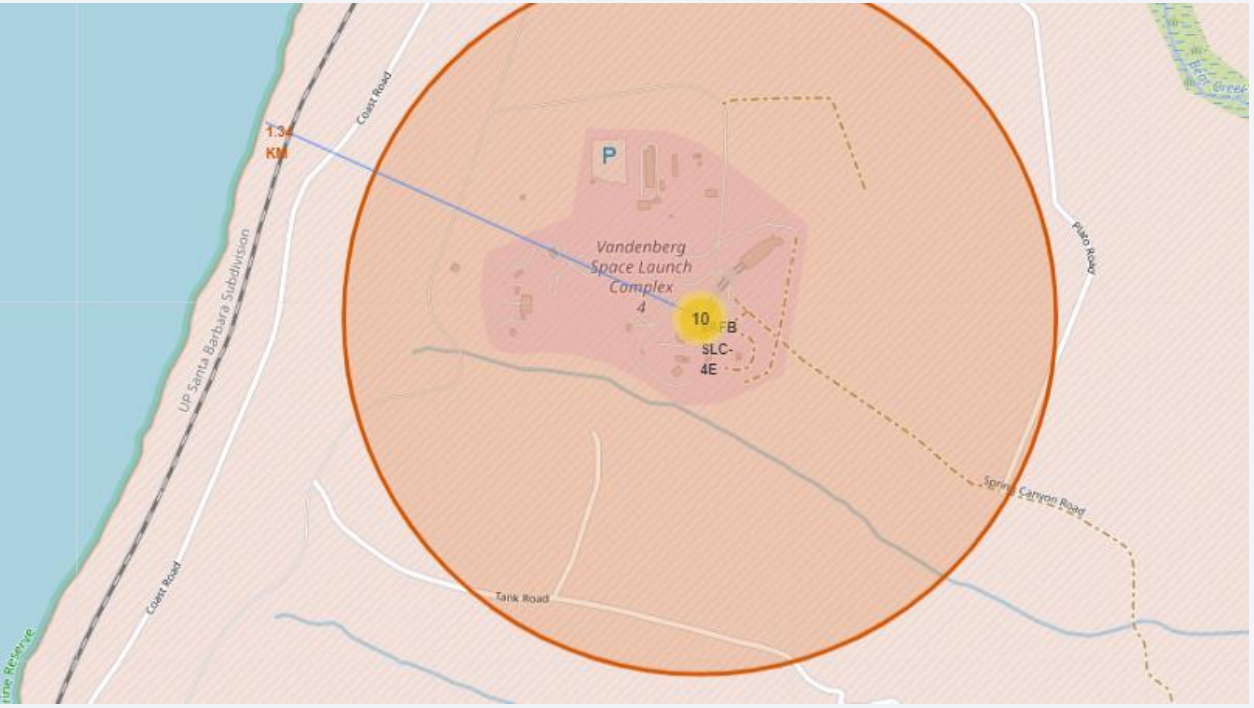






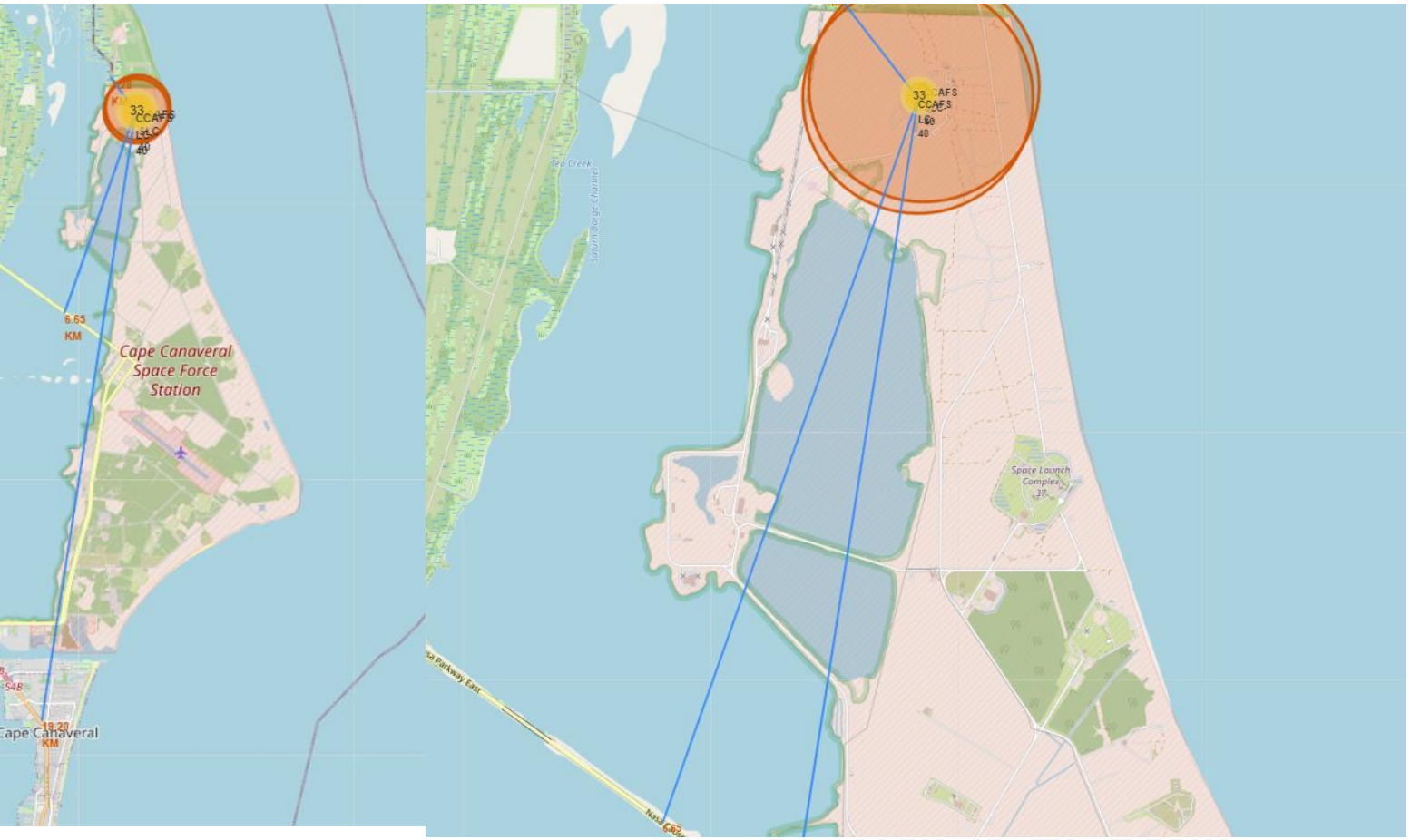
- We used markers to highlight the success and faultier attempts of landing each site (green successful and red is a faultier)

Launch Site Distance to landmarks



- Here we highlighted the distance between this site (SLC 4E) and the cost line and it’s 1.3 KM

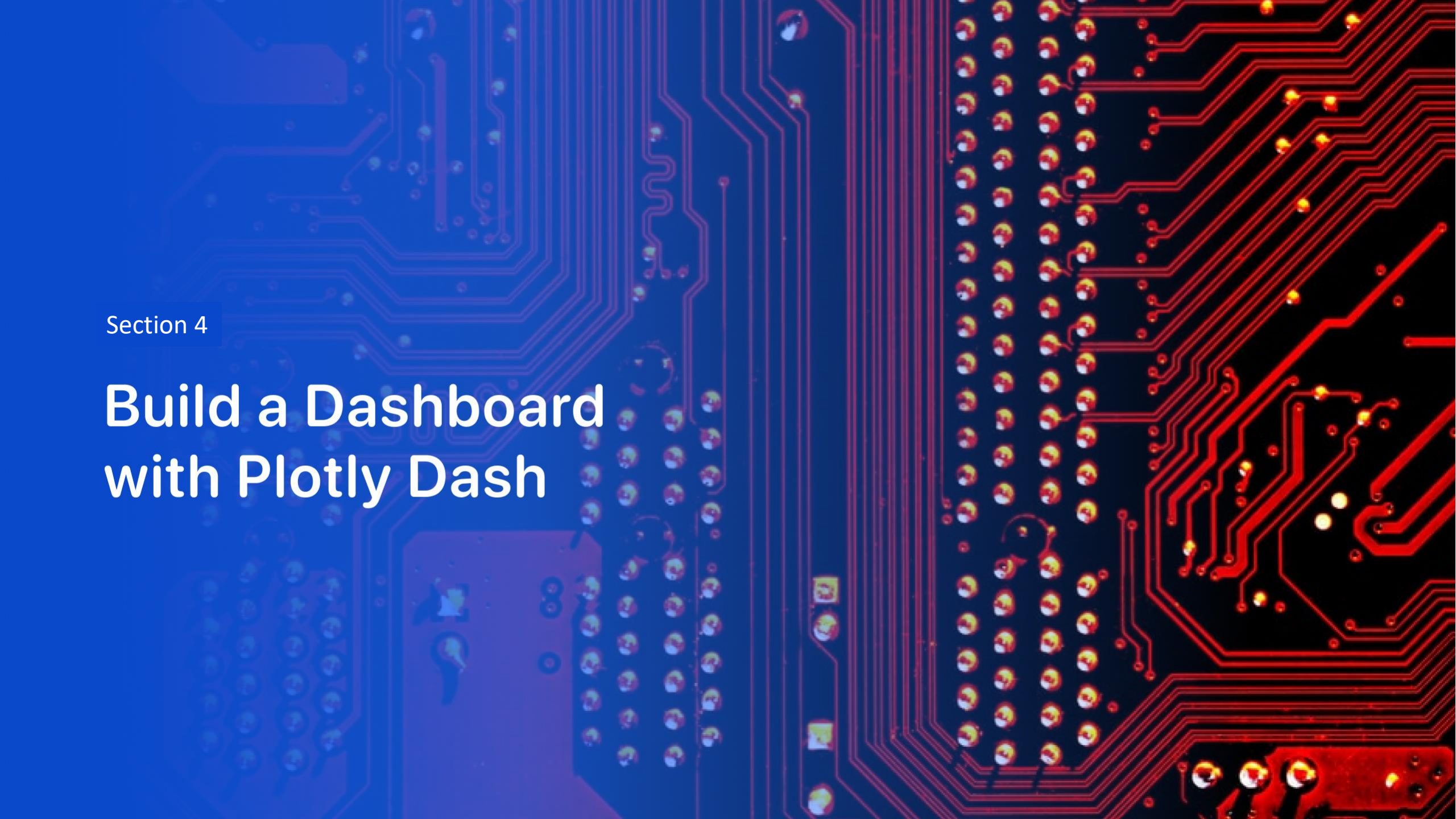
Launch Site Distance to landmarks



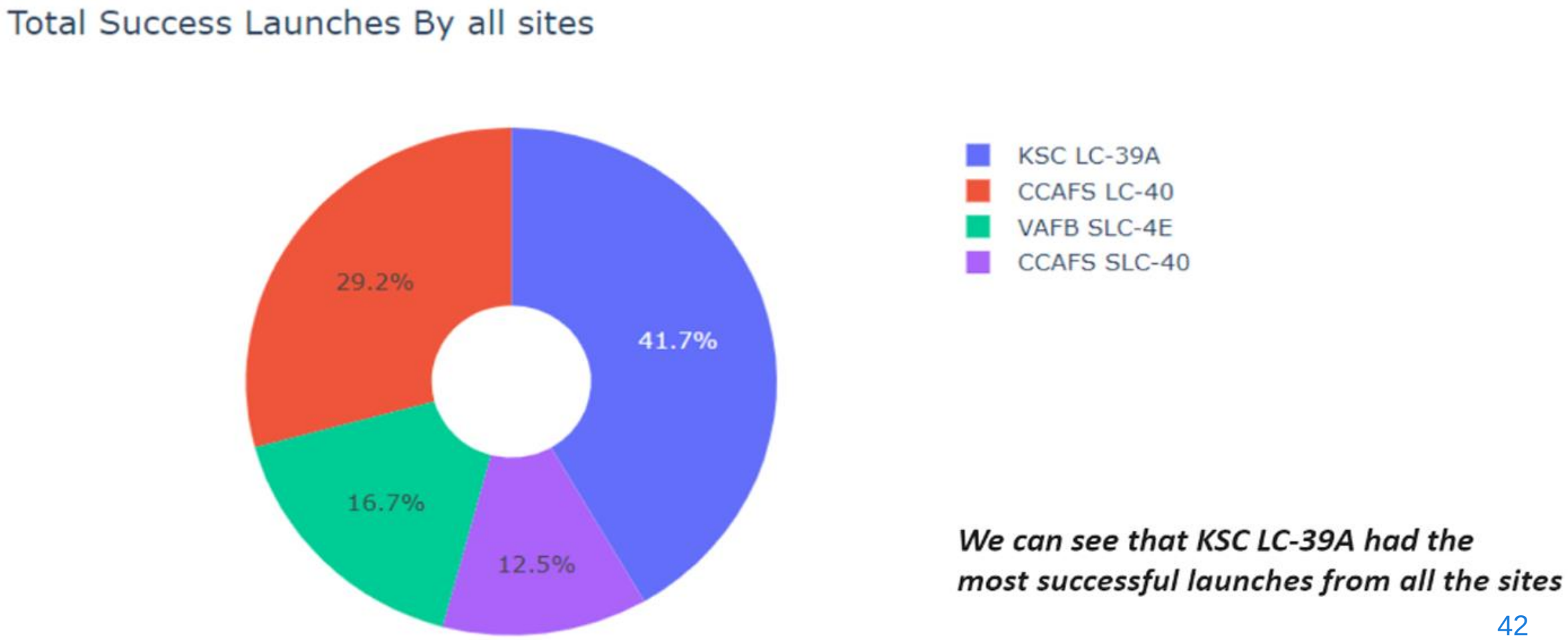
- Here we highlighted the distance between the other two sites and the landmarks (closes city , closes highway and railway) using the Mouse Position to get the coordination in the map and add_to(marker_cluster) function to add multiple markers to a map

- Here is the link for the notebook :

https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/lab_jupyter_launch_site_location.ipynb

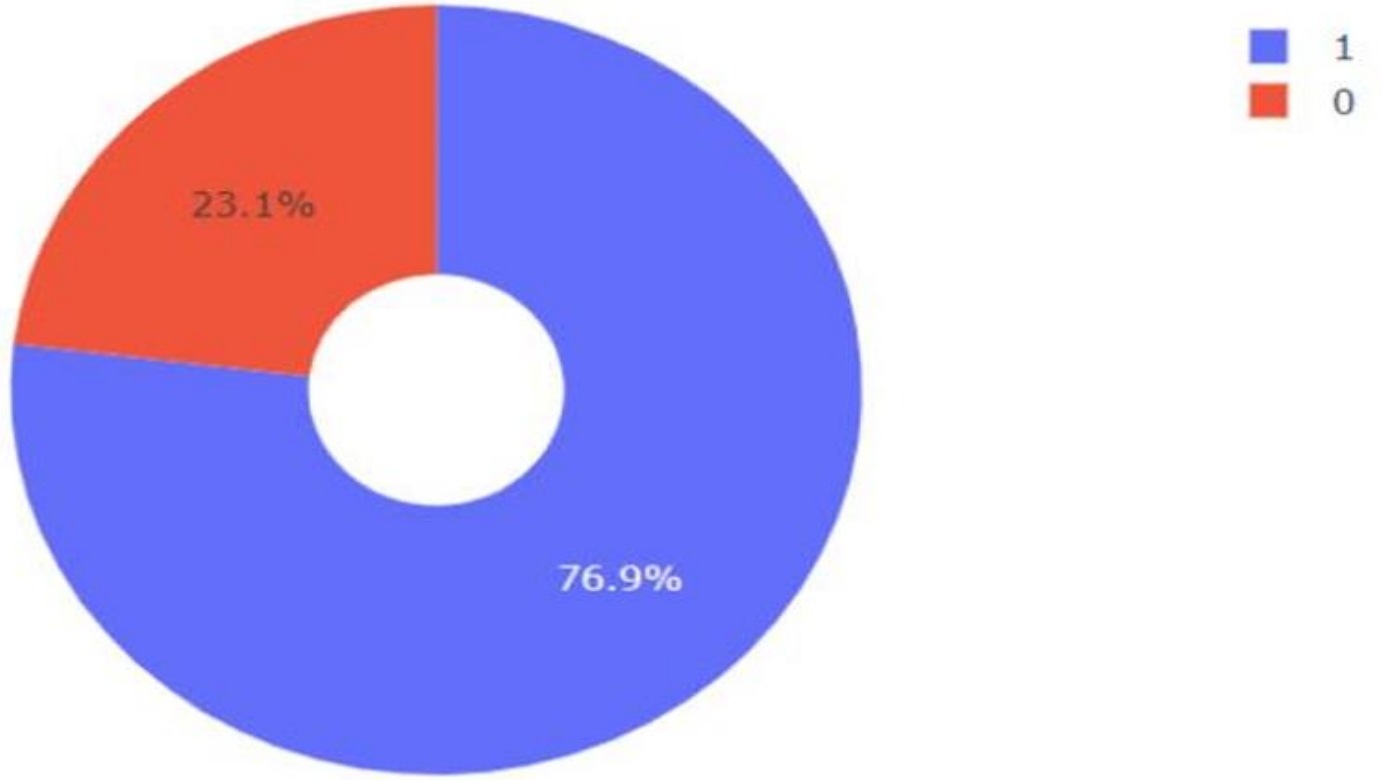

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



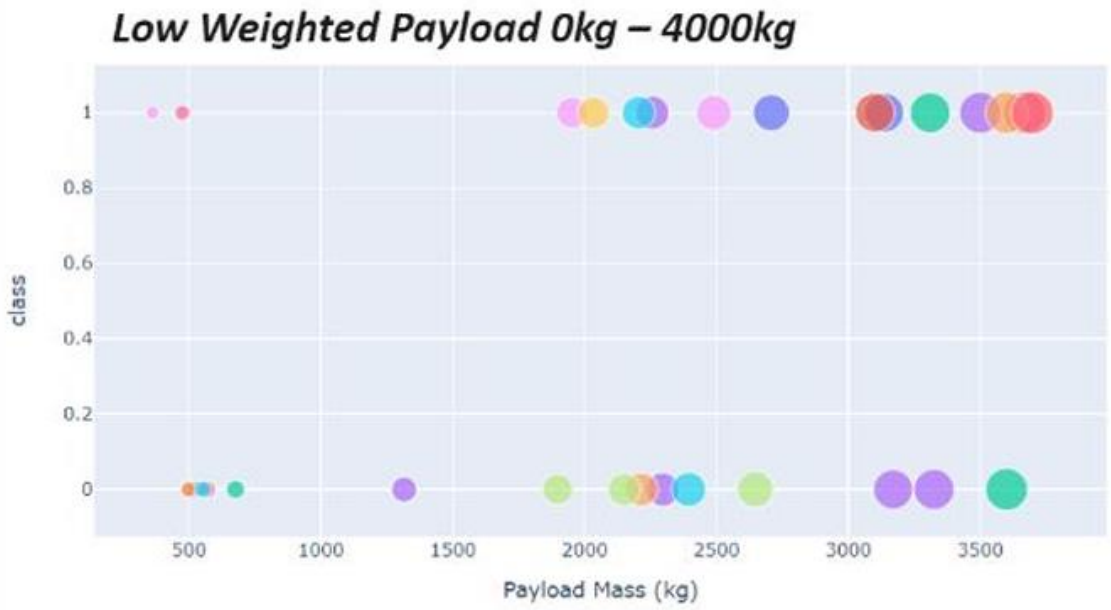
We can see that KSC LC-39A had the most successful launches from all the sites

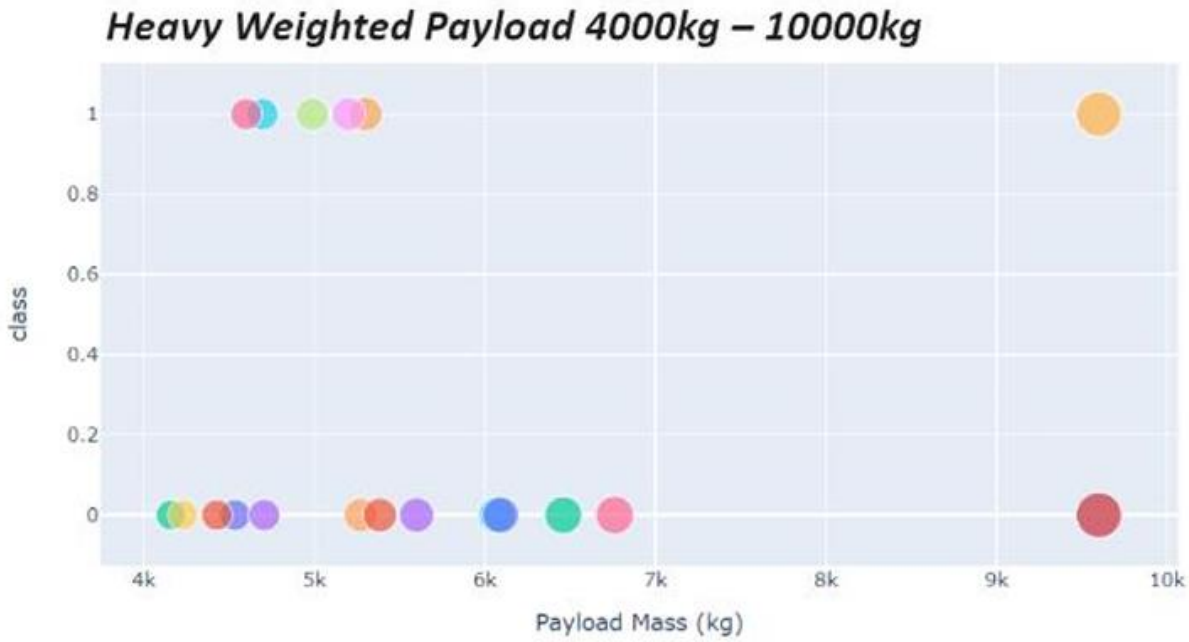
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





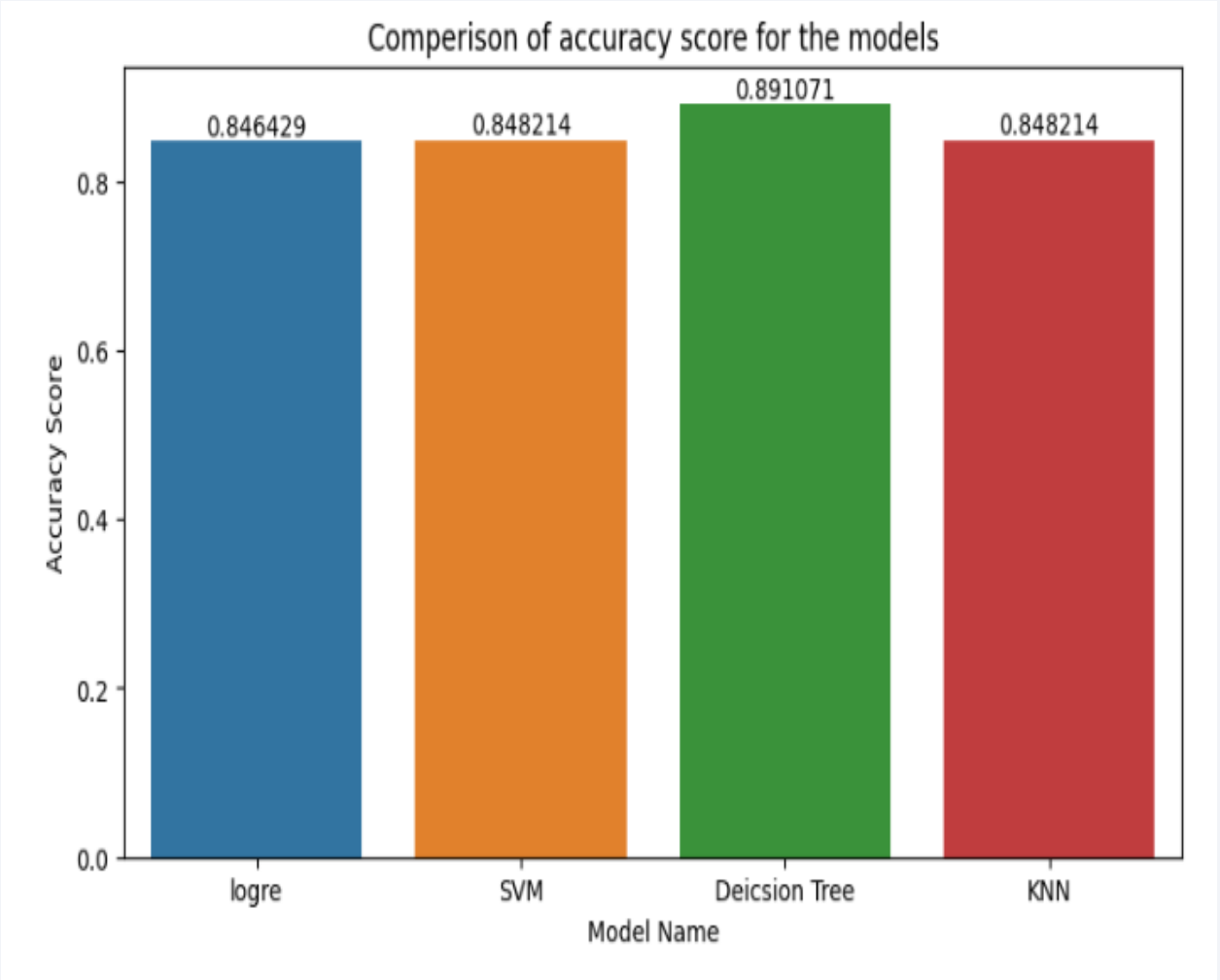
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

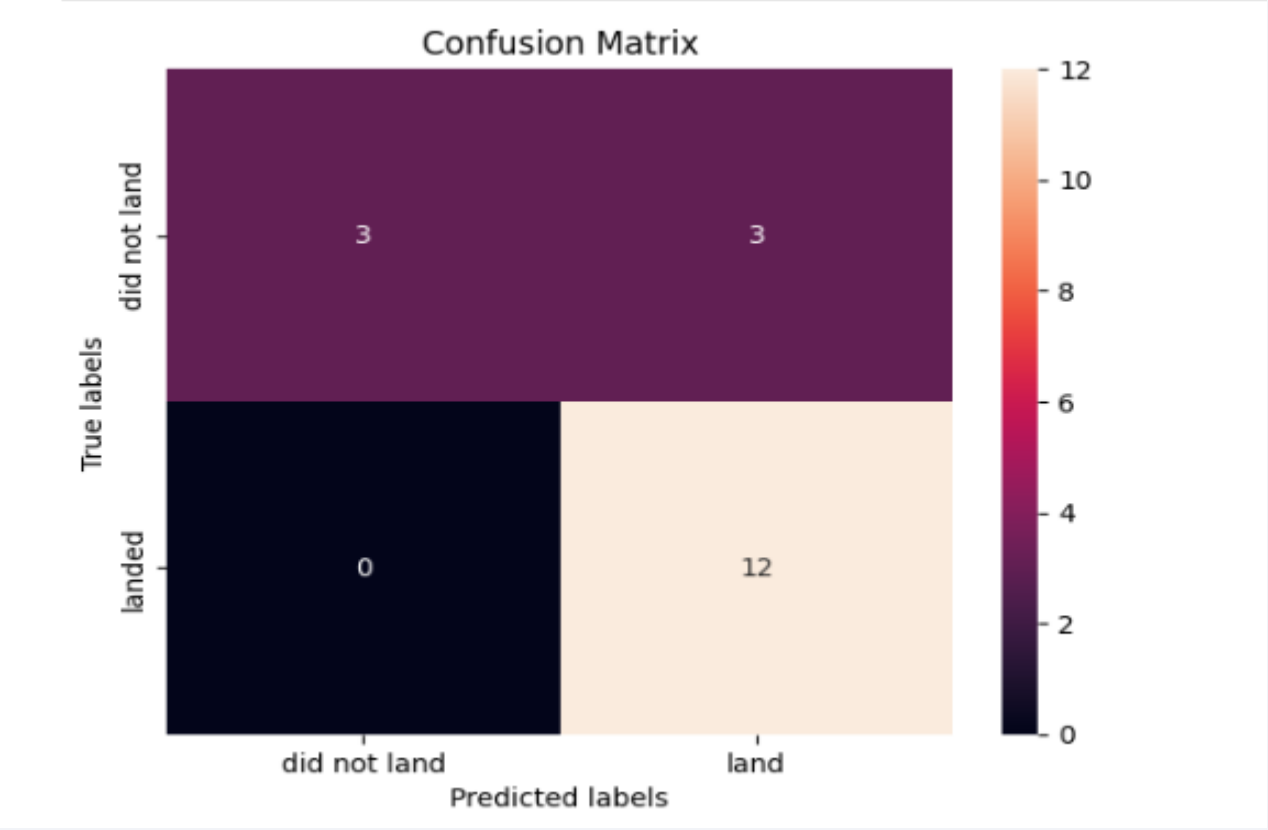
Classification Accuracy

- Used the Hyper parameter function GridSearchCV to find the best parameters and then find the best score for each model and then compared each model score using bar plot as showing the graph
- The decision tree classifier is the model with the highest classification accuracy
- The link for the notebook is :https://github.com/BaselMoh/MY-IBM-Data-Scientist-LAbs/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.
- Here we can see that 3 cases the model has put them as land and in reality they are did not land and this is the model Error



Conclusions

The project successfully developed a predictive model that can help Space Y optimize their launch processes and reduce costs. The insights gained from the EDA and the interactive visualizations provide valuable information for decision-making and strategic planning. The continuous improvement of the model and further analysis of additional features will enhance the accuracy and reliability of the predictions and here is the main points :

- The larger the flight amount at a launch site, the greater the success rate at launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!