Winning Space Race with Data Science

Alexis Yanes Sanz
14.01.2024

Outline

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

Executive Summary

Summary of methodologies

- Collect Space X data by using SpaceX API and Web Scraping.
- Perform data wrangling using Pandas and EDA using visualization and SQL.
- Create interactive visual analytics using Folium and Plotly Dash.
- Predict analysis using classification models in Scikit-learn library.

Summary of all results

- The best classification model is Decision Tree Classifier model with accuracy of 94.44%.
- From the confusion matrix, Decision Tree Classifier can distinguish between the different classes, but have false positives as a major problem.

Introduction

Project background and context

• In this capstone, I am a data scientist working for a new rocket company named “Space Y” that would like to compete with SpaceX.

- My job is to determine the price of each launch by gathering information about Space X and creating dashboards.
- I also determine if SpaceX will reuse the first stage by training a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Problems you want to find answers

- Determine if Space Y should reuse the first stage rocket based on machine learning model, trained using Space X data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting Space X data by using SpaceX API and Web Scraping.
- Perform data wrangling.
  - Analyzing and Cleaning data using Pandas library.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models.
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using Scikit-learn library.

Data Collection

Describe how data sets were collected.

- SpaceX API.
- Web Scraping.

You need to present your data collection process use key phrases and flowcharts.

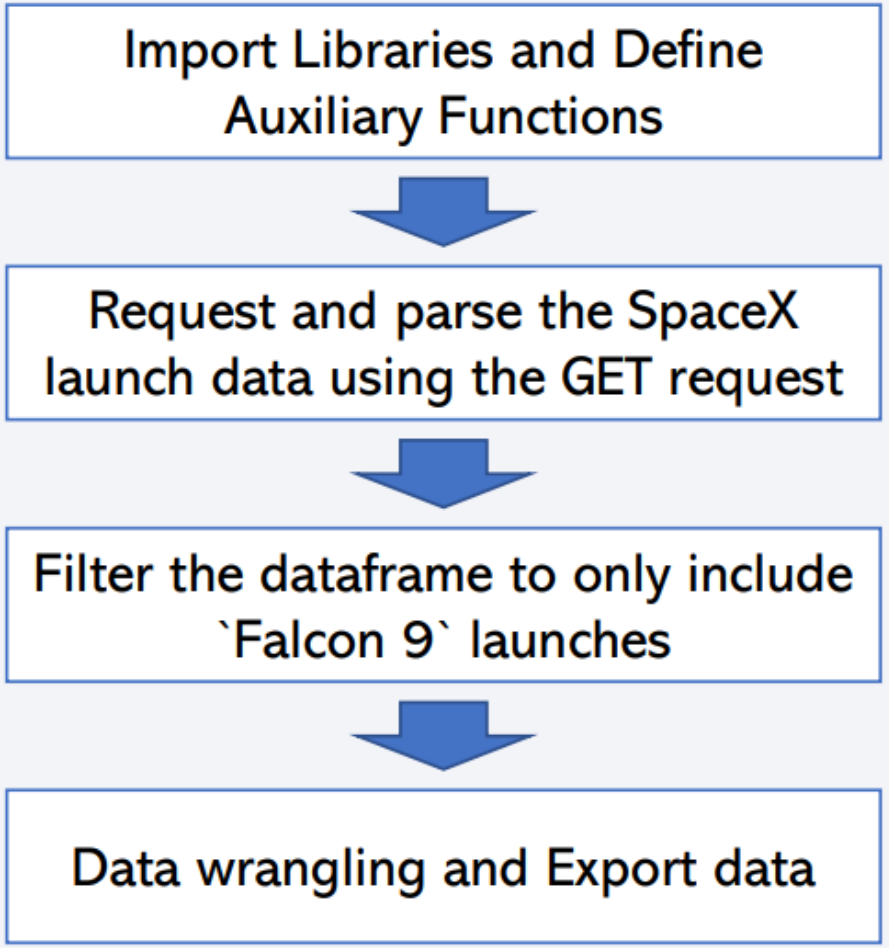
- GitHub URL as an external reference and peer-review purpose.

Data Collection – SpaceX API

- SpaceX data were collected by SpaceX REST calls API as described in the flowchart.

- GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/1%20Collecting%20the%20data.ipynb

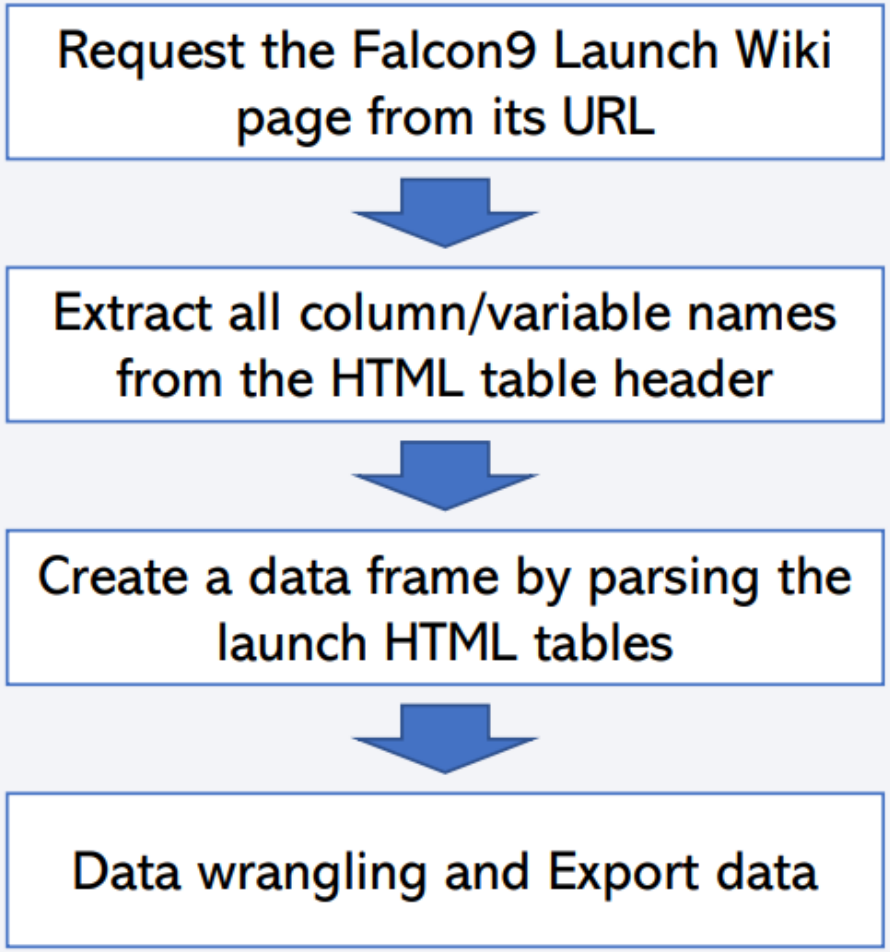


Data Collection - Scraping

- SpaceX data were collected by web scraping as described in the flowchart.

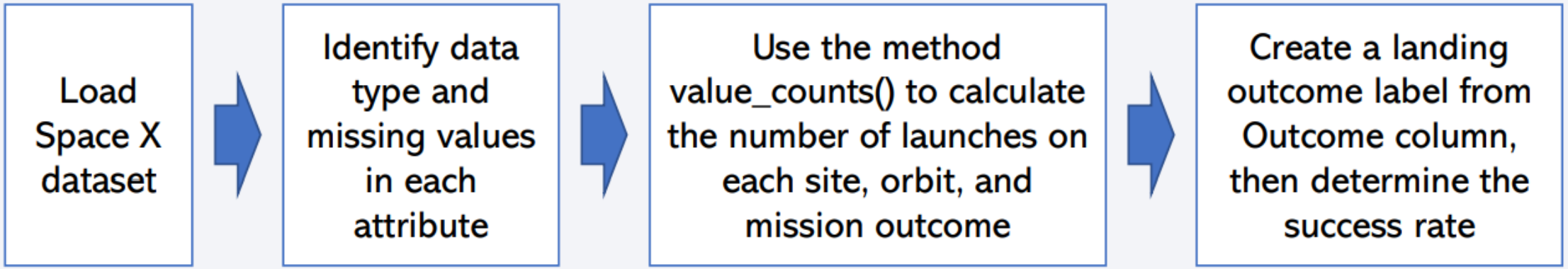
- GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/2%20Web%20scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20from%20Wikipedia.ipynb



Data Wrangling

- Analyzing and Cleaning data using Pandas library.
- Data wrangling process flowcharts.



- GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/3%20Data%20wrangling.ipynb

EDA with Data Visualization

Summary

- Use scatter plot to visualize the relationship between Flight Number, Payload, Launch Site and Orbit type.
- Use bar plot to visualize the relationship between success rate of each orbit type.
- Use line plot to visualize the launch success yearly trend.

GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project

EDA with SQL

Summary.

- Load data set to DB2 server.
- Query the names of the unique launch sites in the space mission.
- Query the total payload mass carried by boosters launched by NASA (CRS).
- Query average payload mass carried by booster version F9 v1.1.
- Query the date when the first successful landing outcome in ground pad was achieved.
- Query the total number of successful and failure mission outcomes.

GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/5%20SQL%20Notebook%20for%20Peer%20Assignment.ipynb

Build an Interactive Map with Folium

Summary:

- Mark all launch sites on a map using circle and marker objects.
- Mark the success/failed launches for each site on the map using marker cluster objects.
- Calculate the distances between a launch site to its proximities using MousePosition, then mark on a map using marker and polyline objects.

GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/6.%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Summarize what plots/graphs and interactions you have added to a dashboard:

- Create dashboard with 4 components including dropdown menu, pie chart, slider, and scatter plot.

Explain why you added those plots and interactions:

- Dropdown menu for selecting launch sites.
- Pie chart to visualize success rate in each launch site.
- Slider to select payload range.
- Scatter plot to visualize relationship launch site, payload, and booster version.

GitHub URL:

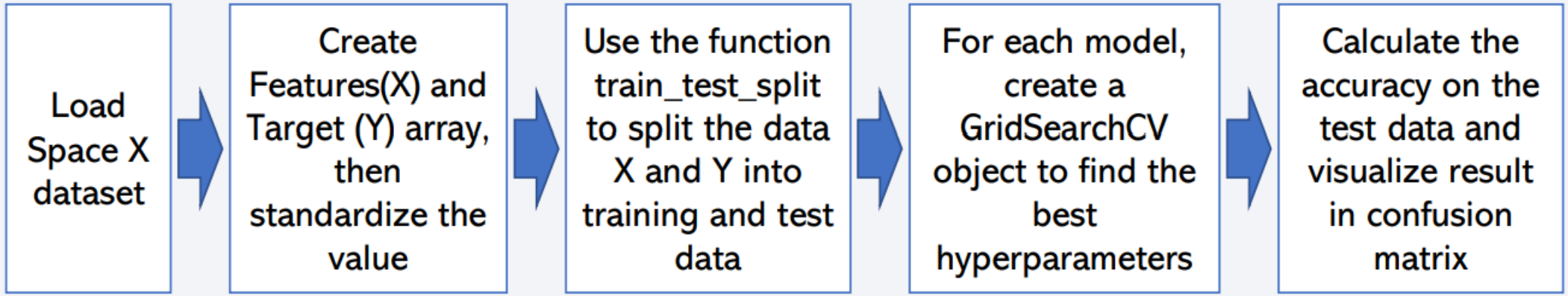
https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/7.%20Dashboard%20Application%20with%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

Summary:

- Develop best model from SVM, Classification Trees and Logistic Regression using Scikitlearn library to determine if Space Y should reuse the first stage rocket.

Model development process flowchart:



GitHub URL:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project/blob/main/8%20Machine%20Learning%20Prediction.ipynb

Results

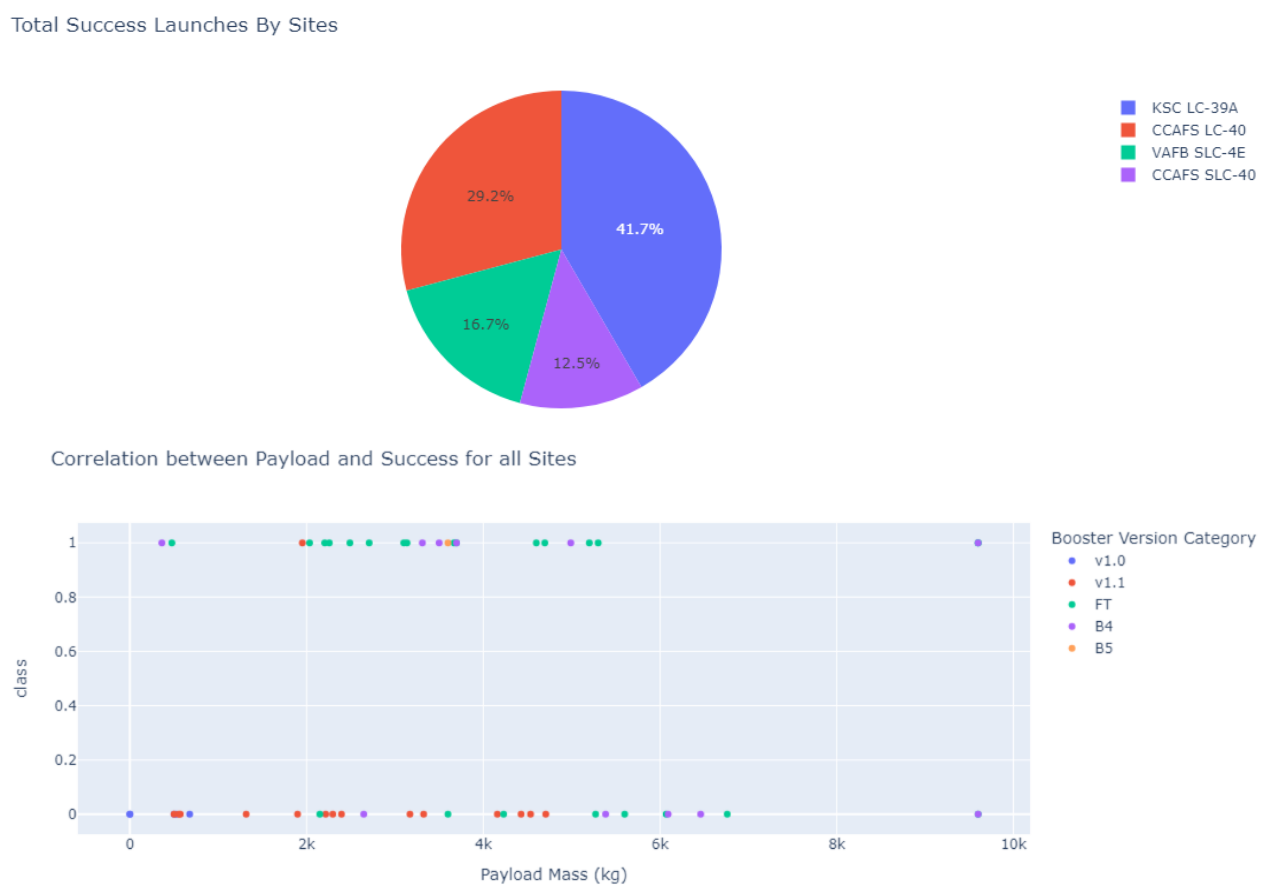
Exploratory data analysis results:

- Visualize trend and relationship between each attribute.
- Query for specific information.

Interactive analytics demo in screenshots:

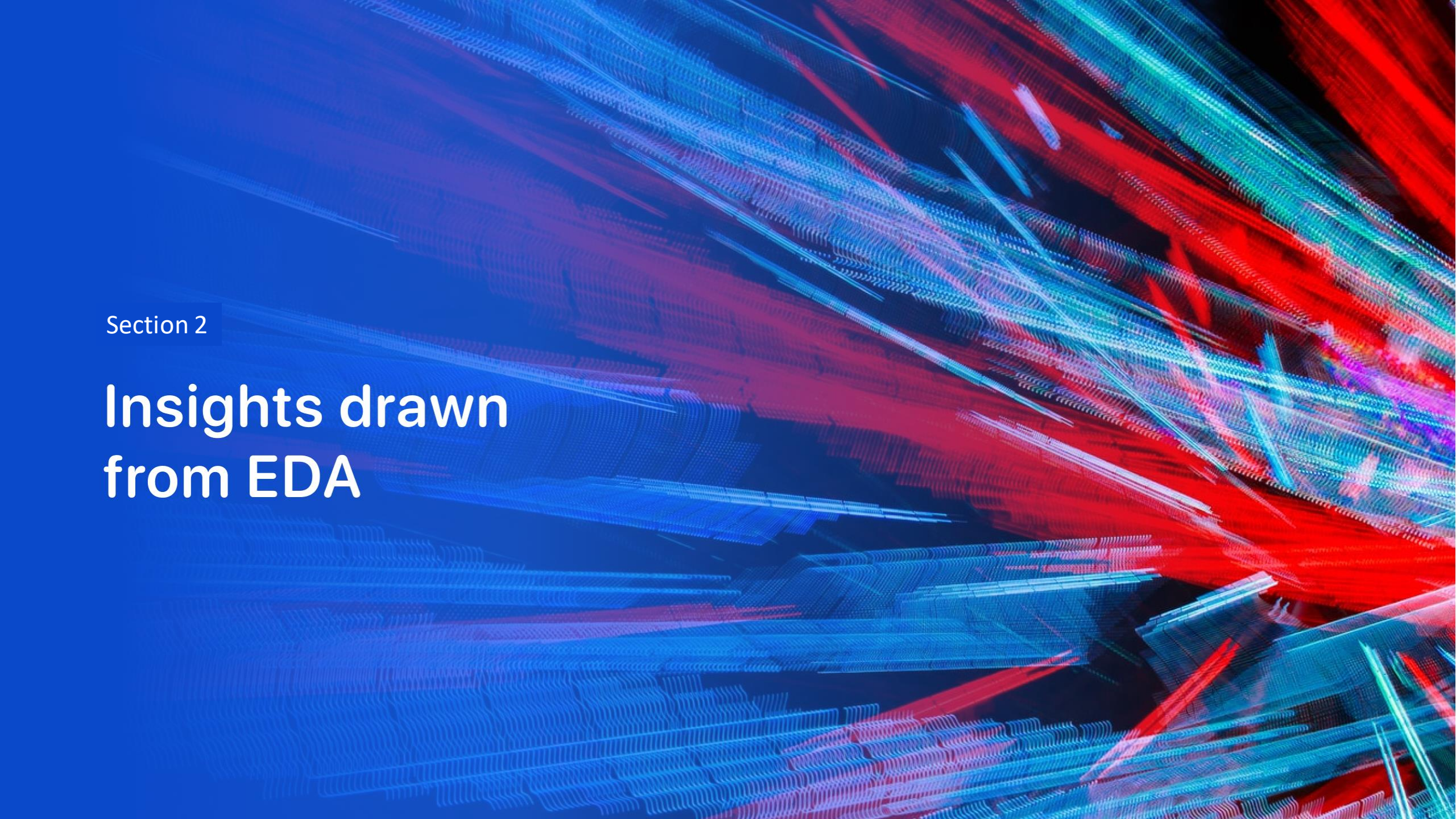
Predictive analysis results:

- The best classification model is Decision Tree Classifier model with accuracy of 94.44%.



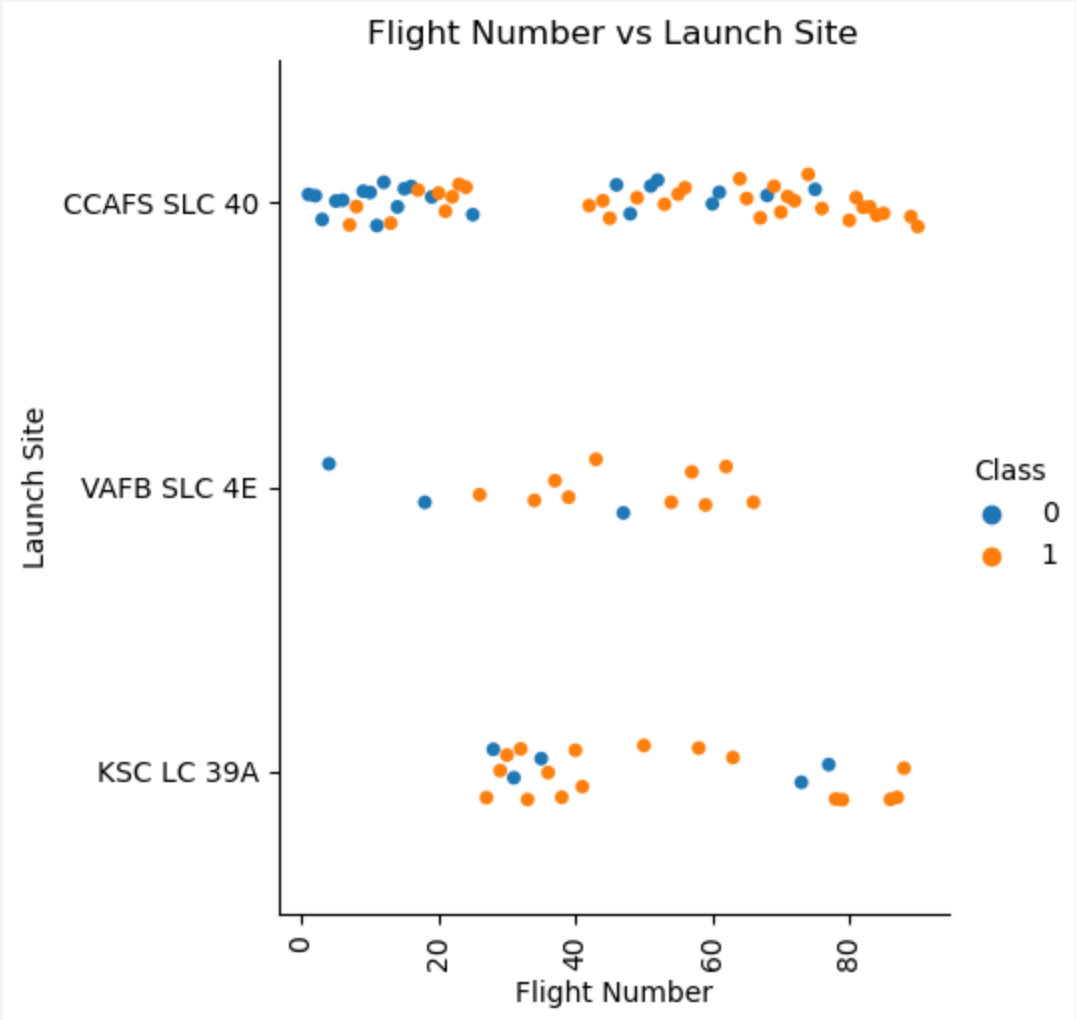
Section 2

Insights drawn from EDA



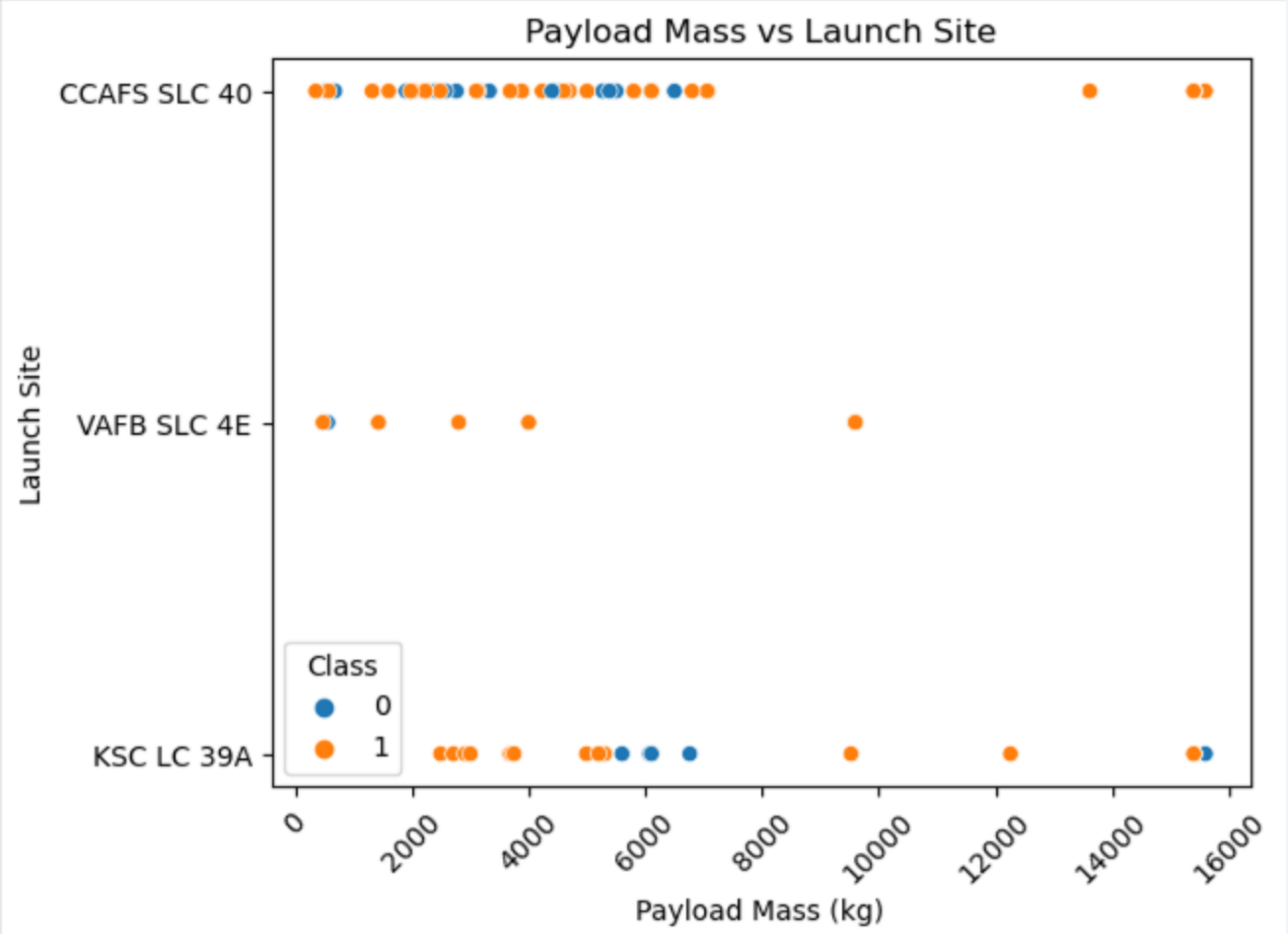
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site.
- Launch Site CCAFS SLC 40 has the most number of launch.



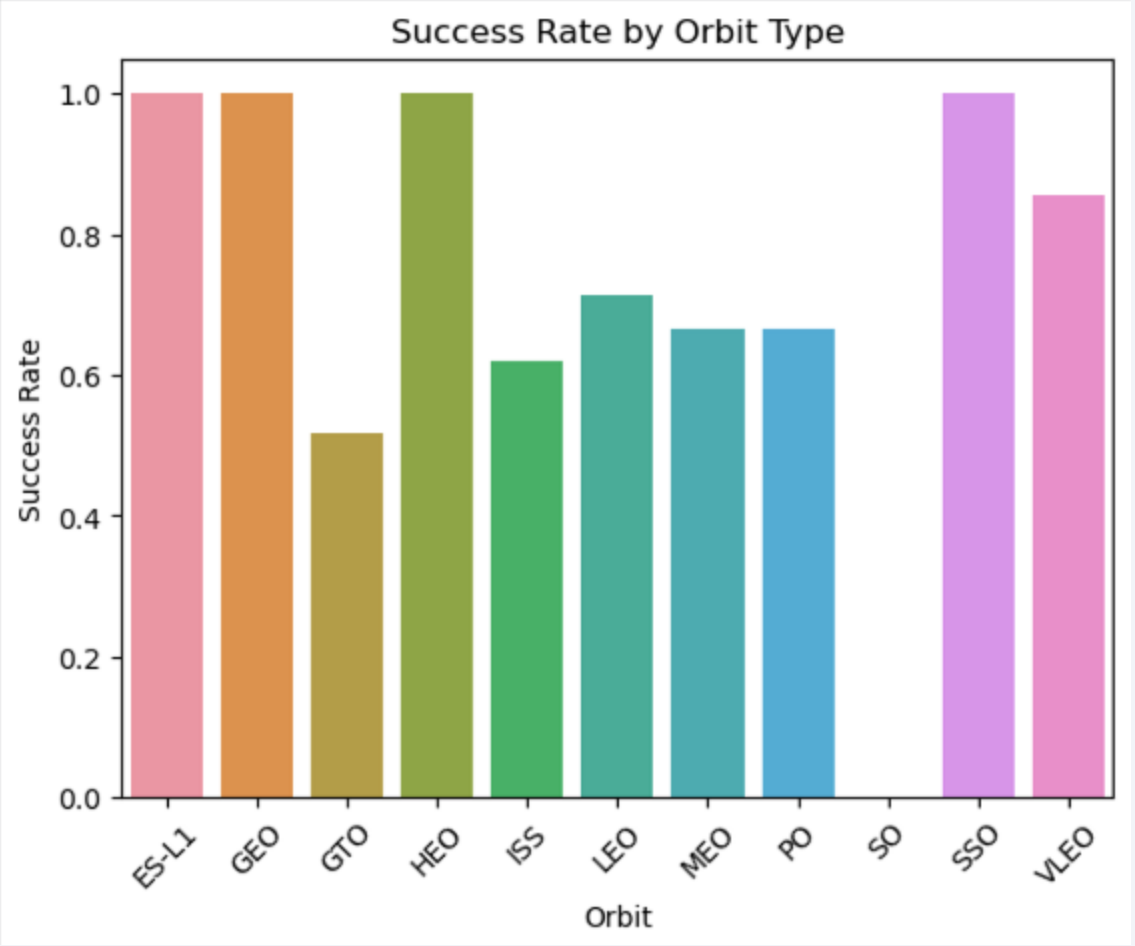
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site.
- Launch Site VAFB SLC 4E do not have launch that have pay load mass more than 10,000 kg.



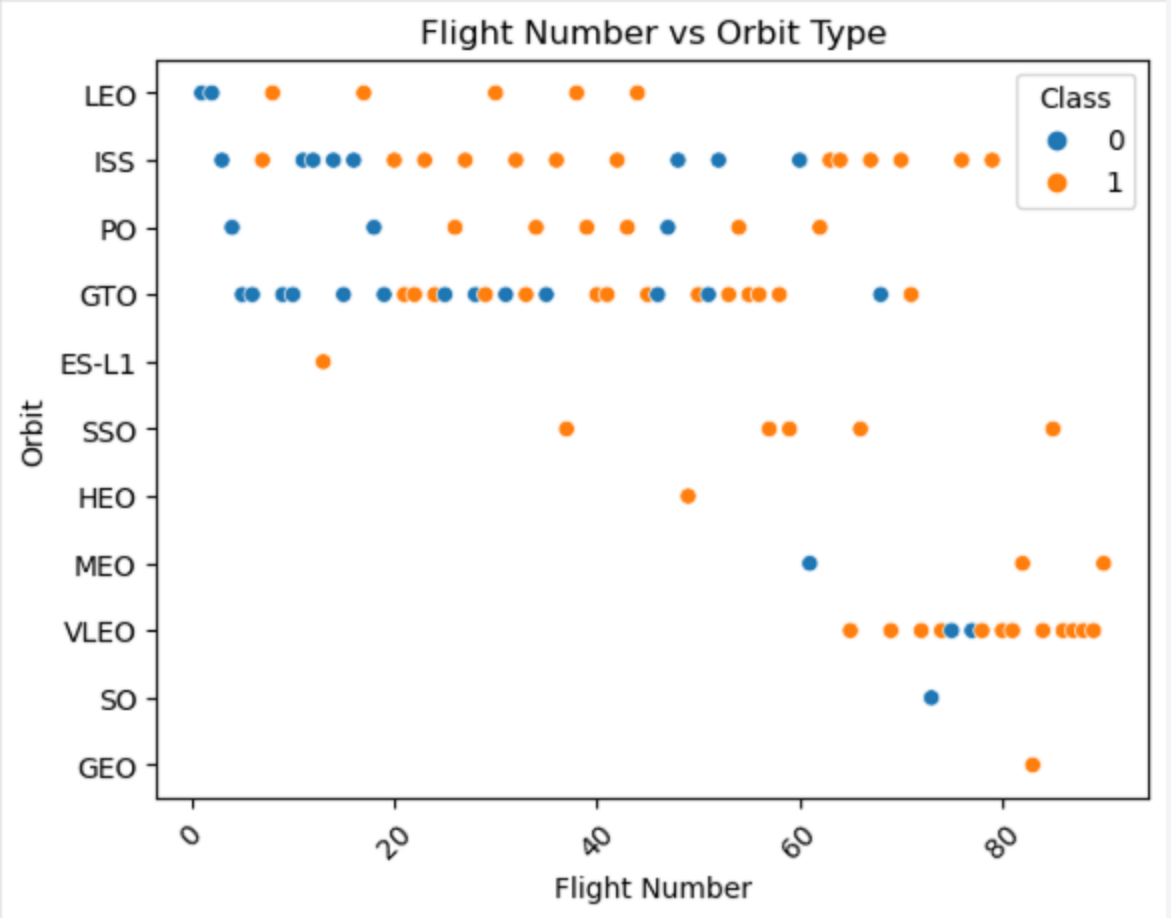
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- Orbit ES-L1, GEO, HEO, SSO have the best success rate.



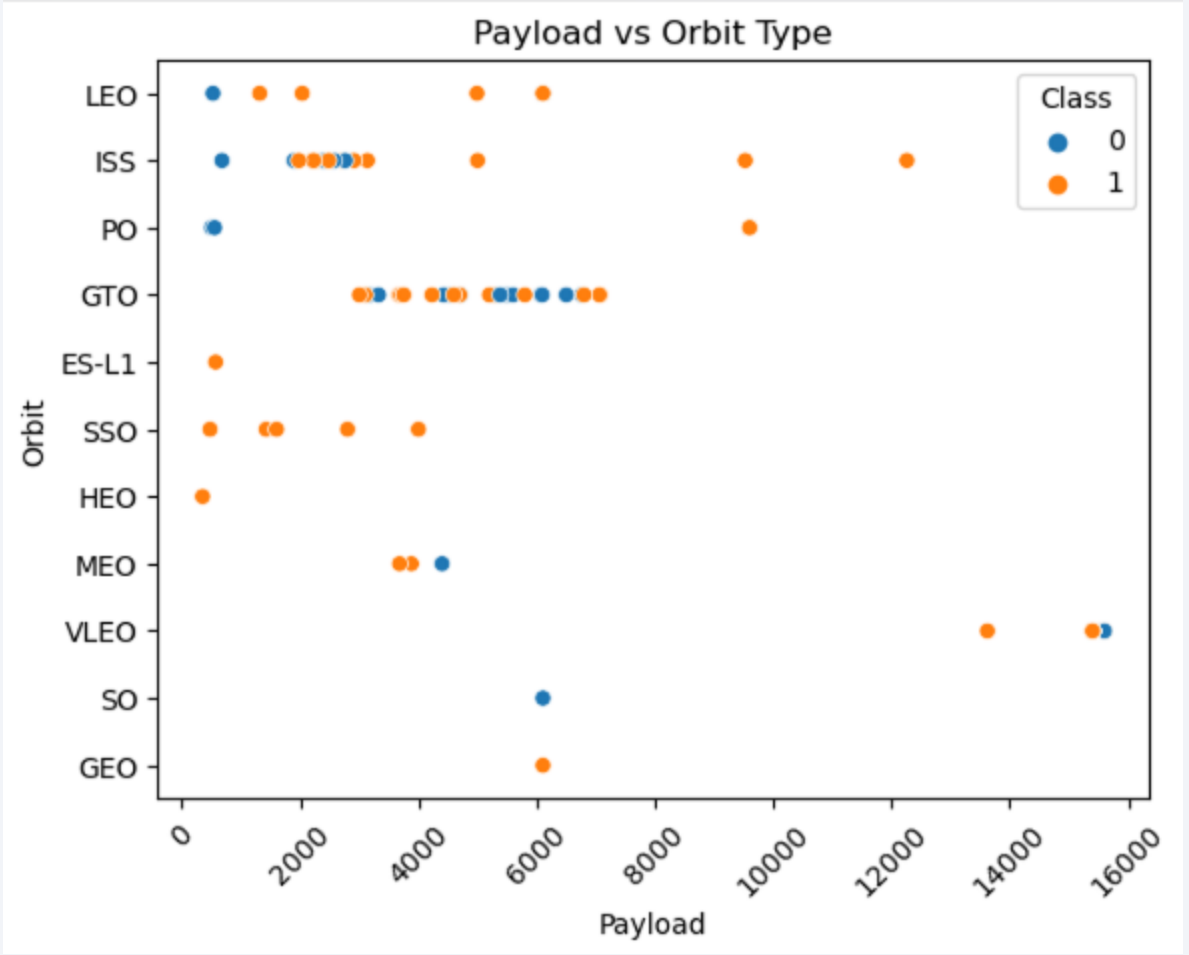
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type.
- Latest flight number usually have launch to VLEO orbit.



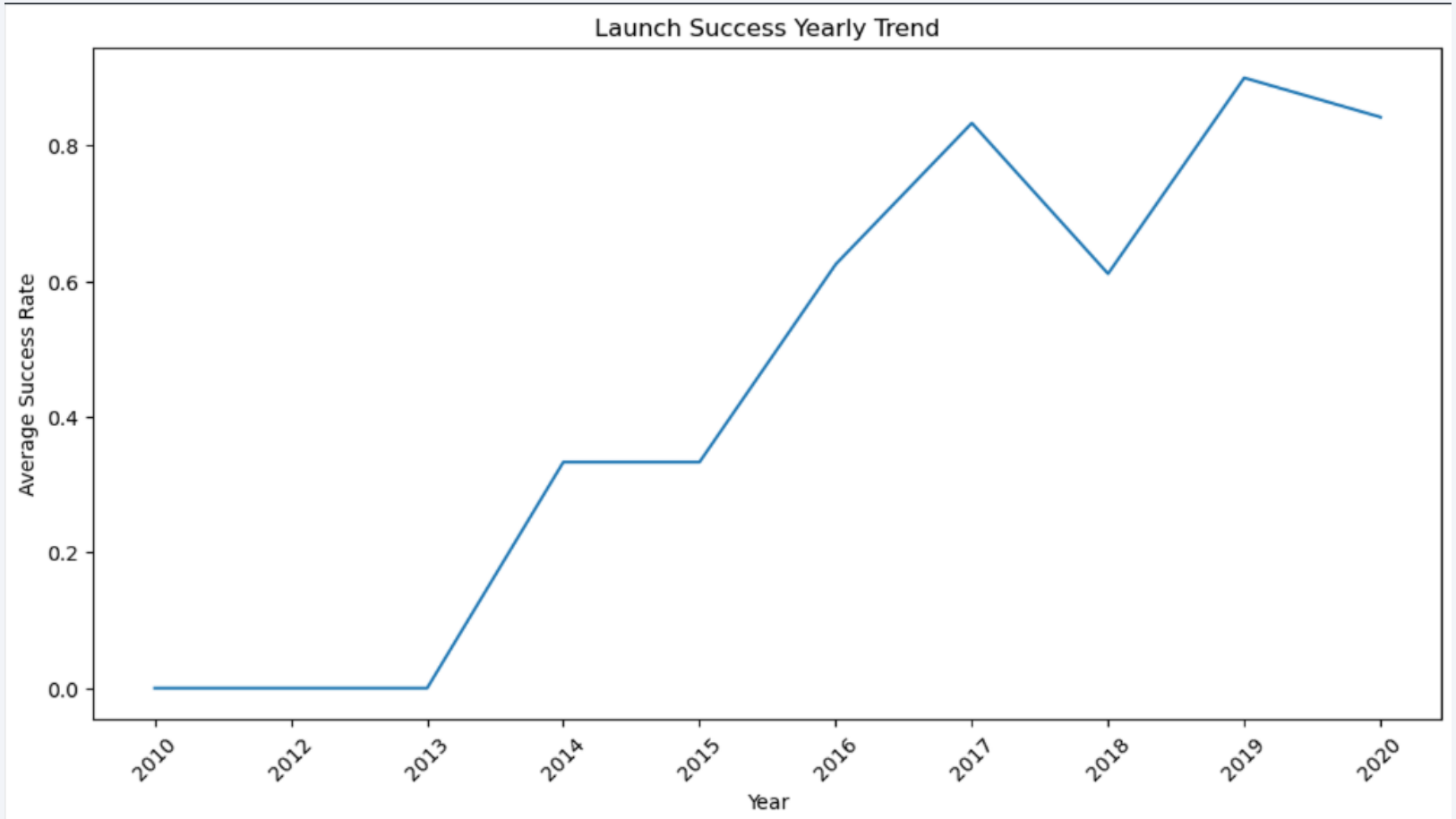
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type.
- VLEO orbit has the heaviest pay load mass launch.



Launch Success Yearly Trend

- Line chart of yearly average success rate.



- The success rate tend to improve as the year pass.

All Launch Site Names

- Names of the unique launch sites.

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

- Space X have 4 launch sites name above.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

- The above are 5 records where launch sites begin with `CCA`.

Total Payload Mass

- Total payload carried by boosters from NASA.

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)';
```

✓ 0.0s Python

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

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload carried by boosters from NASA is 45,596 kg.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1.

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

✓ 0.0s Python

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad.

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';
```

✓ 0.0s Python

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

MIN(Date)
2015-12-22

- The first successful landing outcome on ground pad was in 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

✓ 0.0s Python

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

Booster_Version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- Total of 7 boosters named above have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes.

```
%sql SELECT "Mission_Outcome", COUNT(*) as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

✓ 0.0s Python

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

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

- Most of the mission are successful.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass.

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

✓ 0.0s Python

```
 * 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

- There are 12 total booster which have carried the maximum payload mass.

2015 Launch Records

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

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- There are 2 failed landing outcomes in drone ship in year 2015.

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

- 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.

```
%sql SELECT "Landing_Outcome", COUNT(*) as Outcome_Count FROM SPACEXTBL WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome
```

✓ 0.0s Python

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

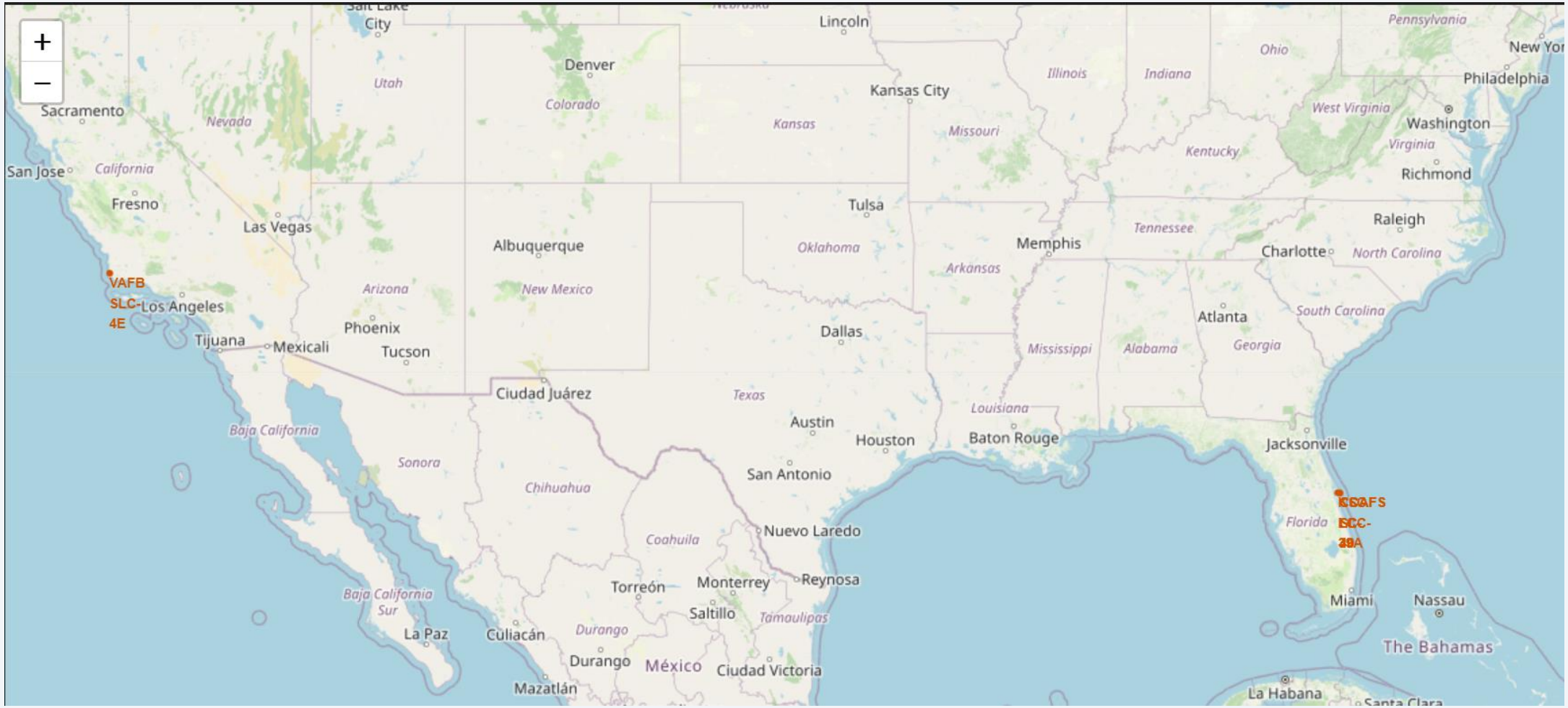
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- There are 8 landing outcome between the date 2010-06-04 and 2017-03-20, where the most common outcome is “No attempt”.


Section 3
Launch Sites
Proximities Analysis

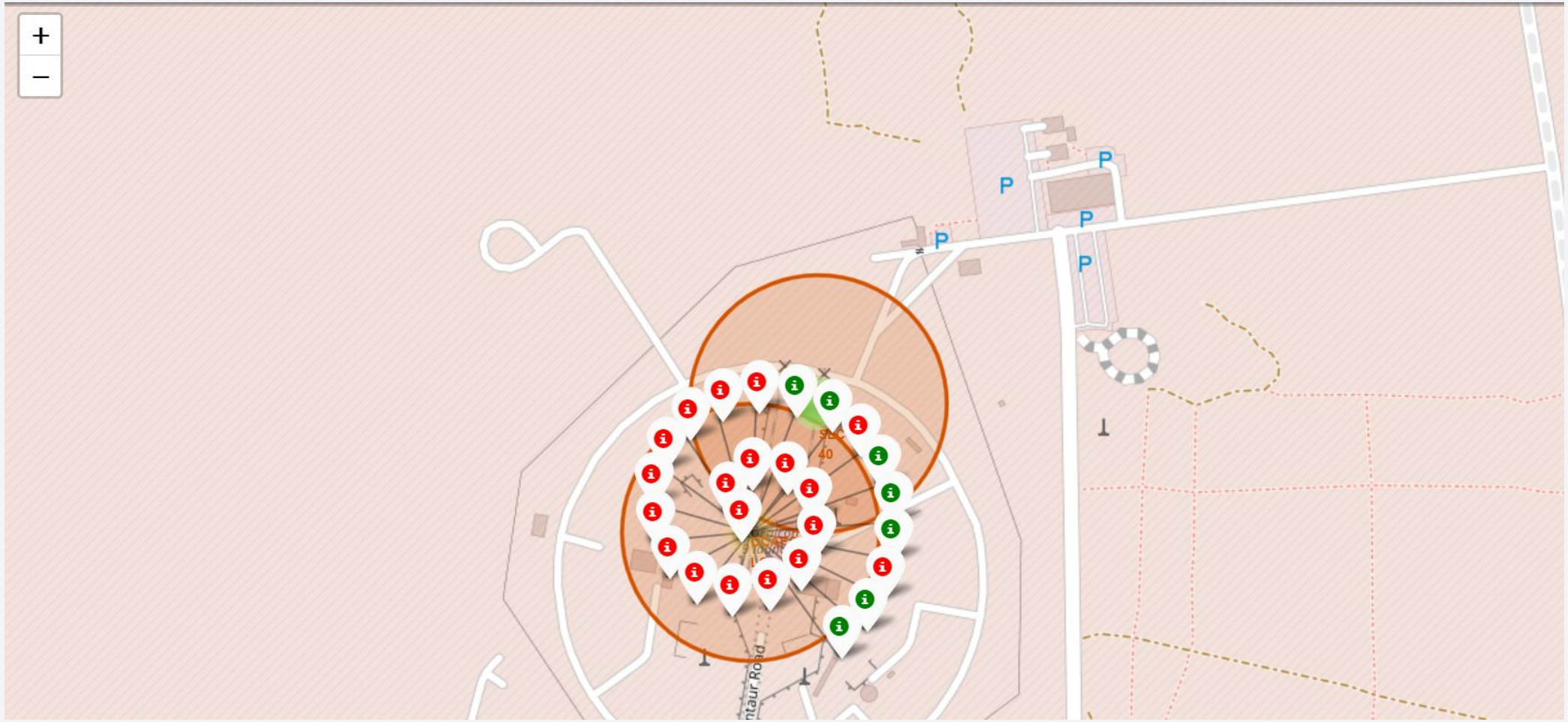
Folium Map of all Launches Sites

- Create a folium map to mark all launch sites location with circle and marker label.



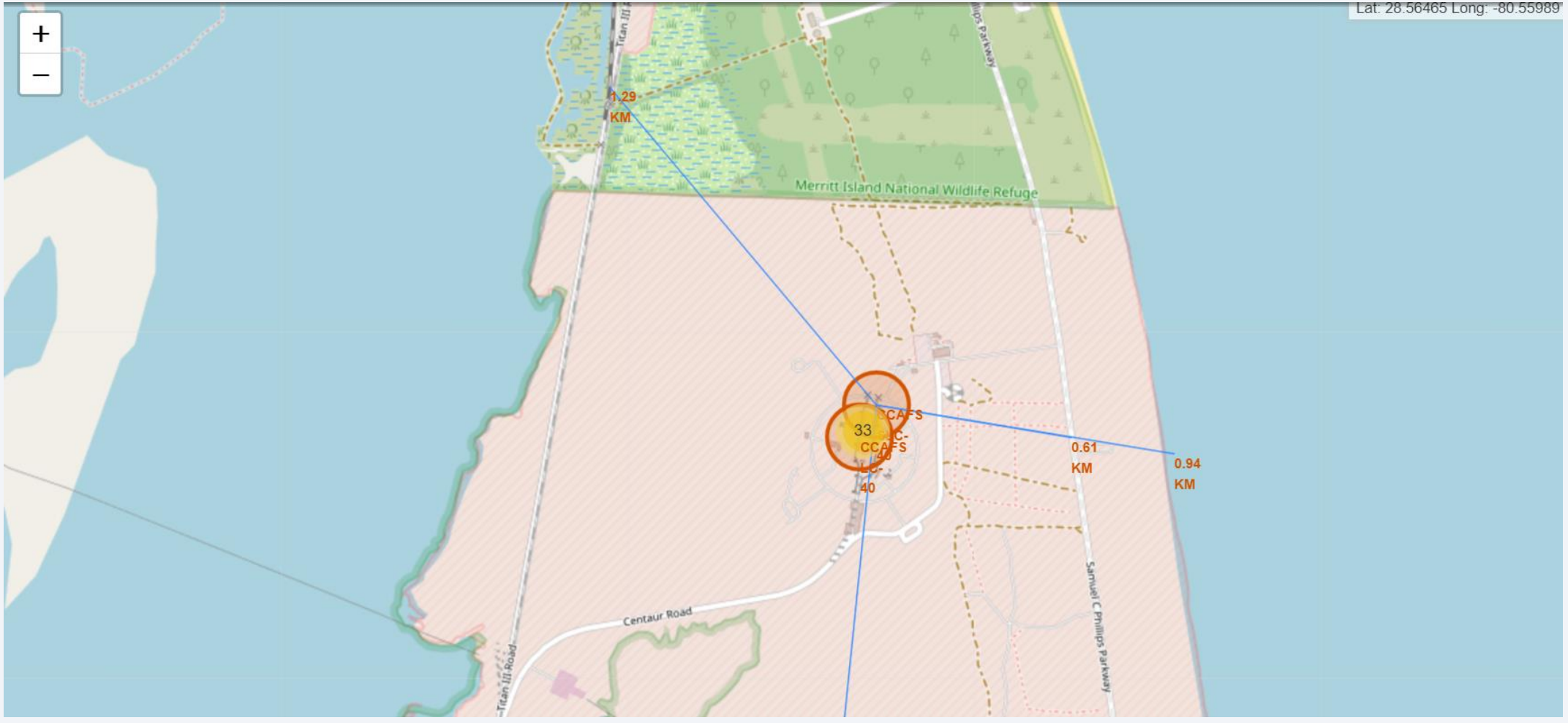
Folium Map of success/failed launches

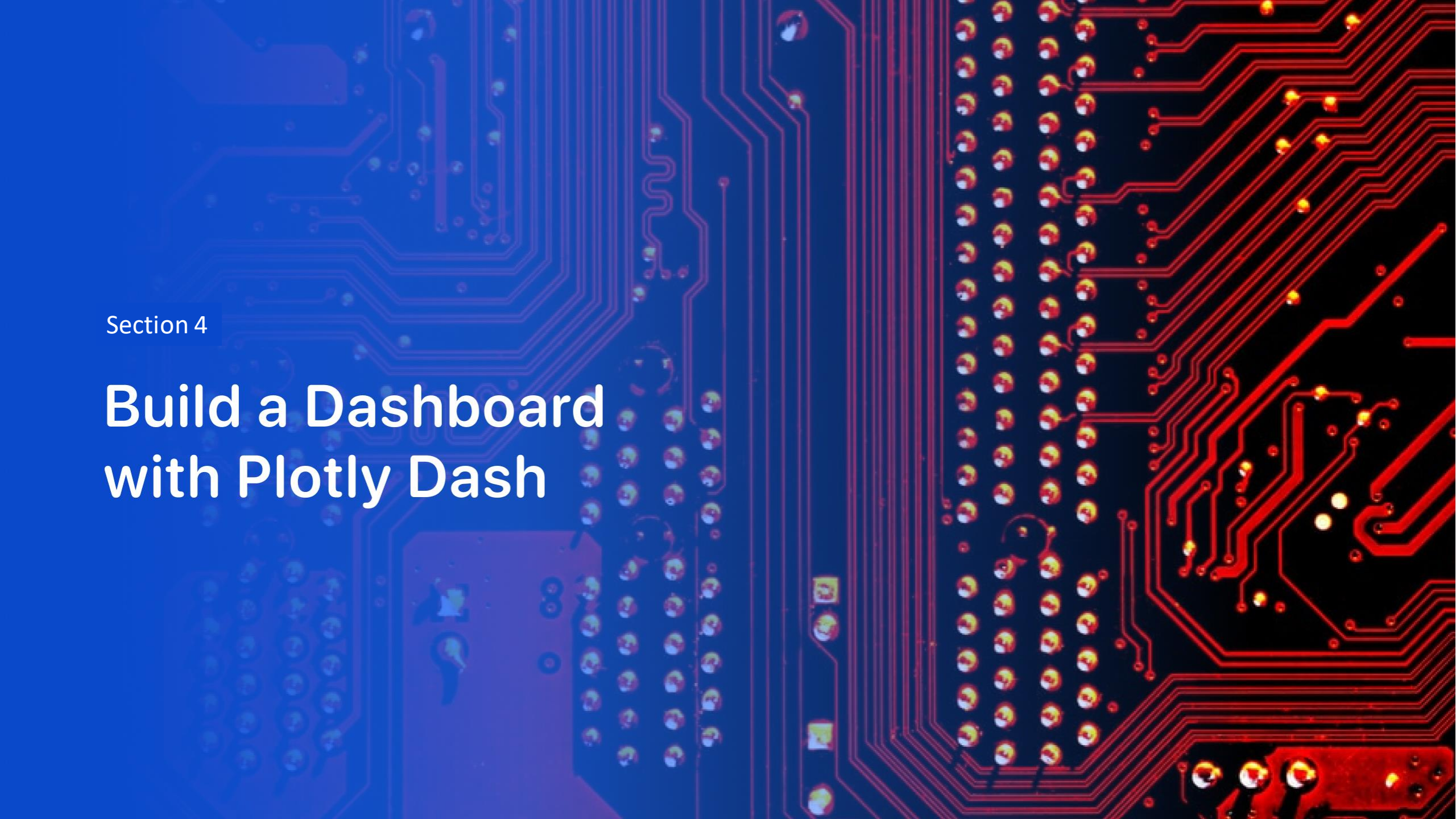
- Mark the success & failed launches for each site on the map using marker cluster objects.



Map with the distances between a launch site to its proximities

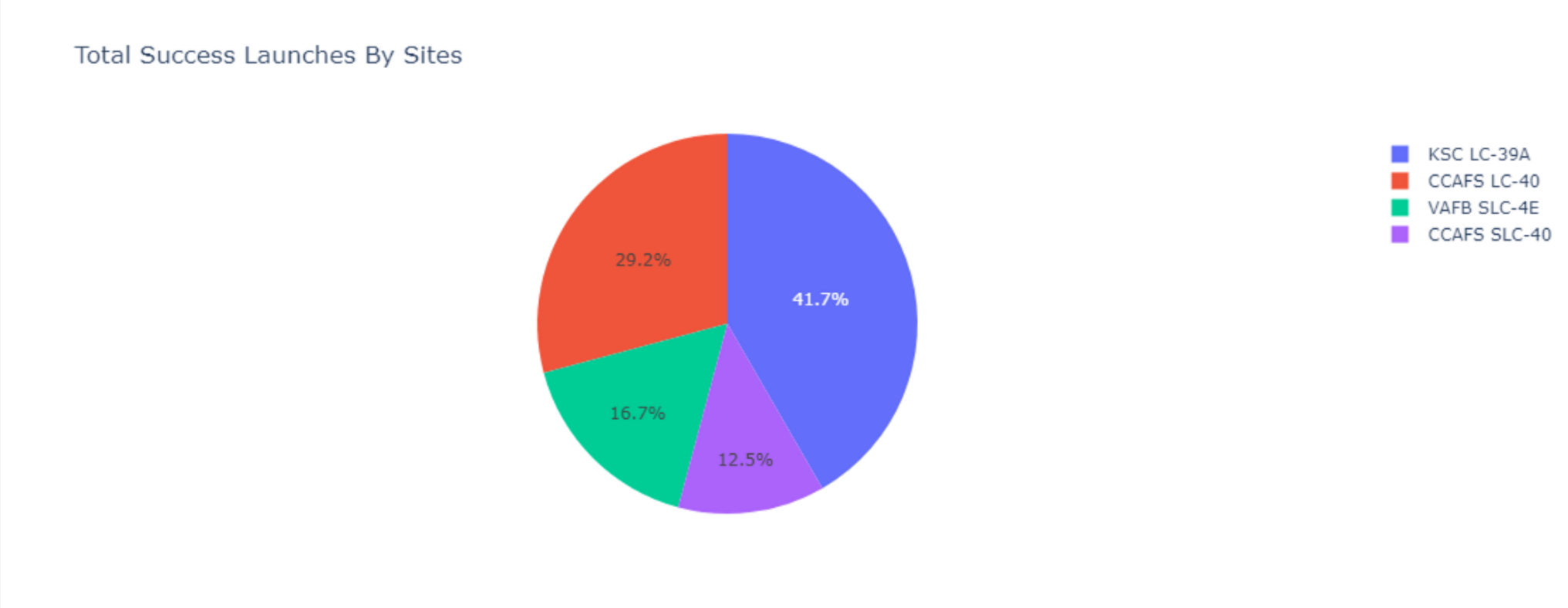
- Calculate the distances between a launch site to its proximities, then mark on a map using marker and polyline objects.




Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Sites

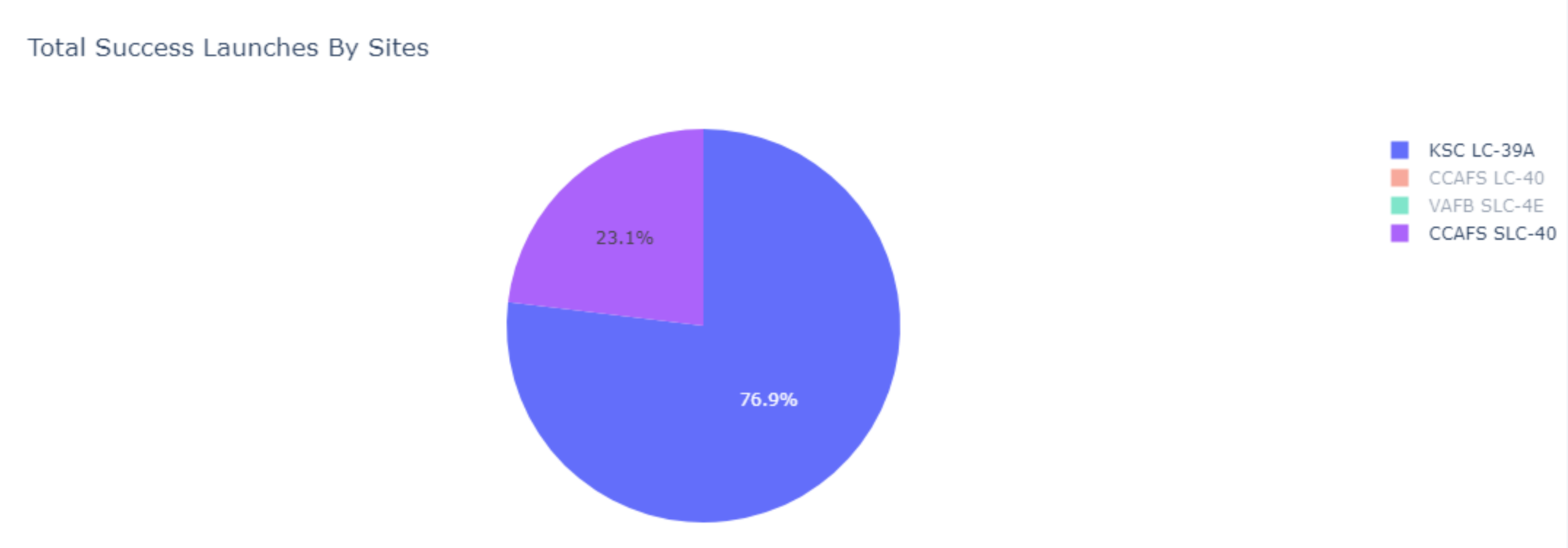
- Dashboard indicate launch success count for all sites in a pie chart.



- Launch site KSC LC-39A has the most success launch.

Launch Site with highest Launch Success ratio

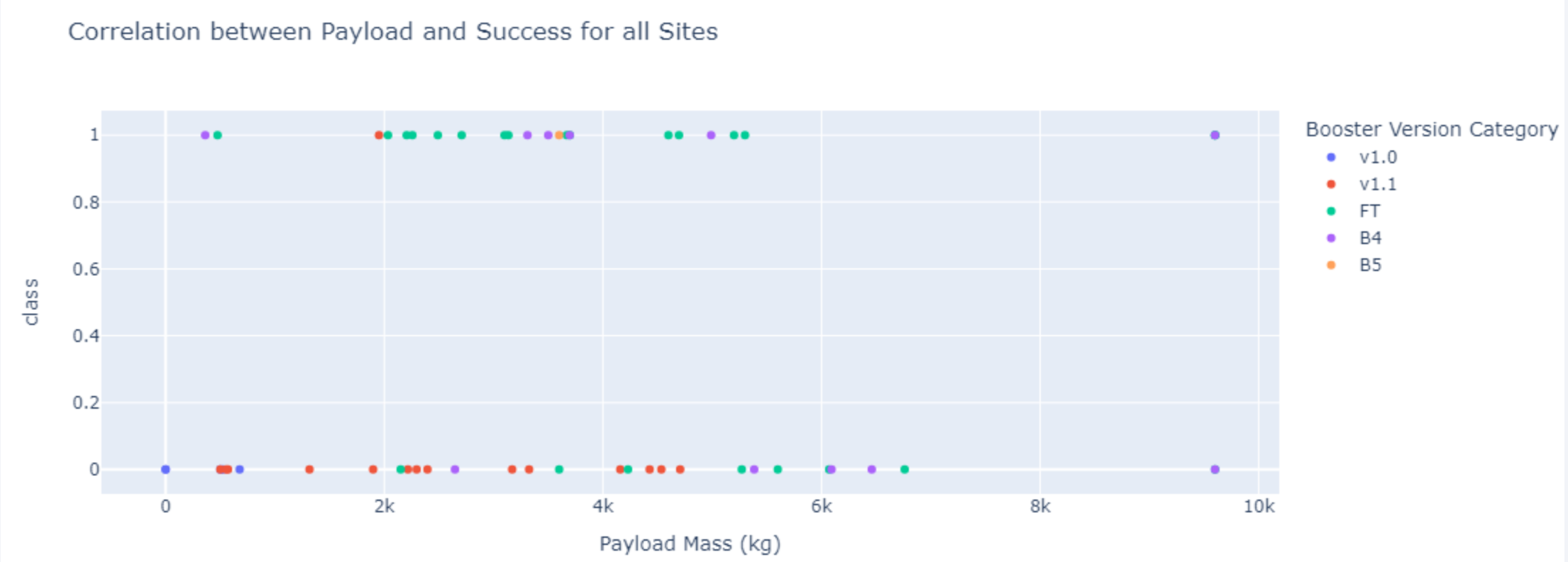
- Dashboard indicate the launch site with highest launch success ratio in pie chart.



- Launch site KSC LC-39A has the highest launch success ratio of 76.9%.

Correlation between Payload and Success for all Sites

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider.



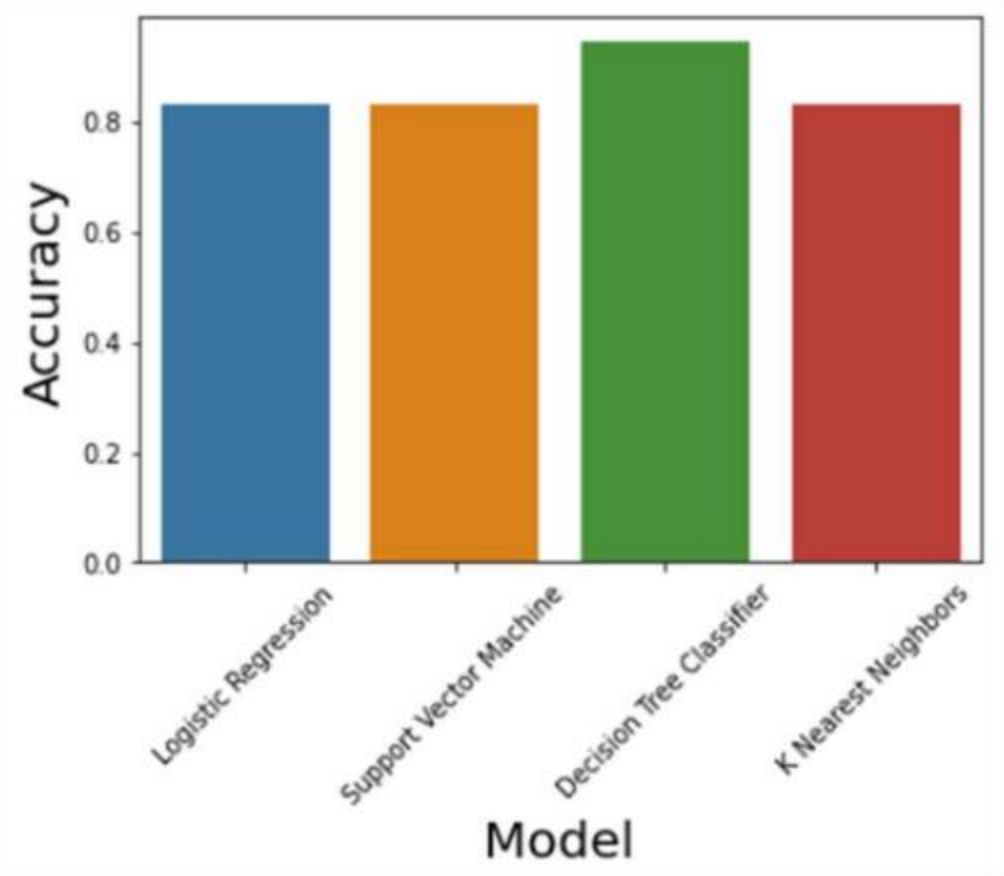
- Booster version FT has the highest success rate in payload range 2,000 – 6,000 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

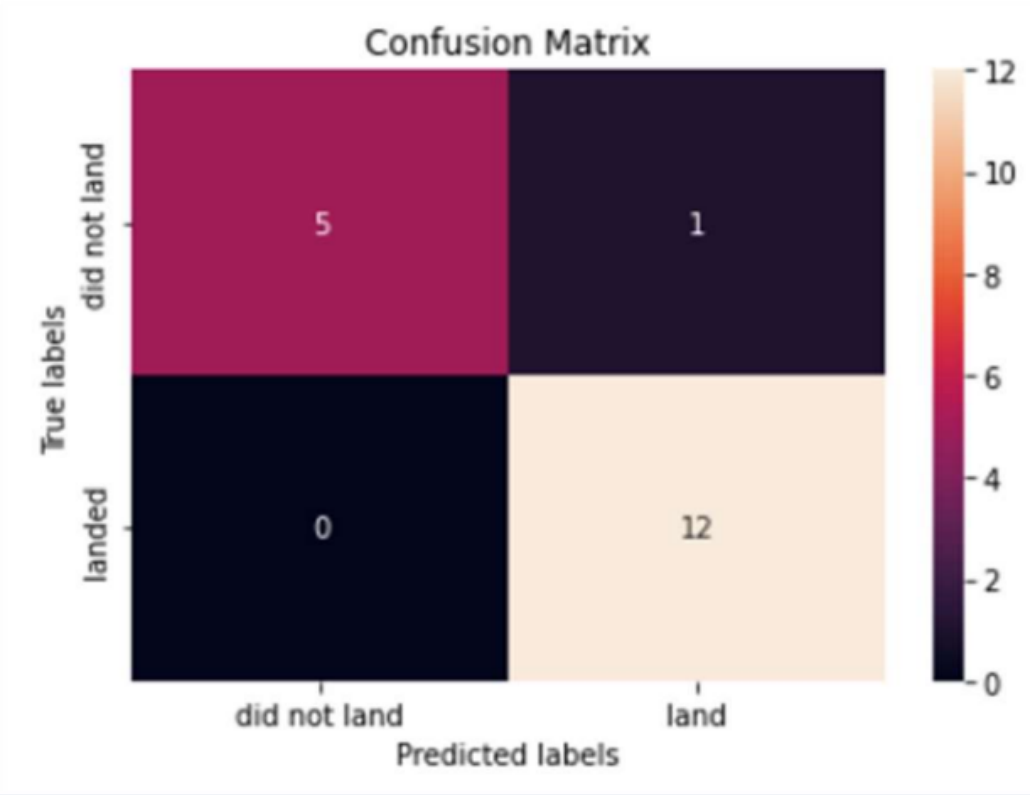
- Visualize the built model accuracy for all built classification models, in a bar chart.



- Decision Tree Classifier model has the highest classification accuracy.

Confusion Matrix

- Confusion matrix of the best performing model (Decision Tree Classifier).



- From the confusion matrix, we see that Decision Tree Classifier can distinguish between the different classes. However, the major problem is false positives.

Conclusions

EDA with SQL and Visualization

- Orbit ES-L1, GEO, HEO, SSO have the best success rate.
- The success rate tend to improve as the year pass.
- The average payload mass carried by booster version F9 v1.1 is 2,534 kg.
- There are 12 total booster which have carried the maximum payload mass.

Folium Map & Dashboard

- Calculate the distances between a launch site to its proximities .
- Launch site KSC LC-39A has the highest launch success ratio of 76.9%.

Predictive Analysis

- The best classification model is Decision Tree Classifier model with accuracy of 94.44%.
- From the confusion matrix, Decision Tree Classifier can distinguish between the different classes, but have false positives as a major problem.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project.
- GitHub URL for the IBM Data Science Capstone Project:

https://github.com/ayanes1991/IBM-Data-Science-Capstone-Project

Thank you!