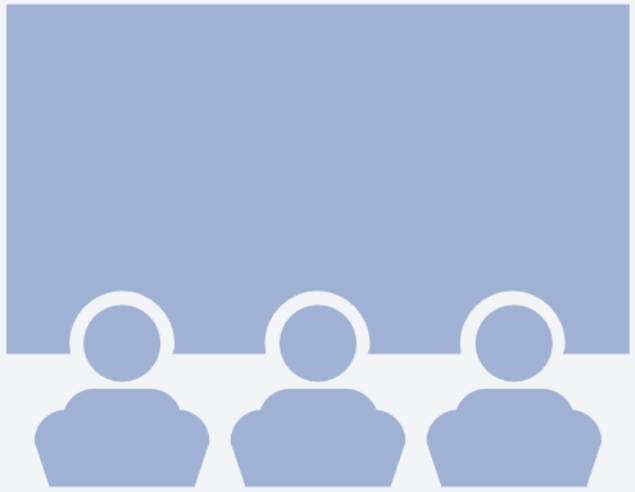
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Maina Duseja

12th Sep 2023

Outline



- Executive Summary
- Introduction
- Methodology
- Data Collection
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Insights drawn from EDA
  - Launch Sites Proximities analysis
  - Dashboard with Plotly Dash
  - Predictive Analysis(Classification)
- 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 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, 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 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.

- **Problems you want to find answers**

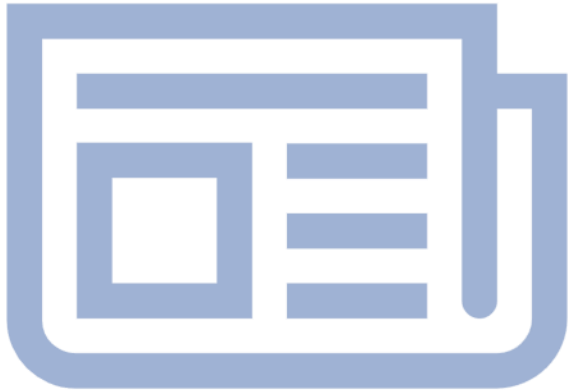
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary



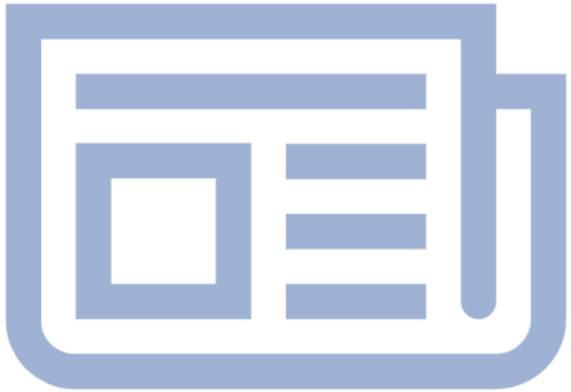
1) Data collection methodology :

- Data from SpaceX was obtained from two sources:
  - SpaceX API -> https://api.spacexdata.com/v4/ (having endpoints /capsules and /cores and launches/past)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

To be continued.

Methodology

Executive Summary



2) Perform data wrangling :

- One-hot encoding was applied to categorical features
- Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.

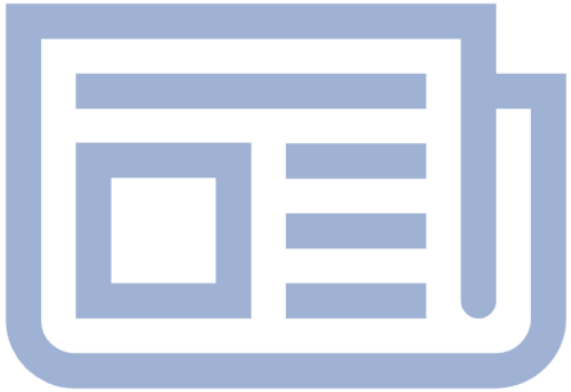
3) Perform exploratory data analysis (EDA) using visualization and SQL

4) Perform interactive visual analytics using Folium and Plotly Dash

To be continued.

Methodology

Executive Summary



5) Perform predictive analysis using classification models :

- Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

- **Describe how data sets were collected.**

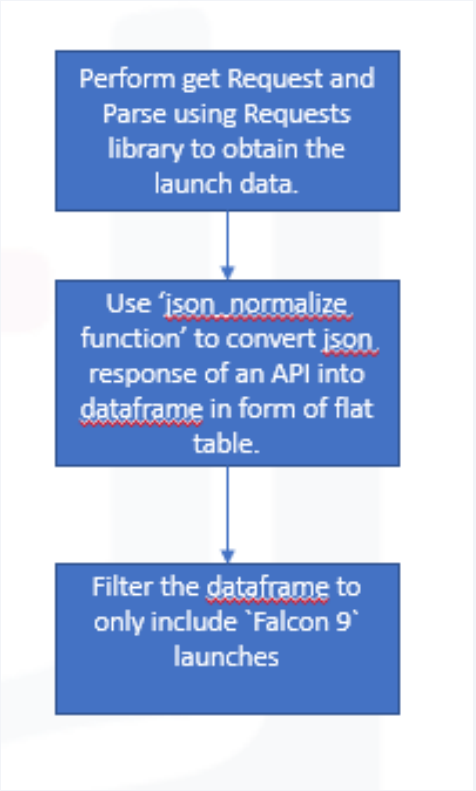
  Data from SpaceX was obtained from two sources:

  - SpaceX API -> https://api.spacexdata.com/v4/ (having endpoints /capsules and /cores and launches/past)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



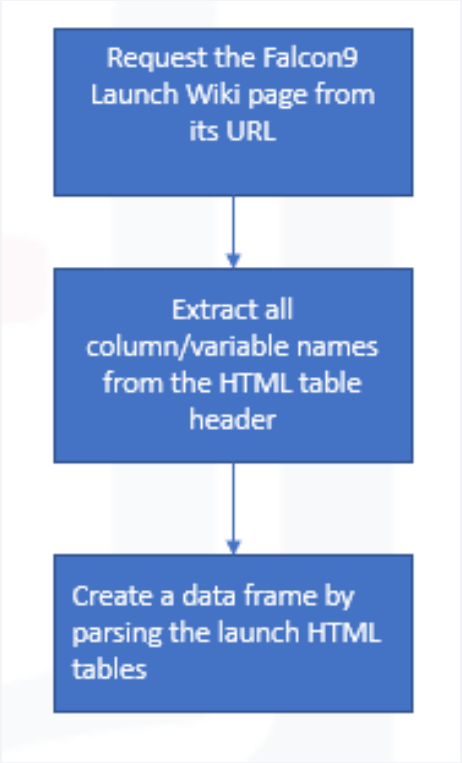
Data Collection - Scraping

Data from SpaceX launches can also be obtained from Wikipedia;

- Data are downloaded from Wikipedia according to the flowchart and then persisted.

Source code Reference:

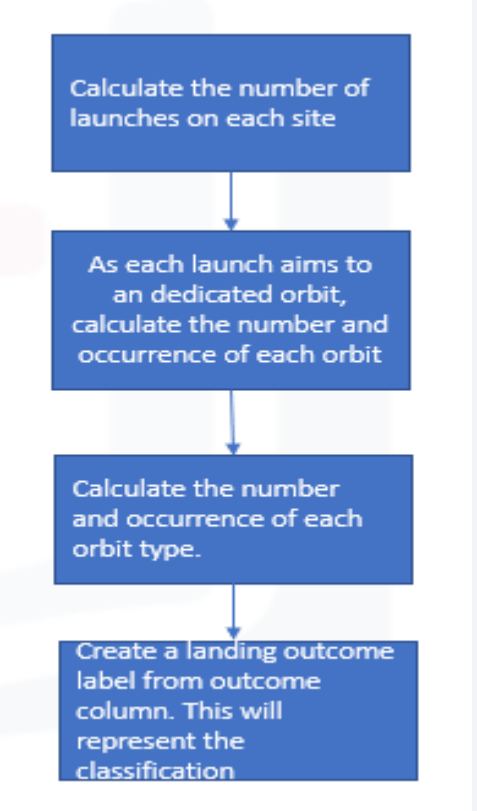
https://github.com/maina-duseja/DS-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Initially Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- In this we performed exploratory Data analysis and Feature engineering using Pandas and Matplotlib.
- To explore data, scatterplots and bar plots were used to visualize the relationship between following features:
  - Payload Mass v/s Flight Number, Launch Site v/s Flight Number, Launch Site v/s Payload Mass, Orbit v/s Flight Number, Payload v/s Orbit type.
- Visualize the relationship between success rate of each orbit type and launch success yearly trend.
- Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begins with the string 'CCA';
  - Total pay load mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing out comes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- With EDA visualization we discovered some preliminary correlations between the launch sites and success rates. With folium we can perform more visual analytics on geospatial data such as launch sites and its proximities.
- 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 to visualize those locations by pinning them on map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success to enhance the launch outcome for each site.
- 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 to see if we can easily find any railway, coastline or highway etc. .
- Source code Reference:
  https://github.com/maina-duseja/DS-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

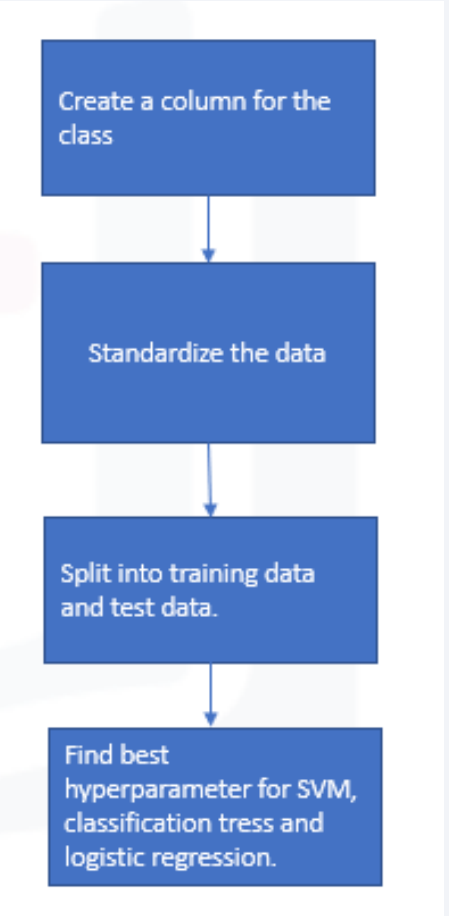
Build a Dashboard with Plotly Dash

- 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.
- These are added to allow users to perform interactive visual analytics on spaceX launch data in real-time and to gain insights such as largest successful launches with highest success rates with F9 booster version.
- Source code Reference:
  https://github.com/maina-duseja/DS-Capstone/blob/main/spacex_dash_app.py

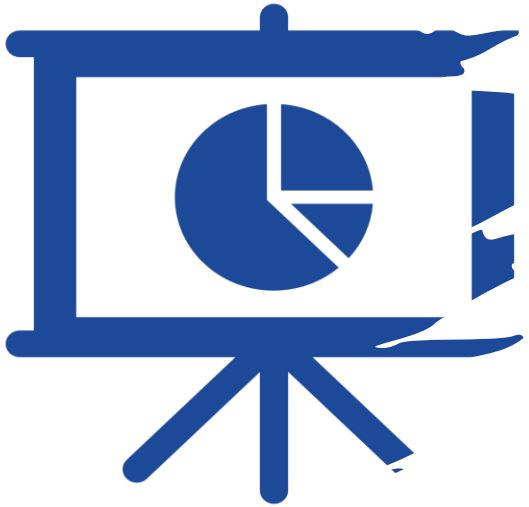
Predictive Analysis (Classification)

- 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.
- Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



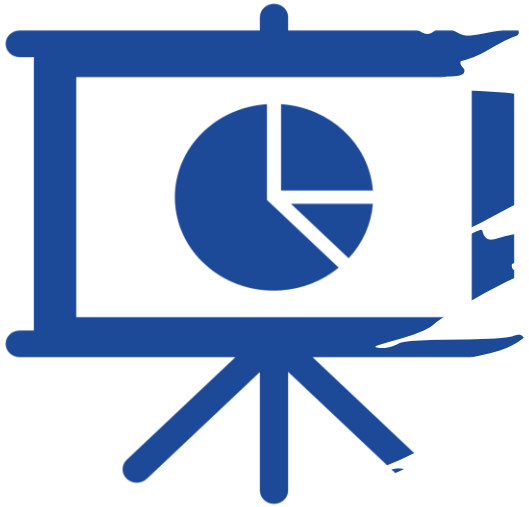
Results



Exploratory data analysis results

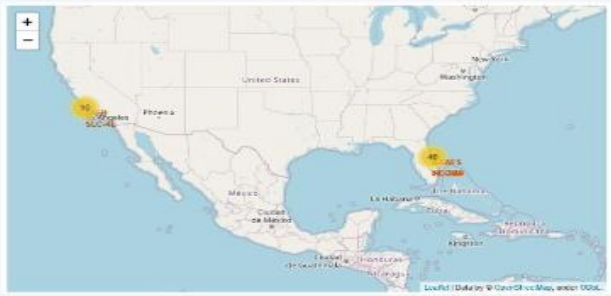
- **CCAFS LC-40**, **VAFB SLC-4E**, **KSC LC-39A** and **CCAFS SLC-40** are launch sites in space mission.
- Total payload mass carried by boosters launched by Nasa(CRS) is **45596 KG**.
- Avg. payload mass carried by booster version **F9 v1.1** is **~ 2534 KG**.
- The first successful landing outcome in group pad was achieved on **22nd December 2015**.
- Total number of successful outcome is 100 and failure outcome is 1
- As the flight number increases, the first stage is likely to land successfully.
- The more massive the payload, the less likely the first stage will return.
- **CCAFS LC-40**, has a success rate of **60%**, while **KSC LC-39A** and **VAFB SLC 4E** has a success rate of **77%**.
- **VAFB-SLC** launch site there are no rockets launched for heavy payload mass(greater than 10000)
- With heavy payloads the successful landing or positive landing rate are more for **Polar**, **LEO** and **ISS**.
- Success rate since 2013 kept increasing till 2020.

Results

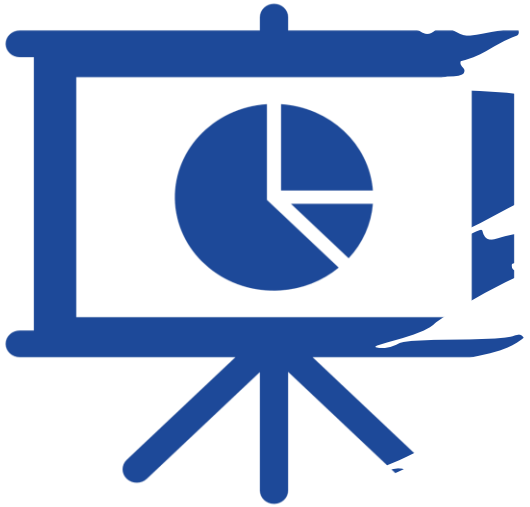


Interactive analytics demo in screenshots

- Launch sites are in close proximity to equator to minimize fuel consumption by using Earth eastward spin to help spaceships get into orbit.
- Launch sites are in close proximity to coastline so they can fly over the ocean during launch, for at least two safety reasons:
    1. Crew has option to abort launch and attempt water landing
    2. Minimize people and property at risk from falling debris.
- Launch sites are in close proximity to highways, which allows for easily transport required for people and property.
- Launch sites are in close proximity to railways, which allows transport for heavy cargo.
- Launch sites are not in close proximity to cities, which minimizes danger to population dense areas.

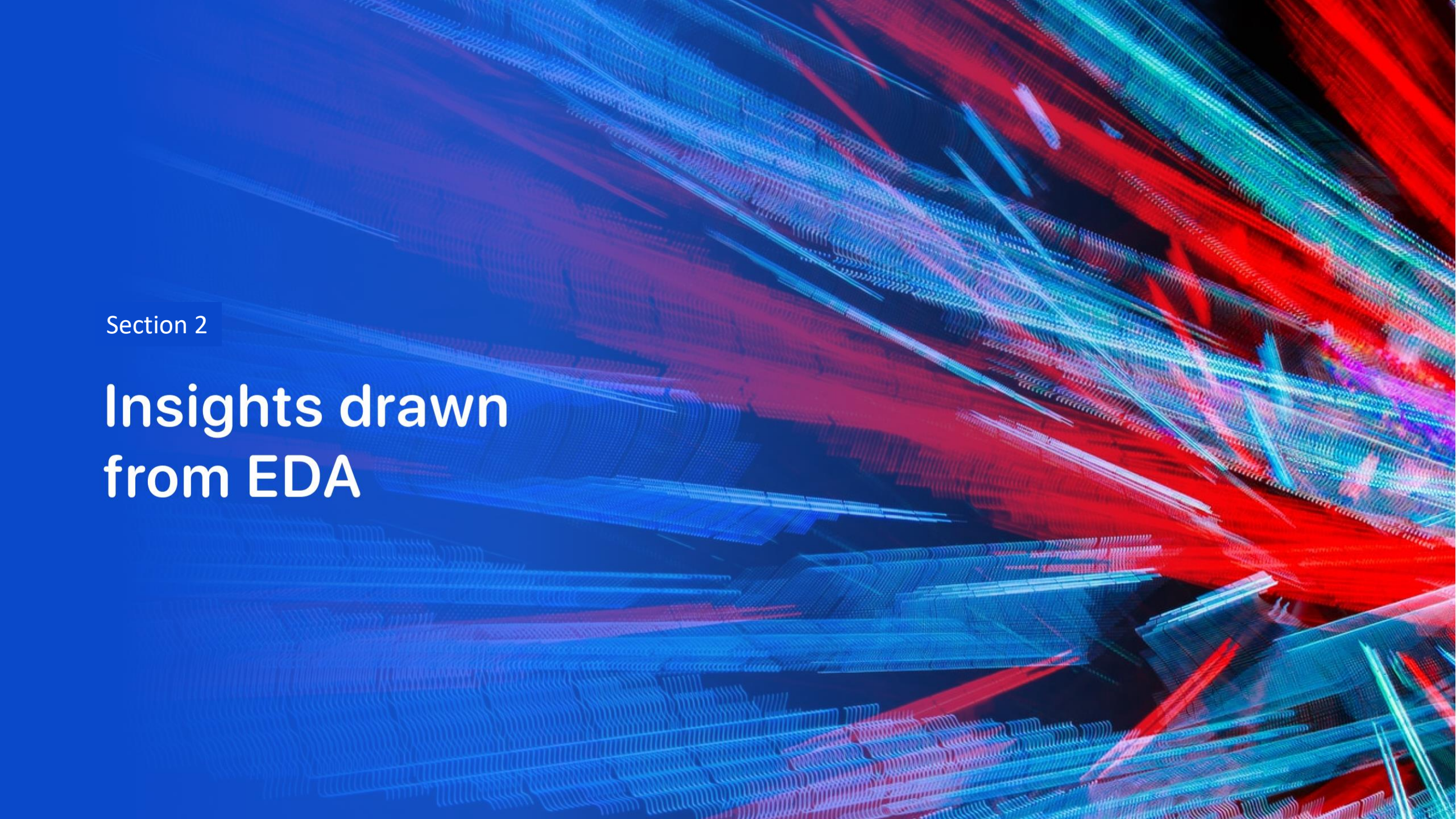


Results



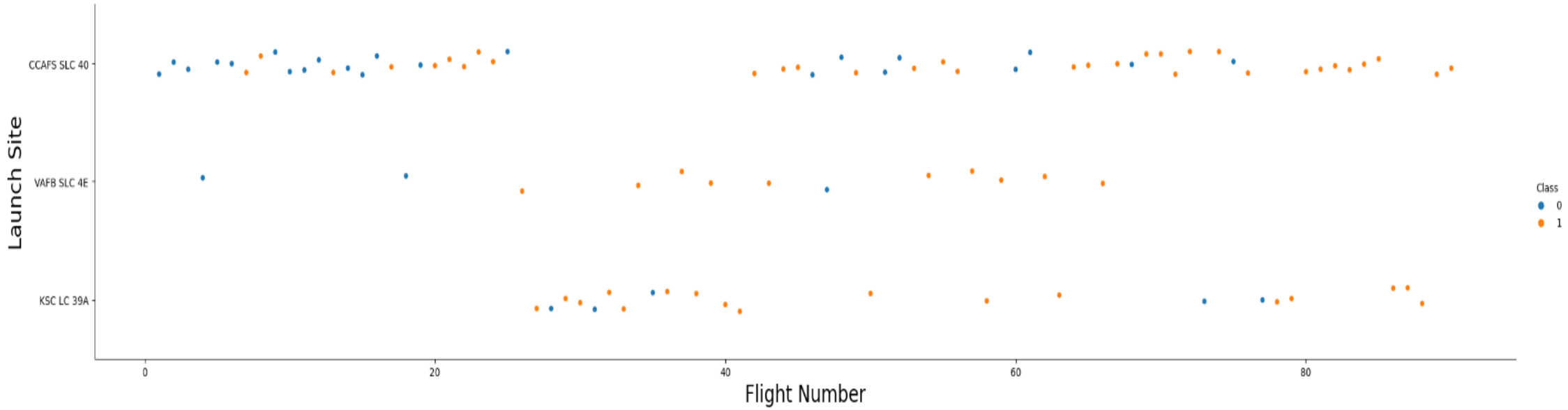
Predictive analysis results

- With FlightNumber, PayloadMass, Orbit and Launch site of Falcon 9 booster version as input we created model with Column Class representing success/failure outcome.
- After comparing accuracy of Logistic Regression, Support Vector machines, Decision Tree Classifier and K-nearest neighbors algorithms we found they all perform the same.


Section 2
Insights drawn from EDA

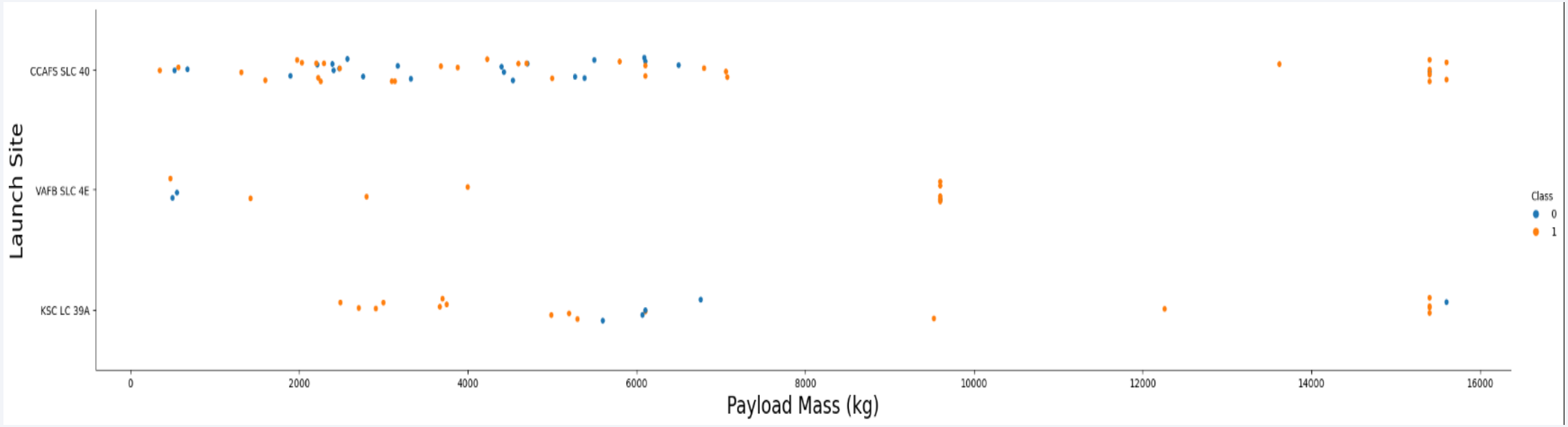
Flight Number vs. Launch Site

- From the scatter plot below, we can say that higher the flight number, the greater the success at launch site.



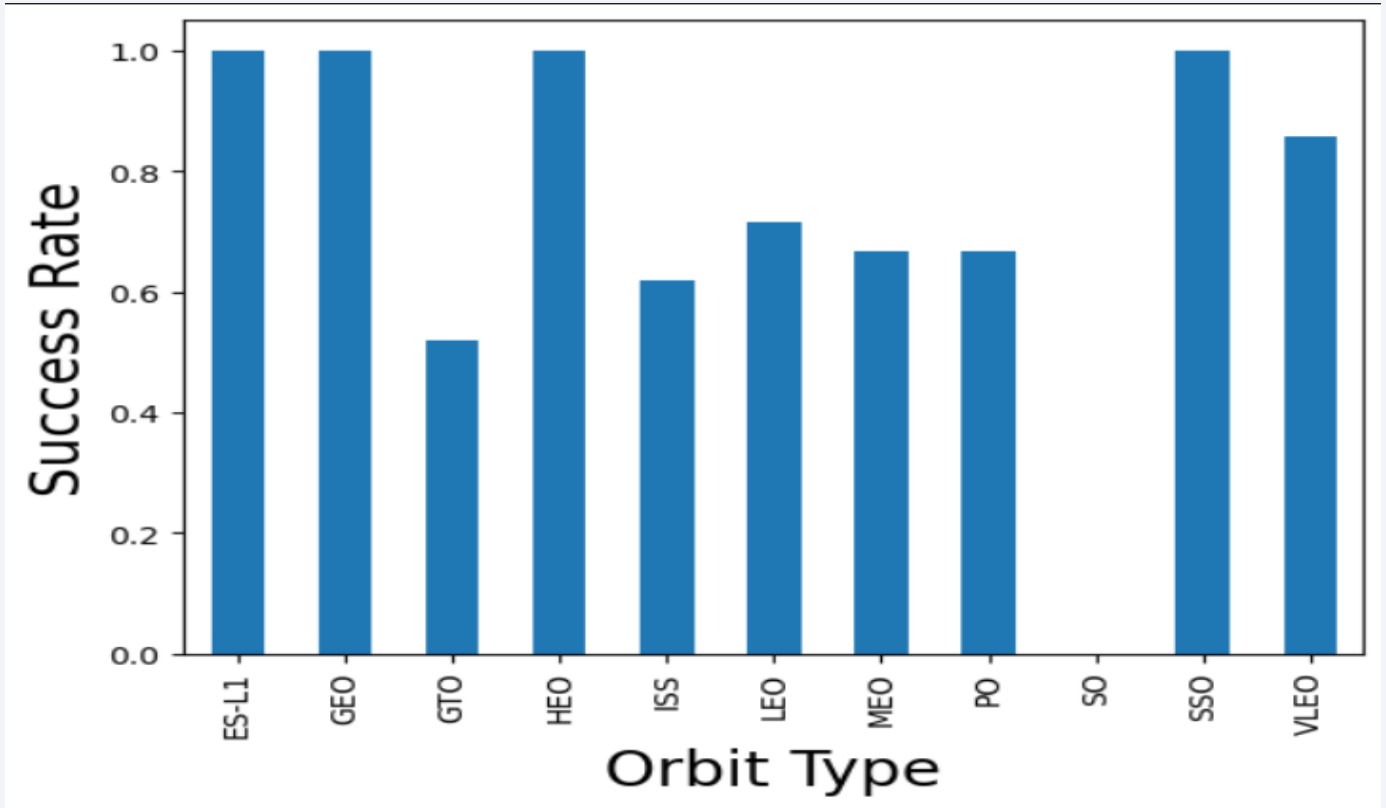
Payload vs. Launch Site

- From the scatter plot below we can say that with heavy payloads the successful landing or positive landing rate are more and for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



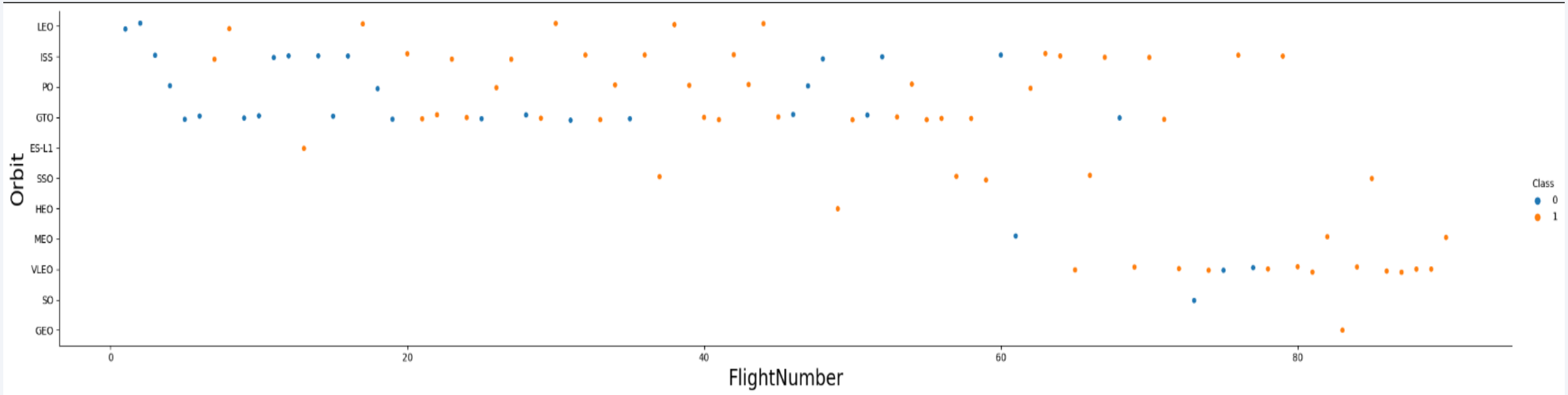
Success Rate vs. Orbit Type

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



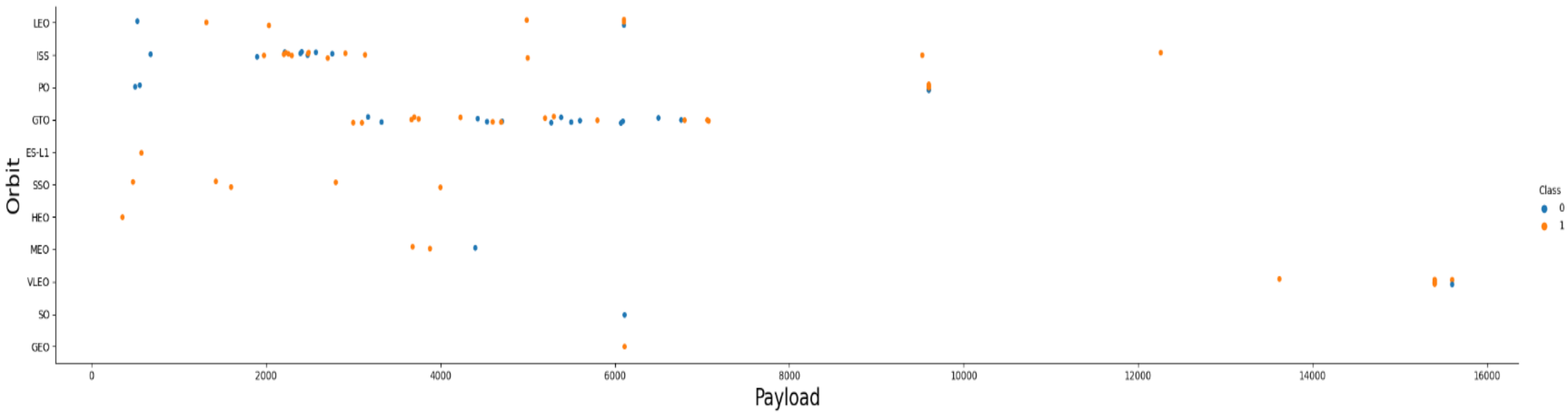
Flight Number vs. Orbit Type

- From the scatter plot below we can see 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.



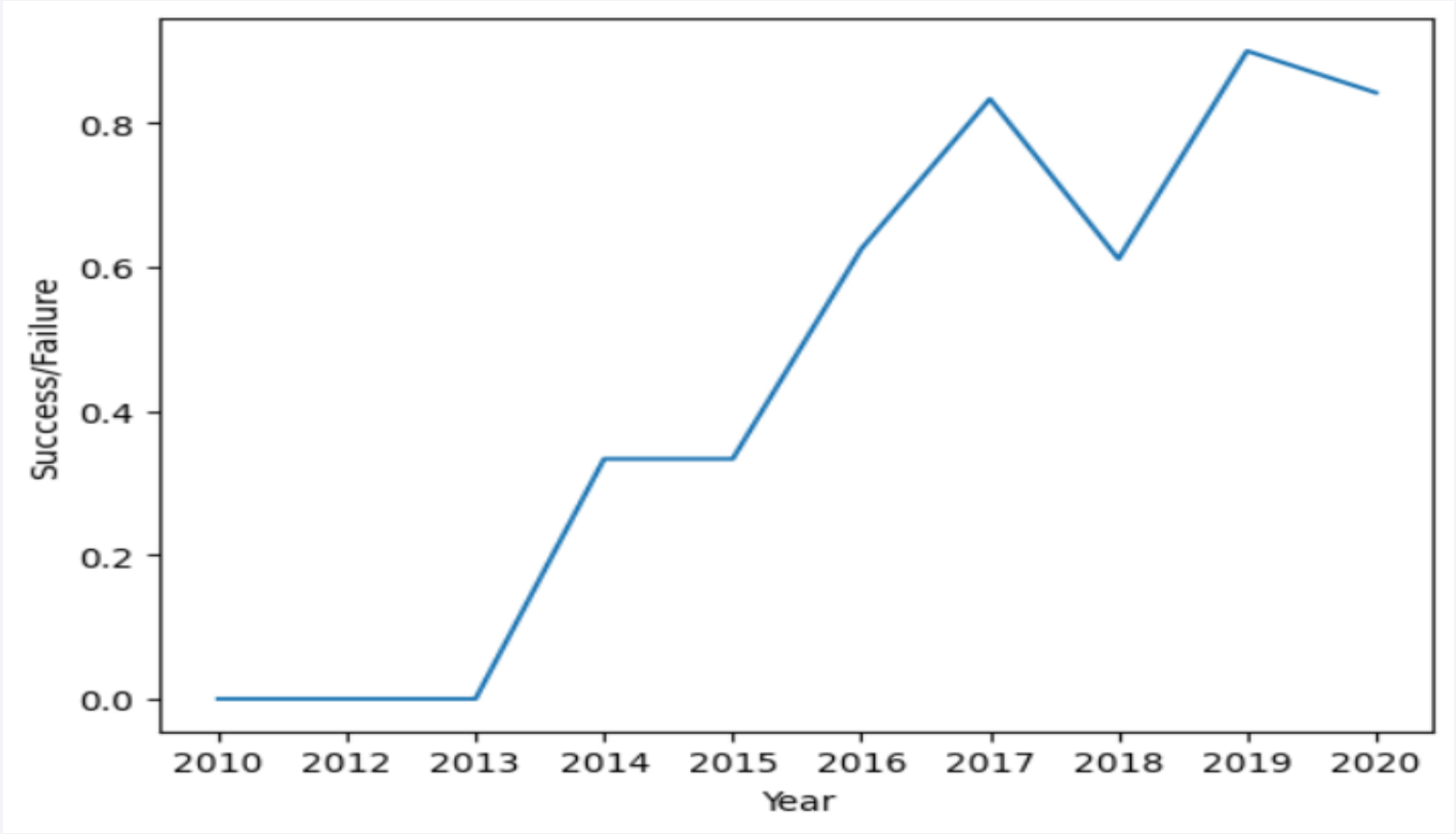
Payload vs. Orbit Type

- 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) both are there.



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- With Distinct keyword on launch site we can get the unique list.

Display the names of the unique launch sites in the space mission

```
[8]
%%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'

- To begin launch sites with `CCA` we have used the like operator with ‘CCA%’ expression. % indicates any letters words after CCA.
- For displaying only 5 records we used Limit as 5 to restrict displaying only 5 records.

Task 2

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

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

[10] Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 used sum function on Payload_Mass-KG_ column to get the total sum.

Task 3

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

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

[13]

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

Total_Payload_Mass_KG_
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 with Avg function on ‘PAYLOAD_MASS__KG_’ column.

Task 4

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

```
[14] %%sql
     SELECT AVG(PAYLOAD_MASS__KG_) as 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

- To get the first successful landing outcome date on ground pad we used min function on Date column.

Task 5

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

Hint:Use min function

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

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

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

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

[23]

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- To get the total number of successful and failure mission outcomes, we used the count function on “Mission_Outcome’ column and done group by on same.

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

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

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 on payload mass column.

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

```
%%sql
SELECT DISTINCT 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

- To get the records for 2015 year, we used strftime method which represents a string representing date and time using date, time or datetime object.

Task 9

List the records which will display the month names, failure landing_ou

Note: SQLLite does not support monthnames. So you need to use substr(Date,

```
[39] %%sql
     SELECT Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE
     FROM SPACEXTBL
     WHERE Landing_Outcome = 'Failure (drone ship)'
         AND strftime('%Y', Date) = '2015';
```

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

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

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.

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY TOTAL_NUMBER DESC
```

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

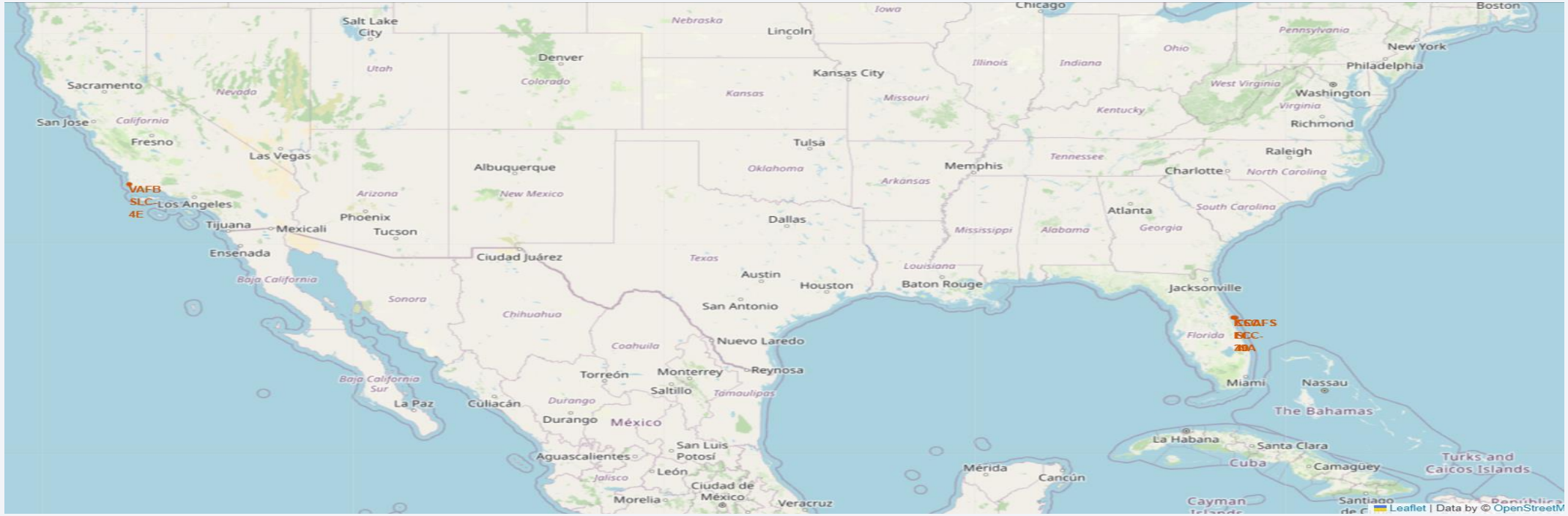
Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

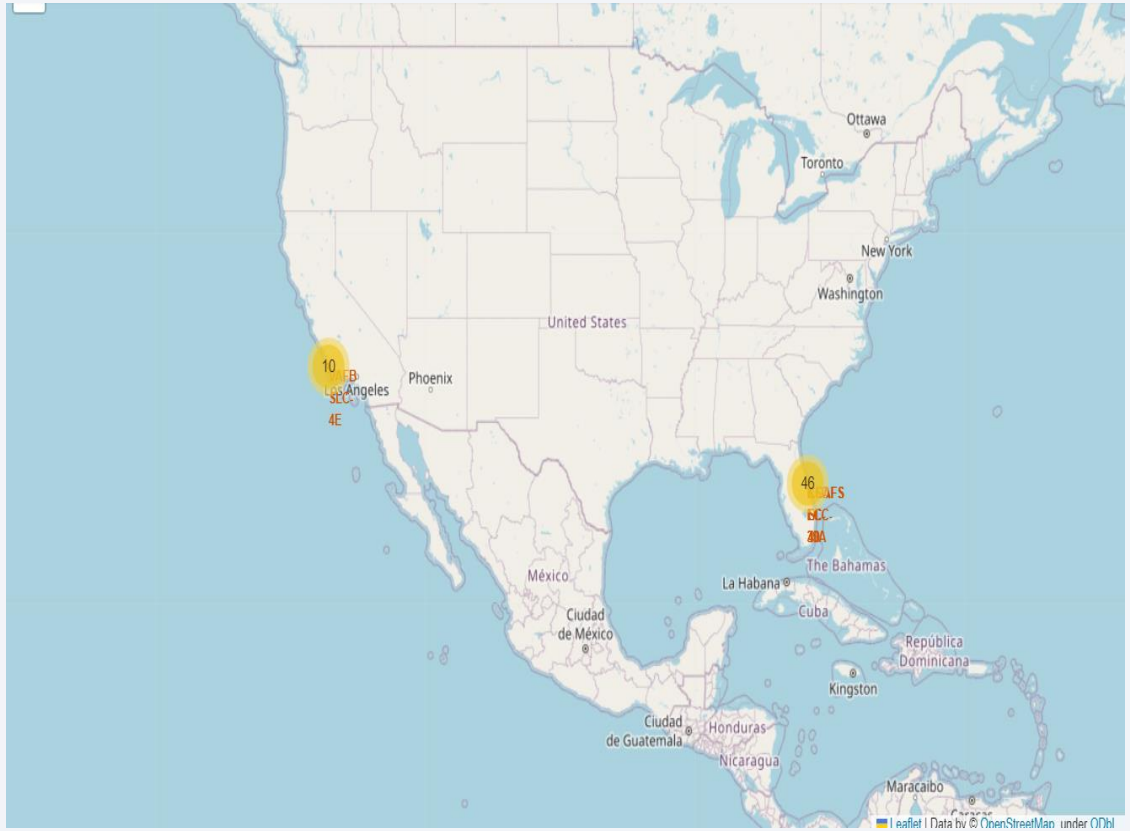
SpaceX Launch Sites

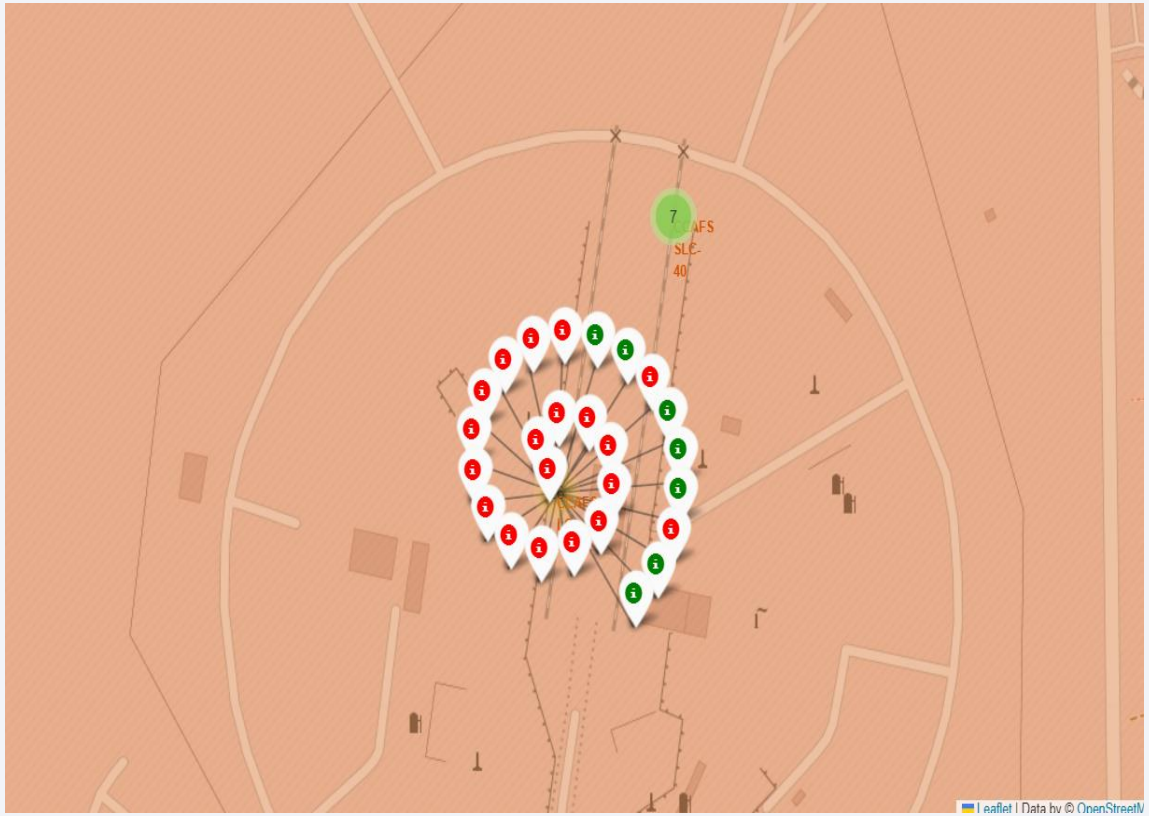
- We can see SpaceX launch sites are at The Unites states of America's coasts California and Florida.
- All launch sites are in proximity to the equator and the coast.
- The launch sites in close proximity to the coast are for safety reasons.



Success/Failed Launch Sites

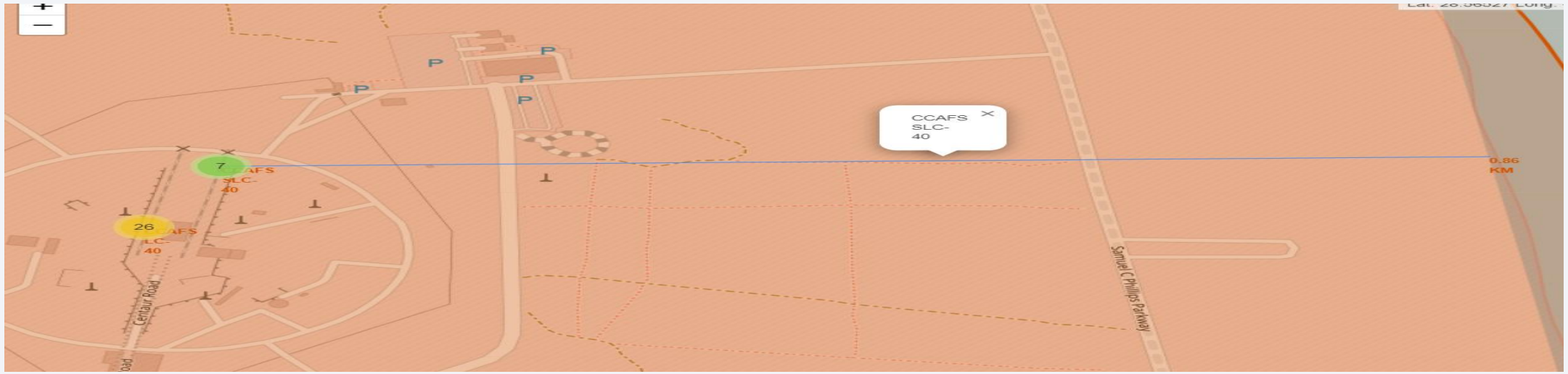
- We enhanced the map by adding the launch outcomes for each site to see which sites have high success rates. The numbers in yellow marker shows the total outcomes per site and green marker shows successful launches and Red Marker shows Failures.

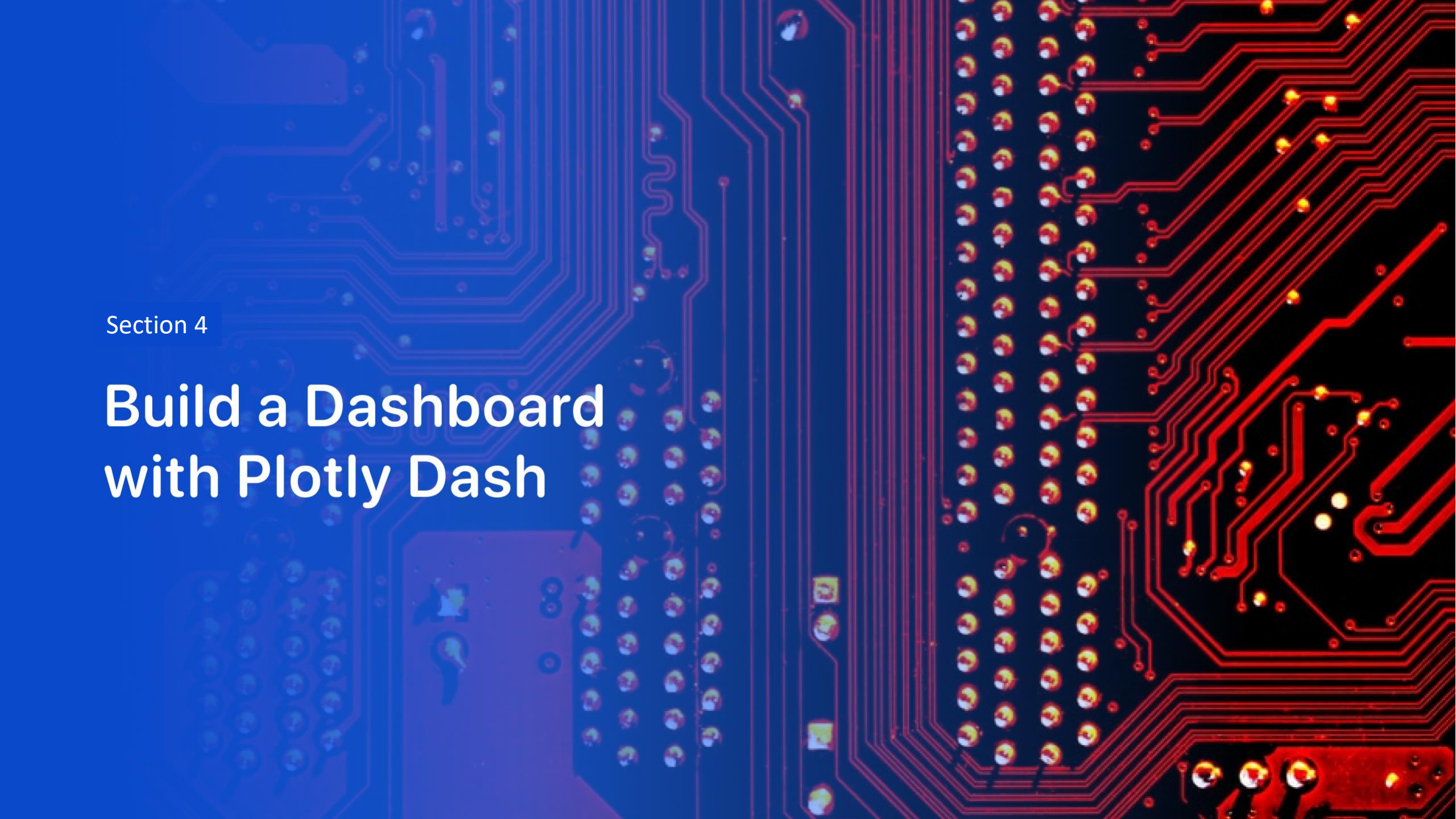




Distance between Launch Sites to proximities

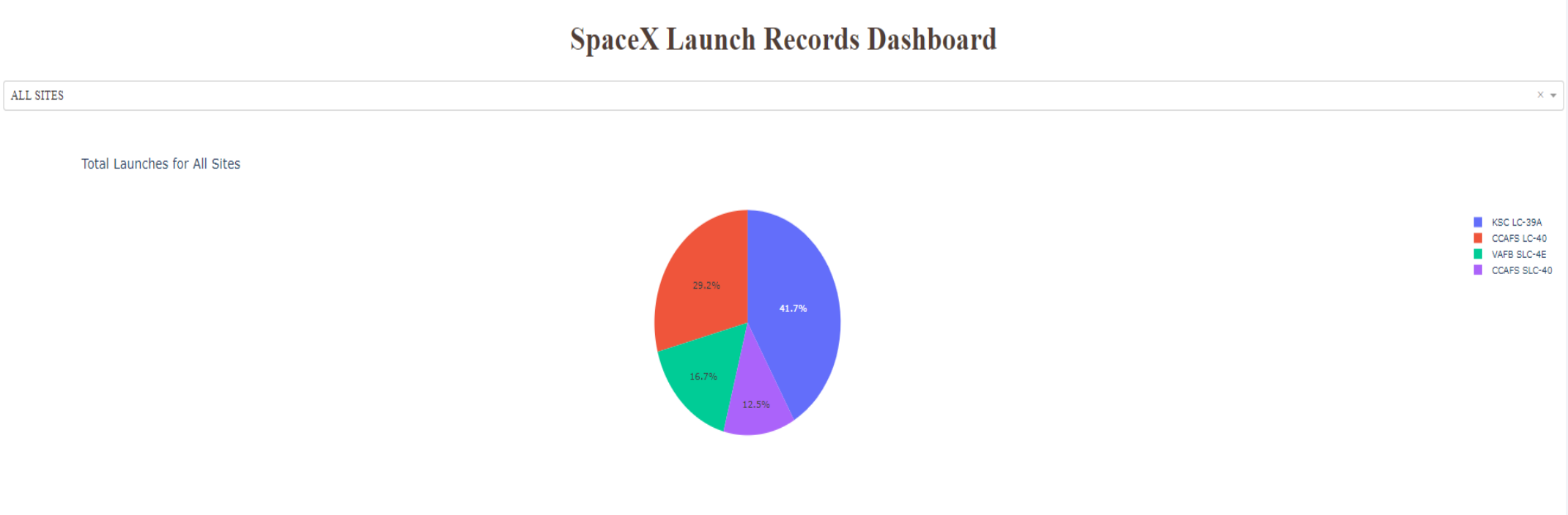
- We explored launch site and its closed proximities to find nearest railway, coastline and highway
- As mentioned before, launch sites are in close proximity to equator to minimize fuel consumption by using Earth eastward spin to help spaceships get into orbit.
- Launch sites are in close proximity to coastline so they can fly over the ocean during launch.
- Launch sites are in close proximity to highways, which allows for easily transport required people and property.
- Launch sites are in close proximity to railways, which allows transport for heavy cargo.
- Launch sites are not in close proximity to cities, which minimizes danger to population dense areas.




Section 4
Build a Dashboard with Plotly Dash

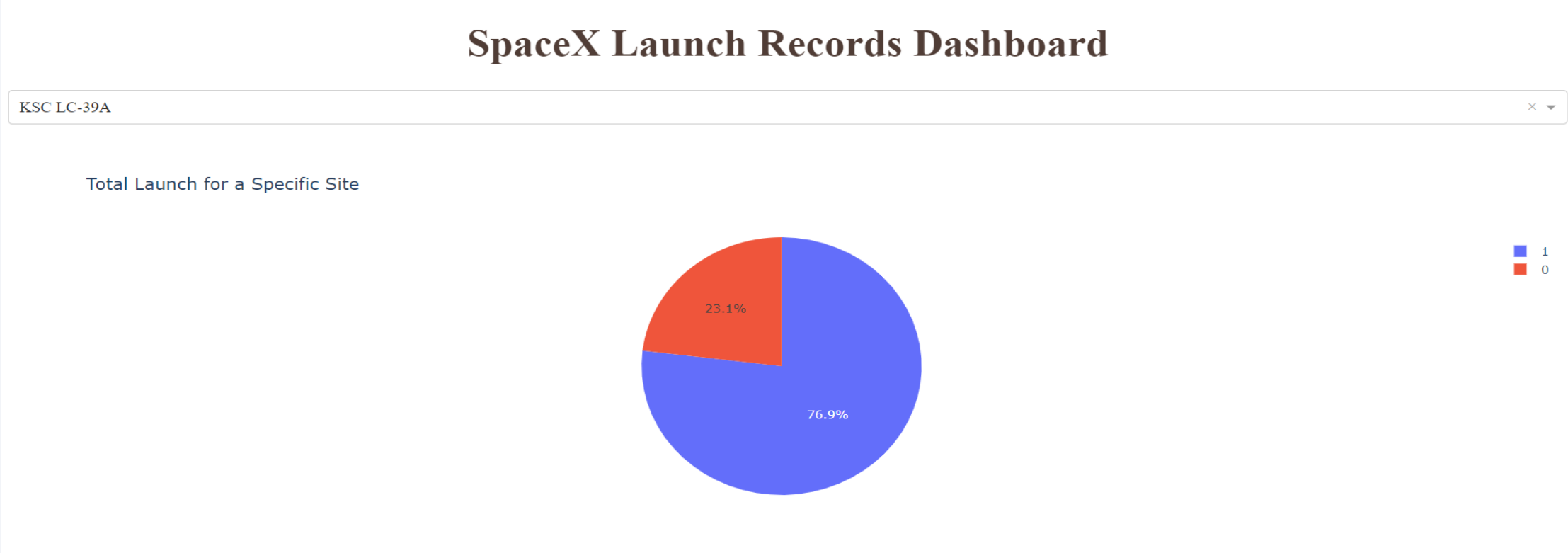
Success by all Launch Sites

- Launch success count for all sites



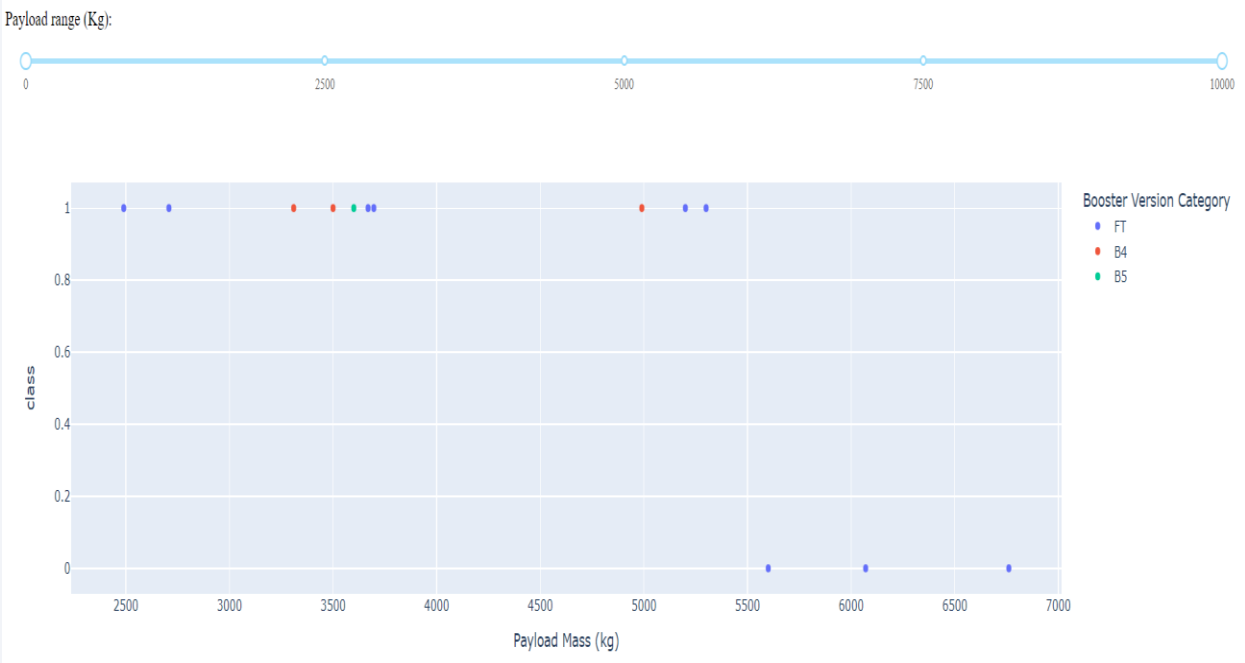
Launch site with highest success Ratio

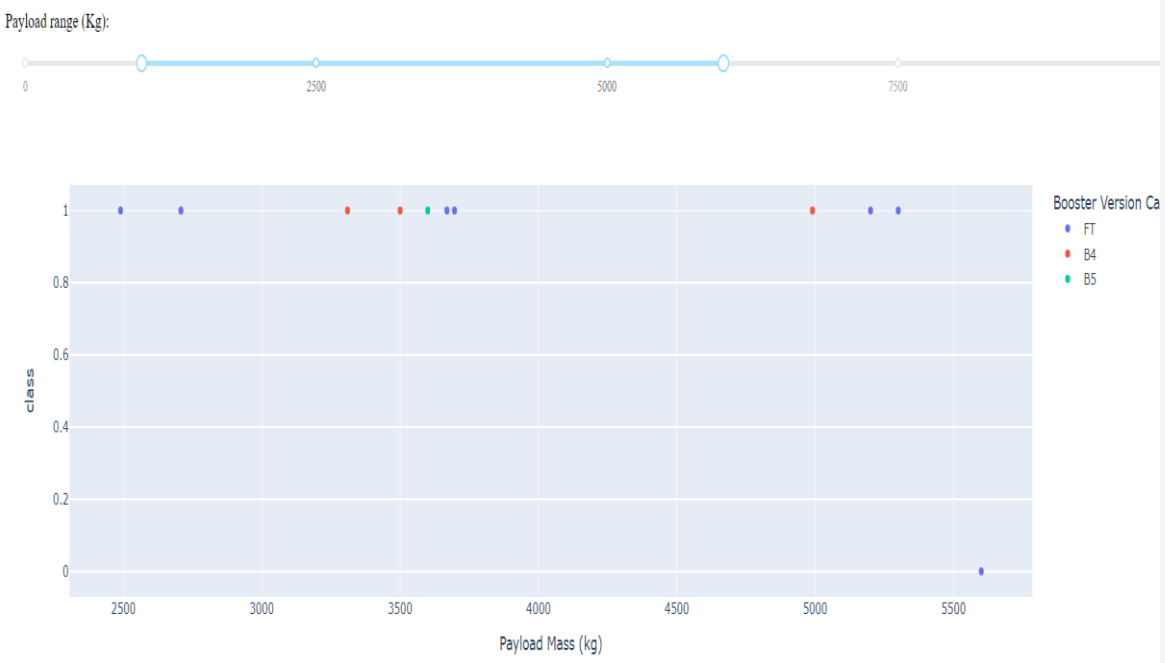
- KSC LC-39A has the highest success rate of 76.9%



Payload vs. Launch Outcome scatter plot

- Payload range between 2490 to 5300 has the highest success rate.
- Between above range FT booster version has the highest success rate.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The accuracy of all the models looks same.

TASK 12

Find the method performs best:

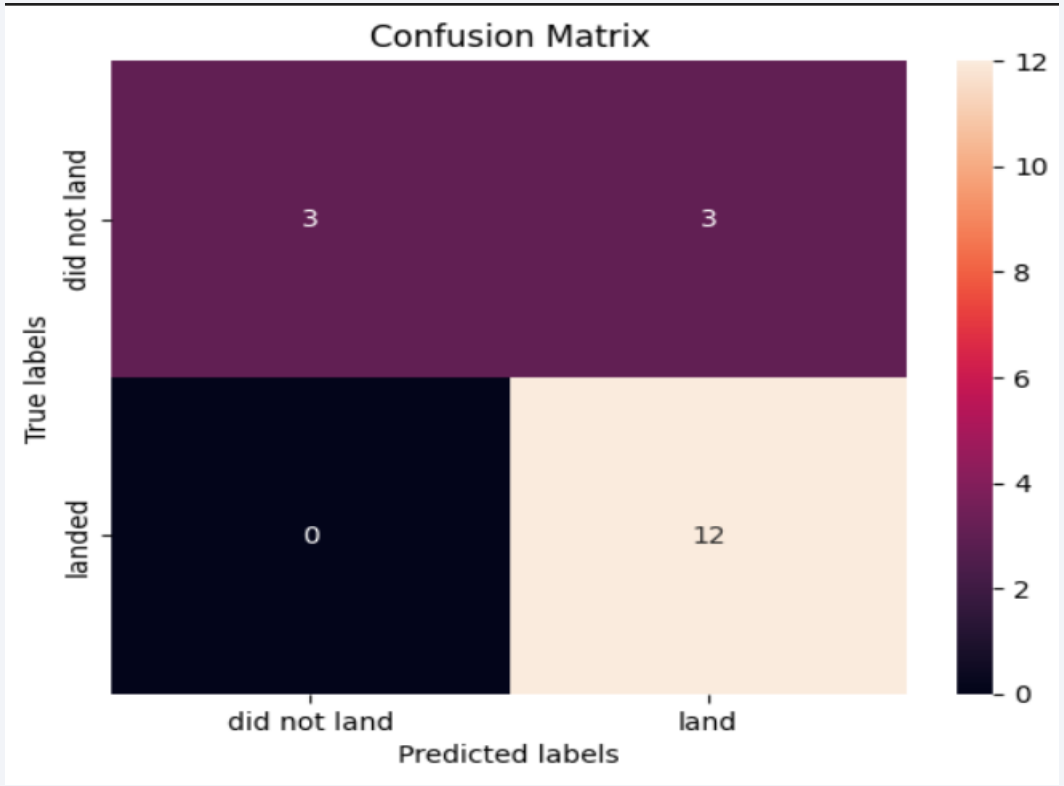
```
scores = [logreg_cv.score(X_test, Y_test),svm_cv.score(X_test, Y_test),tree_cv.score(X_test, Y_test),knn_cv.score(X_test, Y_test)]
print(scores)
```

```
[0.8333333333333334, 0.8333333333333334, 0.8333333333333334, 0.8333333333333334]
```

After comparing accuracy of above methods, they all preformed practically the same

Confusion Matrix

- The confusion matrix for all the model 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.



Conclusions

- CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40 are launch sites in space mission.
- The larger the flight amount, the greater the success rate at a launch site.
- Success rate since 2013 kept increasing 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.
- Logistic Regression, Support Vector machines, Decision Tree Classifier and K-nearest neighbors models gave similar accuracy and anyone can be chosen for machine learning algorithm model and tuning

Appendix

- Githhub Project Source code Reference:

  https://github.com/maina-duseja/DS-Capstone/tree/main

Thank you!