IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Collected the data related to SpaceX rocket launch from SpaceX website and Wikipedia Page. Performed Exploratory data analysis on the dataset to see how the columns are correlated and to find the best labels to perform the data analysis. Then created a map using Folium to see which rocket launch sites has most successful rate and at last utilized the machine learning algorithms Such as GridSearch to find the accuracy.

- The used machine learning models are: GridSearch, Decision Tree, Support Vector Machine, K-Nearest Neighbor. The predicted accuracy of most of models was around 83.33%.

Introduction

Background

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Task

- To train a machine learning model and use public information to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX Public API and Wikipedia Page
- Perform data wrangling
  - Calculated the number of launch on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurrence of mission outcome per orbit
  - Created a landing outcome from outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Utilized GridSearch

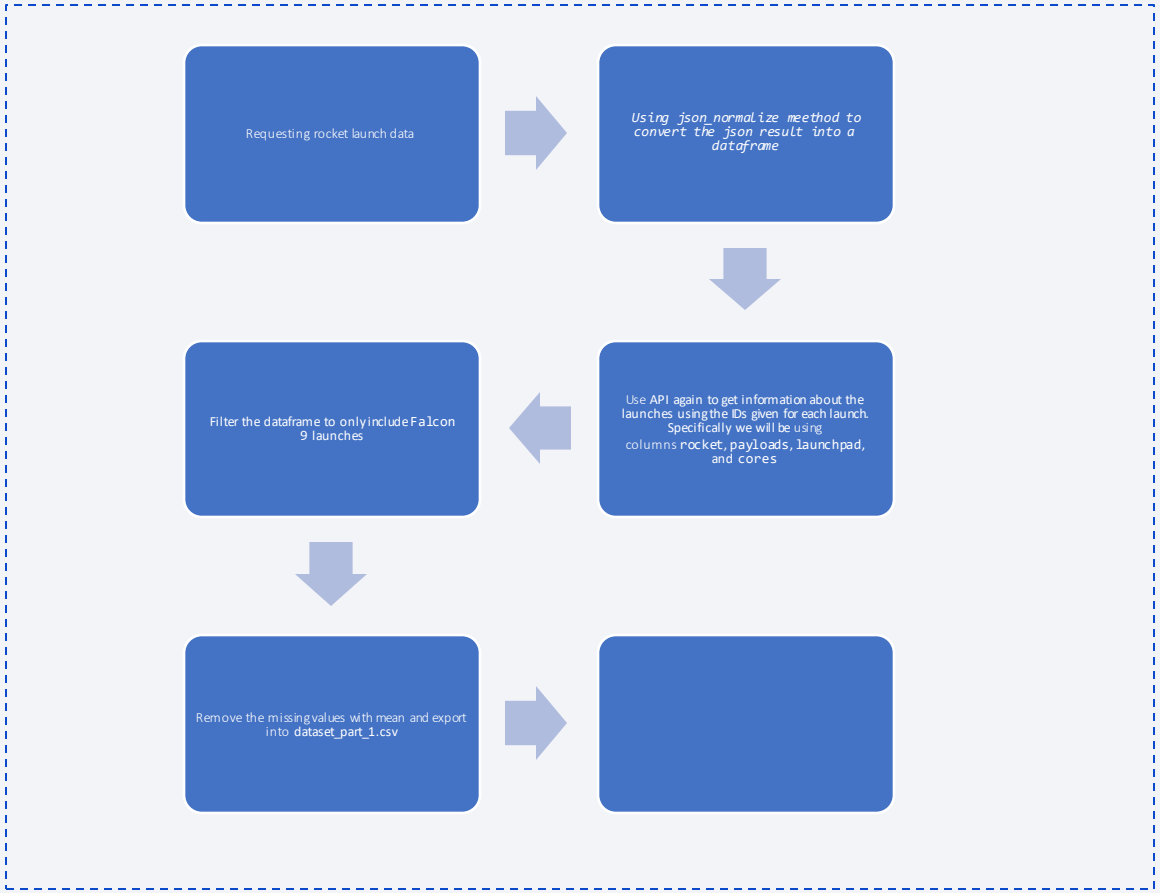
Data Collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia.
- Space X API Data Columns:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
  - Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
  - Flight No., Launch site, Payload,

Data Collection – SpaceX API

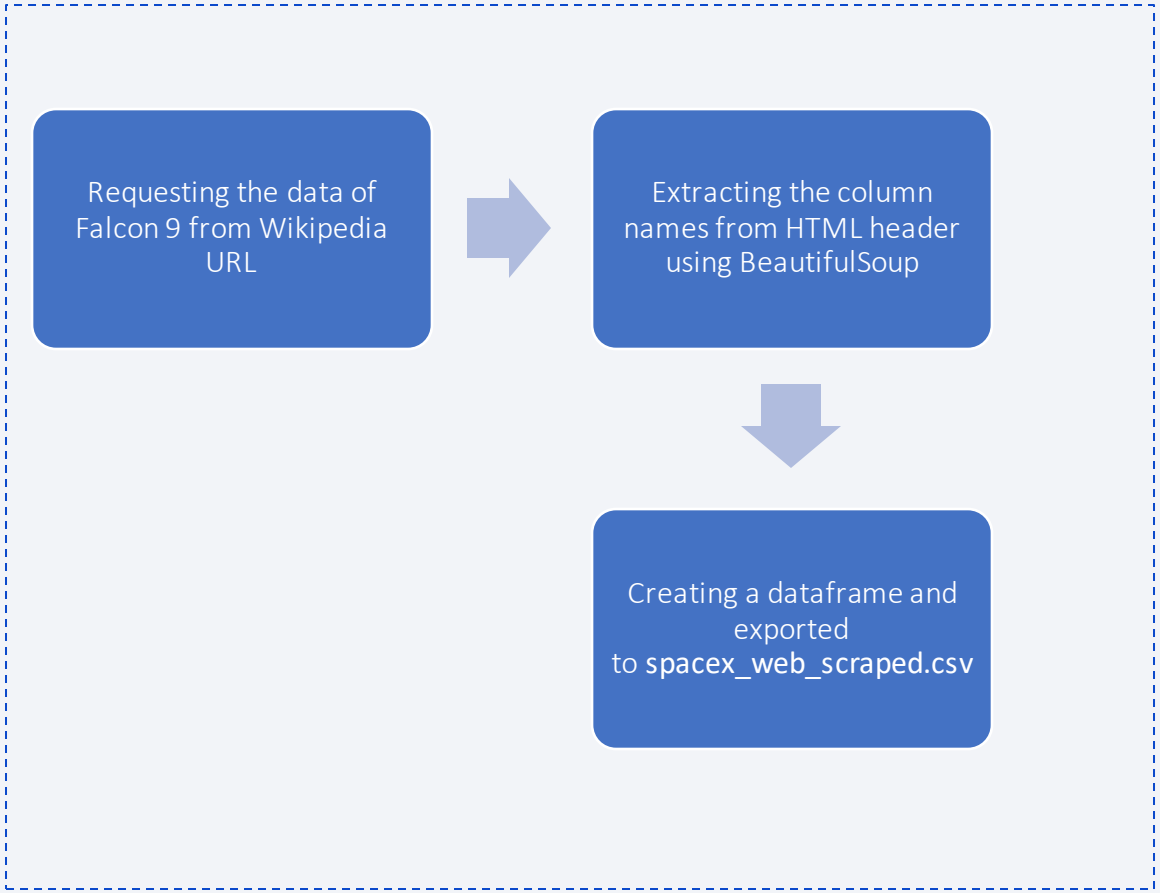
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

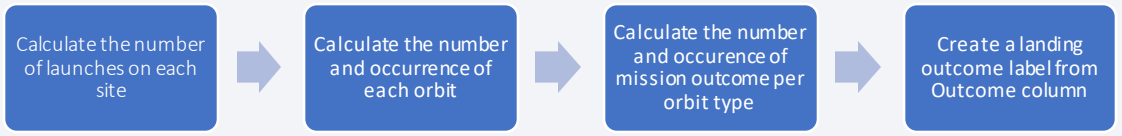
- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- After extracting the data from the dataset, the total percentage of missing values are counted. Then the total number of launches on each site is calculated using value_counts(), then occurenece of each orbit is calculated and then using the method .value_counts() on the column Outcome to determine the number of landing_outcomes.Then assign it to a variable landing_outcomes. The another column is created where 0 means bad outcome and 1 means landed successfully

- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site.
- Visualize the relationship between Payload and Launch Site
  - VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload and Orbit type
- The sucess rate since 2013 kept increasing till 2020
- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL Queries performed are:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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
- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

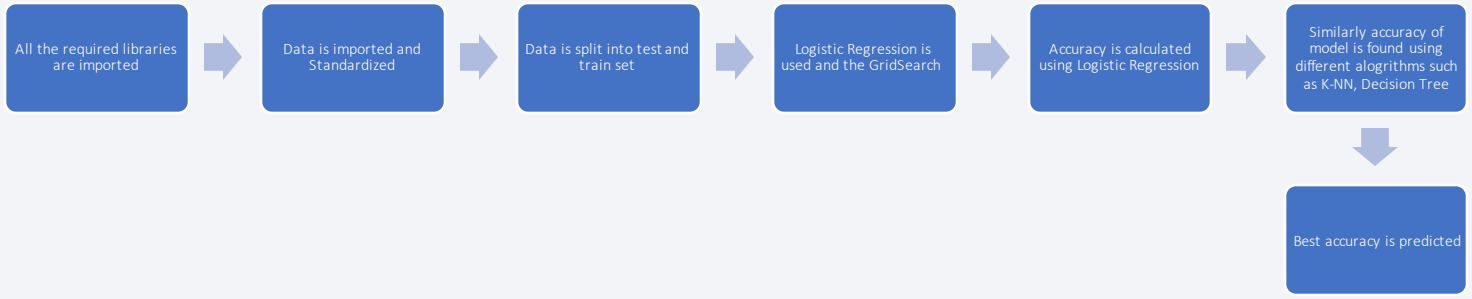
- Marker with Circle, Popup Label of NASA Johnson Space Center using the latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Added Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added Lines to show distances between the Launch Site and Railway, Highway, Coastline and Closest City.
- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Scatter Plot is added to show the interaction with payload mass and class, and success is counted with different payload massess.
- https://github.com/SauravDeb-98/IBM-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The dataset is divided into test and train set, and then GridSearch is used to find the accuracy, the different algorithms used are Support vector machine, Decision Tree, K-Nearest neighbor, and then all of these are compared to find which algorithm yield the best score.



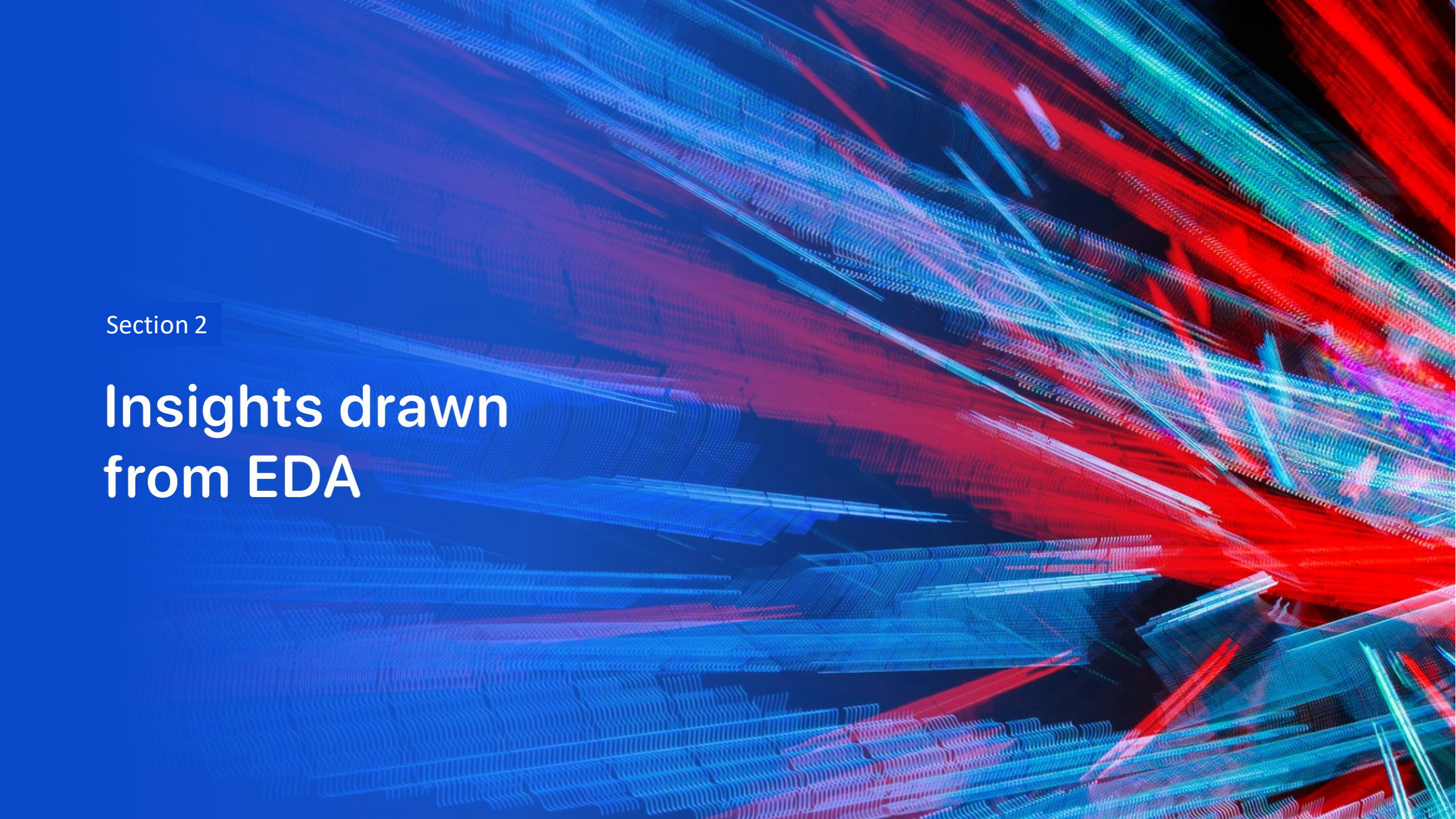
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

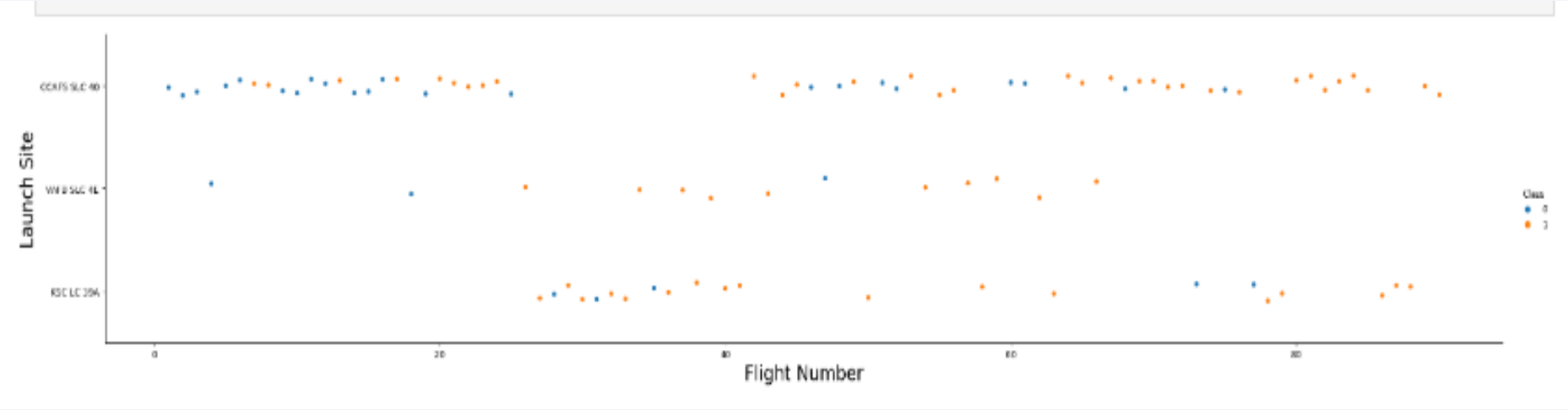
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

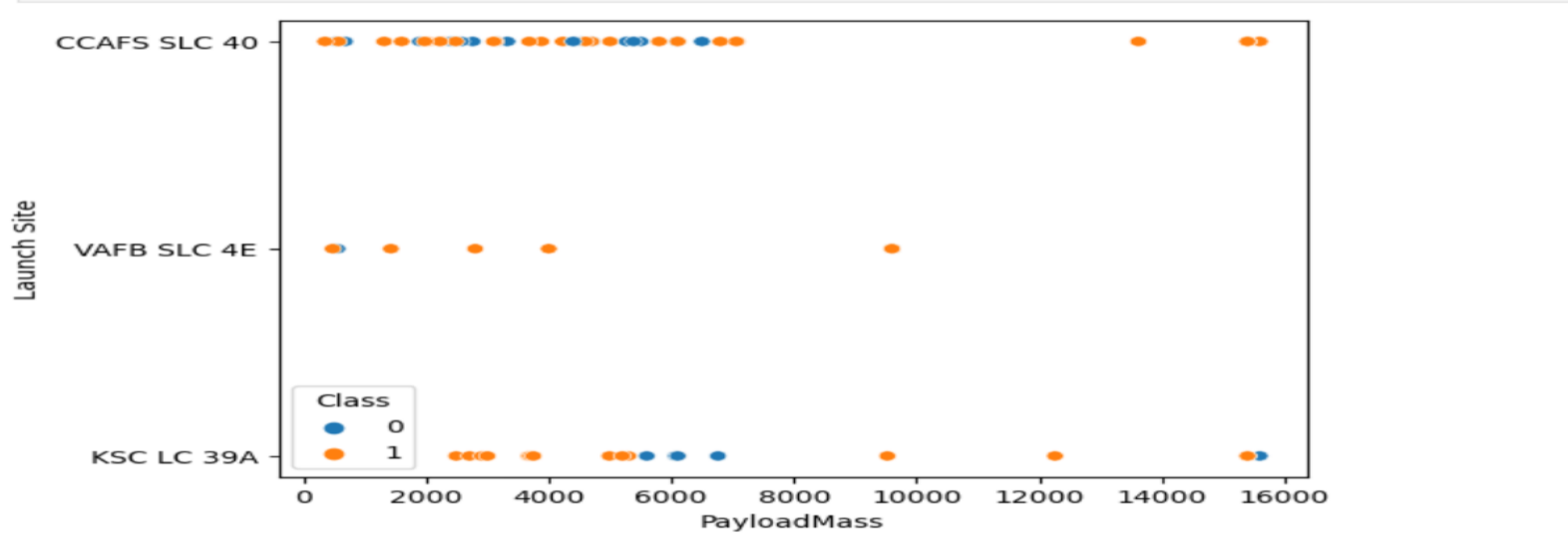


Flight Number vs. Launch Site



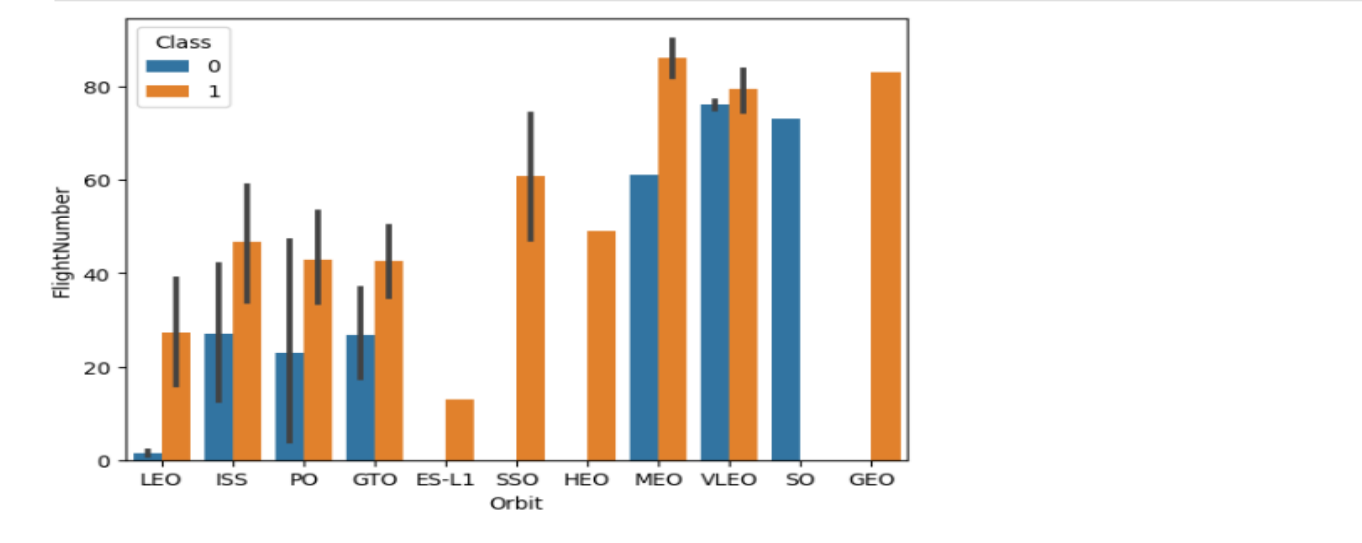
- Flights at the beginning failed however, with time the number of successful flights increased
- CCAFS SLC 40 has highest number of successful flights

Payload vs. Launch Site



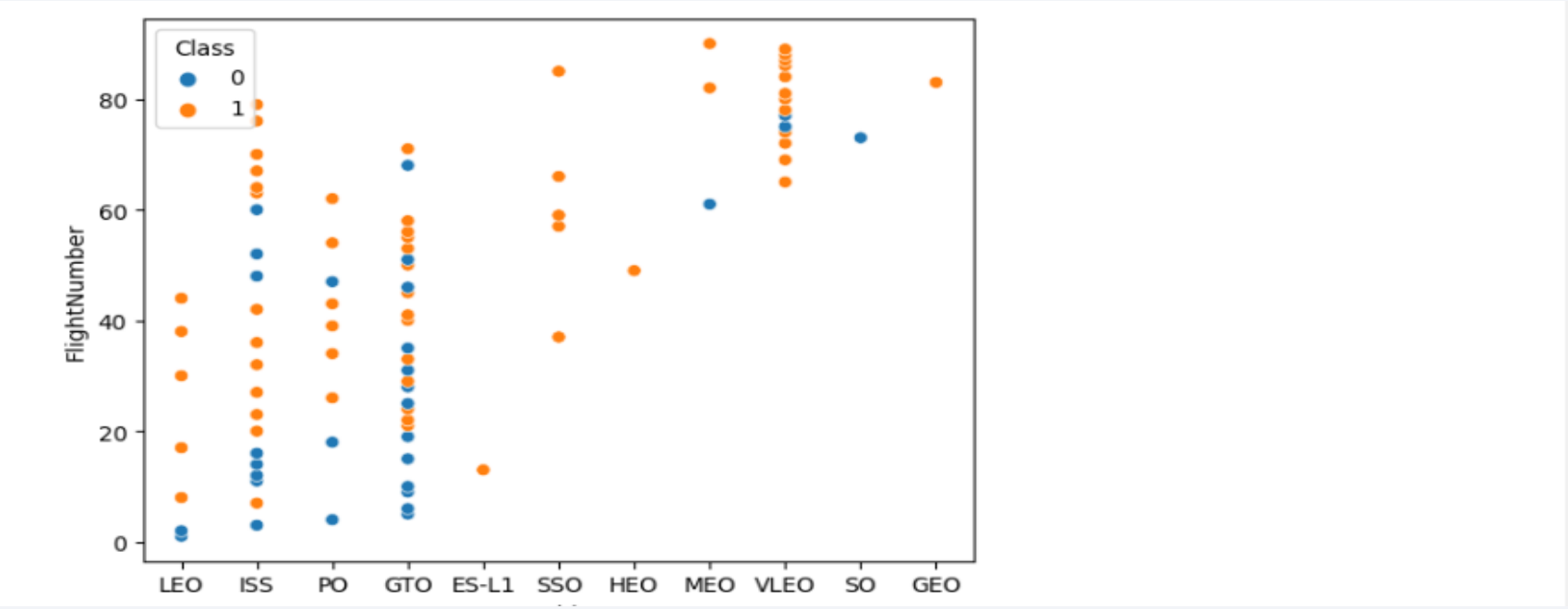
- The class is labelled as 0 which means failure; and 1 means Success
- For payload mass between 2000 to 6000, CCAFS SLC 40 has highest success
- KSC –LC-39 A has higehst success rate overall

Success Rate vs. Orbit Type



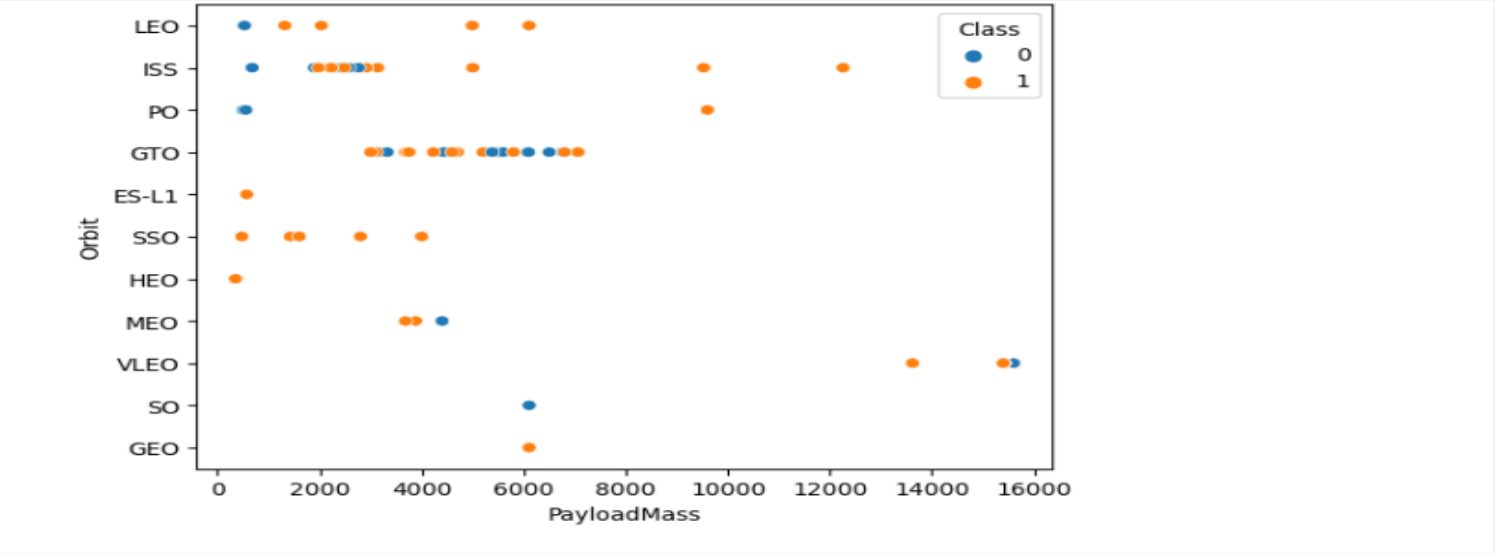
- Geo, SSO, ES-L1 has 100% success rate
- SO has 0% success rate

Flight Number vs. Orbit Type



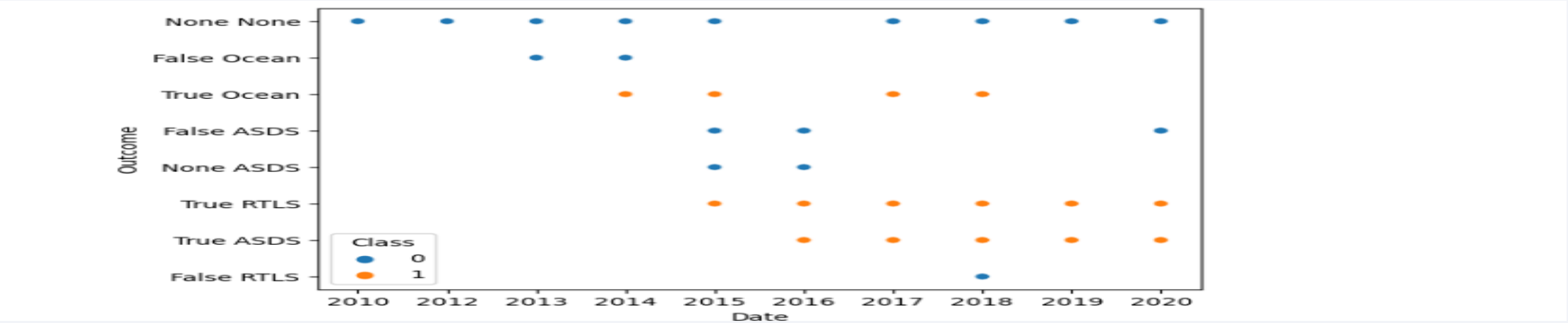
- Geo, SSO, ES-L1 has 100% success rate
- SO has 0% success rate

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



- Year 2017 and 2018 has more successful outcomes
- From year 2010 to 2013 there were no successful outcomes

All Launch Site Names

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

```
In [12]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

- The names of the unique launch sites

Launch Site Names Begin with 'CCA'

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

In [16]:
```
%sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Out[16]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Names of Launch Site begin with 'CCA'

Total Payload Mass

```
In [23]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
Out[23]: Total_Payload
         619967
```

- Total Payload Mass

Average Payload Mass by F9 v1.1

Task 4

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

In [24]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version ='F9 v1.1'
```

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

Out[24]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average Payload Mass of F9 v1.1

First Successful Ground Landing Date

Hint:Use min function

In [25]:
```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Mission_Outcome = 'Success'
```

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

Out[25]:

MIN(Date)
2010-04-06

- First date of successful ground landing

Successful Drone Ship Landing with Payload between 4000 and 6000

In [32]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4
```

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

Out[32]:

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

Task 7

- The names of boosters which 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

List the total number of successful and failure mission outcomes

In [36]:
```
%sql SELECT COUNT(Mission_Outcome) as Number_Of_Outcomes, Mission_Outcome AS Outcomes FROM SPACEXTABLE GROUP BY(Mission_Out
```

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

Out[36]:

Number_Of_Outcomes	Outcomes
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

- 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

```
In [39]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

Out[39]:

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 names of the booster which have carried the maximum payload mass

2015 Launch Records

```
In [43]: %sql SELECT substr(Date, 4,2), Landing_Outcome as Outcome, Booster_Version, Launch_Site as Site FROM SPACEXTABLE WHERE subs
 * sqlite:///my_data1.db
Done.
```

Out[43]:

substr(Date, 4,2)	Outcome	Booster_Version	Site
5-	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
6-	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
6-	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
6-	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

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

substr(Date,7,4)= 2015 for year.

In [43]:
```
%sql SELECT substr(Date, 4,2), Landing_Outcome as Outcome, Booster_Version, Launch_Site as Site FROM SPACEXTABLE WHERE subs
```

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

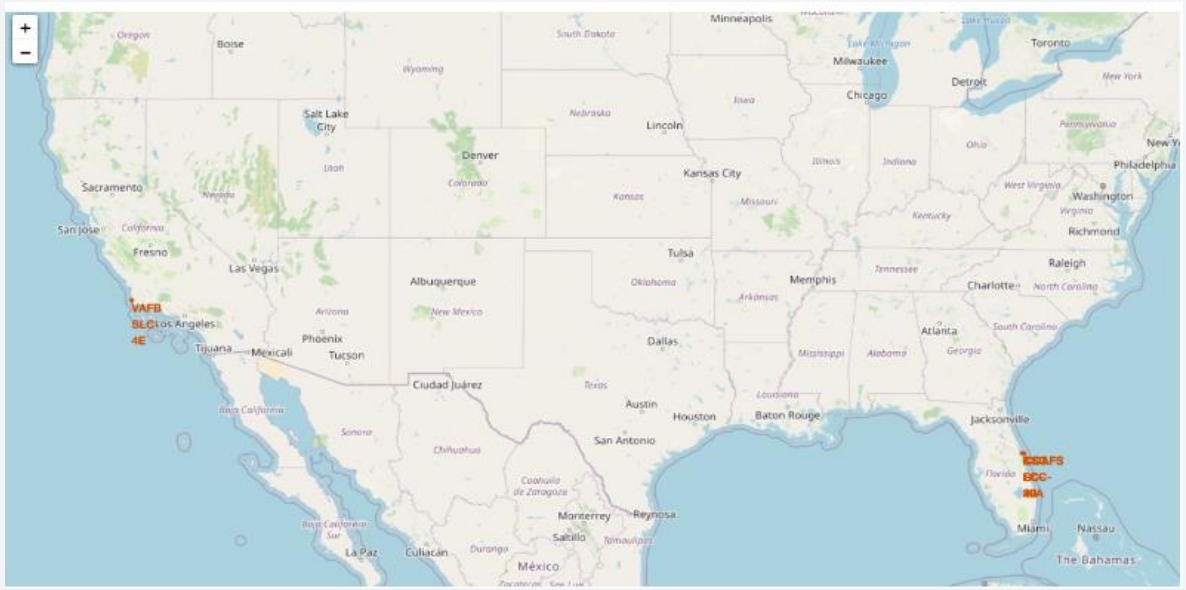
Out[43]:

substr(Date, 4,2)	Outcome	Booster_Version	Site
5-	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
6-	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
6-	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
6-	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

- 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

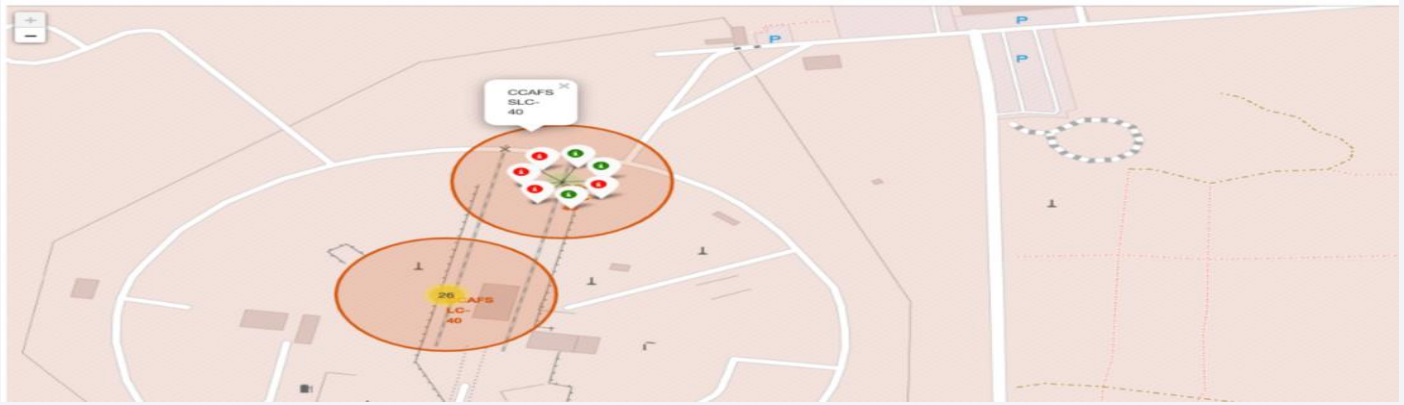

Section 3
Launch Sites
Proximities Analysis

All launch sites



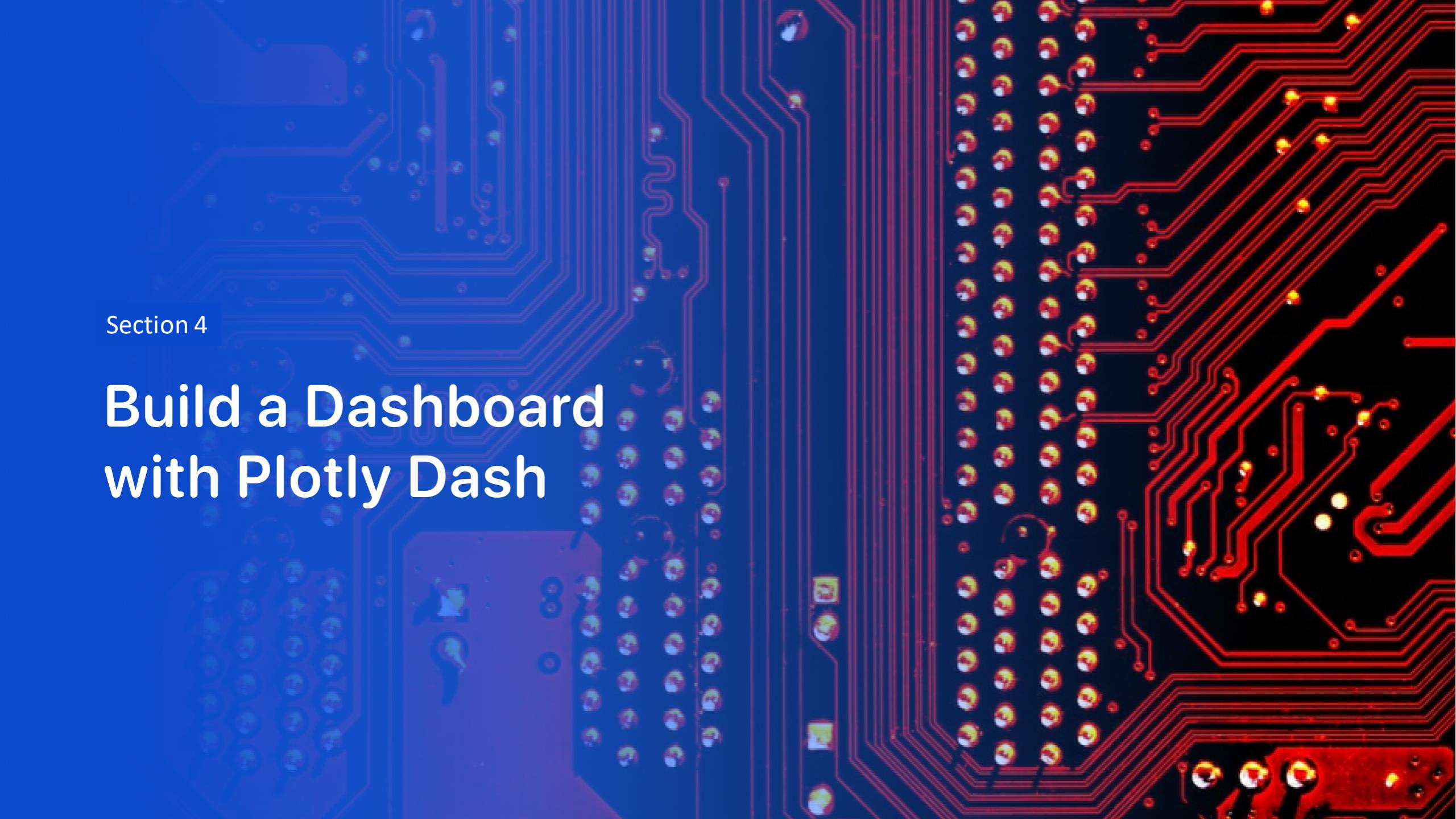
- The launch sites are marked with red color. And we can decipher that all launch sites are closer to the coast.

Labelled Launch Sites



- Green Label = Successful Sites Launch
- Red Label = Failure Site Launch

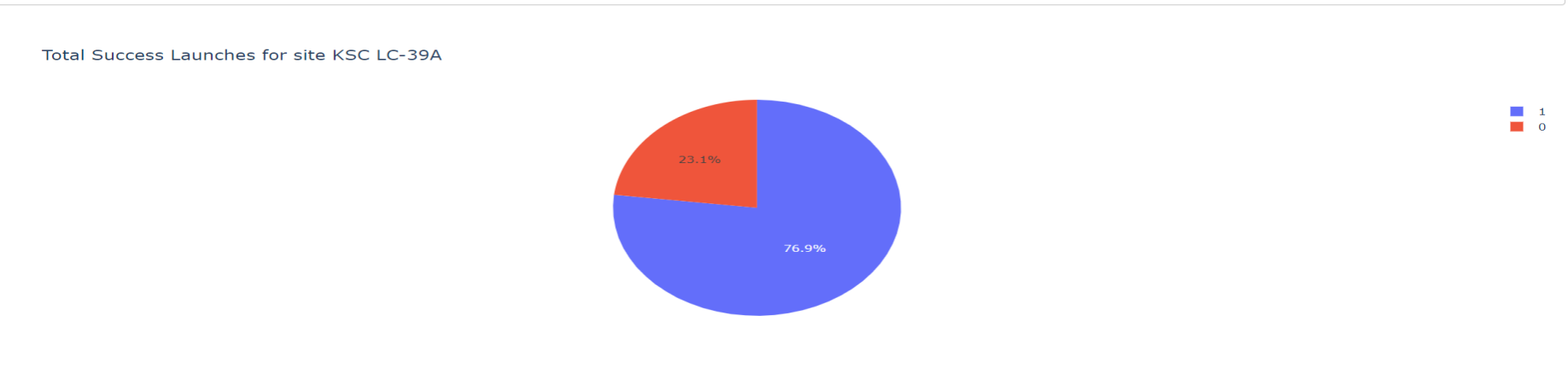
Distances of various points from launch sites


Section 4
Build a Dashboard with Plotly Dash

Pie Chart showing the Success Count of each site

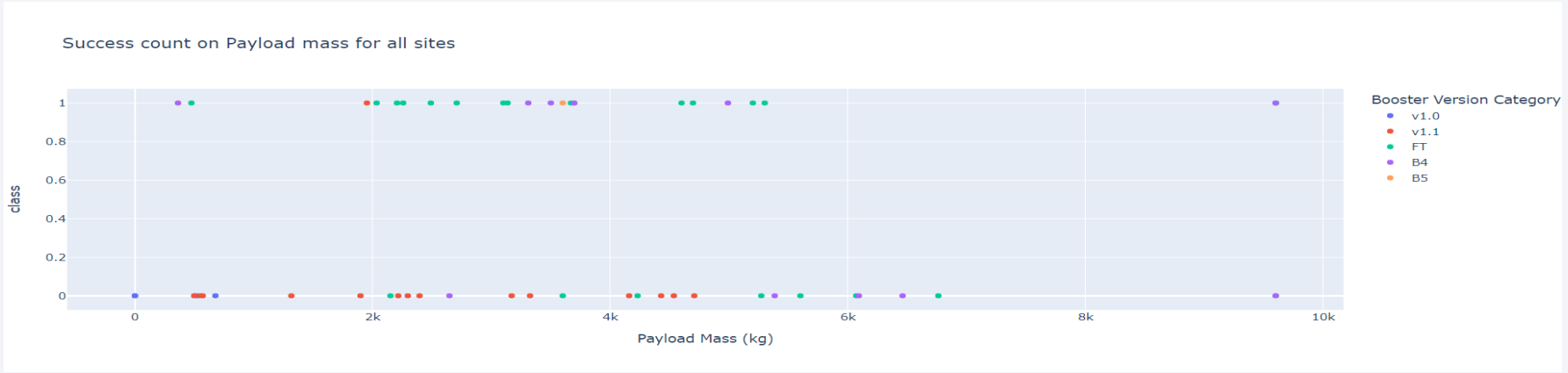
- KSC-LC-39A has highest success rate
- CCAFS-LC-40 has lowest success rate

Pie chart of the launch site with highest successful rate



- Success rate is 76.9% and Failure is 23.1%

Payload Vs Launch Outcome Scatter Plot for all sites



- Payload Mass between 2000 to 5000 has highest successful rate

Section 5

Predictive Analysis (Classification)

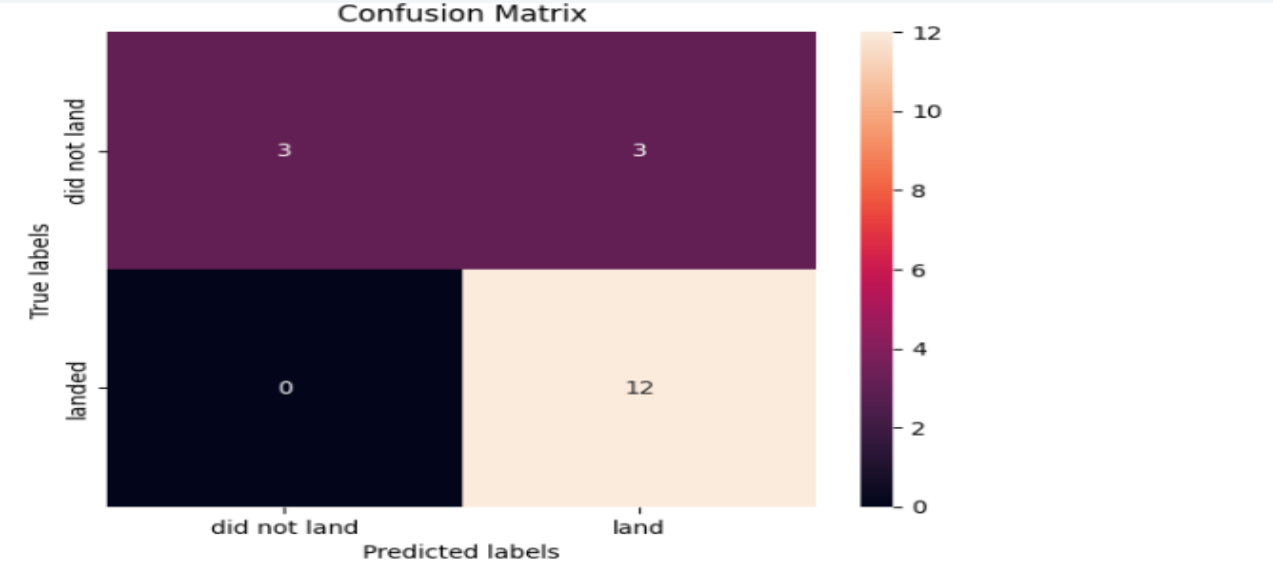
Classification Accuracy

```
[14]: print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
      print("accuracy :",logreg_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8196428571428571
```

- Accuracy of Logistic regression

Confusion Matrix



- False positive is the major issue.

Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass.

- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Special Thanks to IBM and Coursera

Thank you!