IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeffrey Lim Lee Keong
22 Nov 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection/ Perform data Wrangling a and preprocess data
  - Exploratory data analysis using visualization and SQL
  - Foilum and Ploty Dash for interactive visual analytics
  - Different classification models for predictive analysis
- Summary of all results
  - Most of the successful recovery of the 1st stage of the launcher landed on KSC LC-39A with a payload of approx. 2034Kg to 5300Kg
  - Results from different classification models yield to have approximate the same accuracy on the test data set

Introduction

- Project background and context
  - To reduce the cost on launching a rocket, it is important to recover the 1st stage of the launcher which largely affect the cost of launching a rocket as this stage is quite and expensive. Unlike other rocket launcher, Space X's Falcon 9 can recover the 1st stage. However, the attempt to recover the 1st stage is not always successful.
  - Space Y would like to complete with Space X. Therefore, in this project, it is to gather and analyze the information from Space X with regards to number of successful/unsuccessful recovery of 1st stage. Then we will train a machine learning model and use public information to predict if Space X will reuse the 1st stage.
- Problems you want to find answers
  - Where and how to gather data?
  - What are the features that correlate to the successful of 1st stage recovery?
  - Which launch site has the highest successful recovery?
  - Which machine learning model is the best?

Section 1

Methodology

Methodology

Executive Summary

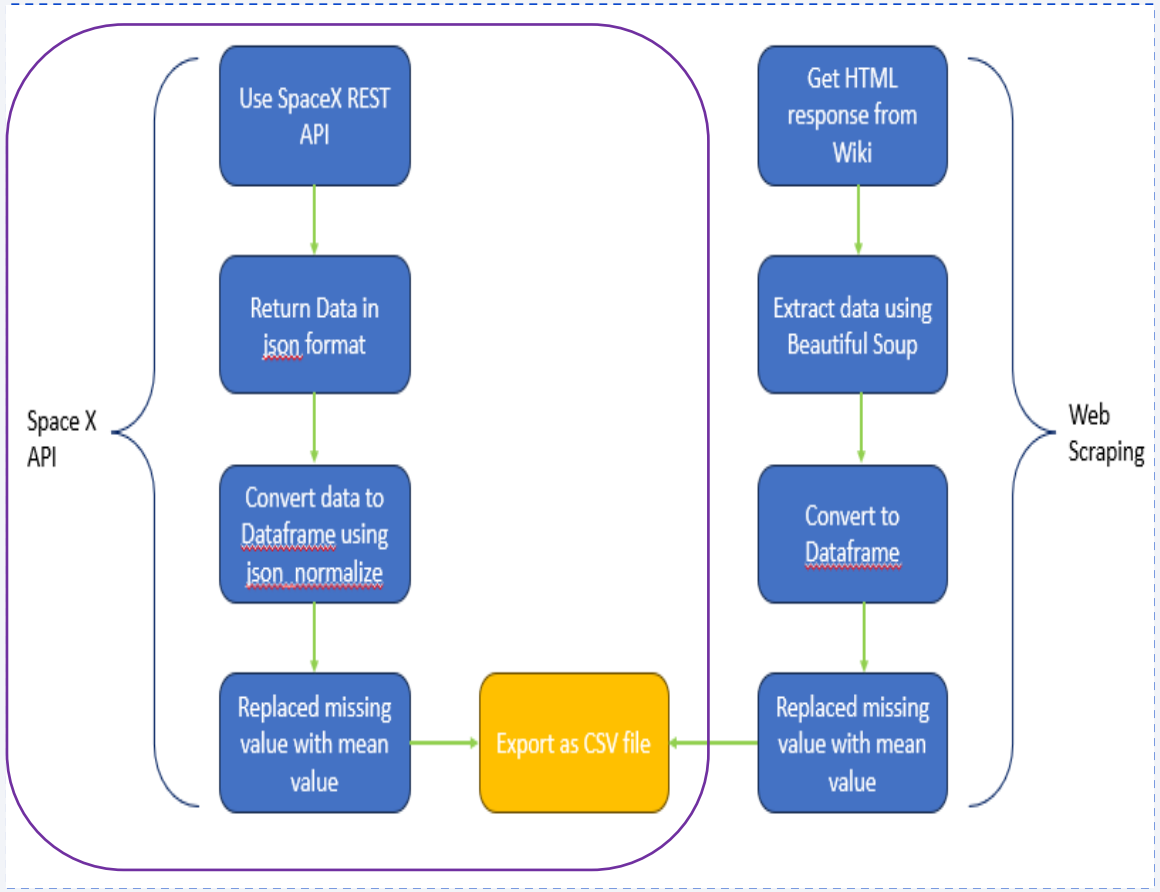
- Data collection methodology:
  - Using Request API to access rockets/launch pad/payload and cores data
  - Web scraping from wikipedia
- Perform data wrangling
  - Dealing missing value data as well as one hot encoding categorical data field
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build different classification models using Logistic Regression/ SVM/Decision Tree/KNN
  - Fine tune models by using GridSearchCV function to find the best parameters for each model

Data Collection

- Rocket launch data are collected using Request API with url "https://api.spacexdata.com/v4/launches/past". Different functions are created that help us use APU to extract information using numbers in the launch data. There are:
  - getBoosterVersion
  - getLaunchSite
  - getPayloadData
  - getCoreData
- Web scraping Falco 9 and Falcon Heavy Launches Records from Wikipedia by using beautifulSoup library
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

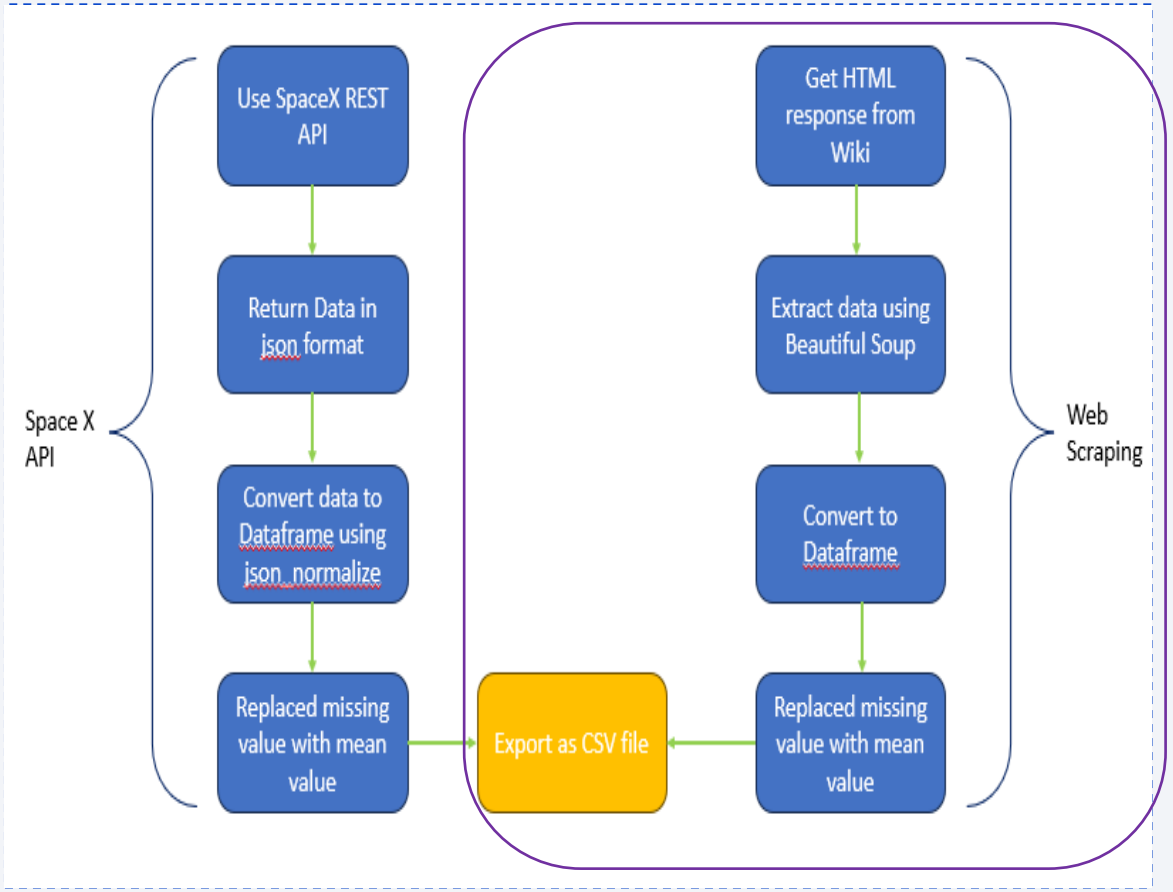
Data Collection – SpaceX API

- Gathered data from SpaceX API "https://api.spacexdata.com/v4/launches/past"
- Convert data to Dataframe using json_normalize function
- Replaced missing value with mean value
- Export as csv file format



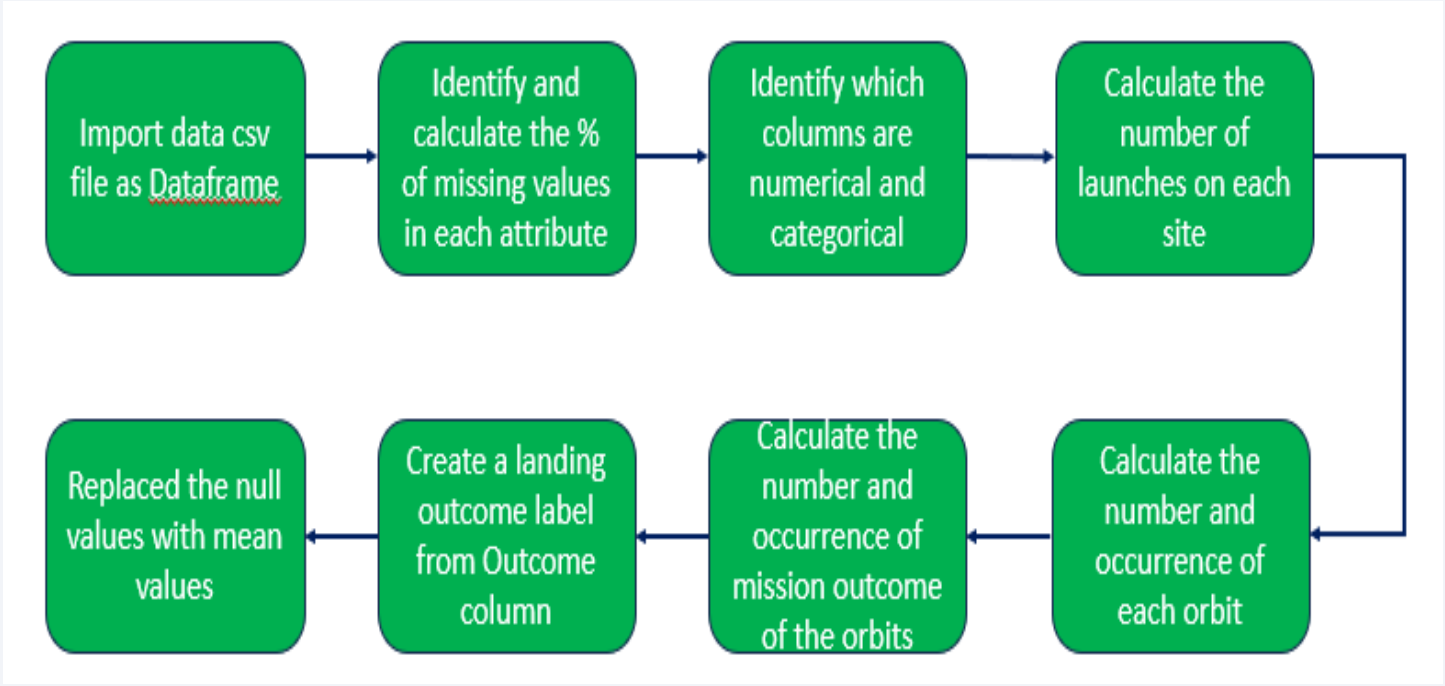
Data Collection - Scraping

- Get HTML response from Wikipedia https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Using Beauitful Soup library to extract necessary data
- Create a dictionary so as to convert to DataFrame
- Export as csv file format

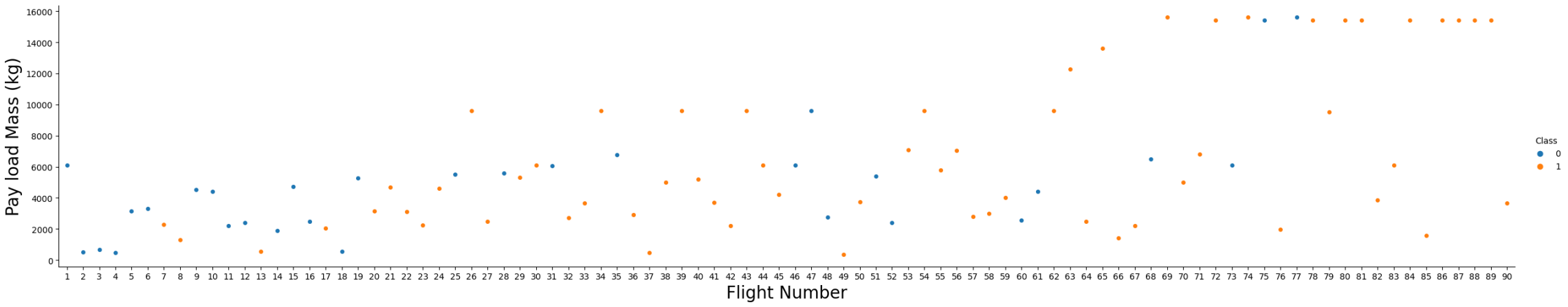


Data Wrangling

- Import data as dataframe
- Identify missing values in each attribute
- Identify which columns are numerical & categorical
- Calculate the following:
  - Number of launches on each site
  - Number & occurrence of each orbit
  - Number & occurrence of mission outcome
- Create a landing outcome label
- Handle null values by replacing it with mean value

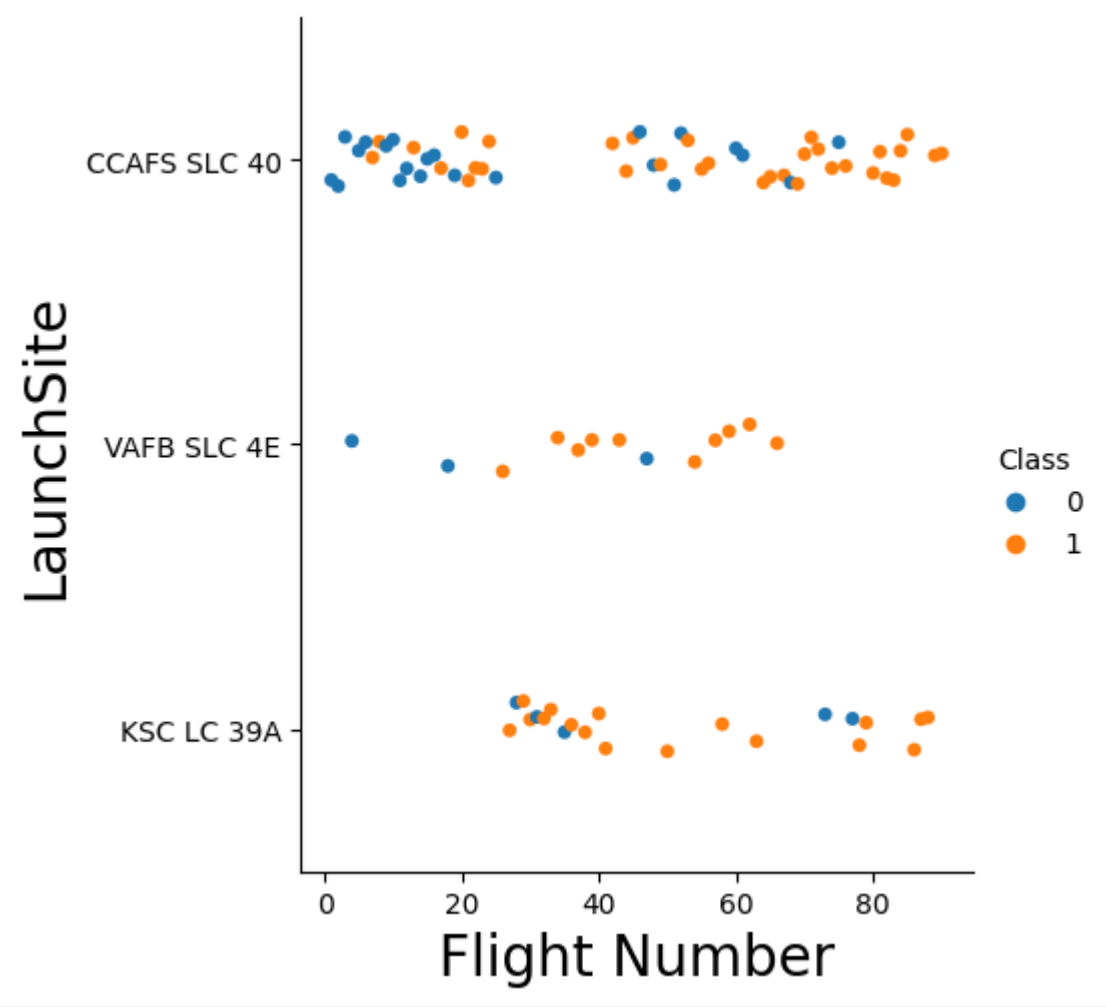


EDA with Data Visualization



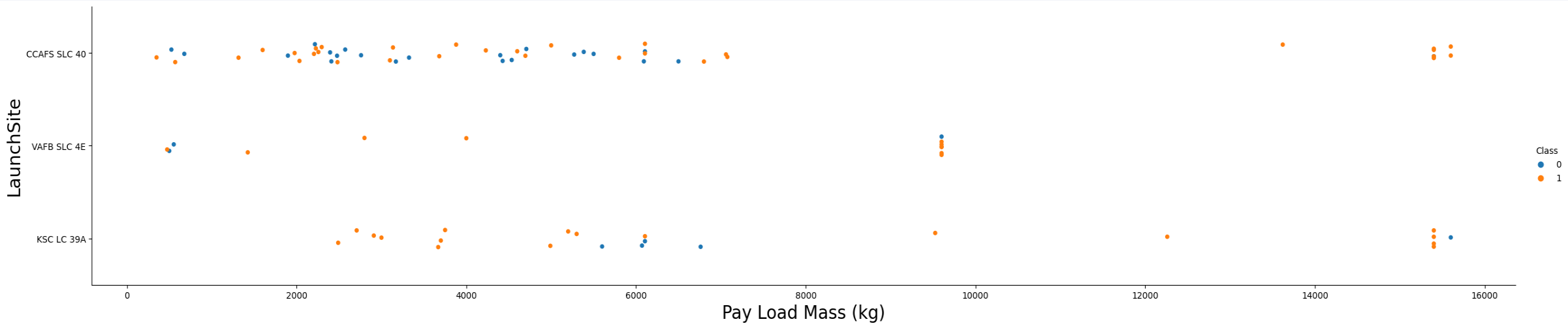
- Categorical plot is used to plot Flight number vs Payload Mass and overlay the outcome of the launch
- With increase flight number, the 1st stage is more likely to land successfully
- With increasing payload mass, the less likely the 1st stage will land successfully

EDA with Data Visualization



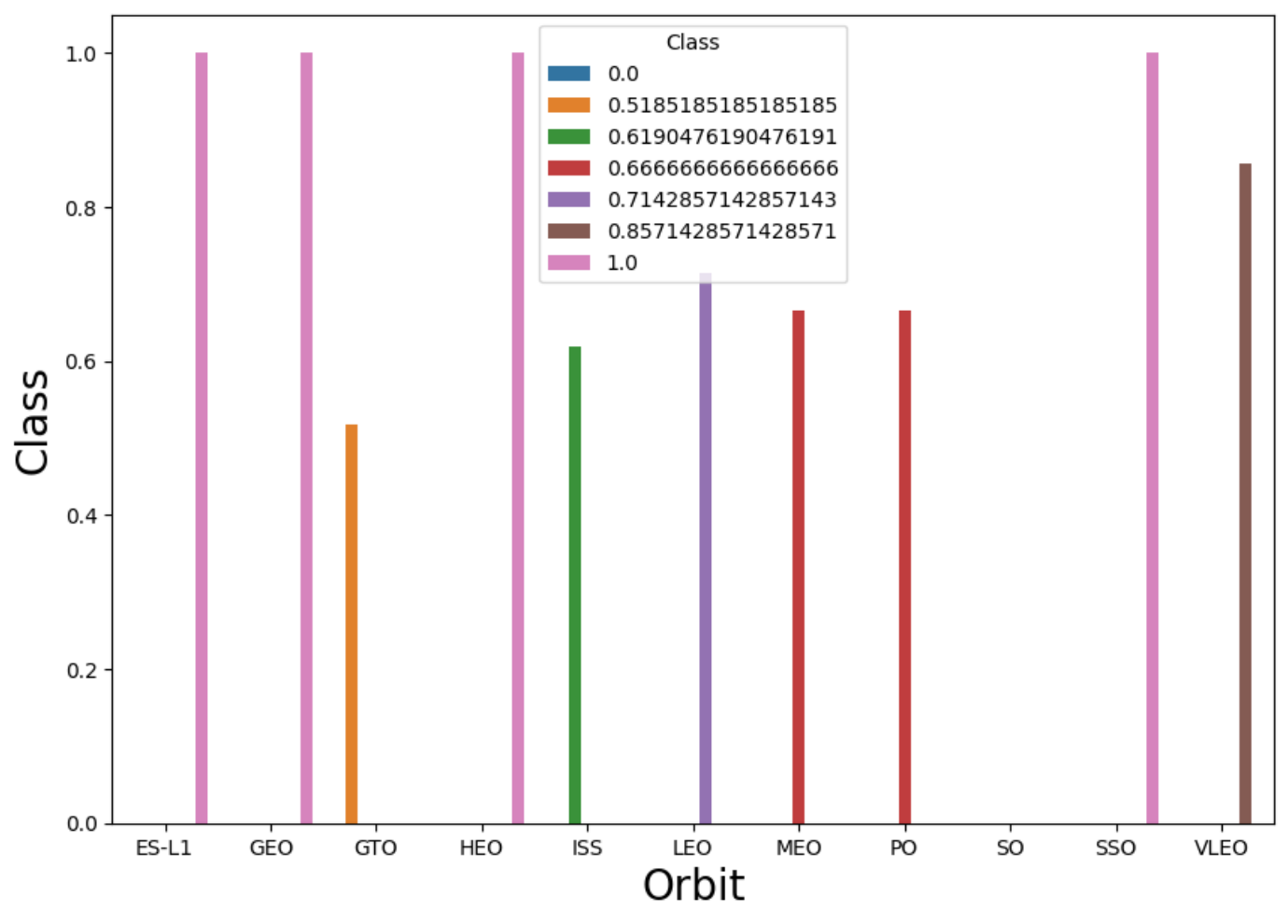
- Categorical plot is used to plot Flight number vs Launch Site and overlay the outcome of the launch
- KSC LC 39A seemed to have a higher successfully landing as compare to the rest of the sites
- For CCAFS SLC 40, the success rate increased with more flight number

EDA with Data Visualization



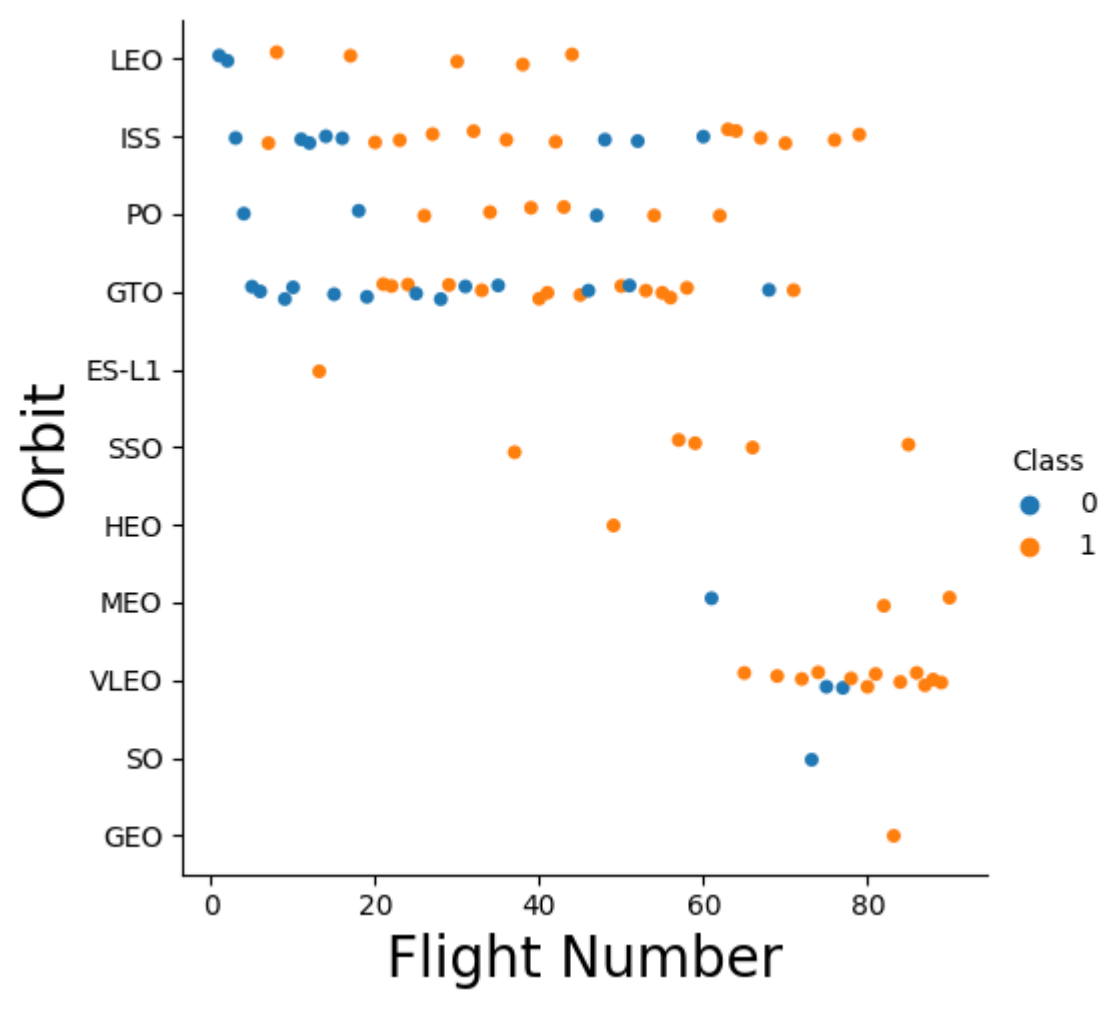
- Categorical plot is used to plot Payload Mass vs Launch Site and overlay the outcome of the launch
- There are no launch for VAFB-SLC with payload above 10000 kg

EDA with Data Visualization



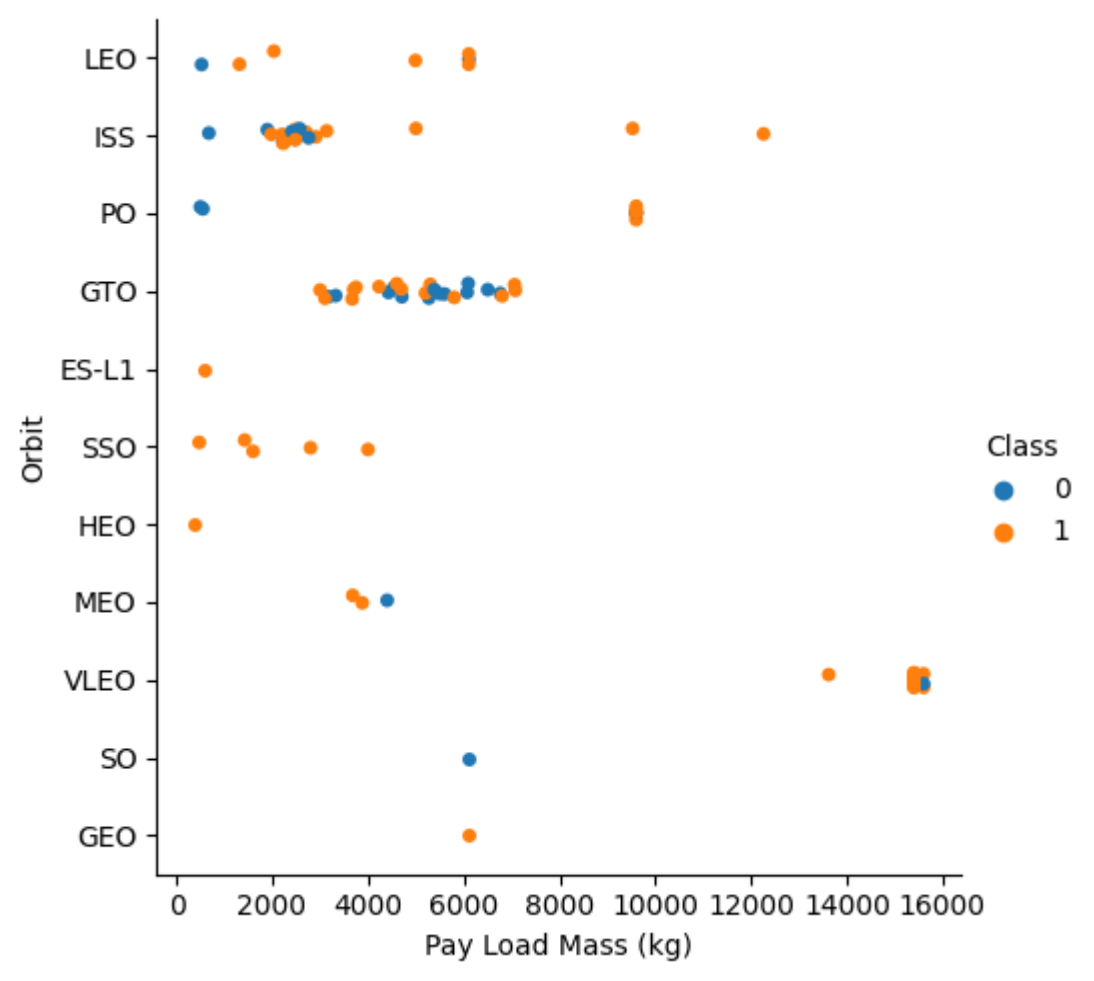
- Bar plot is used to plot Orbit vs the mean of the success landing outcome
- Orbits ES-L1/GEO/HEO and SSO have high success rate

EDA with Data Visualization



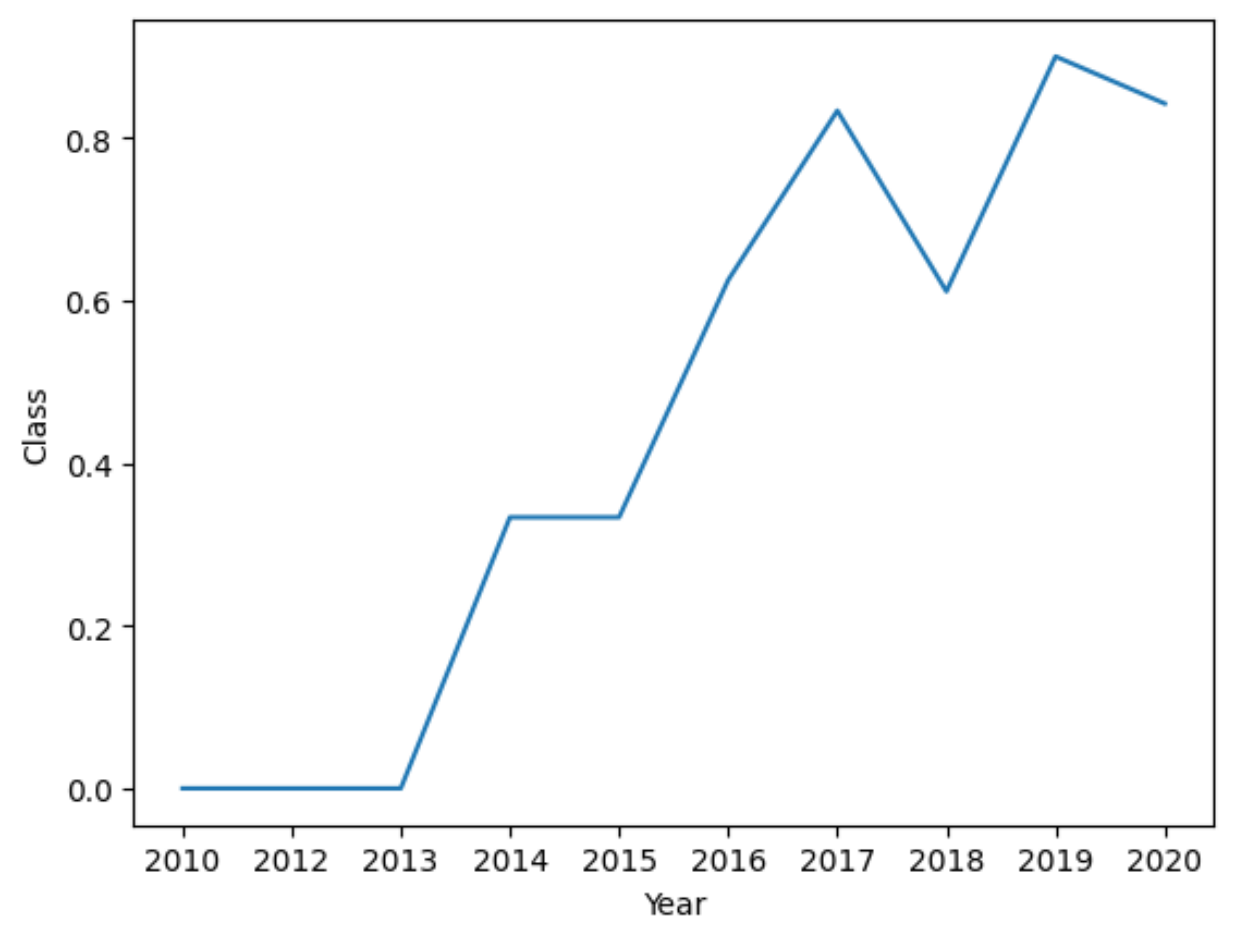
- Categorical plot is used to plot Flight number vs Orbit and overlay the outcome of the launch
- For LEO, the success increased with flight number
- For GTO, there is no concrete relationship between flight number

EDA with Data Visualization



- Categorical plot is used to plot Pay Load Mass vs Orbit and overlay the outcome of the launch
- With heavy payloads, the successful landing are more for LEO/PO & ISS
- For GTO, unable to distinguish as both successful and unsuccessful landing are scattered near to each other

EDA with Data Visualization

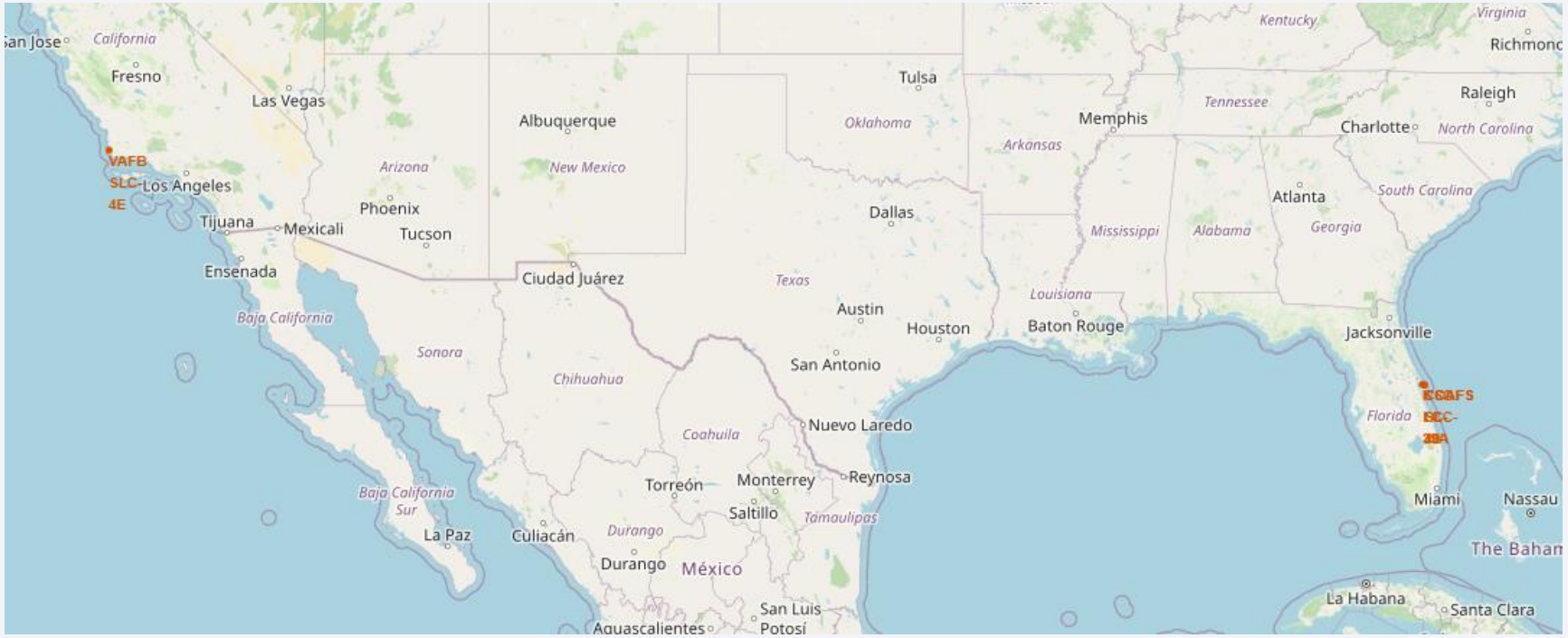


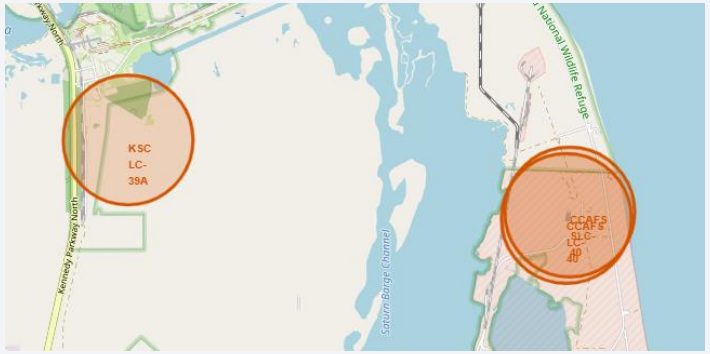
- Line plot is used to plot Year vs the mean of the success landing outcome
- The success rate increased from Year 2013 till 2017 (stable in 2014, started to increase in Year 2015).

EDA with SQL

- SQL queries performed
  - Display names of the unique launches sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - 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

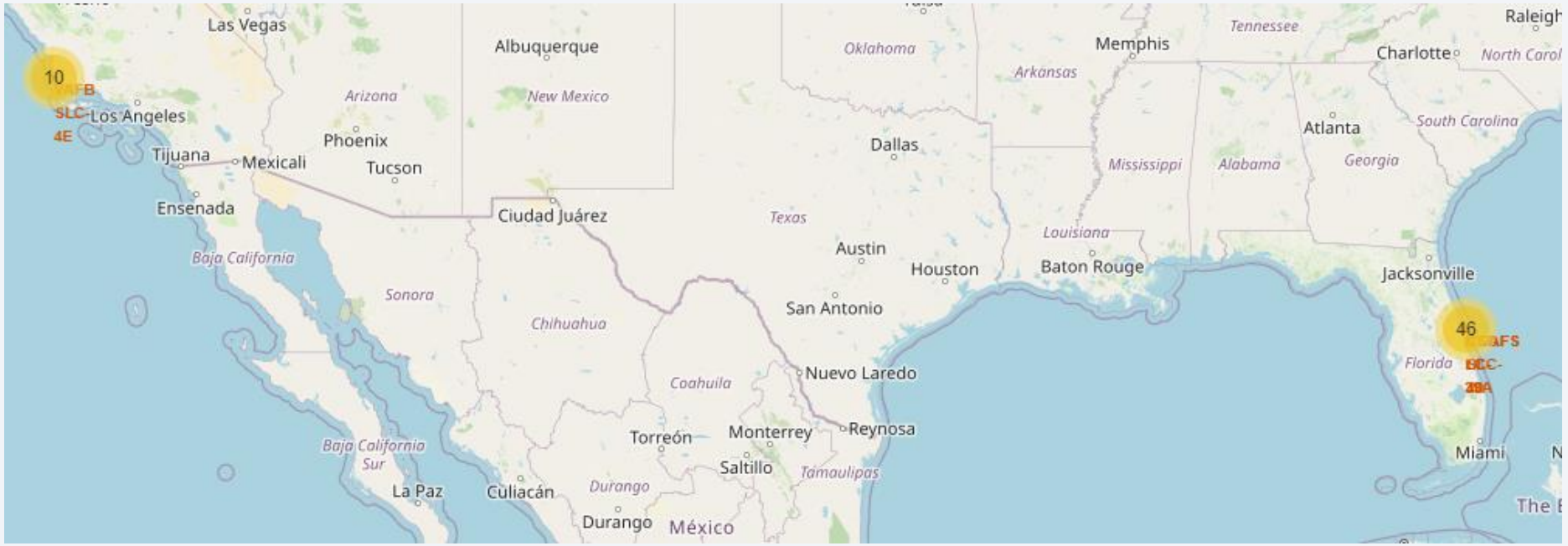
Build an Interactive Map with Folium

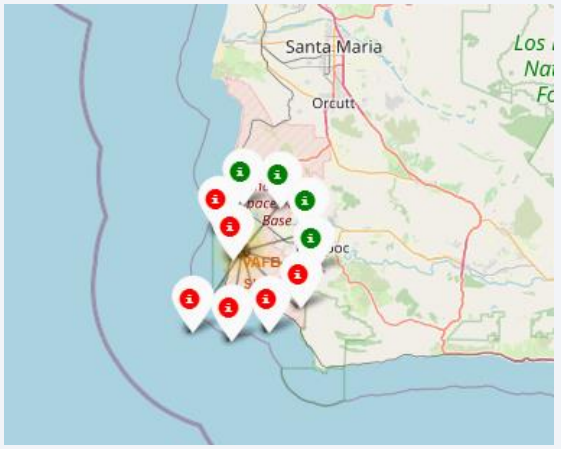


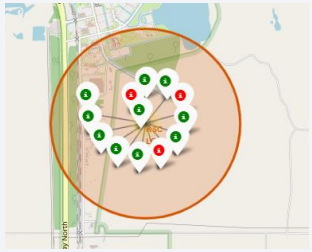


- Using circle to indicate the coordinates of launch sites
- Using text label as a marker for the name of the launch sites

Build an Interactive Map with Folium

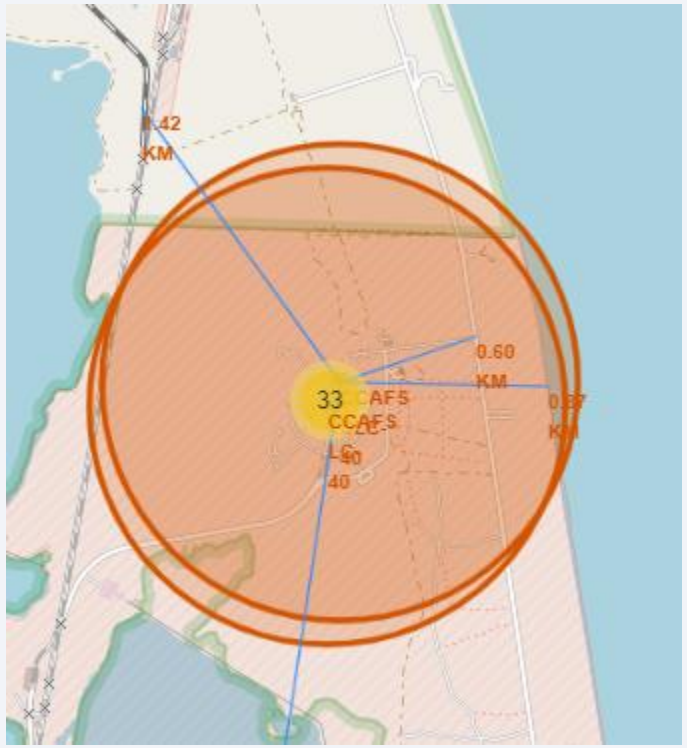


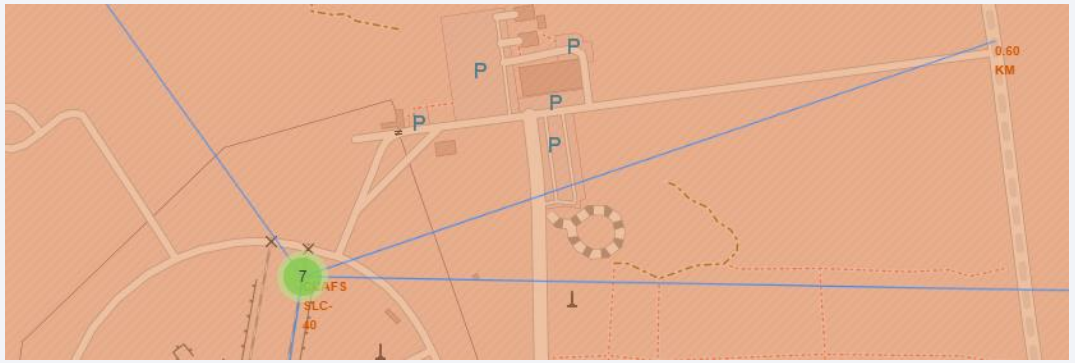


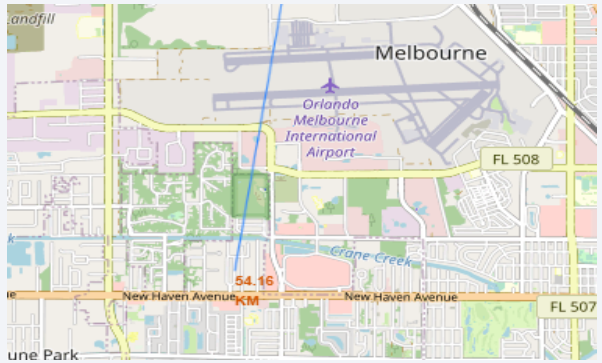


- Enhanced map by adding launch outcomes for each site (red marker – Unsuccessful, Green marker – Successful)
- Using marker cluster to indicate the green/red marker on each of the sites

Build an Interactive Map with Folium

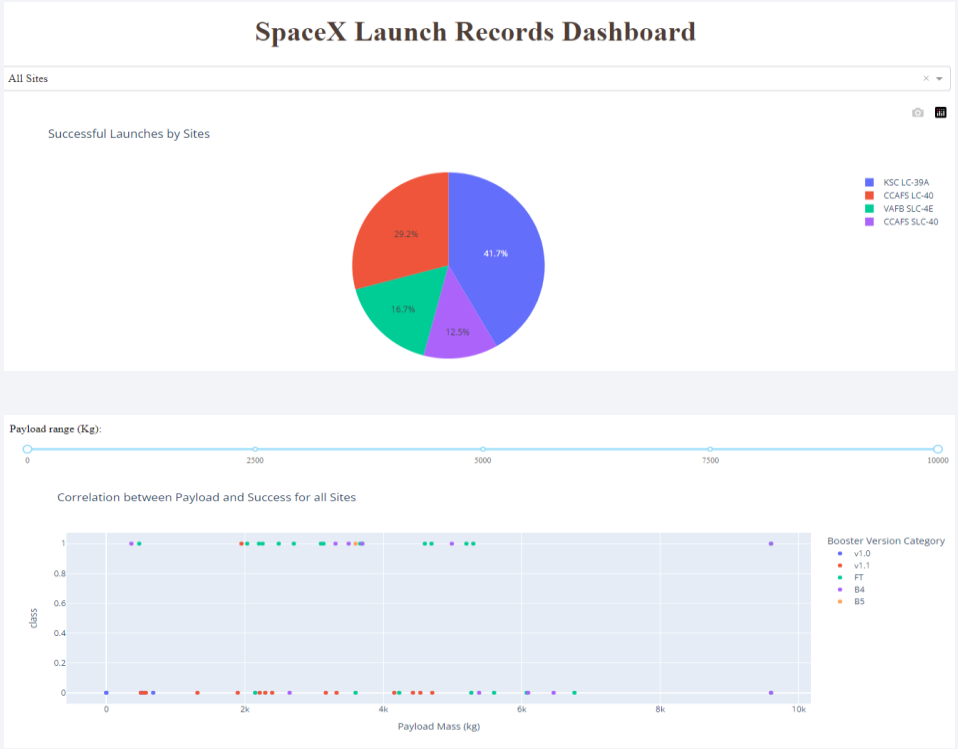


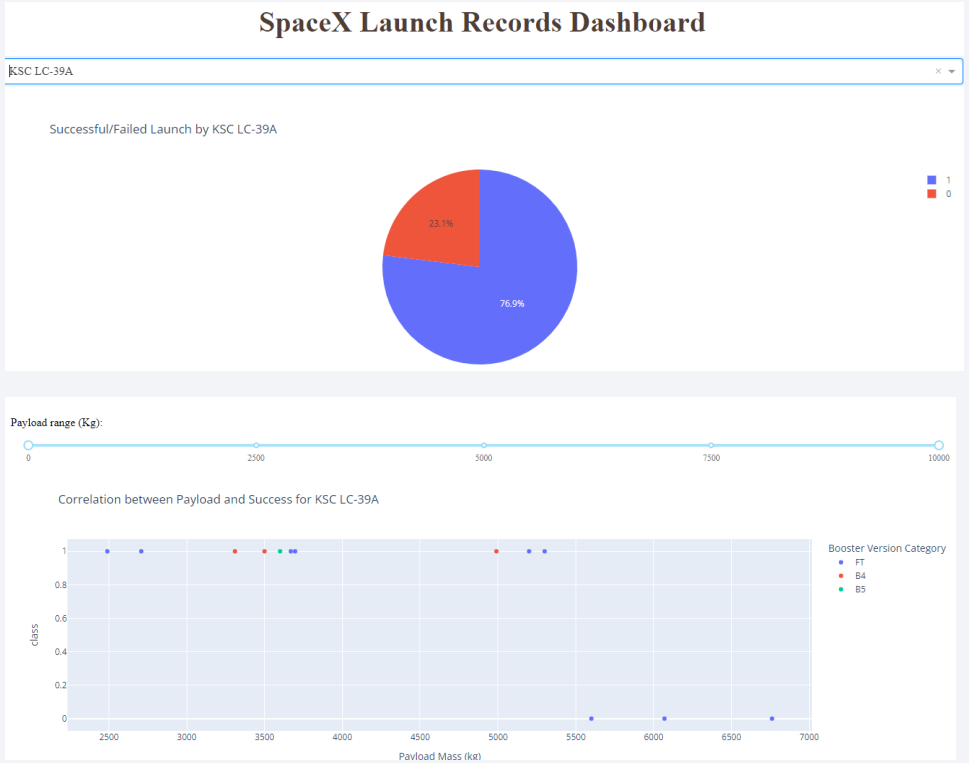




- Further explore and analyze the proximities of launch sites
- Calculation of distances from the launch site CCAFS SLC-40 to nearest highway/ railway/coastline as well as city
- Using Mouse position to highlight the coordinates of the nearest highway/railway/coastline as well as city
- Marker used on the highway/railway/coastline as well as city will be the text showing the distance
- Line is drawn from the launch sites using ‘Polyline’
- Distance from launch site to coastline is approx. 0.87km
- Distance from launch site to railway is approx. 1.42km
- Distance from launch site to highway is approx. 0.6km
- Distance from launch site to city Melbourne is approx. 54.16km

Build a Dashboard with Plotly Dash

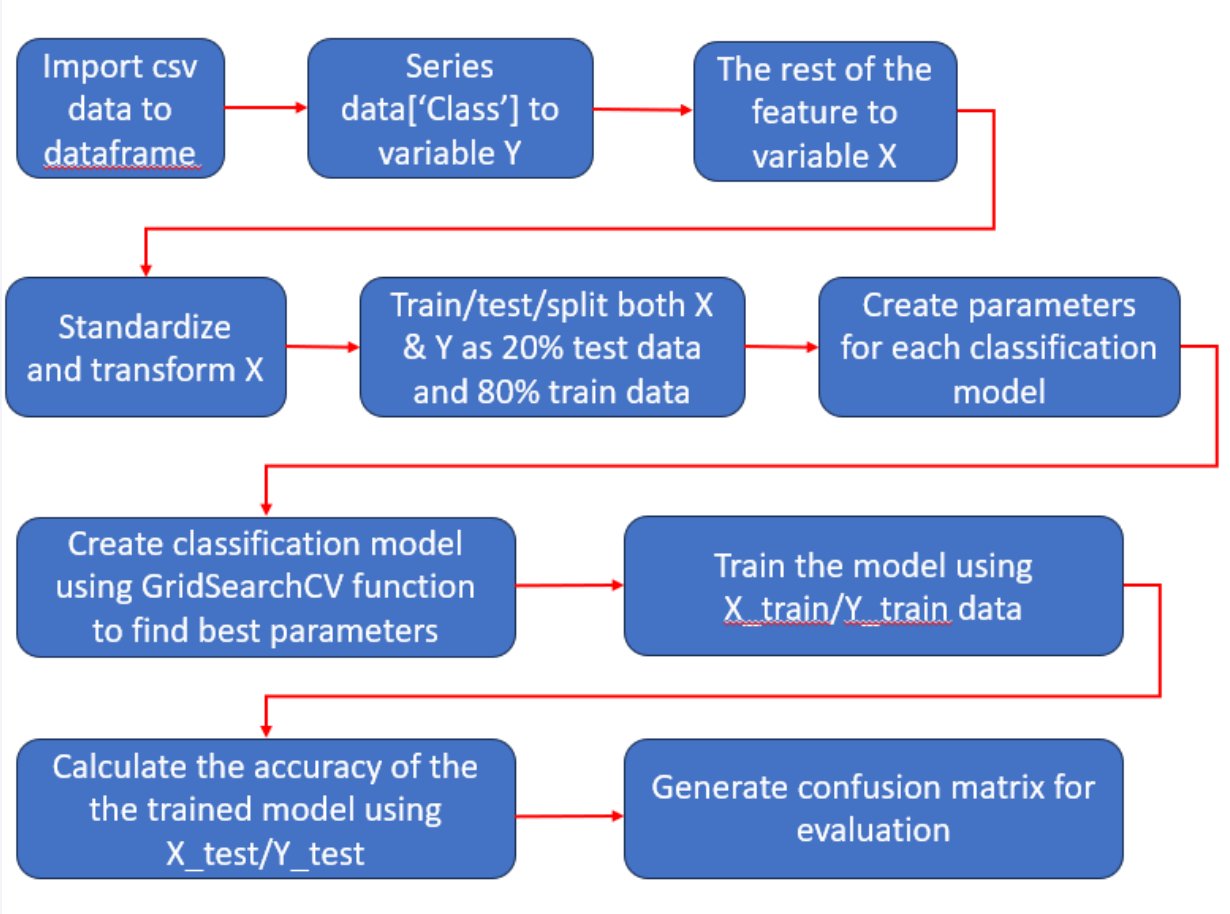




- A pie chart which show the total successful launches for all sites and if a specific launch sites was selected, it show the Success vs Failed counts for the site
- A scatter chart to show correlation between the payload and launch success.
- This interaction dashboard will helps to understand the successful launch for each sites as well as correlation of payload mass with the launches.

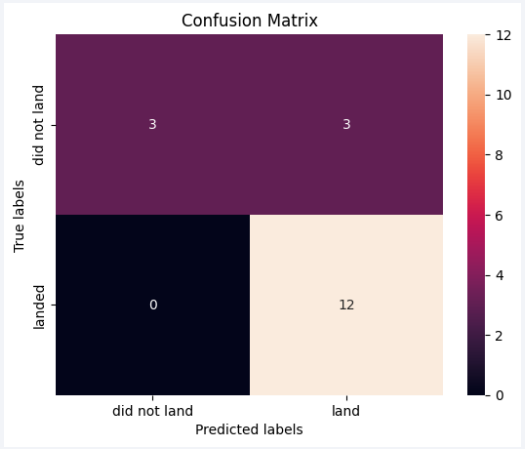
Predictive Analysis (Classification)

- Import the data to dataframe and assign the column ‘class’ to variable Y. The rest of the features assign to variable X.
- Normalized the X variable following by using train_test_split function to X_train/X_test/Y_train/Y_test
- Create parameters for each classification model & use GridSearchCV function to find the best parameter
- Train the model using X_train/Y_train
- Evaluate the model by calculating the accuracy of the model using X_test/Y_test and plot confusion matrix

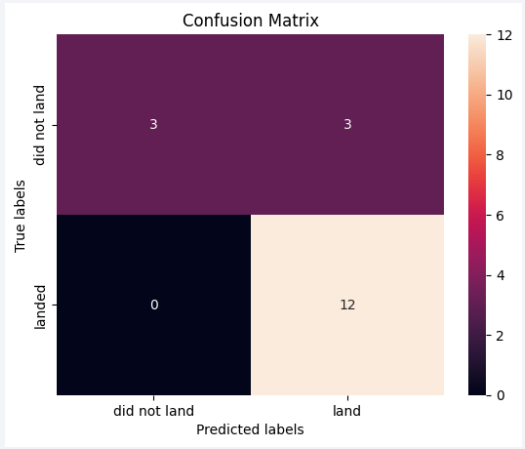


Predictive Analysis (Classification)

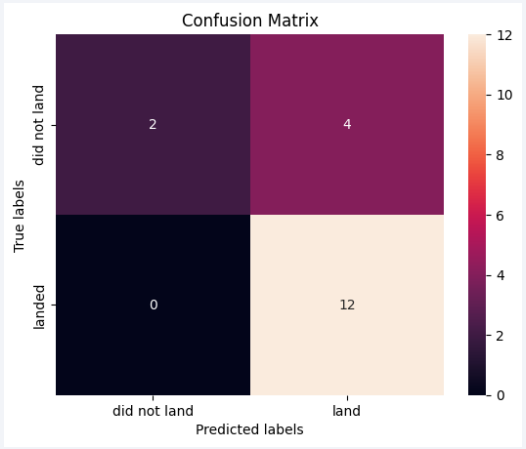
- Accuracy for Logistics Regression method: 0.833
- Accuracy for Support Vector Machine method: 0.833
- Accuracy for Decision tree method: 0.777
- Accuracy for K nearest neighbors method: 0.8333



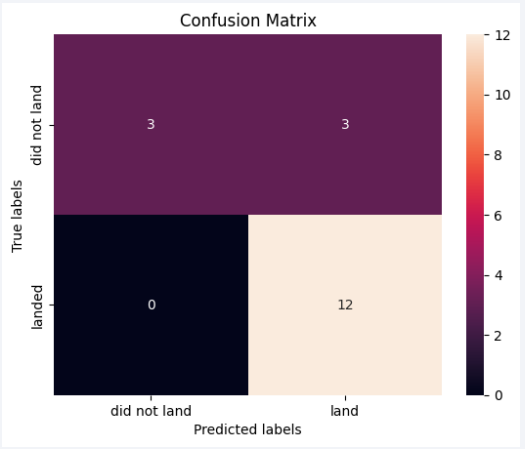
Logistic Regression



Support Vector Machine



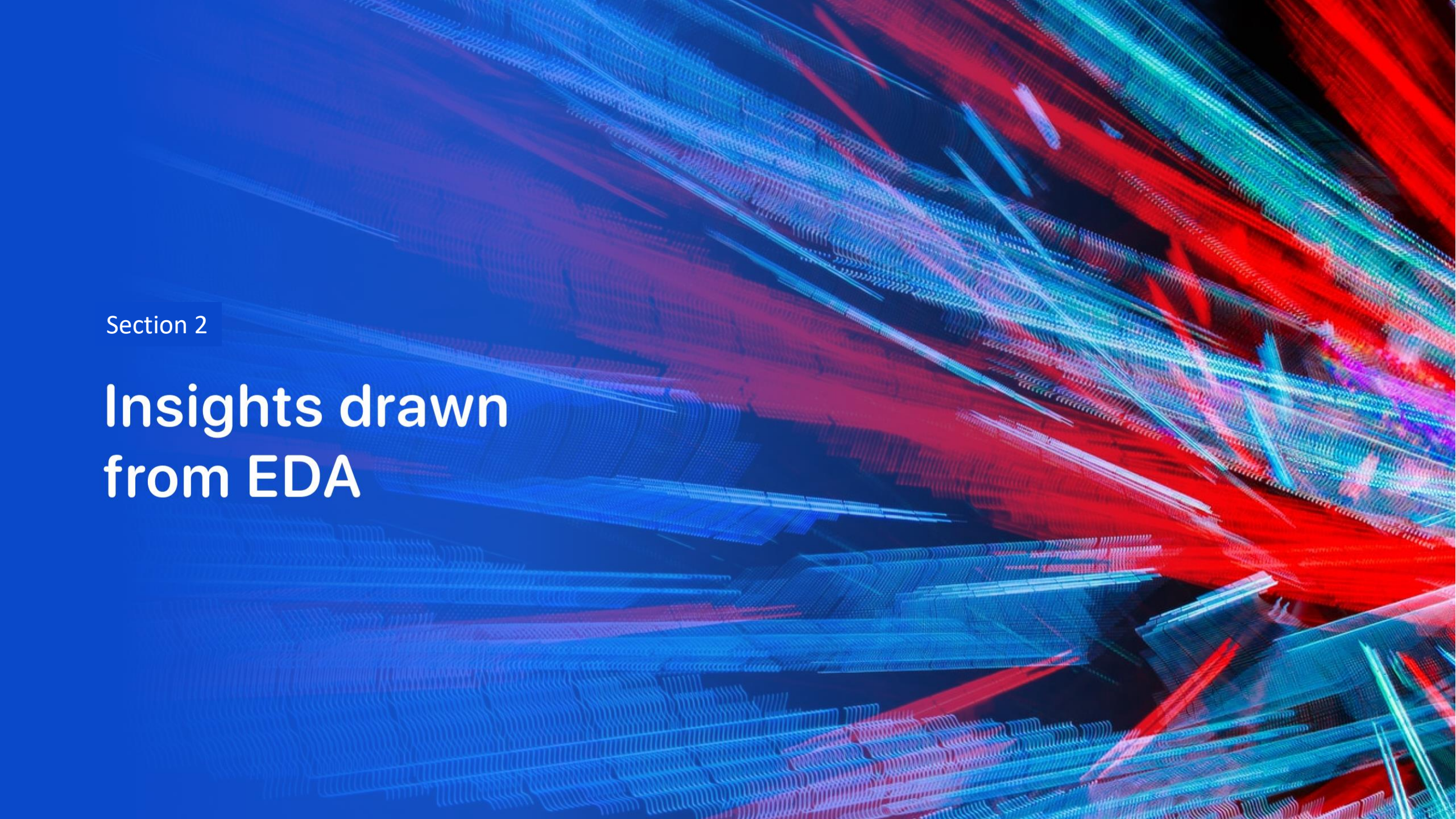
Decision Tree



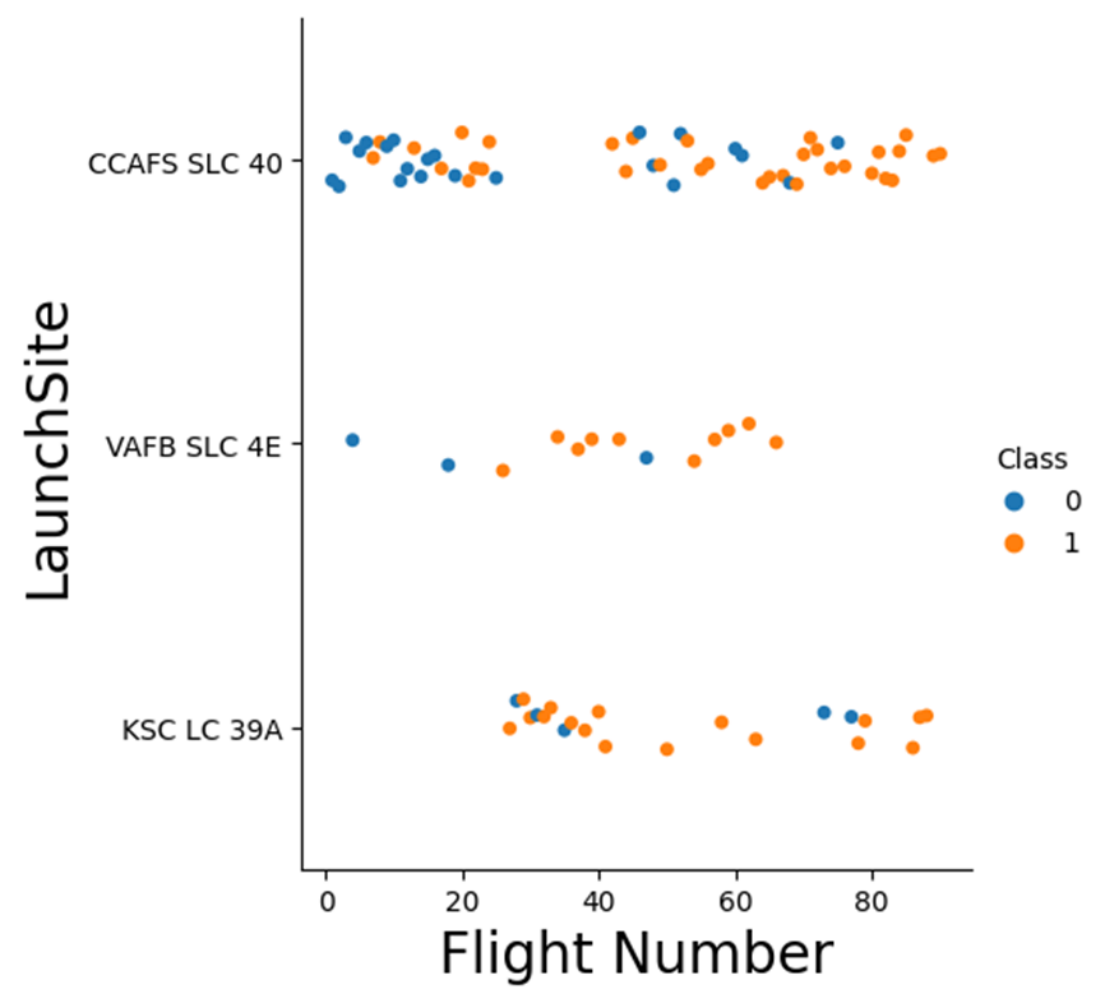
KNN

Results

- The success rate for SpaceX improved/increased from year 2013 till 2017
- With increasing payload mass, the less likely the 1st stage will land successfully
- Orbits ES-L1/GEO/HEO and SSO had high success rate
- KSC LC 39A had the most successful launches
- The Logistic Regression/Support Vector Machine/KNN have the best accuracy result of 83% on test dataset.

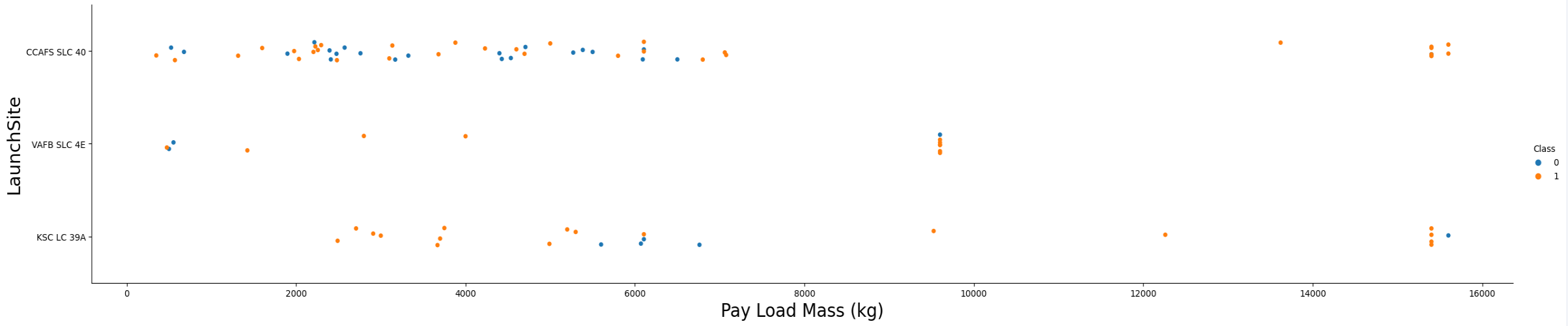

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



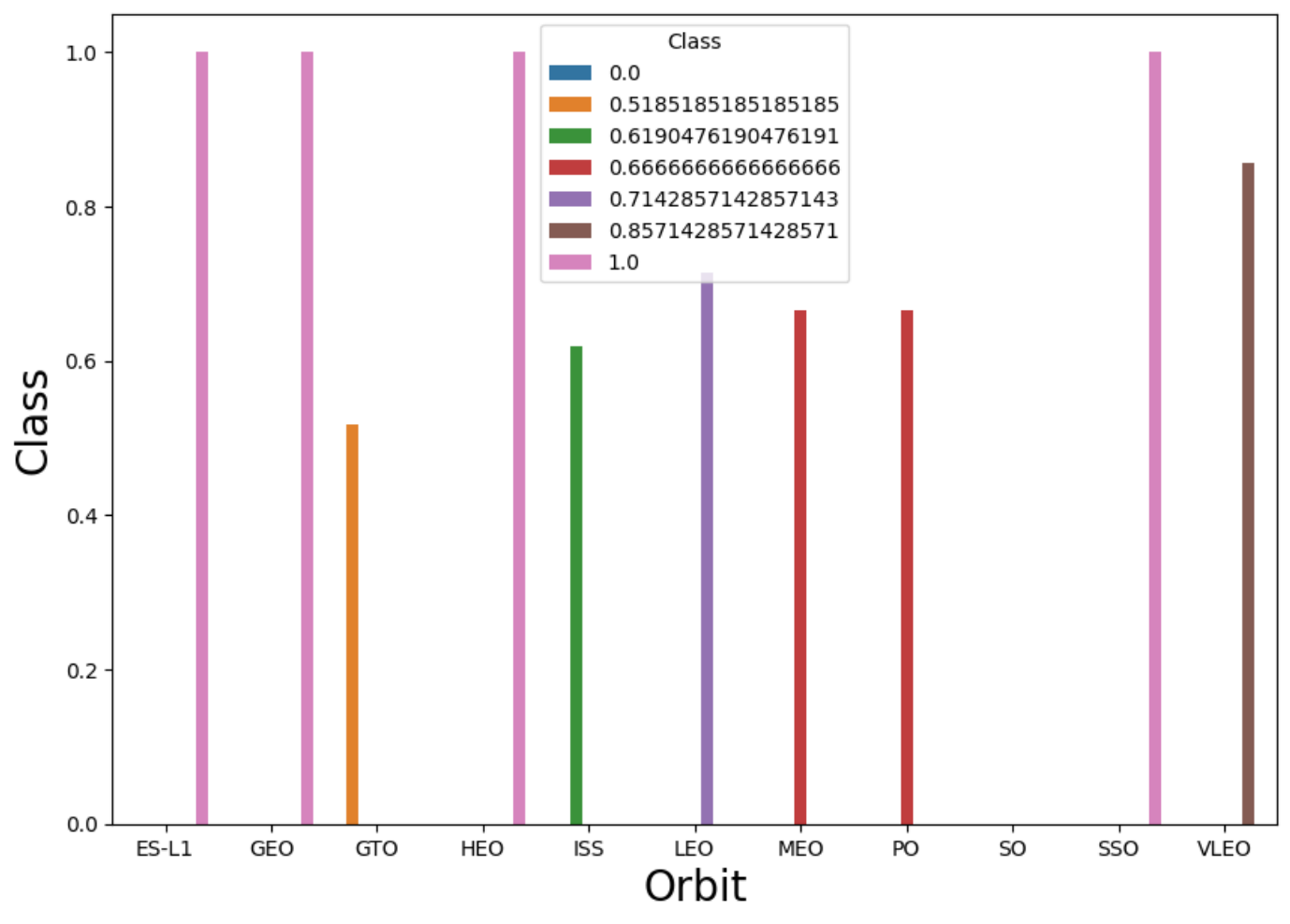
- KSC LC 39A seemed to have a higher successfully landing as compare to the rest of the sites
- For CCAFS SLC 40, the success rate increased with more flight number
- CCAFS SLC 40 has more launches as compared to other sites

Payload vs. Launch Site



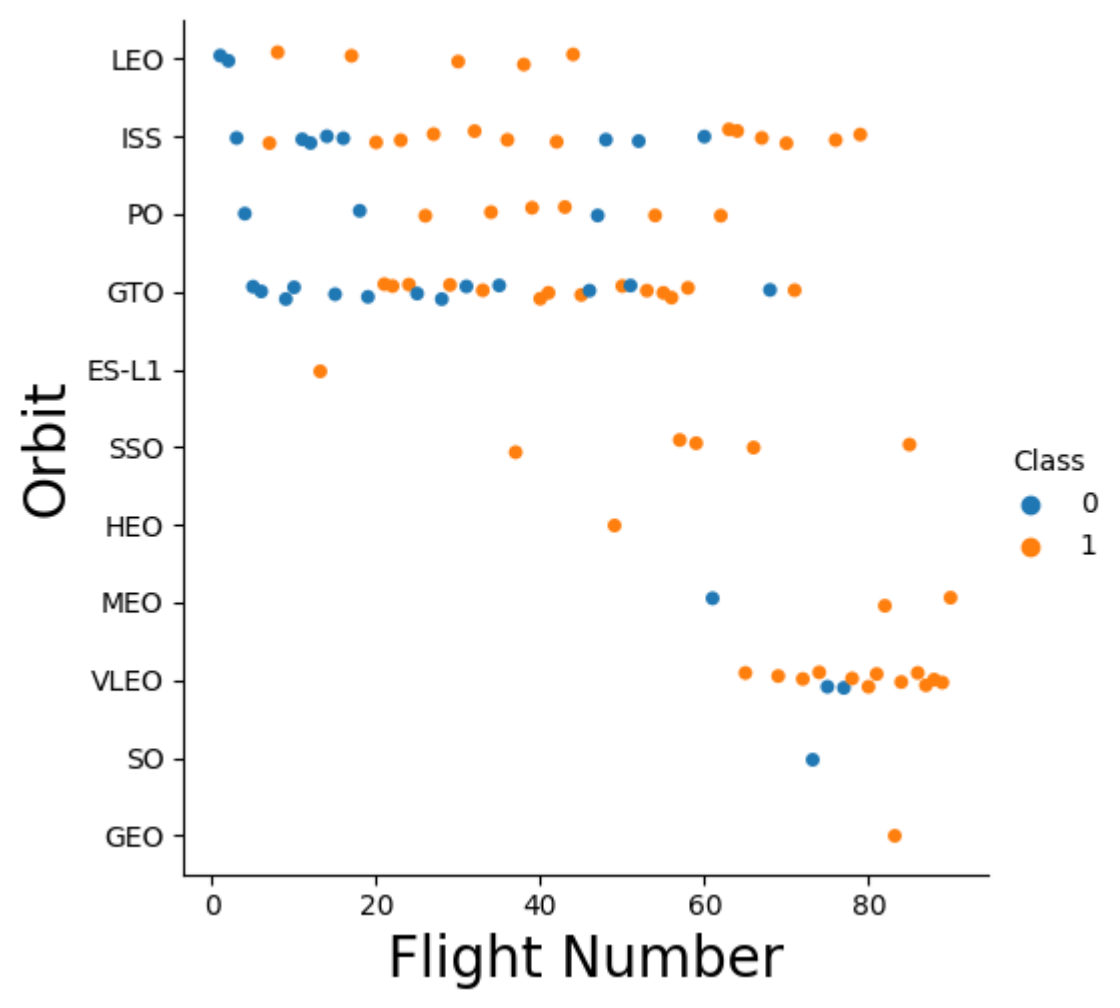
- There are no launch for VAFB-SLC with payload above 10000 kg
- More lower payload mass launched from CCAFS SLC40
- Higher rate of successful launches with payload mass of 9000 kg and above in all sites

Success Rate vs. Orbit Type



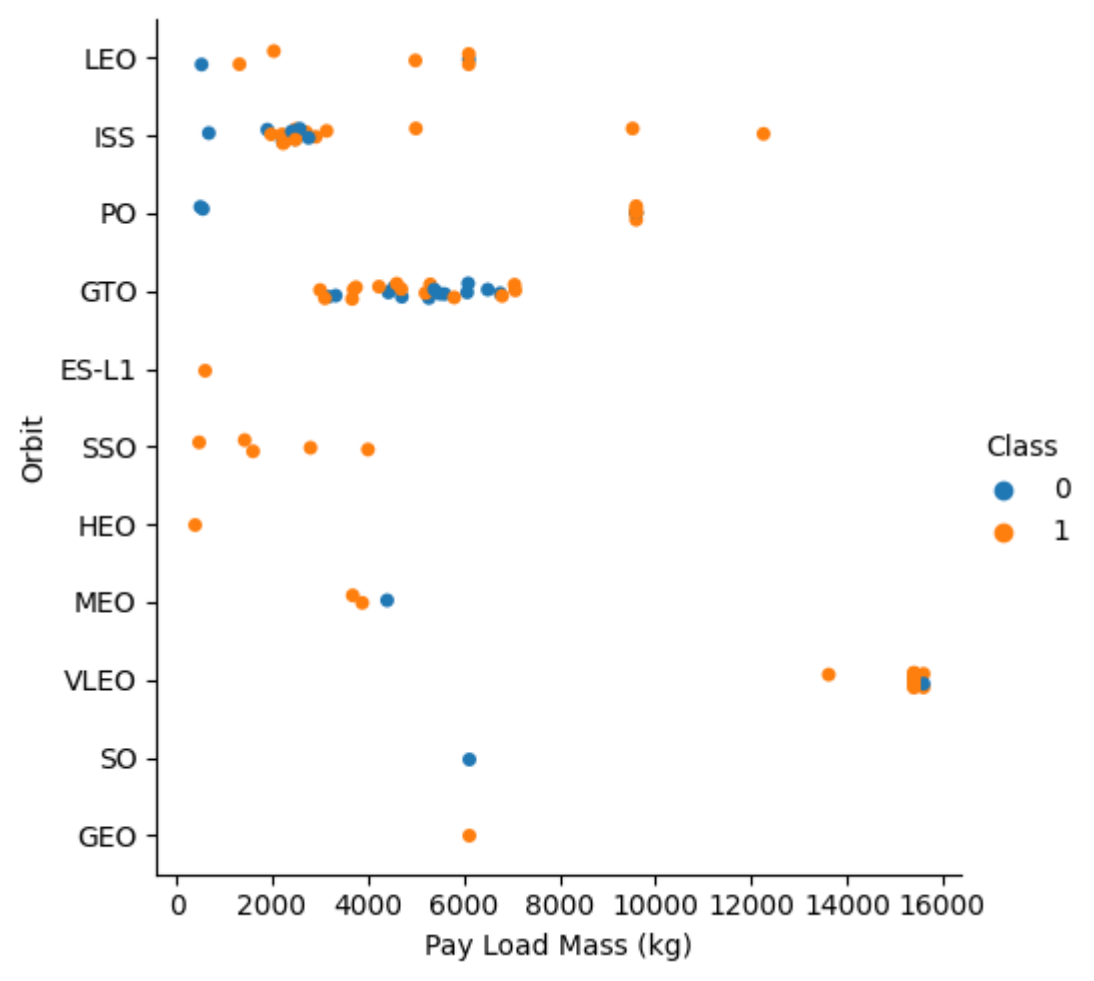
- Orbits ES-L1/GEO/HEO and SSO have high success rate

Flight Number vs. Orbit Type



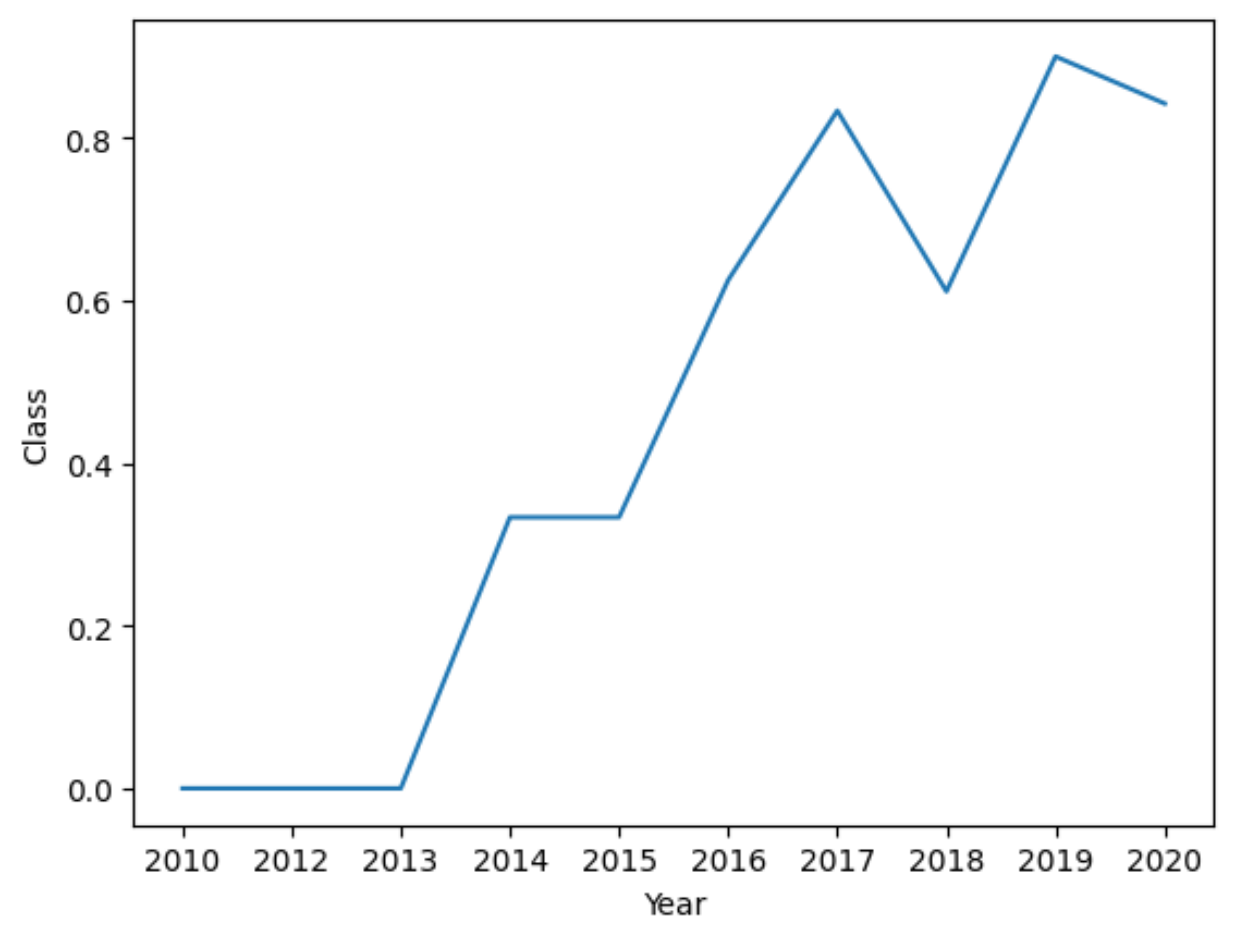
- For LEO, the success increased with flight number
- For GTO, there is no concrete relationship between flight number

Payload vs. Orbit Type



- With heavy payloads, the successful landing are more for LEO/PO & ISS
- For GTO, unable to distinguish as both successful and unsuccessful landing are scattered near to each other
- SSO has 100% successful landing

Launch Success Yearly Trend



- The success rate increased from Year 2013 till 2017 (stable in 2014, started to increase in Year 2015).

All Launch Site Names

- Find the names of the unique launch sites

```
%sql select DISTINCT Launch_Site from SPACEXTABLE
```

```
 * 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'

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

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' LIMIT 5
```

```
 * 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 00:00:00	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12 00:00:00	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 00:00:00	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10 00:00:00	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03 00:00:00	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer == 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

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 SPACEXTABLE where Booster_Version == 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome == 'Success (ground pad)'
```

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

min(Date)
2015-12-22 00:00:00

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the 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 SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select count(Mission_outcome), Mission_outcome from SPACEXTABLE group by Mission_outcome
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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

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

```
%sql select substr(Date, 6, 2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5) = '2015' and Landing_Outcome like '%Failure
```

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

substr(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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, Date from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' order by Date DESC
```

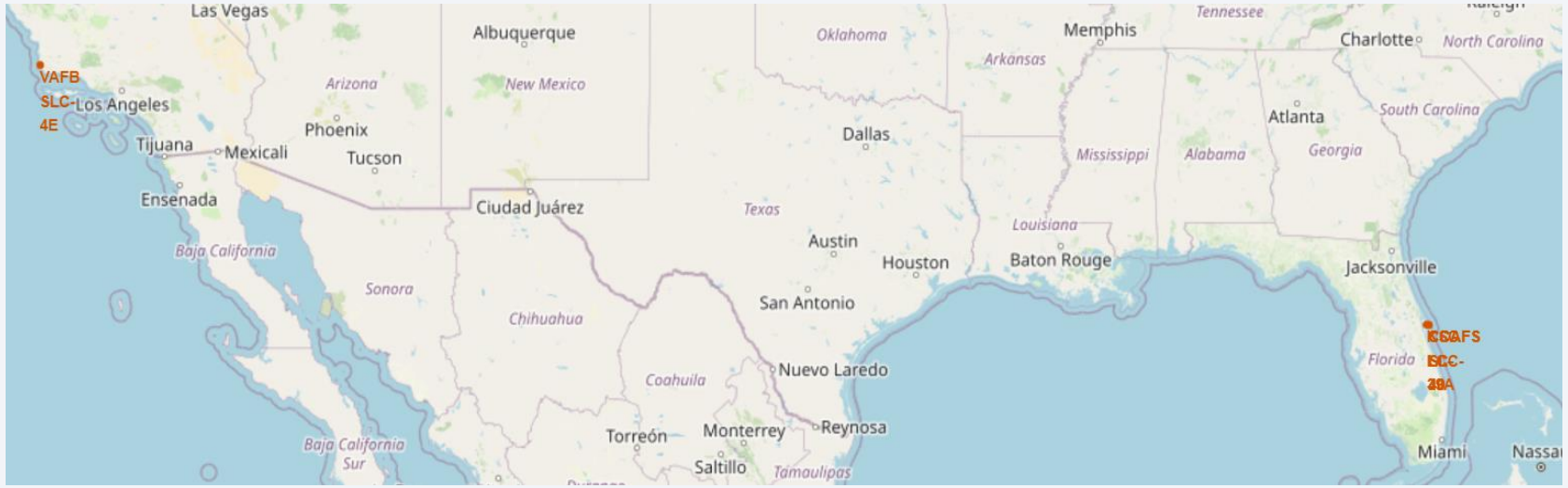
```
 * sqlite:///my_data1.db
Done.
```

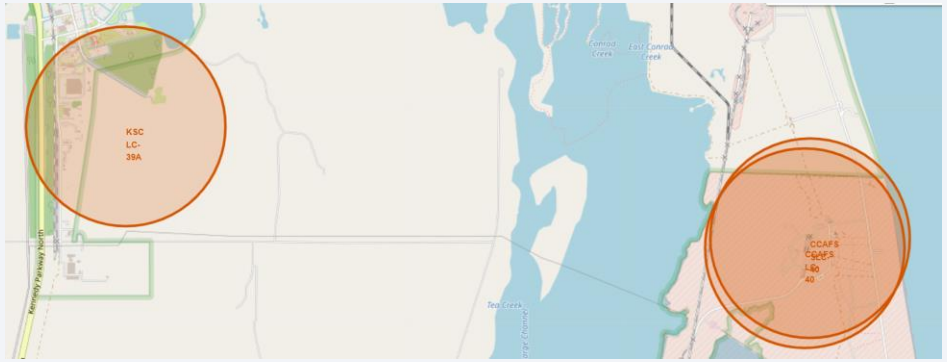
Landing_Outcome	Date
No attempt	2017-03-16 00:00:00
Success (ground pad)	2017-03-06 00:00:00
Success (ground pad)	2017-02-19 00:00:00
Success (drone ship)	2017-01-14 00:00:00
Success (ground pad)	2017-01-05 00:00:00
Success (drone ship)	2016-08-14 00:00:00
Success (drone ship)	2016-08-04 00:00:00
Success (ground pad)	2016-07-18 00:00:00
Failure (drone ship)	2016-06-15 00:00:00
Success (drone ship)	2016-06-05 00:00:00
Success (drone ship)	2016-05-27 00:00:00
Failure (drone ship)	2016-04-03 00:00:00
Failure (drone ship)	2016-01-17 00:00:00
Success (ground pad)	2015-12-22 00:00:00
Controlled (ocean)	2015-11-02 00:00:00
Failure (drone ship)	2015-10-01 00:00:00
Precluded (drone ship)	2015-06-28 00:00:00
No attempt	2015-04-27 00:00:00
Failure (drone ship)	2015-04-14 00:00:00
No attempt	2015-02-03 00:00:00
Uncontrolled (ocean)	2014-09-21 00:00:00
Controlled (ocean)	2014-07-14 00:00:00
No attempt	2014-07-09 00:00:00
No attempt	2014-06-01 00:00:00
No attempt	2014-05-08 00:00:00
Controlled (ocean)	2014-04-18 00:00:00
Uncontrolled (ocean)	2013-09-29 00:00:00
No attempt	2013-03-12 00:00:00
No attempt	2013-01-03 00:00:00
No attempt	2012-08-10 00:00:00
No attempt	2012-05-22 00:00:00
Failure (parachute)	2010-08-12 00:00:00

Section 3

Launch Sites Proximities Analysis

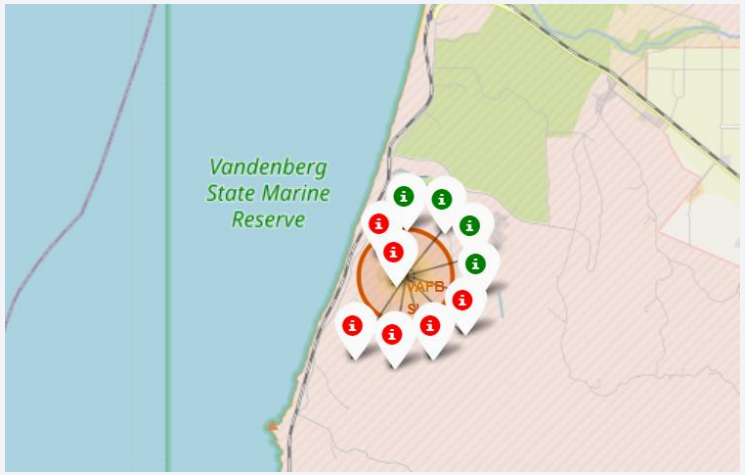
Folium Map SpaceX Launch Location



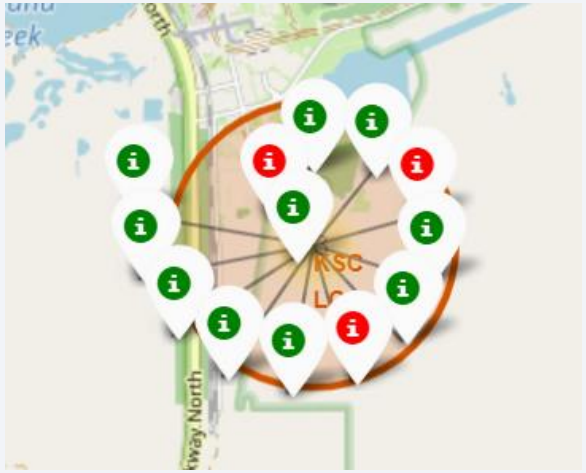


- All 4 launch sites are situated near to the coastline and ocean
- Launch site VAFB in the western part of US and launch sites KSC LC-39A/CCAFS LC-40 and CCAFS SLC-40 in eastern part of US

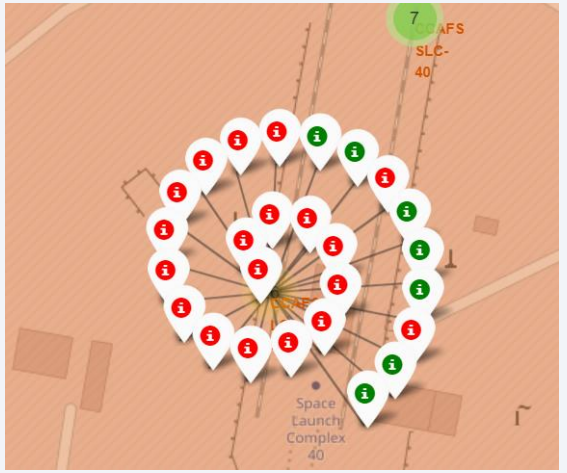
Folium Map SpaceX Launch Site Outcome



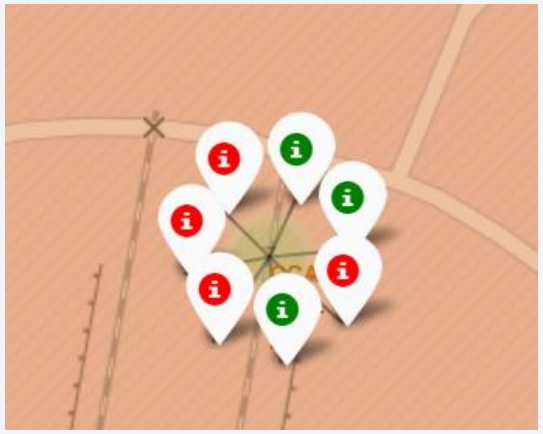
VAFB SLC-4E



KSC LC-39A



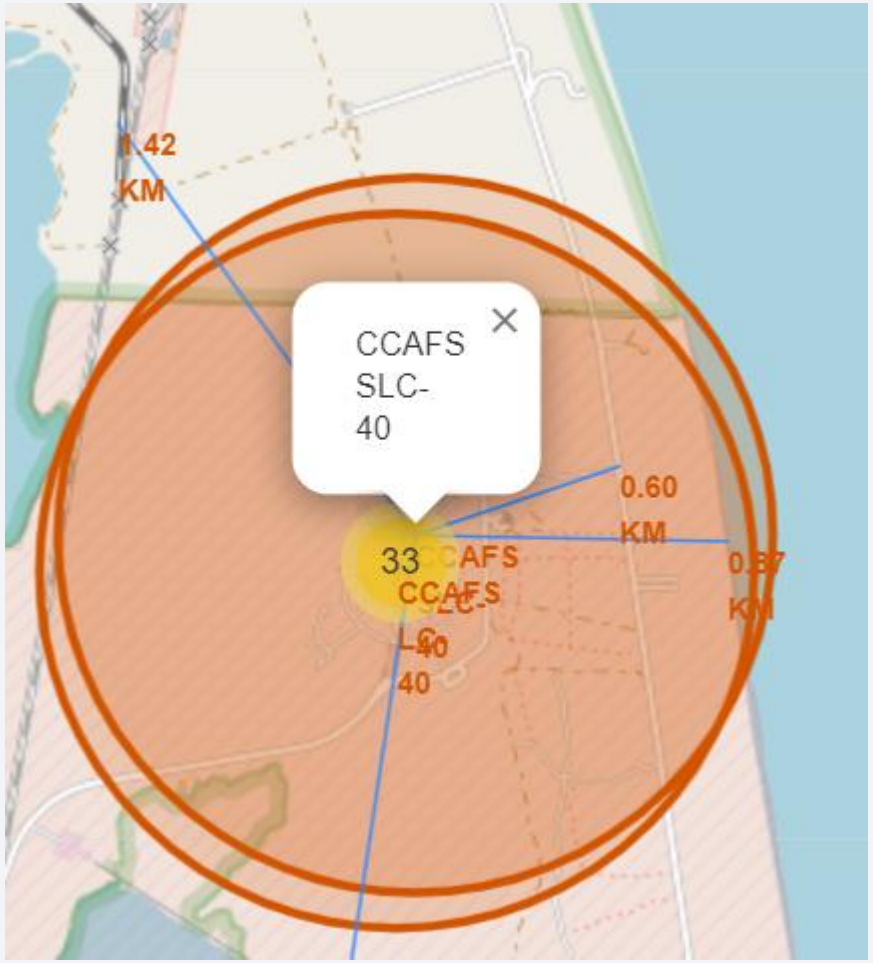
CCAFS LC-40



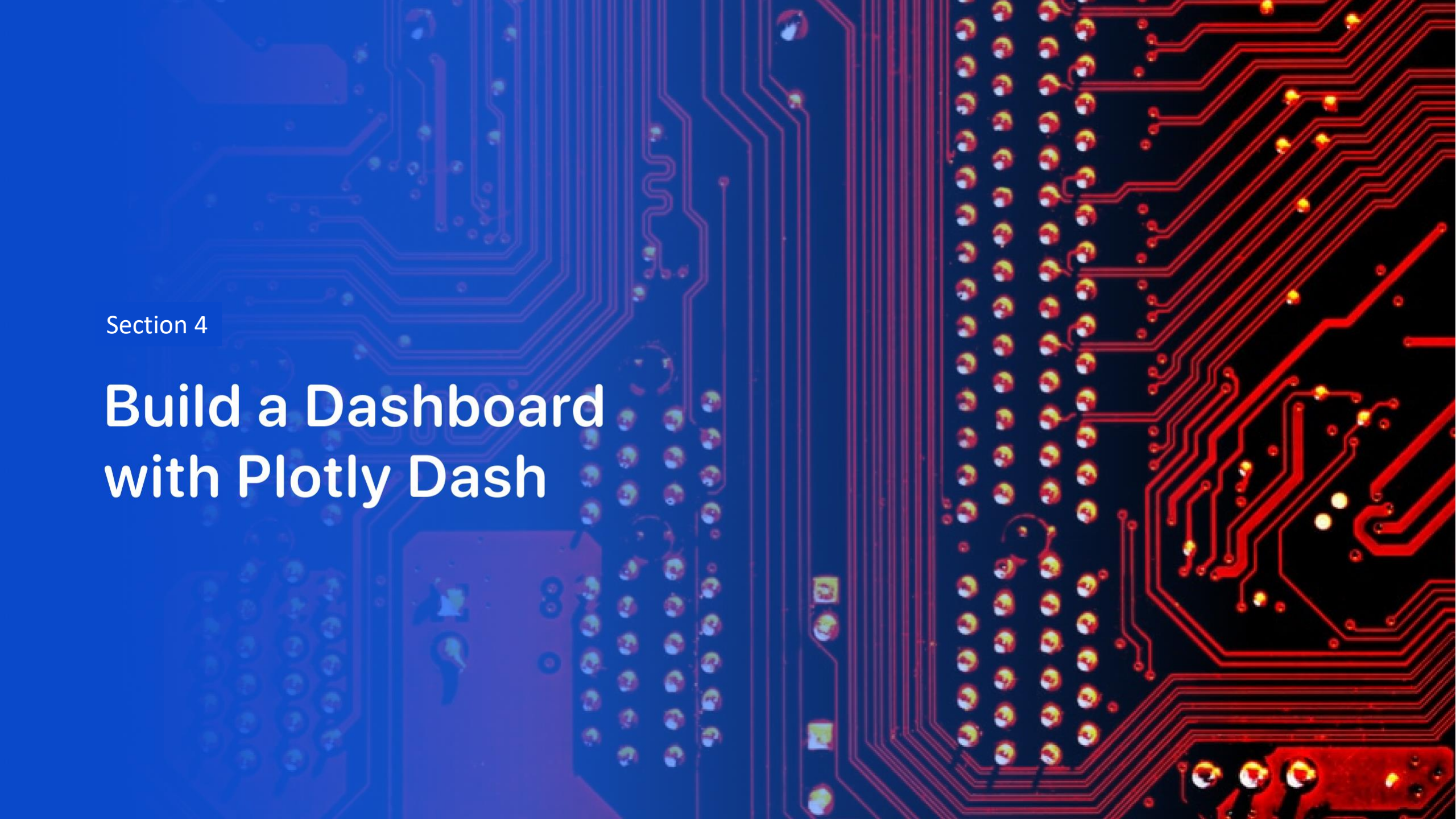
CCAFS SLC-40

- Based on the markers launch outcomes from each site, KSC LC-39A had more successful recovery of 1st stage
- CCAFS LC-40 had more launches as compared to the rest, however, the successful recovery of 1st stage was low

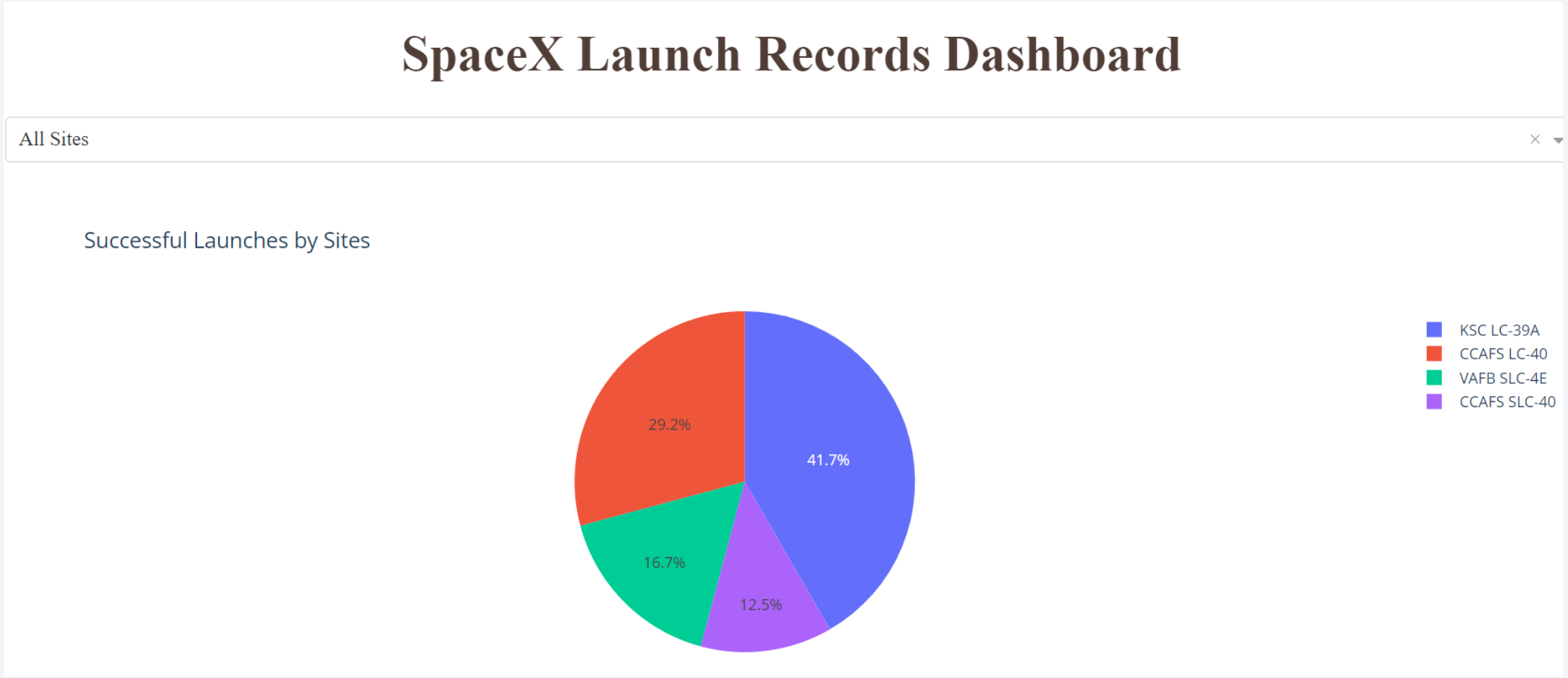
Folium Map Launch Site CCAFS SLC-40 Proximity



- The distance from the site to the nearest coastline is approx. 0.87km
- The distance from the site to the nearest highway is approx. 0.6km
- The distance from the site to the nearest railway is approx. 1.42km

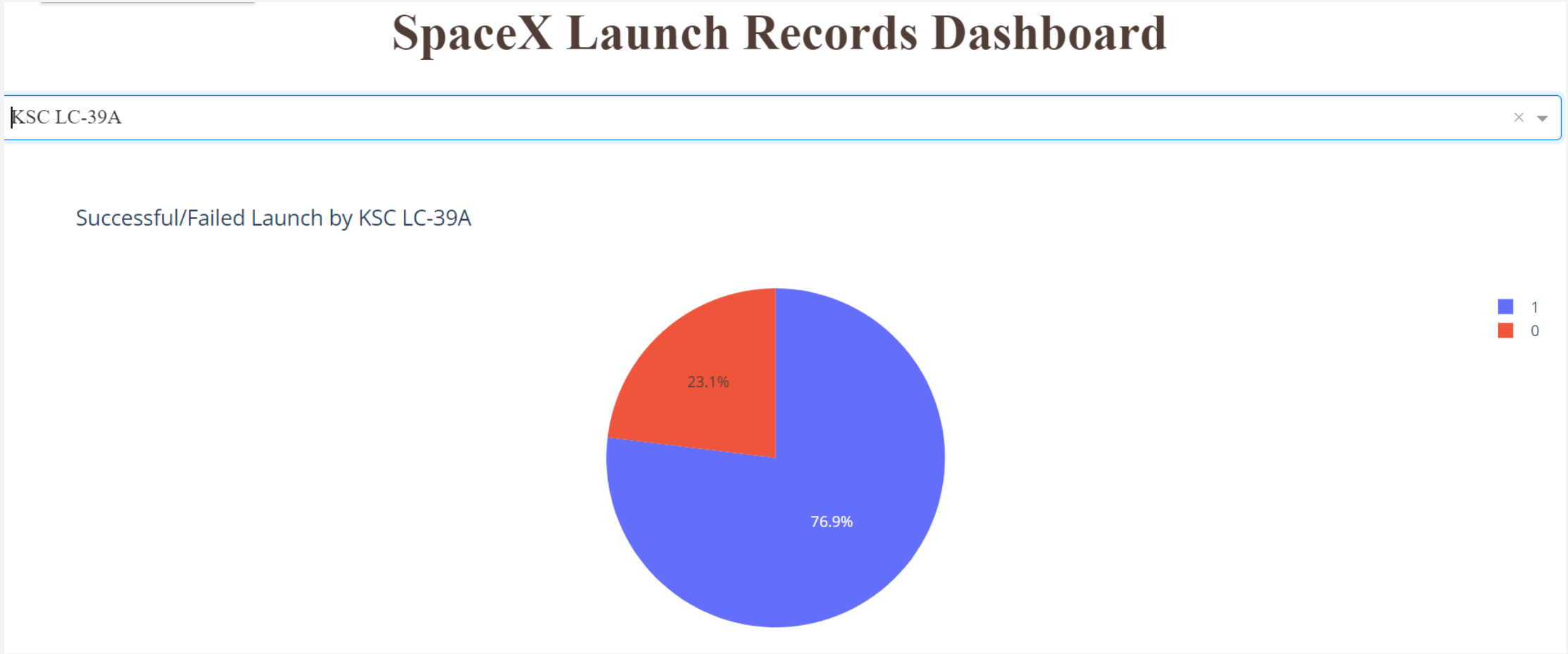

Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



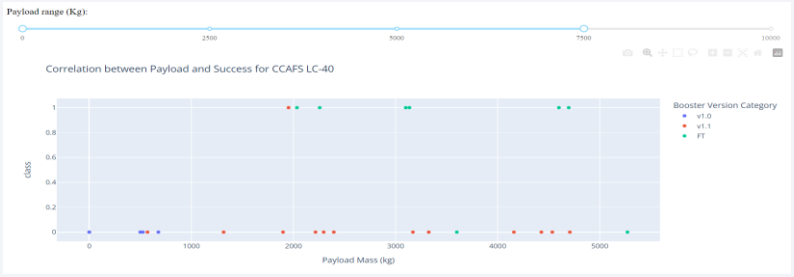
- Based on the piechart, KSC LC-39A had the highest ratio of successful recovery of 1st stage as compared to the rest. (41.7%)

SpaceX Launch Records Dashboard KSC LC-39A



- Successful recovery of 1st stage from site KSC LC-39A – 76.9%
- Unsuccessful recovery of 1st stage from site KSC LC-39A – 23.1%

SpaceX Launch Records Dashboard Payload vs Launch Outcome

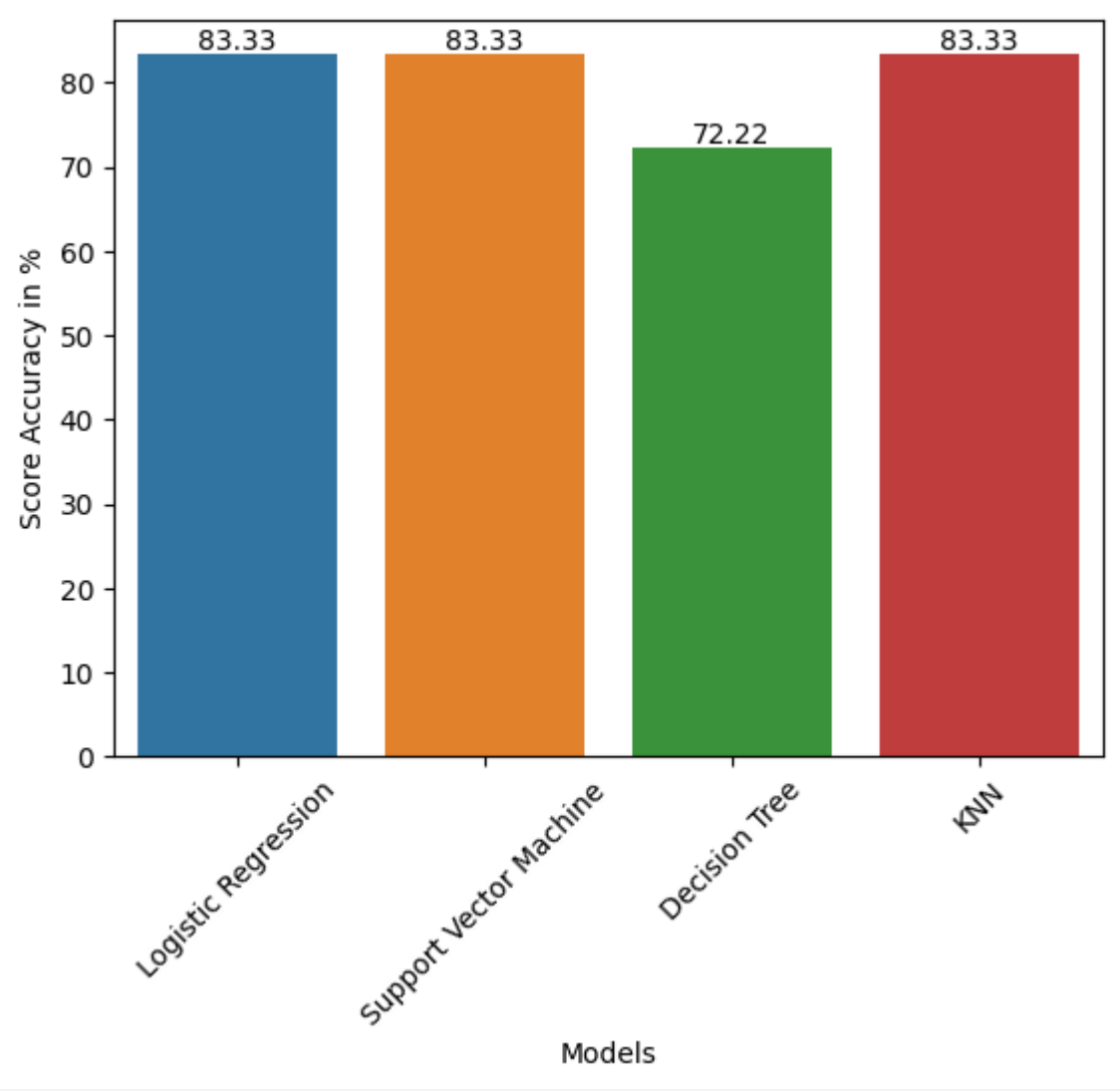


- Based on the scatter plot above, Booster Version Cat FT had a higher successful recovery of 1st stage between payload mass of 2034Kg and 5300Kg

Section 5

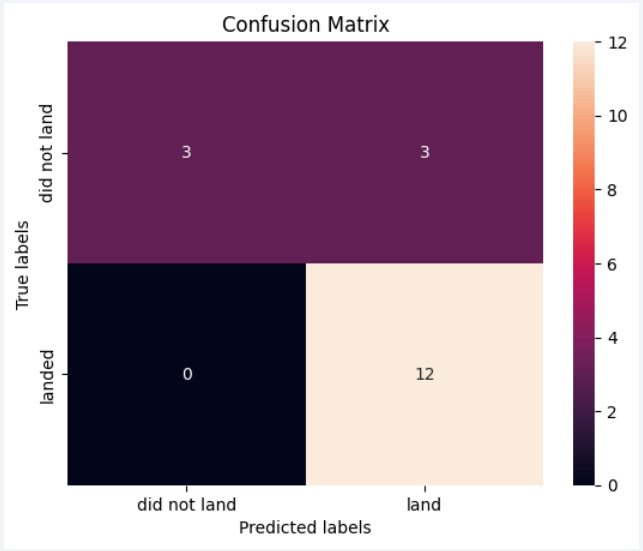
Predictive Analysis (Classification)

Classification Accuracy

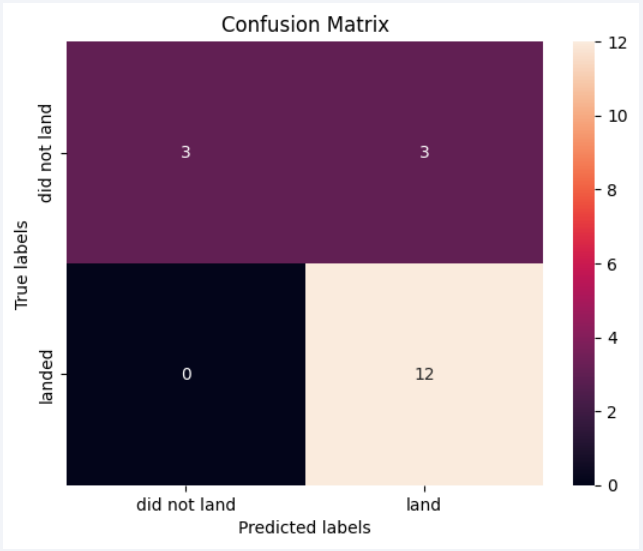


- Logistic Regression/ Support Vector Machine and KNN have the same high accuracy of score 83.33%

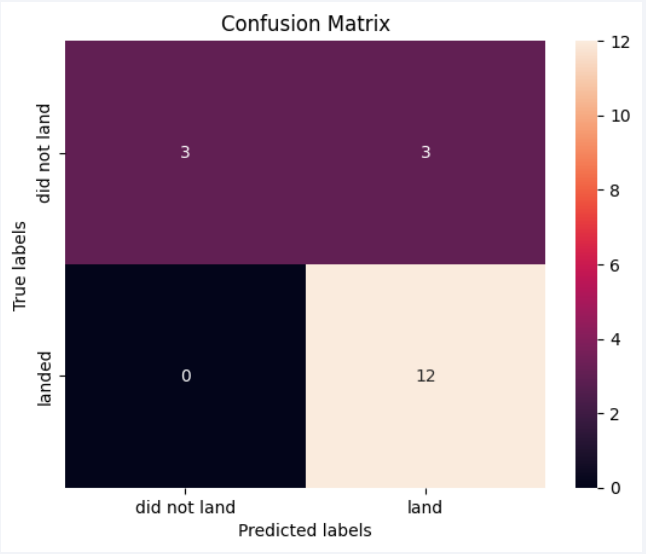
Confusion Matrix



KNN



Support Vector Machine



Logistic Regression

- All the 3 models above score a same accuracy of 83.33%
- The confusion matrix for this 3 models are identical also
- There are 3 data rows on the X_test dataset were predicted wrongly (false positive)
- These 3 data rows true class is “did not land” but yet predicted “land”.

Conclusions

- The launch sites are situated near to the coastline/highway and railway. However, they are far away from the cities area.
- The success rate increased from Year 2013 till 2017 (stable in 2014, started to increase in Year 2015) based on the plot Year vs the Mean of the success landing outcome. This success rate will keep increasing with more launches over the years.
- Rocket launches for Orbits ES-L1/GEO/HEO and SSO have high success rate of recovery the 1st stage
- Data shows that launch site “KSC LC 39A” had the most successful landing as compare to the rest of the sites.
- Booster Version Cat “FT” had a higher successful recovery of 1st stage between payload mass of 2034Kg and 5300Kg
- Logistic Regression/SVM/KNN models scored the same highest accuracy of 83.33%

Appendix

- Data Collection
  - https://github.com/johnloh47/testrepo/blob/main/jupyter_labs_spacex_data_collection_api.ipynb
- Webscraping
  - https://github.com/johnloh47/testrepo/blob/main/jupyter_labs_webscraping.ipynb
- Data Wrangling
  - https://github.com/johnloh47/testrepo/blob/main/labs_jupyter_spacex_Data_wrangling.ipynb
- Explore Data Analysis and Visualisation
  - https://github.com/johnloh47/testrepo/blob/main/jupyter_labs_eda_dataviz.ipynb
- Explore Data Analysis and SQL
  - https://github.com/johnloh47/testrepo/blob/main/jupyter_labs_eda_sql_coursera_sqllite.ipynb
- SpaceX Dashboard App
  - https://github.com/johnloh47/testrepo/blob/main/spacex_dash_app.py
- Folium Map Launch Sites Location
  - https://github.com/johnloh47/testrepo/blob/main/lab_jupyter_launch_site_location.ipynb
- Machine Learning Prediction
  - https://github.com/johnloh47/testrepo/blob/main/SpaceX_Machine_Learning_Prediction_Part_5_jupyterlite%20(1).ipynb

Thank you!