IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shaik Shoaib Aslam
24th Aug 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API
  - Data Wrangling
  - EDA with SQL
  - EDA with Interactive Maps and Dashboard
  - Predictive Analysis using Machine Learning
- Summary of all results
  - EDA visualizations
  - Results of prediction models

Introduction

Project background and context

SpaceX 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 SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

Problems you want to find answers

Using this information we are trying to predict if the first stage will land or not.This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The primary source of the data is the SpaceX API
- Perform data wrangling
  - Filtering required data
  - Dealing with Missing values
  - One hot encoding to convert text data to numbers for Classification
- 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, tune, evaluate classification models

Data Collection

- Raw data collected using SpaceX API and SpaceX wikipedia.
- Considered only selective columns that are useful to perform classification.
- Applied filtering and formatting on each of these columns to format and get the data that is required.
- The data is then collated into a single dataset which is further used for data preprocessing and data wrangling

Data Collection – SpaceX API

Request Data from SpaceX API and using json and normalize functions for formatting

↓

Applying custom functions to transform data to a dataframe

↓

Filtering dataframe to include only Falcon 9 launches

↓

Replacing missing values with mean of the same column

https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/Data-collection-api.ipynb

Data Collection - Scraping

Request Data from SpaceX API and using json and normalize functions for formatting

↓

Applying custom functions to transform data to a dataframe

↓

Filtering dataframe to include only Falcon 9 launches

↓

Replacing missing values with mean of the same column

https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/Webscraping.ipynb

Data Wrangling

- Performed data analysis on the created dataset and performed some calculations that help to understand the data better.
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbits
- Convert the landing outcome label to a numerical value for analysis purposes

- https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/Data%20wrangling.ipynb

EDA with Data Visualization

- Various charts were plotted to understand and get better insights on the data and relationships between the columns.
- Some of the charts plotted include
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Success Rate vs. Orbit Type
  - Payload vs. Orbit Type
  - Launch Success Yearly Trend
- Bar charts showed relationships between the categories like what is the success rate on each of the Orbit type, while scatterplot showed how trend changed over time
- https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/Data-visualization%20EDA-api.ipynb

EDA with SQL

- Using SQL, performed a deep dive into the dataset gaining insights and facts that help us to understand various user-cases related to the launch sites or Payload mass, like
  - Average payload mass carried by booster version F9 v1.1
  - The 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 greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - All the booster versions that have carried the maximum payload mass
  - 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/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/EDA-sql.ipynb

Build an Interactive Map with Folium

- Folium a python module is used to observe and get more geographical details related to Launch Sites like
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculating the distances between a launch site to its proximities

https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/Data-visualization-Folium.ipynb

Build a Dashboard with Plotly Dash

- Plotly dashboard is used to observe and gain overview on categorical data using Pie charts and Scatter plots
  - Total site launches by sites
  - Total Success Launches for KSC LC-39A
  - Correlation Between Payload Mass and Sites

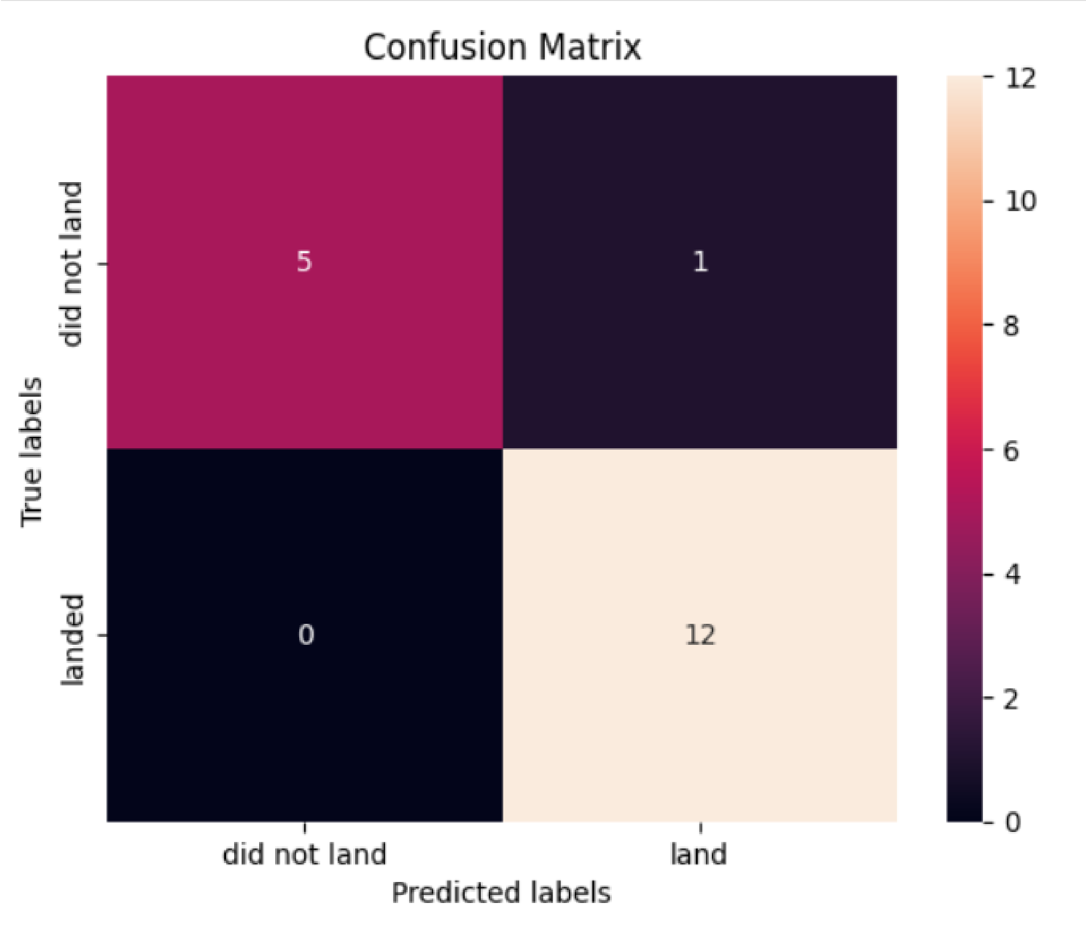
http://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/spacex-dash-app.py

Predictive Analysis (Classification)

- Initially converted the target variable column to a pandas series before fitting the model
- Standardized the categorical columns data using Standard Scaler, these are required before fitting prediction models because numerical data is required
- Divided the data into X_train, X_test, Y_train, Y_test
- Used GridSearchCV and created objects for each of the machine learning models, plotted confusion matrix on all models and Calculate the accuracy on the test data.
    - logistic regression
    - support vector machine
    - decision tree classifier
    - k nearest neighbors
    - https://github.com/ShoaibShaik786/Coursera_Assignments/blob/main/IBM-DataScience-Capstone/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

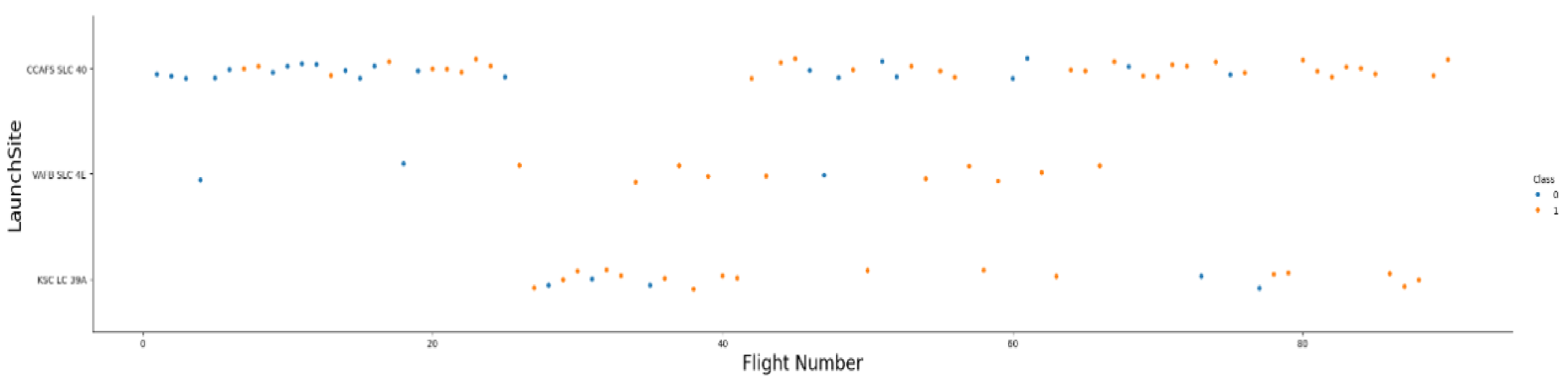
- Decision Tree algorithm provided best accuracy on test data among the other models
- From the confusion matrix of decision tree classifier we can see it has incorrectly predicted only 1 value



Section 2

Insights drawn from EDA

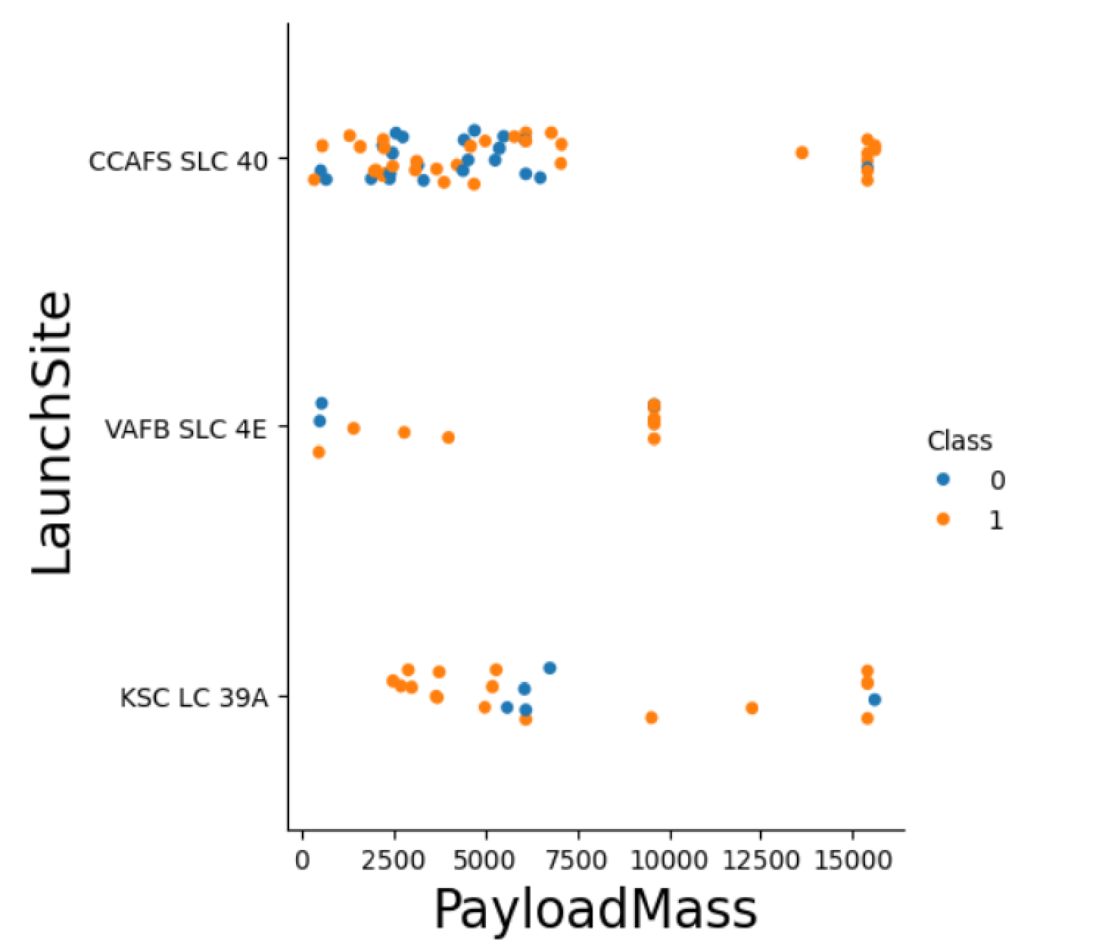
Flight Number vs. Launch Site



- CCAFS SLC 40 launch site has more number of launches
- VAFB SLC 4E has higher success rates compared to other launch sites
- Earlier Flight numbers had some failure rate and its relatively reduced over new Flight Numbers

Payload vs. Launch Site

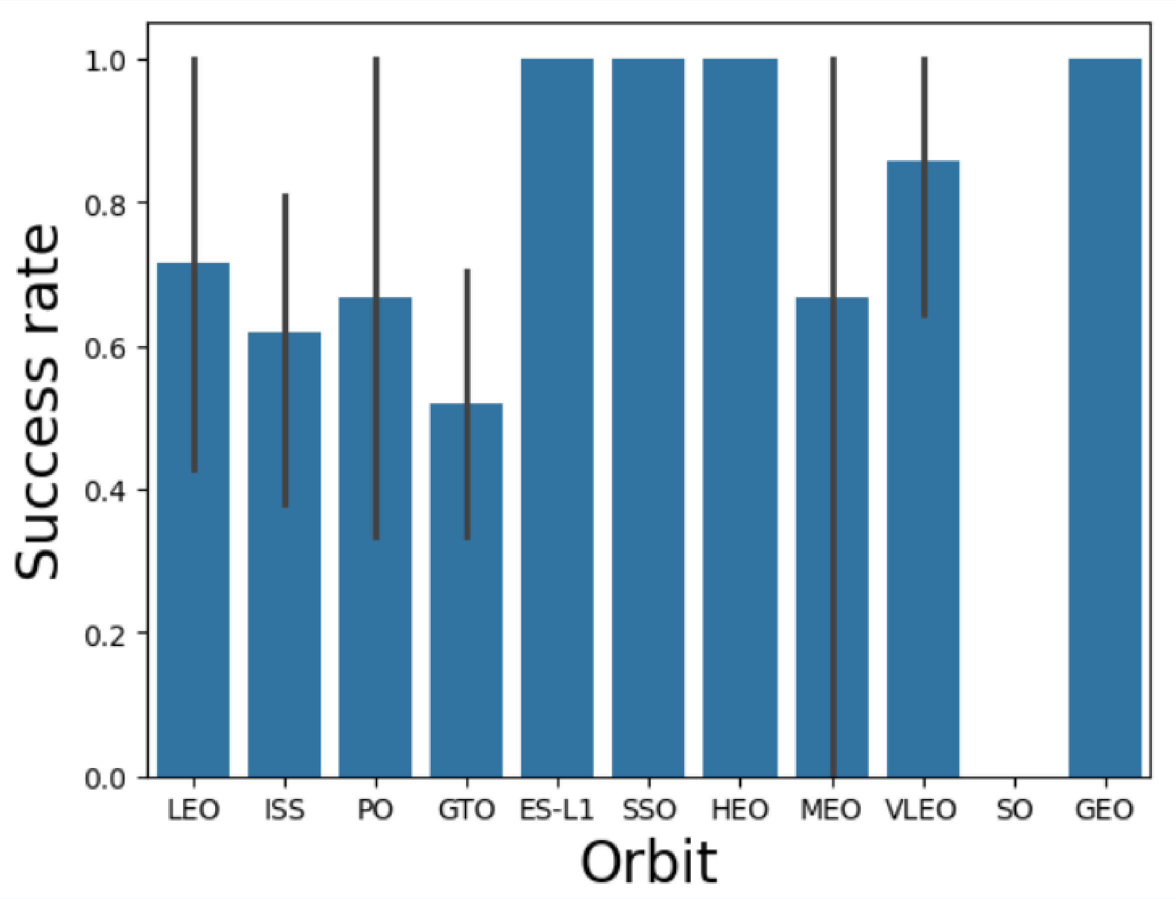
- Success rate is higher when the payload mass higher
- At 10,000 Payload Mass there are no failure cases
- KSC LC 39A has more success rate for under 5000 Payload Mass



Success Rate vs. Orbit Type

- ESL1,SSO,HEO and GEO has 100% success rate

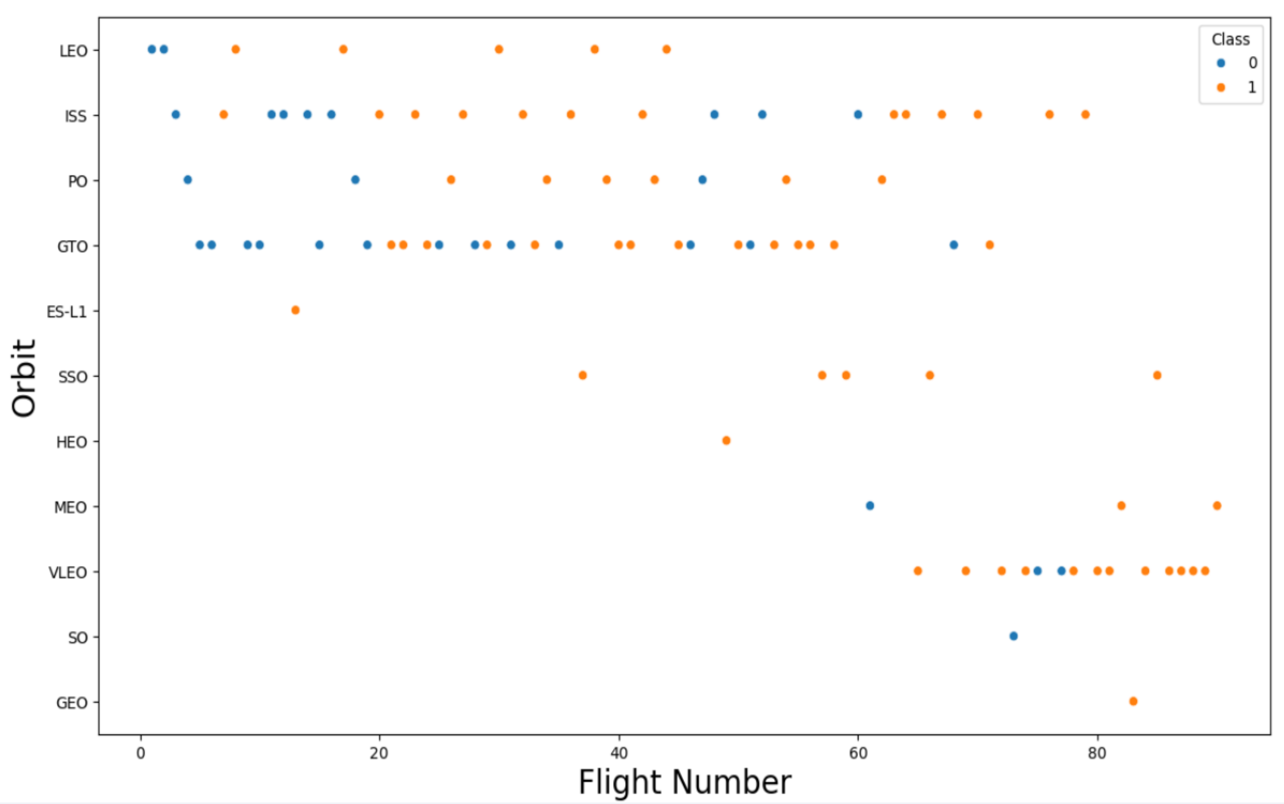
- SO orbit has 0% success rate



Flight Number vs. Orbit Type

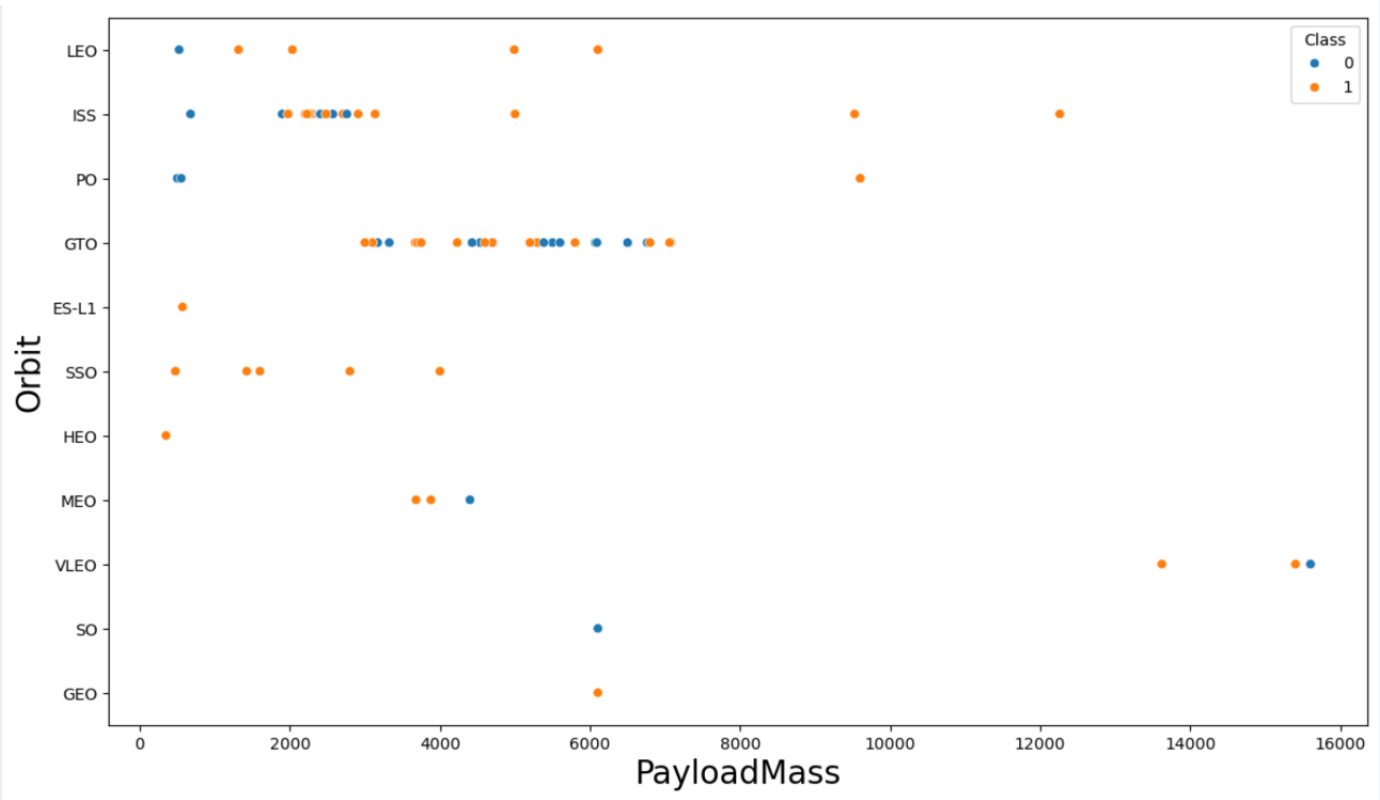
- In LEO orbit, the success rate appears more as the Flight number is increased

- There is no much relationships observed with Flight number and orbit type



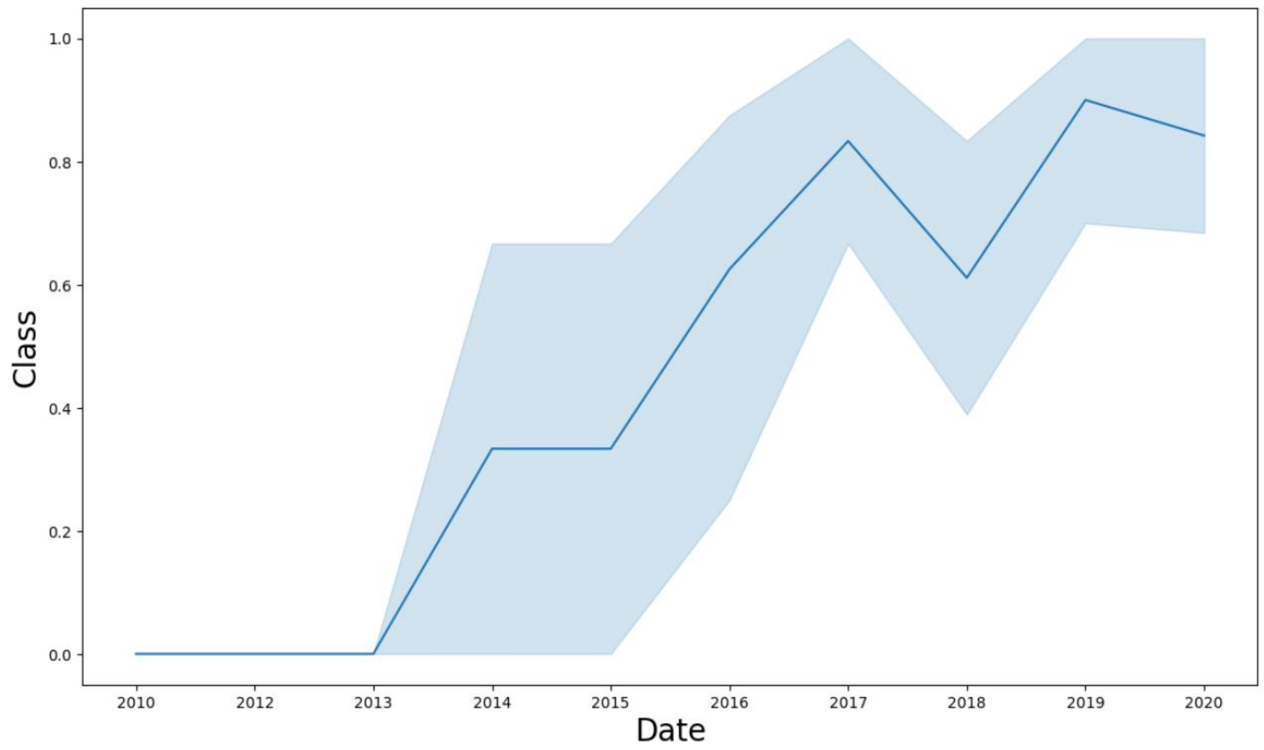
Payload vs. Orbit Type

- Higher Payload seems to have more success rate in all the orbits
- SSO orbit has no failure rate with lesser Payload mass
- In GTO orbit, the higher payload mass has seemingly higher failure cases



Launch Success Yearly Trend

- After 2013, the success rate increased exponentially till 2017



All Launch Site Names

- Names of the All 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'

```
%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
2010-06-04	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

5 records where launch sites begin with `CCA`

Total Payload Mass

- The total payload carried by boosters from NASA

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

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

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as "AVG_PAYLOAD_MASS" from SPACEXTABLE where Booster_Version ='F9 v1.1'
```

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

AVG_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

- On 22nd Dec 2015,the first successful landing outcome on ground pad was observed

```
%sql select min(Date) from SPACEXTABLE 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

- 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

- Total number of successful and failure mission outcomes
- Query text adjust to capture screen

```
%%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes
FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%') success_table,
(SELECT COUNT(*) number_of_failure_outcomes FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%') failure_table
```

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

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- 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

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

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

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranked 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 count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

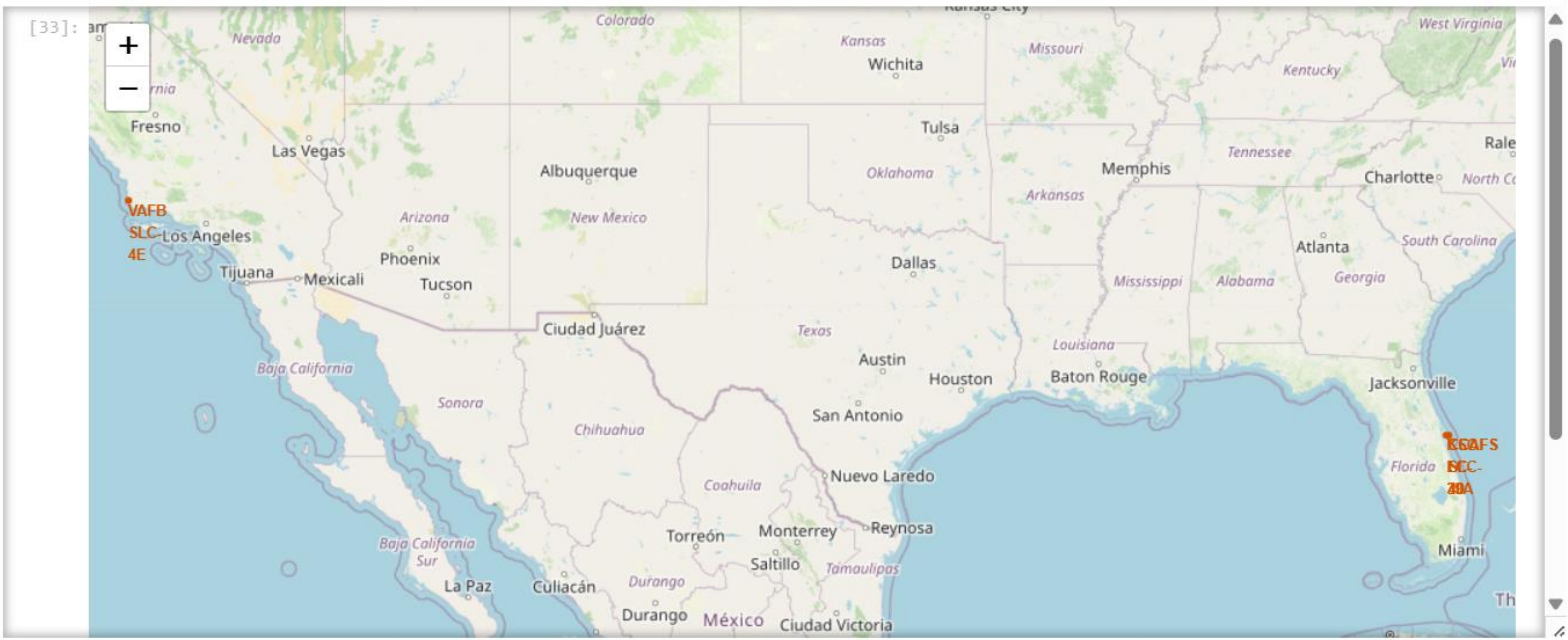
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
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

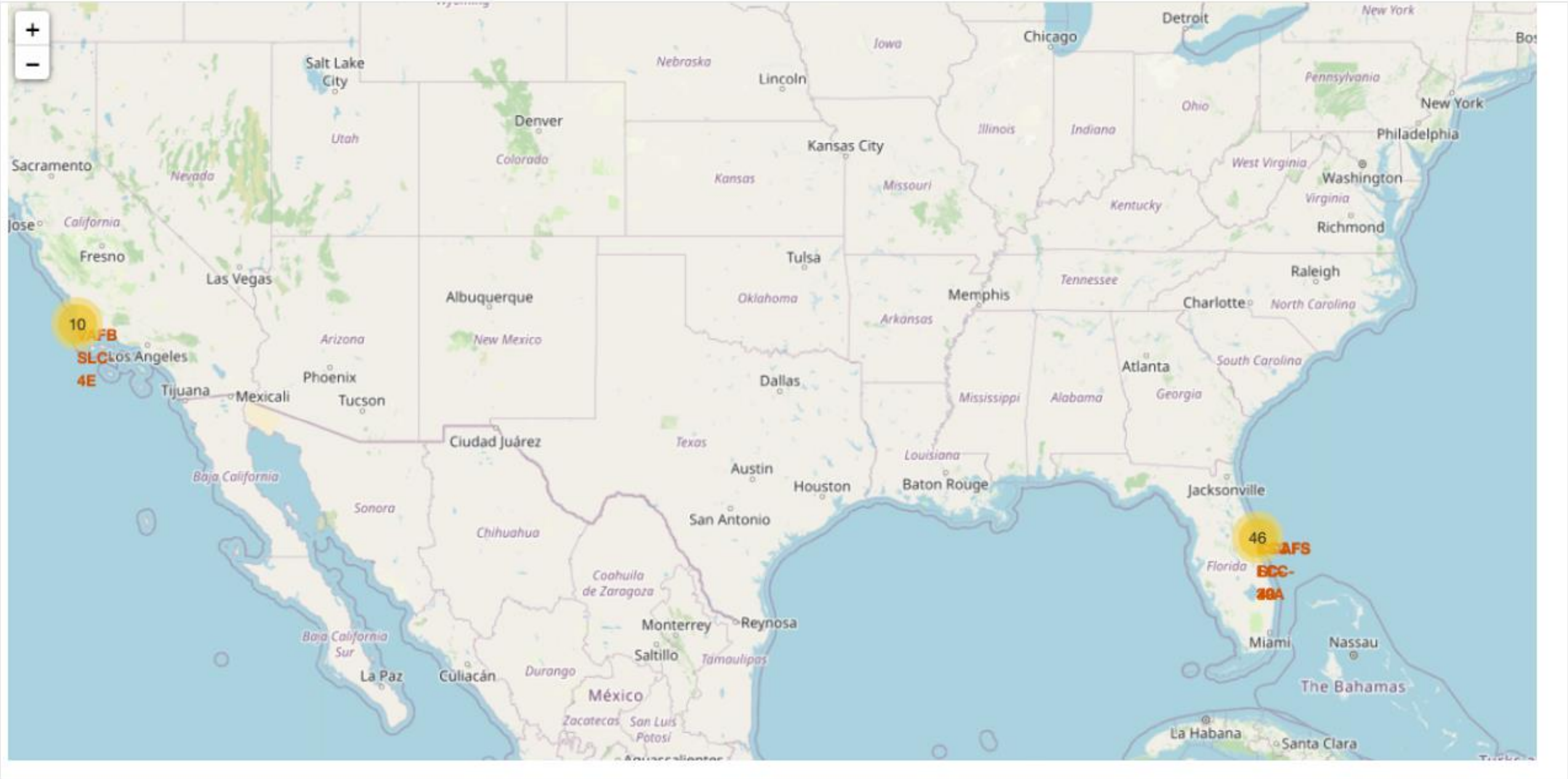
Section 3

Launch Sites Proximities Analysis

Launch Sites Marked on a global map using Folium



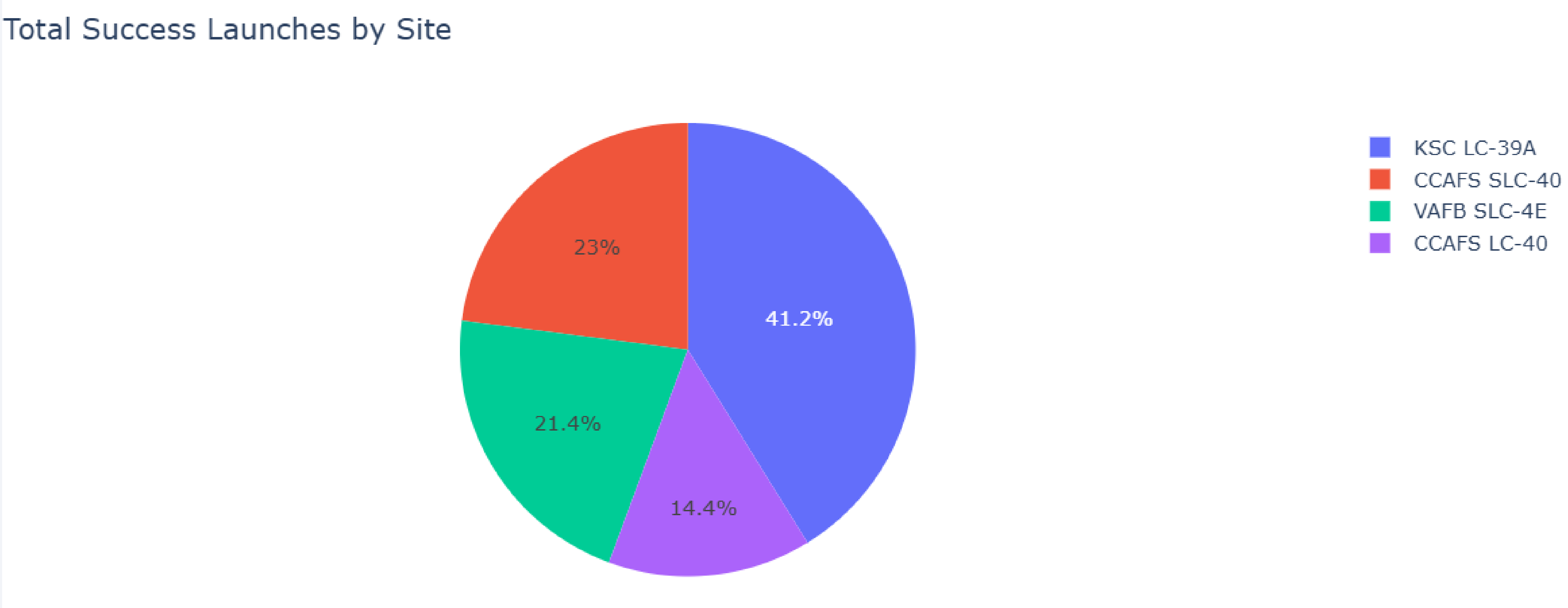
Map showing cluster of launch sites



Section 4

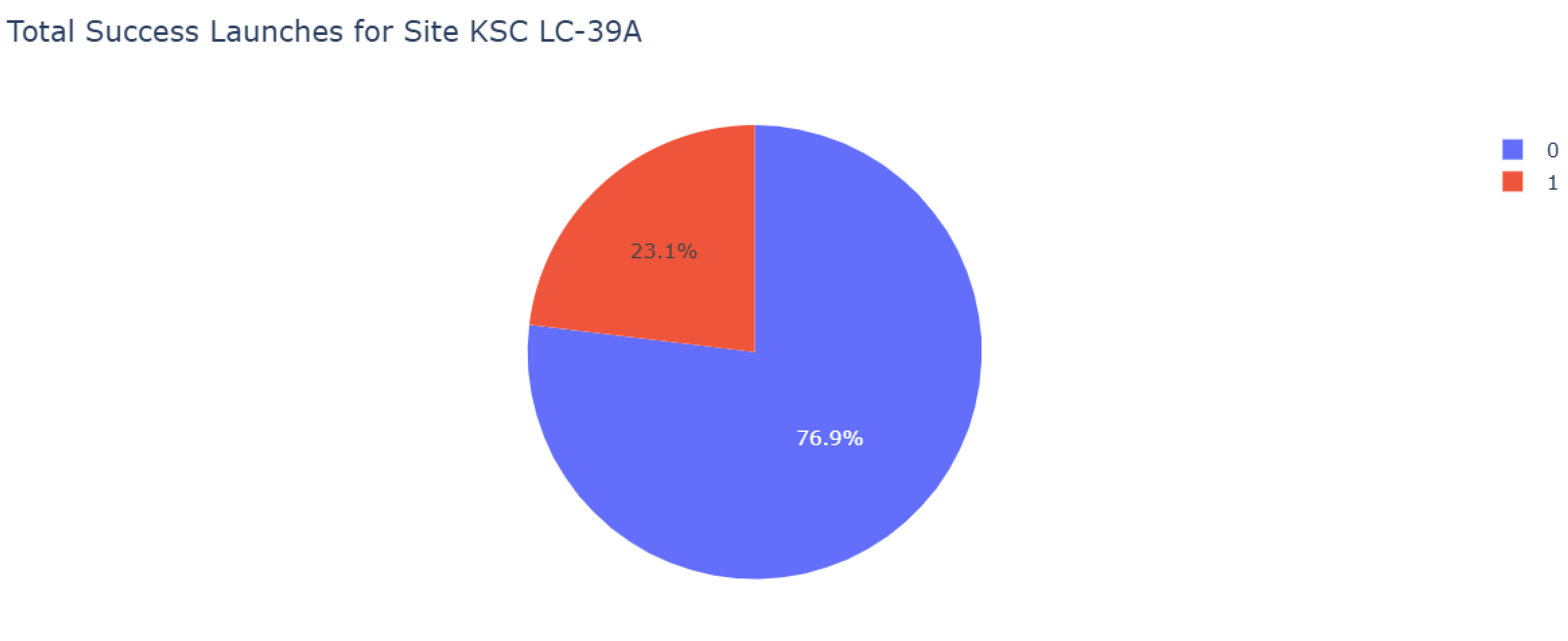
Build a Dashboard with Plotly Dash

Total site launches by sites



- KSC LC-39A has more number of launches

Total Success Launches for KSC LC-39A



- KSC LC-39A has 76.9% of Success launches

Correlation Between Payload Mass and Sites

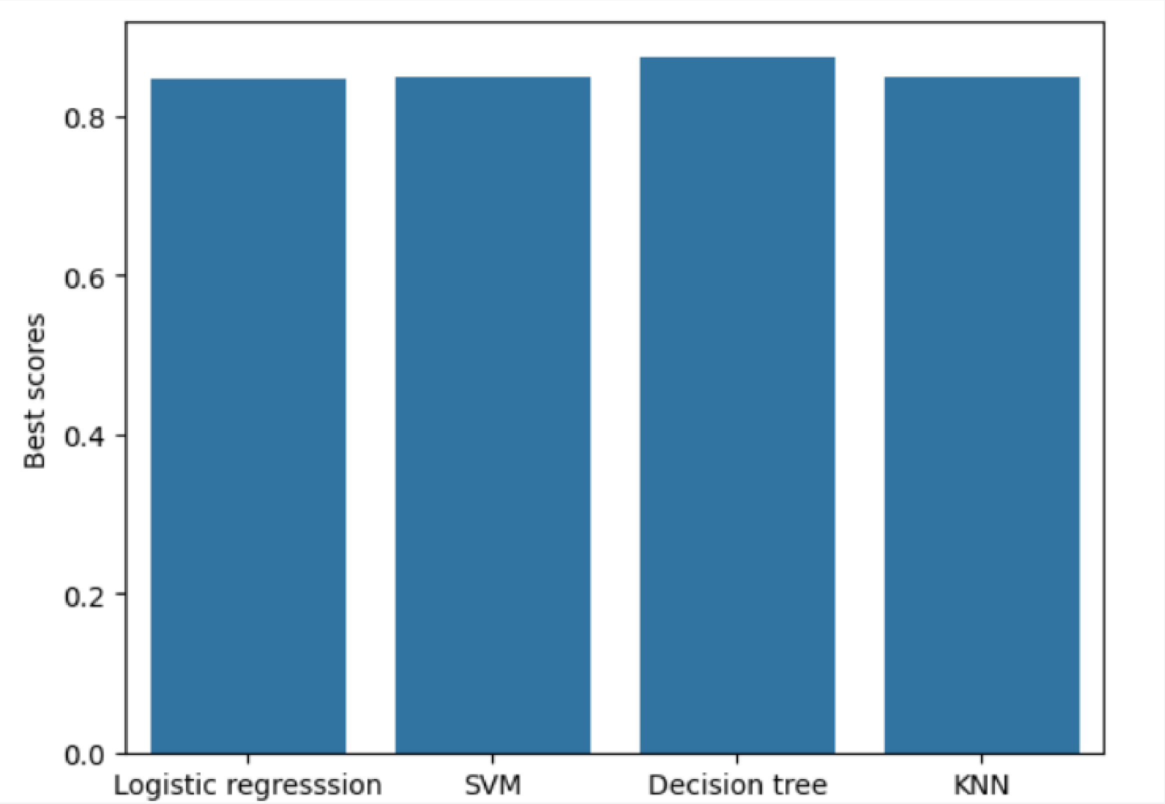
- FT Is the Booster version with High success rate with Payload mass below 5000kg

Section 5

Predictive Analysis (Classification)

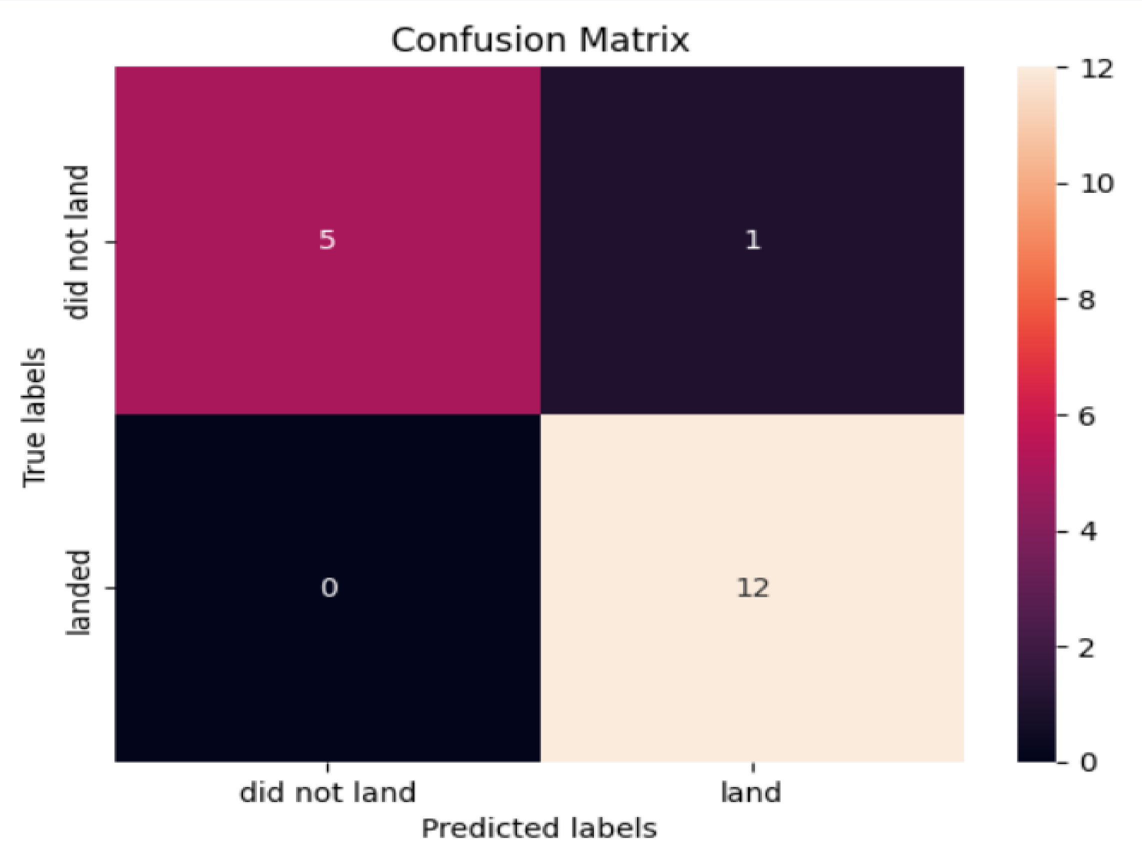
Classification Accuracy

- Decision Tree Classifier has shown higher accuracy on provided data than other classification models



Confusion Matrix

- From the matrix of the Decision Tree model we can see it has predicted most of the output accurately for most outcomes and has predicted 1 value incorrectly.



Conclusions

- Decision Tree algorithm provided best accuracy on test data among the other models
- The success rate of the launches kept increasing over the years
- KSC LC-39A has the highest success rate of launches from all the sites
- Launches with low payload mass achieved more success than high payload mass

Thank you!