IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ralph Arvin De Castro
12-20-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL
  - Exploratory Data Analysis for Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis
  - Interactive Analytics
  - Predictive Analysis

Introduction

- Project background and context
    - SpaceX launches Falcon 9 rockets at a price of $62 million, while other providers cost more. They save money by reusing the first stage. Therefore, we can estimate the launch cost by determining the success of the first stage. This information can be valuable to competitors such SpaceY. The project's objective is to train a machine learning model to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
    - Factors that determine landing success

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Gather launch data from SpaceX REST API
  - Obtain Falcon 9 Launch data from web scraping Wiki pages.
- Perform data wrangling
  - Perform Exploratory Data Analysis (EDA) to find patterns
  - Determine labels for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data

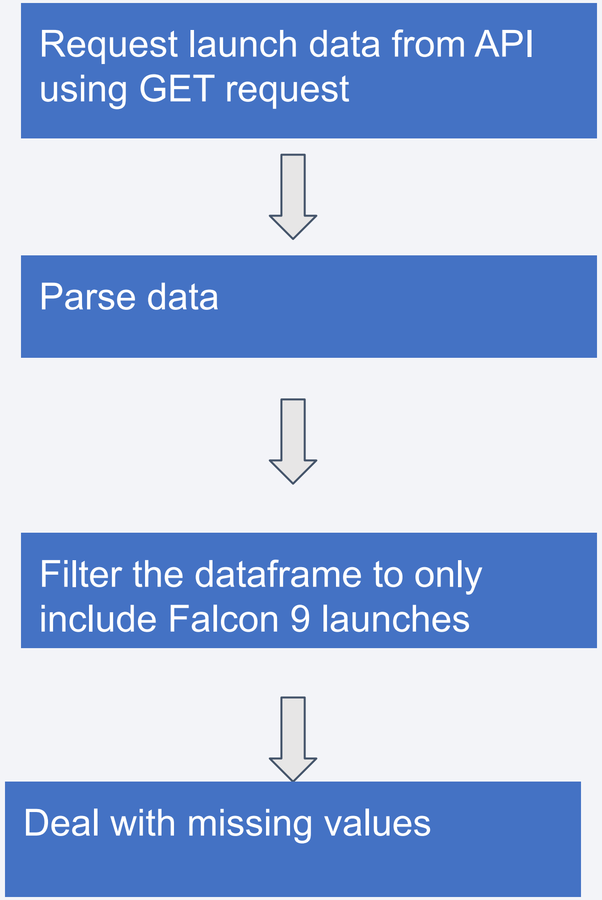
Data Collection

- Gathered launch data from SpaceX REST API
  - https://api.spacexdata.com/v4/rockets/
- Obtained Falcon 9 Launch data from web scraping Wiki pages.
  - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Requested launch data from API using GET request
- Parsed data
- Filtered the dataframe to only include Falcon 9 launches
- Dealt with missing values

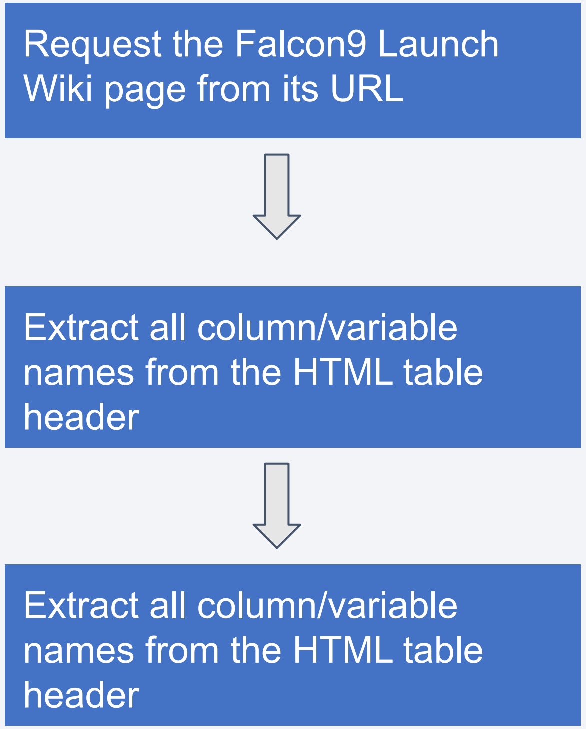
- Add the GitHub URL
    - https://github.com/imralpharvin/data-science-ibm/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

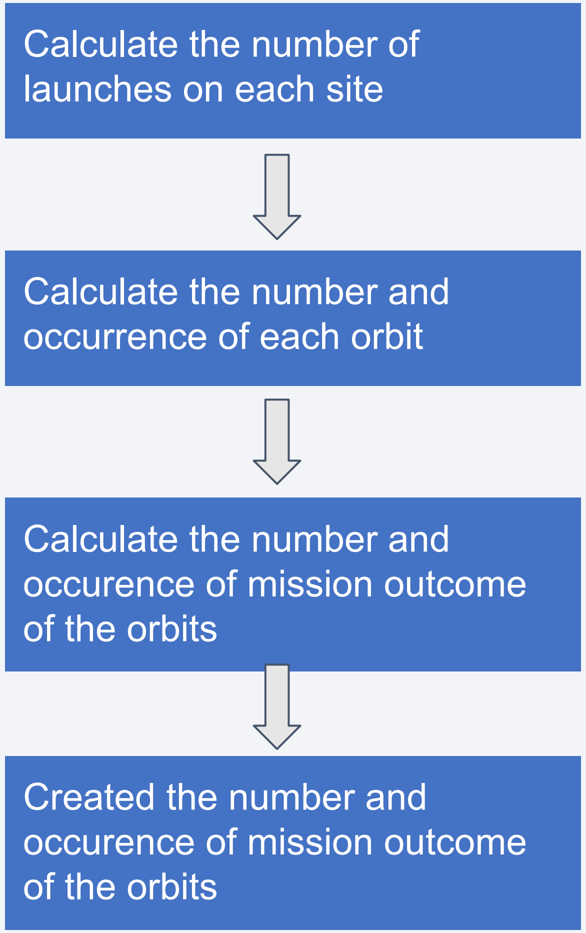
- Requested the Falcon9 Launch Wiki page from its URL
- Extracted all column/variable names from the HTML table header
- Created a data frame by parsing the launch HTML tables

- GitHub URL:
    - https://github.com/imralpharvin/data-science-ibm/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Calculated the number of launches on each site
- Calculated the number and occurrence of each orbit
- Calculated the number and occurence of mission outcome of the orbits
- Created a landing outcome label from Outcome column
- GitHub URL:
    - https://github.com/imralpharvin/data-science-ibm/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

1. Visualized the relationship between Flight Number and Launch Site
2. Visualized the relationship between Payload and Launch Site
3. Visualized the relationship between success rate of each orbit type
4. Visualized the relationship between FlightNumber and Orbit type
5. Visualized the relationship between Payload and Orbit type
6. Visualized the launch success yearly trend
7. Created dummy variables to categorical columns
8. Casted all numeric columns to `float64`

- https://github.com/imralpharvin/data-science-ibm/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Visualize the relationship between
- Flight Number and Launch Site
- Payload and Launch Site
- success rate of each orbit type
- FlightNumber and Orbit type
- Payload and Orbit type

↓

Visualize the launch success yearly trend

↓

Create dummy variables to categorical columns

↓

Cast all numeric columns to `float64`

EDA with SQL

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

- https://github.com/imralpharvin/data-science-ibm/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map
- Calculated the distances between a launch site to its proximities

- https://github.com/imralpharvin/data-science-ibm/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

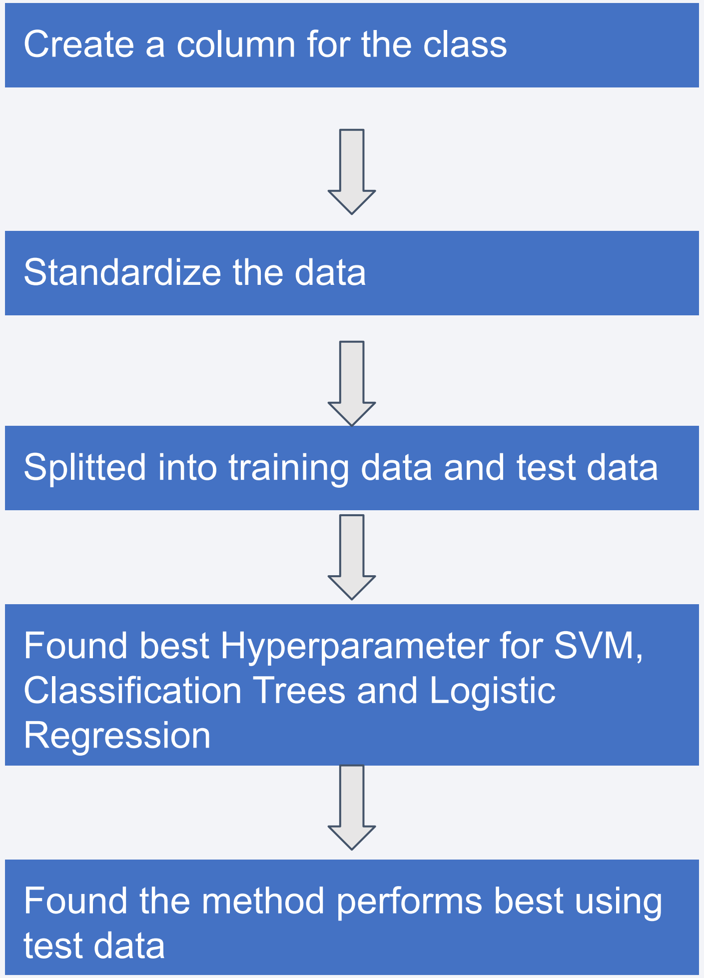
- Added pie chart of total success launches by sites
- Added pie charts of launch success ratios of sites
- Added a dropdown list for users to select launch sites
- Added scatter plot of payload mass, booster versions and success/failure launch
- Added a range slider for uses to select payload mass range

- https://github.com/imralpharvin/data-science-ibm/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Created a column for the class
- Standardized the data
- Splitted into training data and test data
- Found best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Found the method performs best using test data

- https://github.com/imralpharvin/data-science-ibm/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



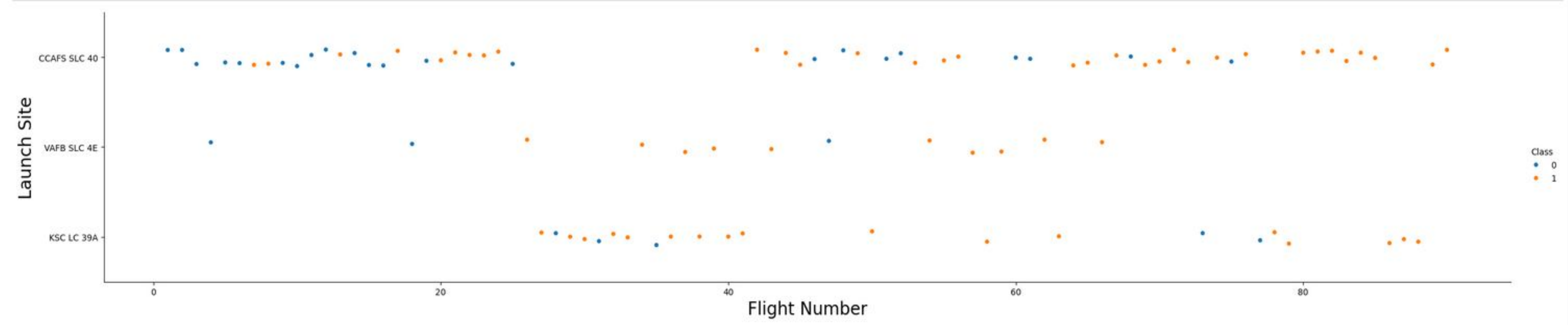
Results

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

Section 2

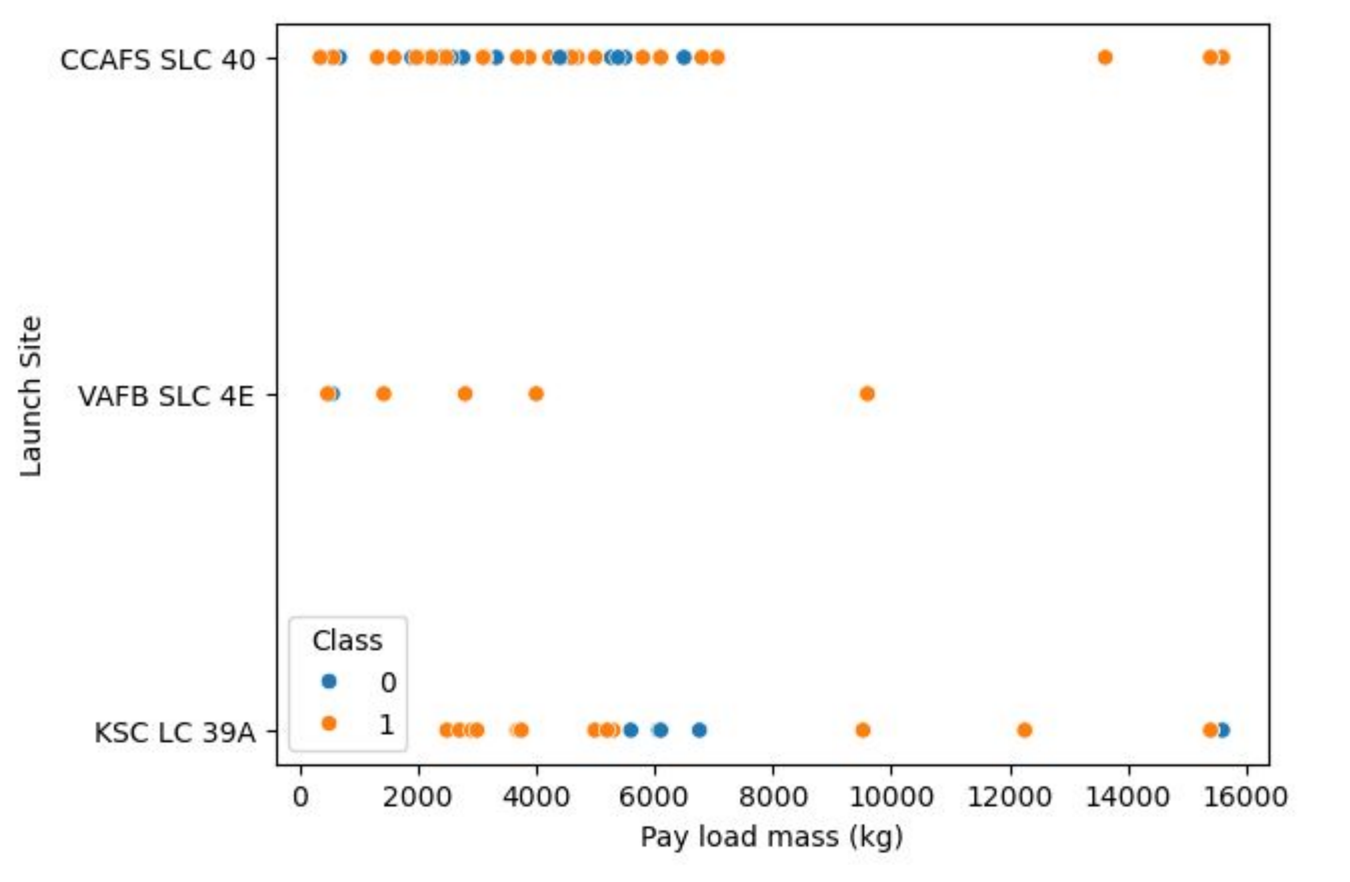
Insights drawn from EDA

Flight Number vs. Launch Site



- The more flights the site has, the higher the success rate is.

Payload vs. Launch Site



Most rockets launched have payload mass below 8000kg.

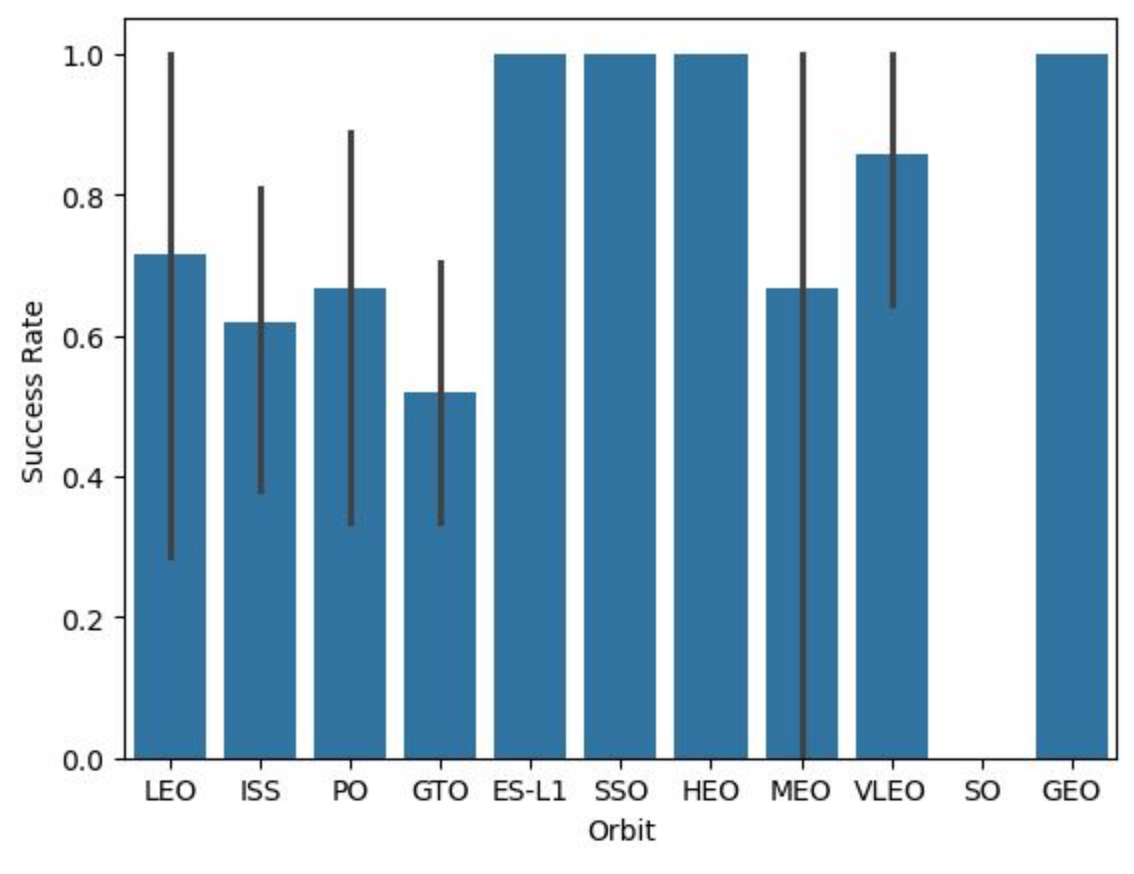
For CCFAS SLC 40 site, the success and failure rate is evenly distributed below 8000kg payload mass. All 3 attempts above 12000kg payload mass failed.

For VAFB SLC 4E site, only one landing was successful with a payload less than 2000kg.

For KSC LC 39A site, most successful landing are in the range of 5000 to 7000 kg payload mass. One landing with more than 15000kg payload mass was successful

In conclusion, trends are different among sites when it comes to payload mass and success rates.

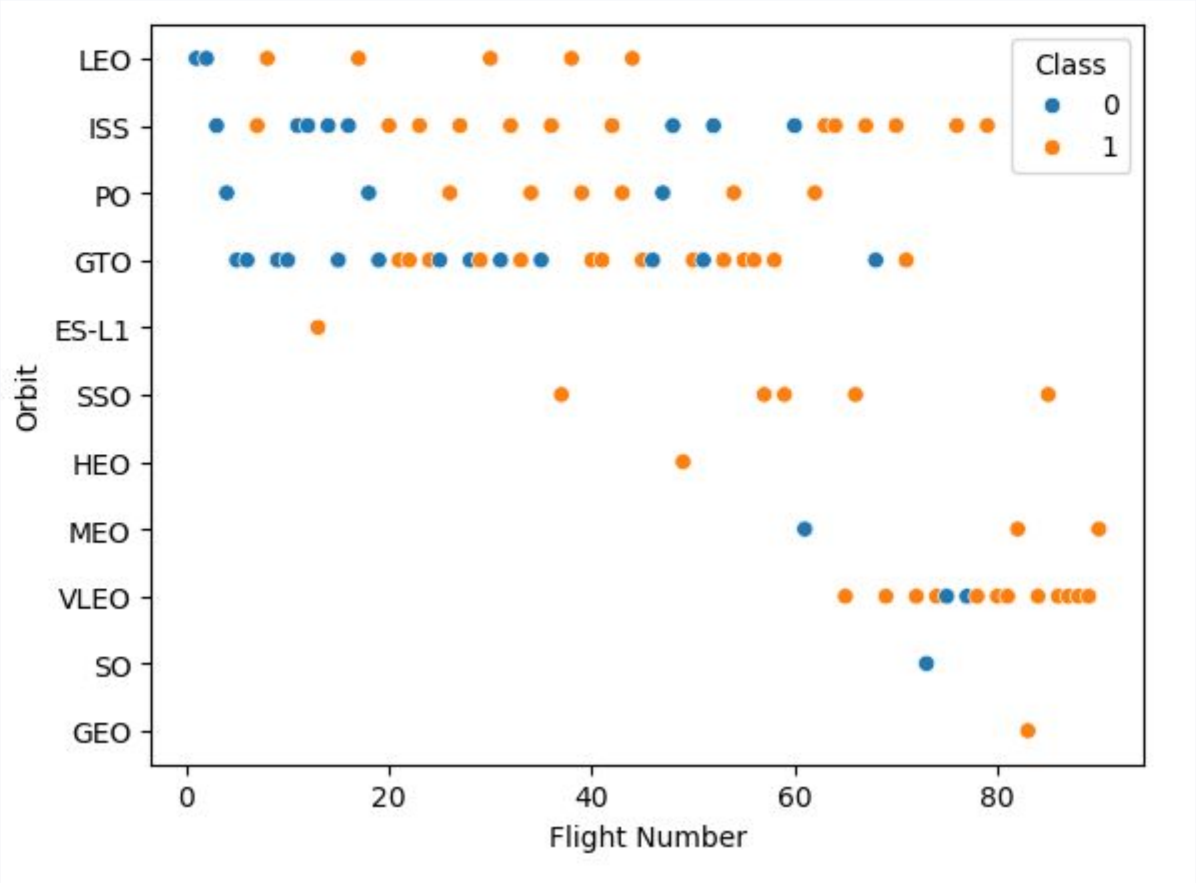
Success Rate vs. Orbit Type



- ES-L1, SSO, HEO and GEO orbits have the highest success rates
- SO orbit has the lowest success rate.

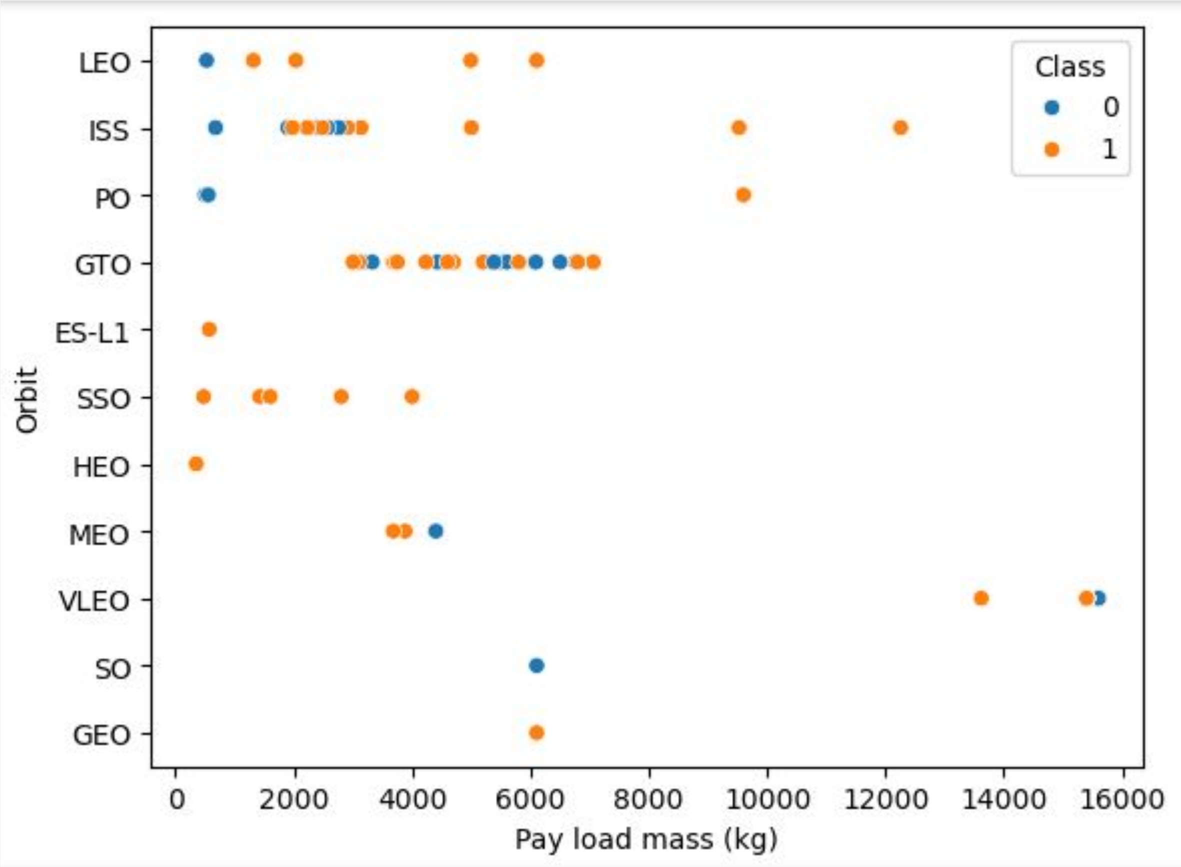
Flight Number vs. Orbit Type

- 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.



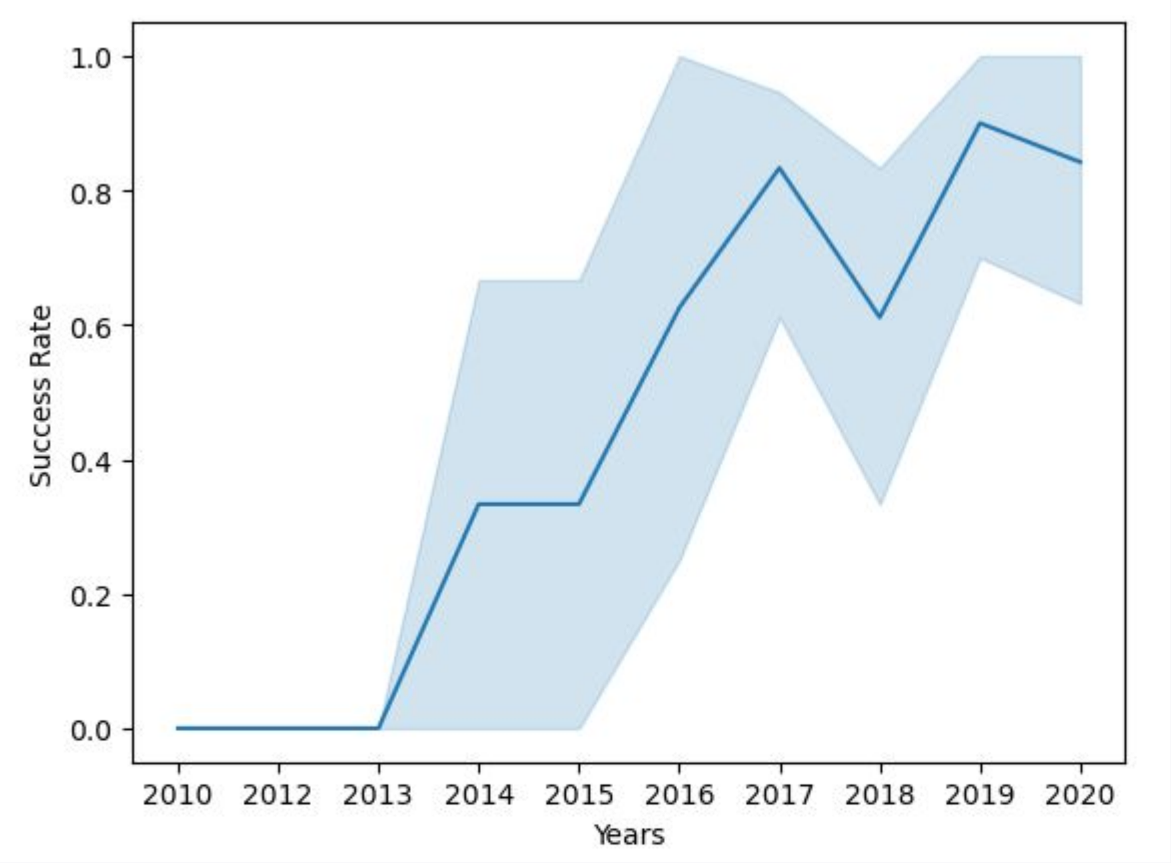
Payload vs. Orbit Type

- 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) are both there here.



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

- SpaceX has 4 launch sites listed in the screenshot

```
[10]: %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;

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

[10]:

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

Launch Site Names Begin with 'CCA'

- We queried 5 records where launch sites begin with `CCA` using LIMIT keyword

```
[9]: %sql SELECT * \
         FROM SPACEXTBL \
         WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;

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

[9]:

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

Total Payload Mass

- Total payload carried by boosters from NASA is 45596kg and found using SUM keyword

```
[11]: %sql SELECT SUM(PAYLOAD_MASS__KG_) \
          FROM SPACEXTBL \
          WHERE CUSTOMER = 'NASA (CRS)';

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

[11]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 using AVG keyword

```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) \
          FROM SPACEXTBL \
          WHERE BOOSTER_VERSION = 'F9 v1.1';

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

[12]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 2015-12-22 using MIN keyword

```
[15]: %sql SELECT MIN(Date) AS FirstSuccessfull_landing_date \
          FROM SPACEXTBL \
          WHERE Landing_Outcome LIKE 'Success (ground pad)'

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

[15]:

FirstSuccessfull_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
[16]: %sql SELECT Booster_Version FROM spacextbl WHERE landing_outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN 4000 and 6000

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

[16]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is 100 and 1 respectively

```
[14]: %sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Success%'

 * sqlite:///my_data1.db
Done.

[14]: COUNT(*)
        100

[15]: %sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Failure%'

 * sqlite:///my_data1.db
Done.

[15]: COUNT(*)
          1
```

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass (see screenshot)

```
[16]: %sql SELECT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl)

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

[16]:

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

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

```
[17]: %sql SELECT date, landing_outcome, booster_version, launch_site FROM spacextbl WHERE landing_outcome LIKE 'Failure (drone ship)' AND DATE LIKE '%2015%'

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

[17]:

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

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

- 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 (see screenshot)

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.

[18]:
```
%sql SELECT landing_outcome, COUNT(*) AS counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY counts DESC
```

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

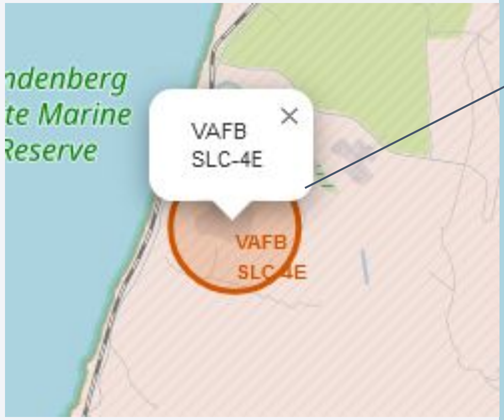
[18]:

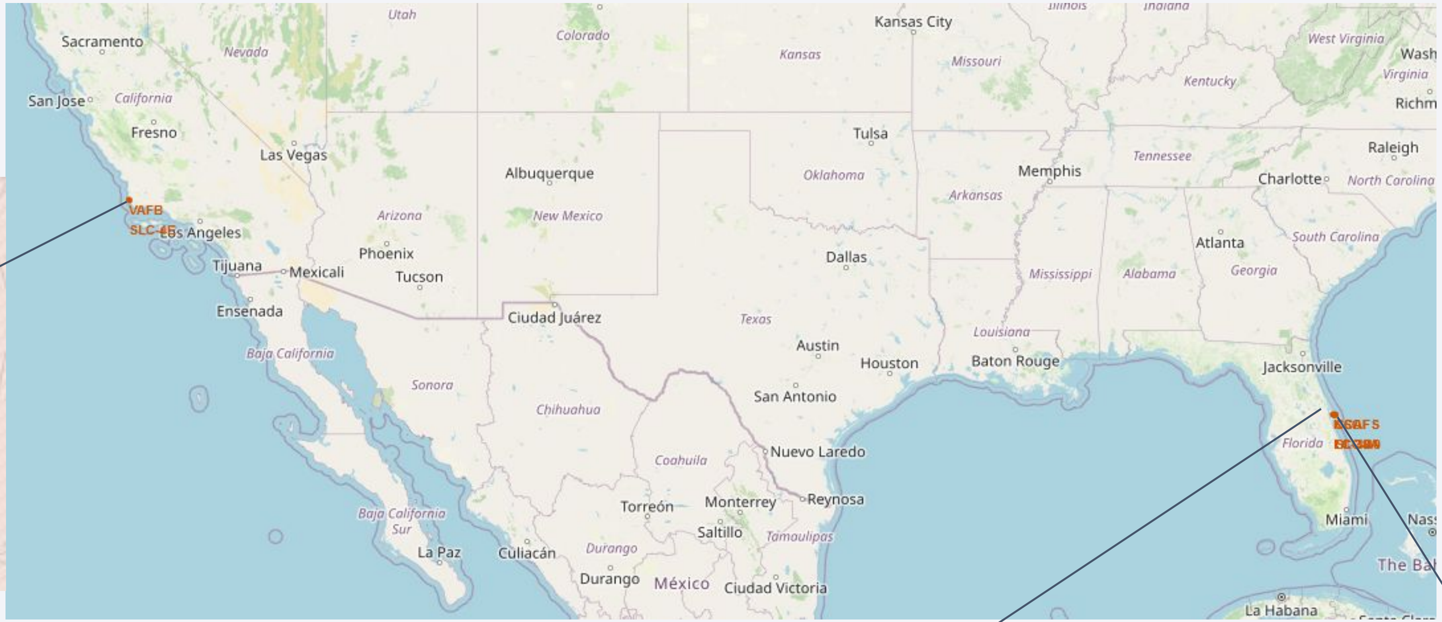
Landing_Outcome	counts
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

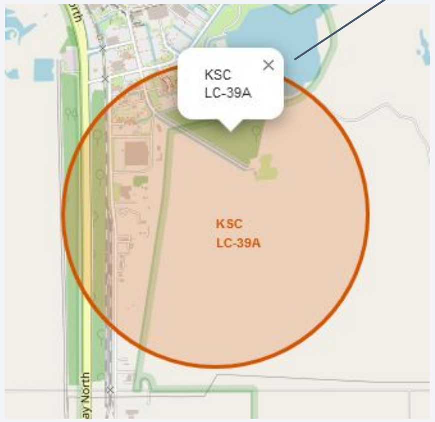
Launch Sites

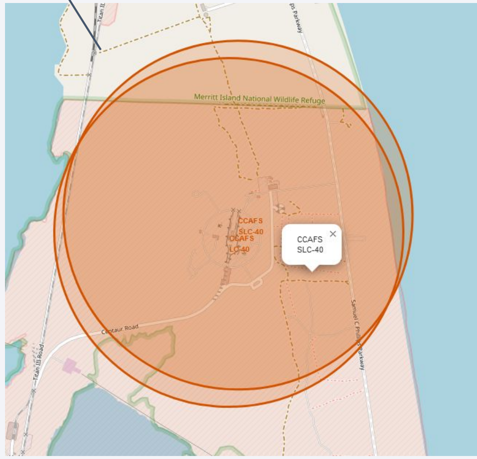




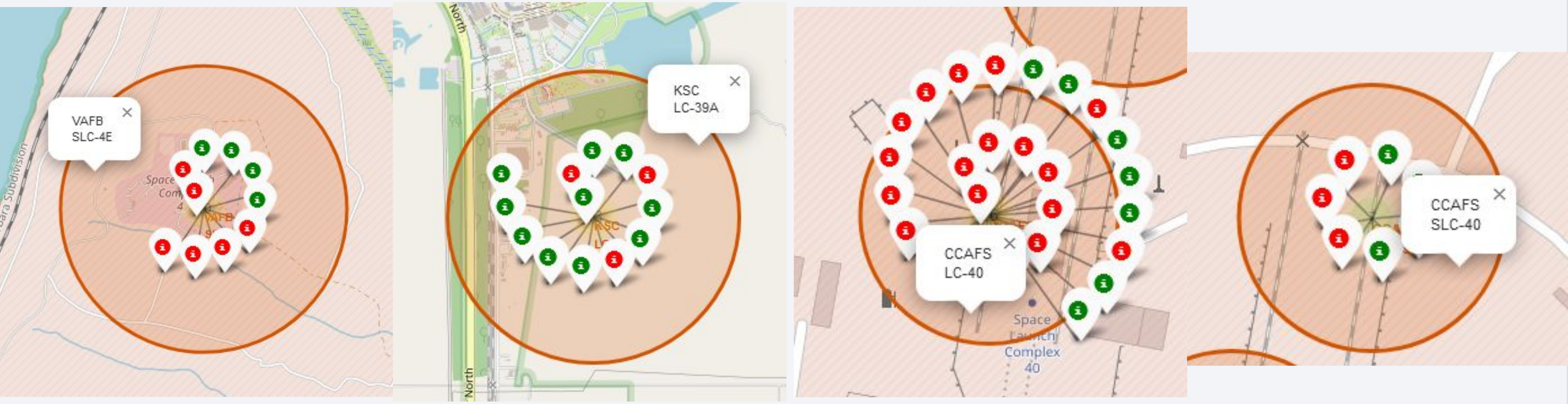
4 sites:

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





Launch Outcomes

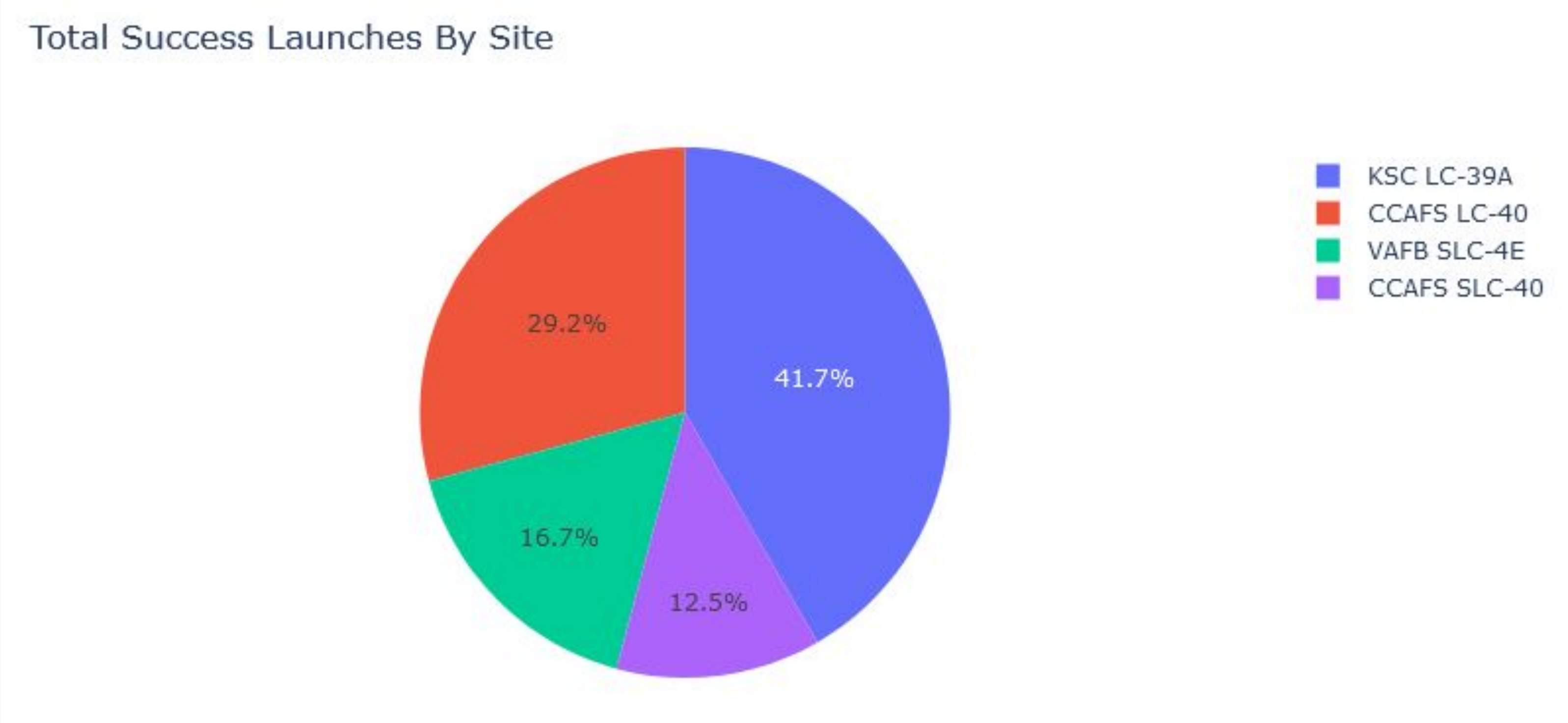


Green: Success
Red: Failure

Section 4

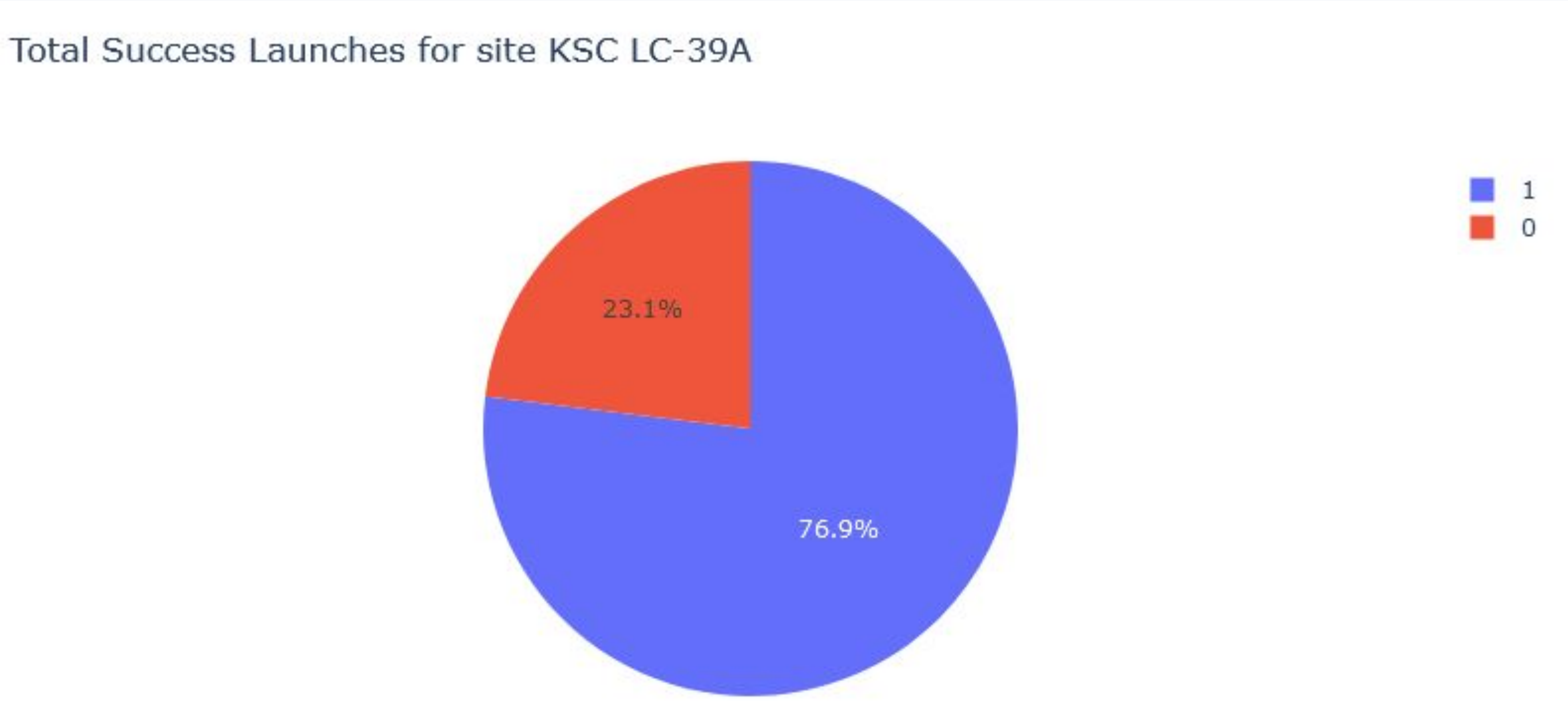
Build a Dashboard with Plotly Dash

Launch Success Percentage for All Sites



KSC LC-39A has the most launches while CCAFS SLC-40 has the least launches

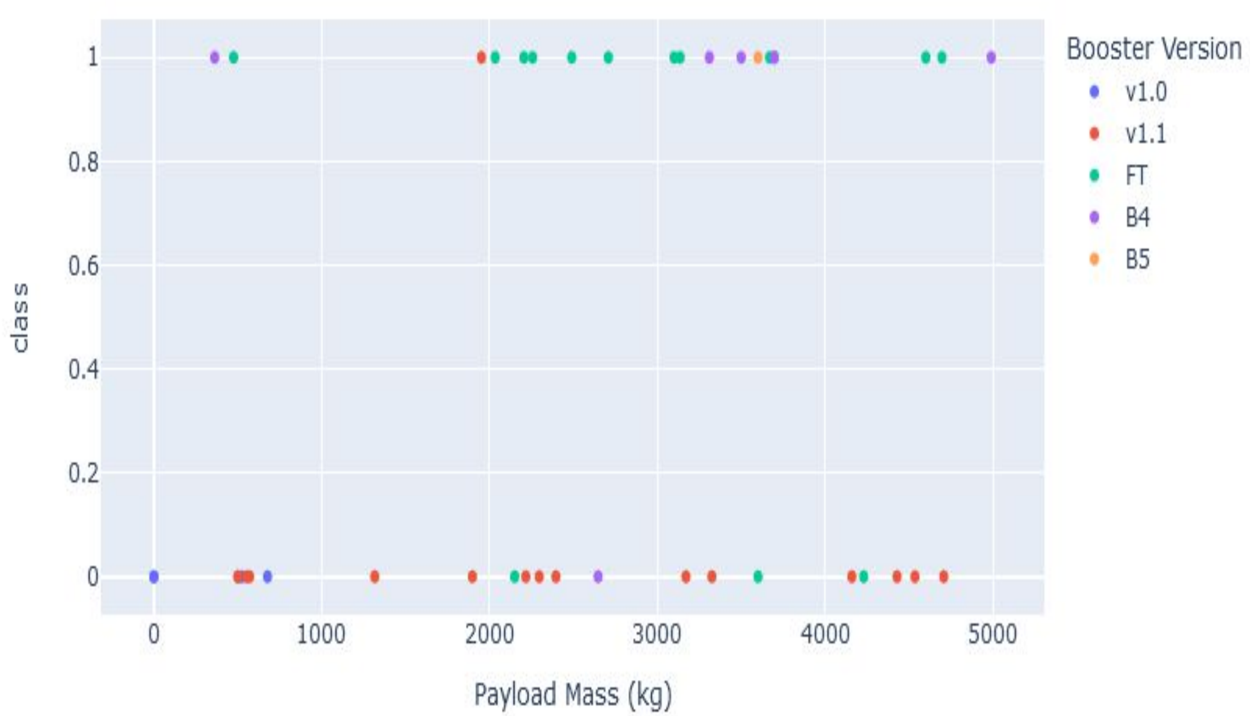
Launch Site with Highest Success Launch Ratio



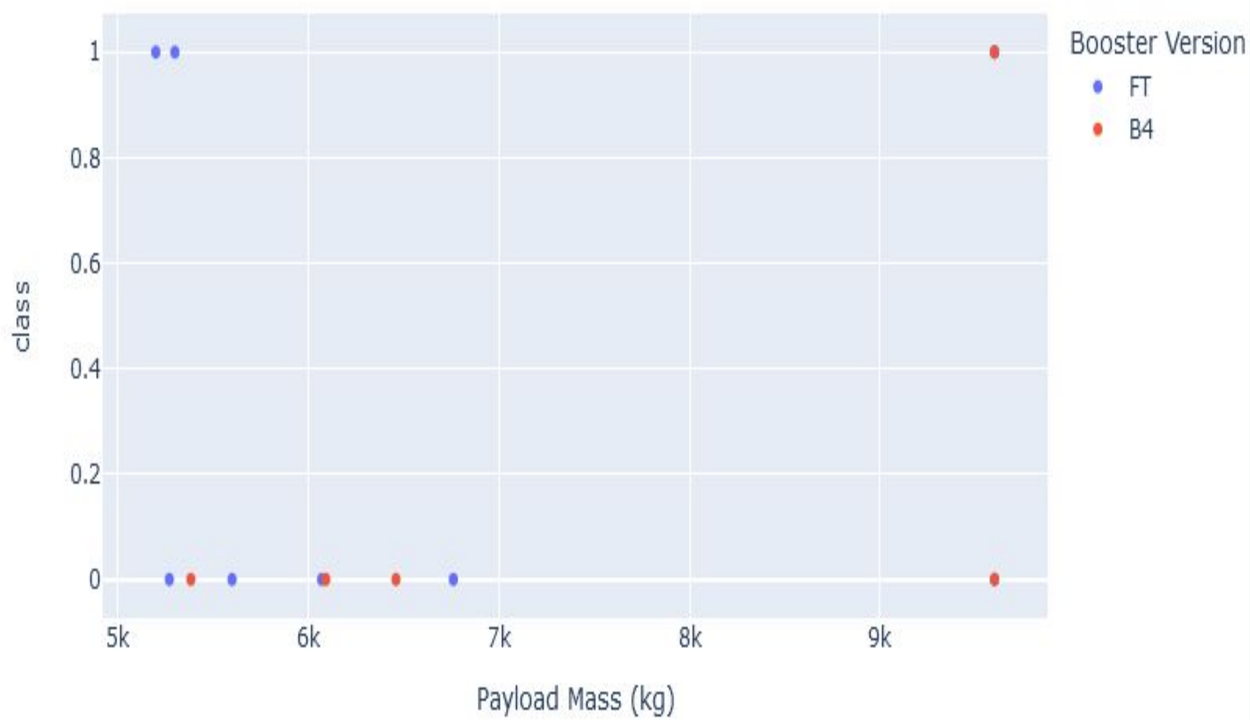
KSC LC-39A has highest success launch ratio of 76.9%:23.1%.

Payload Mass and Booster Version

Payload Mass <5000kg



Payload Mass >5000kg



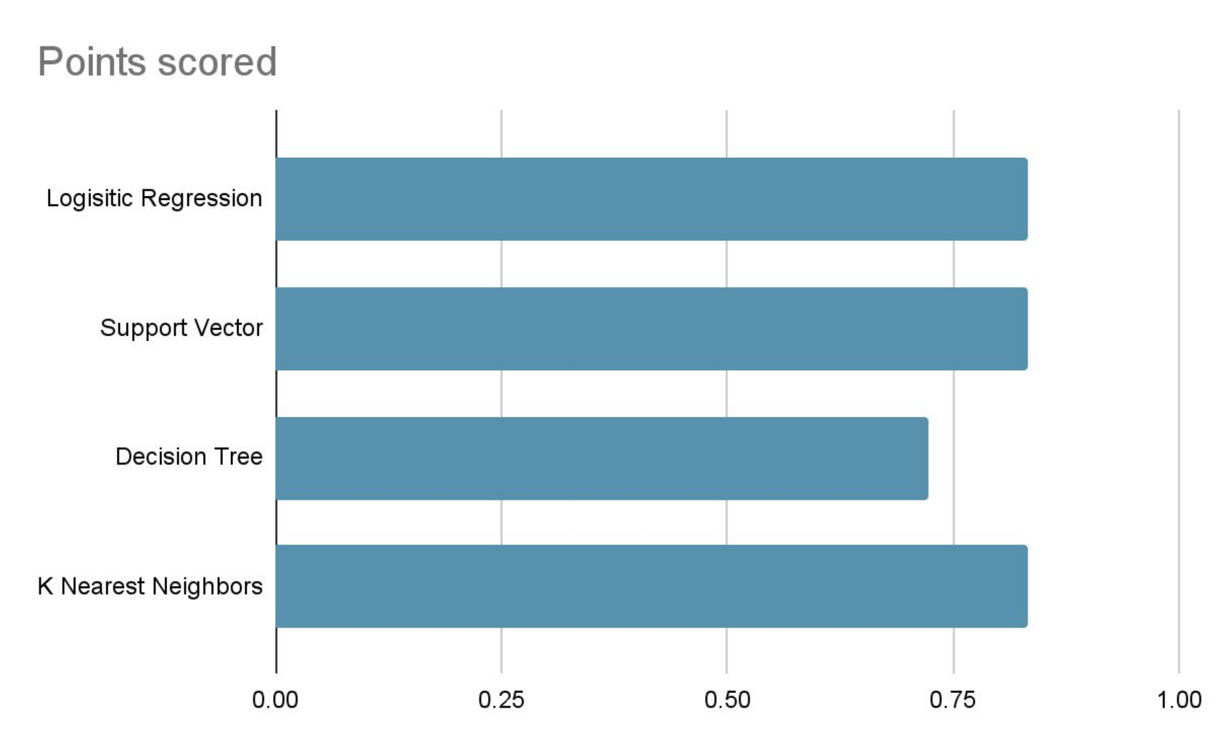
Lower payload mass below 5000 kg have higher success
B5 booster version has 100% success ratio but has only one launch
FT has the highest success ratio among booster versions that has more than 5 launches

Section 5

Predictive Analysis (Classification)

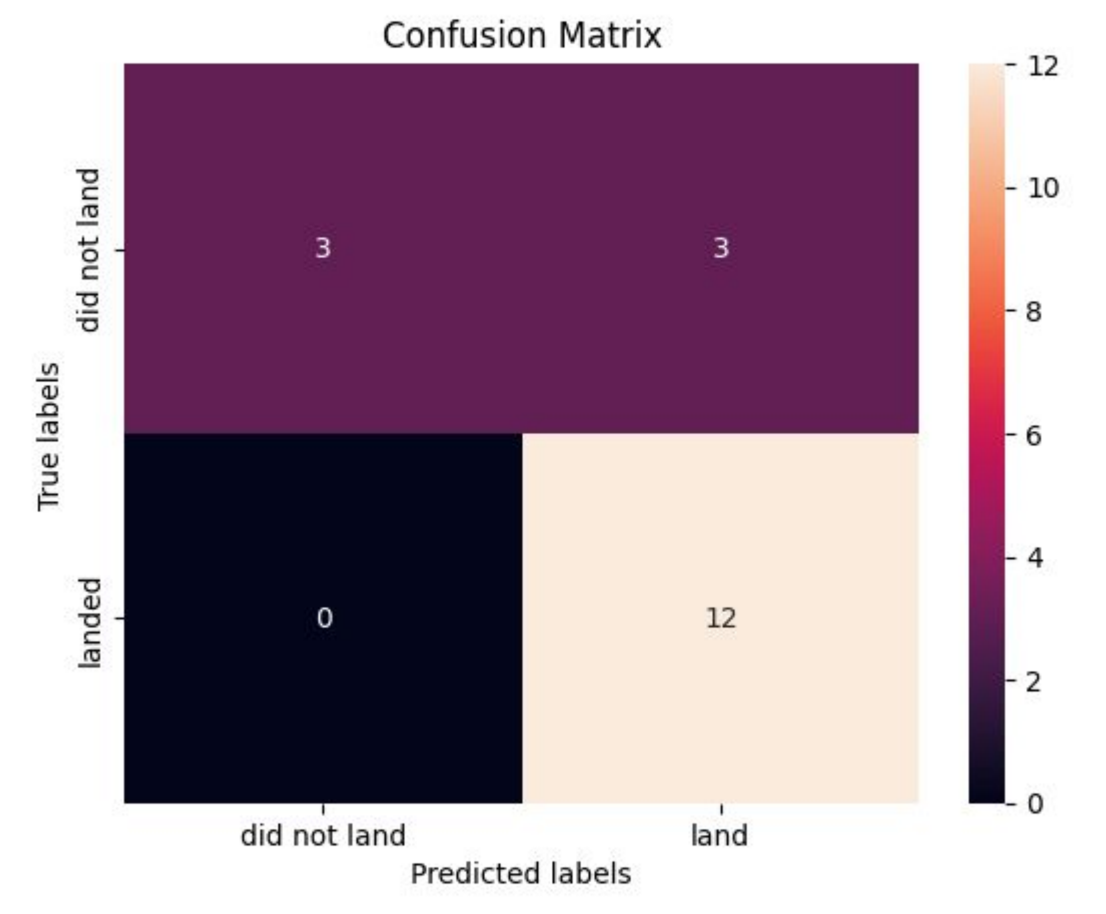
Classification Accuracy

Logistic Regression, Support Vector, K Nearest Neighbors have the highest classification accuracy of 0.8333334



Confusion Matrix

All 3 classification models have the same confusion matrices and can distinguish between different classes. The major problem is false positives or when the classifier classifies failed landings as successful landings.



Conclusions

- The more flights the site has, the higher the success rate is.
- ES-L1, SSO, HEO and GEO orbits have the highest success rates. SO orbit has the lowest success rate.
- The success rate since 2013 kept increasing till 2020
- KSC LC-39A had the most successful launches of any sites.
- Lower payload mass below 5000 kg have higher success.
- Logistic Regression, Support Vector, K Nearest Neighbors have the highest classification accuracy

Thank you!