IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mattis K
2024-03-08

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 mUSD; other providers of the same services could charge a cost of upwards of 165 mUSD each. The differences in cost is primarily due to SpaceX being able to reuse the first stage of their rockets. Therefore, if one can determine if the first stage will land successfully and thus be reused, one can determine the cost and savings of a launch.

  The predictions made in this project can be used if an alternate company wants to bid against SpaceX for a rocket launch

- Problems you want to find answers

  Which factors contributes towards predicting whether the first stage of the rocket will land successfully?

  The interaction between the factors to determine the success rate of a successful landing of the first stage of the rocket.

  What conditions that needs to be in place to secure a successful landing and provide alternate bidders with intelligence of when and how to launch their rockets in order to successfully recover and reuse the first stage of their rockets.

Section 1

Methodology

Methodology

Executive Summary

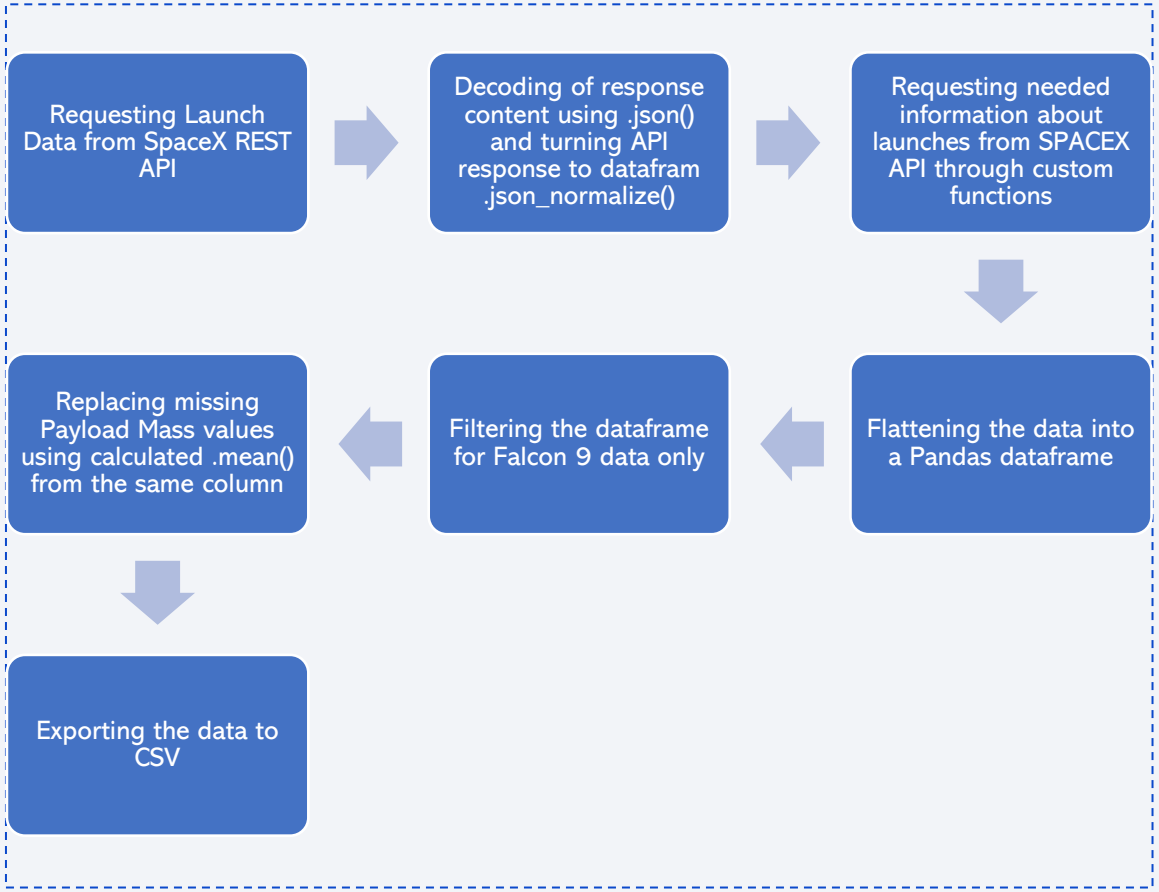
- Data collection methodology:
  - Data has been collected using the SpaceX API (Open data) and using web scraping on SpaceX data from Wikipedia.
- Perform data wrangling
  - Missing data has been handled by calculating the means for the continuous variables in the dataset and replacing the missing values with the calculated mean.
  - In order to fit categorical variables to our predictive models, One-hot encoding was applied to all categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- API and Web scraping
  - The base data used in this project stem from the SpaceX REST API (https//.docs.spacexdata.com)
    - Specifically, the /launches/past endpoint was used to collect information about SpaceX's past launches and collect information on whether the landing of the first stage rocket was successful or not to train the models used to predict optimal conditions to launch an alternate rocket in.
    - Data from the API required some cleaning, checking for missing values filling these missing values where necessary using the mean for continuous variables collected from the API response.
  - The base data was supplemented with scraped data from Wikipedia's page about the Falcon 9 rocket launches.
    - The data was residing in a HTML table and which was scraped using Python's web scraping library "BeautifulSoup".

Data Collection – SpaceX API

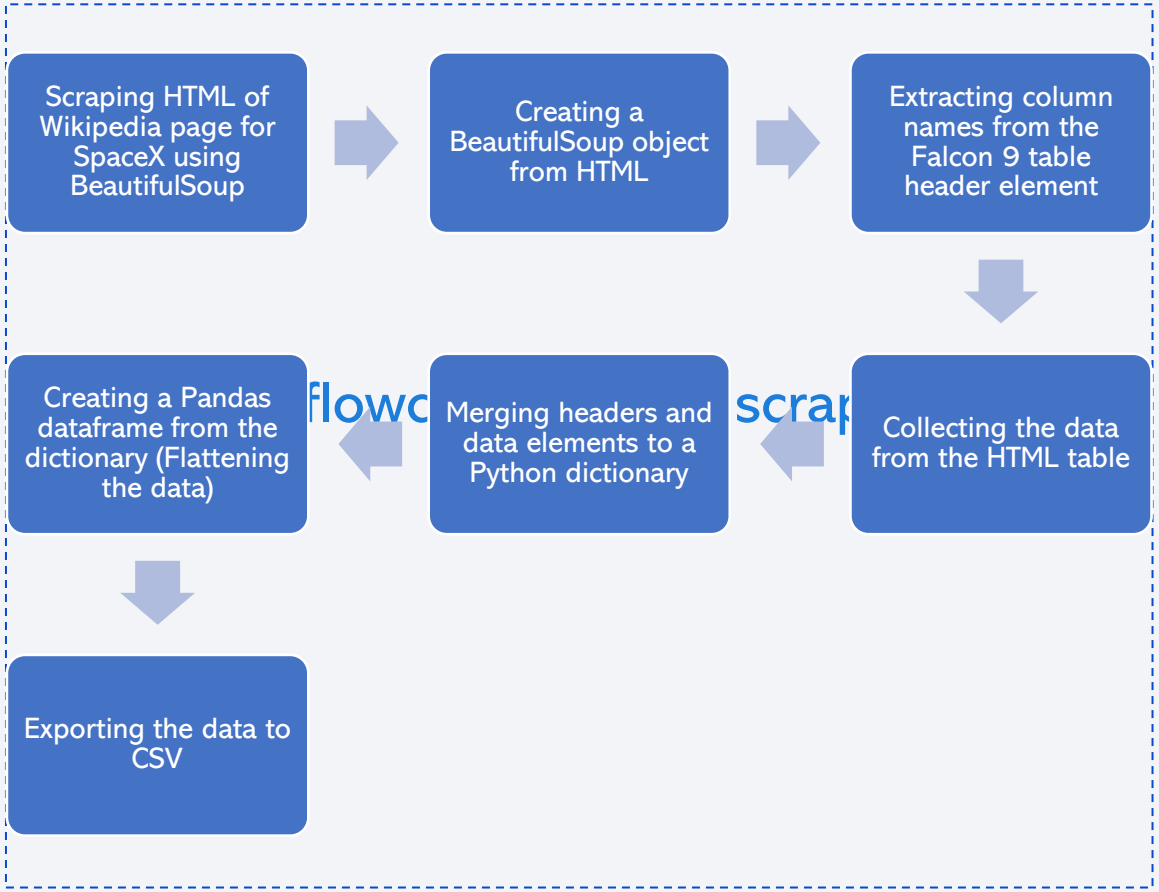
- Using Python's Requests library and Get method we accessed the /launches/past endpoint from the SpaceX REST API.
- Using the REST API JSON response we then normalized and flattened the data into a Pandas dataframe making it ready for further processing.
- GitHub Notebook: [dscap/data_collection_api.ipynb at main · mattan87/dscap (github.com)]



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- GitHub Notebook: [dscap/data_collection_scrape.ipynb at main · mattan87/dscap (github.com)]



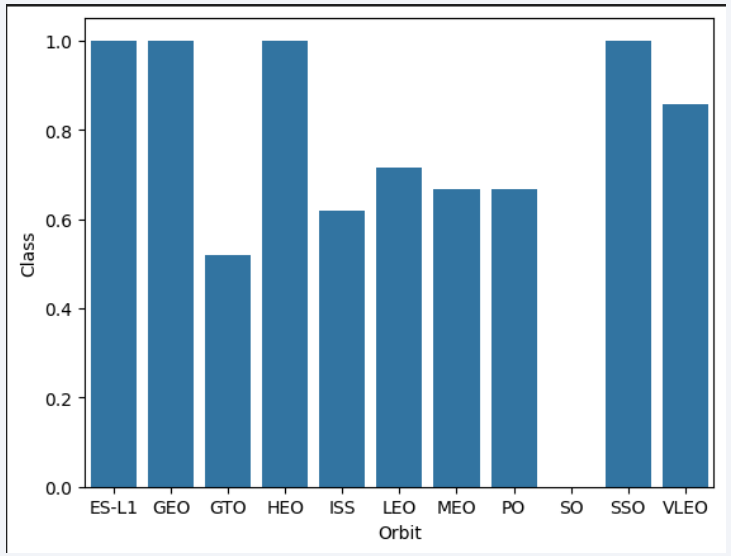
Data Wrangling

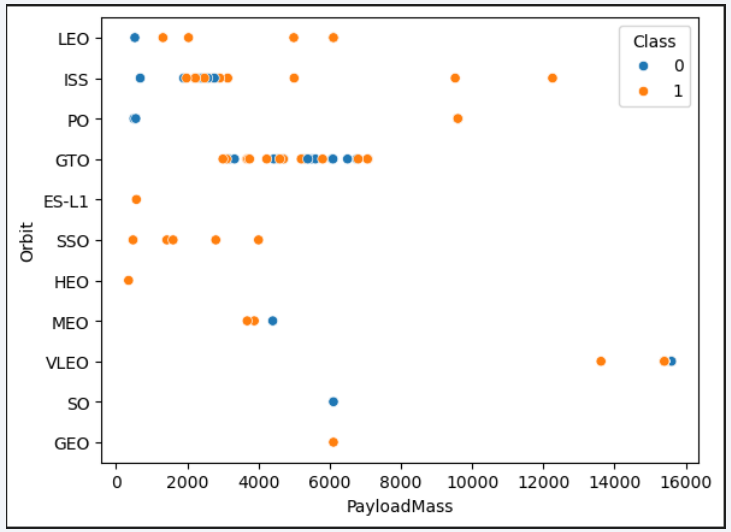
- First step of the process was to identify whether the collected data contained null values to understand whether we had gaps in the data that needed to be handled prior to fitting any model
- Secondly, we analyzed the whether the data was categorical or numerical to understand whether we had to standardize the data
- Third, we explored the occurrences of the different cateogircal variables, such as
  - Landing Outcomes (How the First Rocket Stages landed)
- Four, based on the third step we classified whether a landing outcome was a success or a fail to be able to get a label on which we could train and also test our models and their respective algorithms.

dscap/eda-dataviz.ipynb at main · mattan87/dscap (github.com)

EDA with Data Visualization

- Data was explored by visualizing the relationship between flight number and launch site, payload and launch site, success rate of each orbit type etc.





[dscap/eda-dataviz.ipynb at main · mattan87/dscap (github.com)]

EDA with SQL

SQL Queries to Support Explorative Data Analysis:

- Display the names of each unique launch site
- Display 5 records with launch site name starting with 'CCA'
- Display total payload mass carried by all 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 on a ground pad landing platform was performed
- Listing the names of boosters that successfully landed on a drone ship with a pay load mass between 4000 and 6000 kg
- Listing the total number different mission outcomes
- Listing the names of booster versions that have carried the maximum payload mass
- Listing the failed landing outcomes on a drone ship for the months in 2015 including
  - Booster version
  - Launch site names
- Creating a rank based on the count of landing outcomes between 2010-06-04 and 2017-03-20 (Descending order)
  - No landing attempt was the most frequent occurring outcome with 21 counts and Precluded (drone ship) was the least occurring with 1 count

dscap/eda-sql-coursera_sqllite.ipynb at main · mattan87/dscap (github.com)

Build an Interactive Map with Folium

- 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.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- 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. We answered some question for instance:
    - Are launch sites near railways, highways and coastlines. Which they were given there is a need to transport larger object to rockets as well as fuel for the rockets.
    - Do launch sites keep certain distance away from cities. Which they do, since the nature of a launch is quite dangerous to do in close proximity to an inhabited area should something go wrong.

- dscap/launch_site_location.jupyterlite.ipynb at main · mattan87/dscap (github.com)

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.

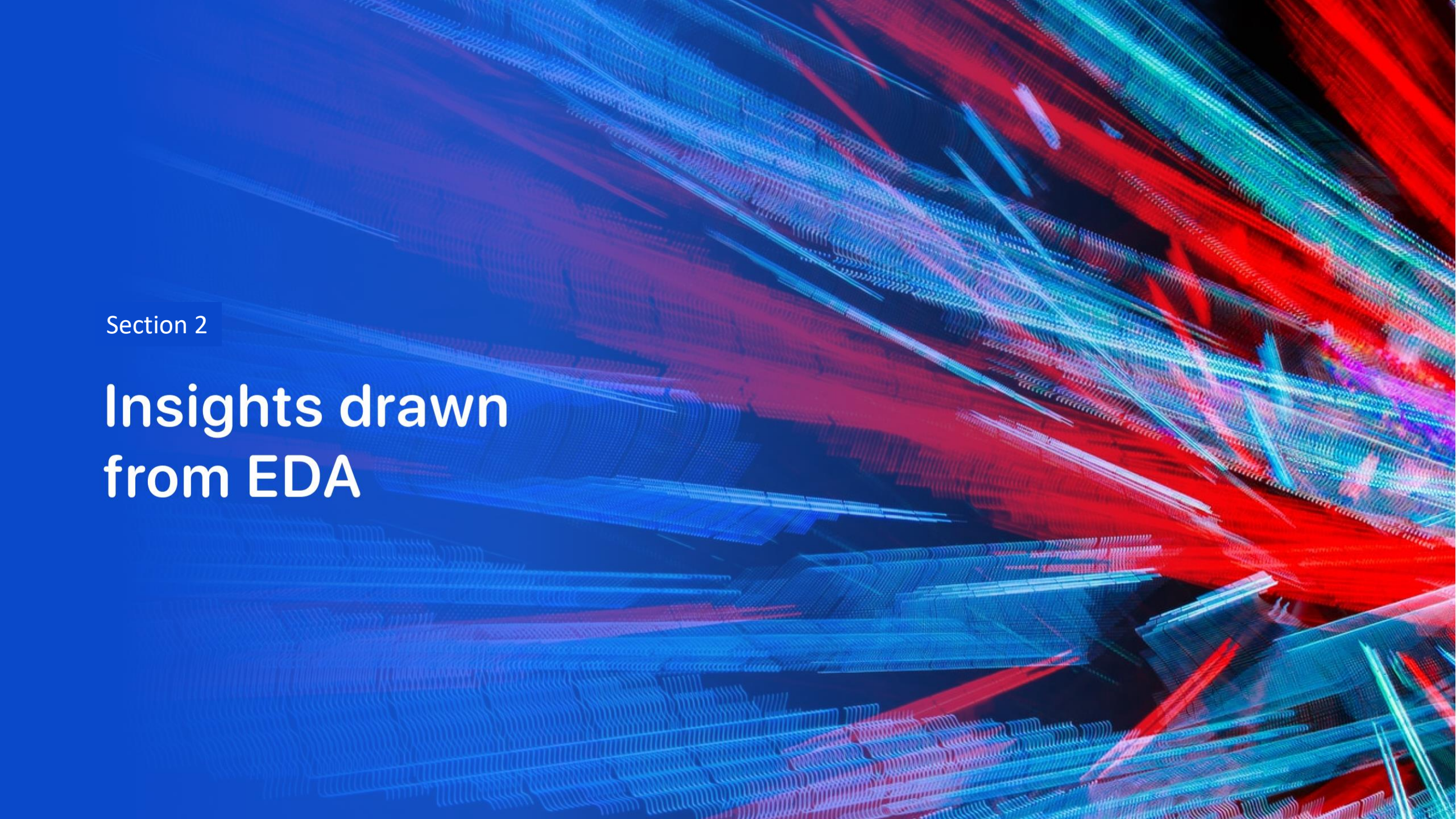
Notebook with application: dscap/spacex_dash_app.py at main · mattan87/dscap (github.com)

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.
- dscap/SpaceX_Machine_Learning_Prediction.ipynb at main · mattan87/dscap (github.com)

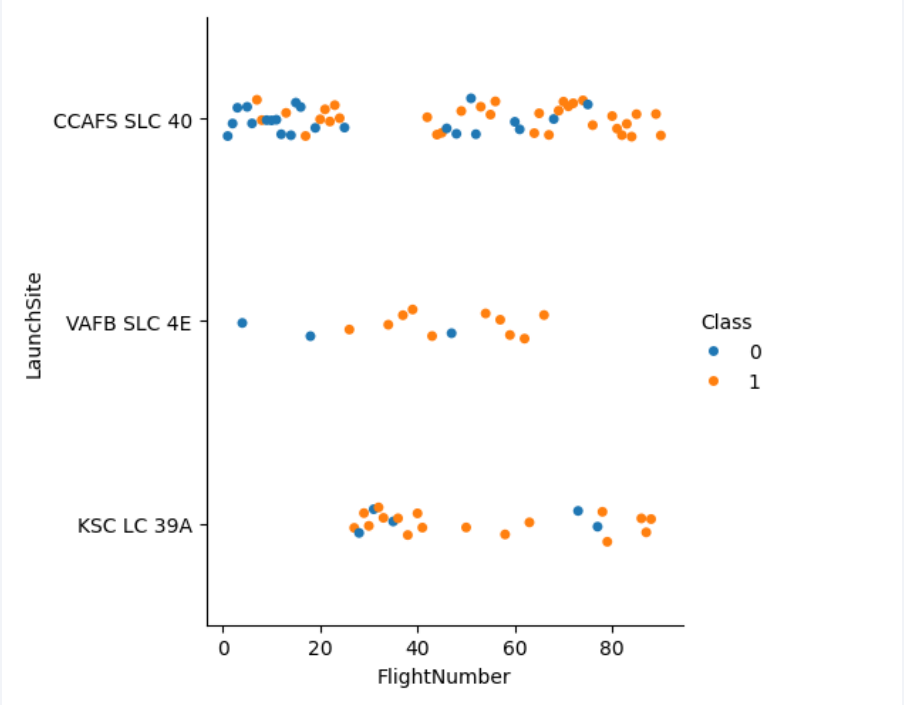
Results

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


Section 2
Insights drawn from EDA

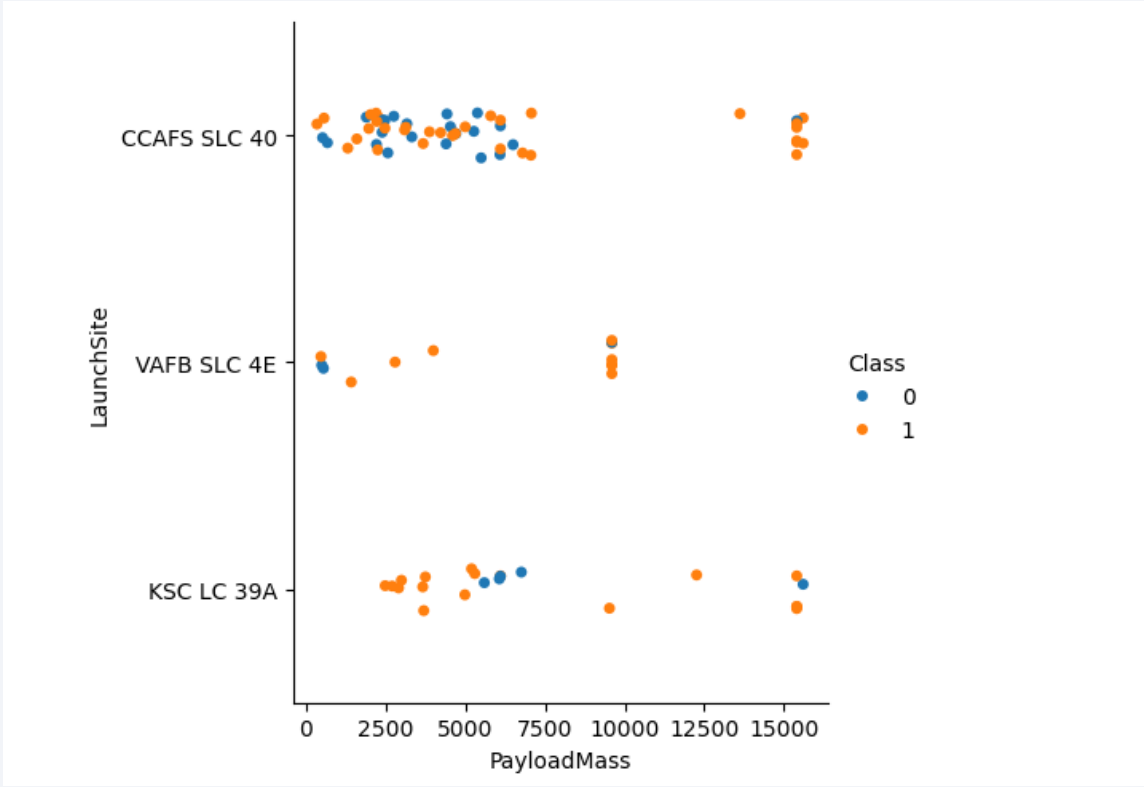
Flight Number vs. Launch Site

- From the plot, we found that the greater the flight number at a lunch site the more likely it would be a success
- Given a later flight number one can assume that learnings has been incorporated in to later flights which would also contribute towards more successful launches



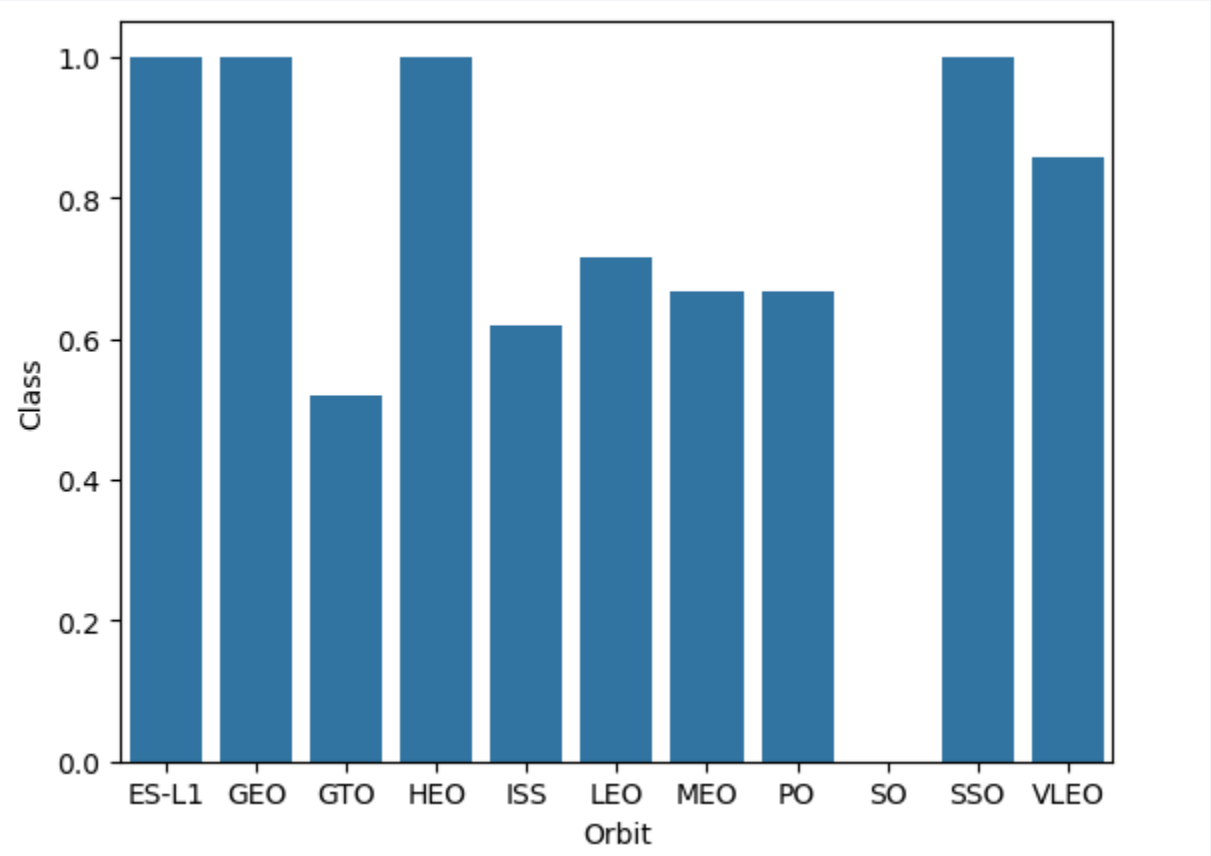
Payload vs. Launch Site

- The greater the payload mass for CCAFS SLC 40 the more likely it is a success.



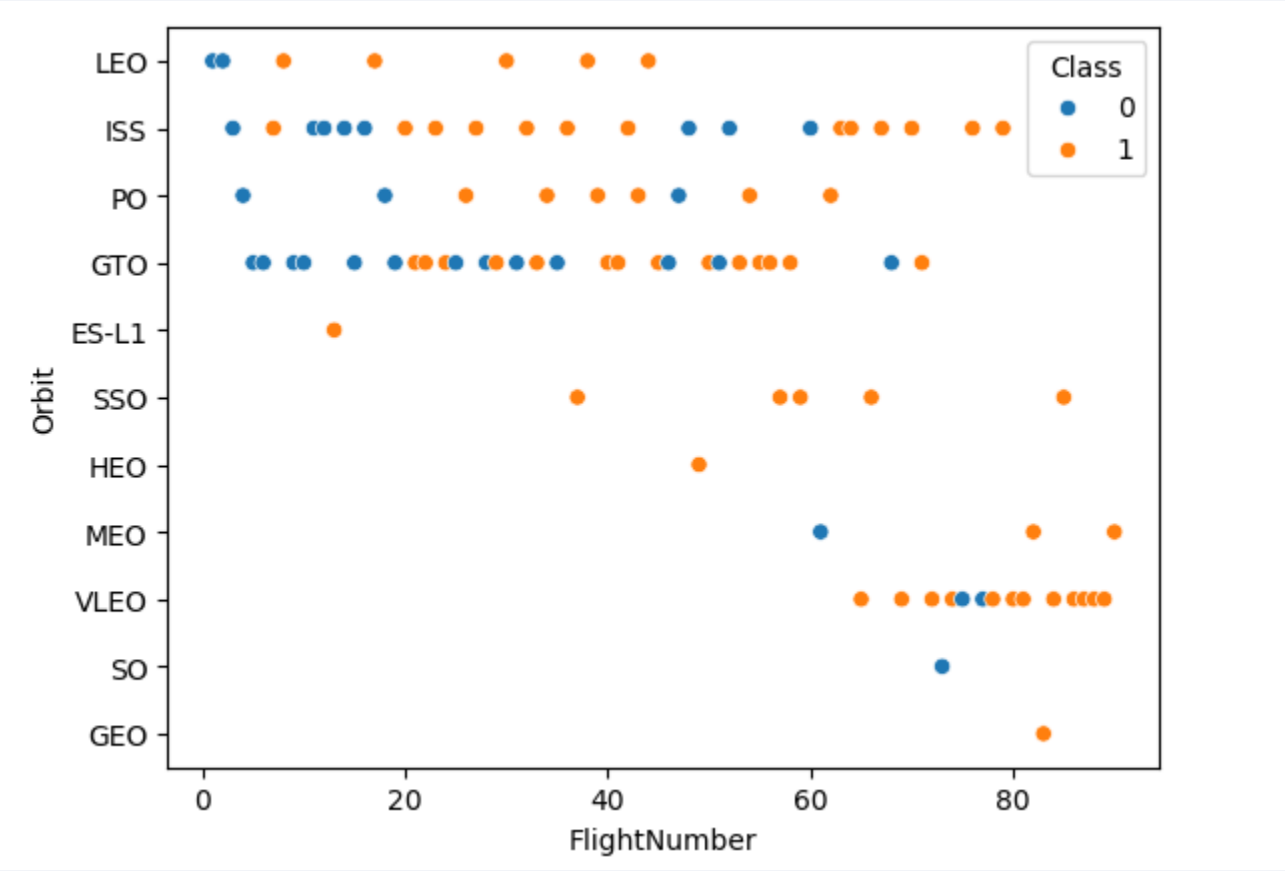
Success Rate vs. Orbit Type

- From the chart, ES-L1, GEO, HEO, SSO, VLEO are the orbit types with the greatest success rates.



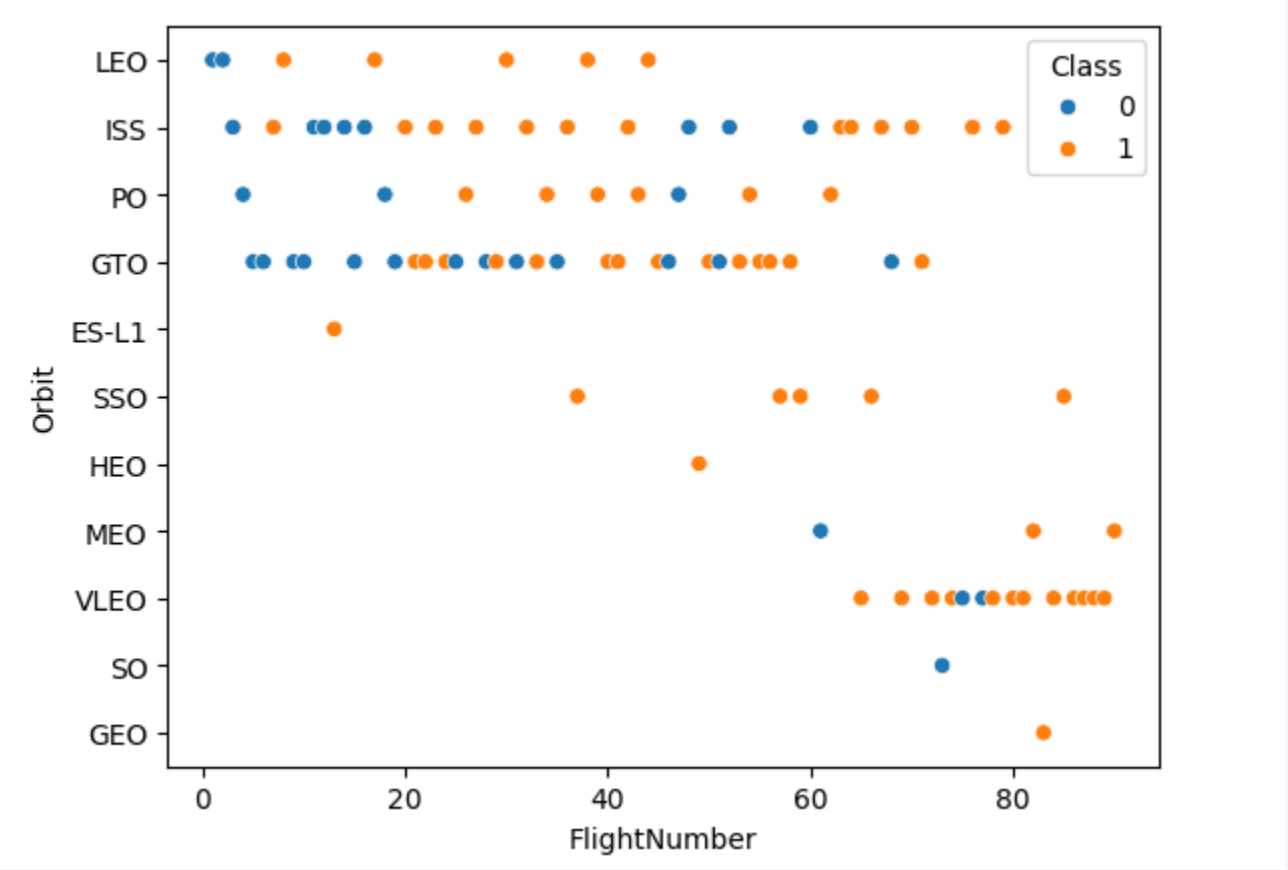
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



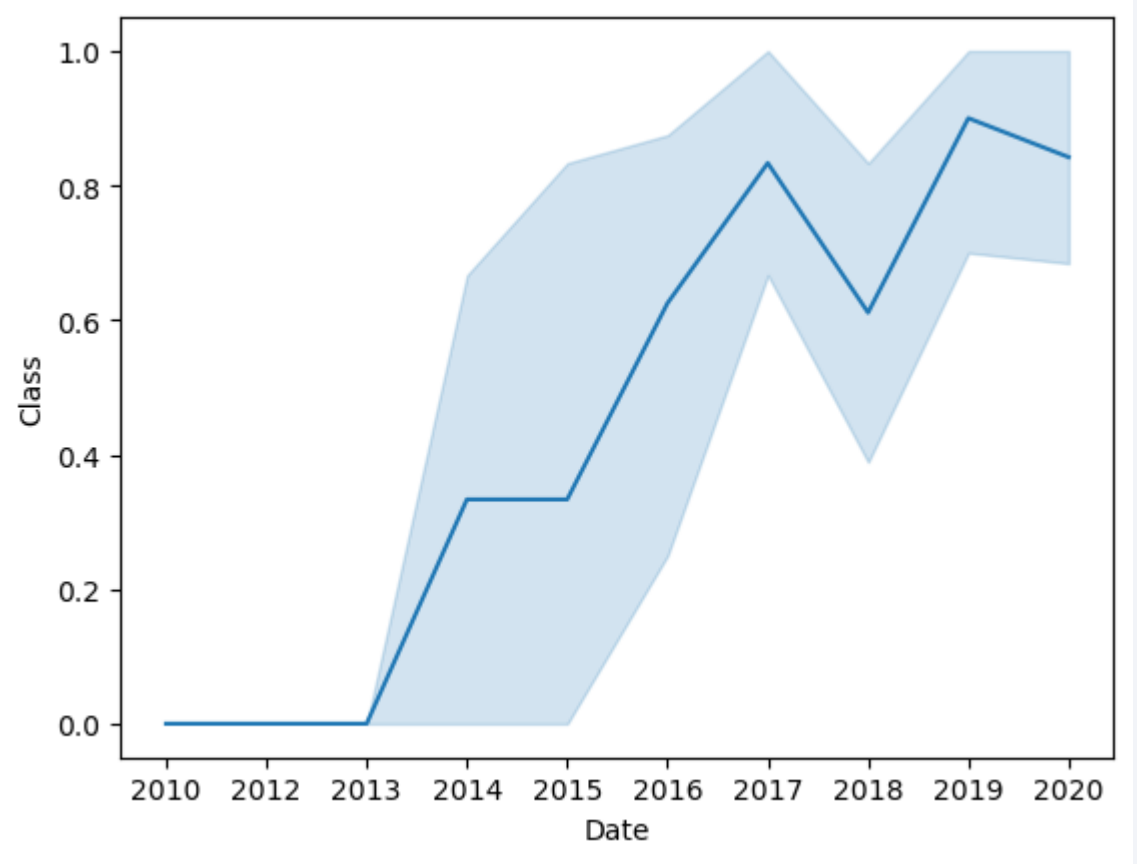
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.
- Launces from 2010 up until the end of 2012 were unsuccessful.



All Launch Site Names

- We used the key word **DISTINCT function** to show the unique launch sites from the SpaceX data.

```
In [11]: %sql SELECT DISTINCT launch_site FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

- We used the query below to display 5 records where launch sites begin with `CCA`

```
In [15]: %sql SELECT * FROM SPACEXTABLE WHERE launch_site LIKE "CCA%" LIMIT 5
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

- We calculated the total payload carried by boosters from NASA as 99980 kg using the query below

```
In [16]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE "NASA%"
 * sqlite:///my_data1.db
Done.
Out[16]: SUM(PAYLOAD_MASS__KG_)
                          99980
```

Average Payload Mass by F9 v1.1

- The average payload mass for booster version F9 V1.1 was calculated to 2534.6 kg

```
In [17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 V1.1%"
 * sqlite:///my_data1.db
Done.
Out[17]: AVG(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
%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

- Using a WHERE and BETWEEN clauses in the data we determined which booster versions that successfully landed on a drone ship with a payload mass between 4000 and 6000 kg

```
In [42]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ B
 * sqlite:///my_data1.db
Done.
```

Out[42]:

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

Total Number of Successful and Failure Mission Outcomes

- We calculated the mission outcomes using sub-queries. As one can tell almost all missions were successful whereas only one was unsuccessful.

```
In [41]: %sql SELECT DISTINCT (SELECT COUNT(MissIon_Outcome) FROM SPACEXTABLE WHERE Mission_Outcome LIKE "%Success%") AS No_Success,
 * sqlite:///my_data1.db
Done.
```

Out[41]:

No_Success	No_Fail
100	1

Boosters Carried Maximum Payload

- Using below query and sub-query we were able to derive a list of all the booster versions carrying the MAX payload mass.

```
In [46]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

Out[46]:

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

- Using a combination of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes on drone ships, their booster versions, and launch site names for the year of 2015

```
In [58]: %sql SELECT SUBSTR(date,0,5) AS Year, SUBSTR(date, 6,2) AS Month_Name, Landing_Outcome, Booster_Version, Launch_Site FROM SP
 * sqlite:///my_data1.db
Done.
```

Out[58]:

Year	Month_Name	Landing_Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	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.

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Count FROM SPACEXTABLE
GROUP BY Landing_Outcome HAVING DATE(date) BETWEEN "2010-06-04" AND "2017-03-
20" ORDER BY Count DESC
```

Out[65]:

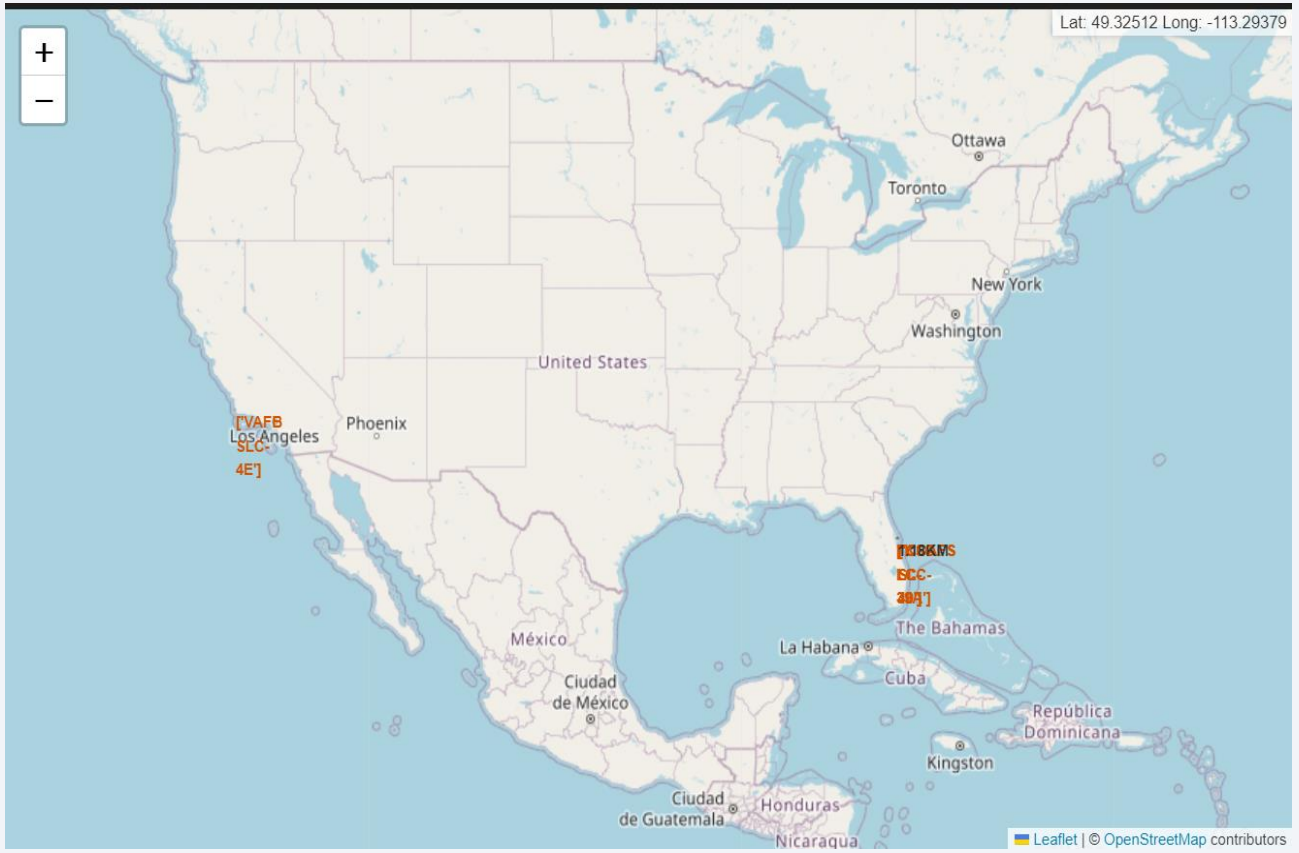
Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

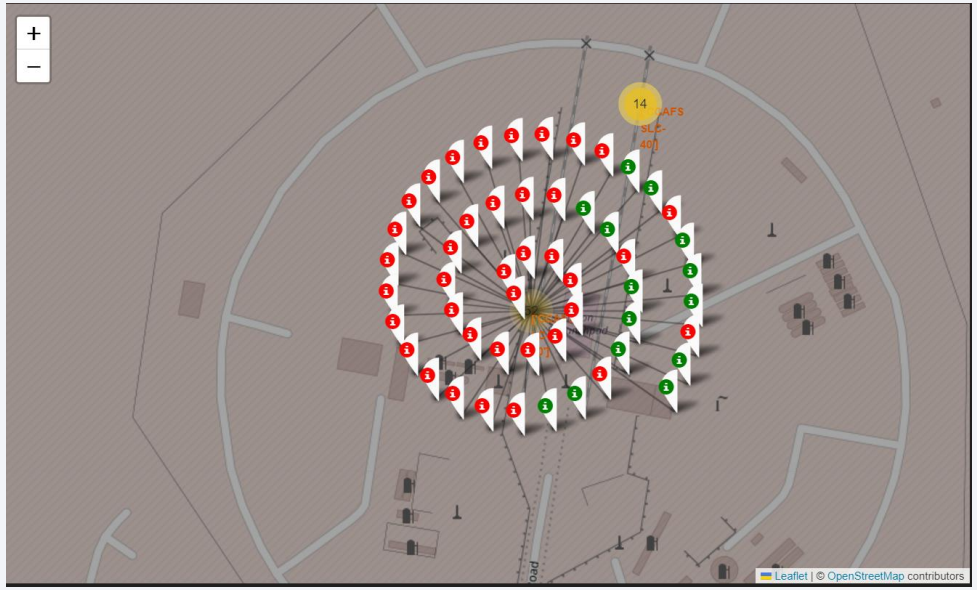
SpaceX Launch Sites

- SpaceX Launch Sites are located the United Sates
- All Launch Sites are close to the ocean and in states that offer a warm climate.
- Launch Sites are in close proximity to the ocean due to the use of both ditching landing methods (Crash landings in the sea) and by remotely operated sea landing drones.



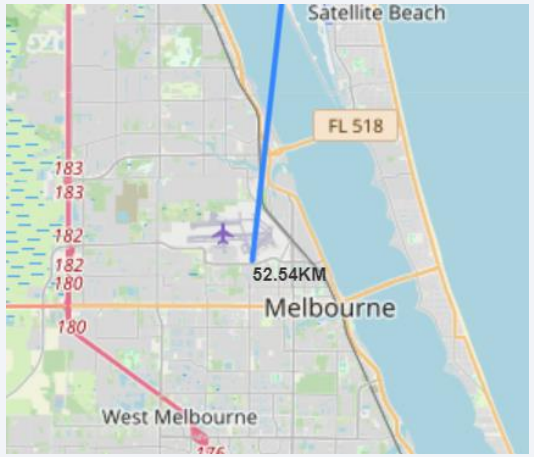
Successful & Failed Launches

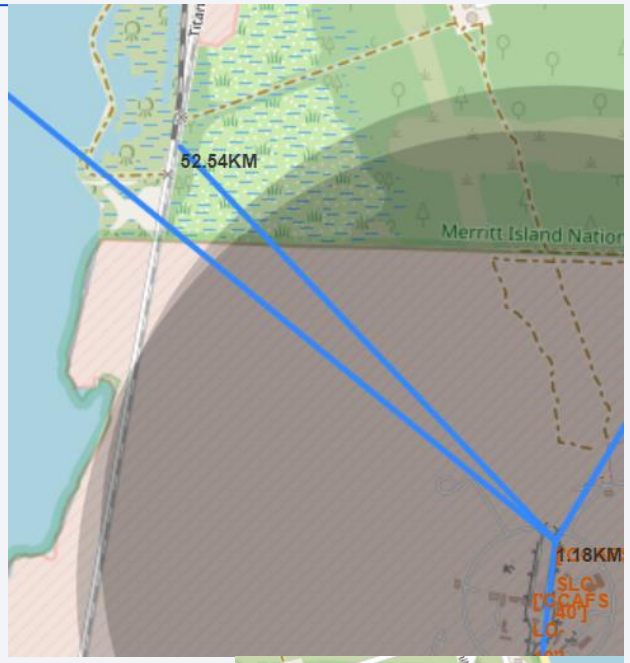
- Green markers represent the successful launches and the Red Marker represents the failures.

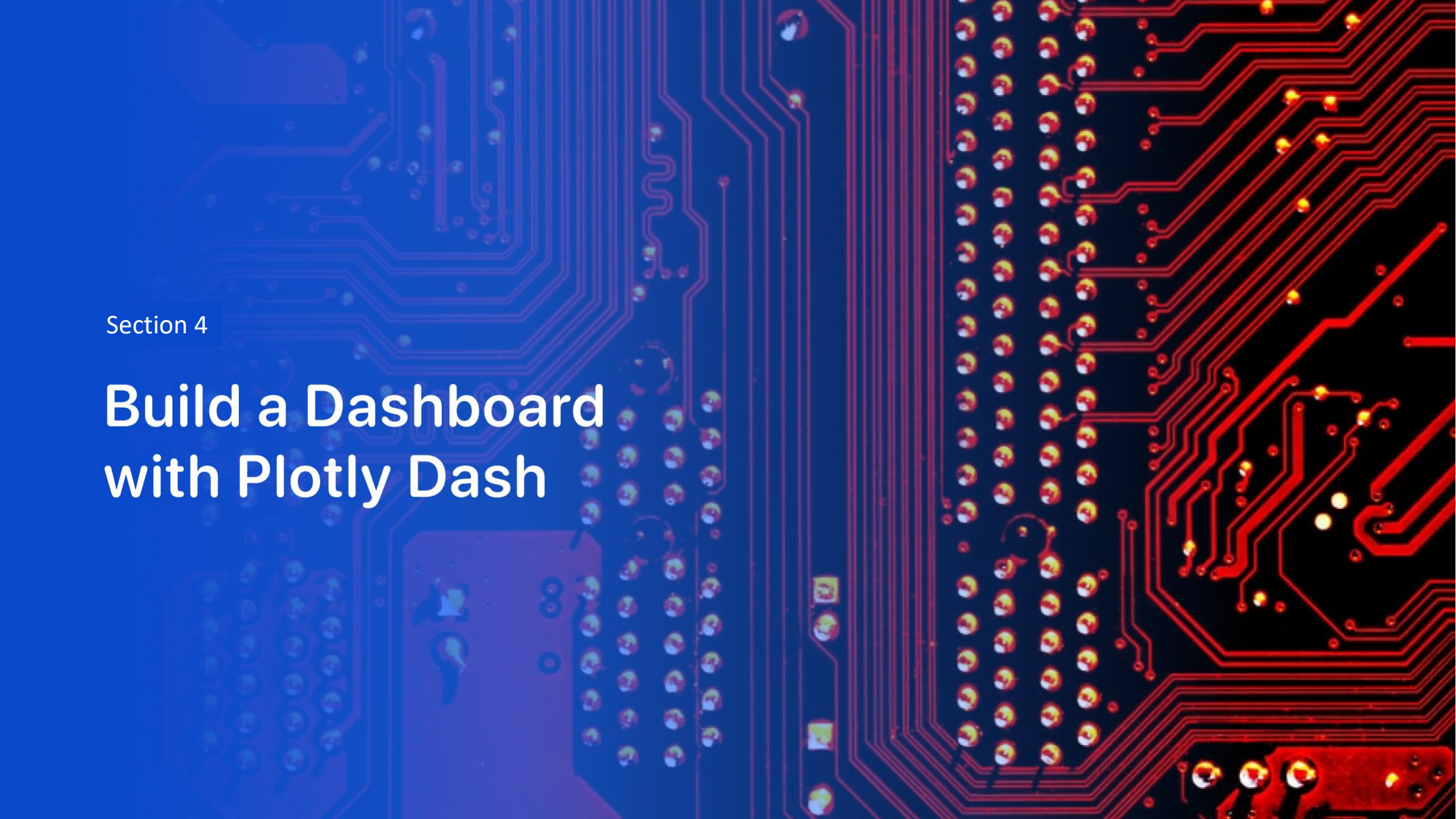


Launch Site distance to landmarks

- Normally Launch Sites are located in the vicinity of different landmarks such as major cities (Melbourne) and other key infrastructure objects such as waterways, rail roads and major roads as demonstrated by the maps to the right.

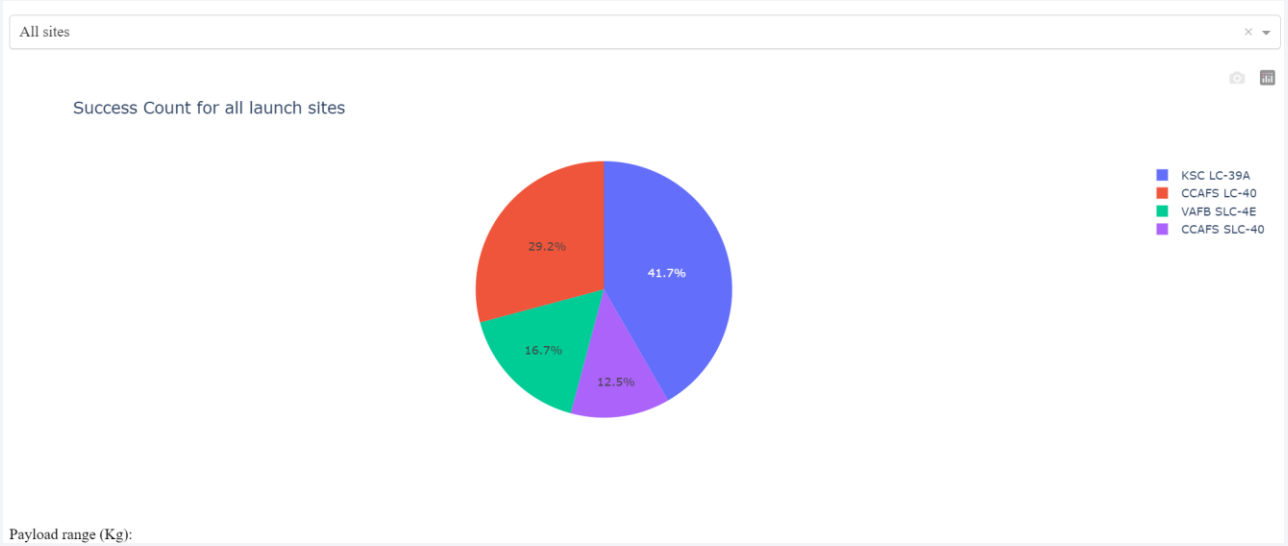





Section 4
Build a Dashboard with Plotly Dash

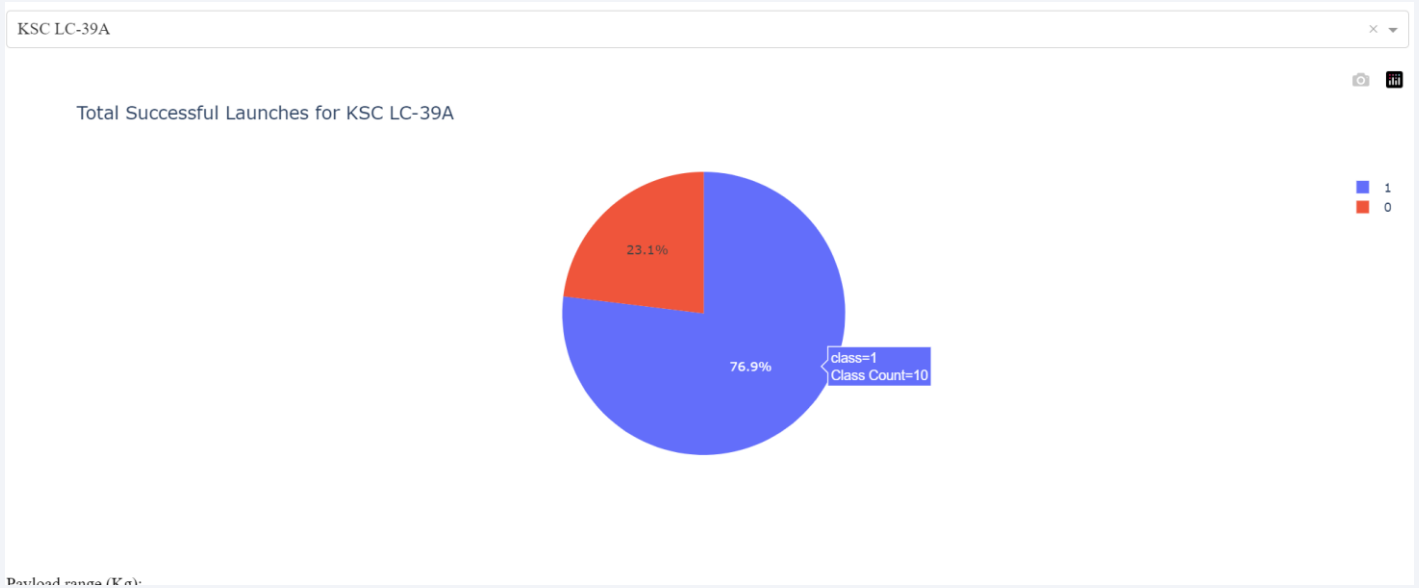
Success Percentage by Launch Site

- KSC LC-39A is the site with the highest number of successful launches of all sites



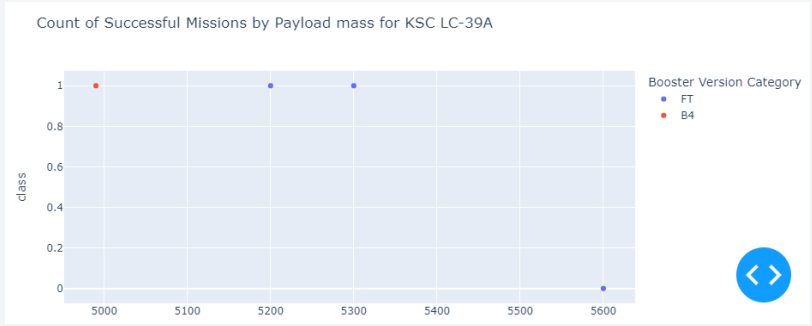
KSC LC39A – The most successful site

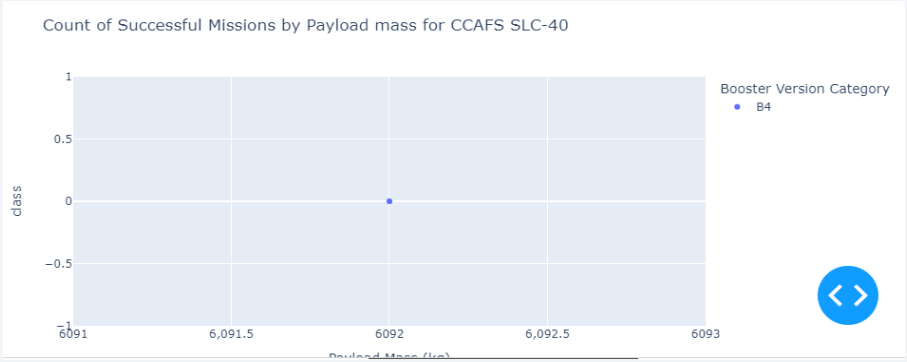
- Launch site KSC LC-39A has a 76.9% success rate and only has a 23.1% failure rate

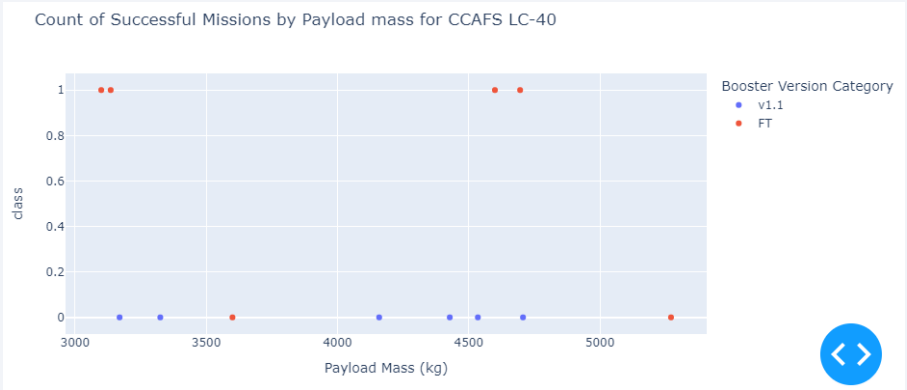


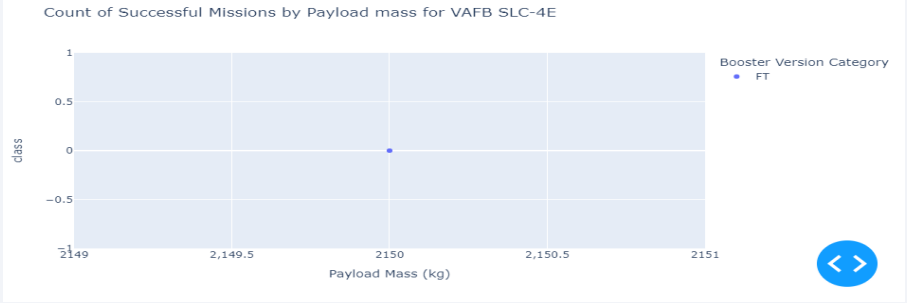
Payload mass for all sites

- Data looks different for all sites.
  However, a majority of the cases include Booster Version Category FT.







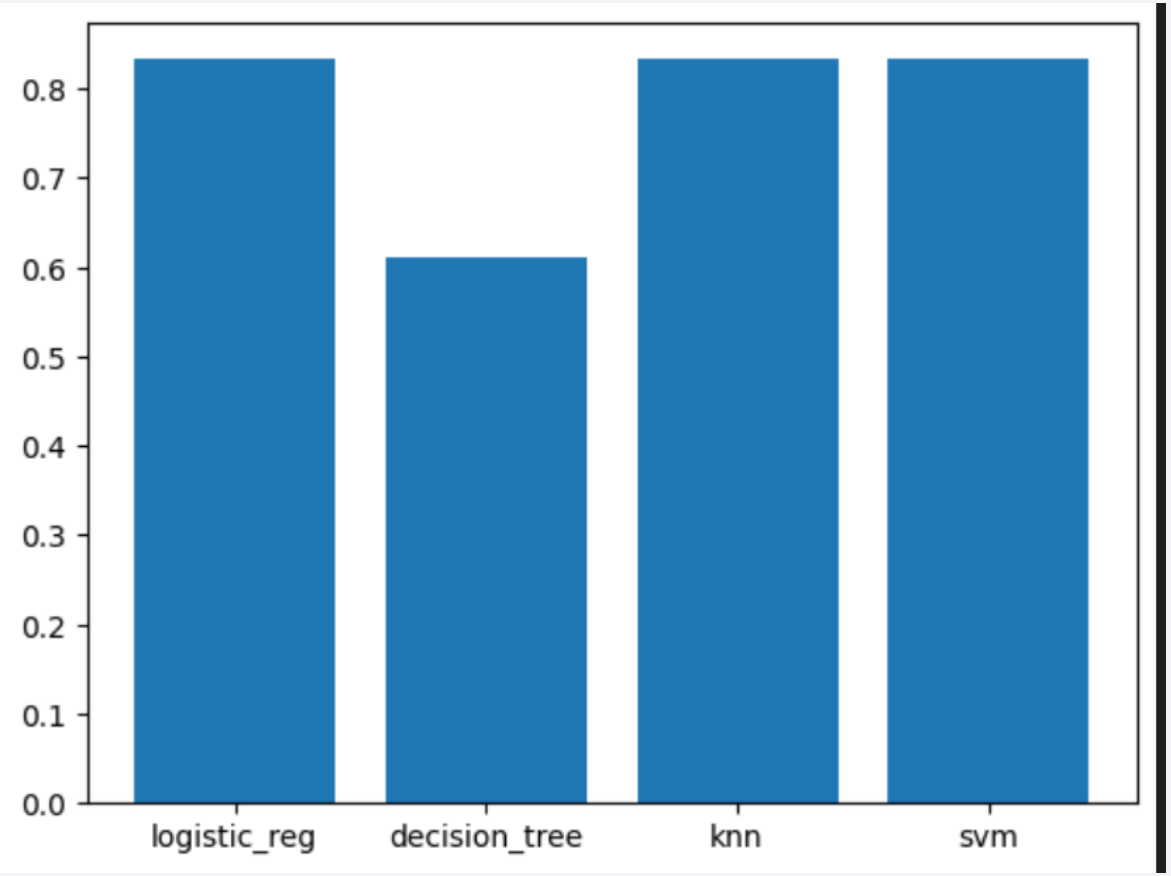


Section 5

Predictive Analysis (Classification)

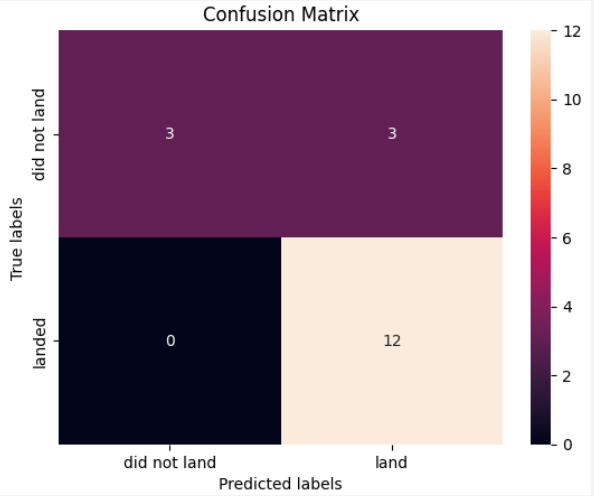
Classification Accuracy

- The model, or models in this case, with the highest classification accuracy (~83%):
  - K-nearest neighbour, Logisitc Regression and Support Vector Maching (SVM)



Confusion Matrix

- The confusion matrix for KNN, SVM and Logistic Regression all look the same. It shows that the classifier can distinguish between the different classes (Successful, Unsuccessful landing). The major problem with our model, impacting the model's accuracy, are the false positives.
- The false positive problems could be due to overfitting the model since only 20 % of our dataset is used for test / evaluation of the model, whereas 80 % of the data is used to train the model.



Conclusions

Based on the findings made we conclude that:

- The greater number of launches taken place at a site, the greater the success rate is at the launch site.
- Launch success rate start to increase from 2013 up until 2020
- Orbits ES-L1, HEO, SSO, VLEO and GEO had the largest success rate of all of the different orbit types
- KSC LC-39A is the launch site the the most successful launches of any site
- There are three classification models that are just as accurate in predicting wether recovery of first stage will be successful or not. These are:
    - KNN, SVM, and Logistic Regression

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!