IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Syu As
20 June 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost up to 165 million dollars each, of which SpaceX could save more and they can reuse the first stage

- Problems you want to find answers

The project task is to predict if the first stage of the SpaceX Falcon 9 rocket will successfully land

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Through SpaceX Rest API and Web scrapping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, kNN, SVM, DT models have been built and evaluated for the best classifier

Data Collection

- Describe how data sets were collected.

Download a .json file containing rocket launch data from the SpaceX API

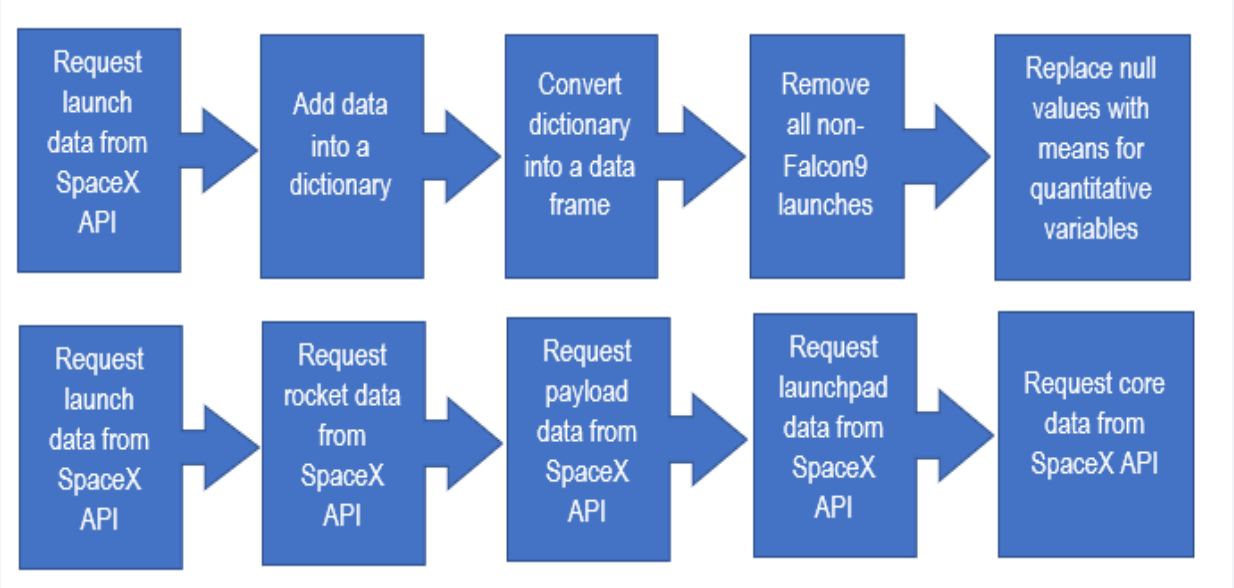
Parse the information into dataframe

Convert null values of quantitative variables into the mean of the rest of the column

- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

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



- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- Github URL: Data Collection - SpaceX API

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

Github URL: Web scrapping

Data Wrangling

- Describe how data were processed

The project is to determine whether a recovery was successful or not

According to data frame, there are 8 different outcomes:

1. True RTLS: Successful landing on a ground pad
2. True ASDS: Successful landing to drone ship
3. True Ocean: Successful landing in ocean
4. None ASDS: Failed to land
5. None None: Failed to land
6. False RTLS: Failed landing on a ground pad
7. False ASDS: Failed landing to drone ship
8. False Ocean: Failed landing in ocean

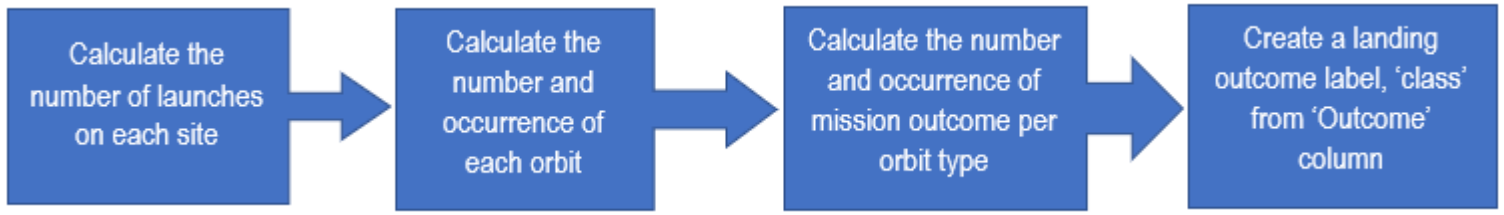
We could create a new column, 'class' to delineate between successful and unsuccessful recoveries

1 – successful recovery

2 – unsuccessful recovery

Data Wrangling

- You need to present your data wrangling process using key phrases and flowcharts



- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

  GitHub URL: Data wrangling

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

  Charts plotted:
  - Flight number vs Launch Site (Cat plot / Scatter point chart)
  - Flight number vs Payload Mass (kg) (Cat plot / Scatter point chart)
  - Success Rate vs Orbit type (Bar plot)
  - Orbit type vs Flight Number (Cat plot / Scatter point chart)
  - Orbit type vs Payload (Cat plot / Scatter plot)
  - Success rate vs Time in years (Line plot)

  Why:
  - Cat plot / Scatter point chart – shows relationship between a numerical and categorical variables
  - Bar plot – shows the mean value and represents an estimate of central tendency for a numeric variable
  - Line plot - shows an estimate of the central tendency and a confidence interval for that estimate.

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

  GitHub URL: EDA with visualization

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Launch site
  - Payload Mass (kg)
  - Mission Outcome
  - Booster version
  - Date
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

  GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - View the launch site of each Falcon 9, represented by a circle
  - Learn how many launches occurred at each location, represented by markers – green markers represent a successful recovery, red markers represent unsuccessful recovery
  - Determine distances to the closest coastline, city, railway and highway, represented by a blue line
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

GitHub URL: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions

  Built an interactive dashboard with Plotly Dash

  Plotted pie charts showing the total launches by a certain sites

  Plotted scatter graph showing the relationship between Outcome and Payload Mass (kg) for the different booster version

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

  GitHub URL: Plotly Dash

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

  - Loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
  - Built different machine learning models and tune different hyperparameters using GridSearchCV.
  - Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
  - Found the best performing classification model.

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

GitHub URL: Prediction analysis

Results

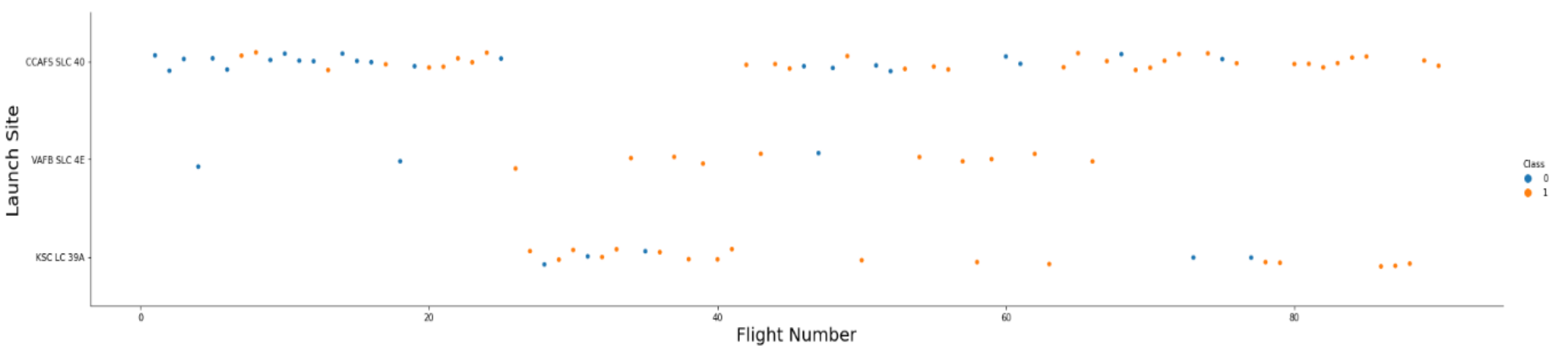
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

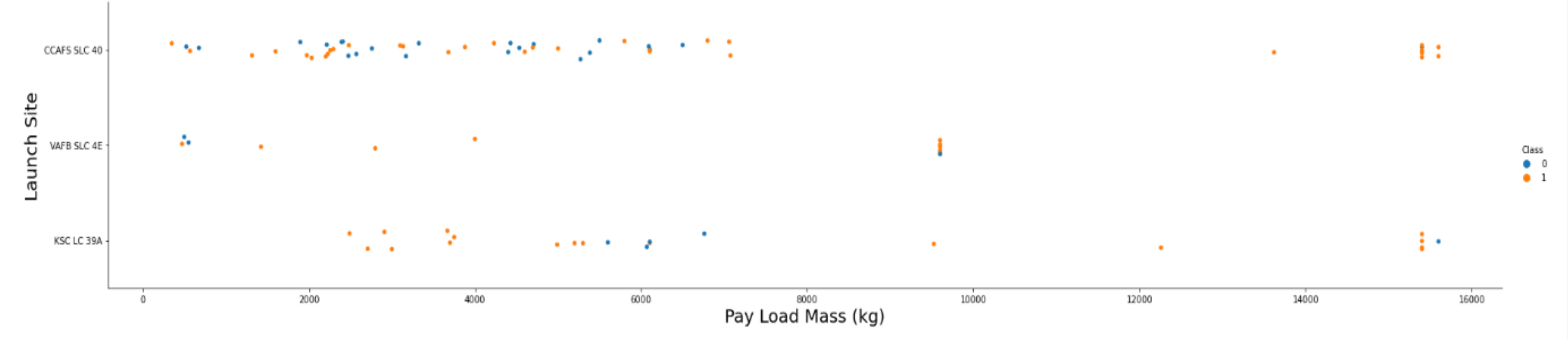


- Show the screenshot of the scatter plot with explanations

The larger the flight amount at a launch site, the greater the success rate at a launch site.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

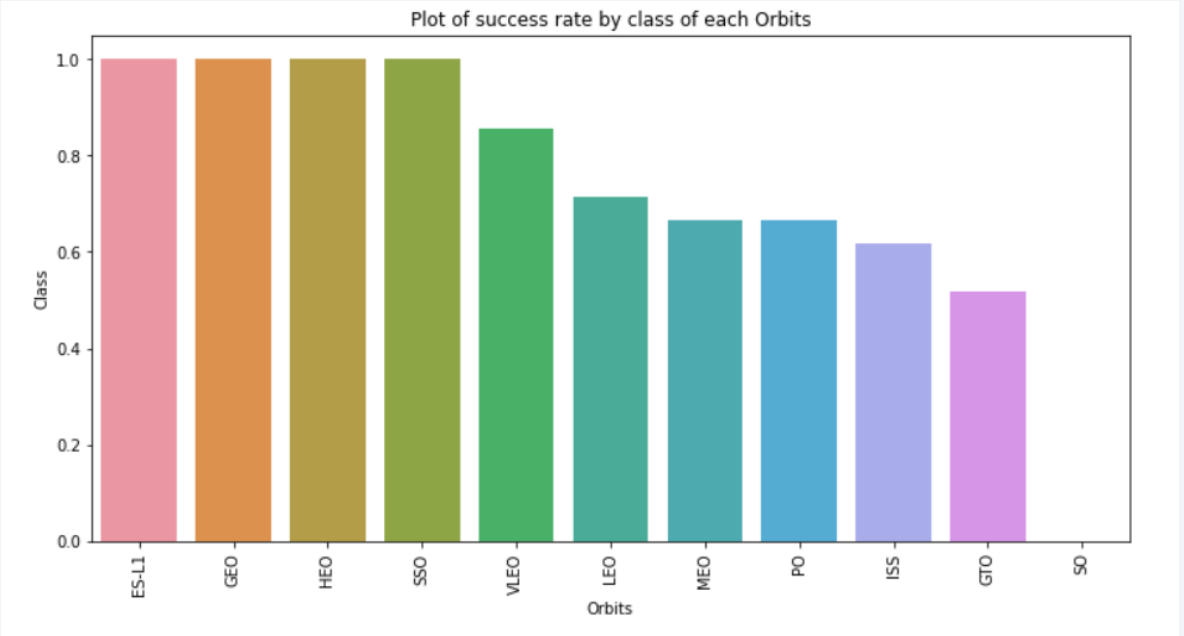


- Show the screenshot of the scatter plot with explanations
  - No rockets launched for heavy payload mass (greater than 10000 kg) for VAFB-SLC

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

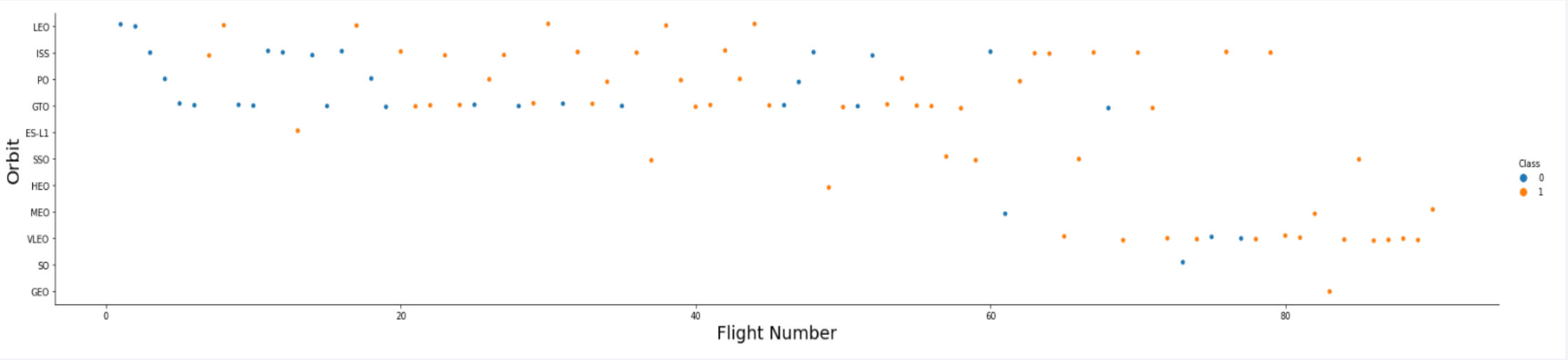
- Show the screenshot of the scatter plot with explanations



ES-L1, GEO, HEO, SSO, and VLEO are the Orbits that have high success rate.
The SO has the least success rate amongst the orbits.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

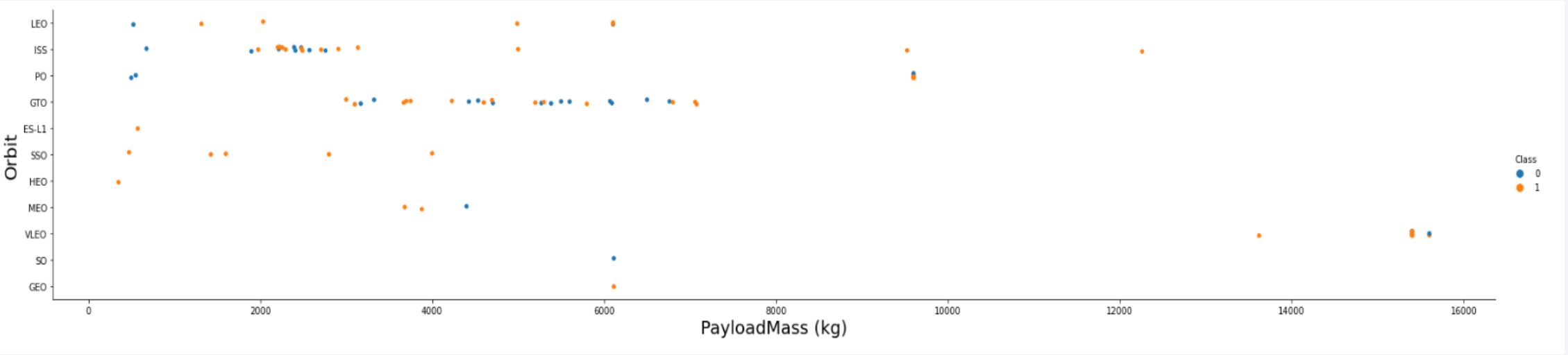


- Show the screenshot of the scatter plot with explanations

In 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

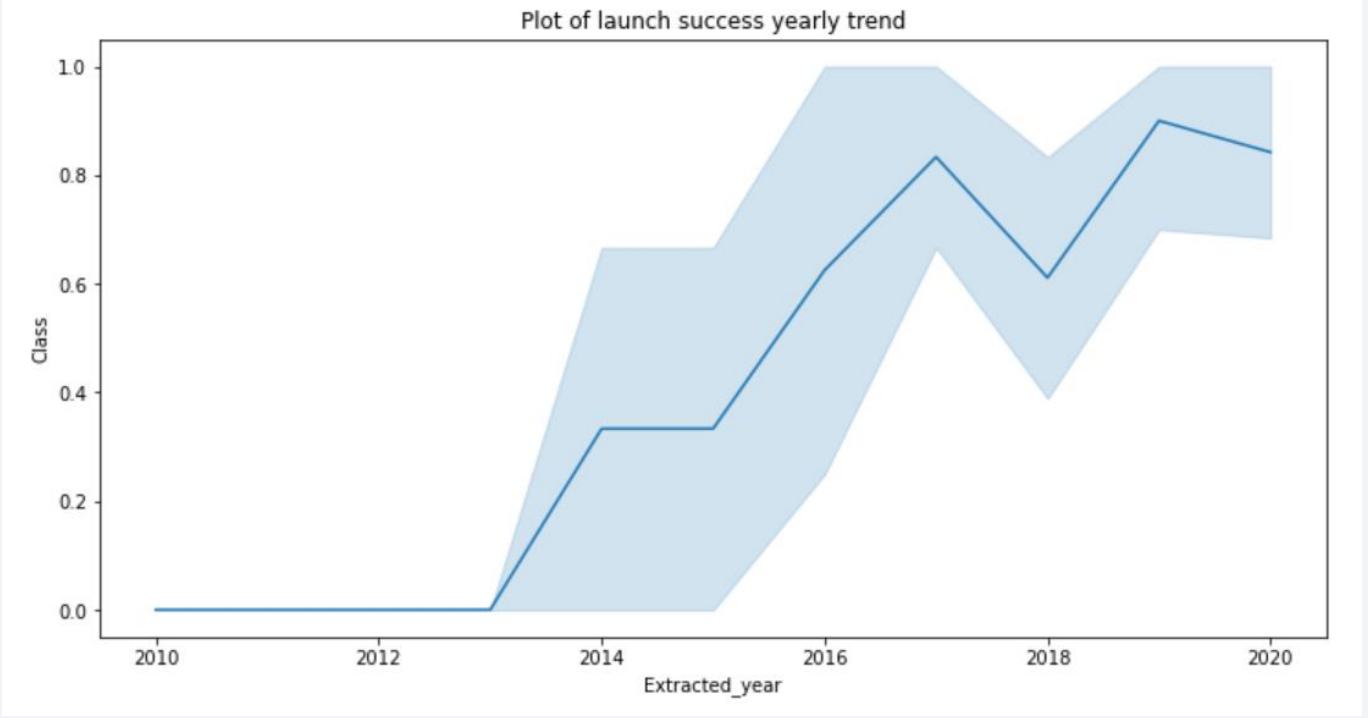
- Show a scatter point of payload vs. orbit type



- Show the screenshot of the scatter plot with explanations
  - 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

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



Success rate is on an increasing trend, from 2013 to 2020

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

```
%sql SELECT DISTINCT(LAUNCH_SITE) from SPACEXTBL order by LAUNCH_SITE ASC;
```

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

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

```
%sql select * from SPACEXTBL 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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
- Present your query result with a short explanation here

Display the total payload mass carried by boosters launched by NASA (CRS)

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

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

sum (PAYLOAD_MASS__KG_)
45596

The total payload carried by boosters from NASA is 45596 kg as per query above

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version ='F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9 v1.1 is 2928.4 kg as per query above

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(Date) from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)';
```

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

min(Date)
2015-12-22

The first successful landing outcome on ground pad was on 22-12-2015

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
- Present your query result with a short explanation

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version from SPACEXTBL where [Landing _Outcome] = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

List the total number of successful and failure mission outcomes

```
%sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome like '%Success%';
```

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

count(Mission_Outcome)
100

```
%sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome like'%Failure%';
```

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

count(Mission_Outcome)
1

Used wildcard like '%' to filter for WHERE Mission_Outcome was a success or failure

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max (PAYLOAD_MASS__KG_) from SPACEXTBL) order by Booster_Version ASC;
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here

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

```
%sql select substr(Date, 4, 2) as Month , substr(Date, 7, 4) as Year, Date, Mission_Outcome, [Landing _Outcome], Booster_Version, Launch_Site from SPA
```

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

Month	Year	Date	Mission_Outcome	Landing _Outcome	Booster_Version	Launch_Site
01	2015	10-01-2015	Success	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	14-04-2015	Success	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
- Present your query result with a short explanation here

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], COUNT (*)
FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017'
group by [Landing _Outcome]
order by COUNT (*) DESC
```

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

Landing _Outcome	COUNT (*)
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

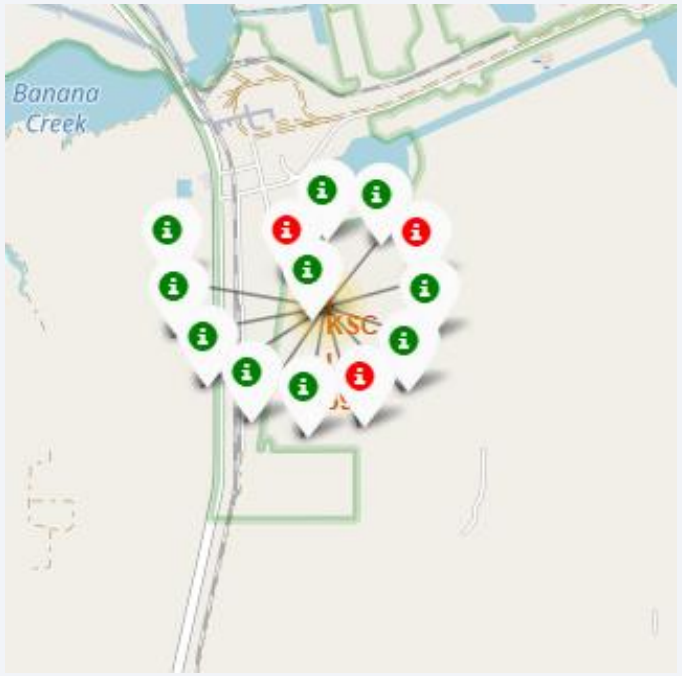
All launch sites global map markers

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

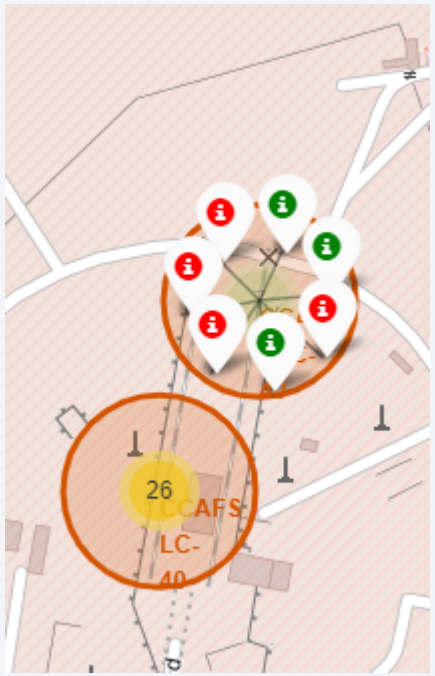
- Explain the important elements and findings on the screenshot

  All SpaceX launch sites are in United States of America coasts, Florida and California

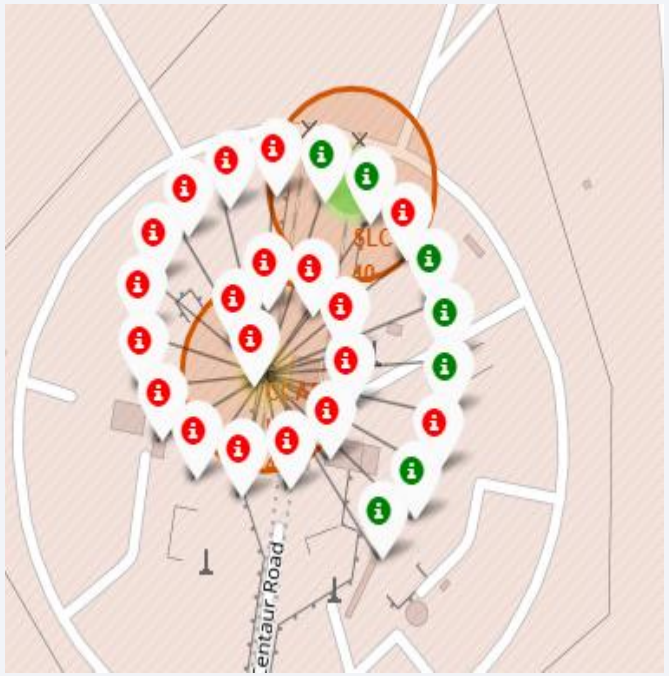
Success/Failed launches for each site



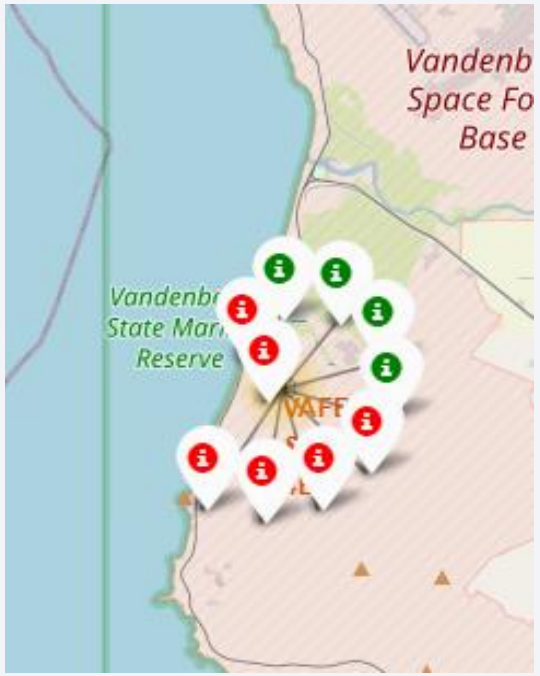
KSC LC-39A



CCAFS SLC-40



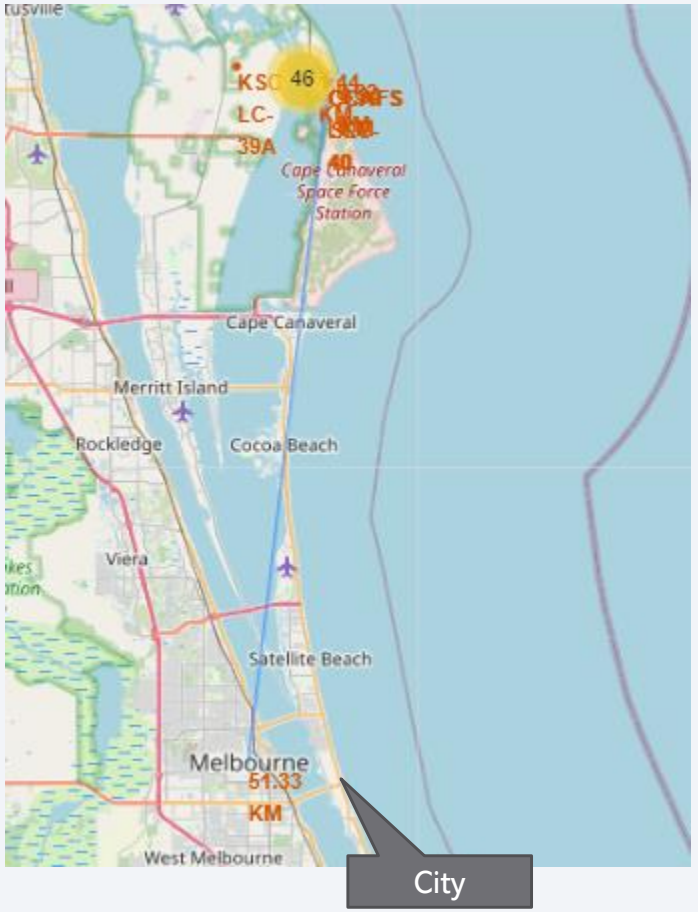
CCAFS LC-40



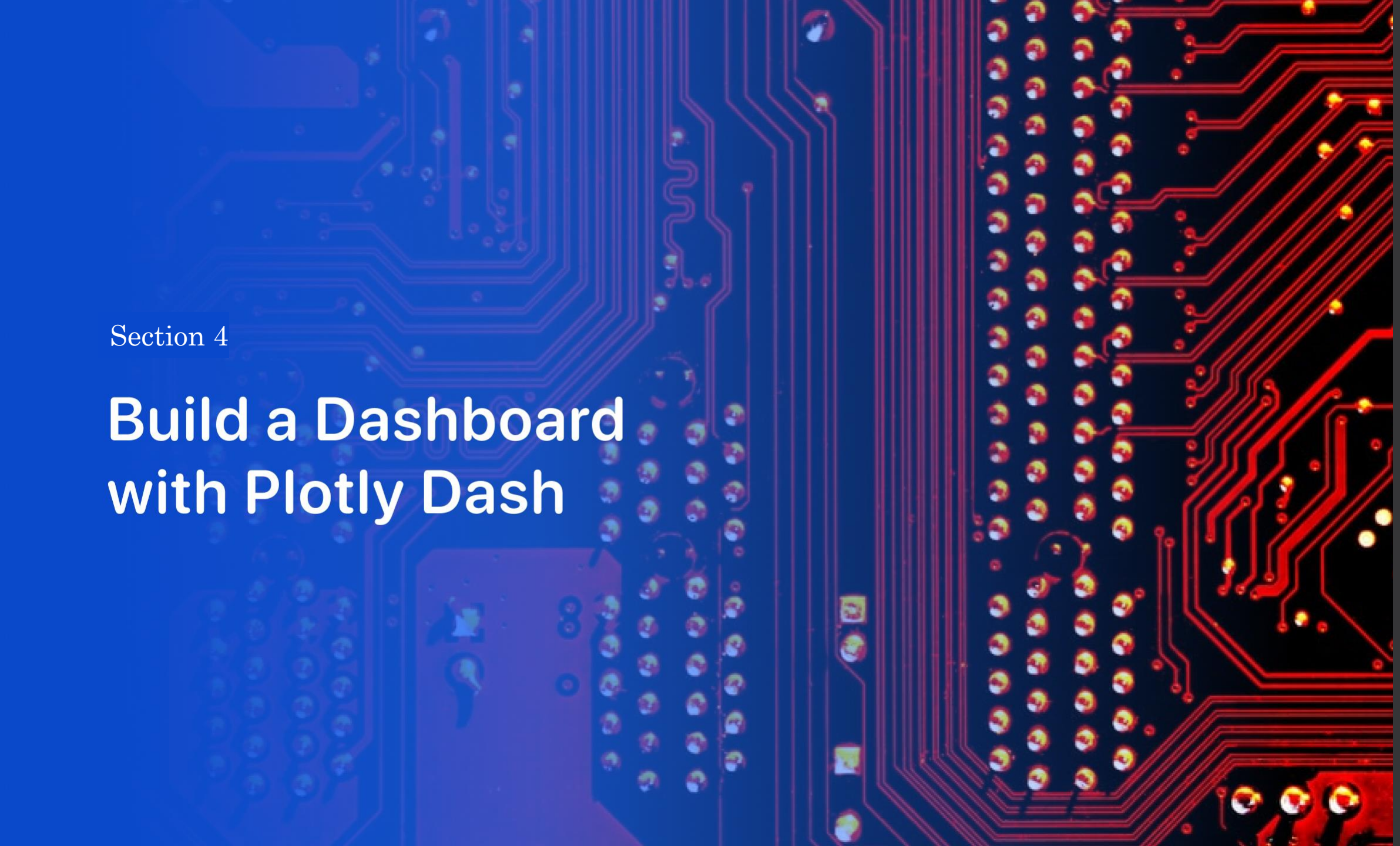
VAFB SLC-4E

Green Marker – Successful launches
Red Marker – Failed launches

Launch Site distance to its proximities



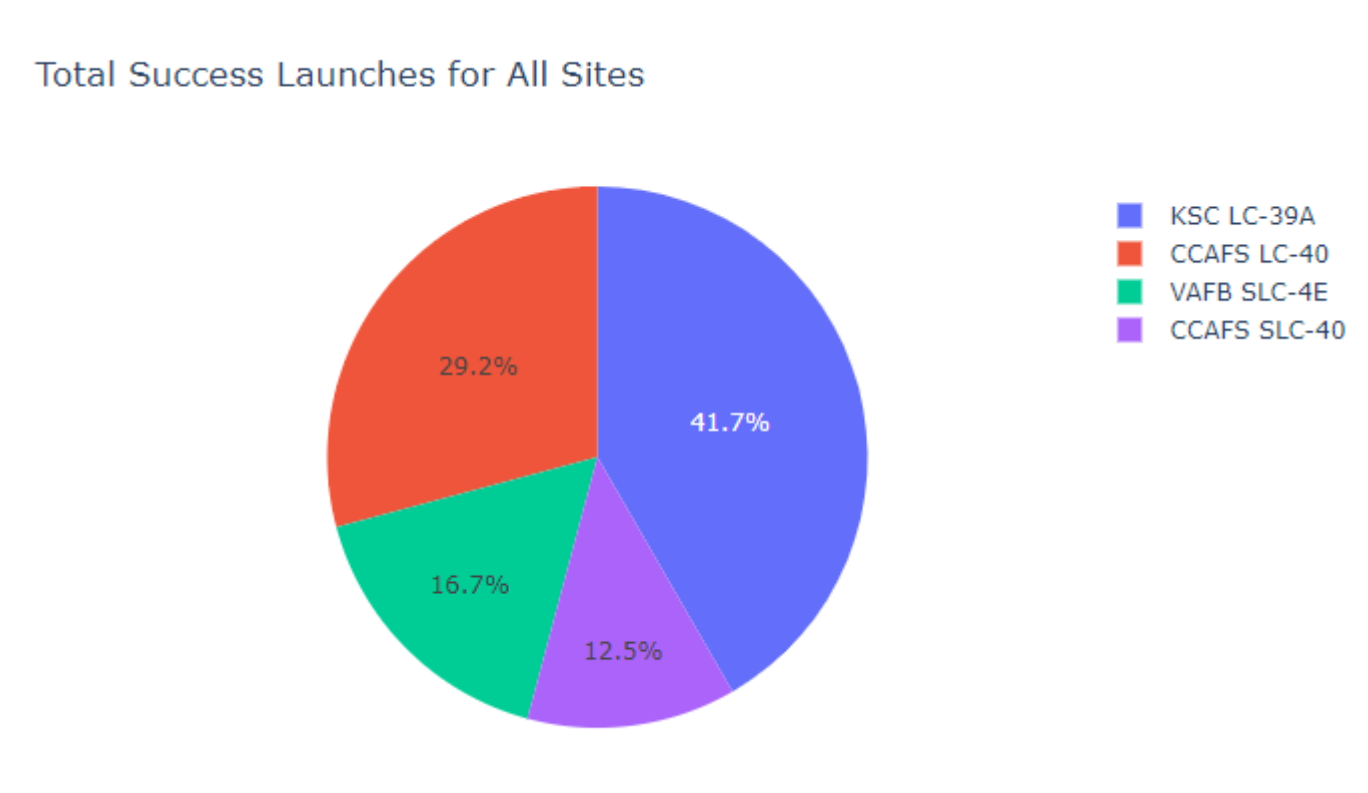
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



Section 4

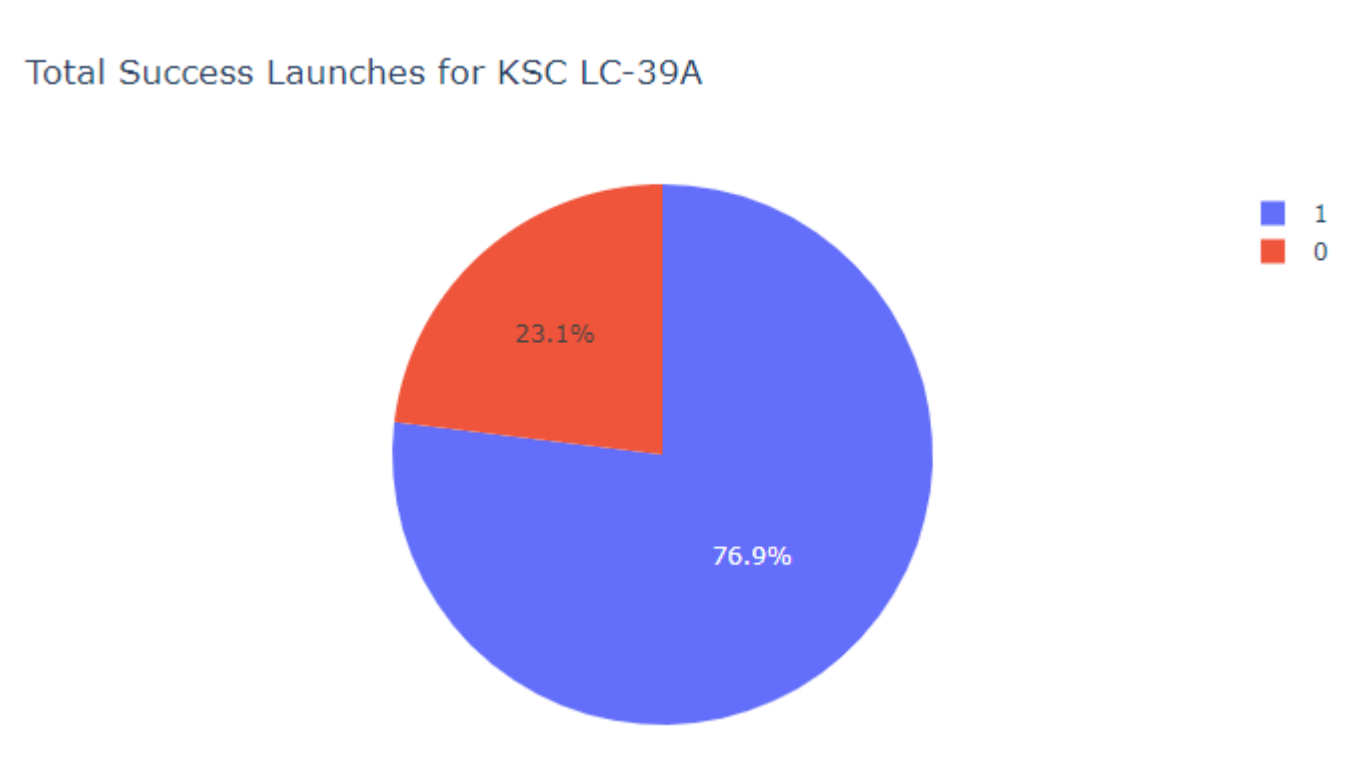
Build a Dashboard with Plotly Dash

Success launch rate for all sites



- KSC LC-39A was the site with highest success rate (41.7%)
- The remaining 3 launch sites success rate were below 30%

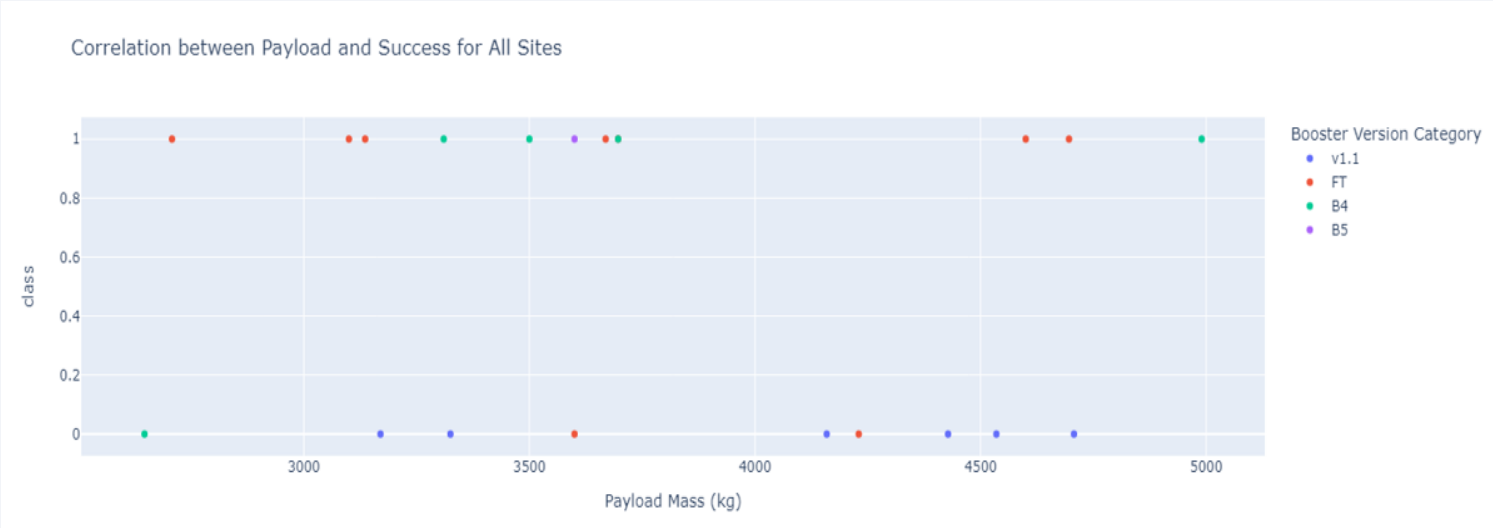
Launch site with highest success rate: KSC LC-39A



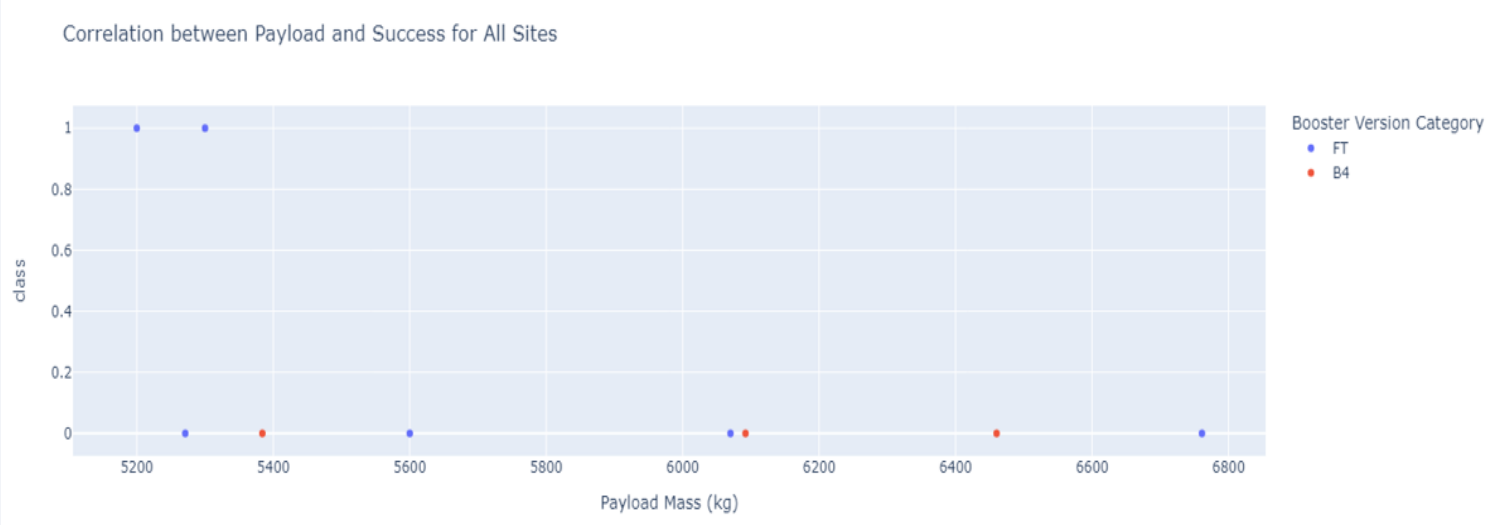
- Success Rate for KSC LC-39A was at 76.9% while the failure rate was at 23.1%

Payload Mass (kg) vs Success Rate for all sites

Low weighted payload, 2500 kg – 5000 kg



Heavy weighted payload, 5000 kg – 7500 kg



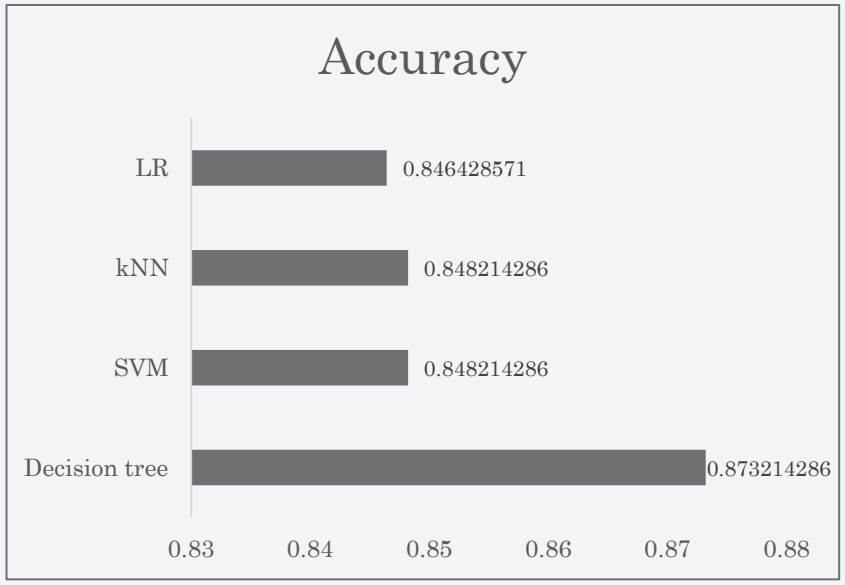
- Success rate for low weighted payload is higher than the heavy weighted payload

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy



Find the method performs best:

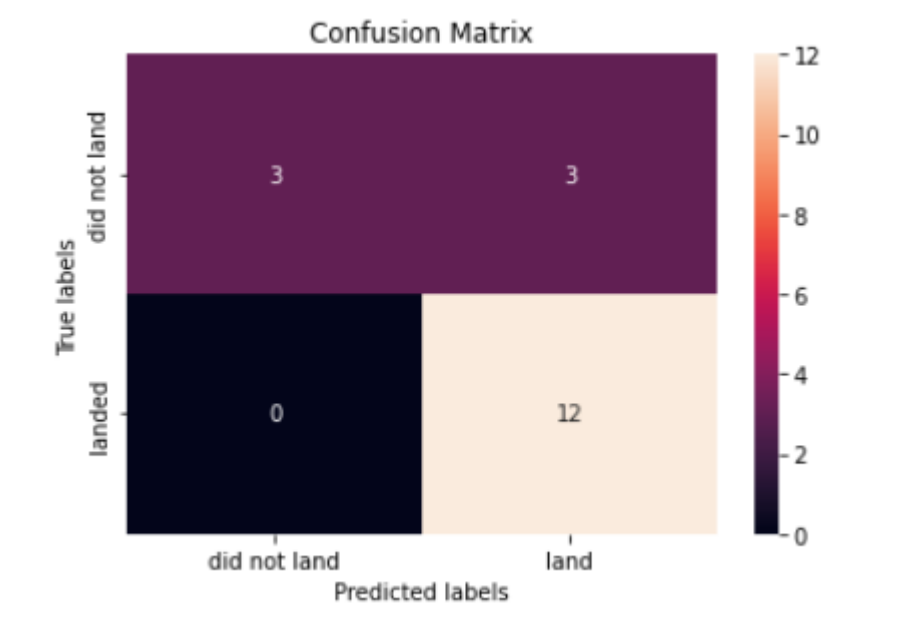
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5,
'splitter': 'random'}
```

- Best model: Decision tree with the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.

Problem: False positive i.e unsuccessful landing marked as successful landing by the classifier.

Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Success rate is on an increasing trend from 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this project.

Thank you!