IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Achimugu A . S

20 May 2022

Outline

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

Executive Summary

- This project data were collected via SpaceX REST API and web scrapping.
- Analysis of data specifically for different launching site shows that the maximum number of launches for Falcon 9 Rockets were from Cape Carnival site and GTO orbit accounted for maximum deployment.
- Decision Tree algorithm performed best among other classification algorithms to predict the successful landing of Falcon 9 Rockets.
- KSC LC launch site showed the highest success rate while Orbit of payload deployment and payload mass were found to have good correlation with the success of the mission.

Introduction

- **Project background and context**
- We predicted if the Falcon 9 first stage will land successfully. Space X advertises Falcon 9 rocket on it website, with a cost of 62 million dollars; other provider cost upward of 165 million dollar each, much of the saving is because Space X can reuse the first stage. Therefore, if we can determine it the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to be against Space X for a rocket launch. .
- **Problems you want to find answers**
- What factor affect successful landing of rockets
- The effect of each relationship of rocket variables to outcome
- Which conditions will enable spaceX in a successful landing

Section 1

Methodology

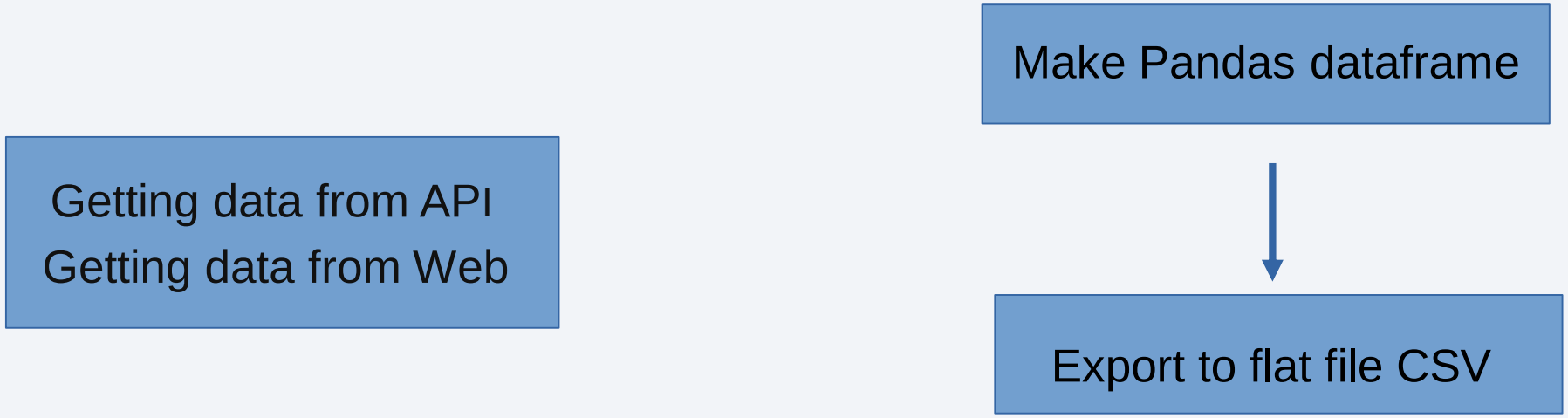
Methodology

Executive Summary

- Data collection methodology
  - Via Spacex Rest API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - By dropping irrelevant columns and perform one hot encoding
- Perform exploratory data analysis (EDA)
  - Using Bar graph and Scatter plot to show relationship between variables
- Perform predictive analysis using classification models
  - Used several classification model, tuned each model using gridsearchcv for best parameters and evaluated model accuracy score and confusion matrix

Data Collection

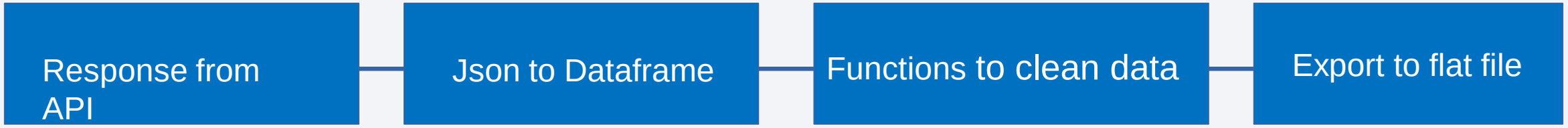
For the purpose of finding the cost of launches, we needed amount of rocket launches data of at least of 5 years. Only falcon 9 Rockets was considered for the purpose of study and related fields like date of launch, location of launch, payload mass, booster version, orbit of deployment, etc. were taken into consideration. Data Collection was done through REST API from SpaceX API website website and web scrapping was carried out from wikipeadia



Data Collection – SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```




```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Notebook

Data Collection - Scraping

```
static url = "https://en.wikipedia.org/w/index.php?title=List of Falcon 9 and Falcon Heavy launches&oldid=1027686922"
```

Getting response from HTML

```
response= requests.get(static_url)
```

Creating Beautifulsoup object

```
soup= BeautifulSoup(response.content, 'html.parser')
```

Extracting all columns

```
html_tables=soup.find_all('table')
first_launch_table.find_all('th')
for row in first_launch_table.find_all('th'):
    name =extract_column_from_header(row)
    if name != None and len(name) > 0:
        column_names.append(name)
```

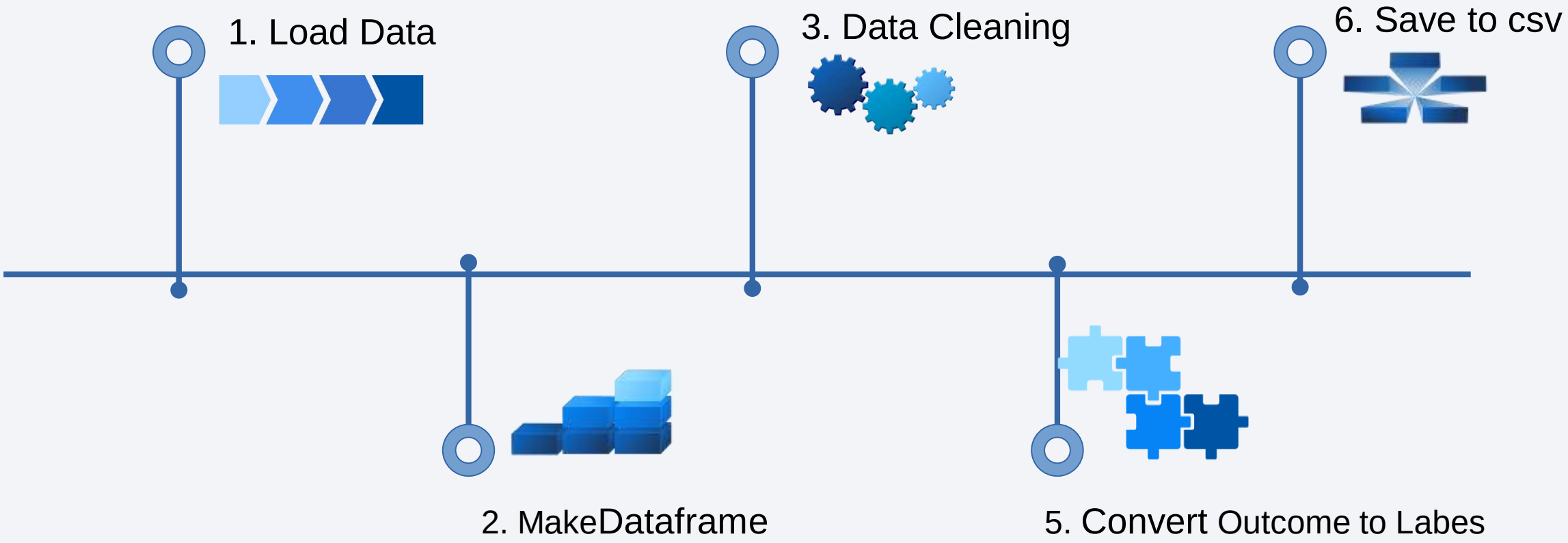
Creating pandas dataframe

```
df=pd.DataFrame(launch_dict)
```


[Notebook]

Data Wrangling

The majoy cleaning here was converting the outcome to labels where 1 and 0 mean sucessfull and unsucssfull landing



Notebook

EDA with Data Visualization

Visualisation was done Seaborn Scatter plot, Bar plot and Line plot;

Scatter plot was used to show the relationship between fight number and launch sight, payload and launch sight, flight number and orbit type, and pay load and orbit type

Bar chart was used to show the comparisons between mean success rate for each orbit. Oribit is a categorical data hence the choice of Bar chart.

Line plot was used to show the the trend in successfull landing from 2013 to 2020

Build an Interactive Map with Folium

Using Folium libarary;

We took the latitude and longitude co-ordinates at each launch site and added a circle marker around each launch site with a label of the name of the launch site inother to visualize the launch data on an interactive map.

We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster

Using Haversine’s formula we also calculated the distance from the launch Site to various landmarks to find various trends about what is around the Launch Site to find patterns.

Predictive Analysis (Classification)

1. Load and standardize data

3. Fit model and tune hyperparameter

5. Choose the best model

2. Split data into train and test data

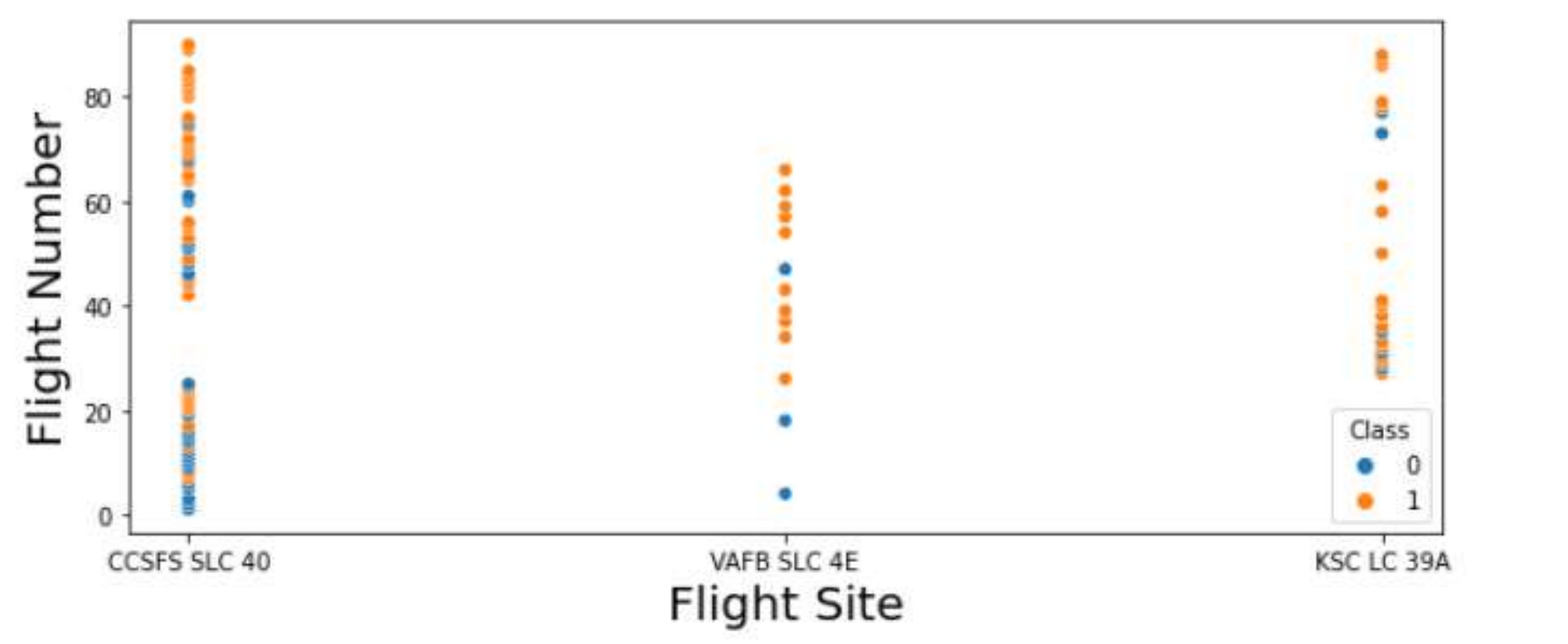
4. Evaluate Model

Notebook

Section 2

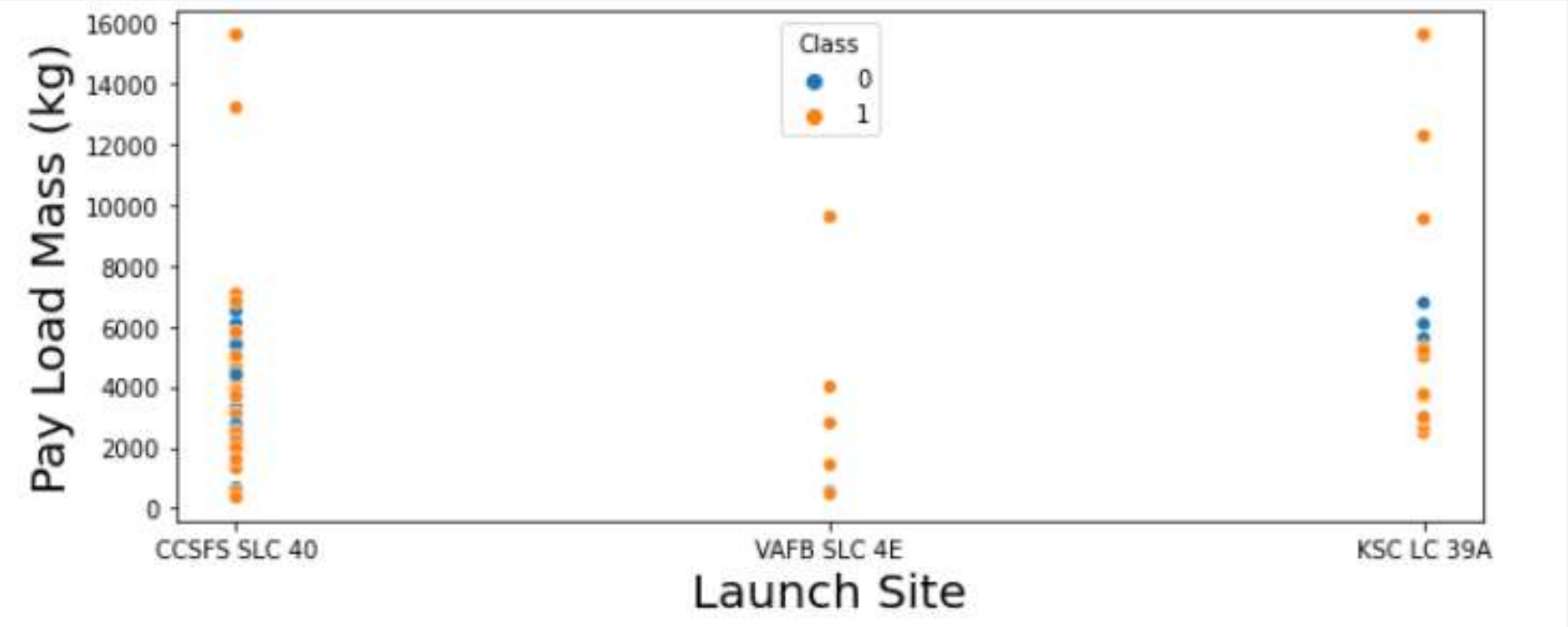
Insights drawn from EDA

Flight Number vs. Launch Site



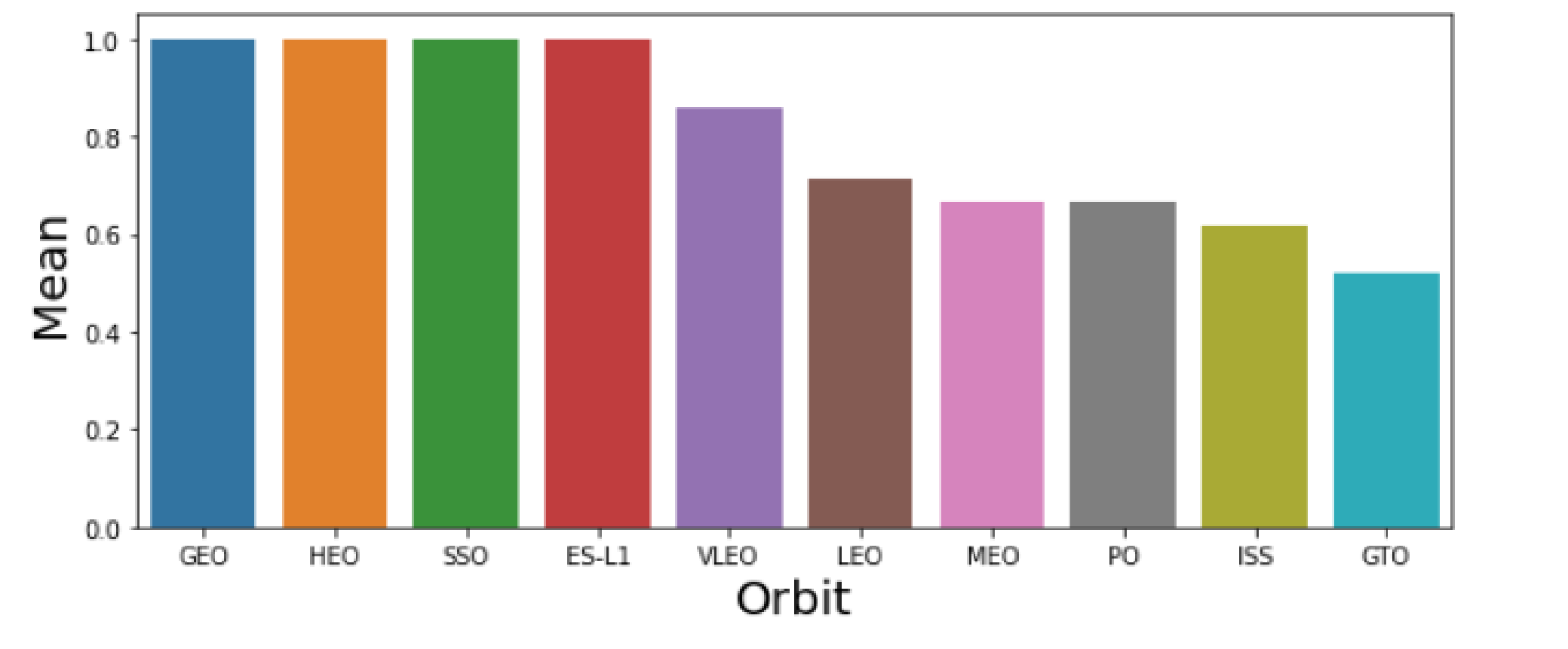
The more amount of flights at a launch site the greater the success rate at a launch site.

Payload vs. Launch Site



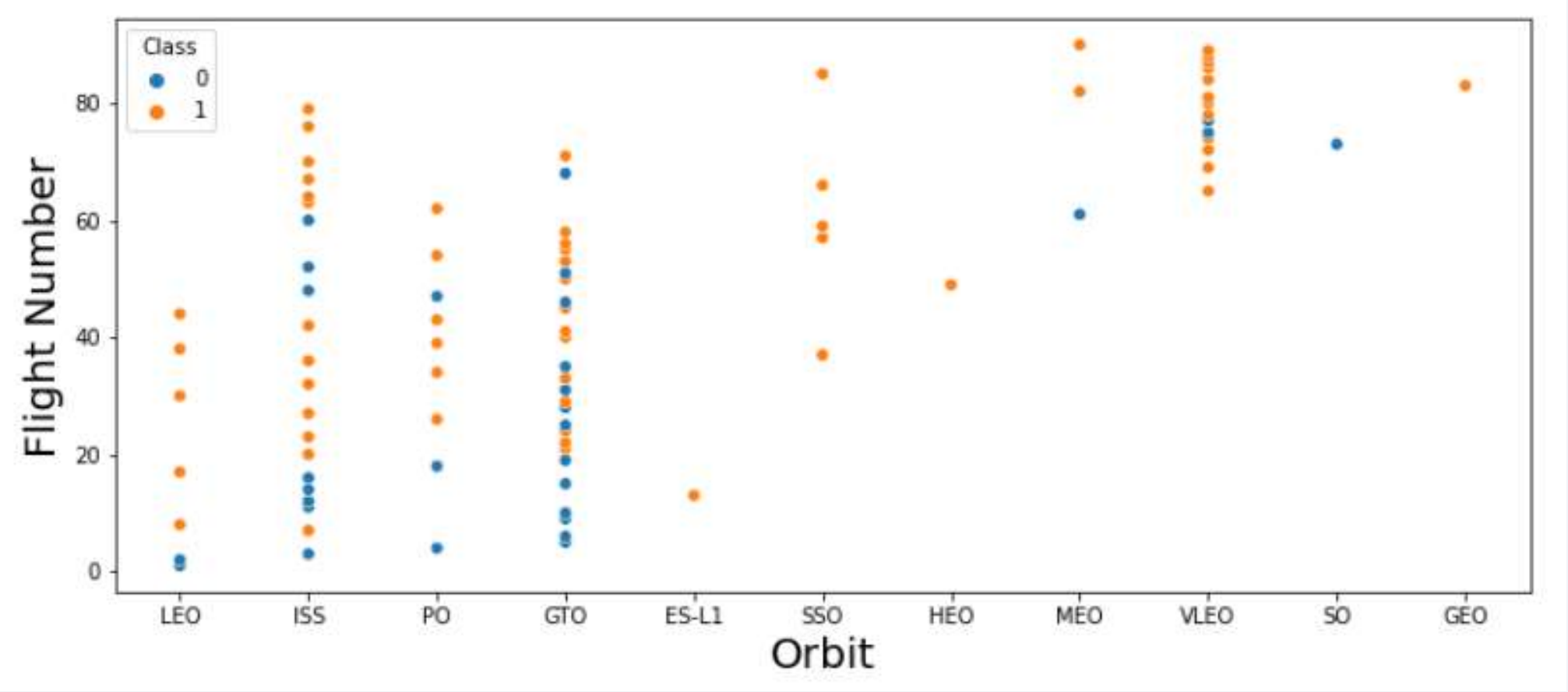
There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

Success Rate vs. Orbit Type



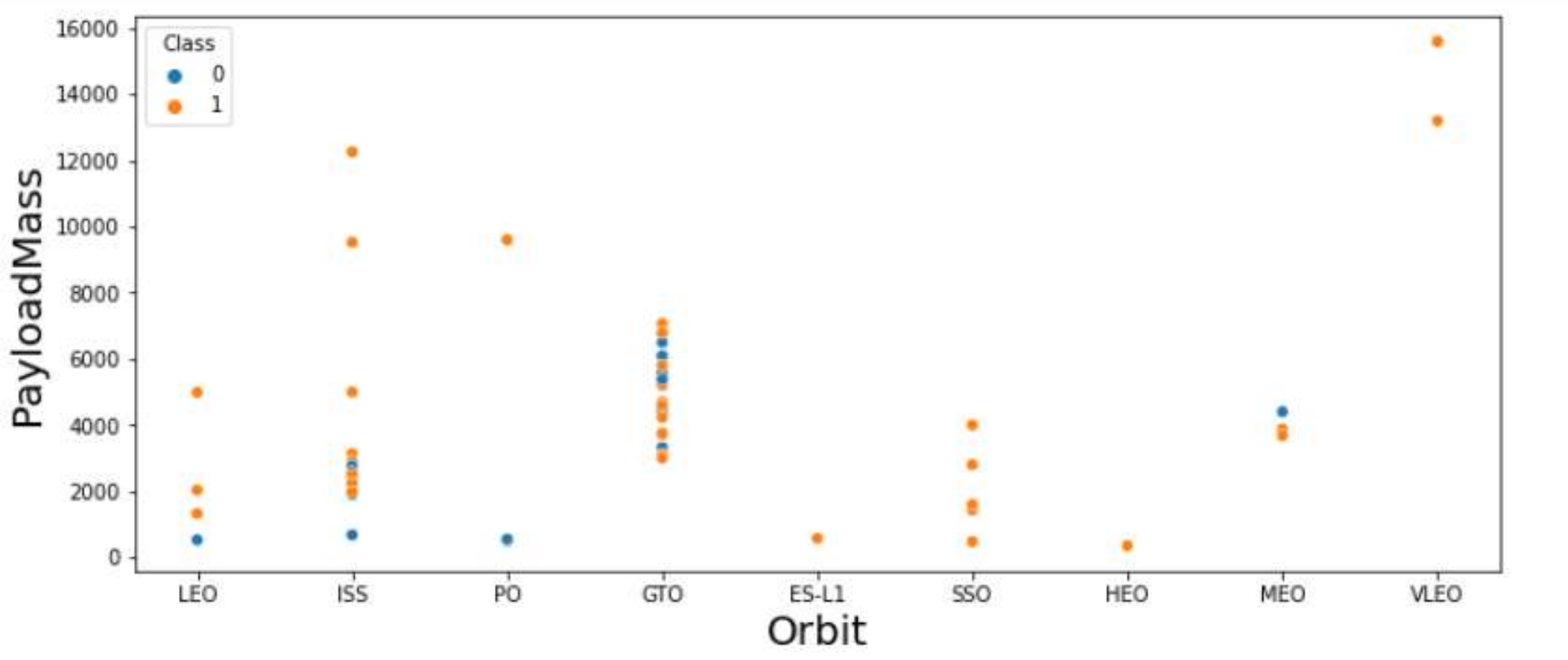
Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Flight Number vs. Orbit Type



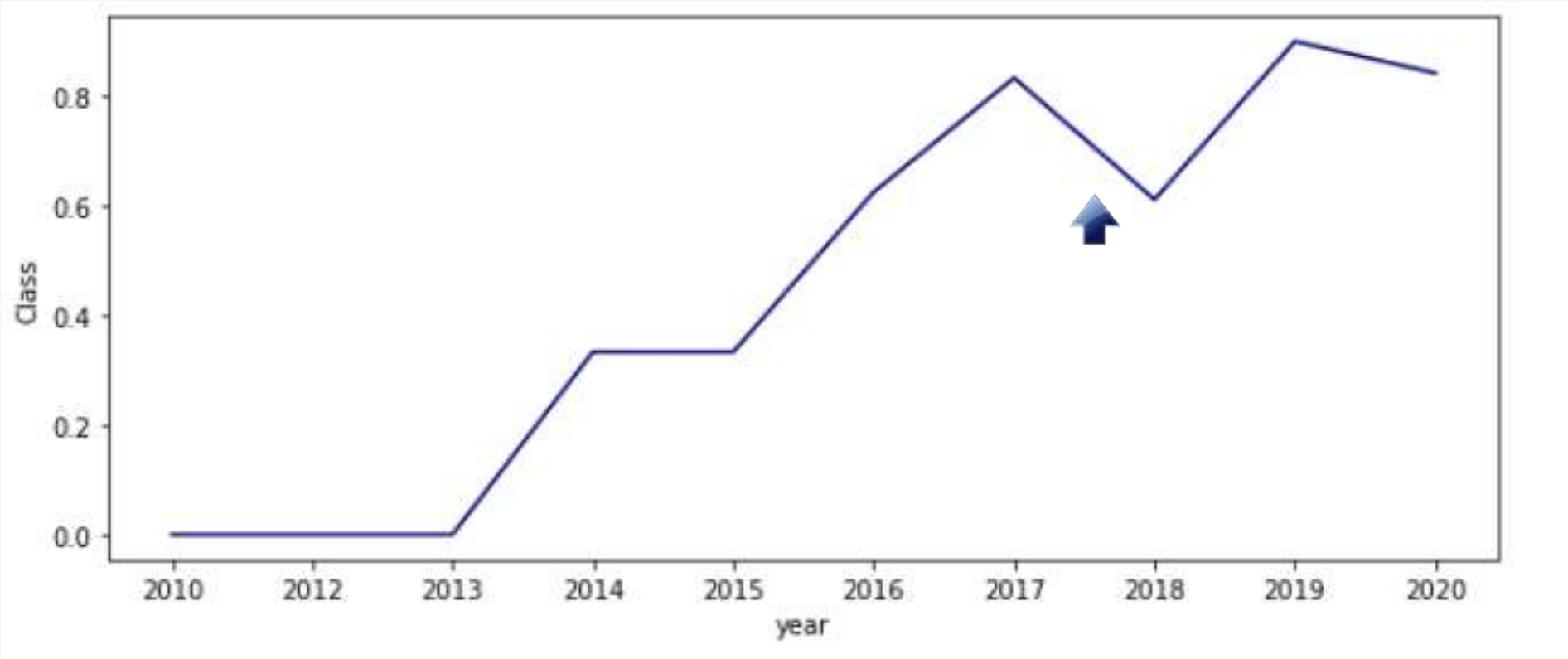
There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



Heavy payloads have a negative influence on GTO orbits and positive influence LEO, ISS orbits.

Launch Success Yearly Trend



From 2013 they has been steady increase in the successfull lauches although they is a deep in 2017 but picked backup in 2018

Section 3

Launch Sites Proximities Analysis

Launch Sites

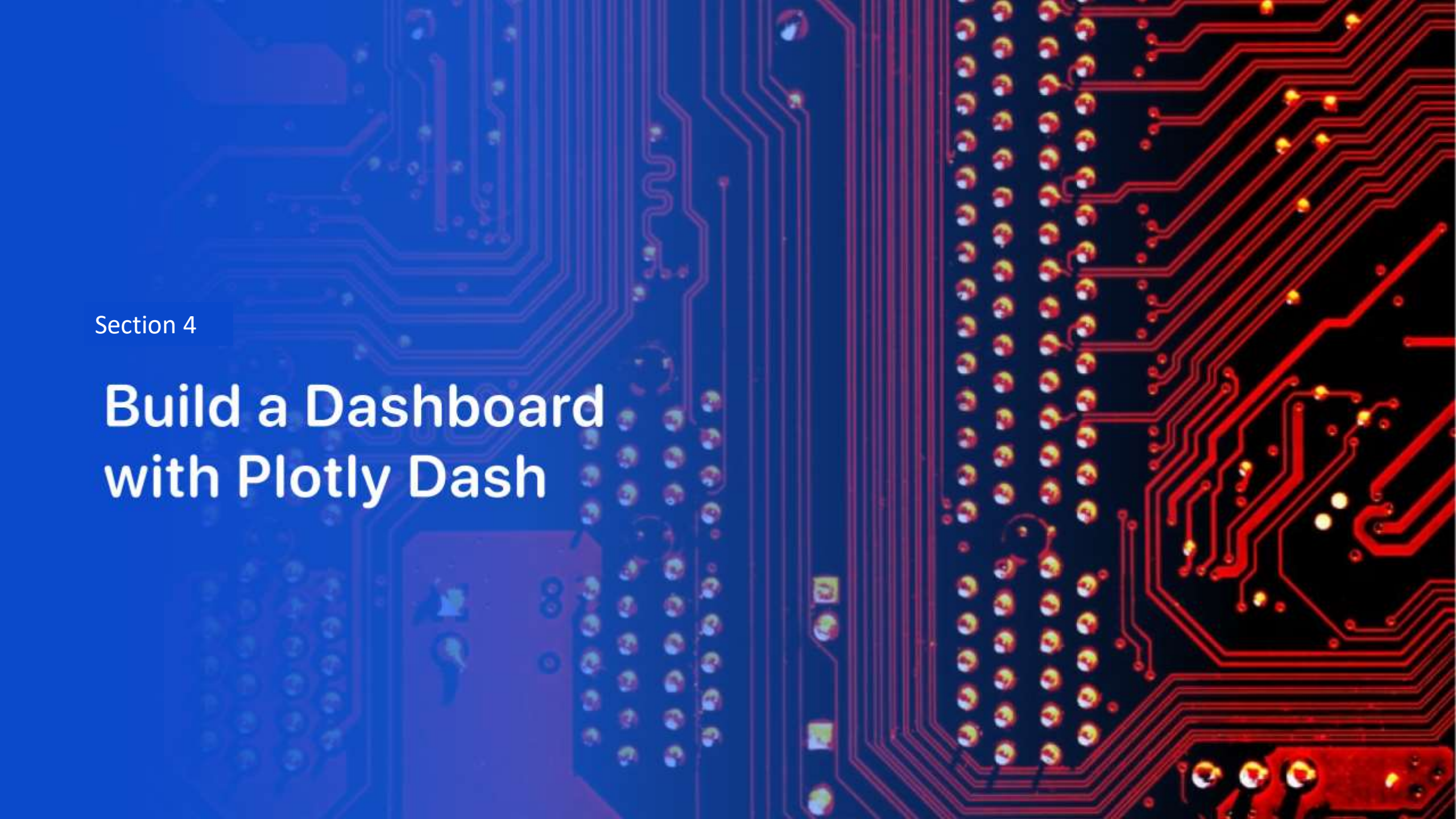
Launch sites are in very close proximity to the coastline

CCAFS SLC-40 from Cities and Highway

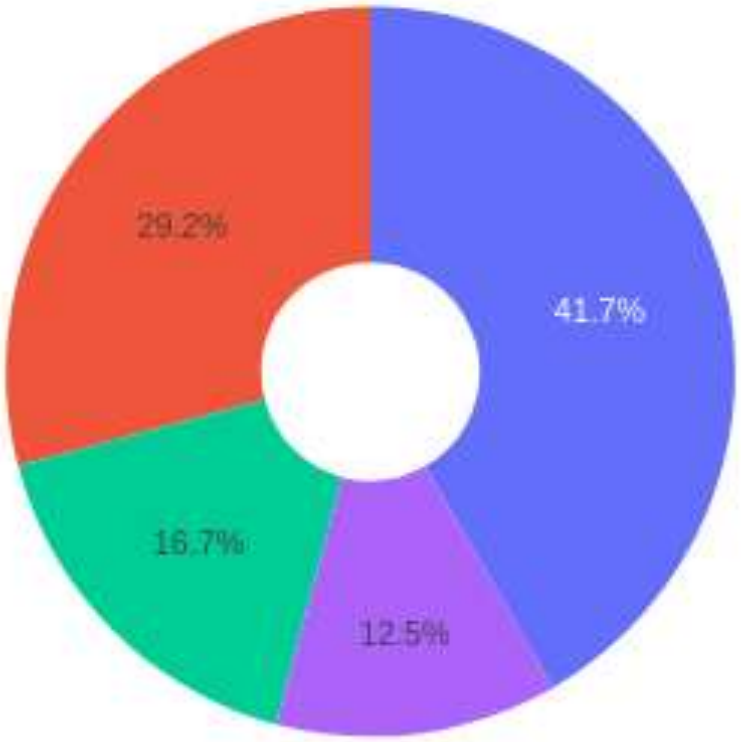
From CCAFS SLC-40 to the coastline is 0.09km while to Orlando city is 78.45km and to the nearest high way is 29.21km

Section 4

Build a Dashboard with Plotly Dash



Launch Success Count by Sites

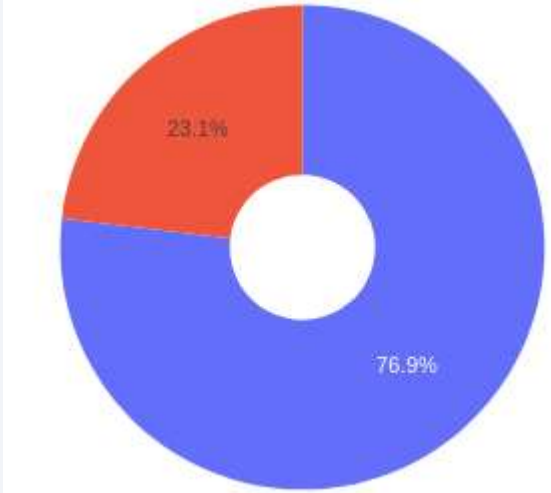


KSLC LC-39A **41.7%**
CCAFS LC-40 **29.2%**
VAFB SLC-4E **16.7%**
CCAFS SLC-40 **12.5%**

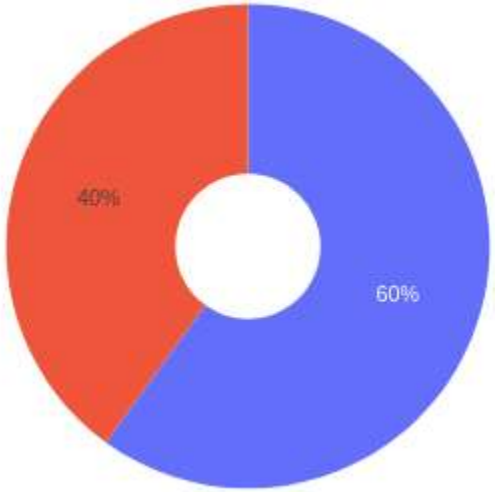
KSLC LC-39A has the largest number of successful launches

Site Outcomes

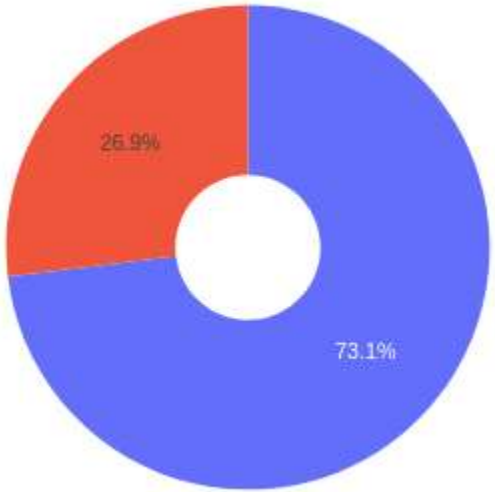
KSLC LC-39A



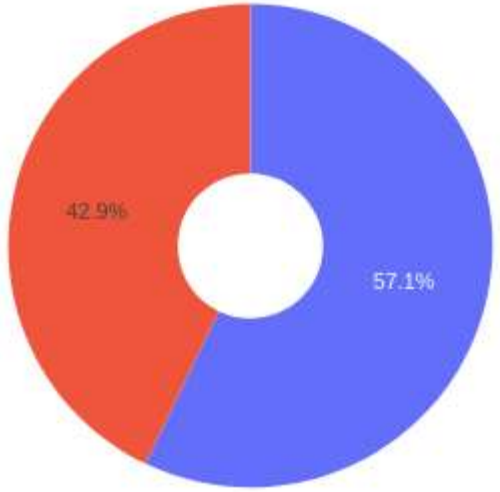
VAFB SLC-4E

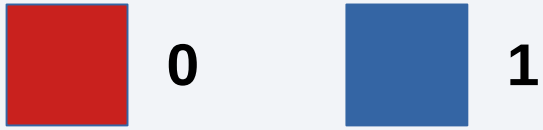


CCAFS LC-40



CCAFS SLC-40



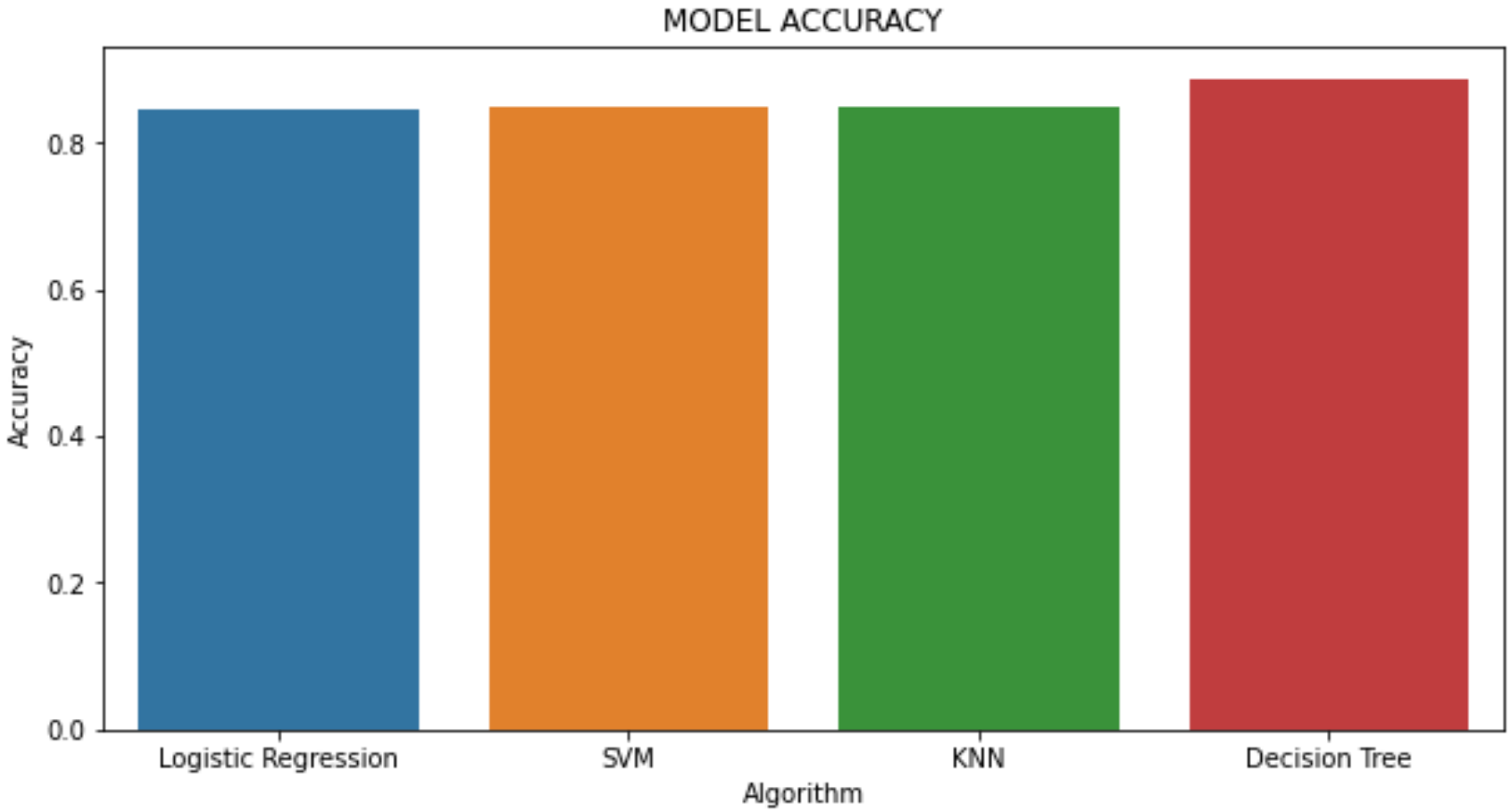


KSC LC 39A has highest success rate among all sites

Section 5

Predictive Analysis (Classification)

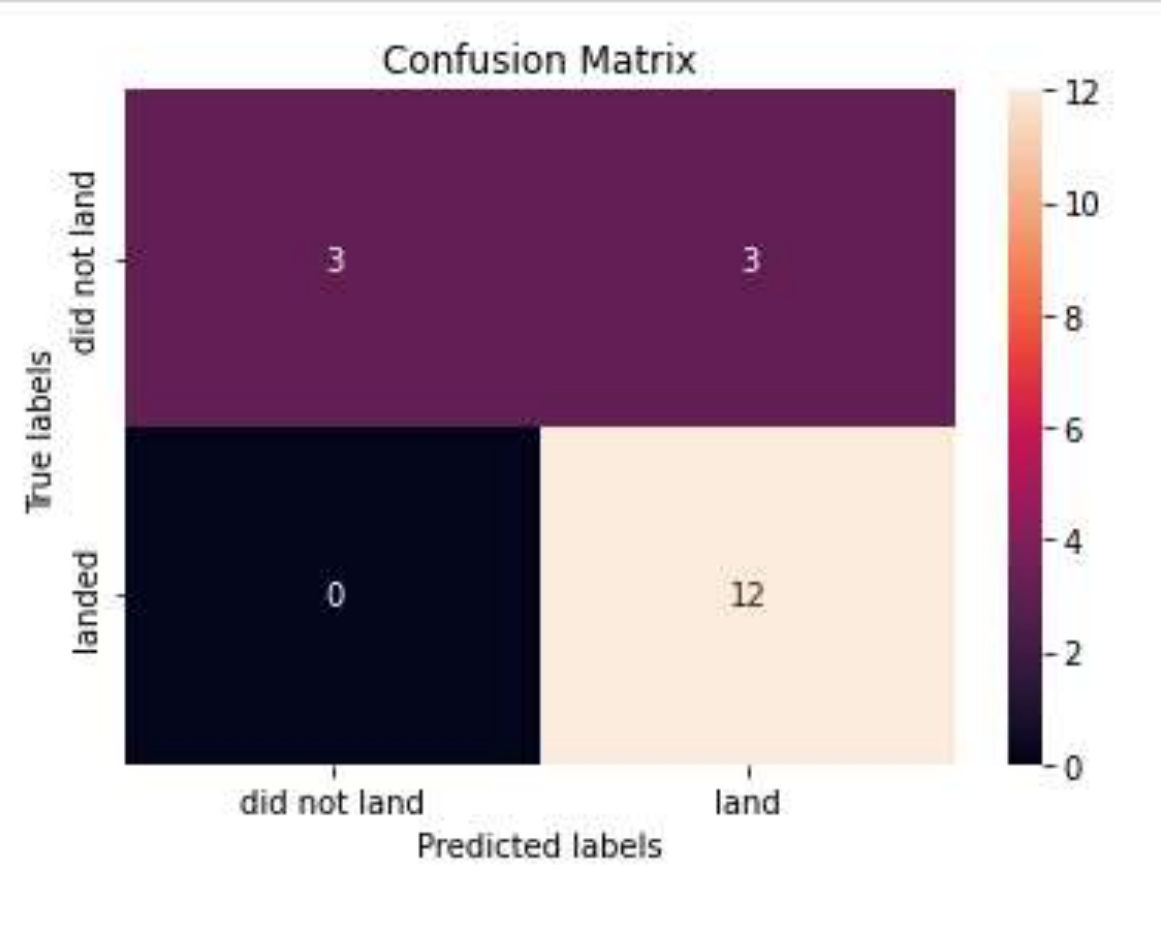
Classification Accuracy



Decision Tree Classifier performed best with an accuracy of 88.87%

criterion: gini
max_depth: 10,
max_features: sqrt
min_samples_leaf: 1
min_samples_split: 2
splitter: random

Confusion Matrix



False Positive 3 means the model predicted 3 unsessfull landing as succss full. True nagetive false nagative and true positive all ok

Conclusions

1. Falcon 9 rockets launched from KSC will be more successful than other launch sites and any payload mass between 2k-4k will have higher chances of successful mission

2. Landing success depends on many features like the physical working attributes of a rocket, orbit of payload

3. The success rate for SpaceX is directly proportional to time in years. With more time, they will eventually perfect the launches

4. The best classifier is Decision tree with an accuracy of 88.87%

…

Appendix

Folium
Plotly

Thank you!