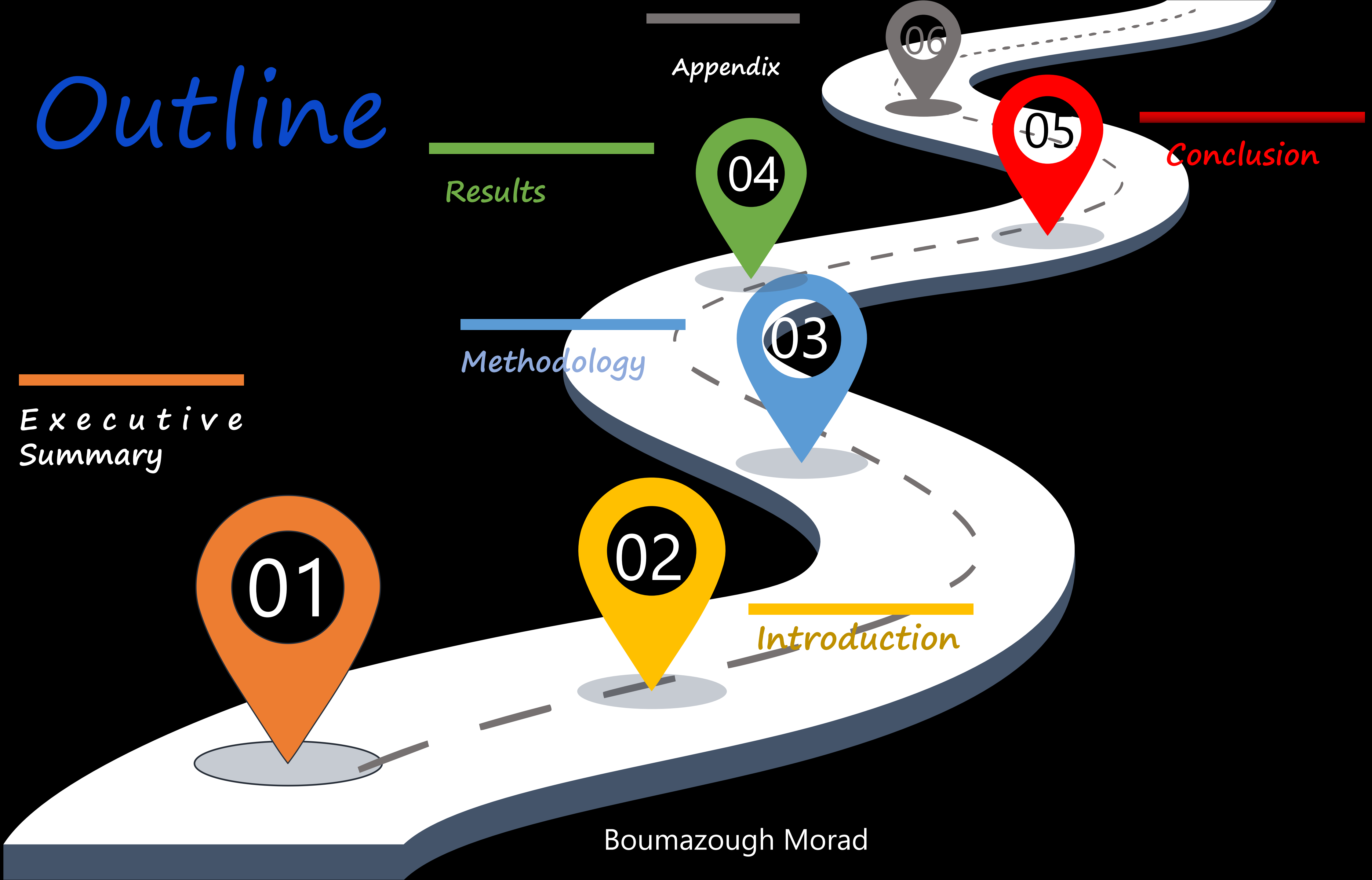
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Boumazough Morad
2024/04/11

Outline



01
Executive Summary
02
Introduction
03
Methodology
04
Results
05
Conclusion
06
Appendix

Executive Summary

- **SpaceY is a new commercial rocket launch provider who wants to bid against SpaceX.**
- **SpaceX advertises launch services starting at $62 million for missions that allow some fuel to be reserved for landing the 1st stage rocket booster, so that it can be reused.**
- **SpaceX public statementsindicate a 1st stage Falcon 9 booster to cost upwards of $15 million to build without including R&D cost recoupment or profit margin.**
- **Given mission parameters such as payload mass and desired orbit, the models produced in this report were able to predict the first stage rocket booster landing successfully with an accuracy level of 83.3%.**
- **As a result, SpaceY will be able to make more informed bids against SpaceX by using 1st stage landing predictions as a proxy for the cost of a launch.**

Introduction

- This report has been prepared as part of the Applied Data Science Capstone course.
- In this capstone, I take the role of a data scientist working for a new rocket company called SpaceY.
- The objective is to evaluate the viability of the new company Space Y to compete with Space X.

<u>BUSINESS PROBLEM :</u>

- ➢ SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars when the first stage of their rockets can be reused.
- ➢ The first stage is estimated to cost upwards of 15 million to build without including R&D cost recoupment or profit margin.
- ➢ Sometimes SpaceX will sacrifice the first stage due to mission parameters such as payload, orbit, and customer.
- ➢ Therefore this report aims to accurately predict the likelihood of the first stage rocket landing successfully as a proxy for the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data from Space X was obtained from 2 sources:

  - Space X API(https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling

  - Collected data was enriched by creating a landing outcome label based on outcome data

  after summarizing and analyzing features

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

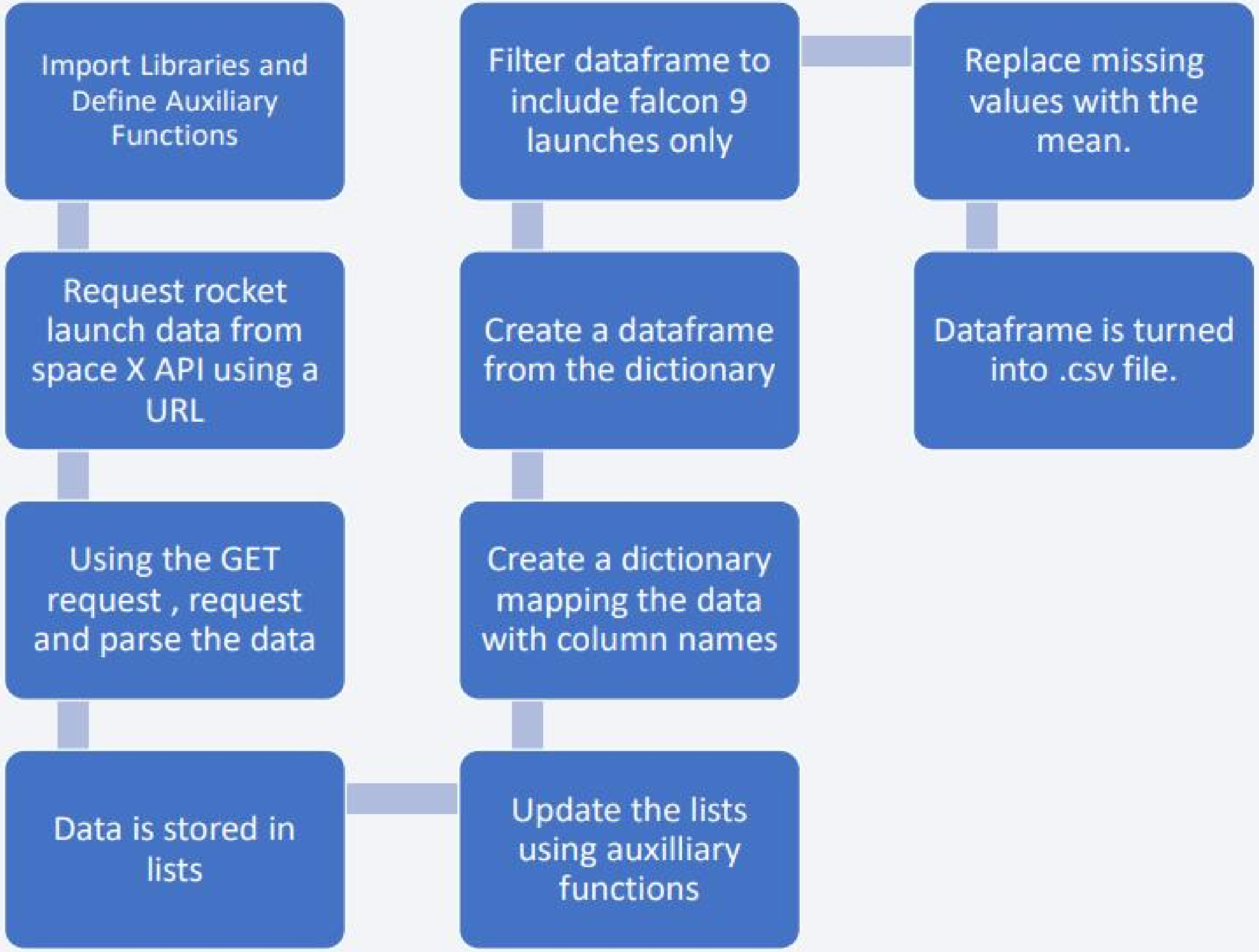
Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being he accuracy of each model evaluated using different combinations of parameters.

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) using web scraping technics.

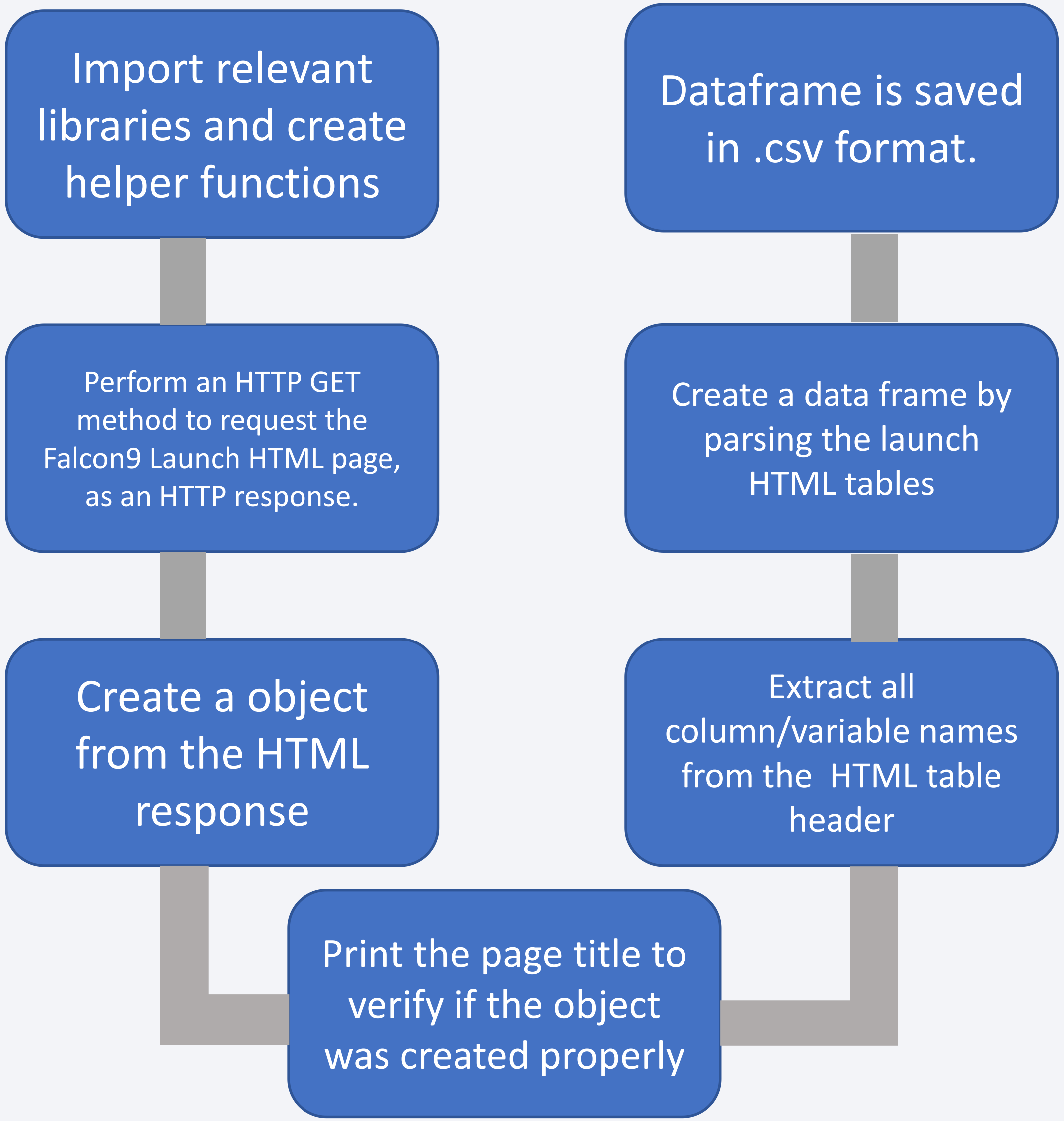
Data Collection – SpaceX API

- Launch data is requested using a URL.
- The data is parsed and stored in a list.
- The data is stored in a dataframe, filtered, wrangled and stored in .csv format.
- https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch-Bidding/blob/main/1.%20Data%20collection%20API.ipynb



Data Collection - Scraping

- The rocket launch data is requested using the HTTP GET method.
- The data is stored in a data frame.
- The data is then saved in a .csv format.
- https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch-Bidding/blob/main/2.%20Data-collection-webscraping%20.ipynb

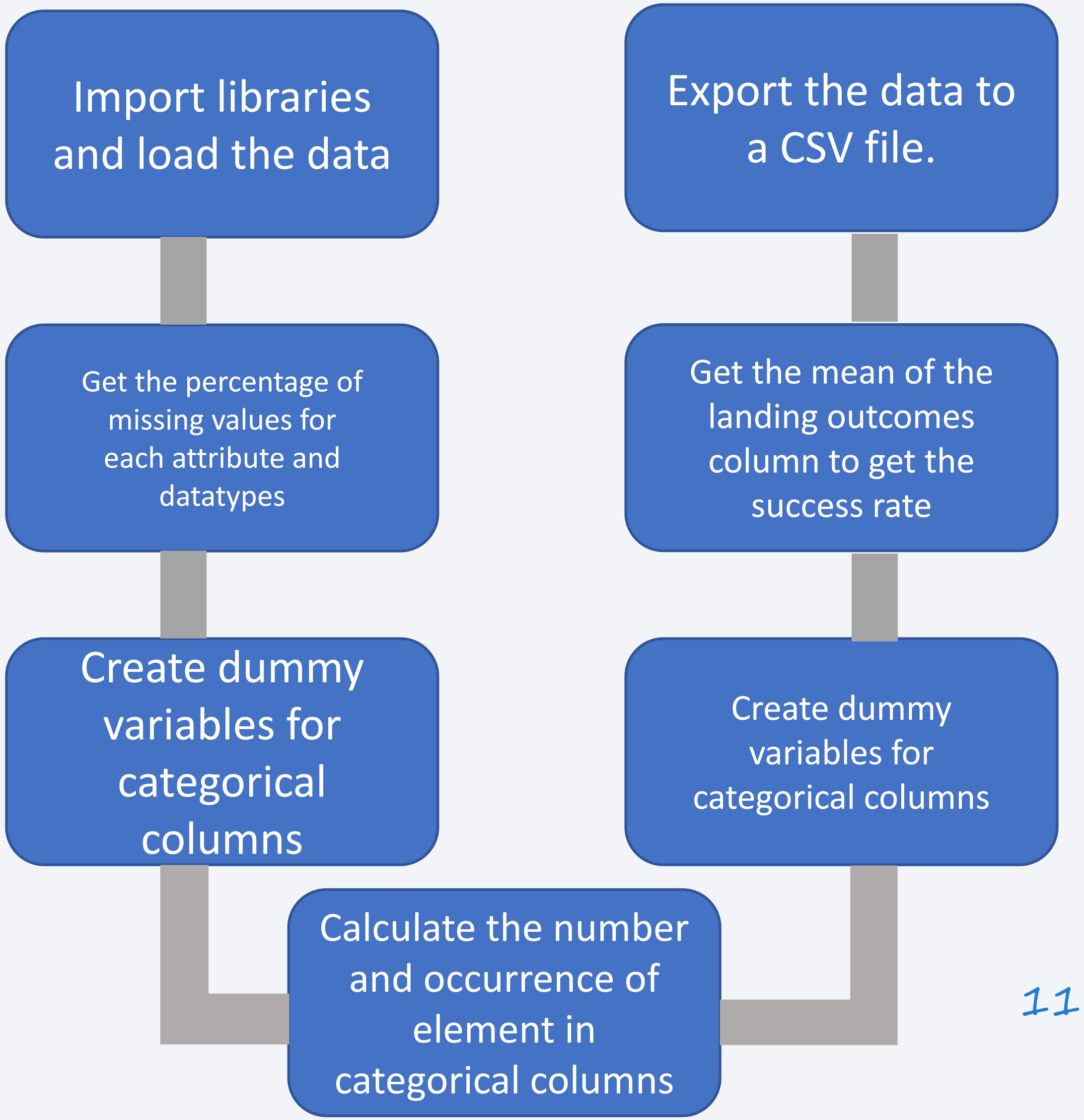


Data Wrangling

- The missing values were found and replaced by mean values.
- The categorical variables where presented and replaced with dummy variables.
- Determination of success rate of

landing outcome.

- https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch-Bidding/blob/main/3.%20Data_wrangling.jupyterlite%20(1).ipynb



EDA with Data Visualization

• Numerous scatterplots were created to aid data visualization.

• The scatterplots were created in order to understand the relationships between variables and determine the extent of their effect on the outcome of the launch.

• The plot of the flight number vs payload mass shows that the first stage is more likely to succeed with a higher flight number and less likely to succeed with higher payload mass.

• A bar chart was created to visualize the success rate of each orbit.

• Based on the information gotten from the data visualization, a new table was created with the variables that can help predict the outcome of the first stage.

• https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch Bidding/blob/main/5.%20EDA%20%26%20dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The dataset was downloaded and loaded into the database table.
- A connection to the database was established.
- Exploratory analysis was carried out on the data to gain key insights.

Build an Interactive Map with Folium

- **Launch success may also depend on the location and proximity of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors, we discovered some of the factors by analyzing the existing launch site locations.**

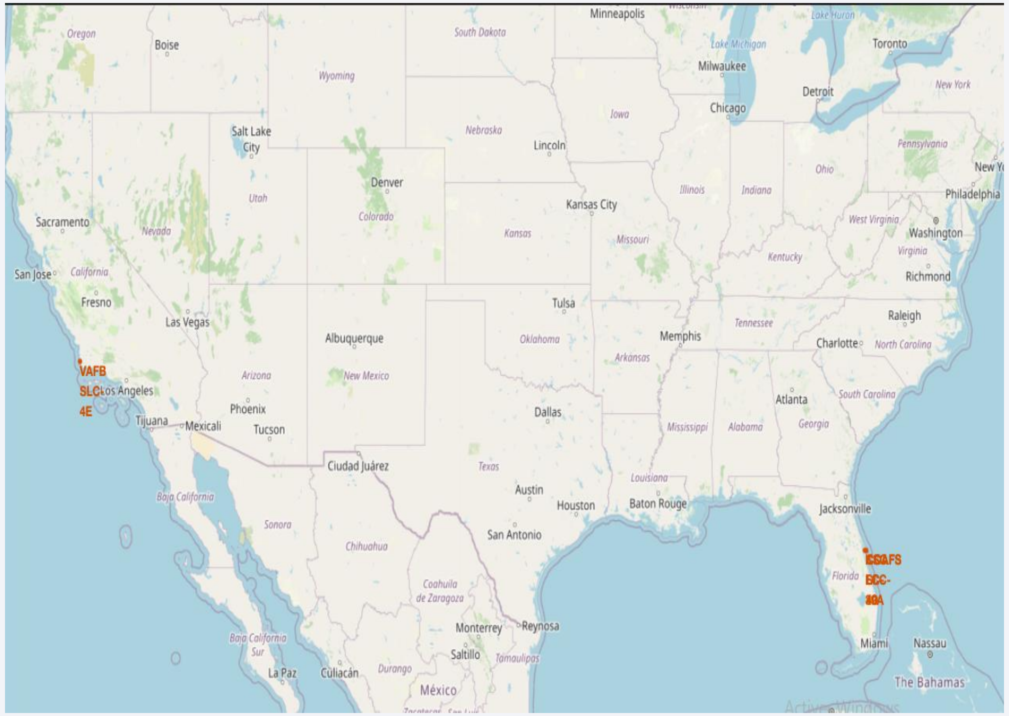
- **Markers were created to visualize each of the launch sites on the map. The circle was created to highlight the site locations and aid visualization.**

- **A marker color object for each launch outcome with unique color identifiers was also added to enhance the map and see which sites have high success rates.**

- **A mouse pointer was added to enhance the ease of getting the longitude and latitudes of the launch sites.**

- **A line was added to visualize the distance between launch sites and significant structures like railways.**

- **https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch-Bidding/blob/main/Interactive%20visual%20analytics-folium.ipynb**

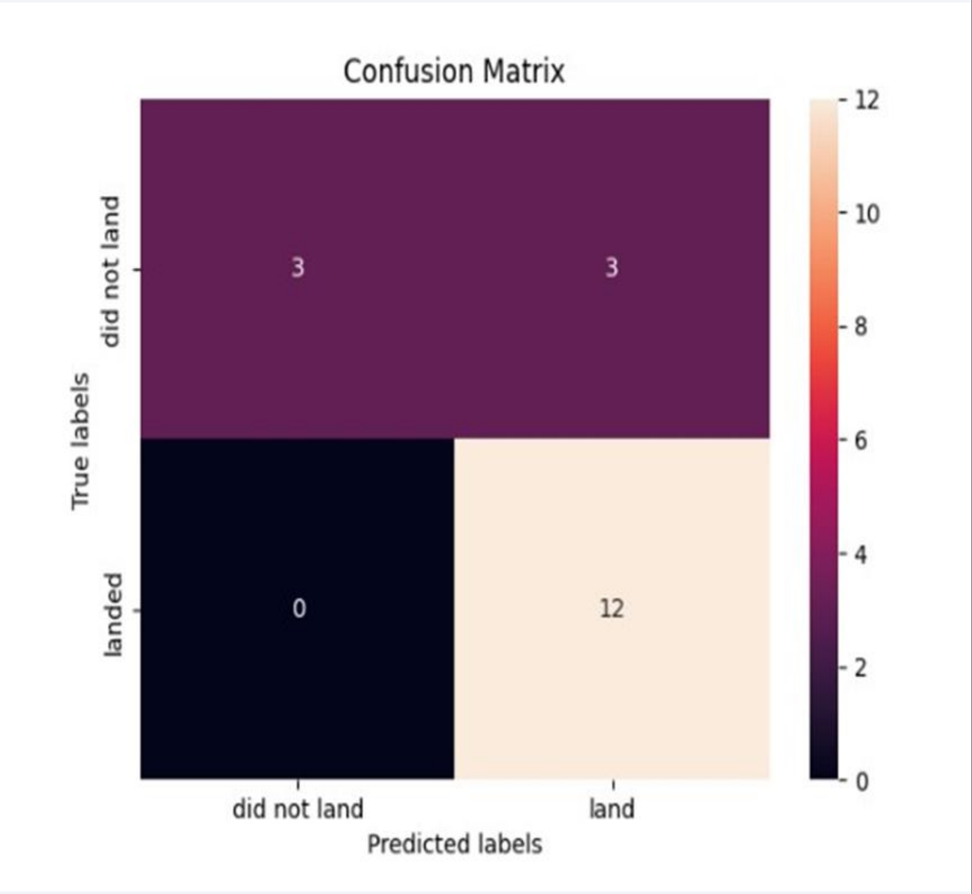


Build a Dashboard with Plotly Dash

- A dropdown list was added to enable Launch Site selection.
- A pie chart was added to show the total successful launches count for all sites.
- Added a scatter chart to show the correlation between payload masses and launch success.
- https://github.com/moradBoumazough/IBM-Predicting-Falcon-9-First-Stage-Landing-Success-for-Cost-Efficient-Rocket-Launch-Bidding/blob/main/Interactive%20visual%20analytics-plotly.ipynb

Predictive Analysis (Classification)

- **Imported libraries and defined function to create confusion matrix (Pandas, Numpy, Matplotlib,Seaborn,Sklearn)**
- **Loaded the dataframe created during data collection**
- **Created a column for our training label 'Class' created during data wrangling**
- **Standardized the data**
- **Split the data into training data and test data**
- **Fit the training data to various model types (Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors Classifier)**
- **Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model**
  - **Enabled by Scikit-learn library function GridSearchCV**
- **Evaluated accuracy of each model using test data to select the best model**



Results

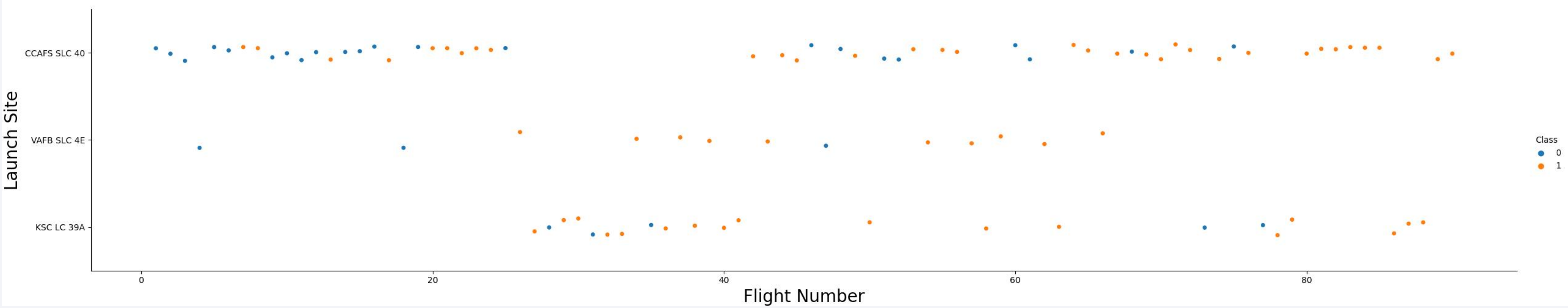
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

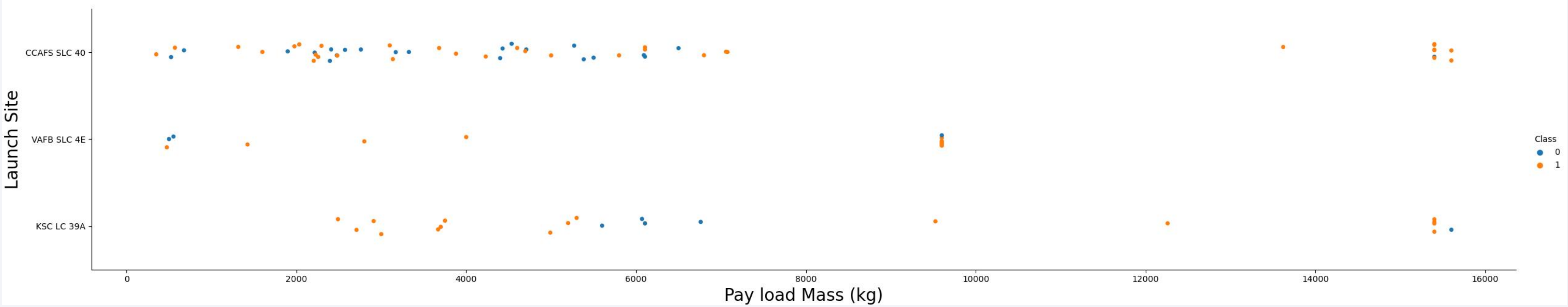
Flight Number vs. Launch Site

- This scatter plot shows that as the flight number increases the first stage is more likely to be successful.
- KSC LC 39A had more successful launches than unsuccessful.
- CCAFS SLC 40 had more unsuccessful launches at the initial flight attempts, the launches became more successful with increasing flight number.
- VAFB SLC 46 had more successful launches than unsuccessful.



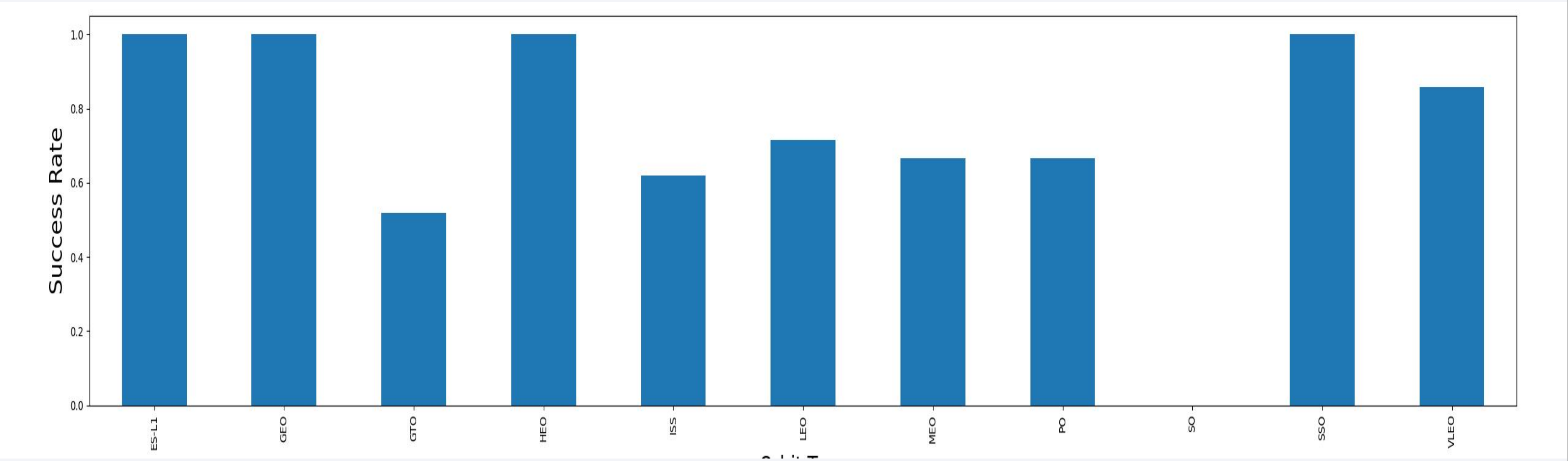
Payload vs. Launch Site

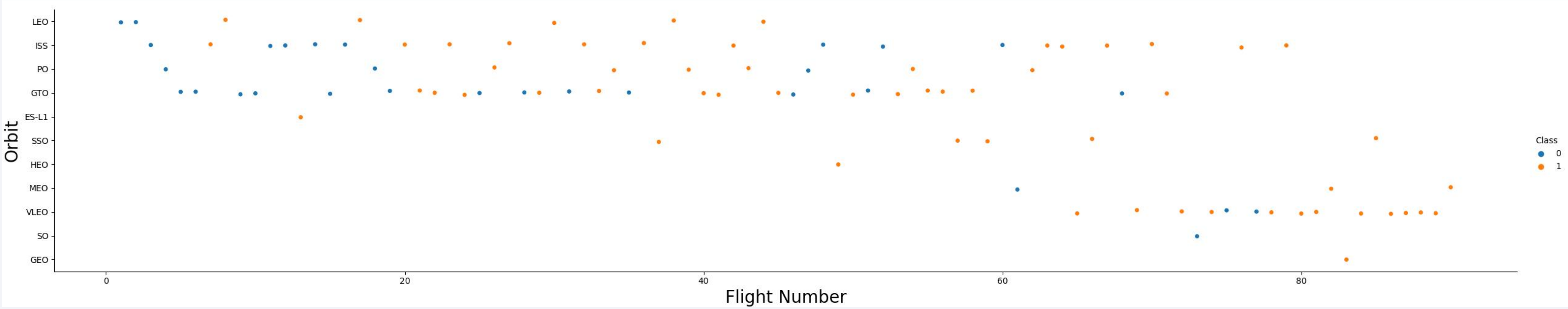
- **VAFB SLC 46 had fewer launches than the other sites but had more successful launches.**
- **KSC LC 39A had more launches with a wide range of payload mass but had more successful launches.**
- **CCAFS SLC 40 had most of the launches with lower pay load mass but has a nearly fair number of successful and unsuccessful launches.**
- **More launches generally took place with lighter payload mass.**



Success Rate vs. Orbit Type

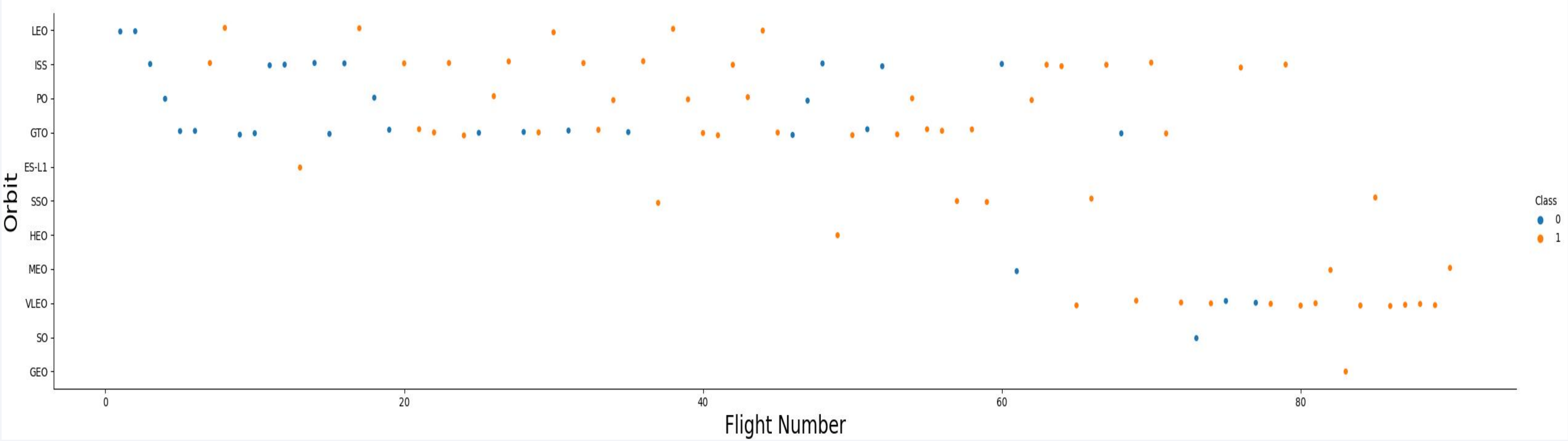
- The success rate of LEO, PO are moderate whereas VLEO success rate is a little higher.
- ES –L1, HEO, SSO and GEO had very high success rates.





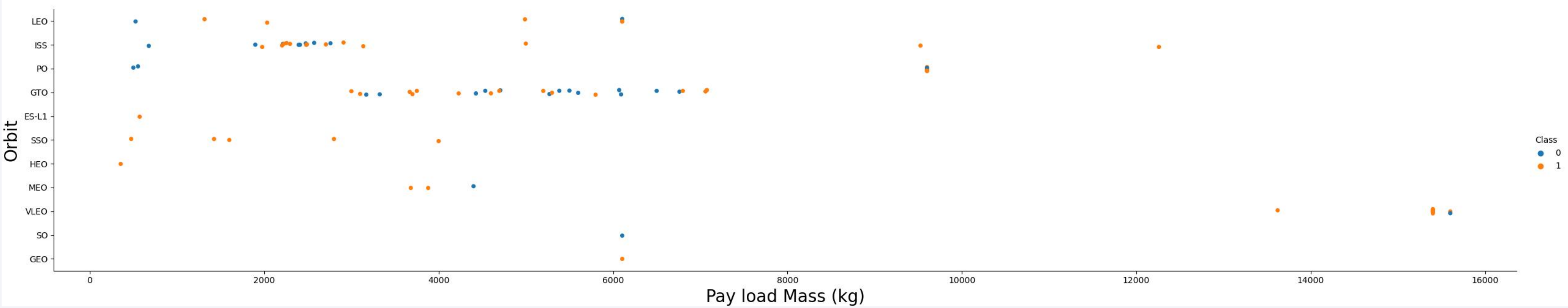
Flight Number vs. Orbit Type

- The success rate of LEO and PO are mostly related to their flight numbers whereas GTO success rate has no correlation with it's flight number.
- ES –L1, HEO, SSO and GEO had very few launch attempts with later flights which explains their success rate.

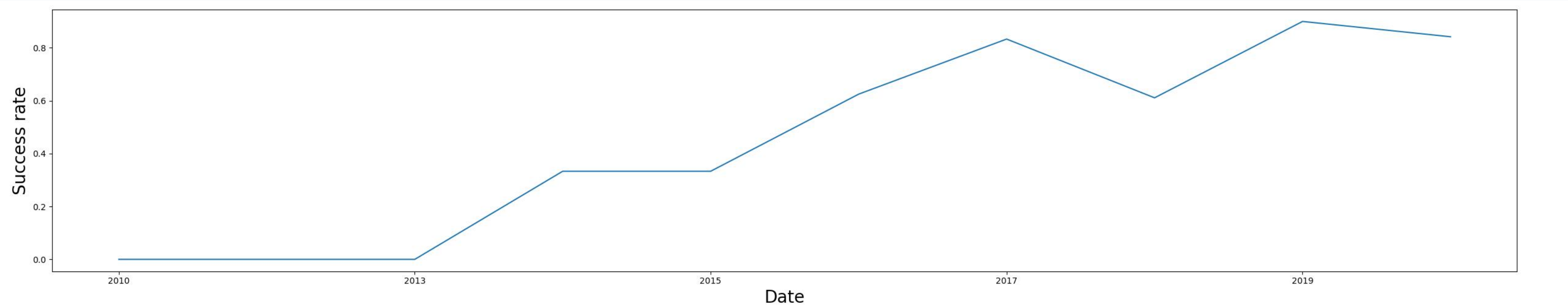


Payload vs. Orbit Type

- With heavy payloads, the rate of successful landings is higher 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.
- ES L1, SSO and HEO have launches with only light payload masses which were all successful.



Launch Success Yearly Trend



- The success rate was constant between 2010 and 2013.
- From 2013, the success rate had a continual increase till 2015, where there’s a slight decrease, there’s a continued increase after till 2017 ,where there’s a
- decrease then another increase from 2018 till 2019 then a decrease past 2019.

All Launch Site Names

- There are four unique launch sites

```
[12]: %%sql
select distinct LAUNCH_SITE from "SPACEXTBL"

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

[12]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

19]:
```
%%sql
select launch_site
from "SPACEXTBL"
where launch_site like "CCA%"
limit 5
```

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

19]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Task 3

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

[22]:
```
%%sql
select sum(payload_mass__kg_)
from SPACEXTBL
where customer="NASA (CRS)"
```

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

[22]:

sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

Task 4

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

[23]:
```
%%sql
select avg(payload_mass__kg_)
from SPACEXTBL
where booster_version="F9 v1.1"
```

 * sqlite:///my_data1.db
Done.

[23]:

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

[61]:
```
%%sql
select min(DATE)
from SPACEXTBL
where "Landing _Outcome"  = 'Success (ground pad)'
```

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

[61]:

min(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

[63]:
```
%%sql
select booster_version,payload_mass__kg_
from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)" and payload_mass__kg_ between 4000 and 6000
```

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

[63]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[72]:
```
%%sql
select count(mission_outcome)
from SPACEXTBL
where mission_outcome like "Success%"
```

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

[72]:

count(mission_outcome)
100

Boosters Carried Maximum Payload

```
[75]: %%sql
select booster_version,payload_mass__kg_
from SPACEXTBL
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

[75]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[78]:
```
%%sql
select substr(Date, 4, 2),booster_version,"Landing _Outcome",launch_site from SPACEXTBL
where substr(Date,7,4)='2015'
and "Landing _Outcome" = "Failure (drone ship)"
```

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

[78]:

substr(Date, 4, 2)	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[100]:
```
%%sql  SELECT Date, "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT
from SPACEXTBL
where   substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'and '20170320'
and  "Landing _Outcome" like "%Success%"
group by "Landing _Outcome"
order by count("Landing _Outcome")
desc
```

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

[100]:

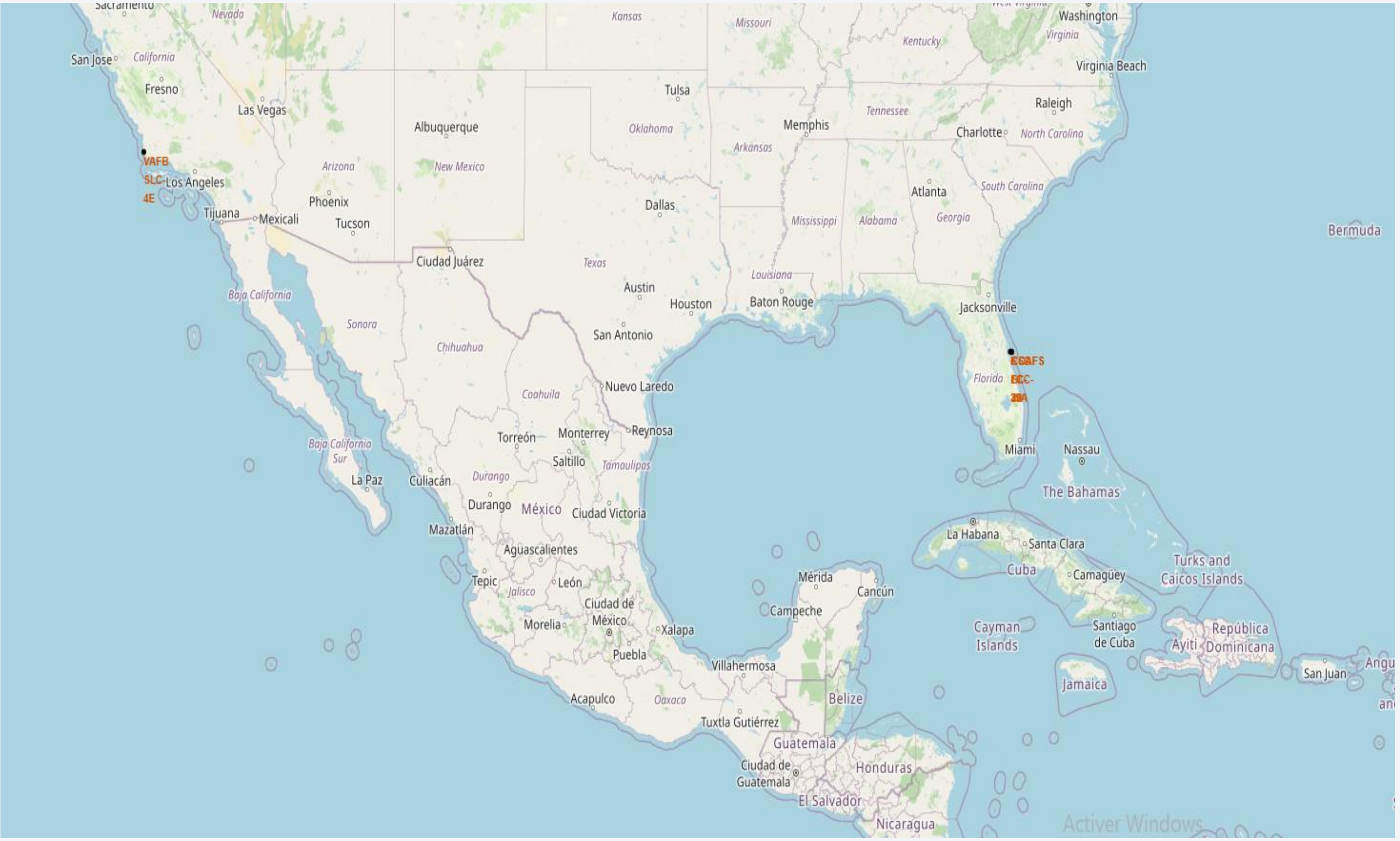
Date	Landing _Outcome	LANDING_OUTCOME_COUNT
08-04-2016	Success (drone ship)	5
22-12-2015	Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

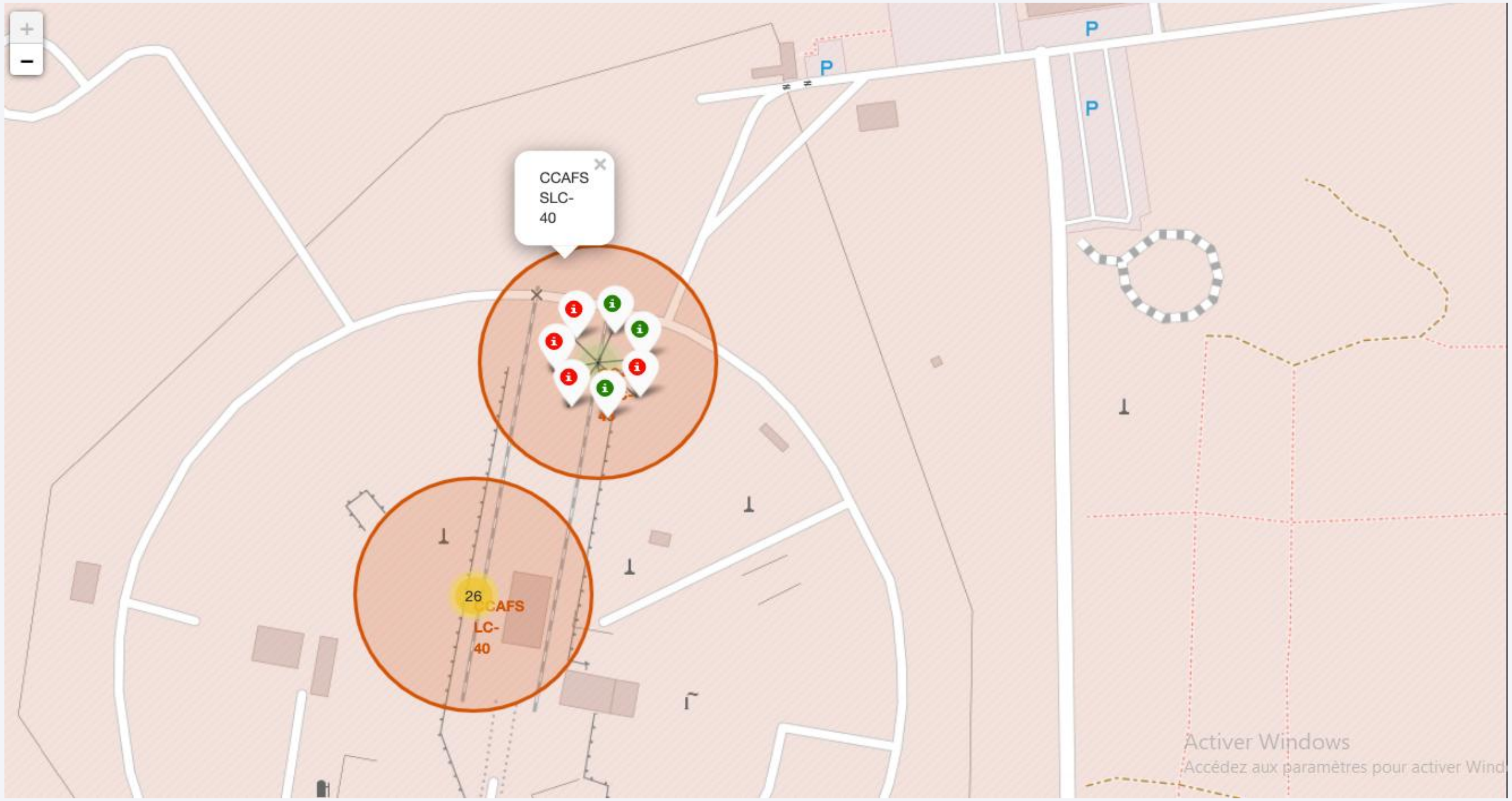
LAUNCH SITE LOCATION ANALYSIS

- This map shows the locations of all the launch sites on a global map.
- The launch sites are all at close proximity to the coastline which is safe area because it is far away from cities, highroads and other areas where it could be harmful to humans.



Folium map displaying the launch outcomes of the launch sites

- The map shows the color-labeled launch outcomes for each site, where green represents successful launch outcomes and red represents the unsuccessful outcomes.



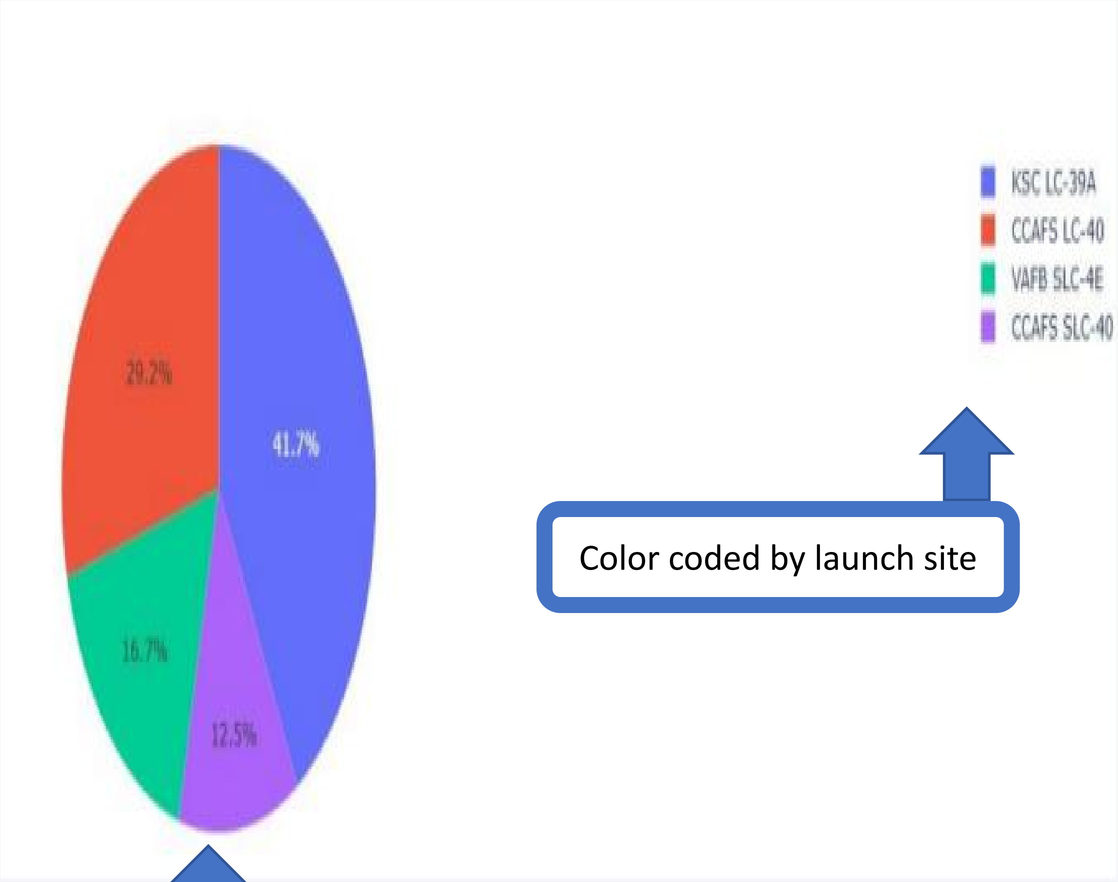
Folium map displaying the distance of VFAB SLC-4E from the coastline

- This shows the distance of VFAB SLC-4E from the coastline.

Section 4

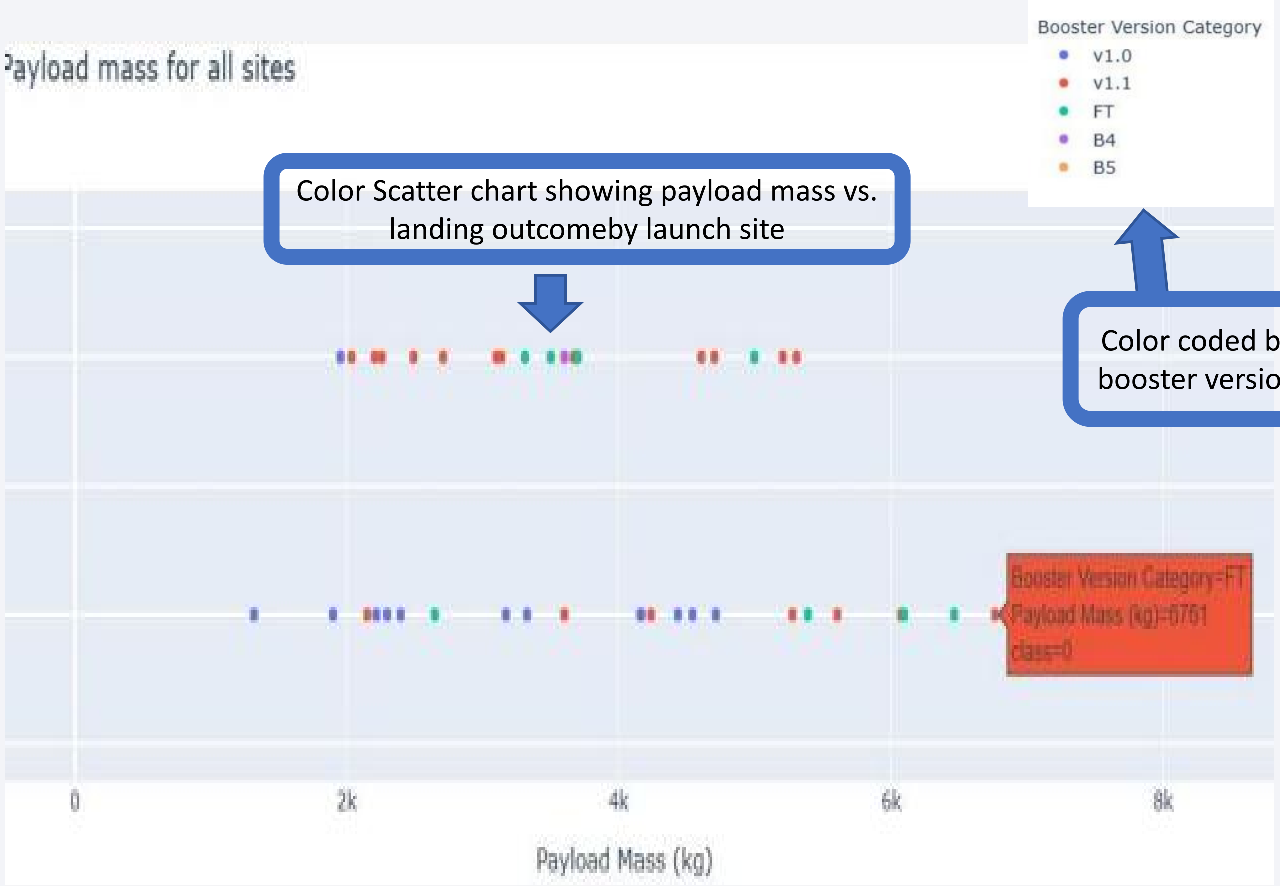
Build a Dashboard with Plotly Dash

LAUNCH RECORDS DASHBOARD



Color coded by launch site

Pie chart showing booster landing success rate

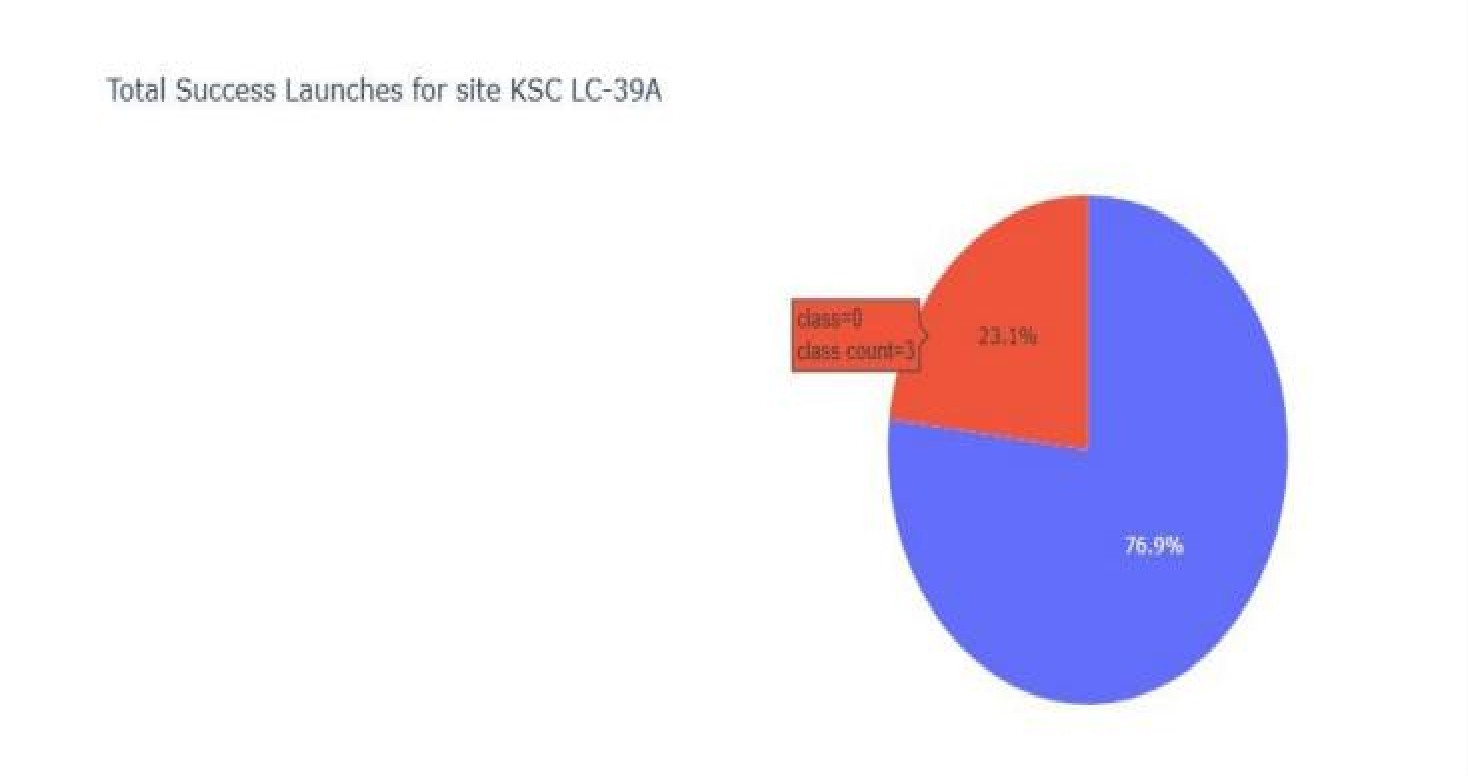


Color Scatter chart showing payload mass vs. landing outcomeby launch site

Color coded by booster version

Pie chart of the KSC LC -39A

- The KSC LC -39A launch site has the most successful launches overall with 76.9% of the launches being successful.

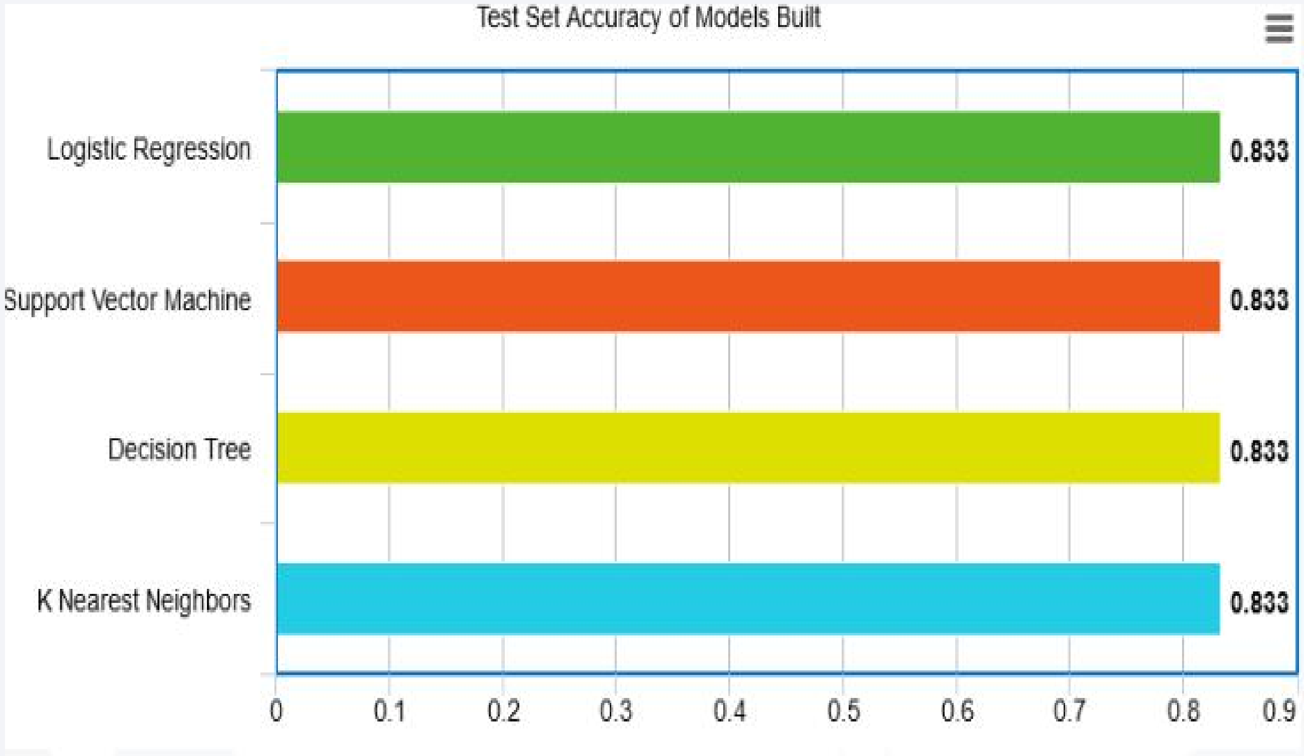


Section 5

Predictive Analysis (Classification)

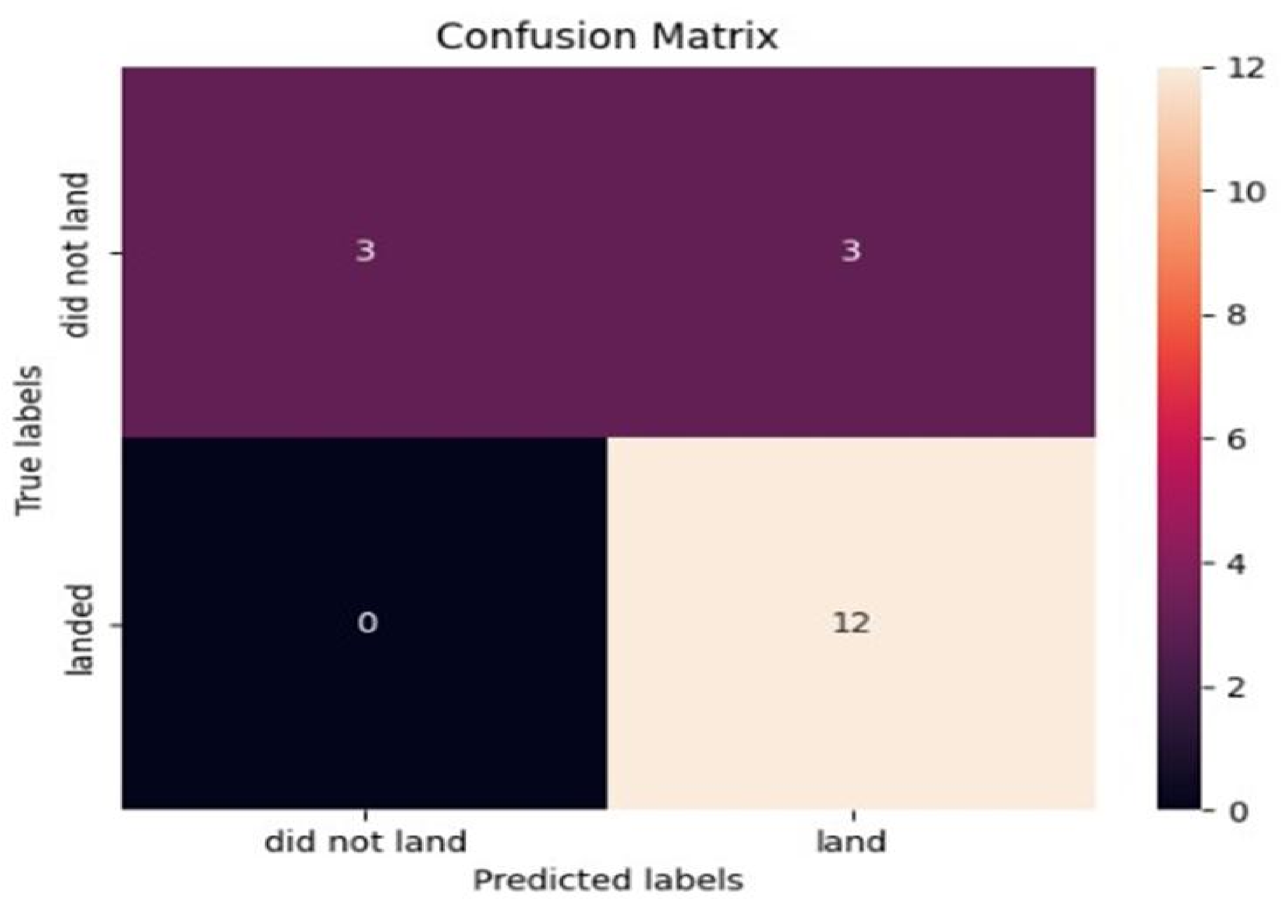
Classification Accuracy

- Each of the four models built came back with the same accuracy score, 83.33%



Confusion Matrix

- The confusion matrices of the best performing models (4-way-tie) are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set



Conclusions

- **This dataset has unveiled intriguing findings, showcasing how the choice of booster version carrying the payload impacts the landing outcome of the launch.**
- **With increasing flight numbers, there is a corresponding increase in the success rate.**
- **Booster v1.1 demonstrates a significant number of successful outcomes across a broad range of payloads.**
- **KSC LC-39A stands out as the launch site with the highest number of successful launches, potentially influenced by factors such as orbit, boosters, and payload mass.**
- **The decision tree model emerges as the optimal choice for machine learning due to its impressive accuracy level of 0.94.**
- **Combining a higher flight number, booster v1.1, and launch site KSC LC-39A appears to be a winning formula for successful first-stage Falcon 9 launches. Additionally, the orbit and payload mass exhibit varied effects on launch success, dependent on other influencing factors.**

Appendix

Thank you to IBM for creating the course and materials

Thank you!