Winning Space Race with Data Science

Roongpraiwan Supat
6/12/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Collect Space X data by using SpaceX API and Web Scraping
  - Perform data wrangling using Pandas and EDA using visualization and SQL
  - Create interactive visual analytics using Folium and Plotly Dash
  - Predict analysis using classification models in Scikit-learn library
- Summary of all results
  - The best classification model is Decision Tree Classifier model with accuracy of 94.44%
  - From the confusion matrix, Decision Tree Classifier can distinguish between the different classes, but have false positives as a major problem

Introduction

- Project background
  - In this capstone, I am a data scientist working for a new rocket company named “Space Y” that would like to compete with SpaceX.
  - My job is to determine the price of each launch by gathering information about Space X and creating dashboards.
  - I also determine if SpaceX will reuse the first stage by training a machine learning model and use public information to predict if SpaceX will reuse the first stage.
- Goal
  - Determine if Space Y should reuse the first stage rocket based on machine learning model, trained using Space X data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting Space X data by using SpaceX API and Web Scraping
- Perform data wrangling
  - Analyzing and Cleaning data using Pandas library
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using Scikit-learn library

Data Collection

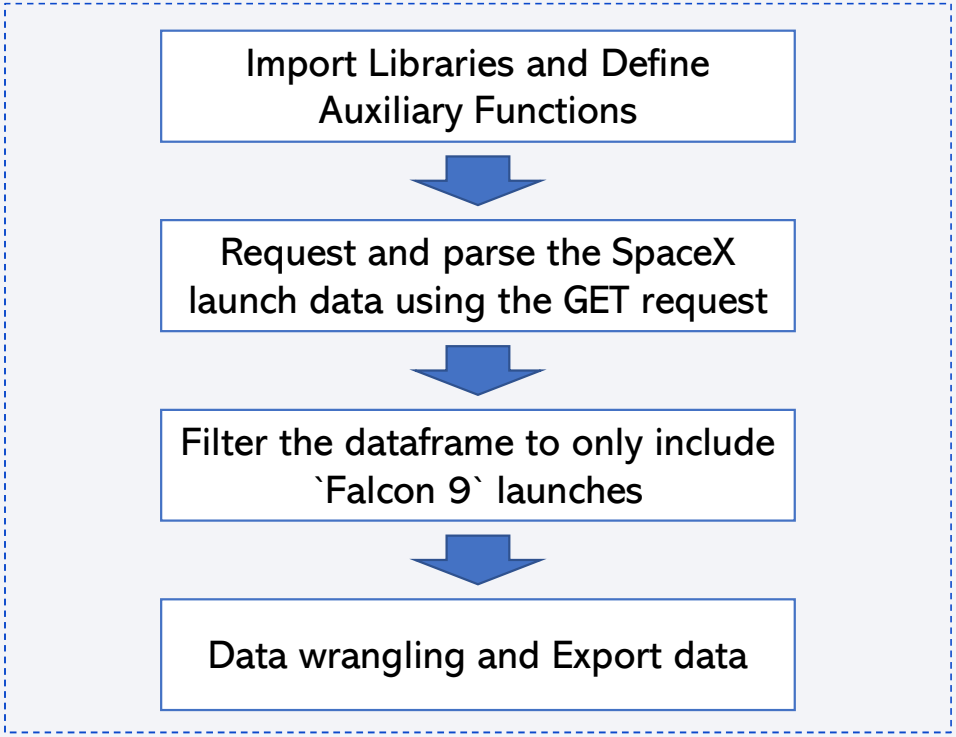
- Data collection methodology
  - SpaceX API
  - Web Scraping

- Present data collection process use key phrases and flowcharts
- GitHub URL as an external reference and peer-review purpose

Data Collection – SpaceX API

- SpaceX data were collected by SpaceX REST calls API as described in the flowchart

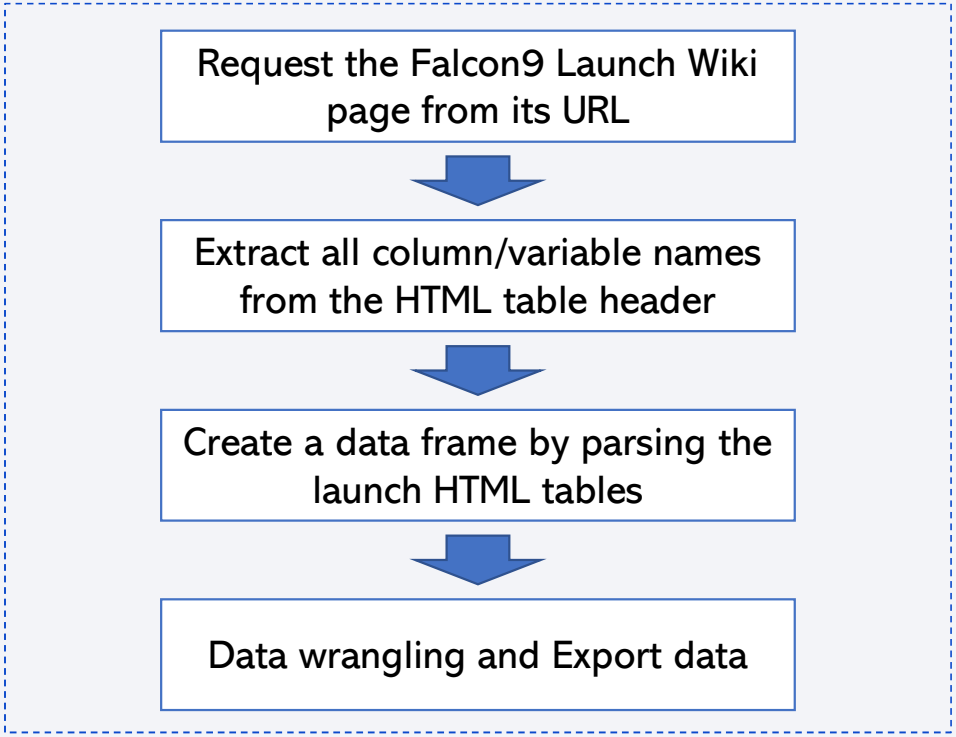
- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/1d1e548d895a03e0eec25f3550d1d5c417137c57/1.%20Data%20Collection%20API.ipynb



Data Collection - Scraping

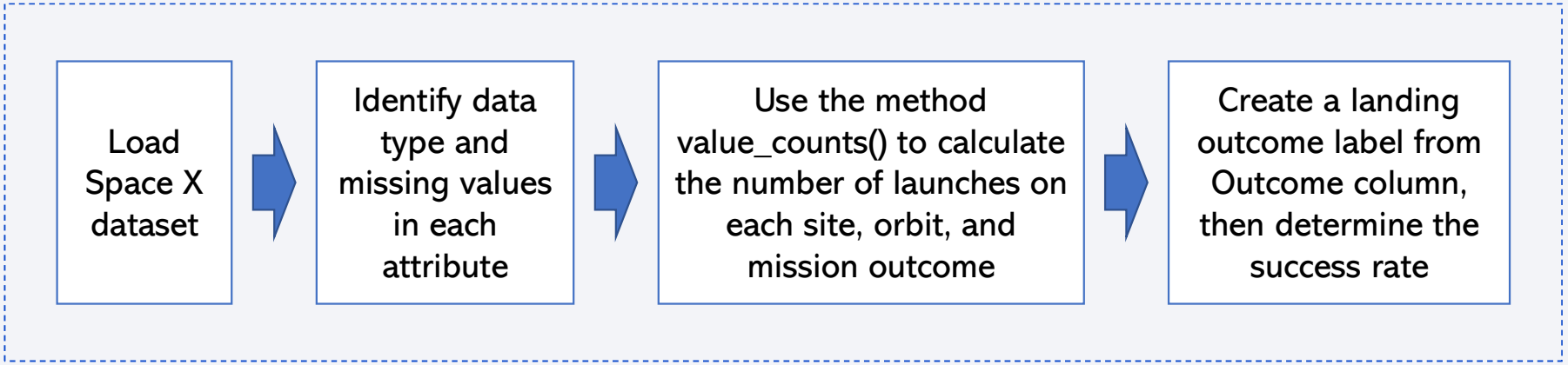
- SpaceX data were collected by web scraping as described in the flowchart

- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/1d1e548d895a03e0eec25f3550d1d5c417137c57/2.%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Analyzing and Cleaning data using Pandas library
- Data wrangling process flowcharts



- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/1d1e548d895a03e0eec25f3550d1d5c417137c57/3.%20EDA%20with%20pandas.ipynb

EDA with Data Visualization

- Summary
  - Use scatter plot to visualize the relationship between Flight Number, Payload, Launch Site and Orbit type
  - Use bar plot to visualize the relationship between success rate of each orbit type
  - Use line plot to visualize the launch success yearly trend
- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/1d1e548d895a03e0eec25f3550d1d5c417137c57/5.%20EDA%20with%20data%20visualization.ipynb

EDA with SQL

- Summary
  - Load data set to DB2 server
  - Query the names of the unique launch sites in the space mission
  - Query the total payload mass carried by boosters launched by NASA (CRS)
  - Query average payload mass carried by booster version F9 v1.1
  - Query the date when the first successful landing outcome in ground pad was achieved
  - Query the total number of successful and failure mission outcomes
  - etc.
- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/03055e087e7901a758dee4ff8933cde2ad62cc74/4.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

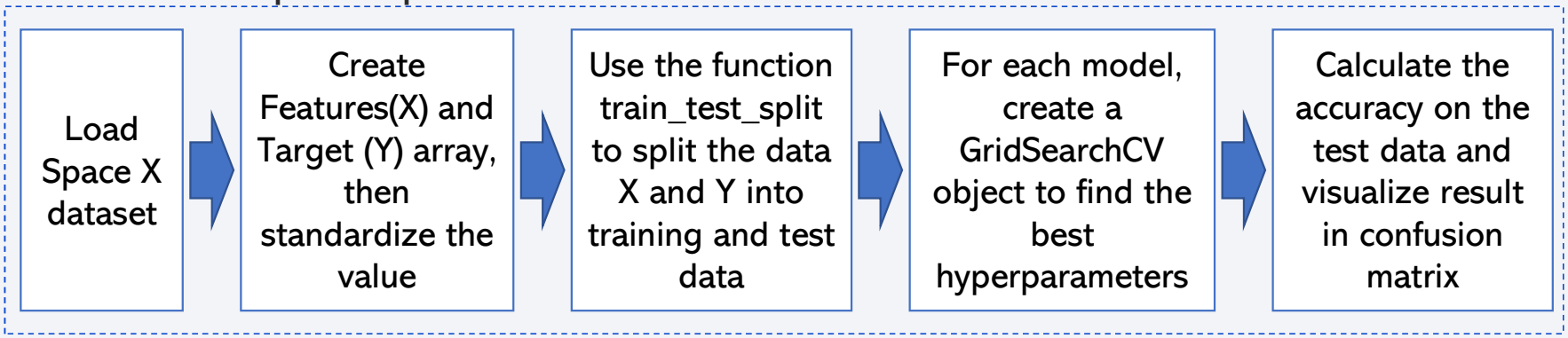
- Summarize map and explain objects
  - Mark all launch sites on a map using circle and marker objects
  - Mark the success/failed launches for each site on the map using marker cluster objects
  - Calculate the distances between a launch site to its proximities using MousePosition, then mark on a map using marker and polyline objects
- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/19a557e1f99aa76b22db08b014a700b531755791/6.%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Summarize dashboard
  - Create dashboard with 4 components including dropdown menu, pie chart, slider, and scatter plot
- Explain plots and interactions
  - Dropdown menu for selecting launch sites
  - Pie chart to visualize success rate in each launch site
  - Slider to select payload range
  - Scatter plot to visualize relationship launch site, payload, and booster version
- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/19a557e1f99aa76b22db08b014a700b531755791/7.%20Dashboard%20Application%20with%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

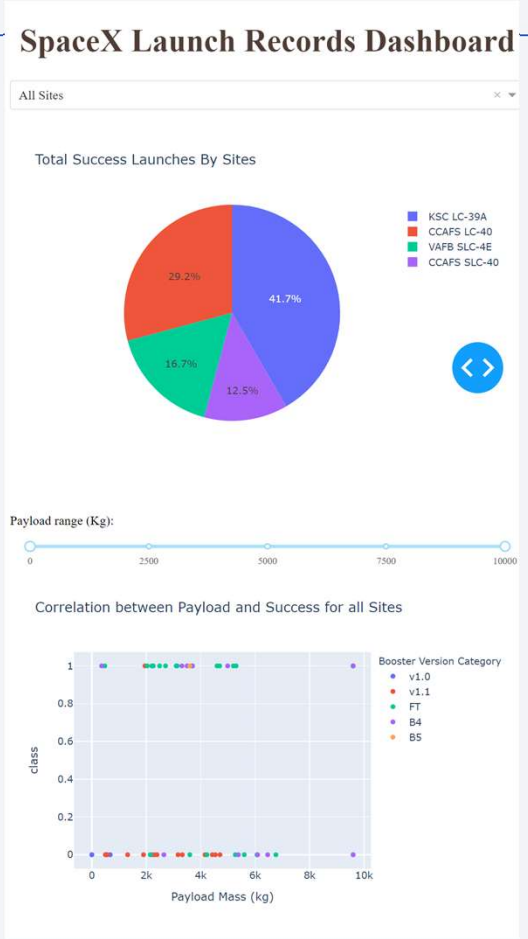
- Summary
  - Develop best model from SVM, Classification Trees and Logistic Regression using Scikit-learn library to determine if Space Y should reuse the first stage rocket
- Model development process flowchart

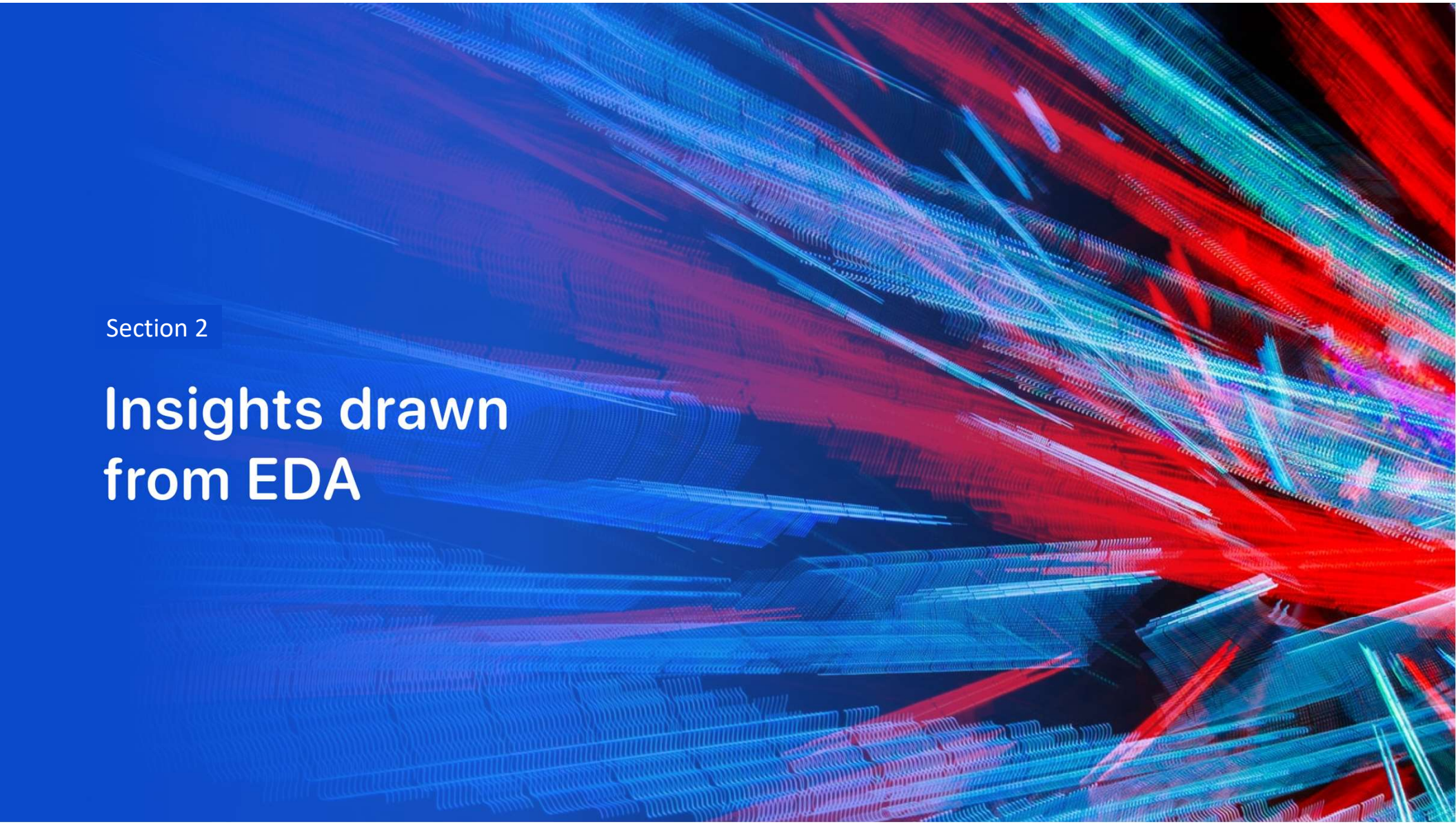


- GitHub URL
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project/blob/9cf3cb83c9b276ec9cb17f9395ef0c5fc3378cfd/8.%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Visualize trend and relationship between each attribute
  - Query for specific information
- Interactive analytics demo in screenshots
- Predictive analysis results
  - The best classification model is Decision Tree Classifier model with accuracy of 94.44%



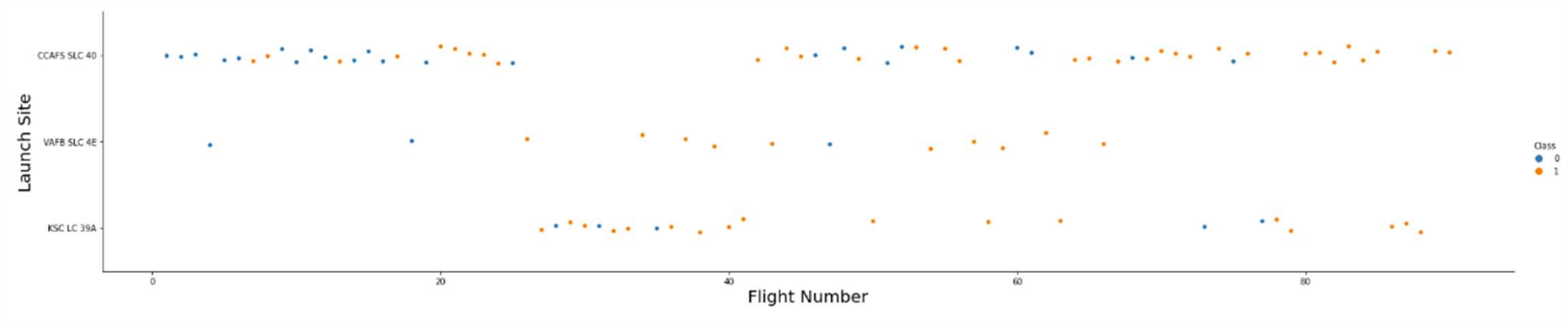


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

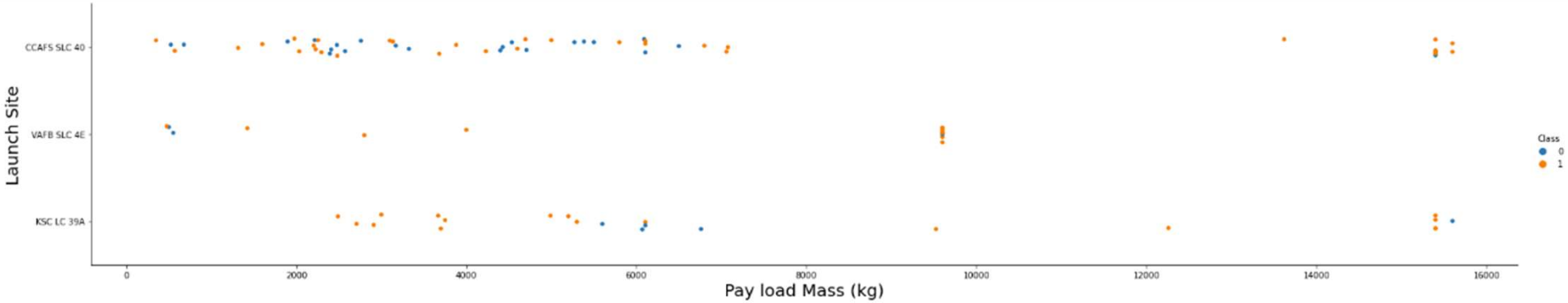
- Scatter plot of Flight Number vs. Launch Site



- Launch Site CCAFS SLC 40 has the most number of launch

Payload vs. Launch Site

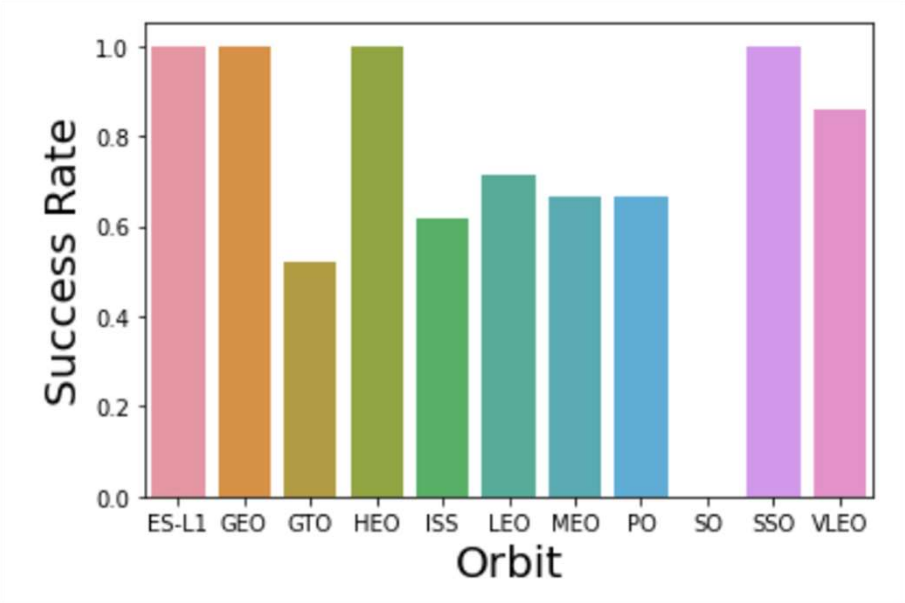
- Scatter plot of Payload vs. Launch Site



- Launch Site VAFB SLC 4E do not have launch that have pay load mass more than 10,000 kg

Success Rate vs. Orbit Type

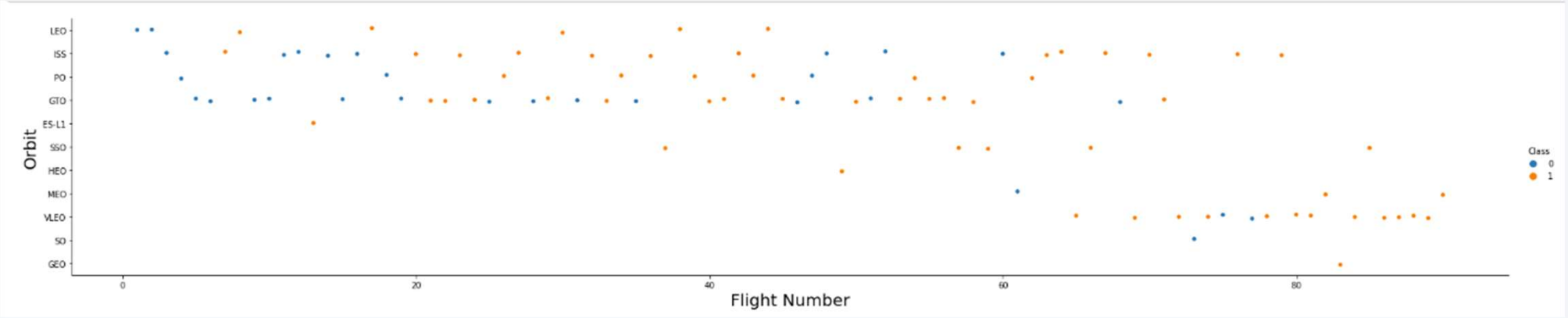
- Bar chart for the success rate of each orbit type



- Orbit ES-L1, GEO, HEO, SSO have the best success rate

Flight Number vs. Orbit Type

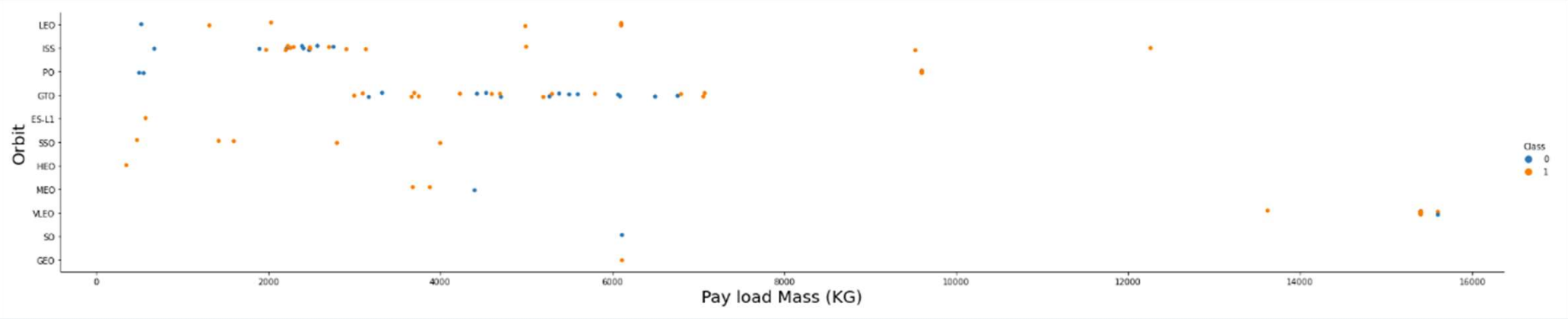
- Scatter point of Flight number vs. Orbit type



- Latest flight number usually have launch to VLEO orbit

Payload vs. Orbit Type

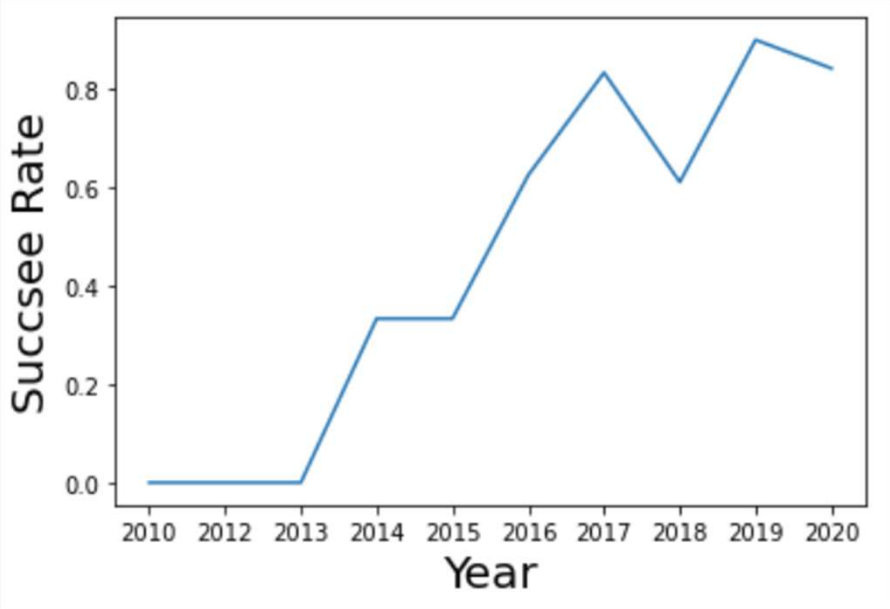
- Scatter point of payload vs. orbit type



- VLEO orbit has the heaviest pay load mass launch

Launch Success Yearly Trend

- Line chart of yearly average success rate



- The success rate tend to improve as the year pass

All Launch Site Names

- Find the names of the unique launch sites

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

- Space X have 4 launch sites name above

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- The above are 5 records where launch sites begin with `CCA`

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://tgk79204:***@98
Done.
```

1
45596

- The total payload carried by boosters from NASA is 45,596 kg

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://tgk79204:***@98538591-7
Done.
```

1
2534

- The average payload mass carried by booster version F9 v1.1 is 2,534 kg

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (ground pad)'
```

```
 * ibm_db_sa://tgk79204:***@98538591-7217-4024-b0
Done.
```

1
2015-12-22

- The dates of the first successful landing outcome on ground pad is 22 December 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

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

- Total of 4 boosters named above have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Most of the mission are successful

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

- There are 12 total booster which have carried the maximum payload mass

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

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

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- There are 2 failed landing outcomes in drone ship in year 2015

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

- There are 8 landing between the date 2010-06-04 and 2017-03-20, where the most common outcome is "No attempt"

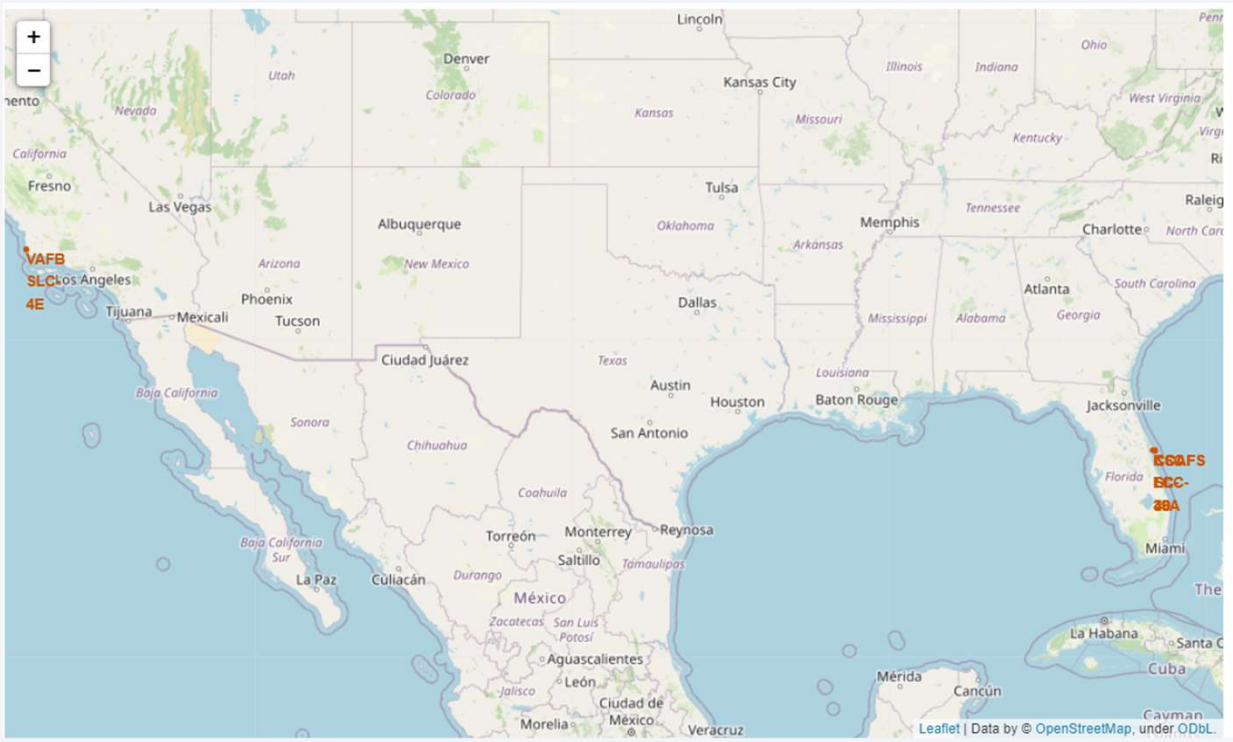
Landing _Outcome	c
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

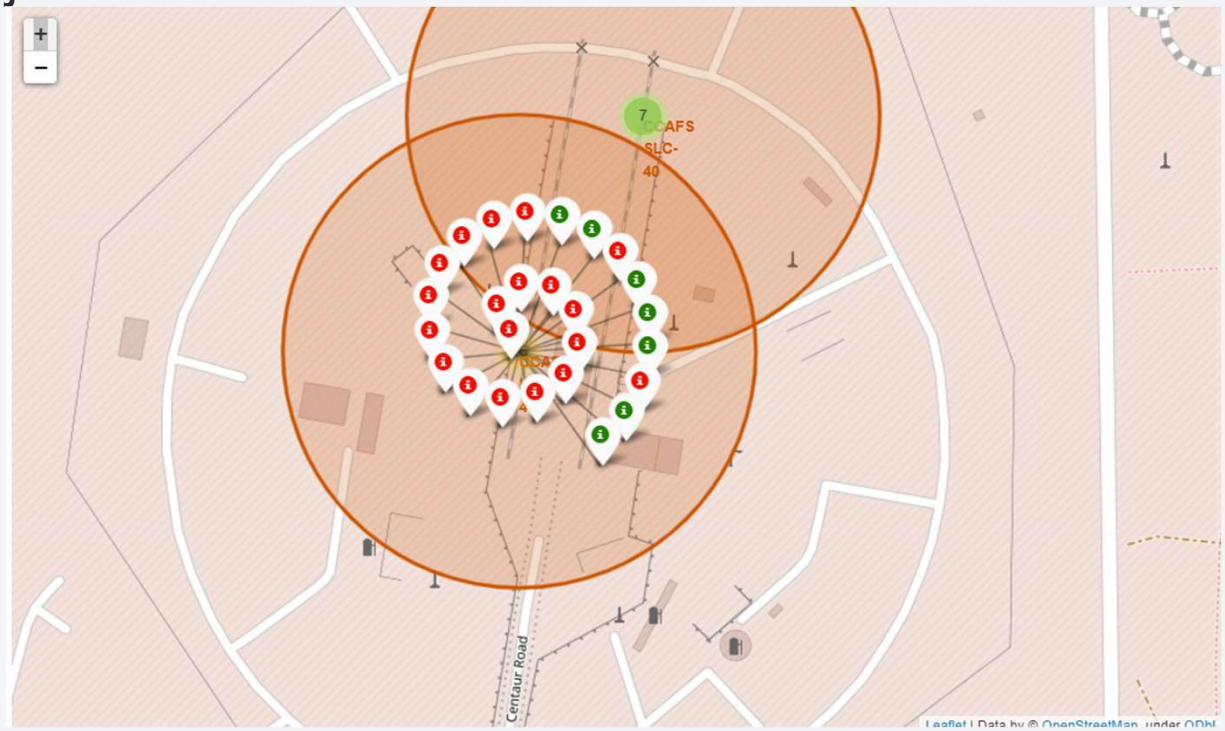
Folium Launch Sites Map

- Create a folium map to mark all launch sites location with circle and marker label



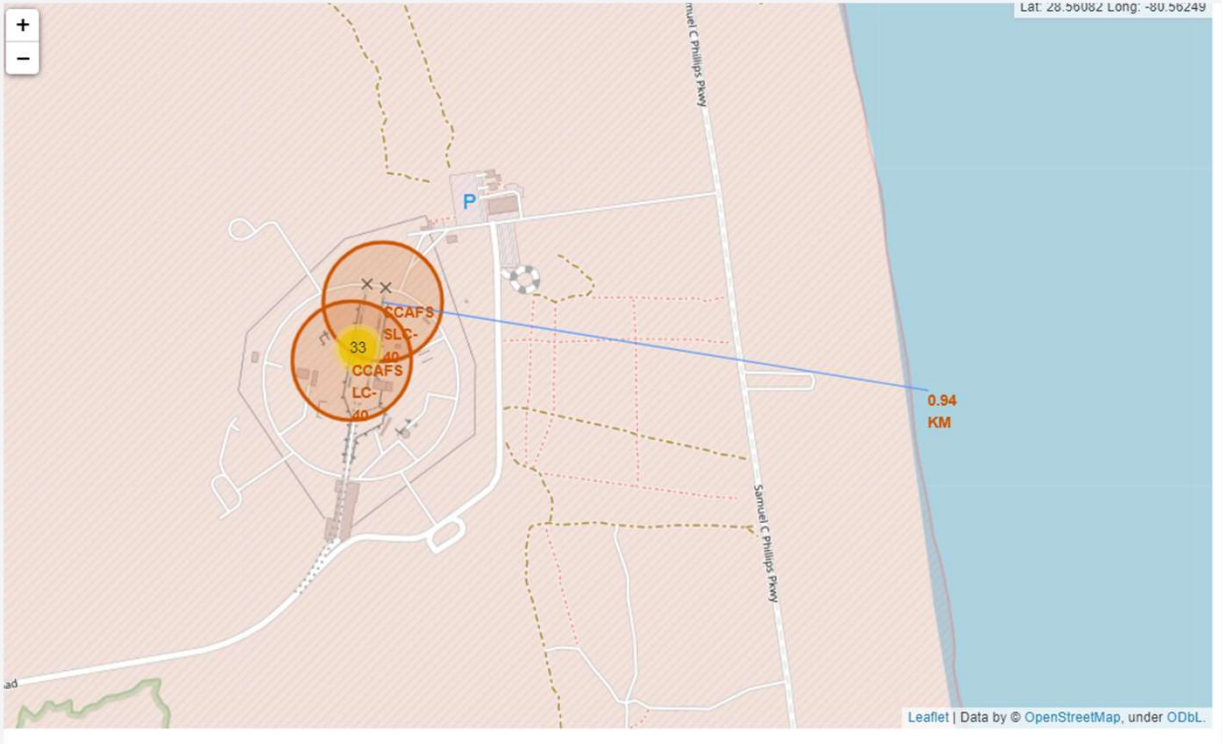
Success & Failed Launches Map Marker

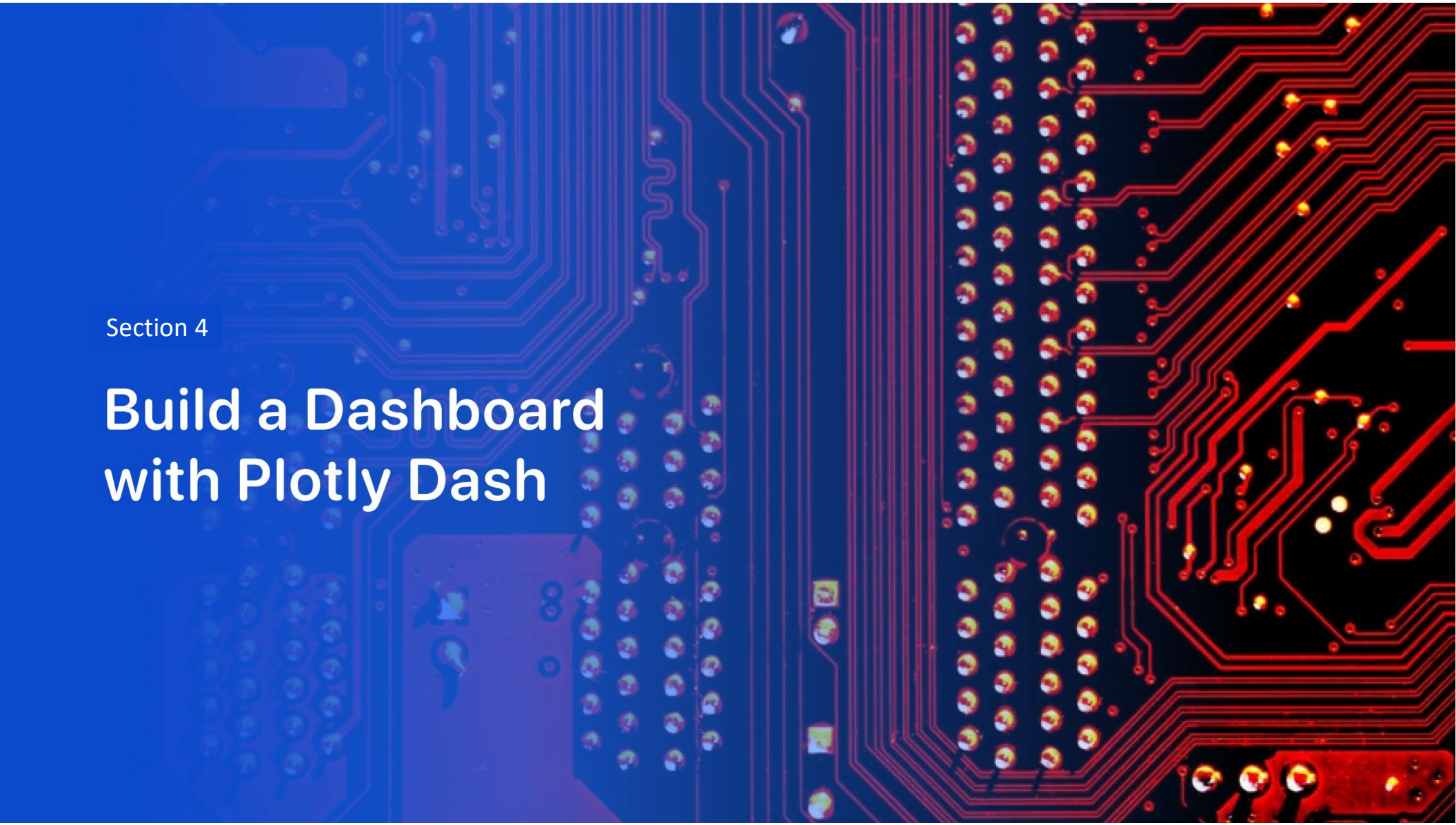
- Mark the success & failed launches for each site on the map using marker cluster objects



Distances between a Launch Site to its Proximities

- Calculate the distances between a launch site to its proximities, then mark on a map using marker and polyline objects



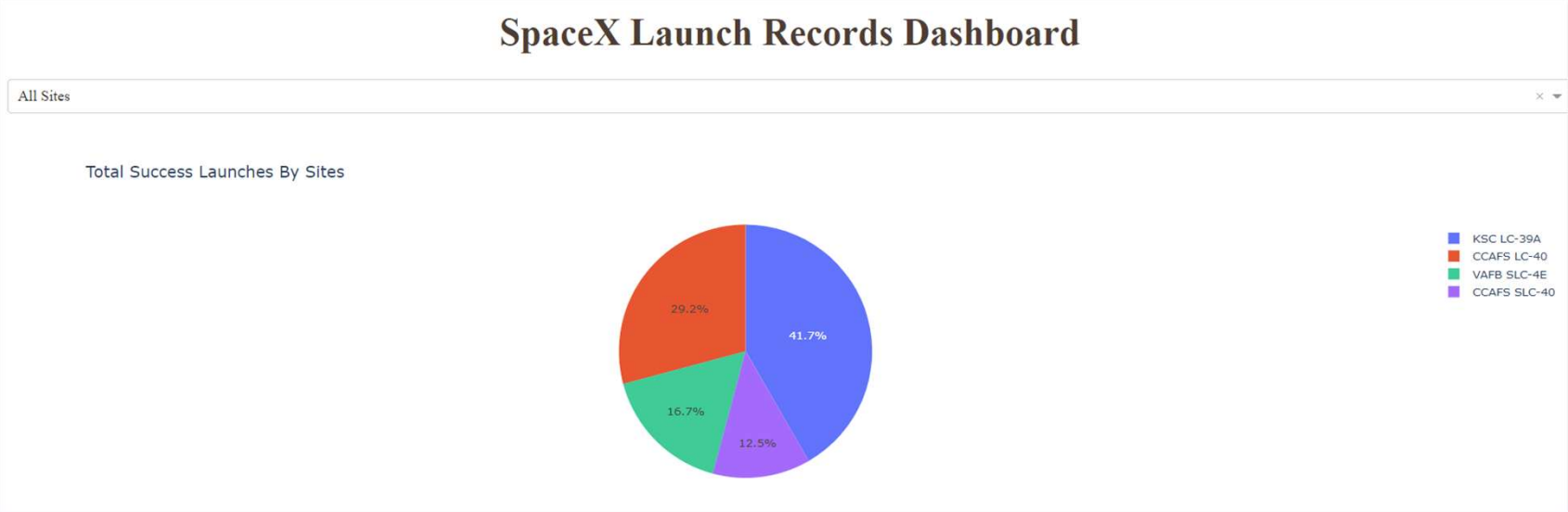


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

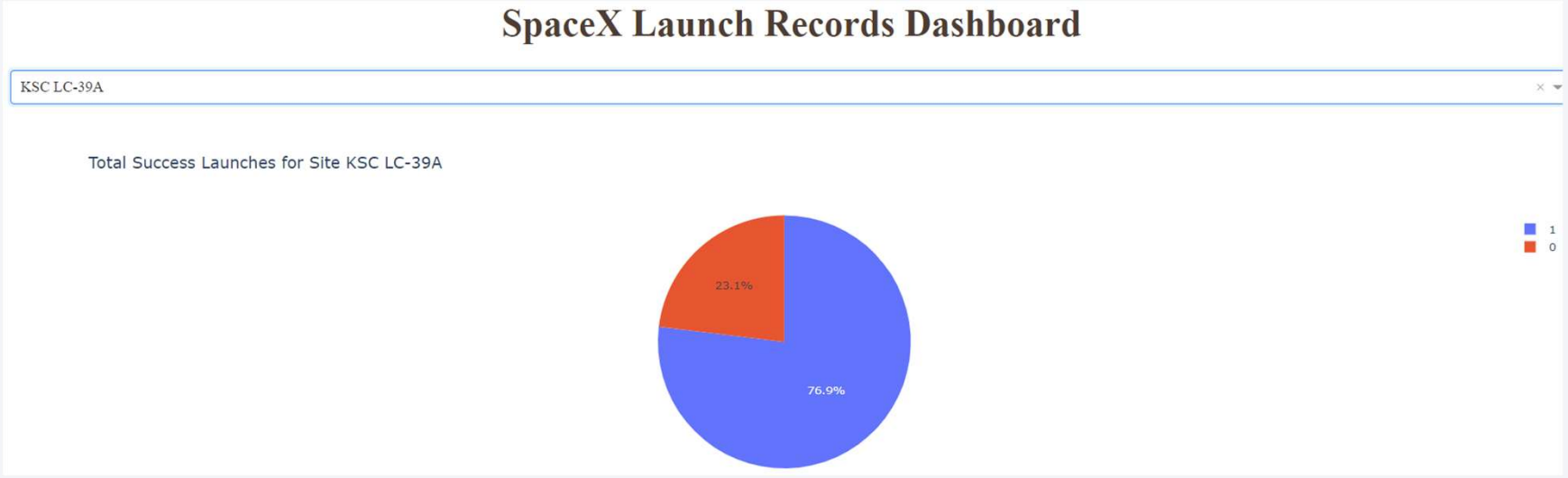
- Dashboard indicate launch success count for all sites in a pie chart



- Launch site KSC LC-39A has the most success launch

Launch Site with Highest Launch Success Ratio

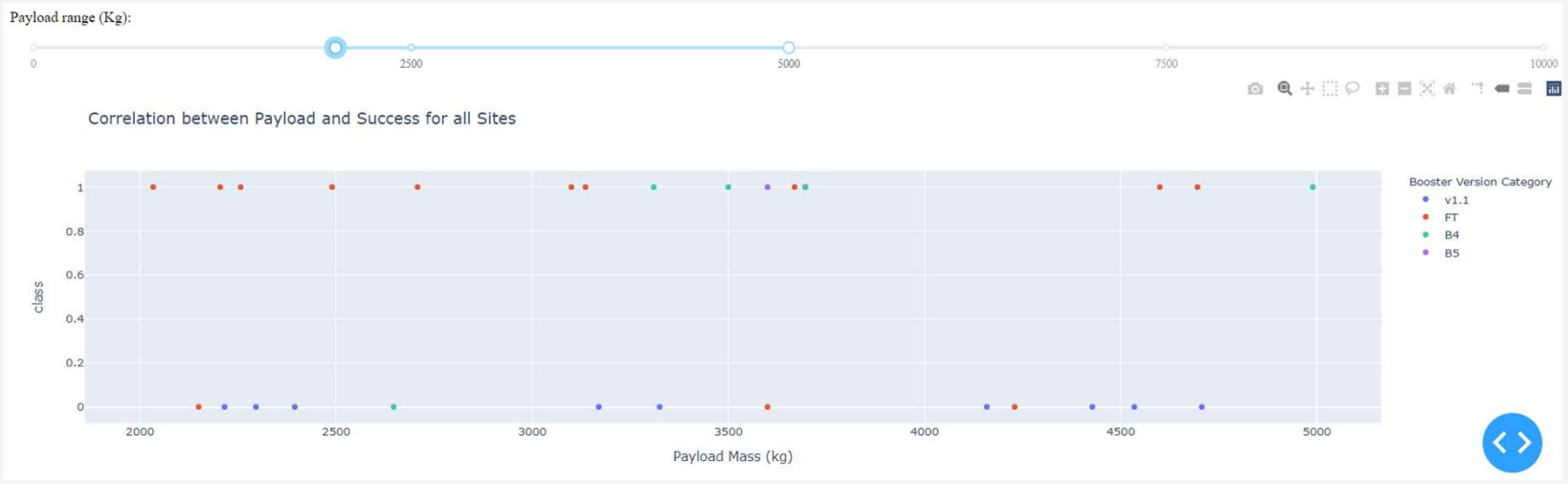
- Dashboard indicate the launch site with highest launch success ratio in pie chart



- Launch site KSC LC-39A has the highest launch success ratio of 76.9%

Payload vs. Launch Outcome Dashboard

- Dashboard indicate Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



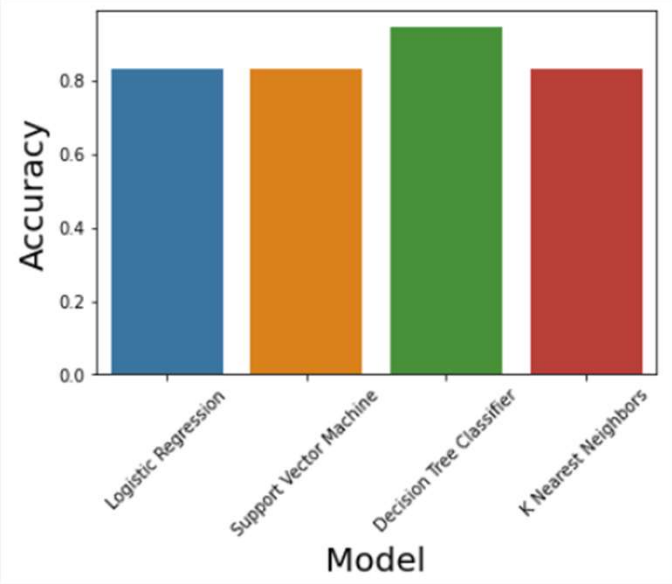
- Booster version FT has the highest success rate in pay load range 2,000 – 5,000 kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

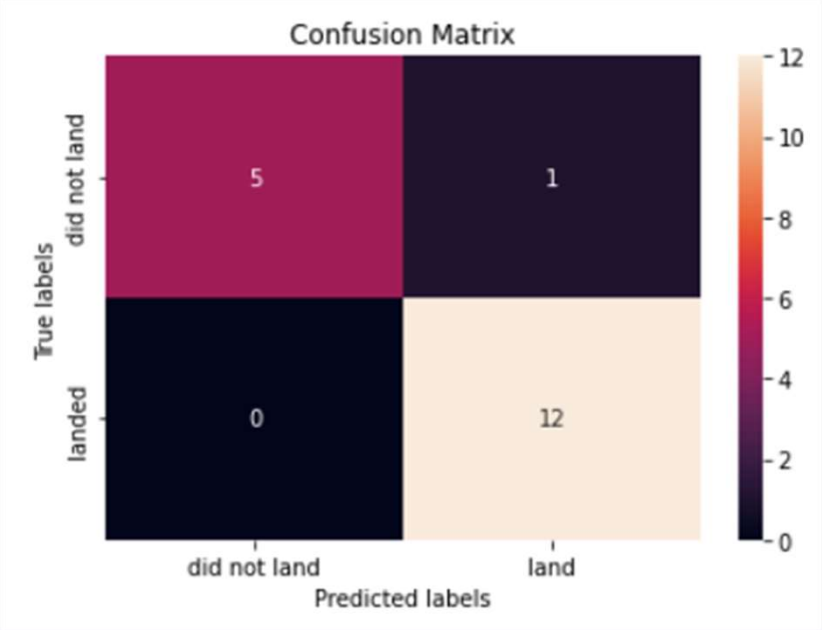
- Visualize the built model accuracy for all built classification models, in a bar chart



- Decision Tree Classifier model has the highest classification accuracy

Confusion Matrix

- Confusion matrix of the best performing model (Decision Tree Classifier)



- From the confusion matrix, we see that Decision Tree Classifier can distinguish between the different classes. However, the major problem is false positives.

Conclusions

- EDA with SQL and Visualization
  - Orbit ES-L1, GEO, HEO, SSO have the best success rate
  - The success rate tend to improve as the year pass
  - The average payload mass carried by booster version F9 v1.1 is 2,534 kg
  - There are 12 total booster which have carried the maximum payload mass
- Folium Map & Dashboard
  - Calculate the distances between a launch site to its proximities
  - Launch site KSC LC-39A has the highest launch success ratio of 76.9%
- Predictive Analysis
  - The best classification model is Decision Tree Classifier model with accuracy of 94.44%
  - From the confusion matrix, Decision Tree Classifier can distinguish between the different classes, but have false positives as a major problem

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
- GitHub URL for this IBM Data Science Capstone Project
  - https://github.com/supat-roong/IBM-Data-Science-Capstone-Project

Thank you!