IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ersin Onur Yukay
15.04.2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scrapping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visualization with Folium and Plotly
  - Machine Learning Prediction
- Summary of all results
  - Results from EDA
  - Interactive Visualization Screenshots
  - Results of the predictive analysis.

Introduction

- Project background and context

  SpaceX advertises Falcon 9 launches with a cost of 62 million dollars whereas its competitions are spending nearly 100 million dollars more. The reason behind this situation is SpaceX can reuse its first stage. This concludes that, if the first stage can be reused, the cost decreases significantly. This information is going to used by an alternative company called SpaceY to bid against SpaceX for rocket launches. Therefore, in this project a machine learning pipeline will be built to predict it the first stage will land successfully.

- Problems you want to find answers

  - What factors are going to affect our analysis ?

  - What conditions must be satisfied in terms of our success ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX API and Web Scrapping from Wikipedia.
- Perform data wrangling
  - In order to implement machine learning algorithms, categorical variables were encoded with one-hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree Classifier and KNN algorithms are tuned by using GridSearchCV.

Data Collection

- The data for this project was collected in two ways.
  - Firstly, a get request was sent to SpaceX API.
  - After obtaining the SpaceX Data, it was transformed into a pandas DataFrame using .json() methods.
  - Then necessary data cleaning was handled and missing values were dealt with.
  - Secondly, using BeautifulSoup library, from Wikipedia, web scrapping was done in order to obtain Falcon 9 launch records.
  - Those records are extracted as HTML tables, which are then parsed and converted into a pandas DataFrame.

Data Collection – SpaceX API

- SpaceX Data was obtained as follows :
- Notebook link https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Data%20Collection%201.ipynb

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

REQUEST AND PARSE THE SPACEX LAUNCH DATA USING GET REQUEST

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_
response.status_code
# HTTP response code '200' means success.
```

```
200
```

```
data = pd.json_normalize(response.json())
data.head()
```

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	...	links.reddit.media	links.r
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'mer...	Engine failure at 33 seconds and loss of vehicle	[]	[]	...	None	
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'har...	Successful first stage burn and transition to ...	[]	[]	...	None	

Data Collection - Scraping

- Web Scrapping was done using BeautifulSoup to extract Falcon 9 launch records.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Data%20Collection%202.ipynb

WEB SCRAPPING FROM WIKIPEDIA FOR MORE DATA

In [3]:
```
static_url_wiki = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response_wiki = requests.get(static_url_wiki)
```

In [4]:
```
soup = BeautifulSoup(response_wiki.text, 'html.parser')
soup.title
```

Out[4]:
```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

In [5]:
```
html_tables = soup.find_all('table')
first_launch_table = html_tables[2]
first_launch_table
```

Out[5]:
```
<table class="wikitable plainrowheaders collapsible" style="width: 100%;">
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Version,<
br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-Dragon-12">[c]</a></sup>
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
```

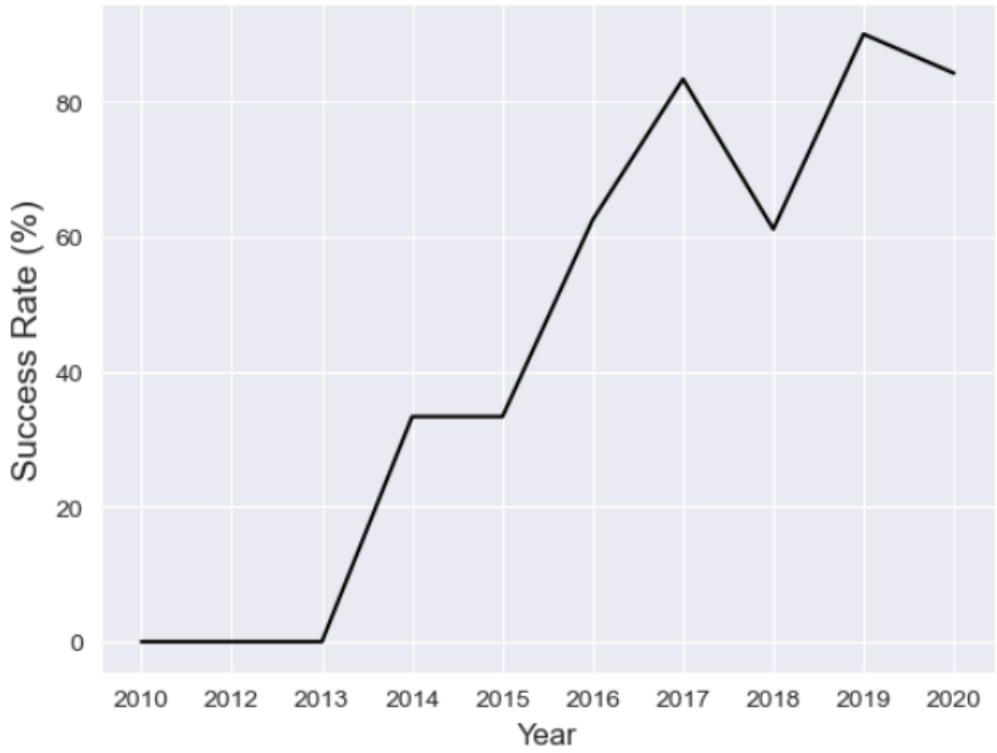
Data Wrangling

- In this part of the project,
  - Number of occurences of each orbit was calculated.
  - Number of launches from each launch site was calculated.
  - Using the outcome column, a new column has been created and named ‘class’, which indicates if the launch was successful or not by encoding it as zeros and ones. (0 -> not successful, 1-> successful)

- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- The data was explored by visualizing flight numbers' relationship between payload mass and launch site, success rate of each orbit, orbit-flight number relationship, orbit-payload mass relationship and lastly, yearly trend of success rate which is increasing as time passes as it should be.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Exploratory%20Data%20Analysis.ipynb



EDA with SQL

- ipython-sql was used to query the data of Falcon 9 launches.
- Various information was obtained by those SQL queries. Some of them are :
  - The total payload mass carried.
  - The average payload mass carried by booster version F9 v1.1
  - Total number of successful missions.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20EDA%20with%20SQL.ipynb

```
In [3]: %load_ext sql

In [4]: %sql sqlite://

In [5]: %%sql
CREATE TABLE spacex(
   Date                DATE  NOT NULL PRIMARY KEY
  ,Time_UTC            VARCHAR(8) NOT NULL
  ,Booster_Version     VARCHAR(14) NOT NULL
  ,Launch_Site         VARCHAR(12) NOT NULL
  ,Payload             VARCHAR(61) NOT NULL
  ,PAYLOAD_MASS_KG_    INTEGER  NOT NULL
  ,Orbit               VARCHAR(11) NOT NULL
  ,Customer            VARCHAR(57) NOT NULL
  ,Mission_Outcome     VARCHAR(32) NOT NULL
  ,Landing_Outcome     VARCHAR(22) NOT NULL
);
```

Build an Interactive Map with Folium

- All launch sites are marked, map objects such as markers, lines and circles were used to enhance data visualization.
- To see how many successful and unsuccessful launches were occurred in each site, marker cluesters are used and they were colored by using the class of each launch.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Folium.ipynb

Build a Dashboard with Plotly Dash

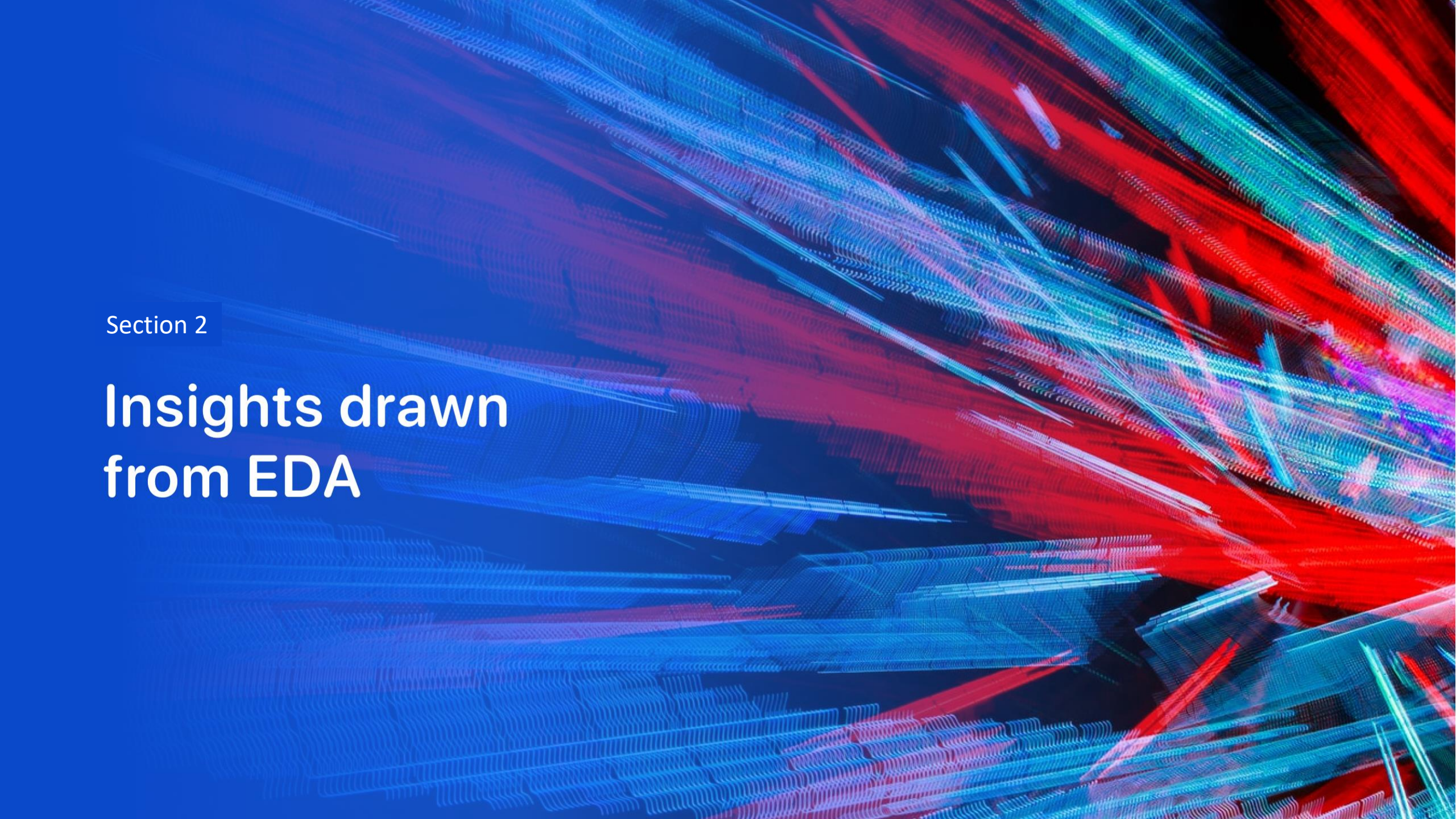
- An interactive dashboard was created using Plotly Dash.
- Pie charts showing launch information from each sites were presented in Dash.
- A scatter graph showing the relationship with Outcome and Payload Mass for different booster versions were added.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Plotly.ipynb

Predictive Analysis (Classification)

- The data was loaded using numpy and pandas libraries, then it was scaled(transformed), split into test and training data using train test split.
- Different machine learning models were used and those models were tuned in terms of their hyperparameters using GridSearchCV.
- After determining the best hyperparameters, by looking the accuracy of each model, the best model was selected.
- Notebook Link : https://github.com/OnurYukay/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Capstone%20Project%20-%20Machine%20Learning%20Prediction.ipynb

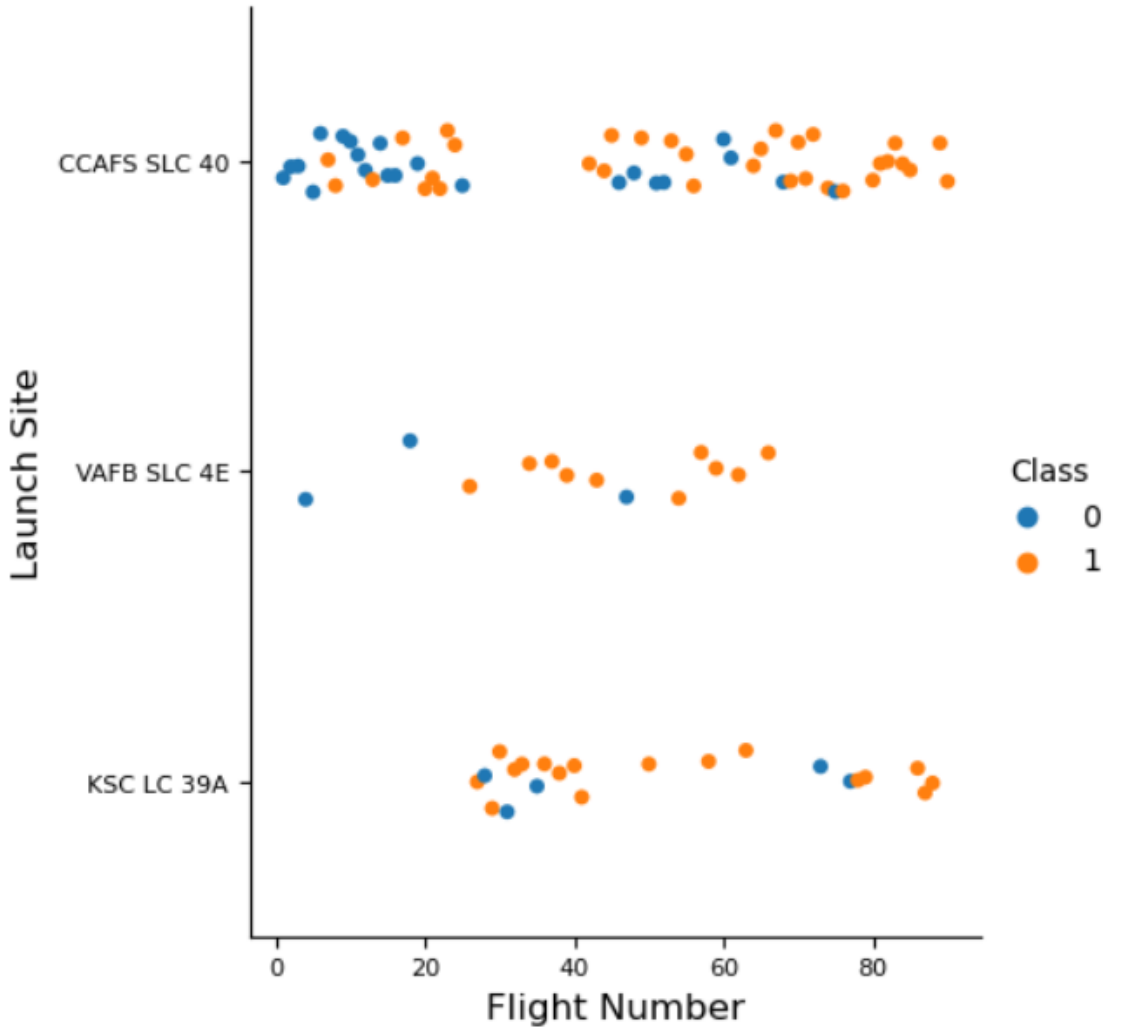
Results

- After EDA, it was clear to see which features should be selected in our model.
- Visualizations will be shown in the next section of this presentation.
- Decision Tree Classifier was selected as the best performing model in terms of its accuracy.


Section 2
Insights drawn from EDA

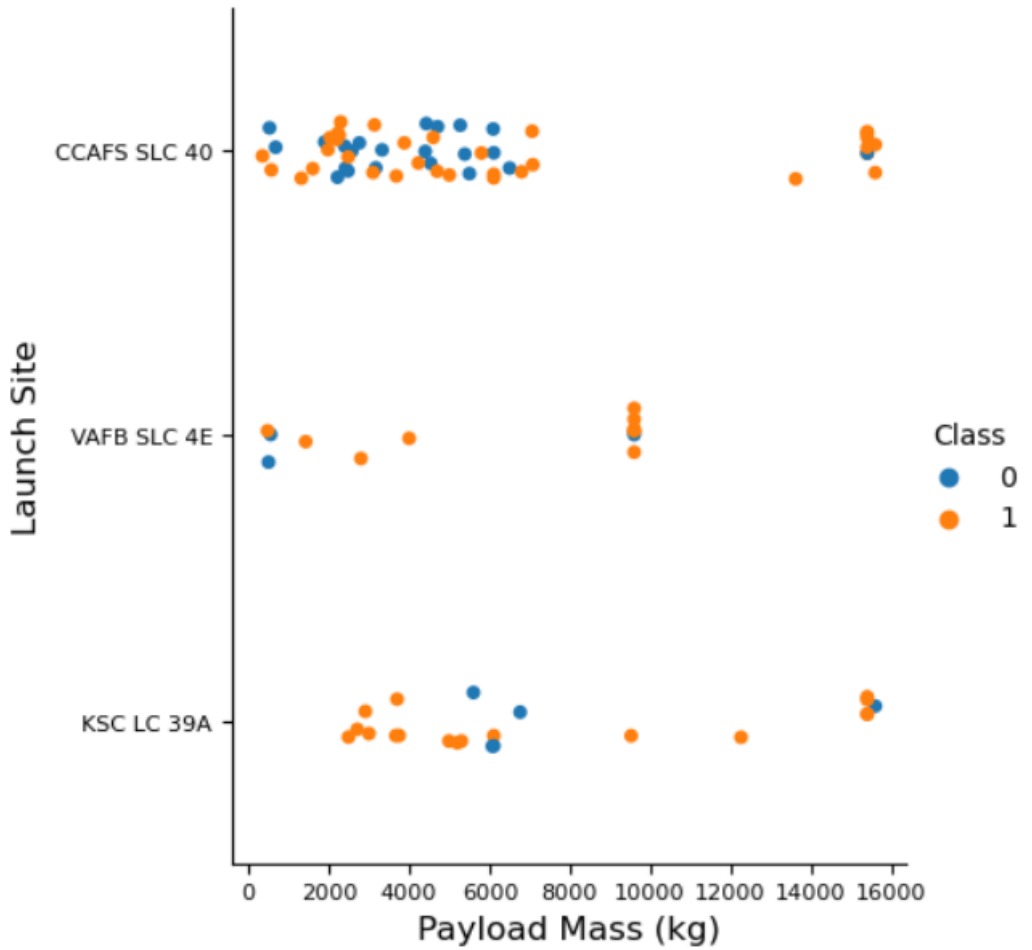
Flight Number vs. Launch Site

- As the flight number increases the success rate is visibly increasing.
- Flight number is the order of that flight. This means that, as the experience increases the success is also increasing.



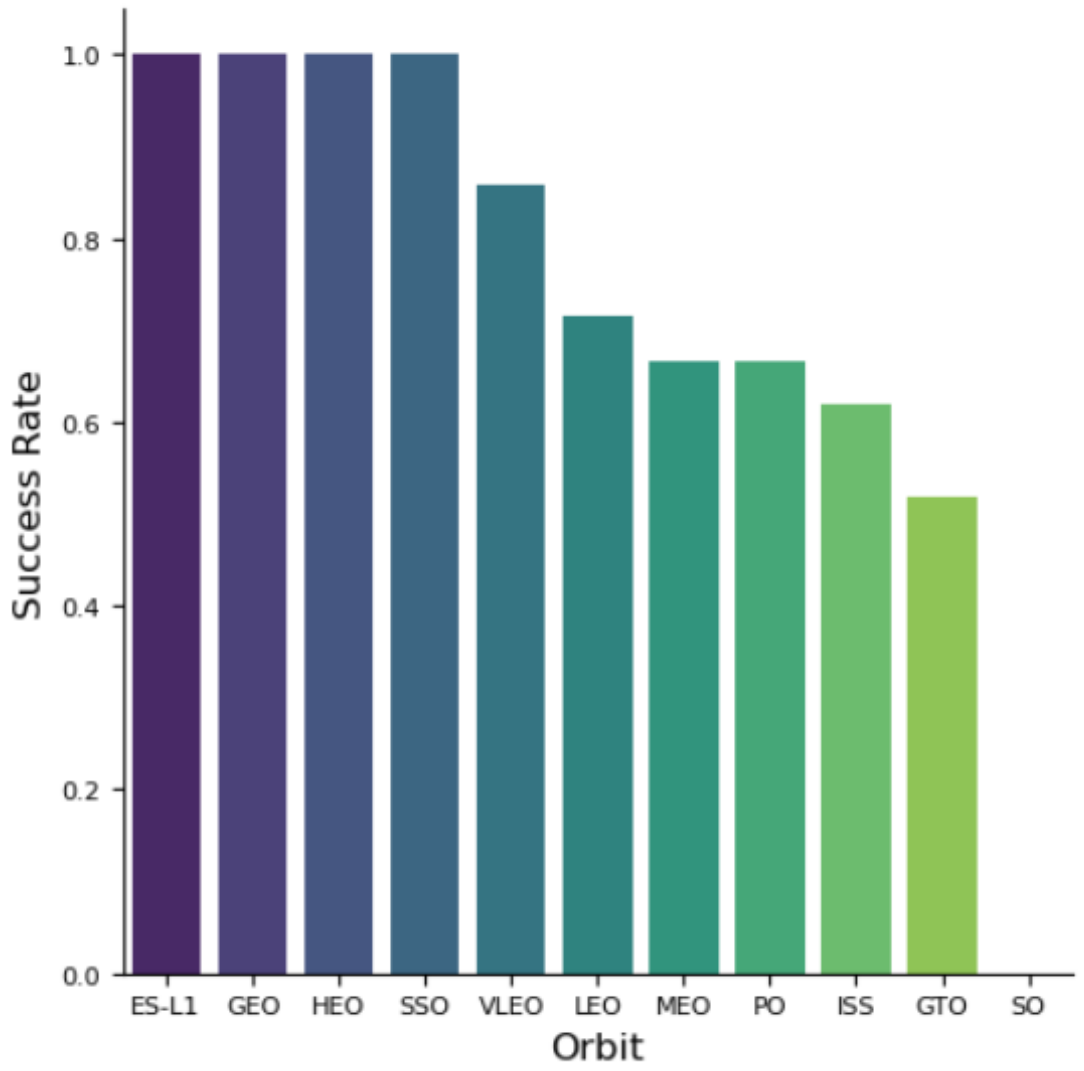
Payload vs. Launch Site

- There is a general trend of increasing success rate with increasing payload mass.
- However, it is more clear on CCAFS SLC 40 launch site.



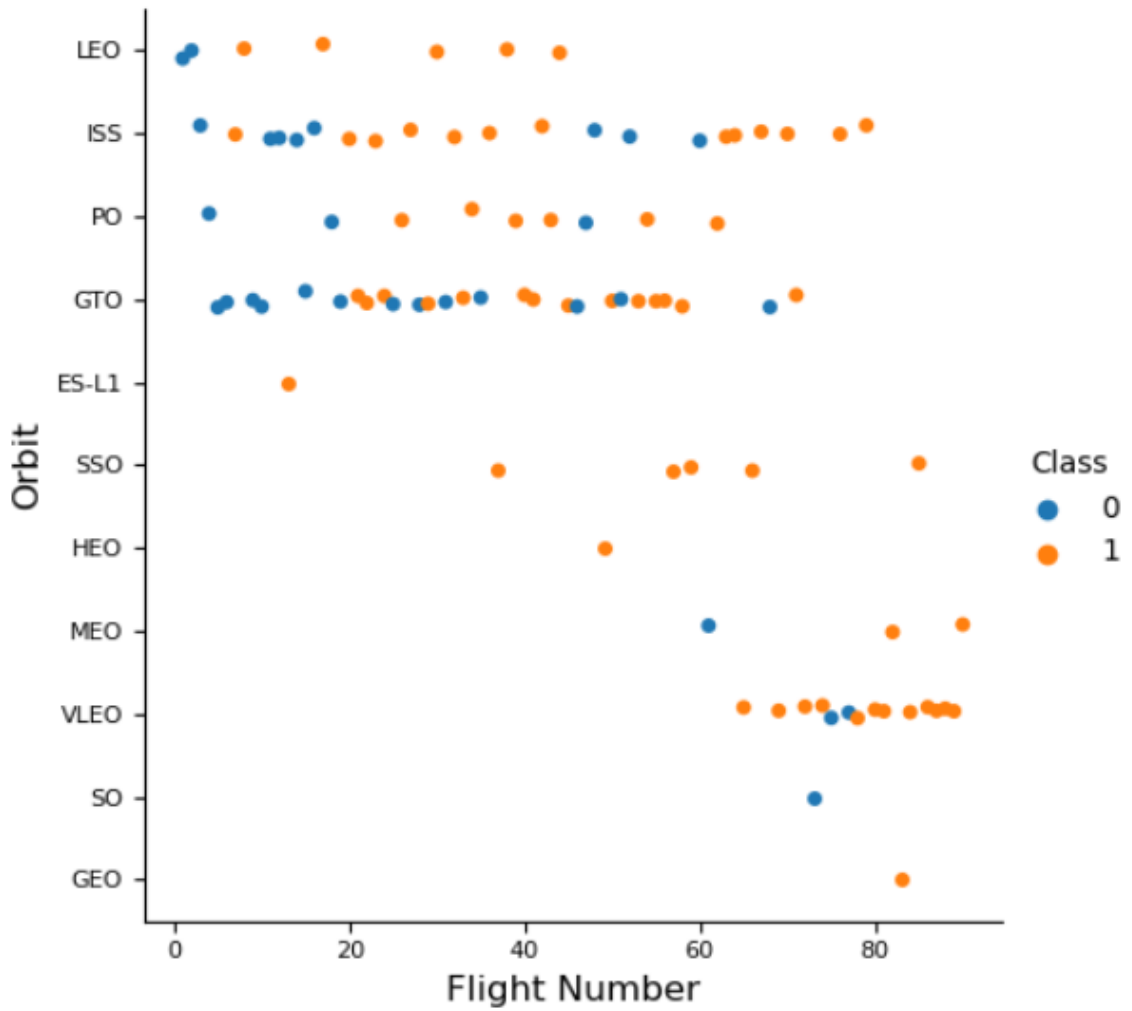
Success Rate vs. Orbit Type

- As shown in the figure, ES-L1, GEO, HEO and SSO orbits have 100% success rate.
- SO orbit might be avoided according to the analysis.



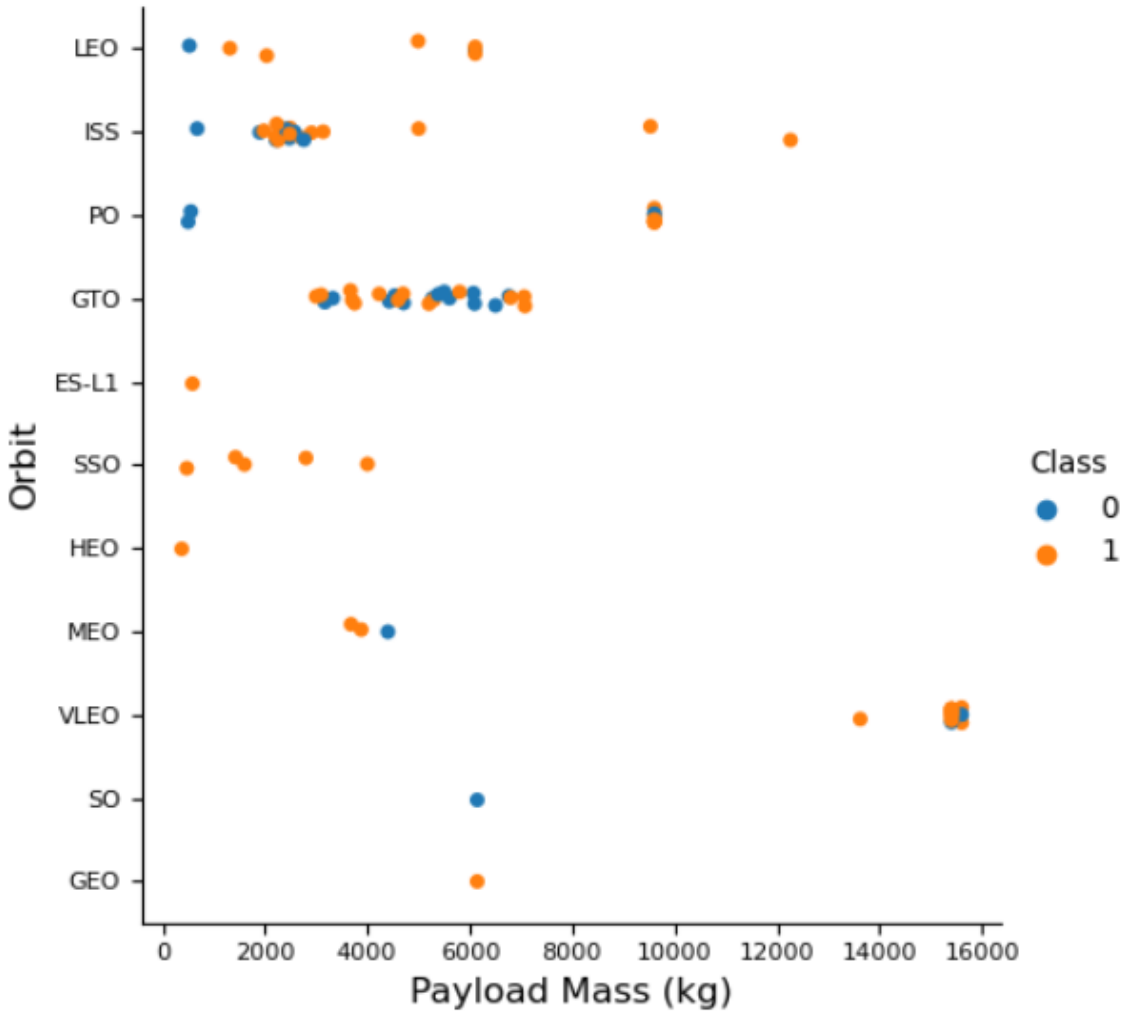
Flight Number vs. Orbit Type

- In LEO orbit, it is clearly seen that success increases with flight number.
- MEO, VLEO, SO, GEO orbits were only used on the last quarter of the flights.



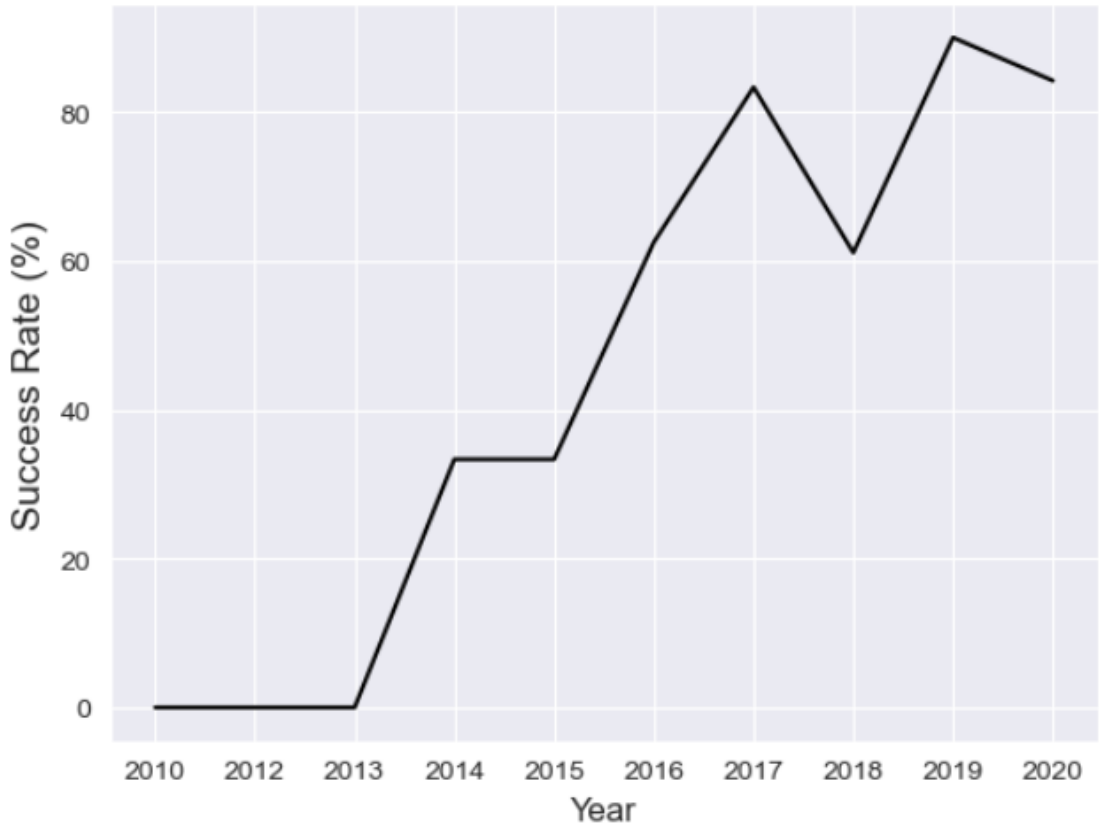
Payload vs. Orbit Type

- For ISS, LEO and VLEO, higher mass resulted in success.



Launch Success Yearly Trend

- As it can be expected, with time the success rate increased due to previous trials.



All Launch Site Names

- Launch site names are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.
- The query was built using a DISTINCT key word to only filter unique launch site names.

```
%%sql

SELECT
    DISTINCT Launch_Site AS Distinct_Launch_Site
FROM spacex
```

```
 * sqlite://
Done.
```

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

Launch Site Names Begin with 'CCA'

- LIKE operator was used to filter Launch Site names.
- It is limited with 5 records using LIMIT keyword.

```
%%sql

SELECT *
FROM spacex
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

```
 * sqlite://
Done.
```

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

- Total Payload Mass was calculated as 45596 kg.

```
%%sql

SELECT
    SUM(PAYLOAD_MASS_KG_) AS PAYLOAD_MASS_KG
FROM spacex
WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite://
Done.
```

PAYLOAD_MASS_KG
45596

Average Payload Mass by F9 v1.1

- Average Payload Mass by F9 v1.1 was calculated as 2928.4 kg .

```
%%sql

SELECT
    AVG(PAYLOAD_MASS_KG_) AS AVG_MASS_KG
FROM spacex
WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite://
Done.
```

AVG_MASS_KG
2928.4

First Successful Ground Landing Date

- First successful landing on the ground pad was occurred on 22 December 2015.

```
%%sql

SELECT
    (MIN(Date)) AS FIRST_DATE
FROM spacex
WHERE Landing_Outcome = 'Success (ground pad)'
```

```
 * sqlite://
Done.
```

FIRST_DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Associated Booster Versions are as follows :
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

```
%%sql

SELECT
    Booster_Version AS Booster_Name
FROM spacex
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000
```

```
 * sqlite://
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Out of 101 missions, 100 of them are resulted in success !

```
%%sql

SELECT
    SUM(X) AS SUCCESS,
    COUNT(1)- SUM(X) AS FAILURE
FROM(SELECT
        CASE
            WHEN Mission_Outcome LIKE '%Success%'
            THEN 1
            ELSE 0
        END AS X
     FROM spacex)
```

```
 * sqlite://
Done.
```

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

- There are 12 different Boosters which have carried the maximum payload.

```
%%sql

SELECT
    DISTINCT Booster_Version AS Boosters_Carried_Max_Mass
FROM spacex
WHERE PAYLOAD_MASS_KG_ = (
    SELECT
        MAX(PAYLOAD_MASS_KG_) AS MAX_LOAD
    FROM spacex
    )
```

```
 * sqlite://
Done.
```

Boosters_Carried_Max_Mass
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

- There are 2 records for Failure(Drone Ship) in 2015.
- Both of them are launched from CCAFS LC-40 launch site.

```
%%sql

SELECT
    Booster_Version,
    Launch_Site
FROM spacex
WHERE Landing_Outcome = 'Failure (drone ship)' AND strftime('%Y', Date) = '2015'
```

```
 * sqlite://
Done.
```

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- After grouping the missions according to their landing outcomes, this is the scenerio between 2010-06-04 and 2017-03-20.

```
%%sql

SELECT
    Landing_Outcome,
    COUNT(Landing_Outcome) AS Occurence
FROM spacex
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY 2 DESC
```

```
 * sqlite://
Done.
```

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

Section 3

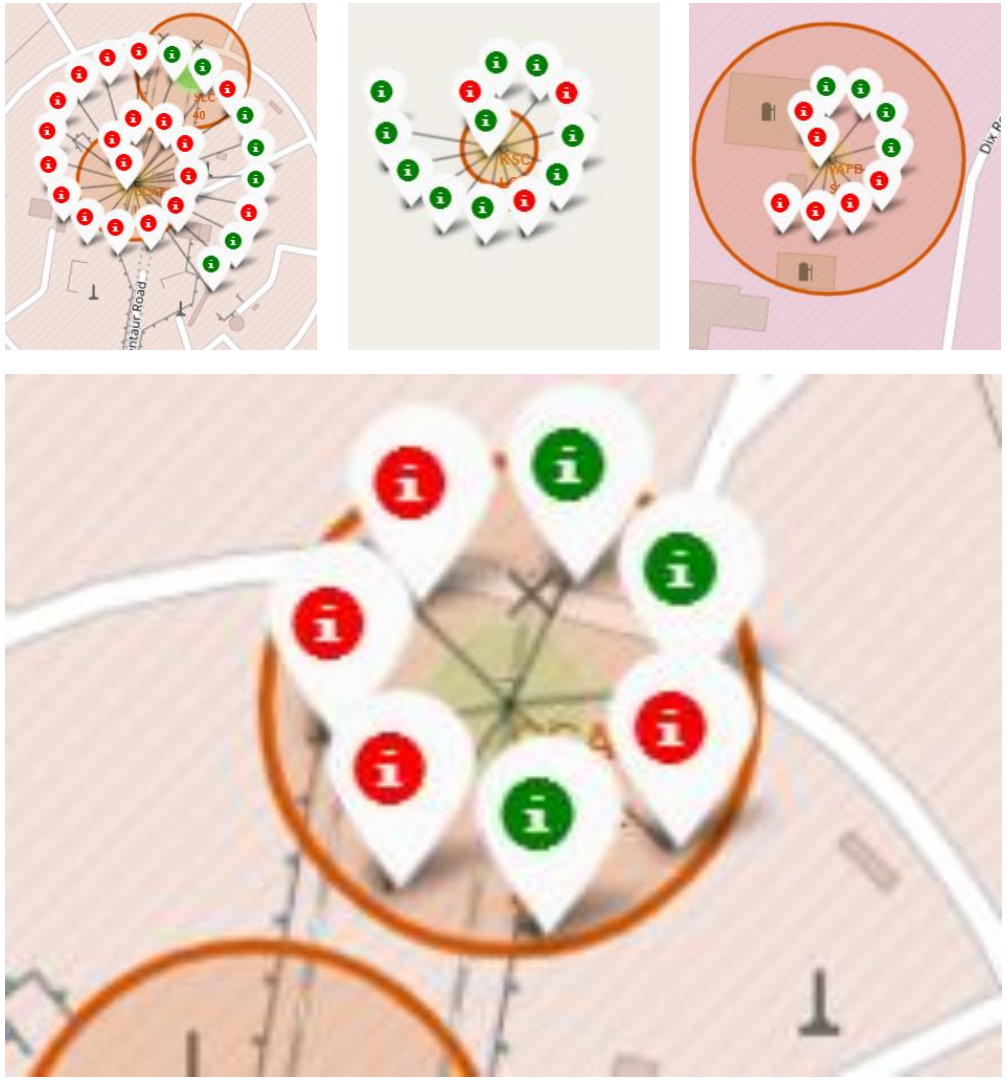
Launch Sites Proximities Analysis

Launch Sites in the Folium Map

- As it can be seen from that figure all 4 of the launch sites are located near the ocean.

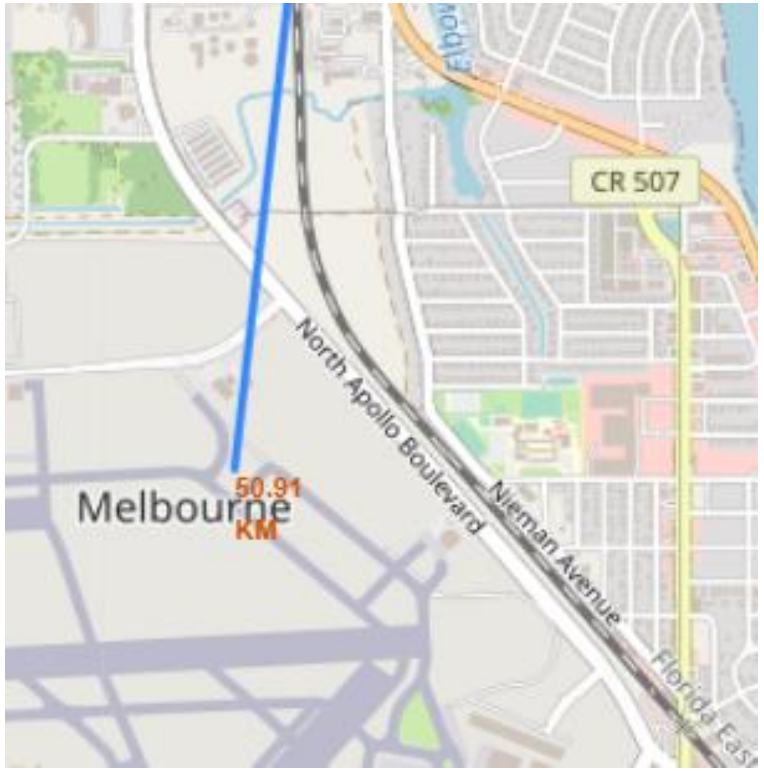
Marker Clusters for Mission Outcomes

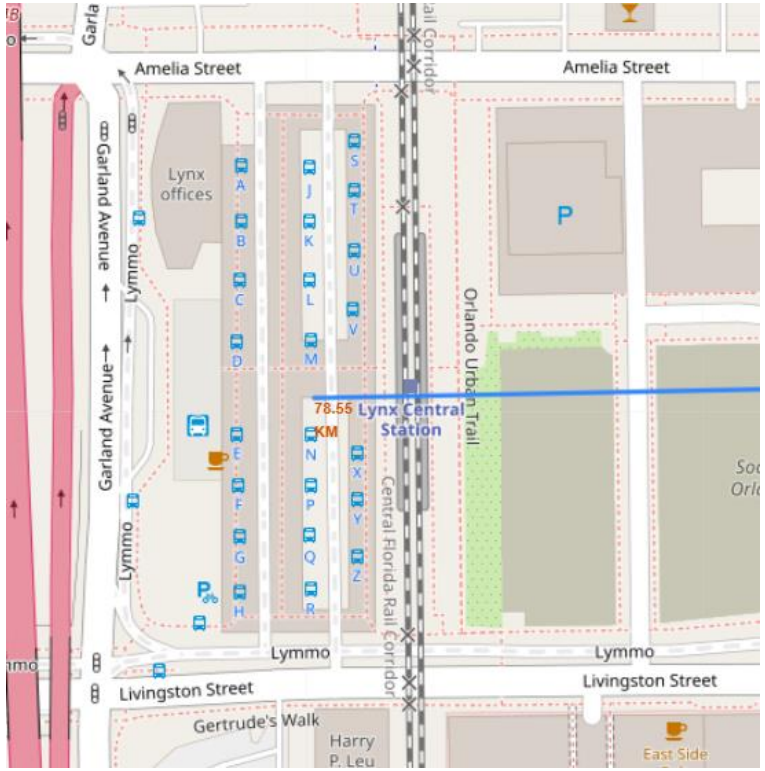
- Marker Clusters indicating mission outcomes for the 4 distinct launch sites are shown in the figures.
- Green indicates success whereas red indicates failure.

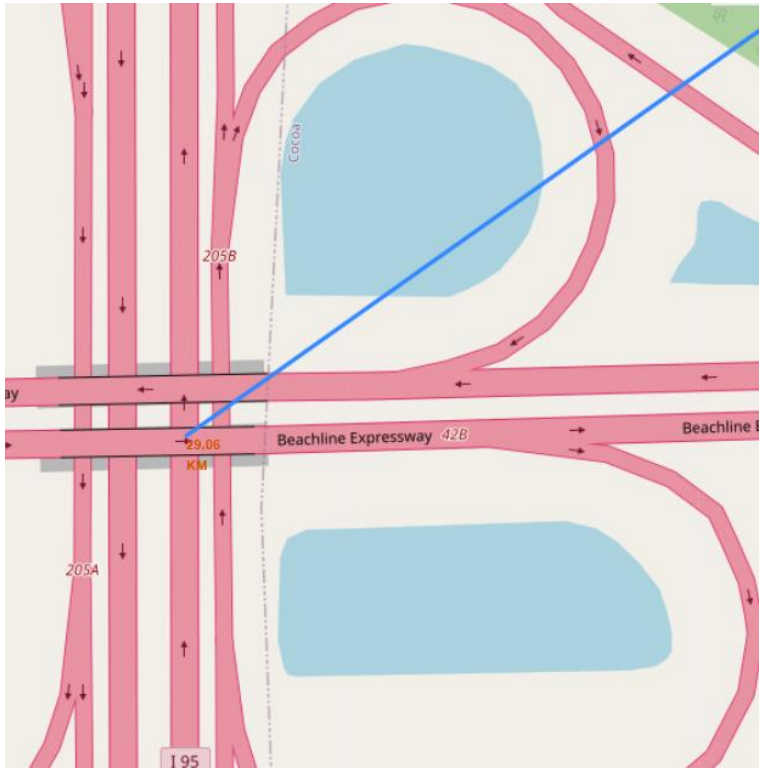


Launch Site Proximity to Highways, Railways and Cities

- From the figures, the closest Highway is Beachline Expressway which is 29 KM out from the launch site.
- The closest Railway Station is in Orlando, Lynx which is located 78.55 KM out from the launch site.
- The closest city center is in Melbourne which is more than 50 KM.





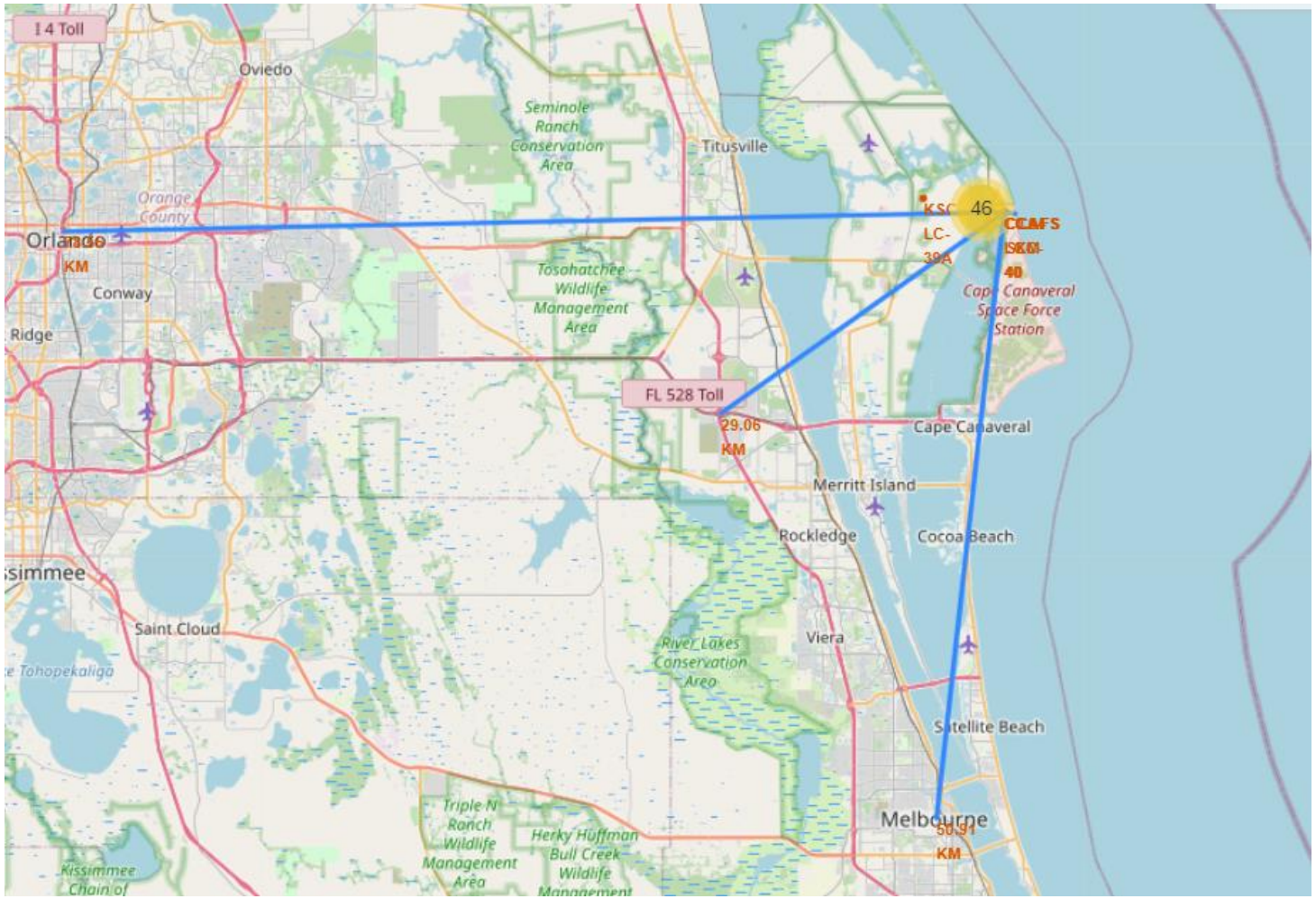


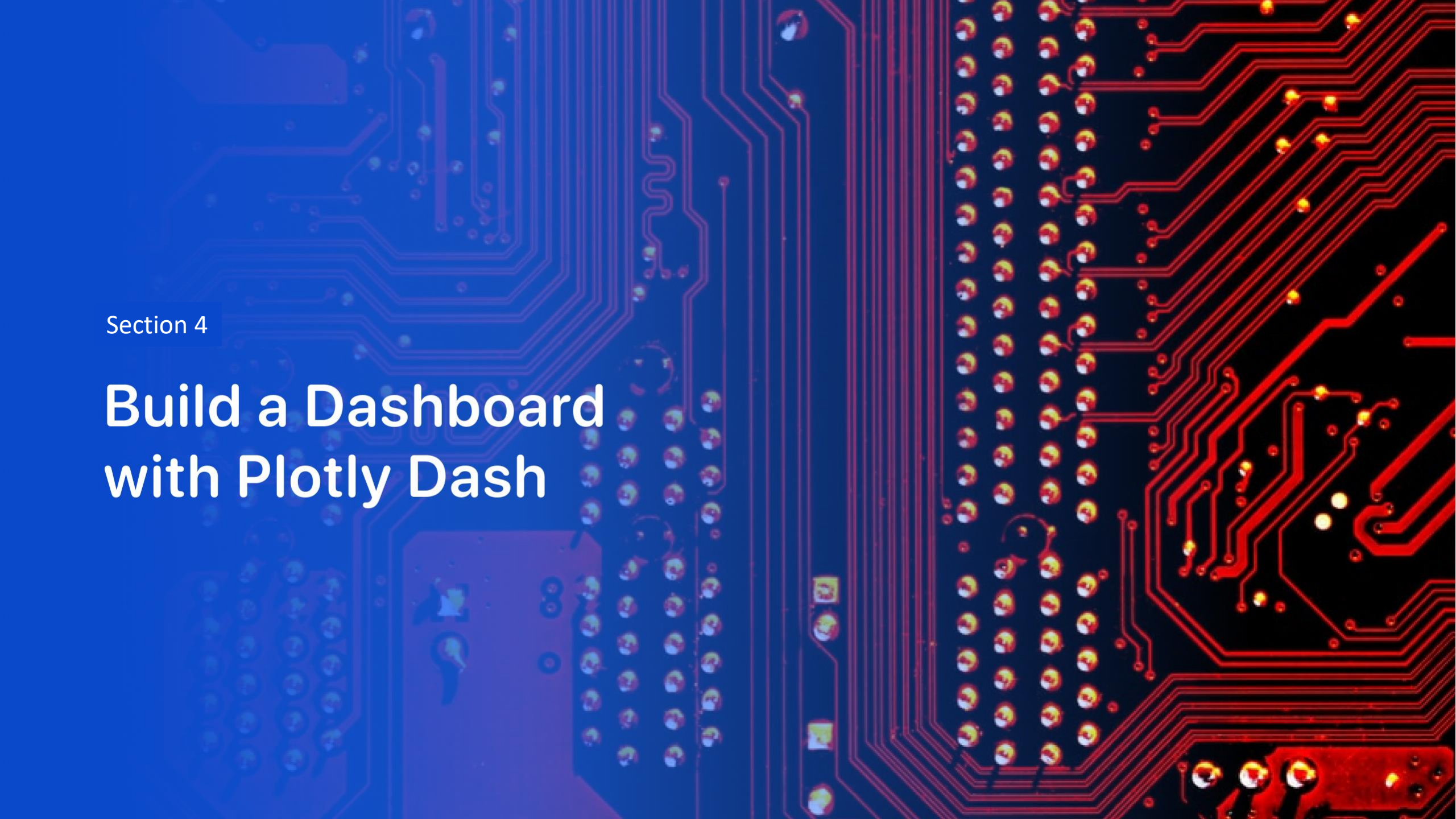
Launch Site Proximity to the Coastlines

- Launching over a large body of water is important for safety, as it reduces the risk of a falling rocket part hitting someone on the ground or damaging someone's property. It's the reason why US launches occur in coastal areas, such as Cape Canaveral, Florida, or the Vandenberg Air Force Base in California.

Launch Site a General Look to the Big Picture

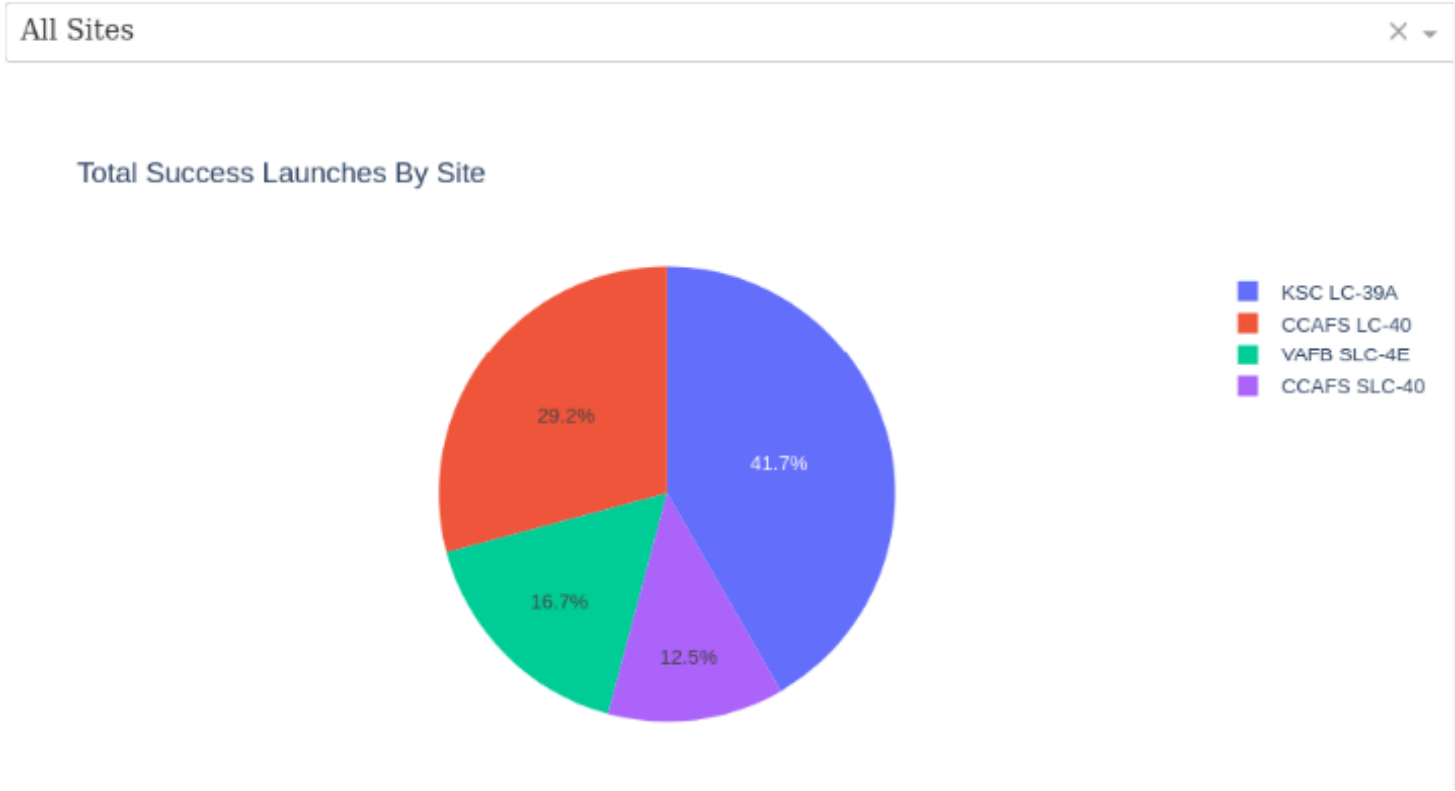
- In this figure it is demonstrated better where the location of a launch site is and how far it is to important places. This is reasonable concerning the safety issues.




Section 4
Build a Dashboard with Plotly Dash

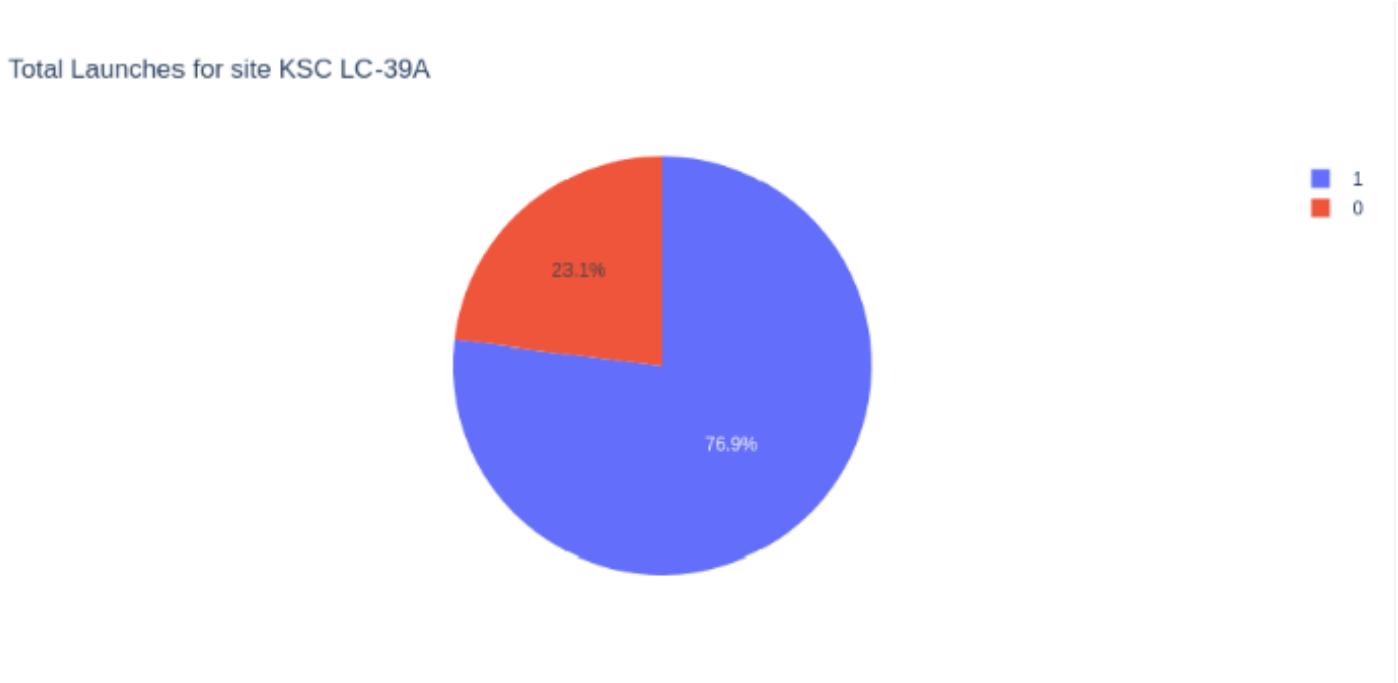
Ratio of Successive Launches Over Launch Sites

- Most of the successful launches are launched from KSC LC-39A.
- Followed by CCAFS LC-40.
- Other 2 launch sites have a little contribution to the overall successive launches.



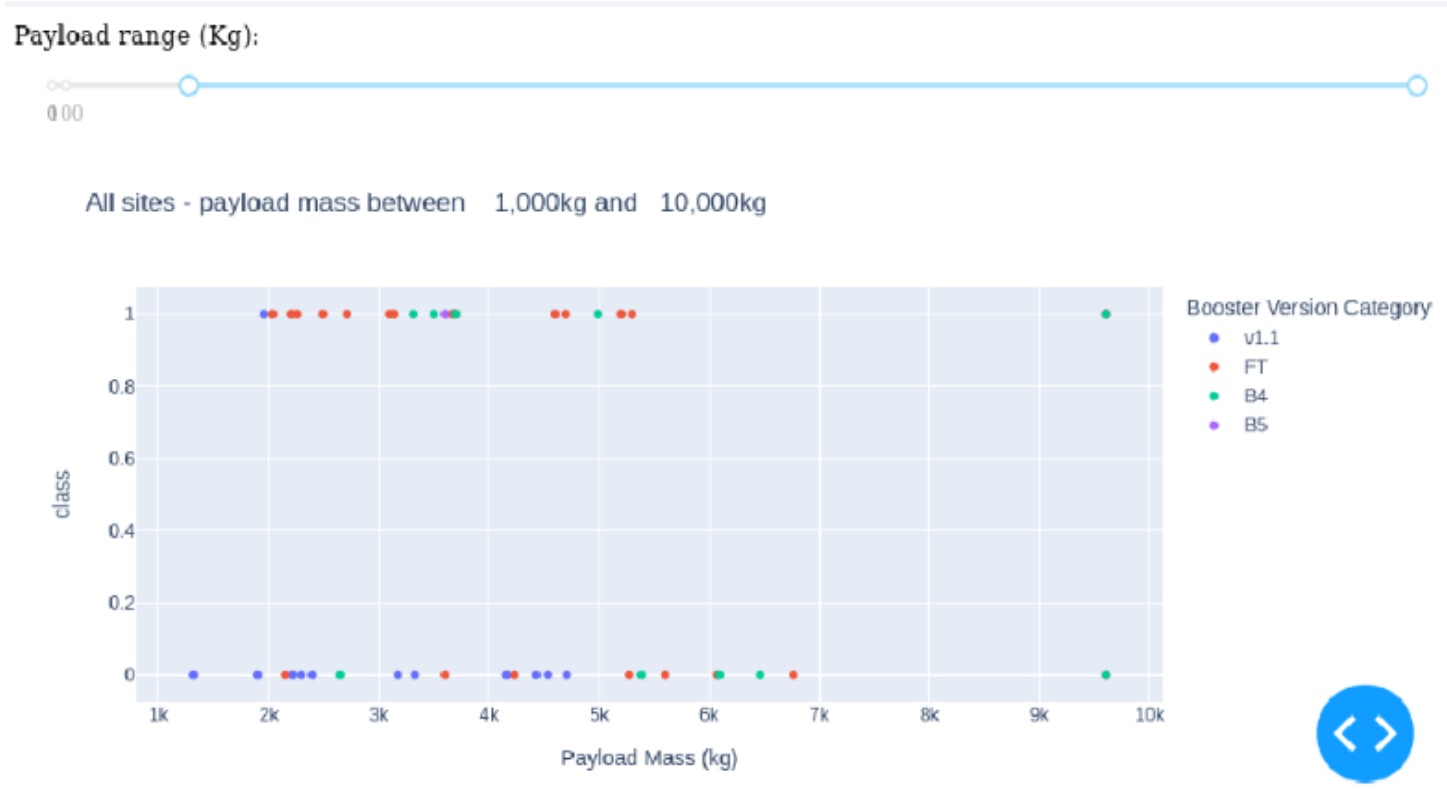
Launch Site with the Highest launch success ratio

- Most successful launch success ratio belongs to KSC LC-39A with a ratio of 76.9%.



Scatterplot of Payload vs. Launch Outcome for all sites

- Payloads under 6 tons are seem to be successful.
- With a combination of FT Boosters it is even better.

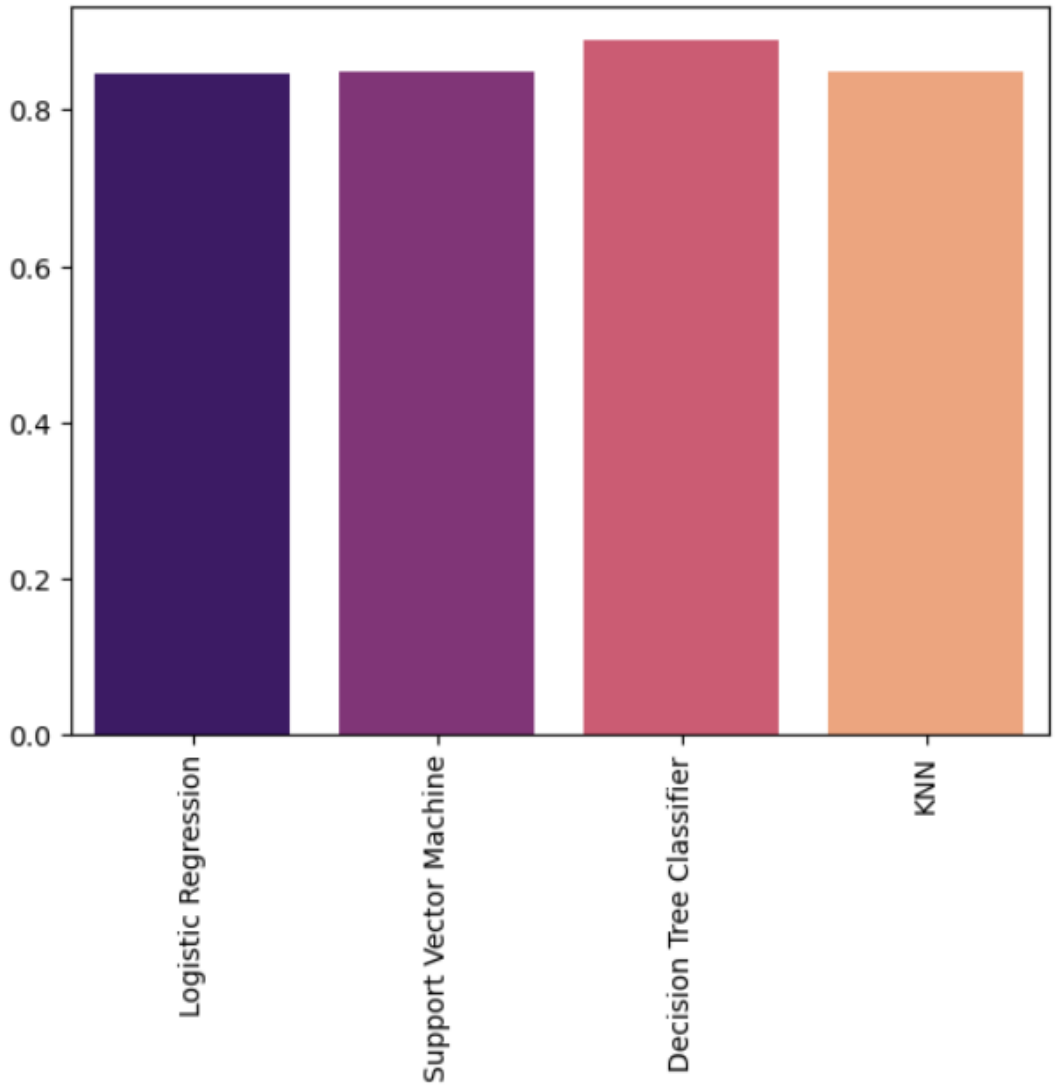


Section 5

Predictive Analysis (Classification)

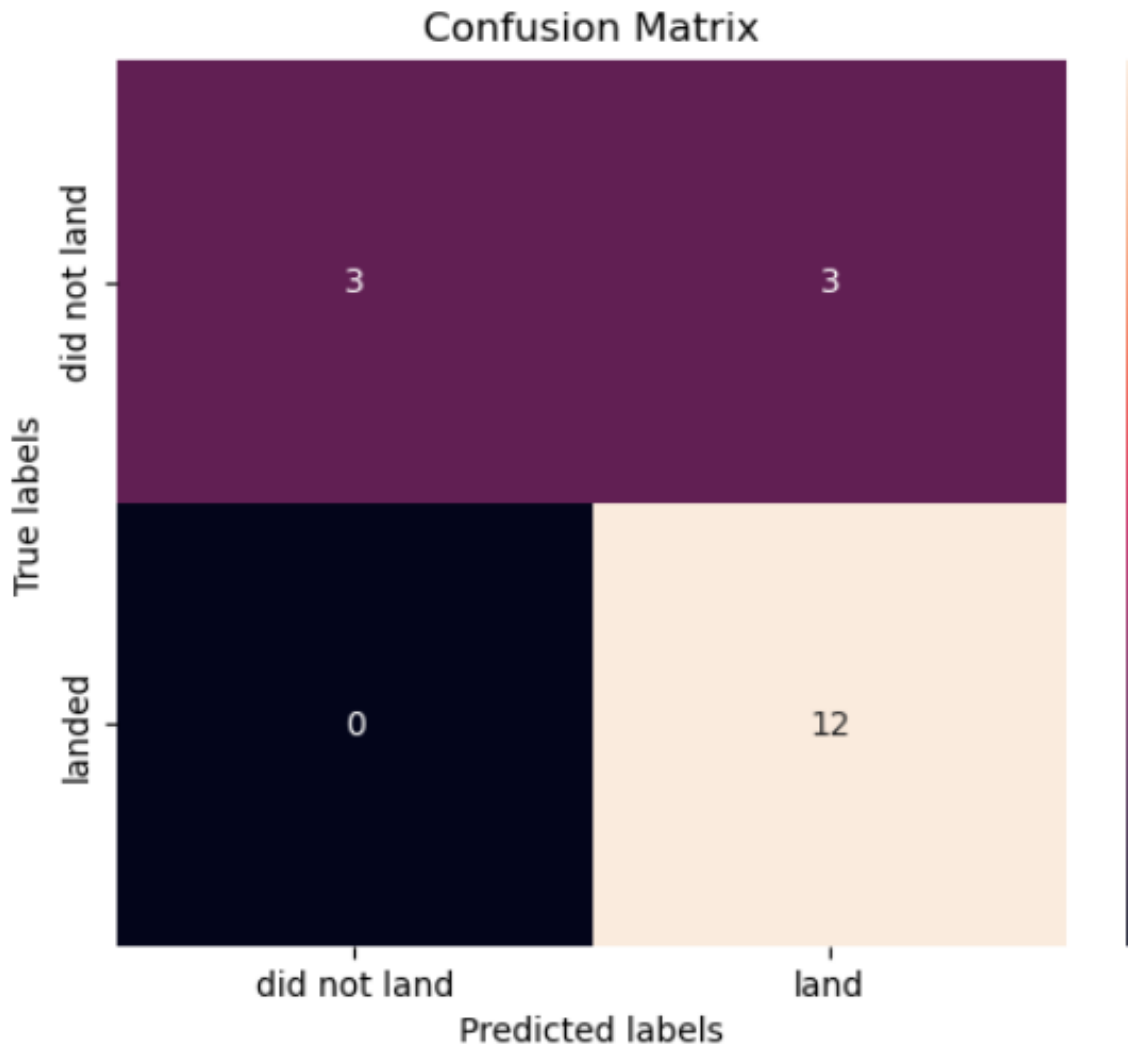
Classification Accuracy

- Decision Tree Classifier is the highest accuracy model therefore it has been selected as the algorithm to use in this project.



Confusion Matrix

- The confusion matrix for decision tree classifier is shown here. As it can be seen from the figure, the major problem is false negatives (Type II).



Conclusions

- Decision Tree Classifier is the best algorithm for our purpose.
- Most successful launch site is KSC LC-39A
- Although there is not enough data to decide, it seems that launches above 7000 kg are less risky.
- Success improves significantly over time.
- Orbits ES-L1, GEO, HEO, SSO and VLEO has the most success rate.

Thank you!