IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Damilola Ibikunle

27th June 2023

Outline

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

Executive Summary

- Summary of methodologies are:
  - Data collected from Space X API and Web scraping.
  - Data Wrangling and processing
  - Exploratory analysis using SQL magic and Python visualization tools.
  - Data analytics using Interactive maps and dashboards
  - Predictive analysis using Machine Learning Classification algorithms.
- Summary of all results:
  - The results from Exploratory Data Analysis
  - Results from interactive dashboards and maps
  - Model evaluation results and interpretation.
  - Determining the best classification model for success prediction

Introduction

- Space X rockets are popularly known for their reusable technology and successful vertical landings.
- Space X Falcon 9 rockets have a successful track record, launching payloads into space with high reliability and efficiency although occasionally, there are instances when it may experience crashes.
- This Applied data science Capstone project for Space X Falcon 9 rockets is aimed at predicting the successful outcomes of the rocket launches for the first stage using classification models.
- The objective is to predict whether or not the first stage would be reused.

Introduction

- Problems to find answers to:
  - Does the size of the payload affect the outcome of the rocket launch?
  - Does the location of the launch affect the outcome of the rocket launch?
  - Does the orbit affect the outcome of the rocket launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describes how data was collected using Space X API and web scraping from Wikipedia website
- Perform data wrangling
  - Describes how the data was processed by finding and replacing missing values and also converting columns with categorical values to numerical
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Describes how to build, tune, and evaluate classification models

Data Collection

- This process describes how Space X Falcon 9 rocket data was collected through different processes.
  - Data collection using Space X API involved:
    - Using the request.get() method on the Space X API URL
    - Decoding response content using .json() method
    - Converting the data to a Pandas data frame using .json_normalize()
    - Extracting further information such as 'booster name', 'payloads', 'launch site', 'landing outcome', 'flight number’, and 'date' using other Space X APIs
    - Filtering the data frame to only include Falcon 9 launches and replacing missing values with the mean values

Data Collection

- Data collection using Web scraping involved:
  - import requests, BeautifulSoup and Pandas library
  - Request data from Wikipedia URL using requests.get() method
  - Using BeautifulSoup() constructor to create a beautiful soup object on the HTML response data
  - Extracting all tables from the Soup object using soup.find_all()
  - Retrieving the third table and Iterate through to extract header/column names
  - Creating a data frame by iterating through each of the HTML tables to extract 'Flight No’, 'Launch site’, 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Version Booster', 'Booster landing', 'Date', 'Time'.

Data Collection – SpaceX API

- The flow chart in Figure 1 shows the data collection process using Python’s library requests and Space X API.
- The final data frame is saved as a CSV file using df.to_csv()
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Data%20collected%20by%20APIs/jupyter-labs-spacex-data-collection-api.ipynb

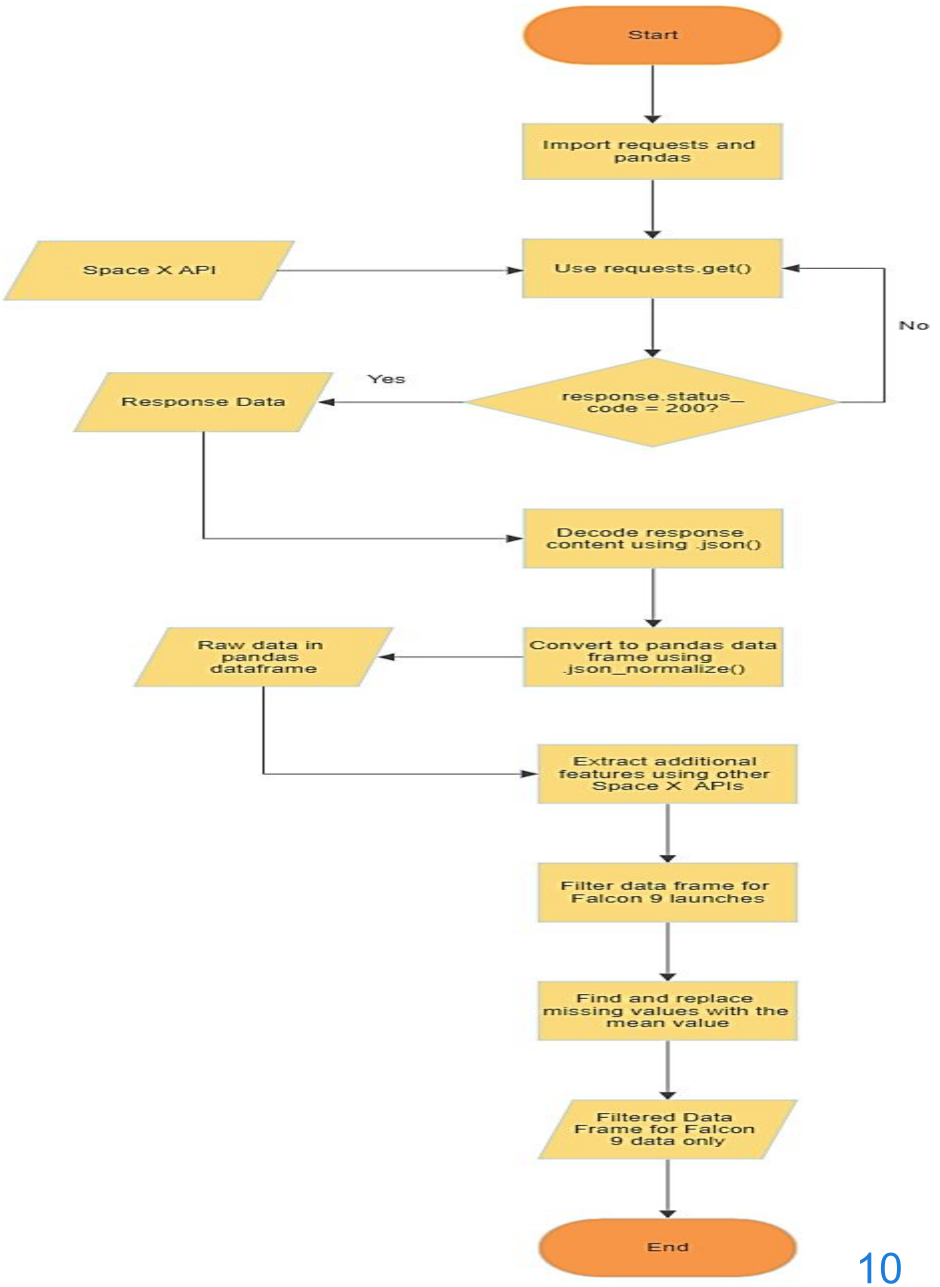

Figure 1. Flow chart of Data collection process using Space X API

Data Collection - Scraping

- The flow chart in Figure 2 shows the data collection process using Python's library requests and Beautiful Soup

- The final data frame is saved as a CSV file using data_falcon9.to_csv()

- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Web%20Scrapping/jupyter_labs_webscraping.ipynb

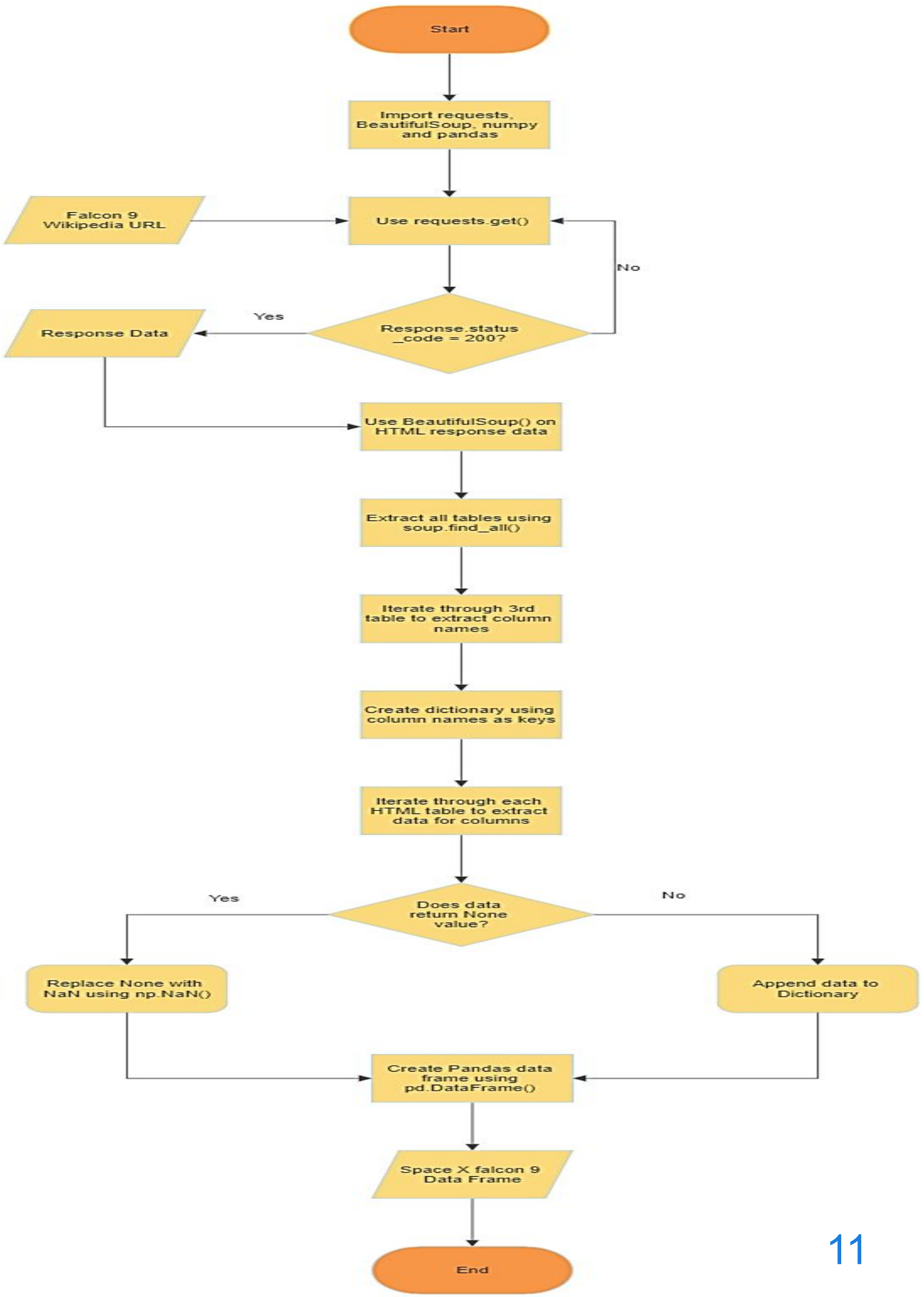

Figure 2. Flow chart of Data collection process using web

Data Wrangling

- This involved conducting some Exploratory Data Analysis on the data
- Also, converting the outcomes/class into training labels with 1 meaning successfully landed and 0 otherwise.
- Figure 3 shows a flow chart of the data wrangling process
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

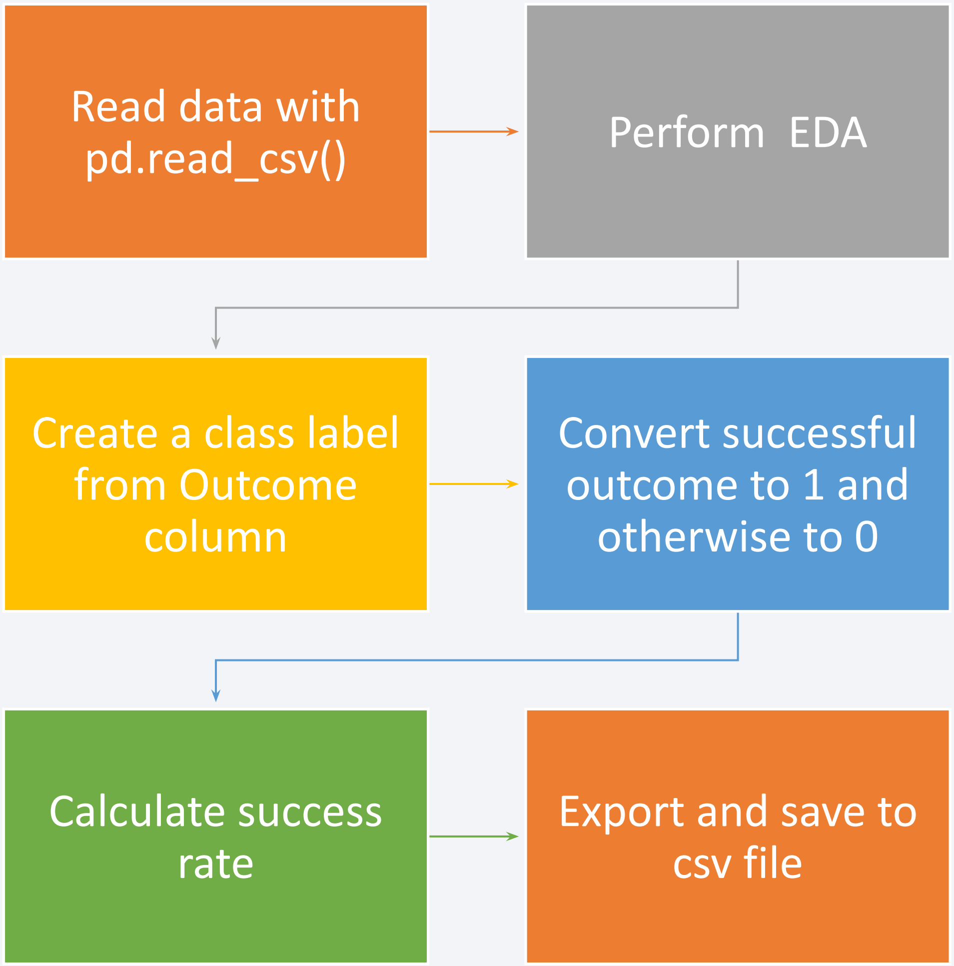

Figure 3. Flow chart of Data Wrangling process.

EDA with Data Visualization

- In the exploratory data analysis, we use scatter plots to visualize how some variables would affect the launch outcome and also the relationships between these variables e.g. Payload mass vs flight number, flight number vs launch site, and payload vs launch site.
- We also use a bar chart to visualize the success rate of each orbit type and scatter plots to visualize the relationship between flight number and orbit type, payload, and orbit type
- Lastly we use a line plot to visualize the yearly trend of the launch successes.
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/EDA%20with%20Python/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Exploratory data analysis conducted using SQL involved the following:
  - Display unique launch sites in the space mission
  - Explore launch sites that begin with ‘CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - Explore the first successful landing outcome in ground pad.
  - Display the successful boosters in drone ship and have a payload between 4000kg and 6000kg
  - Display the total number of successful and failed mission outcomes

EDA with SQL

- Exploratory data analysis (cont'd):
  - Display names of the booster versions that have carried the maximum payload mass
  - Display month names, failure landing outcomes in drone ship, booster versions, and launch sites for the months in the year 2015.
  - Rank the count of successful landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order.
- The URL below is the GIT repository containing the Jupyter notebook for EDA with SQL

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/EDA%20with%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Certain folium map objects were used such as:
  - folium.Circle used to add a highlighted circle area of NASA JSC as an initial centre location
  - folium.map.Marker used to create a marker at a specific launch location on the map
  - MarkerCluster() used to create cluster markers of successful and failed launches for a particular site
  - MousePosition() provides a way to display the latitude and longitude coordinates of the mouse cursor’s position on a map. Used to calculate the distance of the launch sites to the coasts.
  - Folium.PolyLine() is used to create a series of connected line segments on the map to mark the distance of the launch sites to the coast, railways, highways, and major cities

Build an Interactive Map with Folium

- The URL below is the GIT repository containing the Jupyter Notebook for Folium Map

  https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Folium%20Maps/lab_jupyter_launch_site_location.ipynb

- To view rendered Folium Map use the link below:

  https://nbviewer.org/github/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Folium%20Maps/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The plots and graphs added to the dashboard include:
  - Drop-down input containing all the launch sites.
  - A rendered pie chart showing the success rates based on the launch site input selected.
  - A range slider to select the payload mass.
  - A scatter plot showing the correlations between the payload mass and the success for the launch sites selected.
- The URL below is the GIT repository containing the Jupyter Notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Plotly%20Dashboard/dash_interactivity.ipynb

Predictive Analysis (Classification)

- The following are the steps taken in building and evaluating the classification models used.
  - The data was loaded and split into features and target
  - The features columns were normalized and the target column was converted to a NumPy array
  - Data was split into training and test set
  - GridsearchCV was used on all classification algorithms which helped in determining the best parameters and best scores using the .best_params_ and .best_score_ respectively
  - accuracy of the test set was also calculated using .score() method.
  - Lastly, a confusion matrix was plotted to visualize the results.

Predictive Analysis (Classification)

- Figure 4 shows a flow chart of building and evaluating the classification models
- The URL below is the GIT repository containing the Jupyter Notebook

https://github.com/Damiieibikun/Data-Science-Capstone-Project-IBM---Space-X-Falcon-9-landing-predictions/blob/main/Predictive%20Analysis%20using%20ML/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

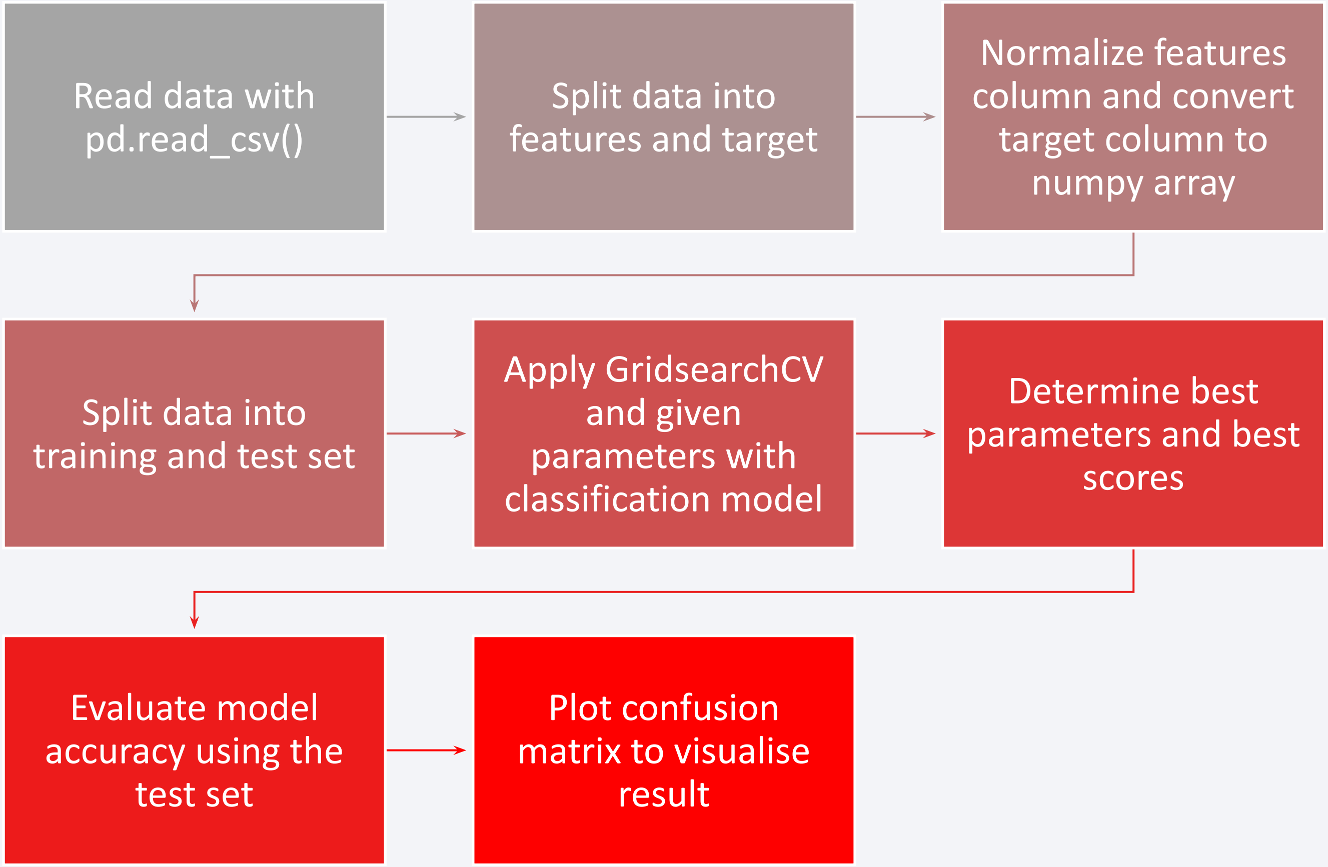

Figure 4. Flow chart of Model development and Evaluation process

Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- From the figure below, it can be noted that the launch site CCAFS SLC 40 has launched the highest number of rockets compared to the other sites.
- Also, it is shown that the later flights from launch sites VAFB SLC 4E and KSC LC 39A showed a higher success rate compared to the earlier flights.

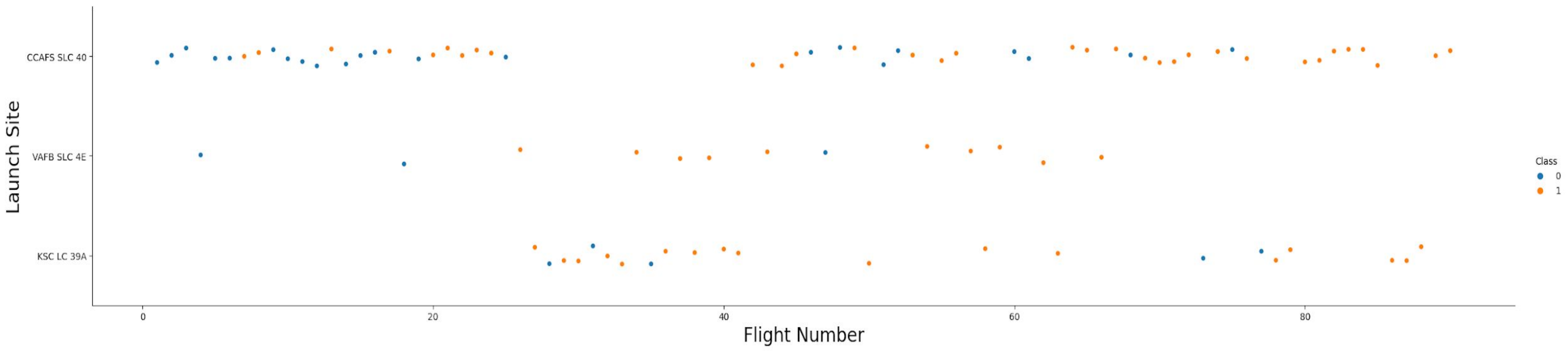

Figure 5. Scatter plot of Flight Number Vs. Launch Site

Payload vs. Launch Site

- From the figure below, it is observed that VAFB-SLC 4E launch site has no rockets launched for heavy payload mass greater than 10000kg.
- It is also observed that most of the rockets launched in all launch sites have a payload mass of less than 9000kg.
- Compared to VAFB-SLC 4E and KSC LC 39A, CCAFS SLC 40 has a higher success rate for rockets launched with a heavy payload mass of 14000kg and 16000kg.

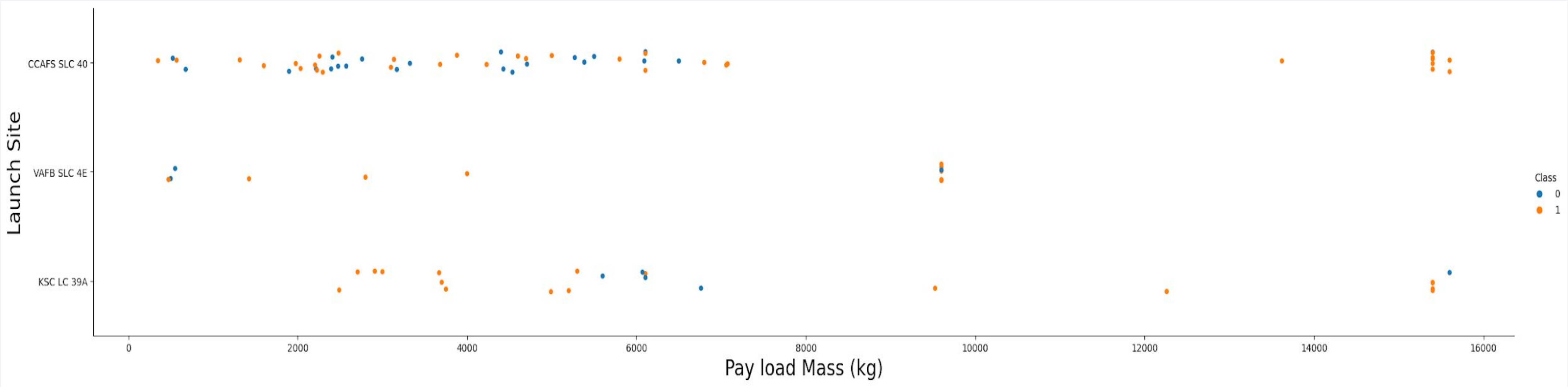

Figure 6. Scatter plot of Payload Vs. Launch Site

Success Rate vs. Orbit Type

- From the figure, it is observed that orbits VLEO, ES-L1, GEO, HEO, and SSO have the highest success rates compared to the other orbit types
- It is also observed that orbit SO has the least success rate.

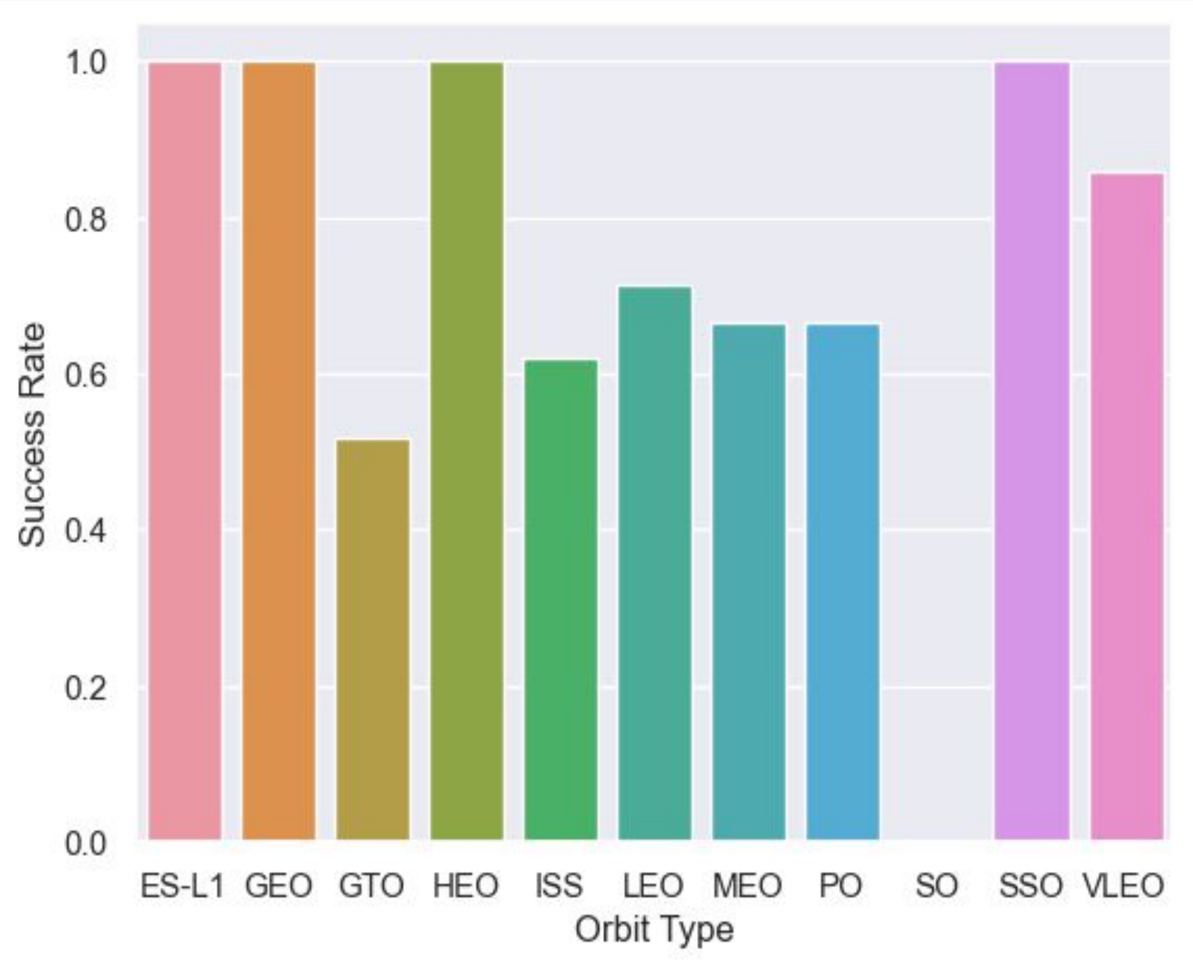

Figure 7. Bar plot of Success rate Vs. Orbit Type

Flight Number vs. Orbit Type

- From the figure below, it is observed that more rockets were launched in LES ISS PO GTO and VLEO
- It is also observed that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number and other orbit.

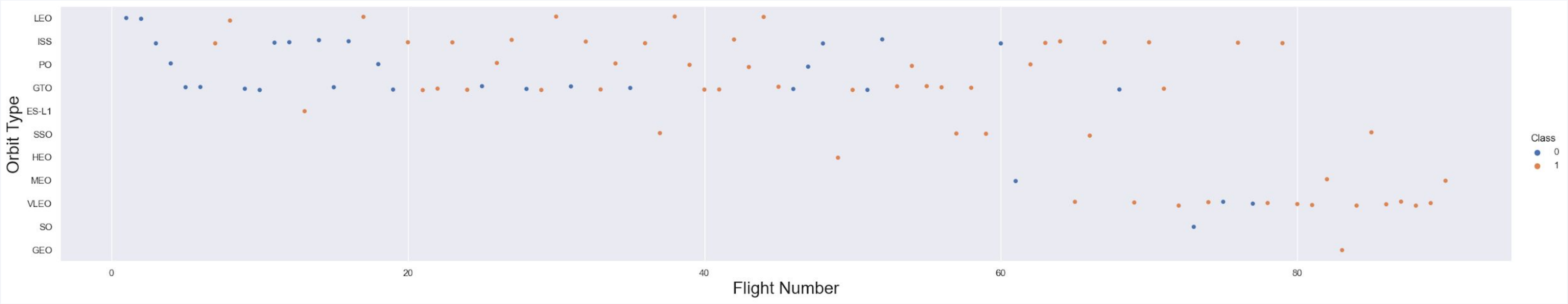

Figure 8. Scatter plot of Flight Number Vs. Orbit type

Payload vs. Orbit Type

- There is a higher success for rockets with heavy payloads launched in PO, LEO, and ISS.
- Rockets launched in SSO and MEO orbits on the other hand have a high success rate with lighter payloads
- Rockets launched in GTO have both positive landing rates and negative landing rates regardless of the size of the payload.

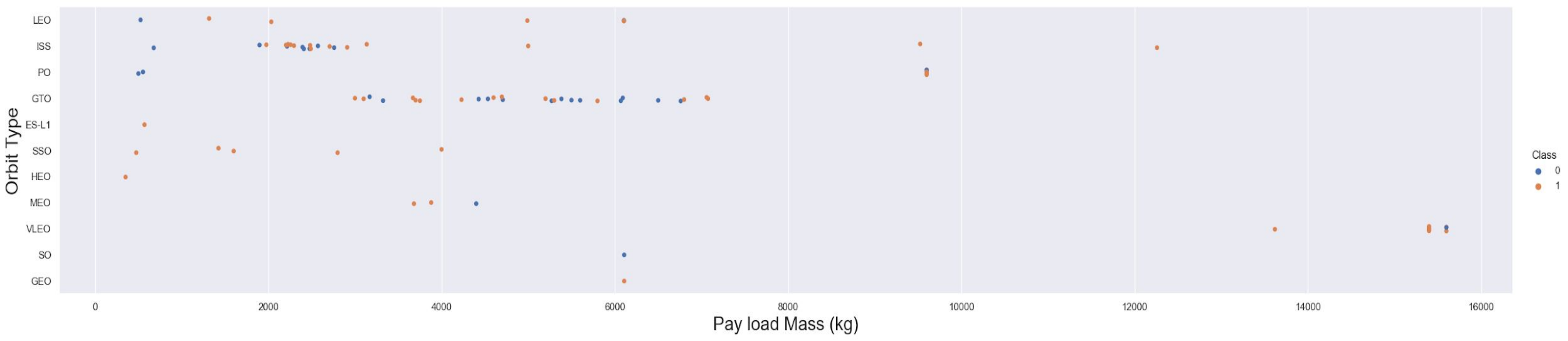

Figure 9. Scatter plot of Payload Vs. Orbit type

Launch Success Yearly Trend

- It is observed in Figure 10 that the success rate since 2013 kept increasing till 2020

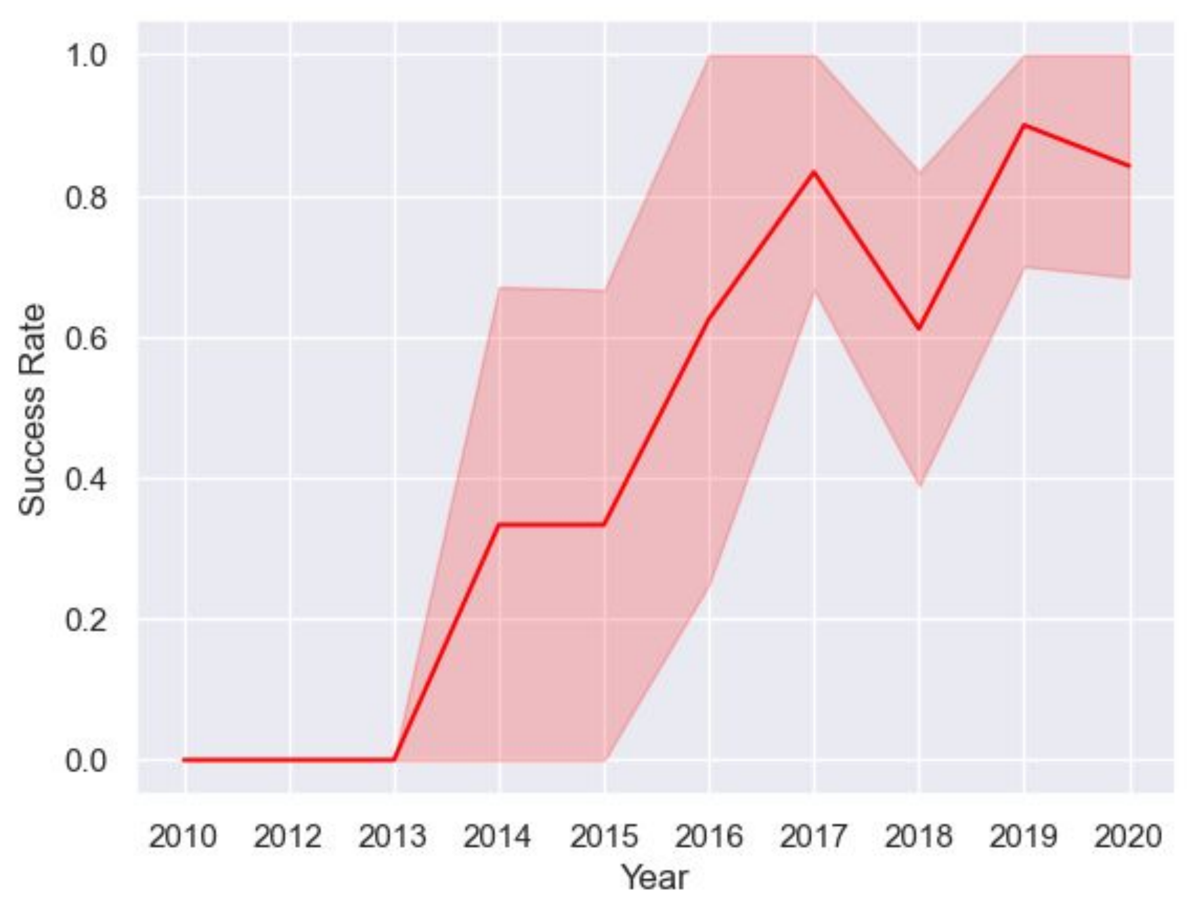

Figure 10. Line plot of Launch success Yearly Trend

All Launch Site Names

- An SQL table called SPACEXTBL using the existing data frame.
- To find the Unique Launch Sites, the keyword DISTINCT was used on the column.

```
%%sql

SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Figure 11. SQL query for unique launch site names

Launch Site Names Begin with 'CCA'

- Keyword LIKE `CCA%` was used to get launch site names beginning with `CCA`.
- LIMIT 5 keyword was used to display only 5 records.

```
%%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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Figure 12. SQL query for Launch site names beginning with CCA

Total Payload Mass

- SUM function was used to calculate the total payload mass of customers with the name ‘NASA (CRS)’.

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_payload_NASA_CRS
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

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

Total_payload_NASA_CRS
45596.0

Figure 13. SQL query for Total payload Mass for NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using the AVG function.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_F9V1_1
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

Average_Payload_F9V1_1
2534.6666666666665

Figure 14. SQL query for Average payload Mass for F9 V1.1

First Successful Ground Landing Date

- SQL query was run for the first successful landing on ground pad.
- The result shows that 22nd of December 2015 was the date for the first successful ground landing.

```
%%sql

SELECT Date FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'
ORDER BY Date DESC
LIMIT 1
```

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

Date
22/12/2015

Figure 15. SQL query for First successful ground landing date

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using keywords BETWEEN and AND, the names of boosters that have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg were displayed.
- The result shows 4 rockets.

```
%%sql
SELECT Booster_Version FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

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

Figure 16. SQL query for successful drone ship landing between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- The COUNT function was used to could the total number of successful missions and failed missions
- The results show that there were 100 successful missions and 1 failed mission.

```
%%sql
SELECT COUNT(Mission_Outcome)
AS Success_missions
FROM SPACEXTBL
WHERE Mission_Outcome LIKE '%Success%'
```

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

Success_missions
100

```
%%sql
SELECT COUNT(Mission_Outcome)
AS Failure_missions
FROM SPACEXTBL
WHERE Mission_Outcome LIKE '%Failure%'
```

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

Failure_missions
1

Figure 17. SQL query for successful drone ship landing between 4000 and 6000.

Boosters Carried Maximum Payload

- A sub-query with the MAX function was used to retrieve the boosters that carried the maximum payload.
- Results show that there are 12 in total.

```
%%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.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

Figure 18. SQL query for boosters that carried the maximum payload

2015 Launch Records

- Substr() was used to extract the month and year from the Date column.
- The WHERE and AND keyword was used to get launch records of failed drone ship landings in 2015.
- The result shows that the failed landings occurred in the months of April (04) and October (10).

```
%%sql

SELECT substr(Date, 4, 2) as Month, Booster_Version, Landing_Outcome, Launch_Site
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 7, 4) = '2015'
```

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

Month	Booster_Version	Landing_Outcome	Launch_Site
10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

Figure 19. SQL query for failed drone ship landings in 2015

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

- Keywords such as GROUP BY, ORDER BY, and DESC as well as functions like substr() and COUNT() were used to rank the count landing outcomes between 2010-06-04 and 2017-03-20, in descending order.
- The results show that there are high numbers for no attempt (10), success on drone ship (5) as well as ground (5)
- Also one instance where there was a failure in parachute deployment

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Number
FROM SPACEXTBL
WHERE substr(Date,7)||substr(Date,4,2)||substr(Date,1,2)
BETWEEN '20100604' and '20170320'
GROUP BY Landing_Outcome
ORDER BY Number
DESC
```

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

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

Figure 20. SQL query ranking landing outcomes between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

Launch Site Locations for Space X Falcon 9

- All launch sites as shown in the figure are located in coastal cities of the United States of America.

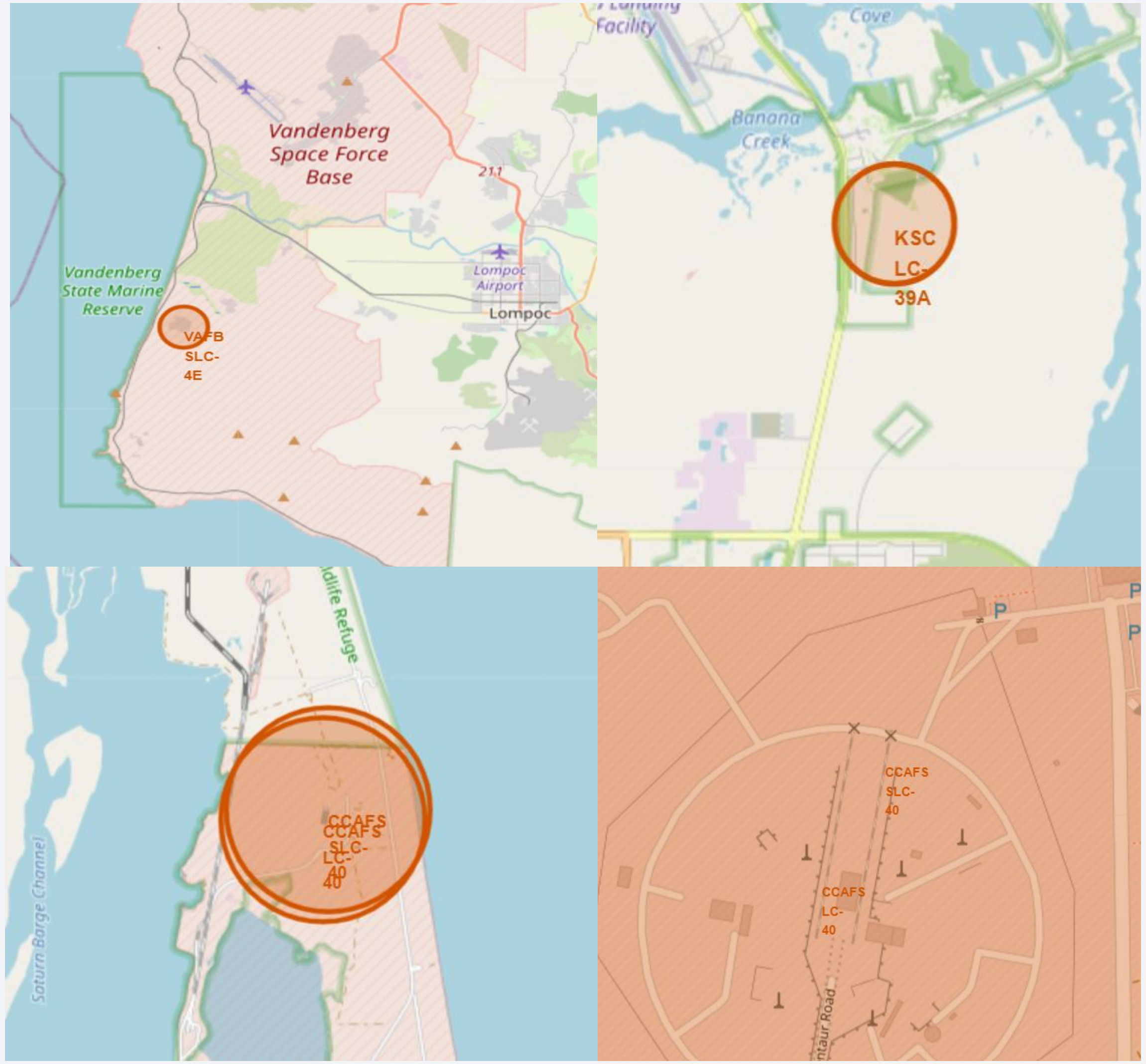

Figure 21. Folium Map showing locations of Space X Falcon 9 launch sites

Launch Outcomes for Space X Falcon 9

- The figure shows the launch outcomes for various launch sites;
  - Top left: VAFB SLC-4E
  - Top right: KSC LC-39A
  - Bottom left: CCAFS SLC-40
  - Bottom Right: CCAFS LC-40

- Red icons indicate the failed outcomes and the green icons indicate successful outcomes.

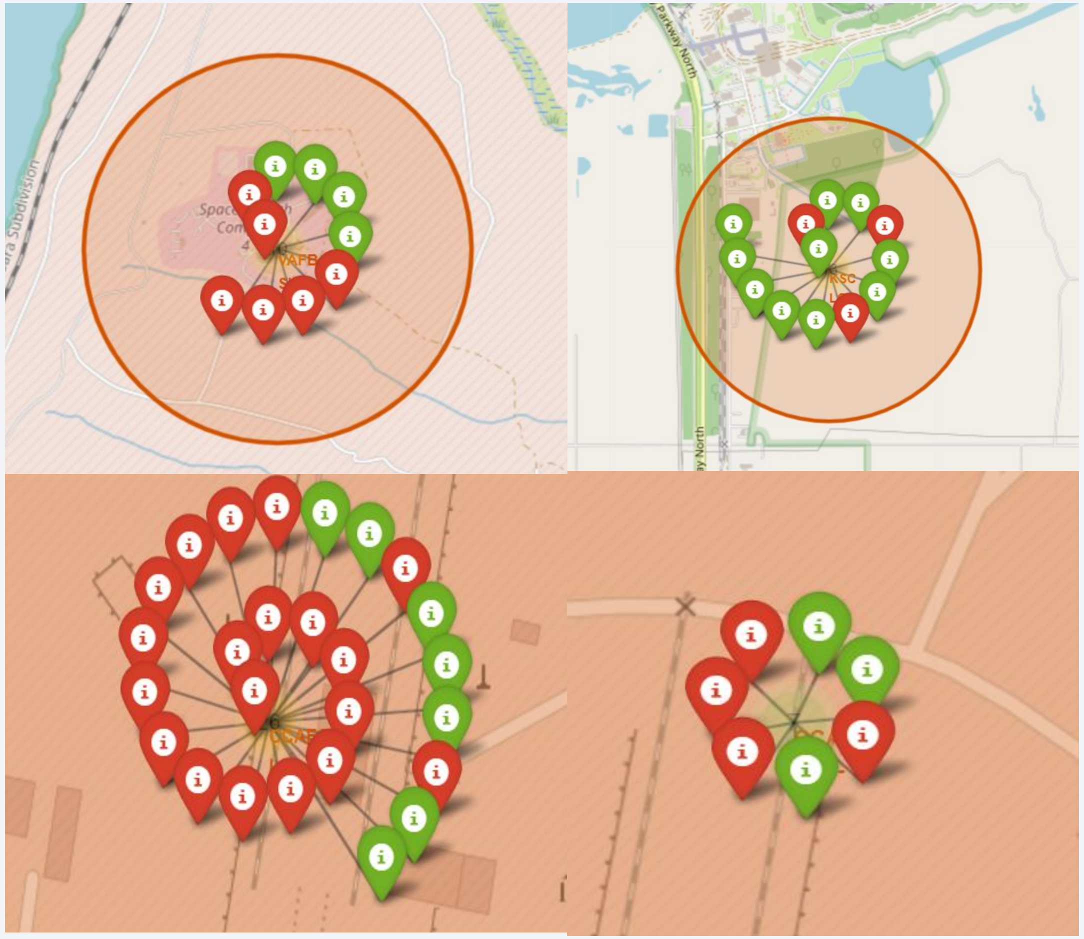

Figure 22. Folium Map showing Launch outcomes in their various launch sites

Launch site distance from coastline, cities, railways and highways.

- From the figure it is shown that launch sites are located very close to the coast i.e. 0.95km from CCAFS SLC 40 and 1.52km from VAFB SLC 4E

- The same can not be said for some railways and highways

- It is also evident that launch sites are located far from major cities, i.e. VAFB SLC 4E is 38.16km away from its closest city Santa Maria and CCAFS SLC 40 is 56.04km away from Melbourne

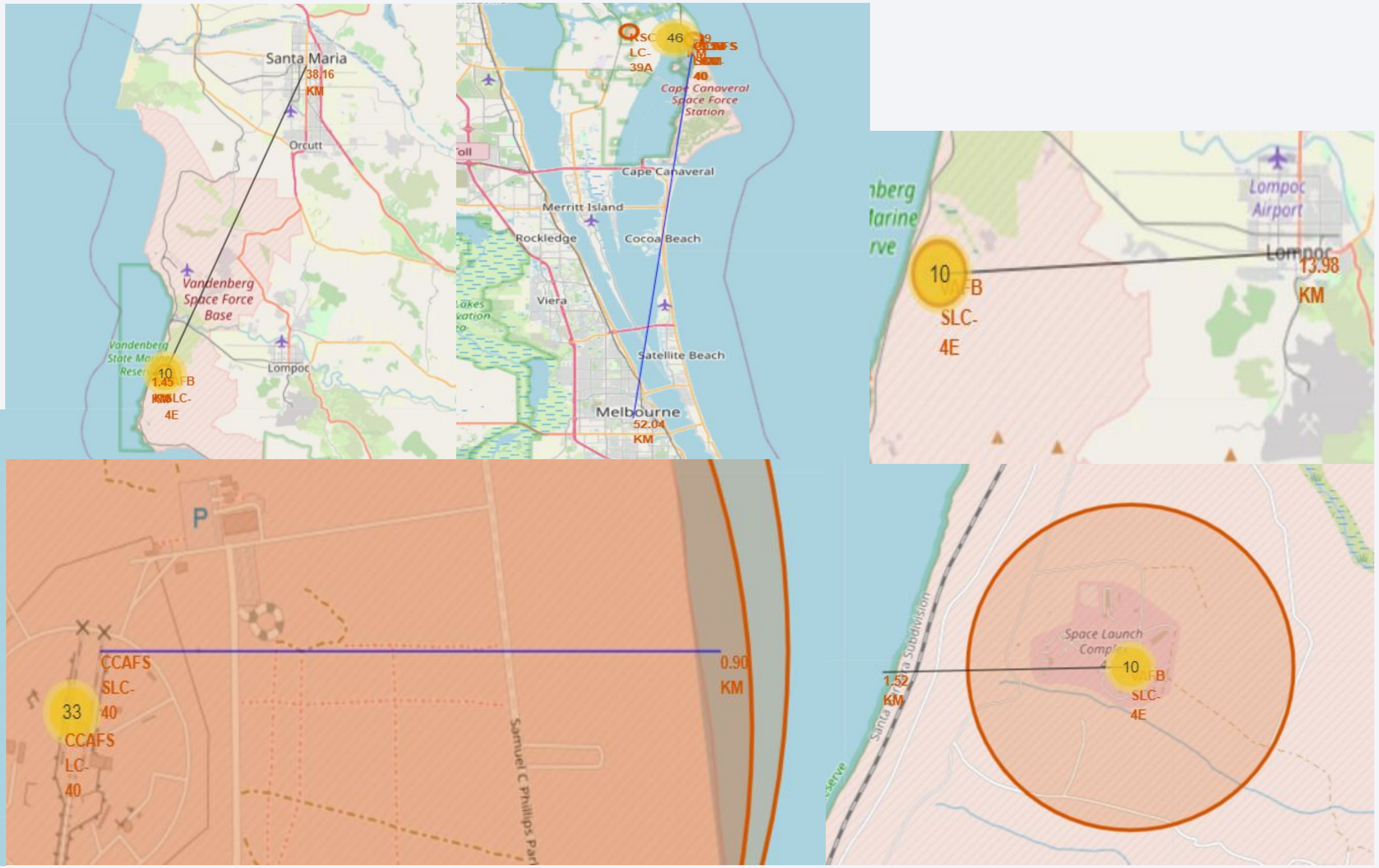

Figure 23. Folium Map showing Launch site distances from coastlines, cities, railways and highway

Section 4

Build a Dashboard with Plotly Dash

Pie Chart of Launch Success for all Sites

- From Figure 24, it is shown that KSC LC-39A has the largest success rate with about 41.7% of the total success ratio with other sites.

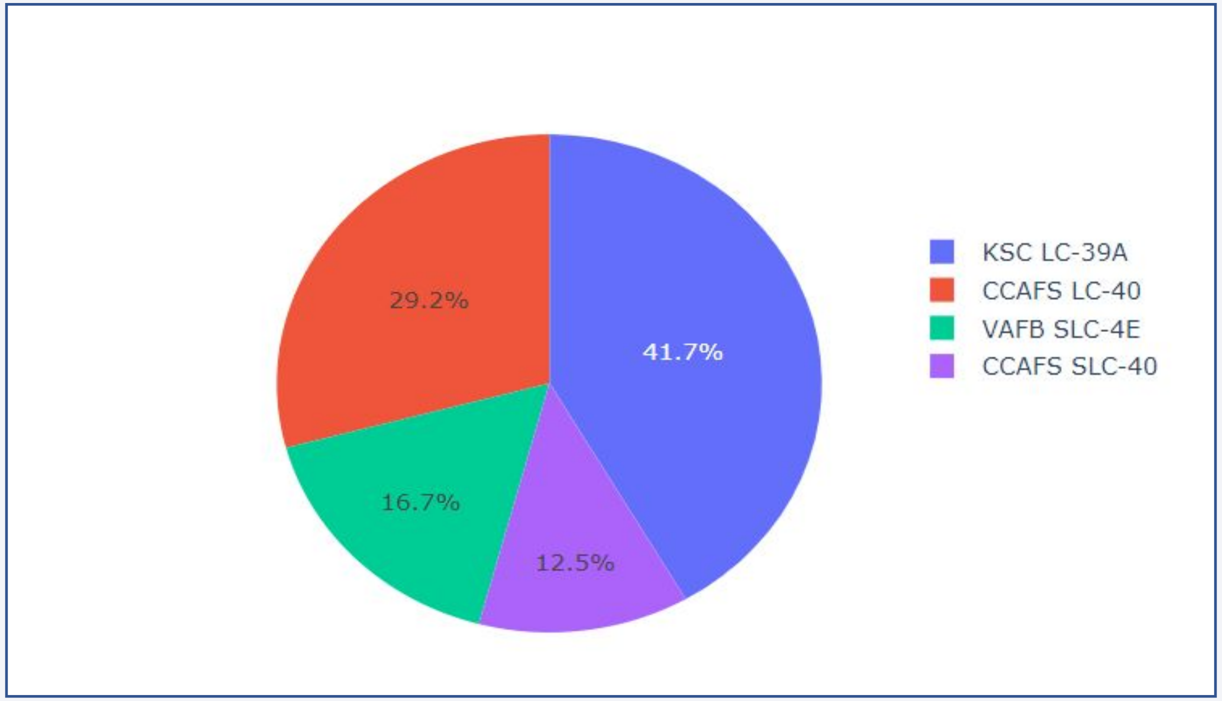

Figure 24. Pie showing the Success rate of all Launch sites

Pie chart of Launch site with highest success ratio

- It is also evident from Figure 25 that KSC LC-39A has the highest success ratio with about 76.9%, compared to the other sites;
  - 73.1% for CCAFS LC-40
  - 60% for VAFB SLC-4E
  - 57.1% for CCAFS SLC-40

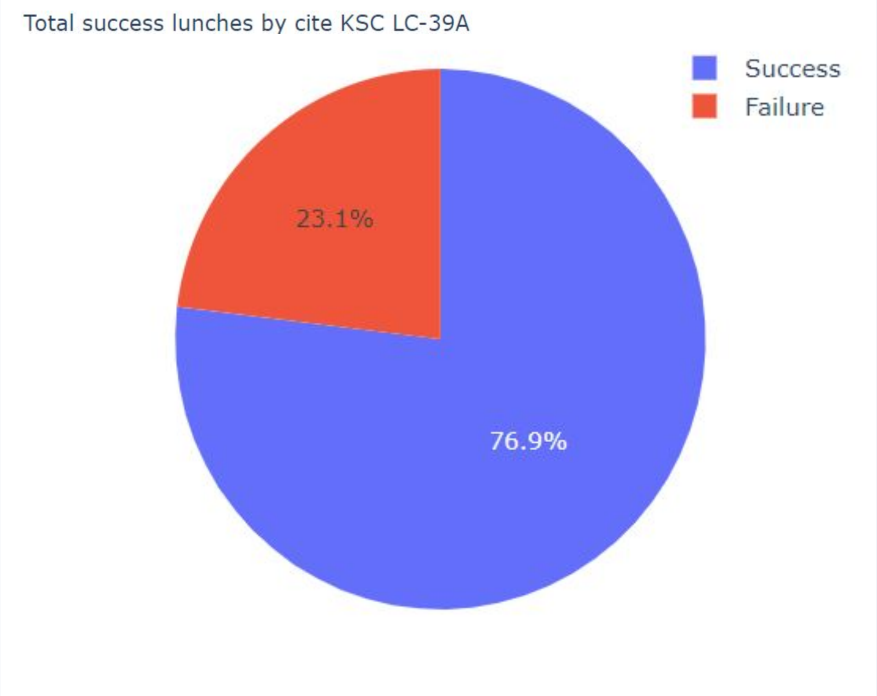

Figure 25. Pie showing the launch site with the highest success ratio

Payload vs Launch outcome for all sites

- From the figures below, Booster version FT has the highest success rate with its payload mass of about between 700kg to 5,500kg.
- It is also shown that rockets with payload mass above 5,500kg have a lower success rate, which means the heavier the payload, the slimmer the chance of a successful outcome.

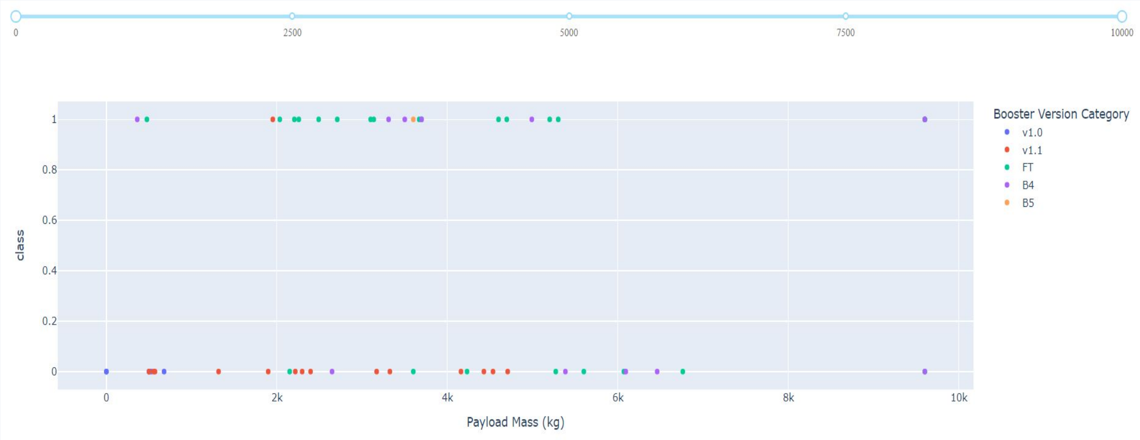

Figure 26. Scatter plot showing the booster versions with different payload mass for all launch sites

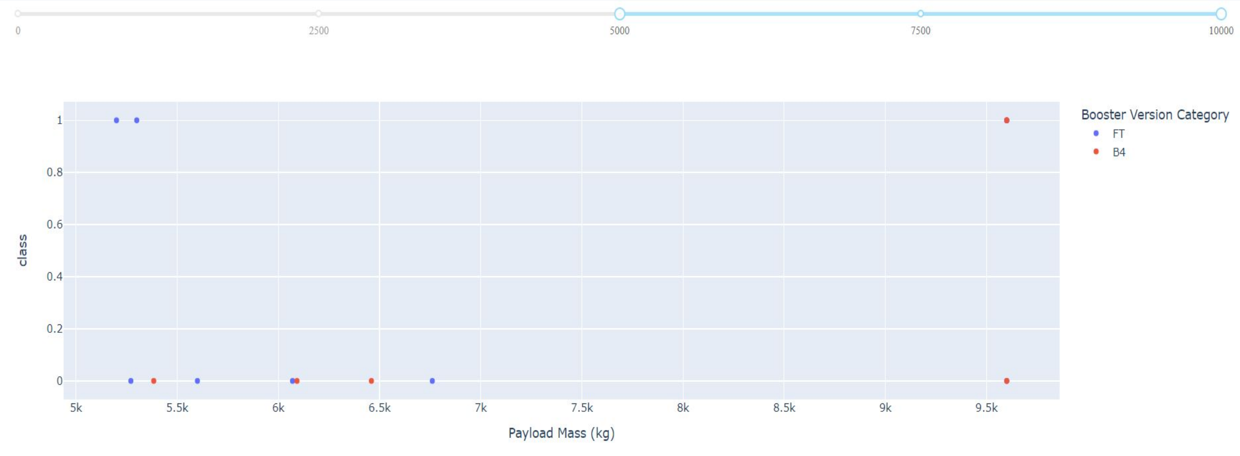

Figure 27. Scatter plot showing the booster versions of different payload mass greater than 5,500kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- From the bar chart in Figure 28, Decision tree classifier performed the best with an accuracy score of approximately 0.875 or ~ 86%.

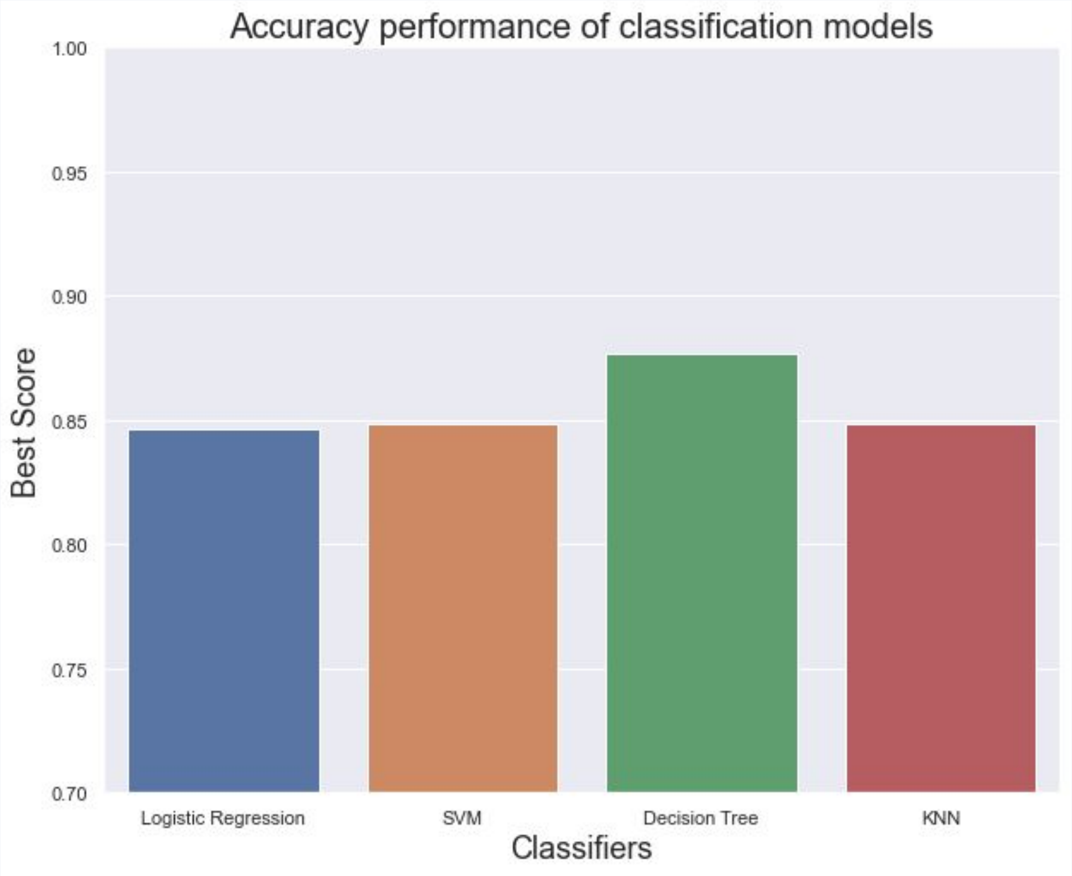

Figure 28. Bar plot showing different classifiers and their accuracies.

Confusion Matrix

- After the dataset was spilt into training and test set, we ended up with only 18 test samples.
- From the test set, the decision tree classifier was able to correctly predict 12 observations that landed (12 True positives) and also 3 observations that did not land (3 True Negatives).
- The classifier also had 0 false negatives because it did not wrongly predict any successful landings.
- However, it had 3 false positives as it predicted wrongly for 3 observations that the outcome was successful.

Figure 29. Confusion matrix of decision tree classifier.

Conclusions

- For a successful mission, the mass of the payload should be considered as rockets with smaller payload had a higher success rate.
- Orbit type should also be considered because rockets launched to certain orbits (VLEO, ES-L1, GEO, HEO, and SSO) had higher success rates compared to others.
- Launch sites are located in coastal cities for easy retrieval/recovery and far from busy areas like major highways and cities to minimize casualties in the event of a failure.
- In recent years the outcome has been more successful as later flight launches had a higher success rate.
- Compared to other classification algorithms, decision tree classifiers had the best performance of approximately 87% making it a good model for landing outcome prediction.

Thank you!