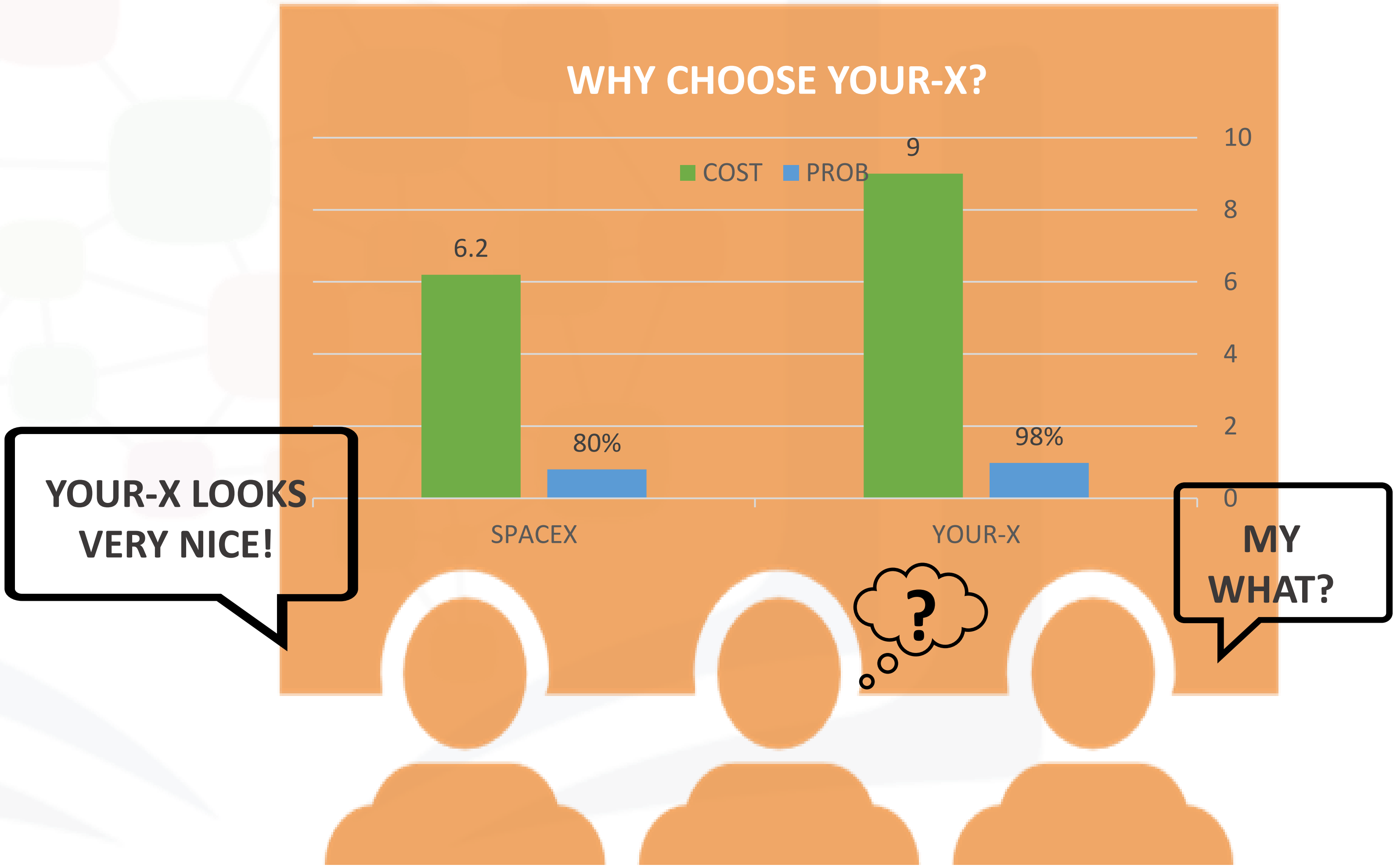
Estimating Rocket Launch Costs With Data

Theophilus Ankrah

15th June 2023

OUTLINE

- ❖Executive Summary
- ❖Introduction
- ❖Methodology
- ❖Results
  - ❖Visualization – Charts
  - ❖Dashboard
- ❖Discussion
- ❖Conclusion



EXECUTIVE SUMMARY

- This project aimed to predict the cost of rocket launches using SpaceX launch data. To better predict this, we needed to determine if the re-usable first stage rockets which cost around 100 million dollar less than conventional rockets can be used.
- In this respect, Space Y is a genuine rival, even a leader in commercial space exploration, actionable knowledge by figuring out the cost of each launch and if they will recycle the first stage.
- I employed several approaches to do this, including web scraping, data visualizations, dashboards to highlight key areas of interest, and data collecting using Space X's APIs. I also created machine learning algorithms to forecast launch results. This research' findings indicate that rockets with a payload between 2.5 and 5.5 kg are quite successful. These boosters are identical to the Falcon9 FT. Launch pad 39-A at KSC.
- Regarding predictive analysis, I observed that most models would be equally appropriate but settled on logistic regression models for this case as it also allows us to predict the probability of the of the prediction itself occurring.

INTRODUCTION

- ❖Project background and context
  - ❖Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- ❖Problems you want to find answers
  - ❖What factors determine if the rocket will land successfully?
  - ❖The interaction amongst various features that determine the success rate of a successful landing.
  - ❖What operating conditions needs to be in place to ensure a successful landing program.


METHODOLOGY

METHODOLOGY

DATA COLLECTION METHODOLOGY

PERFORM DATA WRANGLING

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 Collection – SpaceX API

SpaceX rocket data was collected using the SPACEX REST_API link:

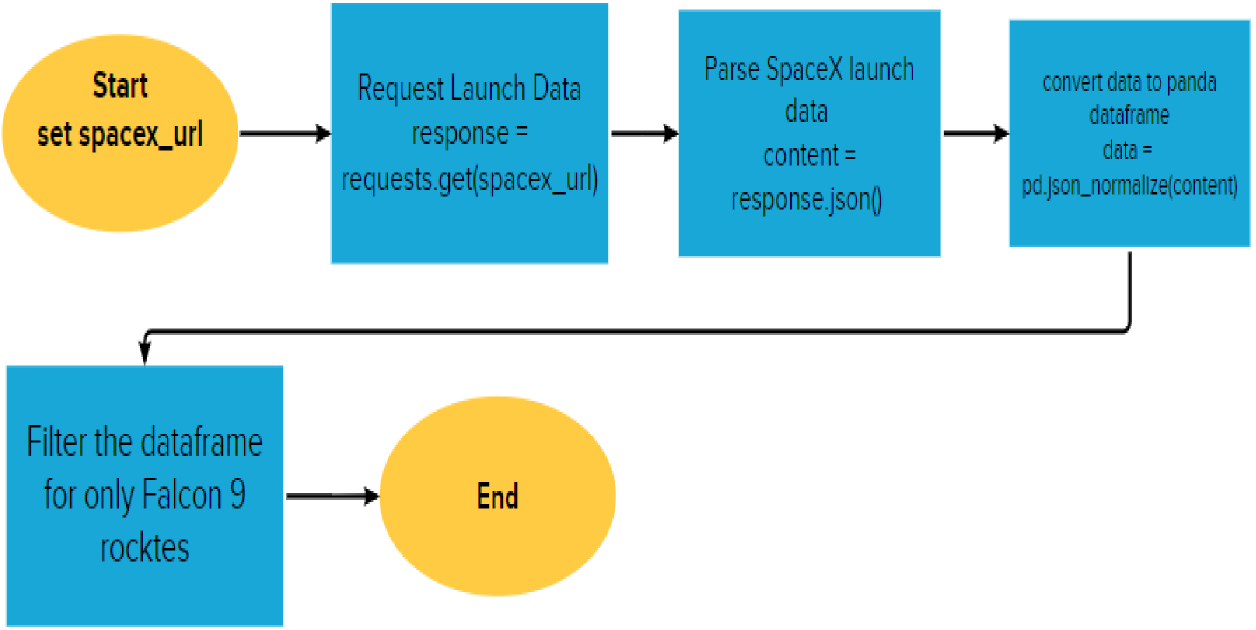
➢ https://api.spacexdata.com.v4/launches/past

STEPS:

❖ SpaceX API was selected as the primary data source – Achieved using get requests.

❖ Data was collected converted in a pandas dataframe. – Achieved by normalizing json formatting

❖ Collected data was cleaned, checked for missing values and replaced them. - Using pandas



GitHub URL to notebook detailing steps:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

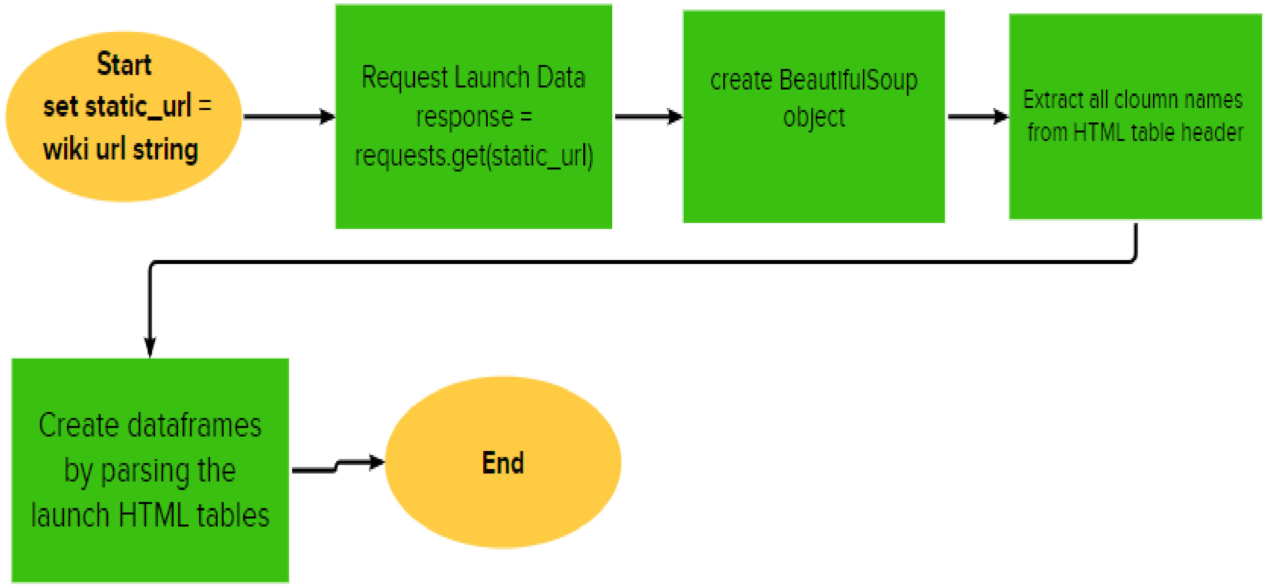
Data Collection – Web Scrapping

Flacon 9 launch data was collected using the bs4 Beautiful soup package

- ❖ from Wikipaedia. –

  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- ❖ The collected data was converted into a tabular dataframe for further analysis. – Using pandas



GitHub URL to notebook detailing steps:

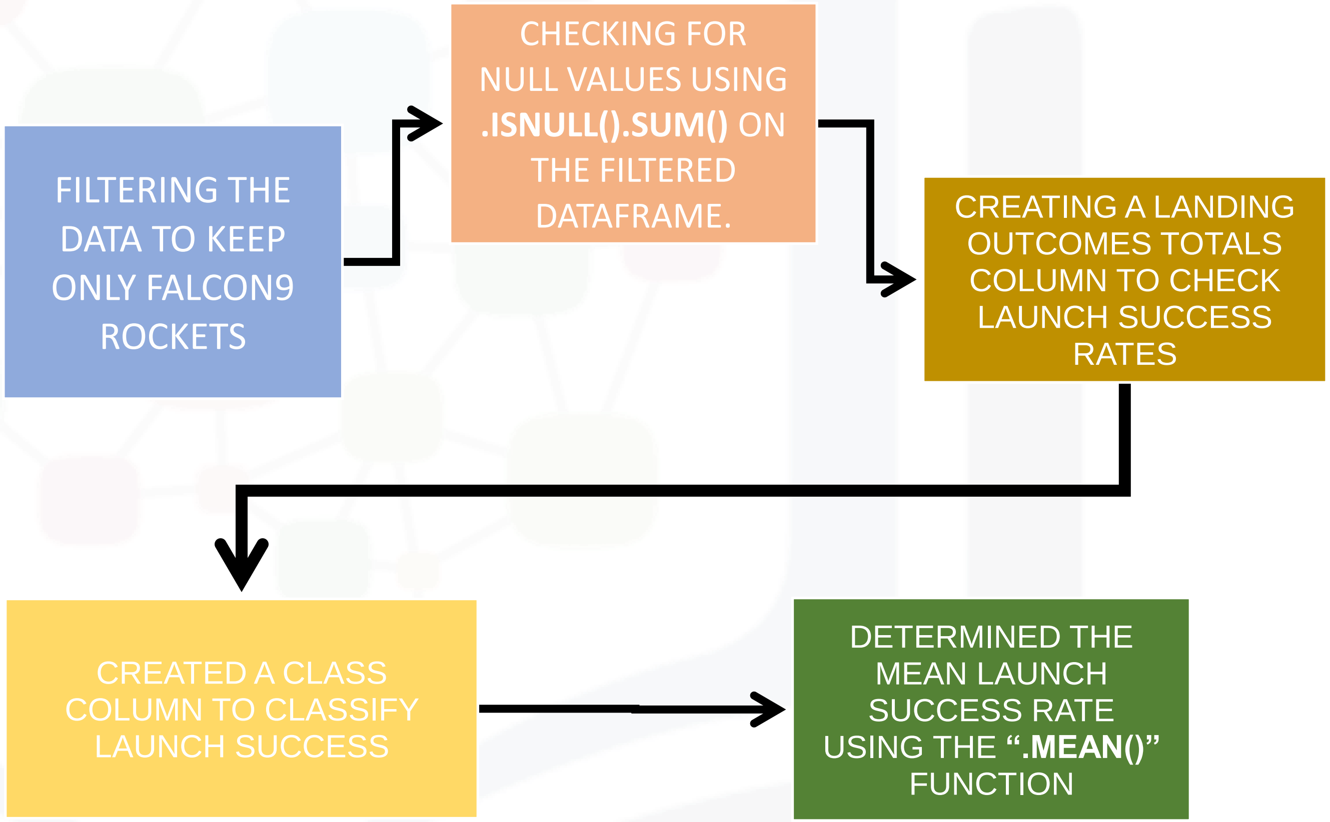
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-webscraping%20(1).ipynb

Data Wrangling

- ❖ Exploratory data analysis was initiated to clean extracted data and determine training labels for our models.
- ❖ Calculated columns for the landing outcome totals for all Falcon 9 launch sites and target orbits were found and included for further analysis
- ❖ Landing outcomes were categorized based on retrieved data and placed in a target column for our model

GitHub URL to notebook detailing steps:

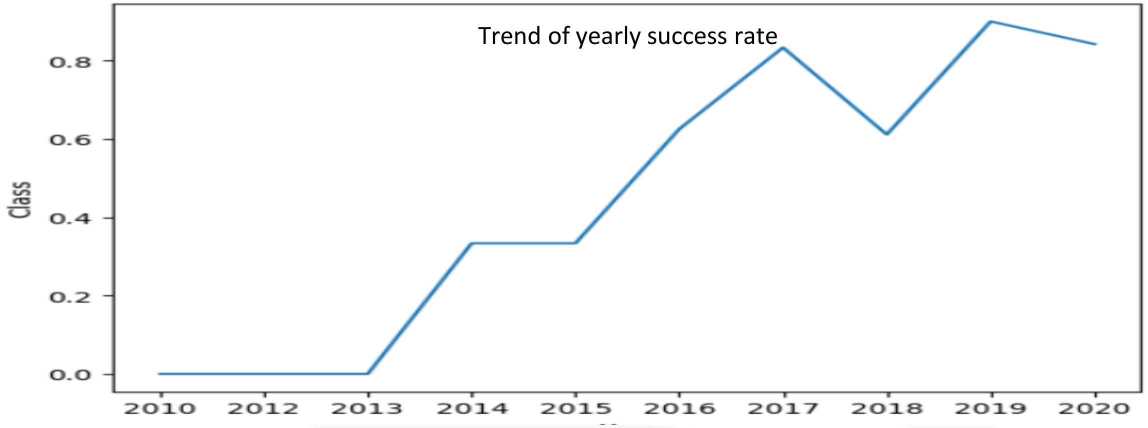
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



EXPLORATORY DATA ANALYSIS VIA DATA VISUALIZATION

Using data visualizing the following relationships were discovered between:

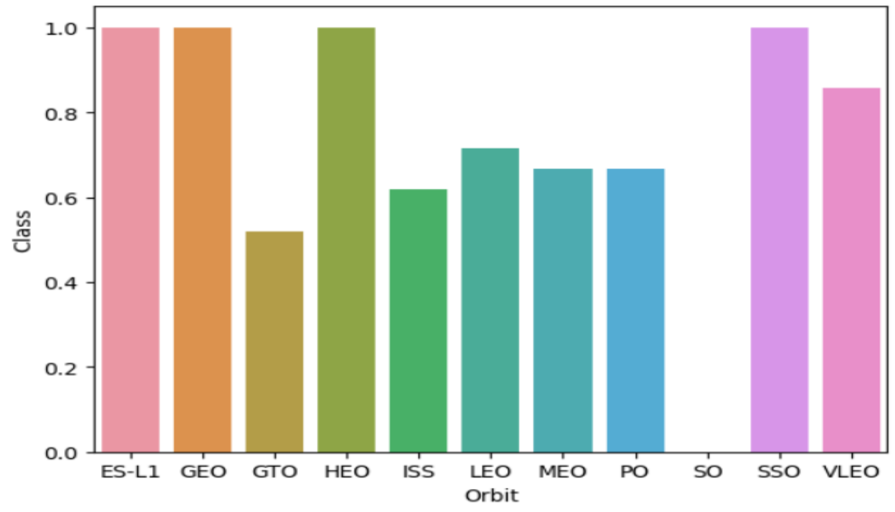
- flight numbers and launch sites
- payloads and launch site,
- success rate of each orbit type,
- flight number and orbit type,



Key take aways :

1. An upward trend was observed for yearly launch success trend.
2. Target orbits had varying success rates.

Plot of success rate of each target orbit



GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Exploratory Data Analysis with SQL

Using SQL, I ran queries of the Space X data set to :

1. Select a distinct ordered list of launch sites
2. Filter launch sites beginning with ‘CCA’ (limited to 5 results)
3. Display the average payload mass carried by booster F9 v1.1
4. List the date when the first successful landing outcome in ground pad was achieved.
5. List the names of the boosters which have success in drone ship and have a payload mass between 4000 and 6000 kg.
6. List the total number of successful and failure mission outcomes.
7. List the names of the booster versions which have carried the maximum payload mass.
8. List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in the year 2015.
9. Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

EXPLORATORY DATA ANALYSIS VIA SQL

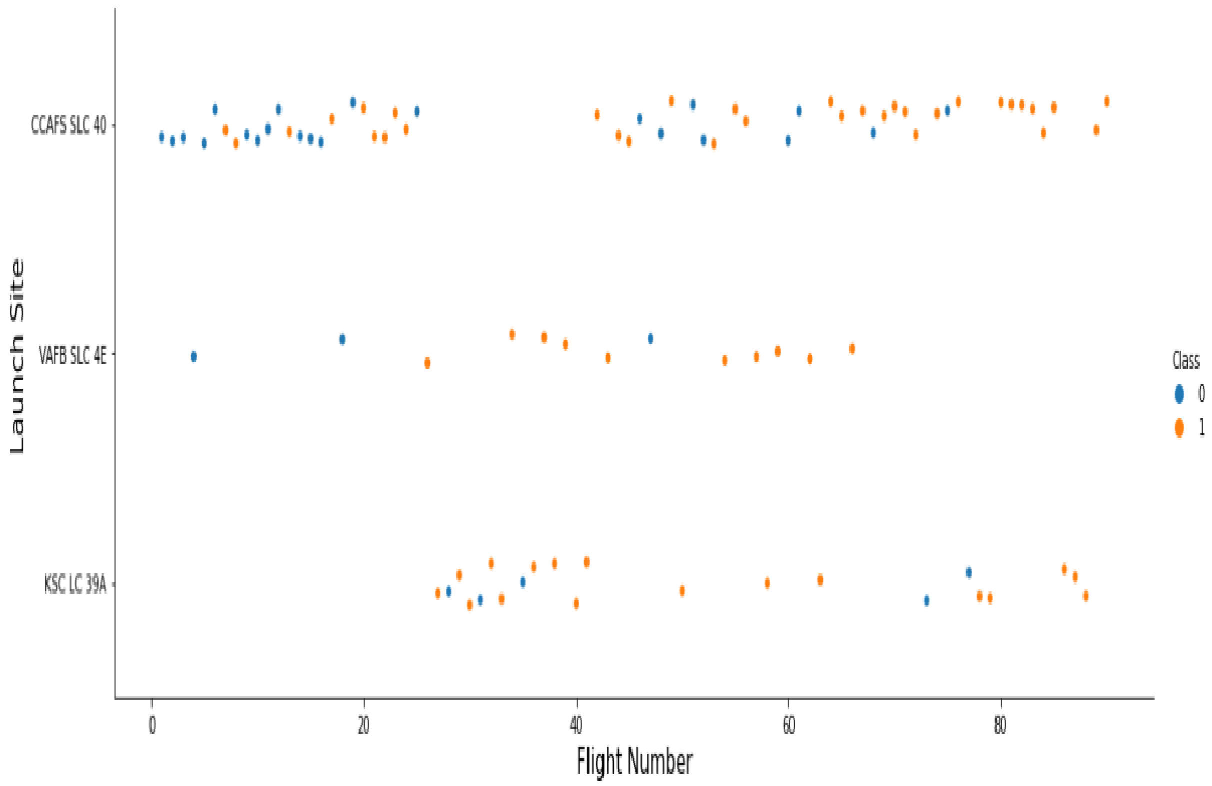
Further exploratory analysis was carried out using SQL via an IBM DB2 connection

Key insights:

1. Average payload for Falcon 9 v1.1 = 2,928.4 kg
2. Total Successful missions for Falcon 9 = 100
3. Total Failed missions for Falcon 9 = 1

Insights drawn from EDAs

Flight Number vs. Launch Site

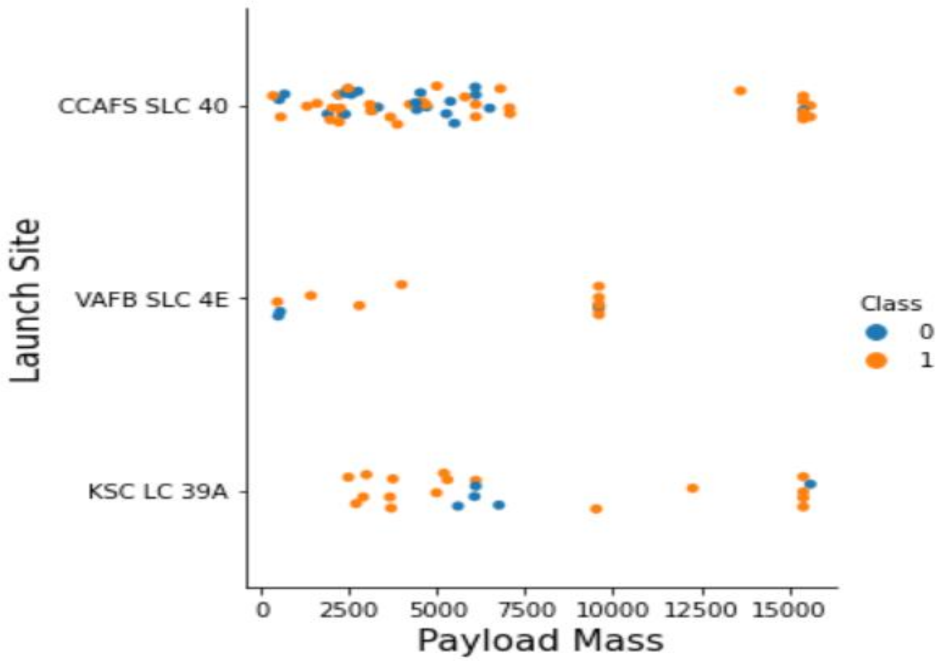


- We can see from the scatter chart below that flight successful rates seemed to increase as the number of flights increased

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Payload vs. Launch Site

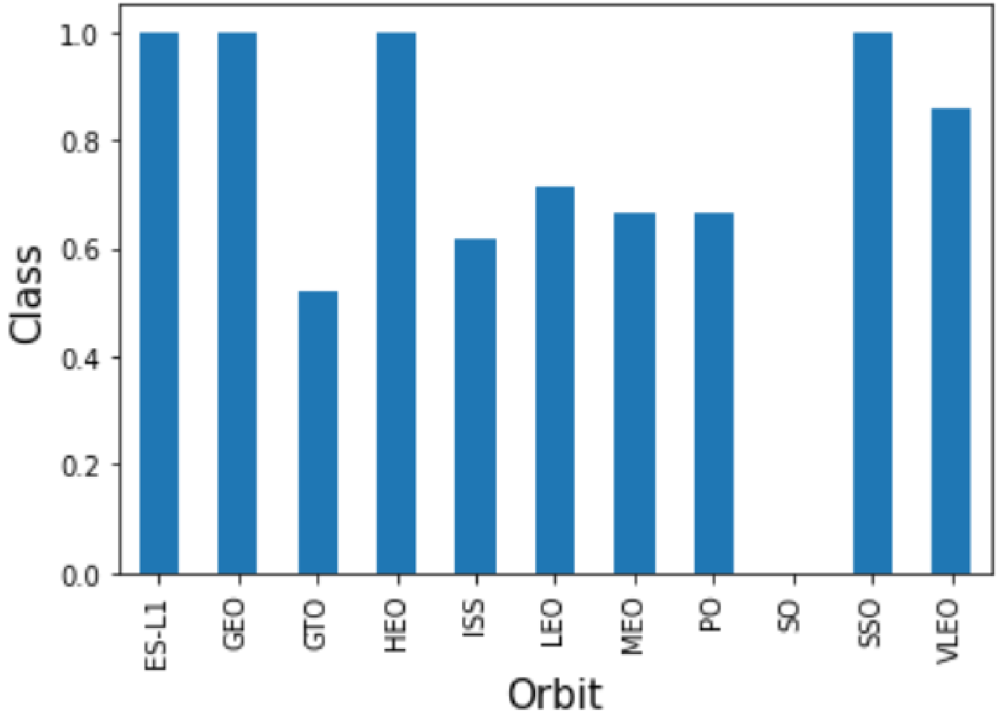


- Launch site CCAFS SLC 40 showed mixed results as to the success of flights when compared to payload mass
- However, we do see at KSC LC 39A a very high success rate between 25000 and 5500 Kg

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Success Rate vs. Orbit Type

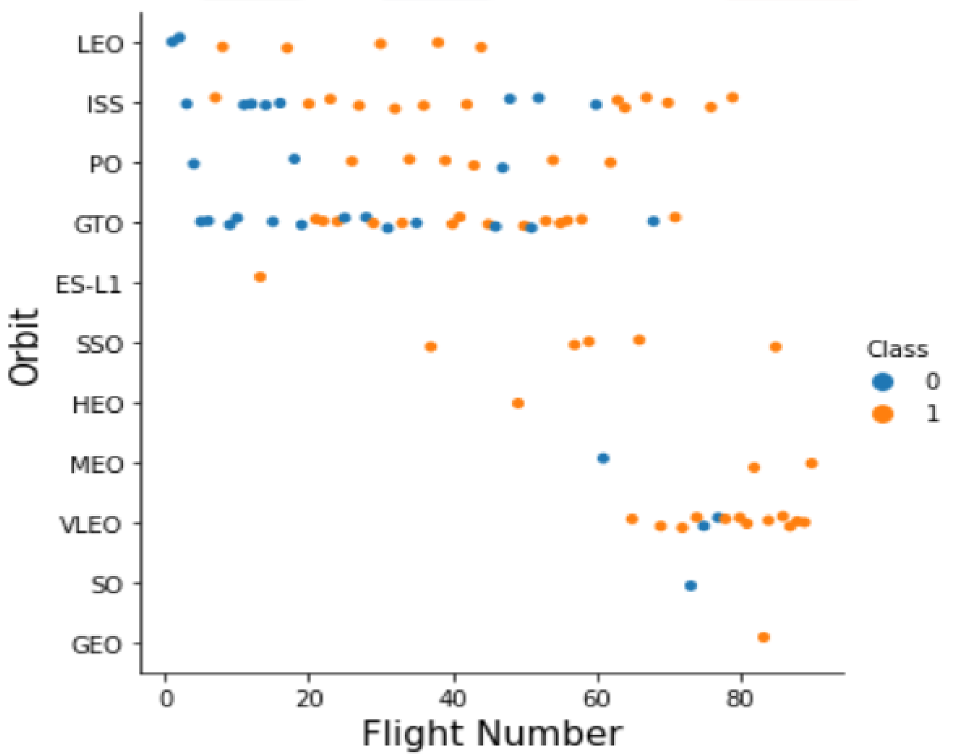


- Does orbit impact success rate? We do see that the following orbits performed better:
- **ES-L1**
- **GEO**
- **HEO**
- **SSO**

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Flight Number vs. Orbit Type

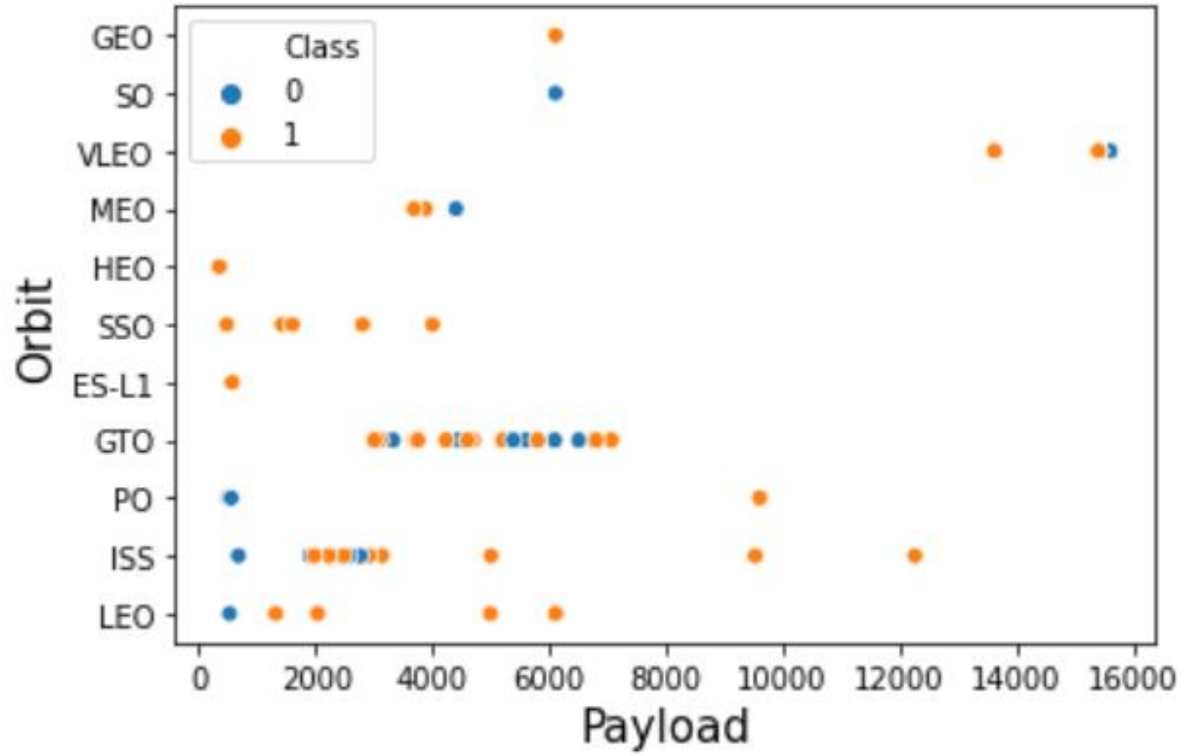


- We see different results at different orbits when related to flight number.
- Leo seems to improve with the number of flights
- GTO appears to have no relation with a balanced mix of success and failures
- Although the SSO didn't have many flights, all being successful, the number of flights didn't seem to have an impact on that success.

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Payload vs. Orbit Type

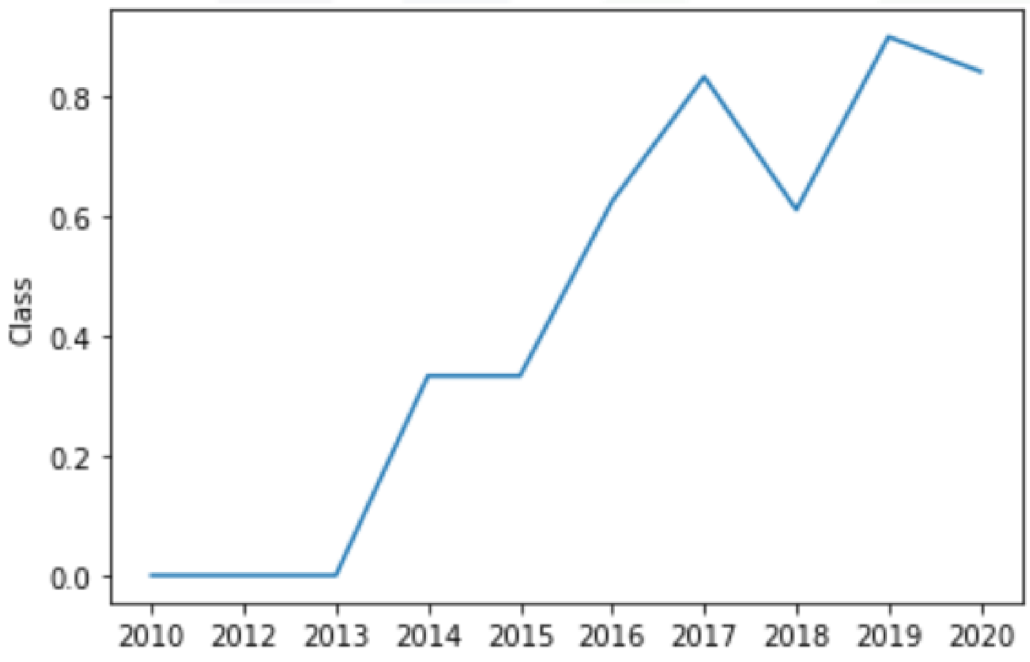


- Here we see that heavier payloads correlate to successful landings at the LEO ISS and PO orbits.
- The GTO and SSO show that payload isn't much of a factor for successful landings.

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

Launch Success Yearly Trend



We can observe a steady increase of successful attempts since 2013.

GitHub URL to the notebook detailing the EDA analysis:

https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

All Launch Site Names

In [16]:

```
%%sql

SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

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

Out[16]:

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

- Using SQL to select a list of all the launch site names.

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Total Payload Mass

In [40]:

```
%%sql
SELECT Customer , SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD
FROM SPACEXTBL
WHERE Customer = "NASA (CRS)"
```

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

Out[40]:

Customer	TOTAL_PAYLOAD
NASA (CRS)	45596.0

- The total payload mass carried by boosters launched by NASA (CRS) is 45,596 KG

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Average Payload Mass by F9 v1.1

In [50]:
```
%%sql
SELECT BOOSTER_VERSION, AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD
FROM SPACEXTBL
WHERE BOOSTER_VERSION LIKE "%F9 V1.1"
```

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

Out[50]:

Booster_Version	AVERAGE_PAYLOAD
F9 v1.1	2928.4

- The calculated average payload mass carried by booster version F9 v1.1 is 2,928.4 KG

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000 KG

In [43]:

```
%%sql
SELECT [Date], [Booster_Version], [Landing_Outcome], [PAYLOAD_MASS__KG_]
FROM SPACEXTBL
WHERE [PAYLOAD_MASS__KG_] > 4000
    and [PAYLOAD_MASS__KG_] < 6000
    and [Landing_Outcome] = 'Success (drone ship)'
```

```
 * sqlite:///my_data3.db
Done.
```

Out[43]:

Date	Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
05/06/2016	F9 FT B1022	Success (drone ship)	4696.0
14/08/2016	F9 FT B1026	Success (drone ship)	4600.0
30/03/2017	F9 FT B1021.2	Success (drone ship)	5300.0
10/11/2017	F9 FT B1031.2	Success (drone ship)	5200.0

- Here, we find the data for boosters which have successfully landed on drone ship and had payload mass (4000 < 6000)kg.
- We observe that the FT boosters were the ones that were successful.

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Total Mission Outcomes

In [108…

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_OUTCOME
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

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

Out[108…

Mission_Outcome	TOTAL_OUTCOME
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Data was queried to retrieve an insight into successful and failed mission outcomes
- Only one mission failure was observed

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Boosters Carried Maximum Payload

In [112...

```
%%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.
```

Out[112...

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

- Data was queried to return the names of the booster which have carried the maximum payload mass
- The B5 booster was observed to carry the biggest payload.

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Ranking Successful Landing Outcomes Between June 2010 and March 2017

In [136…

```
%%sql

SELECT LANDING_OUTCOME, COUNT(*) AS NO_SUCX_LANDINGS
FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE "%Success%"
AND DATE BETWEEN "04/06/2010" AND "20/03/2017"
GROUP BY LANDING_OUTCOME
```

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

Out[136…

Landing_Outcome	NO_SUCX_LANDINGS
Success	20
Success (drone ship)	8
Success (ground pad)	7

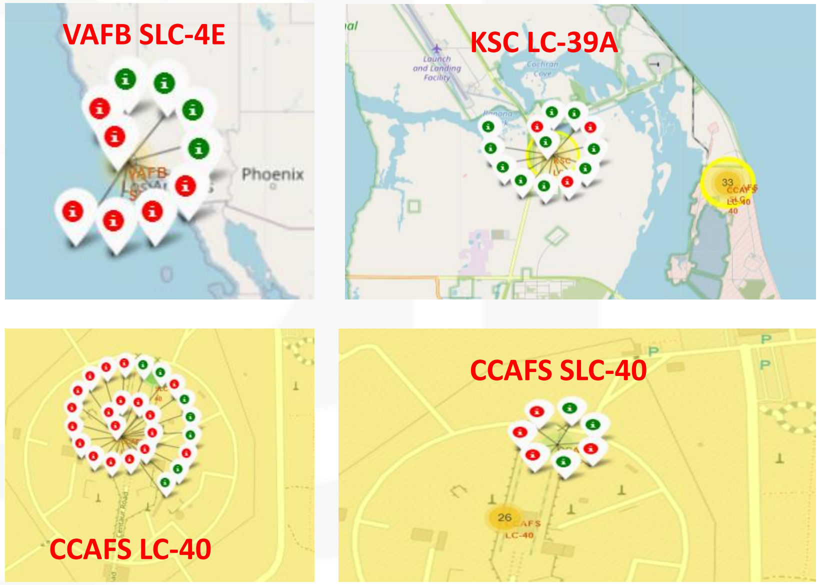
- The count of successful landing outcomes for drone ship or ground pad between the date 2010-06-04 and 2017-03-20, are shown here.
- Data also includes unassigned landing platform successes.

GitHub URL to the notebook detailing the SQL query steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1)%20(1).ipynb

Building Interactive Maps with Folium

- All launch sites were marked on the map – Observation made for all launch site being close to the beach and sea.
- Successful launches were marked with green markers with a class assignment of 1 and 0 & red markers for unsuccessful launches. – KSC LC – 39A had the highest success rate
- Site success rates were emphasized through marker clusters that clearly tell the general success rate for each site.
- Nearness to key landmarks was also investigated for possible relevance.

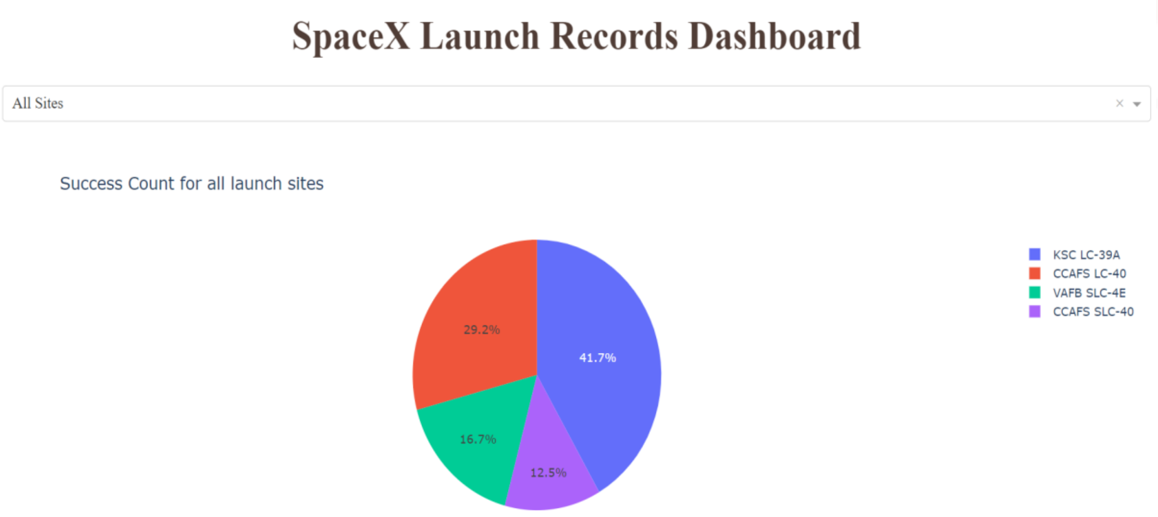
Success rates by Launch Sites via marker clusters



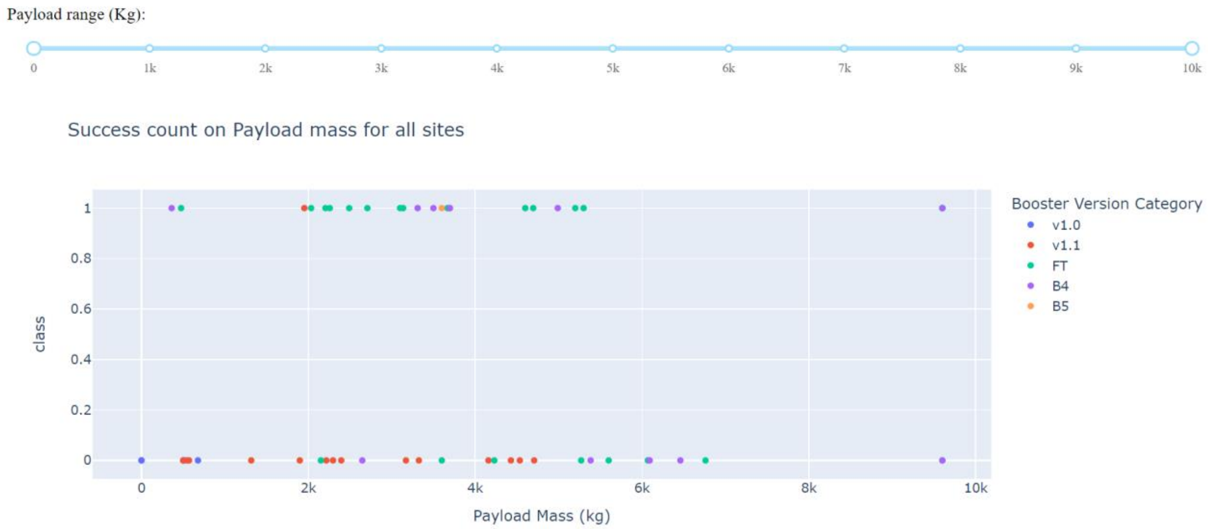
PLOTLY DASHBOARD

An interactive dashboard was developed with the following properties:

1. A pie chart to communicate successful launches by certain sites.



2. A scatter plot to show the relationship between Mission Outcome and Payload Mass for the different booster versions



Predictive Analysis

- The objective of this analysis was to produce a predictive model that can predict with a high accuracy whether a launch will be successful or not.
- The collected data is transformed and normalized before being split into training and testing data – 80% used for training and 20% for testing.
- Several models were developed, and their parameters tuned to obtain the highest performing model parameters.
- The prediction accuracy of all the models was compared and found as:

155]:

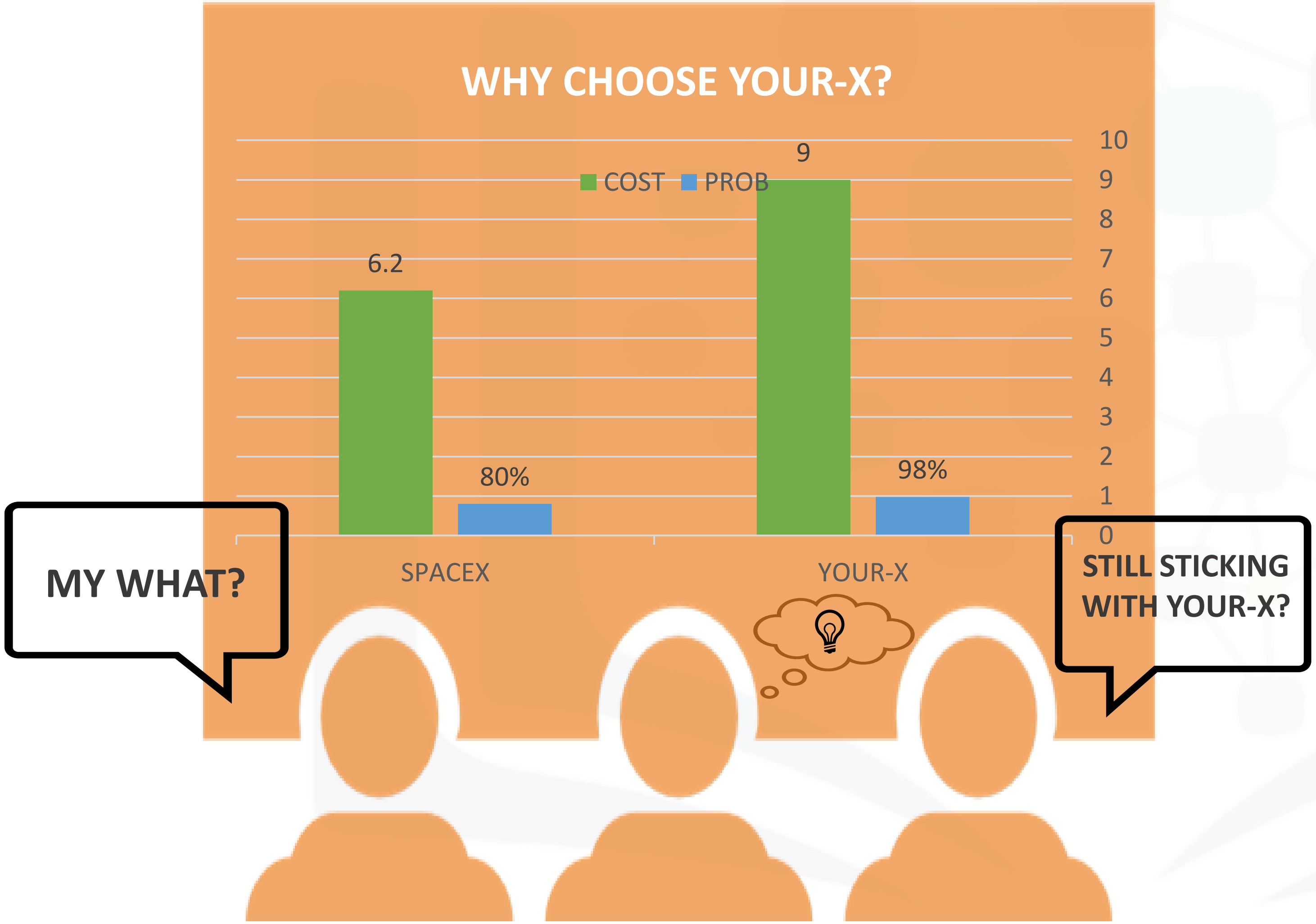
	Model	Accuracy Score
0	LOGISTIC REGRESSION	0.833333
1	SUPPORT VECTOR MACHINE	0.833333
2	DECISION TREE	0.888889
3	K-NEAREST NEIGHBOR	0.833333

Key insights:

1. Almost all models had similar accuracy scores
2. In this case, Decision tree fared best after multiple runs
3. However, Decision tree accuracy changes at every run and is sometimes the lowest
4. All models turned out high accuracies

GitHub URL to the notebook detailing the modelling steps:
https://github.com/TheophilusAnkrah/ANALYSING-SPACEX-ROCKET-LAUNCH-DATA/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

DISCUSSION



- Was enough data collected for the project?
- Was data credible and unbiased?
- Did the methodology lack in any way?
- Were the dependent variables enough to inform the model?
- Were the right models selected?
- Did the tuning improve model delivery?
- Can the model be improved?
- Was the objective of this project achieved?

CONCLUSION

We can conclude that:

- ❖ SPACEX gets better success rates with increasing number of launches
- ❖ KSC LC-39A had the most successful launches of any sites.
- ❖ Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- ❖ Launch success rate started to increase in 2013 till 2020.
- ❖ The Decision tree classifier is not the best machine learning algorithm for this task as with each run , the accuracies keep changing.
- ❖ Logistic regression model is recommended as we can also check for the probability of a prediction and has equally high accuracy as the others.