Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies

Data Collection

Data Wrangling

EDA With Data Visualization and SQL

Interactive Map with Folium

Dashboard with Plotly Dash

Predictive Analysis

- Summary of all results

Exploratory Data Analysis Results

Interactive Analytics Demo in Screenshots

Predictive Analysis Results

Introduction

- Project background and context
- We predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars while other providers quote cost of 165 million dollars each.

- Problems you want to find answers

- Problems you want to find answers

What influences the successful landing of the rocket

The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.

What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

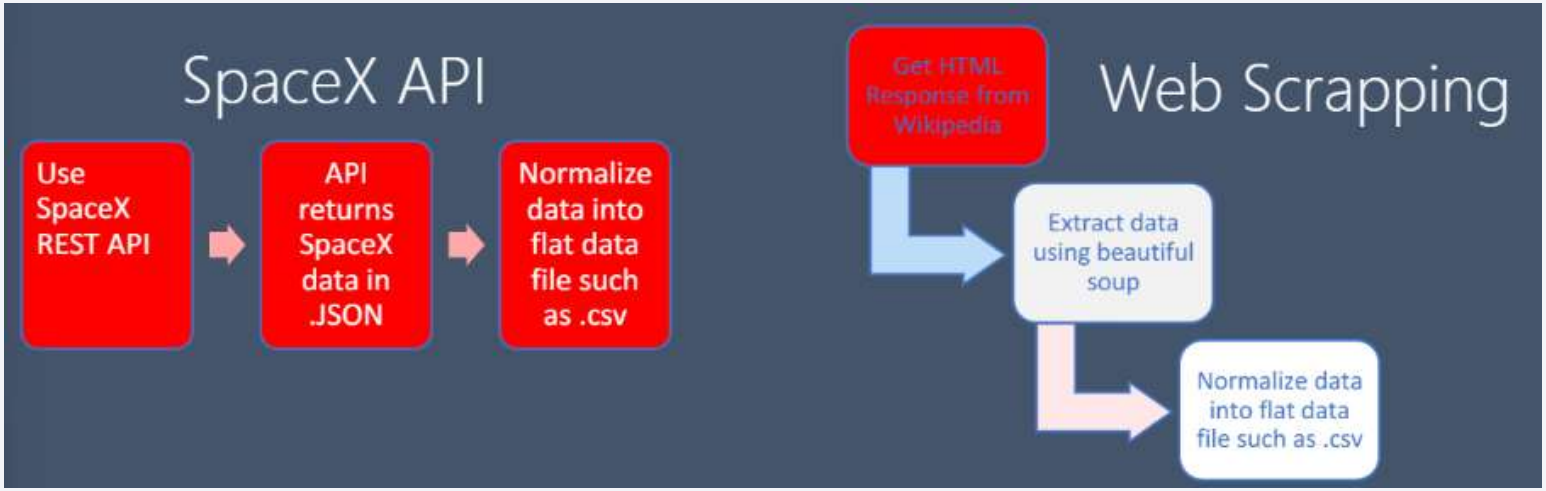
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API and Web Scraping from Wikipedia
- Perform data wrangling
  - One hot encoding data fields for machine learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Plotting Scatter plots and bar graphs to show relationship between variables
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Cross Validation with 10 folds

Data Collection

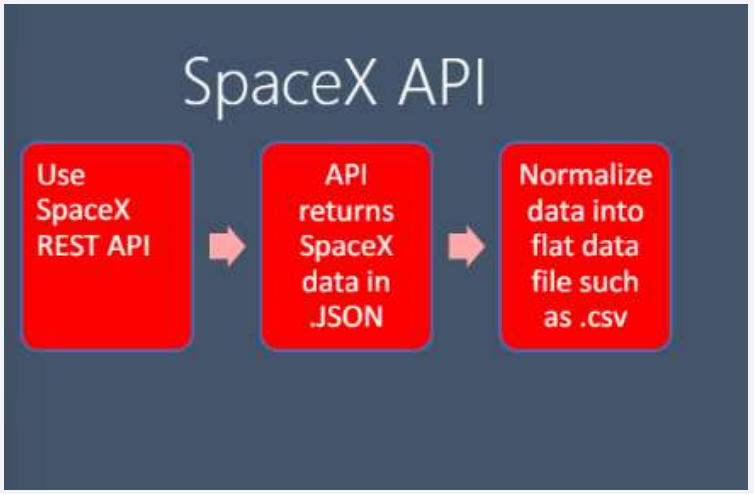
- Describe how data sets were collected.
- Data collected using SpaceX's REST API and Webscraping from Wikipedia Website



Data Collection – SpaceX API

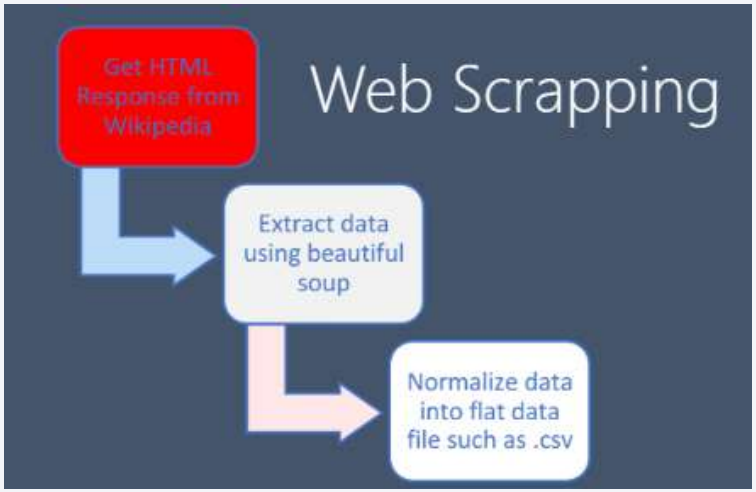
- Requests.get() The API link, convert it into a JSON File, and clean that data and normalize it into a .csv file

https://github.com/raunaqjabbal/IBM-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

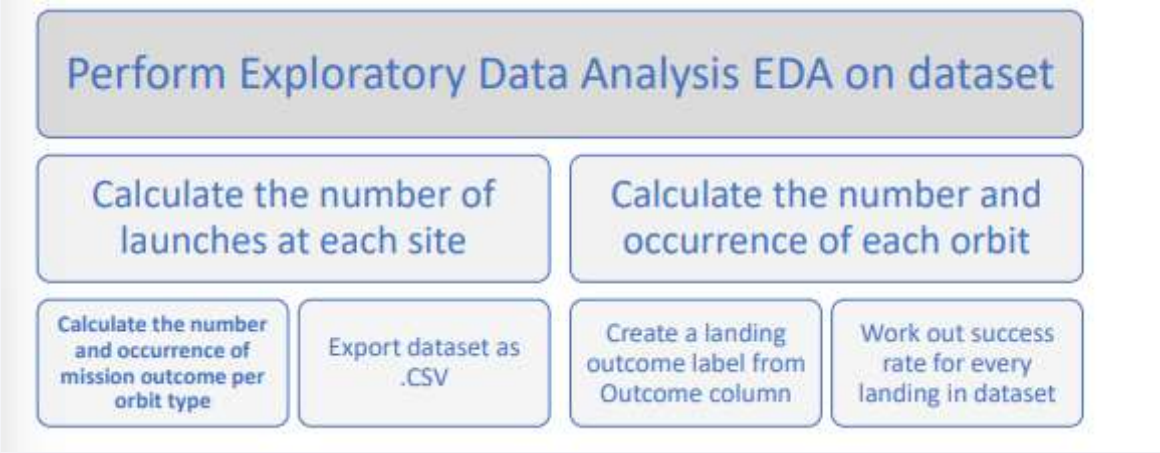
- Get response from HTML
- Create BeautifulSoup object
- Find table
- Get column names
- Append data to keys and convert to .csv
- Create dictionary and convert to dataframe



- https://github.com/raunaqjabbal/IBM-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- https://github.com/raunaqjabbal/IBM-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter Graphs being drawn:
- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass
- Bar Graph being drawn: Mean VS. Orbit
- Line Graph being drawn: Success Rate VS. Year
- https://github.com/raunaqjabbal/IBM-Project/blob/main/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

https://github.com/raunaqjabbal/IBM-Project/blob/main/EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks
- Example of some trends in which the Launch Site is situated in.

•Are launch sites in close proximity to railways? No

•Are launch sites in close proximity to highways? No

•Are launch sites in close proximity to coastline? Yes

•Do launch sites keep certain distance away from cities? Yes

https://github.com/raunaqjabbal/IBM-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Used Python Anywhere to host the website live 24/7 so your can play around with the data and view the data –
- The dashboard is built with Flask and Dash web framework.
- Graphs - Pie Chart showing the total launches by a certain site/all sites - display relative proportions of multiple classes of data. - size of the circle can be made proportional to the total quantity it represents.
- Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions - It shows the relationship between two variables. - It is the best method to show you a non-linear pattern. - The range of data flow, i.e. maximum and minimum value, can be determined. - Observation and reading are straightforward.
- https://github.com/raunaqjabbal/IBM-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- BUILDING MODEL

• Load our dataset into NumPy and Pandas and transform Data

• Split our data into training and test data sets and check how many test samples we have

• Decide which type of machine learning algorithms we want to use

• Set our parameters and algorithms to GridSearchCV and fit our datasets into the GridSearchCV objects and train our dataset.

- EVALUATING MODEL

• Check accuracy for each model

• Get tuned hyperparameters for each type of algorithms and plot Confusion Matrix

- IMPROVING MODEL

- • Feature Engineering and Algorithm Tuning

- FINDING THE BEST PERFORMING CLASSIFICATION MODEL

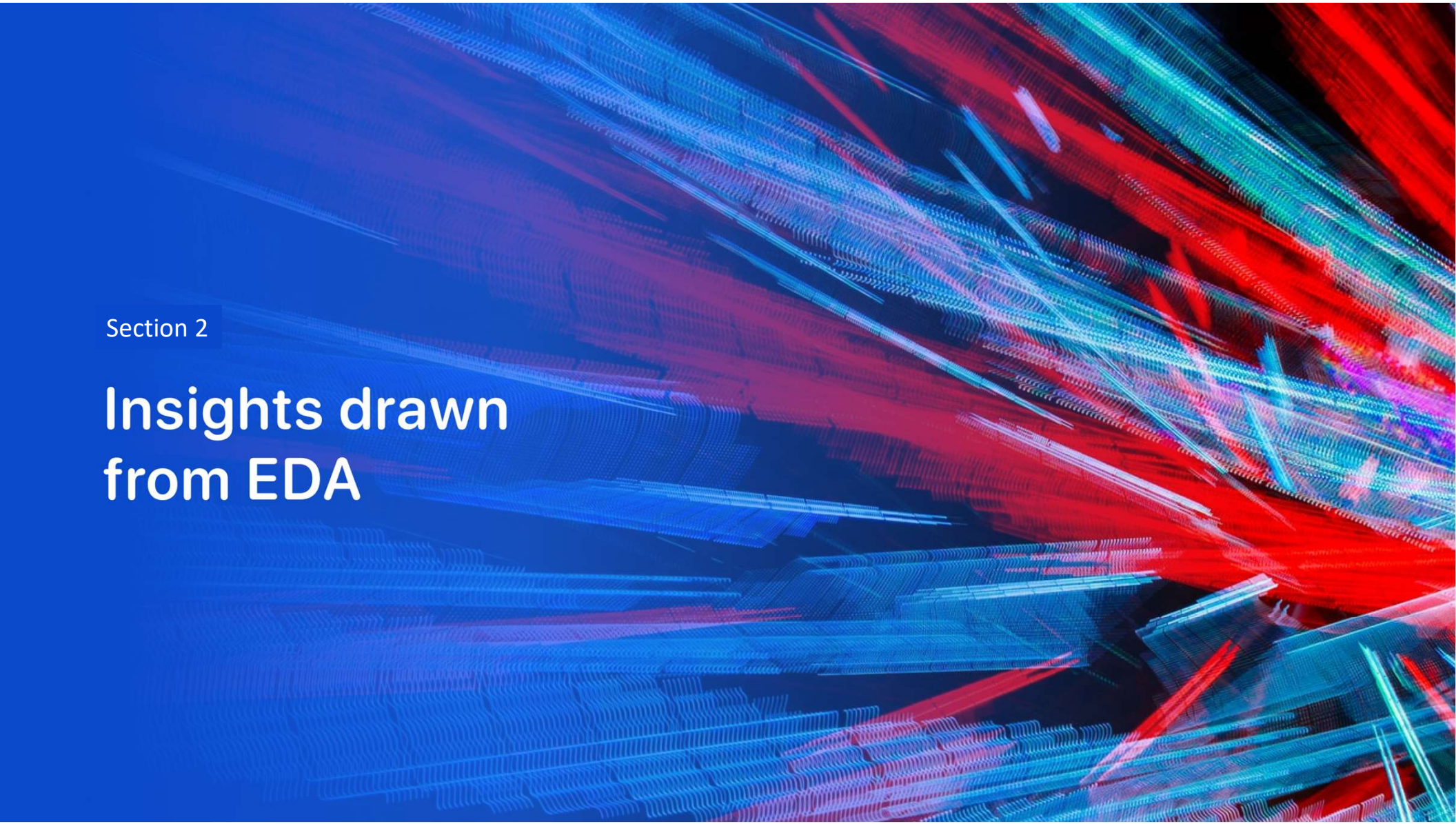
• The model with the best accuracy score wins the best performing model

• In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

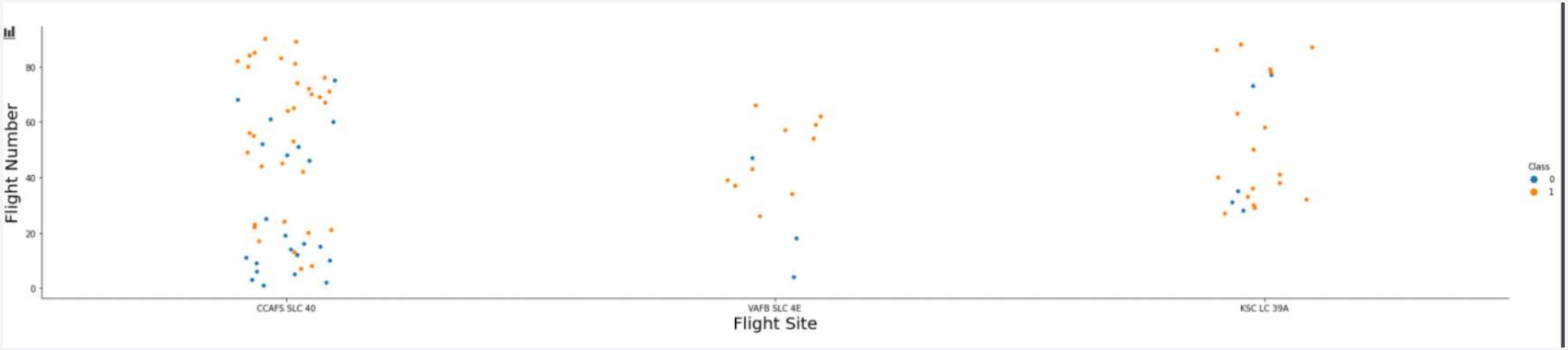
https://github.com/raunaqjabbal/IBM-Project/blob/main/Machine%20Learning%20Prediction%20lab.ipynb

Results

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

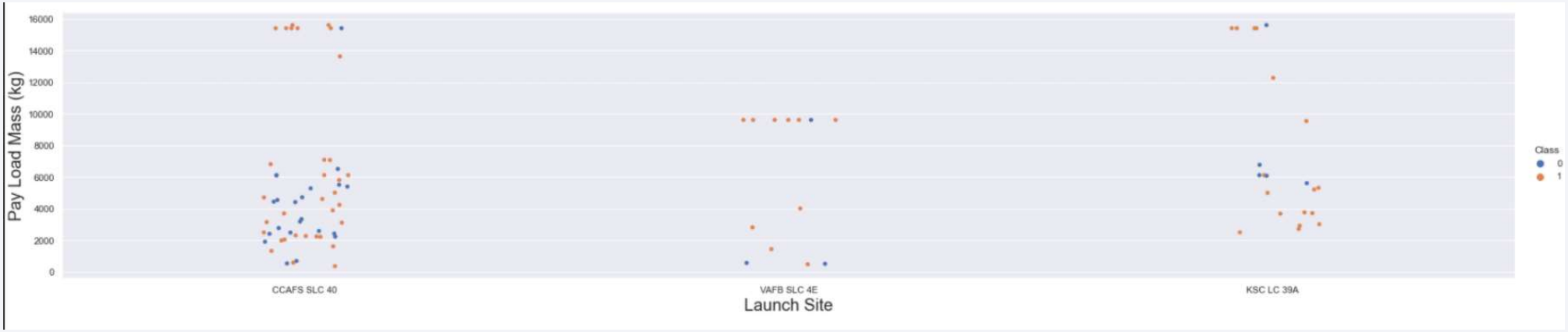

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- The more amount of flights at a launch site the greater the success rate at a launch site.

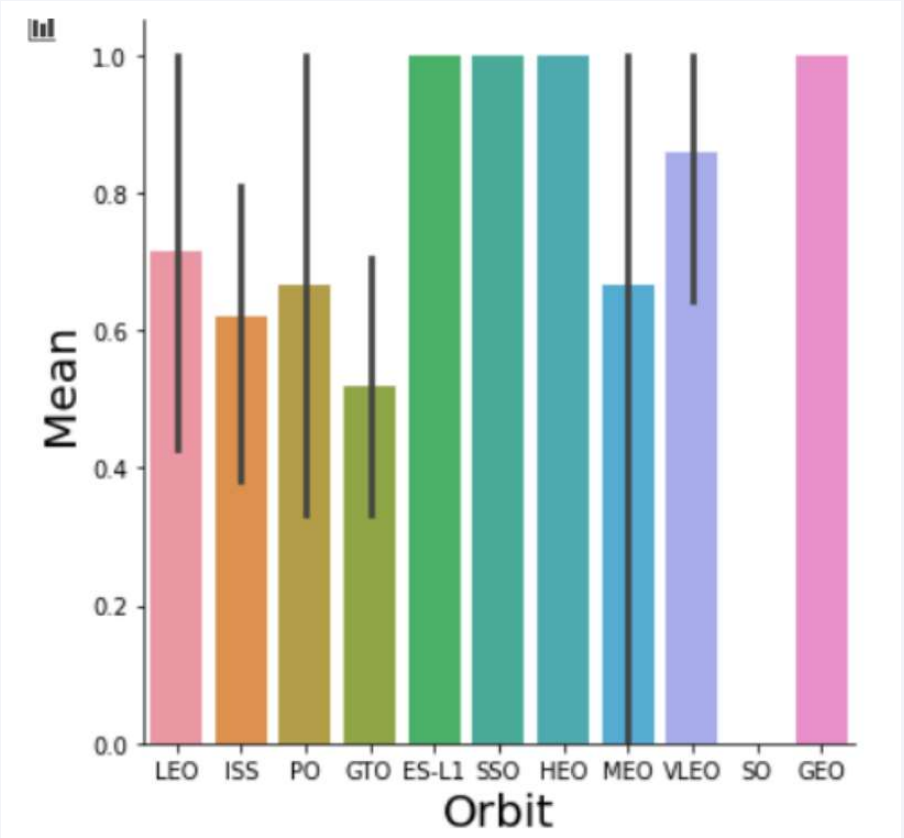
Payload vs. Launch Site



- The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

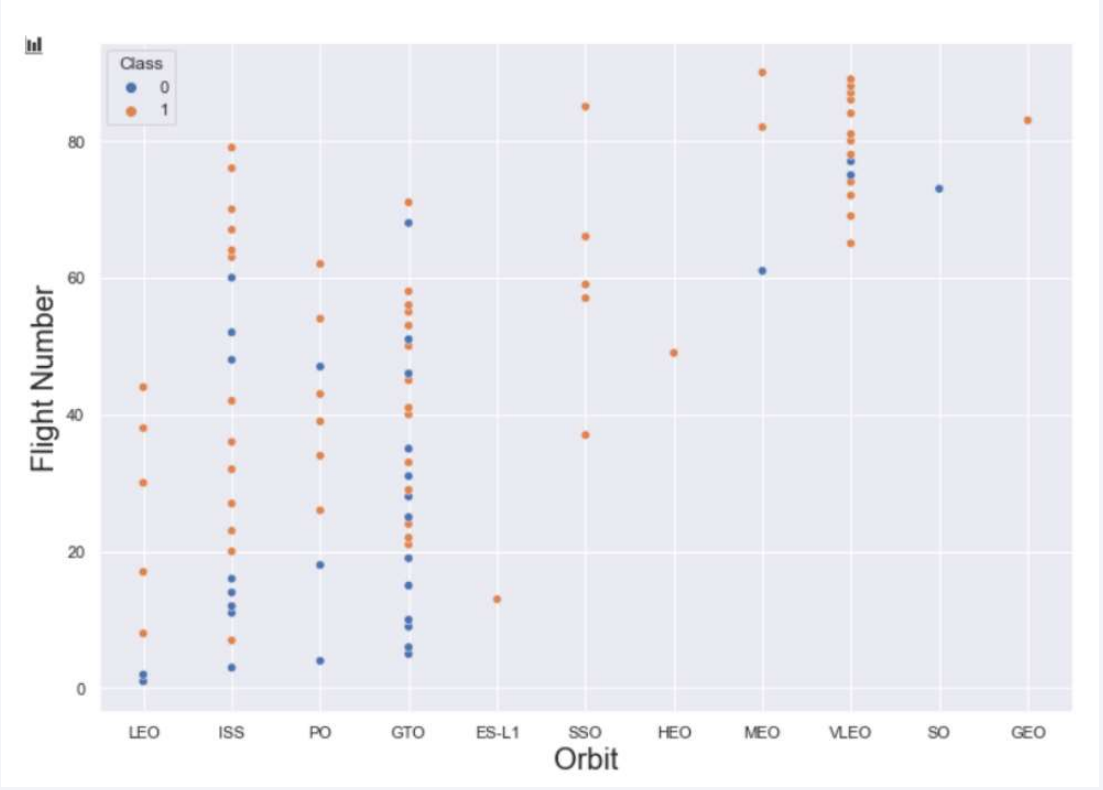
Success Rate vs. Orbit Type

- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



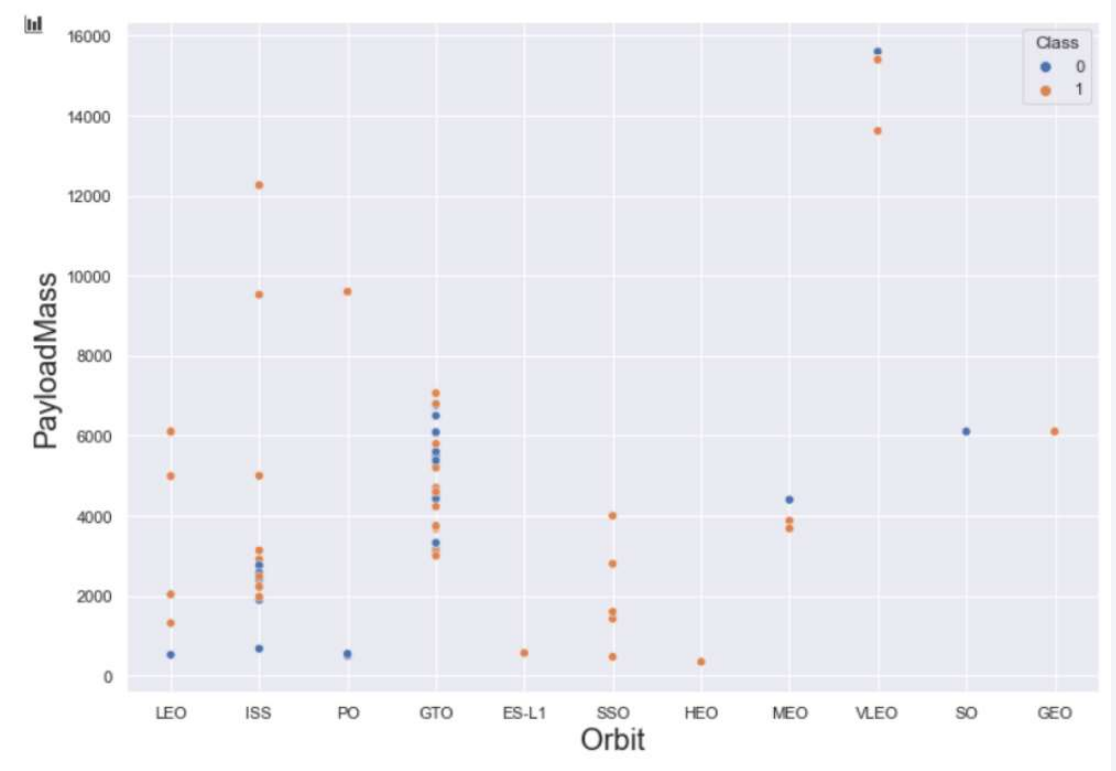
Flight Number vs. Orbit Type

- You should see 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 when in GTO orbit.



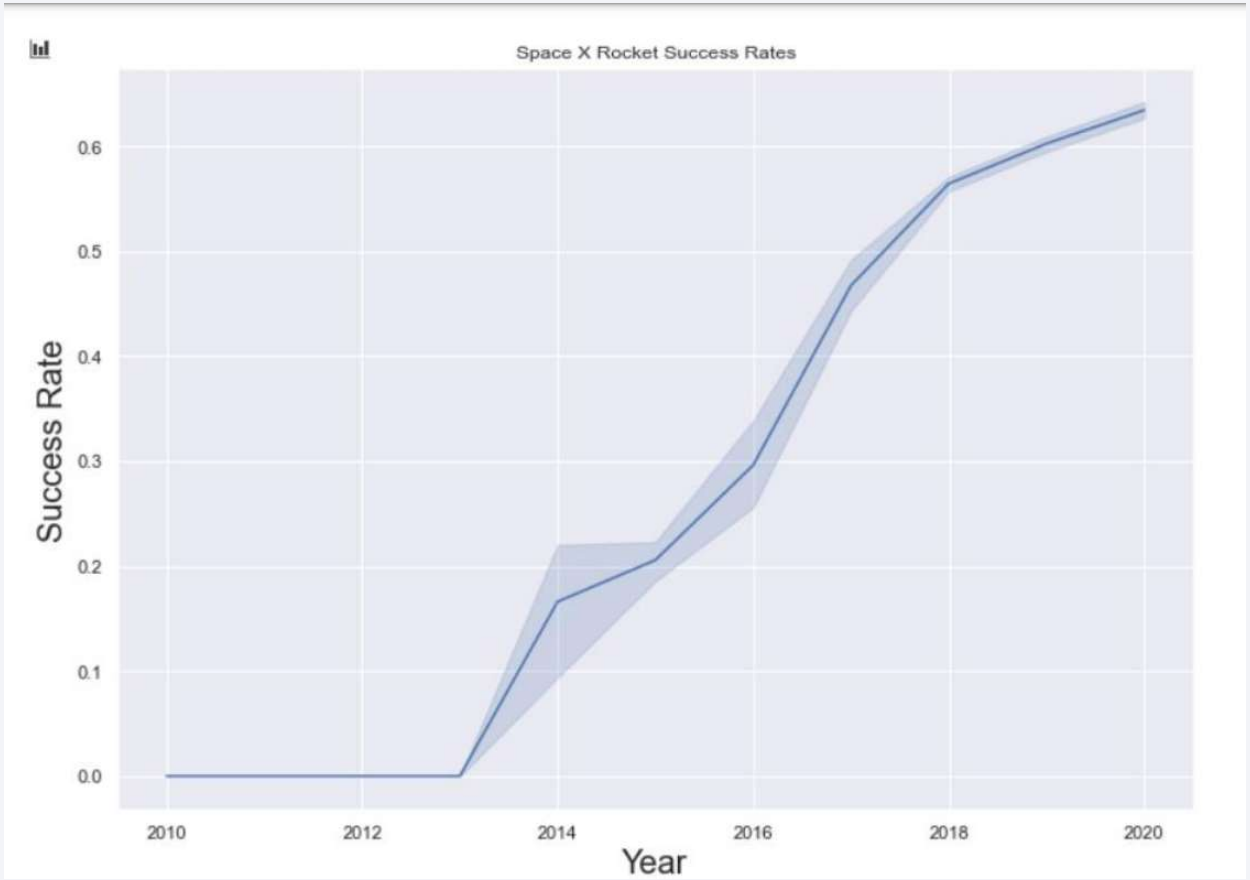
Payload vs. Orbit Type

- You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits



Launch Success Yearly Trend

- you can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- select DISTINCT Launch_Site from tblSpaceX
- Unique Launch Sites CCAFS LC-40 CCAFS SLC-40 CCAFS SLC-40 KSC LC-39A VAFB SLC-4E

- Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from tblSpaceX

Launch Site Names Begin with 'CCA'

- select TOP 5 * from tblSpaceX WHERE Launch_Site LIKE 'KSC%’
- Using the word TOP 5 in the query means that it will only show 5 records from tblSpaceX and LIKE keyword has a wild card with the words ‘KSC%’ the percentage in the end suggests that the Launch_Site name must start with KSC.

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.0000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.0000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.0000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.0000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.0000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass : **45596**

- Using the function SUM summates the total in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Customer NASA (CRS)

Average Payload Mass by F9 v1.1

- select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX where Booster_Version = 'F9 v1.1' : **2928**

- Using the function AVG works out the average in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

First Successful Ground Landing Date

- select MIN(Date) SLO from tblSpaceX where Landing_Outcome = ”Success (drone ship)” : **06-05-2016**

- Using the function MIN works out the minimum date in the column Date The WHERE clause filters the dataset to only perform calculations on Landing_Outcome Success (drone ship)

Successful Drone Ship Landing with Payload between 4000 and 6000

- select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (ground pad)' AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000
- Selecting only Booster_Version The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship) The AND clause specifies additional filter conditions Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Success%') as Successful_Mission_Outcomes, (SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Failure%') as Failure_Mission_Coutcomes **100 Success 1 Failure**
- a much harder query I must say, we used subqueries here to produce the results. The LIKE '%foo%' wildcard shows that in the record the foo phrase is in any part of the string in the records for example. PHRASE "(Drone Ship was a Success)" LIKE '%Success%' Word 'Success' is in the phrase the filter will include it in the dataset

Boosters Carried Maximum Payload

- SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS _KG_) AS [Maximum Payload Mass] FROM tblSpaceX GROUP BY Booster_Version ORDER BY [Maximum Payload Mass] DESC

- Using the word DISTINCT in the query means that it will only show Unique values in the Booster_Version column from tblSpaceX GROUP BY puts the list in order set to a certain condition. DESC means its arranging the dataset into descending order

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows x 2 columns

2015 Launch Records

- SELECT DATENAME(month, DATEADD(month, MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM tblSpaceX WHERE (Landing_Outcome LIKE N'%Success%') AND (YEAR(CONVERT(date, Date, 105)) = '2015')

- a much more complex query as I had my Date fields in SQL Server stored as NVARCHAR the MONTH function returns name month. The function CONVERT converts NVARCHAR to Date. WHERE clause filters Year to be 2015

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

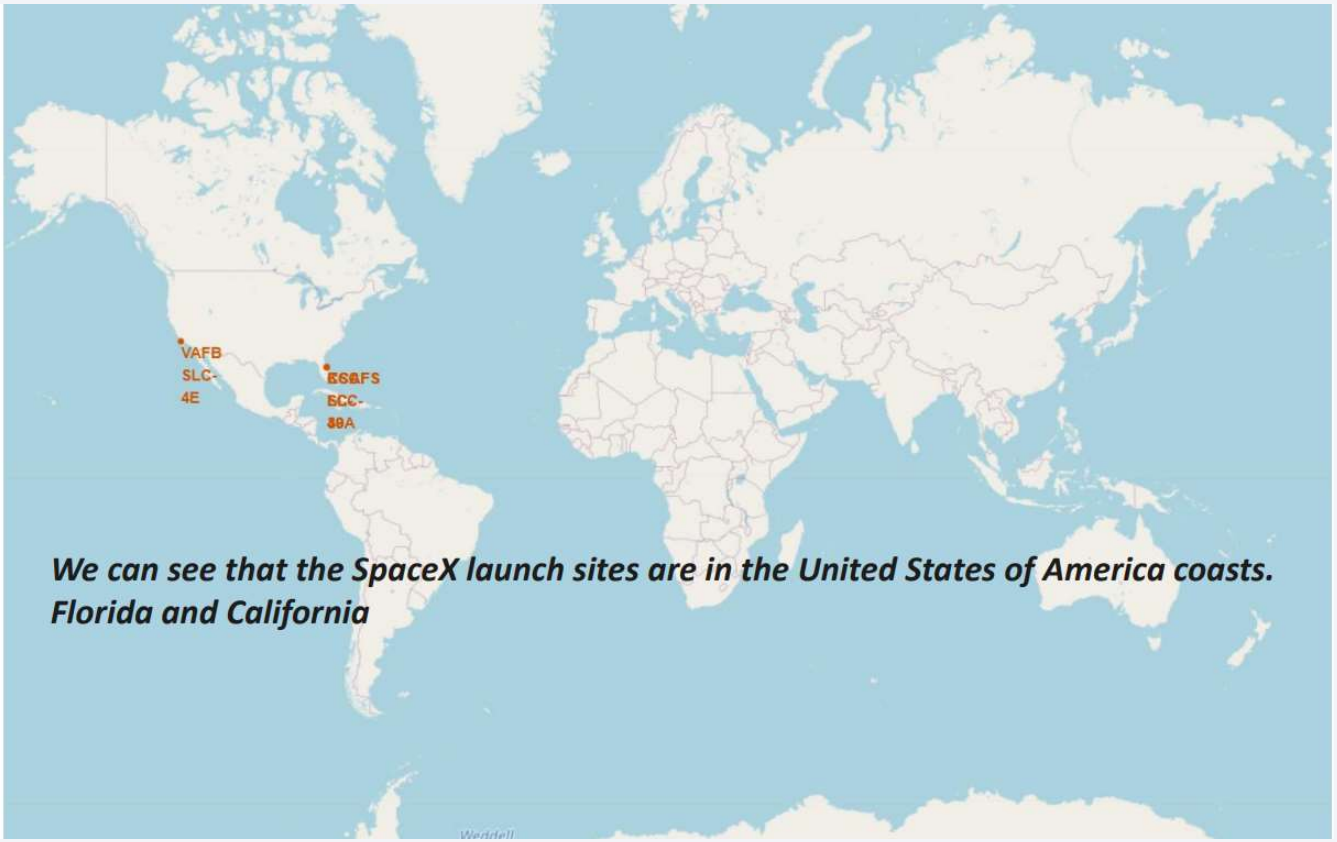
- SELECT COUNT(Landing_Outcome) FROM tblSpaceX WHERE (Landing_Outcome LIKE '%Success%') AND (Date > '04-06-2010') AND (Date < '20-03-2017') : **34**

- Function COUNT counts records in column WHERE filters data LIKE (wildcard) AND (conditions) AND (conditions)

Section 3

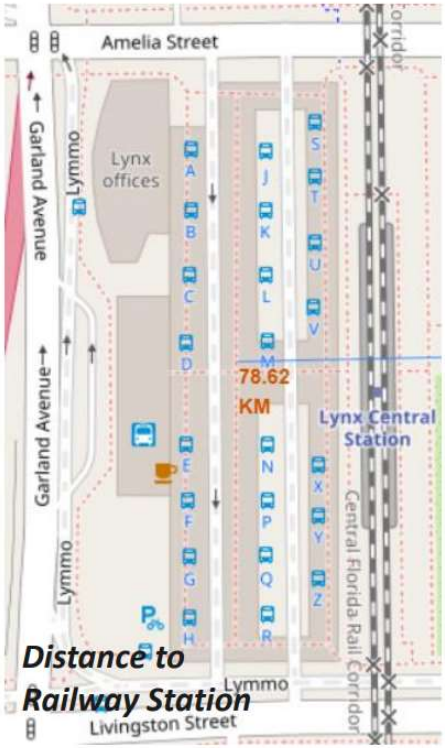
Launch Sites Proximities Analysis

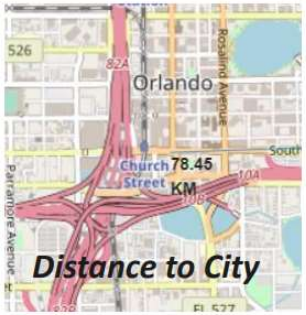
All launch sites global map markers



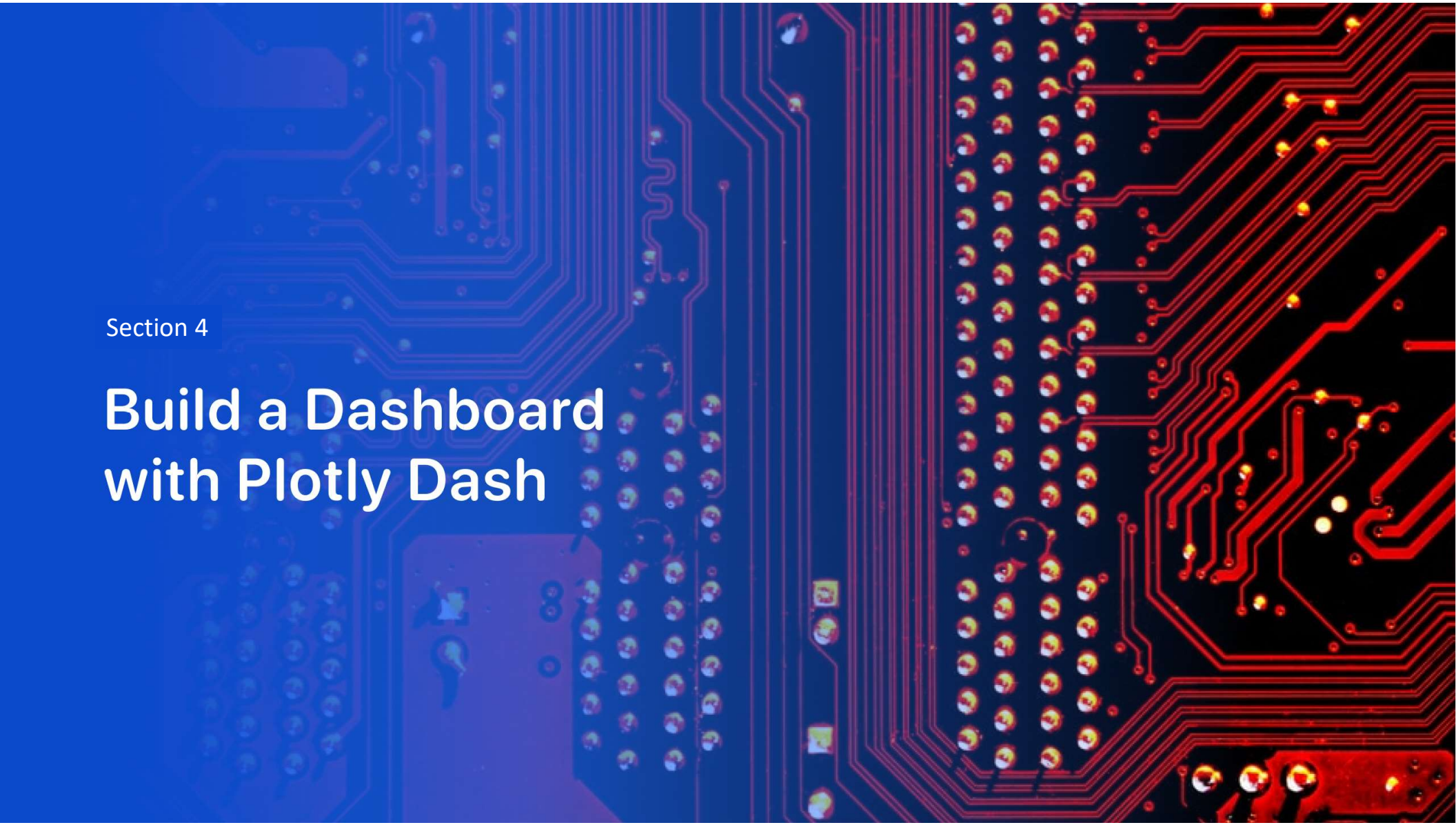
Colour Labelled Markers

Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference





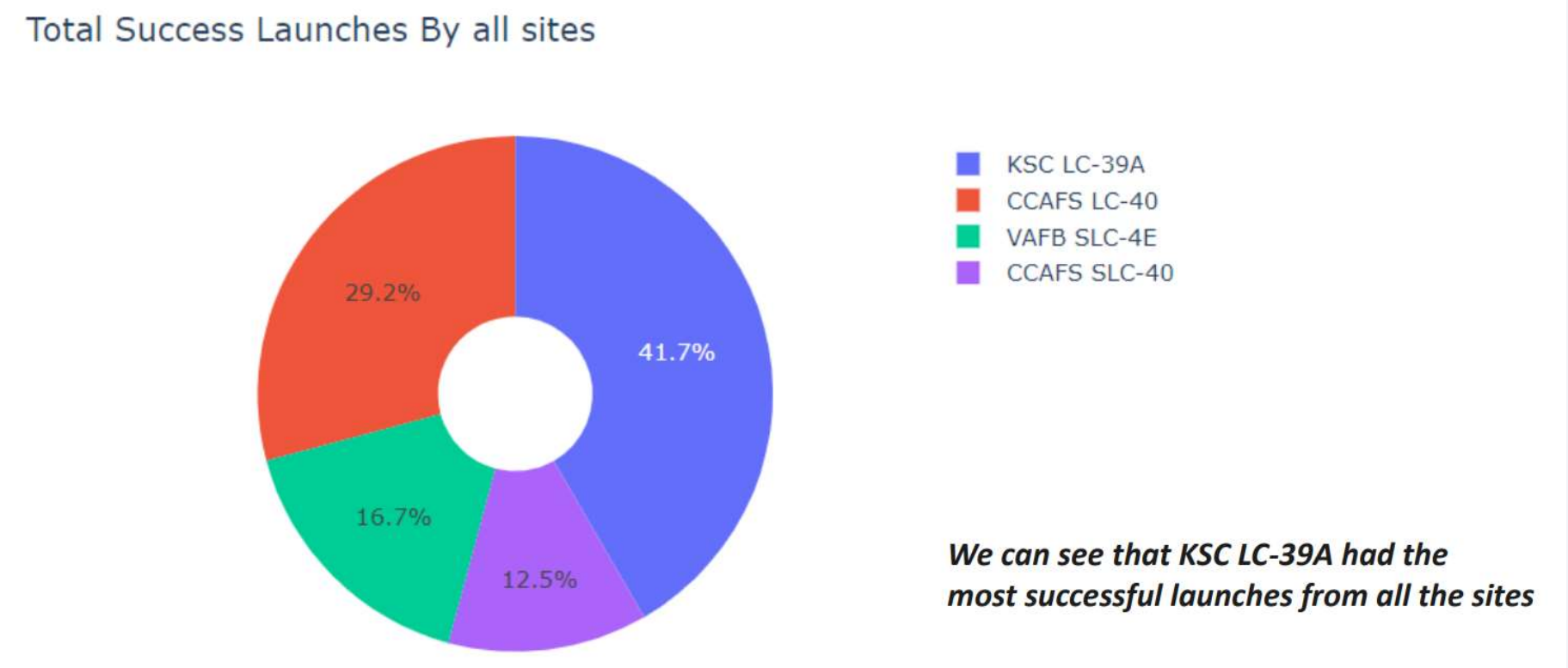
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



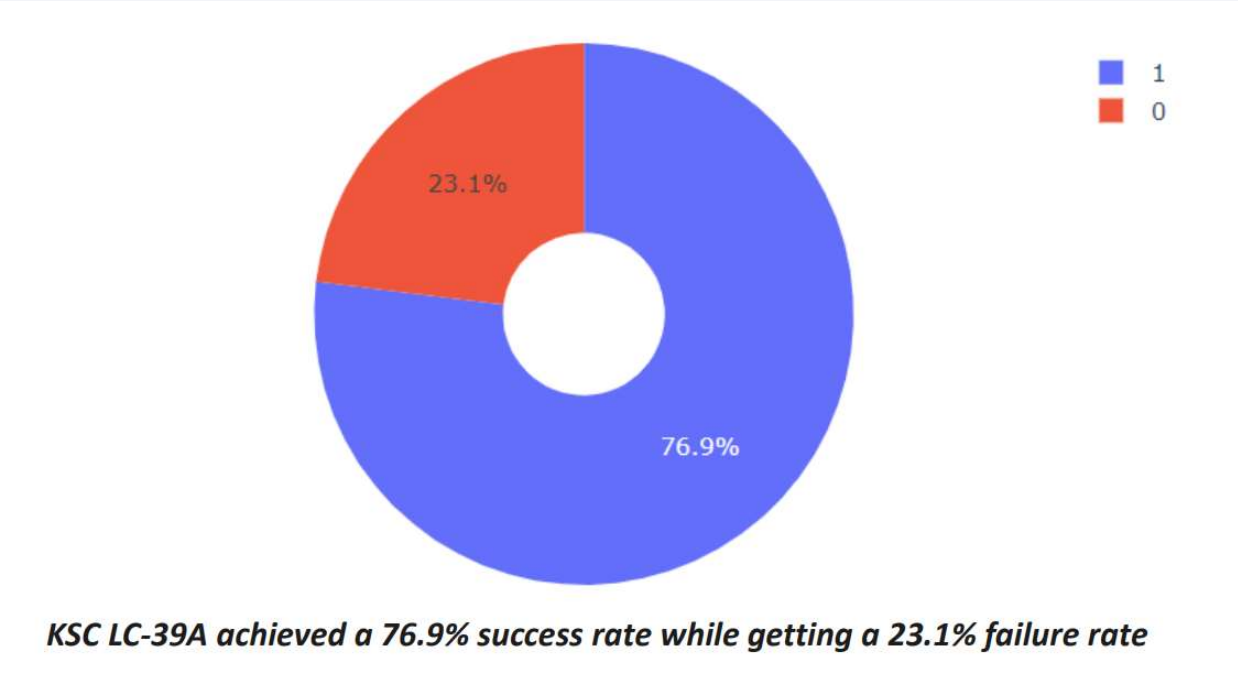
Section 4

Build a Dashboard with Plotly Dash

DASHBOARD – Pie chart showing the success percentage achieved by each launch site

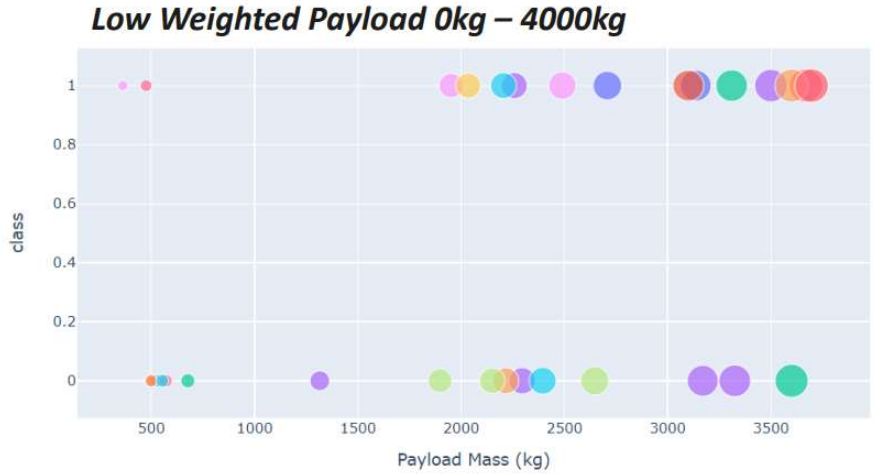


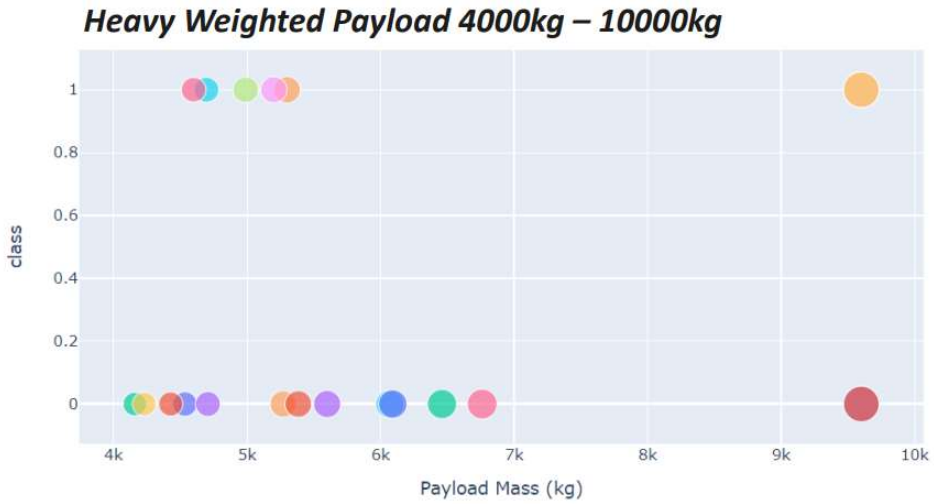
DASHBOARD – Pie chart for the launch site with highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

<Dashboard Screenshot 3>





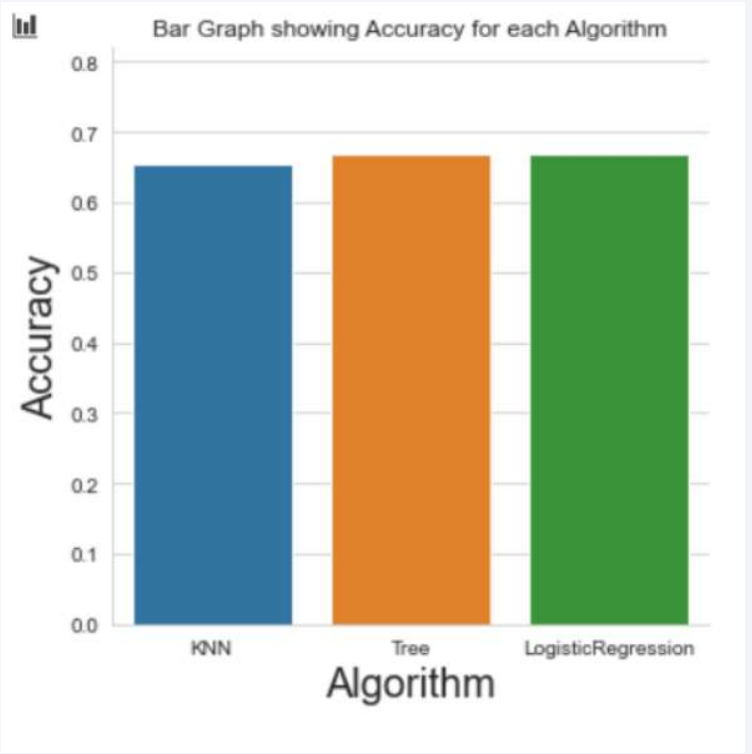
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

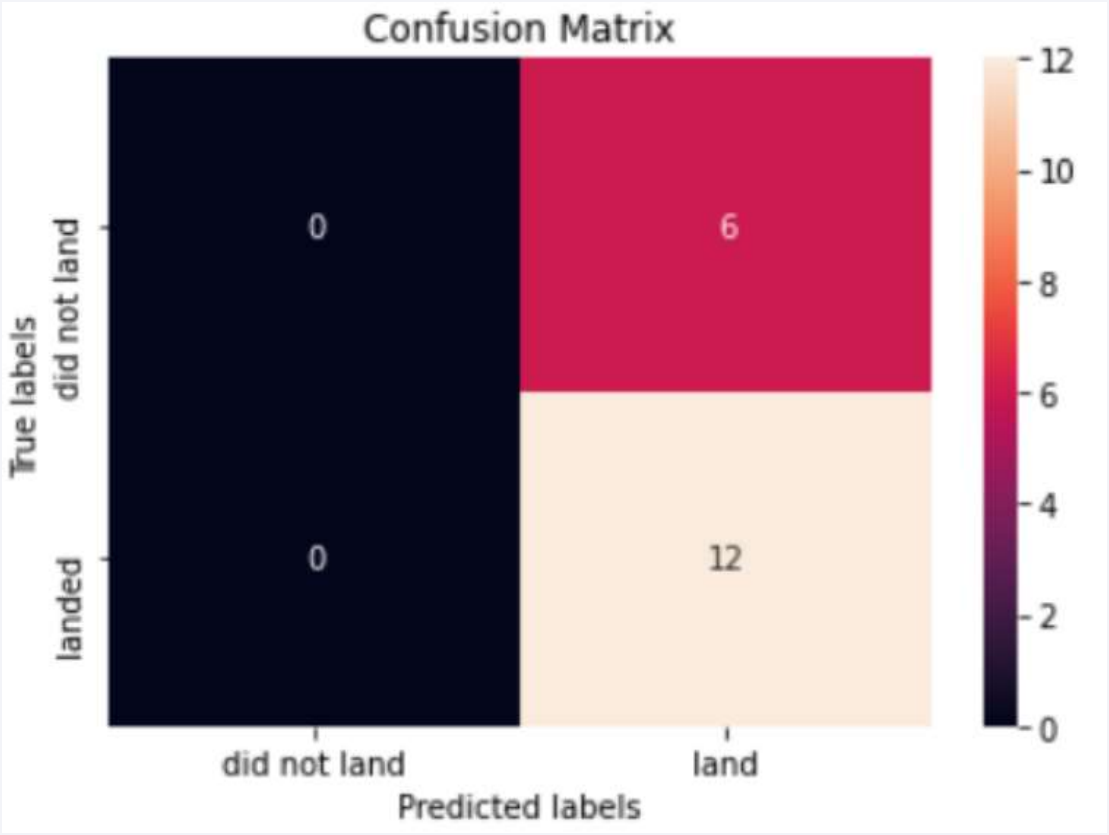
Classification Accuracy

- As you can see our accuracy is extremely close but we do have a winner its down to decimal places! The tree algorithm wins!!
- After selecting the best hyperparameters for the decision tree classifier using the validation data, we achieved 83.33% accuracy on the test data.



Confusion Matrix

- Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!