IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriel Marcondes
21th January/2022

Outline

Executive Summary

- Summary of methodologies
  - Data Collection with API, SQL and Web Scraping
  - Data Wrangling and Analysis
  - Interactive Maps with Folium
  - Predictive Analysis for each classification model
- Summary of all results
  - Data Analysis along with Interactive Visualizations
  - Best model for Predictive Analysis

Introduction

SpaceX is a revolutionary company who has disrupt the space industry by offering a rocket launches specifically Falcon 9 as low as 62 million dollars; while other providers cost upward of 165 million dollar each. Most of this saving thanks to SpaceX astounding idea to reuse the first stage of the launch by re-land the rocket to be used on the next mission. Repeating this process will make the price down even further. As a data scientist of a startup rivaling SpaceX, the goal of this project is to create the machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

Problems that needed solving:

- What influences if the rocket will land successfully?
- 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

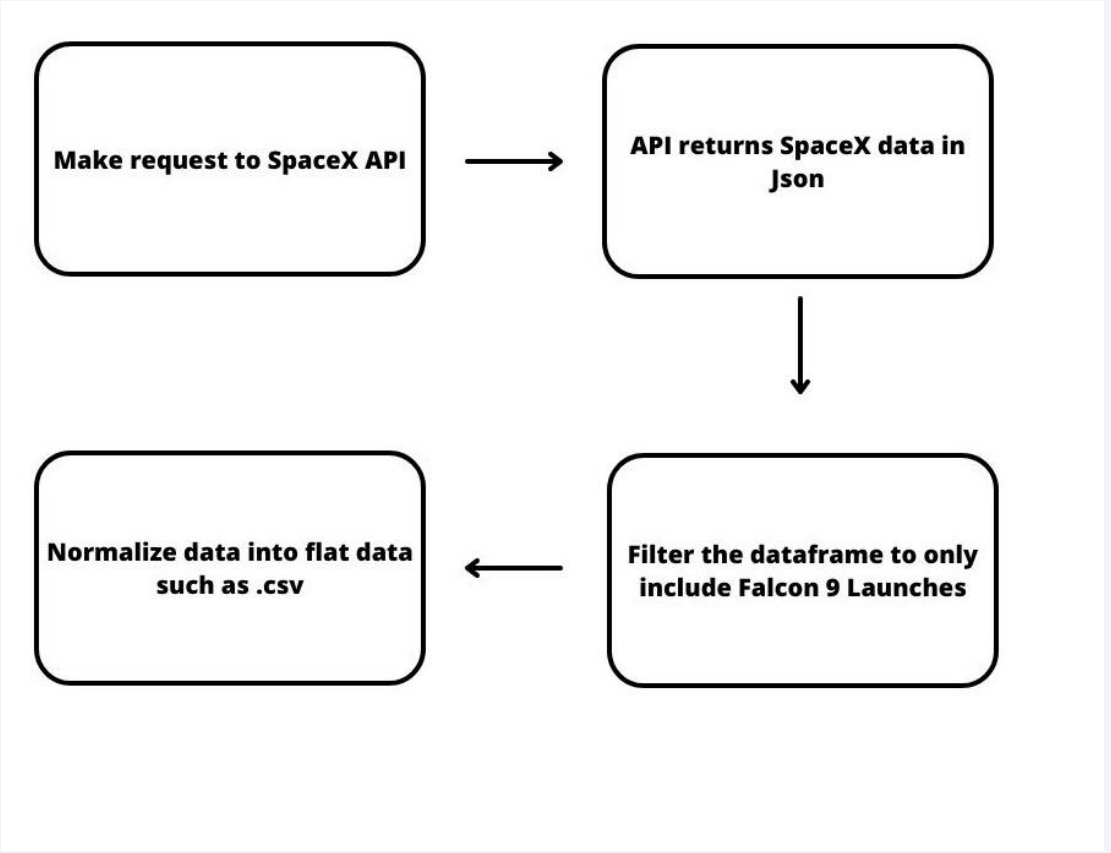
- Data collection methodology:
  - SpaceX Rest API
  - (Web Scrapping) from Wikipedia
- Perform data wrangling
  - Clean the data and explore it to find patterns in the data to determine the labels for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land given the data.

Data Collection

- Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. As mentioned, the dataset was collected by REST API and Web Scrapping from Wikipedia For REST API, its started by using the get request. Then, we decoded the response content as Json and turn it into a pandas dataframe using json_normalize(). We then cleaned the data, checked for missing values and fill with whatever needed.
- For web scrapping, we will use the BeautifulSoup to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for further analysis

Data Collection – SpaceX API

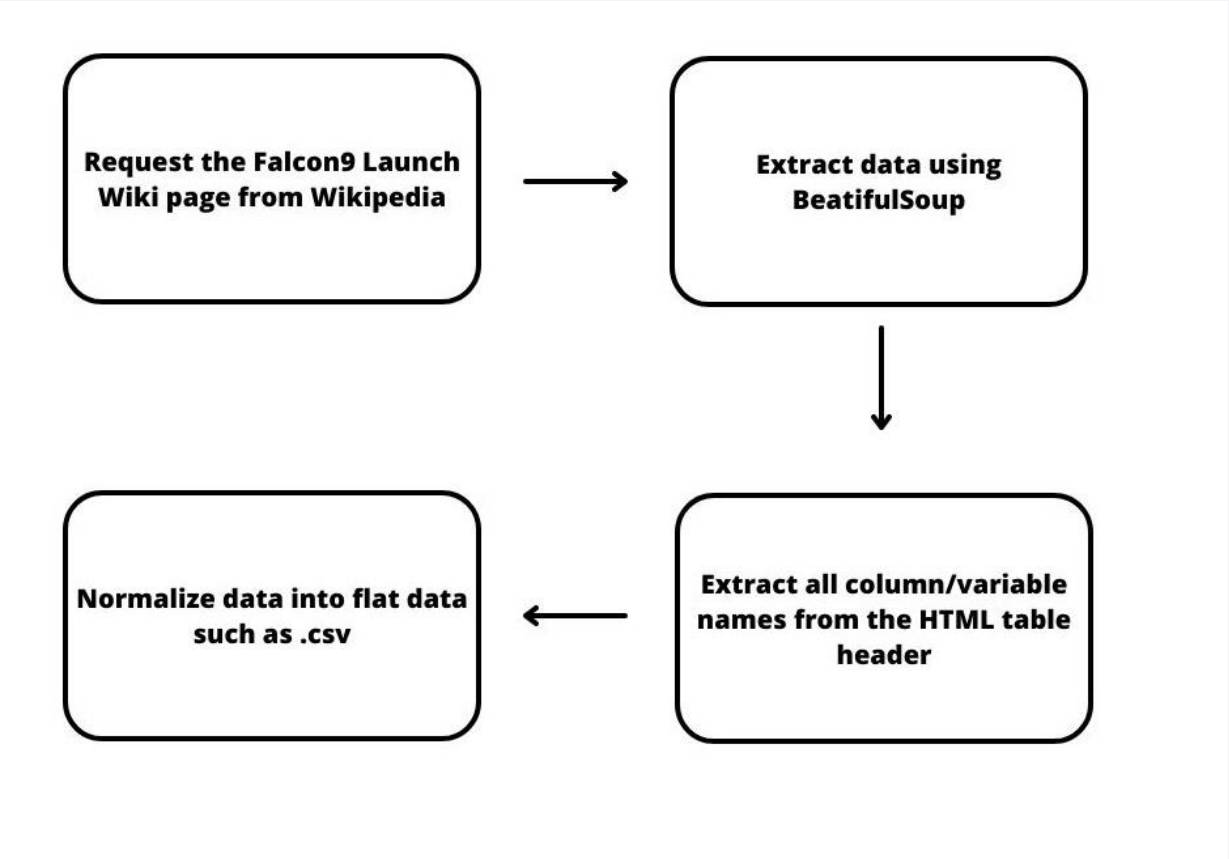
- Make a request to SpaceX API and make sure the data is in the correct format.
- Perform some basic data wrangling and formatting in order to clean the requested data.
- Convert our data frame into a CSV dataset.



- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Data_Collection-SpaceX-API.ipynb

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from Url
- Create a BeautifulSoup from the HTML response
- Extract all column/variable names from the HTML header
- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Data_Collection-SpaceX-Scraping.ipynb



Data Wrangling

- Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA).
- We will first calculate the number of launches on each site, then calculate the number and occurrence of mission outcome per orbit type.
- We then create a landing outcome label from the outcome column. This will make it easier for further analysis, visualization, and ML. Lastly, we will export the result to a CSV

- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Data_Wrangling-SpaceX.ipynb

EDA with Data Visualization

- Data visualization helps us understand data by curating it into a form that's easier to understand, highlighting the trends and outliers. Several types of charts were used in the visualization of the data:
- Cat plots and scatter plots were used to view the relationships of categorical variables like Launch Site and Orbit.
- A bar chart was used to visualize the success rate of each orbit type.
- A line chart was used to visualize the launch success yearly trend

- GitHub link: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/EDA-Data_Visualization-SpaceX.ipynb

EDA with SQL

Using SQL, we had performed many queries to get better understanding of the dataset, Ex:

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string ‘CCA’.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload massgreater 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 failed landing_outcomes in drone ship, their booster versions, and launch sitesnames for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and
- 2017-03-20, in descending order.
- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/EDA_SQL-SpaceX.ipynb

Build an Interactive Map with Folium

- Folium Markers were used to show the SpaceX launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to their nearest land marks.
- Furthermore, Folium Circles were used to highlight circle area of launch sites.
- In order to mark the success/failed launches for each site, marker clusters were used on the map. Whereby Red represents rocket launch failures while Green represents the successes.
- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Interactive_Map_Folium-SpaceX.ipynb

Build a Dashboard with Plotly Dash

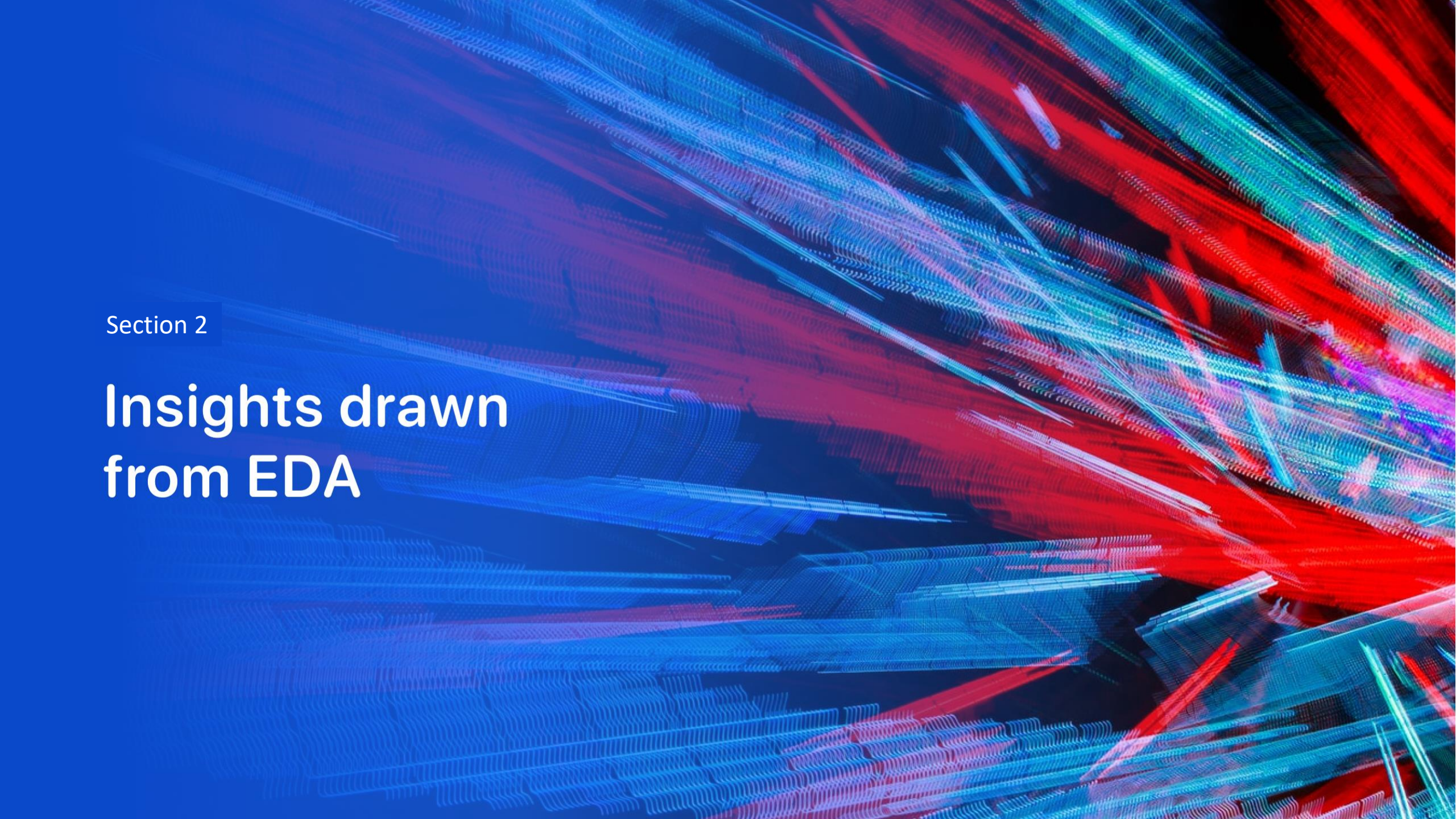
- We built an interactive dashboard with Plotly dash which allowing the user to play around with the data as they need.
- We plotted pie charts showing the total launches by a certain sites.
- We then plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Dashboard_Plotly_Dash-SpaceX.ipynb

Predictive Analysis (Classification)

- BUILDING MODEL
- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets 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
- 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
- Plot Confusion Matrix
- IMPROVING MODEL
- Feature Engineering
- Algorithm Tuning
- FINDING THE BEST 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
- GitHub URL: https://github.com/DerPestarzt/IBM-DataScience-SpaceX-Capstone/blob/main/Predictive_Analysis_Classification-SpaceX.ipynb

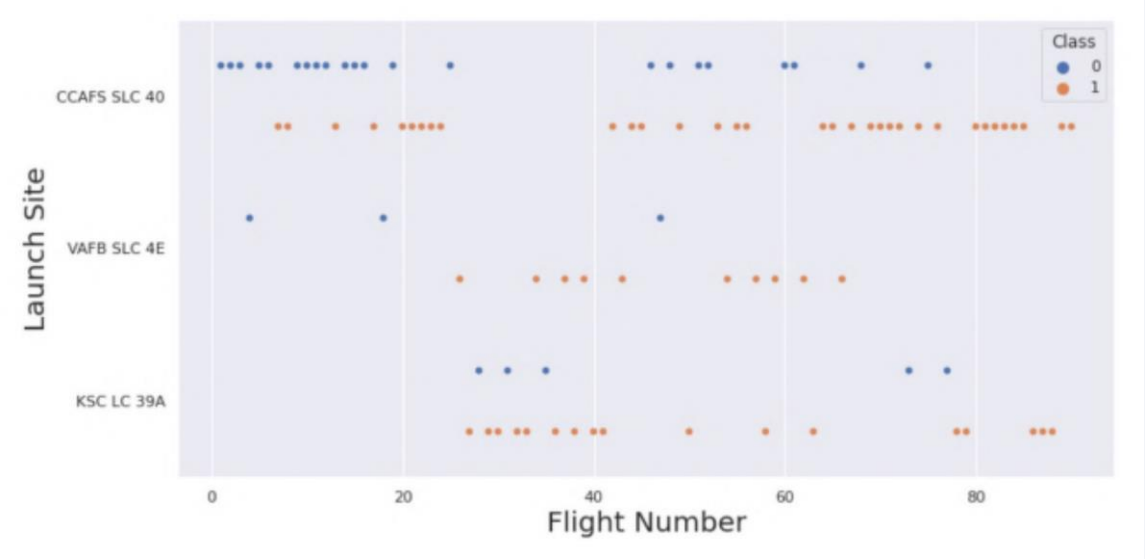
Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- • All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%. This accuracy can be increased in future projects with more data.


Section 2
Insights drawn from EDA

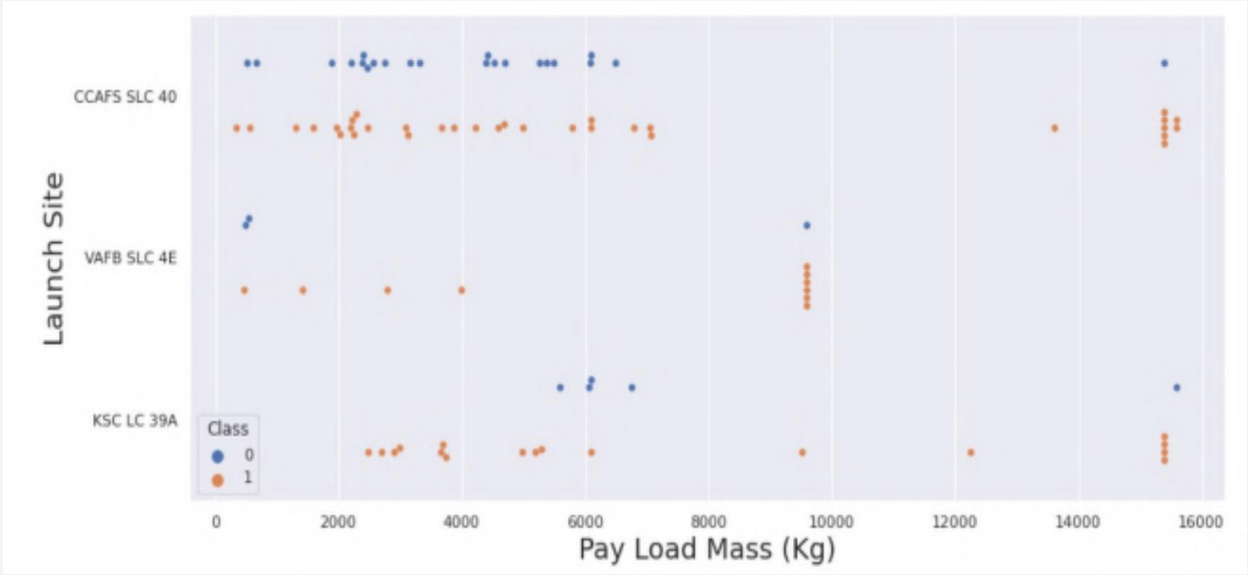
Flight Number vs. Launch Site

- This scatter plot shows that the larger the flights amount of the launch site, the greater the the success rate will be. However, site CCAFS SLC40 shows the least pattern of this.



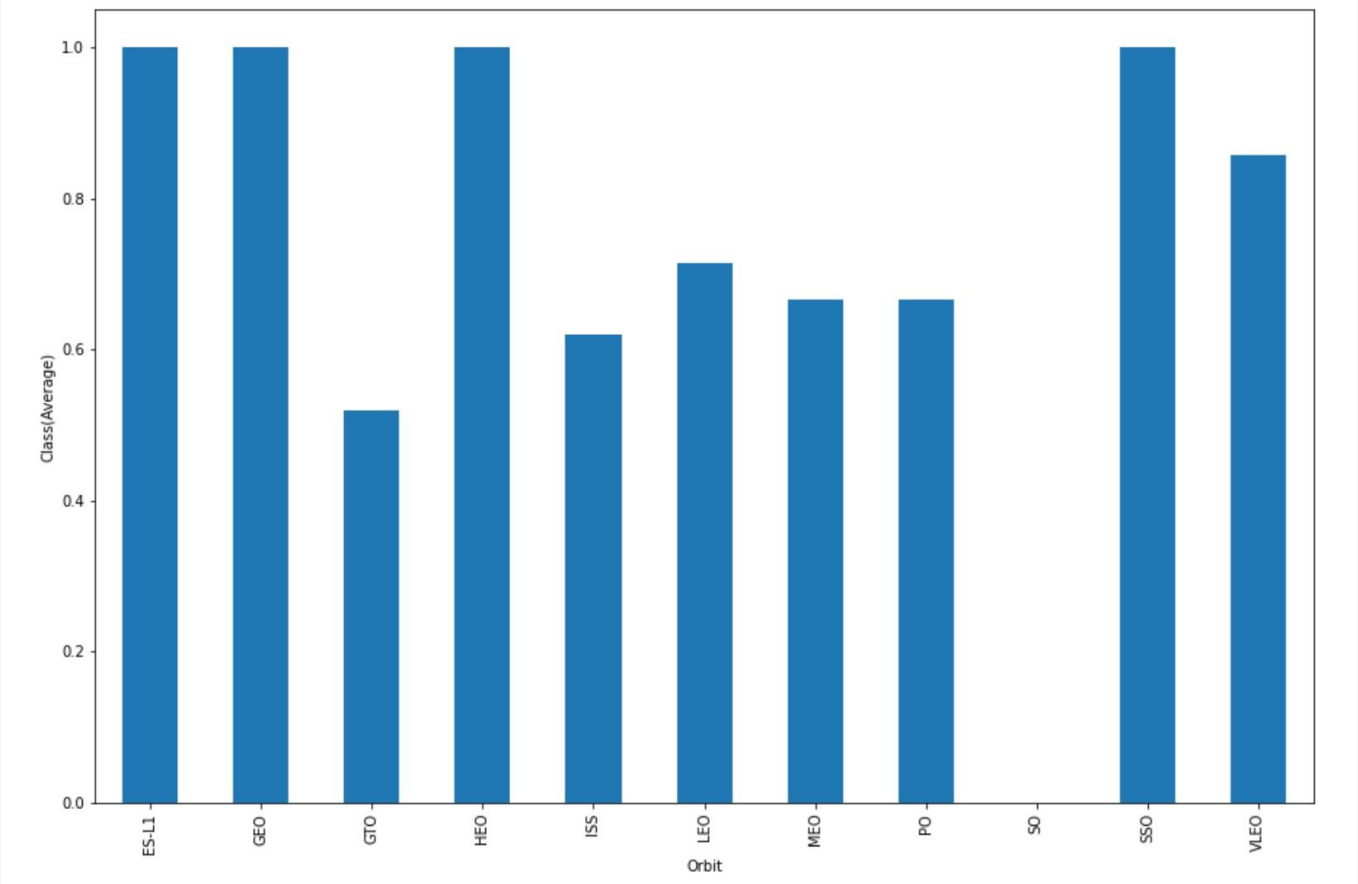
Payload vs. Launch Site

- 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



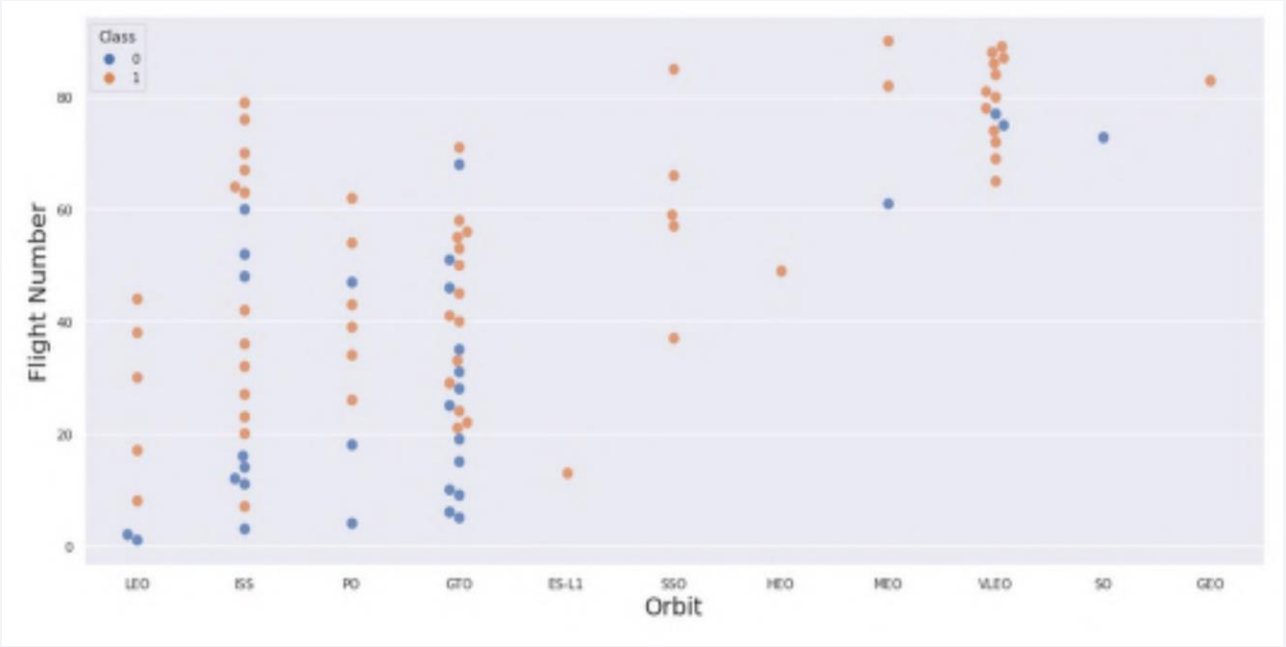
Success Rate vs. Orbit Type

- The orbit types SSO , HEO , GEO and ES -L1 had the highest success rate.



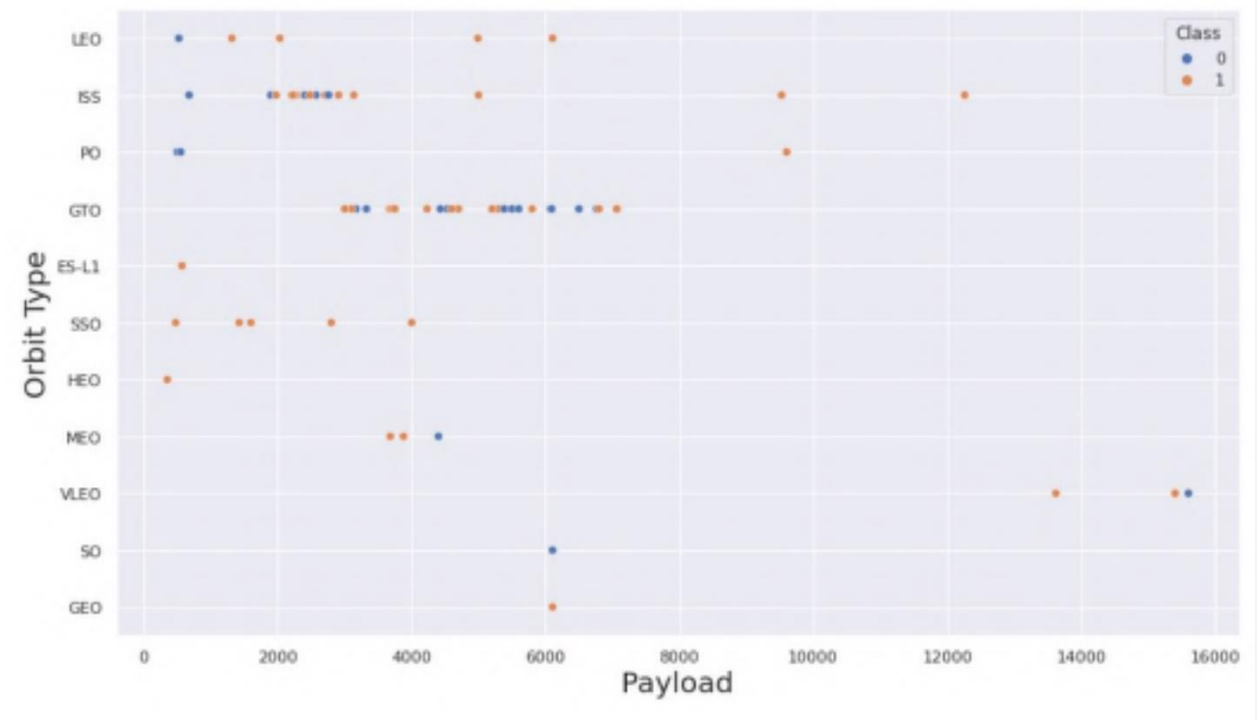
Flight Number vs. Orbit Type

- This scatter plot shows that generally, the larger the flight number on each orbits, the greater the success rate (especially LEO orbit) except for GTO orbit which depicts no relationship between both attributes. Orbit that only has 1 occurrence should also be excluded from above statement as it's needed more dataset.



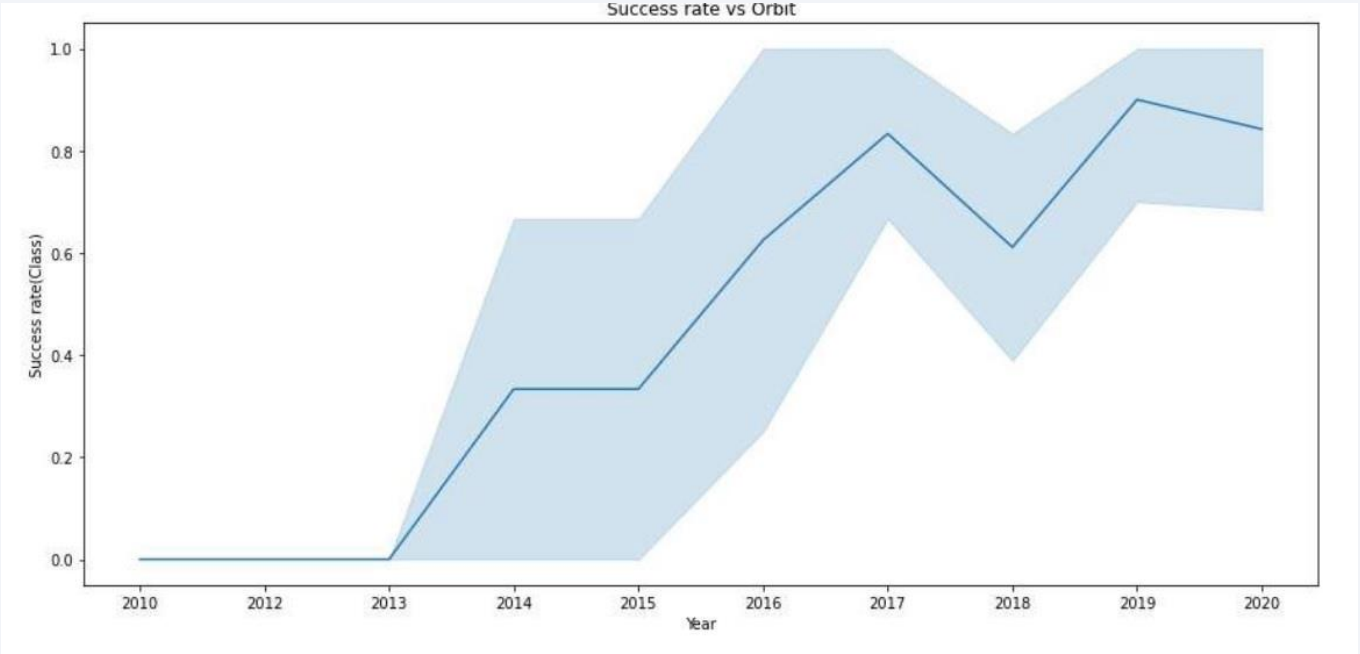
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.



Launch Success Yearly Trend

- This figures clearly depicted and increasing trend from the year 2013 until 2020. If this trend continue for the next year onward. The success rate will steadily increase until reaching 1/100% success rate.



All Launch Site Names

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

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEX;
```

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

Launch Site Names Begin with 'CCA'

- 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	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)
```

Total Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- 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

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass by Booster
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Average Payload Mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

- Use the min() function to find the result, observe that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
%sql SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

First Succesful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- It appears that missions generally tend to be successful with the exception of one failure.

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "Successful Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Success%';
```

Successful Mission
100

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "Failure Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Failure%';
```

Failure Mission
1

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function

Booster Versions which carried the Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- • It appears that 2 boosters failed to land at the beginning of the year.. • The first successful landing took place later that year in December as we saw earlier.

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

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

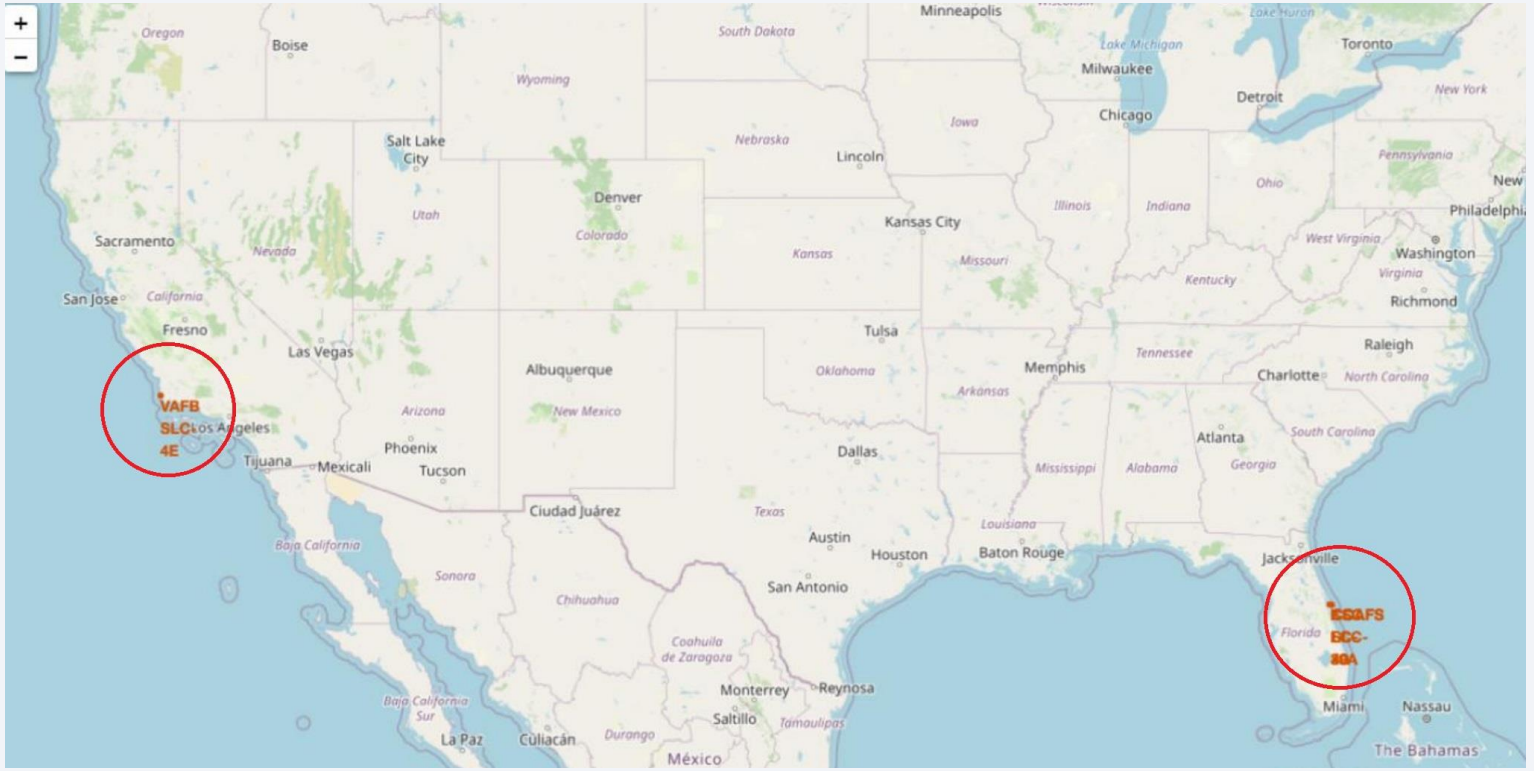
- It appears that 2 boosters failed to land at the beginning of the year.
- The first successful landing took place later that year in December as we saw earlier.

Landing Outcome	Total Count
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

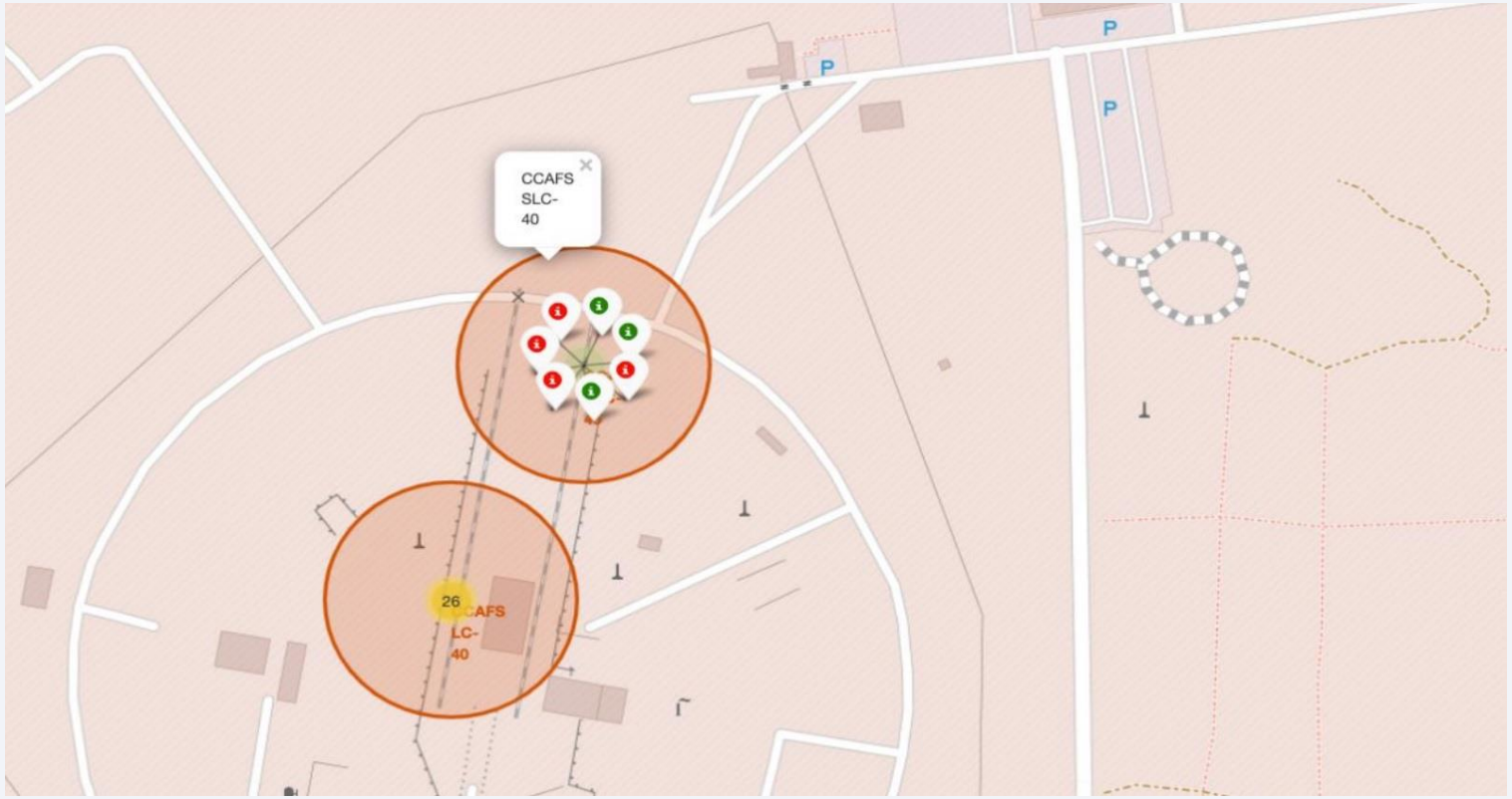
Launch Sites Proximities Analysis

Location of all the Lauch Sites



- We can see that all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.
- It is interesting to see that most launch sites are concentrated near Miami

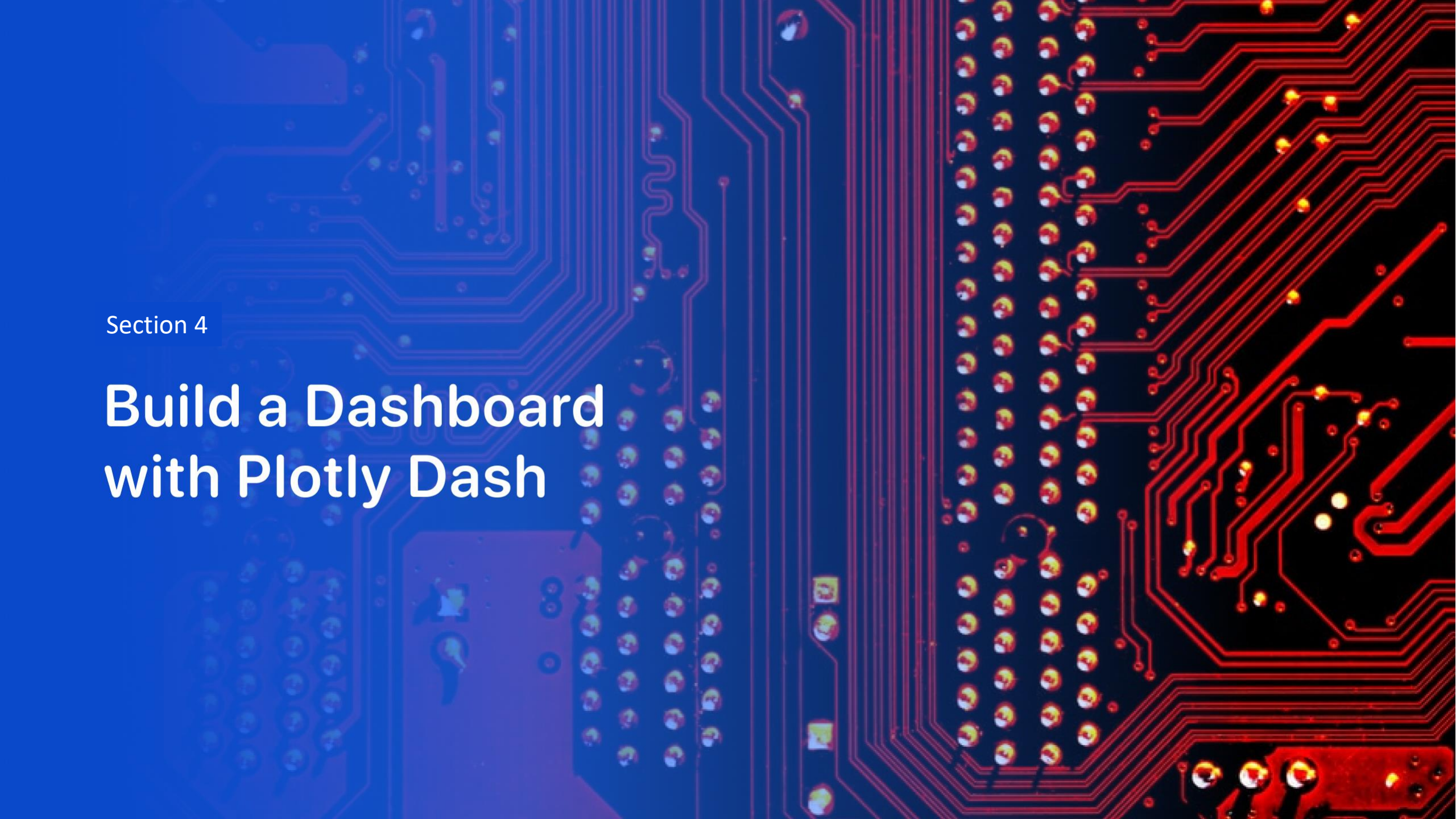
Success rate of Rocket Launches



- The successful launches are represented by a green marker while the red marker represents failed rocket launches.
- It appears that KSC LC39A had the highest success rate of rocket launches compared to other launch sites

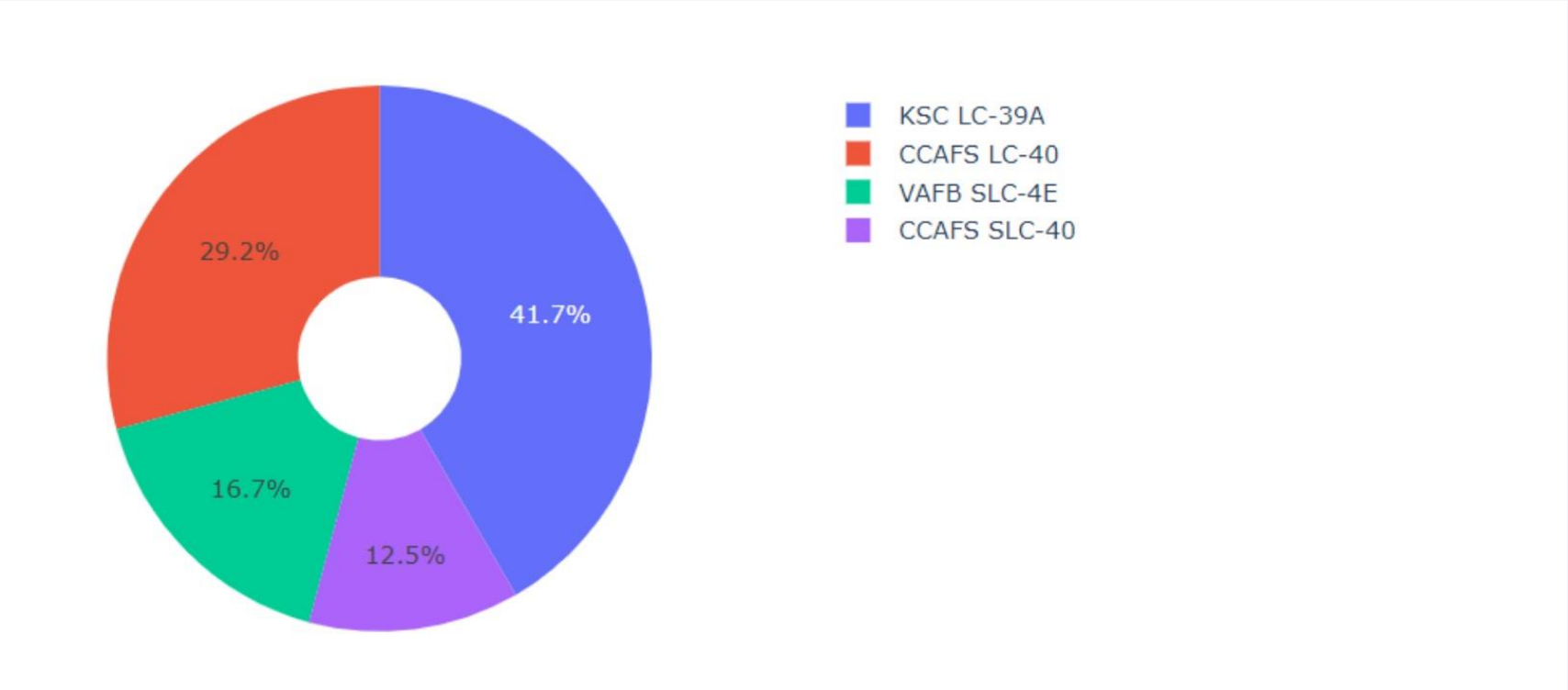
Surrounding Landmarks

- The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean .


Section 4
Build a Dashboard with Plotly Dash

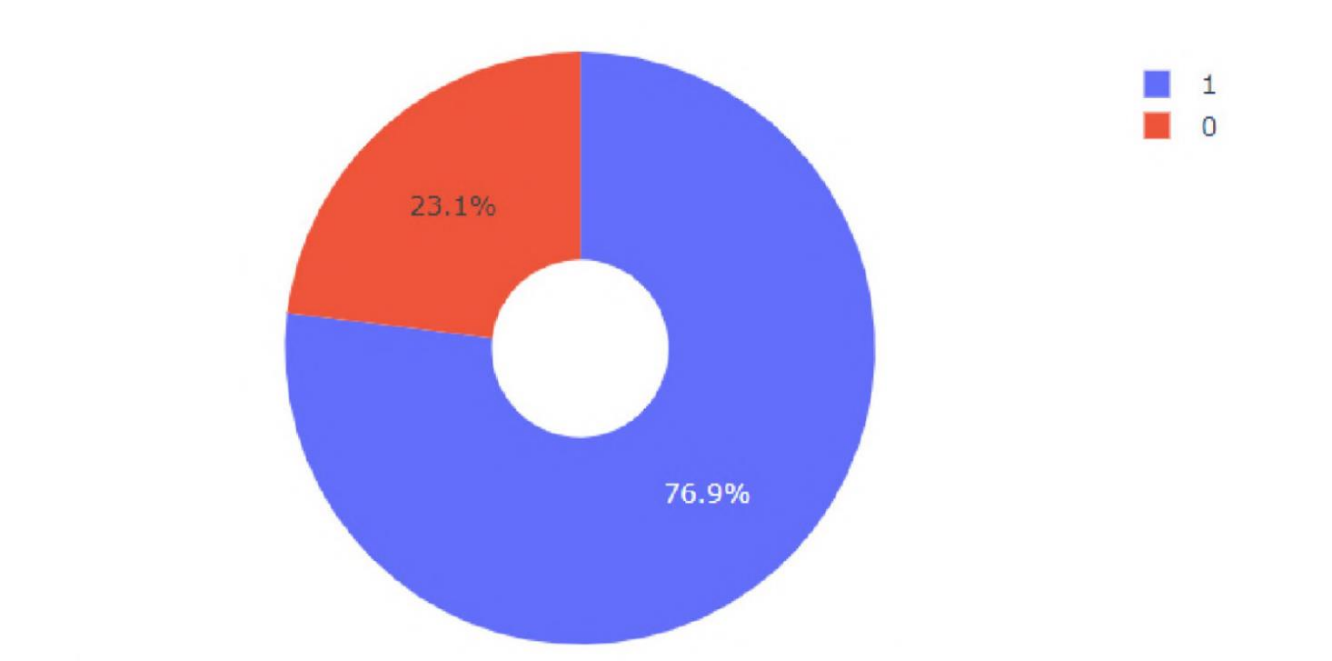
Graphic with success percentage by each sites

- Site KSC LC-39A has the largest successful launches as well the highest launch success rate.
- More investigation may be needed to determine why KSC LC-39A is the preferred launch site.



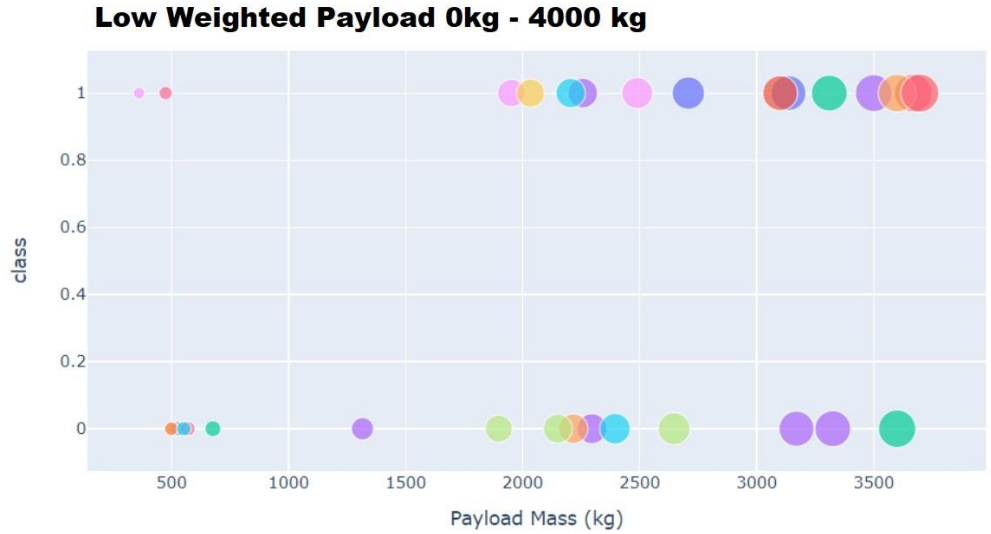
The highest success ratio

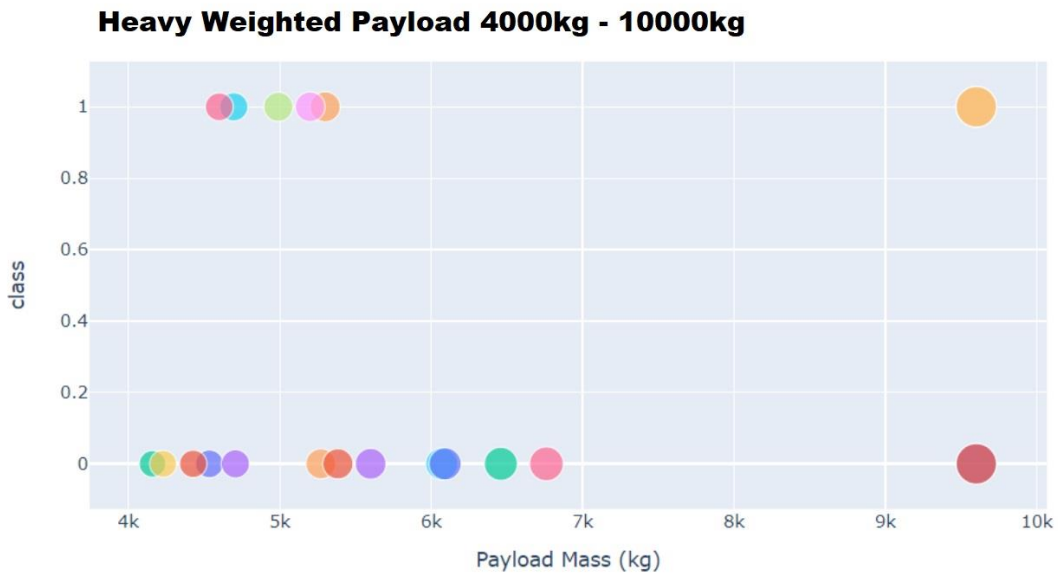
- As we can see, 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.
- However, this success rate was only around 3% higher than the runner up; site CCAFS LC-40.



Payload vs Launch Outcome Scatter Plot

- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was also dramatically low between the payload range of 0kg and 2500kg. Perhaps very low masses decrease launch success.
- The booster version FT, seems to have a higher success rate than other booster versions





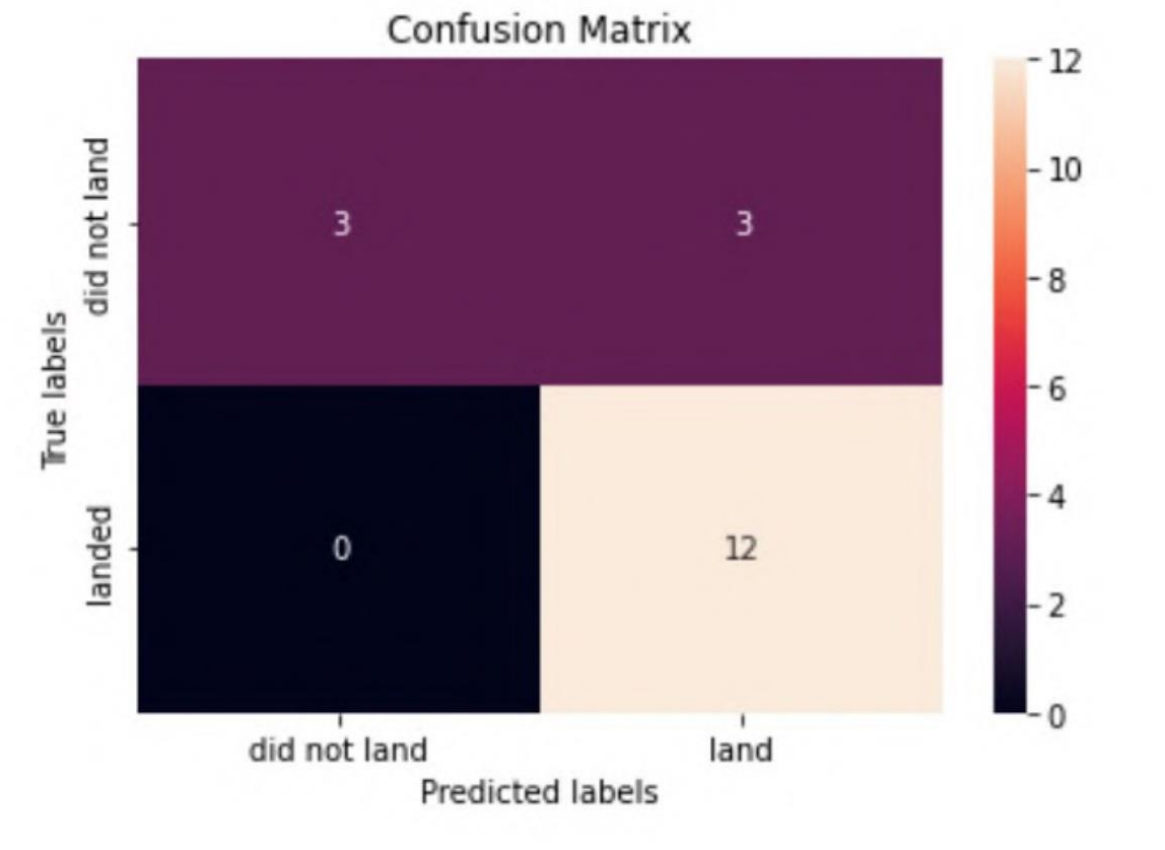
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We could identify that the best algorithm to be the Tree Algorithm which have the highest classification accuracy.

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
- The low weighted payloads (which define as 4000kg and below) performed better than the heavy weighted payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- KSC LC-39A have the most successful launches of any sites; 76.9%
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Appendix

- Interactive Plotly
- Folium MeasureControl Plugin Tool
- Folium Custom Title Layers with Labels
- IBM Cognos Vizualization Tool
- Basic Decision Tree Construction

Thank you!