Data Science

By Hesham Magdy

https://github.com/hesham-elomari/Applied-Data-Science-Capstone

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Techniques that were used to analyze data:**

1. Data collecting using Web Scrapping and Space X API.

2. Exploratory Data Analysis like data wrangling , data visualization ,and interactive visual analytics.

3. Machine Learning Prediction.

- **Results of the techniques that have been used:**

1. Wanted Data can now be seen by visualizing it and how their attributes can relate to each other.

2. Three similar models of the highest accuracy having a value of 83.3% (Support Vector Machines, Decision Tree, K Nearest Neighbors) for doing the best predictions so far.

Introduction

- **Companies** are making space travel more affordable such as Space X which is owned by Elon Musk. Multiple things that Space X do make it more successful than other companies such as having better prices for rocket ,and the ability to recover these rockets.

- **Space Y** is a company that is trying to compete with Space X. Instead of using rocket science , Space Y will be hiring us as a data scientist to gather data about Space X in order to determine launch price ,and the reuse of the first stage if its successful or not.

Methodology

- **Perform Data Collection:**

Collect data from SpaceX public API and use Web Scrapping of SpaceX Wikipedia page.

- **Perform Data Wrangling:**

Analyze the features of the data , we were able to identify the landing outcome of the first stage.

- **Perform exploratory data analysis (EDA) using visualization and SQL**

- **Perform interactive visual analytics using Folium and Plotly Dash**

- **Perform predictive analysis using classification models:**

Tuning classification models to the choose the best one for predicting output

Data Collection Space X API

1. Request Space X API
2. Normalize json data to a dataframe
3. Choose necessary features for dataframe
4. Filter the dataframe to only include Falcon 9 launches
5. Deal with missing values

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Data_Collection_API.ipynb -

Data Collection With Web Scraping

- 1. Request Falcon9 Launch wiki page
- 2. Extract all column/variable names from the HTML table header using BeautifulSoup
- 3. Create a data frame by parsing the launch HTML tables

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Data_Collection_with_Web_Scrapping.ipynb

Data Wrangling

1. Exploratory Data Analysis is performed
2. Calculated the number and occurrence of each orbit
3. Calculated the number and occurrence of mission outcome per orbit type
4. Created a landing outcome label from Outcome column

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Data_Wrangling.ipynb

EDA with SQL

- **SQL queries that have been done**:

1. Names of the unique launch sites in the space mission
2. Top 5 records where launch sites begin with the string 'CCA'
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date when the first successful landing outcome in ground pad was achieved.
6. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Total number of successful and failure mission outcomes
8. Names of the Booster Versions which have carried the maximum payload mass.
9. Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
10. Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Explanatory_Data_Analaysis_Using_SQL.ipynb

EDA with Visualization

- **Visualizations performed:**

1. Relationship between Flight number and Launch Site
2. Relationship between success rate of each orbit type
3. Relationship between Flight number and Orbit type
4. Relationship between Payload and Orbit type
5. The launch success yearly trend

- **Performed feature engineering on important variables that would affect the success rate**

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Explanatory_Data_Analysis_With_Visualization.ipynb

Interactive Map with Folium

- **Markers, circles, lines and marker clusters were used with Folium Maps**:

1. Markers indicate points like launch sites;
2. Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Centre
3. Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
4. Lines are used to indicate distances between two coordinates.

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/Interactive_Visual_Analytics_With_Folium.ipynb

Interactive Dashboard with Poly

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot shows : all sites or individual site and payload mass on a slider between 0 and 10000 kg.

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Prediction Analysis by Classification

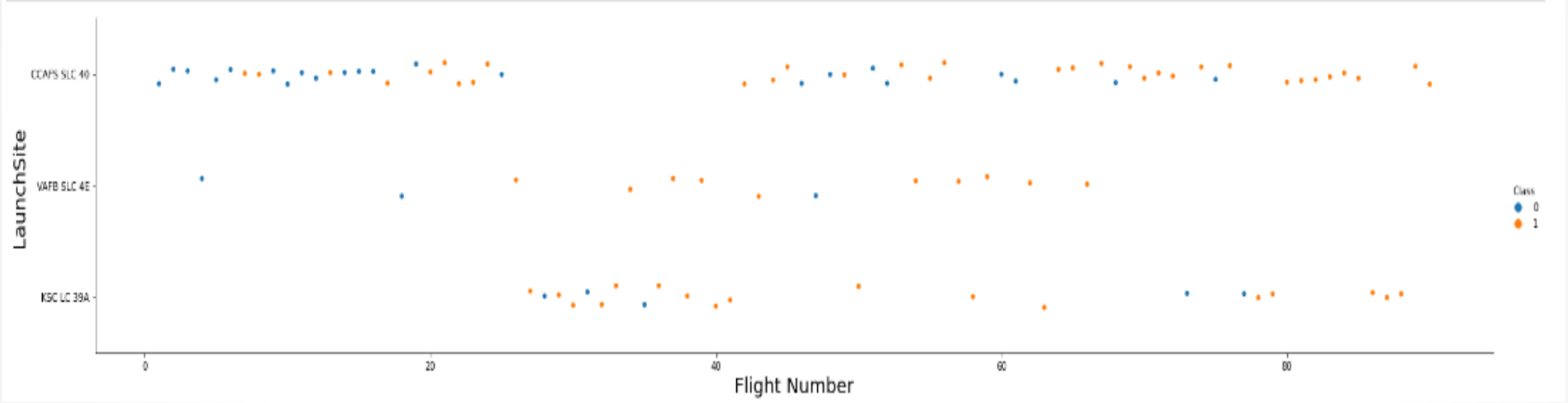
We preprocessed the data then tested to see which of the four models (Decision Tree, K Nearest Neighbor, Support Vector Machine ,and Logistic Regression) along with trying different hyper parameters for each model to achieve the best accuracy out of the model.

https://github.com/hesham-elomari/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

EDA with visualization

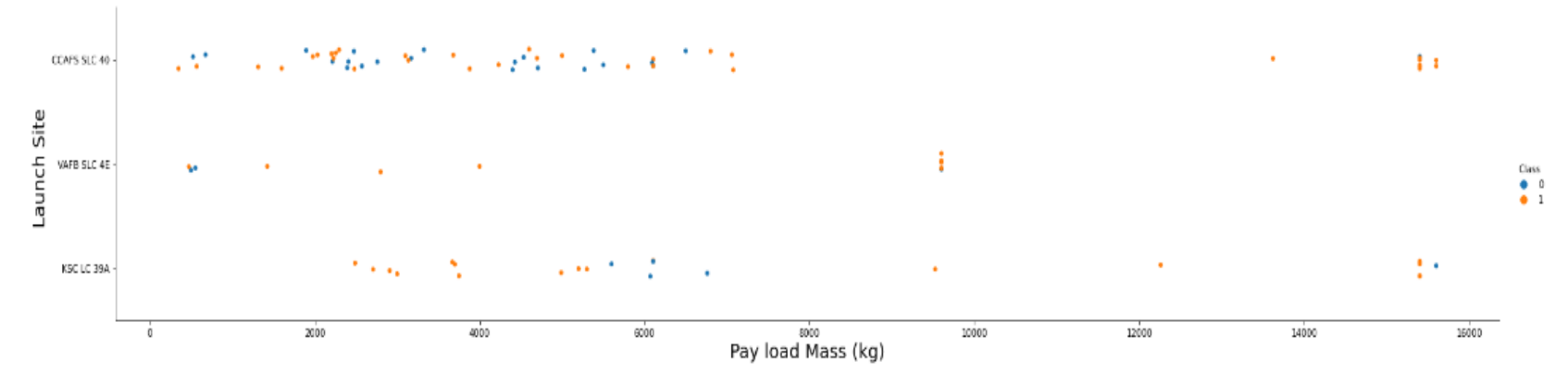
- Relationship between **Flight Number and Launch Site**



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
- You can also see that the general success rate improved over time.

EDA with visualization

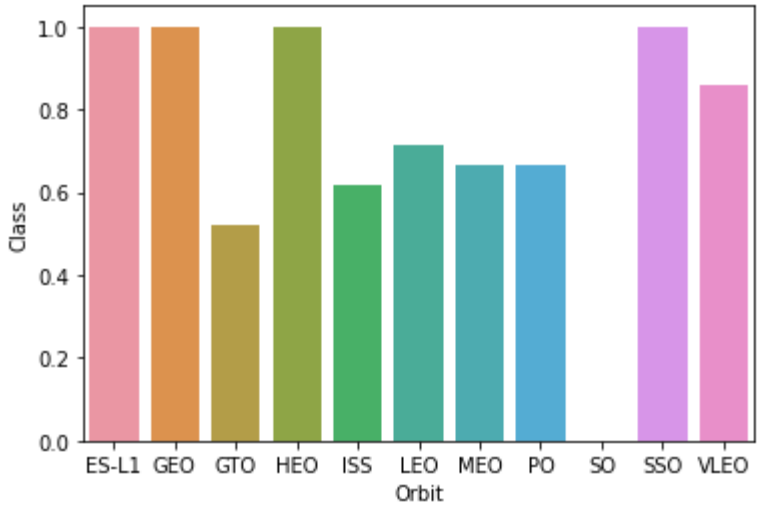
- Relationship between **Payload and Launch Site**



- Payload mass appears to fall mostly between 0-6000 kg.
- Payloads over 10,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

EDA with visualization

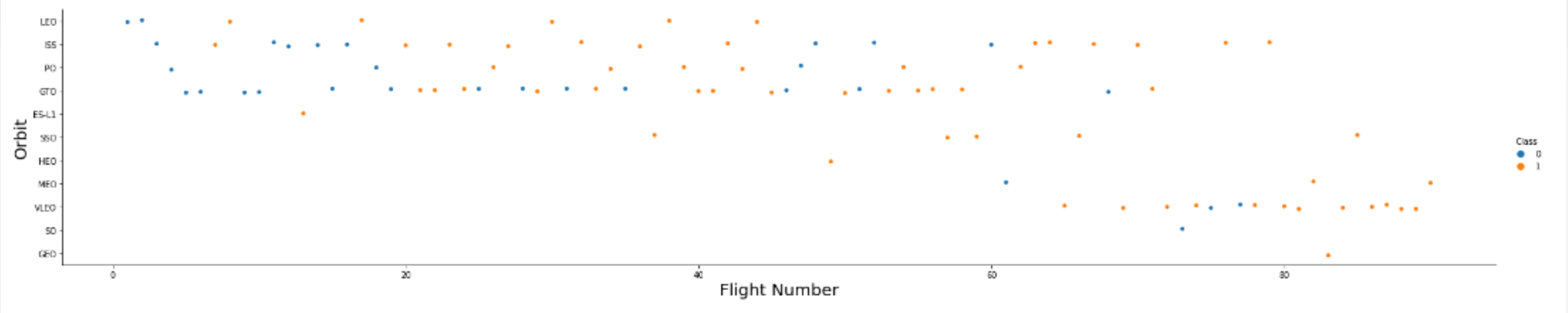
- Relationship between **Success rate of each orbit type**



- ES-L1, GEO, HEO, SSO has a success rate of 100%
- GTO is approximately above 50%
- ISS is almost 63%
- LEO is almost 70%
- MEO and PO are almost 66%
- SO is 0%
- VLEO is almost 85%

EDA with visualization

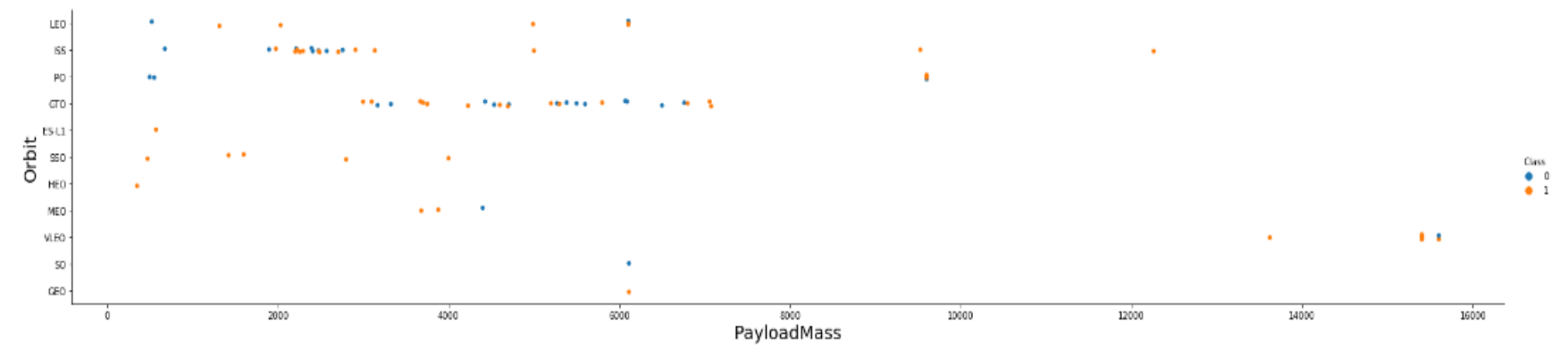
- Relationship between **Flight Number and Orbit Type**



- LEO success increases with flight number
- VLEO becomes better than the others when there are more than 60 flights

EDA with visualization

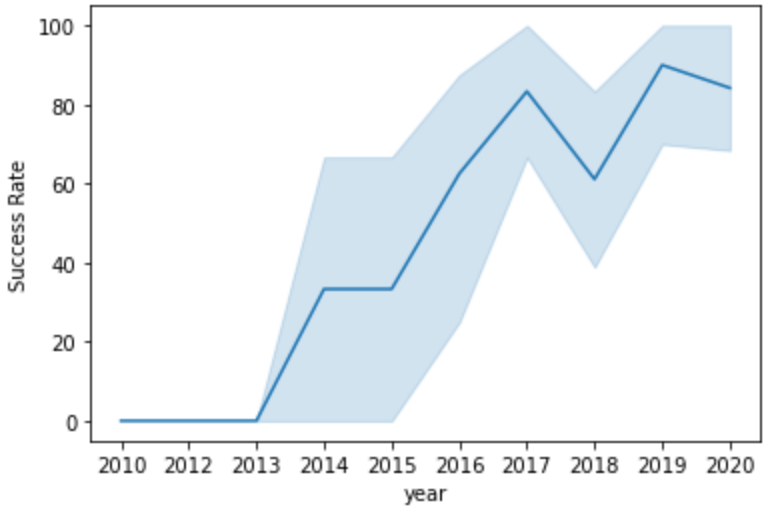
- Relationship between **Payload mass and Orbit type**



- There are more Orbits when Payload mass is less than 8000

EDA with visualization

- **Launch success yearly trend**



- After year 2013, Success rate increases over time by a huge amount (from 0-80 percent)

EDA with SQL

- **Launch Sites Unique Names**

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

- **5 samples of Launch sites that begin with the string 'CCA'**

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

EDA with SQL

- **Total payload mass carried by boosters launched by NASA (CRS)**

Total_Payload_Mass
111268

- **Average payload mass carried by booster version F9 v1.1**

Average_Payload
2928.4

- **Date when the first successful landing outcome in ground pad was achieved.**

DATE
01-05-2017

EDA with SQL

- **Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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**

Mission_Outcome	Number of successful and failure
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

EDA with SQL

- **Names of the booster versions which have carried the maximum payload mass**

Booster_Version
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

- **The month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.**

Months	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

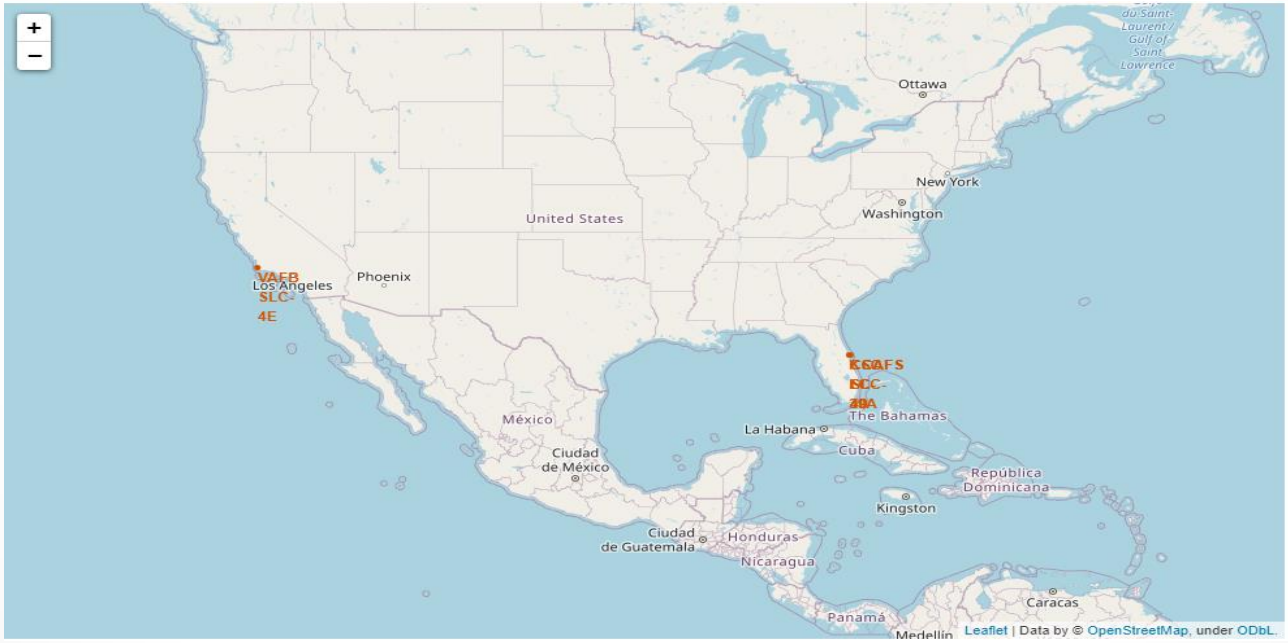
EDA with SQL

- **Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017**

Landing _Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Interactive Map with Folium

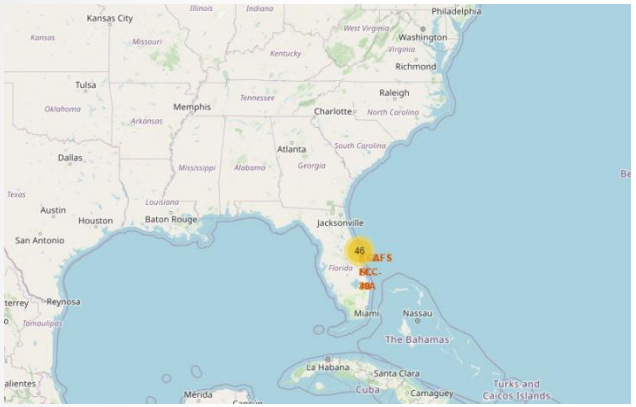
- **All launch sites on a map**

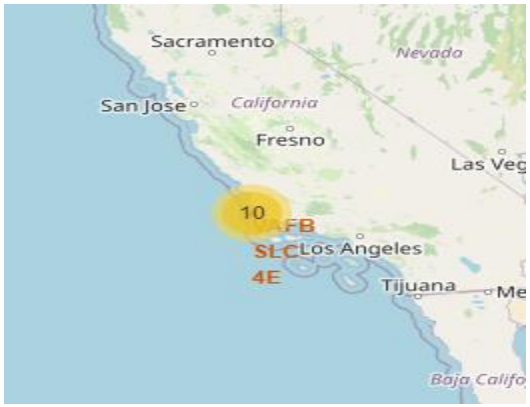


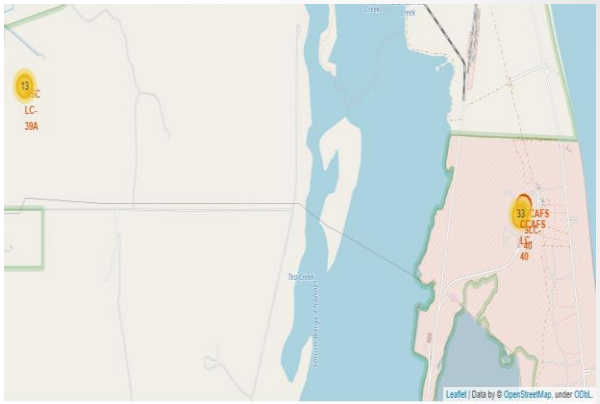
- **As you can see, all sites are near sea for safety**

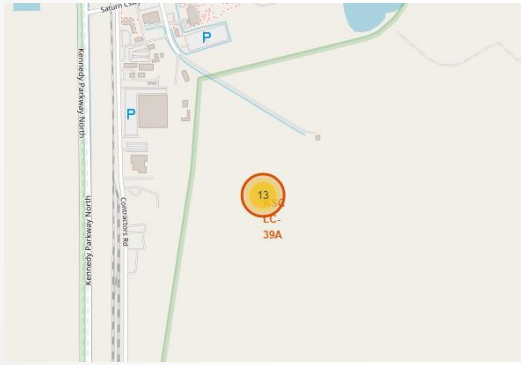
Interactive Map with Folium

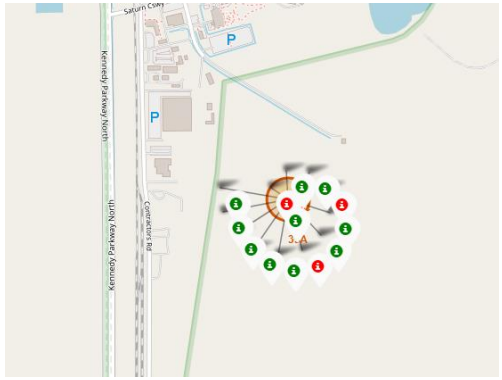
- Example of KSC LC-39A launch site launch outcomes





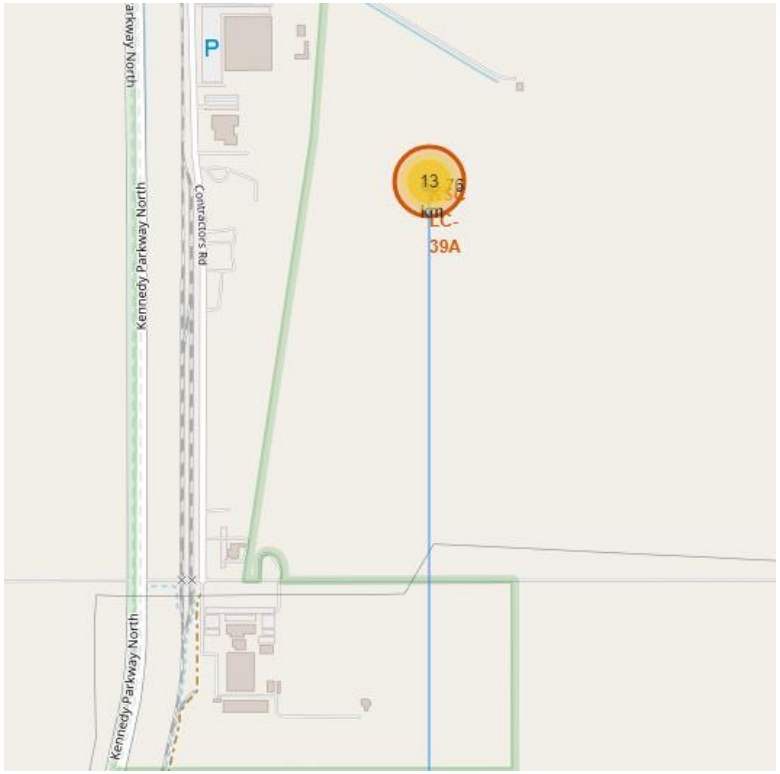






- Green markers indicate successful and red ones indicate failure.

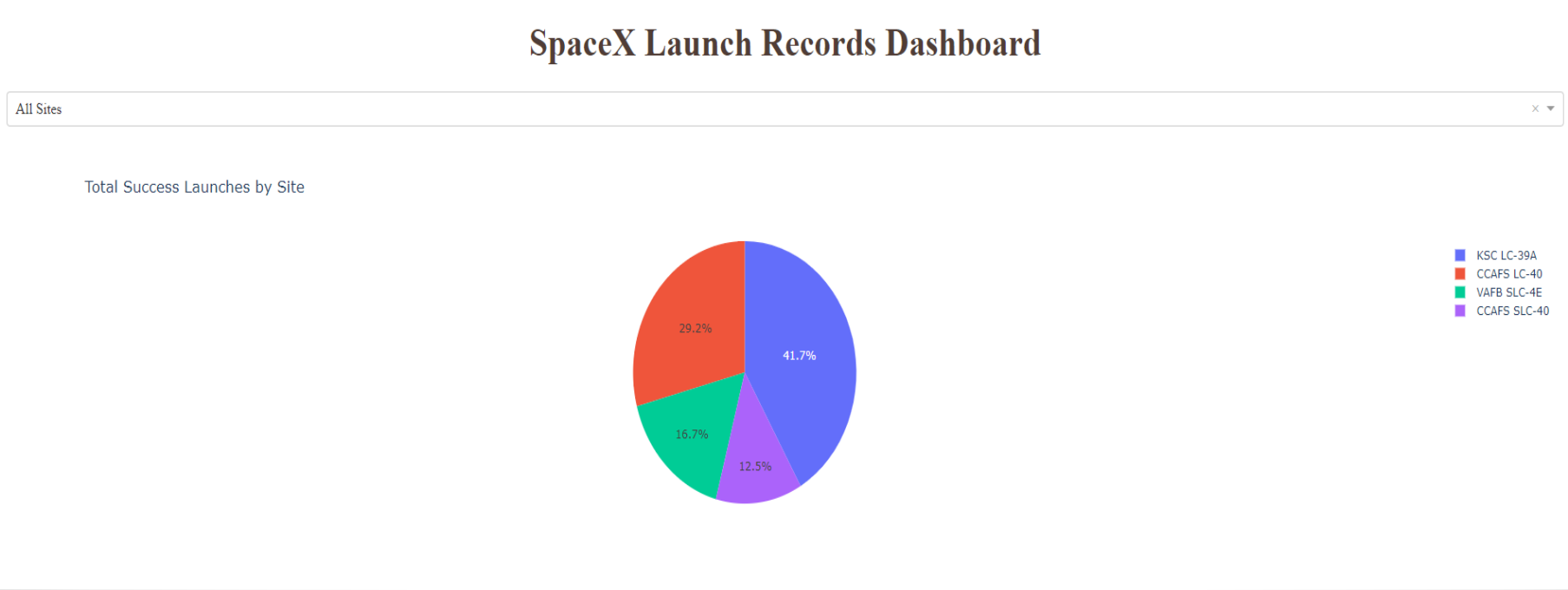
Interactive Map with Folium



- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Plotly Dash dashboard

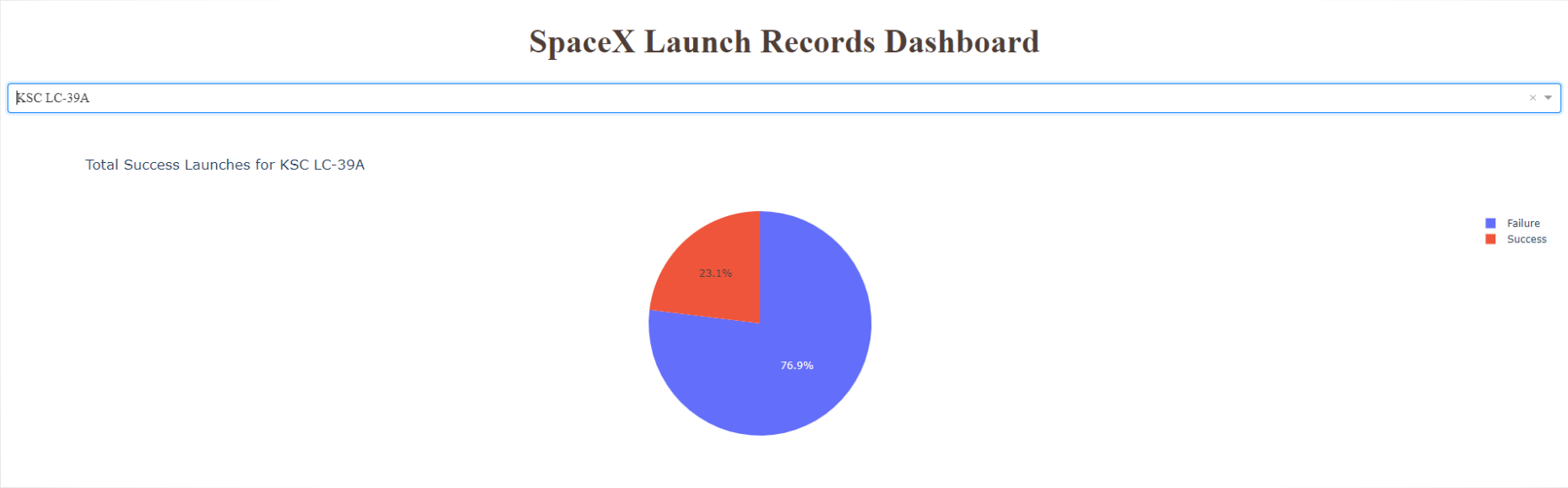
- **Successful Launches by Site**



- **The place from where launches are done seems to be a very important factor of success of missions.**

Plotly Dash dashboard

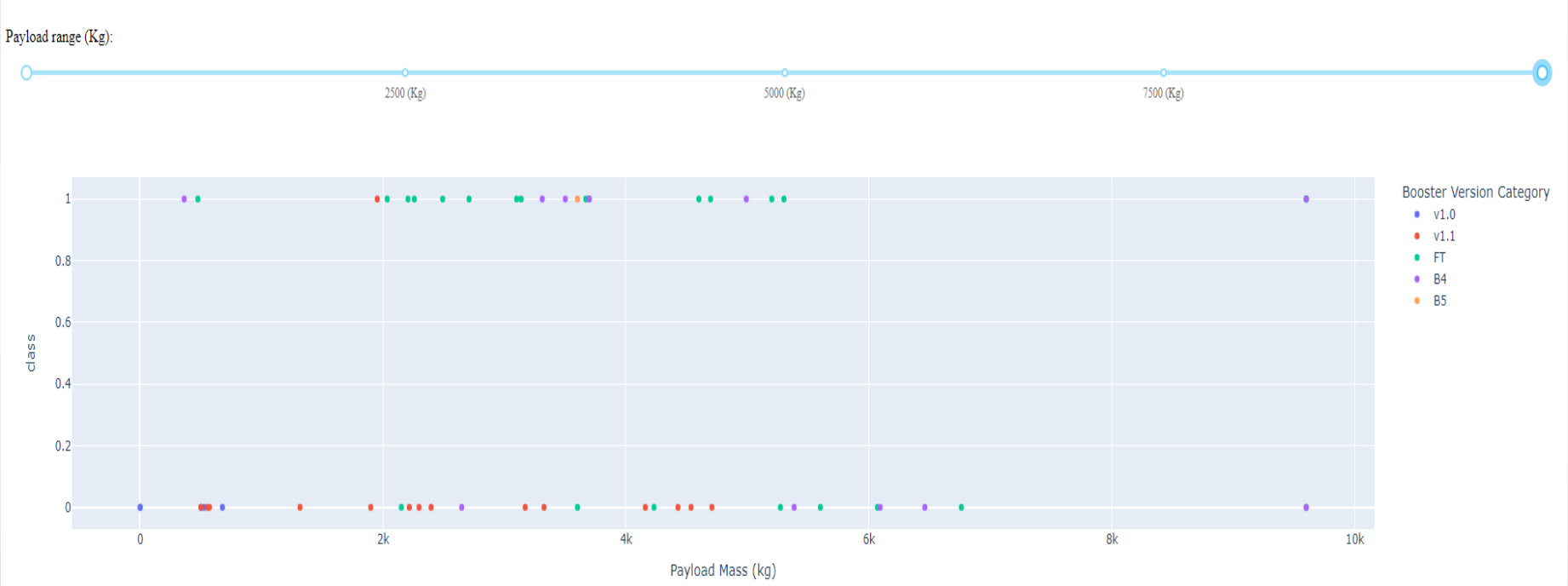
- **Launch Success Ratio for KSC LC-39A**



- **76.9% of launches are successful in this site.**

Plotly Dash dashboard

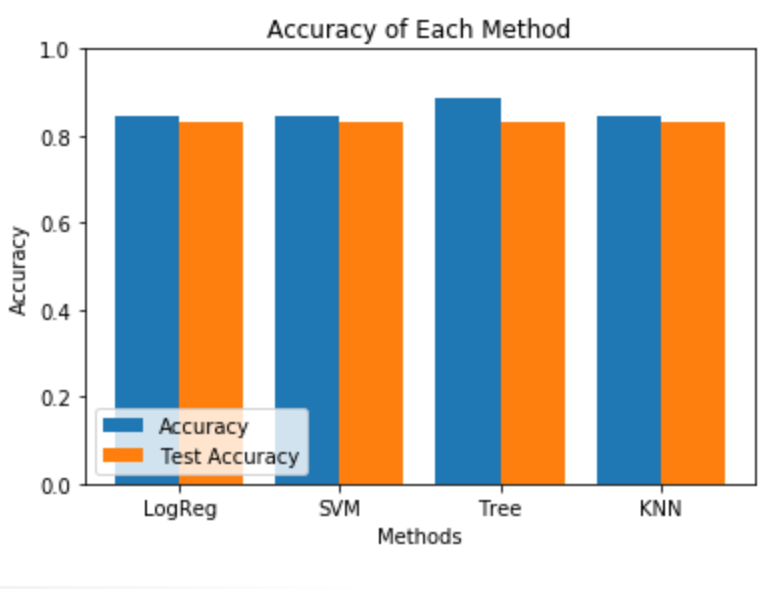
- **Launch Success Ratio for KSC LC-39A**



- **Payloads under 6,000kg and FT boosters are the most successful combination.**

Prediction Analysis by Classification

- K Nearest Neighbors, Support Vector Machine, Decision Tree, and Logistic Regression have been used

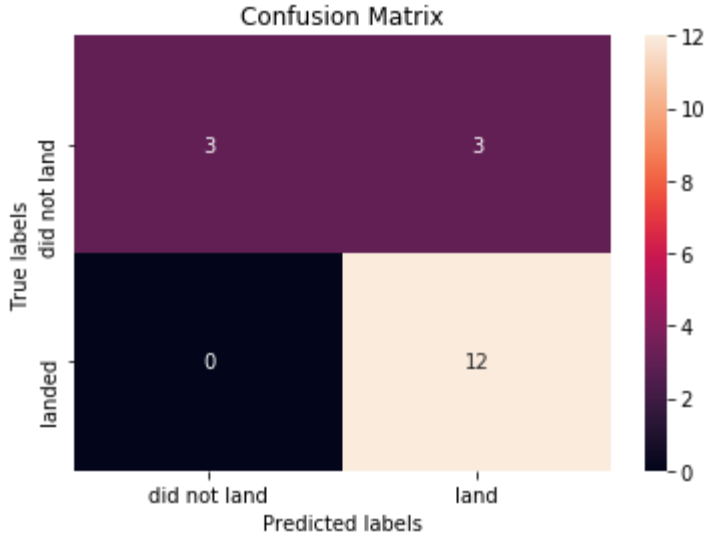


Model	Accuracy	TestAccuracy
LogReg	0.84722	0.83333
SVM	0.84722	0.83333
Tree	0.88889	0.83333
KNN	0.84722	0.83333

- All have the same test accuracy ,
- Decision Tree has the highest accuracy score

Prediction Analysis by Classification

- Since all have the same test accuracy , their confusion matrix will also be the same



- All have the same test accuracy ,
- Decision Tree has the highest accuracy score

Conclusion

- **The best launch site is KSC LC-39A**
- **Payload mass appears to fall mostly between 0-6000 kg.**
- **Decision Tree Classifier can be used to predict successful landings and increase profits.**
- **More data needs to be collected so that the machine learning models would have better accuracy**

APPENDIX

- **GITHUB LINK:**

https://github.com/hesham-elomari/Applied-Data-Science-Capstone