IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Deepak Singh
25/03/2022

Outline

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

Executive Summary

- Summary of methodologies

i. Data Collection (API)

ii. Data Collection (Web Scraping)

iii. Data Wrangling

iv. EDA with Data Visualization

v. EDA with Data Visualization

vi. Build an Interactive Map with Folium

vii. Build a Dashboard with Plotly Dash

- Summary of all results

i. Predictive Analysis (Classification)

ii. Results

Introduction

- Project background and context

Throughout history, space programs have always been of high costs. This is not only due to the high levels of research and development, but also due to the wastage that occurs during the space missions. Space X has become a staple aerospace company due low costs which has been directly caused by Space X technology allowing them to reuse the first stage of rocket. Other competitors in the modern age can charge up to $162,000,000, whereas a Space X launch price is around the region of $62,000,000. It is vital for the initial stage of the mission recovery to be successful to keep the costs down. This project entails of the use of Machine Learning to allow the predictions of whether the first stage will land successfully.

- Problems you want to find answers

i. Factors that have an affect on the mission success.

ii. Environmental and External operating conditions needed for a successful mission.

Section 1

Methodology

Methodology

Executive Summary

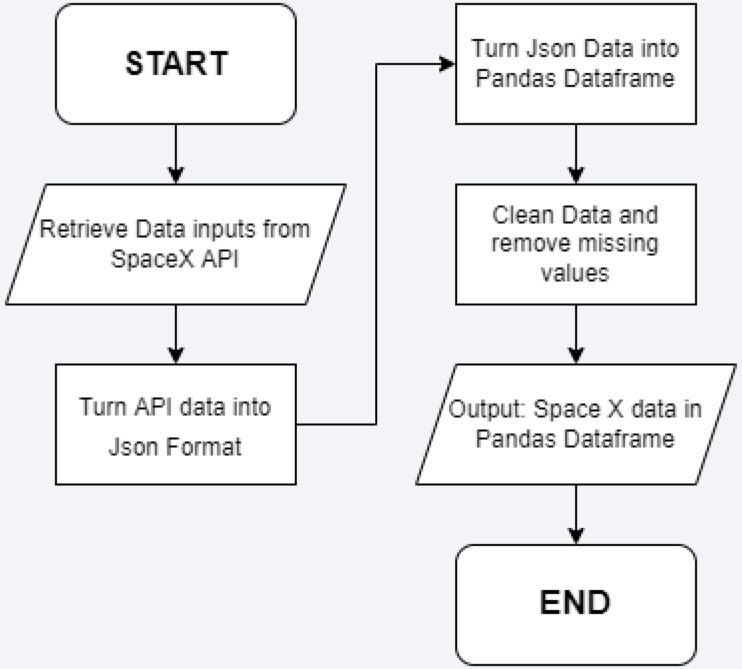
- Data collection methodology:
  - Collection of data was uploaded to SpaceX API. Also uploaded to Wikipedia, web scraping occurred.
- Perform data wrangling
  - One-hot encoding applied to the categorial features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The initial data collect was done using Space X API.
- SpaceX API data was stored via Json format; therefore, the python Json Library was used to turn the data from Json format to a Pandas Dataframe.
- Preprocessing on all the different methodologies occurred which allowed for data to be cleaned.
- Web scraping methods were used, most notably the Beautiful Soup Library on python which allowed for data on Wikipedia to be retrieved.

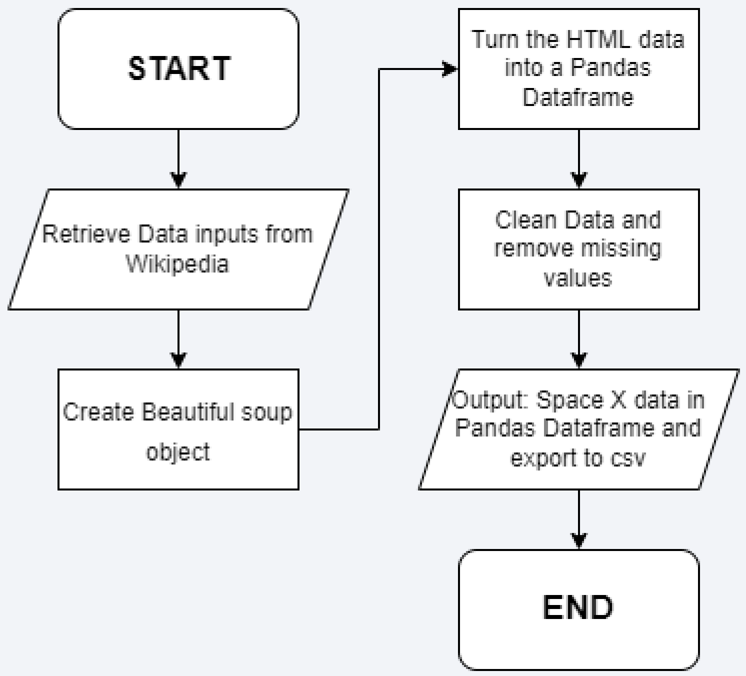
Data Collection – SpaceX API

- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/jupyter-labs-spacex-data-collection-api%20(2)%20(1).ipynb



Data Collection - Scraping

- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) was performed to find patterns.
- The training supervisor models were also determined
- The number of launches were calculated at each site. Where each launch site was a dedicated orbit. Examples include:

i. Low Earth Orbit (LEO)

ii. Geosynchronous Orbit (GTO)

- The number and occurrence of each orbit was done using the Pandas function ".value_counts()"
- The final data wrangling stages encompassed calculating the number and occurrence of mission outcome per orbit type and also creating a landing outcome label.
- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

Data Visualization allowed for analysis containing plots:

i. Flight Number vs Payload Mass (Catplot)

ii. Flight Number vs Launch Site (Catplot)

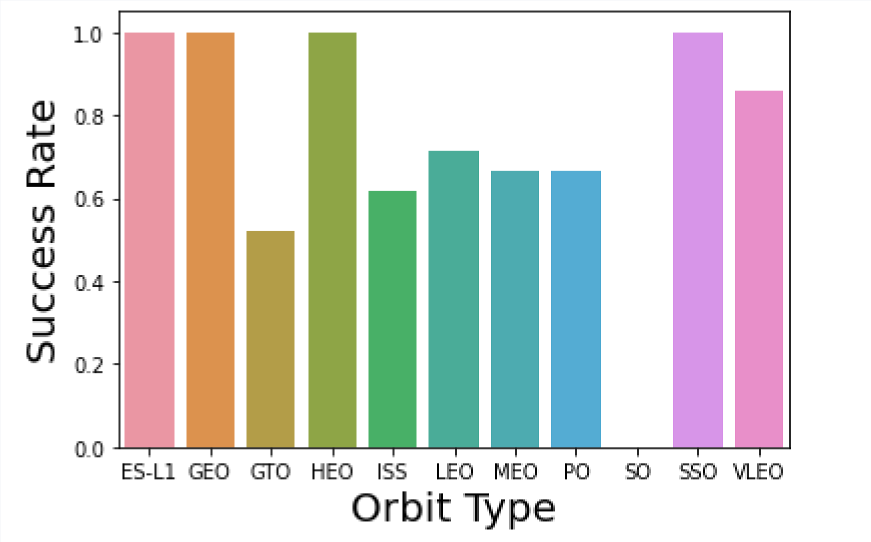
iii. Payload Mass vs Launch Site (Catplot)

iv. Launch Site Success rate Bar Chart

v. FlightNumber vs Orbit type (Catplot)

vi. Payload Mass vs Orbit tpe (Catplot)

vii. Success Rate by Year (Line Chart)



https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

There were 10 Dataset Queries including:

i. Display the names of the unique launch sites in the space mission

ii. Display 5 records where launch sites begin with the string 'CCA'

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

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

v. List the date when the first successful landing outcome in ground pad was achieved.

vi. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

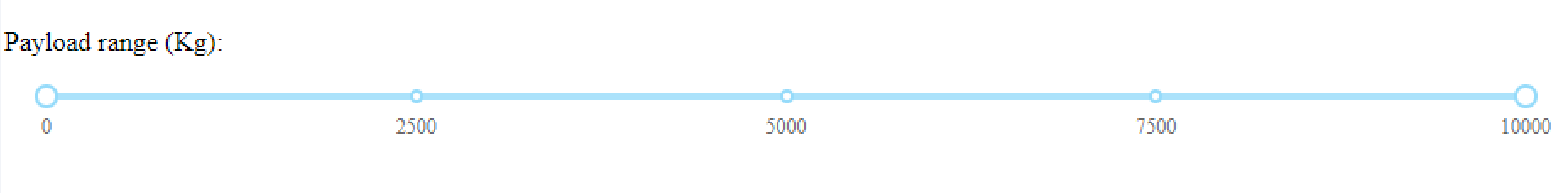
- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/jupyter-labs-eda-sql-coursera%20(2)%20(2).ipynb

Build an Interactive Map with Folium

- Launch sites were marked with circles to, markers and lines. These were used to visually display whether the launch sites were successful.
- Coloring of the different markers/ circles were used to differentiate between whether outcomes were successful or not.
- Lines were used to display the distances of important proximities for instance, the launch sites to the coast and the launch sites to the main highways.
- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

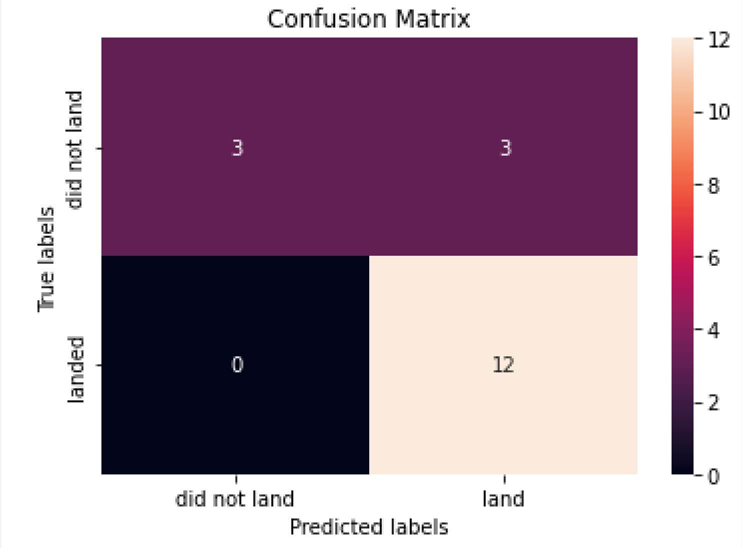
- Pie Chart showing the total launches by selective sites.
- Interactive Scatter graph displaying the relationship between Launch Outcome and Payload Mass (kg)
- Interactive Scatter graph includes a slider ranging from 0 – 10000 Kg



- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data is imported from IBM cloud. Where Numpy and Pandas Libraries are used to prepare data for training and testing, which requires the data to be standardized.
- Multiple parameters for the different Machine Learning models are investigated. A GridSearchCV object was used to identify the best parameters.
- Machine Learning models include Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifer, K Nearest Neighbors (KNN).
- Confusion Matrix for each of the models were made.
- https://github.com/DeepakTSingh1998/Capstone-Project-IBM/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb



Results

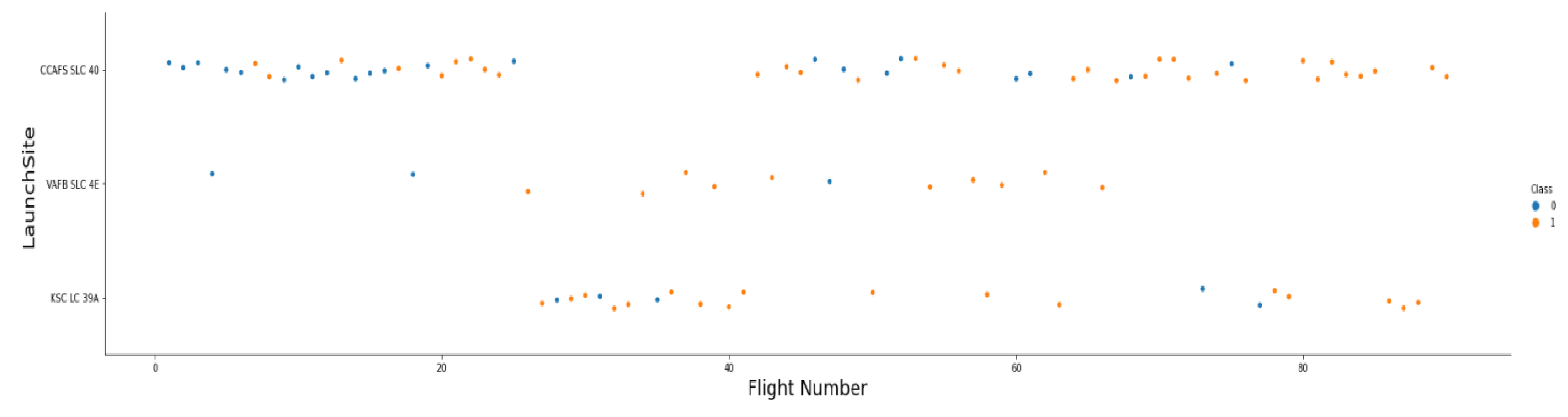
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

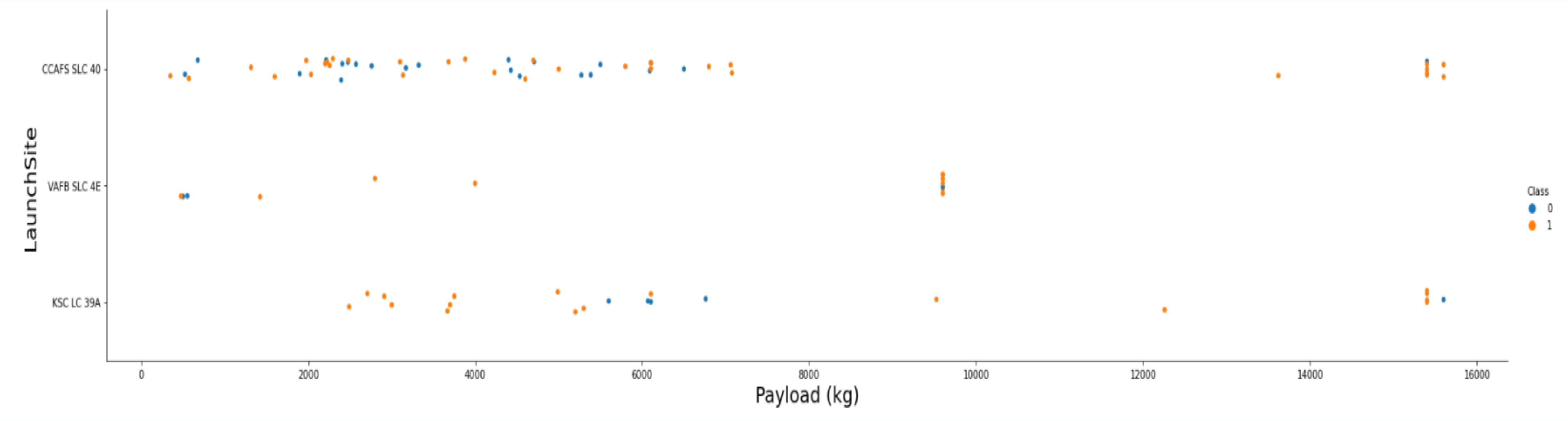
Flight Number vs. Launch Site

- Scatter plot displays the Flight number with the Launch site. Displaying if the mission was successful (1) or a failure (0)
- VAFB SCL 4E launch site seems to have the highest ratio of mission failures.



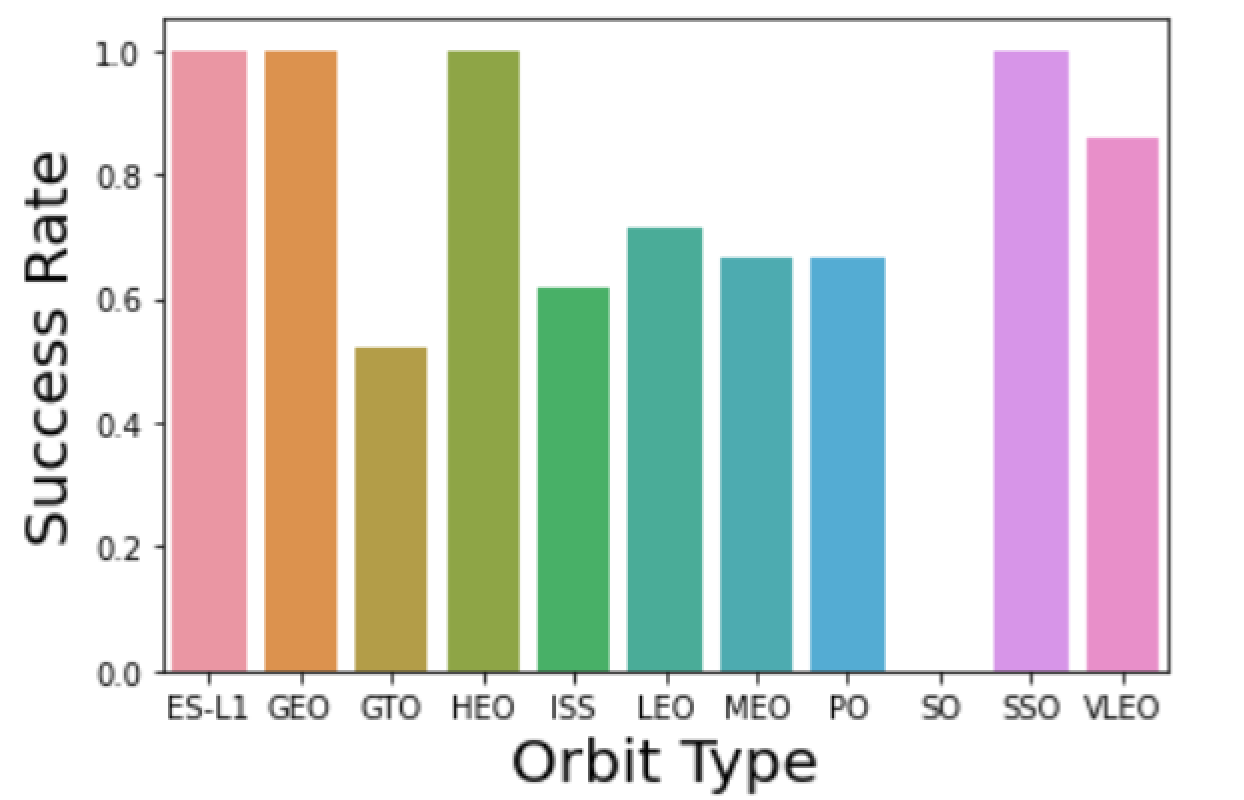
Payload vs. Launch Site

- Scatter plot displays the different payloads (kg) of the missions at the different launch sites. Showing the successful missions as class 1 and failure as class 0.
- Above a Payload of around 10000 kg, the chances of a successful mission are small.



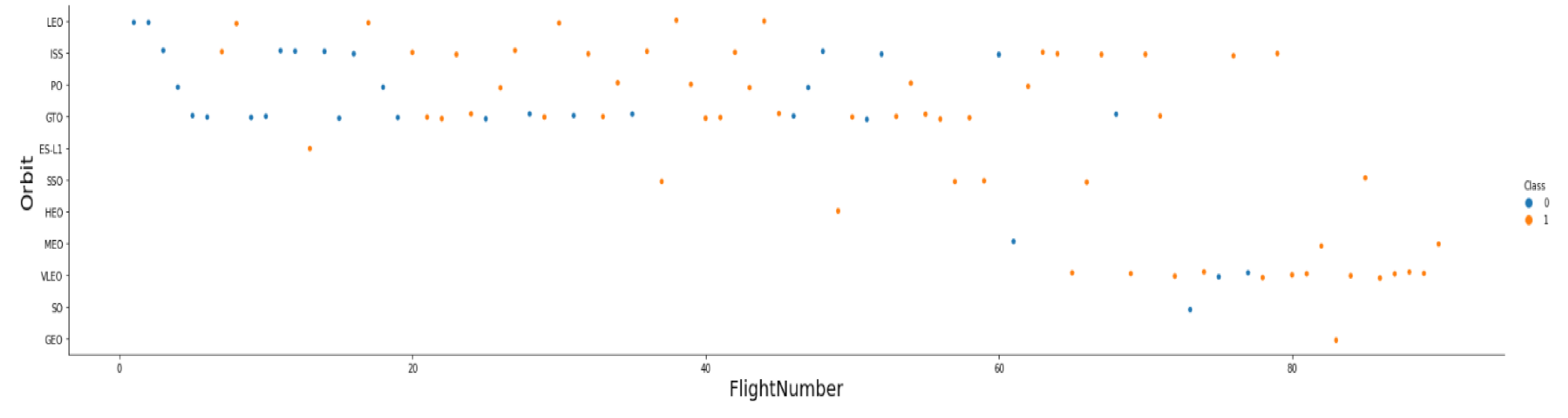
Success Rate vs. Orbit Type

- Barchart displays the Success Rate of the different Orbit types. With a success rate of 1.0 meaning that the orbit had a 100% success rate.
- Orbits ES-L1, GEO, HEO, SSO, VLEO all have an success rate of 80% or over.



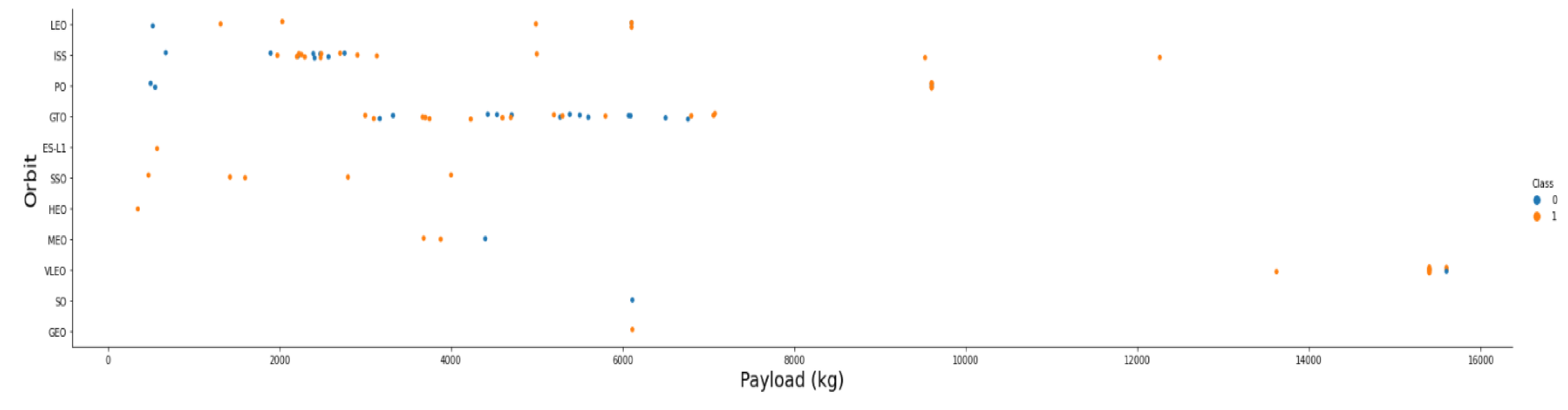
Flight Number vs. Orbit Type

- The scatter plot displays the Orbit type with the flight number, displaying whether the mission was a success (1) or a failure (0).



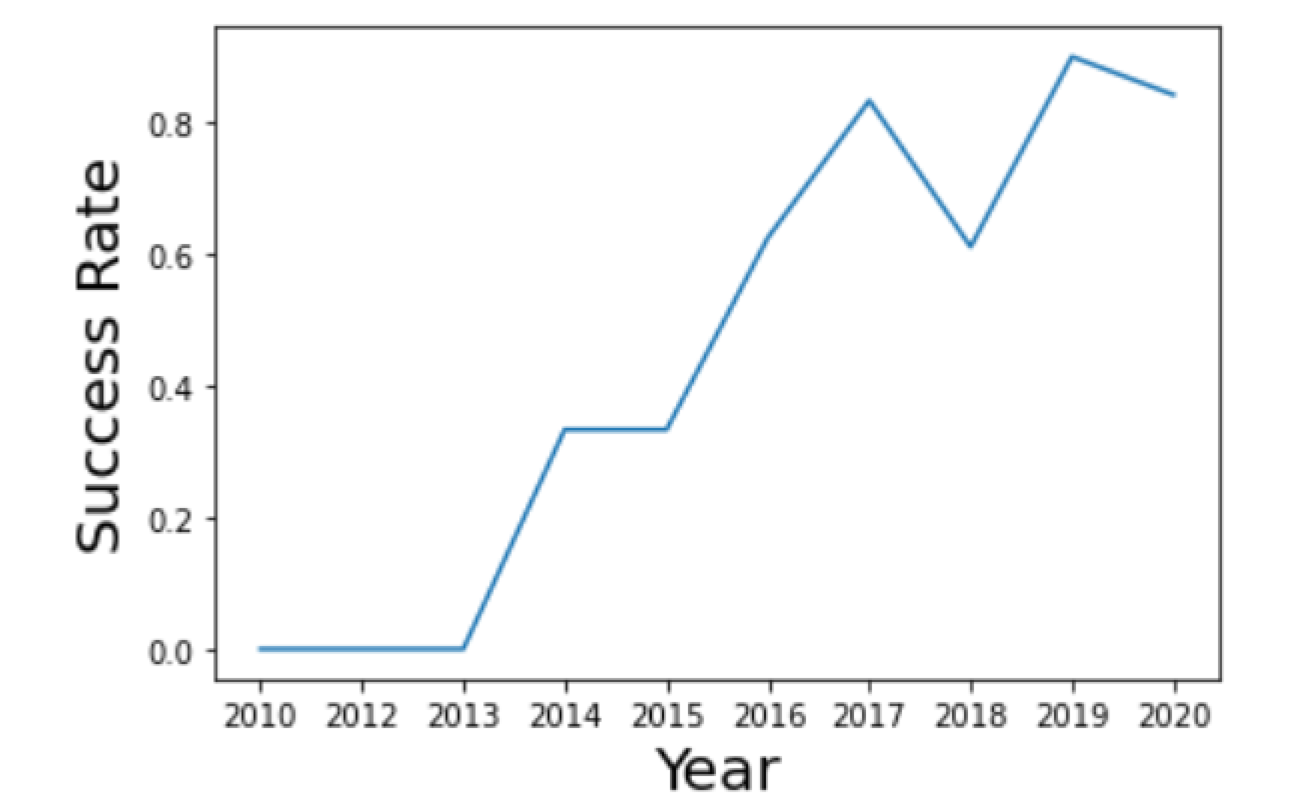
Payload vs. Orbit Type

- The scatter plot displays the Orbit type with the Payload (kg). An successful mission is displayed with class 1. Whereas a mission with failure is shown with class 0.



Launch Success Yearly Trend

- The Line Chart displays the success rate with each year since 2010 to 2020.
- Positive uptrend since the year of 2013.
- Slight dip from the major uptrend in 2017, with the success rate recovering back in 2019.



All Launch Site Names

- SQL query allows for the names of the launch sites to be retrieved.

In [10]:

```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

Out[10]:

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

Launch Site Names Begin with 'CCA'

- All Launch Sites with the characters containing “CCA” can be found. Displaying the data held within the SQL databases for the different missions.

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

Out[11]:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- SQL query allows for the summation of the Payload Mass (kg) carried by the boosters launched by NASA (CRS). The value obtained is 45596 (kg)

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

In [12]:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

Out[12]:

1
45596

Average Payload Mass by F9 v1.1

- SQL query allowed for the average Payload Mass (kg) to be found carried by the booster versions “F9 v1.1”. The average Mass is 2928 kg.

Task 4

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

In [13]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

Out[13]:

1
2928

First Successful Ground Landing Date

- SQL Query finds when the first date when there was a successful landing in the ground pad was achieved. The date found was the 22/12/2015.

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [14]:
```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL query finds the boosters which have a successful drone ship mission deployment with a Payload Mass greater than 4000kg, but less than 6000 kg (4000 kg < x <6000 kg)

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTC
```

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

Total Number of Successful and Failure Mission Outcomes

- The total amount of mission outcomes, whether successful or failure, can be found using the SQL query. The total amount is equal to 101.

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

1
101

Boosters Carried Maximum Payload

- SQL query is used to find all the booster versions where the Payload Mass is the highest amount. All the Booster versions are type “F9 B5.”

In [17]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Out[17]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Complex SQL query allows for the failed landing outcomes in the drone shop. The corresponding booster versions, launch sites for 2015 can be found as shown by the output below.

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Complex SQL query allows for the count of the landing outcomes between 2010 – 06 – 04 and 2017 – 03 -20 to be ranked as shown below.

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SPACEXTBL \
    WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND LANDING__OUTCOME LIKE '%Success%' \
    GROUP BY "DATE" \
    ORDER BY COUNT(LANDING__OUTCOME) DESC
```

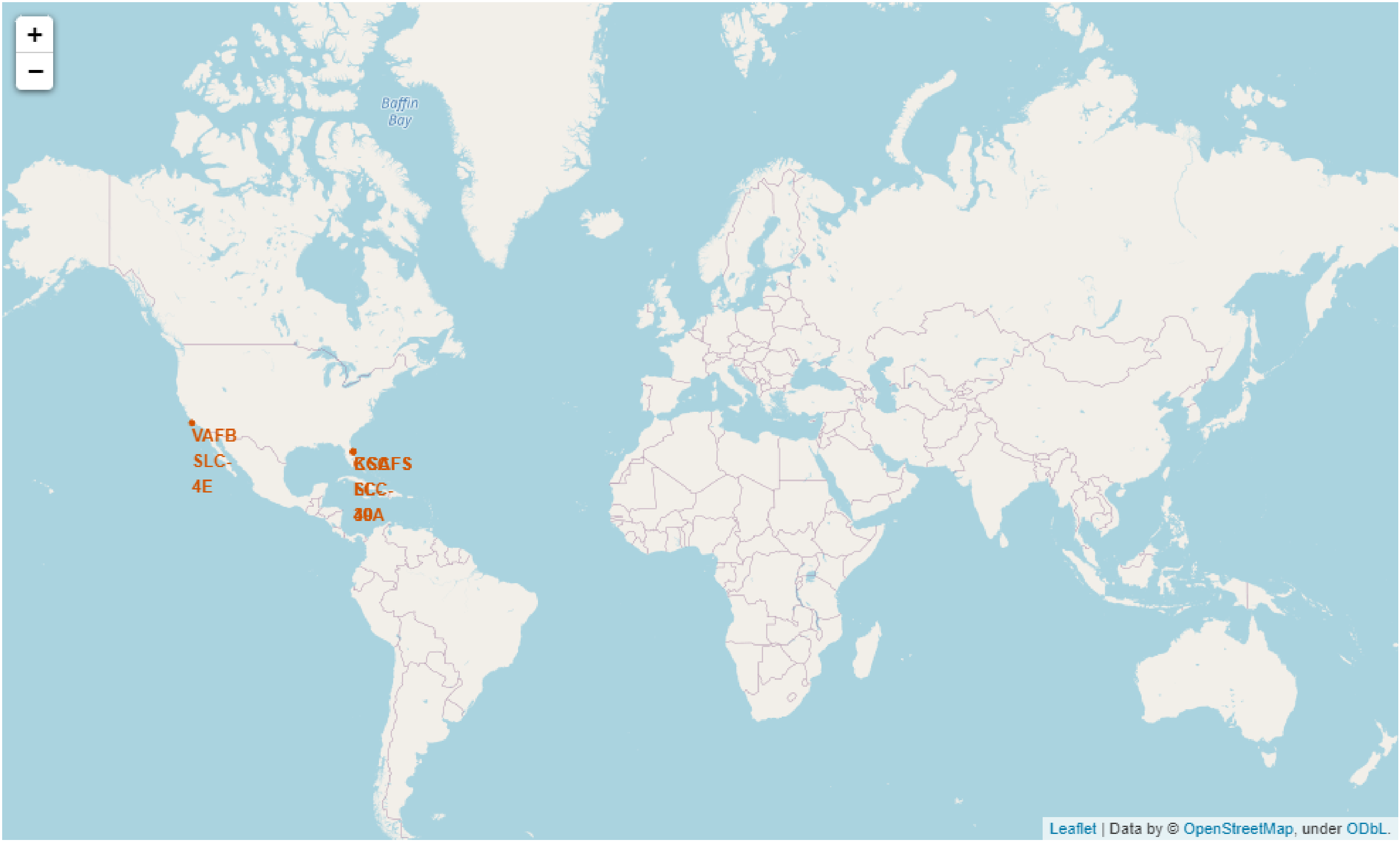
DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Section 3

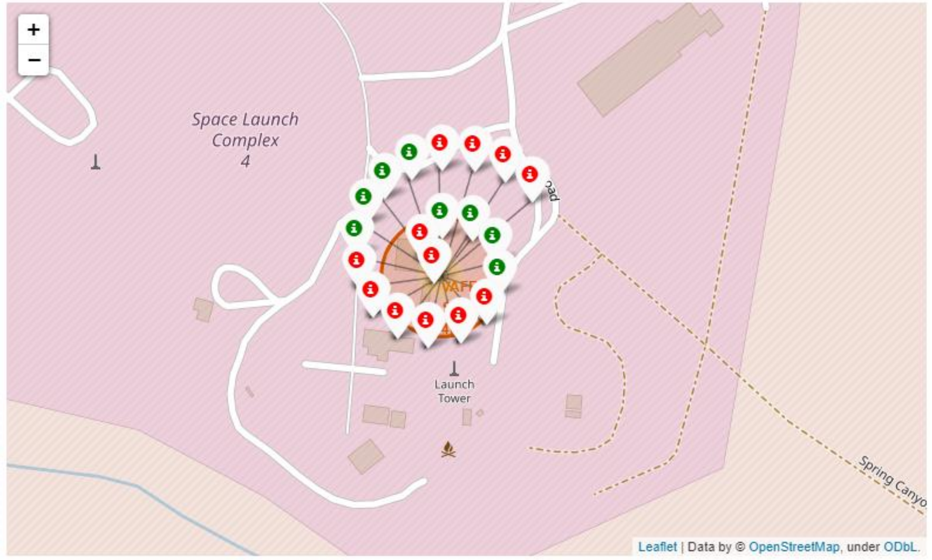
Launch Sites Proximities Analysis

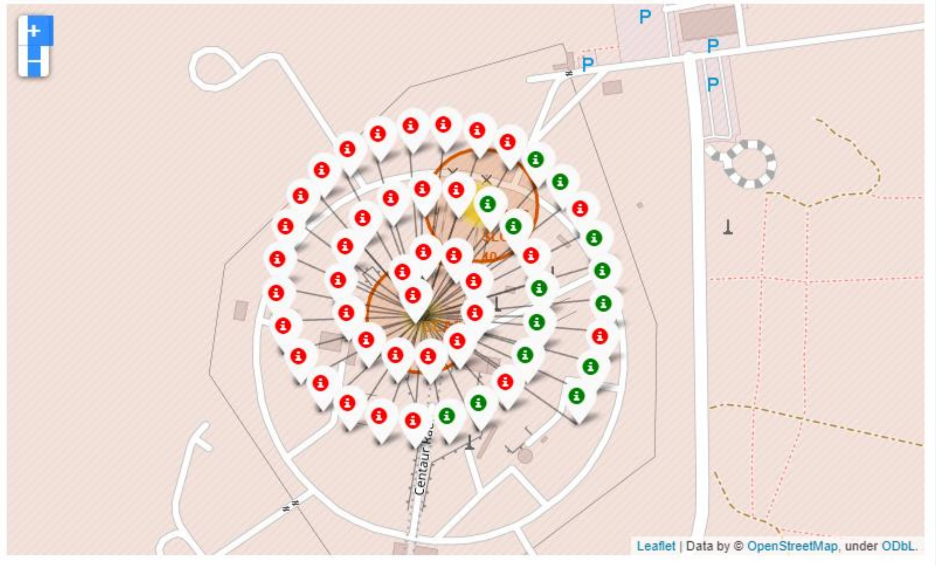
Global Map displaying Launch Sites

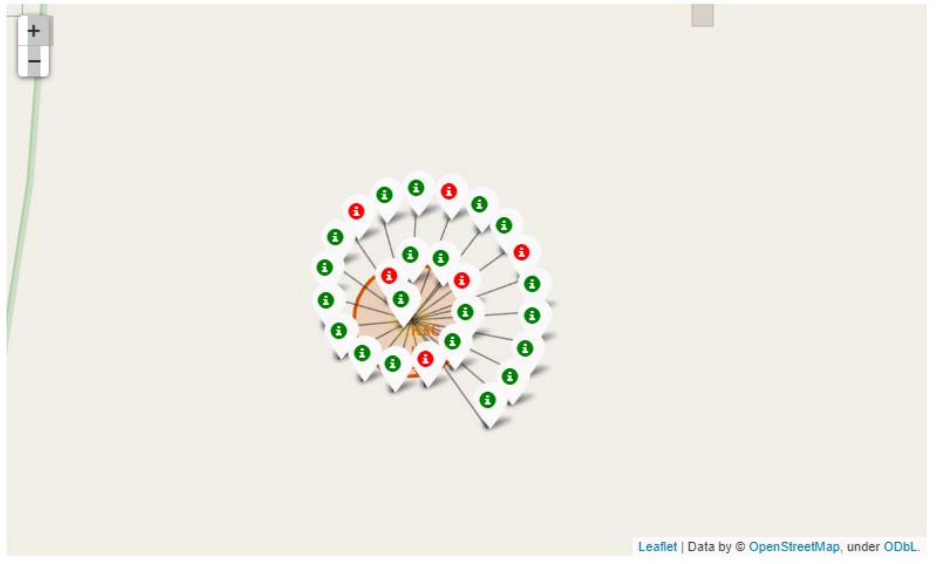
- Folium Map displays the Launch Sites on a global Scale. The maps shows that the launch sites are just located in the United States of America.



<Folium Map Screenshot 2>

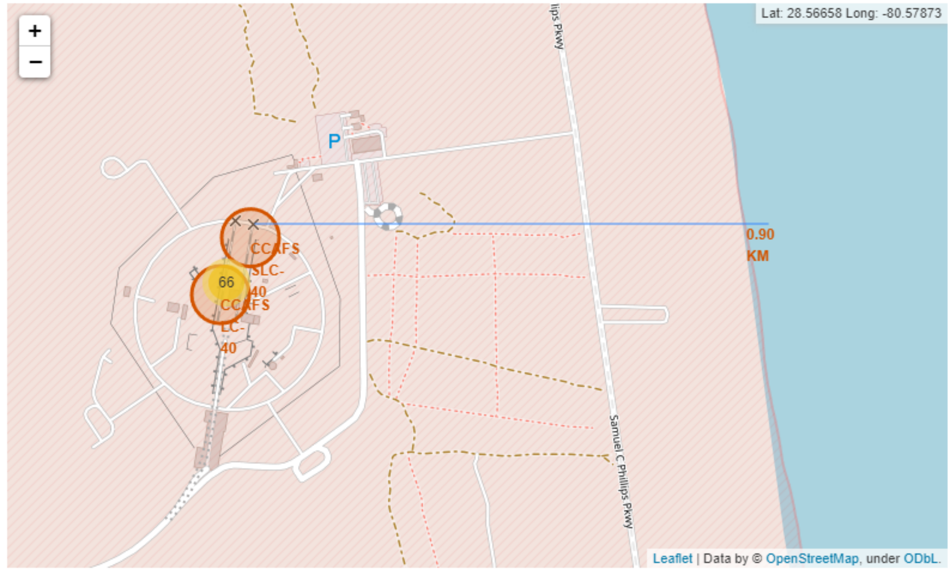


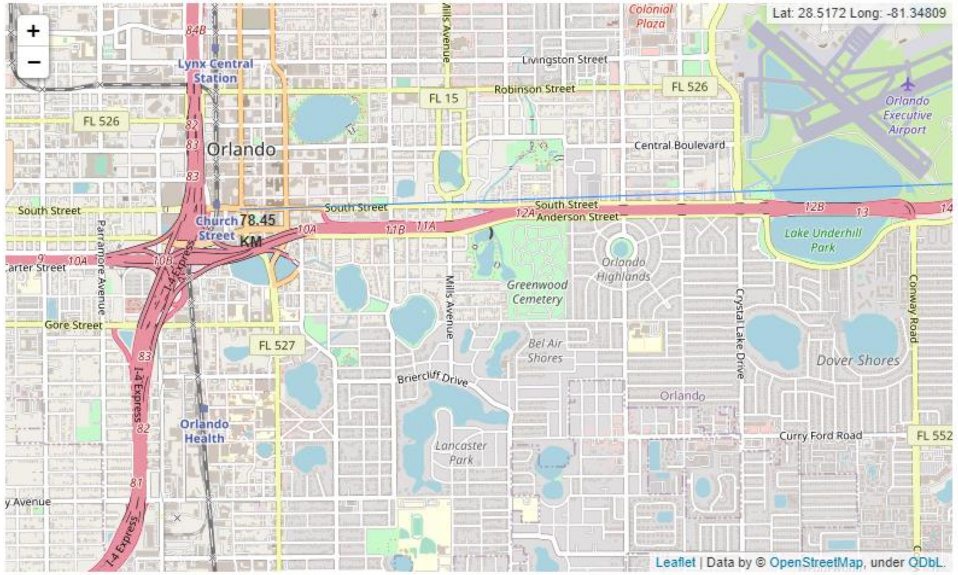


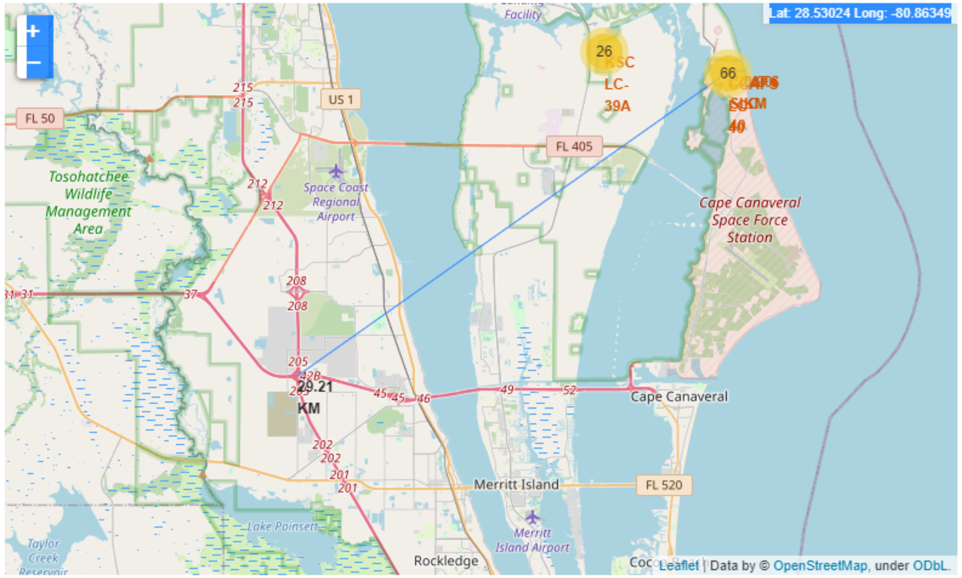


- Three images display launch sites. The top right image displays two Launch sites located in Florida. The markers display whether if the launches were either Successful (green) or failures (red).

Distances from Florida Launch site to Proximities







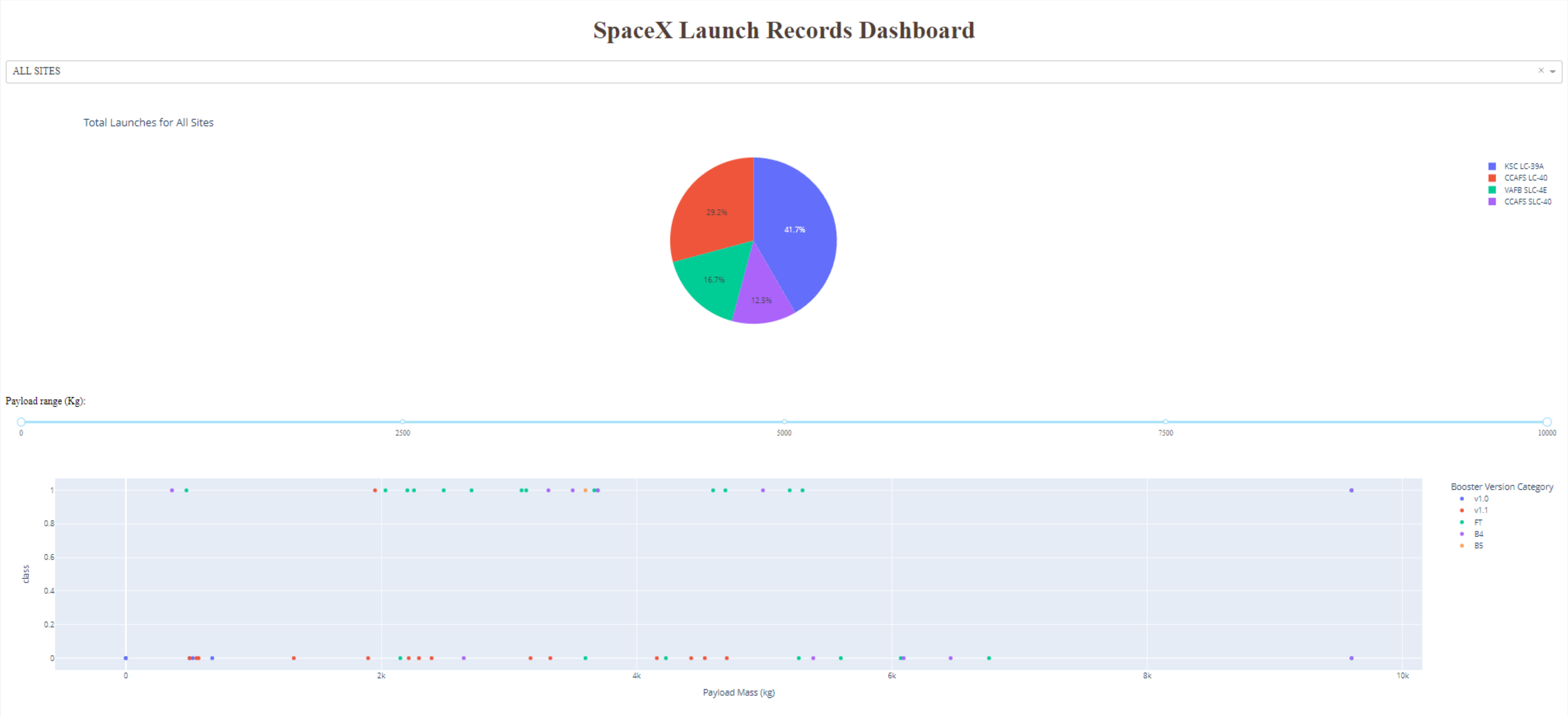
- Top left Image shows the distance to the coastline which is 0.90 km
- Top Right Image shows the distance to the City, which is 78.45 km
- Top Left Images shows the distance to the Main Highway which is 28.21 km

Section 4

Build a Dashboard with Plotly Dash

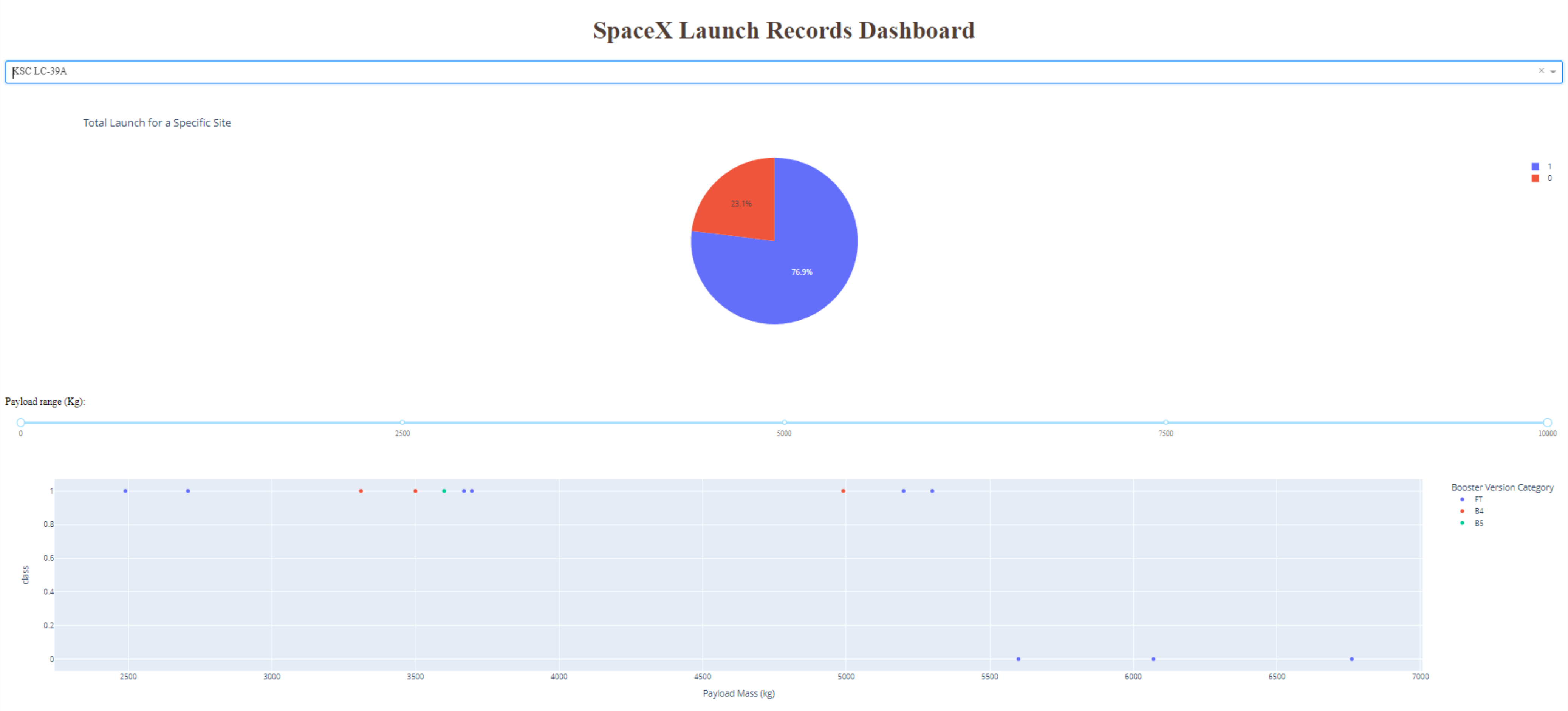
Launch Success for all Sites

- Plotly Screenshot displays total launches for the four different launch sites. The ratios are displayed in a pie chart format as shown below.



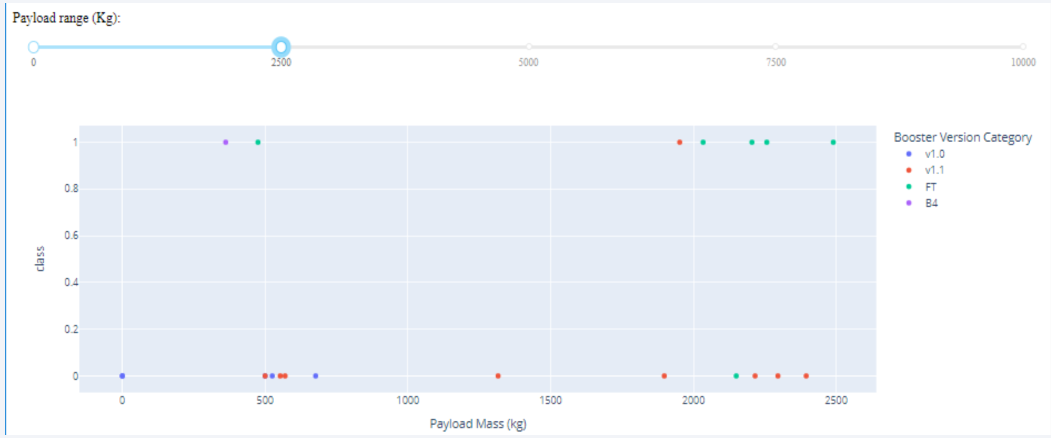
Launch Sites with the most Successful Launches

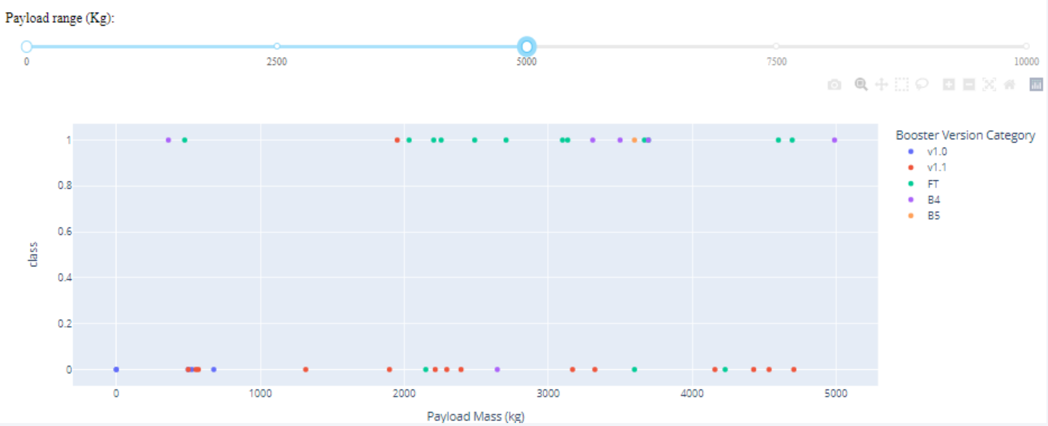
- The Launch sites with the highest success rate is the KSC LC-39A. The Pie Chart shows that the success rate is 76.9%.

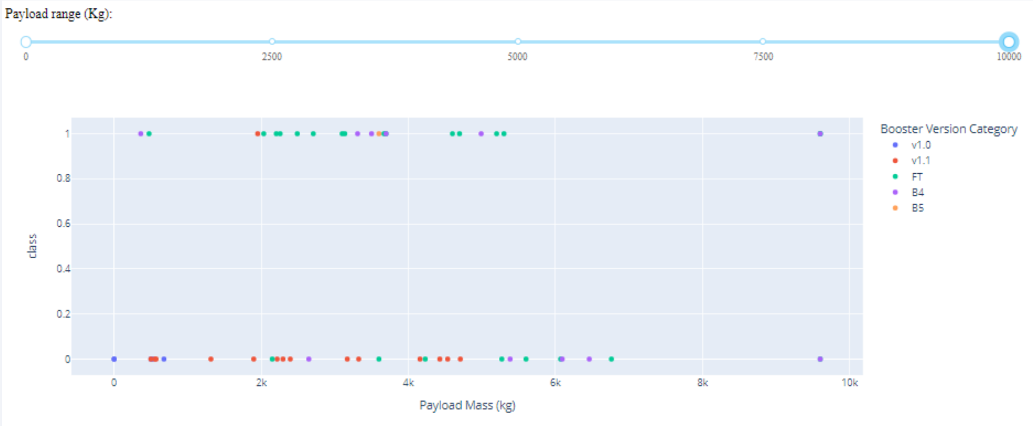


Different Payload Mass (kg) for All Sites

- Four images display the Booster Versions at different payloads. The most failures occur with the payload's mass at the lowest. For the other payload masses there seems to be a similar amount of success rates.







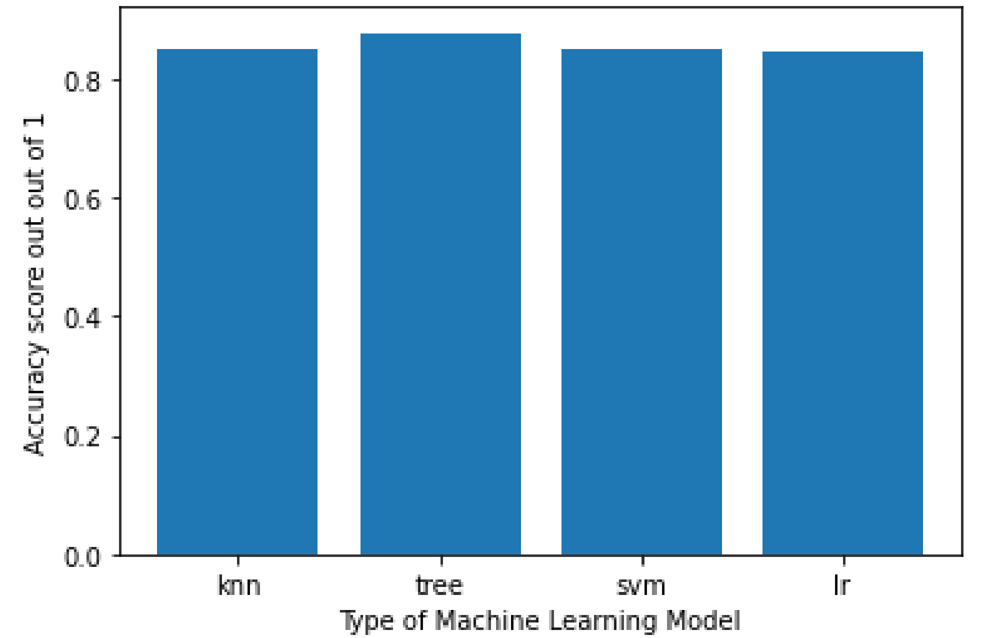
Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
Methods = {'knn':knn_cv.best_score_,
           'tree':tree_cv.best_score_,
           'svm':svm_cv.best_score_,
           'lr':logreg_cv.best_score_}
Best_method = max(Methods, key=Methods.get)
print(Best_method)
```

```
tree
```

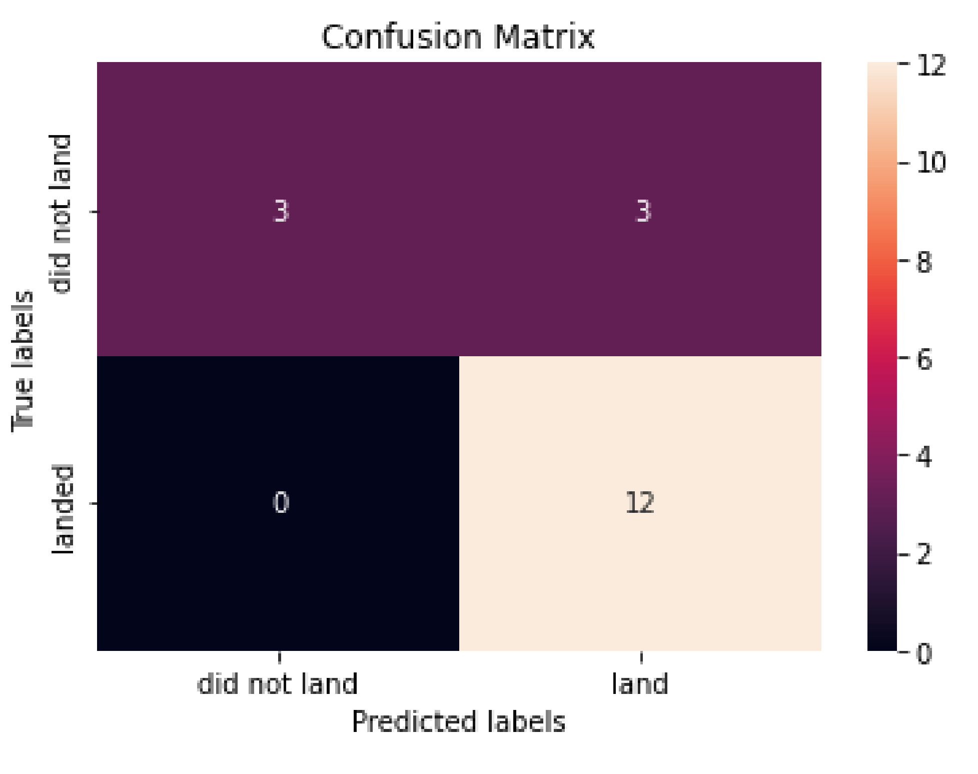


- The bar show shows the best Machine Learning model by the accuracy. The model found to have the highest accuracy is the Tree model with an accuracy of 87.6%.
- The lowest accuracy is the model based on Logistic Regression with an accuracy of 84.6%

```
KNN: 84.8%
Tree: 87.7%
SVM: 84.8%
LR: 84.6%
```

Confusion Matrix

- The confusion Matrix for the Decision Tree Model shows whether there have been any false positives. This can be shown in the top right corner to be 3. There 0 false negatives as shown by the bottom left corner.



Conclusions

- The launch rate success rate has been steadily increasing in a positive trend since 2013.
- The less Payload Mass, the more likely failure to occur as found in the Plotly dash interactive results.
- The orbit types with the most successful launches was the ES-L1, GEO, HEO, SSO, VLEO, which all has a success rate higher than 80%
- The KSC-LC-39A Booster had the most successful launches.
- The Decision Tree Classifier had the most accurate Machine Learning model.

Appendix

- No Appendices are required.

Thank you!