IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ahmed Meleha
01.08.2024

Outline

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

Executive Summary

- Project-Aim:
  - Predict whether SpaceX's Falcon 9 first stage rockets will land successfully using several machine learning classification algorithms.
- Methodologies:
  - Data collection and wrangling
  - Exploratory data analysis (EDA)
  - Interactive data visualization and dashboarding
  - Machine learning prediction
- Results:
  - The SVM, KNN and Logistic regression models are the best in terms of predicting the launch outcomes of Falcon 9 first stage rockets.

Introduction

- Project background and context
  - SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

- Research Question at hand:
  - Is it possible, given various Falcon 9 rocket specifications (e.g. payload mass, orbit type, booster version, launch site etc.) to predict the success of first stage landing attempts?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping of additional Falcon 9 historical launch records from Wikipedia
- Perform data wrangling
  - Filtering data for Falcon 9 observations, replacing Payload-Mass null values with mean values, One-Hot-Encoding the landing outcomes for Machine Learning (ML) 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
  - Evaluating and testing Linear Regression (LR), KNN, SVM, DT models to determine the best classifier

Data Collection

SpaceX REST API

- The SpaceX REST API (Endpoints: api.spacexdata.com/v4/) is called to retrieve data for Falcon launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - SpaceX REST API returns data as JSON and is then turned into a Pandas df.
  - The created df is filtered for Falcon 9 lauches.
  - Missing PayloadMass values (null values) are replaced by the mean PayloadMass of all launches.

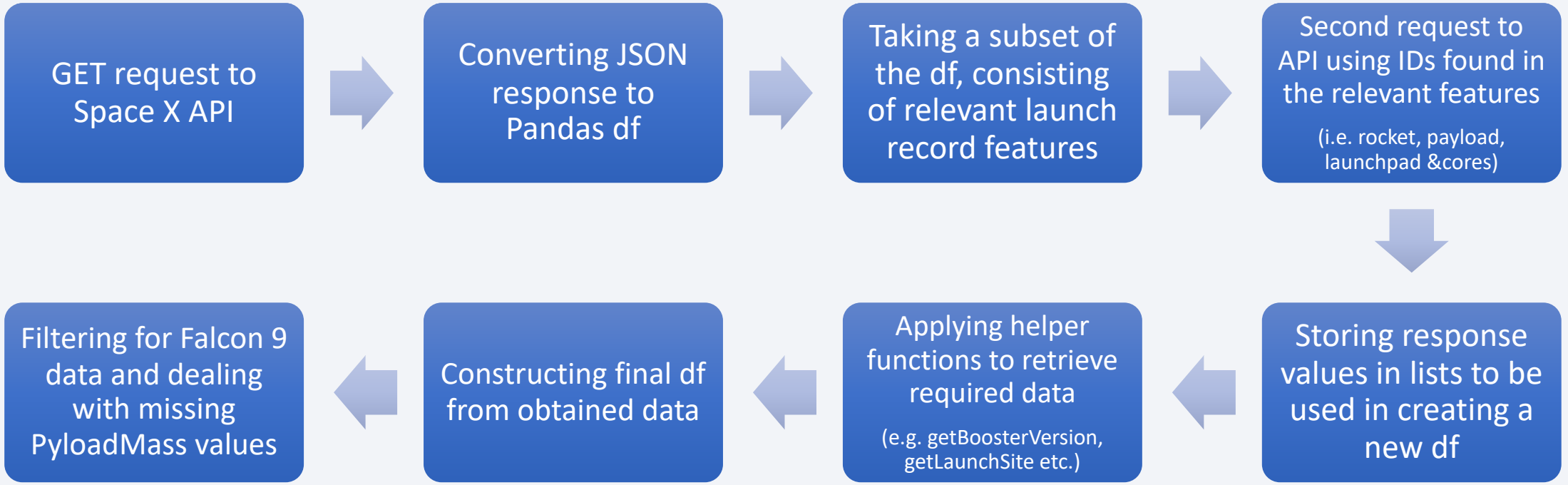
Web-Scraping

- Additional Falcon 9 Launch records are web scraped from the dedicated Wiki page using BeautifulSoup.
  - The HTML Tables found on the Falcon 9 Wiki page are parsed and stored as a Pandas df.

Export as CSV

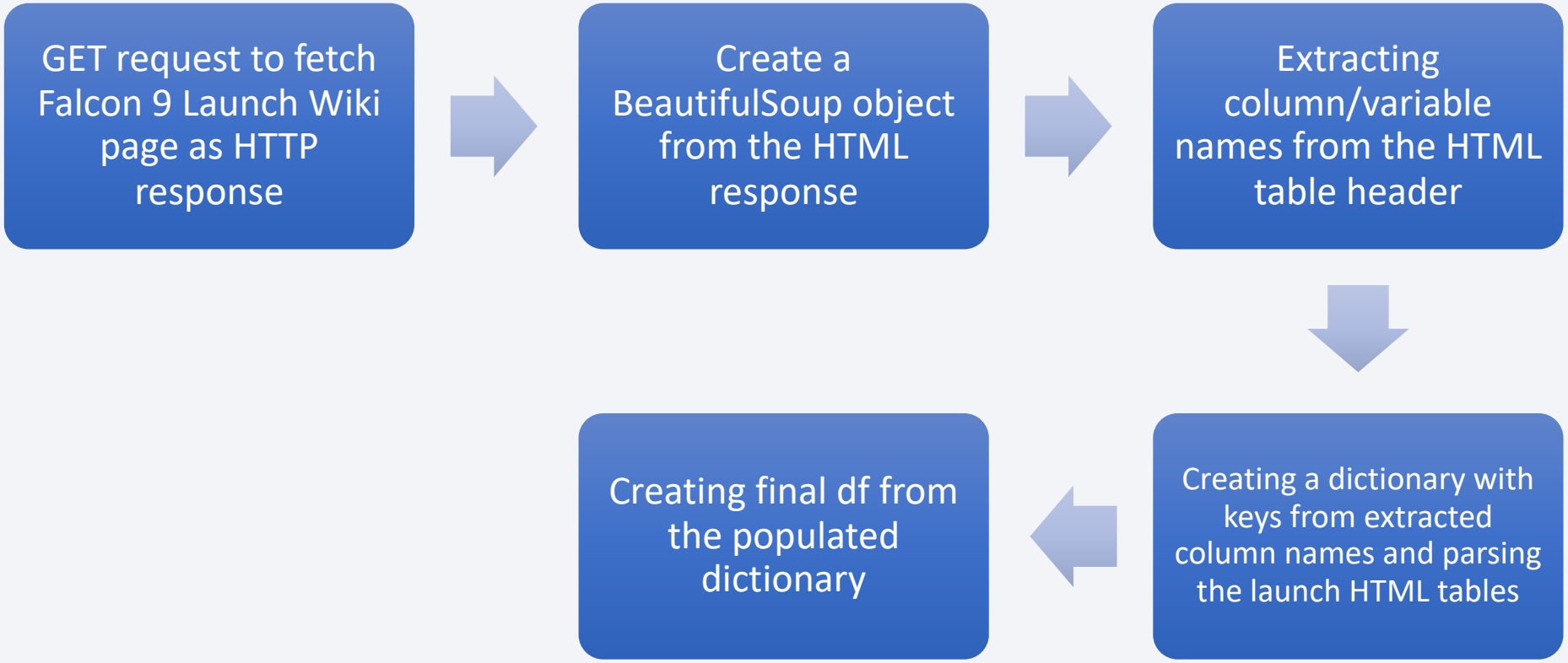
- The resulting datasets / dataframes from the SpaceX REST API and Web-Scraping proccess are exported as CSV files for further data wrangling and exploration.

Data Collection – SpaceX API



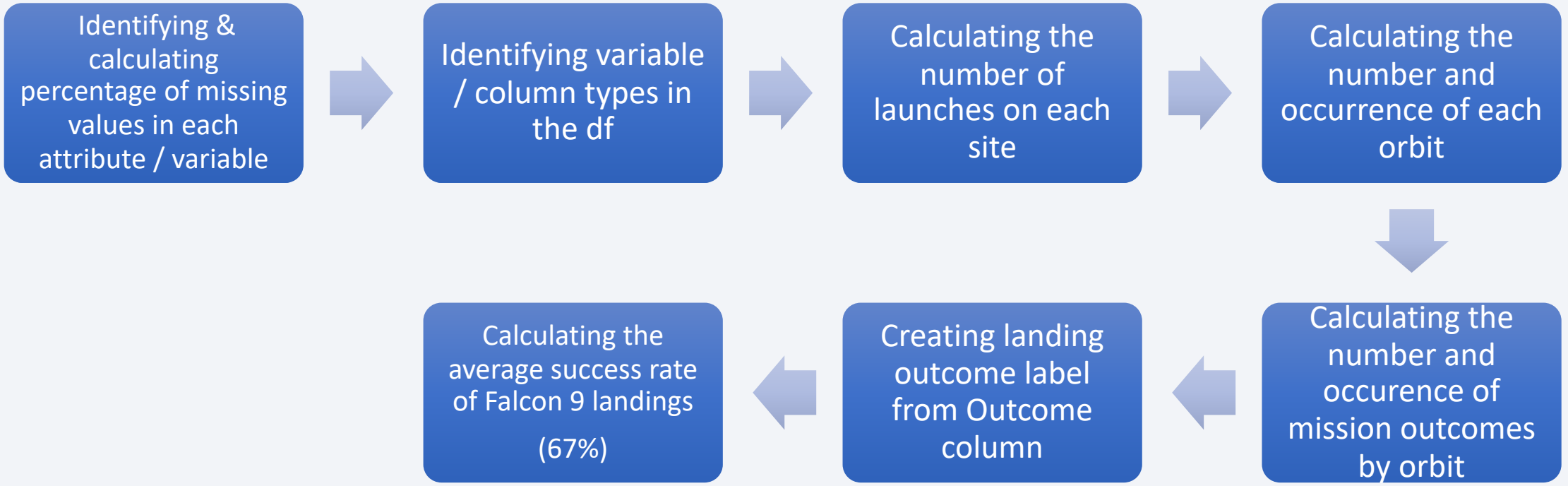
https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping



https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling (Preliminary EDA)



https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To perform EDA with Data visualization mainly the Matplotlib & Seaborn libraries were utilized (in addition to Pandas & Numpy).
- Specifically:
  - Seaborn's Catplot was used to visualize the relationship between flight number and payload mas.
  - Seaborn's Catplot was used to visualize the relationship between flight number and launch site.
  - Seaborn's Catplot was used to visualize the relational sip between payload mass and launch site.
  - Seaborn's Bar-chart was used to visualize the success rate of the different orbits.
  - Seaborn's Catplot was used to visualize the relationship between flight-numbers / flight-frequency and orbit type.
  - Seaborn's Catplot was used to visualize the relationship between payload mass and orbit type.
  - A simple Line-chart was used to visualize the average yearly success rate trend for Falcon 9 first stage rockets.

https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries performed include:

- Displaying the name of the unique launch sites in the space mission.
- Displaying five record where launch sites begin with the string “CCA”.
- Displaying the total payload mass carried by boosters 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 successful landings on drone ships and have a 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 of failure landing outcomes on drone ships, including boosters version and launch site for the months in the year 2015.
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

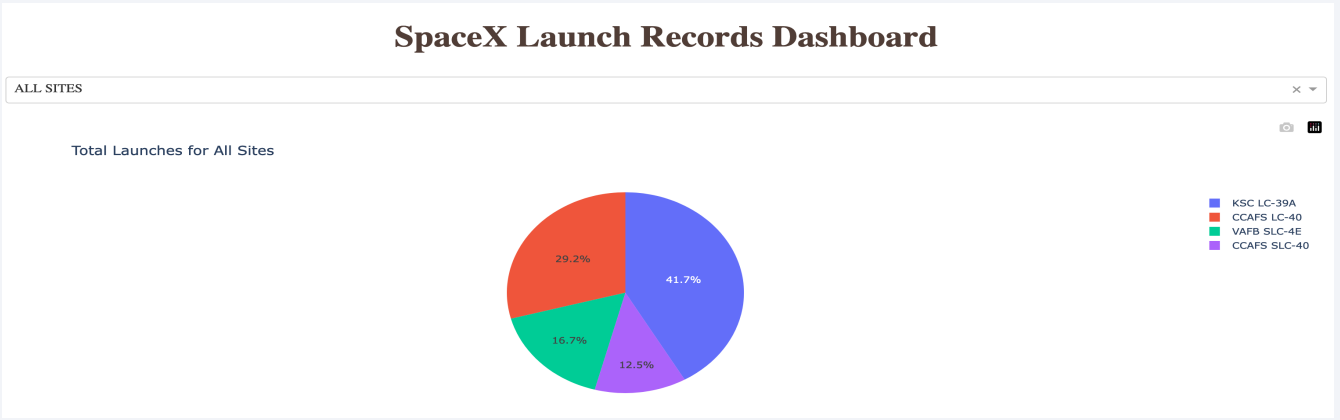
Build an Interactive Map with Folium

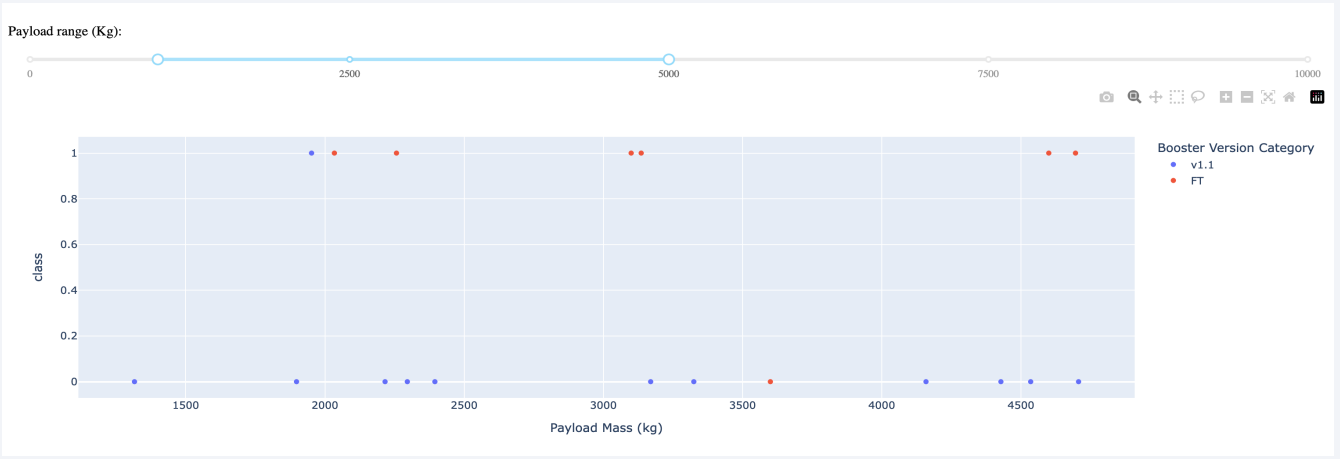
- The Folium package was utilized to perform an additional interactive visual analysis, aiming at identifying possible correlations between the launch sites and their success rate.
- Accordingly, the visual analysis included:
  - Marking all launch sites on the map to identify any similarities, like for example that the launch sites are in proximity to the Equator line and in very close proximity to the coast.
  - Marking the successful/failed launches for each launch site on the map to identify the most successful launch sites, for example launch site KSC LC-39A.
  - Calculating the distances between a launch site to its proximities, for example to calculate the proximity to the nearest railway, highway, coastline, etc.

https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A pie chart with a drop down menu (including all launch sites) was added to the dashboard, to highlight the site with the most successful launches / success rate.
- As scatter plot with a range slider for the payload mass was added, to determine which payload mass range has the lowest / highest success rate.
- The scatter plot also helps in determining which Falcon 9 Booster Version has the highest launch success rate.

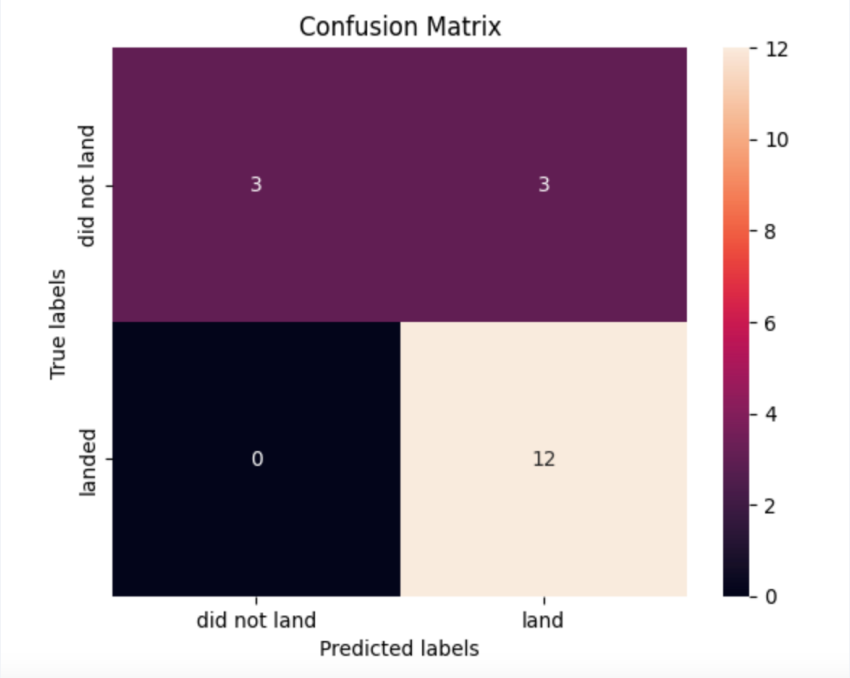




https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The “Class” data (representing the launch outcomes) were assigned to the dependent variable Y
- The independent variables were standardized and assigned to variable X
- The X and Y data were then split into a training and testing test, whilst setting the test size parameter to 0.2 and random state to 2.
- Grid Search is then performed for all ML models (Logistic Regression, Support Vector Machine (SVM) , Decision Tree and K-Nearest Neighbors (KNN)) to identify the best tuning parameters.
- The test accuracy is then calculated to identify the most accurate classification model. Confusion Matrix is also produced to highlight each model’s accuracy.



https://github.com/ameleha/IBM_ds_capstone_submission/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Exploratory Data Analysis (EDA) Results

- The EDA shows that as the flight number increases, the first stage is more likely to land successfully.
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- Launch sites: KSC LC-39A and VAFB SLC 4E have the highest first stage landing success rate with 77%
- Launches aimed towards ES-L1, GEO , HEO and SSO orbits have the highest landing success rate among other orbits.
- The more Space X continues to launch rockets over the years the more successful landings they achieve. It seems therefore that each launch adds valuable experience in regards to optimizing the launches and respective landing procedures.

Interactive Analytics Results

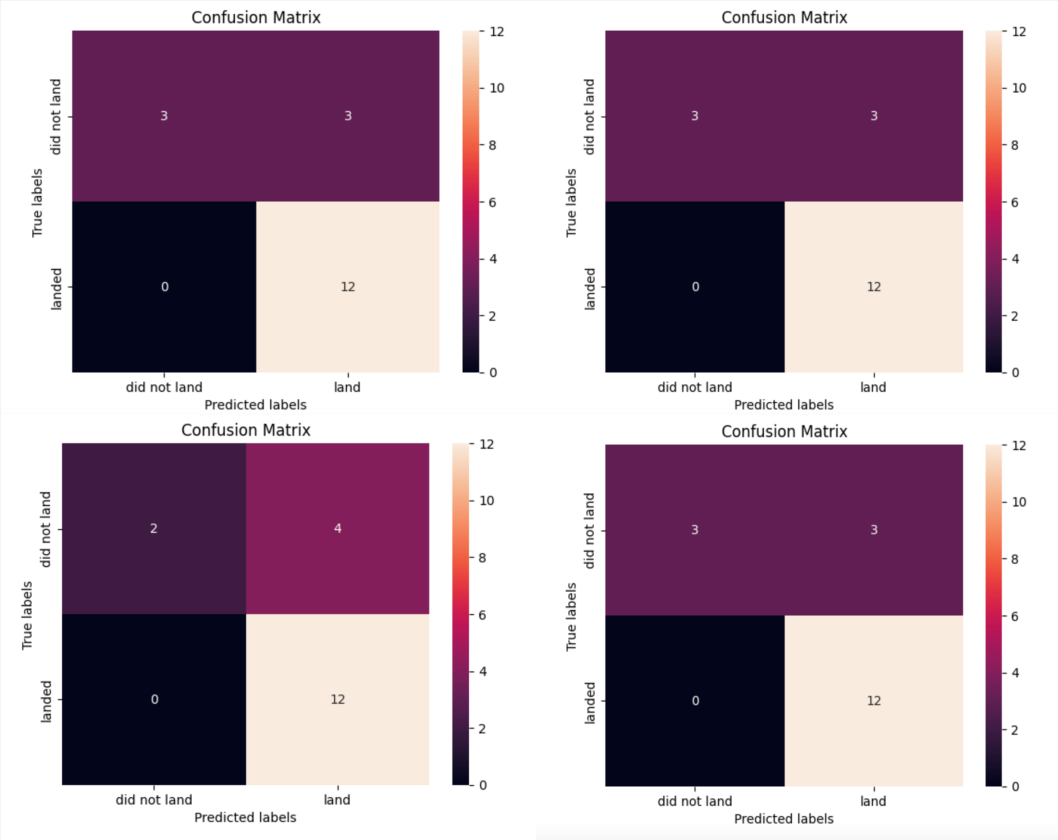
- Interactive analytics reveal the launch sites are in proximity to the Equator line and in very close proximity to the coast.

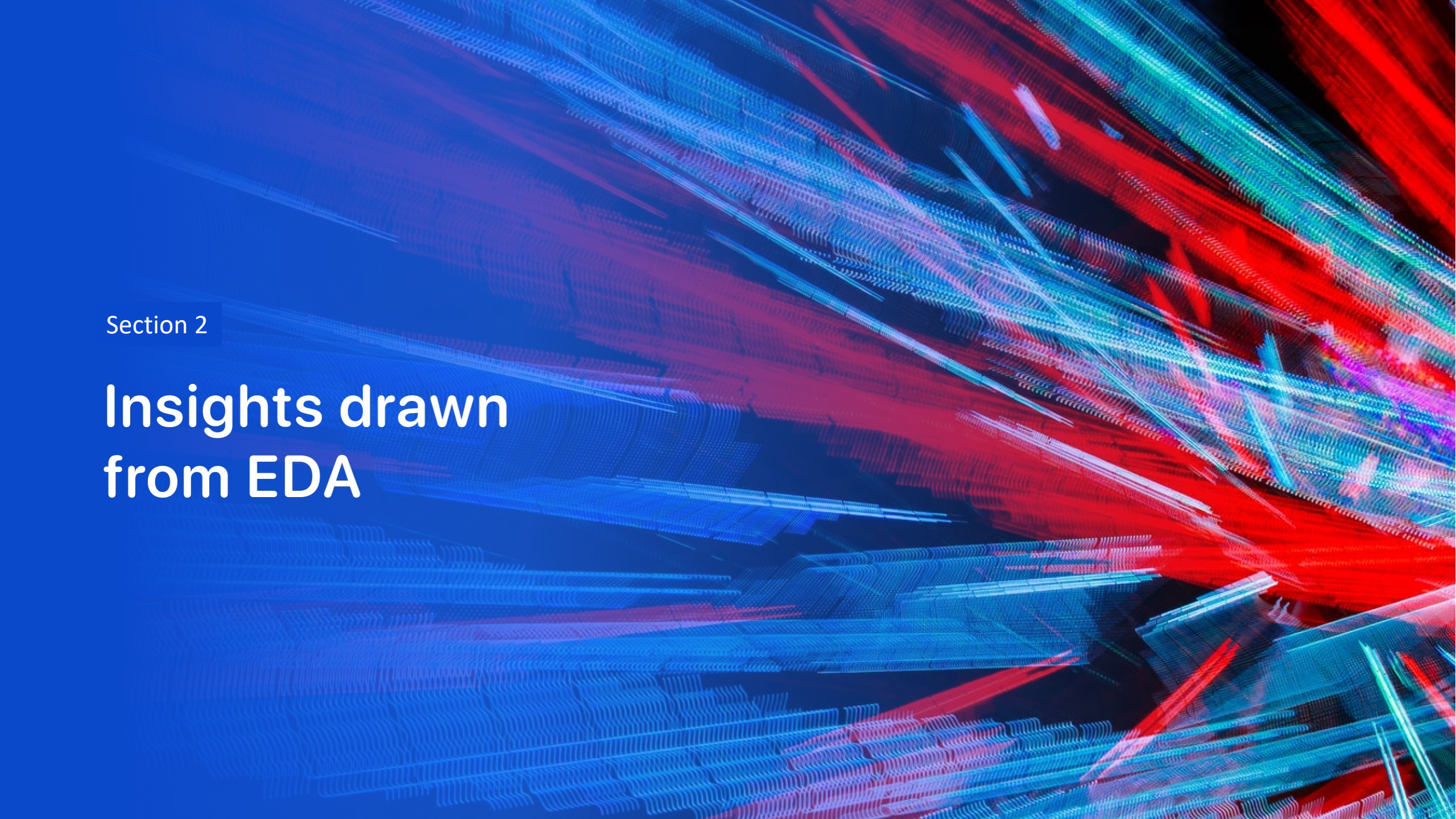
- According to the count of successful /failed launches (green/red tags) for each launch site on the map, launch site KSC LC-39A is the most successful.

Predictive Analysis Results

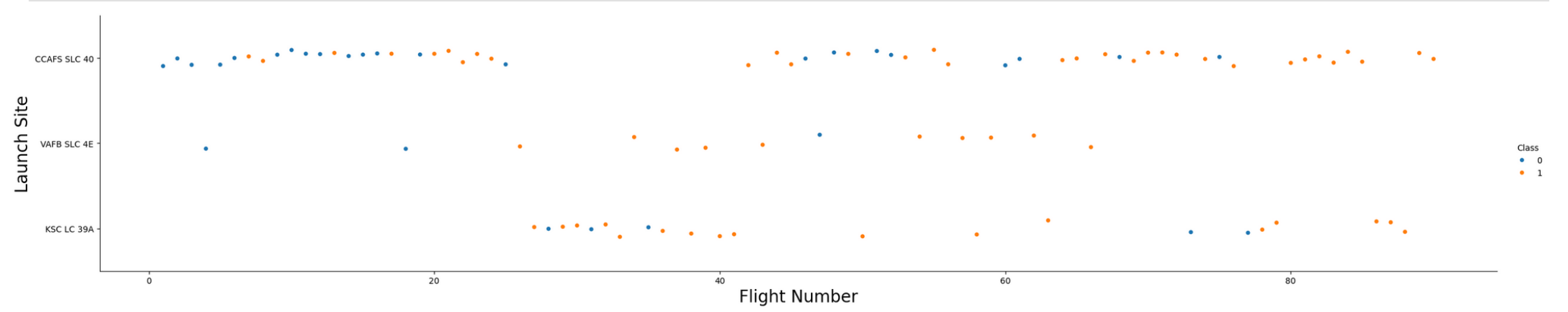
The SVM, KNN and Logistic regression models are the best in term of prediction
accuracy of the dataset with an accuracy of 83.3%.

Whereas the Decision Tree model only scored an accuracy of 77.8%




Section 2
Insights drawn from EDA

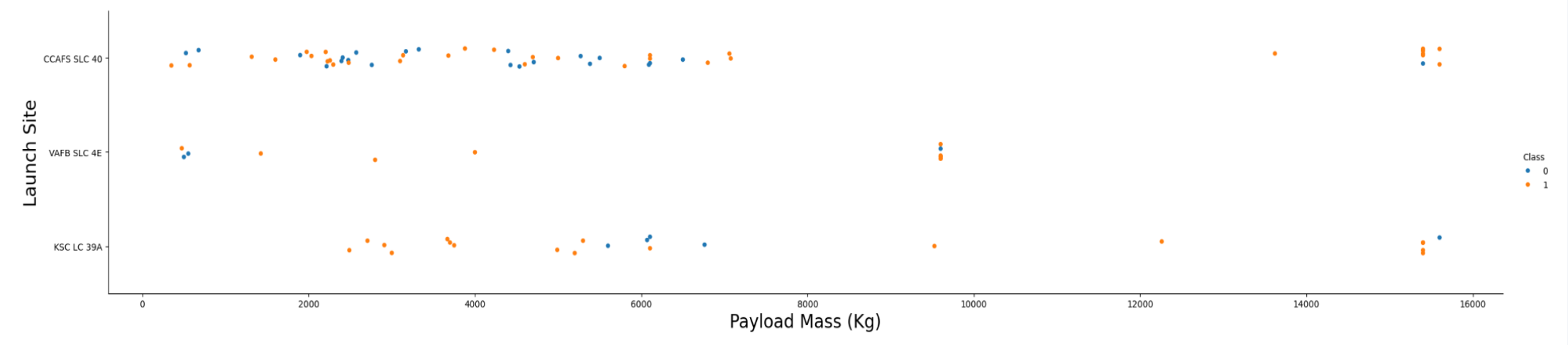
Flight Number vs. Launch Site



- The above scatter plot shows the relationship between the Flight Number and Launch Outcome (i.e. success / failure) with respect to each Launch Site.

- The plot reveals a possible correlation between flight frequency and landing success. Meaning that Falcon 9 first stage rockets seem to land more successfully, the more launch attempts are being made (a good example for that is launch site: VAFB SLC 4E).

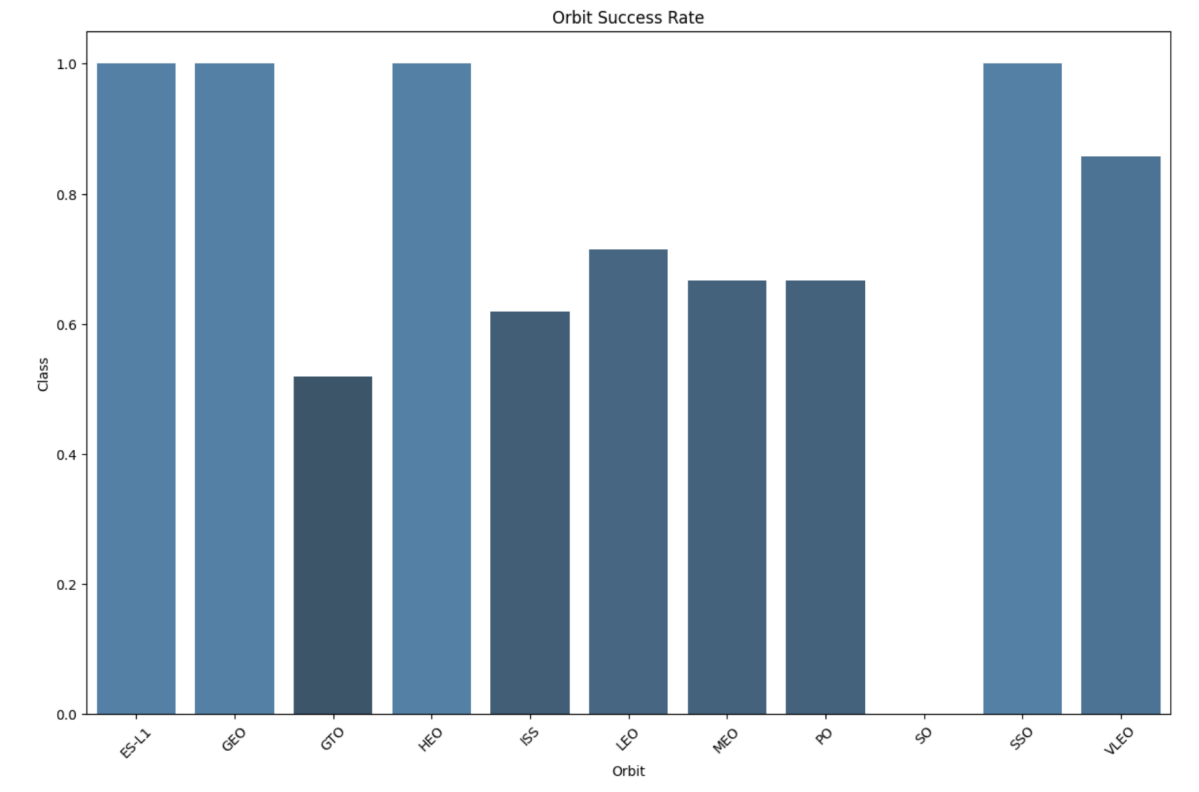
Payload vs. Launch Site



- The above scatter plot shows the relationship between the Payload Mass and Launch Outcome (i.e. success / failure) with respect to each Launch Site.

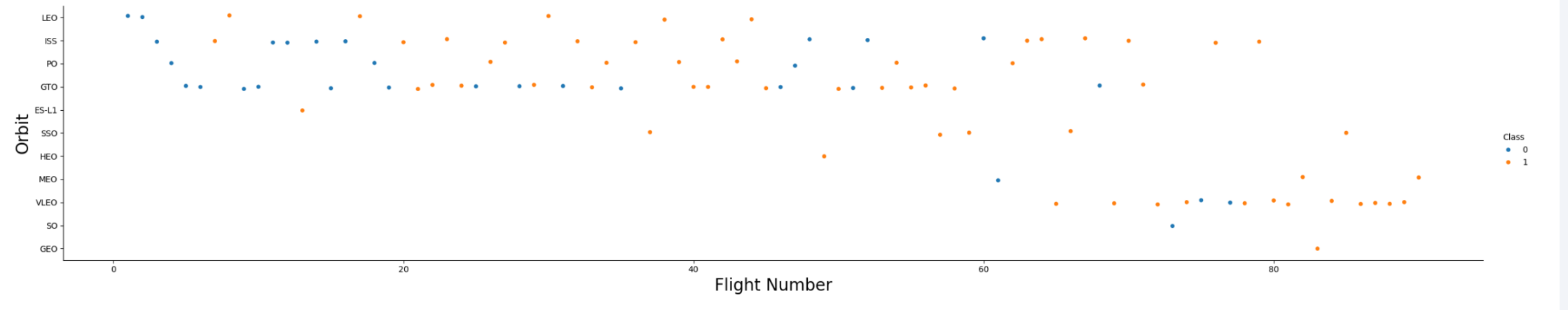
- The plot also reveals that generally most Falcon 9 launches, have a payload below 8'000 Kg and that the VAFB-SLC site launches no rockets with a payload mass greater than 10'000 Kg.

Success Rate vs. Orbit Type



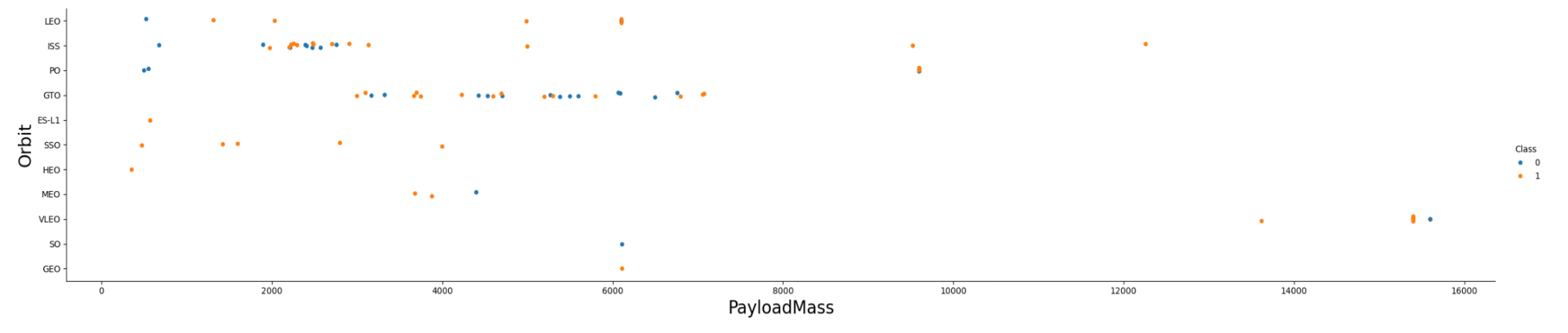
- The bar chart above shows the success rate of each orbit.
- The chart shows that launches aimed towards ES-L1, GEO , HEO and SSO orbits have the highest landing success rate among other orbits.

Flight Number vs. Orbit Type



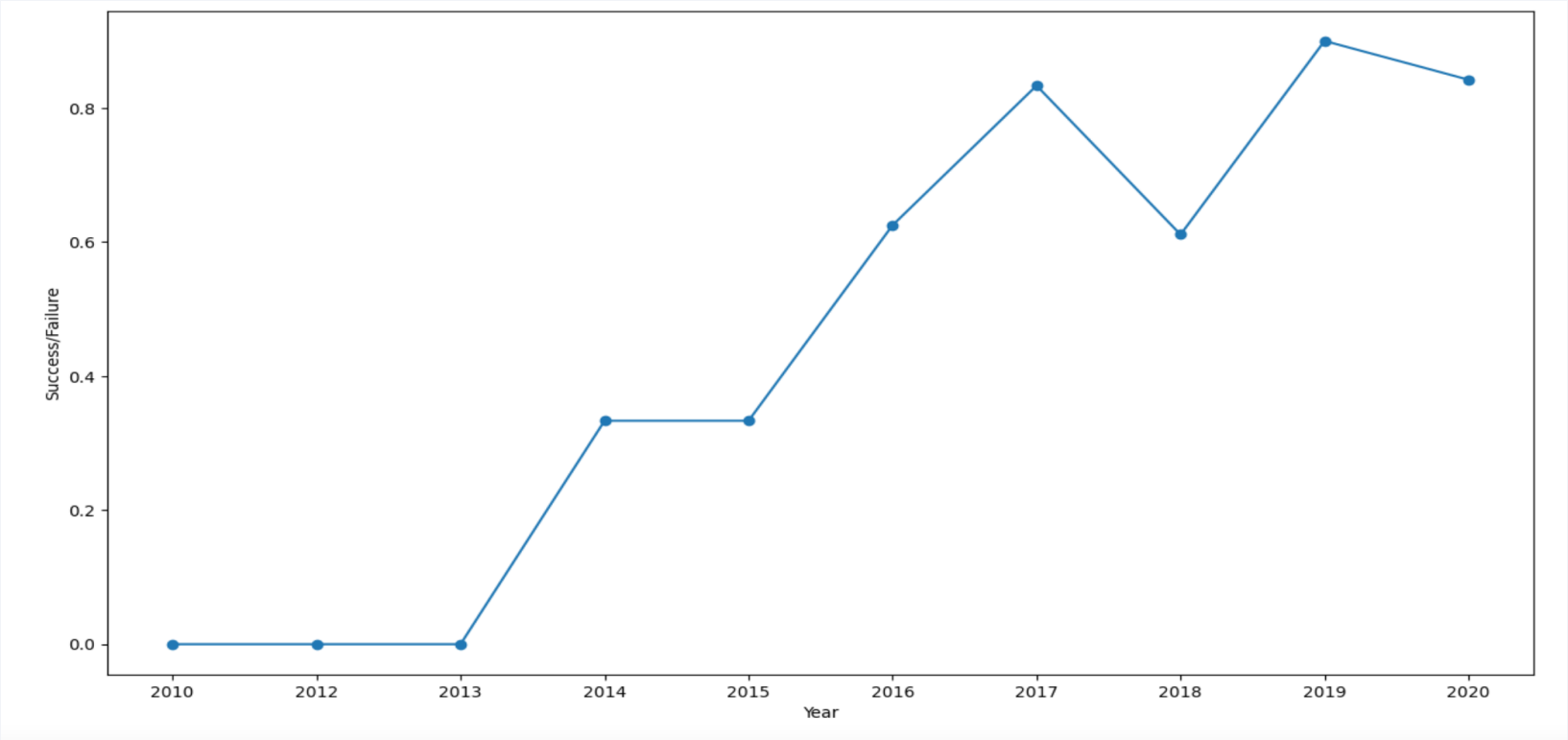
- The above scatter plot shows the relationship between the Orbit and the launch frequency in terms of the success / failure of landings.
- The LEO orbit success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in the GTO orbit.

Payload vs. Orbit Type



- The above scatter plot shows the relationship between the Orbit and the Payload Mass in terms of the success / failure of landings.
- The plot reveals a possible positive correlation between the Orbit and Payload Mass for heavy payloads launched in Polar, LEO and ISS orbits. stage rockets seem to land more successfully, the more launch attempts are being made
- For the GTO orbit we cannot distinguish any correlation as the success/failure rate of launches with a payload range between 3’000 and 6’500 KG seem to be similarly frequent.

Launch Success Yearly Trend



The more Space X continues to launch rockets over the years the more successful landings they achieve. It seems therefore that each launch adds valuable experience in regards to optimizing the launches and respective landing procedures.

All Launch Site Names

```
%%sql SELECT DISTINCT Launch_Site
    FROM SPACEXTABLE;
```

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

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

- The query retrieves the names of the unique launch sites of Falcon 9 first stage rockets.

Launch Site Names Begin with 'CCA'

```
%%sql SELECT *
    FROM SPACEXTABLE
    WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	La
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fa
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	Fa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

- The query retrieves 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

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

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

SUM(PAYLOAD_MASS__KG_)
45596

- The query calculates the total payload mass (of 45‘596 Kg) carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
    WHERE Booster_Version = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- The query calculates the average payload mass (of 2‘928.4 Kg) carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
%%sql SELECT MIN(Date)
    FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

- The query retrieves the date when the first succesful landing outcome on a ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT DISTINCT Booster_Version
    FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

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

- The query lists the names of boosters which have successfully landed on a drone ship and had a payload between 4’000 and 6’000 Kg.

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS  count
    FROM SPACEXTABLE
    GROUP BY 1
    ORDER BY 2 DESC;
```

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

Mission_Outcome	count
Success	98
Success (payload status unclear)	1
Success	1
Failure (in flight)	1

- The query lists the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%%sql SELECT DISTINCT Booster_Version
    FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (SELECT Max(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

- The query lists the names of the booster versions that have carried the maximum payload mass.

2015 Launch Records

```
%%sql SELECT CASE
        WHEN SUBSTR(Date, 6, 2) = '01' THEN 'January'
        WHEN SUBSTR(Date, 6, 2) = '02' THEN 'February'
        WHEN SUBSTR(Date, 6, 2) = '03' THEN 'March'
        WHEN SUBSTR(Date, 6, 2) = '04' THEN 'April'
        WHEN SUBSTR(Date, 6, 2) = '05' THEN 'May'
        WHEN SUBSTR(Date, 6, 2) = '06' THEN 'June'
        WHEN SUBSTR(Date, 6, 2) = '07' THEN 'July'
        WHEN SUBSTR(Date, 6, 2) = '08' THEN 'August'
        WHEN SUBSTR(Date, 6, 2) = '09' THEN 'September'
        WHEN SUBSTR(Date, 6, 2) = '10' THEN 'October'
        WHEN SUBSTR(Date, 6, 2) = '11' THEN 'November'
        WHEN SUBSTR(Date, 6, 2) = '12' THEN 'December'
    END AS Month,
Landing_Outcome,
Booster_Version,
Launch_Site
FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Failure (drone ship)'
    AND substr(Date,0,5)='2015';
```

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

Month	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

- The query lists the failed landing outcomes on drone ships, their booster versions, and launch site names for in year 2015

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

```
%%sql SELECT Landing_Outcome,
    COUNT(Landing_Outcome) AS count
        FROM SPACEXTABLE
        WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
        GROUP BY 1
        ORDER BY 2 DESC;
```

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

Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The query ranks 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.


Section 3
Launch Sites
Proximities Analysis

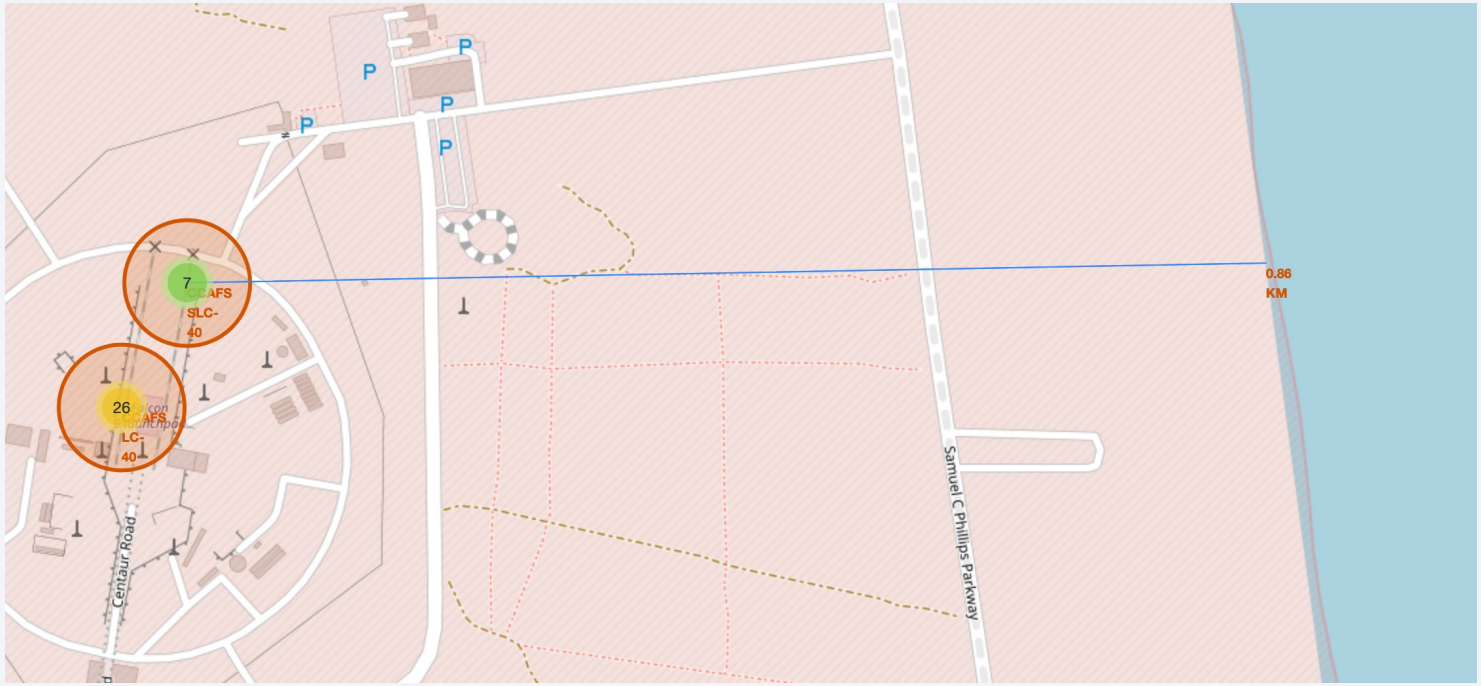
SpaceX's Global Launch Sites

- The above map marks all launch sites. Accordingly we can derive some similarities between launch sites, like for example that the launch sites are in proximity to the Equator line and in very close proximity to the coast.

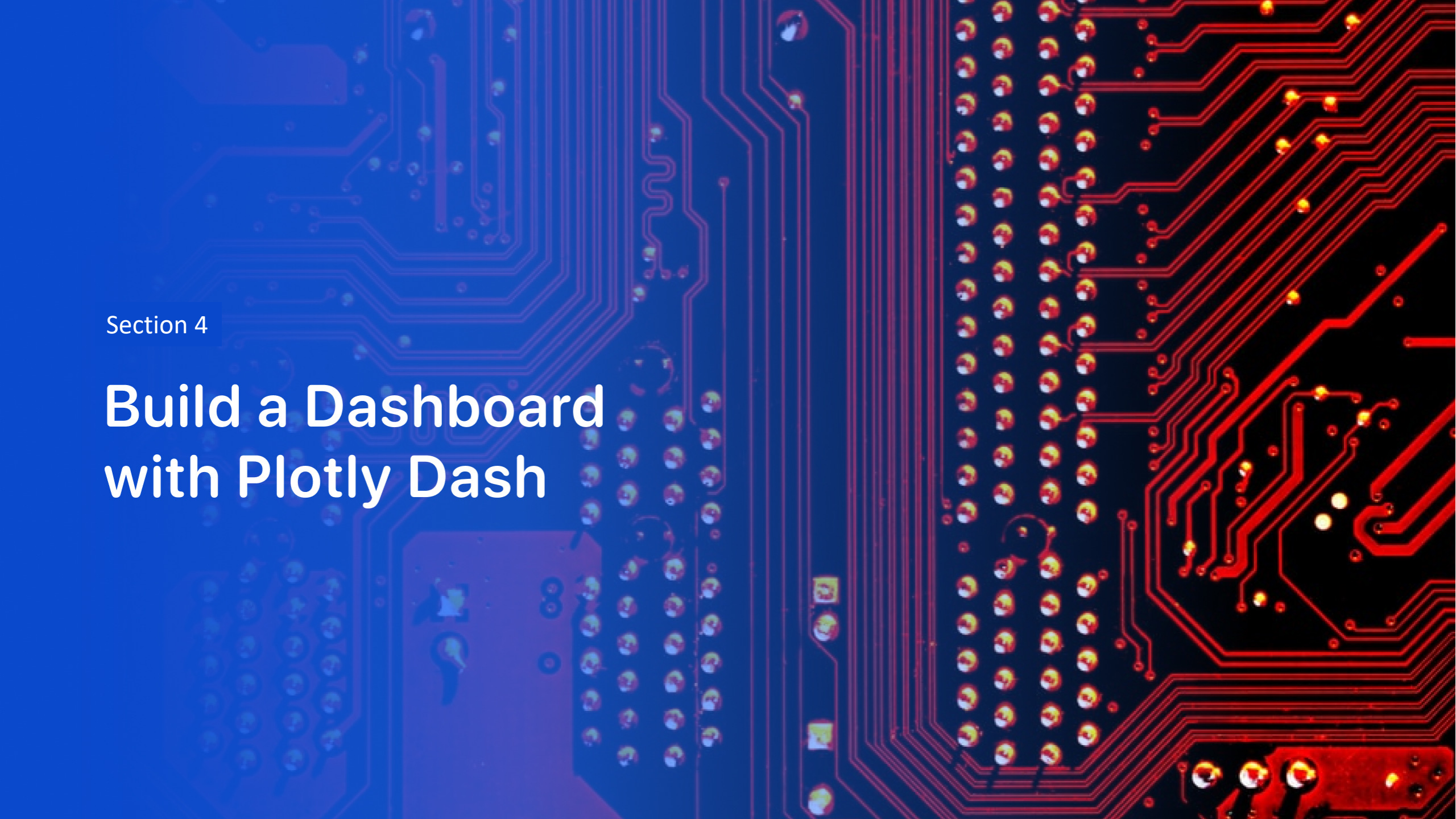
Successful/Failed Launches for each Site

- The above image shows an example for marking the successful/failed launches (green/red tags) for each launch site on the map to identify the most successful among launch sites, which in the case of Falcon 9 first stage rockets is launch site KSC LC-39A.

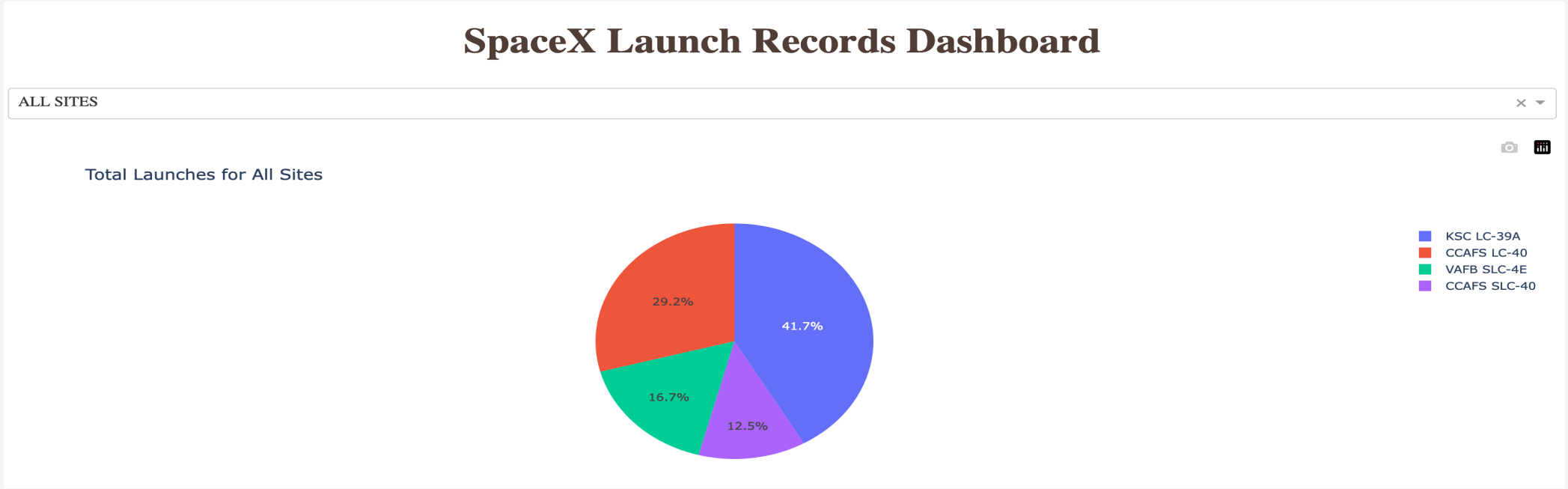
Proximities to Launch Sites



- The above image shows an example of calculating the distances between a launch site to its proximities, for example to the proximity to the nearest coastline.

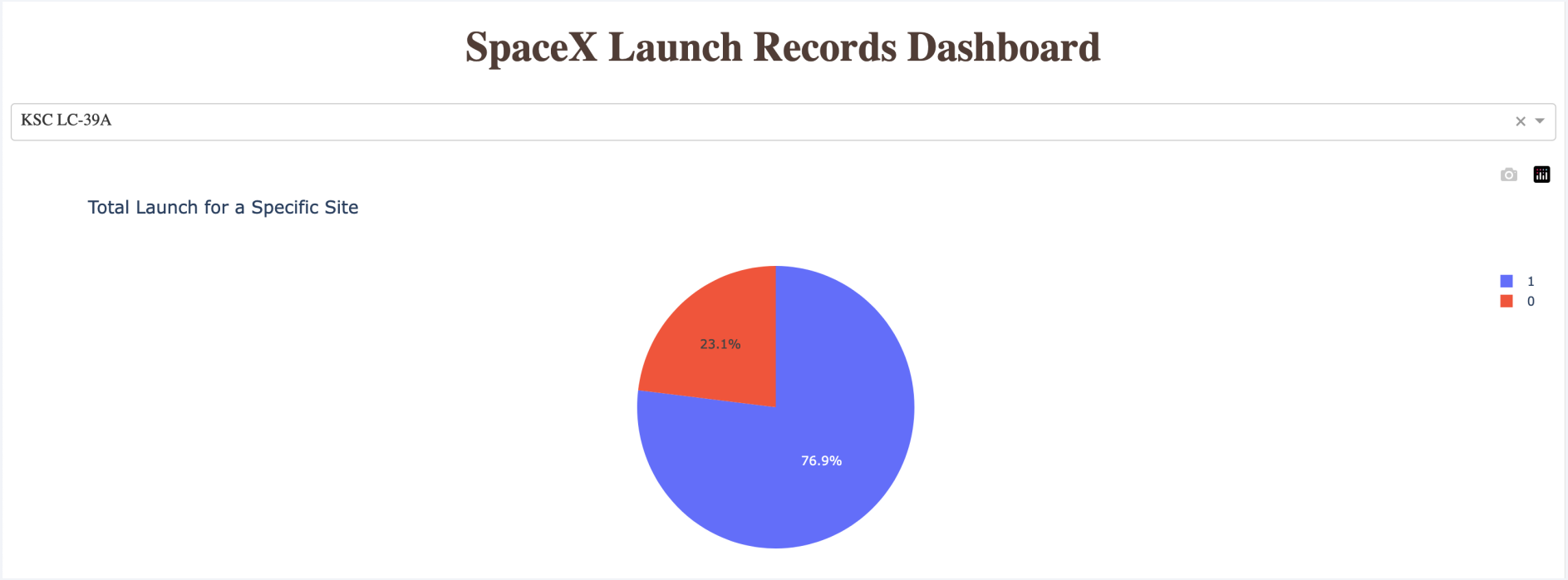

Section 4
Build a Dashboard with Plotly Dash

Share of Launch Site in Total Successful Launches



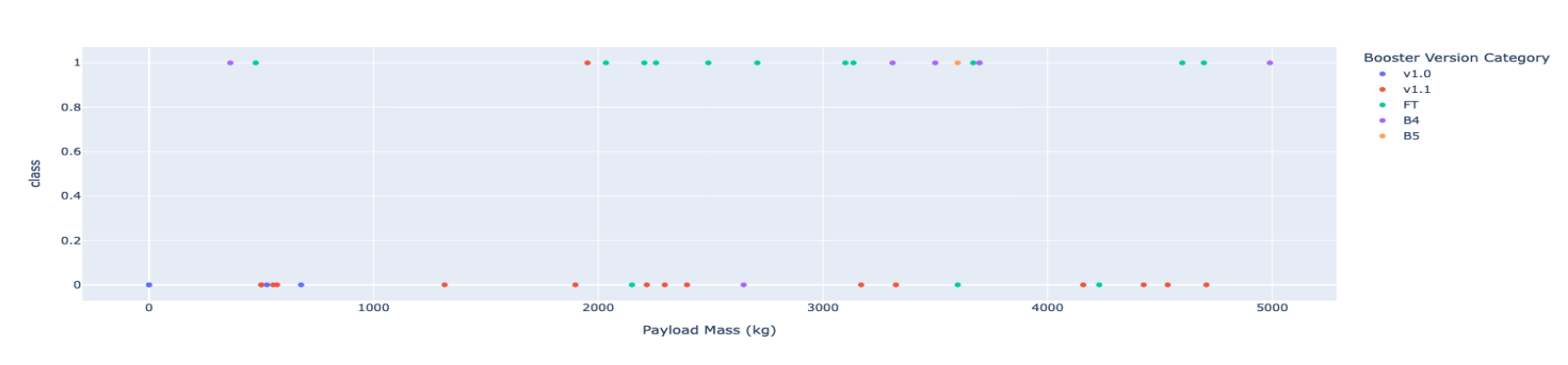
- The above image shows a dashboard depicting a pie chart highlight the site with the most successful launches / success rate.
- As shown in the pie chart, launch site KSC LC-39A is the most successful in terms of Falcon 9 launches.

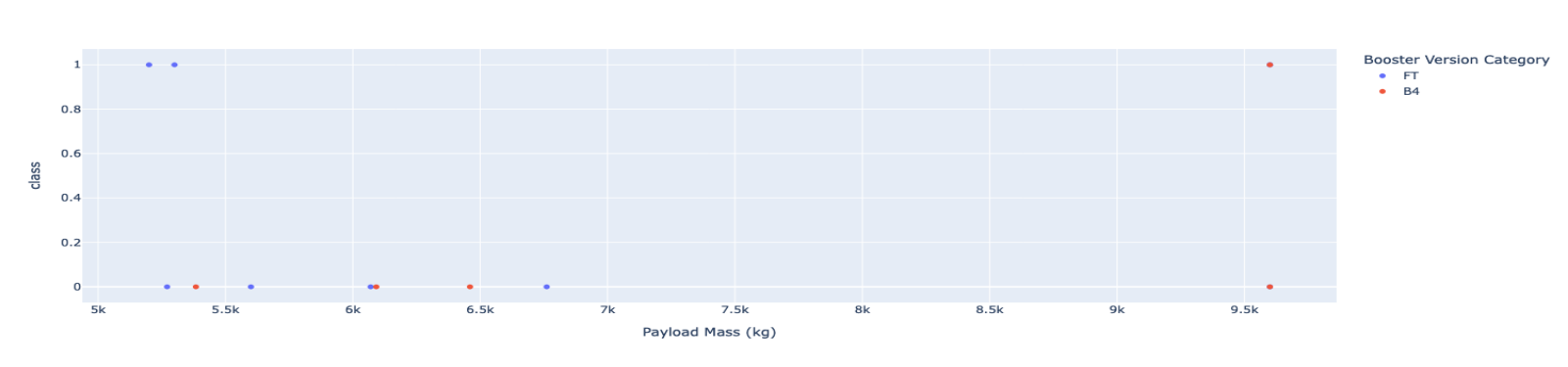
Success Rate of individual Launch Sites



- The same dashboard can be filtered for the mentioned most successful KSC LC-39A site, in order to view the site's individual success rate (76.9%) .

Effect of Payload Mass on Launch Outcome



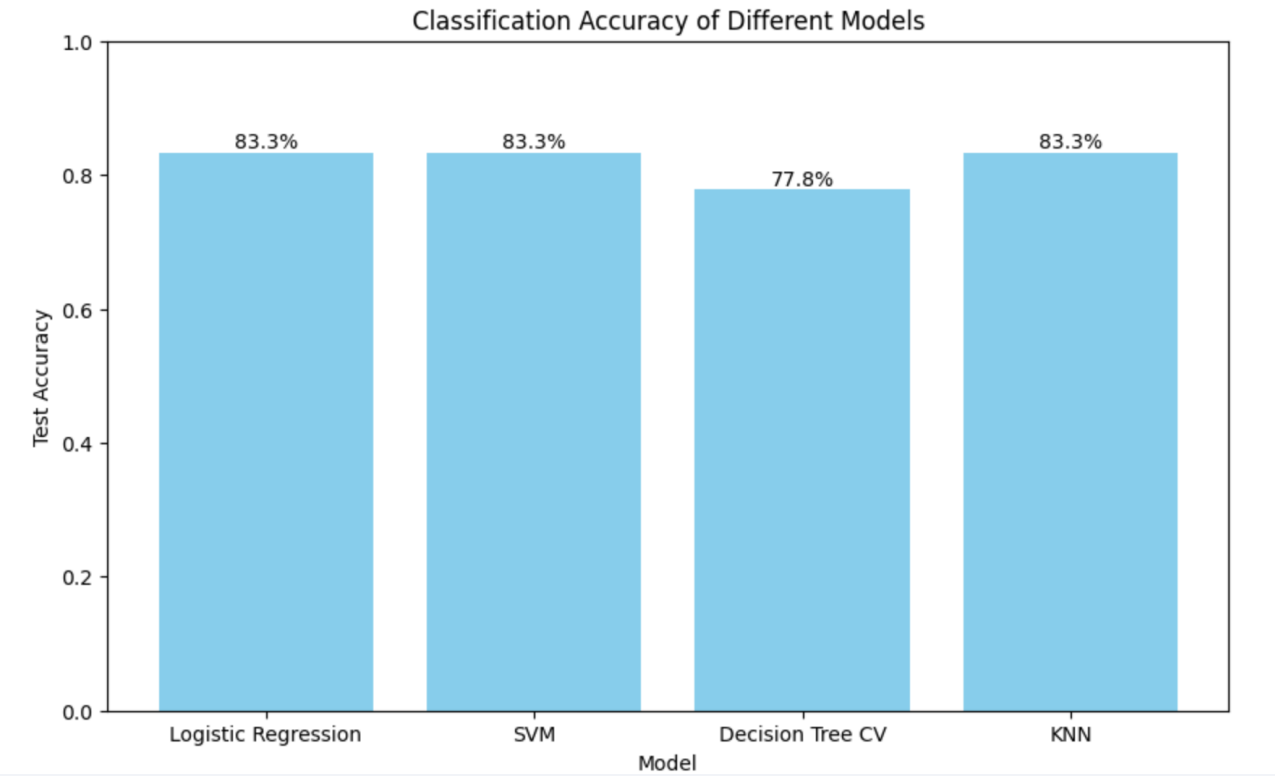


- The above scatter plots obviously shows that the smaller/lighter the payload mass the higher the success chances of launches are.
- Whereas for payloads from 5'000 Kg and above, only Booster Versions FT & B4 are able to land successfully after launch.

Section 5

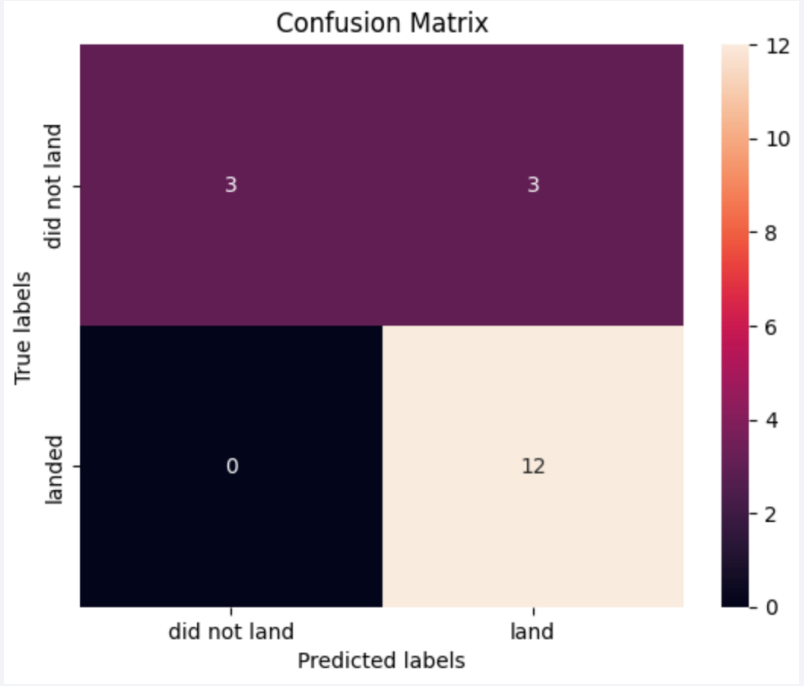
Predictive Analysis (Classification)

Classification Accuracy



The SVM, KNN and Logistic regression models are the best in term of prediction accuracy of the dataset with an accuracy of 83.3%.

Confusion Matrix



- The above confusion matrix corresponds to SVM, KNN and Logistic regression models, which are the best in terms of prediction accuracy of the dataset with an accuracy of 83.3%.
- As the matrix highlights, all three models suffer from a degree of accuracy loss or better said false positive predicted outcomes.

Conclusions

- The EDA shows that as the flight number increases, the first stage is more likely to land successfully.
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- Launch sites: KSC LC-39A and VAFB SLC 4E have the highest first stage landing success rate with 77%
- Launches aimed towards ES-L1, GEO , HEO and SSO orbits have the highest landing success rate among other orbits.
- The more Space X continues to launch rockets over the years the more successful landings they achieve. It seems therefore that each launch adds valuable experience in regards to optimizing the launches and respective landing procedures.
- In terms of predicting the launch outcomes of Falcon 9 first stage rockets, the SVM, KNN and Logistic regression models are the best in term of prediction accuracy of the dataset with an accuracy of 83.3%.

Appendix

- For the lab notebooks and related outputs, please visit the following GitHub repository: https://github.com/ameleha/IBM_ds_capstone_submission

Thank you!