
IBM Developer
SKILLS NETWORK
Winning the Space Race with Data Science
Shashank Khobragade
25.10.2025

Outline

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

Executive Summary

Summary of Methodologies

1. **Data Collection**
   - API and Web Scraping was used to collect launch data for SpaceX
2. **Data Wrangling**
   - Data was cleaned by removing redundant data
   - Target variable was identified and converted to dummy
3. **EDA with Data Visualization**
   - Visualization was used to identify key patterns and relationships between various parameters
4. **EDA with SQL**
   - SQL queries were used to gain further insights on the launch data
5. **Interactive Map with Folium**
   - An interactive map was built using Folium to identify factors relevant to the launch site locations
6. **Dashboard with Plotly Dash**
   - An interactive visualization dashboard was used to investigate factors affecting successful launches
7. **Predictive Analysis (Classification)**
   - Four classification models were compared to find the most efficient one

Summary of Results

EDA Results

- Identified relationships between parameters affecting successful launch
- Compared correlation between geographical location of launch sites and their impact on success rate

Model Performance

- The four classification models - Logistic Regression, Support Vector Machine, Decision Tree Classifier, and k-Nearest Neighbors - returned an accuracy score of 8.33
- None of the models was found to be better than the other

Key findings

- Launch site, launch orbit and payload mass affect the success outcome of a launch
- Further methods and/or metrics must be used to find the best classification model, or more models must be built and tested

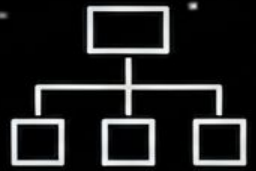
Introduction

Project Background and Context

This project seeks to identify the factors leading to the successful landing of Falcon 9 rockets. SpaceX's reusable rockets have made them a reliable and cost-effective launch provider. Understanding their success factors can inform strategies for investments, launch sites, and orbit selection to compete in the space industry.

Problems we need to find answers to

- ➔ Which factors determine the outcome of the landing of Falcon 9?
- ➔ Can we use machine learning to predict the landing outcome accurately?
- ➔ Which machine learning model performs best?



Methodology



Methodology

Executive Summary
This project analyzes SpaceX data to predict the landing outcome of Falcon 9 rockets. The process includes Data Collection, Data Wrangling, Exploratory Data Analysis, Interactive Visualizations, and Predictive Analysis using Classification models.

Data Collection Methodology
SpaceX API and Wikipedia table for Falcon 9 launches was used as the data source.

Perform Data Wrangling
Data cleaning was done by handling the missing values, identifying the target variable and converting to dummy for binary classification.

Perform Exploratory Data Analysis using Visualization and SQL
Initial insights were gained through data visualization and SQL queries, and various relationships were investigated between the parameters and their impact on the landing outcome.

Methodology

Perform interactive visual analytics using Folium and Plotly Dash

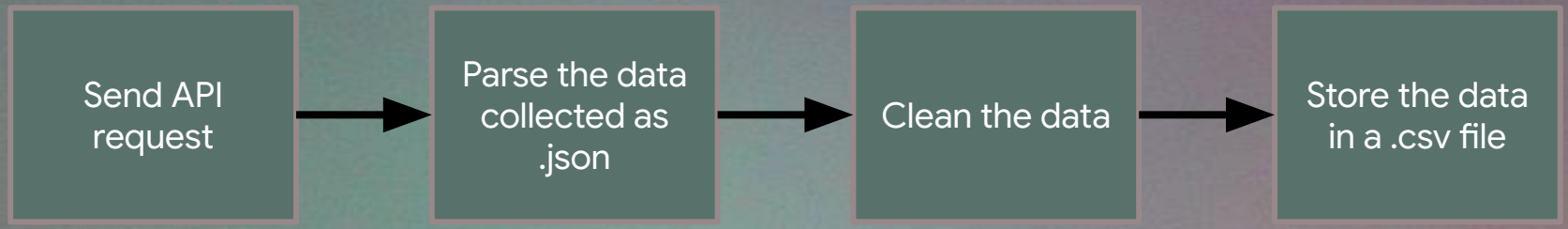
Folium and Plotly Dash was used to create interactive visualizations for quick insights on relationships between launch site locations.

Perform predictive analysis using classification models

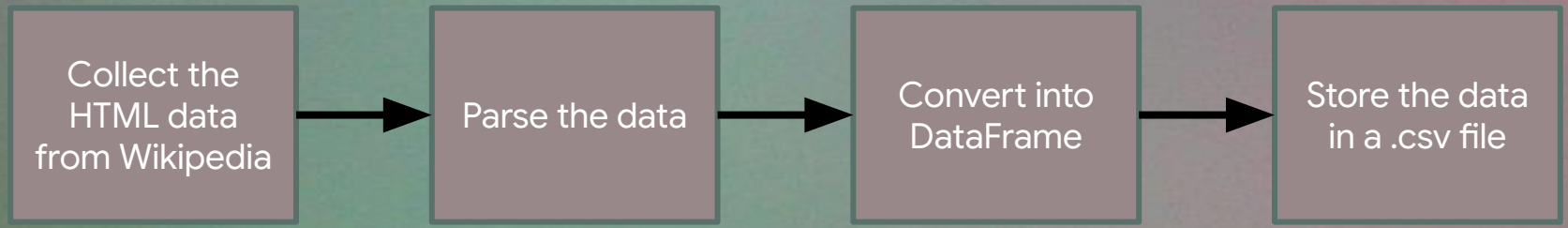
- Four classification models - Logistic Regression, Support Vector Machine, Decision Tree Classifier, and k-Nearest Neighbors - were tested
- The models were tested each with different hyperparameters using GridSearchCV
- The best set of parameters were found for each of them
- Accuracy scores and Confusion matrices were used to evaluate model performance

Data Collection

SpaceX API



Web Scraping



Data Collection: SpaceX API

Step 1: Send API request

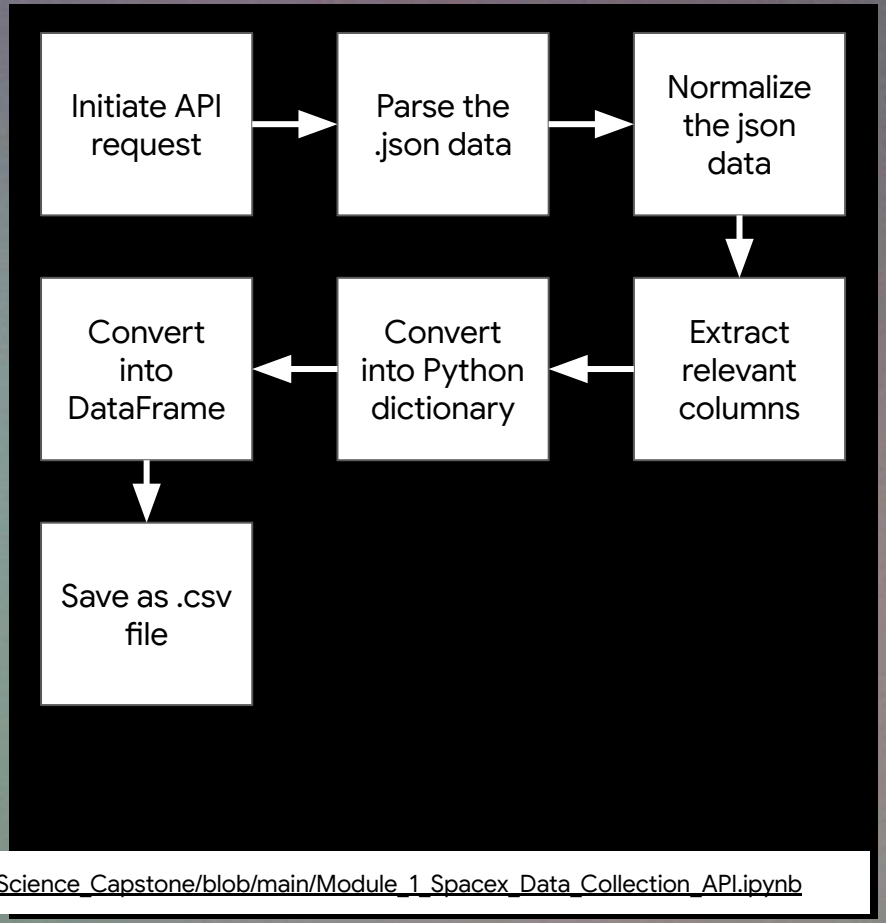
- Use Python request get method to import the data
- Connect to the SpaceX API endpoint: https://api.spacexdata.com/v4/launches/past

Step 2: Parse the response

- Normalize the .json structure
- Convert the data for the relevant columns into Python dictionary

Step 3: Store data in .csv file

- Convert the data into a DataFrame
- Save the DataFrame as .csv file



GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_1_Spacex_Data_Collection_API.ipynb

Data Collection: Web Scraping

Step 1: Import the HTML table

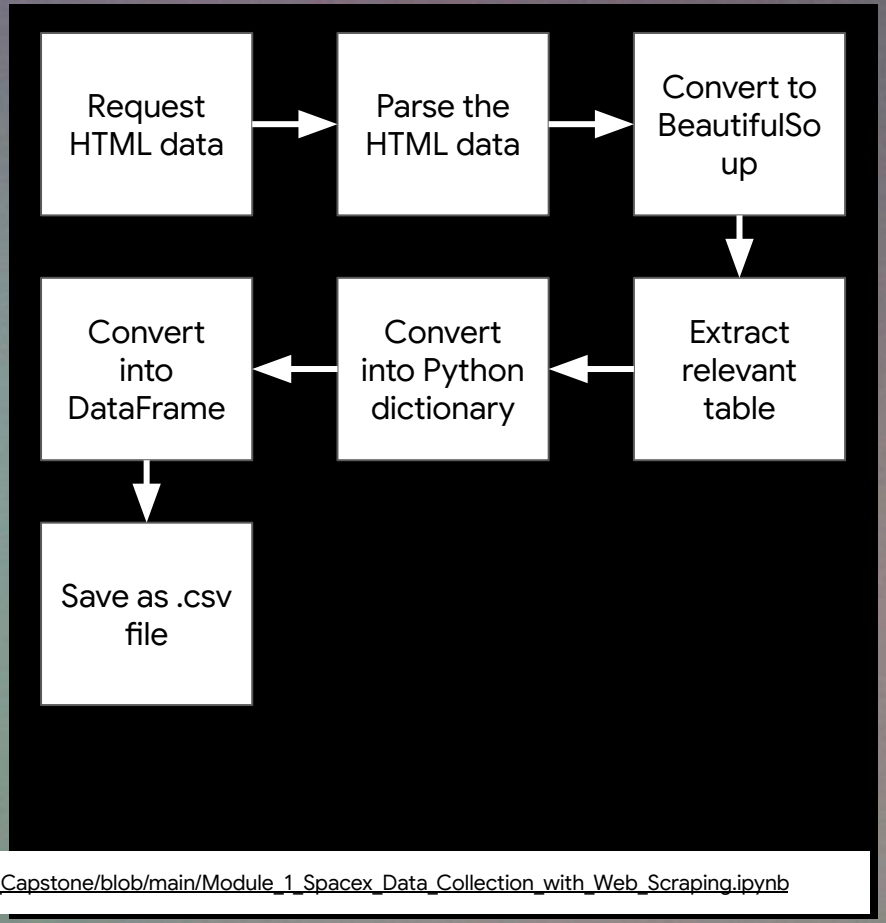
- Use Python request get method to import the HTML table
- Wikipedia URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Step 2: Parse the HTML content

- Use BeautifulSoup for parsing the HTML content
- Extract the table data for Falcon 9

Step 3: Store data in .csv file

- Convert the data into a DataFrame
- Save the DataFrame as .csv file



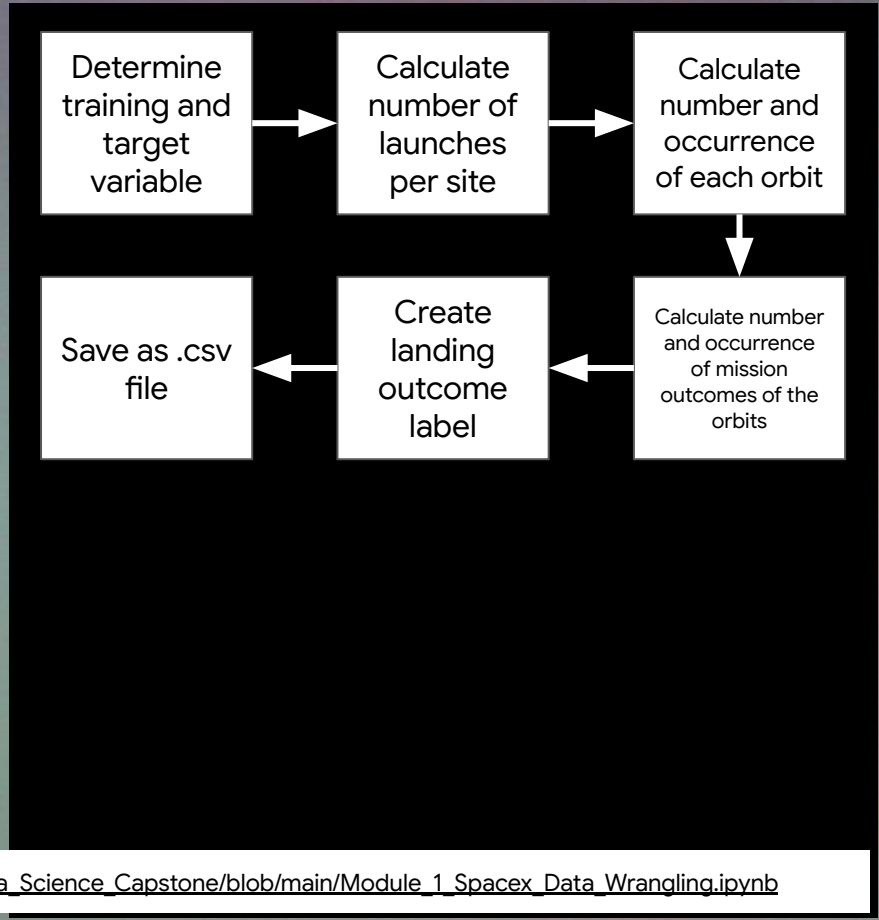
GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_1_Spacex_Data_Collection_with_Web_Scraping.ipynb

Data Wrangling

Step 1: Data Cleaning
- Replace missing values in the dataset

Step 2: Data Transformation
- Standardize the data types
- Convert the target variable into dummy **1** and **0**

Step 3: Store data in .csv file
- Save the DataFrame as .csv file



GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_1_Spacex_Data_Wrangling.ipynb

EDA with Data Visualization

Data Visualization with various charts was used to find patterns in the dataset visually

Scatter Plot with Hue

- Scatter plots were used to identify the relationships between two variables at a time
- The outcome was used as hue to visualize their relationship with the target variable

Bar plot

- Bar plot was used to visualize the average success rate in each orbit

Line chart

- Line chart was used to visualize the yearly trend of the outcomes

GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_2_EDA_with_Visualization.ipynb

EDA with SQL

SQL querying provided understanding of the SpaceX dataset

Finding unique values

- DISTINCT querying was used to find unique launch site and landing outcome details

Aggregated data

- SUM and AVG were used to find Payload Mass data
- COUNT was used to calculate the number of successful and failed outcomes

Conditional queries

- WHERE was used in combination with the above queries to restrict data filtering based on certain conditions
- AND and LIKE were predominantly used to implement the conditions

GitHub URL:https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_2_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

Map Objects

Markers

- Indicate specific launch site locations.
- Also show success and failure counts.

Circles

- Highlight the proximity area for launch sites.

Lines

- Visualize distances from launch sites to the coastline, railway, highway, and nearest city.

Reasons to add Map Objects

Markers

- Specify exact launch site locations for initial spatial orientation.

Circles

- Highlight the immediate area surrounding the launch sites.

Lines

- Help find other factors determining launch site locations.

GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_3_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

Plots/Graph and Interactions Added

Pie Chart with Dropdown

- Pie chart shows the success distribution for launch sites
- The dropdown helps narrow down the success - failure distribution for the selected launch site

Scatter Plot with Range Slider

- Scatter plot shows the relationship between Payload Mass and Outcome
- The Range Slider helps select the Payload Mass range to narrow down the relationship

Reasons to add Map Objects

Pie Chart with Dropdown

- The top level pie chart shows a clear distribution of which launch site has better success rate
- The dropdown helps to go one level down to see the exact distribution for the site

Scatter Plot with Range Slider

- The top level scatter plot shows the entire relationship for one parameter and the target variable
- The slider helps in visualizing this relationship for various payload ranges

GitHub URL:
https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_3_Interactive_Dashboard_with_Plotly_Dash.py

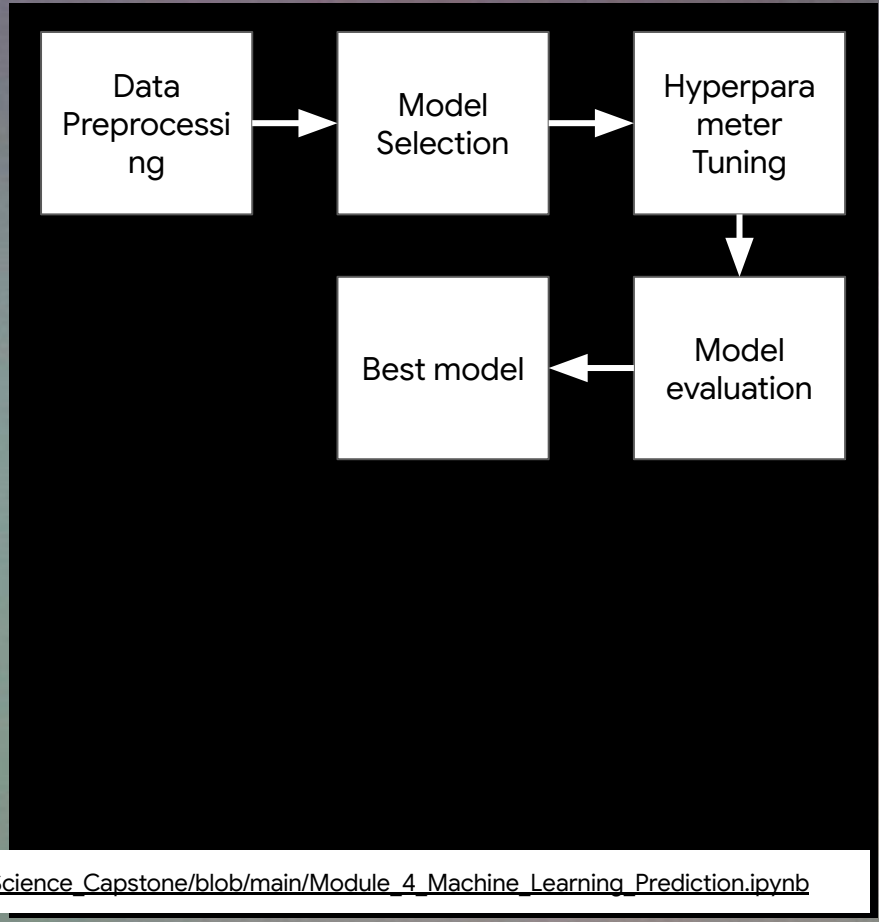
Predictive Analysis (Classification)

Data Preprocessing
- Standardize the data
- Split the data into training and testing

Model Selection
- Test four classification models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, k-Nearest Neighbors

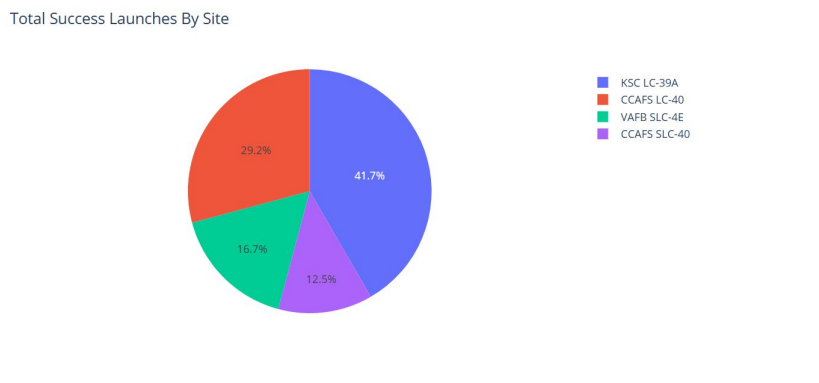
Parameter Hypertuning
- Use GridSearchCV to hypertune the respective model parameters

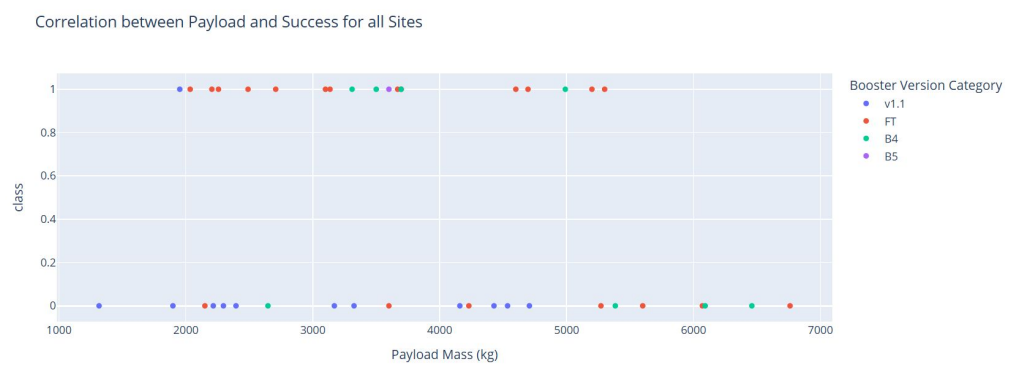
Model Evaluation
- Use Accuracy score and Confusion matrix to compare model performance



GitHub URL: https://github.com/FootlooseNFree/Coursera_IBM_Applied_Data_Science_Capstone/blob/main/Module_4_Machine_Learning_Prediction.ipynb

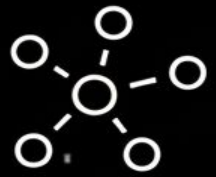
Results





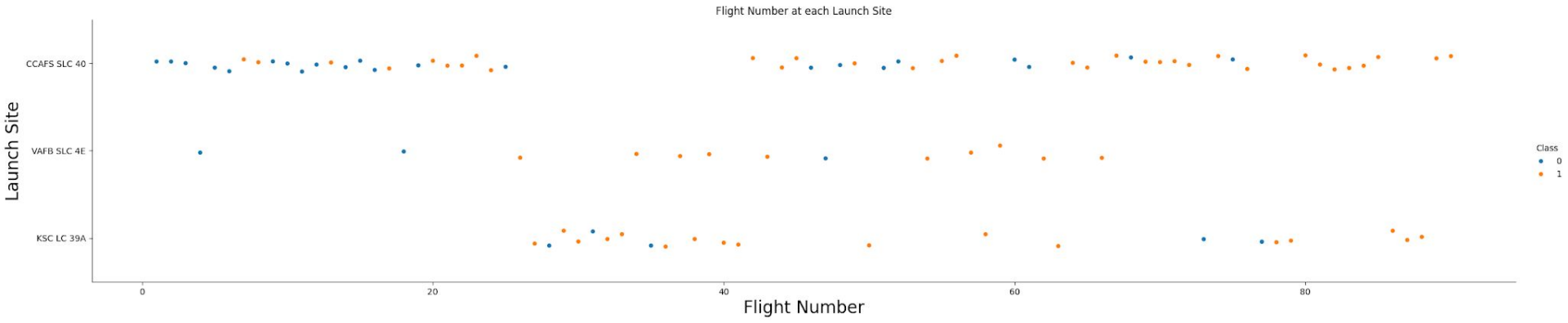

```
models = pd.DataFrame({'Model': ['Logistic Regression', 'Support Vector Machines', 'Decision Tree', 'K-Nearest Ne
print(models.sort_values(by='Accuracy', ascending=False))

sns.barplot(data = models, y='Model', x='Accuracy', hue='Model', palette=['red','blue','yellow','green'])
plt.title('Accuracy of Different Models', color='blue')
plt.show()
```

```
                   Model  Accuracy
0    Logistic Regression  0.833333
1  Support Vector Machines  0.833333
2          Decision Tree  0.833333
3    K-Nearest Neighbors  0.833333
```



Exploratory Data Analysis

Flight Number v/s Launch Site

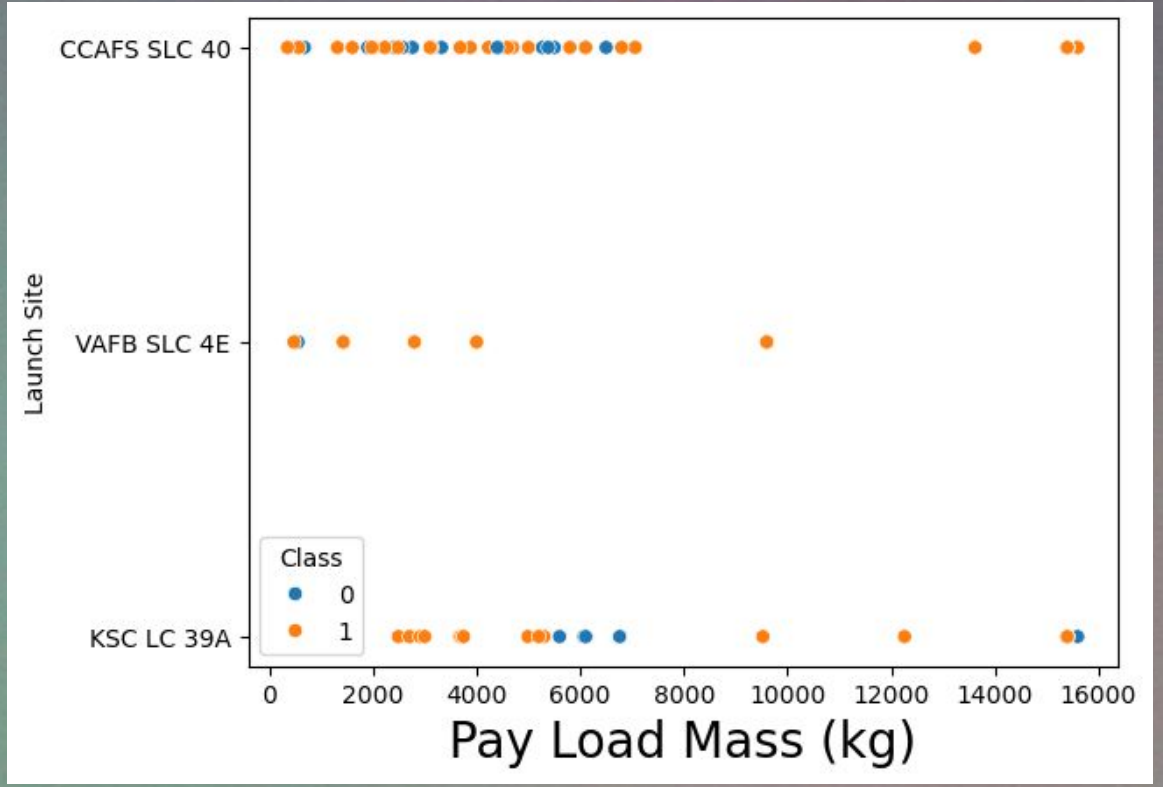


Flight Number v/s Launch Site

With the increasing flight numbers, the outcome at each launch site appears to to be positive. The launches at VAFB SLC 4E seems to have peaked at around 60 flights.

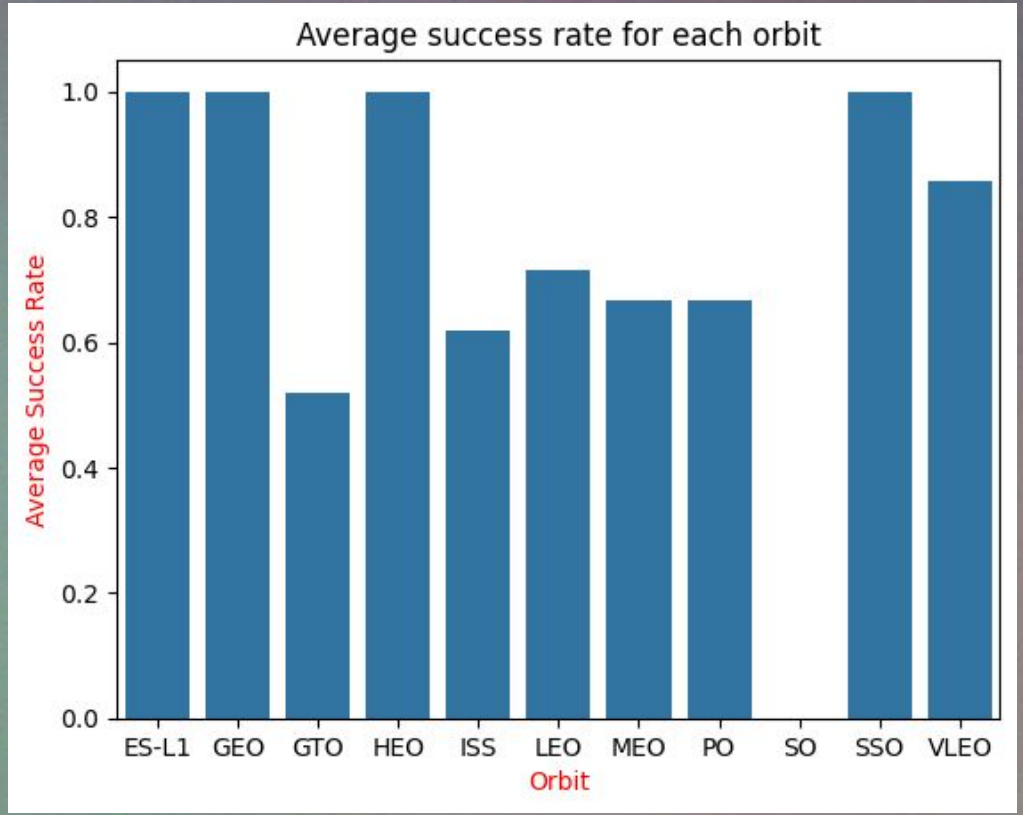
Payload Mass v/s Launch Site

Payload Mass v/s Launch Site
There seems to be no direct correlation between the payload mass and success outcome at any of the launch sites.



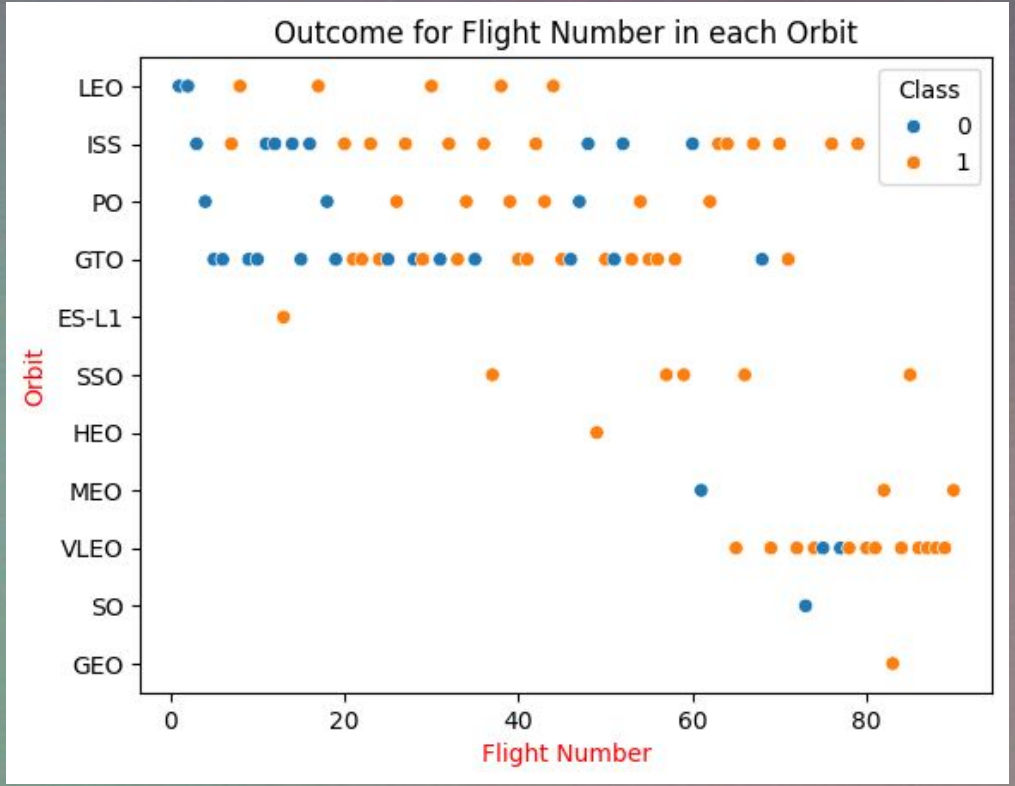
Success Rate v/s Orbit Type

Success Rate v/s Orbit Type
Launches in the orbit types ES-L1, GEO, HEO and SSO guarantee successful outcomes, while GTO orbits have the lowest success rates. Others are moderately successful.



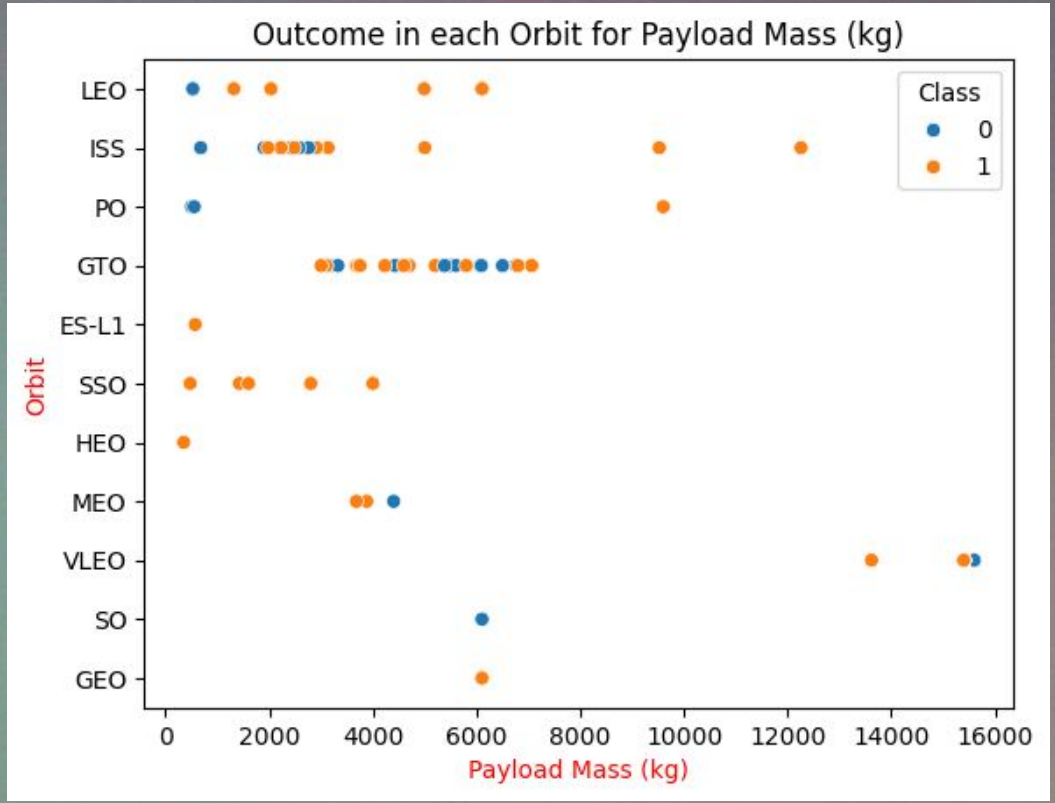
Flight Number v/s Orbit Type

Flight Number v/s Orbit Type
The scatter plot suggests that the Orbits ISS, SSO and VLEO can be trusted for successful launches, especially after having launched more than 70 rockets in them.



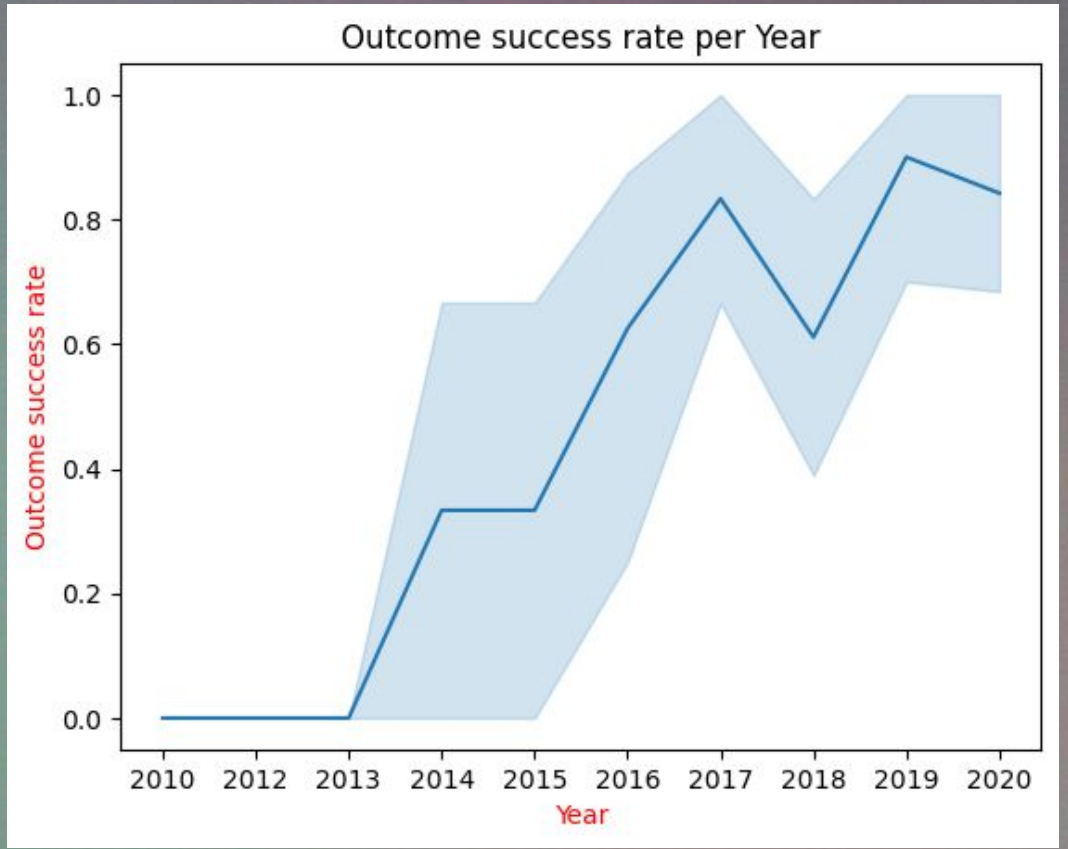
Payload v/s Orbit Type

Payload v/s Orbit Type
The appears to be no direct correlation between the Payload Mass carried in each Orbit type and the launch outcome. For SSO, ES-L1 and HEO all the launches have been successful at lower payloads, but the number of launches is very low.



Launch Success Yearly Trend

Launch Success Yearly Trend
The launches have a better success rate between the years 2017 and 202 with the exception of 2018. This could be attributed to factors like better technology, higher investments, etc.



All Launch Site Names

There are four distinct launch sites used by SpaceX for rocket launches. Two of them appear to be in close vicinity to each other namely CCAFS SLC and LE 40.

```
%%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

Launch Site Names Begin with ‘CCA’

The initial 5 launches for CCAFS LC-40 had a successful mission outcome but the landing outcome was negative. The dates show it was between 2010 and 2013 suggesting it was still an early stage in reusable rocket technology.

```
%%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	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

NASA (CRS) is a trusted customer of SpaceX which can be seen by the payload mass it has carried for them i.e. 45596 kg.

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

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

Total_payload_mass_NASA_CRS_kg
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by F9 v1.1 rockets is relatively small, indicating that it is perhaps not so suitable for higher amount of payload mass.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_payload_mass_F9v11
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Avg_payload_mass_F9v11
2534.6666666666665

First Successful Ground Landing Date

The first successful landing outcome for ground pad was in late in the year 2015, which could have triggered a better success rate in the following years.

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

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

First_landing_date_successful
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Four specific booster versions have been able to make a successful drone ship landing for payload mass between 4000 kg and 6000 kg.

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

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

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

Total Number of Successful and Failure Mission Outcomes

There has been only one failed mission outcome. This does not give any information on the landing outcome, however. Hence, understanding how mission and landing outcomes are measured is important.

```
%%sql
SELECT COUNT(Mission_Outcome) AS Total_Successful_Mission_Outcomes
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%Success%';
```

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

Total_Successful_Mission_Outcomes
100

```
%%sql
SELECT COUNT(Mission_Outcome) AS Total_Failure_Mission_Outcomes
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%Failure%';
```

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

Total_Failure_Mission_Outcomes
1

Boosters Carried Maximum Payload

With 12 different Booster Versions being able to carry the maximum payload, it can be said that the technology can be generally trusted with large payload.

```
%%sql
SELECT 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

2015 Launch Records

The only two failed launches in the year 2015 came in January and April.

```
%%sql
SELECT Booster_Version, Launch_Site, substr(Date, 6, 2) AS 'Month',
substr(Date, 1, 4) AS 'Year'
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure%' AND substr(Date, 1, 4) = '2015';
```

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

Booster_Version	Launch_Site	Month	Year
F9 v1.1 B1012	CCAFS LC-40	01	2015
F9 v1.1 B1015	CCAFS LC-40	04	2015

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

Between 2010 and 2017, the success and failure landing outcome was more or less equal. Ground ship landing seemed to have more success, while parachute landings failed twice.

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

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

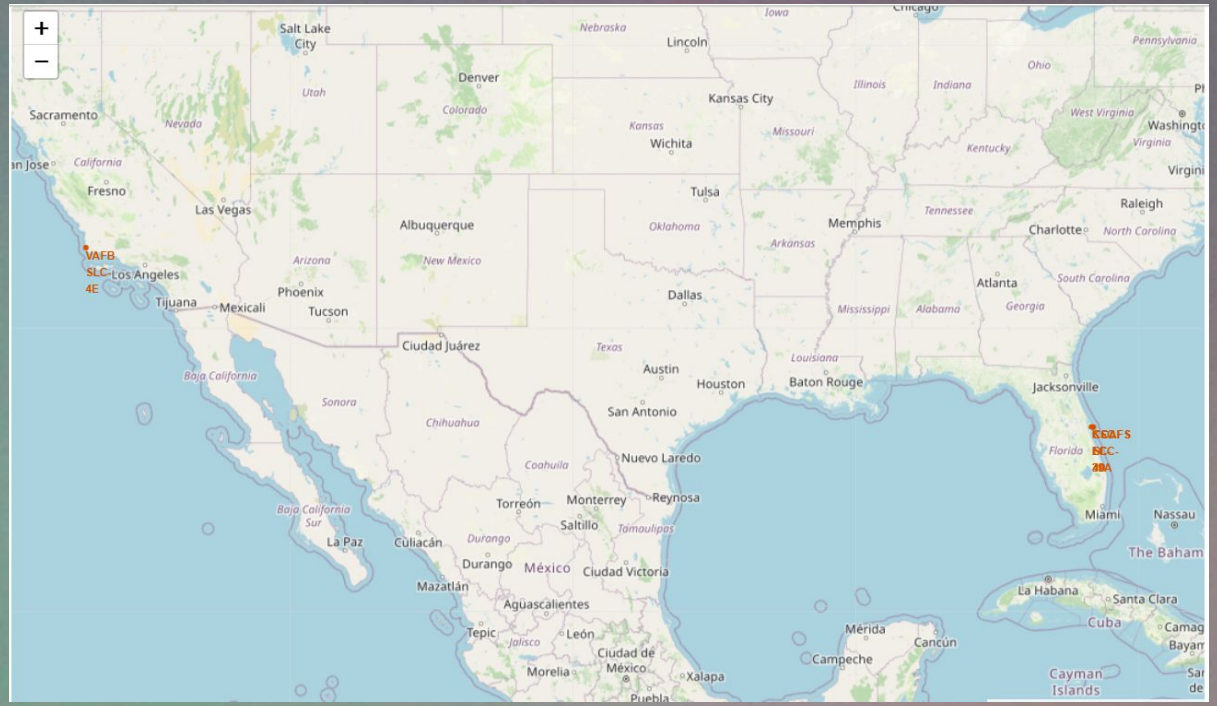
Landing_Outcome	Count_landing_outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Launch Sites Proximities Analysis

Launch Site Locations

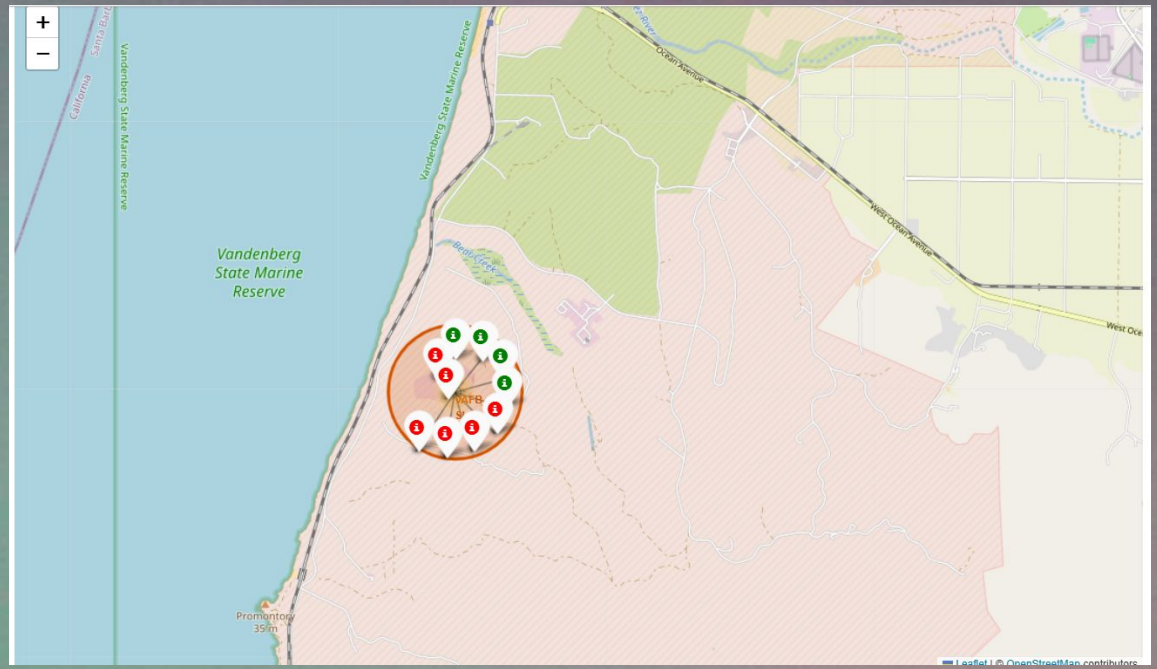
The launch site locations can be seen on the map on the Eastern and Western coastlines of the USA.

Their location further South indicate the intention to be as close as to the Equator.



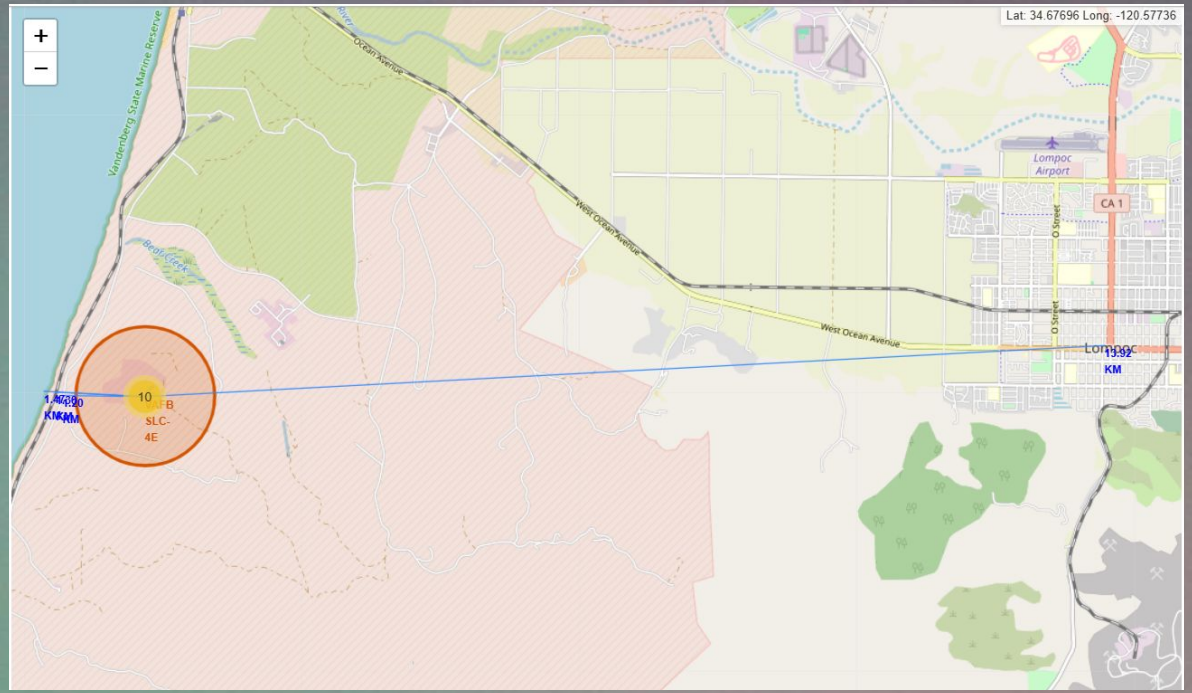
Launch Outcomes for VAFB SLC-4E

The launch site VAFB SLC-4E has a relatively low success rate. It is the only site among the four located at the Western US-Coast.



Launch Site Proximities for VAFB SLC-4E

VAFB SLC-4E is located less than 2 km away from the nearest coastline, railway and highway, while the closest city Lompac is around 20 km away.

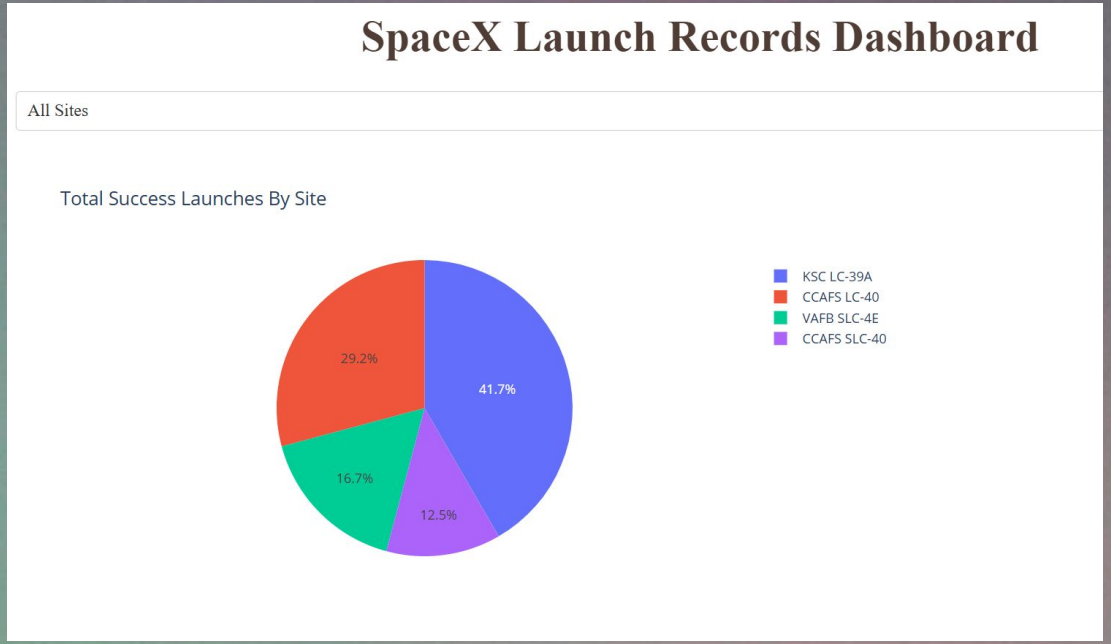


Dashboard with Plotly Dash

Success Launches Distribution for All Sites

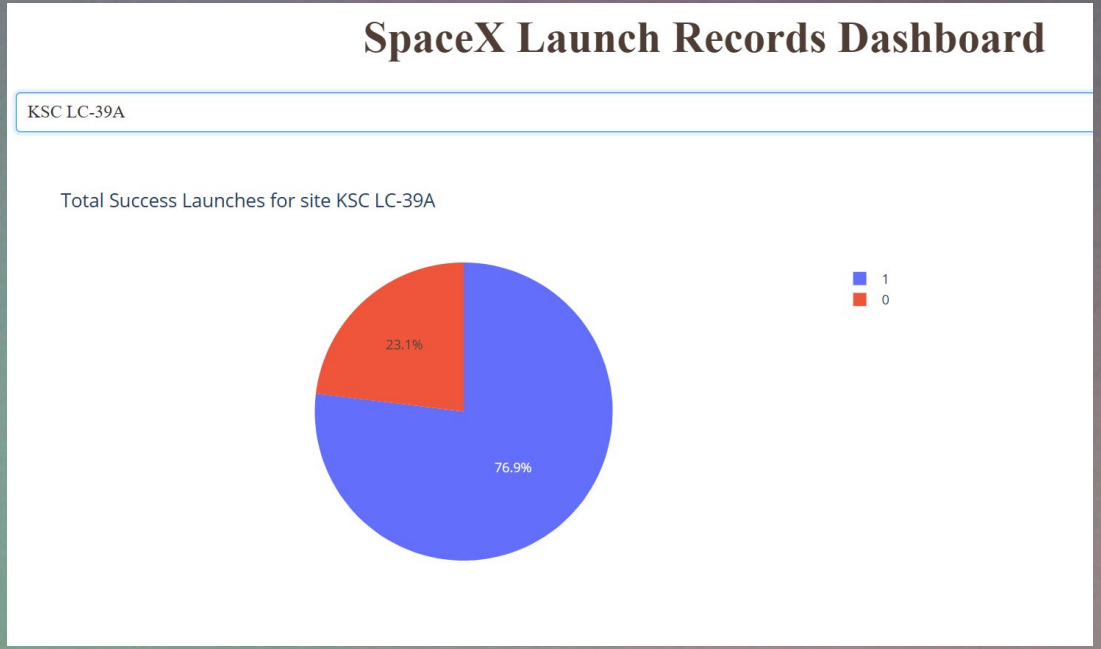
The pie chart shows that launches at KSC LC-39A have the majority of the share with 41.7% successful launches, followed by CCAFS LC-40.

The other two sites had a rather low success share.



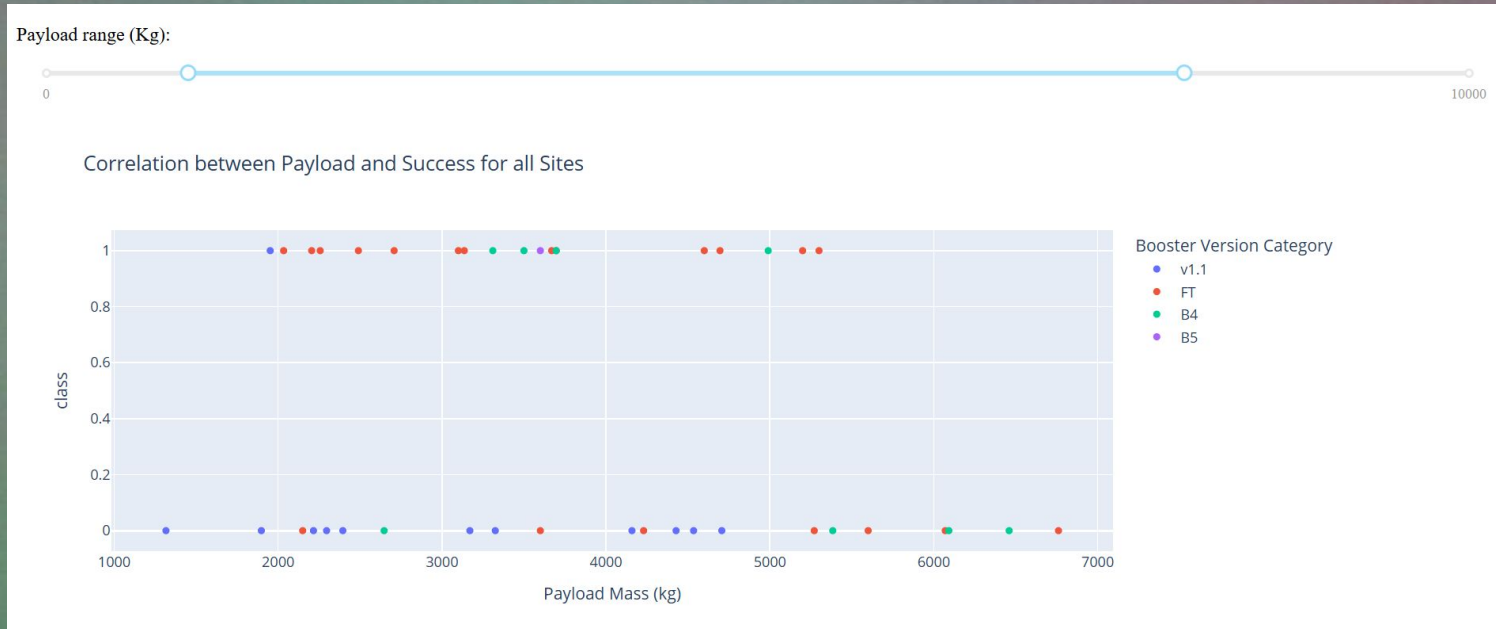
Launch Success Yearly Trend

A closer look at the launch outcomes at KSC LC-39A shows that more than 75% launches have a positive outcome, which is significantly high even if not perfect.



Launch Success Yearly Trend

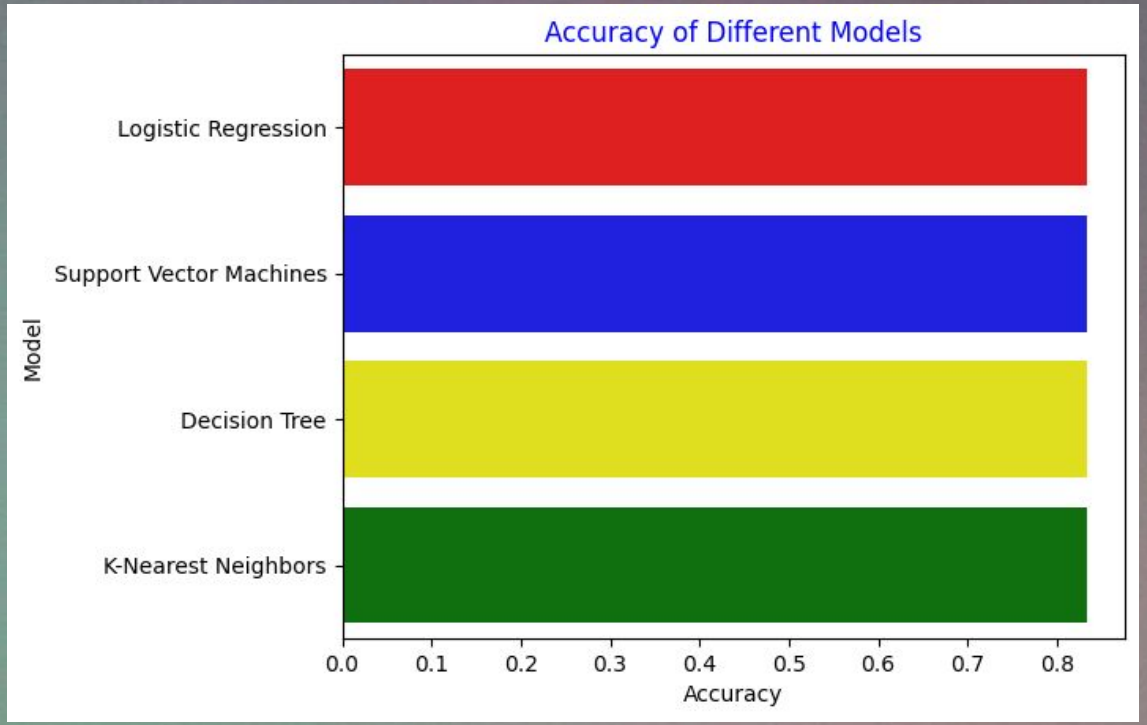
The relationship between Payload Mass and Success Outcome show no correlation at all at any of the launch sites. Hence, this parameter may not be the most ideal candidate while selecting model features.



Predictive Analysis (Classification)

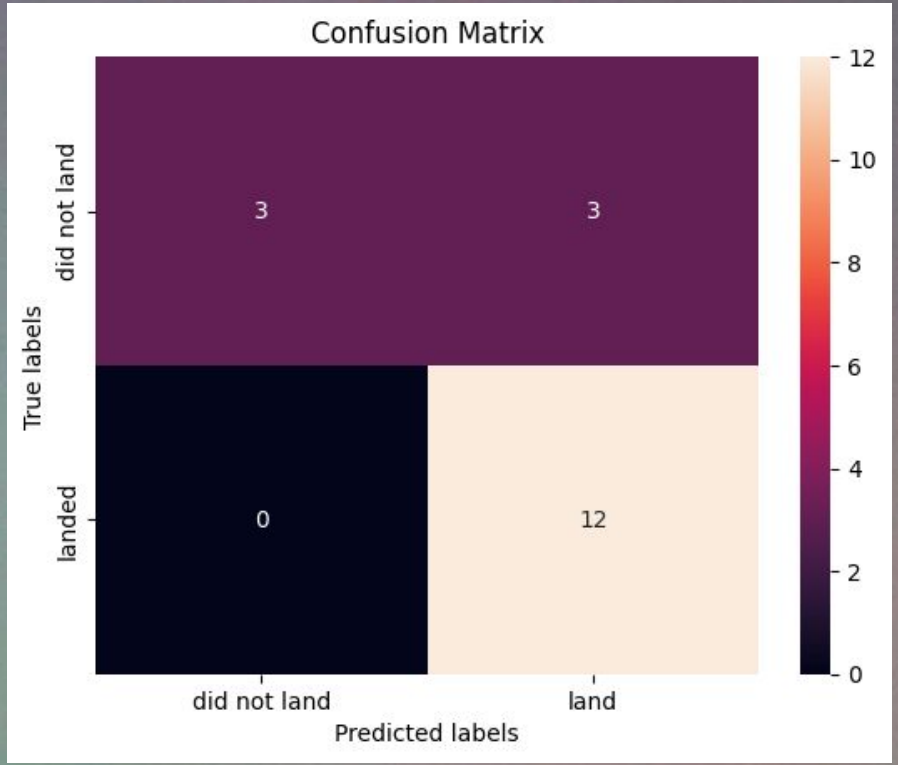
Classification Accuracy

The bar chart comparing the accuracy scores for the four classification models shows that there is not a clear winner.
All of them performed with the same accuracy of 8.33



Confusion Matrix

The confusion matrix represents the classification for all the four models, which is identical. They all are good at identifying True Positives and True Negatives, while showing three cases of False Positives.



Conclusions

Point 1
With more flights, the launch seem to get better i.e. there are more chances of a successful outcome with more experience.

Point 2
The payload mass does not seem to have any influence on the success outcome

Point 3
ISS, VLEO and SSO orbits have better success rates, which is again independent of the payload mass

Point 4
The launch sites have close proximity to the Equator. They are all very close to the coastline, and the Eastern coast appears to favour the launch outcome more. They all also have a railway, highway and cities located very near to them.

Conclusions

Point 5
None of the four classification models outperformed the other. This could be due to the small test size 18.

Point 6
Other classification models can be tested and evaluated, or the current models can be trained and evaluated with other parameters and/or train-test split size, or further cross validation methods.

Appendix

Tools

- Jupyter Notebook
- GitHub repository
- Google Slides

Python Libraries/Modules

- Pandas
- NumPy
- Scikit-learn
- Matplotlib
- Seaborn
- Sqlite3
- Folium
- Plotly Dash

Images

- Background: https://tinyurl.com/bgimageproject


Thank you!