Applied Data Science Capstone Project

Fábio Clemente Vilela

07/12/2024

Executive Summary

- SpaceX, the most successful company in the commercial space age, has revolutionized space travel by significantly reducing costs. While other providers charge upwards of $165 million per launch, SpaceX offers Falcon 9 rocket launches for $62 million. This cost efficiency is primarily due to the reusable first stage of the Falcon 9 rockets.

- This project aims to predict the likelihood of the Falcon 9 first stage landing successfully and being reused. Leveraging public information and machine learning models, this analysis will provide valuable insights for cost estimation and strategic planning for space missions. The outcome will support SpaceX's mission to make space travel more affordable and benefit the broader commercial space industry.

Table of Contents

Introduction

- SpaceX has emerged as a leader in the commercial space industry by making space travel more affordable. By reusing the first stage of its Falcon 9 rockets, SpaceX can offer launches at $62 million, significantly lower than the $165 million charged by other providers. Predicting whether the first stage will successfully land and be reused is crucial for determining the cost-effectiveness of each launch.

- This project uses public information and machine learning models to predict the likelihood of the Falcon 9 first-stage landing successfully. The analysis aims to answer the following questions:

    - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
    - Does the rate of successful landings increase over the years?
    - What is the best algorithm for binary classification in this case?

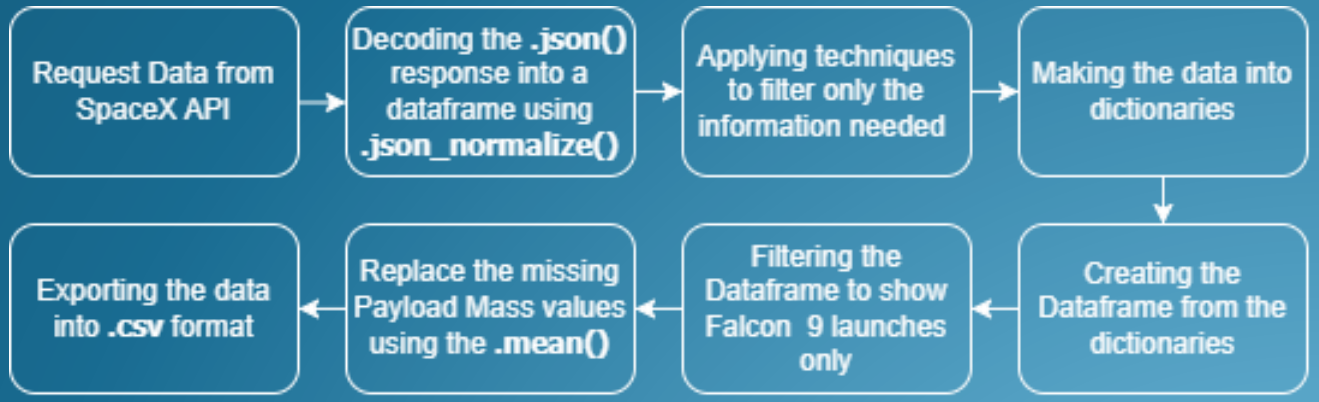
Methodology

To achieve these goals, the following steps will be undertaken:

1. Data Collection: Gathering data on Falcon 9 launches using SpaceX Rest API and web scraping.
2. Data Wrangling: Cleaning and preparing the data for analysis.
3. Exploratory Data Analysis (EDA): Analyzing the data to identify patterns and relationships.
4. Data Visualization: Creating interactive dashboards to visualize launch records and landing success.
5. Model Development: Developing machine learning models including Support Vector Machines (SVM), Decision Tree Classifiers, and K-Nearest Neighbors (k-NN).
6. Model Evaluation: Evaluating the performance of the models to determine the best one for predicting landing success.
7. Reporting: Compiling the findings into a comprehensive report for stakeholders.

Data Collection – API

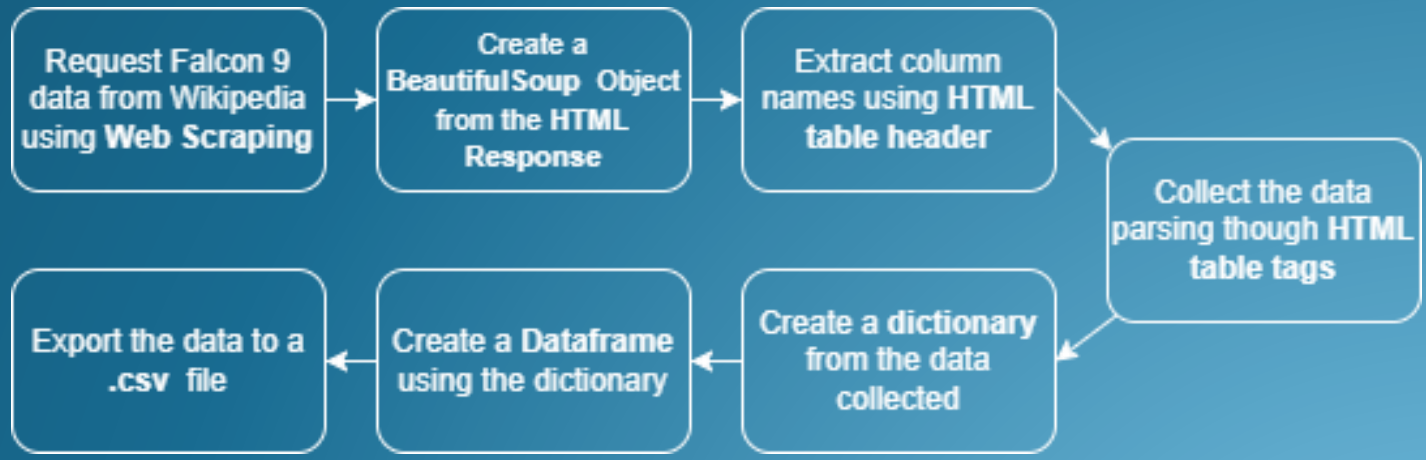
- In order to collect the Data, we used two sources: **SpaceX API**
- Decode the **.json()** file to a Dataframe using **.json_normalize()**
- Request information about launches using functions
- Create dictionaries from the data gathered
- Make a Dataframe from the Dictionaries
- Filter only Falcon 9 launches
- Replace the missing Payload Mass with the calculated **.mean()**
- Export the data to a **.csv** file



Github Link: Data Collection - API

Data Collection – Web Scraping

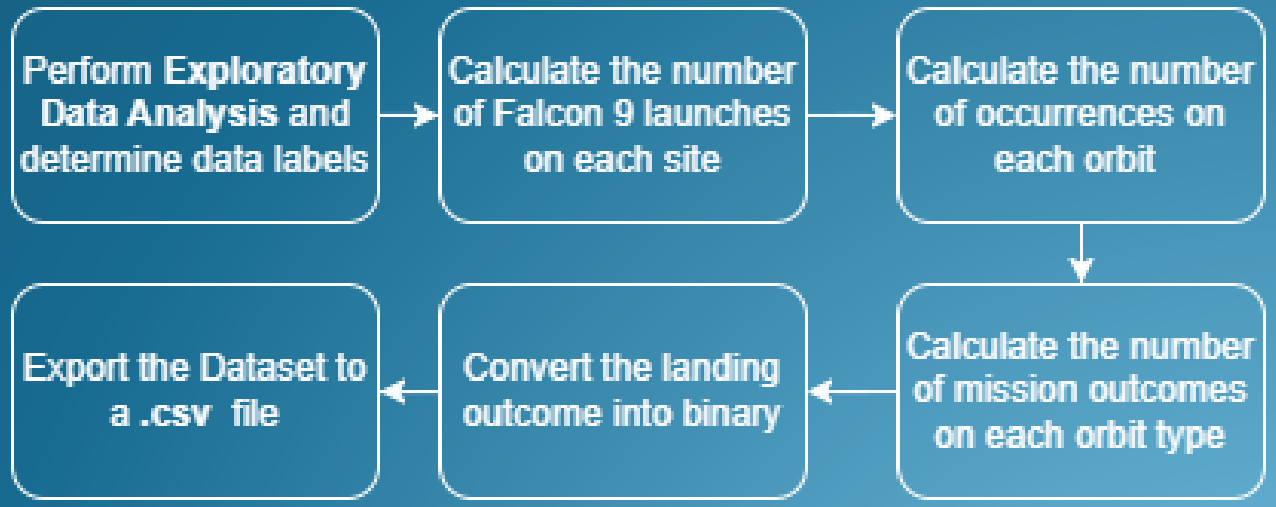
- Request Falcon 9 Data from Wikipedia using **Web Scraping**
- Create a **BeautifulSoup** object from the **HTML response**
- Extract column names using **HTML table header**
- Collect the Data parsing through HTML table tags
- Create a **dictionary** from the data collected
- Create a **Dataframe** using the dictionary
- Export the data to a **.csv** file



Github Link: Data Collection – Web Scraping

Data Wrangling

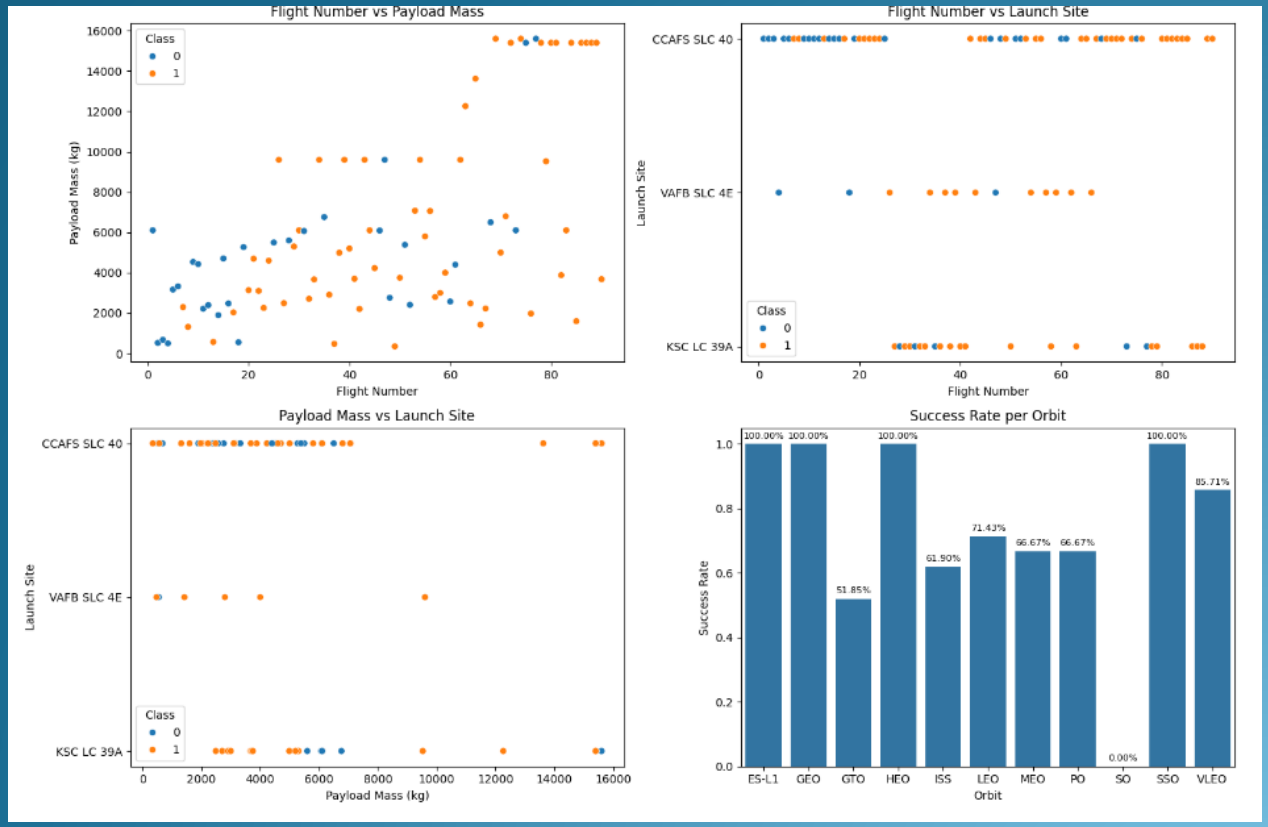
- Perform **Exploratory Data Analysis** and determine data labels
- Calculate the number of Falcon 9 Launches on each site
- Calculate the number of occurences on each orbit
- Calculate the number of mission outcomes on each orbit type
- Convert the landing outcome into binary
- Export the Dataset to a **.csv** file



Github Link: Data Wrangling

Data Visualization

- Charts Plotted:
  - Flight Number vs. Payload Mass(kg)
  - Flight Number vs. Launch Site
  - Payload Mass(kg) vs. Launch Site
  - Payload Mass(kg) vs. Orbit Type
  - Success Rate Yearly Trend
- **Scatter plots** are used to see relationships between variables
- **Bar charts** are used to compare discrete and categorical variables
- **Line charts** are used to view time series trends



Github Link: EDA – Data Visualization

SQL

- Queries:
  - The name of **unique** launch sites
  - 5 first records of launch sites beginning with "**CCA**"
  - Total Payload Mass carried by boosters launched by **Nasa (CRS)**
  - Average Payload Mass carried by **Falcon 9 v1.1 booster**
  - Date when the **First Booster** landed successfully on a ground pad
  - List of boosters that successfully landed in drone ship and have a payload mass **between 4.000 and 6.000 kg**
  - List of **total number** of failures and successful **mission outcomes**
  - List of names of the booster versions which have carried the **maximum payload mass**
  - List of **failed landing outcomes in drone ships**, their booster versions and launch site names for the **months in the year 2015**
  - Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Github Link: SQL

Interactive Map - Folium

- Lanch sites markers:
  - Added **markers** with a **circle**, **popup label,** and **text label** of all launch sites using their latitude and longitude coordinates to show their geographical locations

- Colored markers based on the outcomes:
  - Added **Green** markers for successful launches and **Red** for failed launches using **Marker Clusters** to show success rates

- Distance to its proximities:
  - Added additional lines and markers to show the **launch site proximities** with railroads, highways, coastline, and nearest city

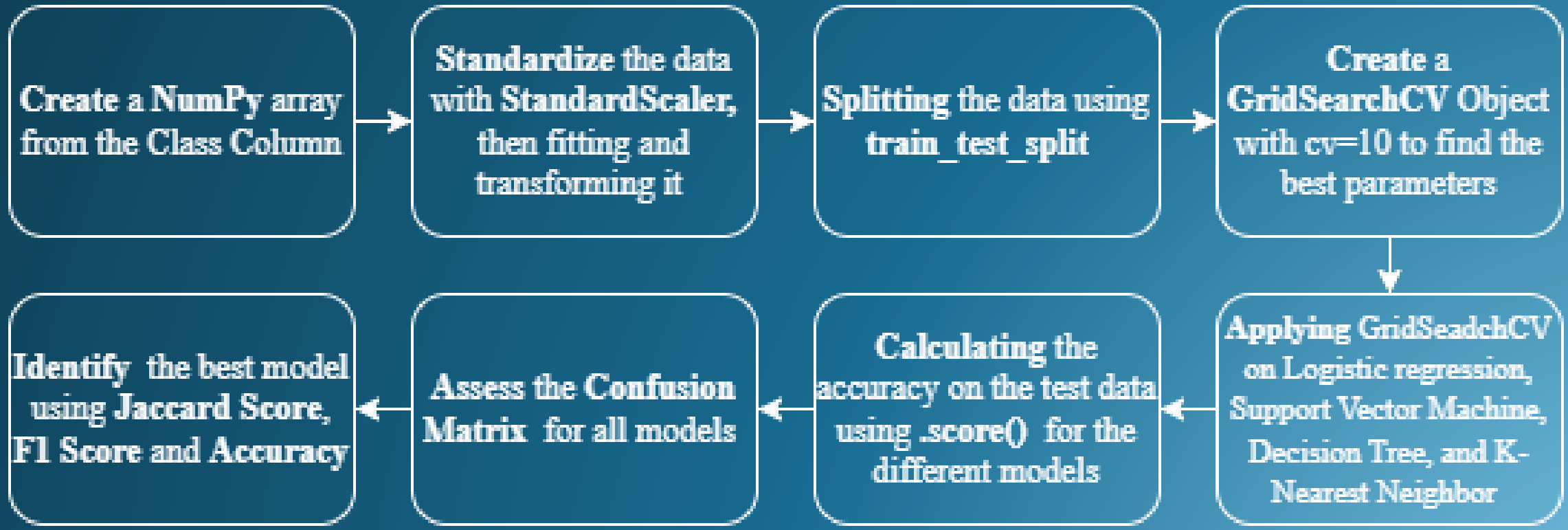
Github Link: Interactive Map - Folium

Dashboard with Plotly Dash

- **Drowpdown** Launch Sites list:
  - The user can select a specific launch site
- **Pie Chart** showing successful launches:
  - Allows the user to see the percentage of suceesses and failures of all or specific launches
- **Slider** of Payload Mass range:
  - Allows the user to choose a specific Payload Mass range
- **Scatter Chart** of the Payload Mass vs Success Rate by Booster Version
  - Allows the user to see the **correlation** between Payload Mass and Launch Success

Github Link: Dashboard with Plotly Dash

Predictive Analysis



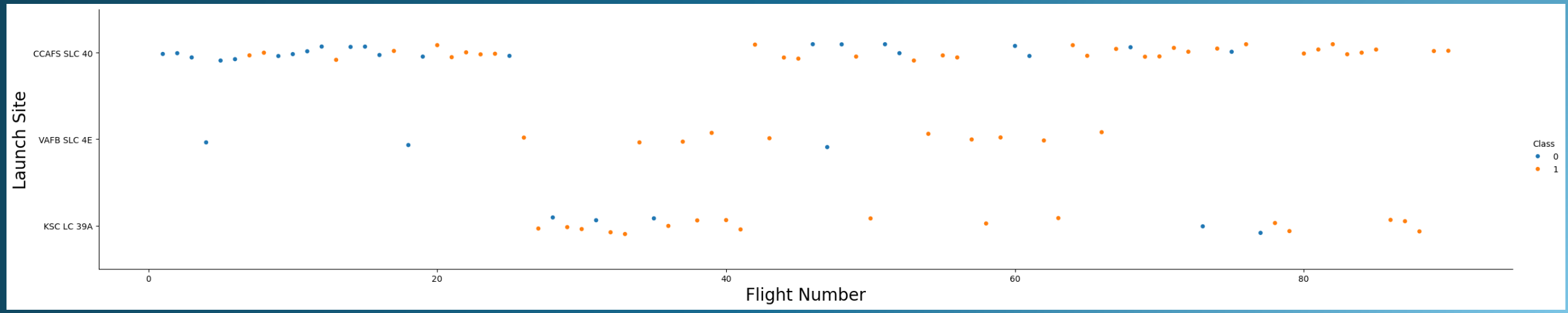
Github Link: Predictive Analysis

Results

Exploratory Data Analysis

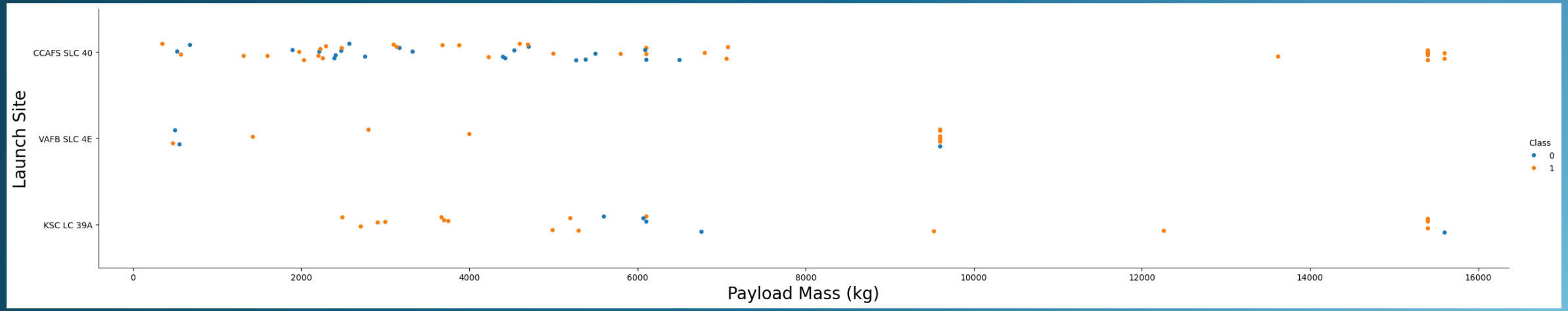
- The Falcon 9 rockets were launched on 3 Sites:
  - Cape Canaveral Air Force Station - Space Launch Complex (CCAFS SLC-40)
  - Vandenberg Air Force Base - Space Launch Complex (VAFB SLC 4E)
  - Kennedy Space Center - Launch Complex (KSC LC 39A)

- Around Half of the launches were from the CCAFS SLC 40 launch site
- Early flights have a higher fail rate (**Blue**)
- Later flights have a higher success rate (**Orange**)
- Based on the analysis, it's safe to infer that new launches have a higher chance of success



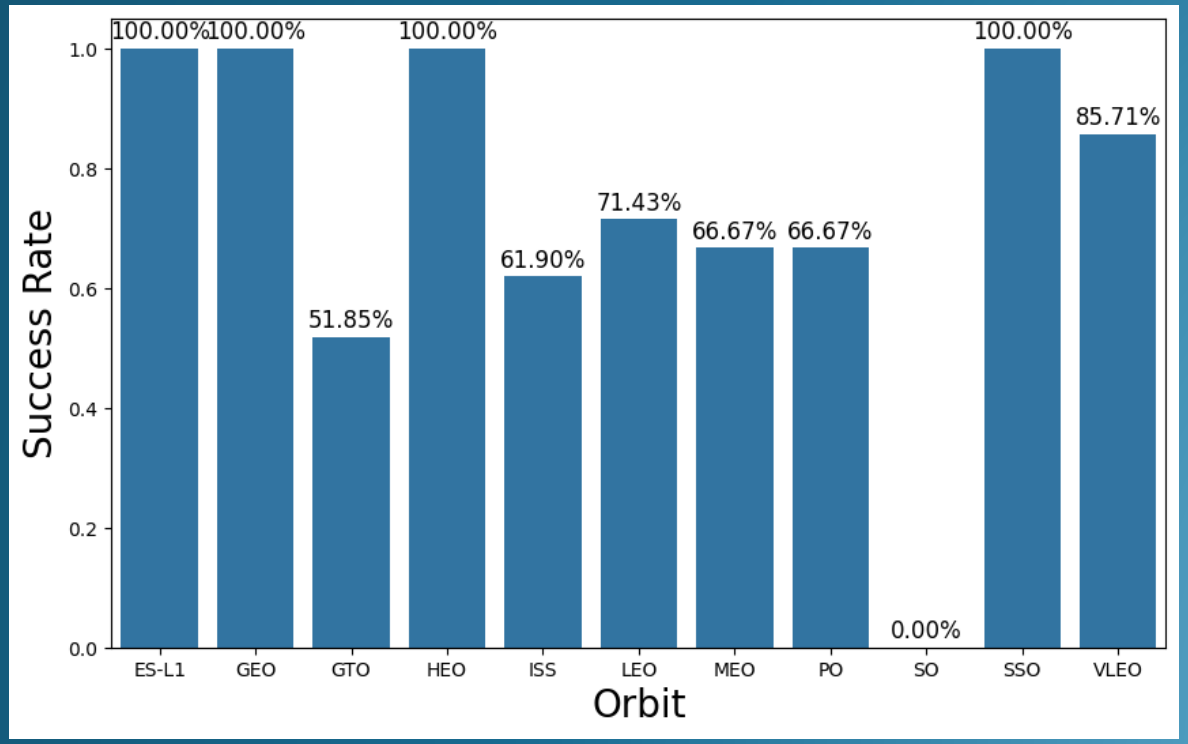
Exploratory Data Analysis

- We can say that, the higher the Payload Mass(kg), the higher the success rate
- Payload Masses of 7.000 kg or higher have a higher success rate
- On KSC LC 39, all lanches lower than 5.500 kg were successful
- VAFB SLC 4E hadn’t launched any rockets greater than 10.000 kg



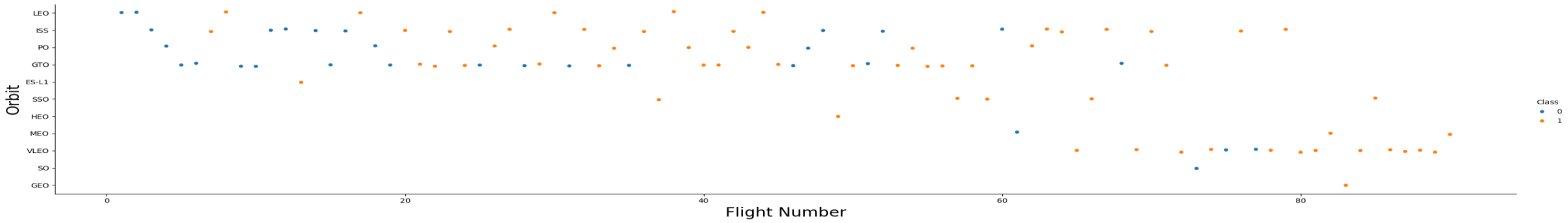
Exploratory Data Analysis

- Relationship between each Orbit type:
  - ES-L1, GEO, HEO, and SSO had a **100% Success Rate**
  - GTO, ISS, LEO, MEO, and VLEO had **between 51.85 to 85.71% Success Rate**
  - SO had a **0% Success Rate**

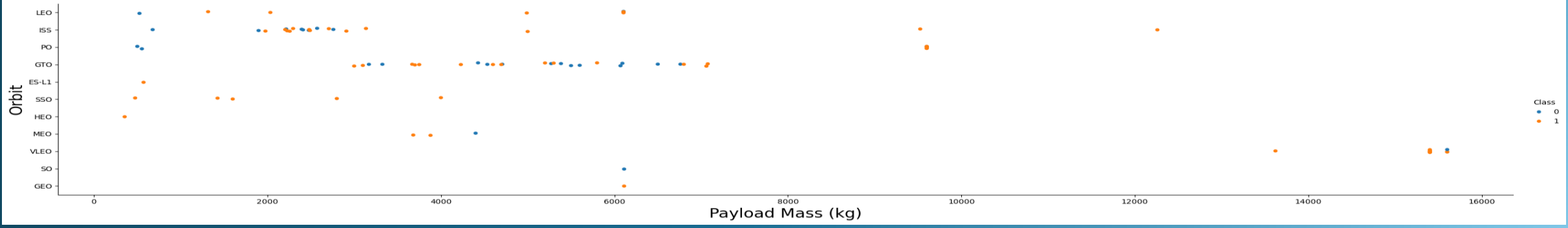


Exploratory Data Analysis

- There is the trend of success increase with each orbit type as the number of flights increase
- However, the GTO orbit doesn't follow this trend

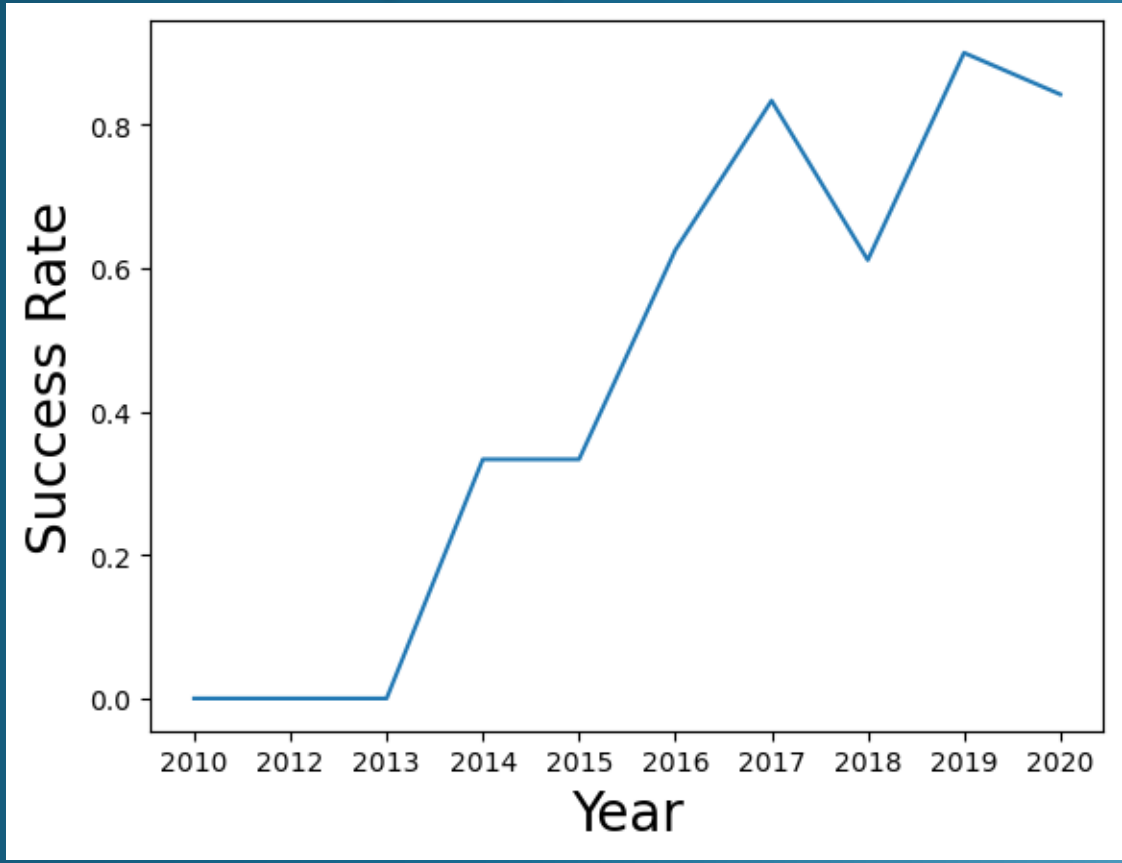


- Heavier Payload Mass had more success with LEO, ISS, and PO Orbits
- However, GTO had mixed results



Exploratory Data Analysis

- The yearly success was increased from **2013 to 2017** and from **2018 to 2019**
- However, it had some decrease in the years **2013** and **2020**
- Overall, we can say that the Success Rate increased throughout the years



Exploratory Data Analysis - SQL

- There are four launching sites in the dataset, although Falcon 9 was launched from only 3 of them

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

- Records with the first 5 Launch Sites records started with “CCA”:

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

Exploratory Data Analysis - SQL

- The **Total Payload Mass** carried by boosters launched by NASA(CRS) was **48.213 kg**

Total_Payload_Mass
48213

- The **Average Payload Mass** carried by Falcon 9 v1.1 was around **2.534,66 kg**

Average_Payload_Mass
2534.6666666666665

- The **First Successful Lnading** was on **December 22, 2015**

First_Successful_Landing
2015-12-22

- The name of the boosters that landed on a drone ship that has a payload between 4.000 and 6.000 kg was

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

Exploratory Data Analysis - SQL

- Total number of successful and failed mission outcomes

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

- Boosters that carried the maximum payload

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

Exploratory Data Analysis - SQL

- Boosters that Failed landing on drone ship in the months of 2015

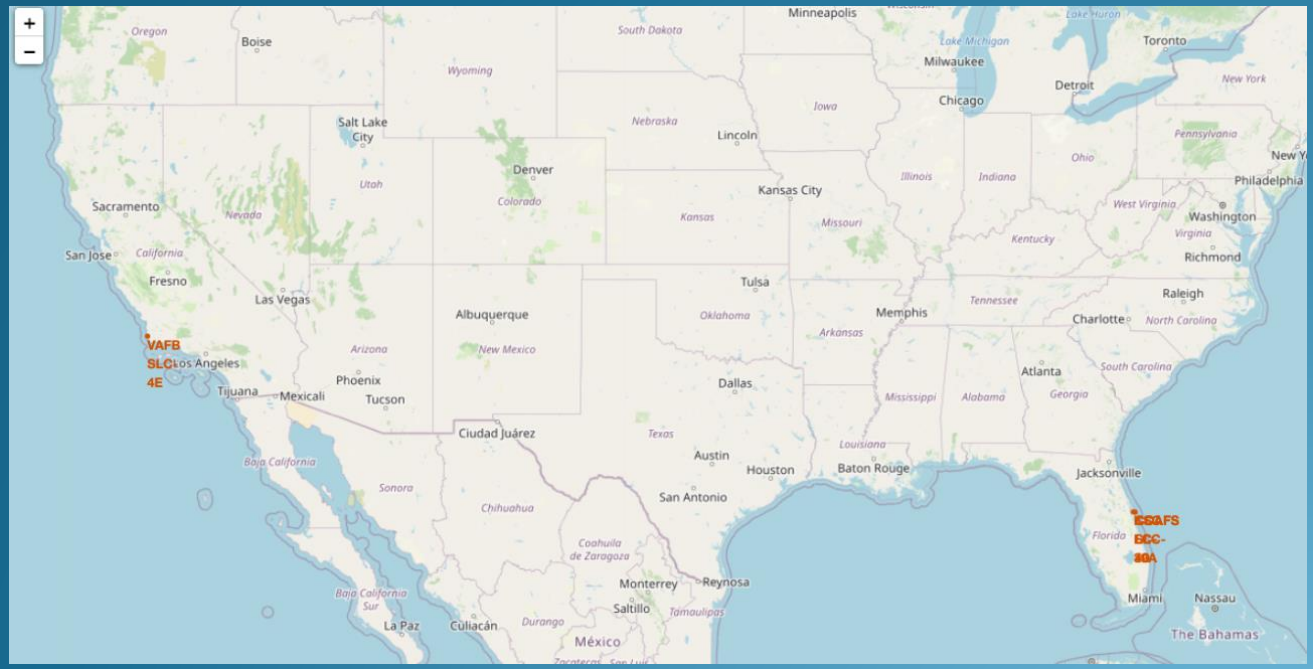
Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Count of landing outcomes from 2010-06-04 to 2017-03-20 in descending order

Landing_Outcome	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

Interactive Map with Folium

- All launch sites are close to the coastlines, as launches towards the ocean to minimize the risk of debris dropping or exploding near residential areas.

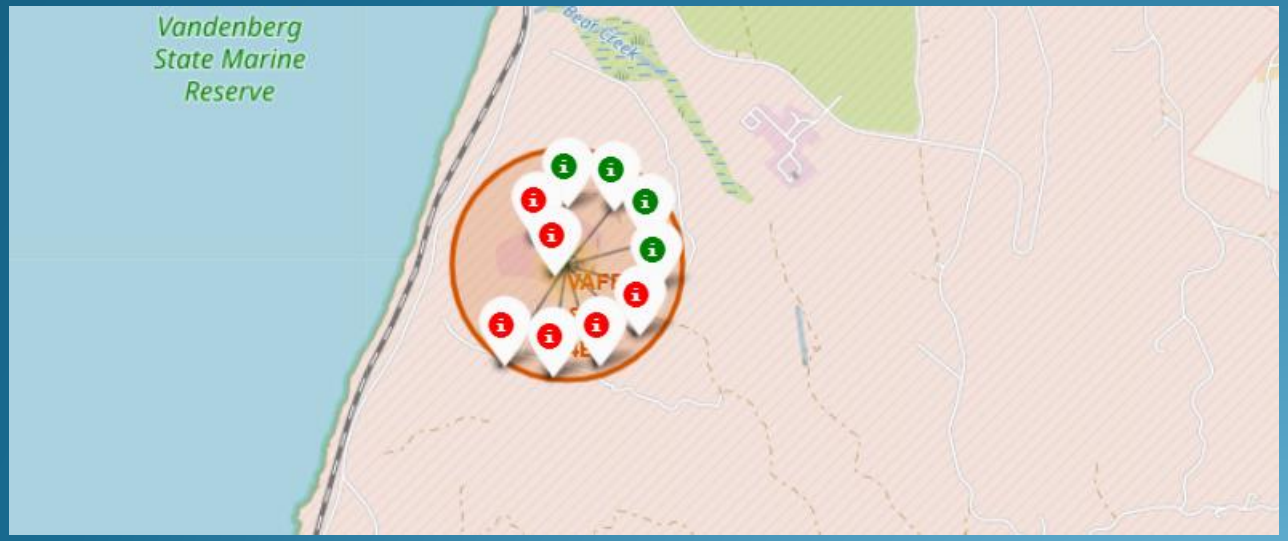


- Many launch sites are situated close to the Equator.

This is because the Earth's rotation is fastest at the Equator, with surface speeds reaching 1.670 km/hour. When a spacecraft is launched from this region, it ascends into space while retaining the horizontal velocity it had due to Earth's rotation. According to Newton's first law of motion, this inertia helps the spacecraft achieve and maintain the necessary speed to stay in orbit.

Interactive Map with Folium

- Each launch was clustered with the number of launches



- When the number is clicked, we can see that the Failure are marked in **Red** and Success is marked in **Green**

- By Looking at the number of launches, we can identify the success rate

Interactive Map with Folium

- **KSC LC-9:**

Distance from the Coastline: 7.42km
Distance from the Highway: 0.63km
Distance from the Railway: 0.66km
Distance from the Nearest City: 16.24km

- **CCAFS SLC-40:**

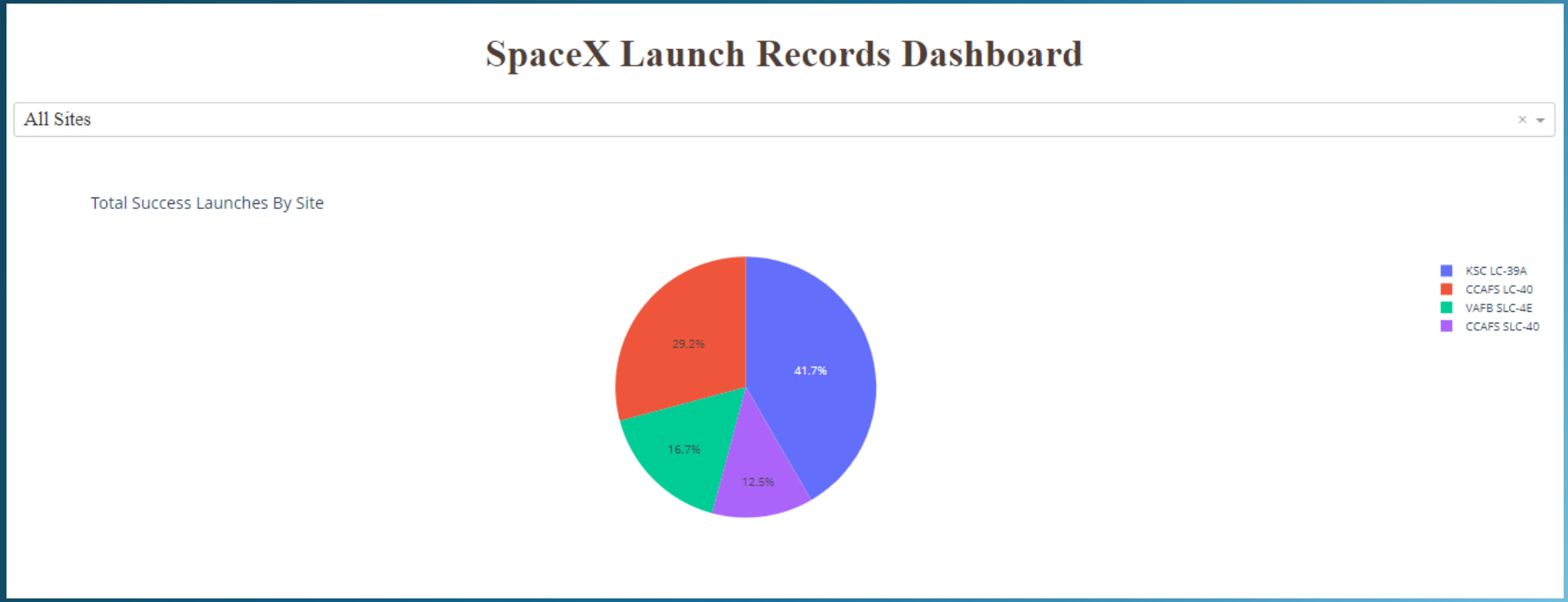
Distance from the Coastline: 0.94km
Distance from the Highway: 0.66km
Distance from the Railway: 0.01km
Distance from the Nearest City: 23.17km

- **VAFB SLC-4E:**

Distance from the Coastline: 1.34km
Distance from the Railway: 0.02km
Distance from the Nearest City: 14.02km

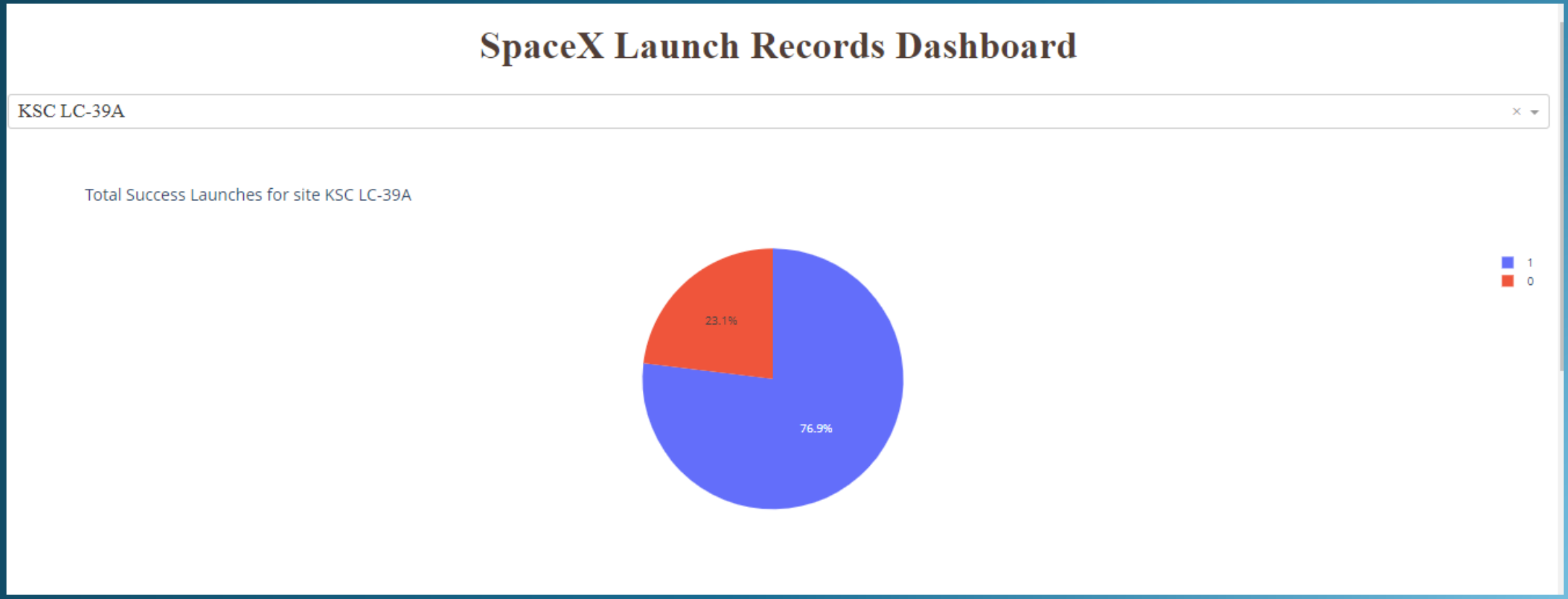
Dashboard with Plotly Dash

- By Looking at the chart, we can see that **KSC LC-39** has the most successful launches (**41.7%**)



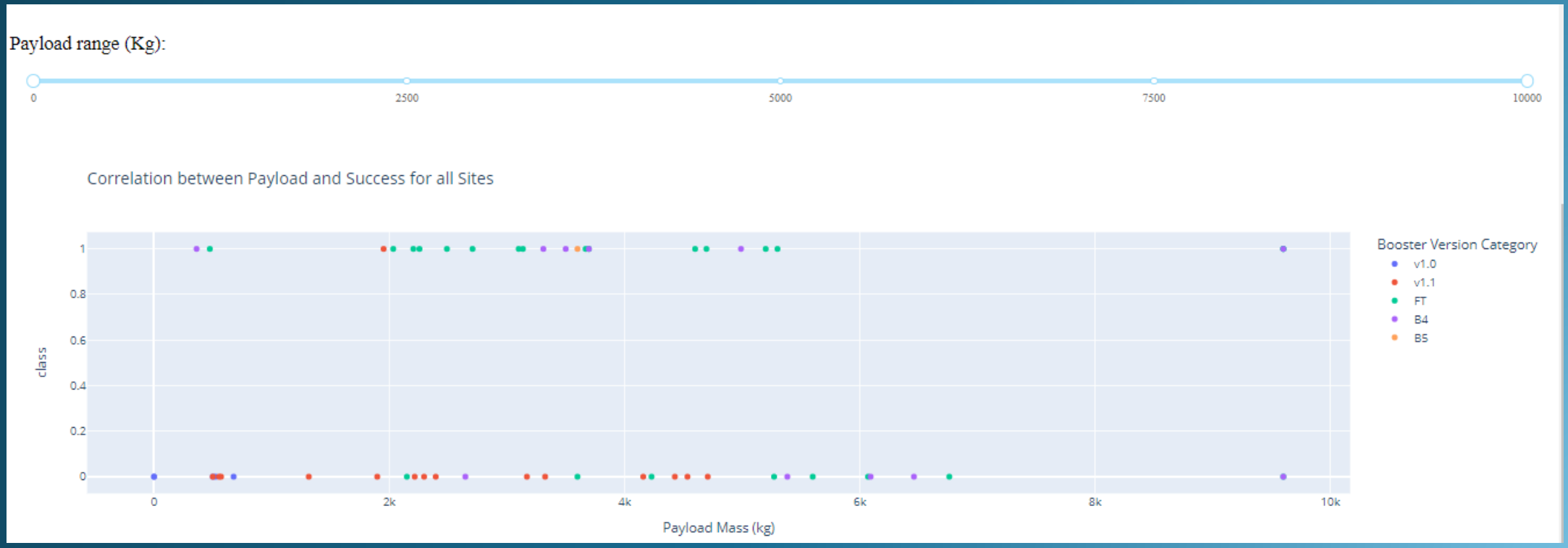
Dashboard with Plotly Dash

- **KSC LC-39A** has the most successful launches among all launching sites (**76.9%**)
- There are 10 successful launches and 3 failed launches.



Dashboard with Plotly Dash

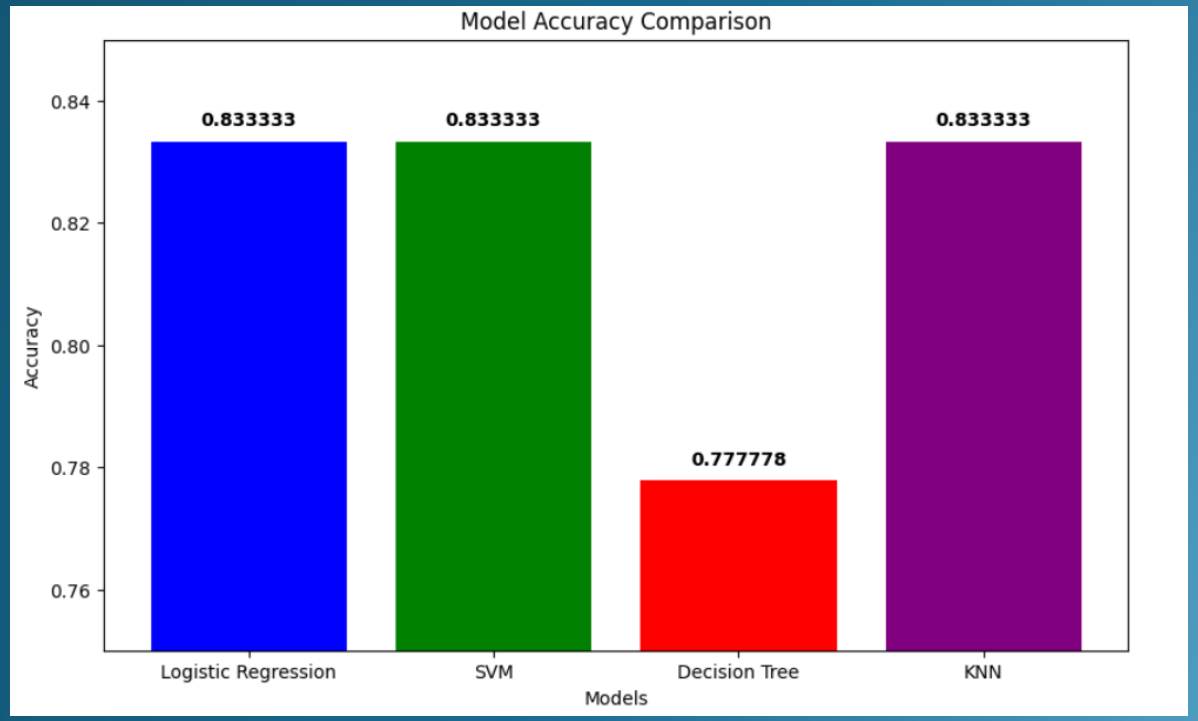
- Payload masses between 2.000 and 5.000 kg have the highest success rates among all (0 is **Failure** and 1 is **Success**)



Predictive Analysis - Classification

- Logistic Regression, SVM, and KNN share the highest classification accuracy at **83.33%**. The Decision Tree model has the lowest accuracy at **77.78%**

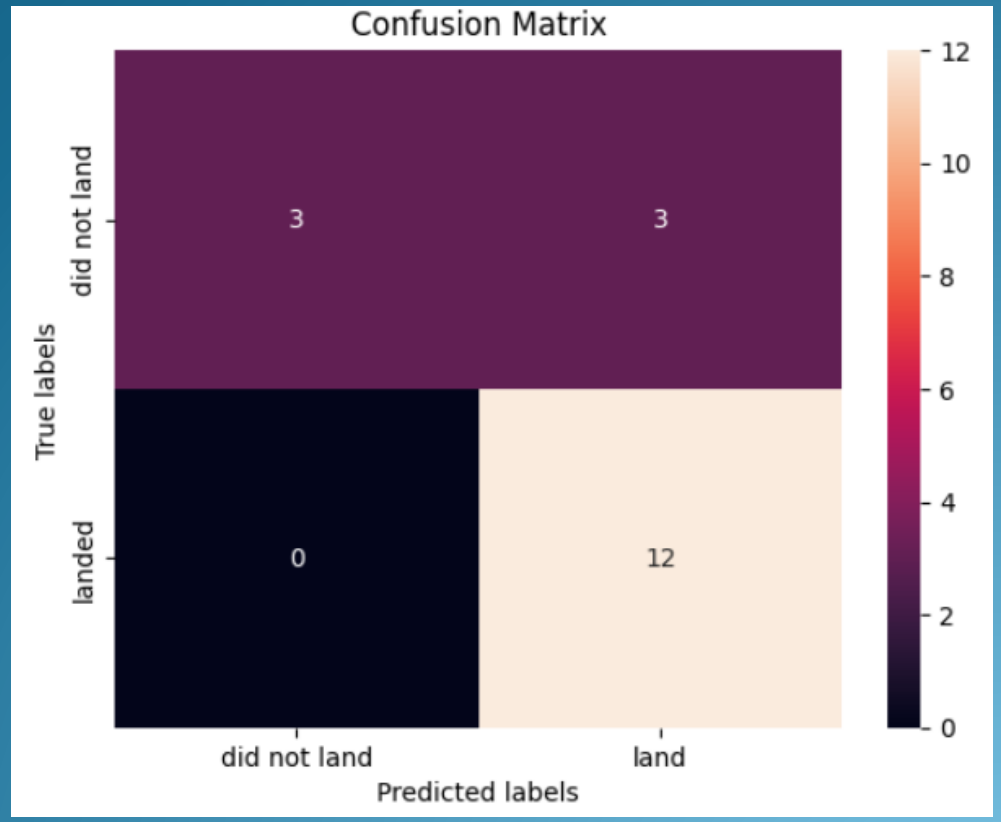
	Algorithm	Jaccard Score	F1 Score	Accuracy
0	Logistic Regression	0.800000	0.888889	0.833333
1	SVM	0.800000	0.888889	0.833333
2	Decision Tree	0.692308	0.818182	0.777778
3	KNN	0.800000	0.888889	0.833333



Predictive Analysis - Confusion Matrix

- While Looking at the Confusion Matrix, we can see that the major prroblem si the false positives, also know as type 1 error.

- Outputs:
  - 12 True Positives
  - 3 True Negatives
  - **3 False Positives**
  - 3 False Negatives



Conclusion

- All launches are close to **Coastilnes** to prevent accidents. Also, mos of them near the equator line
- **Lanch Success** increased over time.
- **KSC LC-39A** has the highest success rates among all launching sites and has 100% success rate for launches under 5.500 kg.
- The higher the **payload mass**, the higher the success rate.
- All models have **similar performance**, except Decision Trees.
- The small dataset might contribute to the similar model performance and false negatives. A larger dataset might help mitigate these issues.
- **ES-L1**, **GEO**, **HEO**, and **SSO** have **100%** success rates.

Appendix

Data Sources

•SpaceX API Data:

- URL: **SpaceX API**
- Description: Data collected from the SpaceX REST API, including launch details, payload information, and landing outcomes.

•Wikipedia Data:

- URL: **Wikipedia Falcon 9**
- Description: Data scraped from the Falcon 9 Wikipedia page, including historical launch data.

- I would like to thank **IBM**, **Coursera** and all the Instructors for this series of courses.
- Special thanks to my fiance who supported me throughout this journey