IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pieter van Hooft
10-04-23

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API/Webscraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Machine Learning
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Dashboards
  - Predictive Analytics results

Introduction

- Project background and context
  - The main objective is to determine the ability of the company Space Y to compete with SpaceX
- Problems you want to find answers
  - Find a way to accurately predict the cost of launches by predicting the success rate
  - What is the best place to launch rockets from

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical 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

- Describe how data sets were collected.
    - Datasets were collected using the SpaceX API and webscraping from Wikipedia

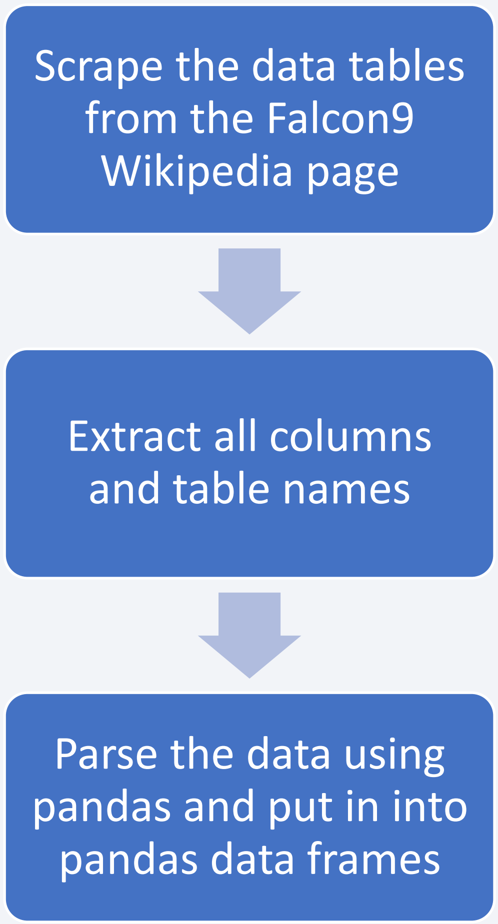
Data Collection – SpaceX API

- Made use of the SpaceX public API where the data was obtained and then used for analysis

- Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Data%20collection%20API.ipynb

Data Collection - Scraping

- BeautifulSoup was used to webscrape the Falcon 9 launch records from Wikipedia
- The data was then parsed and put into a pandas data frame
- Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Data%20collection%20webscaping.ipynb



Data Wrangling

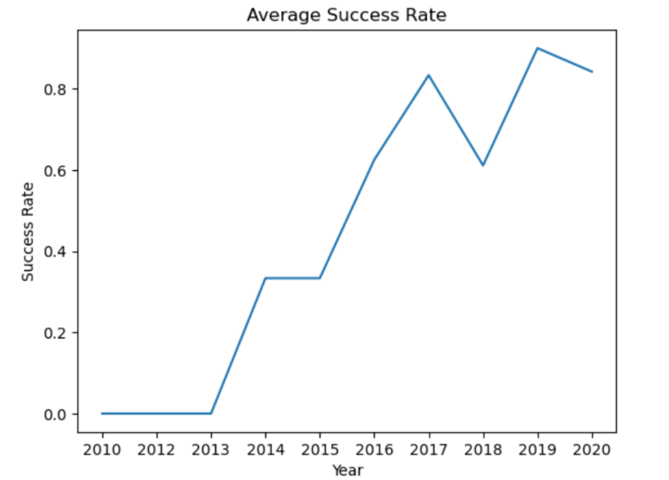
- Exploratory data analysis was conducted to determine the training labels.
- The count of launches at each site and the number and frequency of orbits were computed.
- The landing outcome label was derived from the outcome column, and the results were exported to a CSV file.

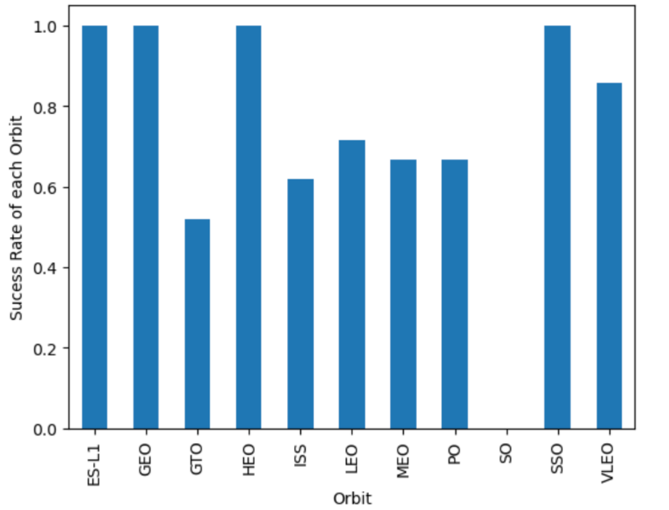
Exploratory data analysis → Launch and orbit analysis → Landing label exported to csv

Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The relationship between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, and the launch success yearly trend were explored through data visualization.
- Data visualization was used to analyze the relationship between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, and the launch success yearly trend.
- Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Exploratory%20Analysis%20using%20Pandas%20and%20Matplotlib.ipynb





EDA with SQL

- The SpaceX dataset was loaded into a PostgreSQL database within the Jupyter notebook.
- SQL-based EDA was utilized to gain insights from the data, such as querying:
  - Unique launch site names in the space mission.
  - Total payload mass carried by NASA-launched boosters (CRS).
  - Average payload mass carried by F9 v1.1 booster version.
  - Total number of successful and failed mission outcomes.
  - Failed landing outcomes in drone ship, with booster version and launch site names
- Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/EDA.ipynb

Build an Interactive Map with Folium

- Launch sites were marked on the folium map
- Map objects such as markers, circles, and lines were added to indicate launch success or failure for each site
- A feature was assigned to class 0 and 1 to represent launch outcomes, with 0 indicating failure and 1 indicating success
- Using color-labeled marker clusters, launch sites with relatively high success rates were identified
- Distances between launch sites and their proximities were calculated
- Questions were answered, such as whether launch sites were located near railways, highways, and coastlines, and whether they maintained a certain distance from cities
- Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

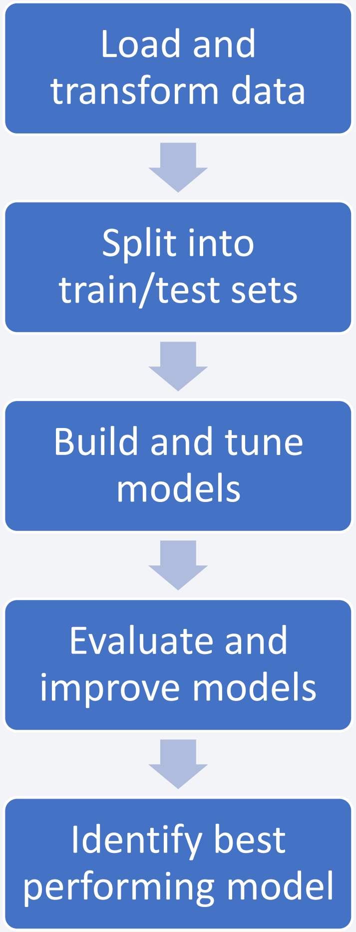
Build a Dashboard with Plotly Dash

- Creation of an interactive dashboard using Plotly Dash
- Inclusion of pie charts displaying the aggregate launches for specific sites
- Generation of scatter graphs to visualize the correlation between Outcome and Payload Mass (Kg) for each booster version
- Use of Plotly Dash to develop the interactive features of the dashboard

- Python file: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/Plotly_dash_app.py

Predictive Analysis (Classification)

- Loaded data using numpy and pandas
- Transformed the data
- Split data into training and testing sets
- Built multiple machine learning models
- Tuned hyperparameters using GridSearchCV
- Used accuracy as the metric for model evaluation
- Improved the model through feature engineering and algorithm tuning
- Identified the best performing classification model



Jupyter Notebook: https://github.com/PetervHooft/Data-Scientist-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

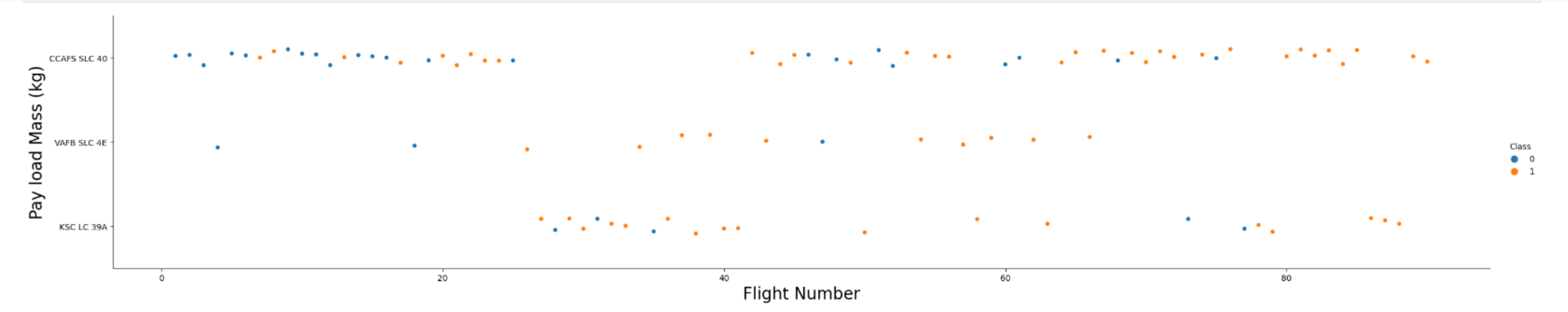
Results

- Exploratory data analysis results
  - The first successful landing outcome happened in 2015, five years after the first launch.
  - Almost 100% of the mission outcomes were successful.
  - Many Falcon 9 booster versions were successful at landing on drone ships, having payload above average.
  - Space X uses 4 different launch sites.
  - The average payload of F9 v2.2 booster is 2,928kg
- Interactive analytics demo in screenshots
  - Interactive analytics can identify the safety and infrastructure of launch sites, such as those located near the sea.
  - Launch sites on the east coast are the most frequently used.
  - The safety and infrastructure of launch sites are crucial factors in the selection process.
- Predictive analysis results
  - According to predictive analysis, the decision tree classifier model is the most effective in predicting successful landings, with an accuracy rate of over 89%.

Section 2

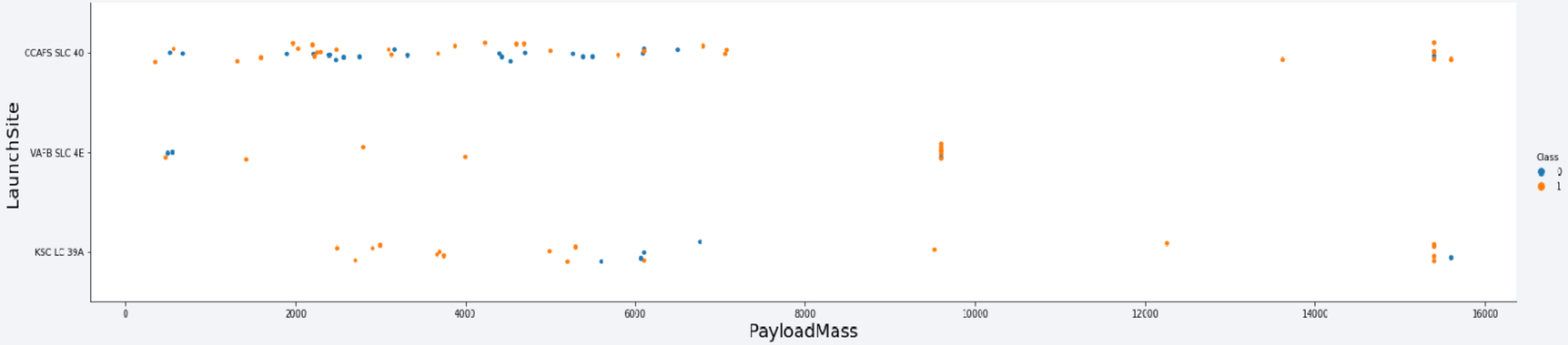
Insights drawn from EDA

Flight Number vs. Launch Site



- The plot indicates that there may be some relationship between flight number and launch site location, with flights having similar launch site coordinates appearing to be clustered together

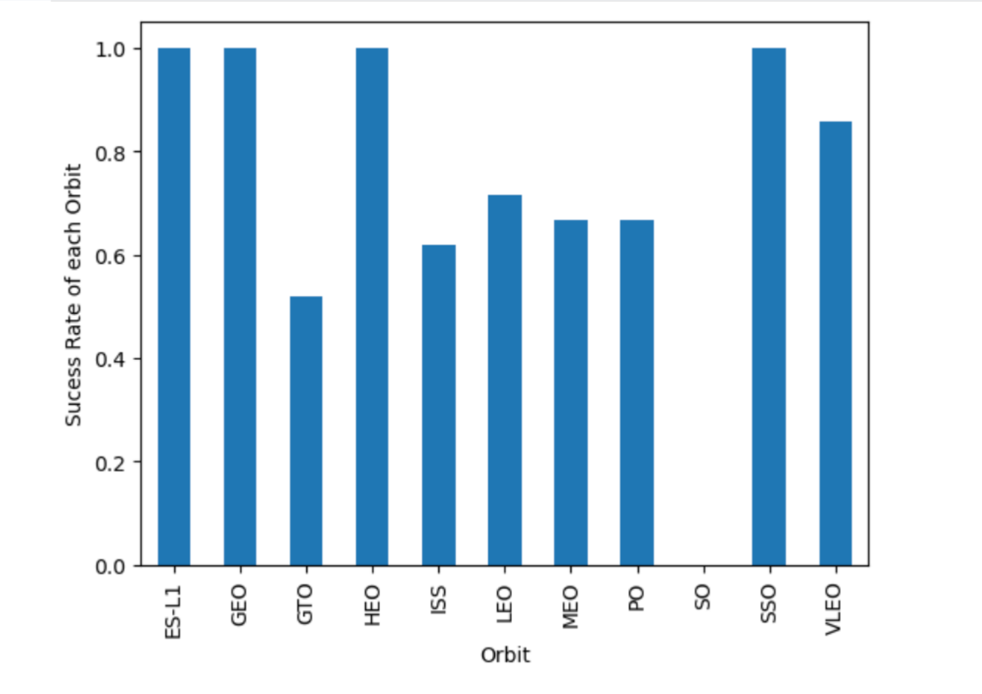
Payload vs. Launch Site



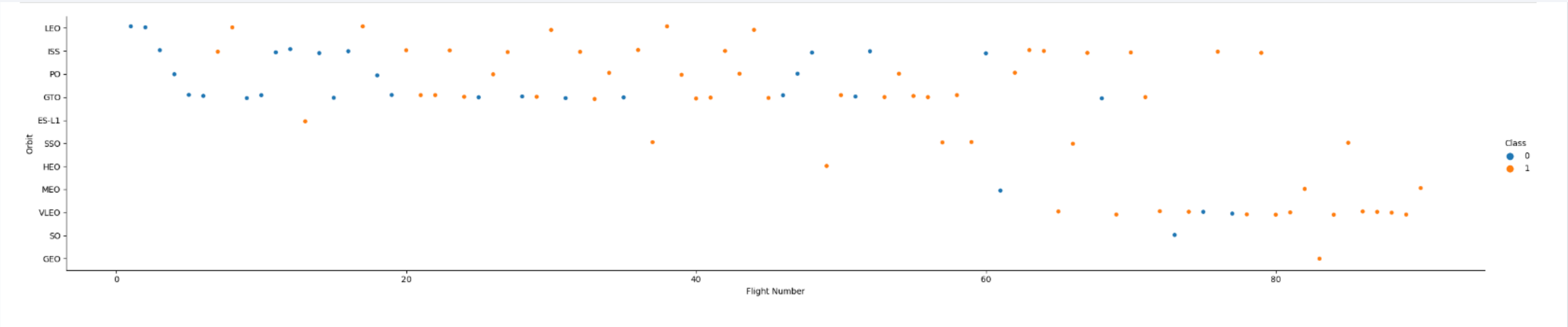
- Launch site location plays a significant role in determining the maximum payload weight that can be launched on a Falcon 9 rocket.
- The latitude and longitude of a launch site should be carefully considered when planning a mission with high payload requirements.

Success Rate vs. Orbit Type

- The success rate of Falcon 9 orbital missions has been steadily increasing over the past decade, indicating ongoing improvements in the rocket's design and launch processes
- Although there is some variation in success rates between launch sites, overall, the Falcon 9 has demonstrated a high level of reliability across all sites
- The most recent years in the chart show a consistent success rate of over 90%, suggesting that the Falcon 9 rocket has reached a high level of maturity and reliability in its design and operations

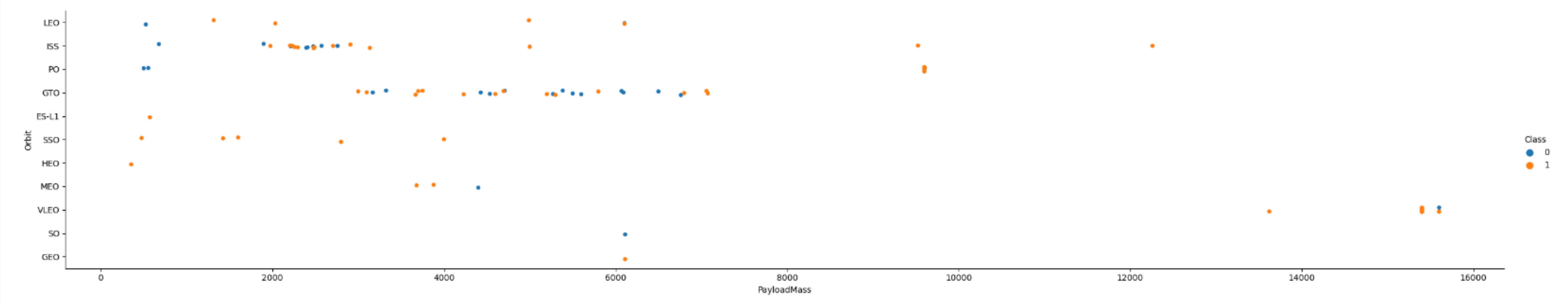


Flight Number vs. Orbit Type



- Low Earth Orbit missions are the most popular type of Falcon 9 launch and have increased in frequency over time, suggesting a growing demand for satellite launches and other space missions in this orbit
- Geostationary Transfer Orbit launches have remained relatively consistent in number and represent a smaller proportion of Falcon 9 launches compared to LEO missions
- The number of Polar Orbit launches is relatively low but has recently shown an increase, potentially reflecting a growing demand for satellites and other missions that require this type of orbit

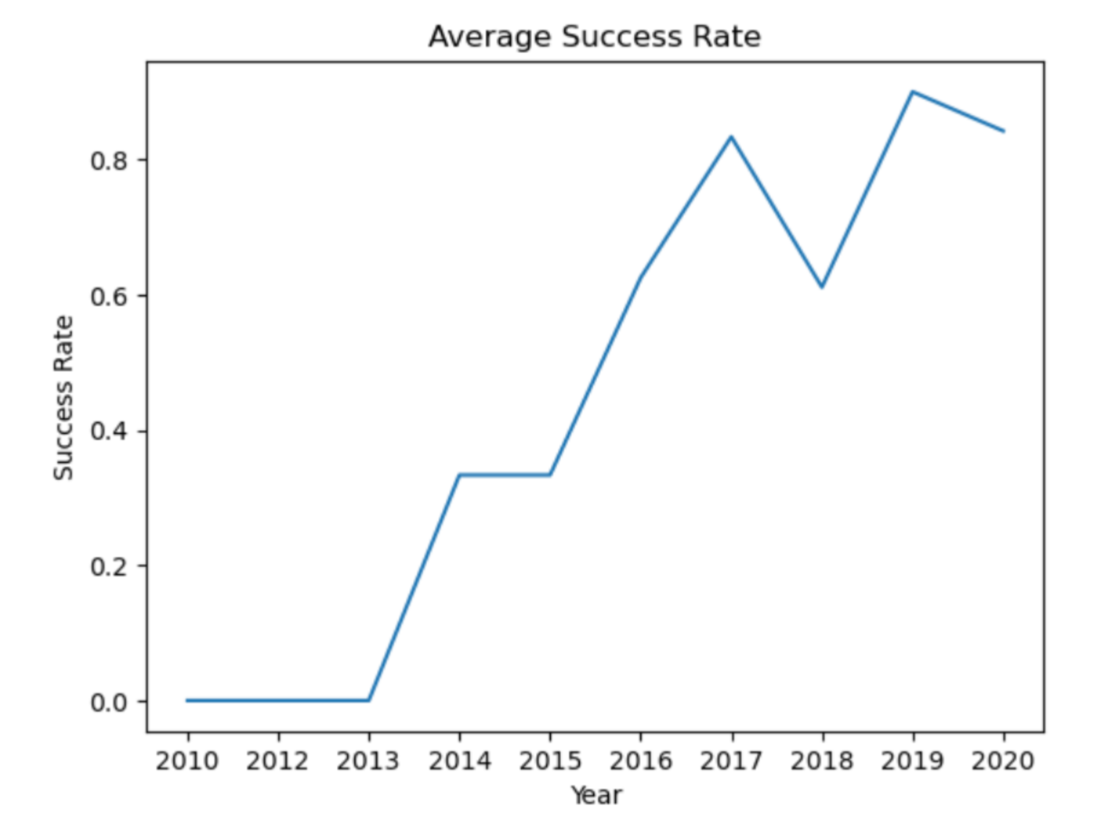
Payload vs. Orbit Type



- The Falcon 9 rocket has been used to launch a range of payloads, with the majority of launches falling into the Low Earth Orbit category, suggesting a high demand for this type of mission
- Geostationary Transfer Orbit launches have supported a significant number of heavy payloads, which may reflect the orbit's use for communication and weather satellites that require larger payloads

Launch Success Yearly Trend

- The success rate of Falcon 9 launches has steadily increased over the years, indicating improvements in SpaceX's rocket technology and launch processes
- The years 2015-2016 had lower success rates compared to the other years, which may reflect the challenges of testing new rocket technology and the learning curve for SpaceX during these early years of Falcon 9 launches



All Launch Site Names

- To display only unique launch sites from the SpaceX data, we utilized the DISTINCT keyword

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

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Using the query below, we displayed five records where launch sites begin with CCA

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

In [11]:

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Using the following query, we computed that the total payload carried by boosters from NASA was 45596

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

In [12]:

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- Using this calculation, we determined that the average payload mass carried by booster version F9 v1.1 was 2928.4.

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

In [13]:

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- Our observation indicated that the first successful landing outcome on a ground pad occurred on December 22nd, 2015

```
In [14]:  task_5 = '''
                  SELECT MIN(Date) AS FirstSuccessfull_landing_date
                  FROM SpaceX
                  WHERE LandingOutcome LIKE 'Success (ground pad)'
                  '''
          create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To filter for boosters that have landed successfully on drone ships, we employed the WHERE clause and applied the AND condition to determine successful landings with a payload mass greater than 4000 but less than 6000

```
In [15]: task_6 = '''
                 SELECT BoosterVersion
                 FROM SpaceX
                 WHERE LandingOutcome = 'Success (drone ship)'
                     AND PayloadMassKG > 4000
                     AND PayloadMassKG < 6000
                 '''
         create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- To filter for instances where MissionOutcome was either a success or a failure, we utilized the wildcard symbol '%' with the WHERE clause

List the total number of successful and failure mission outcomes

In [16]:

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''

print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

Out[16]:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- By utilizing a subquery within the WHERE clause and the MAX() function, we were able to identify the booster that carried the highest payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [17]:
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- To filter for failed landing outcomes in drone ship, booster versions, and launch site names for the year 2015, we used a combination of the WHERE clause, LIKE, AND, and BETWEEN conditions

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [18]:

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Selected Landing outcomes and the COUNT of landing outcomes from the data and filtered for landing outcomes between 2010-06-04 to 2010-03-20 using the WHERE clause
- Grouped the landing outcomes and ordered the grouped landing outcome in descending order using the GROUP BY and ORDER BY clauses

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

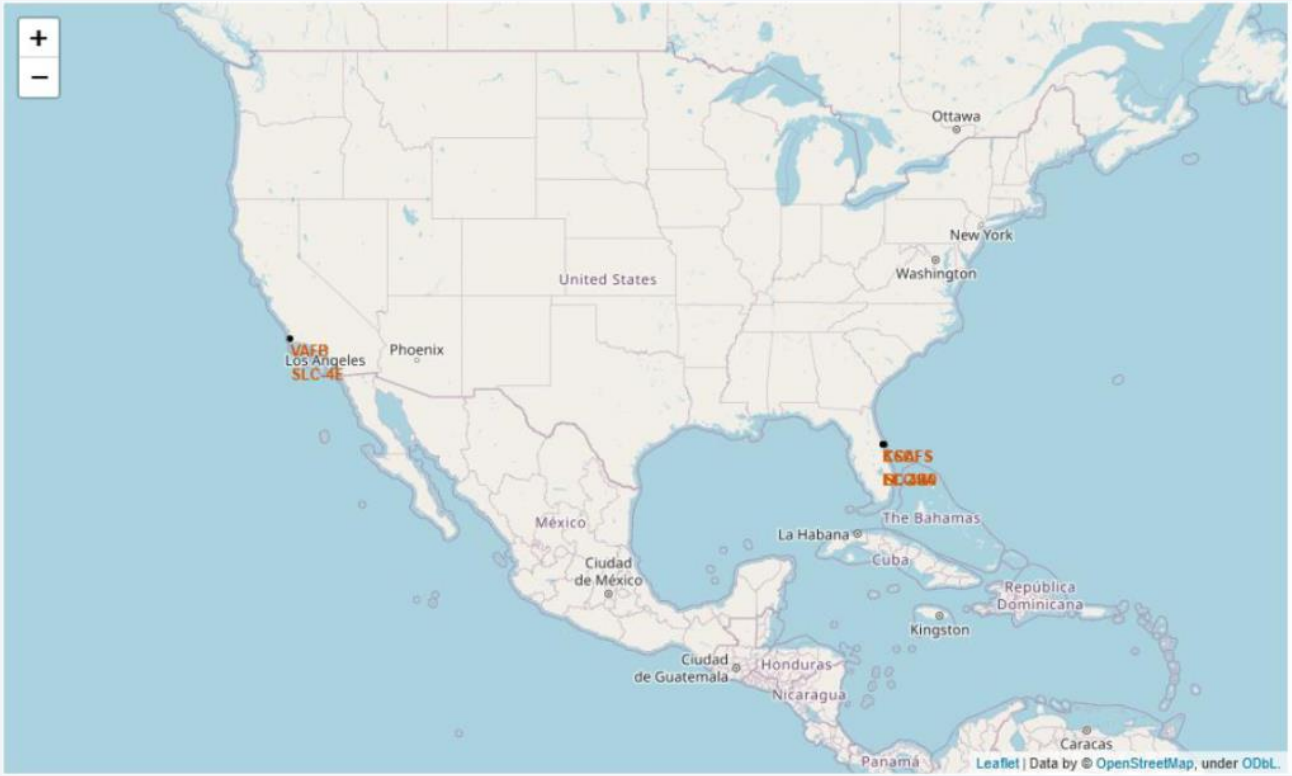
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

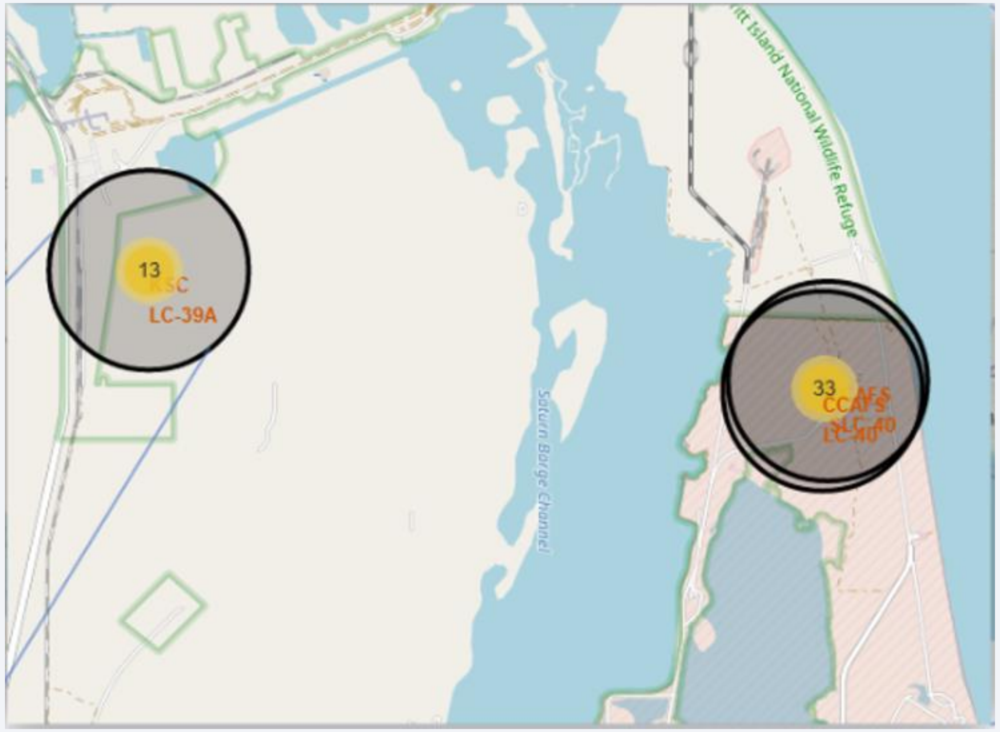
Global Map Markers of All Launch Sites

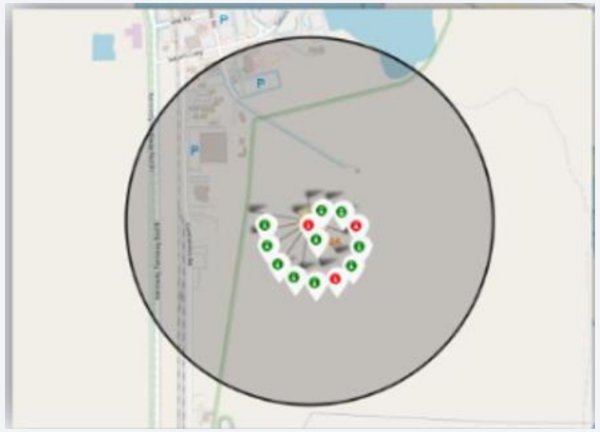
- The reason for zooming in on the US is that all launch sites are located within the country



Launch Outcomes at KSC LC-39A

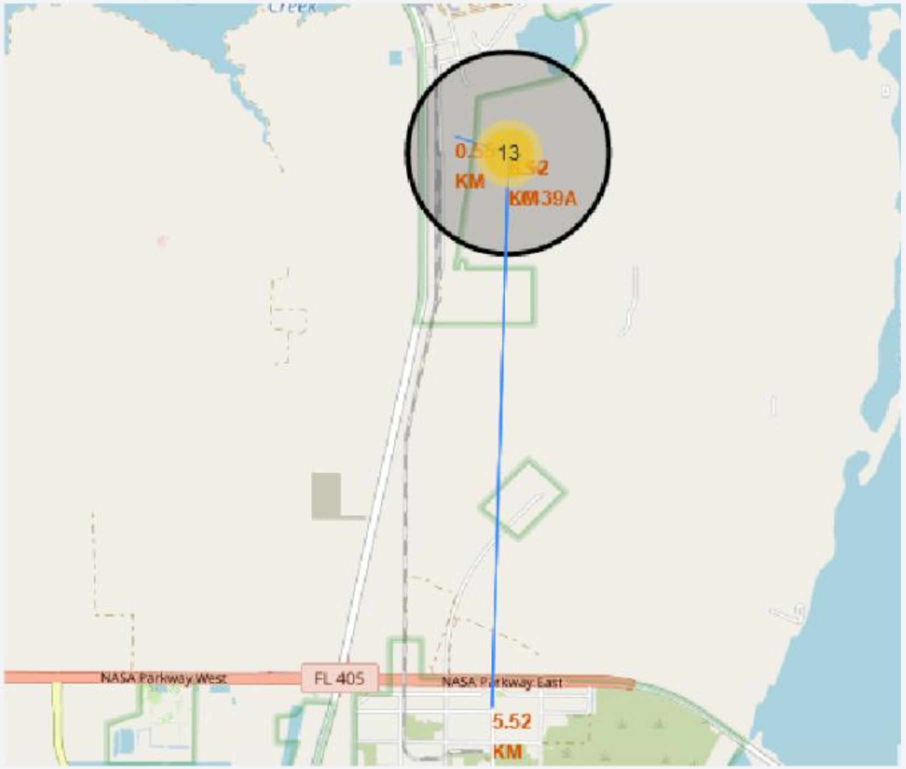
- This is an example of KSC LC-39A launch site outcomes, with green markers indicating successful launches and red markers indicating failures





Logistical analysis of KSC LC-39A launch site

- The KSC LC-39A launch site has favorable logistic characteristics due to its proximity to a railroad and road infrastructure and relatively remote location from populated areas

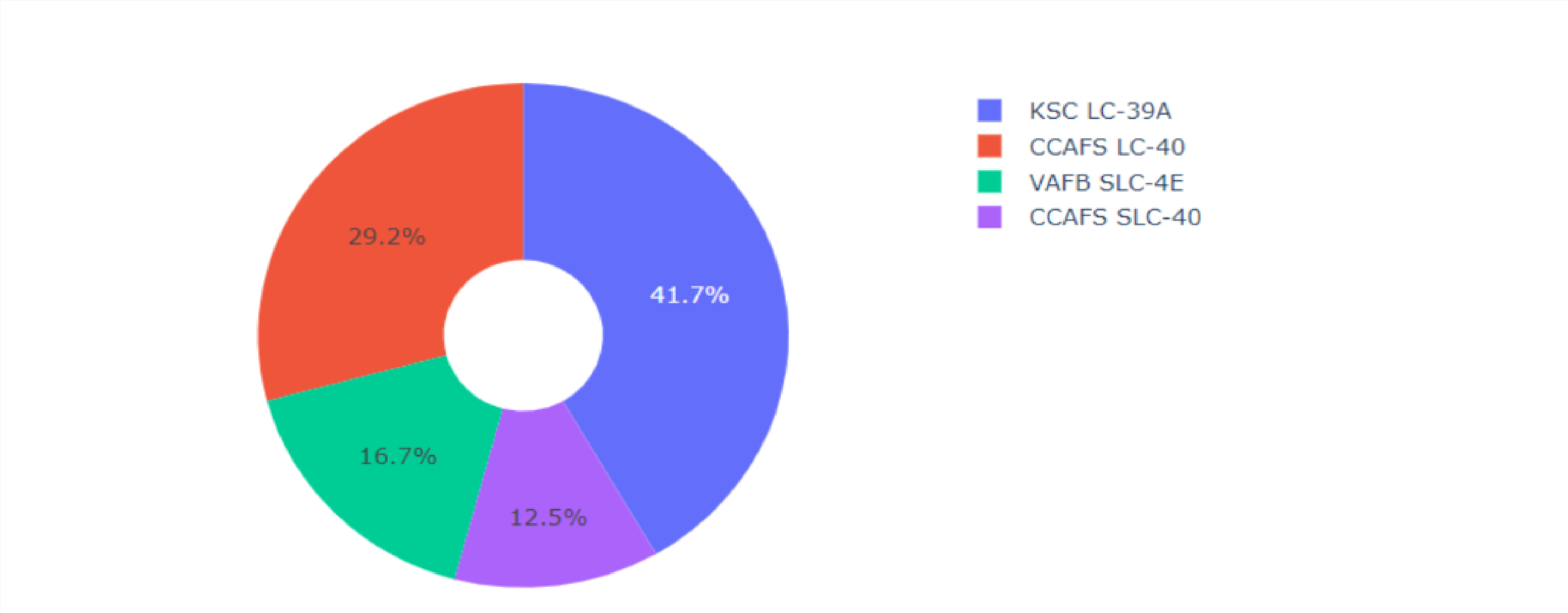


Section 4

Build a Dashboard with Plotly Dash

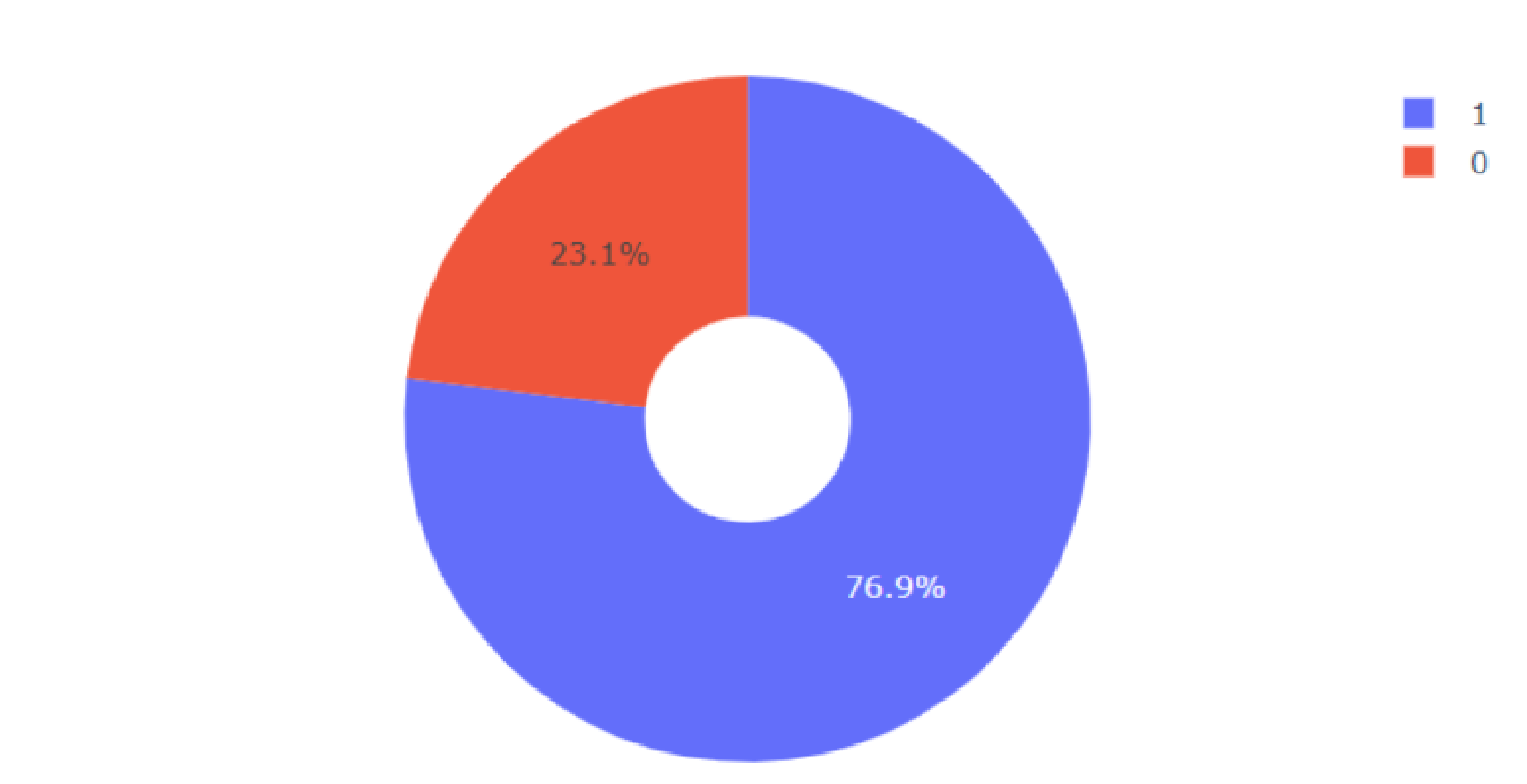
Successful Launches at Different Launch Sites

- KSC LC-39A accounted for the highest number of successful launches, contributing to 41.7% of the total successful launches



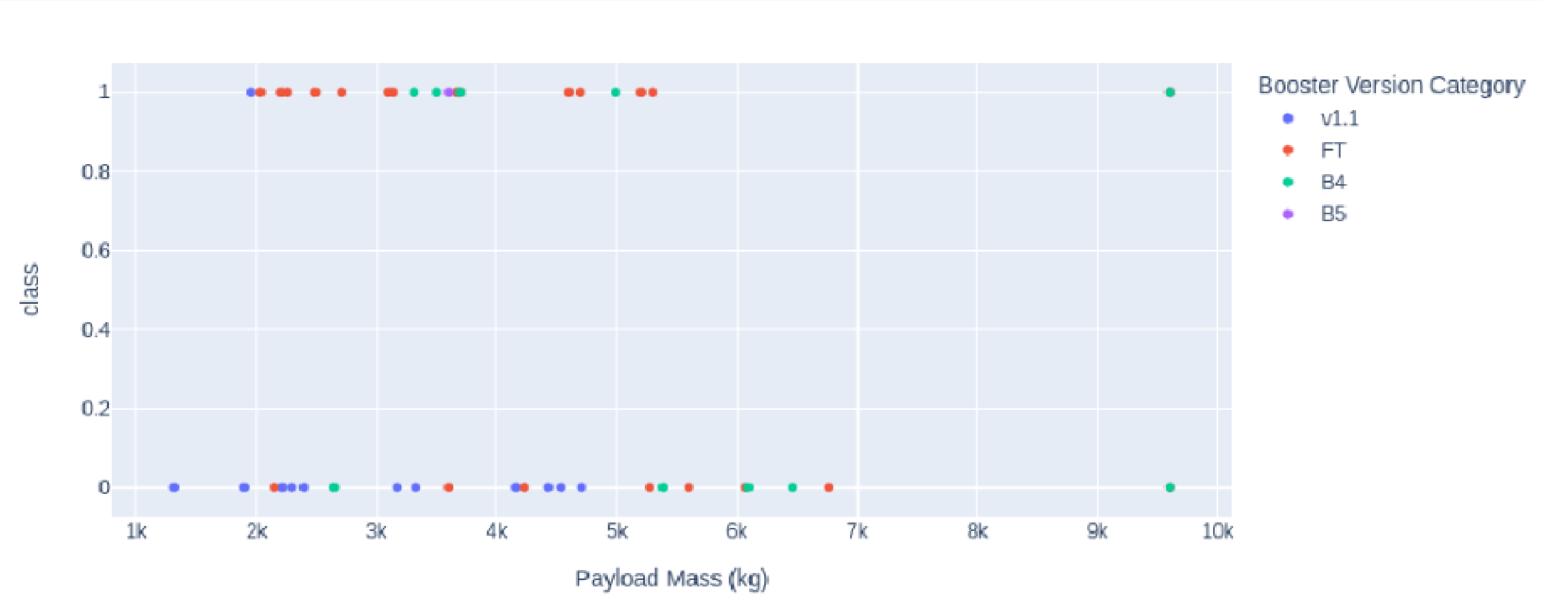
Success rate for KSC LC-39A

- 76.9% of launches were successful



Relationship between Payload Mass and Booster Versions

- There appears to be a positive correlation between payload mass and launch success, as the successful launches tend to have higher payload masses
- There are a few outlier points where successful launches had relatively low payload masses and failed launches had relatively high payload masses, suggesting that other factors beyond payload mass may also play a role in launch success

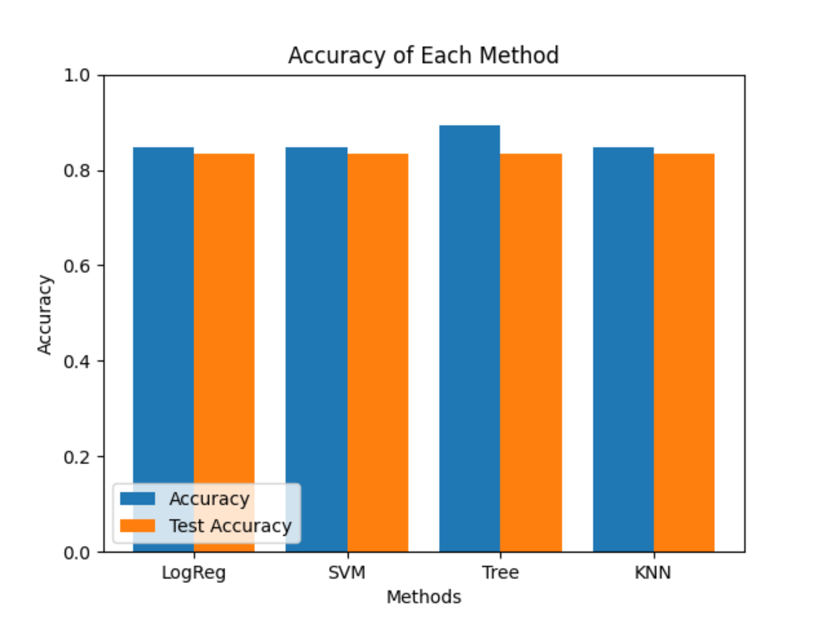


Section 5

Predictive Analysis (Classification)

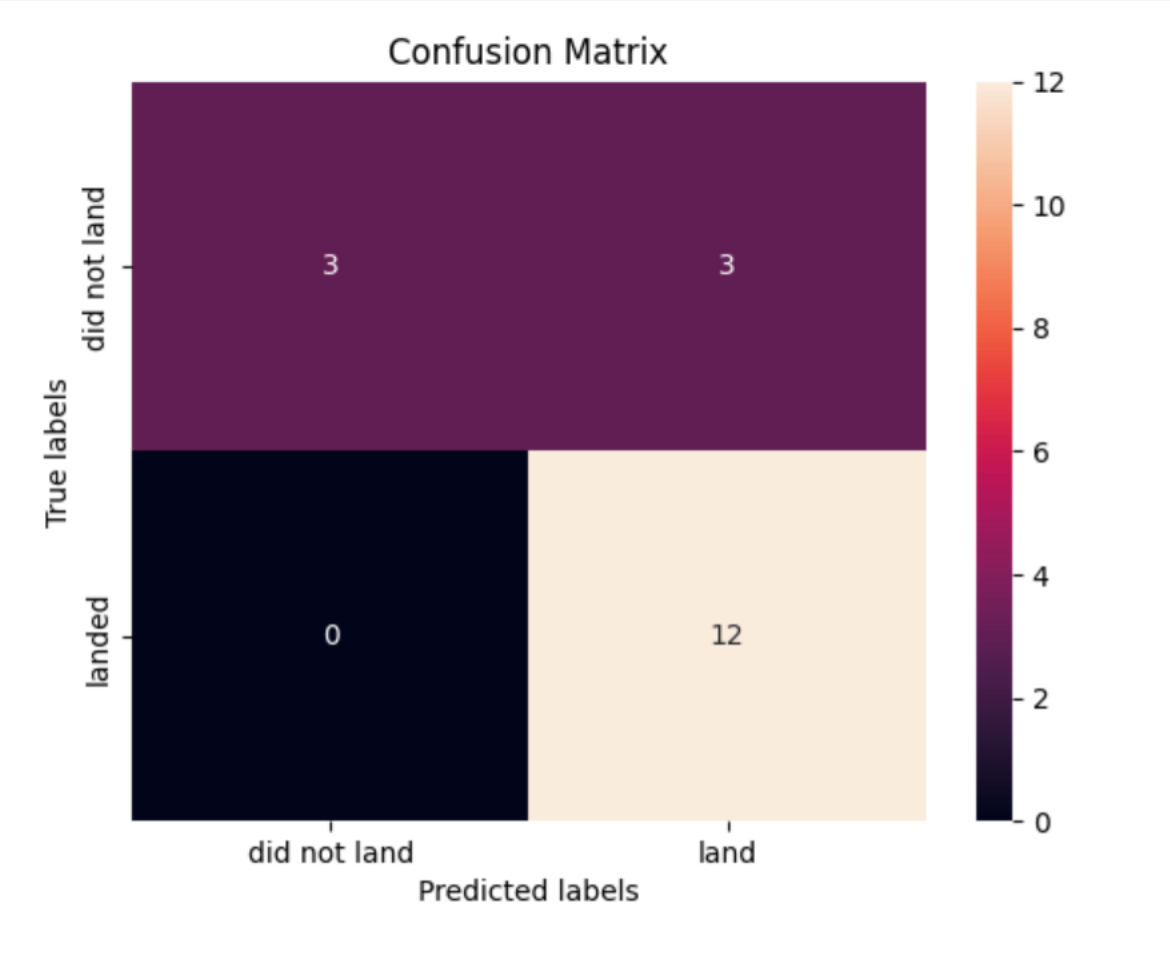
Classification Accuracy

- The Decision Tree is the most accurate out of all the predictive classification models



Confusion Matrix

- We can observe that the Decsision Tree model had a 100% accuracy in predicting successful landings, but only a 50/50 accuracy for unsuccessful landings



Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!