# Winning Space Race with Data Science

Bamidele Olawale Ajayi
24 June 2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - This Data Science project used various Python Libraries for analysis and visualization. Pandas, NumPy, Scikit Learn were used to analyze the data while Folium and Plotly Dash were used to visualize the data.
- Summary of all results
  - KSC LC 39A led the pack of successful launches with 41.7% while launches from CCAFS SLC-40 were the least successful with 12.5%
  - Booster Version FT category had the most successful launches overall regardless of payload mass while V1.1 was the least successful.

# Introduction

- Project background and context
  - The main objective is to analyze SpaceX's launch data to determine factors that influence the costs and success of each launch.
  - The project will leverage machine learning models to predict whether the first stage of a Falcon 9 launch will land successfully. This estimation is important to planning and budgeting.
- Problems I want to find answers
  - What are the specific factors that significantly influence the cost of launches for SpaceX, and how can these be quantified for Space Y?
  - Can we predict the success of Falcon 9 first stage landings based on previous launch data?
  - What variables most strongly influence the likelihood of the first stage landing successfully?

Section 1

# Methodology

# Data Collection – SpaceX API

- Called SpaceX REST API with requests .get() method.
- Turned the response object to a Pandas dataframe using the .json_normalize() method.
- Extracted important details such as Booster Version, Payload Mass etc.
- Create headers and construct a new dataframe with the extracted details.
- GitHub Here

Call API→Create DataFrame with response→ Extract Important Details →Add Column Headers→ Filter to remove Falcon 1 Launches → Replace missing payload mass with payload mean → Store in a CSV file.

# Data Collection - Scraping

- Scraped Wikipedia page with requests .get() method.
- Parsed response with BeautifulSoup.
- Extracted Column headers from tables with predefined functions.
- Created a dictionary and a dataframe from data extracted
- GitHub

Scraped the Wikipedia page for HTML tables → Parse the response with BeautifulSoup and extract Column names from table headers → Create a dictionary with the HTML tables using the table headers as column headers → Create a dataframe from the dictionary → Save in a CSV file.

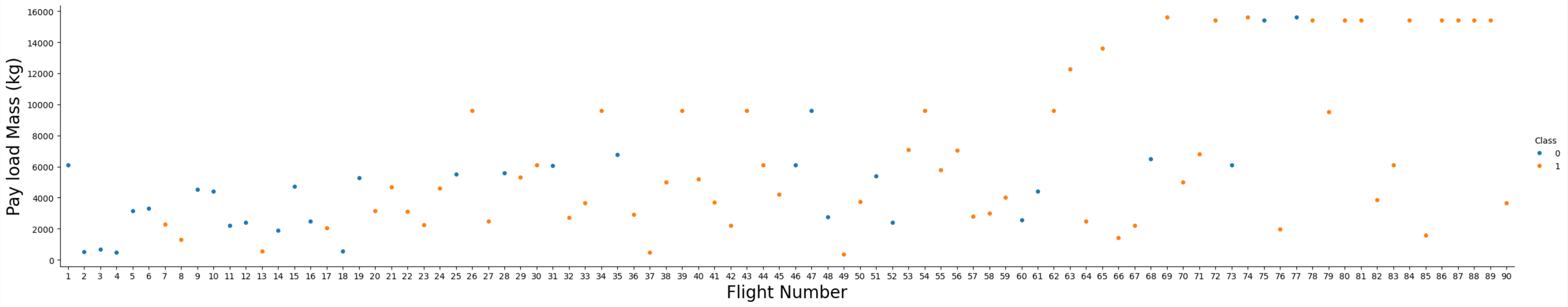
# Data Wrangling

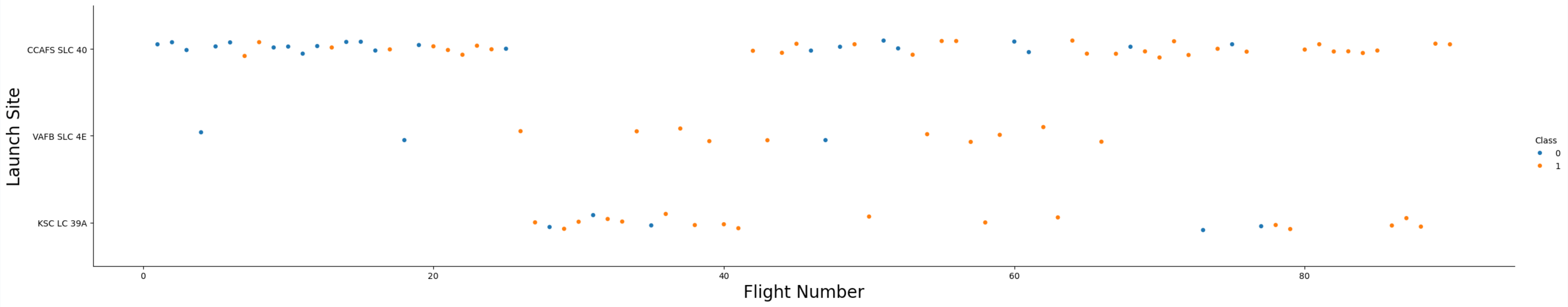
- Description of  how data were processed
  - The number of launches for each launch site was counted using the .value_counts() method.
  - The percentage of each attribute that was missing was determined.
  - Th data types for each attribute was determined to know which attributes were categorical and which were numerical
  - The landing outcomes were also dummy-coded to 0 if bad and 1 if good.

- Counted number of launches for each site → determined the nature(numerical or categorical) of each attribute → dtypes of each column was determined → Landing outcomes dummy-coded to 0 or 1 depending on outcome.
- GitHub

# EDA with Data Visualization

- Some Charts



# EDA with Data Visualization

- The first chart shows that as Payload Mass and Flight number increases the chances of a successful landing increases
- From the second chart, it is clear that CCAFS SLC-40 had the highest number of launches.

- GitHub

# EDA with SQL

- SQL Queries performed
  - `%sql select distinct Launch_Site from SPACEXTBL`
  - `%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5`
  - `%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'`
  - `%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1'`
  - `%sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'`
  - `%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000`
  - `%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome`
  - `%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)`



- GitHub

# Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - A map of the NASA Space station was created. Circles and markers were added to indicate launch sites and the outcome (red for failure and green for success) of each launch for the launch sites
- Explain why you added those objects
  - These objects were added to help us determine if location played a role in the success or otherwise of the landing. We are also able to determine that launch sites are located far away from infrastructures and cities.
- GitHub

# Build a Dashboard with Plotly Dash

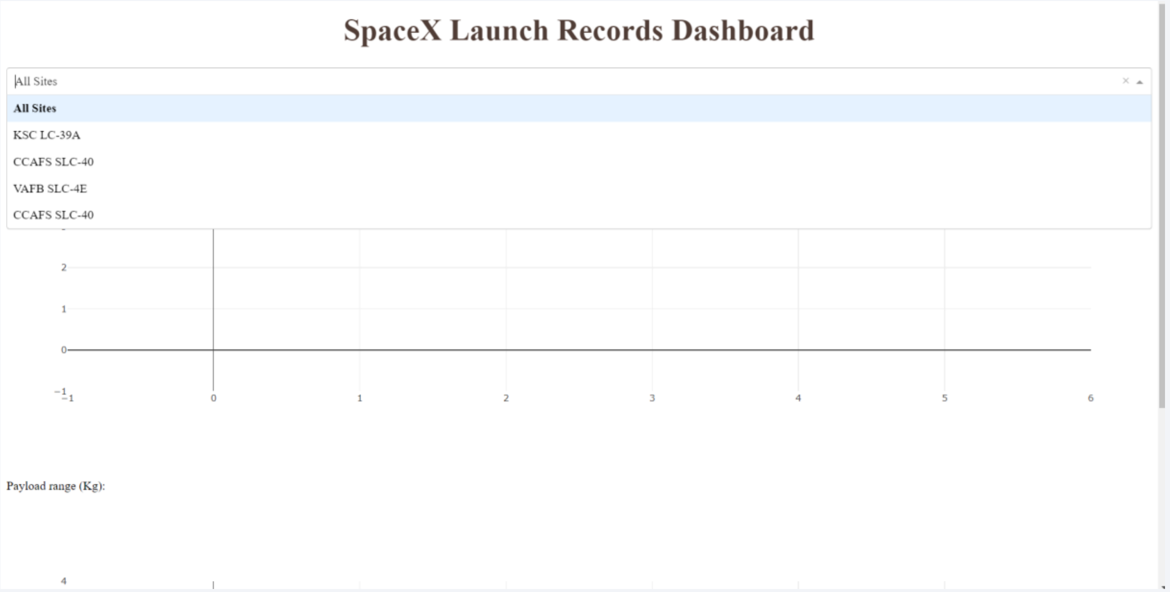
- Dashboard features
  - A dropdown of launch sites from which any launch site or all launch sites can be selected
  - A pi chart that changes based on what is selected from the above
  - A rangeslider to choose a range of payload mass
  - A scatter chart that changes according to the payload mass range selected
- Explain why you added those plots and interactions
  - These interactive features helped to determine how the success of the launches are influenced by payload mass and the percentage of successful launches from each site.
- GitHub

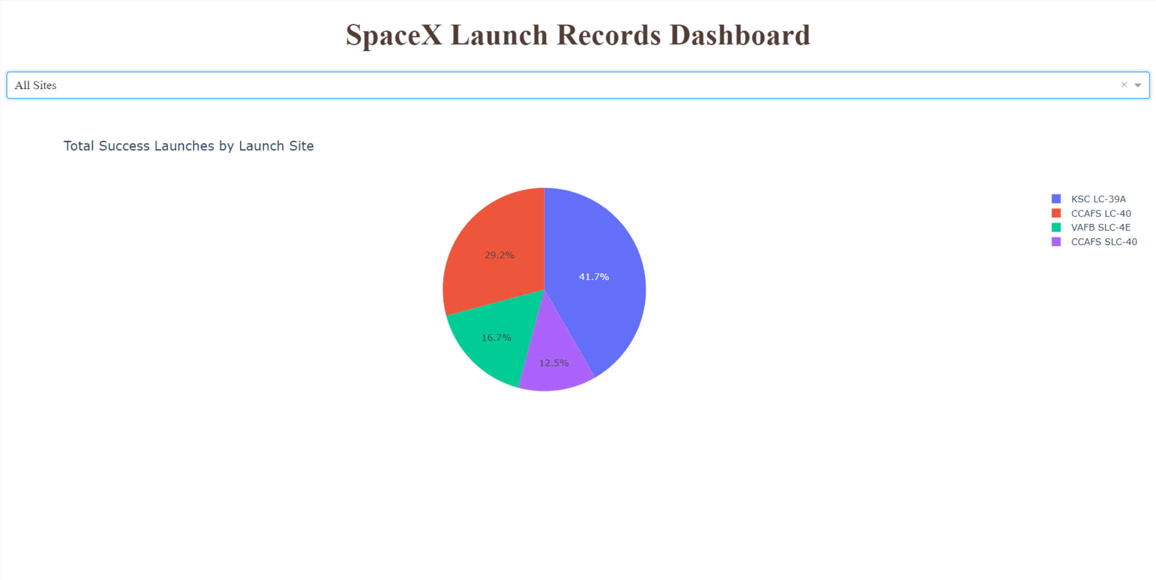
# Predictive Analysis (Classification)

- Building, evaluating, improving, and finding the best performing classification model
  - The data was split into training and testing data sets and then fitted with Logistic regression, Support Vector Machine, Decision Tree Matrix and k-Nearest Neighbor models.
  - The Decision Tree Model performed the best
- You need present your model development process using key phrases and flowchart
  - Splitting → creating a classification model object and parameters dictionary → Creating a GridSearch Object with the parameters and object created and fitting the object to get the best performing parameters → apply the best parameters on the test set
- GitHub

# Results

- Exploratory data analysis results
  - ES-L1, GEO, HEO, and SSO are the orbits with the best success rates
  - Launches with Payload mass greater than 10000 are more likely to result in failure
- Interactive analytics



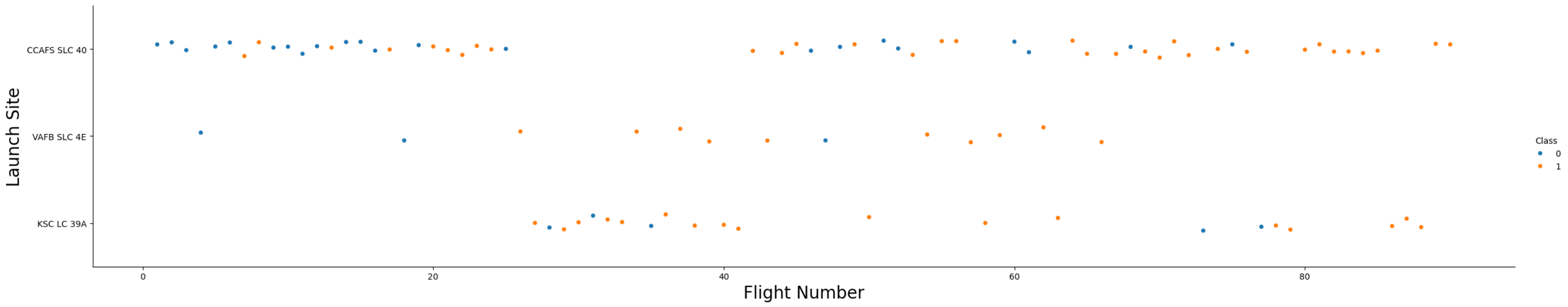



Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

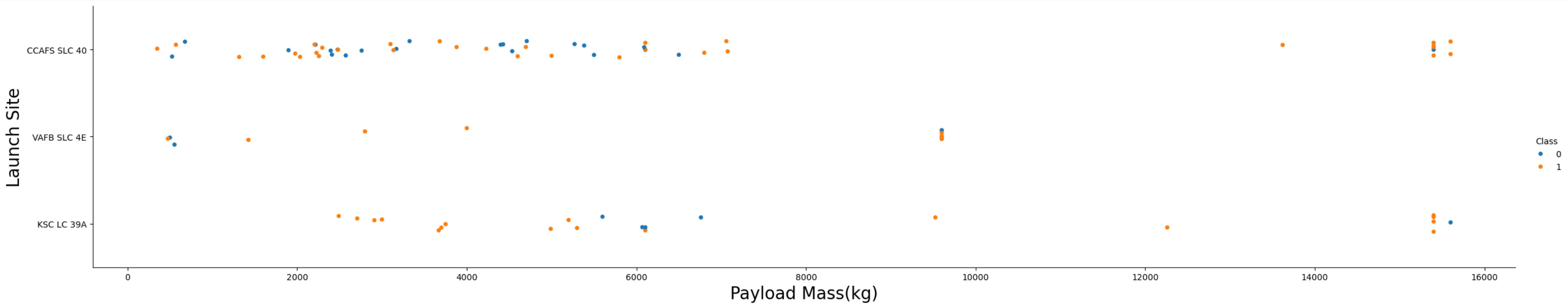
- Scatter plot of Flight Number vs. Launch Site




- The chart shows that the highest success rate is achieved at flight numbers below 20 and launch site CCAFS SLC-40

# Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

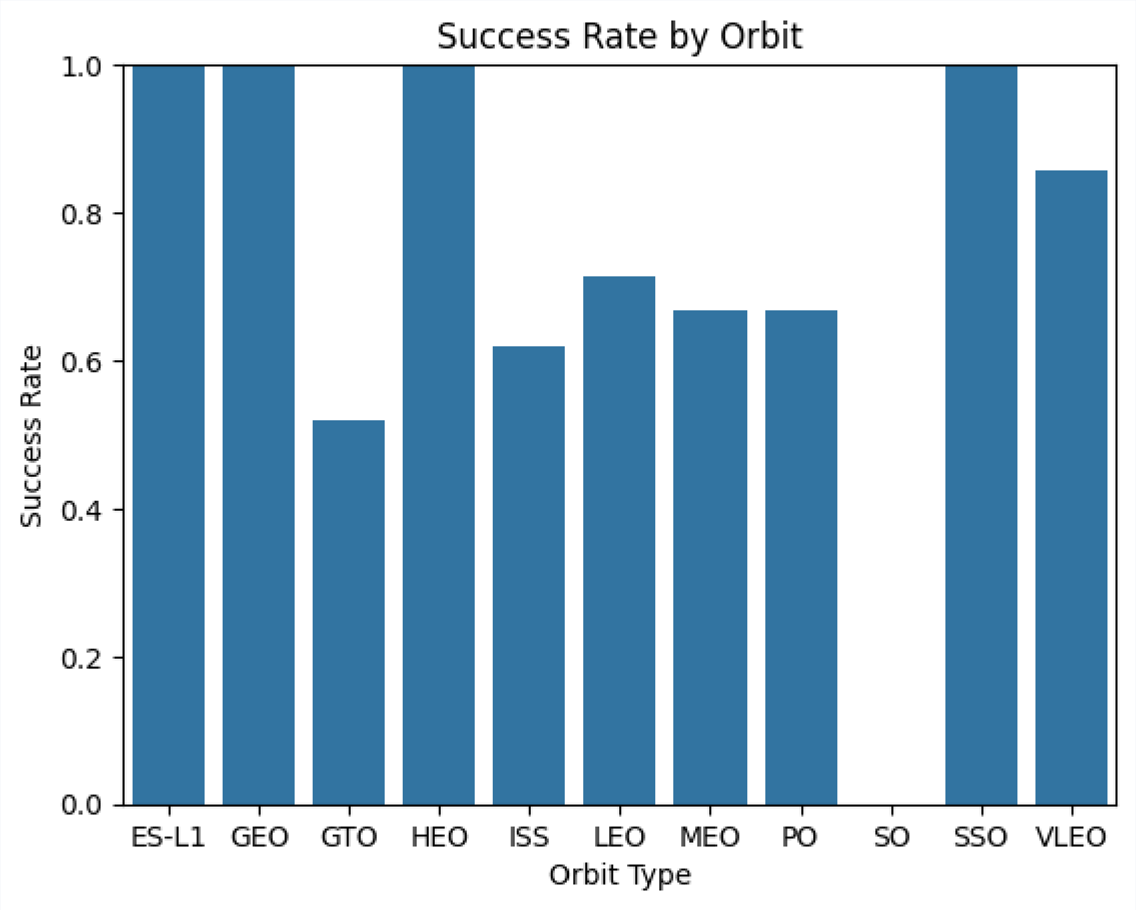



- Payload mass greater than 10000kg reduces the chances of success regardless of launch site.

# Success Rate vs. Orbit Type

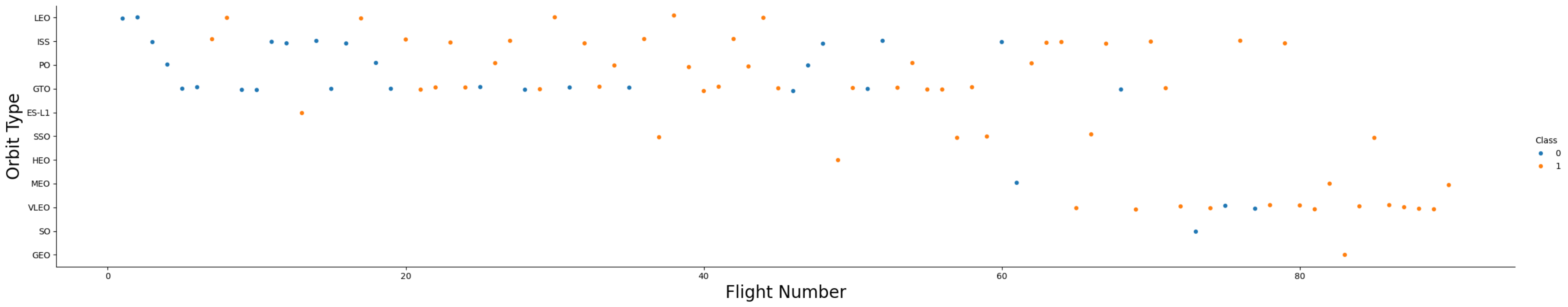
- Bar chart for the success rate of each orbit type

- ES-L1, GEO, HEO and SSO has 100% success rates while GTO is the lowest at about 50%



# Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type




- Flight numbers greater than 60 lands on the SSO and larger orbits while lower flight numbers land on the smaller orbits.

- Smaller orbits also appear to have higher success rates.

# Payload vs. Orbit Type

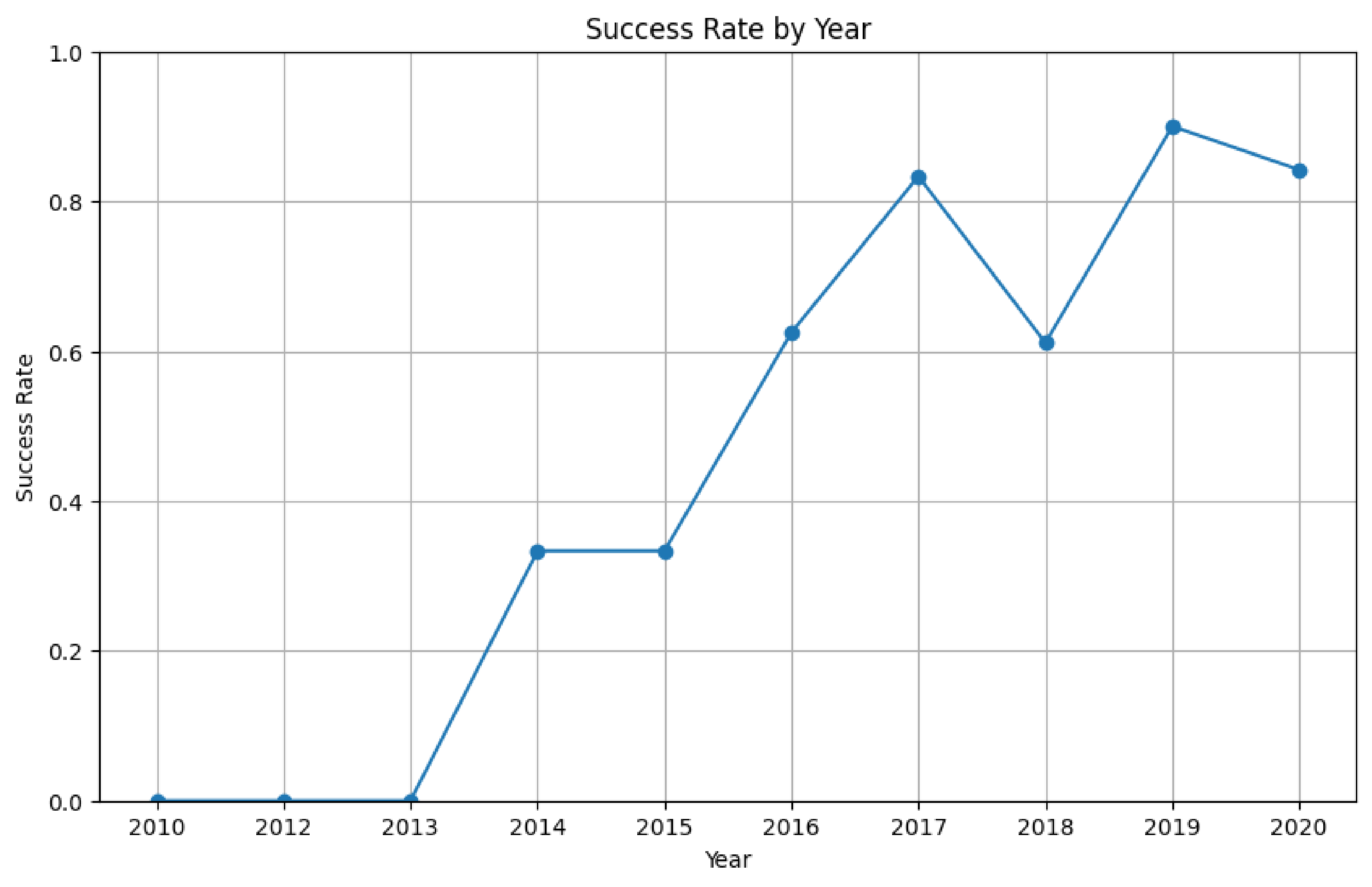
- Scatter point of payload vs. orbit type


- Payload mass of 10000kg and heavier mostly lead to failures

- Smaller orbits have higher landing and success rates.

# Launch Success Yearly Trend

- Line chart of yearly average success rate
- 2019 recorded the highest ever success rates at about 90%



# All Launch Site Names

- Unique launch sites
  - CCAFS SLC 40
  - VAFB SLC 4E
  - KSC LC 39A
  - CCAFS LC-40

  - %sql select distinct Launch_Site from SPACEXTBL .
  - The query selects unique launch sites from the SPACEXTBL table

# Launch Site Names Begin with 'CCA'

- Query result
- %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
- The query selects the top five results from the SPACEXTBL table where the launch site name starts 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 attemp |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attemp |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attemp |

# Total Payload Mass

- Total payload carried by boosters from NASA
  - %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'
- Query result
  - 45596Kg
  - It shows that the payload mass for NASA totaled 45596kg.

# Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
  - %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1’
- Query result with a short explanation
  - 2928.4Kg
  - The Average payload mass for Booster Version F9 v1.1 is 2928.4kg

# First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
  - %sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)’
- Query result with a short explanation
  - 2015-12-22
  - The first successful landing on a ground pad occurred on the 22nd December 2015.

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - %sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

- Query result with a short explanation
  - Only the shown booster version have successful drone ship landings with payload between 4000 and 6000kg

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

# Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
  - %sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome
- Query result with a short explanation
  - The figures show the total number of successful and failure outcomes

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

# Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
  - %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- Query result with a short explanation
  - The result shows the booster versions that have 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 |

# 2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - %sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'

- Query result with a short explanation
  - The 2015 launch records is shown in the results

| 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 |

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

- Ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
  - %sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by 1 order by 2 desc

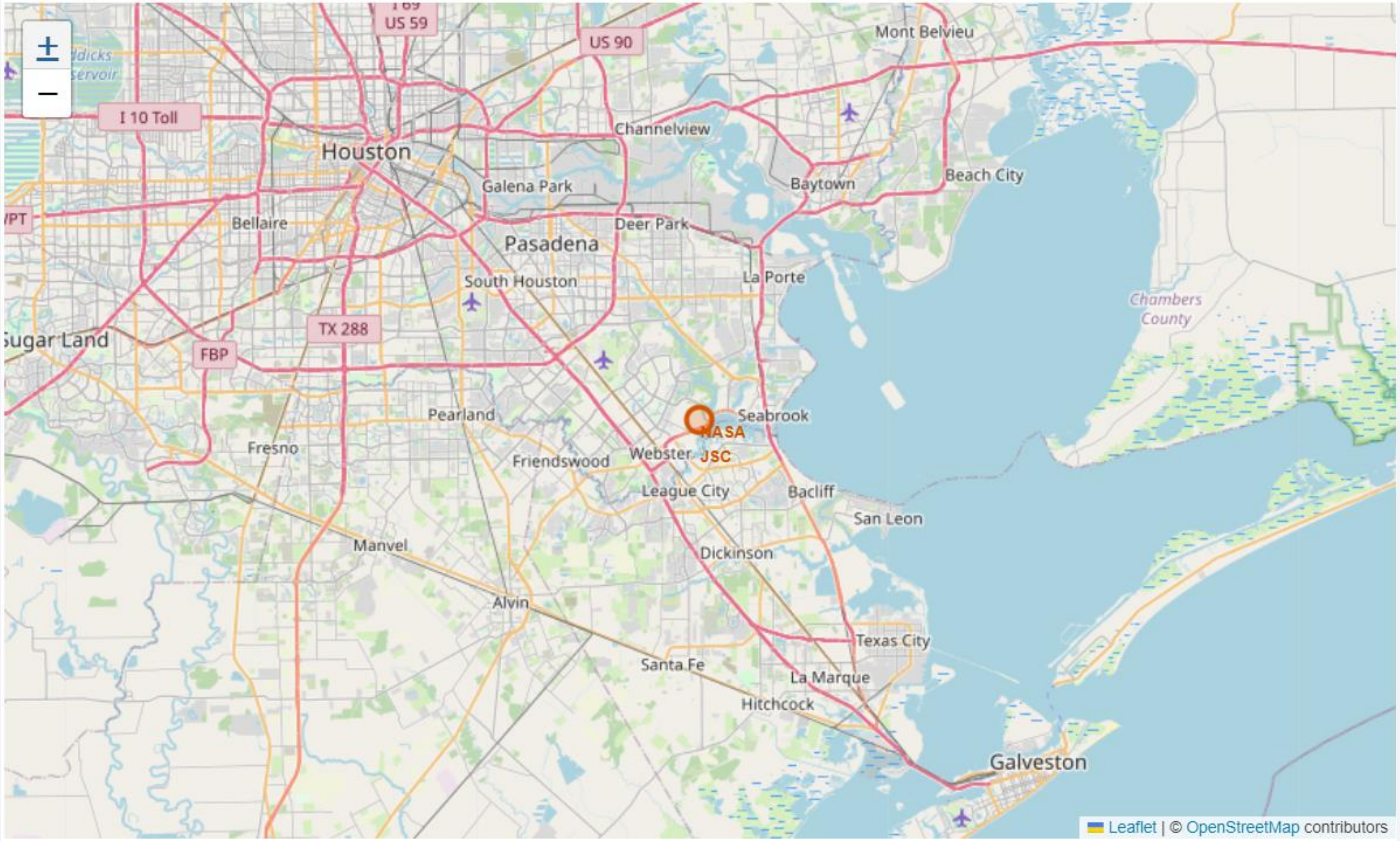
| Landing_Outcome | count(Landing_Outcome) |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

- Query result with a short explanation
  - The ranking is shown in the results.
  - There were 10 records with no attempts

Section 3

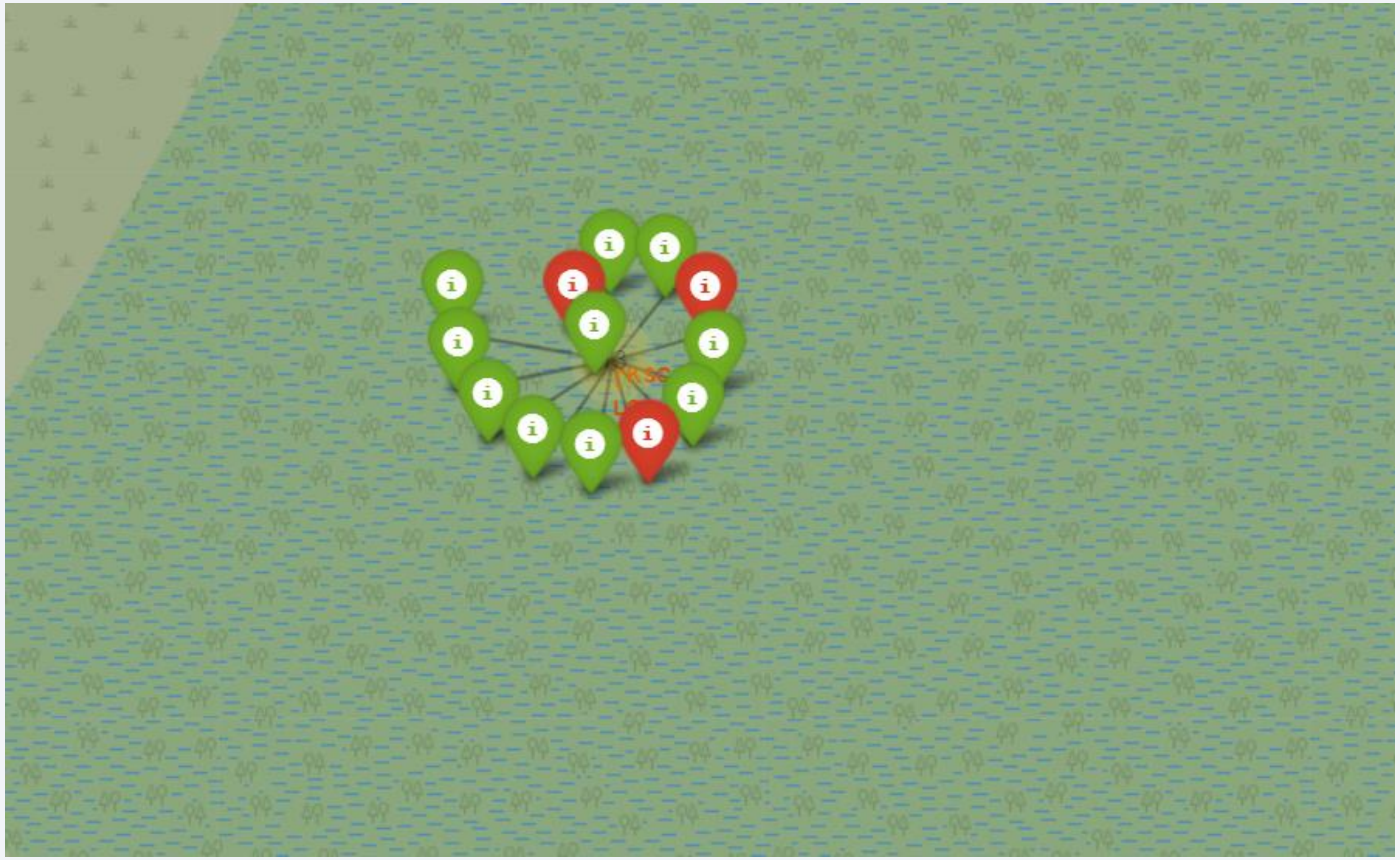
# Launch Sites Proximities Analysis

# Map Showing NASA Johnson Space Center at Houston, Texas.



# Map Showing the Launch Sites

# Map showing the KSC LC-39A launch site

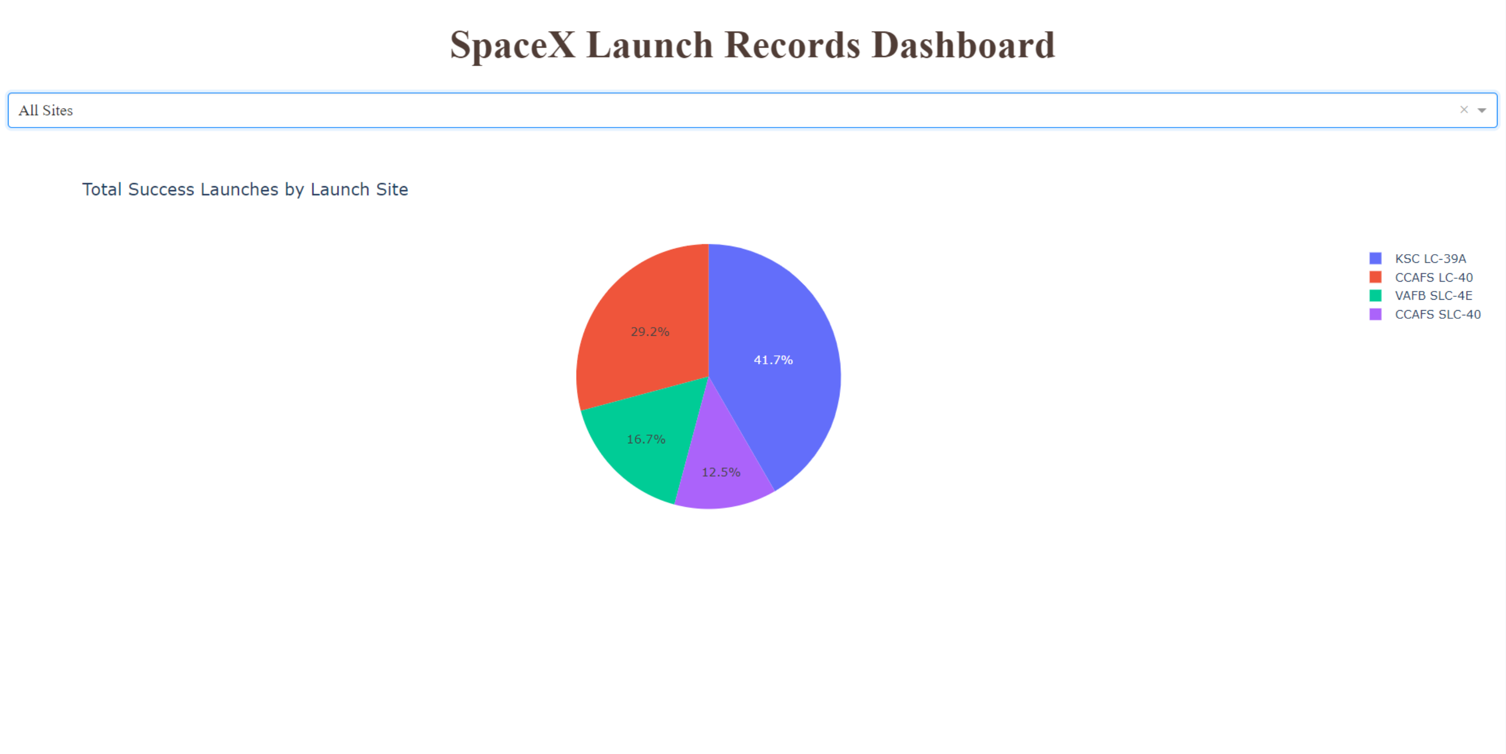


The green markers indicate successful launches while the red markers represent failures

Section 4

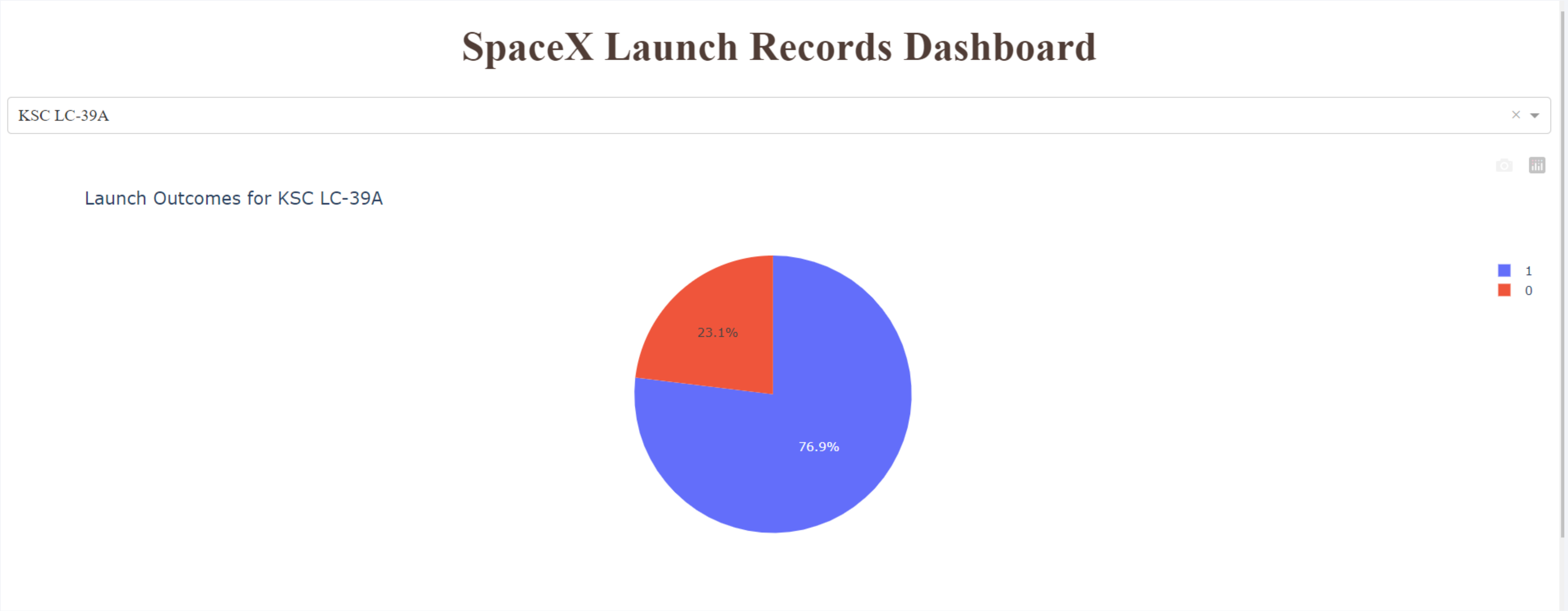
# Build a Dashboard with Plotly Dash

# Pie Chart showing Successful Launches by Launch site



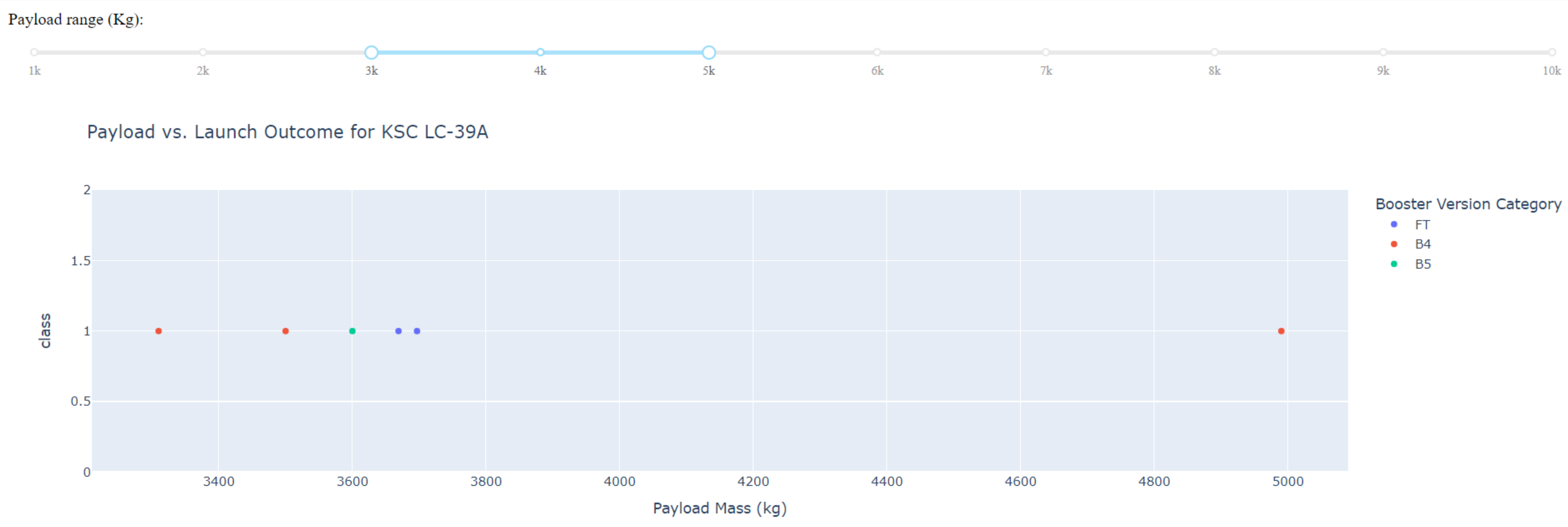

The chart shows the percentage of all successful launches contributed by each site. KSC LC-39A contributed the most successful launches.

# Dashboard Screenshot of the KSC LC-39A Launch Site




The KSC LC-39A has the highest success launch ratio with almost 77% success rate.

# Dashboard Screenshot of Payload vs Launch Outcome for KSC LC-39A



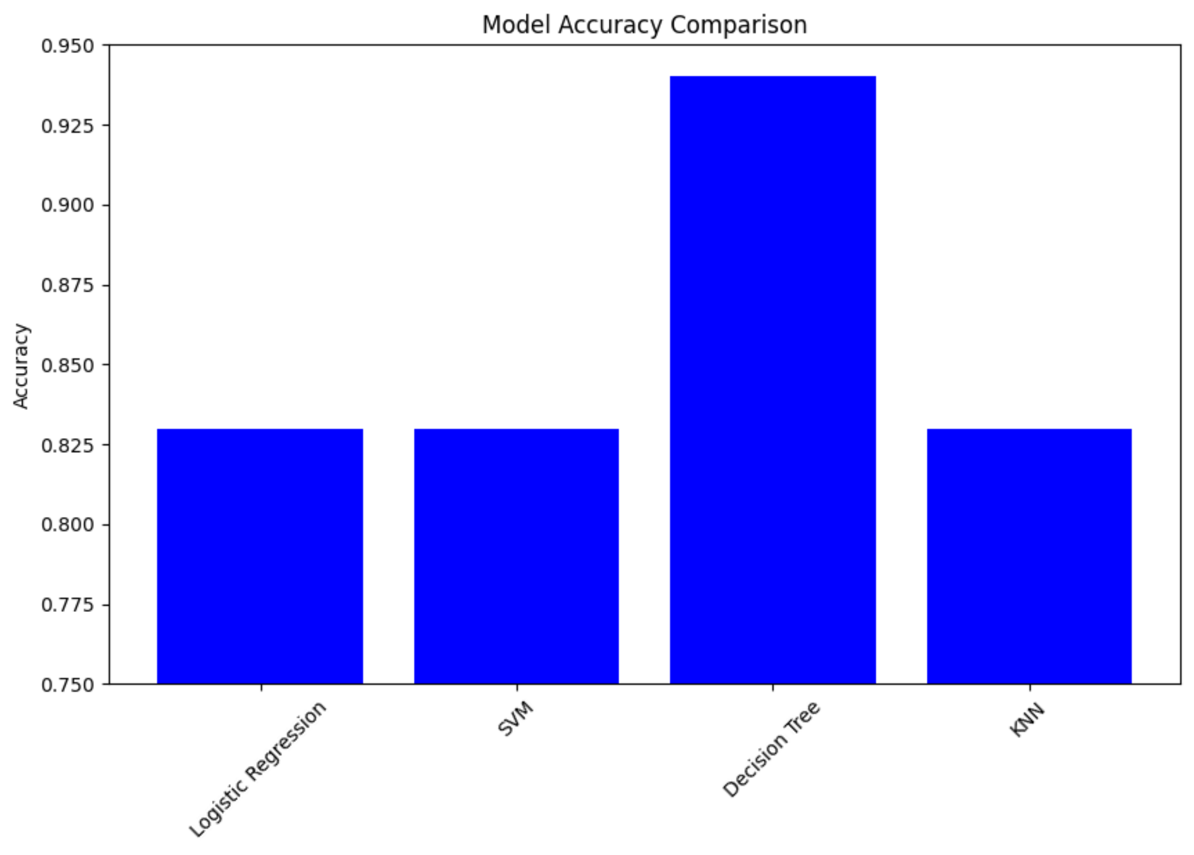

All launches with payload in the range 3000 – 5000kg were successful for the launch site.

Section 5

# Predictive Analysis (Classification)

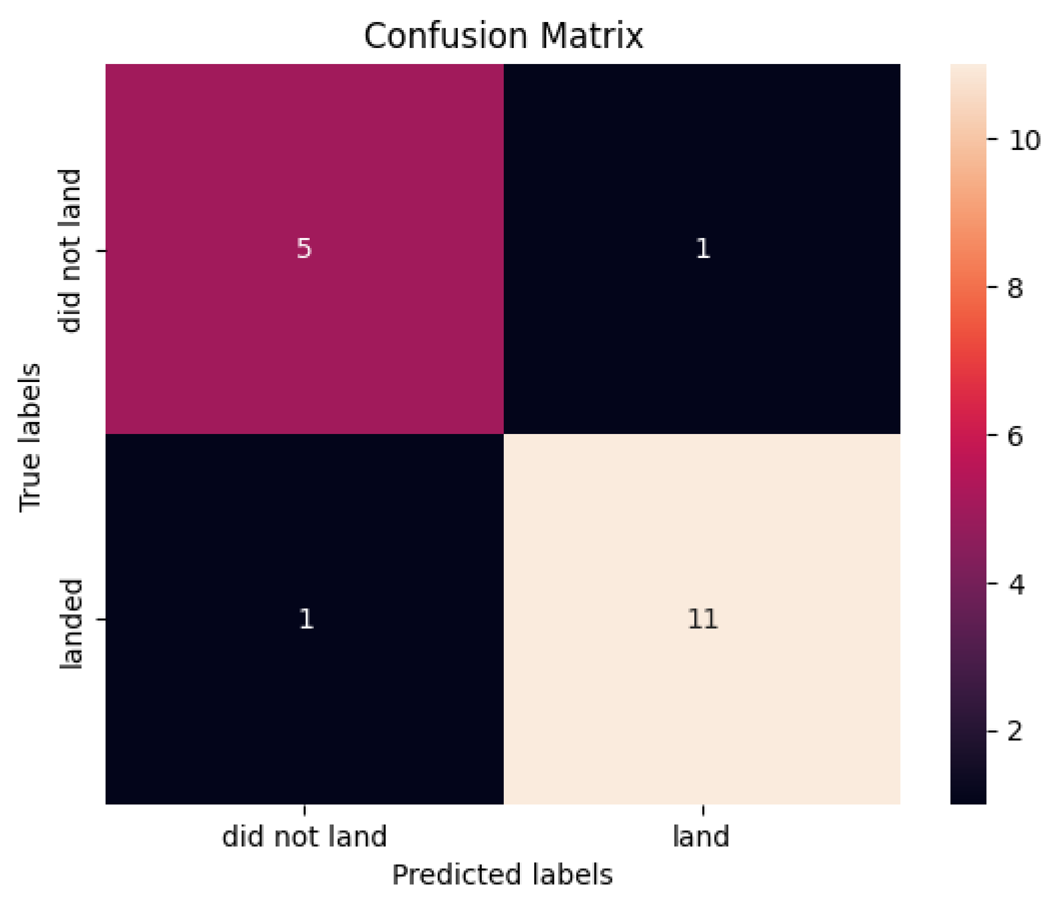
# Classification Accuracy

- The Decision Tree model performs the best. All other models have equal performance.



# Confusion Matrix

- This confusion matrix shows a model with high accuracy, precision, recall, and F1 score, suggesting it performs well in correctly identifying both 'land' and 'did not land' cases. This is shown in the high values in the True Positive and True Negative quadrants.



# Conclusions

- Decision Tree is the best model to use for the prediction.
- The KSC LC-39A seem the most probable to give a successful launch
- It is advisable to keep the payload mass under 5000Kg
- The FT Booster Category is most likely to lead to a successful outcome.
- …

# Appendix

Find the method performs best:

```
data1 = {
    'Model': ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN'],
    'Accuracy': [0.83, 0.83, 0.88, 0.83]
}
model_df = pd.DataFrame(data1)
model_df
```

| | Model | Accuracy |
|---|---|---|
| **0** | Logistic Regression | 0.83 |
| **1** | SVM | 0.83 |
| **2** | Decision Tree | 0.88 |
| **3** | KNN | 0.83 |

```
plt.figure(figsize=(10, 6))
plt.bar(model_df['Model'], model_df['Accuracy'], color='blue')
plt.ylabel('Accuracy')
plt.title('Model Accuracy Comparison')
plt.xticks(rotation=45)
plt.ylim(0.75, 0.95)
plt.show()
```



New dataframe created to visualize the best performing model.

# Thank you!