IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Miranda Khairunnisa
May 2024

# Outline

# Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scrapping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data visualization
  - Interactive Visual Analytics with Folium lab
  - Interactive Dashboard with Ploty Dash
  - Machine Learning Prediction
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

# 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. The main objective of the project is to create a Machine Learning model that is able to predict if the first stage of the rocket launches will land successfully.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program?

Section 1

# Methodology

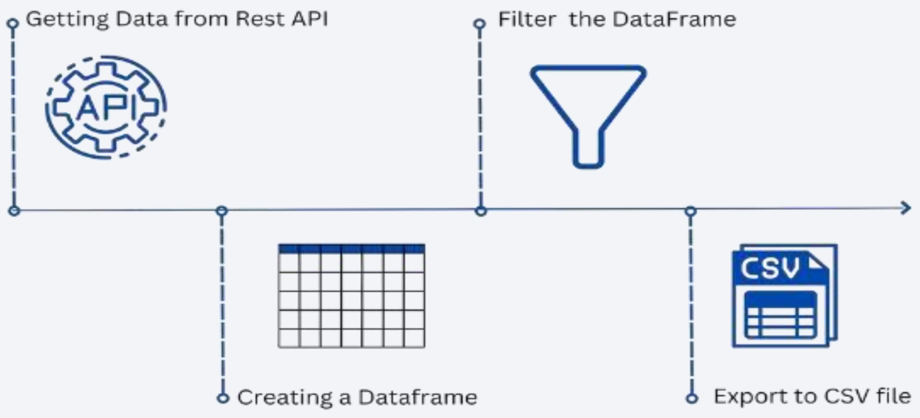
# Methodology

## Executive Summary

- Data collection methodology:
  - Using SpaceX API and Web Scraping
- Perform data wrangling
  - One-hot encoding on 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

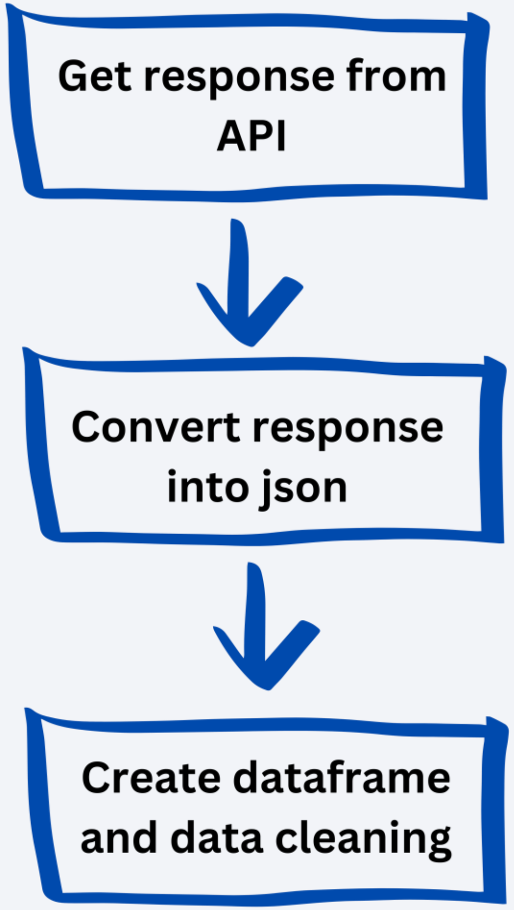
# Data Collection

- The data was collected using various methods:
  - SpaceX launch data collected from SpaceX REST API.
  - This API gave data about launches, including information about the rocket used, payload delivered, launch specification, landing specifications, and landing results.
  - Next, data was normalized and cleaned of missing values.
  - Another data collection method that was performed is web scraping from Wikipedia for Falcon 9 launches records using BeautifulSoup package. The objective was to extract the launch records as an HTML table then parse the table and convert it to a pandas dataframe for future analysis.



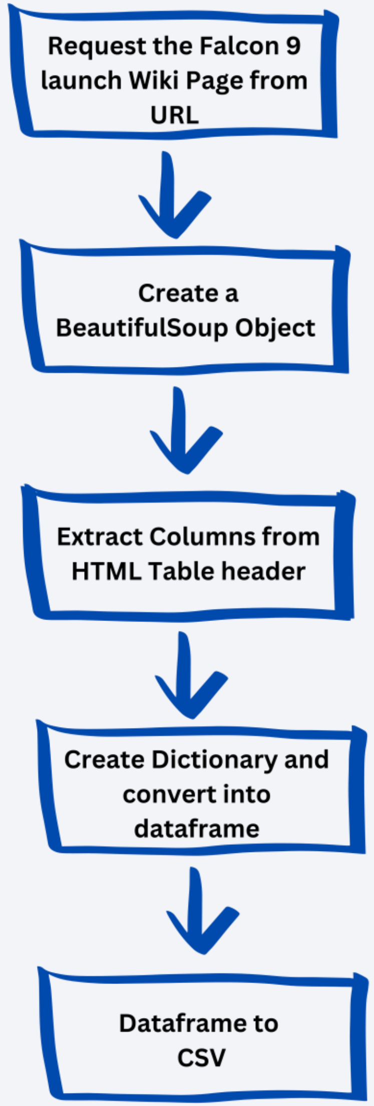
# Data Collection - SpaceX API

- A get request was to the SpaceX API to collect and clean the requested data. Some data wrangling and formatting was required to perform data analysis.
- Link to the notebook: Data Collection through API



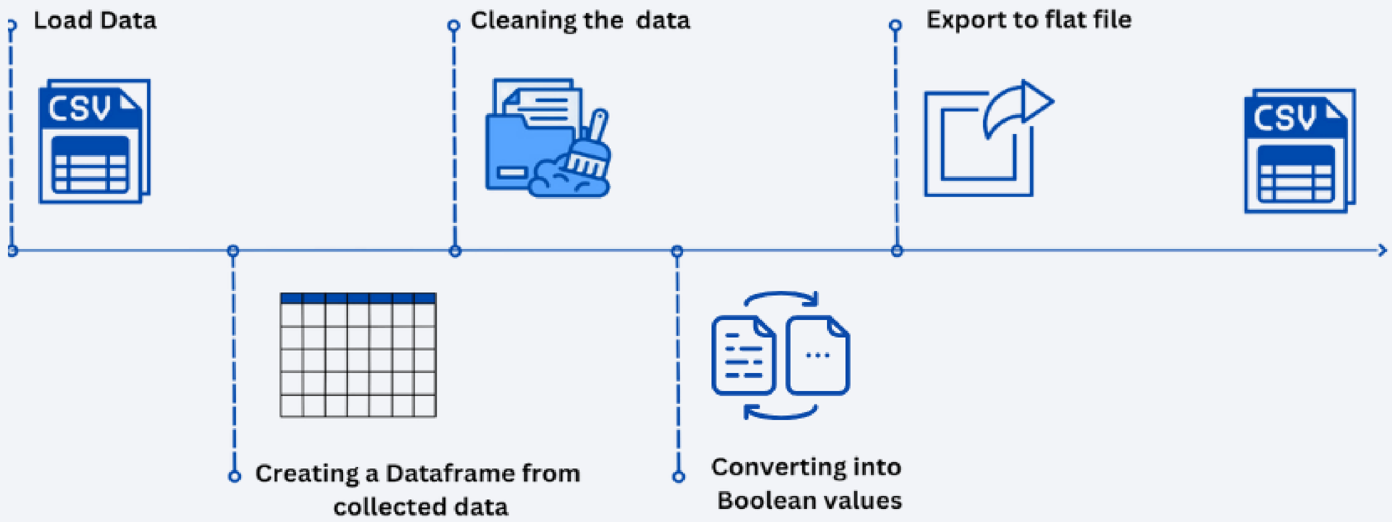
# Data Collection - Scraping

- Web scrapping is used on the Falcon 9 launch records with BeautifulSoup.
- Link to notebook: [Data Collection through Web Scraping]



# Data Wrangling

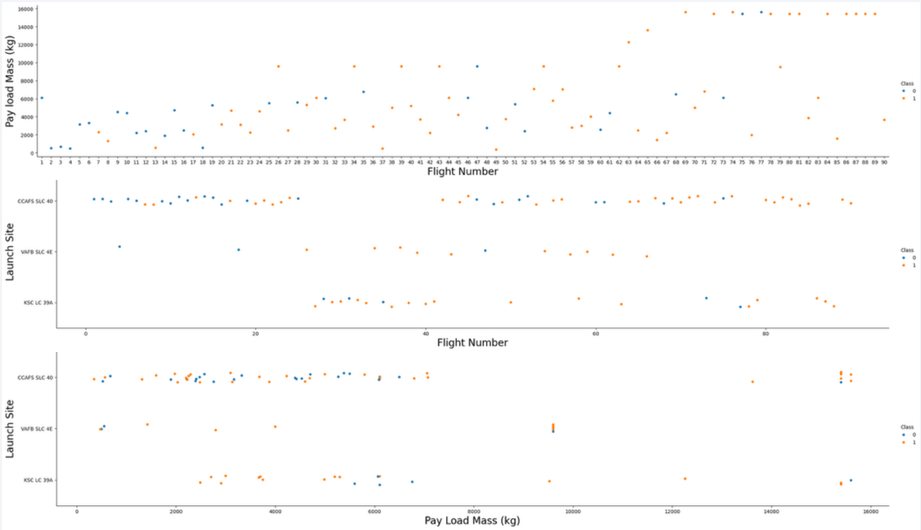
- Exploratory data analysis is performed to determine the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv.
- Link to notebook: [Data Wrangling]()

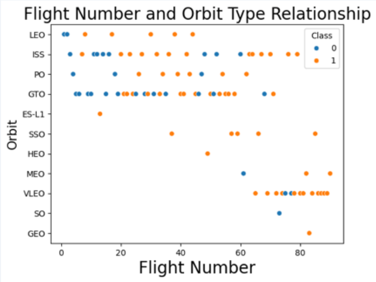


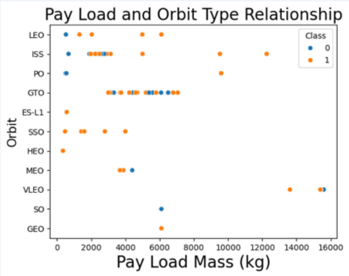
# EDA with Data Visualization

## Scatter Graphs :

- Payload vs Flight Number
- Launch Site vs Flight Number
- Launch Site vs Payload
- Flight Number vs Orbit
- Payload vs Orbit



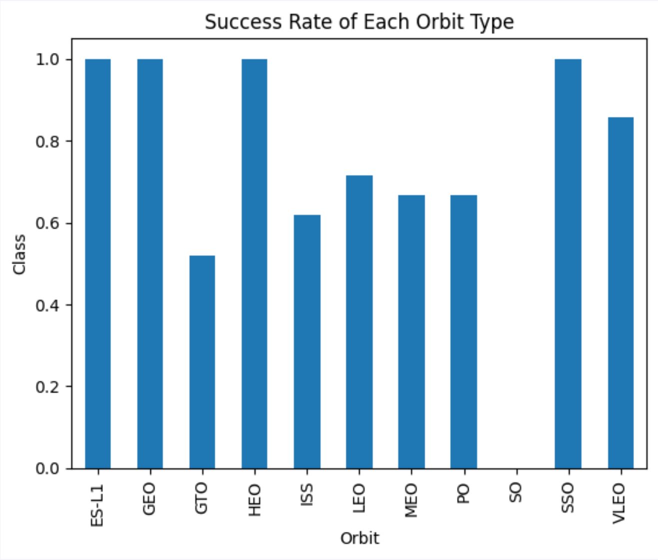




## Bar Graphs :

Success Rate vs Orbit Type

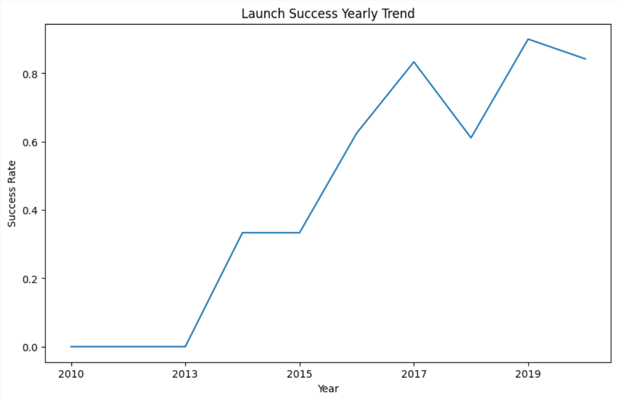
Bar graphs are best suited to represent relation between two categorical variables. In this project it is used to find relation between Success rate and Orbit type.




## Line Chart:

Launch Success Yearly Trend

Line chart is used in this project to plot the average launch success trend against previous years which helps in prediction of future launch outcomes.



# EDA with SQL

SQL is designed for a specific purpose to query data contained in a relational database. Due to this it is an indispensable tool for data scientist to deal with real world data driven problems. In this project, IBM's DB2 was used for Cloud as database which is a fully managed SQL service.

- SQL queries were used to:

  - Display the names of the unique launch sites in the space mission.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the names of the boosters which have success in drone ship.
  - List the total number of successful and failure mission outcome.
  - List the names of the booster versions and launch site names.
  - Rank the count of landing outcomes such as Failure on drone ship or Success on ground pad.

EDA with SQL

# Build an Interactive Map with Folium

- All launch sites were marked, and added map objects such as markers, circles, lines to mark the success of failure of launches for each site on the folium map.

- The feature launch outcomes (success and failure) were set to class 1 and 0.
  i.e, 1 for success and 0 for failure.

- Using the color-labeled marker clusters, the launch sites having relatively high success rate were identified.

- The distances were calculated between a launch site to its proximities. The following questions were answered:
  - Are launch sites near railways, highways and coastlines ?
  - Do launch sites keep certain distance away from cities ?

[Interactive Visual Analytics with Folium]

# Build a Dashboard with Plotly Dash

- An interactive dashboard was built with Plotly dash using IBM's Theia Platform.
- Pie charts were plotted showing the total launches by a certain sites.
- Scatter graph were plotted showing the relationship with Outcome and Payload  Mass(kg) for the different booster version.

[Dashboard with Plotly]

# Predictive Analysis (Classification)

- The data was loaded using Numpy and pandas, transformed the data, split the data into training and testing.
- Four different Machine Learning Models were built and tune the hyperparameters using GridSearchCV.
- Used accuracy as metrics for the model, improved the model using features engineering and algorithm tuning.
- The best performing classification model were identified.

[Machine Learning Prediction]

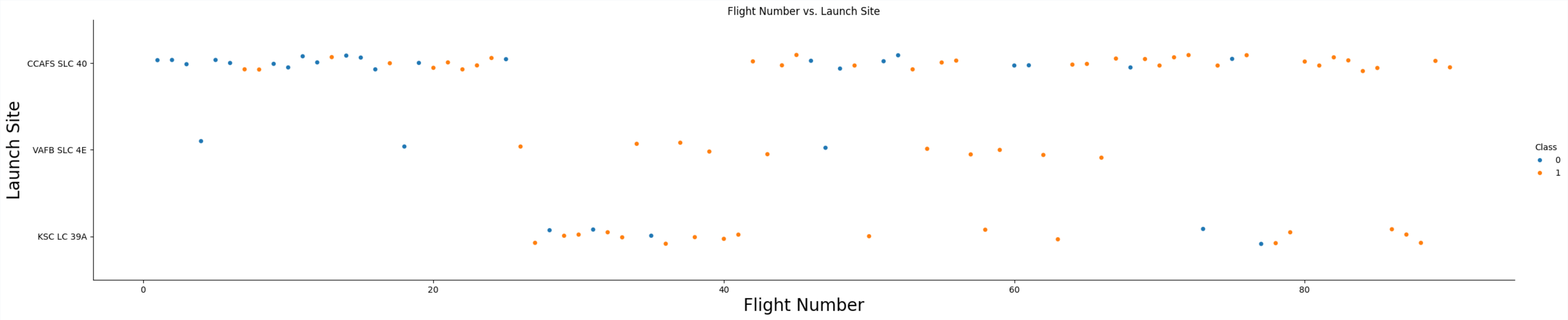
# Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSCLC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Section 2

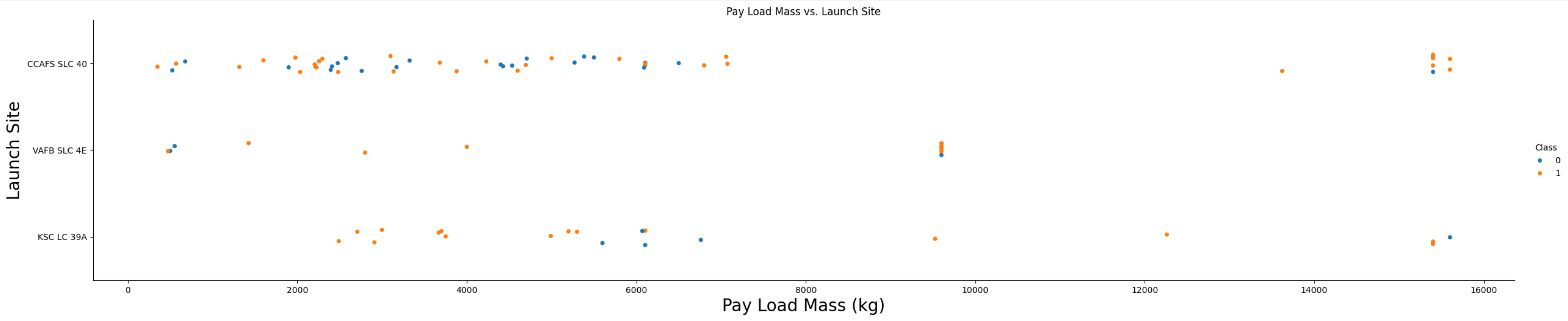
# Insights drawn from EDA

# Flight Number vs. Launch Site



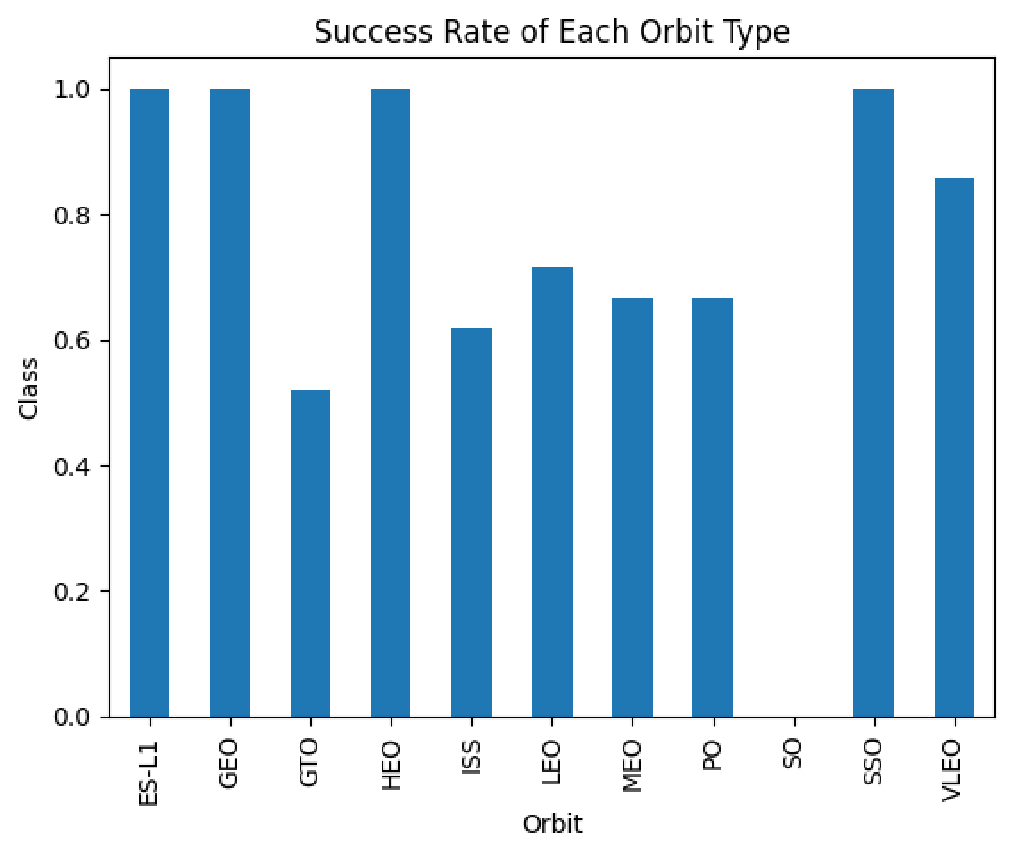

Most launches were conducted at CCSFS SLC 40, followed by KSC LC 39A and VAFB SLC 4E, with CCSFS SLC 40 showing a relatively higher frequency of launches. CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

# Payload vs. Launch Site



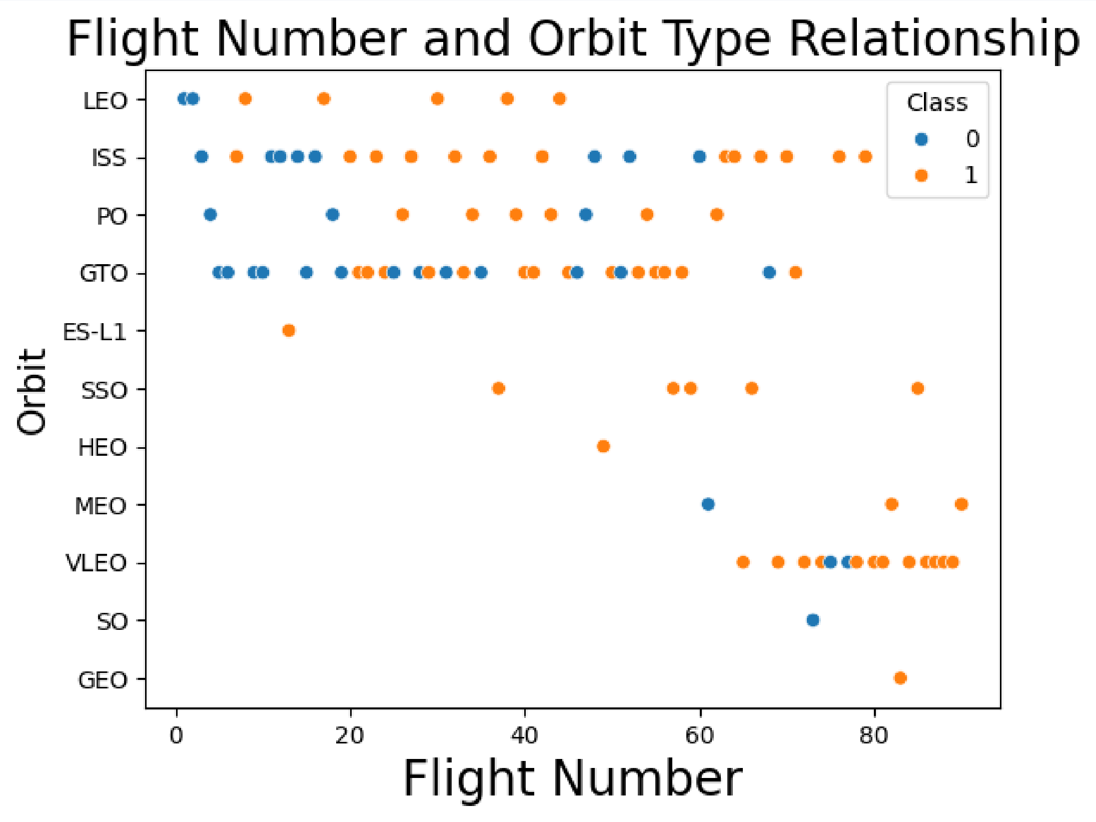

The scatter plot shows that CCSFS SLC 40 handled most of the launches across a wide range of payload masses, especially in the lower to mid-range payloads (below 10000 kg). KSC LC 39A is associated with heavier payloads, with several launches exceeding 10000 kg, mostly showing successful outcomes. In contrast, VAFB SLC 4E primarily managed lighter payloads and had fewer overall launches compared to the other sites.

# Success Rate vs. Orbit Type



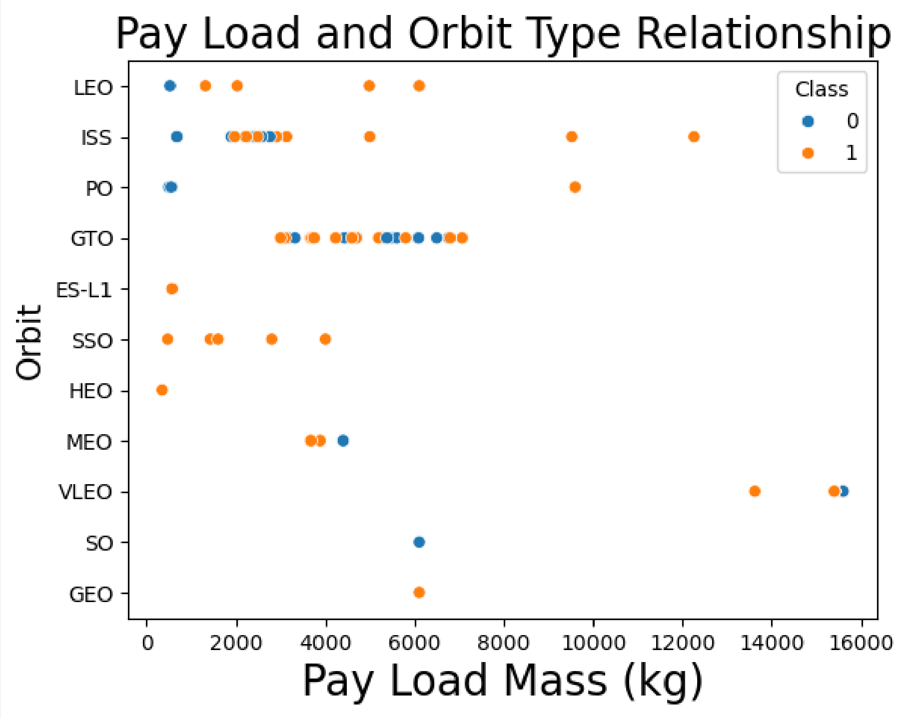

The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.

# Flight Number vs. Orbit Type



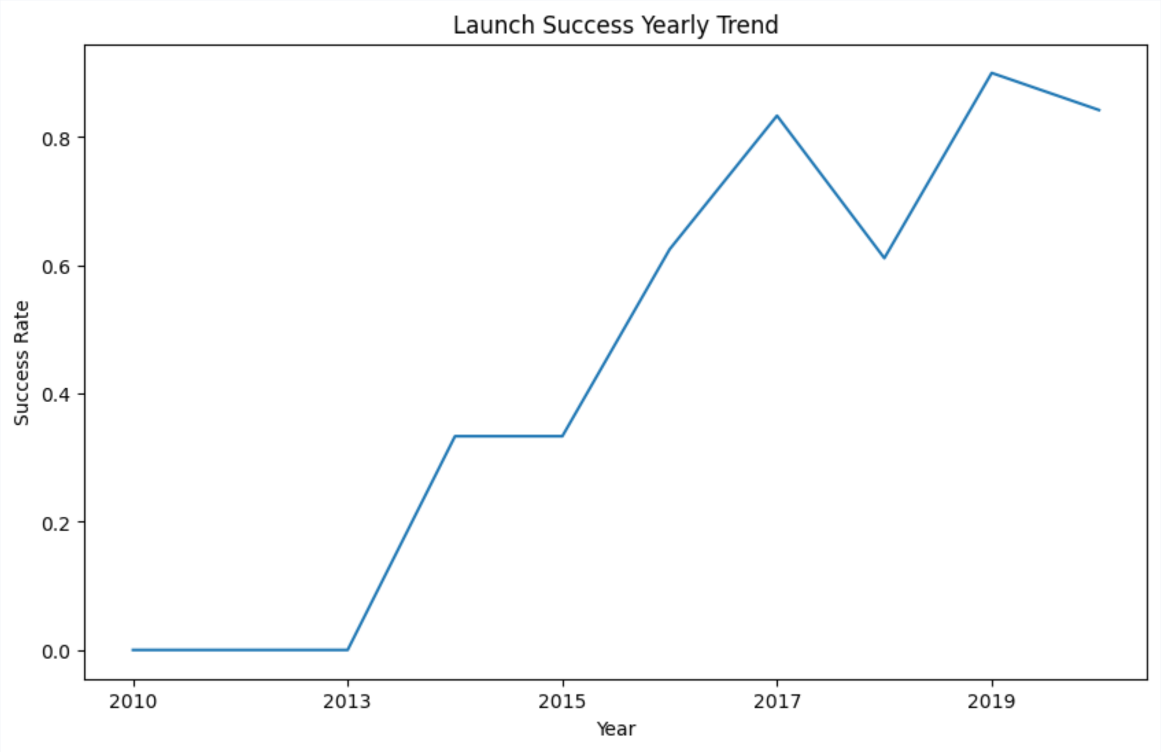

The LEO orbit the success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number in GTO orbit.

# Payload vs. Orbit Type




For heavier payloads, successful landings are more common in PO, LEO, and ISS orbits. Successful landings are also frequent for lighter payloads, particularly for missions targeting ES-L1, SSO, and HEO orbits. However, for GTO missions, the outcomes are mixed, with both successful and unsuccessful landings observed across different payload masses.

# Launch Success Yearly Trend




Launch success rate has increased significantly since 2013 and has stabilized since 2019, potentially due to advance in technology and lessons learned.

# All Launch Site Names

- The key word DISTINCT was used to show only unique launch sites from the SpaceX data.

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

```
%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 query below is used to display the 5 records where launch sites 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 | Mis |
|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | |
| 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 | |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | |

# Total Payload Mass

- The total Payload carried by boosters from NASA was calculated as 45596.

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

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

| SUM(PAYLOAD_MASS__KG_) |
| ---: |
| 45596 |

# Average Payload Mass by F9 v1.1

- The average Payload mass carried by booster version F9 v1.1 was calculated as 2928.4

```
%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 |

# First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was observed at 22 December 2015.

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

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

| MIN(Date) |
|---|
| 2015-12-22 |

## Successful Drone Ship Landing with Payload between 4000 and 6000

- Used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000.

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)'AND PAYLOAD_MASS__KG_
```

```
 * 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

```
%%sql SELECT Total_Successful_Mission_Outcomes, Total_Failure_Mission_Outcomes FROM
    (SELECT COUNT(*) AS Total_Successful_Mission_Outcomes FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%'),
    (SELECT COUNT(*) AS Total_Failure_Mission_Outcomes FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%')
```

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

| Total_Successful_Mission_Outcomes | Total_Failure_Mission_Outcomes |
|---|---|
| 100 | 1 |

# Boosters Carried Maximum Payload

- The booster that have carried the maximum Payload was determined using a subquery in WHERE clause and the MAX() function.

```
%%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 |

# 2015 Launch Records

- A combination of WHERE clause, LIKE, AND, and BETWEEN conditions were combined to filter for failed landing outcome in drone ship, their booster versions, and launch site names for year 2015

```
%%sql SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
    WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)'
```

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

| 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

```
%%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Landing_Count FROM SPACEXTBL
   WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DES
```

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

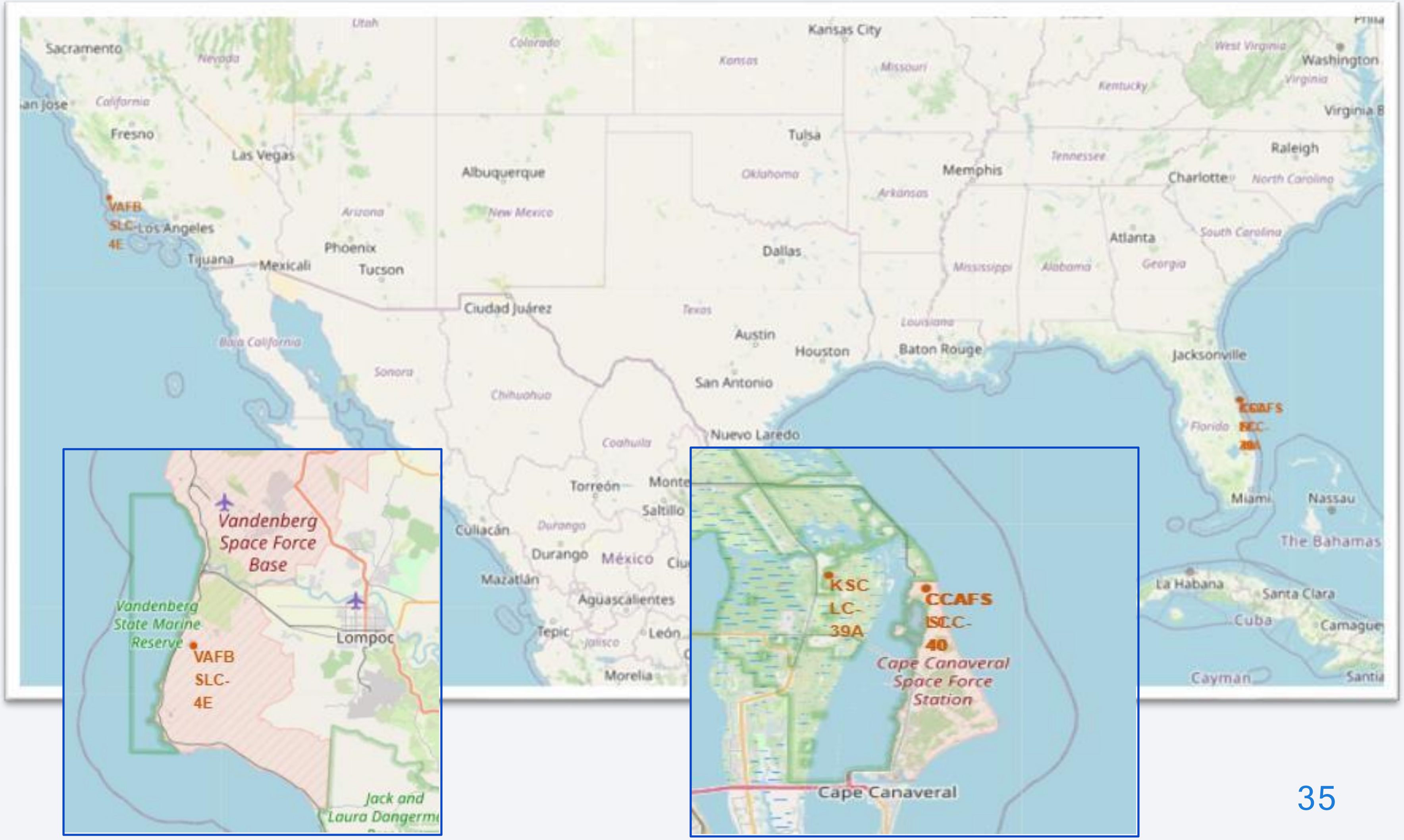
| Landing_Outcome | Landing_Count |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

# All Launch Sites Markers on a Global Map

- The SpaceX launch sites are close to the United States of America coasts i.e., Florida and California region.



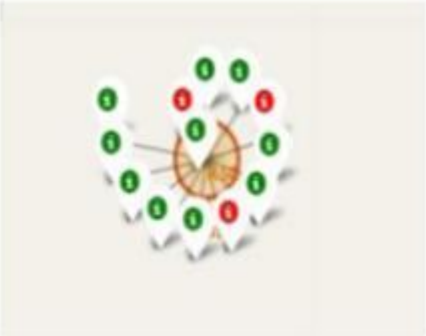
# Markers showing Launch Sites with Color Labels


**Green Marker** shows successful launches and **Red Marker** shows failures.

From these screenshots, **it can be easily sighted that KSC LC-39A has the maximum success rate.**


VAFB SLC-4E

KSC LC-39A



CCAFS LC-40

CCAFS SLC-40

# Launch Site Distance to its Proximities

- Are launch sites in close proximity to railways?

  Yes (Less than 2km)

- Are launch sites in close proximity to highways?

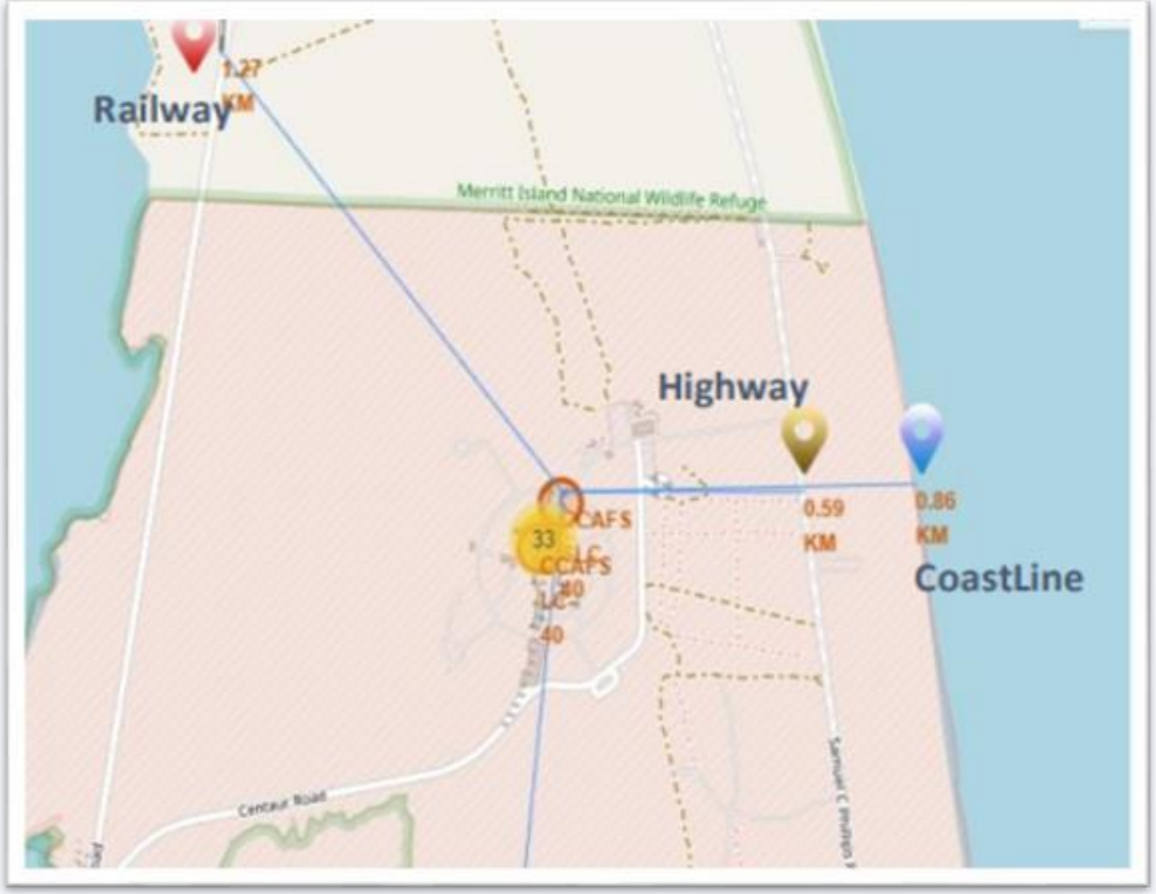
  Yes (Less than 2 Km)

- Are launch sites in close proximity to coastline?

  Yes (Less than 5 Km)

- Do launch sites keep certain distance away from cities?

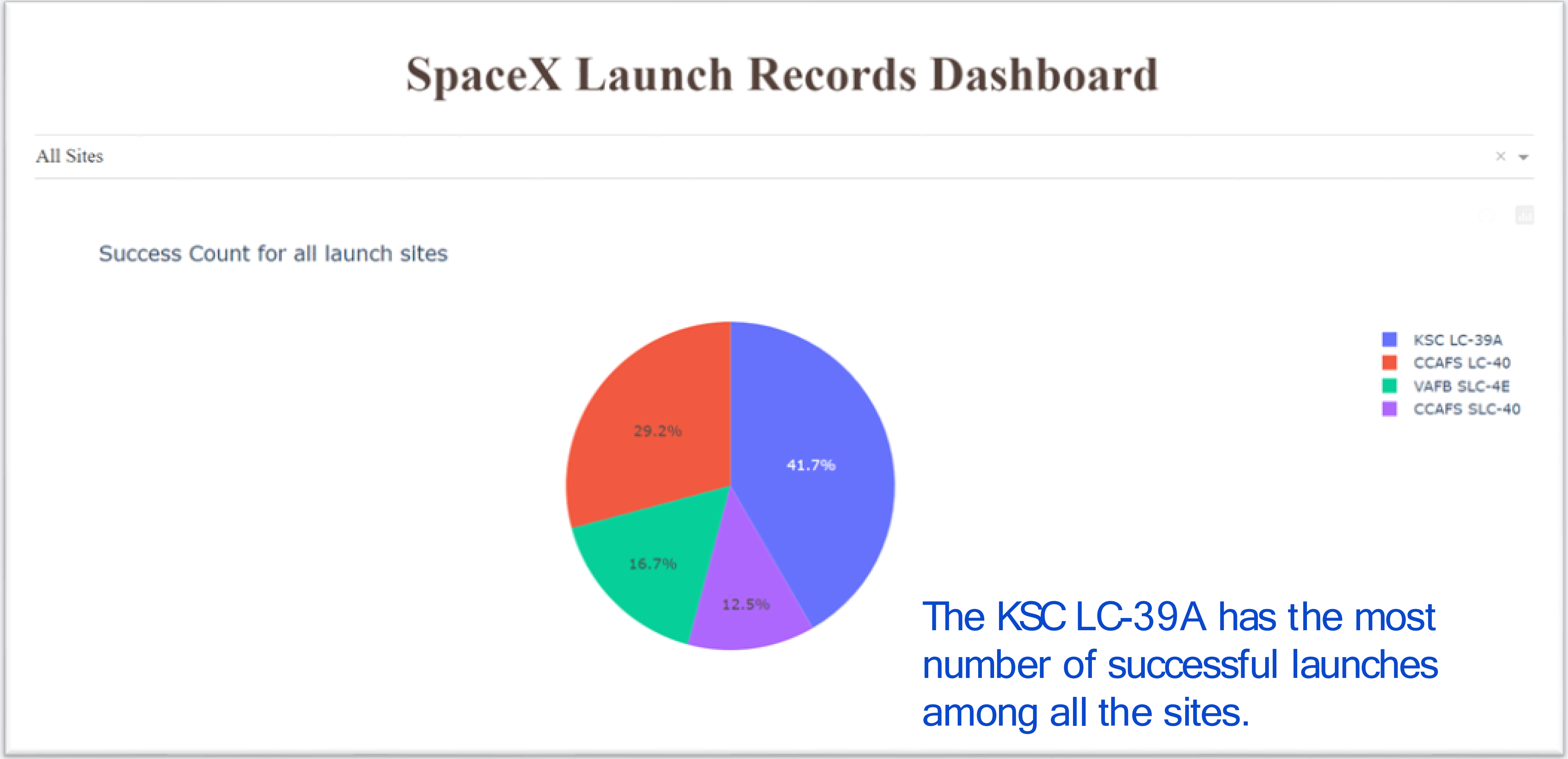
  Yes (More than 15 Km)



Section 4

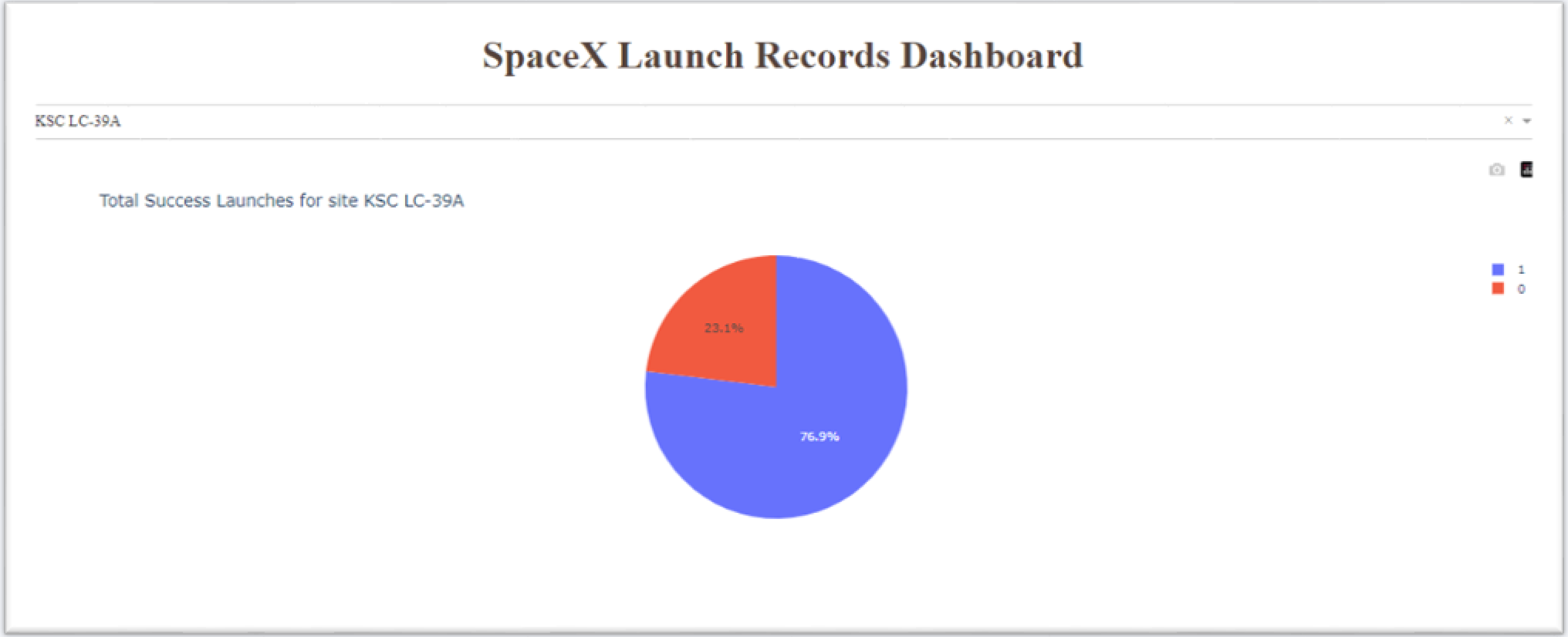
# Build a Dashboard with Plotly Dash

# Pie Chart showing the Success Percentage of Launches




The KSC LC-39A has the most number of successful launches among all the sites.

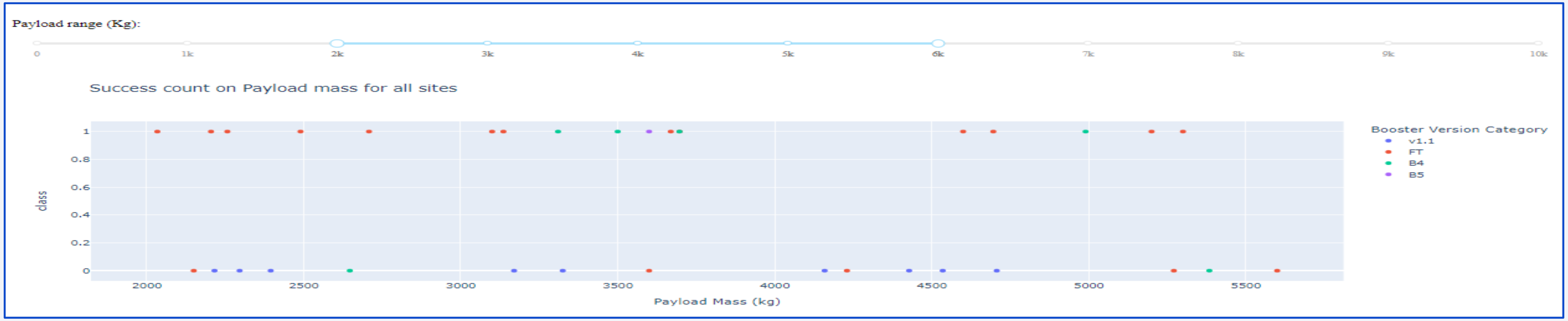
# Pie Chart showing Success ratio by Launch Site

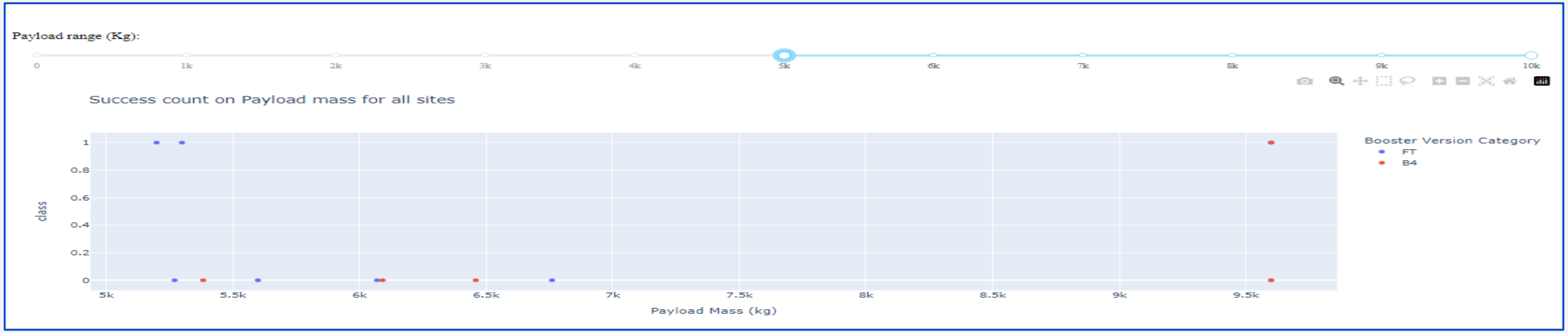



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

# Scatter Plot of Payload Vs Launch Outcome for all Sites

- We can see the success rate for low weighted Payloads is higher than that of heavy weighted Payloads.



Section 5

# Predictive Analysis (Classification)

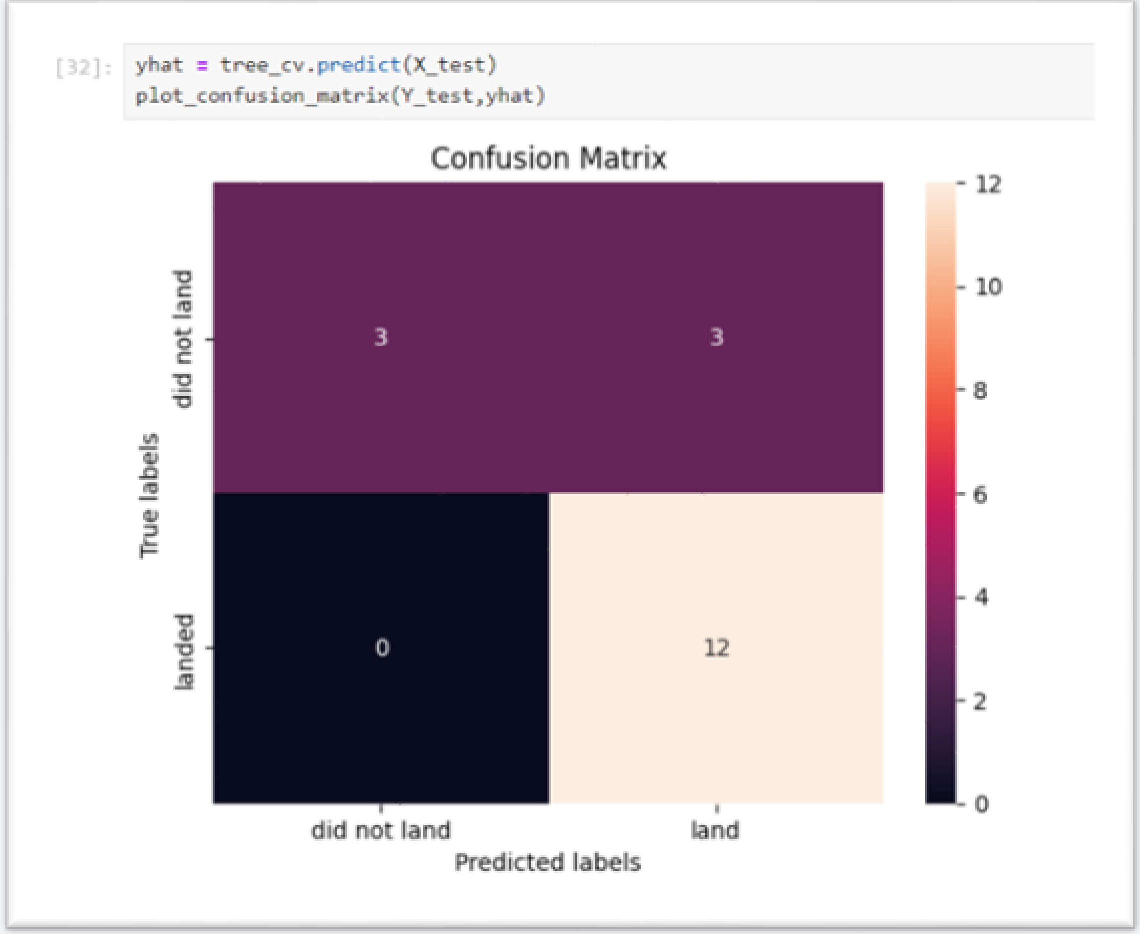
# Classification Accuracy

- Among them, Logistic Regression has slightly better training accuracy and lower risk of overfitting compared to Decision Tree. Decision Tree has the highest training accuracy but lowest test accuracy, indicating overfitting.

| | Logistic Regression | SVC | Decision Tree | KNN Classifier |
|---|---|---|---|---|
| **Train** | 0.846429 | 0.835714 | 0.875000 | 0.848214 |
| **Test** | 0.833333 | 0.833333 | 0.777778 | 0.833333 |

# Confusion Matrix

The confusion matrix for the logistic regression shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.




- **Accuracy**: (TP+TN)/Total = (12+3)/18 = 0.8333
- **Misclassification Rate**: (FP+FN)/Total = (3+0)/18 = 0.1667
- **True Positive Rate**: TP/Actual Positive =12/12 = 1
- **False Positive Rate**: FP/Actual Negative = 3/6 = 2
- **True Negative Rate** : TN/Actual Negative = 3/6 = 2
- **Precision**: TP/Predicted Positive = 12/15 = 0.8
- **Prevalence**: Actual Positive/Total = 12/18 = 0.6667

# Conclusions

- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best Success Rate.
- Success rates for SpaceX launches has been increasing relatively with time.
- Logistic Regression is best suited Machine Learning Model for the given data set.

# Appendix

- All Python code snippets, SQL queries, charts, Notebook outputs, and data sets that are created during this project is listed on my GitHub repository.

[Winning Space Race]

# Thank you!