IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Darshan Kapadia
October 17, 2022

# Outline

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

# Executive Summary

- The following methodologies are applied to analyze the SpaceX data:
  - Data is collection using data collection API with web scrapping and SpaceX API,
  - Exploratory Data Analysis (EDA) which includes data wrangling, data visualization, and interactive visual analytics,
  - Predictive analysis using Machine Learning techniques.
- Summary of all results
  - It is possible to collect the data from public sources, and with suitable data processing, it is ready to analyze,
  - Exploratory Data Analysis (EDA) allows to investigate the data set and summarize their main characteristics,
  - Various Machine Learning techniques allow to get the best models with the high accuracy.

# Introduction

- The objective of this project is to predict how an alternate company wants to bid against SpaceX for a rocket launch.
- Problems to find answers:
  - Estimate the total cost of launch by predicting successful landings of first stage of Falcon 9 rockets,
  - Find the best place to launch the rockets.

Section 1

# Methodology

# Methodology

## Executive Summary:

- Data Collection Methodology:
  - Data is collected from 2-sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets, https://api.spacexdata.com/v4/launchpads, https://api.spacexdata.com/v4/payloads, https://api.spacexdata.com/v4/cores)
    - Web Scrapping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform Data Wrangling:
  - Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

# Methodology

## Executive Summary:

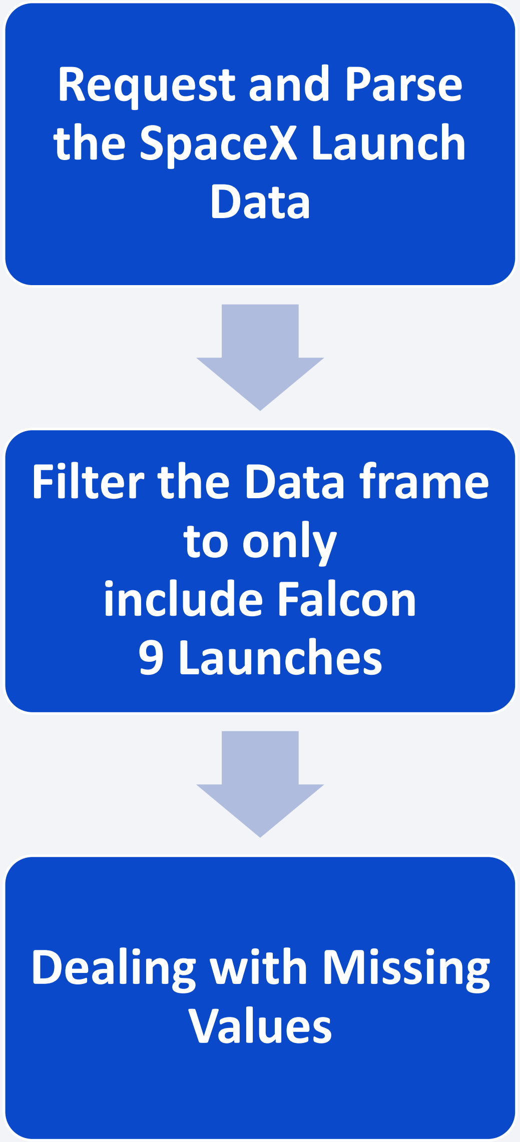
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were standardized,
  - Divide the data set into training and test data sets, and evaluate the data by four different classification models,
  - Evaluate the accuracy of each model using different combinations of parameters.

# Data Collection

- Data is collected from 2-sources:
  - SpaceX API (https://api.spacexdata.com/v4/rockets, https://api.spacexdata.com/v4/launchpads, https://api.spacexdata.com/v4/payloads, https://api.spacexdata.com/v4/cores)
  - Web scraping to collect Falcon 9 historical launch records from a Wikipedia page title “List of Falcon 9 and Falcon Heavy launches”. (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

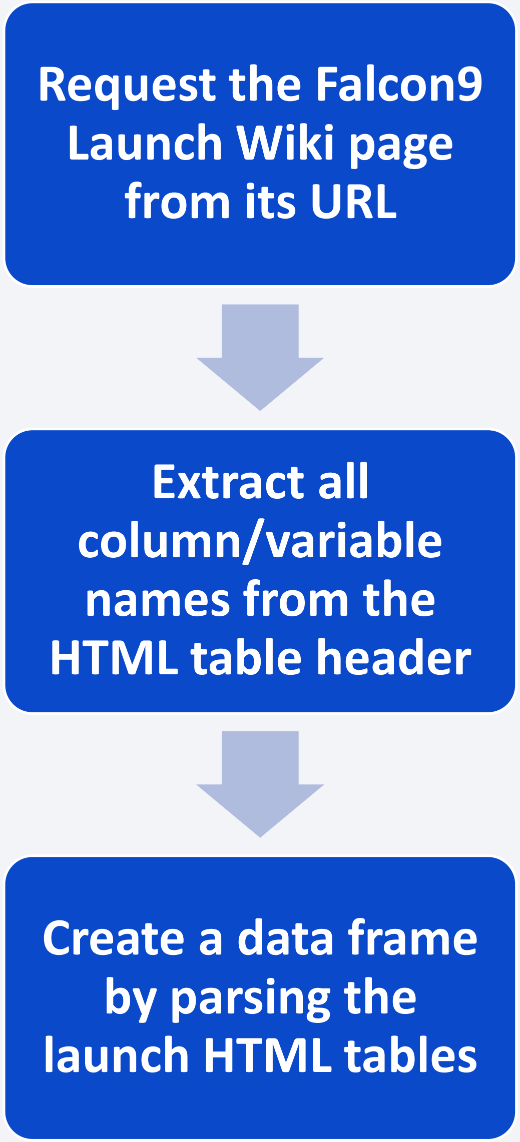
# Data Collection – SpaceX API

- Processing of data using SpaceX API is shown in the flowchart here.
  - Request and Parse the SpaceX Launch Data using GET request to make the requested JSON results more accurate,
  - Filter the data frame by removing Falcon 1 launches,
  - Dealing the missing values by replacing the mean value.



# Data Collection - Scraping

- Processing of data using web scraping is shown in the flowchart here.
    - Request data from its Wikipedia page using its URL,
    - Collect all relevant column names from the HTML table header,
    - Create a data frame by parsing the launch HTML tables.



# Data Wrangling

- Exploratory Data Analysis (EDA) is performed to find some patterns in the data,
- Also, determines what would be the label for training supervised models.

**Calculate the number of launches per site**

↓

**Calculate the number and occurrence of each orbit and mission outcome per orbit type**

↓

**Create a landing outcome label from outcome column**

# EDA with Data Visualization

- Several scatterplots and bar charts are plotted to know the relationship among various features of the data. Few plots are:
    - Flight Number vs. Pay Load Mass, Flight Number vs. Launch Site, Launch Site vs. Pay Load Mass, Obrit vs. Class, Flight Number vs. Orbit, and Pay Load Mass vs. Orbit.

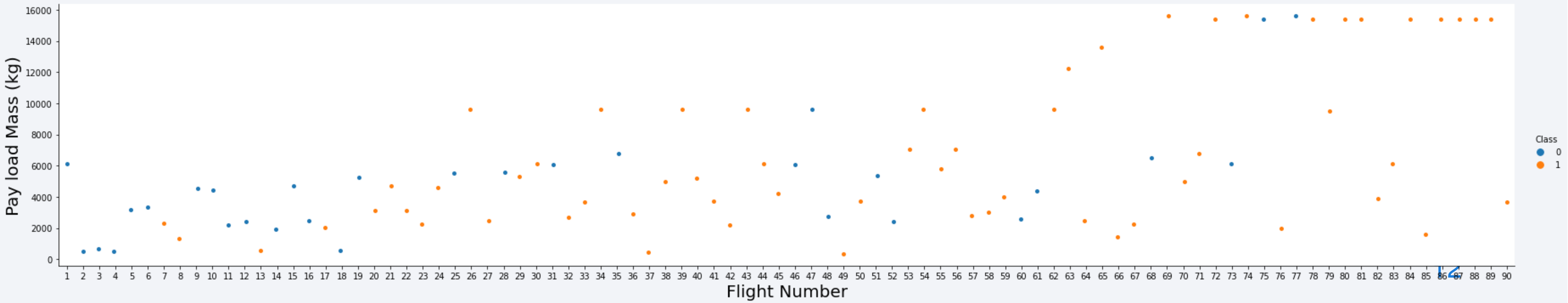



Fig. 1: Scatter Plot of Flight Number and Pay Load Mass

# EDA with SQL

- The following SQL queries performed:
  - Display the names of the unique launch sites in the space mission,
  - Display 5 records where launch sites begin with the string 'CCA',
  - 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 date when the first successful landing outcome in ground pad was achieved,
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
  - List the total number of successful and failure mission outcomes,
  - List the names of the booster versions which have carried the maximum payload mass,
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015,
  - Rank 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.
-

# Build an Interactive Map with Folium

- Different markers, circles, lines, and marker clusters are used with Folium Maps.
    - Markers are used to indicate points such launch sites,
    - Circle are used to highlight the areas around specific coordinates,
    - Lines are used to get the distances between two coordinates,
    - Marker clusters are used to indicate groups of events in each coordinate.

# Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

# Predictive Analysis (Classification)

- Perform Exploratory Data Analysis and determine training labels to find best hyperparameter for Logistic Regression, Support Vector Machine, Tree Classifier, and k-Nearest Neighbors.

# Results

- Exploratory Data Analysis (EDA) results:
  - Space X uses 4 different launch sites,
  - Launch Site CCAFS SLC 40 has the high success rate compare to others,
  - Orbit ES-L1, GEO, HEO, and SSO has the 100% success rate,
  - Orbit SSO has the highest rate of success for less than 4000 kg of payload, and VLEO has the highest rate of success for over 14000 kg of payload,
  - From year 2013 to 2017, consecutive rate of success is shown by the graph,
  - The number of landing outcomes became as better as years passed.

# Results

- Interactive Analytics results:
  - Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around,
  - Most launches happens at east cost launch sites.

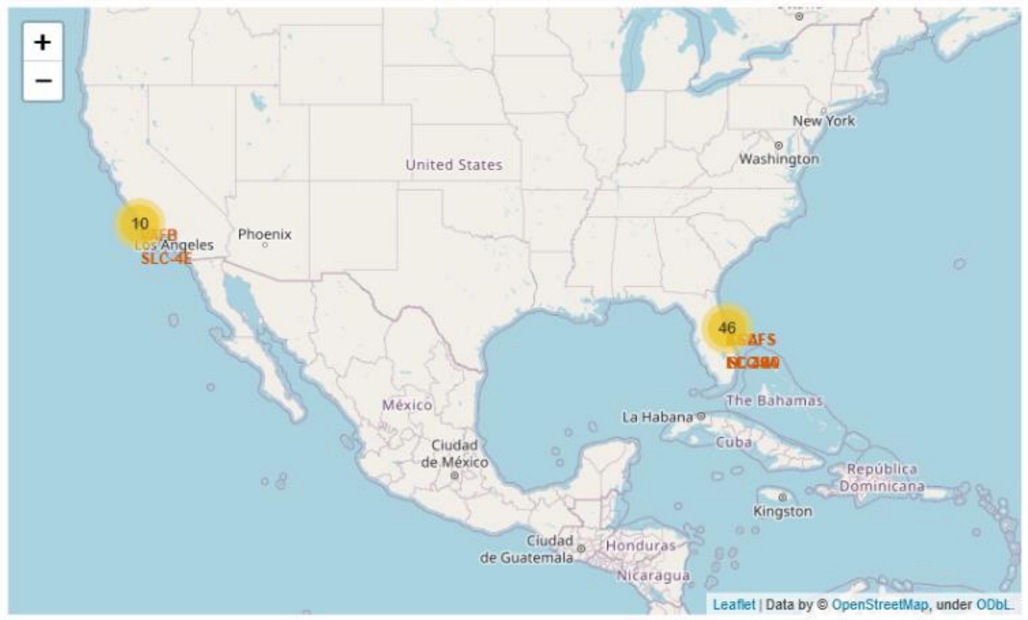



Fig. 2: Interactive Analytics show success rate for different launch sites on both coasts

# Results

- Predictive Analytics results:
  - Using different predictive analytics, Decision Tree Classifier shows the highest accuracy in both data sets.
  - It has accuracy of 87% and accuracy for test data is of 94%.

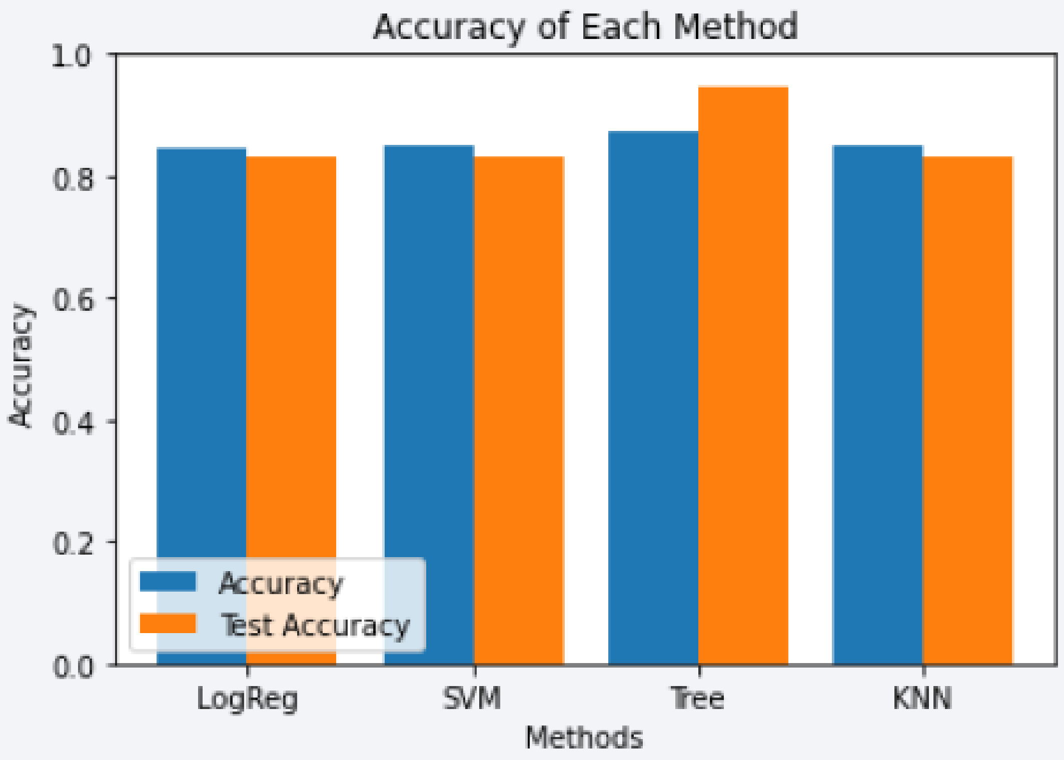



Fig. 3: Accuracy of Different Methods

Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

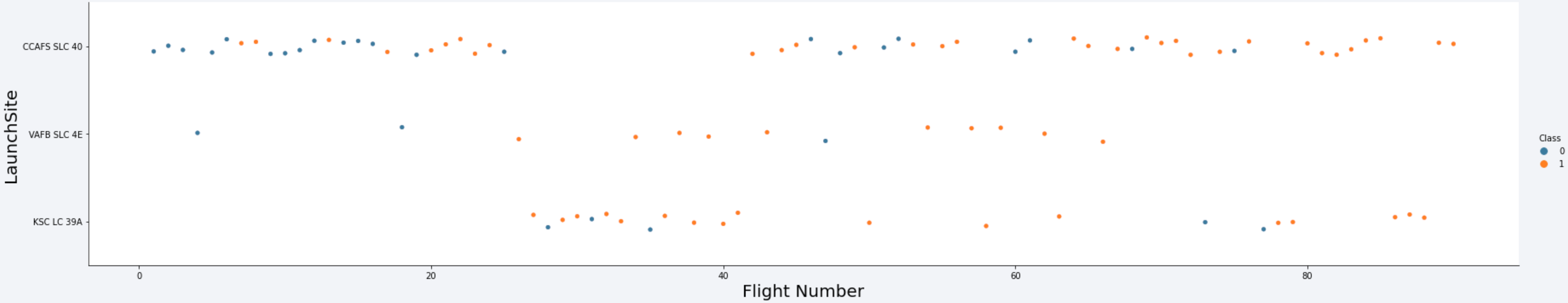



Fig. 4: Scatter Plot of Flight Number vs. Launch Site

- From the above plot, it can see that Launch Site CCAFS SLC 40 has the high success rate compare to other twos.
- Over the time, rate of success is also increased.

# Payload vs. Launch Site

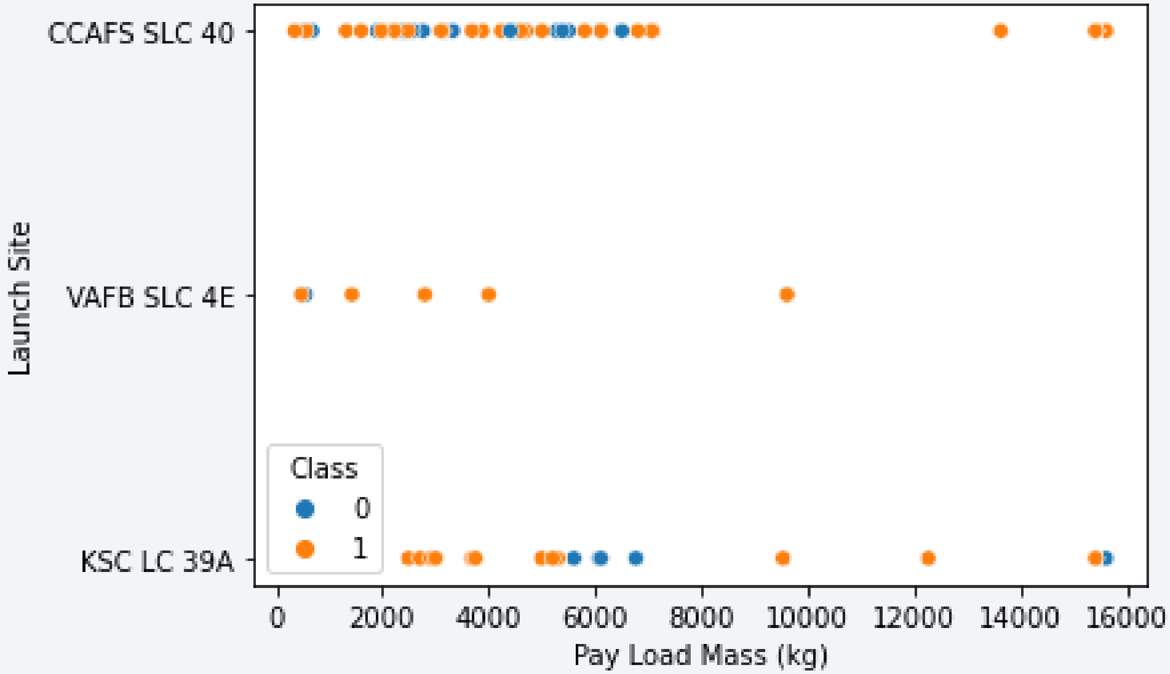



Fig. 5: Scatter Plot of Pay Load Mass (kg) vs. Launch Site

- From the above plot, it can see that CCAFS SLC 40 has high success rate for less than 8000 kg of payload mass.
- Launch site VAFB SLC 4E has high success rate for pay load up to 10000 kg.

# Success Rate vs. Orbit Type

- From the bar chart, it is observed that ES-L1, GEO, HEO, and SSO has the 100% success rate,
- Orbit SO has the lowest rate of success.

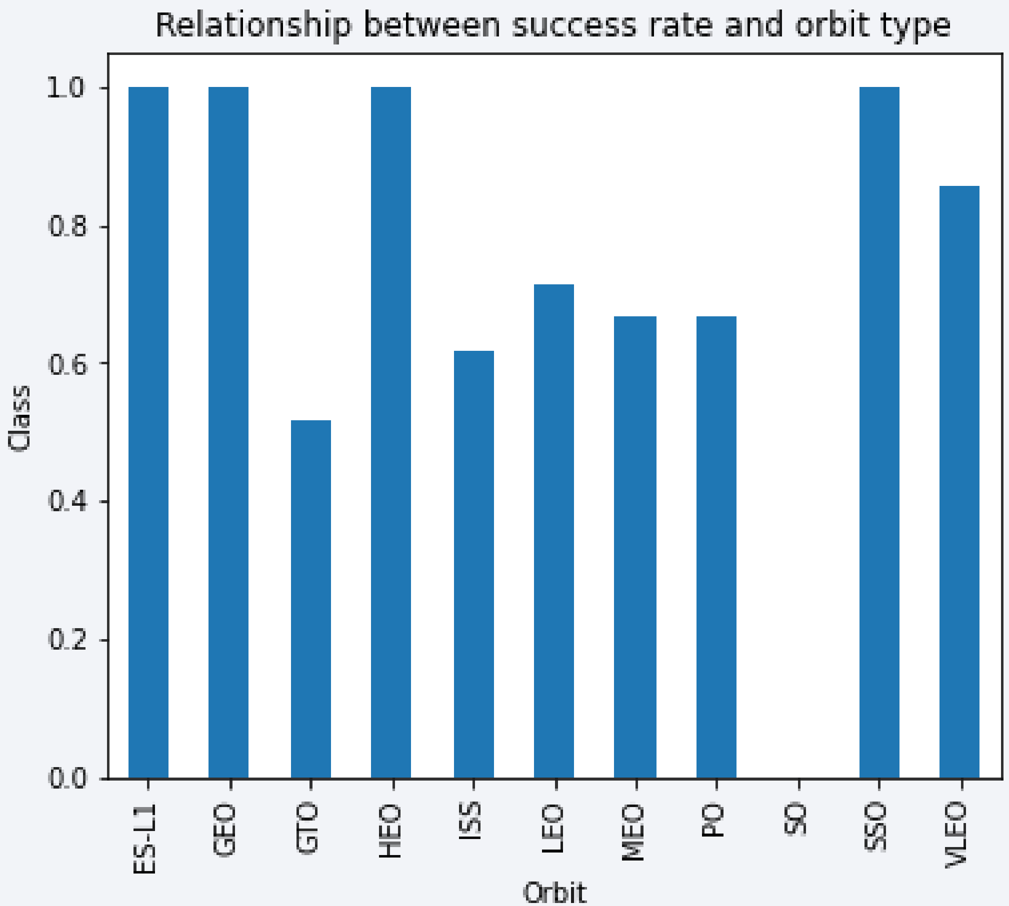



Fig. 6: Bar Chart of Success Rate vs. Orbit Type

# Flight Number vs. Orbit Type

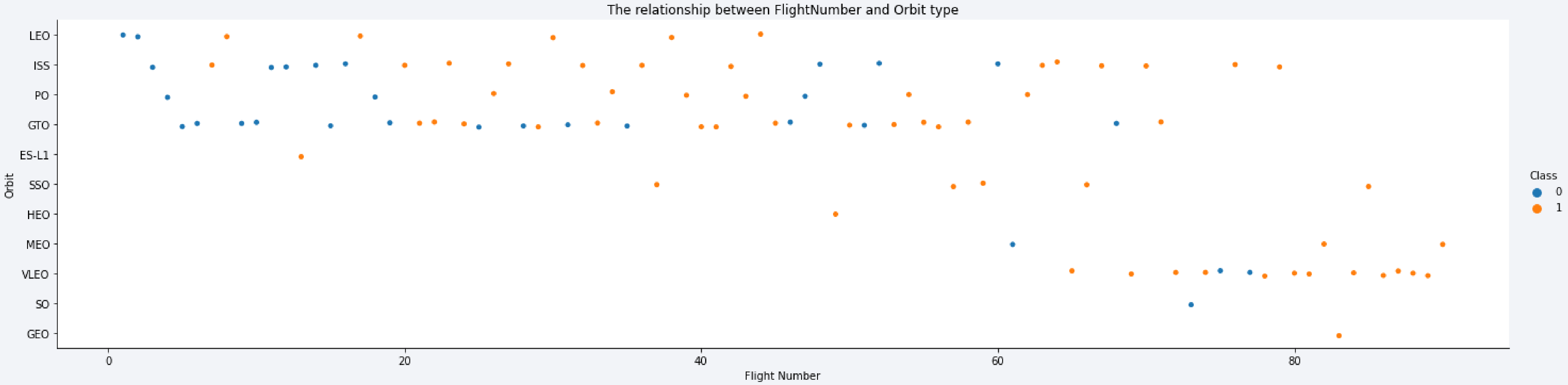



Fig. 7: Scatter Plot of Flight Number vs. Orbit Type

- SSO and VLEO has the highest rate of success followed by LEO, and PO,
- Over the time, rate of success has increased.

# Payload vs. Orbit Type

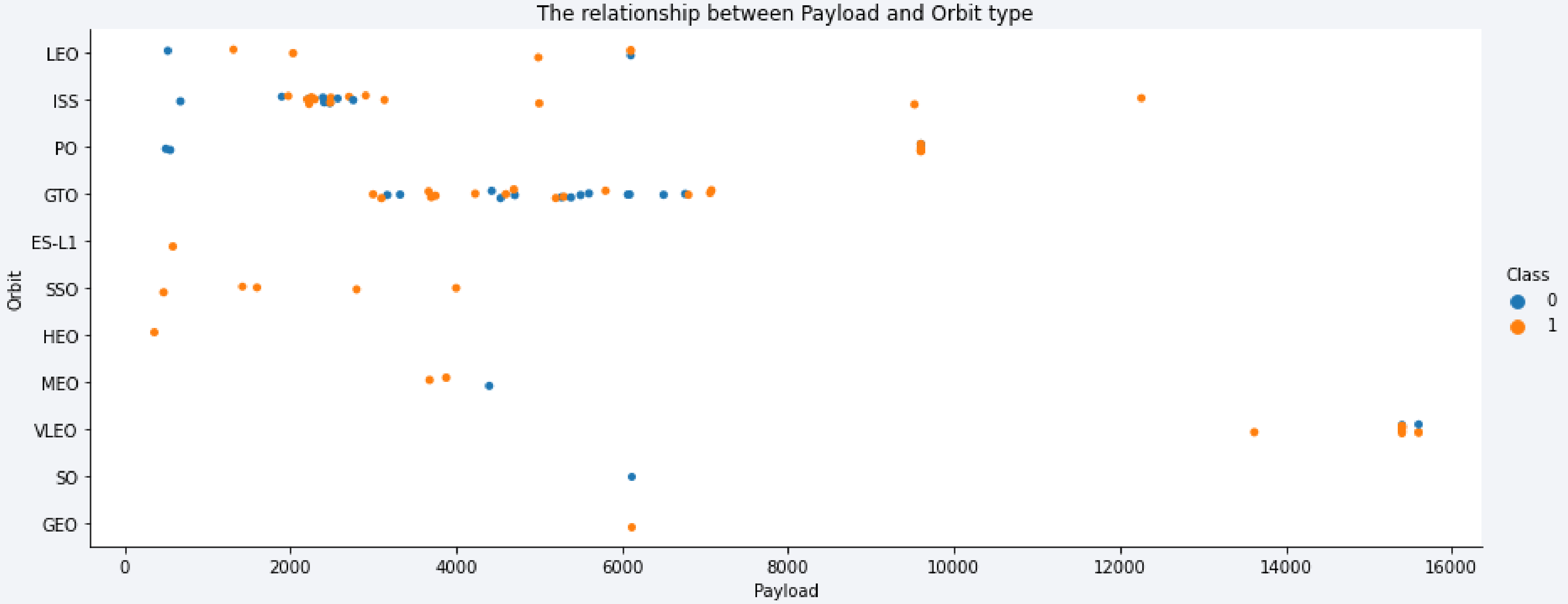



Fig. 8: Scatter Plot of Pay Load vs. Orbit Type

- SSO has the highest rate of success for less than 4000 kg of payload, and VLEO has the highest rate of success for over 14000 kg of payload,
- There is a mixed relation for ISS and GTO.

# Launch Success Yearly Trend

- Over the time, success rate is increased,
- From year 2013 to 2017, consecutive rate of success is shown by the graph,
- From year 2017-18 and 2019-20, few failures are observed.

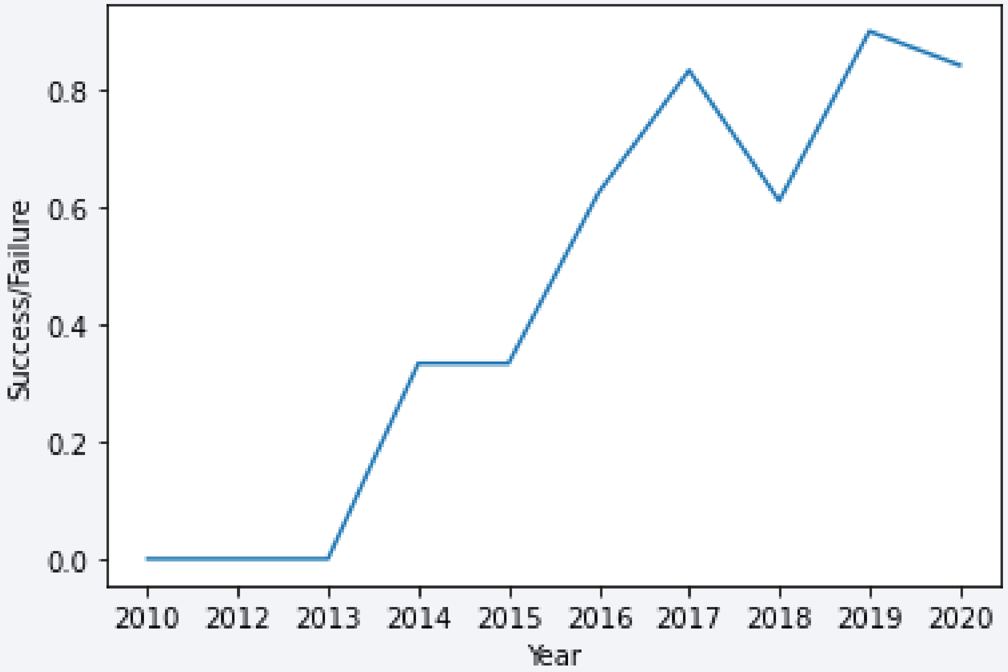



Fig. 9: Year vs Success/Failure

# All Launch Site Names

- All launch site names are as per the table-1,
- Distinct launch sites are obtained by selecting "Launch Site" by data frame.

Table – 1: Launch Site Names

| Launch Site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

# Launch Site Names Begin with 'CCA'

- All launch site names begin with ‘CCA’ are as per the table-2,
- Launch sites names with ‘CCA’ are obtained by selecting “Launch Site” by data frame.

Table – 2: Launch Site Names Begin with ‘CCA’

| Date | Time | Booster Version | Launch Site | Pay Load | Payload Mass (kg) | Orbit | Customer | Mission Outcome | Landing Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-04-06 | 18:45:00 | F9 v1.0 B0003 | **CCA**FS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-08-12 | 15:43:00 | F9 v1.0 B0004 | **CCA**FS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2012-08-10 | 00:35:00 | F9 v1.0 B0006 | **CCA**FS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-01-03 | 15:10:00 | F9 v1.0 B0007 | **CCA**FS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-12 | 22:41:00 | F9 v1.1 | **CCA**FS LC-40 | SES-8 | 3170 | GTO | SES | Success | No attempt |

# Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS) is shown in the table-3,
- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Table – 3: Total Payload Mass

| Total Payload Mass (kg) |
| --- |
| 22007 |

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is shown in the table-4,
- Average payload mass by F9 v1.1 is obtained by averaging all payloads whose codes contain ‘F9 v1.1’, which corresponds to Booster Version.

Table – 4: Average Payload Mass by F9 v1.1

| Average Payload (kg) |
|---|
| 3676 |

# First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad is achieved and shown in the table-5,
- First successful ground landing date is obtained by minimizing the date whose code contains Success (ground pad), which corresponds to Landing Outcome.

Table – 5: First Successful Ground Landing Date

| Min. Date |
| --- |
| 2017-01-05 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 is achieved and shown in the table-6,
- Successful drone ship landing with payload between 4000 and 6000 kg is obtained by selecting success (drone ship), which corresponds to Landing Outcome.
- Also, payload mass is selected between 4000 and 6000 kg.

Table – 6: Successful Drone Ship Landing with Payload between 4000 and 6000

| Booster Version |
|---|
| F9 FT B1022 |
| F9 FT B1031.2 |

# Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is achieved and shown in the table-7,
- Grouping mission outcomes and counting records for each group led us to the summary above.

Table – 7: Total Number of Successful and Failure Mission Outcomes

| Mission Outcomes | Frequency |
|---:|:---:|
| Success | 44 |
| Success (payload status unclear) | 1 |
| **Total** | **45** |

# Boosters Carried Maximum Payload

- The names of the booster versions which have carried the maximum payload mass is achieved and shown in the table-8,
- There are five booster versions which are carried maximum payload.

Table – 8: Boosters Carried Maximum Payload

| Booster Version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1060.2 |
| F9 B5 B1058.3 |

# 2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 is achieved and shown in the table-9,
- Only one outcome is achieved by the query here.

Table – 9: The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

| Booster Version | Launch Site | Landing Outcome | Date |
| --- | --- | --- | --- |
| F9 v1.1 B1012 | CCAFS LC-40 | Failure (Drone Ship) | 2015-10-01 |

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

- Rank 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 is achieved and shown in the table-10,

Table – 10: Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

| Landing Outcome | Frequency |
|---|---|
| No attempt | 7 |
| Failure (drone ship) | 2 |
| Success (drone ship) | 2 |
| Success (ground pad) | 2 |
| Controlled (ocean) | 1 |
| Failure (parachute) | 1 |
| Success | 0 |
| **Total** | **15** |

Section 3

# Launch Sites Proximities Analysis

# Launch Sites

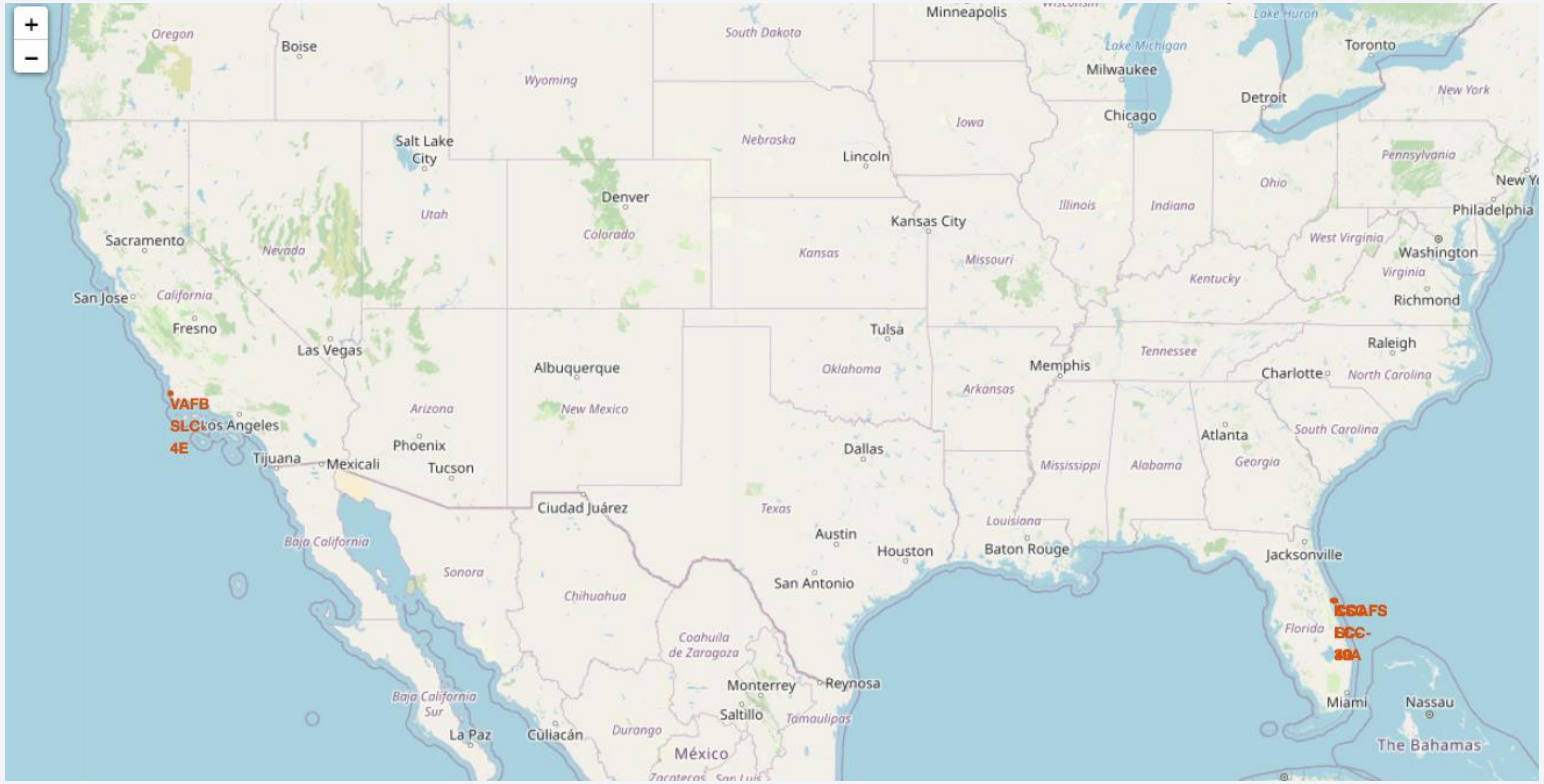



Fig. 10: All launch sites

- All launch sites are located on east-coast or on west-coast near ocean.

# Launch Site by Success Rate

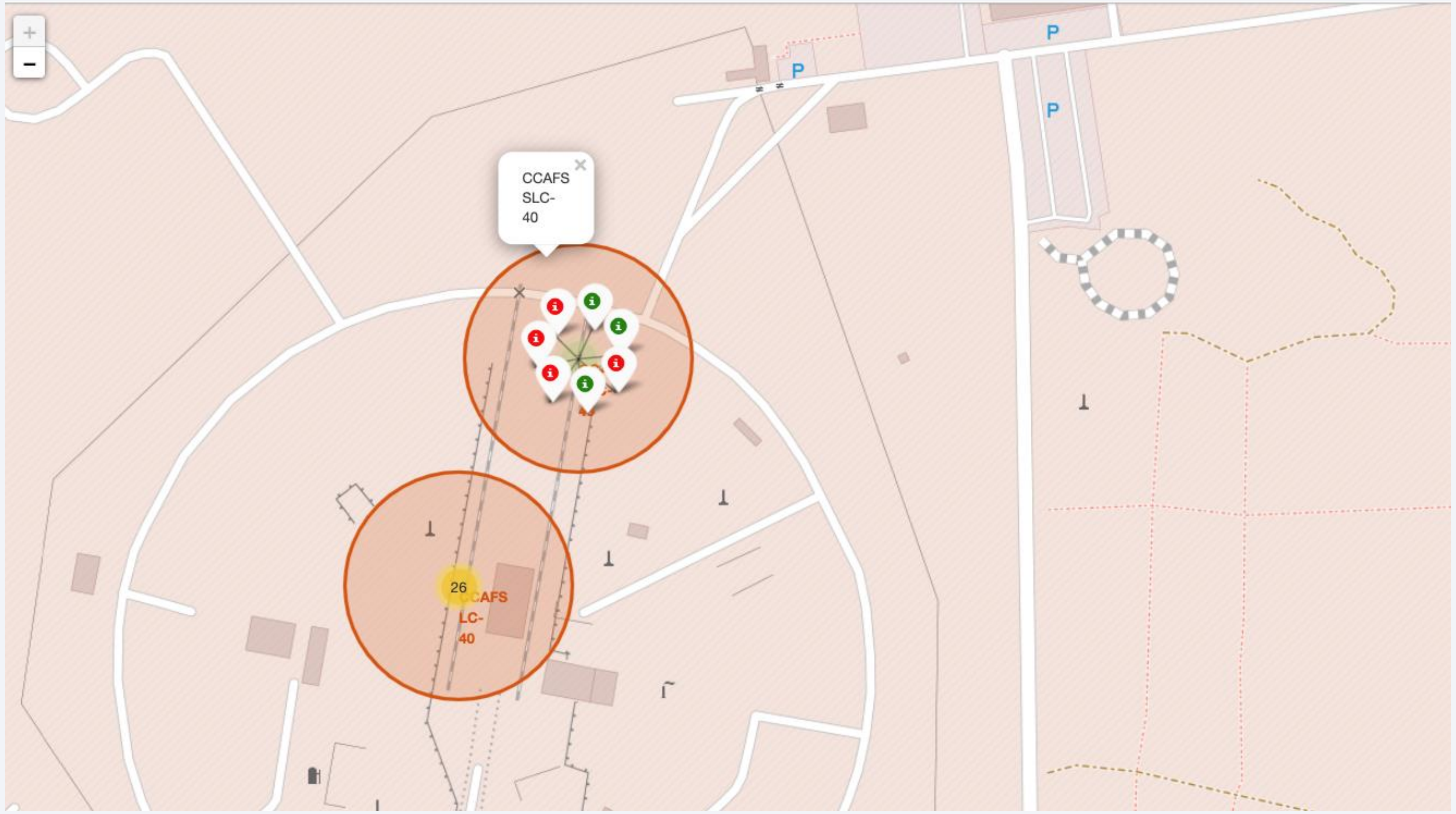



Fig. 11: Launch Site by Success Rate

- Launch site CCAFS SLC-40 is shown in the above figure with its success rate. Red color shows failure and green color shows success.

# Map with Highways

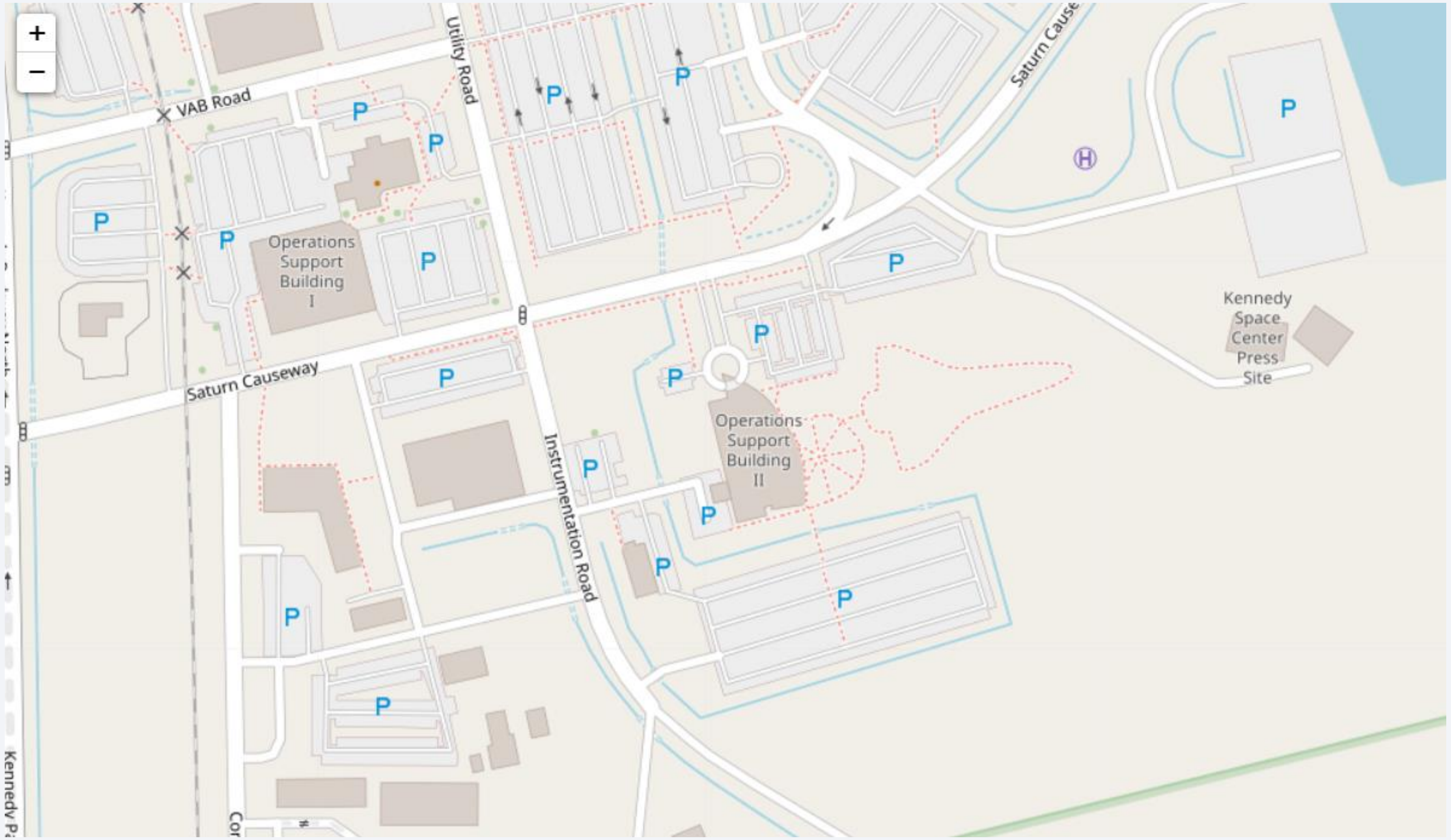



Fig. 12: Map with Highways

- Above diagram is shown map with highway label.

Section 4

# Build a Dashboard with Plotly Dash

# SpaceX Launch Records Dashboard

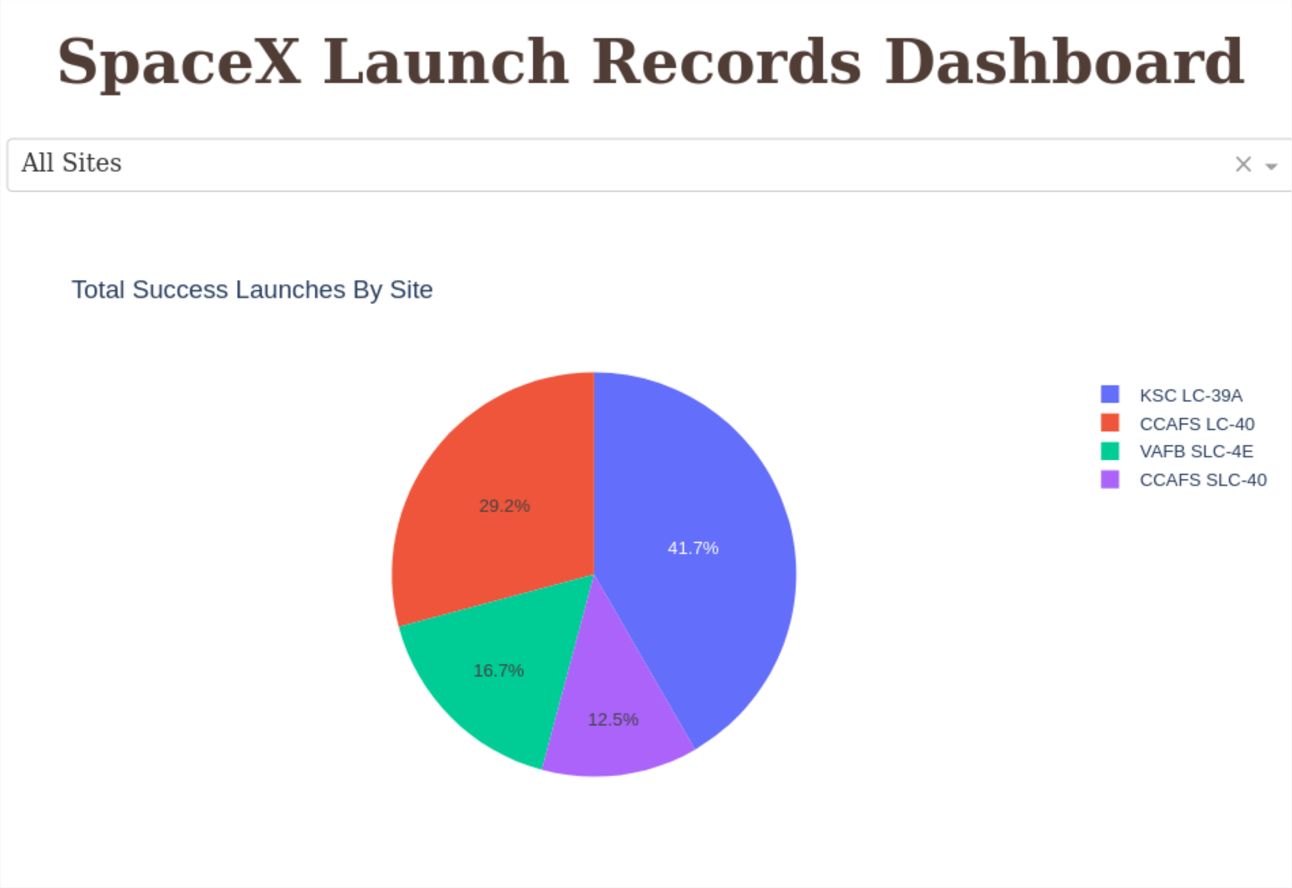



Fig. 13: SpaceX Launch Records Dashboard

- The pie chart shows the total successes lunches by sites.

# Launch Success Ratio for KSC LC-39A

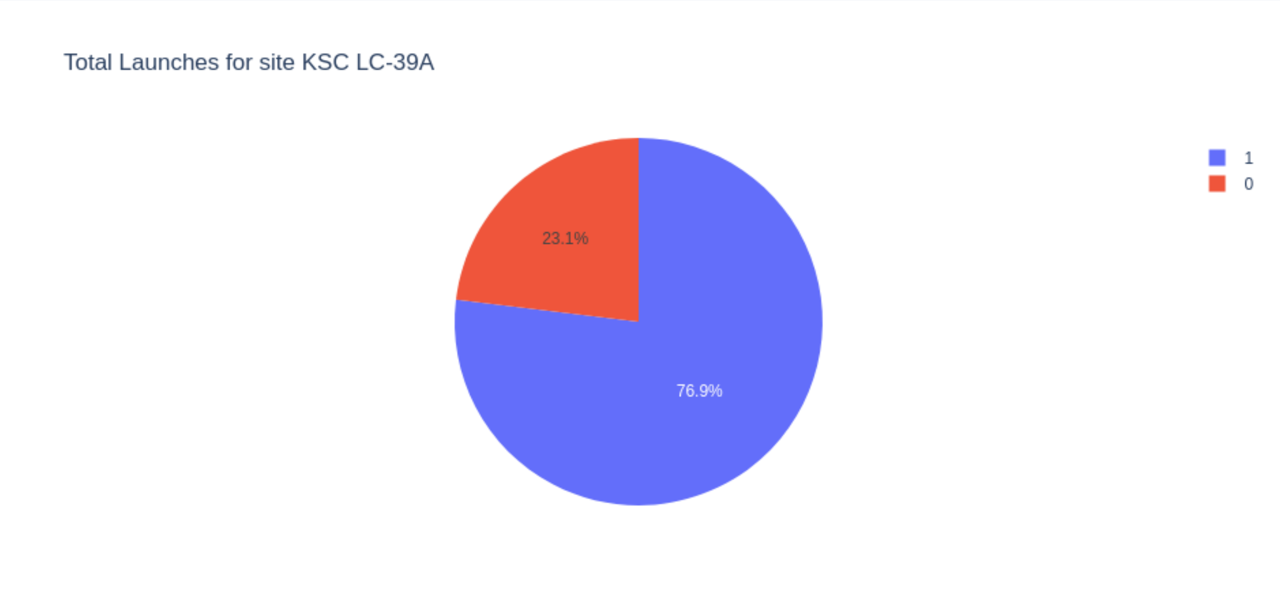



Fig. 14: Launch Success Ratio for KSC LC-39A

- The success ratio of KSC LC-39A is 76.9%.

# Payload vs. Launch Outcome

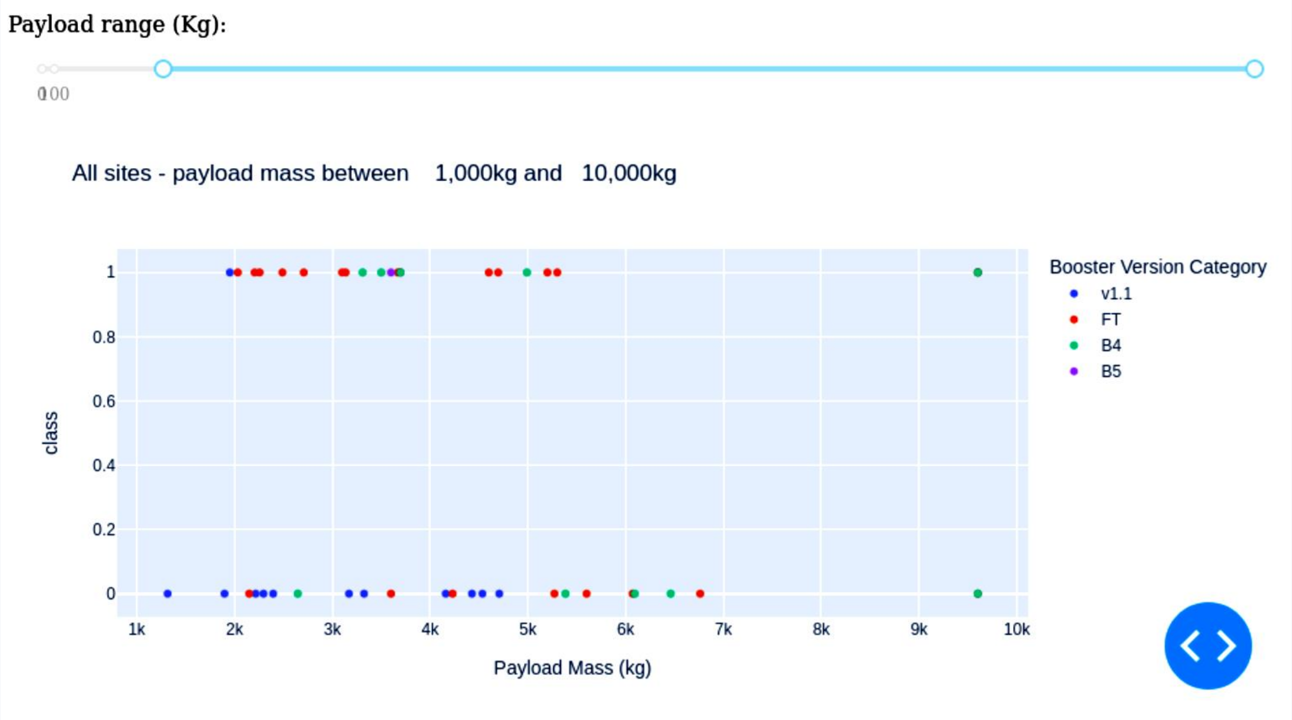



Fig. 15: Payload vs. Launch Outcome

- FT boosters have the highest rate of success among all four booster version.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside with the help of bar chart,
- As per the chart, Decision Tree Classifier has the highest accuracies in both categories.

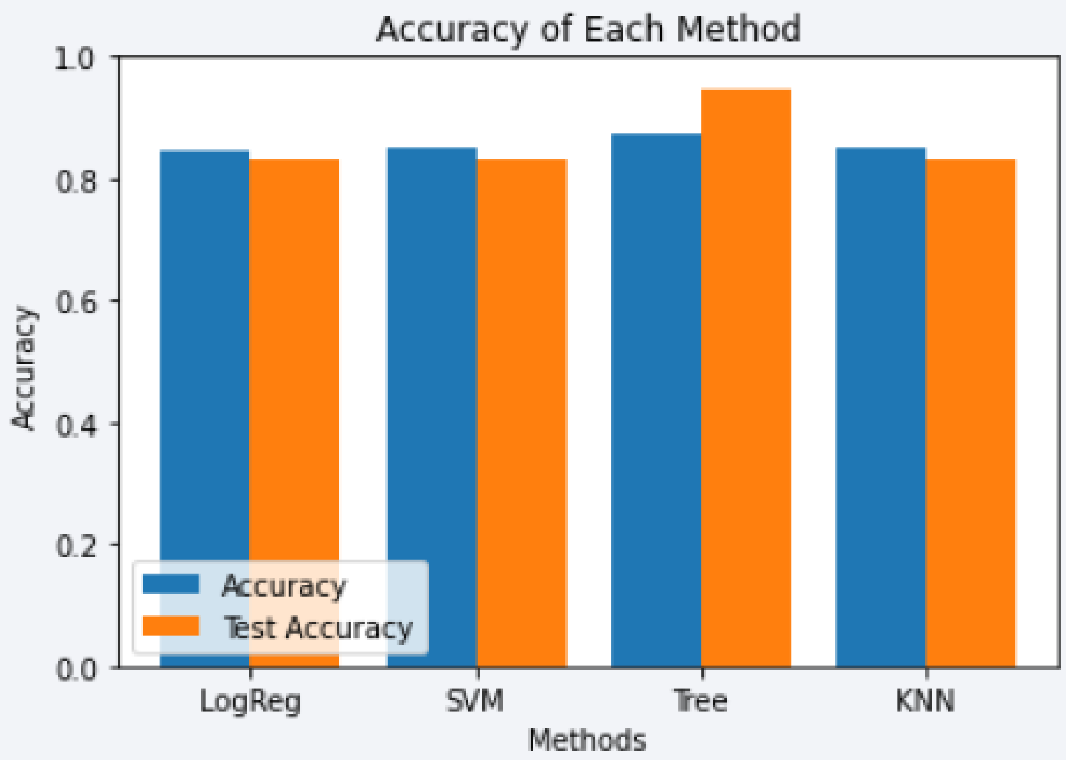



Fig. 16: Accuracy of each method

# Confusion Matrix

- Confusion Matrix is a performance measurement for machine learning classification.
- Confusion Matrix of the Decision Tree Classifier is as shown in fig.


Fig. 17: Confusion Matrix of Decision Tree Classifier

# Conclusions

- Different data sources were analyzed, refining conclusions along the process,
- The best launch site is KSC LC-39A,
- Launches above 7,000 kg are less risky,
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets,
- Decision Tree Classifier can be used to predict successful landings and increase profits.

# Thank you!