IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Devang Ghela
July 3, 2023

# Outline

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

# Executive Summary

For this project, SpaceX Falcon 9 Launch data was collected using the SpaceX API and by webscraping the Wikipedia page for Falcon 9 and Falcon Heavy. This data was lightly processed and was sent to be explored and analyzed by creating various visualizations for the data. Various SQL queries were also ran for the exploratory data analysis. From there, the data was used to create a map with the different launch sites and their respective rocket launches. Next, a dashboard app was also built that allows for the data to be filtered to create visualizations very quickly. Finally, four different classification methods were tested for predicting launches.

From all of this, a couple of conclusions were drawn: launch site KSC LC-39A yields the highest successful landing rate, booster version FT, when launched with a payload mass between 2000 and 4000 kilograms yields a high successful landing rate, and finally, all classification models yielded the same accuracy of about 0.833 meaning there is no “best” model.

# Introduction

There are many companies providing the service of space launches. These launches can cost upwards of 165 million dollars each, but SpaceX's Falcon 9 rockets have been shown to cost about 62 million dollars each. This competitive cost is based on the idea that they reuse the first-stage of their rockets. However, these savings could be lost if the first-stage does not land successfully.

This project aims to answer a few questions like:

- Which launch site yields a high successful landing rate?
- Which booster version has the highest successful landing rate?
- What payload mass range has the highest successful landing rate?
- Can it be predicted whether a landing will be successful?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

- Data sets were collected from two sources:
  - **Source 1:** SpaceX API - https://api.spacexdata.com/
    - Data is obtained through various endpoints from SpaceX API using HTTP GET requests
    - Data is converted to a Pandas DataFrame and filtered to the Falcon 9 launch data
    - Data is wrangled to replace null values of numerical columns with the mean
  - **Source 2:** Wikipedia Page for Falcon 9 / Heavy launches - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Data is obtained through multiple tables from the Wikipedia page using HTTP GET requests
    - Data is parsed using BeautifulSoup object and the .find_all() function on various HTML tags for tables, table headers, and data cells.
    - Data is converted to a Pandas DataFrame

# Data Collection – SpaceX API

| | |
|---|---|
| Source the data | • Data is gathered using the SpaceX API<br>• https://api.spacexdata.com |
| Launch Data | • Gather launch data from ‘launches’ endpoint using the API<br>• Many columns in from this endpoint are just IDs so create helper functions |
| Use Helper Functions | • The helper functions use the IDs to obtain corresponding data from other endpoints from the API<br>• Other endpoints include: ‘rockets’, ‘launchpads’, ‘payloads’, and ‘cores’ |
| Create DataFrame | • Create dictionary using all of the data obtained from the various endpoints<br>• Convert the dictionary to DataFrame using pd.DataFrame() |
| Filter | • DataFrame contains data for Falcon 1 launches, so filter the data to exclusively Falcon 9 data<br>• Data is ready for wrangling |

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%201/jupyter-labs-spacex-data-collection-api.ipynb

# Data Collection - Scraping

**Source the data**

- Data is gathered from HTML Tables on Wikipedia page using HTTP GET method
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

**Find Tables**

- Create and use a BeautifulSoup object to parse the content of the response
- Find the appropriate tables using the object using soup.find_all(‘table’)

**Extract Column names**

- Use the BeautifulSoup object to obtain all of the column/variable names from tables using soup.find_all(‘th’)
- Create a dictionary from the list of names

**Extract Data from Tables**

- Loop through tables, loop through rows using soup.find_all(‘tr’), and use soup.find_all(‘td’) to append data to dictionary

**Create DataFrame**

- Use pd.DataFrame and pass dictionary as the argument to convert it to a DataFrame

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%201/jupyter-labs-webscraping.ipynb

# Data Wrangling

**Find Null Values**

- Use .issnull().sum() on the DataFrame
- This returns a series of the total number of null values for each column

**Replace Null Values**

- Calculate the mean of the columns with null values using .mean()
- Use .replace() to replace the null values with the mean

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%201/jupyter-labs-spacex-data-collection-api.ipynb

# EDA with Data Visualization

- Scatter plots were used for the reason of visualizing the following relationships:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit
  - Payload Mass vs. Orbit
- A bar chart was used for the reason of visualizing the success rate of each orbit type
- A line chart was used for the reason of visualizing the yearly trend of launch success rate

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%202/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

# EDA with SQL

- SQL queries were created to do the following tasks:
  - Display the names of the unique launch sites
  - Display 5 records of launch sites that begin with the string 'CCA'
  - Display the total payload mass carried by the boosters launched by NASA (CRS)
  - Display the average payload mass carried by the booster version F9 v1.1
  - List the date when the first successful ground pad landing outcome occurred
  - List the names of the boosters with payload mass between 4000kg and 6000kg with success drone ship landings
  - List the total number of successful and failed mission outcomes
  - List the names of booster versions that carried the maximum payload mass
  - List the month, booster version, and launch site of missions with failed landing outcomes for the months in 2015
  - Rank the count of each type of landing outcome between 2016-06-04 and 2017-03-20 in descending order

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

- In the Interactive Map, the following map objects were used
  - Circles were used to show the general areas of each launch site on the map
  - Markers were used to display the names of the launch sites on the map, for showing each launch at its corresponding site, and for showing the numerical distance from a launch site to a nearby point of interest
  - MarkerClusters were used to bunch the launch markers to the corresponding launch site on the map
  - Lines were used to visualize the distance from a launch site to a nearby point of interest

https://github.com/gheladevang/SpaceXCapstone/blob/126e412bade235ee83bfb7b730c89772bf647939/Module%203/lab_jupyter_launch_site_location.jupyterlite.ipynb

# Build a Dashboard with Plotly Dash

- In the dashboard, the following plots were added:
  - Pie chart
    - To show the success rate of each launch site
    - To show which site has the highest count of successful launches
  - Scatter plot to show the correlation between Payload Mass and Success
- There is a dropdown interaction which filters the data for both plots by the Launch Site
- There is a slider interaction which filters the data for the scatter plot to a specified range for the Payload Mass

https://github.com/gheladevang/SpaceXCapstone/blob/40f04ec152d3b90c5e60bae444f4a66cd8a77a60/Module%203/spacex_dash_app.py

# Predictive Analysis (Classification)

For the predictive analysis, four different classification algorithms were used. They are Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors Classifier. Each of the models were built and evaluated using the flow shown below:

**Building**

- Data is first split using train_test_split() on the data with test_size=0.2 and random_state=2
- Create a model object for each algorithm

**Improvement**

- Pass the model object into a GridSearchCV object with a dictionary of the parameters to optimize and cv=10
- Fit this on the training data obtained from the split
- GridSearchCV automatically refits the algorithm on the best parameters

**Evaluation**

- Use the .score() function on the test data to obtain the testing score for each model
- Plot a confusion matrix to visualize this score

https://github.com/gheladevang/SpaceXCapstone/blob/40f04ec152d3b90c5e60bae444f4a66cd8a77a60/Module%204/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

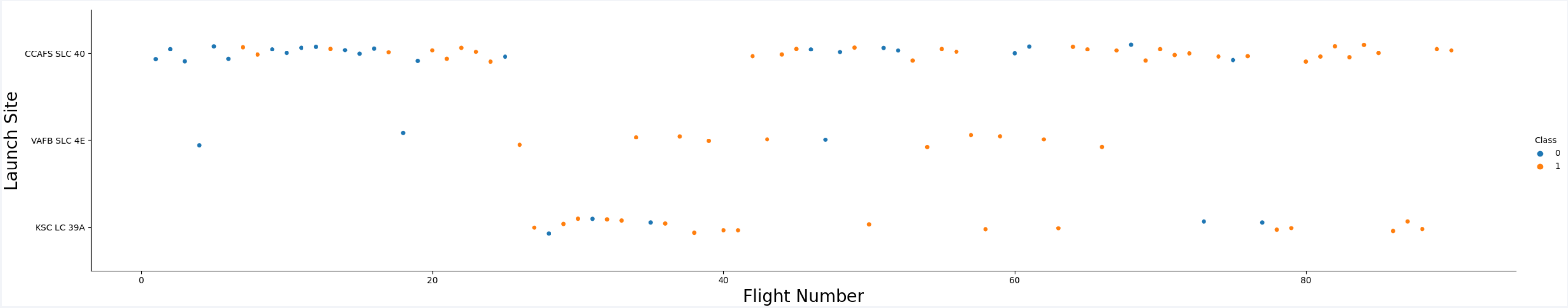
# Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

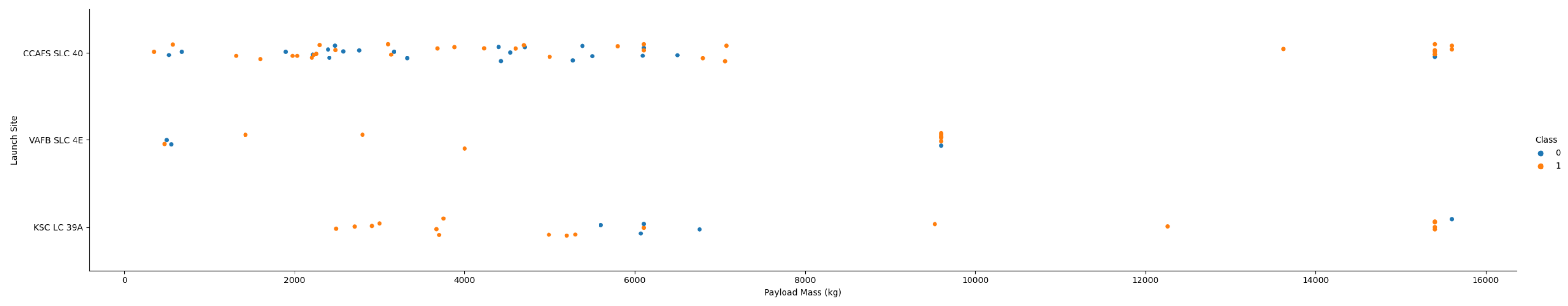
# Insights drawn from EDA

# Flight Number vs. Launch Site



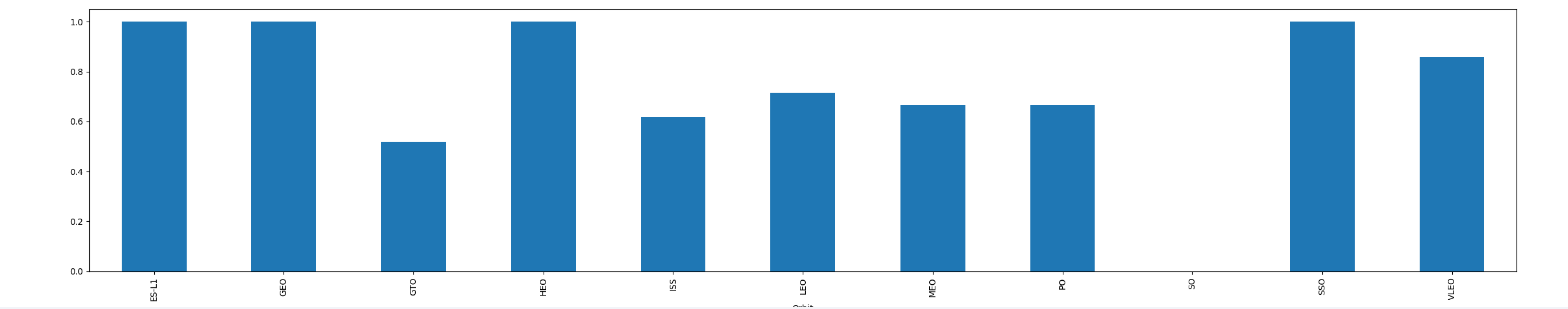

Observation: Each launch site seems to have less failed first-stage-landings as they conduct more launches. Launch site CCAFS SLC 40 also seems to have the most launches conducted of the three sites.

# Payload vs. Launch Site



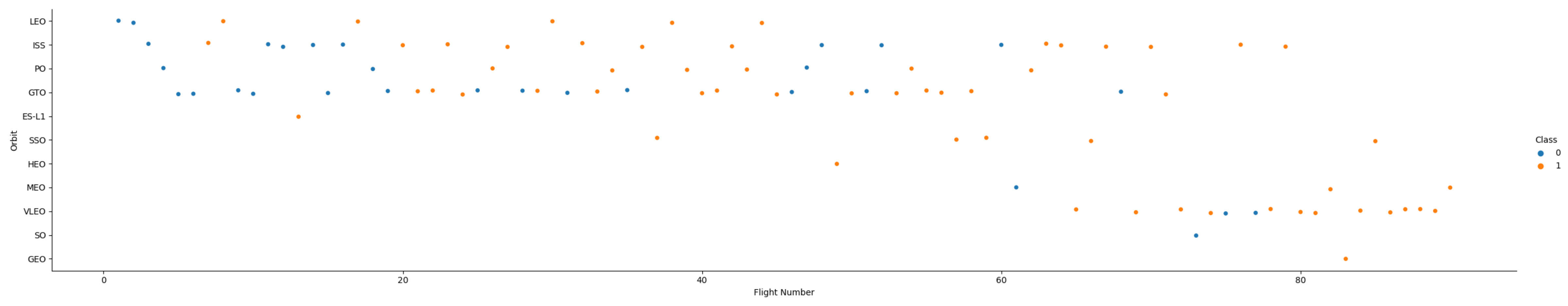

Observation: For each launch site, most launches have payload masses of less than 8000kg, but launch site VAFB SLC 4E does not have any launches above 10000kg.

# Success Rate vs. Orbit Type



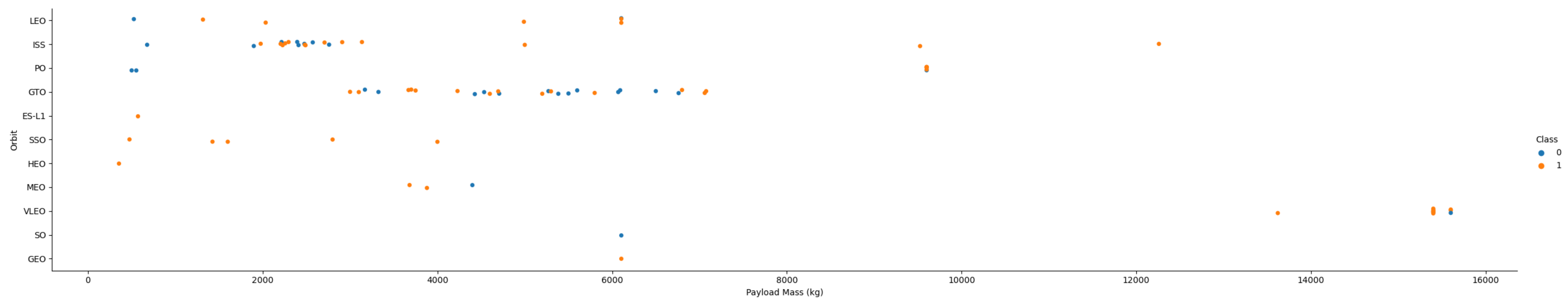

Observation: Orbit types ES-L1, GEO, HEO, SSO, and VLEO all have success rates above 80%, but the first four of those have success rates of 100%.

# Flight Number vs. Orbit Type



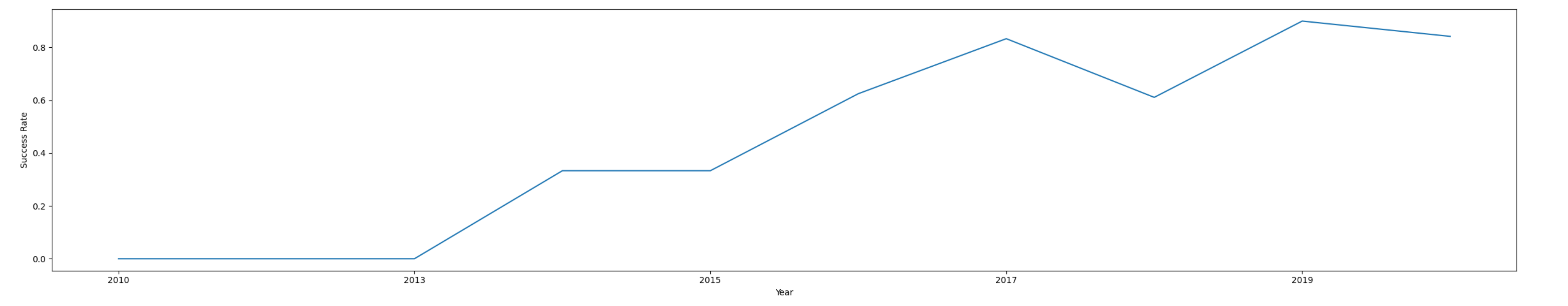

Observation: For Orbit type LEO, every launch after the first two has a successful landing of the first-stage. All other Orbit types (that don't have 100%) have at least one failed landing after the first two launches.

# Payload vs. Orbit Type




Observation: For heavier payloads (>6000kg), LEO, ISS, and PO orbit type have higher success than the GTO orbit type.

# Launch Success Yearly Trend




Observation: Landing success rate increased from 2013 to 2017, decreased in 2019, increased again in 2019, and decreased in 2020.

# All Launch Site Names

| Launch_Site |
| --- |
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |
| None |

There are 4 unique launch site names

# Launch Site Names Begin with 'CCA'

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 06/04/2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0.0 | LEO | SpaceX | Success | Failure (parachute) |
| 12/08/2010 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0.0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 22/05/2012 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525.0 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 10/08/2012 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 03/01/2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |

5 records with launch sites with CCA in the name

# Total Payload Mass

| SUM(PAYLOAD_MASS__KG_) |
| --- |
| 45596.0 |

The total payload mass carried by the boosters launched by NASA (CRS) is 45596 kg

# Average Payload Mass by F9 v1.1

| AVG(PAYLOAD_MASS__KG_) |
| --- |
| 2534.6666666666665 |

The average payload mass carried by the booster version F9 v1.1 is about 2534.67 kg

# First Successful Ground Landing Date

| MIN("Date") |
| --- |
| 01/08/2018 |

The date of the first success landing on the ground pad is
01/08/2018

# Successful Drone Ship Landing with Payload between 4000 and 6000

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

There are four booster versions with launches that had payload mass between 4000 and 6000kg and had successful landings on the drone ship

# Total Number of Successful and Failure Mission Outcomes

| Mission_Outcome | Total |
|---|---|
| None | 0 |
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

There are 100 launches with a successful mission outcome and
1 launch with a failed mission outcome

# Boosters Carried 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 |

There are 12 different booster versions that carried the maximum payload mass

# 2015 Launch Records

| Month | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|
| 10 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

There were two failed landings on the drone ship in 2015 and they occurred in April and October from the Launch Site. They had different Booster Versions

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

| Landing_Outcome | Total |
|---|---|
| Success | 20 |
| No attempt | 10 |
| Success (drone ship) | 8 |
| Success (ground pad) | 7 |
| Failure (drone ship) | 3 |
| Failure | 3 |
| Failure (parachute) | 2 |
| Controlled (ocean) | 2 |
| No attempt | 1 |

In the launches between 2010-06-04 and 2017-03-20, the highest-ranked Landing Outcome was Success and the lowest-ranked Landing Outcome was (technically) Controlled (ocean)

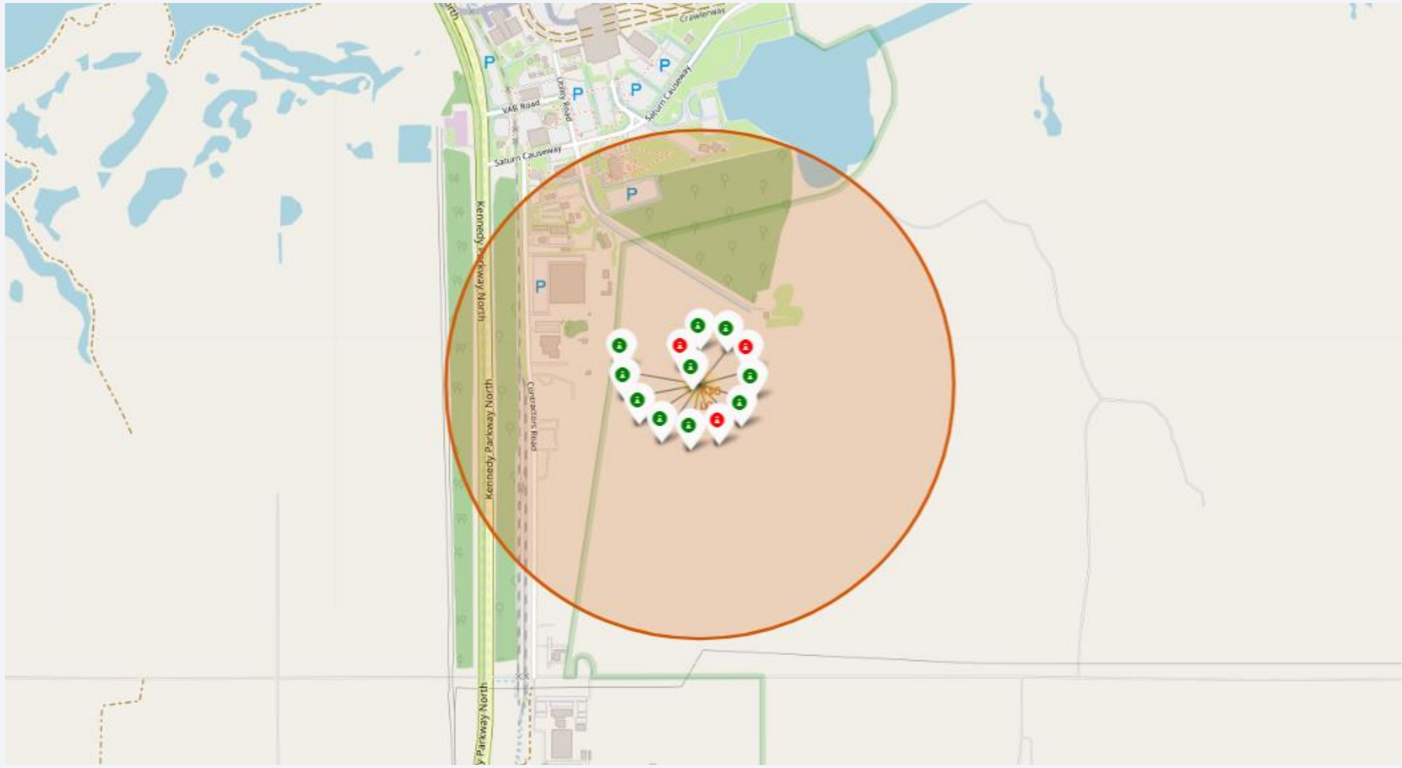
Section 3

# Launch Sites Proximities Analysis

# Locations of each Launch Site

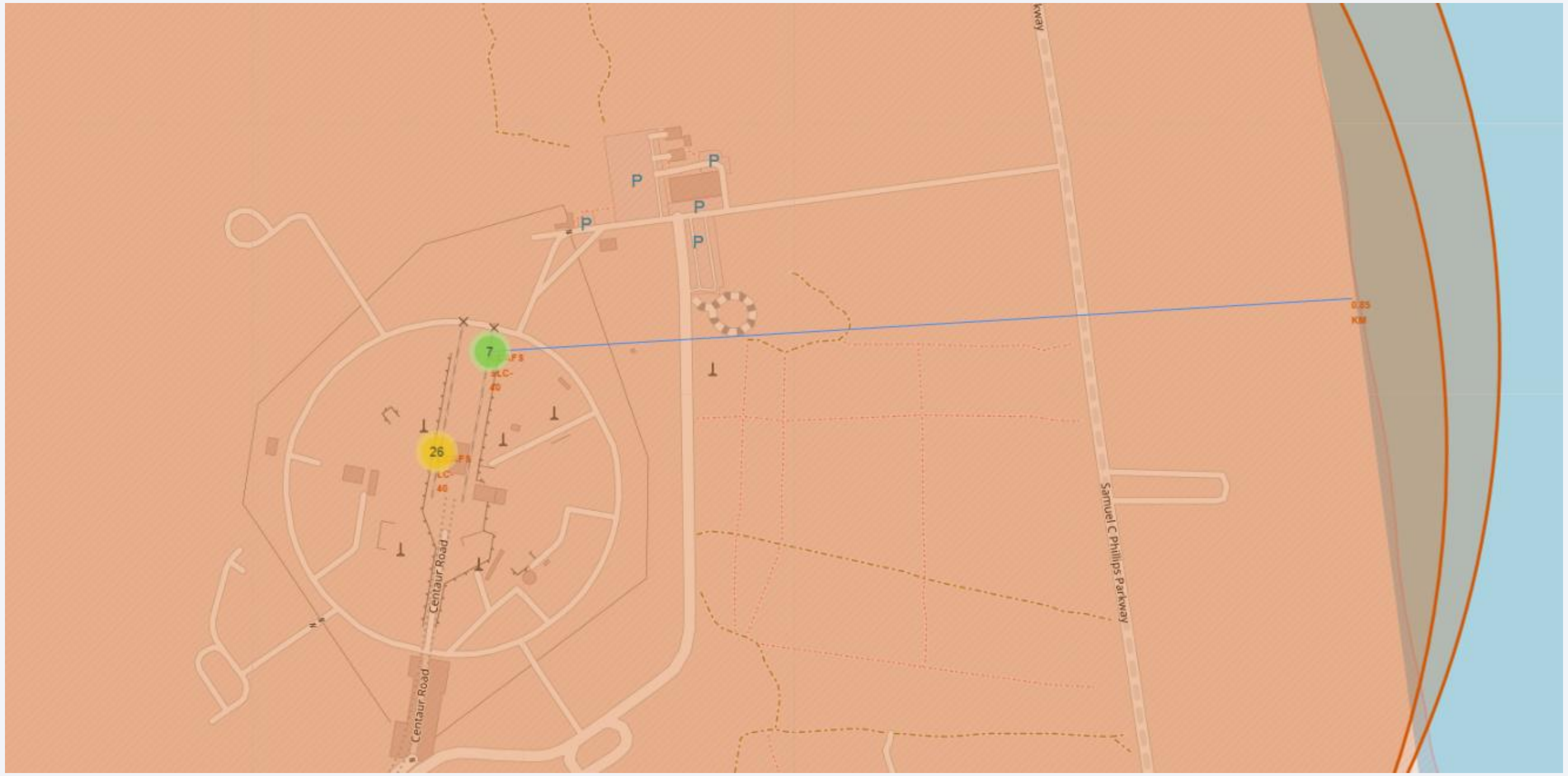

There are four launch sites displayed on the map; one is on the west coast, in the state of California, and the other three are in the state of Florida

# Outcomes of Launch Site KSC LC-39A




At Launch Site KSC LC-39A there were 13 launches; of the launches, 10 had successful landings while 3 failed

# Point of Interest Near Launch Site CCAFS SLC-40

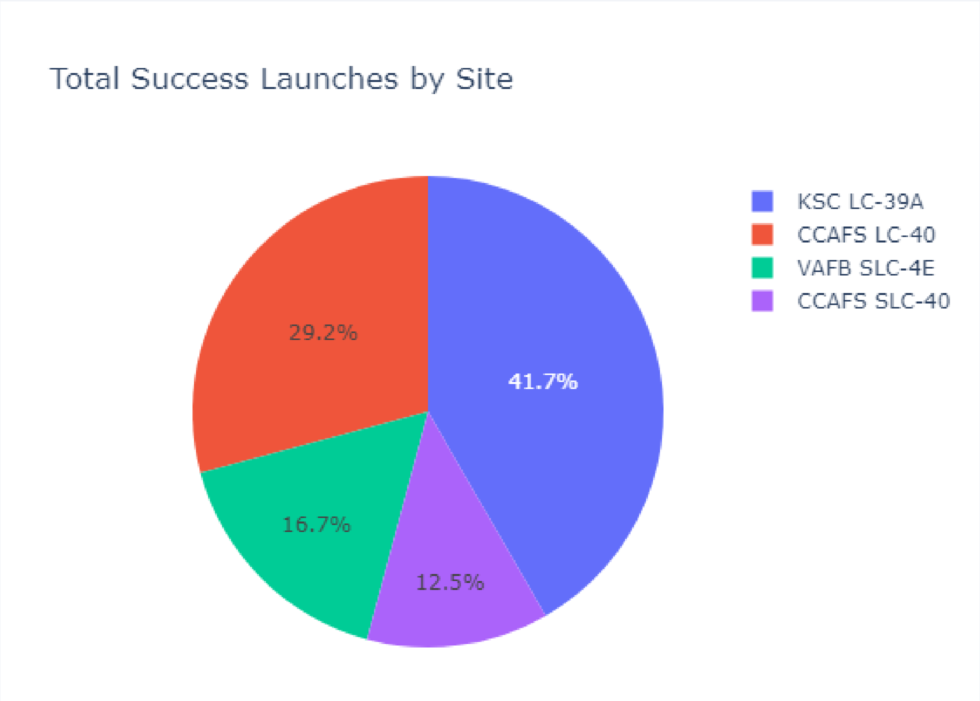



The distance between Launch Site CCAFS SLC-40 and the coastline is about 0.85 km; this distance is shown by the blue line

Section 4

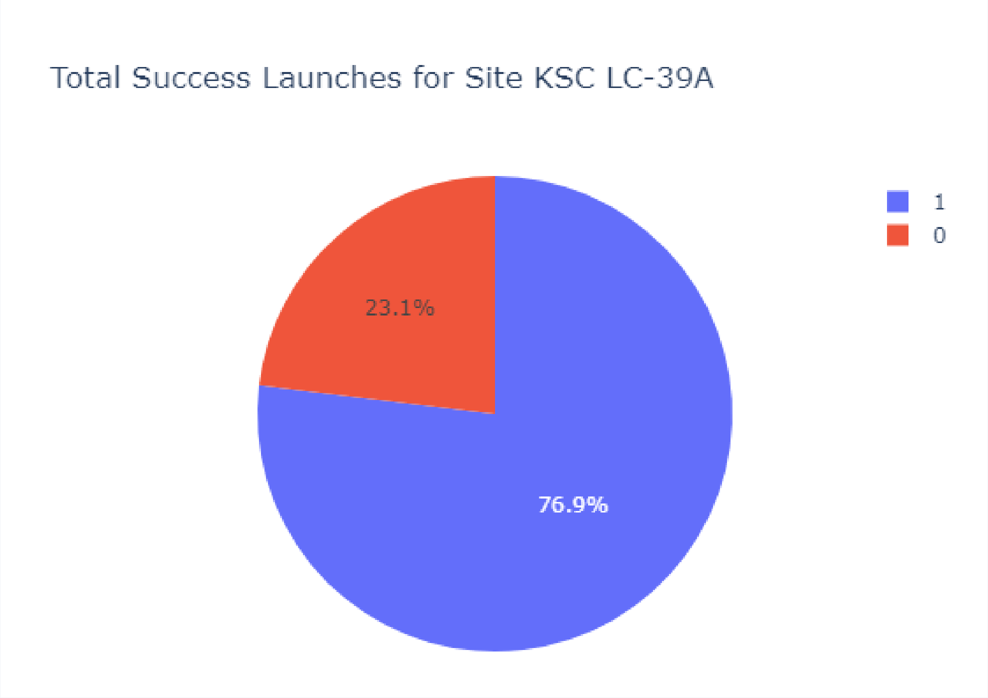
# Build a Dashboard with Plotly Dash

# Dashboard - Launch Success Count



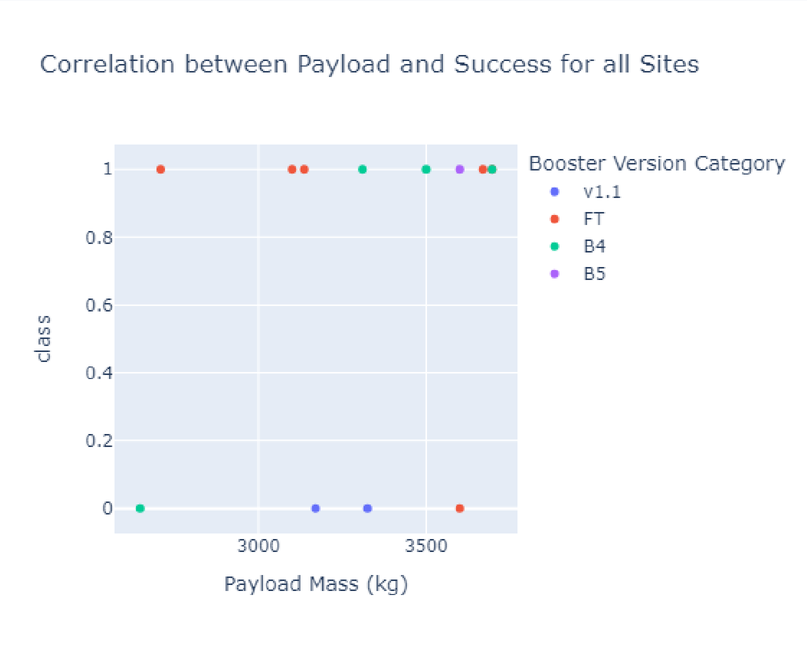

Out of all sites, Launch Site KSC LC-39A has the highest launch success count, and accounts for 41.7% of the total of all sites

# Dashboard - Highest Launch Success Rate

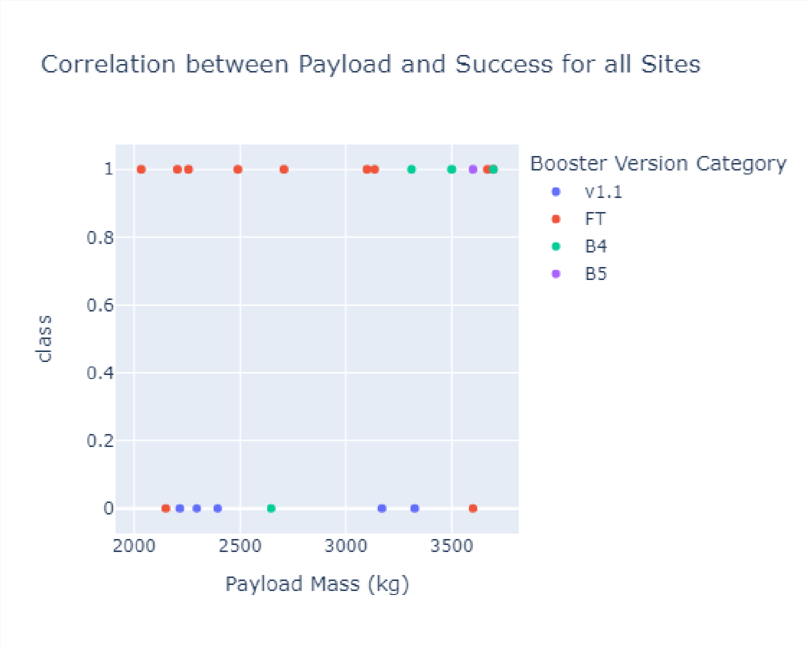



Out of all sites, Launch Site KSC LC-39A has the highest launch success rate (76.9%)

# Dashboard – Payload Mass vs. Launch Outcome




The payload range of 2500-4000 kg seems to have the highest success rate of 67%

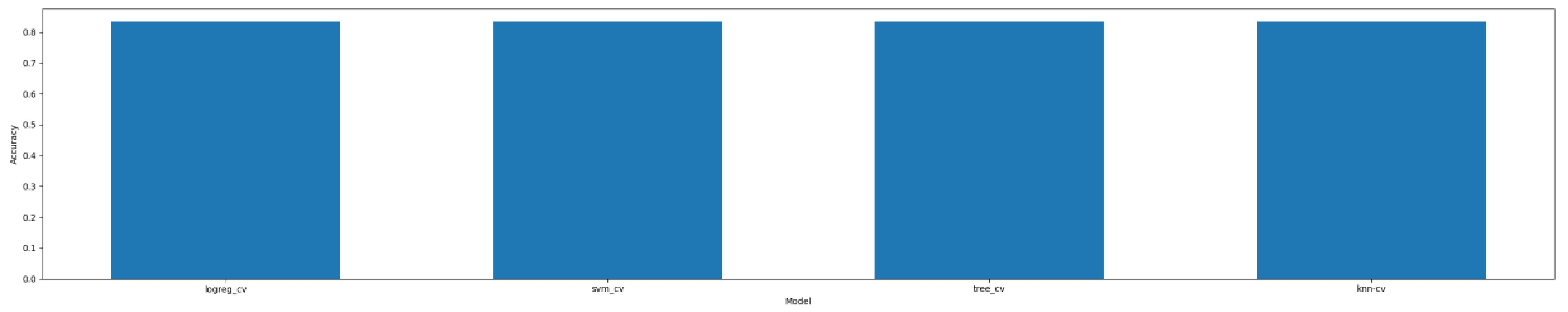



In the payload range of 2000-4000 kg, booster version FT has the highest success rate (80%) compared to the other boosters in that range (except version B5)

Section 5

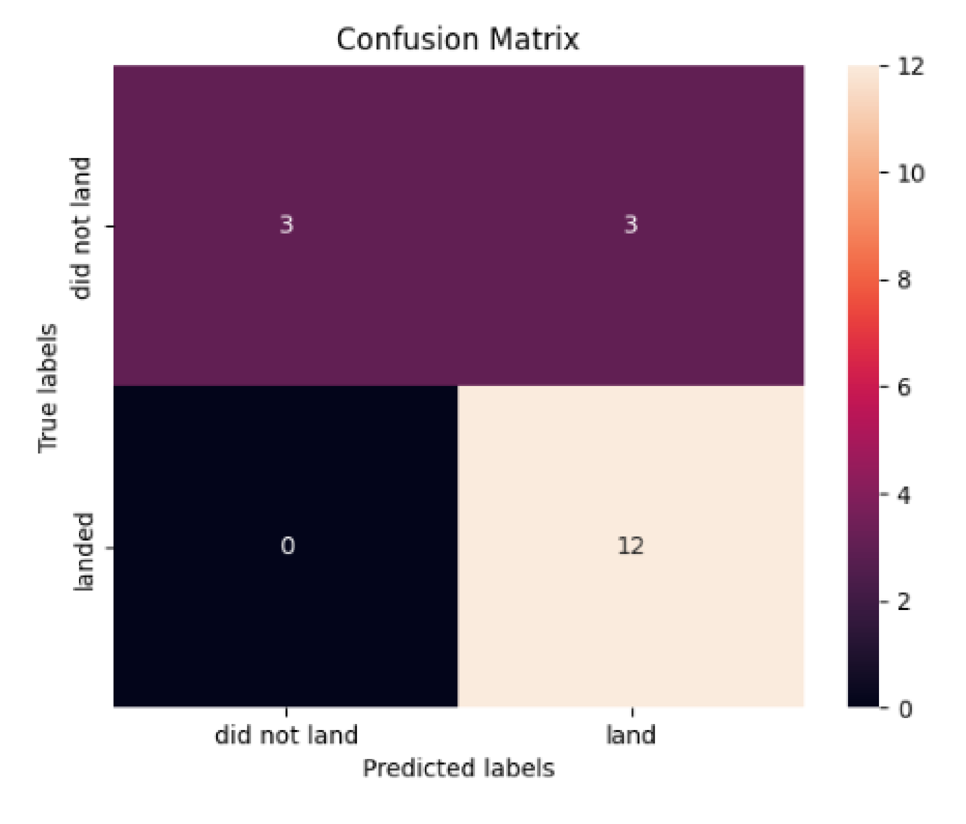
# Predictive Analysis (Classification)

# Classification Accuracy




All models yield the same accuracy of 0.833

# Confusion Matrix




The confusion matrix is the same for all models

# Conclusions

- Launches from site KSC LC-39A has the highest landing success rate and count of successful landings of all launch sites.
- Booster FT seems to yield high landing success rate whenever it launches with a payload between 2000 and 4000 kilograms.
- Out of the four classification models tested on the dataset, there is no specific one that is the best as they all yield the same accuracy on testing data (~0.83).

# Appendix

- GitHub repository of Jupyter Notebooks and Python files for this project:
    - https://github.com/gheladevang/SpaceXCapstone

# Thank you!