# Going to Space with SpaceY

Joey Laffort

03/01/2024

# Table of Contents

# Executive Summary

## Summary of Methodologies

- **Data Collection |** Make get request to the SpaceX API, and use data wrangling and formatting.
- **Data Wrangling |** Find patterns in the collected data, and determine what would be the label for training supervised models.
- **Exploratory Data Analysis |** Determine the cost of a launch based on whether the first stage will land.
- **Building an Interactive Map with Folium |** Analyze launch locations to discover factors that affect the launch success rate.
- **Building a Dashboard with Plotly Dash |** For users to perform interactive visual analytics on SpaceX launch data in real-time.
- **Predictive Analysis (Classification) |** Create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.

## Summary of Results

**Exploratory Data Analysis**

- KSC LC-39A has the highest success rate among launch sites.
- Launch success has improved over time.

**Exploratory Data Analysis**

- Majority of the launch sites are located near the equator, and all are close to the coast for various reasons.

**Exploratory Data Analysis**

- The decision tree model slightly outperformed the other models, however all model's results were almost identical.

# Introduction

## Project Background & Context

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps one of the most successful companies of the commercial space age is SpaceX. One of the reasons that makes SpaceX so successful, is because their rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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. We will be gathering information about Space X and creating dashboards for our team. We will also determine if SpaceX will reuse the first stage. In order to accomplish this, we will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

## Project Goals

- Predict the likelihood successful first stage landing.
- Determine how payload mass, launch site, number of flights, and orbits affect a successful first-stage landing.
- Determine the best algorithm that can be used for binary classification to find out whether the successful landings increase over time.

# Methodology

# Methodology

## Summary of Methodologies

- **Data Collection**

  - Utilize SpaceX API.

  - Web scraping to collect Falcon 9 historical launch records from a Wikipedia.

- **Data Wrangling**

  - Filter the data.

  - Handling missing values and prepare the data for analysis and modeling.

- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics**

  - Build an interactive map with Folium

  - Build a dashboard with Plotly Dash

- **Perform Predictive Analysis (by using classification models)**

# Data Collection

## SpaceX API - Steps

- Request rocket launch data data from SpaceX API.
- Decode response using .json() and convert to a data frame using .json_normalize().
- Request relevant information regarding the launches from SpaceX API by utilizing custom functions.
- Create dictionary based on the data we obtained.
- Create a Pandas data frame from the dictionary.
- Filter the data frame to only include Falcon 9 launches.
- Replace missing values of Payload Mass with calculated .mean().
- Export the data into a csv file.

## Web Scraping - Steps

- Request Falcon 9 launch data from Wikipedia.
- Create BeautifulSoup object from HTML response.
- Extract column names from HTML table header.
- Collect data from parsing HTML tables.
- Create dictionary from the data.
- Create a Pandas data frame from the dictionary.
- Export the data into a csv file.

# Data Wrangling

## Steps

- Perform EDA to find patterns in the data and determine what would be the label for training supervised models.
- Calculate the number of launches on each site.
- Calculate the number and occurrence of each orbit.
- Create a landing outcome label from outcome column.
- Export the data into a csv file.

## Landing Outcome

- **True Ocean** indicates a successful landing in a specific region of the ocean.
- **False Ocean** indicates an unsuccessful landing outcome in a specific region of the ocean.
- **True RTLS** indicates a successful landing to a ground pad.
- **False RTLS** indicates an unsuccessful landing to a ground pad.
- **True ASDS** indicates a successful landing to a drone ship.
- **False ASDS** indicates an unsuccessfully landing to a drone ship.
- The outcomes are converted into training labels with:
  - **`1`** means the booster successfully landed .
  - **`0`** means it was unsuccessful.

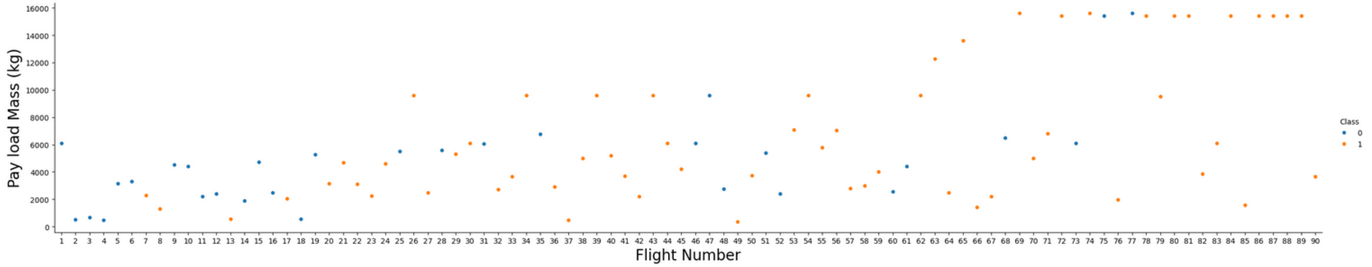
# EDA Results

# Exploratory Data Analysis

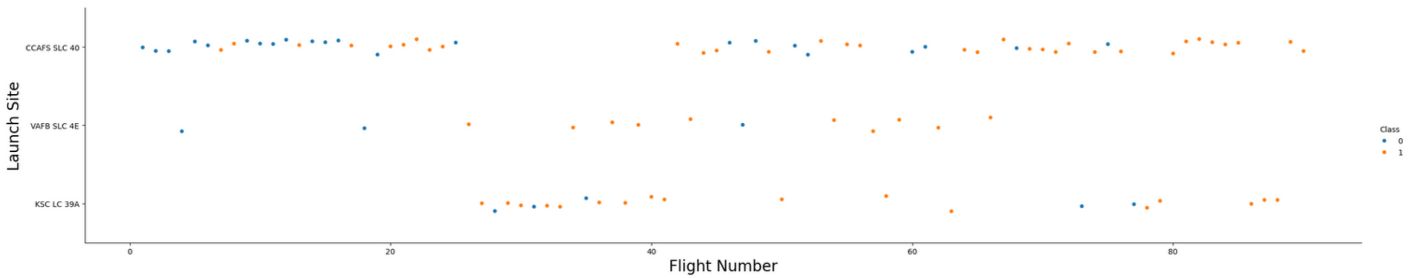
## EDA with Visualization - P1/2



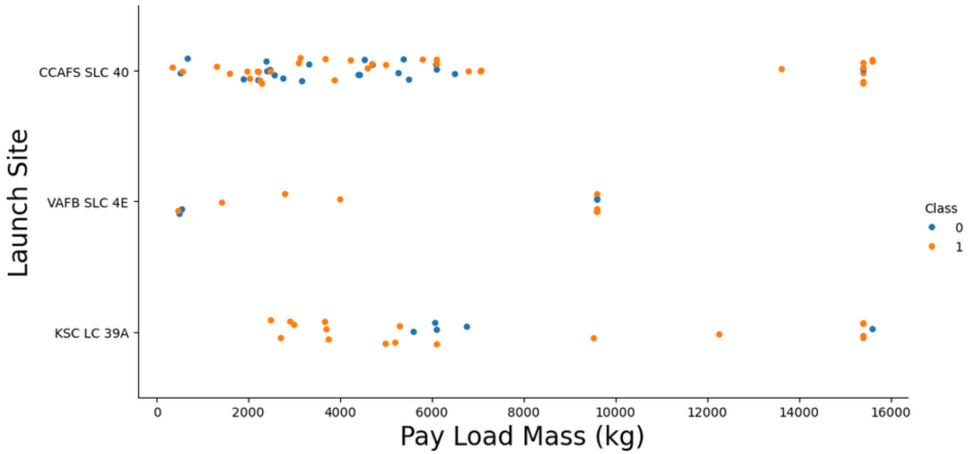
FlightNumber vs PayloadMass - 1st stage landing success positively correlated with continuous launch attempts, while negatively correlated with payload mass:




FlightNumber vs LaunchSite - CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place:



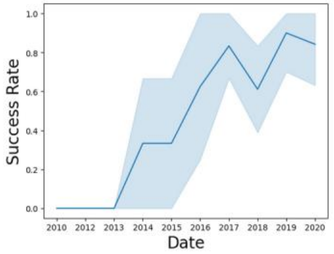

PayloadMass vs LaunchSite - CCAFS SLC 40 and KSC LC 39A appear to be favored for heavier payloads:



# Exploratory Data Analysis

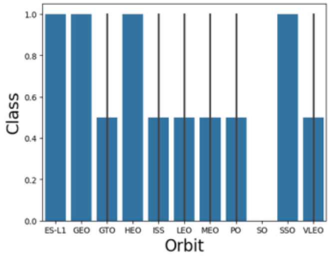
## EDA with Visualization - P2/2

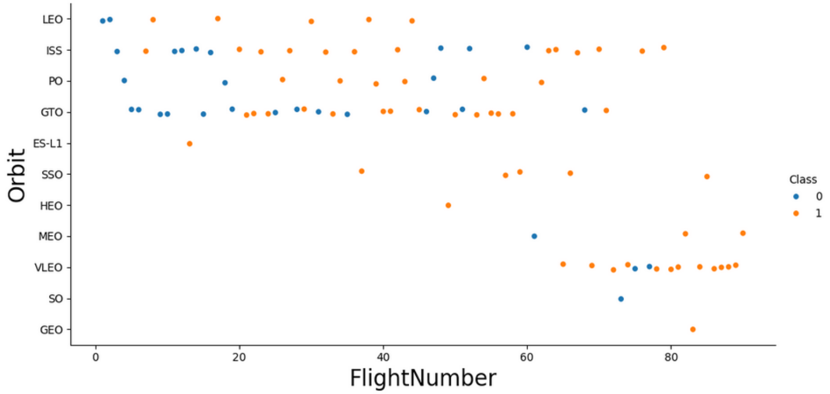




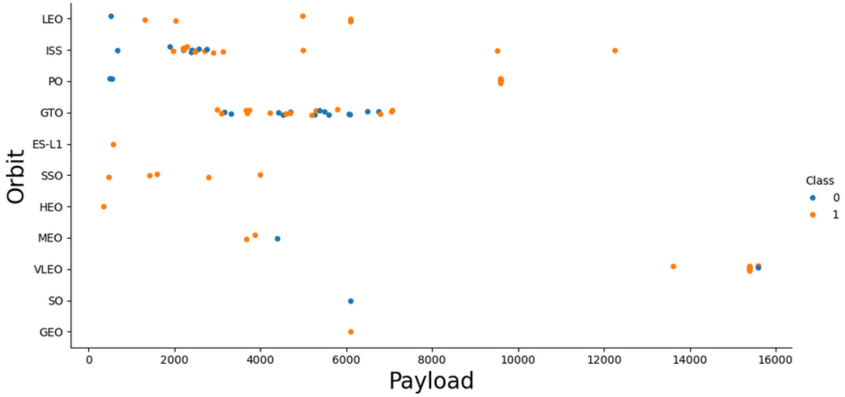

Year vs Success rate - Success rate trending positively on a yearly basis since 2013.

Orbit type vs Success rate - All orbit types except 'SO' have had successful 1st stage landings.




FlightNumber vs Orbit type - Flight number positively correlated with 1st stage recovery for all orbit types.




PayloadMass vs Orbit type - Heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits.

# Exploratory Data Analysis

## EDA with SQL - P1/3



SpaceX Launch Sites:

- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

Examine launch site and date records that start with the string 'CCA':

- Last launch from CCAFS LC-40 was 2016-08-14
- First launch from CCAFS SLC-40 was 2017-12-15

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

| TOTAL PAYLOAD MASS |
|---|
| 45596 |

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

| F9 v1.1 AVG PAYLOAD MASS |
|---|
| 2534.6666666666665 |

List the date when the first successful landing outcome in ground pad was achieved:

| FIRST SUCCESFUL GROUND PAD LANDING |
|---|
| 2015-12-22 |

# Exploratory Data Analysis

## EDA with SQL - P2/3



Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:

| Booster_Version | Payload_Mass |
|---|---|
| F9 FT B1022 | 4696 |
| F9 FT B1026 | 4600 |
| F9 FT B1021.2 | 5300 |
| F9 FT B1031.2 | 5200 |

Total number of successful and failure mission outcomes:

- 1 - Failure (in flight)
- 99 - Success
- 1 - Success (payload status unclear

Names of the booster_versions which have carried the maximum payload mass:

| Booster Version w/ Max Playload Mass | PAYLOAD_MASS__KG_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1060.3 | 15600 |
| F9 B5 B1049.7 | 15600 |

# Exploratory Data Analysis

## EDA with SQL - P3/3

Records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015:

| 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 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:

| Landing_Outcome | Total |
|---|---|
| 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 |

# Interactive Visual Analytics
# Results

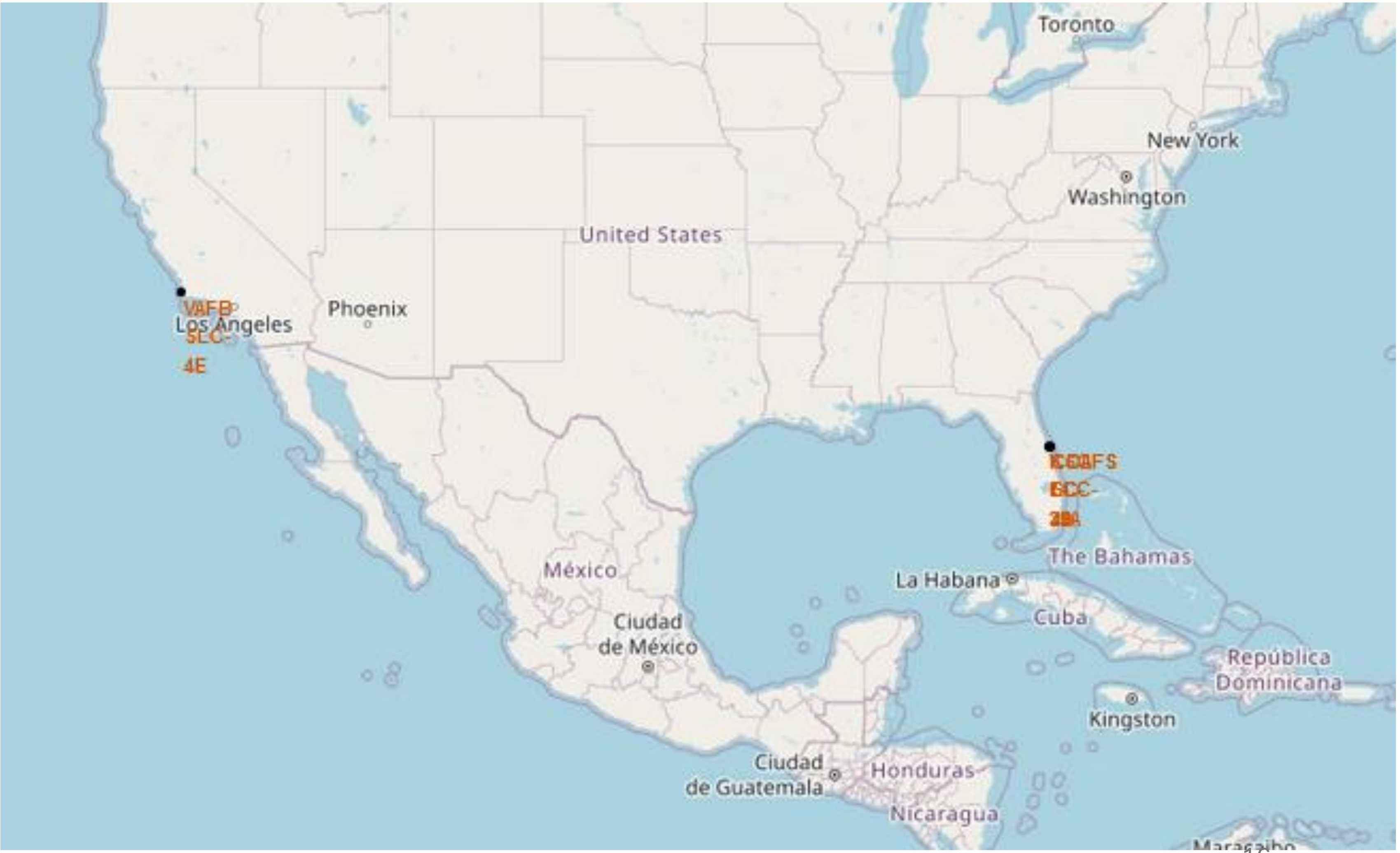
# Interactive Map with Folium

## Folium P1/3 - Launch Site Locations



**Insights**

All launch sites are located in North America, and are located near to coastlines - specifically the coasts of Florida and California.



# Interactive Map with Folium

## Folium P2/3 - Labeled Launch Outcomes

IBM Developer

### Insights

Visualizing the launch outcomes for each launch site highlights which launch sites have relatively high success rates.

- **Green** label highlights Successful launch outcome
- **Red** label highlights unsuccessful launch outcome

# Interactive Map with Folium

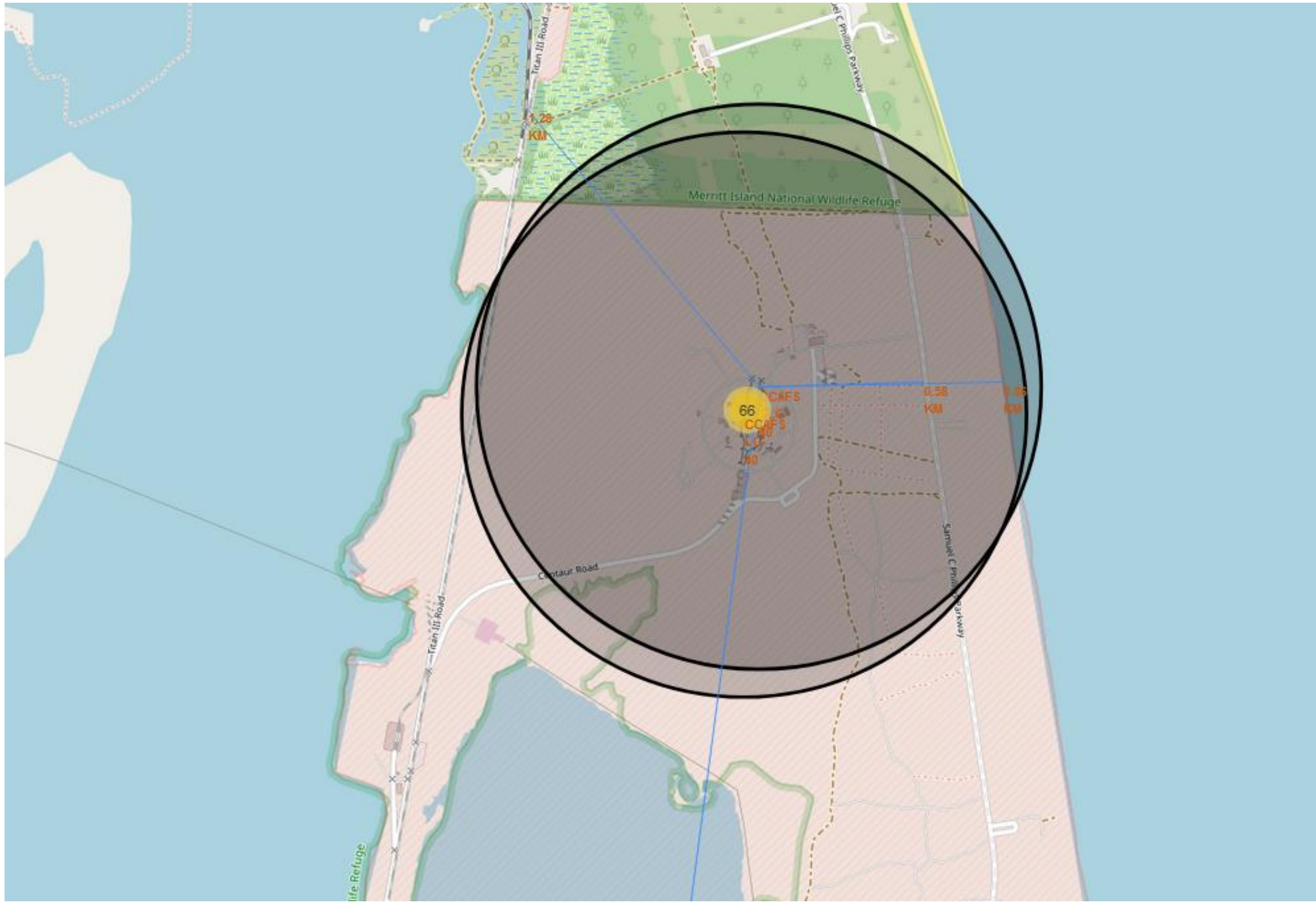
## Folium P3/3 - Launch Site Locations



### Insights

Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each is, for example, proximities for CCAFS SLC-40:

- Railway: 1.28 km
- Highway: 0.58 km
- Coastline: 0.86 km
- City: 51.43 km



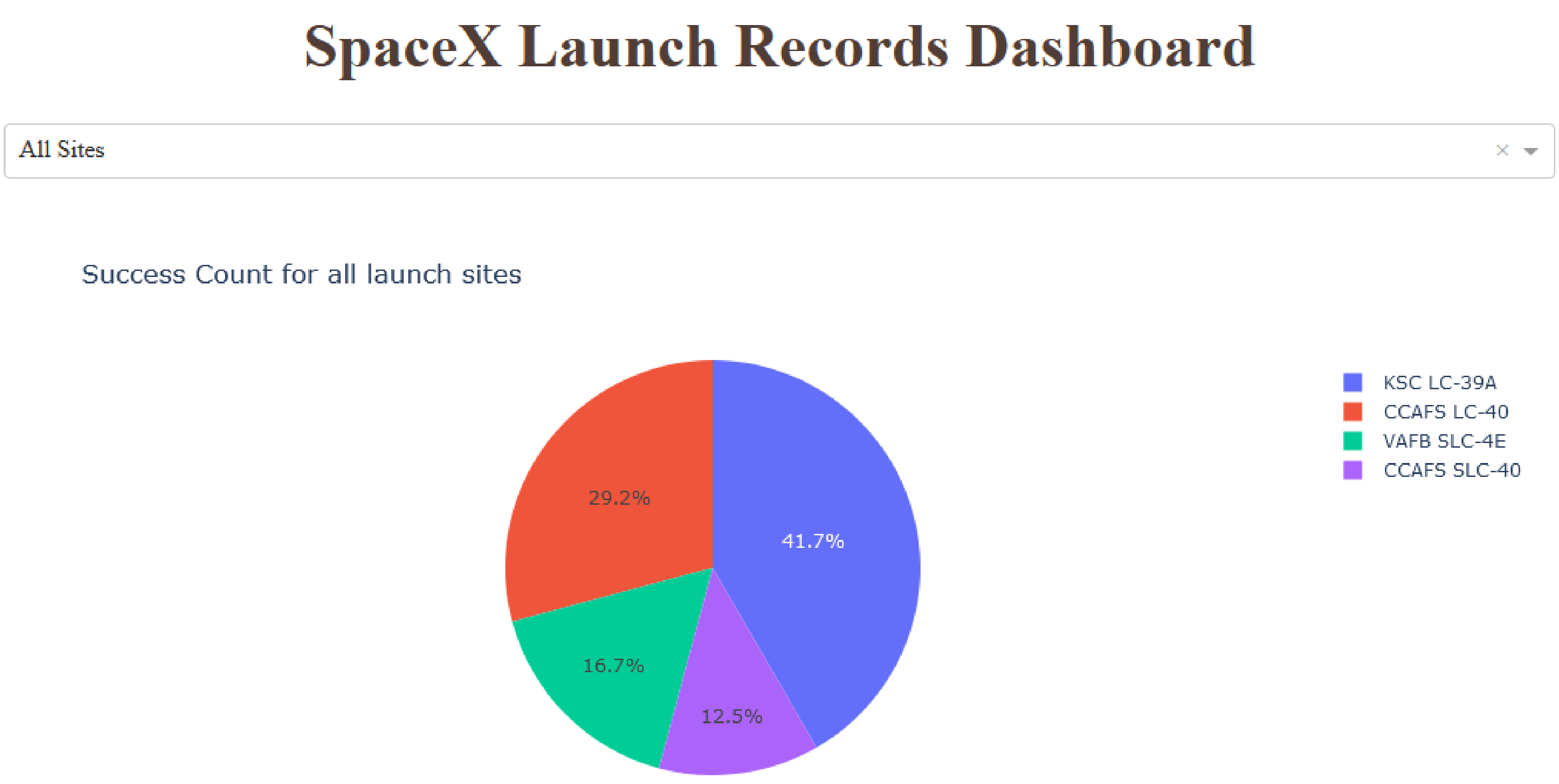
# Dashboard with Plotly Dash

## Plotly Dash P1/3 - Successful Stage 1 Landings by Launch Site



**Insights**

most successful landings were launches from KSC LC-39A. The least successful landings were launches from CCAFS SLC-40



# Dashboard with Plotly Dash

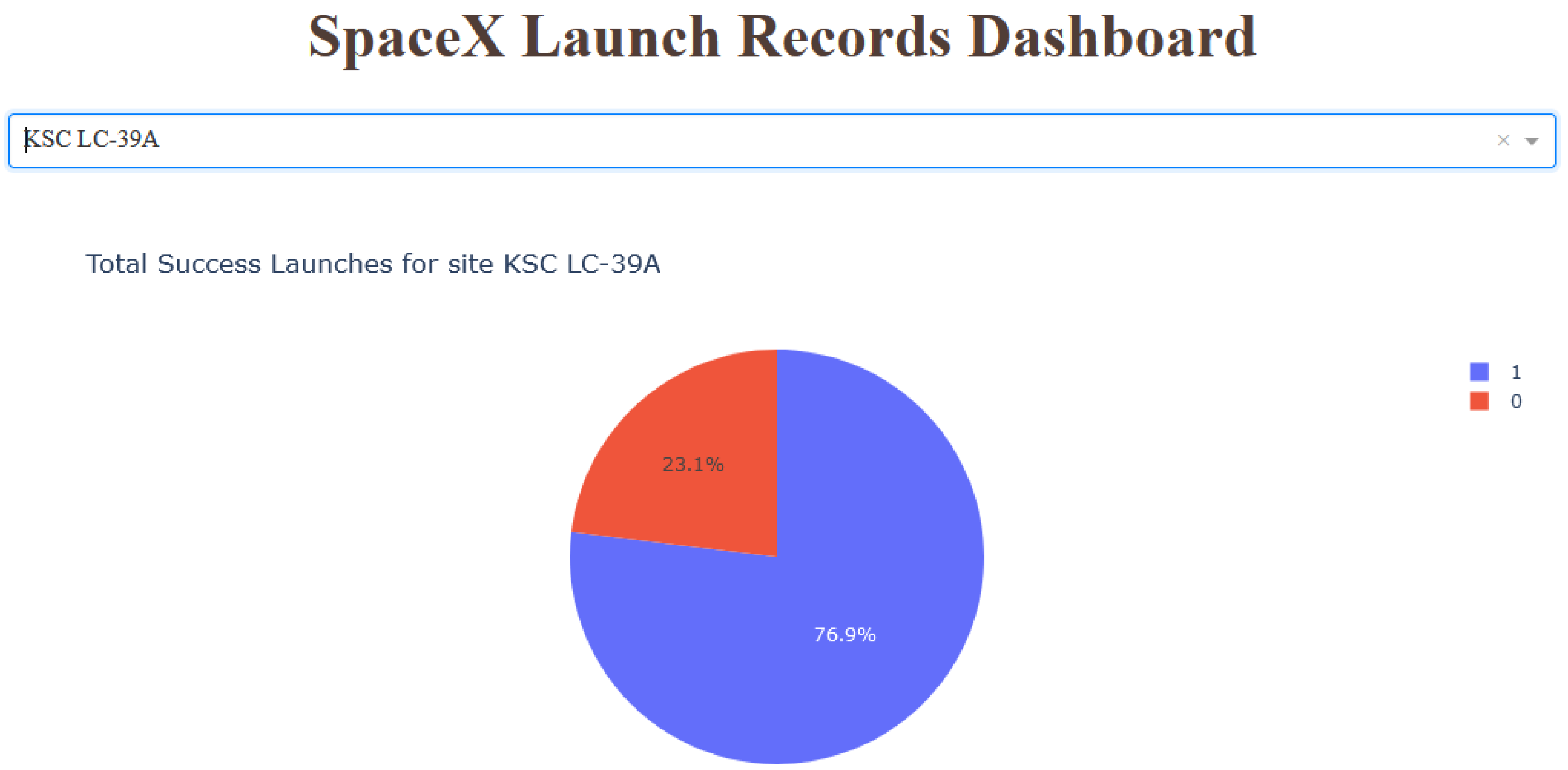
## Plotly Dash P2/3 - Successful Stage 1 Landings for KSC LC-39A

IBM Developer

### Insights

Drilling into the most successful launch site, we can see the success vs failure for KSC LC-39A. Even though many of the population successes are from this launch site, it actually has a low success rate.

- 1 = Unsuccessful Landing
- 0 = Successful Landing



# Plotly Dash P3/3 - Payload Mass vs Success vs Booster Version Category



## Insights

Booster version category FT has many successes and few failures, whereas v1.1 has many failures and few successes.

# Predictive Analysis (Classification) Results

# Predictive Analysis (Classification)

## Classification

### Accuracy

All training models practically had the same accuracy levels (83.33%), which is likely due to the small dataset.

However, the Decision Tree model slightly outperformed the rest when looking at .best_score_ (average of all cv folds for a single combination of the parameters).

```
models = {'KNeighbors': knn_cv.best_score_,
          'DecisionTree': tree_cv.best_score_,
          'LogisticRegression': logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalg = max(models, key = models.get)
print('Best Model =', bestalg, '| score =', models[bestalg])
```

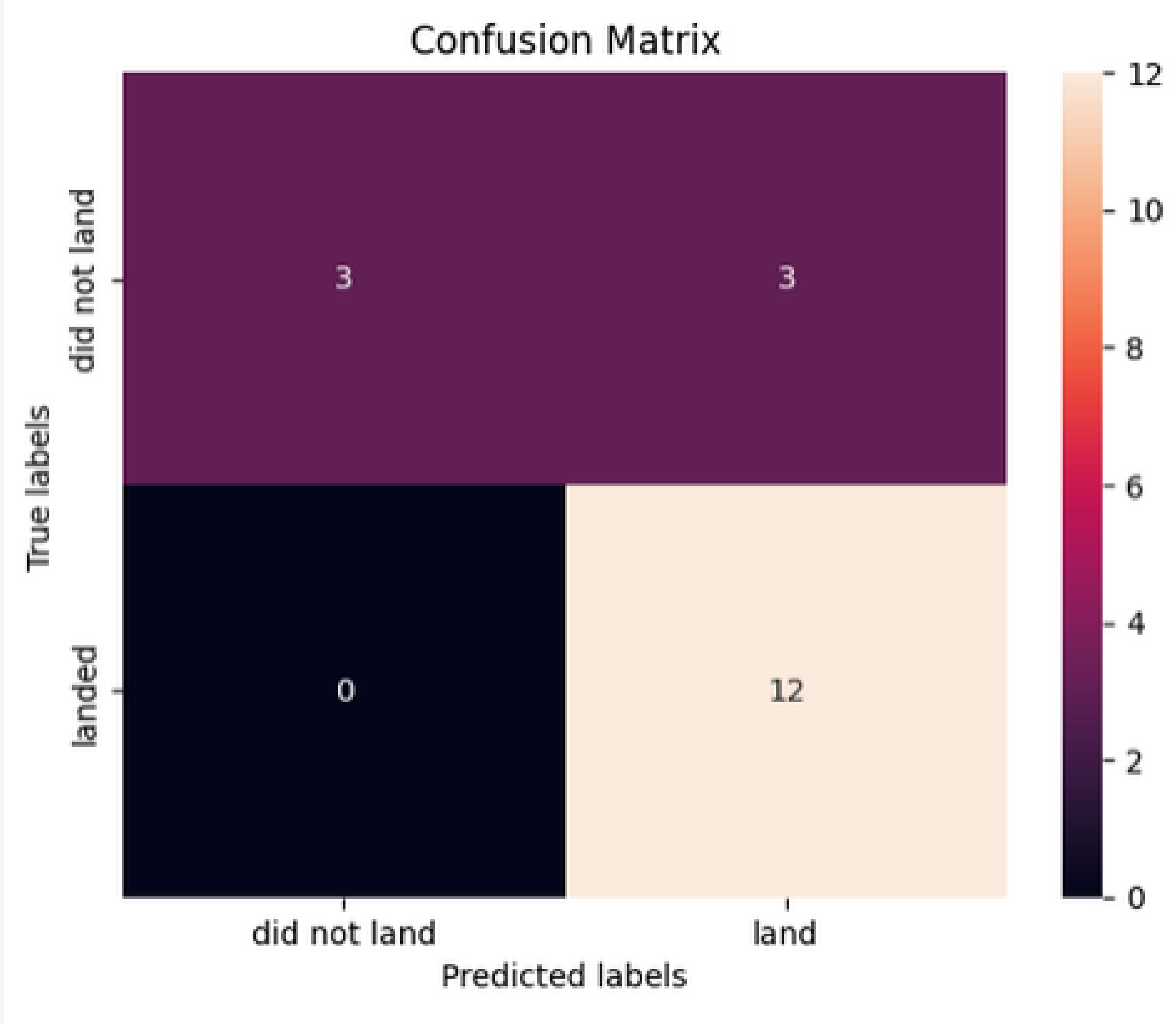
```
Best Model = DecisionTree | score = 0.8607142857142858
```

# Confusion Matrix



**Result**

The confusion matrices of the best performing models (4-way-tie) are identical. The false positives are an issue as indicated by the models, which are incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set.



# Conclusion

## Analysis

- Model Performance - The models performed very similarly on the test set, however the decision tree model had a slight advantage.
- Equator - The majority of the launch sites are located near the equator to take advantage of the rotational speed of earth to minimize fuel cost.
- Coast - All launch sites are near the coast.
- Launch Success increases over time.
- Success Rate - KSC LC-39A has the highest success rate among launch sites (with a 100% success rate for launches less than 5,500 kg).
- Orbits - ES-L1, GEO, HEO, and SSO have a 100% success rate.
- Payload Mass - Higher the payload mass (kg), increases the success rate.

# Appendix

## Work Files

GitHub Repository - https://github.com/JoeyLaffort/Applied.Data.Science.Capstone

| JoeyLaffort Add files via upload | | 718b1c7 · 5 minutes ago 3 Commits |
|---|---|---|
| Collecting Data - Hands-on Lab Complete the ... | Add files via upload | 5 minutes ago |
| Collecting Data - Hands-on Lab Complete the ... | Add files via upload | 5 minutes ago |
| Data Wrangling - Hands-on Lab Data Wranglin... | Add files via upload | 5 minutes ago |
| Exploration Analysis Using Pandas and Matplot... | Add files via upload | 5 minutes ago |
| Exploratory Analysis Using SQL - Hands-on La... | Add files via upload | 5 minutes ago |
| Interactive Visual Analytics and Dashboards - ... | Add files via upload | 5 minutes ago |
| Interactive Visual Analytics and Dashboards - ... | Add files via upload | 5 minutes ago |
| Preditive Analysis (Classification) - Hands-on La... | Add files via upload | 5 minutes ago |
| README.md | Update README.md | 8 minutes ago |

# Thank you!