IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Brendan Monaghan
04/21/2025

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis and Visualization
  - Exploratory Data Analysis with SQL
  - Created an Interactive Map with Folium
  - Created an Interactive Dashboard with Plotly Dash
  - Machine Learning and Predictive Analytics
- Summary of all results
  - Discussion of Exploratory Data Analysis findings
  - Review the interactive analytics
  - Machine Learning and Predictive Analytics results

# Introduction

- Project background and context

  We are in the beginning of the commercial space age and SpaceX is leading the way with the Falcon9 rocket. With its ability to land and be reused, the cost of Falcon is 62 million compared to the 165 million other competitors in the market cost due to the fact they do not reuse the first stage of the rocket like SpaceX. Using SpaceX data to help determine if the first stage will be able to be reused, we can determine the general costs associated with the space travel.

- Problems you want to find answers

  - What is the impact of variables such as payload mass, number of flights, location of launch sites, and orbits have on the success of the first stage landing and ability to reuse?
  - How has the Falcon9 success rate changed over the years?
  - Can we predict the costs of launches based on the success rate of the first stage landing?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data pulled from SpaceX Rest API
  - Data pulled from Wikipedia using Web Scraping
- Perform data wrangling
  - Filtered data to just the Falcon9 launches
  - Cleared any missing data using the mean of existing results
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built Classification models, tuned and tested the models to determine which model would be best to use for predicting results

# Data Collection

There were two primary data collection processes used in this project, data requests from SpaceX REST API and web scraping from the Wikipedia SpaceX page.  By using both methods, we were able to get a more complete data set to review.

- The following fields of data were collected from the REST API:
  - FlightNumber, Date,  BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount,Serial, Longitude, Latitude

- The following fields data were collected from the SpaceX Wikipedia page:
  - Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Luanch outcome, Version Booster, Booster landing, Date, Time

# Data Collection – SpaceX API

Prior to requesting data from the API, we defined 4 functions to parse the data: getBoosterVersion, getLaunchSite, getPayloadData, and getCoreData

Used requests.get to pull the JSON data from the API to a variable. Once the data was pulled in, we further refined it by using json_normalize to put the data into a dataframe

Created global variable to hold the relevant data from the original data frame and used the defined functions to extract the data to the variables

Created dictionary to hold the data and created a data frame called data_falcon9. We further refined the data to just the falcon9 launches

Used the .mean() function to fill in the null values in the PayLoadMass field

Exported the final cleaned data to a CSV file



GitHub: Data Collection API

# Data Collection - Scraping

Used requests.get to get Falcon9 launch data from Wikipedia

Created a BeautifulSoup object called soup from the HTMP data we requested

Extracted all column names and variables from the HTML table header using the .find_all(‘table’) function

Iterated through the table header to get column names from the table

Extracted the relevant data from the table into a data dictionary then created a data frame from the dictionary

Exported the final cleaned data to a CSV file



[GitHub: Data Collection Scraping]

# Data Wrangling

Explored the data to understand what the outcome field represented

Calculated the number of launches per launch site using the value_counts () function

Calculated the number launches grouped by orbit using the value_counts () function

Created a landing_outcomes data frame using .value_counts() then using the enumerate() and .keys() functions we indexed the outcomes

We then created a set of bad_outcomes to allow us to identify a new field called 'Class' in the data frame to assign 0 to failures and 1 to successes

Exported the completed exploratory data analysis to a CSV file



[GitHub:  Data Wrangling]

# EDA with Data Visualization

- Chart types used for exploring the data:
  - Scatter Plots:
    - Flight Number vs. Payload Mass: We observed that as flight number increases, landings are more successful regardless of the Payload Mass
    - Flight Number vs. Launch Site: We observed that CCAFS SLC 40 launch site had the most flights and successful landings
    - Launch Site vs. Payload Mass: We observed that VAFB SLC 4E launch site no Payload Mass greater than 10,000 kg
    - Flight Number vs. Orbit: In the LEO orbit, success appears to be related to number of flights but that does not appear to be so in the GTO orbit
    - Payload Mass vs. Orbit: Polar, LEO, & ISS had a higher positive landing rate while GTO could not be distinguished

Bar Chart:

Orbit Success Rate: We observed in the bar chart that ES-L1, GEO, HEO, & SSO all had the highest success rate

Line Graph:

Yearly Success Rate: We observed a steady increase in success rate from 2013-2019 but a drop in success in 2020



[GitHub: EDA with Data Visualization]

# EDA with SQL

- SQL queries performed on the data:
  - Displayed the names of the unique launch sites in the space mission
  - Displayed 5 records where launch sites begin with the string 'CCA'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date when the first successful landing outcome in ground pad was achieved.
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes
  - Listed all the booster versions that have carried the maximum payload mass. Use a subquery.
  - Listed the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
  - Ranked 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.



[GitHub: EDA with SQL]

# Build an Interactive Map with Folium

- Added a circle marker, pop label and text label of the NASA Johnson Space Center in Huston
- Added circle markers, pop labels and text labels indicating launch sites on the map
- Colored the launch site markers based on outcomes.  Green = Successful; Red = Failure
- Defined a function to calculate the distance between points on the map
- Using site CCAFS SLC- 40 as the base point, we added markers for the nearest coastline, railway, highway, and city.  Used the function to calculate the distances and added them to the map
- Doing this allowed us to visualize the proximities of the launch sites to these critical points.



[GitHub: Interactive Mapping with Folium]

# Build a Dashboard with Plotly Dash

Launch Site Dropdown:

- Added a dropdown list of the launch sites allowing for the ability to look at results of all sites or focus on an individual site

Slider of Payload Mass:

- Added a slider feature to allowing the ability to look at results for all payload masses or results for specific payload ranges

Charts and Graphs:

- Added a pie chart showing the success rate of all sites or specific sites depending on what is chosen in the Launch Site Dropdown

- Added a scatter plot to show any correlation between Payload Mass vs. Success Rate

GitHub: SpaceX Dash App

# Predictive Analysis (Classification)

1. Loaded the dataframe and created a NumPy array from the Class column
2. Standardized the data using StandardScaler. Then used transform.fit_transform to prepare the data
3. Split the data into training set and test set using train_test_split
4. Created the following 4 models: LogReg, SVM, Decision Tree, & KNN
5. Used GridSearchCV with cv=10 on all models to find the best parameters
6. Produced a Confusion Matrix for all 4 models
7. Compared all models using the best_score_ funciton

[GitHub: SpaceX Machine Learning](#)

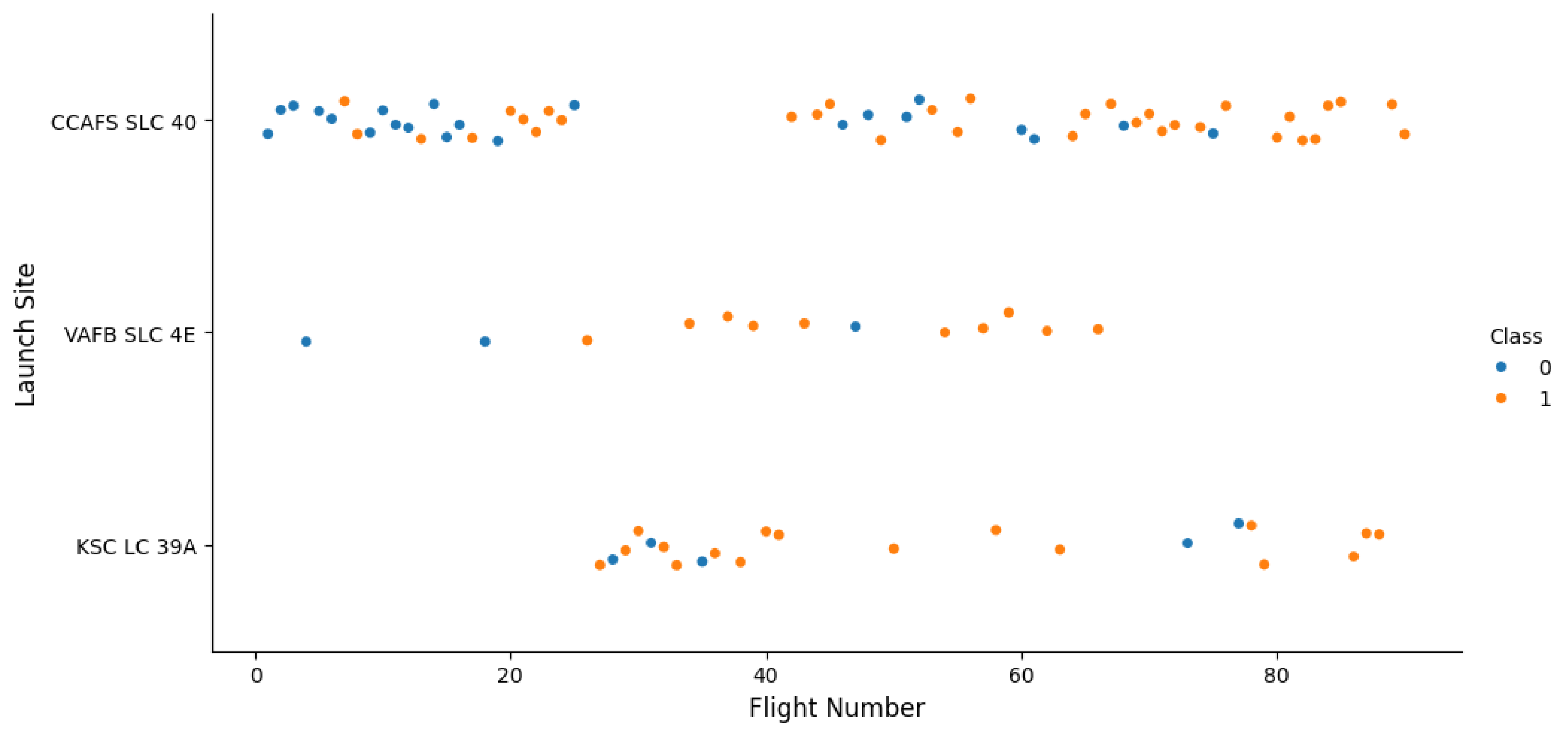
# Results

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

Section 2

# Insights drawn from EDA

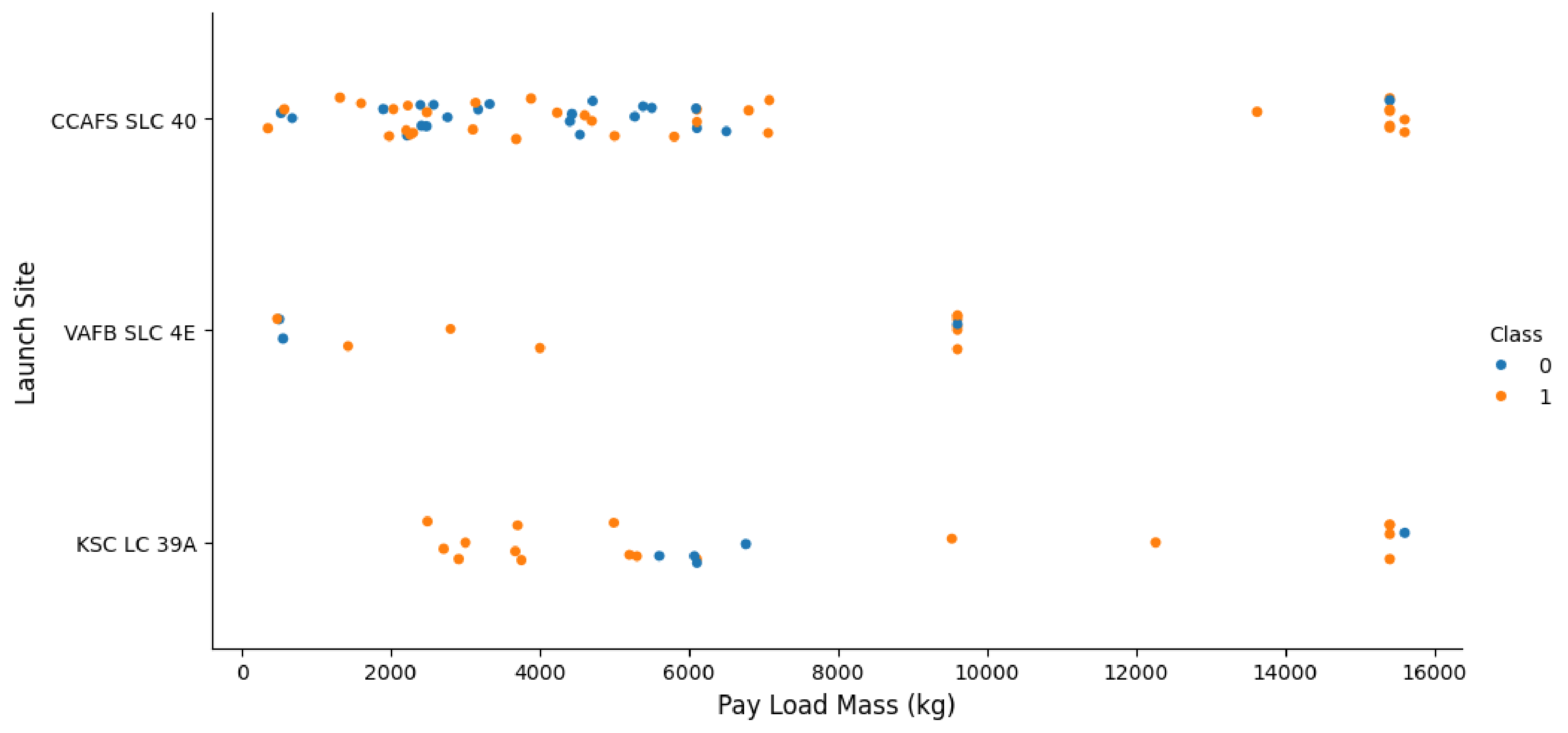
# Flight Number vs. Launch Site




Observations:

- CCAFS SLC 40 has the most flights out of all 3 launch sites and was the primary launch site at the early stages having 23 of the first 25 flights and only 9 were successful. This demonstrates a learning curve; as the flights increased so did the success.
- It shows that after the initial 25 flights and the increase in success, the program expanded into 2 other launch sites with

# Payload vs. Launch Site



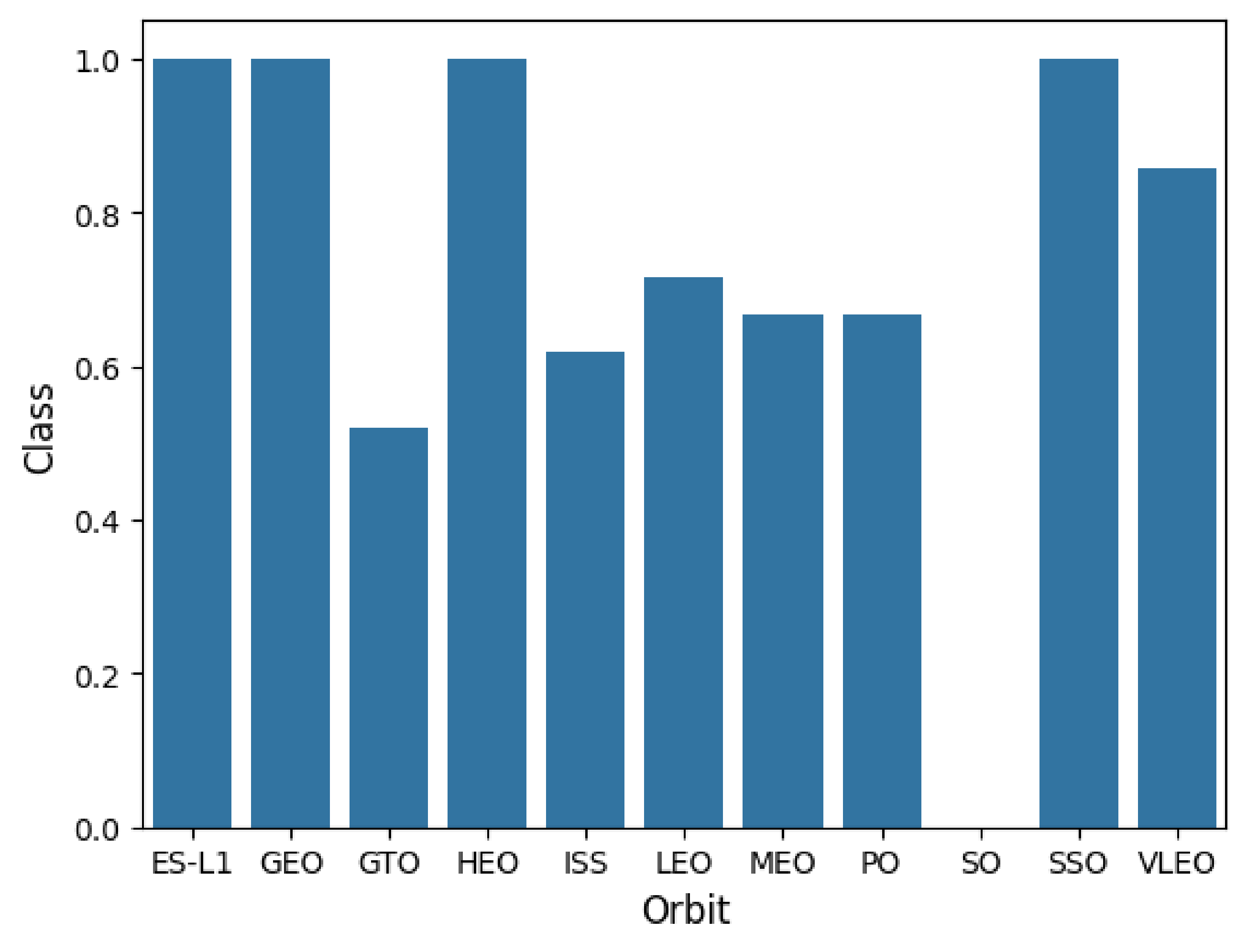

Observations:

- Launch Site VAFB-SLC had no heavy payload (over 10,000kg) launches.
- Most payloads were well below the 10,000kg heavy payload category
- Taken observations from the previous slide into account, it can be surmised that the heavy payloads were later in the program given their high rate of success. As the program matures, the payloads get heavier due to the successful landing of the stage 1 rockets.

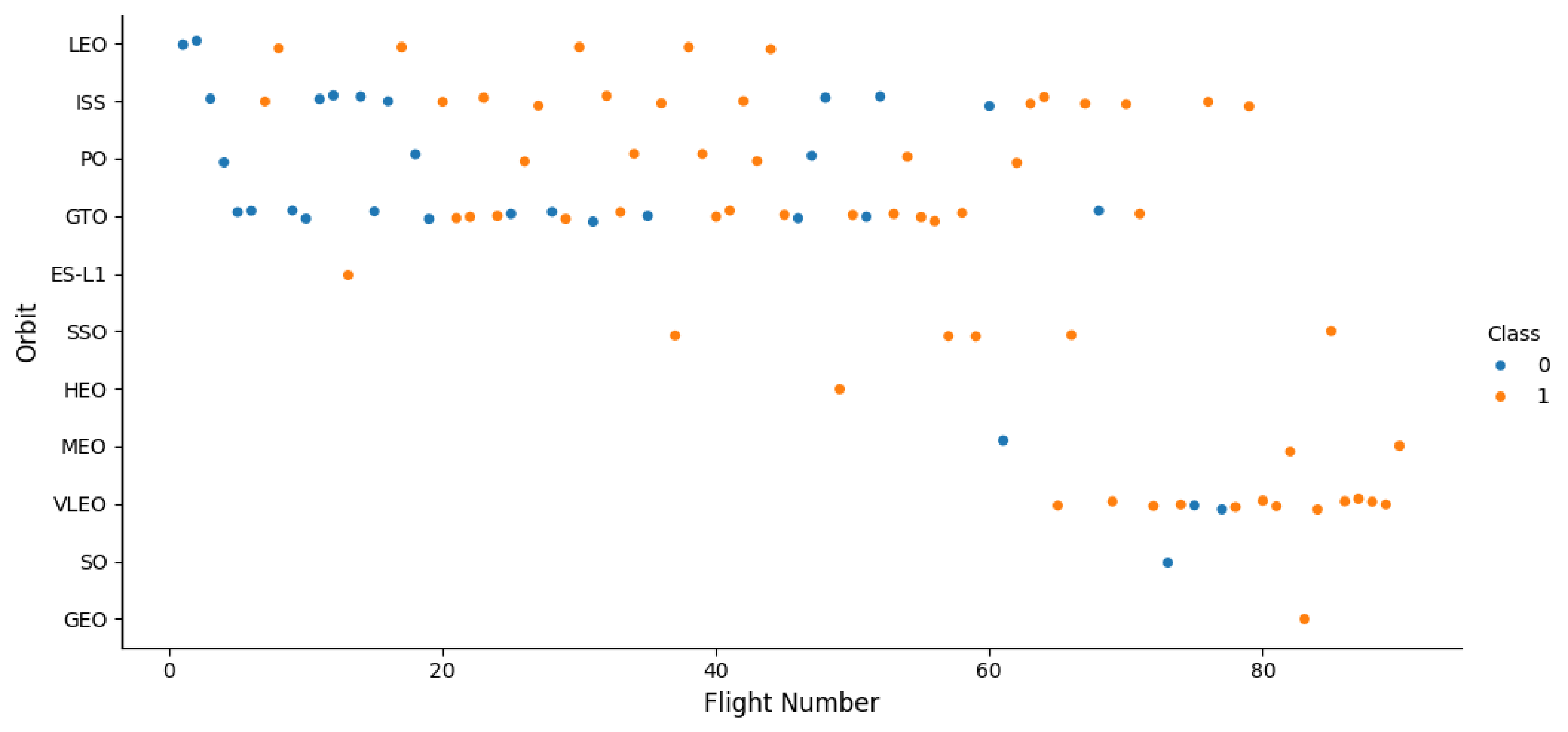
# Success Rate vs. Orbit Type

Observations:

- ES-L1, GEO, HEO, & SSO orbits all have 100% success rate. GEO, ES-L1, and HEO only had 1 flight each
- VLEO had 85% success
- SO had 1 flight and it was unsuccessful
- All other orbits ranged between 50 and 70% successful



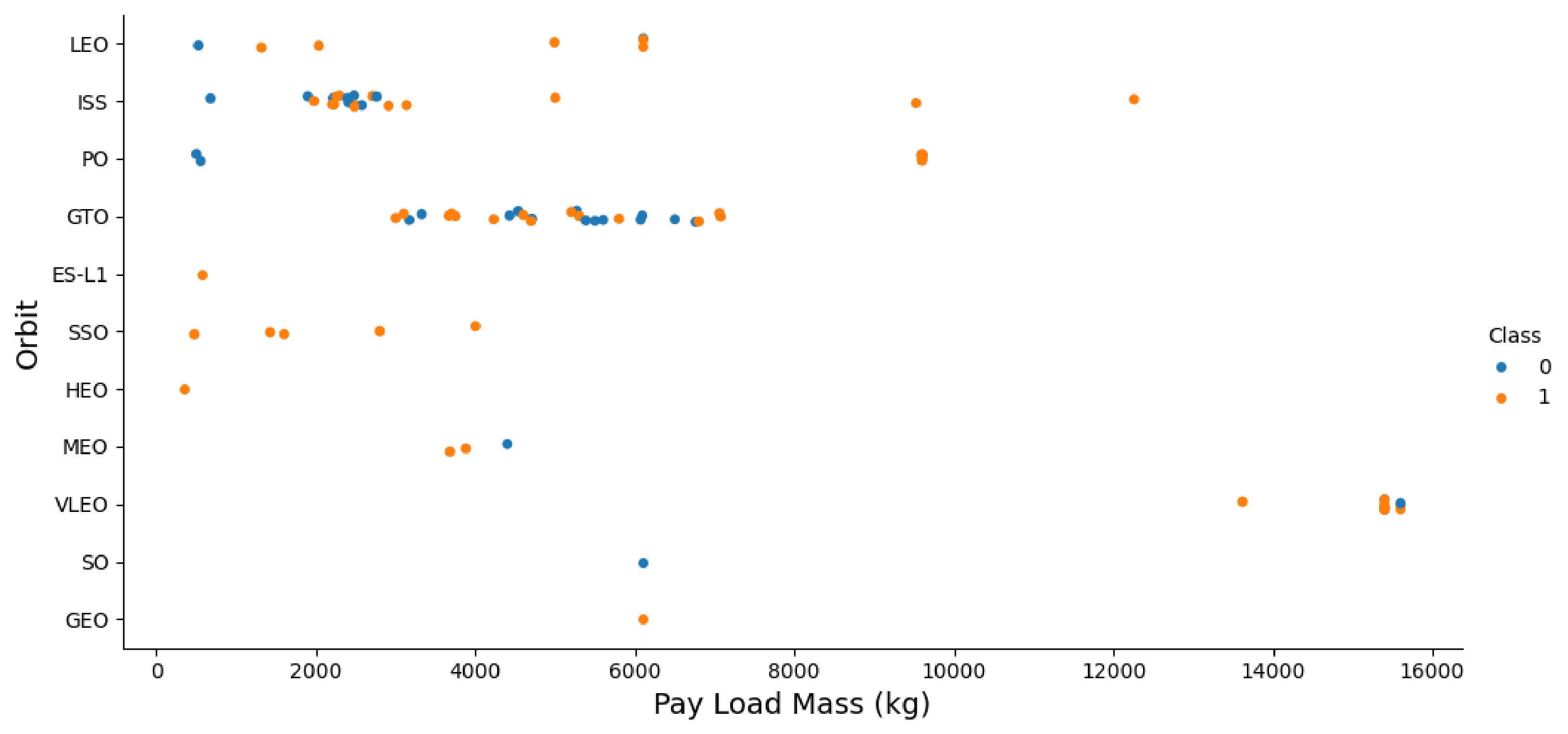
# Flight Number vs. Orbit Type




Observations:

- ISS and GTO had many of the first 25 flights.  Most of the early flights were not successful
- As observed in a previous slide, the increase in flights shows an increase in successful landings.  The same observation can be made hear, as the flights increase in each of the orbits, we observe an increase in successful landings

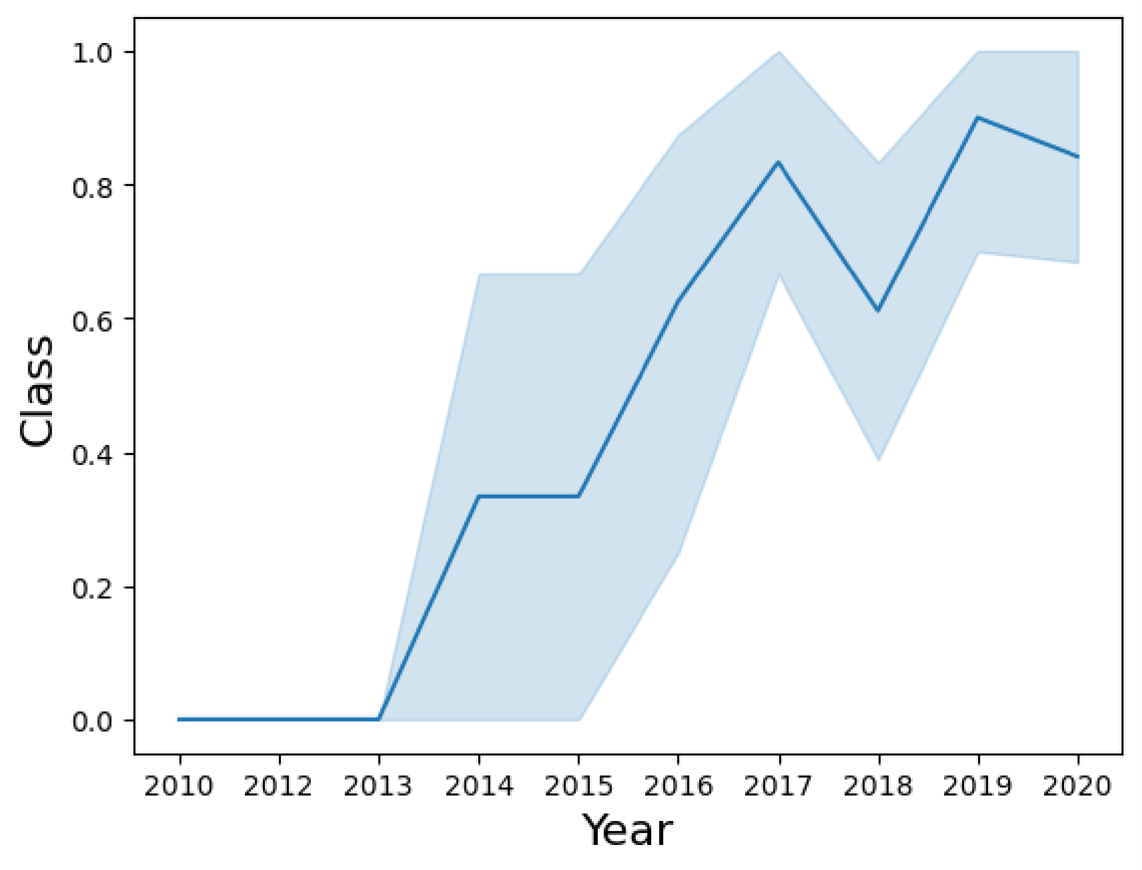
# Payload vs. Orbit Type




Observations:

- As observed in the previous slide, ISS and GTO had many of the first 25 flights which had a low success rate.
- As observed in an earlier slide, the increase in flights shows an increase in successful landings and the heavy payloads (greater than 10,000kg) took place later in the program and shows that the success rate is higher

# Launch Success Yearly Trend




Observations:

- Since 2013, the success rate has been on a relatively steady increase apart from a small drop in 2018 and 2020

# All Launch Site Names

```
%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 |

The SQL code in the screenshot above pulls the unique names of all the Falcon9 launch sites

# Launch Site Names 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 | Mission_Outcome | Landing_ |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (p |
| 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 (p |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | N |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | N |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | N |

The SQL code in the screenshot above pulls the first 5 records where the launch site name begins with “CCA”

# Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)';
```

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

| SUM("PAYLOAD_MASS__KG_") |
|---|
| 45596 |

The SQL code in the screenshot above pulls total payload mass carried by boosters launched by NASA (CRS)

# Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1';
```

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

| AVG("PAYLOAD_MASS__KG_") |
|---|
| 2928.4 |

The SQL code in the screenshot above pulls average payload mass carried by booster version F9 v1.1

# First Successful Ground Landing Date

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

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

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

The SQL code in the screenshot above pulls date of the first successful landing on a ground pad

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

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

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

The SQL code in the screenshot above pulls the names of the Booster Version with successful landings on a drone ship

# Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

| Mission_Outcome | COUNT("Mission_Outcome") |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

The SQL code in the screenshot above pulls the total number of successful and failed mission outcomes.

- Please not that the underlying data had 1 record with a space at the end of the word success in the record which has caused the duplicate “Success” outcome.

# Boosters Carried Maximum Payload

```
%sql SELECT "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 |

The SQL code in the screenshot above pulls a list of booster versions which carried the maximum payload mass.

# 2015 Launch Records

```
%sql SELECT substr("Date",6,2) as Month,substr("Date",0,5) as Year, "Landing_Outcome","Booster_Version", "Launch_Site" FROM SPACEXTABLE \
    WHERE "Landing_Outcome" LIKE 'Failure (drone ship)' AND substr("Date",0,5) = '2015';
```

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

| Month | Year | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|---|
| 01 | 2015 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | 2015 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

The SQL code in the screenshot above pulls a table of failed drone ship landings. It pulls the month, year, landing outcome booster version and launch site.

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

```
%sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") as Count_of_Outcomes FROM SPACEXTABLE WHERE \
    Date between '2010-06-04' and '2017-03-20' GROUP BY "Landing_Outcome"  ORDER BY COUNT("Landing_Outcome") DESC;
```

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

| Landing_Outcome | Count_of_Outcomes |
|---|---|
| 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 |

The SQL code in the screenshot above pulls a table ranking the landing outcomes between 2010-06-04 and 2017-03-20, grouped by landing outcome, in descending order.

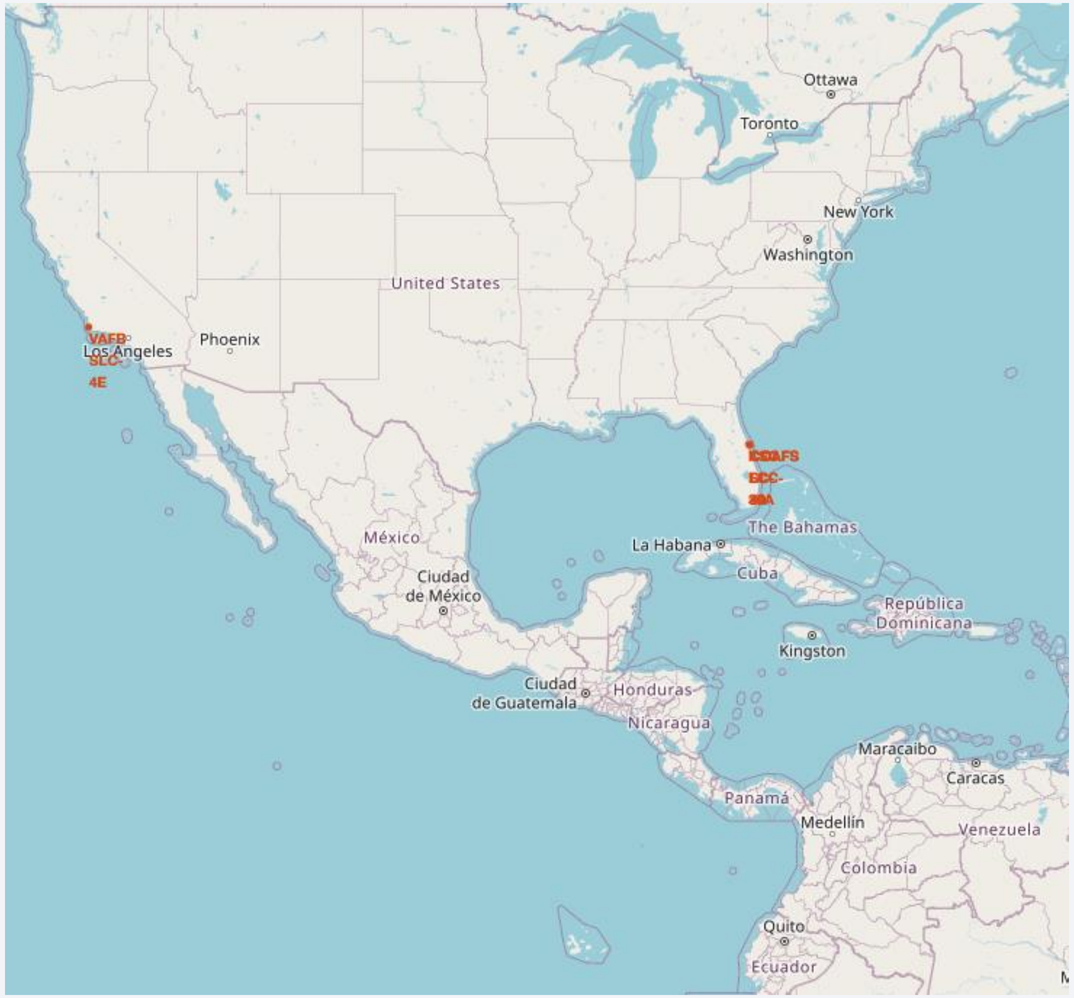
Section 3

# Launch Sites Proximities Analysis

# All Launch Sites Locations on a Global Map

Observations:

- Launch sites are generally closer to the equator
    - Launches use the rotation of the earth to assist in the launch since it is rotating at approximately 1K/mph.  When rocket is launched eastward this acts like a boost which helps reduce the amount of fuel needed to get to orbital speed.
- Launch sites are close to the coastline and far from urban areas
    - Due to the noise and the speed at which rockets go, keeping them away from urban areas and next to coasts helps reduce chances of accidents involving citizens.

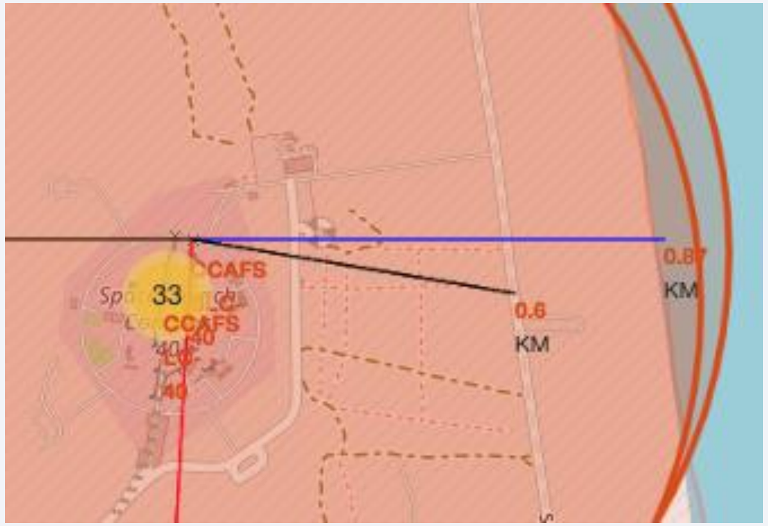


# Color-Labeled Launch Sites

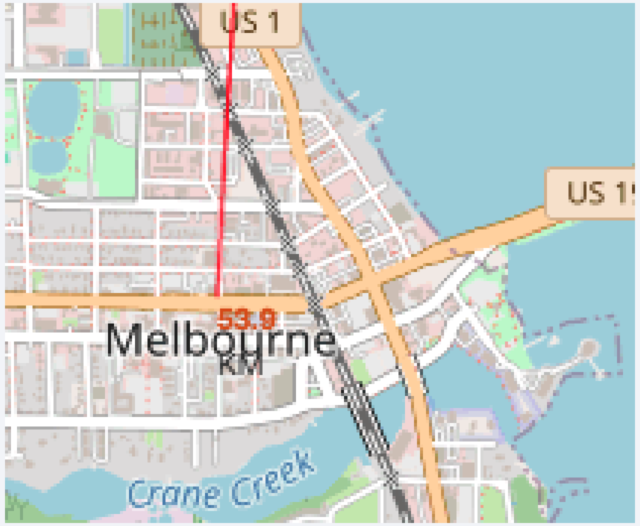
Observations:

- Using Launch Site CCAFS SLC – 40, we plotted color-labeled markers showing launches taken
  - Green = Successful
  - Red = Failure
- The interactive nature of this map allows the user to click to compare results. There were 7 launches from this specific site and 3 had successful landings

# Distances From Launch Sites







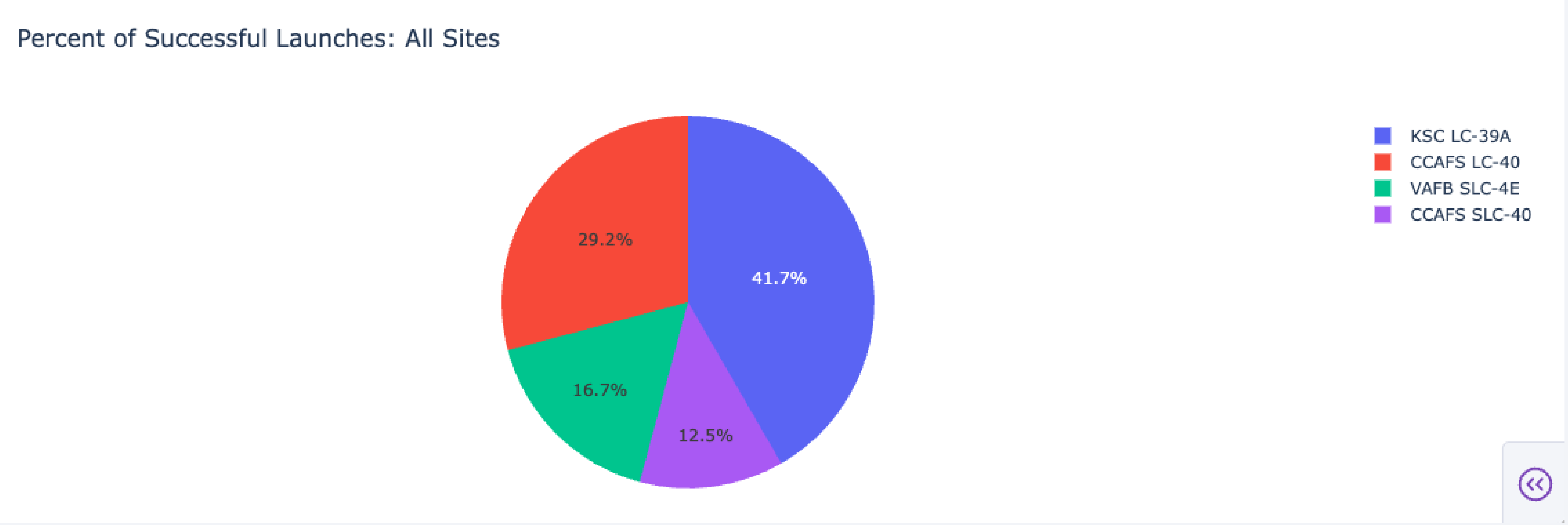

Observations:

- Using CCADS SLC-40 as the example launch site we plotted the following distances:
  - Coastlines, highways, and railways are relatively close at 0.87km, 0.60km, and 22.1km, respectively
  - Cities and large metro areas are further away as in the screenshot with Melbourne being 53.9km away
- Proximity of highways and railways are close to allow for transportation of cargo for the launches whereas the proximity of coastlines and cities are for safety. If something goes wrong with the launch, it is more likely going to land in the ocean than in a highly populated area.

Section 4

# Build a Dashboard with Plotly Dash

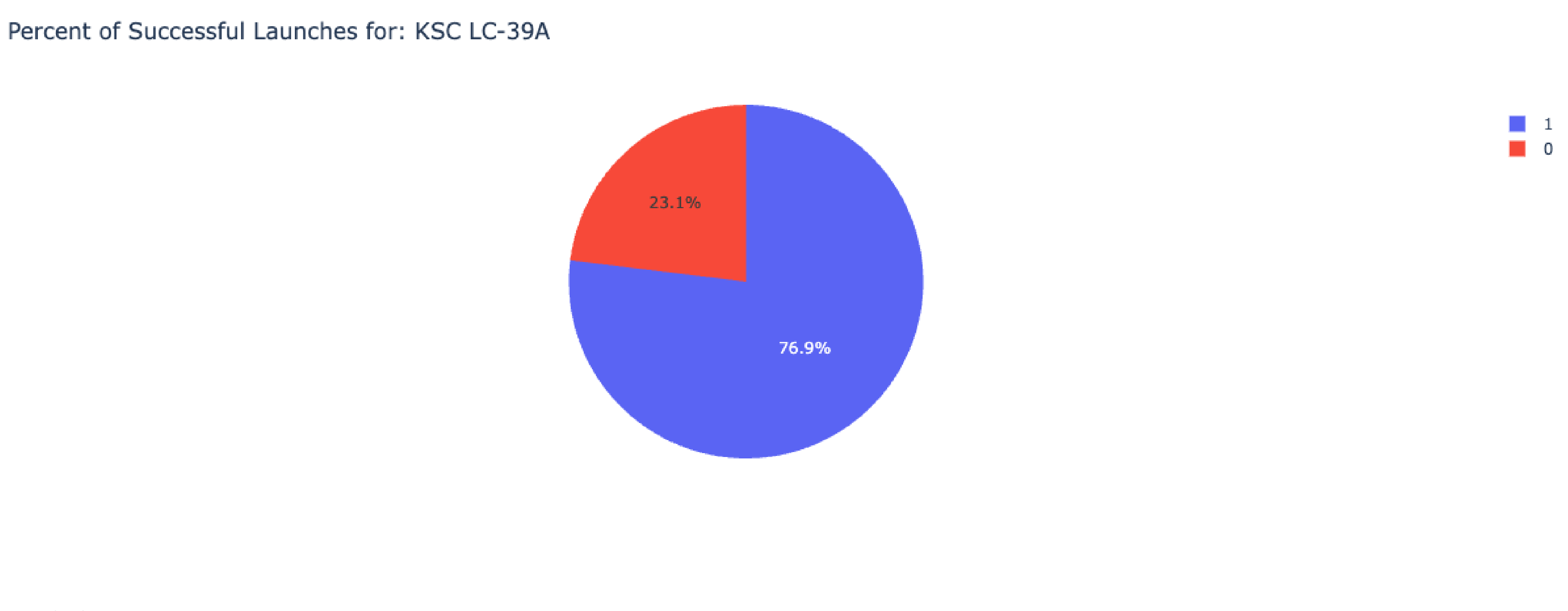
# Percent of Successful Launches:  All Sites




Observations:

- The chart above is showing that site KSC LC-39A has the highest success rate.

# Percent of Successful Launches: KSC LC-39A




Observation:

- The chart show the percentage breakdown of the launches from KSC LC-39A. With 10 successful launches, the site has a 76.9% success rate

# Payload Mass vs. Launch Outcome:  All Sites

Observations:

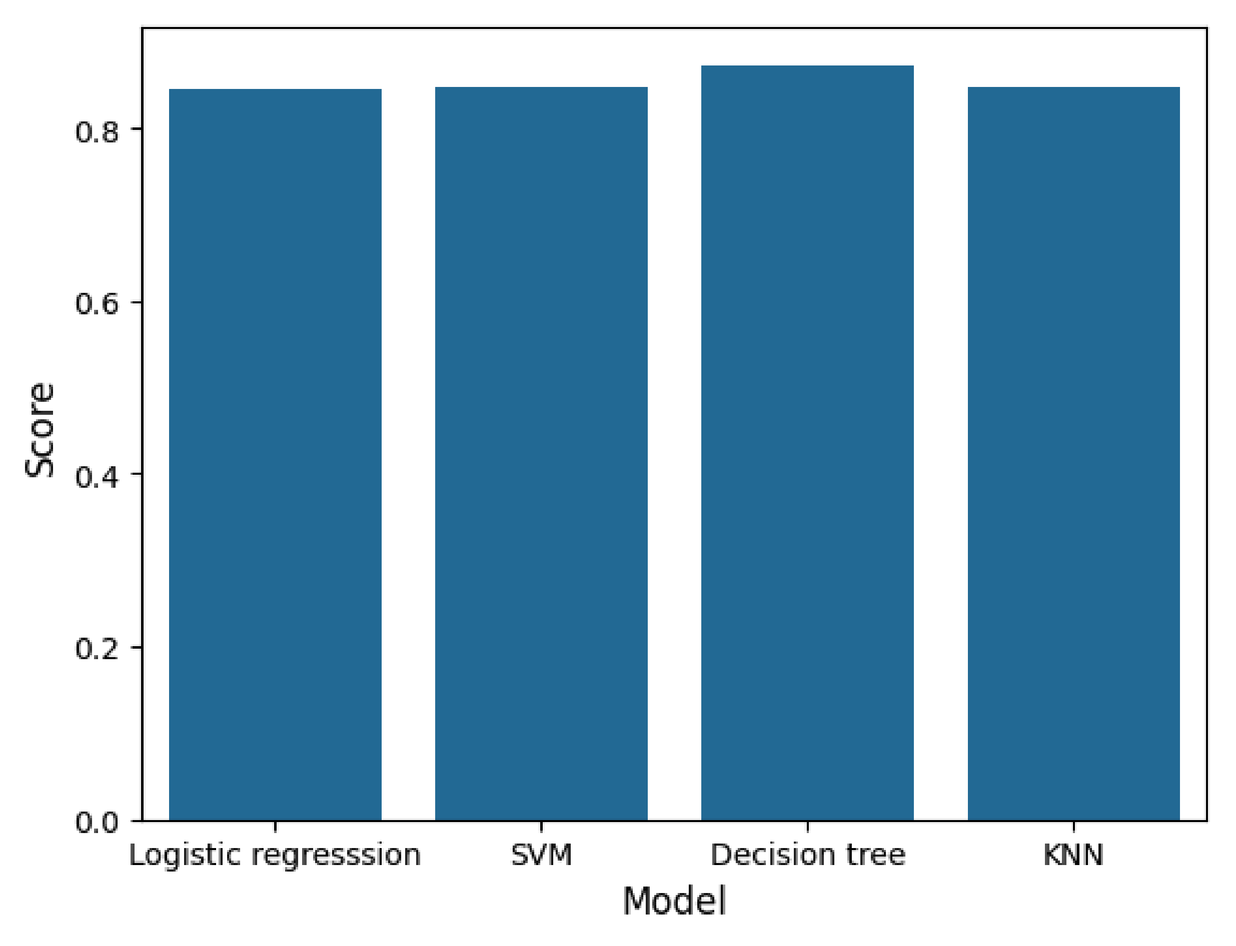
- The charts show that the highest success rate based on payload mass happened between 2500 and 7500 kg


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

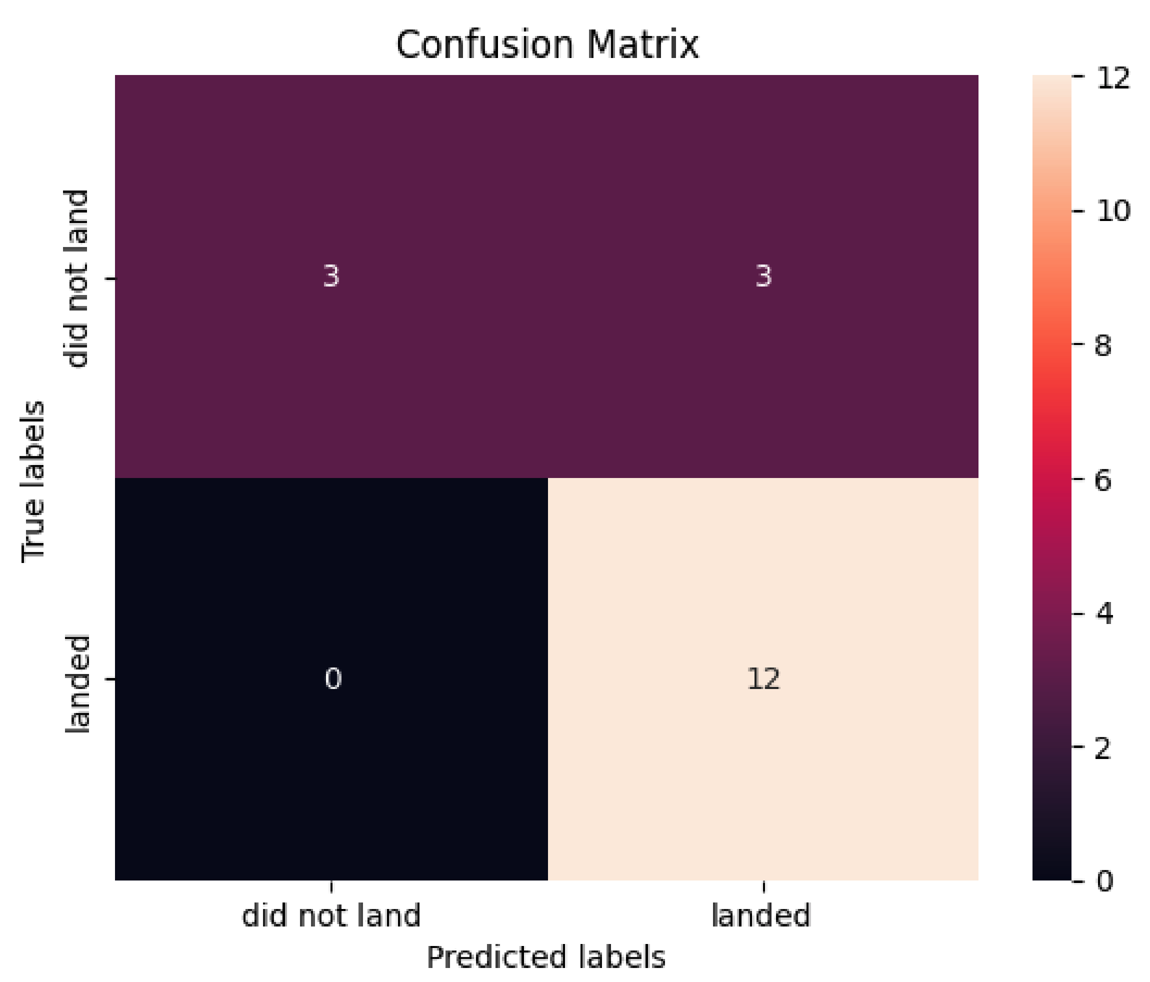
Based on the chart to the right, the Decision Tree has a slight advantage.



# Confusion Matrix

Observations:

- As the matrix shows the model is showing 3 false positives, meaning the model predicted 3 landed when they did not land.



# Conclusions

- The success rate for launches has increased over time
- Most launch sites are near the equator but away from populated areas
- There are fewer heavy payloads greater than 10,000 kg and most launches were between 2,500 and 7,500 kg.
- Looking at one or two variables or rates and drawing a conclusion can cause us to overlook important information. The data shows that the program has become more successful over time. Looking at something like success rate of one site over the another without looking at launch volume and when the site started being utilized could cause an incorrect conclusion.

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!