IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christina B.
April 2023

Outline

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

Executive Summary

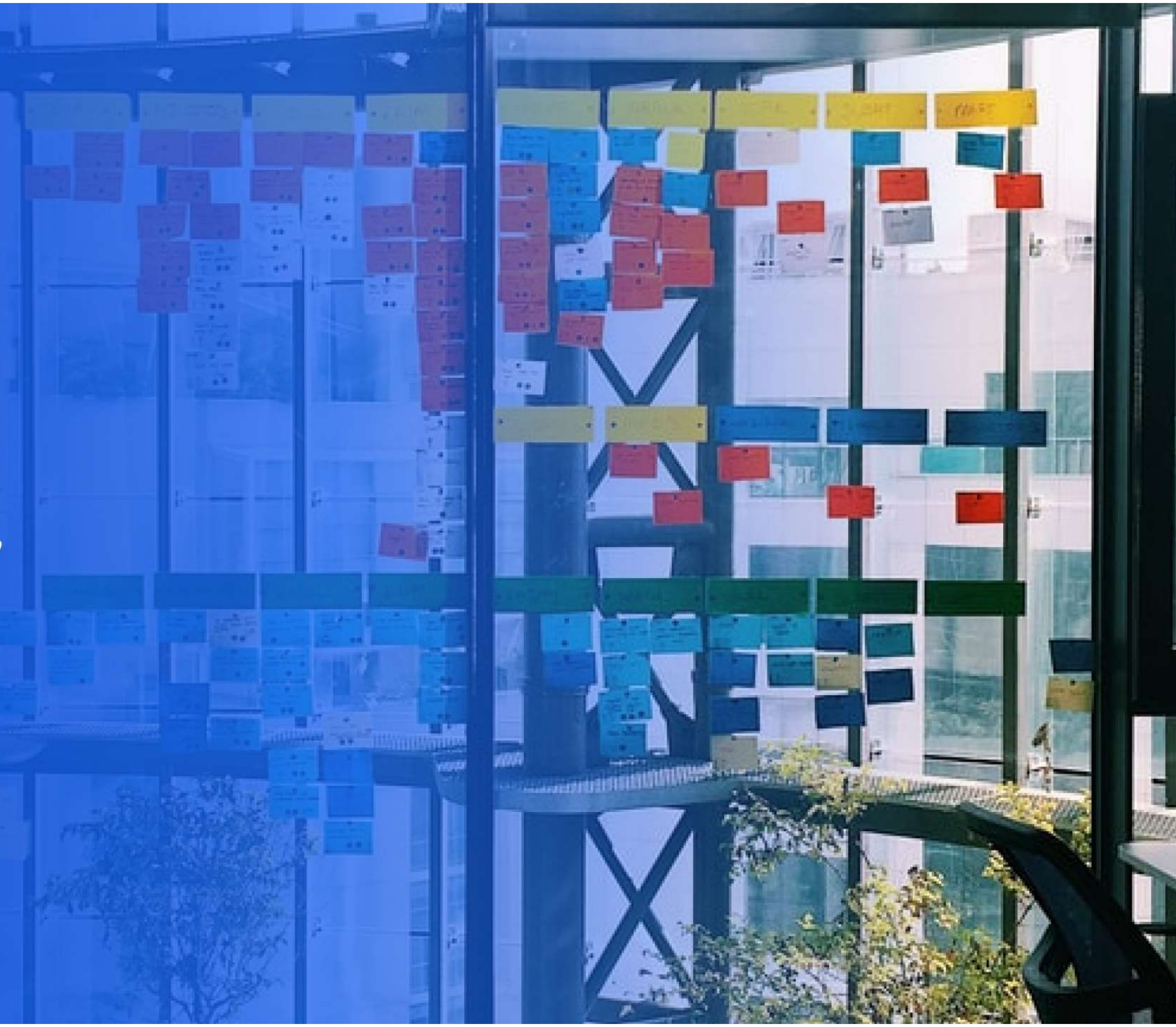
- Summary of methodologies
- Summary of all results

Introduction

- SpaceY is looking to compete against SpaceX, a leader in the commercial space age
- SpaceX (Falcon9)'s leading edge is cost effective launches by reusing the first stage of their rockets when possible
- Various mission parameters (payload, orbit, customer) help indicate whether the first stage will be reused
- By analyzing SpaceX's launches, we will determine if they will resuse the first stage of the Falcon9 launch and its cost

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

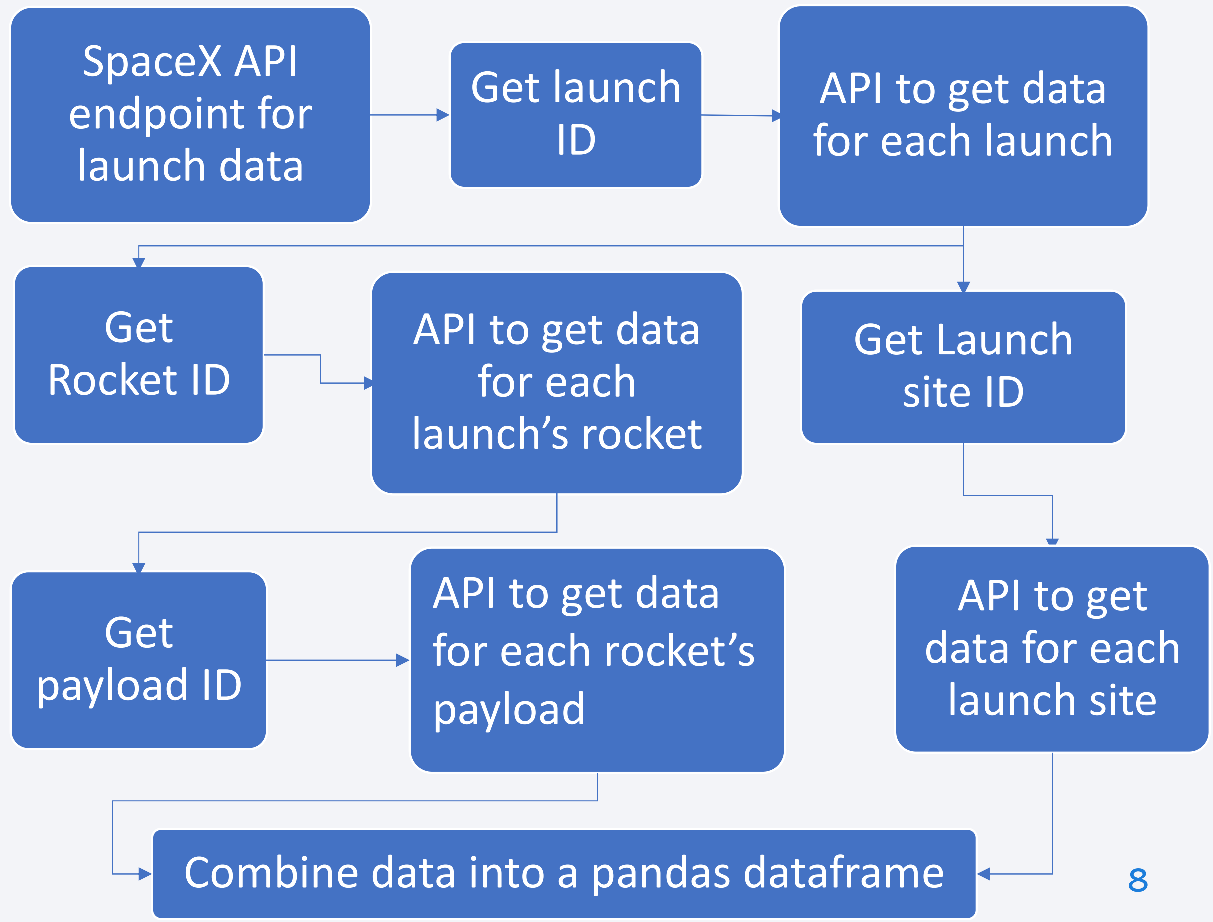
- Rocket launch data was collected from SpaceX API using the url:

  http://api.spacexdata.com/v4/launches/past

Request	Request the data from the API
Decode	Decode contents as a JSON
Gather Additional Data	Use the launch ID's to obtain additional data
Combine	Combine columns into a dictionary
DataFrame	Create a pandas DataFrame from the dictionary
Filter	Filter to only include Falcon9 launches
Replace NaNs (Data Wrangle)	Replace missing values with average values of that column

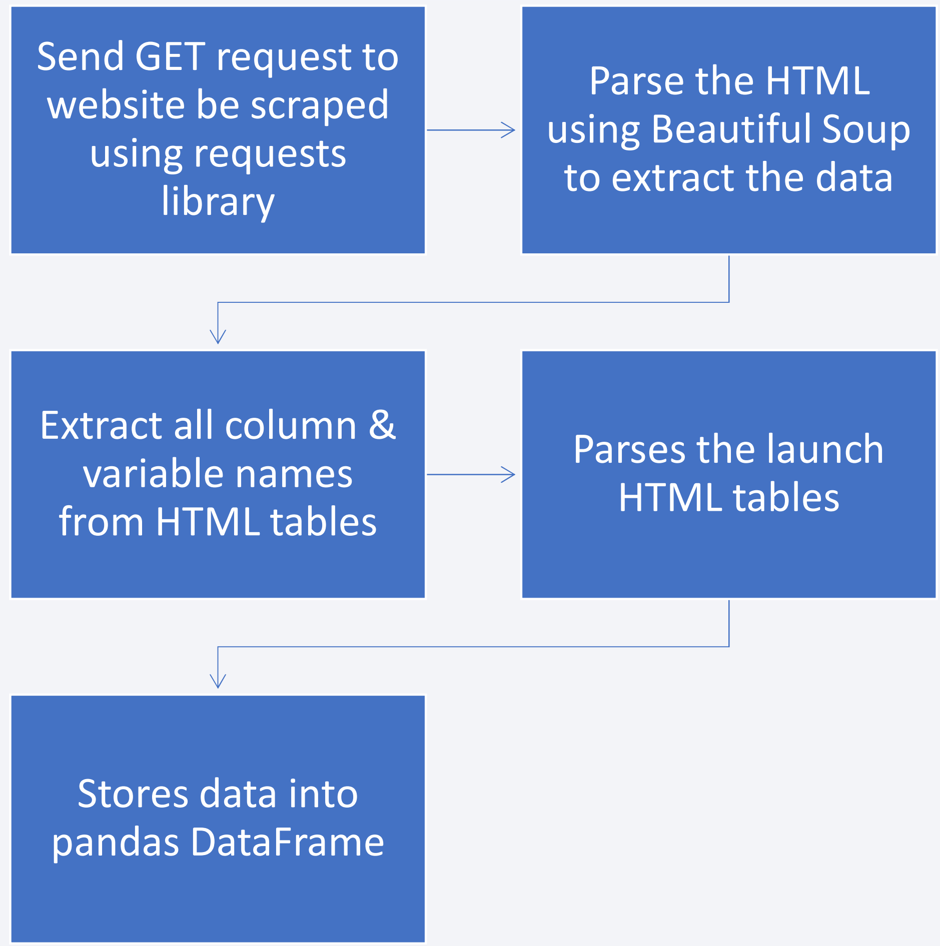
Data Collection – SpaceX API

- Data was collected using a SpaceX API use requests.get() in python and then cleansed
- For details see: https://github.com/CB-git-bit/DataScience_Learning/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



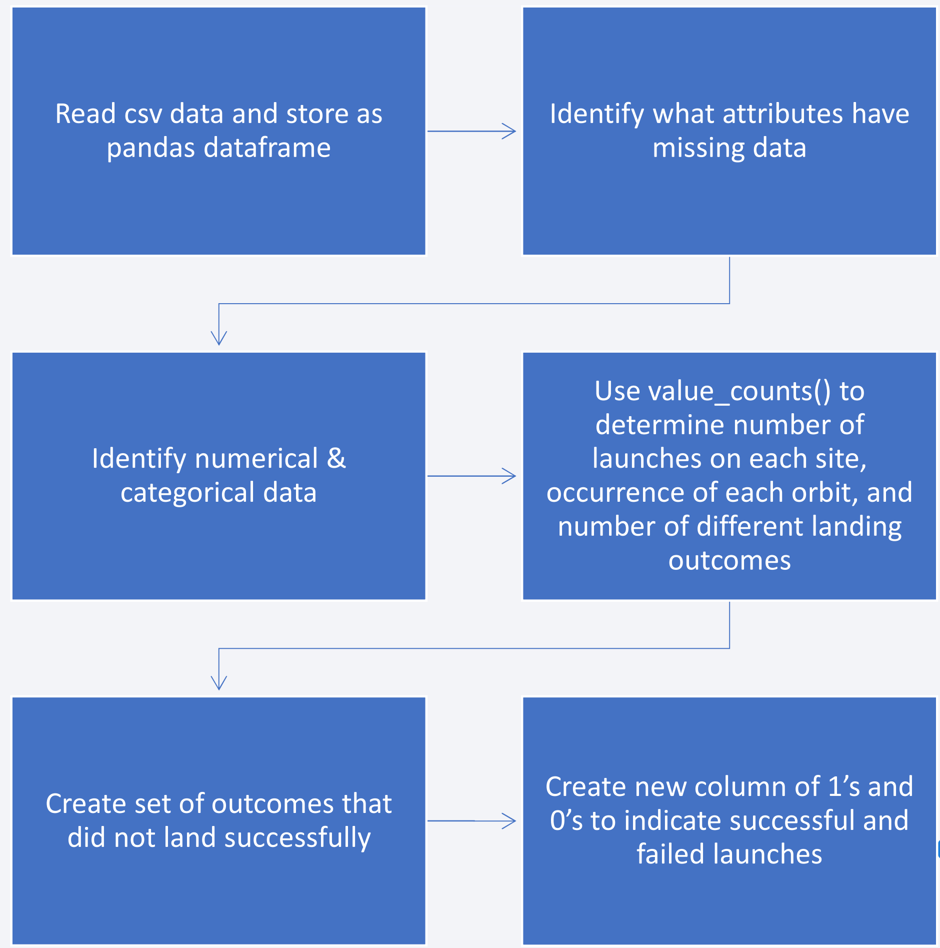
Data Collection - Scraping

- Data is scraped from Wikipedia and stored in a pandas dataframe
- It is then cleaned, visualized, and exported to a CSV for further analysis
- For more details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



Data Wrangling

- Data is first explored by identifying where there is missing data and what are the data types
- Data is processed by classifying orbit types to determine success of landings, which will be used as training labels later
- For more details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Scatter plots were made of the following relationships

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

For more details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Various SQL Queries were performed including but not limited to:

- SELECT – to specify which columns to choose
- DISTINCT – to ensure uniqueness of values
- WHERE – to apply a filter to the data
- GROUP BY – to group rows with the same values into summary rows
- ORDER BY – to sort the results
- SUM(), AVG(), MAX() – to find the sum, maximum value, and average of a column
- COUNT() – to count the number of rows
- LIKE() – to specify a substring

For details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

The following map objects were added to folium map to visually represent the launch sites and provide additional information:

- Markers – each marker represents the longitude & latitude of a launch site and indicate if the launch was successful (green) or failed (red).
- Circles – represent the range of each launch site
- Lines – shows distance between launch sites and proximities (e.g. coastlines, cities, etc.)

For details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

A dashboard was created and includes the following features:

- Line chart – show the success rate of SpaceX launches over an user-specified year range
- Pie chart – show the success/failure rate of SpaceX launches over a user-specified year range
- Dropdown menu – input method for the user to specify a launch site
- Scatter plot – show the relationship between payload mass and mission outcome. The user can hover over each data point to see additional information

These plots and interactions were added to drill down the analysis to individual launch site and to see how different parameters impact success

For details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four Classification methods were developed with optimized parameters using GridSearchCV to determine the best prediction method.

Get Data
- Get X and Y values from two different csvs and store as pandas dataframes

Transform the data (Preprocessing)
- Transform Y values into numpy array
- Transform X values into Standard Scalar

Split Data
- Split the data into training and testing sets

Find optimal parameters
- Use GridSearchCV to find optimal parameters for each of the 4 classification methods

Test Accuracy
- On each method, test accuracy using score
- Show confusion matrix

Determine best performing method

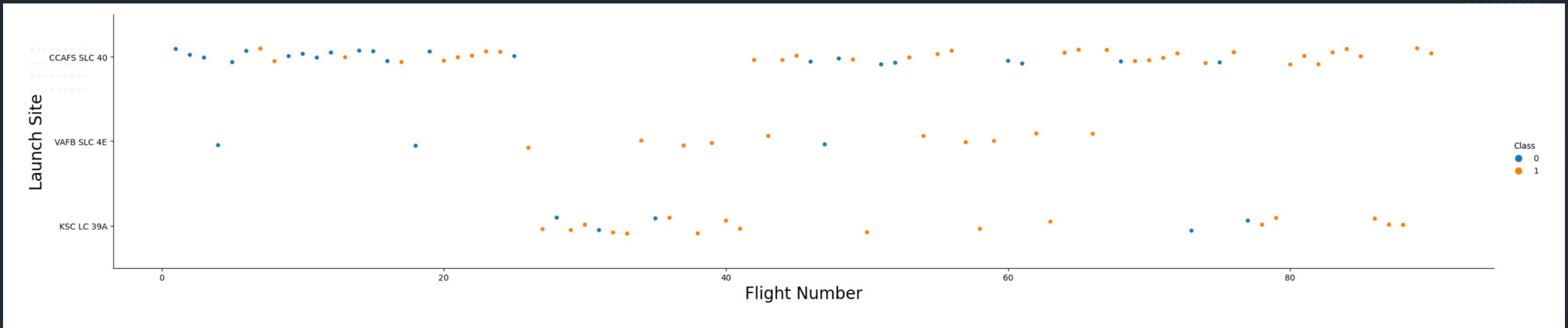
- For details: https://github.com/CB-git-bit/DataScience_Learning/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

1. Exploratory Data Analysis (EDA) Results & Insights Drawn
2. Launch Site Proximitries Results & Analysis
3. Dashboard Results
4. Predictive Analysis (Classification)

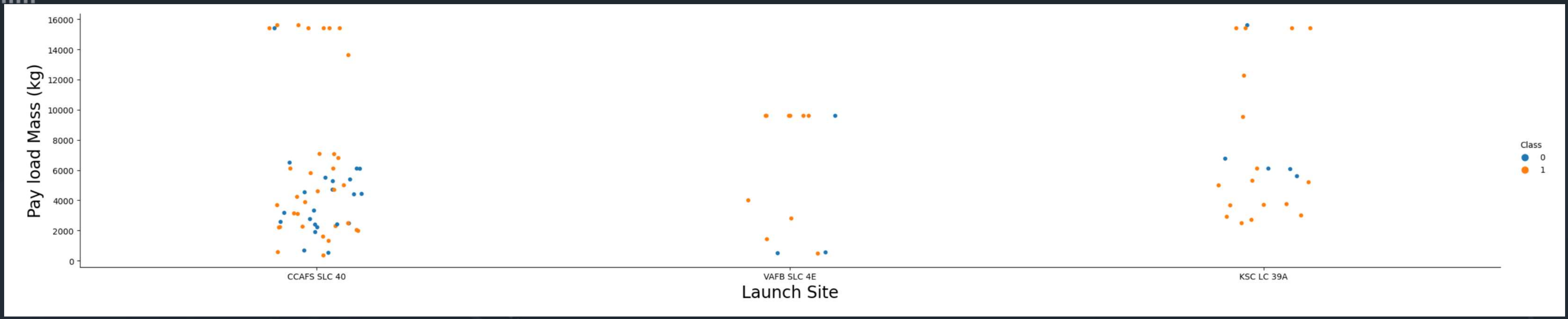
Section 2

Insights drawn from EDA



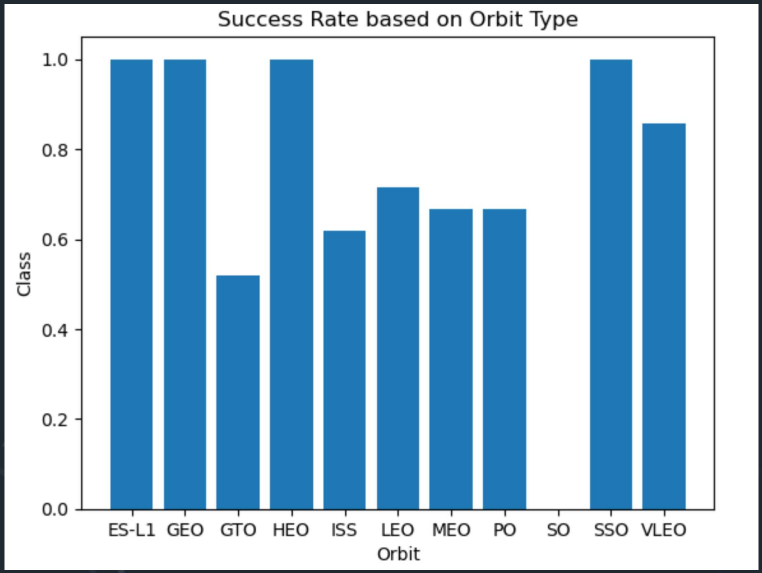
Flight Number vs. Launch Site

- This plot shows where each of the flights were launched and their success by colour
- The increase in data points from left to right indicate an increase in the frequency of SpaceX launches over time
- SpaceX uses 3 main launch sites for their missions (y-axis), with the most popular one being CCAFS SLC 40
- Failed launches occur more often in the lower flight numbers for each of the three launch sites



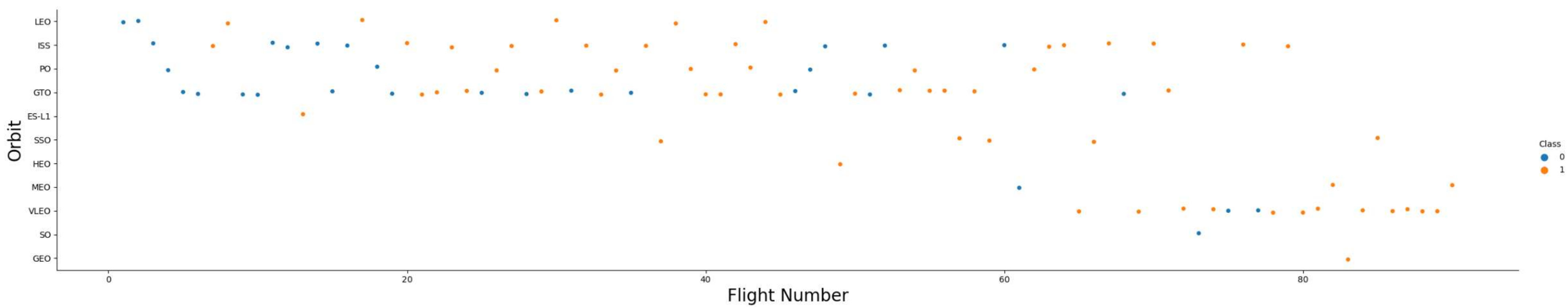
Payload vs. Launch Site

- The scatter plot shows the distribution of the payload weights [kg] for the three different launch sites
- Each data point indicates mission success by colour. Blue for failure and orange for success
- VAFP-SLC launch site has no rockets launched for heavy payloads (10000+)
- Heavier payloads tend to be more successful for all launch sites



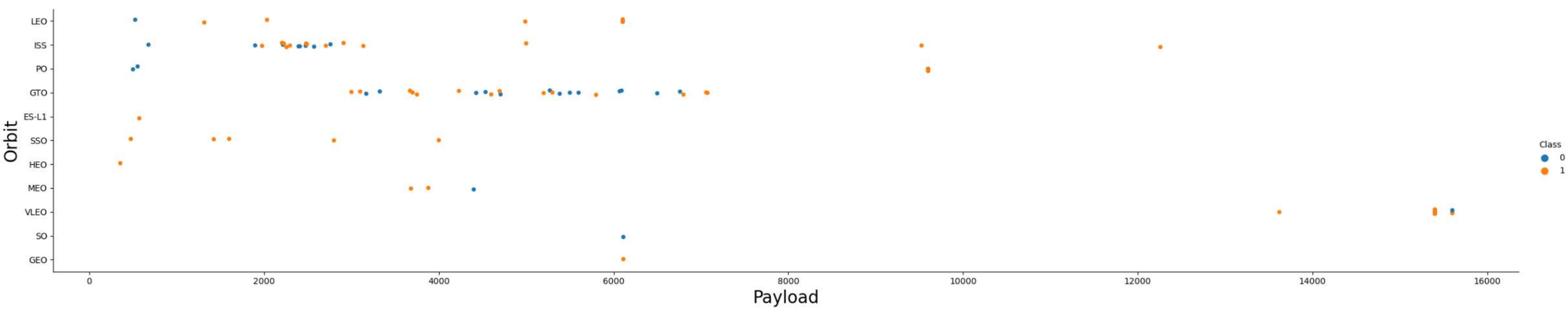
Success Rate vs. Orbit Type

- The success rates (y-axis) of SpaceX launches are shown in the bar chart for different types of orbits (x-axis) such as Low Earth Orbit (LEO), Geostationary Transfer Orbit (GTO) and Polar Orbit (PO)
- ES-L1, GEO, HEO, and SSO orbits had the most success
- SO orbits had no success



Flight Number vs. Orbit Type

- The scatter plot shows the distribution of the orbit types for different flight numbers
- As flight number increases, there are more successful missions
- Show a scatter point of Flight number vs. Orbit type
- More recent flights (i.e. higher flight numbers) are more successful and have different orbits than the more common orbit
- LEO orbit success appears to be related to number of flights
- GTO orbit success appears independent of flight number

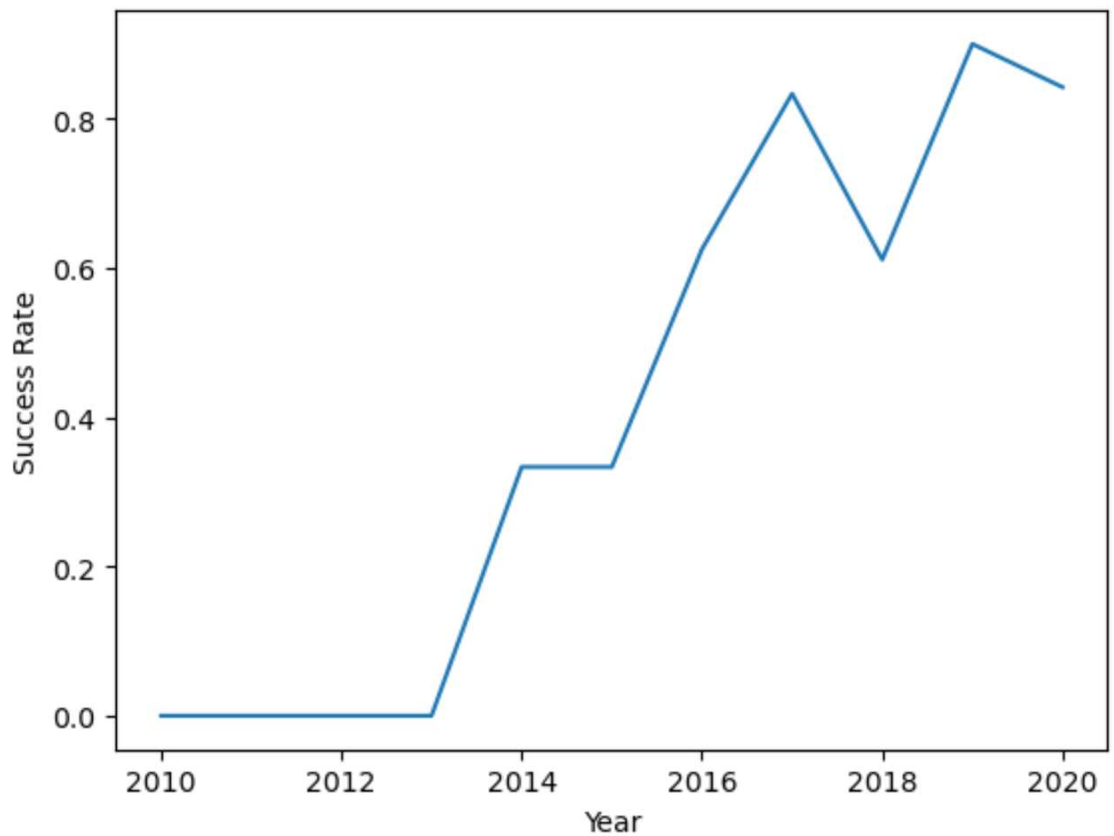


Payload vs. Orbit Type

- The scatter plot shows the distribution of payload weights for different orbit types with the success of launches indicated by orange.
- Successful and heavy payloads are most prominent in Polar, LEO, and ISS orbits
- SSO orbit does not have payloads greater than 6000, but all of its launches have been successful

Launch Success Yearly Trend

- This line chart shows an overall upward trend in successful missions over time
- There is only one dip in 2018, indicating there were mainly failed missions of that year



All Launch Site Names

- The table outlines all the unique launch sites in the dataset
  - It was obtained using the SELECT DISTINCT functions in SQL

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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 (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- This is the data of launches from sites CCAFS LC-40 or CCAFS SLC-40
- All these launches had orbit type LEO and did not include any successful missions

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
48213

- Total payload mass carried by boosters launched by NASA (CRS) is 48 213 kg
- This was achieved by summing the payload mass (kg) column for entries that have CRS in customer

Average Payload Mass by F9 v1.1

- The average payload mass by F9 v1.1 is 2928.4
  - This was calculated by averaging the entries in the column Payload Mass (kg) where the Booster Version is F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first date of successful landing outcome on ground pad occurred on 02-03-19
  - This is the lowest valued date in the table where there is a successful landing outcome

MIN(Date)
02-03-2019

Successful Drone Ship Landing with Payload between 4000 and 6000

- Unique booster versions were called from the table where the payload mass (kg) is between 4000 and 6000 and had success in drone ship
- These are all F9 boosters

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

Total Number of Successful and Failure Mission Outcomes

- 98 successful missions and 1 failed missions
- This is a

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome Like 'Success'
```

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

COUNT(*)
98

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%';
```

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

COUNT(*)
1

Boosters Carried Maximum Payload

- The boosters with maximum payload mass are shown
- These are unique values, which indicate that the max payload for the booster version was only once

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

month_name	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

2015 Launch Records

- These are the booster version, launch site, and landing outcome for launches in 2015
- In 2015, all launches were from CCAFS LC-40 and failed

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

- 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 were ranked
- Outcomes were inconclusive due to technical difficulties

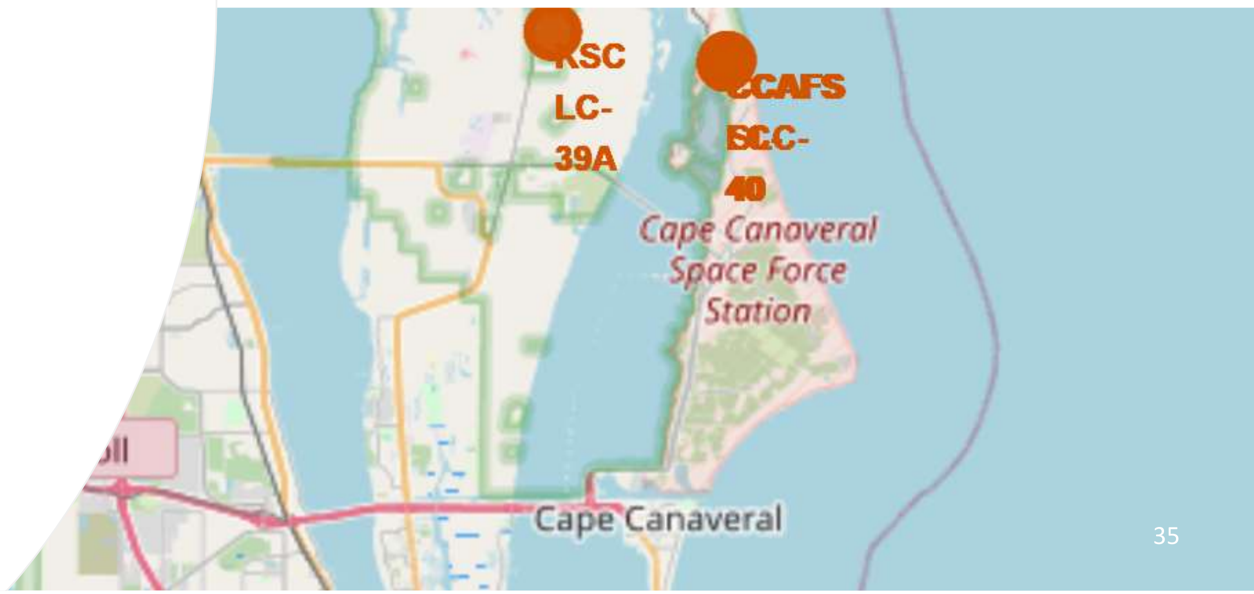
```
%sql SELECT "Landing _Outcome", COUNT(*) AS count_success FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' AND "Landing _Outcome" LIKE
```

Section 3

Launch Sites Proximities Analysis

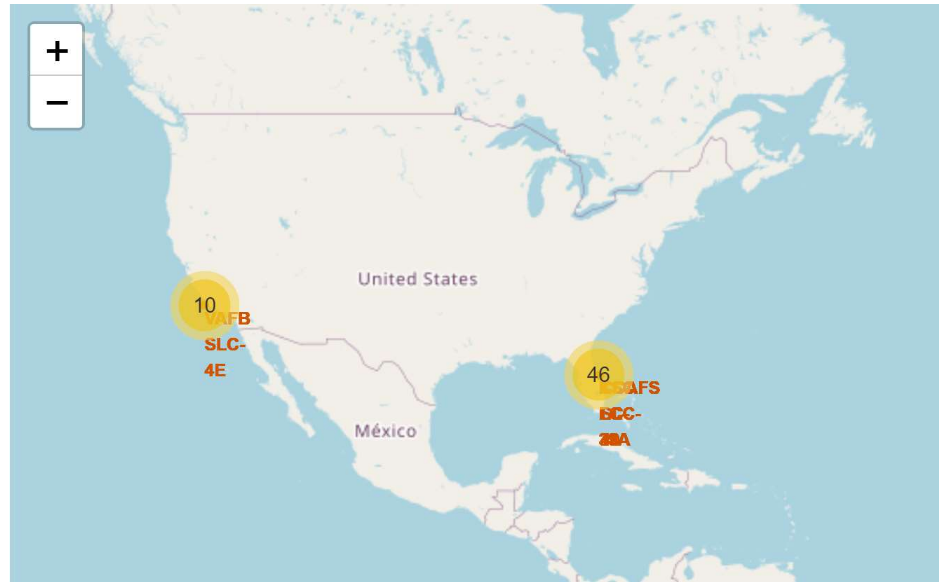
Map of all Launch Sites

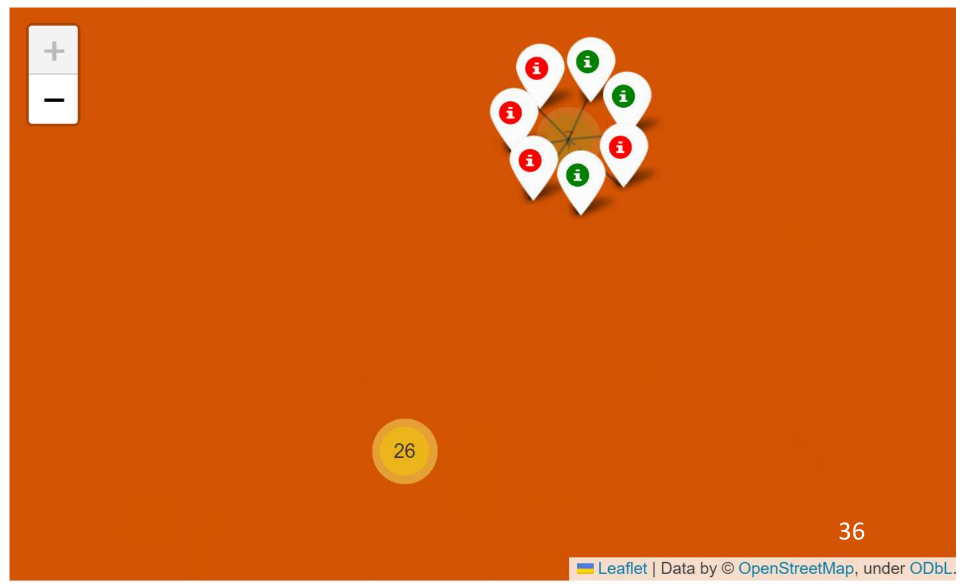
- Most launches are in Florida as shown in the photos on the right
- There is also a site in California
- All sites are based out of the USA



Map of Launch Sites with Success Indicators

- Replace <Folium map screenshot 2> title with an appropriate title
- The map now indicates how many launches are in each location
  - 10 on the West Coast and 46 on the East
- As you zoom in, each launch has its own pin that is coloured according to mission success
  - Green is successful and red is failure
  - Representative photo is of launch site CCAFS SLC-40





Map with Distance Marker

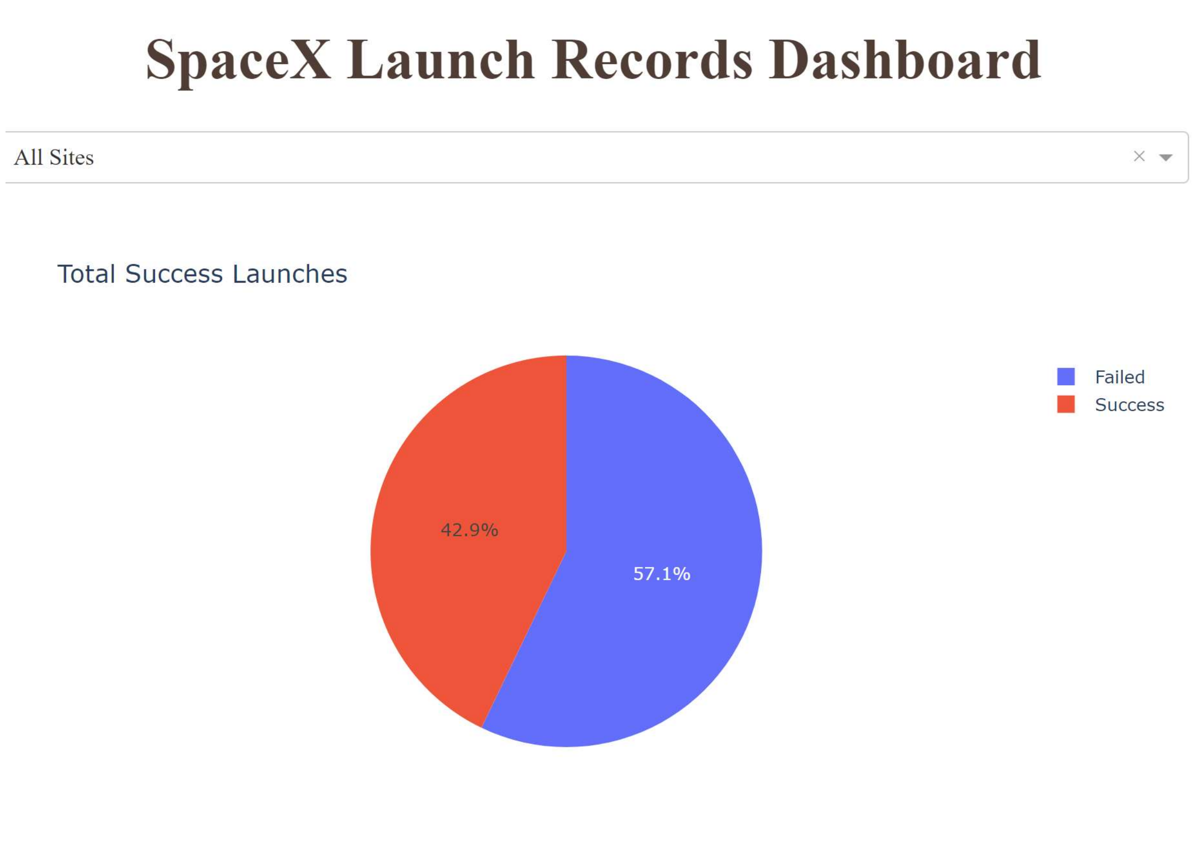
- A line was added to the map from CCAFS SLC-40 to the coast of KSC LC-39A launch site
- This distance is nearly 3 miles

Section 4

Build a Dashboard with Plotly Dash

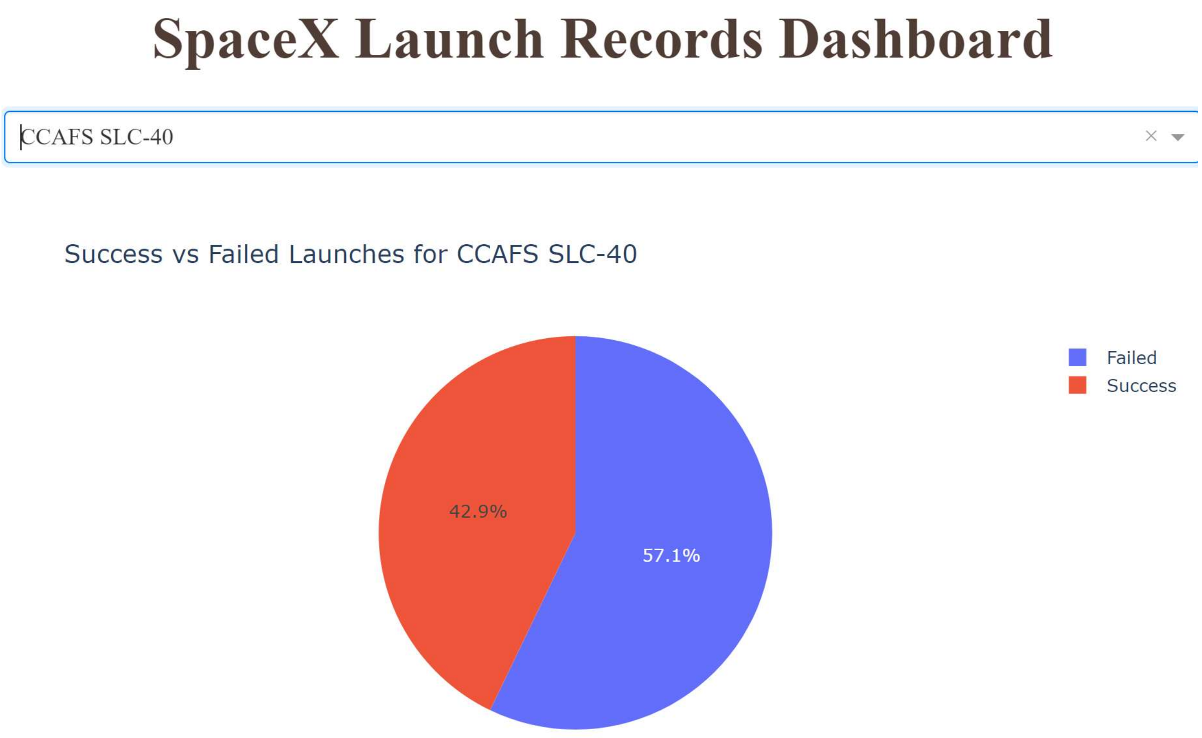
Dashboard: Launch Success for all sites

- The launch success is shown in the pie chart as a percentage of the all the launches SpaceX has made across all its sites, represented in red
- This chart shows that most launches were not successful



Dashboard: Most Successful Launch Site

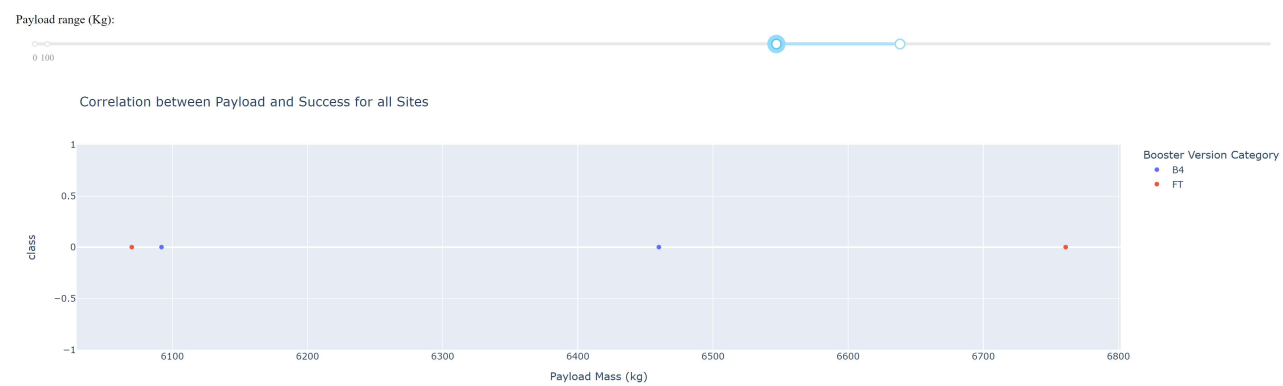
- CCAFS SLC-40 has the highest launch success ratio at 42.9%
- This is the highest launch success ratio compared to all other launch sites and is the same as the overall success ratio across all launch sites
- Despite having the most successful launches out of all the launch sites, CCAFS SLC-40 still has more failed launches



Dashboard: Payload vs Success Rate for All Sites

- A majority of V1.1 boosters are class 0 (failed) launches and a majority of FT boosters are class 1 (success) launches
- The payloads of failed launches spans 0 to 6500kg at somewhat regular intervals
- The payloads of successful launches mainly span 1900kg to 5300kg
- All launches with payloads in the 6000-6800 kg range were unsuccessful across all sites and for all boosters

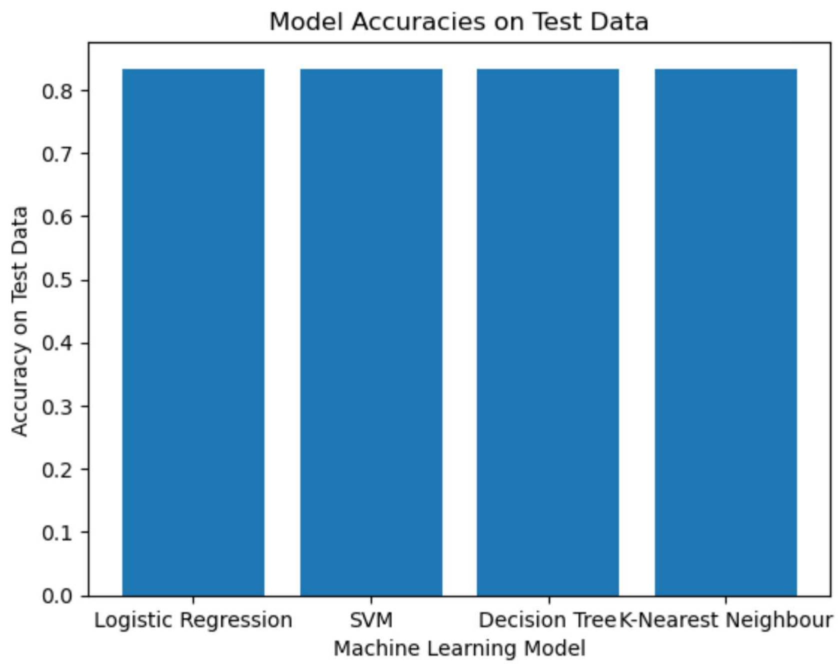
Payload Success Rate for all sites and all payloads

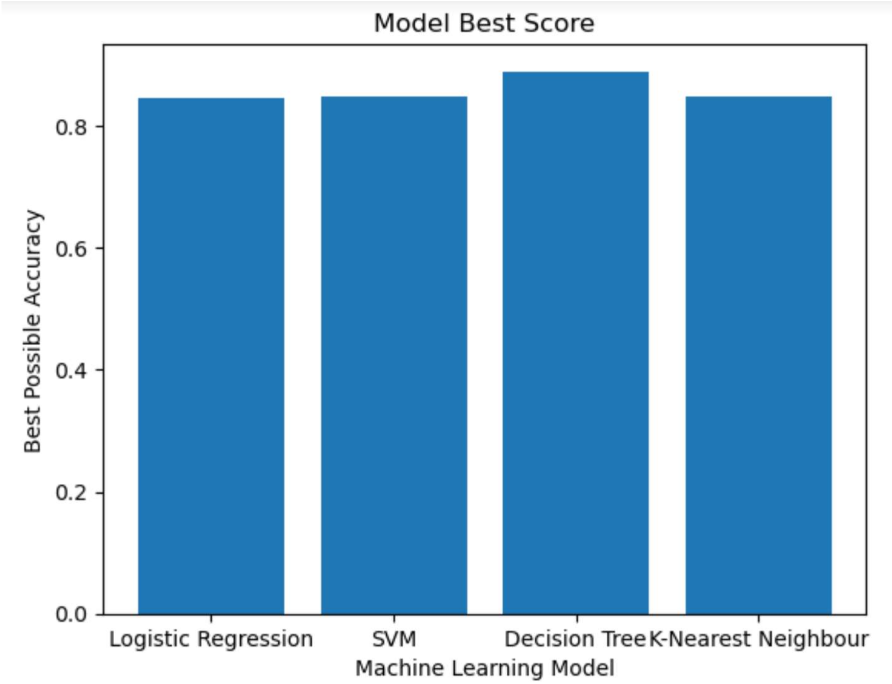


Payload Success Rate for all sites and all payloads in the 6000 to 6800 kg range

Section 5

Predictive Analysis (Classification)



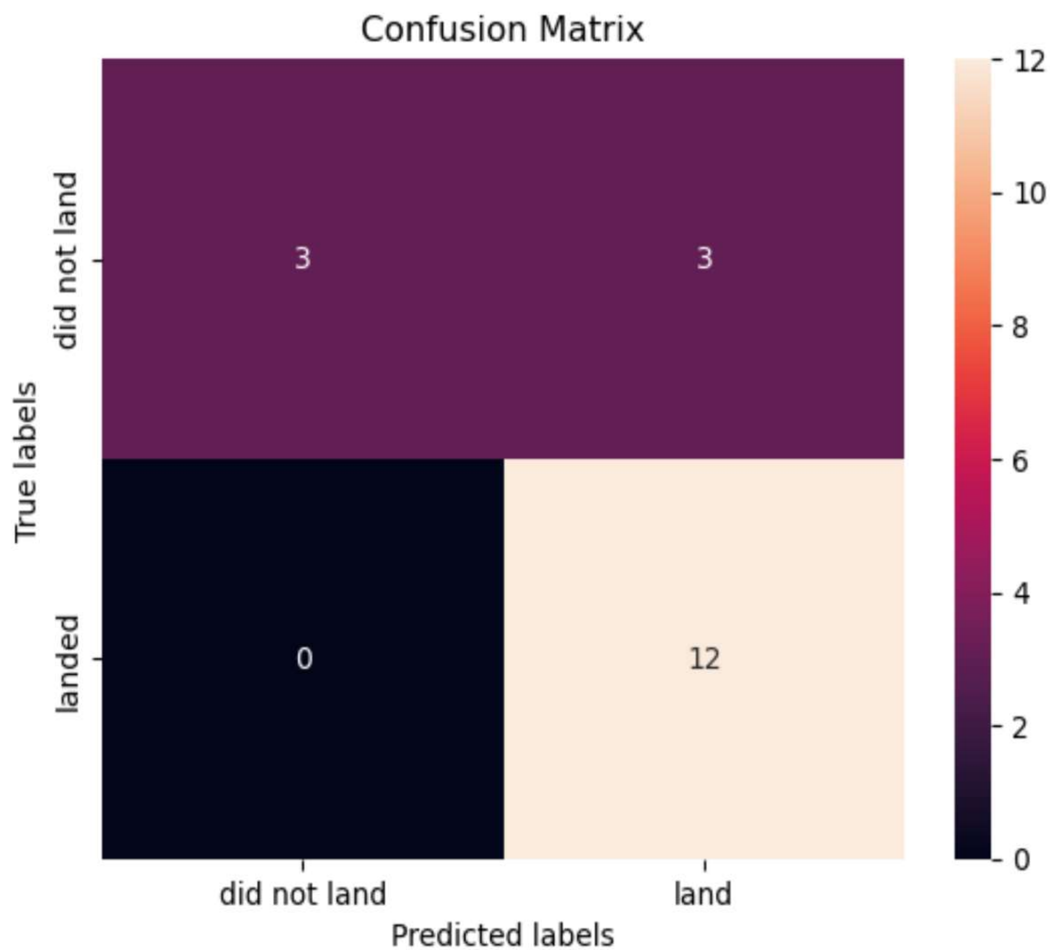


Classification Accuracy

- All models have the same accuracy when tested on the test data as shown in the bar graph on the left
    - This is expected because they all have similar best performances and model accuracy was calculated on the same dataset with the same train_test_split() and the same random see value
- Best scores for the models are shown in the bar graph on the right
    - Decision Tree has the highest accuracy in this method
- Thus Decision Tree has the highest possible accuracy

Confusion Matrix

- This is the confusion matrix of the Decision Tree model which had the best score and a high accuracy on the tested dataset of 83.3%
- There are 0 false negatives and 12 true negatives
- The false and true positives are equal at 3
- This is an confusion matrix, with minimal false negatives, maximum true negatives, and somewhat equal true & false positivess



Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

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!