IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anton L.
2025-06-02

Outline

Executive Summary

- **Methodology**
  - Data on Falcon 9 first stage landings from 2010 to 2020 was obtained from a publicly accessible API (https://api.spacexdata.com/), which is not officially affiliated with SpaceX, as well as from publicly available information on Wikipedia (https://en.wikipedia.org/wiki/SpaceX). Additional datasets were provided as part of the course.
  - The data cleaning and preprocessing process involved extracting landing outcome information to be used as the target variable for machine learning models.
  - SQL queries and a range of visualizations—including static charts, interactive maps, and a dynamic dashboard—were used to explore the dataset and uncover key insights.
  - Predictive modeling was carried out using several classification algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN).

- **Results**
  - The dataset on SpaceX Falcon 9 first stage landings contains information such as flight number, launch date, payload mass, orbit type, launch site, and mission outcome.
  - For prediction tasks, Logistic Regression, SVM, and KNN models demonstrated similar levels of performance on this dataset.

Introduction

- Competing with SpaceX, a rival rocket launch company seeks to predict the success of Falcon 9 first stage landings.
- Key questions to investigate include:
  - What type and scope of data are available regarding Falcon 9 first stage landings?
  - Which machine learning model offers the highest accuracy in predicting landing outcomes for future launches?
  - Can we accurately forecast whether a future Falcon 9 first stage landing will be successful?

Section 1

Methodology

Methodology

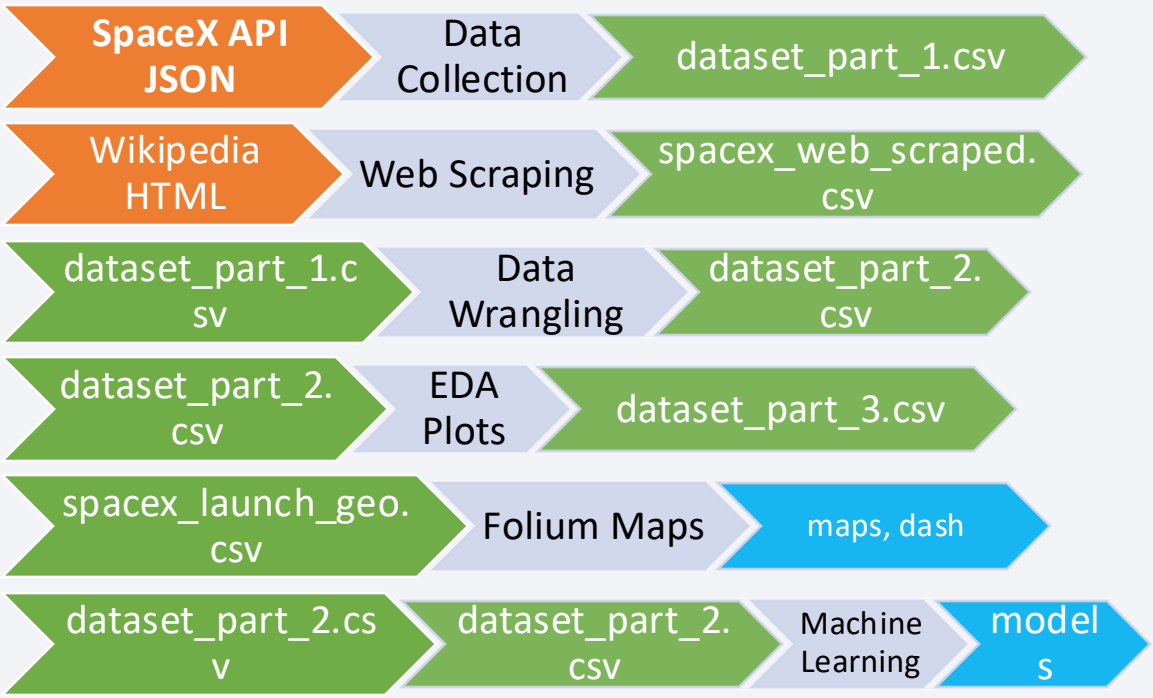
Executive Summary

- Data collection methodology:
  - Information on SpaceX Falcon 9 first stage landings was gathered from a publicly available API not associated with SpaceX, as well as from a Wikipedia article. Supplementary datasets in CSV format were also provided as part of the course materials. escribe how data was collected.
- Perform data wrangling
  - The data was cleaned and preprocessed to prepare it for visualizations, SQL queries, and training machine learning models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis was conducted using classification techniques implemented through machine learning models.

Data Collection

- The datasets were gathered from the following sources:
  - An IBM-hosted copy of launch data retrieved from a publicly available API in JSON format.
  - A static version of a Wikipedia page containing launch data in HTML tables (as of the June 9, 2021 revision).
  - Additional CSV-formatted datasets supplied as part of the course materials.

Flowchart Data Collection



Data Collection – SpaceX API

- SpaceX data API https://api.spacexdata.com/
- Data was extracted from the response from the API and loaded into a Pandas DataFrame for further analysis
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/20bd13eb1a1e77aac47797bbc76dfc65aa3afa32/collecting_the_data.ipynb

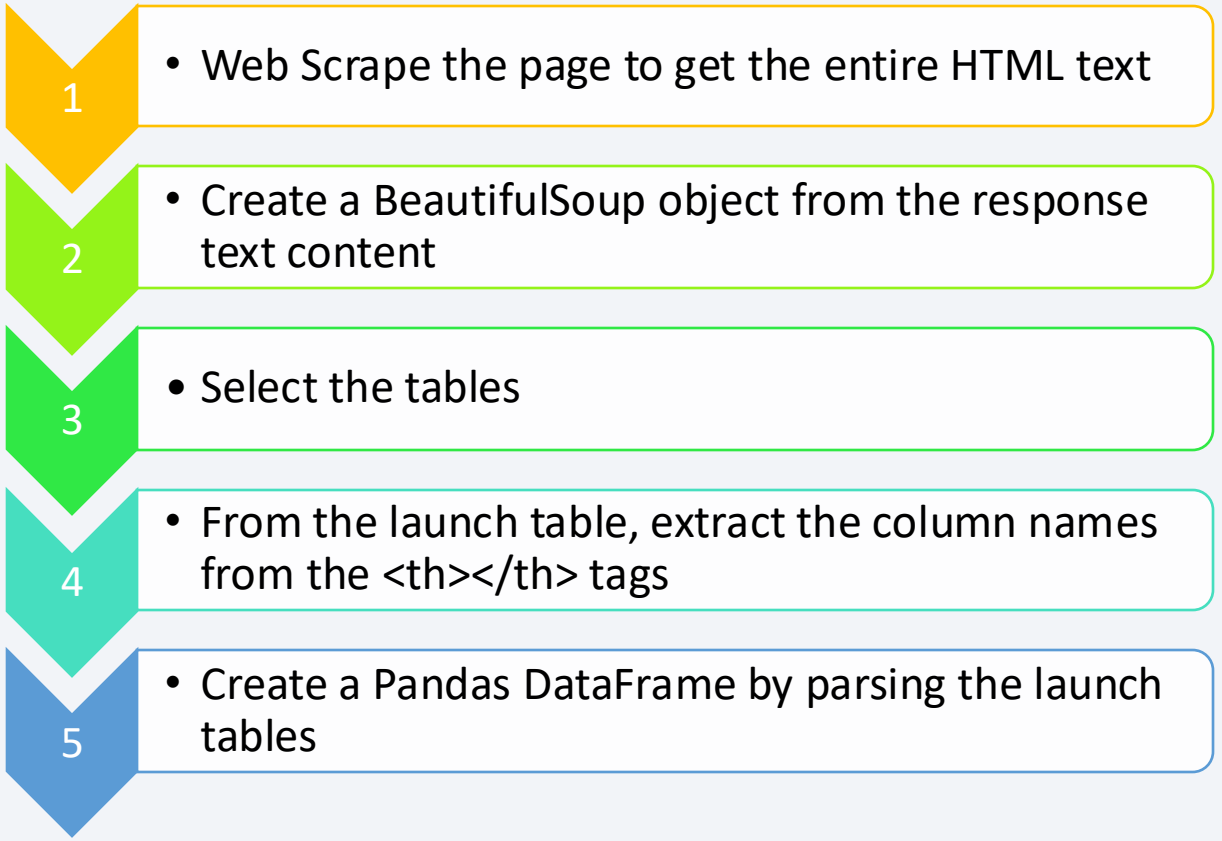
Flowchart of SpaceX API calls

1. Send GET request to API
2. Extract nested data
3. Convert date format
4. Use defined functions to generate specific columns of data
5. Combine separate columns into a DataFrame
6. Filter out all launches with rockets other than the Falcon 9
7. Handle missing values

Data Collection - Scraping

- SpaceX launch data was scraped from HTML tables on a permanently-linked copy of the SpaceX Wikipedia webpage https://en.wikipedia.org/wiki/SpaceX.
- Data was extracted from the tables and loaded into a Pandas DataFrame for further analysis
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/20bd13eb1a1e77aac47797bbc76dfc65aa3afa32/web_scraping.ipynb

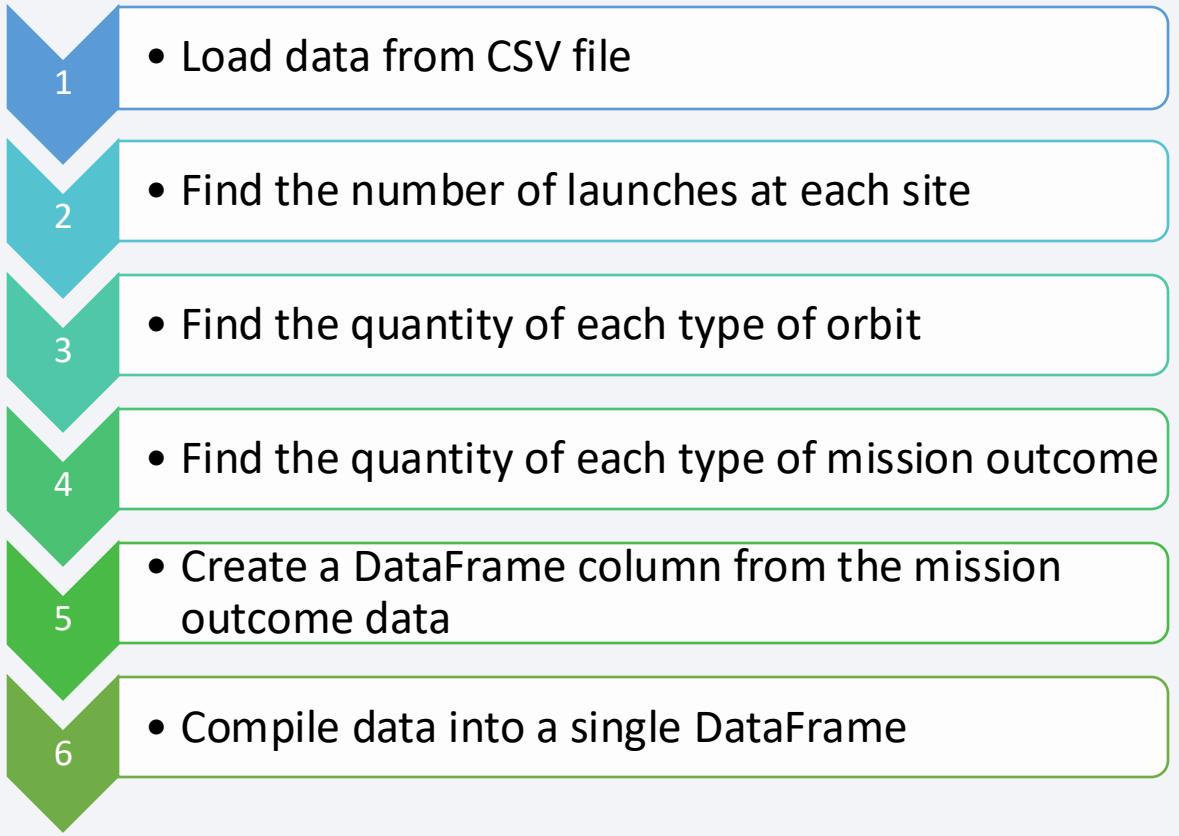
Flowchart of Scraping



Data Wrangling

- The CSV file from the first section contained the data in need of cleaning/wrangling.
- The launch sites, orbit types and mission outcomes were processed and reformatted.
- The mission outcome types were converted to a binary classification (one-hot encoding) where 1 represented the Falcon 9 first stage landing being a success and 0 represented a failure.
- The new mission outcome classification column was added to the DataFrame.
- GitHub Link https://github.com/lehkyi/Applied-Data-

Flowchart Data Wrangling



EDA with Data Visualization

- The following charts were created to look at **Launch Site trends**
  - Scatterplot to see mission outcome relationship split by Launch Site and Flight Number.
  - Scatterplot to see mission outcome relationship split by Launch Site and Payload.
- The following charts were created to look at **Orbit Type trends**
  - Bar chart to see mission outcome relationship with Orbit Type.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Flight Number.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Payload.
- The following chart was created to look at trends based on **Date**
  - Line plot to see mission outcome trend by year.
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/20bd13eb1a1e77aac47797bbc76dfc65aa3afa32/EDA.ipynb

EDA with SQL

- SQL queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/d515d014d4e2f51728cc6f0a12d577a6e21d796b/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Map objects were created and added to the Folium map
  - **Markers** were added for launch sites and for the NASA Johnson Space Center
  - **Circles** were added for the launch sites.
- **Lines** were added to show the distance to the nearby features:
  - Distance to the coastline
  - Distance to the rail line
  - Distance to the perimeter road
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/d515d014d4e2f51728cc6f0a12d577a6e21d796b/folium.ipynb

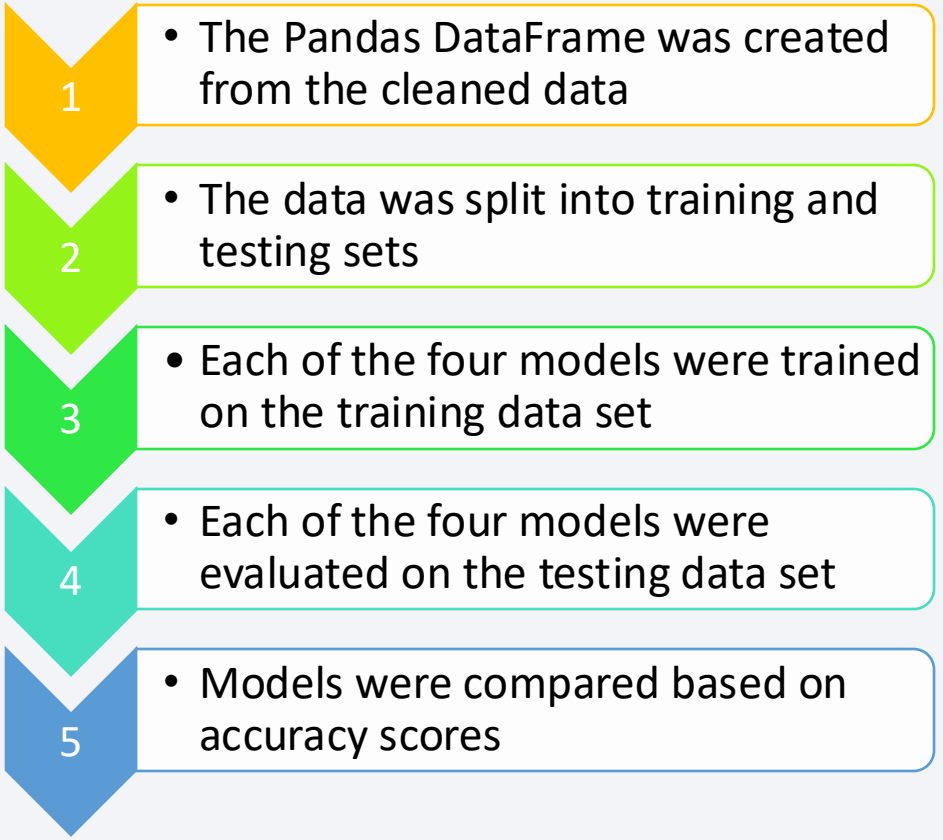
Build a Dashboard with Plotly Dash

- The Plotly Dash dashboard included a dropdown input to select data from ‘one’ or ‘all’ launch sites to display on the pie chart and scatterplot.
- For ‘one’ launch site, the pie chart displayed the distribution of successful and failed Falcon 9 first stage landings for that site.
- For ‘all’ launch sites, the pie chart displayed the distribution of successful Falcon 9 first stage landings between the sites.
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displayed the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/21375b91e0ebddd925a1f921300624791e0fa71e/dash.py

Predictive Analysis (Classification)

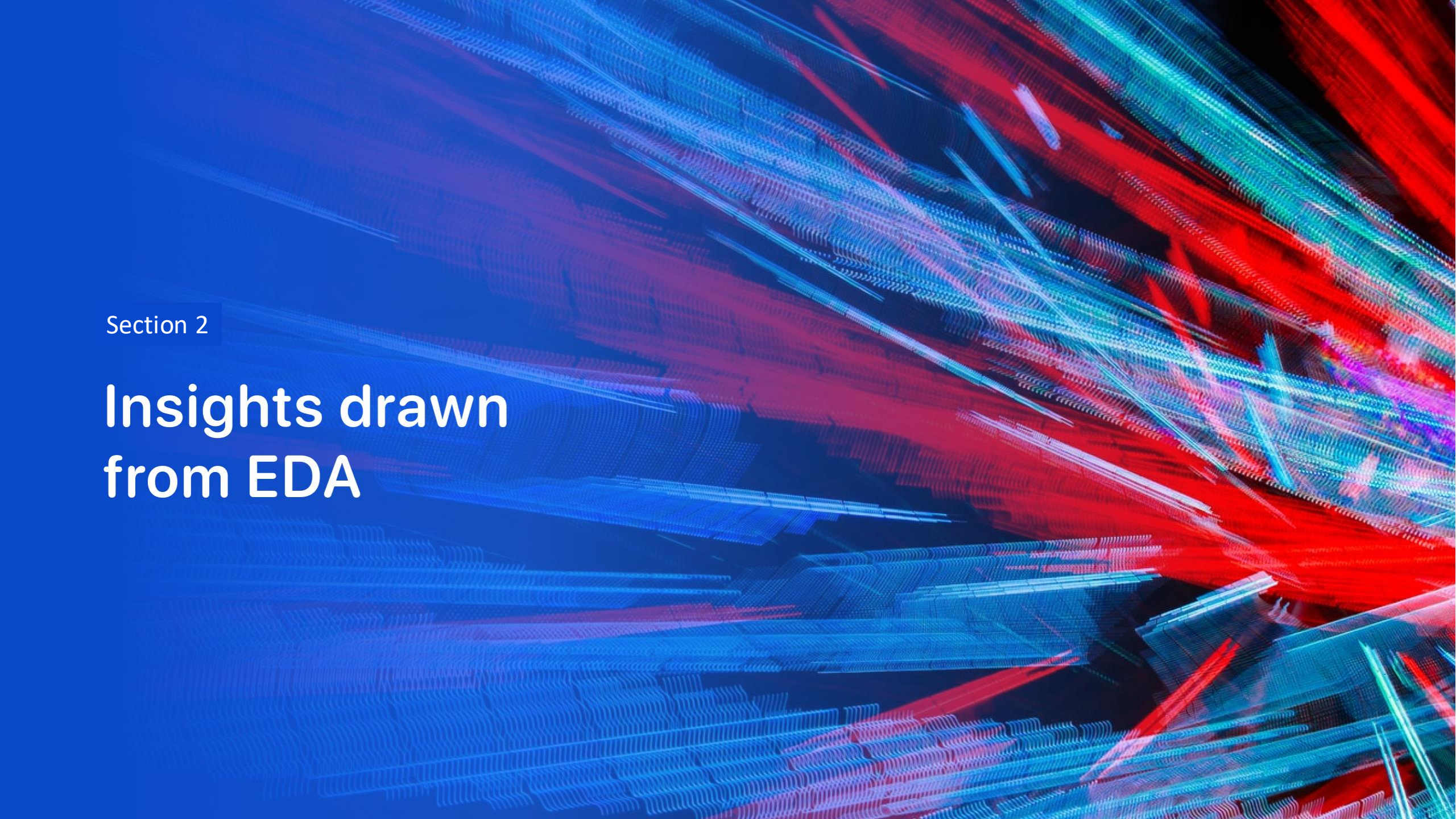
- The dataset was split into training and testing sets.
- The following machine learning models were trained on the training data set:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - KNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using the best_params method.
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.
- GitHub Link https://github.com/lehkyi/Applied-Data-Science-

Flowchart of Predictive Analysis

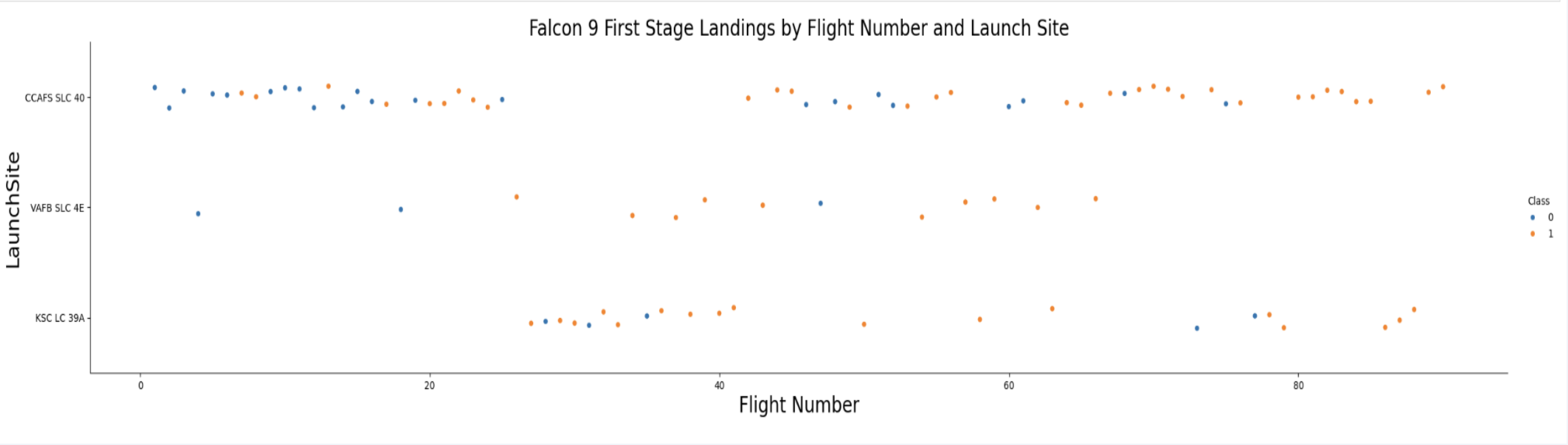


Results

- Insights Drawn from EDA (Exploratory Data Analysis)
  - Exploratory Data Analysis – Data Visualizations
  - Exploratory Data Analysis – SQL Queries
- Launch Sites Proximities Analysis
  - Interactive Folium Maps (Screenshots)
- Build a Dashboard with Plotly Dash
  - Interactive Plotly Dash Dashboard (Screenshots)
- Predictive Analysis (Classification)
  - Predictive Analysis (Classification) – Machine Learning

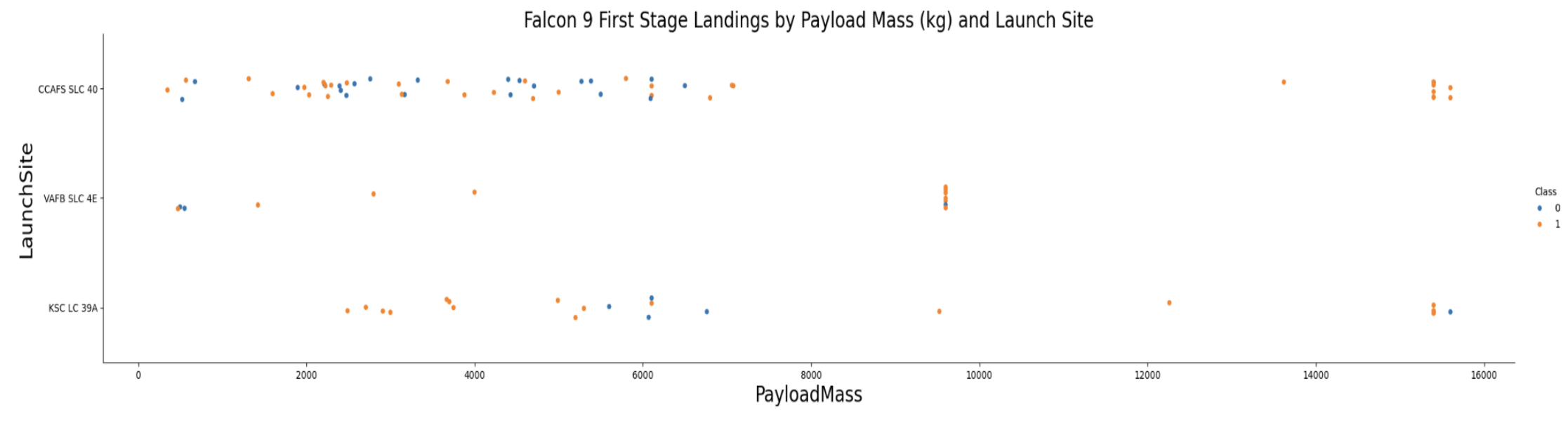

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



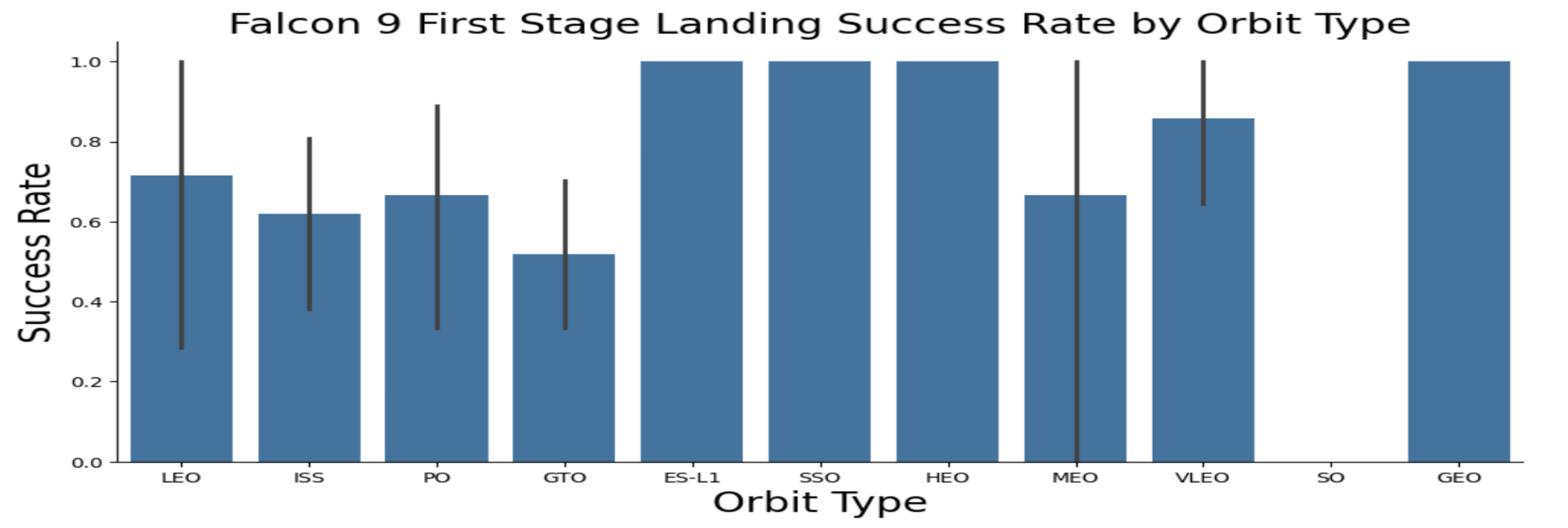
- *Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).*
- *Success rate varied noticeably with launch site.*
- *Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.*

Payload vs. Launch Site



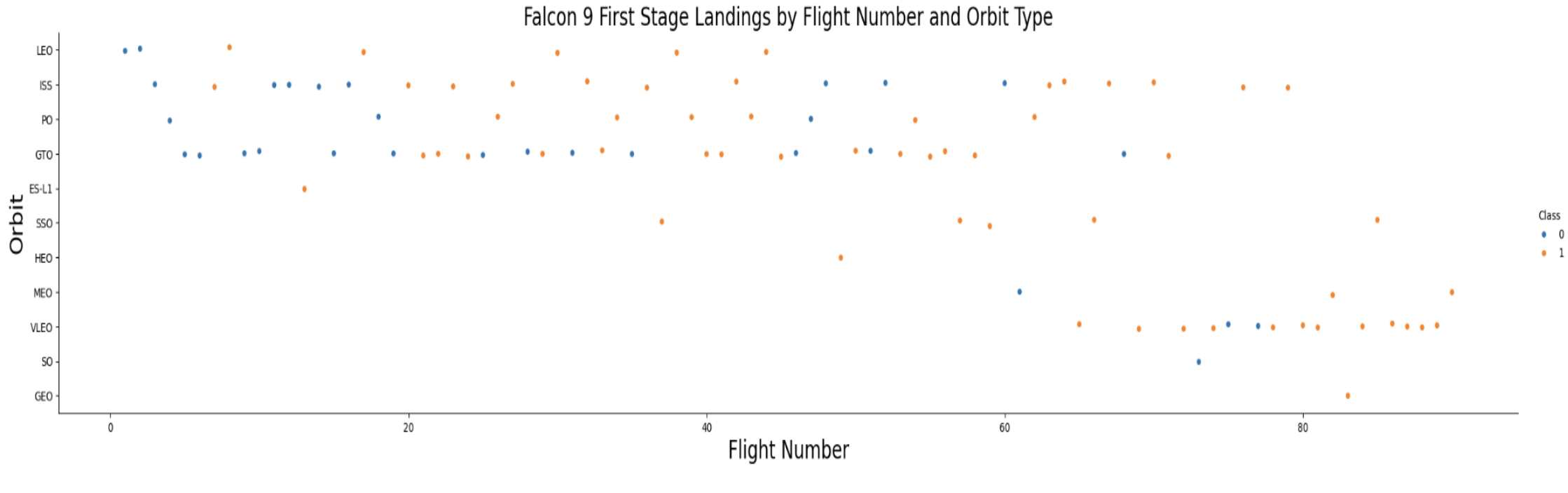
- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.
- The failed landings at the KSC LC 39A launch site are mostly grouped around a narrow band of payload masses.

Success Rate vs. Orbit Type



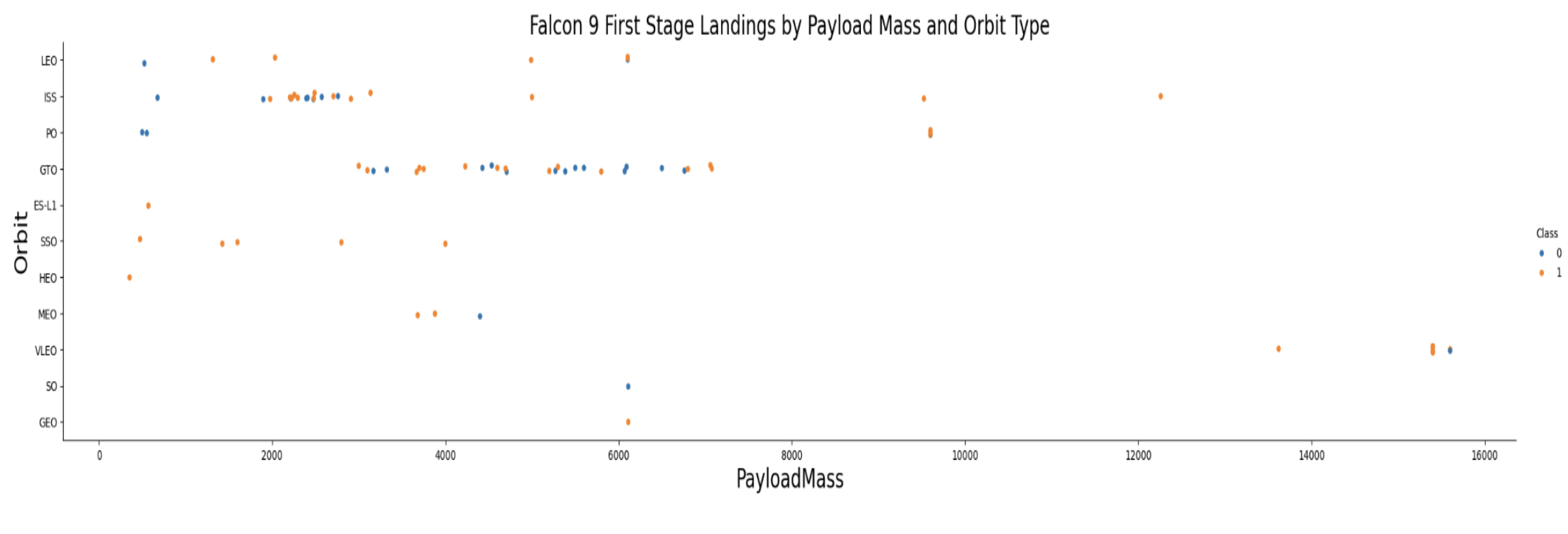
- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.

Flight Number vs. Orbit Type



- There is a positive correlation between flight number and success rate. Larger flight numbers were associated with higher success rates.

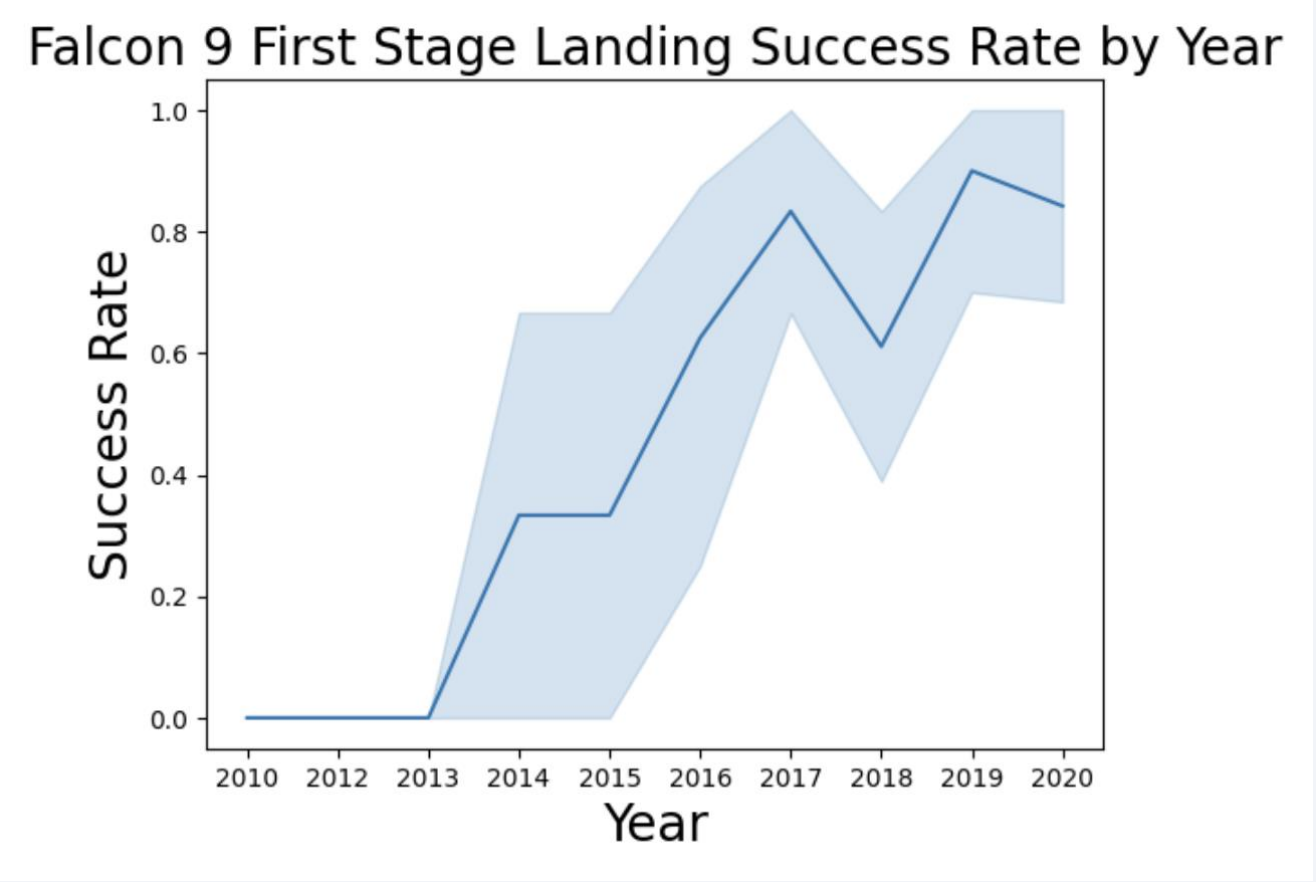
Payload vs. Orbit Type



- Some orbit types showed higher success rates than others.
- Success rate appeared to have no obvious correlation with payload mass.

Launch Success Yearly Trend

- The success rate of the Falcon 9 first stage landings has increased significantly over the selected interval of years.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

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

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

- There are 4 unique launch sites.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are 5 records where launch sites begin with `CCA`

Total Payload Mass

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL PAYLOAD NASA" FROM SPACEXTBL WHERE "Customer" = "NASA (CRS)"
```

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

TOTAL PAYLOAD NASA
45596

- The total payload carried by boosters from NASA is 45, 596 kg

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1"
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2, 928 kg

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

MIN("Date")
2015-12-22

- The first successful landing outcome on ground pad is 22 Dec 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)"
    AND "PAYLOAD_MASS__KG_">4000 AND PAYLOAD_MASS__KG_<6000
```

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

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

- There are 4 names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT 'Success' AS Outcome, COUNT(*) AS Total FROM SPACEXTBL WHERE Mission_Outcome LIKE '%Success%'
    UNION
   SELECT 'Failure' AS Outcome, COUNT(*) AS Total FROM SPACEXTBL WHERE Mission_Outcome NOT LIKE '%Success%';
```

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

Outcome	Total
Failure	1
Success	100

- There are 100 successful and 1 failed mission outcomes

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

```
 * 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

- There are the list of 12 boosters that have carried the max payload mass.

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date, 6,2) as "Month","Booster_Version", "Launch_Site" FROM SPACEXTBL
WHERE "Landing_Outcome" = "Failure (drone ship)" AND substr(Date,0,5)="2015"
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- There were two failed landing outcomes with a drone ship in 2015. Both launched from CCAFS LC-40. One occurred in January and the other in April.

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.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS "Outcome_Count" FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY  "Landing_Outcome" ORDER BY  "Outcome_Count" DESC
```

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

Landing_Outcome	Outcome_Count
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

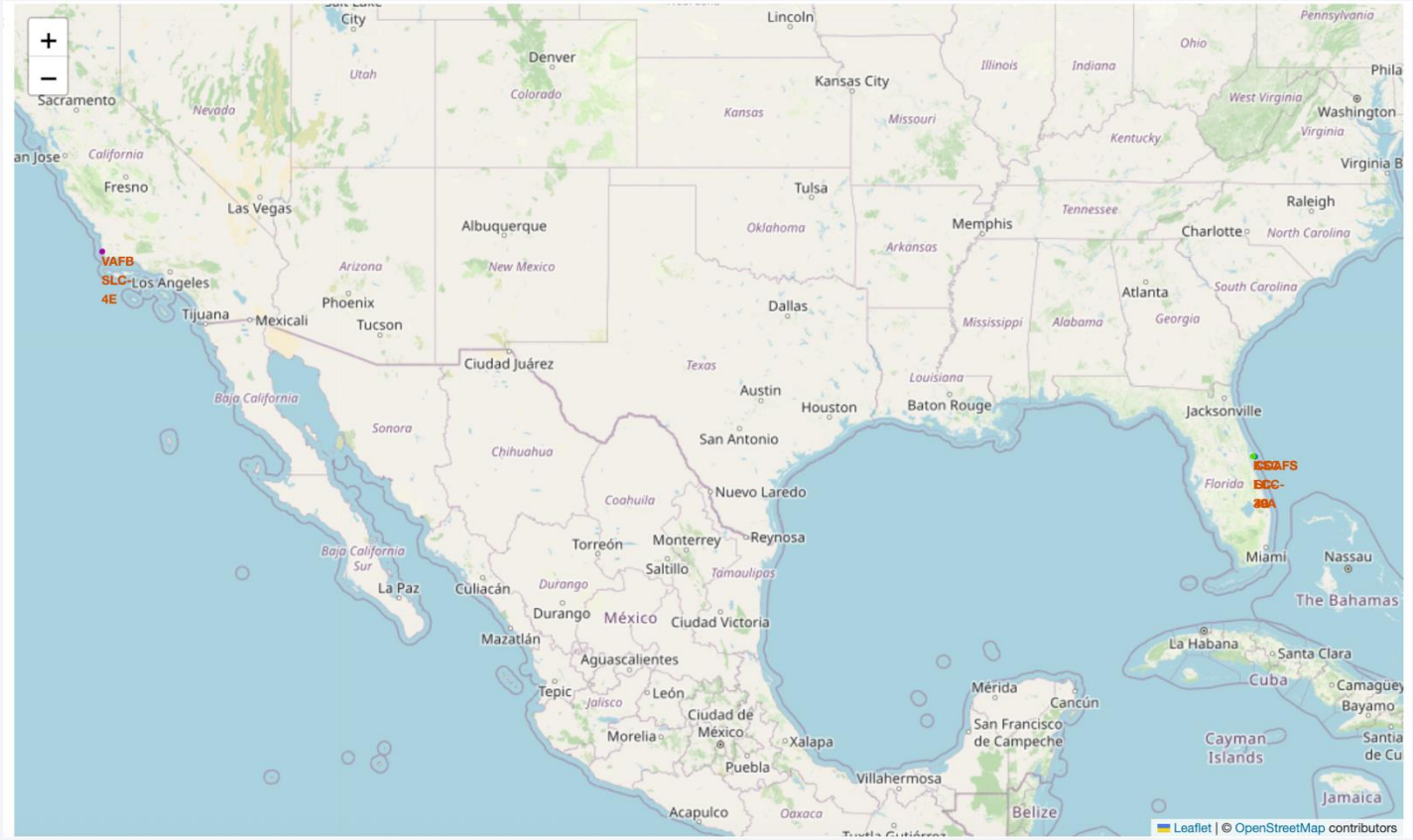
- The most common landing outcome was ‘No attempt’.

Section 3

Launch Sites Proximities Analysis

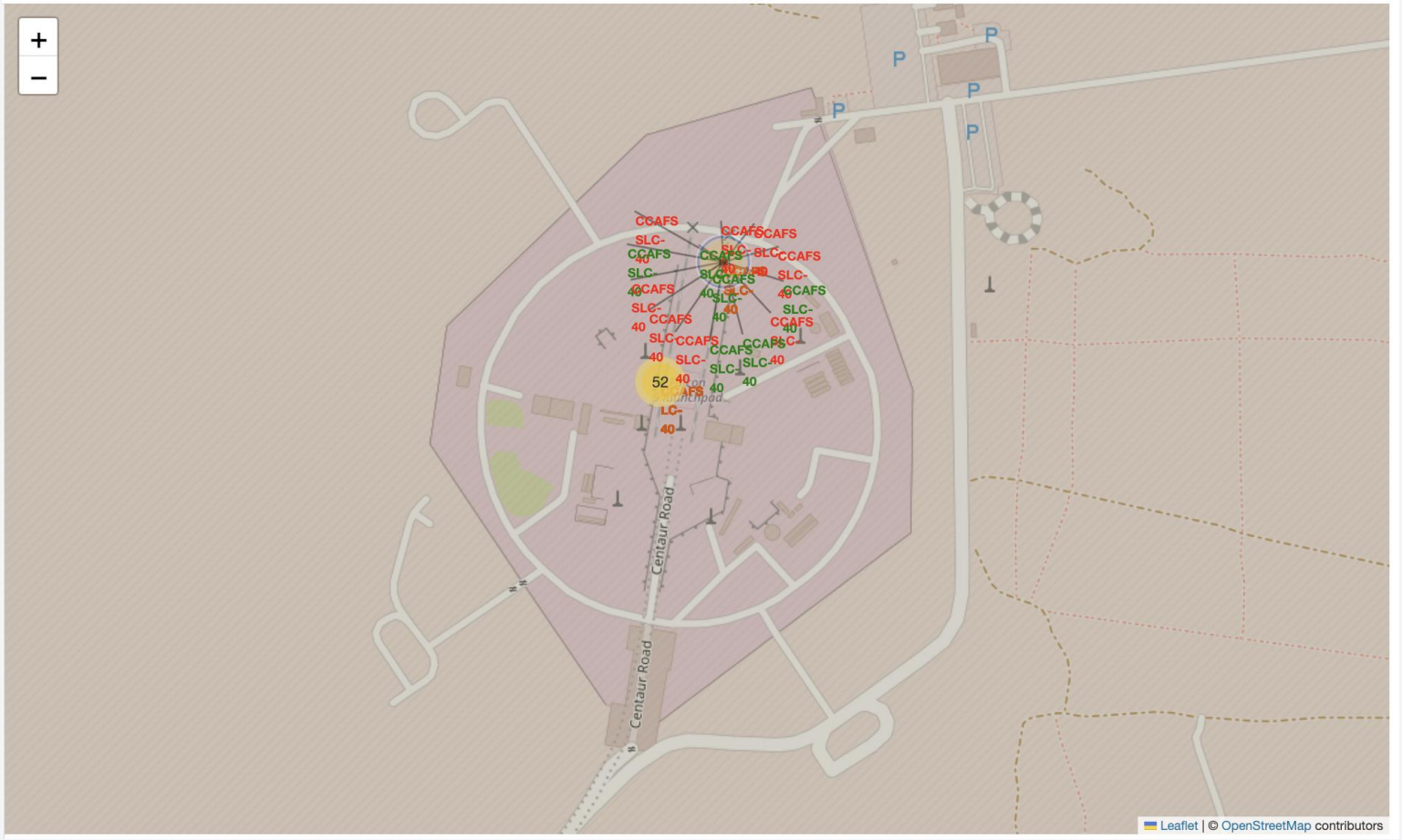
All Launch Sites

- All launch sites in proximity to the Equator line
- All launch sites in very close proximity to the coast



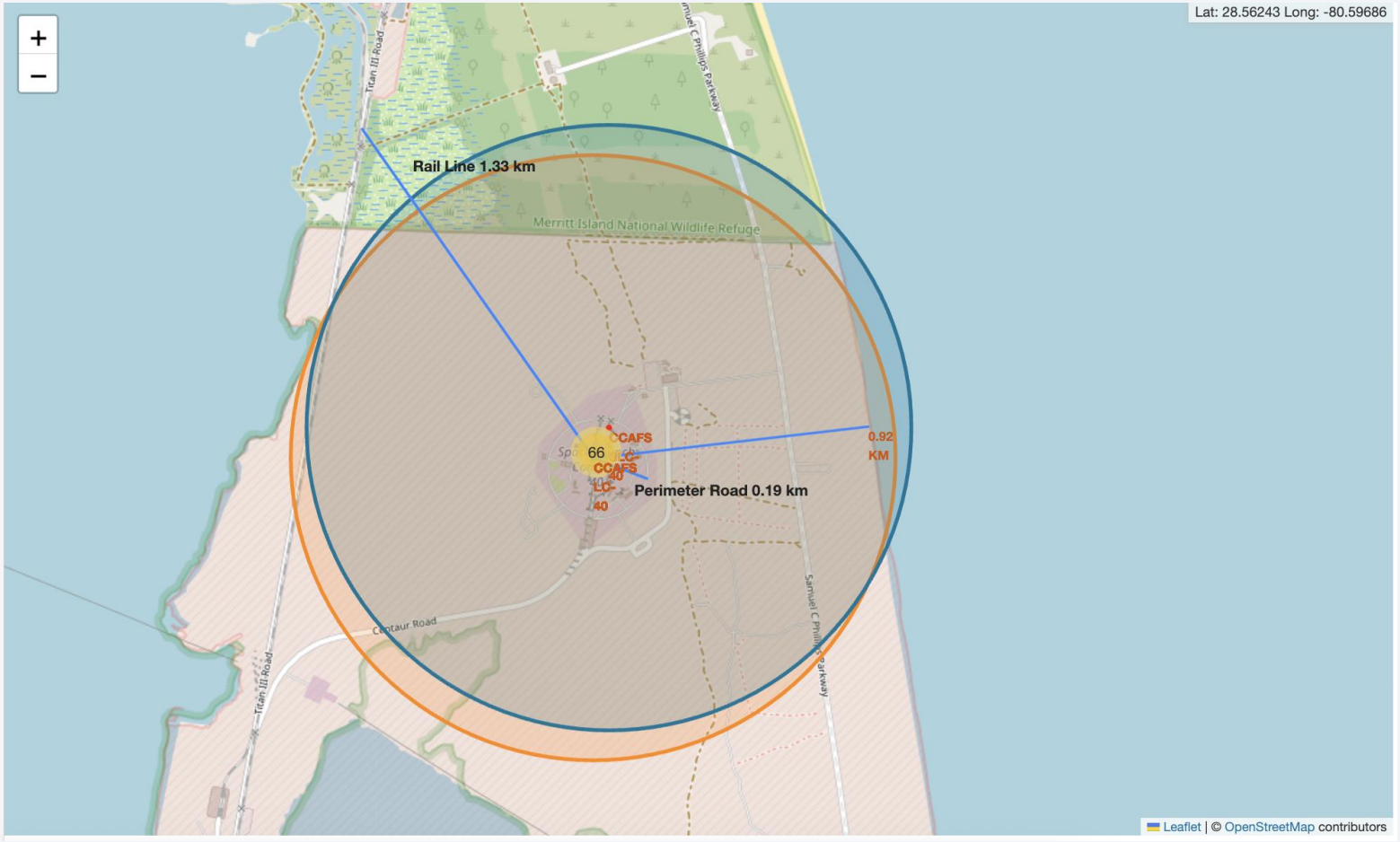
Map Markers of Success/Failed Landings

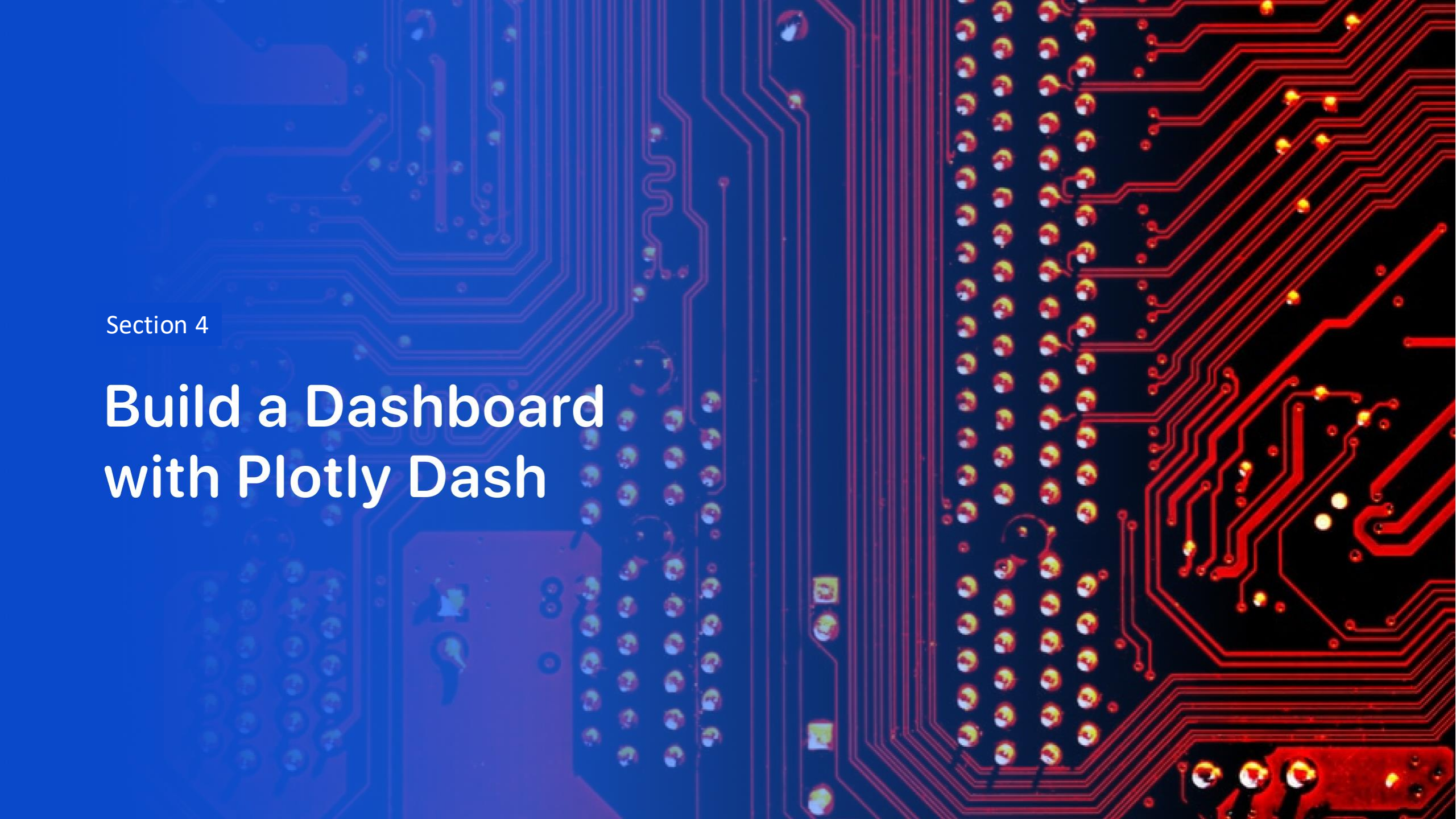
- The markers display the mission outcomes (Success - green/Failure - red) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.



Distance from Launch Site to Proximities

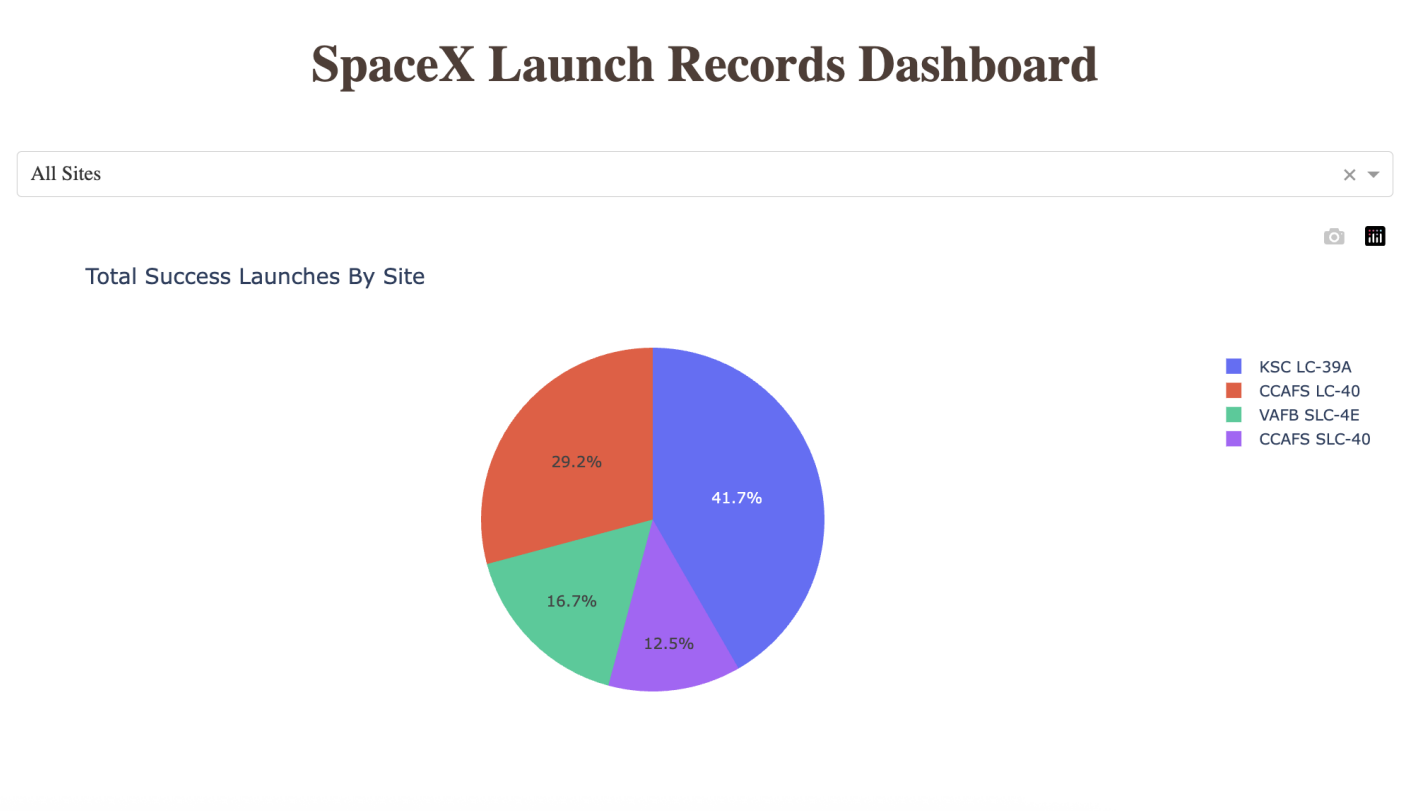
- The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- The coastline is 0.92 km away from CCAFS LC-40.
- The rail line is 1.33 km away from CCAFS LC-40.




Section 4
Build a Dashboard with Plotly Dash

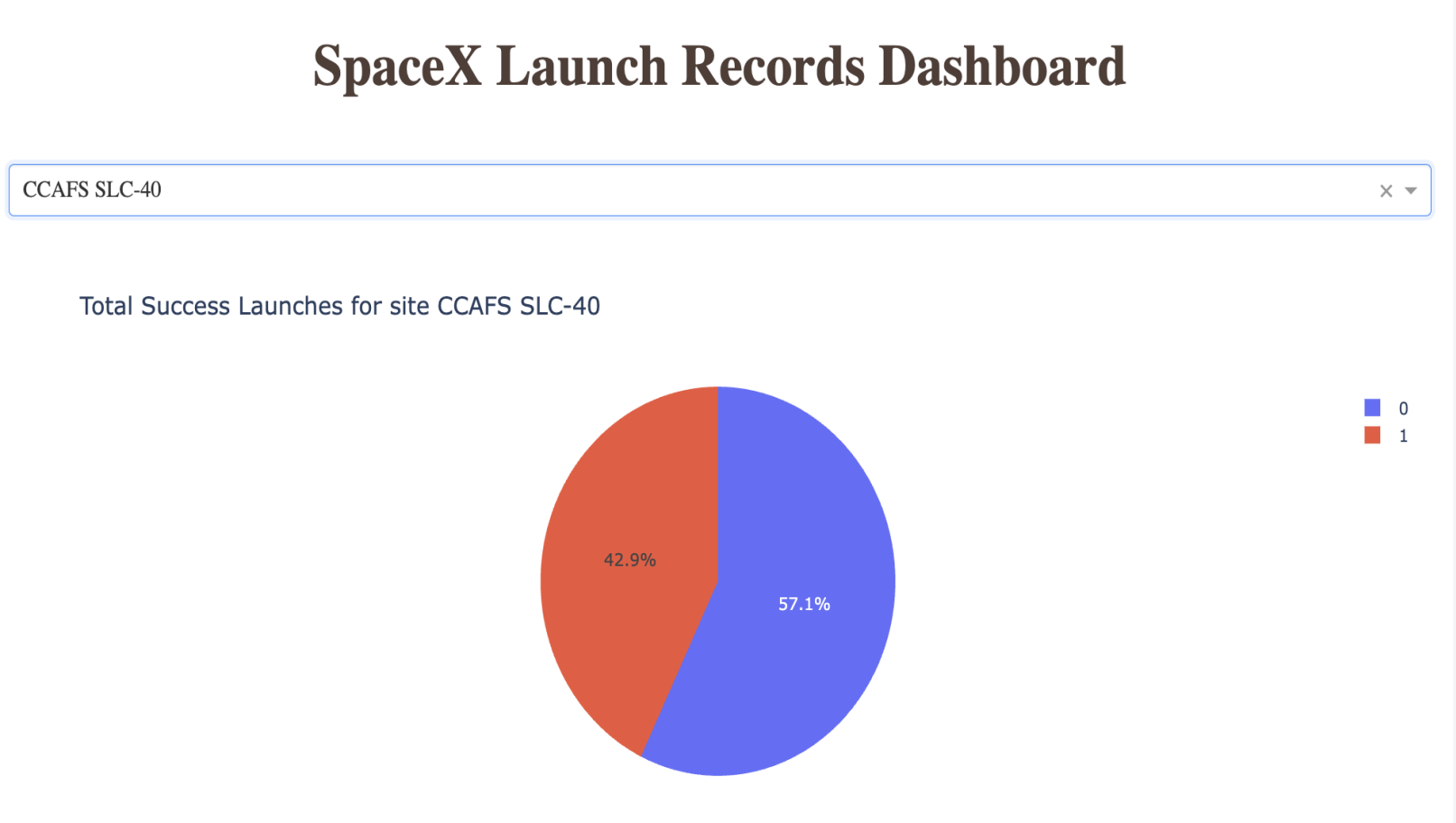
SpaceX Launch Records Dashboard

- The dropdown menu allowed the selection of one or all launch sites.
- With all launch sites selected, the pie chart displayed the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.



Launch Site with Highest Launch Success Ratio

- Falcon 9 first stage failed landings are indicated by the ‘0’ Class (■ blue wedge in the pie chart) and successful landings by the ‘1’ Class (■ red wedge in the pie chart).
- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).



Payload vs. Launch Outcome

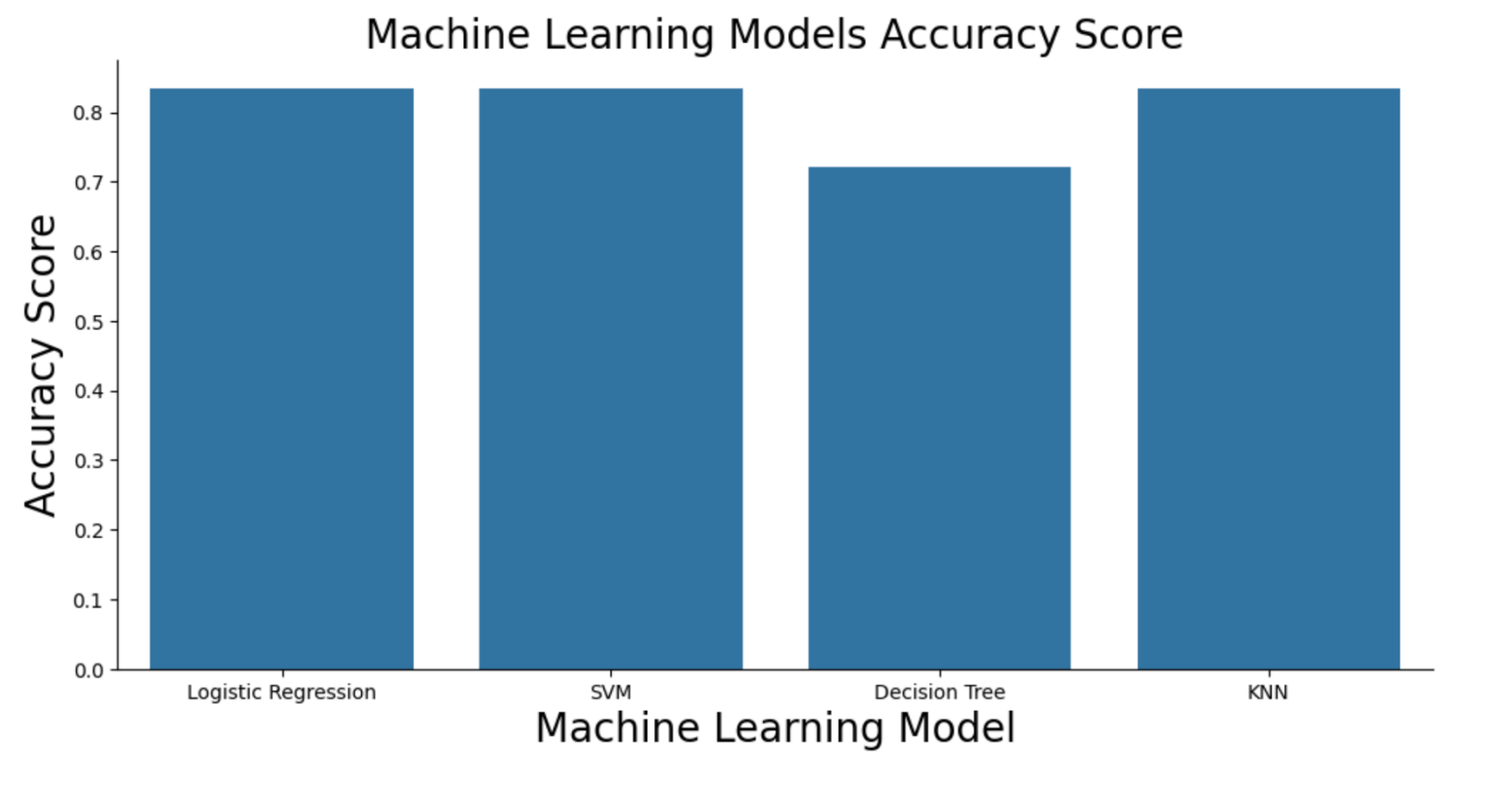
- The ‘FT’ booster version category has the largest success rate

Section 5

Predictive Analysis (Classification)

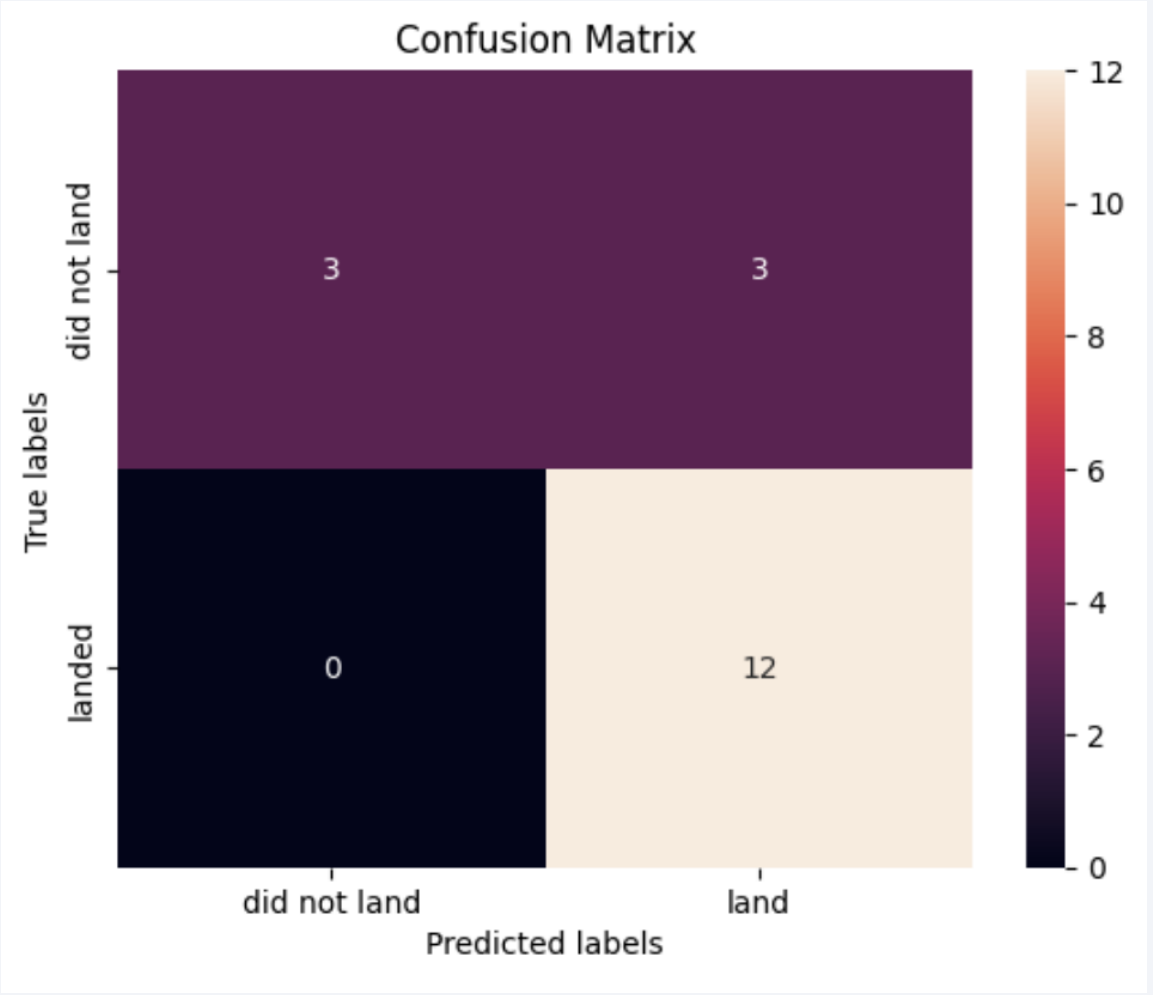
Classification Accuracy

- All models except Decision Tree have hight accuracy



Confusion Matrix

- Shown the confusion matrix of the KNN
- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives



Conclusions

- SpaceX's record for Falcon 9 first stage landing outcomes has improved.
- The trend is toward better performance and greater success as more launches are made.
- The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

Appendix

CSV files	Path
dataset_part_1.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/21375b91e0ebddd925a1f921300624791e0fa71e/dataset_part_1.csv
dataset_part_2.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/21375b91e0ebddd925a1f921300624791e0fa71e/dataset_part_2.csv
dataset_part_3.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/39b8cc566c3a945c6a4fb7002bbc8ec26ba7a19e/dataset_part_3.csv
spacex_launch_dash.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/39b8cc566c3a945c6a4fb7002bbc8ec26ba7a19e/spacex_launch_dash.csv
spacex_launch_geo.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/39b8cc566c3a945c6a4fb7002bbc8ec26ba7a19e/spacex_launch_geo.csv
spacex_web_scraped.csv	https://github.com/lehkyi/Applied-Data-Science-Capstone/blob/39b8cc566c3a945c6a4fb7002bbc8ec26ba7a19e/spacex_web_scraped.csv

Thank you!