Winning the Space Race with Data Science

Jonathan M Clark

2024-11-30 (Updated version)

OUTLINE

Outline

SUMMARY

Summary

- Methodology
  - Data was collected about the Falcon 9 first stage landings from 2010 to 2020 from a public API (https://api.spacexdata.com/) unaffiliated with SpaceX and from the publically available data on Wikipedia (https://en.wikipedia.org/wiki/SpaceX). Additional data sets were provided with the course.
  - Data cleaning / wrangling included extracting landing outcome data to serve as the dependent variable for the machine learning models.
  - SQL queries and data visualizations, including static plots, interactive maps, and an interactive dashboard, were used to discover insights about the data set and to answer various questions.
  - Predictive analysis was performed using the following machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN)
- Results
  - The data of the SpaceX Falcon 9 first stage landings include the flight number, date of launch, payload mass, orbit type, launch site, and mission outcome.
  - Logistic Regression, SVM, and KNN all performed equally well on this dataset for predictive purposes.

INTRODUCTION

Introduction

- In competition with SpaceX, a rival rocket launch company wants to make predictions about the success of SpaceX Falcon 9 rocket first stage landings.
- Questions to explore:
  - What is the nature and extent of the available data about SpaceX Falcon 9 first stage landings?
  - Which machine learning model would work best (have the highest accuracy) to predict the outcome of a Falcon 9 first stage landing from a future launch?
  - Will a future Falcon 9 first stage landing be successful?

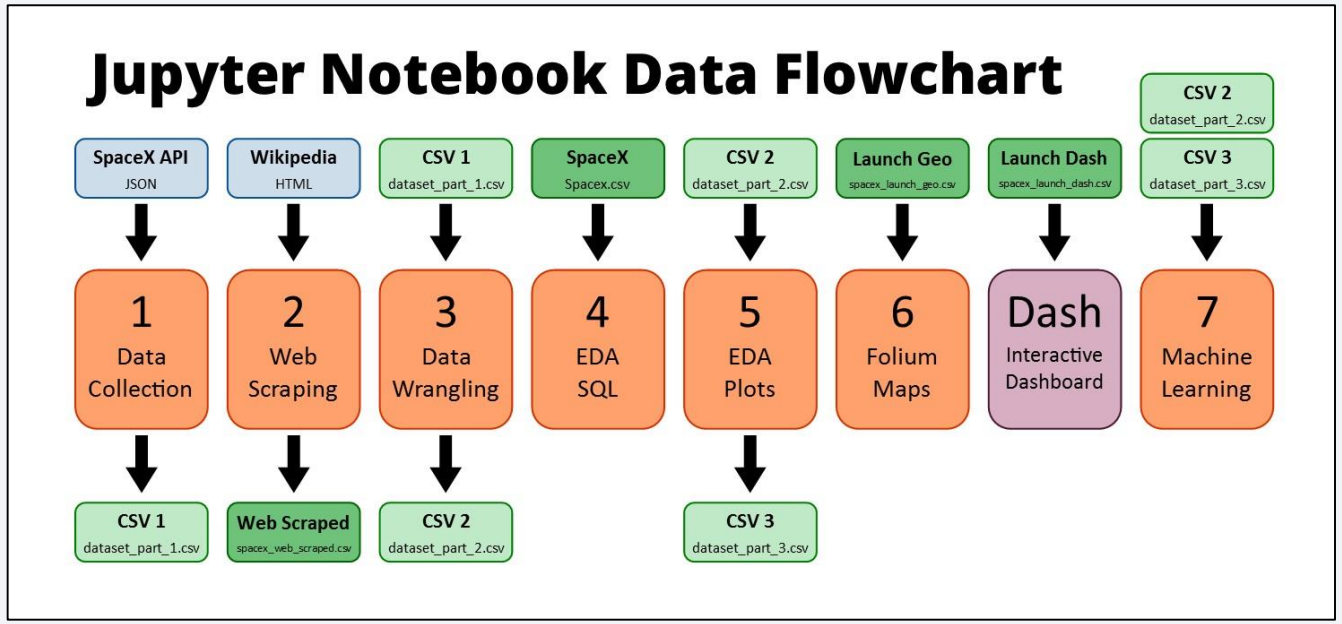
METHODOLOGY

Methodology

- Data on the SpaceX Falcon 9 first stage landings was collected from a public API, unaffiliated with SpaceX, and from a Wikipedia article. Additional data sets were provided with the course in CSV file format.
- Data was wrangled/cleaned in preparation for visualizations, queries, and machine learning model training.
- Exploratory Data Analysis (EDA) was performed using data visualizations and SQL.
- Interactive data visualizations were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done using machine learning models.

Data Collection

- The data sets were collected from:
  - An IBM copy of a response from a publically accessible API with launch data in JSON format.
  - A permanently-linked Wikipedia page with launch data in HTML tables (9 June 2021 revision).
  - Additional data sets were provided with the course in CSV file format (Highlighted as the darker green CSV files in the top row of the diagram below). See appendix for links.

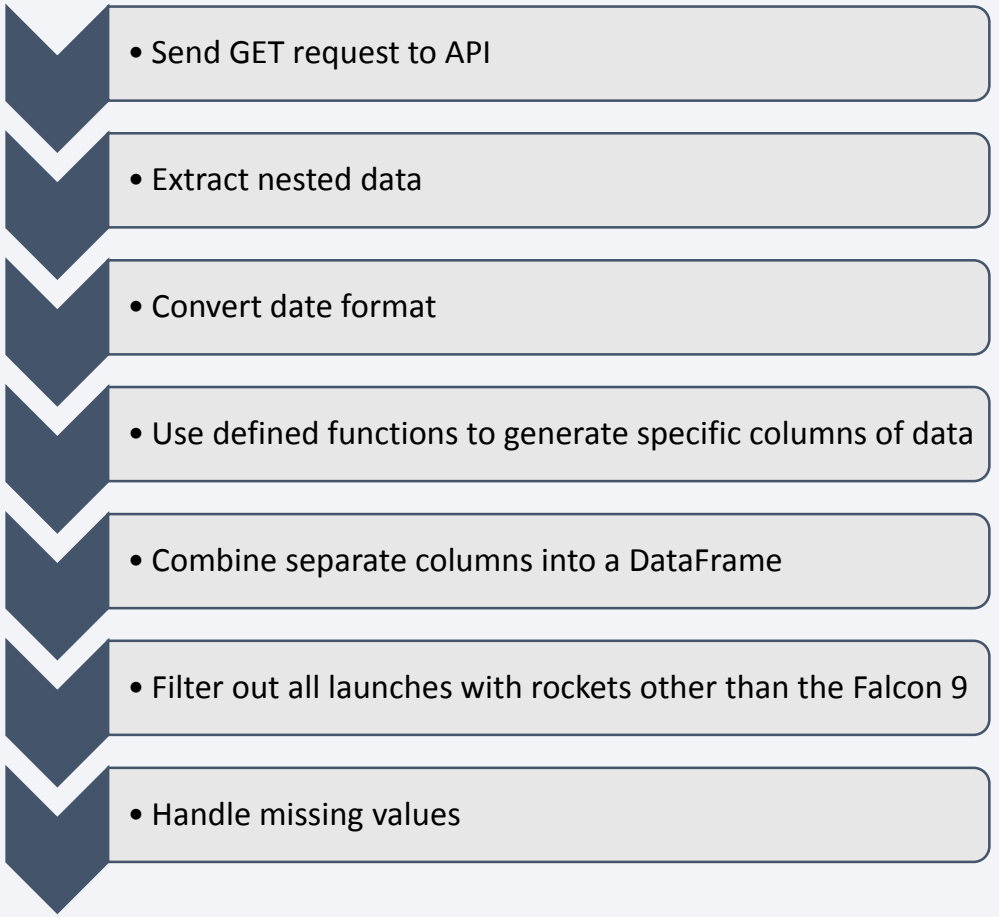


Data Collection – API

- SpaceX data was available publically at the API endpoint: https://api.spacexdata.com/
- Note: This API is not affiliated with SpaceX.
- A copy of the response from this API was made available for the purposes of this project. See appendix for link.
- Data was extracted from the response from the API and loaded into a Pandas DataFrame for further analysis.
- GitHub URL (Data Collection):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/1_Capstone_JonathanMClark_Data_Collection.ipynb

Flowchart of API Data Processing



Data Collection – Wikipedia Web 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).
- Launch data was extracted from these tables and loaded into a Pandas DataFrame for further analysis.
- GitHub URL (Web Scraping):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/2_Capstone_JonathanMClark_Webscraping.ipynb

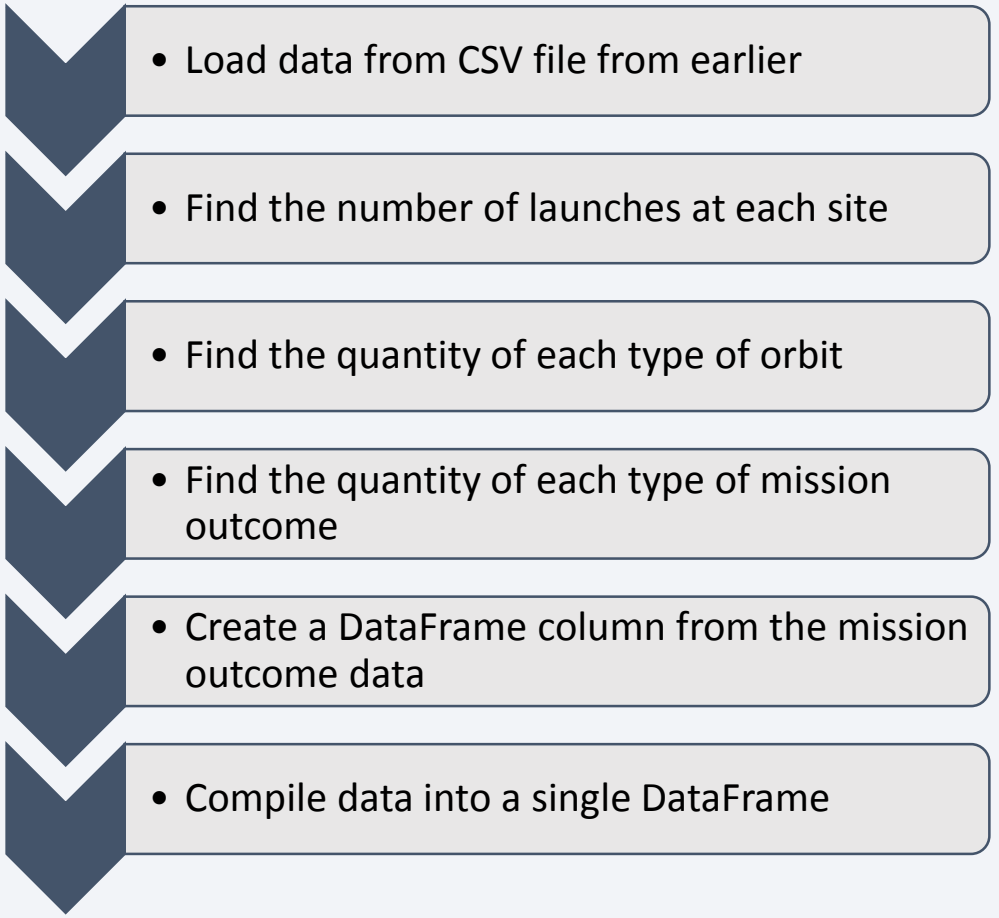
Flowchart of Wikipedia Web Scraping

Data Cleaning / 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 URL (Data Wrangling):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/3_Capstone_JonathanMClark_Data_Wrangling.ipynb

Flowchart of Data Cleaning / 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 time
  - Line plot to see **mission outcome** trend by **year**.
- GitHub URL (EDA with Data Visualization):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/5_Capstone_JonathanMClark_EDA_Data_Visualization.ipynb

EDA with SQL

- SQL queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes

- GitHub URL (EDA with SQL):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/4_Capstone_JonathanMClark_EDA_SQL.ipynb

Interactive Folium Map

- 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 from CCAFS LC-40 to the coastline
    - Distance from CCAFS LC-40 to the rail line
    - Distance from CCAFS LC-40 to the perimeter road
- GitHub URL (Folium Maps):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/6_Capstone_JonathanMClark_Launch_Site_Location.ipynb

Interactive Plotly Dash Dashboard

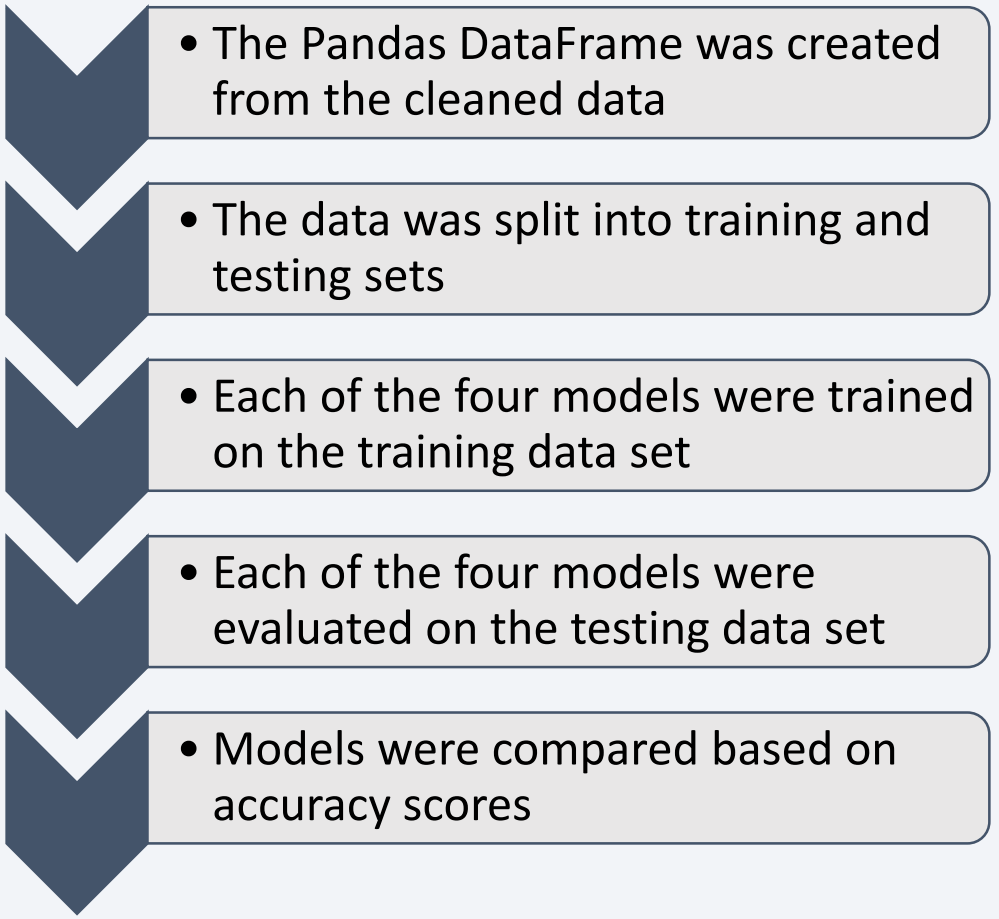
- The Plotly Dash dashboard included a dropdown input to select data from ‘one’ or ‘all’ launch sites to display on the pie chart and scatterplot.
- For ‘one’ launch site, the pie chart displayed the distribution of successful and failed Falcon 9 first stage landings for that site.
- For ‘all’ launch 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 URL (Dashboard File):

https://github.com/JonathanMClark/DataScienceCapstone/blob/main/JonathanMClark_spacex_dash_app.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 URL (Machine Learning): https://github.com/JonathanMClark/DataScienceCapstone/blob/main/7_Capstone_JonathanMClark_Machine_Learning_Prediction.ipynb

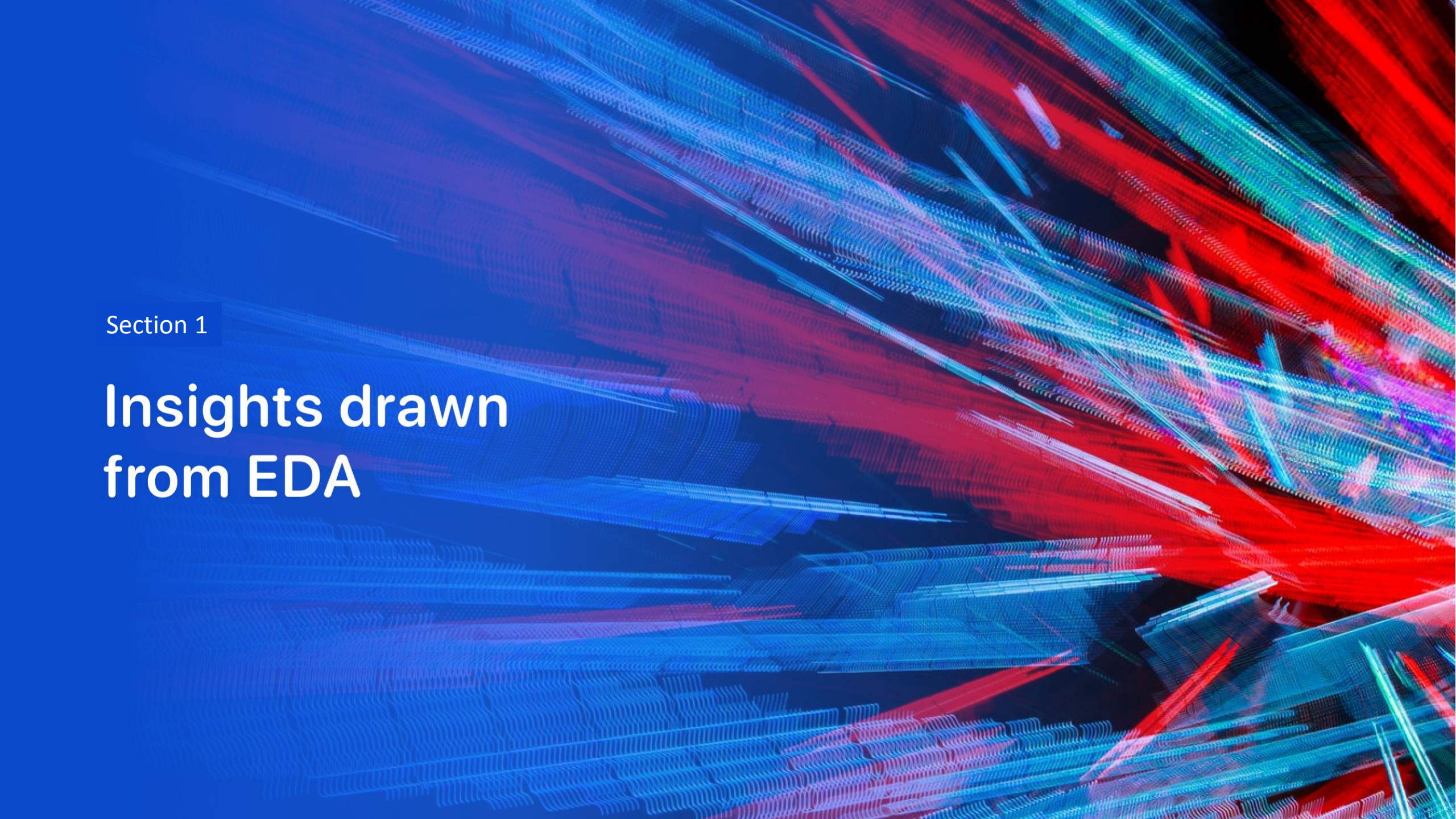
Flowchart of Machine Learning



RESULTS

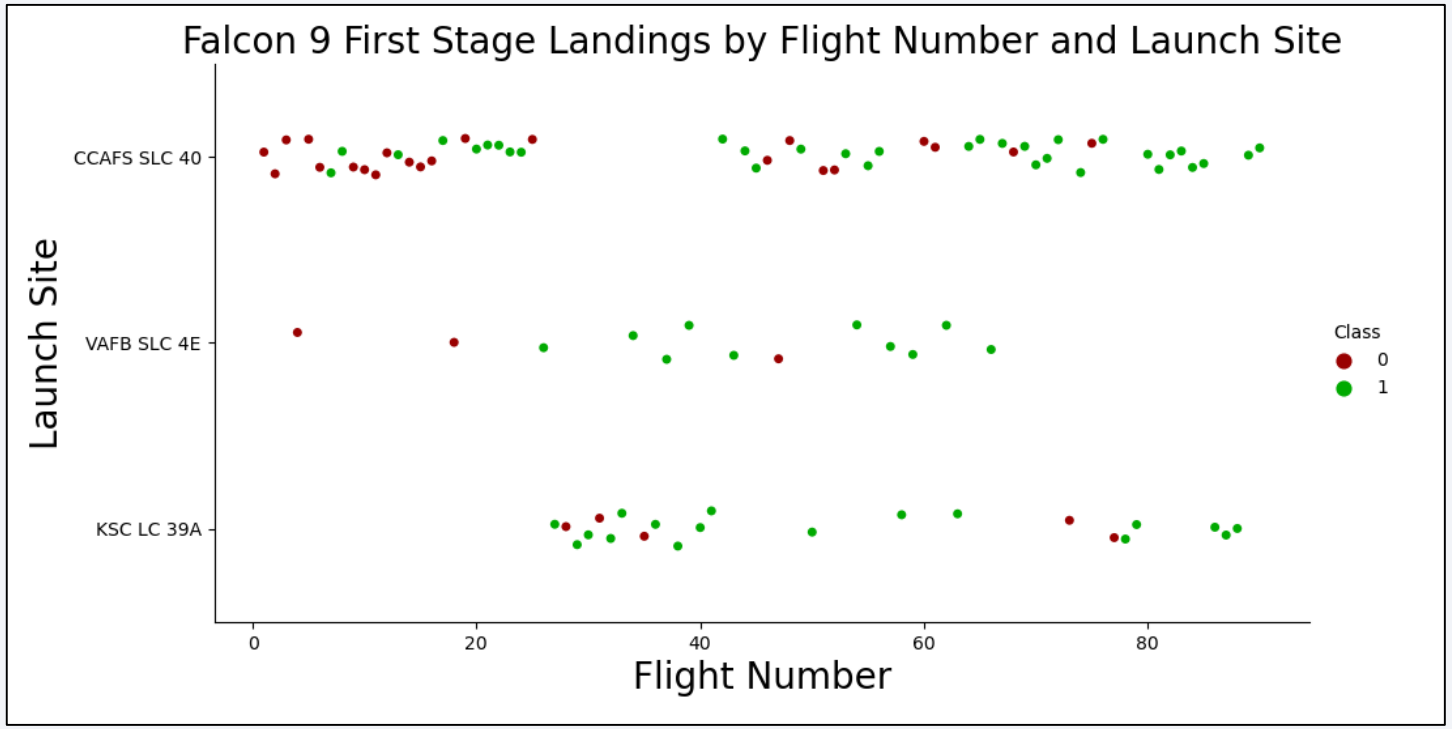
Results

- 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 1
Insights drawn from EDA

Flight Number vs. Launch Site

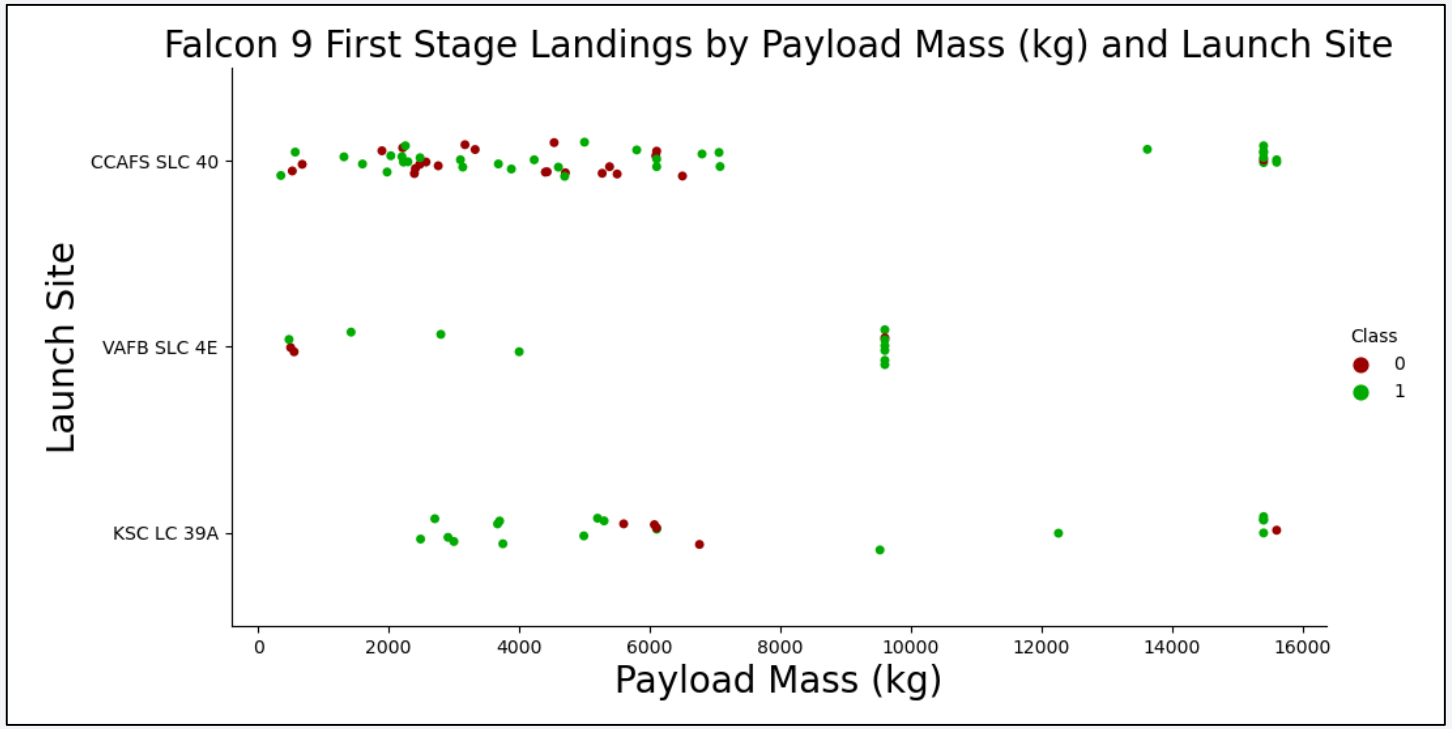
- Success rate varied noticeably with launch site.
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.



- Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).

Payload vs. Launch Site

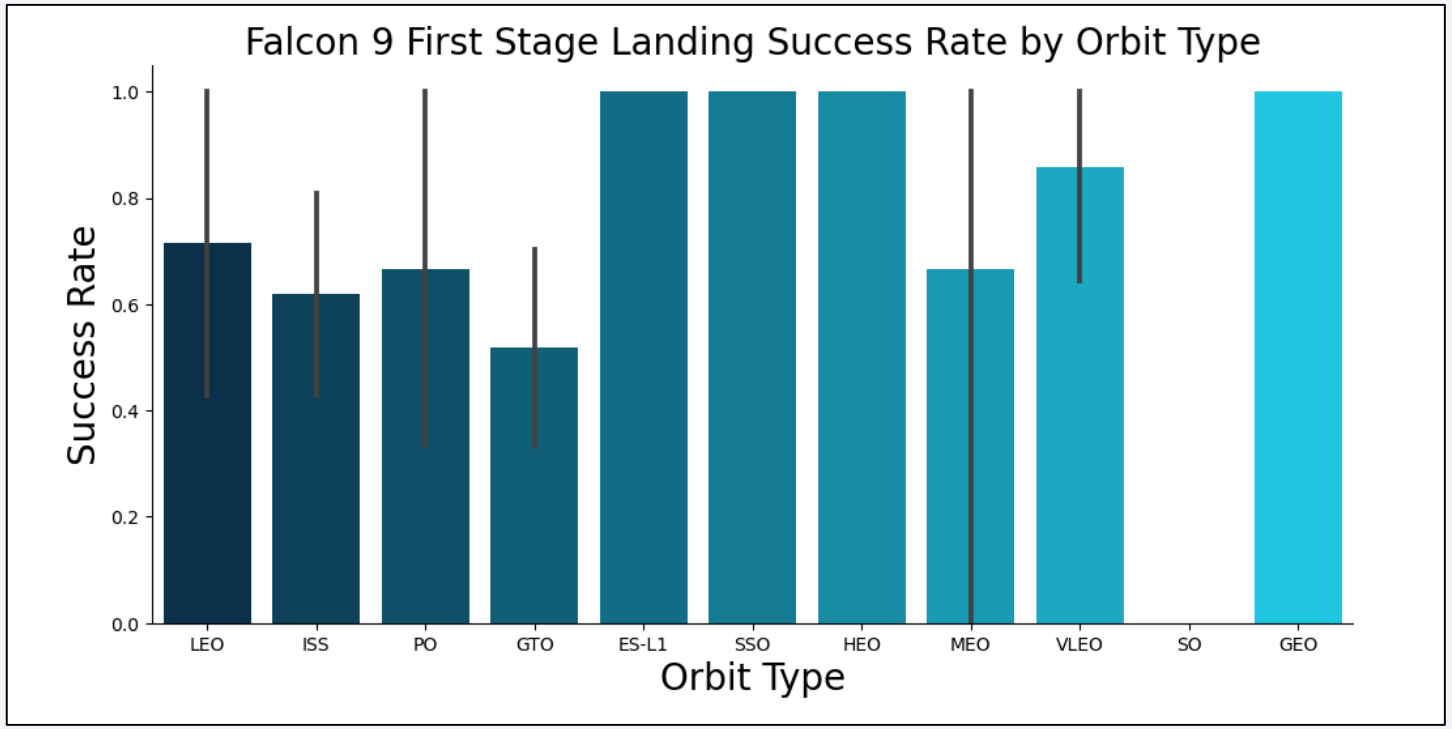
- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.
- The failed landings at the KSC LC 39A launch site are mostly grouped around a narrow band of payload masses.



- Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).

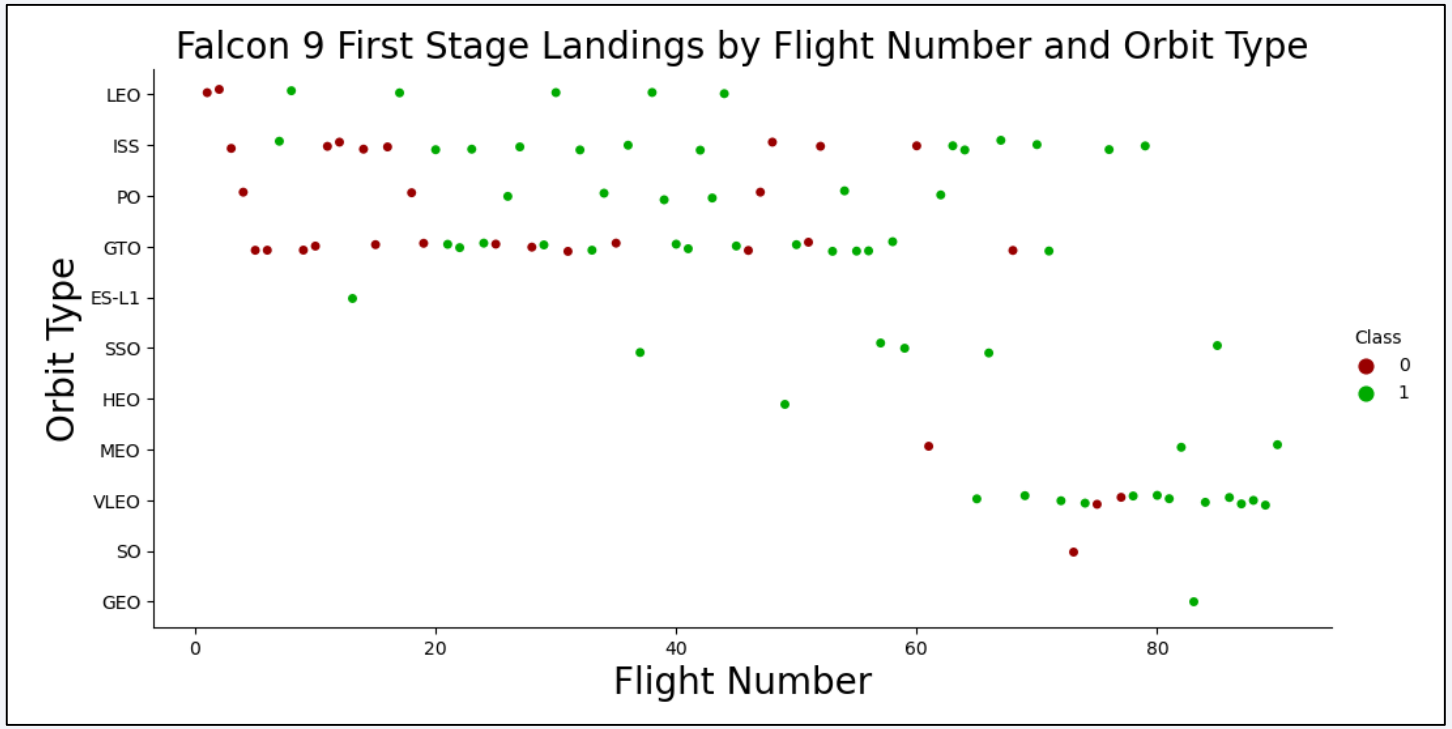
Success Rate vs. Orbit Type

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



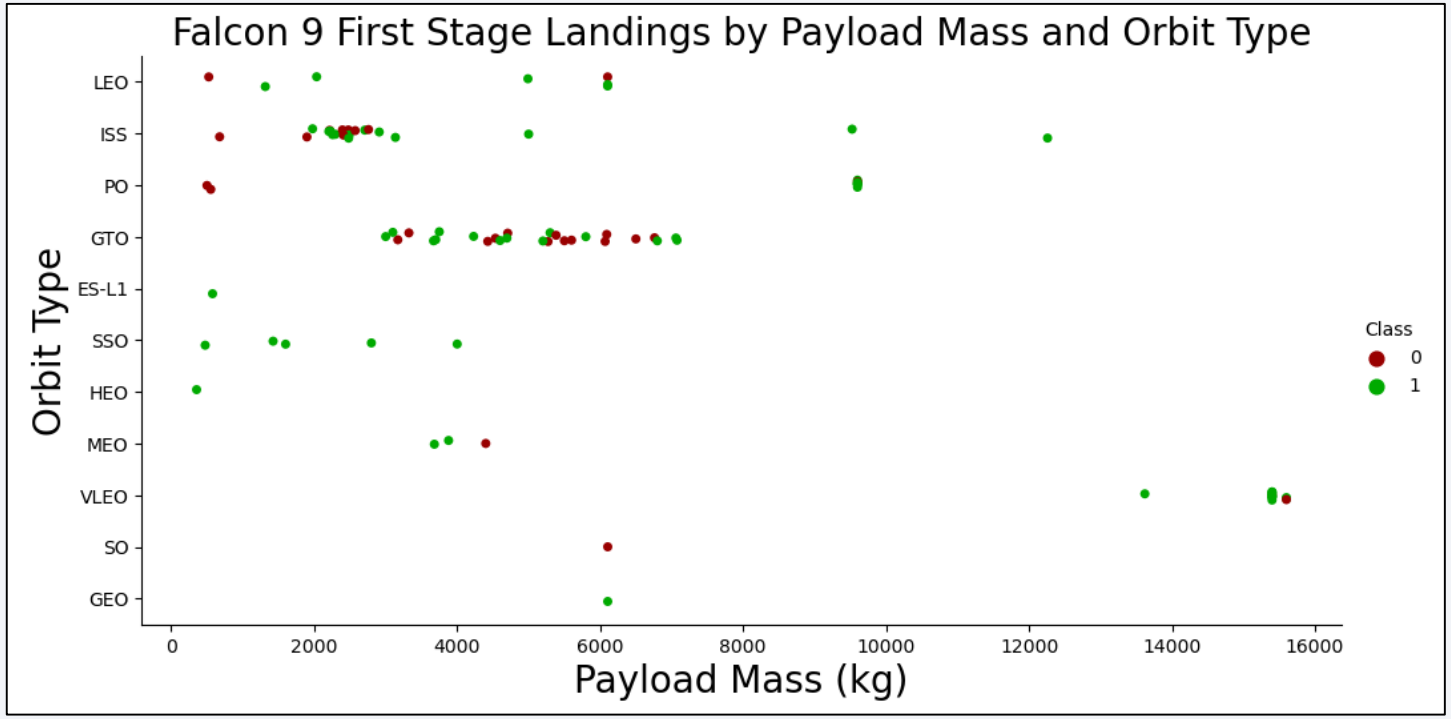
Flight Number vs. Orbit Type

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



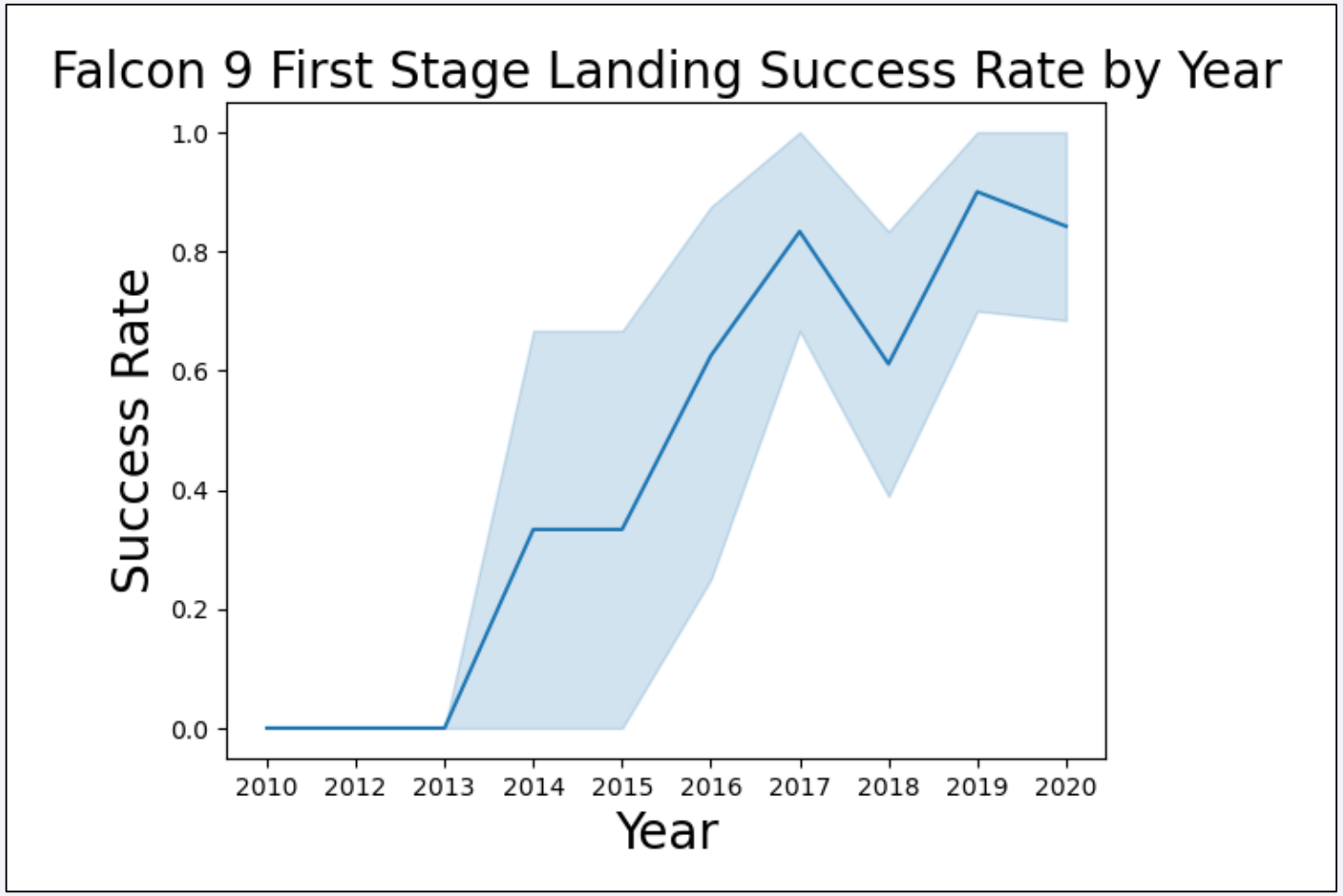
Payload vs. Orbit Type

- 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

- **Question:** What are the names of the unique launch sites?
- **Query:**
```
SELECT DISTINCT `LAUNCH_SITE` FROM `SPACEXDATASET`;
```
- **Result:**

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

- **Explanation:** There are four unique launch sites.

Launch Site Names That Begin with 'CCA'

- **Task:** Find 5 records with launch sites that begin with `CCA`.
- **Query:** SELECT * FROM `SPACEXDATASET` WHERE `launch_site` LIKE 'CCA%' LIMIT 5;
- **Result:**

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- **Explanation:** This is a fairly straightforward sampling mechanism used to gain a sense of the data contained in the database table.

Total Payload Mass

- **Question:** What is the total payload carried by boosters from NASA?
- **Query:**
```
SELECT sum(`payload_mass__kg_`) AS "Total Payload Mass (kg)" FROM `SPACEXDATASET` WHERE `customer` LIKE '%NASA (CRS)%';
```
- **Result:**

Total Payload Mass (kg)
48213

- **Explanation:** The total payload carried by boosters from NASA is 48,213 kg.

Average Payload Mass by F9 v1.1

- **Question:** What is the average payload mass carried by booster version F9 v1.1?
- **Query:**
```
SELECT sum(`payload_mass__kg_`) / count(`payload_mass__kg_`) AS “Average Payload Mass (kg)" FROM `SPACEXDATASET` WHERE `booster_version` LIKE 'F9 v1.1';
```
- **Result:**

Average Payload Mass (kg)
2928

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

First Successful Ground Landing Date

- **Question:** On which date did the first successful landing outcome on ground pad occur?
- **Query:**
```
SELECT min(`DATE`) AS "First Successful Landing Outcome Date" FROM `SPACEXDATASET` WHERE `landing__outcome` LIKE 'Success (ground pad)';
```
- **Result:**

First Successful Landing Outcome Date
2015-12-22

- **Explanation:** The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Question:** What are the names of the boosters which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000?
- **Query:**

```
SELECT DISTINCT `booster_version` FROM `SPACEXDATASET` WHERE `landing__outcome` = 'Success (drone ship)'
AND `payload_mass__kg_` BETWEEN 4000 AND 6000;
```

- **Result:**

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

- **Explanation:** The four booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg are listed above.

Total Number of Successful and Failure Mission Outcomes

- **Question:** What was the total number of successful and failed mission outcomes?
- **Query:**

```
SELECT (SELECT count(*) FROM `SPACEXDATASET` WHERE lcase(`landing__outcome`) LIKE '%success%') AS "Success“, count(*) AS "Failure" FROM `SPACEXDATASET` WHERE lcase(`landing__outcome`) NOT LIKE '%success%';
```

- **Result:**

Success	Failure
61	40

- **Explanation:** There were 61 successful and 40 failed mission outcomes.

Boosters Carried Maximum Payload

- **Question:** What were the names of the boosters which have carried the maximum payload mass?
- **Query:**

```
SELECT `booster_version`, `payload_mass__kg_` FROM `SPACEXDATASET` WHERE `payload_mass__kg_` =
(SELECT max(`payload_mass__kg_`) FROM `SPACEXDATASET`);
```

- **Result:**

booster_version	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

- **Explanation:** The maximum payload mass carried in this dataset is 15,600 kg. Twelve (12) separate Falcon 9 boosters carried this amount of payload mass.

2015 Launch Records

- **Task:** List the failed landing_outcomes in drone ship, their booster versions, and launch site names for records in year 2015.
- **Query:**

```
SELECT MONTHNAME(`DATE`) AS 'Month', `landing__outcome`, `booster_version`, `launch_site` FROM `SPACEXDATASET` WHERE `landing__outcome` = 'Failure (drone ship)' AND YEAR(`DATE`) = 2015;
```

- **Result:**

Month	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- **Explanation:** 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

- **Task:** 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.
- **Query:**

```
SELECT `landing__outcome`, count(`landing__outcome`) AS 'Count' FROM `SPACEXDATASET` WHERE `DATE`
BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY `landing__outcome` ORDER BY count(`landing__outcome`) DESC;
```

- **Result:**

landing__outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- **Explanation:** The most common landing outcome was 'No attempt'.


Section 2
Launch Sites
Proximities Analysis

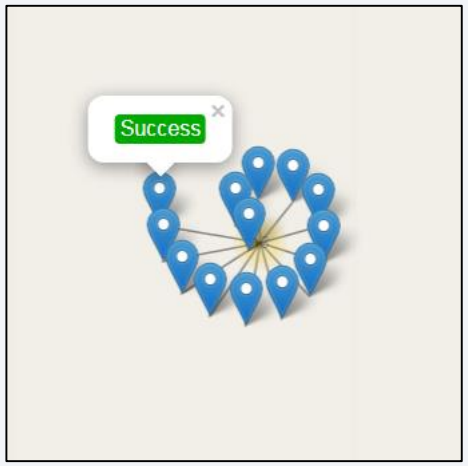
Falcon 9 Launch Site Locations

- California, USA
  - VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E
- Florida, USA
  - KSC LC-39A | Kennedy Space Center Launch Complex 39A
  - CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40
  - CCAFS SLC-40 | Cape Canaveral Air Force Station Space Launch Complex 40
  - *Note: CCAFS LC-40 and CCAFS SLC-40 in the data refer to the same launch site

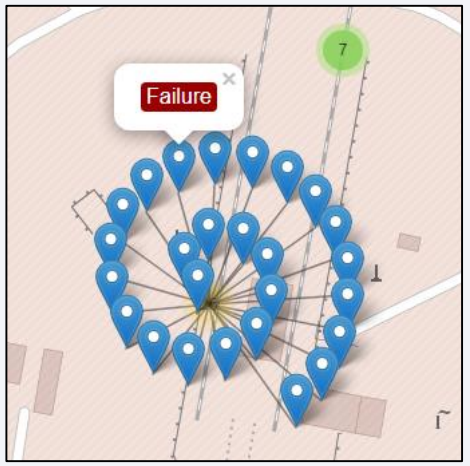
Map Markers of Success/Failed Landings

- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.

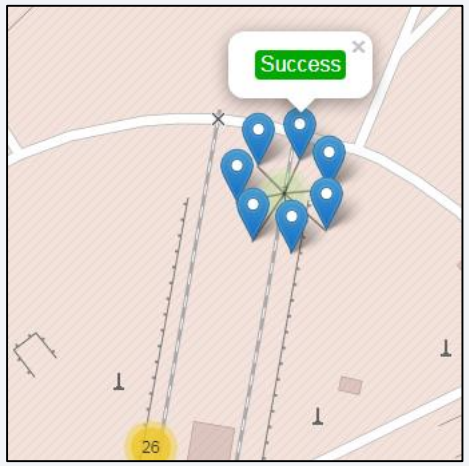
VAFB SLC-4E



KSC LC-39A



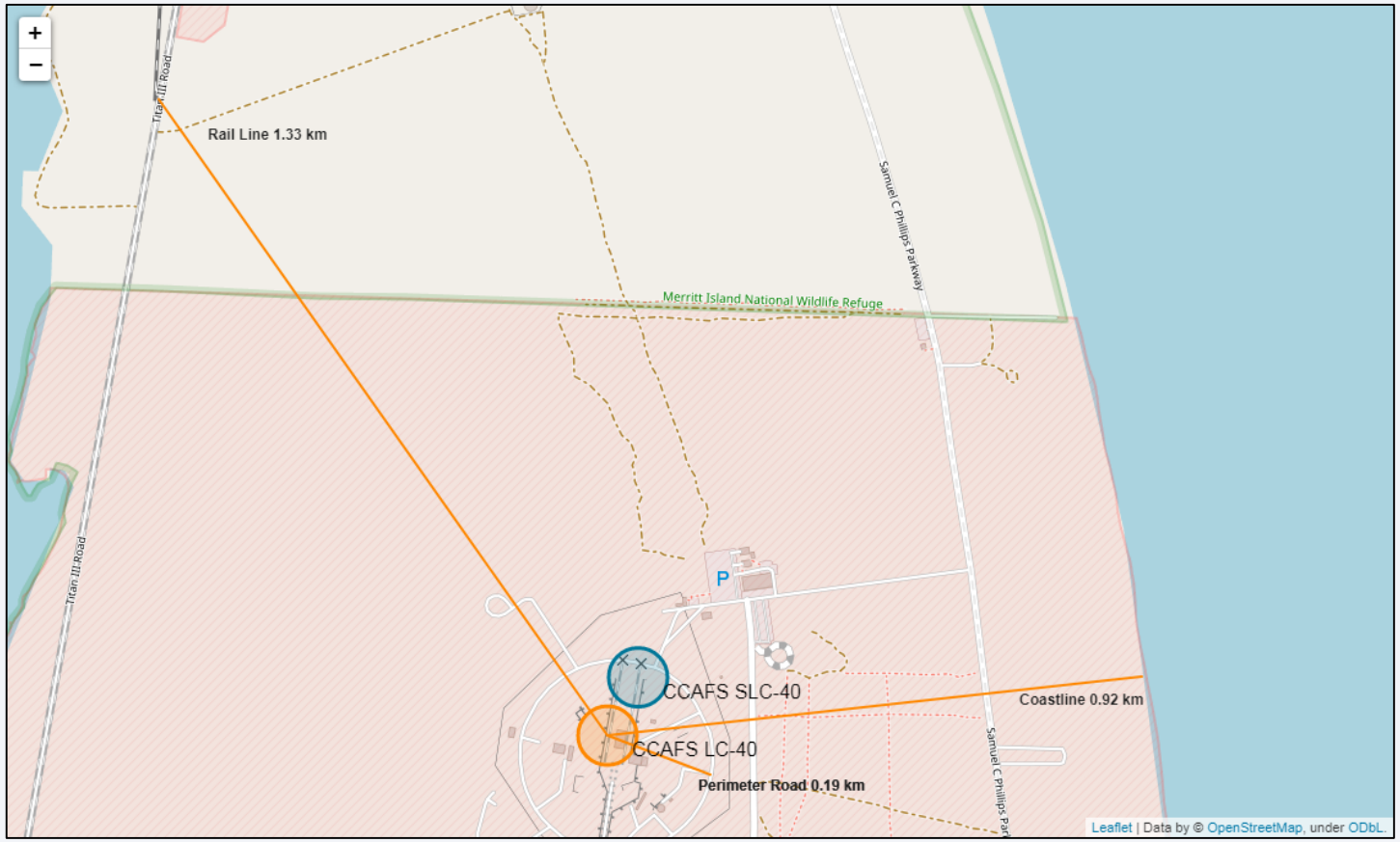
CCAFS LC-40

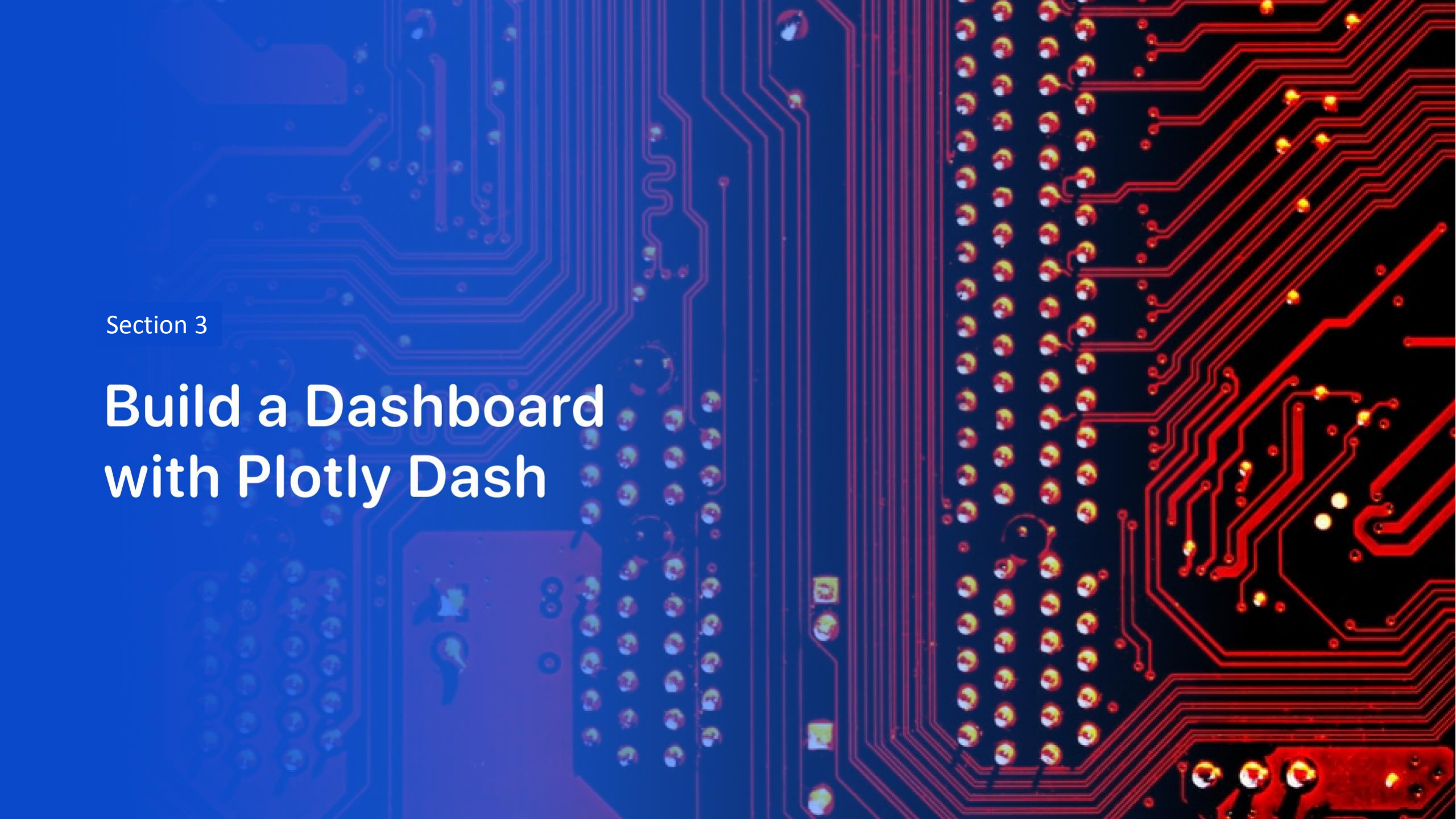


CCAFS SLC-40

Distance from Launch Site to Proximities

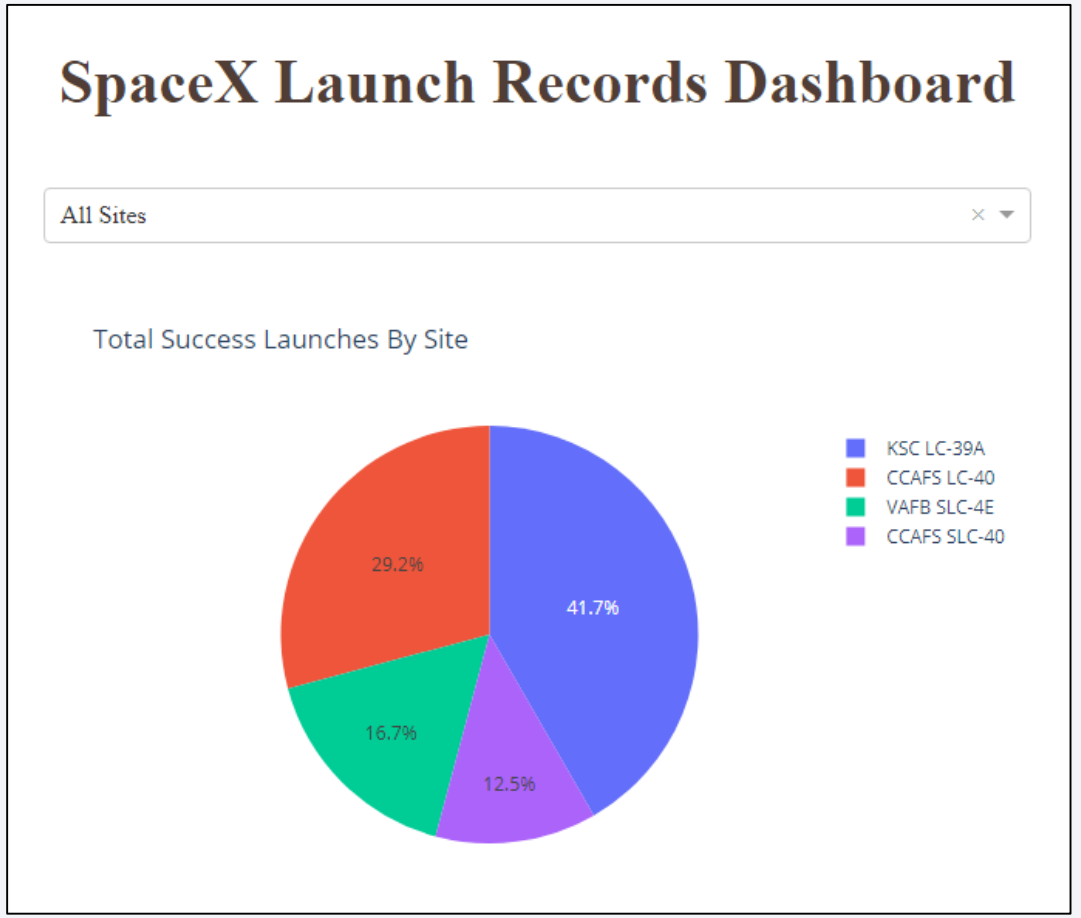
- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being, but are not exactly, right on top of each other.
- The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- The coastline is 0.92 km away from CCAFS LC-40.
- The rail line is 1.33 km away from CCAFS LC-40.




Section 3
Build a Dashboard with Plotly Dash

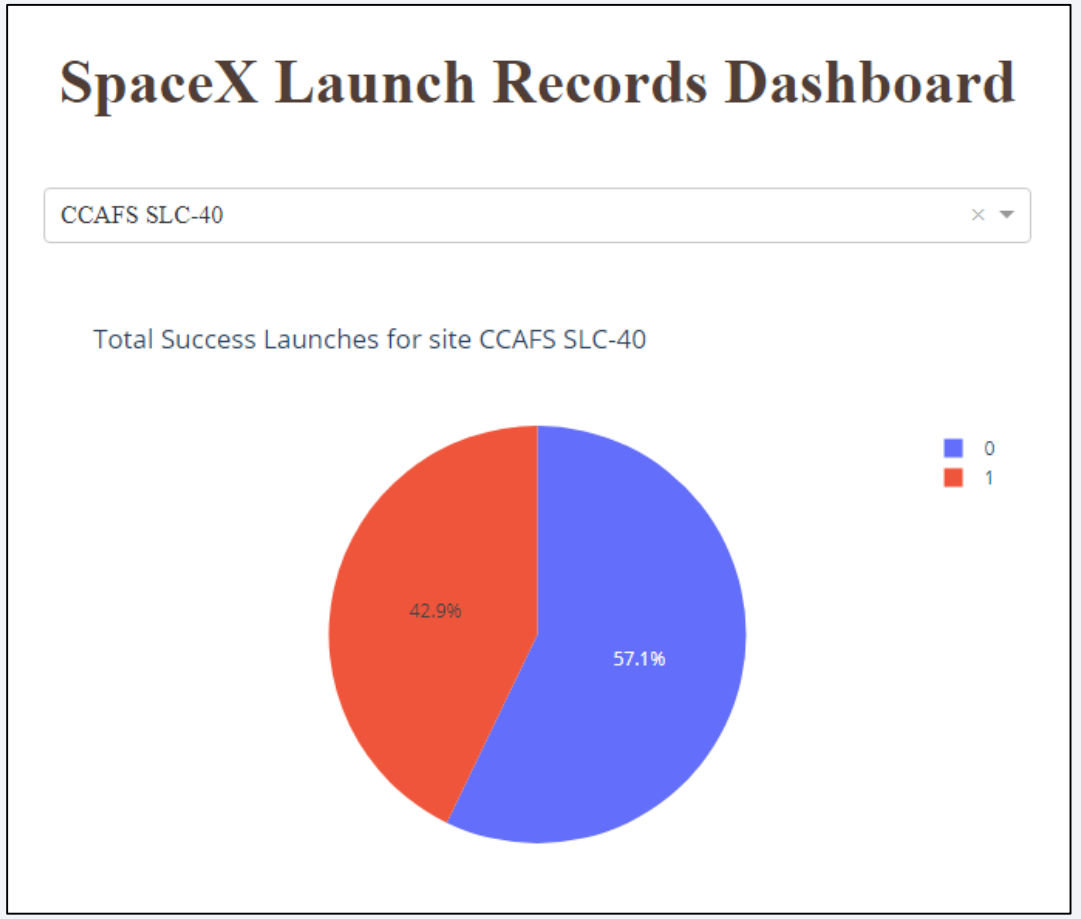
Launch Success Count for All Sites

- 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

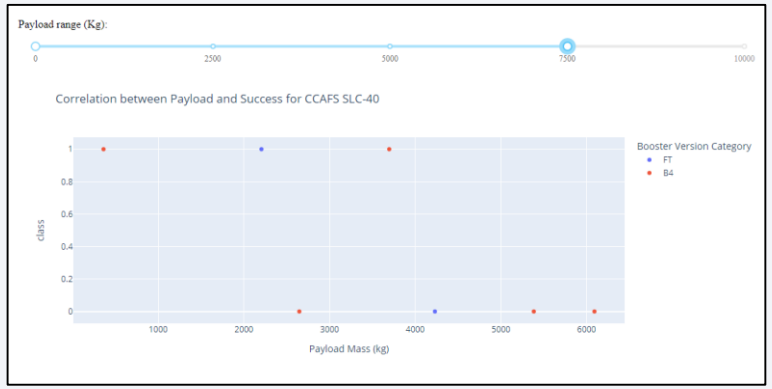
- Falcon 9 first stage failed landings are indicated by the ‘0’ Class (■ blue wedge in the pie chart) and successful landings by the ‘1’ Class (■ red wedge in the pie chart).
- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).



Payload vs. Launch Outcome

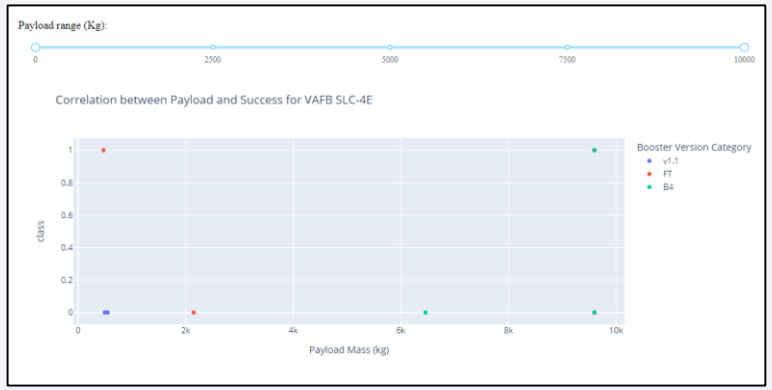
- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate.

CCAFS LC-40



CCAFS SLC-40

KSC LC-39A



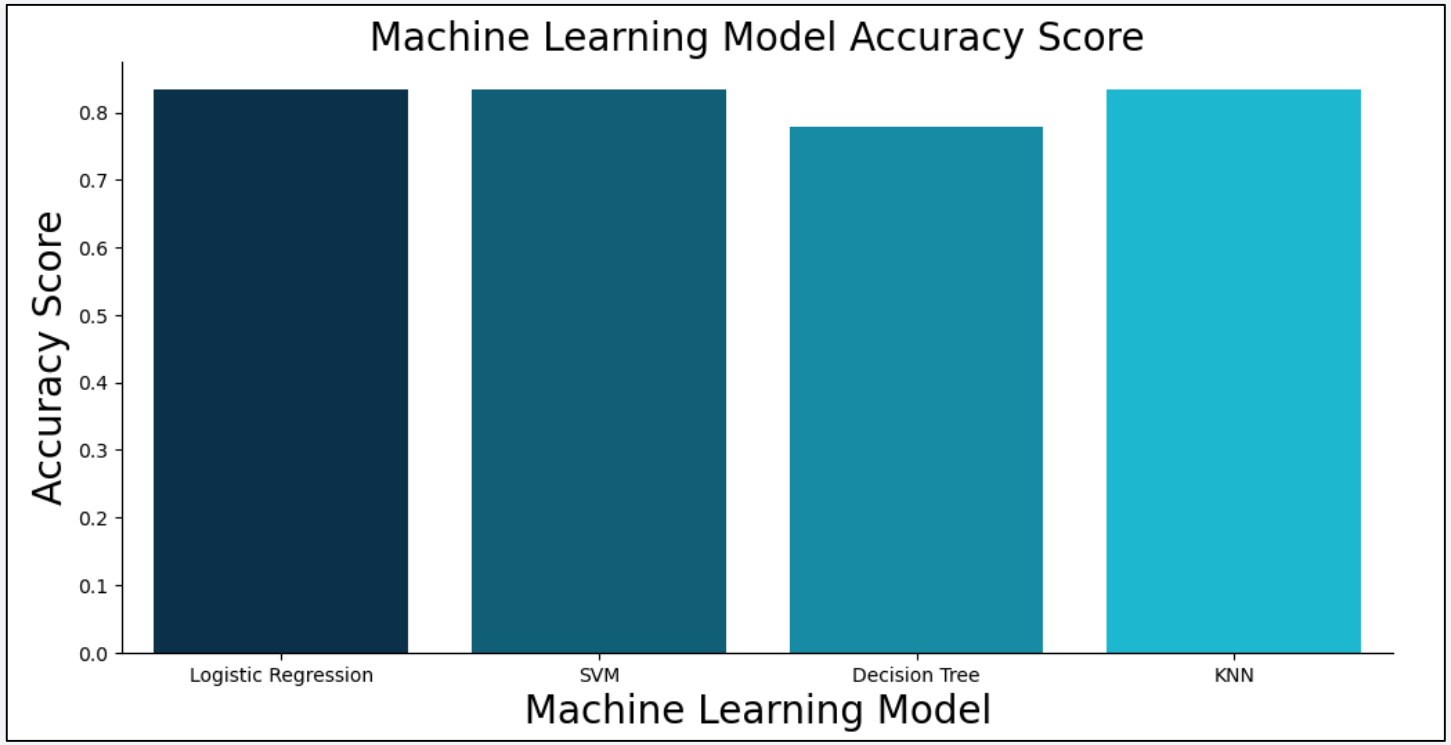
VAFB SLC-4E

Section 4

Predictive Analysis (Classification)

Classification Accuracy

- All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.

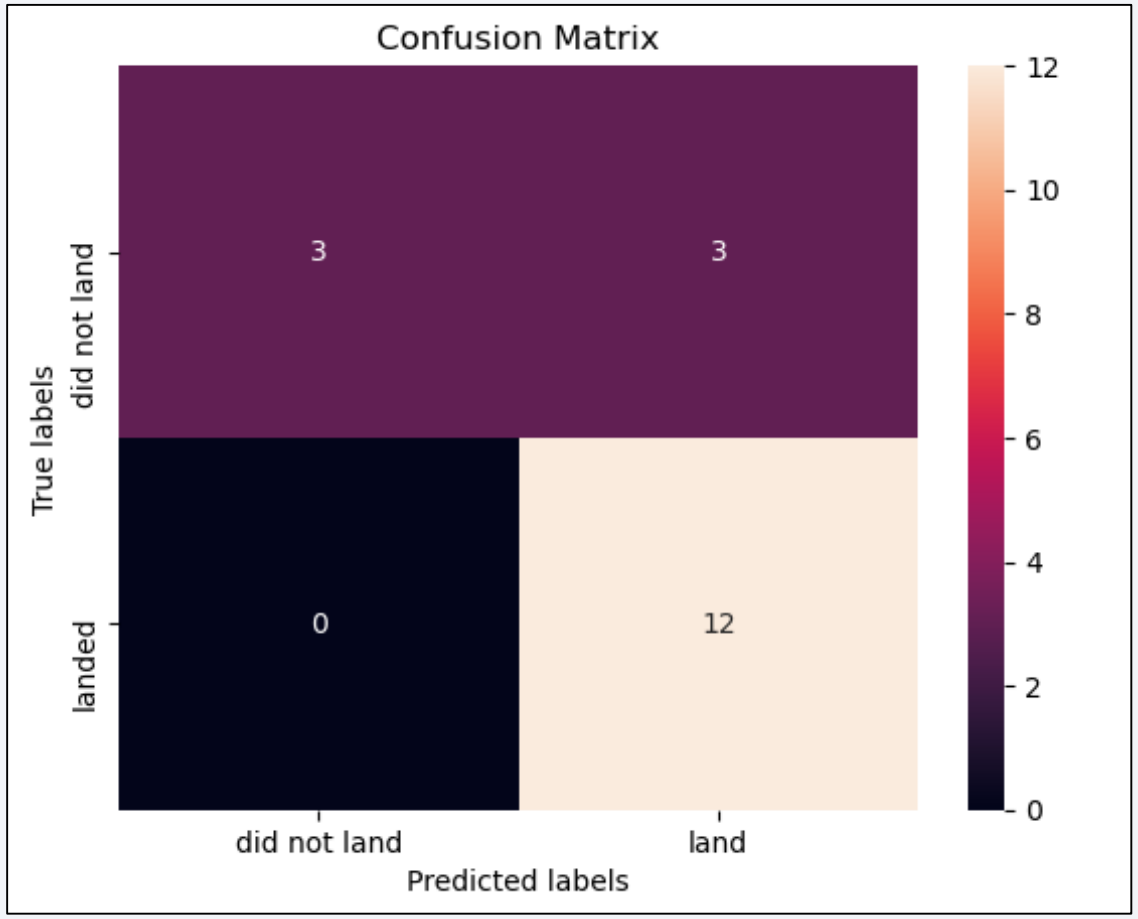


Confusion Matrix

- Shown here is the confusion matrix for the Logistic Regression model.
- Confusion matrices can be read as:

True Negative	**False Positive**
False Negative	**True Positive**

- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives



CONCLUSION

Conclusion

- 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

Initial Data Sources

- **SpaceX API (JSON):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- **Wikipedia (Webpage):** https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- **SpaceX (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- **Launch Geo (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- **Launch Dash (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Processed Data Sets (CSV files)

- **GitHub URL (CSV 1):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_1.csv
- **GitHub URL (Web Scraped):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_web_scraped.csv
- **GitHub URL (CSV 2):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_2.csv
- **GitHub URL (SpaceX):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/Spacex.csv
- **GitHub URL (CSV 3):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/dataset_part_3.csv
- **GitHub URL (Launch Geo):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_launch_geo.csv
- **GitHub URL (Launch Dash):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/spacex_launch_dash.csv

Jupyter Notebooks and Plotly Dashboard File

- **GitHub URL (Data Collection):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/1_Capstone_JonathanMClark_Data_Collection.ipynb
- **GitHub URL (Web Scraping):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/2_Capstone_JonathanMClark_Webscraping.ipynb
- **GitHub URL (Data Wrangling):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/3_Capstone_JonathanMClark_Data_Wrangling.ipynb
- **GitHub URL (EDA with SQL):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/4_Capstone_JonathanMClark_EDA_SQL.ipynb
- **GitHub URL (EDA with Data Visualization):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/5_Capstone_JonathanMClark_EDA_Data_Visualization.ipynb
- **GitHub URL (Folium Maps):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/6_Capstone_JonathanMClark_Launch_Site_Location.ipynb
- **GitHub URL (Dashboard File):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/JonathanMClark_spacex_dash_app.py
- **GitHub URL (Machine Learning):** https://github.com/JonathanMClark/DataScienceCapstone/blob/main/7_Capstone_JonathanMClark_Machine_Learning_Prediction.ipynb

REFERENCE

Reference

- Data Collection
- Exploratory Data Analysis – Data Visualizations
- Exploratory Data Analysis – SQL Queries
- Interactive Folium Maps
- Interactive Plotly Dash Dashboard
- Predictive Analysis (Classification) – Machine Learning

Thank you!