IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Suraj Thakur
29 April 2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data was collected from the SpaceX public API and publicly available data on Wikipedia. Data wrangling included extracting launch outcome information to serve as the dependent variable in the Machine Learning models.
  - SQL queries and data visualizations (static plots, interactive maps, and an interactive dashboard) were created to discover insights about the data set and answer questions.
  - Predictive analysis was pursued using Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) Machine Learning models.
- Summary of results
  - Launch data include info about flight number, date of launch, payload mass, orbit type, launch site, mission outcome and other variables.
  - Logistic Regression, SVM (Support Vector Machine), and KNN (k-Nearest Neighbors) all perform equally well for Machine Learning models on this dataset.

# Introduction

- Project background and context
  - Determining the landing success of the first stage is crucial for cost estimation in rocket launches. SpaceX offers Falcon 9 launches at $62 million, significantly less than competitors, mainly due to their reusable first stage.
  - By predicting first stage landing outcomes, we can accurately assess launch costs. This insight is invaluable for potential competitors seeking to bid against SpaceX for launch contracts.
  - Aim is to predict the successful landing of the Falcon 9 first stage. We will utilize data from Falcon 9 rocket launches as advertised on SpaceX's website.
  - 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?

Section 1

# Methodology

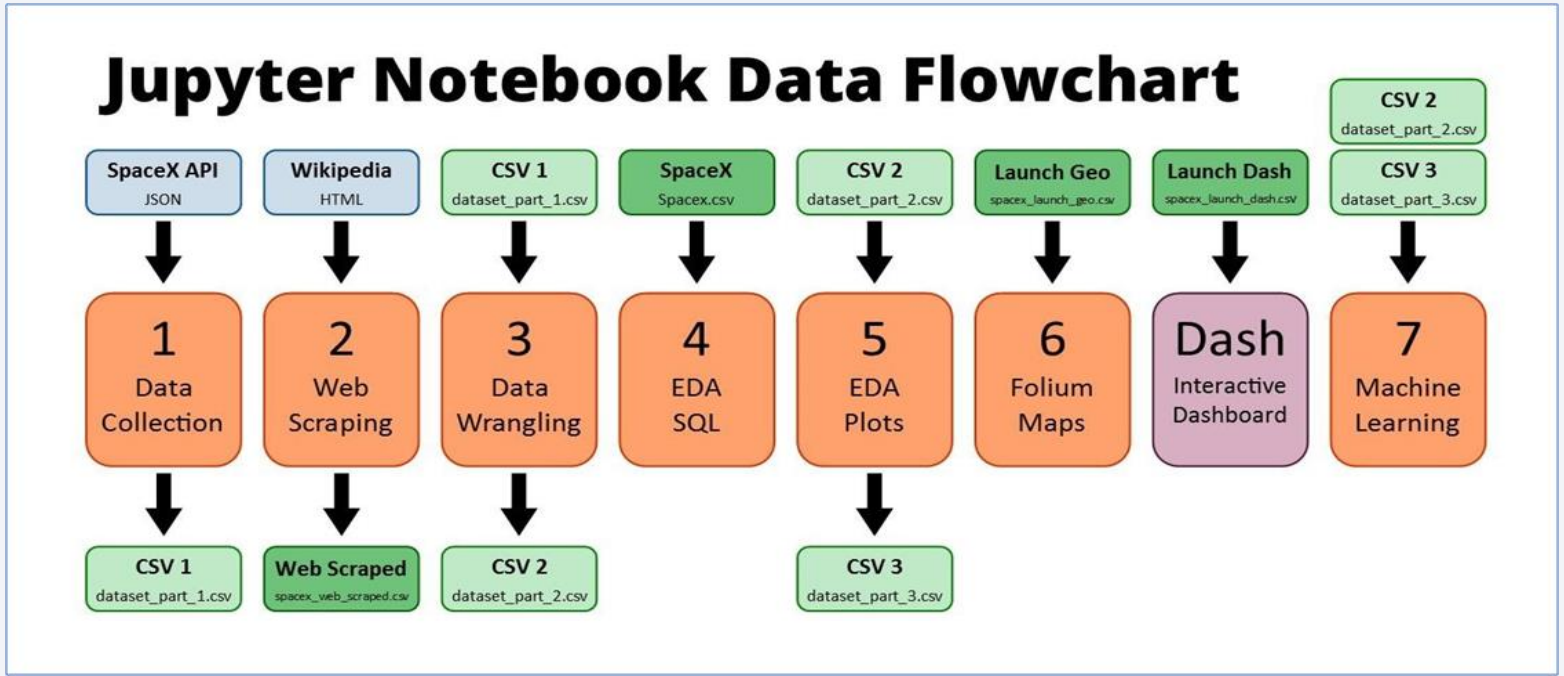
# Methodology

## Executive Summary

- SpaceX API and Wikipedia launch table data was collected.
- Data was cleaned in preparation for visualizations, queries and machine learning model creation.
- Exploratory data analysis (EDA) was done using visualization and SQL.
- Interactive visual analytics were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done.

# Data Collection

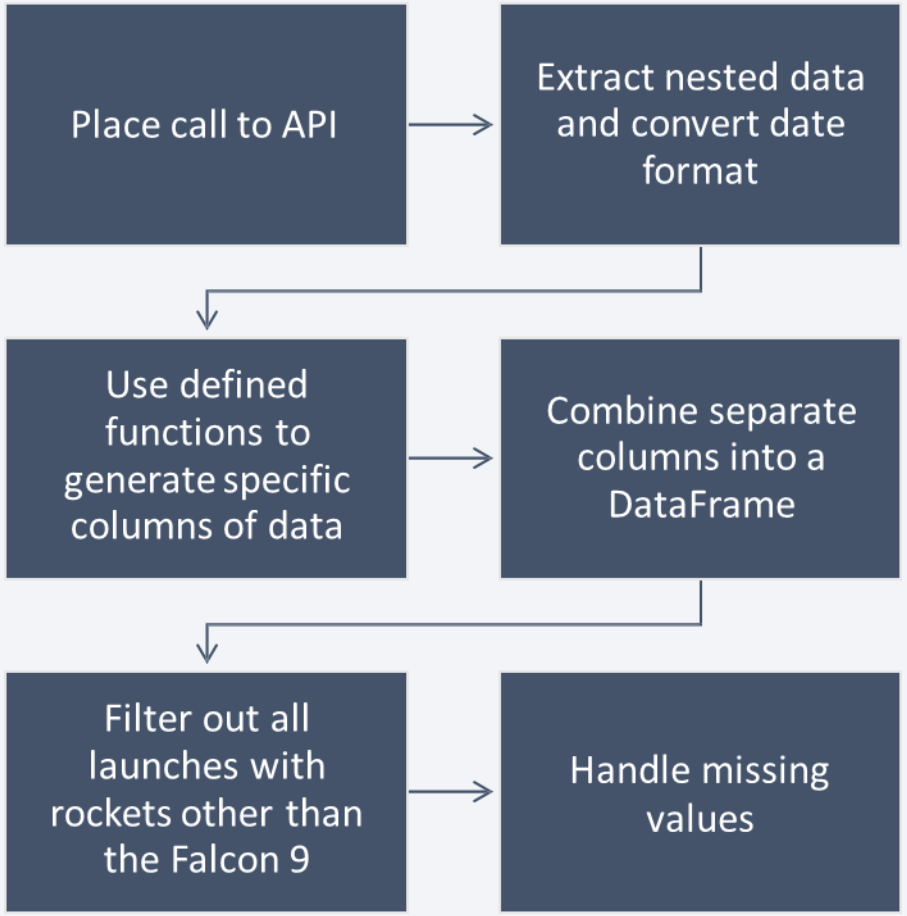
- Data wrangling commenced with a call to the SpaceX API, facilitated by an IBM copy, to extract launch data in JSON format.
- Permanently linked was a Wikipedia page, revised on 9 June 2021, offering essential launch data in HTML tables, augmenting the dataset.
- Additional datasets, noted as darker green .csv files in the diagram's top row, were included to bolster the analysis, potentially providing supplementary variables for predictive modeling.



# Data Collection – SpaceX API

- Data collection from the SpaceX API involved making a GET request to the SpaceX API, which is a RESTful API. The retrieved SpaceX launch data was then parsed and decoded from the GET request response content, resulting in a JSON format. Finally, the JSON data was converted into a Pandas dataframe for further analysis.

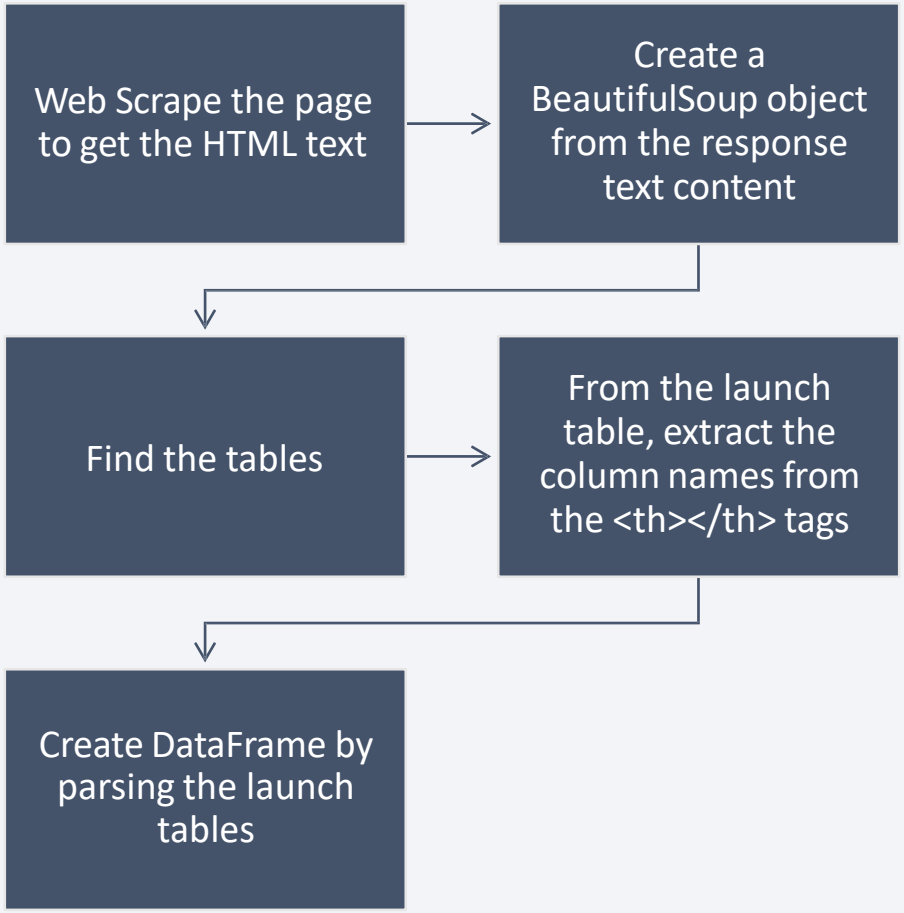
- Here is the GitHub URL of the completed SpaceX API calls notebook "Data Collection for SpaceX using API"




Flowchart of SpaceX API Calls

# Data Collection - WebScraping

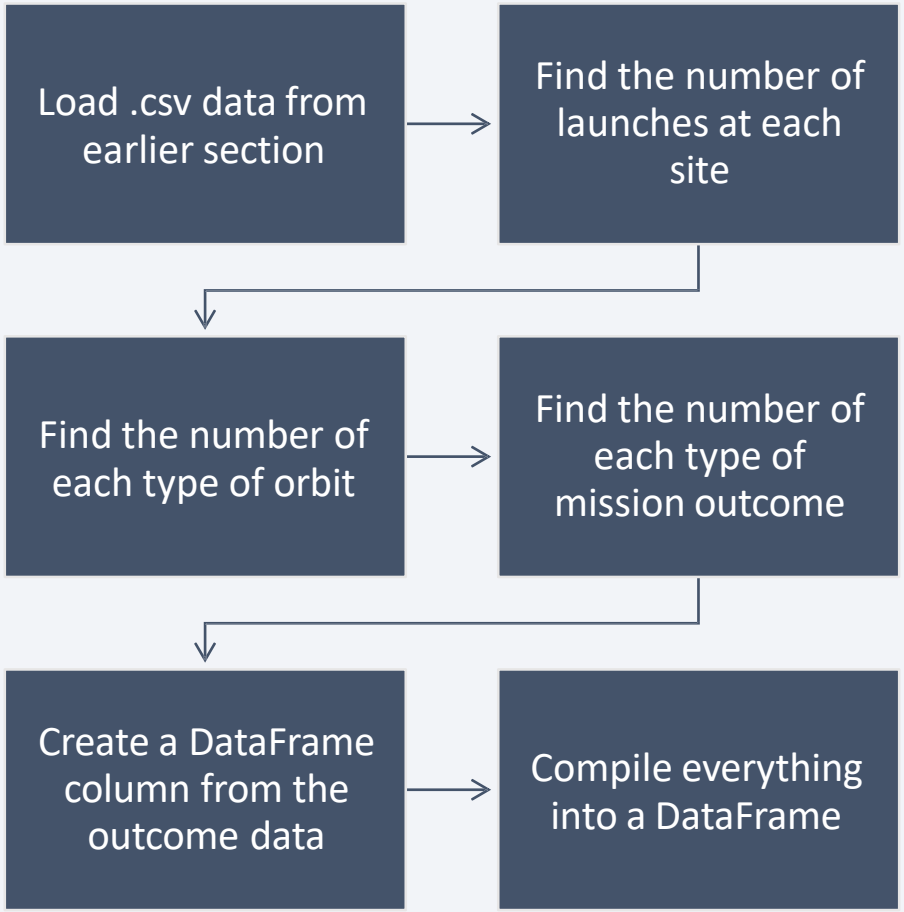
- Wikipedia's SpaceX launch page features tables packed with valuable launch data, ripe for extraction.
- Using web scraping tools like BeautifulSoup, the data can be efficiently retrieved and organized into a Pandas DataFrame for in-depth analysis.
- Following is the GitHub URL of the completed web scraping notebook. "Web scraping SpaceX- Falcon 9 and Falcon Heavy Launches Records using Wikipedia"




Flowchart of Web Scraping

# Data Wrangling

- The .csv file from the initial section required cleaning to ensure data accuracy.
- Cleanup efforts focused on refining launch sites, orbit types, and mission outcomes.
- Mission outcome types were condensed into a binary classification: 1 denoting a successful Falcon 9 first stage landing, and 0 representing failure.
- The updated classification was incorporated into the DataFrame to facilitate subsequent analysis.




Flowchart of 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 time
  - Line plot to see mission outcome trend by year.
- GitHub URL: "Insightful EDA Visualization"

# EDA with SQL

- Queries were written to extract information about:
  - Launch sites: Location query for launch sites data
  - Payload masses: Extracting payload mass details
  - Dates: Querying launch dates information
  - Booster types: Retrieving booster type specifics
  - Mission outcomes: Capturing mission success outcomes

- GitHub URL: “Exploratory Data Analysis with SQL on SpaceX Data”

# Build an Interactive Map with Folium

- Outline the map elements utilized in the Folium map:
  - Markers were incorporated to pinpoint launch sites and the NASA Johnson Space Center.
  - Circles were integrated to denote the launch sites.
  - Lines were drawn to illustrate distances to adjacent 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

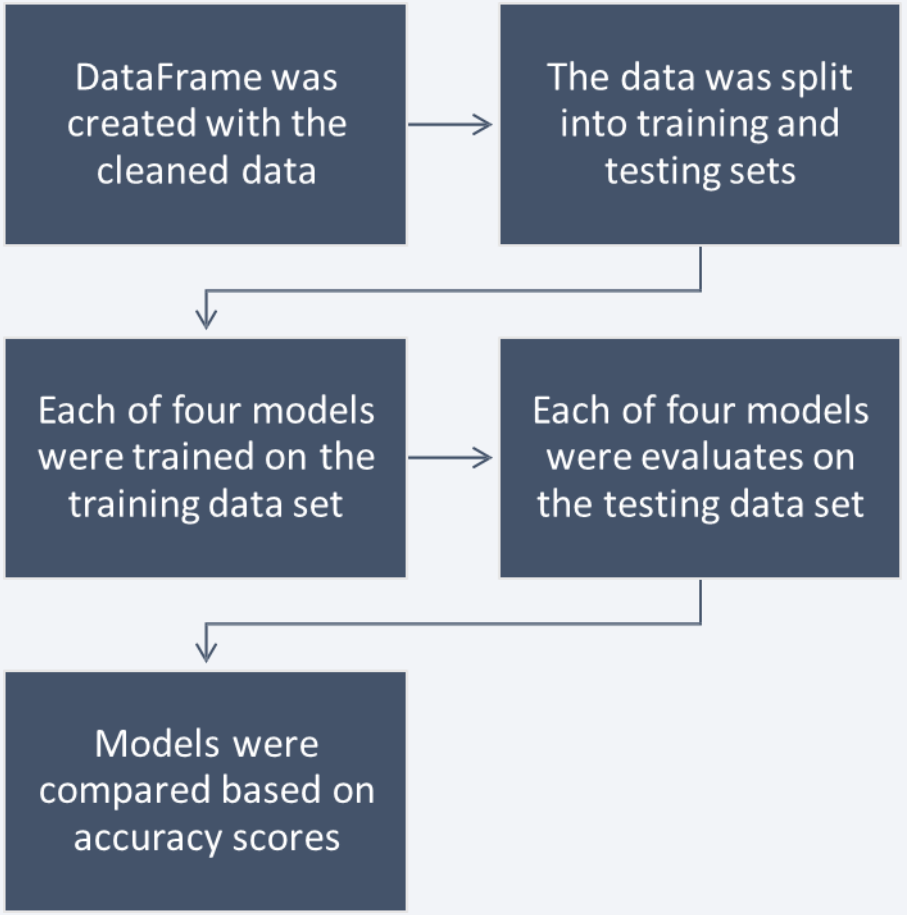
- GitHub URL: “Folium Launch Sites Locations Analysis”

# Build a Dashboard with Plotly Dash

- The input dropdown is used to select one or all launch sites for the pie chart and scatterplot.
- The pie chart displays one of two things:
  - For All Sites – the distribution of successful Falcon 9 first stage landings between the sites
  - For One Site – the distribution of successful and failed Falcon 9 first stage landings for that site
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displays the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.
- GitHub URL: "Interactive Dashboard with Plotly Dash"

# Predictive Analysis (Classification)

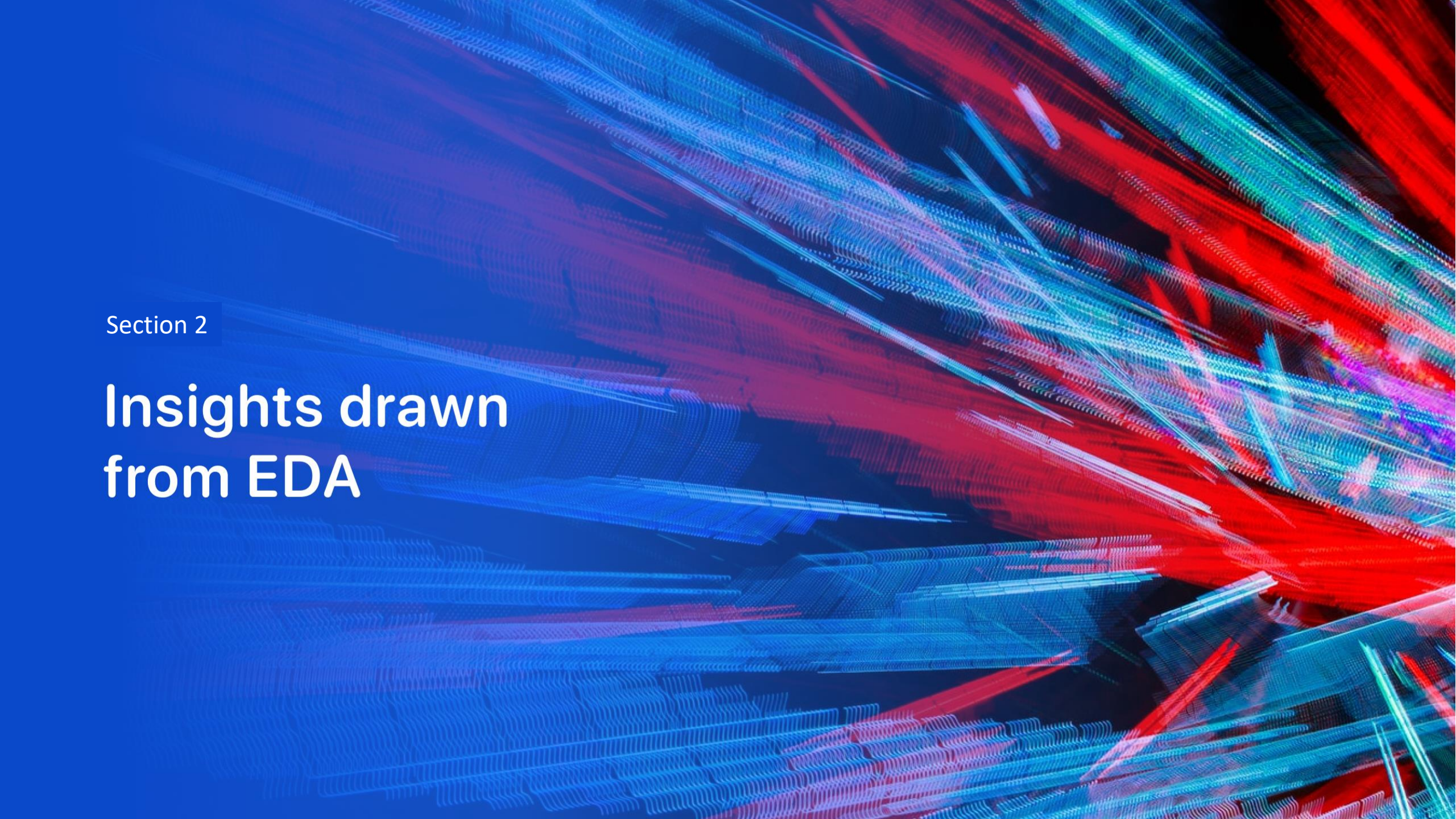
- The dataset underwent division into training and testing sets.
- Machine learning models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN), were trained on the training dataset.
- Hyperparameters were assessed using GridSearchCV(), and the optimal parameters were chosen based on '.best_params_'.
- Utilizing the optimal hyperparameters, each model was evaluated for accuracy on the testing dataset.
- GitHub URL: "SpaceX Machine Learning Prediction"




Flowchart of Machine Learning

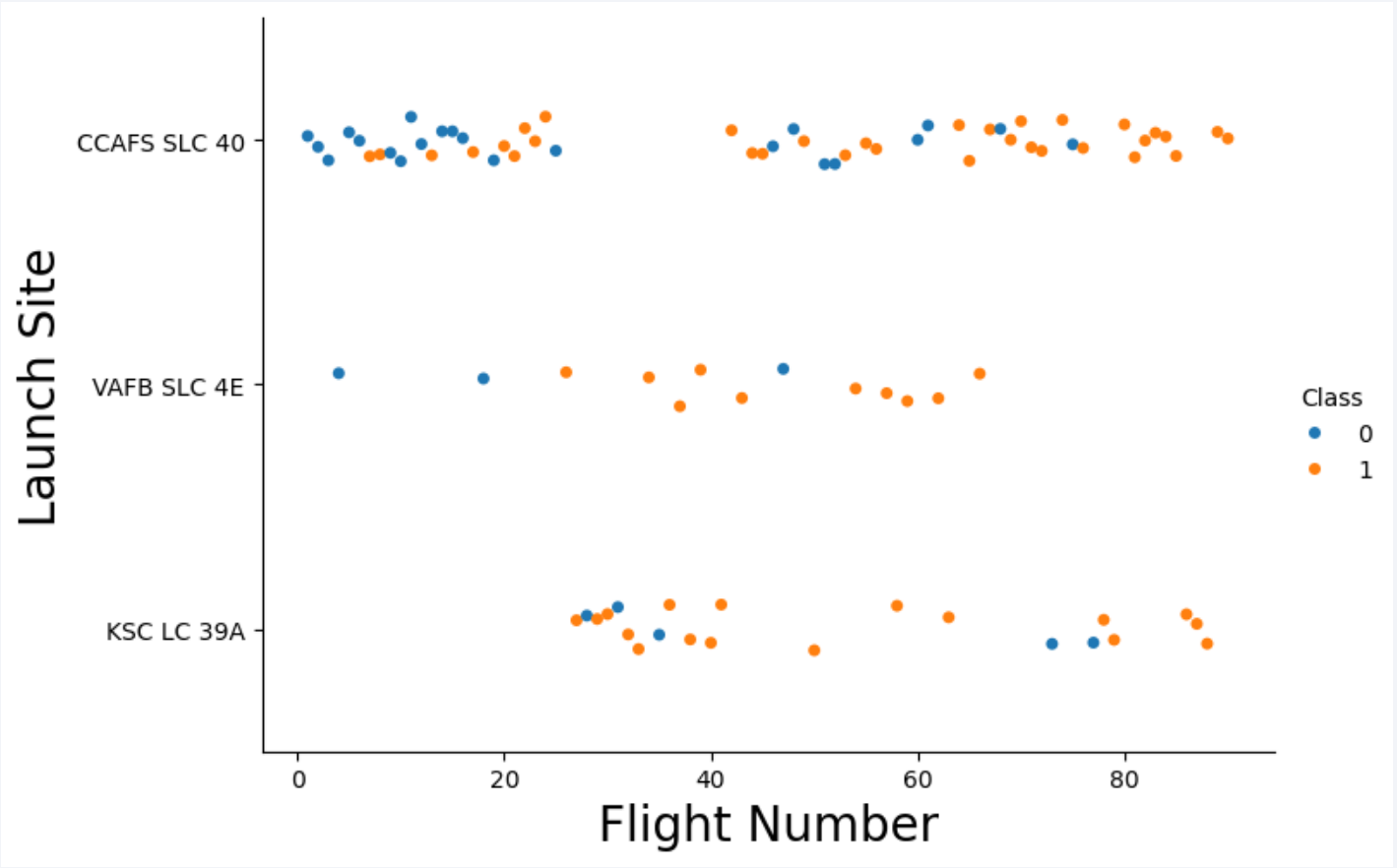
# Results

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


Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site

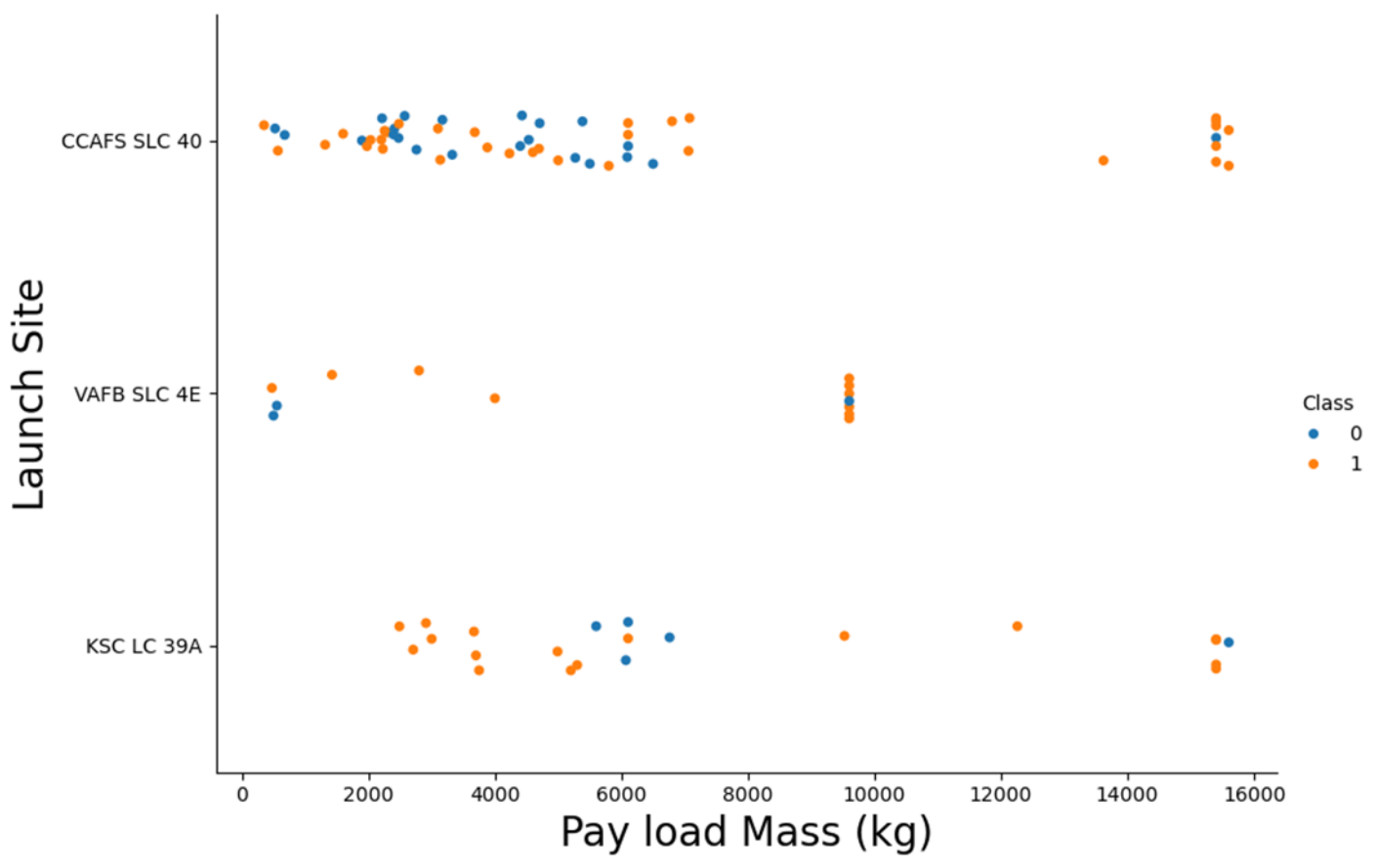
- Success rate varies noticeably with launch site.
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases




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

# Payload vs. Launch Site

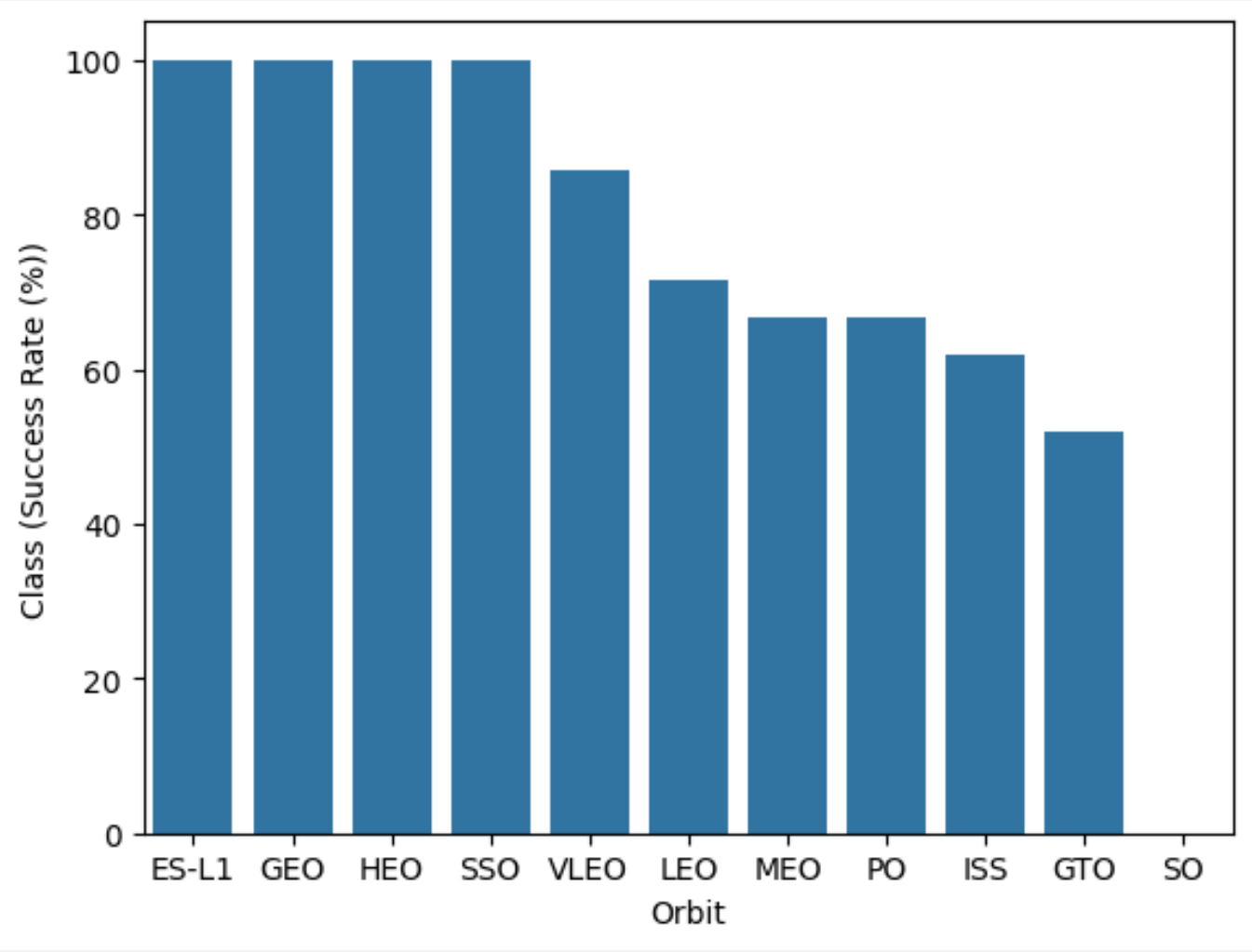
- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.
- The failed landings at the KSC LC 39A launch site are all grouped around a narrow band of payload masses.




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

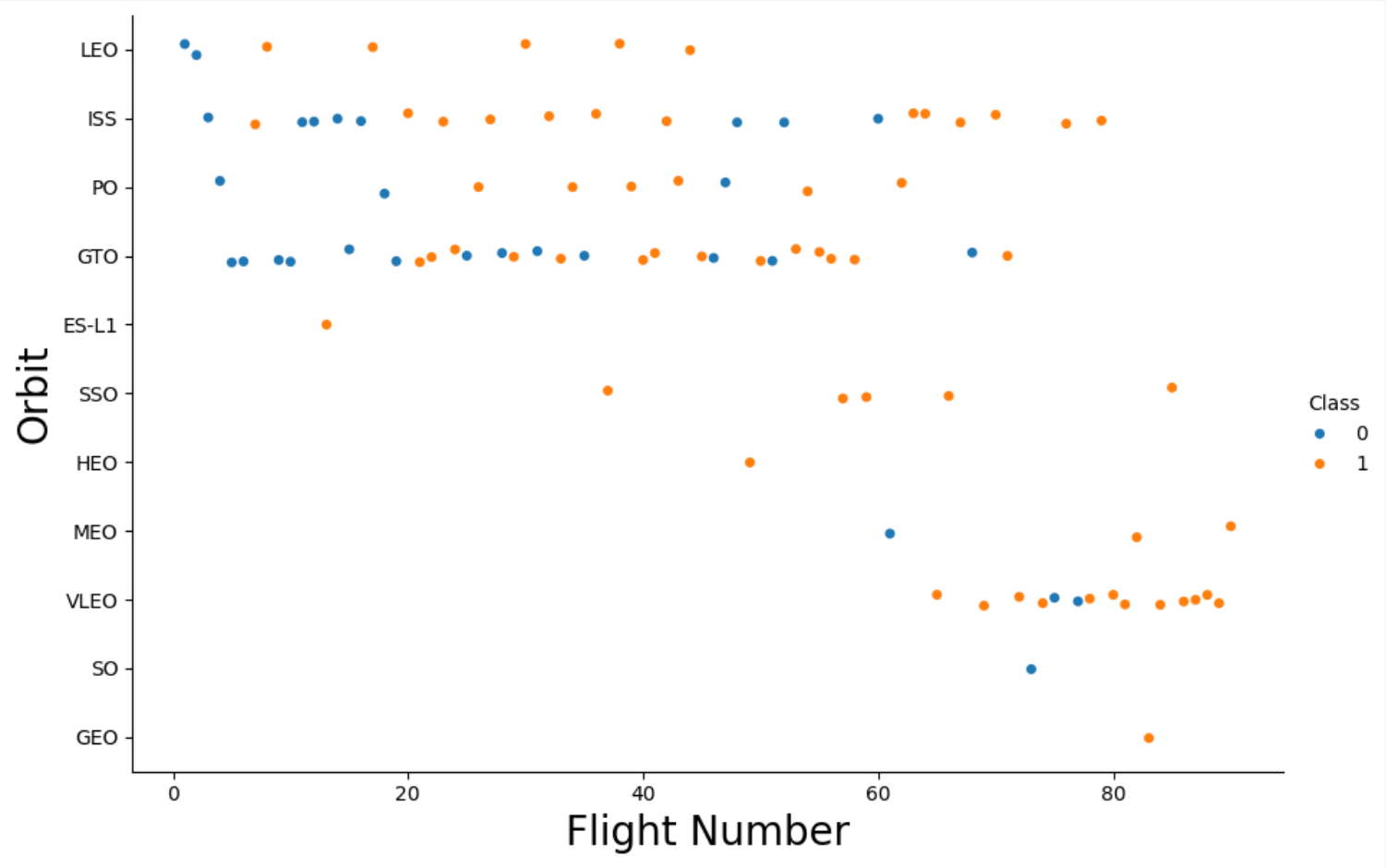
# Success Rate vs. Orbit Type

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



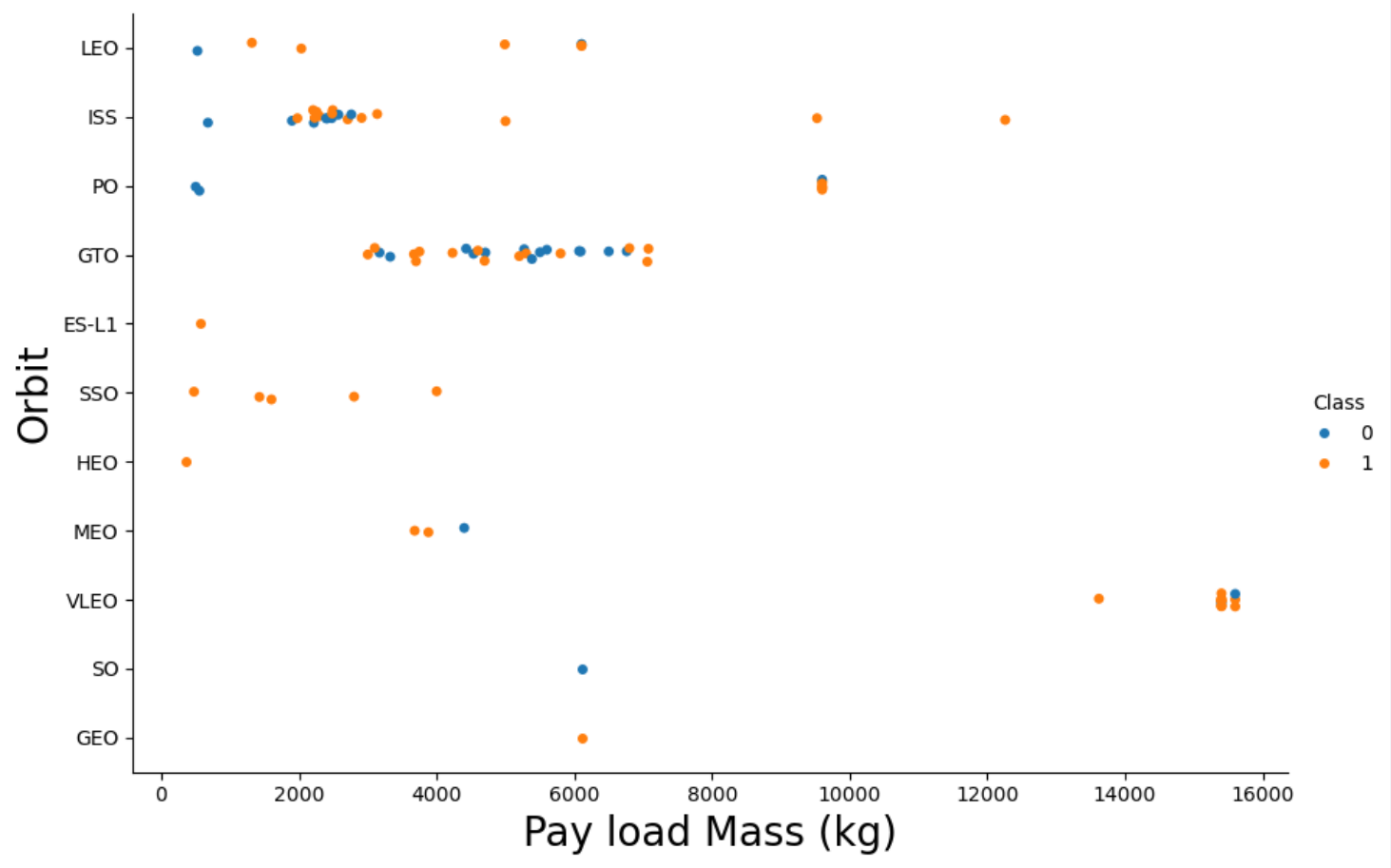
# Flight Number vs. Orbit Type

- There is a correlation between flight number and success rate with larger flight numbers being associated with higher success rates.



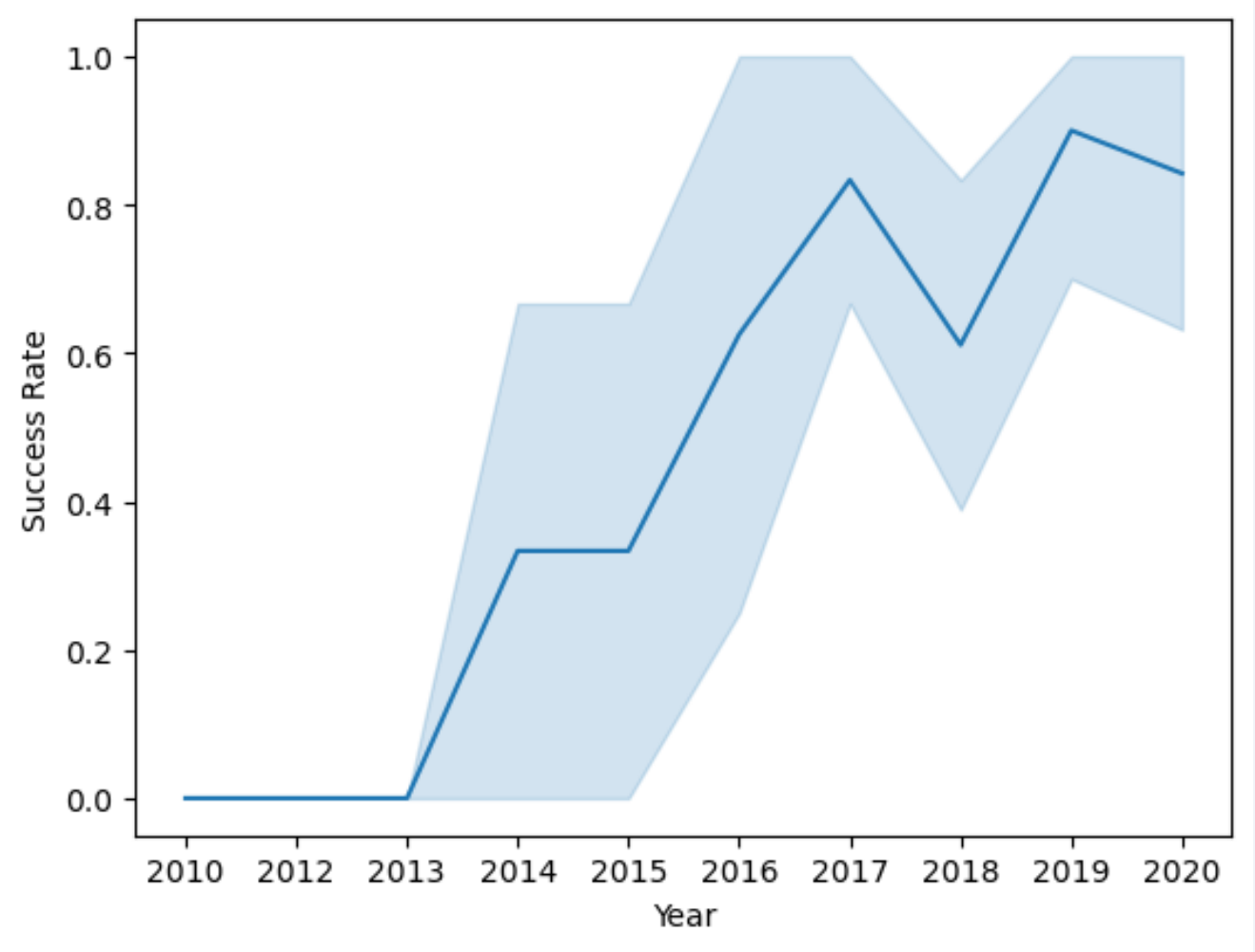
# Payload vs. Orbit Type

- Some orbit types have better success rates than others.
- Success rate appears to have no obvious correlation with payload mass.



# Launch Success Yearly Trend

- The success rate has increased significantly over the years.



# All Launch Site Names

- **Question:** What are the names of the unique launch sites?
- **Query:** SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
- **Result:**

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

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

# Launch Site Names Begin with 'CCA'

- **Task:** Find 5 records with launch sites that begin with `CCA`.
- **Query:** SELECT * FROM 'SPACEXTBL' 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 | 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 |

- **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(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

- **Result:**

| Total Payload Mass(Kgs) | Customer |
|---|---|
| 45596 | NASA (CRS) |

- **Explanation:** The total payload carried by boosters from NASA is 45596 kg.

# Average Payload Mass by F9 v1.1

- **Question:** What is the average payload mass carried by booster version F9 v1.1?
- **Query:**

```
SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE
Booster_Version LIKE 'F9 v1.1%';
```

- **Result:**

| Payload Mass Kgs | Customer | Booster_Version |
|---|---|---|
| 2534.6666666666665 | MDA | F9 v1.1 B1003 |

- **Explanation:** The average payload mass carried by booster version F9 v1.1 is 2,534 kg.

# First Successful Ground Landing Date

- **Question:** On which date did the first successful landing outcome on ground pad occur?
- **Query:**
```
SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
```
- **Result:**

| MIN(DATE) |
| --- |
| 01-05-2017 |

- **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, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 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 "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP
BY "Mission_Outcome";
```

- **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, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

- **Result:**

| Booster_Version | Payload | PAYLOAD_MASS__KG_ |
|---|---|---|
| F9 B5 B1048.4 | Starlink 1 v1.0, SpaceX CRS-19 | 15600 |
| F9 B5 B1049.4 | Starlink 2 v1.0, Crew Dragon in-flight abort test | 15600 |
| F9 B5 B1051.3 | Starlink 3 v1.0, Starlink 4 v1.0 | 15600 |
| F9 B5 B1056.4 | Starlink 4 v1.0, SpaceX CRS-20 | 15600 |
| F9 B5 B1048.5 | Starlink 5 v1.0, Starlink 6 v1.0 | 15600 |
| F9 B5 B1051.4 | Starlink 6 v1.0, Crew Dragon Demo-2 | 15600 |
| F9 B5 B1049.5 | Starlink 7 v1.0, Starlink 8 v1.0 | 15600 |
| F9 B5 B1060.2 | Starlink 11 v1.0, Starlink 12 v1.0 | 15600 |
| F9 B5 B1058.3 | Starlink 12 v1.0, Starlink 13 v1.0 | 15600 |
| F9 B5 B1051.6 | Starlink 13 v1.0, Starlink 14 v1.0 | 15600 |
| F9 B5 B1060.3 | Starlink 14 v1.0, GPS III-04 | 15600 |
| F9 B5 B1049.7 | Starlink 15 v1.0, SpaceX CRS-21 | 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 |

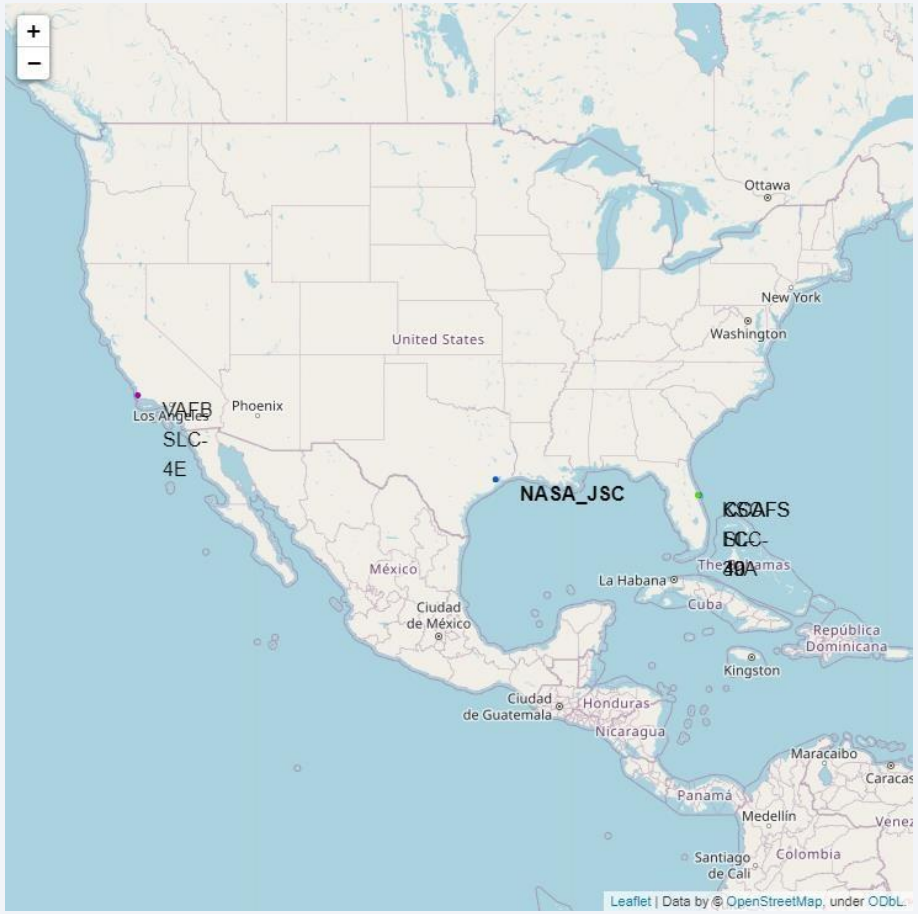
- **Explanation:** The most common landing outcome was ‘not attempted’.

Section 3

# Launch Sites Proximities Analysis

# Falcon 9 Launch Site Locations

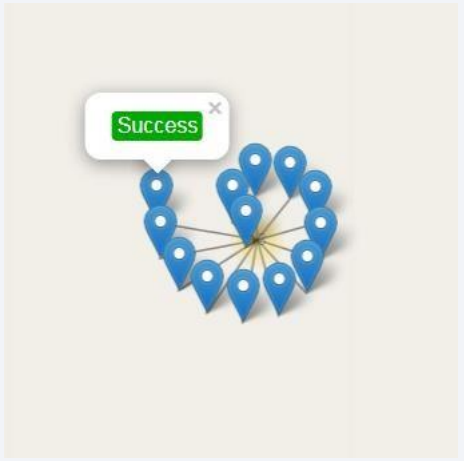
- VAFB SLC-4E (California, USA)
  - Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA)
  - Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Space Launch Complex 40



# Map Markers of Success/Failed Landings

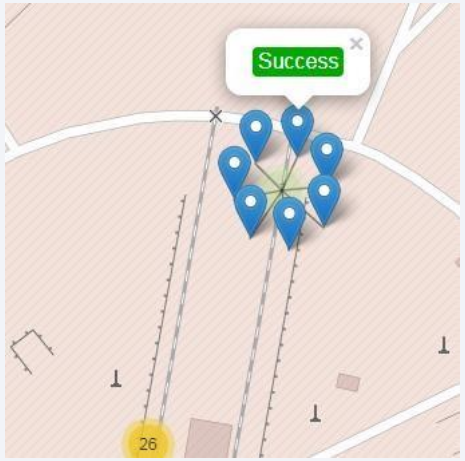
- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative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

VAFB SLC-4E



KSC LC-39A

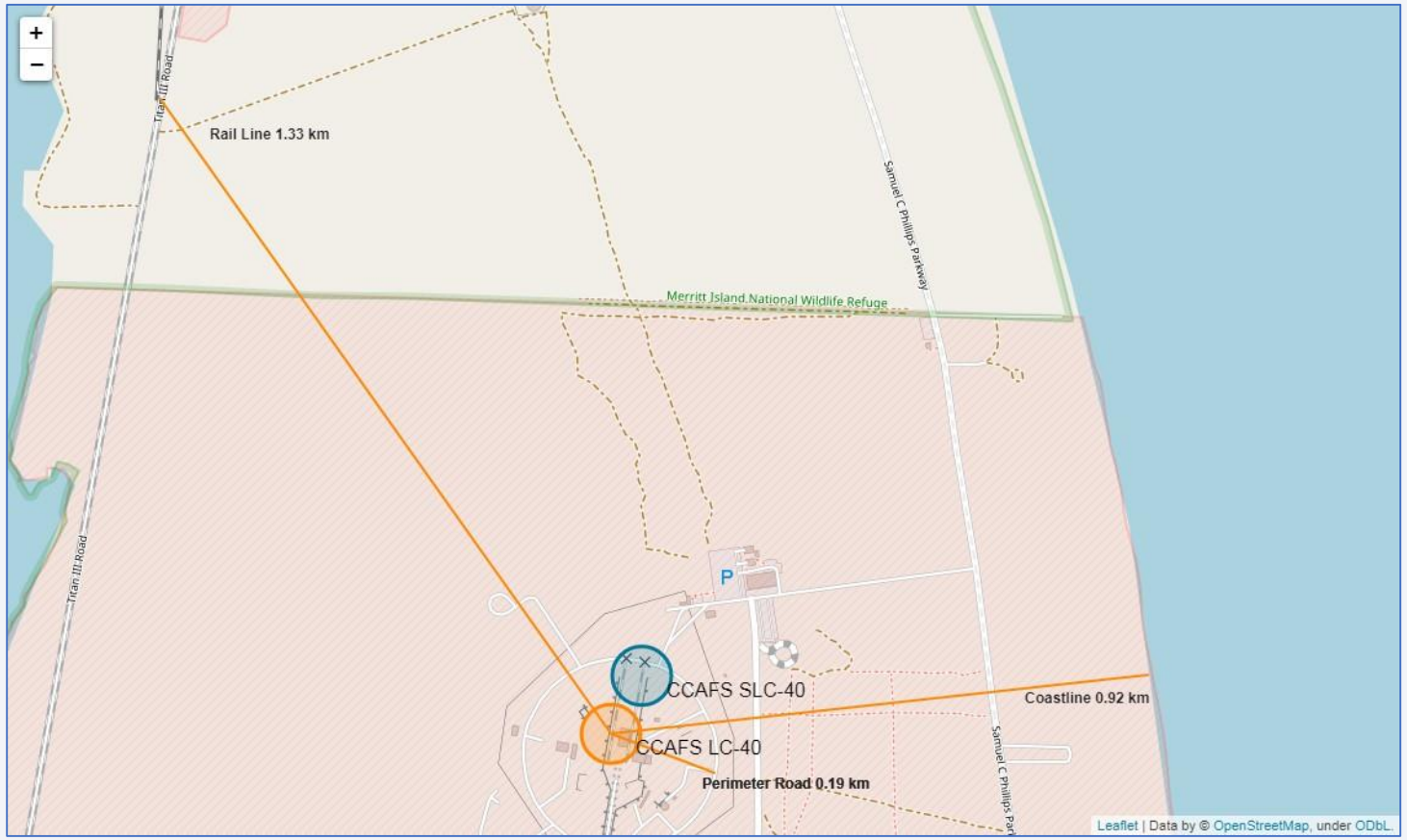
CCAFS LC-40

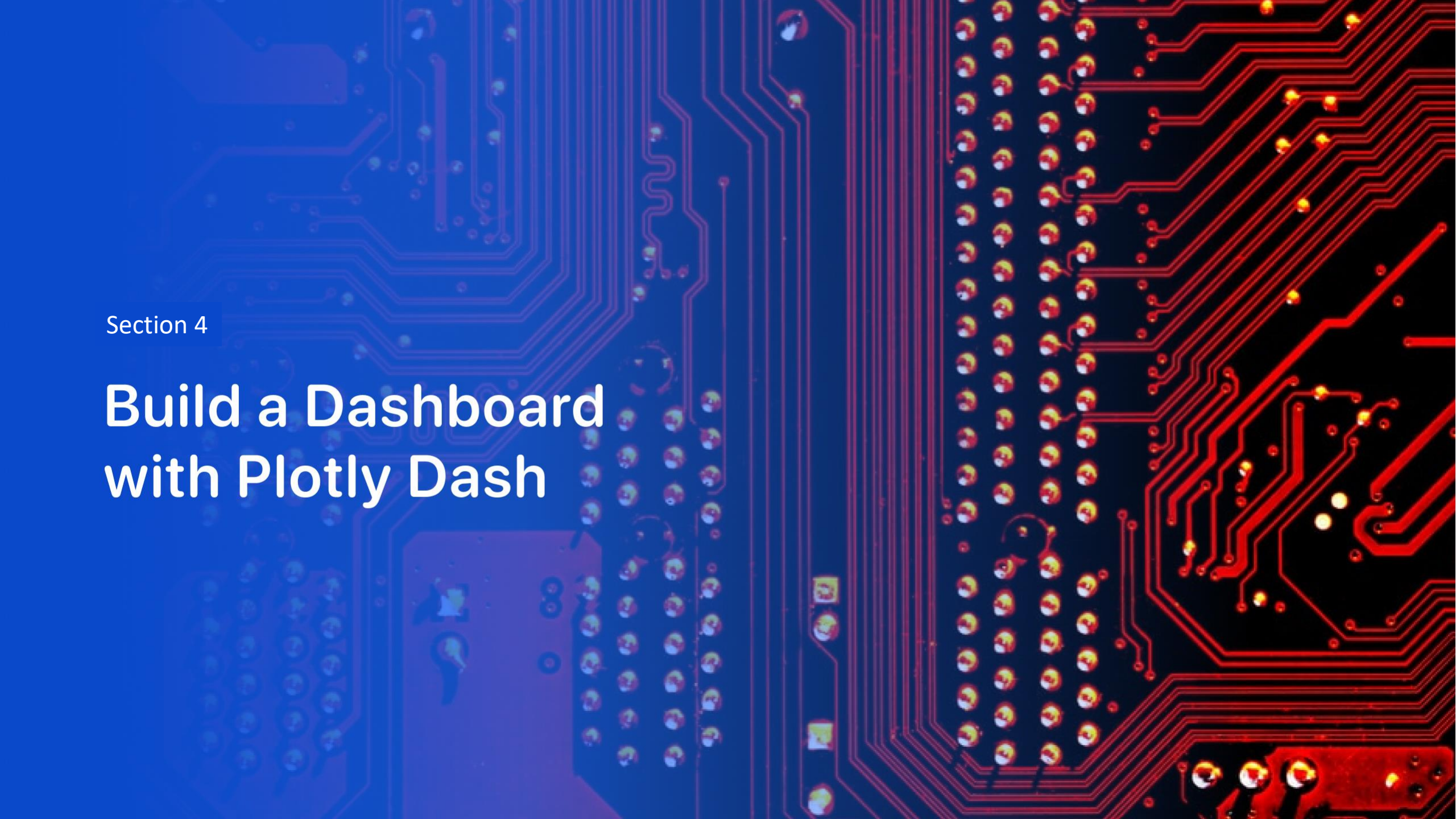


CCAFS SLC-40

# Distance from Launch Site to Proximities

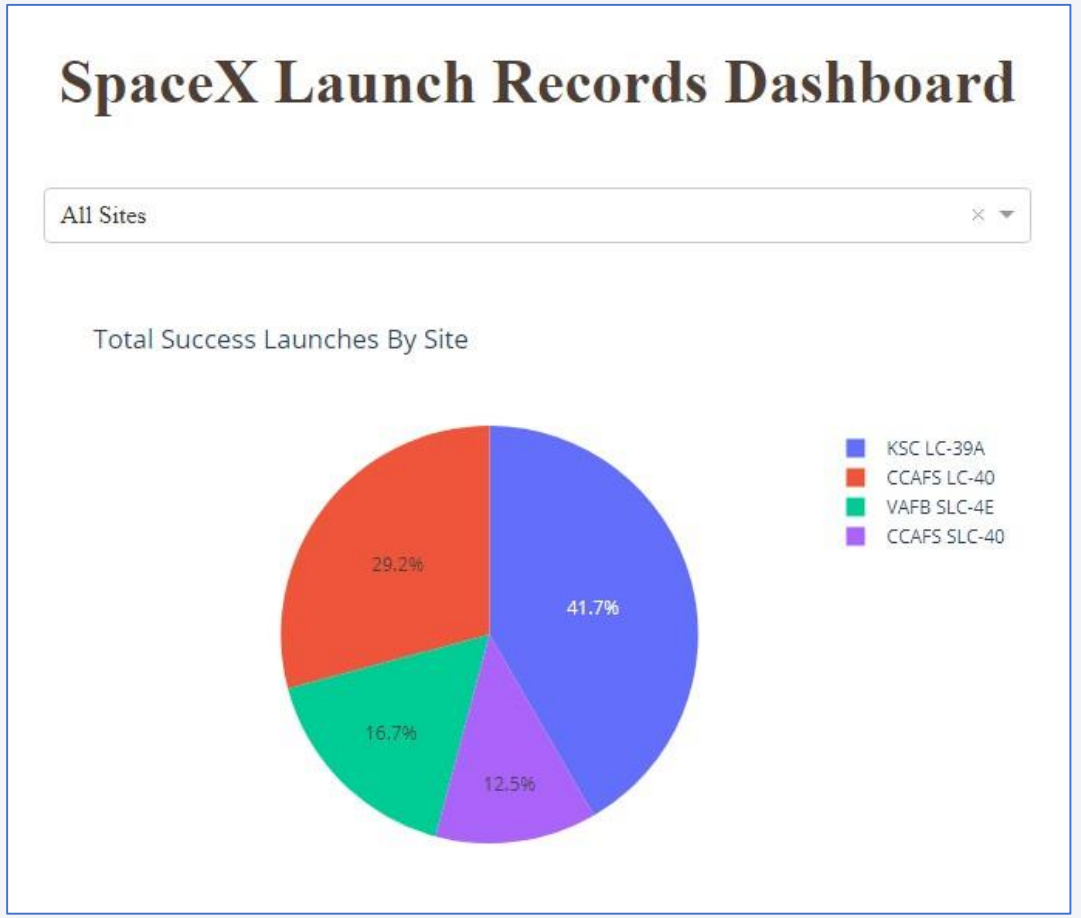
- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being, but are not exactly, right on top of each other.
- The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- The coastline is 0.92 km away from CCAFS LC-40.
- The rail line is 1.33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




Section 4
Build a Dashboard with Plotly Dash

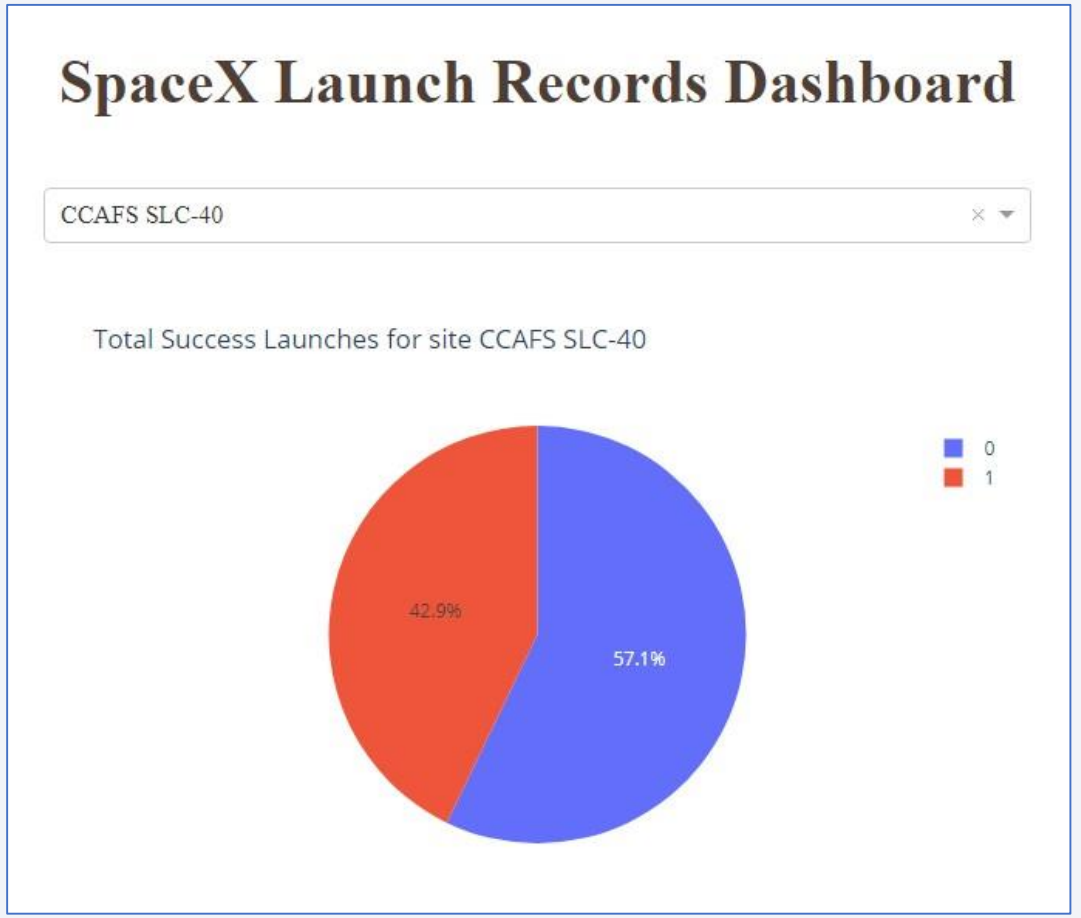
# Launch Success Count for All Sites

- The dropdown menu allows the selection of one or all launch sites.
- With all launch sites selected, the pie chart displays 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 5

# Predictive Analysis (Classification)

# Classification Accuracy

| Method | Test Data Accuracy |
|---|---|
| Logistic_Reg | 0.833333 |
| SVM | 0.833333 |
| Decision Tree | 0.611111 |
| KNN | 0.833333 |

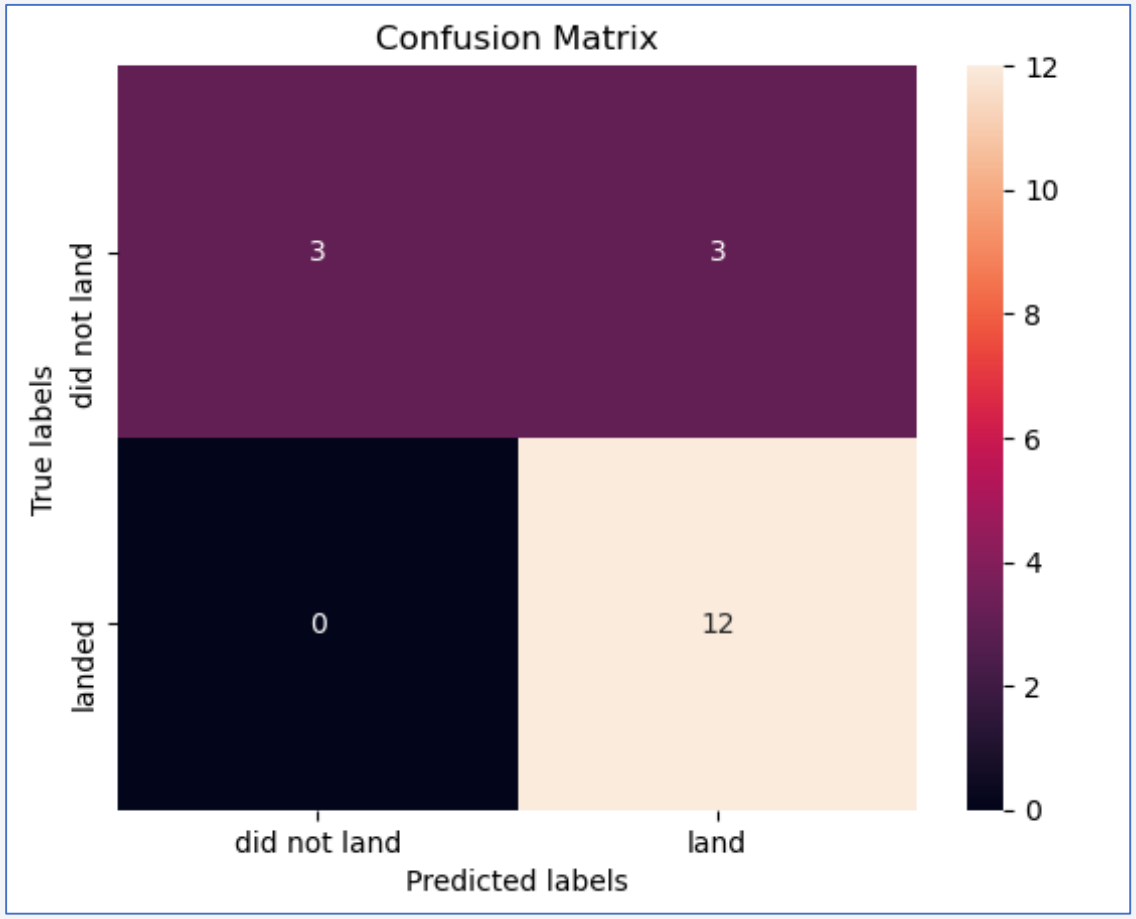
- All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.

# 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



# Conclusions

- SpaceX does not have a perfect track record of Falcon 9 first stage landing
- outcomes.
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards 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 Sets

- 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

## Jupyter Notebooks and Dashboard Python File

- GitHub URL (Data Collection for SpaceX using API): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/1)%20Data%20Collection%20for%20SpaceX%20using%20API.ipynb
- GitHub URL (Web scraping SpaceX- Falcon 9 and Falcon Heavy Launches Records using Wikipedia): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/2)%20Web%20scraping%20SpaceX-%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20using%20Wikipedia.ipynb
- GitHub URL (SpaceX-Data Wrangling): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/3)%20SpaceX-Data%20Wrangling.ipynb
- GitHub URL (Exploratory Data Analysis with SQL on SpaceX Data): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/4)%20Exploratory%20Data%20Analysis%20with%20SQL%20on%20SpaceX%20Data.ipynb
- GitHub URL (Insightful EDA Visualization): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/5)%20Insightful%20EDA%20Visualization.ipynb
- GitHub URL (Folium Launch Sites Locations Analysis): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/6)%20Folium%20Launch%20Sites%20Locations%20Analysis.ipynb
- GitHub URL (Interactive Dashboard with Plotly Dash): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/7)%20Interactive%20Dashboard%20with%20Plotly%20Dash.py
- GitHub URL (SpaceX Machine Learning Prediction): https://github.com/SurajThakur10/Falcon9_1st-Stage_Landing_Prediction/blob/main/8)%20SpaceX%20Machine%20Learning%20Prediction.ipynb

# Thank you!