
# Winning Space Race with Data Science

Arnav Mankad
February 23, 2025

# Outline

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

# Executive Summary

## Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analysis
- Predictive Analysis via Machine Learning Classification

## Summary of Results

- Data collected successfully
- Data Analyzed through various plots
- Map analytics through Folium
- Interactive Dashboard created using Plotly
- Predictive analysis results

# Introduction

## Project Background and Context

- SpaceX has advertised that Falcon rocket launches cost 62 million dollars
  - Other providers cost more (more than 165 million dollars) because their **first stage is not reusable**
- The cost of a launch is dependent on determining if the first stage will land

## Project Goals

- This project involves predicting if the first stage of Falcon 9 will land successfully
  - This will ultimately determine the cost of the overall launch

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Rocket launch data requested from SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Counting launch numbers on each site, number and occurrence of each orbit, number of occurrence of mission outcome per orbit category
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Involves data pre-processing, spitting data into training and test sets, train different types of models and assess their accuracy, and find hyperparameters with GridSearchCV

# Data Collection

- Datasets were collected in three distinct ways:
  - Publicly accessible API with launch data in JSON format was used
  - Data was scraped from a Wikipedia page containing launch data
  - Course provided additional data in CSV format

# Data Collection – SpaceX API

- Data was made publicly available at the following API endpoint: https://api.spacexdata.com/
  - A copy of the API response was made available for the project
- Data was extracted from the API response, loaded into a Pandas dataframe, and analyzed further
- GitHub URL of SpaceX API calls notebook: https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataCollectionAPI.ipynb

## Steps of API Data Processing

API Get Request

Nested Data Extraction and Date Format Conversion

Specific, Relevant Columns of Data Generated

Columns Combined to Create Dataframe

Isolate Falcon 9 Launches

Handle Missing Values

# Data Collection – Web Scraping

- Data was scraped from the SpaceX Wikipedia webpage: https://en.wikipedia.org/wiki/SpaceX
- Data was extracted from the site, loaded into a Pandas dataframe, and analyzed further
- GitHub URL of SpaceX web scraping notebook: https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataCollectionWebScraping.ipynb

## Steps of Web Scraping

Web Scraping of Wikipedia Page

BeautifulSoup Object from Response Text Content

Table Selection

Extraction of Column Names from Launch Table

Parsing Launch Tables to Create Dataframe

# Data Wrangling

- Data extracted from a provided CSV file
- Launch sites, orbit types, and mission outcomes specifically processed/reformatted
- Binary classification of landing outcomes (1 = success, 0 = failure)
- Pandas Dataframe created
- GitHub URL of SpaceX web scraping notebook: https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataWrangling.ipynb

## Steps of Data Wrangling

Load CSV data

Find Launch Numbers per Site

Find Orbit Type Numbers

Find Mission Outcome Numbers

Dataframe from Mission Outcome Data

# EDA with Data Visualization

**Scatterplots:**

- Effect of launch site and flight number on mission outcome
- Effect of launch site and payload mass on mission outcome
- Effect of orbit type and flight number on mission outcome
- Effect of orbit type and payload mass on mission outcome

**Bar Charts:**

- Relationship between orbit type and mission outcome

**Line Plots:**

- Mission outcomes year-by-year

**GitHub URL (EDA with Data Visualization):**

- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_EDAwithVisualization.ipynb

# EDA with SQL

## SQL Queries (Not Exhaustive):

- Names of unique launch sites
- 5 records where launch sites begin with ‘CCA’
- Total payload mass carried by NASA-launched boosters
- Average payload mass carried by F9v1.1
- Listing booster names with ground pad success (mass between 4000 and 6000
- Number of successful and failed lands
- Ranking count of landing outcomes between 03-20-2017 and 03-20-2017

## GitHub URL (EDA with SQL):

- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_EDAwithSQL.ipynb

# Build an Interactive Map with Folium

**A Folium map was created, and the following map objects were added:**

- **Markers** for launch sites and the NASA Johnson Space Center
- **Circles** were added for the launch sites
- **Lines** were added to show distance from CCAFS LC-40 to the coastline, CCAFS LC-40 to rail line, CCAFS LC-40 to perimeter road

**GitHub URL (Folium Map)**

- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_VisualAnalyticsFolium.ipynb

# Build a Dashboard with Plotly Dash

**The dashboard created using Plotly Dash contains a dropdown input to select data from either one or all launch sites**

- **Pie charts:**
  - For one launch site, pie chart displays distribution of successful and failed Falcon 9 first stage landings
  - For all launch sites, pie chart displays distribution of successful Falcon 9 first stage landings for all sites
- **Scatterplot:**
  - The scatterplot displays Falcon 9 first stage landings based on payload mass, mission outcome, and booster version
  - Input slider filters payload masses

**GitHub URL (Plotly Dash Dashboard):**

- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_InteractiveDashboardPlotly.py

# Predictive Analysis (Classification)

## Machine Learning Flowchart:

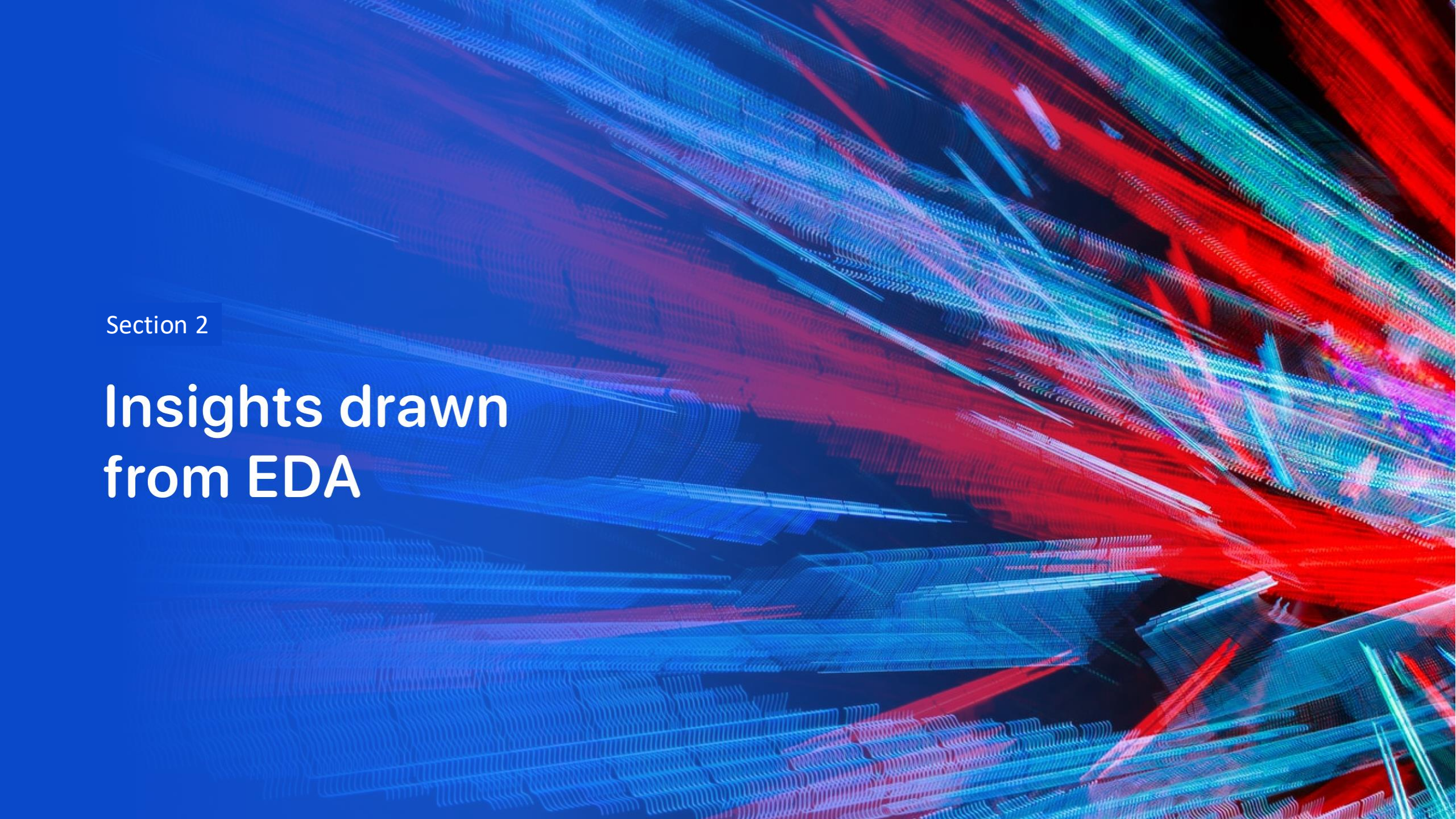
- Pandas Dataframe created from cleaned data
- Data was split into training and test sets
- Four models (Logistic regression, SVM, Decision Tree, KNN) were trained on the training set
- Hyper-parameters were evaluated using GridSearchCV() to select best parameters
- The models were tested with the test set (using optimal hyper-parameters) and accuracy scores were determined

## GitHub URL (Predictive Analysis):

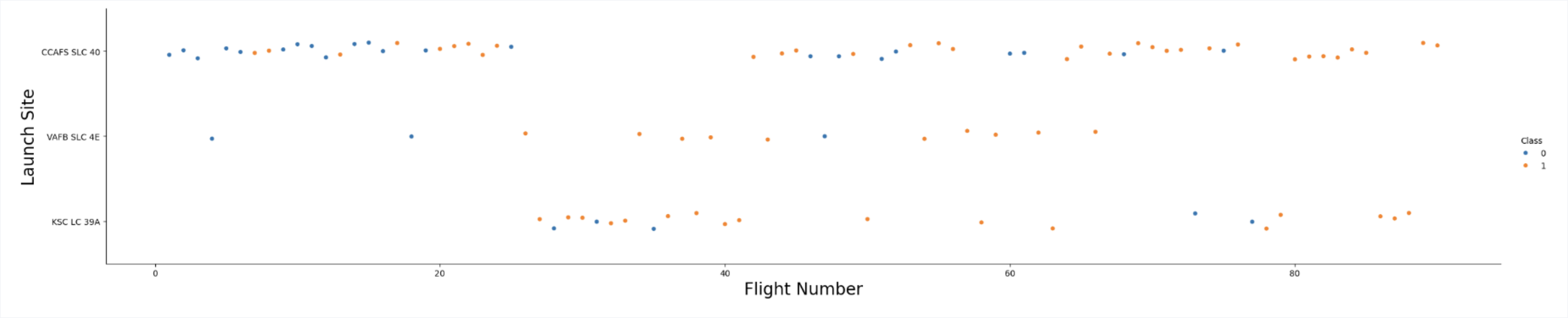
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_MachineLearningPrediction.ipynb

# Results

- Insights drawn from exploratory data analysis
- Launch sites proximities analysis
- Dashboard built with Plotly Dash
- Predictive analysis with classification

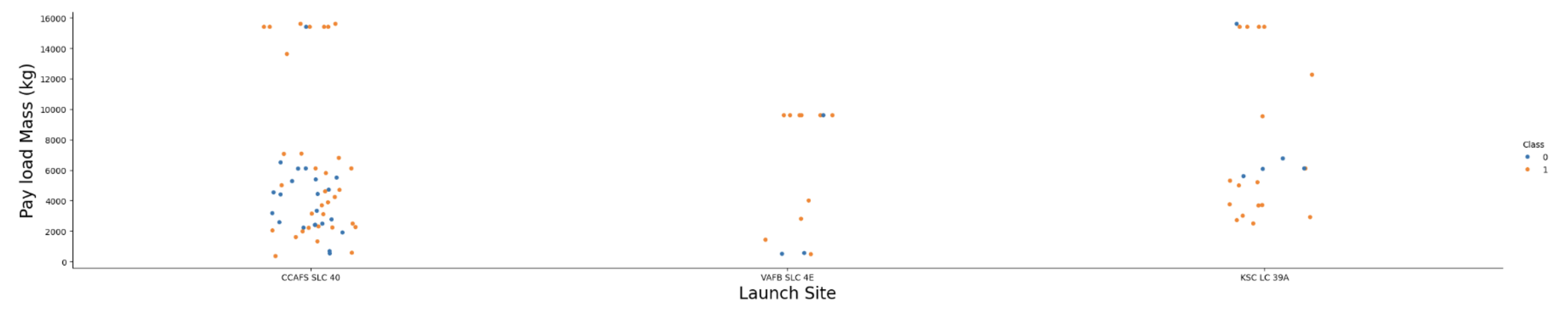

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site




- Launch site plays an impact on success rate of landing
- Increasing flight number results in more successful first stage landings
- Note: failed landings are 0 (green), successful landings are 1 (orange)

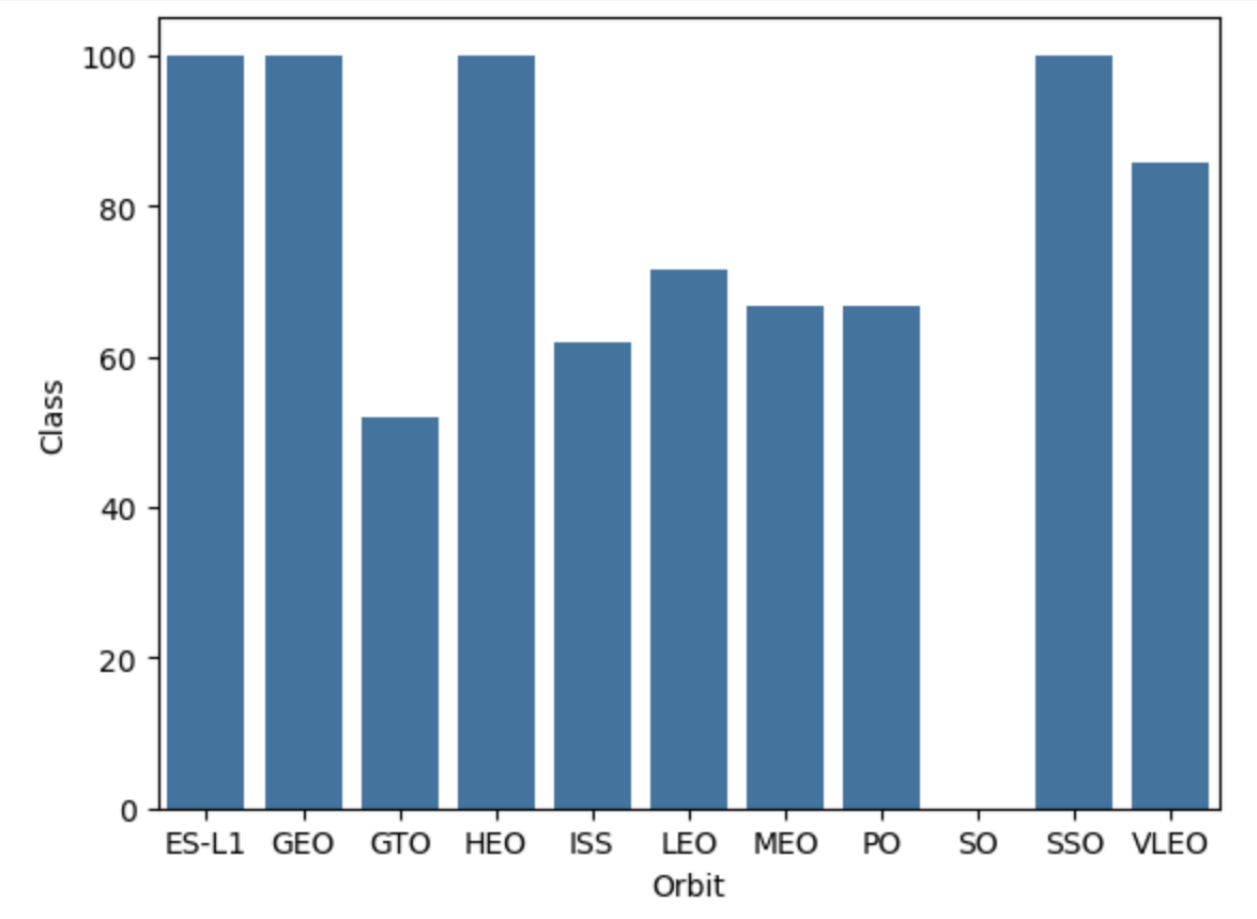
# Payload vs. Launch Site



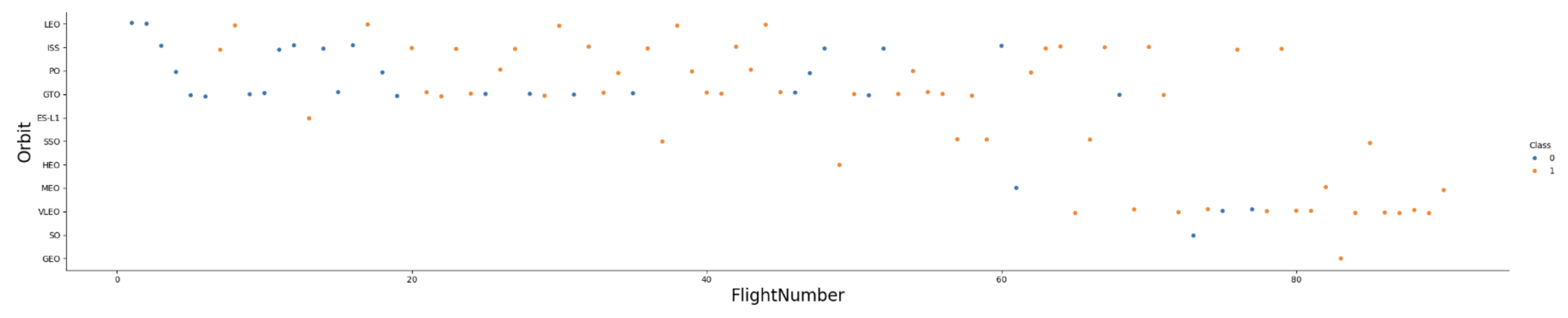

- More successful, but also more failed, landing outcomes for CCAFS SLC 40
- Most failed landings for KSC LC 39A launch site group around small region of payload masses

# Success Rate vs. Orbit Type

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

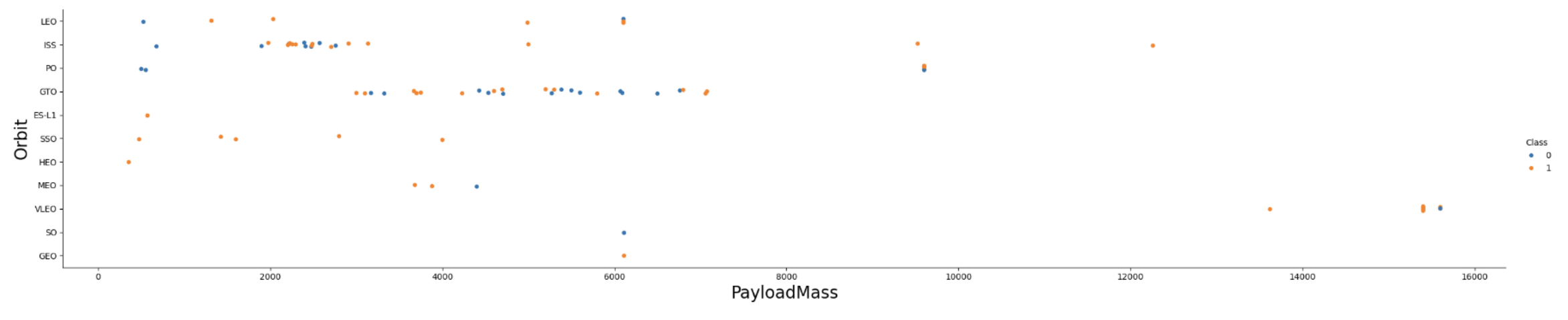


# Flight Number vs. Orbit Type




- The higher the flight number, the more the success rate

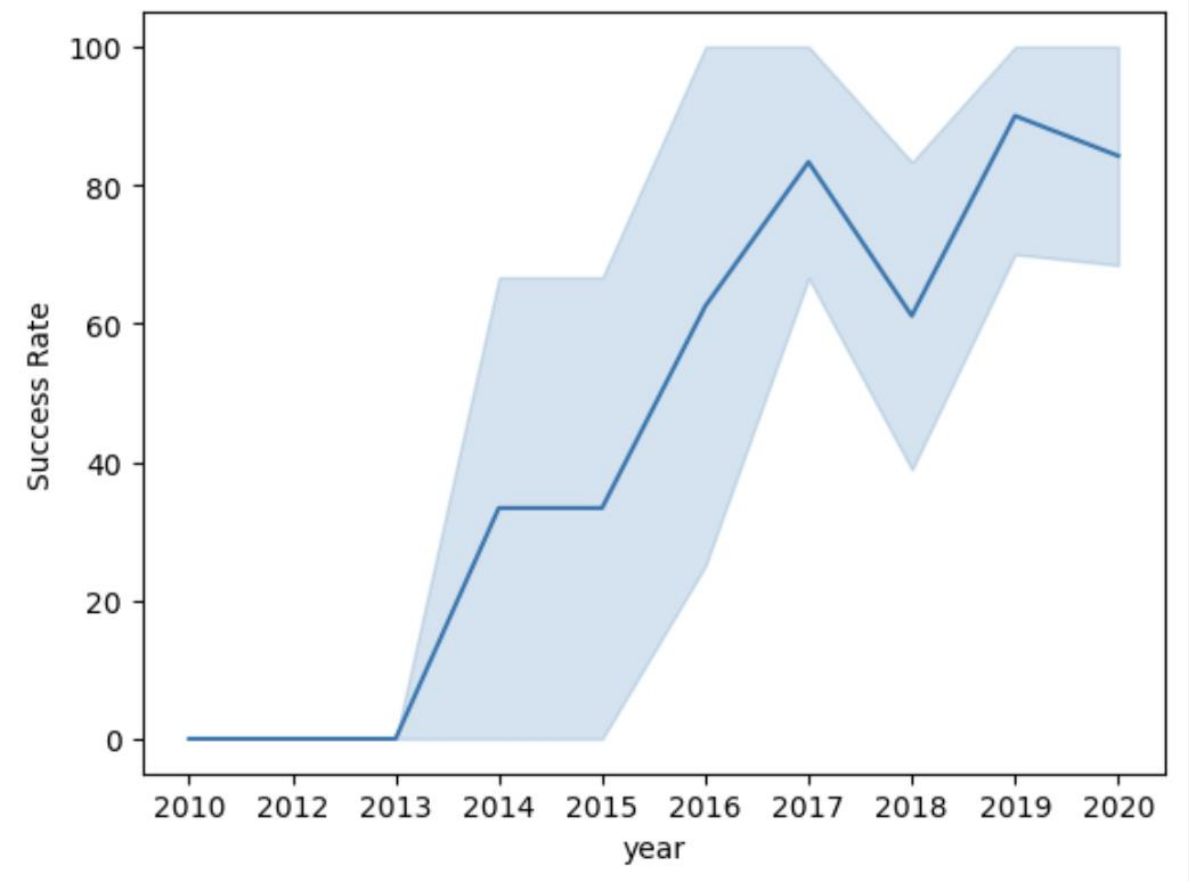
# Payload vs. Orbit Type




- There are more successes in cases of lower payload mass, but there are also many failures
- Difficult to determine the actual correlation between payload mass and orbit

# Launch Success Yearly Trend

- Success rate has increased generally over the years



# All Launch Site Names

## SQL Query

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

## Description

- There are four unique launch sites

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

# Launch Site Names Begin with 'CCA'

## SQL Query

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

## Table

| 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 |

# Total Payload Mass

## SQL Query

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

## Description:

- Total payload mass = 45596 kg

## Payload Mass:

| sum |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

## SQL Query

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

## Average Payload Mass:

| Average |
| --- |
| 2534.6666666666665 |

# First Successful Ground Landing Date

## SQL Query

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

## First Successful Ground Landing Date:

| Date |
| --- |
| 2010-06-04 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

## SQL Query

```
%sql select booster_version from SPACEXTBL where (mission_outcome like 'Success') AND

(payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing_outcome like 'Success (drone ship)')
```

## Names of 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

## SQL Query

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

## Total Success/Failure Number:

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

# Boosters Carried Maximum Payload

## SQL Query

```
maximum = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maximum[0][0]
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)
```

## Boosters with Max Payload:

| 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 |

# 2015 Launch Records

## SQL Query

```
%sql select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL\
where substr(Date, 1, 4) = '2015' and Landing_Outcome = 'Failure (drone ship)'
```

## Failed Landing Outcomes in 2015:

| month | 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 |

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

## SQL Query

```
%sql select Landing_Outcome, count(*) as count_outcomes \
from SPACEXTBL\
where Date between '2010-06-04' and '2017-03-20'\
group by Landing_Outcome\
order by count_outcomes Desc;
```

## Ranked Landing Outcomes:

| Landing_Outcome | count_outcomes |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

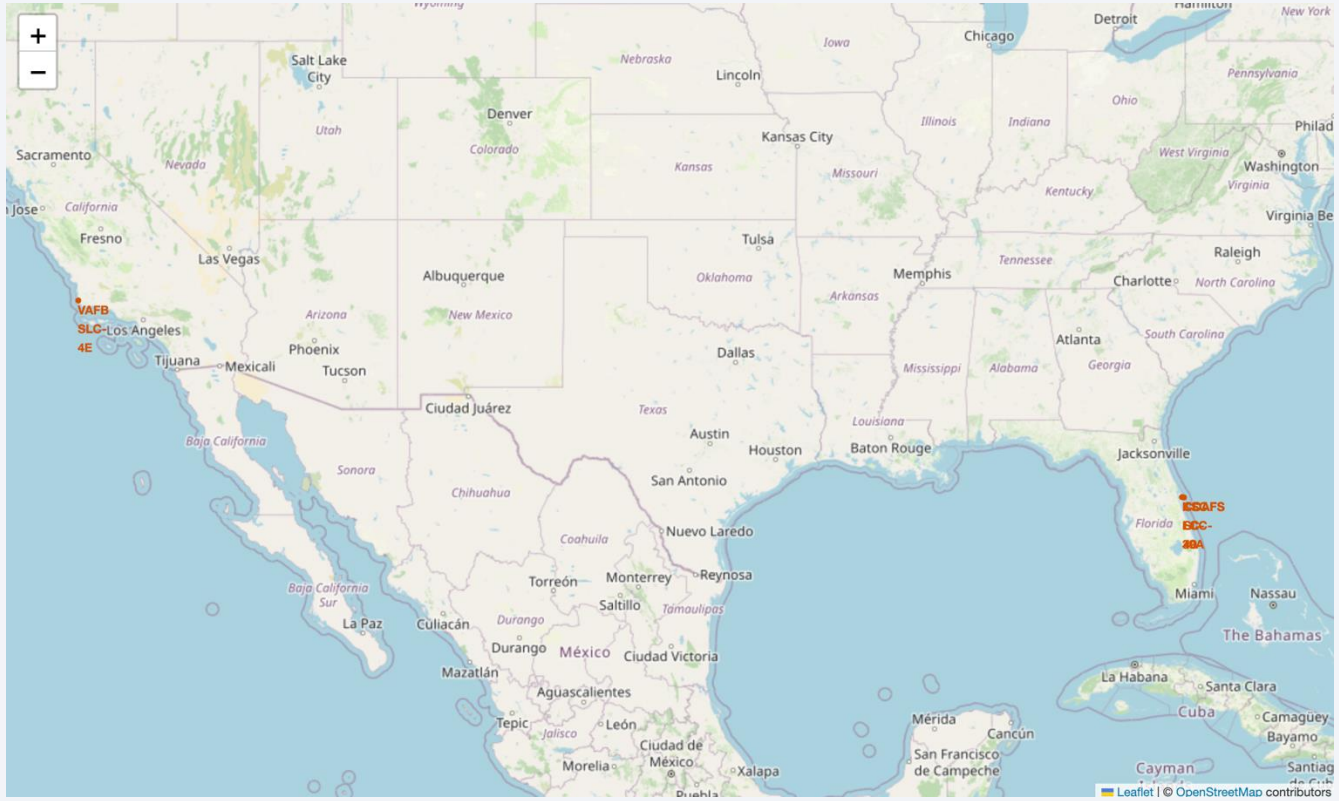
# Falcon 9 Launch Sites

## California

- VAFB SLC-4E ~ Vandenberg Air Force Base Space Launch Complex 4E

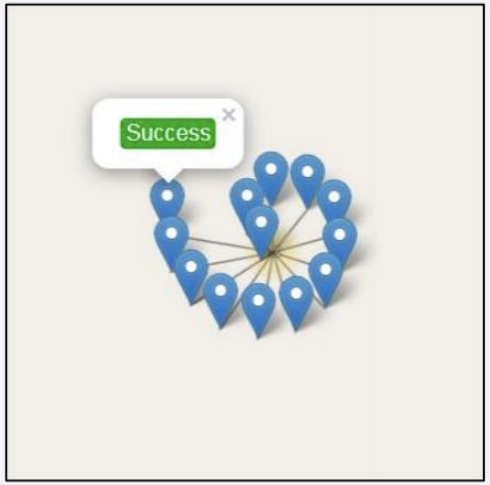
## Florida

- 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



# Success and Failure Map Markers


VAFB SLC-4E




KSC LC-39A

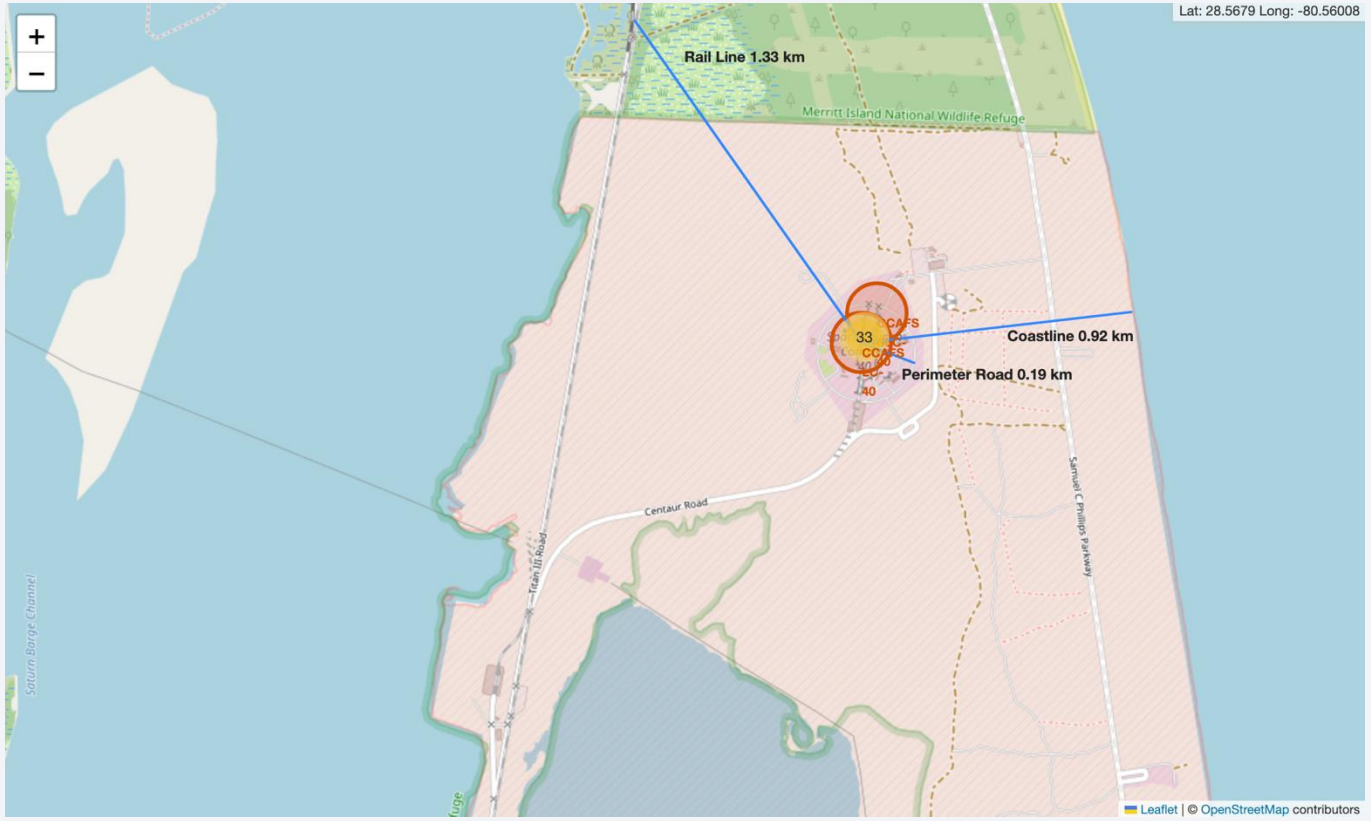

CCAFS LC-40

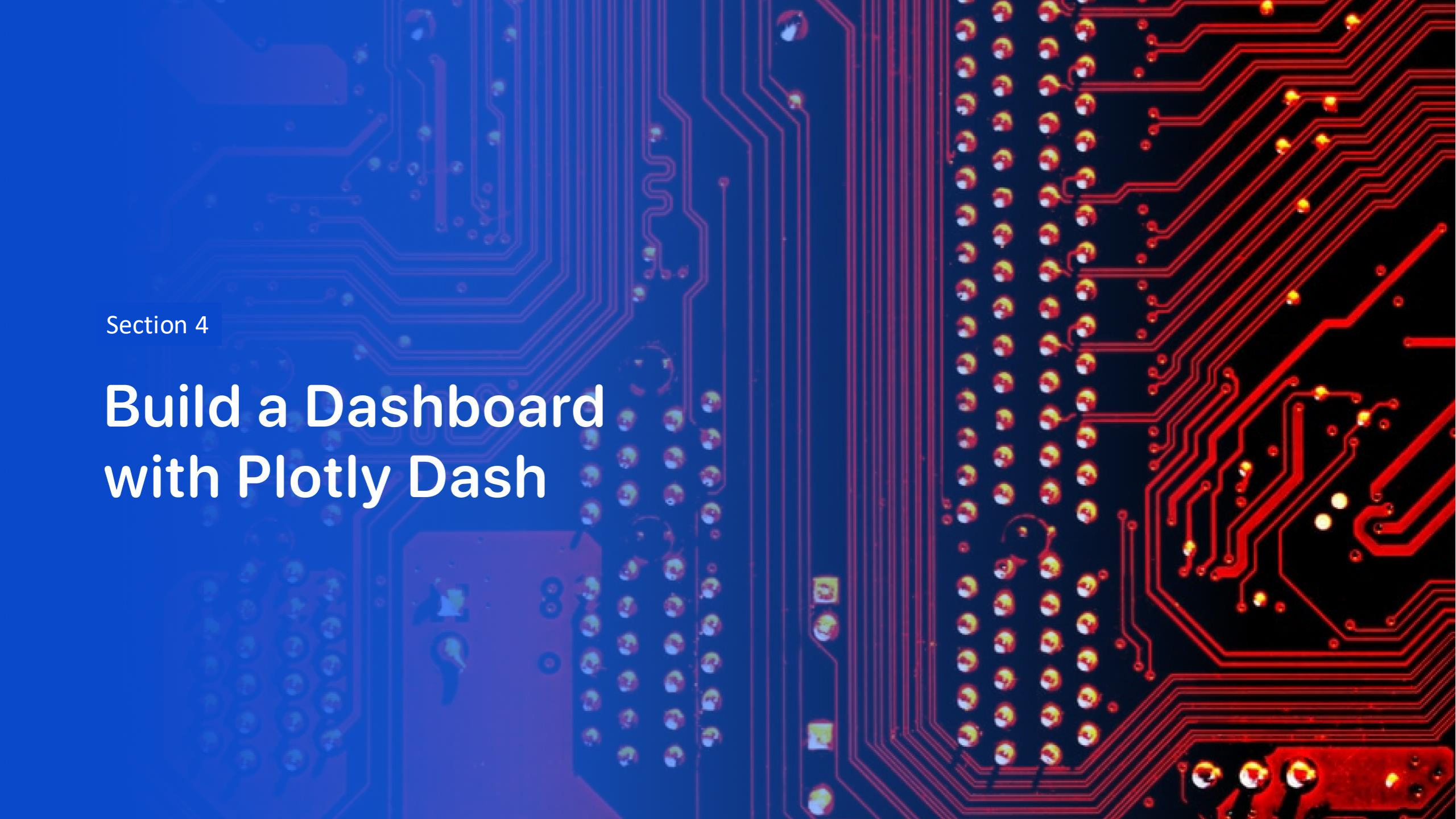

CCAFS SLC-40

- The markers display the outcomes (pass or fail) for Falcon 9 first stage landings
- The four pictures are the four launch sites

# Launch Site to Proximities Distances

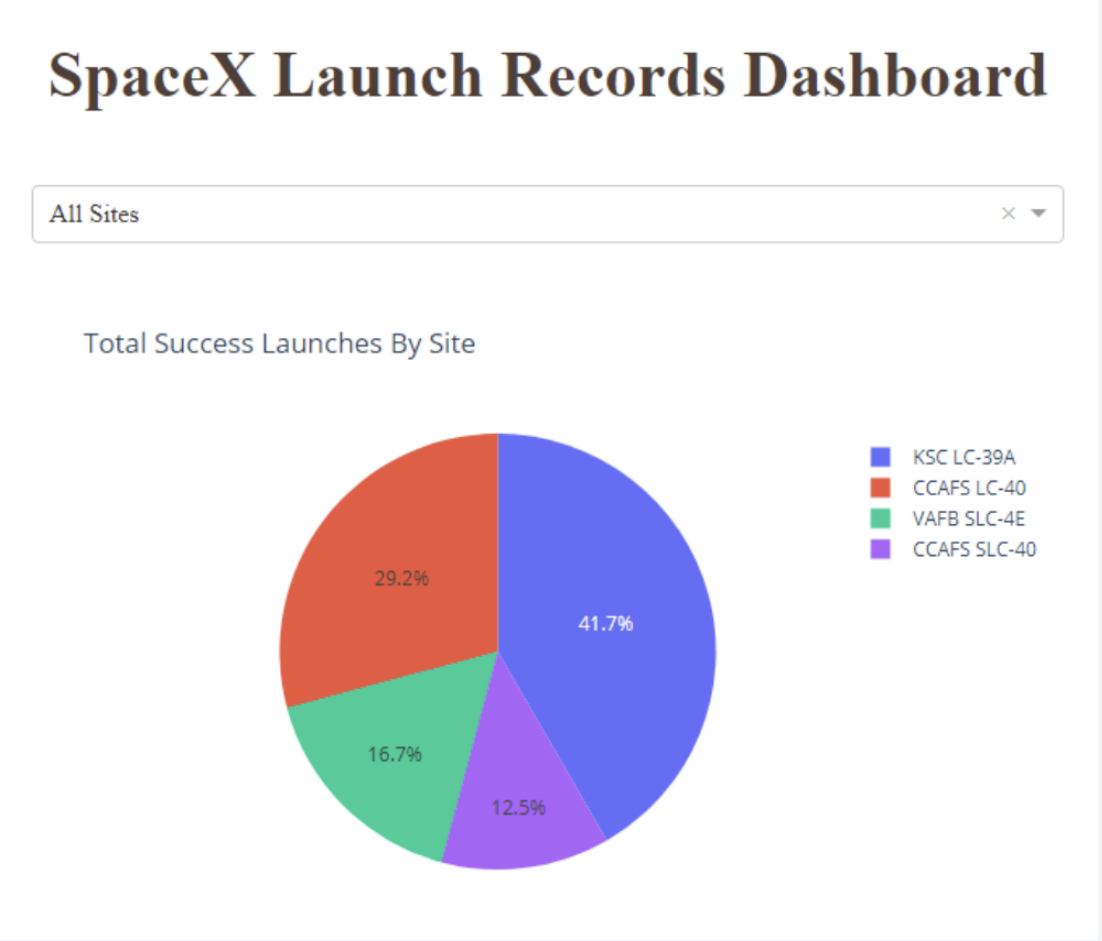
- The coastline is 0.92 km away from CCAFS LC-40
- The rail line is 1.33 km away from CCAFS LC-40
- The perimeter road is 0.19 km away from launch sites




Section 4
Build a Dashboard with Plotly Dash

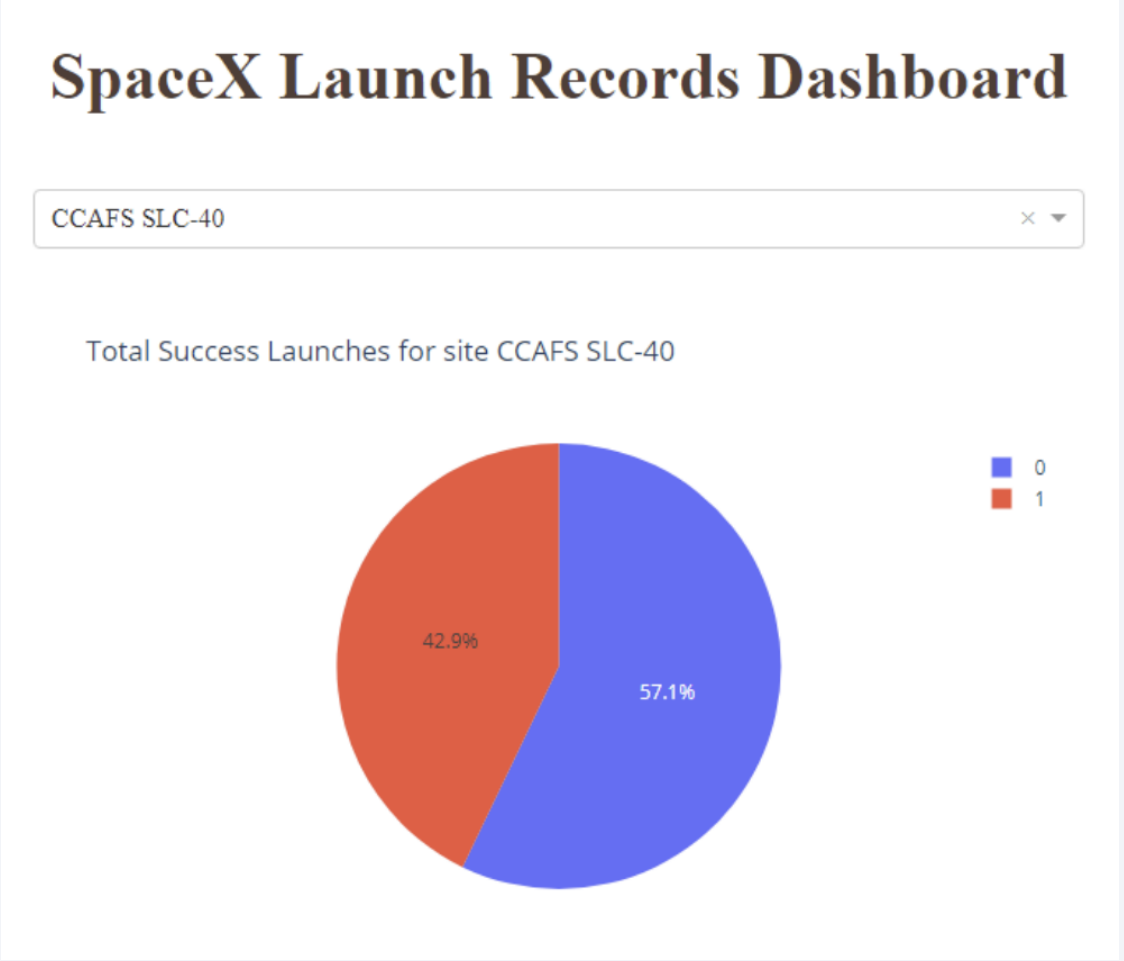
# Launch Success between All Sites

- The greatest share of successful launches occurred from KSC LC-39A
- The lowest share of successful launces occurred from CCAFS SLC-40



# Success/Failure from CCAFS SLC-40

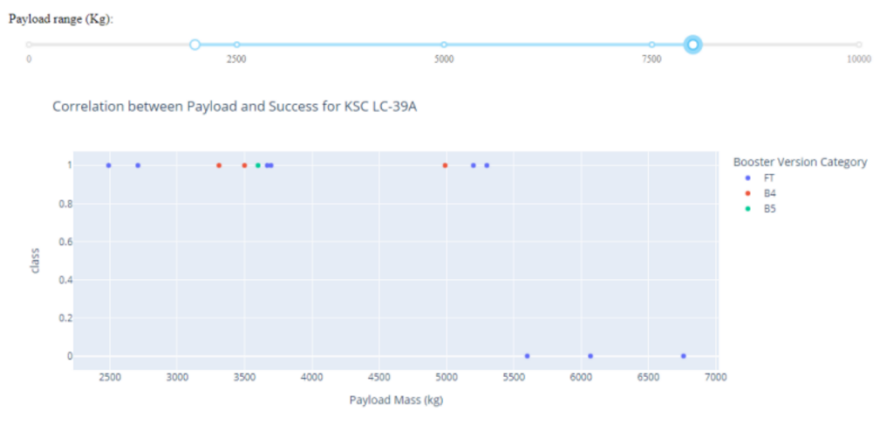
- CCAFS SLC-40 has the highest percentage of success of the four
- 42.9% of launches were successful



# Payload vs. Outcome Scatter Plots

- Overall, the highest success rate came from 2000 to 5000 kg payload



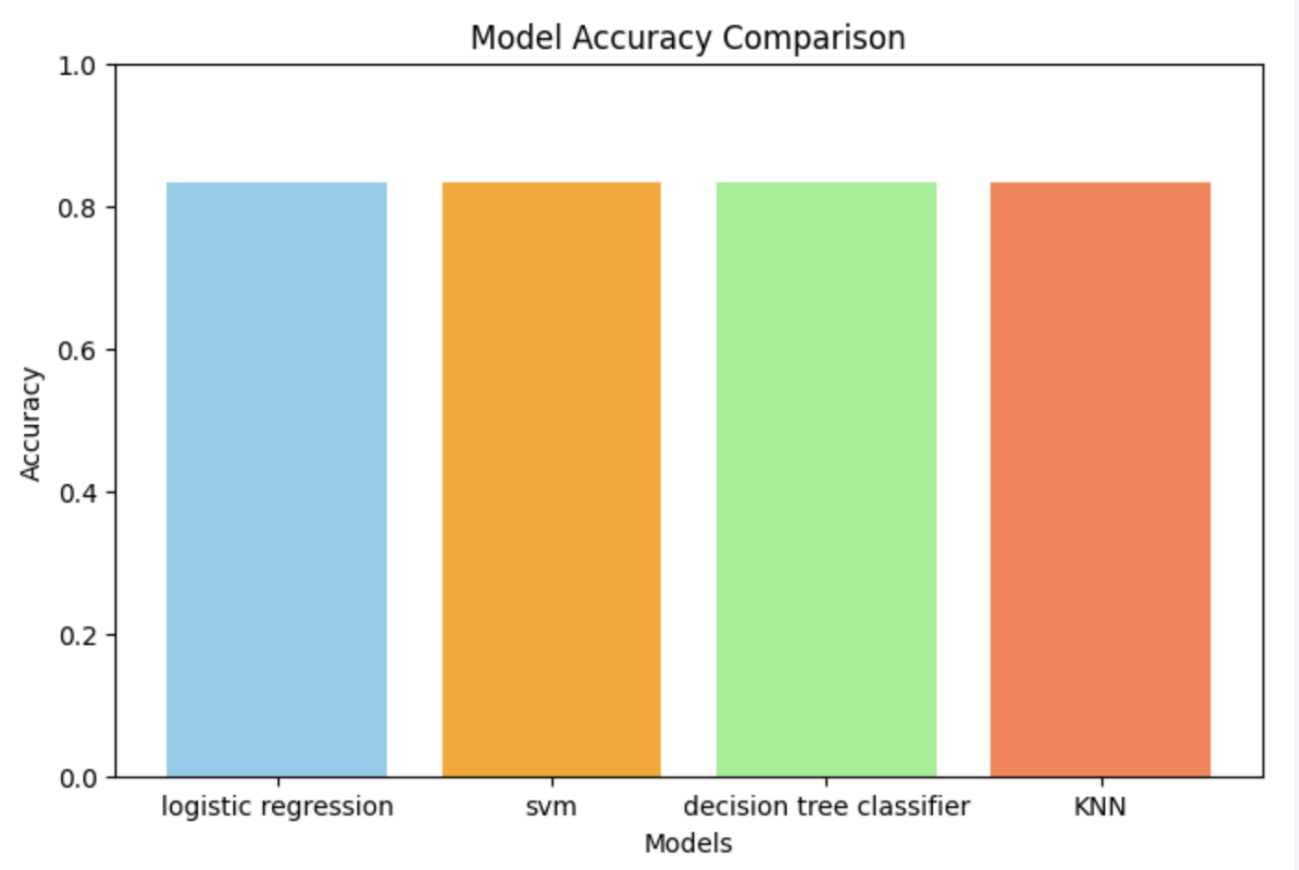



Section 5

# Predictive Analysis (Classification)

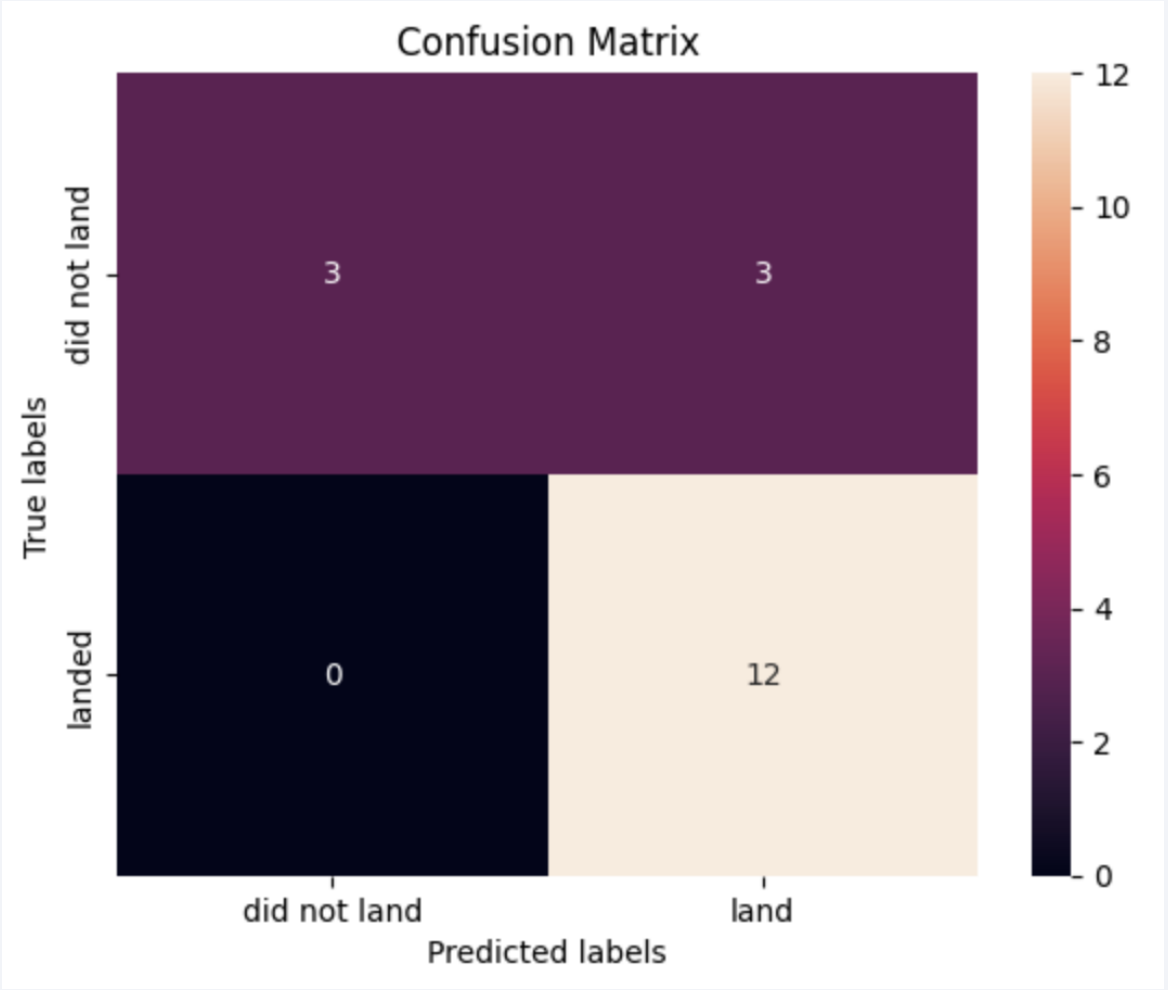
# Classification Accuracy

- All four methods have comparable accuracy (83.33%)



# Confusion Matrix

- The confusion matrix on the right corresponds to knn
- Breakdown is as follows:
  - 12 true positives
  - 3 true negatives
  - 3 false positives
  - 0 false negatives



# Conclusions

- First stage landing outcomes have continued to improve over time
- All classification models used to test have similar accuracy, though there is still room for improvement
  - Random forest, or another more powerful model could improve the accuracy

# Appendix

**GitHub Repository for Capstone Project:**

- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataCollectionAPI.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataCollectionWebScraping.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_DataWrangling.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_EDAwithSQL.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_EDAwithVisualization.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_InteractiveDashboardPlotly.py
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_MachineLearningPrediction.ipynb
- https://github.com/arnavmankad24/DataScienceCapstone/blob/main/Mankad_VisualAnalyticsFolium.ipynb

# Thank you!