IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robert Sizemore
July 4th, 2023

Outline

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

Executive Summary

A rocket launch typically involves two stages, the first of which does the majority of the work in sending a payload into orbit. The cost of launching a rocket into space can be significantly reduced if we can recover and reuse the first stage.

The Falcon 9 rocket series used by SpaceX is notable for reusing the first stage. Using official data relating to the Falcon 9 rocket launches, we can predict whether we can successfully recover the first stage for a given launch with around 85% accuracy.

Introduction

Our goal is to determine which factors contribute to successful launches

- Exploratory Data Analysis
  - Use visualizations to determine relationships between variables
  - What effect does payload mass and orbit type have on launch success?
- Launch Sites
  - Does proximity to cities, railroads, coastlines matter?
- Statistical Modeling
  - Can we accurately predict successful landing using statistical models?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Official API from the SpaceX website
  - Webscraping from Wikipedia
- Perform data wrangling
  - Fix missing values, encode categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

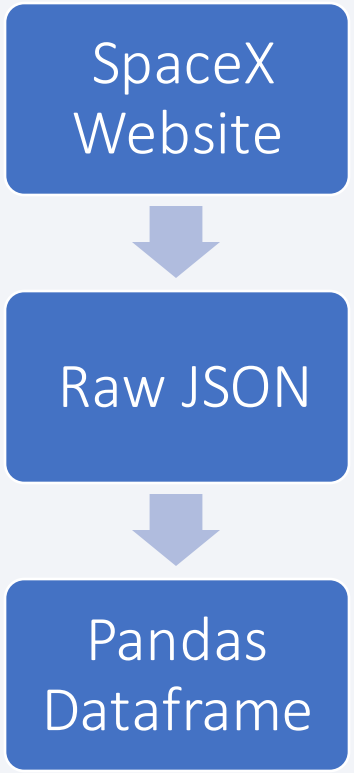
The data we consider in this analysis relate to the Falcon 9 rocket family. We collect the data from two publicly accessible sources:

- The official SpaceX API
- Wikipedia via webscraping

Data Collection – SpaceX API

- The raw data obtained via the official API is in JSON format. We can manipulate this into tabular format using pandas functions.

- Script: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/01_data_collection_spacex_api.ipynb



Data Collection - Scraping

- The raw data obtained via webscraping is HTML code. We can extract the underlying data by looping through the cells of the HTML tables

- Script: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/02_data_collection_web_scraping.ipynb

Data Wrangling

- Missing Values – replace the missing values of **payload mass** with the average payload
- Categorical Variables – encode **orbit**, **launch site**, **landing pad** and **serial** using one-hot encoding via the 'get_dummies' method
- Label Encoding – We create a binary **class** variable indicating if landing was successful
- Script: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/03_data_wrangling.ipynb

EDA with Data Visualization

- Our goal is to predict successful landings, in our first EDA we use visualizations to discover relationships between variables and successful landings. We use color coding to indicate success in the case where we are comparing two independent variables.
- **Payload Mass** vs **Flight Number** – We use a scatterplot to visualize this relationship since flight number is essentially a temporal variable. We see that SpaceX has steadily increased their maximum payload over time, leveling out at about 16000 kg.
- **Payload Mass** vs **Launch Site** – We use a categorical scatterplot to visualize this relationship since there are only 3 launch sites. We see that one launch site appears to be restricted to payloads of 10,000 kg or less.
- Success Rate Per **Orbit Type** – We use a bar chart to visualize this relationship, with the length of the bar indicating the percentage of successful landings.
- EDA URL: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/04_eda_dataviz_seaborn.ipynb

EDA with SQL

- In our second EDA we use SQL queries to extract qualitative information about our data. See below for some example queries.

- Average payload mass carried by the v1.1 Falcon9 boosters
- Find the date of the first successful landing on ground pad
- Which boosters are tasked with carrying the maximum payloads
- Find the counts for each distinct landing outcome
- SQL EDA: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/05_eda_sql.ipynb

Build an Interactive Map with Folium

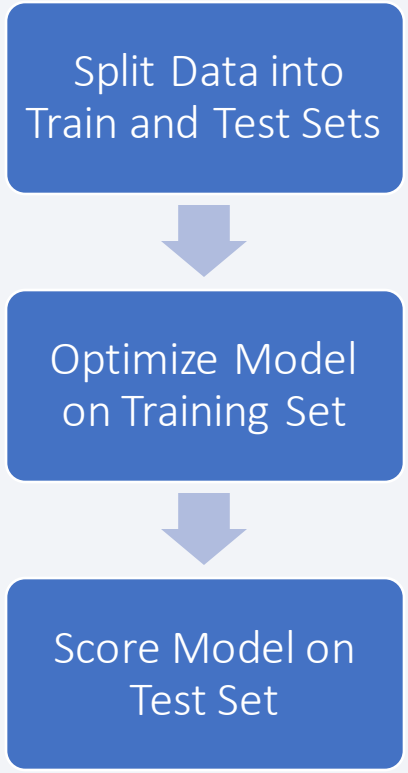
- We use folium to create several interactive maps of the launch sites. Our goal is to explore the launch sites and their proximity to cities, highways, rail access and coasts.
- Each launch site is designated using a colored **circle** object with a popup label given by a **marker** object.
- We use a **cluster** object to create a cluster of **marker** objects around each launch site, each marker represents a separate launch, colored by whether it was successful.
- We use **line** objects to indicate the shortest distance between a launch site and the nearest coast, railway, highway or city.
- Notebook: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/06_interactive_maps_in_folium.ipynb

Build a Dashboard with Plotly Dash

- We explore the relationship between **launch site**, **booster version** and **payload mass** and the recoverability of the first stage, given by the **class** variable.
- We create a dashboard with two interactive elements and two graphs.
- The interactive elements are a dropdown menu with a list of the distinct **launch sites** and a slider from the min and max values of **payload mass**.
- The graphs are a pie chart showing the percentage of successful launches (**class** = 1) in the specified **launch site** and a categorical scatterplot of **class** vs **payload mass** colored coded by the **booster version**, showing only payloads within the range specified by the slider and launches specified by the dropdown menu
- Source Code: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/dashboard/spacex_dash_app.py

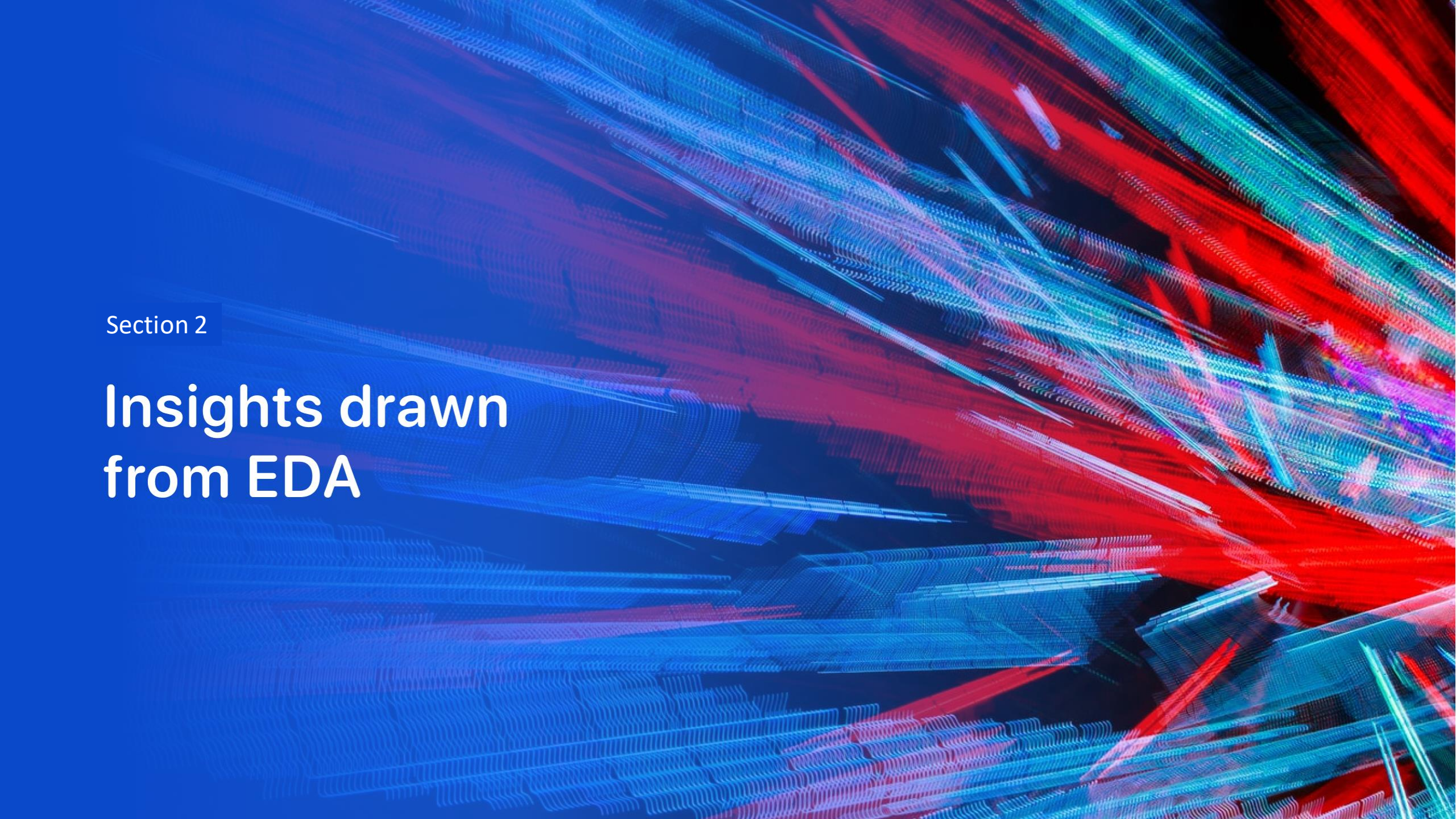
Predictive Analysis (Classification)

- Classification Models
  - Logistic Regression
  - Decision Tree
  - Support Vector Machine
  - K Nearest Neighbors
- Optimization
  - Cross-Validation
  - Grid Search (exhaustive)
- Notebook: https://github.com/rsizem2/IBM-Data-Science-Capstone/blob/main/notebooks/07_predictive_analysis.ipynb



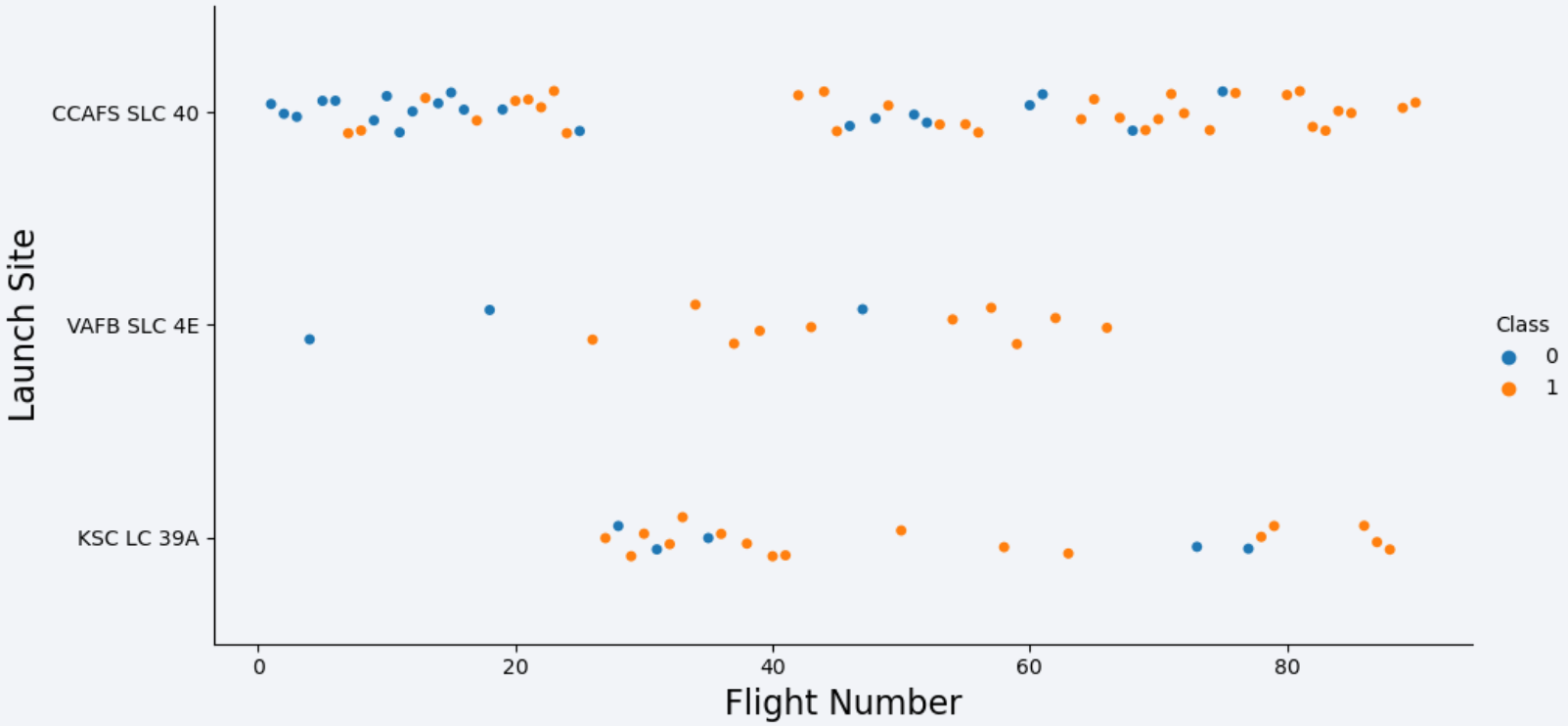
Section 2

Insights drawn from EDA



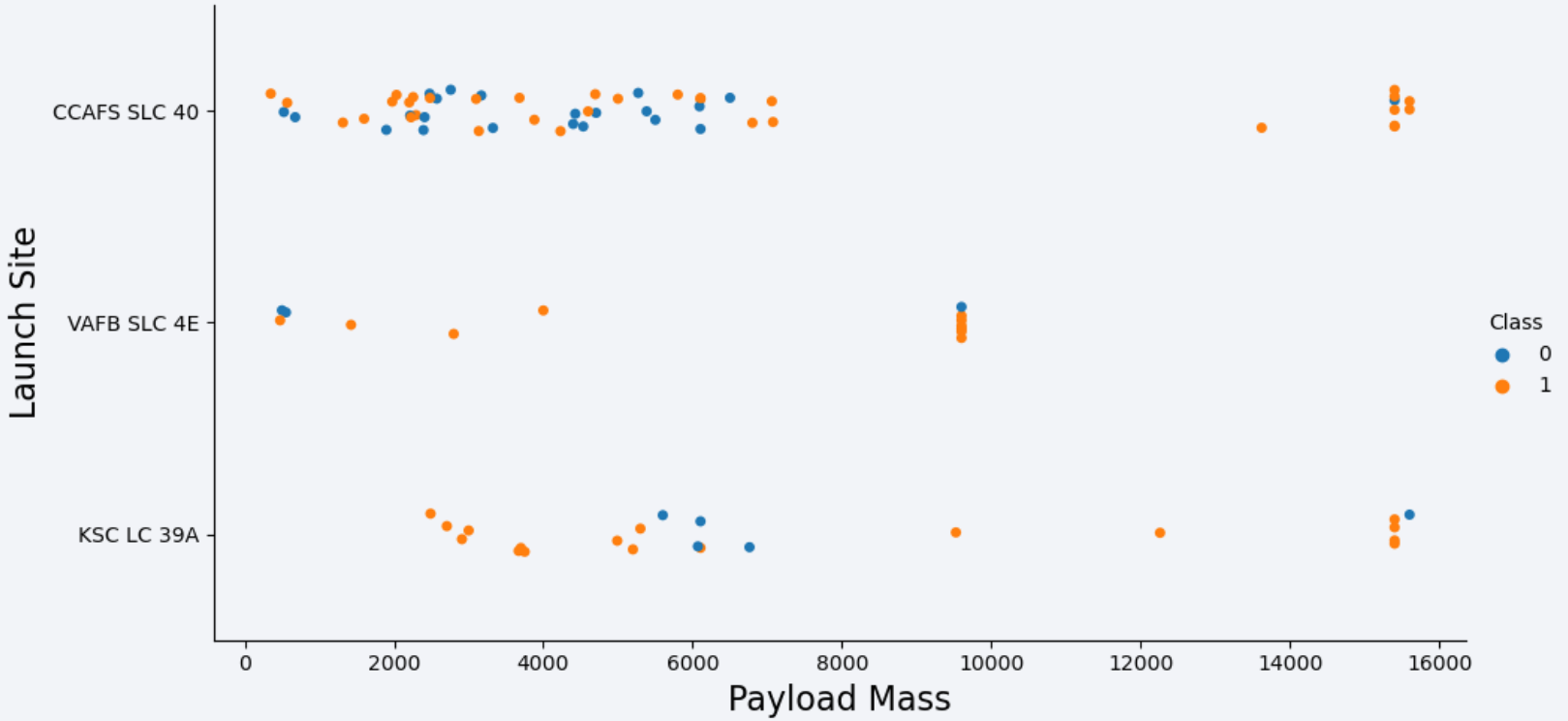
Flight Number vs. Launch Site

- CCAFS SLC 40 is the preferred launch site except for about 20 consecutive flights
- Most "failed" launches at CCAFS SLC 40
- Fewer failed launches as flight number increases



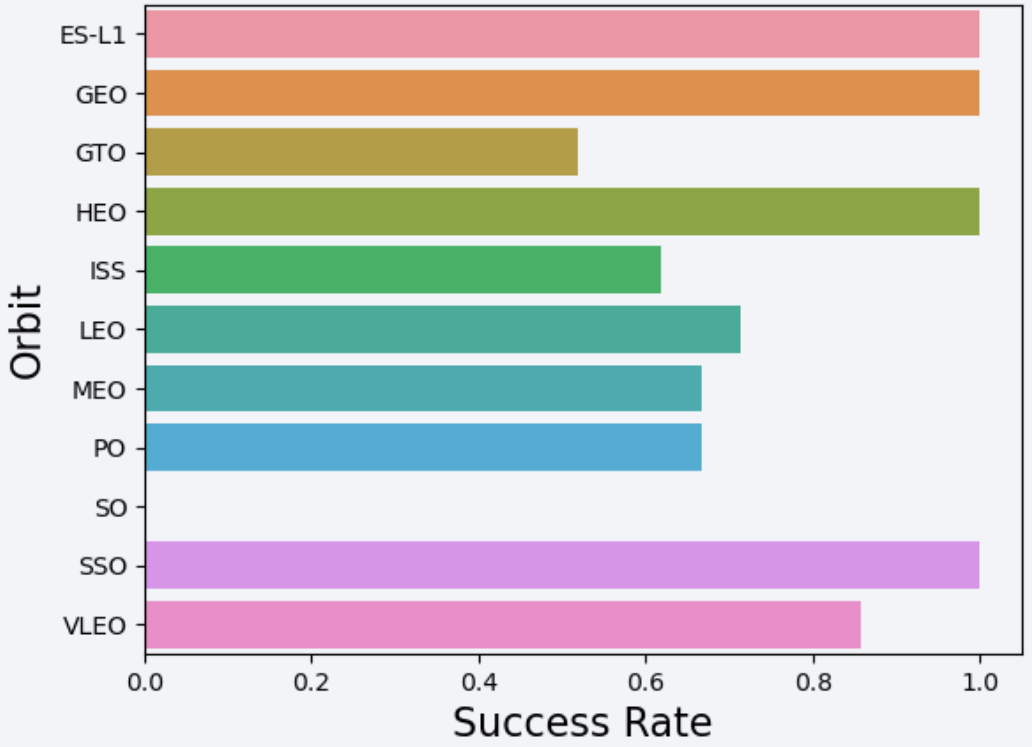
Payload vs. Launch Site

- Majority of flights have payloads under 8000kg
- VAFB SLC 4E flights all have payloads under 1000kg
- KSC LC 39A flights all have payloads over 2000kg



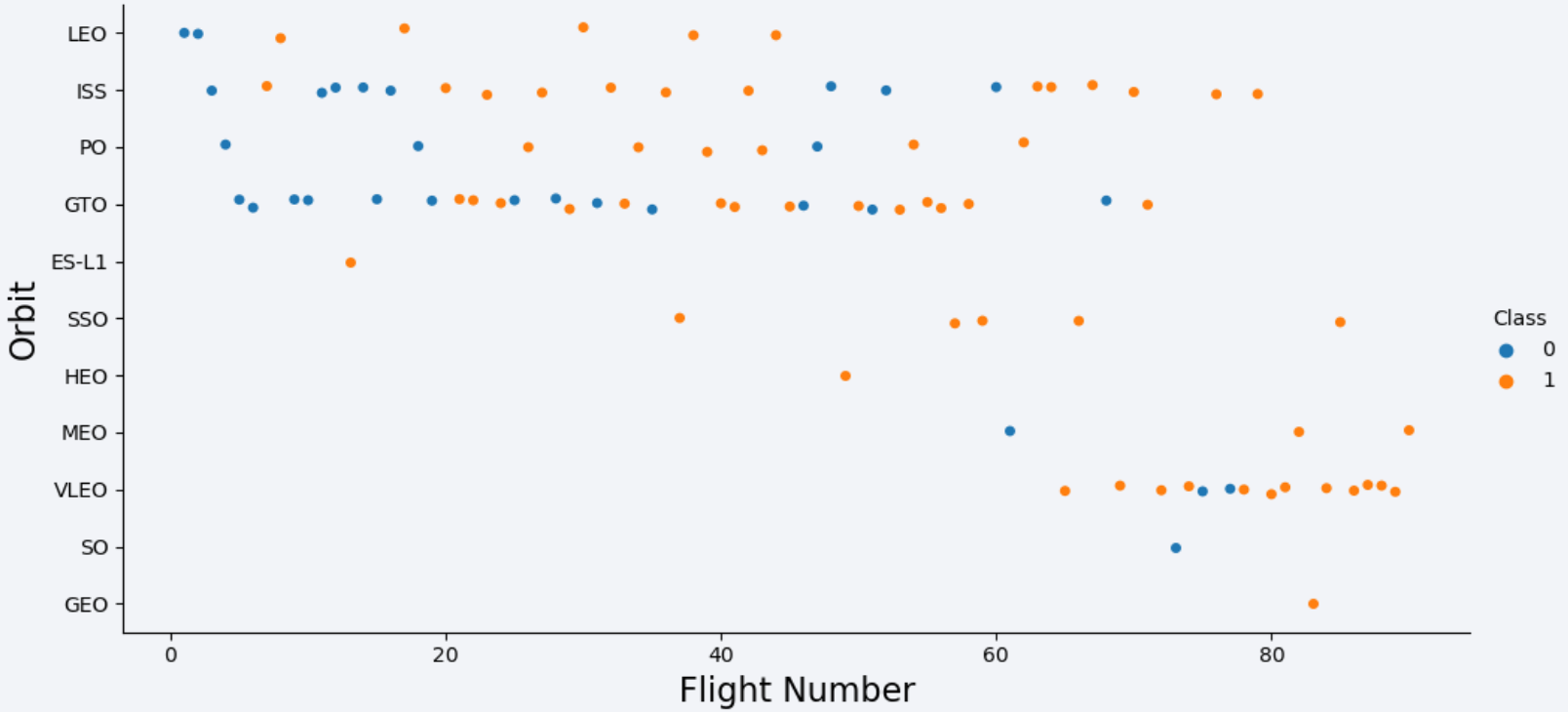
Success Rate vs. Orbit Type

- 7 of the 11 orbit types account for all of the unsuccessful launches
- Only two orbit types have less than 60% success rate



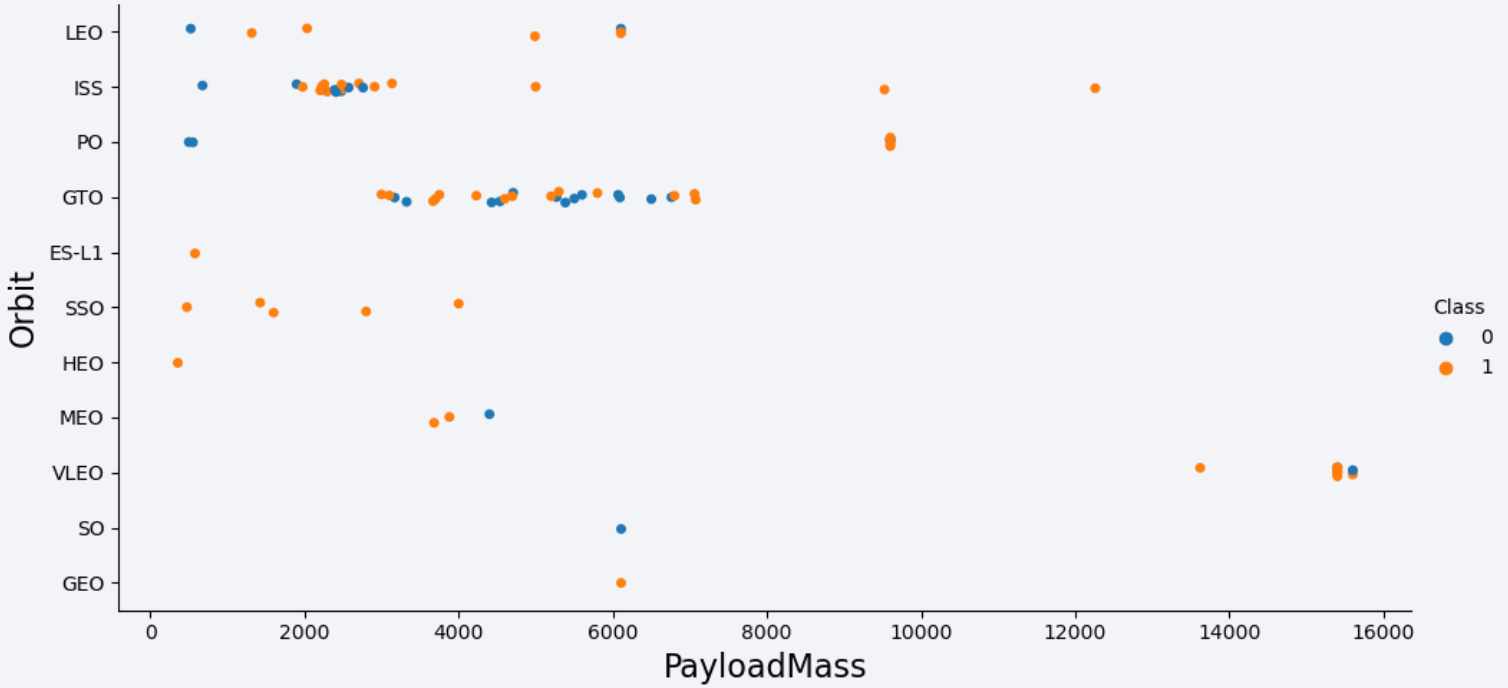
Flight Number vs. Orbit Type

- Orbit types with 0% or 100% success had fewer than 5 flights
- GTO orbit type has relatively high failure rate



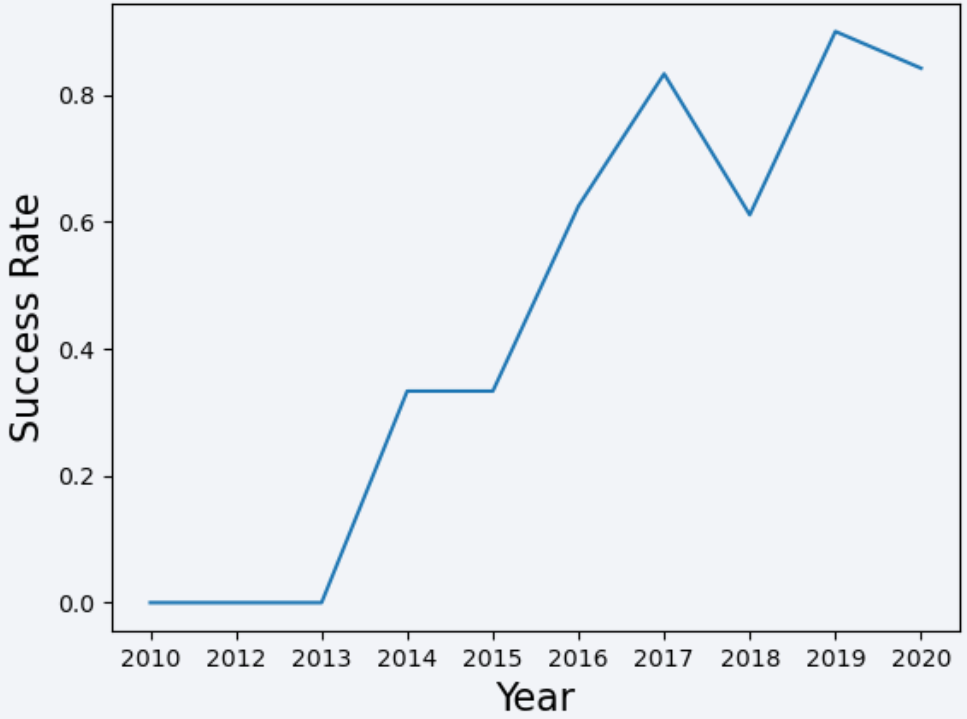
Payload vs. Orbit Type

- Generally, payloads are similar for flights within the same orbit type
- Most GTO flights have payloads between 3k and 7k
- Most ISS flights have payloads between 2k and 3k



Launch Success Yearly Trend

- Success rate has steadily increased since 2013
- 20% dip from 2017 to 2018
- Greater than 50% success from 2016 onwards



All Launch Site Names

- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40
- None (missing value)

- Data stored as table named SPACEEXTBL
- Launch Sites stored in column "Launch_Site"
- SQL Query: "SELECT DISTINCT Launch_Site FROM SPACEXTBL;"
- DISTINCT returns all unique values

Launch Site Names Begin with 'CCA'

- SQL query: SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- The character * selects all columns
- Like 'CCA%' - selects all strings starting with CCA
- LIMIT 5 – returns only the first 5 results

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload
0	06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
1	12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...
2	22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
3	10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
4	03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Total Payload Mass

- Total payload carried by boosters from NASA: **107,010 kg**
- SQL Query: SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE '%NASA%';
- SUM() - returns the total sum of the specified column
- LIKE '%NASA%' selects all strings containing 'NASA' anywhere.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: **2534 kg**
- SQL query: SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
- AVG() - returns the average of the specified column
- LIKE 'F9 v1.1%' - selects all booster versions starting with F9 v1.1

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad: **Dec 22, 2015**
- SQL Query: 'SELECT Date FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)" LIMIT 1;'
- The data is ordered by flight number by default, which is chronological.
- We output the first flight for which the **landing outcome** is "Success (ground pad)"

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query: SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

- We use keyword BETWEEN to indicate the range of values which want to select

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

Total Number of Successful and Failure Mission Outcomes

- Total Successful Outcomes: 100
- Total Failure Outcomes: 1
- SQL Query: SELECT COUNT(Date) FROM SPACEXTBL WHERE Mission_Outcome LIKE '%Success%'
- SQL Query: SELECT COUNT(Date) FROM SPACEXTBL WHERE Mission_Outcome NOT LIKE '%Success%'
- We ignore missing values

Boosters Carried Maximum Payload

- SQL Query: SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
- We a subquery to find the maximum payload
- Total of 12 distinct boosters which carried the maximum payload

- 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: SELECT substr(Date, 4, 2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
  WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date,7,4)='2015';

- The substr() function is used to extract the month from the Date string.

Month	Outcome	Booster	Site
10	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: SELECT Landing_Outcome AS Outcome, COUNT(Landing_Outcome) AS Count FROM SPACEXTBL WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY Landing_Outcome ORDER BY Count DESC;

- We use GROUP BY and COUNT() to count the total outcomes, which we call "Count", we use ORDER BY to sort by the new "Count" column with DESC for descending order.

Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	7

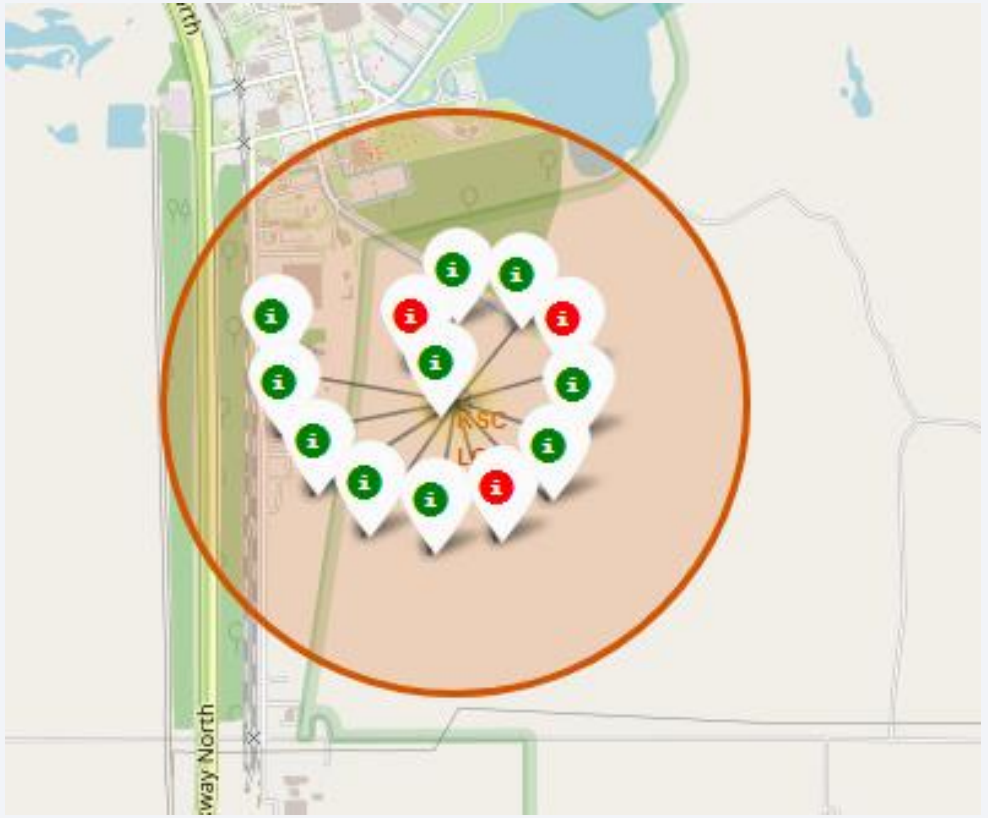

Section 3
Launch Sites
Proximities Analysis

Folium Map 1: Launch Sites

- We use **circle** objects to mark each launch site and **marker** objects to tag each site with it's name
- We notice each site is located close to the coast
- Three of the four launch sites are located in very close proximity on the coast of Florida

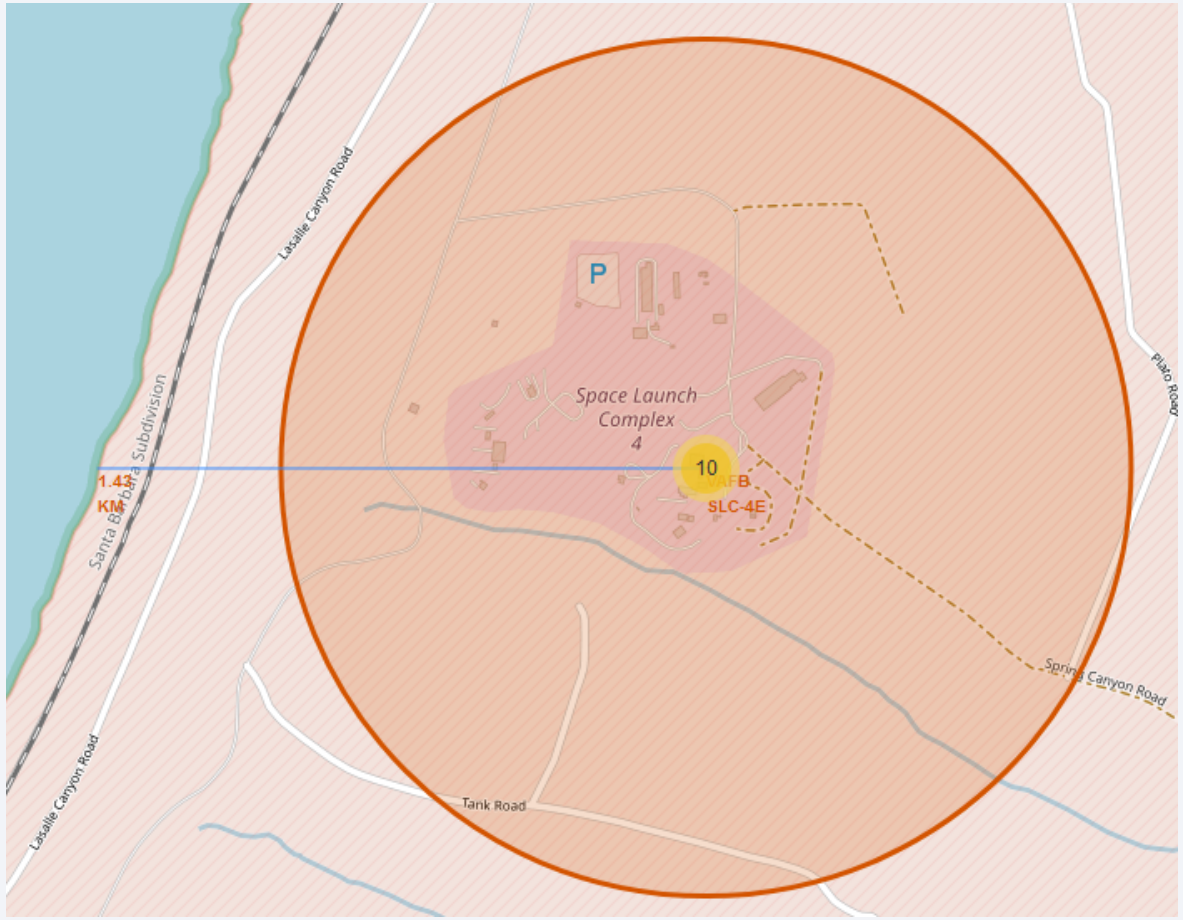
Folium Map 2: Launch Outcomes

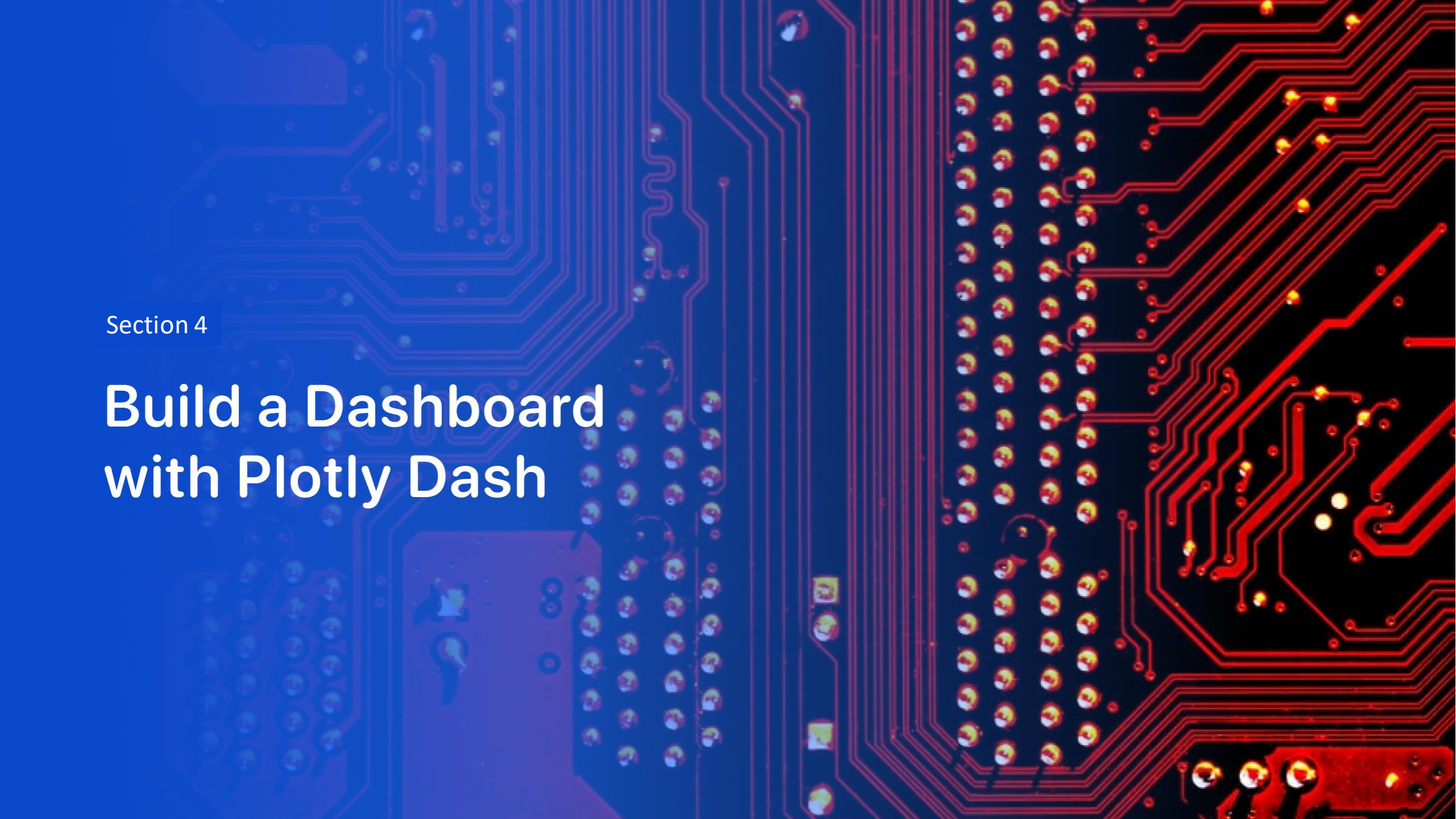
- We use a **MarkerCluster** object to create a cluster of markers around each launch site
- Clicking on a launch site shows a cluster of icons indicating whether a launch was successful or not
- Red icons indicate failed launches and green icons indicate successful launches



Folium Map 3: Launch Site Proximity

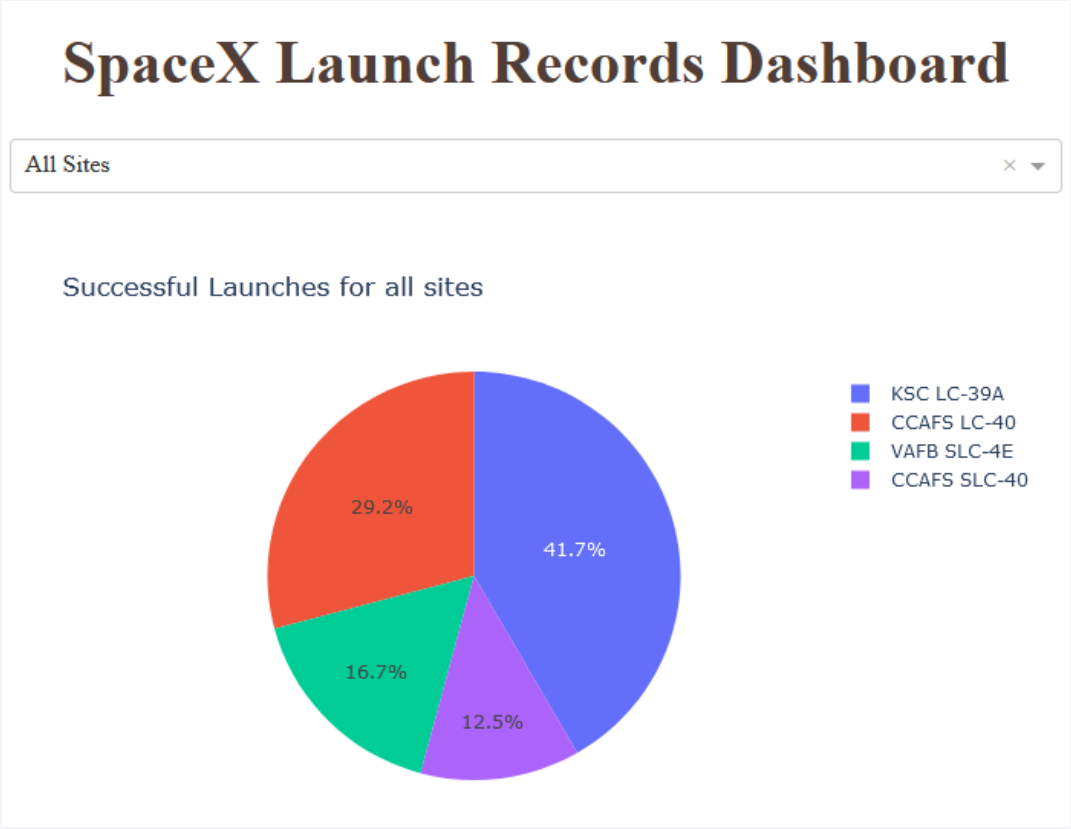
- All 4 launch sites are in close proximity to the coast as well as to railway access
- We use a **Polyline** object to draw a line from the launch site to the nearest point on the coast
- We use a **marker** object to write the distance at the end of the line




Section 4
Build a Dashboard with Plotly Dash

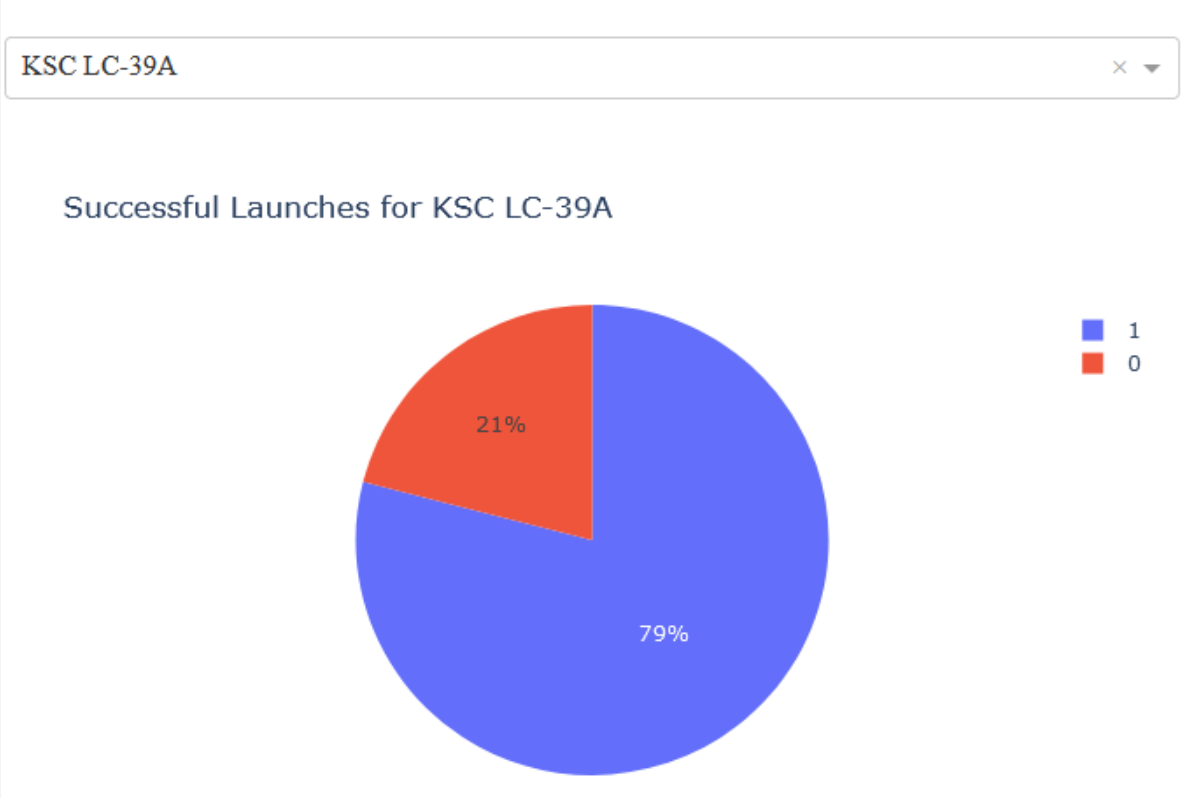
Dashboard 1: All Sites

- The first interactive element of our dashboard is a dropdown menu containing a list of the launch sites as well as an "all sites" option
- By default the item selected is "all sites"
- If "all sites" is selected than the dashboard displays a pie chart of successful launches over all sites



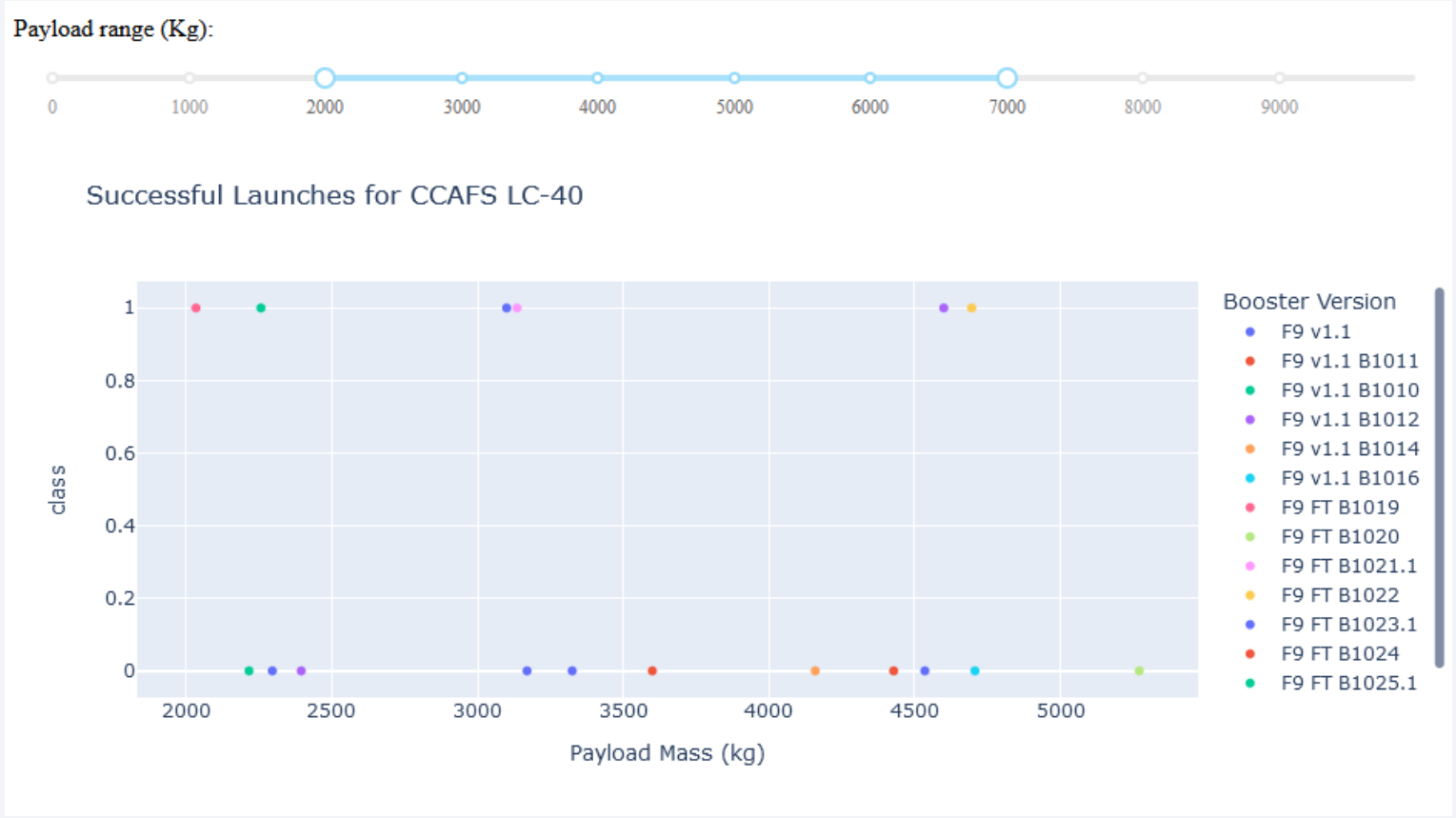
Dashboard 2: Launch Site KSC LC-39A

- If we select a specific site it shows a pie chart of launches (successful/failed) from that launch site
- The site with the highest number of successful launches is KSC LC-39A



Dashboard 3: Payload Slider

- The second interactive element is a slider for the payload range
- The dashboard plots the outcome vs payloads for the range selected by the slider, only including launches at the launch site selected by the dropdown menu

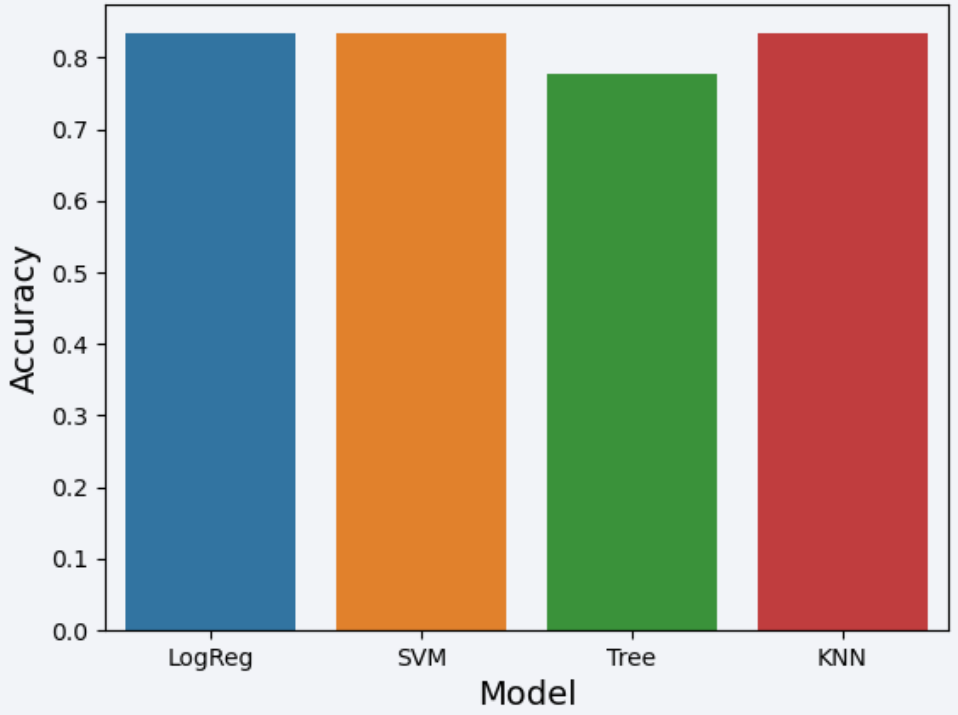


Section 5

Predictive Analysis (Classification)

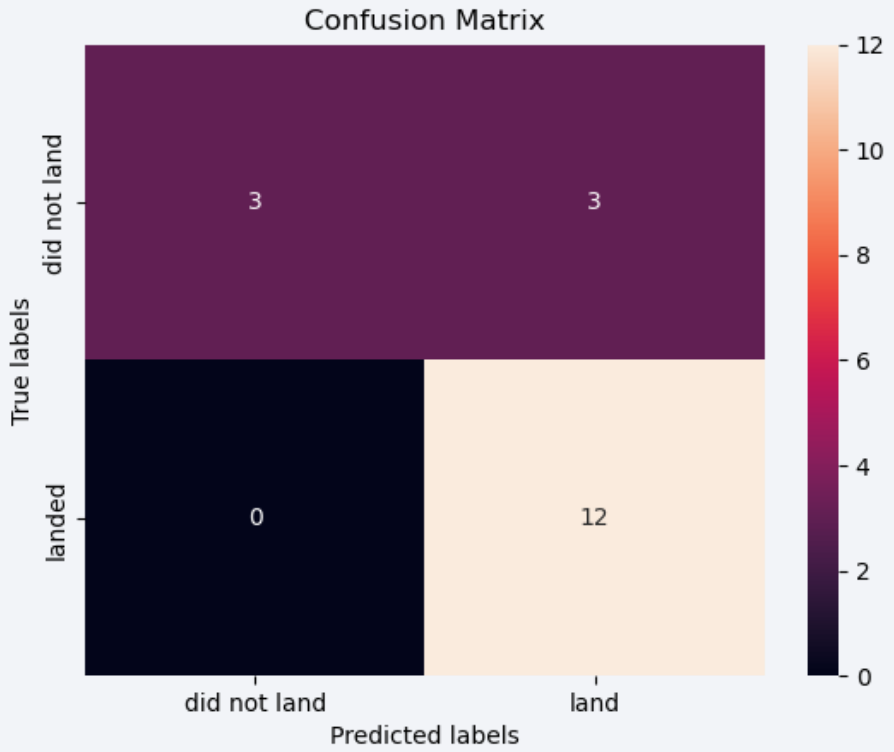
Classification Accuracy

- Three of the four models have the same accuracy on the test set, roughly 85%.
- All things being equal, we should pick Logistic Regression over SVM and KNN since it is the simplest model, fastest to train and easiest to interpret



Confusion Matrix

- We look at the confusion matrix for our logistic regression model
- The confusion matrix allows us to compare the predicted labels with the actual labels
- There are no false negatives (successful flights that were predicted to fail)
- The are three false positives (failed flights that were predicted to succeed)



Conclusions

- Using current data, flight success can be predicted with about 85% accuracy
- Launch sites tend to be in close proximity to the coast and have easy access to railways.
- Launch success is related to certain payload mass ranges and orbit types
- Our models were tested on a small set of data (18 examples), more data is likely needed for further analysis.

Thank you!