IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noman Shaikh

28-06-25

Outline

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

Executive Summary

- This capstone project focuses on predicting the successful landing of the SpaceX Falcon 9 first-stage booster by applying a range of machine learning classification techniques.
- The workflow for the project included:
- **Gathering and preparing the data**
- **Performing exploratory analysis to uncover patterns**
- **Creating interactive visualizations for deeper insights**
- **Building and evaluating predictive models**
- Through our analysis, we identified several features—such as launch site, booster version, and payload mass—that influence the likelihood of a successful landing. Among the models tested, the **Decision Tree algorithm** emerged as a strong candidate, offering the most reliable predictions for Falcon 9 first-stage landing outcomes.

Introduction

- In this capstone project, our objective is to predict whether the Falcon 9 first-stage booster will land successfully. SpaceX offers Falcon 9 launches at a competitive price of $62 million, compared to over $165 million charged by other providers. A key reason for this cost advantage lies in SpaceX's ability to recover and reuse the rocket's first stage. Therefore, being able to anticipate the landing success can provide valuable insight into potential cost savings per launch.
- This predictive capability can be particularly useful for other companies aiming to compete with SpaceX, allowing them to estimate launch costs more accurately and tailor their bids accordingly.
- It's important to note that not all failed landings are accidents—some are intentionally controlled landings in the ocean, carried out as part of mission strategy.
- The core question we aim to answer is:
  Given various launch features—such as payload mass, orbit type, launch site, and booster version—can we accurately predict whether the first stage of a Falcon 9 rocket will land successfully?

Section 1

Methodology

Methodology

The project follows a comprehensive workflow consisting of the following phases:

Data Acquisition and Preparation

- Collected data through the **SpaceX API** and **web scraping**
- Performed data cleaning, wrangling, and formatting to prepare it for analysis

Exploratory Data Analysis (EDA)

- Utilized **Pandas** and **NumPy** for statistical exploration and feature understanding
- Executed **SQL queries** to extract and analyze structured information

Data Visualization

- Created static and comparative plots using **Matplotlib** and **Seaborn**
- Built an **interactive map** with **Folium**
- Developed dynamic dashboards with **Dash** for deeper insight

Predictive Modeling

- Applied various machine learning classification algorithms:
  - **Logistic Regression**
  - **Support Vector Machine (SVM)**
  - **Decision Tree**
  - **K-Nearest Neighbors (KNN)**

Data Collection Wrangling & Formatting

- SpaceX API
  - The API used is https://api.spacexdata.com/v4/rockets/.
  - The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
  - Every missing value in the data is replaced the mean the column that the missing value belongs to.
  - We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Web Scraping

- To supplement our dataset, we performed web scraping on the following Wikipedia page: **List of Falcon 9 and Falcon Heavy launches**
- This page provides detailed information specifically about **Falcon 9 launches**. After extracting and cleaning the data, we obtained a dataset consisting of **121 records** (launch instances) and **11 features**. These features include details such as launch date, mission outcome, booster version, payload, and more.
- Below is a preview of the first few rows of the cleaned dataset:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Collection - Scraping

- The scraped data was further processed to ensure consistency and usability for machine learning. This included handling missing values and applying **one-hot encoding** to transform categorical features into numerical format.
- An additional column named **‘Class’** was introduced to represent the landing outcome of each launch:
- **1** indicates a **successful** landing
- **0** indicates a **failed** landing
- After cleaning and encoding, the final dataset consisted of **90 rows (instances)** and **83 columns (features)**, ready for exploratory analysis and predictive modeling.

Data Collection Wrangling & Formatting

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

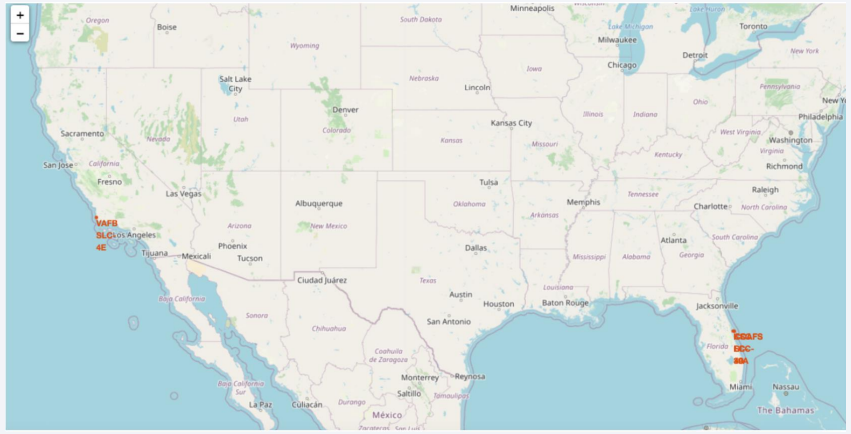
- **Data Exploration with Pandas and NumPy**
- We utilized functions from the **Pandas** and **NumPy** libraries to perform basic exploratory data analysis. Key insights derived include:
- The **number of launches per launch site**
- The **frequency of each orbit type**
- The **count and distribution of mission outcomes**
- These operations helped us understand the structure and composition of the dataset before applying more complex analysis.

EDA with SQL

- **Data Querying with SQL**
- To answer specific analytical questions, we used **SQL queries** on the dataset. Some of the queries included:
- Retrieving the **unique names of launch sites** used in the missions
- Calculating the **total payload mass** carried by boosters launched under **NASA's CRS program**
- Determining the **average payload mass** delivered by **booster version F9 v1.1**
- These queries provided deeper insights into the dataset and helped support feature selection for our machine learning models.

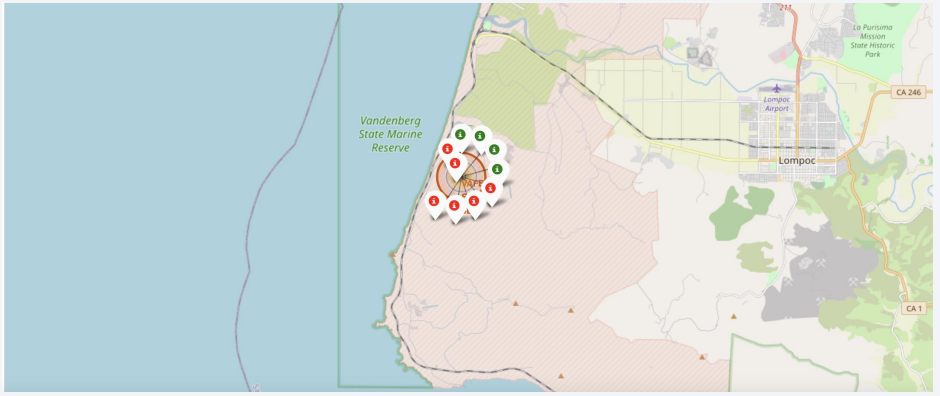
Build an Interactive Map with Folium

- All launch sites on map



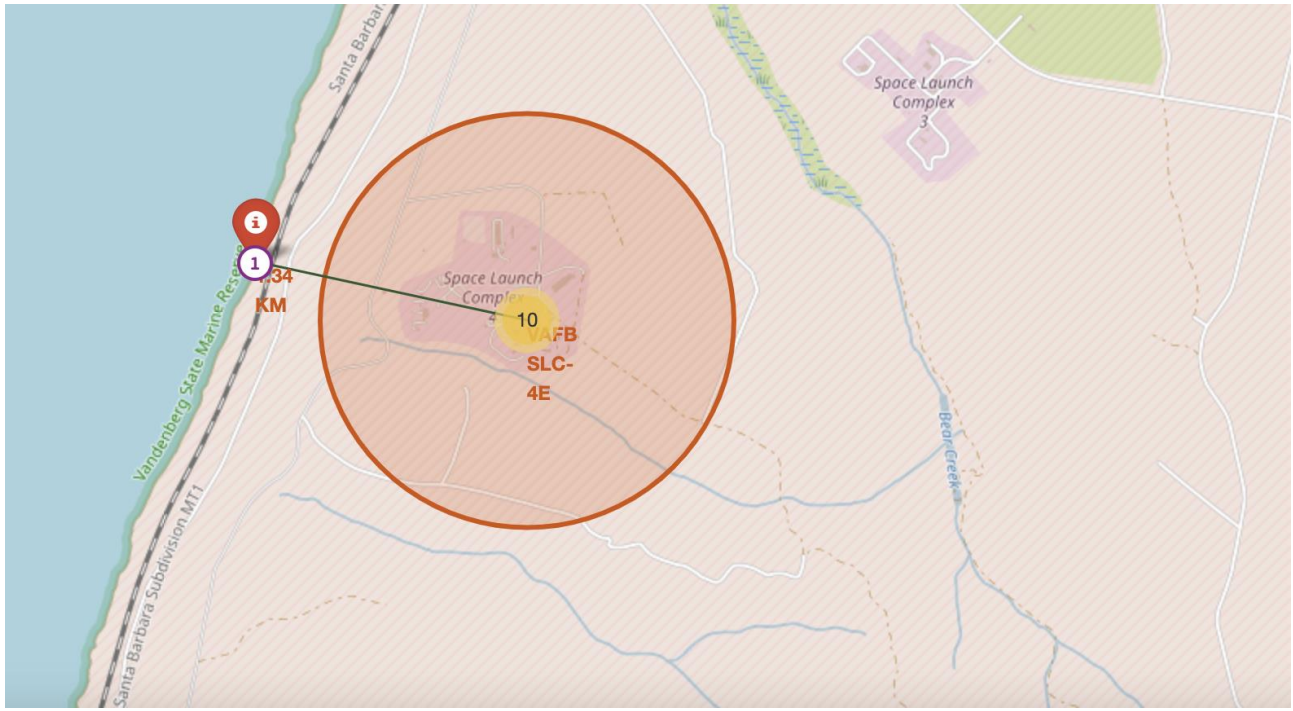
The succeeded launches and failed launches for each site on map

If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



Results Folium

- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site

Predictive Analysis (Classification)

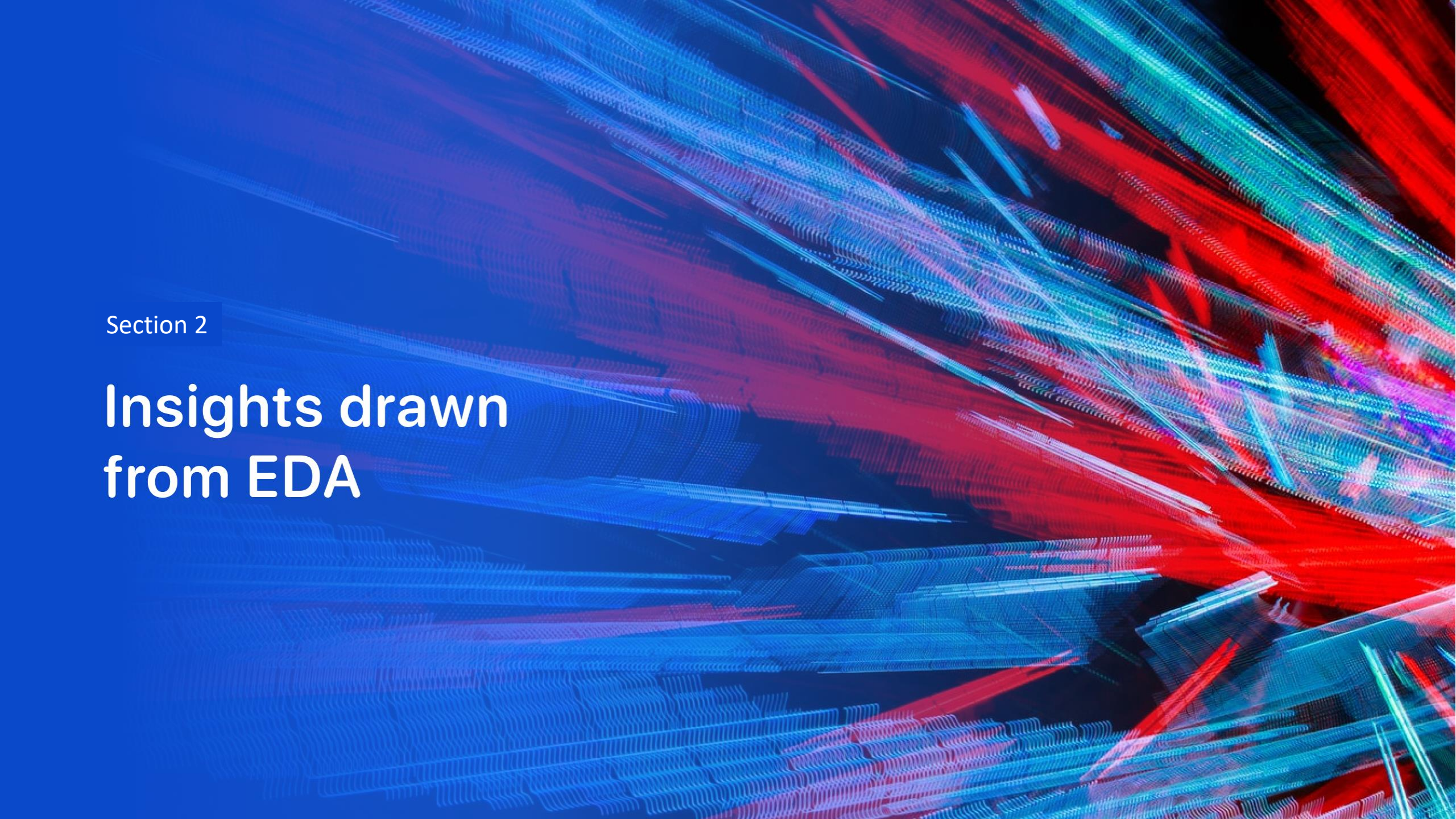
- **Machine Learning Prediction**
- We used the **Scikit-learn** library to build and evaluate various machine learning classification models. The prediction workflow followed these key steps:
- **Data Standardization**
  - Features were scaled to ensure consistent ranges, improving model performance and convergence.
- **Data Splitting**
  - The dataset was divided into **training** and **testing** subsets to evaluate model generalization.
- **Model Development**
  - The following classification algorithms were implemented:
    - **Logistic Regression**
    - **Support Vector Machine (SVM)**
    - **Decision Tree**
    - **K-Nearest Neighbors (KNN)**
- **Model Training**
  - Each model was trained using the training data and fitted to learn the patterns in the features.
- **Hyperparameter Tuning**
  - We explored different combinations of hyperparameters to optimize each model's performance.
- **Model Evaluation**
  - Models were assessed using **accuracy scores** and **confusion matrices** to compare their predictive capabilities.

Results

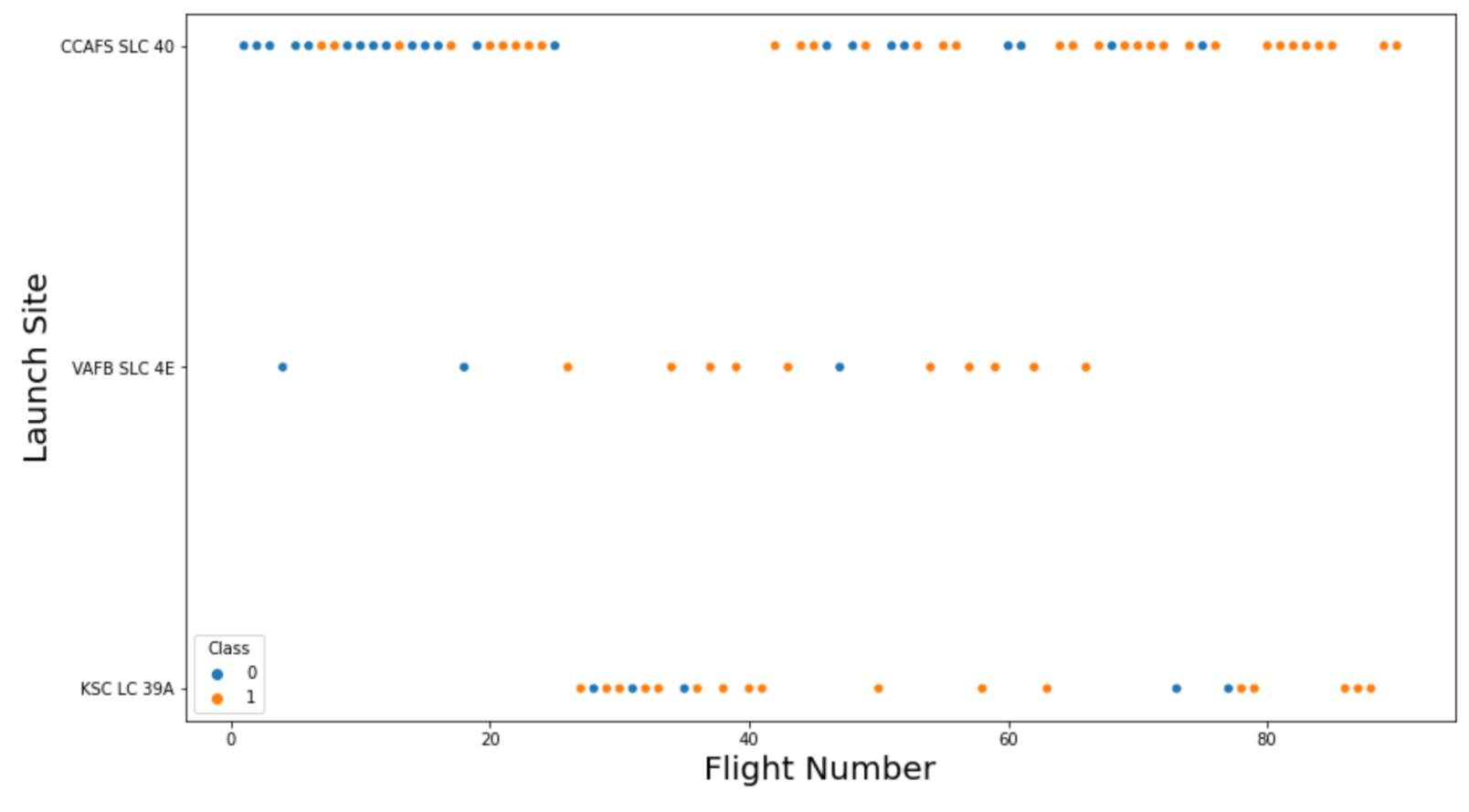
- **Results Overview**
- The project results are presented across five main sections, each highlighting different aspects of the analysis and insights:
- **SQL – Exploratory Data Analysis (EDA with SQL)**
  - Extracted meaningful insights using SQL queries on the structured dataset.
- **Matplotlib and Seaborn – EDA with Visualizations**
  - Visualized trends, correlations, and distributions to better understand the data.
- **Folium – Interactive Mapping**
  - Created geographic visualizations to explore launch site locations and outcomes.
- **Dash – Interactive Dashboard**
  - Developed an interactive web-based dashboard for dynamic data exploration and storytelling.
- **Predictive Analysis**
  - Built and evaluated machine learning models to predict launch success based on input features.
- In all visualizations and analyses, **Class 0** indicates a **failed launch**, while **Class 1** indicates a **successful landing** of the Falcon 9 first stage.

Section 2

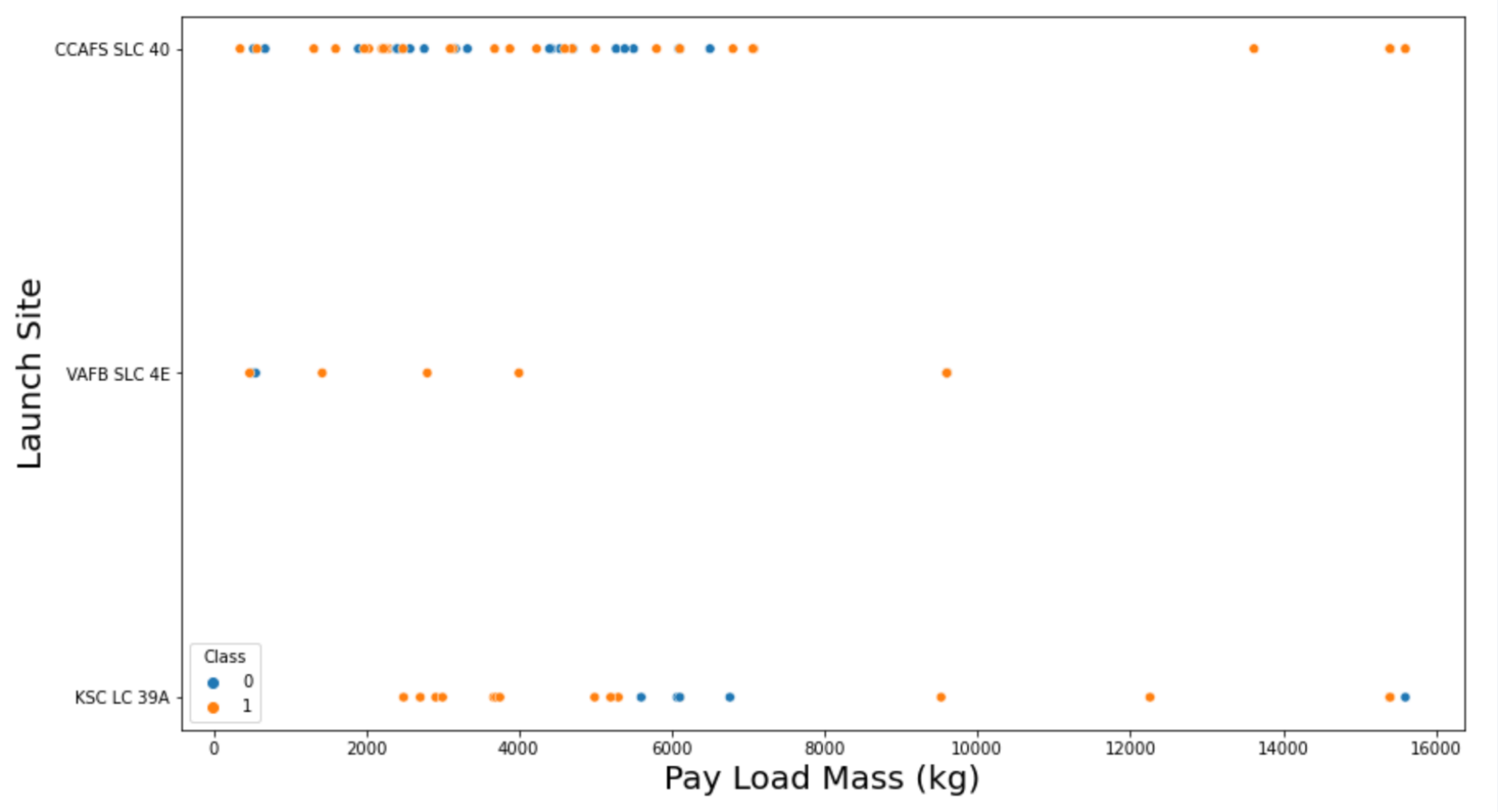
Insights drawn from EDA



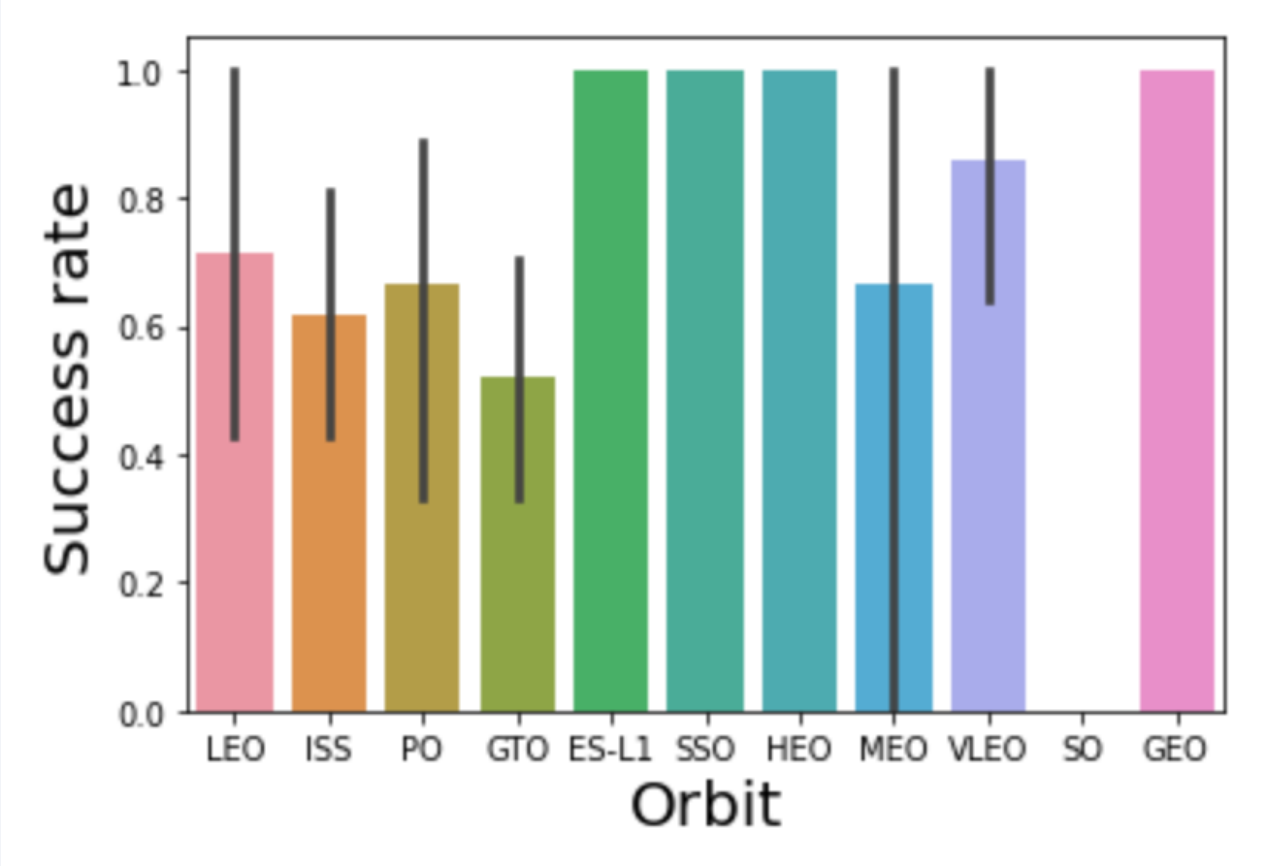
Flight Number vs. Launch Site



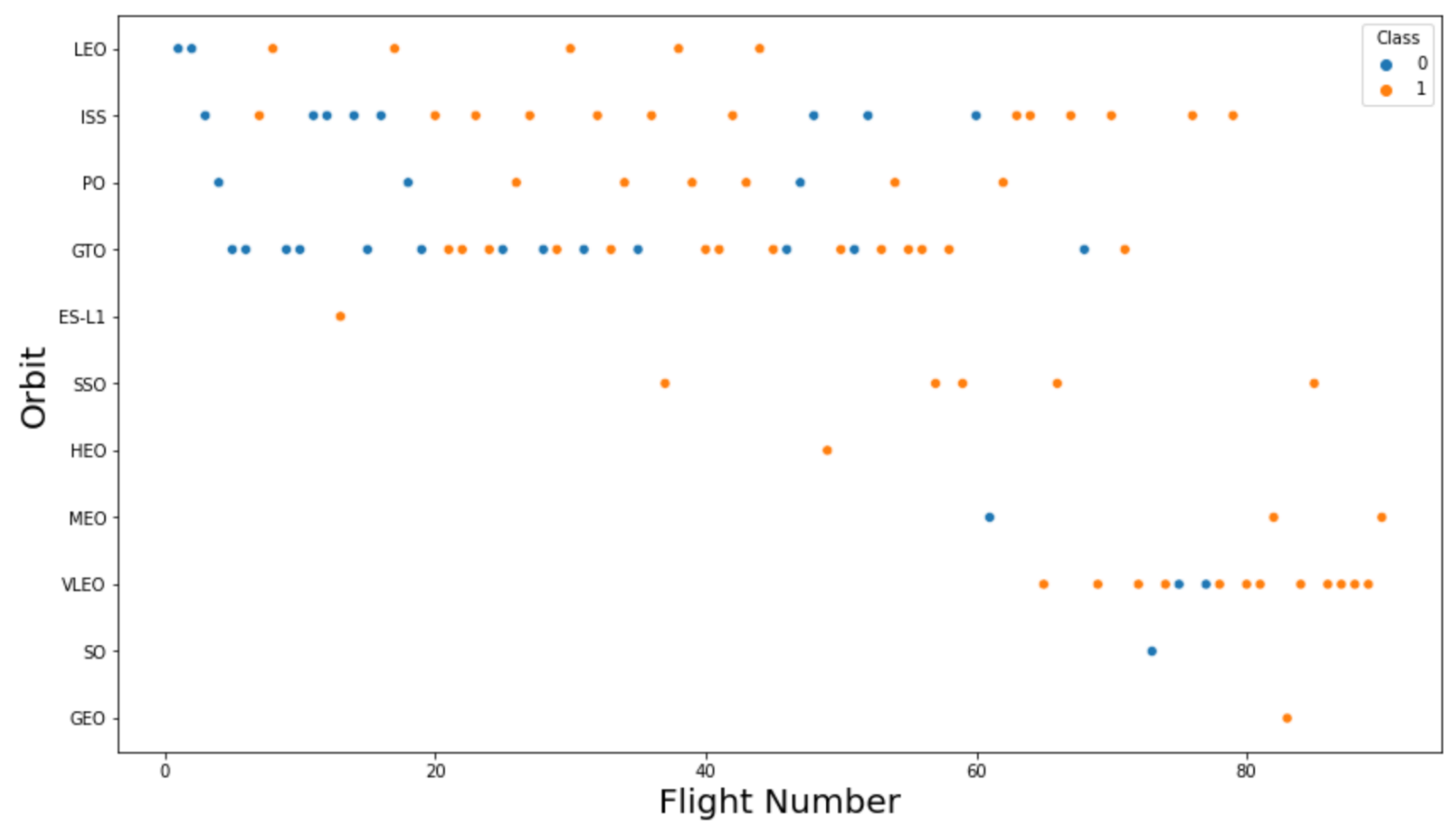
Payload vs. Launch Site



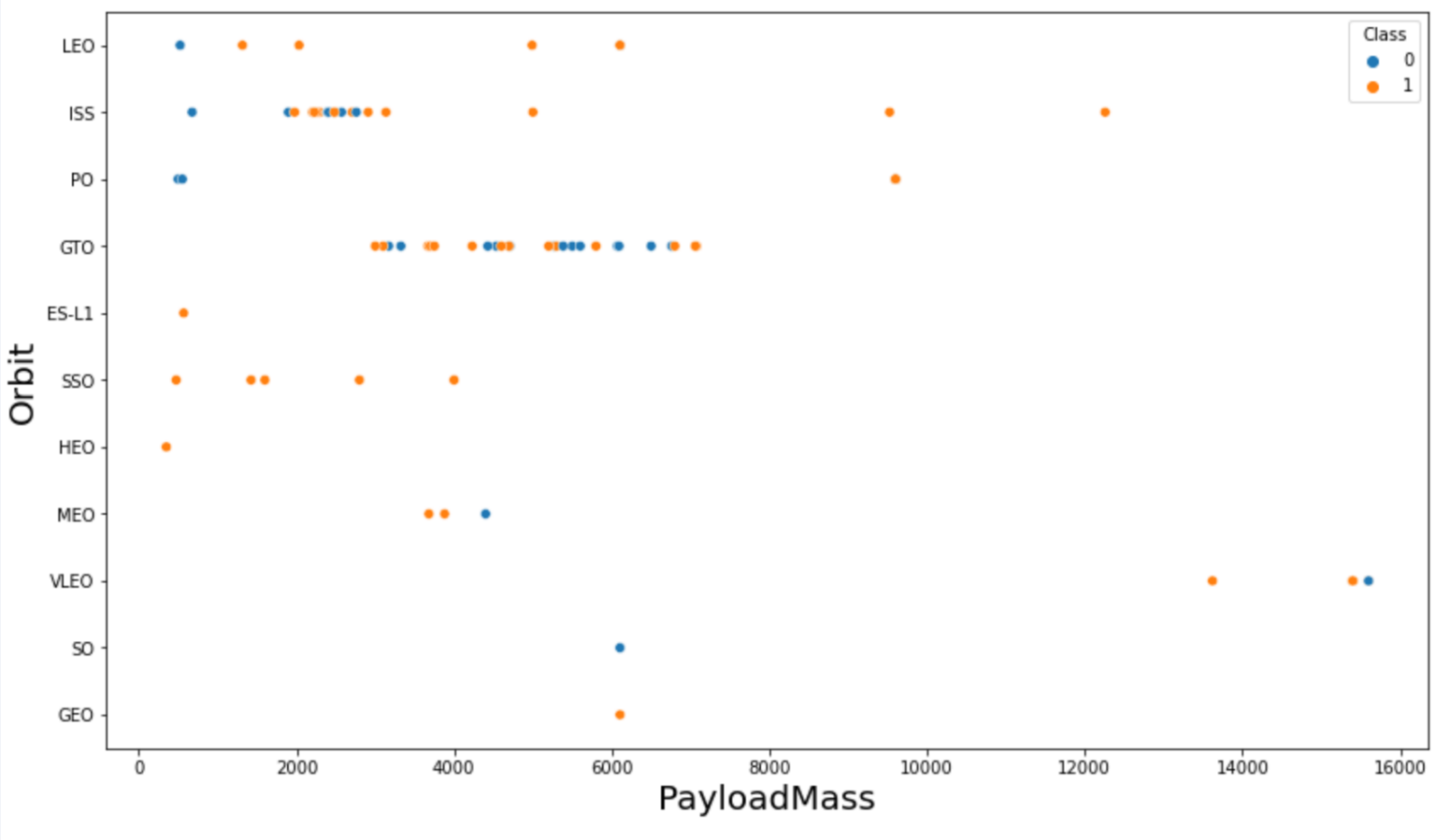
Success Rate vs. Orbit Type



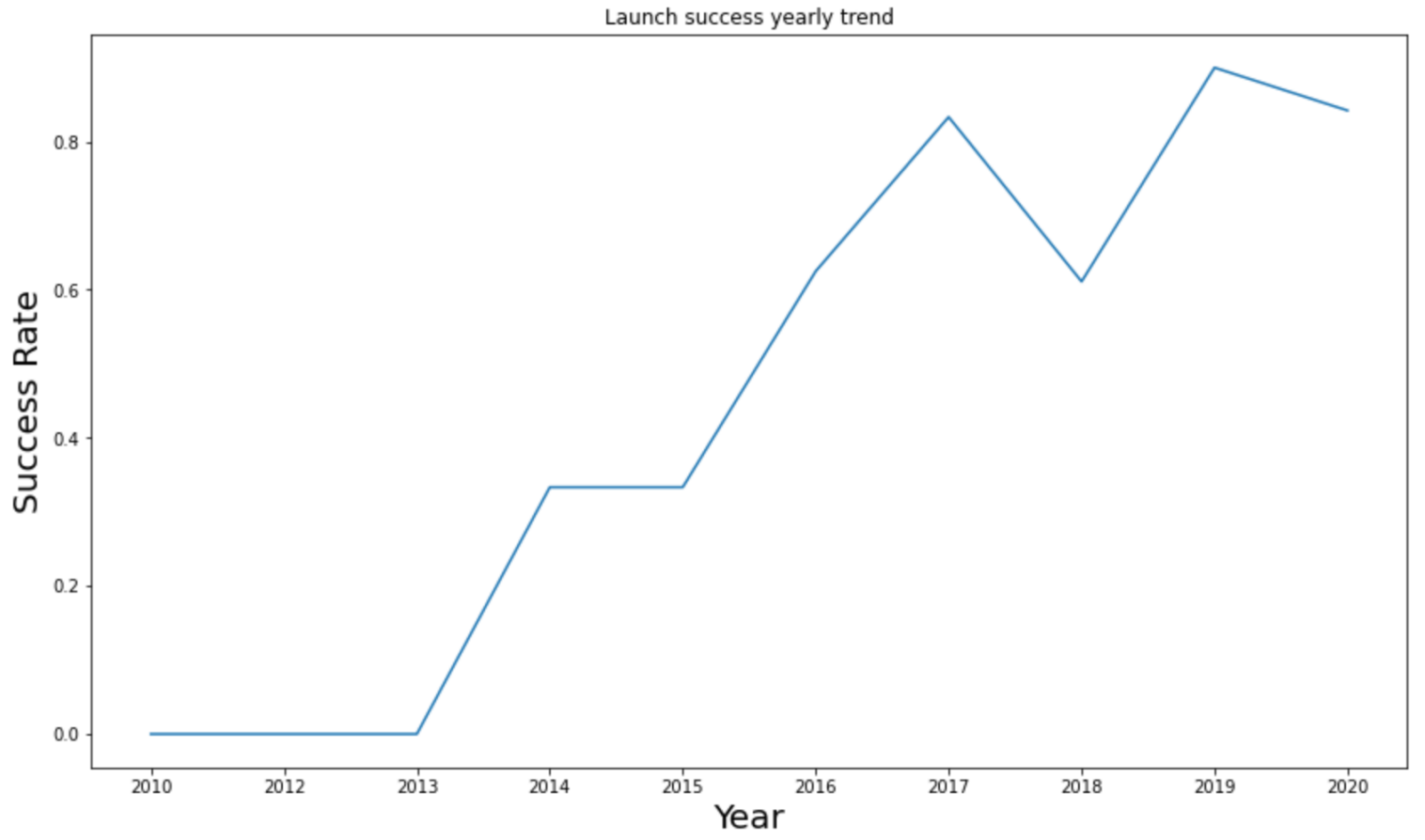
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

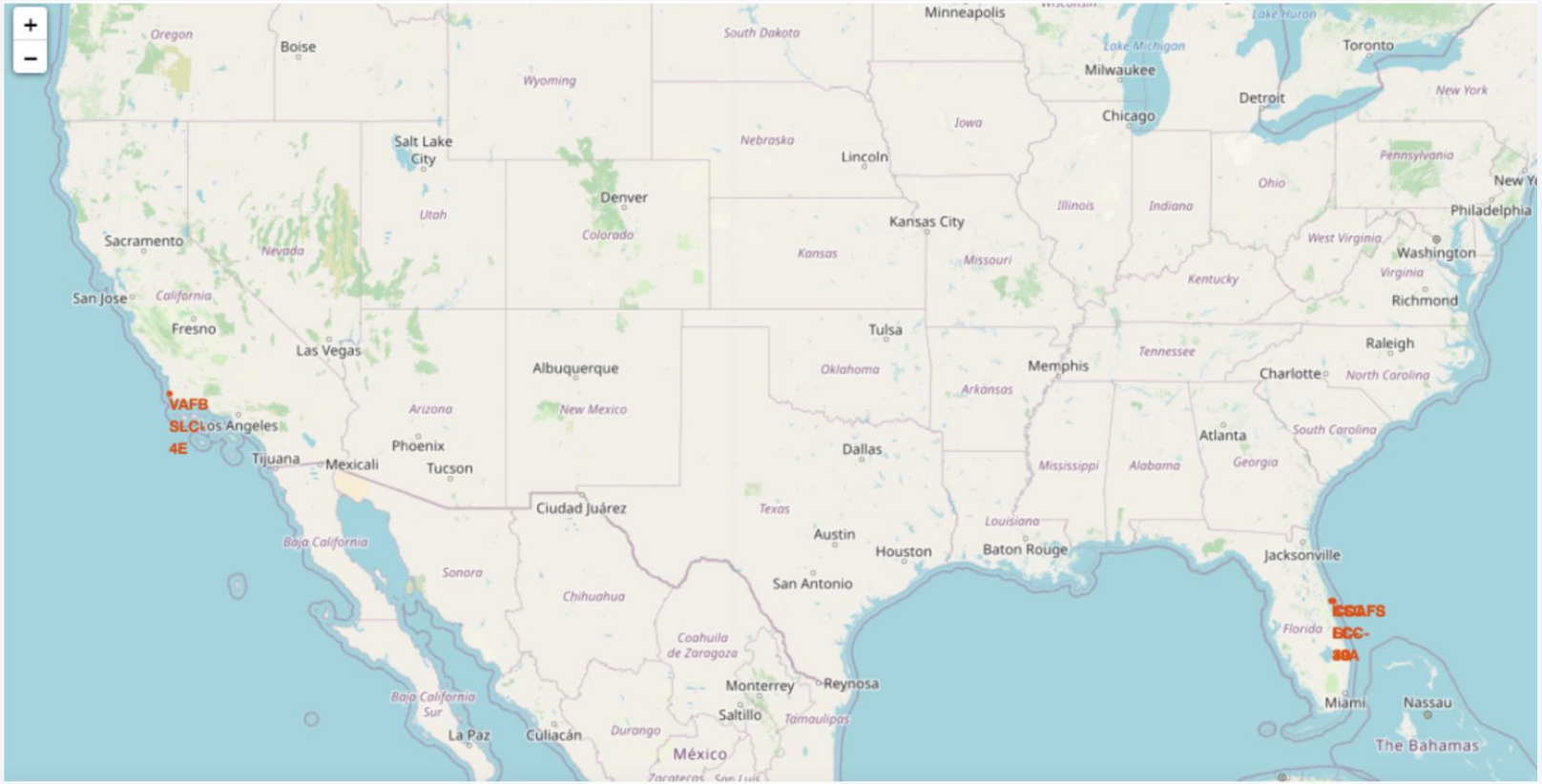
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

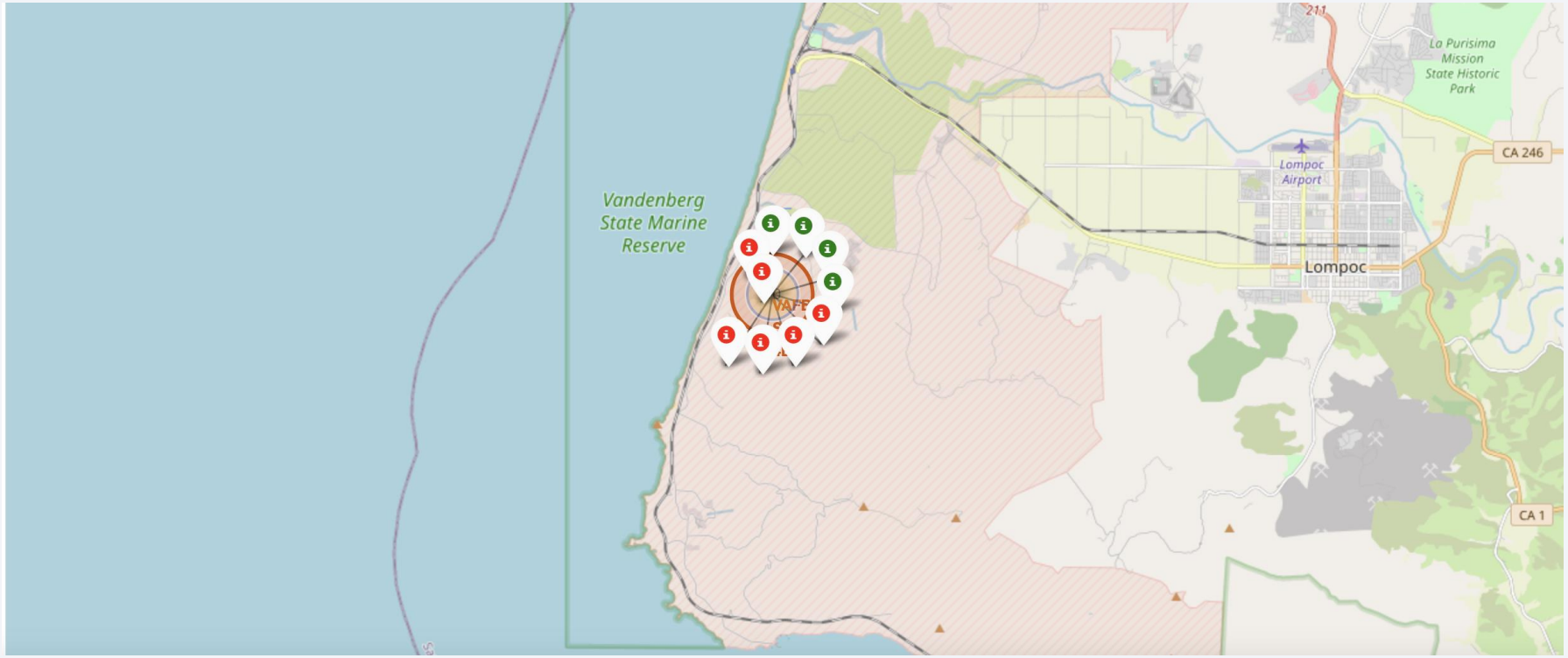
landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

<All launch sites on map>

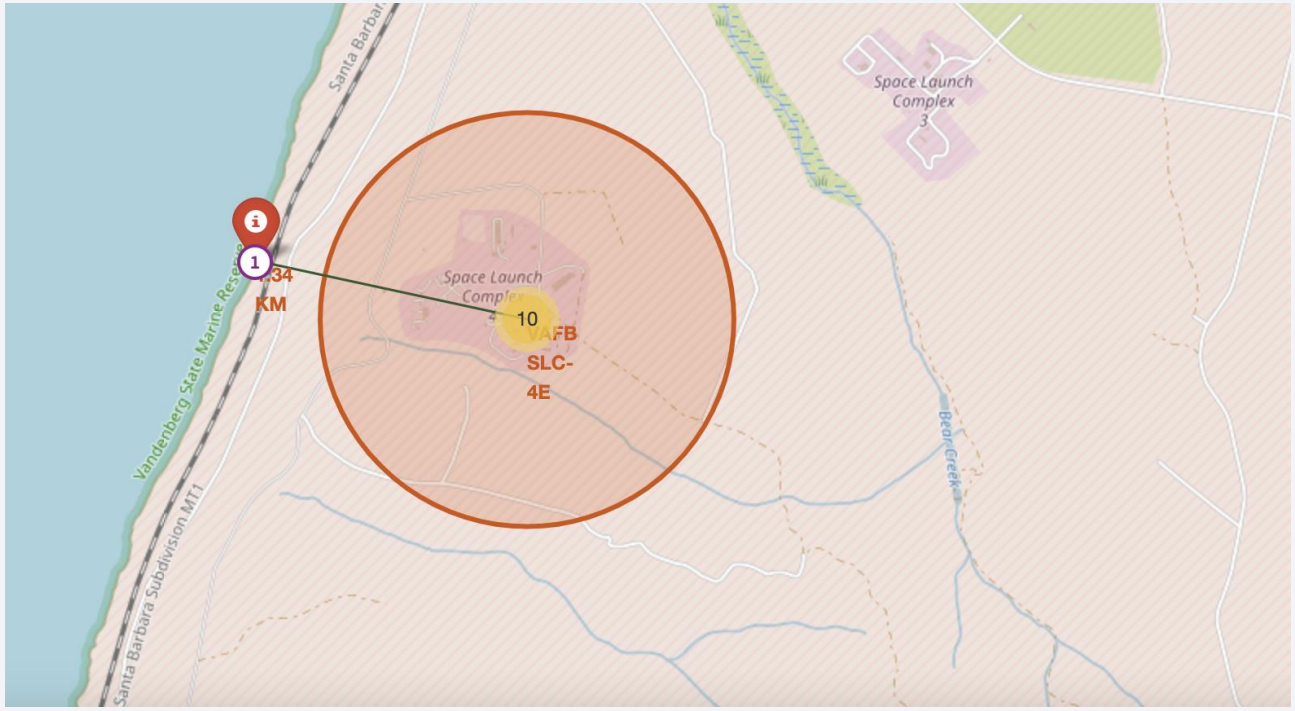


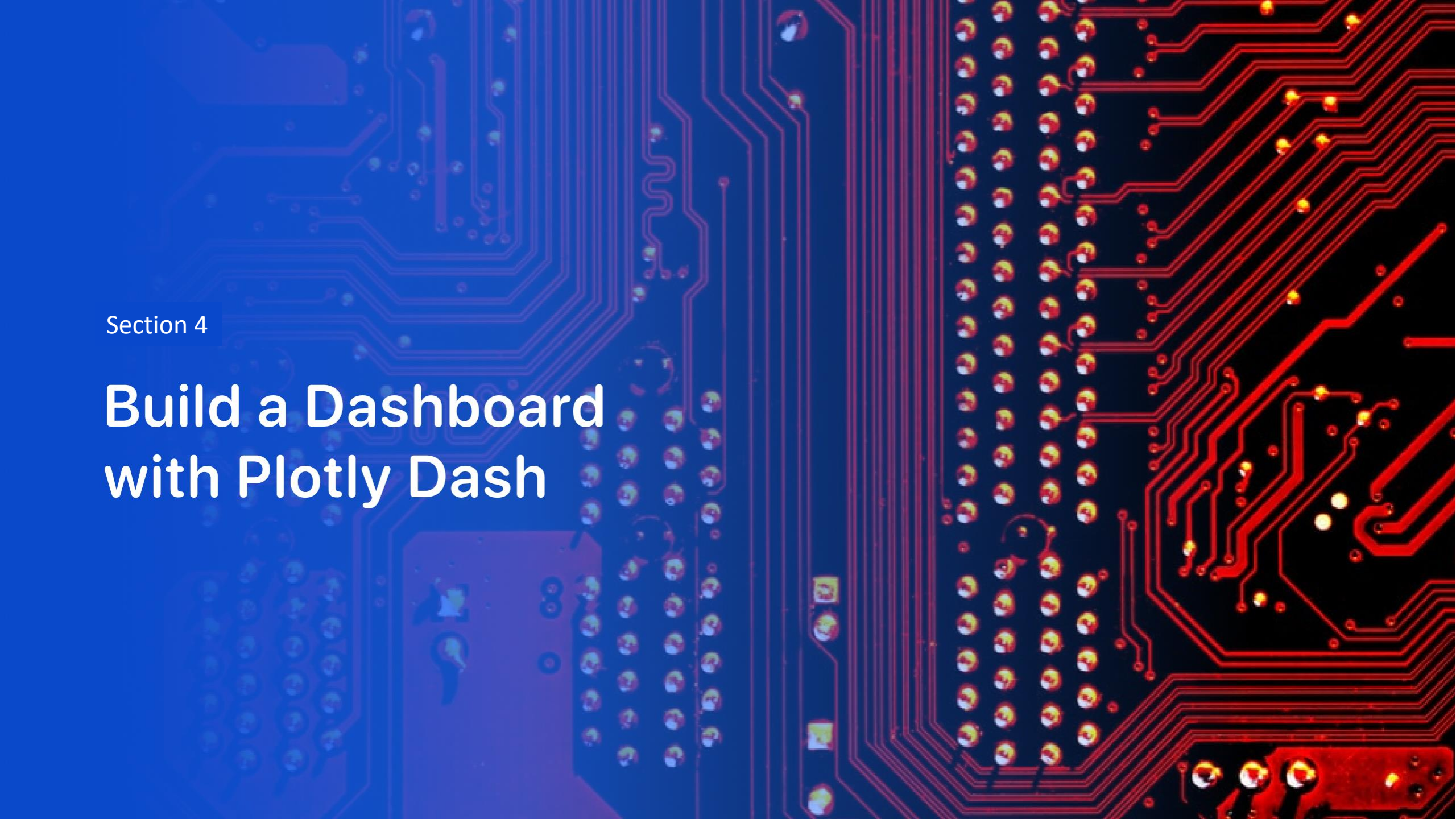
Folium Results



Results Folium

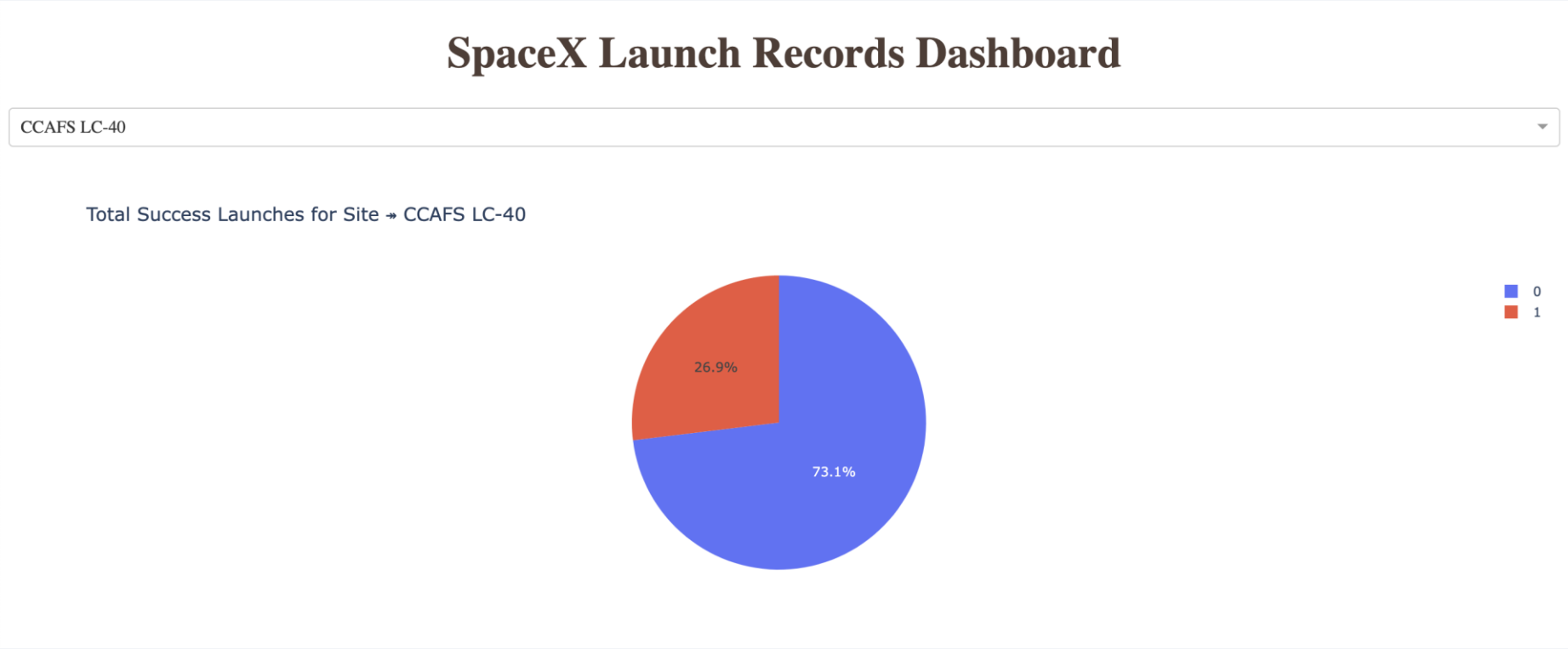
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



Dashboard

- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.etc.

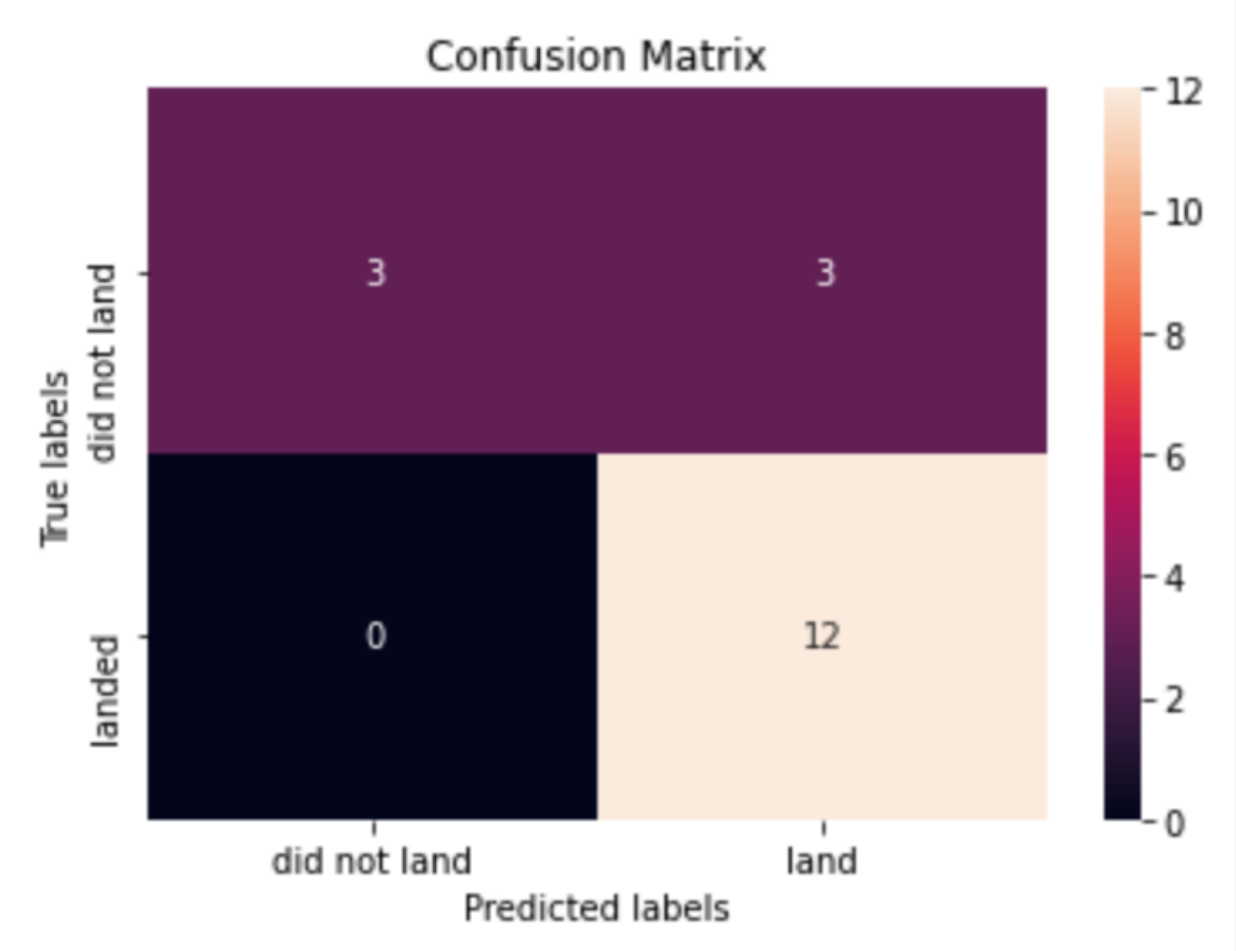
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Logistic regression**
  - **GridSearchCV best score: 0.8464285714285713**
  - **Accuracy score on test set: 0.8333333333333334**

Confusion Matrix



Conclusions

- **Model Comparison and Ranking**
- When evaluated on the test dataset, all four machine learning models produced **identical accuracy scores and confusion matrices**, making it difficult to differentiate their performance based solely on test set results.
- To address this, we used the **best cross-validation scores from GridSearchCV** as a more reliable metric for model comparison. Based on these scores, the models are ranked as follows (from best to worst):
- **Decision Tree** – *Best score: 0.889*
- **K-Nearest Neighbors (KNN)** – *Best score: 0.848*
- **Support Vector Machine (SVM)** – *Best score: 0.848*
- **Logistic Regression** – *Best score: 0.846*
- This ranking suggests that the **Decision Tree classifier** outperformed the others during cross-validation and is therefore considered the most promising model for predicting Falcon 9 landing success.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!