Winning Space Race with Data Science

Cilibiu Nicoleta Mădălina
November 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

This capstone project focuses on predicting whether the first stage of the SpaceX Falcon 9 rocket will successfully land, using various machine learning classification algorithms.

Key project steps include:

- Data collection, cleaning, and formatting
- Exploratory data analysis
- Interactive data visualization
- Machine learning modeling and prediction

The analysis suggests that certain launch features are correlated with landing outcomes (success or failure).

The project concludes that the decision tree algorithm may be the most effective model for predicting the success of the Falcon 9 first-stage landing.

Introduction

In this capstone project, we aim to predict the successful landing of the Falcon 9 rocket's first stage. SpaceX offers Falcon 9 launches for $62 million, compared to other providers that charge upwards of $165 million. This cost reduction is largely due to SpaceX's ability to reuse the rocket's first stage. Predicting the likelihood of a successful landing can help estimate the launch cost, which could be valuable for competitors looking to challenge SpaceX.

Most unsuccessful landings are intentionally planned as controlled ocean landings by SpaceX.

The central question of this project is: given specific features of a Falcon 9 launch, such as payload mass, orbit type, and launch site, can we predict whether the rocket's first stage will land successfully?

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-nearest neighbors (KNN)

Data Collection – 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

Data Collection - Scraping

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches& oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

	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 Wrangling

The data is preprocessed to ensure there are no missing values, and categorical features are converted using one-hot encoding.

- An additional column, named ‘Class’, is added to the dataset to indicate the launch outcome: 0 for a failed launch and 1 for a successful one.
- The final dataset consists of 90 rows (instances) and 83 columns (features).

EDA

Pandas and NumPy:

- Functions from the Pandas and NumPy libraries are used to extract basic insights from the data, such as:
  - The number of launches at each launch site
  - The frequency of each orbit type
  - The count and frequency of different mission outcomes

SQL:

- SQL queries are applied to answer specific questions about the dataset, including:
  - Identifying the unique launch sites involved in the missions
  - Calculating the total payload mass carried by boosters for NASA's CRS missions
  - Determining the average payload mass for the booster version F9 v1.1

Data Visualization

Matplotlib and Seaborn:

- Functions from the Matplotlib and Seaborn libraries are used to create scatterplots, bar charts, and line charts to visualize relationships within the data.
  - These visualizations help explore key relationships, including:
  - The link between flight numbers and launch sites
  - The relationship between payload mass and launch site
  - Success rates across different orbit types

Folium:

- The Folium library is used to generate interactive maps for deeper geographic insights.
- Specifically, Folium is used to:
  - Mark the locations of all launch sites on a map
  - Display successful and failed launches at each site
  - Highlight distances between each launch site and nearby locations, such as cities, railways, and highways

Data Visualization

Dash:

- The Dash library is used to create an interactive interface with dropdown menus and range sliders for dynamic data exploration.
- Through a pie chart and a scatterplot, the dashboard allows users to visualize:
  - The total number of successful launches from each launch site
  - The correlation between payload mass and mission outcome (success or failure) for each launch site

Predictive Analysis

Scikit-learn:

- Functions from the Scikit-learn library are used to develop machine learning models for prediction.

The machine learning process includes the following steps:

- Standardizing the data for consistent scaling
- Splitting the dataset into training and test sets
- Building various machine learning models, including:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- Training each model on the training dataset
- Tuning hyperparameters to find the optimal settings for each model
- Evaluating model performance using accuracy scores and confusion matrices

Results

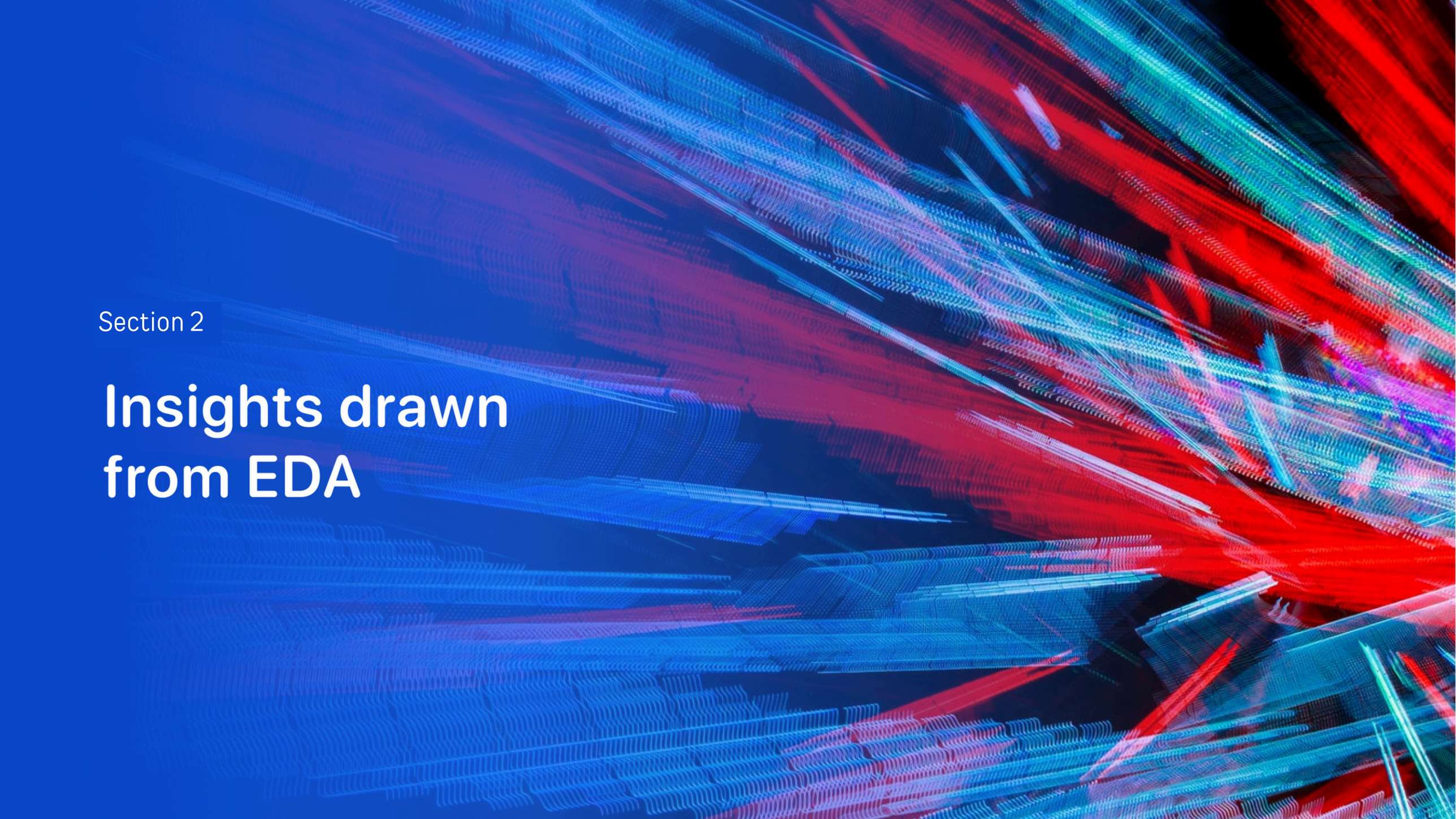
The results are organized into five sections:

- SQL (EDA with SQL): Exploratory data analysis using SQL queries
- Matplotlib and Seaborn (EDA with Visualization): Data visualization for exploratory analysis
- Folium: Interactive maps for geographic data visualization
- Dash: An interactive dashboard for dynamic data exploration
- Predictive Analysis: Machine learning model predictions and evaluation

In all subsequent graphs, Class 0 indicates a failed launch, while Class 1 indicates a successful launch.

Section 2

Insights drawn from EDA



EDA with SQL

- The names of the unique launch sites in the space mission

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

- 5 records where launch sites 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

EDA with SQL

- The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

- The average payload mass carried by booster version F9 v1.1

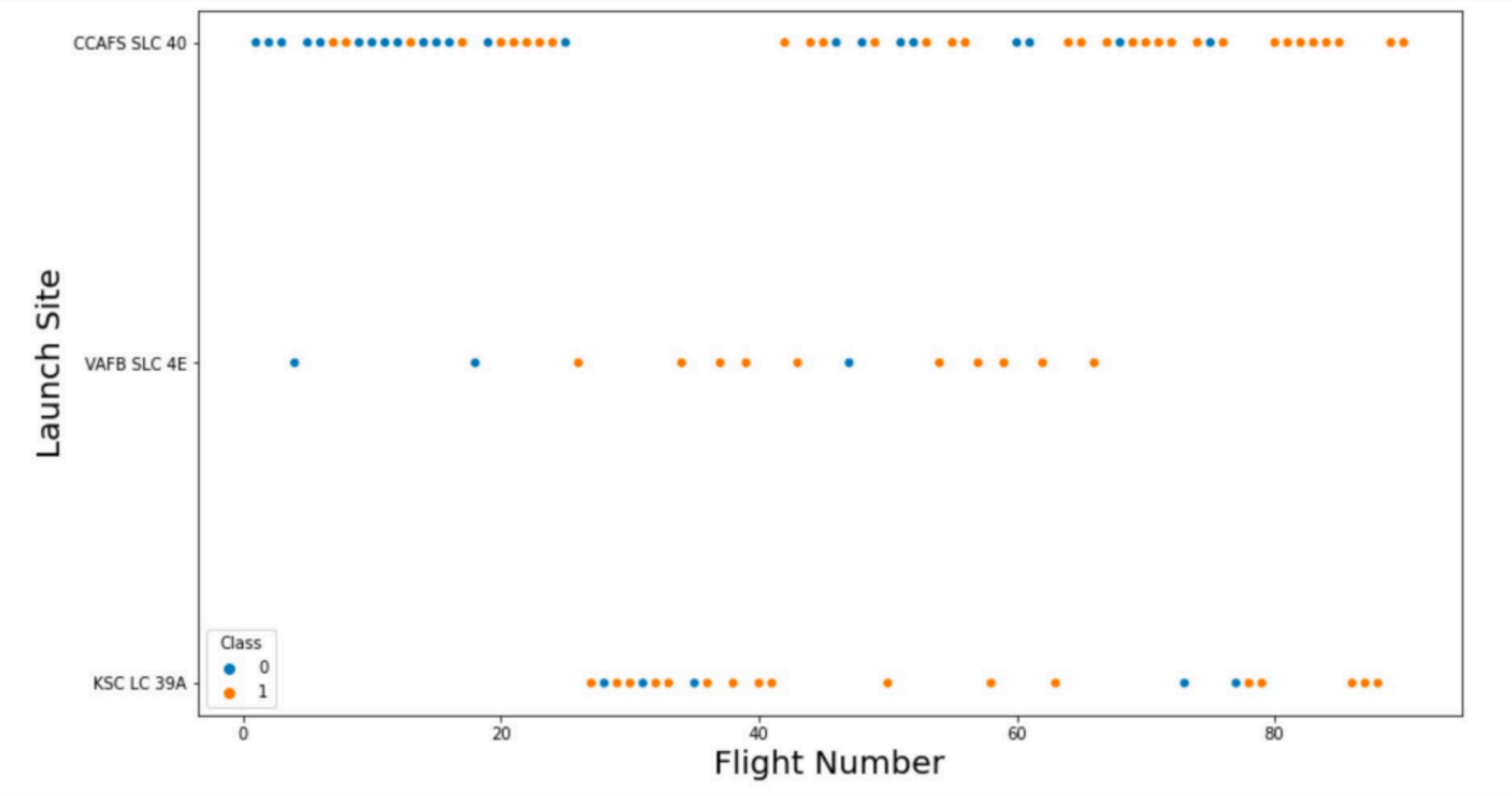
Average payload mass by Booster Version F9 v1.1
2928

- The date when the first successful landing outcome in ground pad was achieved

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

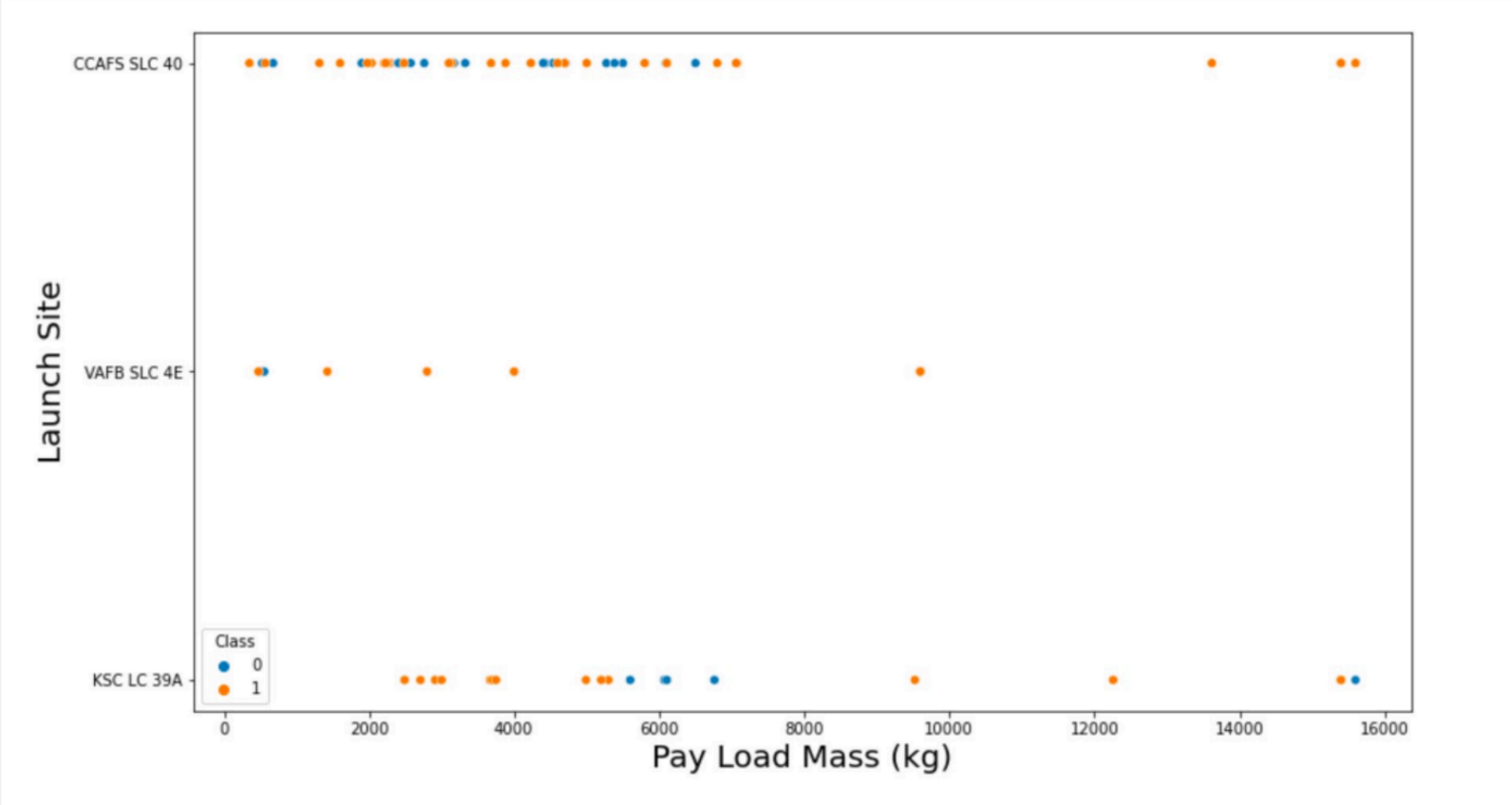
EDA with Visualization

- The relationship between flight number and launch site



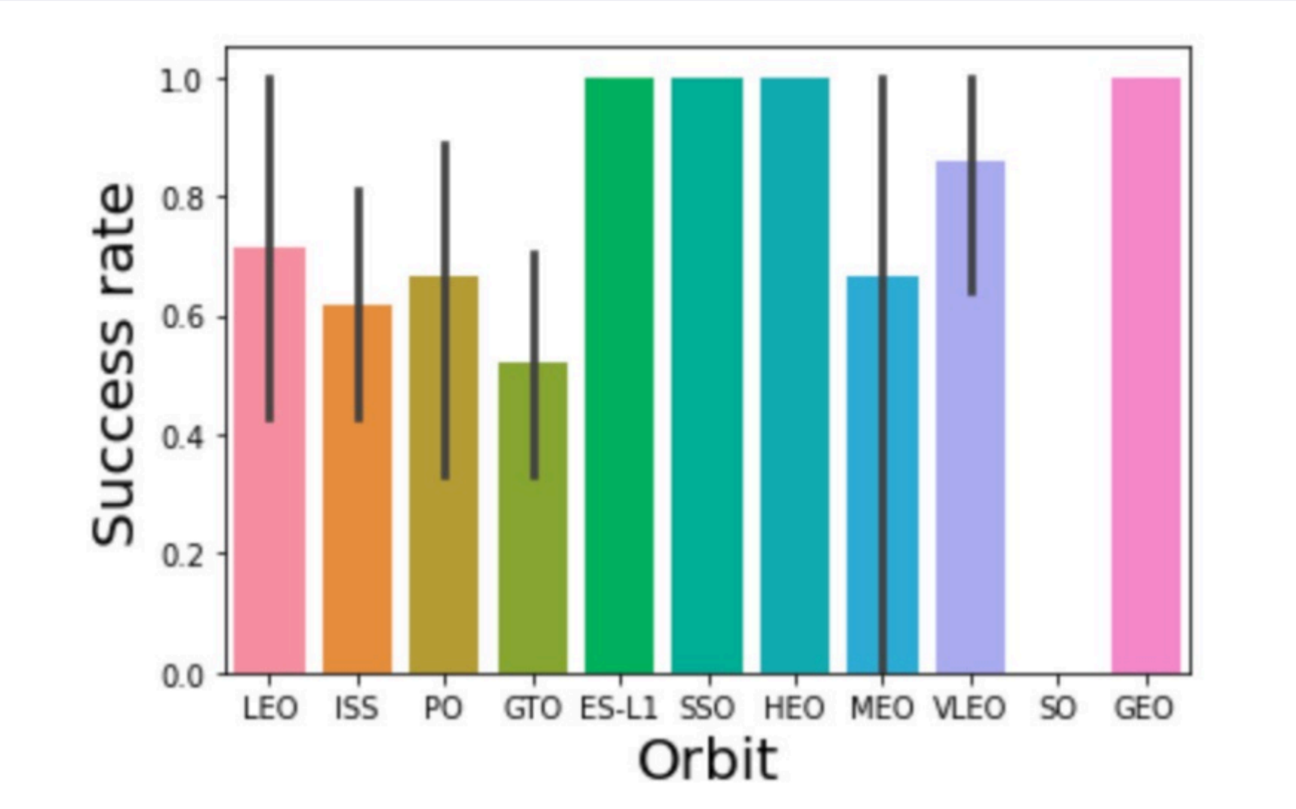
EDA with Visualization

- The relationship between payload mass and launch site



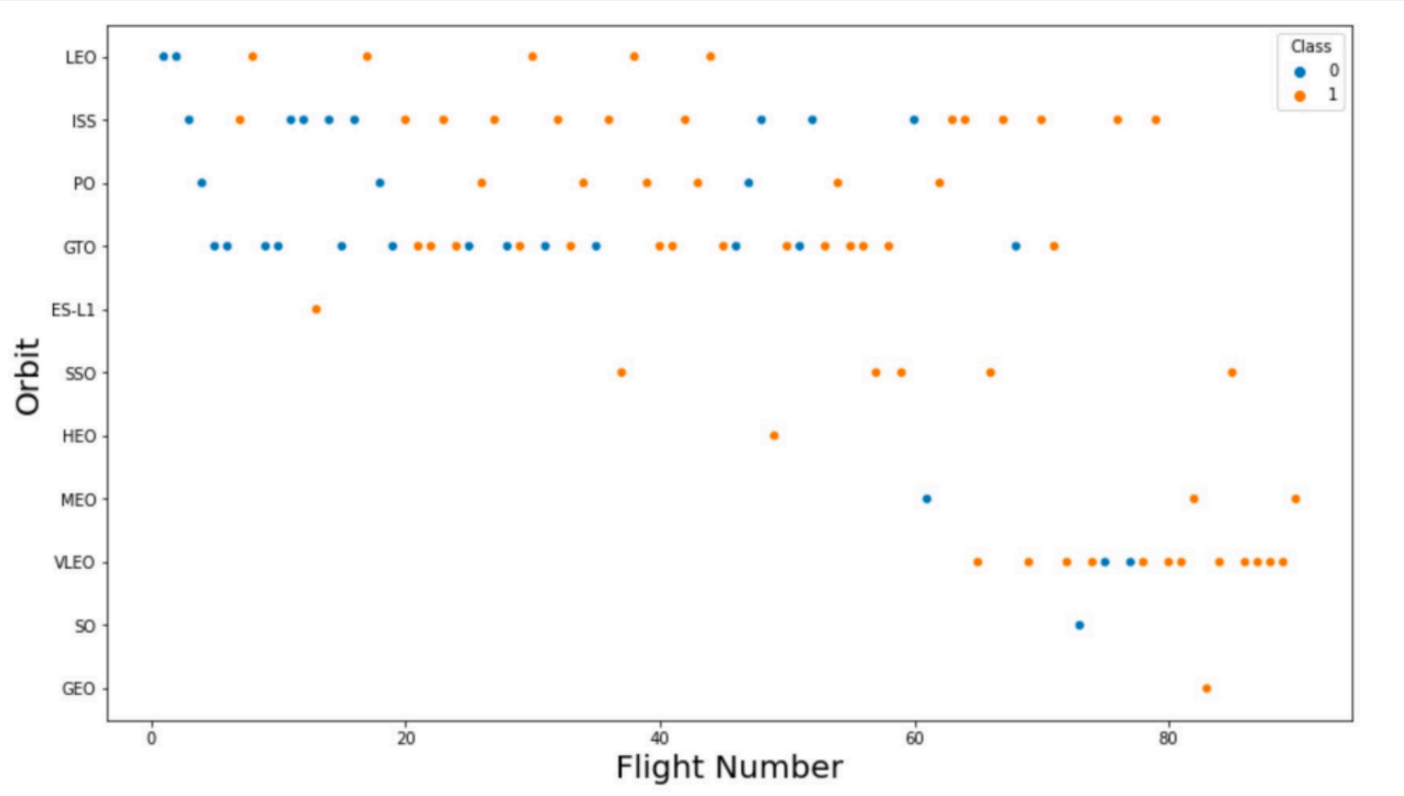
EDA with Visualization

- The relationship between success rate and orbit type



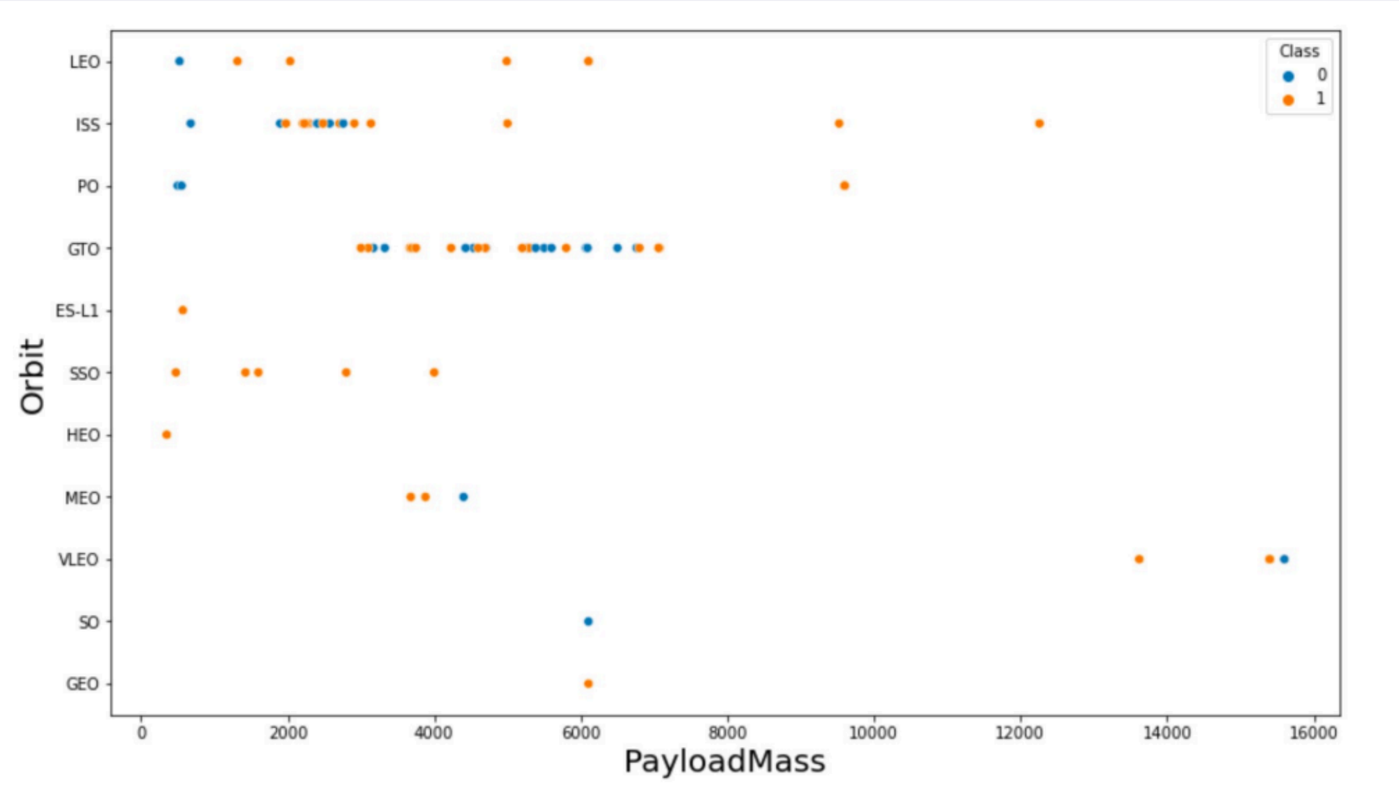
EDA with Visualization

- The relationship between flight number and orbit type



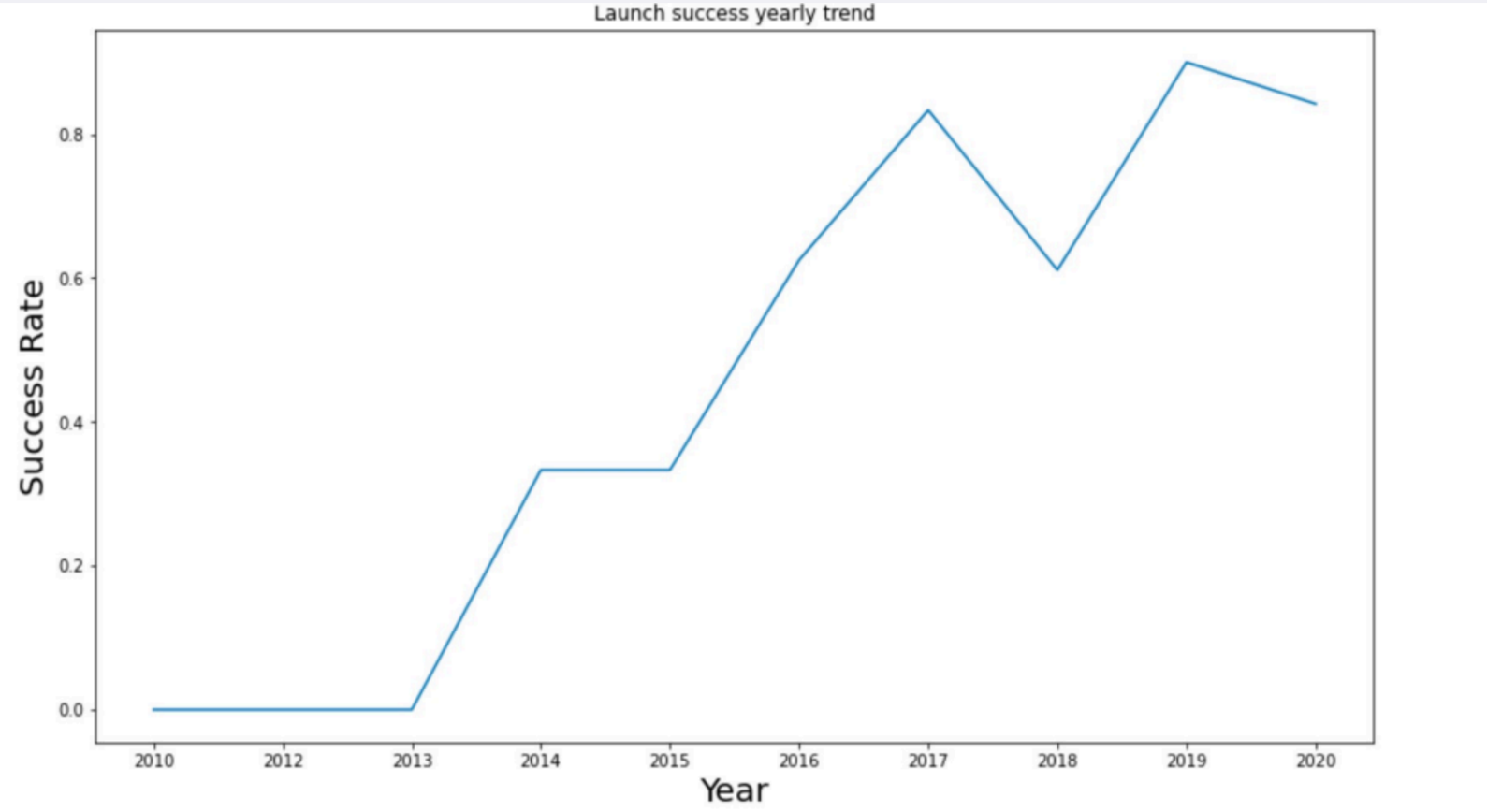
EDA with Visualization

- The relationship between payload mass and orbit type



EDA with Visualization

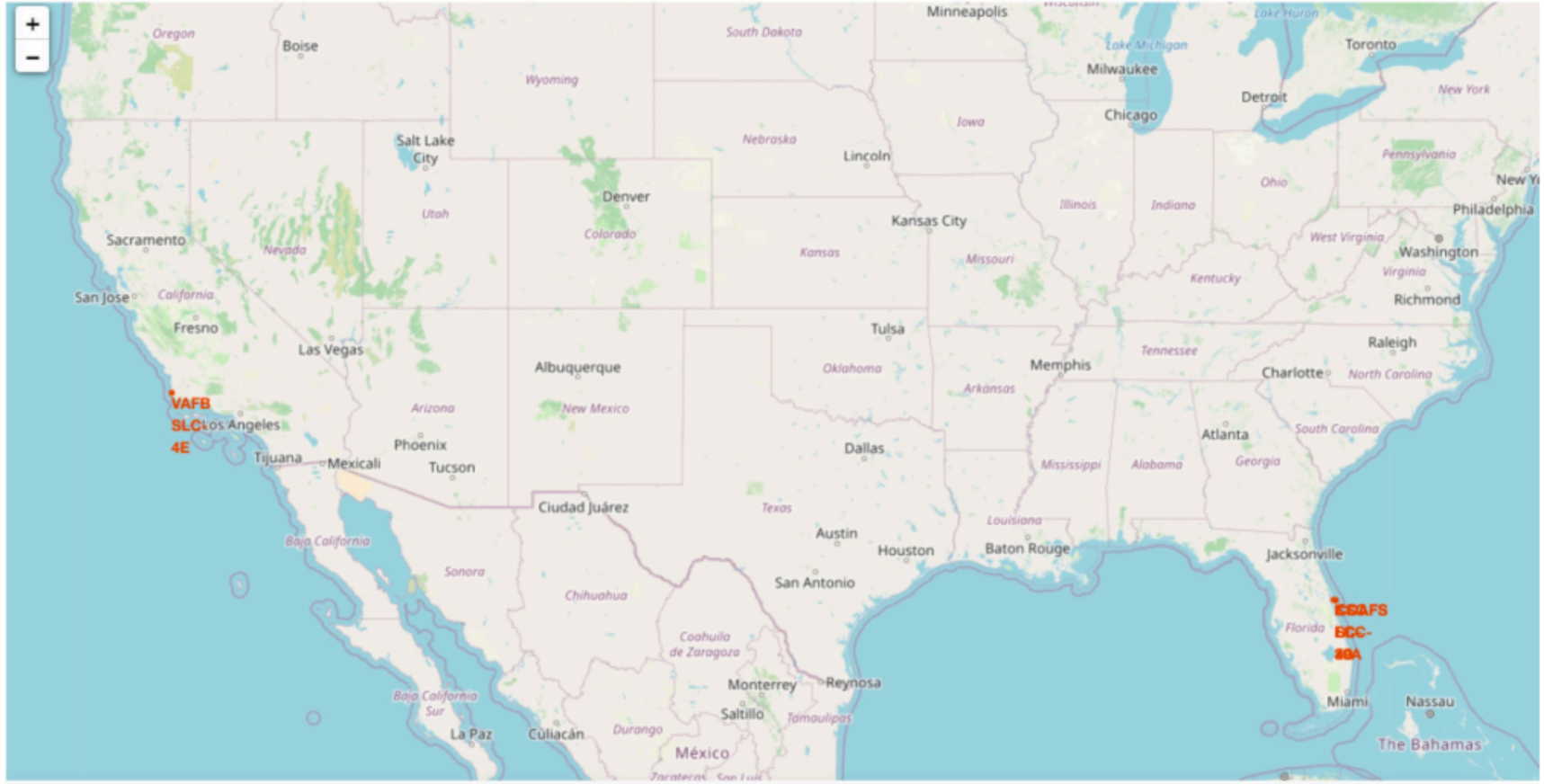
- The launch success yearly trend



Section 3

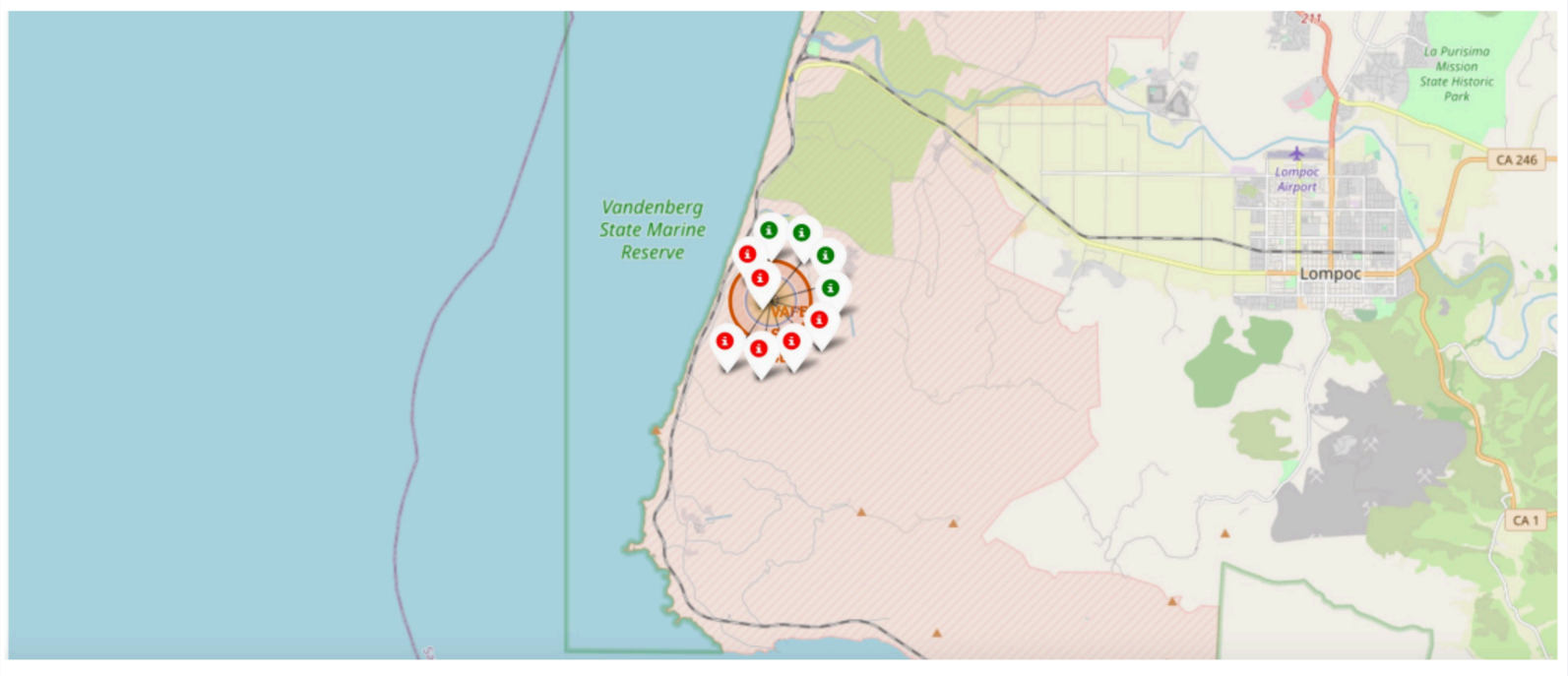
Launch Sites Proximities Analysis

Folium - All launch sites on map



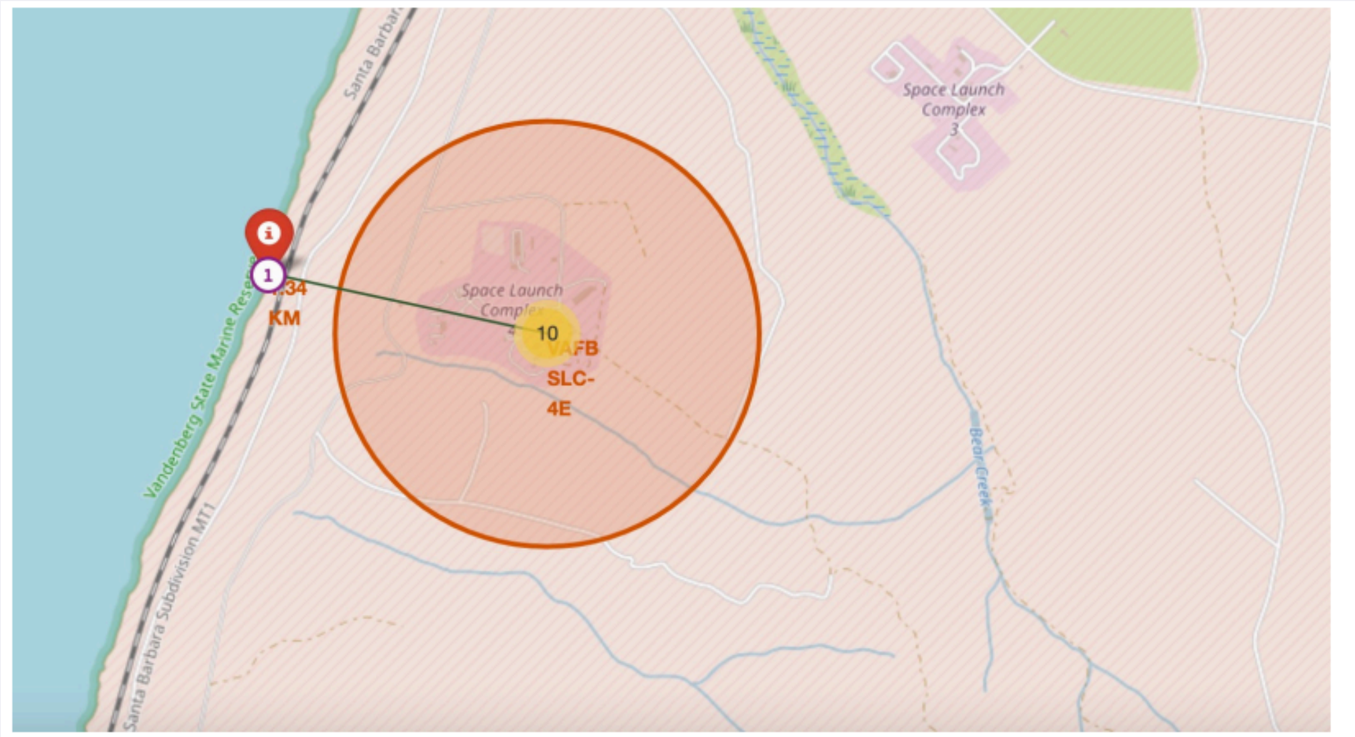
Folium

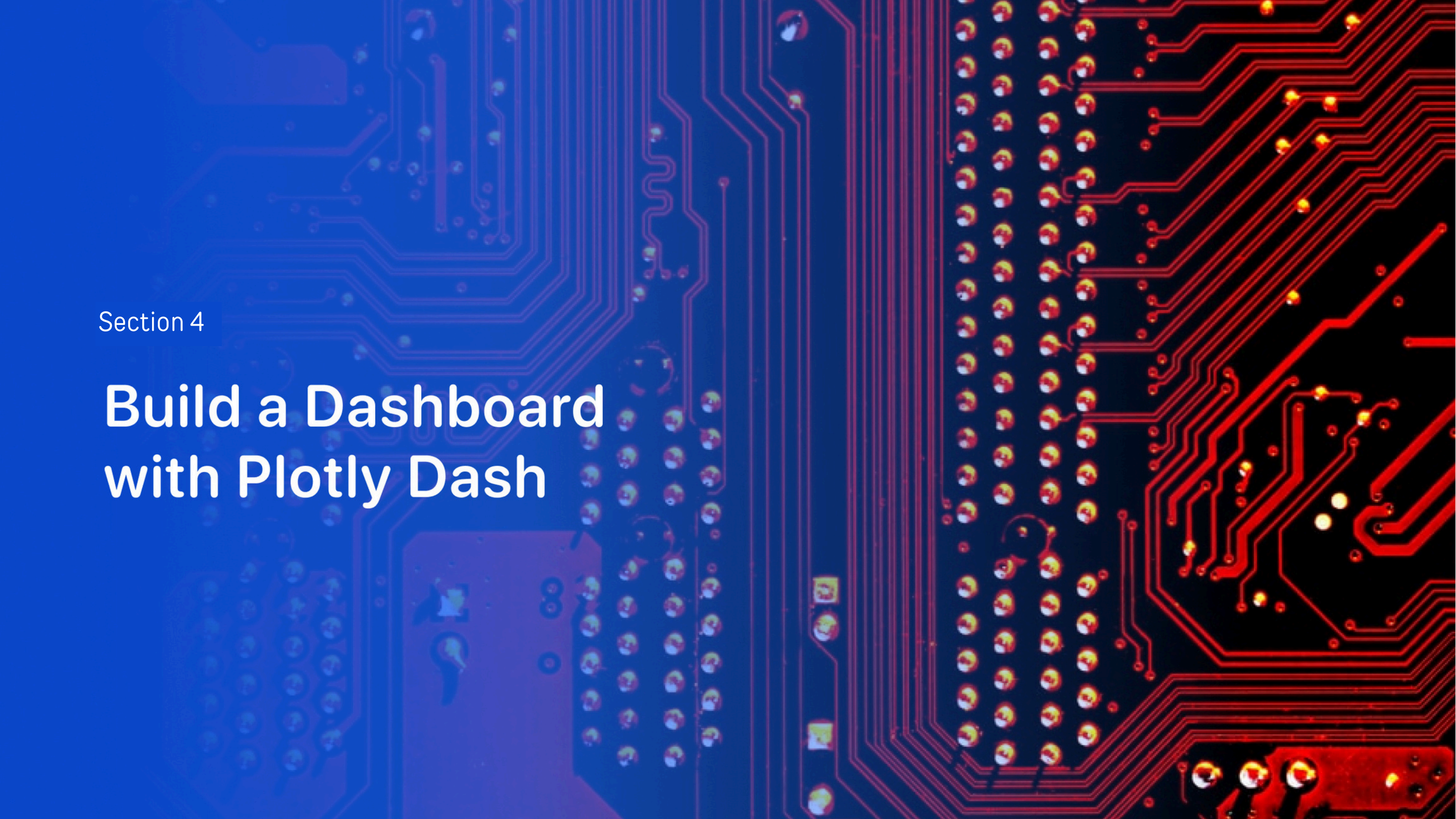
- The succeeded launches and failed launches for each site on map
- If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



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



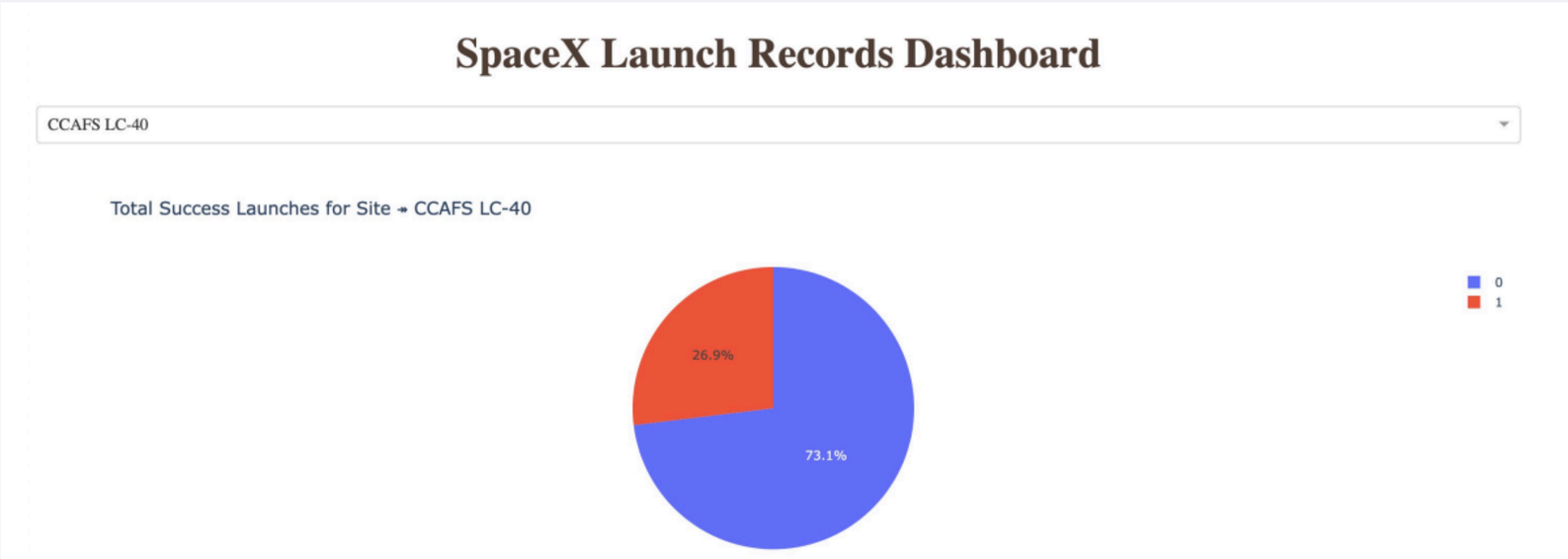


Section 4

Build a Dashboard with Plotly Dash

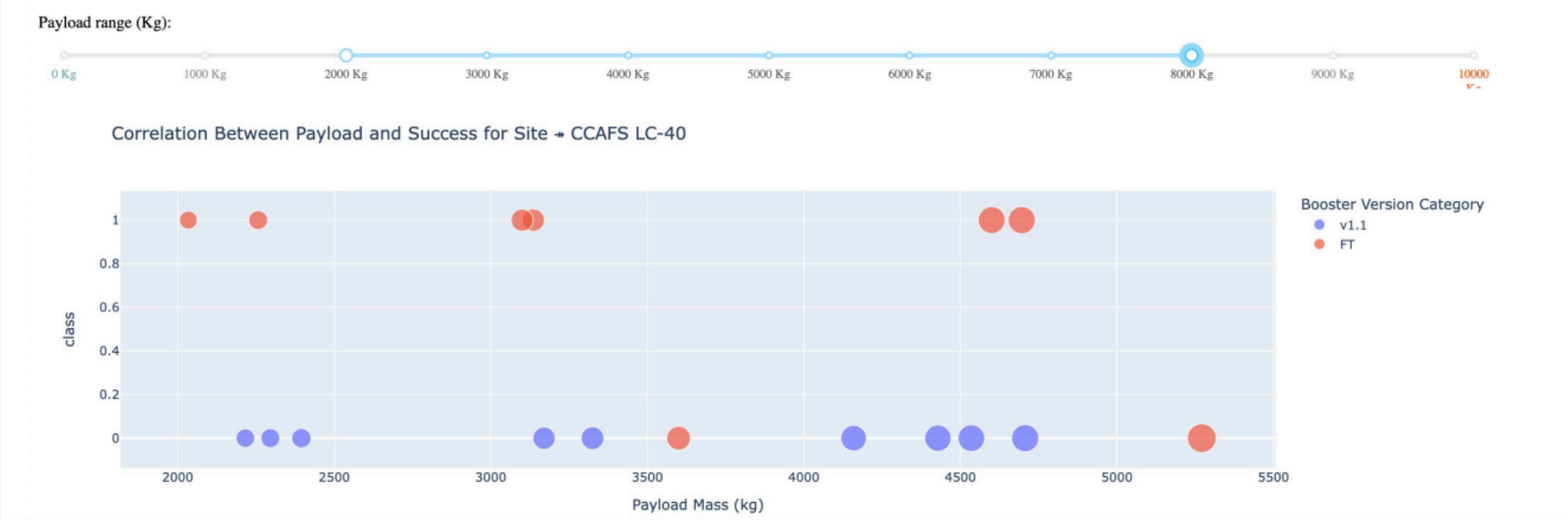
Dashboard 1

- 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 2

- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

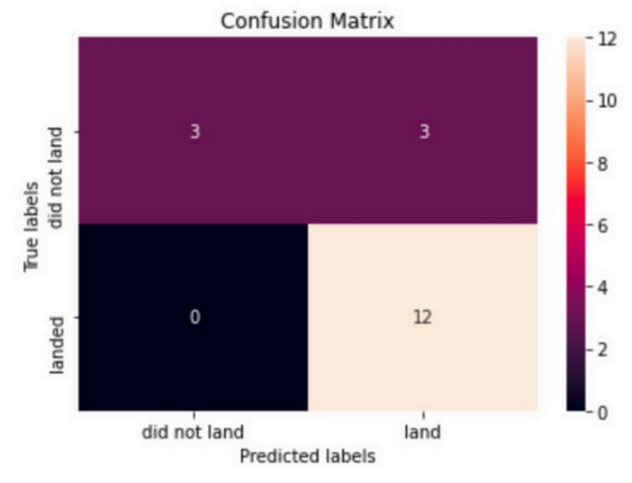


Section 5

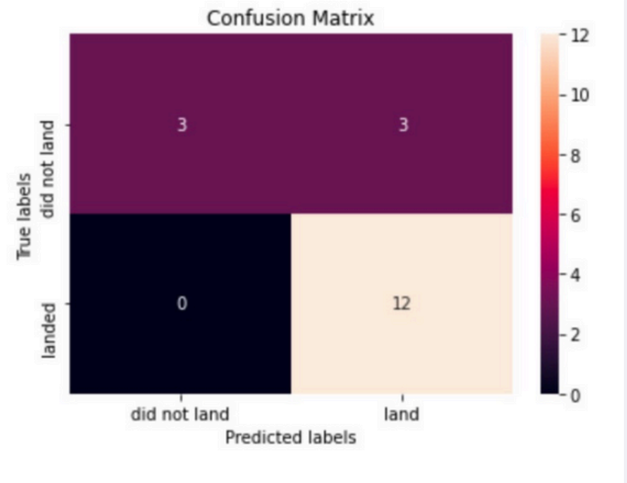
Predictive Analysis (Classification)

Confusion Matrix

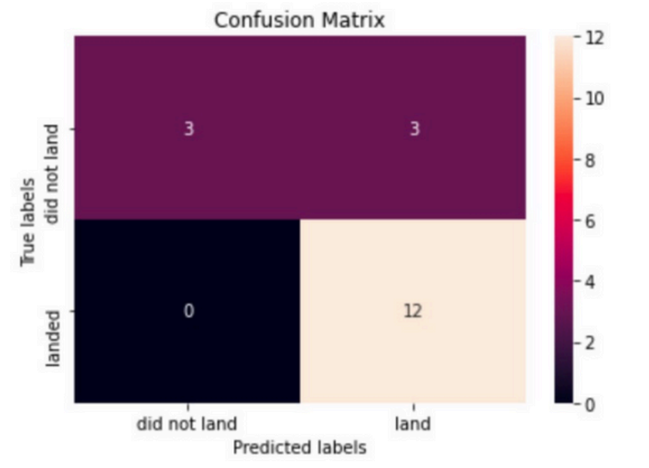
Logistic Regression



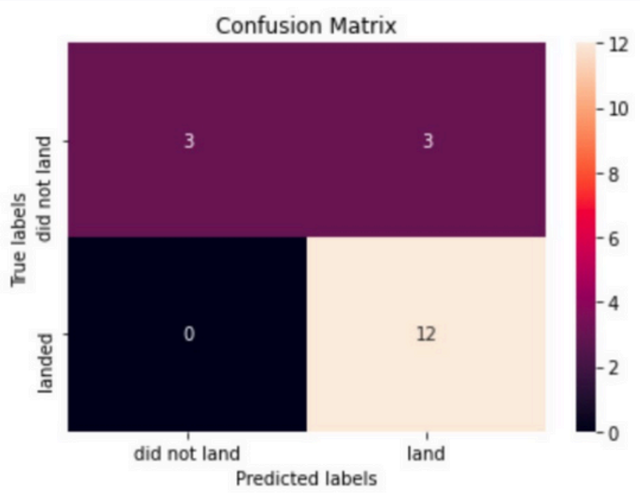
Support vector machine



Decision tree



K nearest neighbors



Conclusions

- When comparing the results of all four models, we observe that they each have the same accuracy score and confusion matrix on the test set.
- To rank the models, we instead use their best scores from GridSearchCV. The models are ranked as follows, from best to worst:
  - Decision Tree – GridSearchCV best score: 0.889
  - K-Nearest Neighbors (KNN) – GridSearchCV best score: 0.848
  - Support Vector Machine (SVM) – GridSearchCV best score: 0.848
  - Logistic Regression – GridSearchCV best score: 0.846

Thank you!