IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Thanh Nguyen>

<November 21 2025>

Outline

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

Executive Summary

In this capstone project, we aim to predict whether the SpaceX Falcon 9 first stage will successfully land by applying a variety of machine learning classification techniques.

The project involves several key stages:

- Collecting, cleaning, and preparing the data
- Performing exploratory data analysis
- Creating interactive visualizations
- Building and evaluating machine learning models

Our visualizations reveal that certain launch characteristics are correlated with the success or failure of the first-stage landing.
Based on our model evaluations, the decision tree algorithm appears to be the most effective at predicting Falcon 9 landing outcomes.

Introduction

- This capstone project focuses on building a model that can **predict the likelihood of a successful Falcon 9 first-stage landing**. The motivation behind this analysis comes from the significant cost advantage SpaceX holds due to its reusable rockets. While competing launch providers charge around **$165 million** per mission, SpaceX advertises launch prices as low as **$62 million**, made possible largely by recovering the first stage. If we can estimate whether the booster will land, we can also estimate the potential cost of a launch, which is useful for any organization looking to compete with SpaceX in commercial launch bids.
- It is also important to recognize that some “unsuccessful” landings are not failures—SpaceX sometimes performs **intentional ocean landings** when booster recovery is not planned.
- The goal of this project is to determine whether we can accurately predict the landing outcome using features from each launch, such as **payload mass, orbit type, launch site, and mission details**. Our central objective is to understand how these variables influence landing success and to develop a model capable of making reliable predictions.

Section 1

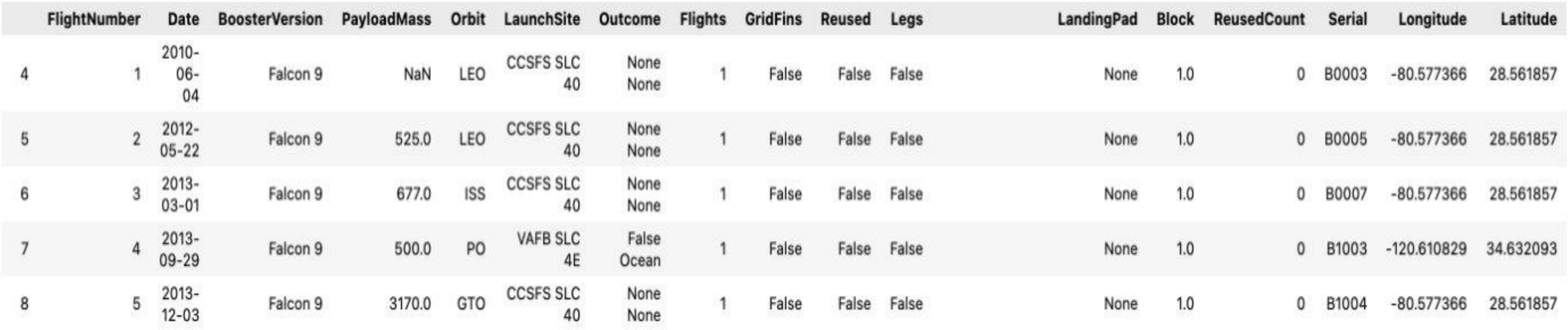
Methodology

Methodology

- The overall methodology includes:
  1. Data collection, wrangling, and formatting, using:
     - SpaceX API
     - Web scraping
  2. Exploratory data analysis (EDA), using:
     - Pandas and NumPy
     - SQL
  3. Data visualization, using:
     - Matplotlib and Seaborn
     - Folium
     - Dash
  4. Machine learning prediction, using
     - Logistic regression
     - Support vector machine (SVM)
     - Decision tree
     - K-nearest neighbors (KNN)

Data Collection

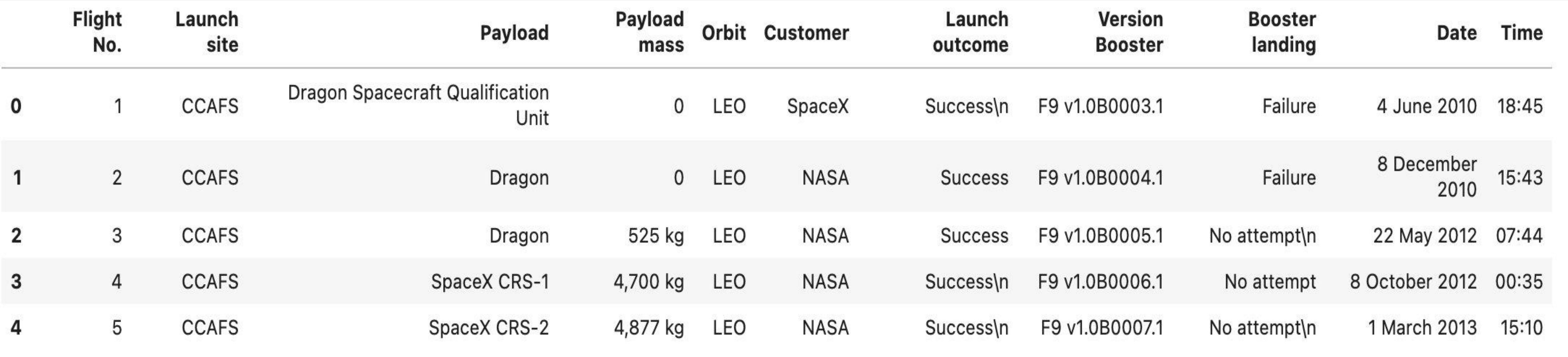
- SpaceX API
  - The API used is https://api.spacexdata.com/v4/rockets/.
  - The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
  - Every missing value in the data is replaced the mean the column that the missing 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 – Web 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

EDA with Pandas

Pandas and NumPy

- We use functions from the Pandas and NumPy libraries to explore and summarize the dataset. These tools help us generate essential insights, such as:
- How many launches occurred at each launch site
- The frequency of different orbit types
- The count and distribution of various mission outcomes

EDA with SQL

SQL

- SQL queries are applied to further analyze and extract information from the data. With SQL, we can answer questions like:
- What launch sites appear in the mission records
- The total payload mass transported by boosters on NASA (CRS) missions
- The average payload mass for launches using the F9 v1.1 booster version

Build an Interactive Map with Matplotlib and Seaborn

We use Matplotlib and Seaborn to create visual charts—like scatterplots, bar charts, and line graphs—to help us better understand the data.

These visualizations show how different features are related, such as:

- How flight number varies by launch site
- How payload mass compares across different launch sites
- How success rates differ by orbit type

Build an Interactive Map with Folium

The Folium library is used to create interactive map visualizations of the launch data. With Folium, we can:

- Plot the locations of all launch sites on a map
- Display markers showing successful and failed launches at each site
- Map the distances from each launch site to nearby features, such as the closest city, railway, or highway

Build a Dashboard with Plotly Dash

Dash is used to build an interactive webpage where users can adjust inputs using a dropdown menu and a range slider.

- The dashboard uses a pie chart and a scatterplot to display:
- The total number of successful launches for each launch site
- How payload mass is related to whether a mission succeeded or failed at each launch site

Machine Learning prediction

We use the Scikit-learn library to build and evaluate our machine learning models.

The prediction process involves several key steps:

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

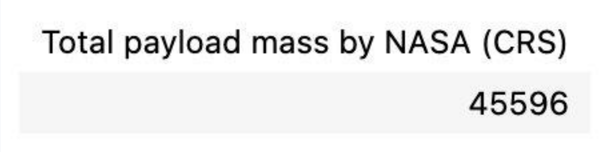
Results

The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

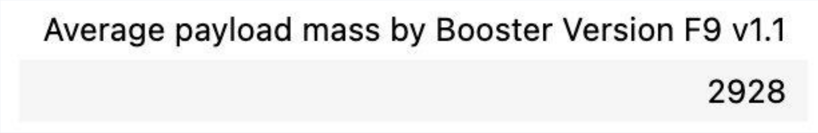
SQL (EDL with SQL)

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



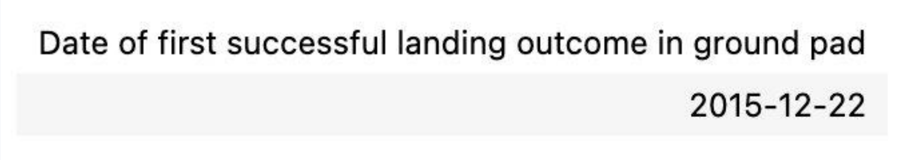
Total payload mass by NASA (CRS)
45596

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



Average payload mass by Booster Version F9 v1.1
2928

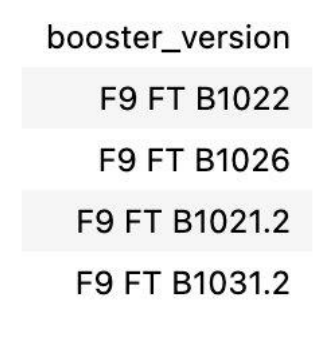
- The date when the first successful landing outcome in ground pad was achieved



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

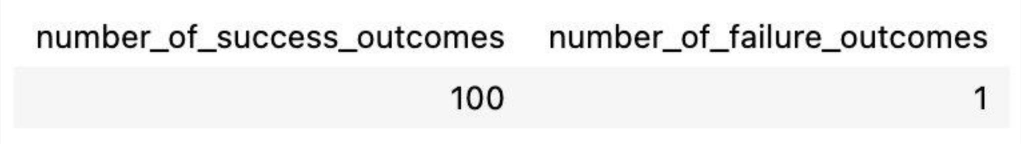
SQL (EDL with SQL)

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000



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

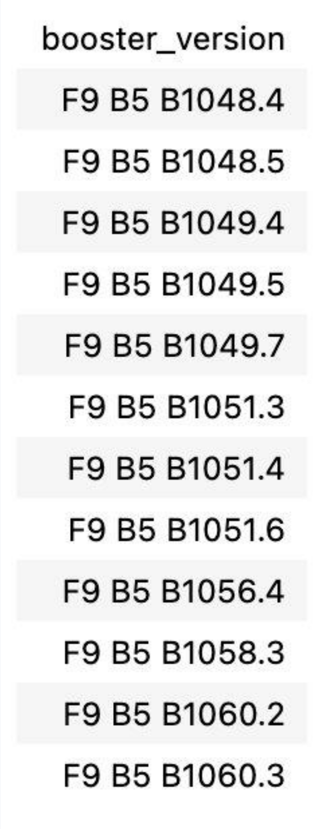
- The total number of successful and failure mission outcomes



number_of_success_outcomes	number_of_failure_outcomes
100	1

SQL (EDL with SQL)

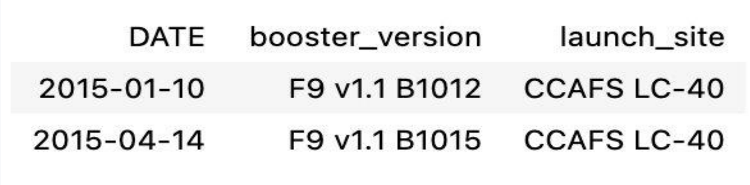
- The names of the booster versions which have carried the maximum payload mass



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

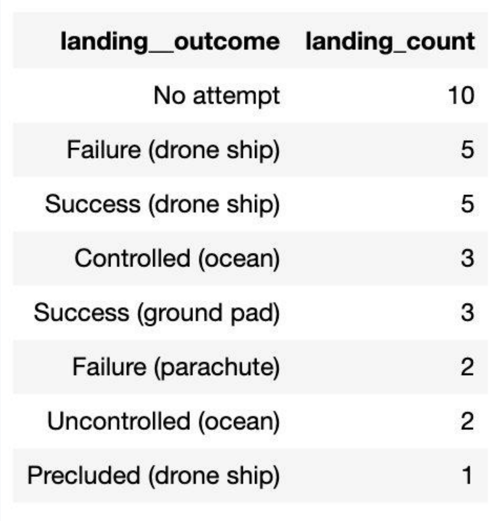
SQL (EDL with SQL)

- 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

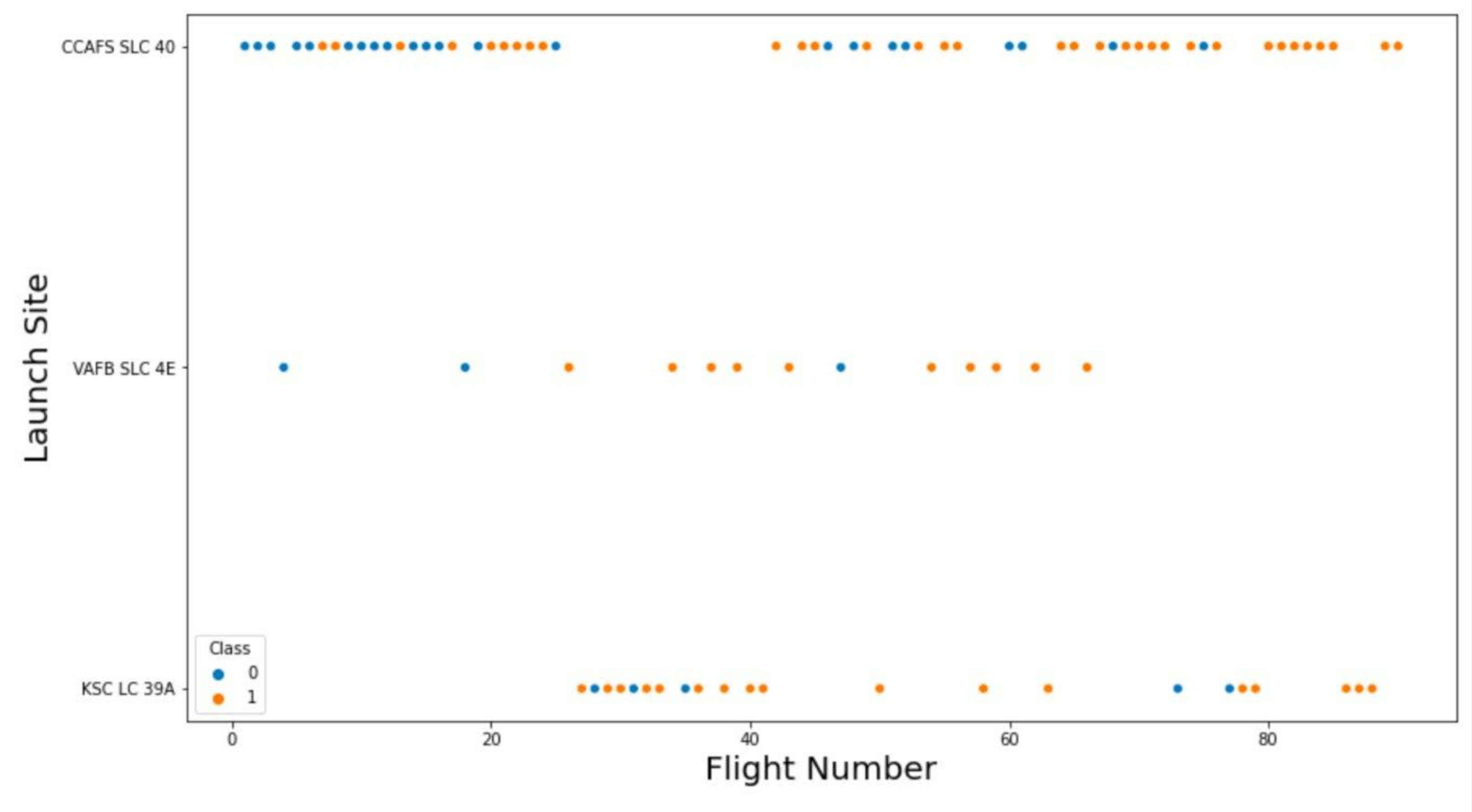
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order



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

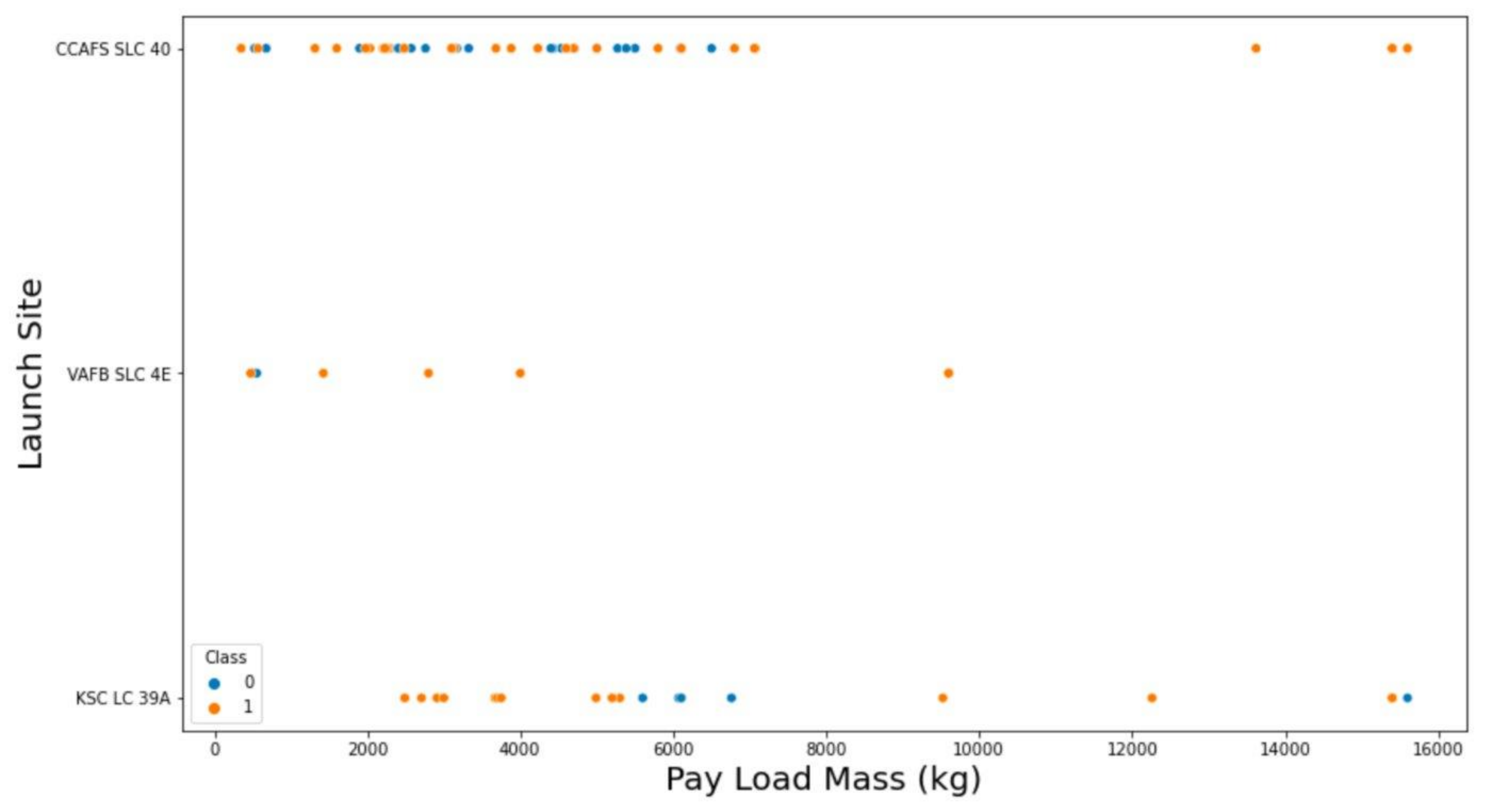
Matplotlib and Seaborn

Flight number vs launch site



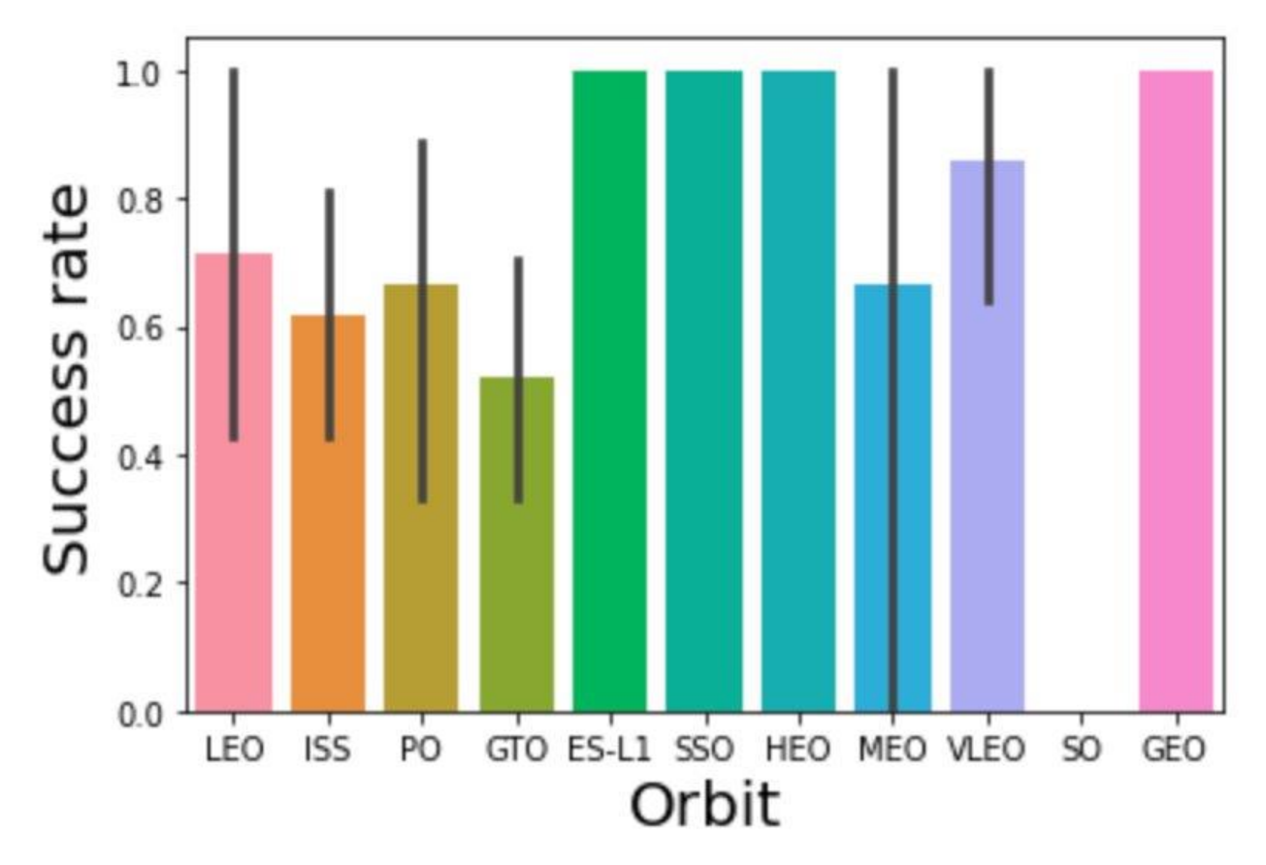
Matplotlib and Seaborn

Payload mass vs launch site



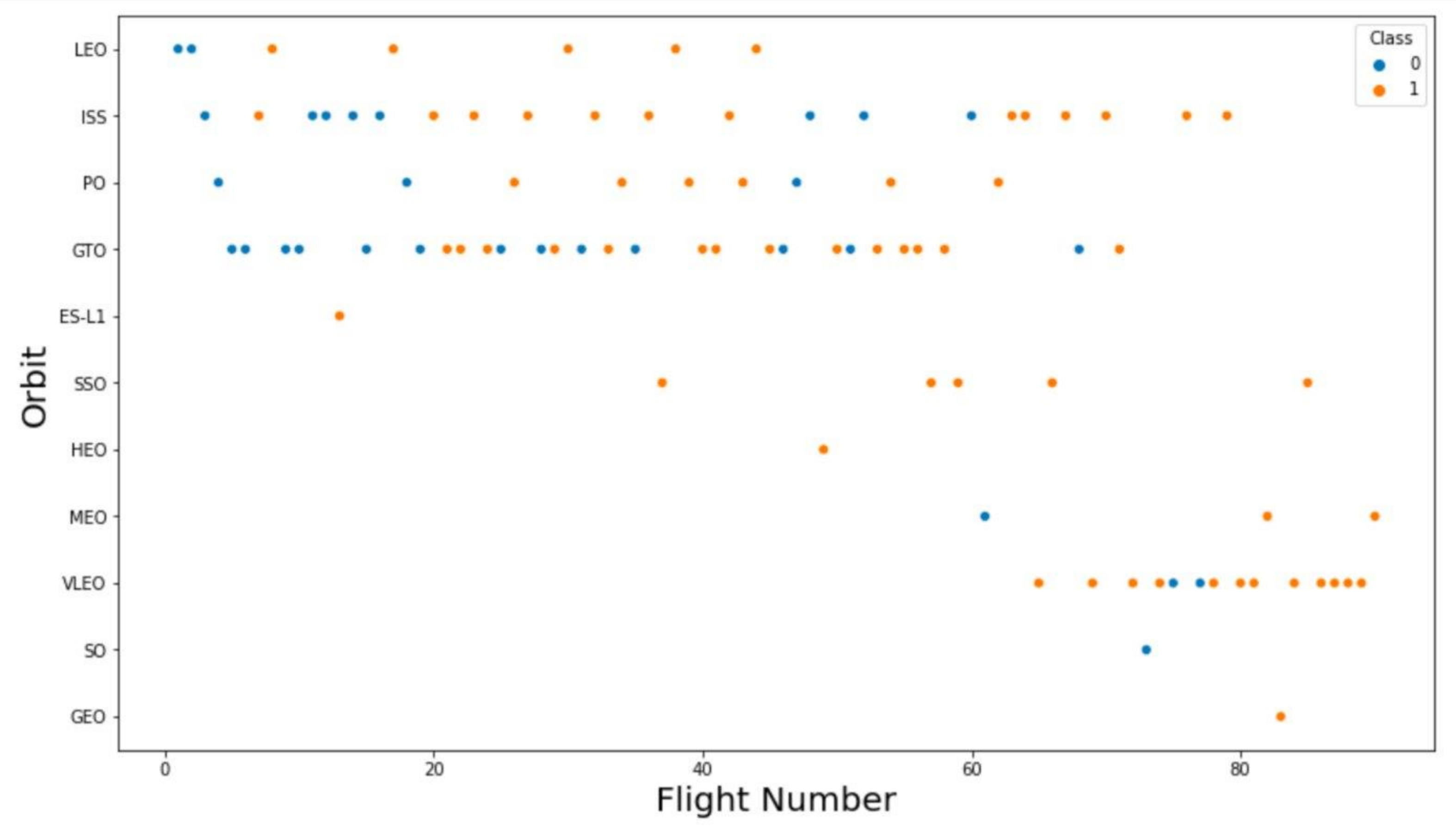
Matplotlib and Seaborn

success rate vs orbit type



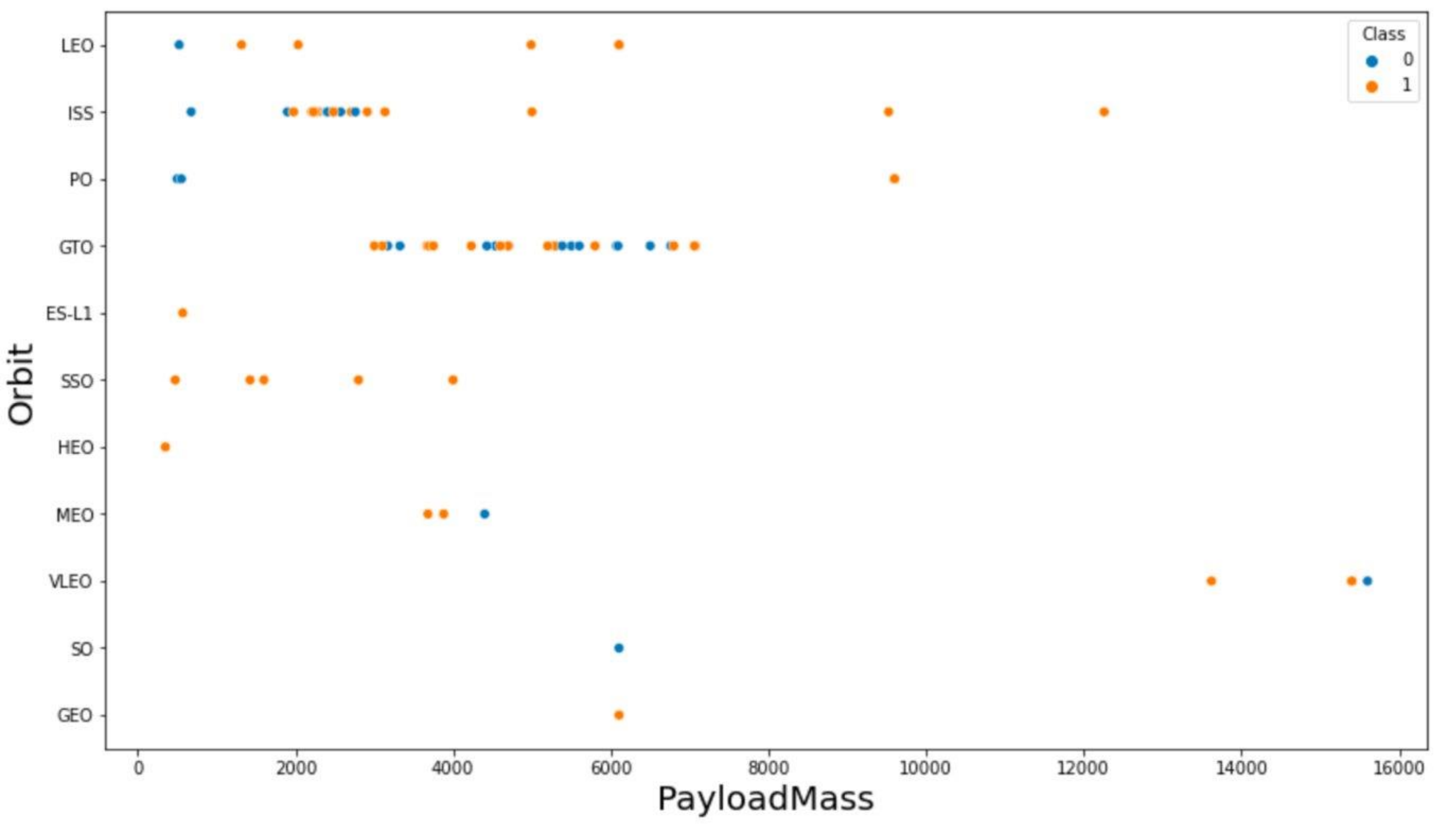
Matplotlib and Seaborn

Flight number vs orbit type



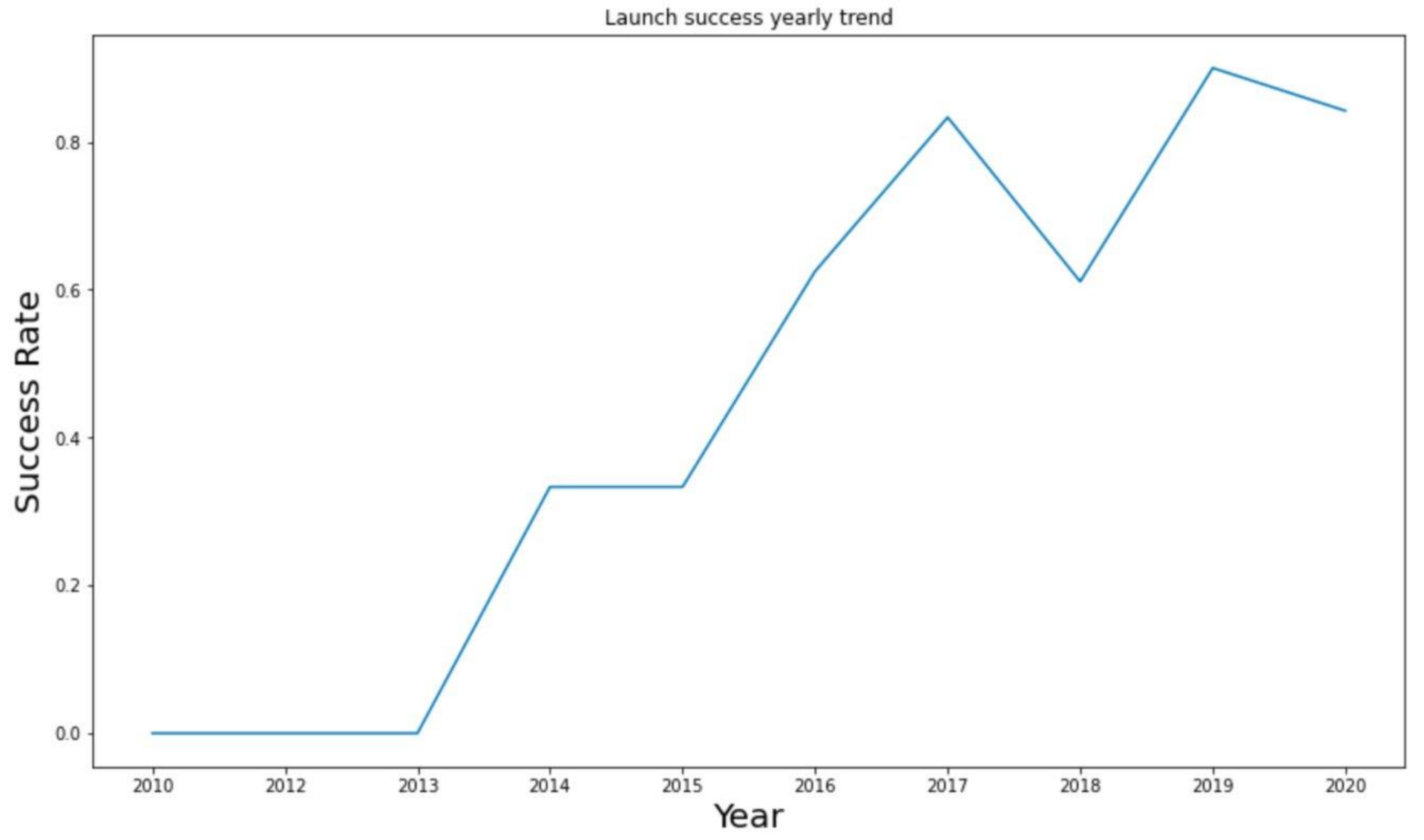
Matplotlib and Seaborn

payload mass vs orbit type



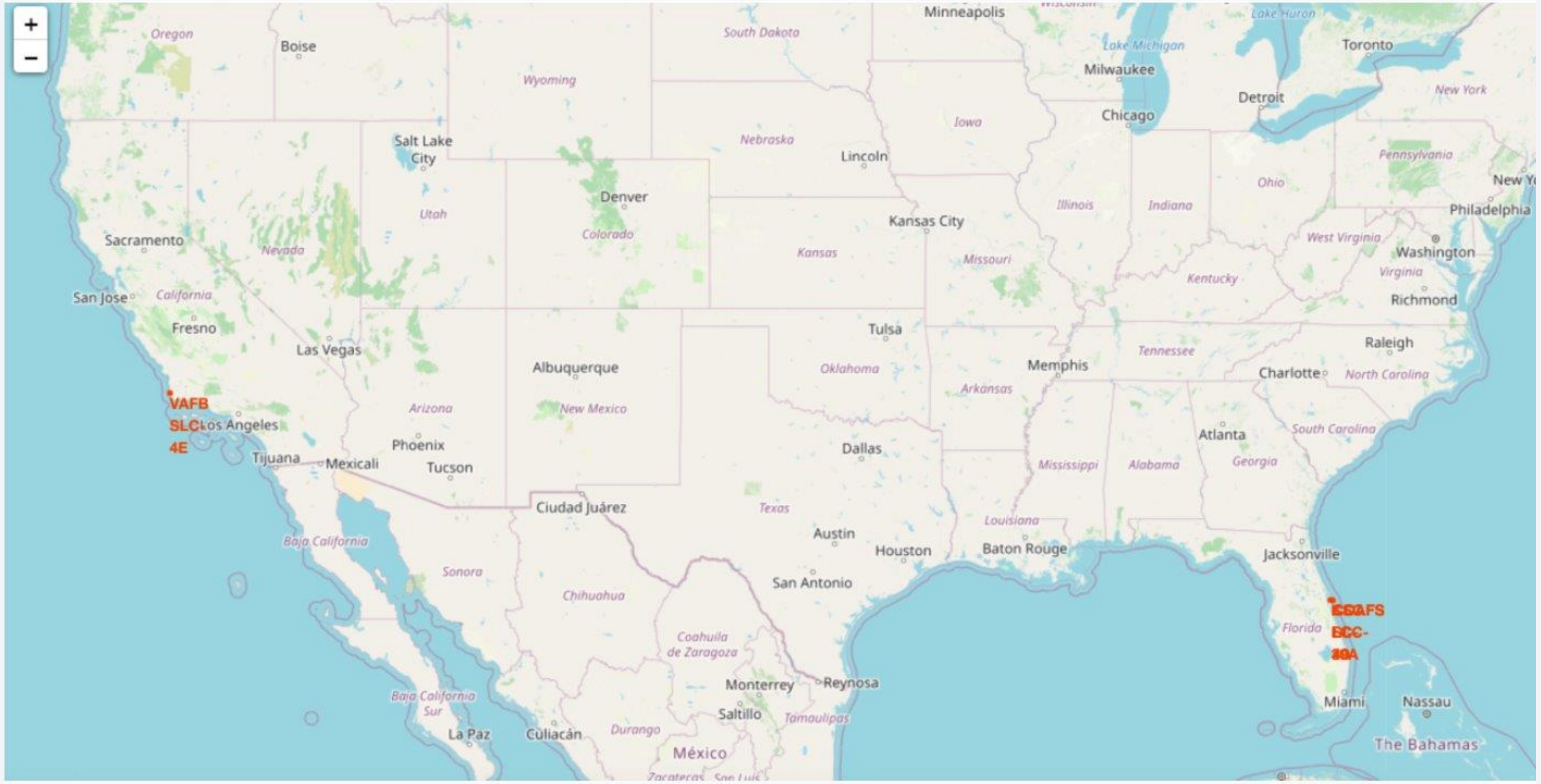
Matplotlib and Seaborn

Launch success yearly trend

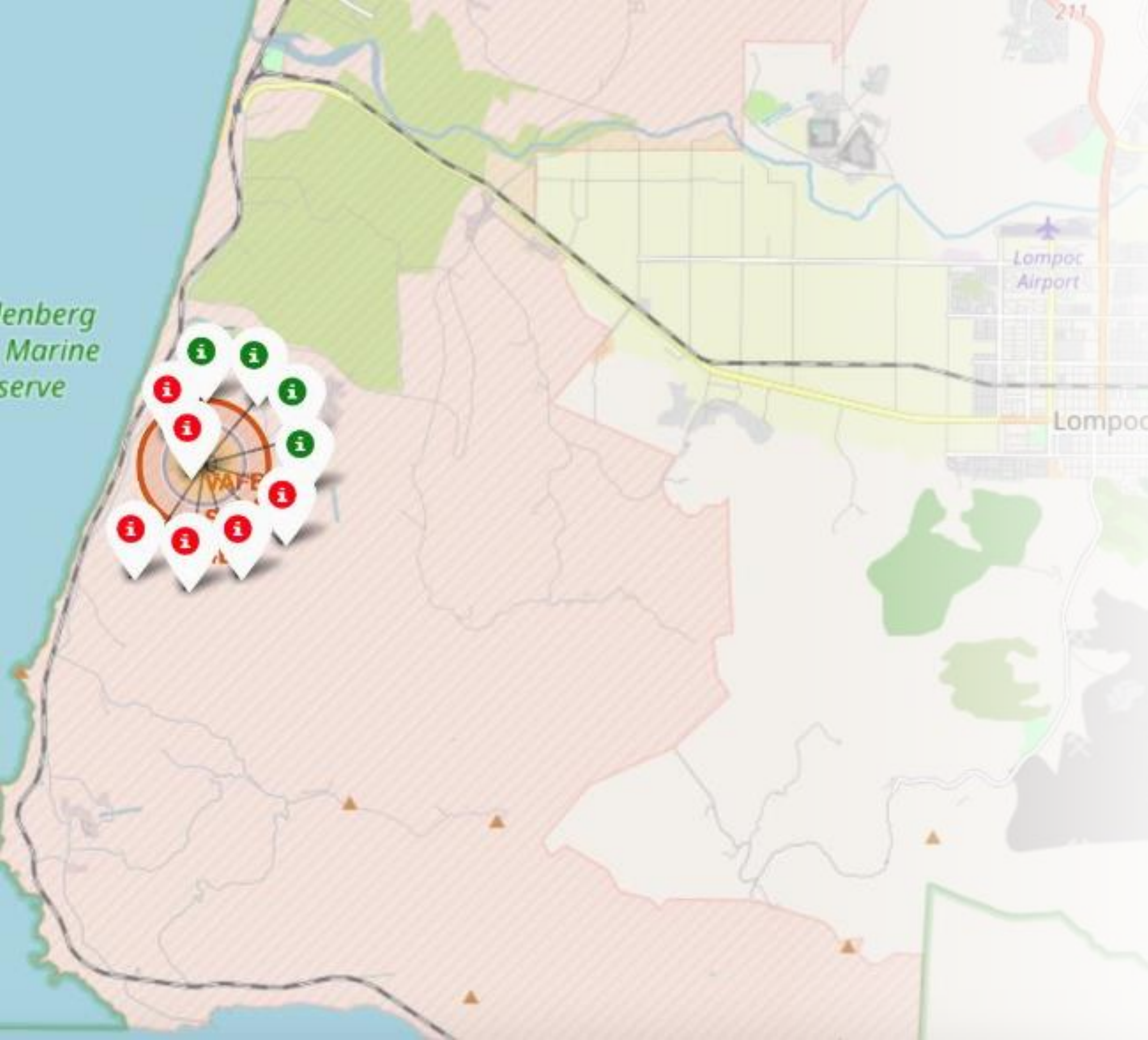


Folium

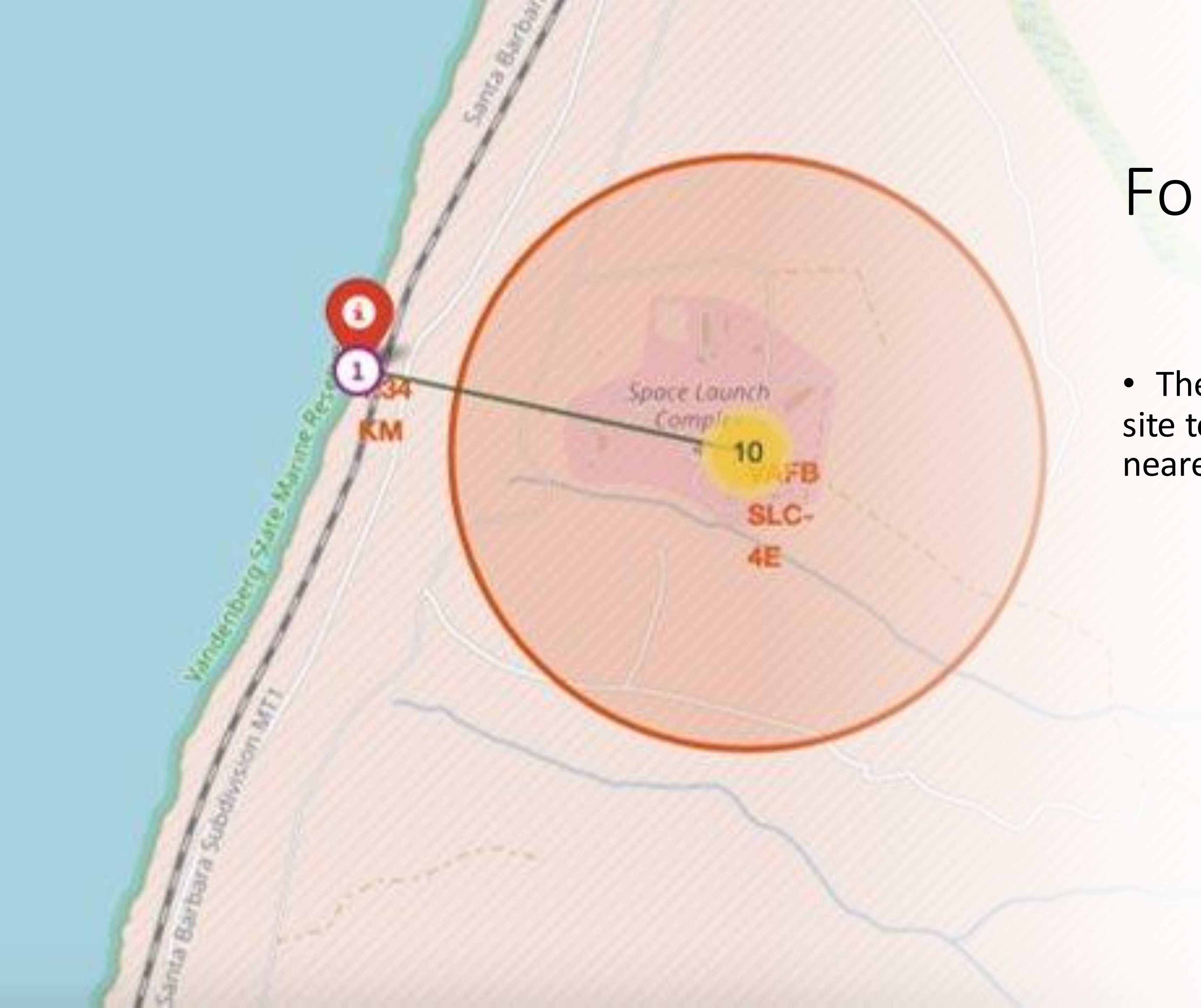
All launch sites



Folium



- The succeeded launches and failed launches for each site on map

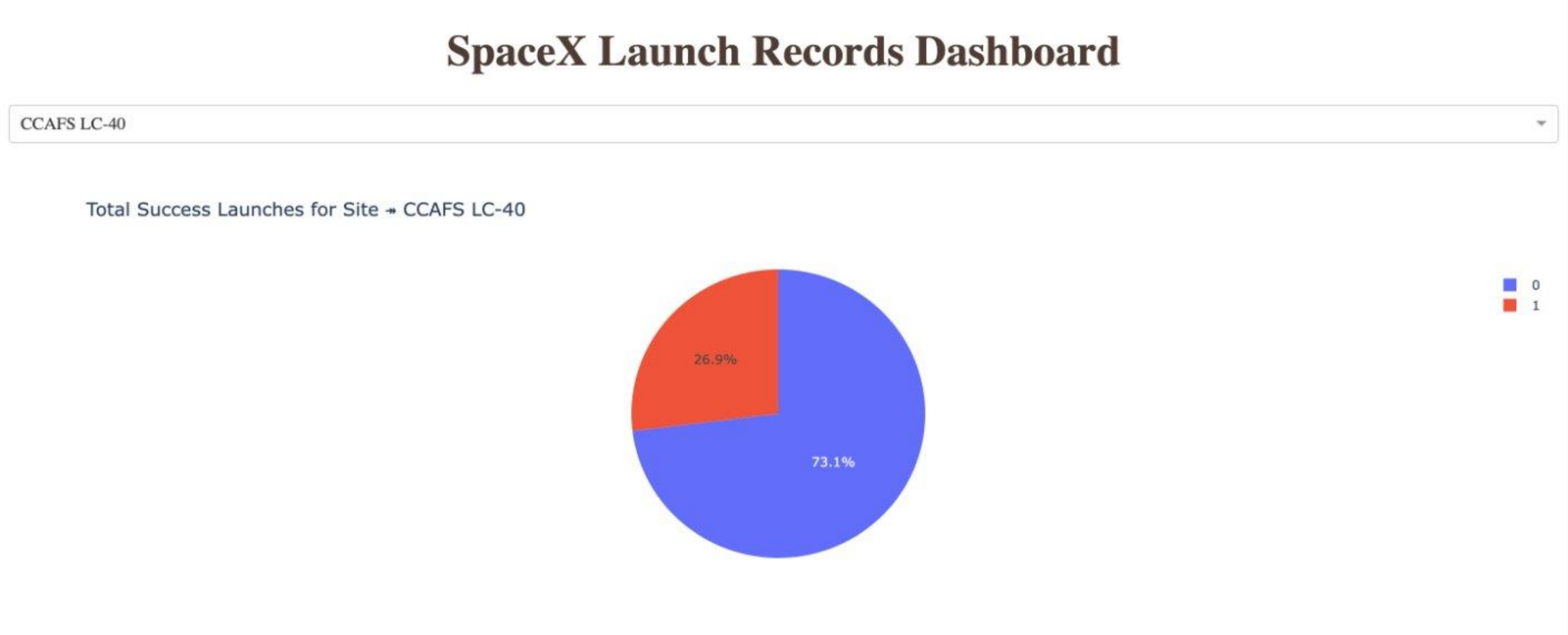


Folium

- The distances between a launch site to its proximities such as the nearest city, railway, or highway

Dash

A pie chart when launch site CCAFS LC-40 is chosen

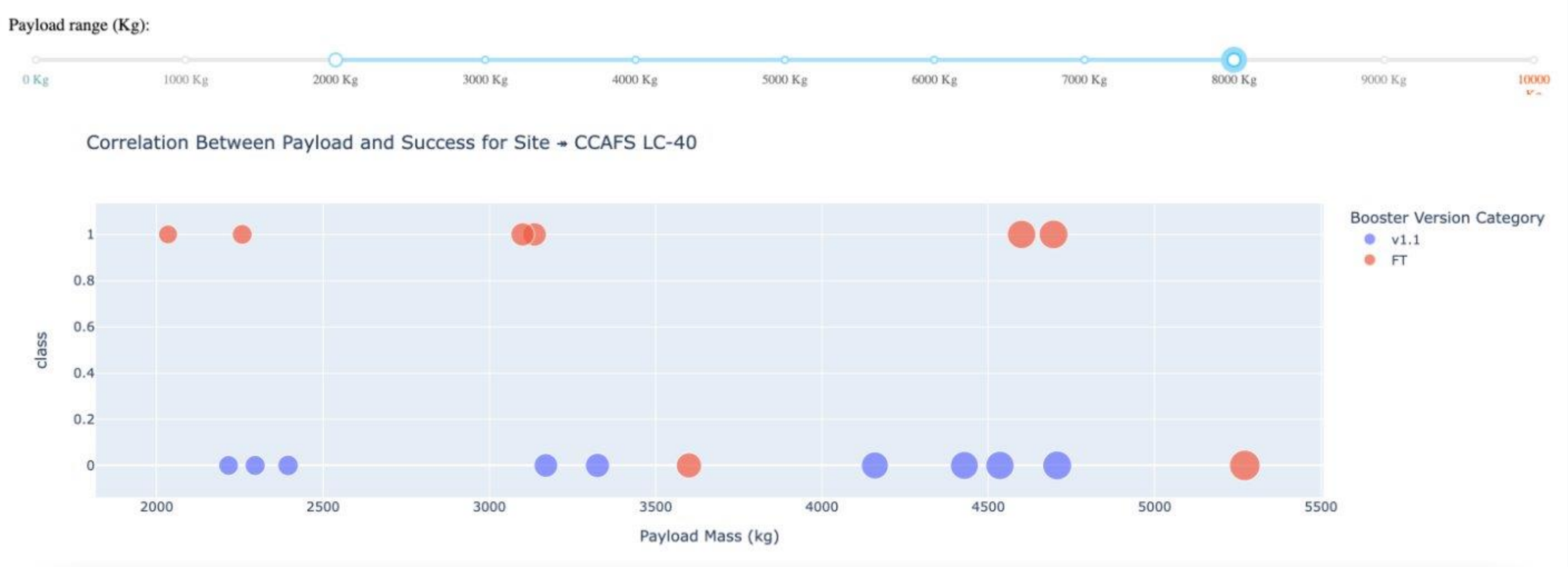


SpaceX Launch Records Dashboard

CCAFS LC-40

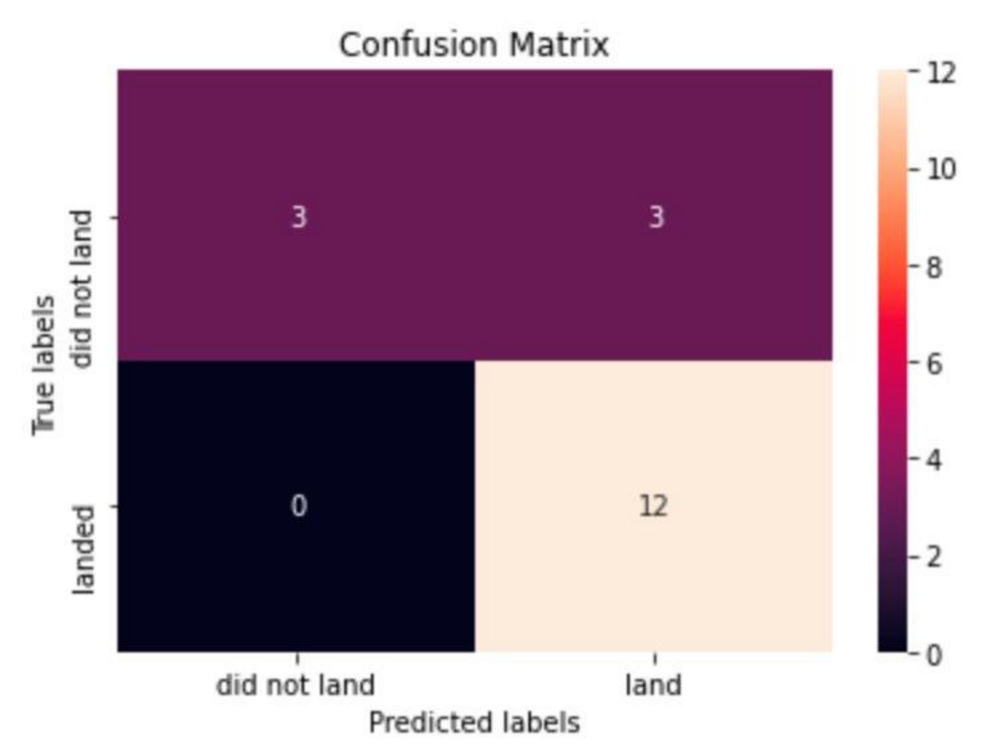
Dash

Scatterplot when the payload mass range is set to be from 2000kg to 8000kg



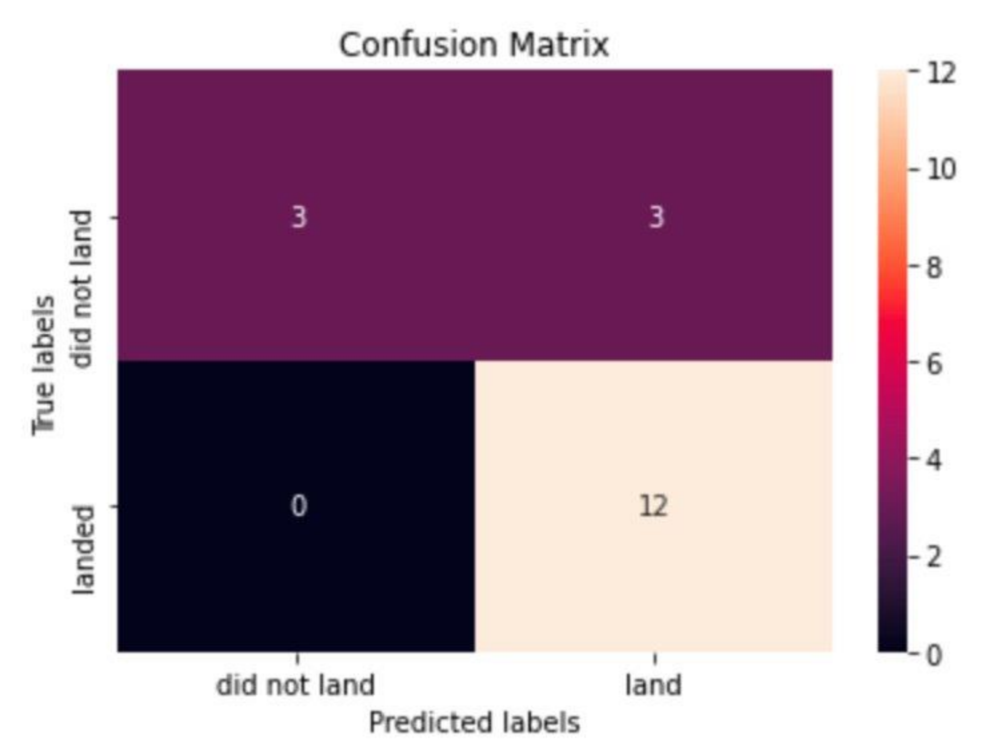
Predictive Analysis (Logistic Regression)

- GridSearchCV best score: 0.84643
- Accuracy score on test set: 0.834
- Confusion matrix:



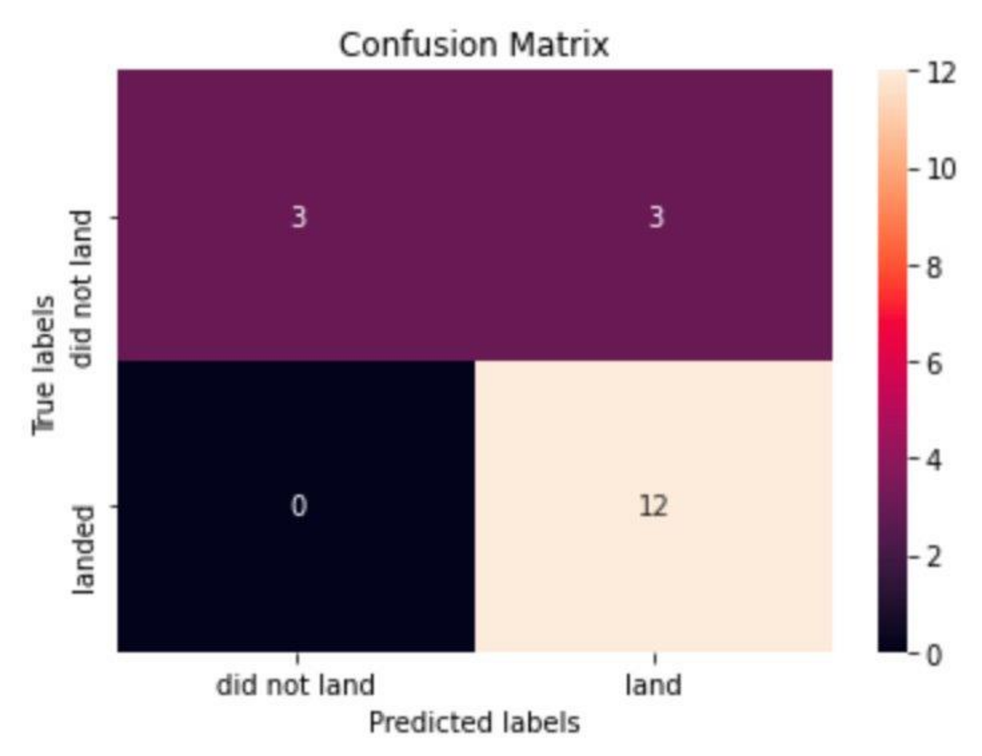
Predictive Analysis (Support Vector Machine)

- GridSearchCV best score: 0.84821
- Accuracy score on test set: 0.8334
- Confusion matrix:



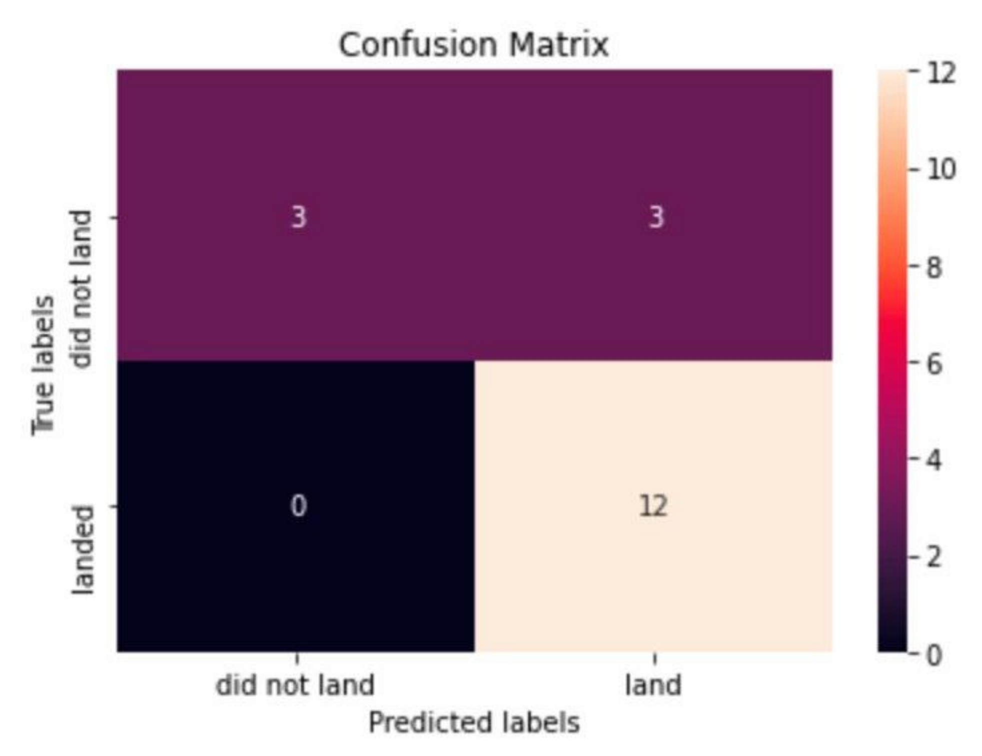
Predictive Analysis (Decision Tree)

- GridSearchCV best score: 0.8893
- Accuracy score on test set: 0.8334
- Confusion matrix:



Predictive Analysis (K nearest neighbors)

- GridSearchCV best score: 0.848214
- Accuracy score on test set: 0.8334
- Confusion matrix:



Predictive Analysis

- Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

1. Decision tree (best score: 0.8893)
2. K nearest neighbors (best score: 0.84821)
3. Support vector machine (best score: 0.84821)
4. Logistic regression (best score: 0.84643)

Conclusion

- In this project, we aim to predict whether the Falcon 9 rocket's first stage will successfully land so we can estimate the launch cost more accurately.
- Different launch features—like payload mass and orbit type—can influence whether the mission succeeds or fails.
- We use several machine learning algorithms to analyze patterns in past Falcon 9 launches and build models that can predict future outcomes.
- Among the four algorithms tested, the decision tree model provided the most accurate predictions.

Thank you!