IBM

Applied Data Science capstone

Chris Konstantinidis

September 21, 2024

TABLE OF CONTENTS

01 EXECUTIVE SUMMARY

02 INTRODUCTION

03 METHODOLOGY

04 RESULTS

05 DISCUSSION

06 CONCLUSION

EXECUTIVE SUMMARY

This project focused on forecasting the outcome of SpaceX Falcon 9 first stage landings by leveraging machine learning classification techniques. The process encompassed data acquisition, preprocessing, exploratory analysis, and developing interactive visualizations, followed by model training and evaluation.

The analysis revealed that certain launch characteristics were linked to the success or failure of the landings. After testing a range of algorithms, the decision tree model stood out as particularly effective in forecasting the outcome of the Falcon 9 first stage landings.

INTRODUCTION

This project explores the use of machine learning models to predict whether the Falcon 9 first stage will successfully land after launch. Successful landings are key to SpaceX’s cost efficiency, as the reusability of the rocket's first stage significantly reduces launch expenses. By analyzing various features such as payload mass, orbit type, and launch site, we aim to identify patterns that influence landing outcomes.

While some landings are intentionally planned for ocean recovery, understanding the factors that contribute to both successful and deliberate unsuccessful landings offers valuable insights for future launches. The results of this project could provide useful guidance for companies seeking to optimize their own launch strategies or compete in the commercial space industry.

METHODOLOGY

The overall methodology includes:

1 **Data Collection, Wrangling, and Formatting:**

- SpaceX API
- Web scraping

2 **Exploratory Data Analysis (EDA):**

- Pandas and NumPy
- SQL

3 **Data Visualization:**

- Matplotlib and Seaborn
- Folium
- Dash

4 **Machine Learning Prediction:**

- Logistic regression
- Support vector machine (SVM)
- Decision tree
- K-nearest neighbors (KNN)

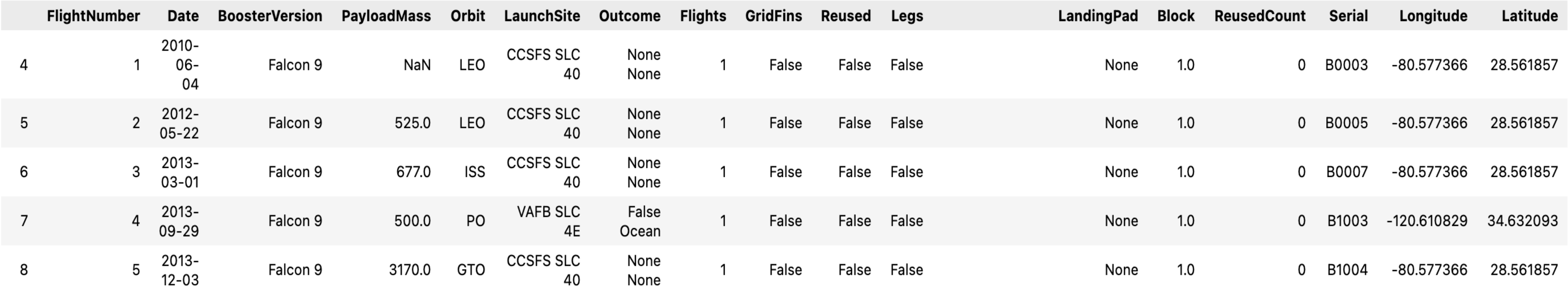
METHODOLOGY

A Data Collection, Wrangling and Formatting

SpaceX API

The SpaceX API utilized for this project is found at https://api.spacexdata.com/v4/rockets/. It provides detailed information on various SpaceX rocket launches. For the purposes of this analysis, we filtered the data to include only Falcon 9 launches.

To handle missing values, we applied a simple imputation method, replacing any missing entries with the mean value of the corresponding column. After preprocessing, the dataset consists of 90 instances and 17 features. The following image displays a snapshot of the first few rows of the dataset:



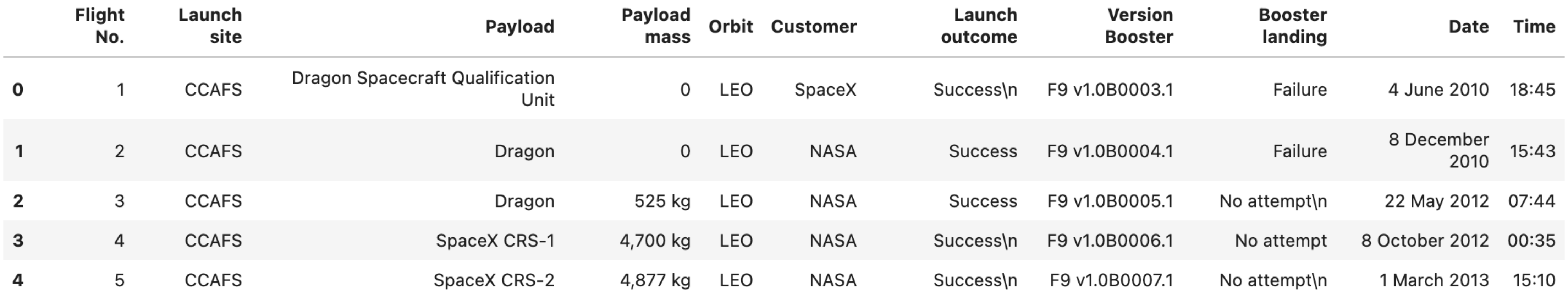
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	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

METHODOLOGY

A Data Collection, Wrangling and Formatting

Web scraping

Data for this project was also collected through web scraping from this Wikipedia page, which provides information exclusively about Falcon 9 launches. After cleaning and organizing the data, we obtained 121 instances with 11 features. Below is a preview of the first few rows of the scraped 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

The data was further processed to ensure there were no missing values, and categorical features were transformed using one-hot encoding. Additionally, a new column labeled ‘Class’ was introduced, indicating the outcome of each launch: 0 for a failure and 1 for a success. After this preprocessing step, the dataset expanded to 90 instances and 83 features.

METHODOLOGY

B Exploratory Data Analysis (EDA)

- **Pandas and NumPy**

  Functions from the Pandas and NumPy libraries were employed to analyze the data and extract basic information. This included determining the number of launches at each site, identifying the frequency of different orbit types, and assessing the occurrence of various mission outcomes.

- **SQL**

  SQL queries were also used to explore the dataset and answer specific questions. These included finding the names of unique launch sites, calculating the total payload mass of boosters launched by NASA (CRS), and determining the average payload mass carried by the F9 v1.1 booster version.

METHODOLOGY

C Data Visualization

- **Matplotlib and Seaborn**
  We utilized functions from the Matplotlib and Seaborn libraries to create various visualizations, including scatterplots, bar charts, and line charts. These plots helped us examine the relationships between key features, such as the correlation between flight number and launch site, the connection between payload mass and launch site, and the success rate across different orbit types.
- **Folium**
  The Folium library was employed to create interactive maps for visualizing the data. It enabled us to mark all launch sites, differentiate between successful and failed launches at each site, and highlight the distances from launch sites to nearby landmarks, including the closest cities, railways, and highways.
- **Dash**
  Functions from Dash were used to develop an interactive web application that allows users to explore the data dynamically through a dropdown menu and a range slider. The site features a pie chart displaying the total successful launches from each launch site, as well as a scatterplot illustrating the correlation between payload mass and mission outcome (success or failure) for each launch site.

METHODOLOGY

D Machine Learning Prediction

Scikit-learn

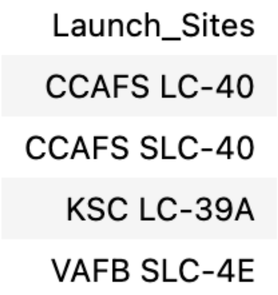
We leveraged functions from the Scikit-learn library to develop our machine learning models. The prediction process involved several key steps. First, we standardized the data to ensure consistent scaling. Next, we split the dataset into training and testing subsets.

We then created several machine learning models, including logistic regression, support vector machine (SVM), decision tree, and K nearest neighbors (KNN). After fitting these models to the training data, we optimized their performance by identifying the best combination of hyperparameters for each model. Finally, we evaluated their effectiveness using accuracy scores and confusion matrices to assess their predictive capabilities.

RESULTS

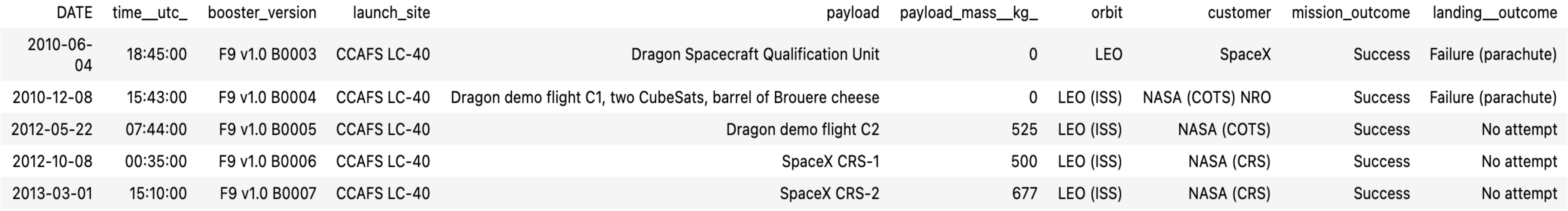
A SQL (EDA with SQL)

- 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’

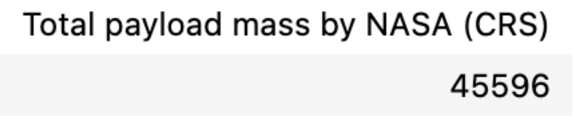


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

RESULTS

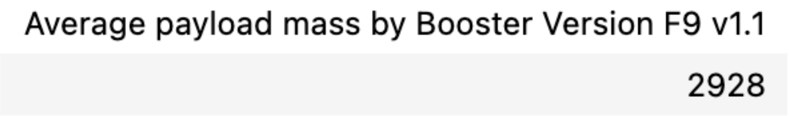
A SQL (EDA with SQL)

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



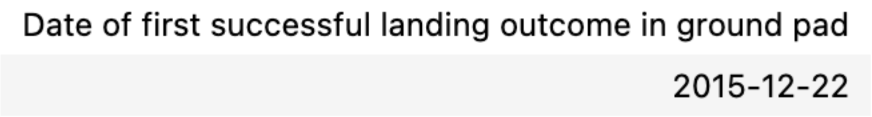
Total payload mass by NASA (CRS)
45596

- 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

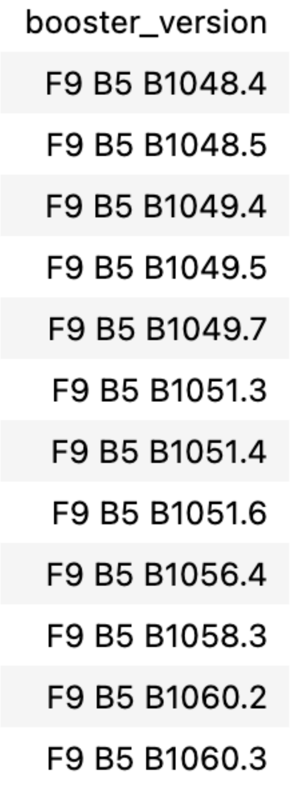


Date of first successful landing outcome in ground pad
2015-12-22

RESULTS

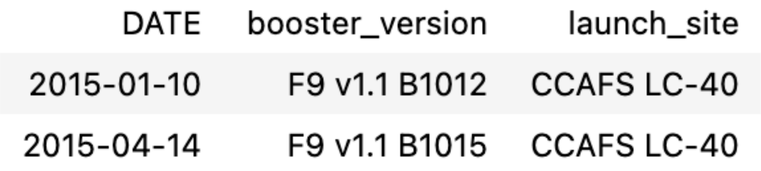
A SQL (EDA with SQL)

- 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

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

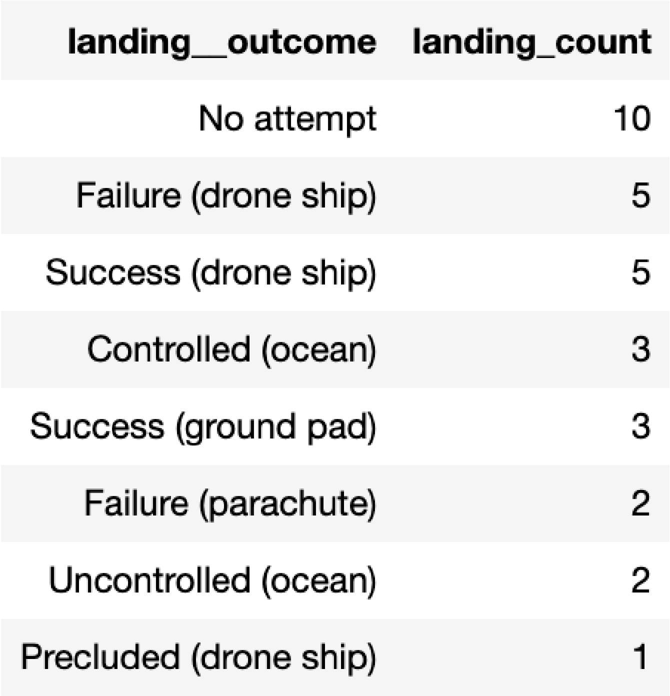


DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

RESULTS

A SQL (EDA with SQL)

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

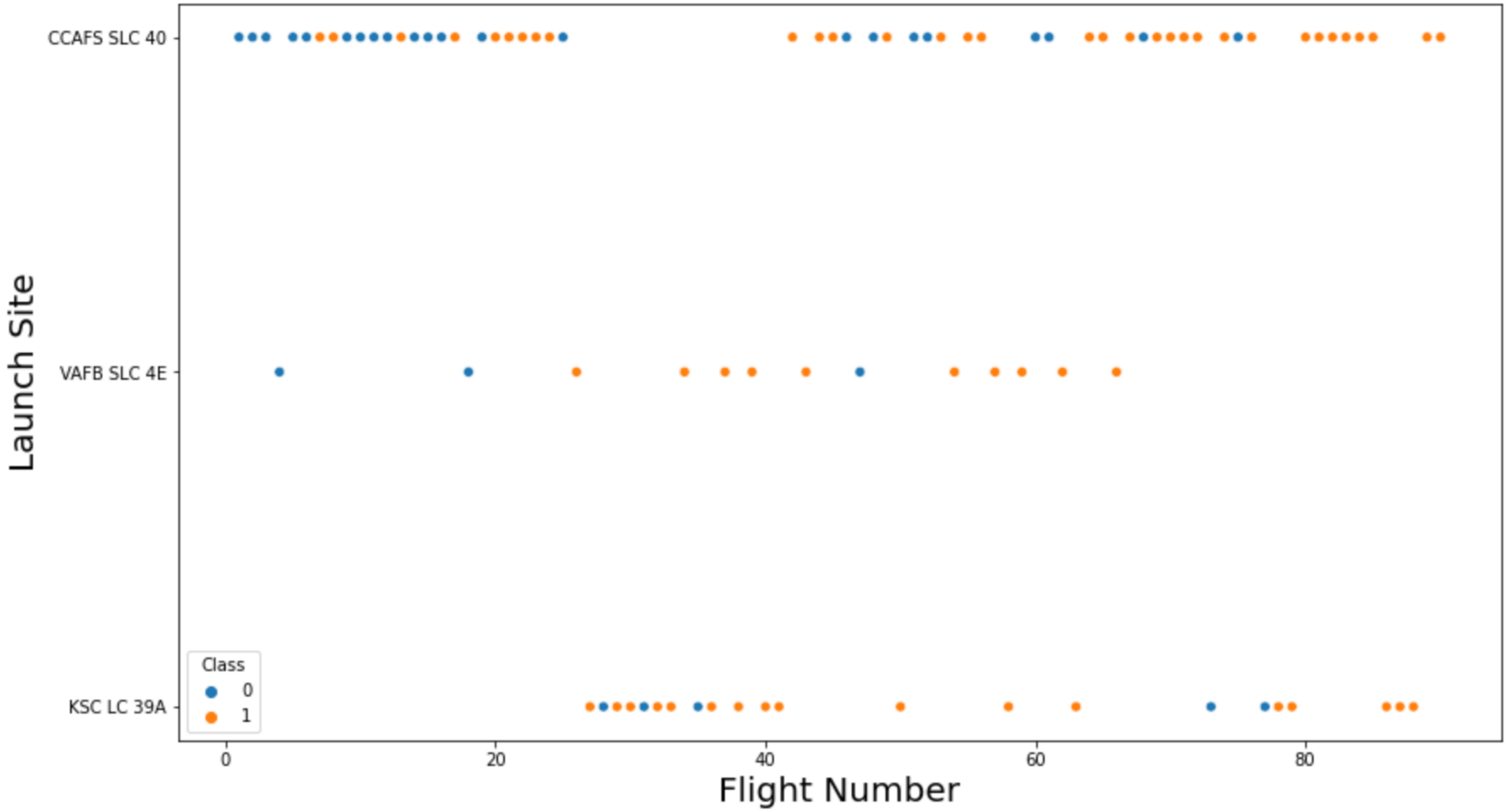


landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

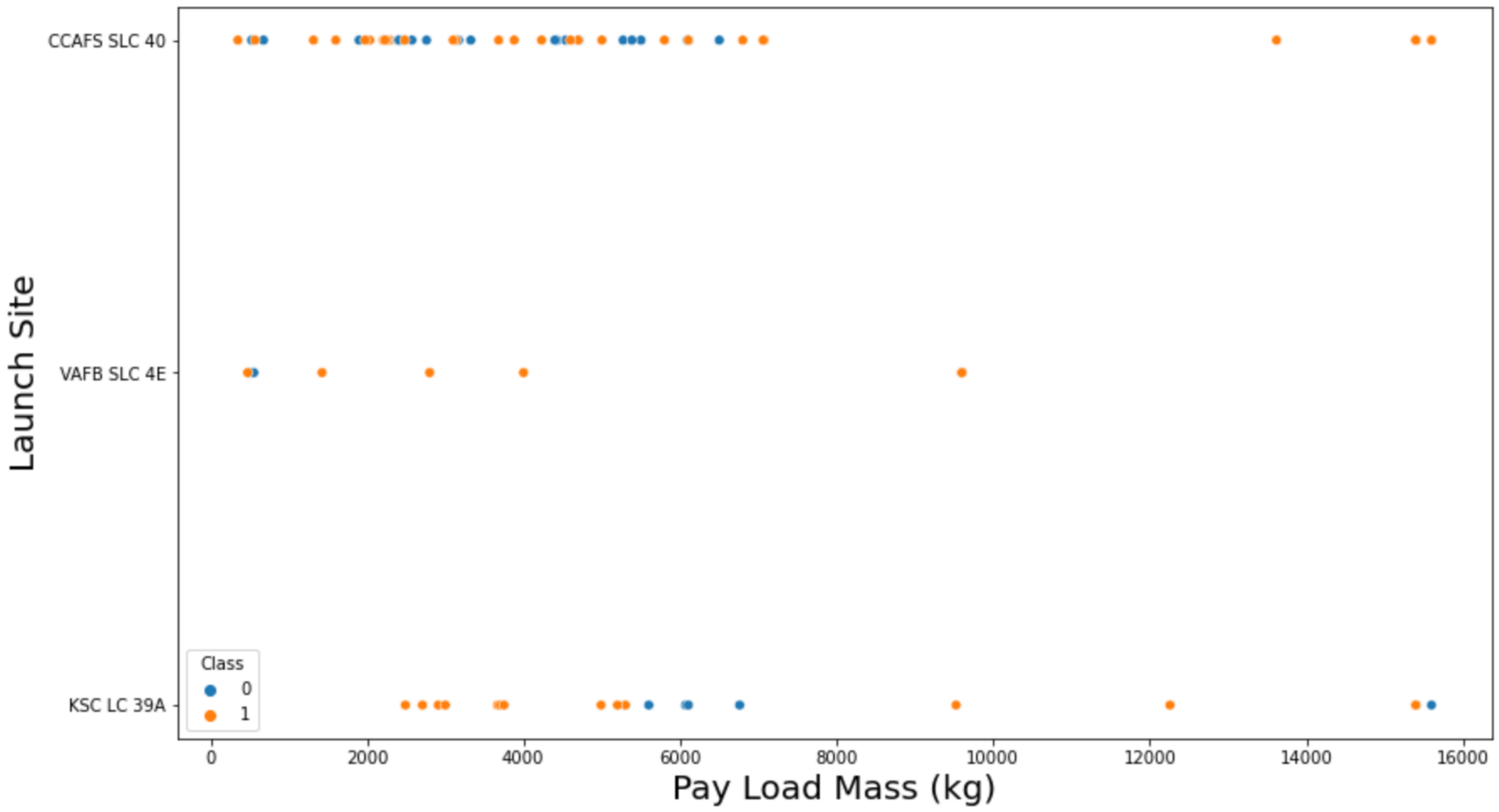
- The relationship between flight number and launch site



RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

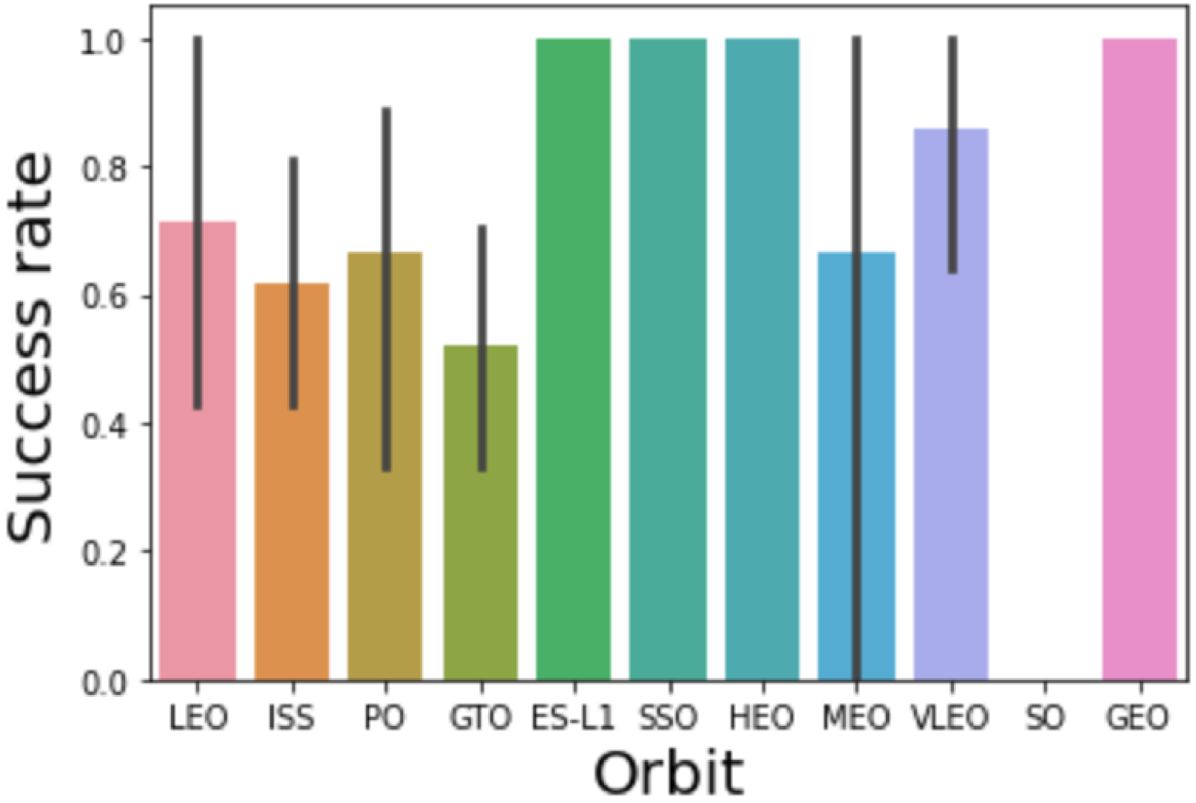
- The relationship between payload mass and launch site



RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

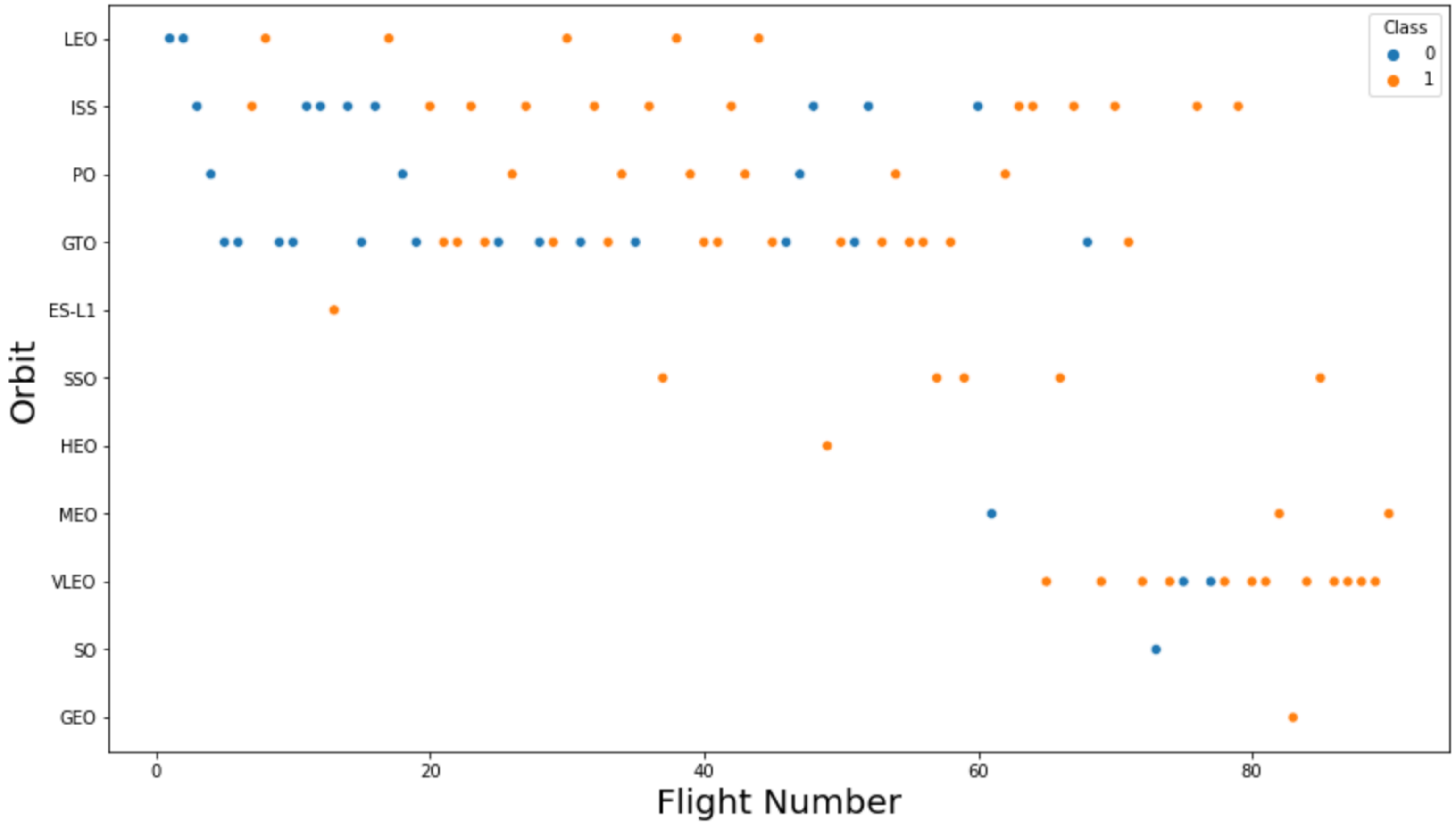
- The relationship between success rate and orbit type



RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

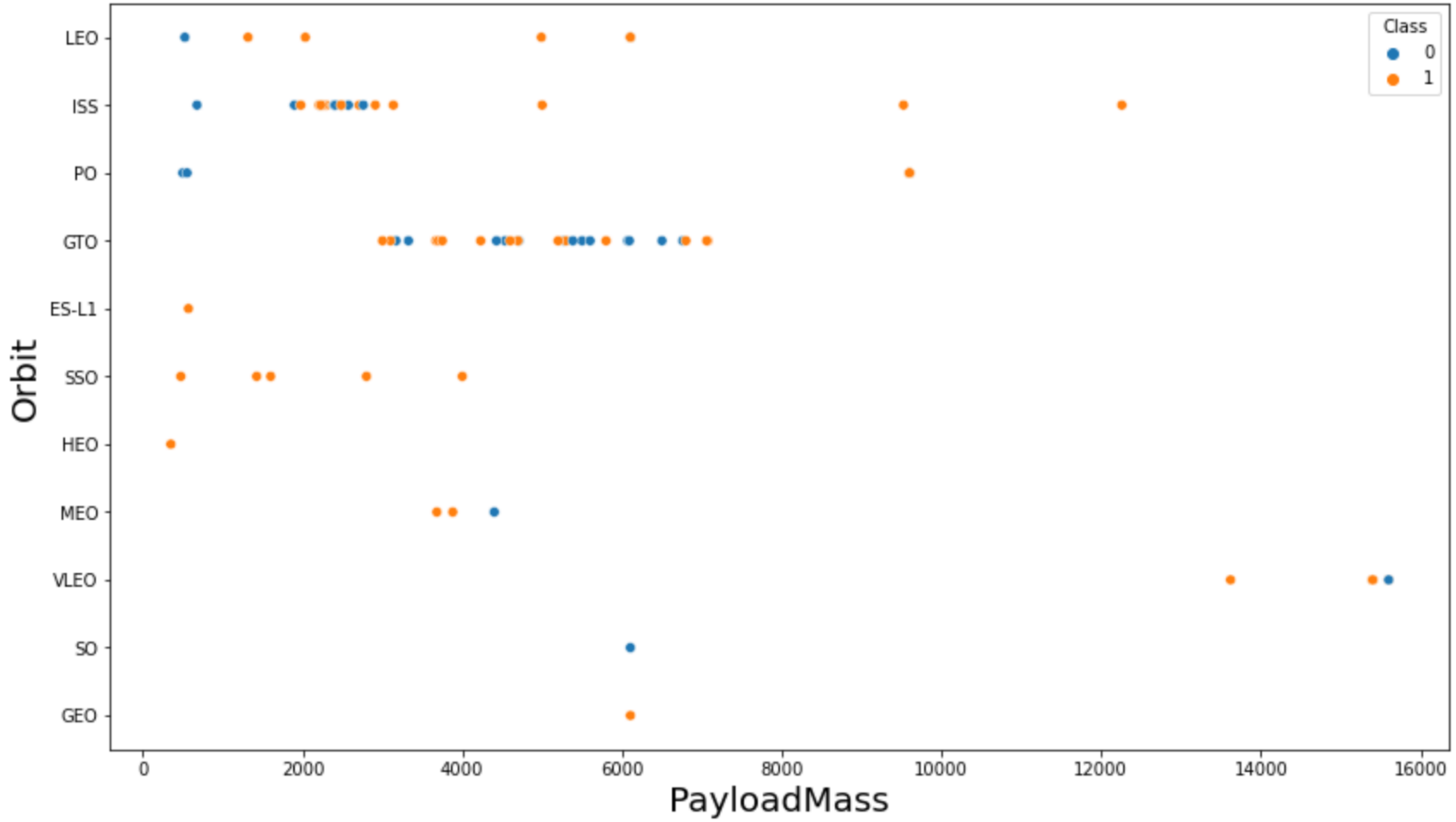
- The relationship between flight number and orbit type



RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

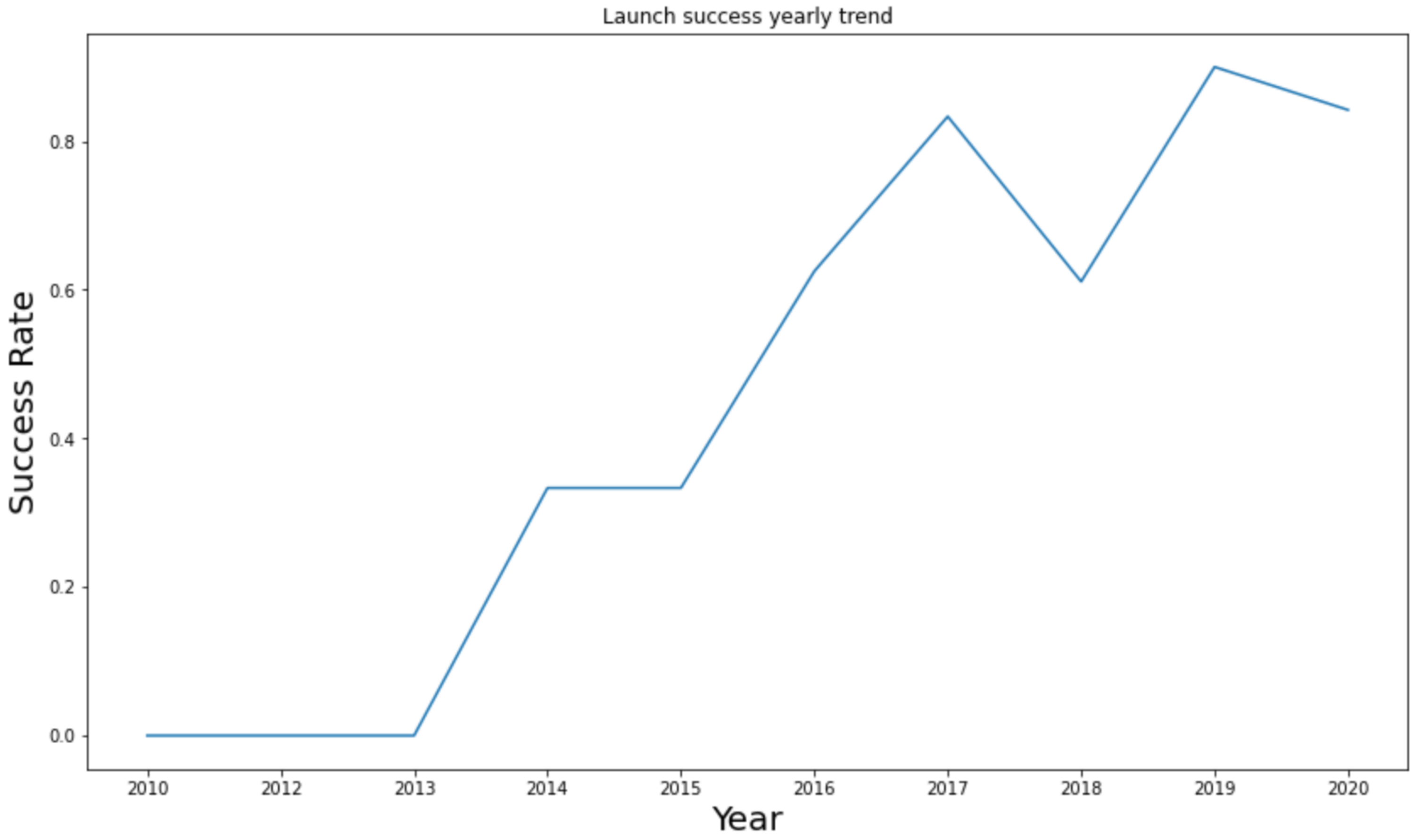
- The relationship between payload mass and orbit type



RESULTS

B Matplotlib and Seaborn (EDA with Visualization)

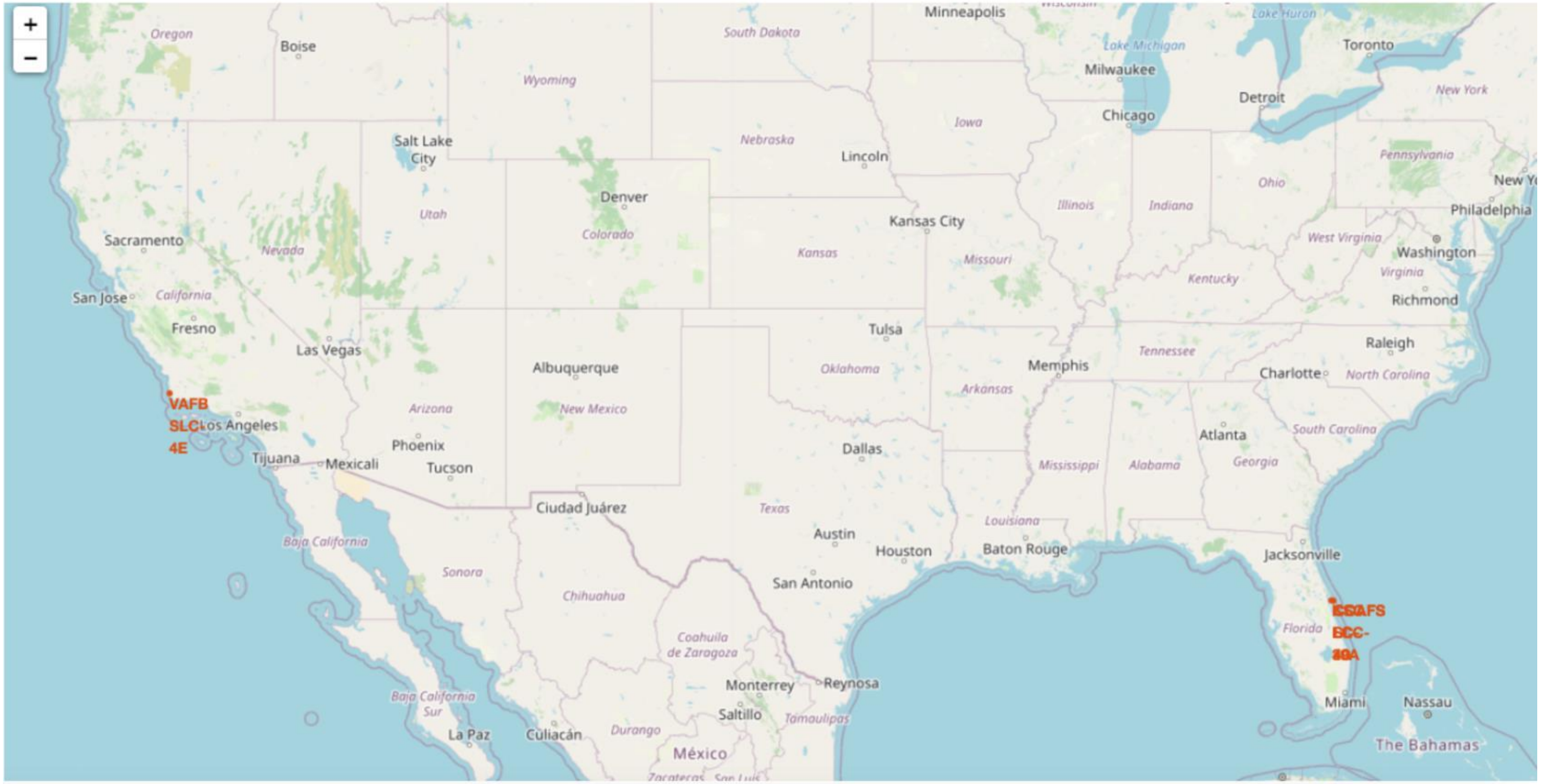
- The launch success yearly trend



RESULTS

C Folium

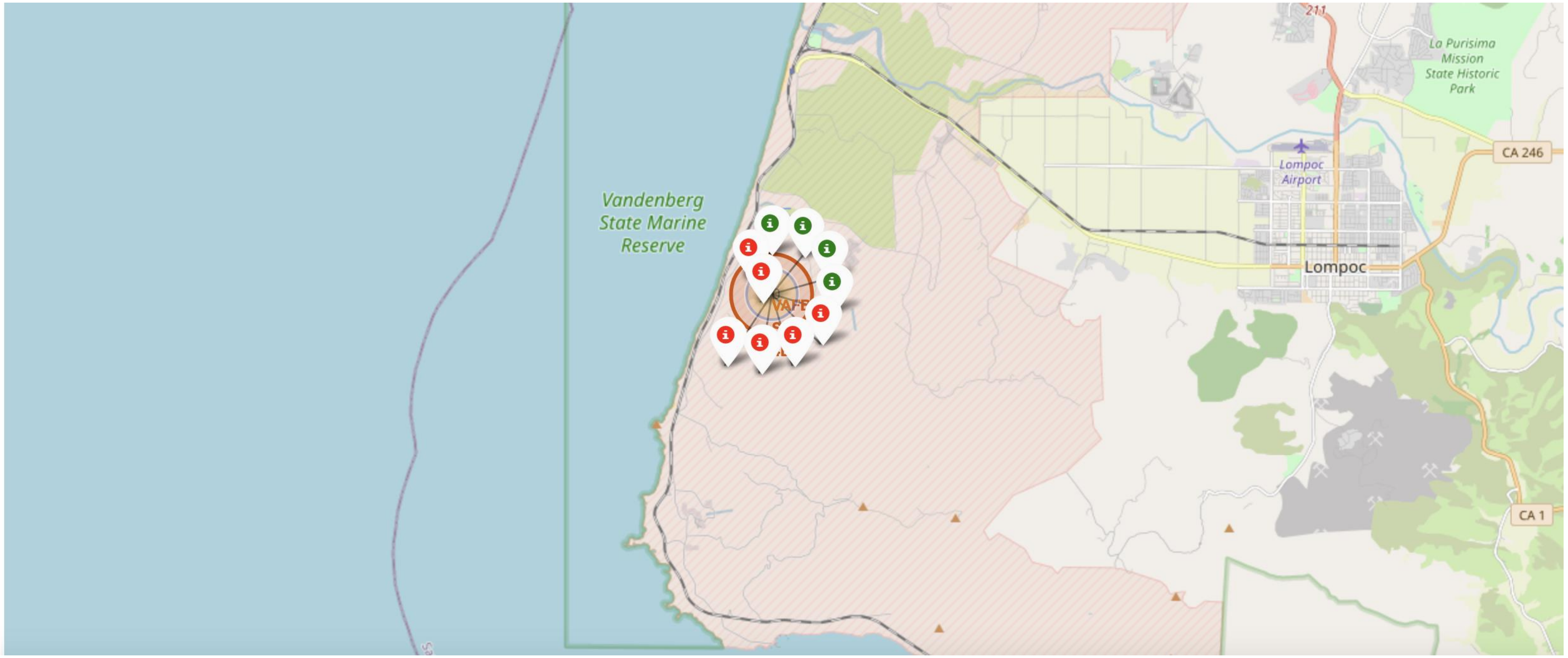
- All launch sites on map



RESULTS

C Folium

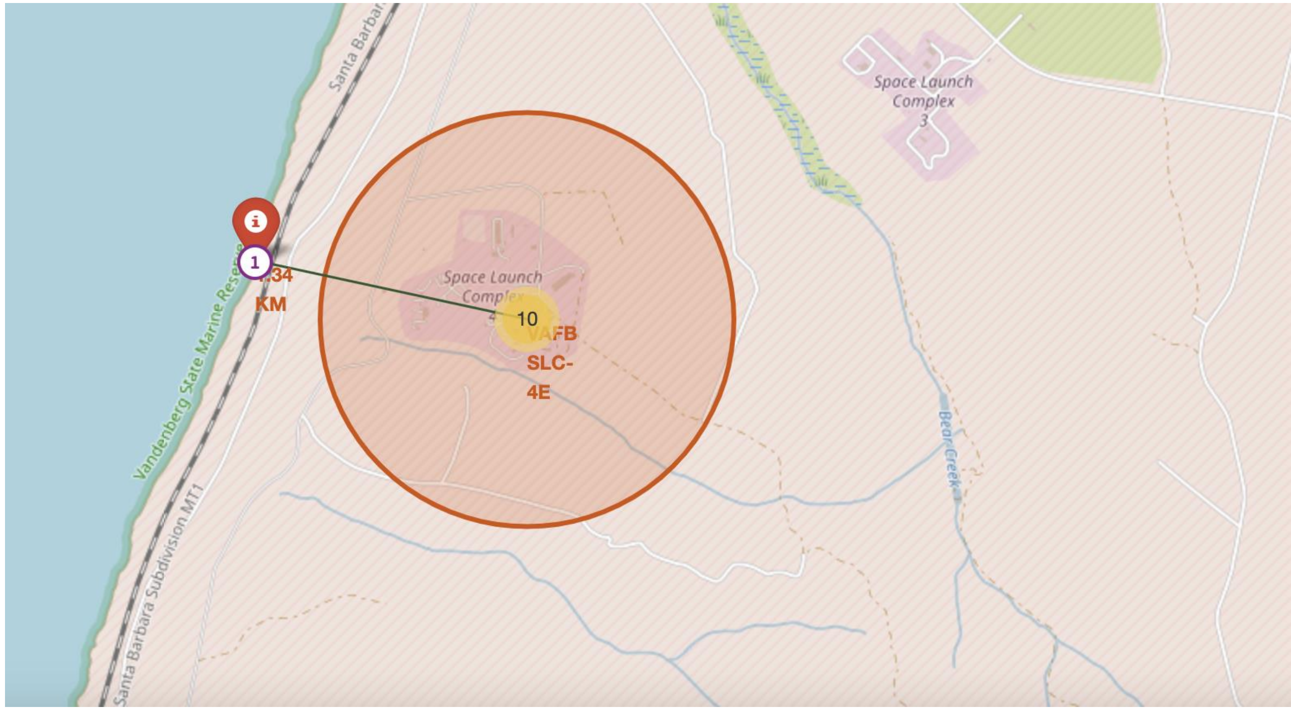
- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch


Vandenberg State Marine Reserve
La Purisima Mission State Historic Park
Lompoc Airport
Lompoc
CA 246
CA 1
211
VAFB

RESULTS

C Folium

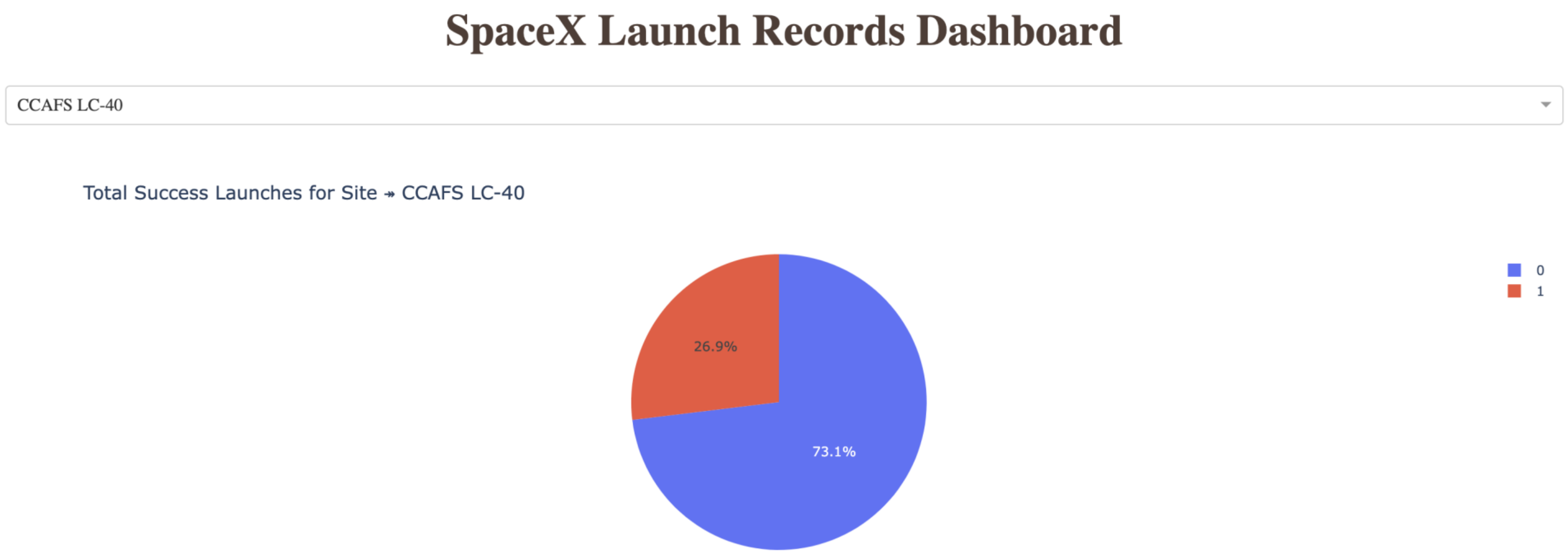
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



RESULTS

D Dash

- 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.



SpaceX Launch Records Dashboard

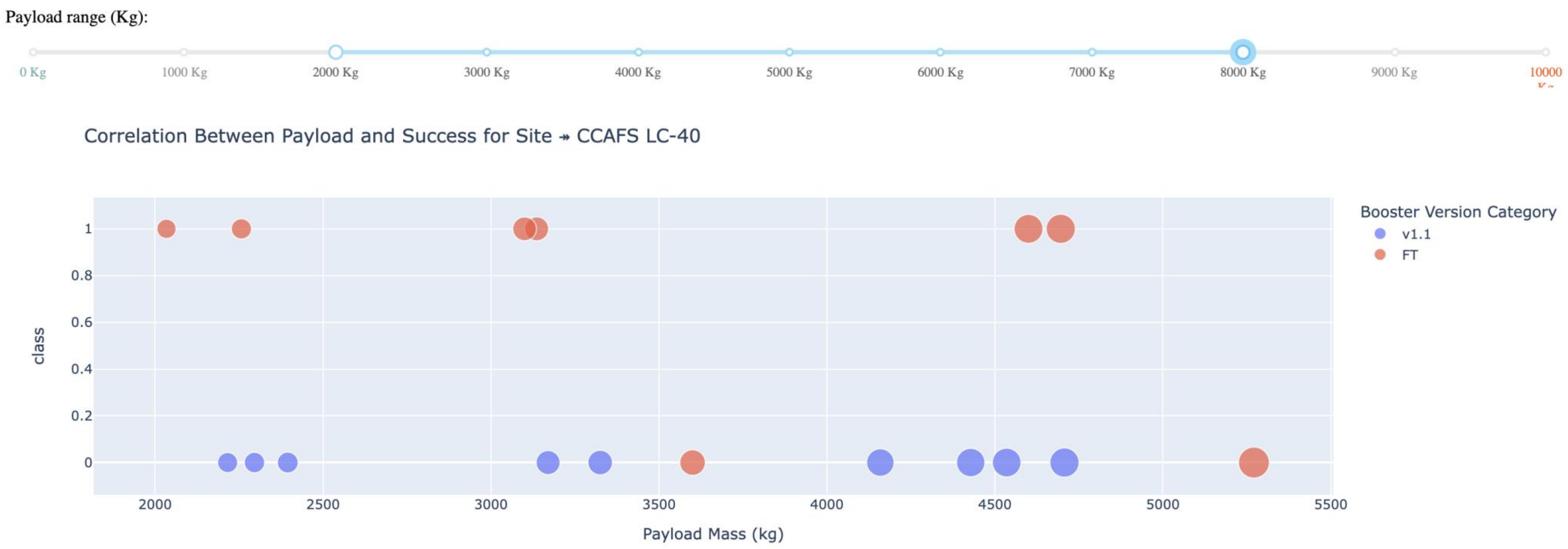
CCAFS LC-40

Total Success Launches for Site → CCAFS LC-40

RESULTS

D Dash

- 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.



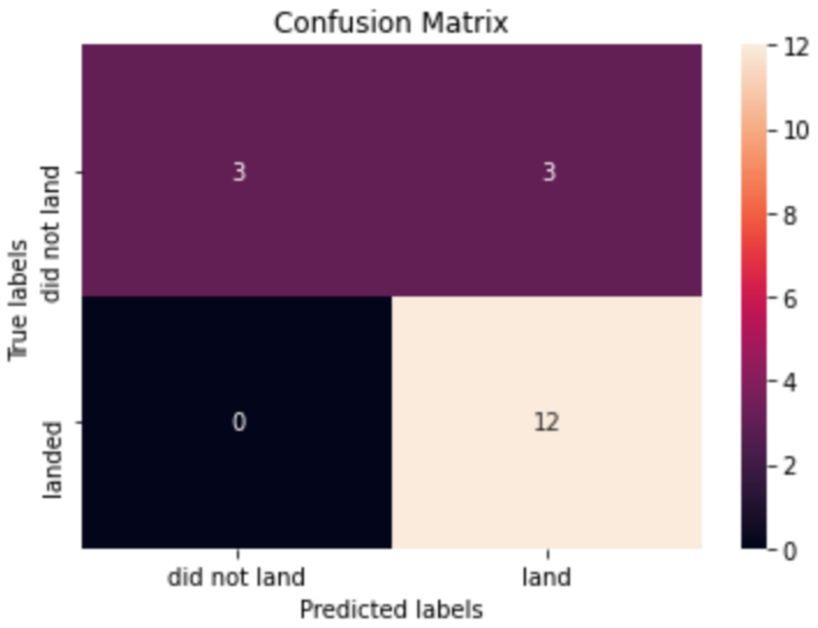
Payload range (Kg):

Correlation Between Payload and Success for Site → CCAFS LC-40

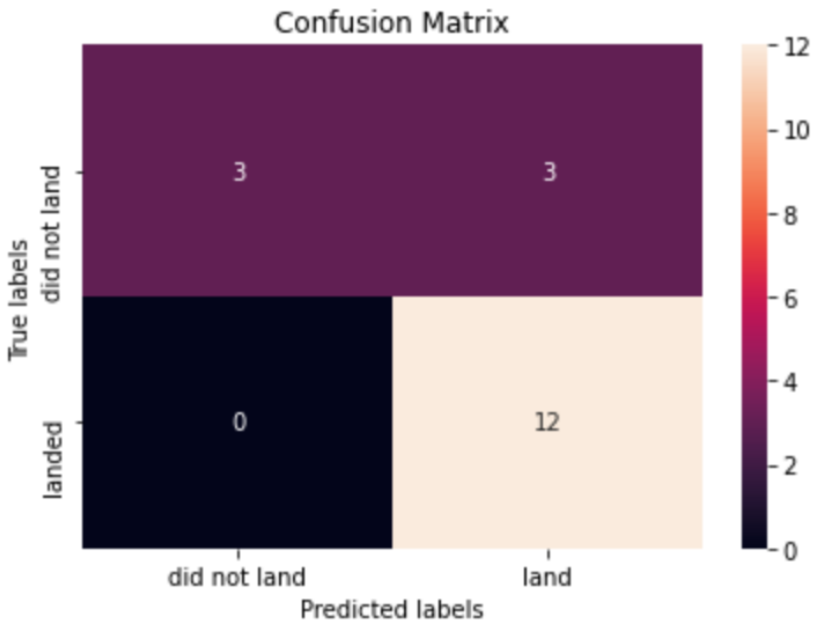
RESULTS

D Predictive Analysis

- **Logistic regression**
  - Best score: 0.8464285714285713



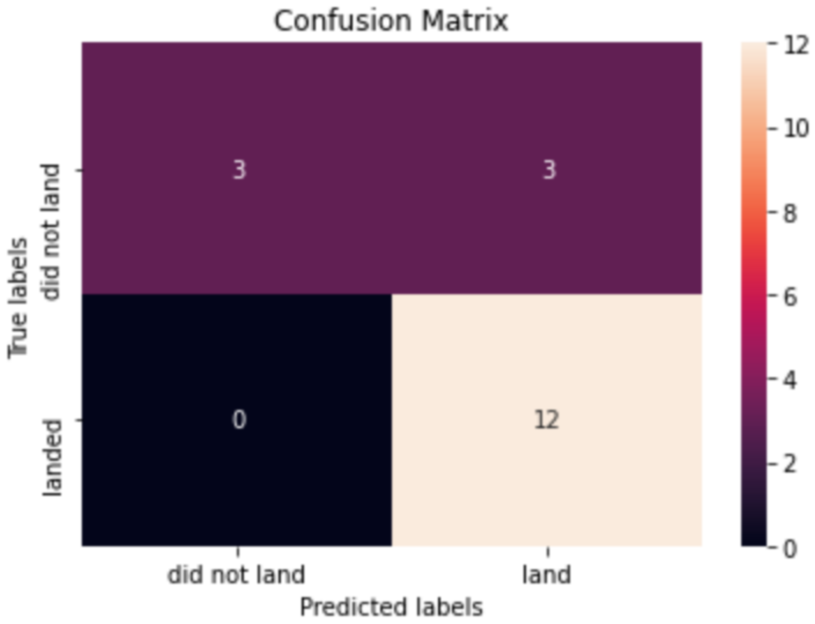
- **Support vector machine (SVM)**
  - Best score: 0.8482142857142856



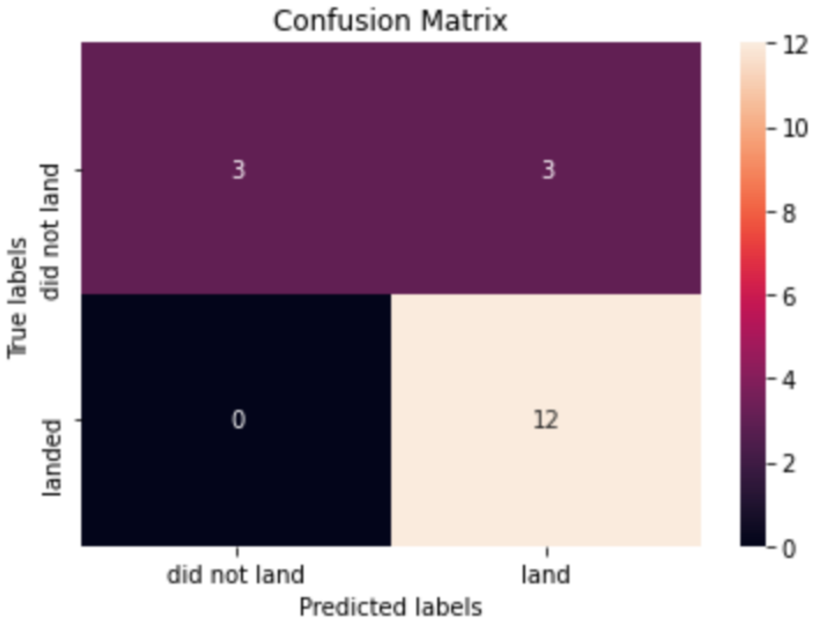
RESULTS

D Predictive Analysis

- **Decision tree**
  - Best score: 0.8892857142857142



- **K nearest neighbors (KNN)**
  - Best score: 0.8482142857142858



RESULTS

D Predictive Analysis

Upon comparing the results of all four models side by side, it was observed that they shared identical accuracy scores and confusion matrices when evaluated on the test set. Consequently, the best scores from GridSearchCV were utilized to rank the models. Based on these scores, the models were ranked as follows, with the top model listed first:

1. Decision Tree (GridSearchCV best score: 0.89)

2. K Nearest Neighbors (KNN) (GridSearchCV best score: 0.85)

3. Support Vector Machine (SVM) (GridSearchCV best score: 0.85)

4. Logistic Regression (GridSearchCV best score: 0.85)

DISCUSSION

Insights gained from the data visualizations suggest that certain features may influence mission outcomes. For instance, missions with heavier payloads often achieve higher success rates in orbits such as Polar, LEO, and ISS. However, the GTO orbit presents a mixed scenario, with both successful and unsuccessful landings recorded.

While the precise impact of each feature on mission success can be intricate, it is evident that they play a role. By leveraging machine learning algorithms, one can analyze historical data to discern patterns that inform predictions about future mission success based on these features.

CONCLUSION

- This project aims to predict whether the first stage of a Falcon 9 launch will successfully land, which can be instrumental in estimating the total cost of the launch. Each feature associated with a Falcon 9 launch, such as payload mass and orbit type, may have a unique influence on the mission's outcome.

- The analysis revealed that certain features significantly impact the landing outcome, allowing for a deeper understanding of the factors at play in rocket launches.

- To develop predictive models, a variety of machine learning algorithms are employed to analyze historical data from Falcon 9 launches. These models are designed to uncover patterns that can assist in forecasting the success of future missions.

- Among the four algorithms tested, the decision tree model outperformed the others, demonstrating the highest accuracy in predicting landing success. This model is particularly effective for forecasting the outcomes of Falcon 9 launches.