IBM

Applied Data Science capstone

Chloe Kastelijn
May 9, 2025

Contents

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

EXECUTIVE SUMMARY

- In this capstone project, we will use a number of machine learning classification methods to forecast whether the SpaceX Falcon 9 first stage will land successfully.
  This project's primary steps are as follows:
  Gathering, organising, and formatting data

- Analysing data exploratorily
  Data visualisation that is interactive
  Prediction using machine learning
  According to our graphs, there is a relationship between some aspects of rocket launches and their success or failure.
  Decision trees may be the most effective machine learning technique for predicting whether the Falcon 9 first stage will land successfully, it is also found.

INTRODUCTION

- We will forecast whether the Falcon 9 first stage will land successfully in this event. Due in large part to SpaceX's ability to reuse the first stage, the company advertises Falcon 9 rocket launches on its website for 62 million dollars, whereas other companies charge upwards of 165 million dollars per. Thus, we can calculate the launch cost if we can predict whether the first stage will land. If a different business want to bid against SpaceX for a rocket launch, they can utilise this information.

- The majority of failed landings are prearranged. SpaceX occasionally does a controlled landing in the ocean.

- Our primary goal is to determine if the Falcon 9 rocket's first stage will land successfully given a specific set of launch parameters, such as the rocket's payload tonnage, orbit type, launch site, and so forth.

METHODOLOGY

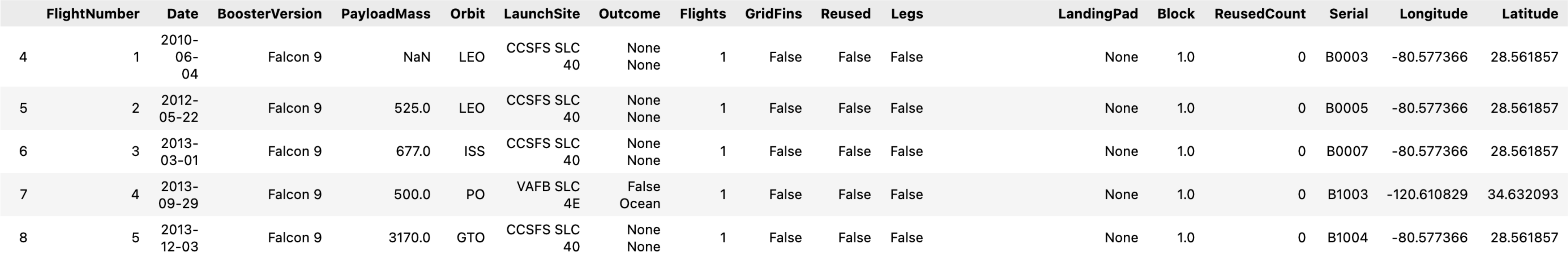
- The overall methodology includes:
  1. Data collection, wrangling, and formatting, using:
     - SpaceX API
     - Web scraping
  2. Exploratory data analysis (EDA), using:
     - Pandas and NumPy
     - SQL
  3. Data visualization, using:
     - Matplotlib and Seaborn
     - Folium
     - Dash
  4.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

METHODOLOGY

1 Data collection, wrangling, and formatting

- API for SpaceX
  The https://api.spacexdata.com/v4/rockets/ API is utilised.
  The API offers information on a variety of SpaceX rocket launches; as a result, the information is filtered to only include Falcon 9 missions.

  The mean of the column to which the missing value belongs is used to replace each missing value in the data.
  In the end, we have 17 columns, or features, and 90 rows, or instances. The first few rows of the data are displayed in the image below:

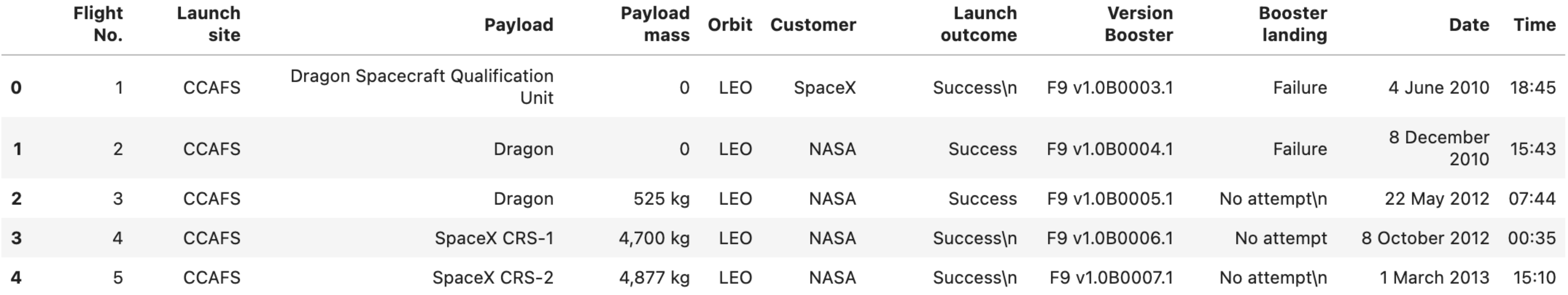


	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO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

1 Data collection, wrangling, and formatting

- 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

METHODOLOGY

1 Data collection, wrangling, and formatting

- One-hot encoding is used to encode categorical features after the data has been processed to ensure that no entries are missed.
- The data frame also gains an additional column named "Class." If a launch is successful, the value in the "Class" column is 1, and if it fails, it is 0.
- Ultimately, we have 83 columns (features) and 90 rows (instances).

METHODOLOGY

2 Exploratory Data Analysis (EDA)

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
    - The number of launches on each launch site
    - The number of occurrence of each orbit
    - The number and occurrence of each mission outcome
- SQL
  - The data is queried using SQL to answer several questions about the data such as:
    - The names of the unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1

METHODOLOGY

3 Data Visualization

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type
- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
    - Mark all launch sites on a map
    - Mark the succeeded launches and failed launches for each site on the map
    - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

METHODOLOGY

3 Data Visualization

- 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

METHODOLOGY

4 Machine Learning Prediction

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
    - Standardizing the data
    - Splitting the data into training and test data
    - Creating machine learning models, which include:
        - Logistic regression
        - Support vector machine (SVM)
        - Decision tree
        - K nearest neighbors (KNN)
    - Fit the models on the training set
    - Find the best combination of hyperparameters for each model
    - Evaluate the models based on their accuracy scores and confusion matrix

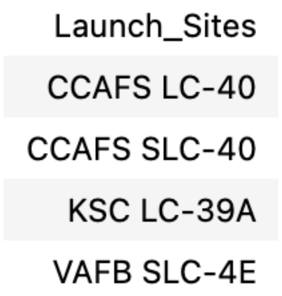
RESULTS

- The results are split into 5 sections:
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

RESULTS

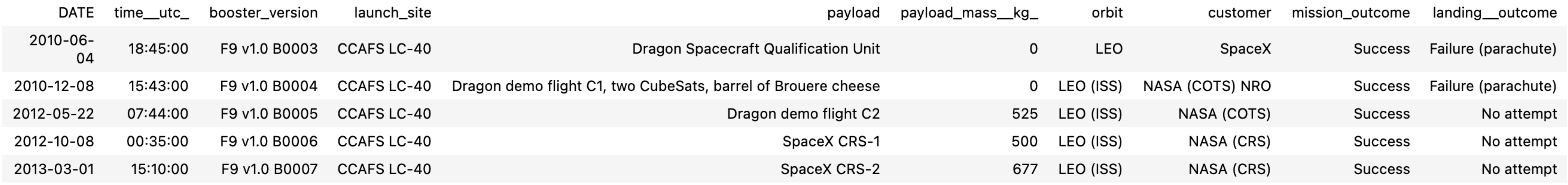
1 SQL (EDA with SQL)

- 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’

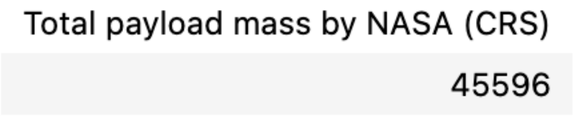


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight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

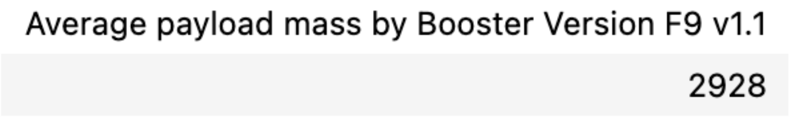
1 SQL (EDA with SQL)

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



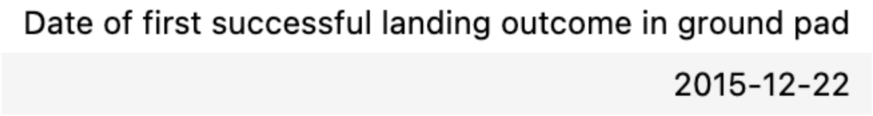
Total payload mass by NASA (CRS)
45596

- 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

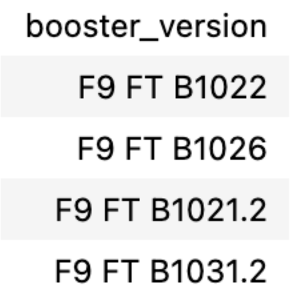


Date of first successful landing outcome in ground pad
2015-12-22

RESULTS

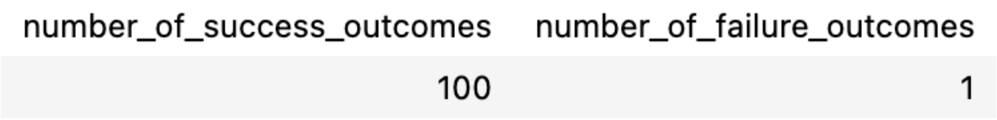
1 SQL (EDA with SQL)

- 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

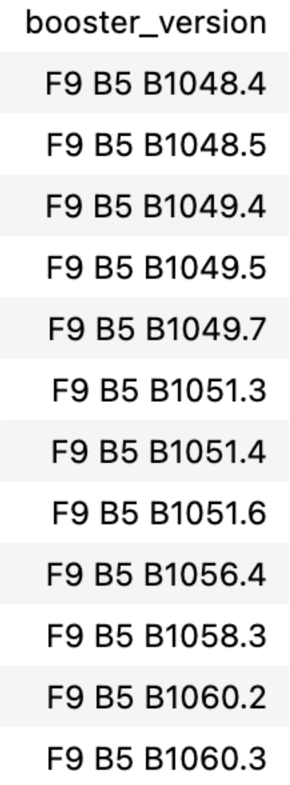


number_of_success_outcomes	number_of_failure_outcomes
100	1

RESULTS

1 SQL (EDA with SQL)

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

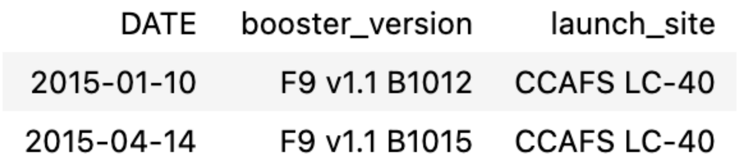


booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

RESULTS

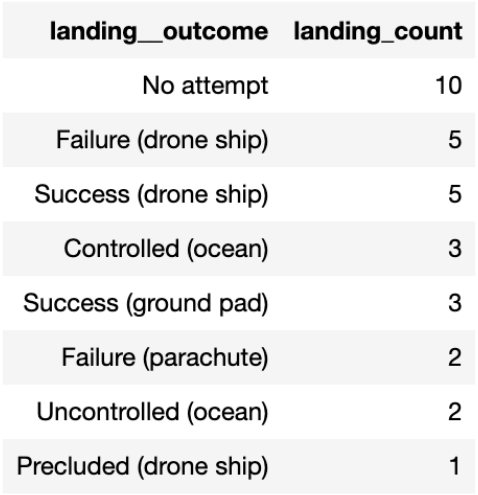
1 SQL (EDA with SQL)

- 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

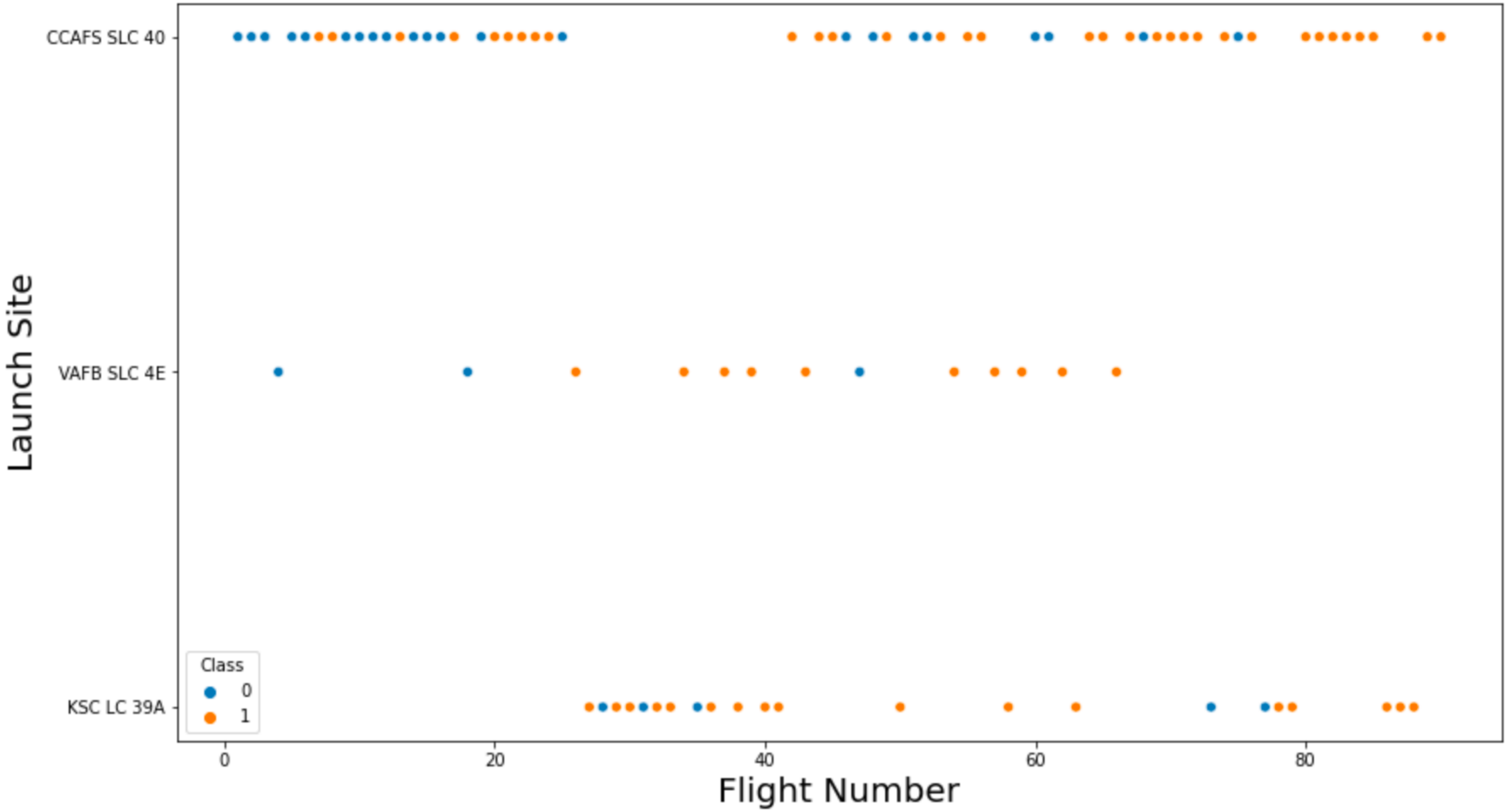


landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

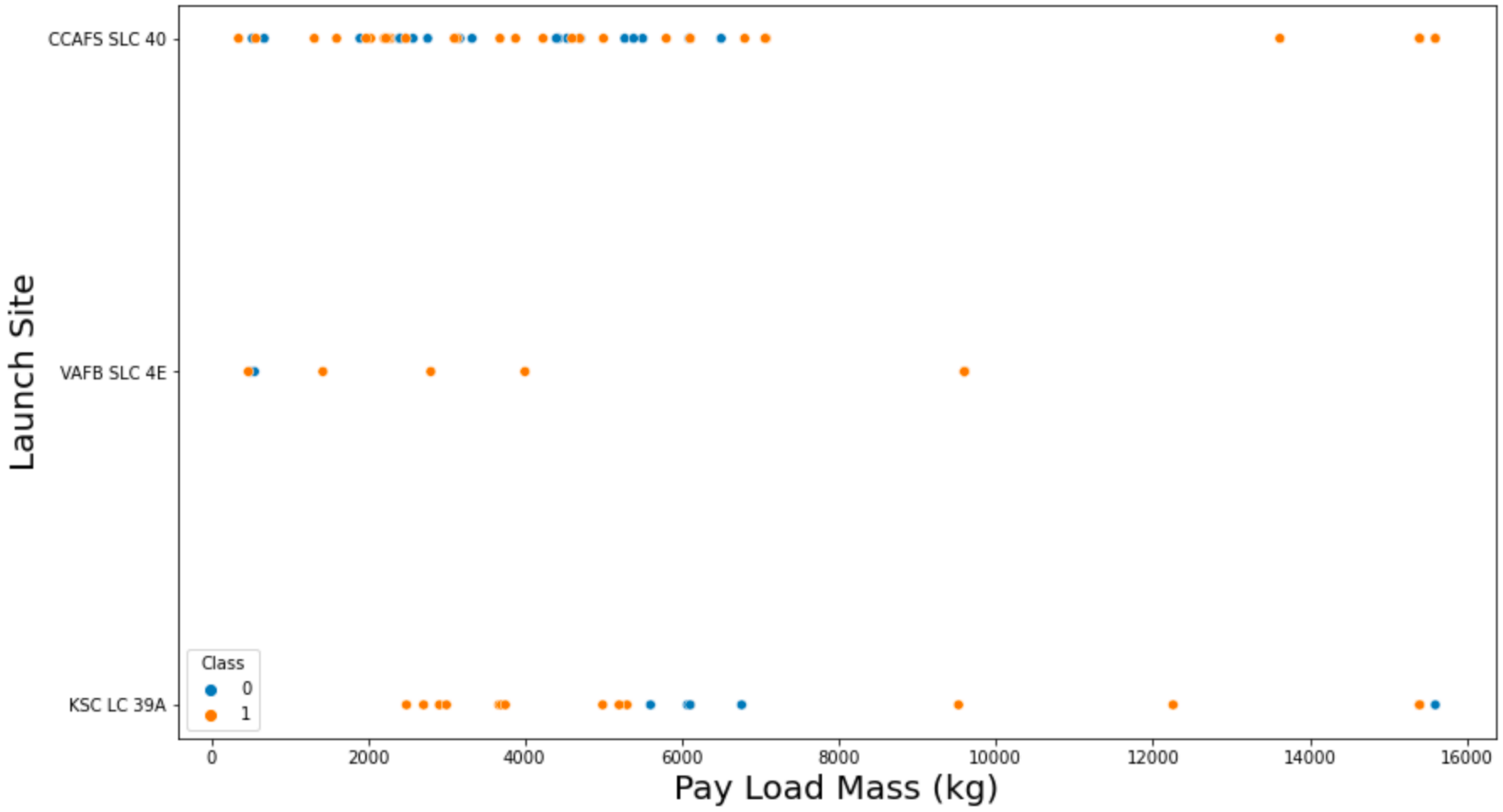
- The relationship between flight number and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

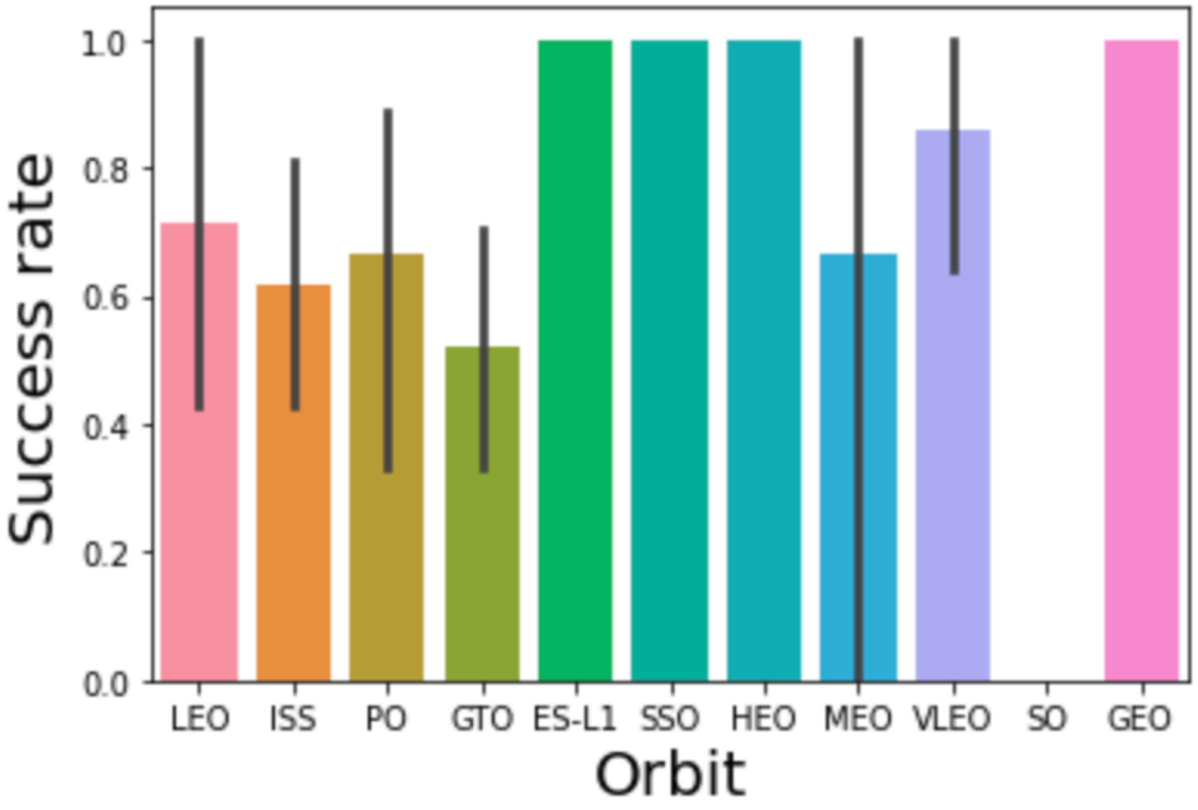
- The relationship between payload mass and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

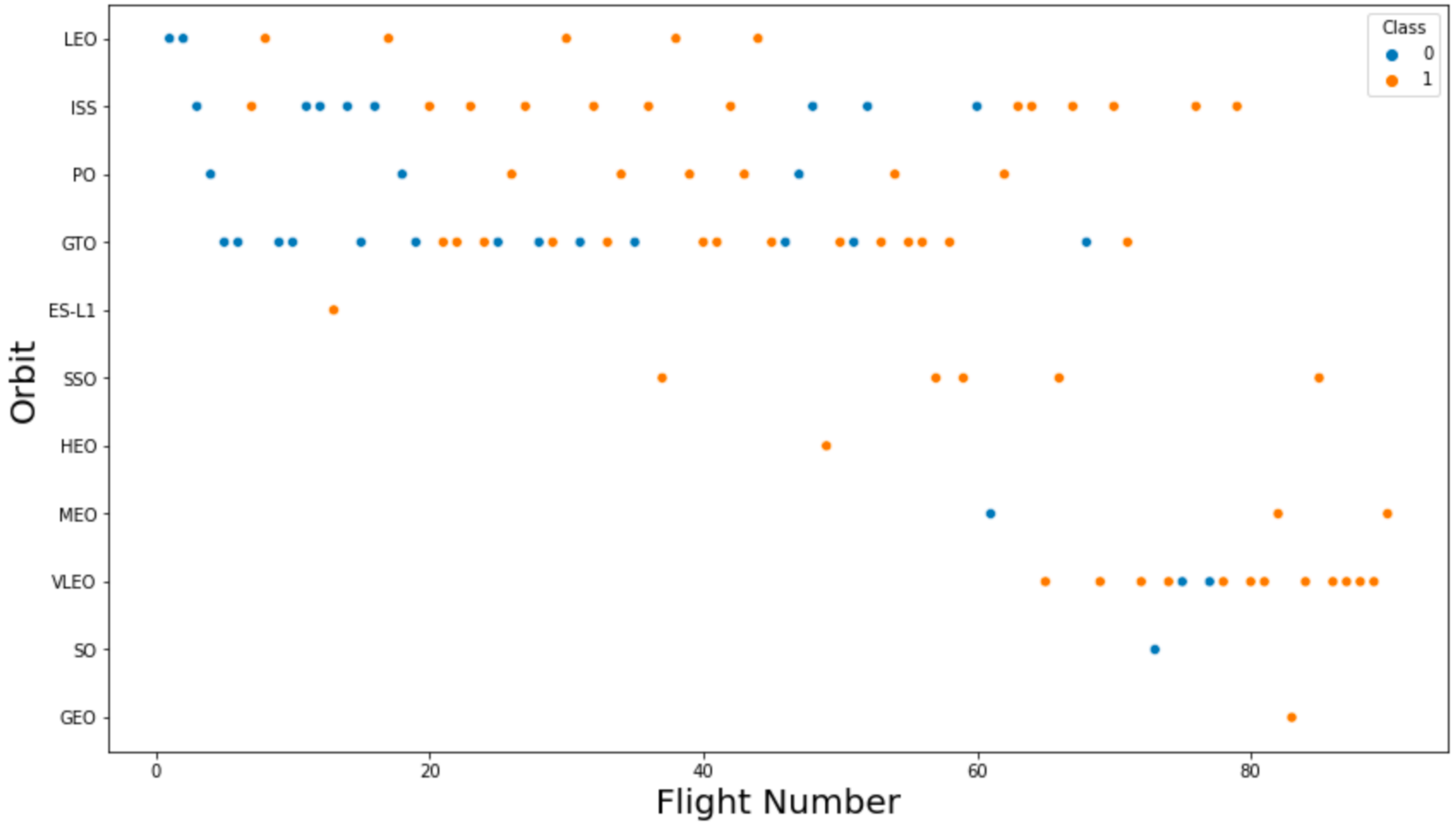
- The relationship between success rate and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

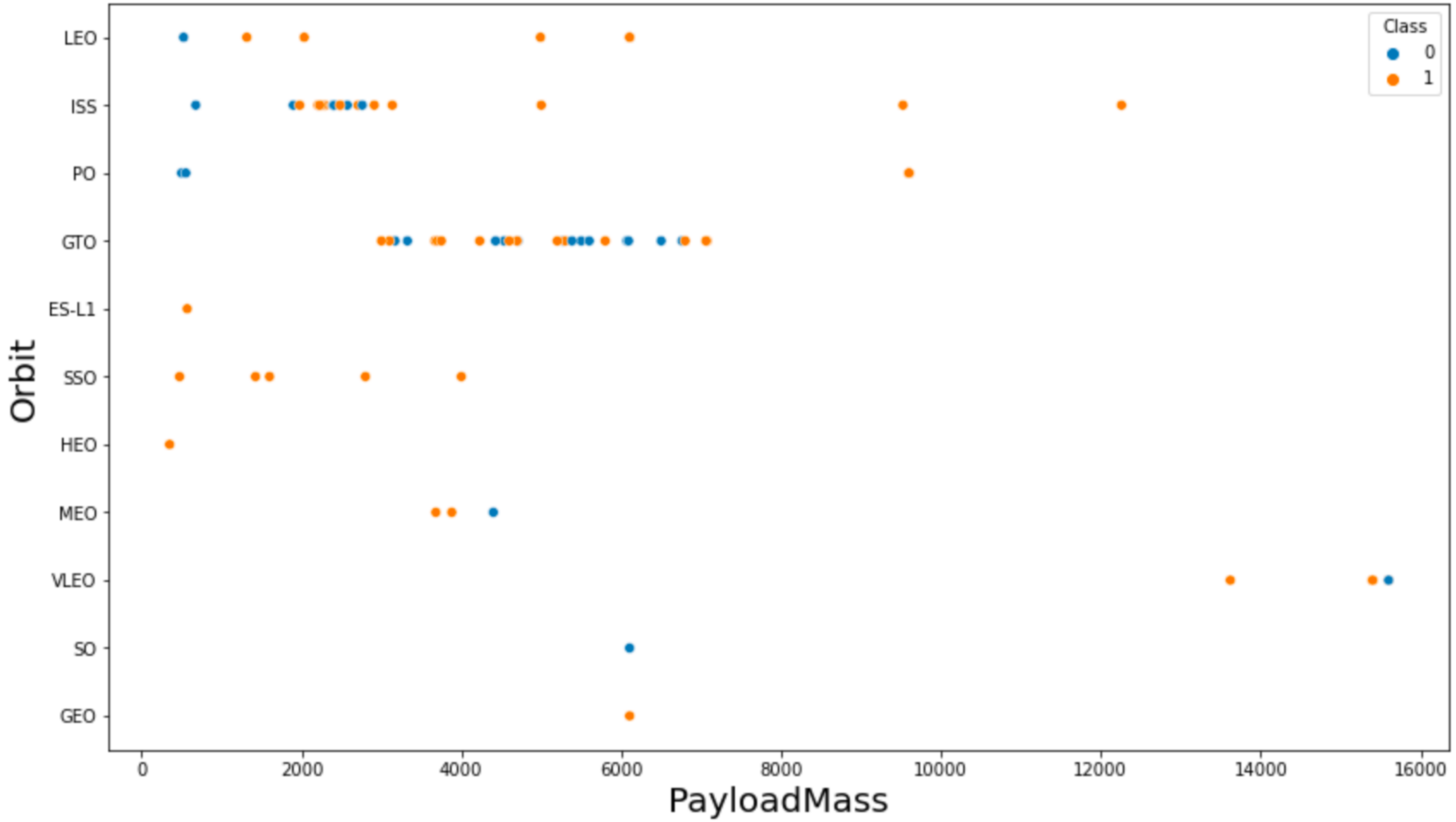
- The relationship between flight number and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

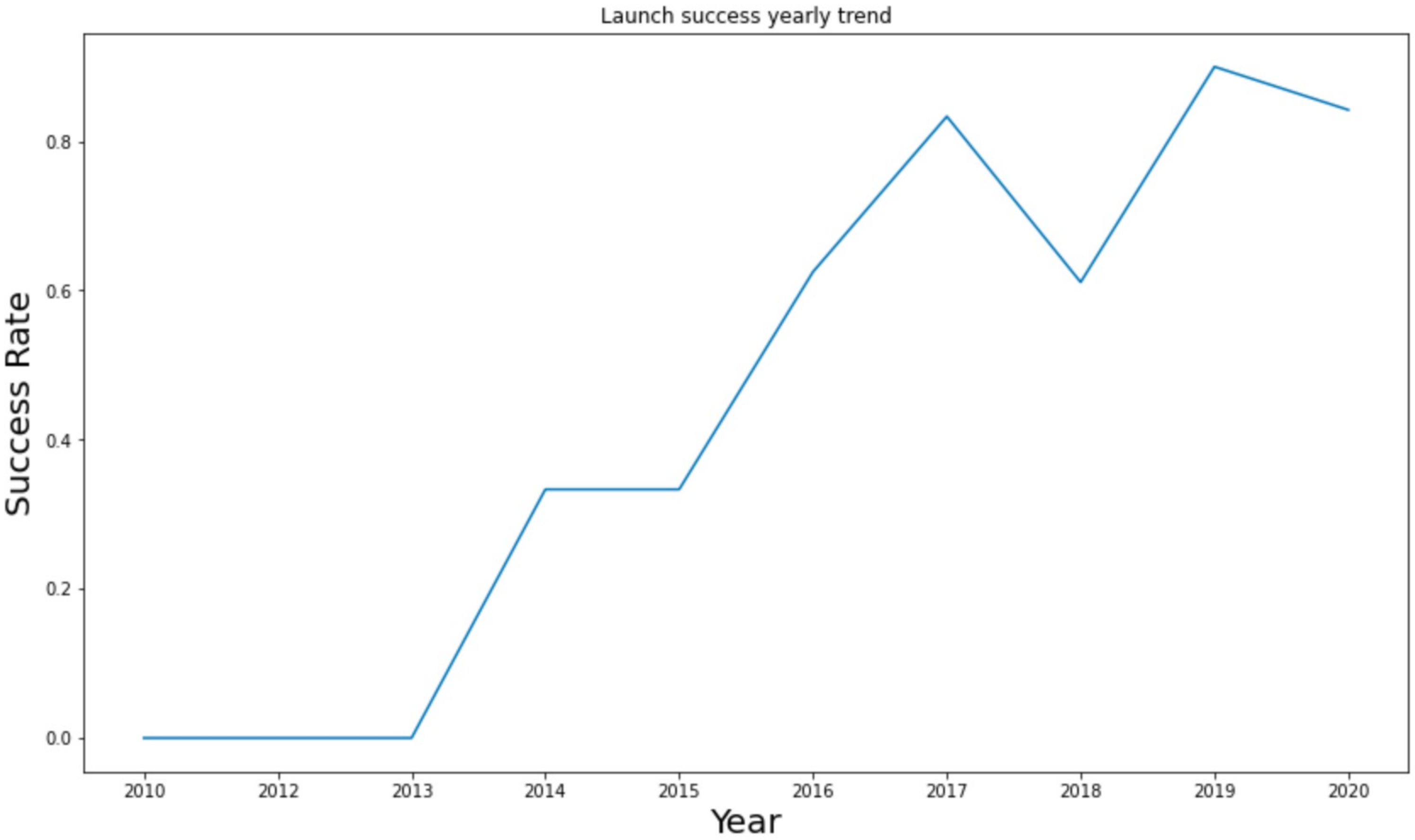
- The relationship between payload mass and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

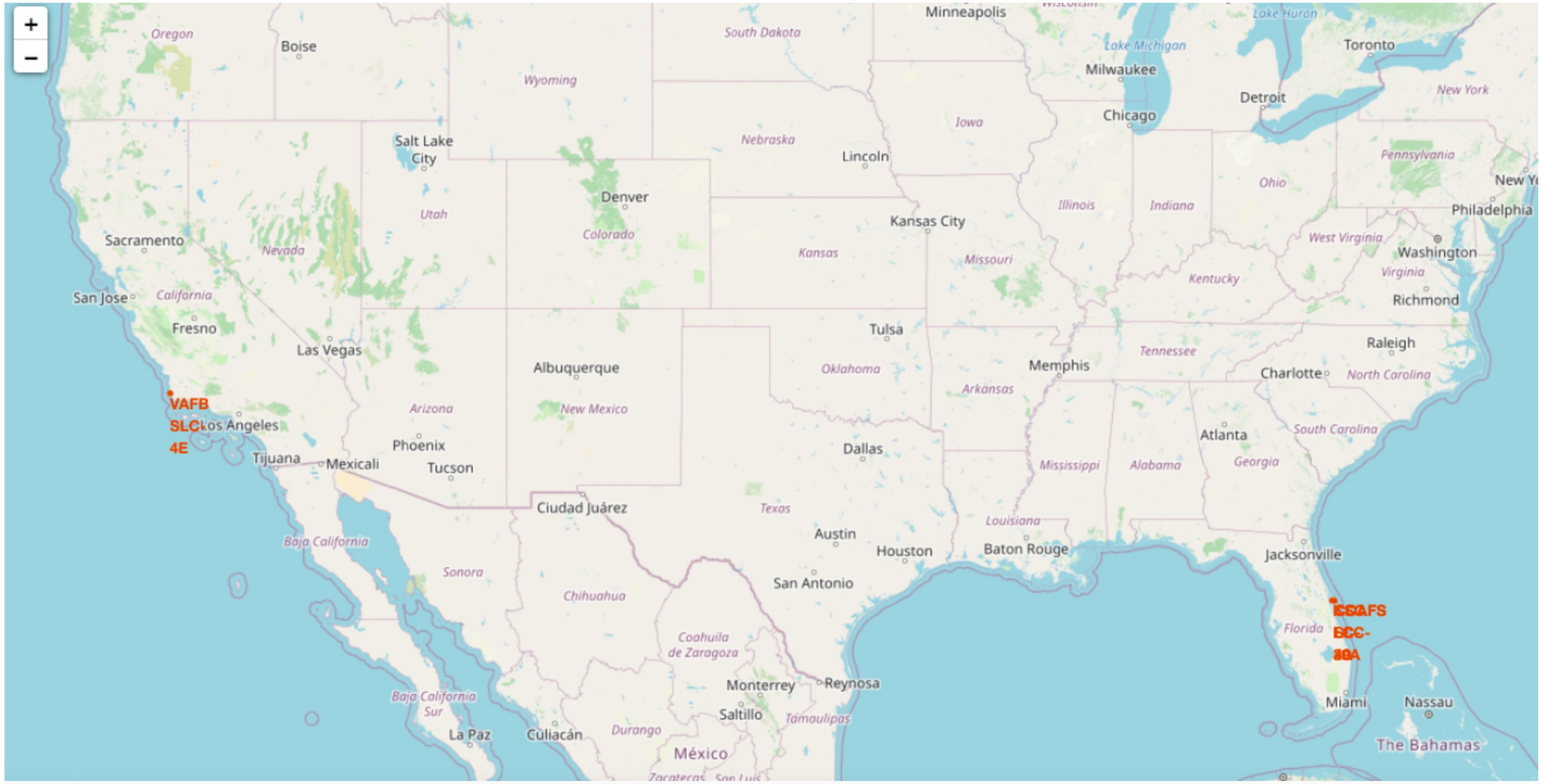
- The launch success yearly trend



RESULTS

3 Folium

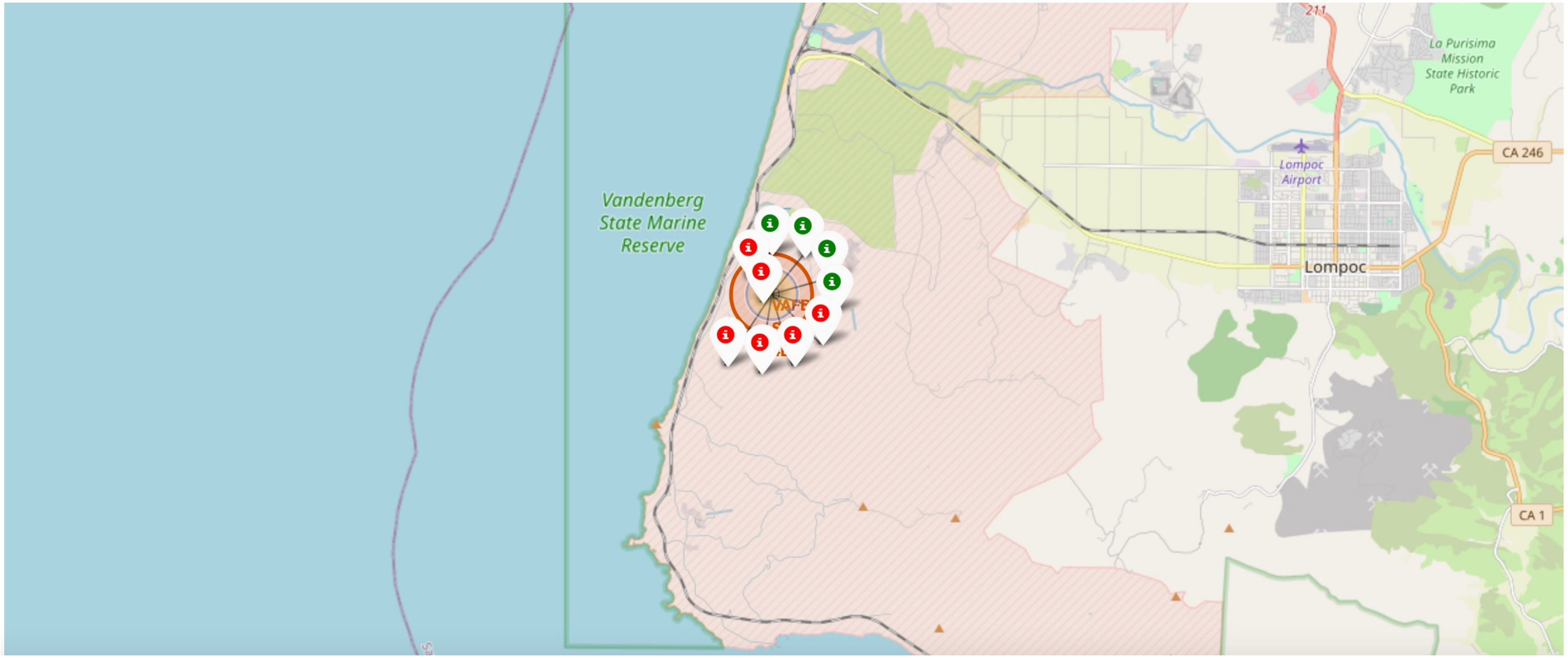
- All launch sites on map



RESULTS

3 Folium

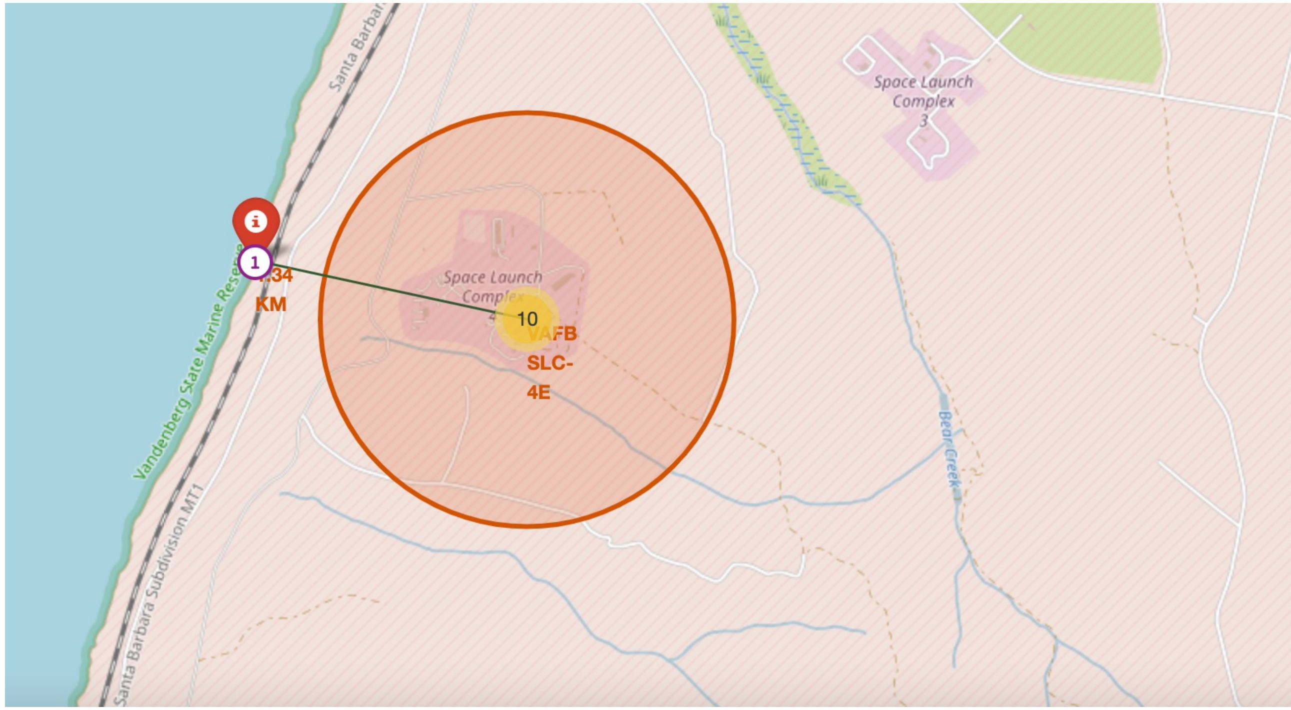
- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch


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

RESULTS

3 Folium

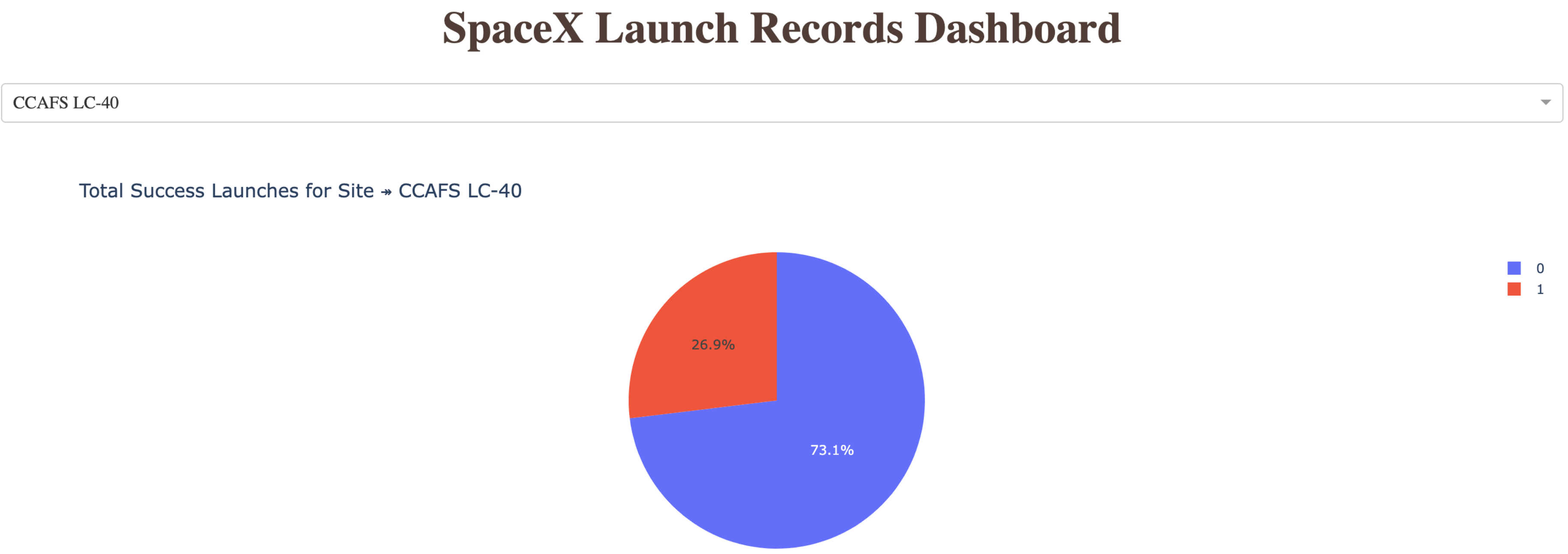
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



RESULTS

4 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

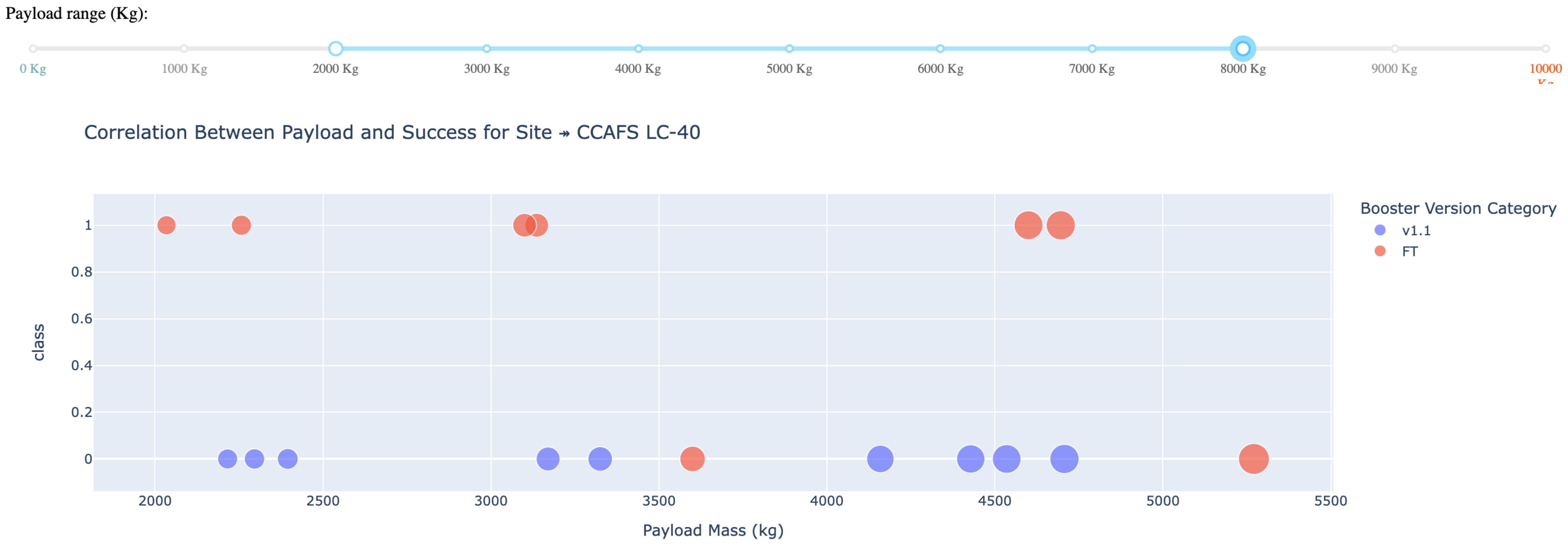
CCAFS LC-40

Total Success Launches for Site → CCAFS LC-40

RESULTS

4 Dash

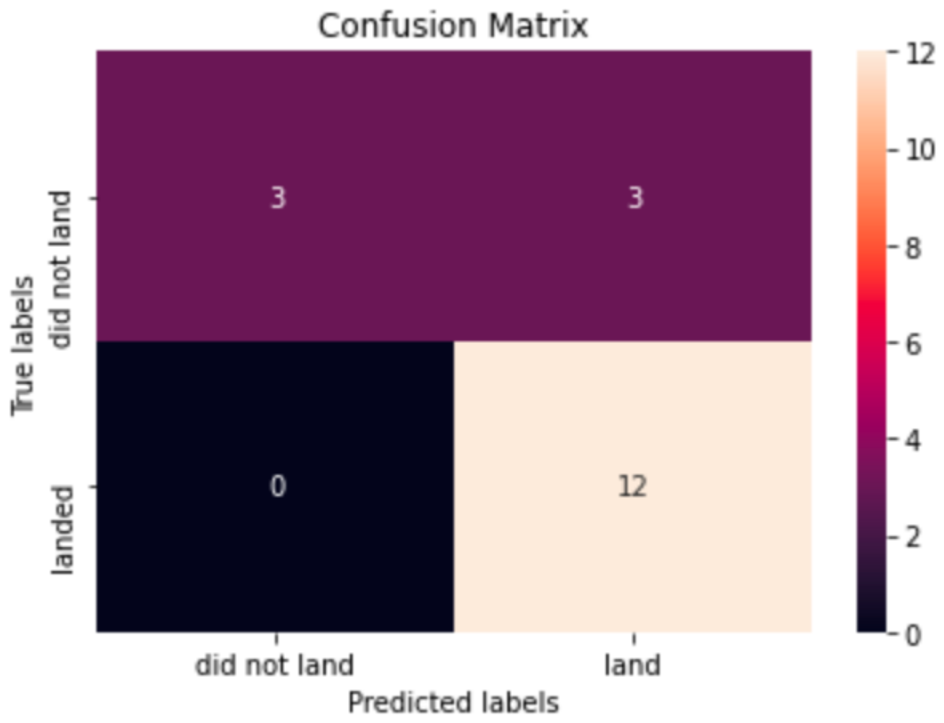
- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.



RESULTS

5 Predictive Analysis

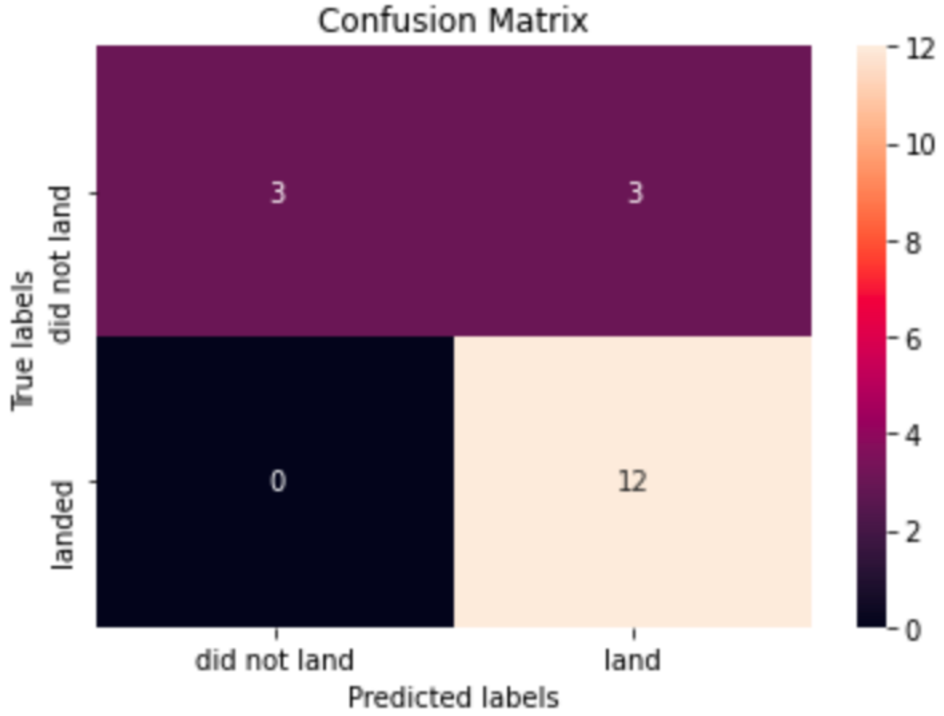
- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

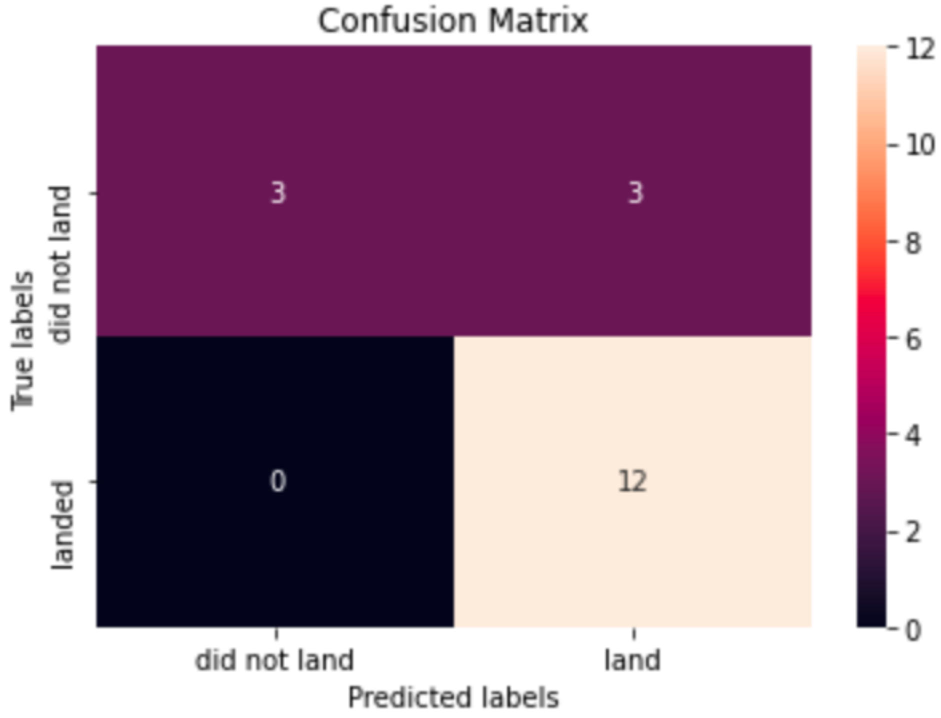
- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

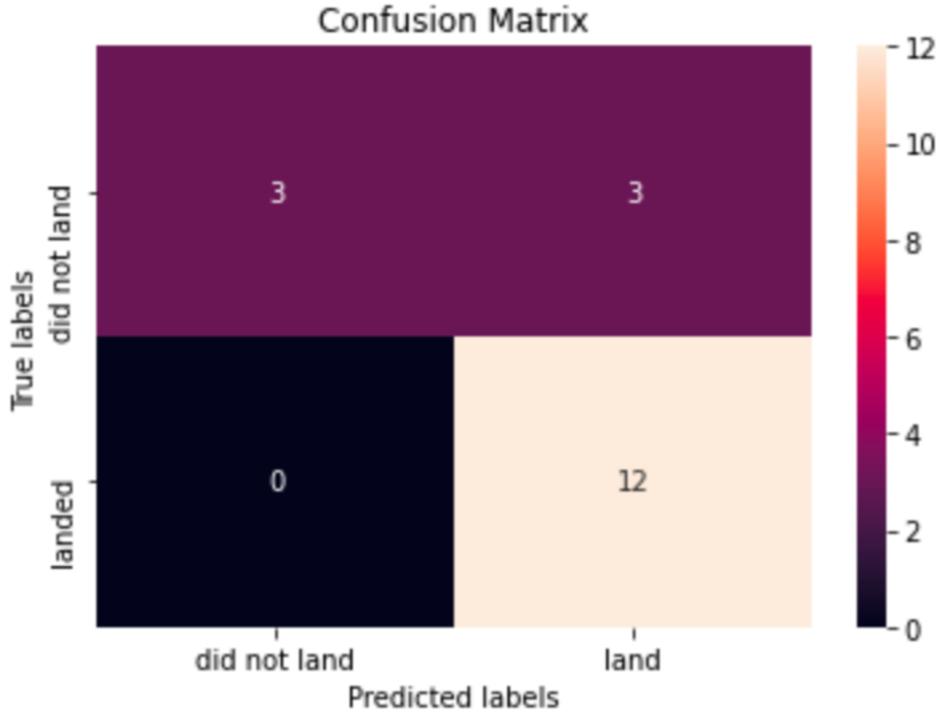
- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

When comparing the outcomes of the four models, we can observe that they all have the same confusion matrix and accuracy score when tested on the test set.
As a result, they are ranked using their GridSearchCV top scores. The models are arranged according to their GridSearchCV best scores, with the top model being the best and the bottom model being the worst:
GridSearchCV's top score for a decision tree is 0.8892857142857142.

(GridSearchCV highest score: 0.8482142857142858) K closest neighbours, KNN
(GridSearchCV highest score: 0.8482142857142856) Support vector machine, SVM
GridSearchCV's optimal score for logistic regression is 0.8464285714285713.

DISCUSSION

- We can observe from the data visualisation section that certain traits might be related to the mission success in a number of ways. For instance, orbit types Polar, LEO, and ISS have higher successful landing or positive landing rates when carrying big payloads. We are unable to differentiate this clearly for GTO, though, because both positive landing rates and negative landings (unsuccessful missions) are present.

- As a result, any feature might influence the mission's ultimate result in some way. It is challenging to determine the precise ways in which each of these characteristics affects the mission outcome. On the basis of the provided features, we can apply machine learning algorithms to identify patterns in historical data and forecast the success or failure of a mission.

CONCLUSION

- In this research, we attempt to estimate the cost of a Falcon 9 launch by predicting whether the first stage of the launch will land.
- Every aspect of a Falcon 9 launch, including the kind of orbit or the mass of the cargo, may have an impact on the mission's success.
- To create prediction models that may be used to forecast the result of a Falcon 9 launch, a number of machine learning algorithms are used to identify patterns in historical Falcon 9 launch data.
- Out of the four machine learning algorithms used, the decision tree algorithm's prediction model outperformed the others.