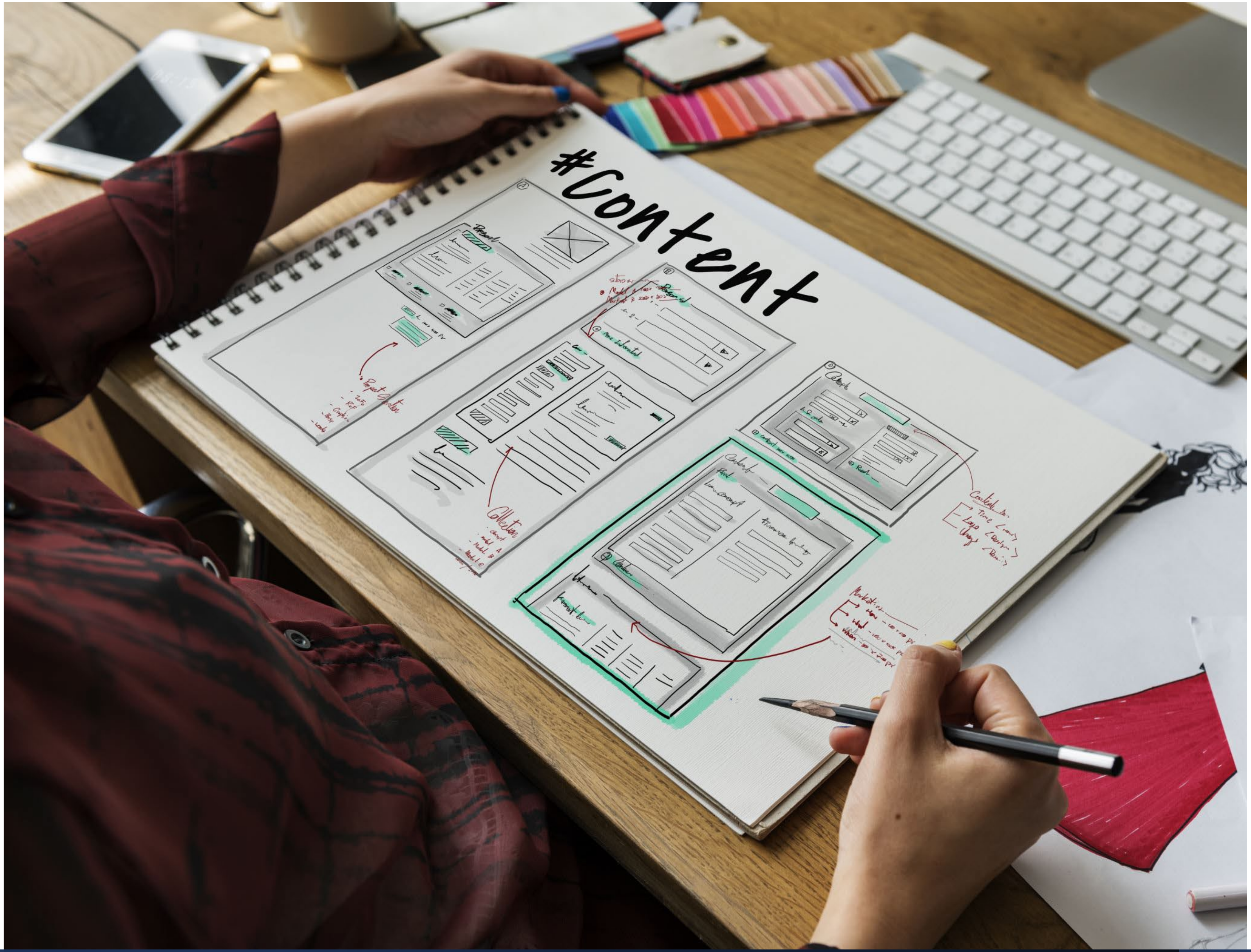
Applied Data Science capstone

Huseyin C

December 8, 2023



OUTLINE

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

Outline

- **EXECUTIVE SUMMARY**

In this capstone project, our objective is to anticipate the outcome of the SpaceX Falcon 9 first stage landing, utilizing a range of machine learning classification techniques. The primary stages of this endeavor encompass:

1.Gathering, refining, and organizing the data.
2.Conducting an exploratory analysis of the data.
3.Creating interactive data visualizations.
4.Employing machine learning for predictive purposes.

Our visual representations reveal that certain aspects of the rocket launches exhibit a connection with the outcome, specifically, whether it was a success or failure. Additionally, our analysis suggests that the decision tree algorithm may offer the most promising results in forecasting the successful landing of the Falcon 9 first stage.

Outline

- **INTRODUCTION**

In this capstone project, our primary goal is to predict the successful landing of the Falcon 9 first stage. It's noteworthy that SpaceX promotes Falcon 9 rocket launches on their website at a cost of 62 million dollars, a significantly more budget-friendly option compared to other providers whose charges can soar as high as 165 million dollars per launch. This cost disparity primarily arises from SpaceX's innovative practice of reusing the first stage of the rocket. Consequently, by determining the likelihood of a successful first stage landing, we can ascertain the overall cost of a launch. This valuable information can be instrumental if another company wishes to compete with SpaceX in bidding for a rocket launch contract.

It's important to note that many instances of unsuccessful landings are actually planned occurrences. SpaceX occasionally conducts controlled landings in the ocean for specific purposes.

The central question we aim to address revolves around the following: Given a set of features related to a Falcon 9 rocket launch, encompassing factors such as payload mass, orbit type, launch site, and more, can we reliably predict whether the first stage of the rocket will achieve a successful landing?

METHODOLOGY

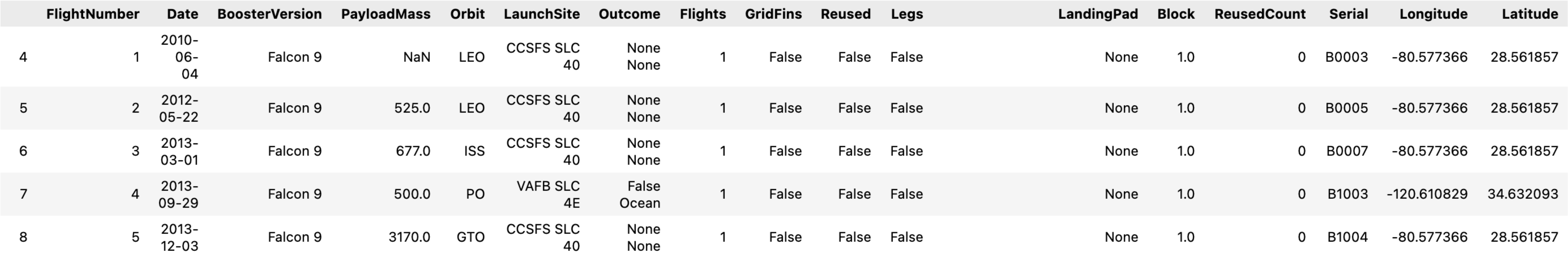
- The comprehensive approach encompasses:

1. Data acquisition, preparation, and structuring, utilizing:
   1. SpaceX API
   2. Web scraping
2. Exploratory data analysis (EDA), incorporating:
   1. Pandas and NumPy
   2. SQL
3. Data representation through visualization, employing:
   1. Matplotlib and Seaborn
   2. Folium
   3. Dash
4. Machine learning forecasting, utilizing:
   1. Logistic regression
   2. Support vector machine (SVM)
   3. Decision tree
   4. K-nearest neighbors (KNN)

METHODOLOGY

1 Data collection, wrangling, and formatting

- SpaceX API: For our data source, we rely on the SpaceX API, accessible via the URL https://api.spacexdata.com/v4/rockets/. This API furnishes a wealth of information regarding various rocket launches conducted by SpaceX. To streamline our analysis, we narrow our focus exclusively to Falcon 9 launches.
- Data Handling: To address missing data points within the dataset, we employ a strategy of imputation. Specifically, any missing values are replaced with the mean value of the respective column to ensure completeness and accuracy.
- Data Dimensions: After this data wrangling process, our dataset comprises a total of 90 rows, each representing individual instances, and 17 columns, which correspond to different features. Below, you can find a snapshot illustrating the initial rows of this 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

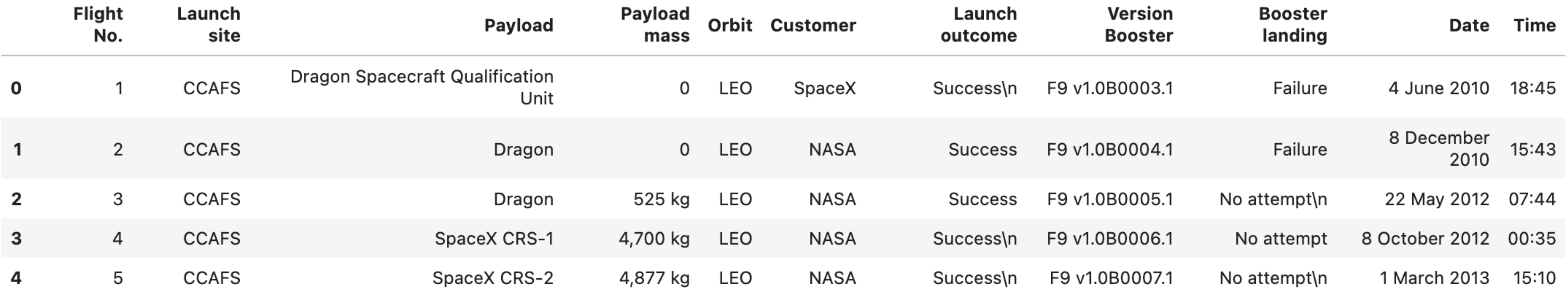
METHODOLOGY

1 Data collection, wrangling, and formatting

-

  Web Scraping:

- To gather additional data for our analysis, we perform web scraping from the webpage located at https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922. It's important to note that this website exclusively contains information pertaining to Falcon 9 launches.
- Data Scope: Following the web scraping process, our dataset expands to encompass 121 rows, signifying distinct instances, and includes 11 columns that represent various features. Please refer to the image below for an initial glimpse of the data's initial rows:



	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

-

  Data Processing and Transformation: Subsequent to data collection, we engage in comprehensive data preprocessing. Our approach includes eliminating any missing entries to ensure data completeness. Additionally, categorical features are encoded using one-hot encoding, a technique that converts them into numerical values for analysis.
- Introduction of 'Class' Column: A pivotal step in our data transformation involves the introduction of an additional column termed 'Class.' Within this column, a value of 0 is assigned to launches classified as failures, while successful launches are denoted with a value of 1. This binary classification serves as a fundamental element of our analysis.
- Final Data Dimensions: Upon completing these data processing steps, our dataset is streamlined to consist of 90 rows, each representing unique instances, and a total of 83 columns, encompassing various features for analysis and prediction.

METHODOLOGY

2 Exploratory Data Analysis (EDA)

Pandas and NumPy: We harness the capabilities of the Pandas and NumPy libraries to perform fundamental data analysis tasks. This includes computing essential statistics and insights such as:

1. Determining the count of launches conducted at each individual launch site.
2. Tabulating the frequency of each specific orbit in our dataset.
3. Establishing the count and prevalence of different mission outcomes.

- SQL Queries: Furthermore, we employ SQL queries to extract valuable information from our dataset, addressing pertinent questions such as:

1. Identifying the distinct names of launch sites participating in space missions.
2. Calculating the total payload mass transported by boosters launched under NASA's CRS (Commercial Resupply Services) program.
3. Evaluating the average payload mass carried by booster version F9 v1.1.

- These analytical approaches, encompassing both Pandas/NumPy and SQL, are pivotal in gaining deeper insights into our data.

METHODOLOGY

3 Data Visualization

- Matplotlib and Seaborn:
- Incorporating functions from the Matplotlib and Seaborn libraries, we employ various visualization techniques, including scatterplots, bar charts, and line charts. These visual representations enable us to explore and comprehend relationships among different features in our dataset, notably:

- 1. Examining the correlation between flight number and launch site.
- 2. Analyzing how payload mass relates to the launch site.
- 3. Investigating the success rate in relation to the type of orbit.

- Folium:
- To enhance our data visualization capabilities, we leverage the Folium library to create interactive maps. This facilitates the following visualizations:

- 1. Pinpointing all launch sites on a map for geographical context.
- 2. Distinguishing between successful and failed launches for each site, offering insights into performance.
- 3. Mapping the distances between launch sites and key proximities such as the nearest city, railway, or highway, aiding in spatial analysis.

- These visualization tools, including Matplotlib, Seaborn, and Folium, play a crucial role in our data exploration and presentation.

METHODOLOGY

3 Data Visualization

Dash Integration:

- We integrate the capabilities of the Dash library to create an interactive web-based platform. This platform features user-friendly elements such as a dropdown menu and a range slider, which allow users to toggle inputs and customize their data exploration experience.

Interactive Site Content:

- Within this interactive site, we present information through a pie chart and a scatterplot, offering valuable insights into our dataset, including:

- 1. Total successful launches originating from each launch site, providing a comprehensive overview of success rates across different locations.
- 2. The correlation between payload mass and mission outcomes (success or failure) for each launch site, using a scatterplot. This allows for a nuanced examination of how payload mass influences mission success at specific launch sites.

Dash empowers users to interact with and explore the data interactively, fostering a deeper understanding of the relationships and trends within our dataset.

METHODOLOGY

4 Machine Learning Prediction

Scikit-Learn for Machine Learning:

- To develop our machine learning models, we leverage functions and tools from the Scikit-Learn library, a robust resource for machine learning tasks. Our machine learning prediction phase encompasses several essential steps:

1. Data Standardization:

- - Ensuring that our data is standardized, which often involves scaling numerical features to have a consistent range.

2. Data Splitting:

- - Segregating the dataset into distinct training and test sets. This division is crucial for training and validating the models.

3. Machine Learning Model Creation:

- - Developing various machine learning models to facilitate predictions. These models encompass:
- - Logistic regression
- - Support vector machine (SVM)
- - Decision tree
- - K nearest neighbors (KNN)

4. Model Training:

- - Fitting these machine learning models on the training set to enable them to learn from the data.

5. Hyperparameter Tuning:

- - Identifying the optimal combination of hyperparameters for each model. This process enhances model performance.

6. Model Evaluation:

- - Assessing the models based on their accuracy scores, which indicate their predictive capability, and utilizing confusion matrices to gain insights into their classification performance.

- These machine learning steps using Scikit-Learn allow us to make informed predictions and draw meaningful conclusions from our data.

RESULTS

- 1. SQL (EDA with SQL):
- - Exploratory Data Analysis conducted using SQL, likely involving data queries and insights extracted from your dataset.

2. Matplotlib and Seaborn (EDA with Visualization):

- - Exploratory Data Analysis facilitated through data visualization techniques using Matplotlib and Seaborn, with a particular focus on the relationships and patterns within the data.

3. Folium:

- - Visualizations and spatial analysis carried out using the Folium library, which likely includes interactive maps and geographical insights.

4. Dash:

- - An interactive web-based platform created with Dash, showcasing insights from your data through pie charts and scatterplots.

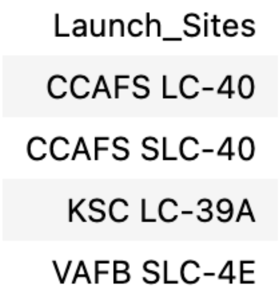
5. Predictive Analysis:

- - The phase where machine learning models, including logistic regression, support vector machine, decision tree, and K-nearest neighbors, are employed for predictive analysis, with a specific focus on classifying launches as either successful (class 1) or failed (class 0).
- In all of these sections, the distinction between class 0 (failed launch outcome) and class 1 (successful launch outcome) is maintained, allowing for a clear interpretation of the results. If you have any specific questions or if there's a particular section you'd like to discuss further, please let me know.

RESULTS

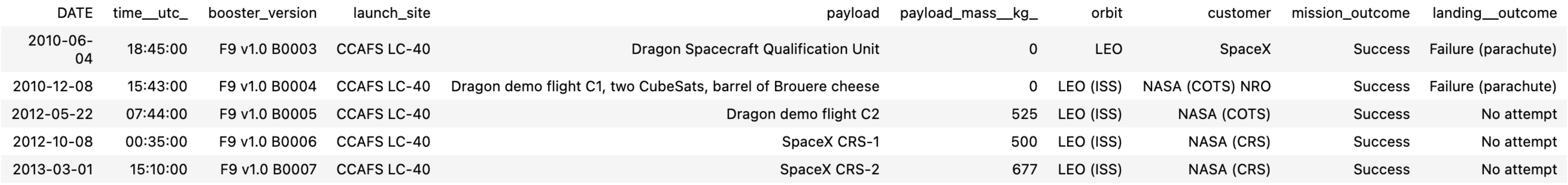
1 SQL (EDA with SQL)

- 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’

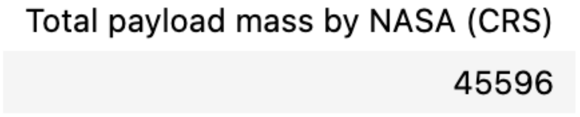


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight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

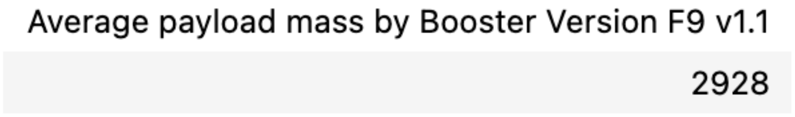
1 SQL (EDA with SQL)

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



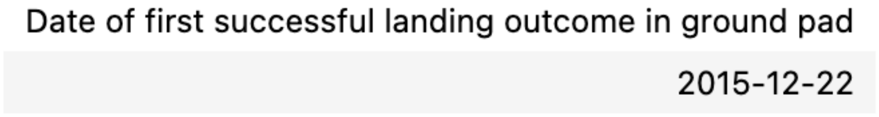
Total payload mass by NASA (CRS)
45596

- 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

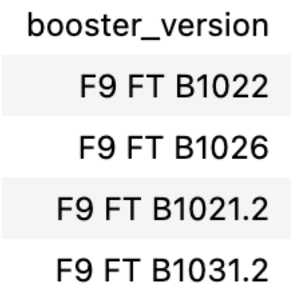


Date of first successful landing outcome in ground pad
2015-12-22

RESULTS

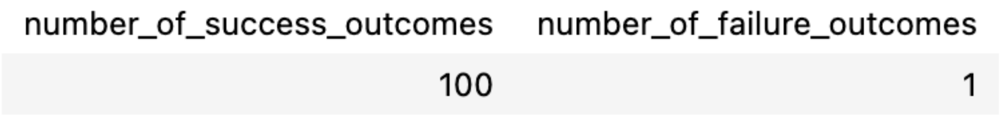
1 SQL (EDA with SQL)

- 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

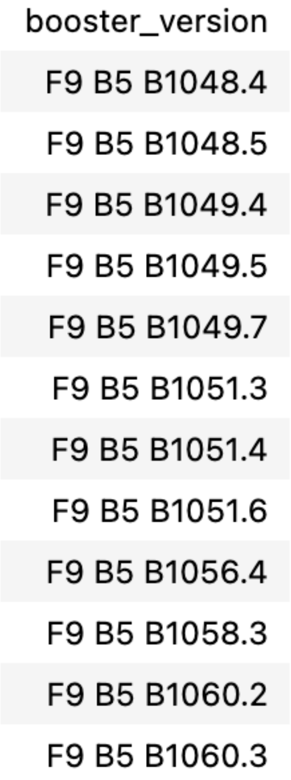


number_of_success_outcomes	number_of_failure_outcomes
100	1

RESULTS

1 SQL (EDA with SQL)

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

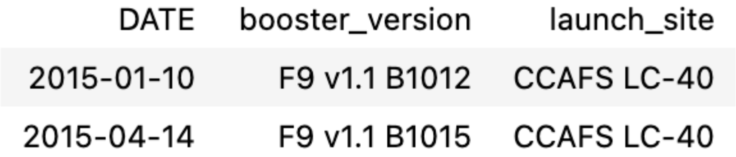


booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

RESULTS

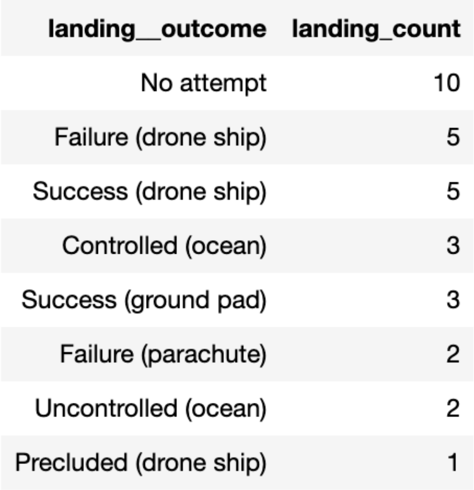
1 SQL (EDA with SQL)

- 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

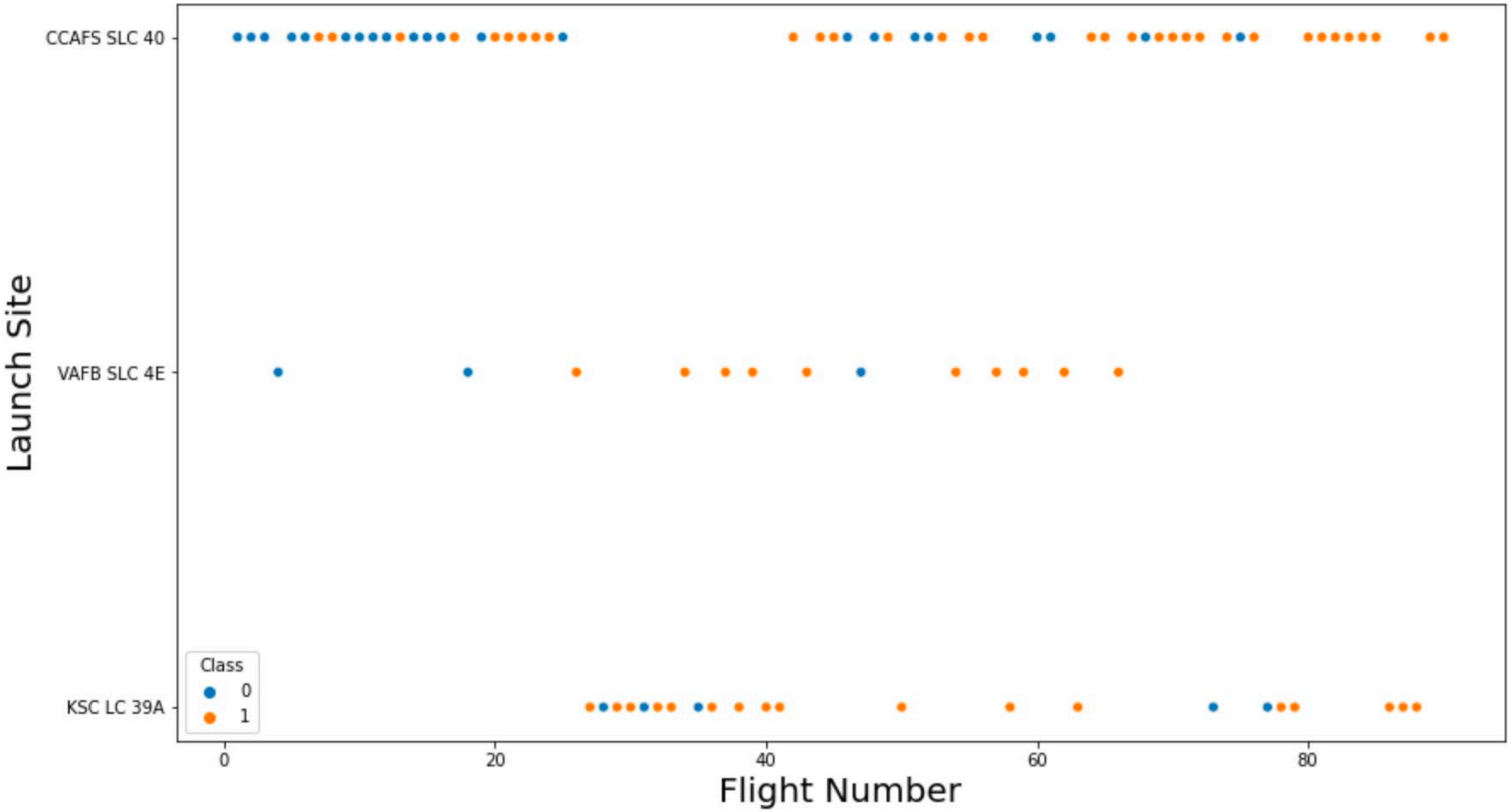


landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

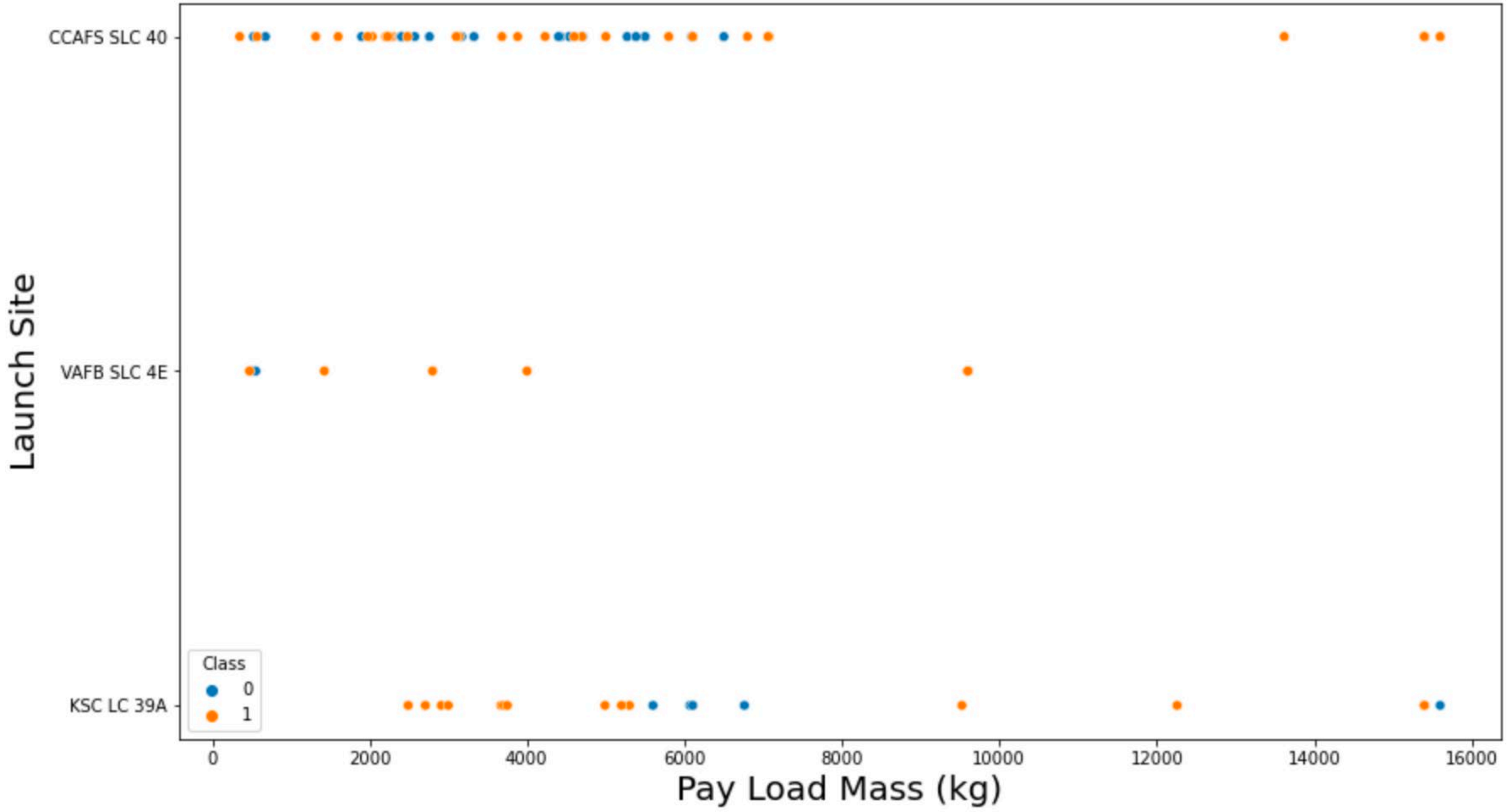
- The relationship between flight number and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

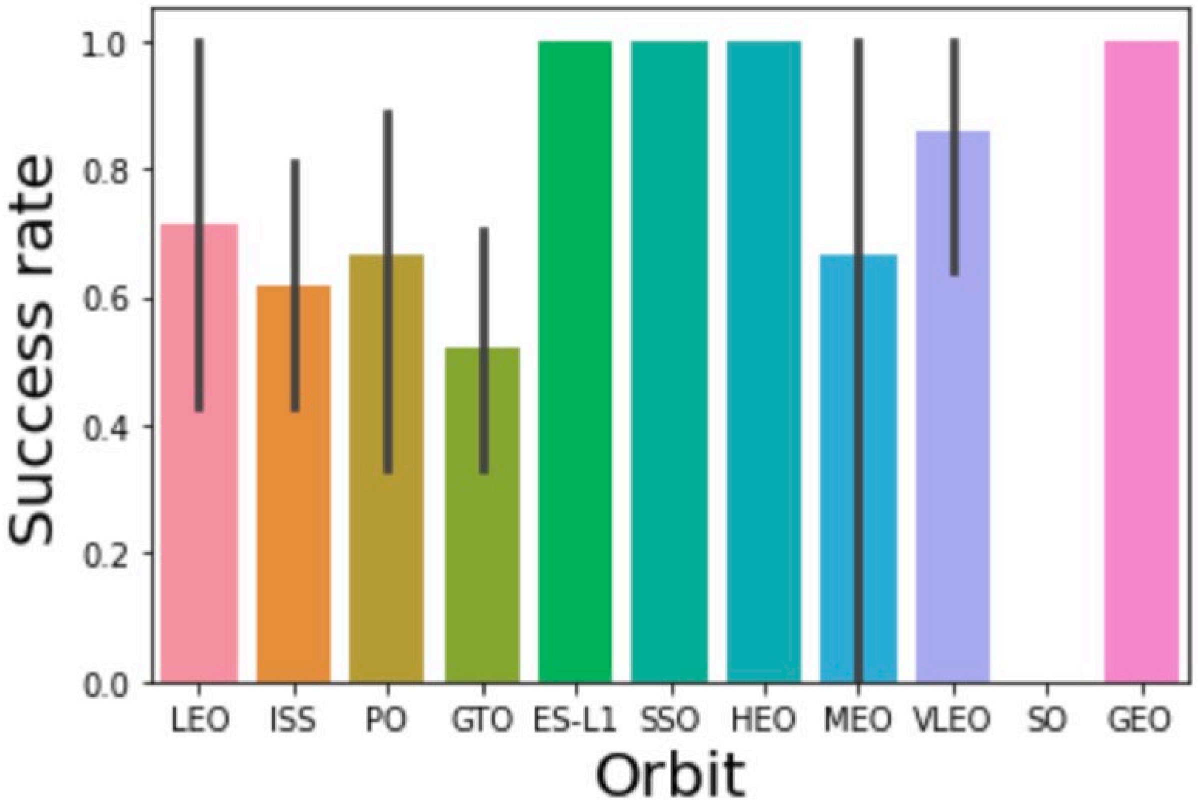
- The relationship between payload mass and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

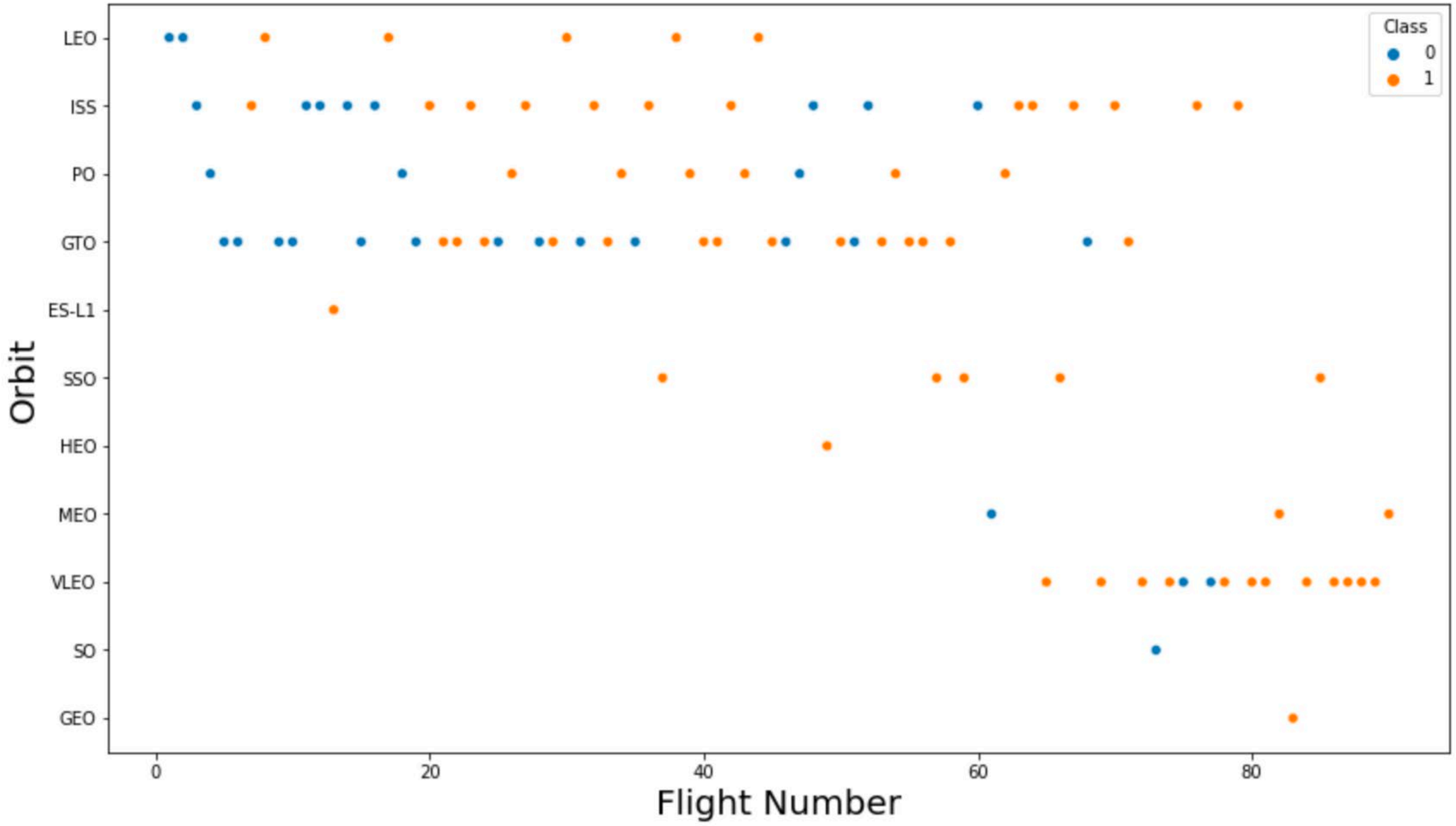
- The relationship between success rate and orbit type



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

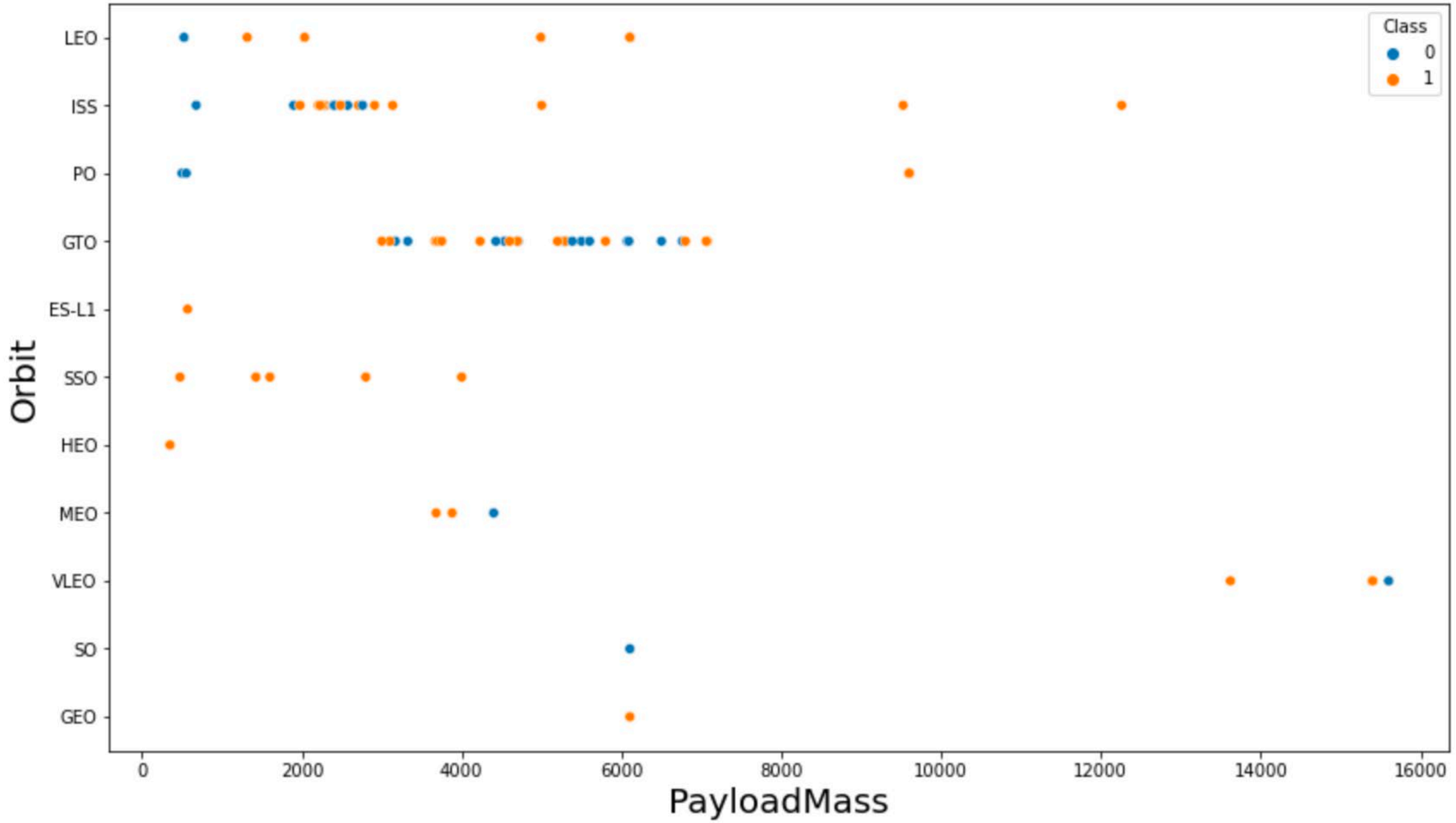
- The relationship between flight number and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

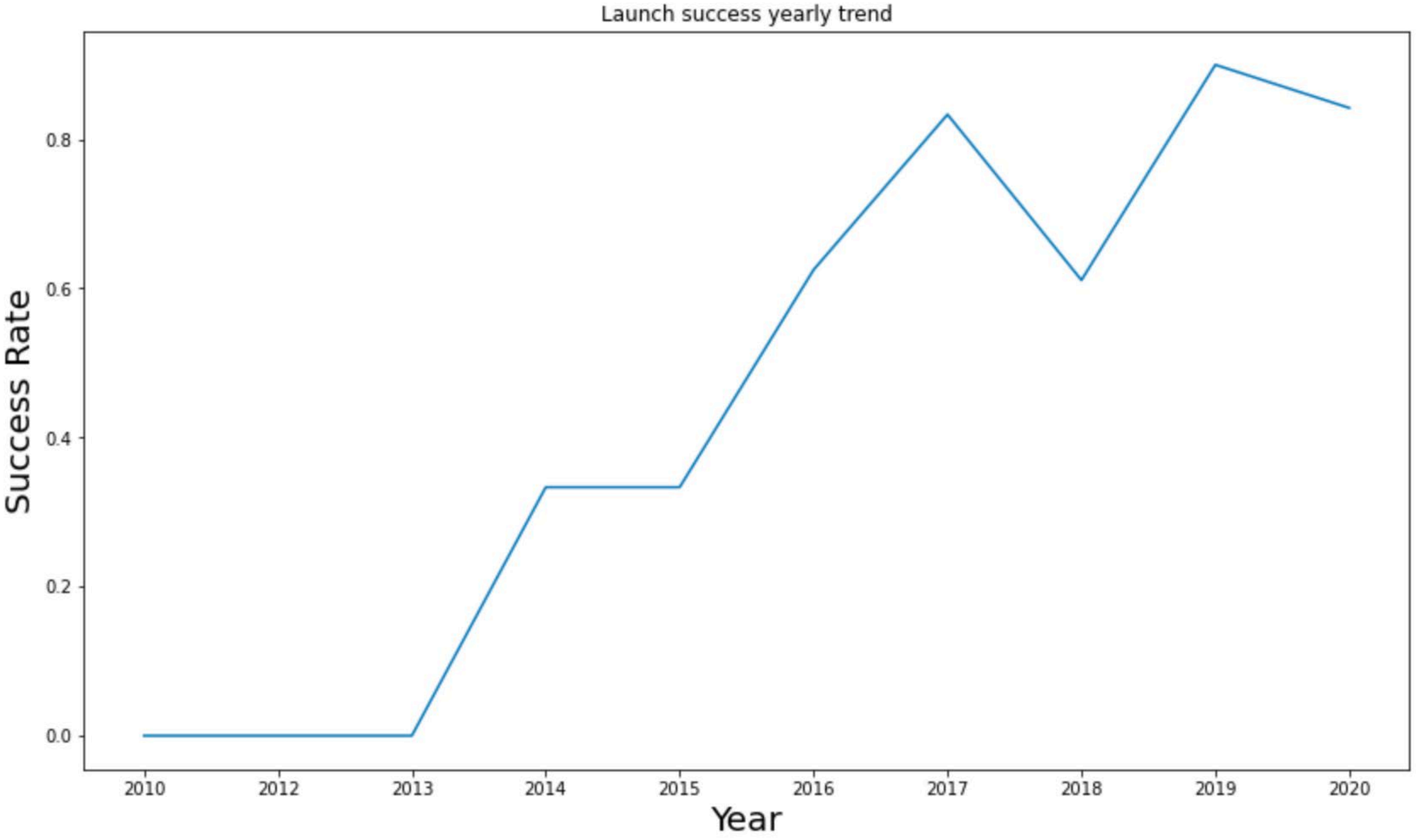
- The relationship between payload mass and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

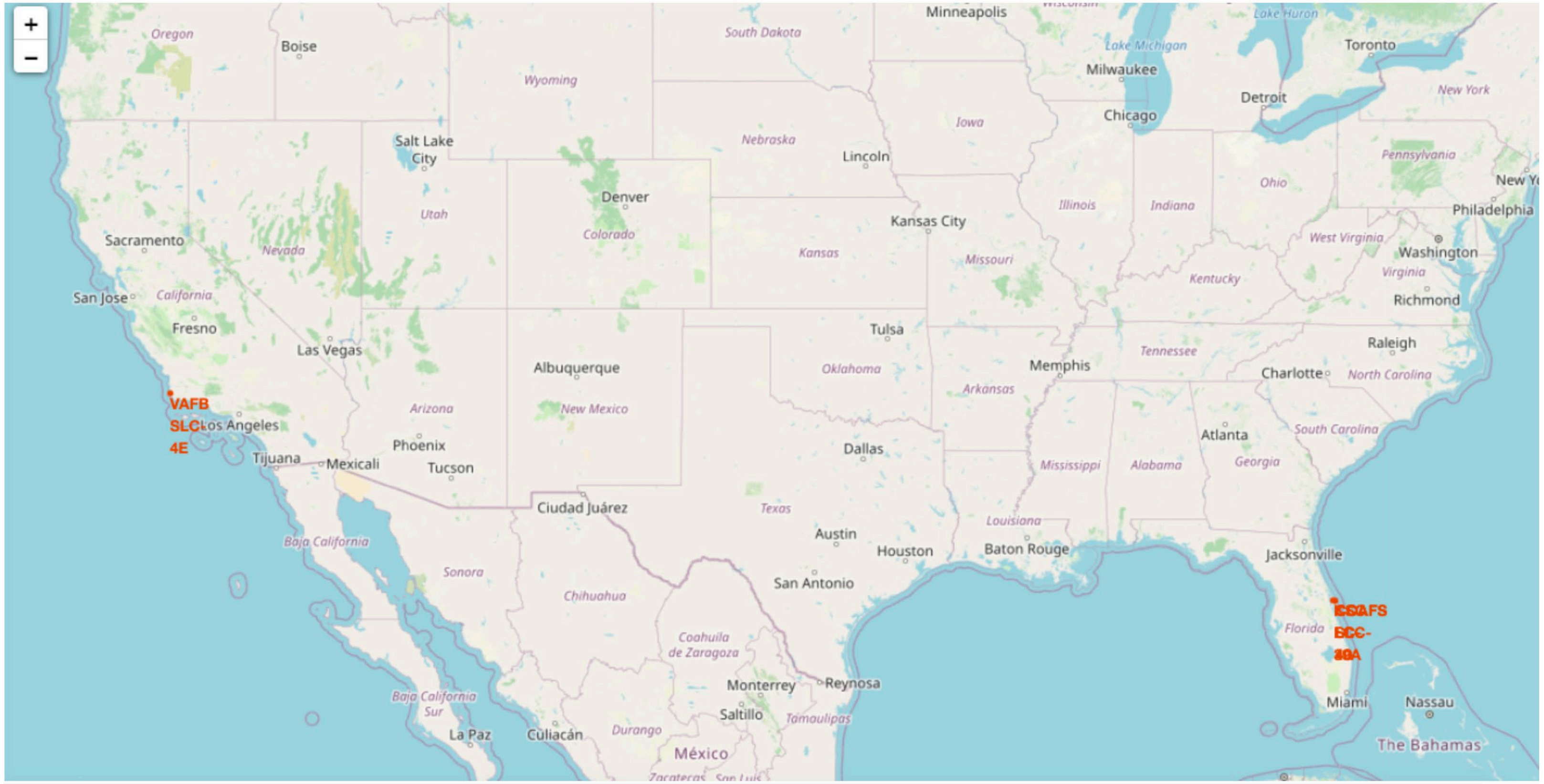
- The launch success yearly trend



RESULTS

3 Folium

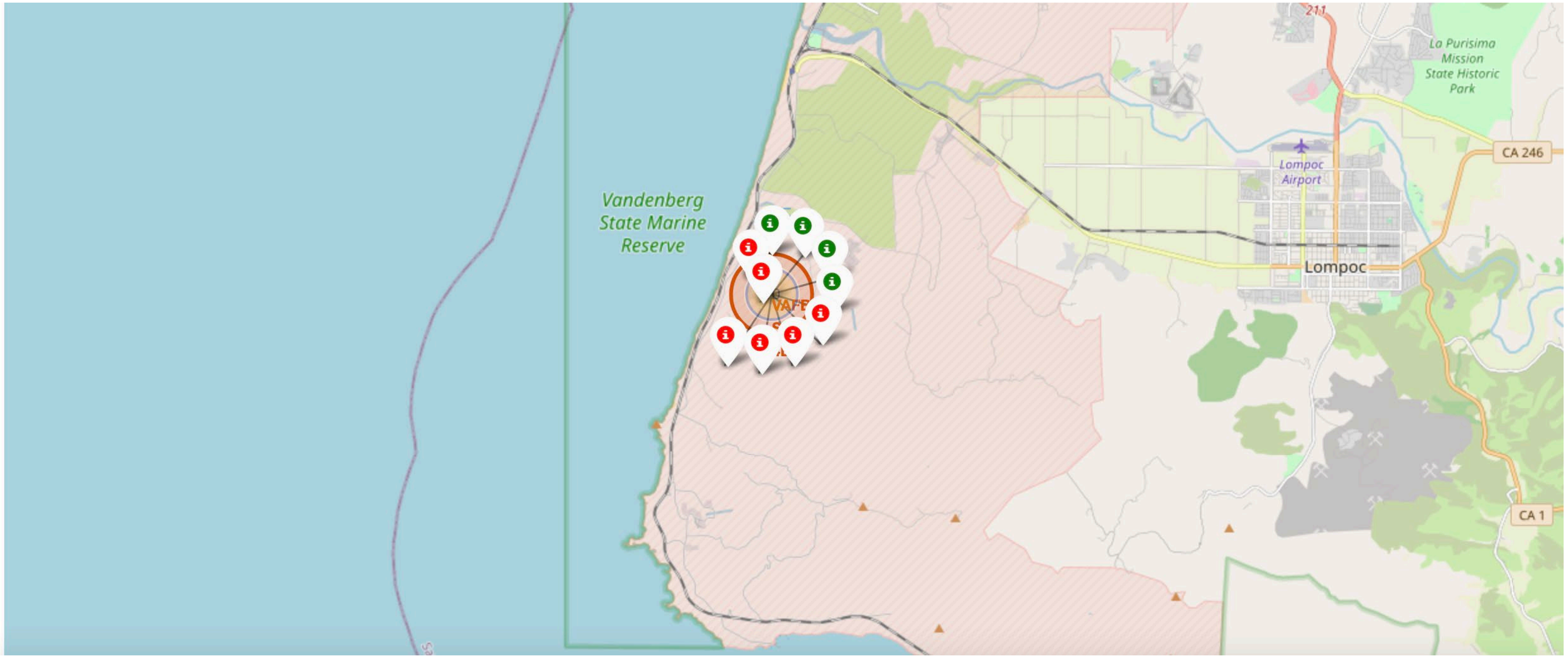
- All launch sites on map



RESULTS

3 Folium

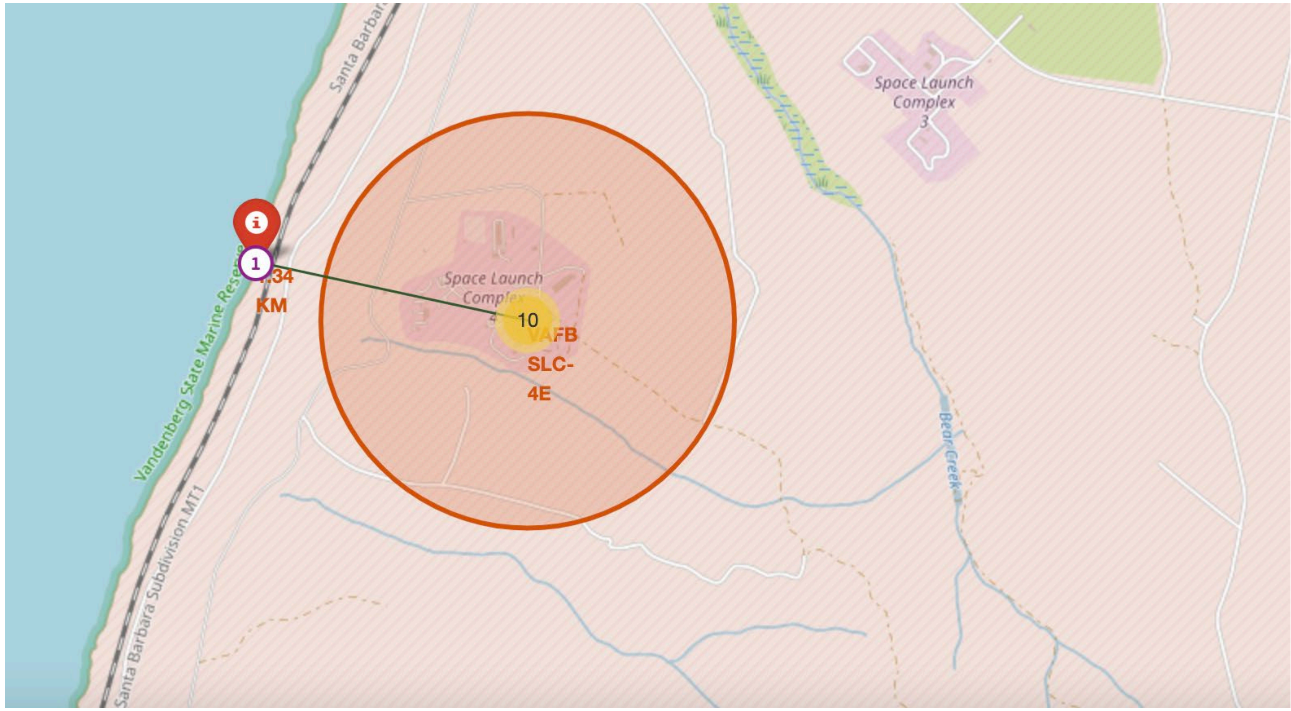
- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch


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

RESULTS

3 Folium

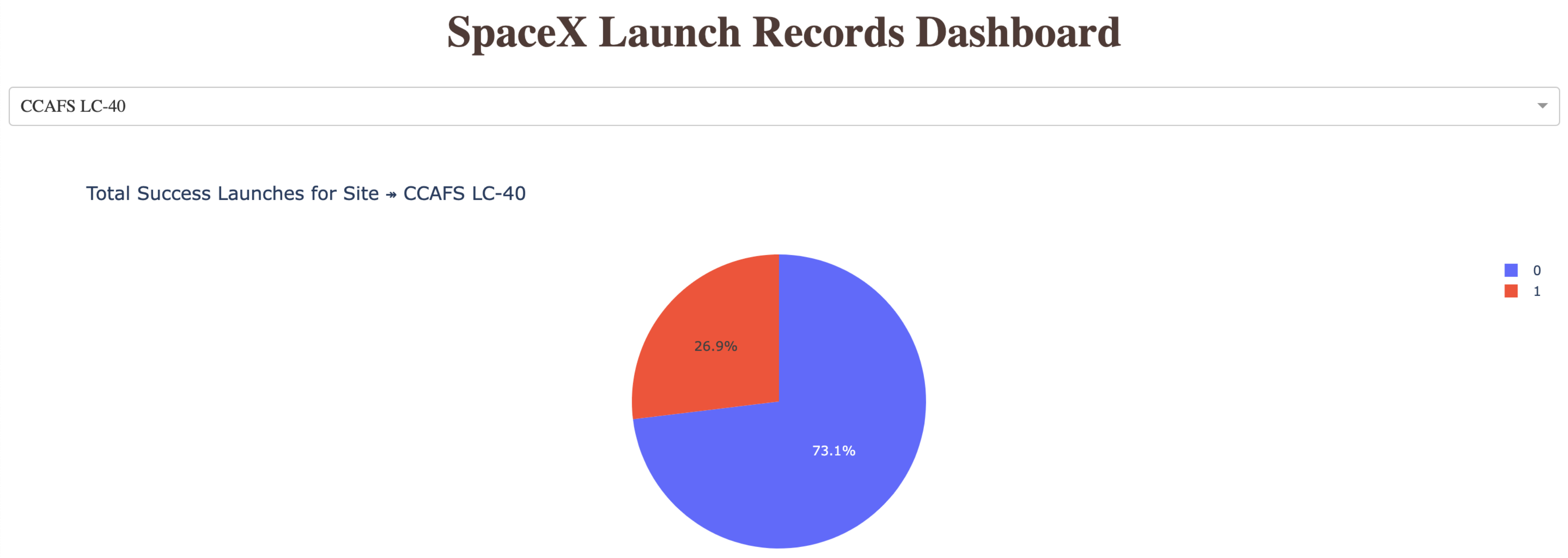
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



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

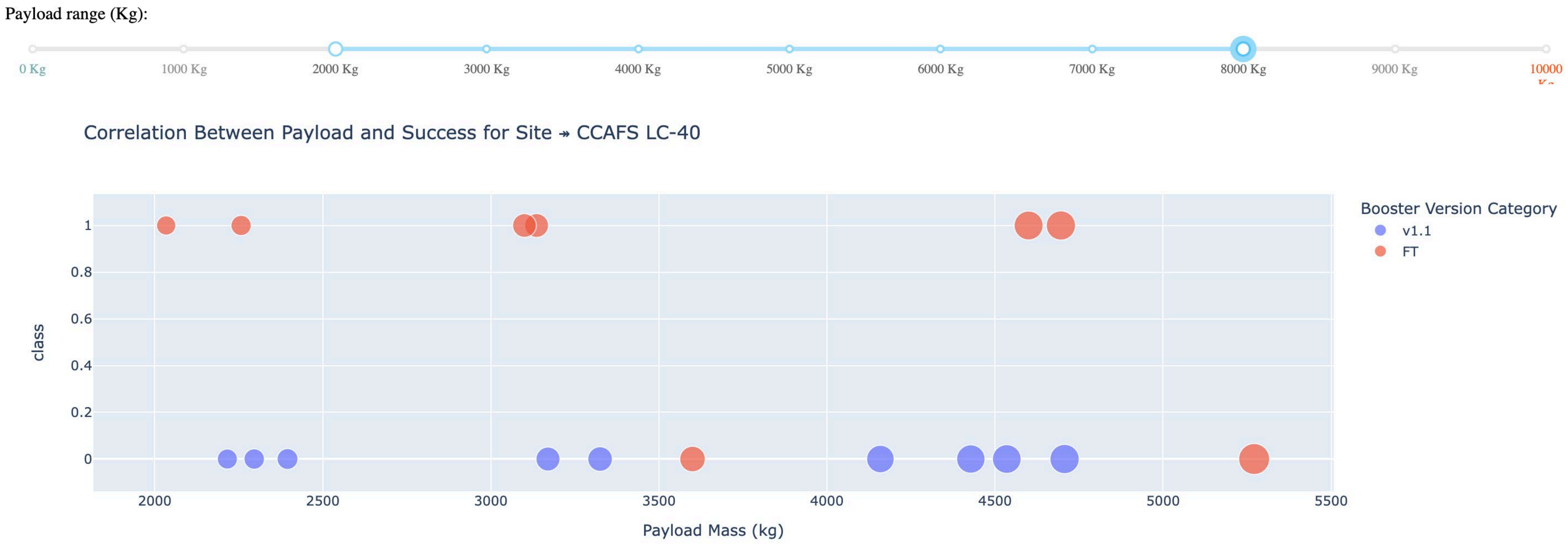
CCAFS LC-40

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.

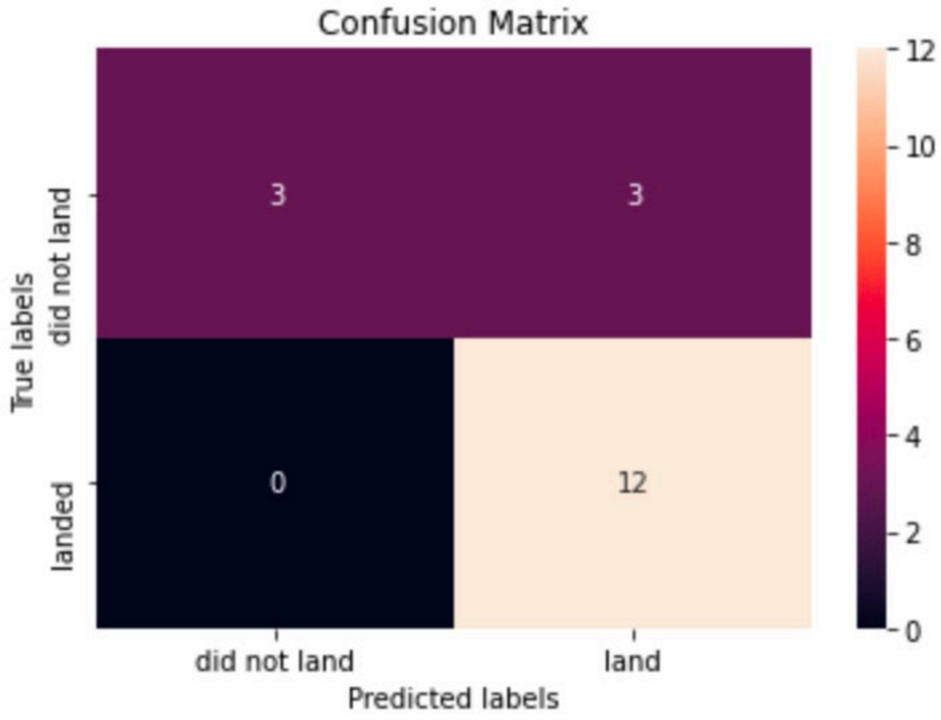


Payload range (Kg):

RESULTS

5 Predictive Analysis

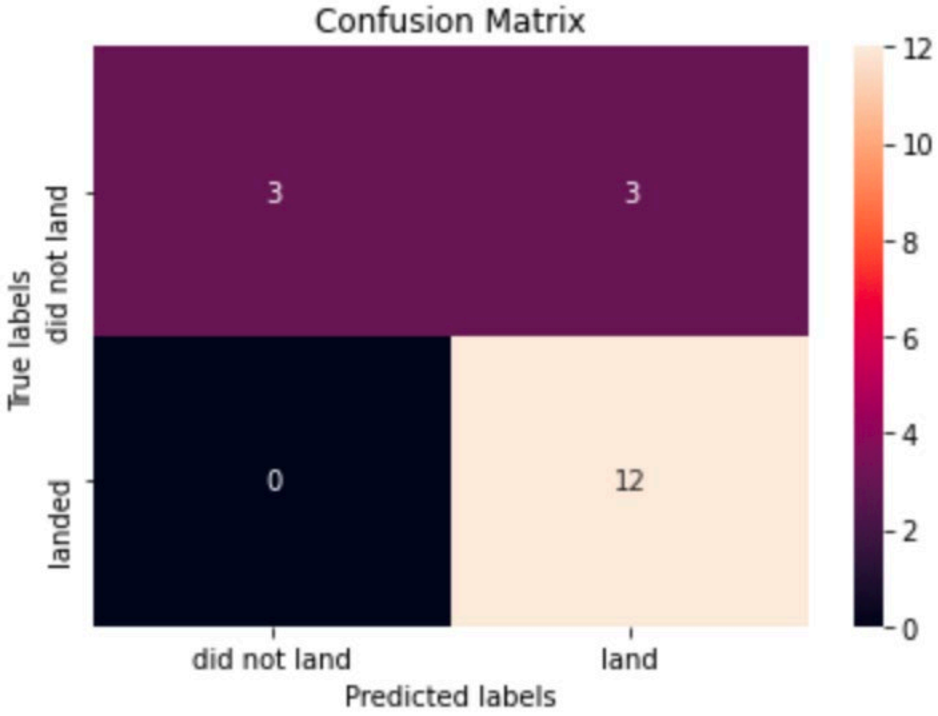
- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

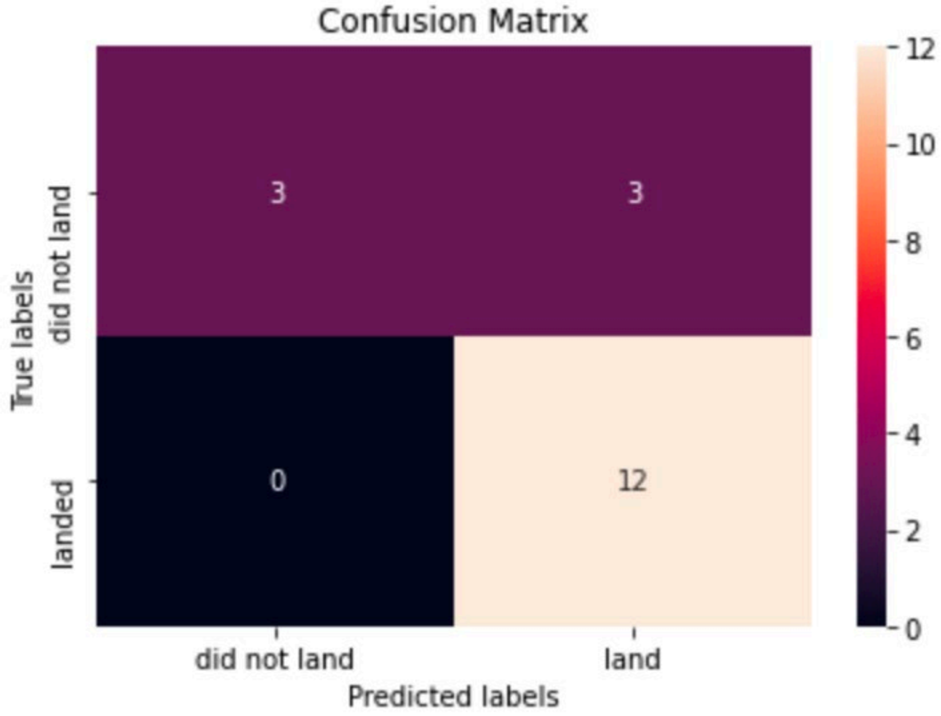
- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

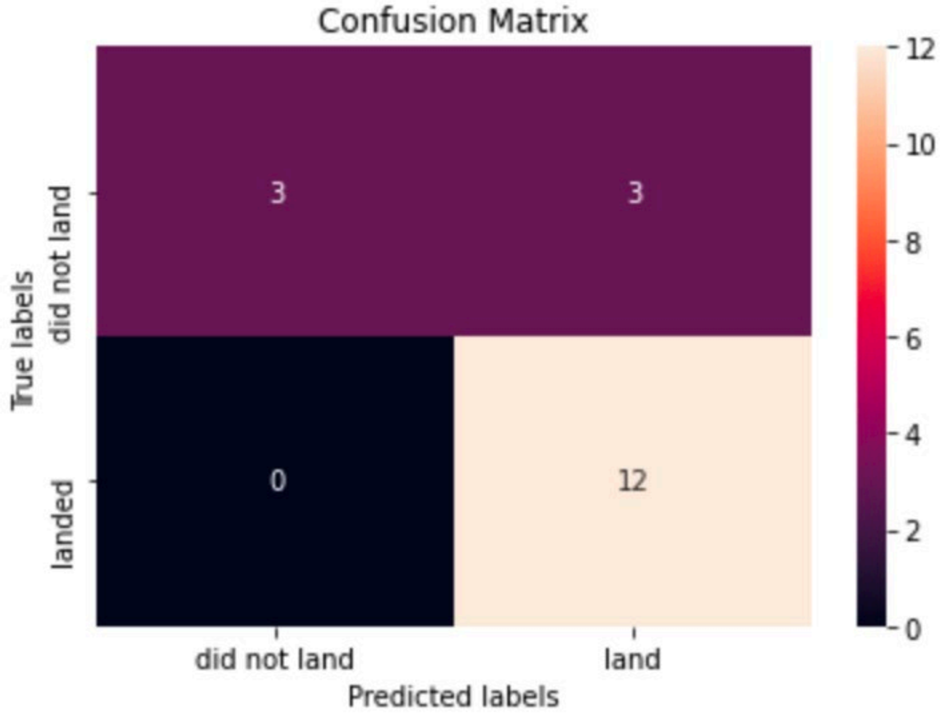
- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



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 we juxtapose the outcomes of all four models, it becomes evident that they exhibit identical accuracy scores and confusion matrices during their evaluation on the test dataset. Consequently, we resort to their GridSearchCV best scores as the determining factor for ranking these models. In accordance with the GridSearchCV best scores, the models are arranged in the following order, with the top performer listed first and the least performer listed last:

1. Decision tree (GridSearchCV best score: 0.8892857142857142)
2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
4. Logistic regression (GridSearchCV best score: 0.8464285714285713)

DISCUSSION

- In the data visualization segment, it becomes evident that certain features exhibit correlations with the mission outcome, manifesting in various patterns. For instance, when dealing with heavy payloads, we observe a higher likelihood of successful landings or positive mission outcomes for specific orbit types, namely Polar, LEO, and ISS. However, when it comes to GTO, distinguishing between these outcomes is more challenging, as both positive landing rates and negative outcomes (unsuccessful missions) coexist.
- As a result, each feature within the dataset appears to exert a distinct influence on the ultimate mission outcome. While deciphering the precise mechanisms by which these features impact mission success can be intricate, we can leverage machine learning algorithms to analyze historical data patterns. By doing so, we can predict whether a mission will achieve success or not based on the provided features.

CONCLUSION

- This project's primary objective is to forecast whether the first stage of a Falcon 9 launch will achieve a successful landing, enabling us to ascertain the cost associated with the launch. Each feature associated with a Falcon 9 launch, be it payload mass or orbit type, is believed to exert a distinct influence on the mission's outcome.
- To achieve this, we deploy a variety of machine learning algorithms to discern patterns within historical Falcon 9 launch data. These algorithms generate predictive models capable of anticipating the outcome of future Falcon 9 launches.
- Among the four machine learning algorithms utilized, the decision tree algorithm emerged as the top performer, yielding the most accurate predictive model.