IBM APPLIED DATA SCIENCE CAPSTONE

PREM KUMAR S

JANUARY 22, 2025

OUTLINE

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

EXECUTIVE SUMMARY

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.
- It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

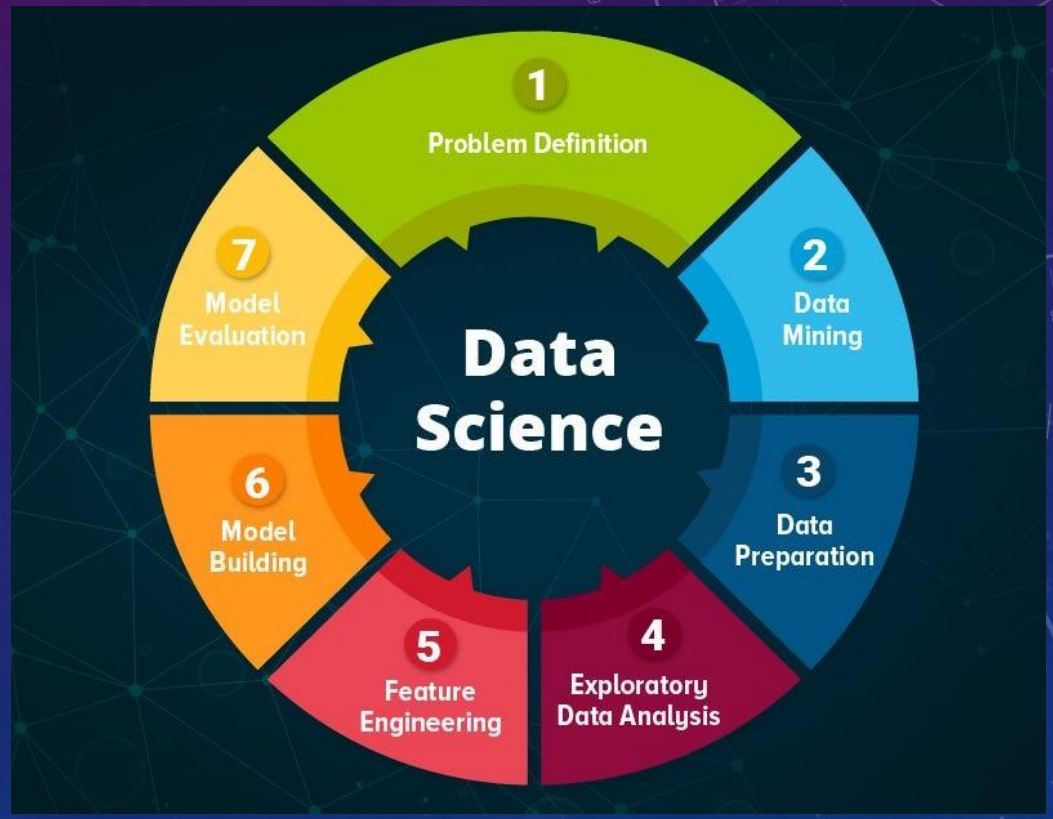
INTRODUCTION

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

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)

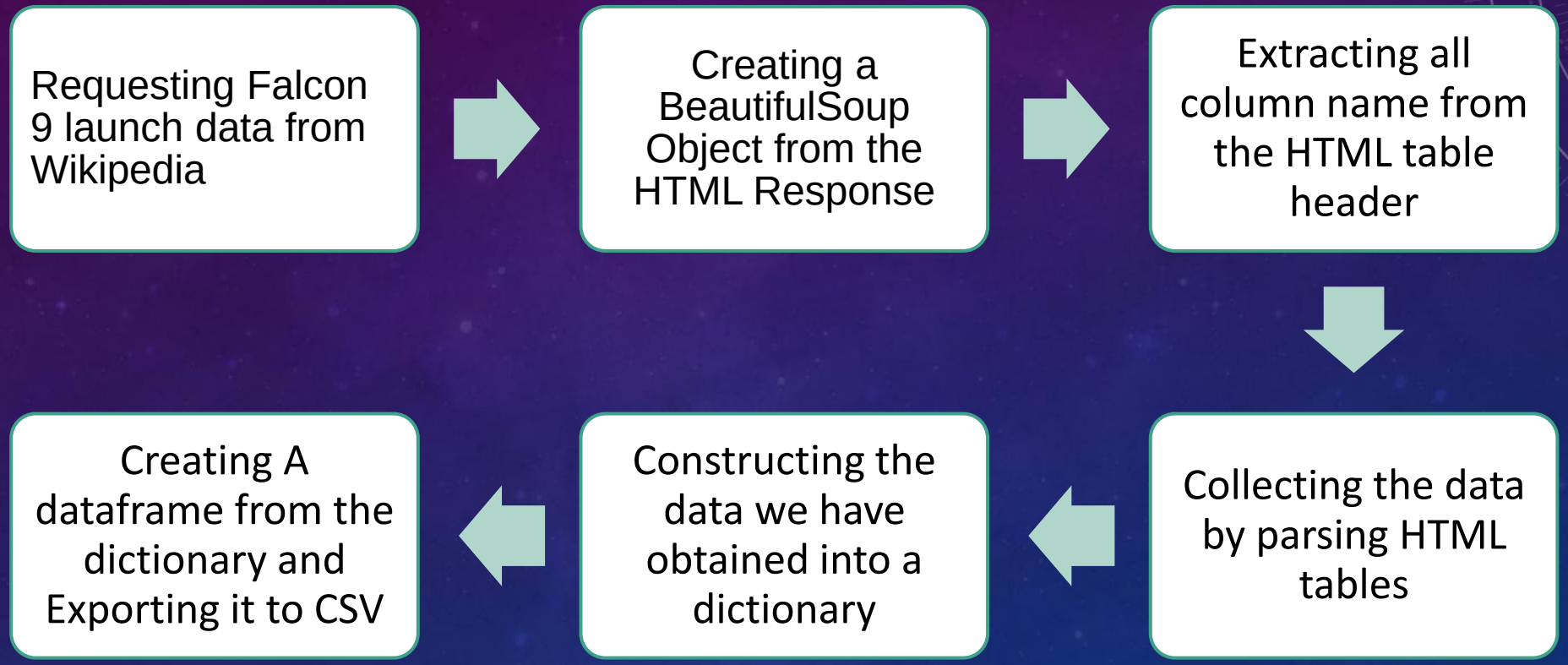


METHODOLOGY

1. DATA COLLECTION - SPACEX API

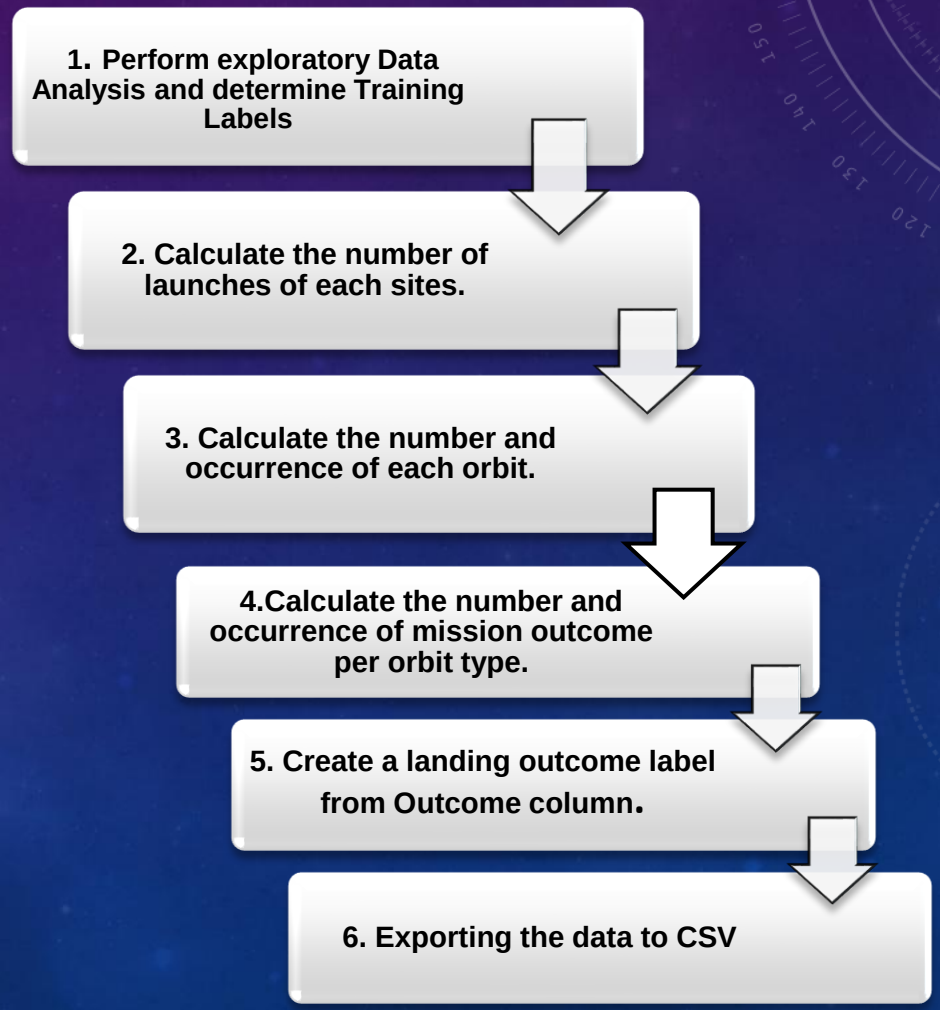
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- Requesting rocket launch data from SpaceX API
- Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
- Requesting needed information about the launches from SpaceX API by applying custom functions.
- Constructing data we have obtained into a dictionary.
- Creating a dataframe from the dictionary.
- Filtering the dataframe to only include Falcon 9 launches.
- Replacing missing values of Payload Mass column with calculated .mean() for this column.
- Exporting the data to CSV

DATA COLLECTION – WEB SCRAPING



DATA WRANGLING AND FORMATTING

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.



2. EXPLORATORY DATA ANALYSIS (EDA)

EDA with data visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series)

EDA WITH SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

3. BUILD AN INTERACTIVE MAP WITH FOLIUM

- Markers of all Launch Sites: -

  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

- Coloured Markers of the launch outcomes for each Launch Site: -

  Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- Distances between a Launch Site to its proximities: -

  Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

BUILD A DASHBOARD WITH PLOTLY DASH

- Launch Sites Dropdown List: -

  Added a dropdown list to enable Launch Site selection.

- Pie Chart showing Success Launches (All Sites/Certain Site): -

  Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

- Slider of Payload Mass Range: -

  Added a slider to select Payload range.

- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions: -

  Added a scatter chart to show the correlation between Payload and Launch Success.

4. MACHINE LEARNING
PREDICTIVE ANALYSIS (CLASSIFICATION)

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

- 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
45596

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

AVG_payload_mass
2928.4

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

RESULTS
1. SQL (EDA WITH SQL)

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

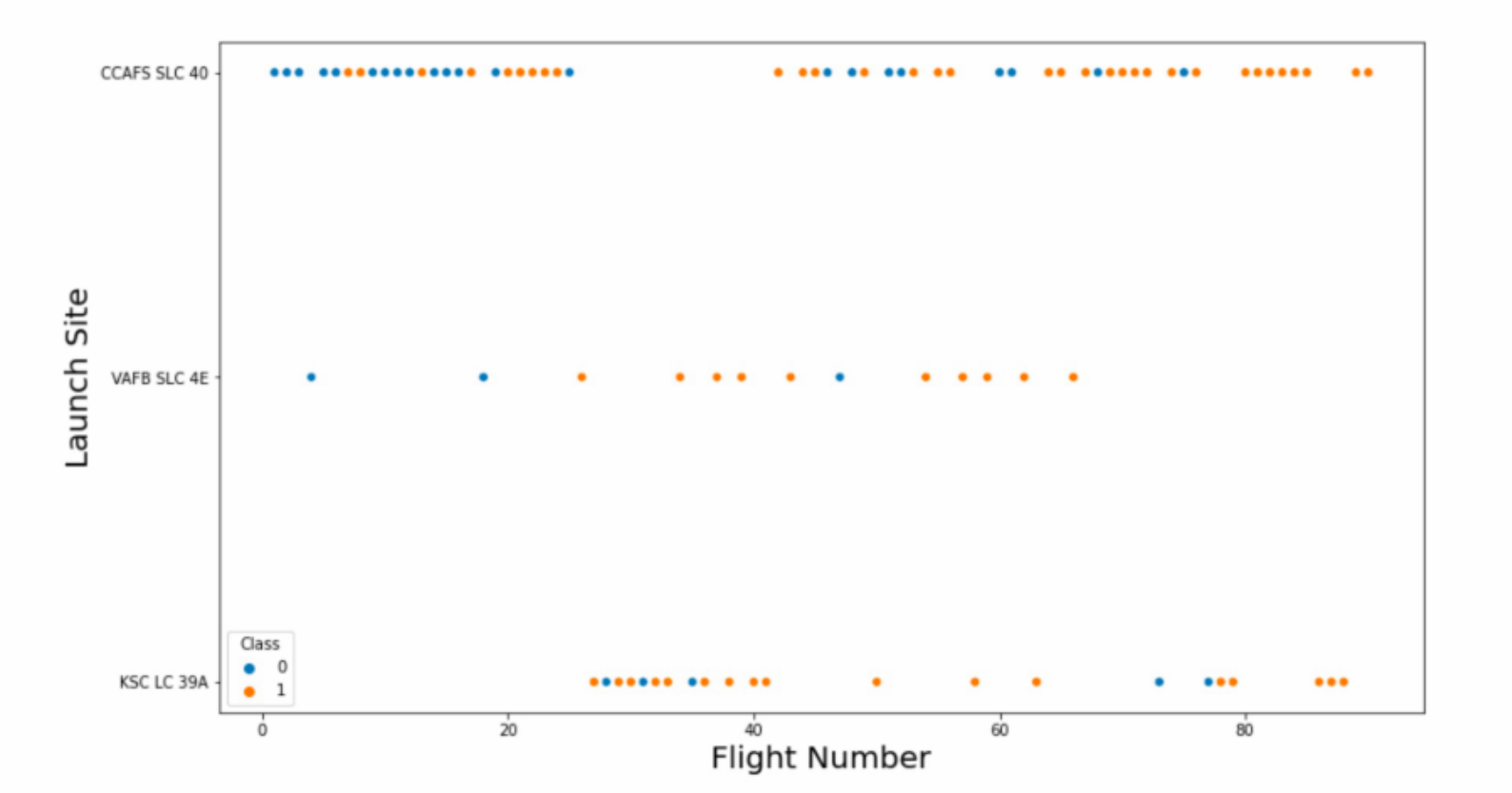
MonthName	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

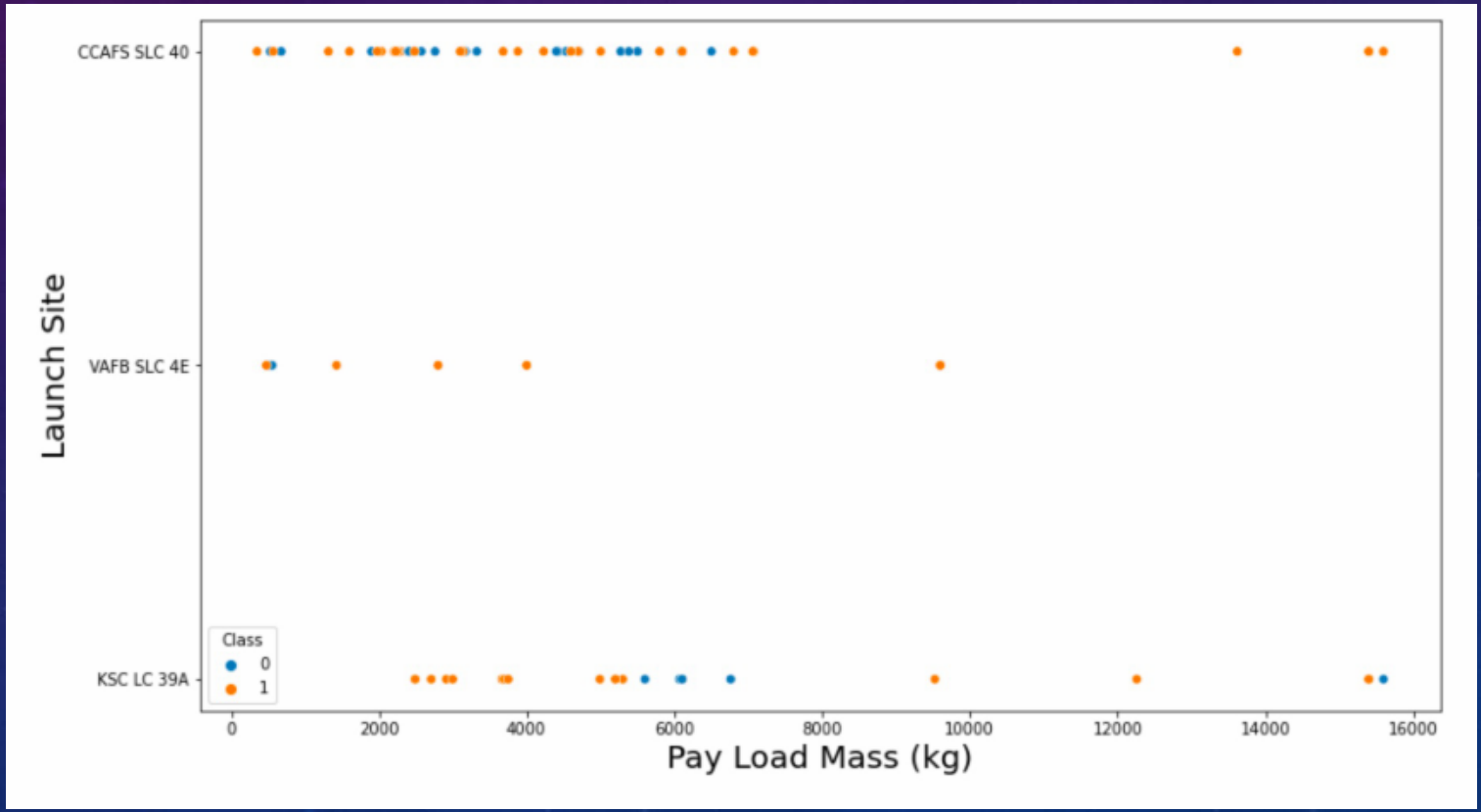
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The relationship between flight number and launch site



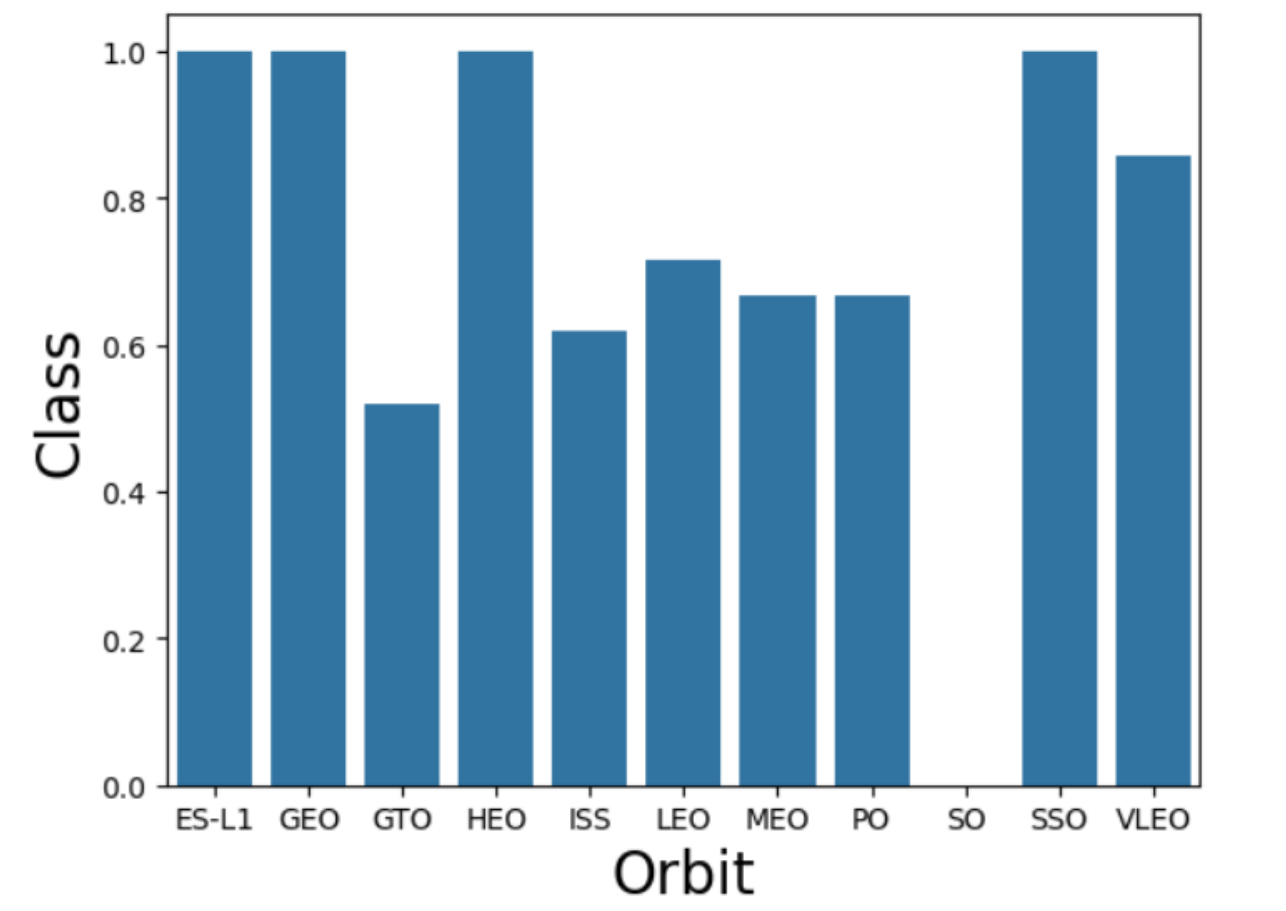
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The relationship between payload mass and launch site



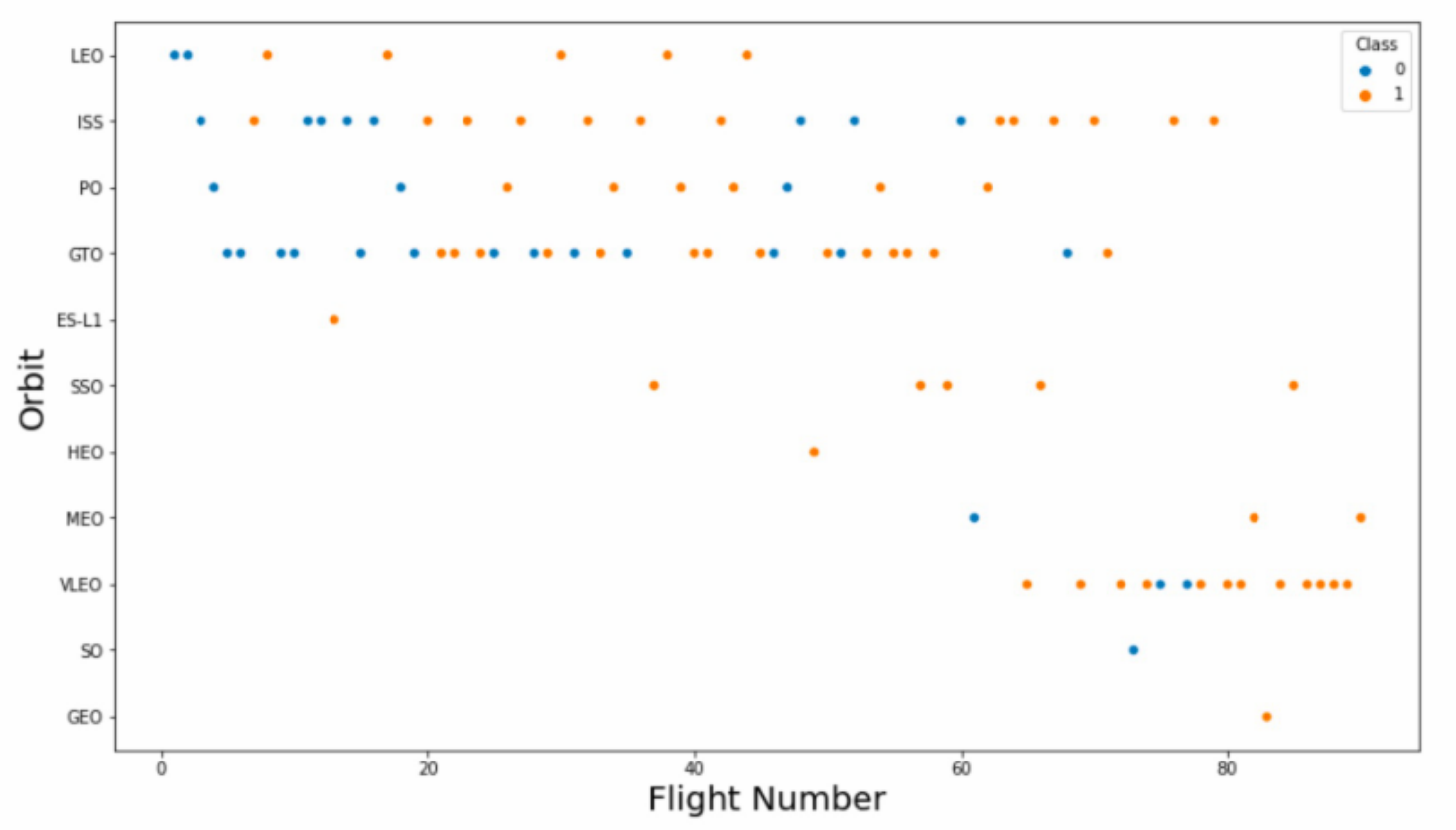
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The relationship between success rate and orbit type



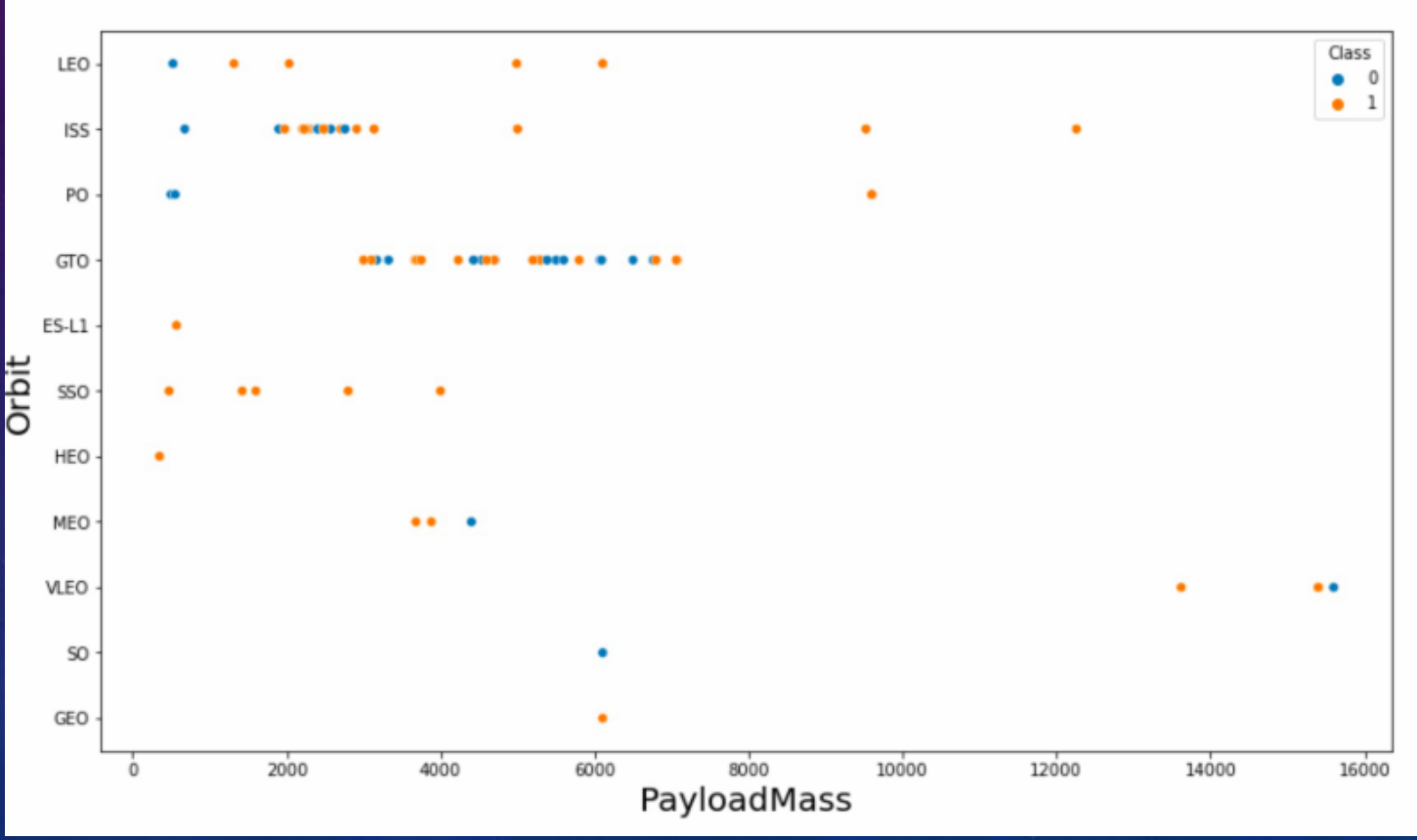
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The relationship between flight number and orbit type



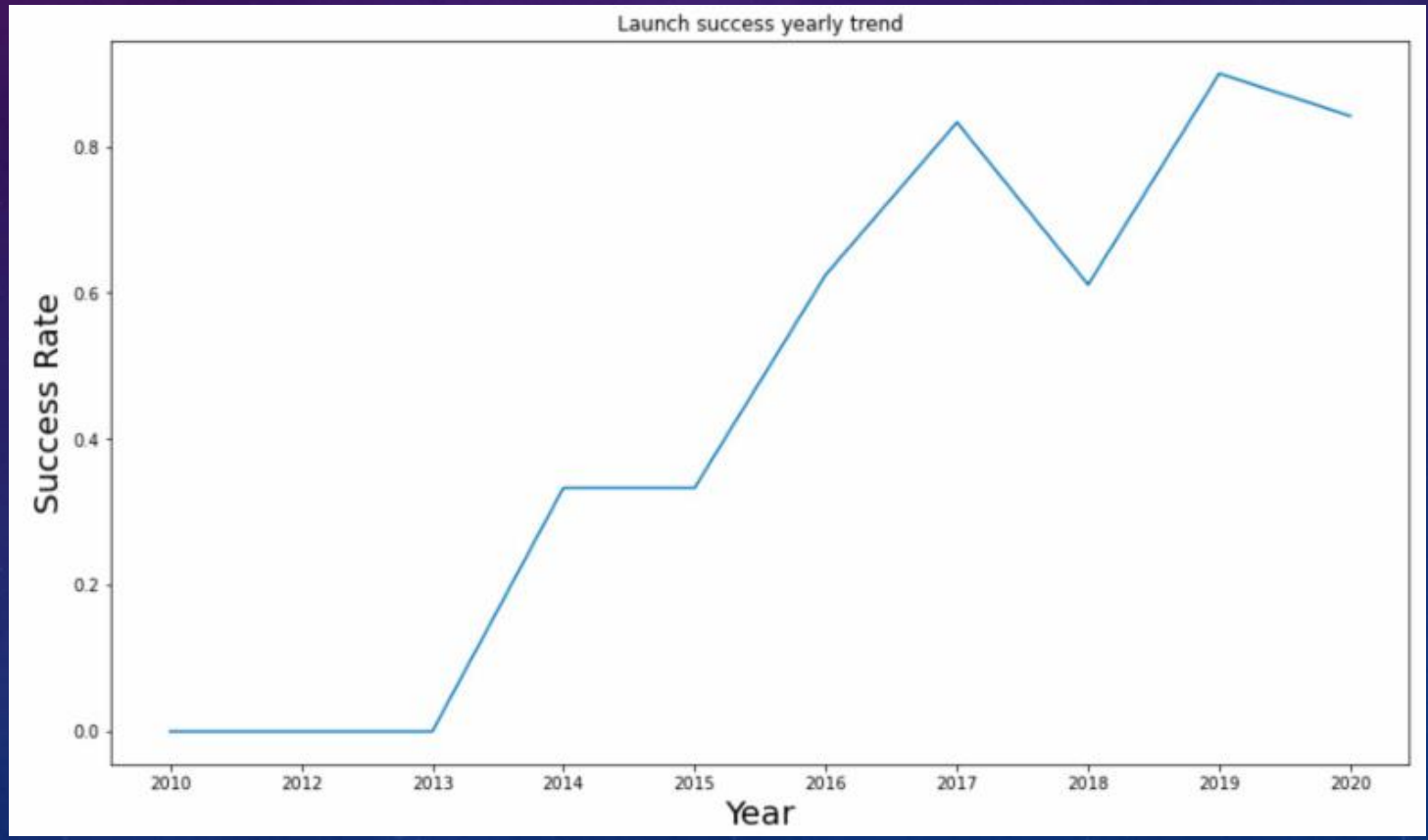
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The relationship between payload mass and orbit type



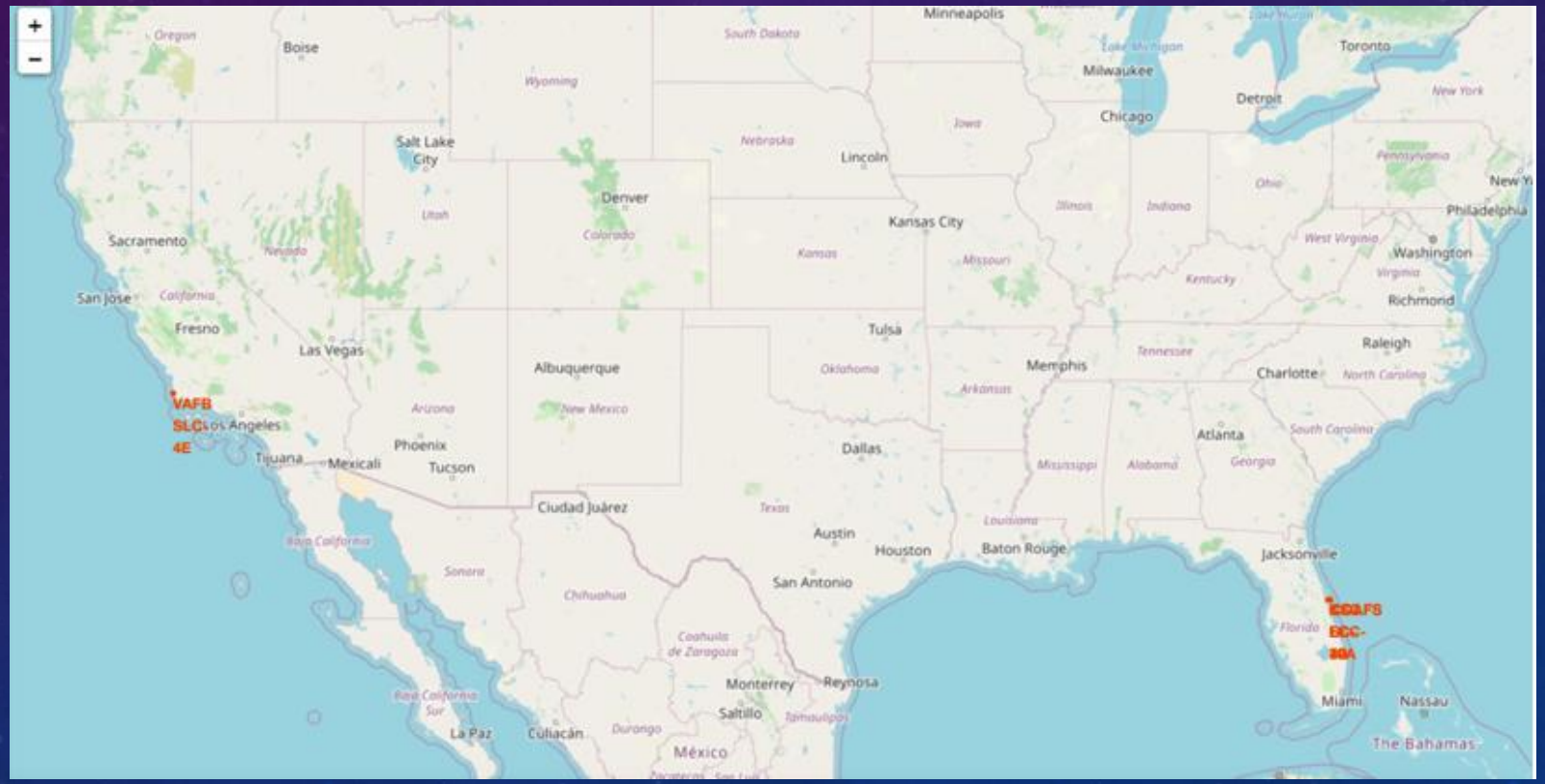
RESULTS
2. MATPLOTLIB AND SEABORN (EDA WITH VISUALIZATION)

- The launch success yearly trend



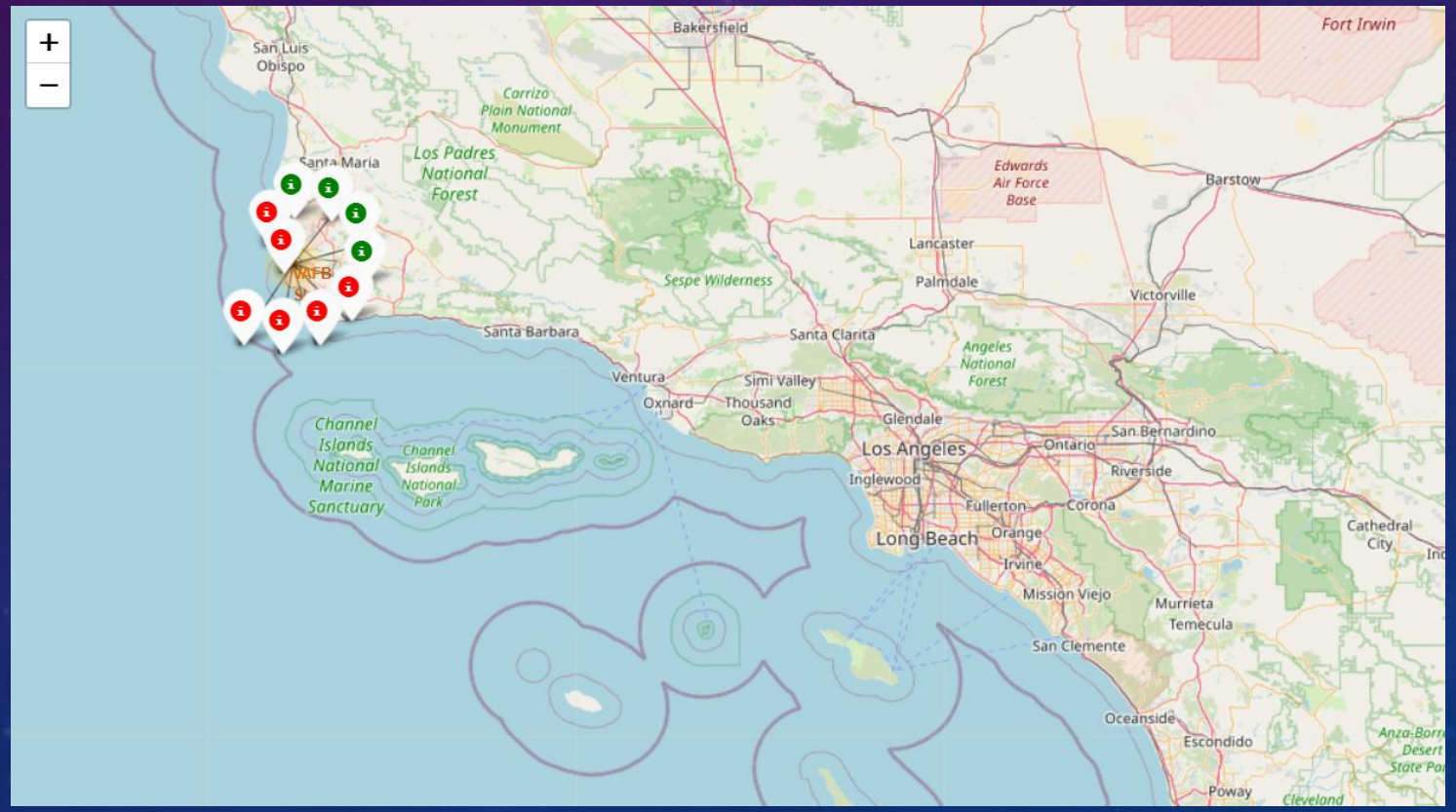
RESULTS
3. FOLIUM

- All launch sites on map



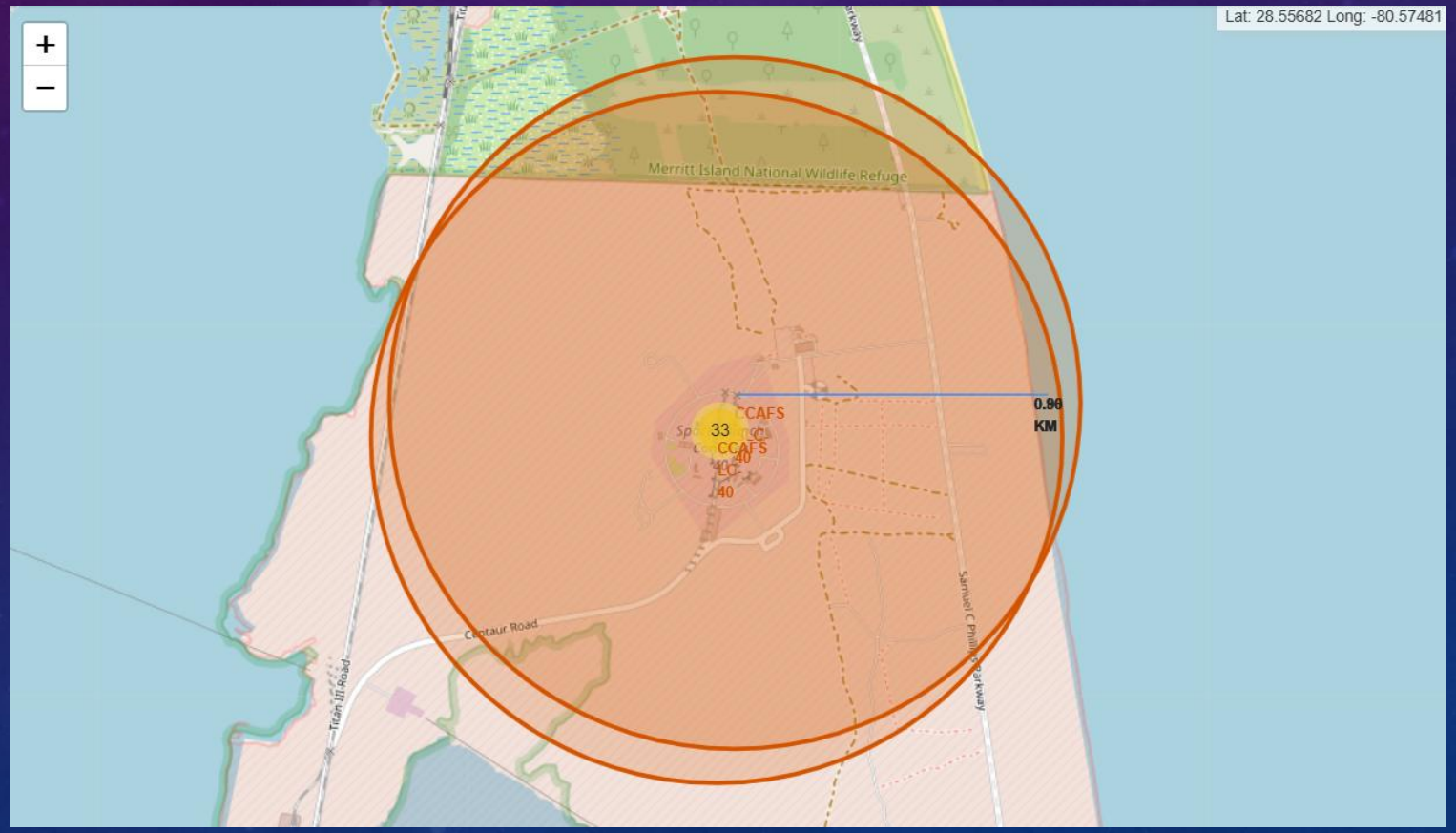
RESULTS
3. FOLIUM

- 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



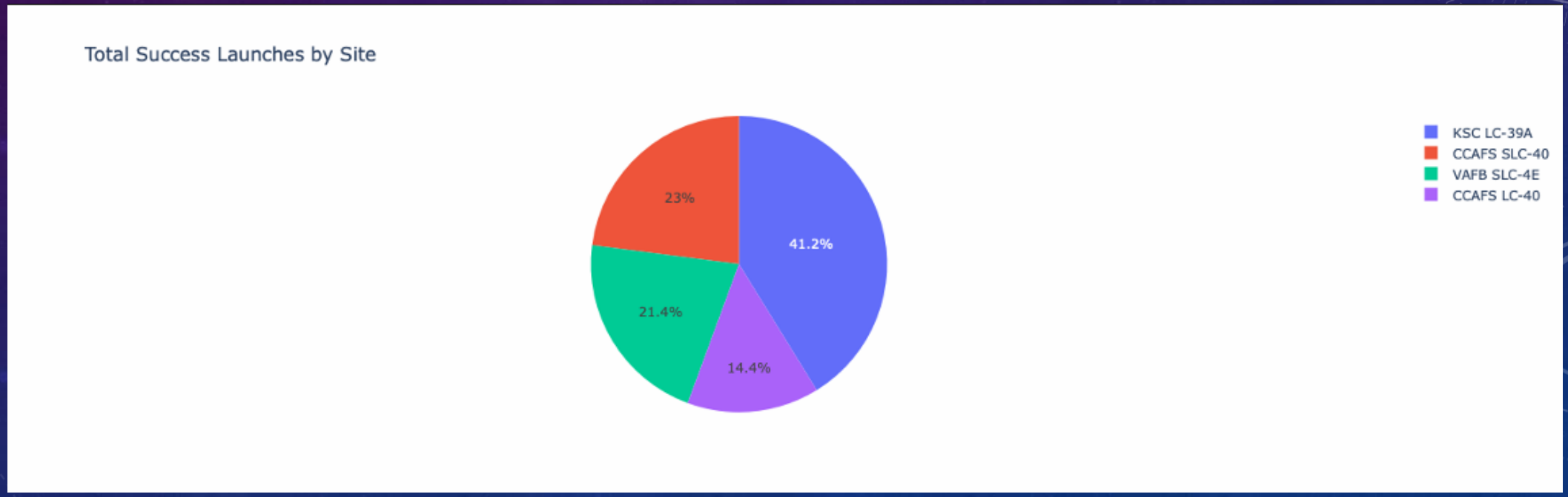
RESULTS
3. FOLIUM

- 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.



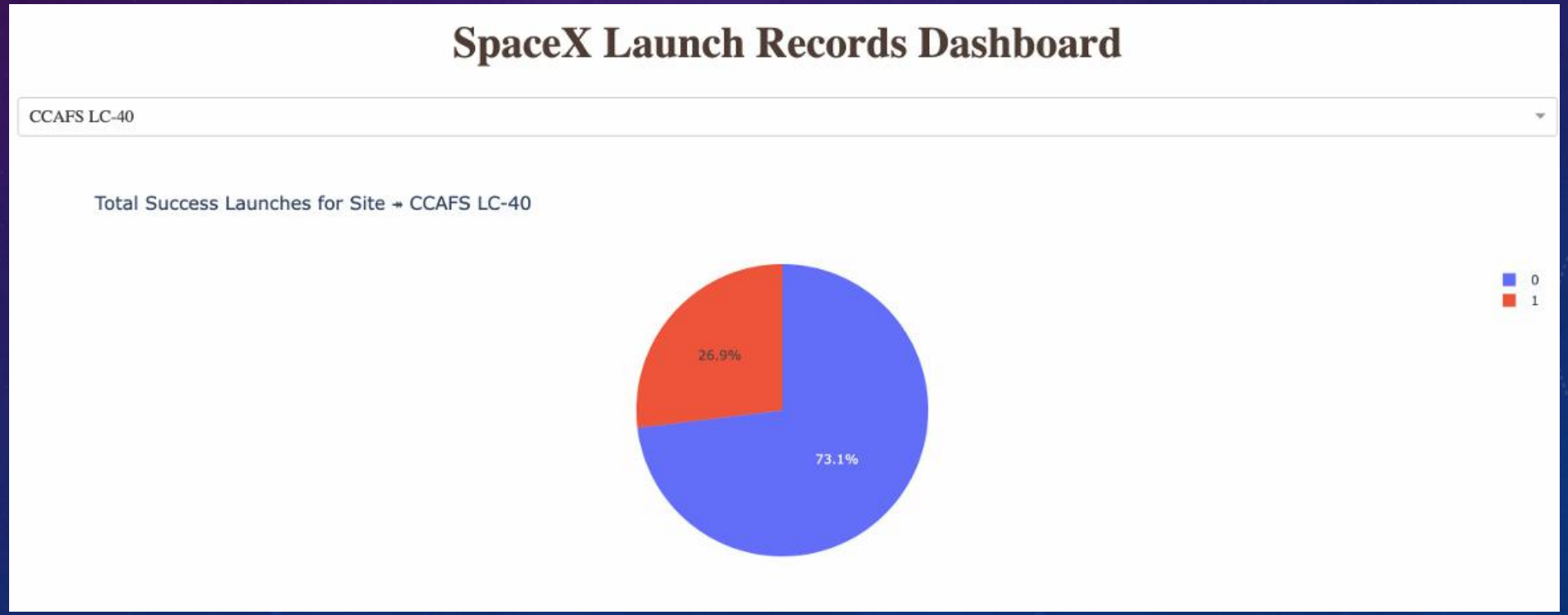
RESULTS
3. DASH

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



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.



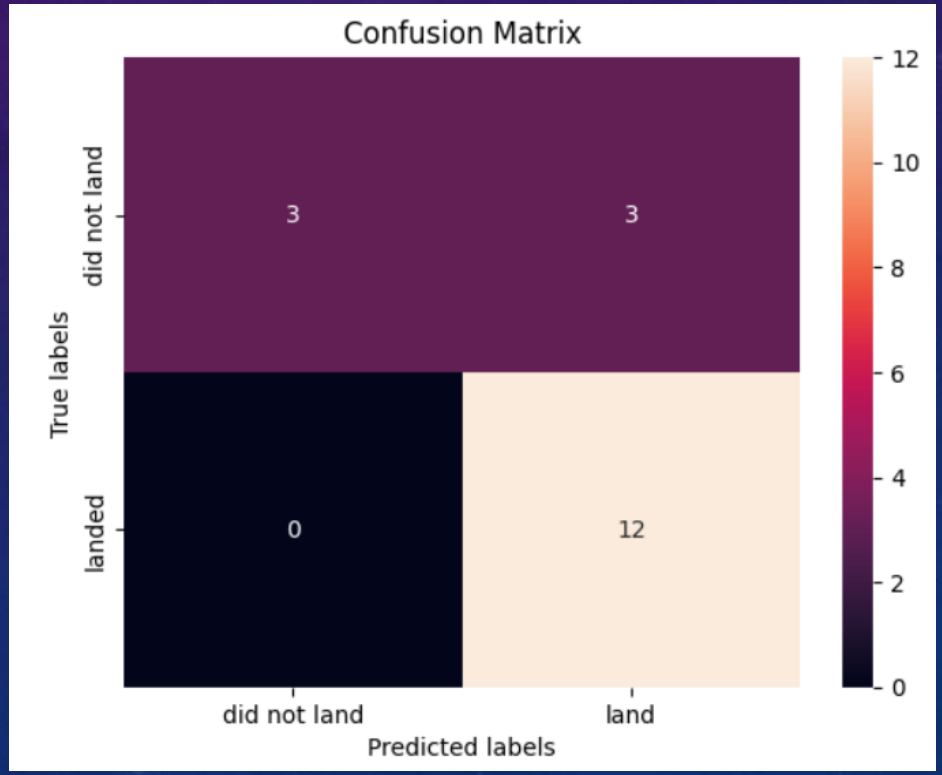
RESULTS

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.

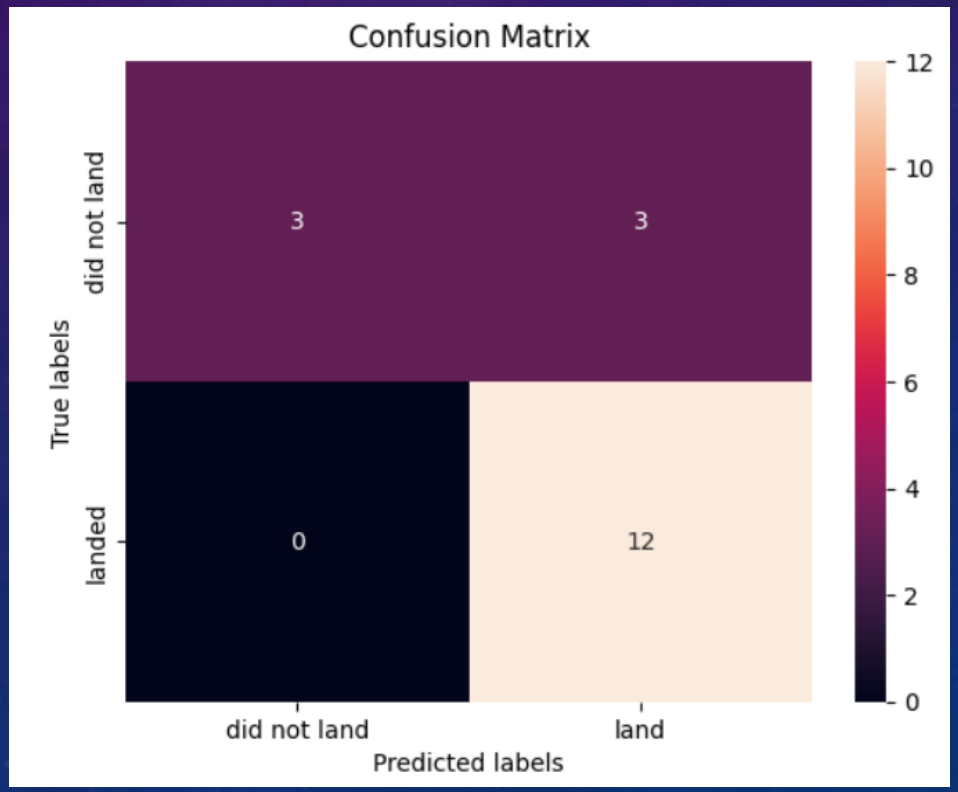
RESULTS
5. PREDICTIVE ANALYSIS

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



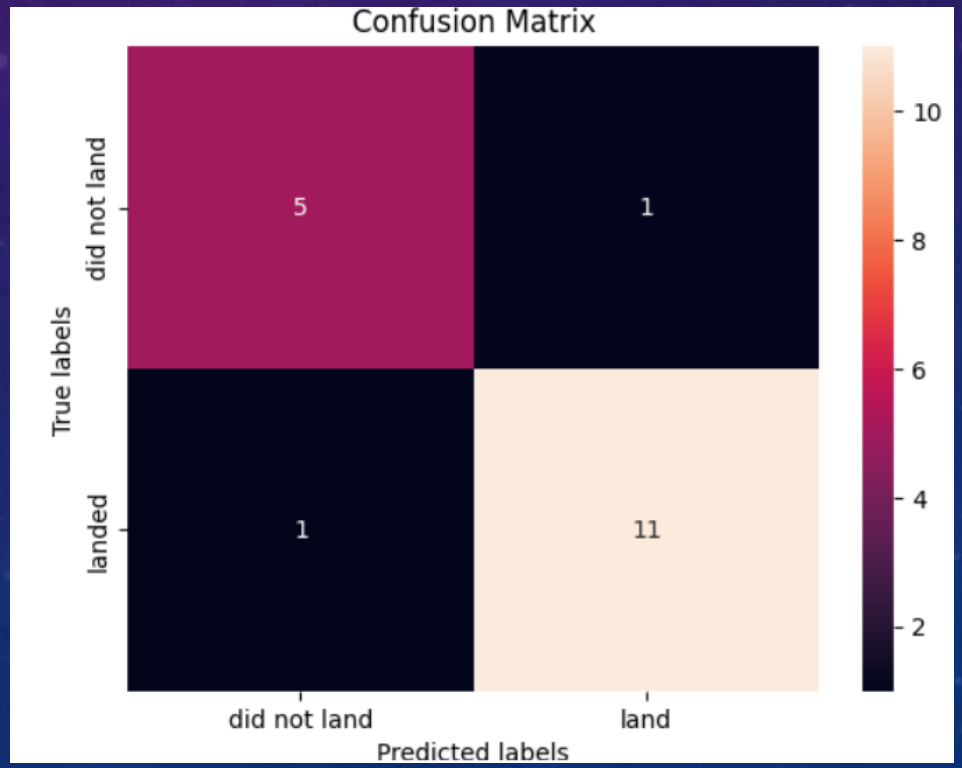
RESULTS
5. PREDICTIVE ANALYSIS

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



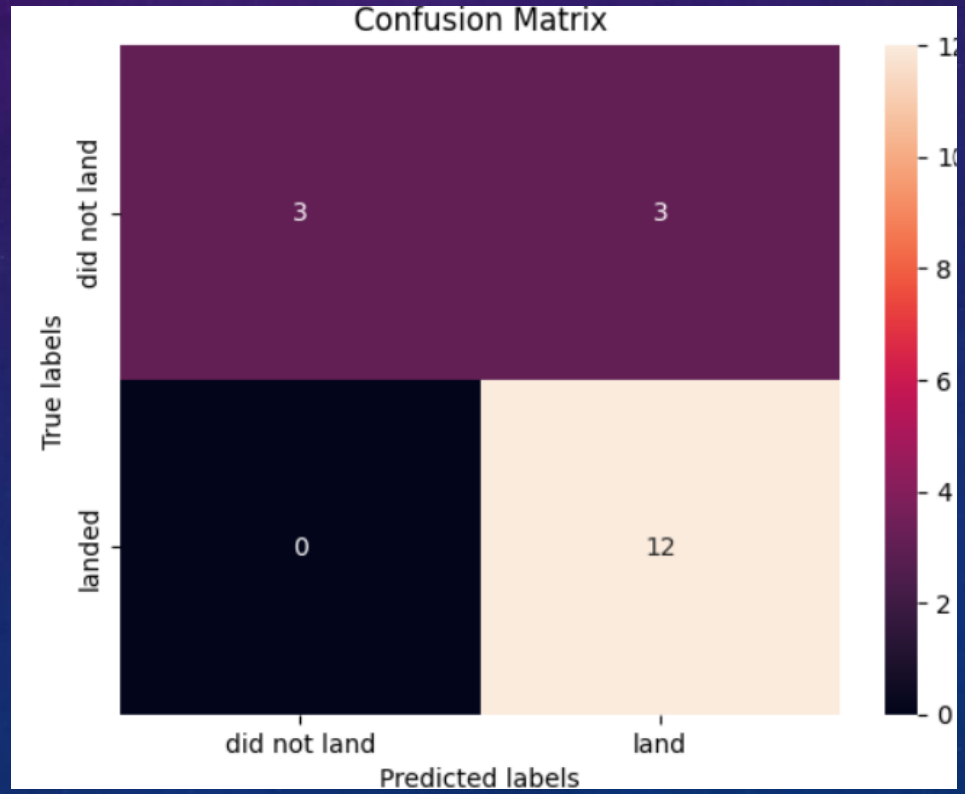
RESULTS
5. PREDICTIVE ANALYSIS

- Decision tree
  - GridSearchCV best score: 0.876785714285714 3
  - Accuracy score on test set: 0.888888888888888
  - 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

- Putting the results of all 4 models side by side, Decision Tree Model has the best accuracy score for this dataset. we can see other 3 models share the same accuracy score and confusion matrix when tested on the test set.
- Therefore, their GridSearchCV best scores are used to rank them instead. 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 (GridSearchCV best score: 0.876857142857143)
    2. Knearest neighbors, KNN (GridSearchCV best score: 0.848214285714285 8)
    3. Support vector machine, SVM (GridSearchCV best score: 0.848214285714285 6)
    4. Logistic regression (GridSearchCV best score: 0.846428571428571 3)

DISCUSSION

- From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Therefore, each feature may have a certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult to decipher. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

CONCLUSION

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate