IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shreya Bhowmick
24th May 2025

Outline

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

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

- This capstone project aims to predict whether the Falcon 9 first stage will land successfully using machine learning.

- Launch cost relevance:
    - SpaceX charges ~$62 million per launch.
    - Competitors charge ~$165 million or more.
    - Reusability of the first stage is key to SpaceX's cost advantage.

- If we can predict landing success, we can estimate launch costs, which is valuable for alternate companies looking to compete with SpaceX.

- Not all failed landings are accidents — some are intentionally controlled ocean landings.

- Main research question:

    - Given features like payload mass, orbit type, launch site, etc.,

    - Can we accurately predict if the Falcon 9 first stage will land successfully?

Section 1

Methodology

Methodology

Executive Summary

The overall methodology includes:

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

Data Collection – SpaceX API

- 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.

Github URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - 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.

GitHub URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/jupyter-labs-webscraping.ipynb

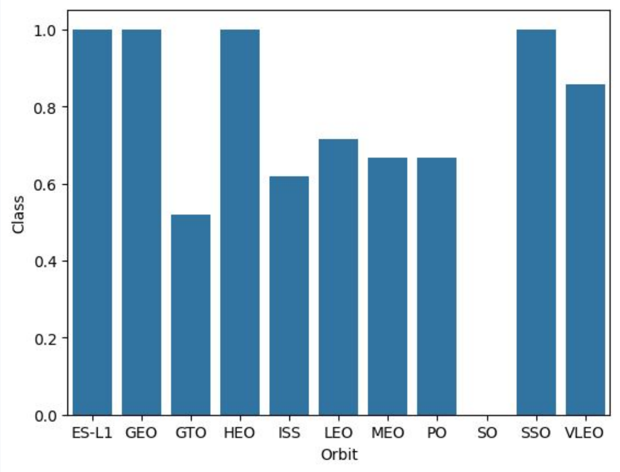
Data Wrangling

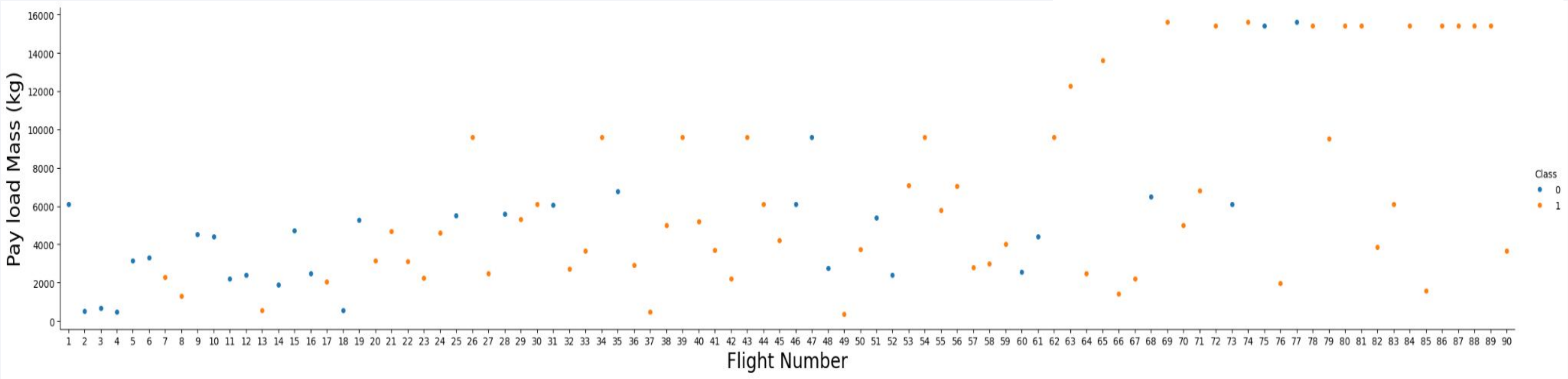
- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits.
- We created landing outcome label from outcome column and exported the results to csv.

GitHub URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.





GitHub URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.

• We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:

- The names of 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
- The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names.

GitHub URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:

- Are launch sites near railways, highways and coastlines.
- Do launch sites keep certain distance away from cities.

GitHub URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

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

Github URL:
https://github.com/Shreya-Bhowmick/IBM_Datascience_Capstone/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

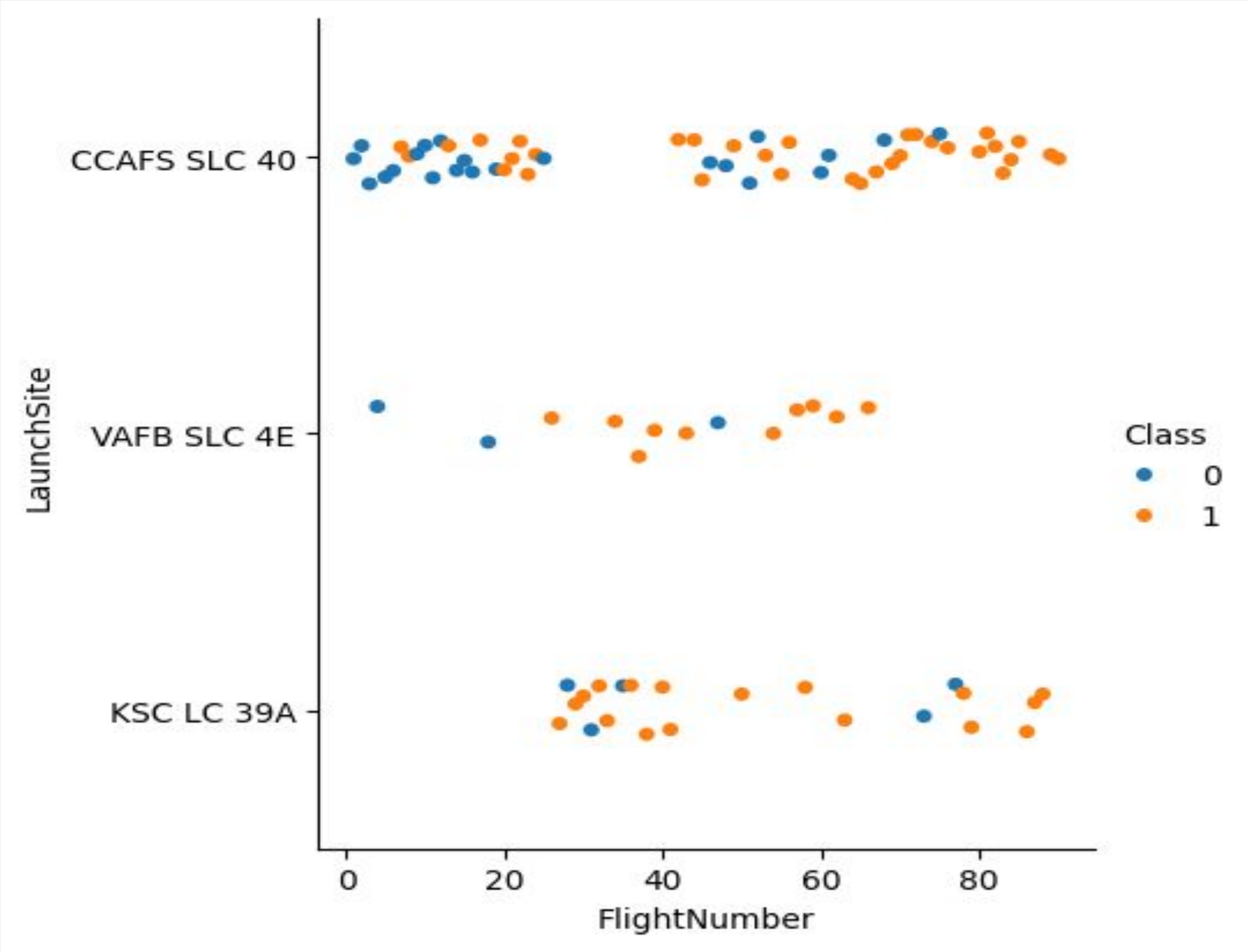
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

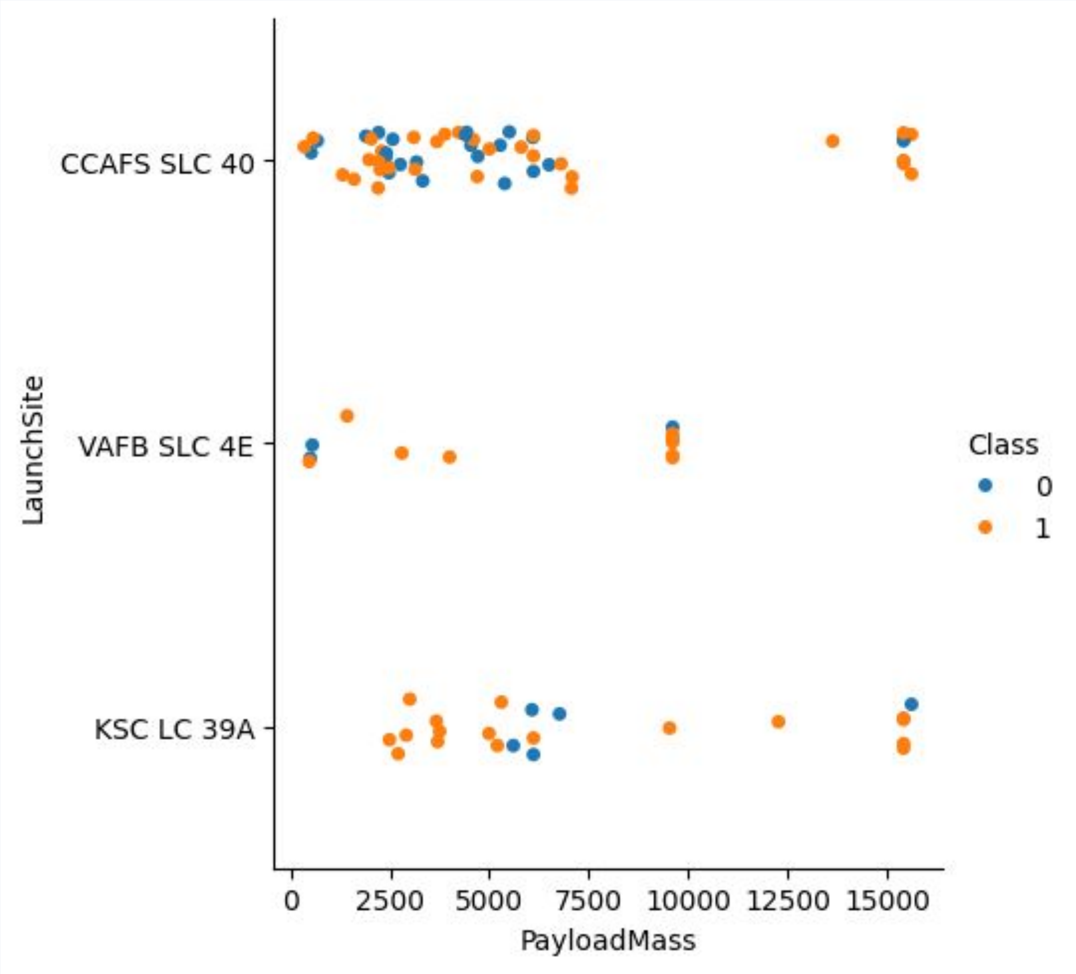
Flight Number vs. Launch Site

From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



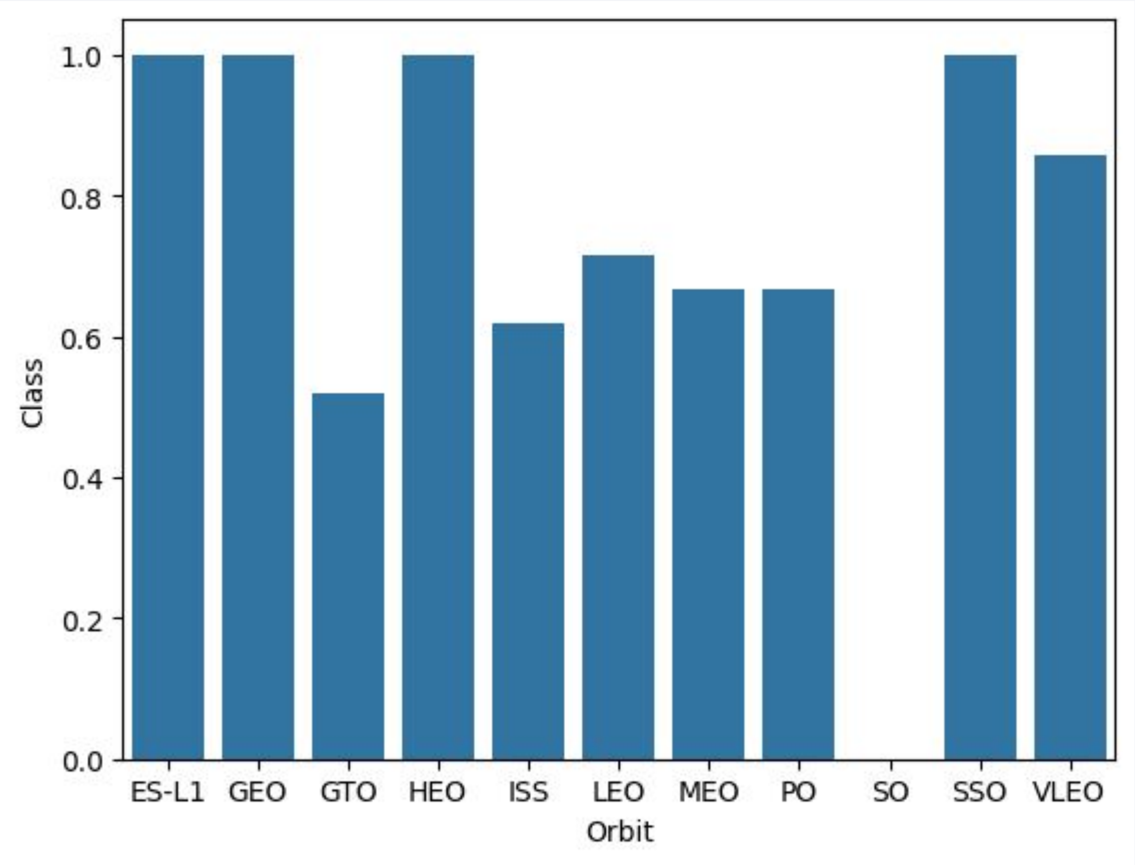
Payload vs. Launch Site

- Greater the payload for launchsite CCAFS SLC 40, higher is the chance of successful landing of the rocket.



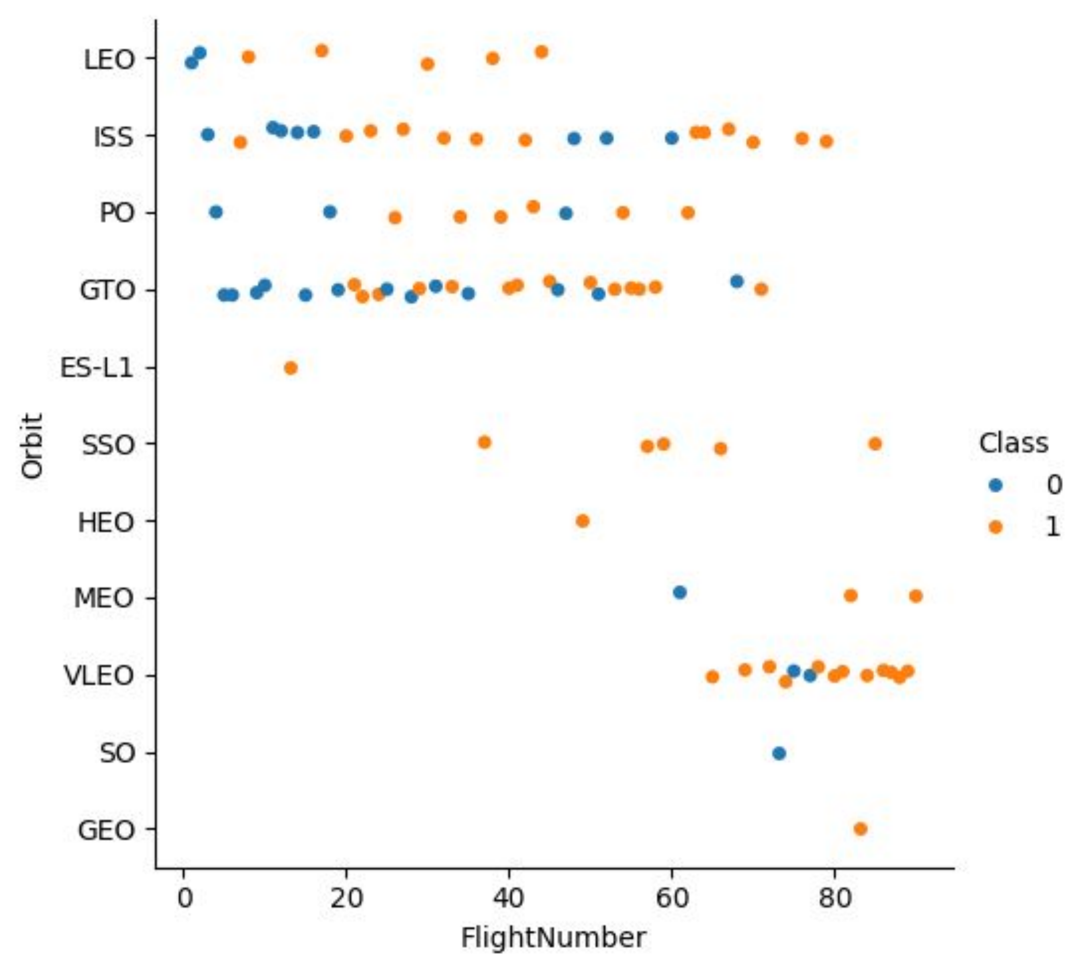
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



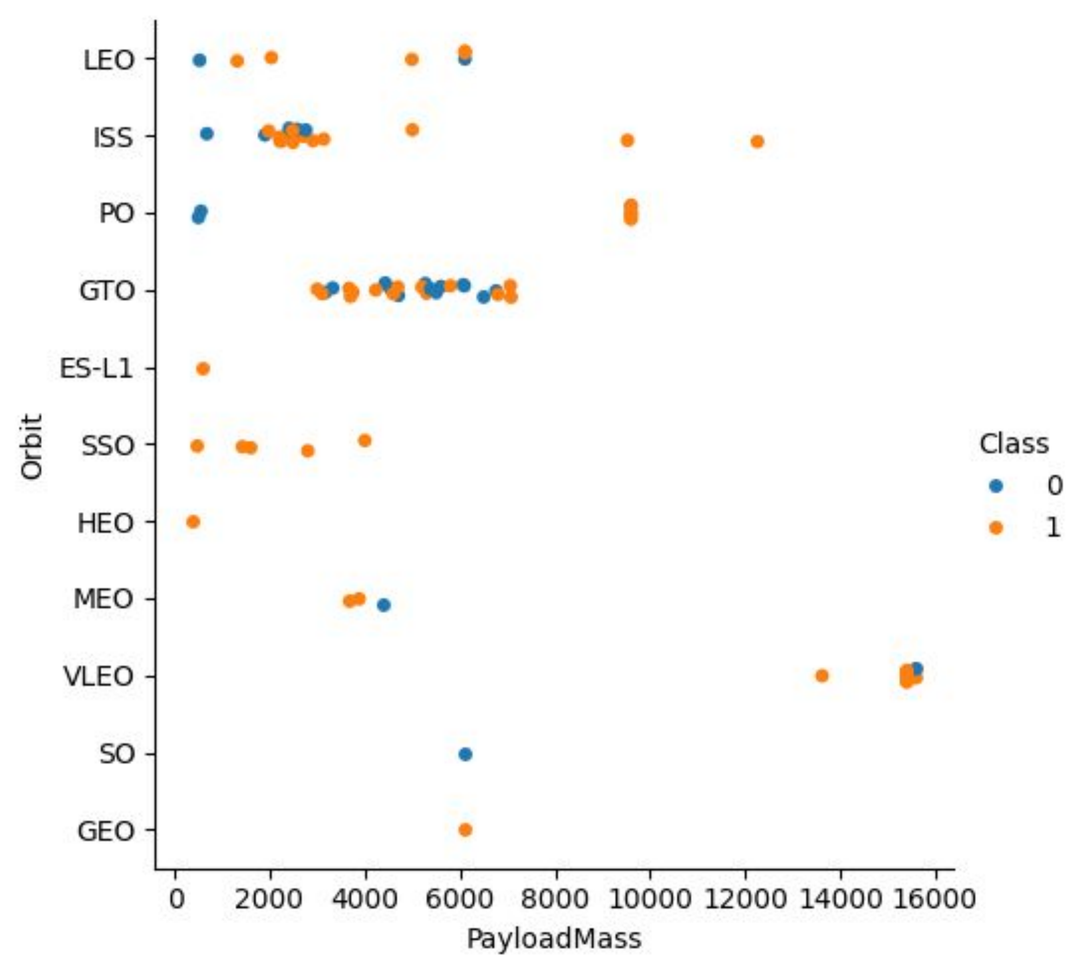
Flight Number vs. Orbit Type

We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



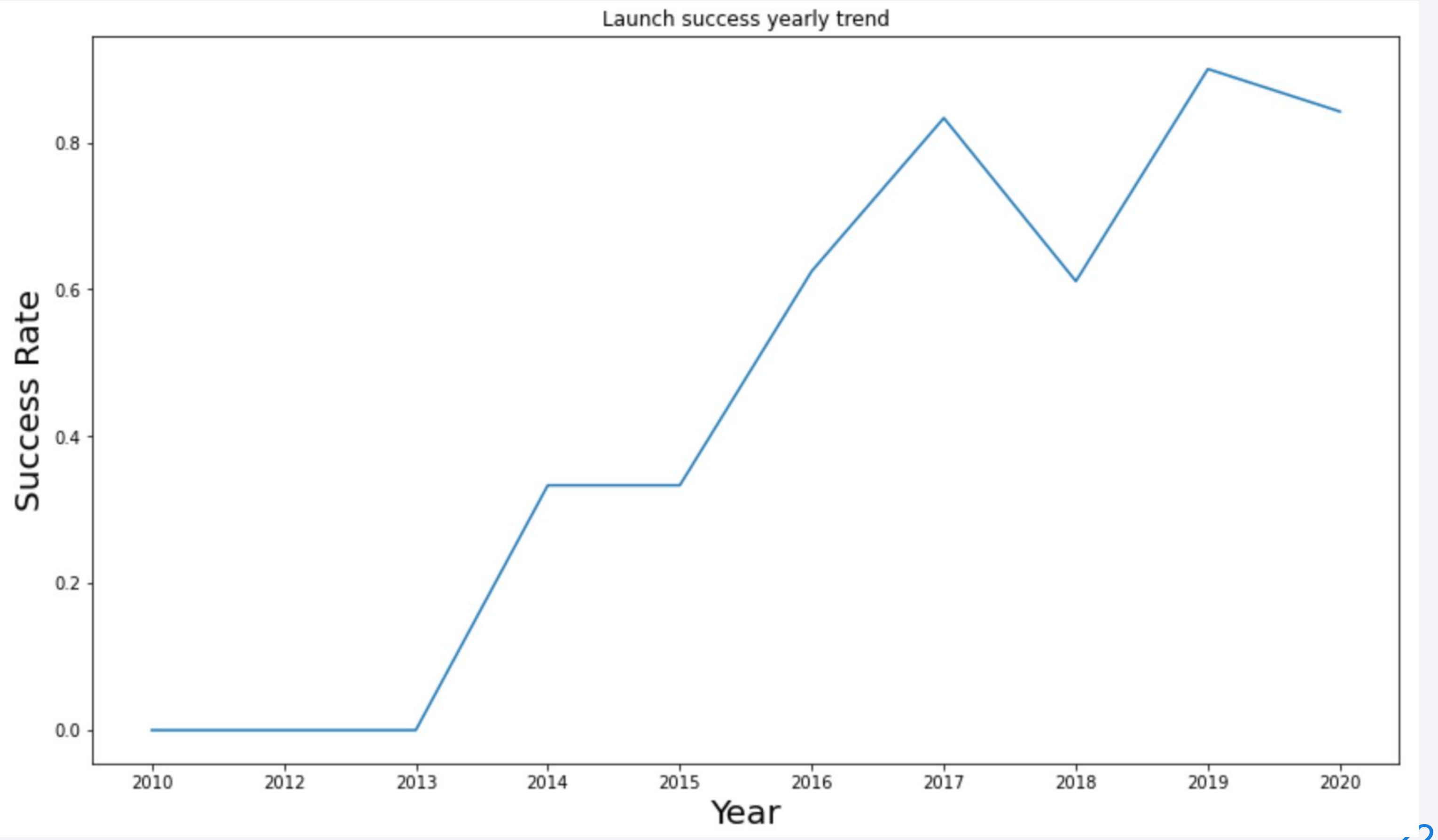
Payload vs. Orbit Type

We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



Launch Success Yearly Trend

We can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

	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

Total Payload Mass

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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

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

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

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

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

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

2015 Launch Records

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

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

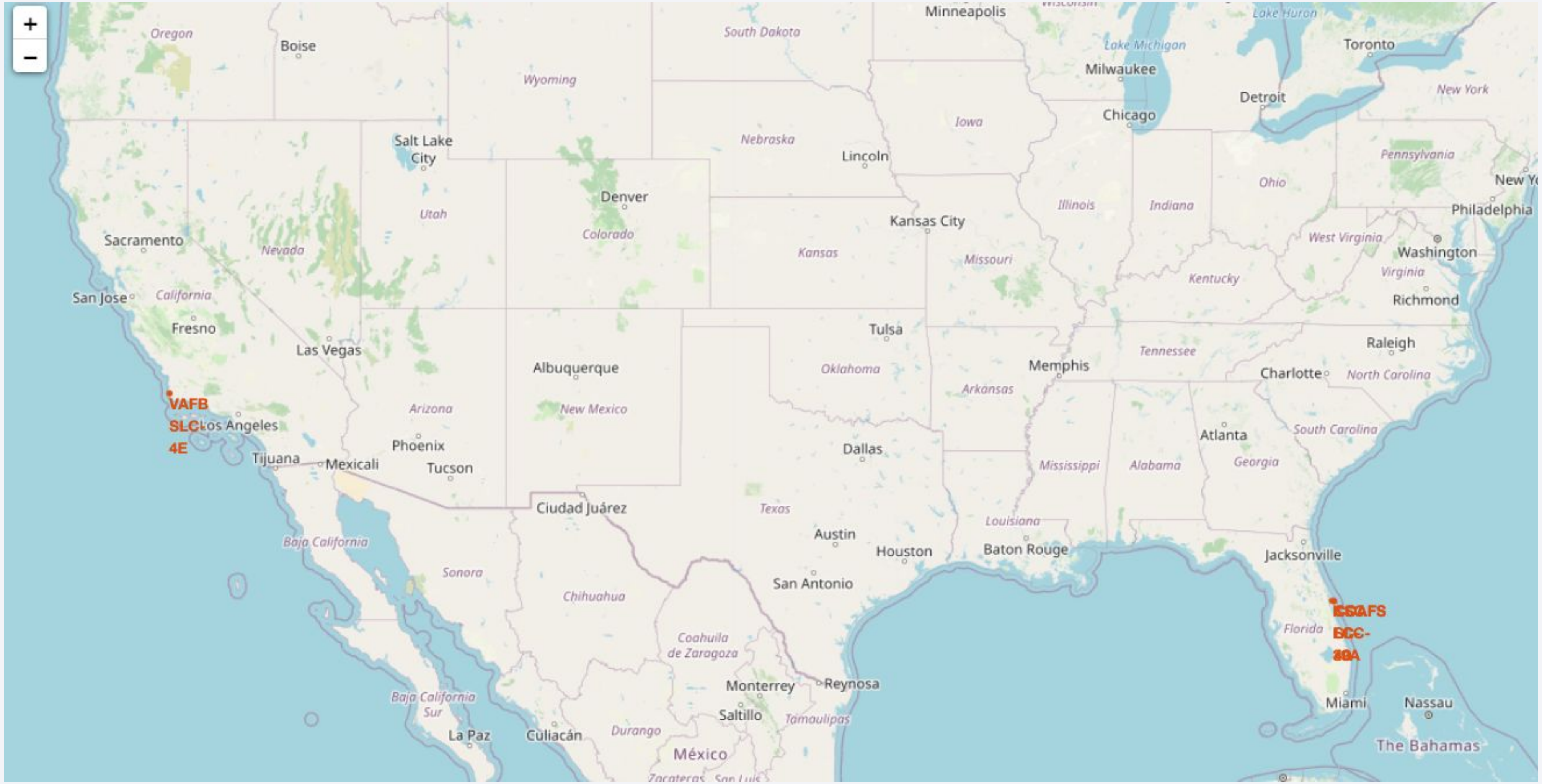
The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

Section 3

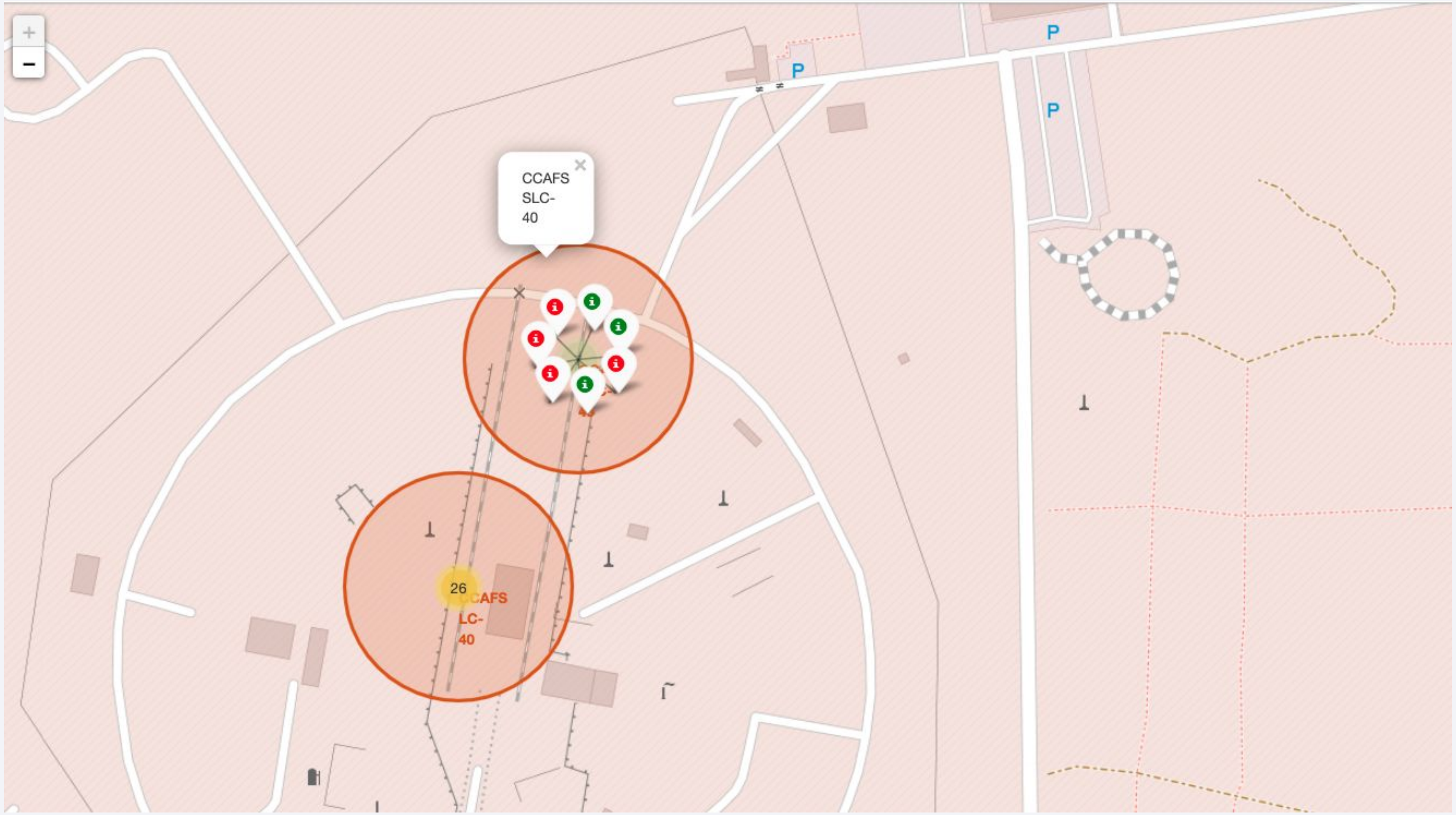
Launch Sites Proximities Analysis

Global Launch Sites



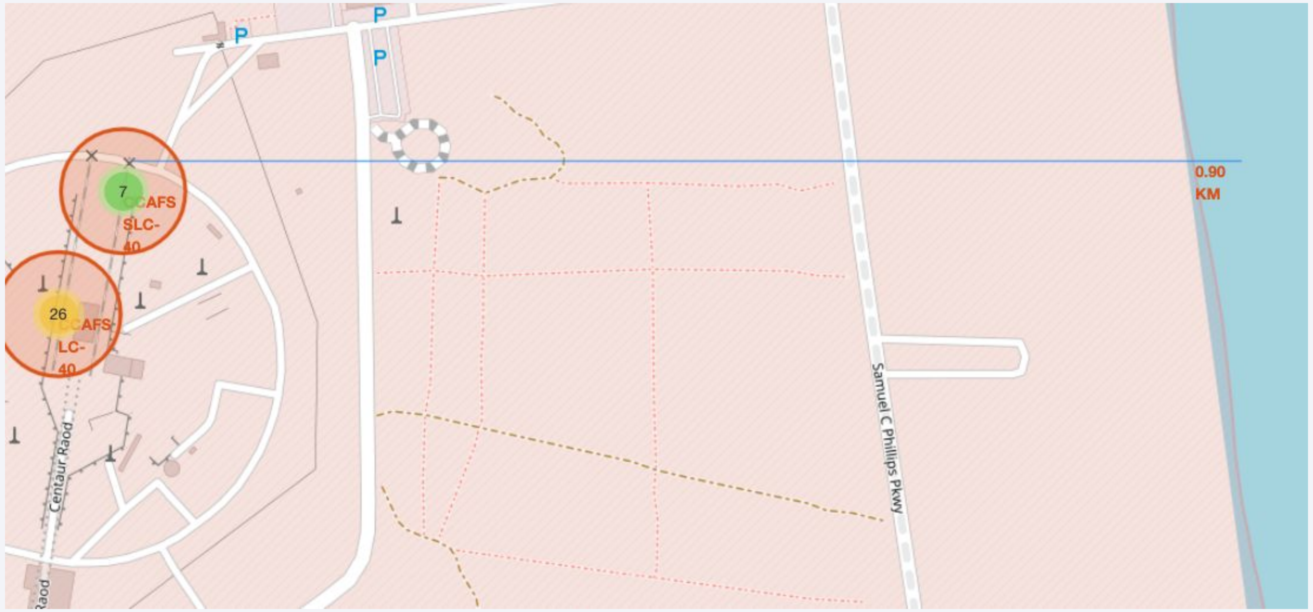
Successful and failed launches

Green markers show successful launches and red marker shows failures.



Distances between a launch site to its proximities

Distance of the launch site from the coastline



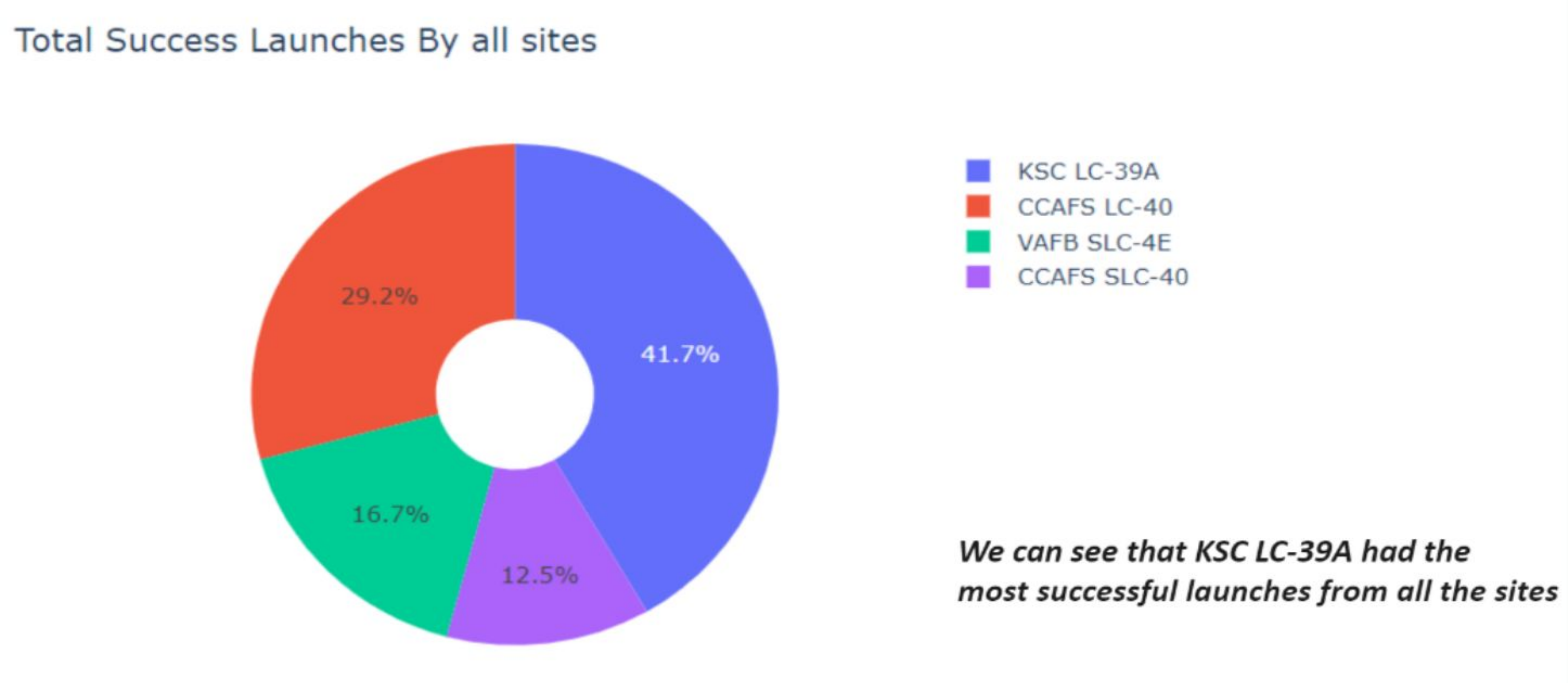
Highway Symbol

City Symbol

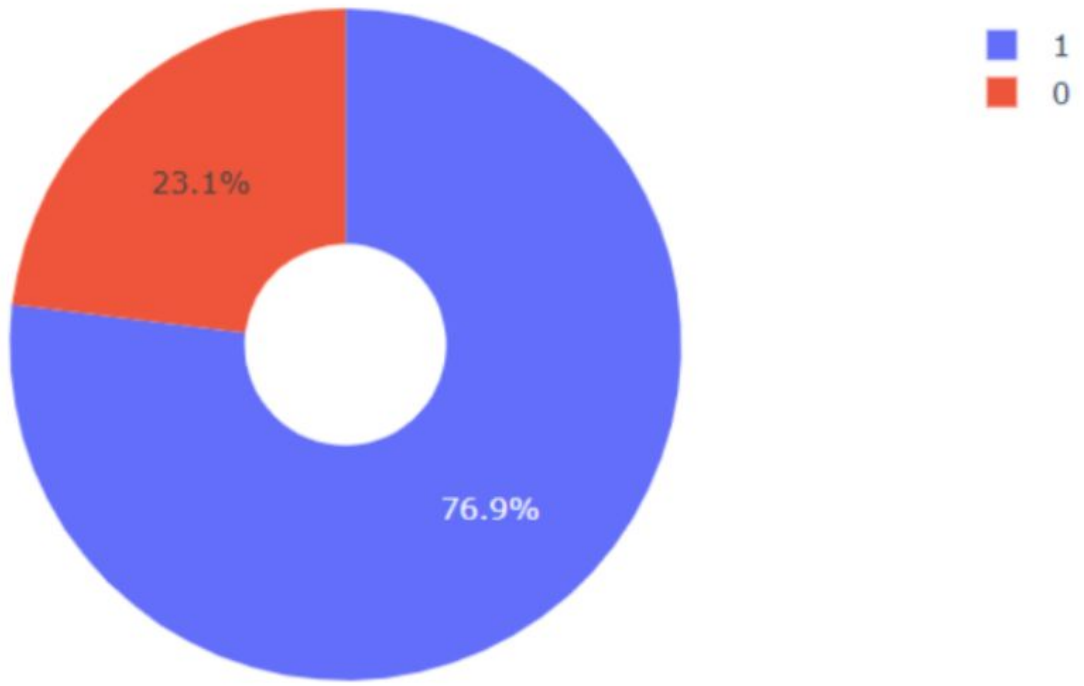
Section 4

Build a Dashboard with Plotly Dash

Total successful launches by all sites in Dashboard

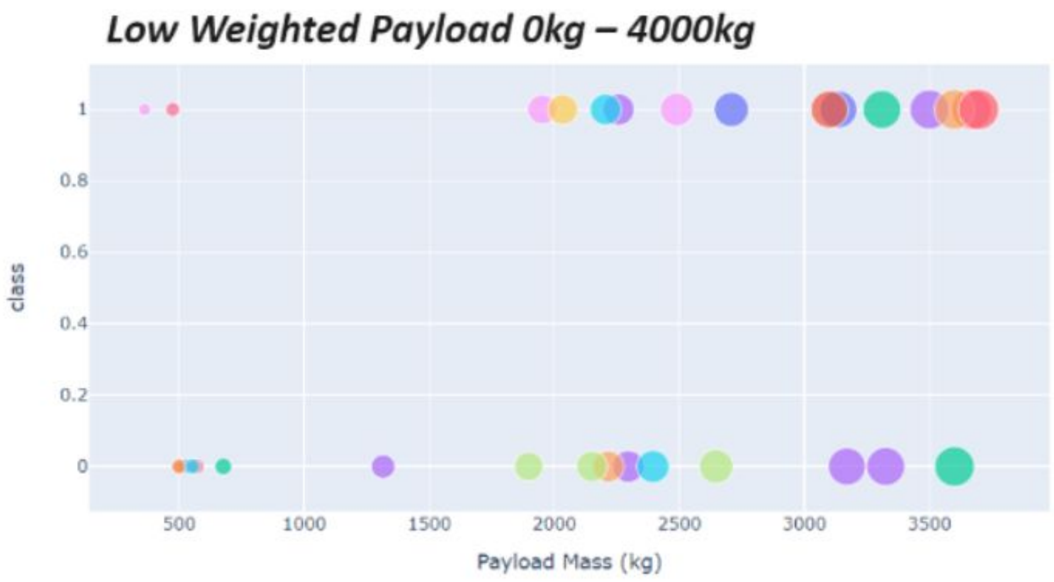


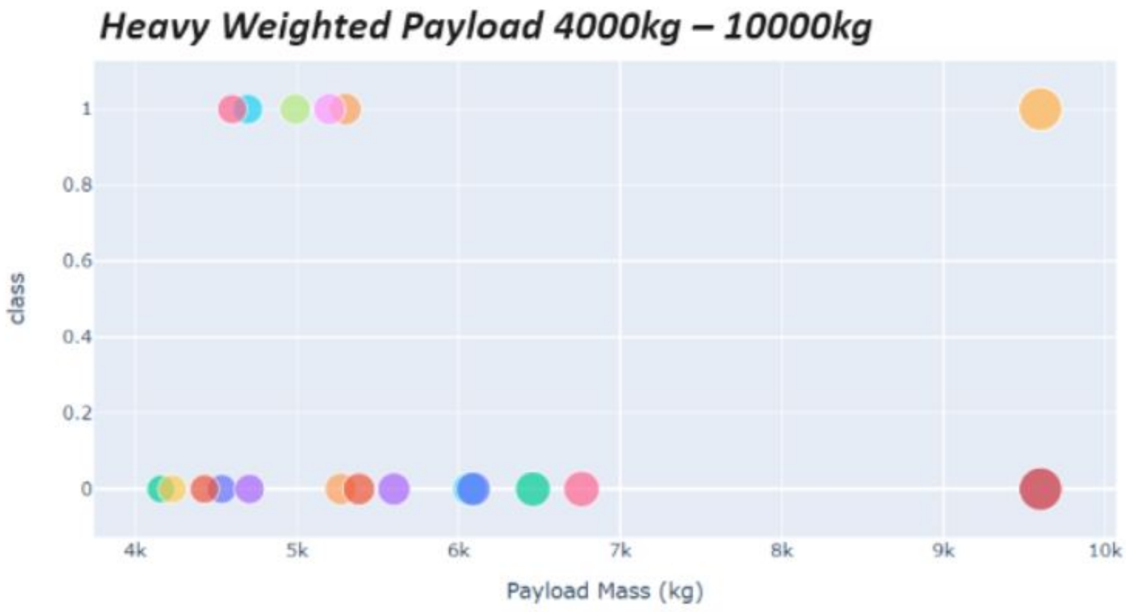
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

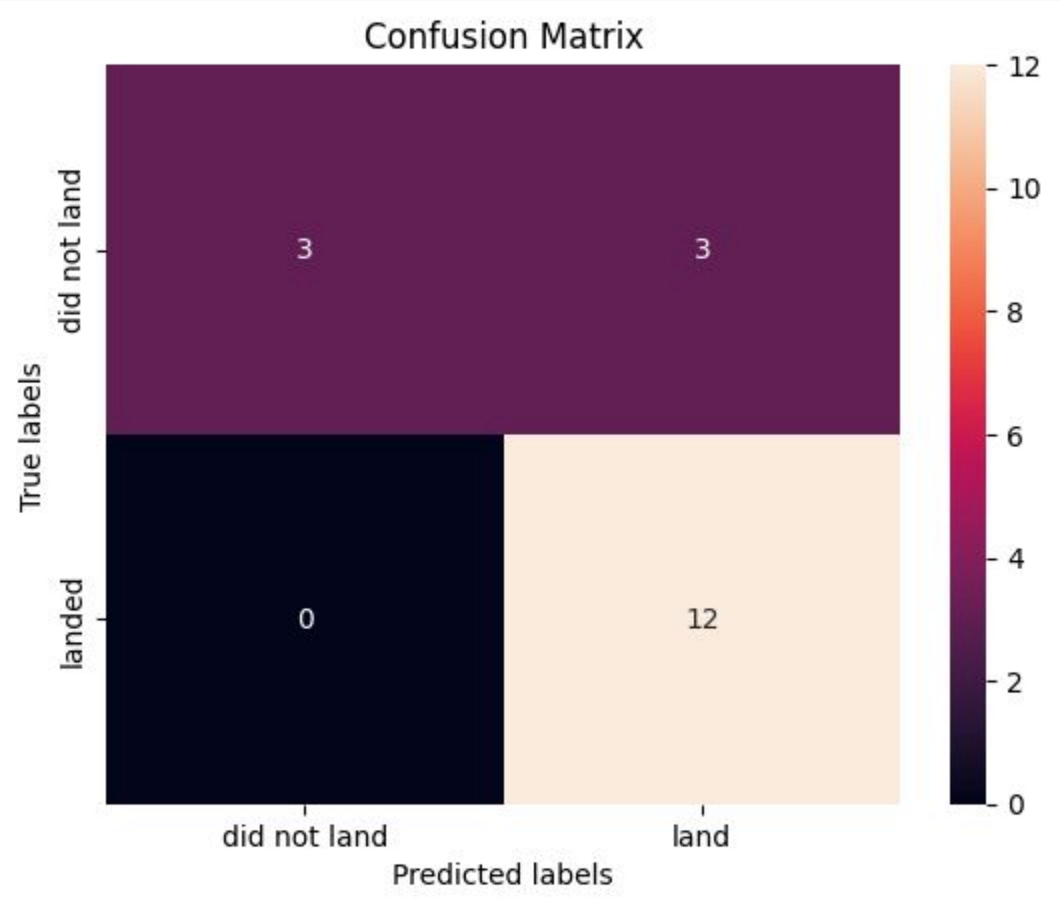
- The classification accuracy of different machine learning models are as follows:
    - Logistic Regression: 0.8464285714285713
    - Support Vector Machine: 0.8482142857142856
    - Decision Tree: 0.8625
    - K-Nearest Neighbors: 0.8482142857142858

The best performing model is Decision tree.

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful

- landing by the classifier.



Conclusions

We can conclude that:

• The larger the flight amount at a launch site, the greater the success rate at a launch

site.

• Launch success rate started to increase in 2013 till 2020.

• Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

• KSC LC-39A had the most successful launches of any sites.

• The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!