Applied Data Science Capstone: An Analysis of SpaceX through Data Science

Angel Barra Muñoz

February 22, 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Collecting the Data
  - Trough API
  - Trough Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA)
  - SQL
  - Visualization
- Interactive Visual Analytics and Dashboard
  - Folium
  - Plotly
- Model predictions
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbor

INTRODUCTION

SpaceX is a space transportation and aerospace manufacturer founded in 2002 by Elon Musk.

Rockets from the Falcon 9 family have been launched 309 times over 14 years, with 91 launches occurring in 2023 alone. The machine learning approach for studying the influence of different variables is essential to predict if a landing will be successful. A significant factor in SpaceX's ability to offer launches at the "reduced" cost of 62 million USD is the reusability of the Falcon 9 rockets.

METHODOLOGY

- Data Collection
  - Using Python the data is collected trough the API from SpaceX and Web Scraping from Wikipedia
- Data Wrangling
  - Processing the data to find patters and determine what is the label for the supervised model
- Evaluation Data Analysis
  - EDA with SQL to study the failures and successes of the landing
  - EDA with visualization to study the different inference of the variables of the data frame in the determination of the failure or success of a landing
- Interactive Analysis
  - See the geographic position of the launchpad and the inference of the ambient in the success of a launch.
  - With Plotly see in a interactive way the success rate of every launchpad and the relation between the payload mass and the success of the mission.
- Model Application to the Data.
  - Determinate the best Hyperparameter for a variety of method, these are Logistic Regression, SVM, Decision Tree and KNN
  - Determinate which of the 4 methods have the best performance.

Data Collection

SpaceX API

- We request data rom the SpaceX API, which is provided in JSON format. This data is then converted into a dataframe for analysis. Since the dataset includes information on both Falcon 9 and Falcon Heavy launches, we apply a filter to isolate only the data pertaining to Falcon 9. Following the filtration process, we export the refined dataframe to a CSV file for further use.
- Link: https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_1_API.ipynb

Web Scraping

- We retrieve Falcon 0 launch data from its Wikipedia page using the specified URL. A BeautifulSoup object is then instantiated to parse the HTML content. Utilizing this object, we extract the columna names and variable identifiers from the HTML headers to strucsture our datset appropriately.
- Link: https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_1_WebScraping.ipynb

Data Wrangling

- During the data wrangling pase, we leveraged pandas to refine the previously obtained dataframe, thereby streamlining subsequent analyses. Our efforts focused on calculating the number of launches for each site in conjuction with the corresponding orbit type at the time of launch. Additionally, we analyzed the mission outcomes to assign a ‘landing outcome’ label, categorizing each evento as either a success or a failure.

- Link:
  https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_2_Data_Wrangling.ipynb

Exploratory Data Visualization with SQL

- Using the %sql magic command, we execute queries to better understand the structure and contents of the database
  - Displaying the names of the launch sites.
  - Displaying the records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying the total average payload mass carried by booster versión F9 1.1
  - 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 les than 6000 kg
  - 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 sites names for in year 2015.
  - Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

  Link:

  https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_2_EDA_SQL.ipynb

Exploratory Data Visualization with Data Visualization

- We conduct an analysis of the relationships between various variables. Shortly, we will present these variables and the corresponding visualizations that illustrate their interconnections.
- The relations are:
  - Flight Number vs Pay Load Mass (kg)
  - Flight Number vs Launch Site
  - Launch Site vs Pay Load Mass (kg)
  - Orbit vs Successful rate
  - Flight Number vs Orbit
  - Pay Load Mass (kg) vs Orbit
  - Year vs Successful rate

Link: https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Wek_2_EDA_DV.ipynb

Data Visualization

- With Folium
  - Our work with Folium involved plotting the launch sites on a map and annotating each sie with markers that signify the success or failure of launches conducted there.
  - We also calculate the distances from the launch site to the nearest cities, highways, and railways to assess proximity and accesibility.
  - Link:
    https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_3_Folium.ipynb

- With Plotly Dash
  - We utilized Plotly to develop and interactive web interface that displays the success rates of every launch site collectively. Users have the ability to filter and view the specific number of successful and failed launches at each site.
  - Addiotionally, users can adjust a slider to see how the payload mass affects the success rate of landings.
  - Link:
    https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_3_Plotly_Dash.ipynb

Predictive Analysis (Classification)

- Our predictive analysis utilized four distinct methods to ensure a comprehensive approach:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- We optimized each method by searching for the best hyperparameters using GridSearchCV.
- Link:
  https://github.com/angelbarram/Applied-Data-Science-Capstone/blob/main/Week_4_Model_Prediction.ipynb

Results (EDA with SQL)

- Display 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

Results (EDA with SQL)

- Display 5 records where launch sites begin with the string ‘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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 (EDA with SQL)

- Display the total payload mass carried by booster launched by NASA (CRS)

payloadmass
619967

- Display average payload mass carried by booster versión F9 1.1

averageloadmass
2928.4

- List the date when the first successful landing outcome in ground pad was achieved.

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

Results (EDA with SQL)

- List the names of the boosters which have success in drone shop and have payload mass greater than 4000 but les than 6000

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- List the total number of successful and failure mission outcomes

Mission_Outcome	missionoutcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Results (EDA with SQL)

- List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

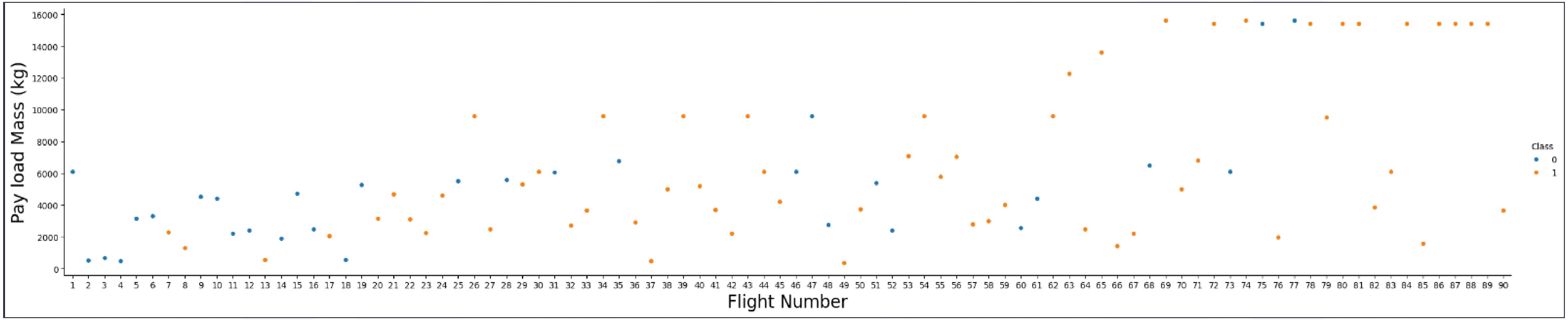
Results (EDA with SQL)

- Rank 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.

Landing_Outcome	count_landing
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 (Data Visualization)

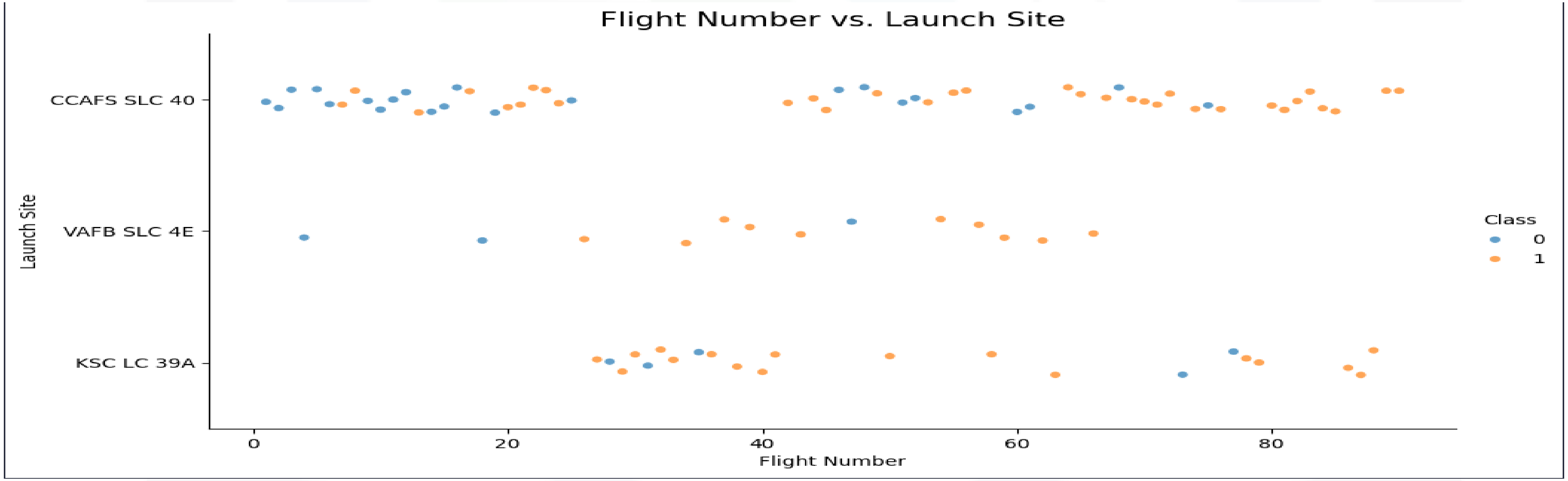
- Flight Number vs Payload Mass (kg)



- CCAFS LC-40 has the lowest success rate which is 60% while the highest is KSC LC 39

Results (Data Visualization)

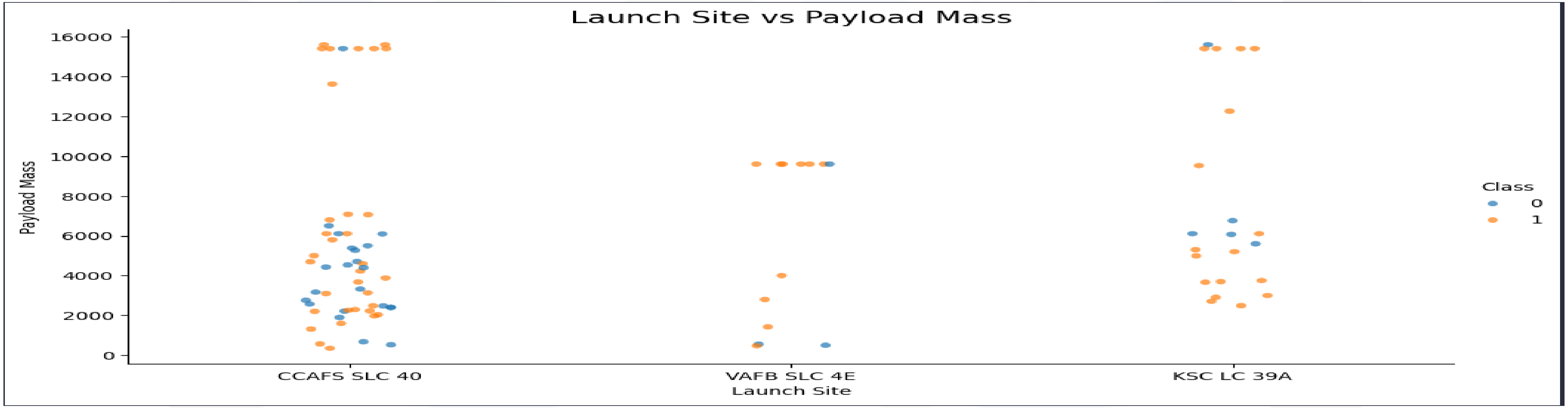
- Flight Number vs Launch Site



- Same as the previous graph, we can see that KSC LC 39A has the highest rate of success and between Flight Number 40 and 70 has only successful landings

Results (Data Visualization)

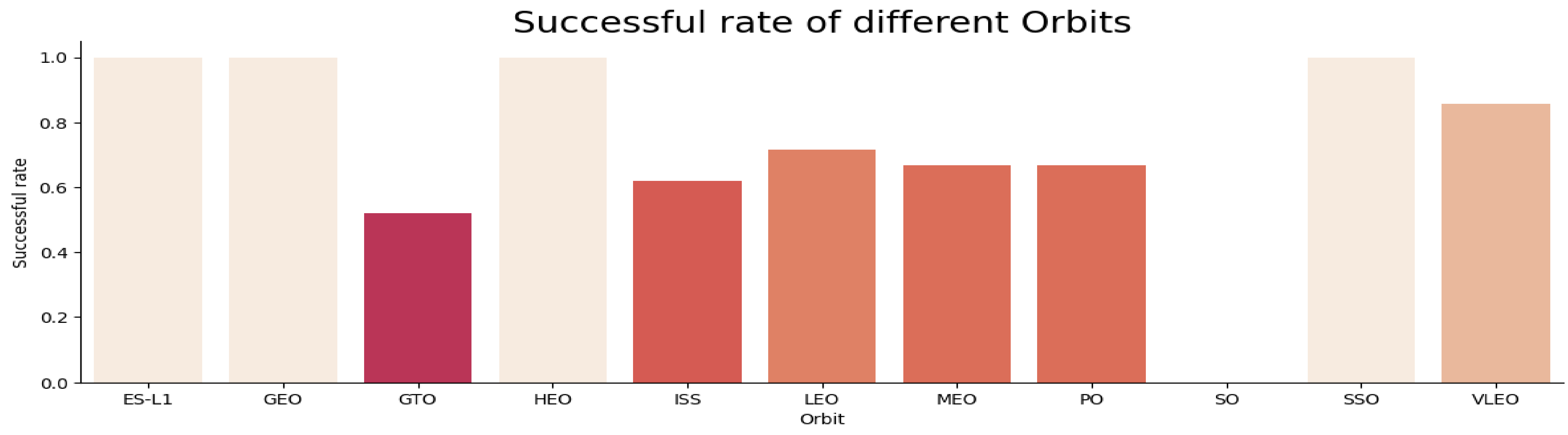
- Launch Site vs Pay Load Mass (kg)



- We can see in the graph that at higher mass the succesfull rate for CCAFS SLC 40 and KSC LC 39A are higher than lower masses while VAFB SLC 4E have a great succesfull rate at any mass.

Results (Data Visualization)

- Successful rate of different Orbits



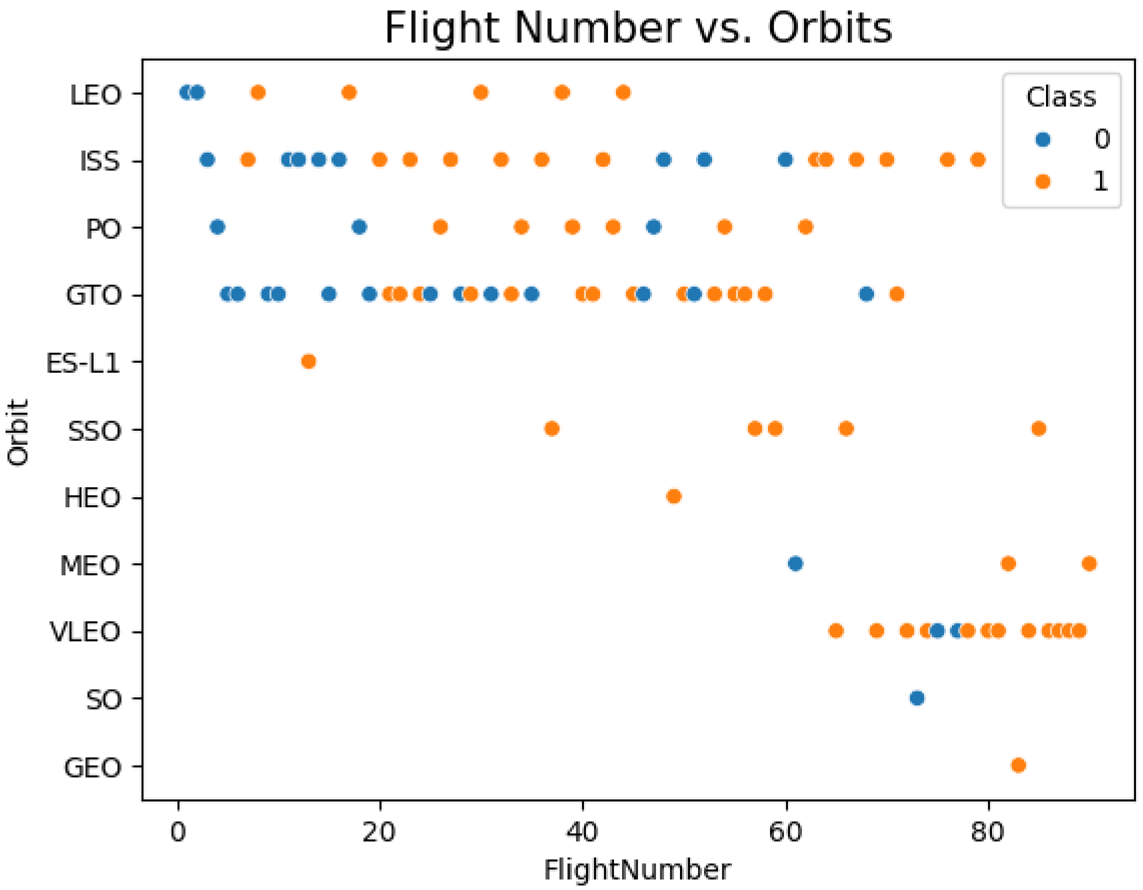
- The orbits ES-L1, GEO, HEO and SSO has the highest succes rate.

Results (Data Visualization)

- Flight Number vs. Orbits

- This graph servers as an addiotional illustration, similar to the previous one. It reveals that the success rates for GEO, HEO and ES-L1 are not reliable because there's just one sample on each orbit, so we can't say that orbit always will end in a successful landing.

- In this case we can see that VLEO has a great successful rate.

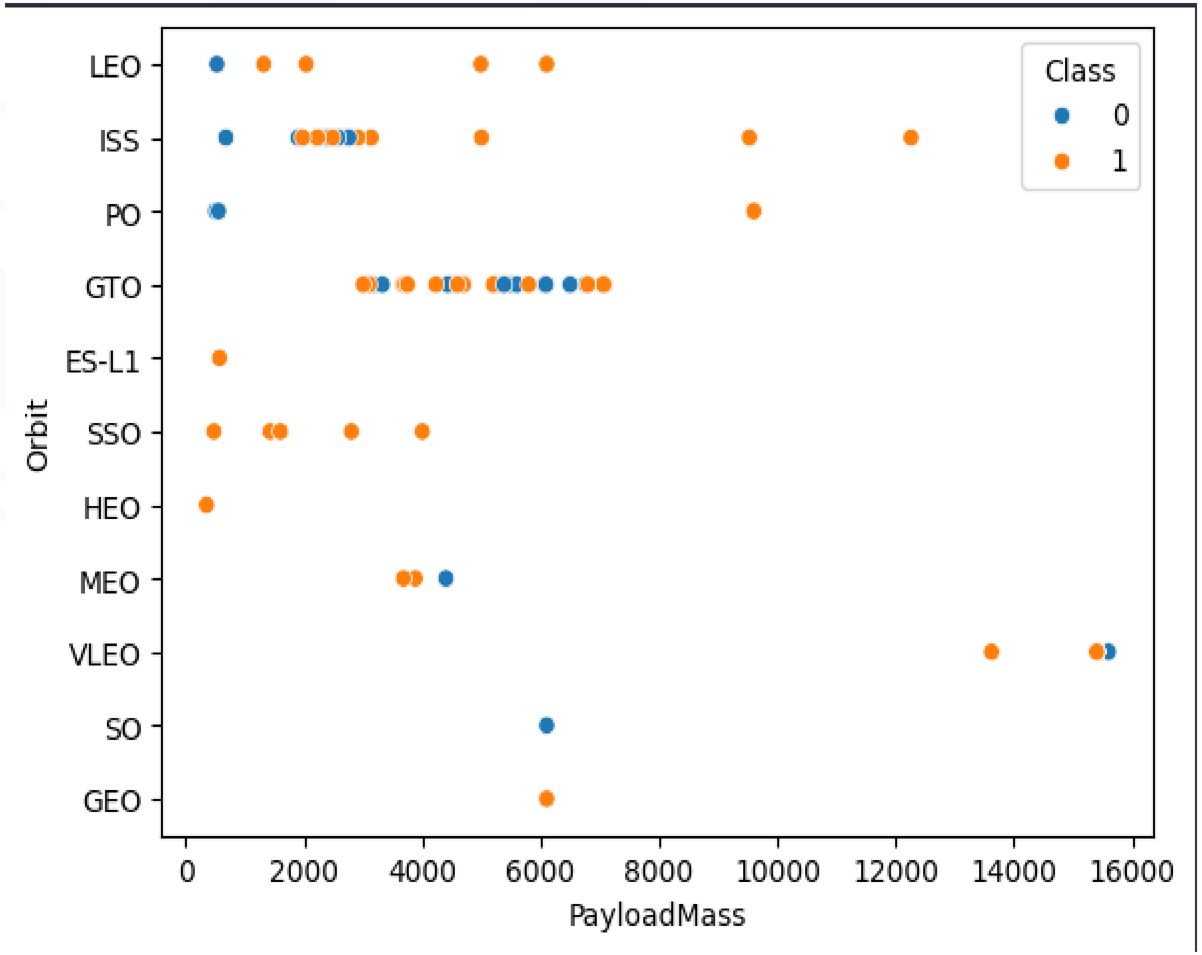


Results (Data Visualization)

- Flight Number vs. Orbits

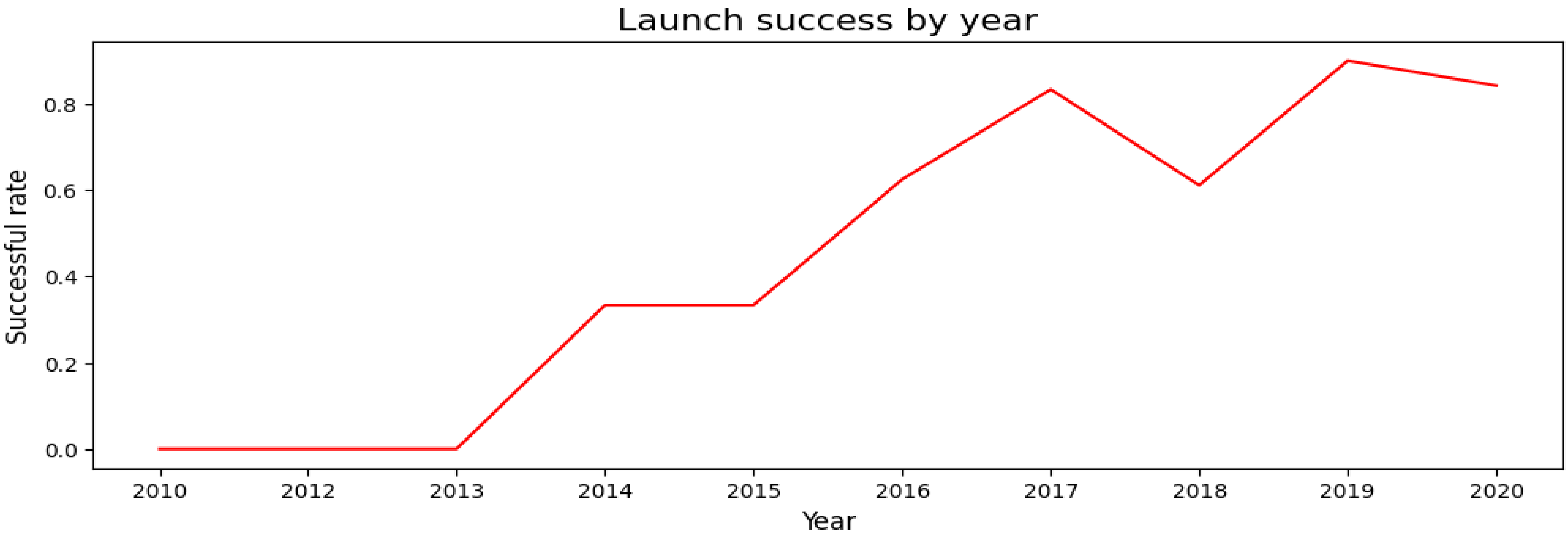
- The analysis of the graph indicates that SSO continues to be a viable orbit with consistent succes. On the other hand, VLEO shows a 66% success rate with only three simples, suggesting a need for more data toa ssess reliability fully.

- For missions to the ISS mass consideration, the success rate appears favorable, similarly to launches to GTO.



Results (Data Visualization)

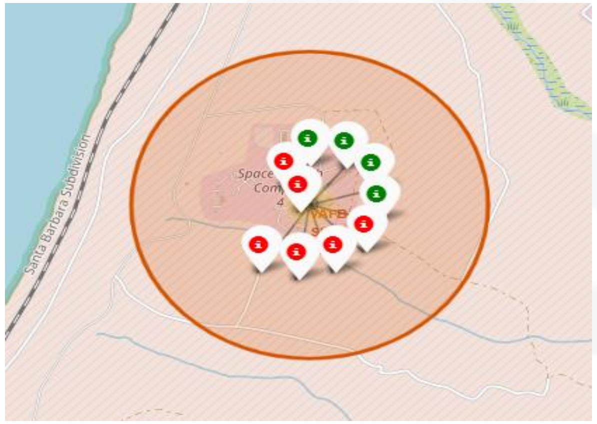
- Launch success per year



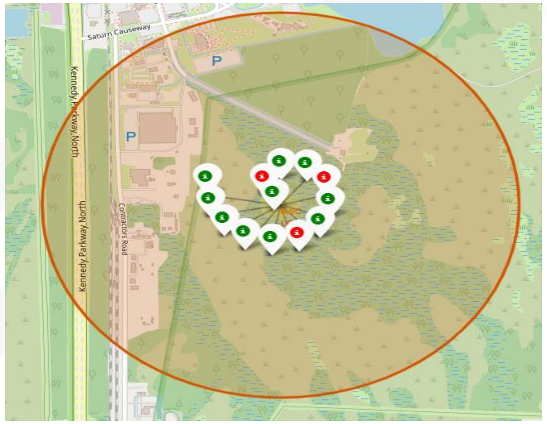
- The graph suggests a year-over-year increase in the success rate of launches, corroborated by the high volume of 91 launches in the previous year, 2023.

Results with Folium

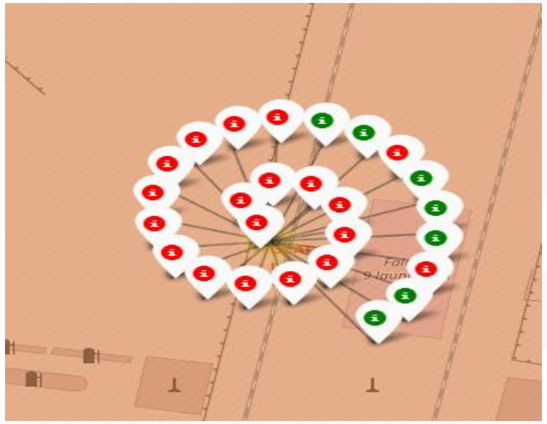
VAFB SLC-4E



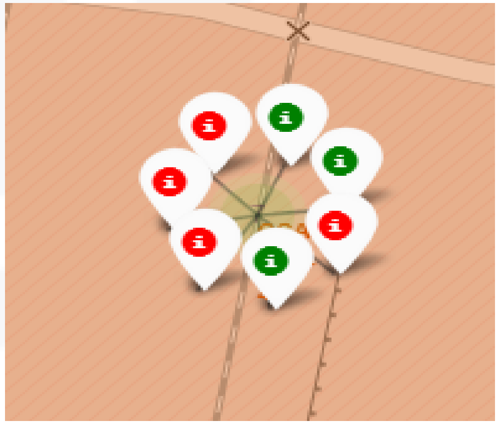
RSC LC-39A



CCAFS LC-40

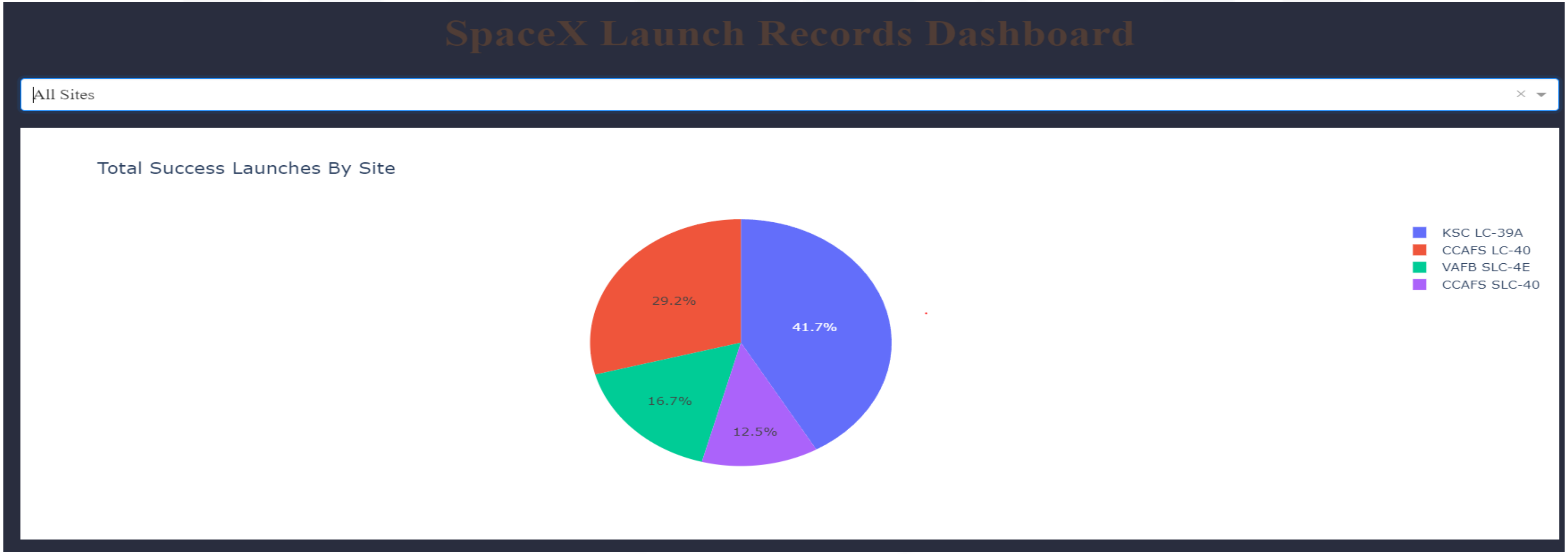


CCAFS SLC-40



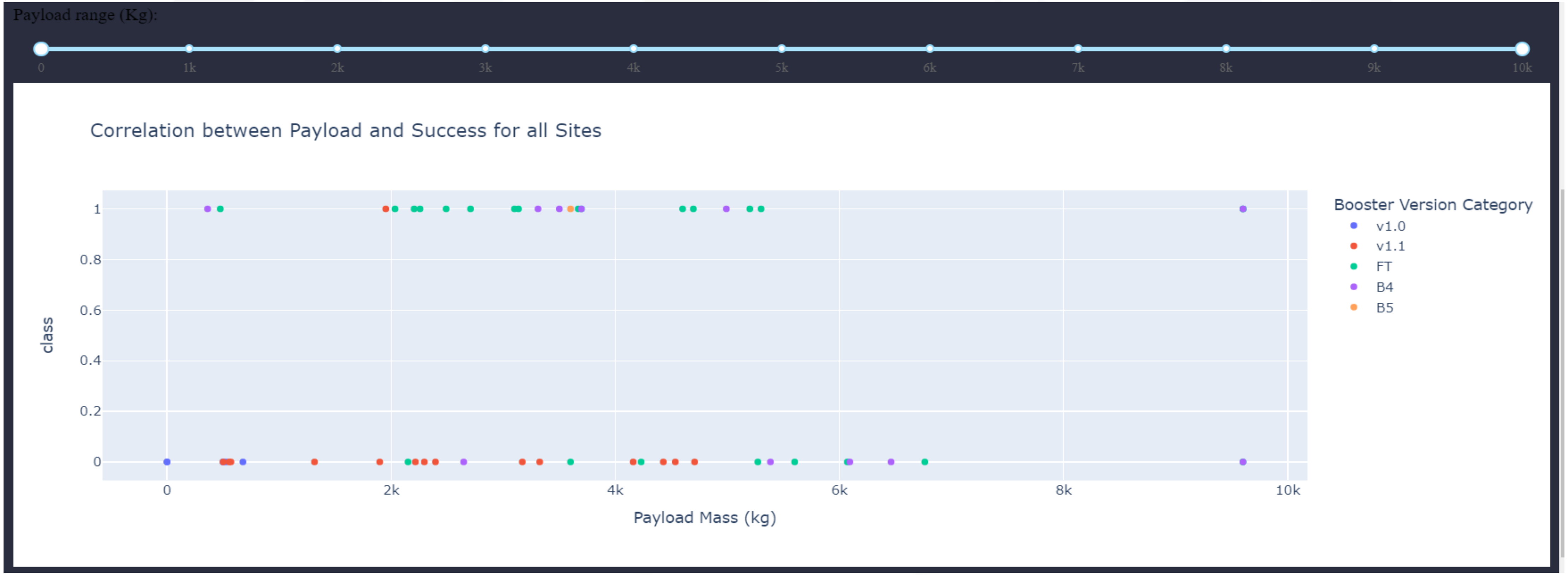
Results with Plotly Dash

- Launch success rate



Results with Plotly Dash

- Payload vs the success of a landing



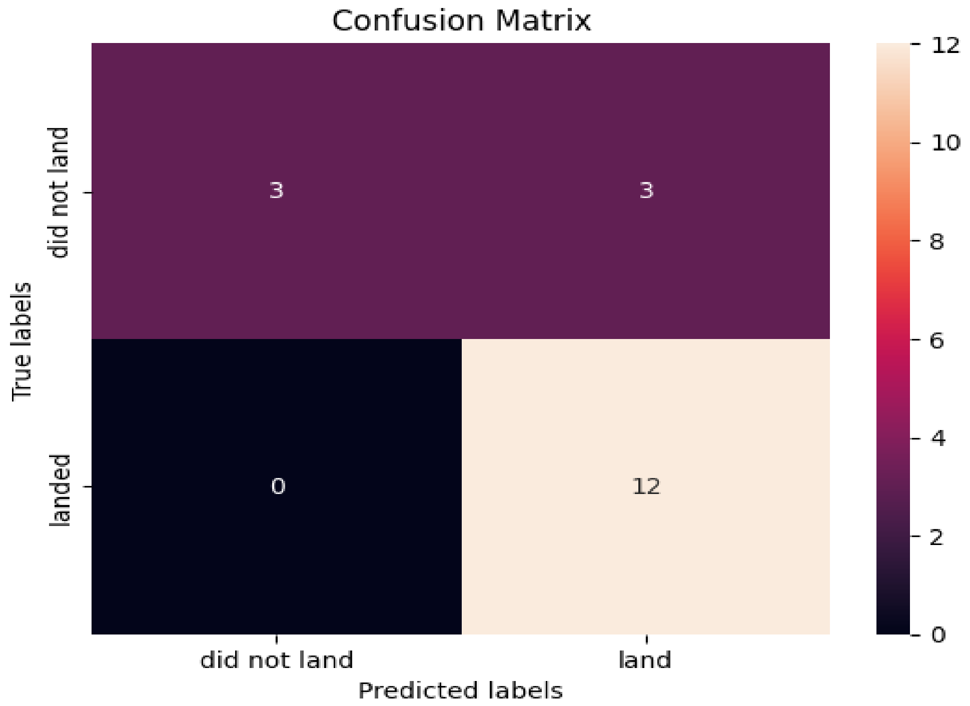
Results (Predictive Models)

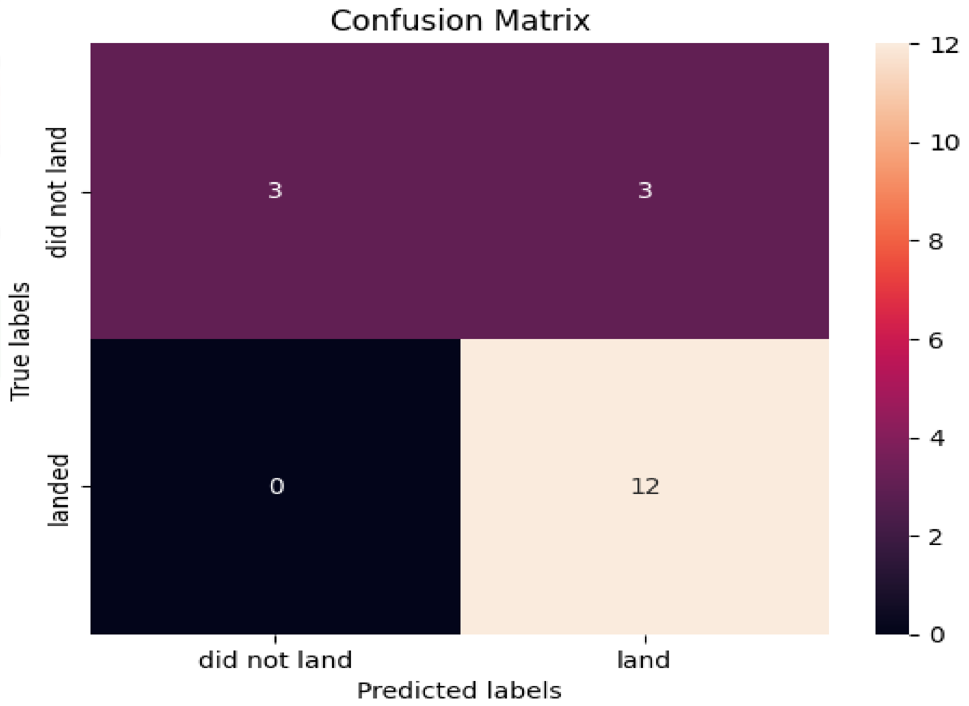
- Logistic Regression:
  - Hyperparameters:

    C: 0.01, penalty: 12, solver:'lbfgs'
    Accuracy: 83.3%

- SVM:
  - Hyperparameters:

    C: 1.0, gamma: 0.03162277660168379, kernel: sigmoid
    Accuracy: 83.3%



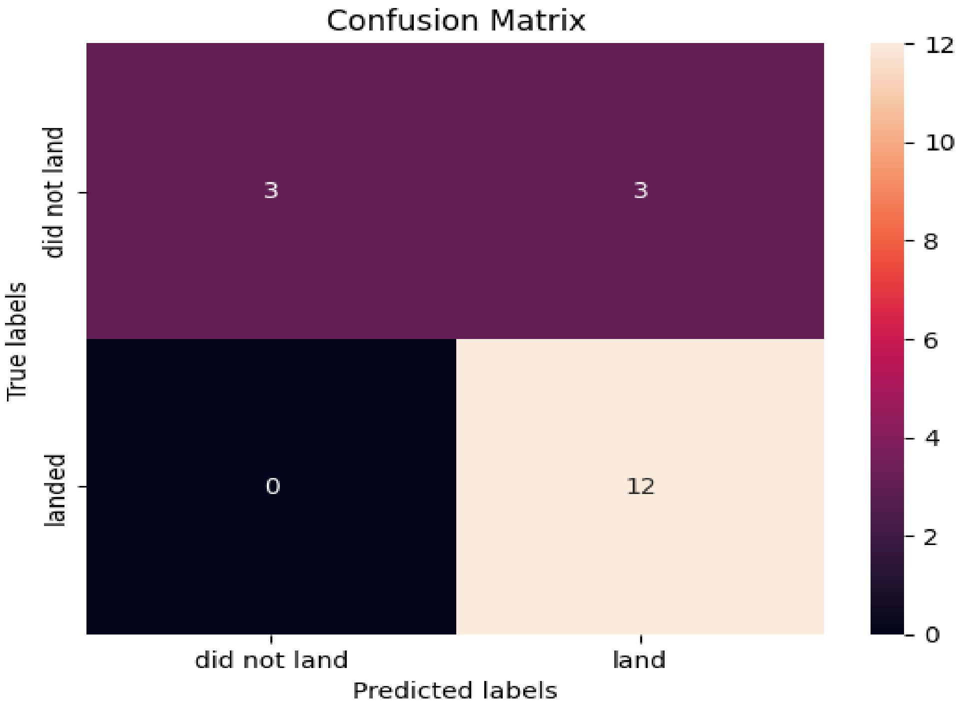


Results (Predictive Models)

- Decision Tree:
  - Hyperparameters:

    criterion: gini, max_depth: 6, max_features: 'sqrt', min_simples_leaf: 4, min_simples_split: 2, splitter: random

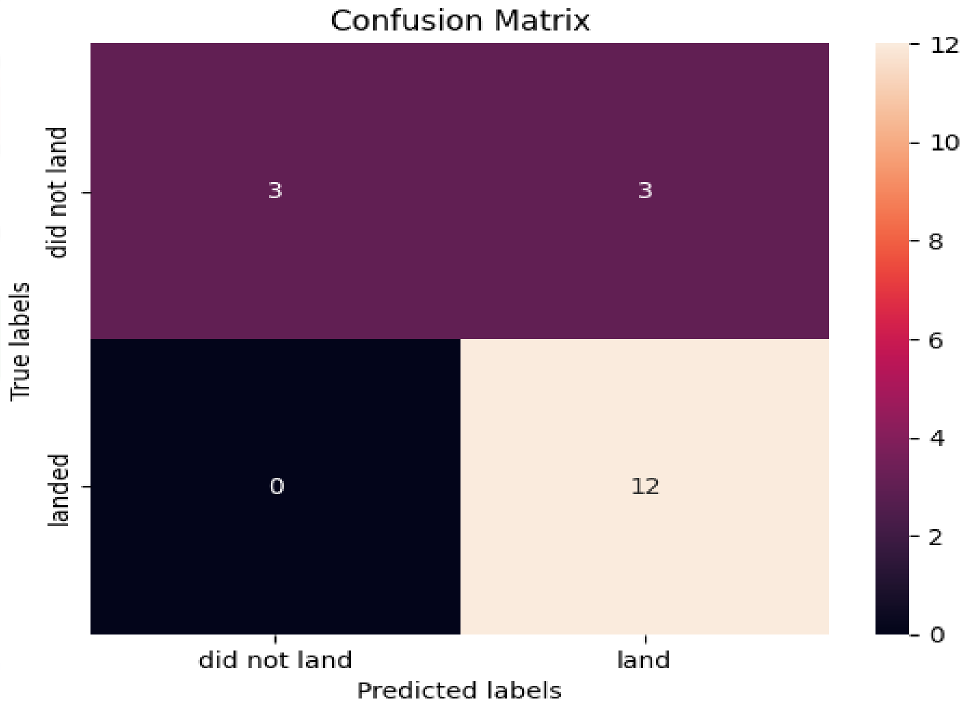
    Accuracy: 83.3%



- KNN:
  - Hyperparameters:

    algoritm: auto, n_neighbors: 10, p: 1

    Accuracy: 83.3%



Conclusion

- The landing pads are strategically placed close to the sea and at a distance from urban áreas and major roads, presumably the risk of civilian casualities in the evento of an incident. A higher payload mass appears to correlate with a reduces risk of launch failure. Specially, the ISS and SSO trajectories demonstrate the highest success rates when considering payload mass. Regarding predictive modeling, all four models tested exhibit comparable accuracy, making each a viable choice for forecasting outcomes for SpaceX's Falcon 9 launches.