Data Science Capstone Project

Marcos Lopez Martinez

12/07/2023

OUTLINE

- Excecutive Summary
- Introduction
- Mathodology
- Results
- Conclusion
- Appendix

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Classification/Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Predictive analysis results
- Interactive analytics

Introduction

- The objective is to evaluate the viability of the new company "Space Y" to compete with Space X

- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting successful landing of the forst stage of rockets
  - Where is the best place to make launches

Methodology

Data collection methodology:

- Space X API
- WebScraping from Wikipedia

Data Wrangling:

- Collected data was enriched by creating a landing outcome label based on autcome data after summarizing and analyzing features

Perform exploratory data analysis (EDA) using tolos and libraries of data visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

Data sets were collected from Space X API and from Wikipedia using web scraping technics

Data Collection: SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used to extract data following the steps bellow:

Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/Data%20Collection%20API.ipynb

Request API and parse the SpaceX launch data → Filter data to only include Falcon 9 launches → Deal with missing values

Data Collection: WebScraping

- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded according to the flowchart and then persisted

Source code: https://github.com/MarcosLopez64/Capstone-Project-/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Request the Falcon 9 launch in the wiki page

Exctract all columns/variables names from the HTML table header

Create a Data Frame by parsing the launch HTML tables

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of misión outcome per orbit type were calculated
- Finally, the landing outcome label was created from Outcome column.

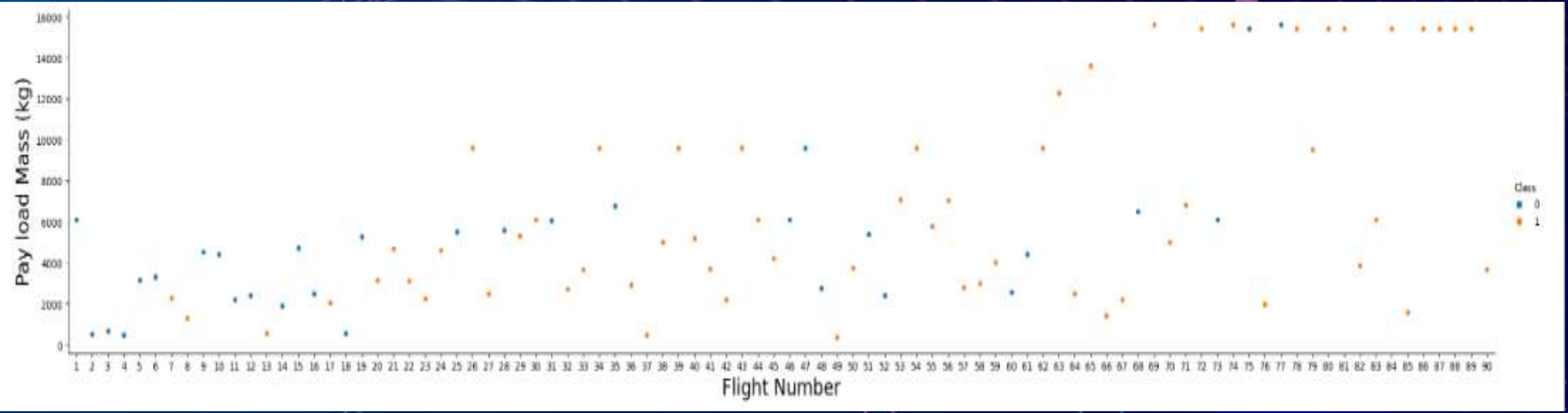
EDA → Summarizations → Creation of Landing Outcome Label

Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight, Payload and Orbit



Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space misión
- Top 5 launch sites whose name begin with the string "CCA"
- Total payload mass carried by boosters launched by NASA (CRS)
- Avarage payload mass carried by boosyer versión F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters wich have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure misión outcomes
- Names of the booster versiona which have carried the máximum playload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in 2015
- Rank of the count of landing outcomes, such as Failure (drone ship) or Success (ground pad) between the 2010/06/04 and 2017/03/20

Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/SQL%20and%20EDA.ipynb

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium Maps:

- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between to coordinates

Source code: https://github.com/MarcosLopez64/Capstone-Project-/blob/main/_Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list anable Launch Site selection

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

- Added a pie chart to show the total successful launches count for al 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

Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/spacex_dash_app.py

Predictive analysis

Creating a NumPy array from the column “Class” in data

Standardizing the data with StandarScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_Split function

Creating a GridSearchCV object with cv=10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the yesy data using the method .score() for all models

Examining the confusión matrix for all models

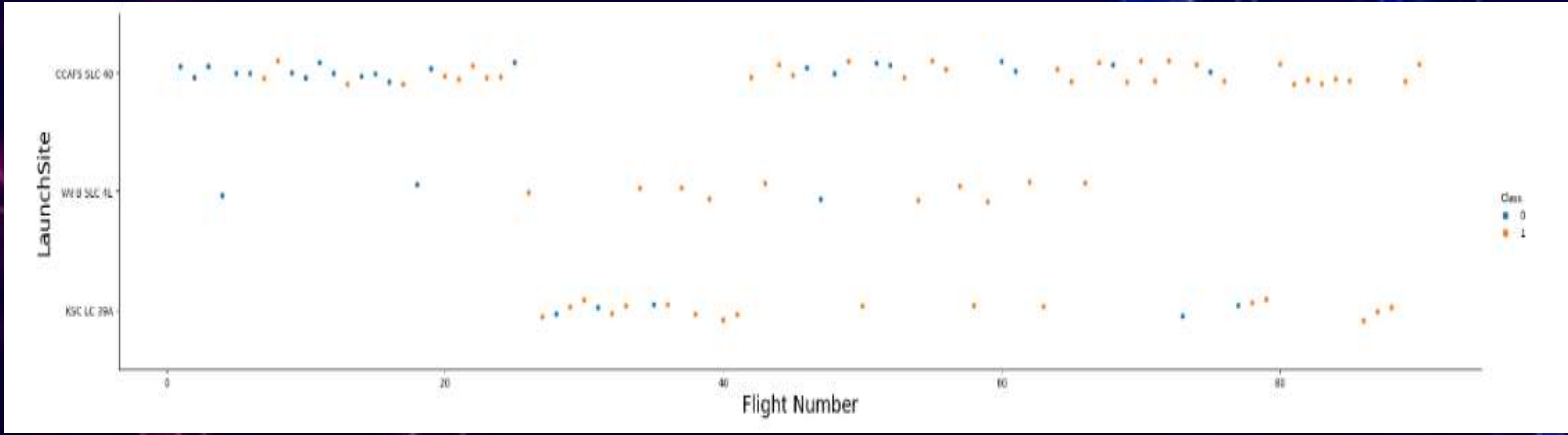
Finding the method performs best by examining the Jaccard_score and F1_score

Source code:
https://github.com/MarcosLopez64/Capstone-Project-/blob/main/Machine%20Learning%20Prediction.ipynb

Results

EDA with Visualization

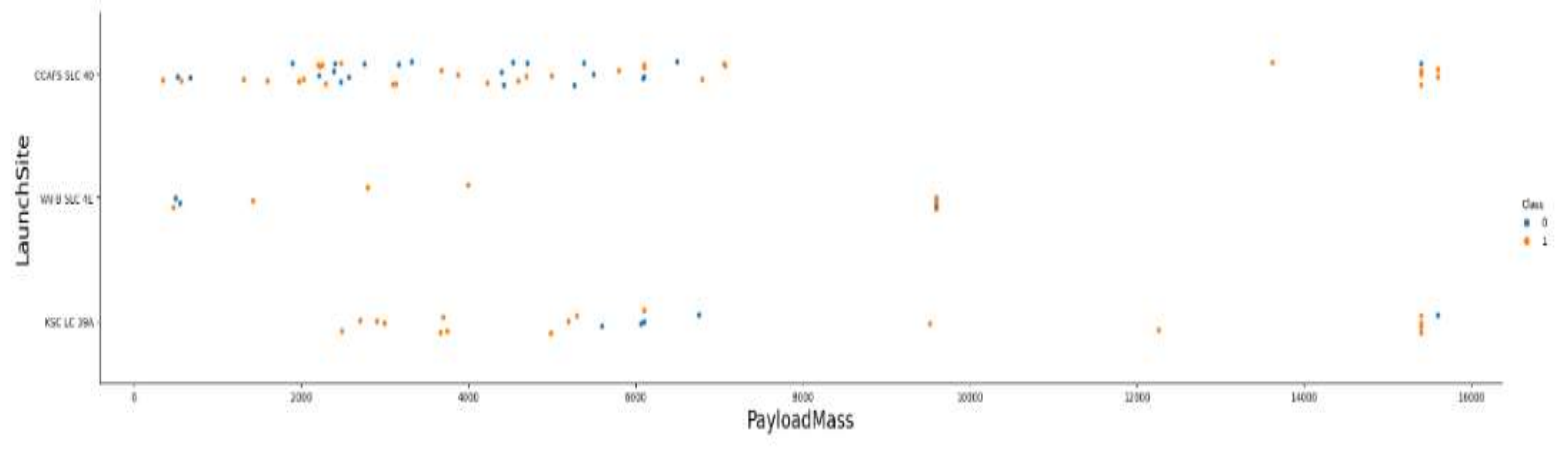
Flight Number vs. Launch Site



Results:

- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about a half of all launches
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success

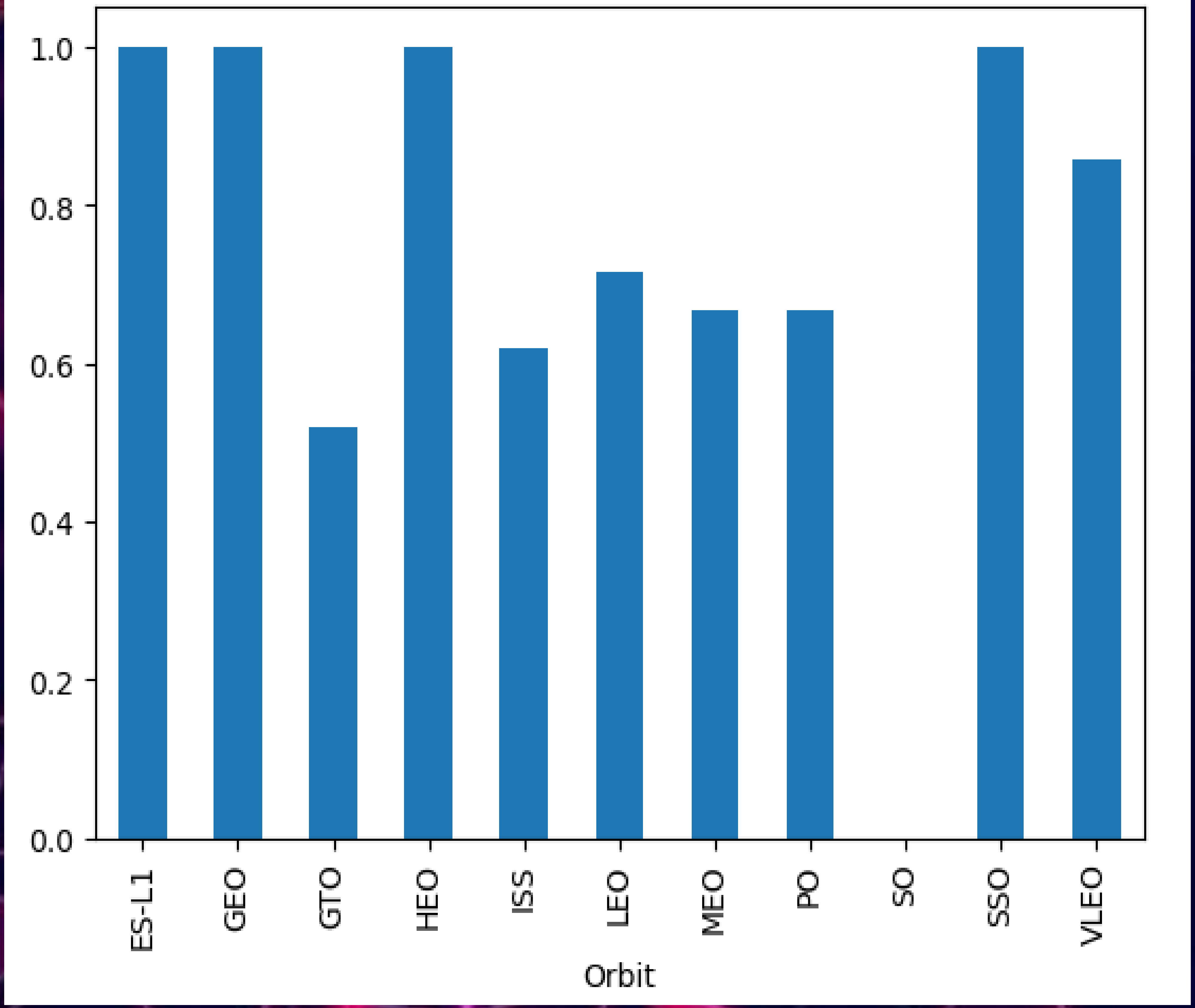
Payload vs. Launch Site



Results:

- For every launch site the higher the payload mass, the higher the success rate
  - Most of the launches with payload mass over 7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

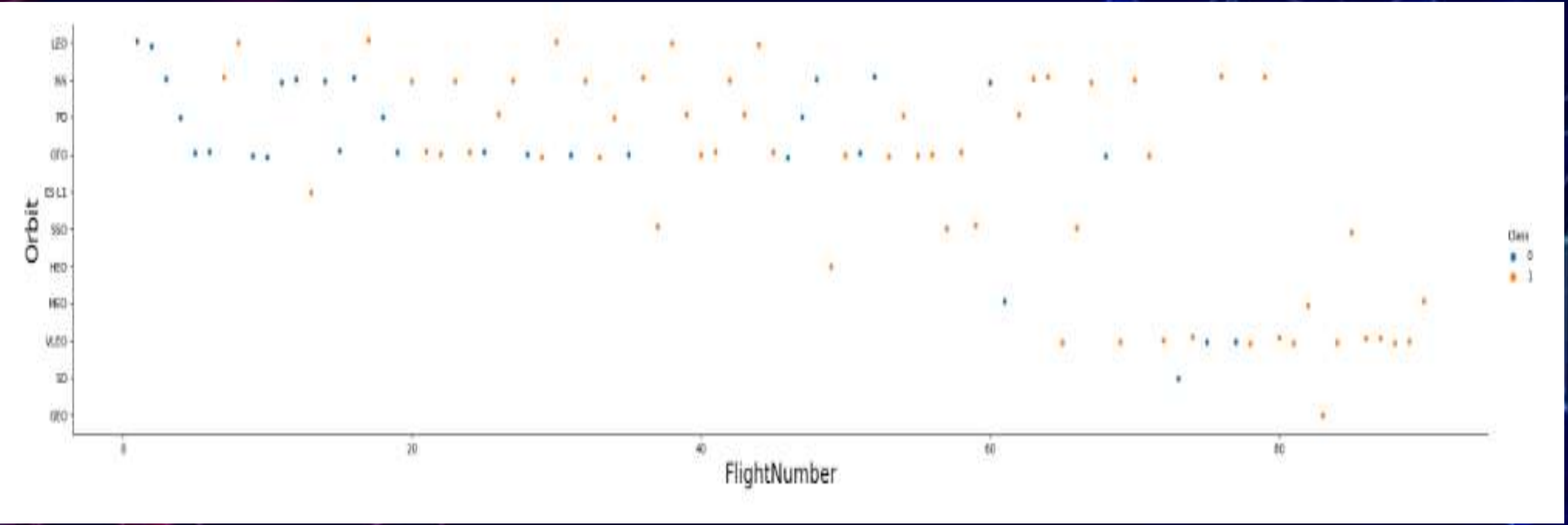
Success rate vs. Orbit type



Results:

- ES-L1, GEO, HEO, SSO = 100% success rate
- SO = 0% success rate
- GTO, ISS, LEO, MEO, PO = 50% - 85% success rate

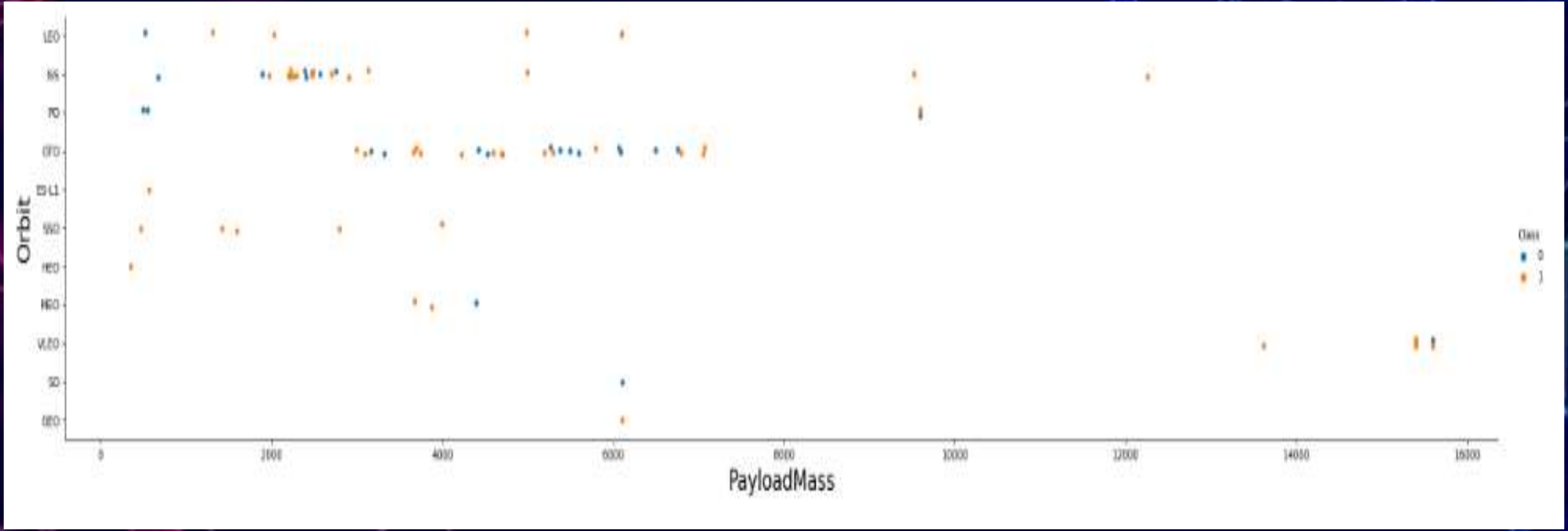
Flight Number vs. Orbit type



Results:

- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

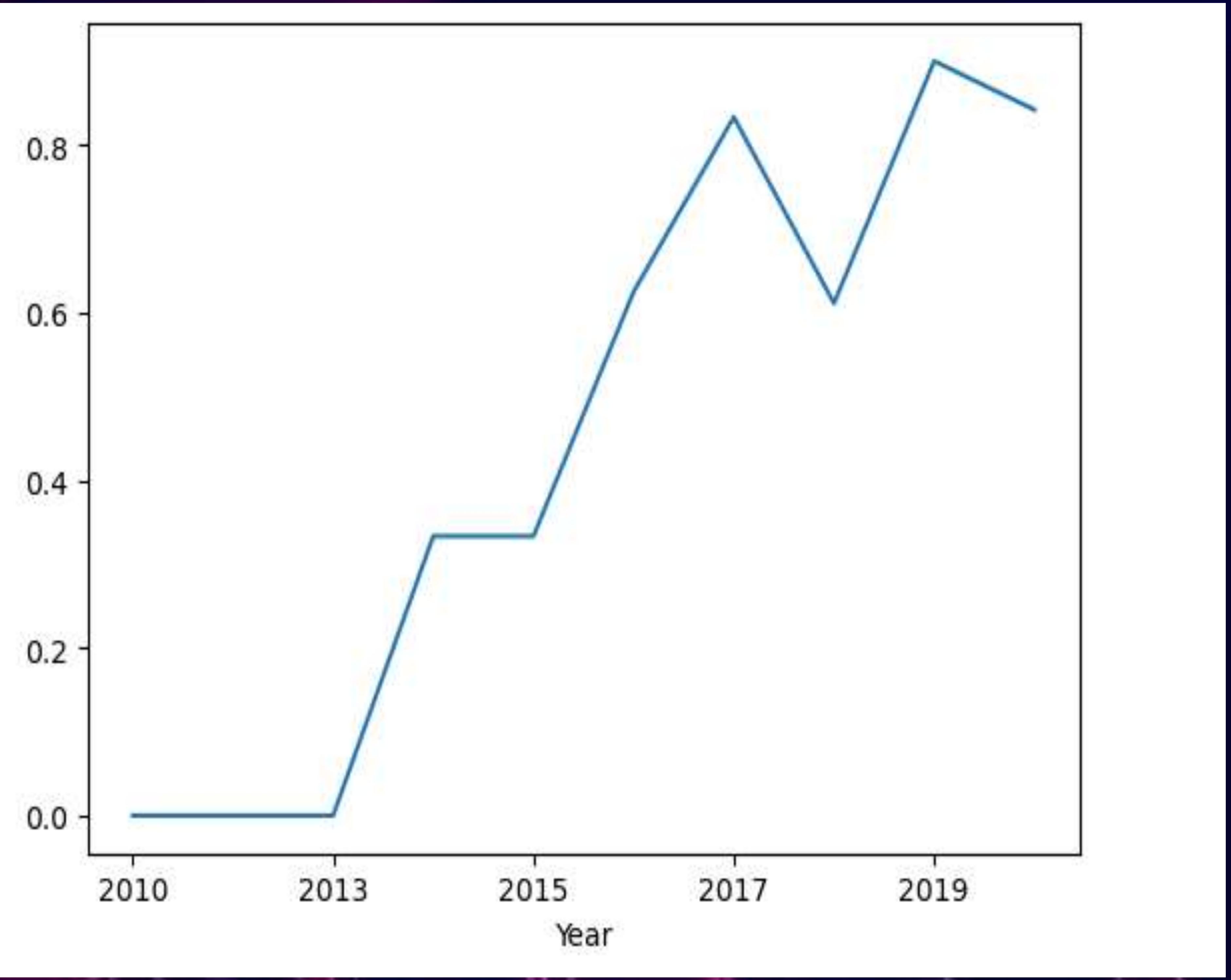
Payload Mass vs. Orbit type



Results:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch success yearly trend



Results:

- The success rate since 2013 kept increasing till 2020

EDA with SQL

All launch site names

Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

They are obtained by selecting unique occurrences of “launch_site” values from the dataset

Launch site names begin with “CCA”

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- Five samples of Cape Canaveral launches

Total Payload Mass

TOTAL_PAYLOAD
111268.0

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

Average payload mass by F9 v1.1

AVG_PAYLOAD
2928.4

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

First successful ground landing date

FIRST_SUCCESS_GP
01/08/2018

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

Successful drone ship landing with payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total number of successful and failure mission outcomes

Mission_Outcome	QTY
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Listing the total number of successful and failure mission outcomes

Successful drone ship landing with payload between 4000 and 6000

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

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

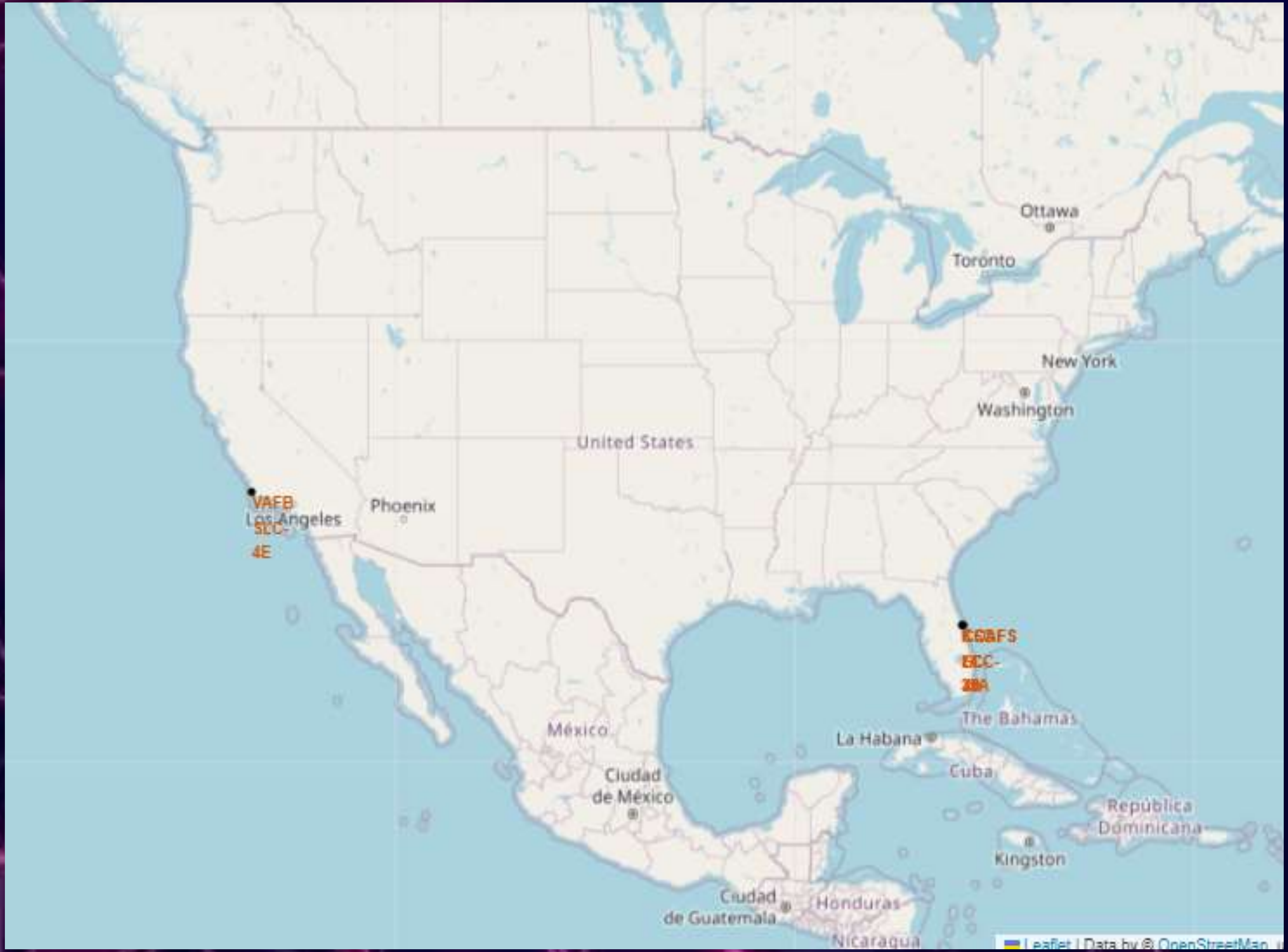
2015 Launch Records

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

Interactive Maps with Folium

All launch sites' location markers on a global map

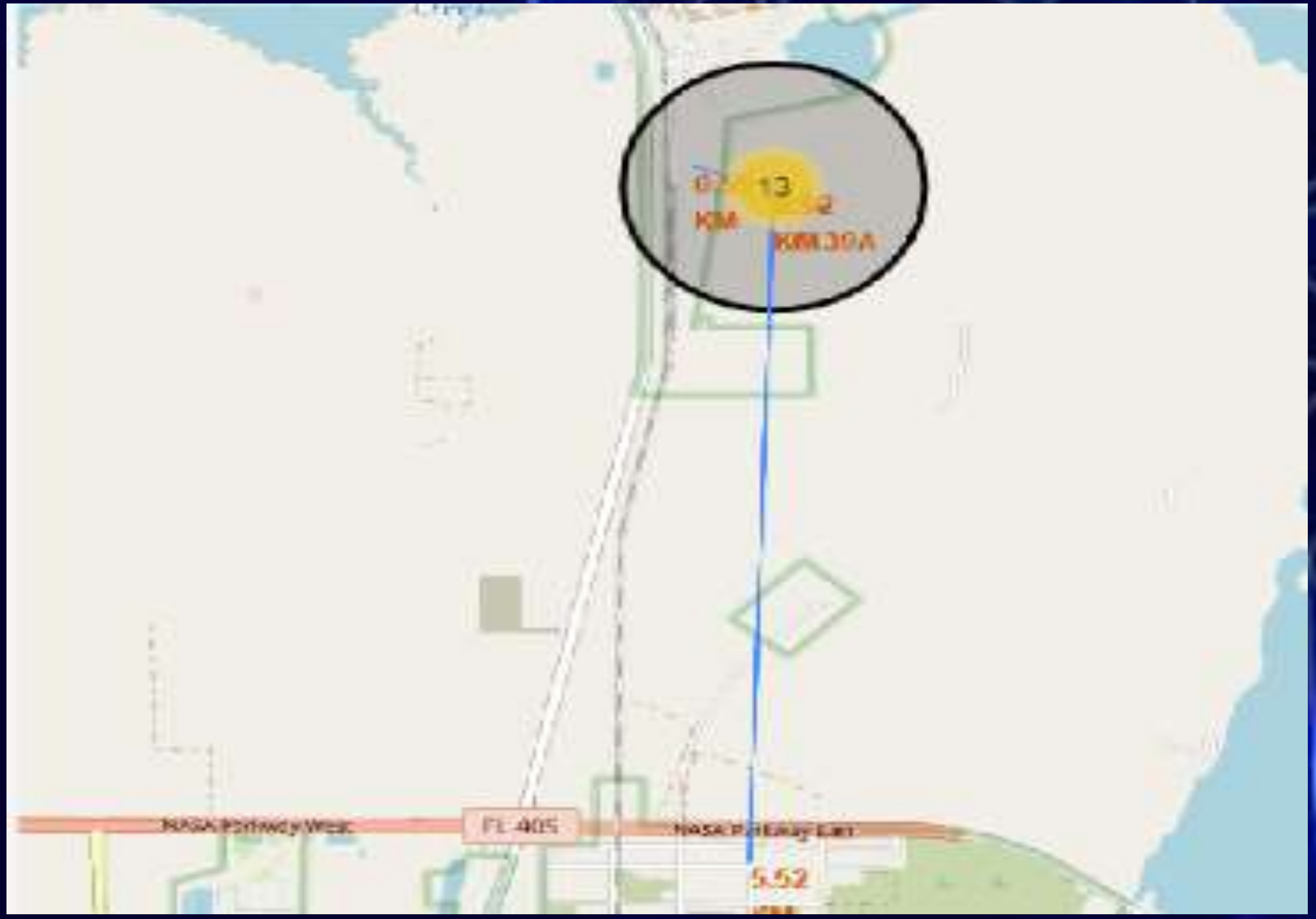


- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people

Launch Outcomes by Site

- In this example green markers indicate successful and red ones indicate failure

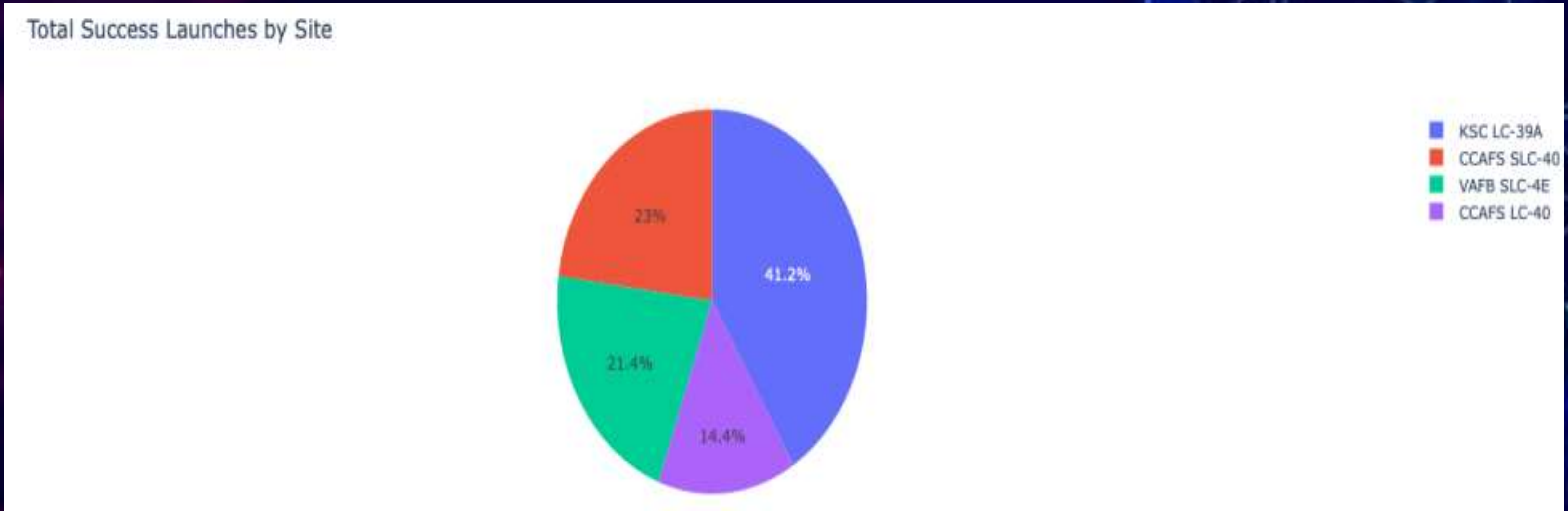
Logistics and Safety



- Launch sites are usually uninhabited areas and close to railroads and roads, so they have optimal logistics and security.

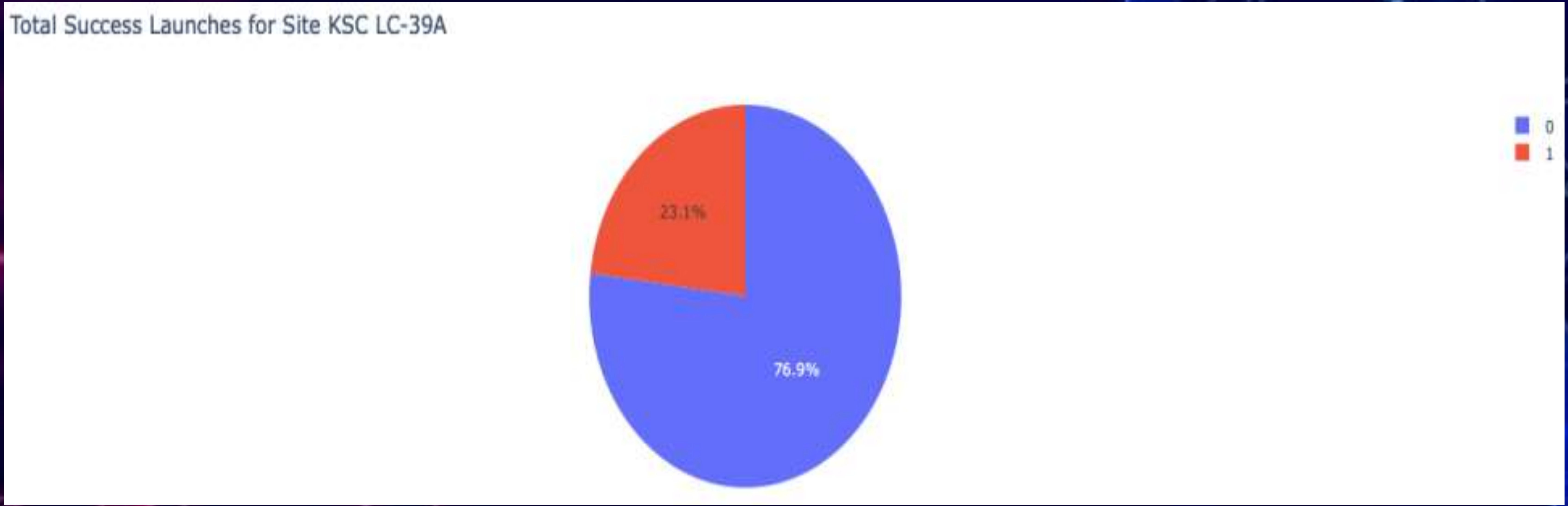
Build a Dashboard with Plotly Dash

Launch success count by site



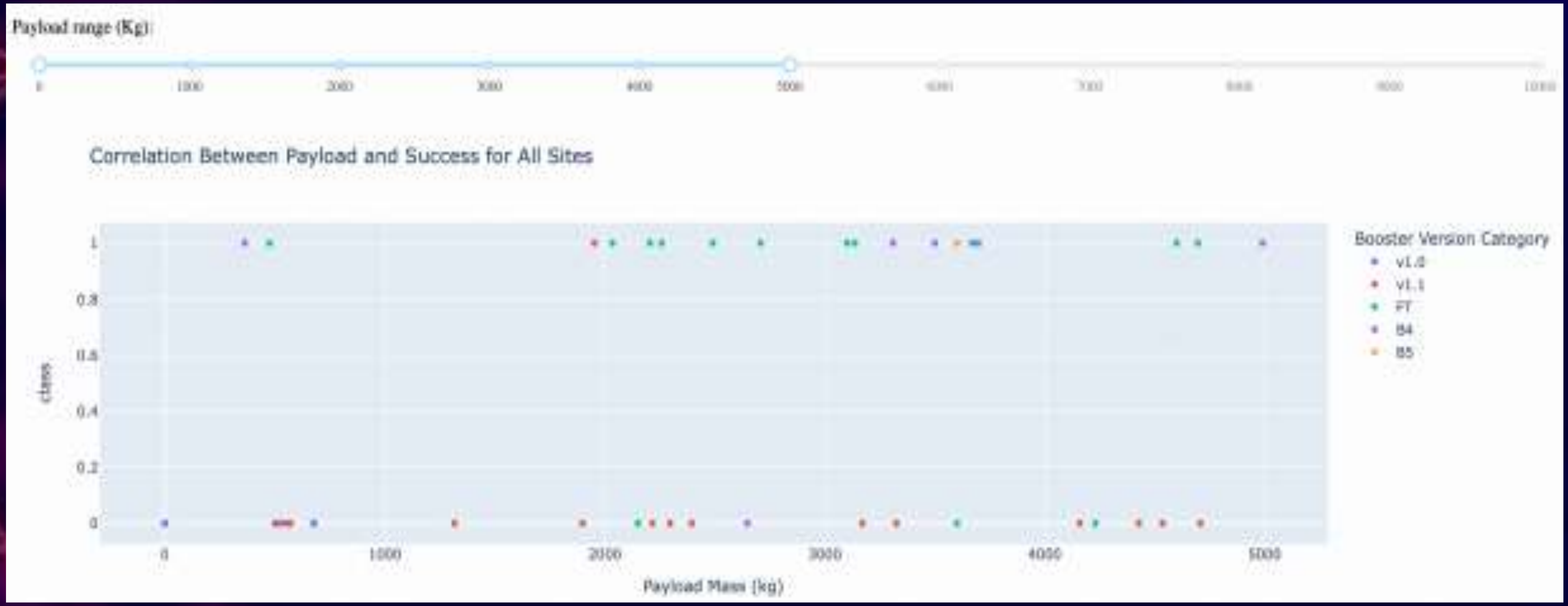
- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

Payload Mass vs. Launch Outcome for all sites



- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

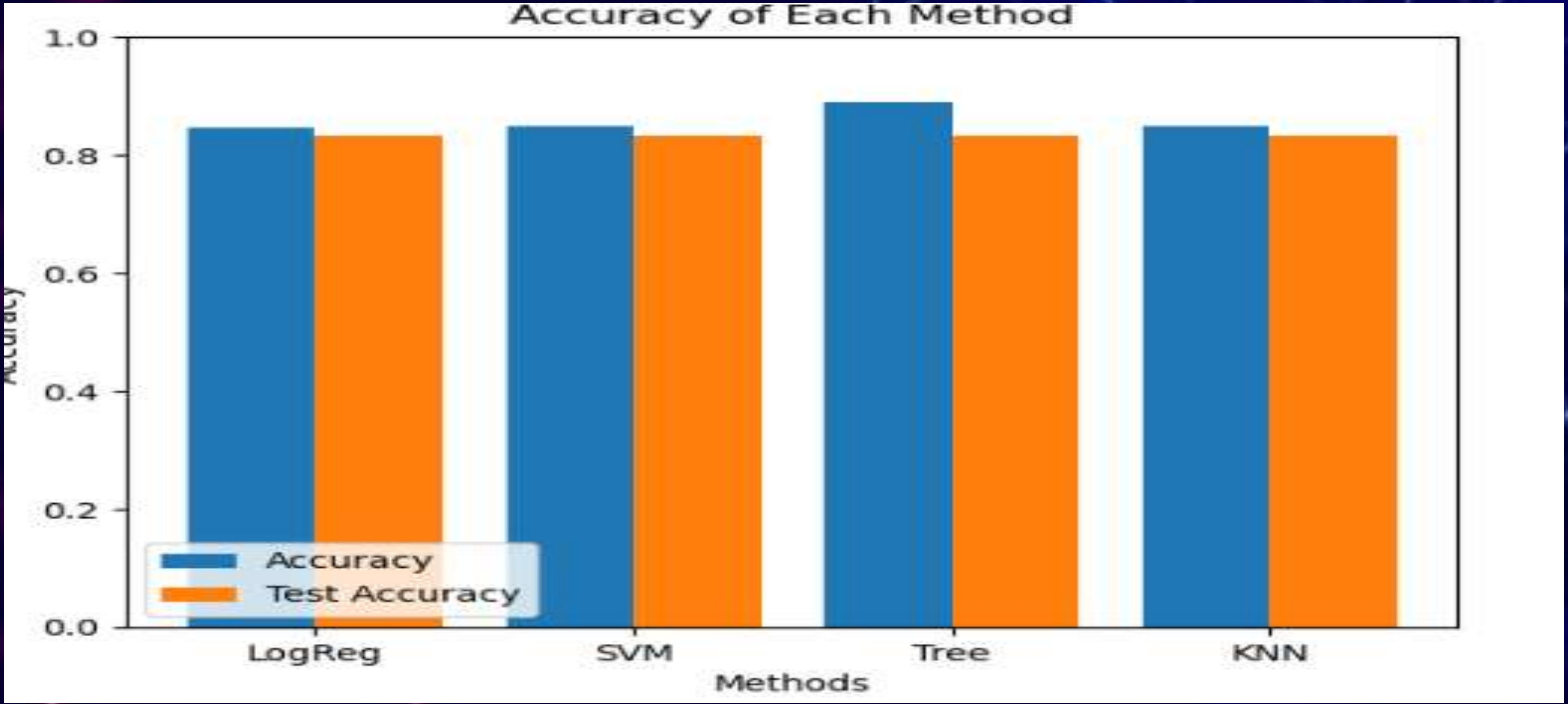
Predictive Analysis (Classification)

Classification Accuracy

As you can see on the next slide, we used four methods:

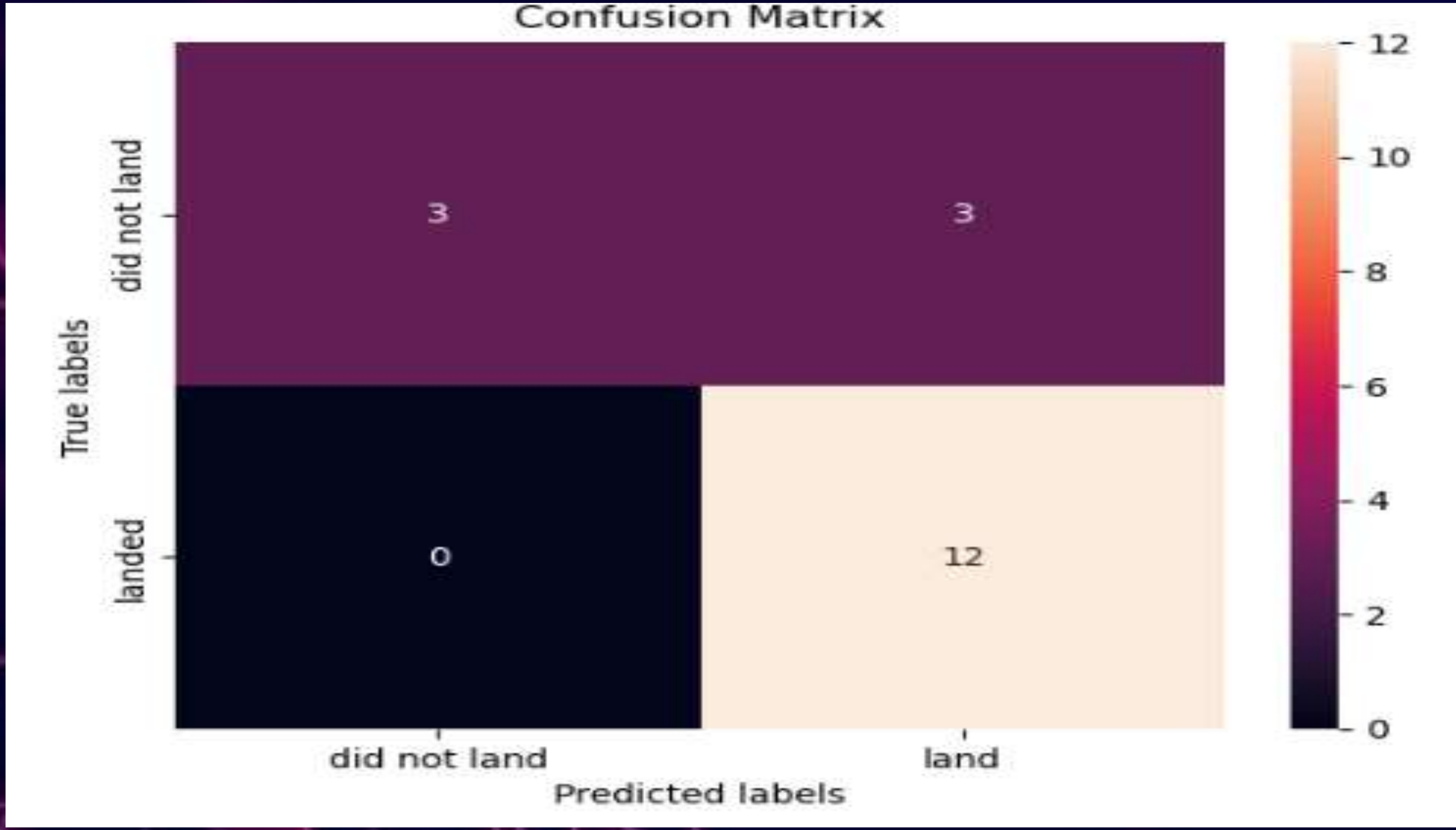
1. Logistic Regression
2. Support Vector Machines
3. Decision Trees
4. K-Nearest Neighbors

These methods obtained similar results. However, the best was the Decision Tree that obtained a success of 89%.



Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333

Confusion Matrix of Decision Tree Classifier



- Confusion matrix of Deicision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to false ones

Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.
- KSC LC-39A has the highest success rate of the launches from all the sites.

Appendix

Special thanks to: