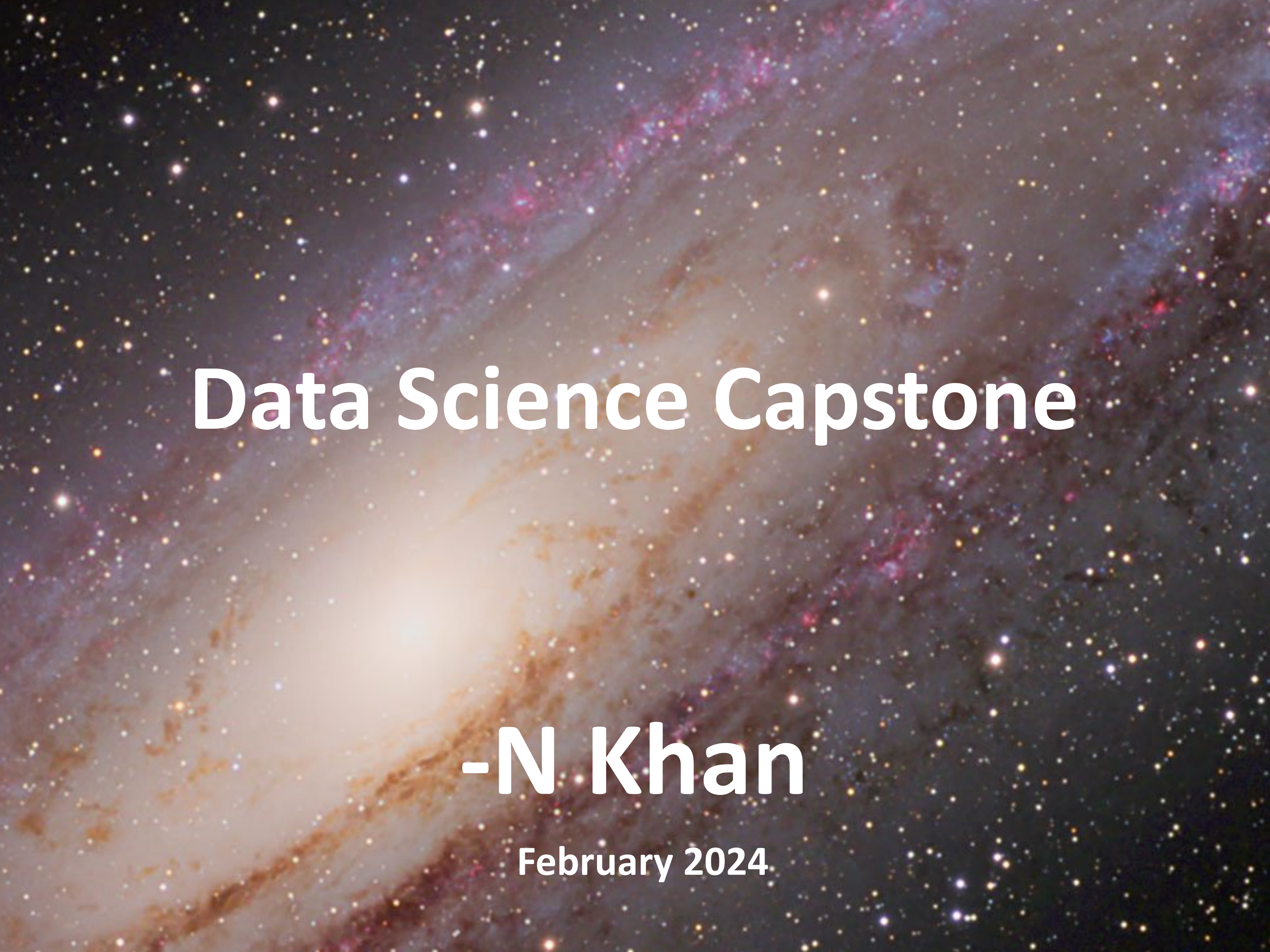


Data Science Capstone

-N Khan

February 2024

Outline:

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

Executive Summary:

- Summary of methodologies:

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

- Summary of all results:

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics results

Part-1
Methodology

Methodology:

Summary:

• Data collection methodology:

- Data was collected using SpaceX API and web scraping from Wikipedia.

• Perform data wrangling:

- One-hot encoding was applied to categorical features

• Perform exploratory data analysis (EDA) using visualization and SQL

• Perform interactive visual analytics using Folium and Plotly Dash

• Perform predictive analysis using classification models

- To build, tune, evaluate classification models

Data Collection:

- The data was collected using various methods:

- Data collection was done using get request to the SpaceX API.

- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize()

- We then cleaned the data, checked for missing values and fill in missing values where necessary.

- The objective was to extract the launch records as HTML table.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- The link to the notebook-

https://github.com/n-khan20/Data-Science-Cap/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup

- The link to the notebook:

https://github.com/n-khan20/Data-Science-Cap/blob/main/DataCollection(a).ipynb

Data Wrangling

- 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
- The link to the notebook:

https://github.com/n-khan20/Data-Science-Cap/blob/main/DataCollection(webScraping-b).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.

- The link to the notebook:

https://github.com/n-khan20/Data-Science-Cap/blob/main/Data%20Visualisation.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.

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.

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)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.

- We built different machine learning models and tune different hyperparameters using GridSearchCV.

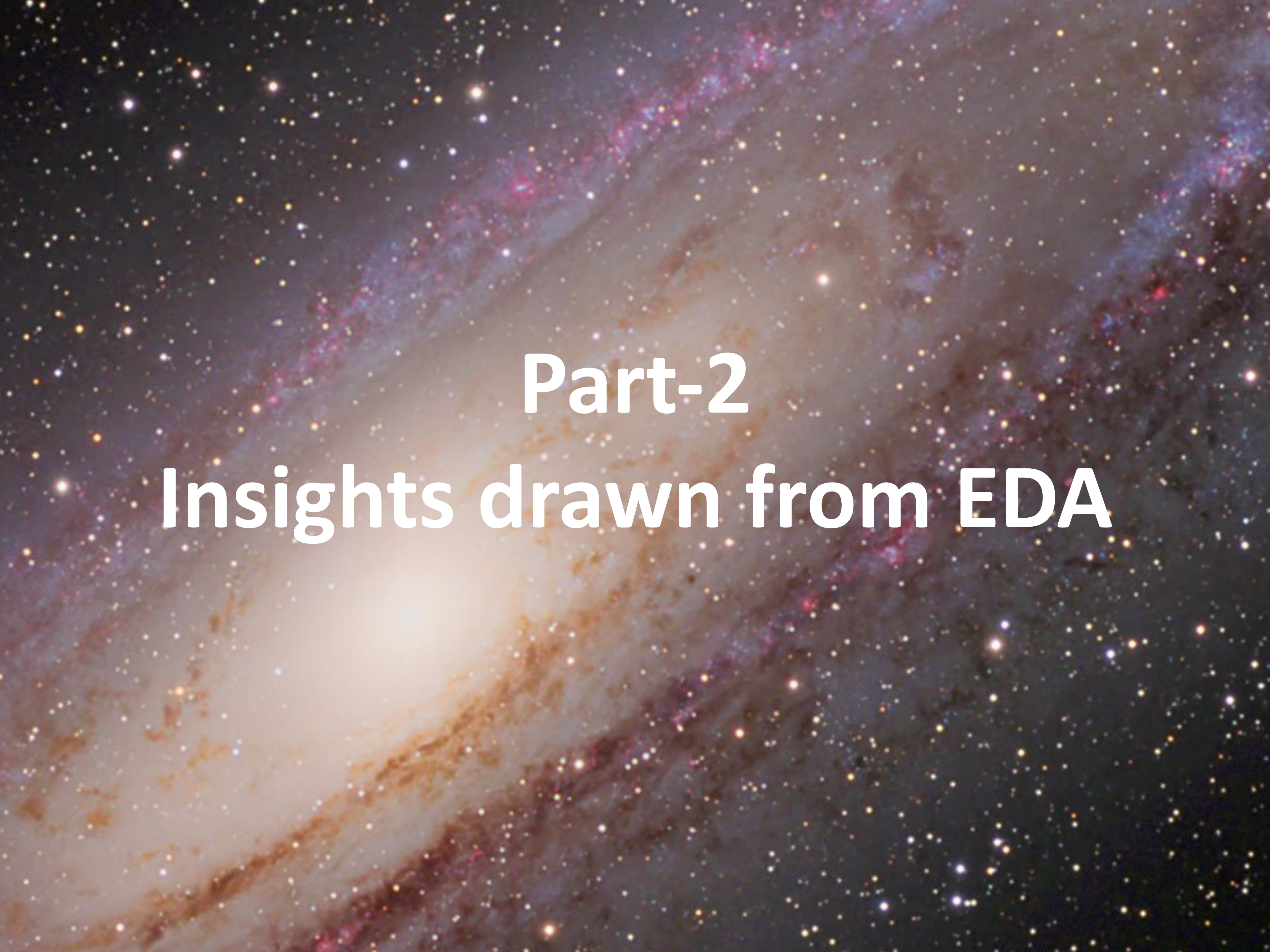
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

- We found the best performing classification model.

- The link to the notebook is https://github.com/chuksoo/IBM-Data-Science-Capstone-SpaceX/blob/main/Machine%20Learning%20Prediction.ipynb

Results

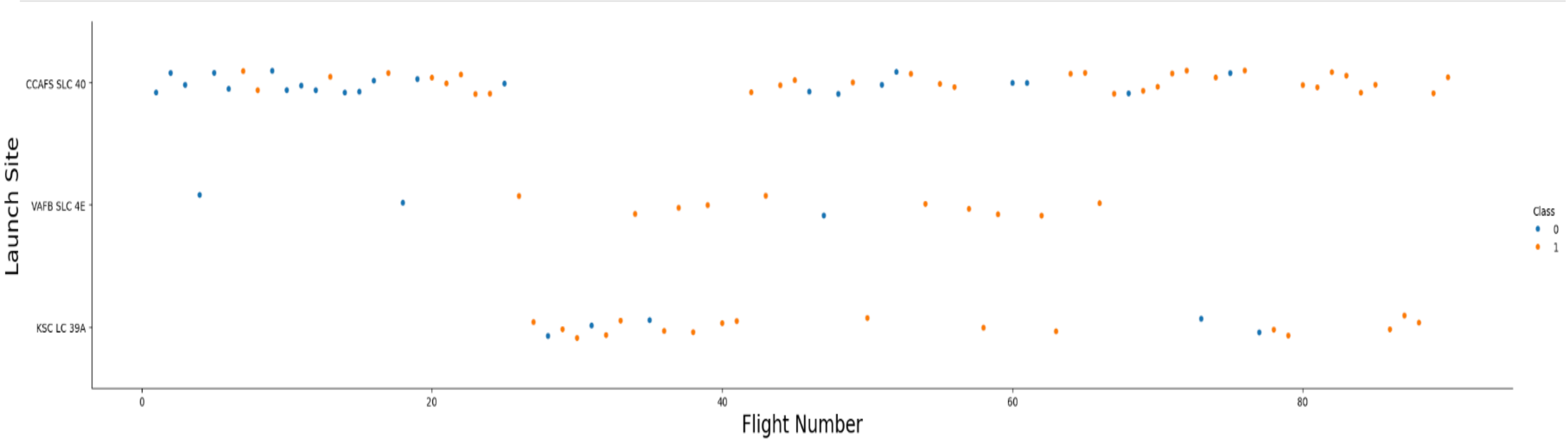
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Part-2
Insights drawn from EDA

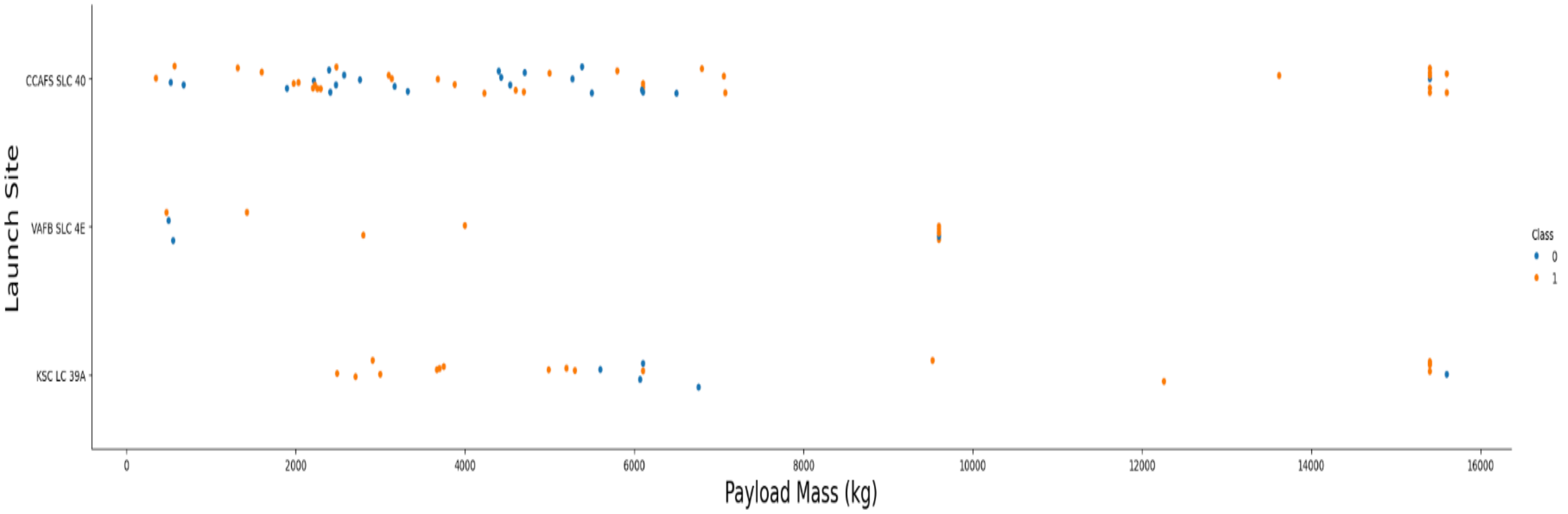
Flight Number vs. Launch Site

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

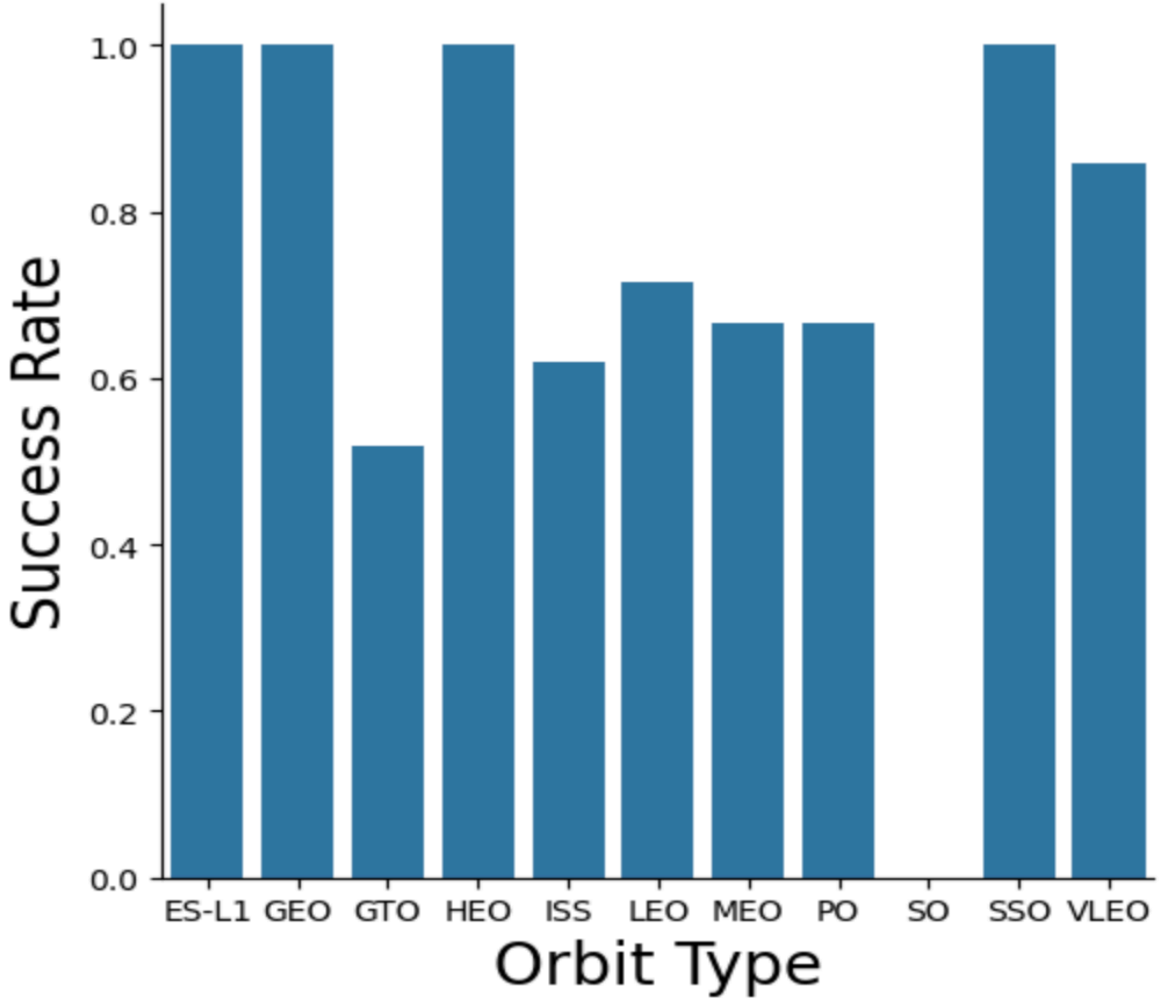


Payload vs. Launch Site

- The greater the payload mass for launch site, the greater is the success rate for the rocket.

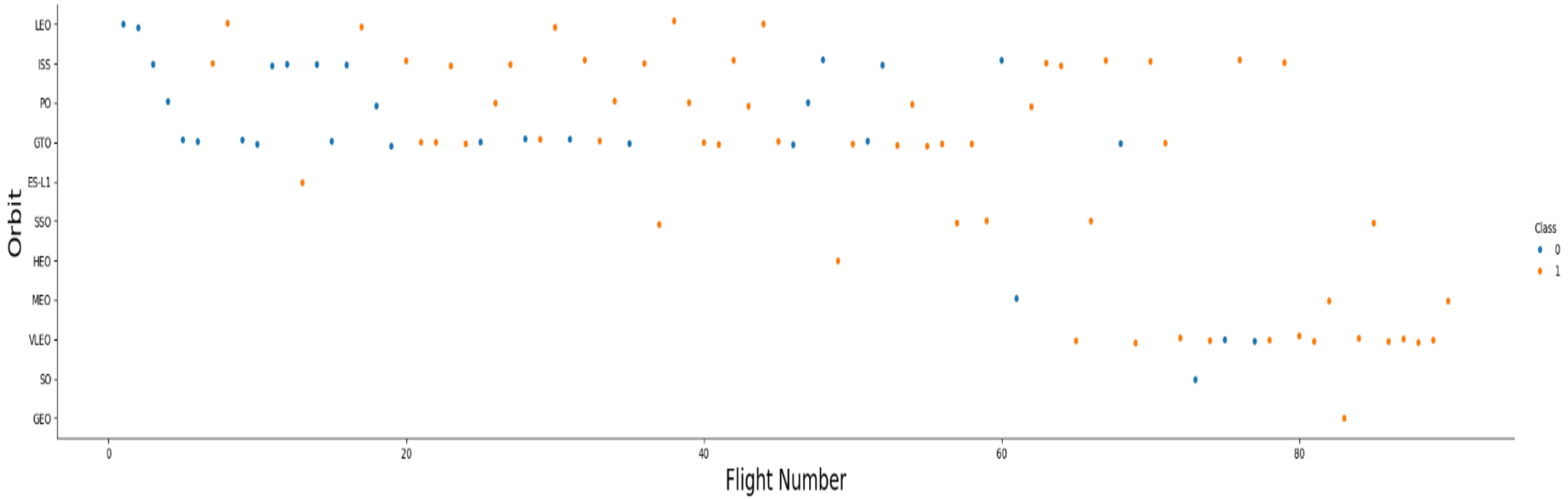


Success Rate vs. Orbit Type



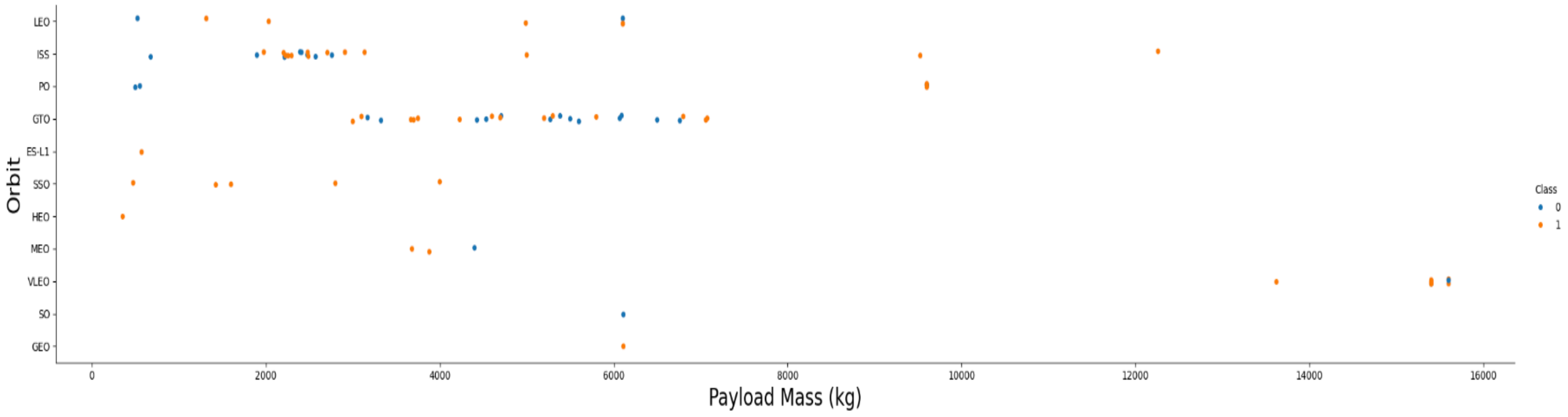
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type.



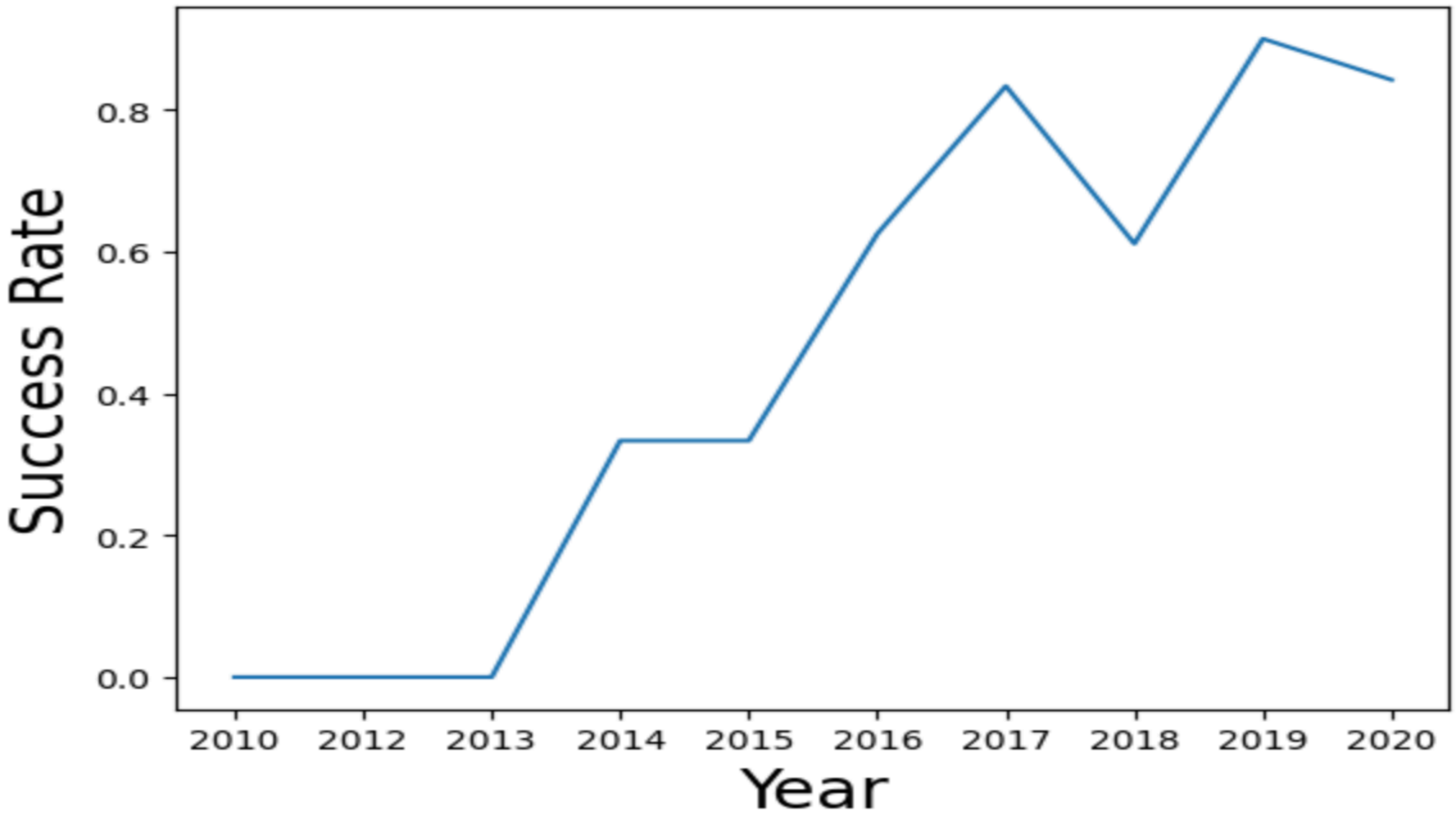
Payload vs. Orbit Type

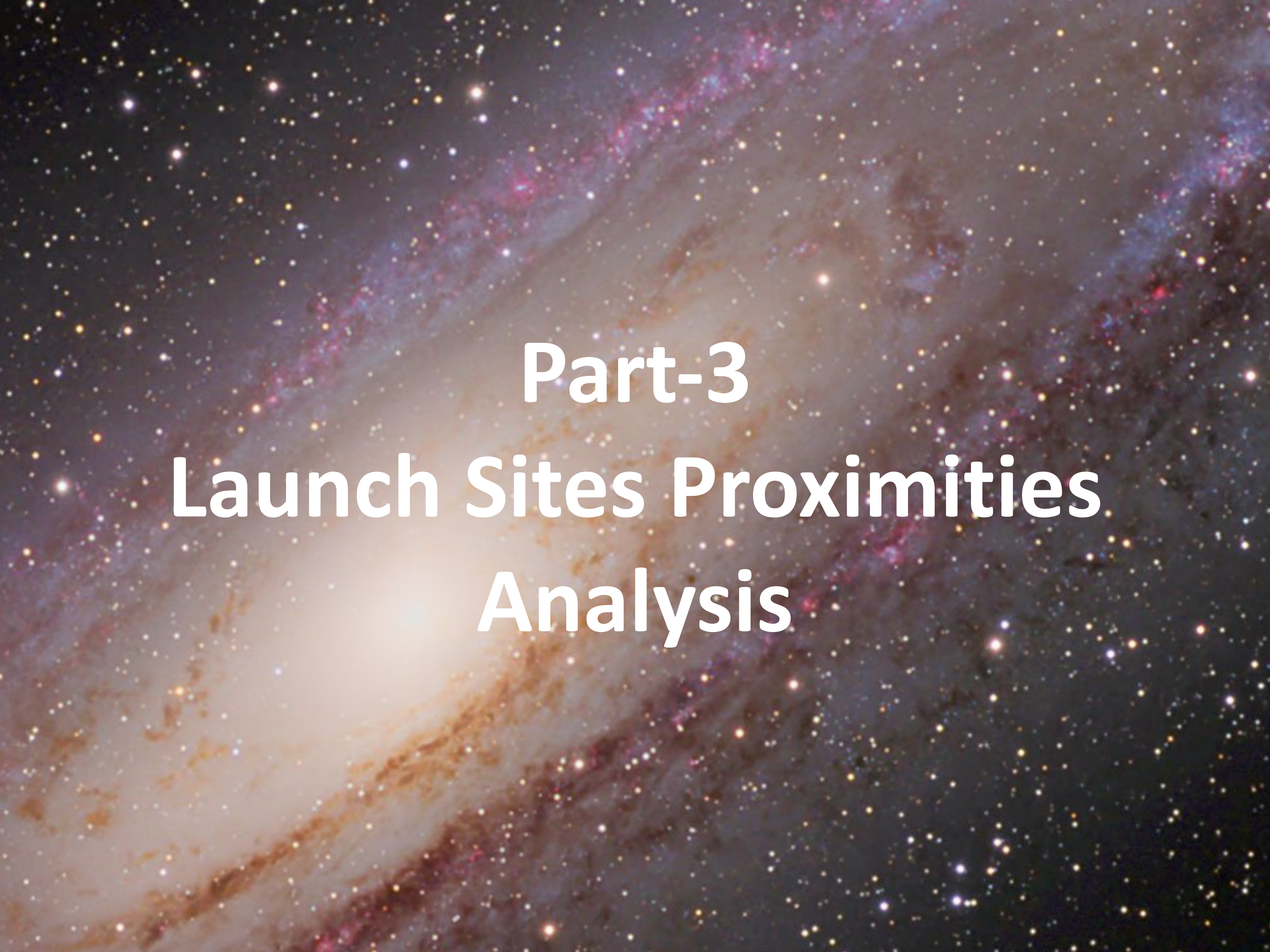
- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



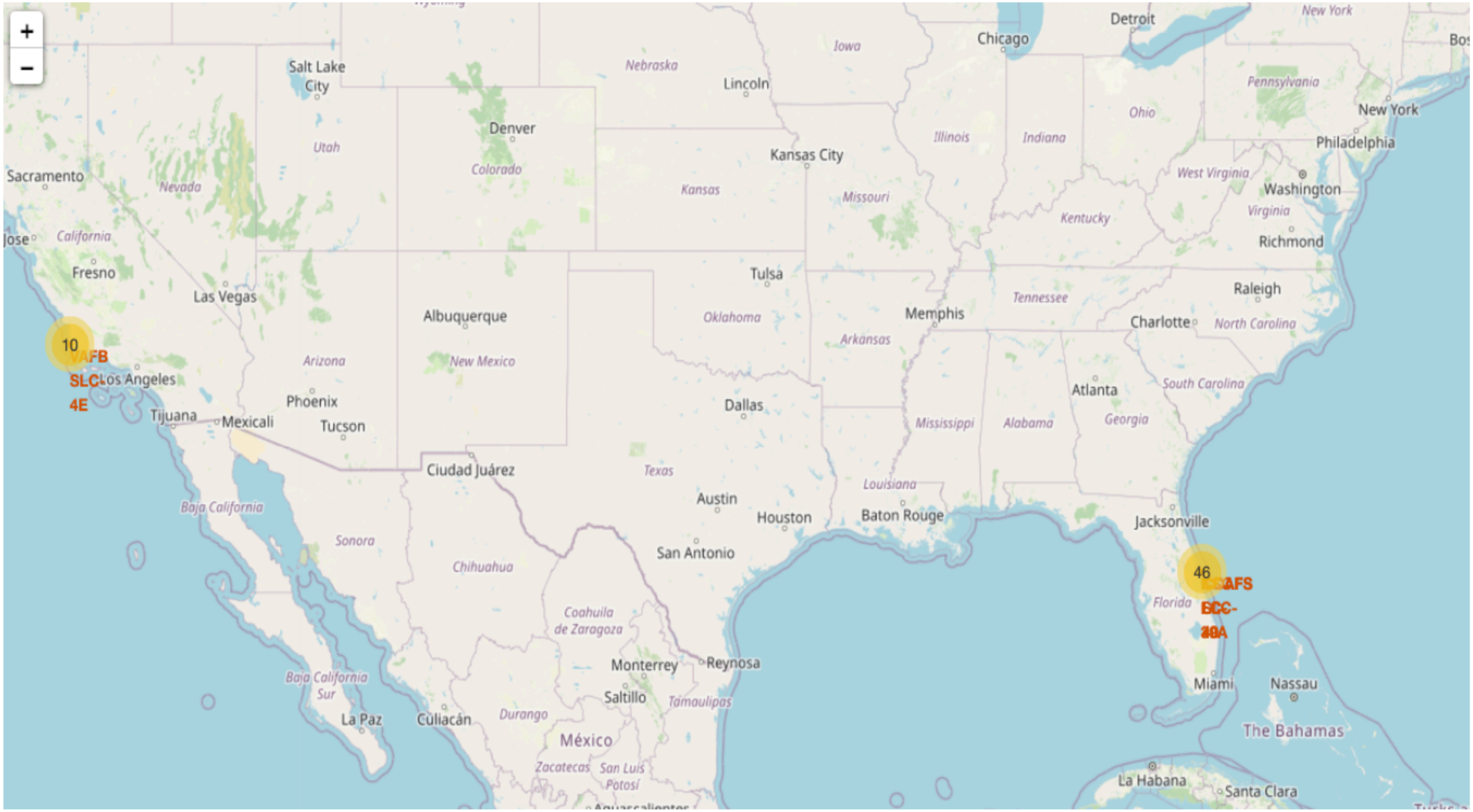
Launch Success Yearly Trend

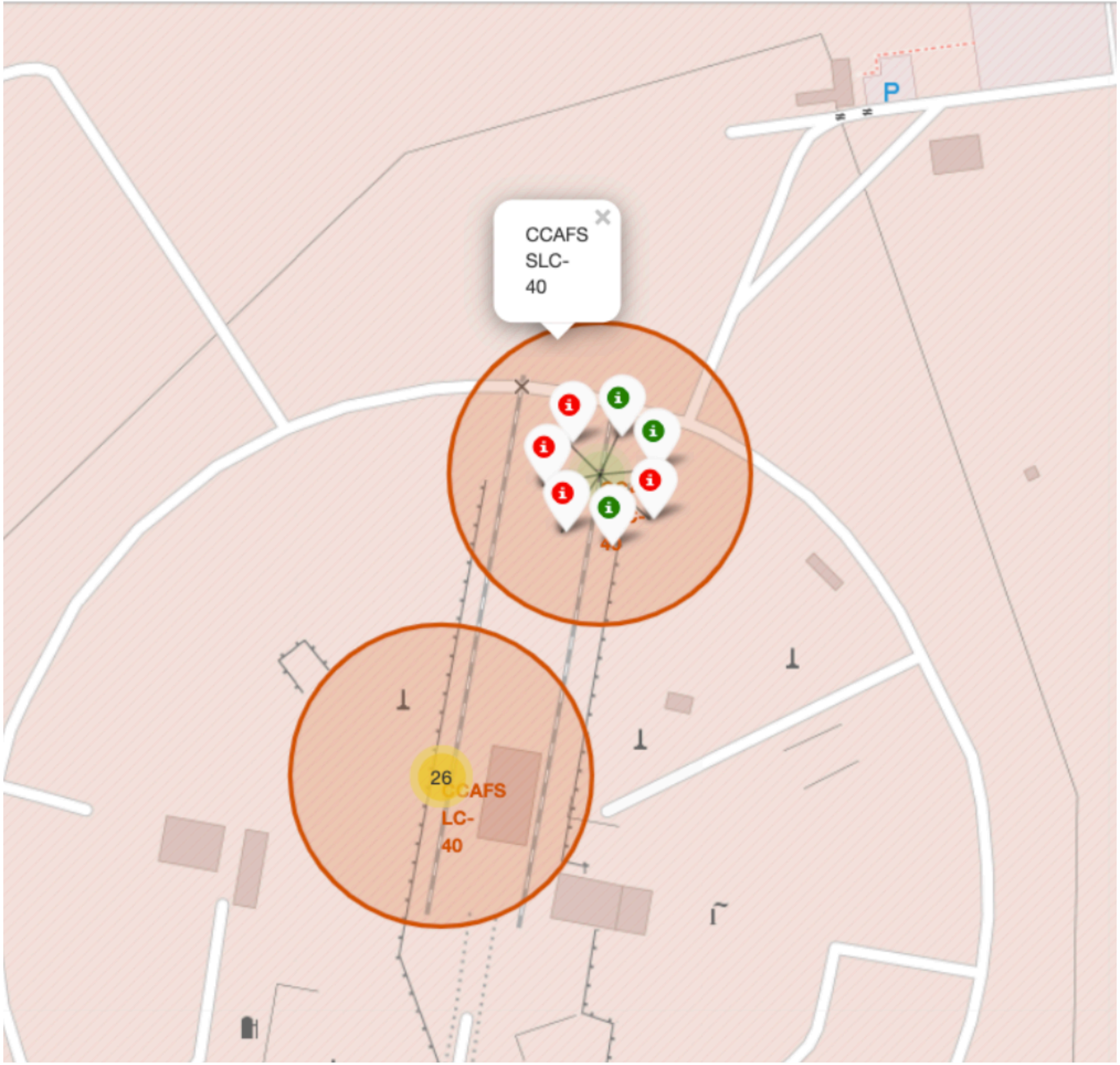
- From the plot, we can observe that success rate since 2013 kept on increasing till 2020

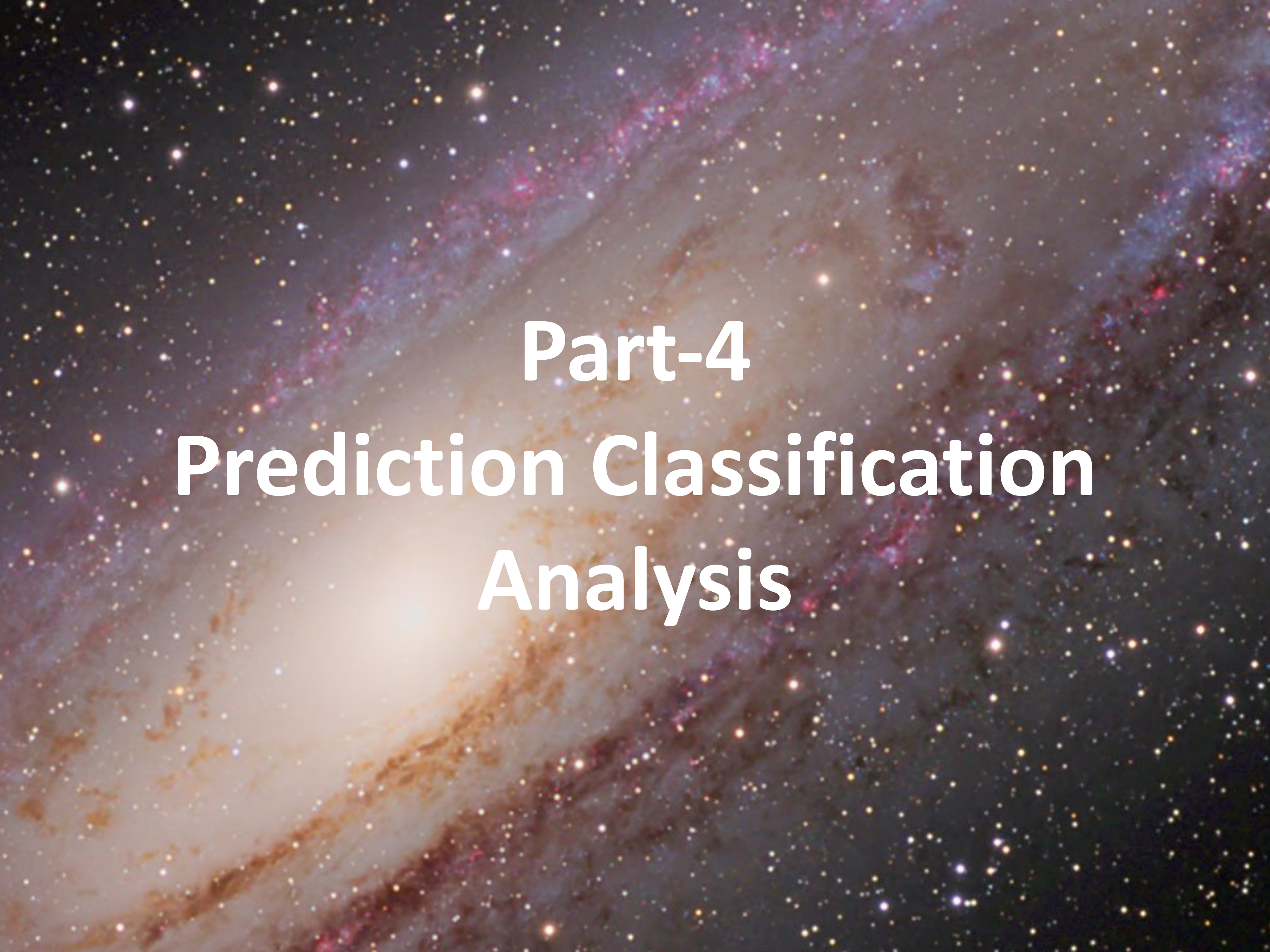




Part-3
Launch Sites Proximities Analysis

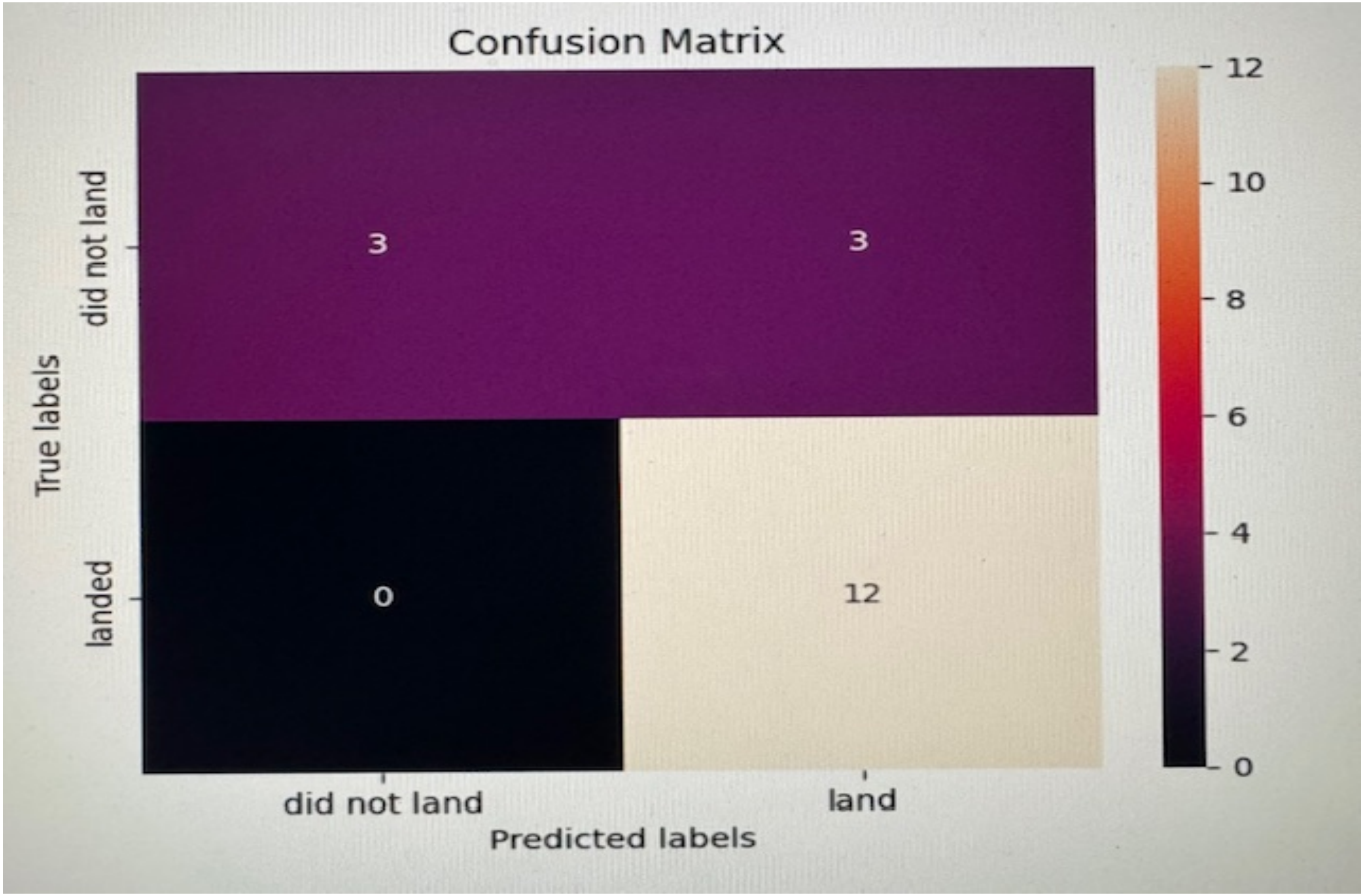

+
−
10
VAFB
SLC-
4E
46
CCAFS
SLC-
39A
Sacramento
Fresno
Las Vegas
Los Angeles
Tijuana
Mexicali
Phoenix
Tucson
Salt Lake City
Denver
Albuquerque
Ciudad Juárez
Lincoln
Kansas City
Tulsa
Dallas
Austin
Houston
San Antonio
Memphis
Baton Rouge
Chicago
Detroit
Atlanta
Charlotte
Raleigh
Richmond
Washington
Philadelphia
New York
Jacksonville
Miami
Nassau
The Bahamas
La Habana
Santa Clara
Monterrey
Saltillo
Reynosa
Culiacán
La Paz
México


CCAFS
SLC-
40
26
CCAFS
LC-
40
P



Part-4
Prediction Classification Analysis

Confusion Matrix for the Decision Tree



	did not land	land
did not land	3	3
landed	0	12

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

Thank you!