Winning Space Race with Data Science

Luciano Nieves
March 31, 2024

Outline

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

Executive Summary

- Different approaches for analyzing data, such as:
  - SpaceX API usage for Data Collection, besides web scraping.
  - Exploratory Data Analysis (known as “EDA”).
  - Data wrangling.
  - Data visualization.
  - Interactive visual analytics.
  - Machine Learning Prediction.

- Summary of all results:
  - All data was collected, standardized, normalized and graphically shown correctly, from launchings (successful and unsuccessful ones) to predictions based on the public information available.

Introduction

- The idea behind this project is to analyze if a new company has possibilities of competing against SpaceX, knowing that SpaceX has plenty of successful landings, big headquarters and a huge amount of budget for each project, besides the experience of its previous attempts and a big market cap.

- Based on the next calculations and predictions, we will be able to check if we can estimate total costs for launches (with both successful and unsuccessful attempts), on top of where would it be a good place for developing the company/project.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used the base SpaceX API → https://api.spacexdata.com/v4/launches
  - Web Scraping: https://en.wikipedia.org/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling:
  - Refined the collected data through different processes for better data understanding, such as normalization and filtering.

Methodology

Executive Summary

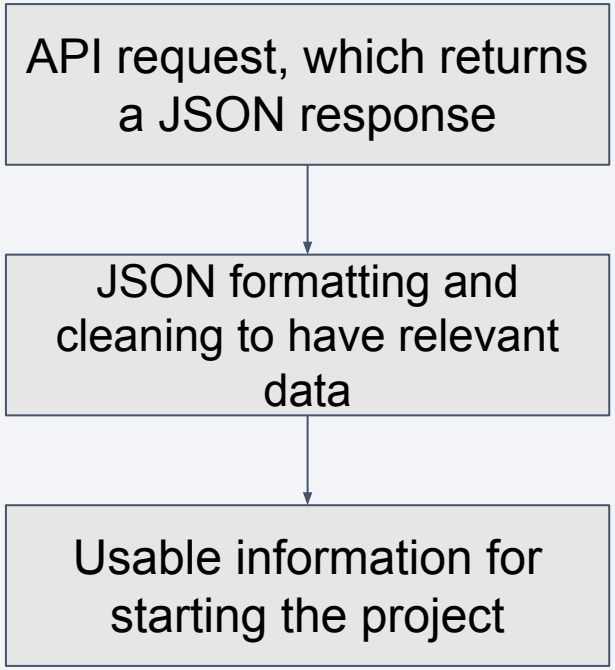
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models:
    - The idea behind this predictive analysis was to use the information in a way that I could train and test the data sets, on top of evaluating them with different models (where the accuracy of each model will depend on the kind of parameters I set when executing them).

Data Collection

- There were two main places where the data was collected from:
  - SpaceX API → https://api.spacexdata.com/v4/launches
  - Web Scraping → https://en.wikipedia.org/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

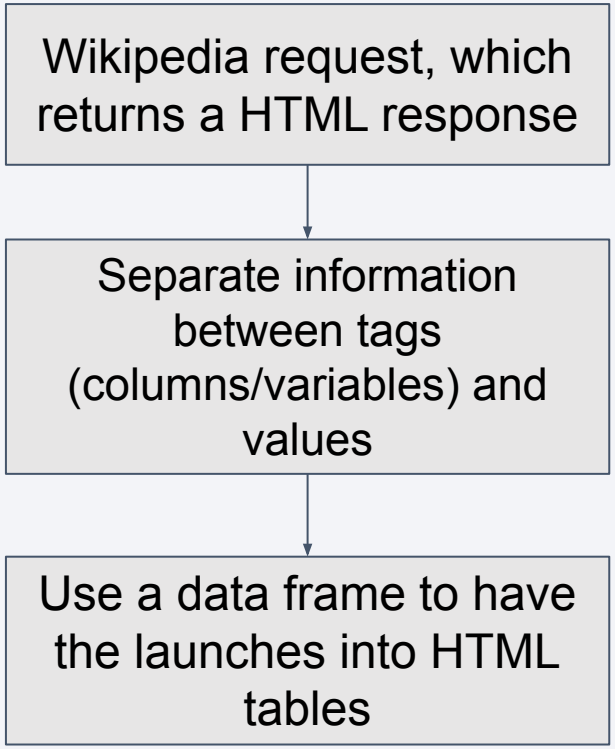
- SpaceX API is accessible to anyone that knows how to retrieve a JSON formatted string and convert it into what they need (no matter if it is with Python, JavaScript, or other languages), to later utilize the data for researching purposes.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%201/COLLECTING%20THE%20DATA/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

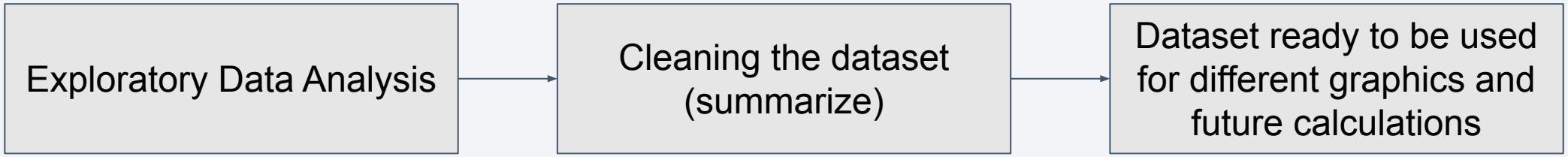
- SpaceX Wikipedia page has different versions for the information that provides, which in this case was to use an old version to have a standard. In this case, we can obtain the information by doing Web Scraping, which is obtaining the HTML to later obtain the tags and information between those tags.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%201/COLLECTING%20THE%20DATA/jupyter-labs-webscraping.ipynb

Data Wrangling

- With EDA, we summarize the information and work with a clean dataset, having the launches per site, outcomes from each one of the missions, what types of orbits were used and/or calculated for the attempts, etc.



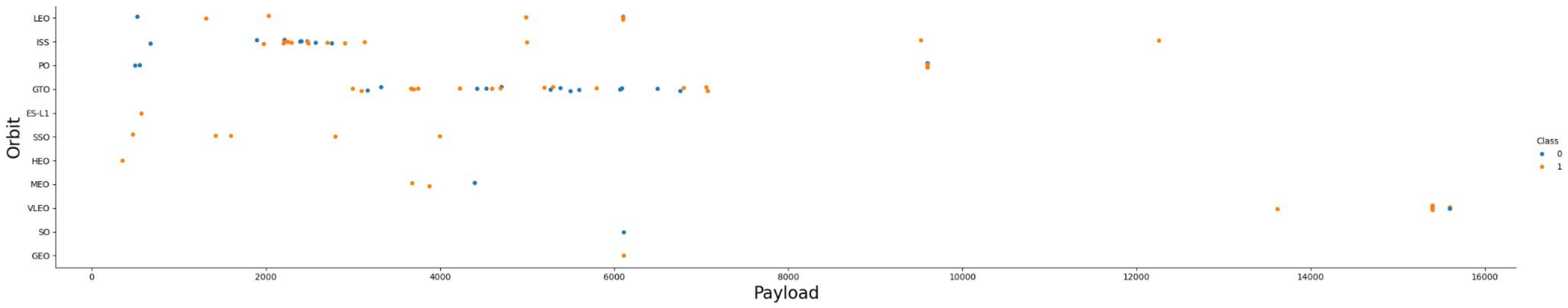
GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%201/DATA%20WRANGLING/labs-jupyter-spacex-Data%20wrangling.ipynb

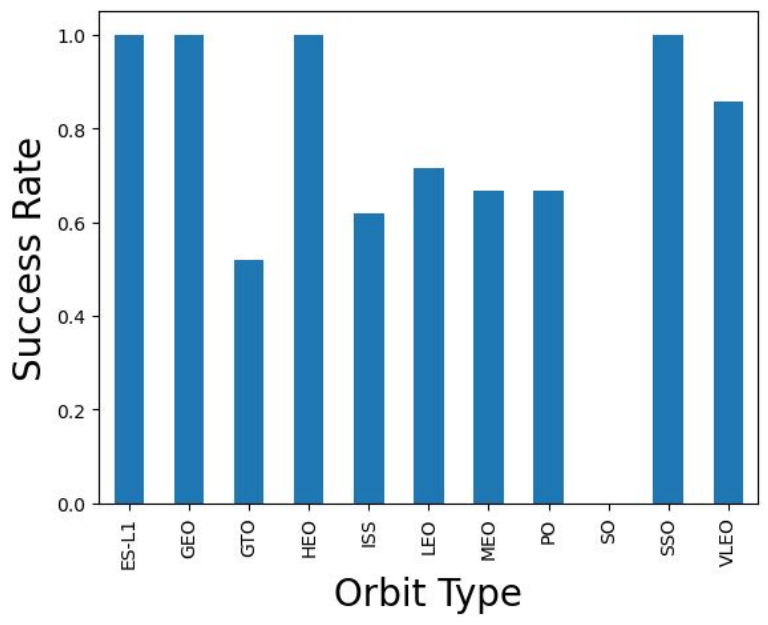
EDA with Data Visualization [1/2]

- There were multiple plots used, with each one of them representing different types of relationships, such as the payload mass (kg) and launch sites, payload mass (kg) and the orbits, and many more. I will attach two different plots for reference.

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

EDA with Data Visualization [2/2]





GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

EDA with SQL

- Some of the SQL queries used were:
  - Names for each one of the launch sites.
  - Total payload mass (kg) carried by boosters launched by "*NASA (CRS)*".
  - Total number of mission outcomes (success and failure).
  - Average payload mass (kg) carried by booster version "*F9 v1.1*".
  - Names of the boosters which have carried the max. payload mass (kg).

Note: full list of SQL queries used are available at the GitHub repository.

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium Maps had:
  - Markers (indicating each launch site).
  - Marker clusters (indicating groups of specific events, like launches in a specific place).
  - Lines (distances between two different points).
  - Circles (denoting the area for each important location).

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/VISUAL%20ANALYTICS%20WITH%20FOLIUM/lab_jupyter_launch_site_location.ipynb

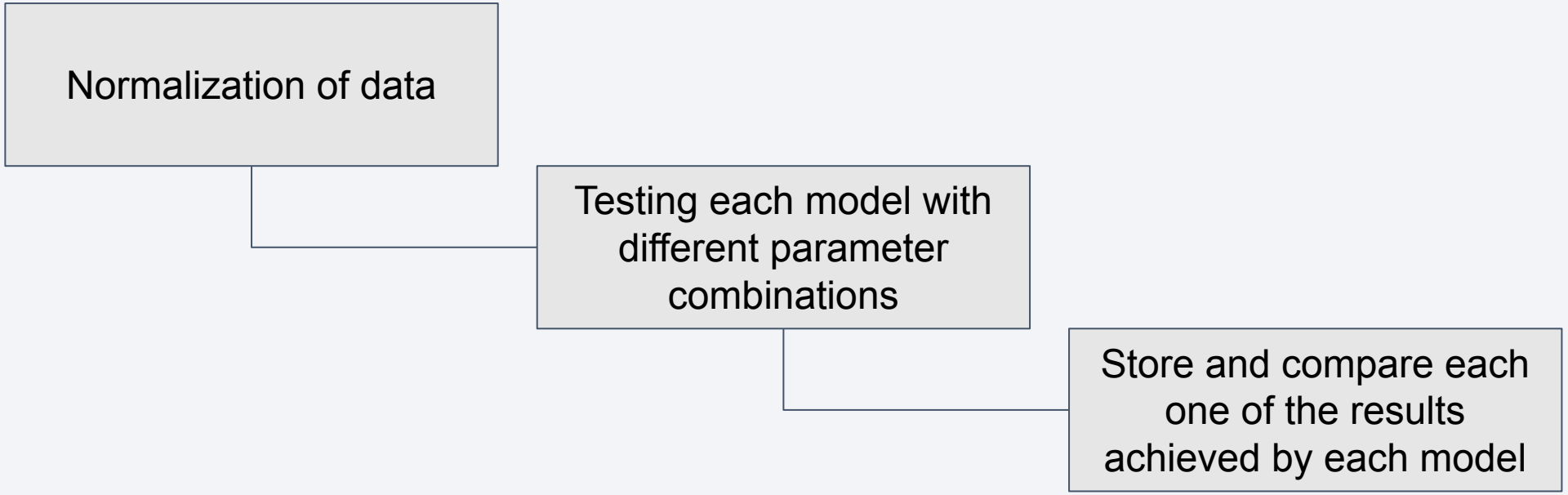
Build a Dashboard with Plotly Dash

- Plotly Dash had:
  - Graphs for visualizing percentages of each launch by specific sites.
  - Payload range (kg) for filtering.
  - Selection (dropdown button) with each one of the sites available to select (which will change the data from the graphics and plots).

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/INTERACTIVE%20DASHBOARD%20WITH%20PLOTY/spacex_dash_app.py

Predictive Analysis (Classification)

- There were a total of four different models for the predictive analysis stage, which were: *“support vector machine”, “decision tree”, “logistic regression”* and *“k nearest neighbor”*.

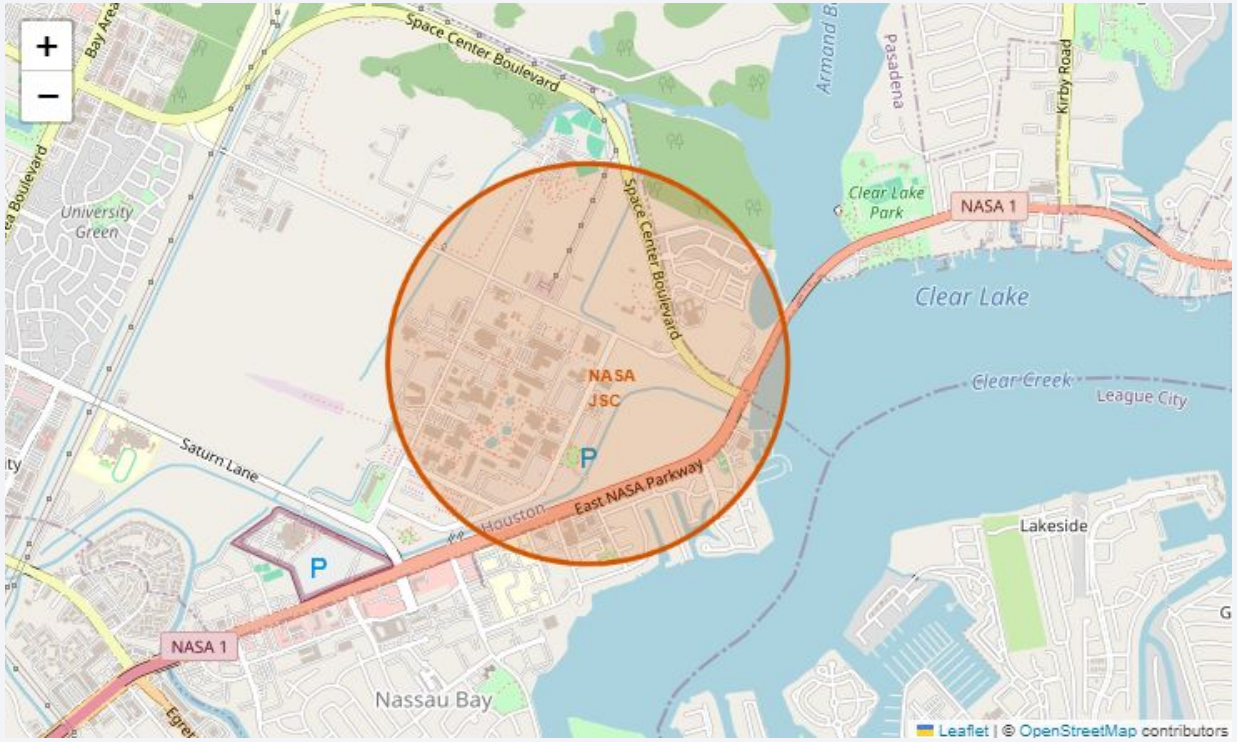


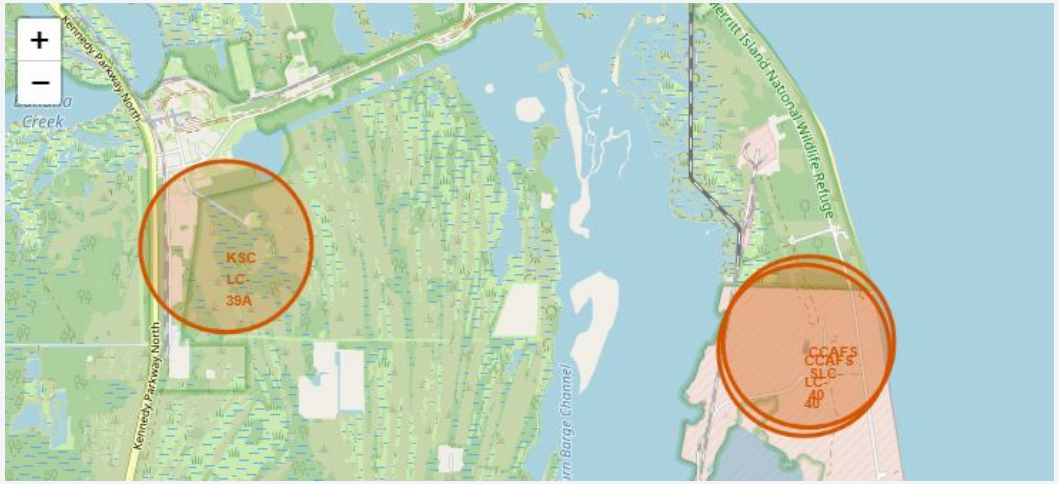
GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%204/MACHINE%20LEARNING%20PREDICTION/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

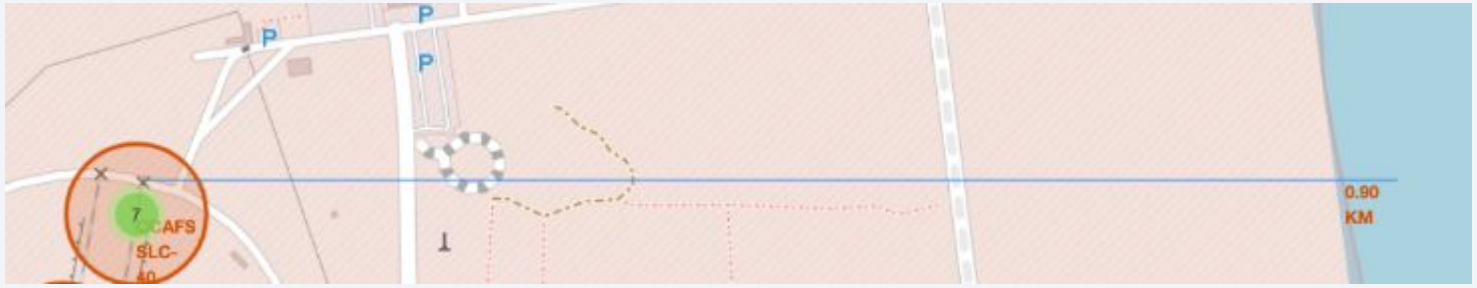
Results [1/2]

- Exploratory data analysis results:
  - There was a high level of successful mission outcomes.
  - The more launches SpaceX made, the more correct landings they achieved (first one achieved on 2015).
  - The total amount of correct landings should increase in the future as well as it was increasing before with all the launches made in the past.
  - Based on the information, two booster versions had problems when landing in drone ships in 2015.

Results [2/2]

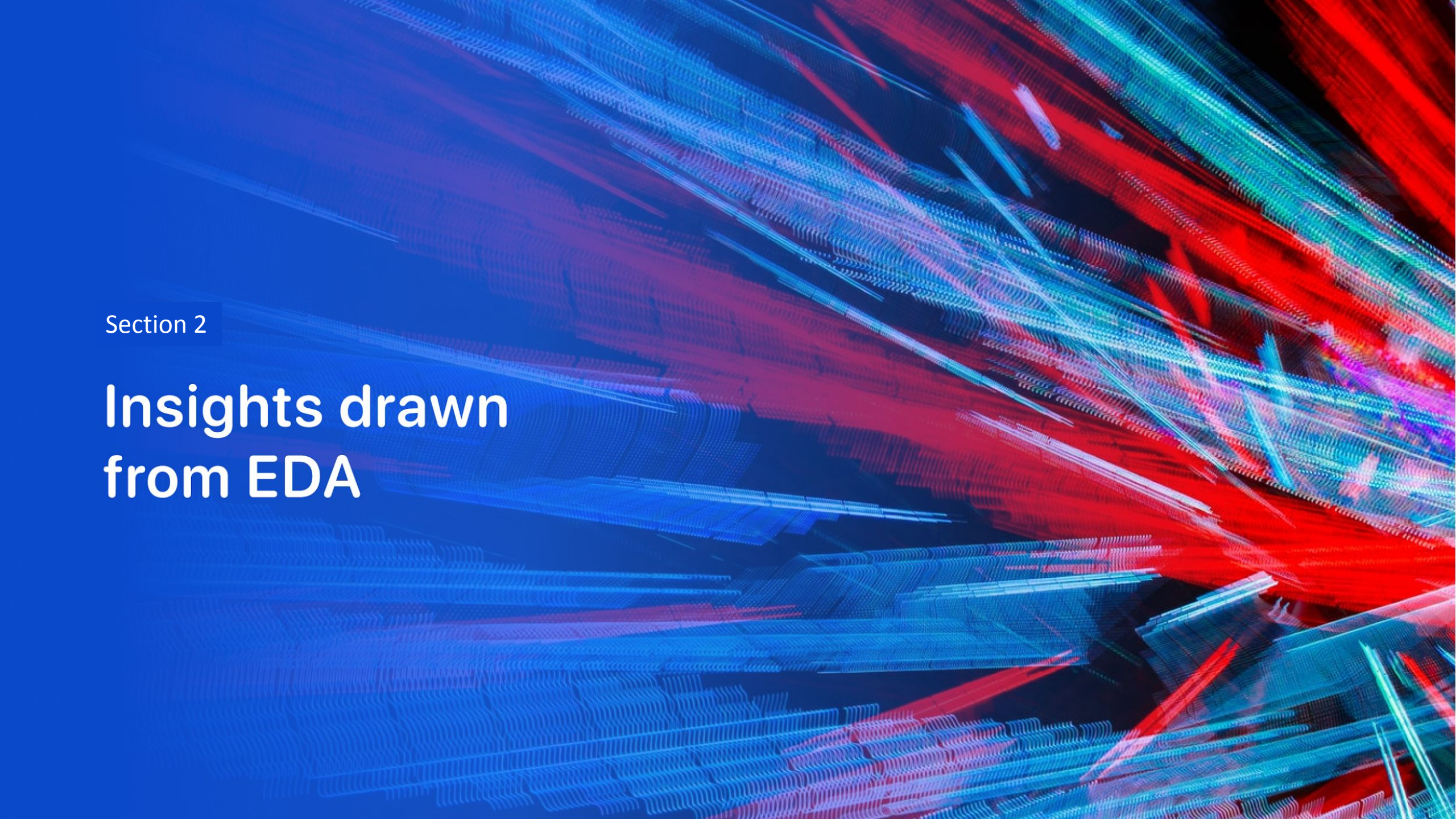

Space Center Boulevard
Armand Bayou
Pasadena
Kirby Road
University Green
Clear Lake Park
NASA 1
Clear Lake
NASA JSC
Clear Creek
League City
Saturn Lane
East NASA Parkway
Houston
Lakeside
NASA 1
Nassau Bay
Leaflet | © OpenStreetMap contributors


Kennedy Parkway North
Merritt Island National Wildlife Refuge
KSC LC-39A
CCAFS SLC-40


CCAFS SLC-40
0.90 KM

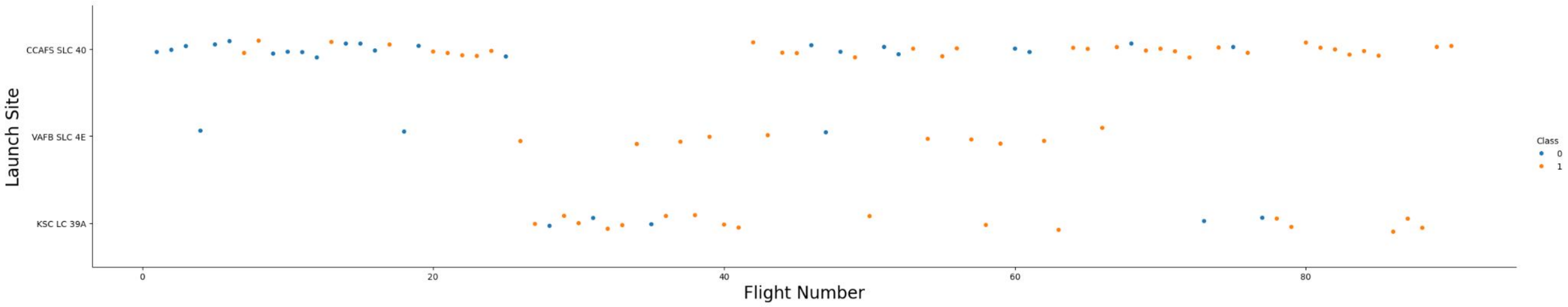
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

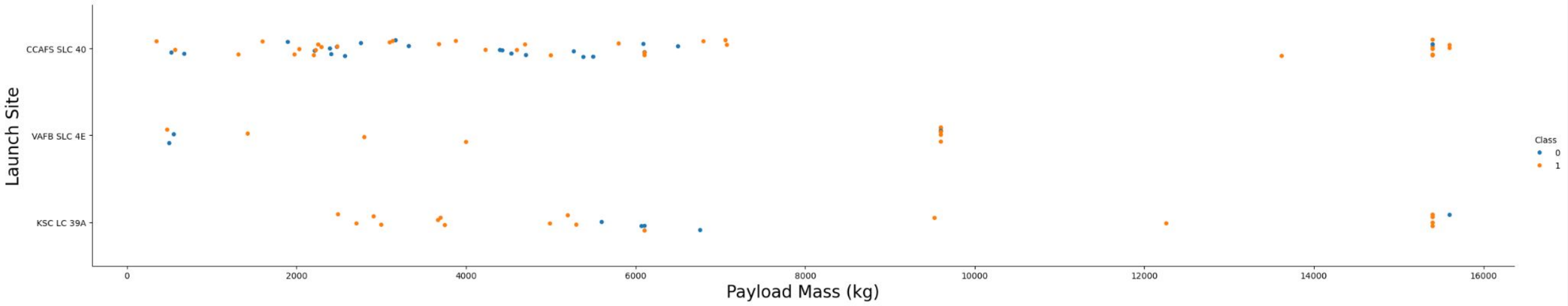
- The following representation between the flight number and the launch sites represents which kind of launch site is (probably) the best place for the company to start missions, where CCAFS SLC 40 ranks at the top with the most flights accomplished, whereas KSC LC 39A is the one with the lowest amount.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

Payload vs. Launch Site

- The following representation between the payload mass (kg) and the launch sites represents which kind of payload mass (kg) would be the most ideal for successful launches. We can see that CCAFS SLC 40 supports payloads greater than 14.000kg, while VAFB SLC 4E could not go over 10.000kg.

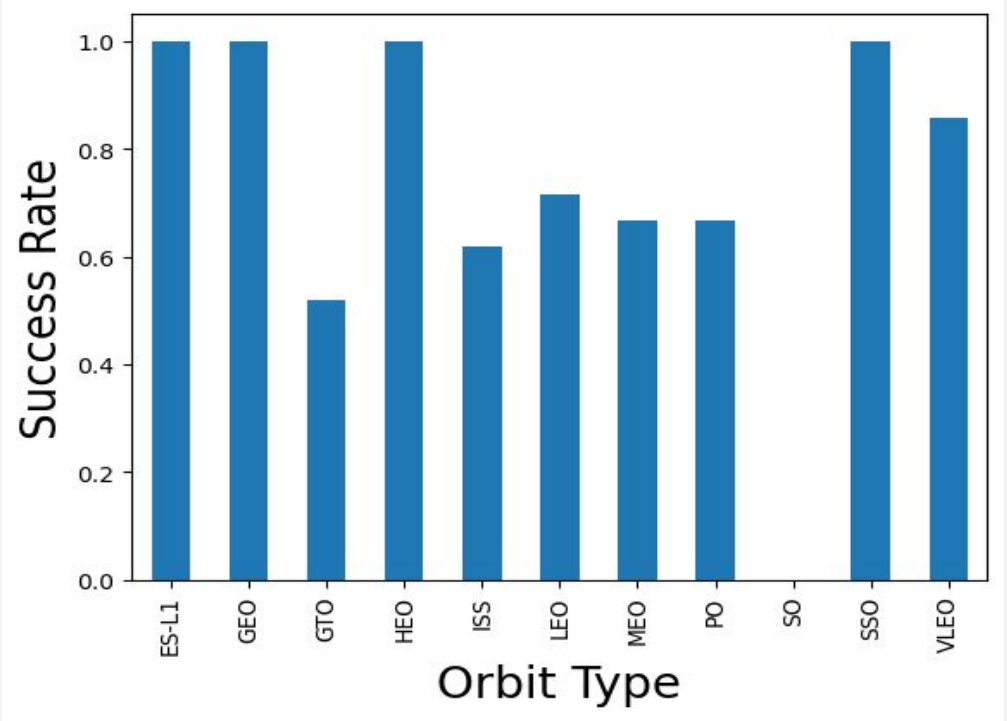


GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

Success Rate vs. Orbit Type

- We can see from the bar graph that the orbit types with more success rate are four: ES-L1,HEO, GEO, and SSO.

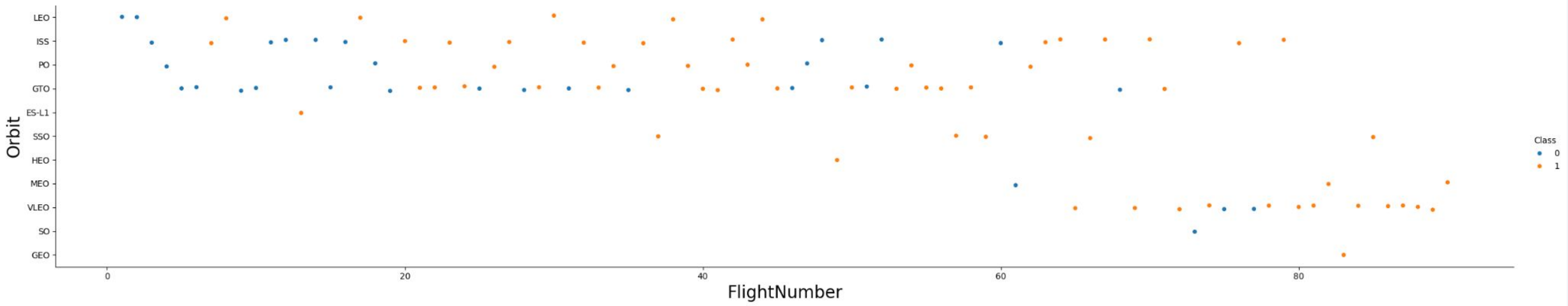
- The orbit type with less success rate is GTO, with less than 50% of success rate.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

Flight Number vs. Orbit Type

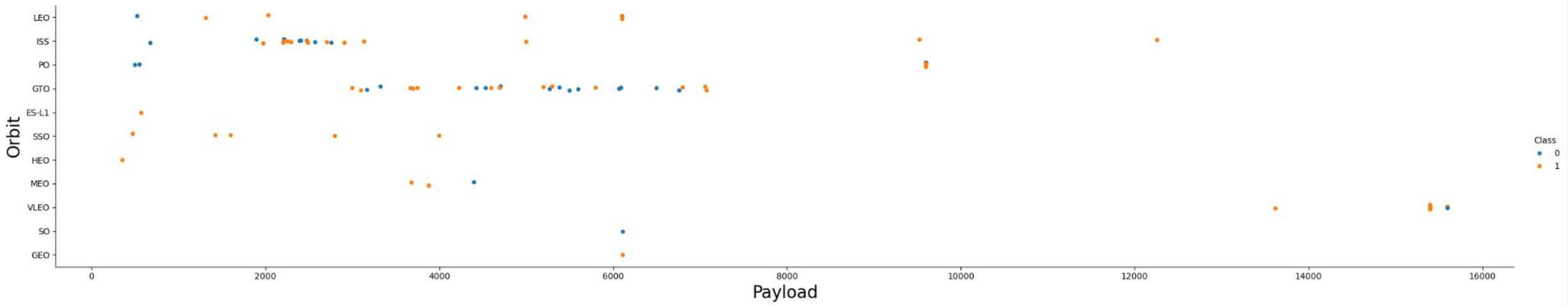
- The following representation between the flight number and the orbit shows that the more flight numbers there were, the more improvements each orbit received over time, increasing the success rate.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

Payload vs. Orbit Type

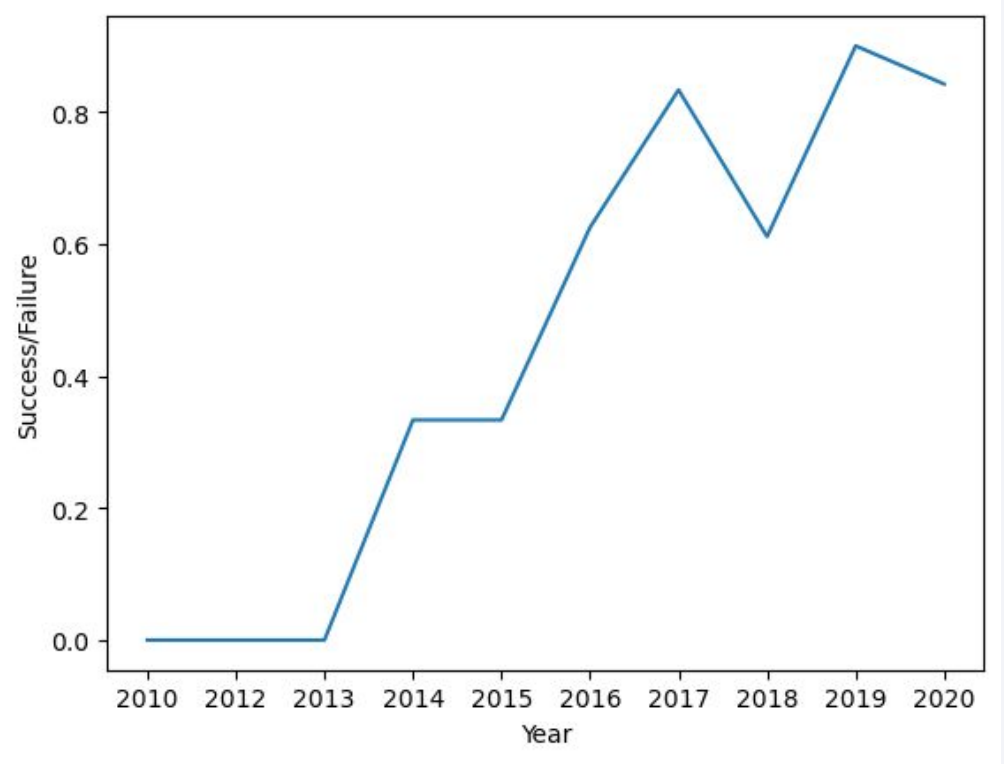
- The following representation between the payload mass (kg) and the orbit shows that ISS can handle a good amount of payload, knowing that there were successful attempts with different amounts of kg used.



GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

Launch Success Yearly Trend

- The success rate increased over time, reaching a maximum value (more than 0.8) on 2019. At the beginning it became a failure after another until reaching a point where the success rate started increasing (successful missions), after 2013.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20FOR%20DATA%20VISUALIZATION/edadataviz.ipynb

All Launch Site Names

- The names of the launch sites are four in total, where each one of them was obtained from the dataset with a SELECT DISTINCT query over the table containing the dataset information.

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

GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Launch Site Names Begin with 'CCA'

- There was a total of five records in the dataset where the launch site names begin with *“CCA”*. The query contains all the information (columns) related to those five launch sites.

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

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

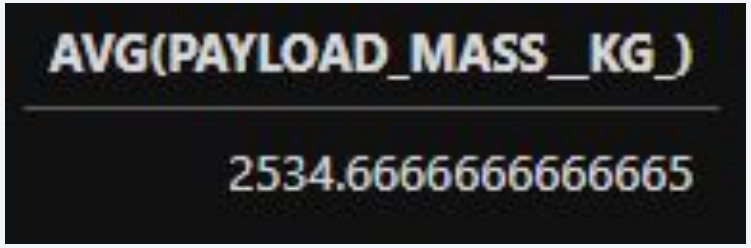
Total Payload Mass

- The total payload mass (KG) that the boosters from NASA can carry is calculated by summing all the payloads where the customer is NASA (CRS).

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Average Payload Mass by F9 v1.1

- The average (avg) payload mass (KG) that the boosters version “F9 v1.1” can carry is calculated by applying the average function from the payload mass (kg) with the boosters with that specific version.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

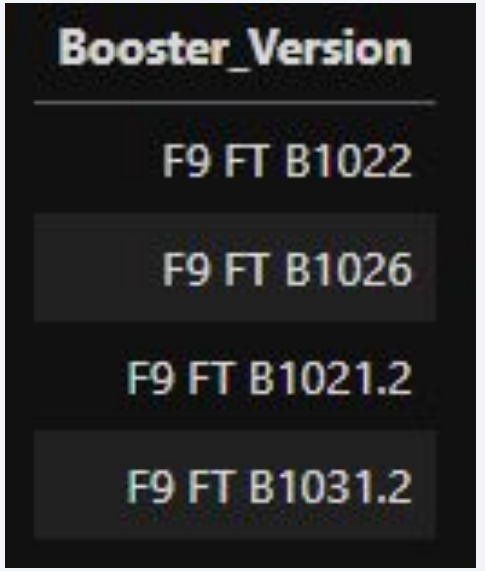
First Successful Ground Landing Date

- The first successful ground landing date is calculated by filtering all the landings that were a success but with ground pad, to later select the minimum date from that result.

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters that have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is calculated by filtering the boosters with successful landing on drone ship and if they had a payload between that range.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Number of Successful and Failure Mission Outcomes

- For the total number of successful/failure mission outcomes I had to group it by its kind and apply the count function to obtain the total amount of each.

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Boosters Carried Maximum Payload

- For the boosters that carried the max. amount of payload mass (kg) I had to first check what was the maximum amount of payload, to later check which booster was able to carry that amount.

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

Booster_Version
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

2015 Launch Records

- The only two cases where there were failed landings happened in 2015 with boosters *“F9 v1.1 B1012”* and *“F9 v1.1 B1015”*, from “”, both from the “CCAFS LC-40” launch site.

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

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

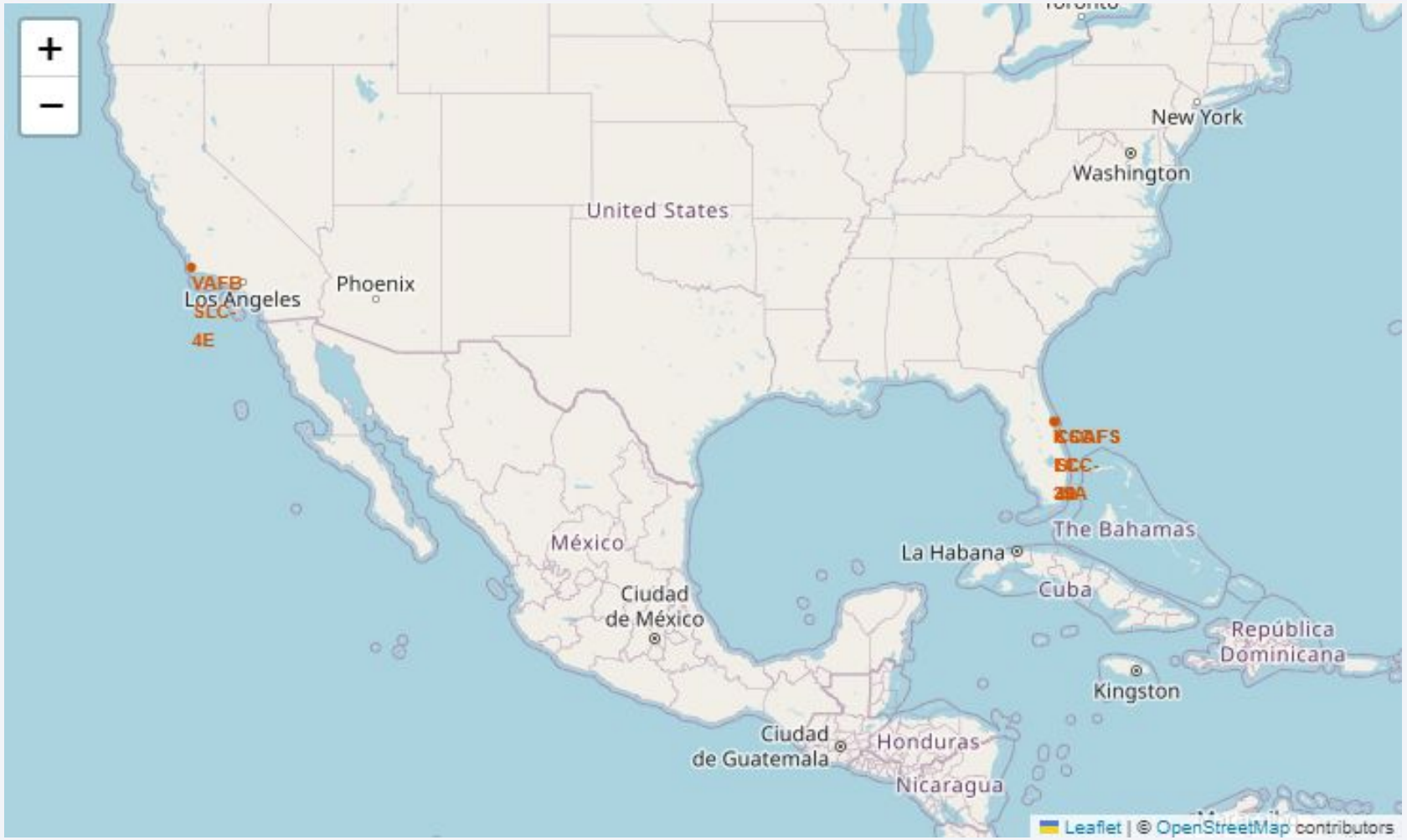
- The results from all landing outcomes between those two dates are shown in the representation, as well as “No attempt” for the ones that do not classify for the other categories.

Landing_Outcome	COUNT(Landing_Outcome)
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

GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%202/EDA%20WITH%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

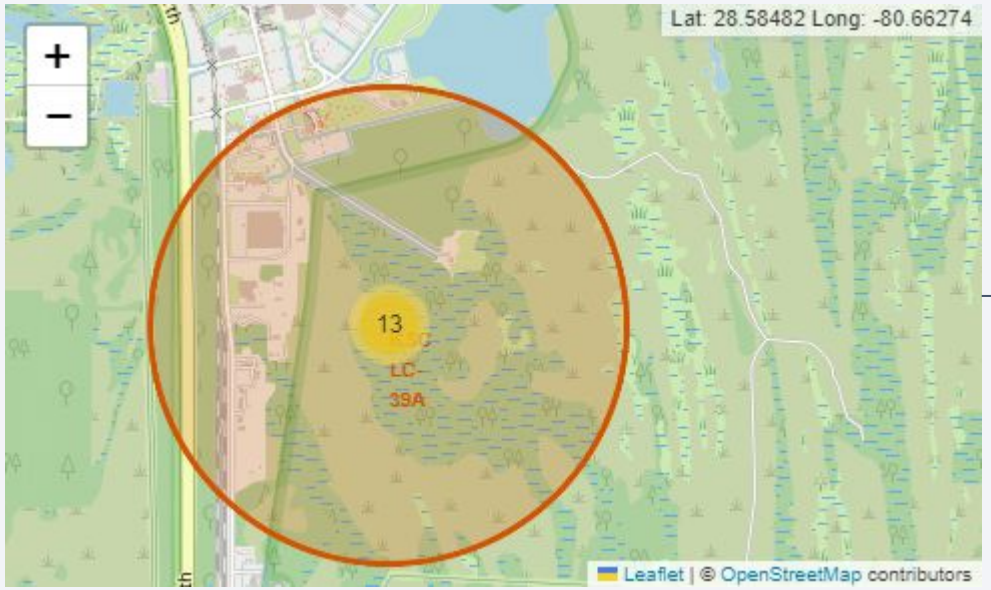

Section 3
Launch Sites
Proximities Analysis

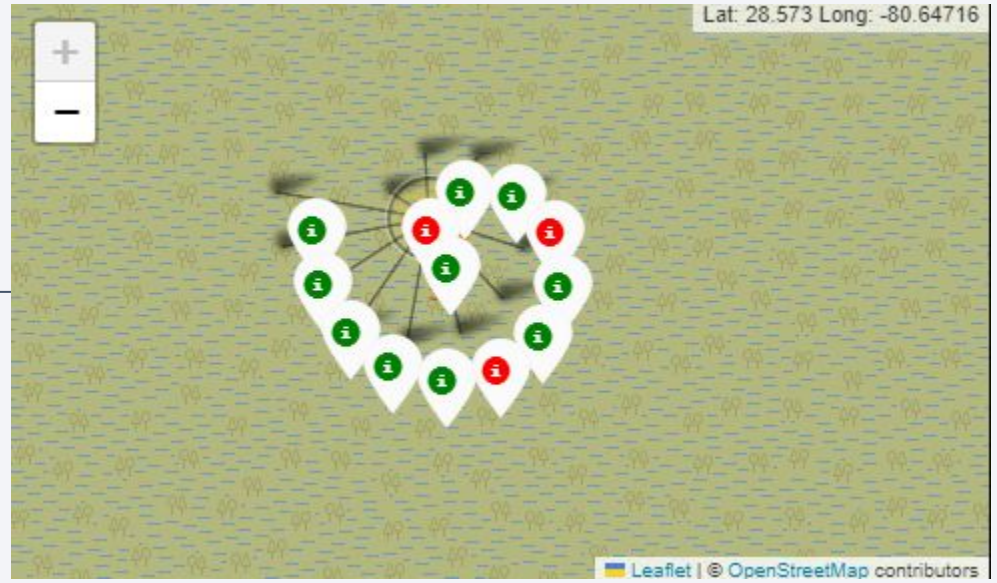
Launch sites from general view



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/VISUAL%20ANALYTICS%20WITH%20FOLIUM/lab_jupyter_launch_site_location.ipynb

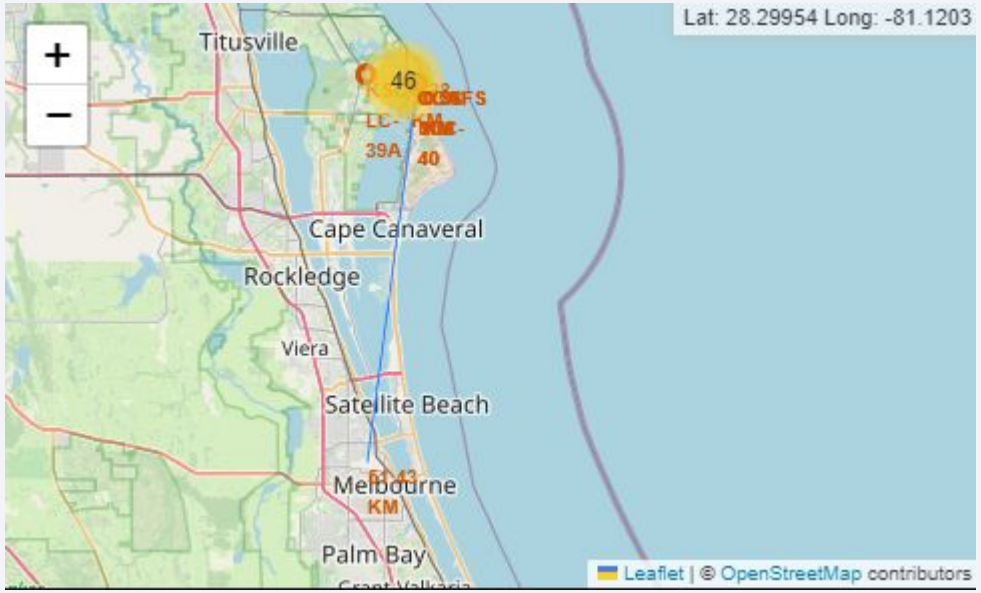
Markers based on launch outcomes

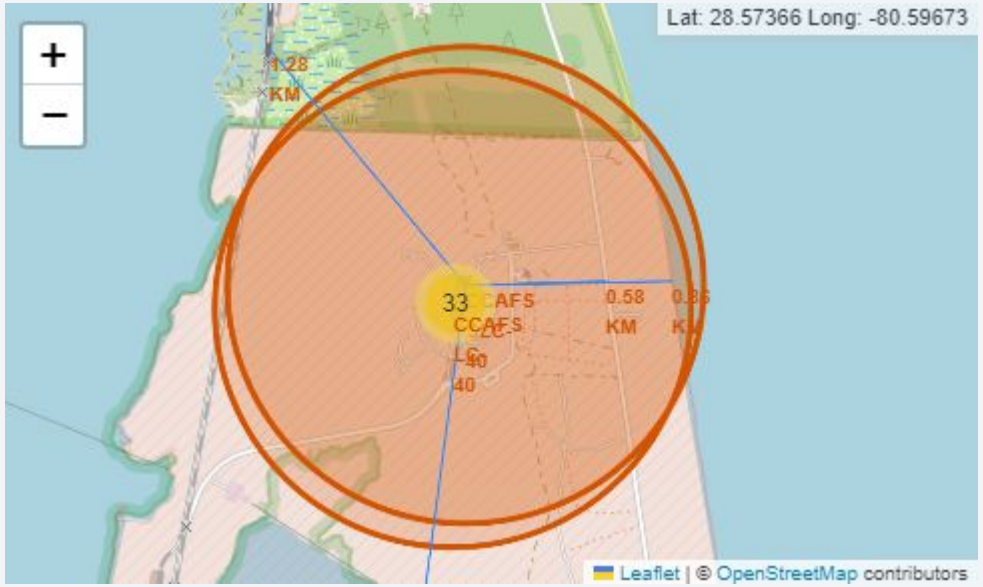




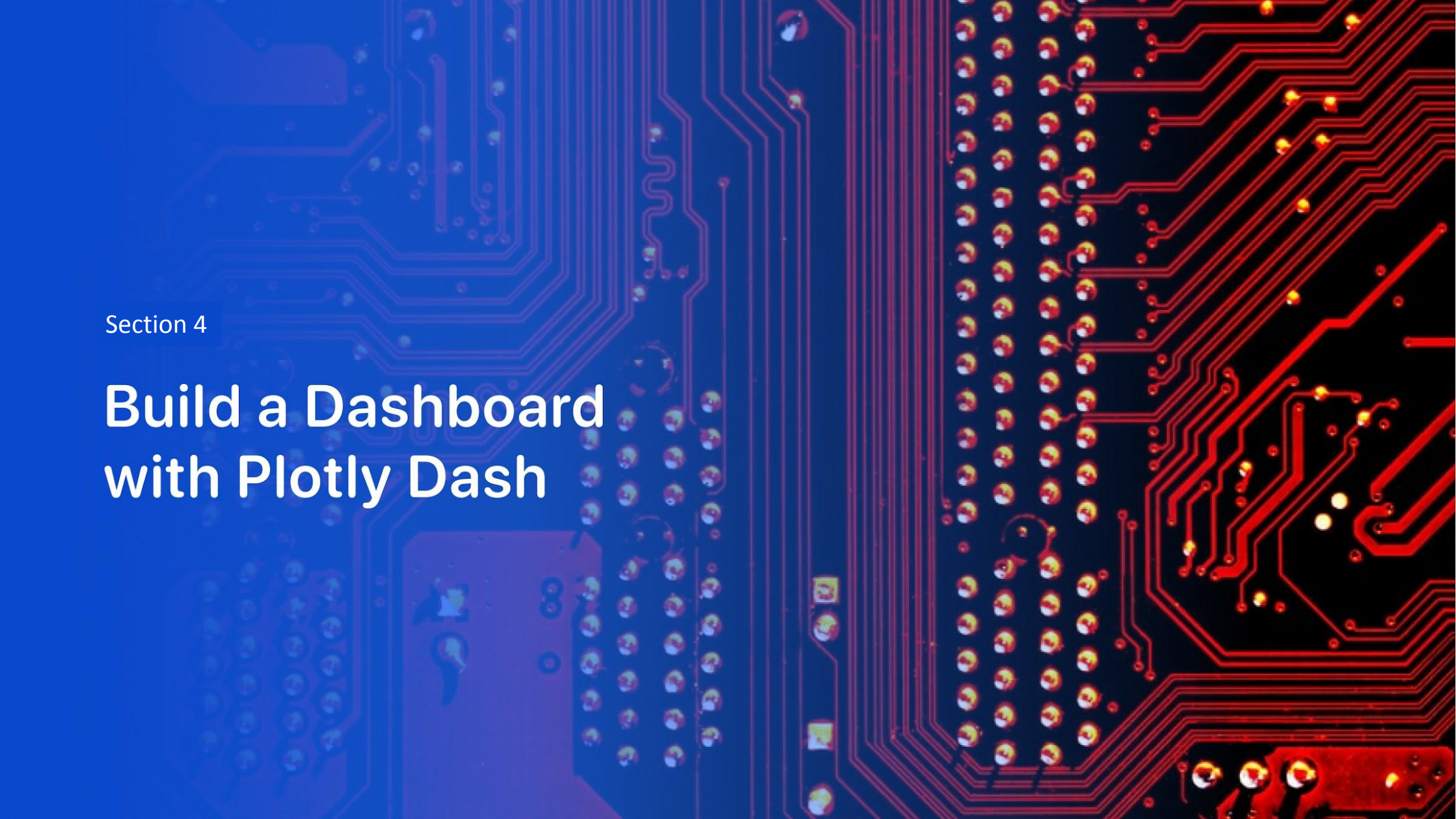
GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/VISUAL%20ANALYTICS%20WITH%20FOLIUM/lab_jupyter_launch_site_location.ipynb

Distances and successful landings between range



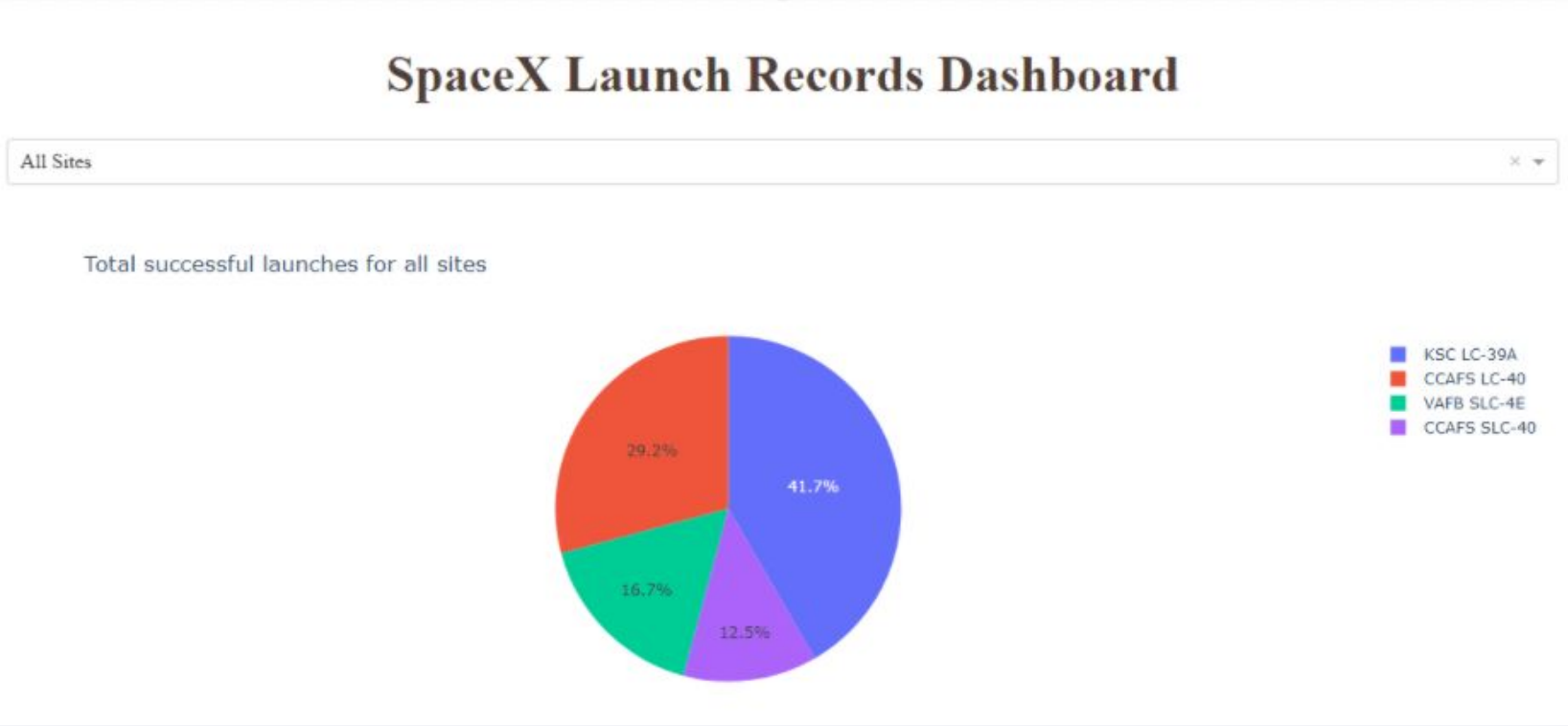


GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/VISUAL%20ANALYTICS%20WITH%20FOLIUM/lab_jupyter_launch_site_location.ipynb


Section 4
Build a Dashboard with Plotly Dash

Successful launches by site

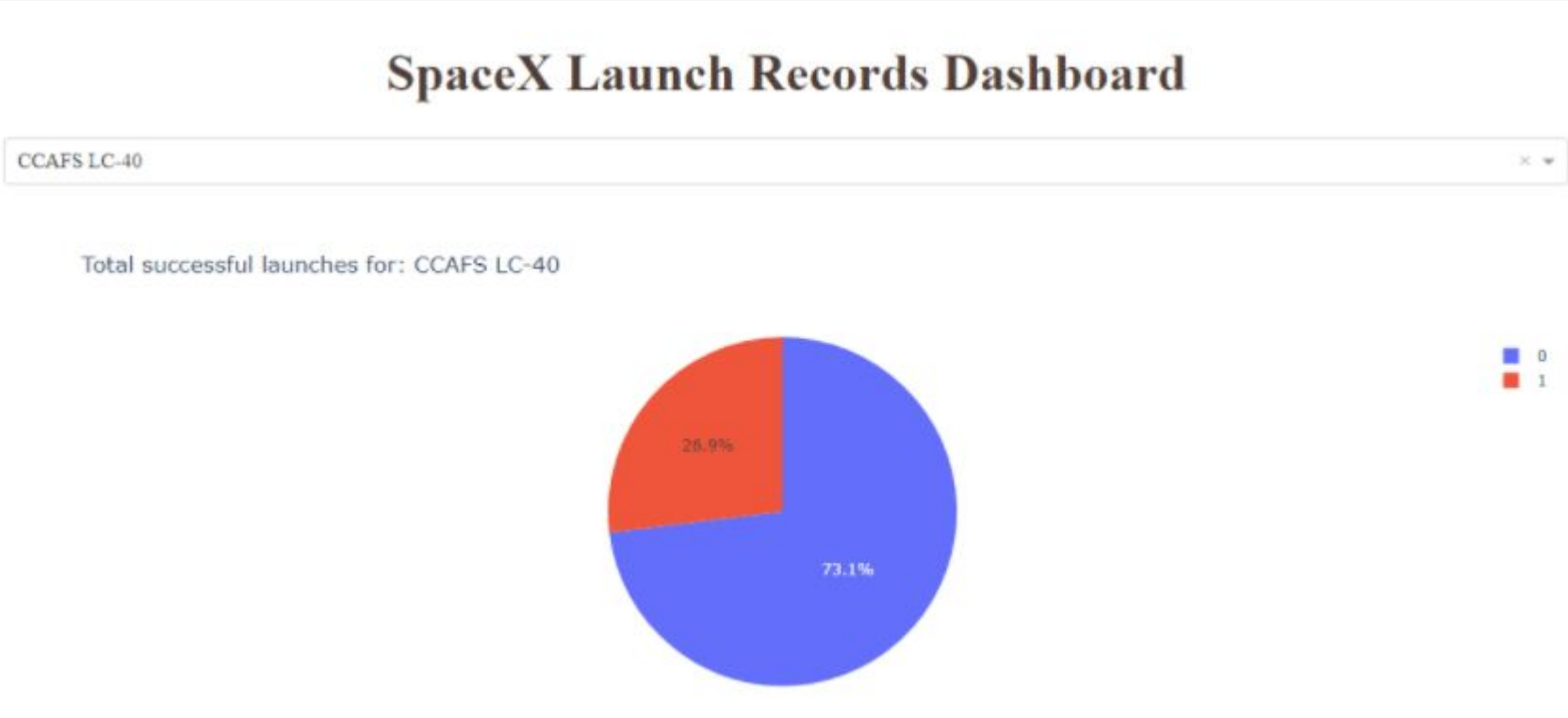
- All launches shown in percentages (where KSC LC-39A takes the lead).



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/INTERACTIVE%20DASHBOARD%20WITH%20PLOTY/spacex_dash_app.py

Successful launches by CCAFS LC-40

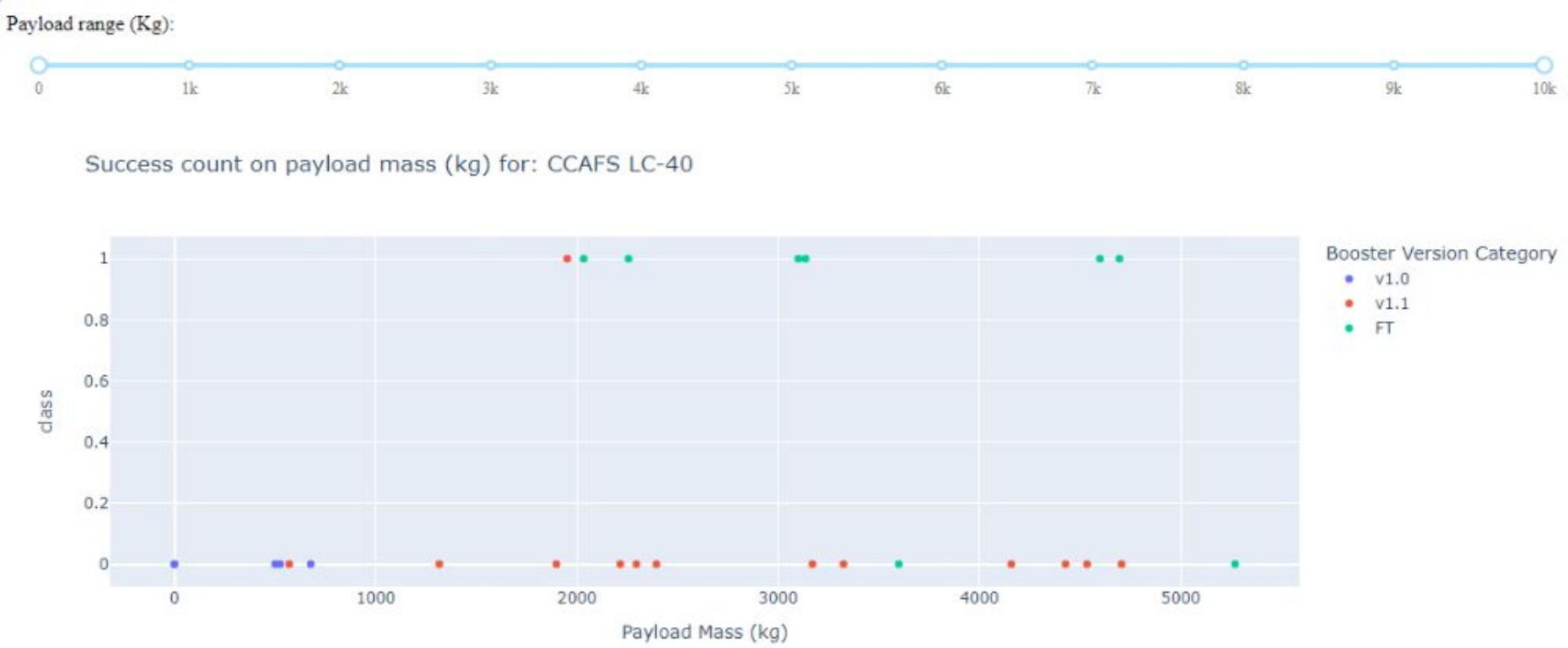
- The successful launches rate is 73.1% against 26.9% of failed ones.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/INTERACTIVE%20DASHBOARD%20WITH%20PLOTY/spacex_dash_app.py

Payload mass (kg) and Launch outcome by CCAFS LC-40

- The payloads that are the most successful for this site are between 2000 and 5000, with specifics that should be avoided (such as ~4500kg).



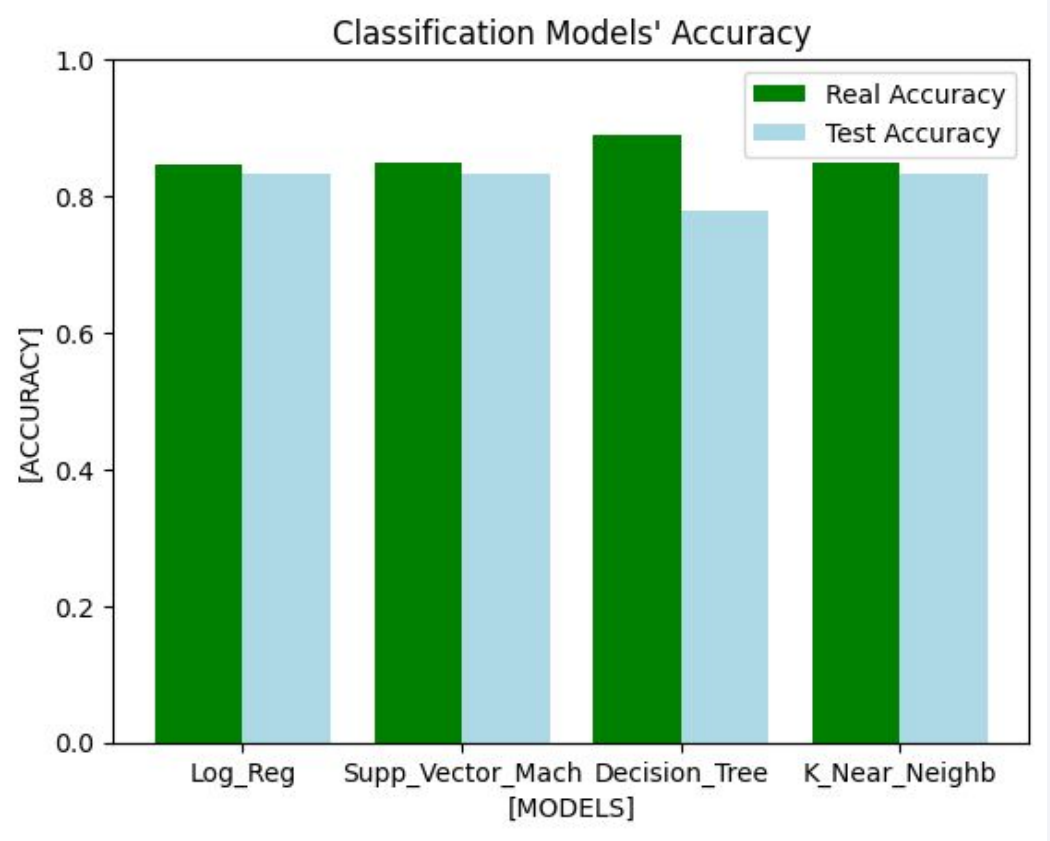
GitHub: https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%203/INTERACTIVE%20DASHBOARD%20WITH%20PLOTY/spacex_dash_app.py

Section 5

Predictive Analysis (Classification)

Classification Accuracy

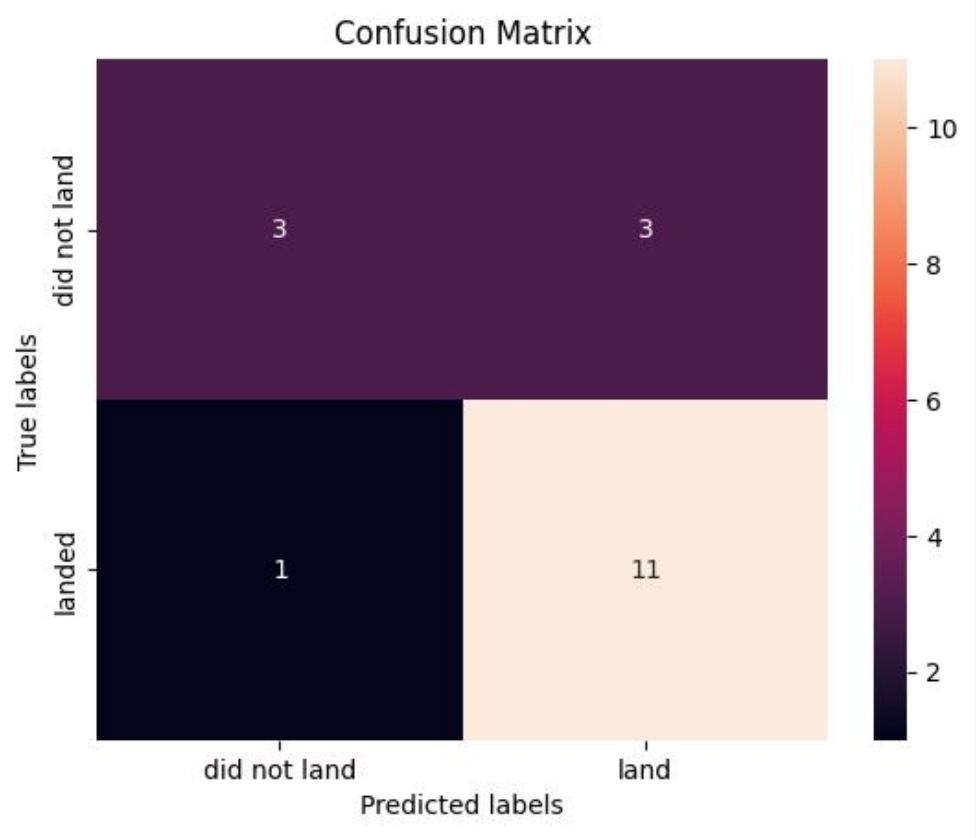
- The four classification models were tested and each one of them represents a accuracy from 0 to 1, where the most variation was from the Decision Tree model, even though it had the most accuracy throughout the tests.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%204/MACHINE%20LEARNING%20PREDICTION/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Confusion Matrix

- The confusion matrix of the Decision Tree model proves its accuracy based on the results achieved throughout the tests.



GitHub:
https://github.com/GOZEBRAHEAD/Applied-Data-Science-Specialization-Capstone/blob/main/WEEK%204/MACHINE%20LEARNING%20PREDICTION/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Conclusions

- Launches that have an (approximately) payload mass (kg) smaller than 7500 kg are less risky in terms of successful and failure.
- The company SpaceX achieved great results over time knowing that when they started they had a three year without successful landing outcomes, meaning that at least until that time, everything was not profit.
- The Decision Tree model for classifications can be used to predict possible outcomes based on the results achieved for the testing part.
- It would be good to have an updated report over the historical information from SpaceX, knowing that we used limited data (for example, website from wikipedia had a date interval).

Appendix

- Anything can be seen on the GitHub repository, following the respective folder structure for each week where the assignments and exercises will be located.

Thank you!

Luciano Nieves
March 31, 2024