Winning Space Race with Data Science

Emirhan Ülgen
21 November 2024

Outline

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

Executive Summary

- Several operations such as web scraping, data collection, data visualization and machine learning were applied to a dataset.
- Data was collected, preprocessed, some relationships between the features of the dataset were explored, and the locations were spotted by maps

Introduction

- A lot of data science techniques was taught and practiced so far. The goal of this project is to apply those methods to the dataset which has significant information about SpaceX.
- There are some features in the dataset such as launch sites, rocket names, launch and landing outcomes as well as their mass weights. Various techniques were used to find relationship among those features.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected using some features' APIs from raw dataset. This process could also be called data engineering or feature engineering.
- The values of a categorical feature in the dataset were converted into numerical values using encoding so that each category was assigned to a number between 0 and 1. Those values were collected and put to an additional column, thus formed a new feature.
- Some key values or points were selected, grouped, and sorted out using SQL queries while some visuals such as scatter plot, bar chart, and line graph were created with data visualization.
- Data was visualized in pie chart and scatter plot, with range slider, with Plotly to see the other relationships and maps were used to find the locations and distances with the help of Folium
- The feature and target variables were defined, the data was split into training and test at 80/20 rate, different parameters were specified with hyperparameter tuning for the model to find the best ones, the models were fitted to the train and test sets, and they were evaluated calculating the accuracy score.

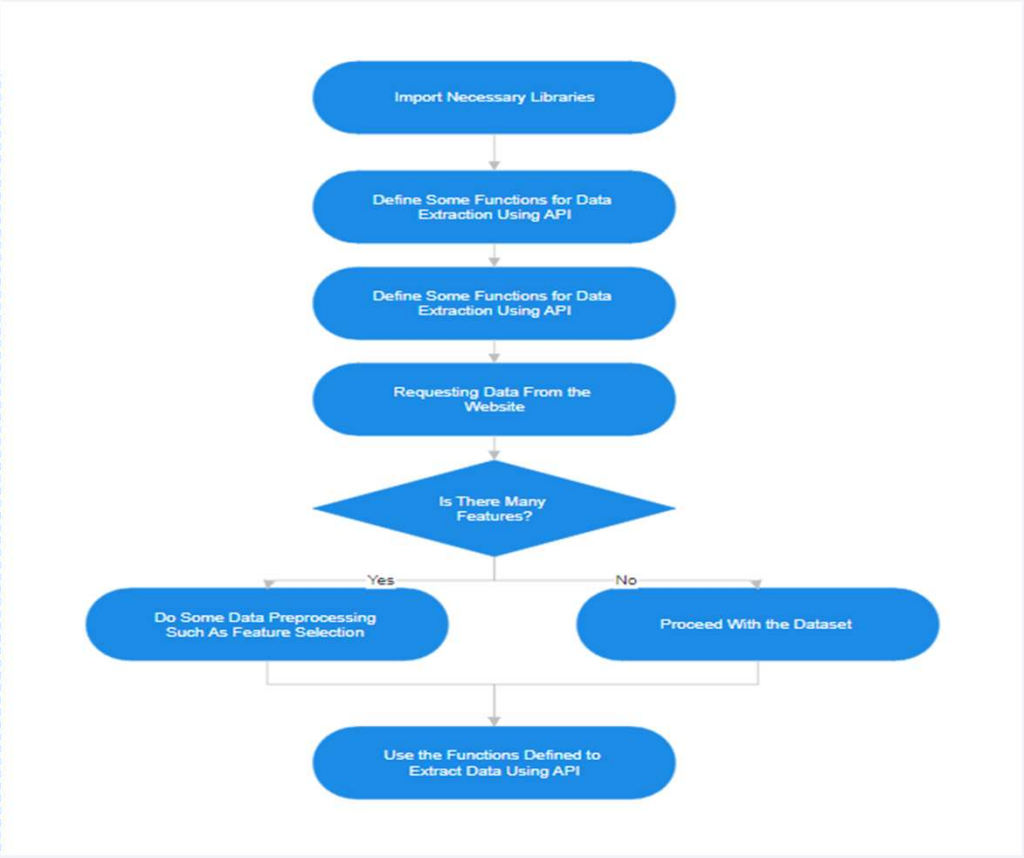
Data Collection

- The datasets were collected using api requests. The pure data was firstly collected with IDs. Some features were taken from it and the data for respective features were extracted using APIs.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

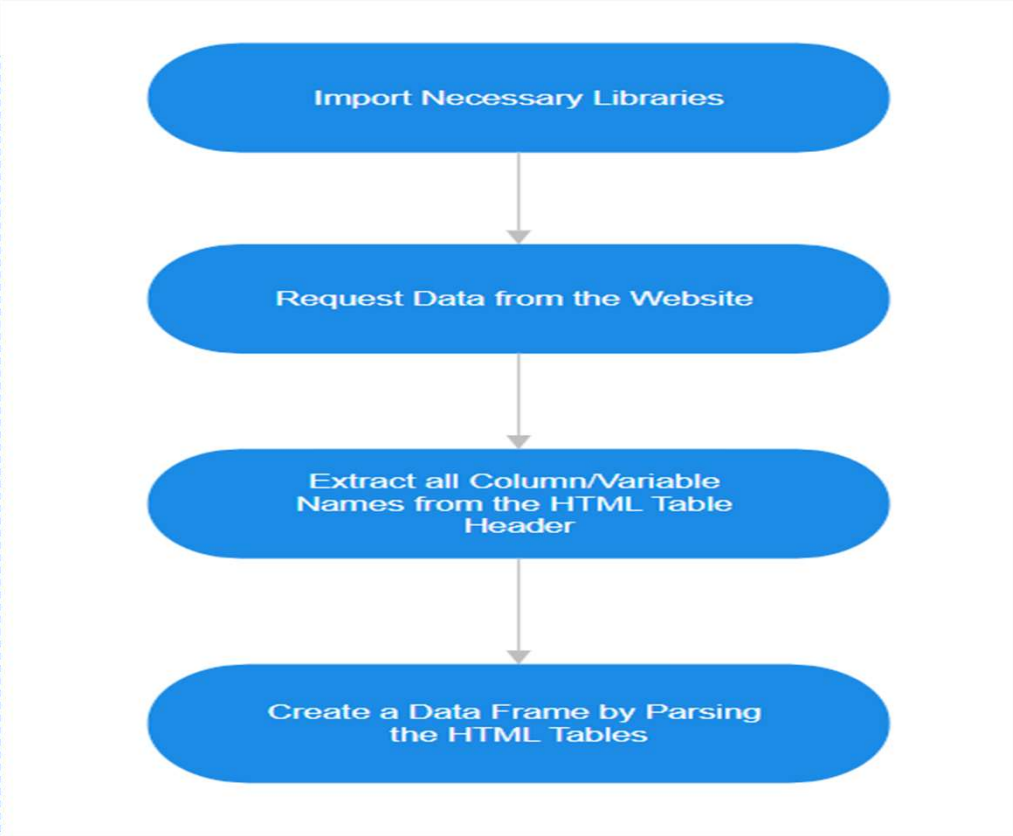
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- The raw dataset was imported, the missing values were calculated, and values in specific features were counted. The values of one categorical feature were counted in order to encode them into numerical values and they are added to the data set with a new column. The encoded feature was used to calculate the rate of success.
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots, bar charts, and line chart were mainly used in order to observe the relationship between the features of the dataset. Scatter plots were mostly used to analyze key points on those relationships.
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Some SQL Codes Used:
  - SELECT DISTINCT(COLUMN_NAME) FROM DATASET;
  - SELECT COLUMN_NAME FROM DATASET WHERE LIKE «xx%» LIMIT x;
  - SELECT SUM(COLUMN_NAME) FROM DATASET WHERE (COLUMN_NAME) = «»;
  - SELECT AVG (COLUMN_NAME) FROM DATASET;
  - SELECT MIN (COLUMN_NAME) AS «» FROM DATASET;
  - SELECT COUNT (COLUMN_NAME) FROM DATASET;
  - SELECT SUM(COLUMN_NAME) FROM DATASET WHERE (COLUMN_NAME) = (SELECT MAX(COLUMN_NAME) FROM DATASET); (Subquerying)
  - SELECT SUBSTR(DATE, 0, 5) FROM DATASET; (Specific Dates such as Months, Days, or Years)
  - SELECT COLUMN_NAME FROM DATASET GROUP BY COLUMN_NAME ORDER BY COLUMN_NAME DESC;

- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(2).ipynb

Build an Interactive Map with Folium

- Markers, circles, clusters, and lines were used to spot specific locations on a specific map. Lines are used to find the distances between two points.
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

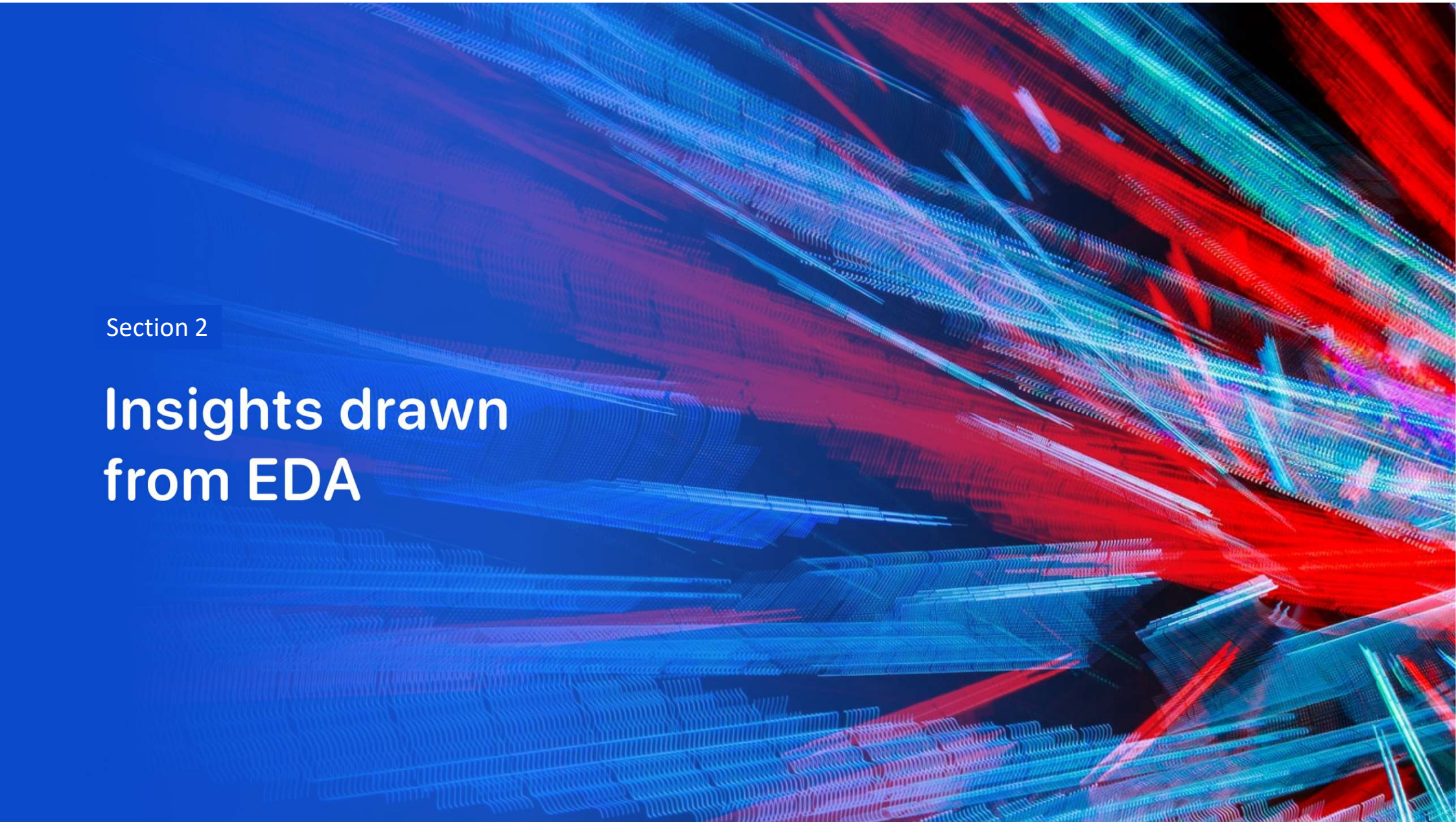
- Pie charts, slide ranger and scatter plots were used in Plotty Dash
- Pie charts were used to show the distribution of variables based on specific value rates while slide ranger was built to specify key points on scatter plot, which showed the relationship between two features, between selected ranges. All of them show relationships both among all features and in one specific feature.
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Three different models were built to predict the target values: Logistic Regression, Decision Tree, and KNN. Support Vector Machine model resulted in timeout failure.
- Logistic Regression and Decision Tree had almost the same prediction performances with high accuracy and same confusion matrix.
- https://github.com/emugre42/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

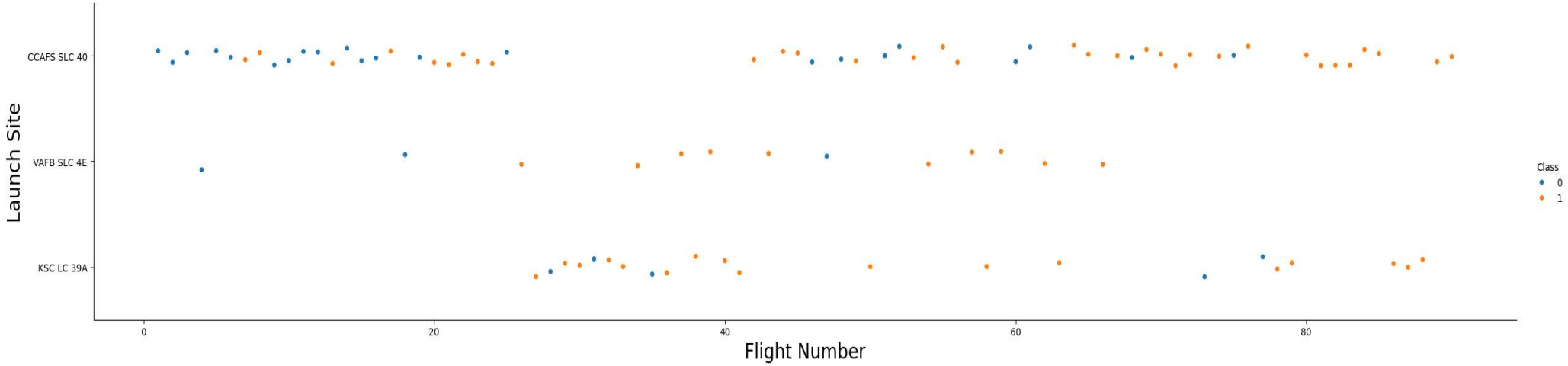
Results

- It was successful to see the specific feature values filtered with SQL and visualize the other features to observe the relationships.
- Using Folium, the locations of SpaceX launch sites were circled and marked, and their distances to specific places such as coasts and roads were drawn.
- The prediction performances of Logistic Regression and Decission Tree were very good while KNN showed a lower performance compared to the other two, but still good enough.

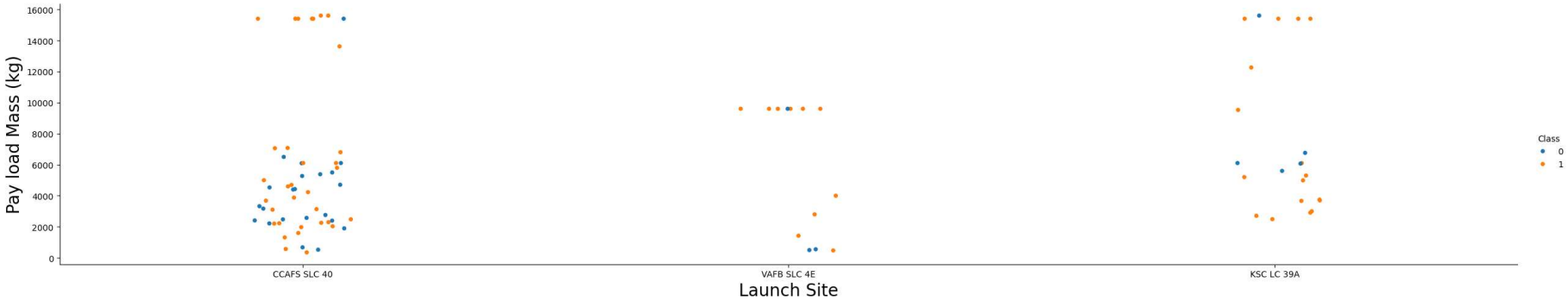

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

- The plot shows that «CCAFS SLC 40» witnessed the launch of first 20 flights and after number 40. The flights after number 25, approximately, were launched in «KSC LC 39A». On the other hand, site «VAFB SLC 4E» has much fewer attempts compared to the other two.



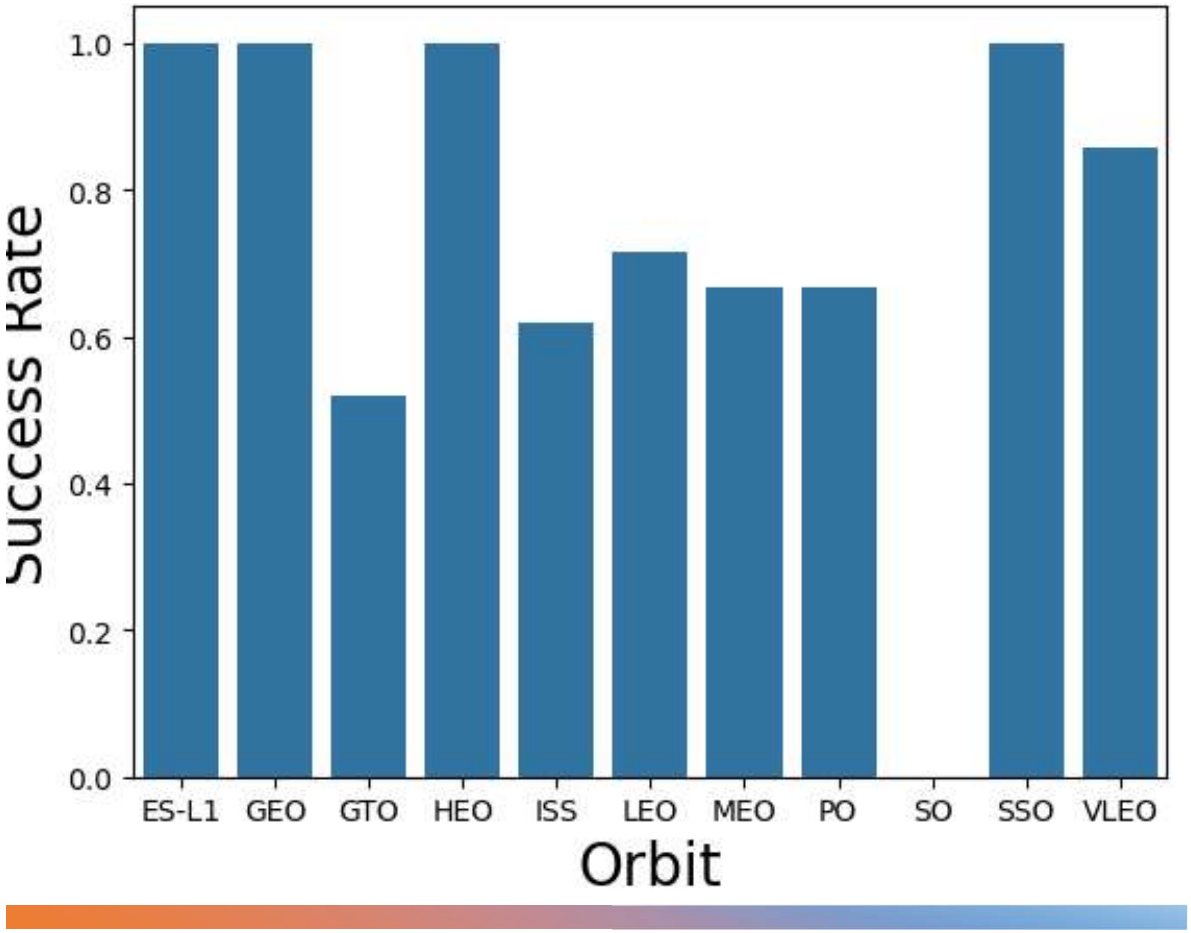
Flight Number vs. Launch Site



Payload vs. Launch Site

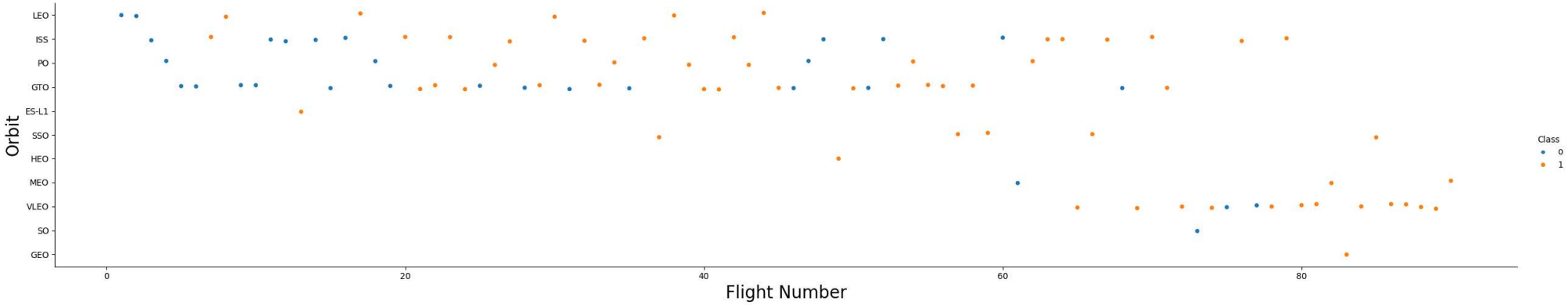
Payload vs. Launch Site

- It can be observed that the «VAFB-SLC» launch site has no rockets launched for heavy payload mass(greater than 10000) while most of the rockets in «CCAFS SLC 40» have weights up to 8000 with some more rockets in a mass around 16000. The situation in «KSC LC 39A» is similar to «CCAFS SLC 40» but the number is much fewer.



Success Rate vs. Orbit Type

- The chart indicates that orbits «ES-L1», «GEO», «SSO», and «HEO» had a completely successful launches while launches in «SO» were totally a failure. The rest except «VLEO», which can be considered successful, have success rates around the average.

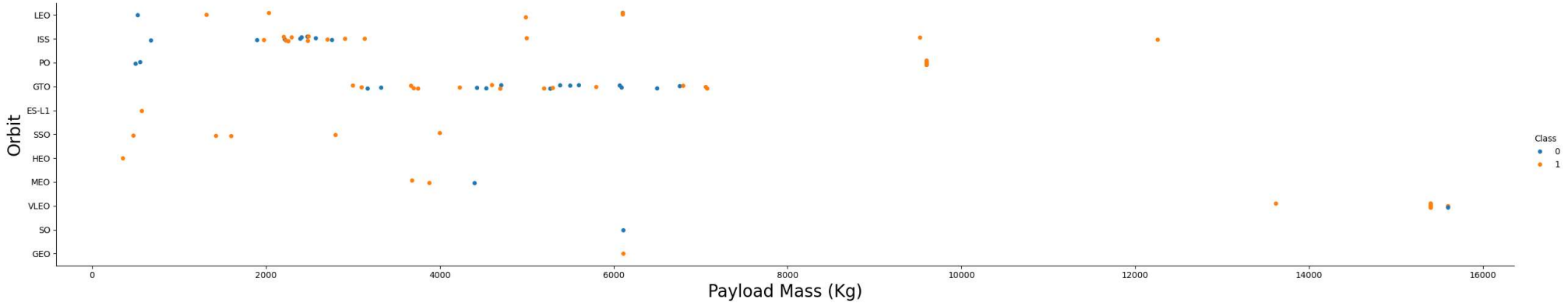


Flight Number vs. Orbit Type

- It can be noted that most of the launches belong to the orbits «LEO», «ISS», «PO», and «GTO» while «VLEO» has a reasonable number of flights and the rest have only few. In the «LEO» orbit, success seems to be related to the number of flights. Conversely, in the «GTO» orbit, there appears to be no relationship between flight number and success.

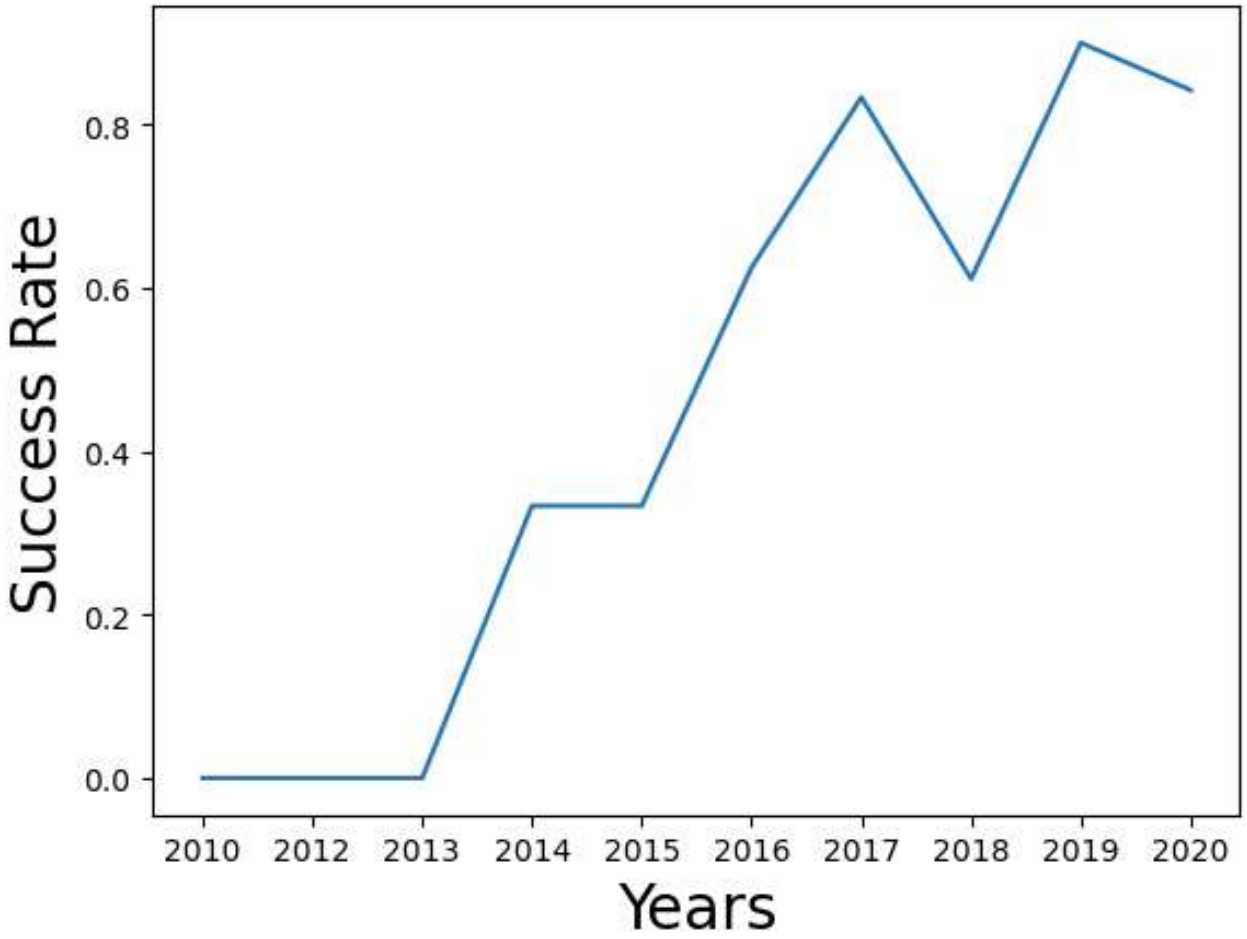
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
  However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- It can be observed that the success rate since 2013 kept increasing till 2020 with a slight decrease between 2017 and 2018.



All Launch Site Names

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

There are 4 different launch sites available in the dataset.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

First Successful Ground Landing Date

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had 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

- Calculate the total number of successful and failure mission outcomes

COUNT(Mission_Outcome)
101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster_Version	Launch_Site	Month
F9 v1.1 B1012	CCAFS LC-40	01
F9 v1.1 B1015	CCAFS LC-40	04

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

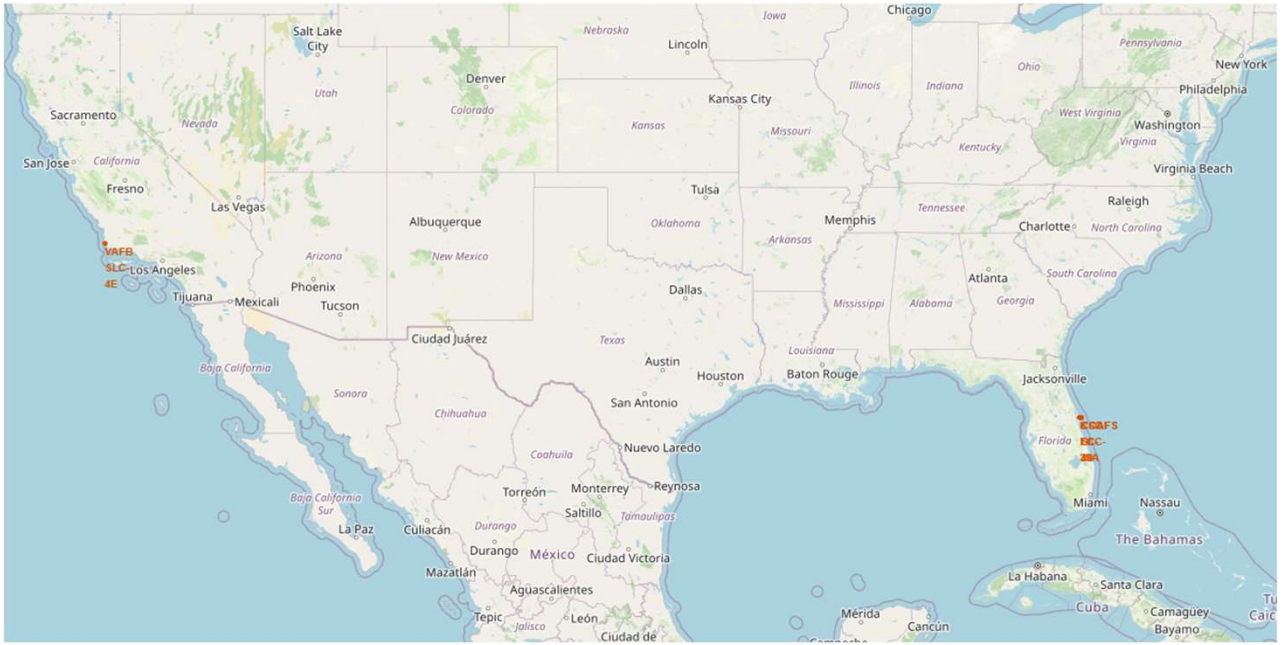
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	COUNT(Landing_Outcome)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)	12
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	12
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)	8
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	5
2014-04-18	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)	4
2013-09-29	16:00:00	F9 v1.1 B1003	VAFB SLC-4E	CASSIOPE	500	Polar LEO	MDA	Success	Uncontrolled (ocean)	2
2015-06-28	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

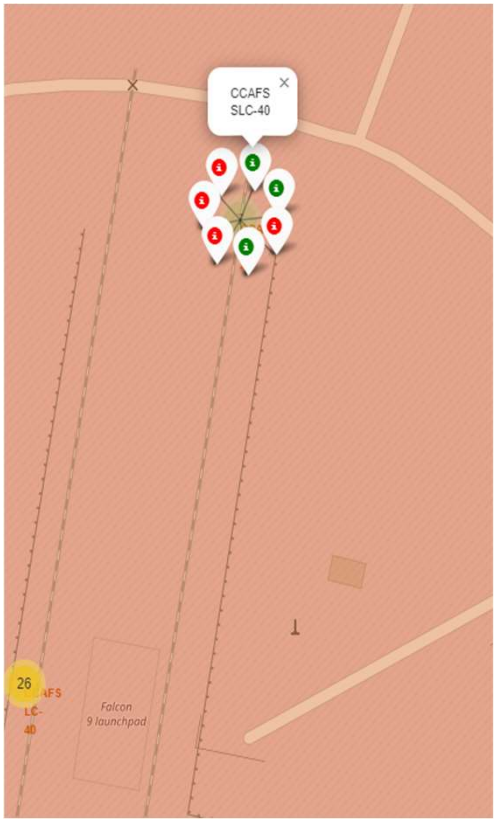
All Launch Sites' Location

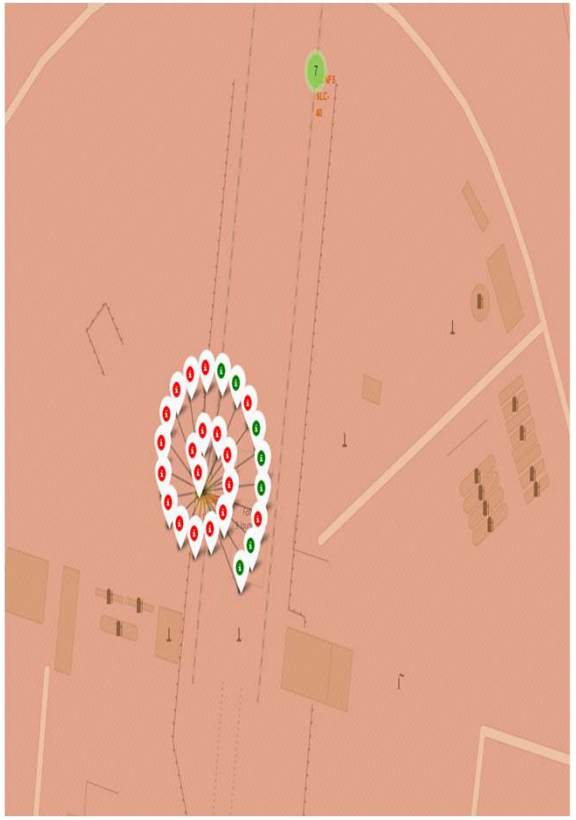
- The launch sites seem to be mostly located in states Florida and California and they are all close to the coastline.



Launch Outcomes on Site CCAFS SLC-40

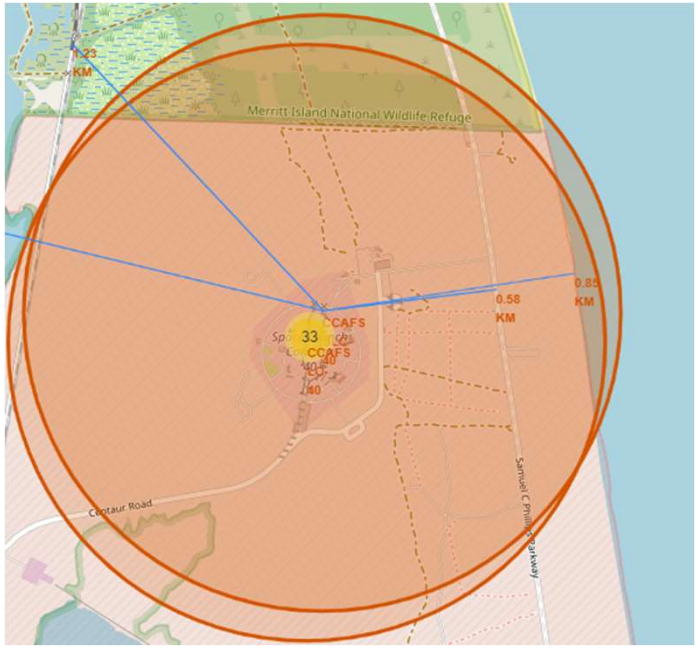
- The maps show that there are 33 launches in CCAFS SLC-40 in near different locations, 7 in one and 26 in another. 3 attempts were successful out of 7 and 7 were successful out of 26 launch attempts.

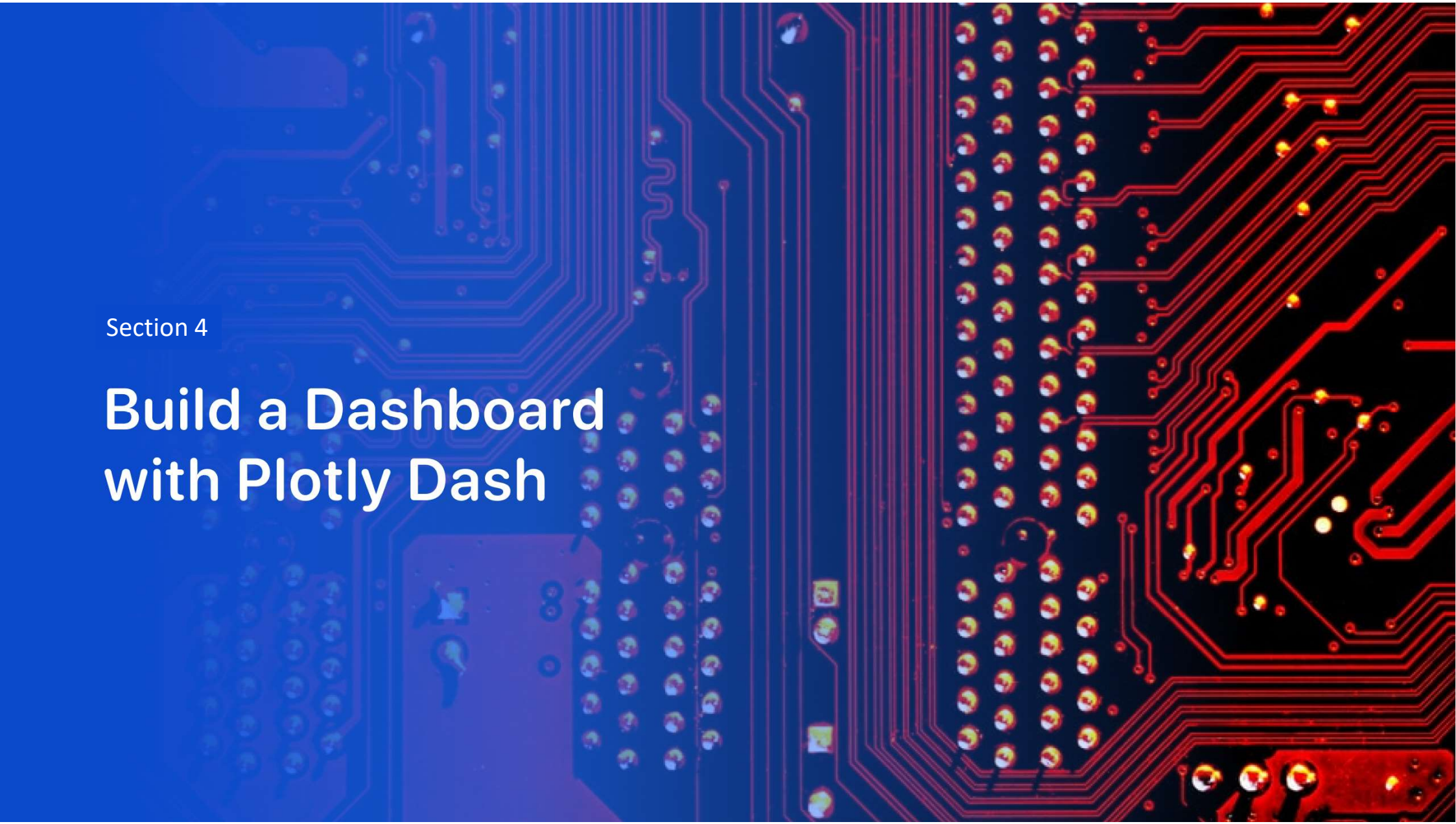




CCAFS SLC-40 Site to Its Proximities

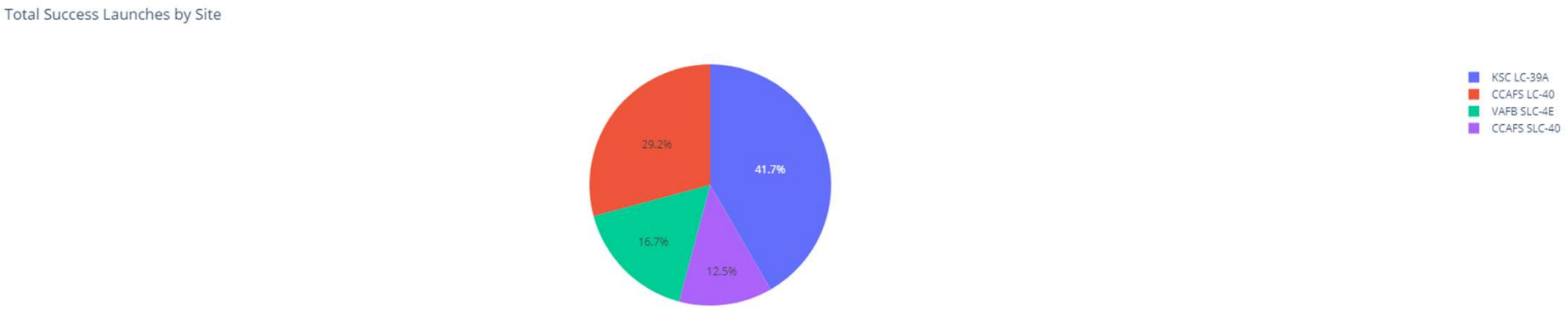
- The distance to the coastline and highway are really close that they are in meters while the distance to the railway is a bit further with more than 1 kilometer. However, the closest city, Titusville in this case, is more than 20 kilometers away from the launch site.



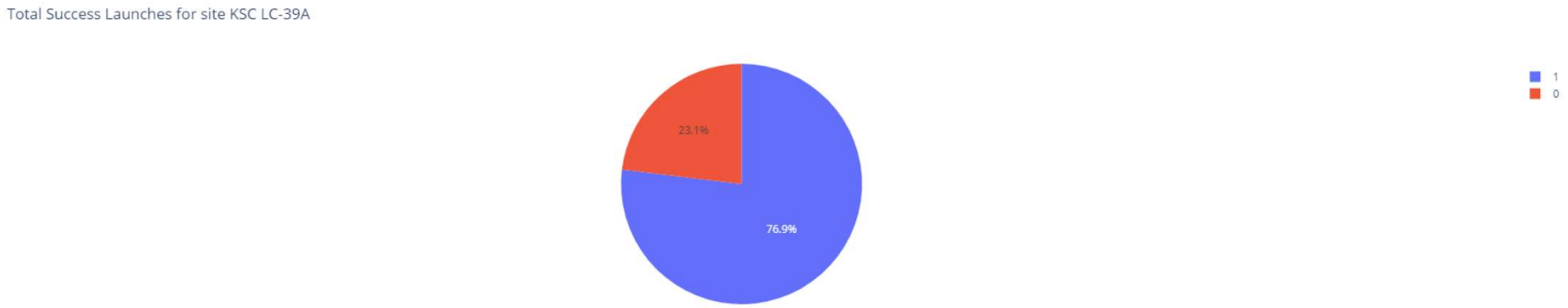

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

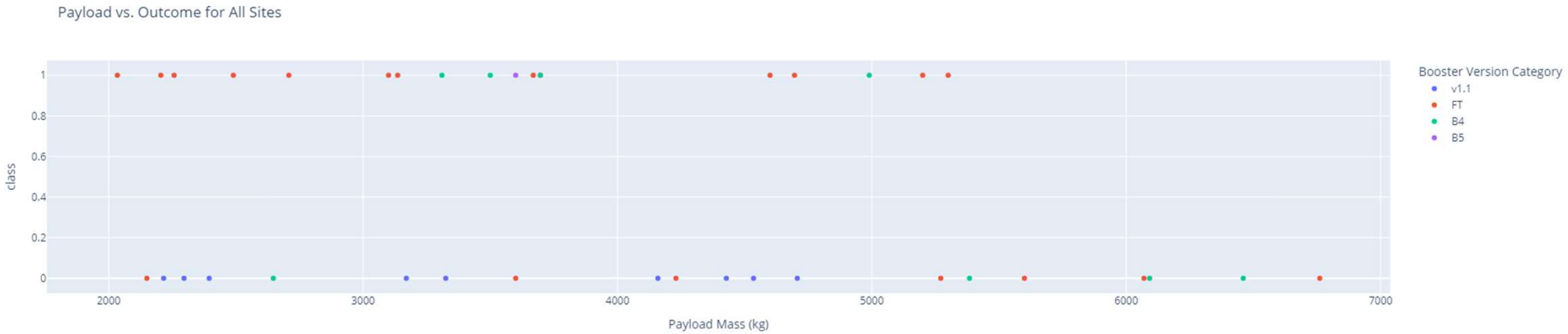
- It can be observed that «KSC LC-39A» had the highest success rate with 41.7% wheras «CCAFS SLC-40» is in the last place with only 12.5% among the launch sites.



Total Success Launches for Site KSC LC-39A



- KSC LC-39A, the launch site with highest launch success ratio, had a rate of 76.9% which more than ¾ of the outcomes and the rate of failure is 23.1%.



Payload vs. Launch Outcome for All Sites in Range Between 2000 and 7000kg

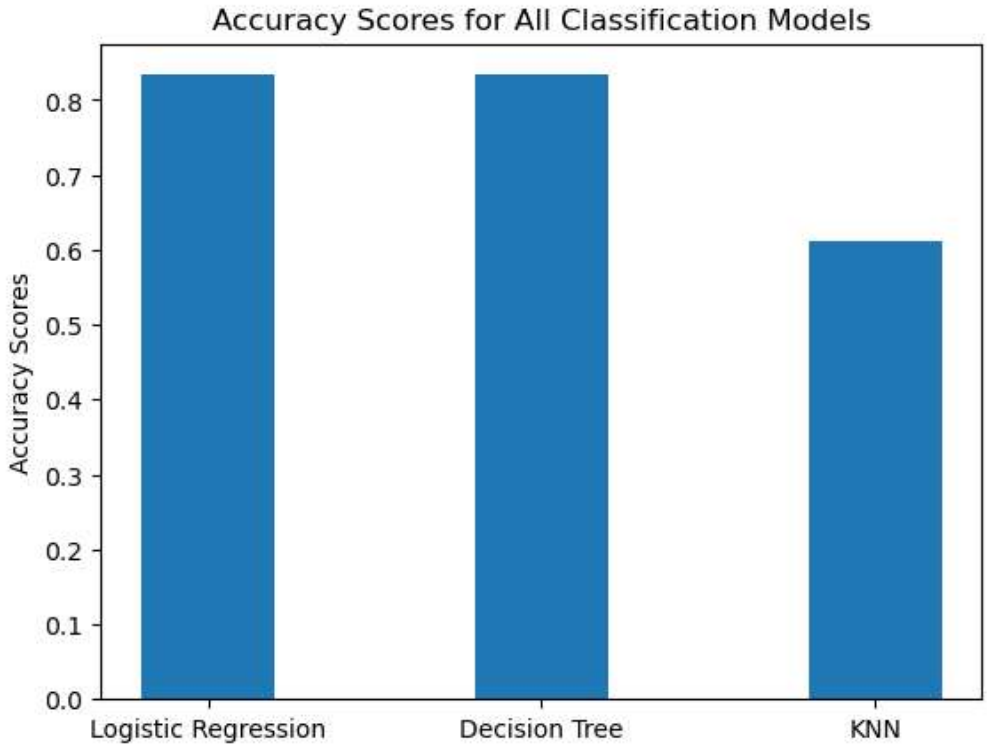
- It can be seen that the number of failed rockets outnumbered those succeeded. Besides, «FT» seems to be the booster with the most successful outcomes and those attempts were achieved by those with the mass between 2000 and a number above 5000kg.

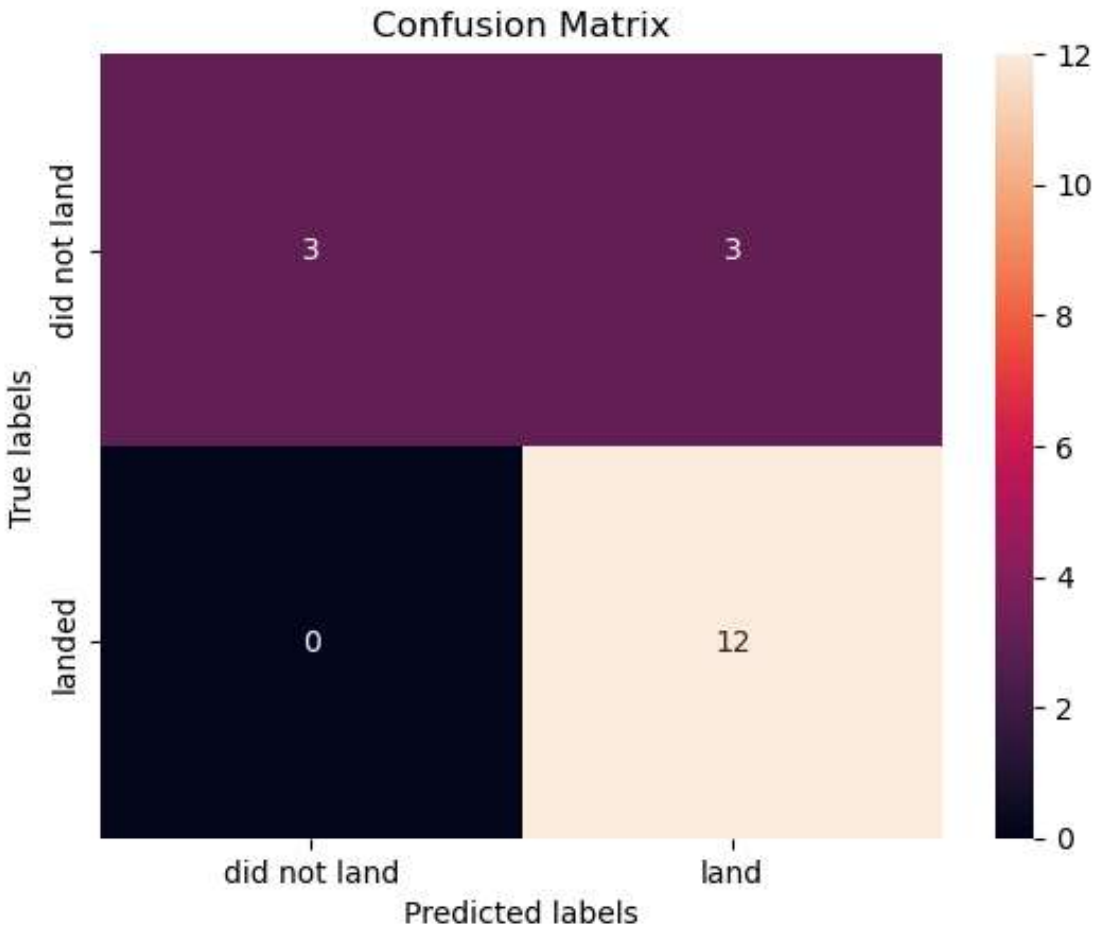
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic Regression and Decision Tree have the same highest score with 0.8%, a powerful score, while the score of KNN is 0.6% which could still be considered good since it is above the average. Support Vector Machine was not applicable because it failed to build the model due to timeout.





Confusion Matrix

This is the confusion matrix of both logistic regression and decision tree classifiers since they have the same score. The models predicted all the values as «landed» correctly but predicted 3 of the values «did not land» as «landed» which are inaccurate.

Conclusions

- A lot of data science techniques like data collection, web scraping, data wrangling, exploratory data analysis using SQL and data visualization, interactive and predictive analysis were taught and those are all used on the SpaceX dataset.
- The easiest methods could be the exploratory data analysis using visuals and predictive analysis using classification algorithms compared to the rest which require a lot of effort and time.
- The most difficult methods were the web scraping and data collection as those involved a little knowledge of programming.
- This project helped to be aware of those different methods and be able to observe the features of the dataset with those techniques.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!