IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexandre Andrade
30/10/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using web scraping and Spacex API;
  - Exploratory Data Analysis (EDA) with data wrangling, data visualization and interactive analysis;
  - Machine Learning to prediction.
- Summary of all results
  - EDA allowed to identify which features are the best predictor for success launchings;
  - Machine Learning showed the best model to discovered which characteristics are important.

Introduction

- The main purpose is to evaluate the viability of the new company Space Y to compete with Space X.
- Principal points to be answered:
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
  - Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

- Executive Summary
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

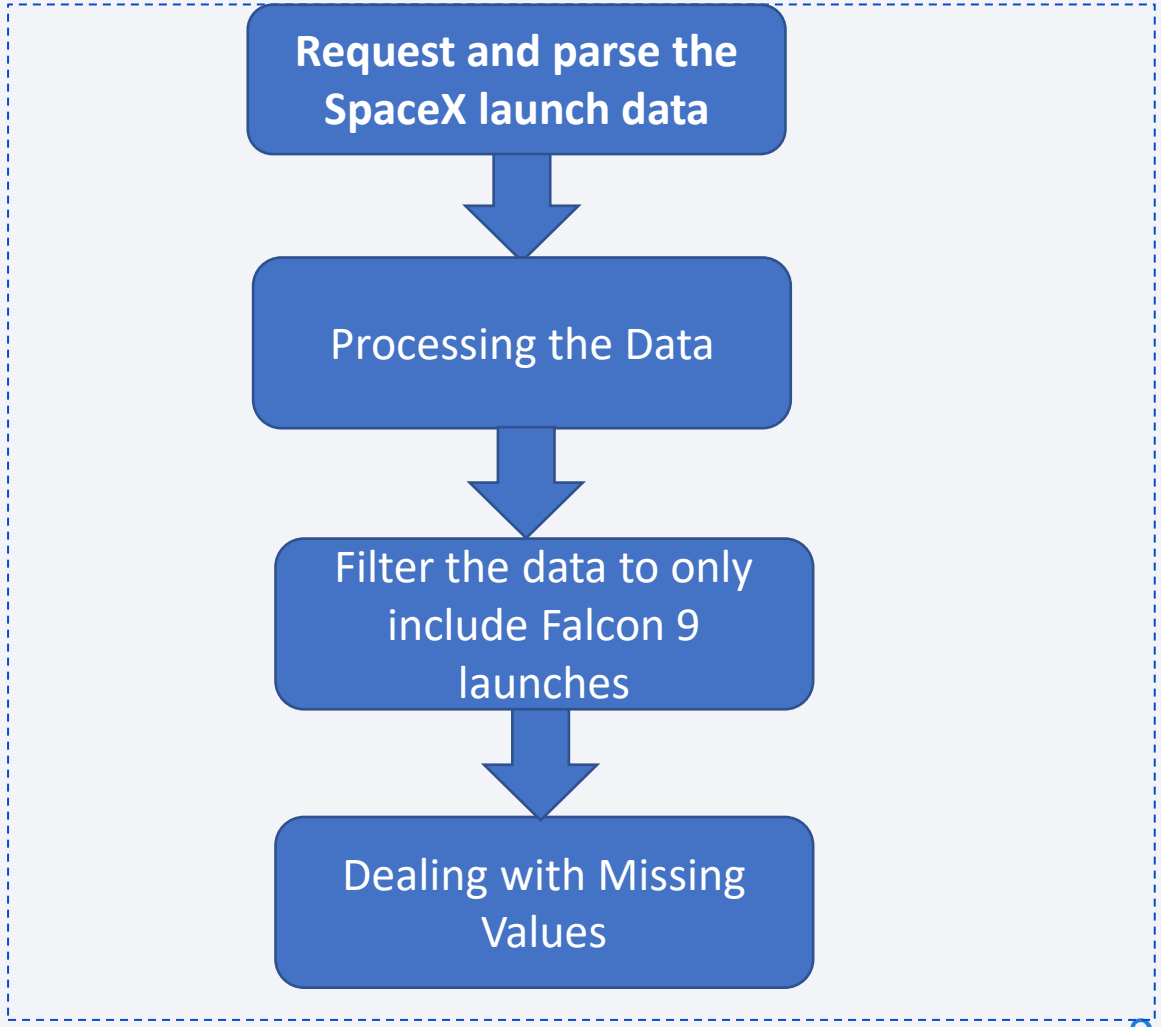
- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), by using web scraping.

- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

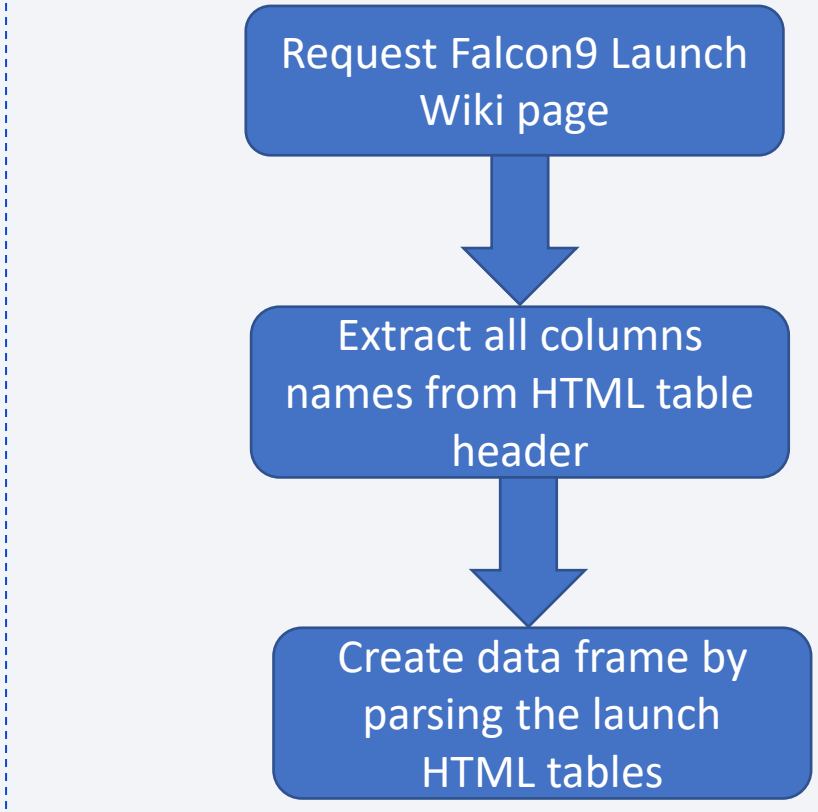
- Public API from SpaceX, where data can be obtained and used;

- Source: https://github.com/AMRAndrade/Corsera/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source: https://github.com/AMRAndrade/Corsera/blob/main/Data%20Collection%20Webscraping.ipynb



Data Wrangling

- We started doing some exploratory data analysis (EDA) on the dataset and determined the training labels.
- Then we calculated the summaries launches per site, occurrences of each orbit and occurrence of mission outcome per orbit type.
- At the end, we created landing outcome label form outcome column and exported the results to a csv file.

- Source: Source: https://github.com/AMRAndrade/Corsera/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- We used scatter point chart, bar chart and line chart.
- Scatter point chart was used to analyze the relationship between pair of features like Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number and Orbit vs Payload Mass.
- Bar chart was used to check if there were any relationship between success and orbit type.
- Line chart was used to visualize the launch success yearly trend.
- Source: https://github.com/AMRAndrade/Corsera/blob/main/EDA%20Using%20Python.ipynb

EDA with SQL

- SQL queries were performed:
  - The names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string ‘CCA’;
  - The total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000Kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Months, booster version and launch site for failed landing outcomes in 2015;
  - Rank of the count of landing outcomes between 04/06/2010 and 20/03/2017;

- Source: https://github.com/AMRAndrade/Corsera/blob/main/EDA%20Using%20SQL.ipynb

Build an Interactive Map with Folium

- We've designated all launch locations and incorporated map elements such as pins, circles, and lines to signify the outcome of launches, whether they were successful or not, on the Folium map.
- We categorized the launch results into two classes: 0 for failure and 1 for success.
- Utilizing color-coded clusters of markers, we pinpointed the launch sites with notably higher success rates
- We conducted distance calculations between each launch site and its immediate surroundings, addressing various inquiries, including:
  - Are these launch sites situated in proximity to railways, highways, or coastlines?
  - Do the launch facilities maintain a specific distance from urban areas?

- Source: https://github.com/AMRAndrade/Corsera/blob/main/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We've constructed an engaging dashboard using Plotly Dash.
- We've generated pie charts illustrating the cumulative launches conducted at specific sites.
- We've created a scatterplot depicting the correlation between the Outcome and Payload Mass (Kilograms) across various iterations of boosters.
- Link to notebook: https://github.com/AMRAndrade/Corsera/blob/main/Space%20Dash.py

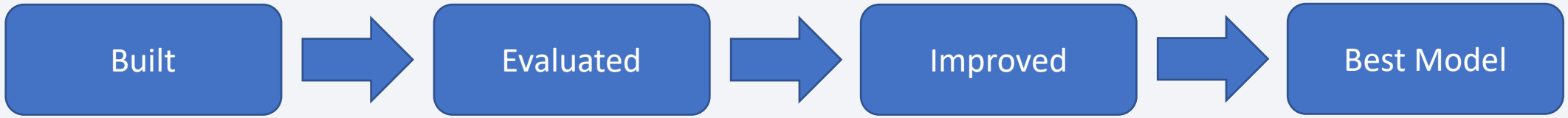
Predictive Analysis (Classification)

- Create numpy array
- Normalize Data
- Split data X and Y into training and test
- Decide each models we will use
- Set and fit datasets into the GridSearchCV

- Calculate the accuracy for each model
- Turned hyperparameters to find the best parameters
- Plot Confusion Matrix

- Feature Engineering
- Algorithm Optimization

- Finding the best model
- Finding the best accuracy
- Finding the best parameters



Link to source:

https://github.com/AMRAndrade/Corsera/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

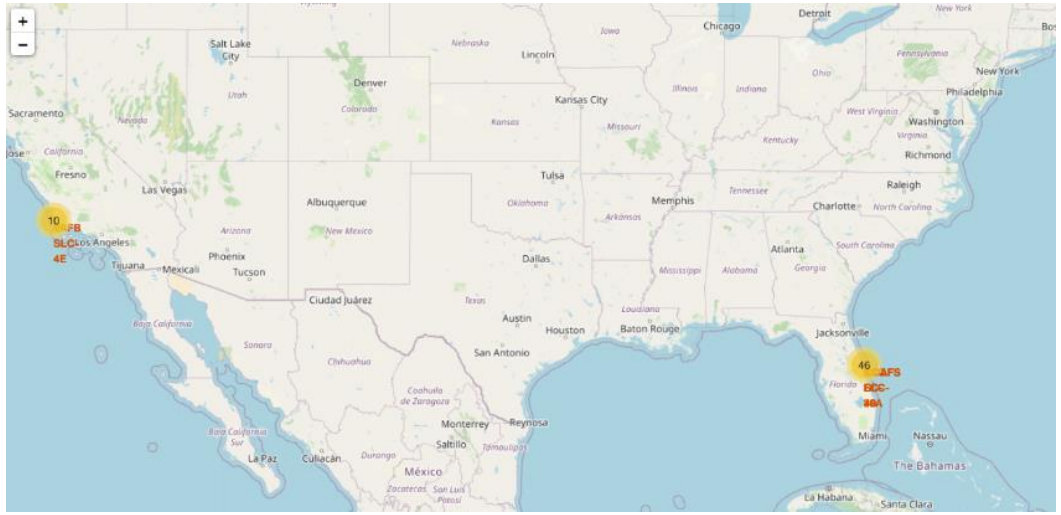
Exploratory data analysis results:

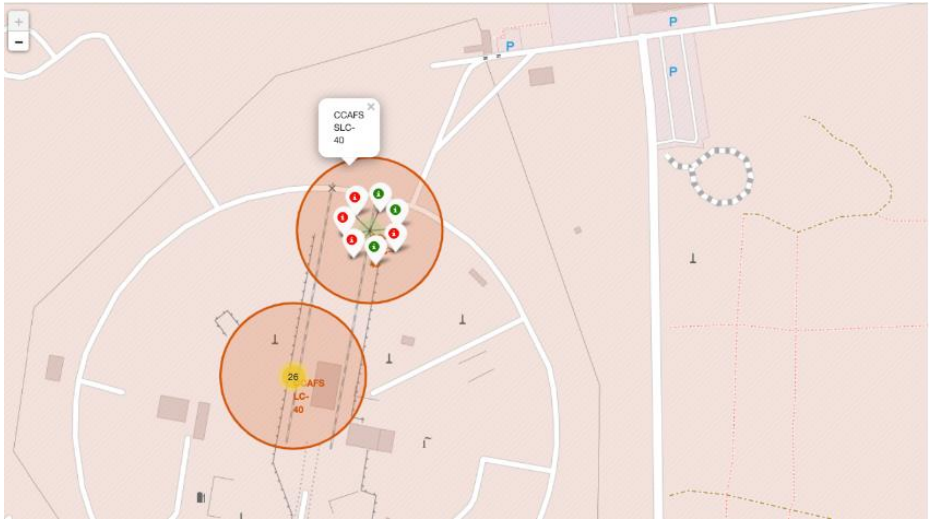
- SpaceX operates from four distinct launch facilities.
- The initial launches were conducted for SpaceX itself and in partnership with NASA.
- The typical payload capacity of the F9 v1.1 booster stands at 2,928 kilograms.
- The first successful landing took place in 2015, five years after the inaugural launch.
- Numerous iterations of Falcon 9 boosters have achieved successful landings on drone ships with payloads exceeding the average.
- Nearly 100% of missions have resulted in success.
- In 2015, two booster versions, namely F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landings on drone ships.
- The frequency of successful landing outcomes has improved over the years

Results

Interactive analytics

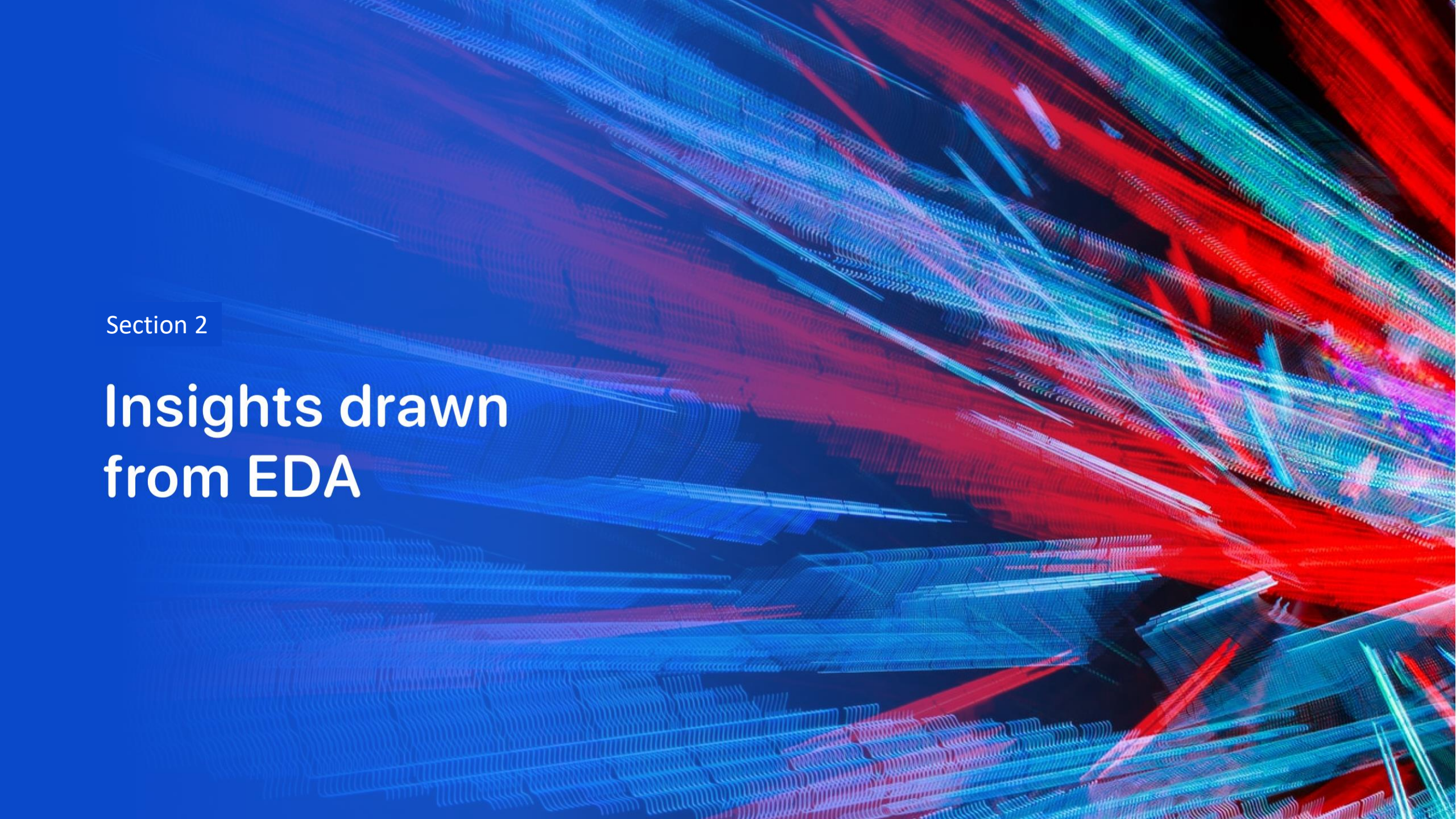
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- • Most launches happens at east cost launch sites.



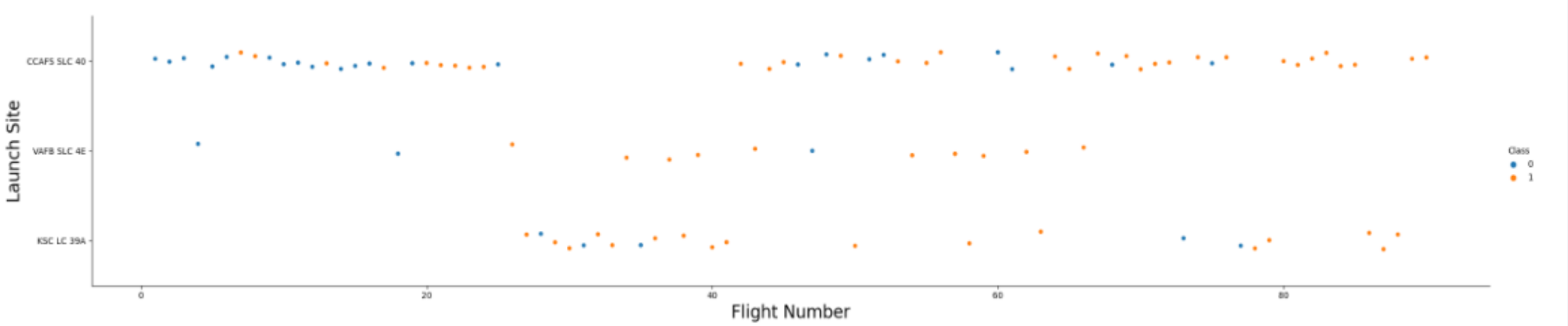


Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%

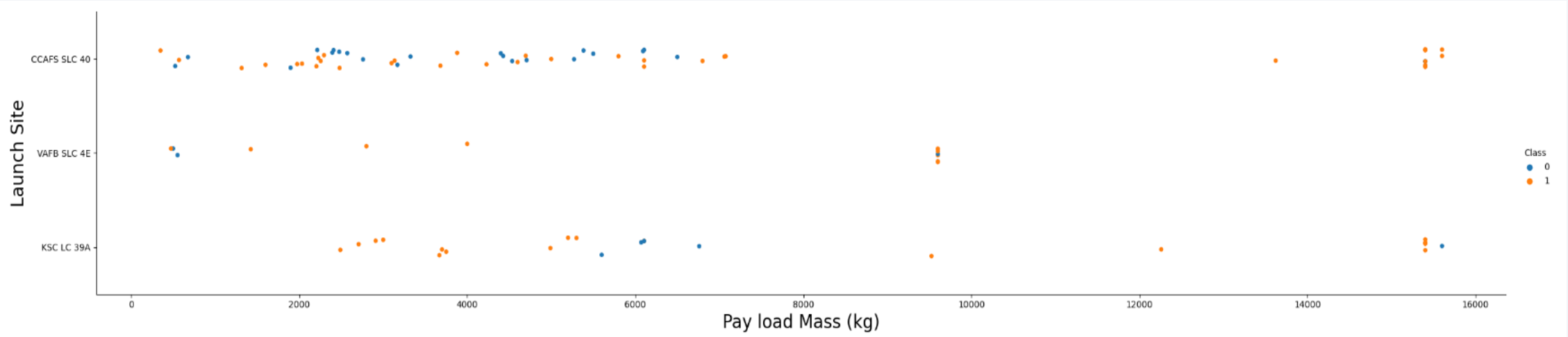

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



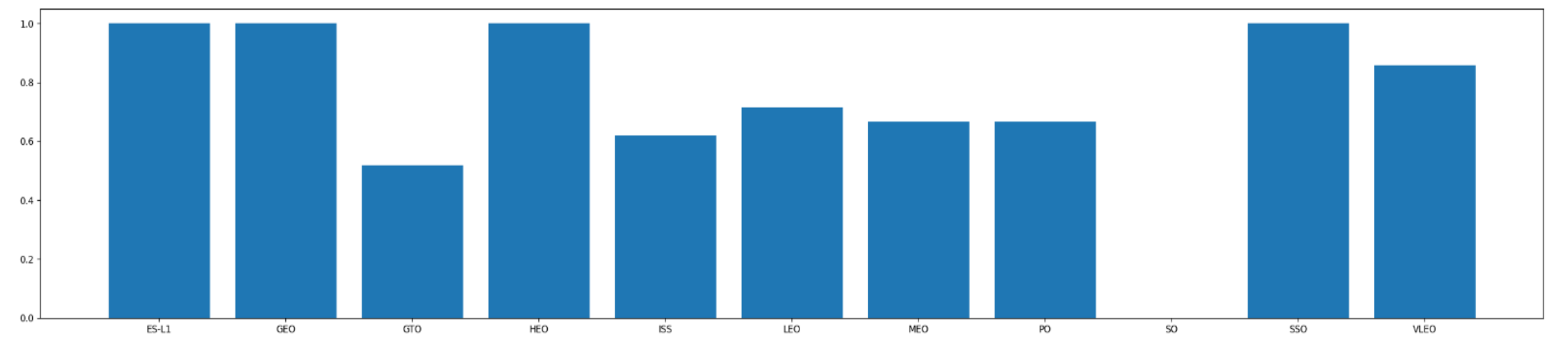
- Based on the provided chart, we can confirm that the optimal launch facility at present is CCAF5 SLC 40, which has witnessed a majority of recent launches achieving success.
- In the second position, we have VAFB SLC 4E, and in the third position, KSC LC 39A.
- Moreover, the overall rate of successful launches has displayed a noticeable improvement as time has progressed.

Payload vs. Launch Site



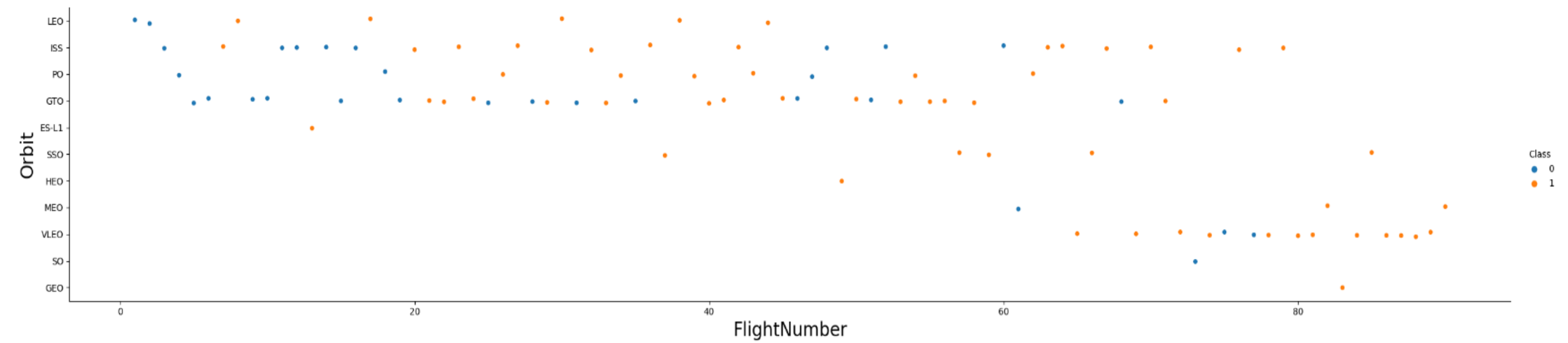
- Payloads exceeding 9,000kg (approximately equivalent to the weight of a school bus) exhibit a commendable success rate.
- Payloads surpassing 12,000kg appear to be feasible primarily at the CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



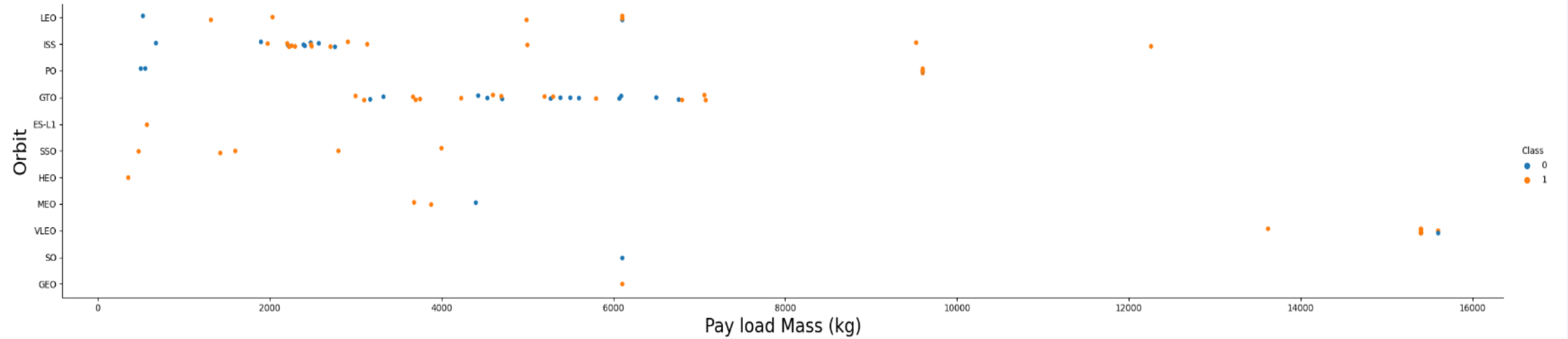
The biggest success rates is: ES-L1; GEO; HEO and SSO

Flight Number vs. Orbit Type



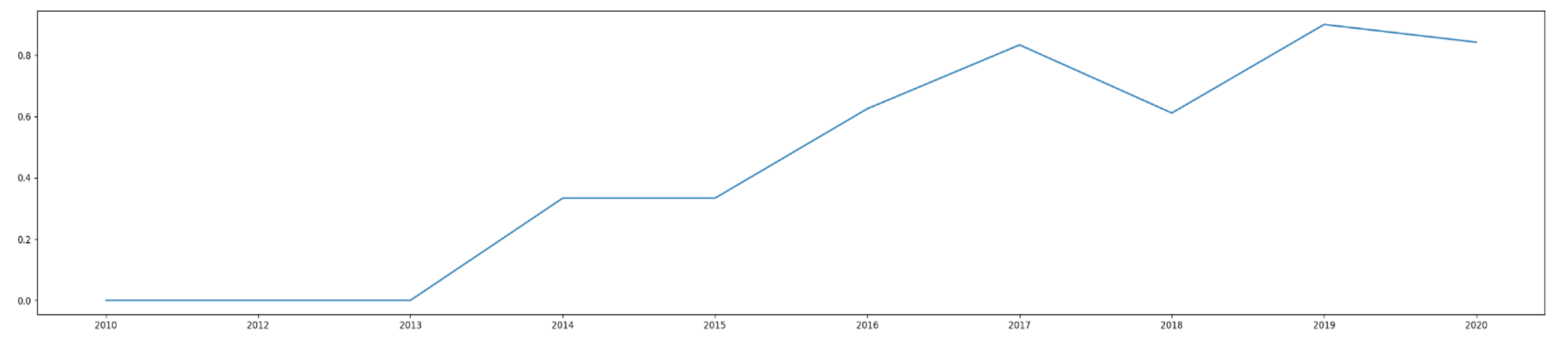
- Evidently, the success rate has shown an improvement over time across all orbital trajectories.
- The Very Low Earth Orbit (VLEO) presents itself as a promising new business opportunity, as its frequency has recently witnessed a notable increase.

Payload vs. Orbit Type



- It seems that there is no discernible correlation between payload and the success rate for the Geostationary Transfer Orbit (GTO).
- The International Space Station (ISS) orbit stands out with the most extensive payload range and a commendable success rate.
- Notably, there are relatively few launches to the Suborbital (SO) and Geosynchronous Equatorial Orbit (GEO) trajectories.

Launch Success Yearly Trend



- The success rate began its ascent in 2013 and continued to rise until 2020.
- It appears that the initial three years marked a period of adjustments and technological improvements.

All Launch Site Names

There are 4 launch sites:

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

We obtained this results locking for the unique occurrences of “launch_site”

Launch Site Names Begin with 'CCA'

- 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
04/06/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08/12/2010	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)
22/05/2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08/10/2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01/03/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

- We select all information of the 5 records of the data set

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload
48213

- We calculate the total payload above, by summing all payload_mass_Kg where customer is Nasa (CRS)

Average Payload Mass by F9 v1.1

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

Avg Payload
2928.4

- We calculate the average payload above, by using the avare of all payload_mass_Kg where Booster_Version is F9 v1.1

First Successful Ground Landing Date

- The first successful landing outcome on ground pad

First Success Landing
2015-12-22

- We select the minimum date where the “Landing_Outcome” is Success (ground pad)

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

- We select the Booster_Version names where “Landing_Outcome” is Success (drop ship) and Payload Mass Kg are between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Mission	Number
Success	100
Failure	1

- We classify the Mission Outcome in “Success” and “Failure” status and count the number of occurrences of each

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 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

- We select the uniques Booster_Version that carried the maximum payload mass.

2015 Launch Records

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

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

- We achieved this result by filtering the information according to two criteria:

Landing_Outcome is failure and Year is 2015

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

- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Landing _Outcome	Number
Success (drone ship)	5
Success (ground pad)	3

```
%%sql SELECT
    "Landing _Outcome"
    ,COUNT(*) as "Number"
    FROM SPACEXTBL
    WHERE DATE(substr(Date,7,4)
    ||'-'
    ||substr(Date,4,2)
    ||'-'
    ||substr(Date,1,2))
    BETWEEN DATE('2010-06-04') AND DATE('2017-03-20')
    AND "Landing _Outcome" LIKE "Success%"
    GROUP BY "Landing _Outcome"
```

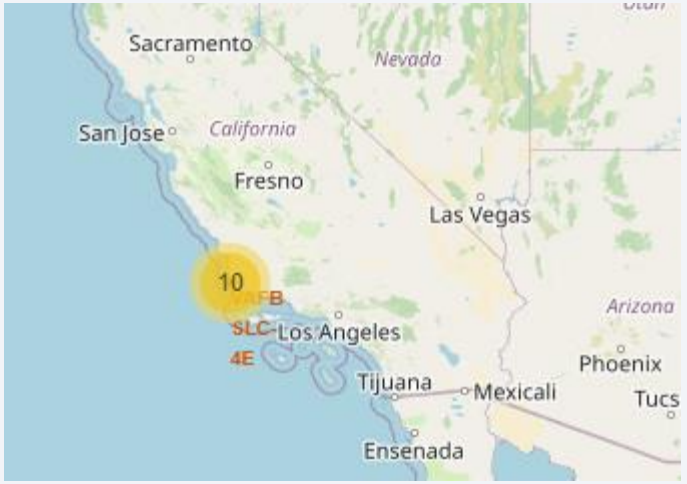
This SQL query summarizes the number of successful landing outcomes within the date range of June 4, 2010, to March 20, 2017, by counting occurrences with "Landing _Outcome" values that start with "Success."

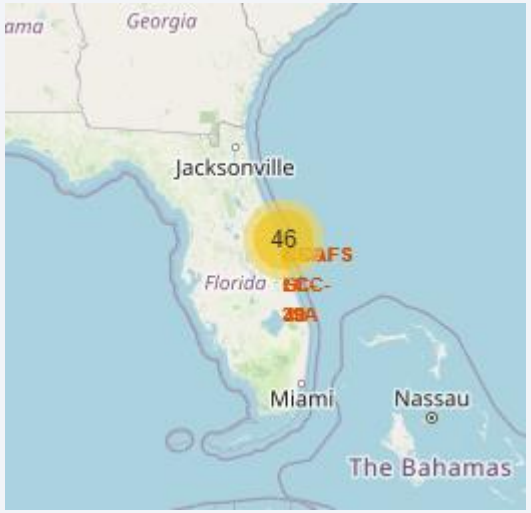

Section 3
Launch Sites
Proximities Analysis

All launch sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

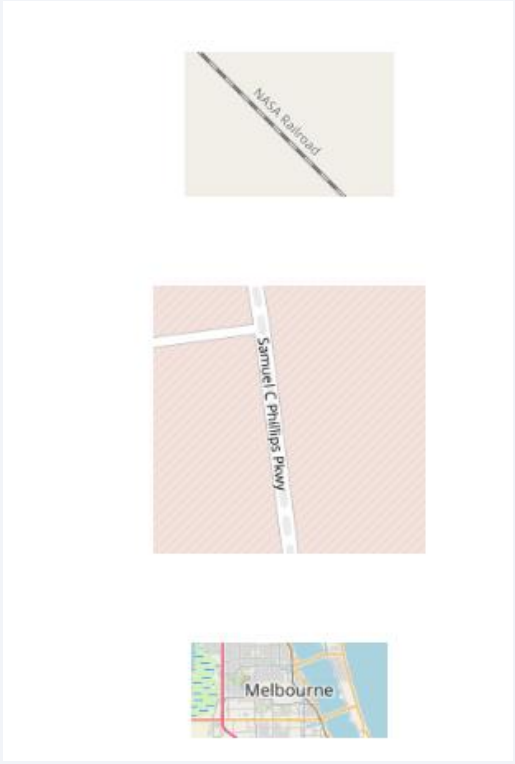
Markers showing launch sites with color labels



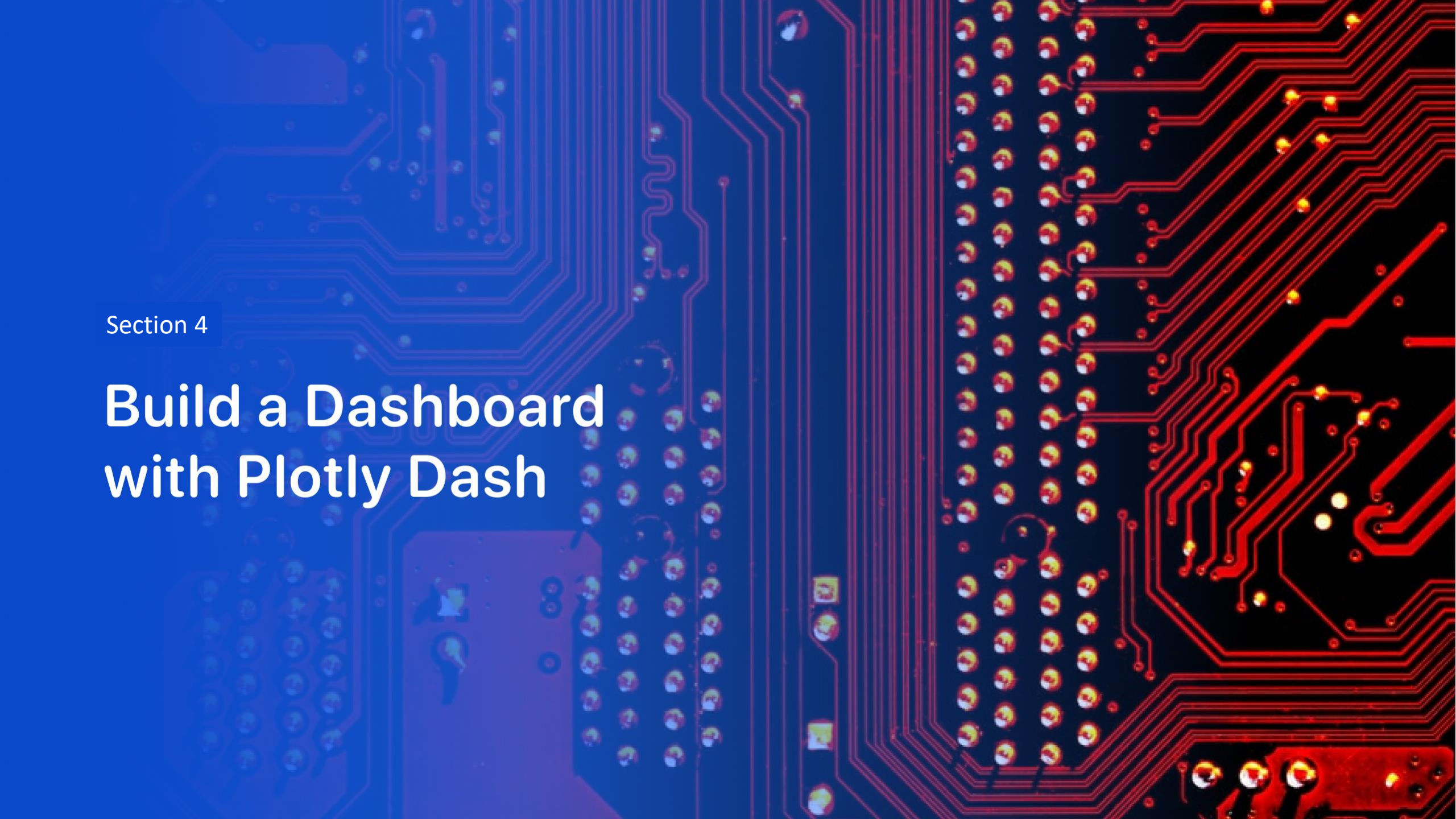


Green markers indicate successful and red ones indicate failure.

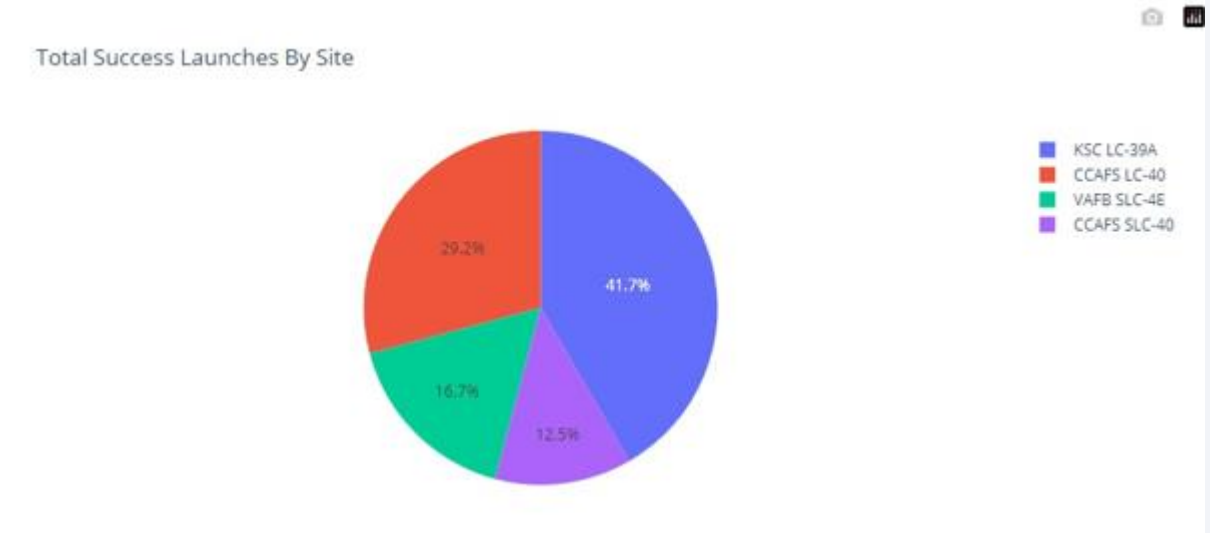
Launch Site distance to landmarks



With this map we can see that all launches were made close to the coastline and far from railways or highways

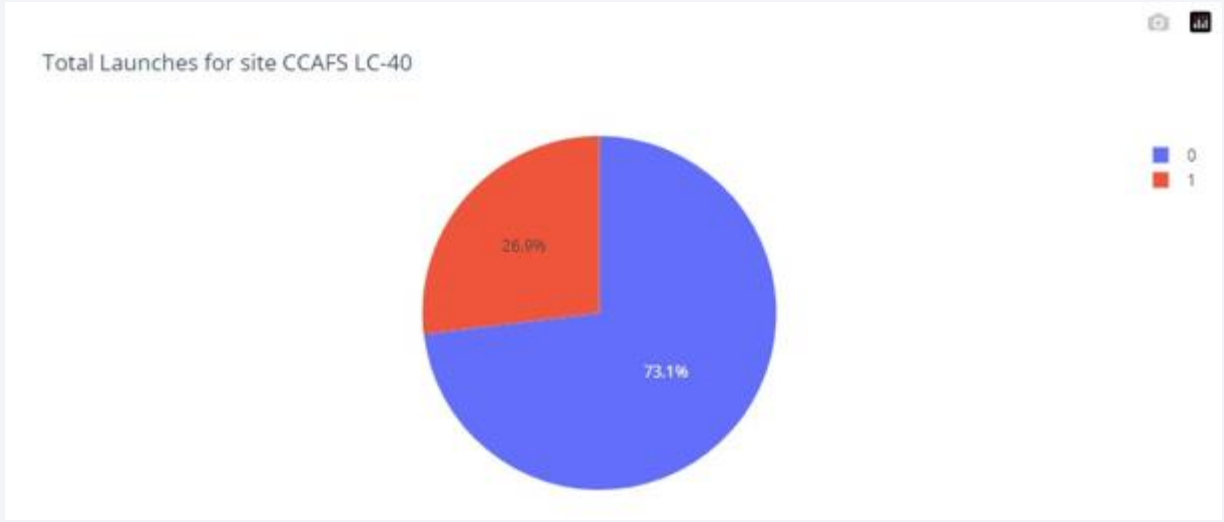

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



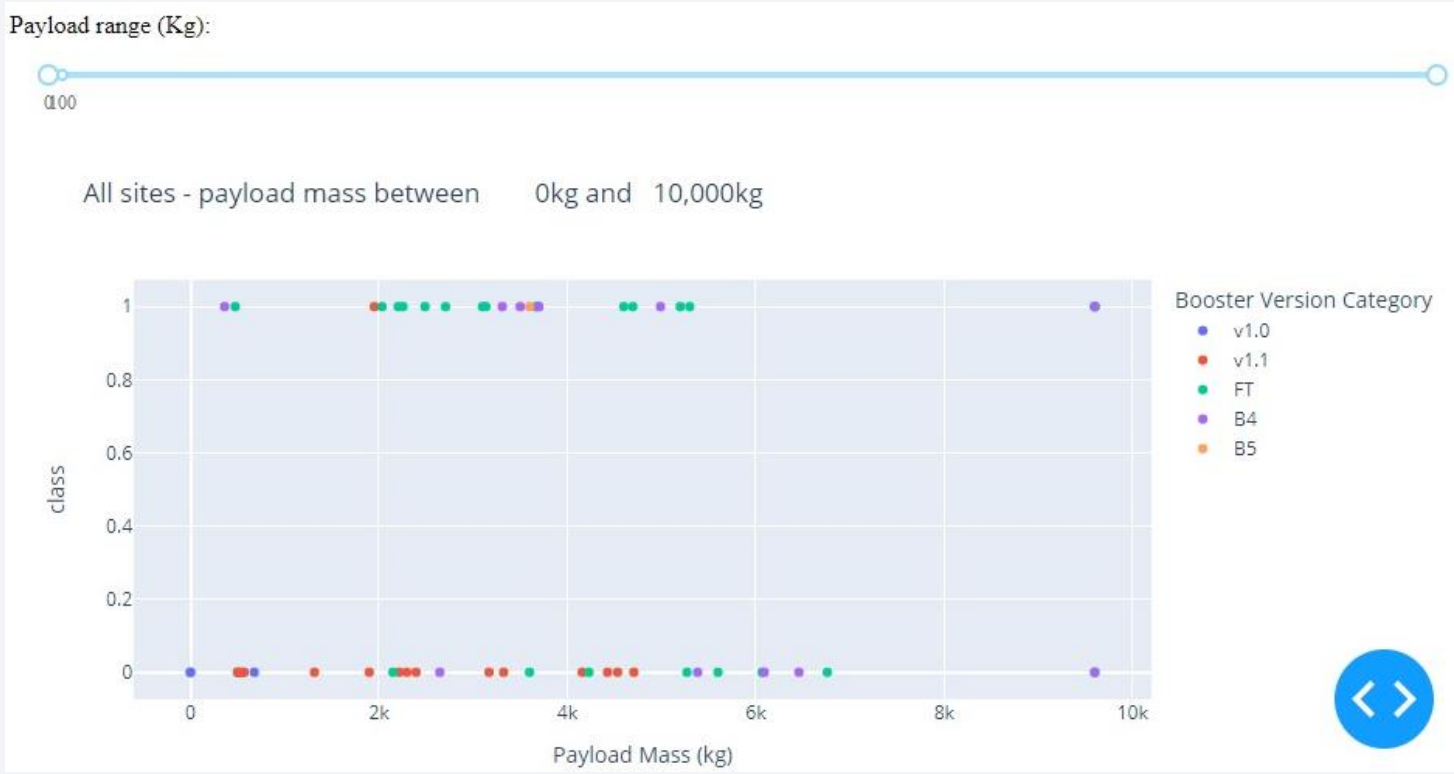
The location where launches originate appears to be a crucial determinant of mission success.

Launch Success Ratio for KSC LC-39A



73.1% of launches are the successful in this site.

Payload vs. Launch Outcome

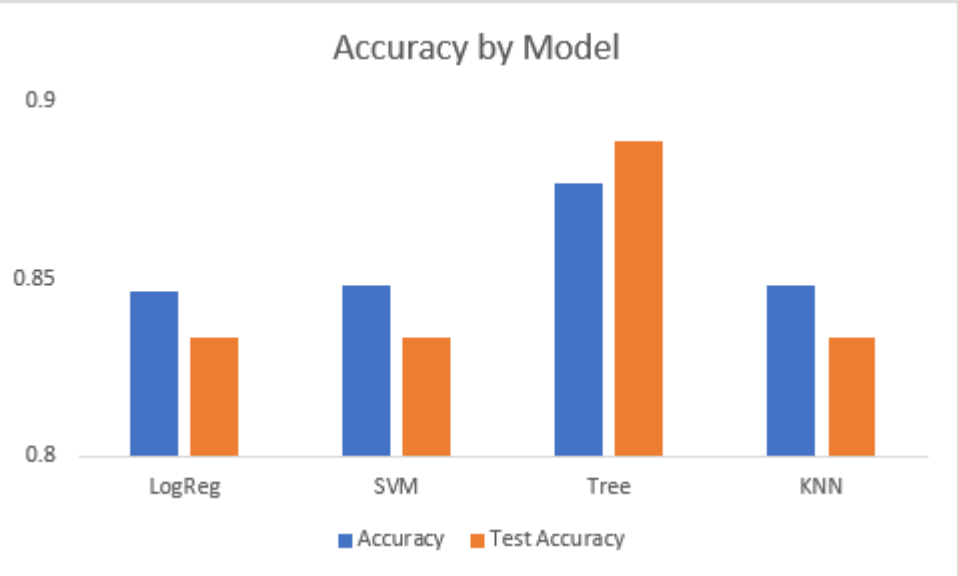


It seems that the best results are obtained using a payload mass between 2000 and 4000 and using the FT Booster version

Section 5

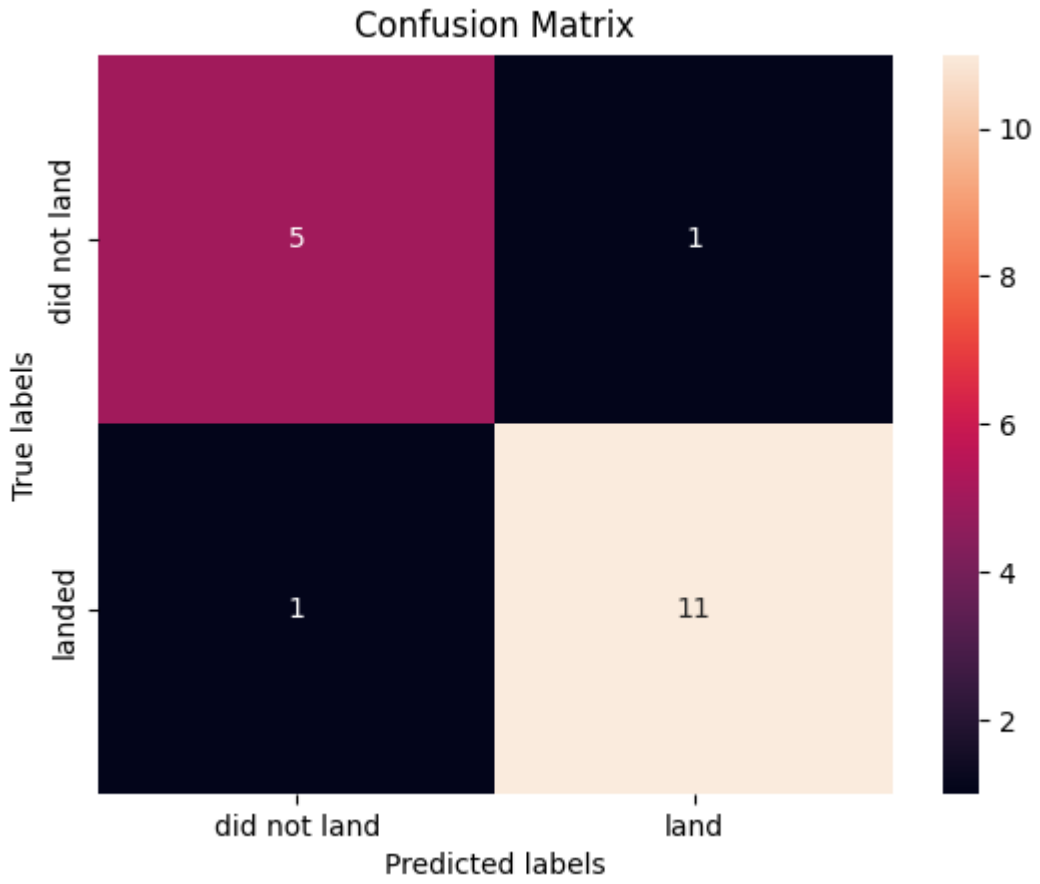
Predictive Analysis (Classification)

Classification Accuracy



- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.

Confusion Matrix



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

Conclusions

- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- 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.

Appendix

See my Githug: https://github.com/AMRAndrade/Corsera/tree/main

Thank you!