IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sakshi Hegde
20/12/2022

Outline

Executive Summary

- Summary of methodologies

- Data Collection using web scraping and SpaceX API.
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics.
- Machine Learning Prediction.

- Summary of all results

- Data was collected from public sources.
- EDA allowed to identify the best features to predict success of launchings.
- Machine Learning Prediction provided the best model to predict which characteristics are important to fit with the given model by using all collected data.

Introduction

- Project background and context:
  - To evaluate the viability of the new company Space Y to compete Space X.

- Solutions related to our problem:
  - The best way to estimate the total cost for launches by predicting the successful landings of the first stage of rockets;
  - Best location for launching rocket.

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/)

  WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform Data wrangling

  Created a landing outcome label based on outcome data after summarizing and analyzing features

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

The trend of data was been explored more by using SQL along with visualization features.

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash

Map was created to mark various rocket launching locations by using folium feature and Dashboard was used to display various trend and how data various according to the trend.

- Perform predictive analysis using classification models

Machine learning model was used in order to obtain the best outcome related to obtained data and hence based on it what would be best possible prediction.

- How to build, tune, evaluate classification models

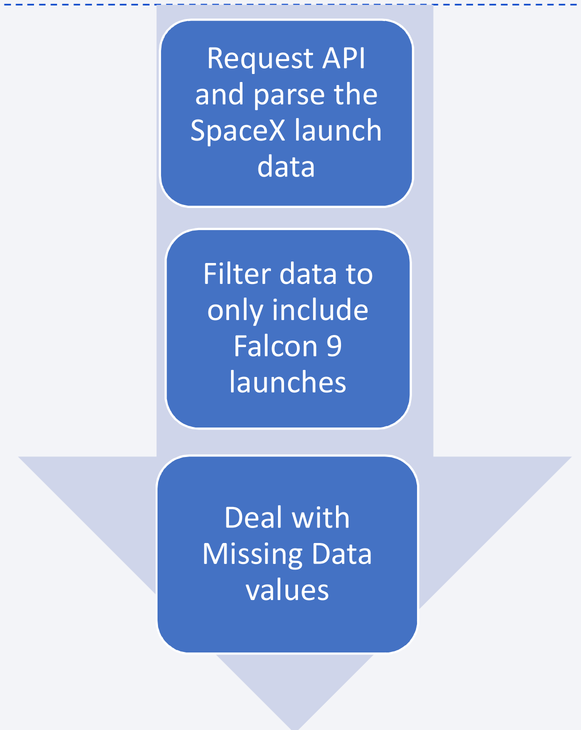
Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters

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), using web scraping technics.

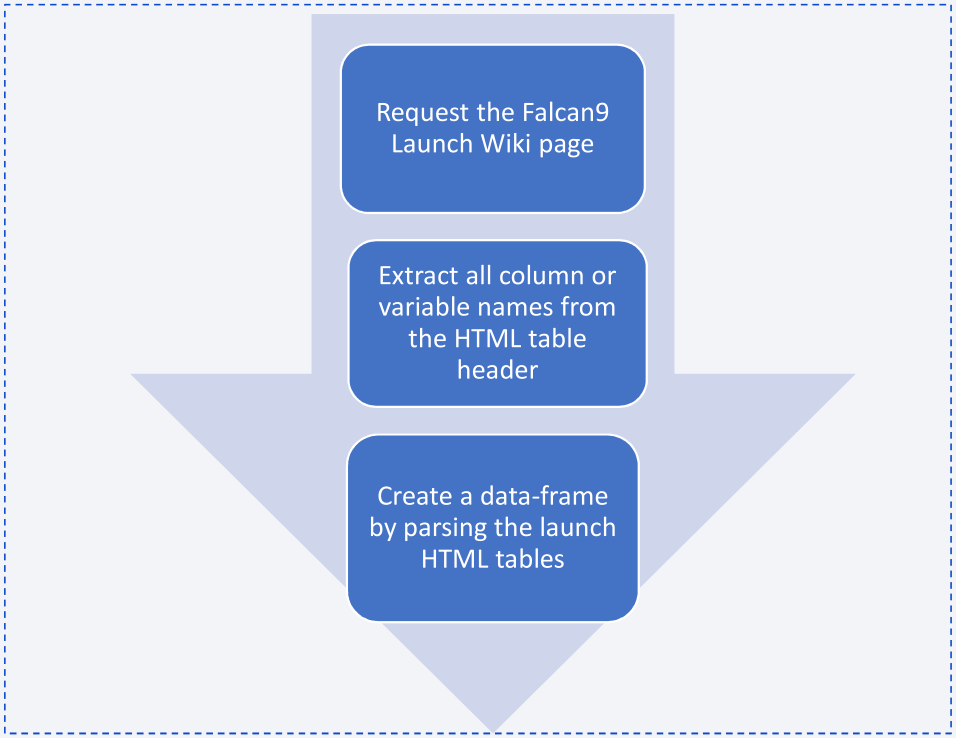
Data Collection - SpaceX API

- SpaceX offers a public API from where data can be obtained and can be parsed.
- This API was used according to the flowchart which is displayed.
- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/API%20Lab.ipynb



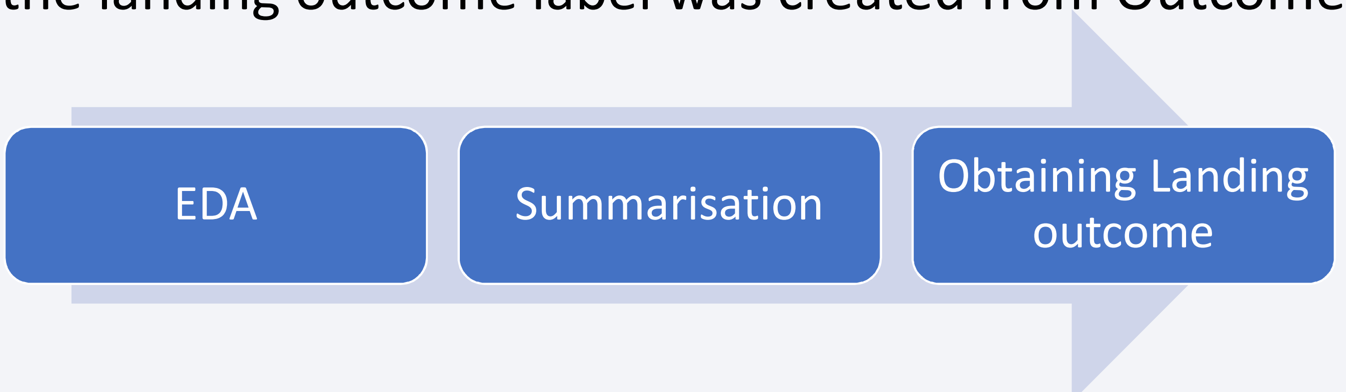
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia.
- Data are downloaded from Wikipedia by following the flowchart and then can be analyzed.
- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/Web%20scraping.ipynb



Data Wrangling

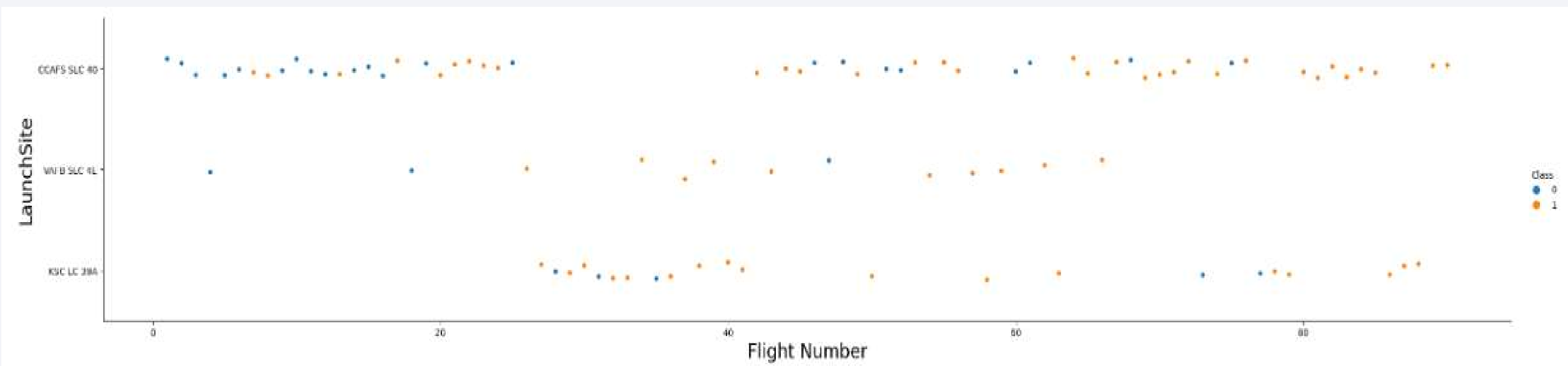
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and bar chart were used to visualize the relationship between pair of features:
- Flight Number Vs Payload Mass, Flight Number Vs Launch Site, Payload Mass Vs Launch Site, Orbit and Flight Number, Payload and Orbit



- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/jupyter-labs-eda-dataviz%20(2).ipynb

EDA with SQL

- The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- 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 between 4000 and 6000 kg
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 04-06-2010 and 20-03-2017

- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- Summarize
- 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; and
- Lines are used to indicate distances between two coordinates
- Explain why those objects were added:
- In order to locate the place of launching marking them on the map we need these objects which were discussed above.
- URL for the code: https://github.dev/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/lab_jupyter_launch_site_location%20(1)%20(1)%20(3).ipynb

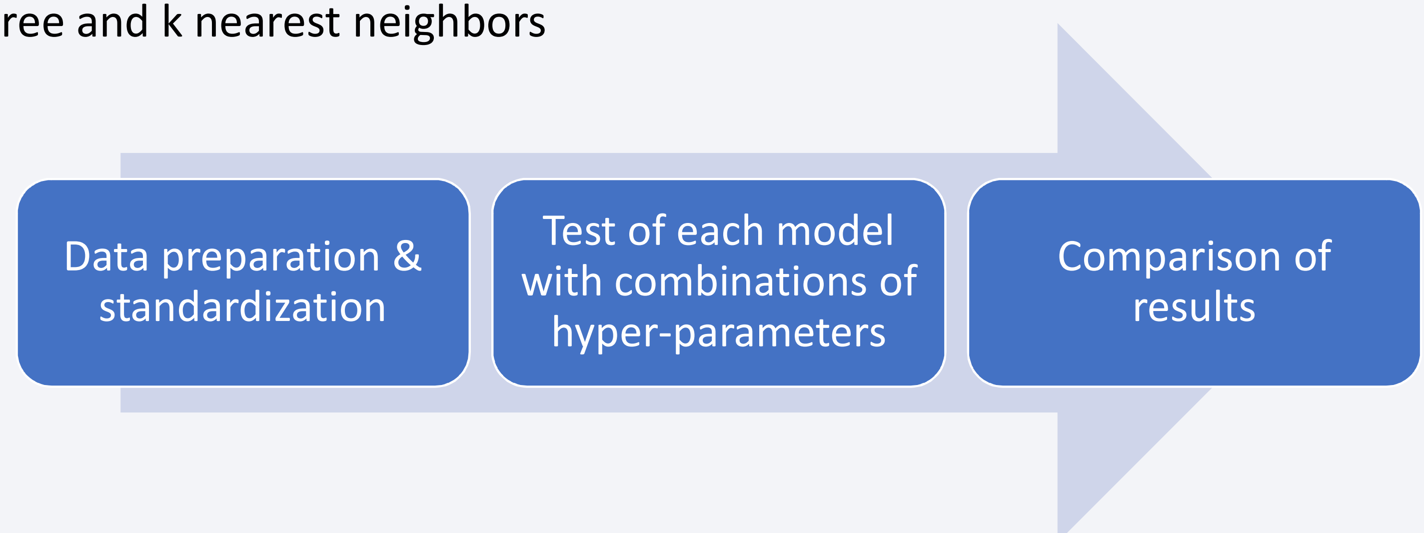
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
- Percentage of launches by site
- Payload range
- Explanation
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads
- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- Summarize

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors



- URL for the code: https://github.com/sakshihegde20/Applied-capstone-Data-Science-projects/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

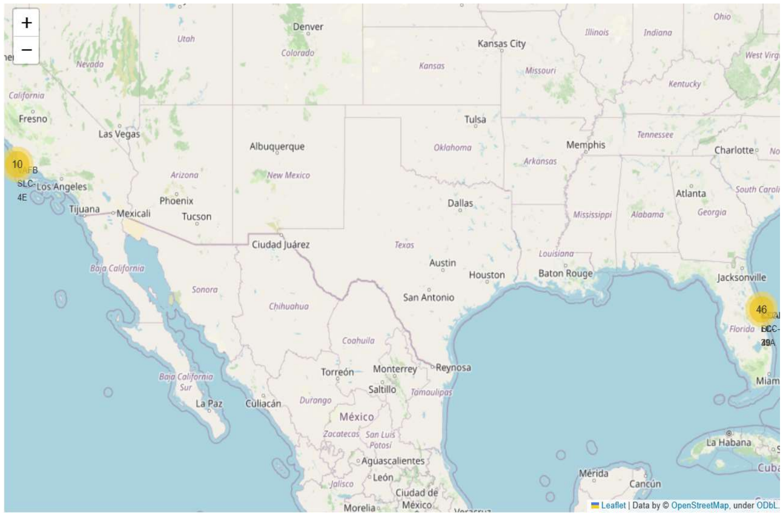
▪ Exploratory data analysis results

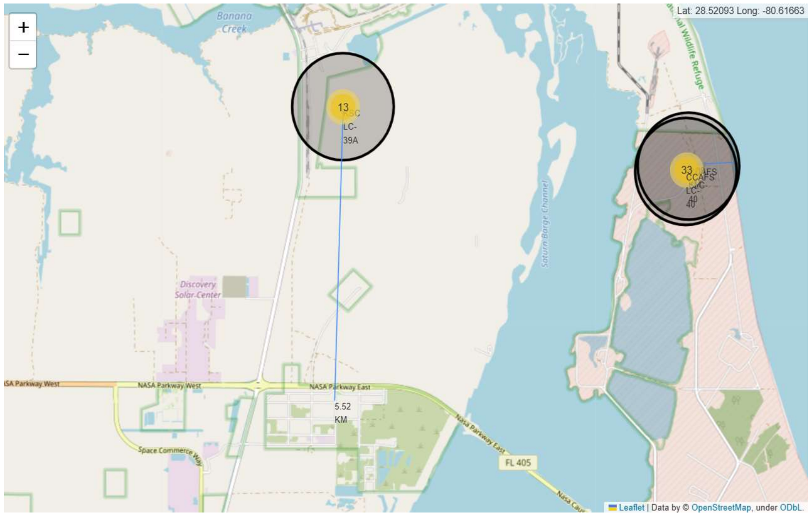
- Space X uses 4 different launch sites
- The first launches were done to Space X itself and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2017 year after the first launch.
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
- Almost 100% of mission outcomes were successful
- One booster version failed at landing in drone ships in 2015: F9 v1.1 B1015
- The number of landing outcomes became as better as years passed

Results

▪ Interactive Analysis:

- 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.

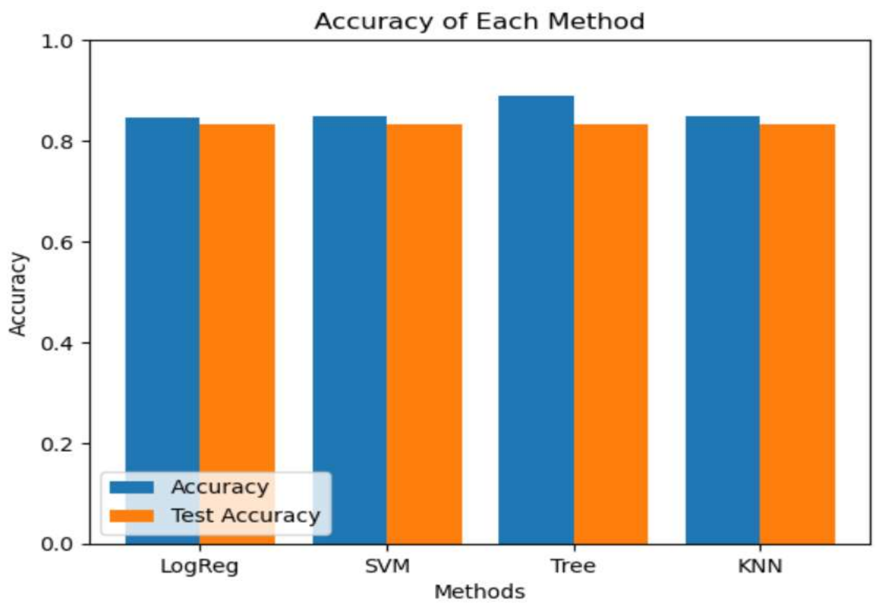




Results

▪ Predictive Analysis:

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over approximately 88.9% and accuracy for test data over approximately 83.3%.

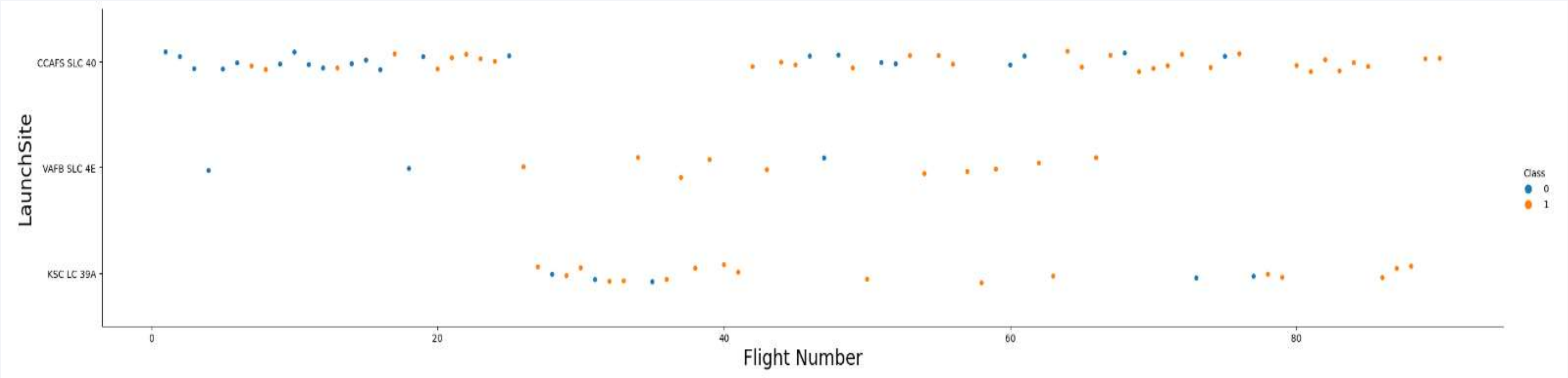


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs Launch Site



- Explanations:

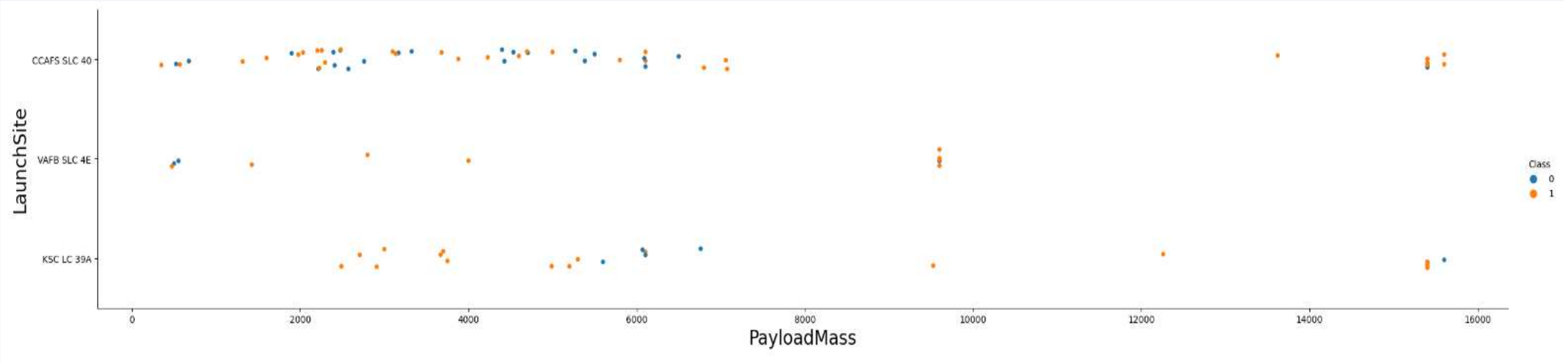
• According to the plot above, it’s possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.

• In second place VAFB SLC 4E and in third place KSC LC 39A.

• It’s also possible to view that the general success rate improved over time.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



- Explanations:

• Payloads over 9,000kg (about the weight of a school bus) have excellent success rate.

• Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

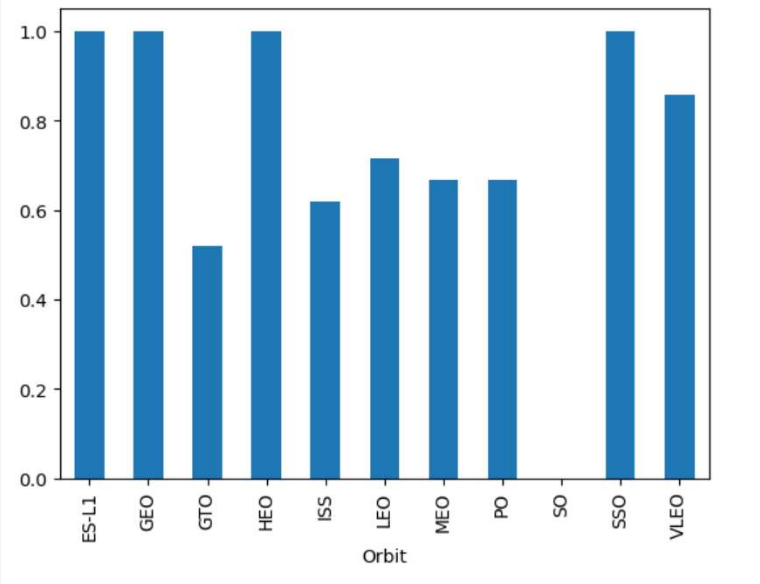
- Explanations:

1. The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO.

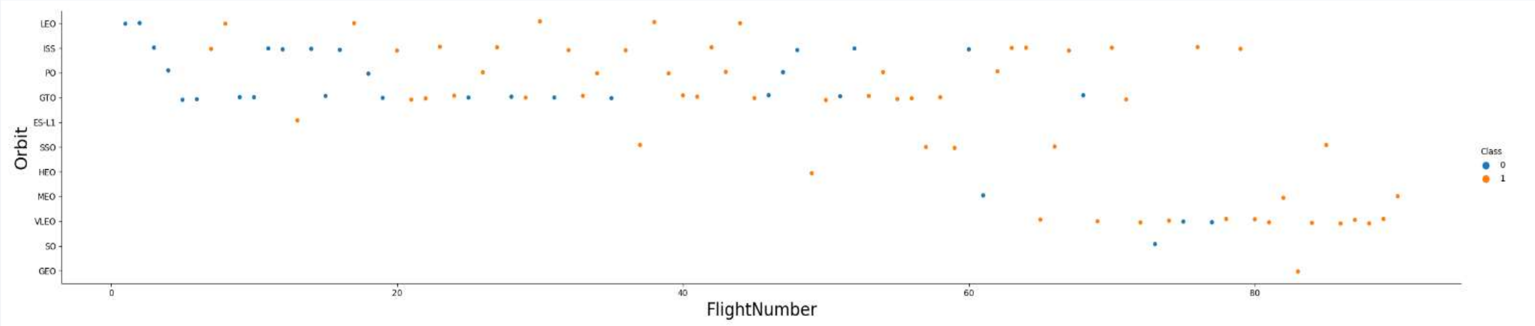
2. Followed by:

- VLEO (above 80%)
- LFO (above 70%).



Flight Number vs. Orbit Type

- Scatter plot of Flight number vs Orbit type

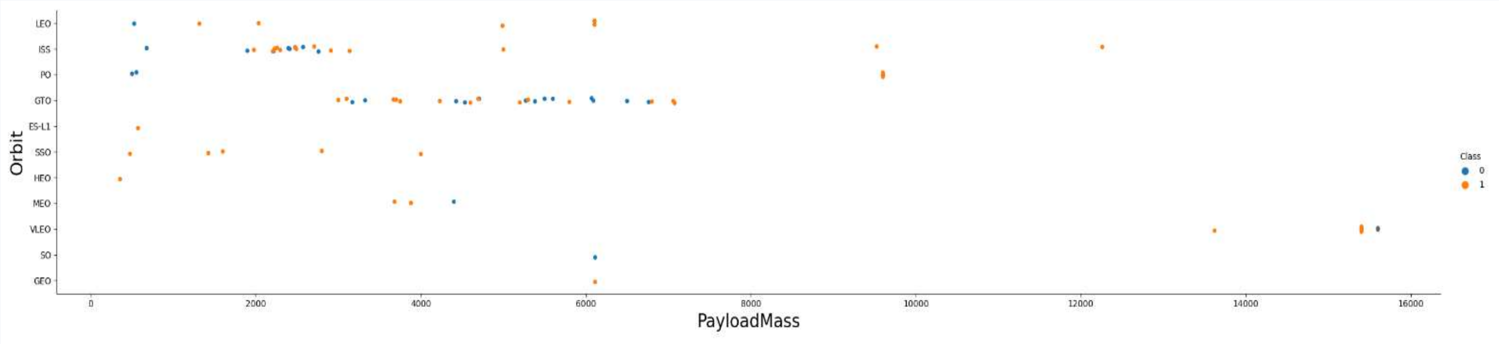


- Explanations:

- Apparently, success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs. Orbit Type

- Scatter plot of payload vs. orbit type



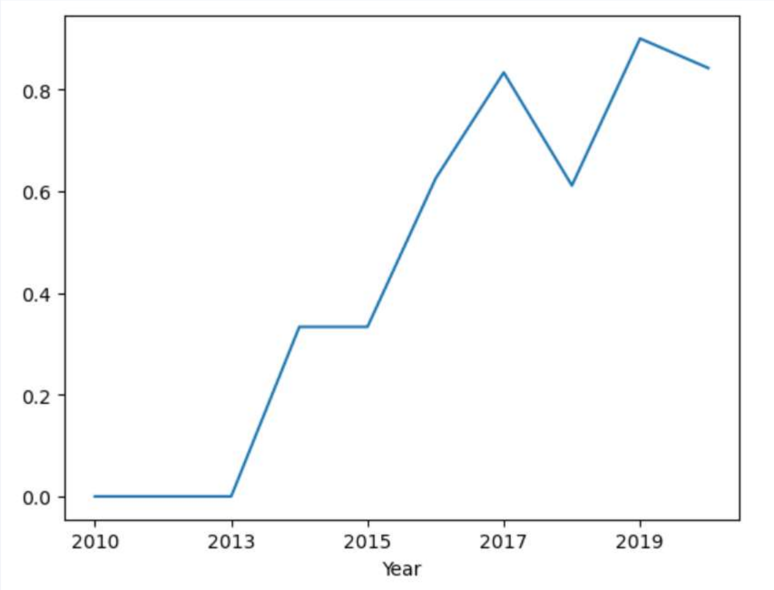
- Explanations:

- There is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020.

- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are five launch sites,

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

- These are obtained by selecting unique occurrences of “Launch_Site” values from the dataset.

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
4/6/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
8/12/2010	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1/3/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Here we can see five samples of Cape Canaveral launches

Total Payload Mass

- Total payload carried by boosters from NASA,

TOTAL_PAYLOAD
111268

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

AVG_PAYLOAD
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad,

FIRST_SUCCESS_GP
1/5/2017

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 1/5/2017.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of 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

- Selecting distinct booster versions according to the filters above, these 4 are the result

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes,

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- List of 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

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

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

Booster_Version	Launch_Site
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only one occurrences.

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

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

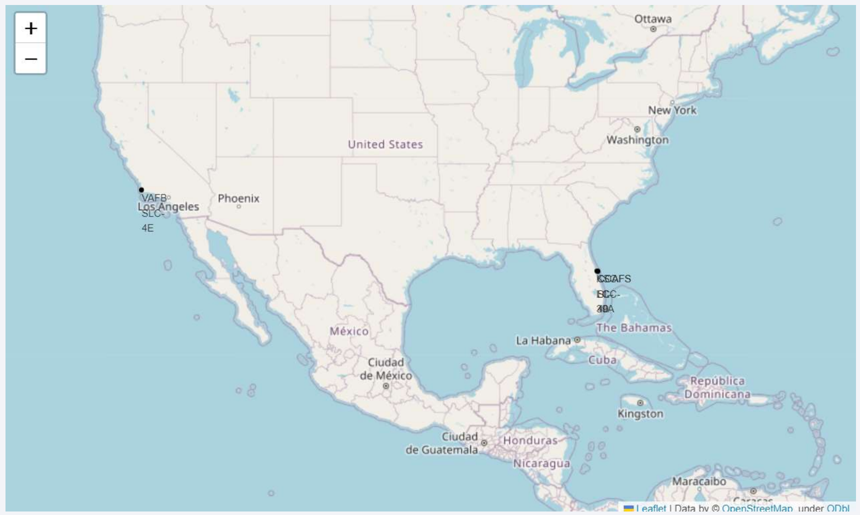
Landing_Outcome	QTY
Success	11
No attempt	6
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	4
Controlled (ocean)	3
Failure	2

- This view of data alerts us that “No attempt” must be taken in account.

Section 3

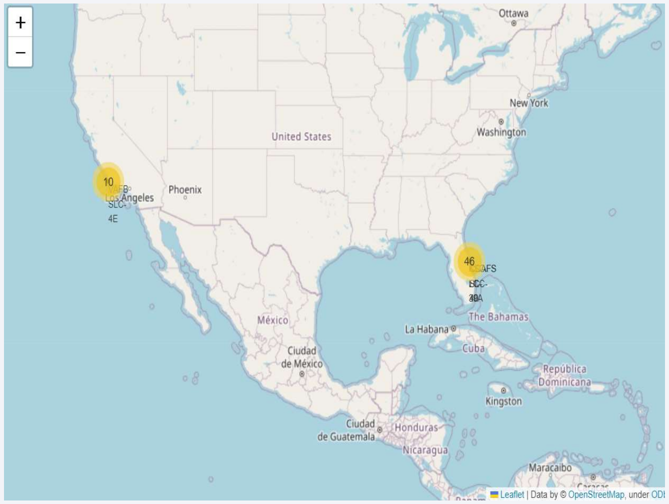
Launch Sites Proximities Analysis

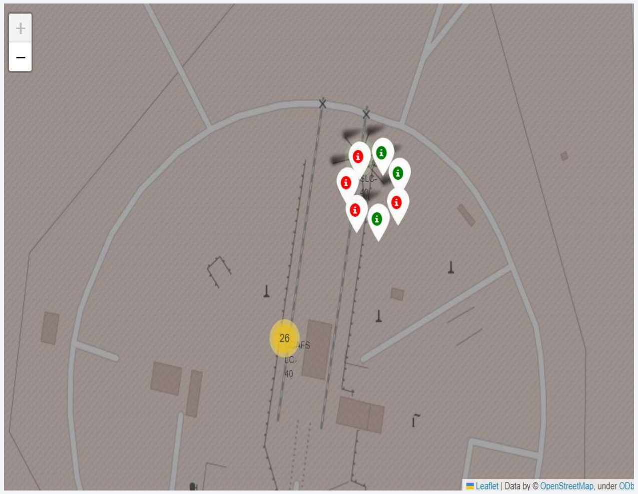
Marked Launch Sites



- Launch sites are near the coastlines, also not too far from the railway and cities.

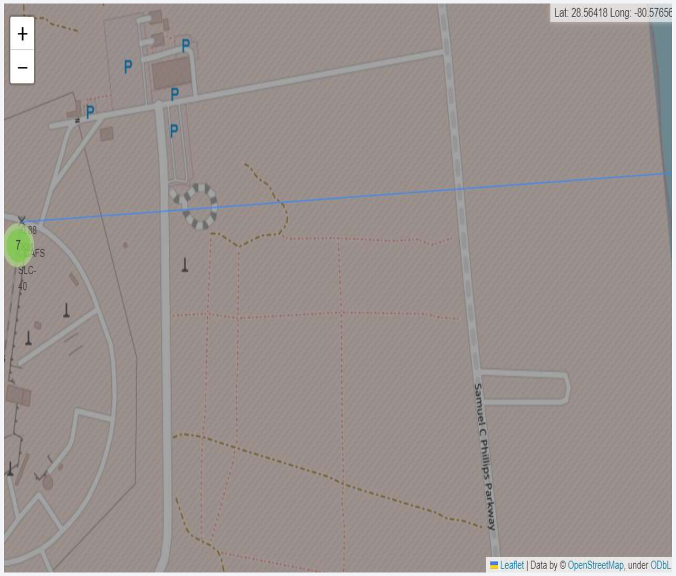
Launch Outcome based on Launch Site

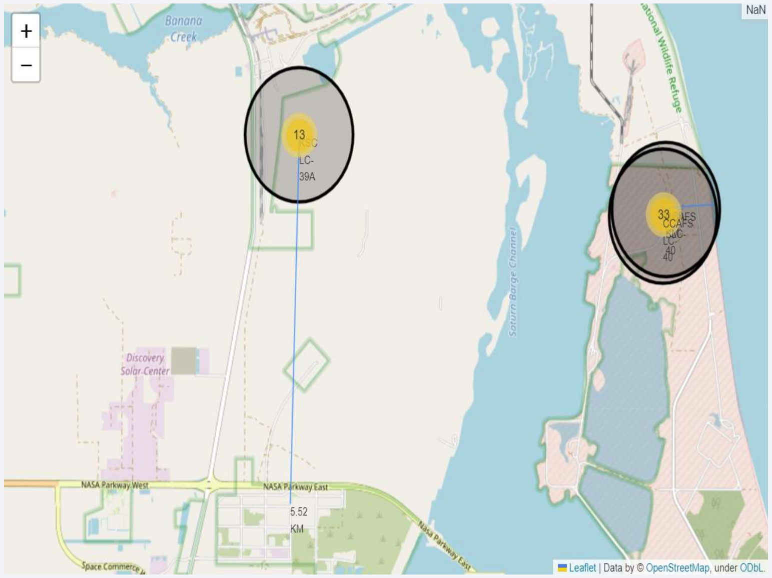




- From third image, Green markers indicate successful and red ones indicate failure

Distances between Launch Site to it's Proximity



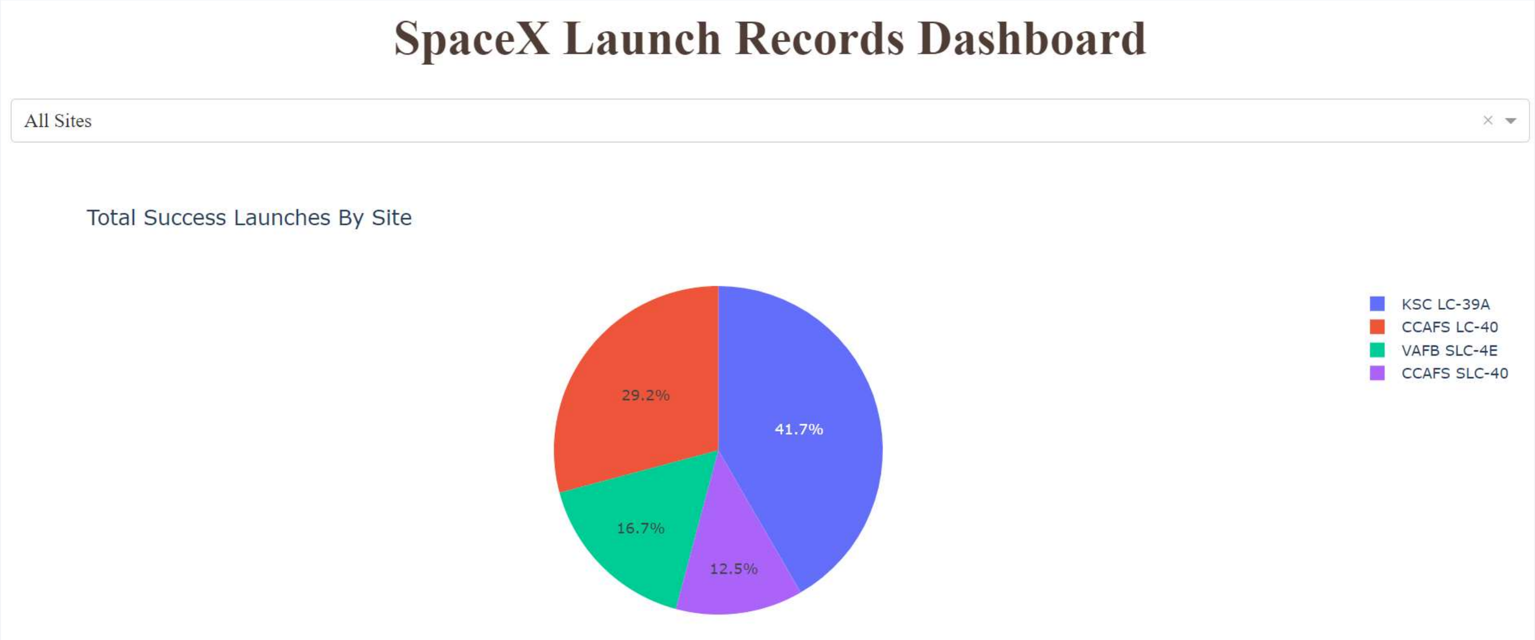


- The launch sites are been near railroad and coastline

Section 4

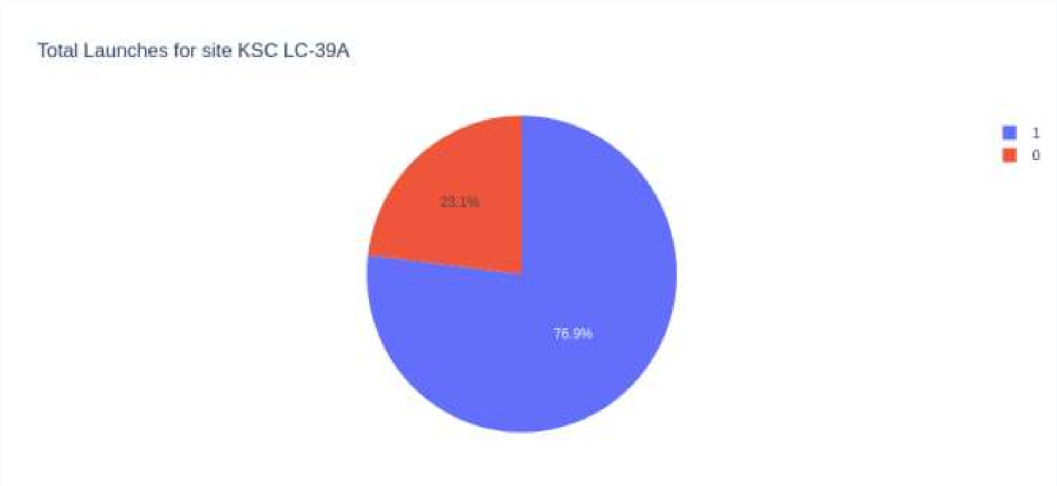
Build a Dashboard with Plotly Dash

Successful Launches by Site



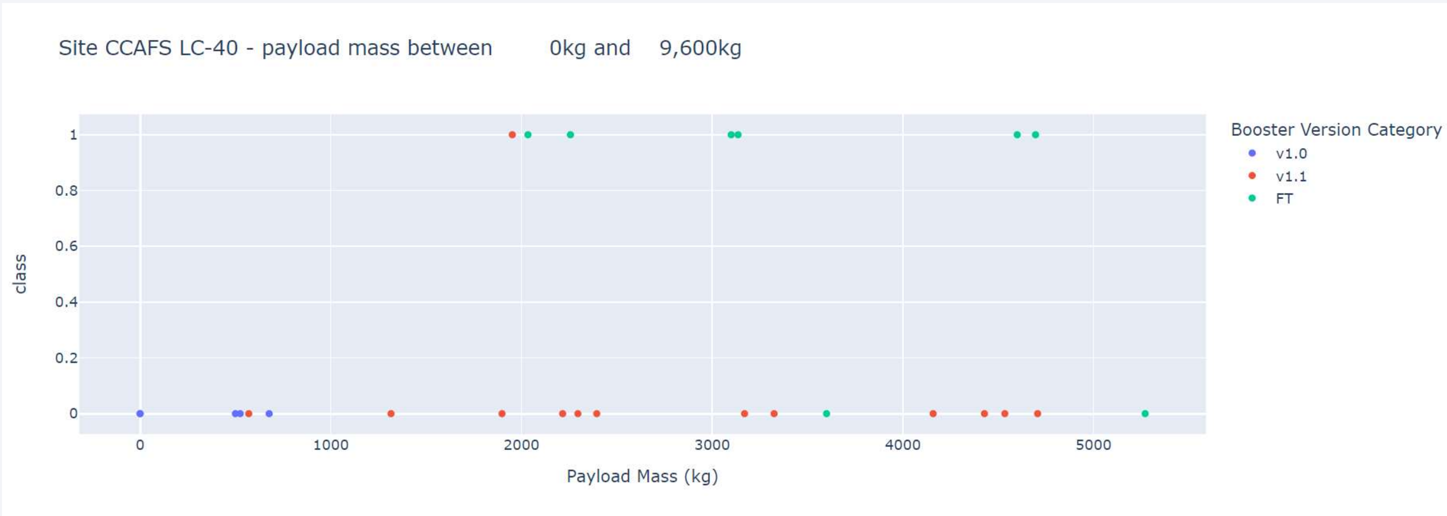
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



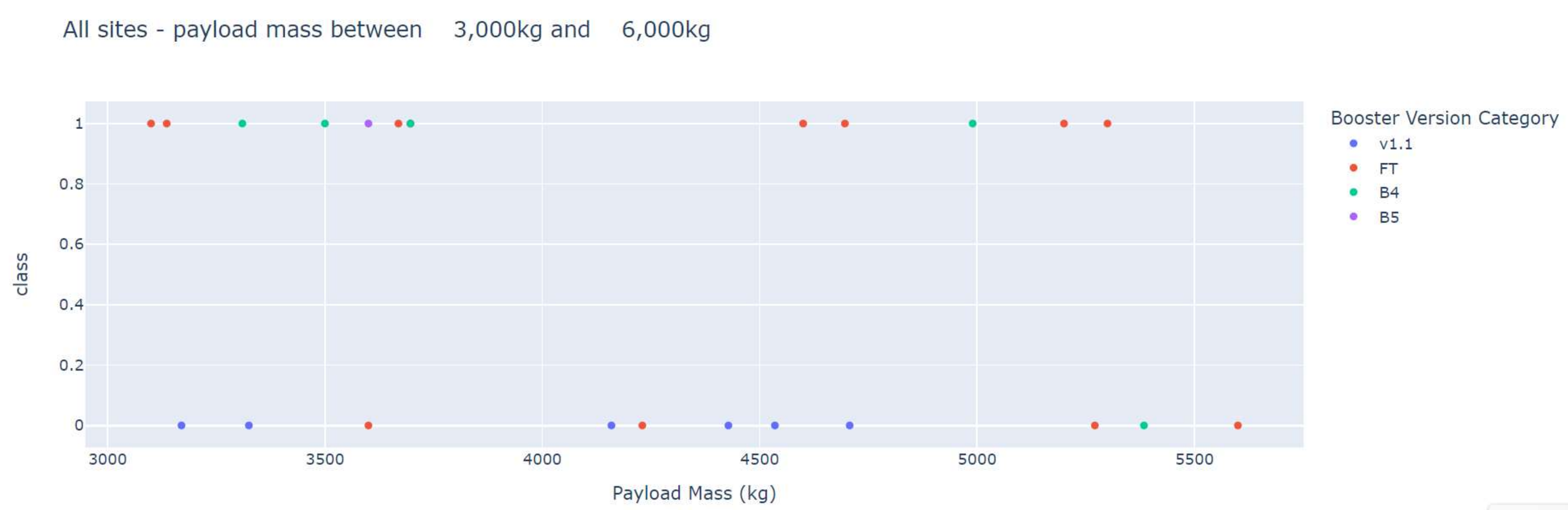
- 76.9% of launches are successful in this site.

Payload vs Launch Outcome



- Payload Mass between 0kg and 9600kg, in this case FT Boosters are successful combinations.

Payload vs Launch Outcome



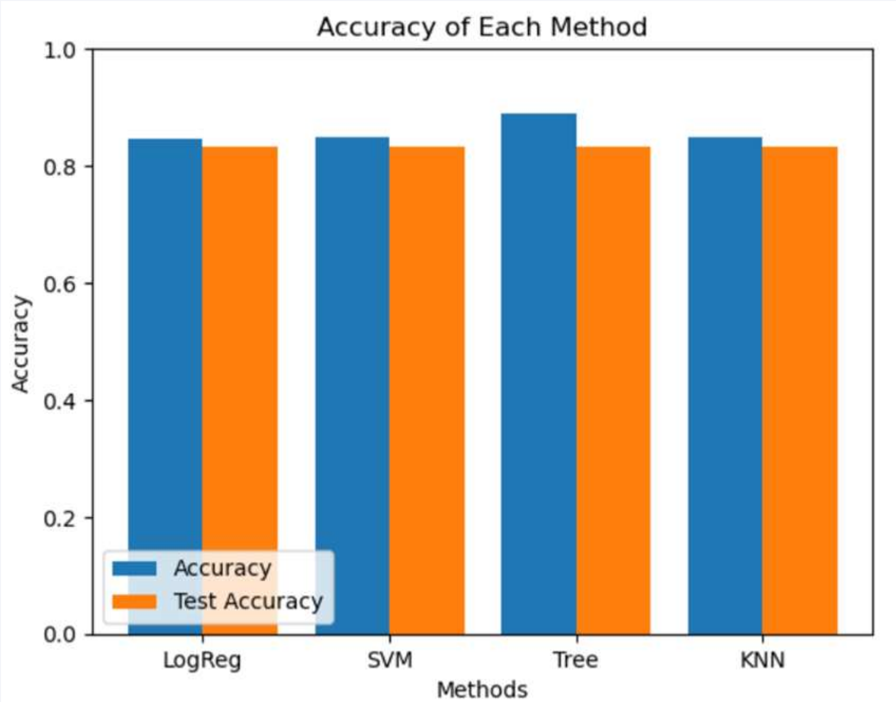
- Payload Mass between 3000kg and 6000kg, in this case also FT Boosters are successful combinations.

Section 5

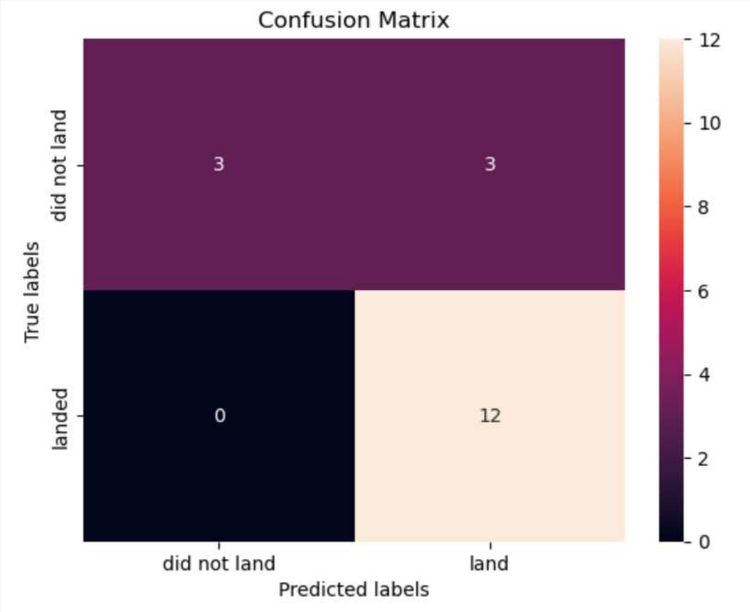
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted.
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 88.9%.



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

- Different data sources were analyzed, refining conclusions along the process.
- The best launch site is KSC LC-39A.
- Launches are analyzed through dashboard as well.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- I have saved the csv file in my directory in order to edit and opened it from the same, since the ‘launch_outcome’ column through the link was not properly defined.


Thank you!