IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Baidal Marco
07/11/2022

Outline

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

Executive Summary

- To analyze data were used these methodologies:
  - Web scraping and the SpaceX API
  - Exploratory Data Analysis (EDA)
  - Machine Learning

- As results:
  - EDA identifies which the best features to predict the level of success in launchings
  - Machine Learning got the best model to predict which characteristics were the most important using the available data

Introduction

- The objective is to know if the new company Space Y could compete against Space X answering:
  - Estimating the total cost of launches by predicting the number of successful landings of the first stage of rockets
  - Where is the best place in a map to make launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was gotten from 2 sources:
  - Space X API
  - WebScraping
- Perform data wrangling
  - The collected data was enriched by creating a landing label based on available data after summarizing and analyzing 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
  - Data collected until this step was normalized and also divided in training and test sets, and after that, evaluated with four different classification models

Data Collection

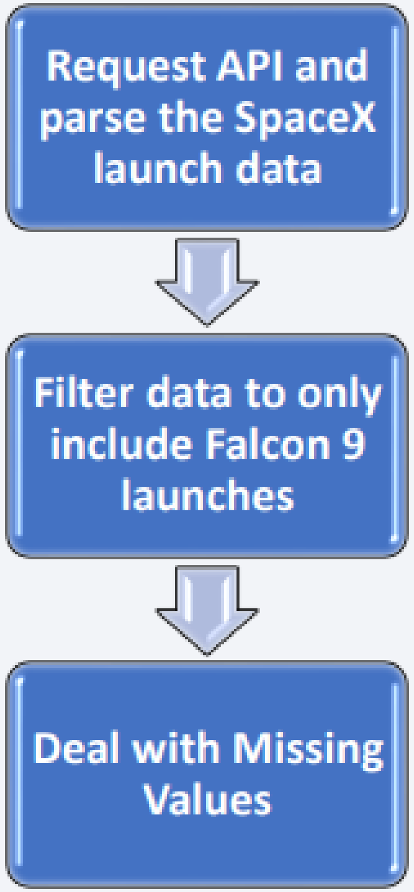
- Data sets were collected from Space X API and from Wikipedia using web scraping technics.

Data Collection – SpaceX API

SpaceX offers a public API from where data can be obtained

Source code:

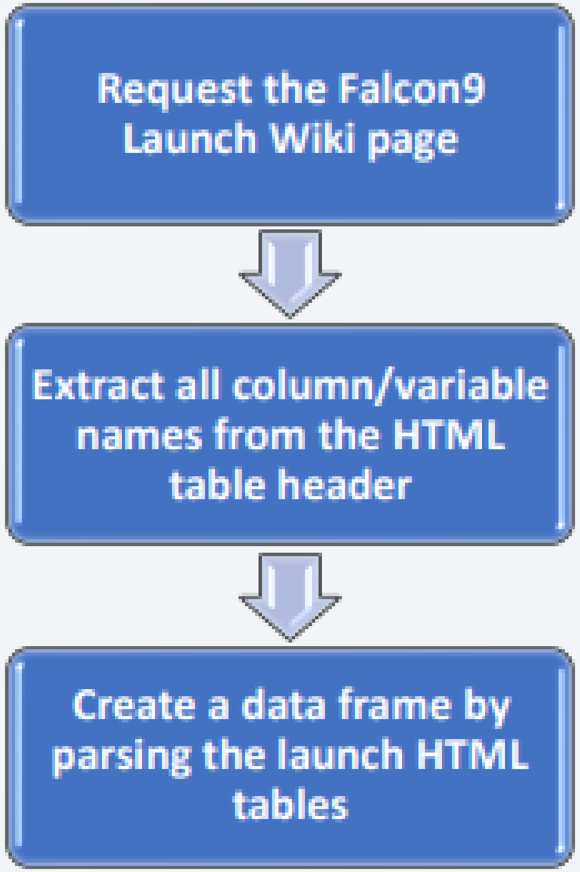
- https://github.com/Wasousky/IBM_Curso/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

- Data from Space X can be obtained from Wikipedia by Webscraping too.

- https://github.com/Wasousky/IBM_Curso/blob/main/Data%20Collection%20Web%20Scraping.ipynb



Data Wrangling

An Exploratory Data Analysis (EDA) is 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.

- https://github.com/Wasousky/IBM_Curso/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

- Payload Mass X Flight Number
- Launch Site X Flight Number
- Launch Site X Payload Mass
- Orbit and Flight Number
- Payload and Orbit

- https://github.com/Wasousky/IBM_Curso/blob/main/EDA%20Data%20Visualization.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;

- https://github.com/Wasousky/IBM_Curso/blob/main/EDA.ipynb

Build an Interactive Map with Folium

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;
- Lines are used to indicate distances between two coordinates.

- https://github.com/Wasousky/IBM_Curso/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data:

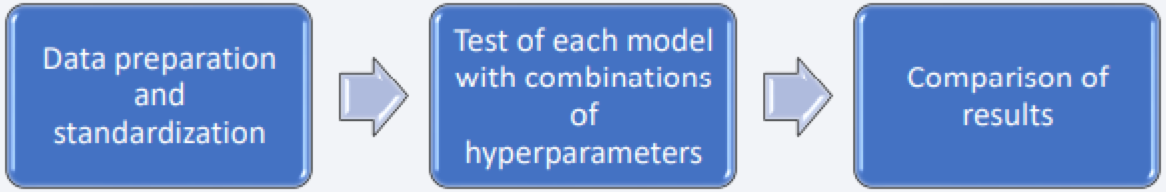
- Percentage of launches by site
- Payload range

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.

Predictive Analysis (Classification)

Four classification models were compared:

- logistic regression
- support vector machine
- decision tree
- k nearest neighbors



- https://github.com/Wasousky/IBM_Curso/blob/main/Machine%20Learning%20Prediction.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 2015 fiver 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
• Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
• The number of landing outcomes became as better as years passed

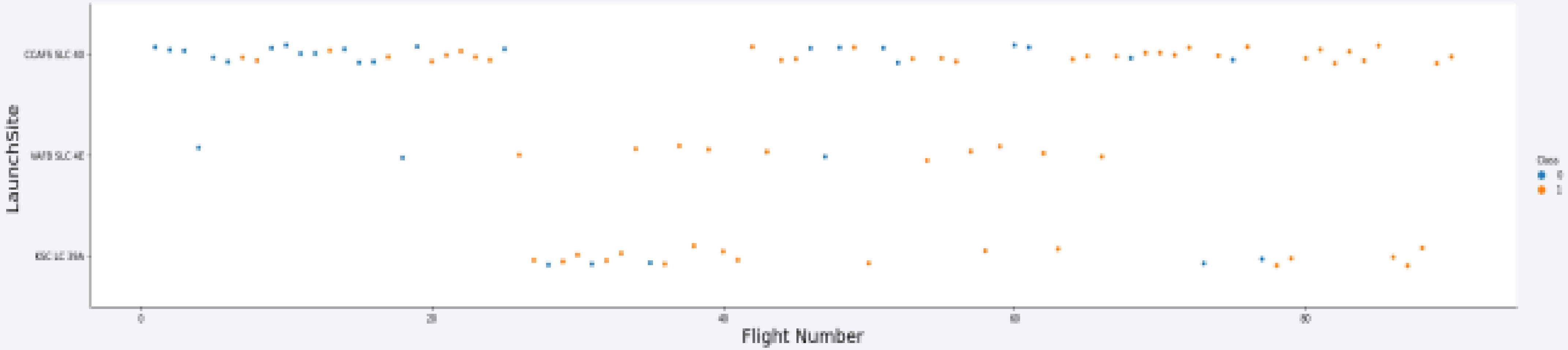
Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.

Most launches happens at east cost launch sites.

Section 2

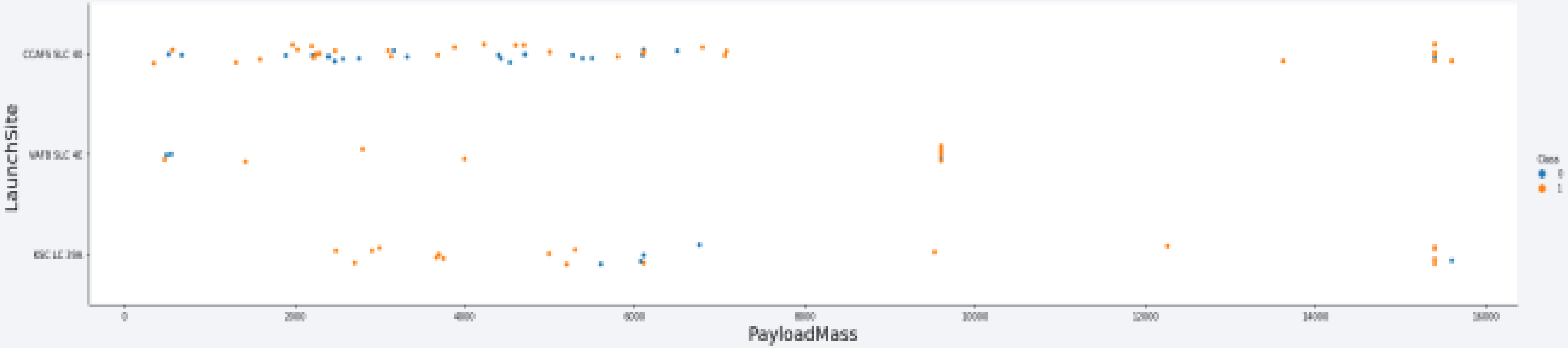
Insights drawn from EDA

Flight Number vs. Launch Site



- 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 third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

Payload vs. Launch Site

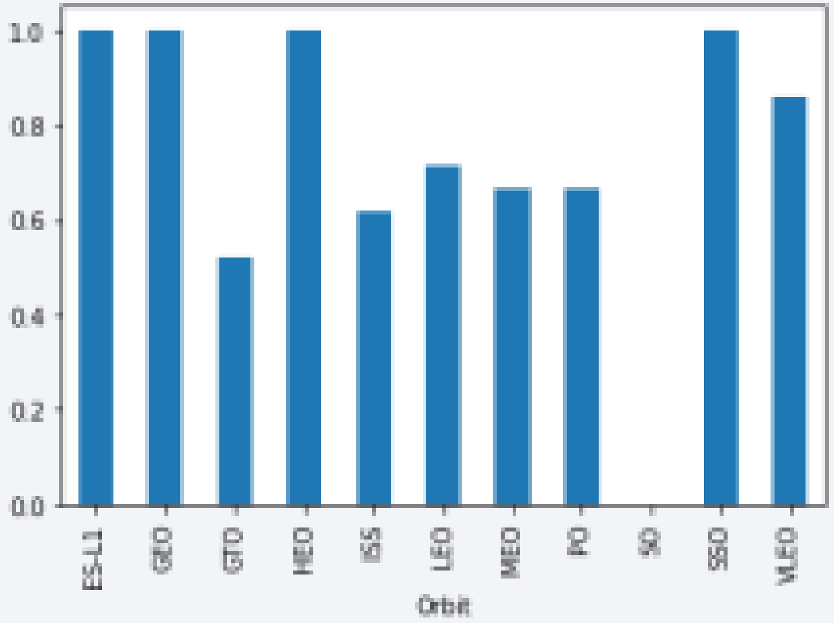


- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO



Followed by:

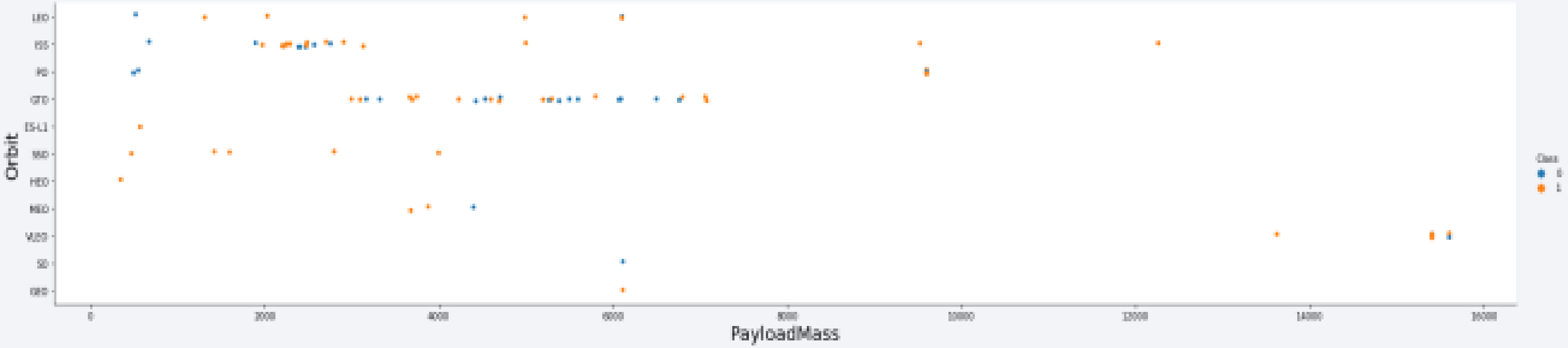
- VLEO (above 80%); and
- LFO (above 70%).

Flight Number vs. Orbit Type

Apparently, success rate improved over time to all orbits.

VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

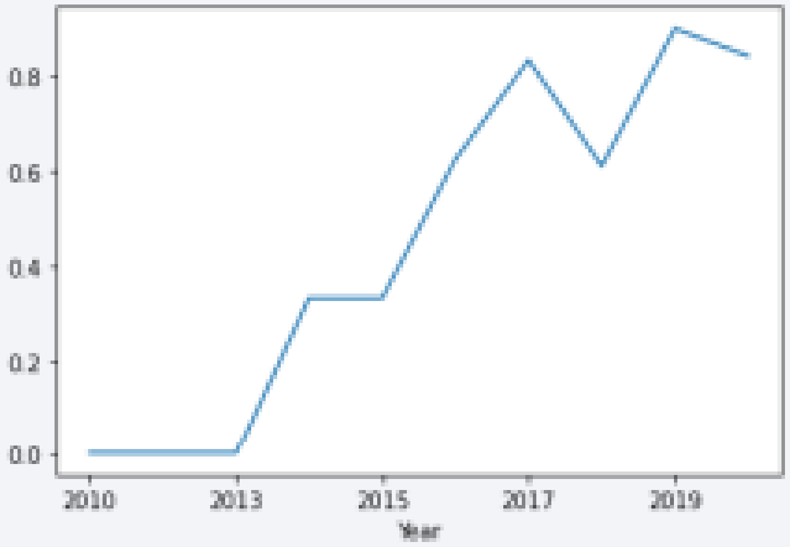
Payload vs. Orbit Type



- Apparently, 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 four launch sites, obtained by selecting unique occurrences of “launch_site” values from the dataset:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

Total Payload Mass

Total Payload (kg)
111.268

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

Average Payload Mass by F9 v1.1

Avg Payload (kg)
2.928

Average payload mass carried by booster version F9 v1.1 by 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

The first successful landing outcome on ground pad 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 12/22/2015

Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 selecting distinct booster versions according to the filters above, these 4 are the result.

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Number of successful and failure mission outcomes grouping mission outcomes and counting records for each group led us to the summary above

Boosters Carried Maximum Payload

Boosters which have carried the maximum payload mass.

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

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

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

The only two failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The rank of 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:

Landing Outcome	Occurrences
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

Section 3

Launch Sites Proximities Analysis

All launch sites

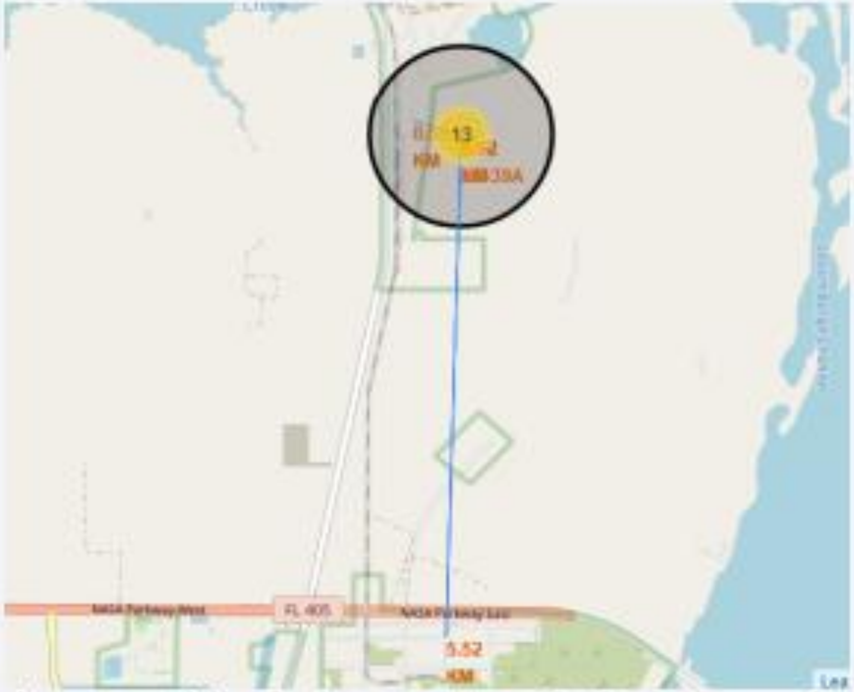
Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes.
- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

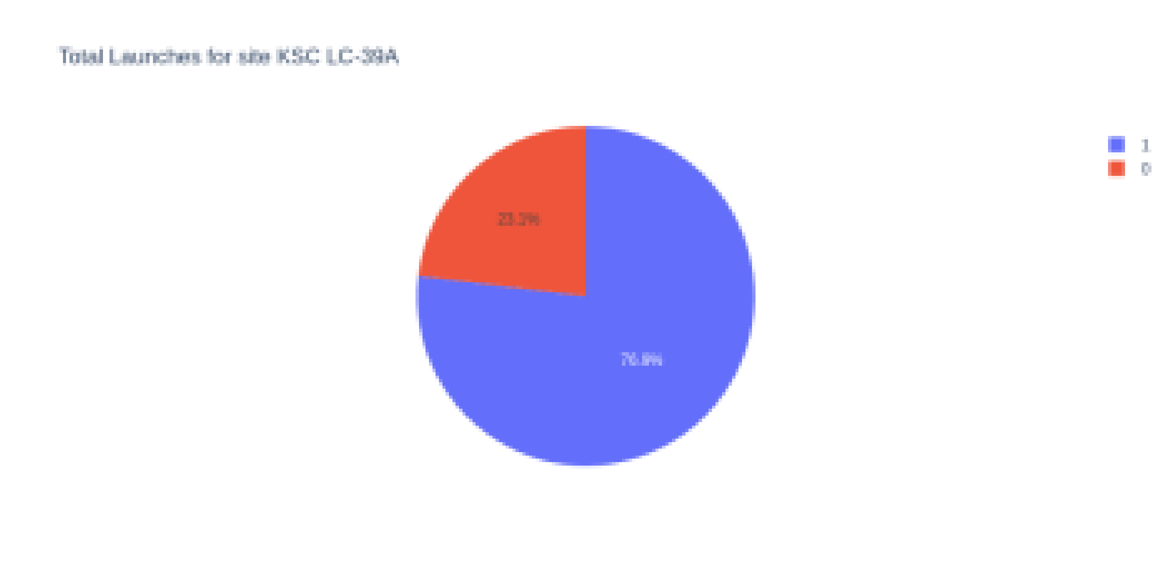
Build a Dashboard with Plotly Dash

Successful Launches by Site

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

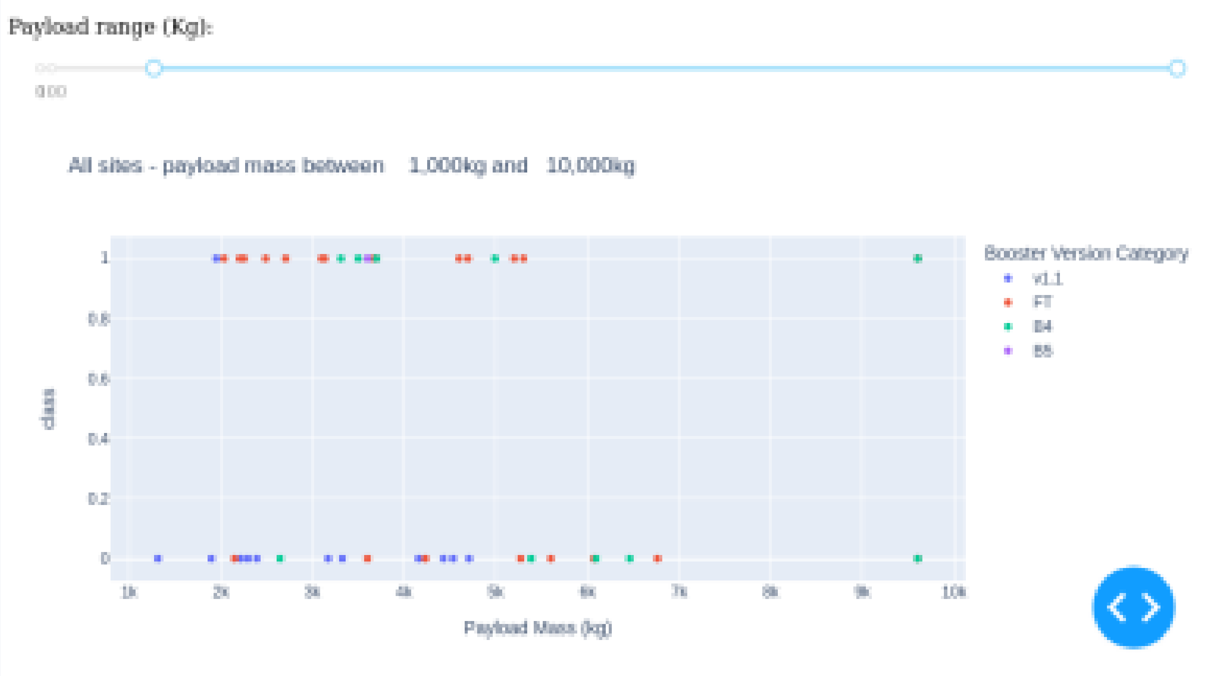
Launch Success Ratio for KSC LC-39A

- The launch site with highest launch success ratio is 76.9%.



Payload vs. Launch Outcome

Payloads under 6,000kg and FT boosters are the most successful combination.

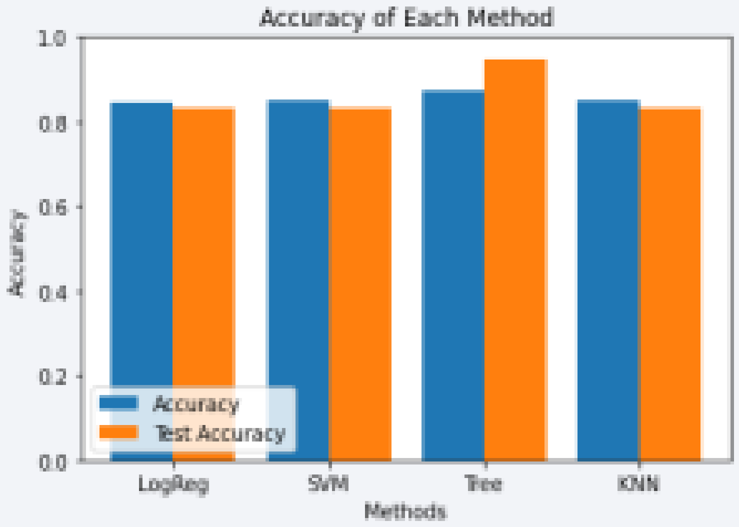


Section 5

Predictive Analysis (Classification)

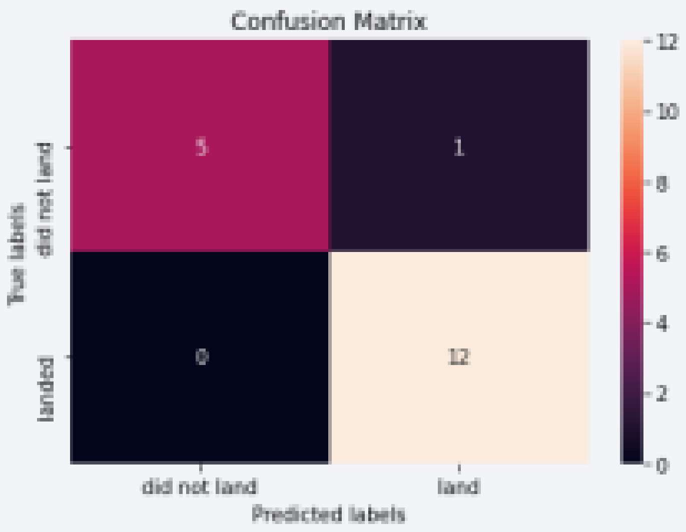
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside.
- 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

Different data sources were analyzed, refining conclusions along the process:

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

As an improvement for model tests, it's important to set a value to "np.random.seed" variable

Thank you!