IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jurijs Bojars
05/23/2023

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), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
  - Find location for 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/)
    - WebScraping (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), using web scraping technics.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used further;
- This API was used according to the flowchart beside and then data is persisted.
- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

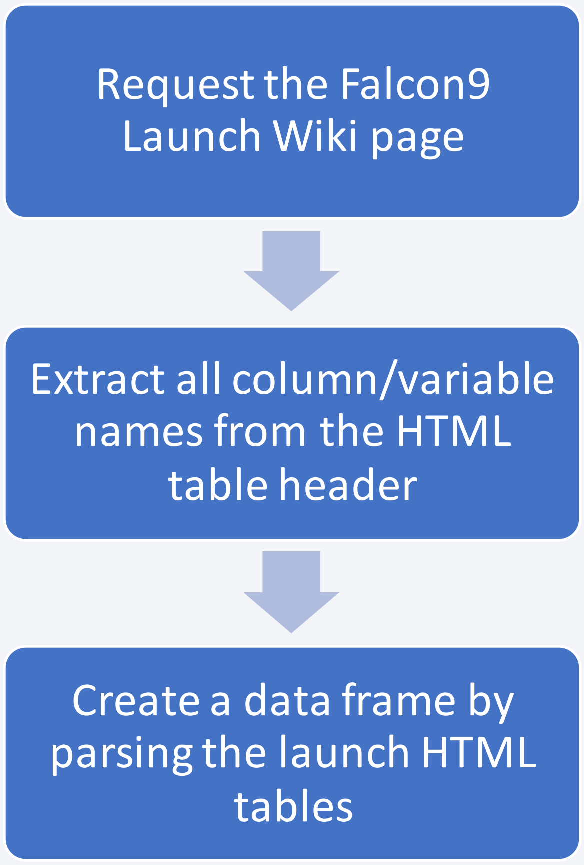
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia

- Data can be downloaded from Wikipedia according to the flowchart and then persist.

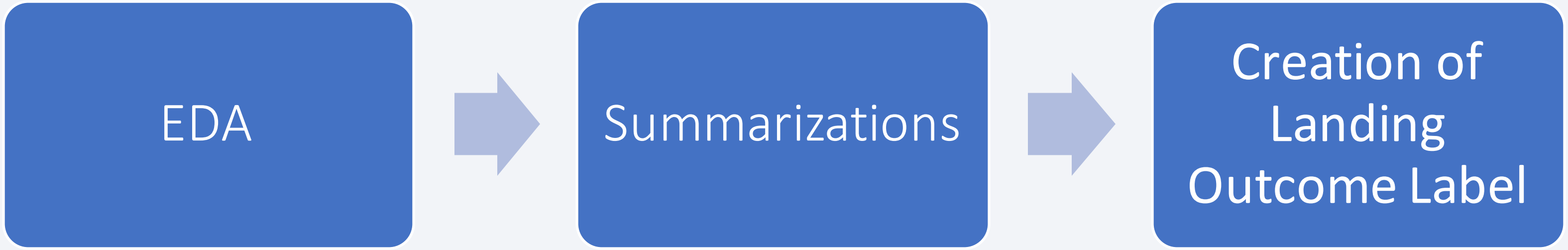
- GitHub URL:

https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.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.

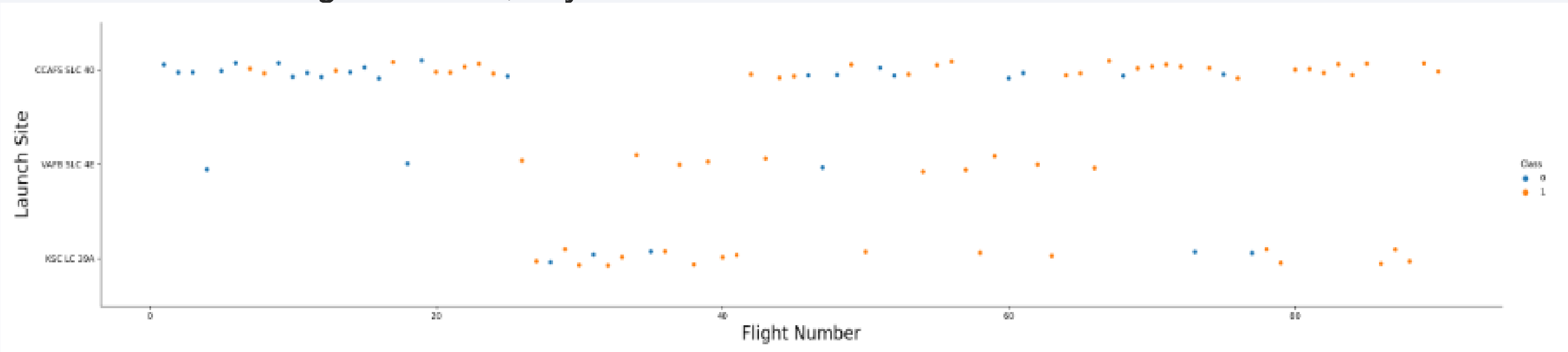


- GitHub url:

https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.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



- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.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 2010-06-04 and 2017-03-20.
- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

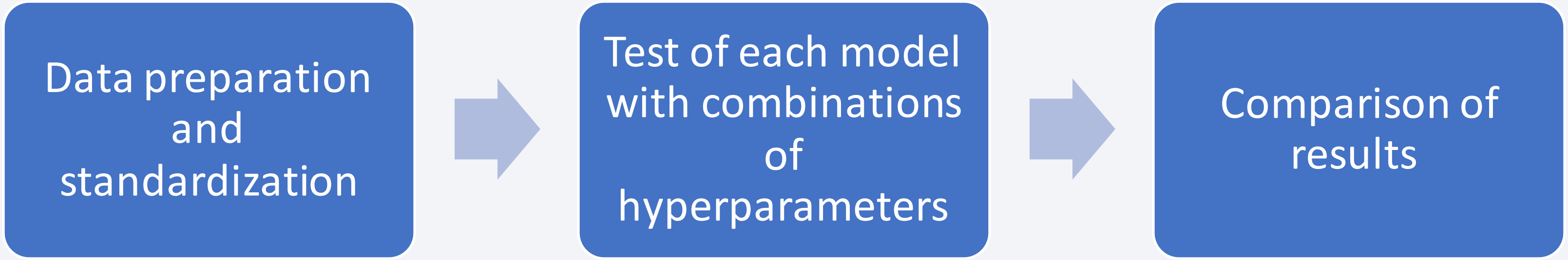
- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.
- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.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.
- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/spacex_dash.app.py

Predictive Analysis (Classification)

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



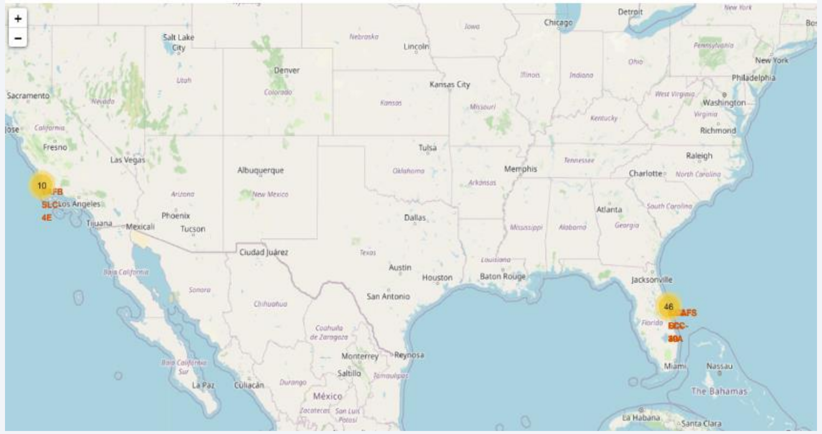
- GitHub URL: https://github.com/JurijsBojars/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

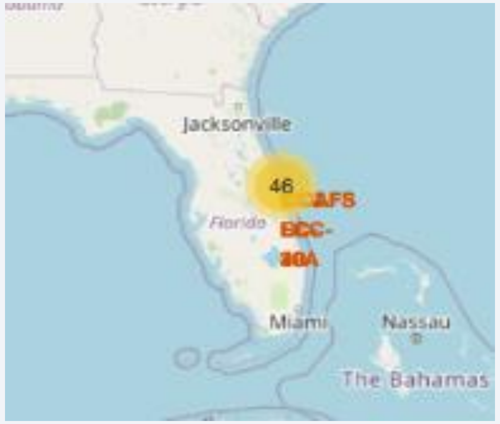
Results

- 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.
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

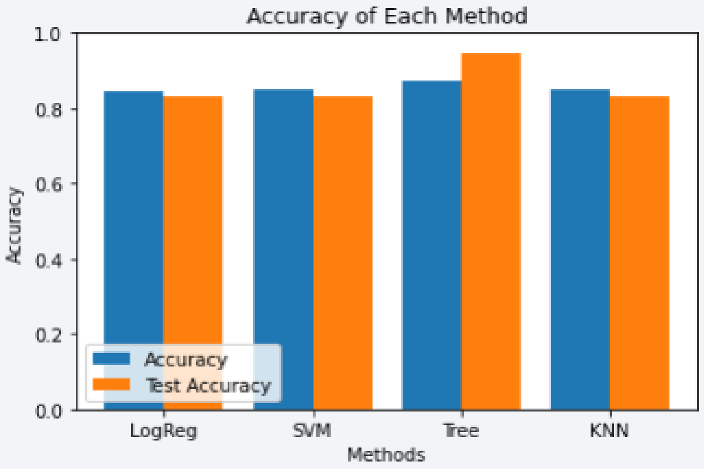
- Interactive analysis results and demo:
  - It was possible to identify that launch sites are located closest to the equator and have a good logistic infrastructure around.
  - Most launches are made at east coast launch sites.





Results

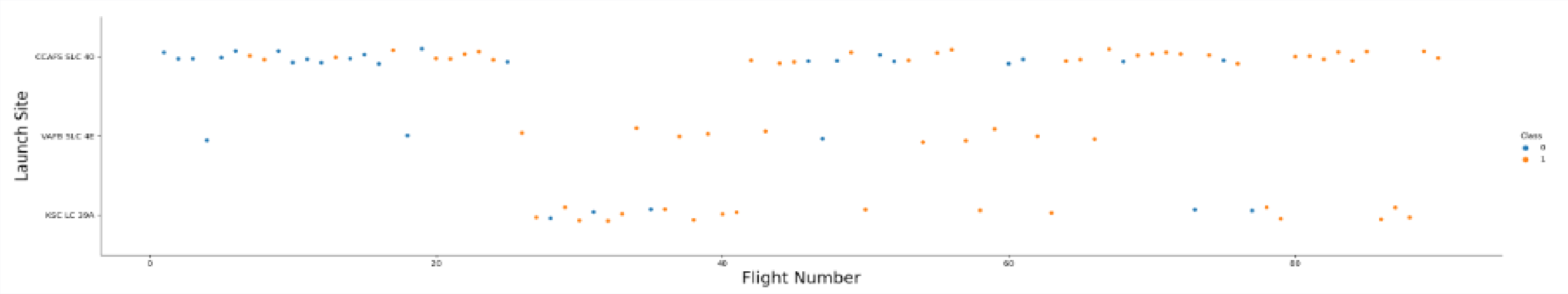
- Predictive analysis results:
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



Section 2

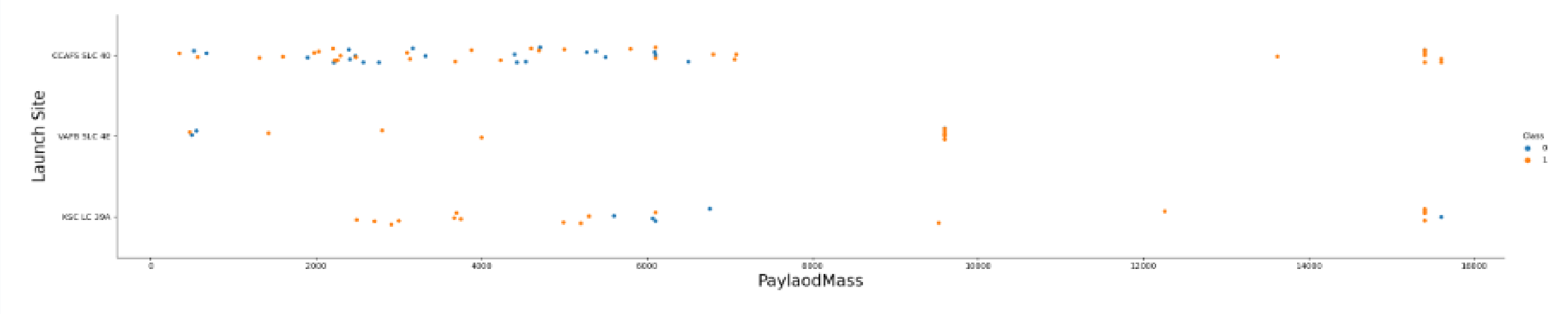
Insights drawn from EDA

Flight Number vs. Launch Site



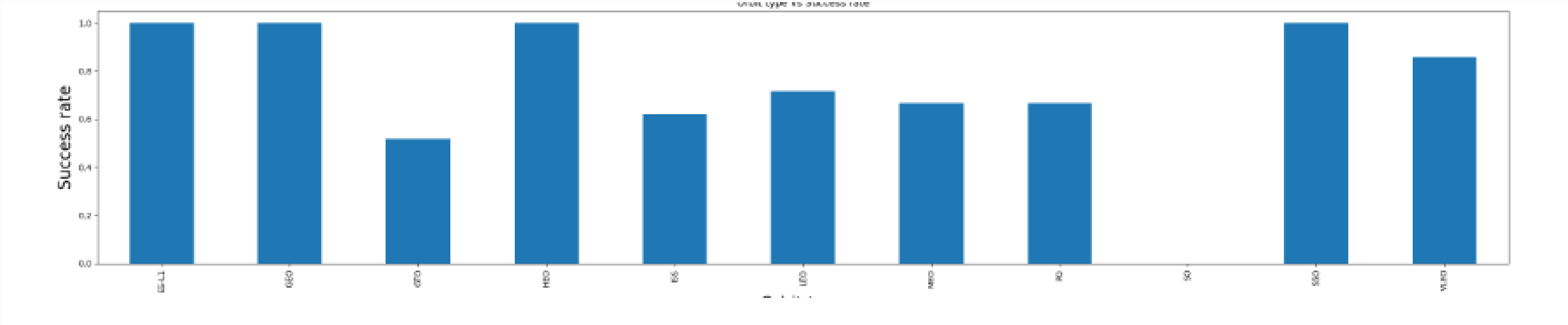
- The general success rate improves over time
- CCAF5 SLC 40 is used the most and it success rate has significantly improved over time.
- VAFB SLC 4E site is used rarely, but it has pretty good success rate and it success rate also increase over time.

Payload vs. Launch Site



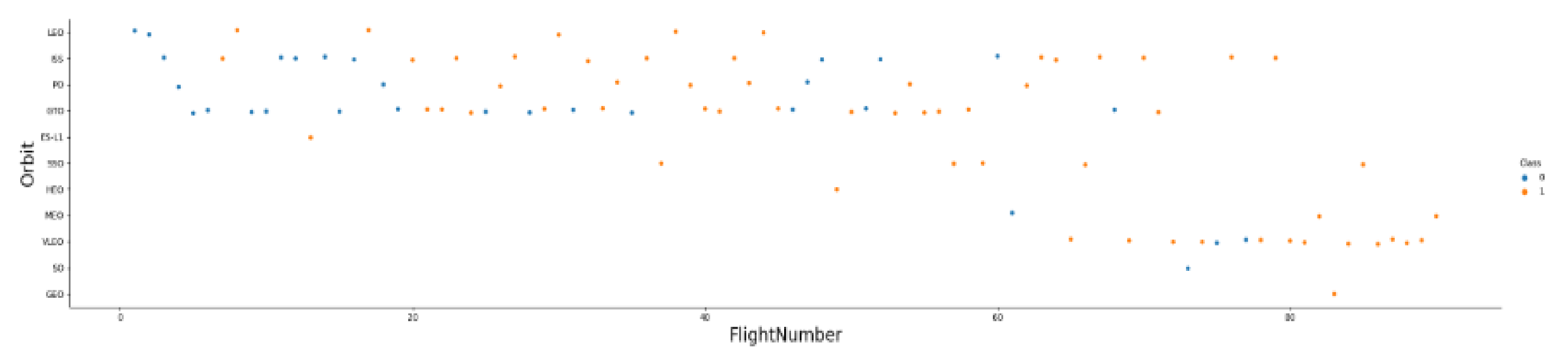
- As payload mass increases success rate also goes up;
- Most heavy launches are made only on two launch sites: CCAF5 SLC 40 and KSC LC 39A;

Success Rate vs. Orbit Type



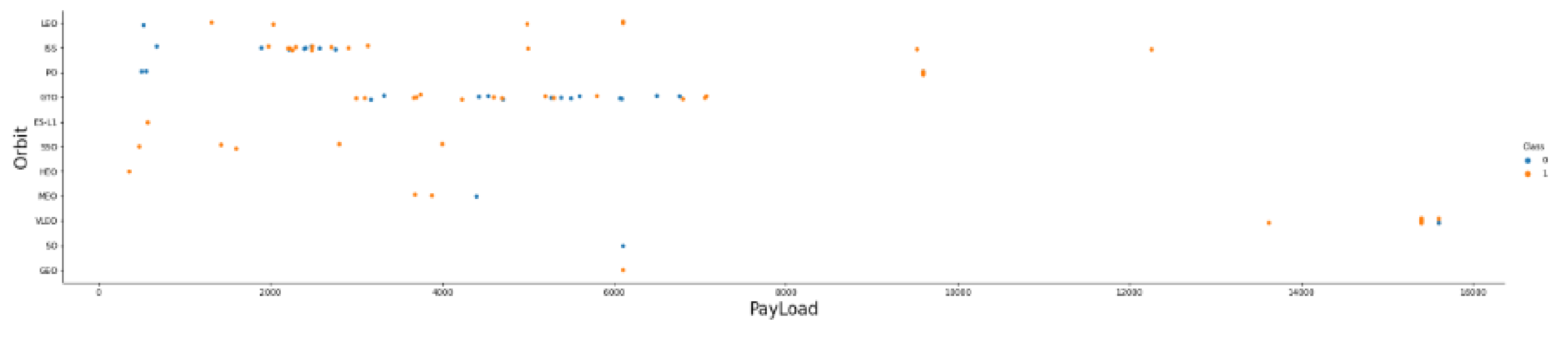
- 4 orbit types has best and equal success rate (ES-L1, GEO, HEO and SSO);
- One orbit type has 0 success rate;

Flight Number vs. Orbit Type



- The LEO orbit the Success appears related to the number of flights;
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Only few launches are made to GEO, SO and ES-11, MEO, HEO orbits. So their success rate should be dropped, then we are comparing orbit type and success rate.

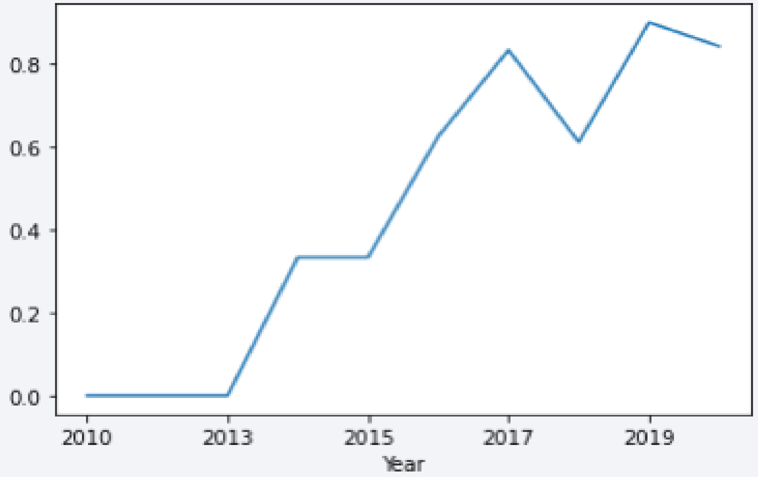
Payload vs. Orbit Type



- All launches heavier then 13000 KG are made only to VLEO orbit;
- All launches to SSO orbit are successful;
- Most launches are made to GTO and ISS orbit;
- There is no relation between payload and success rate to orbit GTO;
- The more heavier launches are has higher success rate for ISS orbit.

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:

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

- They are obtained by selecting unique occurrences of “launch_site” value from the dataset.

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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	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	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	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

SUM(PAYLOAD_MASS__KG_)
45596.0

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

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
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

- Date of the first successful landing outcome on ground pad:

MIN(DATE)
01/08/2018

- 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 01/08/2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- Total number of successful and failure mission outcomes

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

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

Boosters Carried Maximum Payload

- Names of the booster 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 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

- 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

- There are only 2 failures at this particular date.

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

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

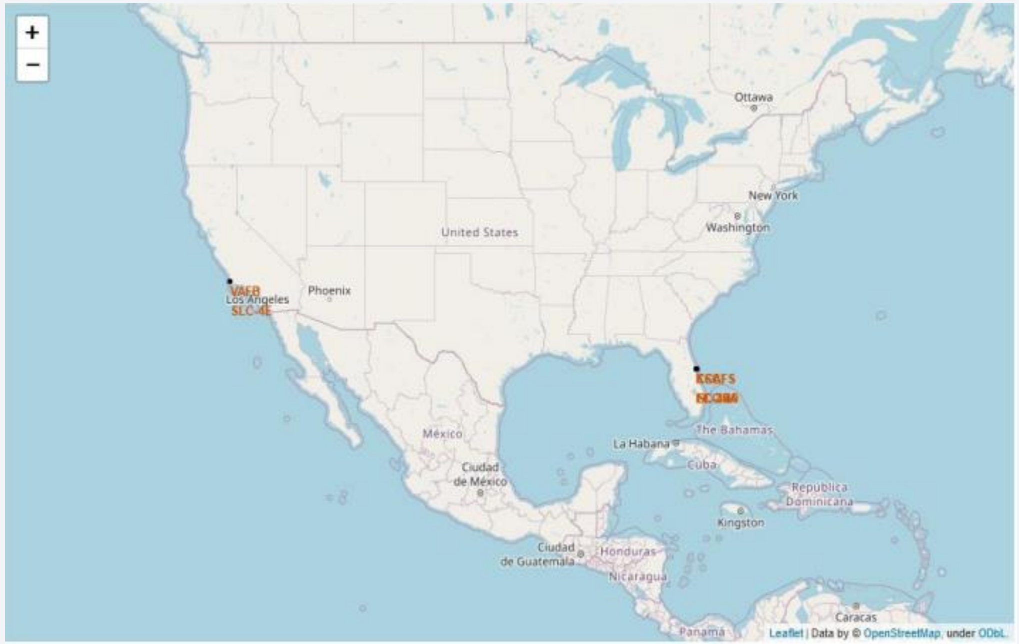
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 3

Launch Sites Proximities Analysis

USA launch sites



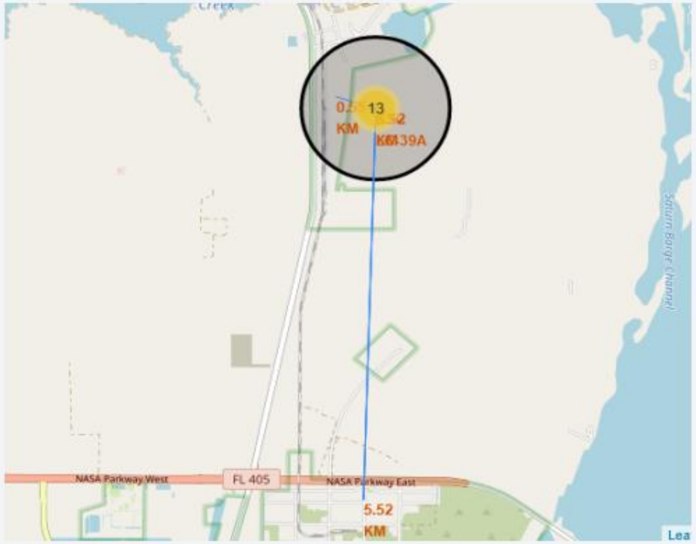
- All launch sites are located near coast line and are placed as close as it possible to equator line.

Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

- Green markers indicates successful launches while red ones indicate failure launches.

Logistics and Safety

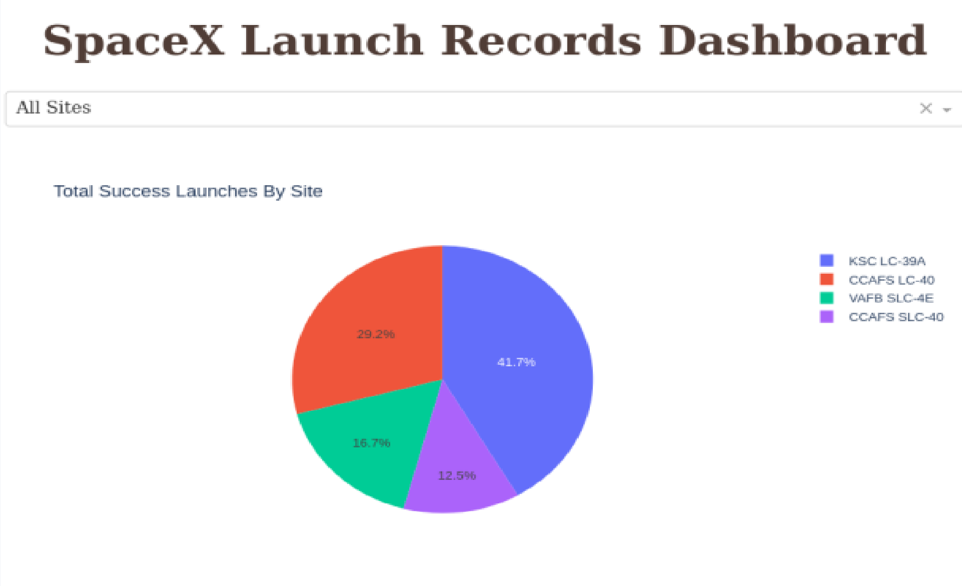


- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 4

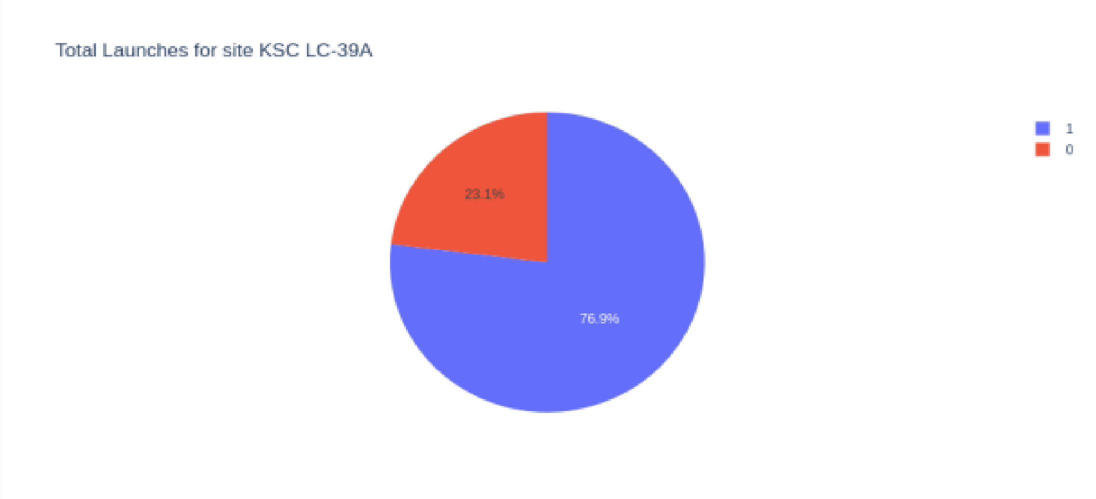
Build a Dashboard with Plotly Dash

Successful Launches by Launch Site



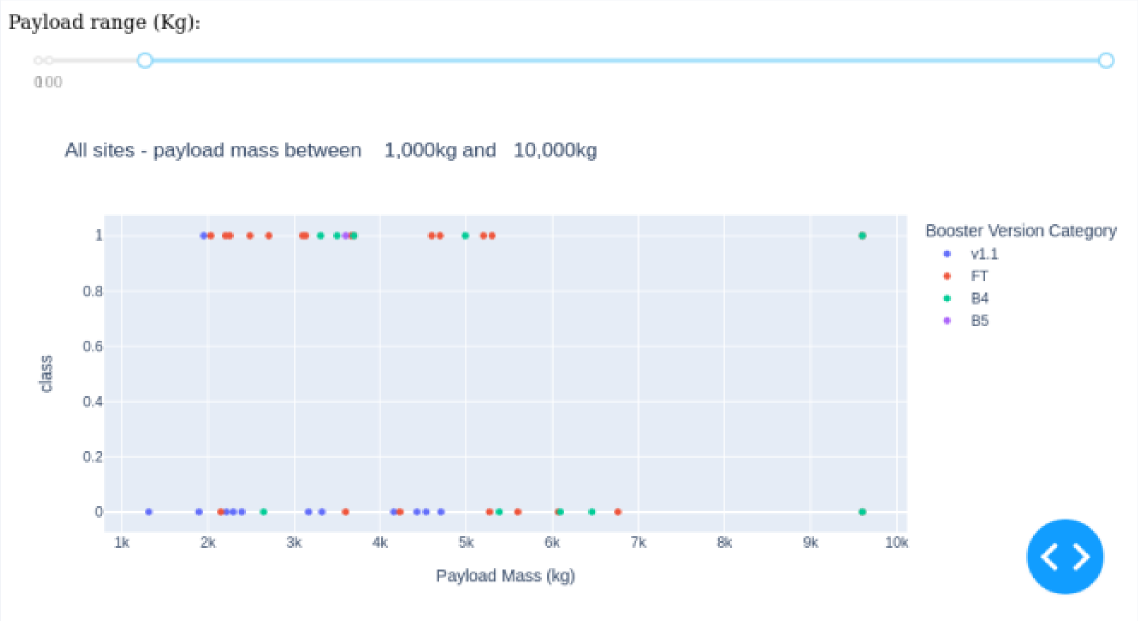
- Biggest part of successful launches are done from KSC LC-39A launch site.
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



- 76.9% of launches are successful in this particular launch site.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.
- Payloads under 6,000kg and v1.1 boosters are the least successful combination.

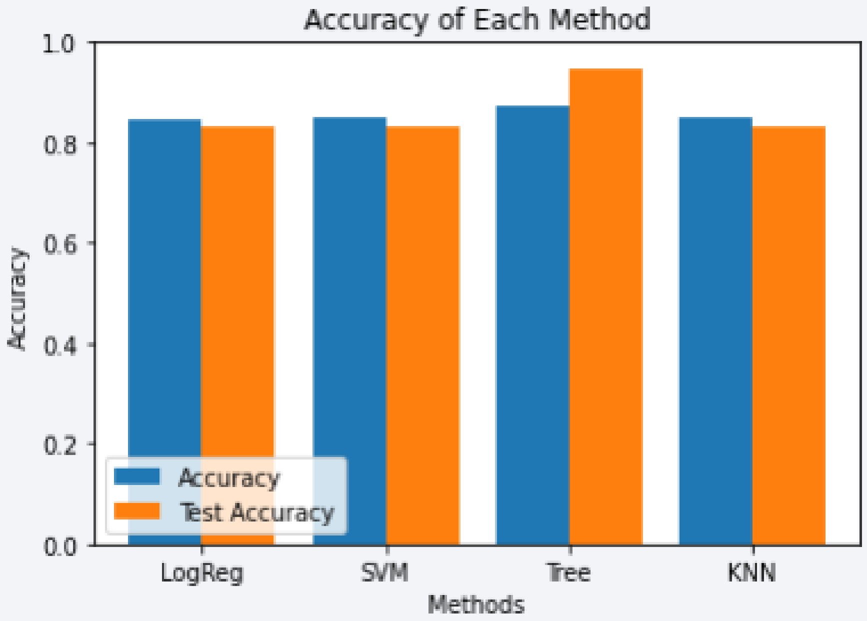
Section 5

Predictive Analysis (Classification)

Classification Accuracy

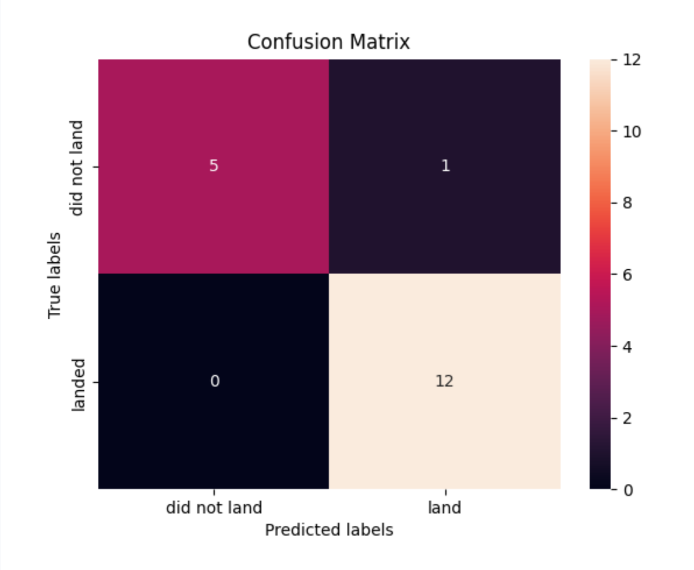
- Four classification models were tested, and their accuracies are plotted beside;

- 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;
- Payloads under 6,000kg and FT boosters are the most successful combination;
- Payloads under 6,000kg and v1.1 boosters are the least successful combination;
- All launch sites are placed near coast line and are located as close as possible to equator line;
- All launch sites are situated near railroad infrastructure;
- 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 rate of launch success.

Appendix

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

Thank you!