IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ayman Hassan
04-09-2022

Outline

Executive Summary

- The following Methodologies were used to analyze the data
  - Data collection using web scrapping and SpaceX api
  - Exploratory data analysis (EDA) includes data wrangling, Data visualization, and interactive Dashboard.
  - Machine learning prediction.
- Summary of all results
  - It was possible to collect valuable data from SpaceX api and Wikipedia pages
  - EDA allowed to find interesting relations of the data as well as helped choose the best ML model

Introduction

- Project background and context
  - The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- Problems you want to find answers
  - Estimate the best place to launch
  - Choose the best rocket model to launch into space
  - Estimate the optimal mass weight each rock should carry for the a successful launch

Section 1

Methodology

Executive Summary

- Data collected from spaceX in two ways:
  - SpaceX REST api (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping from Wikipedia (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

Executive Summary

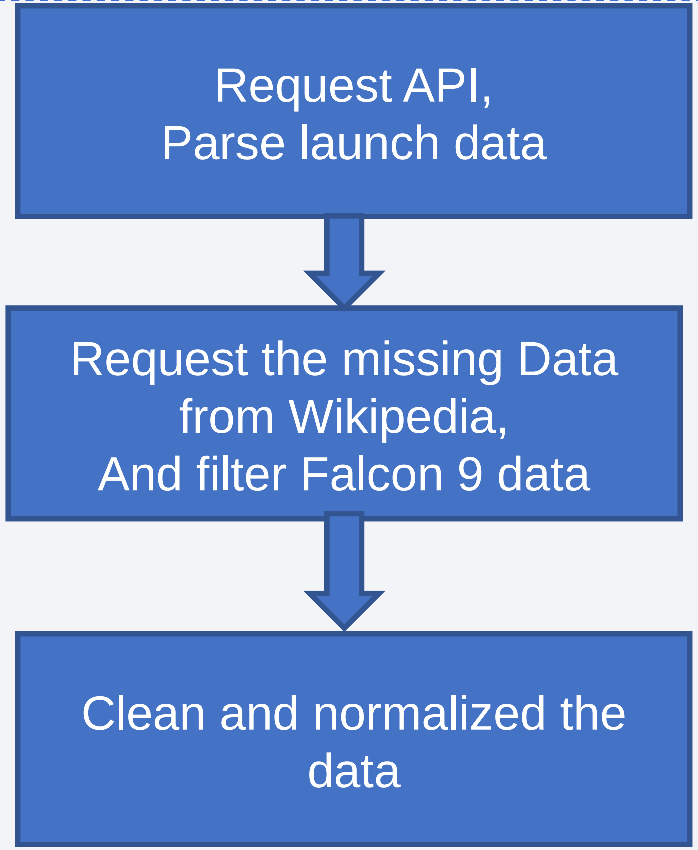
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The collected Data were normalized, cleaned, divided into train and test set, then evaluated using 4 different ML models and the best model was being chosen

Data Collection

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

Data Collection – SpaceX API

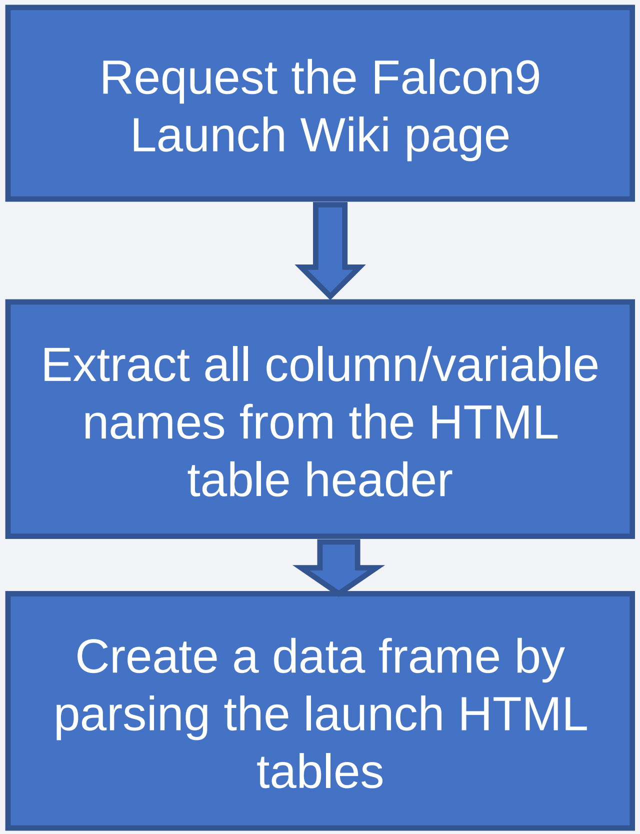
- SpaceX offers a public API from where data can be obtained and then used.
- Add the GitHub URL of the completed SpaceX API calls notebook
- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb)



Data Collection - Scraping

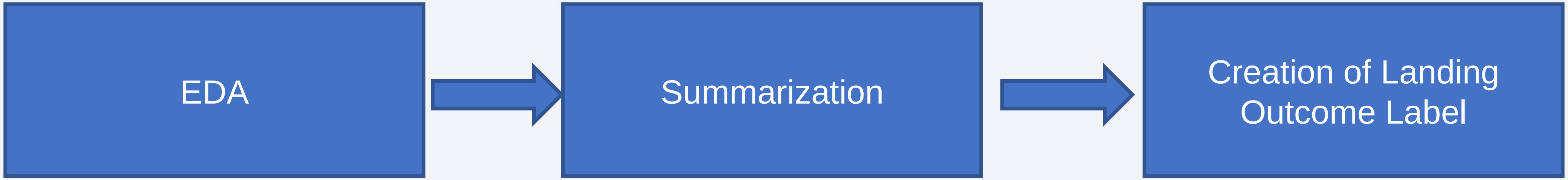
- Missing data are downloaded from Wikipedia according to the flowchart and then persisted.

- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/cc152f7d46c6d2def98795a56a11e2ed479f11c2/labs-jupyter-spacex-Data%20wrangling.ipynb)



Data Wrangling

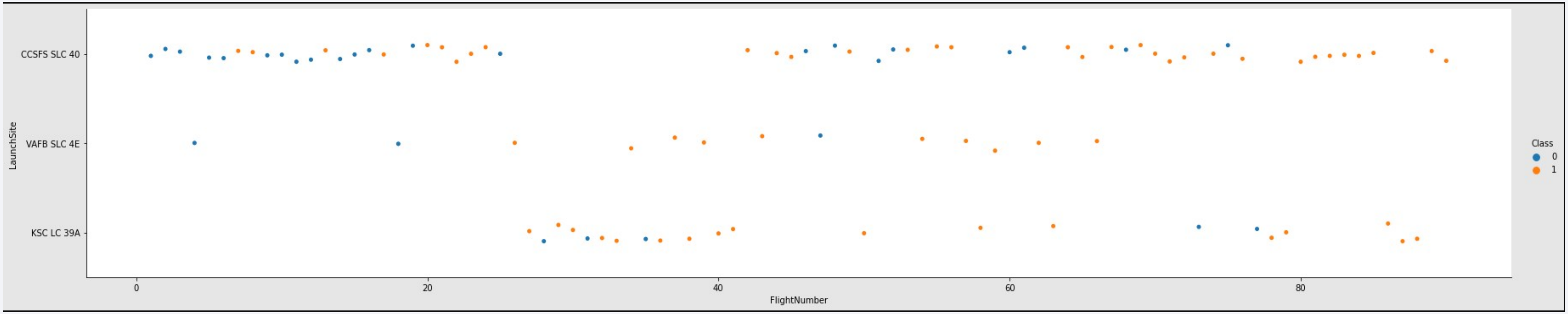
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/cc152f7d46c6d2def98795a56a11e2ed479f11c2/labs-jupyter-spacex-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



- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.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.

- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/cc152f7d46c6d2def98795a56a11e2ed479f11c2/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.

- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/cc152f7d46c6d2def98795a56a11e2ed479f11c2/lab_jupyter_launch_site_location.ipynb)

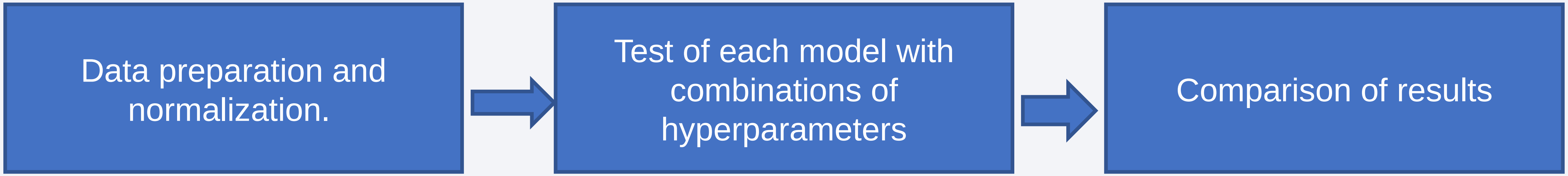
Build a Dashboard with Plotly Dash

- 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

- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/b757ea0bcbc8e5e2f73631ab54ad3b8bb6ebbb49/spacex_dash_app.py)

Predictive Analysis (Classification)

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



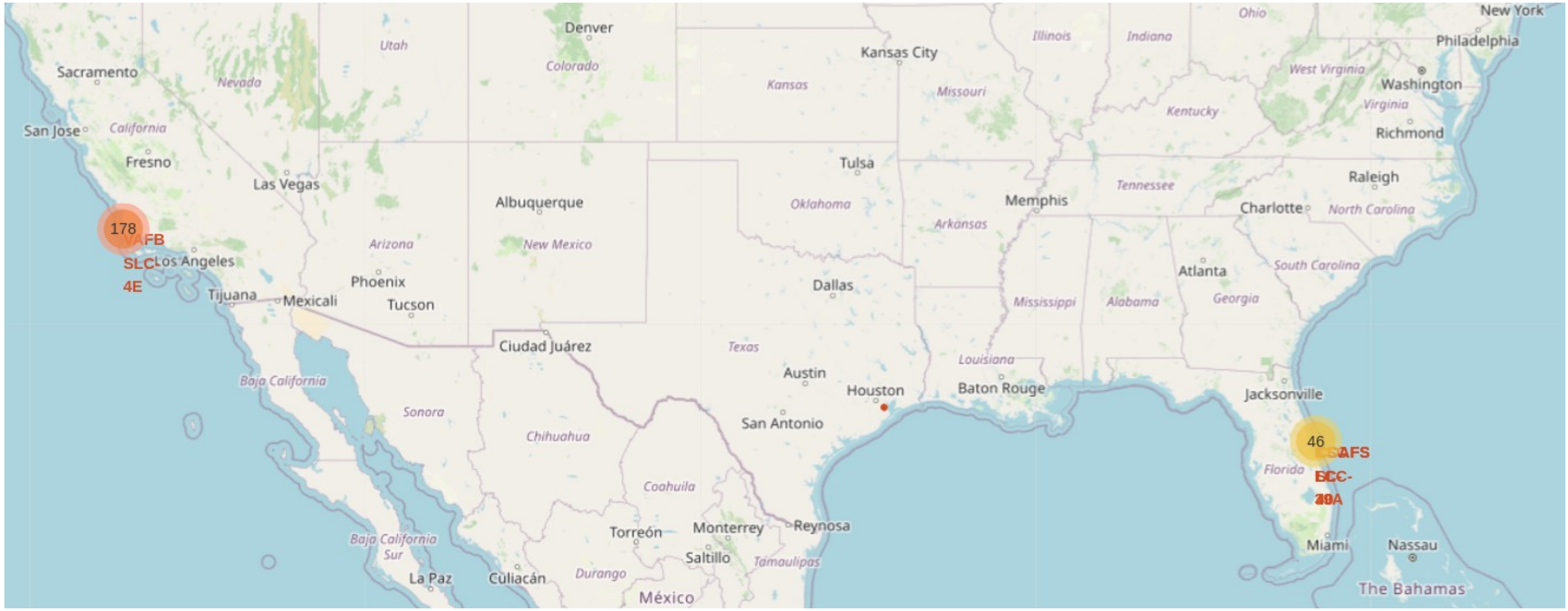
- Source code (https://github.com/aymanred121/IBM-Applied-Data-Science-Capstone/blob/b757ea0bcbc8e5e2f73631ab54ad3b8bb6ebbb49/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

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.

Results

- 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.



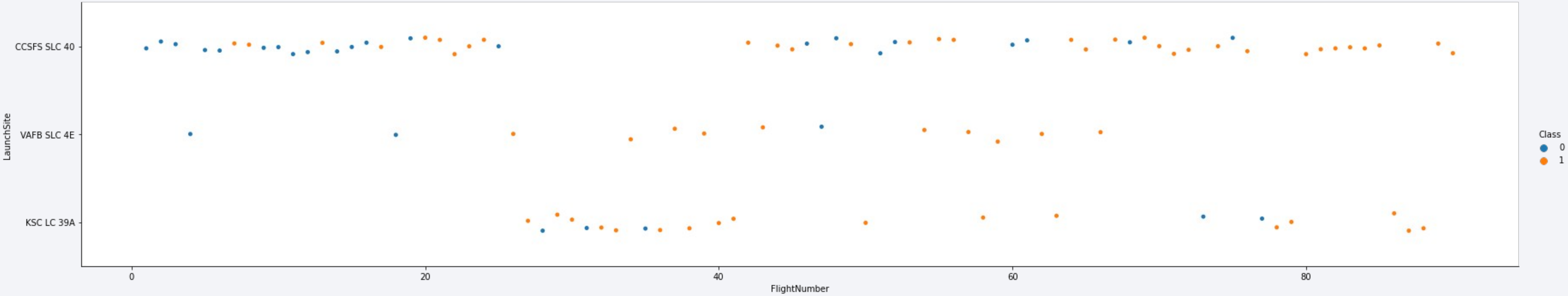
Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 84%

Section 2

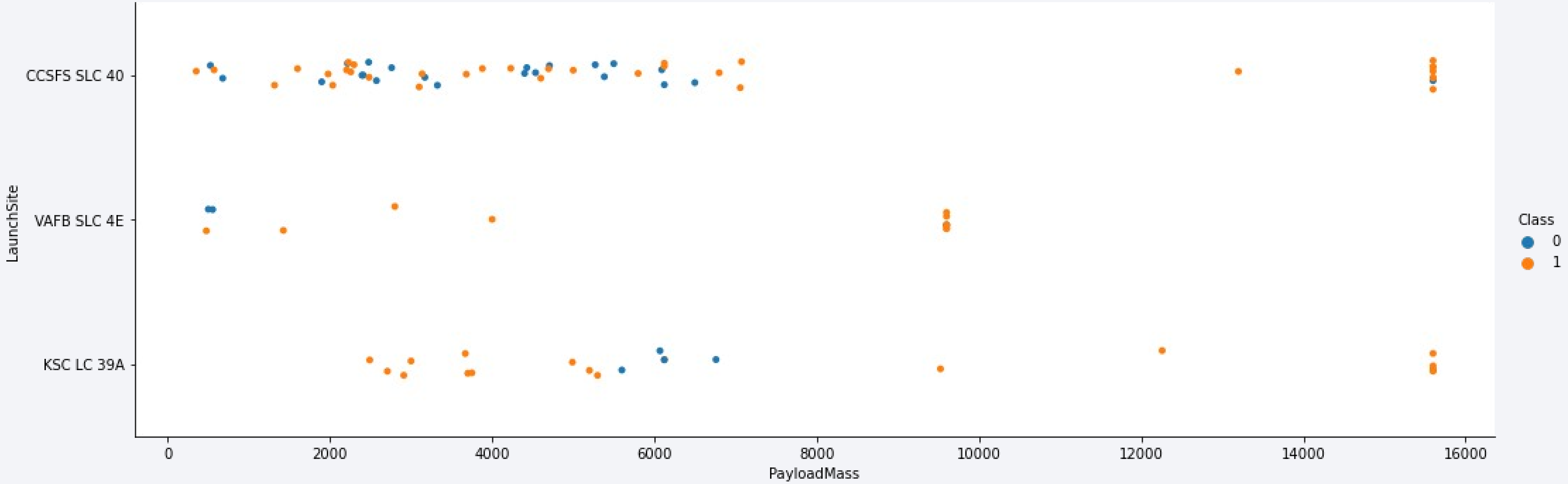
Insights drawn from EDA

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.

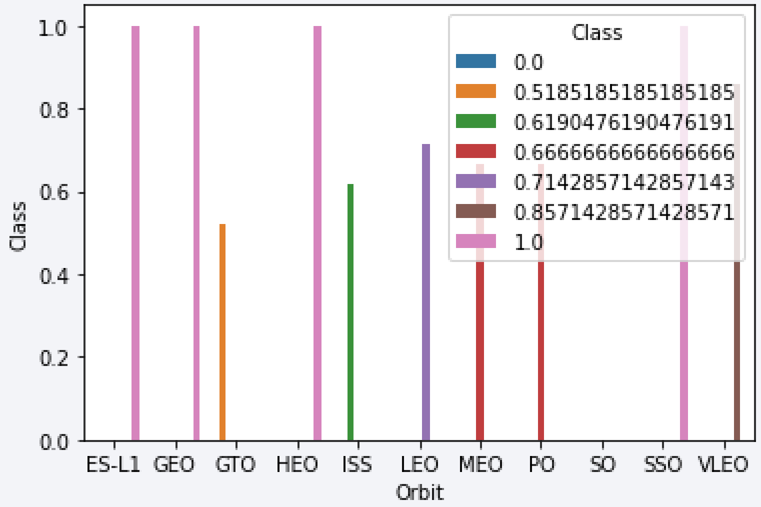
Payload vs. Launch Site



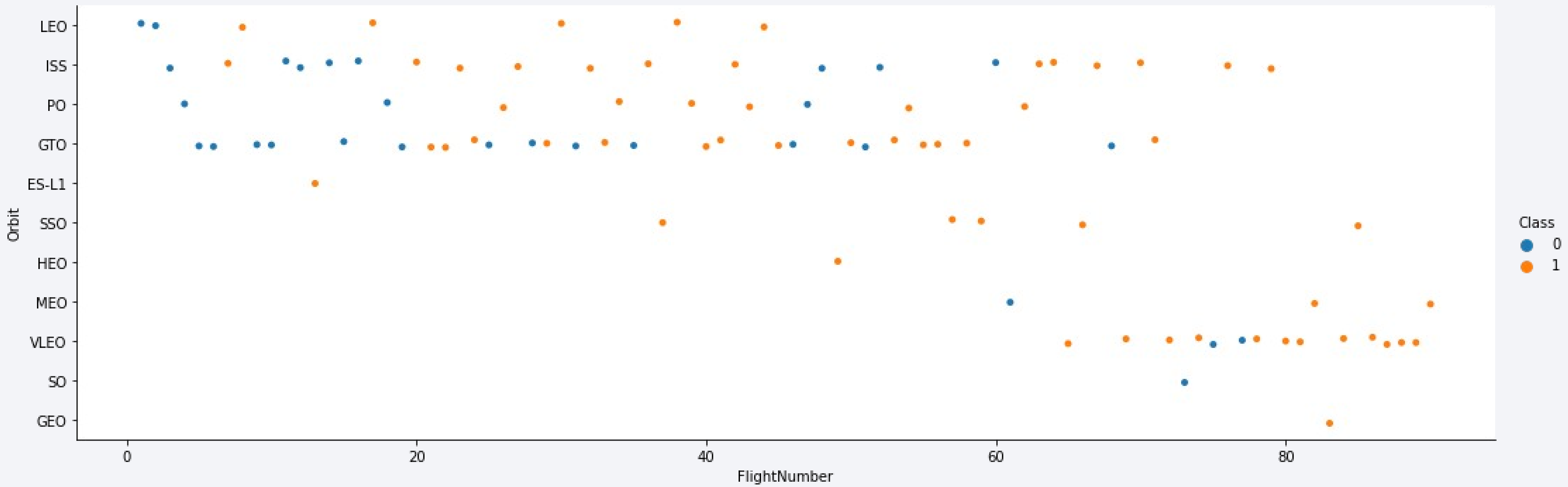
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO.
- Followed by:
  - VLEO (above 80%);
  - LFO (above 70%)

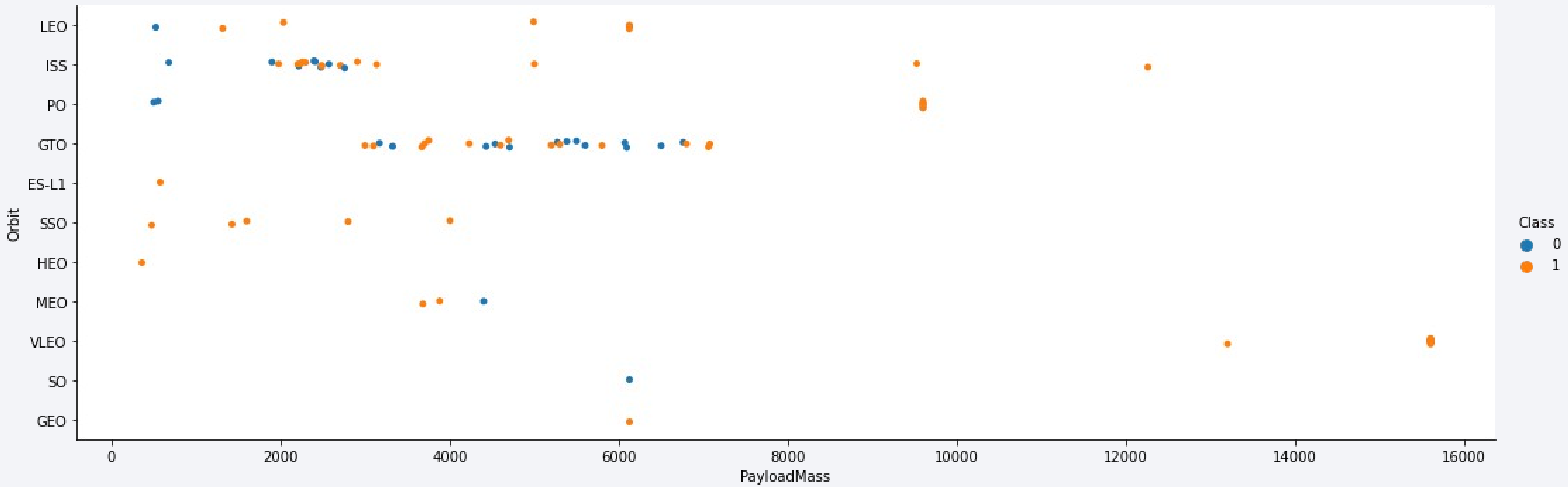


Flight Number vs. Orbit Type



- success rate improved over time to all orbits;
- VLEO seems to be the better spot for now since it has the higher flights

Payload vs. Orbit Type

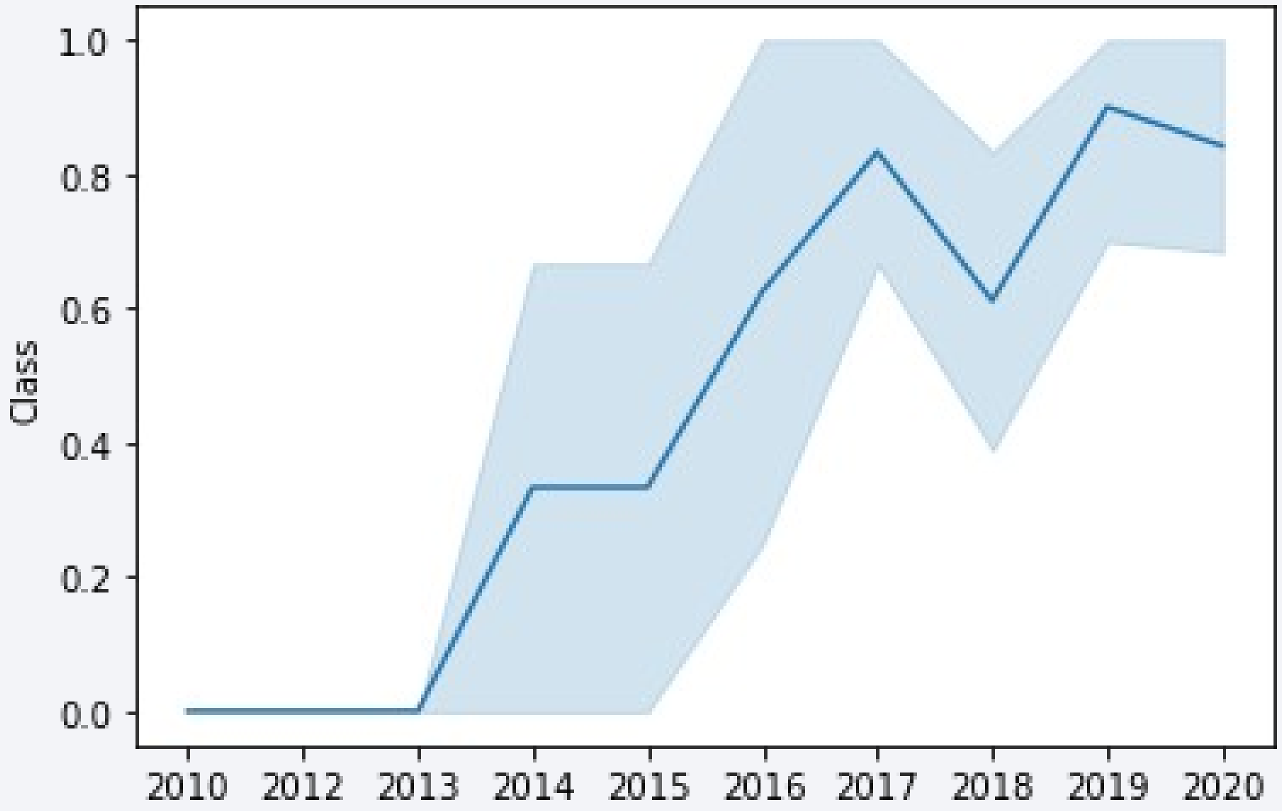


Again it seems like VLEO is the best orbit even on higher masses

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

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

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

Launch Site Names 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	No attempt

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

Nasa_Payload
107010

Total payload calculated above, by summing all payloads that belongs to NASA.

Average Payload Mass by F9 v1.1

avg_F9_v11
2534.6666666666665

Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2534.66kg.

First Successful Ground Landing Date

Date
22-12-2015

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.

Successful Drone Ship Landing with Payload between 4000 and 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

Mission_Outcome	Occurrences
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

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

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

The list above has the only two occurrences.

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

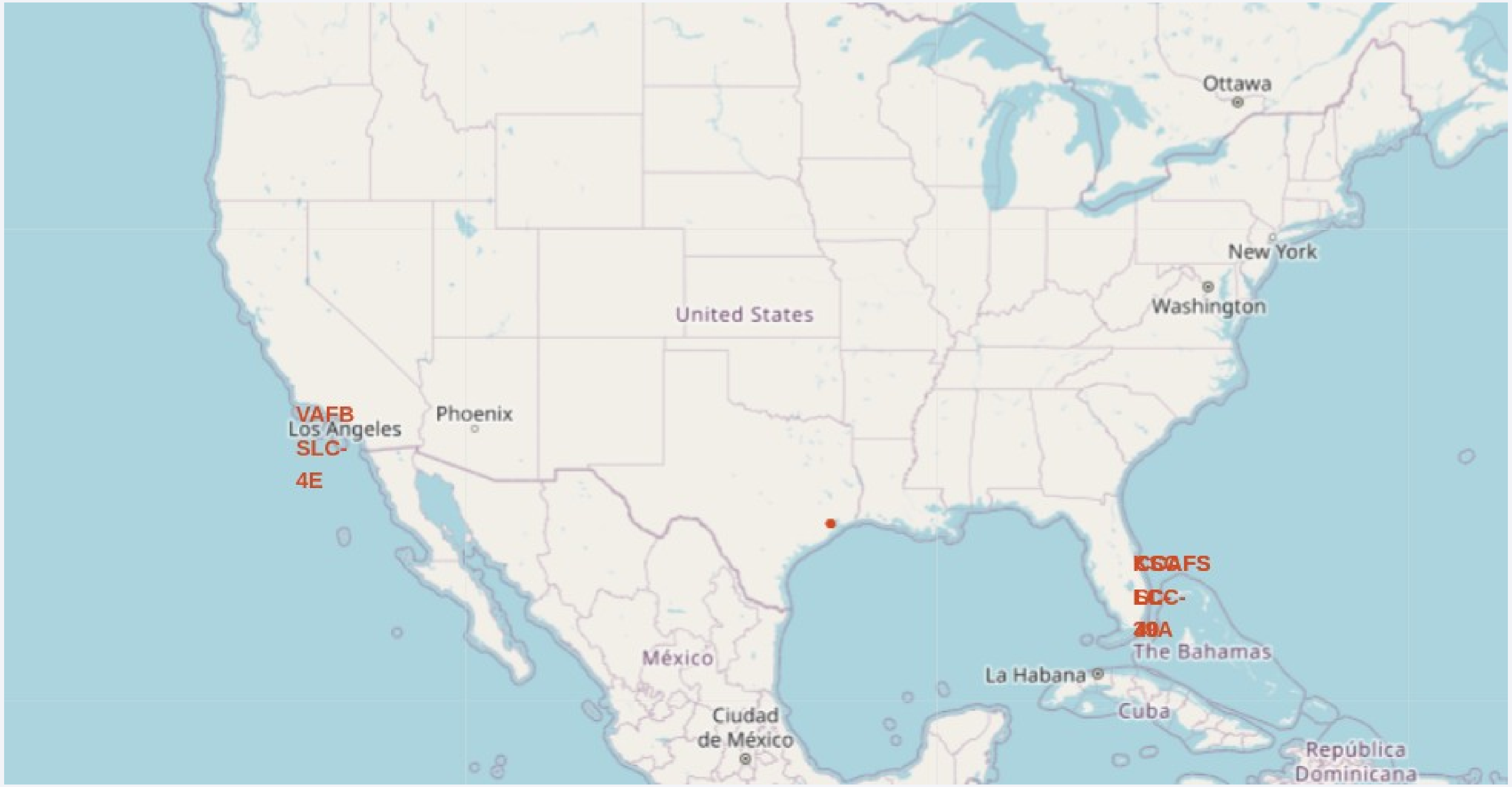
Landing _Outcome	COUNT_LAUNCHES
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

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

Section 3

Launch Sites Proximities Analysis

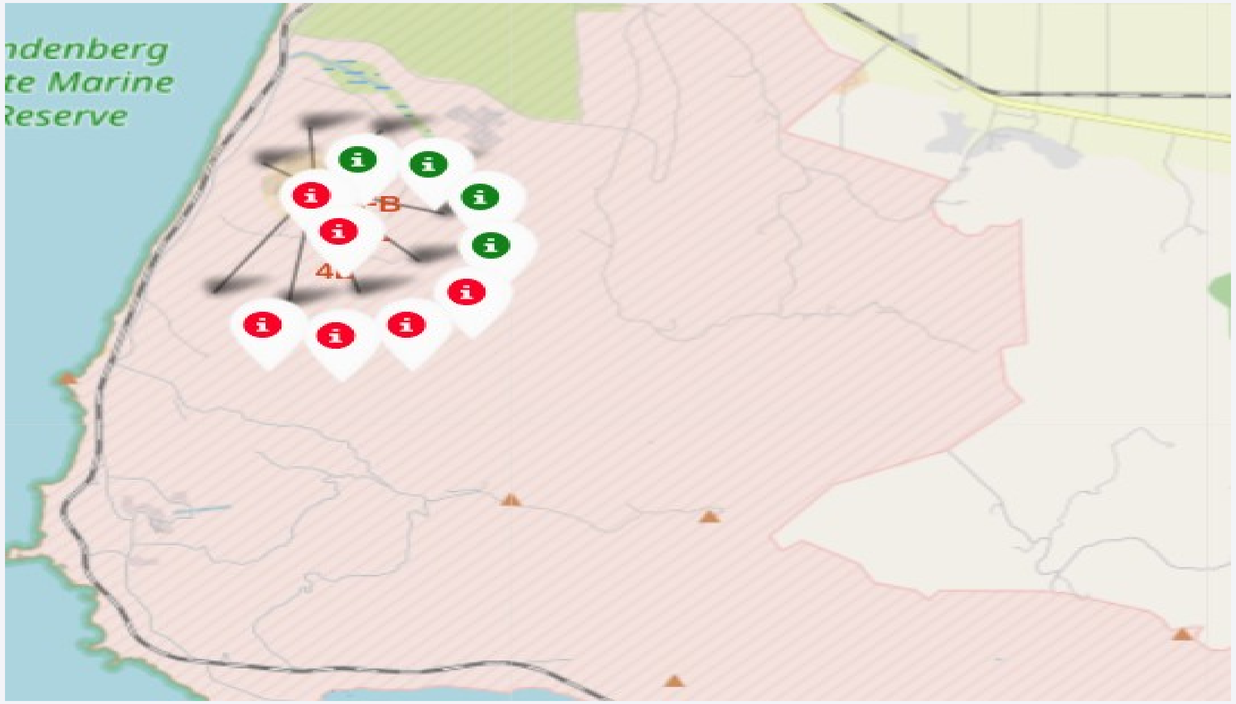
Launch Site



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

Launch Outcomes by Site

Example of VAFB SLC-4E launch site launch outcomes



Green markers indicates success and red markers indicates failure.

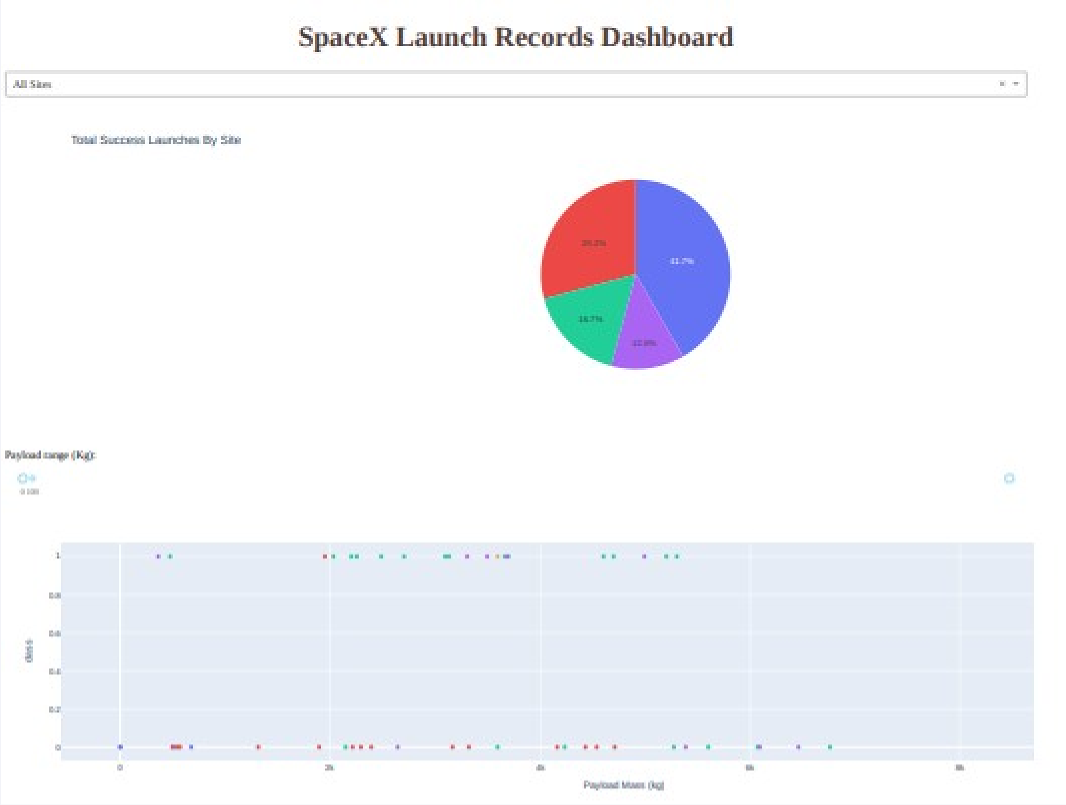
<Folium Map Screenshot 3>

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

Section 4

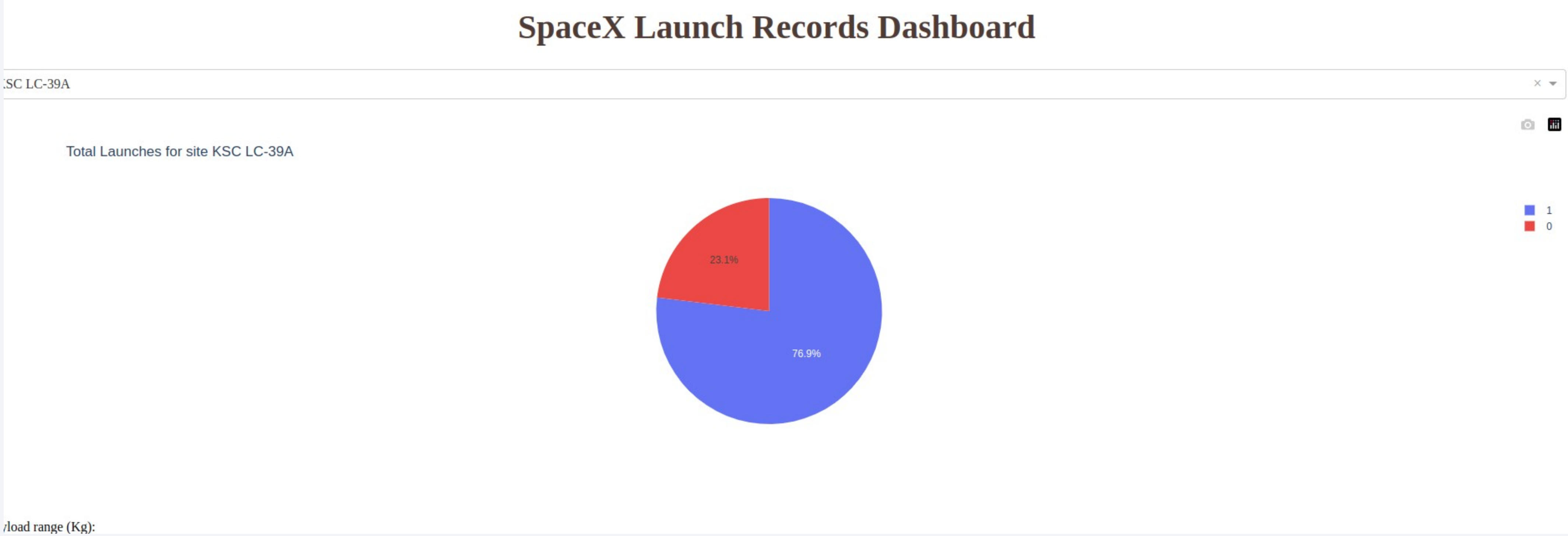
Build a Dashboard with Plotly Dash

Successful Launches by Site



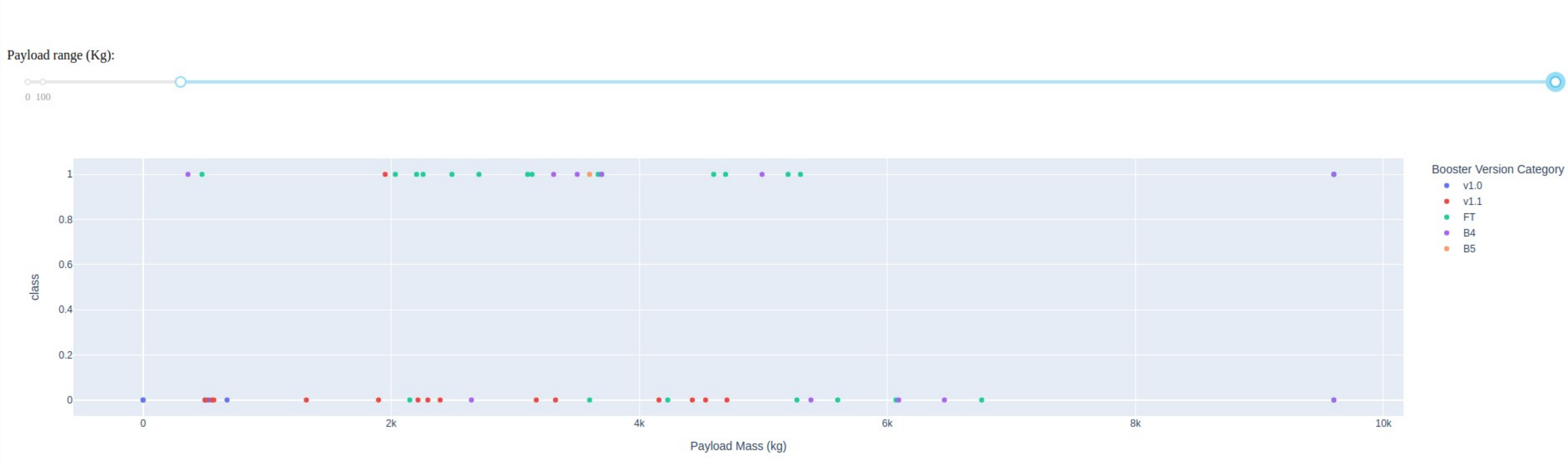
The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



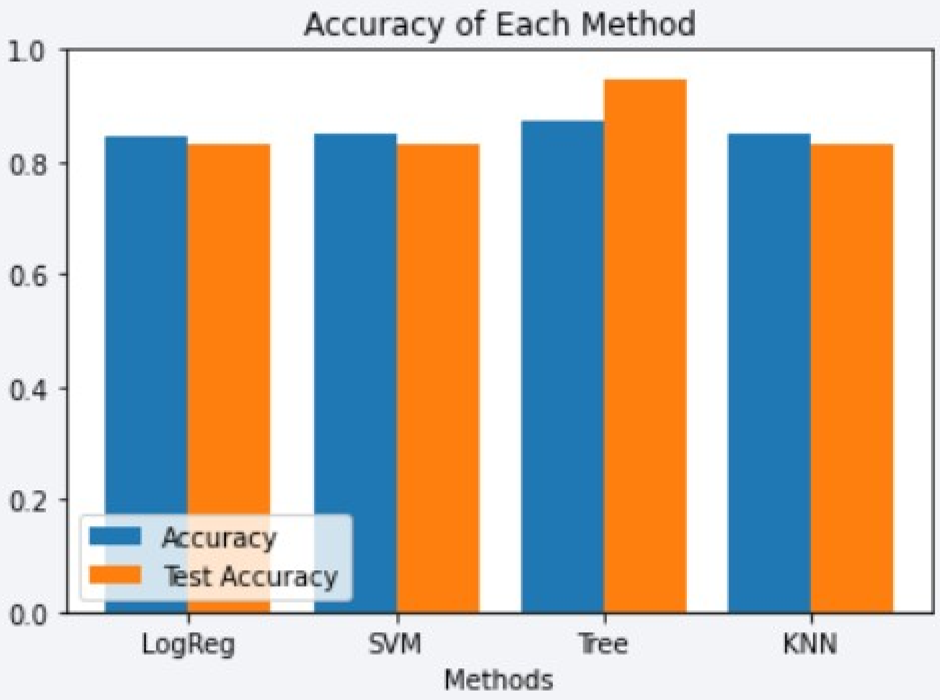
- Payloads under 6,000kg and FT boosters are the most successful combination.

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

Different data sources were analyzed, refining conclusions along the Process;

- The best launch site is KSC LC-39A;

- 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;
- Use github.dev for better interactive with the notebook

Thank you!