Winning Space Race with Data Science

Mustaquim Ahmad
6th May, 2022

Outline

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

Executive Summary

- Below are the methodologies were used to analyze and present the data:
  1. Data collection using web scrapping and SPACE_X API
  2. Data Wrangling
  3. EDA with data visualization
  4. EDA with SQL
  5. Map with Folium
  6. Dashboard with Plotly Dash
  7. Machine Learning Analysis using SVM
- Summary of all results
  1. EDA results
  2. Interactive and Predictive analysis

Introduction

Project background and context

- ❑ The goal of the project is to forecast the successful landing of the Falcon 9 first stage. SpaceX promotes the launches of Falcon 9 rockets on its website, priced at $62 million, whereas other providers charge over $165 million per launch. The significant cost savings of SpaceX is due to their ability to reuse the first stage. Consequently, by predicting the landing outcome, we can estimate the overall cost of a launch.

Problems

- ❑ This project final task is to predict if the first stage of the SpaceX Falcon 9 will land successfully or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Collected data was enriched by encoding data fields for machine learning and data cleaning of null values and unnecessary columns.
- 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

- Datasets were collected from SpaceX REST API & from Wikipedia
  - https://api.spacexdata.com/v4/rockets/
  - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used.
- This API was used according to the flowchart beside and then data is persisted.

- GitHub source url:

https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api_m_a.ipynb

Requested API and parse the SpaceX Launch data

Filter data to only Include Falcon 9 launches

Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from the Wikipedia.
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- GitHub source url:

https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb

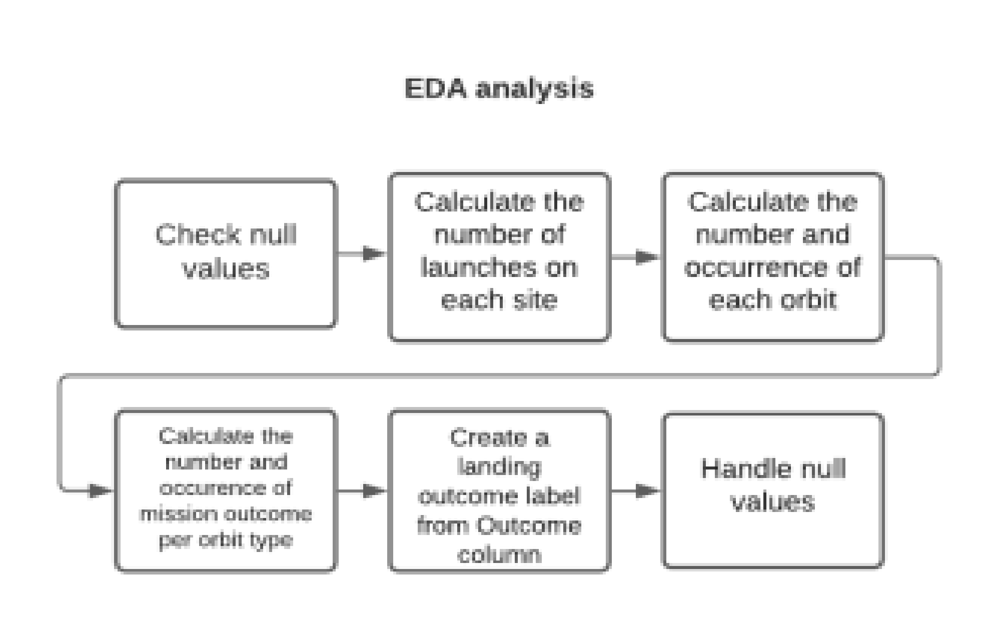
Request the Falcon 9 Launch Wikipedia page

Extract all column/variables names from the HTML table header

Create a data frame by parsing the Launch from HTML table

Data Wrangling

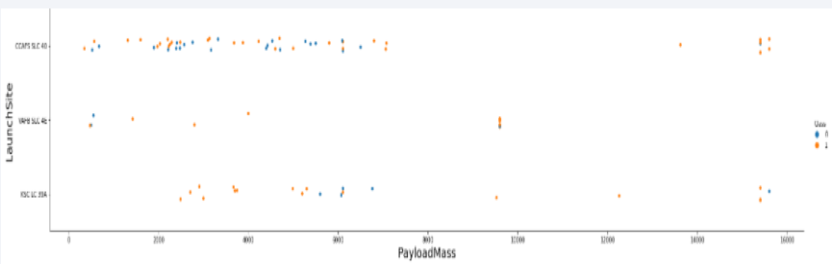
- Initially some exploratory data analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit were calculated.
- Finally, the landing outcome was achieved.

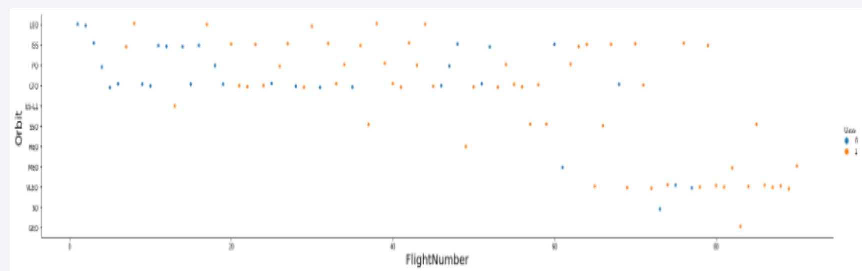


EDA with Data Visualization

- To explore data, scatterplots and bar plots 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/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/jupyter-labs-eda-dataviz.ipynb





EDA with SQL

- The following SQL queries were performed:
  - Names of the unique 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.
  - 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 version which have carried the maximum payload mass.

url: https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker cluster were used with Folium Maps.
  - Markers indicate points like launch sites
  - Circles highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launch site
  - Lines are used to indicate distances between two coordinates.

url: https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

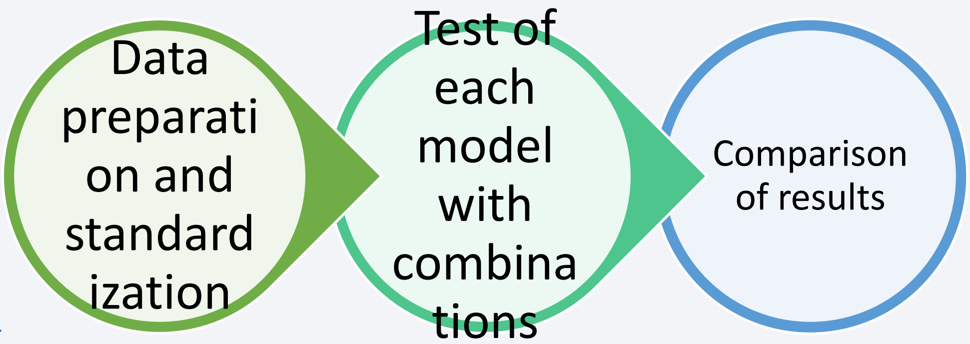
- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is the best place to launch according to payloads.

url: https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/SpaceX%20Dashboard%20App/spacex_dashboard_app.py

Predictive Analysis (Classification)

- Four classification models were compared:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN

  url: https://github.com/mustaquim-ms/IBM_Data_Science_Capstone_Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



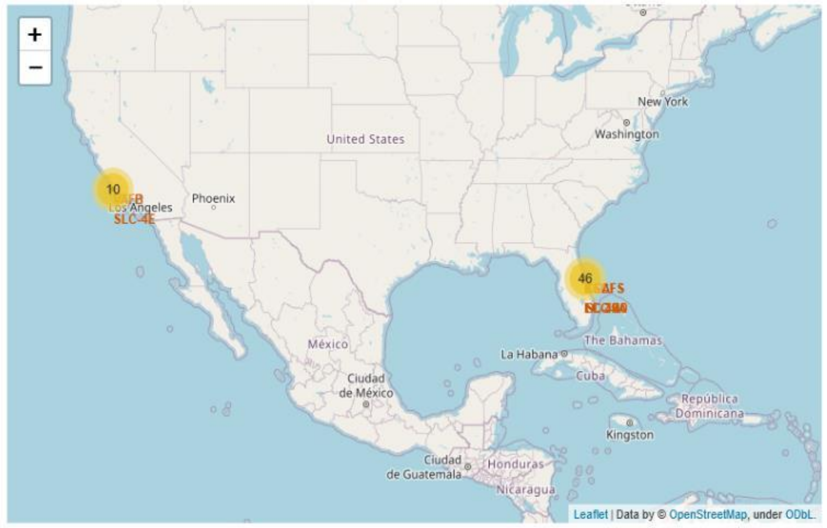
Results

- 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

Results

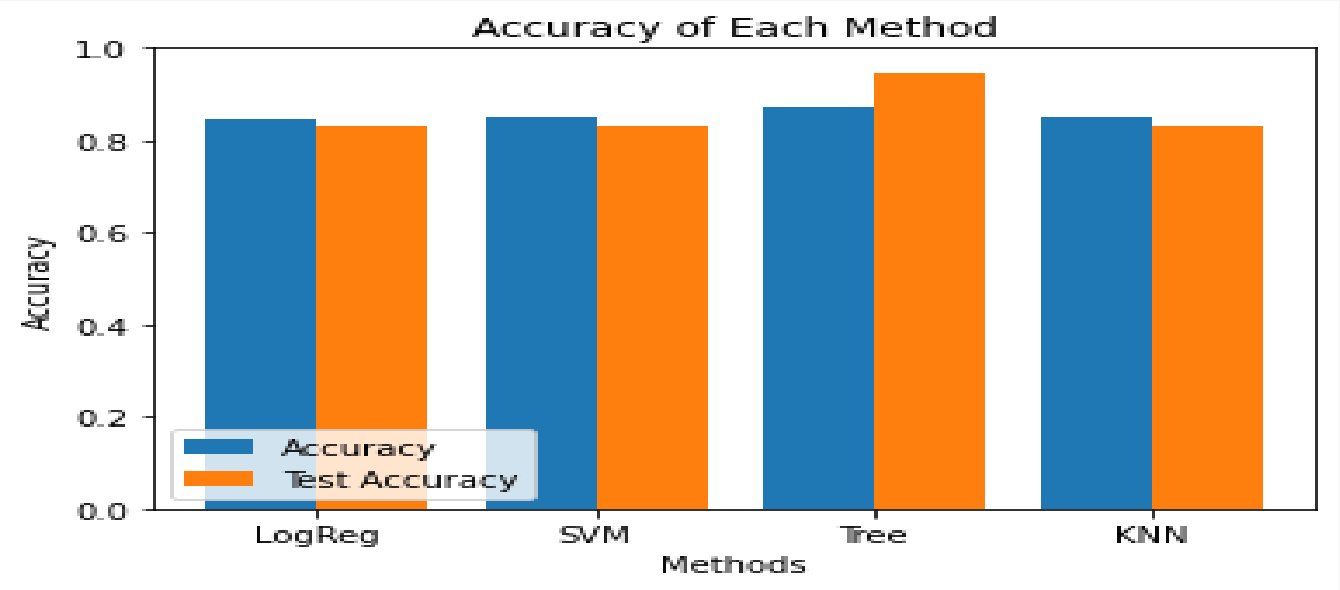
• Using interactive analytics was possible to identify that launch sites used to be in safety places, near sea, i.e., and have a good logistic infrastructure around.

• Most launches happens at east coast launch sites.



Results

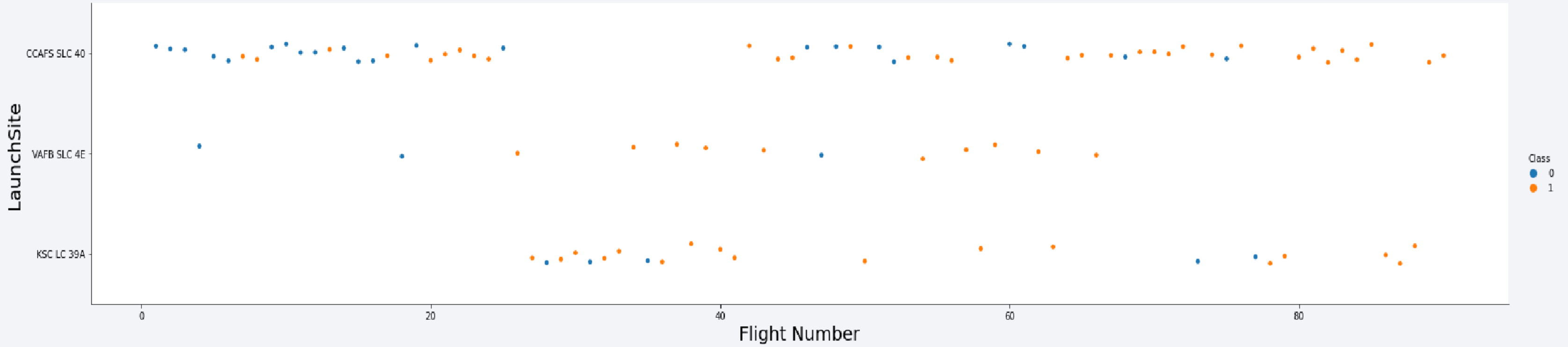
• Predictive analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 85% and accuracy for test data over 94%.



Section 2

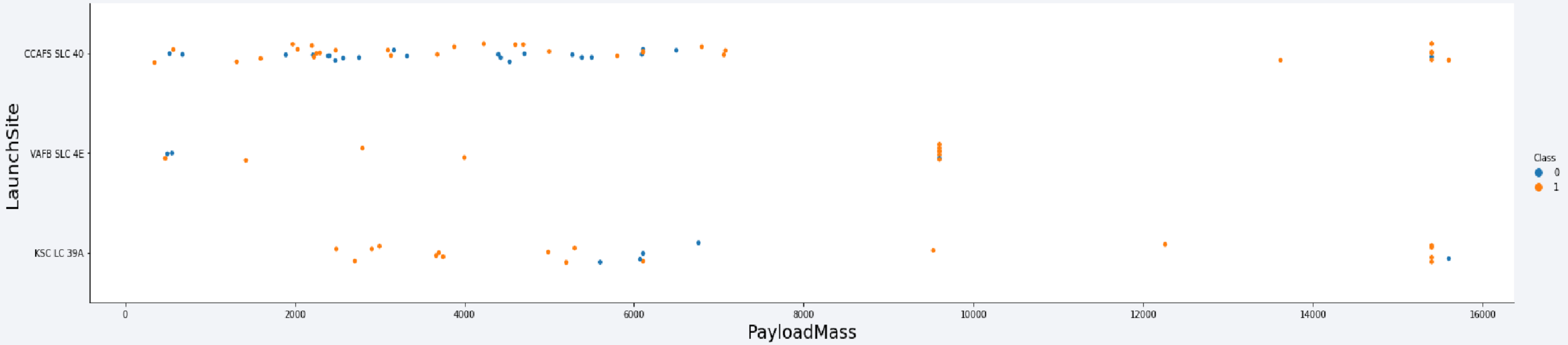
Insights drawn from EDA

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.

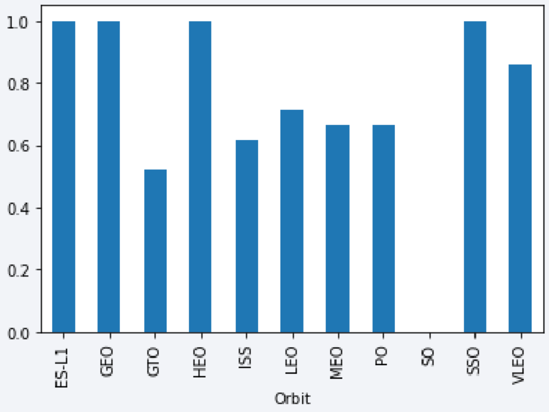
Payload vs. Launch Site



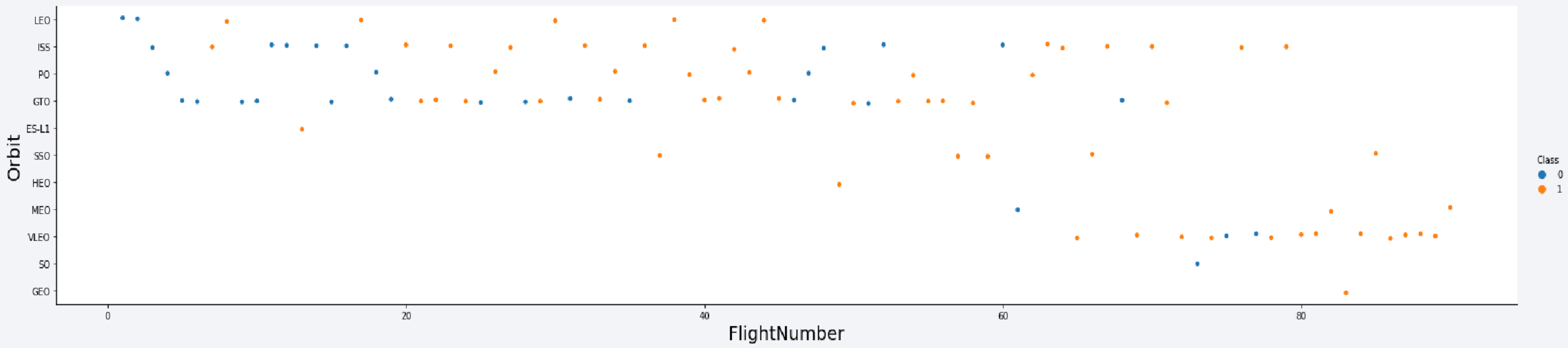
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

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

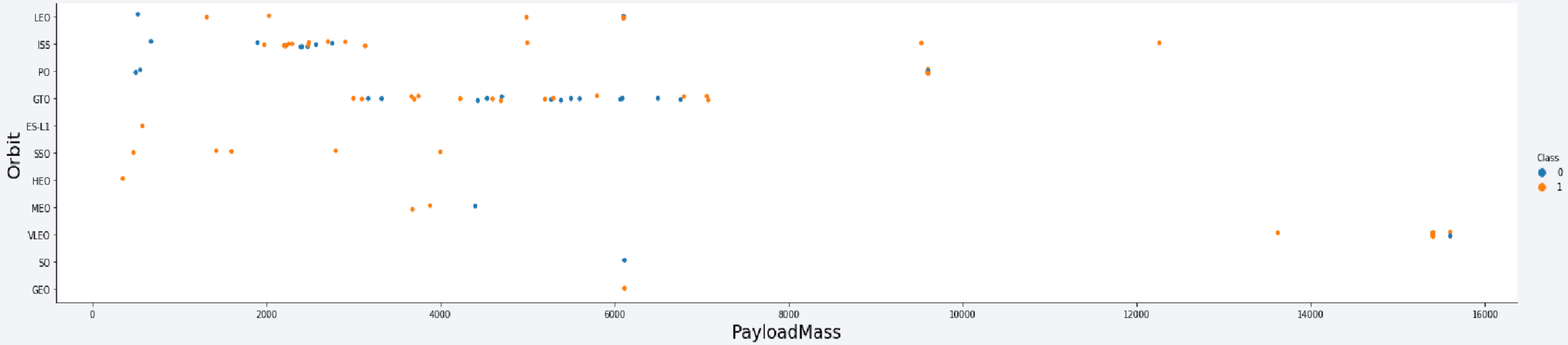


Flight Number vs. Orbit Type



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

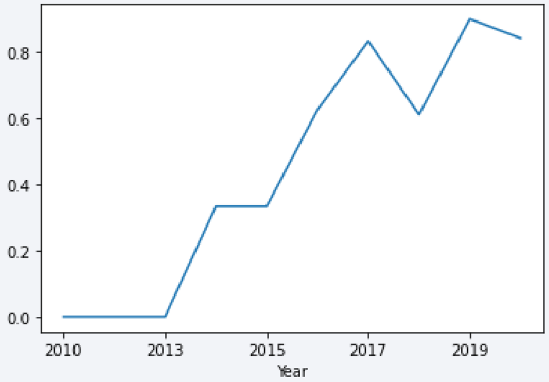
Payload vs. Orbit Type



- 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:

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

- They are obtained by selecting unique occurrences of "launch_site" values 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
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

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

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- 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 (kg)
2.928

- 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

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- 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

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

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

Total Number of Successful and Failure Mission Outcomes

- 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

- Boosters 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

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

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

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

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

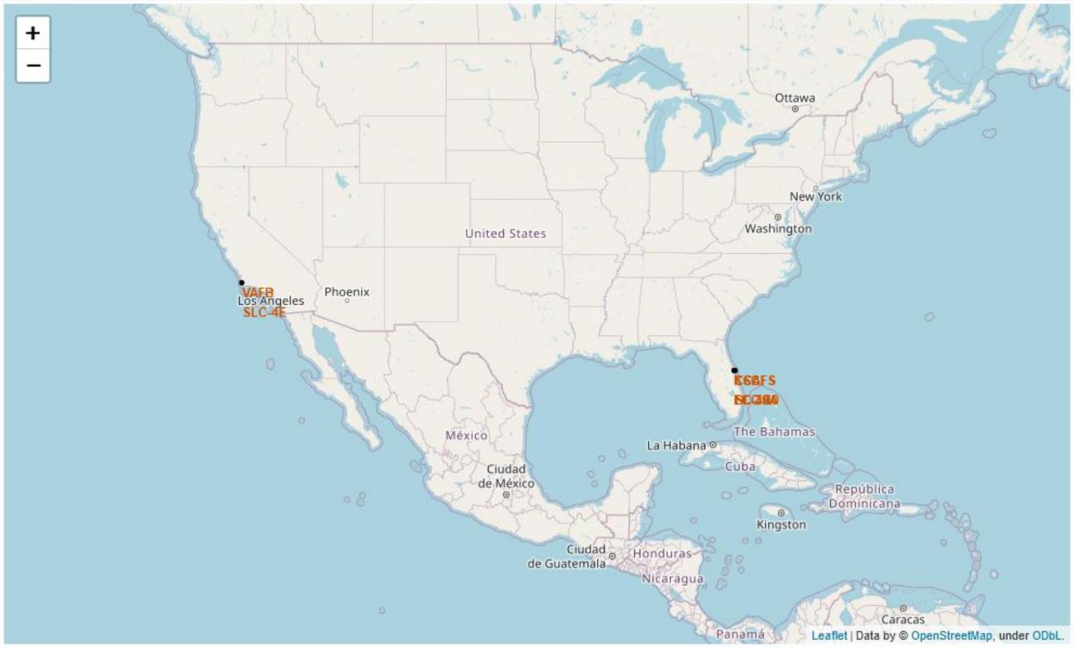
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 4

Launch Sites Proximities Analysis

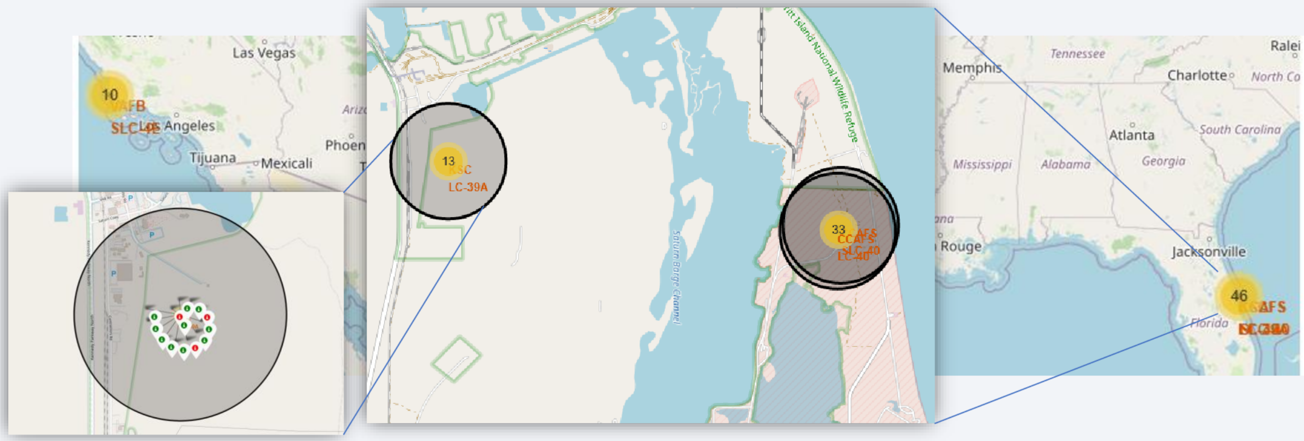
All launch sites



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

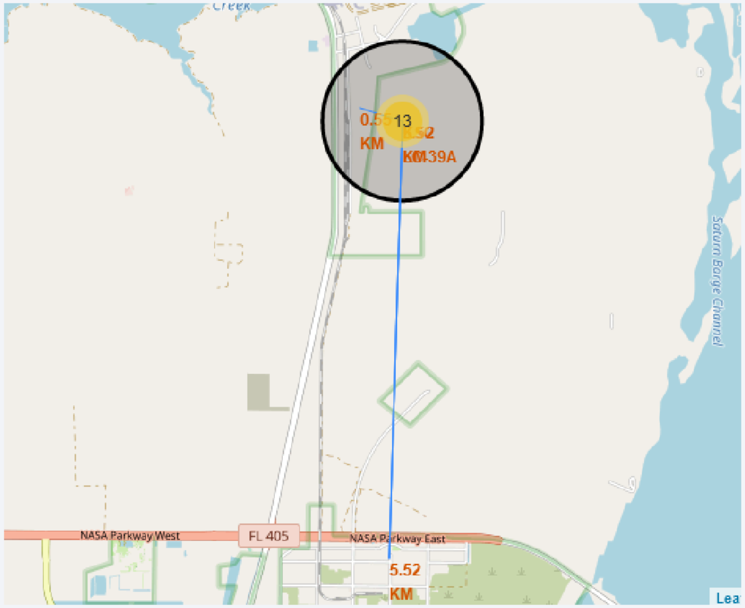
Launch Outcomes by Site

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



- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

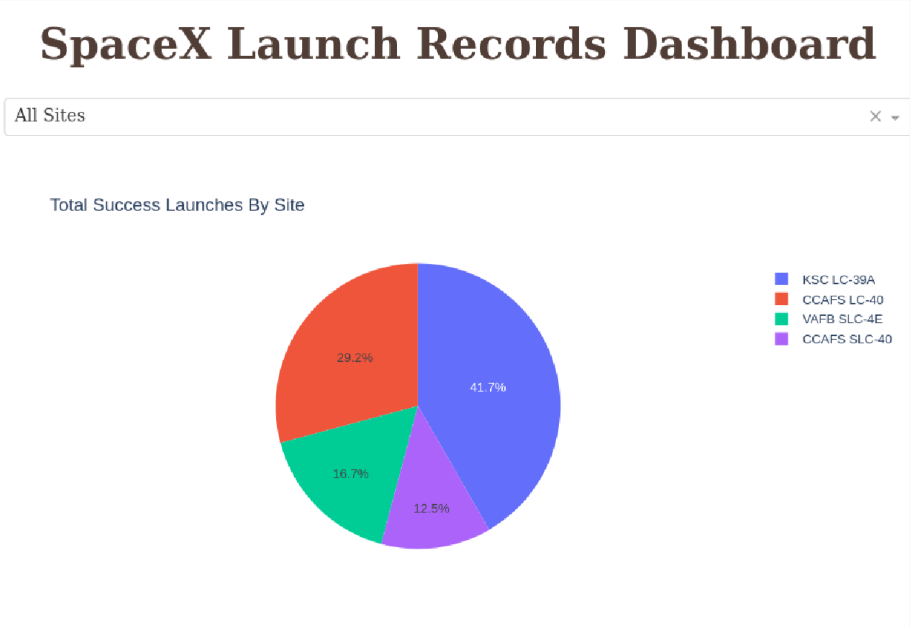


- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 5

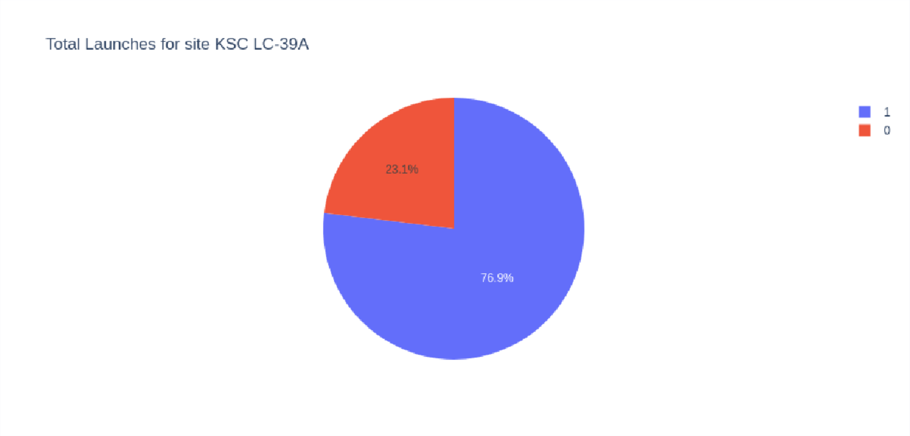
Build a Dashboard with Plotly Dash

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

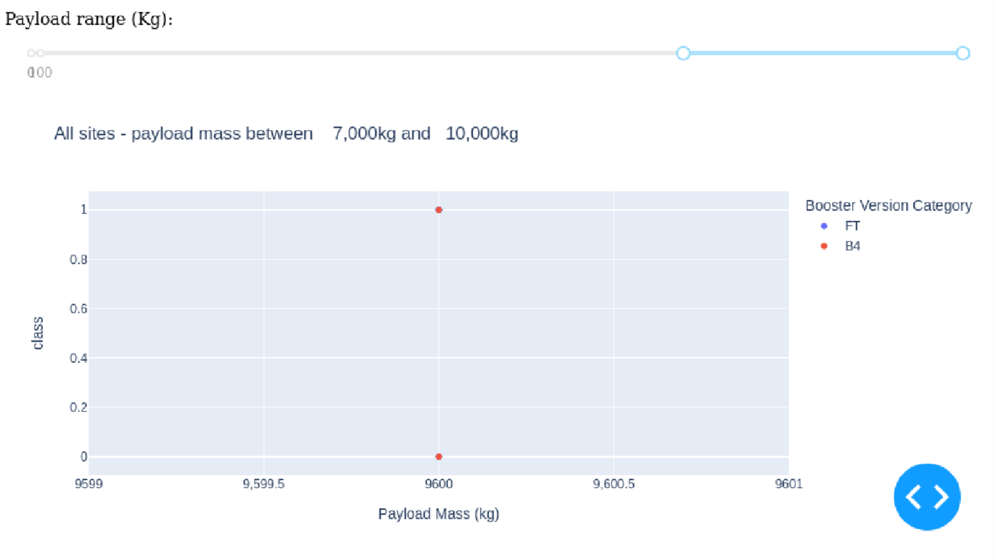


- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome

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

Payload vs. Launch Outcome

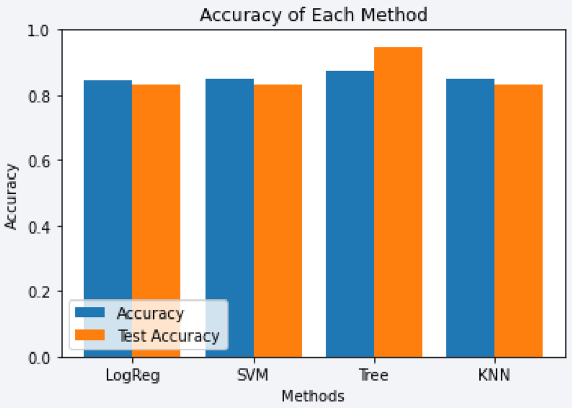


- There's not enough data to estimate risk of launches over 7,000kg


Section 6
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 of Decision Tree Classifier

- 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

- All the data were used appropriately
- Some images did not show up so I had to take screenshots.

Thank you!