IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ghazouani Hichem
20/03/2022

Outline

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

Executive Summary

- For winning Space Race, SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- In this presentation we will predict if the Falcon 9 first stage will land successfully. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Executive Summary

- **The following methodologies were used to analyze data:**
  - Data Collection using web scarping and SpaceX API.
  - Exploratory Data Analysis, 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.
  - ML Prediction showed the best model, using all collected data.

Introduction

- The objective of our project is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully. Therefore if we can accurately predict the likelihood of the first stage rocket landing successfully, we can determine the cost of a launch. So, the competing startup we have been hired by can make more informed bids against SpaceX for a rocket launch.

- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets.
  - Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    - SpaceX API (‘https://api.spacexdata.com/v4/…’)
    - Web Scraping (‘List of Falcon 9 and Falcon Heavy launches – Wikipedia’)
- Perform data wrangling
  - Collected data was processed by creating a landing outcome label based on outcome data after summarizing and analyzing features.

Methodology

Executive Summary

- 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 were normalized, divided in training and test data sets and evaluated by four different models, being the accuracy of each model evaluated using different combinations of parameters for objective to find the best model.

Data Collection

- Data sets were collected from SpaceX API and from Wikipedia using Web Scraping technics.

Data Collection – SpaceX API

- SpaceX offers a public REST API where data can be obtained and used.

- Source code:

  https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/labs-spacex-data-collection-api.ipynb

- Request REST API and parse the SpaceX launch data
- Filter data to only include Falcon 9 launches
- Deal with Missing Values

Data Collection - Scraping

- Data is obtained also from Wikipedia pages.
- Source code:

  https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/labs-webscraping.ipynb

- Request th Falcon 9 launch Wikipedia page
- Extract all data from the launch tables
- Create a DataFrame by parsing the launch tables

Data Wrangling

- Initially some Exploratory Data Analysis was performed on our dataset.
- Then we calculated the summaries launches per site, occurences of each orbit and mission outcome per orbit.
- Finally, the landing outcome label was created using the outcome column.

- Source code:
  https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/labs-spacex-Data-wrangling.ipynb

EDA with Data Visualization

- To explore data and finding relationship between pair of features, we have used scatterplots, barplots and line Chart.
- For example, we use Scatter Plots to visualize [**(y="PayloadMass", x="FlightNumber«) (y="LaunchSite", x="FlightNumber")(y="LaunchSite", x="PayloadMass") (y="Orbit", x="FlightNumber") (y="Orbit", x="PayloadMass")**] and Line chart **(x="Date", y="Class")**

- Source code:

  https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/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 ‘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 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 landiing outcomes in drone ship, their booster versions and launch sites names for in 2015.
  - Rank of the count of landing outcomes between the date 04-06-2010 and 20-03-2017.

  Source code:

  https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/labs-spacex-eda-sql.ipynb

Build an Interactive Map with Folium

- Markers, markers clusters, circles and lines 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/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/lab_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/GhazouaniHichem/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.

Data preparation and standardization → Test of each model with combinations of hyperparameters → Comparison of results

Source code:

https://github.com/GhazouaniHichem/Applied_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

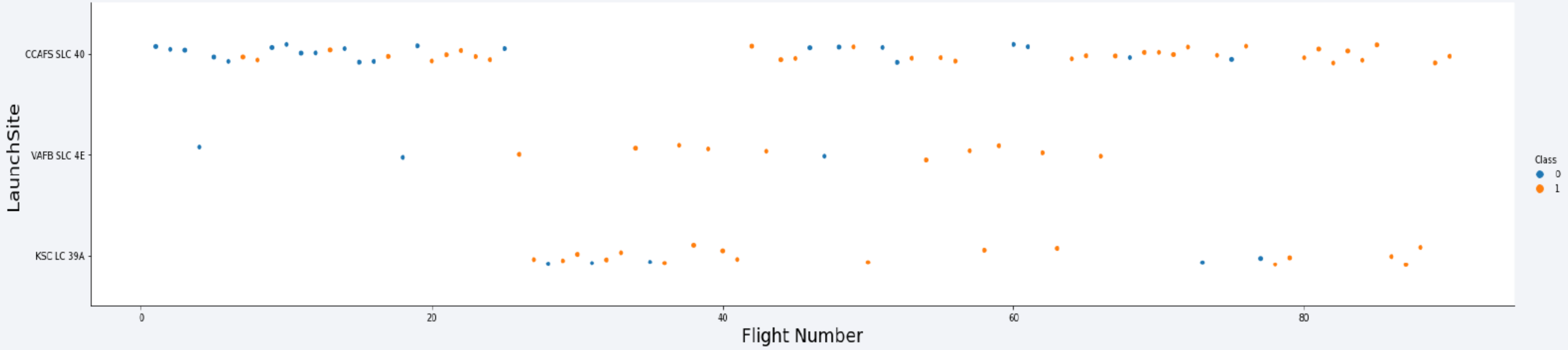
Exploratory data analysis results:

- 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.
- 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

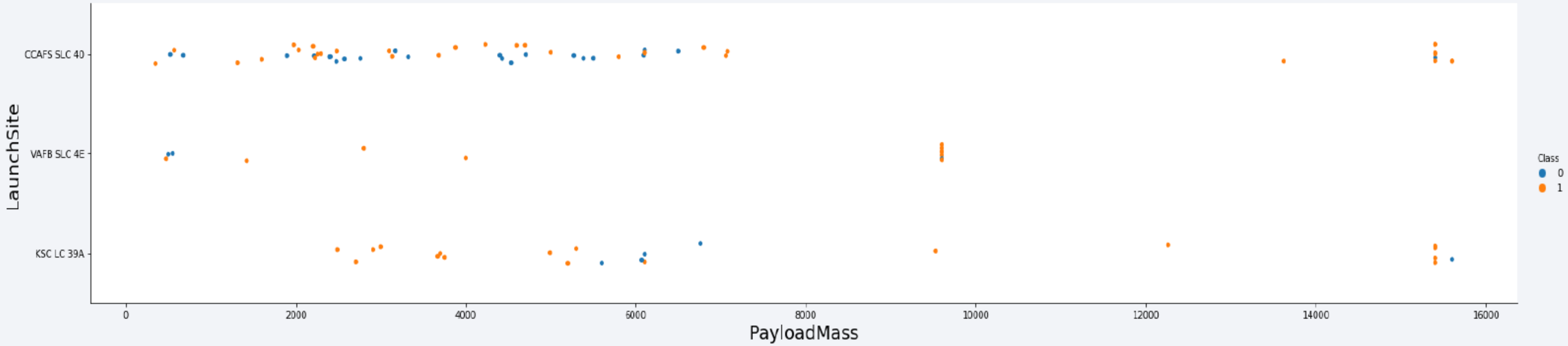


According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.

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.

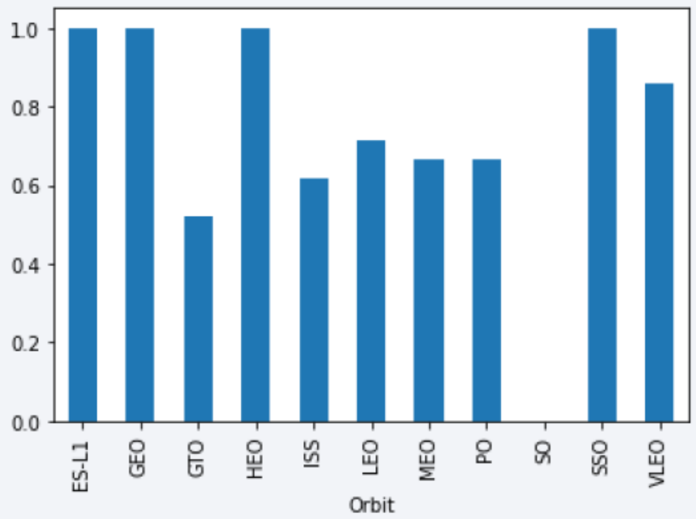
Payload vs. Launch Site



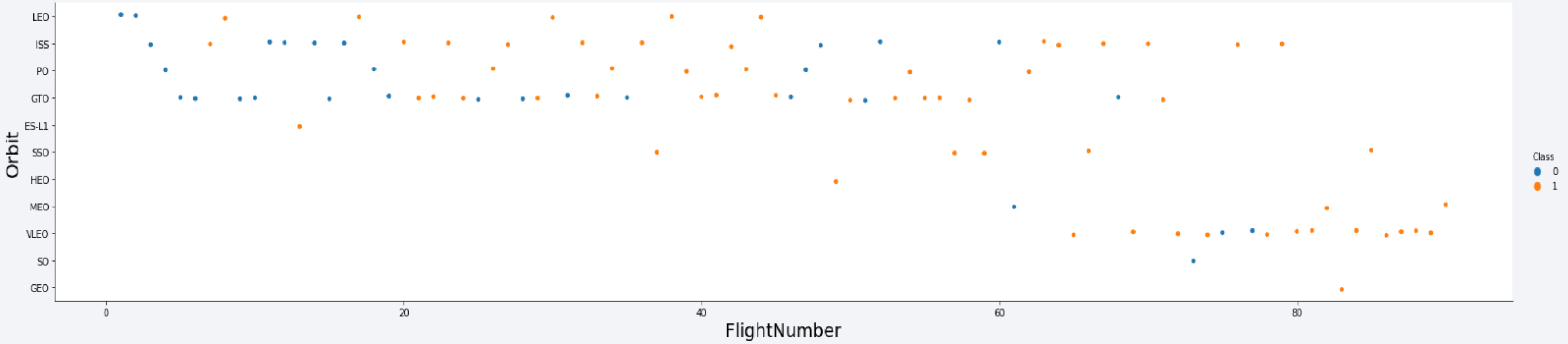
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

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%).

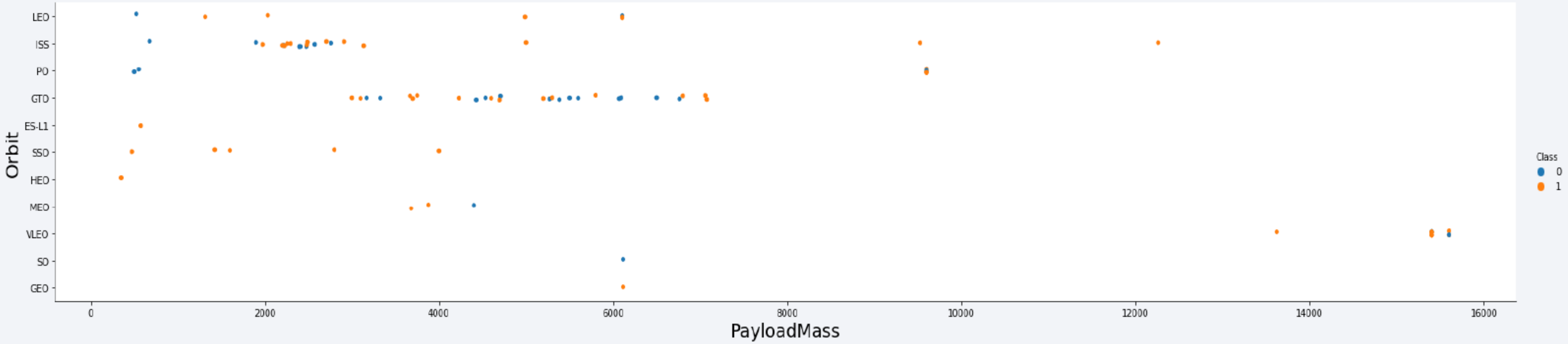


Flight Number vs. Orbit Type



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

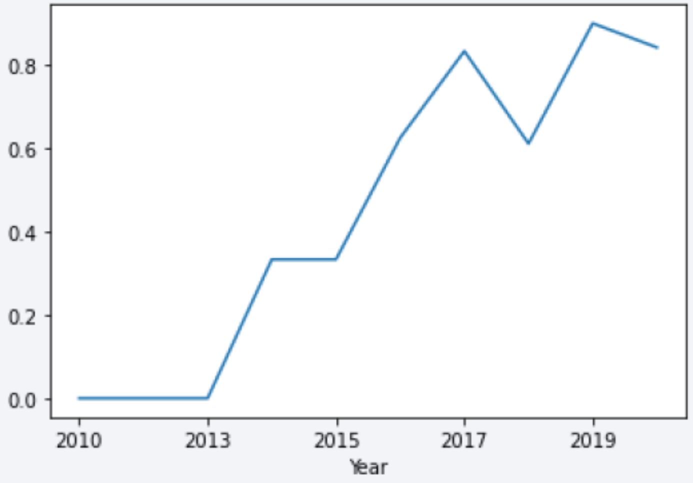
Payload vs. Orbit Type



- 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
- Itseemsthat 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'

- Find 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

- 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:

Average 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
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

- 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 B 1031.2
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 B1051.4
F9 B5 B1048.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1056.4
F9 B5 B1049.5	F9 B5 B1058.3
F9 B5 B1049.7	F9 B5 B1060.2
F9 B5 B1051.3	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, andlaunch 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:

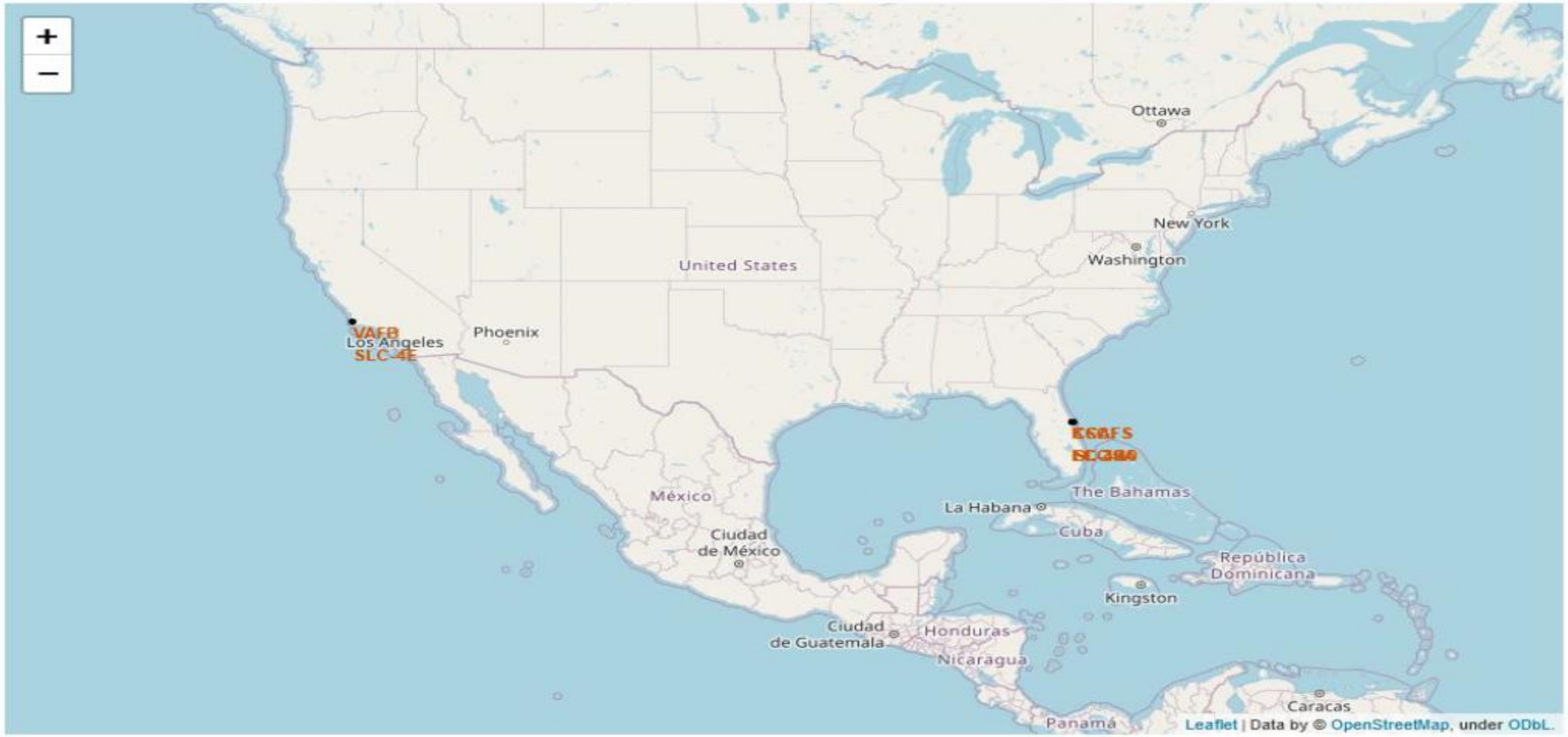
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 3

Launch Sites Proximities Analysis

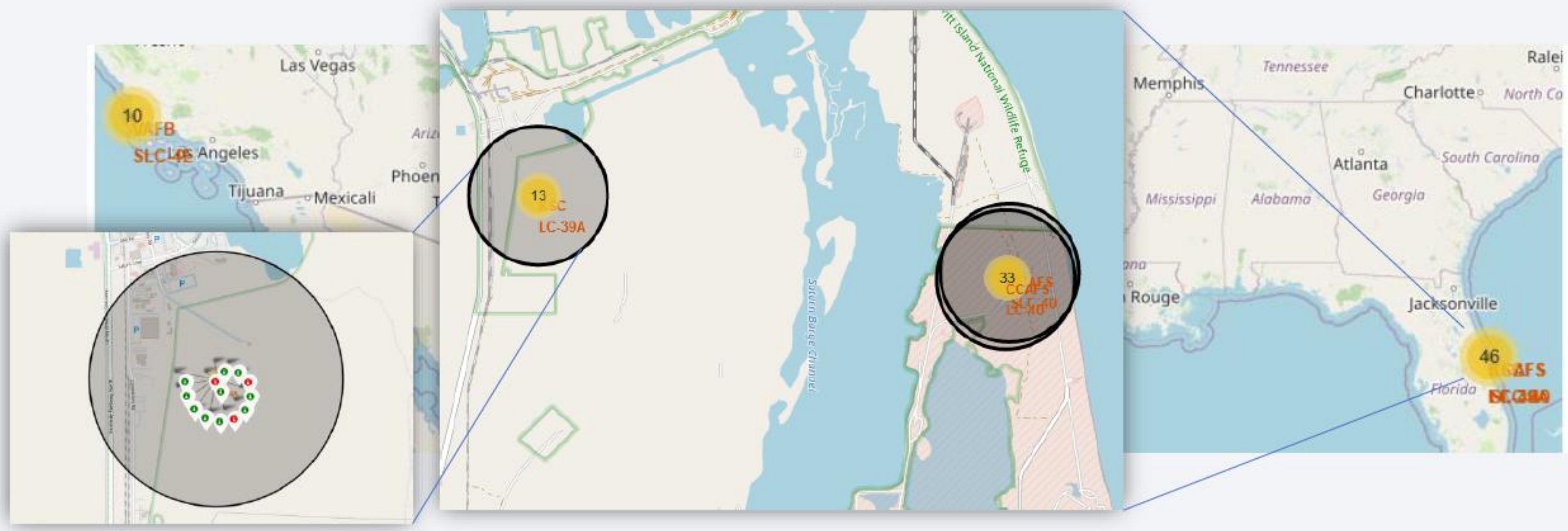
Launch Sites



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

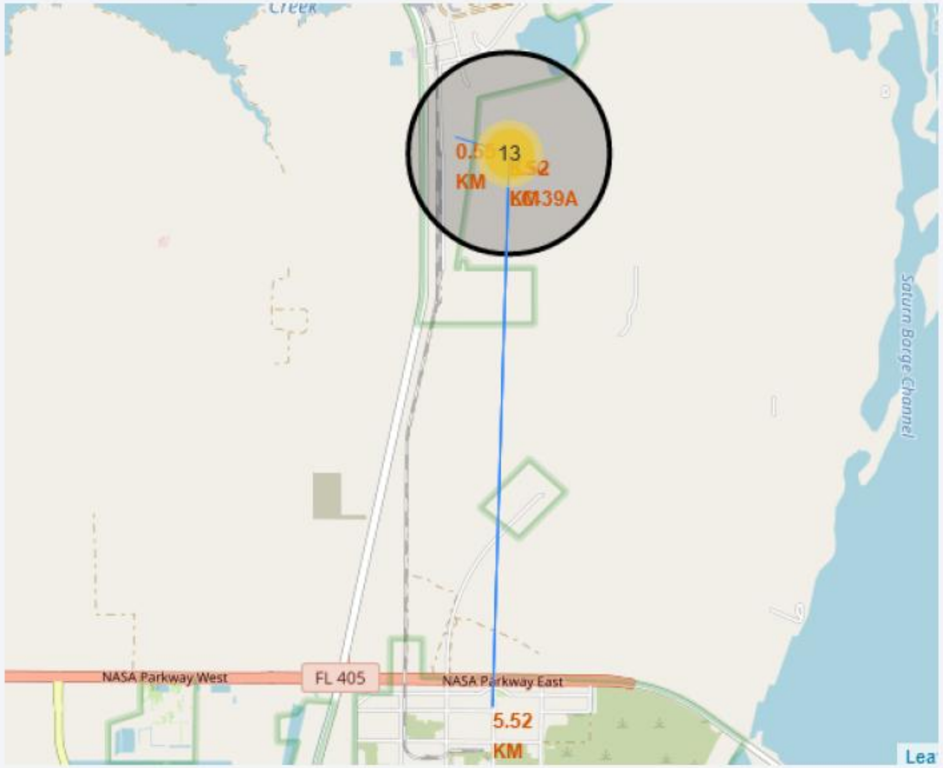
Launch Outcomes by Site

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



- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

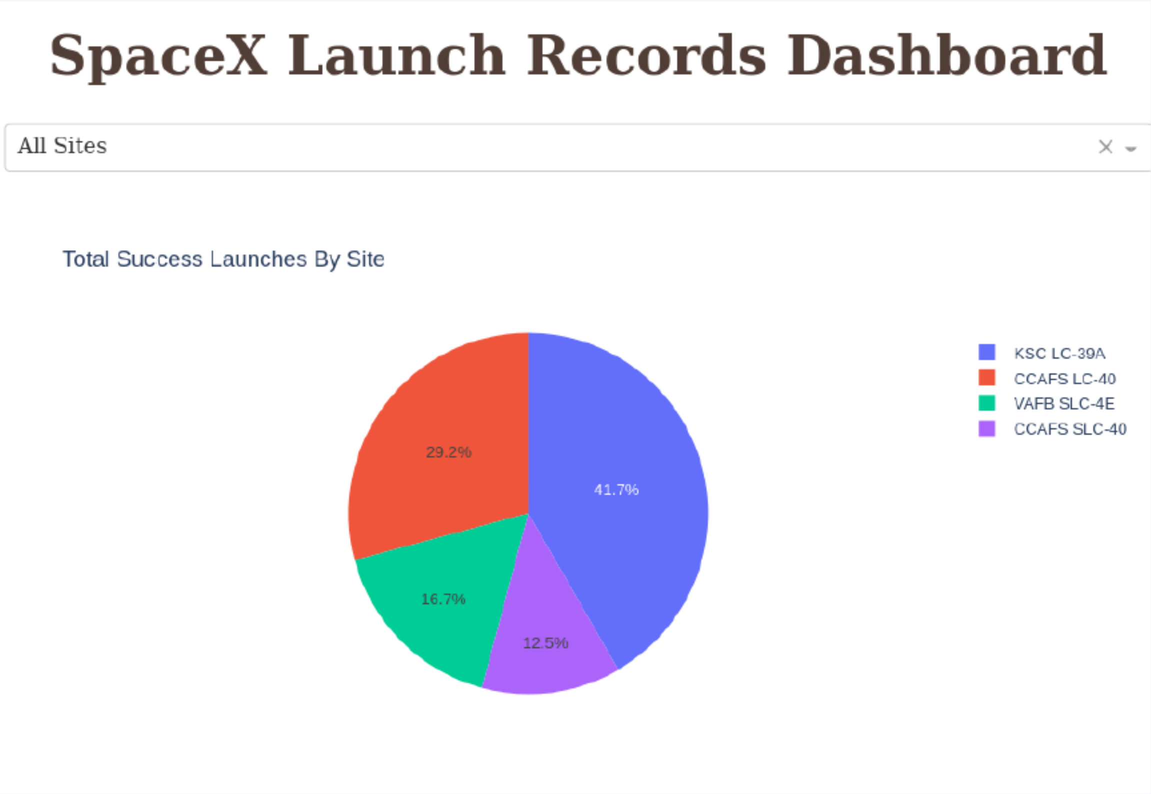


- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 4

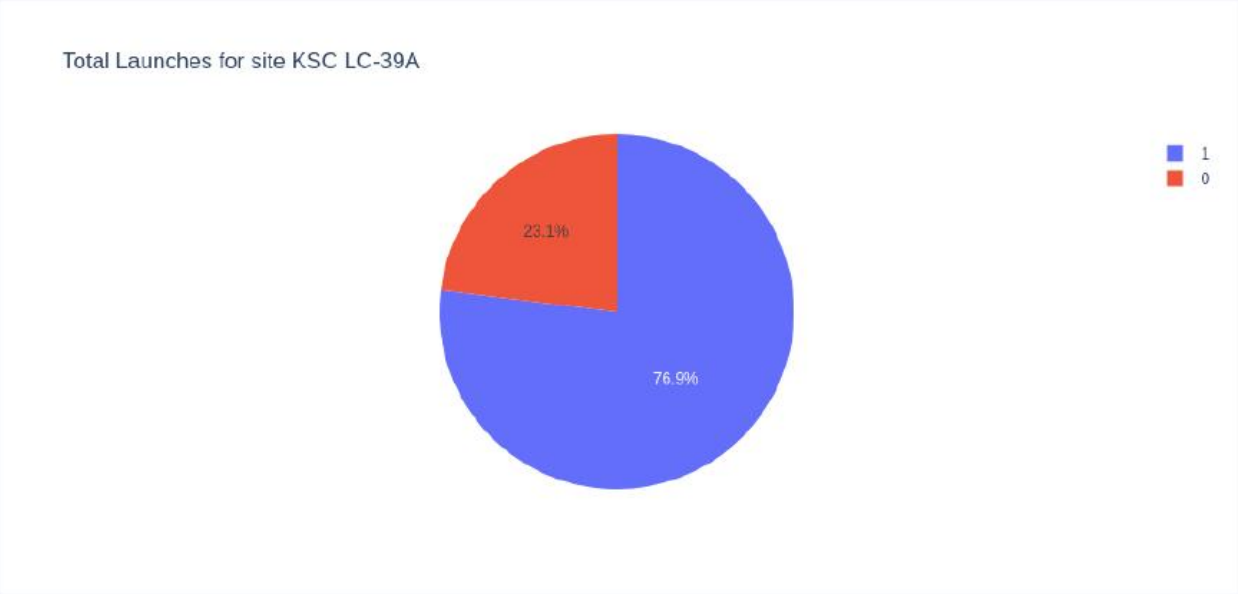
Build a Dashboard with Plotly Dash

Successful Launches by Site



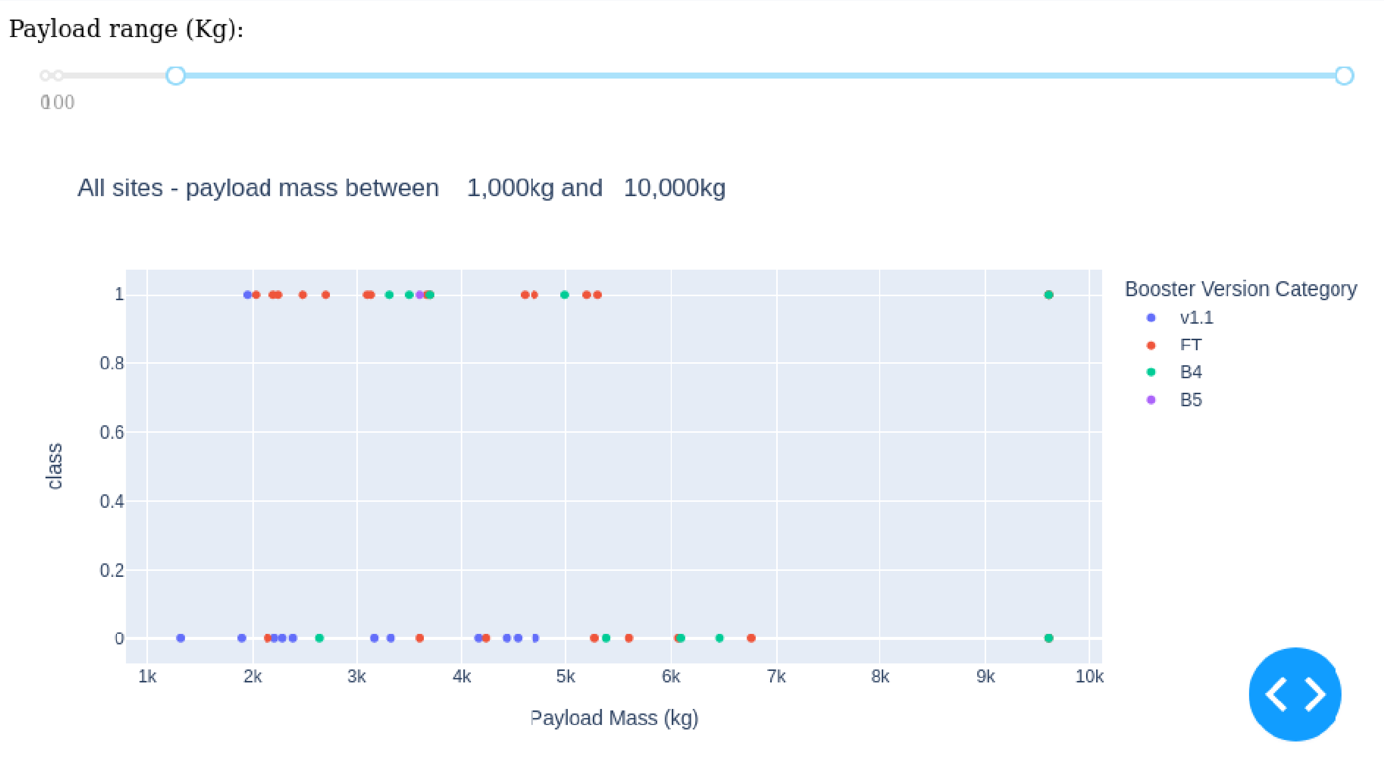
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



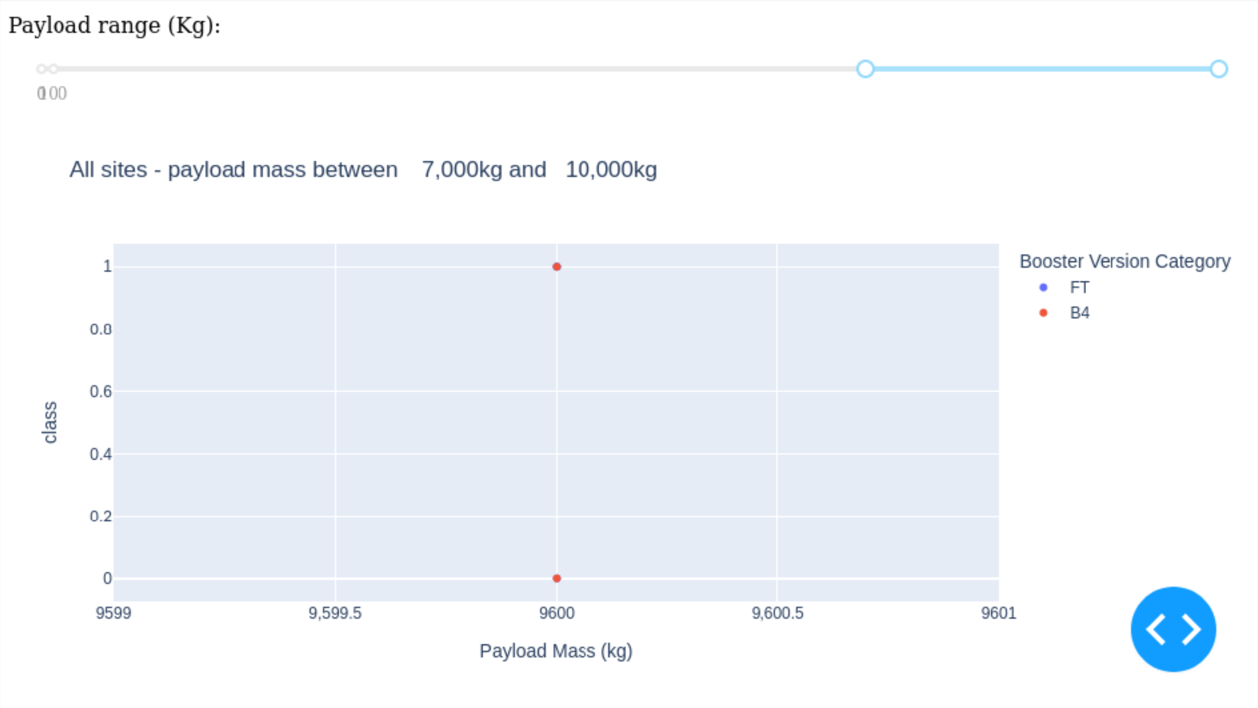
- 76.9% of launches are successful in this site.

Payload / Launch Outcome



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

Payload / Launch Outcome



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

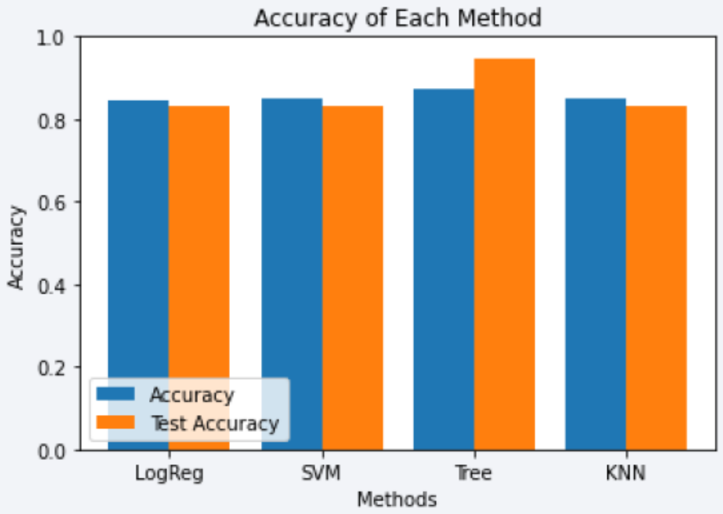
Section 5

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

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!