IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SUSHANT KUMAR
18-10-2023

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
  - Building a Dashboard with Plotly Dash
  - Machine Learning Prediction
- ❑ Summary of all results
  - Exploratory Data Analysis results
  - 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.
  - 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/rockets/)
    - Web Scraping (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

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash

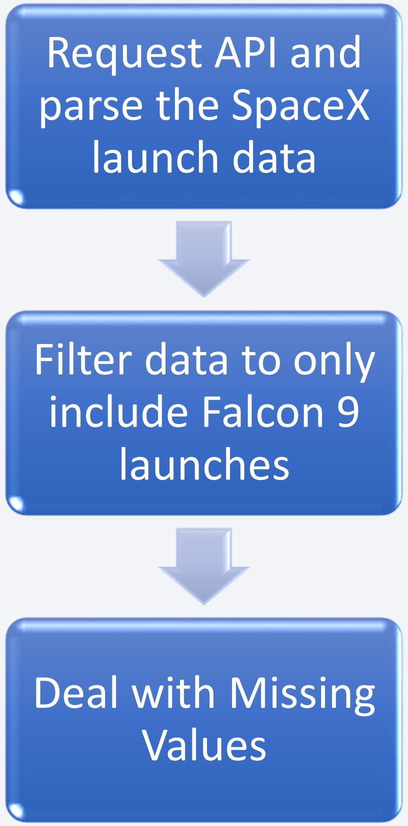
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

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 technique.

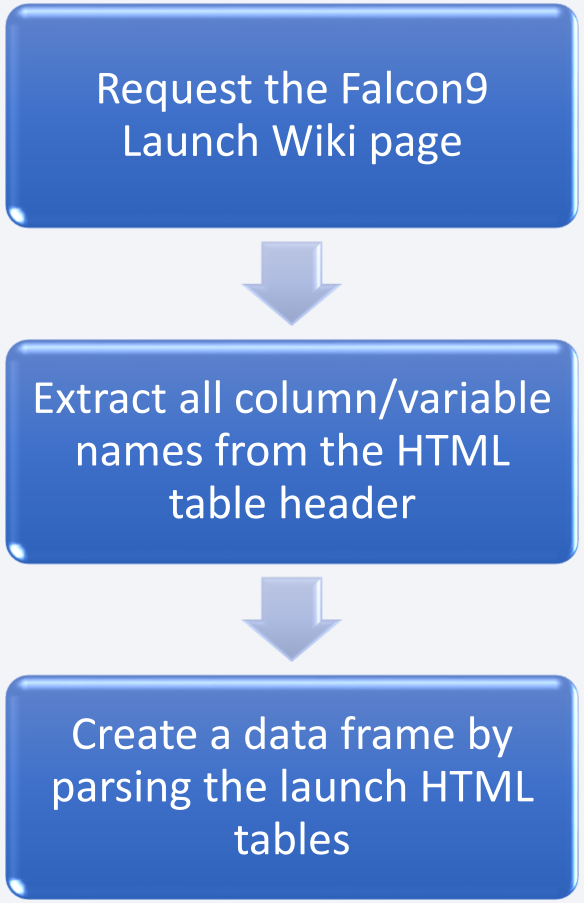
Data Collection – SpaceX API

- 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.
- Source code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/jupyter-labs-spacex-data-collection-api_1.ipynb



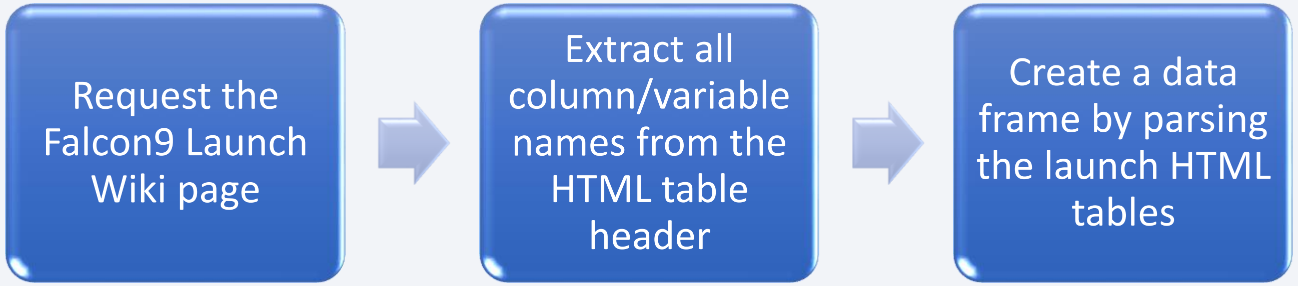
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia.
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source Code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/jupyter-labs-webscraping_2.ipynb



Data Wrangling

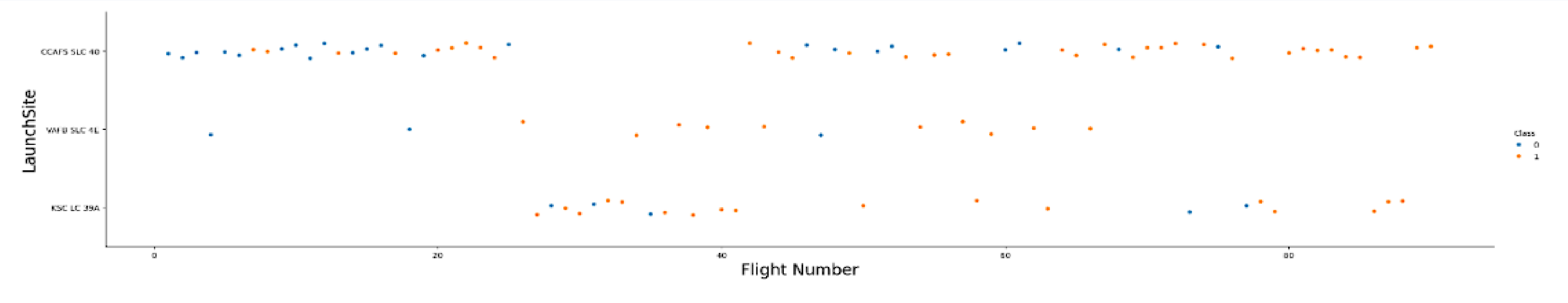
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrence of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- Source Code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling_3.ipynb

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



- Source Code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite_5.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; and
  - 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/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_4.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, and
  - Lines are used to indicate distances between two coordinates.

- Source Code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite_6.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/sushantkumar92940/DS_Final_Assignment_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

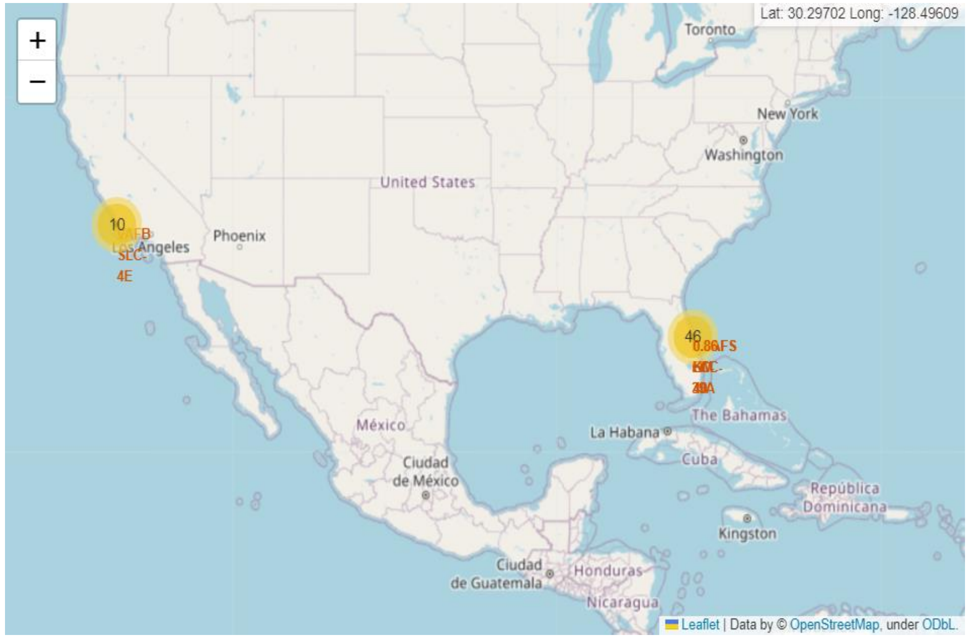
- Source code: https://github.com/sushantkumar92940/DS_Final_Assignment_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite_7.ipynb

Results

- Exploratory data analysis results:
  - SpaceX 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.

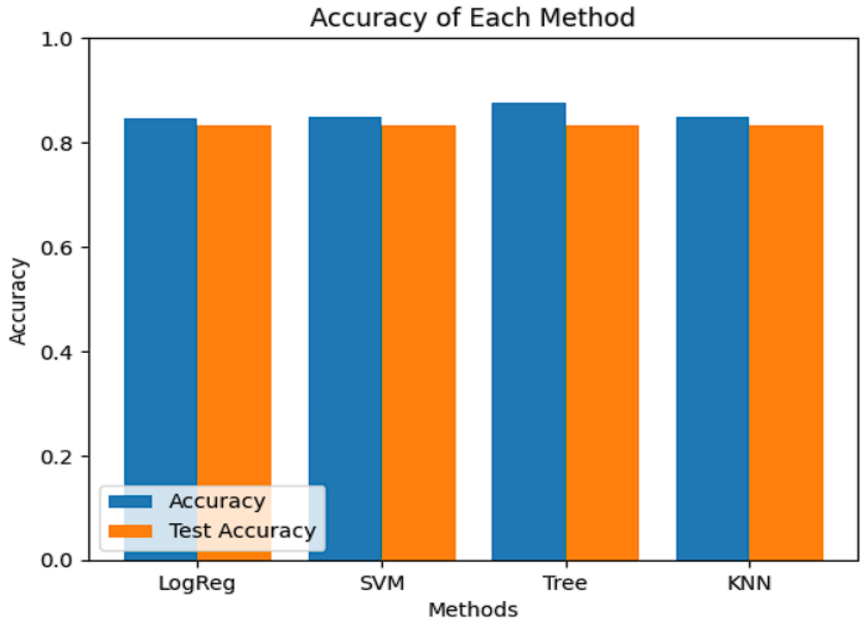
Results

- 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

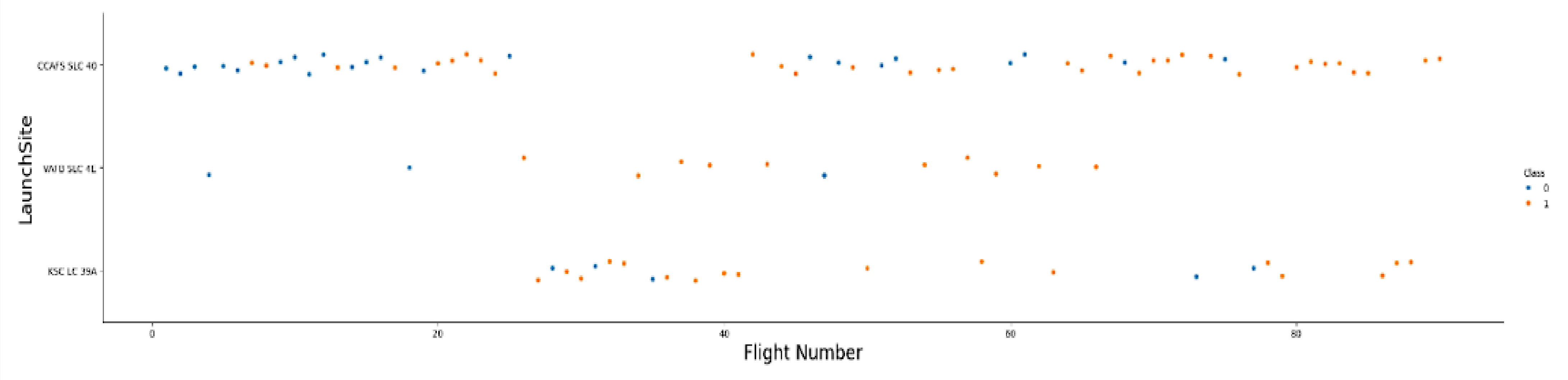
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 83%.



Section 2

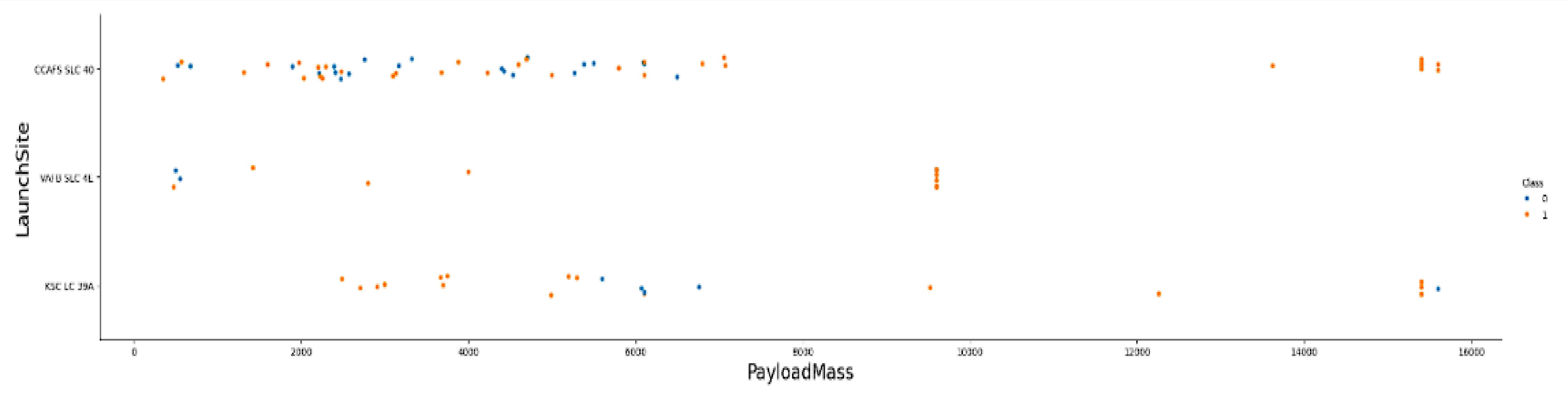
Insights drawn from EDA

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.

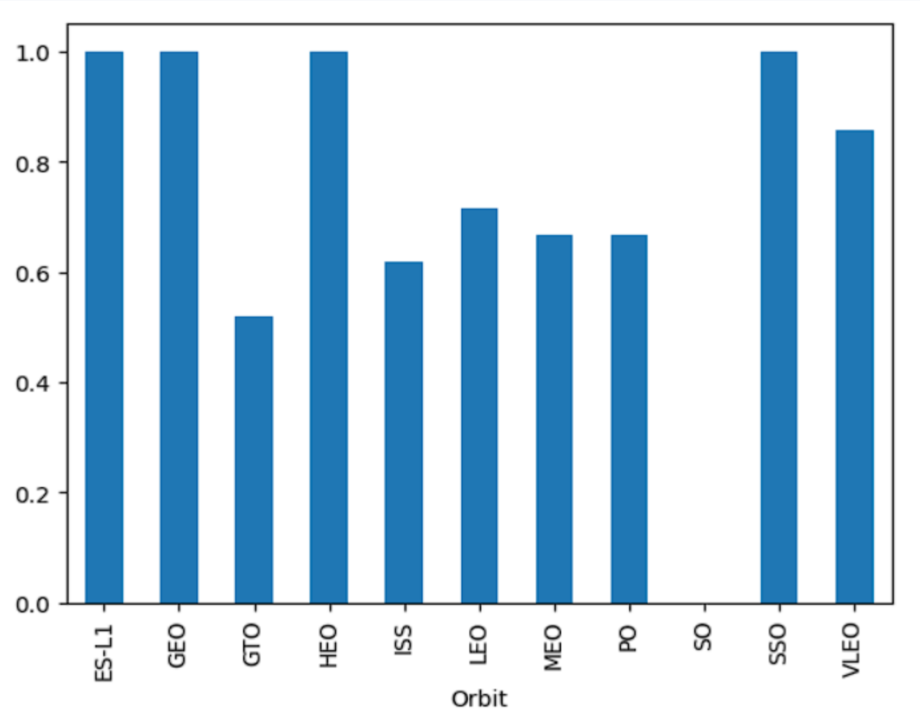
Payload vs. Launch Site



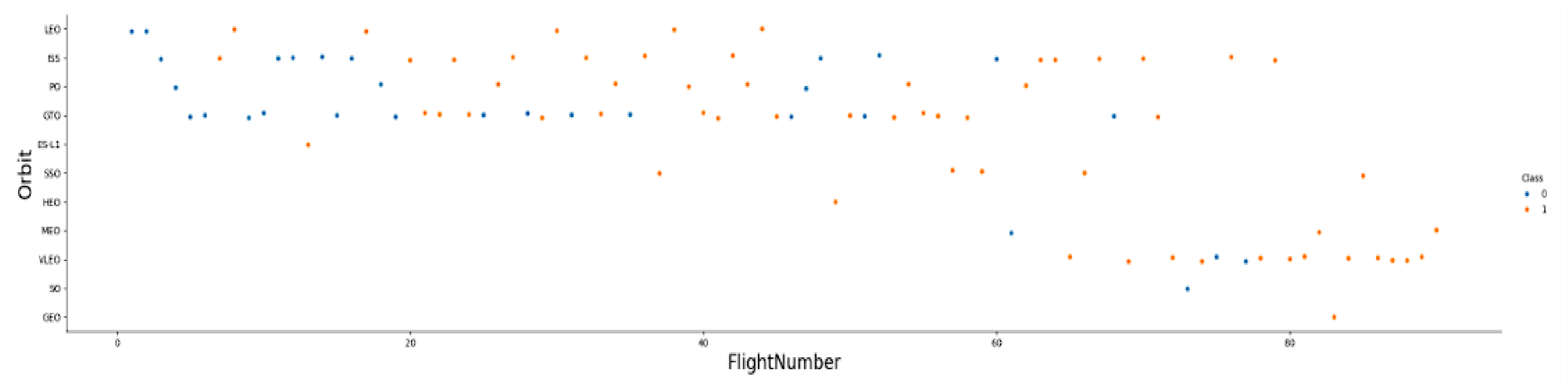
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

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%)

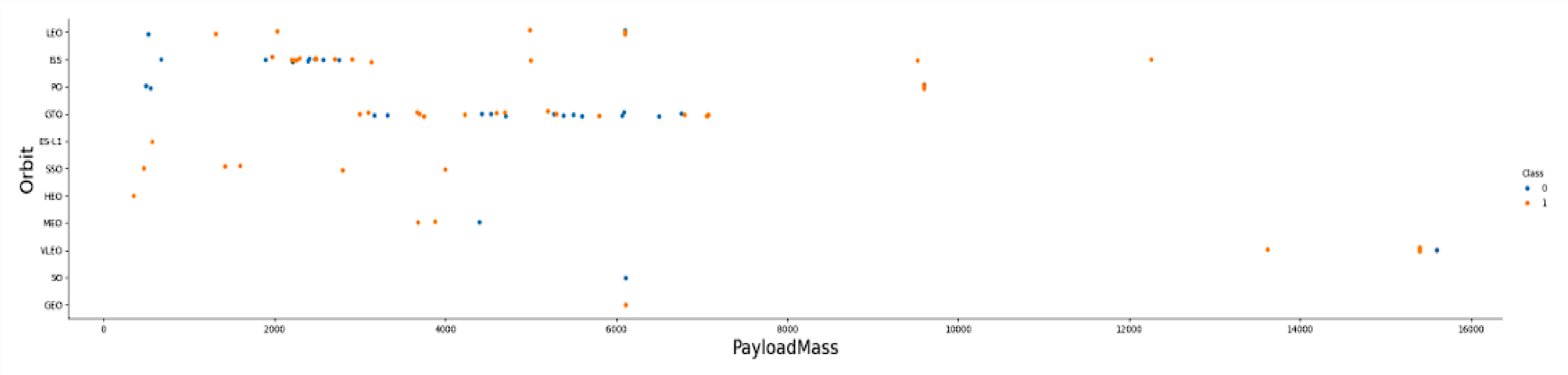


Flight Number vs. Orbit Type



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

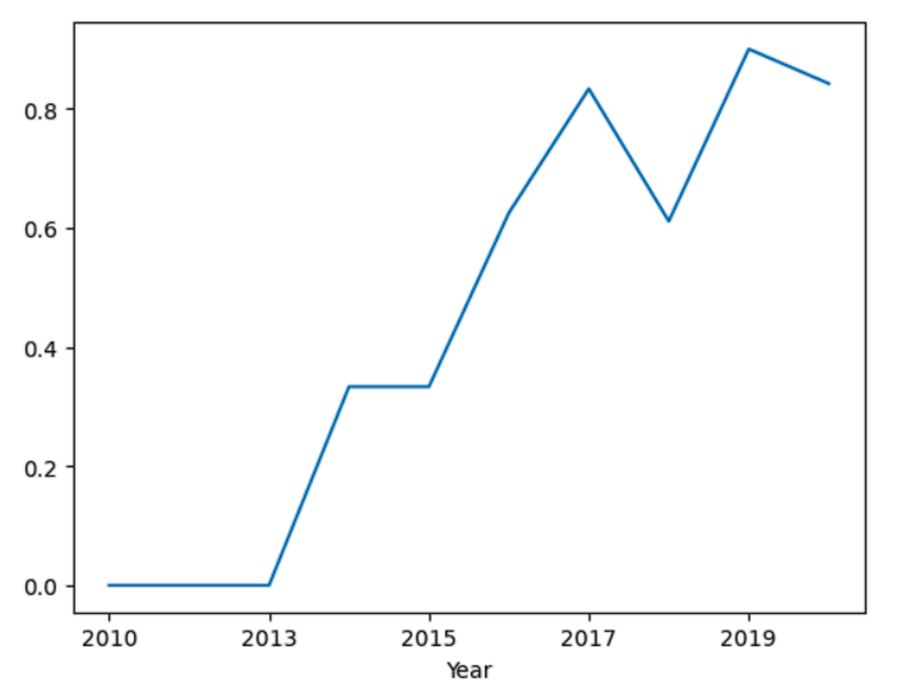
Payload vs. Orbit Type



- 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.
- 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'

- Five records where launch sites begin with `CCA`:

Date	Time (UTC)	Booster Version	Launch_Site	Payload	Payload Mass KG	Orbit	Customer	Mission Outcome	Landing Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Above table are the five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

TOTAL_PAYLOAD
111268

- Total payload calculated above, by summing all payloads whose codes contains “CRS”, which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

AVG_PAYLOAD
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

- First successful landing outcome on ground pad:

FIRST_SUCCESS_GP
2015-12-22

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 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 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

- Number of successful and failure mission outcomes:

Mission_Outcome	QTY
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

- 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
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

- 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

- We used a combinations of the **WHERE** clause, **SUBSTR** to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015.

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, in descending order:

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

- We applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

Section 3

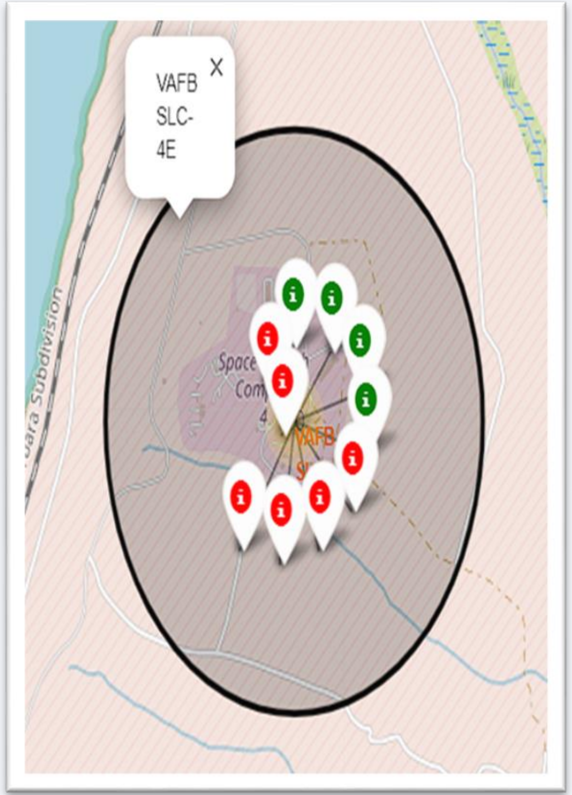
Launch Sites Proximities Analysis

All launch sites

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

Markers showing launch sites with color labels

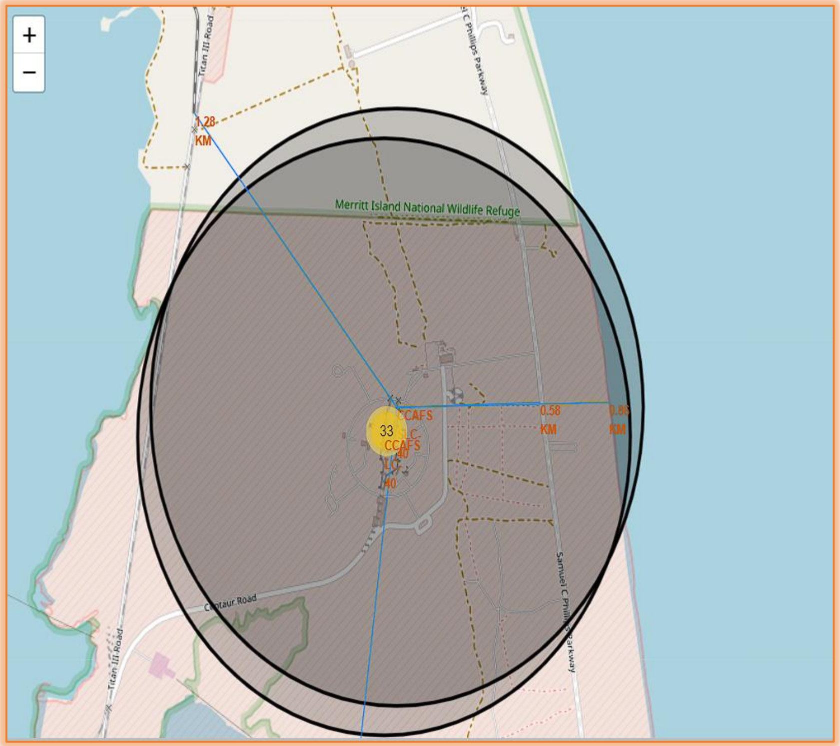
Florida Launch Sites

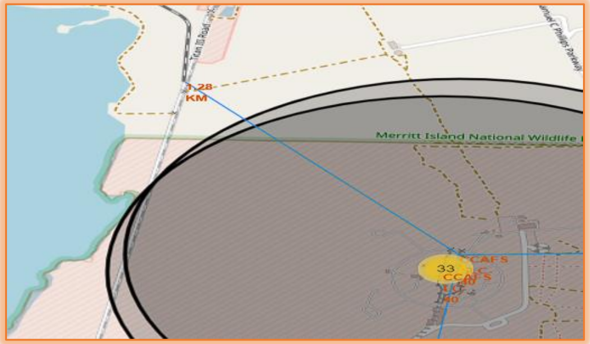


California Launch Site

- Green markers indicate successful and red ones indicate failure.

Launch Site distance to landmarks



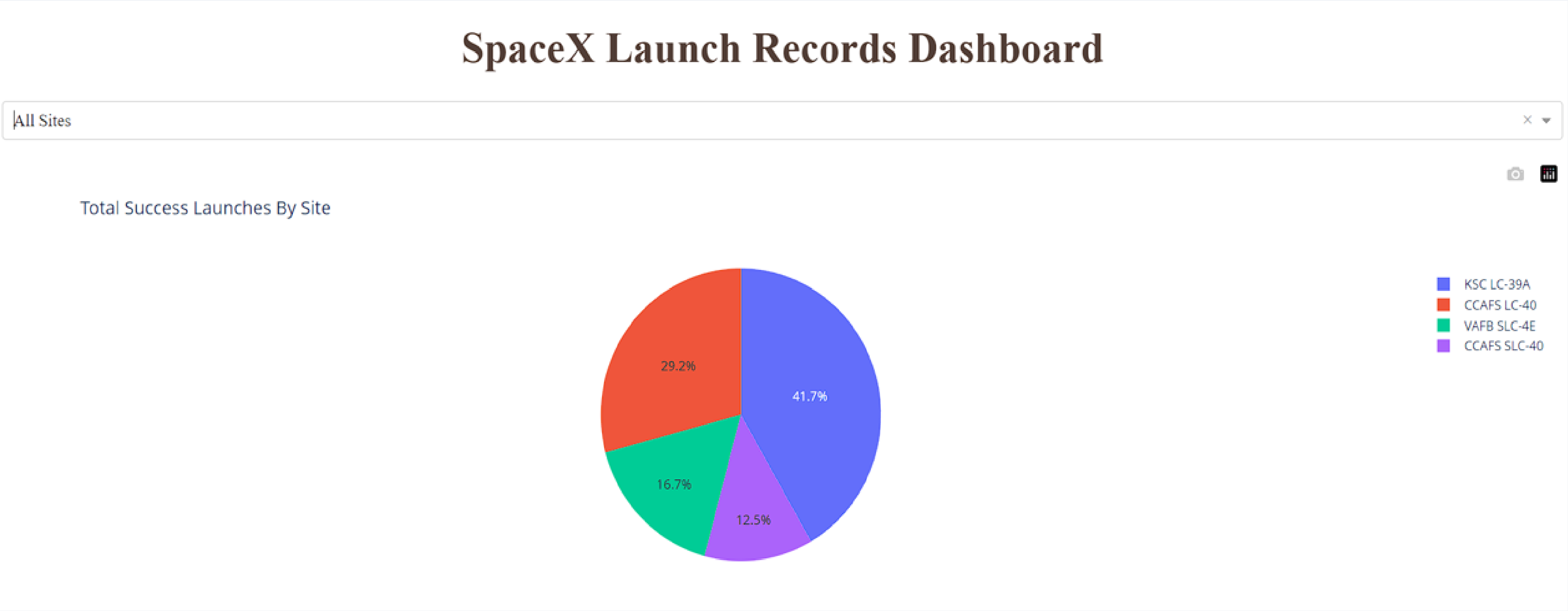


- Launch sites are in close proximity to coastline.
- Launch sites have certain distance away from cities.

Section 4

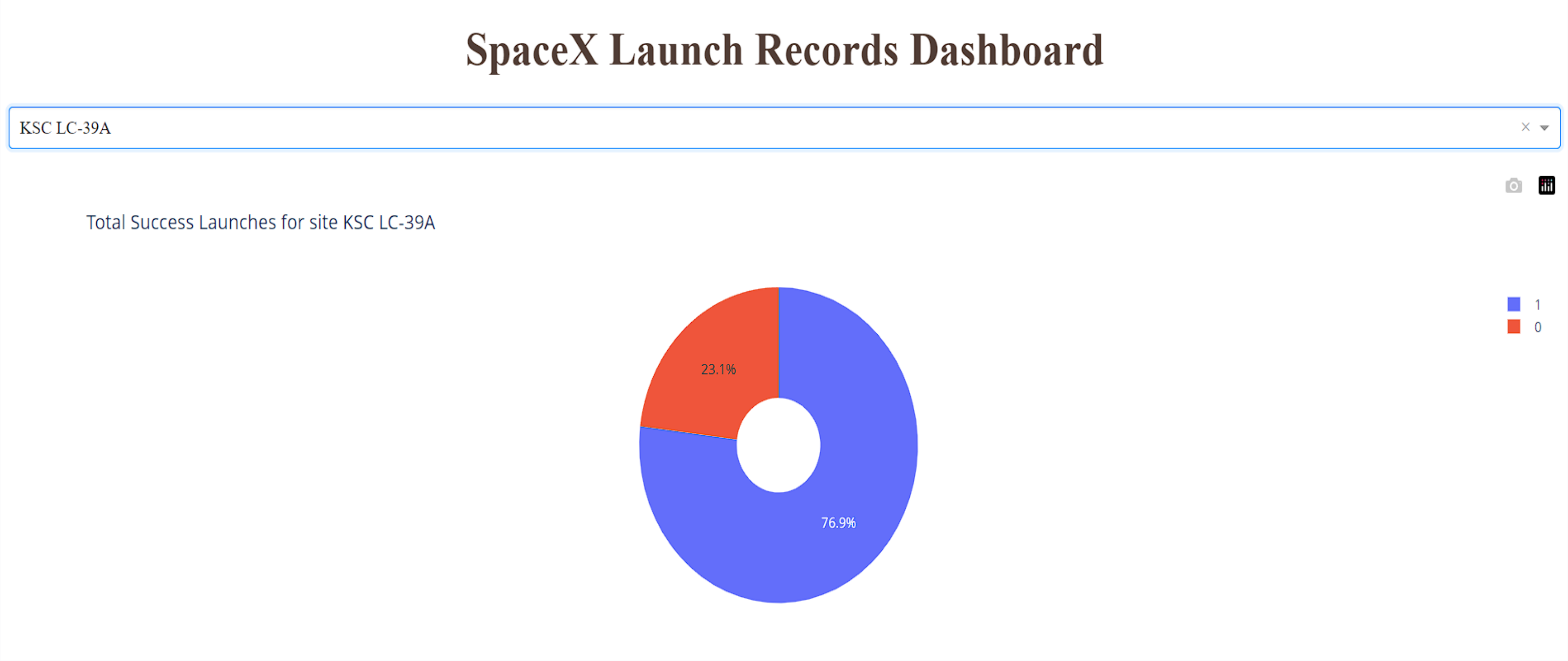
Build a Dashboard with Plotly Dash

Successful Launches by Site



- We can see that KSC LC-39A had the most successful launches from all other sites.

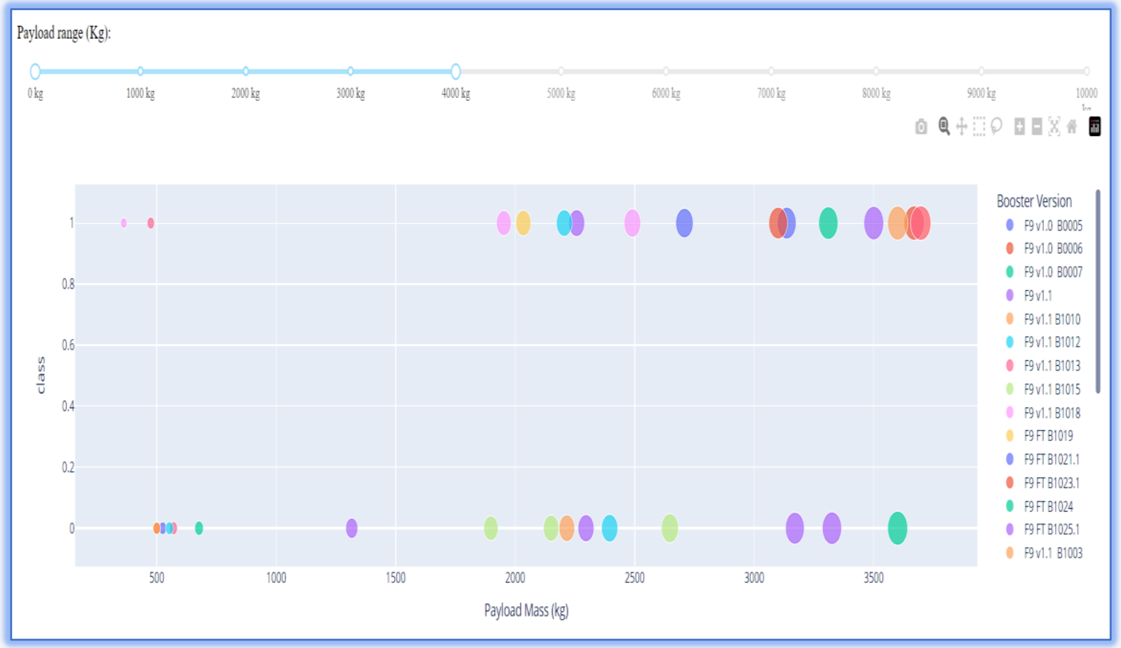
Launch Success Ratio for KSC LC-39A



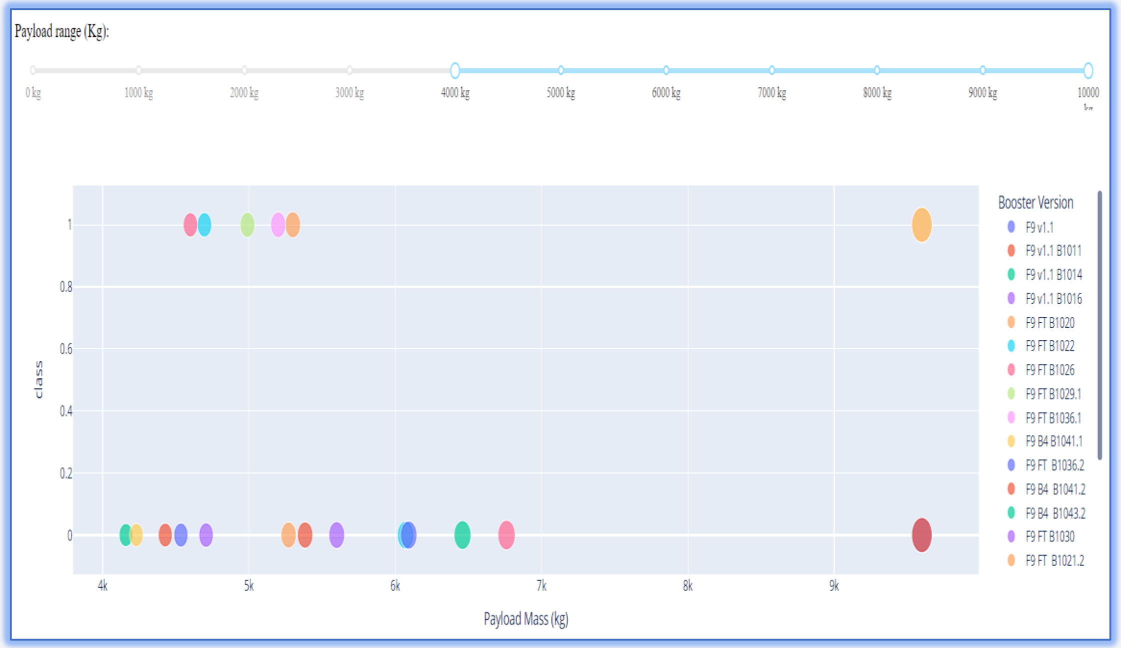
- KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate.

Payload vs. Launch Outcome

Low Weighted Payload 0kg-4000kg



High Weighted Payload 4000kg-10000kg



- We can see the success rates for low weighted payloads is higher than the heavy weighted payloads.

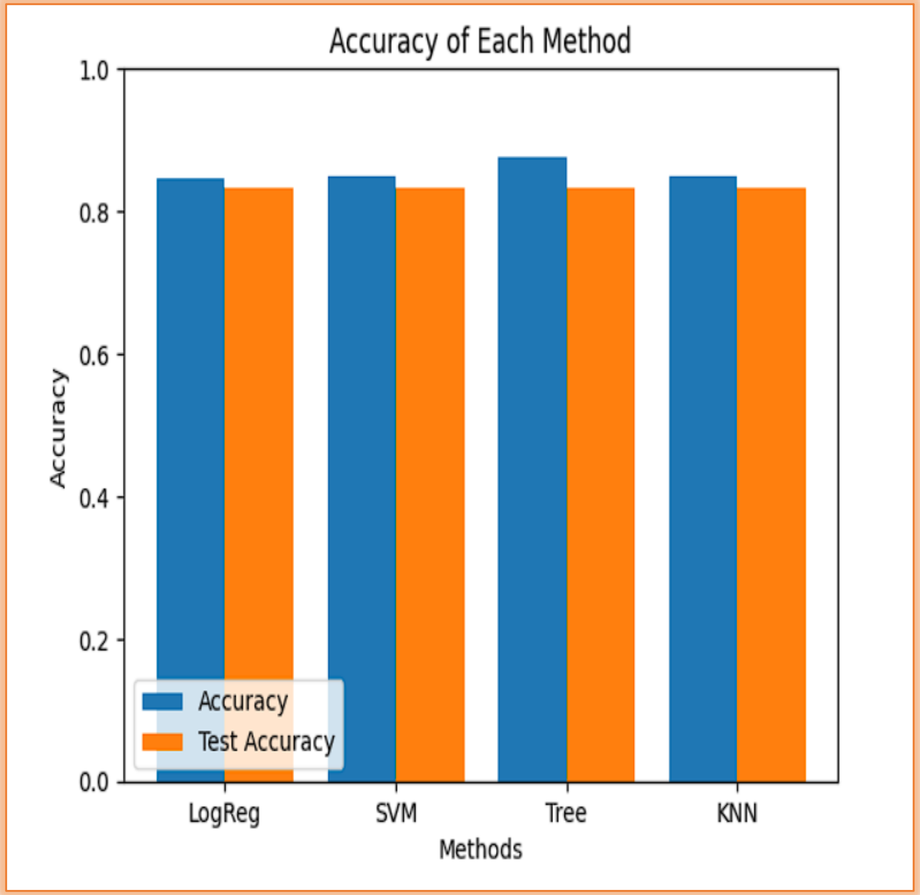
Section 5

Predictive Analysis (Classification)

Classification Accuracy

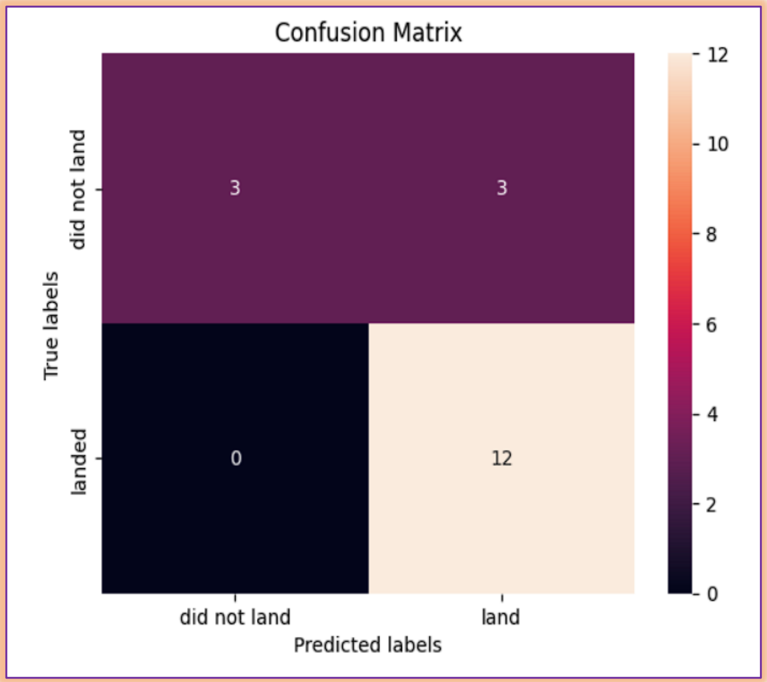
- Four classification models were tested, and their accuracies are plotted beside:

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- Launch success rate started to increase in 2013 till 2020.
- KSC LC-39A had the most successful launches of any sites.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- Different data sources were analyzed, refining conclusions along the process.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.

Appendix

- Folium didn't show maps on GitHub, so I took the screenshots.

Thank you!