IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Linh Nguyen>
<14/09/2023>

Outline

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

Executive Summary

- Summary of methodologies:
  - ✓ Data Collection use web scraping and SpaceX API.
  - ✓ Exploratory Data Analysis (EDA): data wrangling, data visualization, interactive visual analytics.
  - ✓ Prediction analysis with Machine Learning
- Summary of all results
  - ✓ Valuable data was successfully gathered from publicly available sources.
  - ✓ Exploratory Data Analysis (EDA) enabled the identification of the most influential features for predicting the success of launchings.
  - ✓ The Machine Learning Prediction phase revealed the optimal model for determining which characteristics play a crucial role in maximizing the success of this opportunity, leveraging the entirety of the collected data.

Introduction

- Project background and context:

  In this capstone project, our goal is to predict the successful landing of the Falcon 9 first stage. SpaceX prominently features Falcon 9 rocket launches on its website, offering them at a cost of 62 million dollars, while other providers charge upwards of 165 million dollars for similar launches. A significant portion of the cost savings lies in SpaceX's ability to reuse the first stage of the rocket. Therefore, by determining the likelihood of a successful first-stage landing, we can estimate the overall cost of a launch. This valuable information can be particularly useful if another company wishes to compete with SpaceX in bidding for rocket launch contracts.

- Problems you want to find answers:

  ✓ Effectively estimate the overall launch cost by predicting the successful landing of rocket first stages

  ✓ Factors that effect the successful landing of rocket

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX public API and SpaceX Wikipedia page – 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 -> Classifying true landings as successful and unsuccessful otherwise
- 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 were normalized, divided in training and test data sets and evaluated by four different classification models, the accuracy of each model is 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).

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 URL: https://github.com/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-data-collection-api.ipynb

1. Request API and parse the SpaceX launch data
2. Filter data to only include Falcon 9 launches
3. Deal with Missing Values

Data Collection - Scraping

- Scraping data from source that is listed at previous slide
- GitHub URL: https://github.com/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-data-collection-web-scraping.ipynb

1. Request the Falcon9 Launch Wiki page
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables

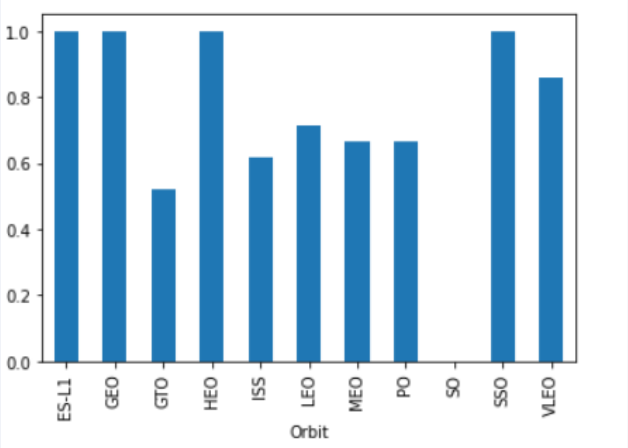
Data Wrangling

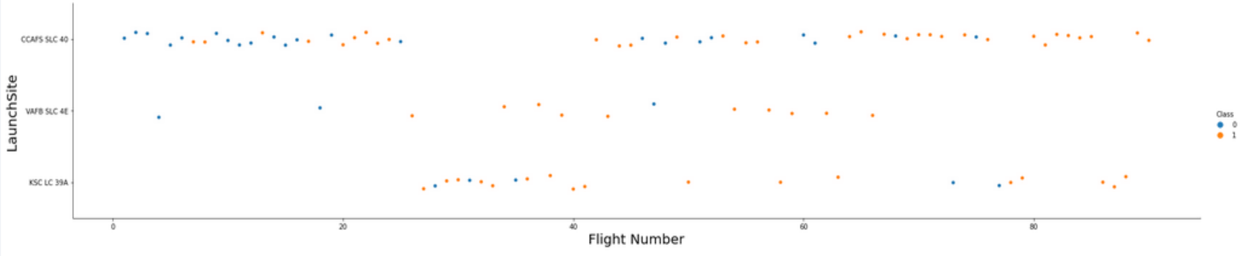
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset -> The summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated -> The landing outcome label was created from Outcome column.
- GitHub URL: https://github.com/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-data-wrangling.ipynb

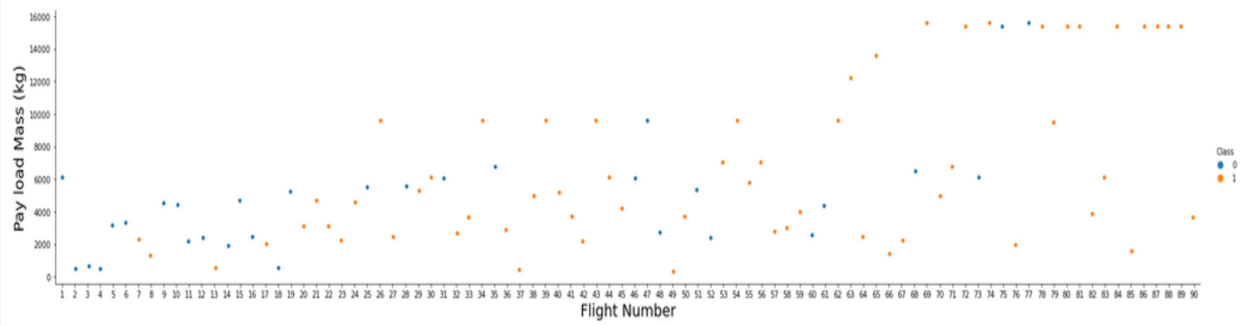
1. EDA
2. Summarization
3. Outcome label

EDA with Data Visualization

- 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/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-EDA.ipynb







EDA with SQL

- 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 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/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-EDA-SQL.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/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-IVA-Folium.ipynb

Build a Dashboard with Plotly Dash

- Use graphs and plots about Percentage of launches by site and Payload range to visualize data -> 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/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-Dash.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/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Spacex-Machine-Learning-Prediction.ipynb

1
- Data preparation and standardization

2
- Test of each model with combinations of hyperparameters

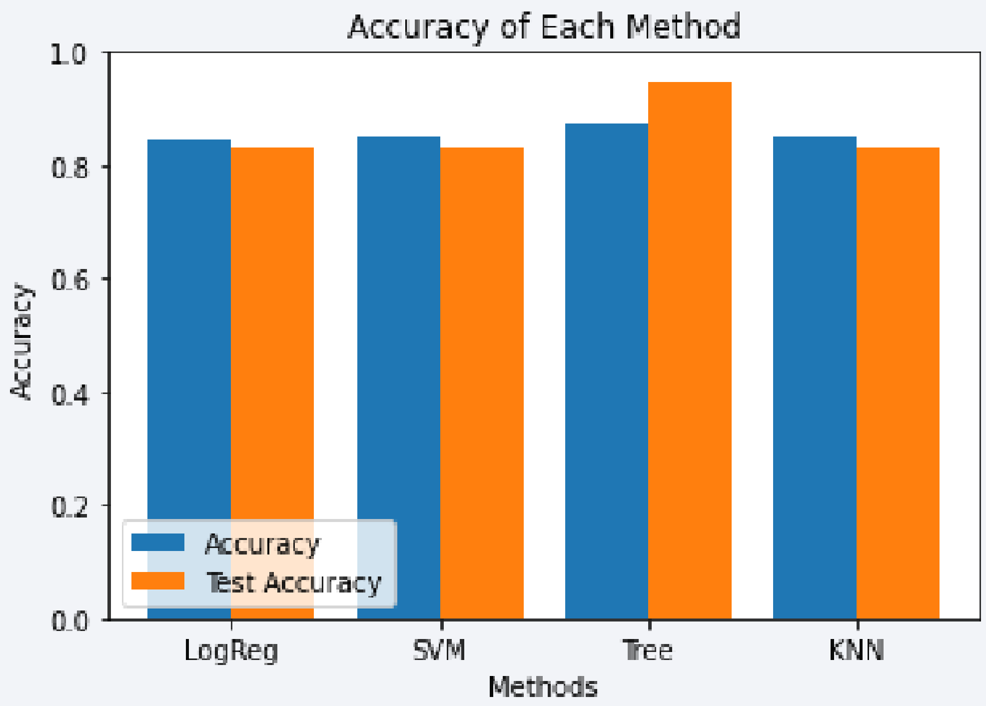
3
- Comparison of results

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;
- Launch sites use to be in safety places, like near seaand have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

Results

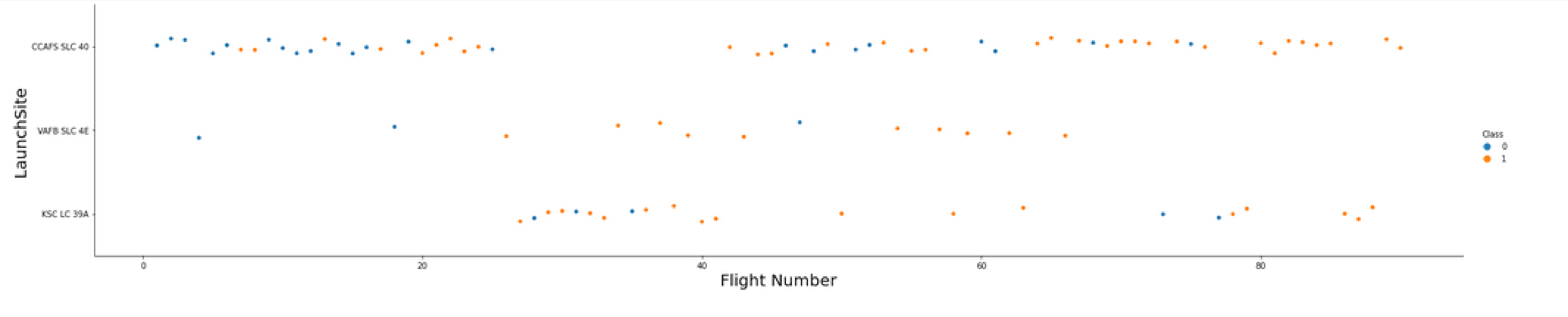
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



Section 2

Insights drawn from EDA

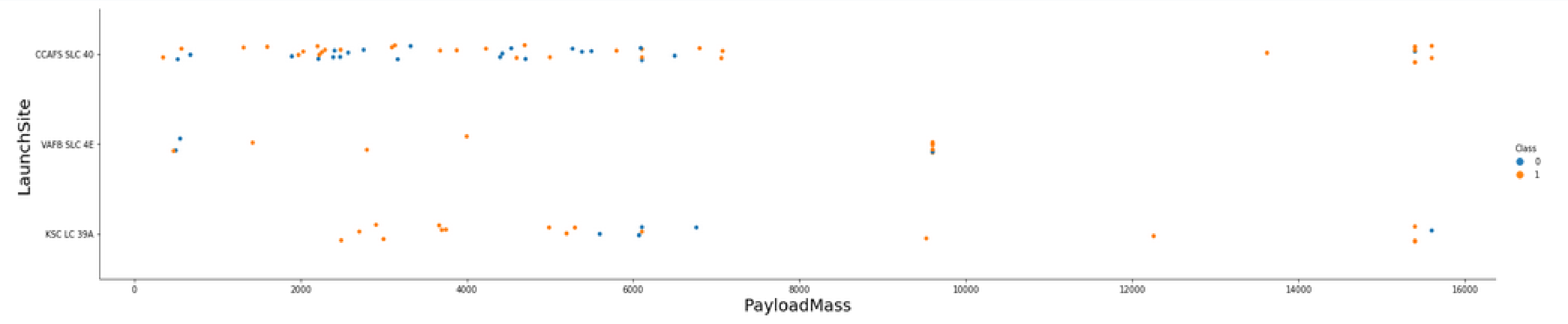
Flight Number vs. Launch Site



Explanation:

- 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



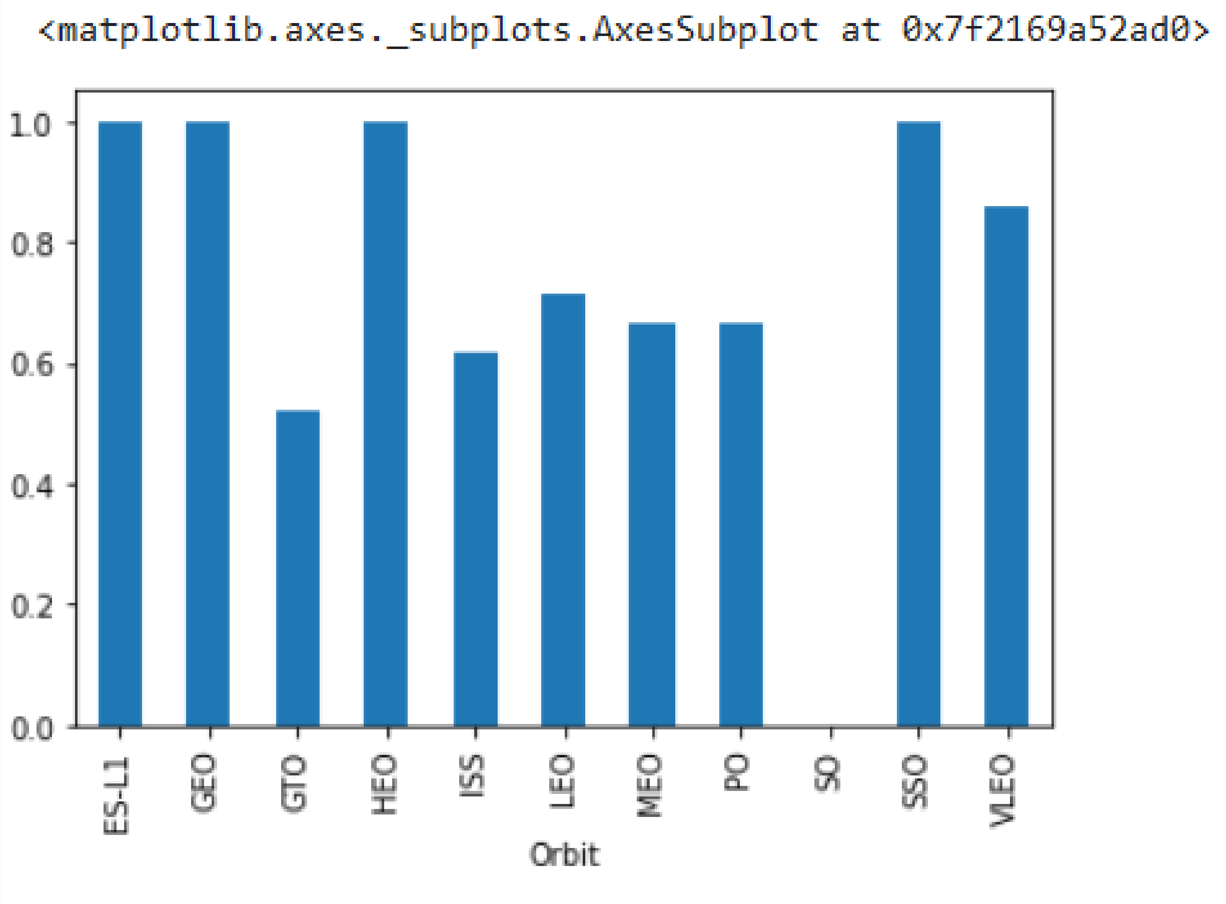
Explanation:

- 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..

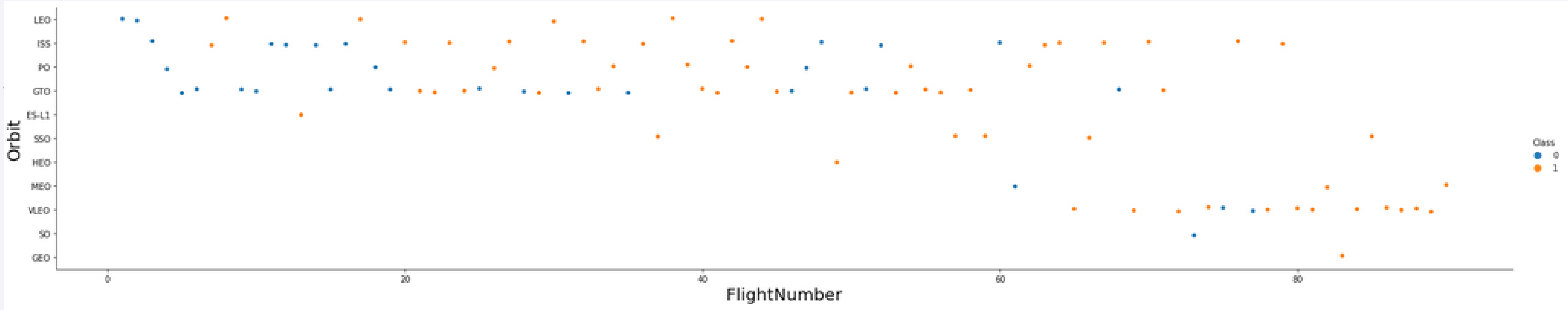
Success Rate vs. Orbit Type

Bar chart for the success rate of each orbit type

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



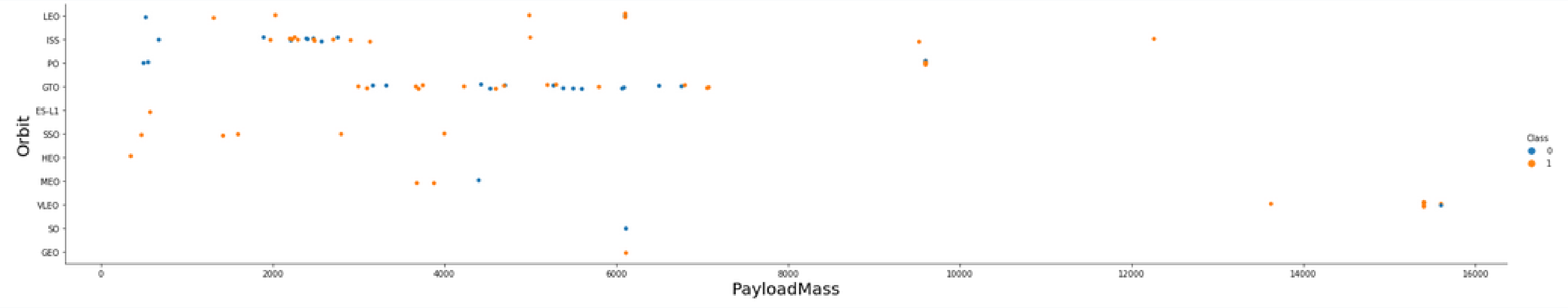
Flight Number vs. Orbit Type



Explanation:

- Success rate improved over time to all orbits;
- We can consider VLEO orbit as a new business opportunity, due to recent increase of its frequency

Payload vs. Orbit Type



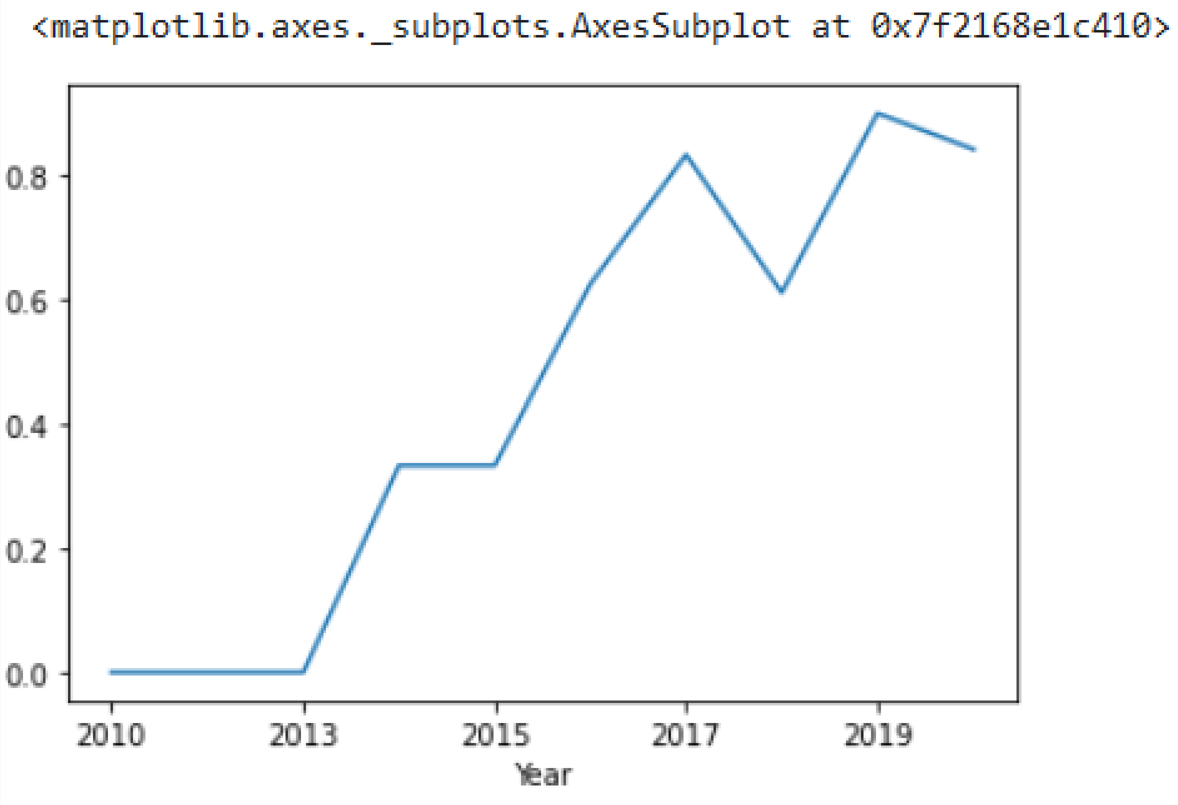
Explanation:

- 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

Explanation:

- Success rate started increasing in 2013 and kept until 2020;
- The first three years were a period of adjusts and improvement of technology since the success rate is evently and significant low.



All Launch Site Names

- There are 04 unique launch sites:

✓CCAFS LC-40

✓CCAFS SLC-40

✓KSC LC-39A

✓VAFB SLC-4E

Out[8]:

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'

Out[9]:

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 attempt

The table above is extracted from SQL query result. we can see five samples of Cape Canaveral launches

Total Payload Mass

- Total payload carried by boosters from NASA
- Query result with a short explanation:

Out[10]:

total_payload
111268

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

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
- Query result with a short explanation

Out[11]:

avg_payload
2928

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

- The dates of the first successful landing outcome on ground pad
- Query result with a short explanation here

Out[13]:

first_success_gp
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
- Query result with a short explanation

Out[14]:

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.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- Query result with a short explanation:

Out[15]:

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query result with a short explanation: These are the boosters which have carried the maximum payload mass registered in the dataset

Out[23]:

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

- Query result:

Out[24]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Query result with a short explanation: This view of data alerts us that “No attempt” must be taken in account

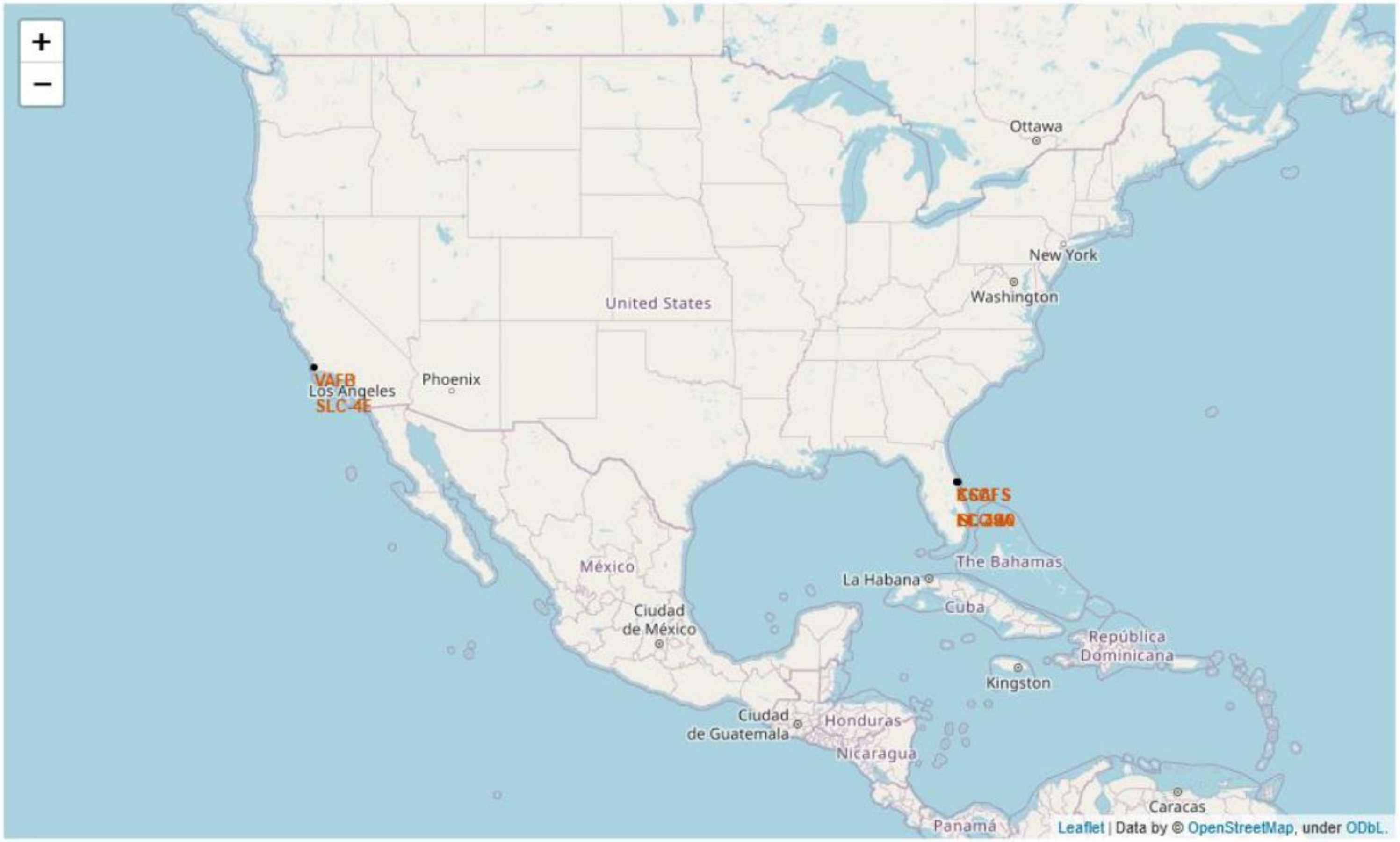
Out[25]:

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

Section 3

Launch Sites Proximities Analysis

All launch sites

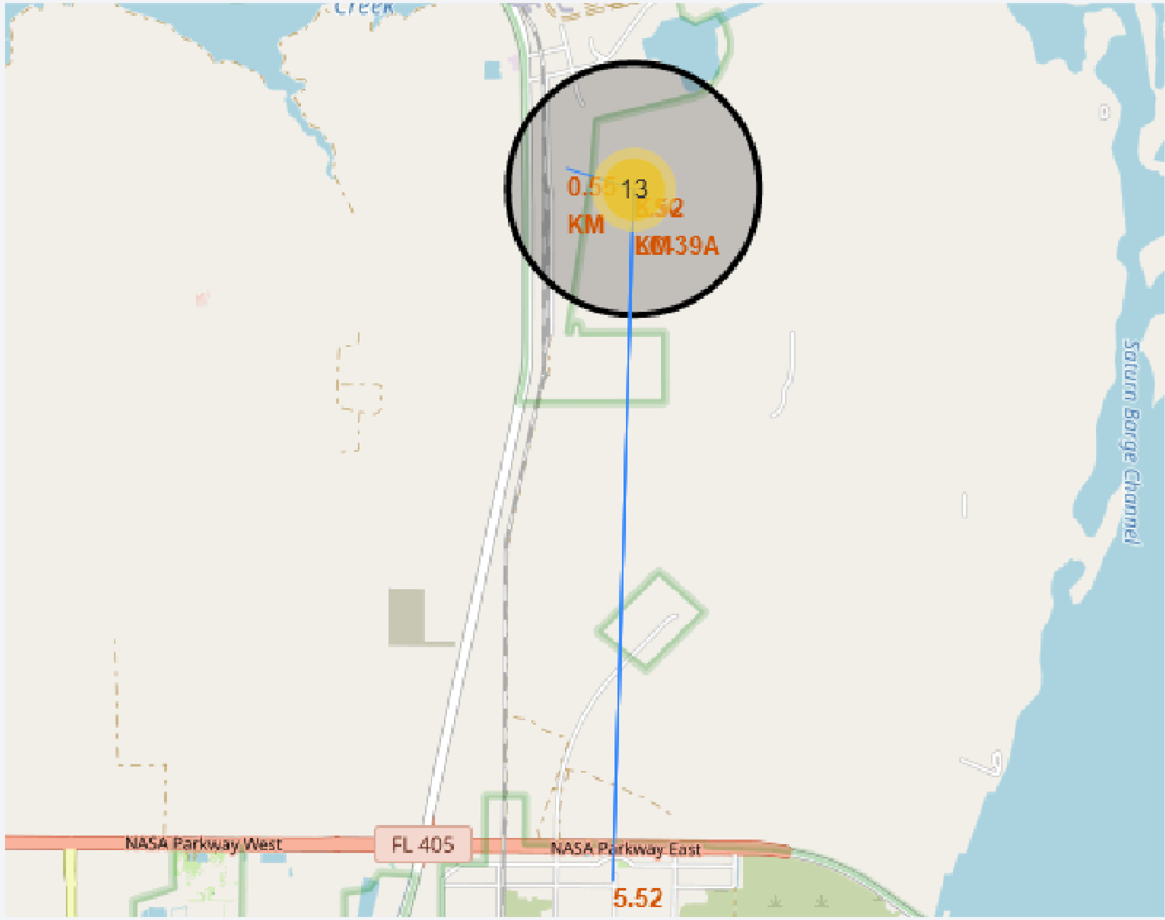


Launch sites are near sea, probably by safety, but not too far from roads and railroads

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

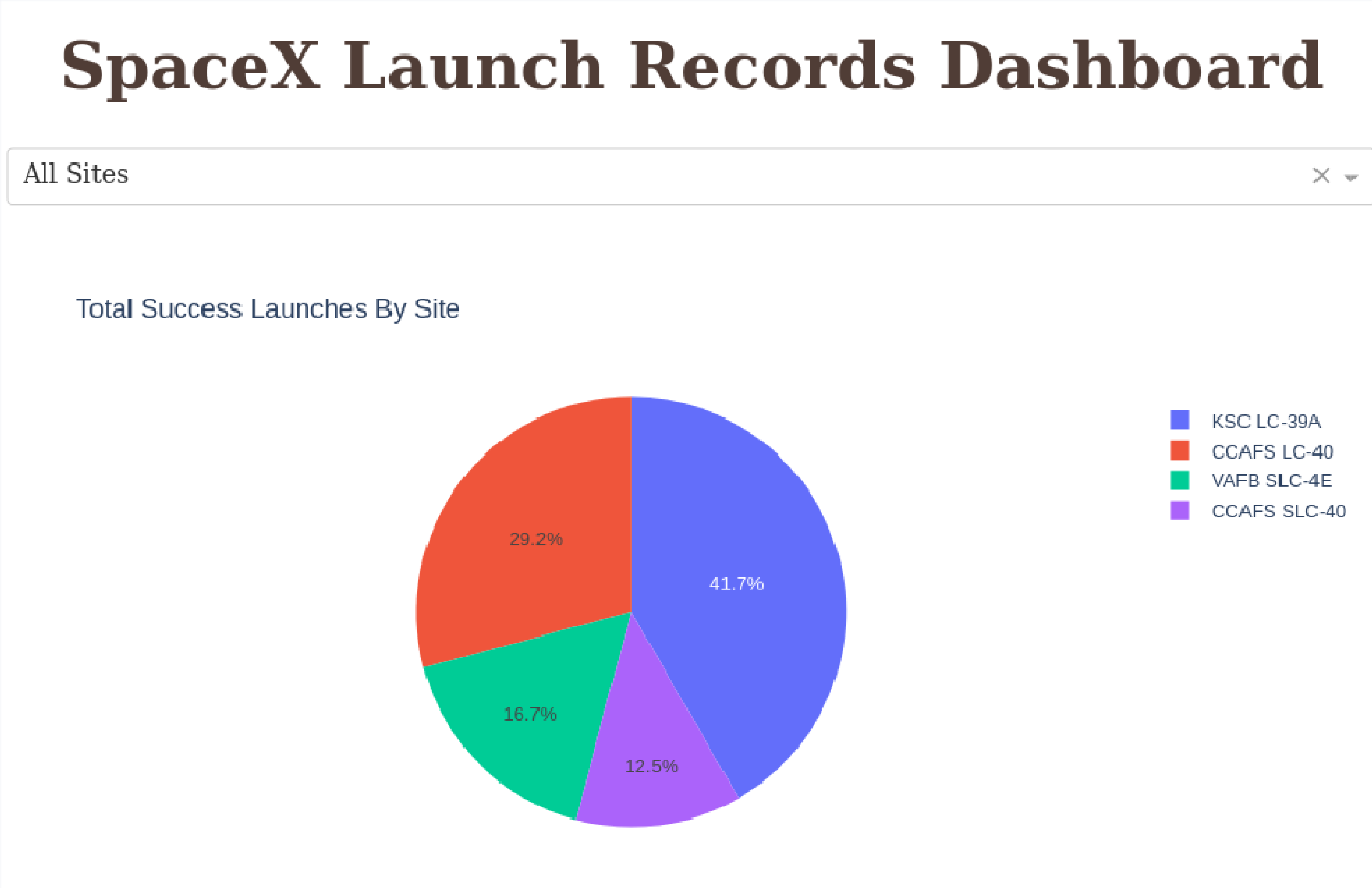


Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas

Section 4

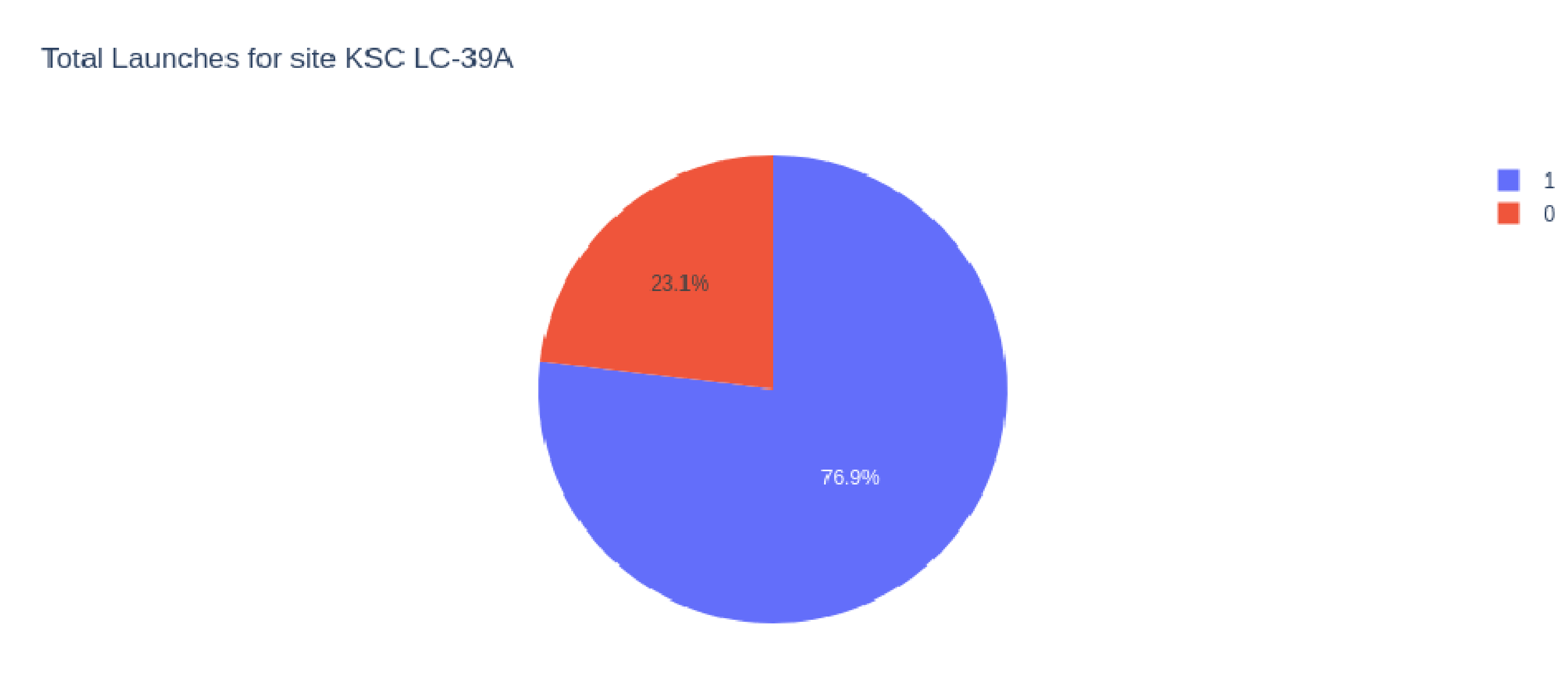
Build a Dashboard with Plotly Dash

Successful Launches by Site



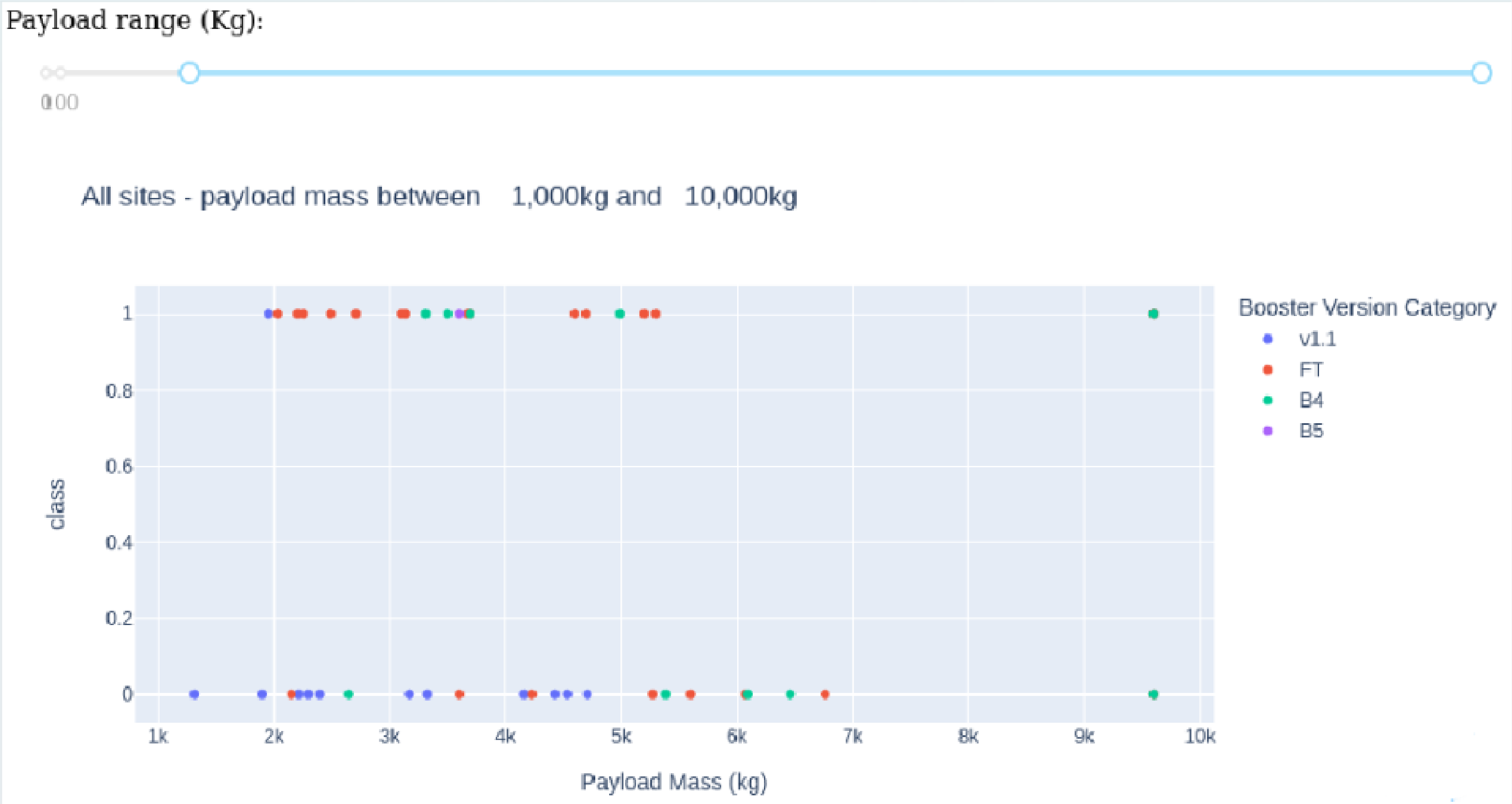
The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



76.9% of launches are successful in this site.

Payload vs. Launch Outcome



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

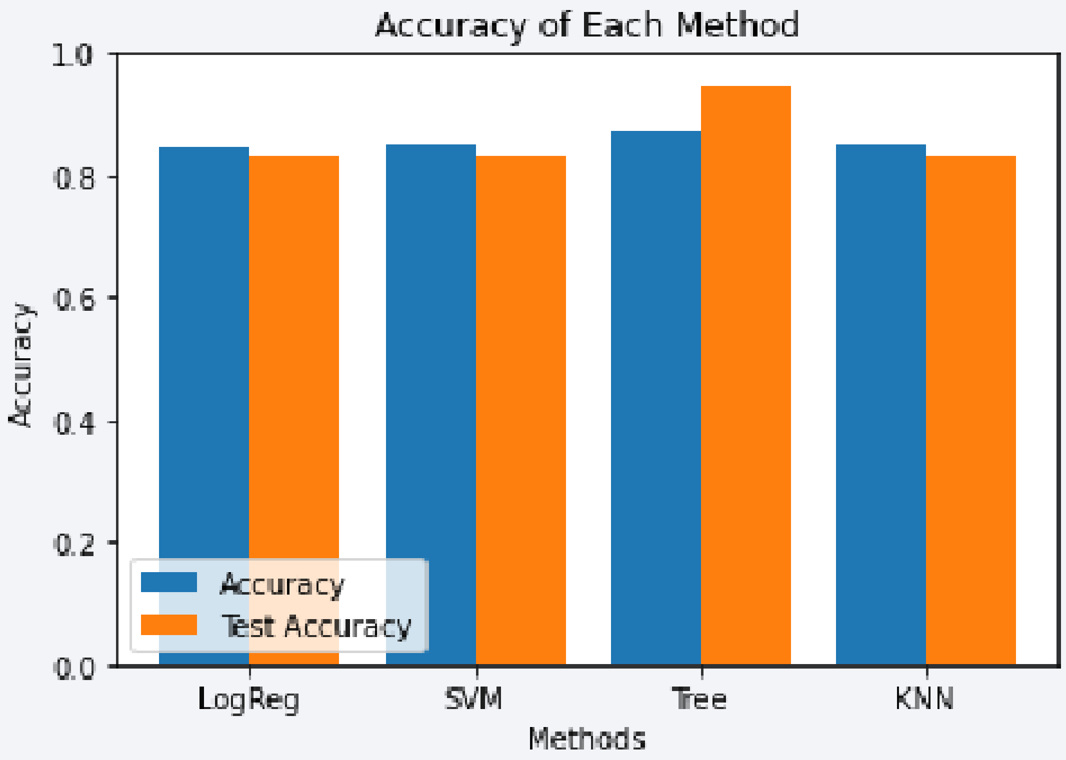
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- 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

- Folium maps can’t show in GitHub, please find screenshot of Folium maps in folder Folium on my GitHub page in following link (I name the pictures with the exactly line numbers in file “Spacex-IVA-Folium”): https://github.com/LinhNguyen-MyLi/Applied-Data-Science-Capstone-IBM/blob/main/Folium.zip

Thank you!