



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sara Vita
11/08/2023



Outline

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

Executive Summary

- The following methodologies were used to analyze data:
 - Data Collection using web scraping and SpaceX API
 - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
 - Machine Learning Prediction.
- Summary of all results
 - It was possible to collect valuable data from public sources;
 - 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.
- What to achieve:
 - 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

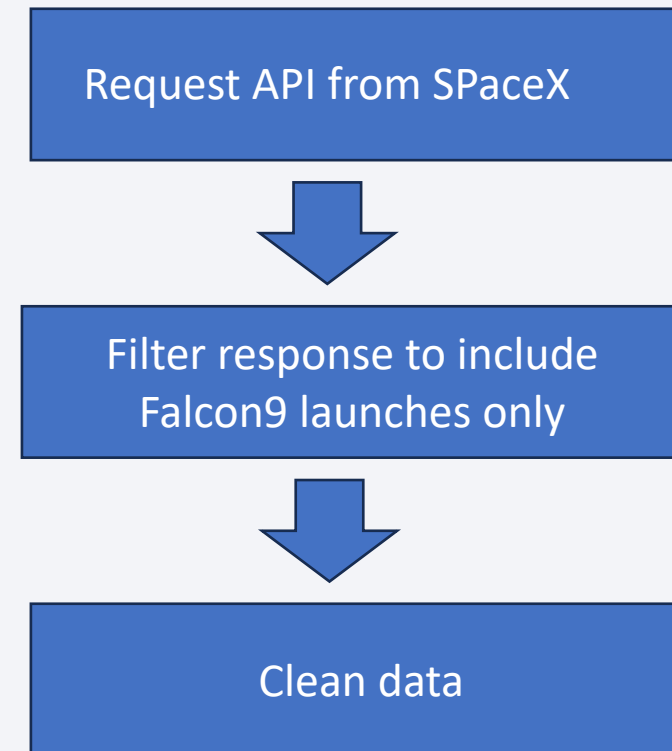
- Data collection methodology:
 - Data from SpaceX was obtained from:
 - SpaceX API: <https://api.spacexdata.com/v4/rockets/>
- Perform data wrangling
 - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Divide data in training and test sets, Used classification models and calculated the accuracy of each.

Data Collection

- Data sets were collected from:
 - SpaceX API: <https://api.spacexdata.com/v4/rockets/>
 - 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 requested and used according to the flowchart beside.

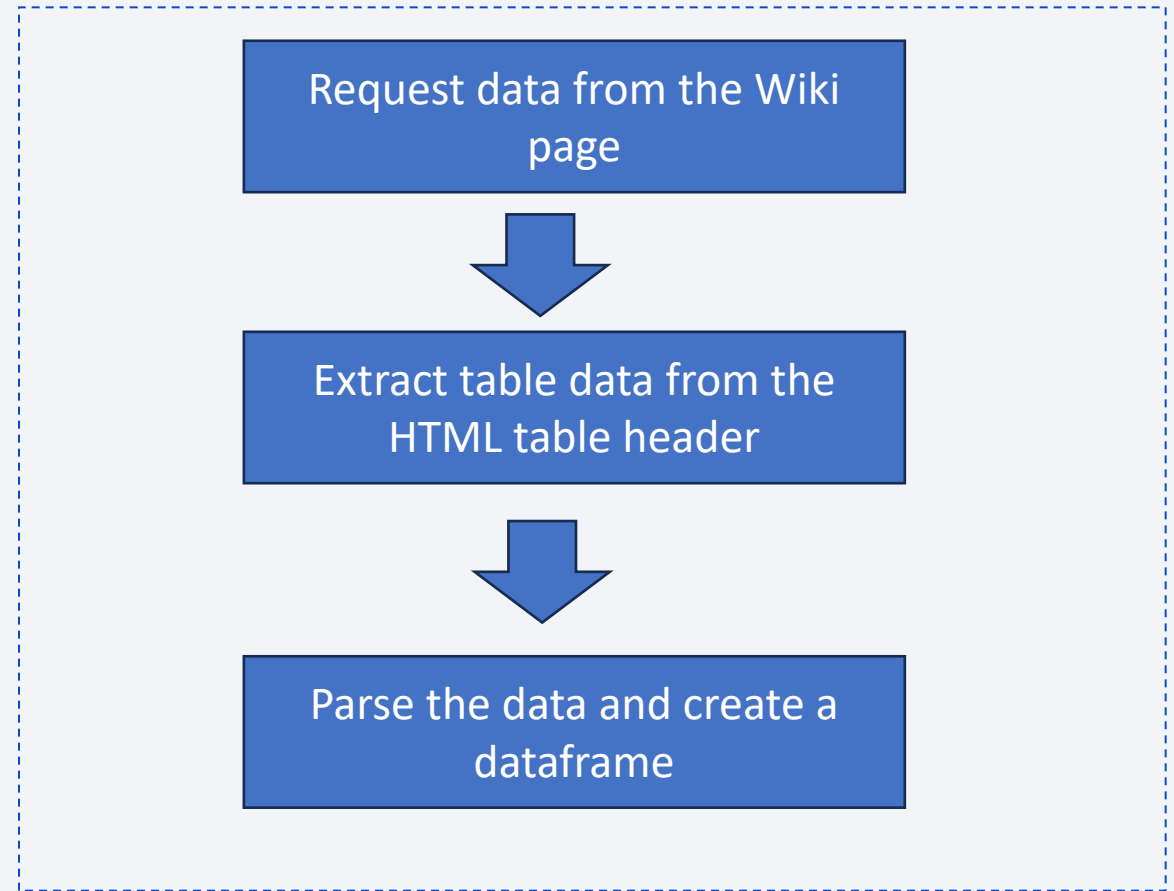


Source code : <https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb>

Data Collection - Scraping

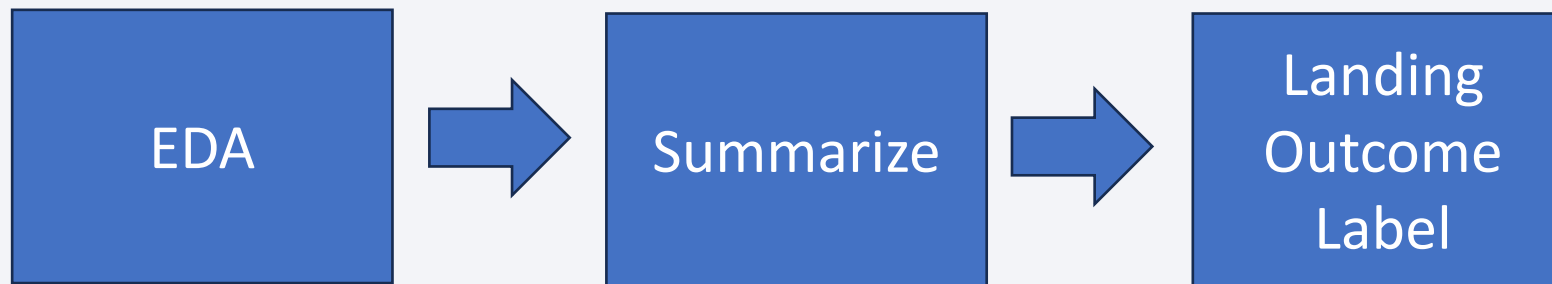
- Data from Wikipedia can be scraped according to the flowchart beside.

Source code: <https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/jupyter-labs-webscraping.ipynb>



Data Wrangling

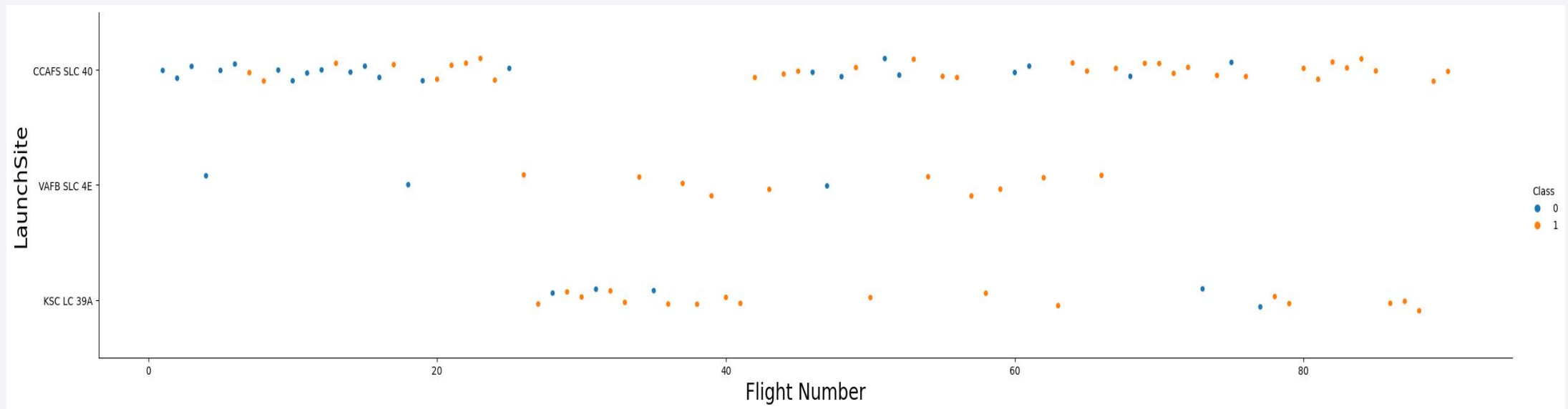
- Exploratory Data Analysis (EDA) was performed first on the dataset
- Summaries launches per site, occurrences of each orbit and mission outcome per orbit were calculated
- Created a landing outcome label



Source code: https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to visualize explored data and the relationship between features
 - (payload, orbit, launch site, etc...)



EDA with SQL

- There were performed a variety of SQL queries, such as:
 - Average payload mass carried by booster version F9 v1.1
 - Total number of successful and failure mission outcomes
 - Count rate of landing outcomes between 2010 and 2017

Source code: https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqlite.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/Sara-Vita/Coursera-Data-Science/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

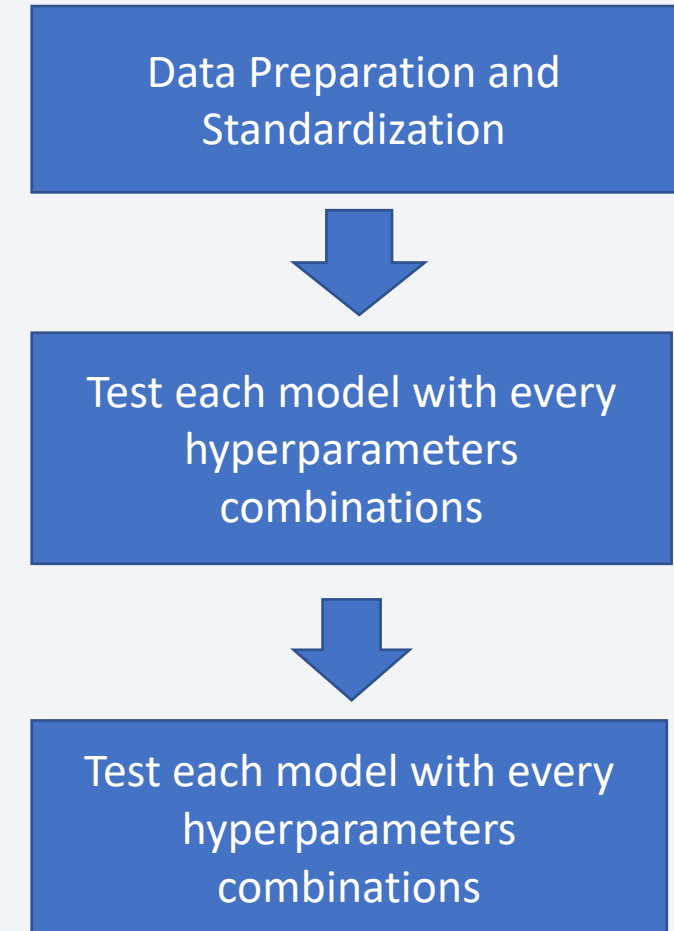
- The following graphs and plots were used to visualize data
 - Percentage of launches by site
 - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source Code: [https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/spacex_dash_app%20\(1\).py](https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/spacex_dash_app%20(1).py)

Predictive Analysis (Classification)

- Four classification models were compared:
 - logistic regression
 - support vector machine (SVM)
 - decision tree
 - k nearest neighbors. (KNN)

Source Code: [https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/Machine%20Learning%20Prediction%20\(1\).ipynb](https://github.com/Sara-Vita/Coursera-Data-Science/blob/main/Machine%20Learning%20Prediction%20(1).ipynb)

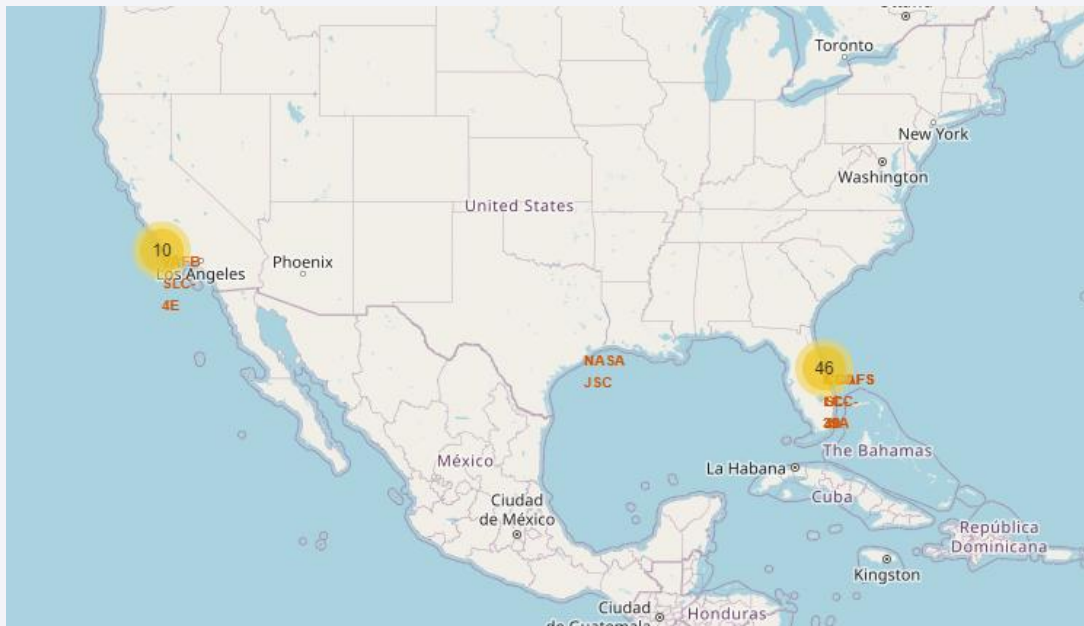


Results

- Exploratory data analysis results:
 - Space X uses 4 different launch sites;
 - 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;
 - 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.



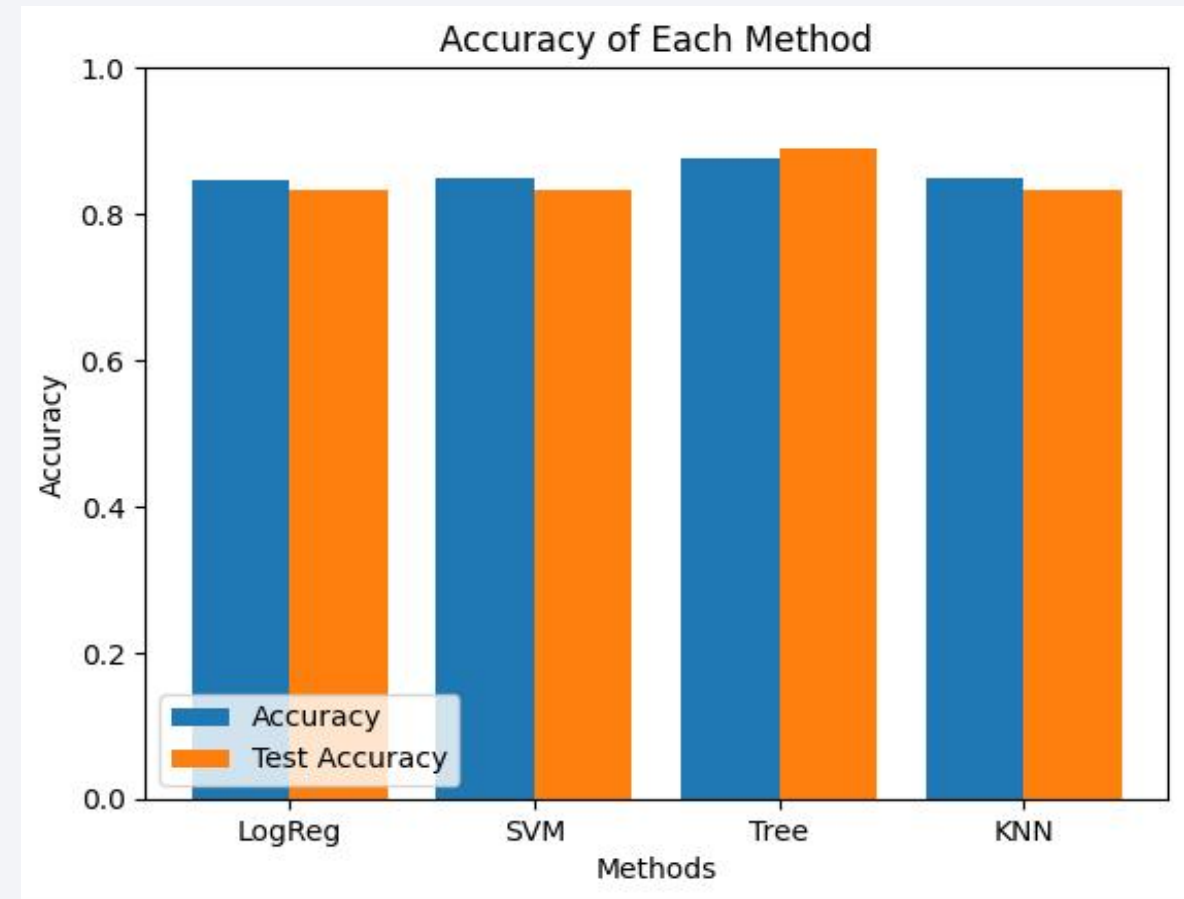
Results

Predictive Analysis showed that:

Decision Tree Classifier

is the best model to predict successful landings,

having accuracy over 87.5% and accuracy for test data of 88%



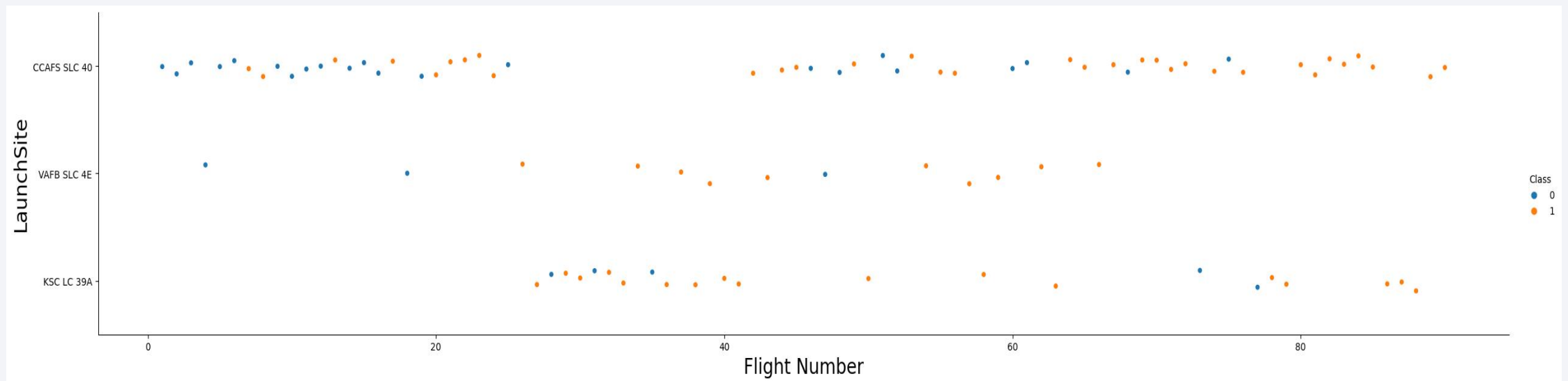
The background of the slide is a complex, abstract composition. It features a dark blue base color on the left, which transitions into a vibrant, multi-colored area on the right. This transition is achieved through a series of diagonal, overlapping bands and streaks in shades of red, teal, and light blue. A fine, grid-like pattern is visible throughout the image, particularly in the teal and red areas, giving it a digital or data-driven appearance. The overall effect is one of dynamic movement and high-tech aesthetics.

Section 2

Insights drawn from EDA

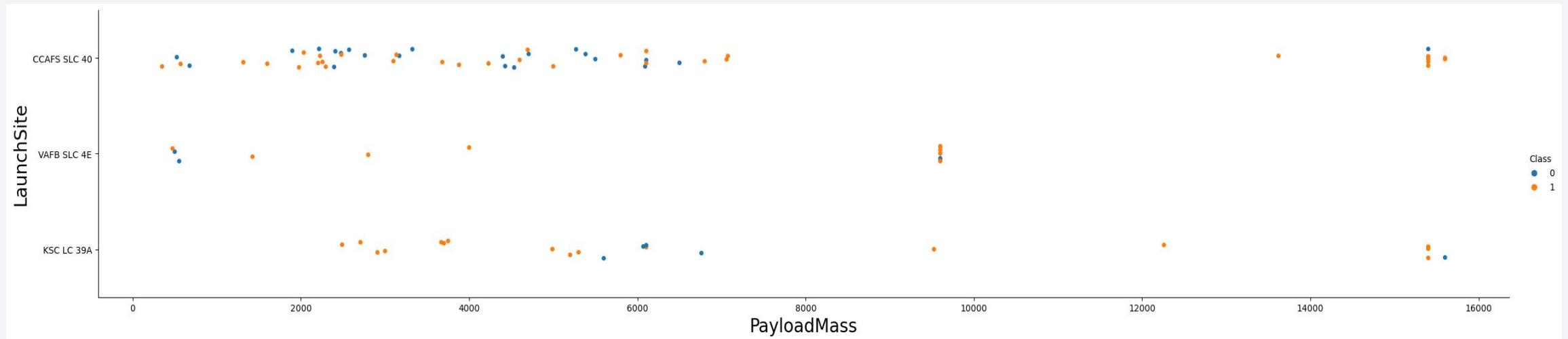
Flight Number vs. Launch Site

- According to the plot above, it's possible to verify that the best launch site nowadays is CCAFS 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.

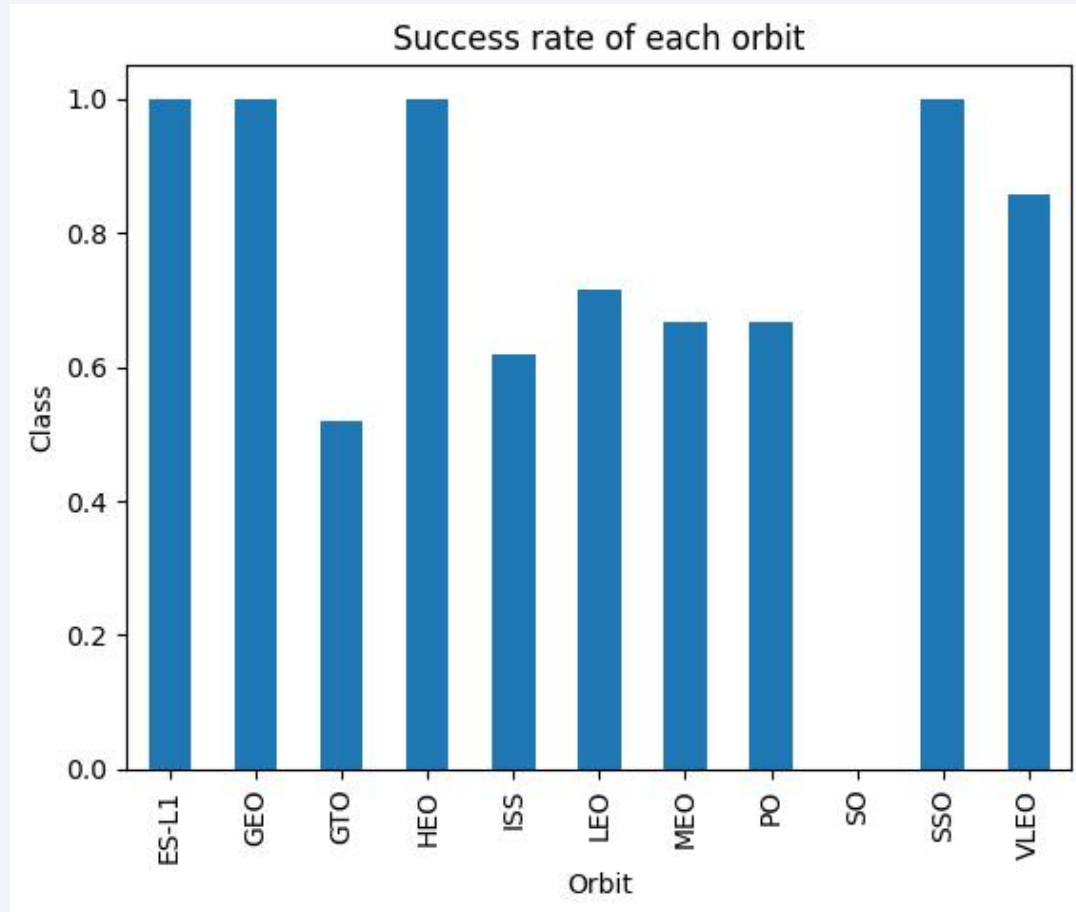


Payload vs. Launch Site

- 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.



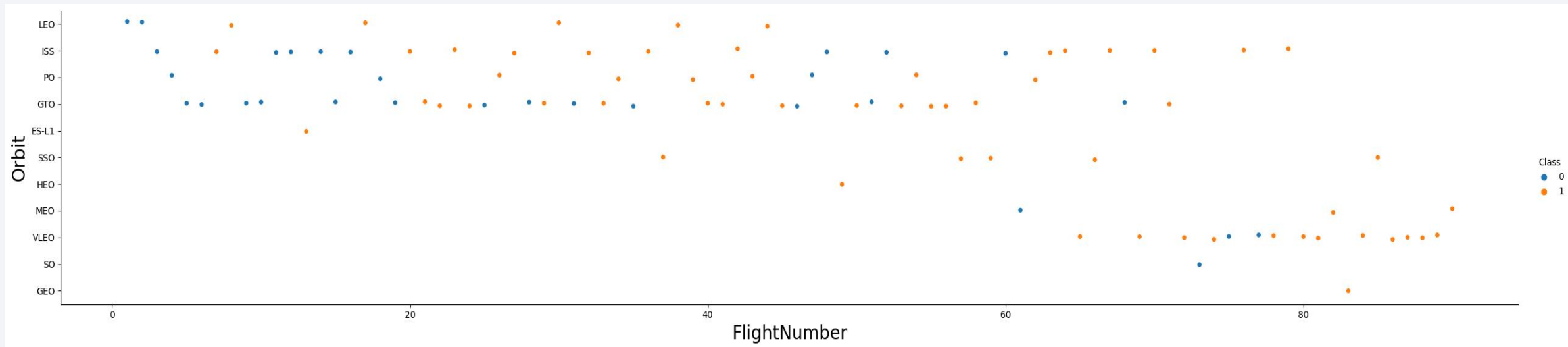
Success Rate vs. Orbit Type



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

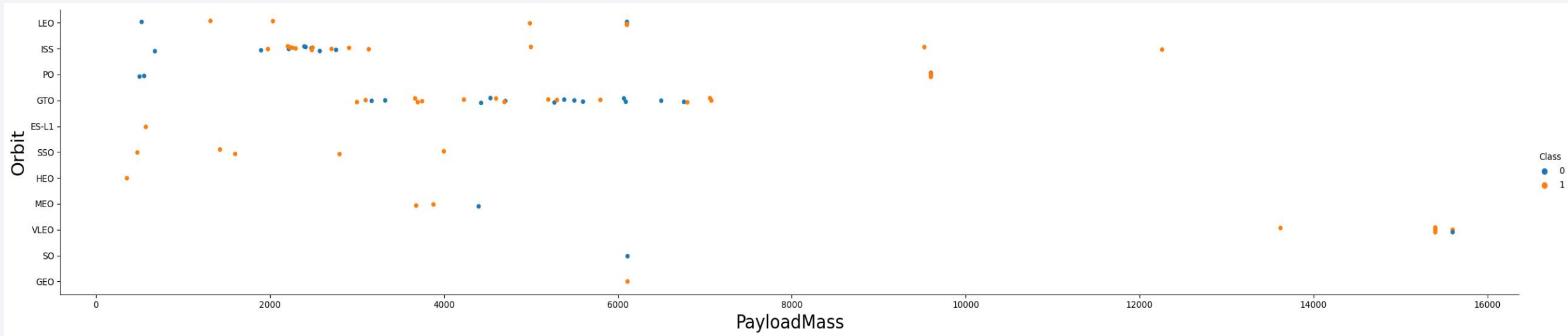
Flight Number vs. Orbit Type

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

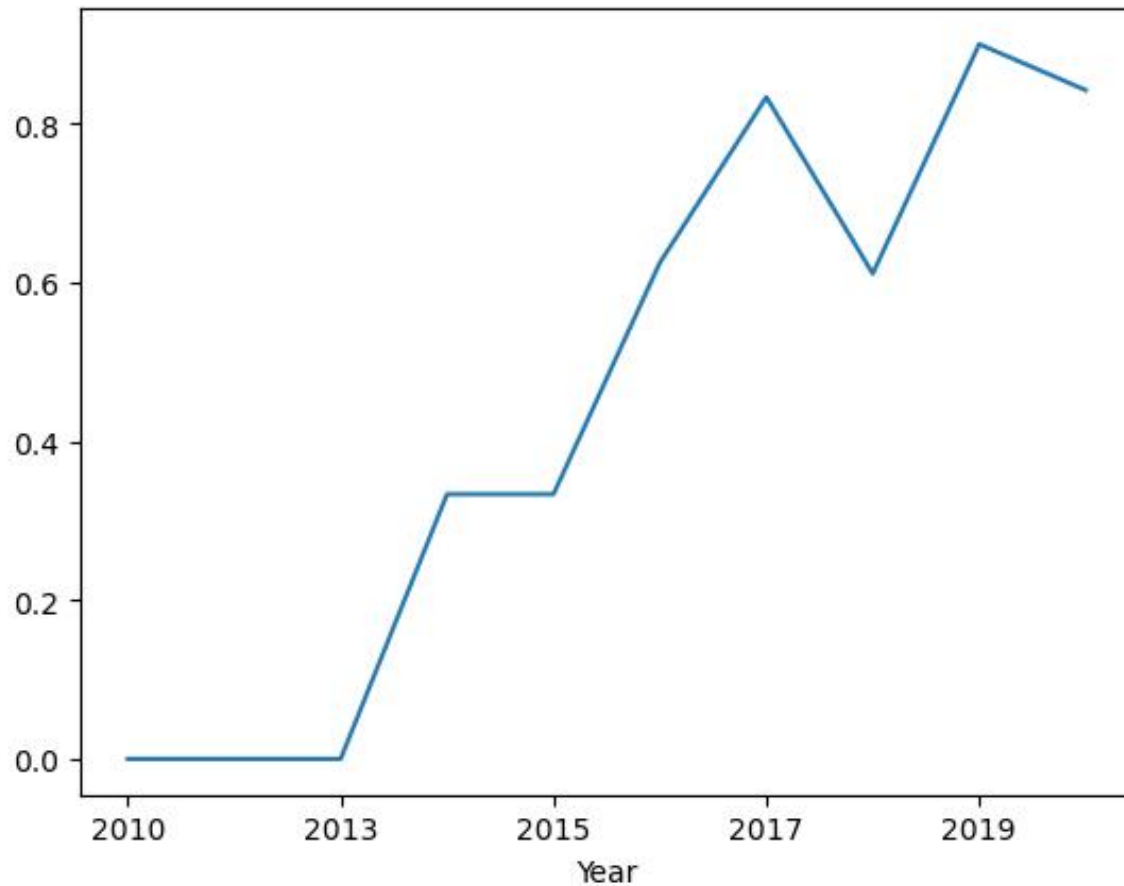


Payload vs. Orbit Type

- 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.



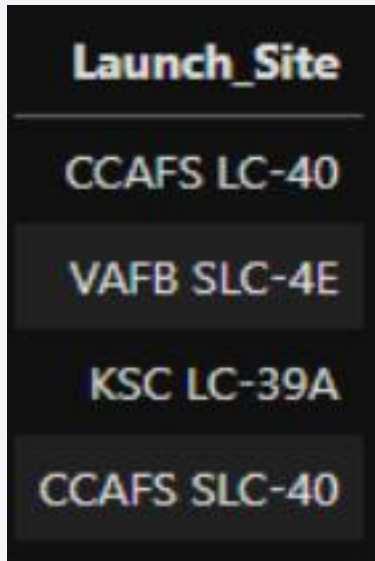
Launch Success Yearly Trend



- 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



A screenshot of a data table with a dark background and light-colored text. The table has a header row labeled 'Launch_Site' and four data rows. The data rows contain the following text: 'CCAFS LC-40', 'VAFB SLC-4E', 'KSC LC-39A', and 'CCAFS SLC-40'.

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'

- 5 records where launch sites begin with 'CCA'
- Here we can see five samples of Cape Canaveral launches.

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

Total Payload Mass

- Total payload carried by boosters from NASA:

```
sum(payload_mass_kg_)
45596
```

- Total payload was calculated by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

Average Payload Mass by F9 v1.1

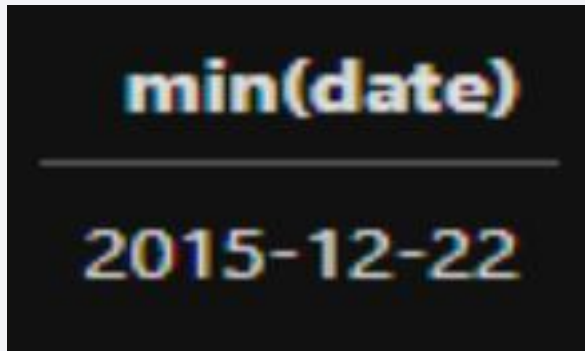
- Average payload mass carried by booster version F9 v1.1:

```
avg(payload_mass_kg_)
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:



```
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 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	count(mission_outcome)
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	PAYLOAD_MASS_KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

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

count(landing_outcome)	Landing_Outcome
10	No attempt
5	Success (ground pad)
5	Success (drone ship)
5	Failure (drone ship)
3	Controlled (ocean)
2	Uncontrolled (ocean)
1	Precluded (drone ship)
1	Failure (parachute)

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky with stars and a view of the Earth's surface from space. The Earth's surface is mostly dark blue, with a thin layer of white clouds. A curved horizon line separates the dark sky from the Earth's surface. In the lower right, there are bright, glowing yellow and orange lights, likely representing city lights or industrial activity. The overall image has a high-contrast, cinematic quality.

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

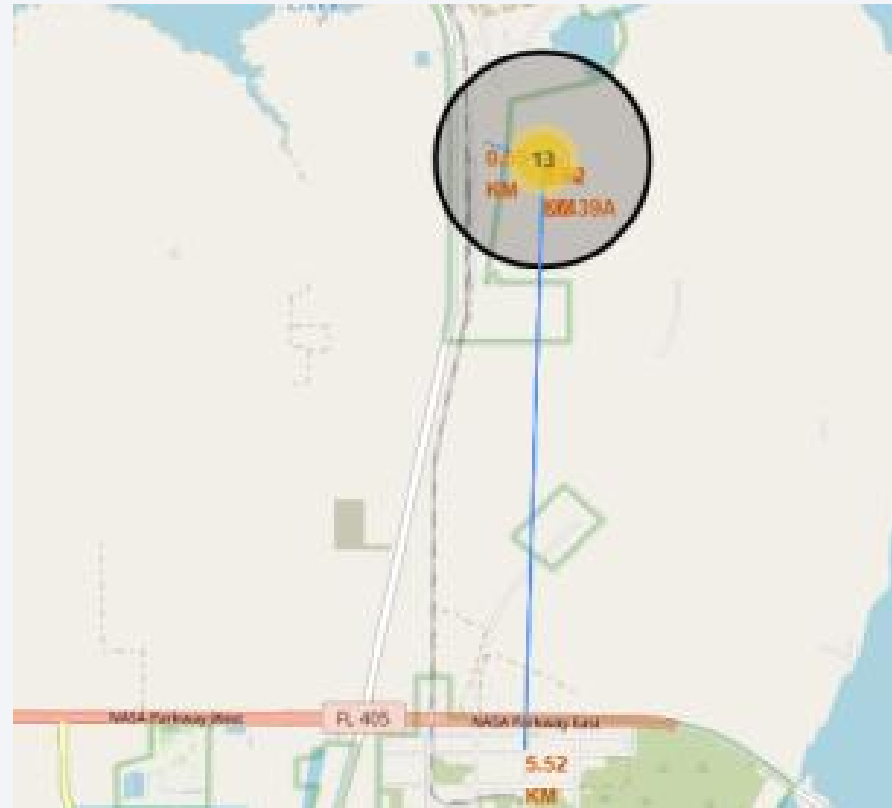
- Example of KSC LC-39A launch site launch outcomes



- 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.



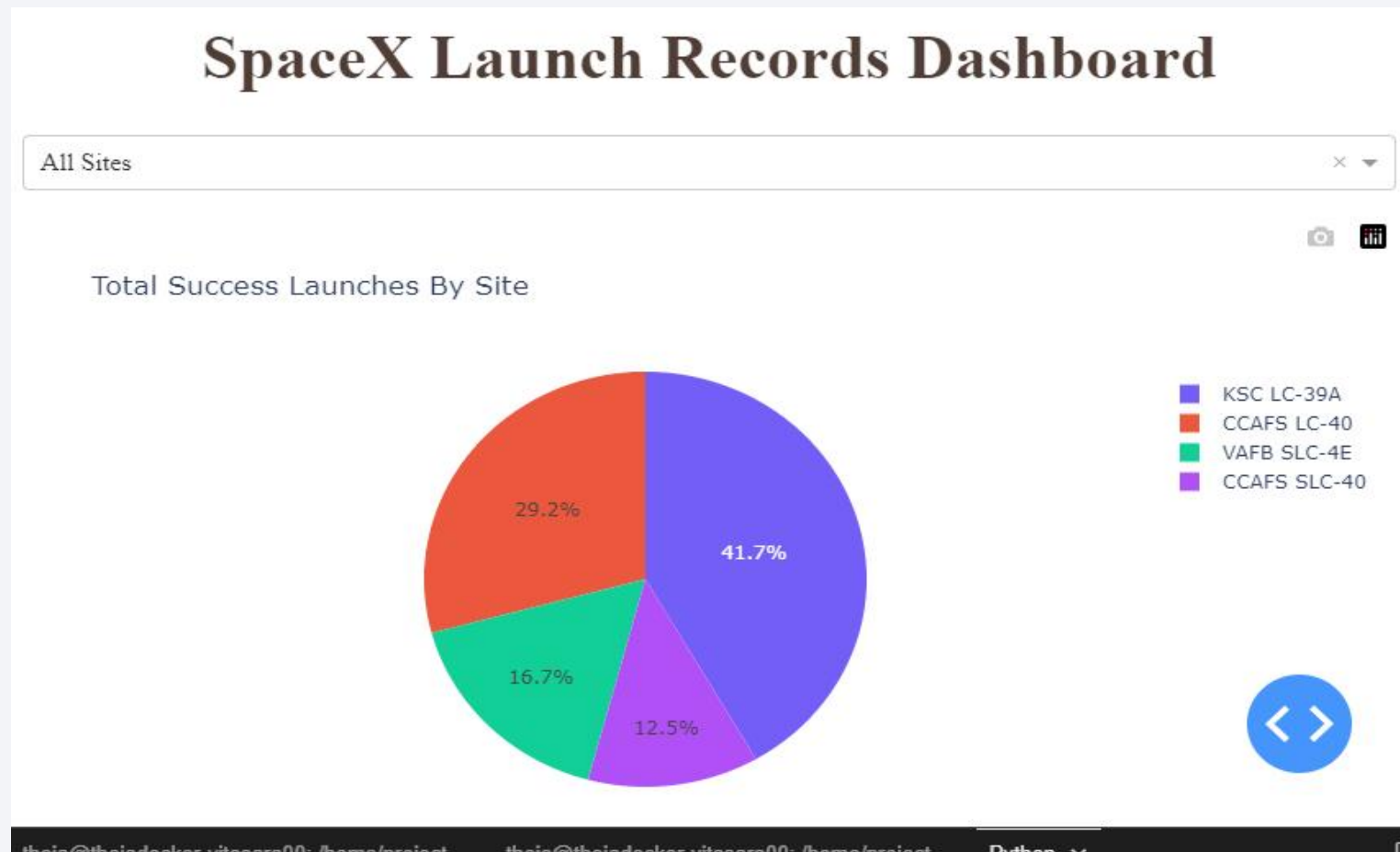


Section 4

Build a Dashboard with Plotly Dash

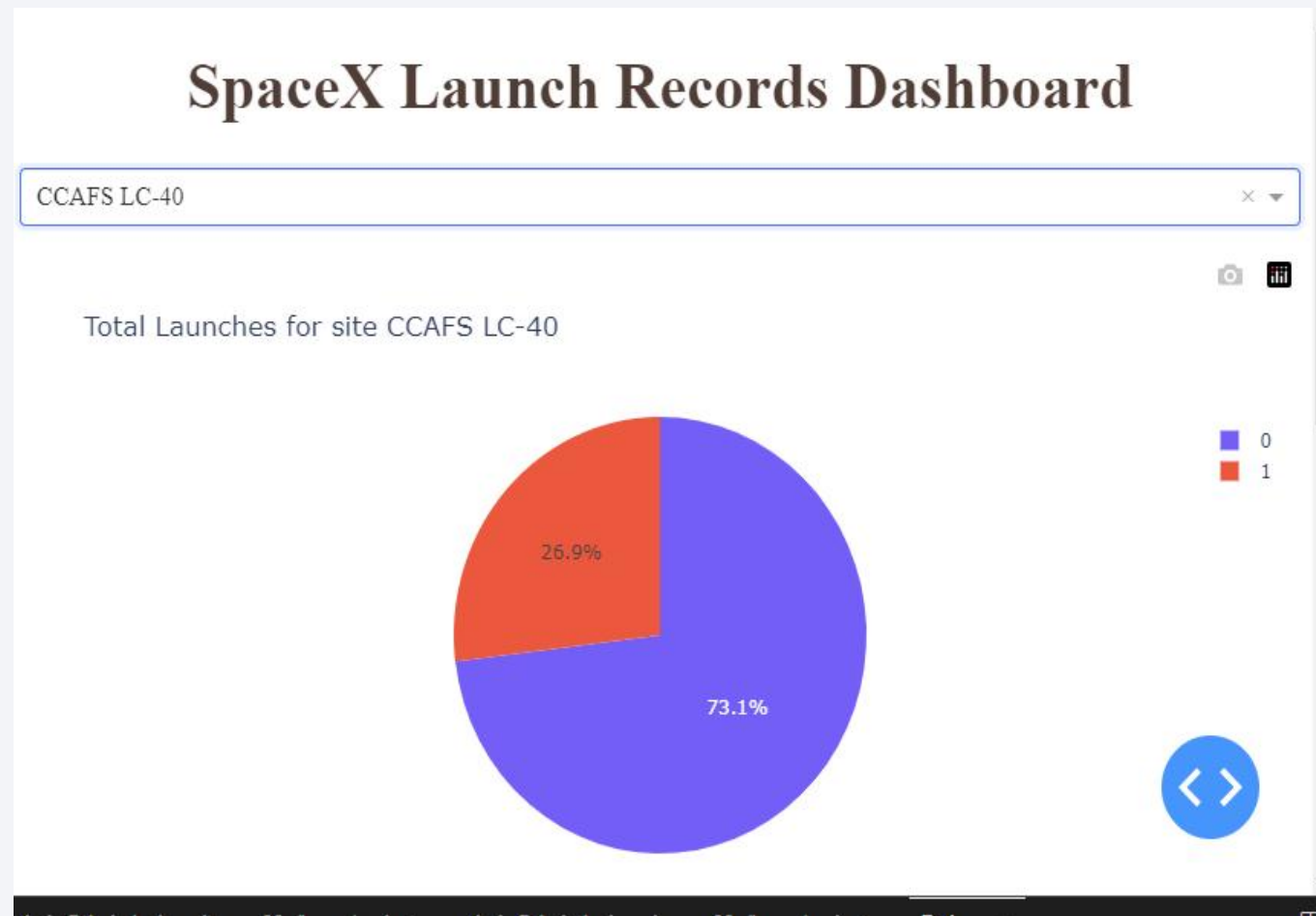
Successful launches by site

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



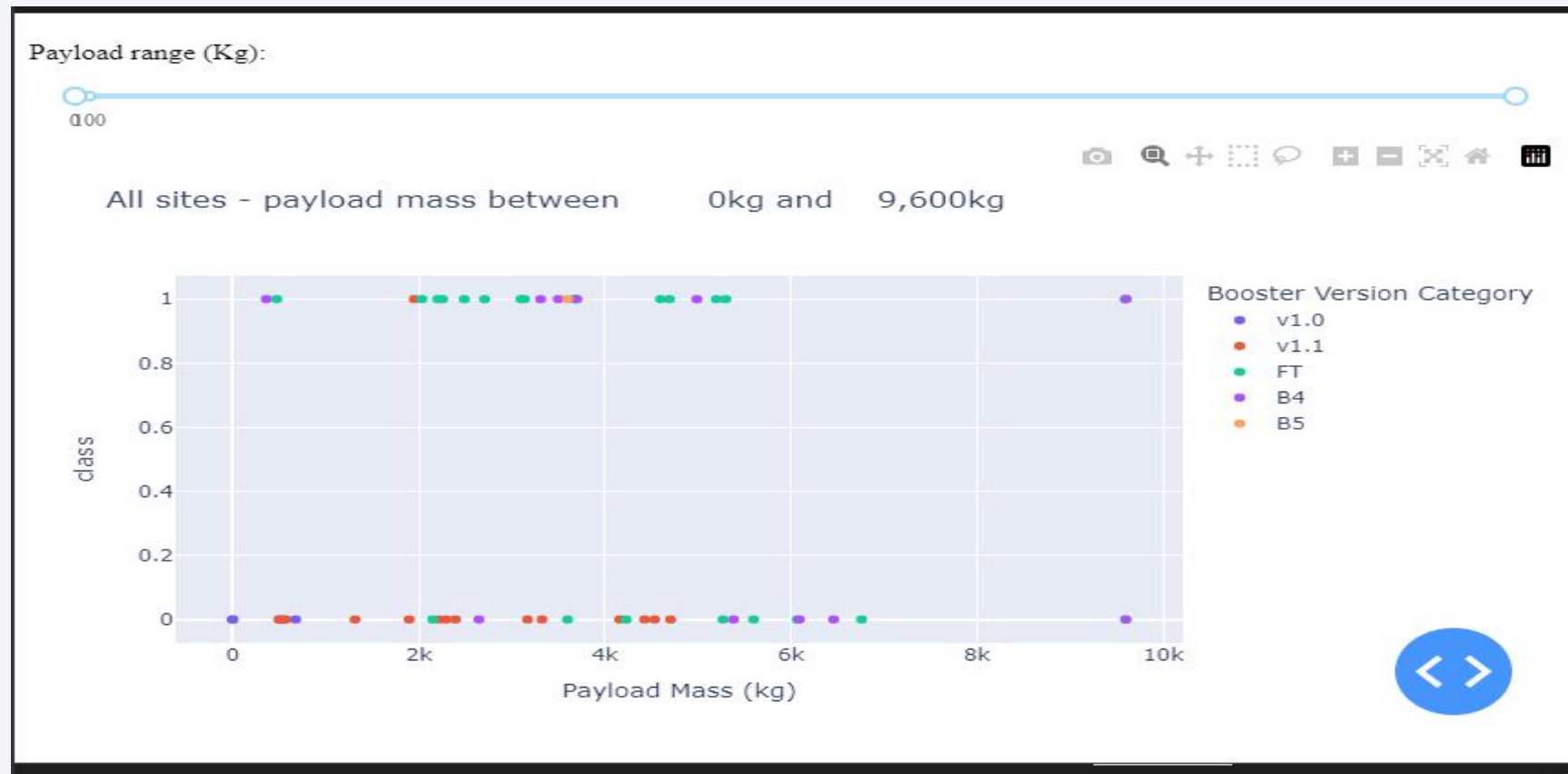
Launch success ratio for CCAFS LC-40

- 73.1% of launches are successful in this site.



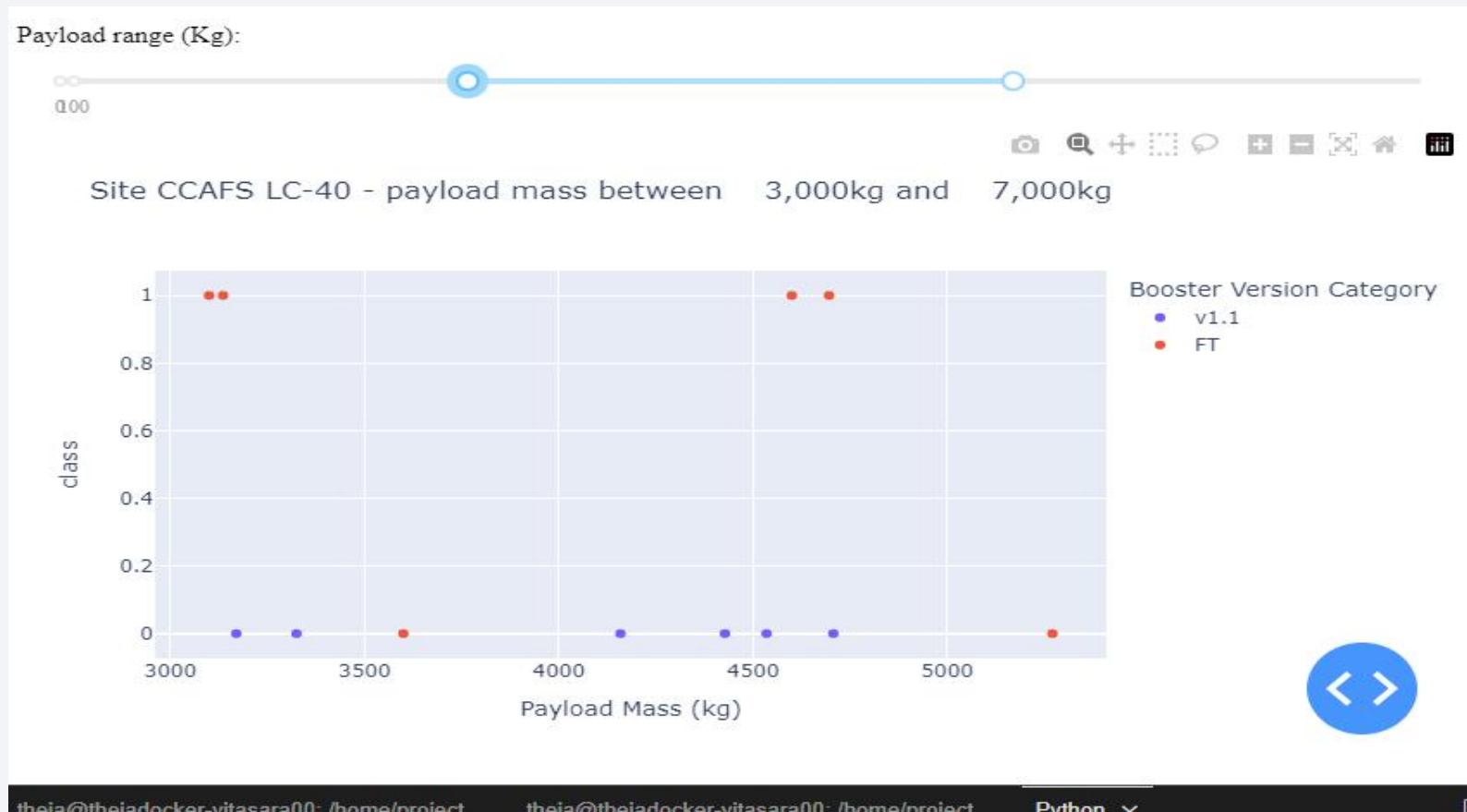
Payload vs Launch Outcome - total range

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



Payload vs Launch Outcome - 3000 to 7000

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

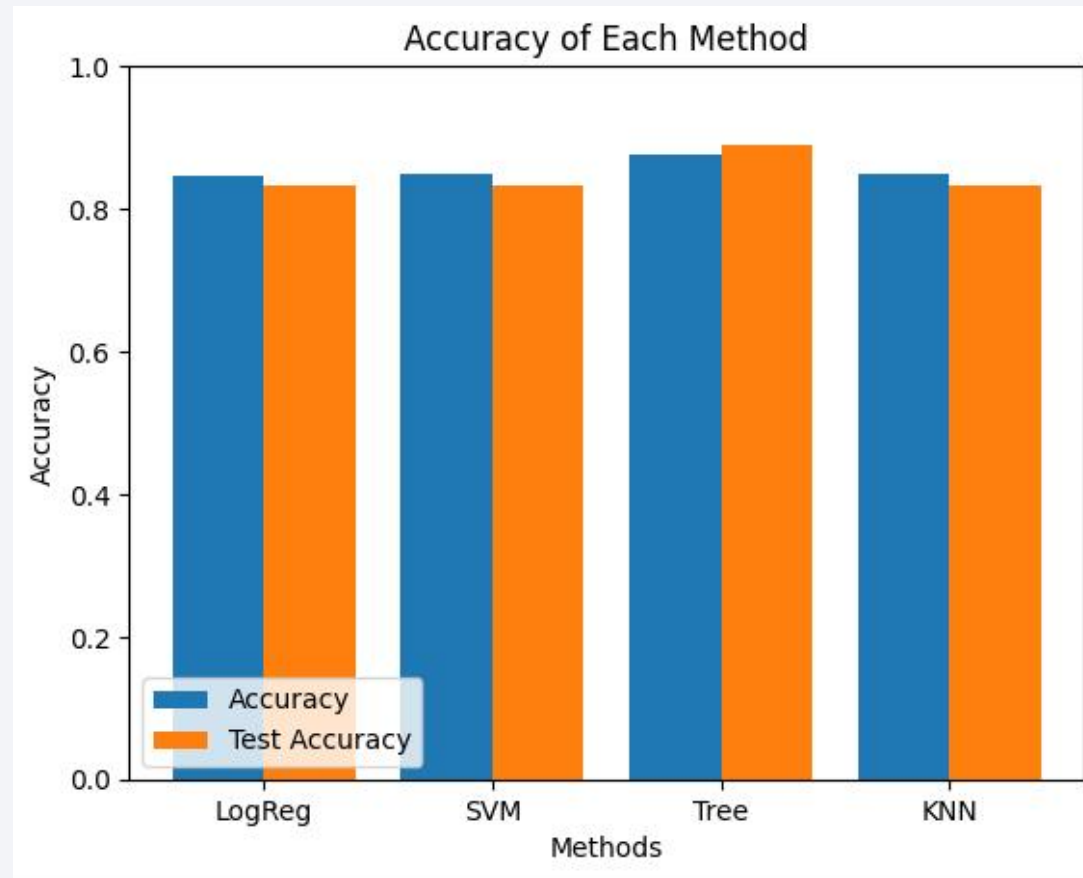




Section 5

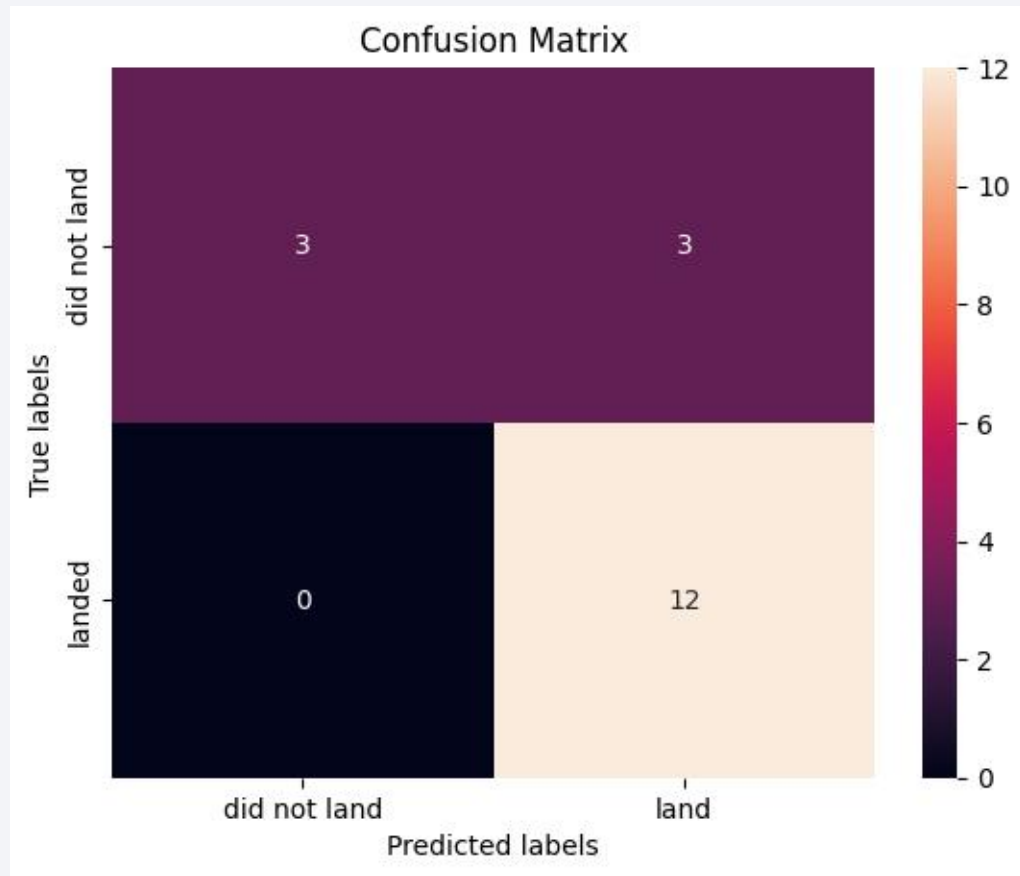
Predictive Analysis (Classification)

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;
- 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 didn't show maps on Github, so I took screenshots.*

Thank you!

