IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fabian Möller
19.09.2022

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection
    - Data Wrangling
    - Exploratory Data Analysis with SQL
    - Exploratory Data Analysis with Data Visualization
    - Interactive Visual Analysis with Plotly, Folium and Dash
    - Machine Learning Prediction
- Summary of all results
    - Exploratory Data Analysis Results
    - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars
  - other providers cost upwards of 165 million dollars per launch
  - SpaceX can reuse the first stage. f we can determine if the first stage will land, we can determine the cost of a launch.
  - Therefore we compare the viability of new company Space Y in competition with Space X
- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - Where is the best place to make launches
  - What operating conditions needs to be in place to ensure a successful landing program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was done to categorical 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
  - How to build, tune, evaluate classification models

Data Collection

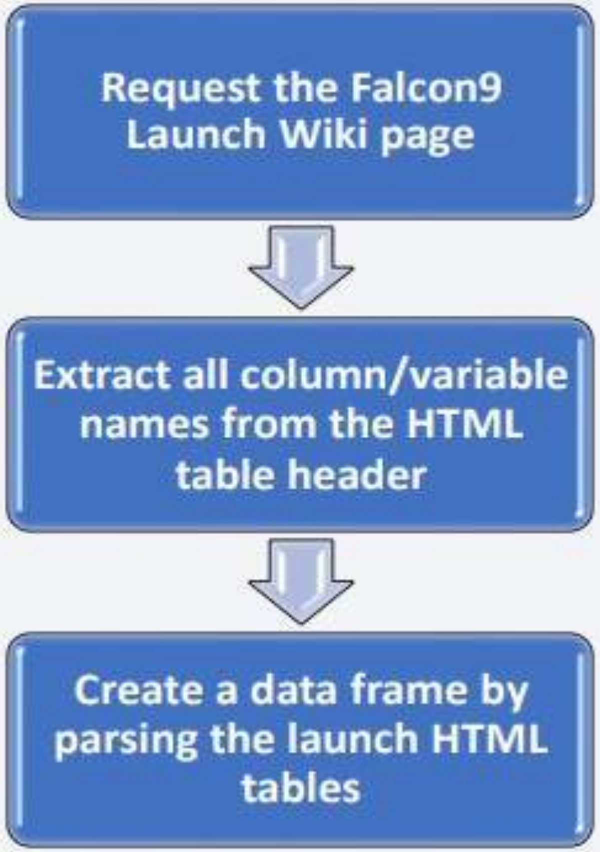
- Data sets were collected from
- Space X API (https://api.spacexdata.com/v4/rockets/) and from
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) using web scraping technics.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- GitHub URL:
  - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

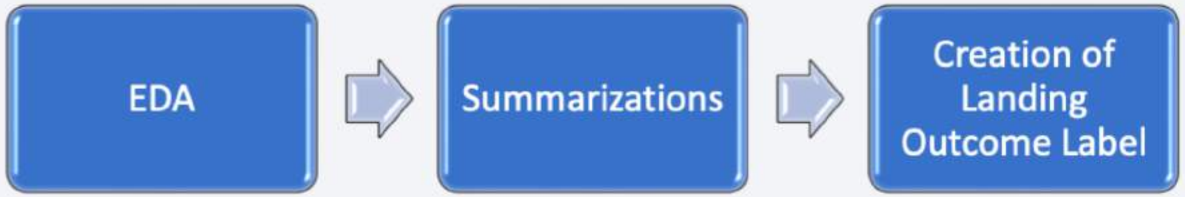
Data Collection - Scraping

- We applied web scrapping to Web scrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas data frame
- GitHub URL:
  - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

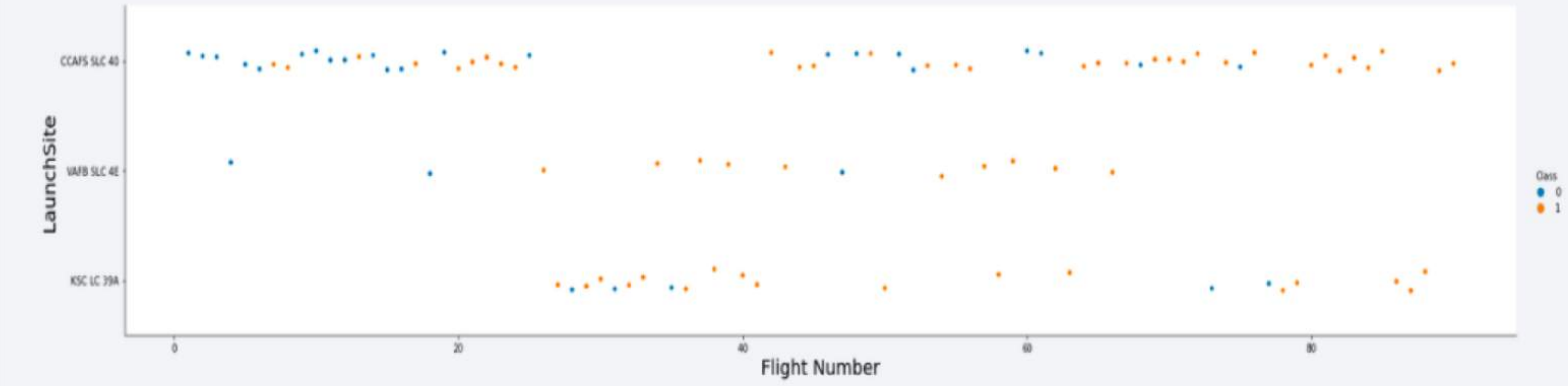
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column. Describe how data were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



- GitHub URL:
  - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly:



- GitHub URL:
  - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
- The names of unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The failed landing outcomes in the drone ship, their booster version and launch site names
- GitHub URL: https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera-2.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map
- We assigned the feature launch outcomes (failure or success) to class 0 and 1. i.e., 0 for failure, and 1 for success
- Using the color-labeled marker clusters, we identified which launch sites have a relatively high success rate
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
    - Are launch sites near railways, highways and coastlines
    - Do launch sites keep certain distance away from cities

- GitHub URL:
    - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics.ipynb

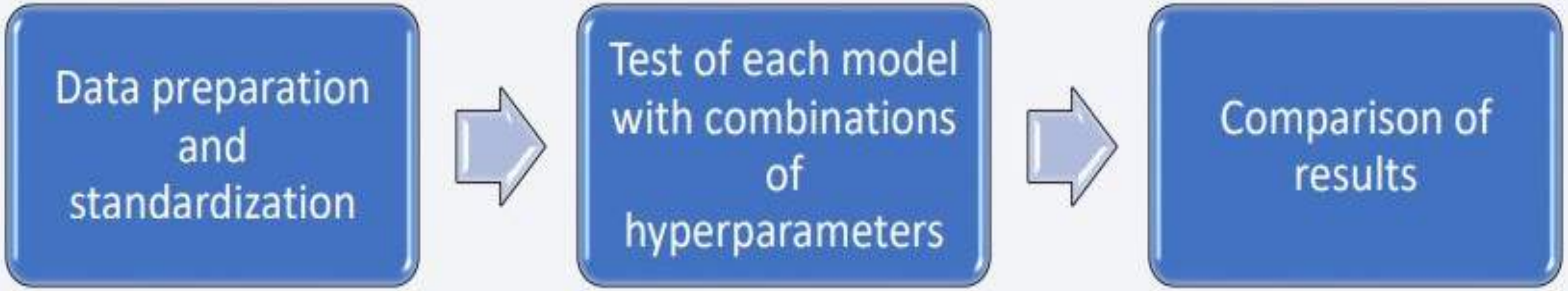
Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain site
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster versions

- GitHub URL:
  - https://github.com/Fubi99/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



- GitHub URL:
  - https://github.com/Fubi99/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory data analysis results

- Space X uses four different launch sites
- The first launches were done by Space X itself and NASA
- The average payload of F9 v1.1 booster is 2928 kg
- The first success landing outcome happened in 2015, fiver years after the first launch
- Many Falcon 9 booster versions were successful at landing on drone ships having a 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 better with time

Results

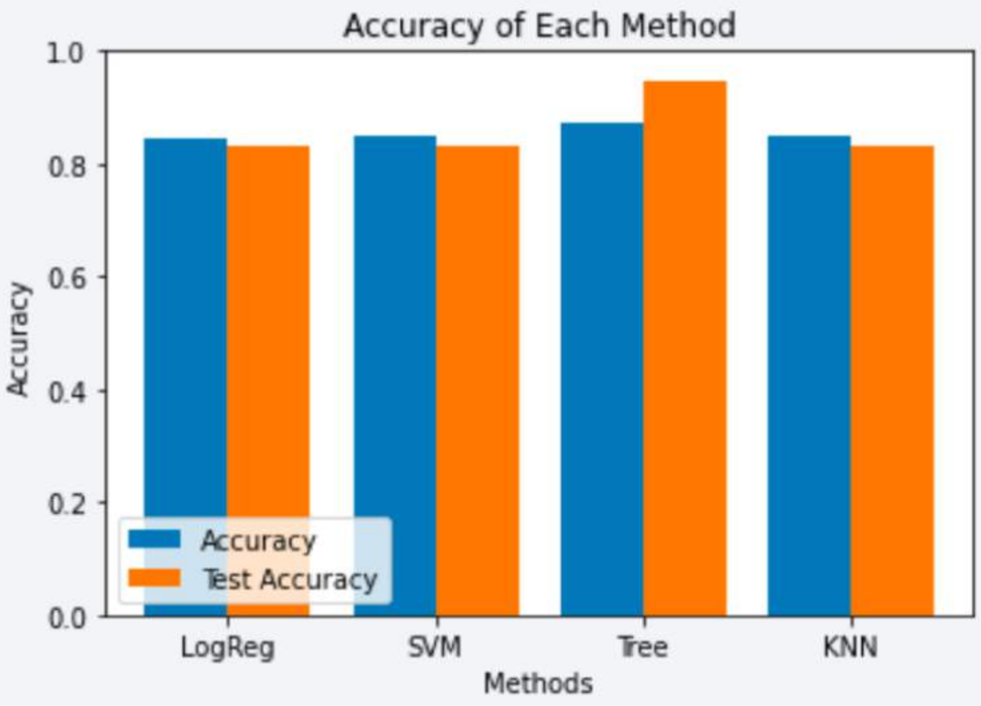
Interactive analytics demo in screenshots

- Using interactive analytics was possible to identify that launch sites are in safety places, near the sea, for example and have a good logistic infrastructure around.
- Most launches happen at east coast launch sites

Results

Predictive analysis results

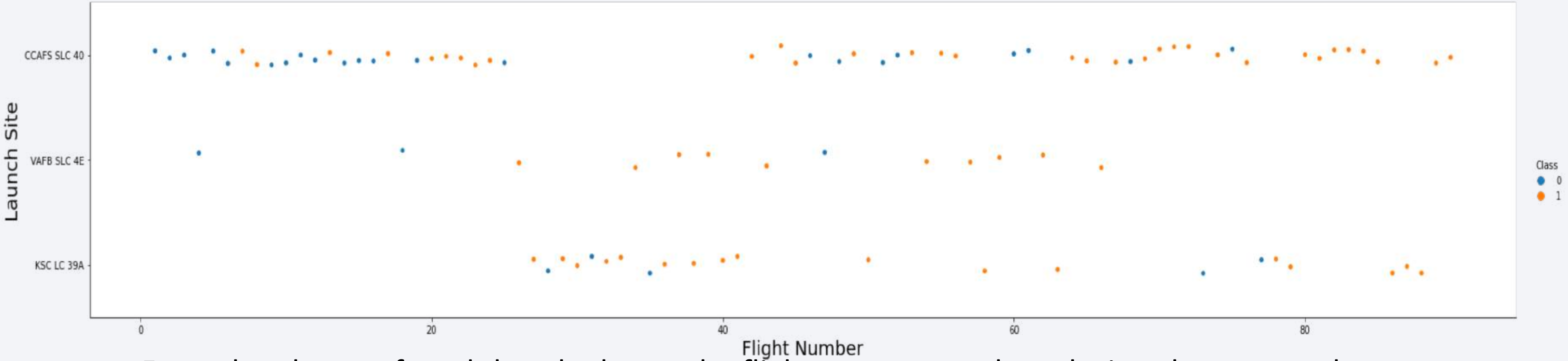
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having an accuracy over 87% and accuracy for test data over 94%.



Section 2

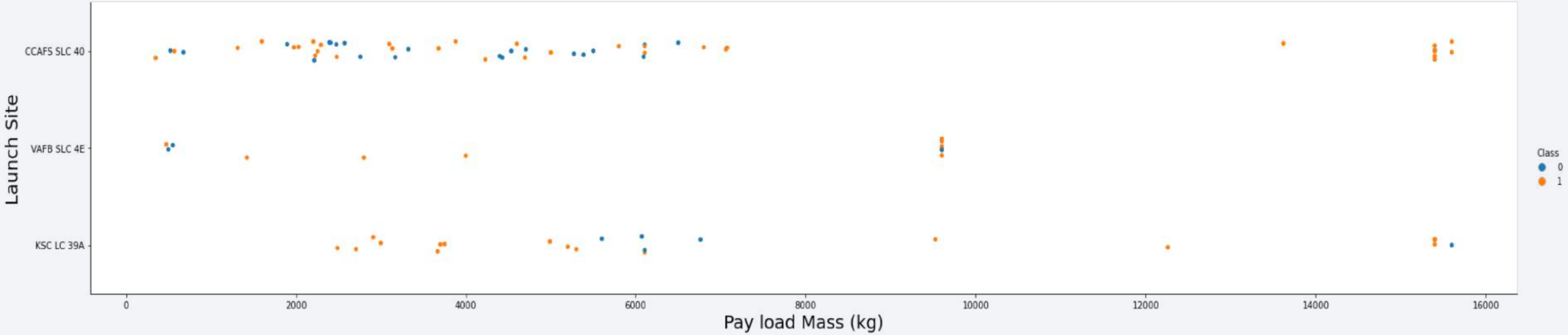
Insights drawn from EDA

Flight Number vs. Launch Site



- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a 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
- general success rate improved over time

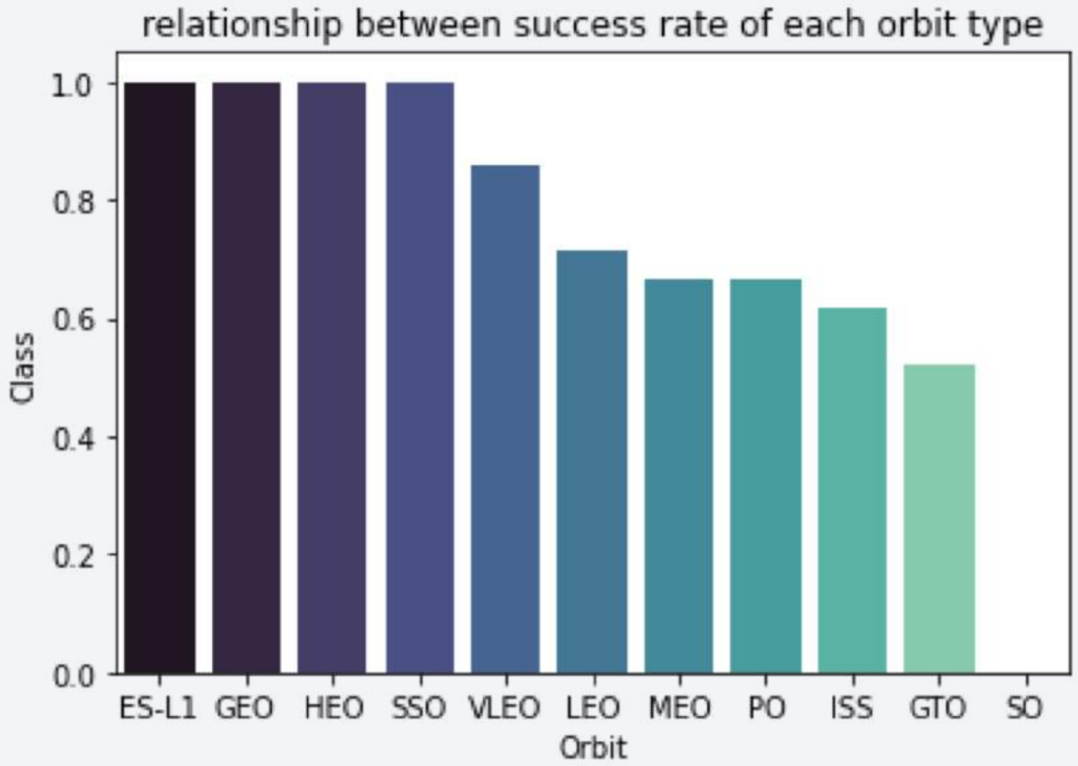
Payload vs. Launch Site



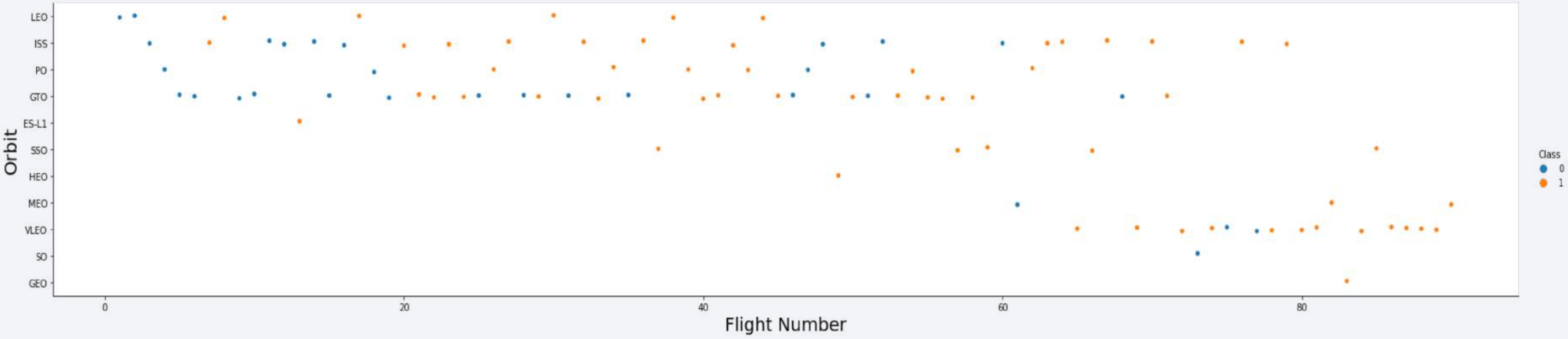
- The greater the payload mass for launch site CCAFS SLC 40, the higher the success rate of the rocket
- Payloads over 9000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the highest success rates

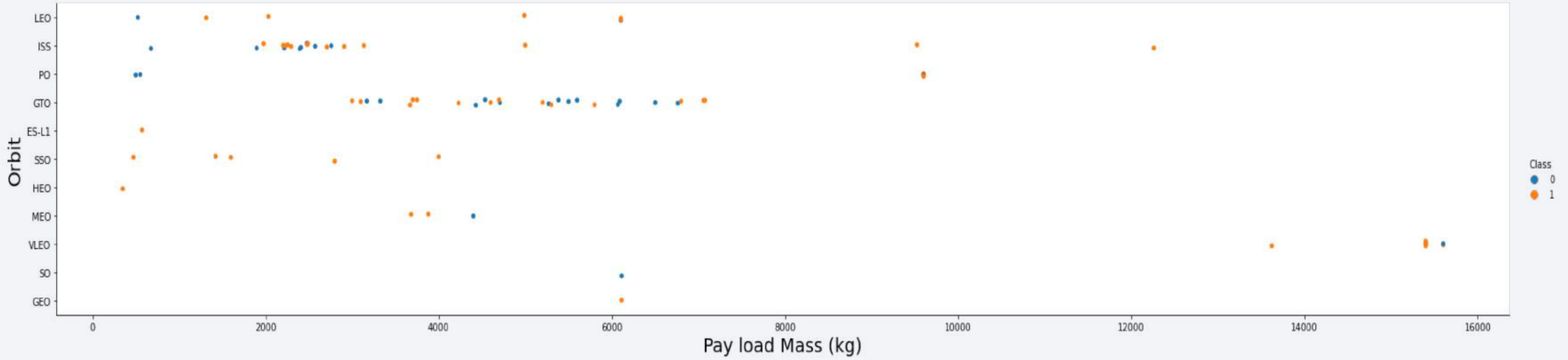


Flight Number vs. Orbit Type



- The plot shows the Flight Number vs. Orbit type
- In the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit

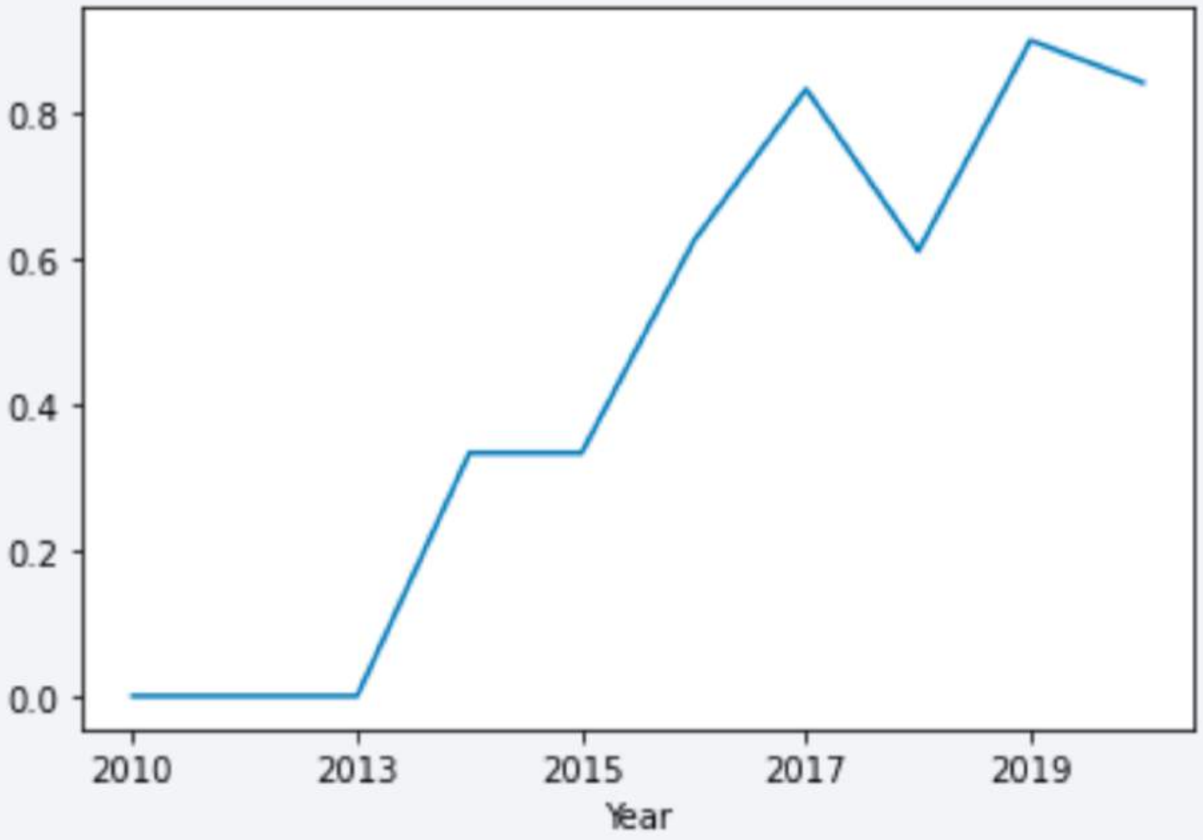
Payload vs. Orbit Type



- We can observe that with heavy payloads, the successful landing are for PO, LEO and ISS orbits

Launch Success Yearly Trend

- From the plot, we can observe the success increasing from 2013 to 2020



All Launch Site Names

- We used the key word **DISTINCT** to show only unique launch sites from the SpaceX data

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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	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

Total Payload Mass

- Total payload carried by boosters from NASA

45596

- 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

2928

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

2015-12-22

- We observed that the dates of the first successful landing outcome on ground pad was 22.12.2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of 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

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

- We used wildcard like '%' to filter for **WHERE** MissionOutcome was a success or a failure

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

- We determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.

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

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The list above has the only two occurrences

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

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20
- 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.

landing_outcome	COUNT
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

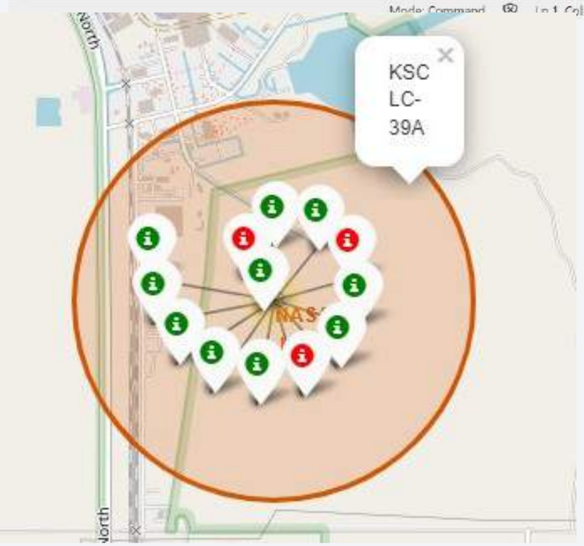
Section 3

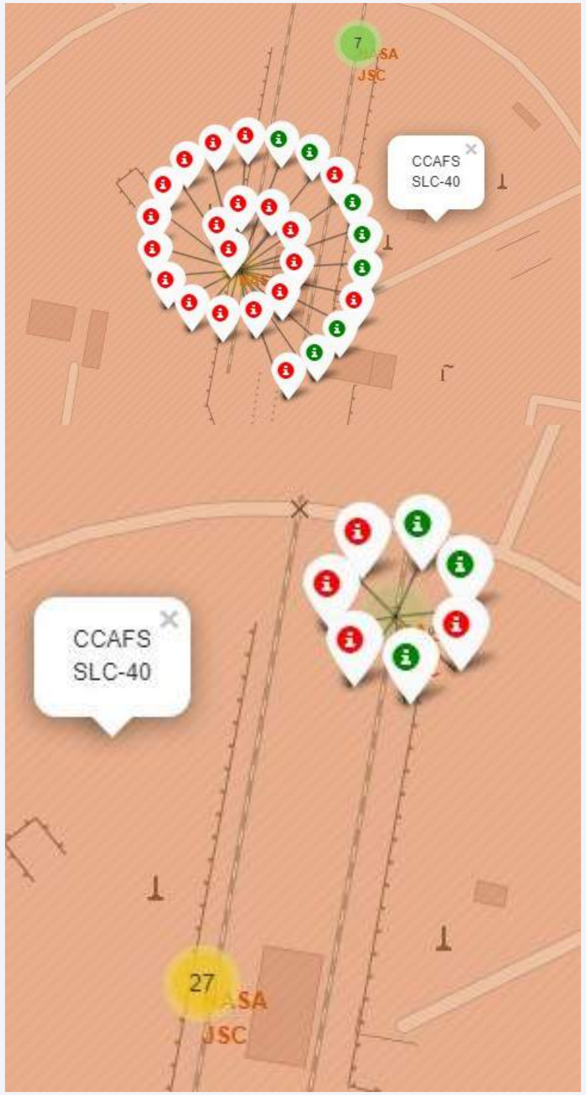
Launch Sites Proximities Analysis

All launch sites global map markers

- Launch Sites are on the Coast of the United States of America

Launch Outcomes by Site





- Green markers show successful launches and red markers show failures

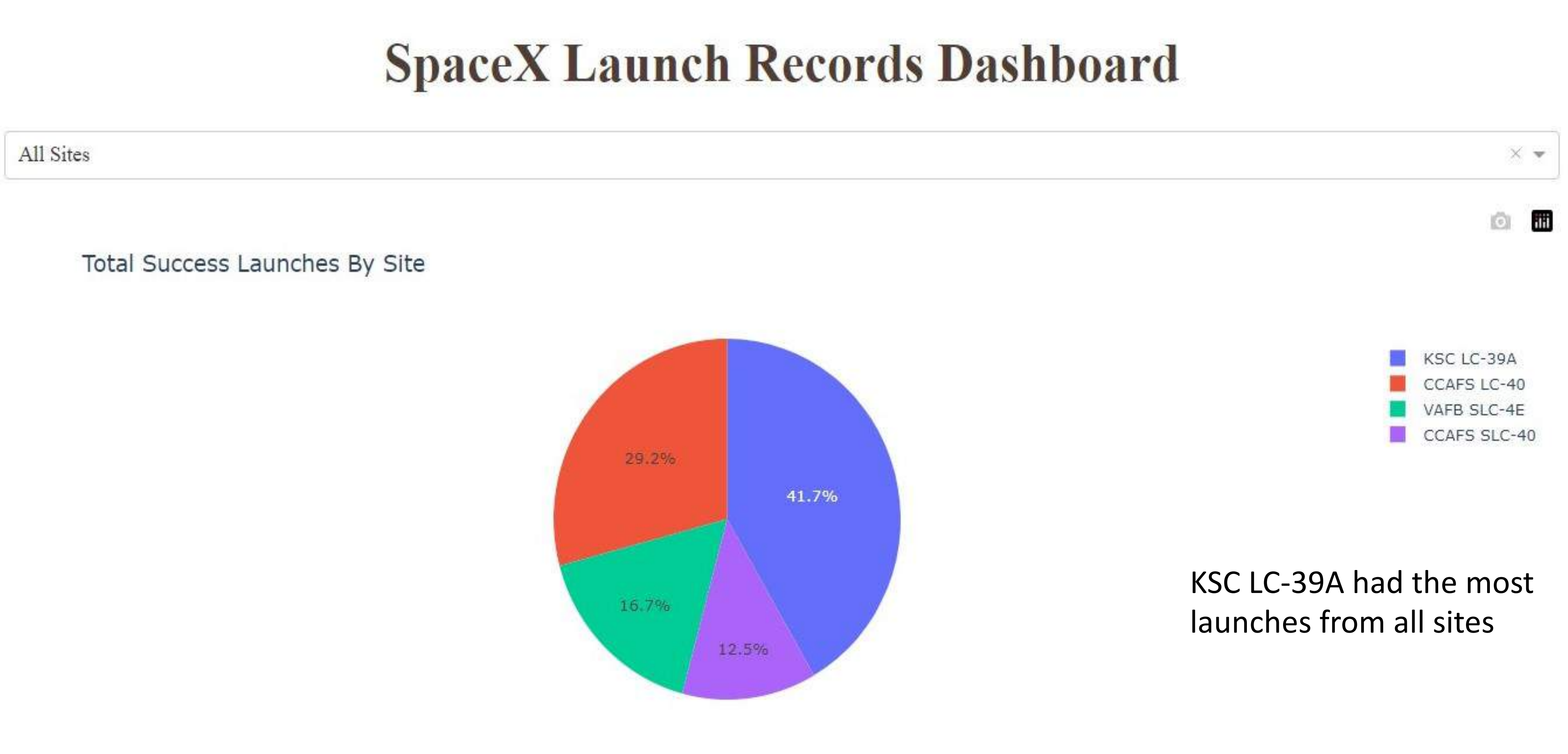
Launch Site distance to Coast

- CCAFS SLC-40 Launch site is close to the coast

Section 4

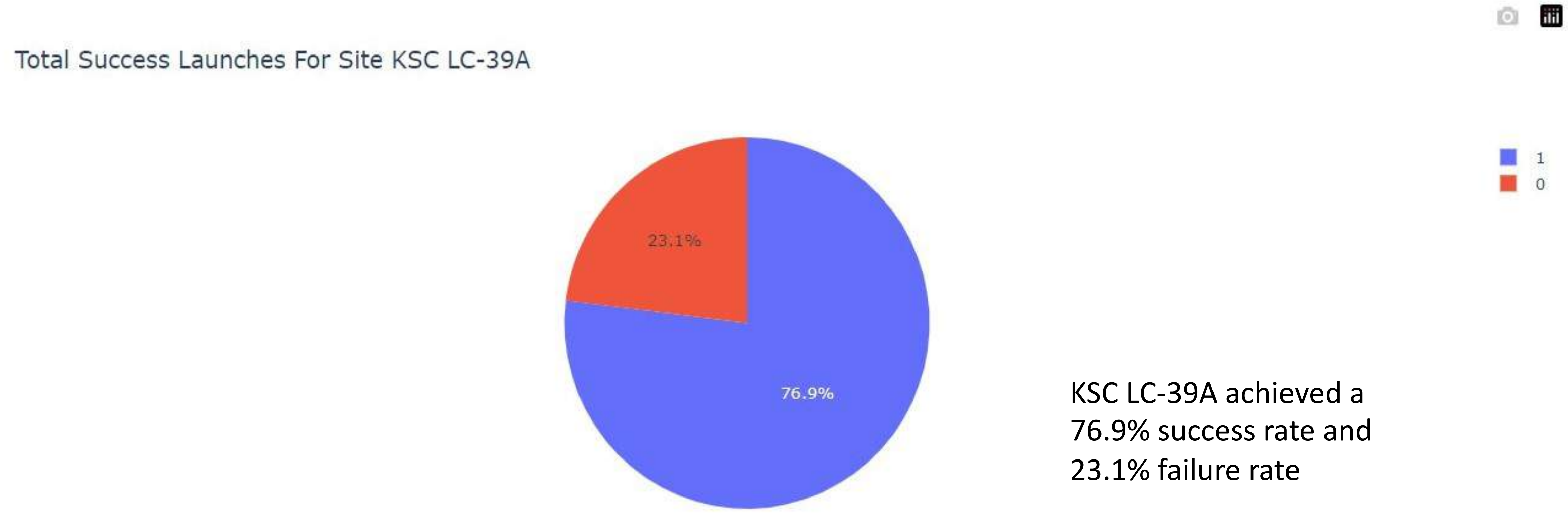
Build a Dashboard with Plotly Dash

Successful Launches by Site



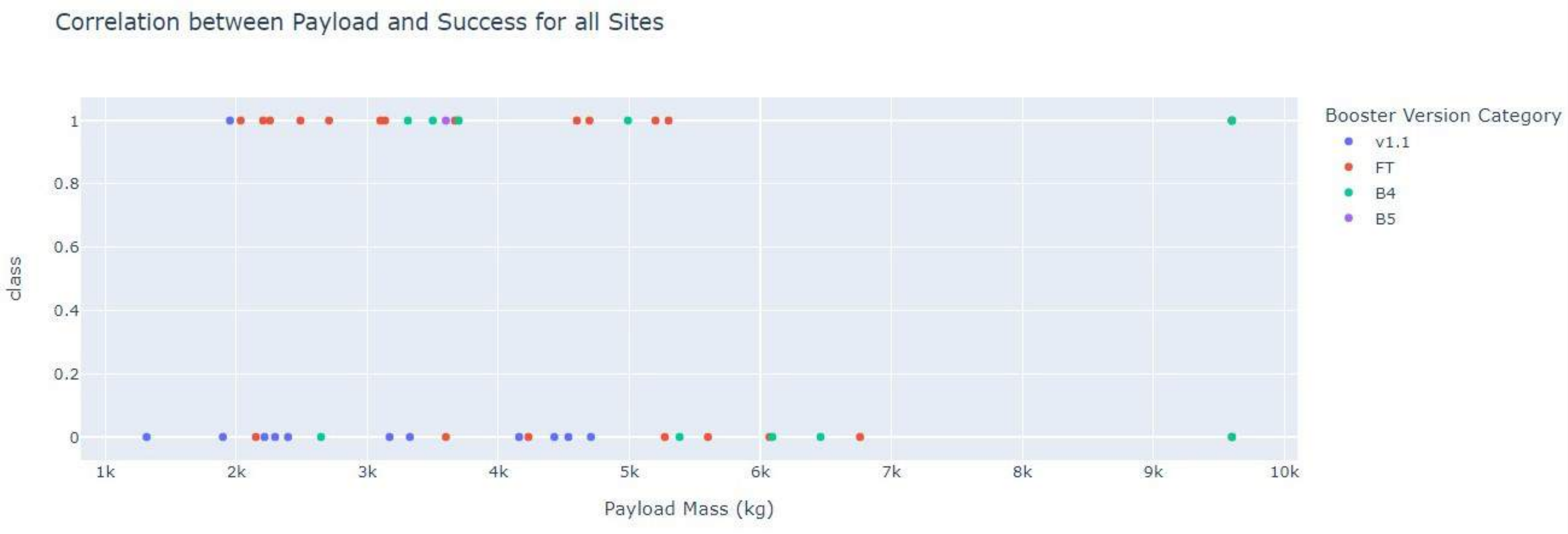
KSC LC-39A had the most launches from all sites

Launch Success Ratio for KSC LC-39A



KSC LC-39A achieved a 76.9% success rate and 23.1% failure rate

Payload vs. Launch Outcome



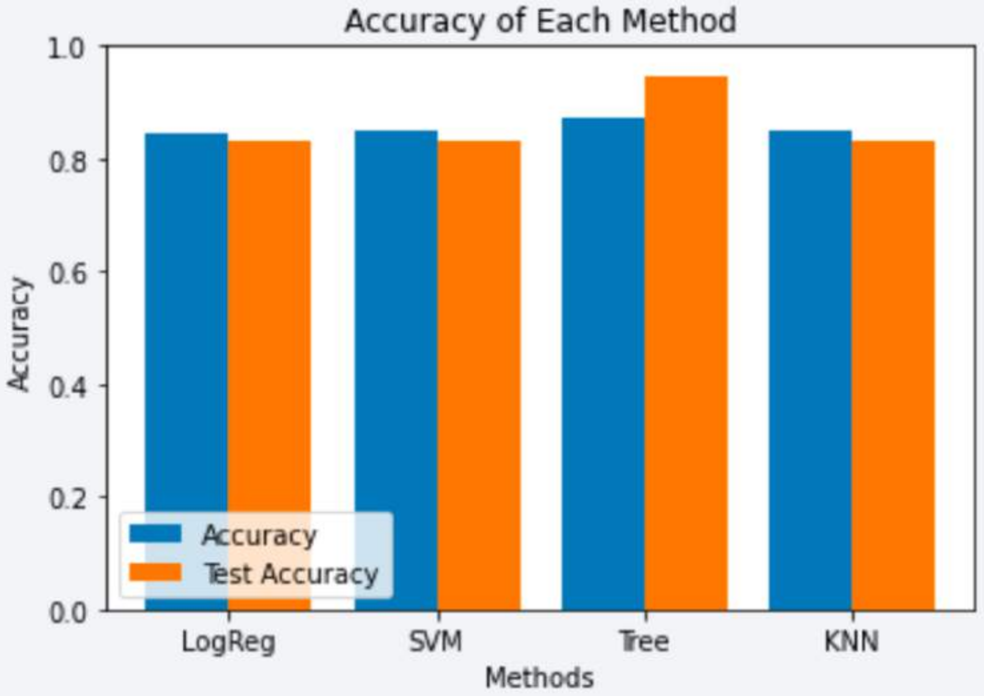
- Payloads under 6000kg and FT boosters are the most successful combination

Section 5

Predictive Analysis (Classification)

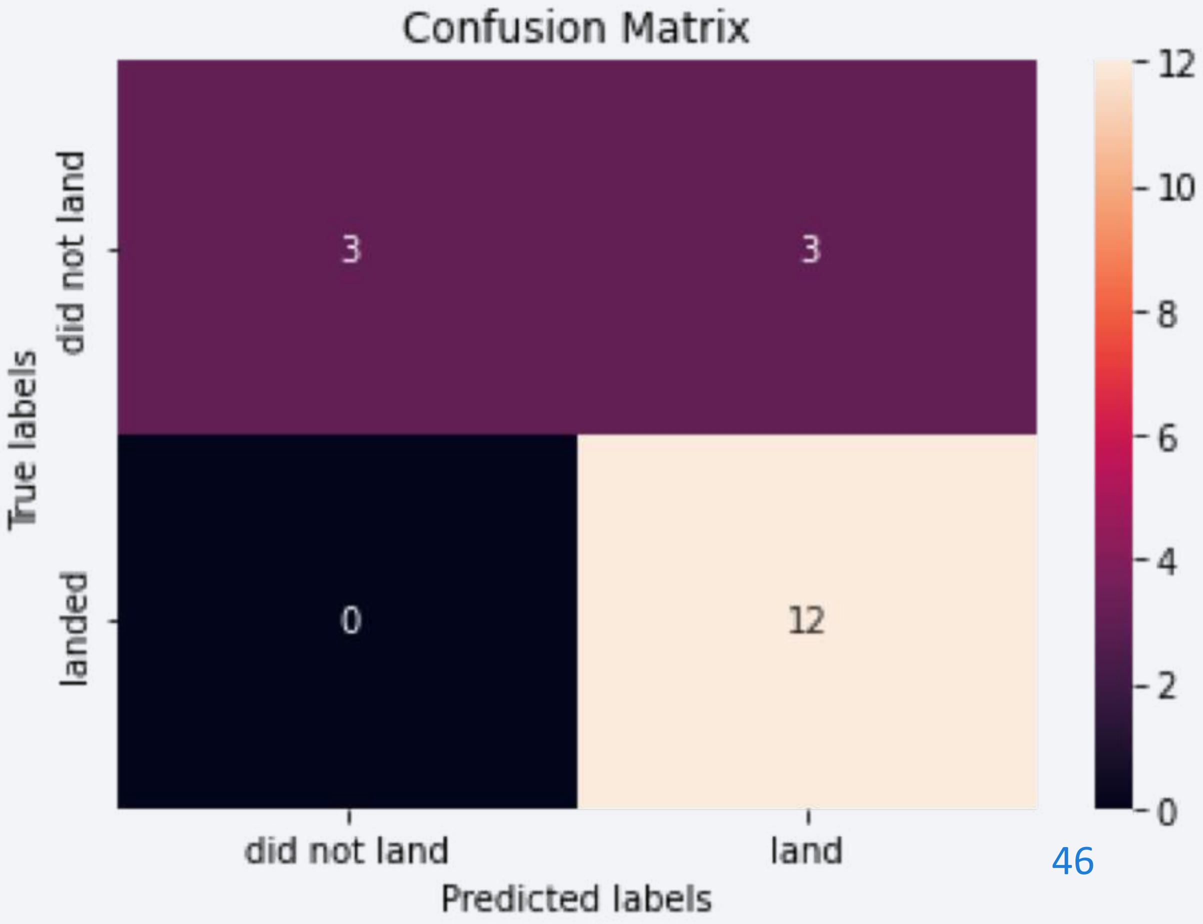
Classification Accuracy

- Four classification models were tested
- Their accuracy is the same on the test data
- The model with the highest classification accuracy is Decision Tree Classifier, which has an accuracy over 87% on the train data



Confusion Matrix

- 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:

- The larger the flight amount at a launch site, the greater the success rate at a launch site
- Launch success rate starts to increase in 2013 until 2020
- Orbits ES-L1, GEO, HEO, SSO, VLEO have the highest success rate
- KSC LC-39A had the most successful launches of all sites
- Launches above 7000kg are less risky
- The Decision tree classifier is the best machine learning algorithm for predictions

Appendix

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

Thank you!