IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franco Stellari
9/11/24

Outline

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

Executive Summary

- We have used publicly available data from Wikipedia and SPaceX API to evaluate the success rate of reusing the 1st stage booster, thus allowing a reduction of launch cost
- We have evaluated several features impacting the success rate, including launch site, payload mass, booster type, targeted orbit, etc.
- The data shows a progressive improvement of the success rate with time/flight number, thus suggesting a possible reduction of risk and cost in the future
- We have been able to train and evaluate several classification model which show a R2 score of ~83% in predicting the success/failure of a launch based on the available feature information

Introduction

- In this project we wanted to evaluate the reusability of SpaceX boosters to reduce launch cost
- Specifically, we wanted to find out if any characteristic of the launch, like Site, Payload, Booster Type, impact the success rate of landing the 1^{st} stage booster and later reuse it

Section 1

Methodology

Methodology

Executive Summary

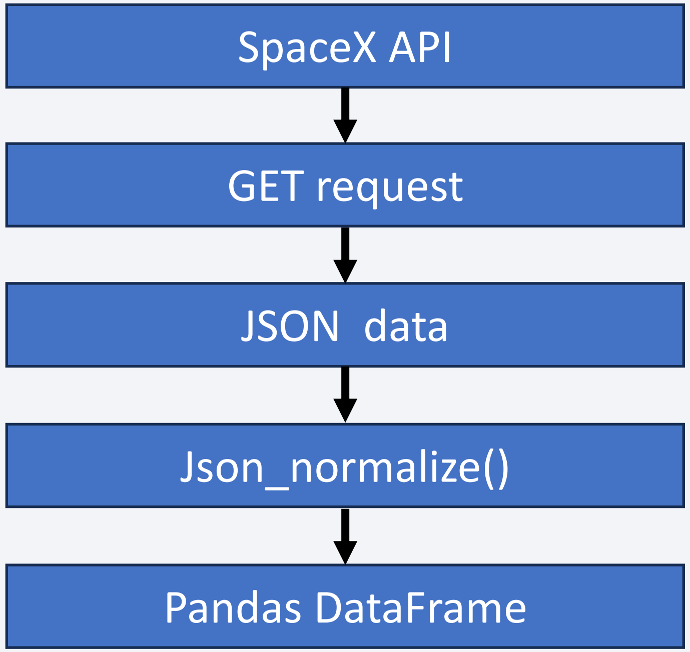
- Data collection methodology:
  - Data was collected from publicly available records of SpaceX launches and SpaceX API
- Perform data wrangling
  - Data was processed to remove missing data points
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification models, such as SVM, KNN, DecisionTree, LinearRegression, we used to model the success rate of booster landing and predict future performance

Data Collection

- Data was collected from public web sites using pull requests
- Tables were extracted from web pages using Beautiful Soup and converted to Pandas DataFrames for further processing

Data Collection – SpaceX API

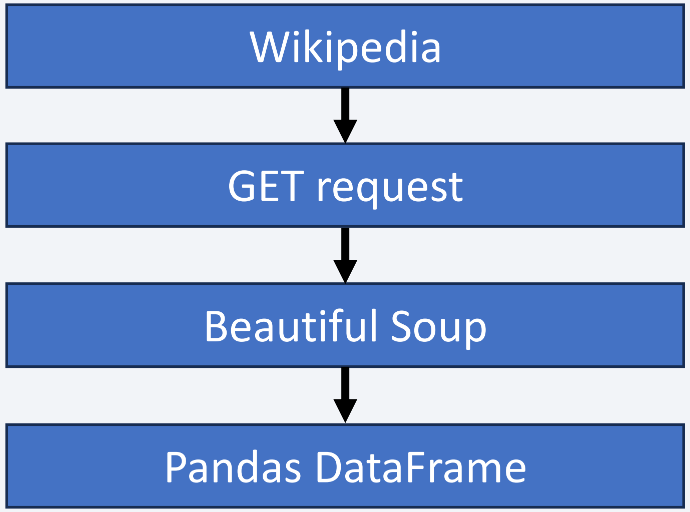
- Past launch information is retrieved from the SpaceX API at https://api.spacexdata.com/v4/launches/past using a GET request which returns a JSON file containing a list of launches
- The JSON data is converted into a Pandas DataFrame using the `.json_normalize()` function



https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/slide%208%20DV0101EN-Exercise-Generating-Maps-in-Python.ipynb

Data Collection - Scraping

- Data was retrieved from Wikipedia web site
- After downloading the relevant page, Beautiful Soup was used to retrieve data from a table embedded in the web page
- Finally, the data is converted to a Pandas DataFrame for further processing and anaysis



https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/jupyter-labs-webscraping.ipynb

Data Wrangling

- The dataset was cleaned by removing record with missing information
- Then a new feature `class` was introduced to report the saucerful=1 or unsuccessful=0 landing of the booster based on other labels in the `Outcome` feature
- The newly created `class` fauture can be used as a target of our modeling

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We make a scatter plot of the FlightNumber vs. PayloadMassand to show that as the flight number increases, the first stage is more likely to land successfully.
- The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.
- By plotting the Payload Mass and Orbit type we can observe more Payload Mass and Orbit type.
- By plotting the overall success rate over the year, we can see a clear improvement over time.

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/edadataviz.ipynb

EDA with SQL

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Task 1: Get a list of Launch Sites
- Task 2: Display 5 records where launch sites begin with the string 'CCA’
- Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)
- Task 4: Display average payload mass carried by booster version F9 v1.1
- Task 5: List the date when the first successful landing outcome in ground pad was achieved.
- Task 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Task 7: List the total number of successful and failure mission outcomes
- Task 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Task 9: List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Task 10: 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.

Build an Interactive Map with Folium

- We created a map with marks for
  - all launch sites on the map
  - the successful/failed launches for each site

- This helps us identify any relationship between the launch site and a successful launch outcome

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- We have created a Dropbox to filter the data by launch site
- We have created a Slider to filter the data by payload mass
- We have generated a Pie chart of the success/failure count controlled by the Dropbox selection
- We have added a scatter plot of the success/failure by payload mass controlled by both the Dropbox and Slider

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/spacex_dash_app.py

Predictive Analysis (Classification)

- After importing the data about SpaceX launches, we applied a StandardScaler to normalize the data
- We split the data into a training and test set with a test size of 20%
- We evaluated several models: Logistic Regression, Support Vector Machine, Decision Tree, and K-Neighbor Classifier
- Finally, we used GridSearchCV to optimize the parameters of each model

https://github.com/francostellari/data-science-class/blob/main/capstone-ppt/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

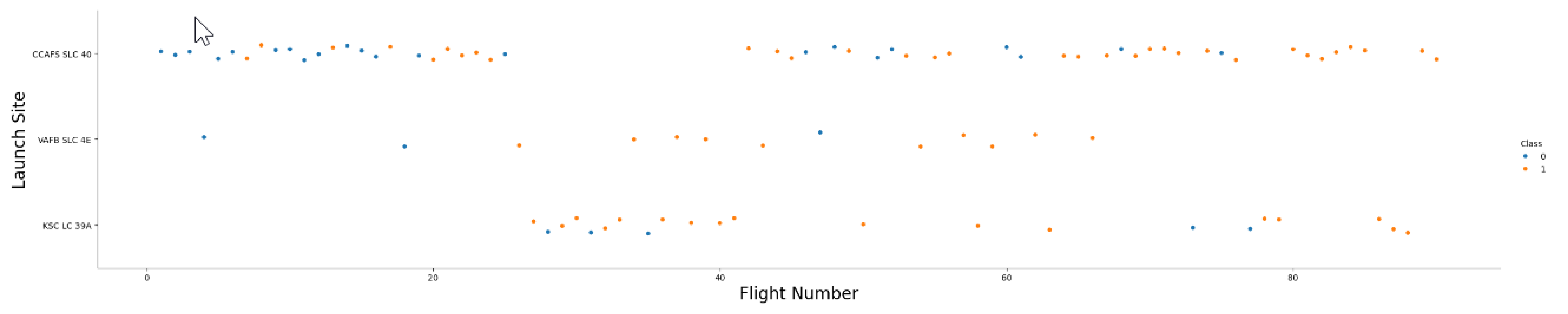
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

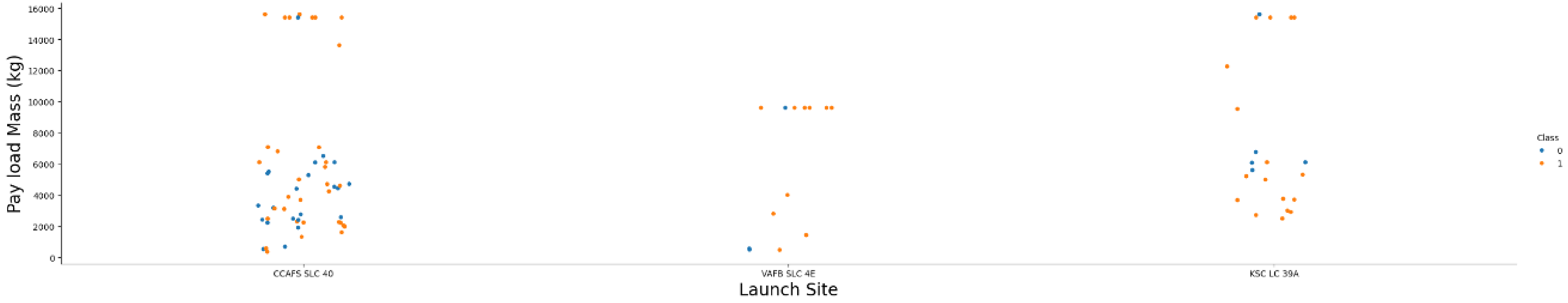
Insights drawn from EDA

Flight Number vs. Launch Site



- We can observe a progressive improvement for each site over time/flight number

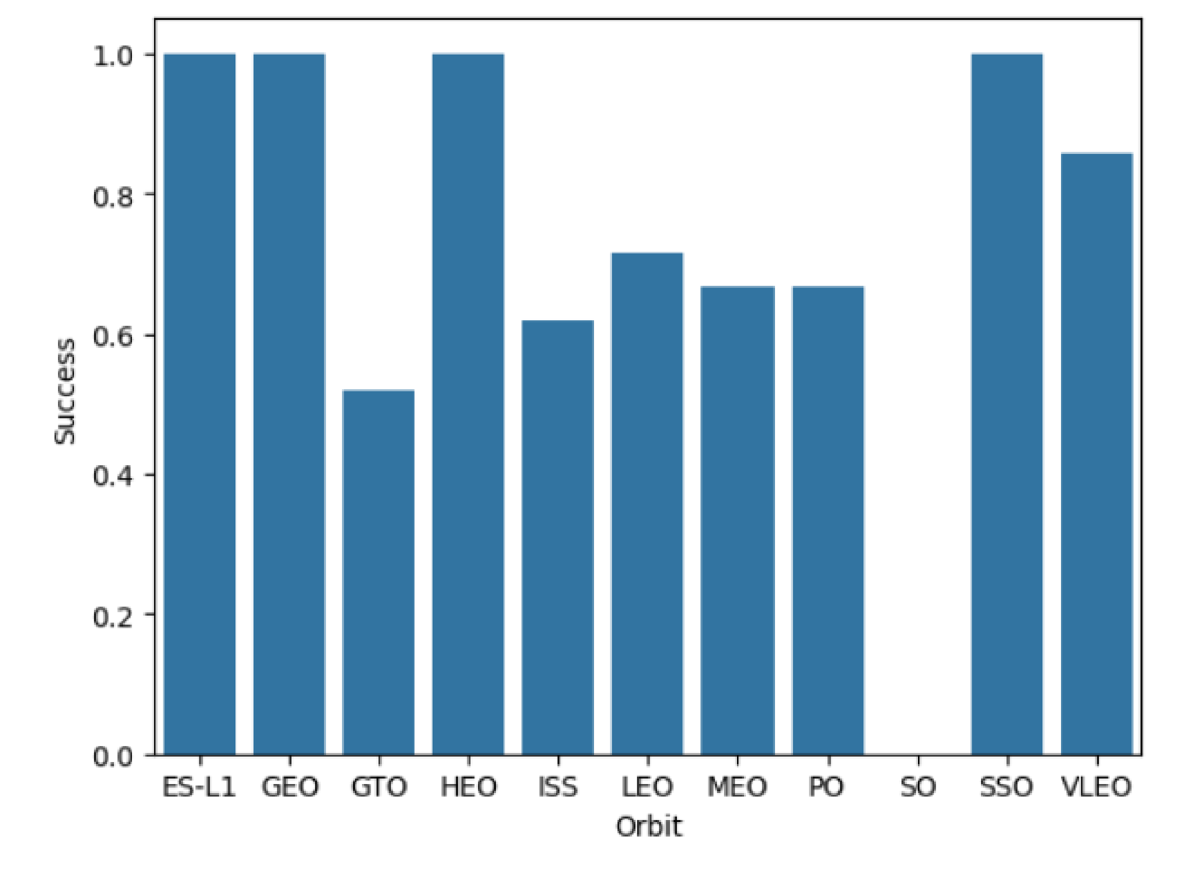
Payload vs. Launch Site



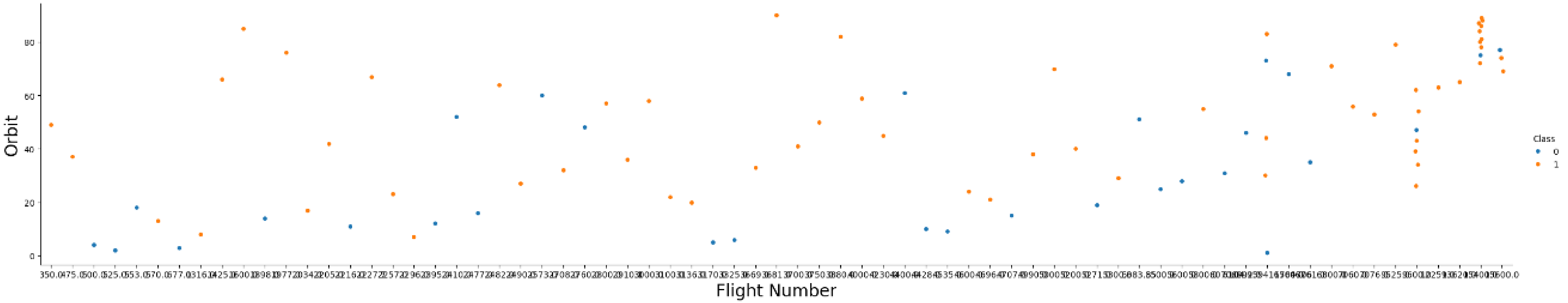
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type

- GTO orbit has the lowest success rate
- ES-L1, GEO, HEO, SSO have all 100% success rate

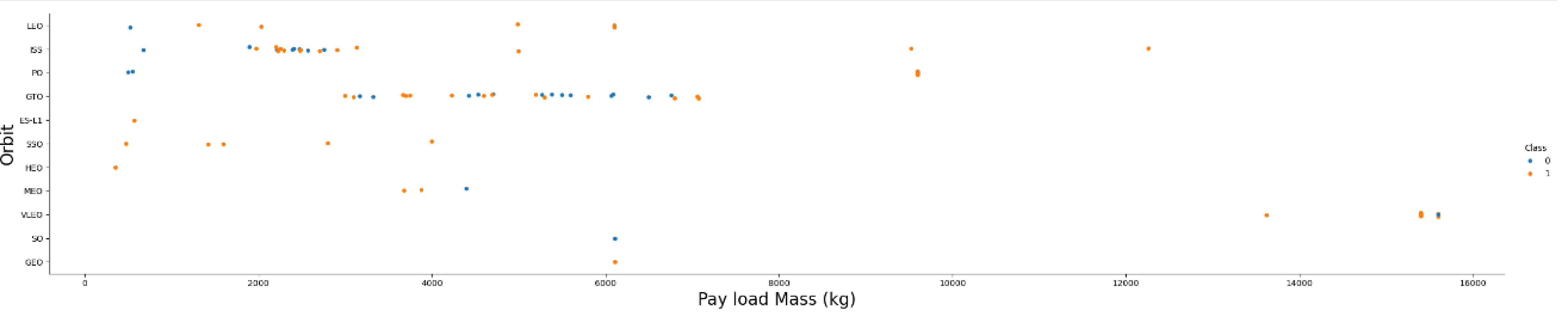


Flight Number vs. Orbit Type



- You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

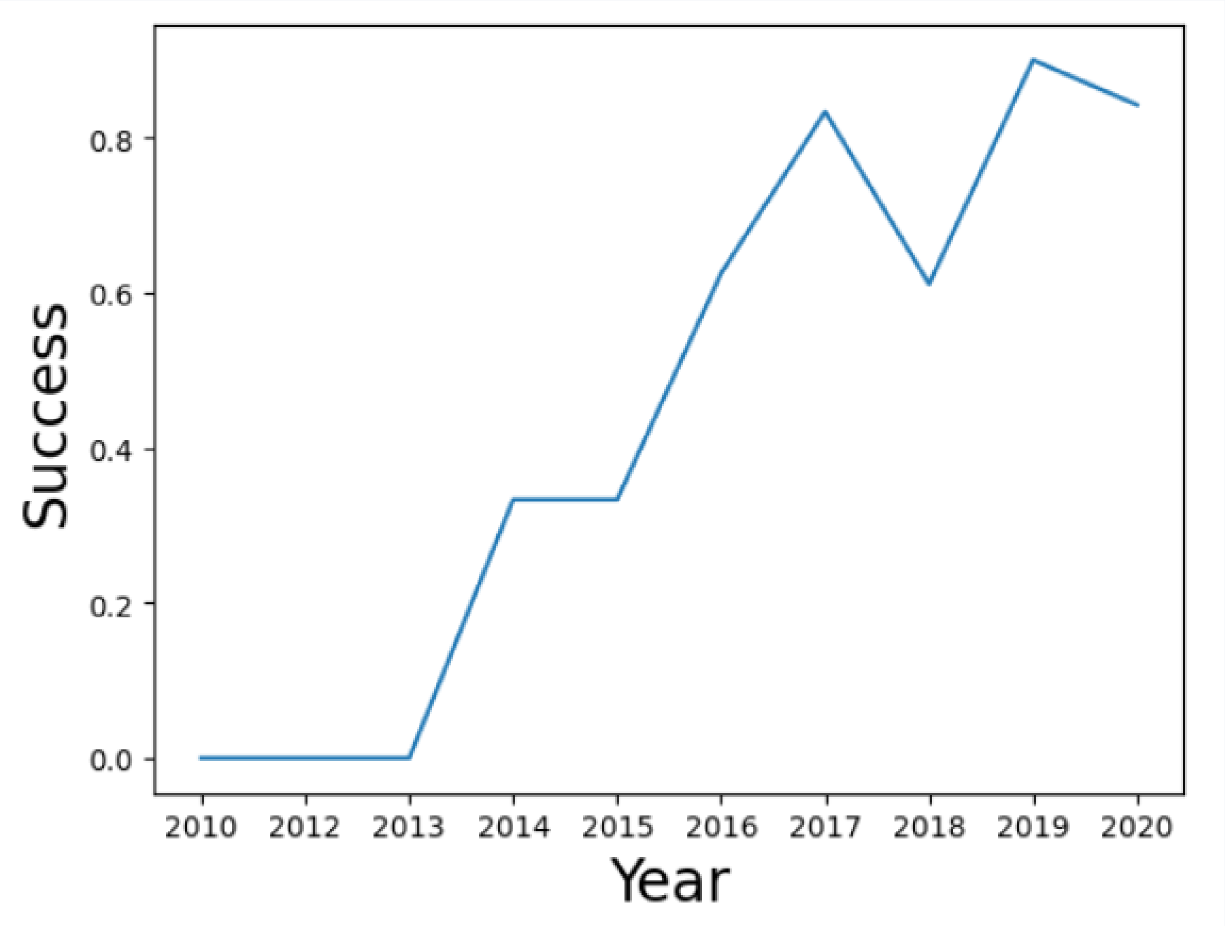
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites
- There are 4 launch sites

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- All successfull

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Calculate the total payload carried by boosters from NASA
- Here is the query: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM "SPACEXTABLE" WHERE "Customer" LIKE "NASA%"

SUM("PAYLOAD_MASS__KG_")
99980

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

```
[40]: %sql SELECT "Date" FROM "SPACEXTABLE" WHERE "Landing_Outcome" IS "Success (ground pad)" ORDER BY "Date" LIMIT 1

 * sqlite:///my_data1.db
Done.
```

[40]:

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

```
[48]: %sql SELECT DISTINCT "Booster_Version" FROM "SPACEXTABLE" WHERE "Landing_Outcome" IS "Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000

 * sqlite:///my_data1.db
Done.
```

[48]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	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

- Present your query result with a short explanation here

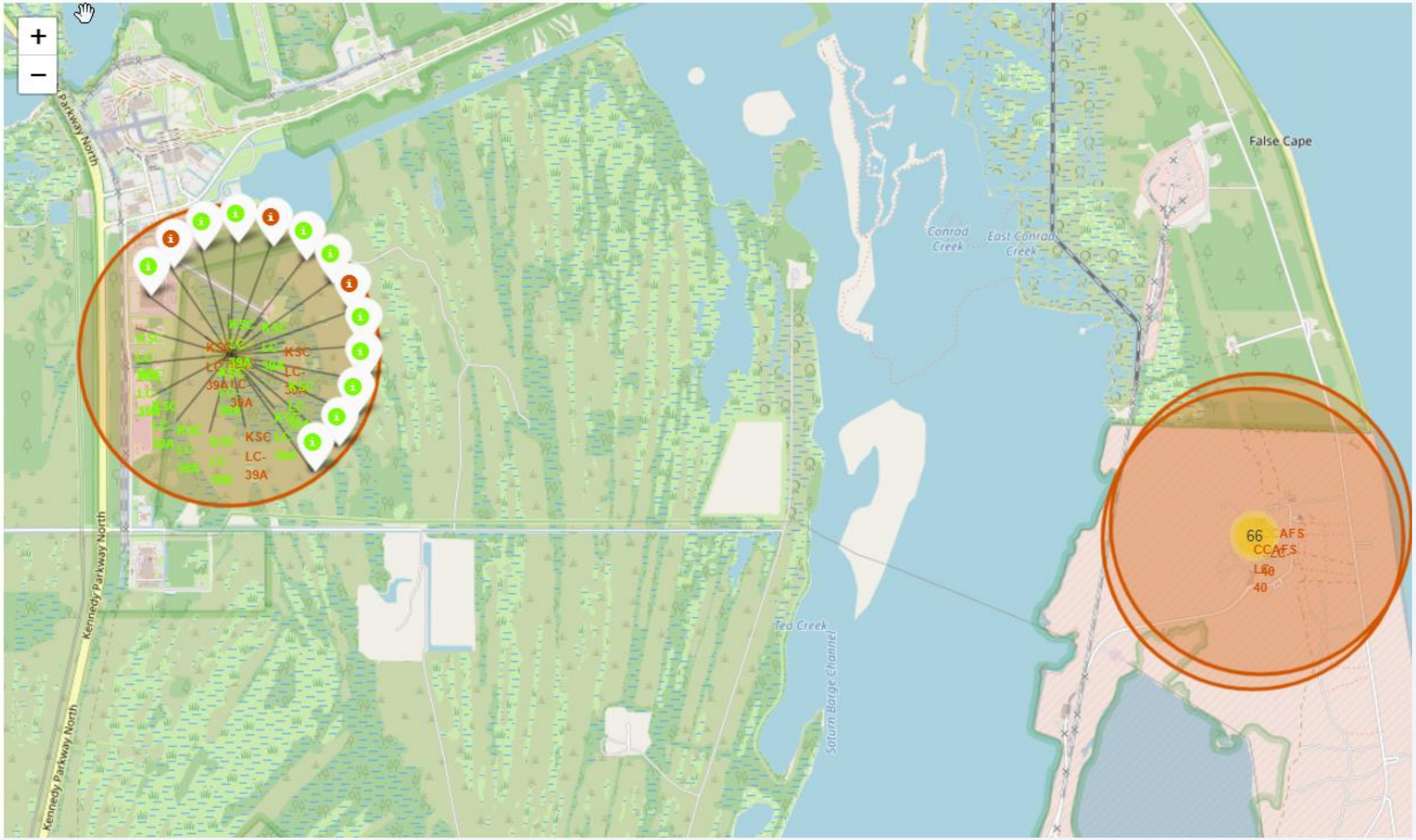
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Locations of the launch sites

Map of launch outcomes for each launch site



<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

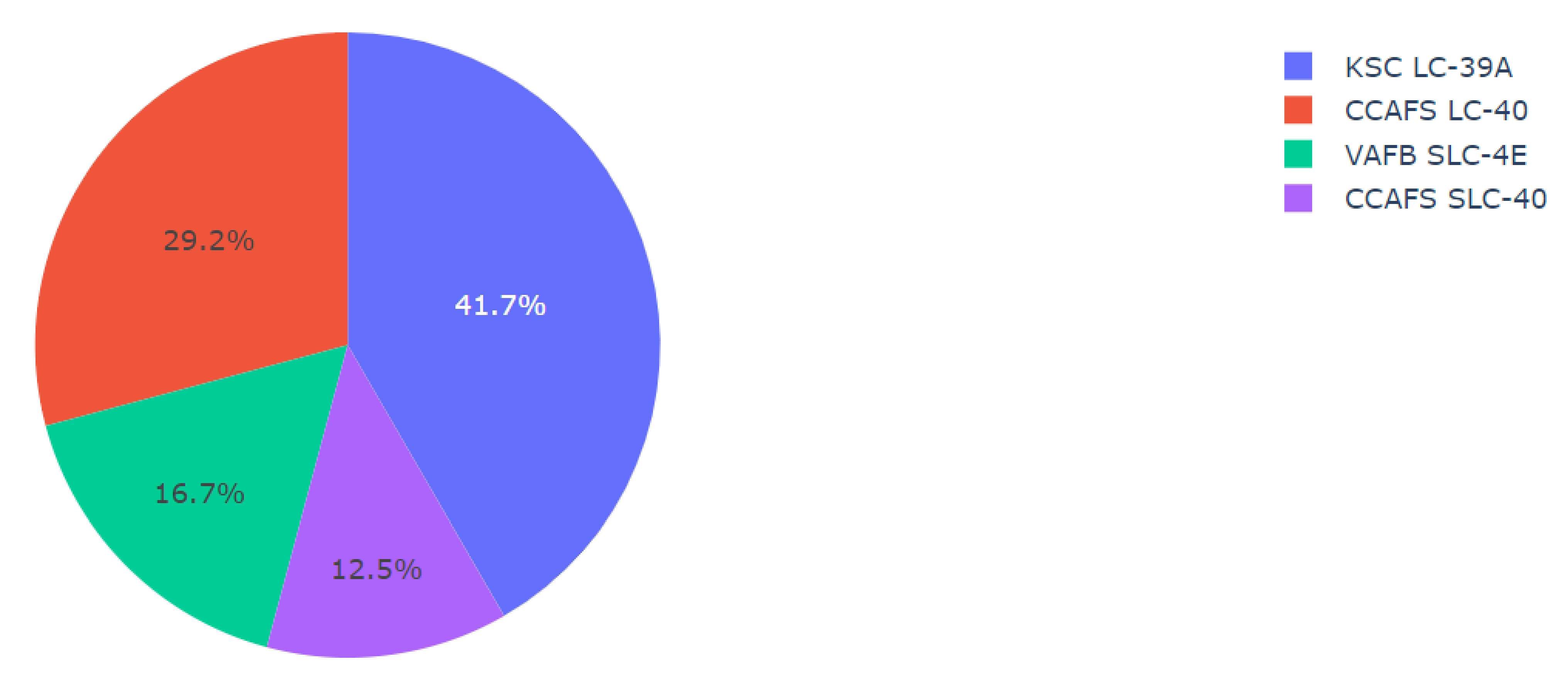
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

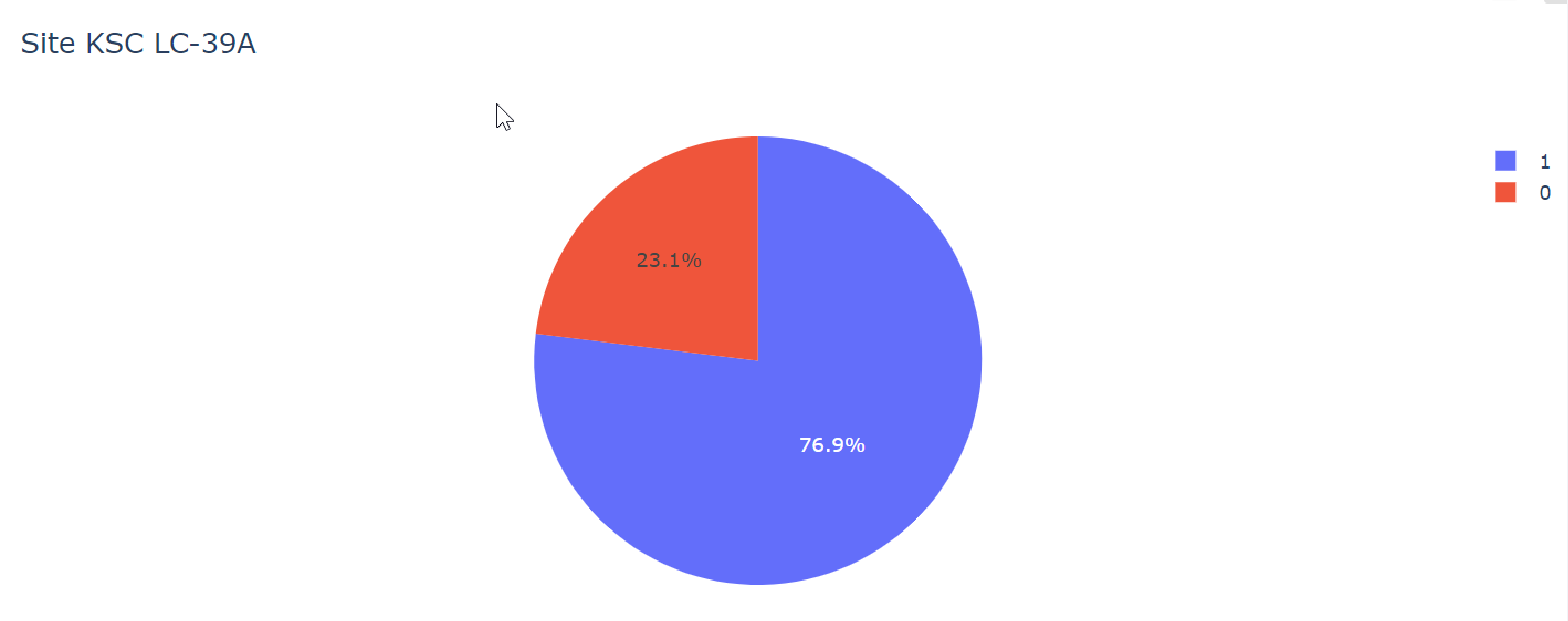
Section 4

Build a Dashboard with Plotly Dash

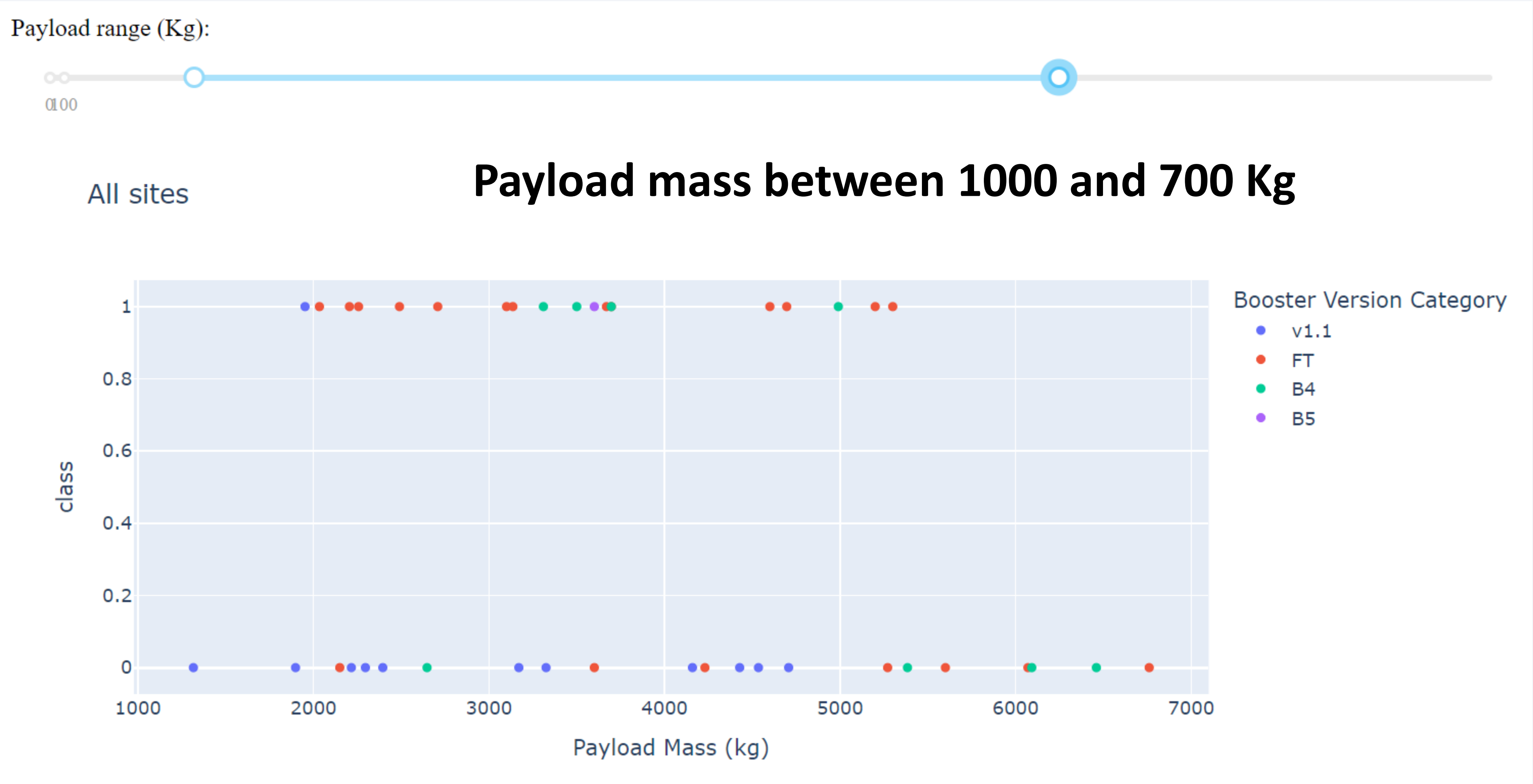
Launch success count for all sites



Launch site with the highest success rate



Success rate for different boosters



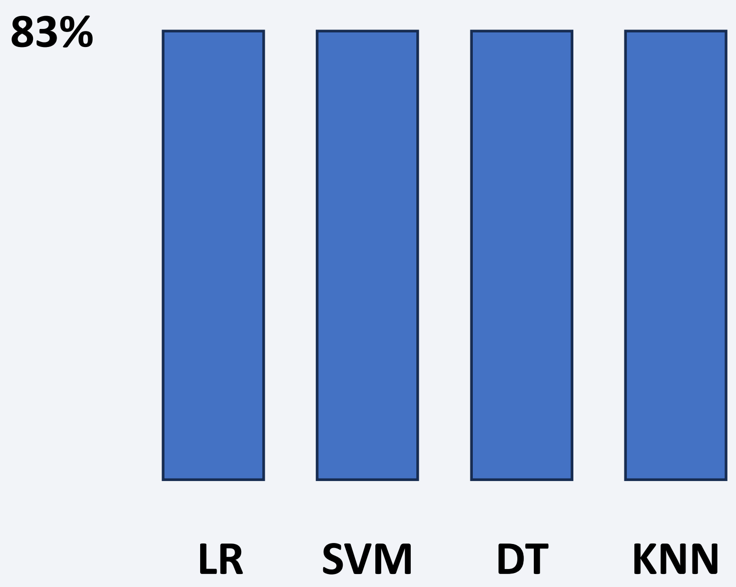
Section 5

Predictive Analysis (Classification)

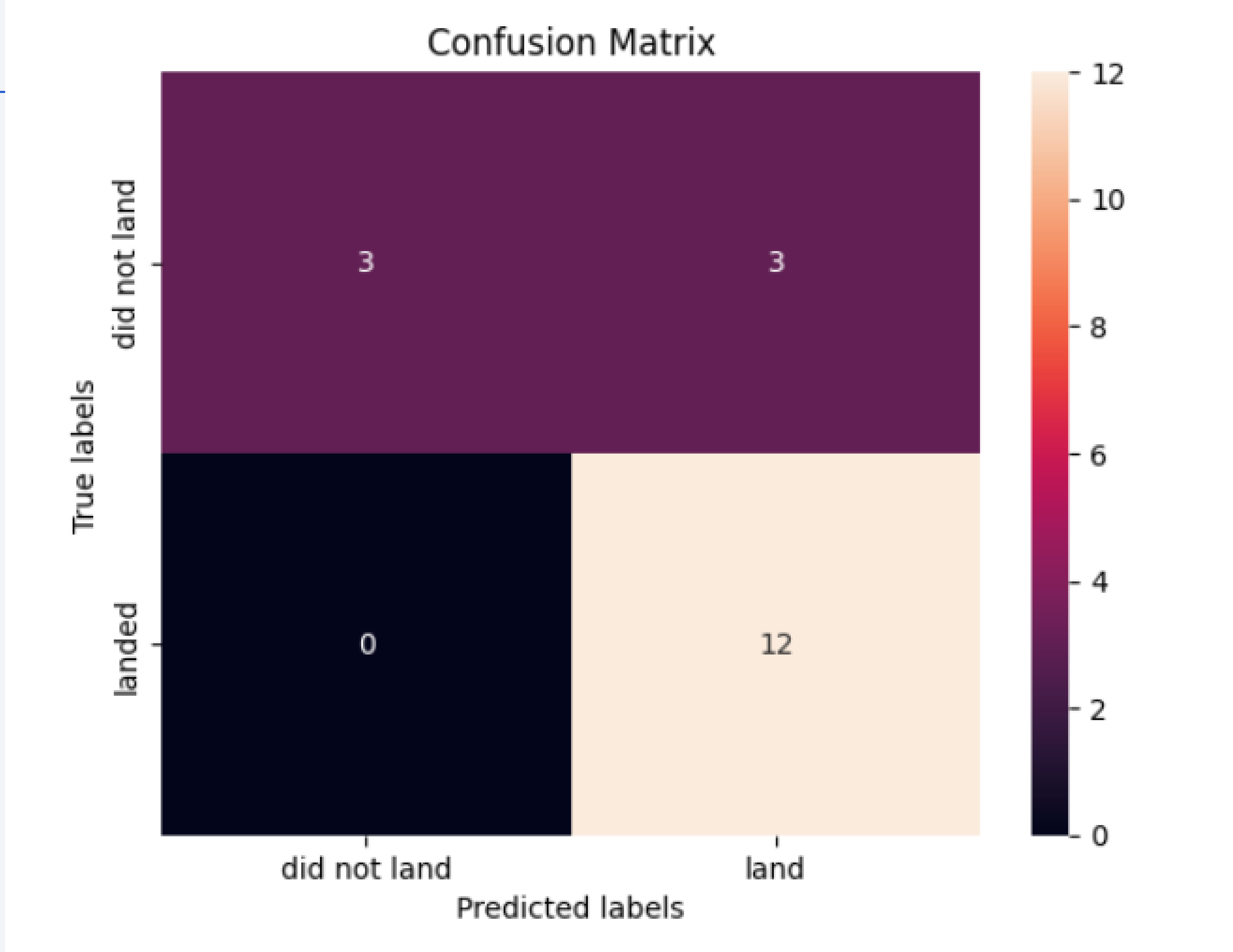
Classification Accuracy

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

- Find which model has the highest classification accuracy



Confusion Matrix



- Confusion matrix of the best performing model

Conclusions

- We can observe a continuous improvement in the success rate over time and flight number
- All classification model evaluated perform roughly the same with a R2 score of 83%

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!