IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abdelrahman Al-Shwayat

Outline

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

Executive Summary

- The project is about predicting if SpaceX's Falcon 9 rockets will land successfully in the first stage. SpaceX's provides it rockets at a much lower cost of 62 million dollars as compared to other providers which is 165 million dollars, this is because SpaceX's can reuse the first stage - IBM.
- In this project I applied techniques such as API, web scraping for initial data collection, data wrangling for restructuring the data, EDA with visualization to get hidden insights, patterns and trends. Finally, I preprocessed the data using feature engineering before applying four different machine learning algorithms for predictive modeling to know whether the first stage of the Falcon 9 landing will be successful.
- The results of the algorithms are almost all the same with an accuracy score of ~83.33%.

Introduction

- I am a data scientist from SpaceY a rocket company founded by billionaire Allon Mask and my job is to use public information about competitor SpaceX to determine whether the first stage landing of their rockets will land successfully. Since SpaceX's Falcon 9 is reusable. So that information from my research will be used by my company to bid against SpaceX's.
- I initially started my research by first collecting publicly available information from SpaceX's API by making a request using the HTTP and BeautifoulSoup then flatten it into a pandas table for further analysis.
- I later wrangle and restructured the data using pandas, the analysis was done with sqlite , matplotlib and seaborn for visual representations of my findings about the relationship between some of features which I used for prediction.
- I then used Folium for visualizing the launch site and dash for interactive dashboard about the hidden relationship of whether the payload affect the successfully landing, or how the Orbit type and Flight Number affects success rate of landing.
- The predictive modeling is mostly classification and binary using Logistic regression, Support vector machine, Decision tree and KNN with GridsearchCv for hyperparameter tuning. However, the dataset has a class ratio of 66.67%(class 1) success rate and how model can accurately predict 83.33% of it correctly.

Section 1: Methodology

Methodology

Executive Summary

Data collection methodology:

The data was collected using SpaceX API and Web scraping from wikipedia

Perform data wrangling

Using pandas, I flatten the data into tabular structure, cleaned missing values.

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Most of my work with models was were tune using GridsearchCv.

Data Collection

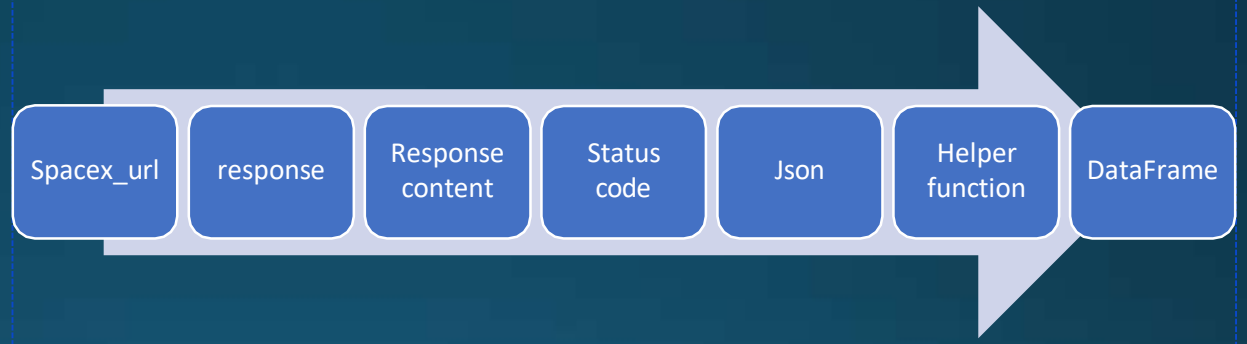
- After a thorough research(reading) about SpaceX, I decided to use their publicly free API using HTTP python library to get the data in a json format. I then converted the json result into a data frame to extract all necessary information that will be useful to my analysis. I filtered all “Falcon 1” data remaining with only “Falcon 9” which is of most interest due to its successfully landing history over the past years.
  I cleaned all missing values at this early stage leaving only LaunchingPad with nulls indicating when no LauchPad was used during the launching process.

- Here is a flow chart to show the process:

Data Collection – SpaceX API

SpaceX URL , response,
Json,
Data Frame

GitHub SpaceX API calls notebook



Data Collection - Scraping

Static URL, Beautiful Soup, Data frame

GitHub URL web scraping notebook

Data Wrangling

- The cleaned data was first imported, and I checked what percentage of missing values is in the LaunchingPad since it was the only column still containing missing value indicating when “no LaunchingPad” was used. I then checked the datatype of each column, there are 4 different datatypes(int64, object, float64 and bool). Further analysis like value count of LaunchSite was also examined for the various facilities Cape Canaveral Space Launch Complex 40 VAFB SLC 4E seem to have the highest count of 55.
- I created a new feature called “class” from the outcome's column were all outcome containing the name “False” and “None” are regarded as bad, therefore a of zero(0) was assigned and one(1) for good outcome.
- I also calculated the success rate of all the good outcome which make up 66.67% of the class feature.

- GitHub URL data wrangling related notebooks

EDA with Data Visualization

I extensive utilized seaborn and matplotlib for the visualization of static charts, plot type such as catplot for displaying the relationship between PayloadMass vs LaunchSite, barplot for orbit vs class and line plot to show the trend of average success rate of rocket launches over time(2010-2020).

- GitHub URL EDA with data visualization notebook

EDA with SQL

- Names of Unique launch sites in the space mission
- Five records where launch sites begin with string ‘CCA’
- Total payload mass by **NASA (CRS)**
- Average payload mass carried by booster version **F9 v1.1**
- First successful landing outcome in ground pad
- Success in drone ship, payload mass **>4000** but **<6000**
- Total number of successful and failure mission outcomes
- Booster_versions with maximum payload mass
- Month names of failure landing outcomes in drone ship in year **2015**
- Landing outcome counts between "2010-06-04" and "2017-03-20“ in descending order

- **GitHub URL EDA with SQL notebook**

Build an Interactive Map with Folium

In the interactive map with Folium, I added objects such as markers which is used to pin the location of the launches, examples is the NASA JSC space station launch site, circle to highlight the area with a text label on a specific coordinate and line for the proximity of the launchsites.

GitHub URL interactive map with Folium map

Build a Dashboard with Plotly Dash

- In the Plotly Dash user application I have added a dropdown list and a range slider to allow a user to interact with a pie chart and the scatter point chart

Predictive Analysis (Classification)

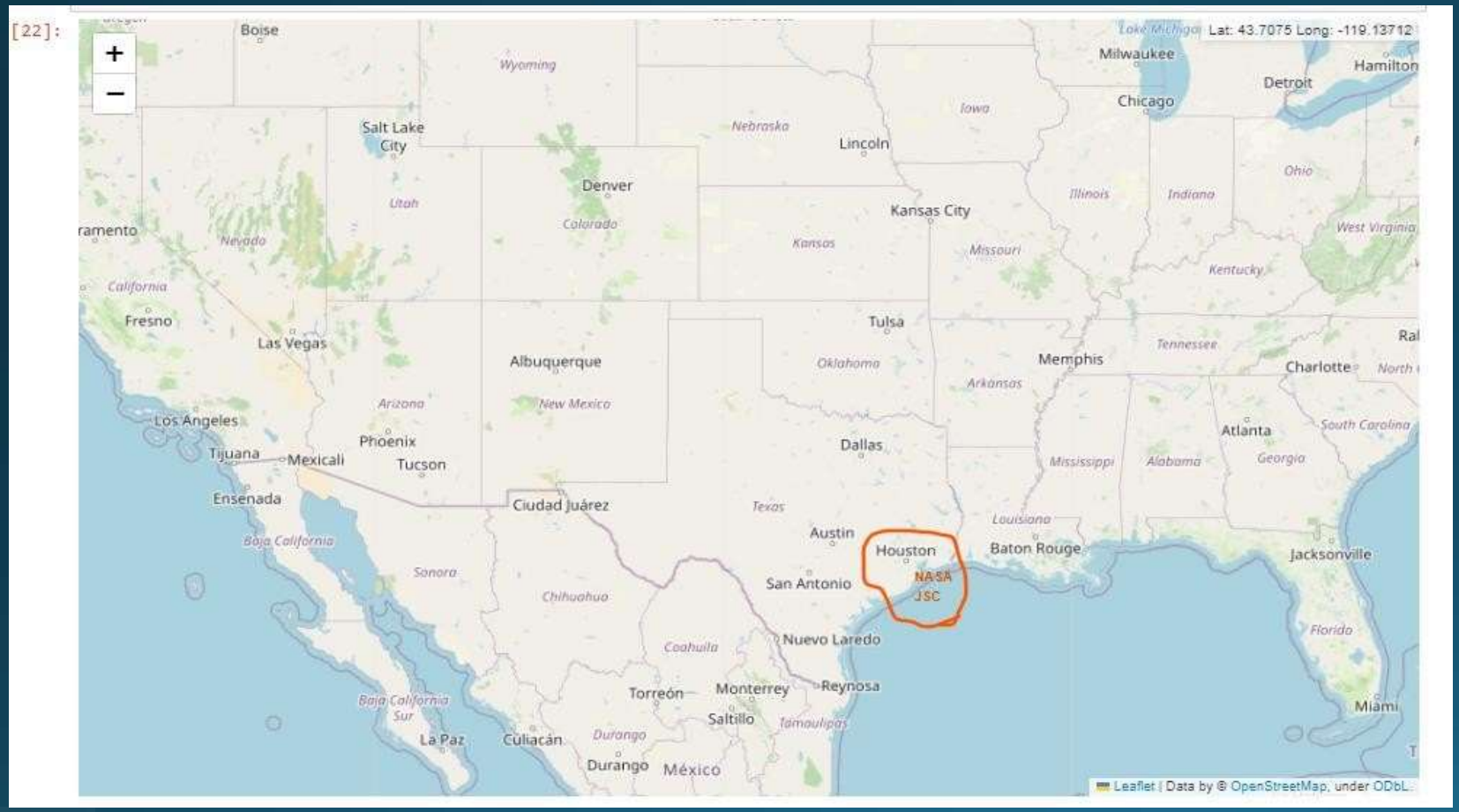
- The cleaned data was imported, and I assigned the features to a variable (X), and target to the variable(Y). The features were then scaled using sklearn's the standard scaler() object, split with inbuilt "train_test_split" with a portion of 20% assigned to the testing.
- The models were instantiated, hyperparameters assigned using "GridSearchCV" for optimal
- best performing parameters.
- Fitting was done on the training-set while the evaluation on the unseen test-set, the "DecisionTreeClassifier" models achieved the best results with an accuracy of 88.88% and F1-score of 88.21% using "GridSearchCV" for hyperparameter tuning.

- GitHub URL predictive analysis lab

Results

- The finding of the exploratory stage were quite interesting with NASA (CRS) having a total “payload mass” of 45,596, average payload_mass of “booster version F9 v1.1” is 2928.4, first successful landing outcome ground pad was achieved in “2015-12-22”, and up to 5 Booster_version having a successful landing_outcome on “drone_ship” with payload_mass between (4000,6000). A total of 99 success mission_outcome was also achieved by SpaceX’s Falcon 9.
- The visualization also depict some interesting findings, with higher success rate for CCAFS SLC 40 as the filight number increases, and higher success rate(100%) of Falcon 9 landing for Obit type “ ES-L1, SSO, HEO and GEO). Also there has been tremendous success in Falcon 9’s first stage landing over time since 2010 to 2020.

Results



Nasa Johnson Space Station (NASA JSC)

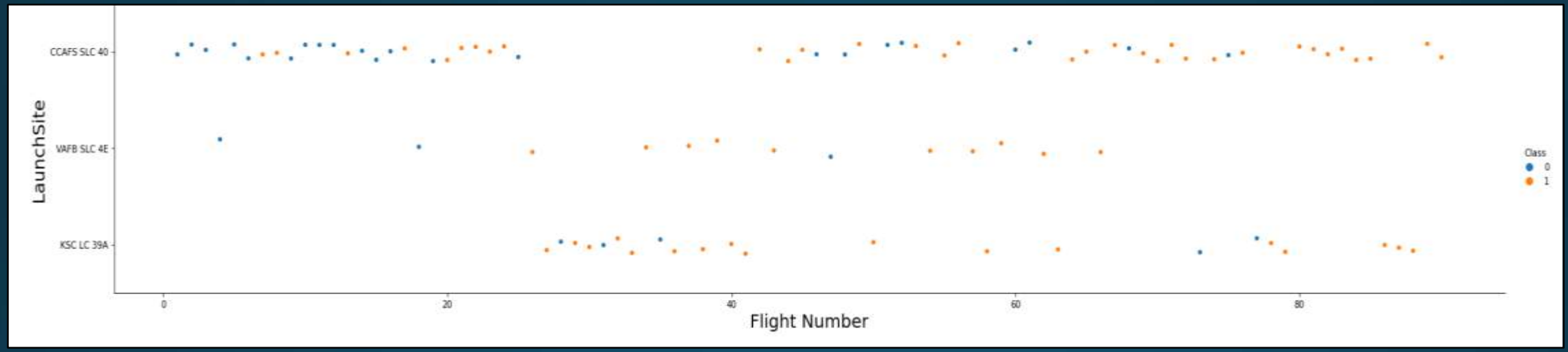
Source: Github

Results

- I used four(4) different machine learning algorithms the baseline Logistic Regression, Support Vector Machine, Decision Tree classifier and KNN. All the models yield a better performance with and accuracy above 83%, while Decision Tree with GridSearchCV achieved the highest accuracy of 88.89% and F1-score of 88.21%. Meaning we are confident enough that 88.89% of the time our model will be able to predict whether Falcon 9 first stage landing will be successful.

Section 2:
Insights Drawn from EDA

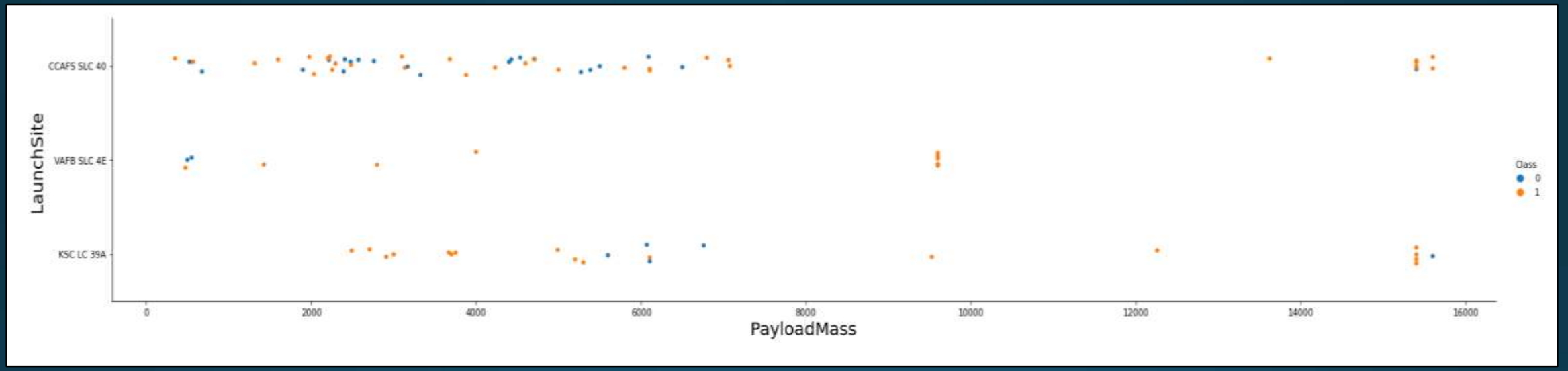
Flight Number vs. Launch Site



As show above the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".

Source: GitHub

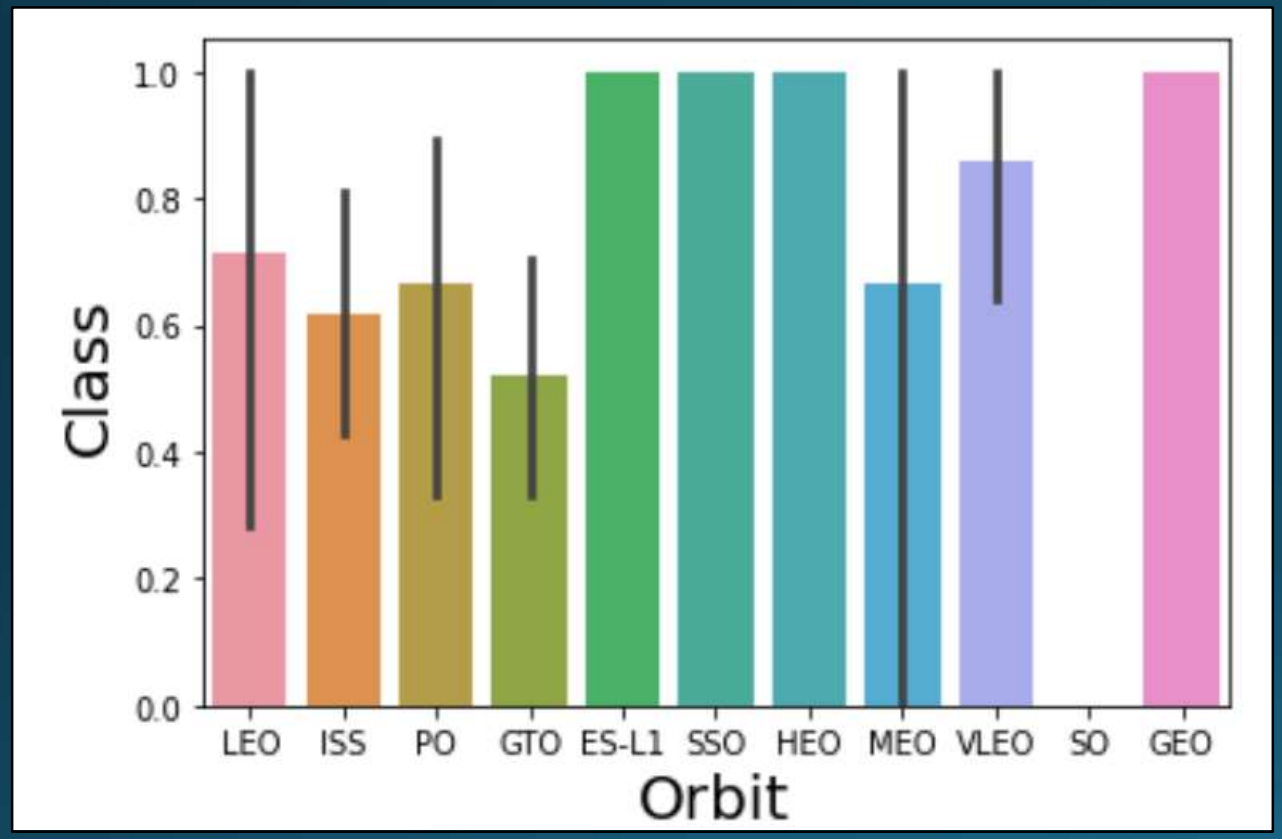
Payload vs. Launch Site



As show above the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E" and more failures than success for "CCAFS SLC 40". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".

Source: GitHub

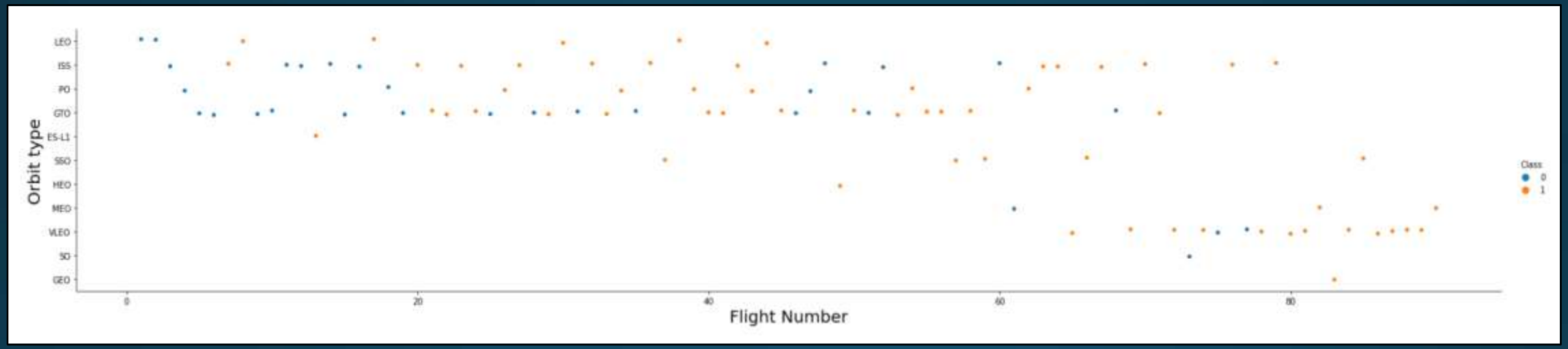
Success Rate vs. Orbit Type



Four(4) Orbit type have the highest success rate of 100% ES-L1, SSO, HEO, and GEO.

Source: GitHub

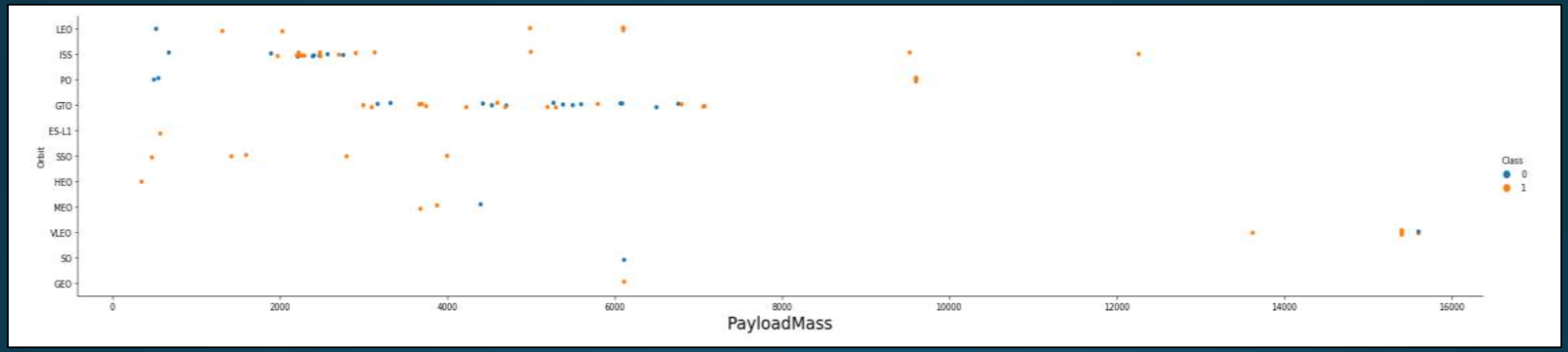
Flight Number vs. Orbit Type



You can see that in the VLEO orbit there seem to be success rate only when the flight number is greater than 60. While there seem to be no relationship in the GTO.

Source: GitHub

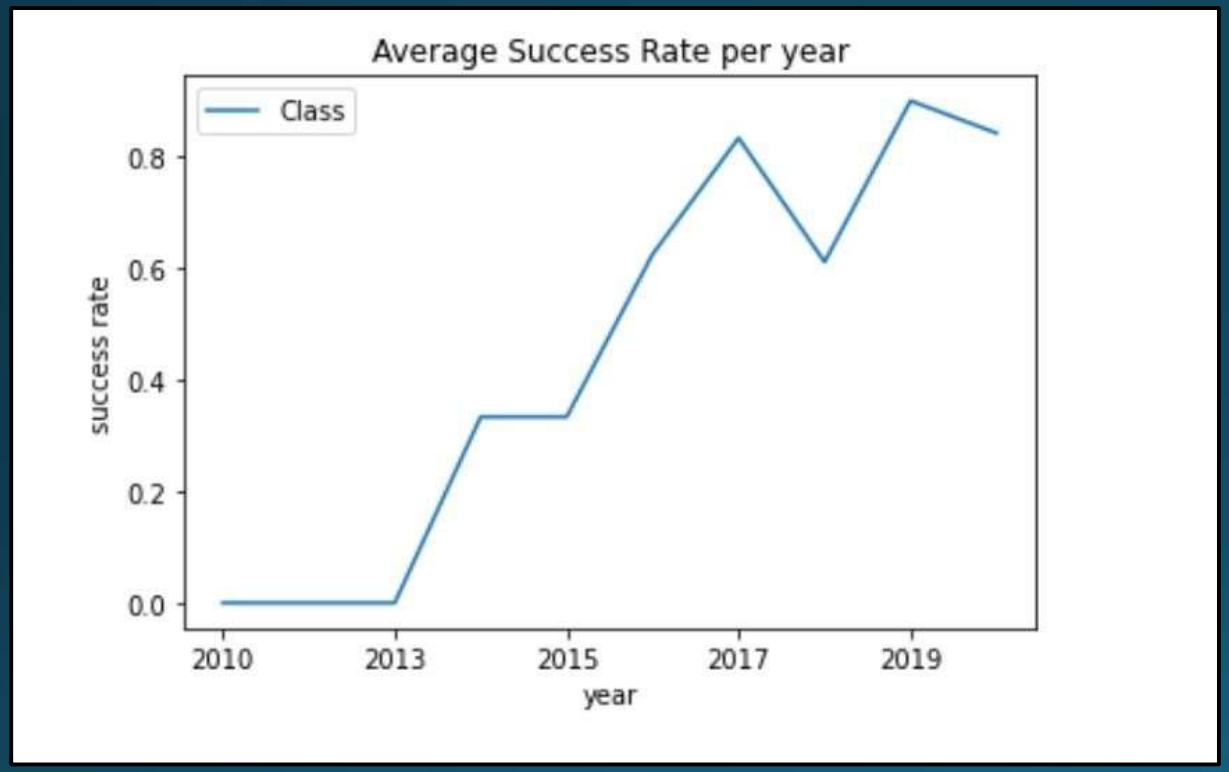
Payload vs. Orbit Type



There is more positive landing rate with heavy payloads for Polar, LEO and ISS, however for GTO there's the presence of both positive and negative landing.

Source: GitHub

Launch Success Yearly Trend



The yearly success rate of Falcon 9 landing kept increasing since 2013 till 2020.

Source: GitHub

All Launch Site Names

- The name of unique launch sites are:

CCAFS LC-40: Space Launch Complex 40 formerly **Launch Complex 40** (**LC-40**) is an orbital launch pad located in northern Cape Canaveral , Florida (Wikipedia).

VAFB SLC-4E: Vandenberg AFB Space Launch Complex 4 is a launch and landing site at Vandenberg Space Force Base, California, U.S. It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 for SpaceX landings(Wikipedia).

KSC LC-39A: Kennedy Space Center Launch Complex 39A Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida(Wikipedia).

Launch Site Names Begin with 'CCA'

Attached below is a list of the first five Launch Site whose names 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

Source: GitHub

Total Payload Mass

The total payload mass of boosters from NASA is 45596.

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

Source: GitHub

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.4

Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

Source: GitHub

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is:

Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

Source: GitHub

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as follows:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Source: GitHub

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes are 99 for success one for Failure (in flight) and 1 for success (payload status unclear).

Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Source: GitHub

Boosters Carried Maximum Payload

List of the booster which have carried the maximum payload mass

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Success	15600
F9 B5 B1049.4	Success	15600
F9 B5 B1051.3	Success	15600
F9 B5 B1056.4	Failure	15600
F9 B5 B1048.5	Failure	15600
F9 B5 B1051.4	Success	15600
F9 B5 B1049.5	Success	15600
F9 B5 B1060.2	Success	15600
F9 B5 B1058.3	Success	15600
F9 B5 B1051.6	Success	15600
F9 B5 B1060.3	Success	15600
F9 B5 B1049.7	Success	15600

Source: GitHub

2015 Launch Records

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

Date	Booster_Version	Launch_Site	Landing_Outcome
2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Source: GitHub

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

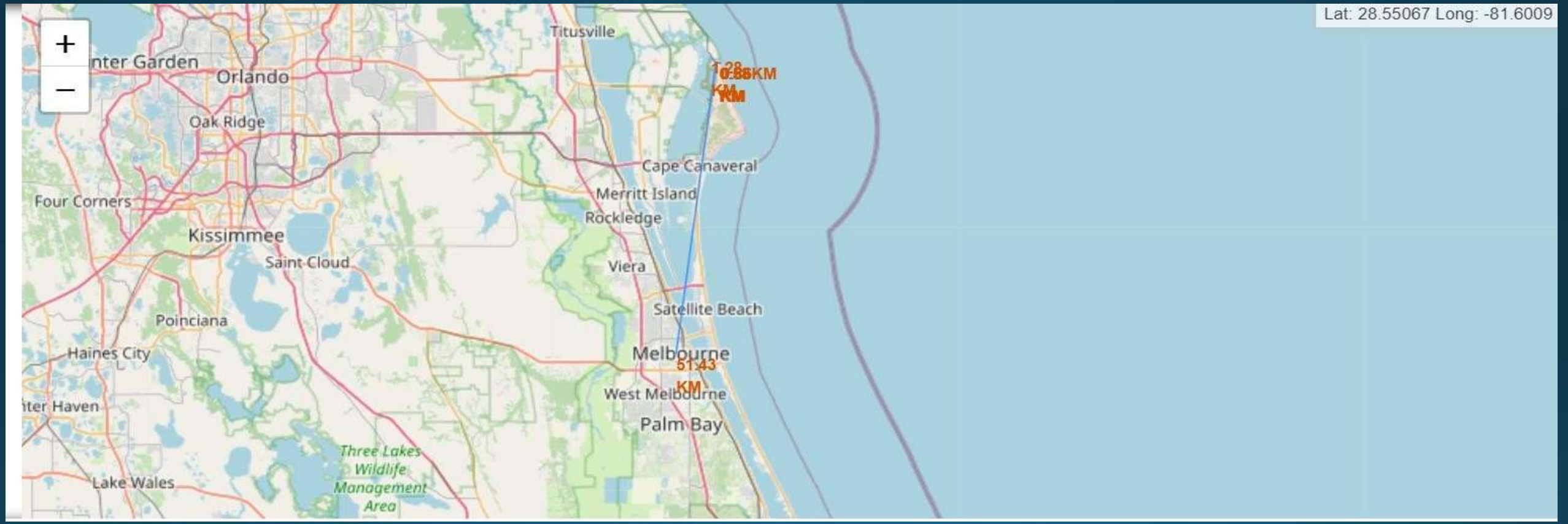
Rank of landing outcomes (such as Failure (drone ship) or Success (ground pad)) count between the date 2010-06-04 and 2017-03-20, in descending order.

Landing_Outcome	Total Count
Success (ground pad)	5
Failure (drone ship)	5

Source: GitHub

Section 3:
Launch Sites Proximities Analysis

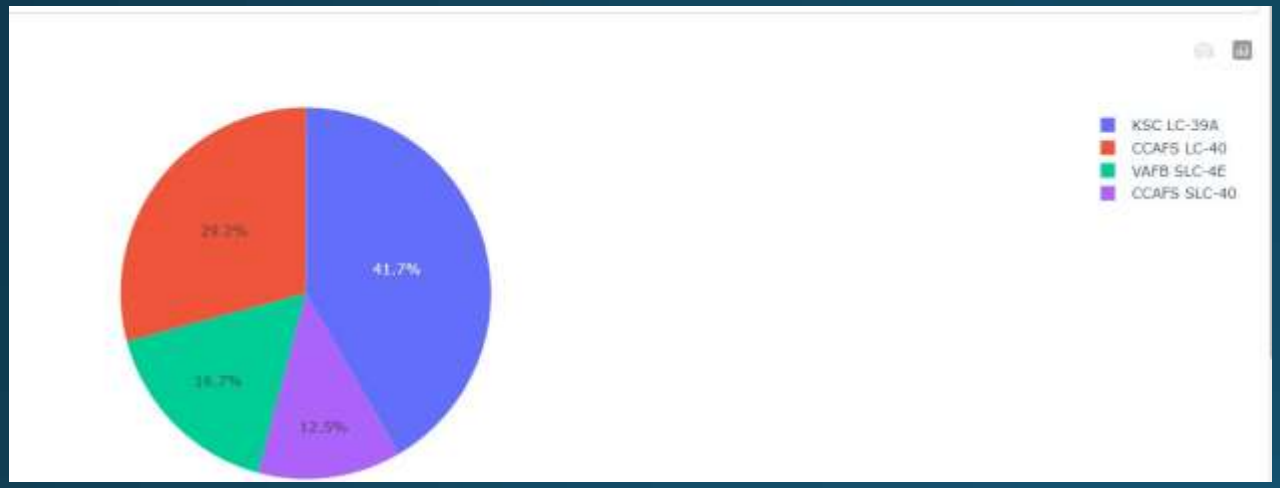
Proximities



proximities such as railway, highway, coastline, with distance calculated and displayed.

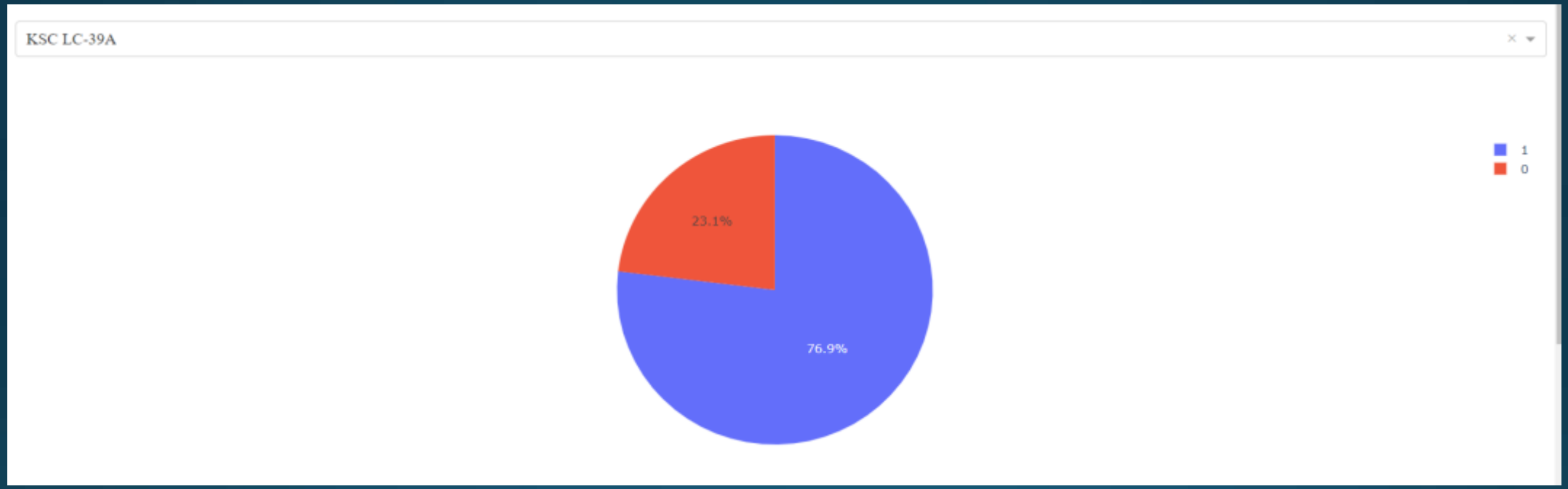
Section 4:
Building a Dashboard with Plotly Dash

Pie chart of all the launch sites



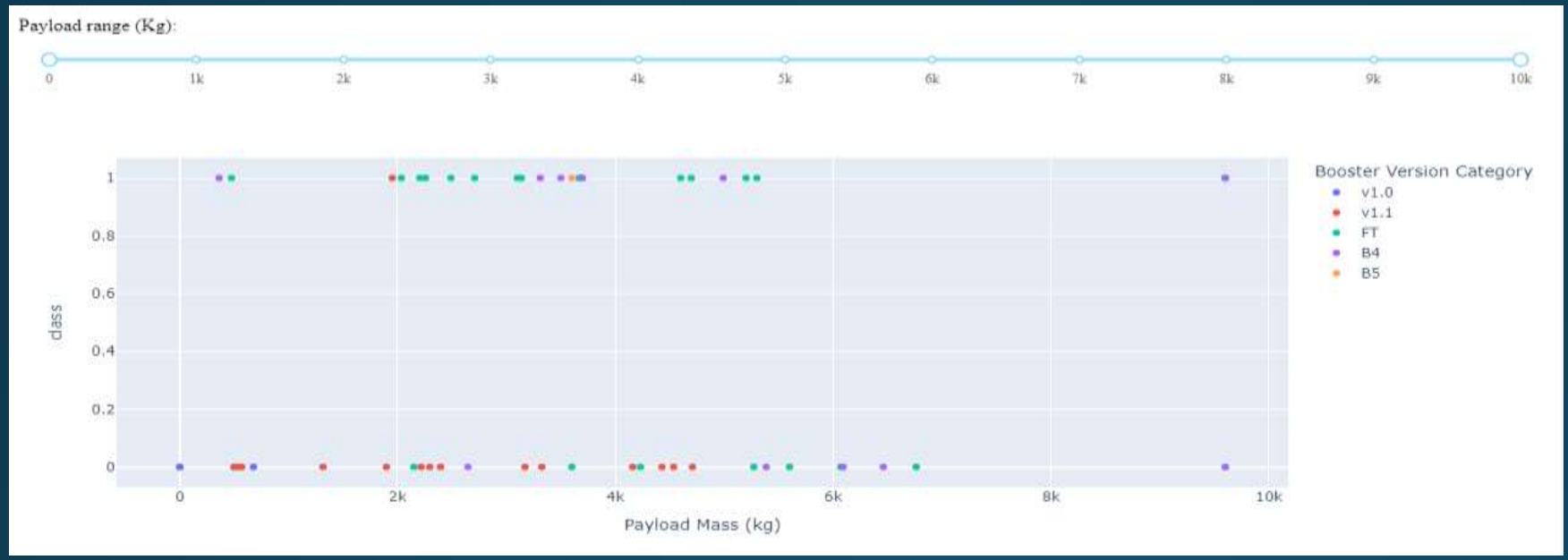
- KSC LC-39A makes up 41.7% of the piechart showing it has more launches than the other sites, followed by CCAFS LC-40 with 29.2%.

Pie chart of KSC LC 39A



Site KSC LC 39A has a higher count with a 76.9% success rate of Falcon 9 landing.

Payload vs. Launch Outcome scatter



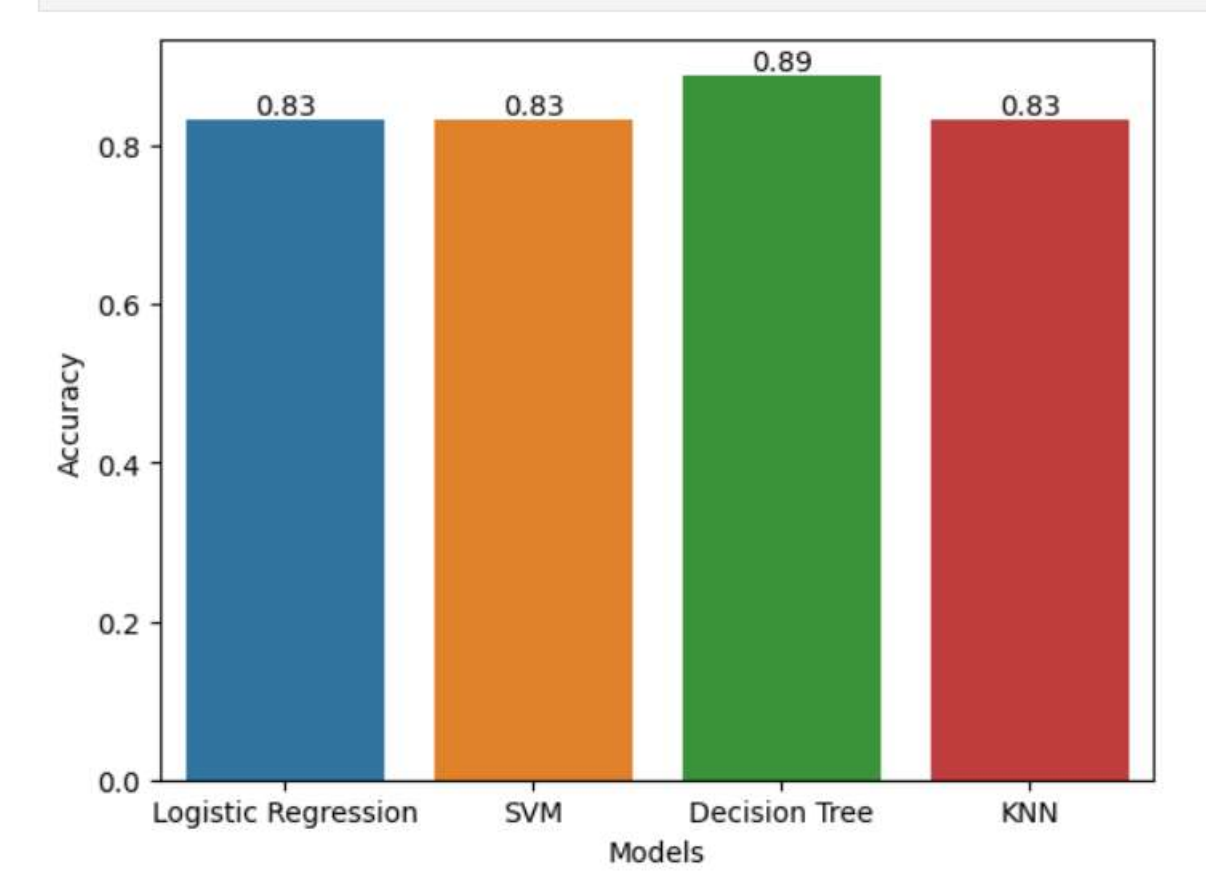
We can see in the scatter plot that as payload mass grow so does the success rate for the booster version FT, while the opposite for booster version v1.1 which has more failures.

Section 5

Predictive Analysis (Classification)

Section 5:
Predictive Analysis
(Classification)

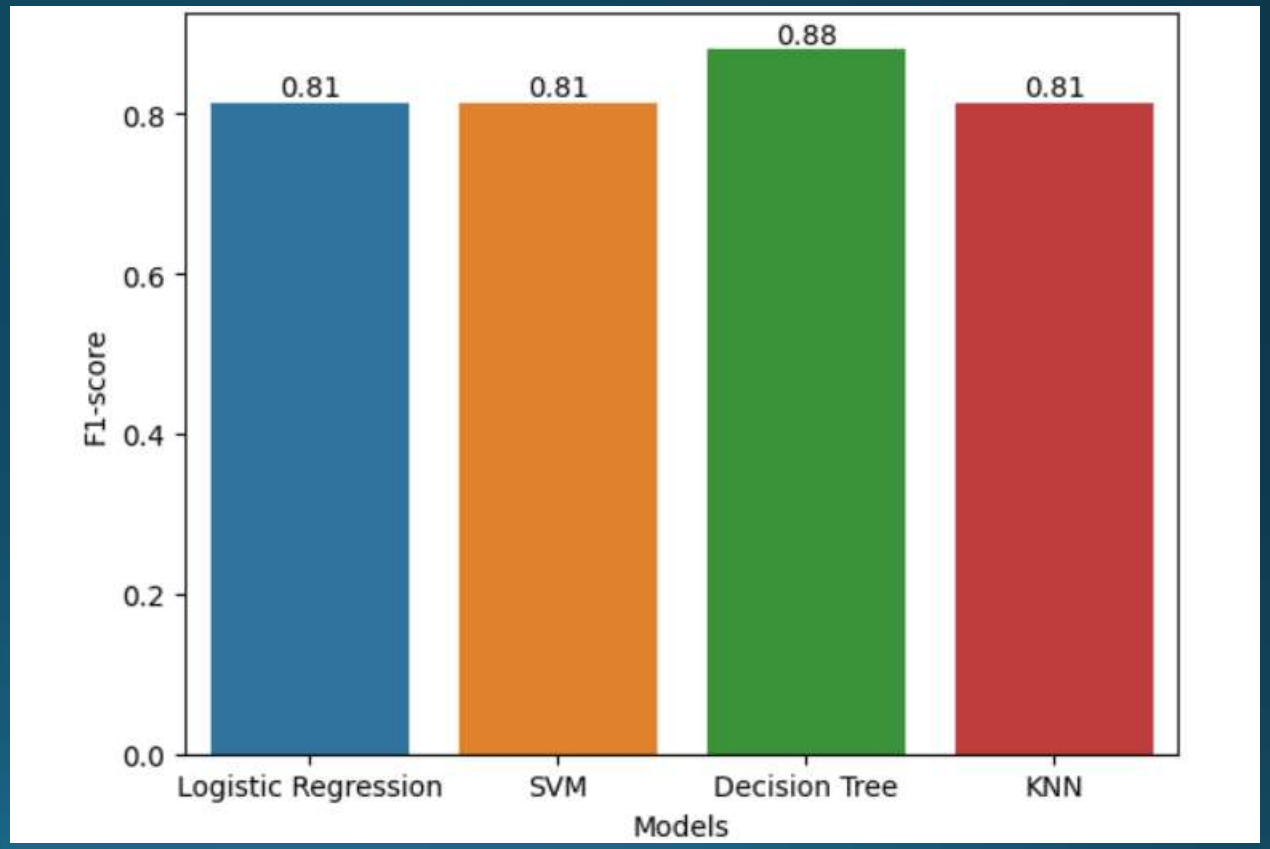
Classification Accuracy



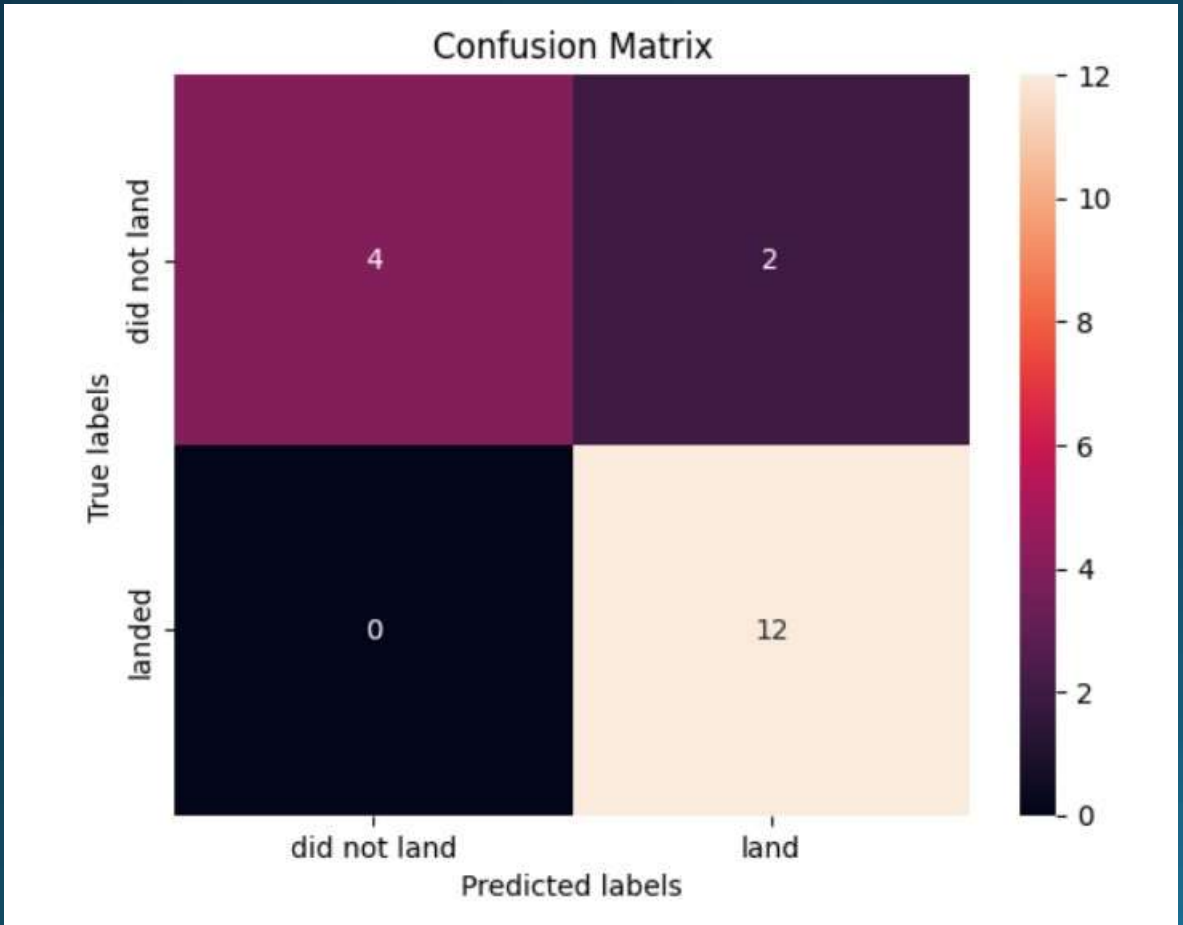
Decision Tree was the best classifier with and accuracy of 88.88%

Classification F1-score

Decision Tree also achieved an F1-score of 88.88%.



Confusion Matrix



The Decision tree model has 2 false-positive and no false negative and achieved a precision accuracy of 85.71% while recall is 100%.

Conclusions

- As the flight number increases there's more success rate for Launch Site CCAFS SLC 40, the same relationship was observed for payload mass.
- There's 100% success rate of landing for Orbit type ES-L1, SSO, HEO, and GEO
- There's been stead increase in the success rate of Falcon 9 landing since 2010-2020
- With a very high accuracy of 88.89%, we know that our model can predict the outcome of the landing.

Appendix

All relevant assets like Python code snippets, SQL queries, charts, and data sets included in this presentation can be found on my **GitHub**.

Thank You!