IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Benjamin Mills
October 21, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Insights drawn from EDA
- Launch Sites Proximity Analysis
- Building a Dashboard With Ploty Dash
- Predictive Analysis (classification)
- Conclusion
- Appendix

Executive Summary

- Data was collected from a static URL with accurate rocket information which allowed us to analyze for creating the rocket landing model
- The data was wrangled to replace missing values with averages and create an outcome column to indicate a successful or failed landing
- Explorative data analysis was done to narrow down on what we should focus on when creating the model, this was done through SQL and creating graphs
- Folium maps were made to see if what physical landmarks were near launch sites and how that may impact success rate
- An interactive dashboard was created to further explore success rate by specific launch sites
- Finally, a predictive model was created to predict success rates, and a decision tree model was ultimately chosen for its higher rate of accurate predictions

Introduction

- The age of space travel is here and with it, space travel for the average person. One major problem with space travel is the expensive cost of single use rockets. To solve this issue, companies have begun making use of multi-use rockets that can land after launch. This is a large leap forward in affordable rocket use, but some questions arise that this presentation will be addressing, which include:
- What is the cost of a launch?
- What data types impacts the success rate of a launch (Payload mass, Orbit type, etc)?
- What physical structures are common between launch sites?
- What model best predicts the success of a launch?

Section 1

Methodology

Methodology

Executive Summary

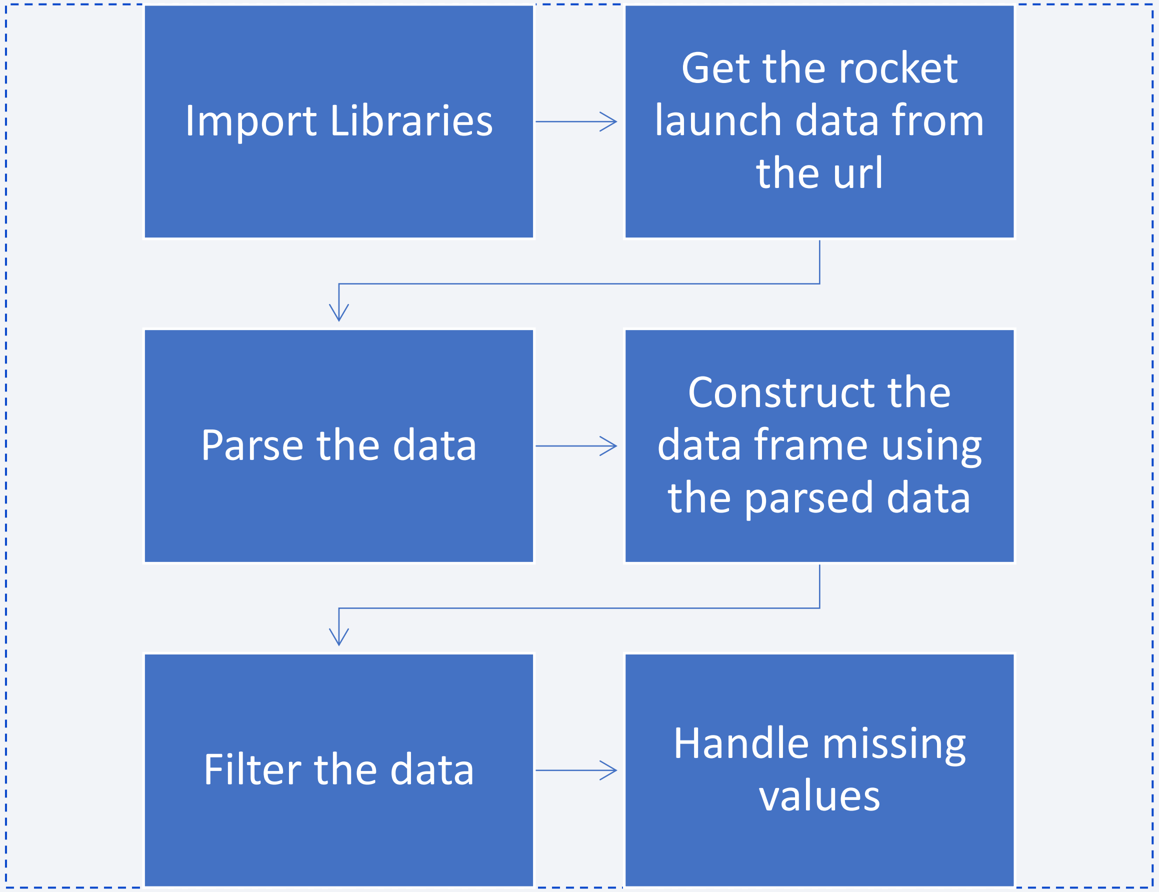
- Data collection methodology:
  - Collected through the provided static URL’s
- Perform data wrangling
  - Processed to replace missing values and create necessary columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models were created and the one with the highest best score and highest test scores was chosen

Data Collection

- Data was collected from the static URL's
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json and https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The data was collected from static URL's to be more consistent but also means that they are missing the most up to data information
- The data was put into data frames for preparation of exploratory data analysis or EDA
- Introductory information like the data variables we will be including in analysis was decided here

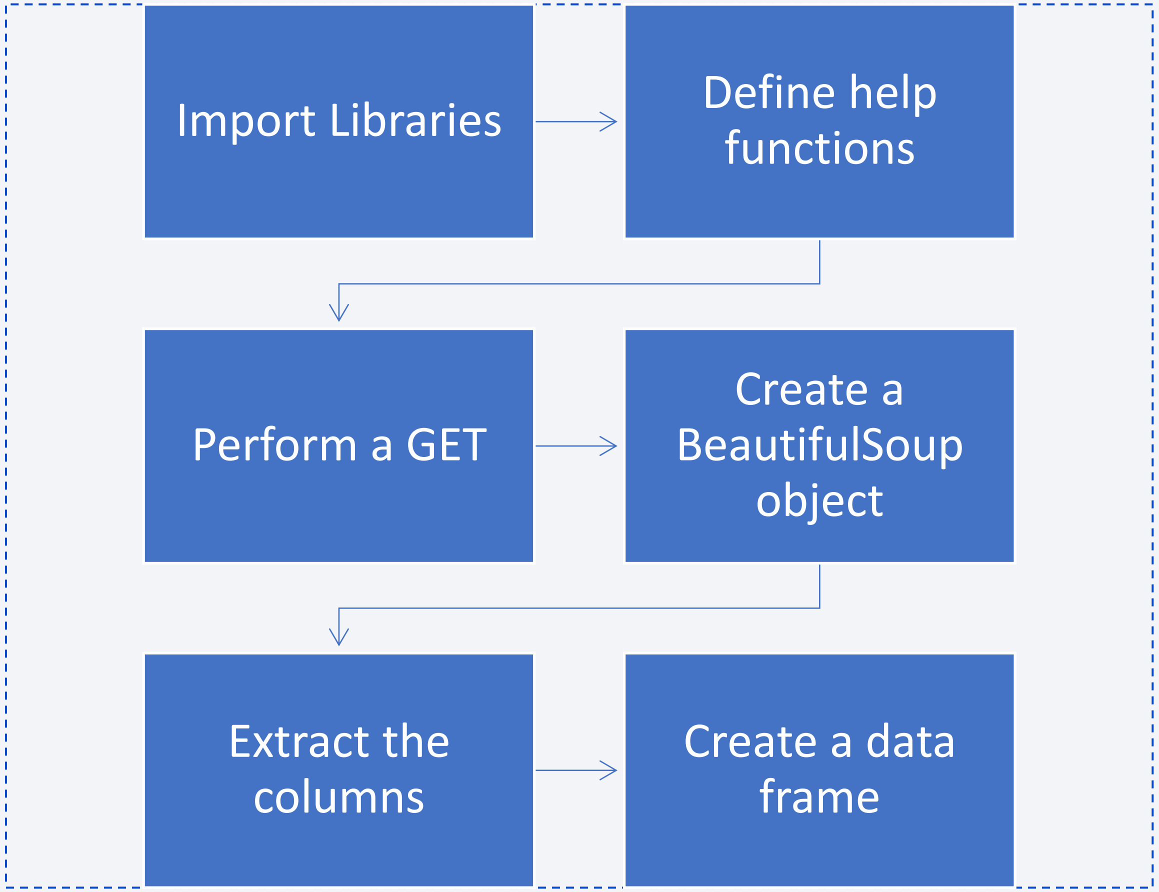
Data Collection – SpaceX API

- First libraries were imported and named for convince
- The rocket data was taken using the url provided
- The data was parsed to get the information we wanted
- A data frame was created with the columns that we wanted for this project
- The data was filtered to only include Falcon 9 launches
- Missing values for PayloadMass was replaced with the average
- https://github.com/Ben625e/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



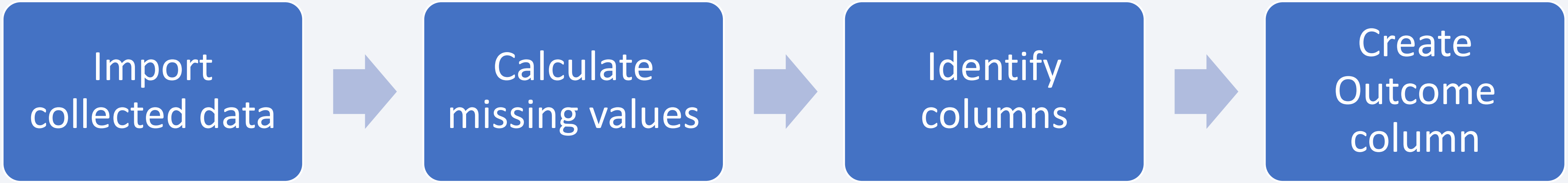
Data Collection - Scraping

- First I begun by importing the libraries that were necessary for this project
- Then defines helper functions that would be used later
- Performed a GET using the provided URL
- Created a BeautifulSoup object from that GET
- Extracted the columns from the BeautifulSoup object by reading for ‘table’
- Created an empty data frame and filled it with the data by reading the soup object
- https://github.com/Ben625e/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- First the data we collected from the previous section was collected
- Then some initial looks at the data like identify column types and finding percentage of missing values was done
- After calculating the number of launches on each site and calculating the number of orbits
- A set of bad outcomes was created which was used to create the Outcome column
- https://github.com/Ben625e/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plots of Flight Number vs Payload mass, Flight Number vs Launch Site, and Flight Number vs Orbit were made to get an idea of how the respective Y variable affected landing outcome
- A scatter plot of Payload mass vs Launch Site was made to compare what payloads of the different launch sites
- A bar graph was made comparing the success rate of the different orbit types to see if certain orbit types were more successful than other's
- A line chart was made comparing the data with the success rate to see how success rate has changed as time went one
- https://github.com/Ben625e/DataScienceCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Queries were made to find the unique launch site names and the first five records where the launch site started with “CCA…”
- The total payload mass and average payload mass of the “F9 v1.1” booster was then made
- The date of the first successful ground pad landing and the list of boosters who had a successful drone ship landing with payloads between 4000kg and 6000kg were made
- The count of the mission outcomes followed by the boosters who carried the maximum payload mass of 15600kg were made
- Finally, failure in drone ship for 2015 and count of the landing outcomes between 2010 and 2017 were made
- https://github.com/Ben625e/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- For the Folium maps, I used markers and polylines to highlight key points of interest and to better gauge distance
- I used markers, circles, and marker clusters for the launch sites to label and allow the launch sites to be interacted with to see the successful and failed landings
- The map also made use of markers for labeling the distance between the launch sites and the key points of interest
- https://github.com/Ben625e/DataScienceCapstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A dropdown containing the launch sites was made to select either a specific site or all of them
- A pie chart indicating the success rate for landings of the selected site(s) was made
- A payload range slider was added for creating a scatterplot of the payload versus the success rate for landing of the selected site(s)
- Unfortunately, I was having an error that made it unable to run the code so I got stopped, I will put the error in the appendix and be uploading pictures from the lab itself to replicate what an answer would be look like
- https://github.com/Ben625e/DataScienceCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was first loaded in and handled to be used for the model building
- The data was split with a test size of 0.2 (20%) and a random state set to 2 to have randomness that could be replicated
- Logistic regression, support vector, decision tree, and k nearest neighbor models was built and tested
- The testing used both best score and compared to test values
- The model which performed the best for both scores was the decision tree
- https://github.com/Ben625e/DataScienceCapstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

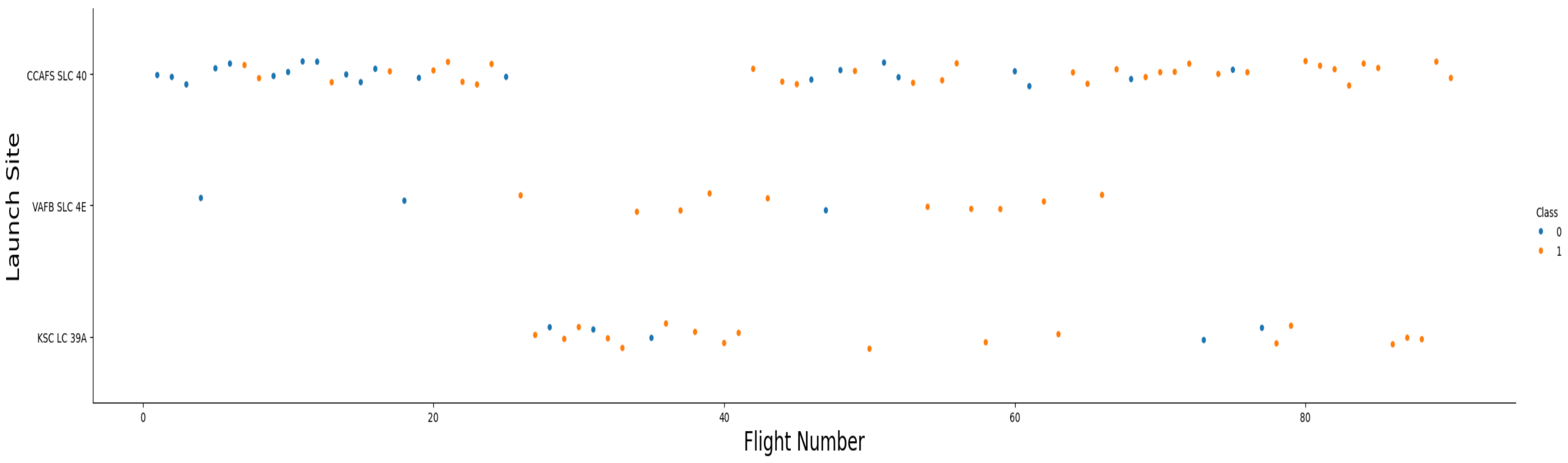
Results

- Through EDA, I found that higher payloads tend to have a higher success rate
- Different orbit missions also impact the success rate with SO having the lowest rate of success and ES-L1, GEO, HEO, and SSO having the highest
- Over the years, successful landings have increased and hit a peak in 2019
- Launch sites tend to be built near coastlines, railways, and highways
- Launch sites tend to be built far away from cities
- Out of all the predictive models built, a decision tree classification model had the highest rate of success

Section 2

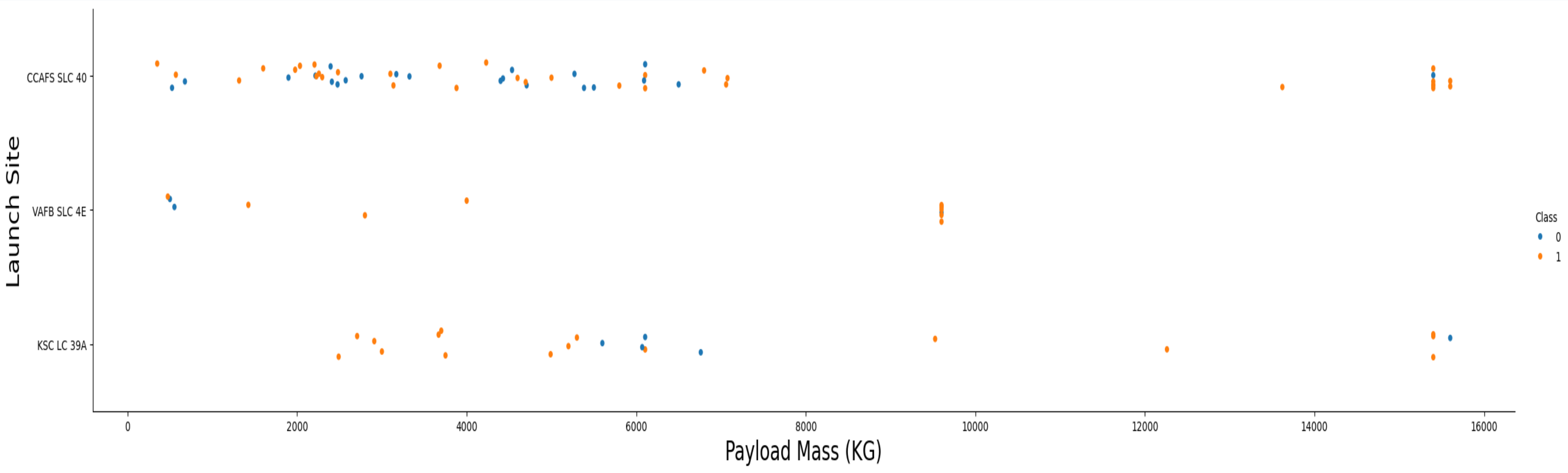
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatterplot shows the flight number by the launch site that flight was taken off at
- It includes the class in coloring, blue to indicate a failure to land and orange to indicate a successful landing
- This graph indicates that VAFB hasn’t had many launches but has a good recorded of successes while CCAFS has had many launches and has a high failure rate

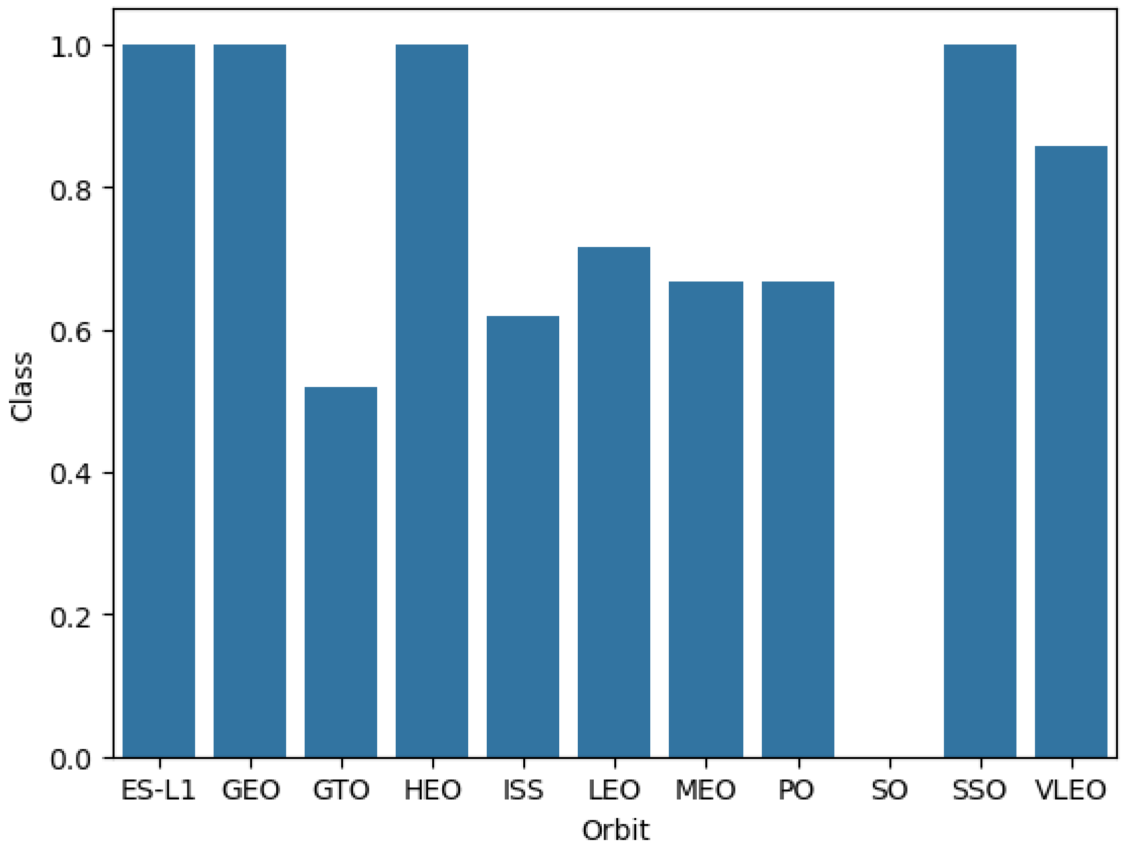
Payload vs. Launch Site



- This scatterplot shows the pay load mass in kilograms by the launch site
- It includes the class in coloring, blue to indicate a failure to land and orange to indicate a successful landing
- This graph indicates that on smaller loads for CCAFS and VAFB, flights tend to fail to land while at higher loads, flights tend to land successfully
- It also shows that VAFB has a low maximum load of under 10kg while CCAFS and KSC can reach loads of almost 16kg

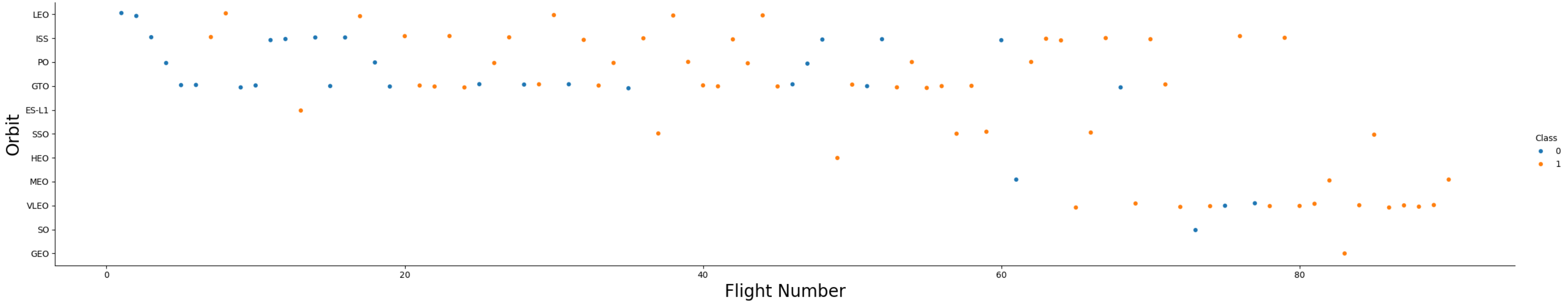
Success Rate vs. Orbit Type

- This bar graph shows the success rate by the orbit type
- This graph indicates that certain orbit types have high success rates like ES-L1, SSO, KEO, and GEO
- It indicates the SO launches have not had a successful landing according to the data
- This graph doesn't consider the number of the launches that have occurred which may skew the results



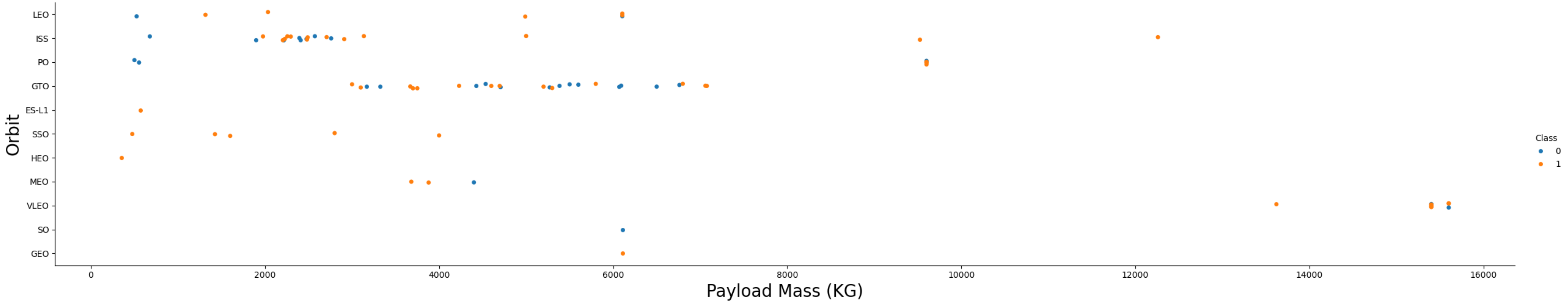
Flight Number vs. Orbit Type

- This scatterplot shows the flight number by the orbit type
- It includes the class in coloring, blue to indicate a failure to land and orange to indicate a successful landing
- This graph shows that certain flights are more successful and that some orbit types have very few flights, such as ES-L1 and SO both having a single flight



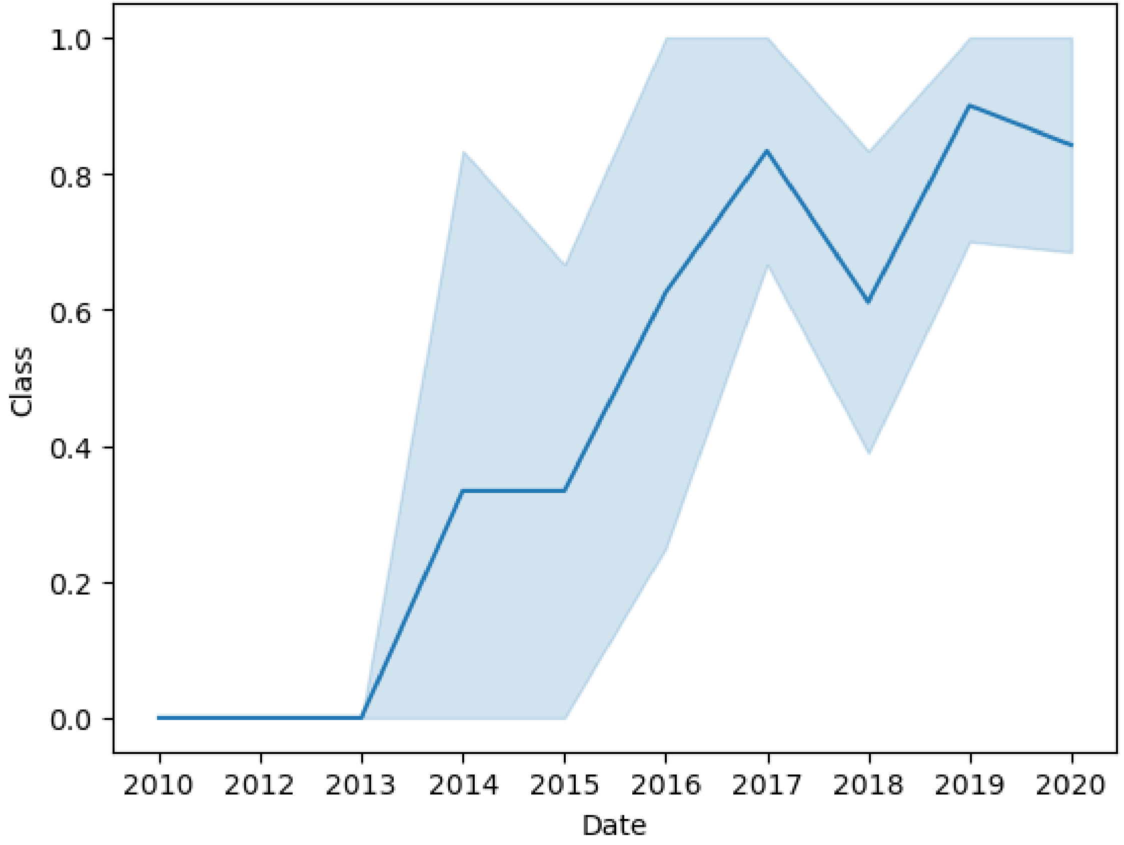
Payload vs. Orbit Type

- This scatterplot shows the payload mass in kilograms by the orbit type
- It includes the class in coloring, blue to indicate a failure to land and orange to indicate a successful landing
- This graph shows that certain orbit types have higher payload masses like VLEO, PO and ISS while others have very low payload masses like HEO and ES-L1



Launch Success Yearly Trend

- This line chart shows the success rate of launches over the years
- It indicates that at the start successful landings did not happen and didn't occur until 2014
- After that, successful landings have had an upward trend reaching a peak in 2019



All Launch Site Names

- The unique names of the launch sites
- The launch sites are displayed on the right and include 4 unique launch sites
- These include CCAFS LC-40, VAFG SLC-4E, KSC LC-39A, and CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

- This is 5 records that the launch site start with "CCA…"
- This query shows off some introductory information like what is included with each data point and what data from "CCA…" launch sites look like

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

- The total payload mass in kilograms carried by the boosters

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass in kilograms carried by the F9 v1.1 boosters

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first date which a ground landing was successful
- Despite launch data going all the way back to 2010, the first successful ground landing happened in 2015

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which had a successful landing with drone ship and had a payload between 4000kg and 6000kg

- Despite the many boosters that exist, only four had a successful drone ship landing between the payload's of 4000kg and 6000kg

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

Total Number of Successful and Failure Mission Outcomes

- The total number of mission outcomes
- Despite the number of unsuccessful landings, the mission outcomes generally ending in a success

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

Boosters Carried Maximum Payload

- List of boosters that carried the maximum pay load mass

- From all the records, only F9 B5 B1000's series boosters carried the maximum pay load of 15600kg

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

2015 Launch Records

- The records where the landing outcome was a failure in drop ship in the year 2015
- The record shows that a failure in drone ship happened on the first month (January) and the fourth month (April), both at the CCAFS LC-40 launch site

SUBSTR(DATE, 6,2)	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

- The count of the landing outcomes between 2010 and 2017 in descending order
- This shows that successful landings were very likely to happen between 2010 and 2017 if an attempt was made

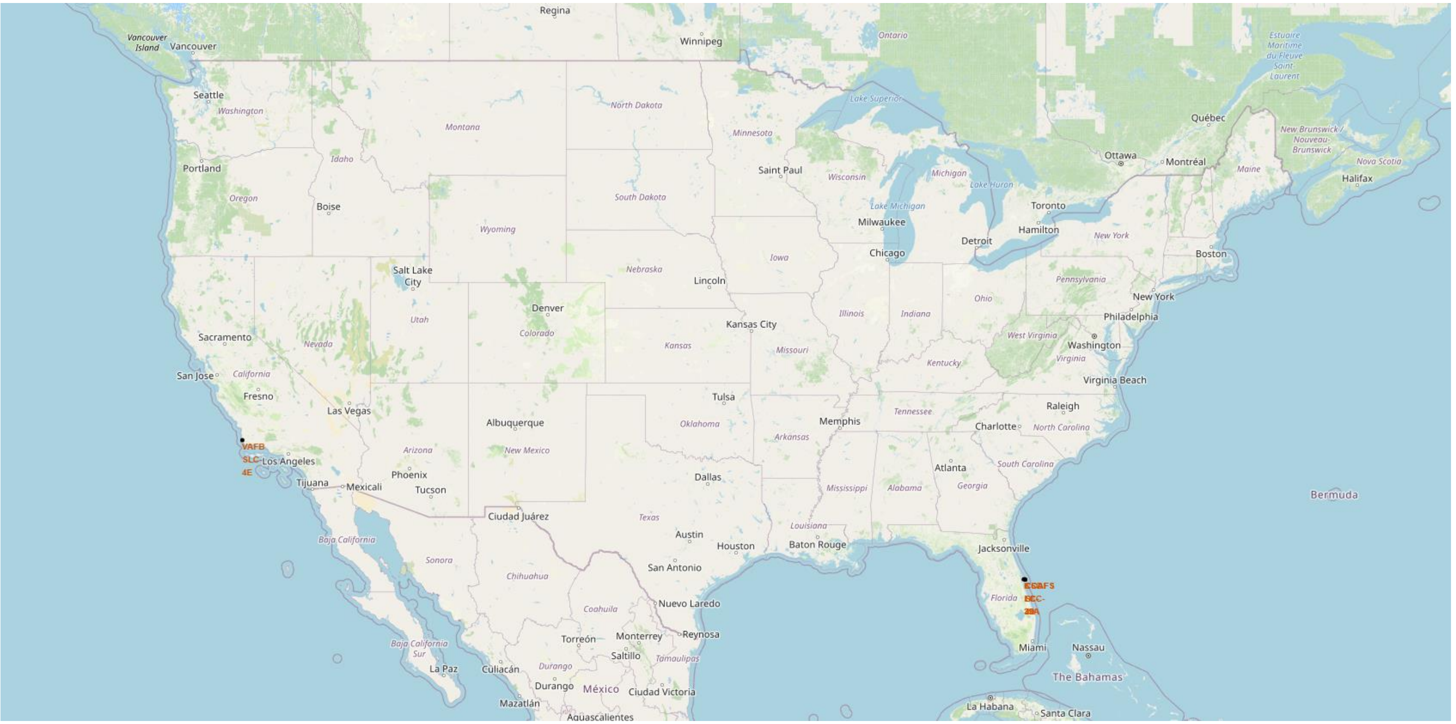
Landing_Outcome	count(Landing_Outcome)
Success (drone ship)	12
No attempt	12
Success (ground pad)	8
Failure (drone ship)	5
Controlled (ocean)	4
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

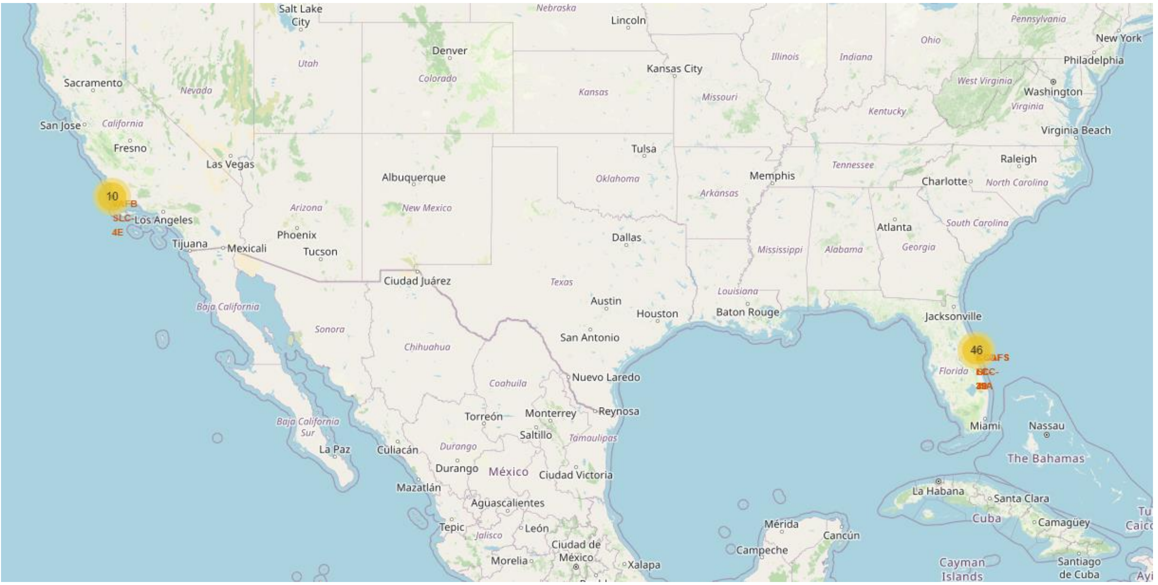
Folium Map of the Launch Sites

- This map shows the all the launch sites found in the data
- We can see that all the launch sites are in the United States
- We can also see that all the launch sites are on the coast and near the equator



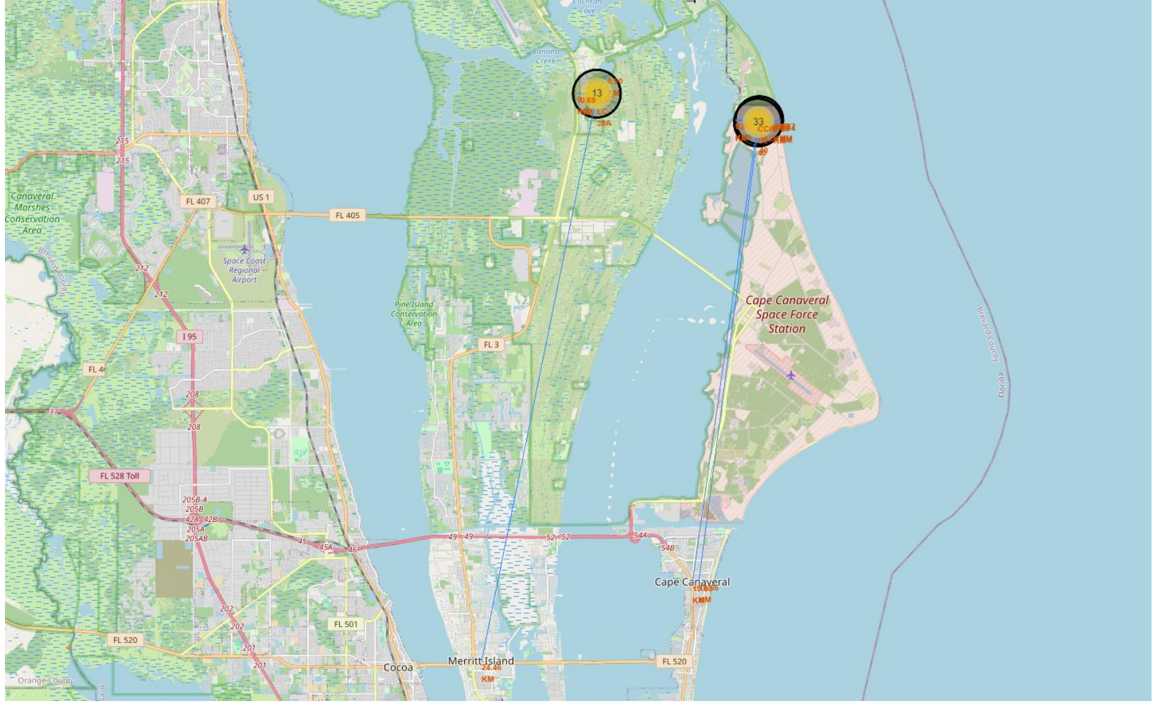
Folium map of the Launch Sites Successes

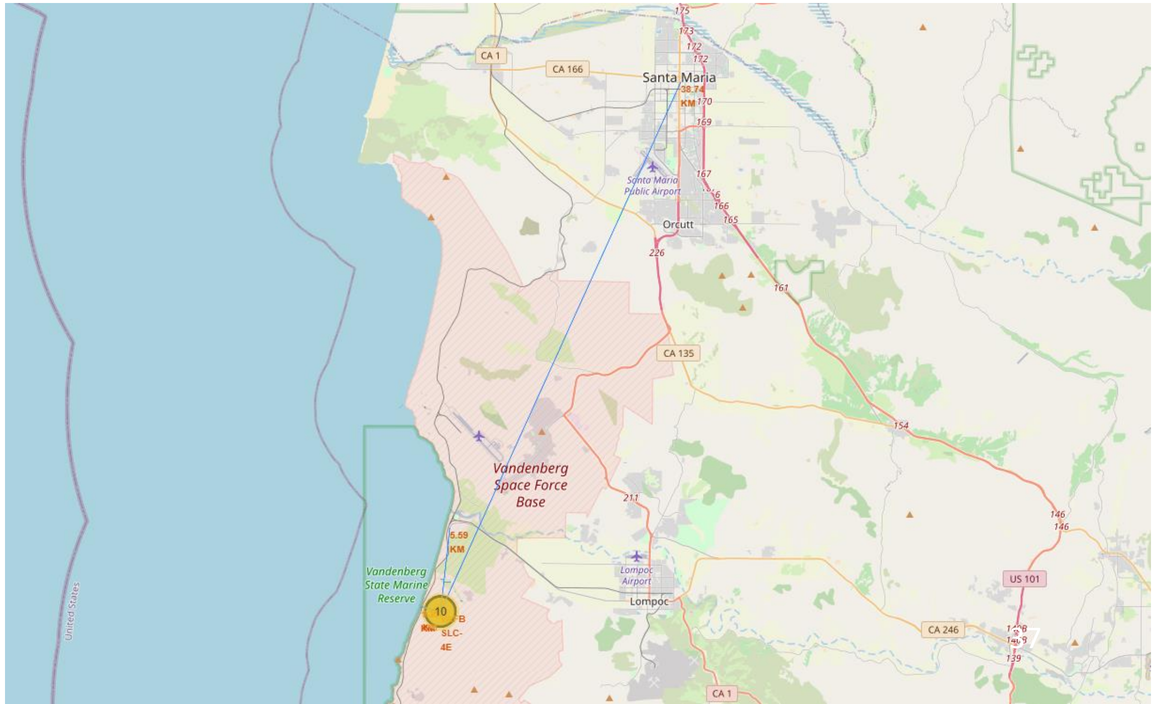
- These screenshots show the launch sites and their successful landings through green and red markers when clicked
- We can see from the screen shots that the west coast launch site does not have as many launches as it’s eastern coast counter parts
- We can also see that the launch site KSC LC-39A had 10 successful landings and 3 failed ones



Folium Map With Notable Nearby Features

- This folium map highlights the key points of interest near all the launch sites including highways, railways, cities, and coastlines
- From this map we can see the launch sites are very close to coastlines, highways, and railways
- We can also see that launch sites tend to be far away from any nearby cities



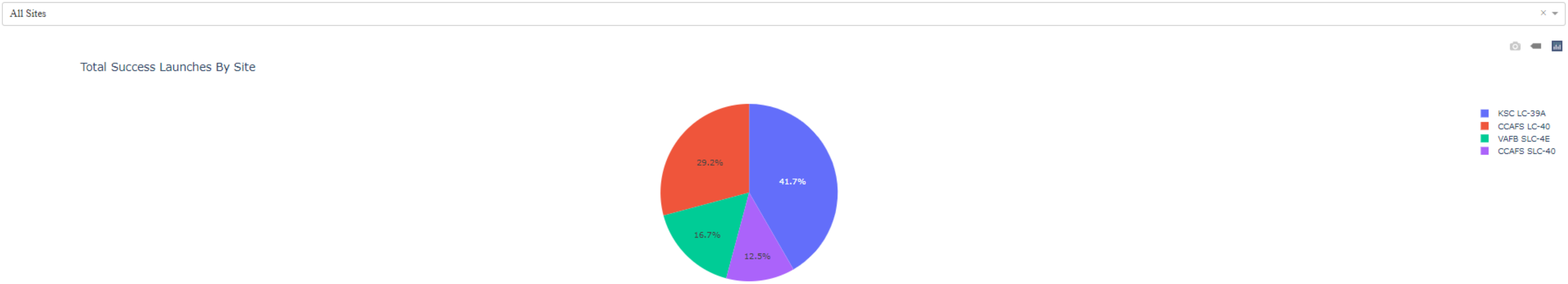


Section 4

Build a Dashboard with Plotly Dash

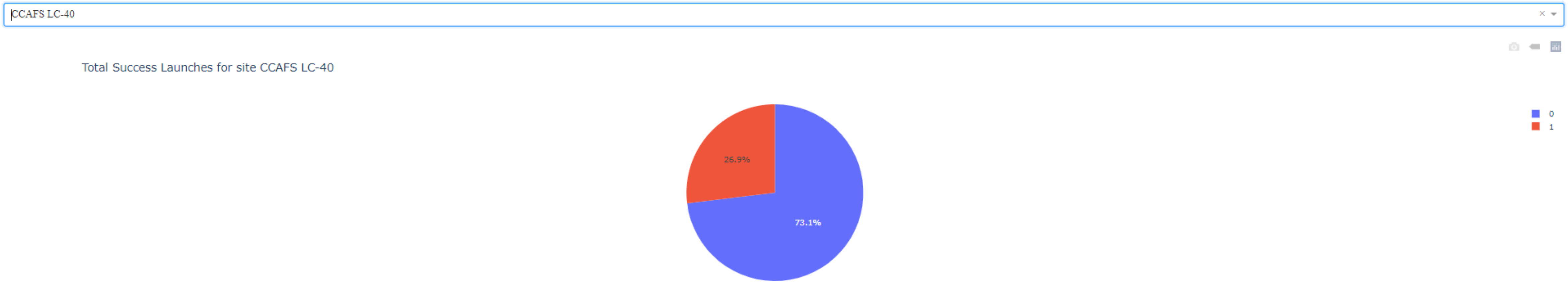
Pie chart of all the launch success counts (FROM LAB, NOT MINE)

- This pie chart shows the successful launches of all the launch sites as a percentage
- It indicates that KSC LC-39A has the highest successful launches but that could be due to having more launches, more information is needed



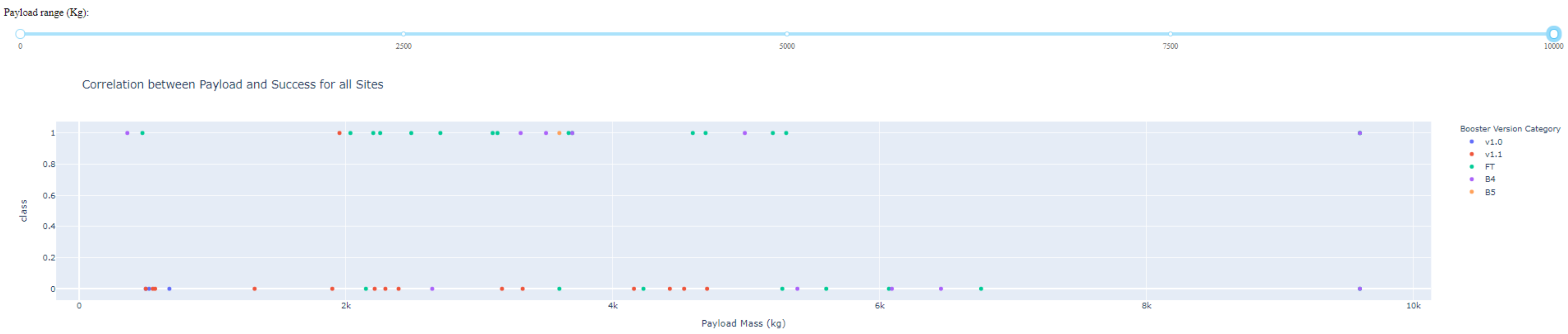
Pie Chart of Launch Site with Highest Success Rate (FROM LAB, NOT MINE)

- This pie chart shows the total successful launches as a percentage for the launch site CCAFS LC-40
- The red indicates a successful launch while blue indicates an unsuccessful launch
- This shows that this site has much more unsuccessful launches than it did successful ones



Payload vs Success Scatterplot (FROM LAB, NOT MINE)

- This scatterplot shows the correlation between payload mass in kilograms vs the success rate for all launch sites at the minimum and maximum range for payload mass
- It seems to indicate that unsuccessful launches are much more prevalent and more successful launches happen between very low payloads to under 6000kg payloads

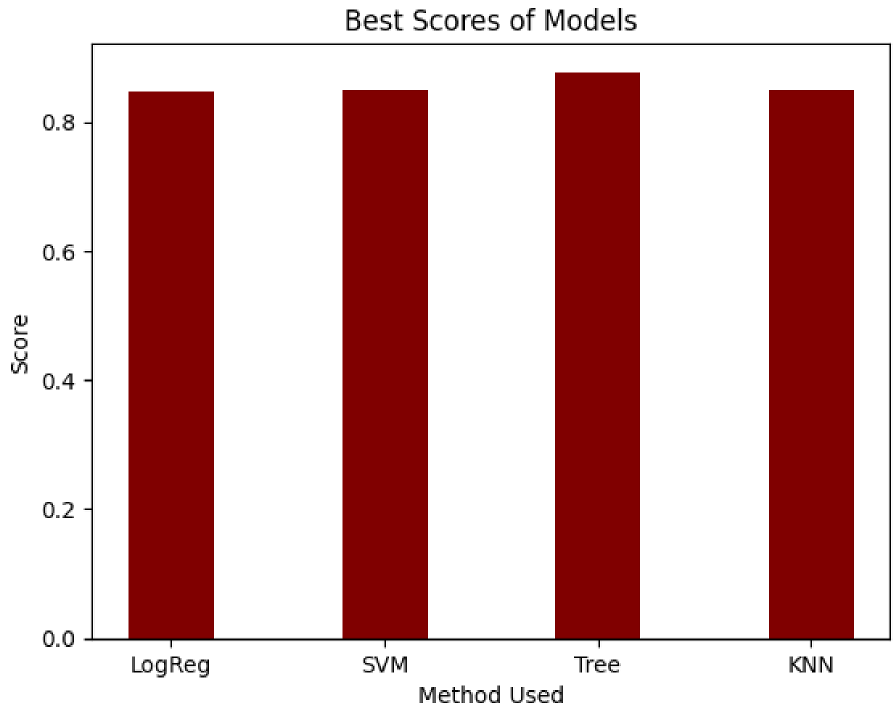


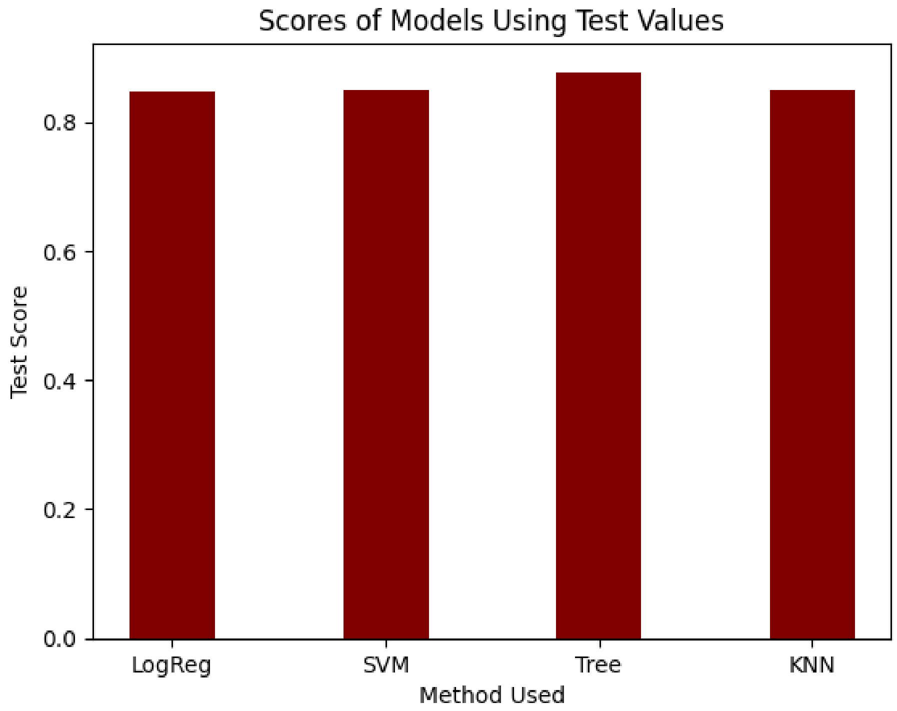
Section 5

Predictive Analysis (Classification)

Classification Accuracy

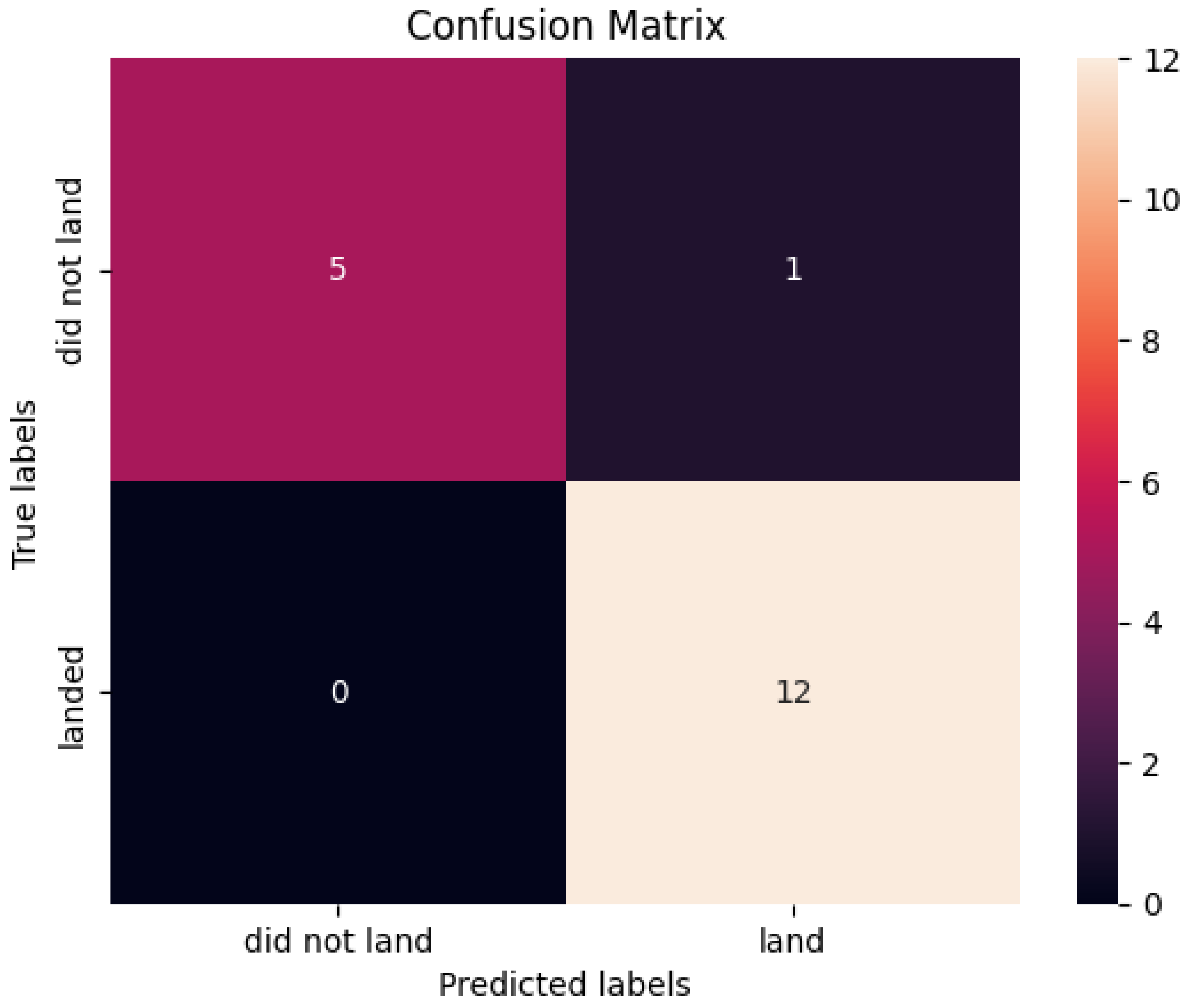
- These two bar charts show the best score and the test scores of the 4 models built
- From these charts, we can see that the Tree method was the most effective





Confusion Matrix

- This is the confusion matrix from the decision tree method
- The coloring indicates the number of times a landing data point fell in the labeled range
- The bottom labels indicates the predicated values, and the y-labels indicate the true values
- From this matrix we can see that method accurately predicts failed landings
- We can see this matrix is less accurate for predicting successful landings as it had one false positive, predicting it would land when it did not

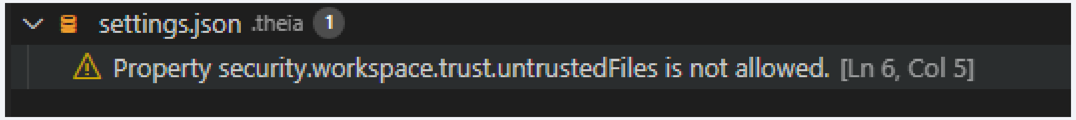


Conclusions

- Using the created decision tree method for predicting, we can accurately predict if a landing will be successful or not
- This will allow us to save costs on testing through trail and error and expedite the process for creating rockets that will land successfully
- Our company will be able to accomplish this by focusing development on rockets that will landing successful more frequently
- This cost saving and higher likelihood of successful landing will make space travel more affordable
- Our low-cost space travel will allow us to compete with other space companies like Space X

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
- The error code I was getting that prevented me from completing the dashboard section



Thank you!