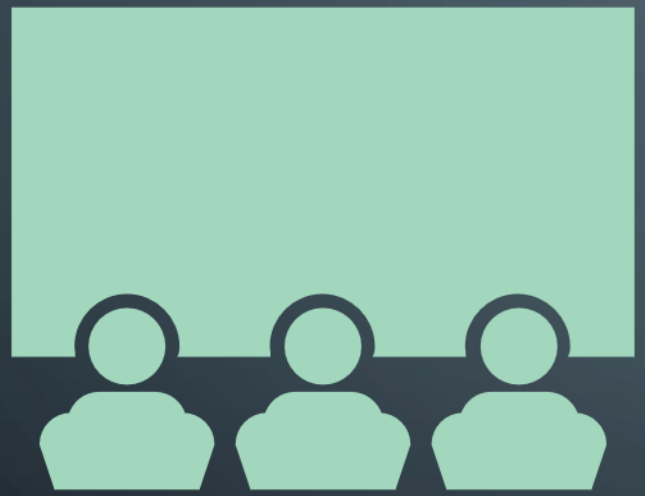
DATA SCIENCE CAPSTONE PROJECT

JESUS GARCIA

08/2021

OUTLINE



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

EXECUTIVE SUMMARY

- Launching a rocket is expensive, so we built multiple classifiers to determine whether the first stage of a Falcon-9 rocket will land, thus determining the cost of the launch.

- Overall, we obtained the following for each model:

1. Logistic Regression Accuracy: 0.8333
2. SVM Accuracy: 0.8333
3. Decision Tree Accuracy: 0.7222
4. KNN Accuracy: 0.8333

INTRODUCTION

- By predicting whether a rocket will land successfully or not we can use that to predict what the cost a launch will be and possibly find features that might affect landing results.
- In this project, we focus specifically on the classification of a launch and whether it will recover the first stage to diminish costs of construction for the first stage Falcon-9 rocket.

METHODOLOGY

- Data collection methodology:
  - Data was collected through previous launches and recorded.
- Perform data wrangling
  - Some of the preprocessing steps taken were one-hot-encoding on attribute as well as some normalization of the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - For evaluation, we utilized built-in scoring from GridSearchCV.

METHODOLOGY

DATA COLLECTION & WRANGLING SPACEX API

1. To retrieve the data necessary, we first define some functions to retrieve essential data such as payload, launch site and booster version data.
2. We then connect to the API by a “get” request and obtain a static json object.
3. Finally, we convert the static json object into a **pandas** DataFrame.

Github URL:
https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/Data%20Collection.ipynb

DATA WRANGLING

- After the data is in a data frame, we replace some of the features missing values by replacing it with the mean of the column.

- Additionally, we add a binary feature named “landing_class” to see whether the recorded rocket landed successfully or not.

- After these pre-processing events are completed, the data is clean and ready to be used for modeling.

- GitHub URL: https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/Data%20Wrangling.ipynb

EDA WITH DATA VISUALIZATION

- This portion of the visualization is purposed to show the relationship among variable utilizing graphs such as scatter plots and bar graphs. We explore the relationship between variables that can be useful.

- Utilizing visualizations can more-easily show the types of relationships among variables and whether they have some sort of correlation or not.

- GitHub URL: https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/EDA%20w_Data%20Viz.ipynb

EDA WITH SQL

- In this section we perform SQL queries to gain some insight on the SpaceX data.
- Since there is not much visualization in this section, please refer to the GitHub below for query results.
- GitHub URL: https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/EDA%20w_SQL.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- With Folium we were able to extract visual map data to see how launch sites were near to open bodies of water and transportation sources like railroads, highways etc.

- In the Folium map, I added markers for launches that occurred for each launch site. The red mark is a failed launch, and a green mark is a successful launch.

- GitHub URL:https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/Folium.ipynb

BUILD A DASHBOARD WITH PLOTLY DASH

- With Dash, I was able to quickly implement and deploy an application built on Python code that visualized data based on launch sites.

- We visualized the data based on scatter-plots and pie charts. We were able to show correlation among variables such as payload mass and whether a launch was successful or not.

- GitHub URL:https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/Plotly.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

- My models were built utilizing **sklearn** classification algorithms. To maintain stable results I utilized a random seed (2) and switched some learning rates occasionally.

- First, we web scraped the SpaceX API for json objects which we then moved the content to a **pandas** data frame and replaced some attributes missing values with the series mean. We then utilized this data to train the model.

- GitHub URL: https://github.com/Jgarcia2048/IBM-Data-Science/blob/main/Data%20Collection.ipynb

RESULTS

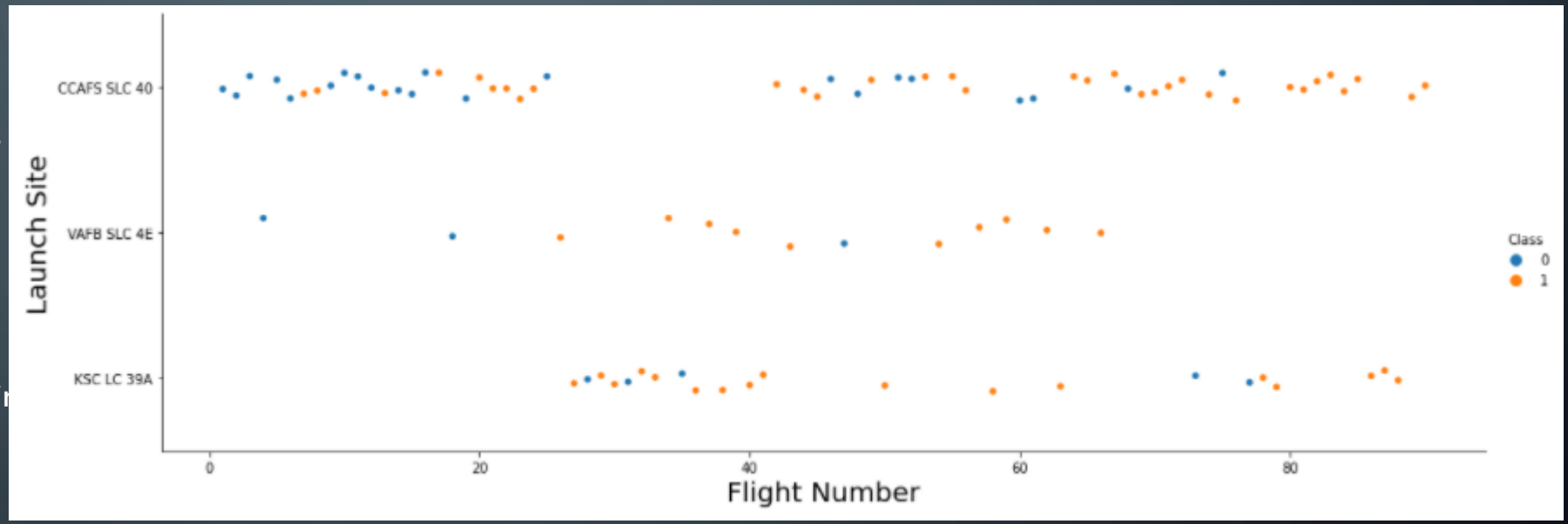
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE

In this graph we can see a increases in successful launches as the flight number increases.

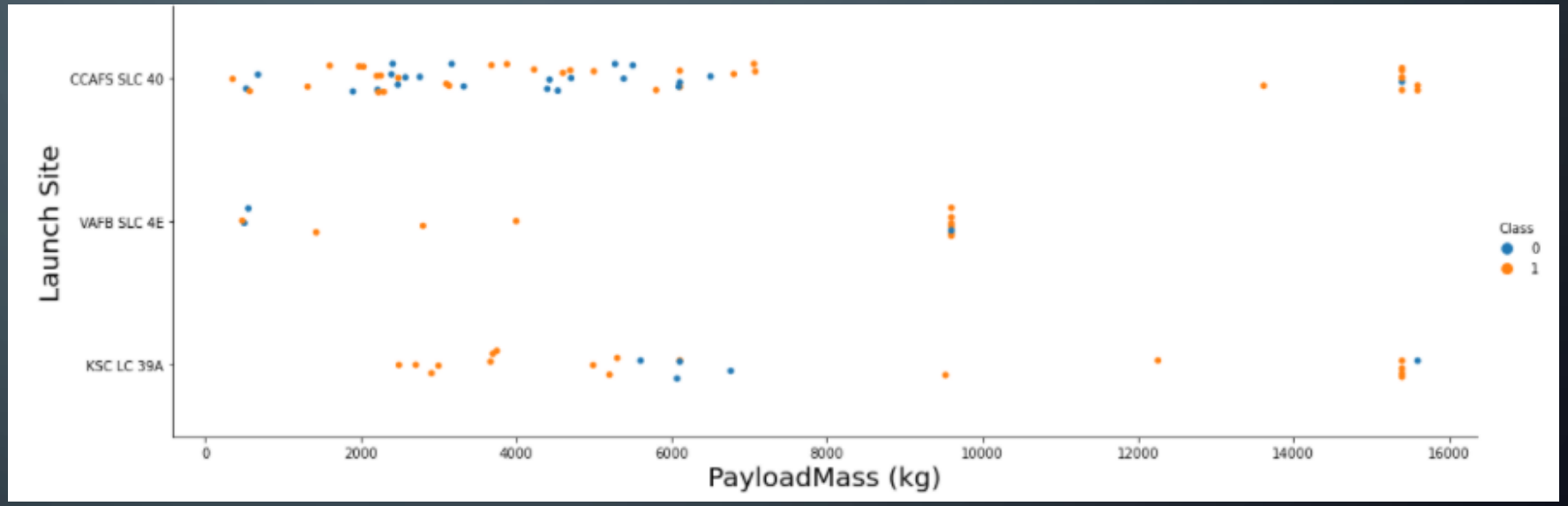
This makes sense since SpaceX attempted to keep improving their tech over time.



PAYLOAD VS. LAUNCH SITE

In this scatterplot we can see that launch sites “SLC 40” and “LC 39A” tend to launch rockets from very minimal weight to very high capacities.

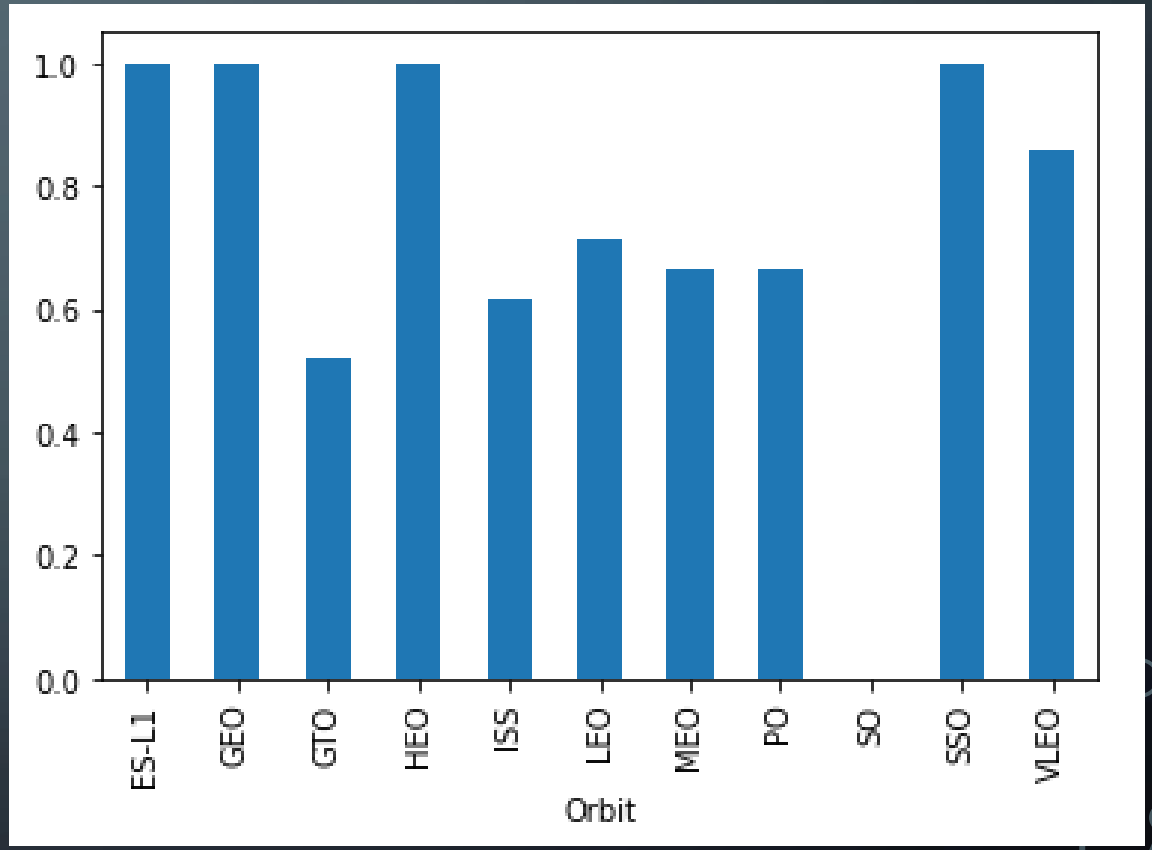
There are less failures in launches that have significantly heavy payloads.



SUCCESS RATE VS. ORBIT TYPE

In this bar graph the Y-axis represent the success rate of each launch to each orbit type.

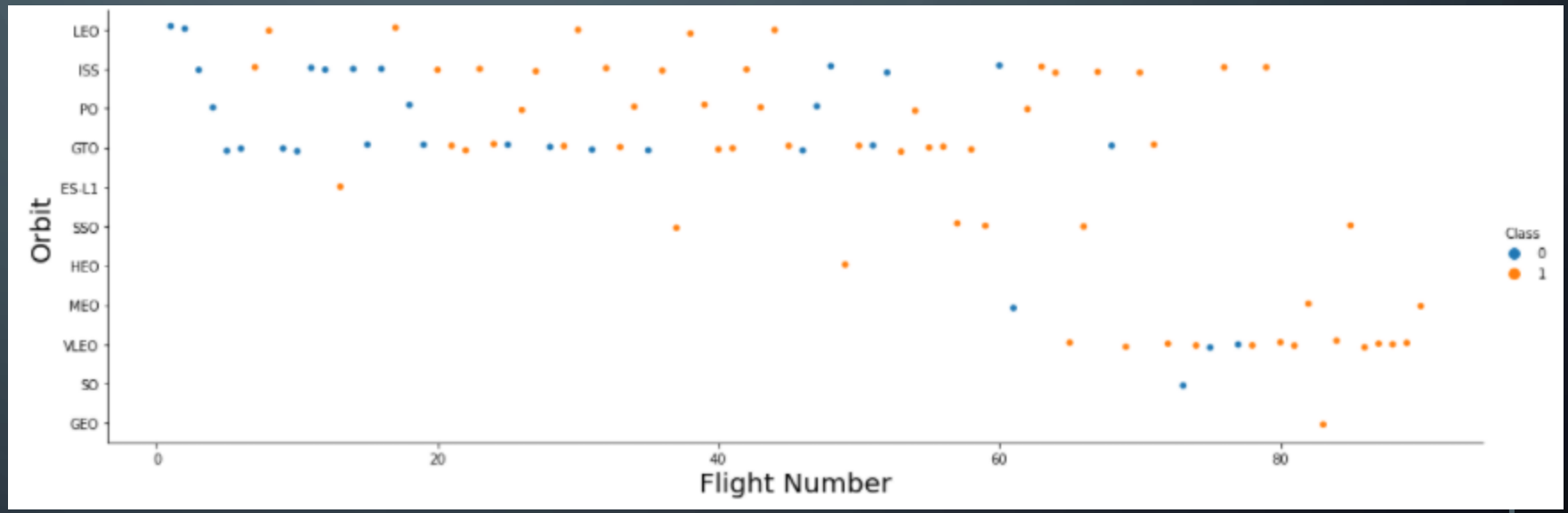
We can see a clear outlier where SO (Sun-Synchronous Orbit) does not have any successful launches.



FLIGHT NUMBER VS. ORBIT TYPE

Here we can see a progression launching to different Orbit types as flight numbers progress.

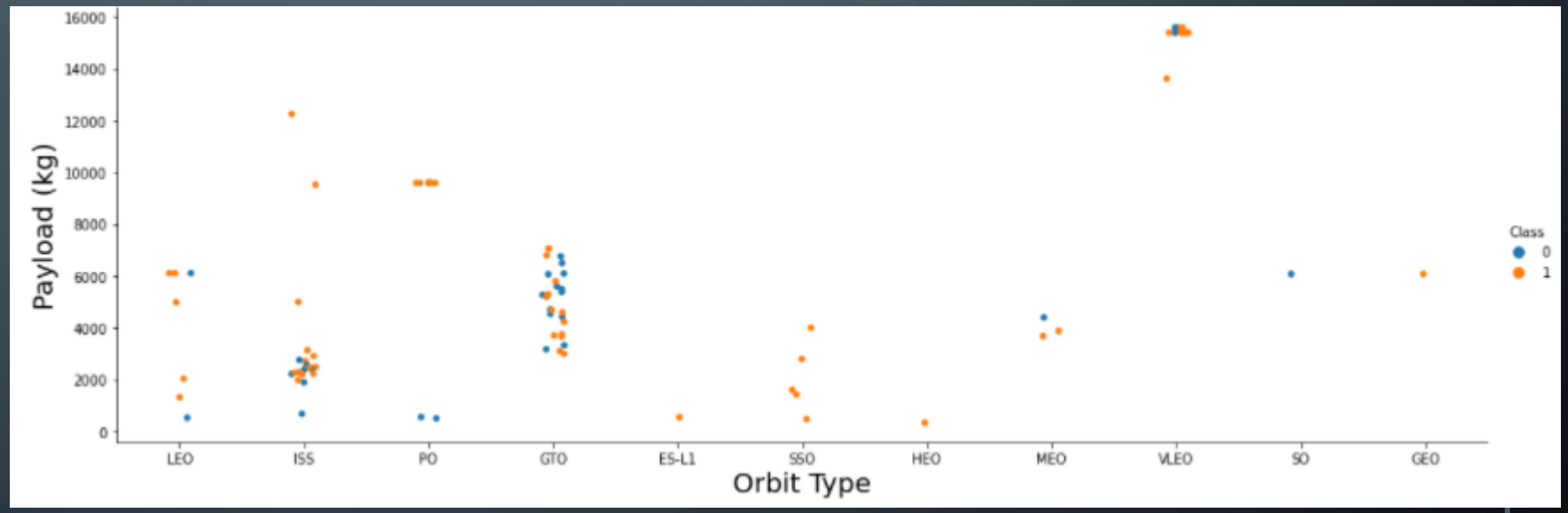
This can be due to advancements in tech that might make it more feasible to travel to different orbits.



PAYLOAD VS. ORBIT TYPE

In this bar graph the Y-axis represent the success rate of each launch to each orbit type.

We can see a clear outlier where SO (Sun-Synchronous Orbit) does not have any successful launches.



EDA WITH SQL

ALL LAUNCH SITE NAMES

- Based on the SQL query, we attempted to find the distinct launch sites which are shown to the right.

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

LAUNCH SITE NAMES BEGIN WITH `CCA`

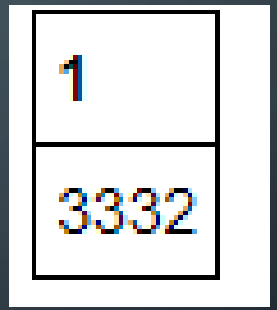
- Here we filtered the database for launch sites that begin with “CCA”

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

TOTAL PAYLOAD MASS

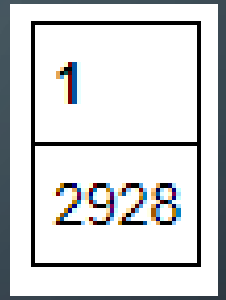
- In this SQL query we calculated the average payload (kg) by the client NASA.

- The average was 3,332kg.



AVERAGE PAYLOAD MASS BY F9 V1.1

- Calculated the average payload mass carried by booster version F9 v1.1 with a SQL query.



- The average was 2,928kg.

FIRST SUCCESSFUL GROUND LANDING DATE

- In this SQL query, we found the first successful landing date.

1
2010-06-04

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- The following boosters were able to successfully land on a drone ship while having a payload between 4,000kg and 6,000 kg.

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

- Successful Outcomes:

SUCCESSFUL OUTCOMES
99

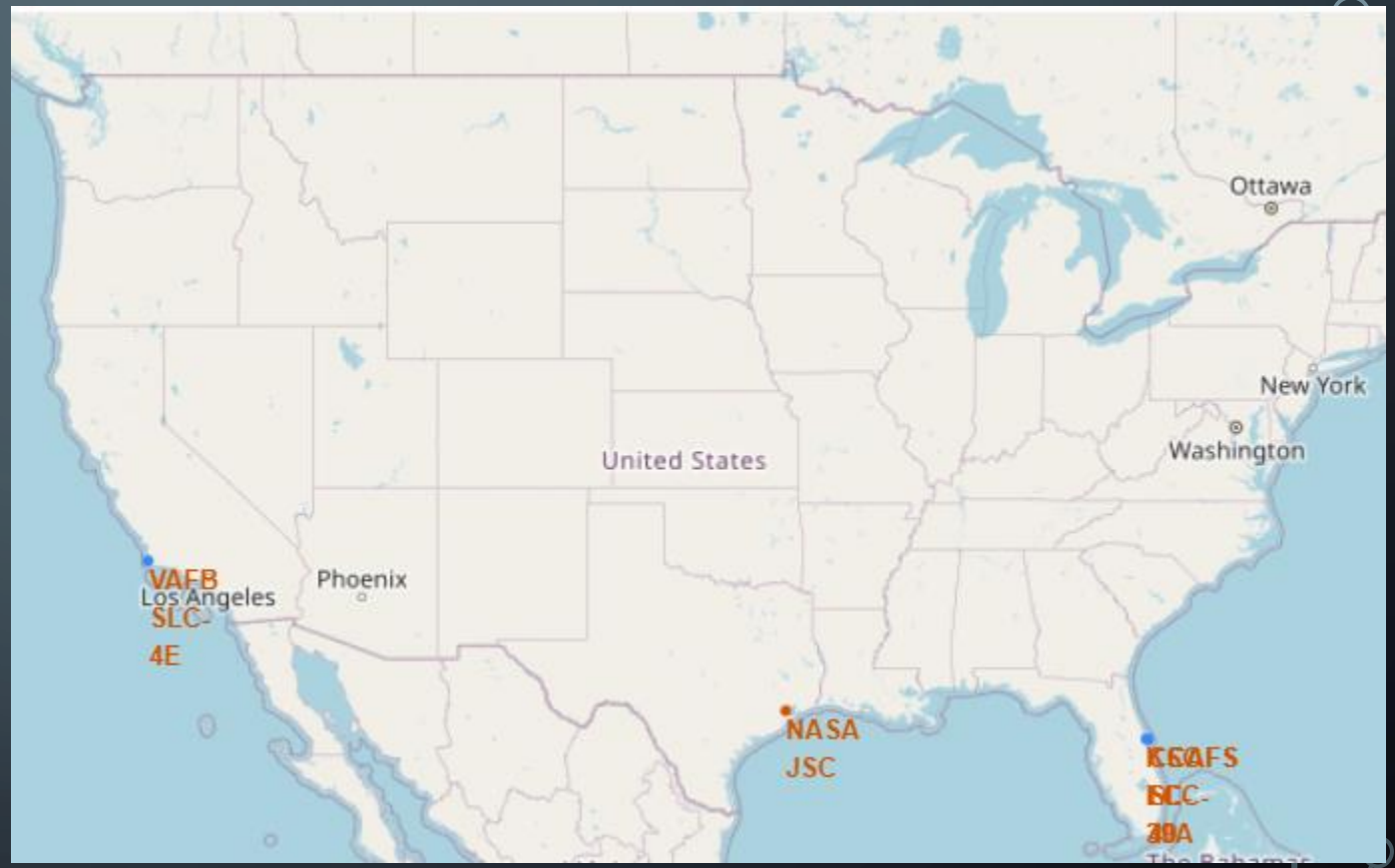
- Failed Outcomes:

FAILED OUTCOMES
1

INTERACTIVE MAP WITH FOLIUM

SPACEX LAUNCH SITES

- Here we can see all the launch sites mostly being in coastal areas near the equator and transportation sources like railroads and highways.



SUCCESS MARKERS IN LAUNCH SITES

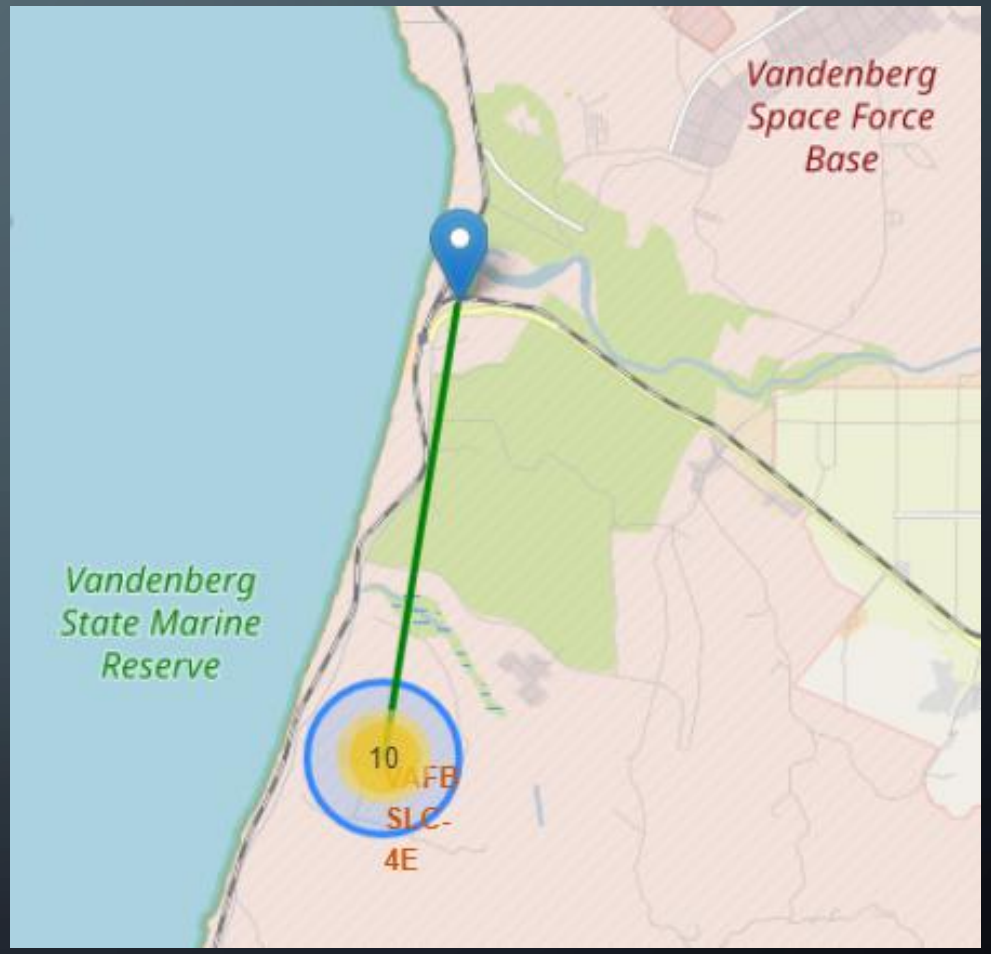
- Top-Left Site: VAF SLC-4E
- Top-Right Site: KSC LC39A
- Bottom-Left Site: CCAFS LC-40
- Bottom-Right Site: CCAFS SLC-40

- We can see that site CCAFS LC-40 has a majority of failed launches, on the other hand KSC LC39A seems to have very successful launches.

STRATEGIC POSITIONING FOR ROCKET LAUNCH SITES

- In this Folium image, we can see that sites tend to be very near coasts and transportation sources like roads and train tracks.

- In this particular site, the location is adjacent to Vandenberg Space Force Base.

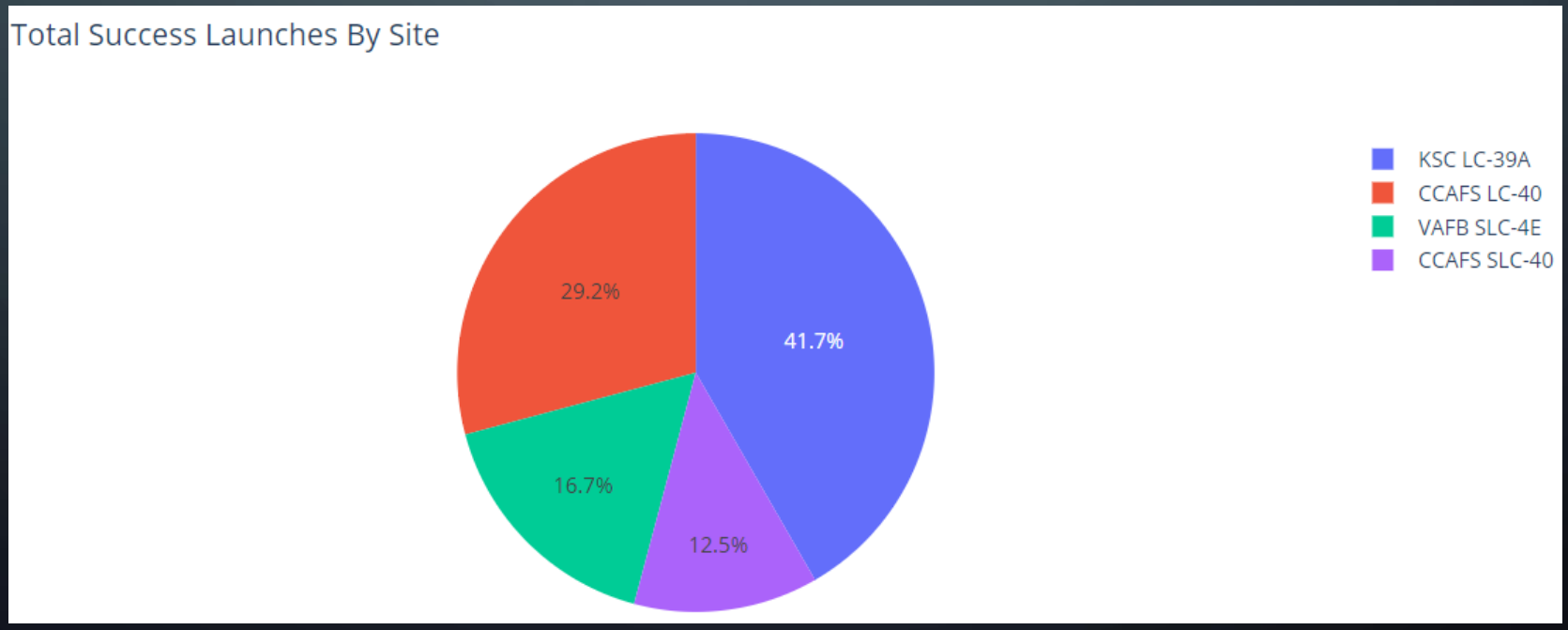


BUILD A DASHBOARD WITH PLOTLY DASH

SUCCESS PER LAUNCH SITE

- This pie graph depicts the share of successful launches per each launch site.

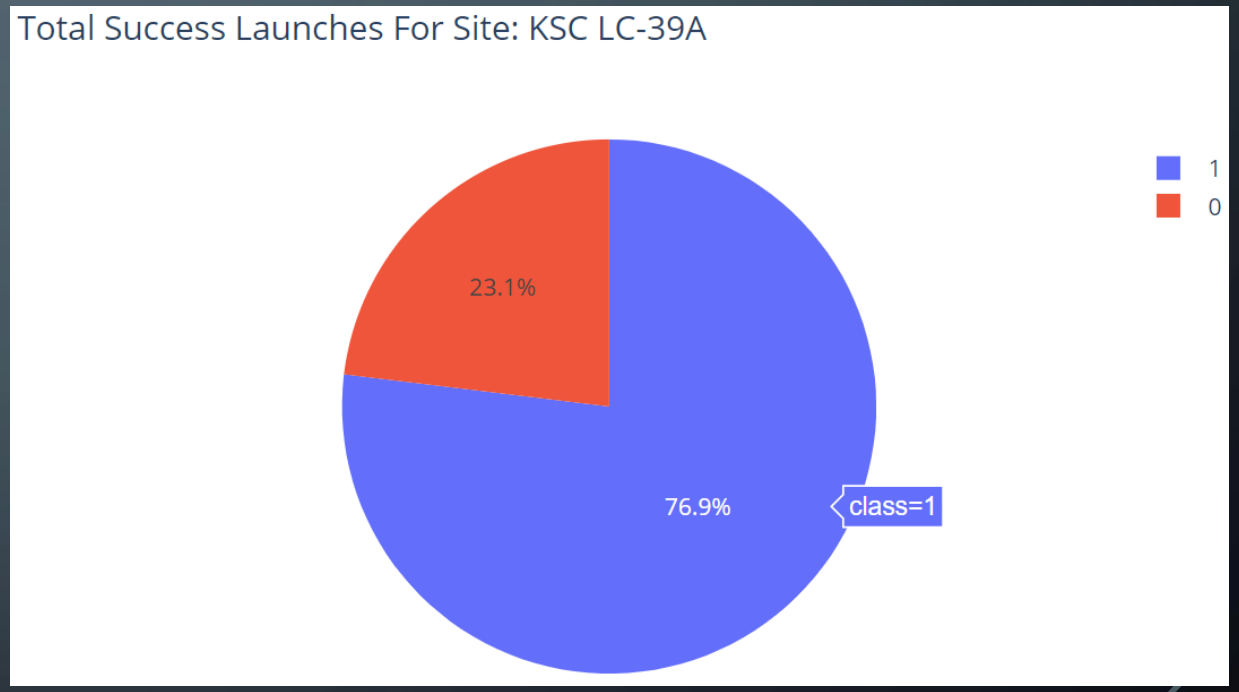
- Here we can again see that site KSC LC-39A has a very high ratio of successful launches compared to every site.



MOST SUCCESSFUL LAUNCH SITE: KSC LC-39A

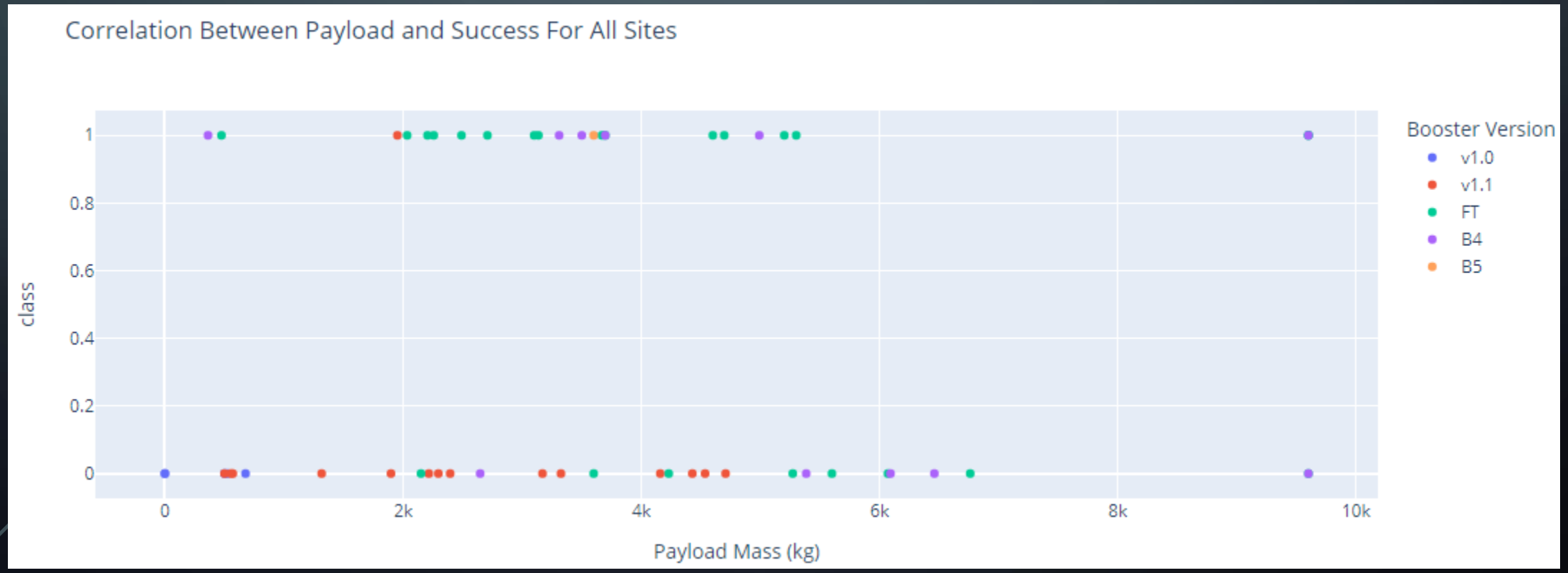
- Here we can again see that site KSC LC-39A has a very high ratio of successful launches compared to every site. (refer slide #31)

- Slightly more than 3 out of 4 launches is successful at this site.



PAYLOAD VS SUCCESSFUL LAUNCH CORRELATION

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

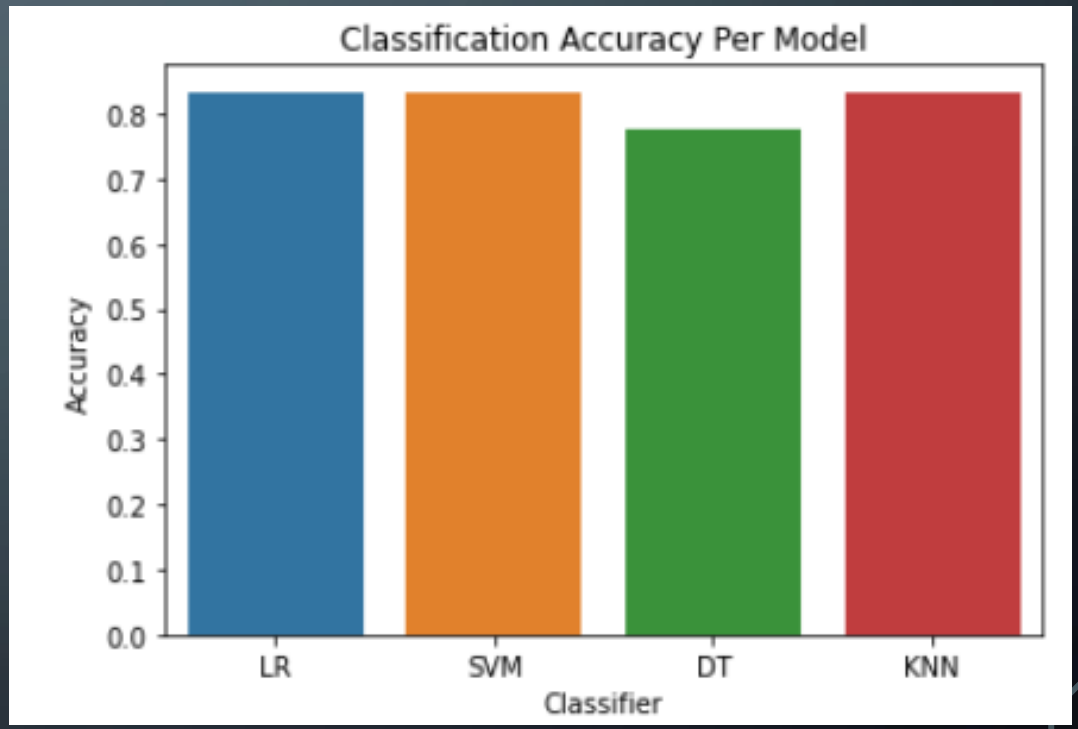


PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

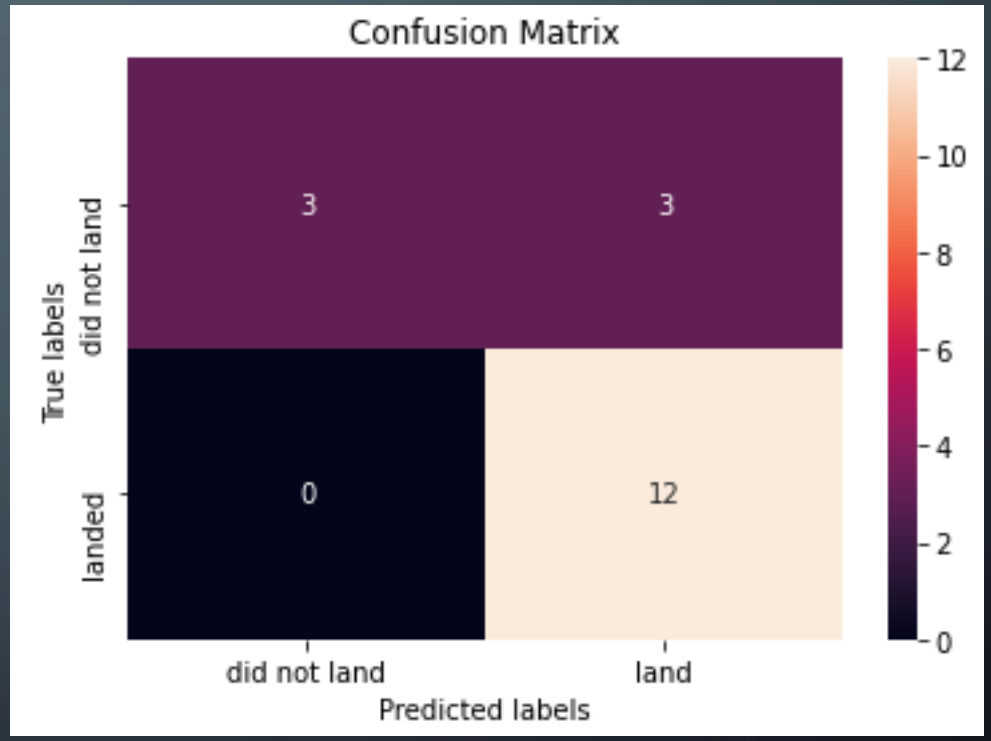
Here we can see all of the classification accuracies implemented through sklearn.

Most of the classifiers actually show identical classification results (83.33%). Decision tree fell out of this category but not by far (77.78%)



CONFUSION MATRIX

Since KNN, SVM and Logistic Regression scored identical, the confusion matrices for the three are exactly the same.



CONCLUSION

1. It is essential to test what sorts of missing value replacement can potential increase your model accuracy.
2. Visualization of data before model training can add a lot of intuition into how the data was collected where it is not so easily seen through a spreadsheet or data object.
3. KNN, SVM & Logistic Regression perform very similar to one another.

APPENDIX

- EDA w/SQL: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/2a297ce3-d3e5-4618-9071-770c6ec219b0/view?access_token=d1b49c703e99bab45896bc1376e9ccfd5d72f9e1ad4cacbe9e199751104b88c1
- EDA w/Data Viz: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/a32b6372-727a-47ad-bb8b-bb1d18c57abb/view?access_token=a21c5898ba312a1c254917f710a740bb0ee33c1fa3b391dbe40aa1477602c790
- Folium Dash: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/01eaf5e8-a2a2-4fd1-b42d-77270b61d794/view?access_token=fd73da0d4875fce077727a9e8a710afcb2f802b36734887ade4ab5817d9ec259
- Data Collection: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/c50c1a68-1d69-4906-b107-b3d5bf8faeb6/view?access_token=6f22fdcdbf99a4d1a22d6c5b31da28b14c1823e1137ae4245f150179e097b1b1
- Data Wrangling: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/5e99a9c6-5982-464c-9346-eef9d0358373/view?access_token=41f708e54a99ff17c5ef03854ad382c0e2582a491d2cafbc431b5f5e8c78a5a4
- Classification: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/288d4c30-4733-4876-bb45-d5cc966b104a/view?access_token=110eb13e31aba8668496733824325 42f47845242f92a218713413011 6be8266d
- GitHub: https://github.com/Jgarcia2048/IBM-Data-Science