Winning Space Race with Data Science

Héctor César Ornelas León
12/01/22

Outline

- Executive Summary

- Introduction

- Methodology

- Results

- Conclusion

- Appendix

Executive Summary

- Summary of methodologies

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Data Collection

- Data wrangling

- EDA with data visualization

- EDA with SQL

- Building a dashboard with Plotly Dash

- Predictive analysis

Summary of all results

- EDA results

- Interactive analytics

- Predictive analysis

Introduction

- Project background and context

On its website, SpaceX promotes Falcon 9 rocket launches for 62 million dollars; other suppliers charge upwards of 165 million dollars for each launch. A large portion of the savings is due to SpaceX's ability to reuse the first stage.

- Problems you want to find answers

The project has the goal of predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully. if we can figure out if the first stage will land, we can figure out how much a launch will cost.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  • SpaceX REST API

  • Web Scraping SpaceX data from Wikipedia

- Perform data wrangling

  • One Hot Encoding of data fields for Machine Learning and data cleaning of null values and irrelevant 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

  • Logistic Regression, SVM, Decision Trees and KNN models have been built and evaluated to determine the best classifier.
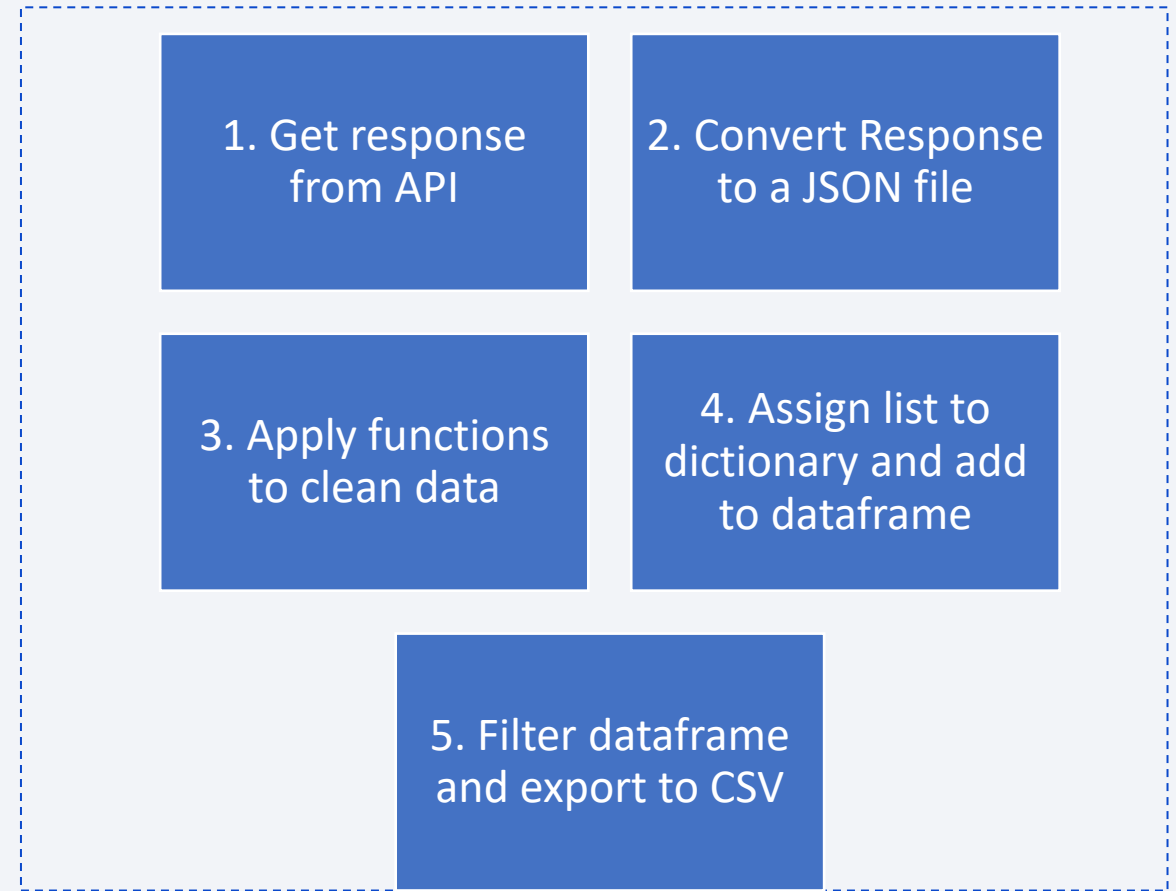
Data Collection

- SpaceX launch data that is generated from the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

- The SpaceX REST API endpoints or URL starts with ap.spacexdata.com/v4/.

- Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.

Data Collection – SpaceX API

Data collection with SpaceX REST calls

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

1. Get response from API	2. Convert Response to a JSON file
3. Apply functions to clean data	4. Assign list to dictionary and add to dataframe

5. Filter dataframe and export to CSV

Data Collection - Scraping

Web Scraping data from Wikipedia

1. Get response from HTML	2. Create BeautifulSoup Object
3. Find Tables	4. Get columns names and create a dictionary
5. Append data to keys	6. Convert dictionary to dataframe and to CSV.

Data Wrangling

1. Check null values

2. Calculate number of launches on each site

3. Calculate the number and occurrence of each orbit

4. Calculate the number and occurrence of mission

5. Create a landing outcome label

6. Handle all null values

EDA with Data Visualization

- The charts used in the EDA include the following:

• Line chart: This chart was used to show the average launch success trend by plotting Year on the x-axis and Average success rate on the y-axis.

• Bar chart: I used a bar chart to visually check if there are any relationship between success rate and orbit type.

• Scatter plot: Multiple scatter plots were plotted in this project showing the linear relationship between the following variables;

 1. Flight Number vs Payload Mass

 2. Payload Mass vs Launch sites

3. Flight Number vs Orbit type

4. Payload vs Orbit type

EDA with SQL

- SQL queries performed include:

• SELECT the names of the unique launch sites in the space mission.

• SELECT 5 records where launch sites begin with the string 'KSC'.

• SELECT the total payload mass carried by boosters launched by NASA (CRS).

• SELECT average payload mass carried by booster version F9 v1.1.

• List the dates where successful landing outcome in drone ship was achieved.

• List the names of the boosters which have success on the ground pad and have payload mass greater than 4000kg but less than 6000kg.

• List the total number of successful and failed mission outcomes.

• List the names of the booster versions which have carried the maximum payload mass.

• SELECT the month names, successful landing outcome on the ground pad, booster versions, launch site for the months in the year 2015

• RANK the count pf successful lamding outcomes between 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

Objects created and added to a folium map include the following:

• Circles: Folium.Circle was used to add a text label and a highlighted circle area at a particular coordinate.

• Markers: By pinning the locations on a map, markers make coordinates in plain numbers easier to understand.

• Markclusters: Mark clusters can be a highly useful tool for streamlining a map with numerous markers that share the same coordinate.

• MousePosition: To obtain coordinates for a mouse over a point on the map, use MousePosition. As a result, you may simply locate the coordinates of any points of interest while exploring the map (such a as highway).

Build a Dashboard with Plotly Dash

The following plots and interactions were added to the dashboard;

• Site-dropdown: This interaction is used to select a desired launch site and renders a pie chart visualizing launch success counts.

• Pie chart: This plot visualizes launch success counts.

• Range Slider: We want to find if variable payload is correlated to mission outcome. This interaction allows you to easily select different payload range and see if we can identify some visual patterns.

• Scatter chart: With "Payload" on the x-axis and "Launch outcome" on the y-axis, we can visually observe how payload may be correlated with mission outcomes for selected site(s)

Predictive Analysis (Classification)

- Performed exploratory Data Analysis and determined Training Labels • Created a column for the class and created a NumPy array from this column by applying the method "to_numpy()" and assigned the array to a variable Y.

- Standardized the data in X then reassign it to the variable X using "transform = preprocessing.StandardScaler()".

- Split the data into training data and test data using the function "train_test_split" and set the parameter test size to 0.2 and random state to 2. The training data is divided into validation data, a second set was used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV.

- Created a logistic regression object then create a GridSearchCV object logreg_cv with cv = 10. Fitted the object to find the best parameters from the dictionary parameters. I output the GridSearchCV object for logistic regression. I displayed the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score_.

- Calculated the accuracy on the test data using the method score and created a confusion matrix. We can tell that logistic regression can distinguish between the various groups by looking at the confusion matrix. We can observe that false positives are the main issue.

- Performed the last two steps for "SVC()", "DecisionTreeClassifier()" and "KNeighborsClassifier()"

- Found the best Hyperparameter for SVM, Classification Trees and Logistic Regression. Using test data, I found the best-performing method.

- The SVM, KNN, Decision Trees, and Logistic Regression model achieved the highest accuracy at 83.3%.

Results

- The SVM, KNN, Decision Trees, and Logistic Regression models are the same in terms of prediction accuracy for this dataset.

- Low weighted payloads perform better than the heavier payloads.

- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.

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

- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Section 2

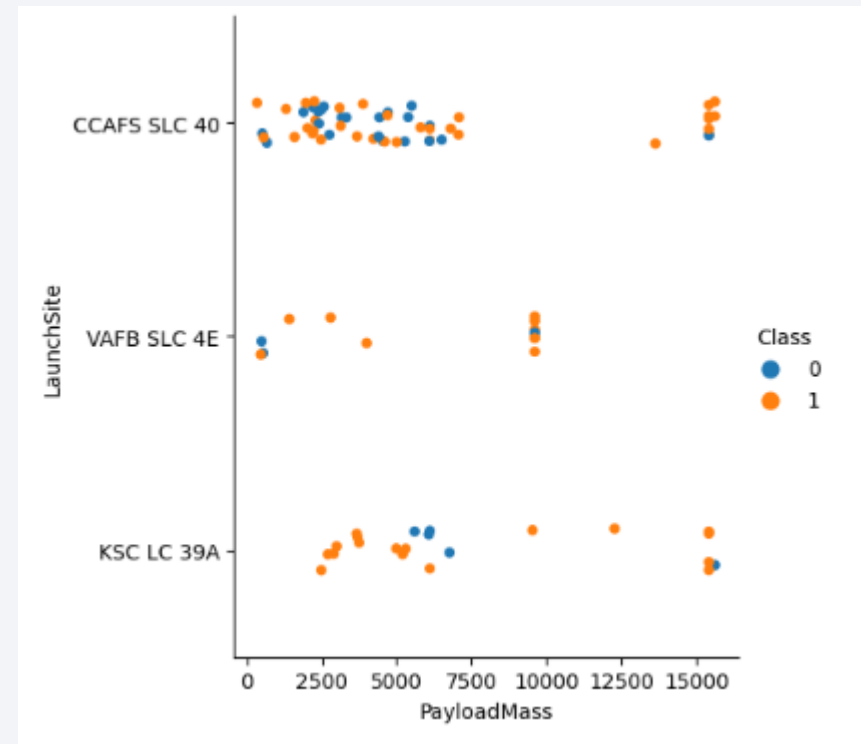# Insights drawn from EDA

Flight Number vs. Launch Site
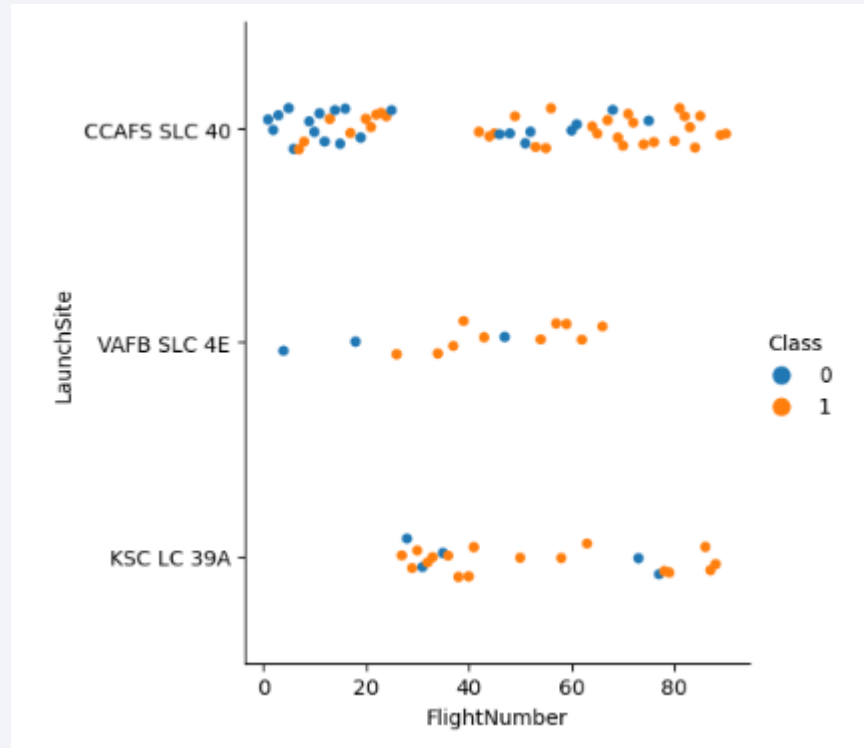
- We see that as the flight number increases, the first stage is more likely to land successfully.

Payload vs. Launch Site

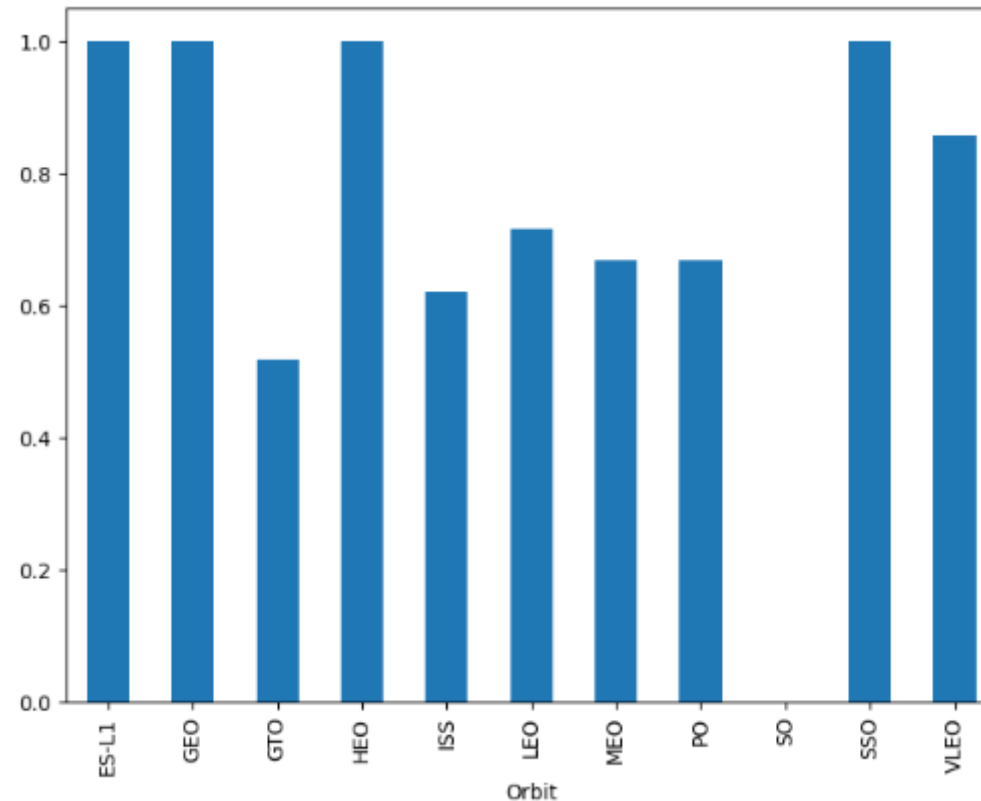For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).
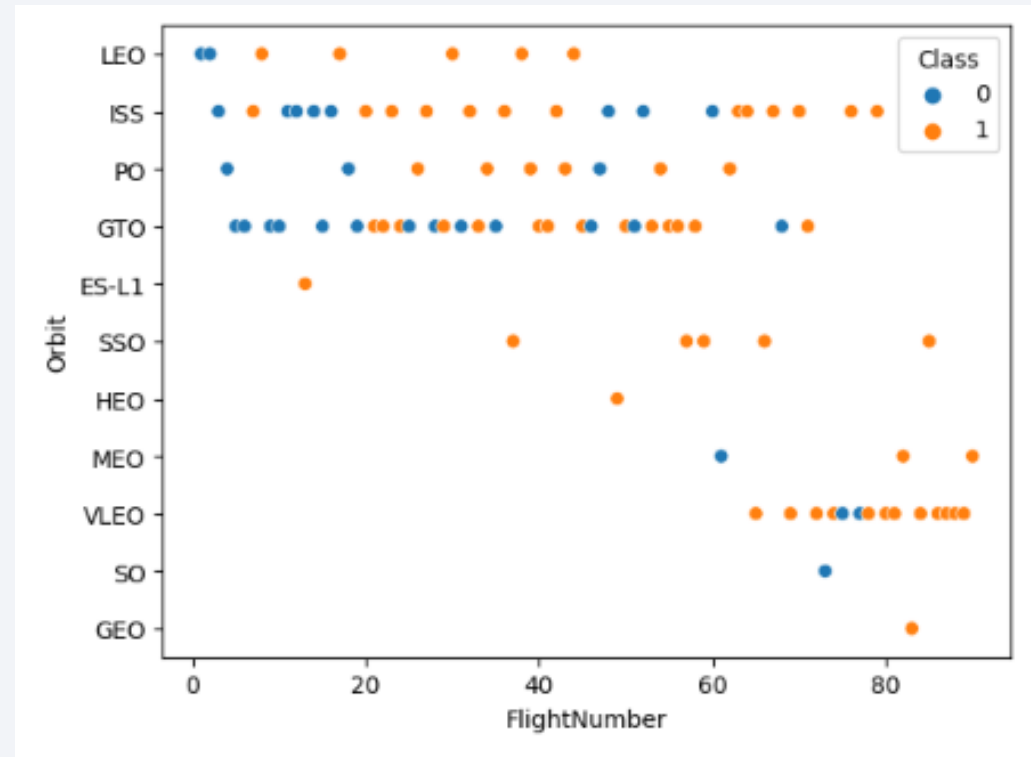
Success Rate vs. Orbit Type

- From the bar chart ES-L1, GEO, HEO and SSO are the orbit types with the highest success rate.

Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
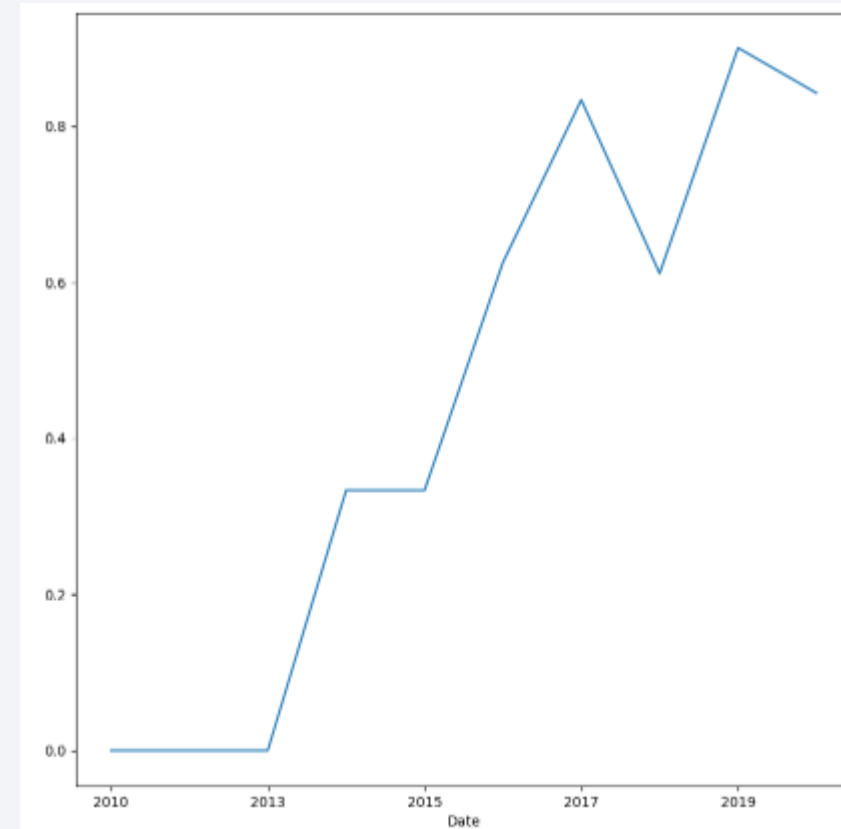
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
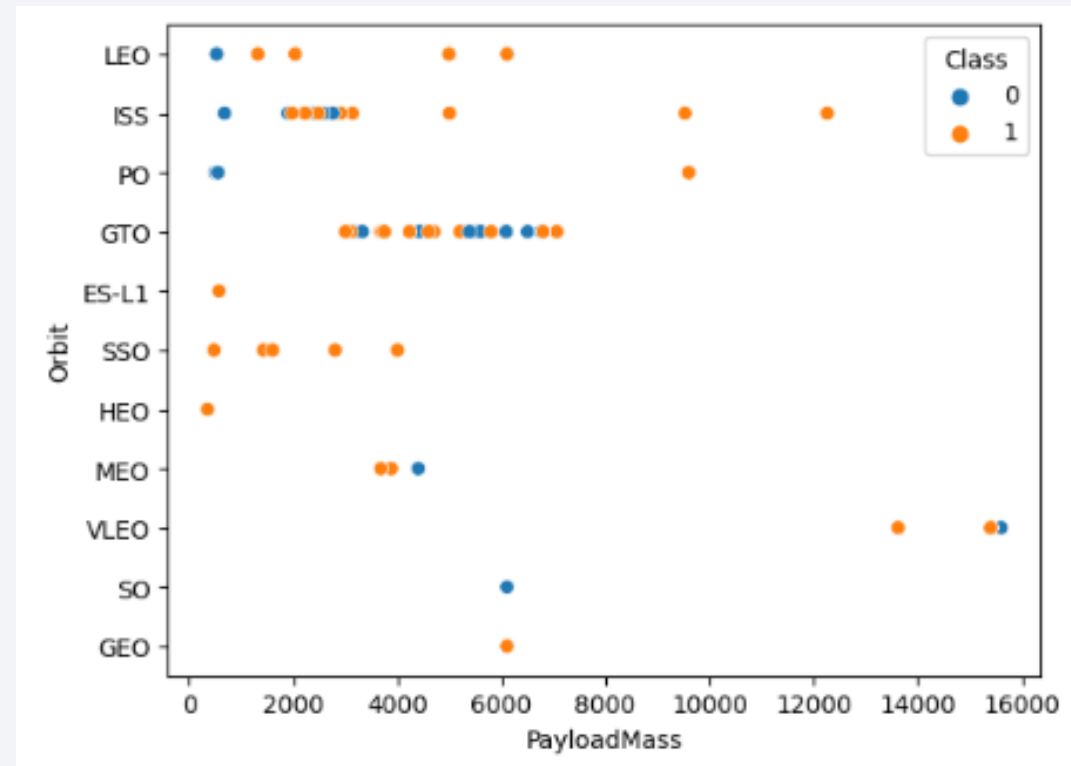
Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- %sql select distinct "launch_site" from SPACEXTBL This query displays the names of unique launch sites in the space mission.

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

Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTBL WHERE launch_site LIKE'CCA%' LIMIT 5 This sql query Displays 5 records where launch sites begin with the string 'CCA'.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Task 3

Total Payload Mass

- %sql select payload_mass__kg_ from SPACEXTBL where customer LIKE 'NASA (CRS)' This query selects the payload mass for NASA(CRS)

PAYLOAD_MASS__KG_
500
677
2296
2216
2395
1898
1952
3136
2257
2490
2708
3310
2205
2647
2697
2500
2495
2268
1977
2972

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%'

- This SQL query Display average payload mass carried by booster version F9 v1.1

AVG(payload_mass__kg_)

2534.666666666665

First Successful Ground Landing Date

- %sql SELECT min(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success (ground pad)'

- This query lists the date when the first successful landing outcome in the ground pad was achieved.

min(DATE)

01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT booster_version FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000

- This query lists the names of the boosters which have success in drone ship and have 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

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT mission_outcome,Count(DATE) as Count FROM SPACEXTBL GROUP BY mission_outcome.

- This query lists the total number of successful and failure mission outcomes.

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

Boosters Carried Maximum Payload

- %sql SELECT booster_version,payload_mass__kg_ from SPACEXTBL WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTBL)

- This query lists the names of the booster_versions which have carried the maximum payload mass.

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

- %sql select substr(Date, 4, 2) as month, "Landing _Outcome", booster_version, launch_site, date FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015'.

- This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

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

- %sql select count("Landing _Outcome") as COUNT, "Landing _Outcome" from SPACEXTBL where WHERE "Landing _Outcome" LIKE 'Success%' and substr(Date,7,4)>'2010-06-04' and substr(Date,7,4)<'2017-03-20' GROUP BY "Landing _Outcome".

- This query ranks 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.

COUNT	Landing _Outcome
4	Controlled (ocean)
5	Failure (drone ship)
12	No attempt
1	Precluded (drone ship)
12	Success (drone ship)
8	Success (ground pad)
2	Uncontrolled (ocean)

Section 3

Launch Sites Proximities Analysis

Launch sites and their locations
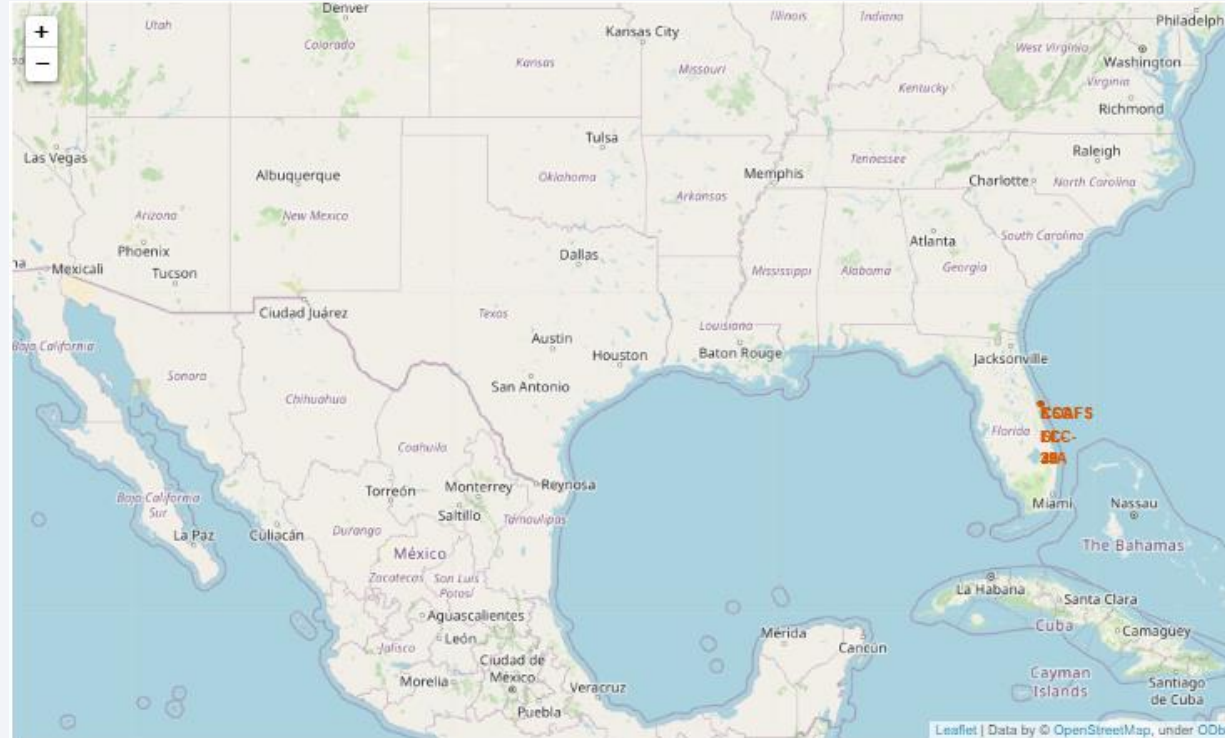
- From the map we can see all launch sites are in close proximity to the coast.
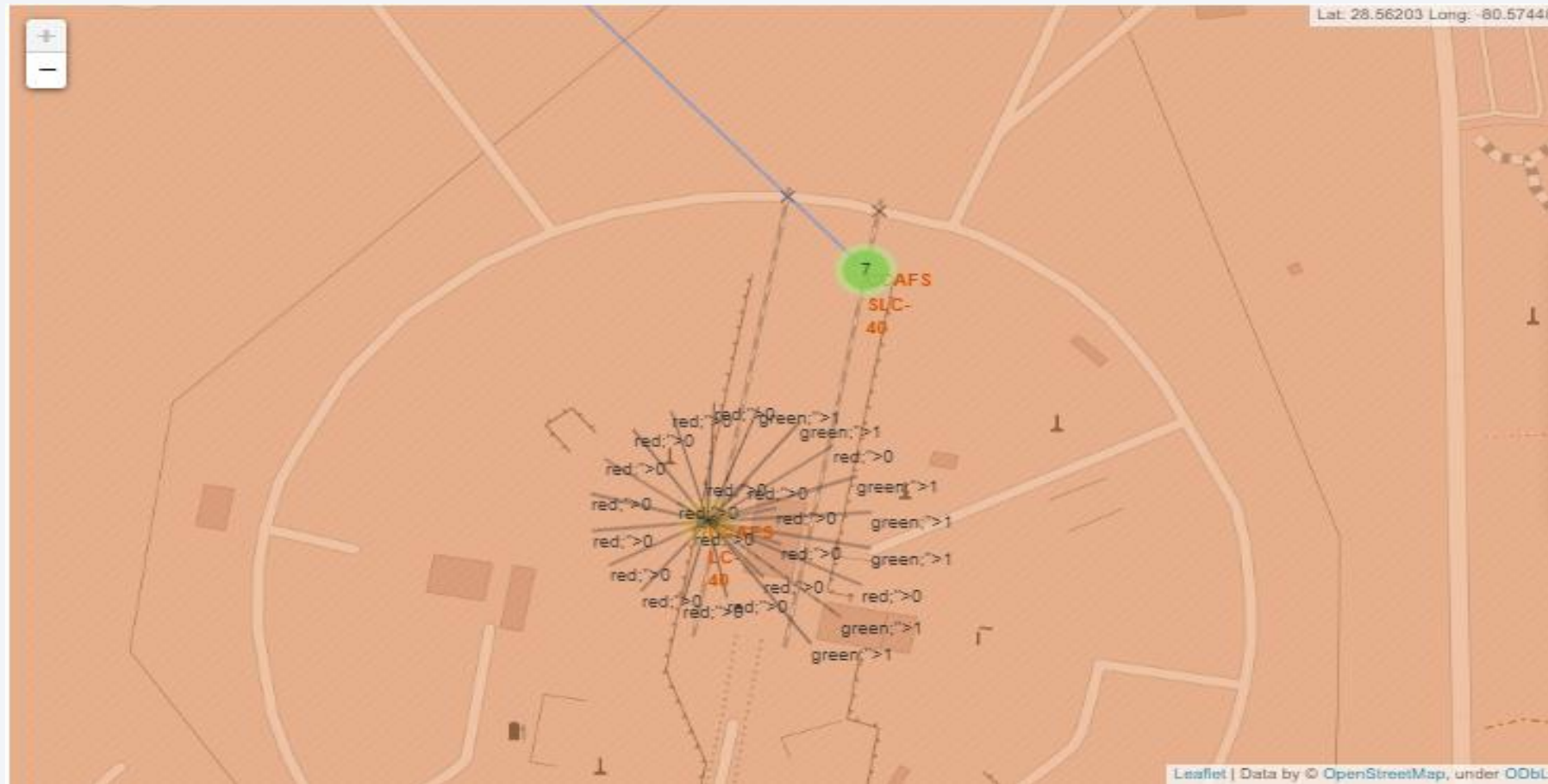
Launch outcomes on Map

- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Launch site and it's proximities

Section 4

Build a Dashboard
with Plotly Dash

Success Launches By Site

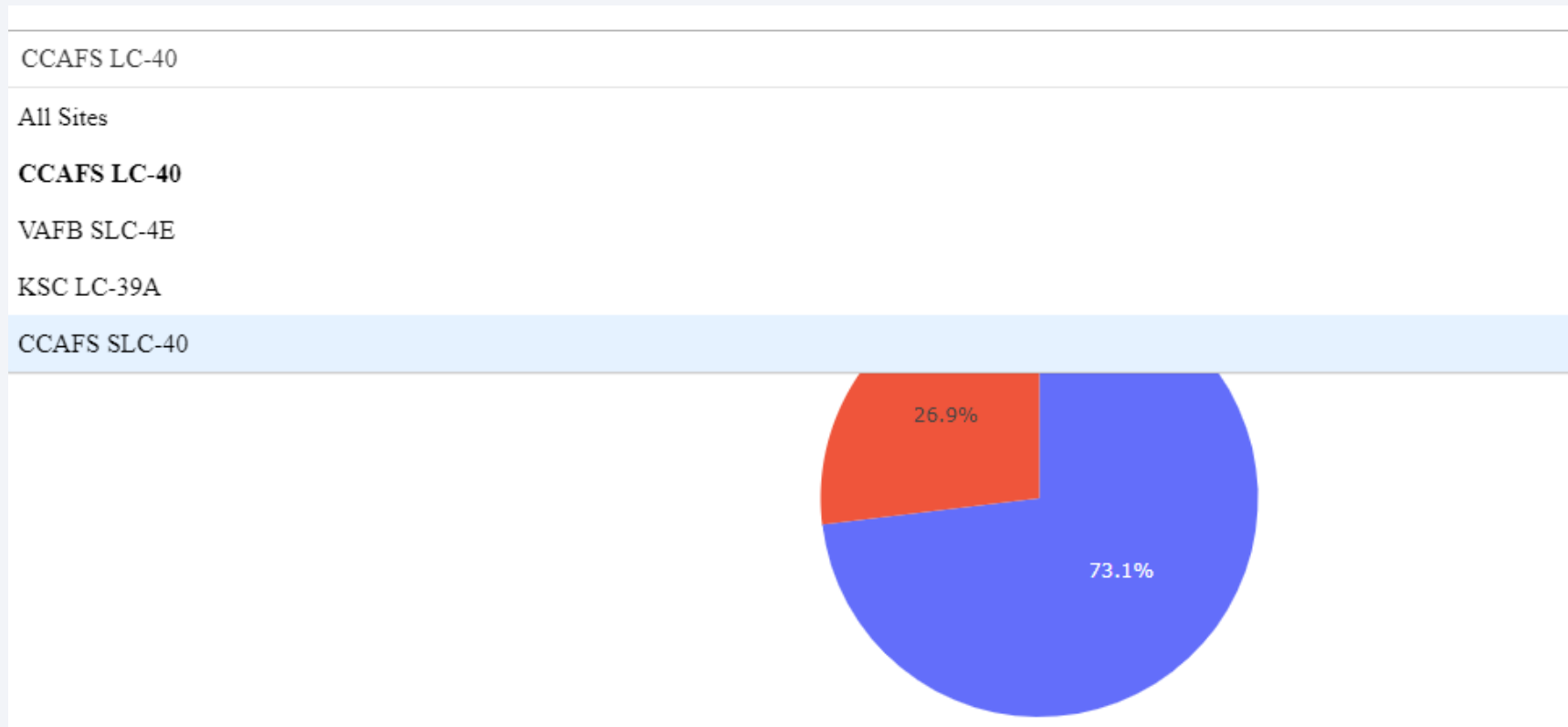
Total Success Launches By Site

KSC LC-39A — 41.7%
CCAFS LC-40 — 29.2%
VAFB SLC-4E — 16.7%
CCAFS SLC-40 — 12.5%

From the pie chart we can see that KSC LC-39A was the site with the most successful launches.

Drop-down feature
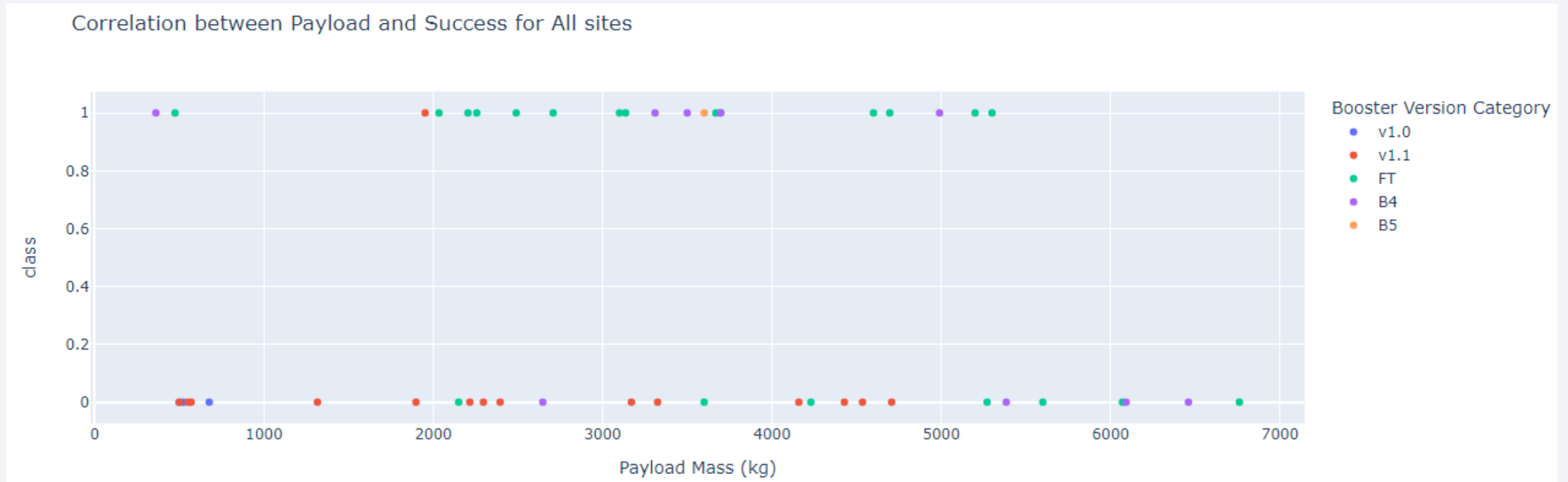

- Interactive drop-down feature of the web app showing all Launch site options.

Payload vs Launch Outcome

Correlation between Payload and Success for All sites

- Overview of Payload vs. Launch Outcome scatter plot for all sites, with all payloads selected in the range slider. From the chart, greater payload mass of above 6000kg have no success.

Section 5

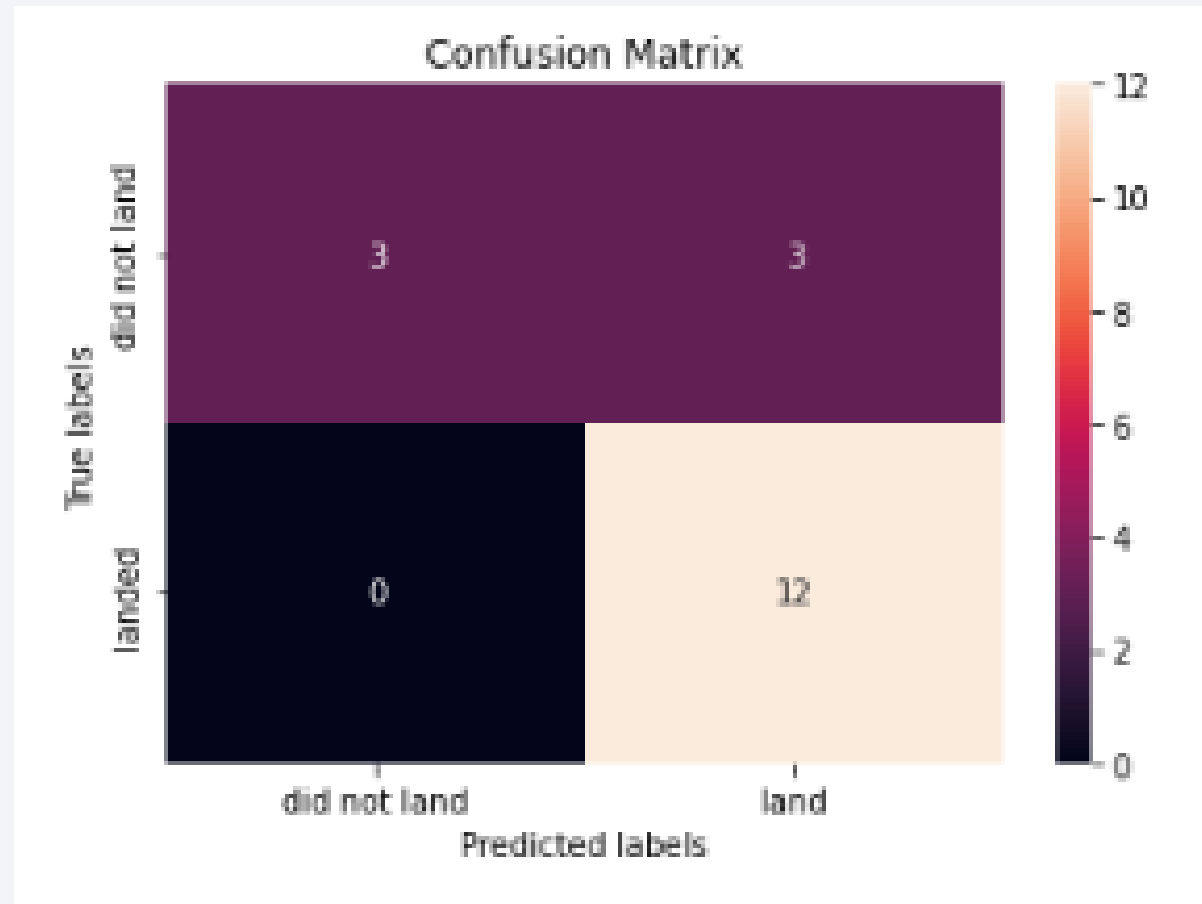**Predictive Analysis (Classification)**

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

Conclusions

• The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.

• Low-weighted payloads perform better than heavier payloads.

• The success rates for SpaceX launches are directly proportional time in years they will eventually perfect the launches.

• KSC LC 39A had the most successful launches from all the sites.

• Orbit GEO, HEO, SSO, ES L1 has the best Success Rate.

Thank you!