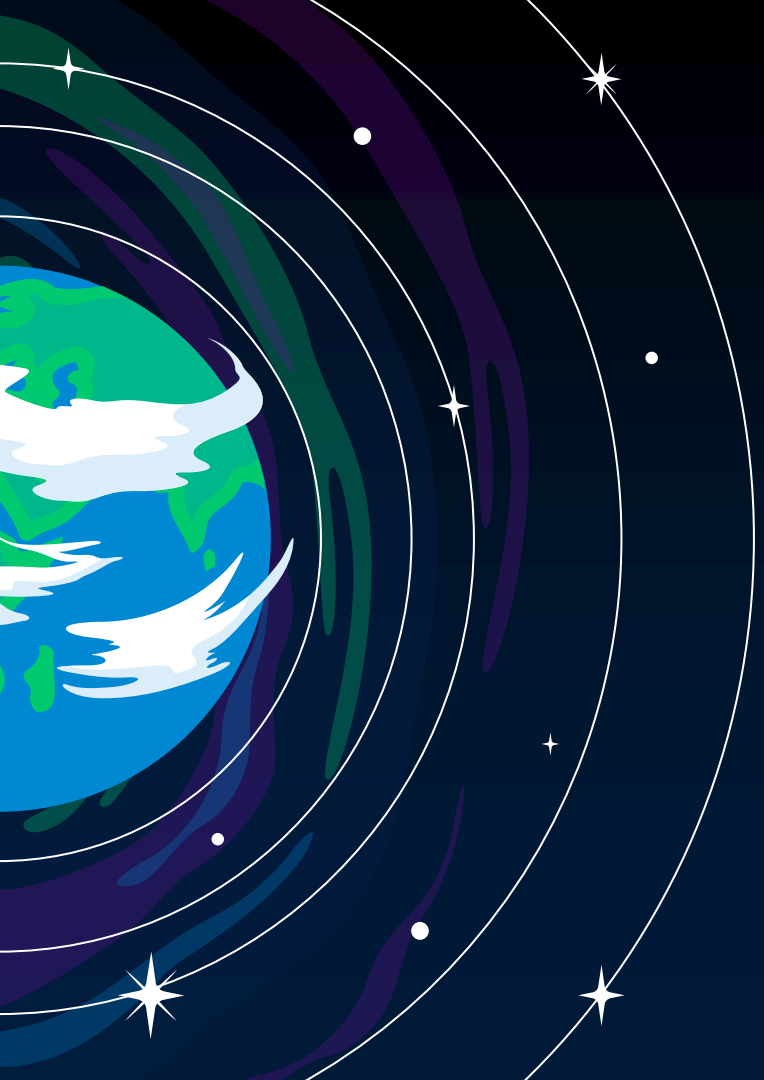


Data Science Capstone

Franco Stratta

November 1st 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclussion
- Appendix

Executive Summary

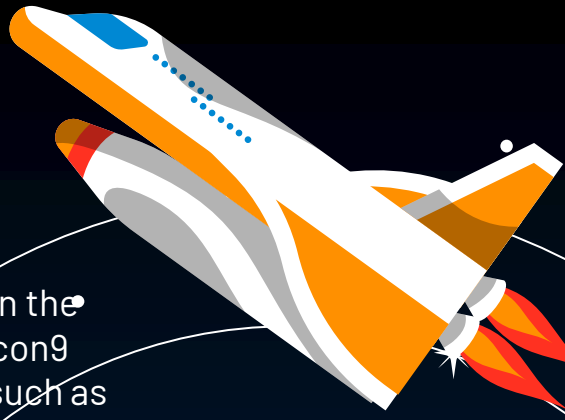
Summary of methodology:

- Data Collection API
- Data Collection with Web Scraping
- Data wrangling
- Explanatory Data Analysis with SQL
- Explanatory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Prediction using Machine Learning

Summary of all results:

- Explanatory Data Analysis results
- Interactive Visual Analytics
- Predictive Machine Learning Analysis results

Introduction



Project Background:

SpaceX is the most successful company of the commercial space age, in the persue of making space travel affordable. The company advertises Falcon9 launches on its website with a cost of 62 Million Dollars, other providers (such as Blue Origin), cost up to 165 Million Dollars, Much of the savings is because SpaceX can reuse the first stage. The goal of the project is to determine if the first stage will land, so we can determine the cost of the launch. We will create a machine learning pipeline to predict the first stage landing outcome.

Problems you will find answers:

- ¿What factors determine if the first stage will land succesfully?
- The interaction amongs various features cthat determine the success rate of a successful landing.
- ¿Does the rate of successful landings increase over the years?

Methodology

- **Data Collection methodology:**

SpaceX API and WebScraping from Wikipedia

- **Performed data wrangling:**

One Hot encoding applied to categorical features

- **Performed Exploratory Data Analysis (EDA) using visualization and SQL**
- **Performed Interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models:**

Building, tunning and evaluating the models to reach the best outcomes

Data Collection

Data collection process involved a combination of API request from SpaceX REST API and Web Scraping data from SpaceX Wikipedia
We have used both of these collection methods in order to get complete information about the launches for out analysis.

Data Columns obtained from SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, Launchsite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCont, Serial, Longitude, Latitude.

Data Columns Obtained from Web Scraping Wikipedia:

Flught No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection - Space X API

Requesting Rocket Launch Data from SpaceX API

Decoding the response content using *.json()* and turning it into a dataframe using *.json_normalize()*

Requesting needed information about the launches from SpaceX API by applying custom functions

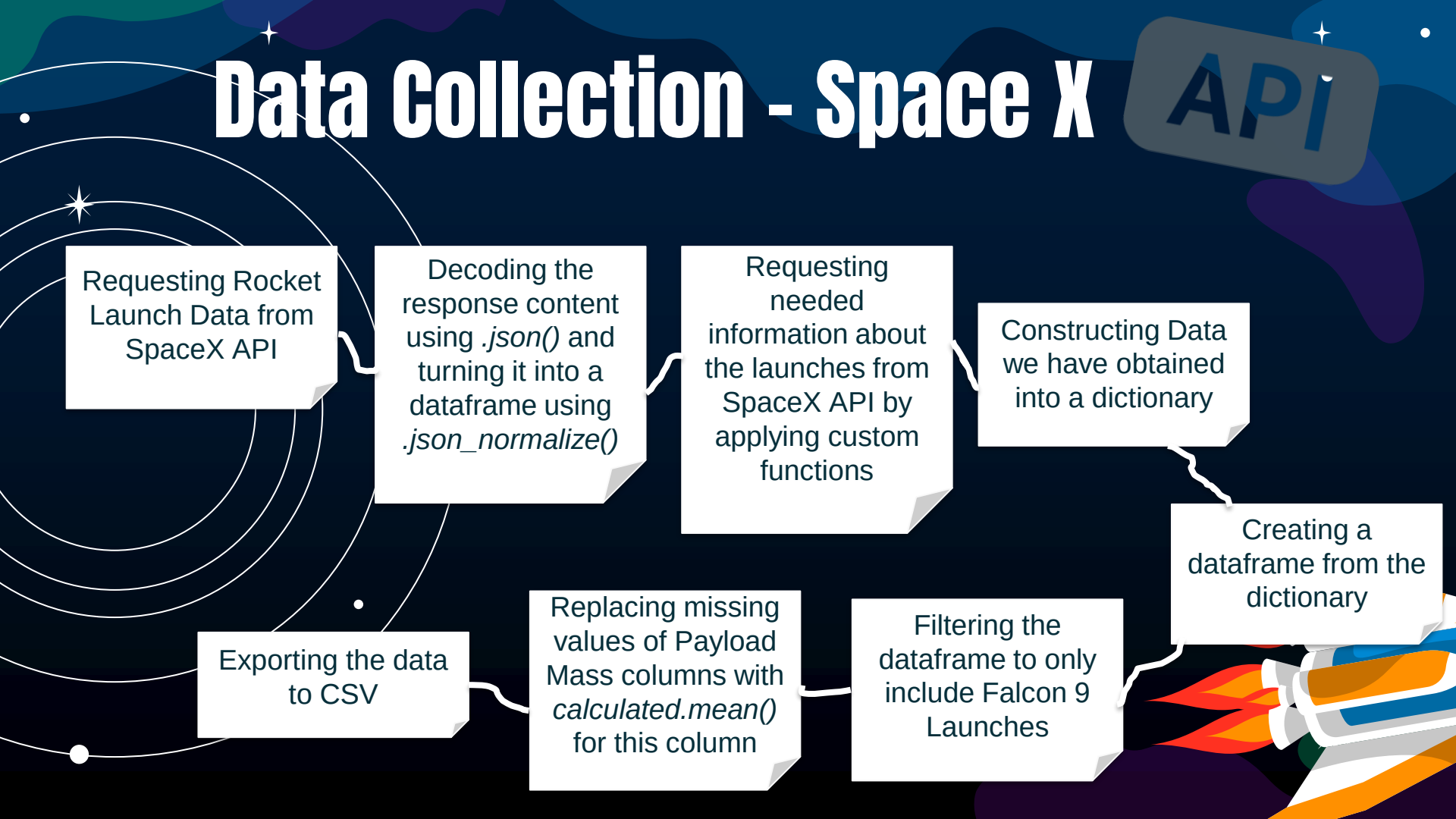
Constructing Data we have obtained into a dictionary

Creating a dataframe from the dictionary

Filtering the dataframe to only include Falcon 9 Launches

Replacing missing values of Payload Mass columns with *calculated.mean()* for this column

Exporting the data to CSV



Data Collection - Space X

Web Scraping

Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all the columna names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV

Data Wrangling

In the dataset, there are several cases where the booster did not land succesfully:

- True Ocean, True RTLS, True ASDS means the mission has been successful
- False Ocean, False RTLS, False ASDS mean the misión was a failure.

We need to convert those outcomes into Training labels with “1” meaning the booster succesfully landed, and “0” meaning the landing was unsuccessful.

Perform explanatory Data Analysis and determine training Labels.

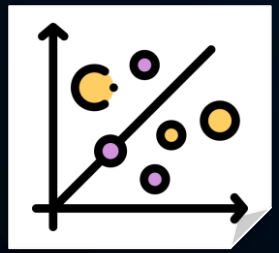
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of misión outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

EDA with Data Visualization

Scatter Graphs

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mass

Scatter plot show relationships between variables. This relationship is called correlation.



Bar Graph

- Success Rate vs. Orbit

Bar graphs show the relationship between numeric and categoric variables.

Line Graph

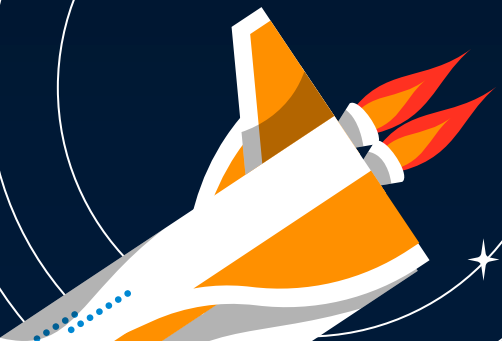
- Success Rate vs. Year

Line graphs shows data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.

EDA with SQL

We performed SQL queries to gather and understand data from dataset:

- Displaying the names of the unique launch sites in the space mission..
- Displaying 5 records where launch sites begin with the string “CCA”.
- Displaying the total Payload Mass carried by boosters launched by NASA (CRS).
- Displaying the average Payload Mass carried by booster version “F9 v1.1”.
- Listing the date when the first successful landing outcome in ground pad was archieved.
- Listing the names of the boosters which have success in drone ship and have a Payload Mass greater than 4.000 but less than 6.000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the Booster Versions wich have carried the maximum Payload Mass.
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in the year 2015.
- Ranking 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.



Buld an interactive Map with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houston, Texas

Red circle at NASA Johnson Space Center's coordinate with label showing its name **(folium.circle, folium.map.Marker).**

Red circles at each launch site coordinated with label showing launch site name **(*folium.Circle, folium.map.Marker, folium,features.DivIcon*).**

The grouping of points ina cluster to display multiple and different information for the same coordinated **(*folium.plugins.MarkerCluster*).**

Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. **(*folium.map.Marker, folium.Icon*).**

Markers to show distance between launch site to key locations (railwaym highway, coastway, city) and plot a line between them. **(*folium.map.Marker, folium.PolyLine, folium.features.DivIcon*).**

These objects are crated in order to understand better the problema and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

Buld DashBoard with Plotly Dash

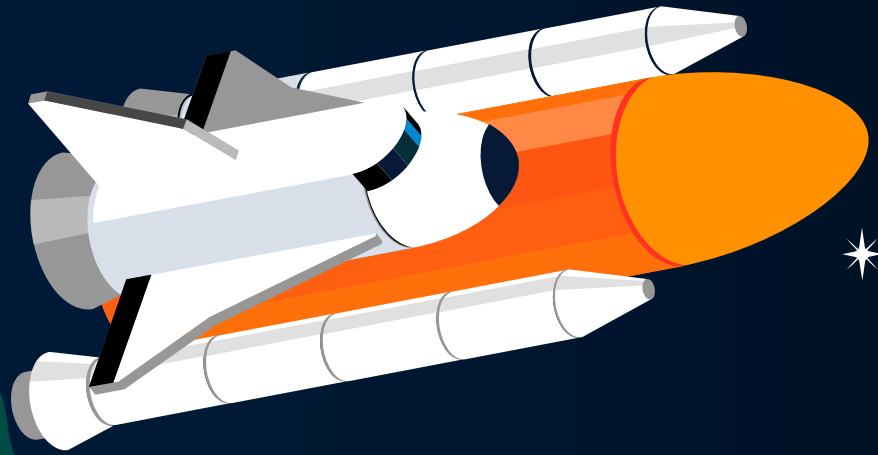
Dashboard has dropdown, pie chart, rangeslider and scatter plot components

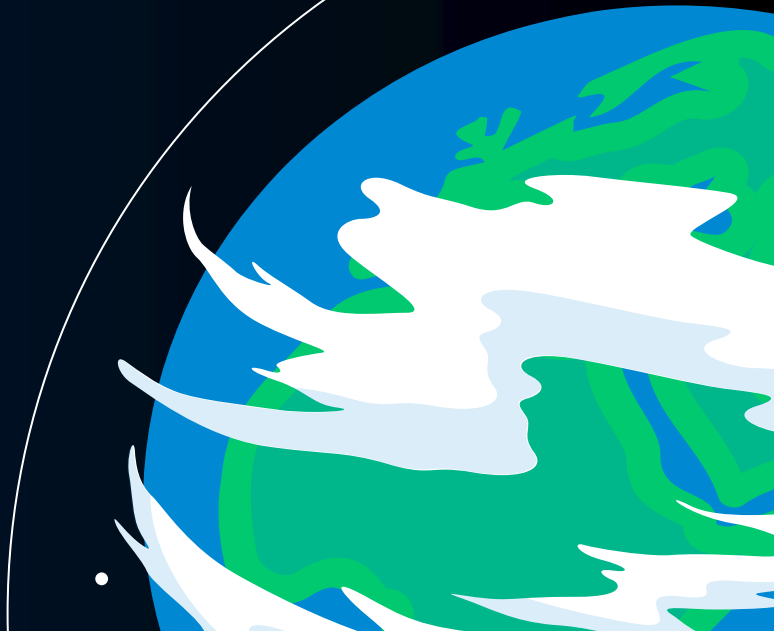
Dropdown allows a user to choose the launch site or all launch sites ***(dash_core_components.dropdown)***

Pie chart shows the total success and the total failure for the launch site chose with the dropdown component ***(plotly.express.pie)***

Rangeslider allows a user to select a Payload Mass in a fixed range ***(dash_core_components.RangeSlider)***

Scatter chart shows the relationship between two variables, in particular Sucess vs Payload Mass ***(plotly.express.scatter)***





Predictive analysis Classification

Creating a NumPy array from the columna “Class” in data

Standardizing the data with StandardScaled, then fitting and transforming it

Splitting the data into training and testing sets with *train_test_Split* function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method *.score()* for all models

Examining the confusión matrix for all models

Finding the best performing method by examining the *Jaccard_score* and *F1_score* metrics

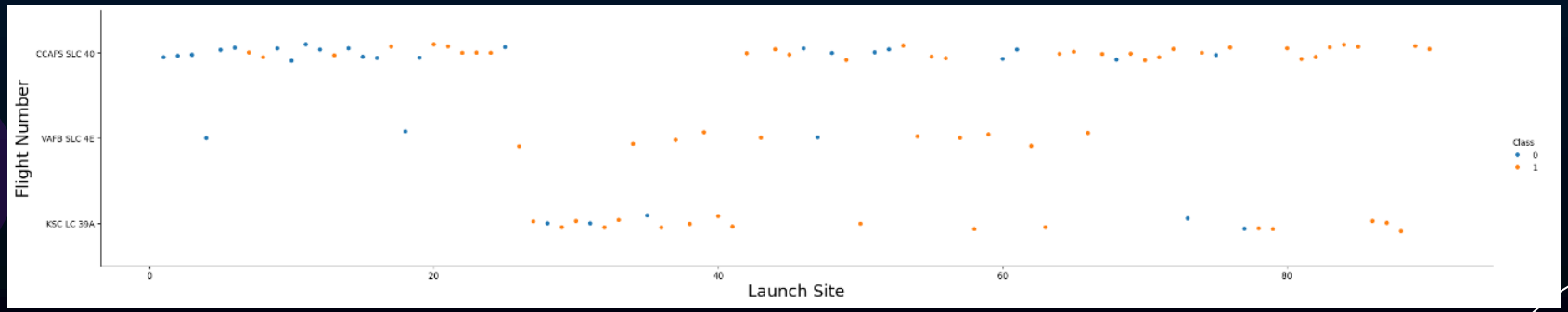
Results

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

EDA with Data Visualization

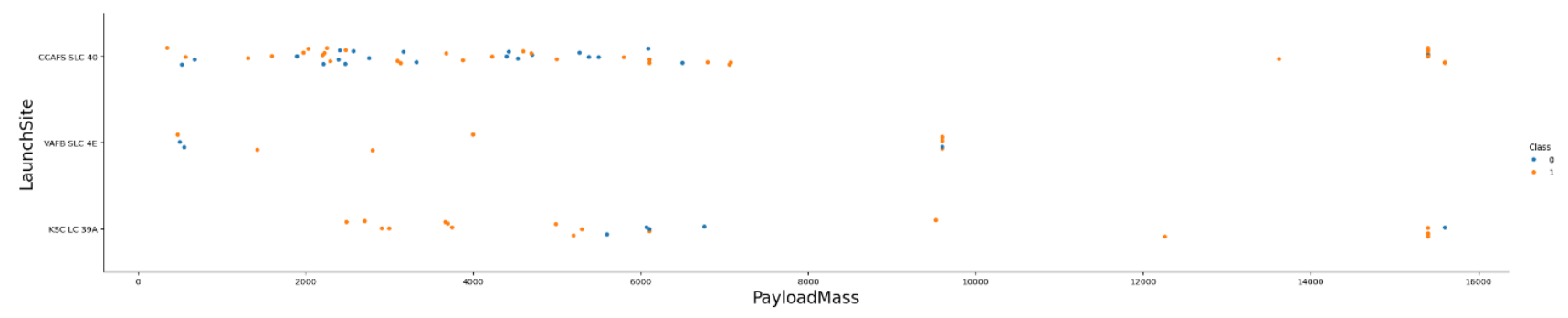
Flight Number vs. Launch Site

- The earliest flights all failed while the latest flights all succeeded.
- The **CCAFS SLC 40** launch site has about a half of all launches.
- **VAFB SLC 4E** and **KSC LC 39A** have higher success rates.
- It can be assumed that each new launch has a higher success rate.

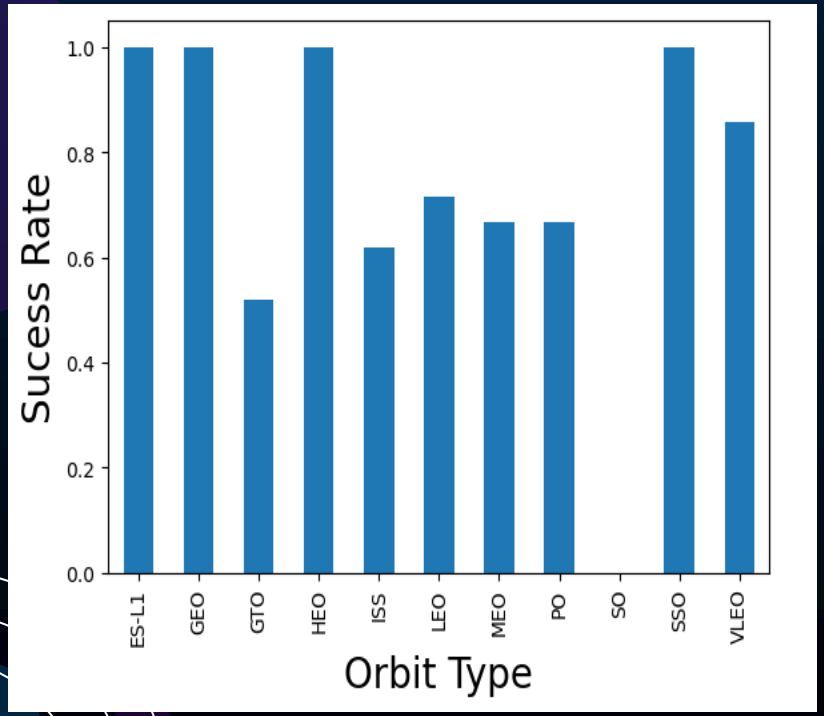


Payload vs. Launch Site

- For every launch site the higher the Payload Mass, the higher the sucess rate.
- Most of the launches with Payload Mass over 7000kg were successful
- **KSC LC 39A** has a 100% success rate for Payload Mass under 5500kg.



Success rate vs. Orbit type



Orbits with 100% success rate:

- ES-L1, GEO, HEO, SSO

Orbits with 0% success rate:

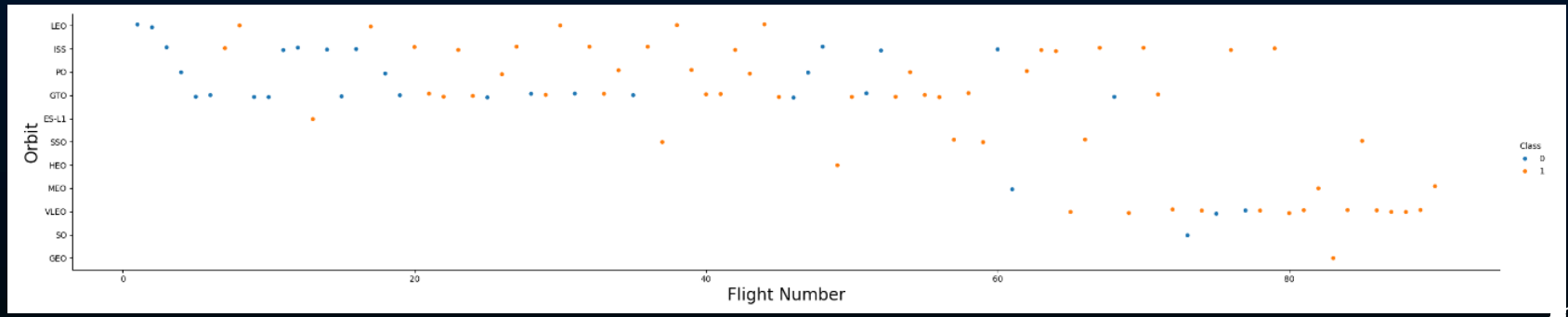
- SO

Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO

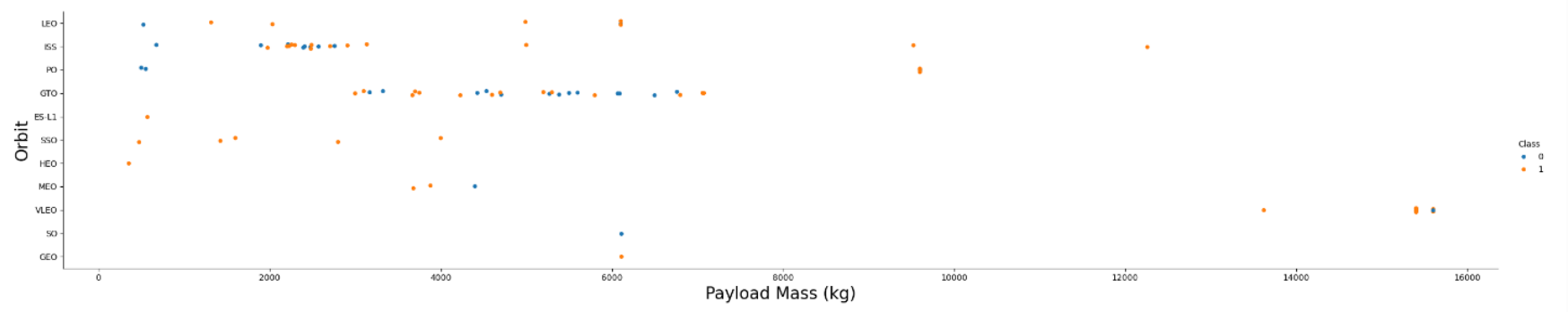
Flight Number vs. Orbit type

- In the LEO orbit the success appears related to the number of fligths; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

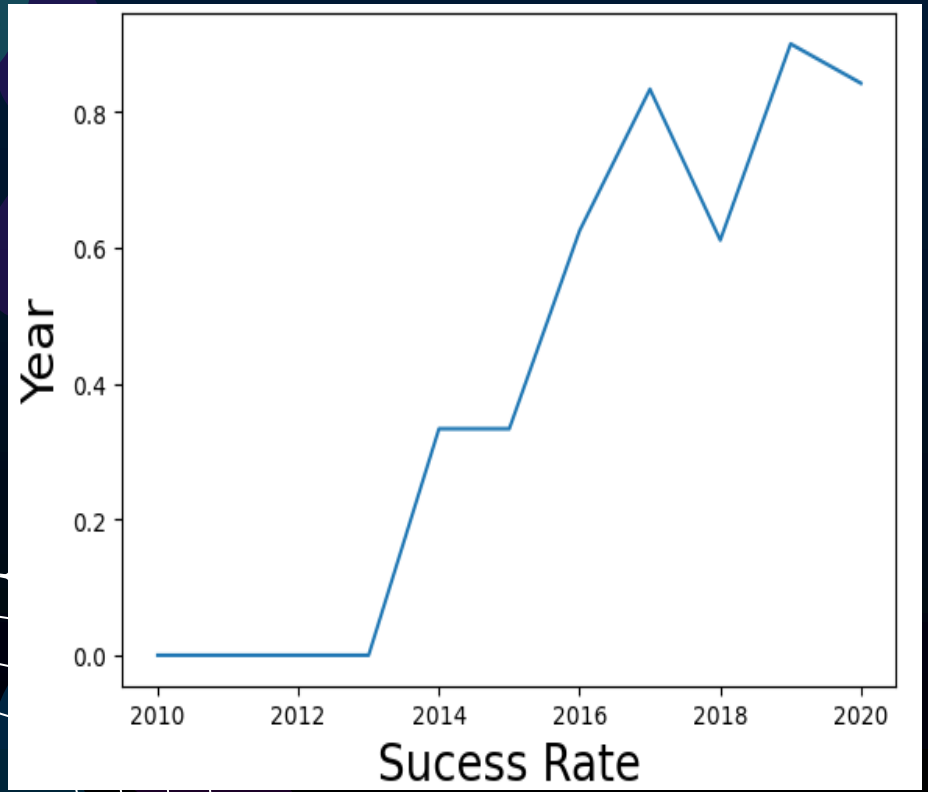


Payload Mass vs. Orbit type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits



Success rate vs. Orbit type



- The success rate since 2013 kept increasing since 2020

EDA with SQL

All launch site names

- Displaying the names of the unique launch sites in the space mission.

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Launch site names that begin with ‘CCA’

- Displaying 5 records where launch sites begin with the string ‘CCA’.

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

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

Total Payload Mass

- Displaying the total Payload Mass carried by boosters launched by NASA (CRS).

```
sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 1.1

- Displaying the average Payload Mass carried by boosters with version F9 1.1

```
sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First successful ground landing date

- Listing the date when the first successful landng outcome in ground pad was achieved.

```
sql SELECT MIN(DATE) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success'
```

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

MIN(DATE)
2010-04-06

Successful drone ship landing with Payload Mass between 4000 kg and 6000 kg.

- Listing the names of the boosters which have success in drone ship and have a Payload Mass greater tan 4000 but less than 6000.

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)'
```

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

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

Total Number of successful and failure mission outcomes

- Listing the total number of successful and failure mission outcomes.

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

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

2015 launch records

- Listing the failed landing outcomes in drone ship in 2015, their booster versions and the launch sites.

```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME='Failure (drone ship)' AND DATE LIKE '2015%'
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Ranking the count of landing outcomes (such as Failure (droneship) or success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

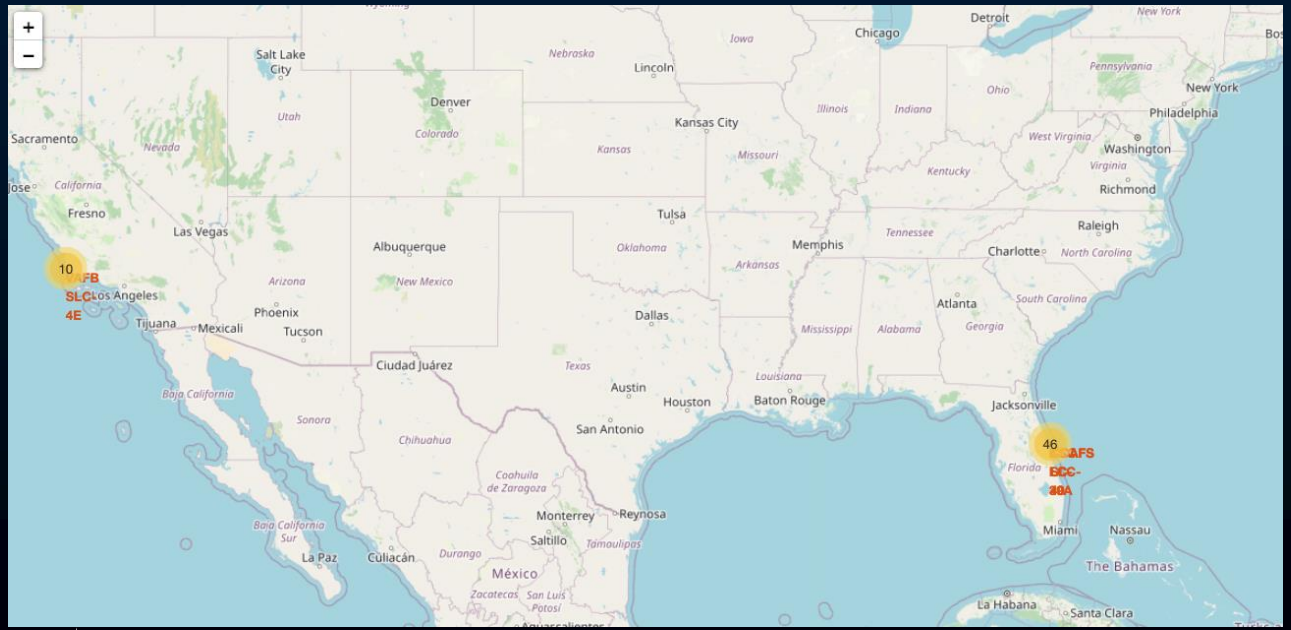
```
sql SELECT landing_outcome, COUNT(*) AS qty FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY qty DESC;
 * sqlite:///my_data1.db
Done.
```

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

Interactive Map with folium

Folium Map - Ground Stations

- We see that Space X launch sites are located on the coast of the United States.



Folium Map - Color Labeled Markers

- Green Marker represents successful launches.
- Red Marker represents unsuccessful launches

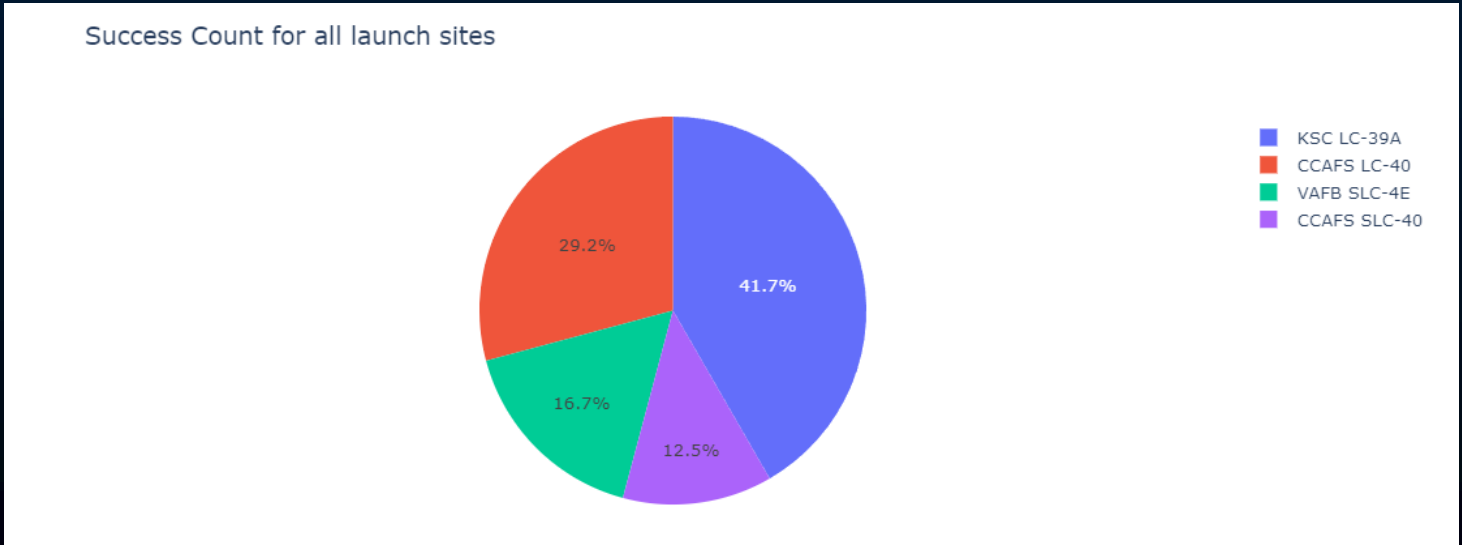
Folium Map - Distances between CCAFS SLC-40 and its proximities

- ¿Is CCAFS SLC-40 in close proximity to railways? **YES**
- ¿Is CCAFS SLC-40 in close proximity to highways? **YES**
- ¿Is CCAFS SLC-40 in close proximity to coastline? **YES**
- ¿Do CCAFS SLC-40 keeps certain distance away from cities? NO

Build a DashBoard with Plotly Dash

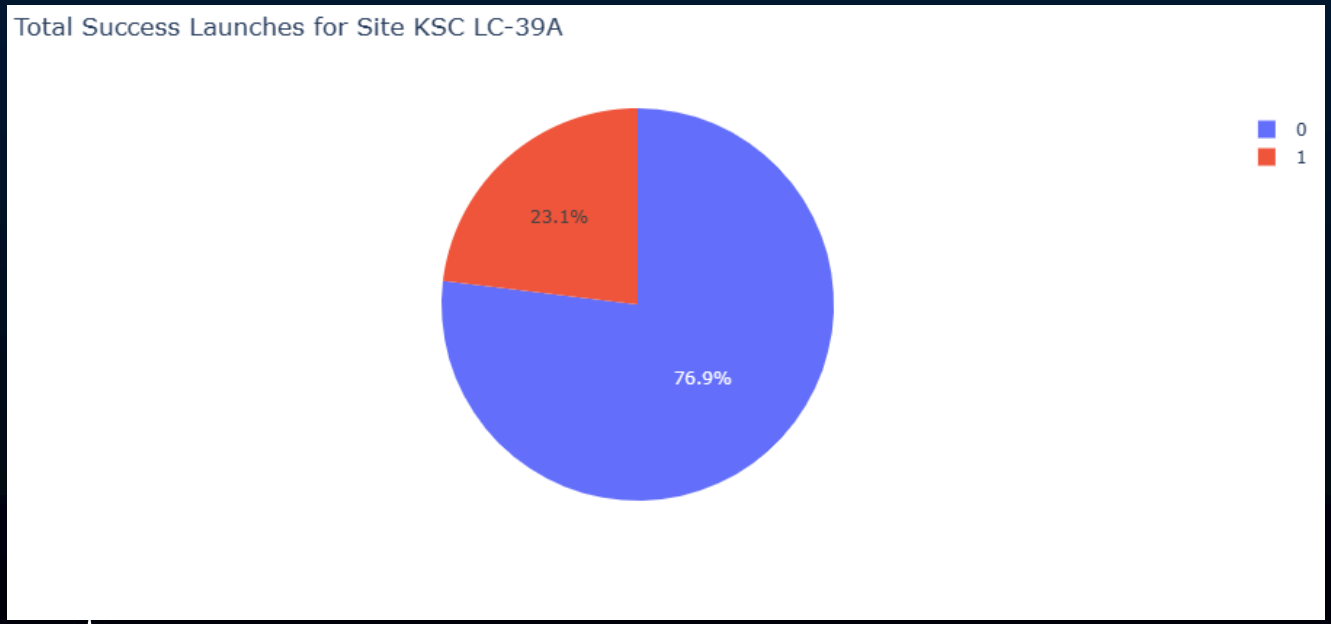
Launch success count for all sites

- The chart clearly shows that from all the sites, KSC LC-39ª has the most successful launches.



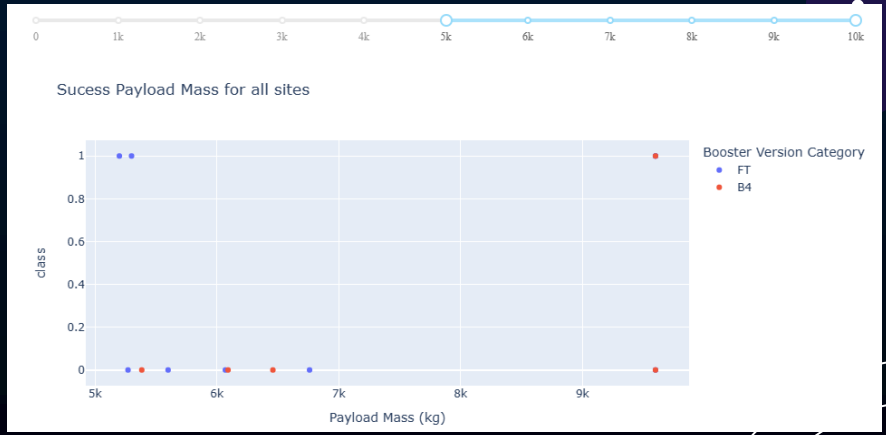
Launch site with highest launch success ratio

- KSC LC-39ª has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings



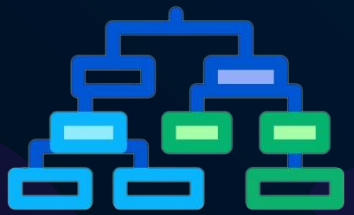
Payload Mass vs. Launch Outcome for all sites

- The charts shows that payloads between 2000kg and 5000kg have the highest success rate.



Predictive Analysis

(Classification)



Classification Accuracy

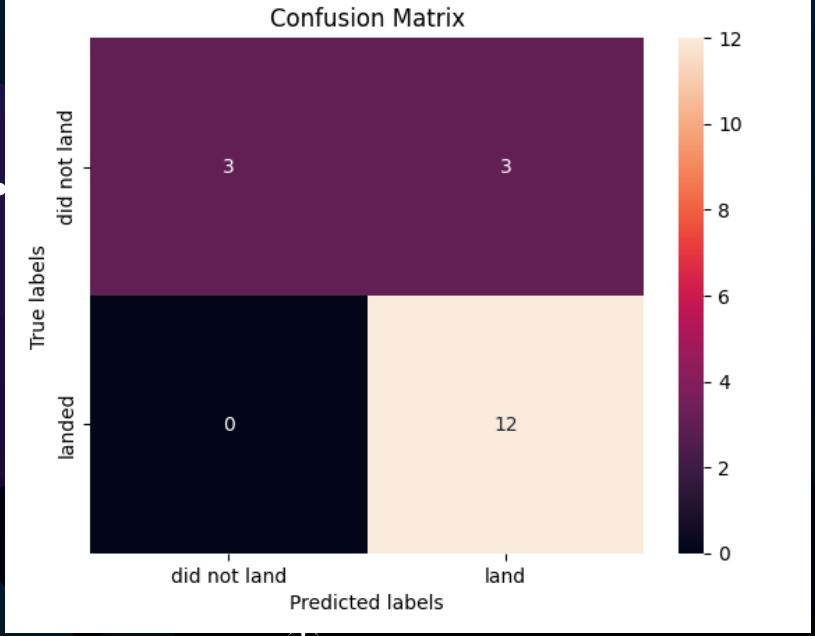
- The decisión tree classifier is the model with the highest classification accuracy.

```
alghorithms= {'KNN':knn_cv.best_score_, 'Tree':tree_cv.best_score_, 'Logistic Regression':logreg_cv.best_score_, 'SVM': svm_cv.best_score_}
best_alghorithm= max(alghorithms, key=alghorithms.get)
print('The best Algorithm is: ', best_alghorithm, 'with a score of:', alghorithms[best_alghorithm])
```

```
The best Algorithm is:  Tree with a score of: 0.9017857142857142
```

Confusion Matrix

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



The confusión matrix for the decisión tree classifier shows that the classifier can distinguish between the different classes.

- The major problem is the false positives (unsuccessful landing marked as successful landing by the classifier).

Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low Payload Mass show better results tan launches with a larger Payload Mass.
- Most of the launch sites are in proximity to the Equator Line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39ª has the highest success rate of all the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% sucess rate.

Appendix

Special Thanks to:

The instructors, i really enjoy it and it was a difficult but beautiful experience.

Coursera, I think it is the best way to learn from courses on the internet, the certificated that you can attain are extremely important in the work market.

IBM, the famous all time classic compute company responsable for financing this courses.

Thank You!!