Winning Space Race with Data Science

Henry Mauricio Rincon Caro

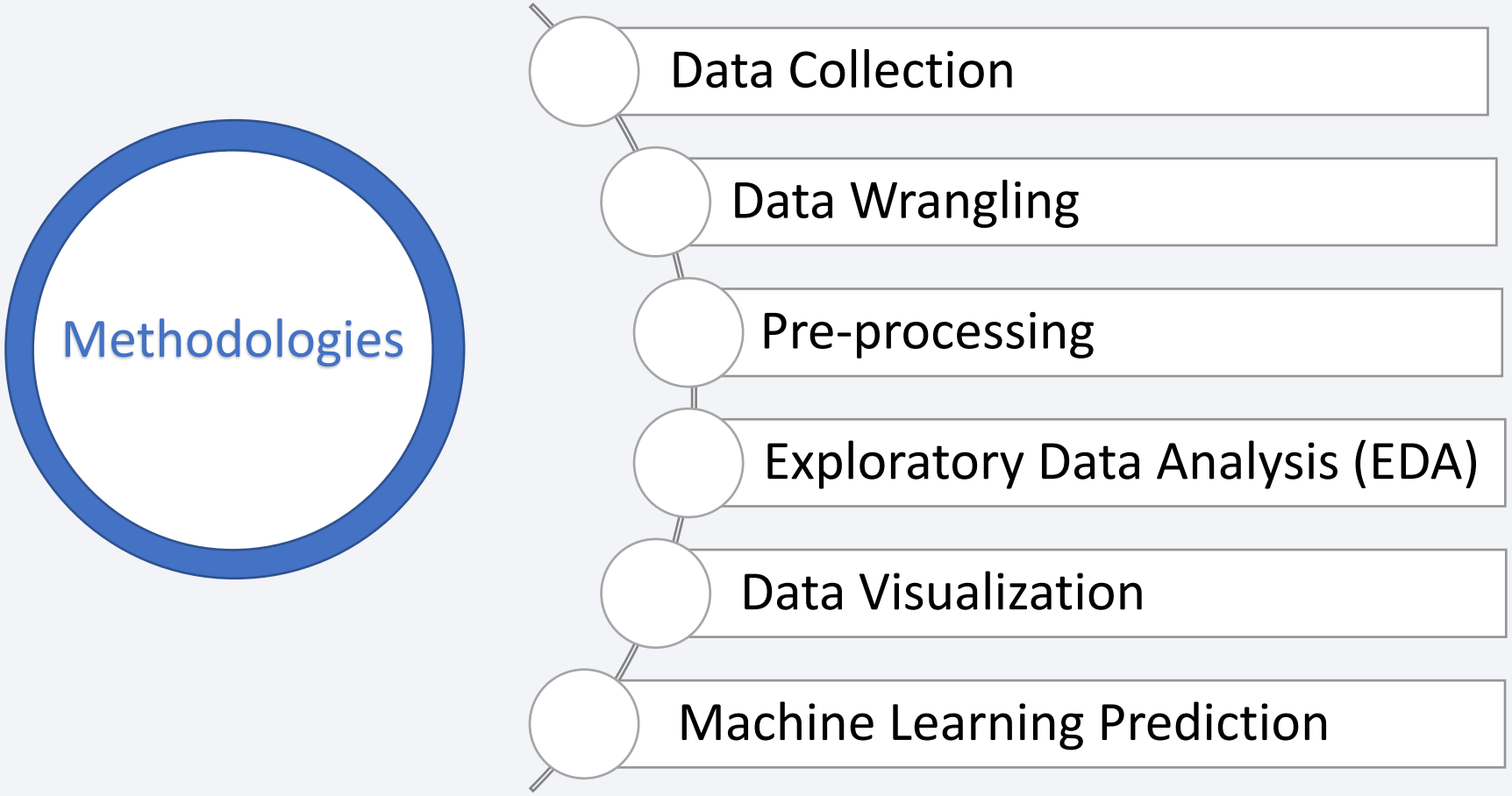
24/02/2023

Outline

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

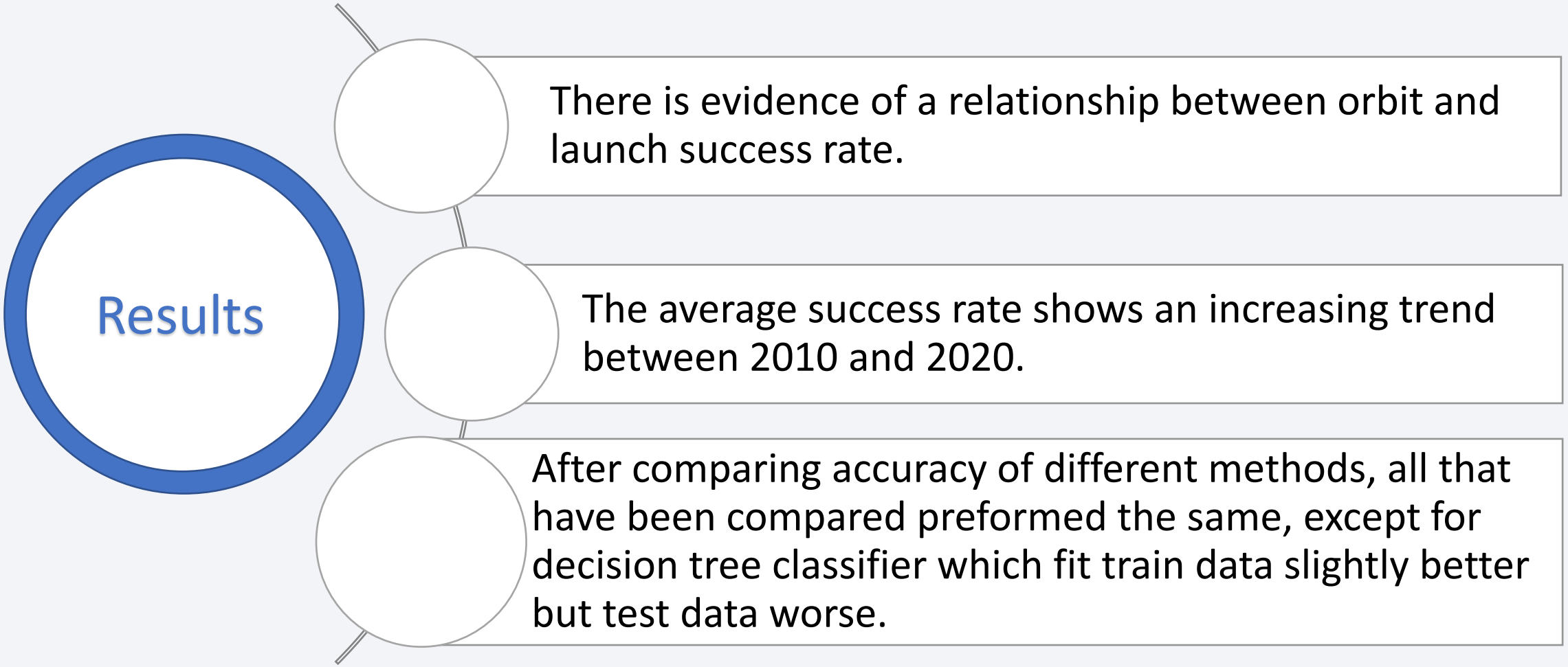
Executive Summary

In relation to the objective of this project, which sought to predict if the Falcon 9 first stage will land successfully, the following methodologies were used:



Executive Summary

Some of the most relevant collusions for the study are presented below:



Introduction

Context

The beginning of the commercial space age is evident, and, in this regard, it is possible to highlight **SpaceX** as the **most successful company**, given that they have been concerned with **low-cost rocket launches** and with clear objectives such as connecting with the International Space Station, and the provision of satellite internet service, in particular the costs have been reduced by around 63%, being a large **part of this saving given by the reuse of the first stage of the rockets**, so if it is possible to determine if the first stage will land, it is possible to determine the cost of a launch.

Problems to answer

Taking the context into account, the **main question** that we try to answer in this project is:

Given the characteristics in which the launch takes place, is it possible to determine if a **future launch of a rocket will be successful relative to the first stage landing**?

Section 1

Methodology

Methodology

Executive Summary

First

In the data collection activity (Collecting the data) two main methods were used, namely, the first compiling the information through an api specifically SpaceX REST API, the second method corresponded to web scraping related Wiki pages using Python BeautifulSoup package.

Second

Regarding the Data wrangling process, it was executed in order to transform and purify the data, for this process the Python Pandas library was used, obtaining raw data to a clean dataset.

Third

In particular, for the exploratory analysis (EDA), SQL queries and the visualization method with the Matplotlib library were used, in which general elements of relevance to the study were identified and in the same way a cast was made to create dummy variables to categorical columns.

Fourth

An interactive analysis was carried out using Folium and Plotly Dash, with which it was possible to explore and manipulate the data in an interactive and real-time way, the first part with Folium the analysis was focused on analyzing launch site geo and proximities and then built an application with the Python Plotly Dash package.

Fifth

In relation to the approach used for the predictive analysis, different methods were evaluated, such as logistic regression, support vector machine, decision tree classifier k nearest neighbors, using the best hyperparameter values, the model with the best accuracy using the training data was determined and the confusion matrix was outputted

Data Collection

Regarding collecting the data, two main approaches were used:

API

- Working with the SpaceX launch data, the request was made to the SpaceX API.

Web Scraping

- Web scraping of the Wikipedia page was used with the BeautifulSoup library

The information obtained corresponds to:

Payload delivered → Launch specifications → Landing specifications → Landing outcome

Data Collection – SpaceX API

1. Request rocket launch data from SpaceX API using get request
2. Parse the SpaceX launch data
3. Use json_normalize method to convert the json result into a dataframe
4. Get only the columns relevant to the project goal using subset of the dataframe
5. Debugging or cleaning the data obtained from the API
6. Apply functions to obtain the required data from the selected columns
7. Build a DataFrame from the data obtained in the previous step using a dictionary
8. Get only the data that corresponds to Falcon 9 using DataFrame filter

GitHub URL
click on this icon

Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/a8e78008b7ad6474e4690b3fa8e40cca886b9e60/1.07%20-%20API%20Lab.ipynb

Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page from its URL performing an HTTP GET method and using a BeautifulSoup object
2. Select the information that corresponds exclusively to the lists of Falcon 9 and Falcon Heavy launches using find_all function
3. Extract all column/variable names from the HTML table header by iterating over each element
4. Create a data frame by parsing the launch HTML tables through a dictionary

GitHub URL
click on this icon

Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/a8e78008b7ad6474e4690b3fa8e40cca886b9e60/1.09%20-%20Web%20Scraping.ipynb

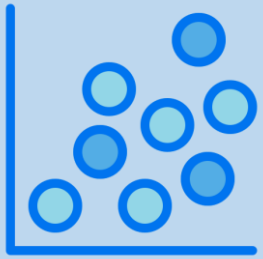
Data Wrangling

In this aspect of data wrangling, some of the main attributes were reviewed by using the value_counts function, attributes such as LaunchSite that includes the different launch sites, orbits that contains different payload orbits, Outcome that contains the first stage successfully landed, with the value_counts function, in the same way, missing data of the null type were completed and the columns that are numeric and cateoric were identified, particularly in the outcome column data were casting from categorical values to an indicator variables (1 or 0).

1. Identify and calculate the percentage of the missing values in each attribute
2. Identify which columns are numerical and categorical
3. Calculate the number of launches on each site
4. Calculate the number and occurrence of each orbit
5. Create a landing outcome label from Outcome column

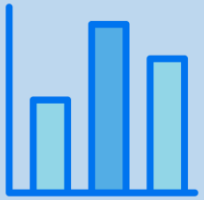
Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/a8e78008b7ad6474e4690b3fa8e40cca886b9e60/1.10%20-%20Data%20wrangling.ipynb

EDA with Data Visualization



The scatter plots were used to determine if there is a relationship between two variables, the pairs of variables selected were:

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



The bar graph sought to establish a comparison of the values obtained in different groups, in particular the landing success rate and the type of orbit.

The line graph seeks to establish the trend generally for a range of time, in this case the trend of the landing success rate between the years 2010 to 2020

Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/a8e78008b7ad6474e4690b3fa8e40cca886b9e60/2.3%20-%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

Identify the names of unique launch sites in the space mission.

Display first 5 records where launch sites begin with the string 'KSC’.

Calculate the total payload mass carried by boosters launched by NASA (CRS)

Calculate average payload mass carried by booster version F9 v1.1

List the date where the first successful landing outcome in drone ship was acheived.

List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

List the total number of successful and failure mission outcomes

List the names of the booster_versions which have carried the maximum payload mass

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

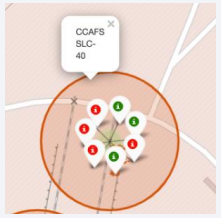
Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/a8e78008b7ad6474e4690b3fa8e40cca886b9e60/2.2%20-%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

Highlighted circle area
Identify initial center location to be NASA Johnson Space Center at Houston, Texas

Highlighted multiple circle area
Identify each launch site in data frame with the information included in the column lunch_sites

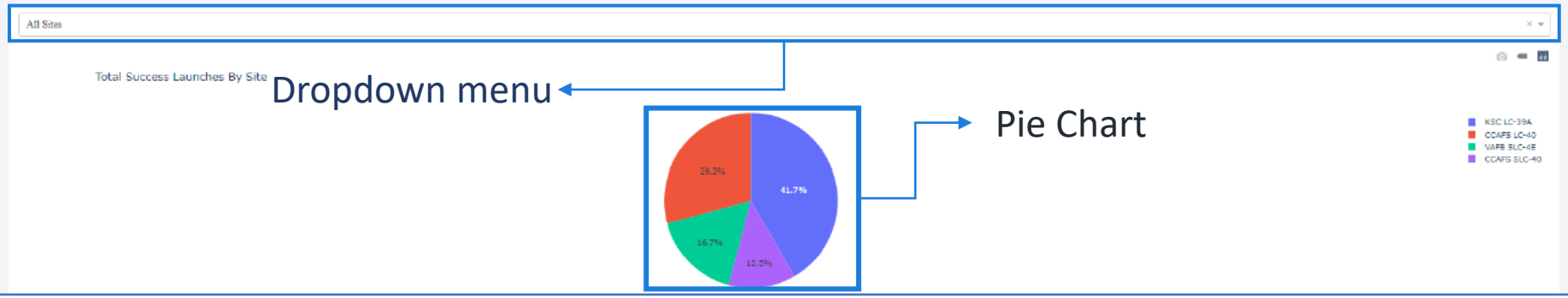
Polyline
Identify the launch site closest to selected coastline point



Color-labeled markers
identify which launch sites have relatively high success rates

Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/c376dd59b5ad728f544e48478c326705d88ecb39/3.1%20-%20Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash



Pie chart

Shows the launches that were successful for each of the locations where the launches were made, this graph has a dropdown menu that allows you to filter the content for each of the different launch sites or by default see all.

Scatter plot

Shows the relationship between the payload and the launch outcome and how payload may be correlated with mission outcomes for selected site, in the same way the Booster version at each dispersion point was labeled with colors, this graph also connects with the dropdown menu that was explained above and included a range slider component for the variable year.

Predictive Analysis (Classification)

From the preprocessing in which the data is divided into training data and test data, it is possible to train the model and perform Grid Search, in this way it is possible to find the hyperparameters that allow a given algorithm to perform best, within the models that are evaluated are:

- Logistic Regression
- Support Vector machines
- Decision Tree Classifier
- K-nearest neighbors

Likewise, the analysis is supported in a complementary way with the confusion matrix.

1. Create a NumPy array from the column in Class in data
2. Standardize the data
3. Split the data X and Y into training and test data
4. Fit the object to find the best parameters, calculate the accuracy on the test data and output the confusion matrix, for Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
5. Analyze the information above and find the method performs best

Link Github: https://github.com/mauriciohmrc/CapstonProjectIBM/blob/c376dd59b5ad728f544e48478c326705d88ecb39/4.1%20-%20Machine%20Learning%20Prediction%20lab.ipynb

Results

Exploratory data analysis

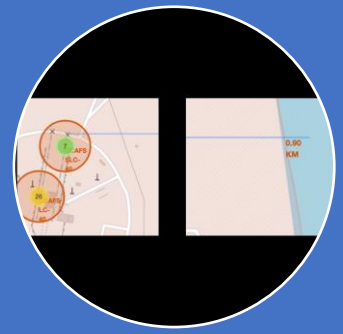
1. It is evident that as the flight number increases, the Pyload mass also increases until reaching maximums of 15,600 kg.
2. It was found that the first launches were carried out in CCAFS SLC-40, after that they focused in VAFB SLC-4E, while the latest launches are distributed between CCAFS SLC-40 and KSC LC-39A.
3. It is evidenced that for the VAFB-SLC launch site there are no rockets with a heavy payload mass more than 10000.
4. There is a relationship between the launch success rate and the orbit, the orbits with the highest success rate are ES-L1, GEO, HEO and SSO, while those with the lowest success rate are GTO and ISS.
5. It is evident that the last launches were carried out more frequently to the VLEO orbit, which has a success rate greater than 0.8, but it is not the one with the highest success rate.
6. The ISS, LEO and ISS orbits are those with the heaviest payloads the successful landing
7. The success rate of the launches presents a constant increasing trend from 2013 to 2020

Results

Interactive analytics

In general, launches are carried out from the west coast mainly at VAFB SLC-4E, and from the east coast at lunch sites KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40.

The largest number of launches occurred on the east coast with 46, while on the west coast in the evaluated period there were a total of 10 launches



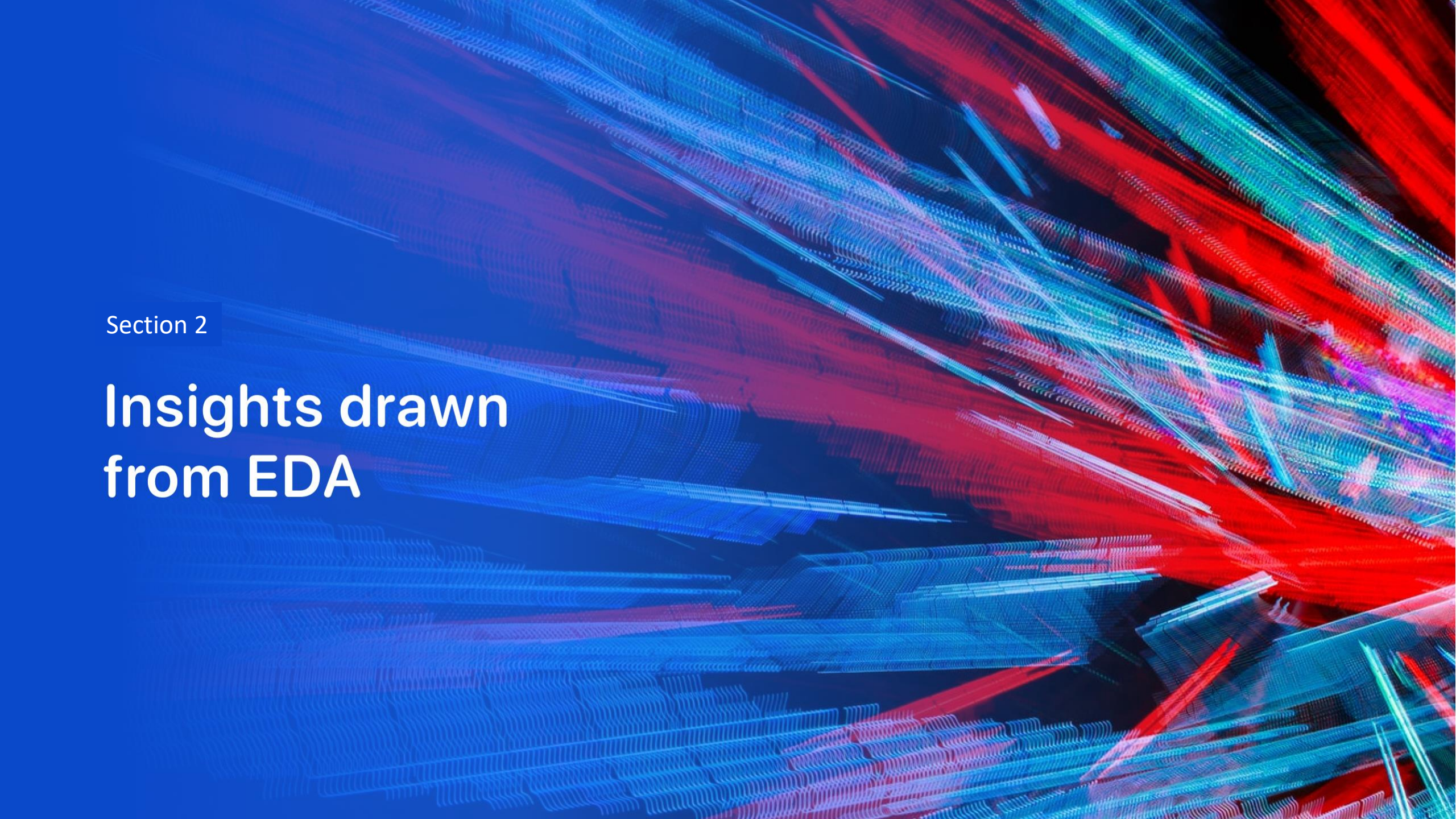
The shortest distance to the coastline is from the Space Launch Complex 40 (CCAFS SLC-40), this distance corresponds to 0.90 km, and is close to railways, highways and the closest city is Melbourne.

Results

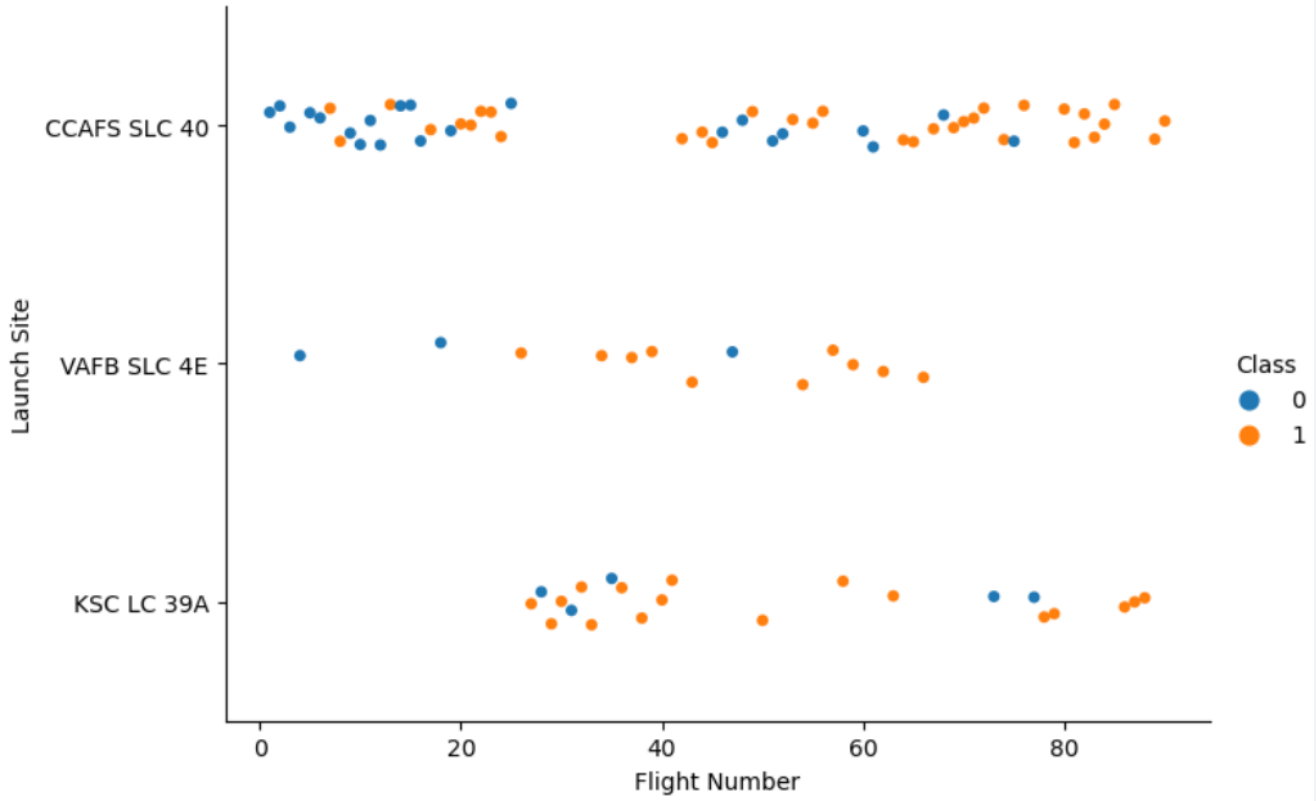
Predictive analysis

Model / Value	Logistic Regression	Support Vector machines	Decision Tree Classifier	K-nearest neighbors
Accuracy Train Data	0.8464285714285713	0.8482142857142856	0.8767857142857143	0.8482142857142858
Accuracy test Data	0.8333333333333334	0.8333333333333334	0.8333333333333334	0.8333333333333334

Comparing the accuracy of the models applied to the test data and visualizing the confusion matrix for each one, we can verify that the models carry out a prediction with similar accuracy, however, the Decision Tree Classifier model fit train data slightly better

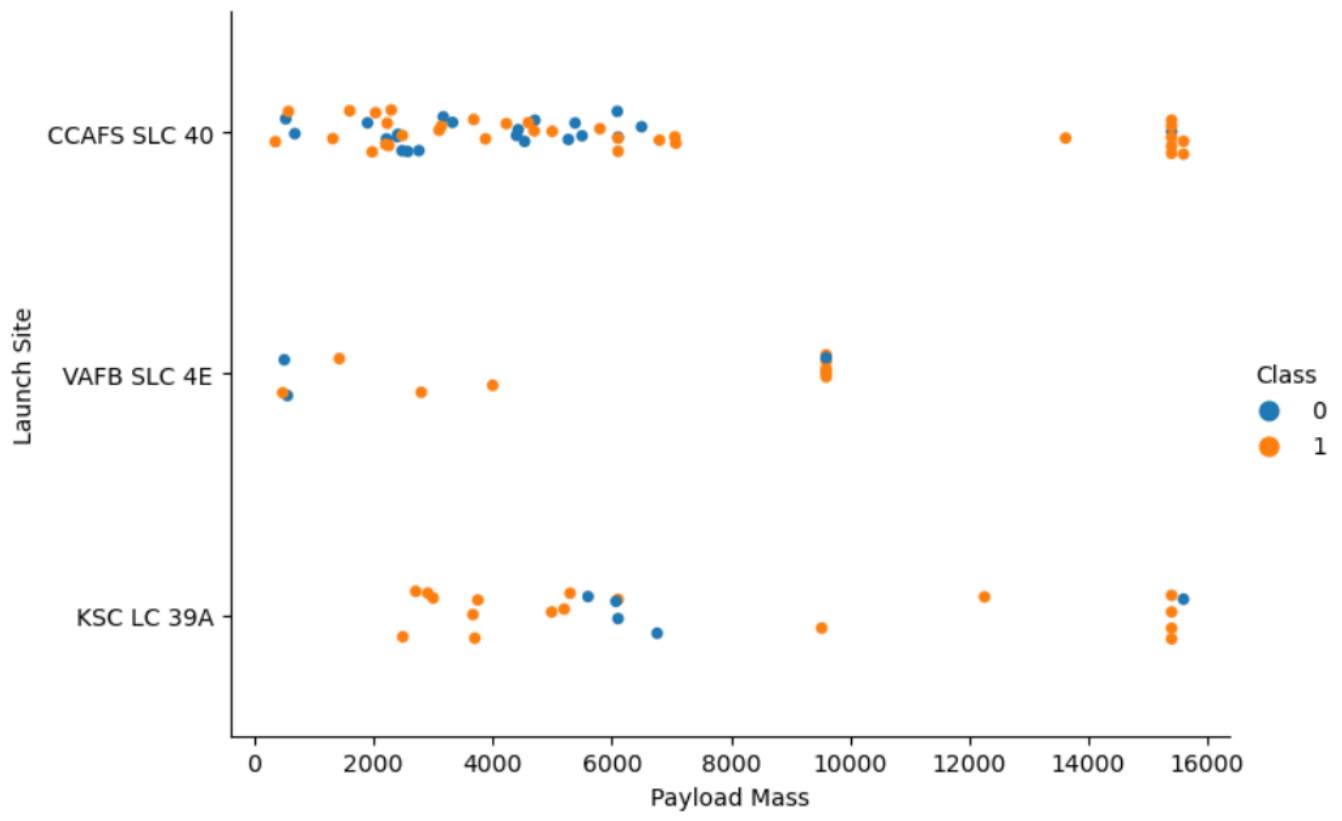

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



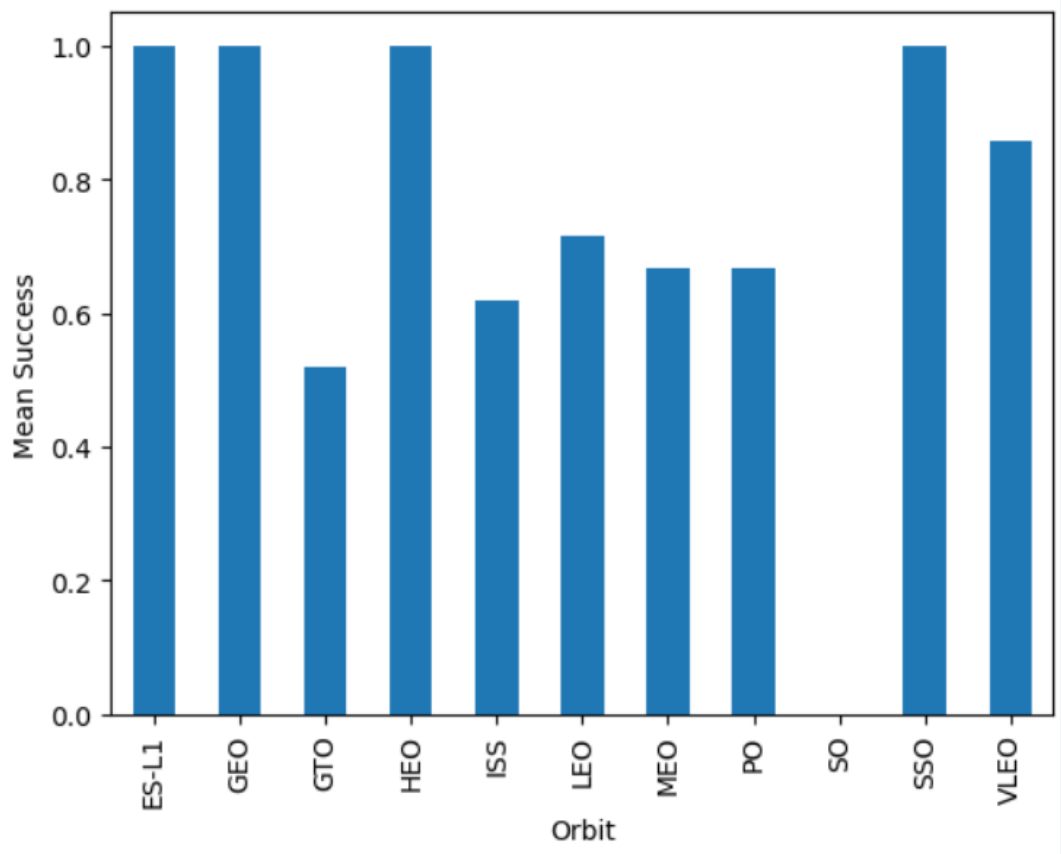
The scatter plot shows that the launches are carried out mainly in CCAFS SLC-40, in the same way it is observed that the success rate increases as the launches have increased, it is important to consider that the blue dots represent successful launches.

Payload vs. Launch Site



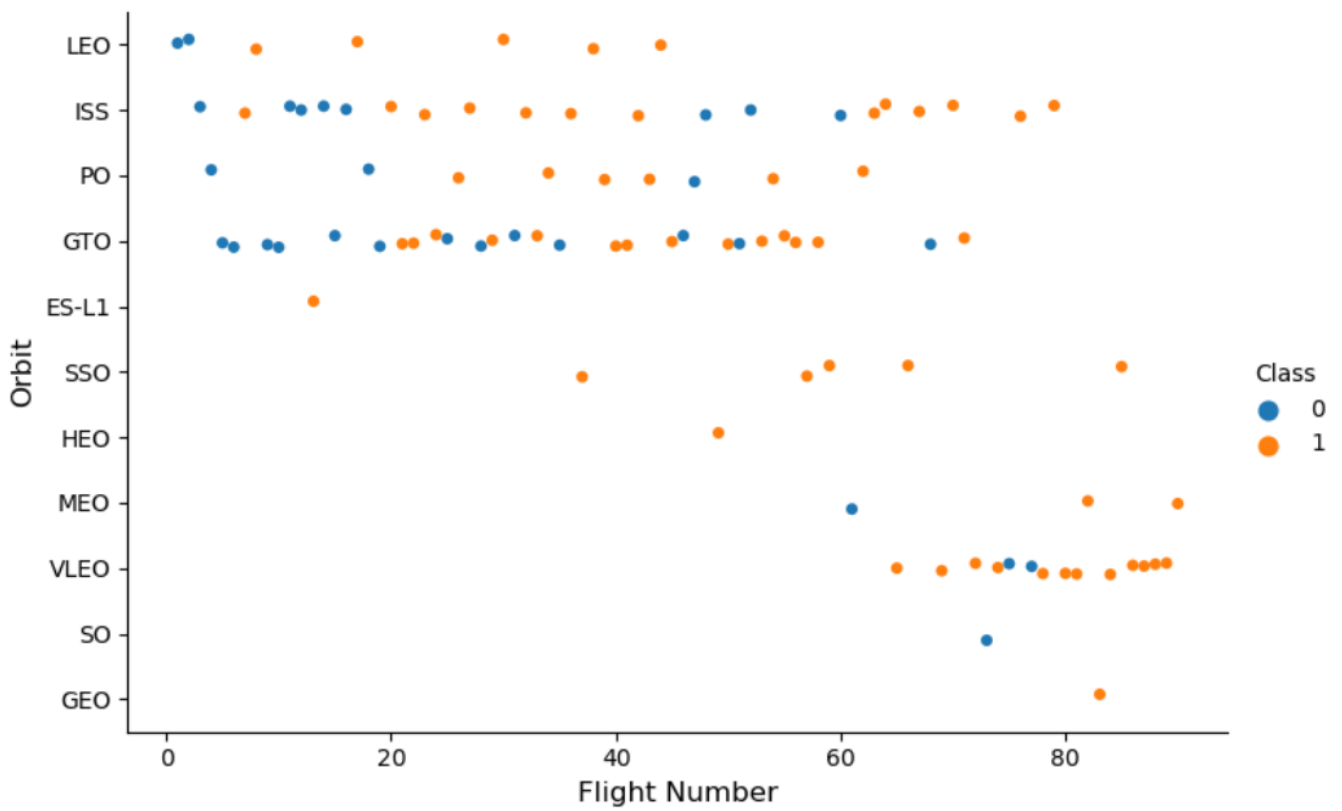
The scatter plot shows that the launch sites with the highest payload mass and that were successful were launched by CCAFS SLC-40 and KSC LC-39A, however, the CCAFS SLC-40 launch site also concentrates the highest number of failed launches with lower payload mass. It is similarly important to note that the blue dots represent successful launches.

Success Rate vs. Orbit Type



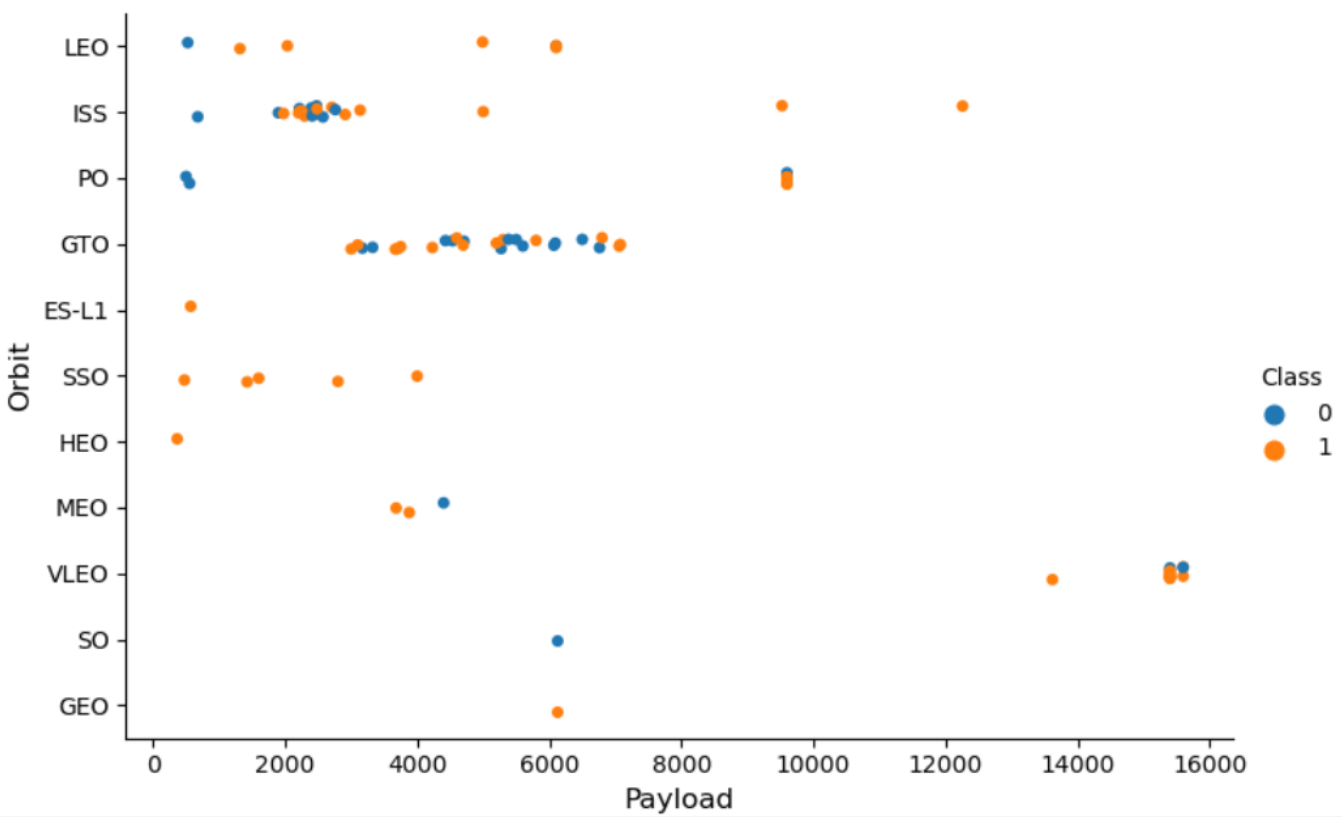
The bar graph shows the launch success rate for each of the orbits, the orbits that show the highest success rate with 1.0 are ES-L1, GEO, HEO and SSO, while the orbit with the lowest success rate is GTO which is below 0.6.

Flight Number vs. Orbit Type



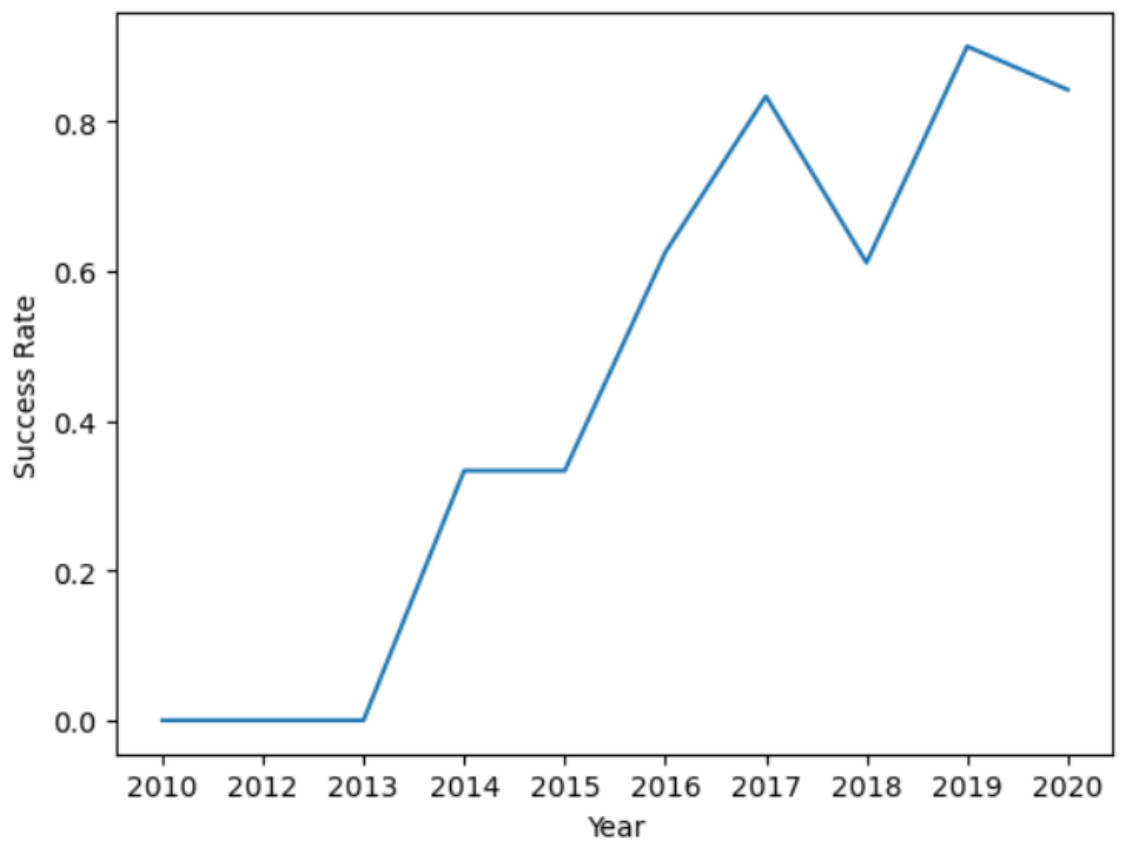
The scatterplot shows that the first launches were carried out mainly to LEO, ISS, PO and GTO orbits, while most of the most recent launches were carried out to VLEO orbit, which has a success rate greater over 0.8.

Payload vs. Orbit Type



The scatter plot shows that the orbit that has the highest amount of payload mass is the VLEO with values close to 16000, however the largest number of launches occurs in the GTO and ISS orbits, whose payload mass is between 1800 and 7000.

Launch Success Yearly Trend



The line graph shows an increasing trend of the success rate of launches between 2013 and 2020.

All Launch Site Names

```
%%sql
   SELECT DISTINCT(LAUNCH_SITE)
       FROM SPACE;
```

The query selects the distinct values with the DISTINCT function applied to the LAUNCH_SITE row of the table named SPACE; the result is the places from where the launches are made.

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

Launch Site Names Begin with 'KSC'

```
%%sql
SELECT DATE, TIME__UTC_, BOOSTER_VERSION,
LAUNCH_SITE
   FROM SPACE
   WHERE LAUNCH_SITE LIKE 'KSC%'
   LIMIT 5;
```

The query takes the DATE, TIME_UTC, BOOSTER_VERSION and LAUNCH_SITE columns from the SPACE table in which the values initially contain the letters 'KSC' and is limited to the first five records.

DATE	time__utc_	booster_version	launch_site
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as "Total
Payload"
   FROM SPACE
   WHERE CUSTOMER = 'NASA (CRS)'
```

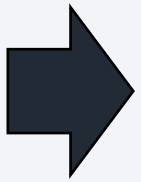
This query adds the values of the PAYLOAD_MASS__KG_ column to the one named "Total Payload" of the SPACE table where the customer is 'NASA (CRS)'.

Total Payload
45596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS
"AVERAGE"
   FROM SPACE
   WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

This query takes the PAYLOAD_MASS__KG_ column from the SPACE table, whose values in the BOOSTER_VERSION column contain the words 'F9 v1.1%' and calculates the average, this result is assigned the name "AVERAGE"

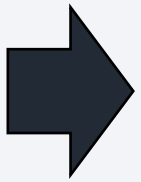


average
2534

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS "First_Success"
   FROM SPACE
   WHERE LANDING__OUTCOME LIKE
'Success%';
```

This query searches into the DATE column of SPACE table for the lowest value or date for this case, in which the value of the column LANDING__OUTCOME begins with the word 'Success'



First_Success
2015-12-22

Successful in ground pad with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION
   FROM SPACE
   WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND
6000
   AND LANDING__OUTCOME = 'Success (ground pad)';
```

This query takes the values of the BOOSTER_VERSION column in the SPACE table which the value in the PAYLOAD_MASS__KG_ column is between 4000 and 6000 and the value of the LANDING__OUTCOME column corresponds to 'Success (ground pad)’

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS
"Number"
   FROM SPACE
   GROUP BY MISSION_OUTCOME;
```

This query counts the rows of the SPACE table, the result is in a column called "Number" and is grouped by the values of the MISSION_OUTCOME column.

mission_outcome	Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT BOOSTER_VERSION,
PAYLOAD_MASS__KG_
   FROM SPACE
   WHERE PAYLOAD_MASS__KG_= (SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACE);
```

The query takes the elements of the BOOSTER_VERSION and PAYLOAD_MASS__KG_ columns of the SPACE table in which the value in the PAYLOAD_MASS__KG_ column is equal to the maximum value of this last column obtained from a nested query.

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

2017 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS "Month",
BOOSTER_VERSION AS "Booster Version", LAUNCH_SITE
AS "Launch Site"
   FROM SPACE
   WHERE LANDING__OUTCOME='Success (ground pad)'
   AND YEAR(DATE)='2017';
```

The query lists the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Month	Booster Version	Launch Site
February	F9 FT B1031.1	KSC LC-39A
May	F9 FT B1032.1	KSC LC-39A
June	F9 FT B1035.1	KSC LC-39A
August	F9 B4 B1039.1	KSC LC-39A
September	F9 B4 B1040.1	KSC LC-39A
December	F9 FT B1035.2	CCAFS SLC-40

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

```
%%sql
SELECT MONTHNAME(DATE) AS "Month",
BOOSTER_VERSION AS "Booster Version", LAUNCH_SITE
AS "Launch Site"
   FROM SPACE
   WHERE LANDING__OUTCOME='Success (ground pad)'
   AND YEAR(DATE)='2017';
```

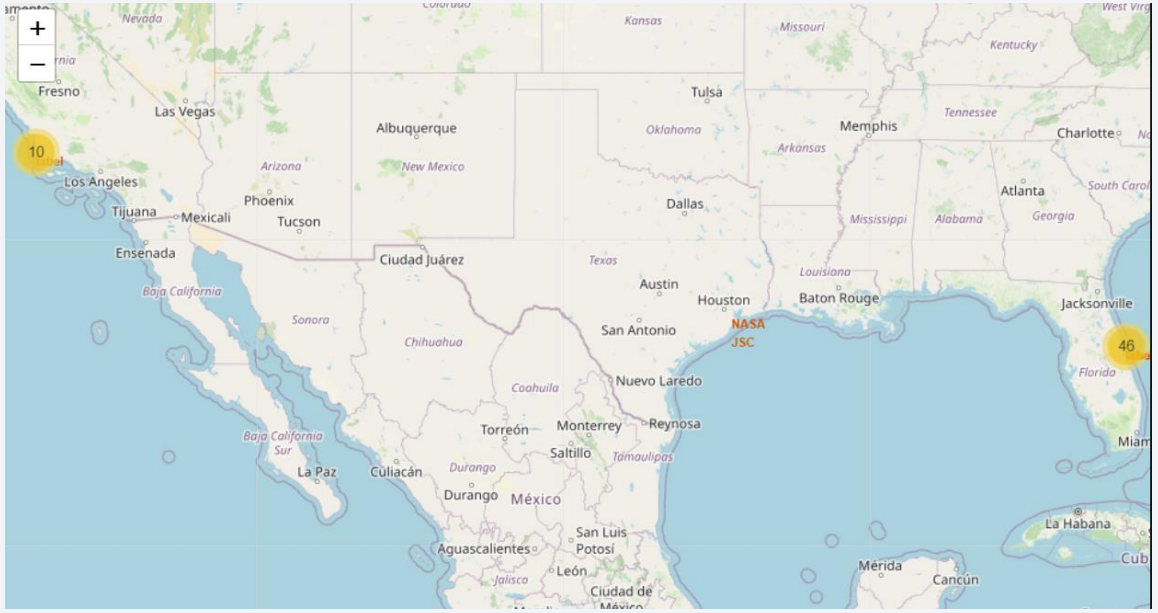
The query lists ranks the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Landing Outcome	Count
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites Locations

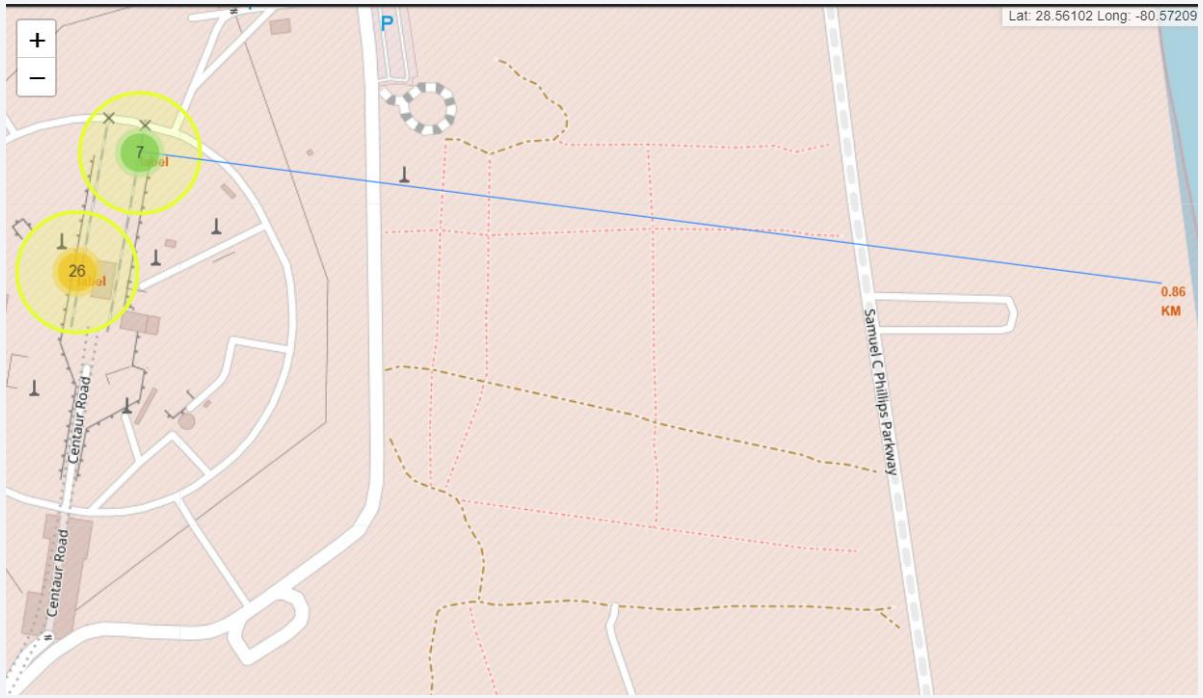


The yellow grouped markers indicate the launch sites of all the Space X missions, it is evident that the largest number are on the East coast with 46 launches compared to 10 on the West coast, in general it can be determined that the launch sites release have been strategically arranged near the coastline.

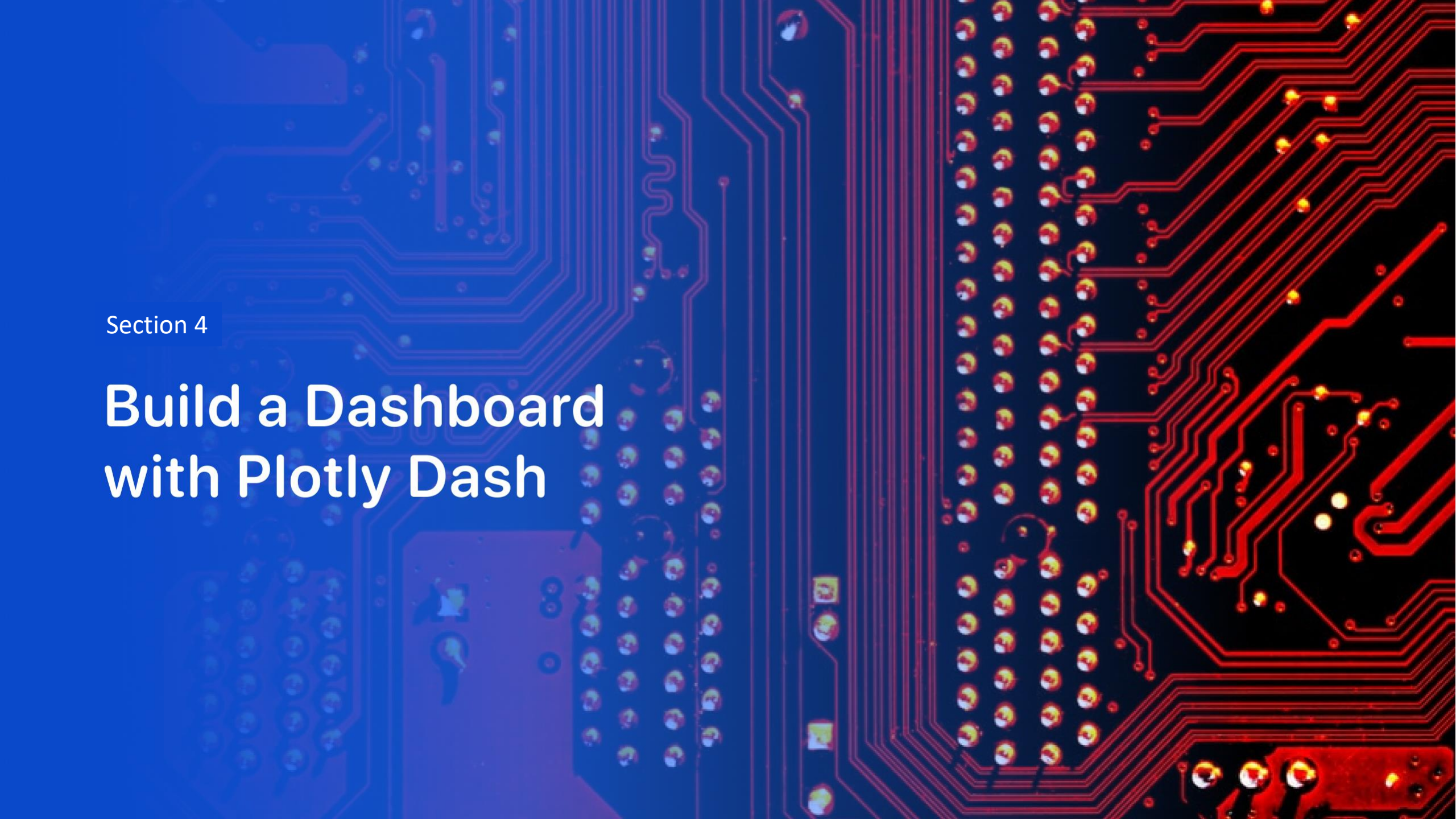
Success or failed launches

By zooming in on the grouped marks it is possible to identify successful launches in green and failures in red.

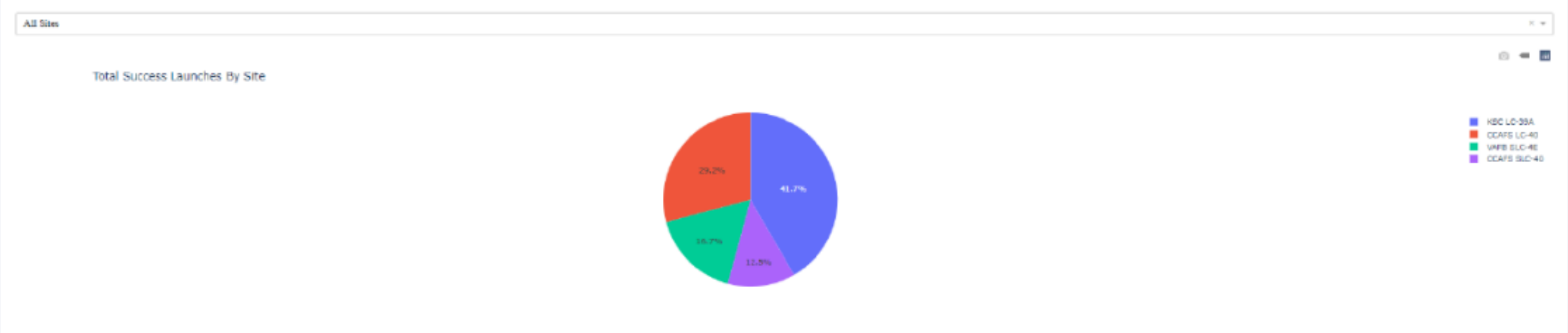
Launch site proximities



The map that was generated shows the points of interest close to the launch sites, in the same way the distance to the coastline are measured and the cities that are adjacent to the launch site are showed.


Section 4
Build a Dashboard with Plotly Dash

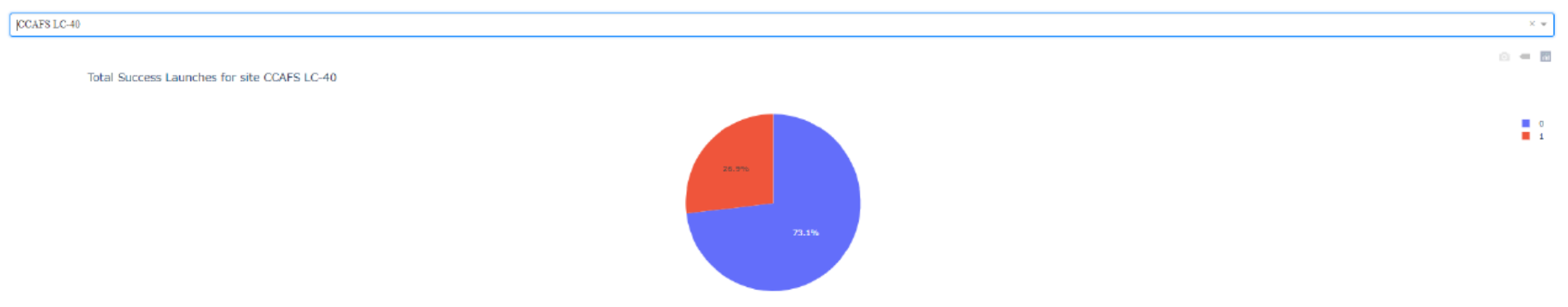
Total Successful Launches By Site



The pie chart shows the total number of successful launches by site, the largest number originating from launch complex 39 (KSC LC-39A) with 41.7% of successful launches.

The graph that has the dynamic feature, presents a drop-down list where you can select the launch site in detail.

Successful or failed launches by site



In this pie chart you can see the details of the failed and successful launches for each of the sites, in this case at the Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40) site, most of them have been successful with a percentage higher than 70%. In the dropdown list you can select the launch sites to find the success or failure rate.

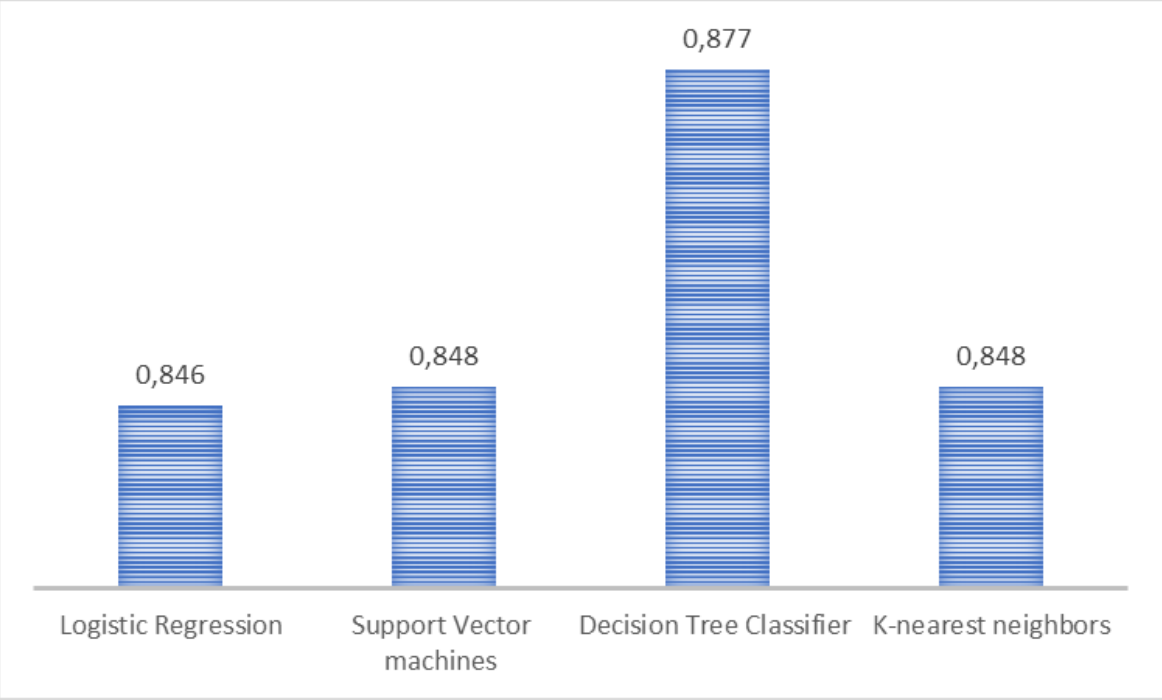
Payloads vs Launch Outcome

In the scatter graph that relates the payload mass in kilograms and the success rate of the different launches, it is evident that The launch success rate for payloads 0-2500 kg is slightly lower than that of payloads 2500-5000 kg. This graph highlights the range slider that allows a dynamic analysis of the ranges that are modified.

Section 5

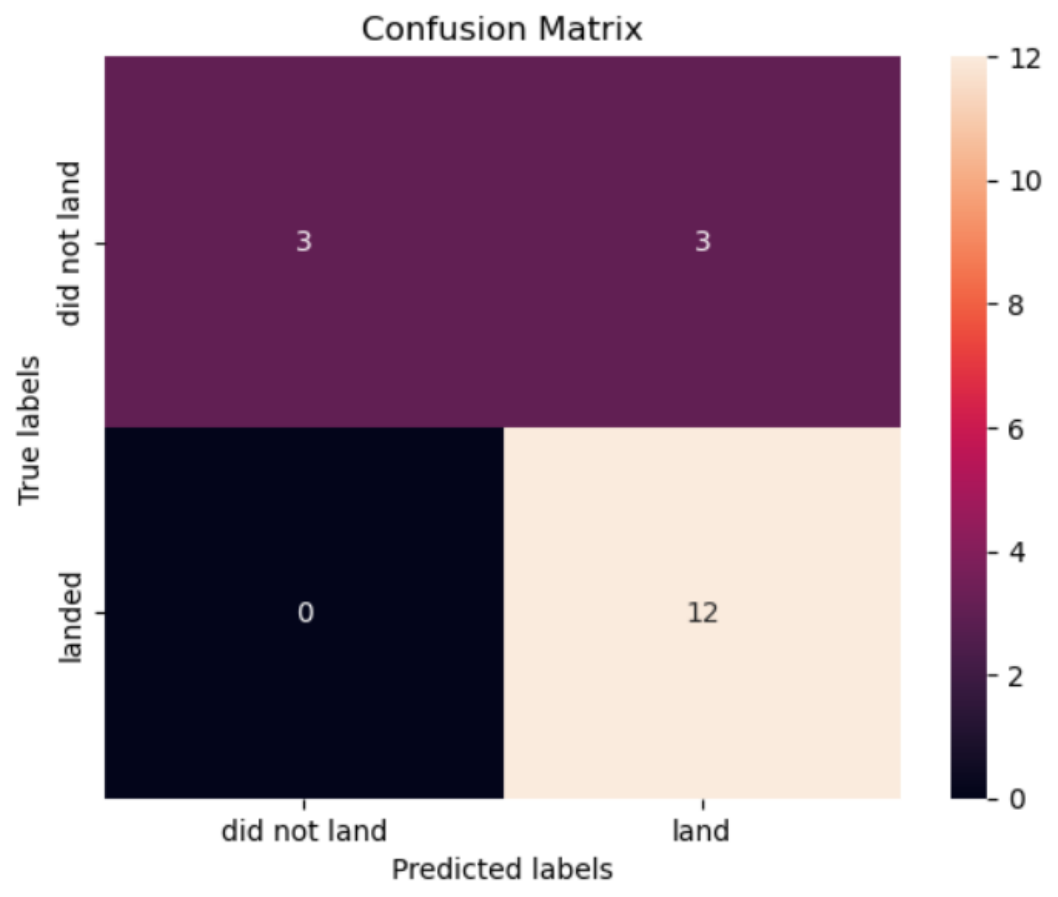
Predictive Analysis (Classification)

Classification Accuracy



It is evident that the model with the highest accuracy is the Decision Tree Classifier, the other three models present a similar behavior.

Confusion Matrix



This model predicted 12 successful landings that did land, 3 landings that should have failed but did land, 3 landings that should not have landed successfully that did not, and none of the launches predicted as unsuccessful that landed correctly.

Conclusions

- It is evident that the launches to the ES-L1, GEO, HEO and SSO orbits present the highest success rate in the launches, while GTO presents the lowest success rate of the orbits to which the launches are made.
- The success rate in the launches comes with a continuous increasing trend since the year 2013.
- The launch sites that present the highest amount of payload mass are CCAFS SLC-40 and KSC LC-39A which are around 16,000 kilograms, however, the CCAFS SLC-40 launch site also concentrates the highest number of failed launches with lower payload mass .
- The launch sites are strategically located near highways and railways for the transport of personnel and cargo, but also far from cities for safety, they are also close to coastlines and are distributed between the East and West coasts.
- The best predictive model to use with this data set is the Decision Tree Classifier, as it had the highest accuracy at 0.877.

Appendix

- Github Repository: https://github.com/mauriciohmrc/CapstonProjectIBM

Thank you!