IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

CESAR CERVANTES JIMENEZ
17 – 05 - 2025

Outline

Executive Summary

- Summary of methodologies

  First, data collection was carried out using API and web scraping. Once the data was collected, different techniques were used for managing this data. With the normalized data and in manageable formats, the data was visualized in order to evaluate it and draw relevant conclusions. And once with this information, the different systems were designed and tested to predict the information.

- Summary of all results

  The best hyperparameters for logistic regression, svm, decision tree and knn classifiers.

Introduction

- Project background and context

  It was determined whether the first stage of SpaceX landed and we assessed the cost of a launch. This information was used to analyze the feasibility of whether another alternative company wants to compete against SpaceX for a rocket launch. In this project, the data was collected, normalized, and predictions were made.

- Problems you want to find answers

  What are the decisive factors for a successful launch?

  What are the chances of completing a full mission?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  For data collection, API and web scraping techniques were used.

- Perform data wrangling

  The data handling was carried out using tools and libraries of the panda language.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  The data was collected, normalized, transformed, evaluated, and then with this information, the groupings and classifications were made and the models were built.

Data Collection

- Describe how data sets were collected.

  Data collection was carried out using two widely used techniques, namely APIs and web scraping, to obtain information from various online sources.

  The API technique allowed access to the structured data provided by a platform.

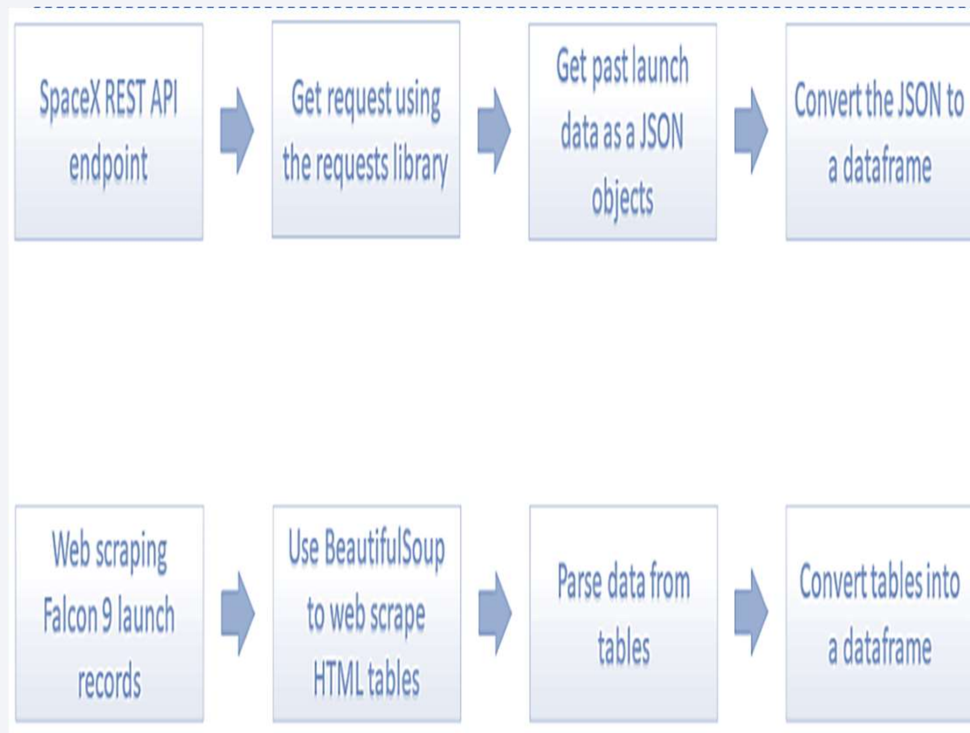
  The web scraping technique involves extracting data directly from web pages.

- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

To make the requested JSON results more consistent, decode the responde content as a Json and turn it into a pandas dataframe

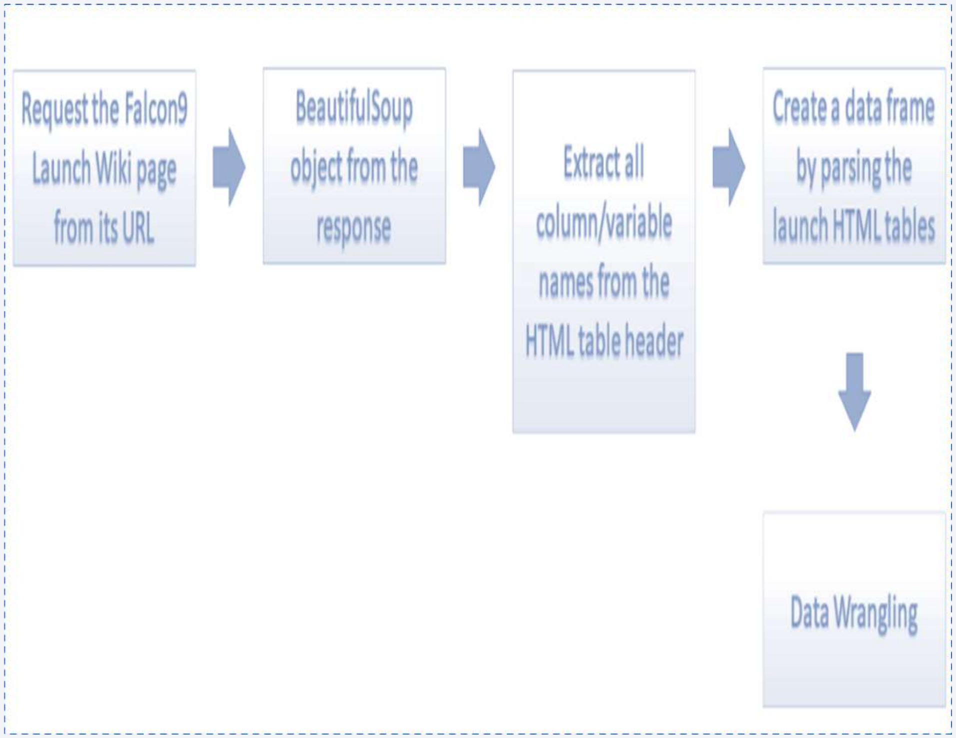
[Castone/1.1 jupyter-labs-spacex-data-collection-api.ipynb at main · blaceta/Castone]



Data Collection - Scraping

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response, create a beautifulsoup object from the html response

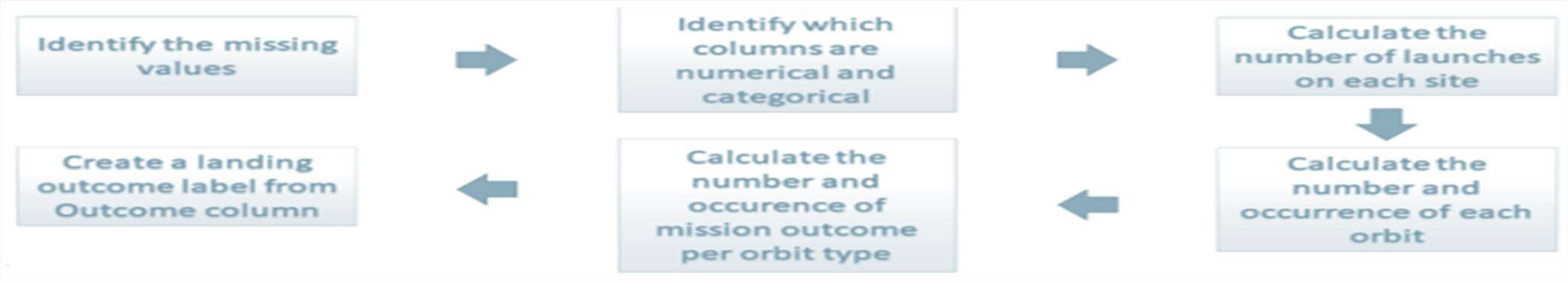
[Castone/1.2 jupyter-labs-webscraping.ipynb at main · blaceta/Castone]()



Data Wrangling

An exploratory data analysis (EDA) was conducted to find some patterns in the data and determine what the label would be to train supervised models.

In the collected dataset, different types of data with very relevant information were found.



[Castone/1.3 labs-jupyter-spacex-Data wrangling.ipynb at main · blaceta/Castone]

EDA with Data Visualization

Catplot: in order to visualize the attempts of failed flights due to the loaded weight to find a match.

Catplot: to graph the relationship between the number of flights and the places where these launches took place, and whether they were successful or not.

Scatterplot: to graph the existing relationship between the load and the places where these launches took place and whether they were successful or not.

Barplot: to visualize the relationship between the success rate of each launch and the type of orbit

Catplot: to visualize visualize the relationship between flightnumber and orbit type

Catplot: to visualize the relationship between the payload mass and the type of orbit

Castone/2.2 edadataviz.ipynb at main · blaceta/Castone

EDA with SQL

- Display the names of the unique launch sites in the space mission:

select DISTINCT LAUNCH_SITE from SPACEXTABLE

- Display 5 records where launch sites begin with the string 'CCA’

select * from SPACEXTABLE where LAUNCH_SITE like 'CCA%' limit 5

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

select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)’

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

select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'

EDA with SQL

- List the date when the first succesful landing outcome in ground pad was acheived.

select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

SELECT booster_version, payload_mass__kg_, landing_outcome FROM SPACEXTBL WHERE landing_outcome='Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)

- List the total number of successful and failure mission outcomes

SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome

- List all the booster_versions that have carried the maximum payload mass. Use a subquery.

SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass__kg_) AS "maximum_payload_mass" FROM SPACEXTBL) FROM SPACEXTBL;

EDA with SQL

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

SELECT landing_outcome, booster_version, launch_site, DATE FROM SPACEXTBL WHERE landing_outcome LIKE '%Failure (drone ship)%' AND (DATE LIKE '2015%') ;

- Rank 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.

SELECT DISTINCT(landing_outcome) FROM SPACEXTBL; SELECT landing_outcome, COUNT(landing_outcome) AS "total" FROM SPACEXTBL WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-20') GROUP BY landing_outcome ORDER BY "total" DESC;

Castone/2.1 jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · blaceta/Castone

Build an Interactive Map with Folium

- Use folium.Circle to add a highlighted circle area with a text label on a specific coordinate.
- Create and add folium.Circle and folium.Marker for each launch site on the site map
- For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster
- Draw a PolyLine between a launch site to the selected coastline point
- draw a line betwee a launch site to its closest city, railway, highway, etc. You need to use MousePosition to find the their coordinates on the map first

Castone/3.1 lab_jupyter_launch_site_location.ipynb at main · blaceta/Castone

Build a Dashboard with Plotly Dash

An application was built with Plotly Dash to perform an interactive visual analysis of SpaceX launch data in real time.This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter plot.

With the Dash callback function in Python to update each time the input component is updated, such as a click or a selection event in the dropdown menu.

Then a scatter plot was created with the x-axis as the payload and the y-axis as the launch result. In this way, it was possible to visually observe how the payload may be correlated with the mission results for the selected site(s).

Castone/3.2 spacex-dash-app.py at main · blaceta/Castone

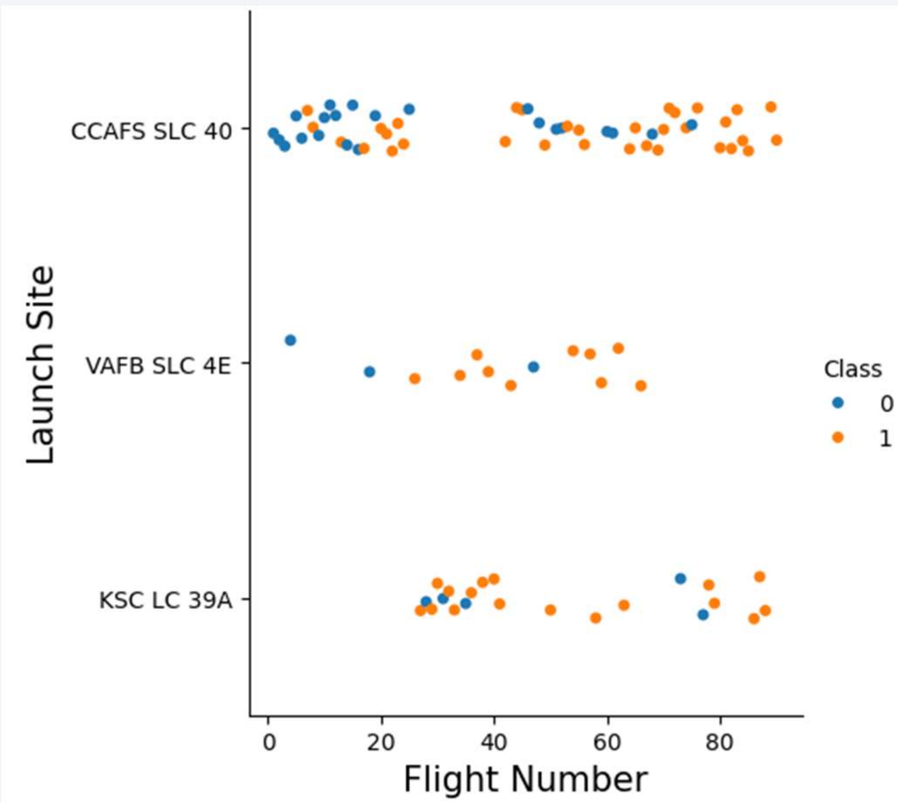
Predictive Analysis (Classification)

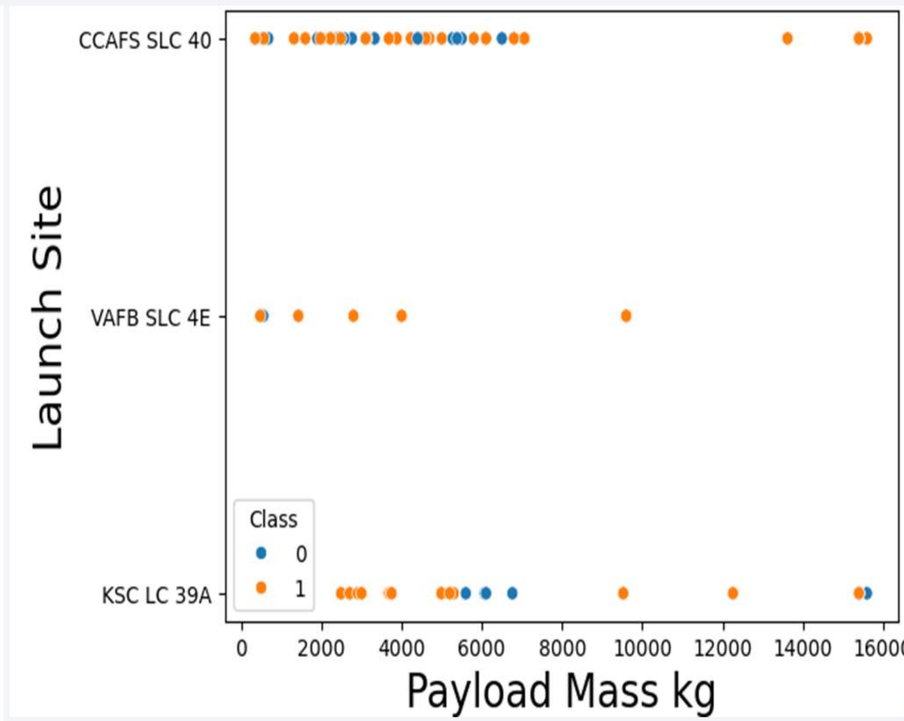
- Create a NumPy array from the column Class in data, by applying the method to_numpy()
- Use the function train_test_split to split the data X and Y into training and test data. Set the parameter test_size to 0.2 and random_state to 2.
- Create a logistic regression object then create a GridSearchCV object logreg_cv with cv = 10. Fit the object to find the best parameters from the dictionary parameters.
- Lets look at the confusion matrix.
- Create a support vector machine object then create a GridSearchCV object svm_cv with cv = 10. Fit the object to find the best parameters from the dictionary parameters.
- Create a decision tree classifier object then create a GridSearchCV object tree_cv with cv = 10. Fit the object to find the best parameters from the dictionary parameters.
- Create a k nearest neighbors object then create a GridSearchCV object knn_cv with cv = 10. Fit the object to find the best parameters from the dictionary parameters.

Castone/4.1 SpaceX_Machine Learning Prediction_Part_5.ipynb at main · blaceta/Castone

Results

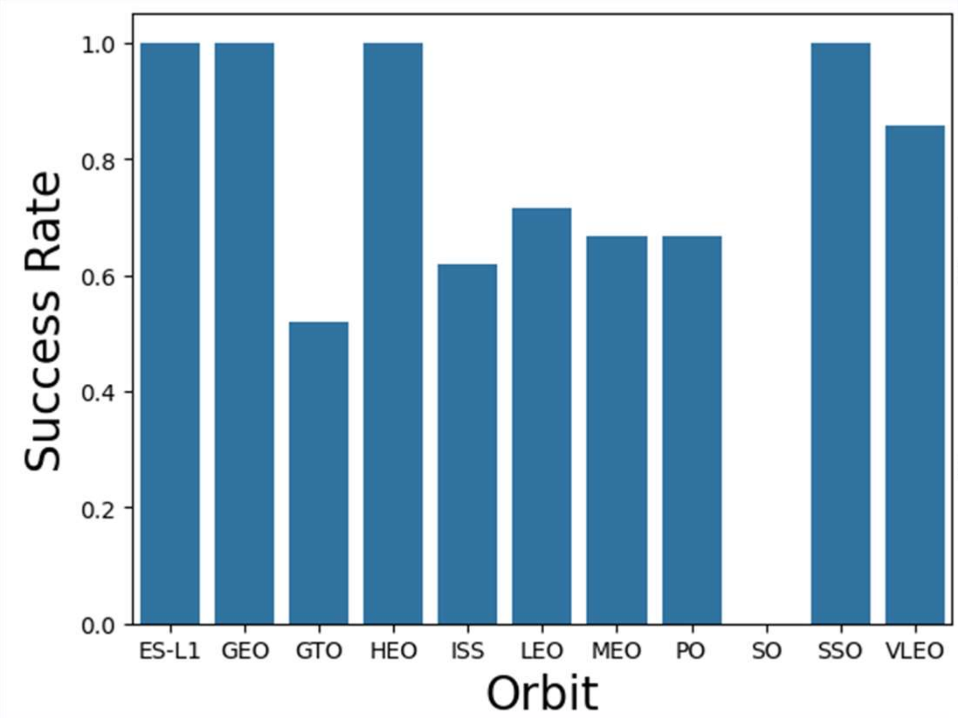
- Exploratory data analysis results

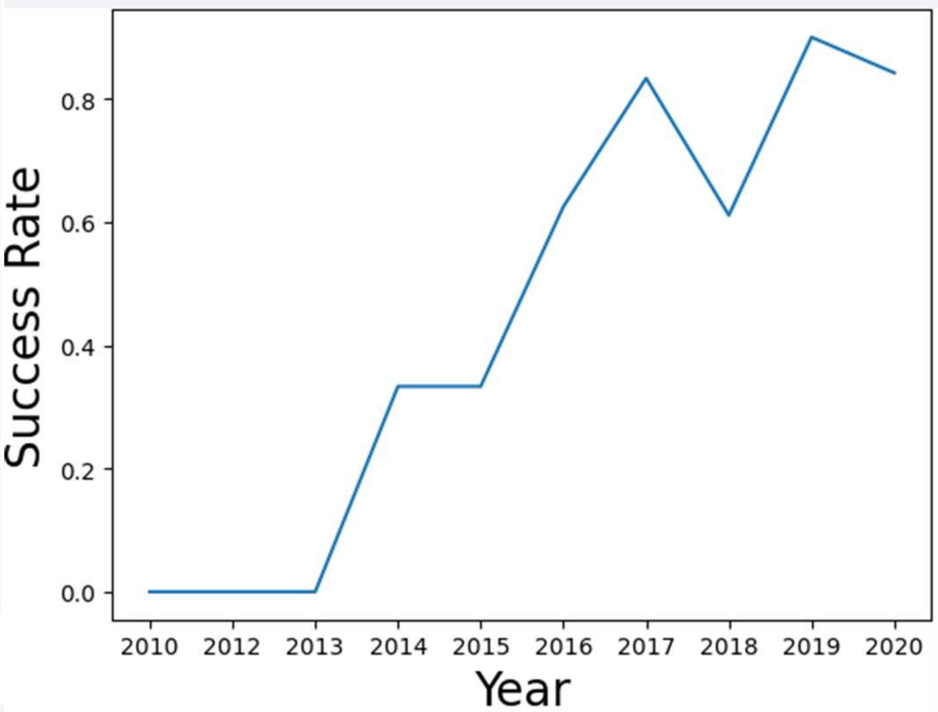




Results

- Exploratory data analysis results



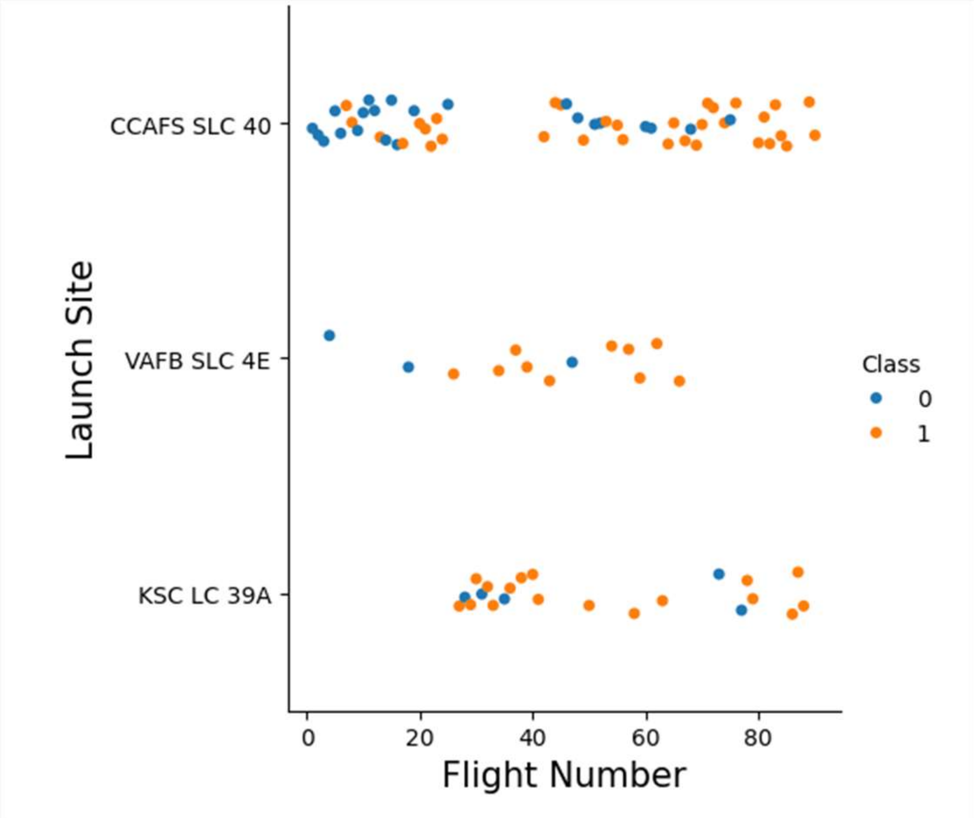


Section 2

Insights drawn from EDA

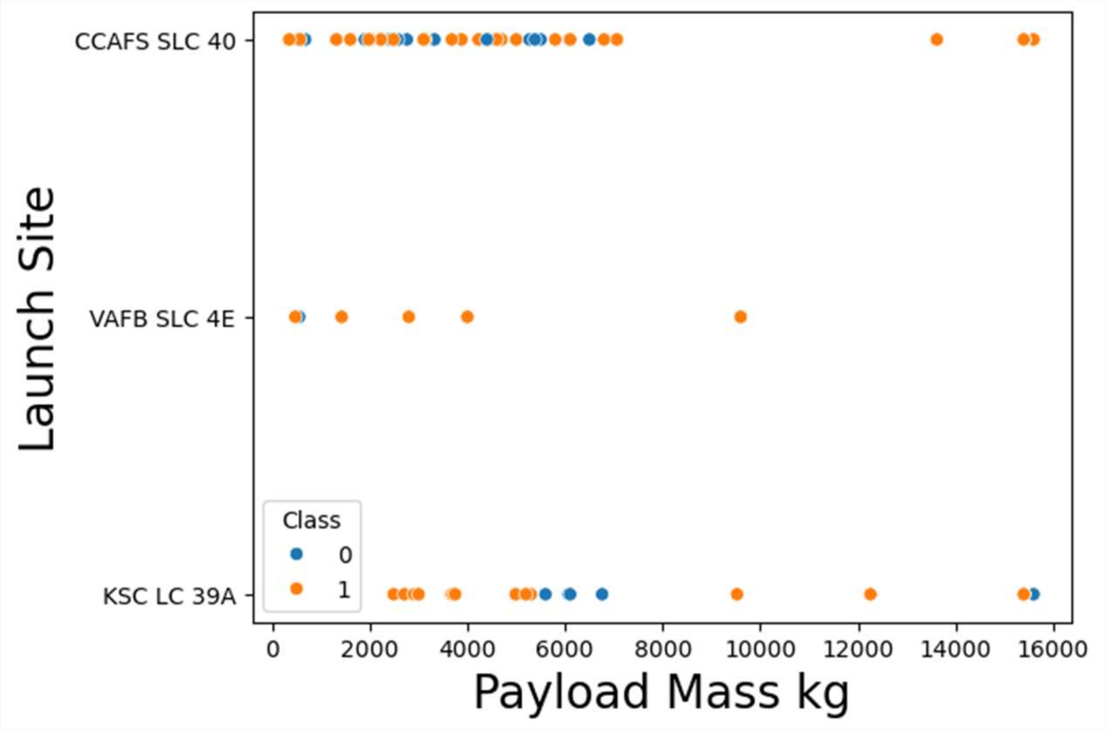
Flight Number vs. Launch Site

In this graph, you can see the number of launches that took place at their respective launch sites, where you can observe those that were successful and those that had problems.



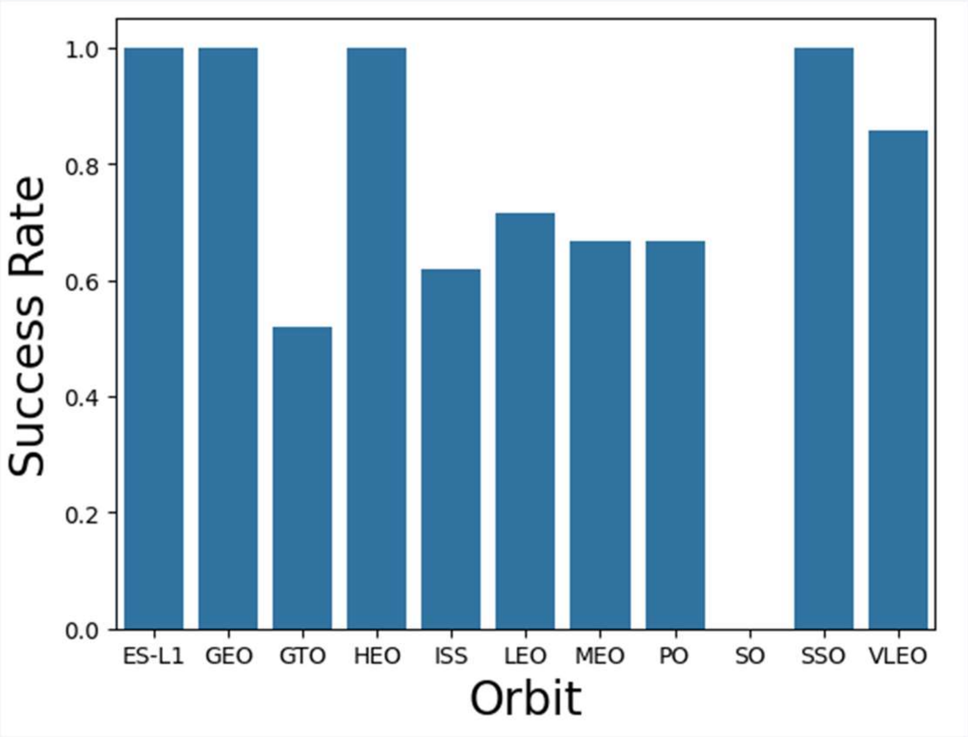
Payload vs. Launch Site

In this graph, you can observe the relationship between the total mass of the launch and the location where the launch took place, as well as whether it was successful or had any issues.



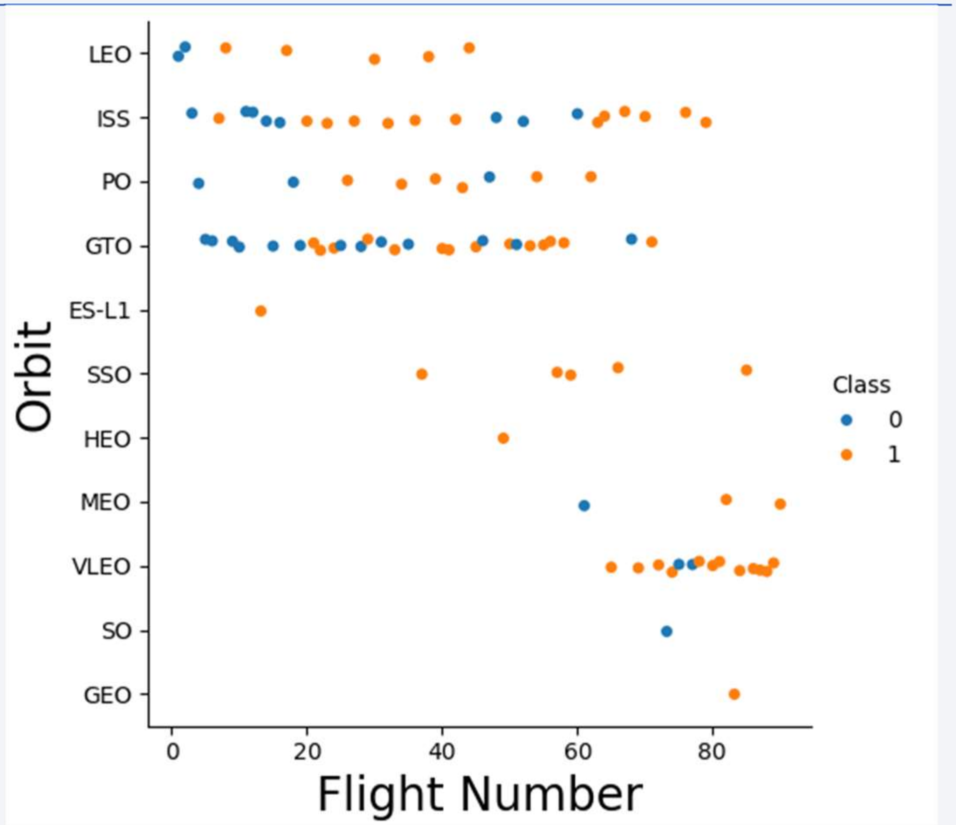
Success Rate vs. Orbit Type

This diagram illustrates the relationship between the success rate and the orbit of the mission. Success was only not achieved in the SO orbit; however, in all other orbits, the success percentage is favorable.



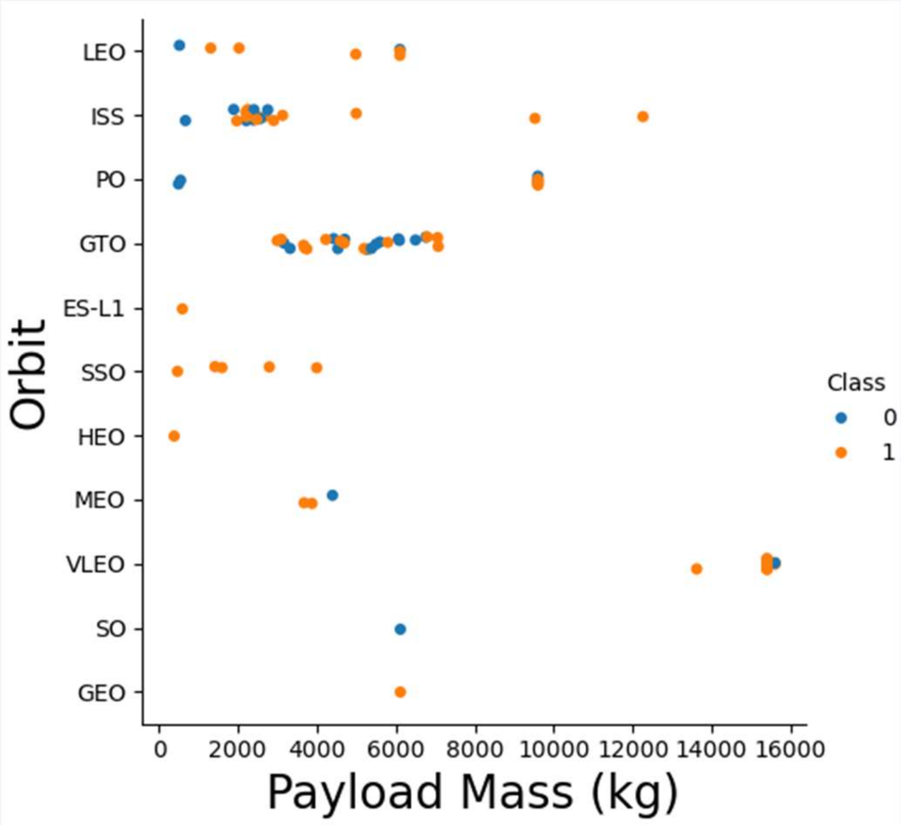
Flight Number vs. Orbit Type

In this graph, we can observe the relationship between the number of flights undertaken and the orbits to which the mission was directed. It can also be evidenced that the number of successful cases is greater, and the SO orbit is the one that does not show success.



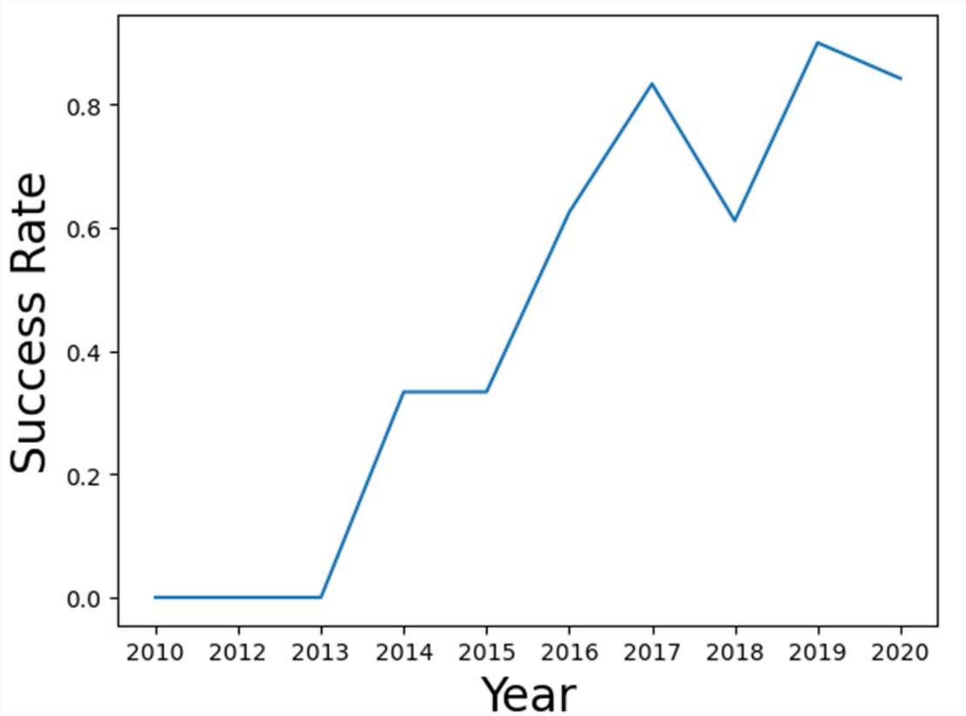
Payload vs. Orbit Type

In the following graph, the relationship between the useful mass of the aircraft and the orbits it was directed to, as well as the cases of successes presented, can be observed.



Launch Success Yearly Trend

The following graph shows the relationship between the success rate of launches and the years they were launched. As can be seen, as experience was gained and technology evolved, the success rate increased.



All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTABLE
```

A consultation was conducted to find the unique locations where the launches took place, four were found, and multiple launches have been conducted from these.

	Launch Site
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where LAUNCH_SITE like 'CCA%' limit 5
```

A query was made to find the five records where the launch sites start with the string 'CCA'. This was done in order to be able to use this information later.

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

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

sum
45596

A consultation was conducted to find the total mass of the payload carried by the rockets launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

Average
2534.6666666666665

A survey was conducted to evaluate the average payload mass carried by the F9 v1.1 rocket version.

First Successful Ground Landing Date

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

Date
2010-06-04

A consultation was carried out to find the date when the first successful landing result was achieved on the terrestrial platform.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT booster_version, payload_mass__kg_, landing_outcome FROM SPACEXTBL \
       WHERE landing_outcome='Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

Booster_Version	PAYLOAD_MASS_KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

The inquiry was made to find the names of the rockets that are successful in the drone ship and if the payload mass was between the range of 4000 and 6000, and four results were found.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

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

The total number of results of successful and failed missions was consulted, where it was evident that the majority were successful and a few had failures.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass__kg_) AS "maximum_payload_mass" FROM SPACEXTBL) FROM SPACEXTBL
```

booster_version	maximum_payload_mass
F9 B4 B1039.2	15600
F9 B4 B1040.2	15600
F9 B4 B1041.2	15600
F9 B4 B1043.2	15600
F9 B4 B1039.1	15600

It was consulted about all the booster versions that have carried the maximum payload mass, and five different ones were found.

2015 Launch Records

```
%sql SELECT landing_outcome, booster_version, launch_site, DATE FROM SPACEXTBL WHERE landing_outcome LIKE '%Failure (drone
```

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

The records were consulted that will show the names of the months, the results of failures in the landing of the drone-ship, the versions of the launcher, and the launch site for the months of the year 2015. It was found that this event has occurred twice.

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

```
%sql SELECT DISTINCT(landing_outcome) FROM SPACEXTBL;
%sql SELECT landing_outcome, COUNT(landing_outcome) AS "total" FROM SPACEXTBL WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-
```

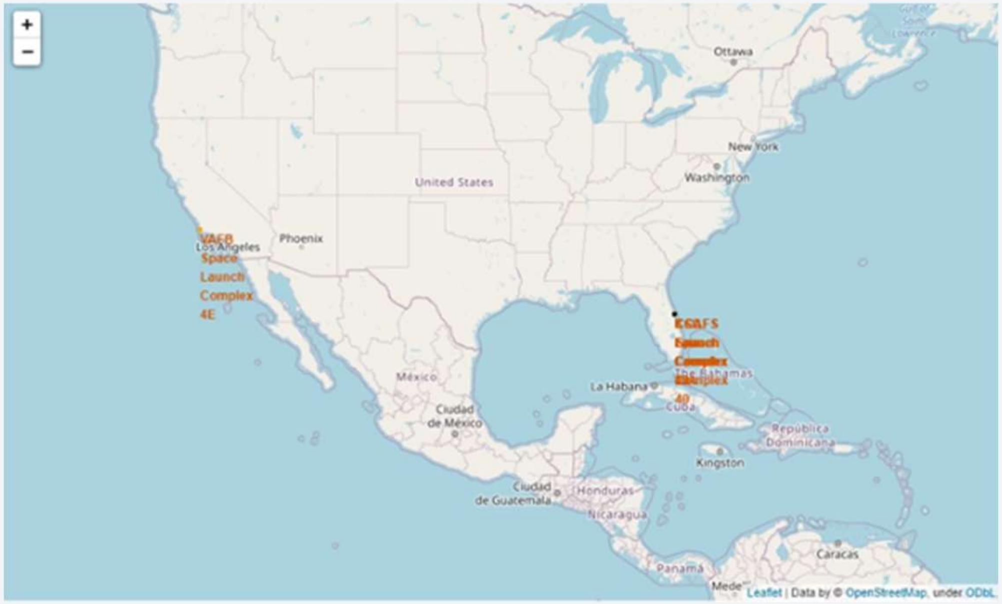
The number of landing results was classified (as Failure (drone ship) or Success (ground platform)) between the dates June 4, 2010, and March 20, 2017, in descending order.

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

Section 3

Launch Sites Proximities Analysis

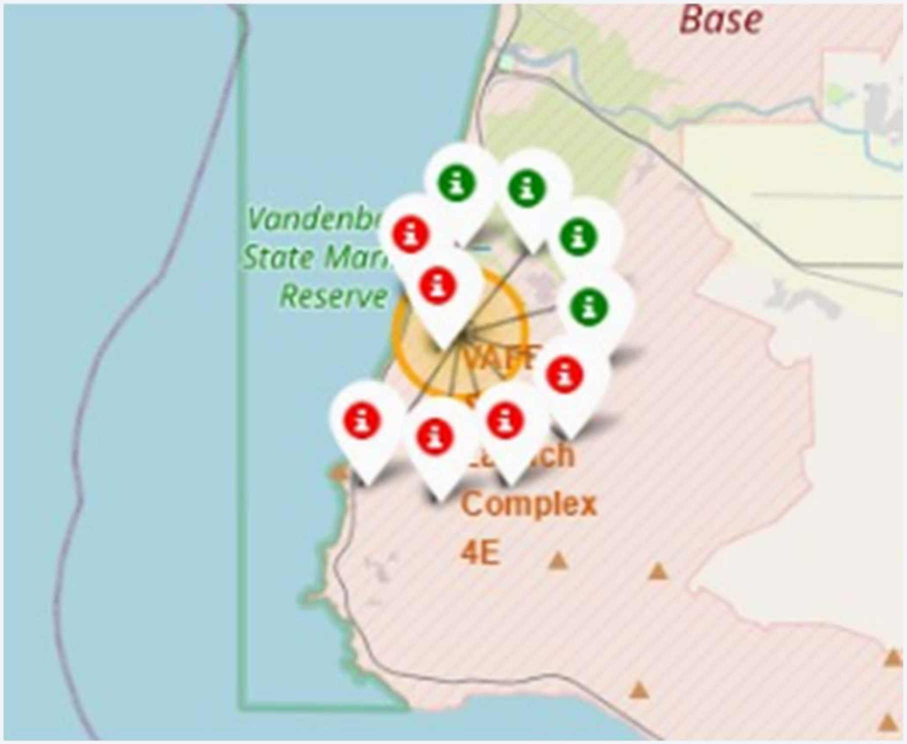
Launch Site on the map



All the sites where the launches have taken place are near the ocean and over restricted areas.

Location of launch sites

The green indicators represent successful launches, and the red indicators represent those that failed. Based on this information alone, we could conclude that there were more failures than successes.



Landmarks to launch sites

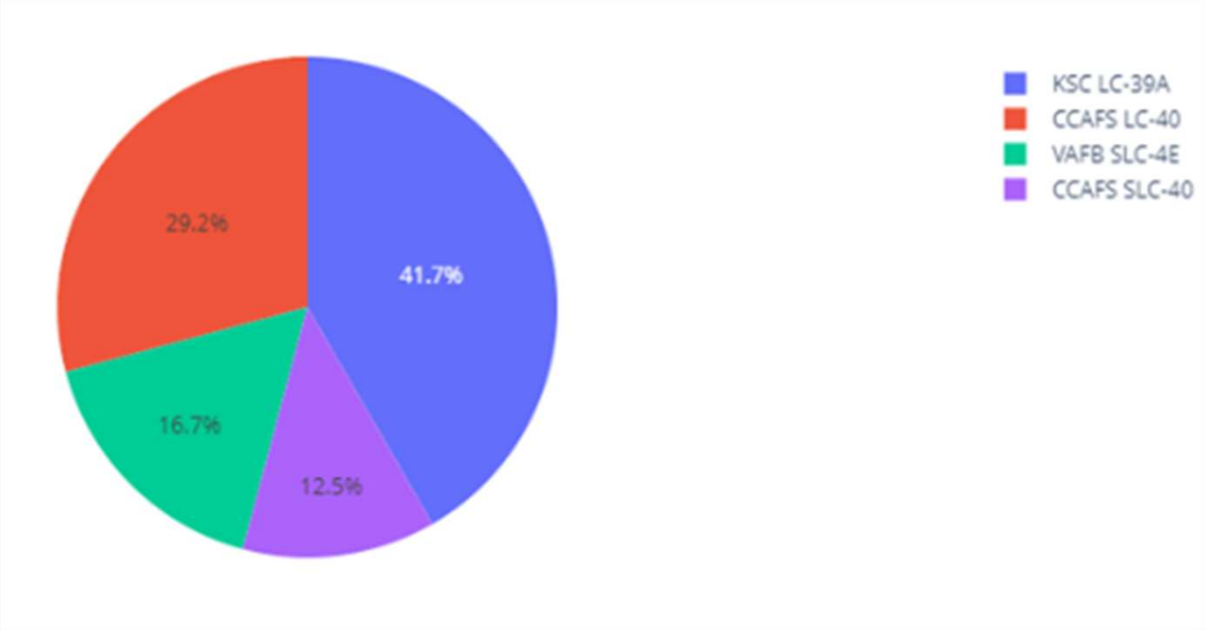
The distances between a launch point and its surroundings were calculated and no striking sites were found nearby because these stations are in hidden locations.

Section 4

Build a Dashboard with Plotly Dash

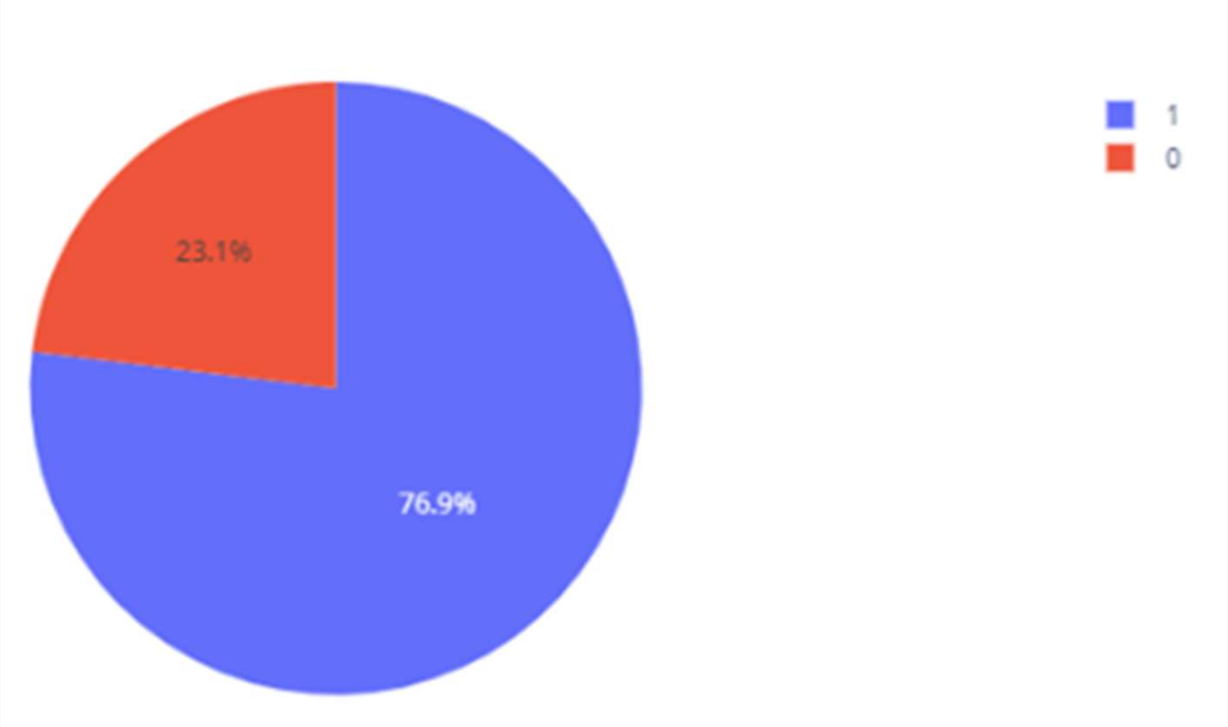
Distribution of site launches

In the following graph you can see the successful distribution of the launches according to the site where the launch was carried out, where the KSC LC-39A station dominates them all.

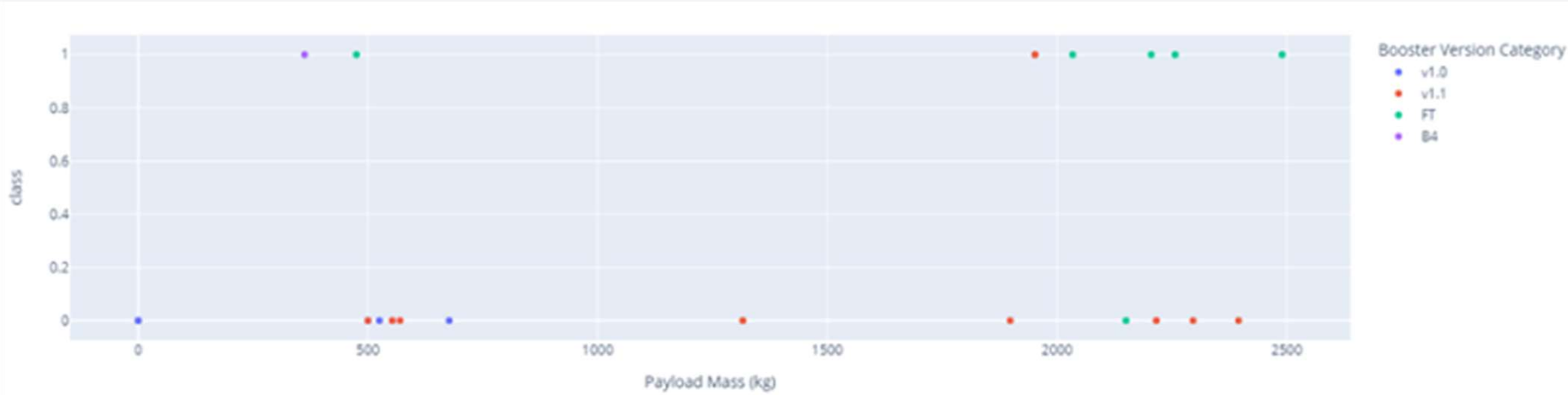


Distribution of KSC LC - 39A

In the previous graph it was determined that the KSC LC 39A station is the most used, therefore in this graph the success rate of the launches at this station is studied where almost 80% of the launches carried out at this station were successful.



Success rate correlation



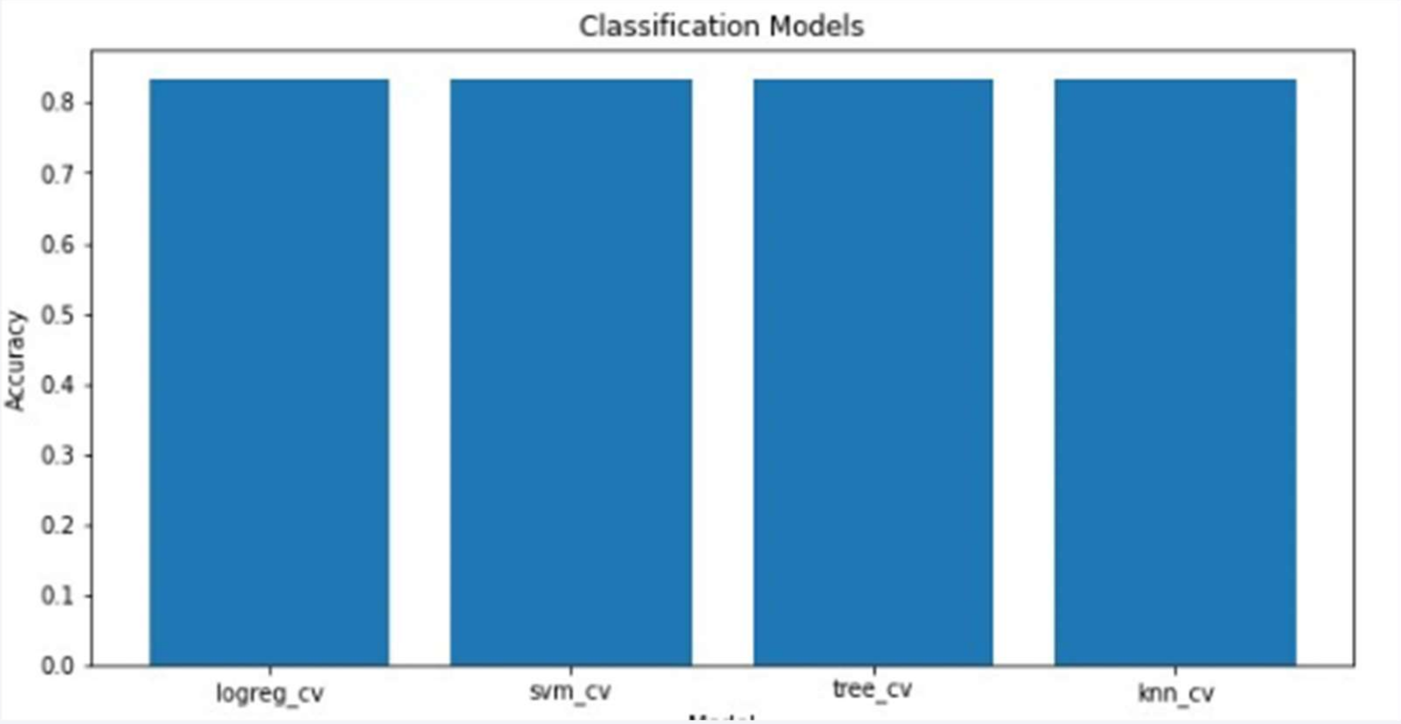
In this graph you can see the correlation between payload and success rate.

Section 5

Predictive Analysis (Classification)

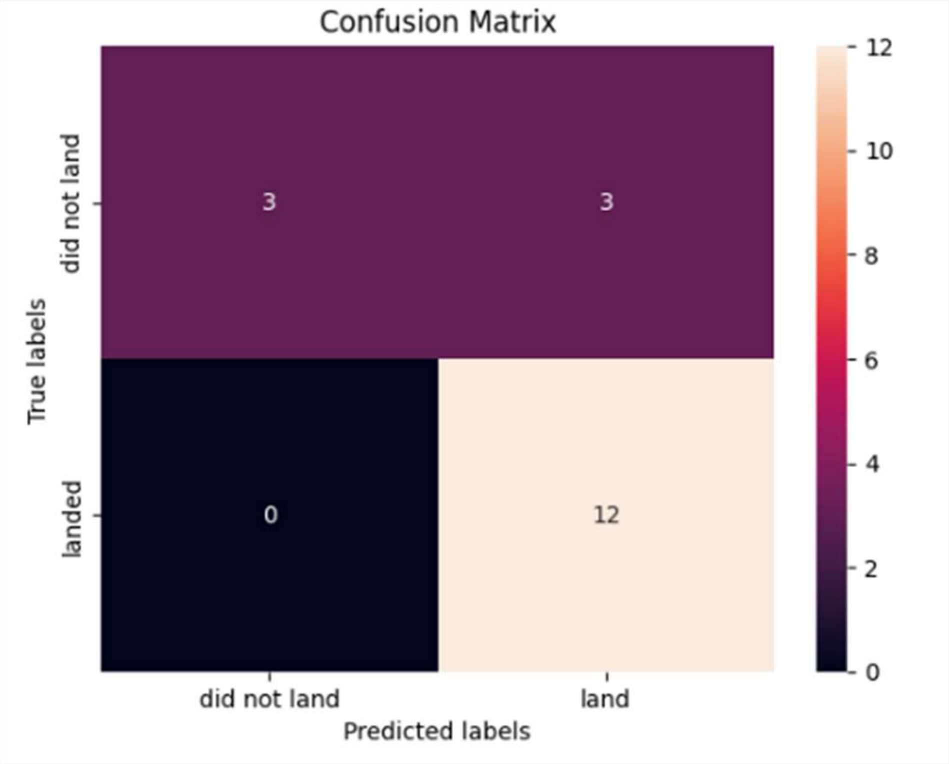
Classification Accuracy

Using the same data and tests performed, it was found that the accuracy is the same.



Confusion Matrix

All models presented the same matrix, which shows that the evidence in the previous graph is relevant.



Conclusions

After conducting this research, the following conclusions were reached:

With the techniques and tools learned, we were able to extract the data and obtain the information needed to reach conclusions about launch success rates and which stations experienced the most failures.

The evaluation also studied the correlation between payload and launch success rates.

Finally, evaluations of different predictive models were conducted, showing that all models presented the same accuracy with the same data.

Appendix

The notebook where you can see the entire repository is located at

[blaceta/Castone]


Thank you!