IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noelia Olivera
12/5/2022

Outline

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

Executive Summary

- In this project we worked with different data extraction methods in order to be able to analyze them later. Different analysis procedures were carried out in order to understand the situation and the relationship between the successes in launching the rockets and the factors involved in it. The primary analysis was performed by EDA with visualization of the relationship between factors using graphs and by data filtering with SQL. Folium was used to analyze the geographical situation and focus on launches and Plotly Dash as a way of visualizing and presenting results. Finally, classification models were generated to be able to predict future results in launches based on the data presented.

Many of the possible causes of ship launch failures were analyzed. Weaknesses in the choice of some orbits and the need to consider the payload mass in the choice of these and the launch booster were identified. The evolution in the accuracy of the launches was observed, which allows us to see the importance of technological advances and knowledge about the problem raised.

Introduction

Currently, the innumerable scientific advances make space travel and missions an everyday thing, recognized companies are dedicated to this, including SpaceX. However, economic inputs will be required, so in case of failures they would represent a loss. SpaceX performs low-cost rocket launches based on the reuse of the first launch stage.

In this project, the information corresponding to previous rocket launches was analyzed to try to understand what caused the failures and therefore how to prevent them. To meet the objective, it was necessary to exhaustively analyze each of the factors involved.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data was extracted by two means: by making a request to the Space X API and through Web Scraping by collecting Falcon 9 launch history records from a Wikipedia page.

- Perform data wrangling:

Information processing consisted of extracting factors of interest, adjusting data formats such as the date, and filling in empty data.

- Perform exploratory data analysis (EDA) using visualization and SQL:

- Catplots were made comparing factors such as payload mass, type of orbit, etc. The temporal evolution of the success rate and the success rate for different orbits was also seen. SQL allowed knowing details such as the date of the first successful landing outcome in ground pad, successful boosters for large mass payloads, etc.

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash

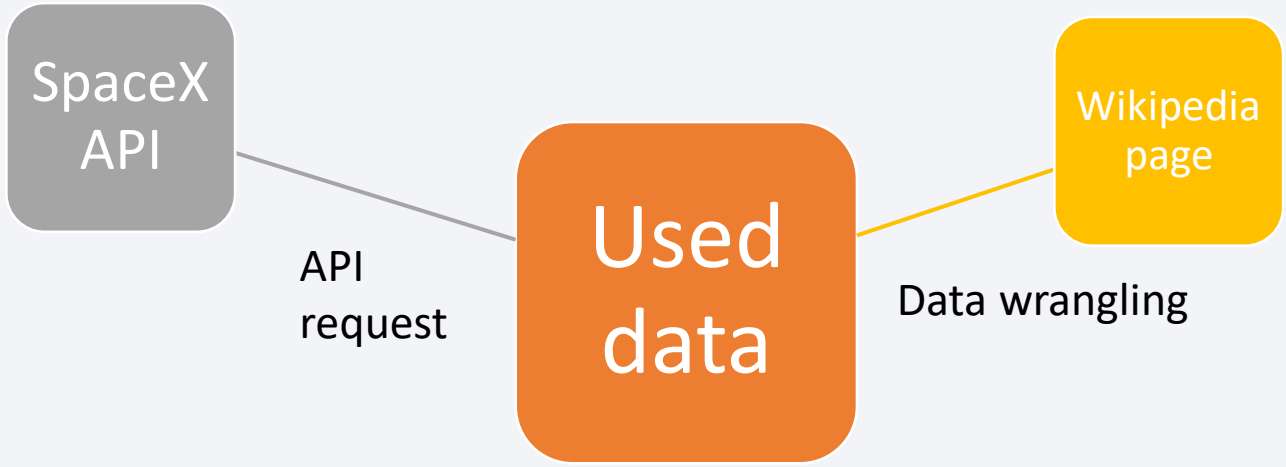
Using Folium, a geographical analysis of the problem was carried out. The location of the launch points was visualized, identifying the successful and unsuccessful launches of each one. The distance to possible elements of interest such as railways, coasts, etc for some centers. With Plotly, interactive content was created that allowed comparing success launches for different types, successes and failures for each place, and the correlation between payload mass and success.

- Perform predictive analysis using classification models

Parameters were determined to be used in different classification models and thus predict the results of possible launches. The models used were K Nearest Neighbors, Support Vector Machine, Decision Tree and Logistic Regression.

Data Collection

- To collect the data, two sources of information and two different methods were used. On the one hand, data was obtained through requests to the SpaceX API and on the other, data wrangling was used on the corresponding Wikipedia page to bring the corresponding information from a table on that page.



Data Collection – SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage
response = requests.get(static_json_url)
```

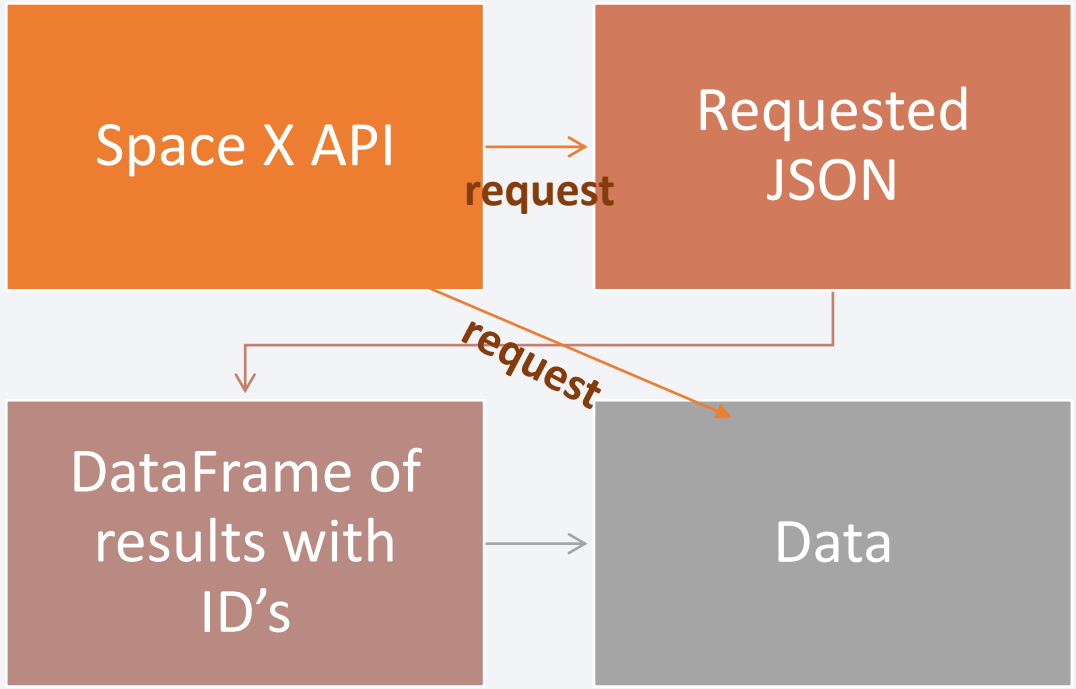
```
# Lets take a subset of our dataframe keeping only the features we
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_n

# We will remove rows with multiple cores because those are falcon
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extrac
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```



As a reference and for peer-review purpose I include my notebook with the process: Here

Data Collection - Scraping

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
Falcon9_HTML_page=requests.get(static_url).content

# assign the response to a object
```

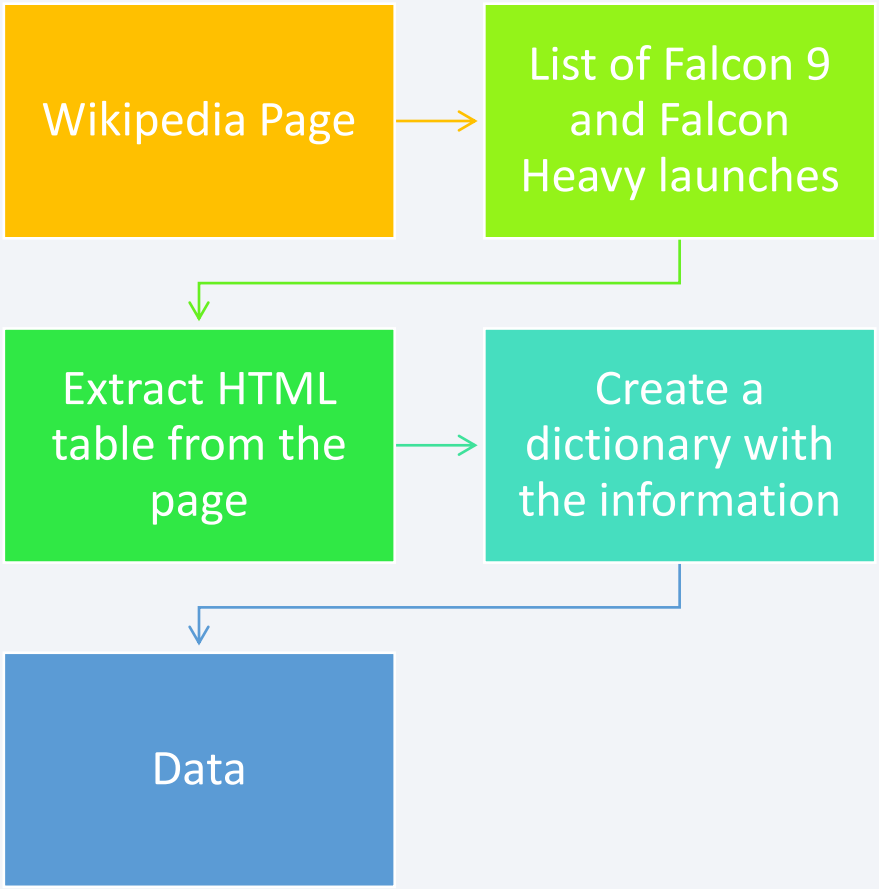
Create a `BeautifulSoup` object from the HTML `response`

```
soup=BeautifulSoup(Falcon9_HTML_page)
```

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```



As a reference and for peer-review purpose I include my notebook with the process: Here

Data Wrangling

- Data was collected using the Space X API, extracted interest factors such as Booster Version, Payload Mass, Orbit, Launch Site, Date, etc. For the date, which was one of the factors involved, it had to be converted from the utc type to the datatime type, in turn the data was restricted to those prior to 11/13/2020. Rows with multiple cores and with multiple payloads on a single rocket were also eliminated. The data was filtered by choosing only data from the Falcon 9 spacecraft, and gaps in factors such as payloads were replaced with averages.

EDA with Data Visualization

Catplots were generated with different data: FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, LaunchSite vs. PayloadMass, Orbit vs. FlightNumber and PayloadMass vs Orbit, visualizing the relationship between both variables and the relationship of each one with the success of the launch. The success rate depending on the type of orbit and on the other hand the success rate depending on the year of launch were also analyzed. These allowed to cover a large number of possibilities or causes in launch failures.

As a reference and for peer-review purpose I include my notebook with the process:
Here

EDA with SQL

The queries made were:

- Names of the launch sites in the space mission
- Selection of information for launch sites begin with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when first succesfull landing outcome in ground pad was achieved
- Names of booster which have success in drone ship with payload in (4000-6000)Kg
- Total number of successful and failure missions outcomes
- Names of booster versions which have carried the maximum payload mass
- Month, booster version and launch site for failure landing outcomes in drone ship in 2015
- Number of successful landing outcomes between 4/6/2010 and 20/3/2017

As a reference and for peer-review purpose I include my notebook with the process: Here

Build an Interactive Map with Folium

Using the coordinates of the different launch sites, they were added to the map, 5 different centers were located, including NASA Johnson Space Center, orange circles were used as icons and name labels were added. Successes and failures were also marked for each site, using green markers for successes and red markers for failures. The distances between different positions, such as between sites and the coastal zone, were calculated and marked with lines.

As a reference and for peer-review purpose I include my notebook with the process:
Here

Build a Dashboard with Plotly Dash

Interactive pie charts and catplots were included. In the application it was allowed to select a specific center to study or all together. The pie chart showed the ratio of successes to failures for the indicated site, or the ratio of successes across all sites.In the catplots, the successes and failures for the sites were studied based on the payload mass, allowing to indicate a desired range for this last variable.

As a reference and for peer-review purpose I include my notebook with the process: Here

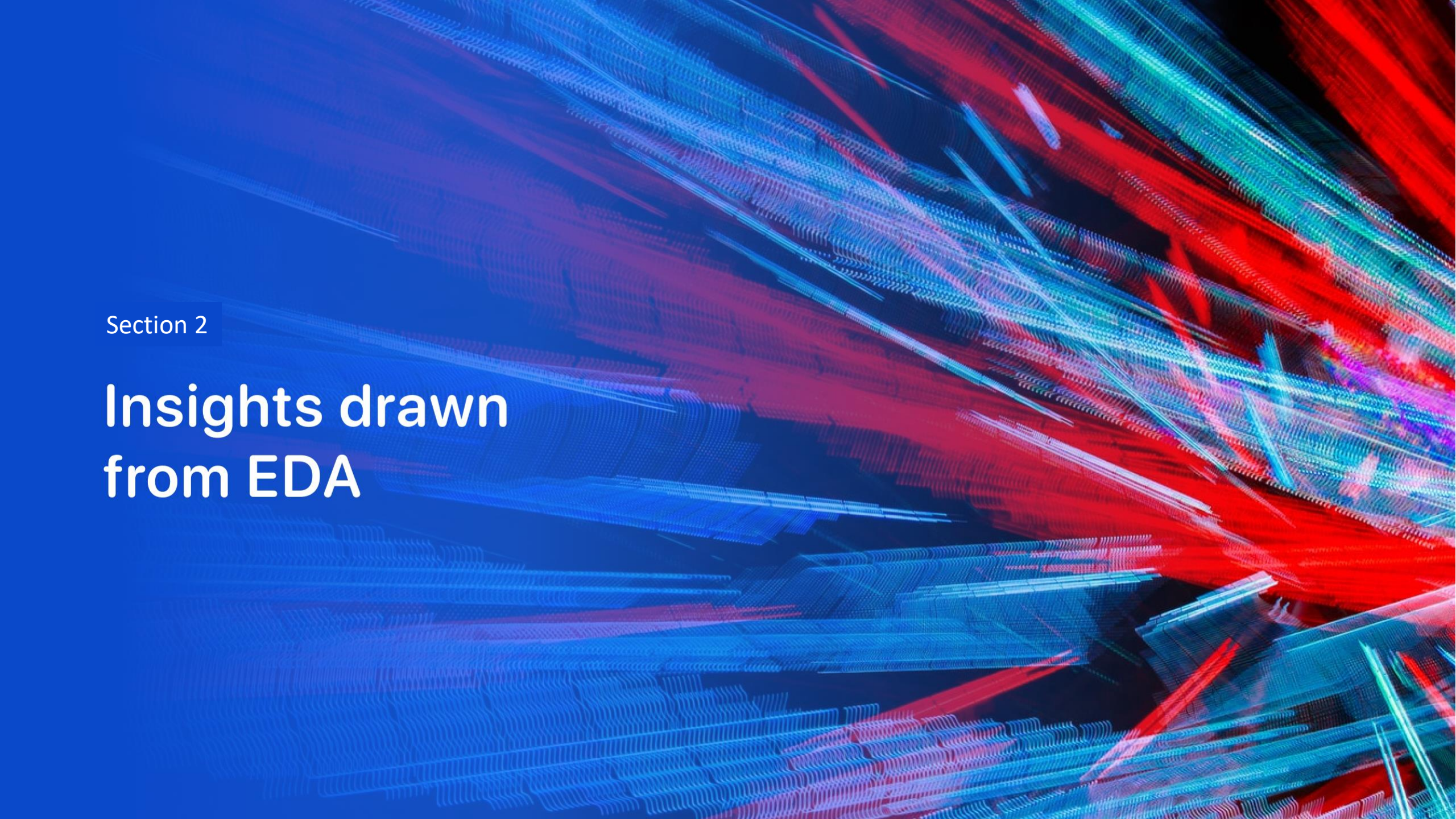
Predictive Analysis (Classification)

- Different classification models were used in order to determine the best one for this case. We work with training and test data generated from the same sample. For each method, a set of parameters was proposed, in order to evaluate them and determine the most suitable set, using GridSearchCV. The models used were K Nearest Neighbors, Support Vector Machine, Decision Tree and Logistic Regression, for each of them the precision was determined and a confusion matrix was elaborated. It was found that the most accurate model was the Decision Tree model.

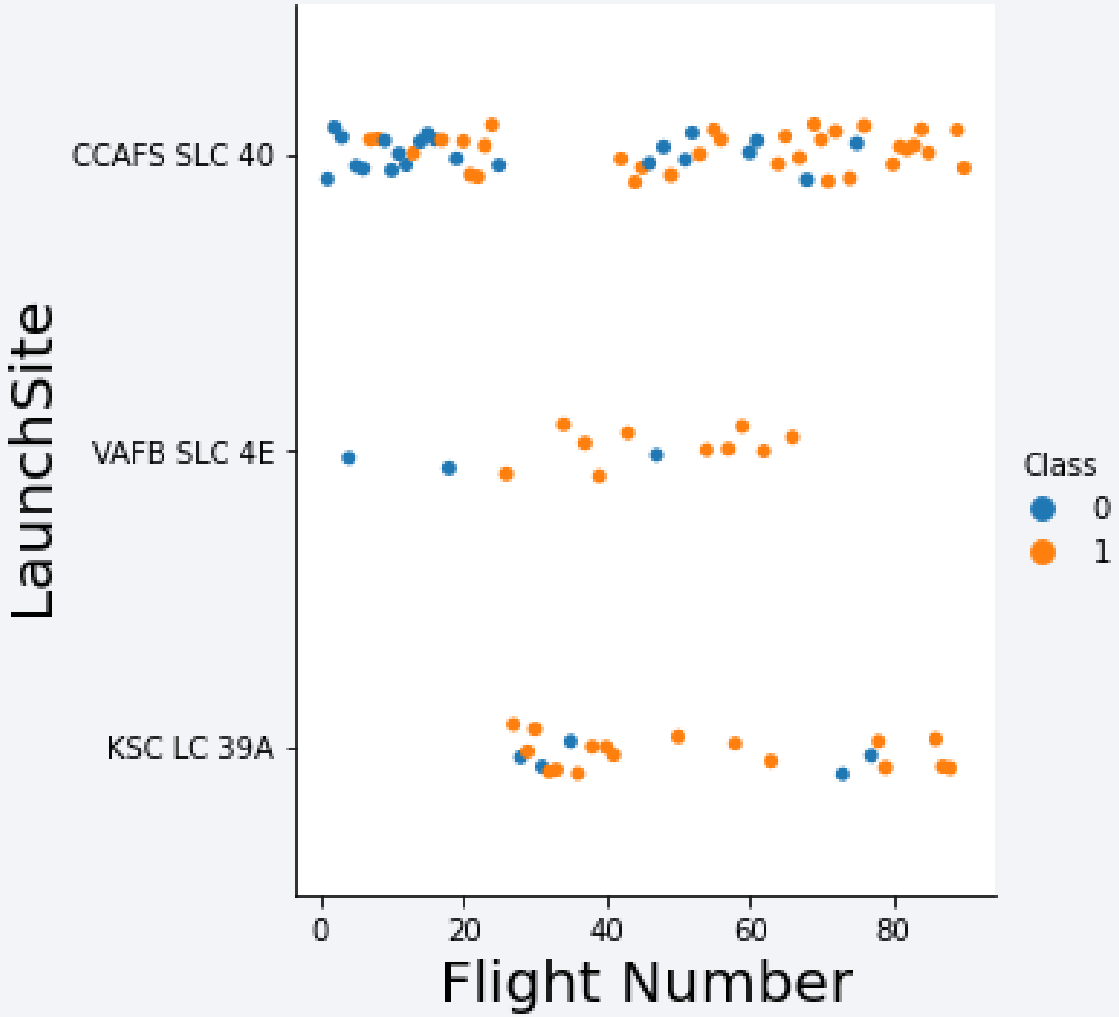
As a reference and for peer-review purpose I include my notebook with the process: Here

Results

- We observe that for flights with a greater number of results, they have been better, therefore this would represent an evolution. This evolution is reflected in the analysis of the success rate over the years. Also another important factor would seem to be the choice of orbit, since the results vary a lot depending on this choice. There are orbits without any success rate and other very successful ones close to 100%. Also, observing the relationship between these and the payload mass, we see that it is necessary to consider this last factor when choosing the orbit.
- The predictive analysis presents a high accuracy, greater than 80% for all the models worked on. In particular, the decision tree allowed to predict 16 launch results correctly, out of 18 situations presented, so it is a good tool to primarily obtain a possible panorama of the situation.


Section 2
Insights drawn from EDA

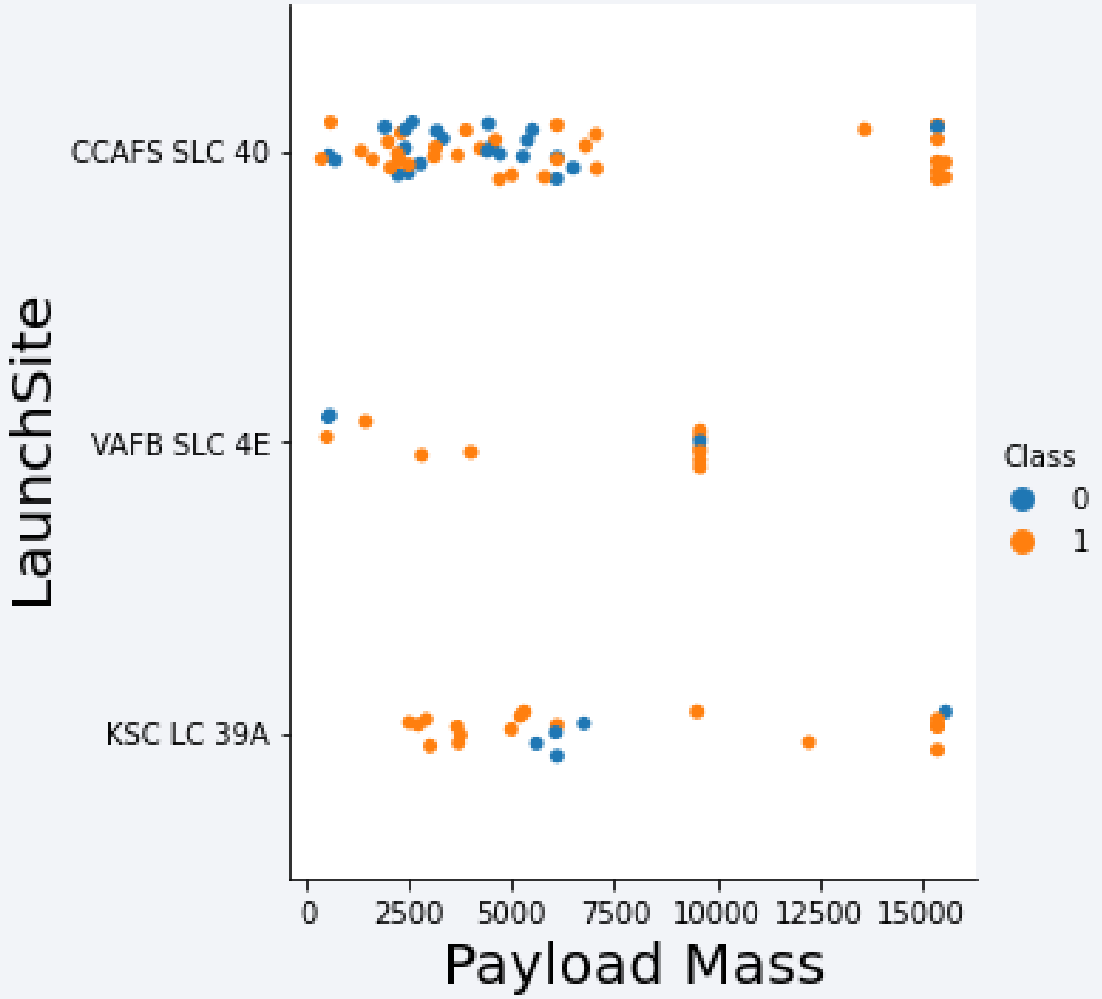
Flight Number vs. Launch Site



Plot of Flight Number vs. LaunchSite with indications of successfull or failed launched.

We see that different launch sites has different success rates and same for differents flight numbers. However, for small flights numbers the number of failures is larger

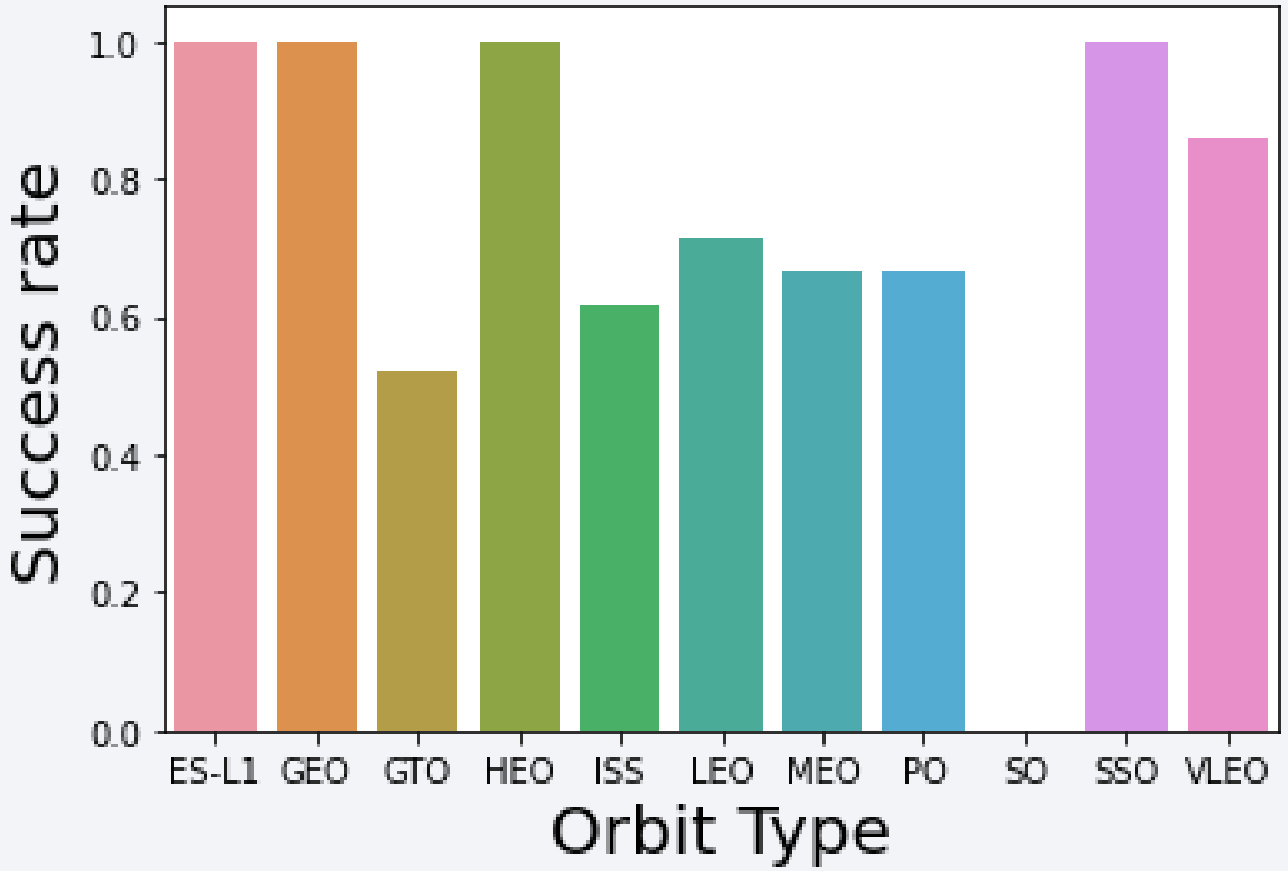
Payload vs. Launch Site



Plot of Payload Mass vs. LaunchSite with indications of successfull or failed launched.

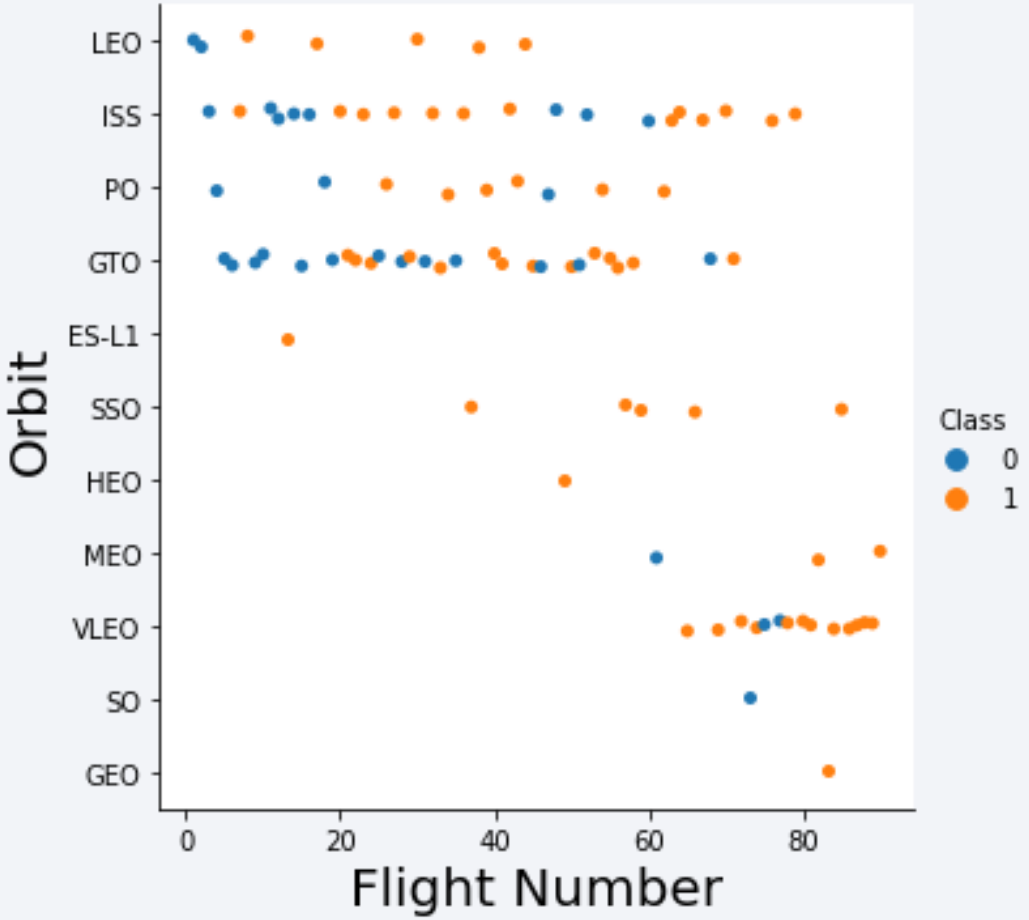
We see that different payload mass has different success rates, for larger payload masses there are fewer launches but the vast majority are successful.

Success Rate vs. Orbit Type



Bar plot of Orbit Type vs. Success rate.
We see that the difference between some orbits is notoriously. SO orbits hasn't got successful results, and some orbits like ES-L1, GEO, HEO and SSO has 100% of successful launches.

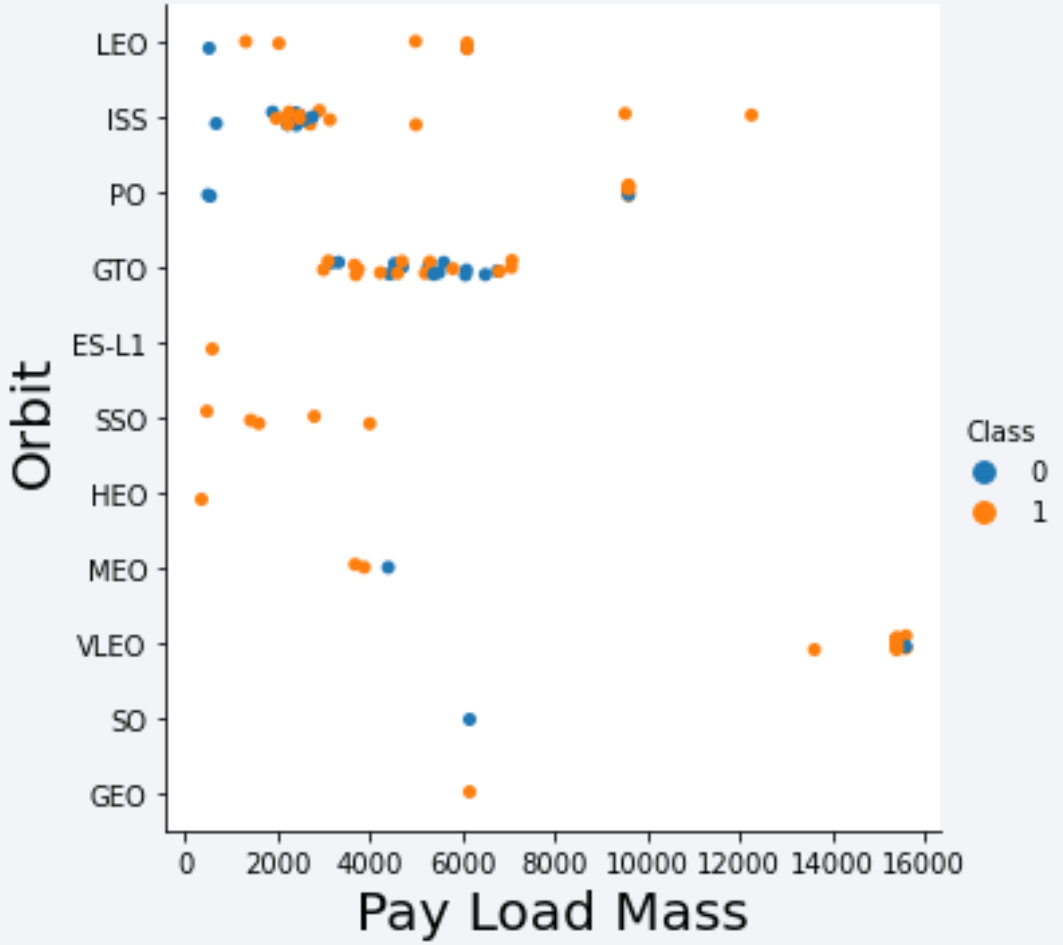
Flight Number vs. Orbit Type



Plot of Flight Number vs. Orbit Type with indications of successfull or failed launched.

We see results similar to the barplot. The mentioned orbits has 100% of successful launches. It is also seen that most orbits have more than 50% success rate.

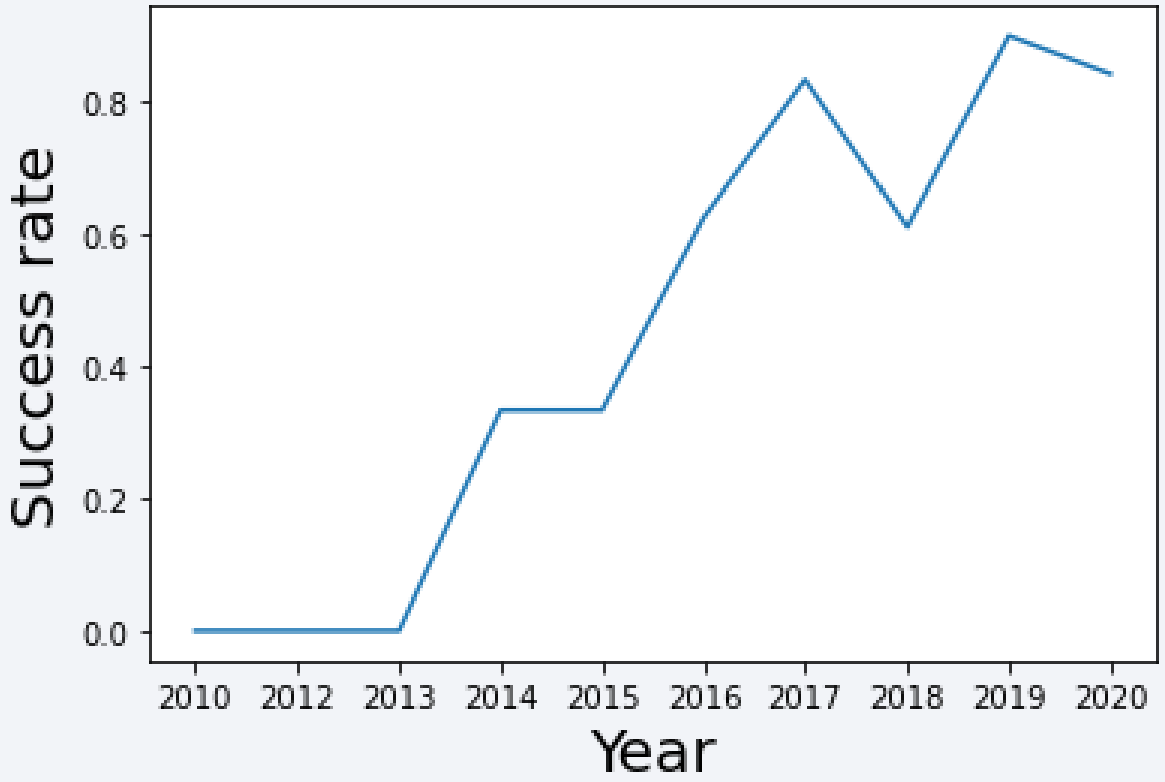
Payload vs. Orbit Type



Plot of Payload vs. Orbit Type with indications of successfull or failed launched.

We see a relationship between the orbit used and the payload mass used. Some orbits would be indicated for certain types of mass. For LEO and ISS orbits it is seen that failures usually occur with small masses.

Launch Success Yearly Trend



Line plot of Year vs. Success rate.

It is seen that the trend of the success rate is to increase as the years go by, although it presents some plateaus or stagnation in some periods.

All Launch Site Names

- Find the names of the unique launch sites

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

Four types of releases are counted. With Folium it was possible to see its proximity to the coasts.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
45596

The payload mass corresponds to the mass transported by the vehicle. In this case we observe that the total mass transported was relatively large.

Average Payload Mass by F9 v1.1

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Calculate 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

- List the names of the booster which have carried the maximum payload mass

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

substr(Date, 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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

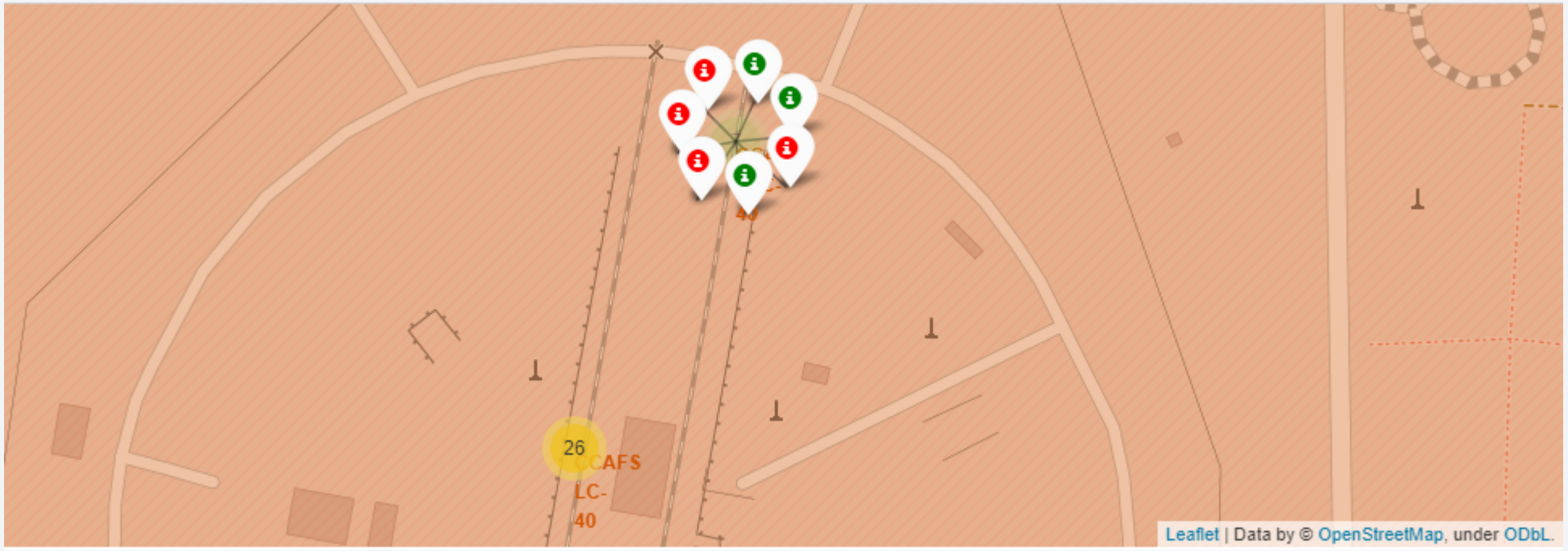
Landing _Outcome	count(`Landing _Outcome`)
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

Launch sites location analysis

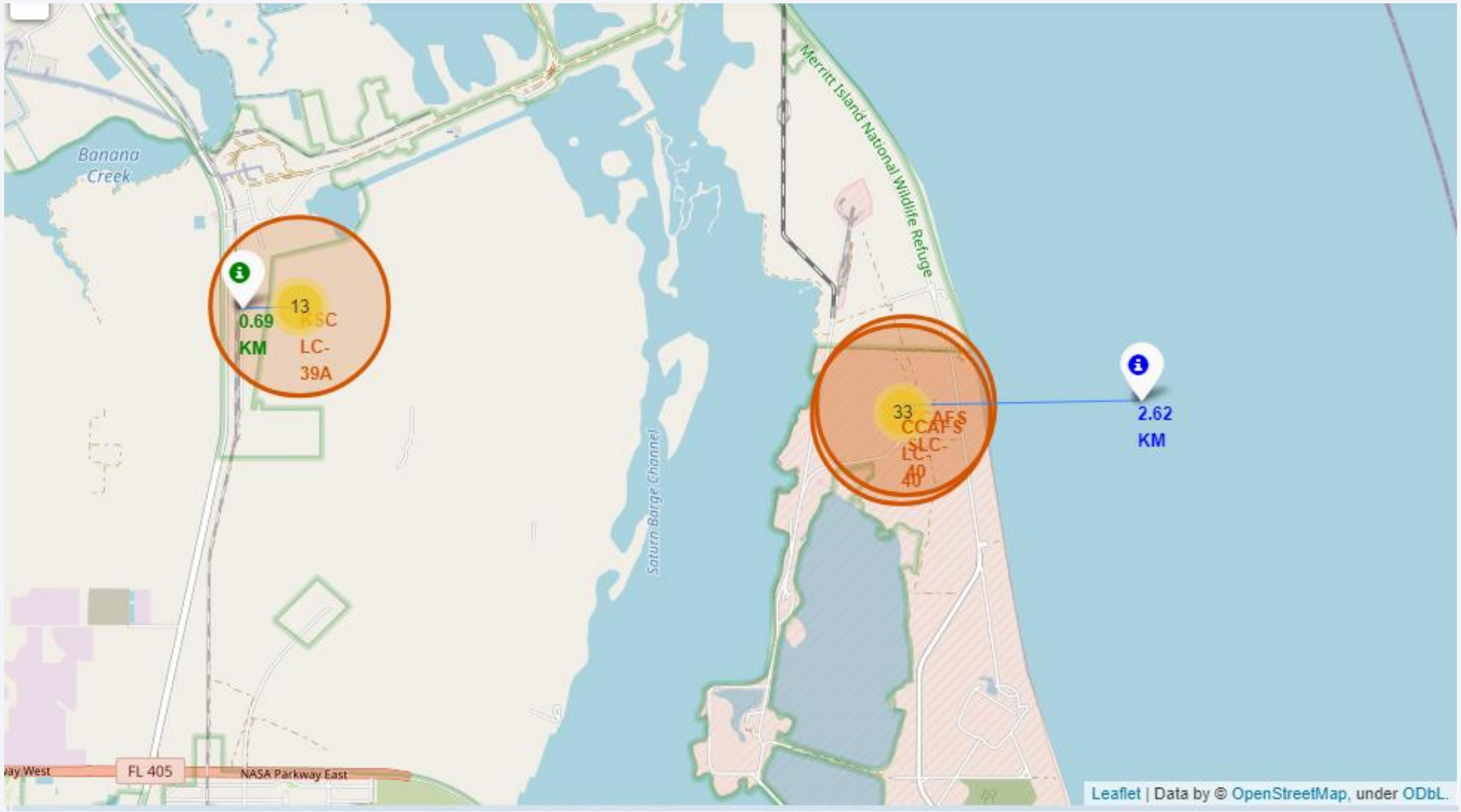
All the different launch sites are marked on the map. As can be seen, they are all close to the coasts and some of them are very close to each other. In turn, sites are found on the east and west coasts of the United States.

Type of launch for a site

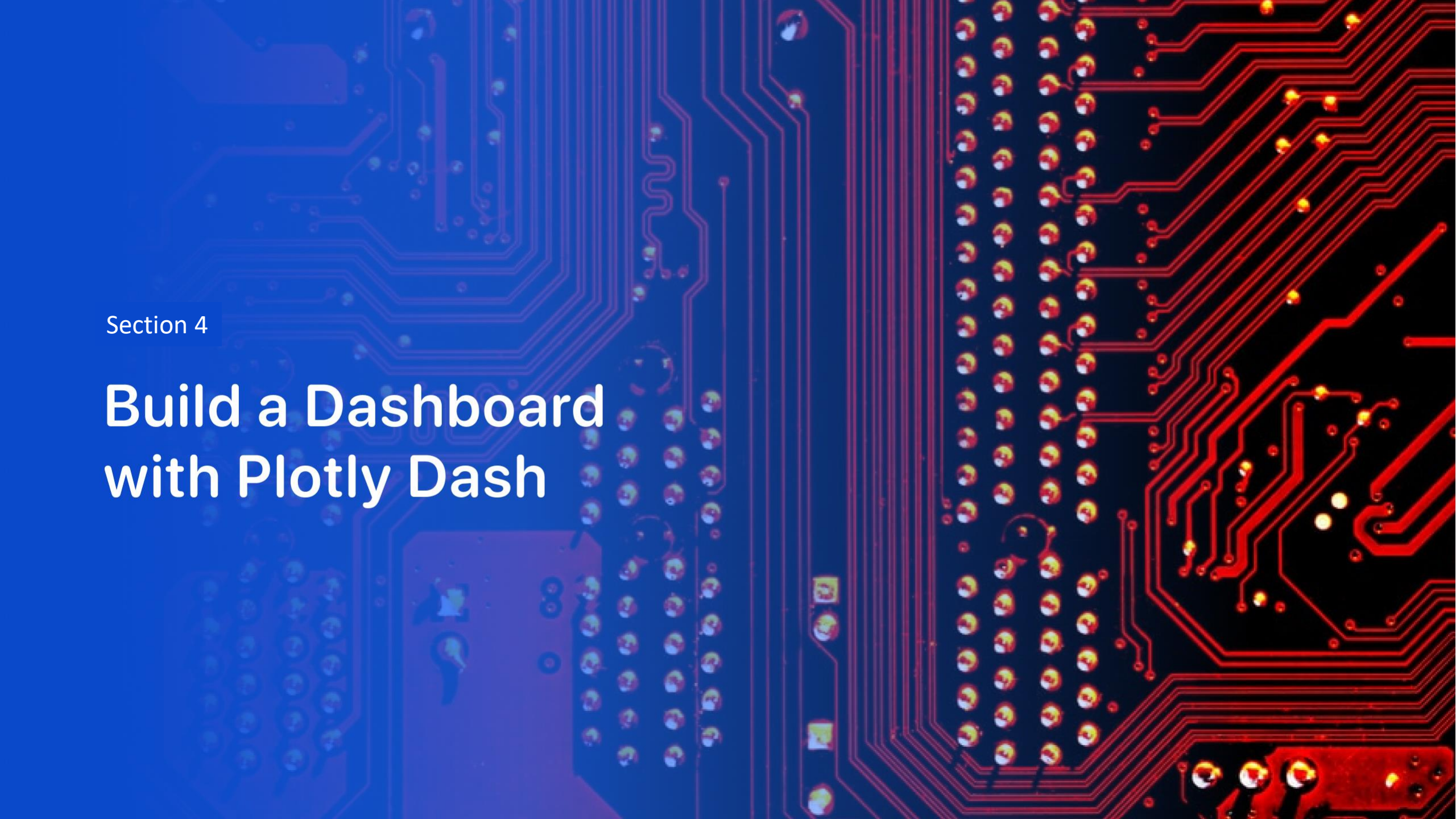


Successes and failures for a given site are displayed. With this procedure it was possible to see that there was not a 100% effective site for launches, although the causes could be independent of the site.

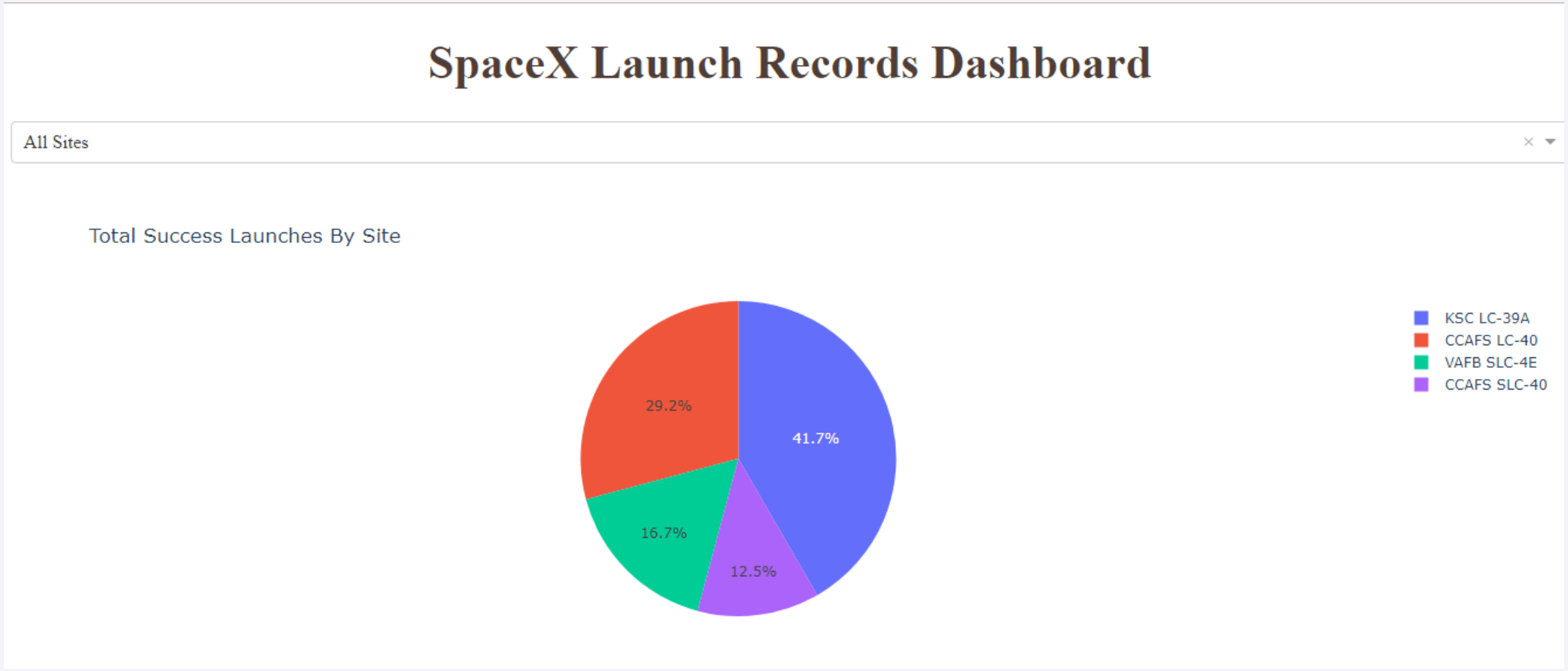
Analysis of proximities for a site



Measuring the distances between the launch points and some centers of interest, it was observed that all of them are relatively close to at least one of these points, be it the coast, railways, highways, etc.

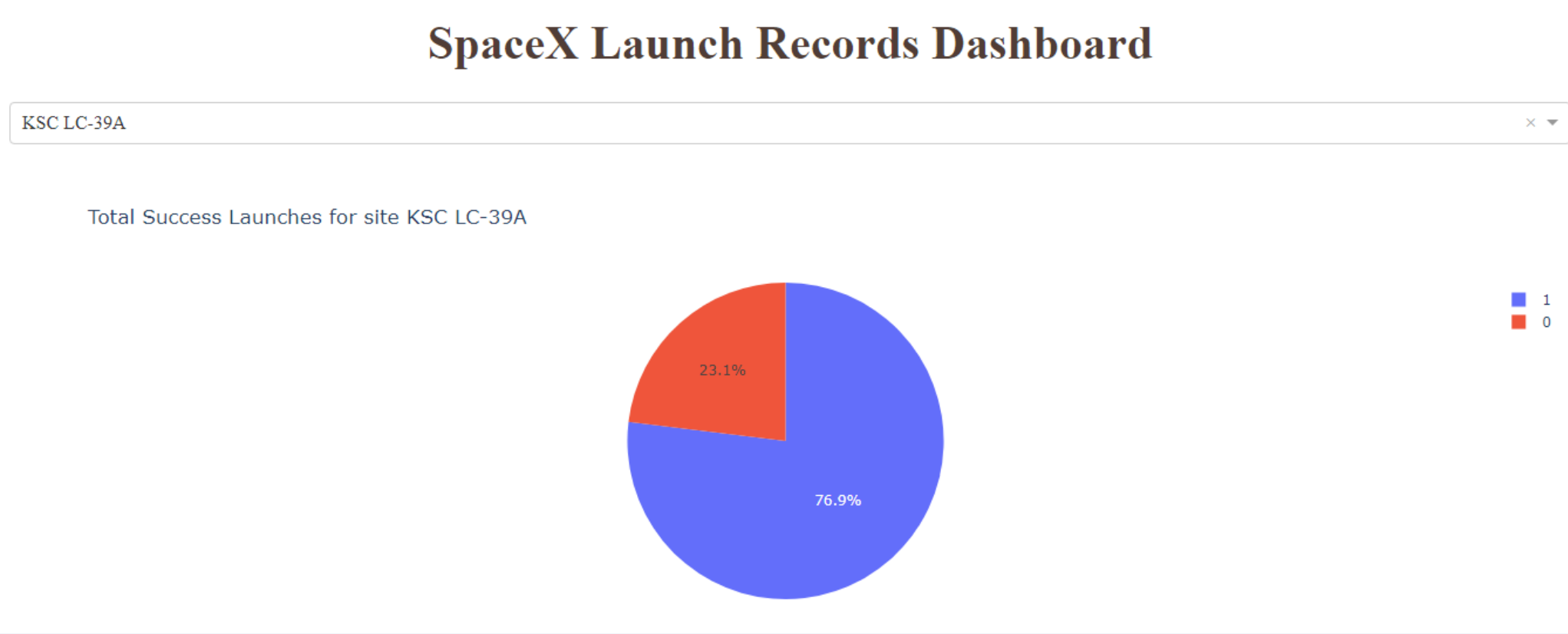

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site



Success rate comparison for all sites. We note two launch sites that have a markedly higher success rate: KSC LC-39A and CCAFS LC-40.

Total Success Launch for a site



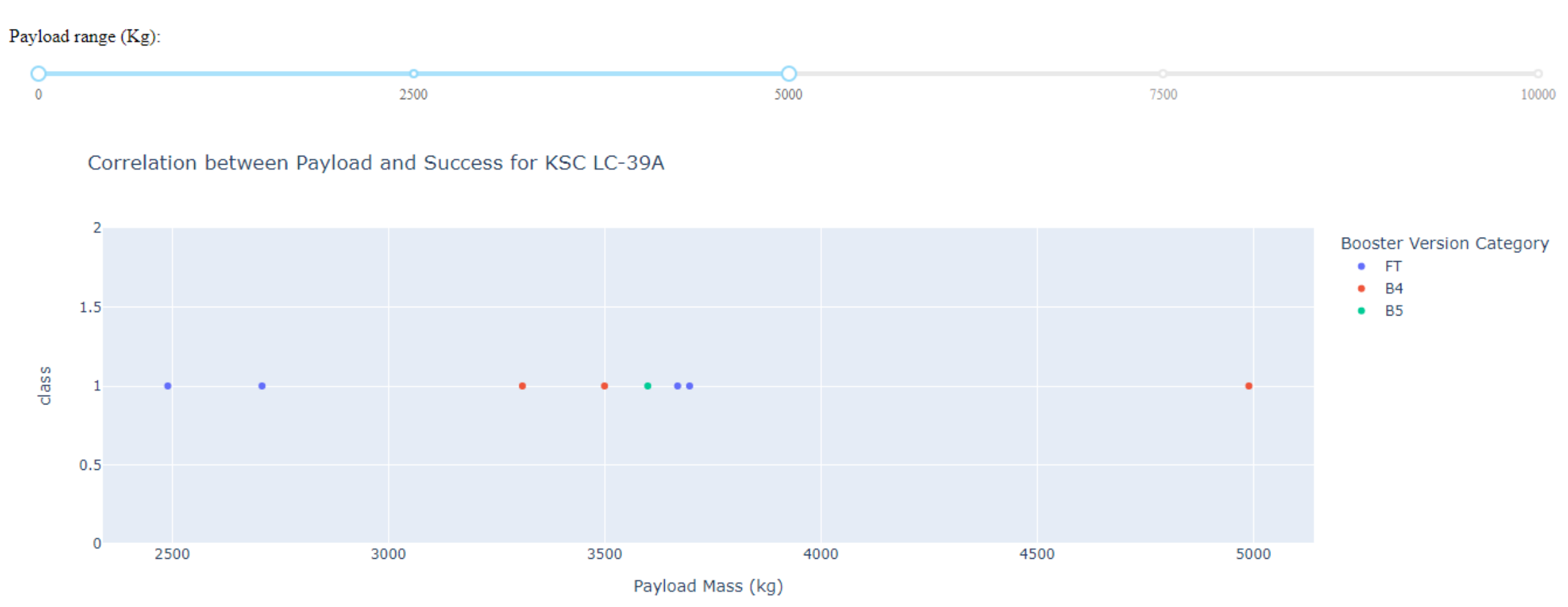
Analysis of success launches for the most successful site. It is noted that more than 30% of launches there are successful.

Relation between payload mass and the result of launches for different booster versions. It is observed that for masses greate

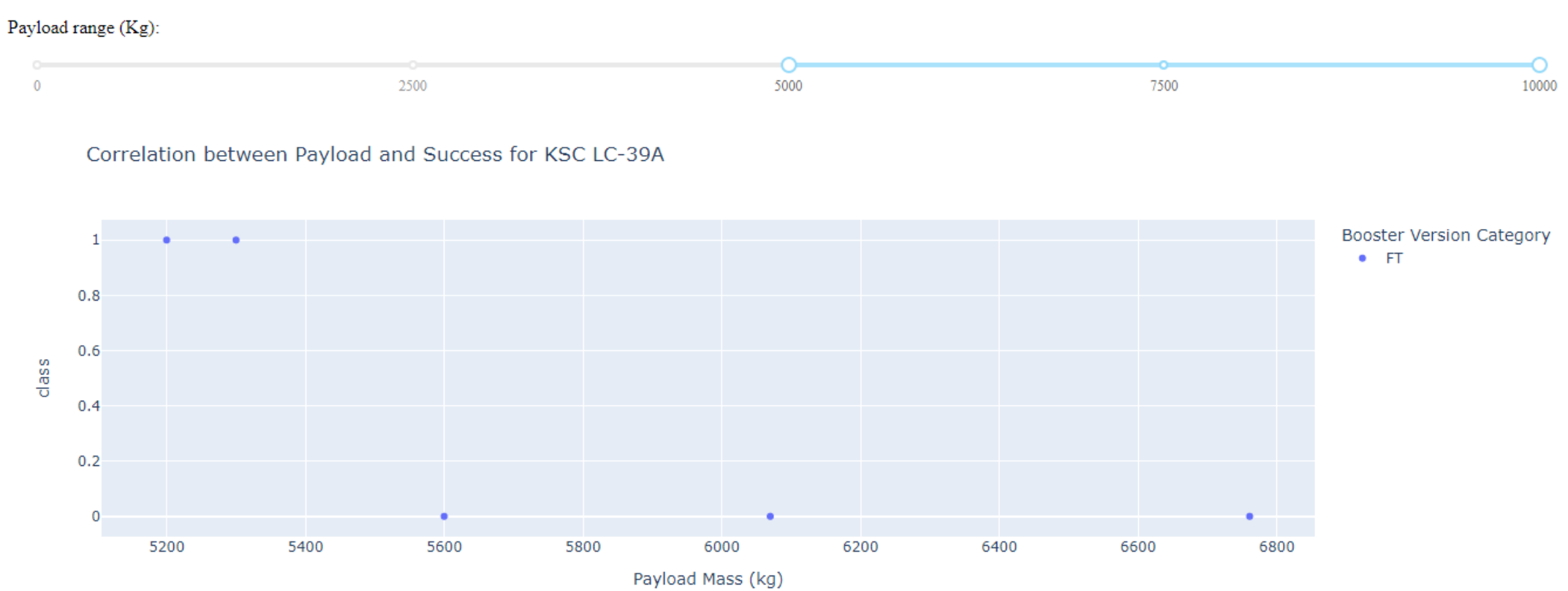
Resultados de traducción

1) Payload-Success Correlation for a site

2) Payload-Success Correlation for a site



3) Payload-Success Correlation for a site

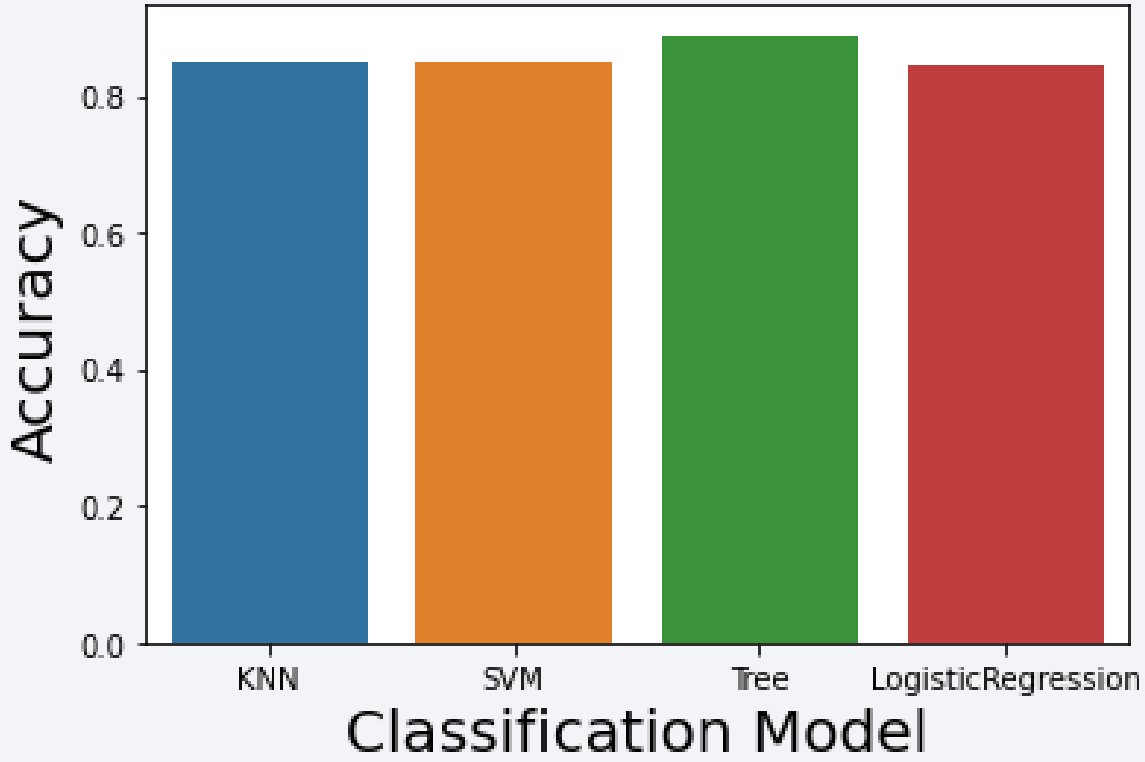


Relation between payload mass and the result of launches for different booster versions. It is observed that for masses greater than 5000Kg, only the FT booster version was used, in turn, in this mass range only two launches were successful and with masses close to 5000, there are no successes for greater masses.

Section 5

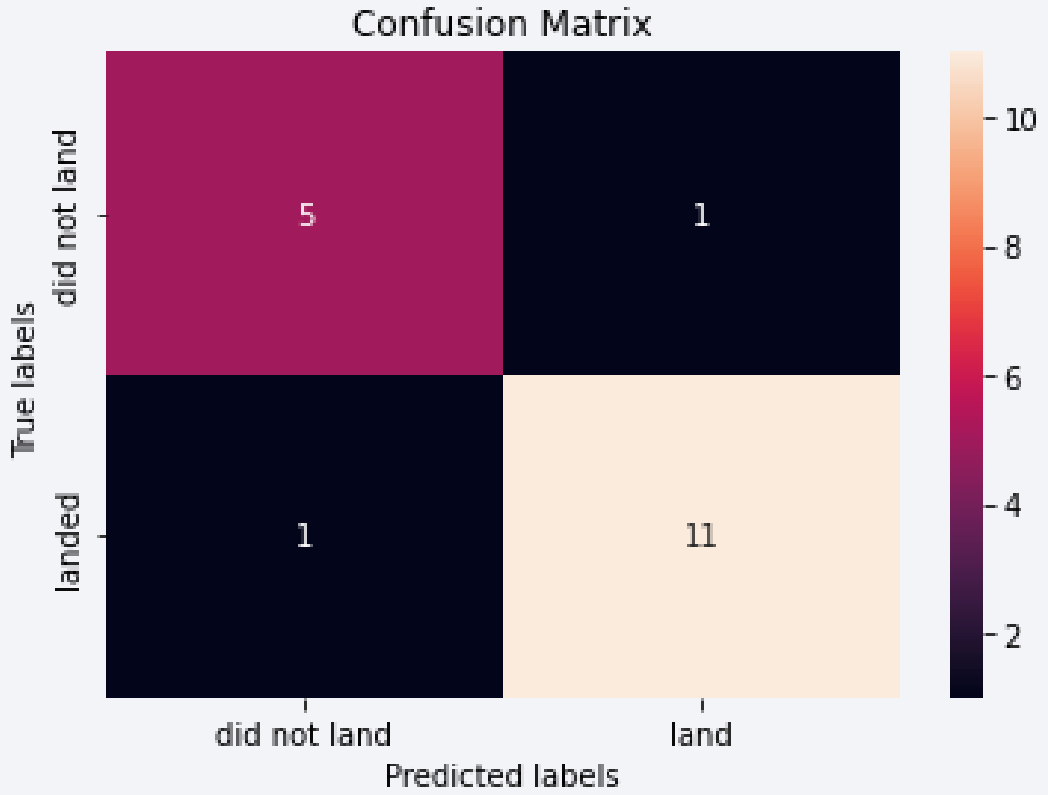
Predictive Analysis (Classification)

Classification Accuracy



Accuracy comparison between the different classification models used. We observe that although the difference in precision is small among the four, the one with the highest precision is the decision tree.
The difference between the remaining three is very small.

Confusion Matrix



Confusion matrix for the test results with the decision tree. It is observed that of the 18 elements to be predicted, 16 were correctly predicted and two were incorrectly predicted. He identified 11 successes correctly and 5 failures correctly. The errors were to identify a success as a failure and a failure as a success.

Conclusions

- The continuous development in the area and the knowledge of the situation and the problems presented mean that the success rate for launches is increasing, this is also seen in the fact that there are more successes in higher flight numbers.
- An important factor for the choice of factors such as the booster version to be used or the type of orbit is the consideration of the objectives and characteristics of the mission (for example, payload mass).
- The type of orbit seems to have an important relationship with the outcome of the mission. There are orbits that could be ruled out due to their low success rate and others that are 100% successful.
- Given the location of the four hubs, it appears that Space X prefers sites closer to shore. In particular, the coasts near the east have the advantage that the impulse of the rocket would take it to the sea, away from populated areas to prevent incidents. Regarding this, the location of one of the centers on the west coast is striking, in plotly dash it is seen that it is one of the least successful. In turn, the chosen points are as close to the Equator as possible, which, as is known, is the place with the highest speed of rotation.
- Searching for the best fit values for the classification models, very good accuracies are obtained for all the models. In particular, the decision tree would be a good tool to evaluate possible results before a possible launch.

Thank you!