Winning Space Race with Data Science

Author _ Jose Alberto Camacho

July 23, 2023

Outline

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

Executive Summary

The commercial space race is a fact. Perhaps the most successful company on is SpaceX. SpaceX's accomplishments include: sending spacecraft, a satellite internet constellation providing satellite Internet access, and sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive is that they reuse the first stage. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because reusing the first stage.

So our goal will be determine if the first stage will land successfully, and for this, we must train a machine learning model and use public information to predict if SpaceX will reuse effectively the first stage.

Welcome to this passionate journey.

Jose Alberto Camacho.

Executive Summary

- **A series of methodologies were used to extract insight from raw data**

  - Data collection: Space X API and webscraping

  - Exploratory Data Analysis (EDA): using Data Wrangling, SQL language and Data Visualization

  - Machine Learning: Prediction of success through classification methods such as Logistic Regression, Support Vector Machine (SVM), Decision Trees and K-Nearest Neighbors (KNN).

- **Summary of all results**

  - In the phase of Data Collection we learn about important features in this study and that the way of extracting them is totally open and free.

  - Thanks to EDA is possible to determine the proper features to predict the success in landings.

  - Machine Learning models give us the prediction about how are related these independent variables when the dependent variable or successful landing.

Introduction

- **Project background and context**

The main objective is determinate how our company, Space Y, could be competitive in the field of commercial space rockets. With this goal in mind, we will try to do reverse engineering, this is, trying to copy successes and to avoid errors from Space X.

- **Problems to find answers**

All the Data Science projects start with some fundamental questions. In our case these are:

What is the real price of a launching taking into account its success?

What are the characteristics of launching sites in order to treat to replicate them in Space Y?

Section 1

Methodology

Methodology

- Data Collection methodology

  Our data was collected by two ways:

  - From the SPACE X Rest API – https://api.spacexdata.com/v4/rockets/
  - Webscraping in Wiki source - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Perform Data wrangling
  - Joining the data collected we created a Class feature for determining success/failure of launches.

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

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models
  - It was needed standardize, split data in train and test data and, finally, evaluated by four different classification models -Logistic Regression, Support Vector Machine (SVM), Decision Trees and K-Nearest Neighbors (KNN)-, determining how of them was the most accurate.

Data Collection

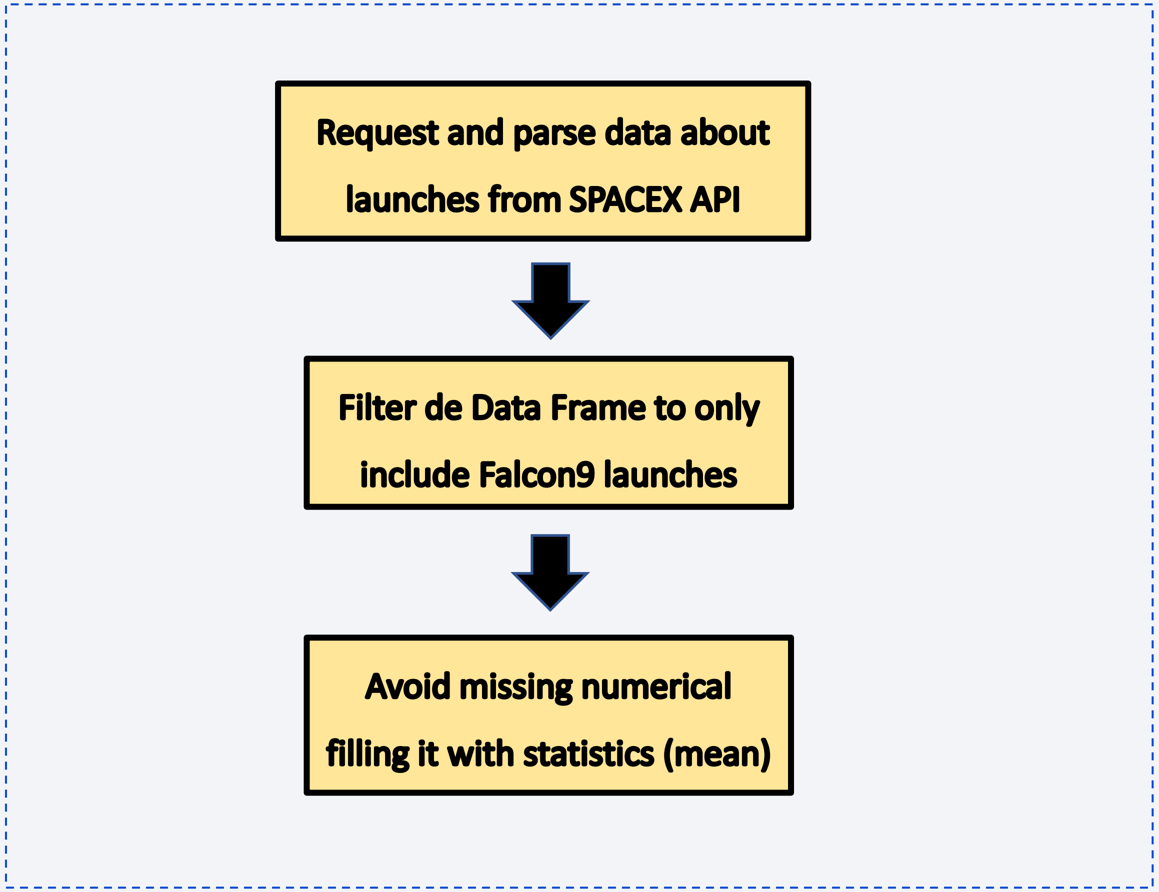
- Data was collected requesting and parsing data to Space X API, and storing by features in a Pandas Data Frame with the Booster version Falcon 9 as interest (source: https://api.spacexdata.com/v4/...)

- In other way, data was also collected with WebScraping from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches thanks to Beautiful Soup package

Data Collection – SpaceX API

- Requesting and parsing data from Space X API
- Storing such data in a Data Frame with columns like Launch Site and PayloadMass, really important for our posterior work.

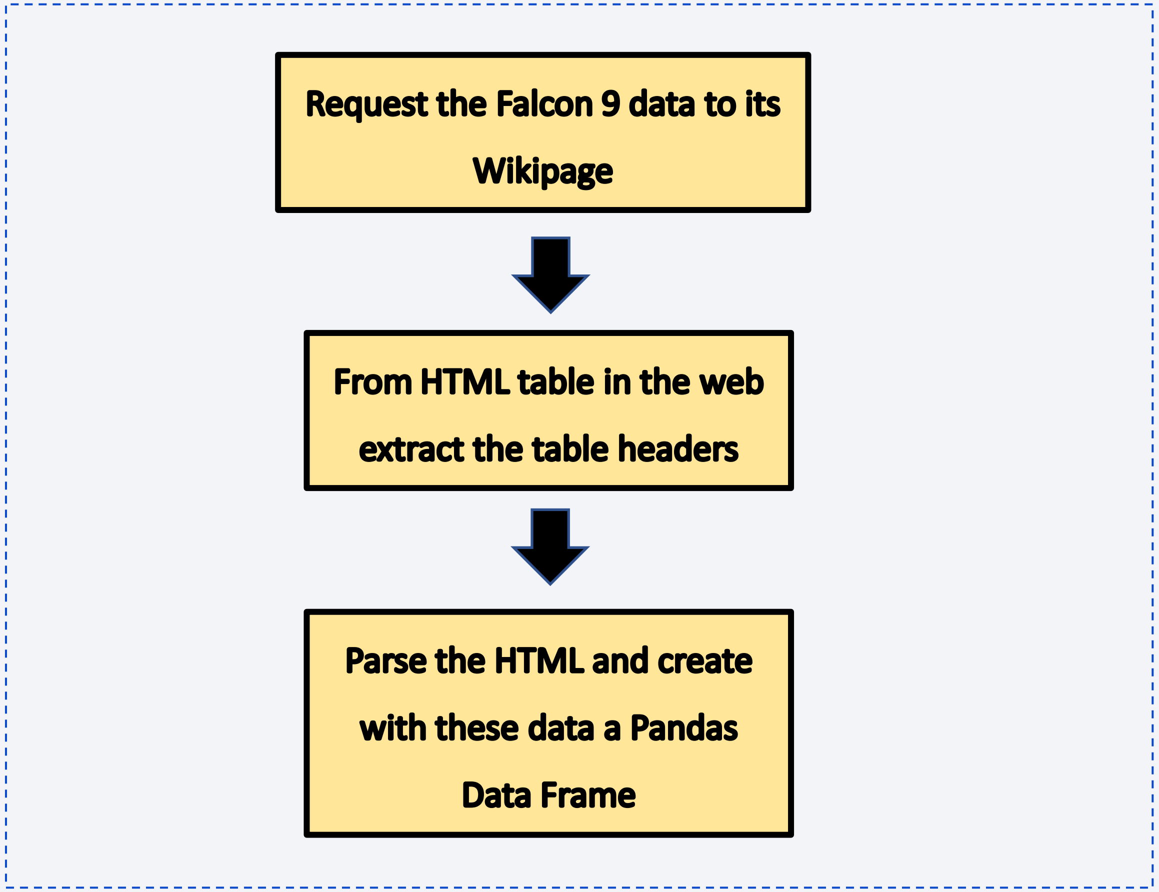
- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Use Beatiful Soup to parse the data
- This data is located in Wikipedia page from SpaceX Falcon 9 launches

- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Webscraping.ipynb



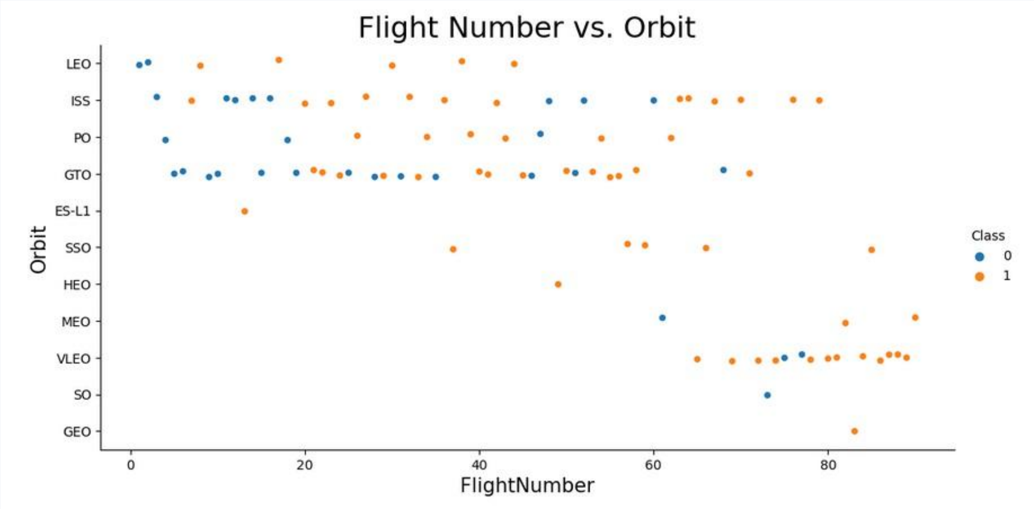
Data Wrangling

- In first place, Data Analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Calculate landing by site, number and occurrence of each orbit, and number and occurrence of mission outcome per orbit type.
- And almost more important, create a class column from Outcome column data based on success and failures.

- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- In order to establish relationship between variables, we select them by pairs and plotted them in scatter plots and bar charts.
- The plots were: Flight number vs. Payload Mass, Flight number vs. Launch Site, Launch Site vs. Payload Mass, Success of Each Orbit, Flight number vs. Orbit, Payload Mass vs. Orbit. One example:



- Github URL https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

Were used these SQL queries:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-06 and 2017-03-20

- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/EDA-SQL.ipynb

Build an Interactive Map with Folium

Markers, circles, lines, and marker cluster were created and added to a folium map

- Markers are used to sign highlighted places in a map. In our project were the Launch Sites, i.e. Cape Canaveral or NASA Johnson Space Center.
- Circles are used to mark special areas around a coordinate, like area of influence a site in our map.
- Marker clusters serve to group places in a coordinate, such as launches in a launch site.
- Lines are used to mark distance between two coordinates, like distance between a launch site and the coast line.
- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash

With Plotly Dash were plotted:

- A Pie Chart ‘Total Success Launches by Site’, to determinate in a glance the success rate of each site (in percentage) and determine what is the most successful site.
- A Scatter Plot ‘Payload range’, which permits to known the correlation between payload and success for all sites.

- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Performance Predictive Analysis (Classification)

- Separate the data in independent and dependent variables
- Standardize the data
- Split the data in train and test data
- Performance each of four classification ML model looking for its best hyperparameters
- Compare four classification models by its accuracy
- Select the most accurate of them –in our case Tree Regression-

Preparation and Standardization of Data → **Search the best hyperparameters in each Classification model** → **Determine what model returns most accurate prediction**

- Github URL: https://github.com/nietoelrelojero/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

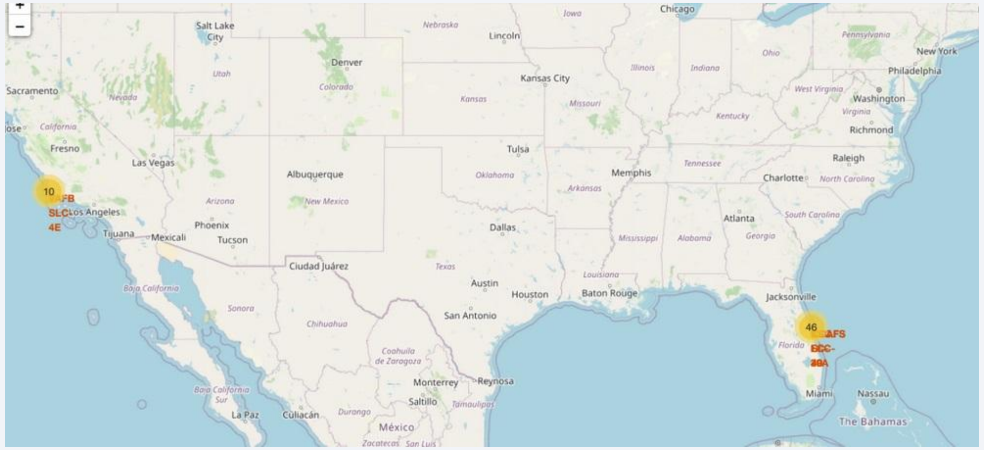
Exploratory data analysis results:

- Space X has 4 bases as launch centers
- Total Payload mass carried by NASA (CRS) was 45596 kg
- Average payload mass carried by booster version F9 v1.1 is 2928,4 kg
- The first launch took place in 2010; but only in 2015 they were able to accomplished a successful landing
- A hundred successful missions by the date of the data
- In 2015 two landing failured in drone ship
- The landing success improves with the pass of years

Results

In our journey through interactive visual analytics were possible to determine a series of insights

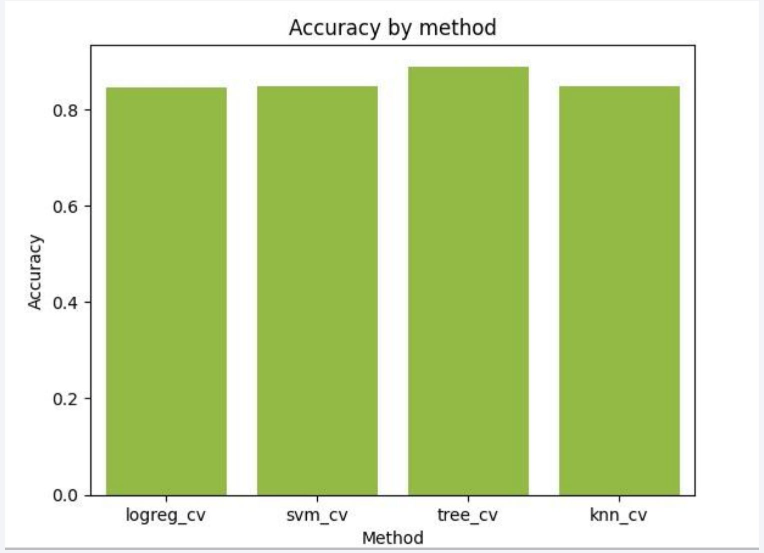
- Launches sites are near the coast line
- Launches sites are in relatively inhabited places
- Launches site are located in industrial areas with good communication



Results

Predictive analysis results

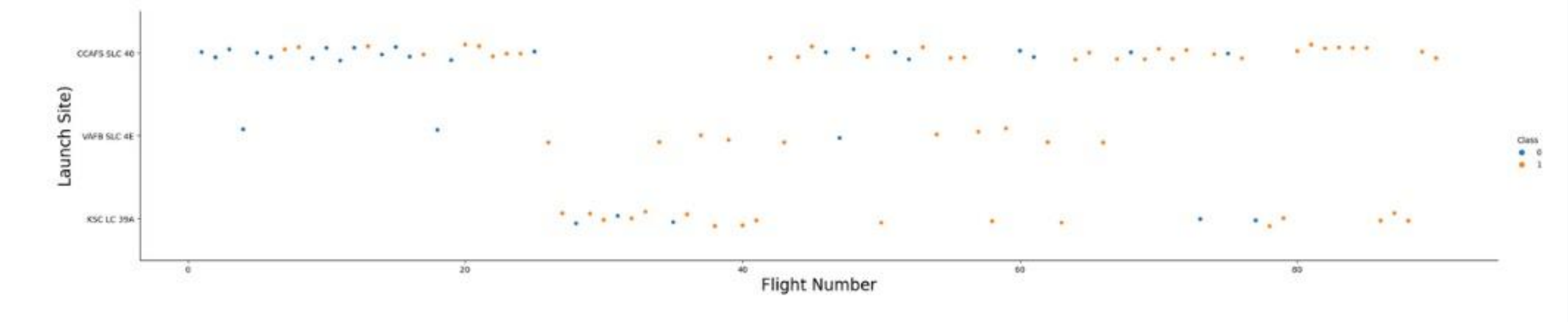
- Thanks to our different classification models performance we obtained that Decision Tree is the best algorithm to predict successful landing. After calculation it showed an accuracy around 87% and an accuracy for test data around 89%



Section 2

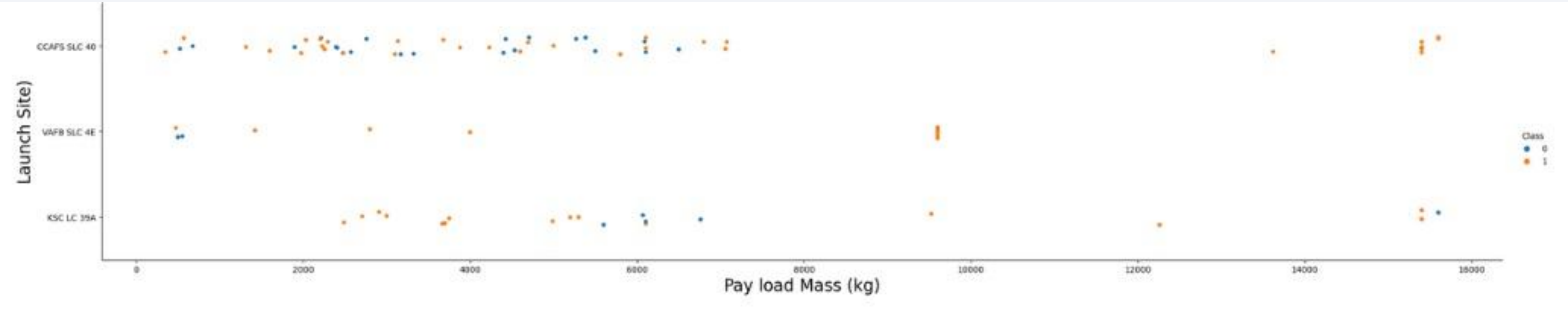
Insights drawn from EDA

Flight Number vs. Launch Site



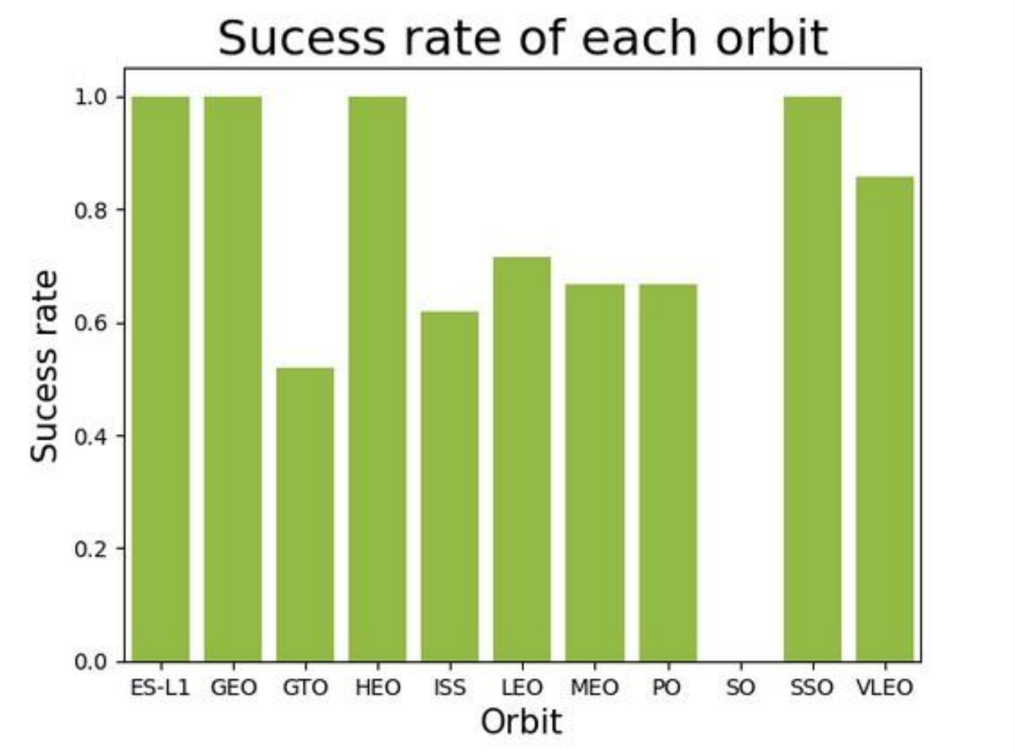
- As first impression it is possible observe the success rate is better now than in the past
- CCAF5 SLC 40 is the best launch site nowadays
- There are no launches in VAFB SLC 4E despite its last successful rate

Payload vs. Launch Site



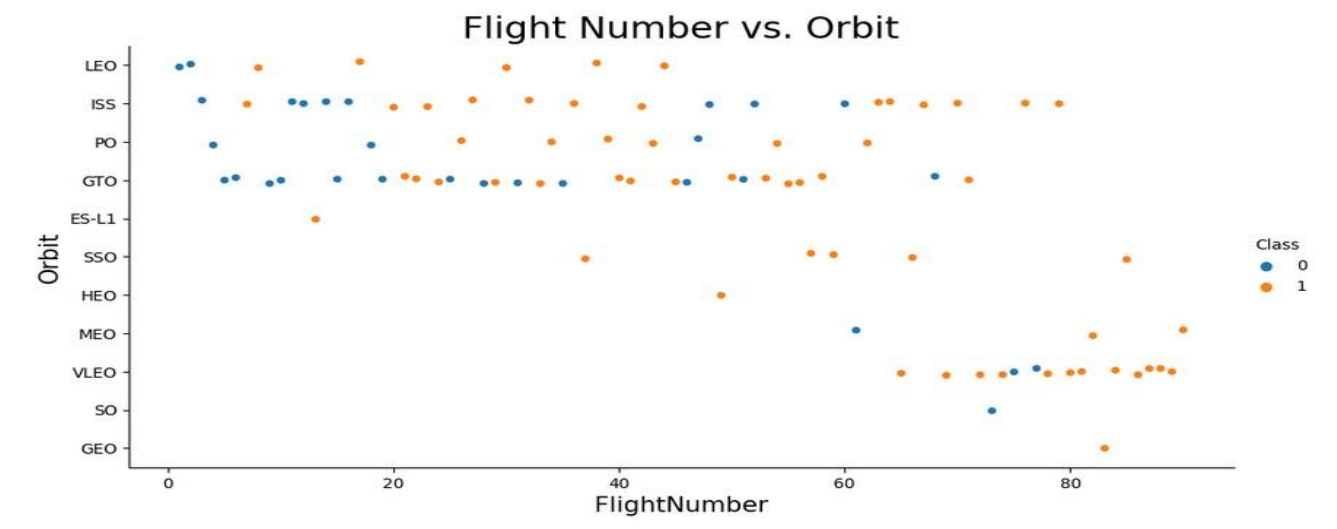
- Payloads of more than 8000 kg. are successful almost always
- VAFB SLC 4E seems are not be prepared for really heavy payloads (more than 9500 kg.)

Success Rate vs. Orbit Type



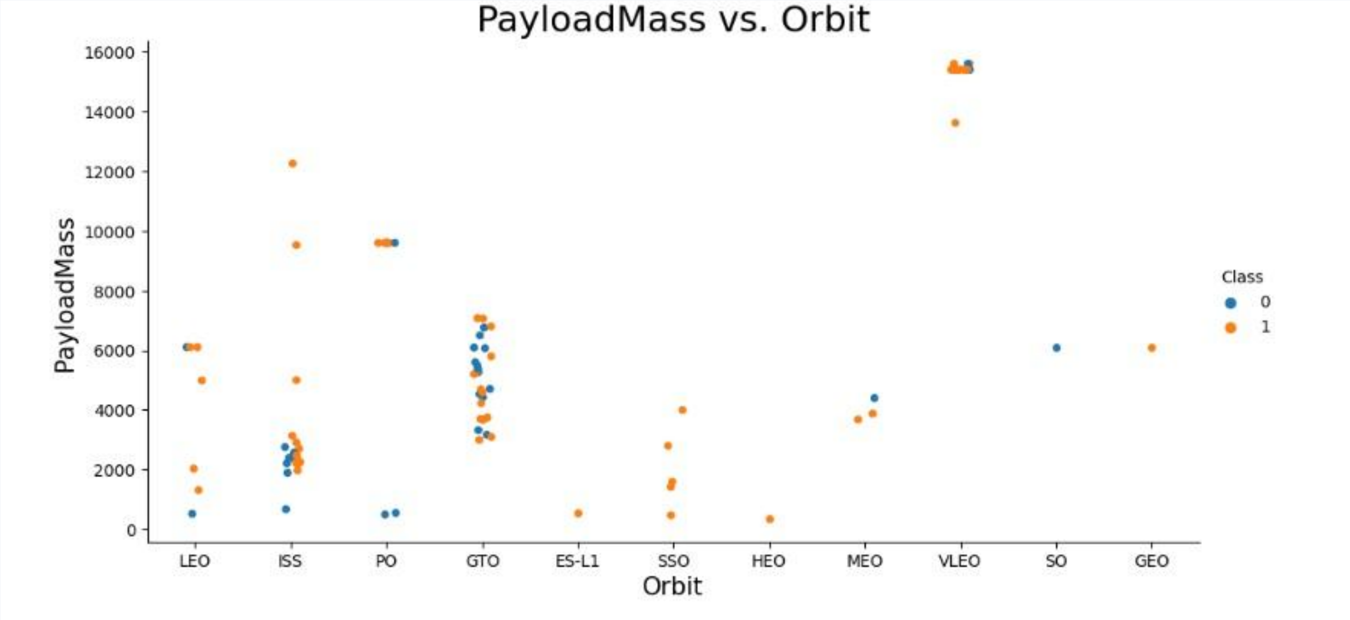
- There are 4 orbits with 100% of success: ES-L1, GEO, HEO and VLEO
- In contrast, GTO has a poor 50%
- SO have no attempts or an awful success rate

Flight Number vs. Orbit Type



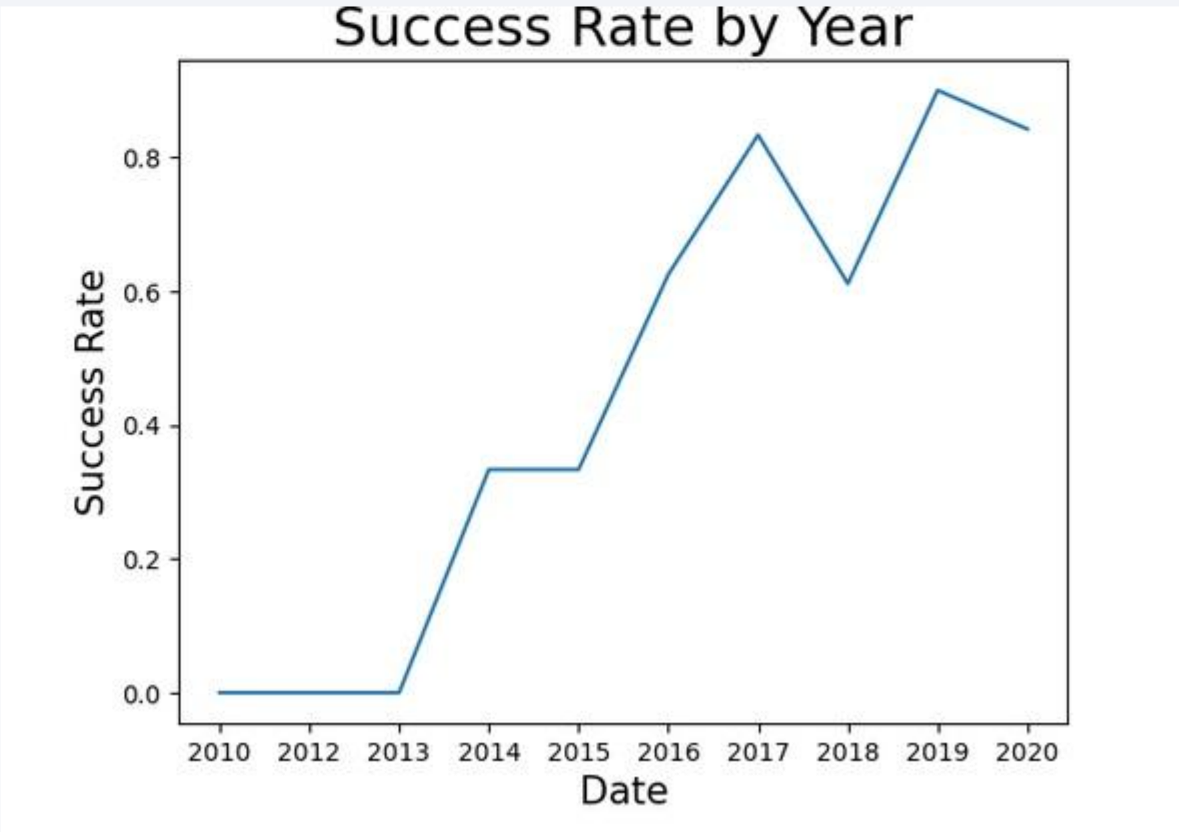
- In general, success rate seems improve in all orbits
- Some orbits (SSO, HEO, SO, GEO) supported a few launches and others are in vogue (above all VLEO)

Payload vs. Orbit Type



- VLEO accepts overweighted launches
- However, ISS has a wide range of payloads
- In almost every orbit overweighted launches are more successful, except in GTO where there is no relation between payload and success.

Launch Success Yearly Trend



- After 2013 Space X has a solid increment of success rate

- Before that year the company could pass a period of testing and researching

All Launch Site Names

- After the query, as unique occurrence of each one, it is showed all sites, 4 in total:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA` : this string is the first three letter for Cape Canaveral as we learned in the Interactive Data Visualization

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Summing all payload mass launched in NASA (CRS), in kg:

Total Payload mass carryed by NASA (CRS)
45596.0

Average Payload Mass by F9 v1.1

- With the average of the column Payload Mass carried and filtering by booster version F9 v1.1 it showed (in kg):

Average Payload
2928.4

First Successful Ground Landing Date

- Filtering data by successful landing outcome on ground pad and use the minimum function date we found this date is:

First succesful landing in ground pad
01/08/2018

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, we obtain four booster versions:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes, group by mission outcome and doing the count:

Mission_Outcome	total number
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. We see there are 12 in total:

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. Two cases for that year, in the months 4 and 10:

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

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. Very elevated the count for ‘No attempts’

Landing_Outcome	Total
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2

Section 3

Launch Sites Proximities Analysis

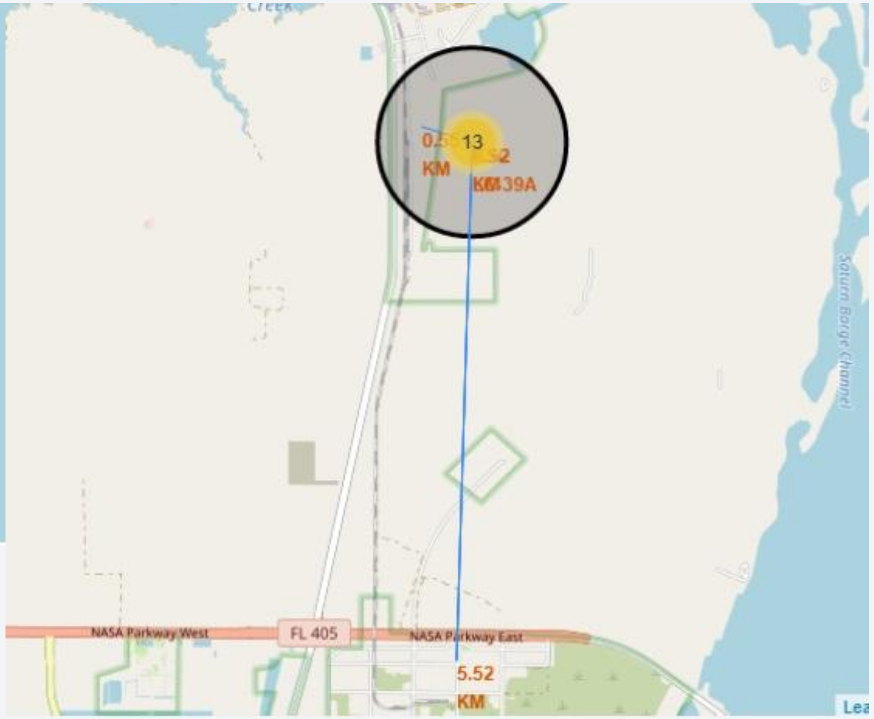
All launch sites

- The map shows all the sites are in coast lines, far away from big cities but maybe close to highways.
- The sites are very distant between them in the United States

Labeled launch outcomes

- An example of Cape Canaveral where green markers indicates success launches and red ones failure launches

Folium Distances

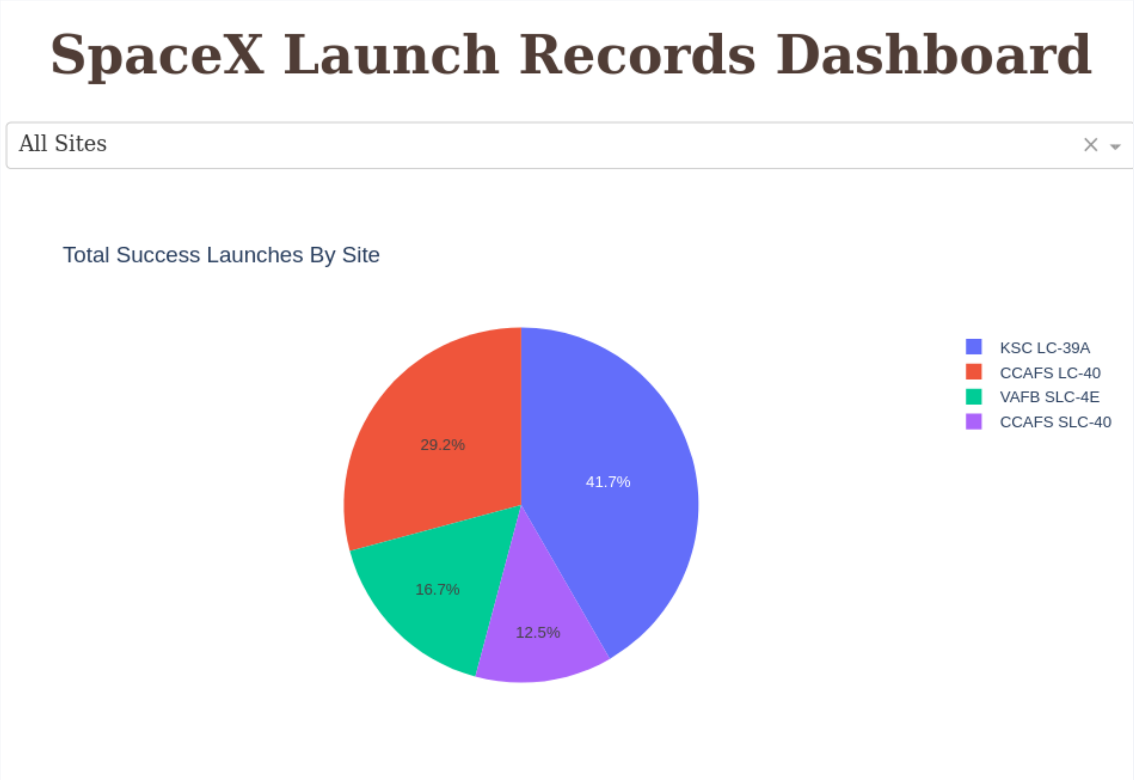


- In the screenshots it is possible determine the distance to highway, infrastructures, and coastline, with the values calculated and displayed
- Launch site is a place alone without urban areas near it

Section 4

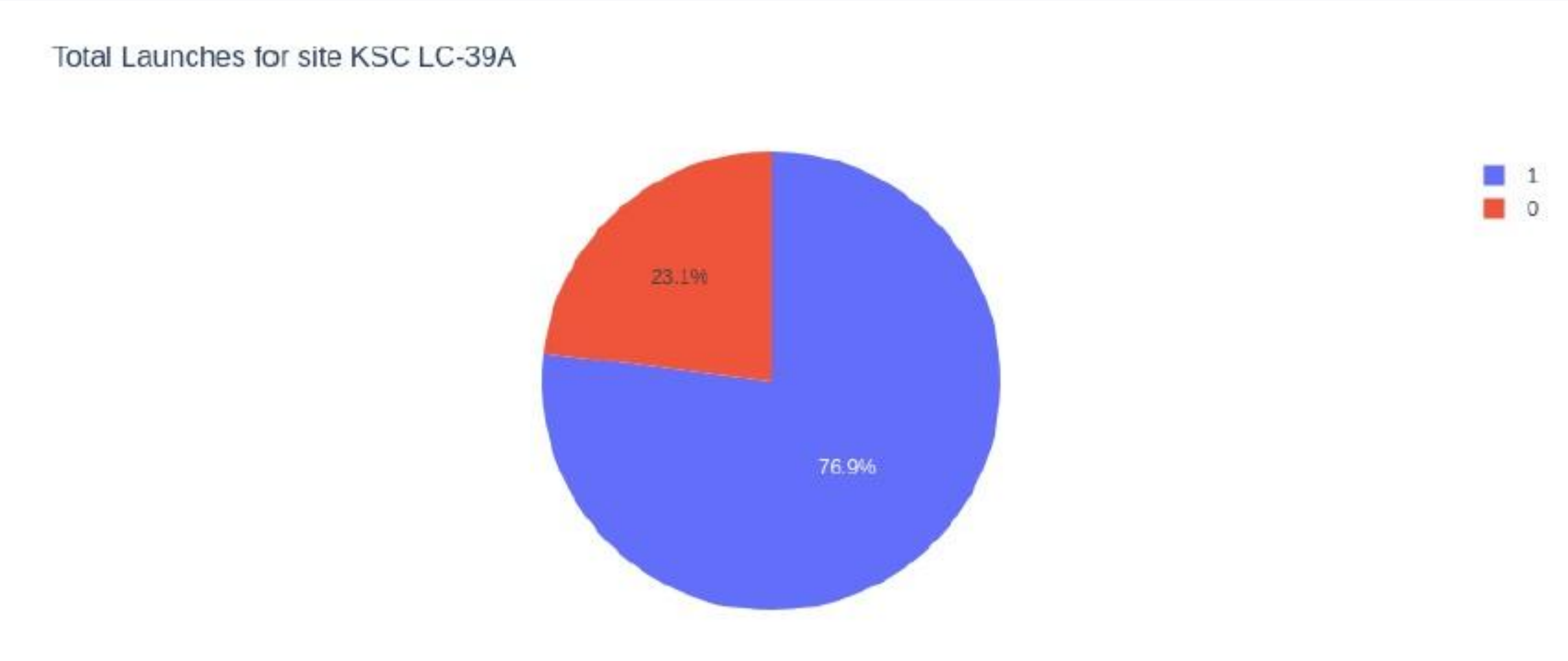
Build a Dashboard with Plotly Dash

Successful Launches by site



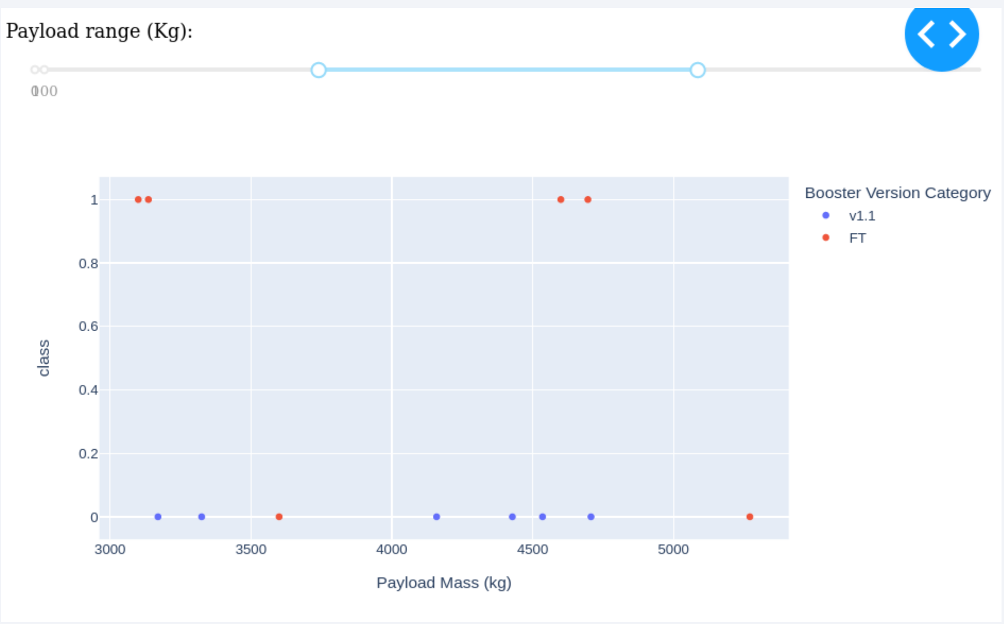
- The site KSC LC-39A is, with some appreciate difference, the most successful site of our comparative

Launch Success Rate for KSC LC-39A



- This launch site has a respectable 76.9 of success

Payload vs. Launch Outcome

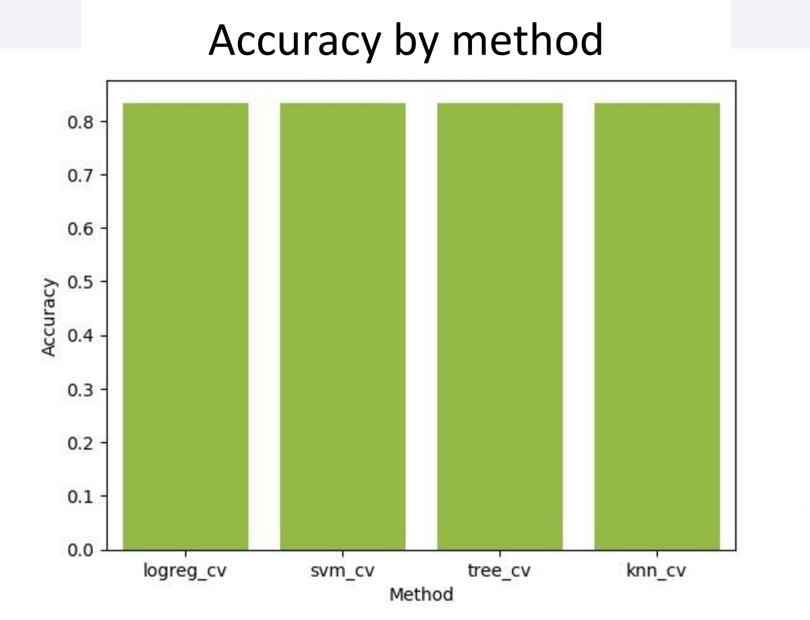


- Payloads of less than 5500 kg are more probabilities of success
- In a superior range (>8000 kg) only Version B4 is able to support the launch

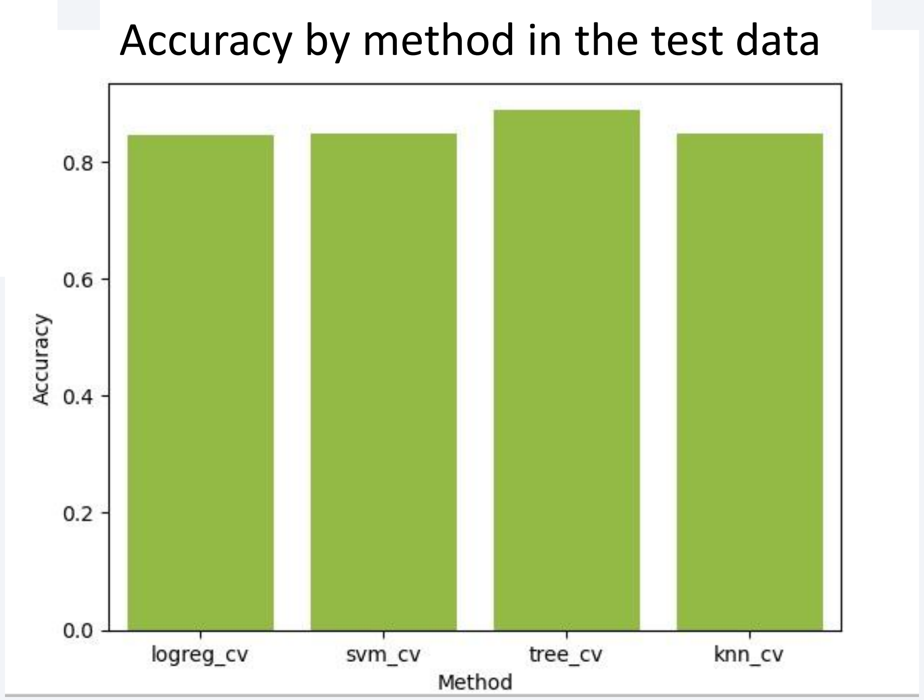
Section 5

Predictive Analysis (Classification)

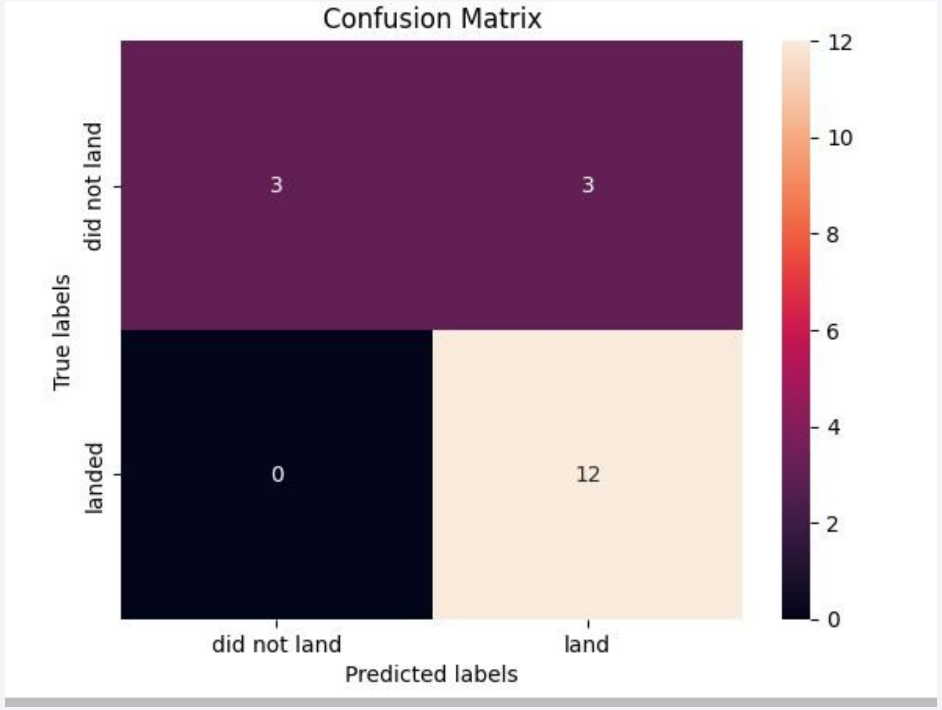
Classification Accuracy



- For our four models - Logistic Regression, Support Vector Machine (SVM), Decision Trees and K-Nearest Neighbors (KNN)- plotted its accuracies and its accuracies in the test data
- We can observe the equality among them though the Decision Tree has a better accuracy method in the test data



Confusion Matrix



- Confusion Matrix for the Decision Tree model: a fantastic true positive prediction (100%), but a fair true negative prediction (50%)

Conclusions

After analyzing a varied amount of data we could establish the following general conclusions:

- The best launching site is KSC LC-39A. We know now its characteristics –one of our doubts of the beginning -page 5-, and we could try to replicate them with the new company SpaceY.
- It is possible to predict the successful landings with a success rate over 80% in the last years. The commercial space rockets seems to be a good business.
- Taking this into account we are able to answer the another previous question: it is realistic adjust the price of our missions to the ones of SpaceX.
- So, finally, SpaceY has everything in its favor to hit the market.

Appendix

- Just as a curiosity, another different way of established the method which performs best our requirements, instead of graphically it is with a for loop. It is the following:

```
In [39]:
#Numerically, with a for loop
methods = [logreg_cv, svm_cv, tree_cv, knn_cv]
best_method = ""
best_accuracy = 0
for method in methods:
    if method.score(X_test, Y_test) > best_accuracy:
        best_accuracy = method.score(X_test, Y_test)
        best_method = method.estimator

print("The method which perfoms best is", best_method, "with an accuracy of", best_accuracy)
```

Thank you!