IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ignacio Gargiulo
October 15, 2022

Outline

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

Executive Summary

- Summary of methodologies

We perform exploratory data analysis using SQL and visualization tools. We georeferenced the launch data and applied machine learning algorithms to predict succesful landings.

- Summary of all results

We can predict with acceptable accuracy if the recovery of a rocket stage of a given launch will succed . The success In recovering the stages is increasing with time and specific regimes where the launches are less successful can be identified.

Introduction

- *Project background and context*

Space X is a company that achieved the reutilization of rocket stages , which reduces the cost of launching satellites to space from 165 million in avg, to 62 million dollars. The first stage, that represents a huge part of the launch cost, can be recovered using retro-jets that help the stage to land again. However not all the launches have a successful first stage recovery.

- *Problems you want to find answers*

We want to predict if a given launch of a Falcon 9 will have a successful recovery of the first stage based on previous launches .

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected from the Space X web API at "https://api.spacexdata.com/v4/...", via common python packages

- Perform data wrangling

  Data was prepared using standard methodology like cleaning empty fields and normalization to facilitate the analysis with machine learning.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We used several classification models, for example: decission trees and logaritmic regression

Data Collection

We downloaded launching data, with the locations of all the launches: payload data, to know the mass and orbit destination; Booster version data, to know the type of booster used in each launch; core data, which stores many different features for each launch with specifics indicators for landings and technical aspects.

Data Collection – SpaceX API

Example of request : Launch Data

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

response = requests.get(spacex_url)

Space X API

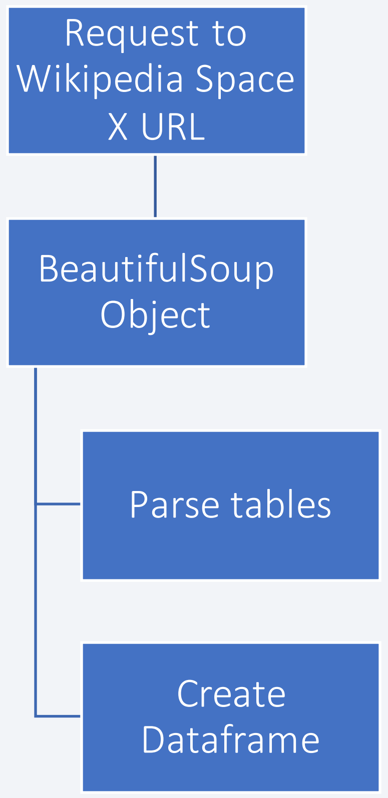
- Launch Data
- Payoad Data
- Booster Version
- Core Data

Data collection notebook at: https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Additionally to request information to the space X API, we performed web scraping to data stored in html tables in the Wikipedia.

We used the BeautifulSoup python package



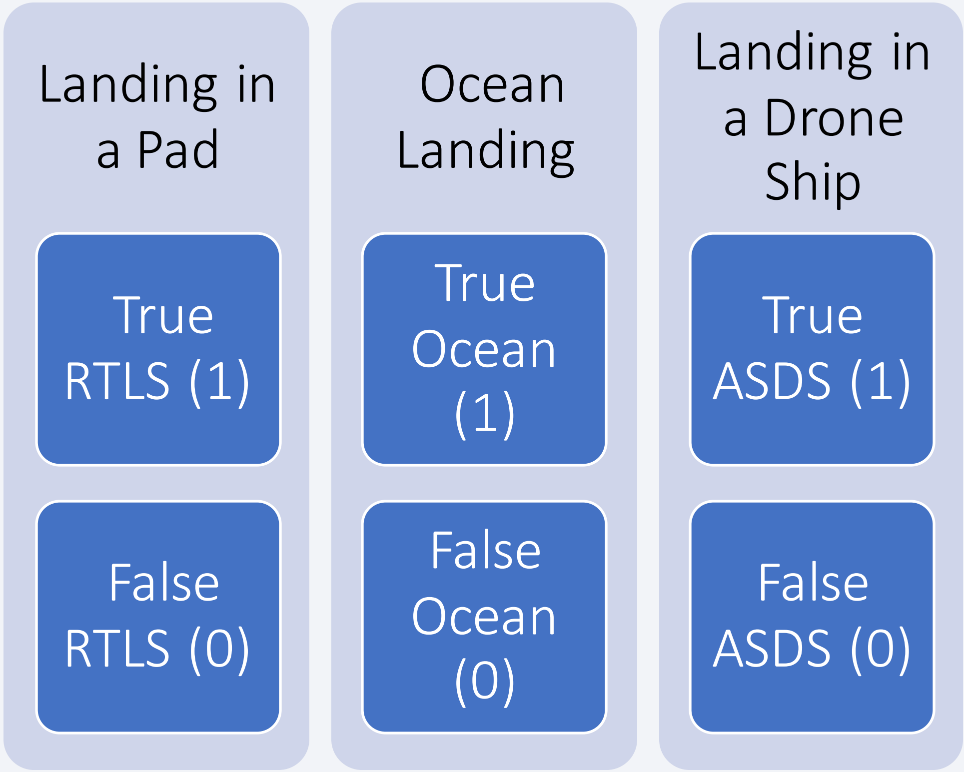
Web scrpaing notebook at: https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

Data needs to be processed after obtaining them.

- We normalized all landing outcomes to succesful or failed, disregarding the platform used.

Possible landing outcomes



1 means succesful , 0 means failed

Data Wranglin notebook available at : https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

We performed an exploratory data Analysis with visualizations.

Launch sites
- Flight number
- Payload

Orbits
- Flight Number
- Payload

Launch Success
- Yearly trend

We explored how launch sites vary as a function of flight number (number of launches in each launchsite) , and the payload .

We also explored how the orbits show success or failure as a function of flight number and payload

Finally , we analized the yearly success trend.

EDA with SQL

```
Select DISTINCT "Launch_Site" from SPACEXTBL
```

```
select "Launch_Site" from SPACEXTBL where "Launch_Site" LIKE "%CCA%" LIMIT 5
```

```
select SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" == "NASA (CRS)"
```

```
select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" == "F9 v1.1"
```

```
select MIN("Date") from SPACEXTBL where "Landing _Outcome" LIKE "%Success%"
```

```
select "Booster_Version" from SPACEXTBL where (("Landing _Outcome" == "Success (drone ship)") & ("PAYLOAD_MASS__KG_" > 4000) & ("PAYLOAD_MASS__KG_" < 6000))
```

```
SELECT "Landing _Outcome", SUM(case when "Landing _Outcome" like '%Success%' then 1 else 0 end) as success_landings, SUM(case when "Landing _Outcome" like '%Failure%' then 1 else 0 end) as failure_landings FROM SPACEXTBL GROUP BY "Landing _Outcome"
```

```
select DISTINCT "Booster_Version" from SPACEXTBL where (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

```
select substr(Date, 4, 2) as month ,"Landing _Outcome", "Booster_Version", "Launch_Site" from SPACEXTBL where ((substr(Date,7,4) == '2015')& ("Landing _Outcome" == "Failure (drone ship)"))
```

```
select "Landing _Outcome", count(*) from SPACEXTBL where (("Date" > '04-06-2010') & ("Date" < '20-03-2017') & ("Landing _Outcome" like '%Success%')) group by "Landing _Outcome" order by (select count(*) from SPACEXTBL where (("Date" > '04-06-2010') & ("Date" < '20-03-2017') & ("Landing _Outcome" like '%Success%')) Group by "Landing _Outcome")
```

See the full sql EDA at : https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We added markers for each launch location of space X and Circles to identify better the locations of the launch sites. We also added a line indicating the shortest distances of launch sites to the coastline.

Please follow the URL below to interact with the maps.

See the full geographical analysis at
: https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We generated a dropdown menu to select between launch sites. The default options is to select all launch sites and show in a pie chart, the proportion of launches in each launch site. Also, we showed the success of launches with respect to the payload mass. We show all the launches if All sites are selected, but only the corresponding launches when a specific launch site is selected in the dropdown menu.

Predictive Analysis (Classification)

- We performed an analysis of several classification algorithms applied to our problem such as : a decission tree, knn-neighbours, Support Vector Machine and logistic regression. For each of these models we performed a cross-validation to fit the best set of hyperparameters from an initial set.
- We selected our best models based on the cross validation. We finally identified the decission tree as our best model , based on the accuracy returned by the cross validation process.

Full machine Learning notebook: https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

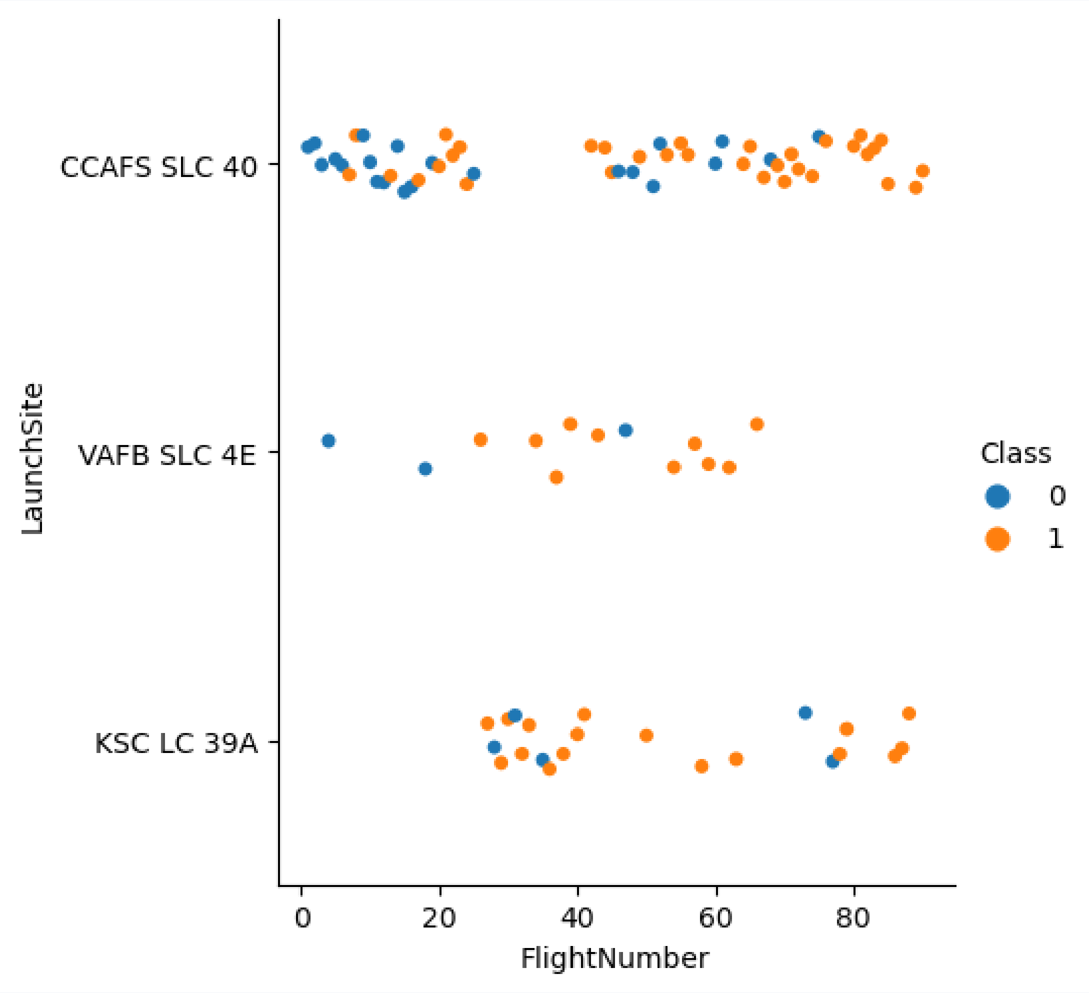
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

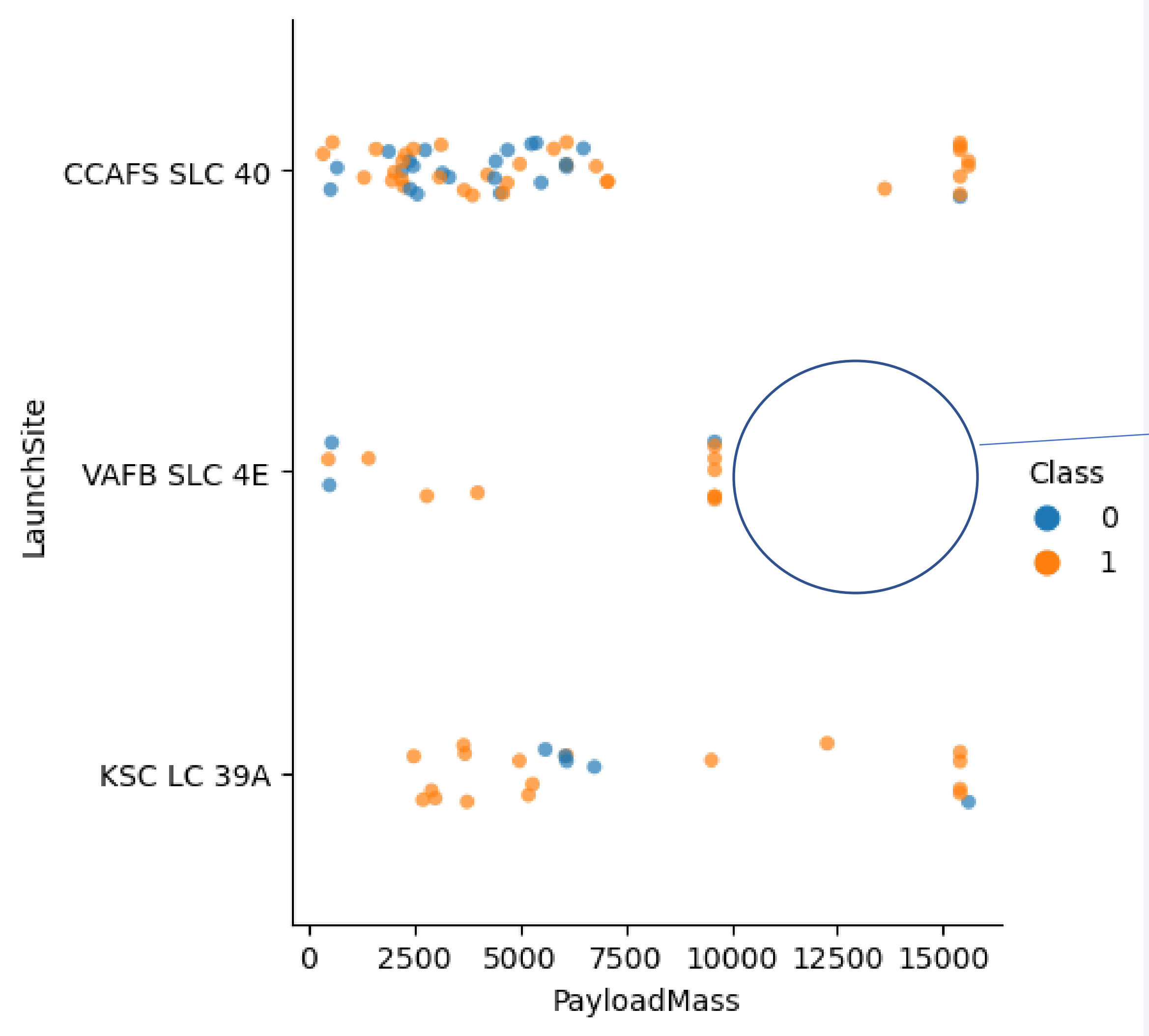
Flight Number vs. Launch Site

*Orange, means success. Blue means failure.



- **VAFB SLC 4E** and **KSC LC 39A** have a large success rate in launches (few blue dots)

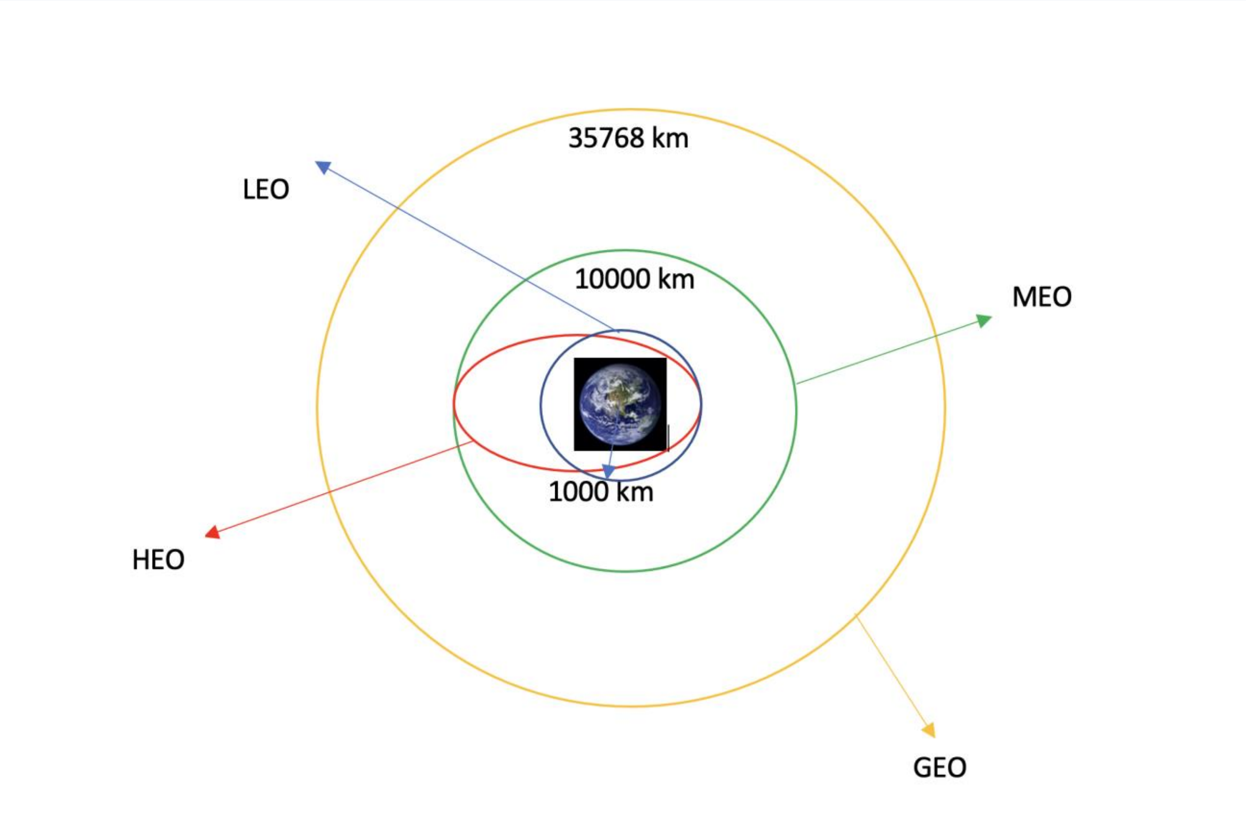
Payload vs. Launch Site

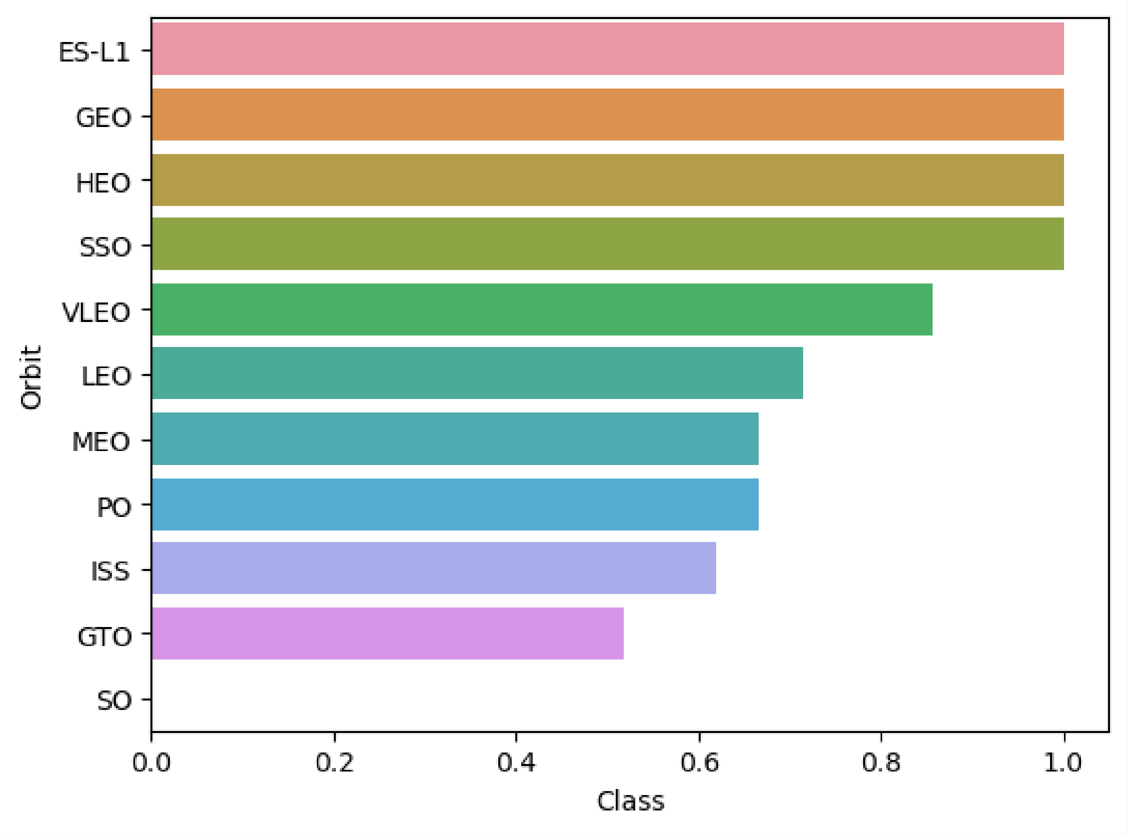


No heavy loaded launches (Mass > 10000) in SLC 4E

Success Rate vs. Orbit Type

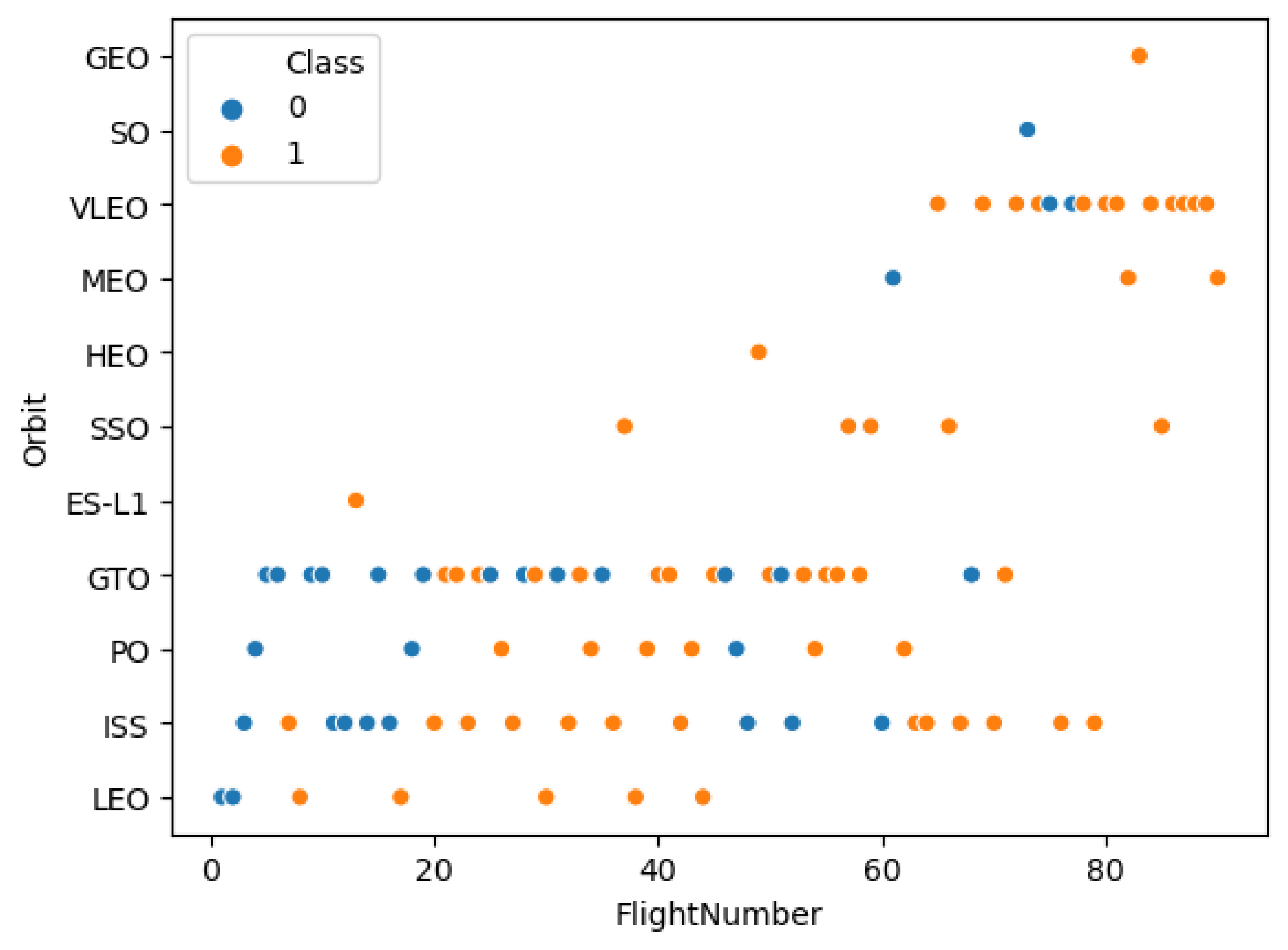
*Some of the types of orbits





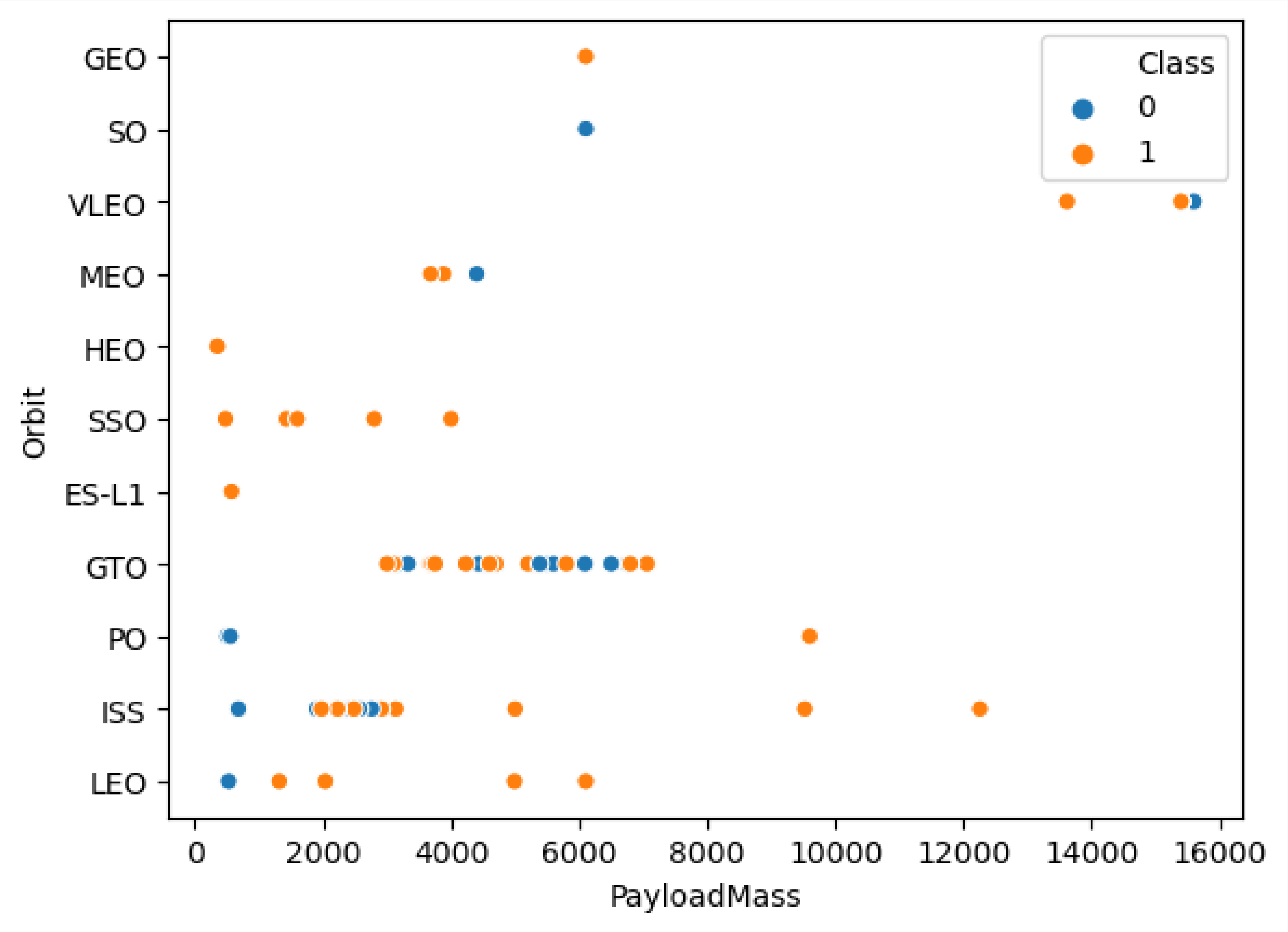
Lagrange point (ES-L1),Geosynchronous (GEO), Elliptical (HEO) and Sun Synchronous (SSO) orbits have a perfect success rate. Geosynchronous GTO type of orbit shows the worst recovery results, only 50%.

Flight Number vs. Orbit Type



The LEO orbit Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

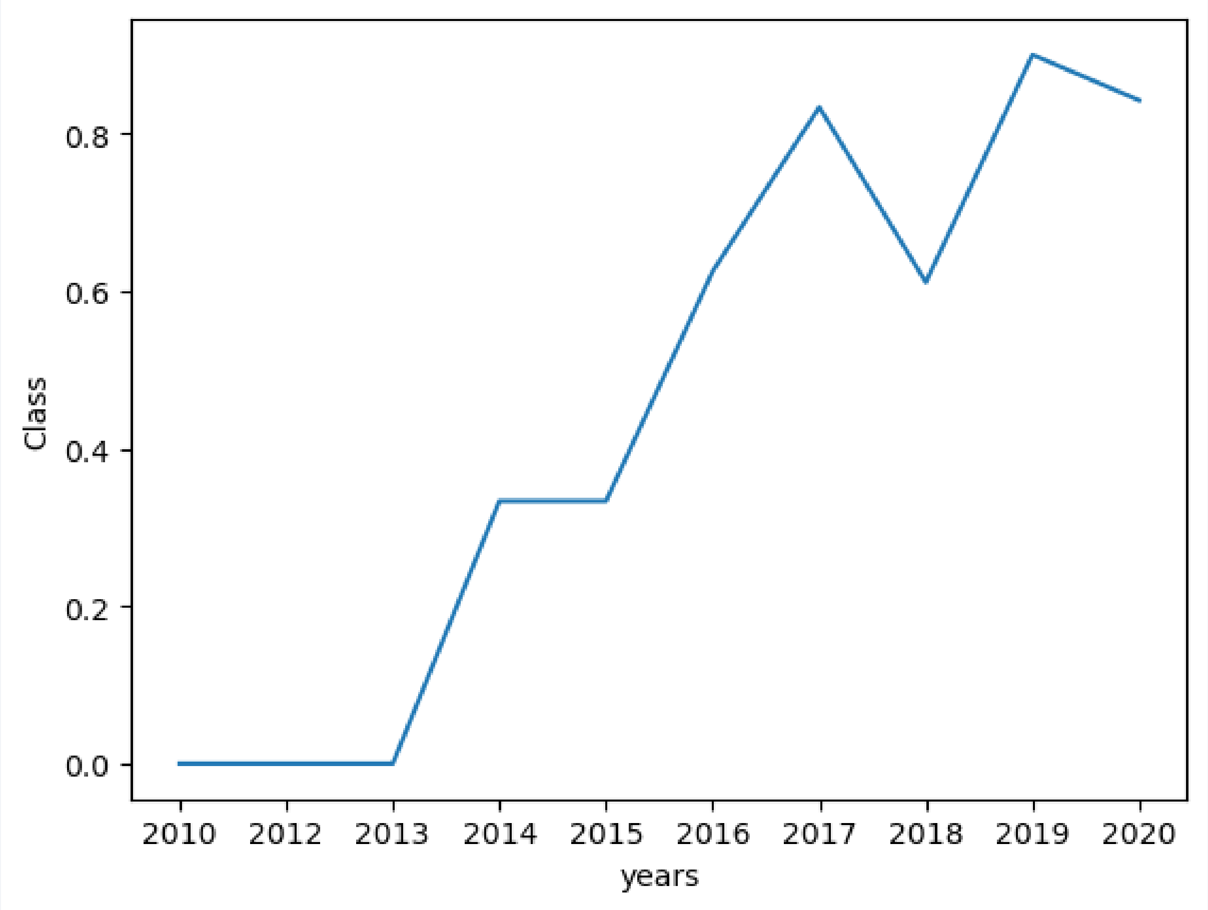


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

They are getting better!



See the full data visualization analysis at:
https://github.com/IgnacioGargiulo/IBMCapstone/blob/master/jupyter-labs-eda-dataviz.ipynb

All Launch Site Names

- Find the names of the unique launch sites

In order to find the unique launch sites we can use the unique method of dataframe objetcs in pandas.

Select DISTINCT "Launch_Site" from SPACEXTBL

Which gives as a result :

Launch_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

CCAFS SLC 40 was in the past CCAFS LC-40, so some launches are indicated to have been ocurred in CCAFS LC-40, but they actually represent the same place geographically.

Launch Site Names Begin with 'CCA'

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

- select "Launch_Site" from SPACEXTBL where "Launch_Site" LIKE "CCA%" LIMIT 5
- Launch_Site
- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40

Records that begin with CCA are the corresponding to CCAFS LC-40 and CCAFS SLC-40 .

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- select SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" == "NASA (CRS)"

SUM("PAYLOAD_MASS__KG_")
45596

We can isolate Nasa boosters using the Customer field, and obtain the sum of the total payload carried by their boosters.

Average Payload Mass by F9 v1.1

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

```
select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" == "F9 v1.1"
```

First Successful Ground Landing Date

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

```
select MIN("Date") from SPACEXTBL where "Landing _Outcome" LIKE "%Success%"
```

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

select "Booster_Version" from SPACEXTBL where (("Landing _Outcome" == "Success (drone ship)") & ("PAYLOAD_MASS__KG_" > 4000) & ("PAYLOAD_MASS__KG_" < 6000))

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

These four boosters achieved a posterior succesful landing on drone ships , and carrying moderately large payloads

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
SELECT "Landing _Outcome", SUM(case when "Landing _Outcome" like '%Success%' then 1 else 0 end) as success_landings, SUM(case when "Landing _Outcome" like '%Failure%' then 1 else 0 end) as failure_landings FROM SPACEXTBL GROUP BY "Landing _Outcome"
```

Landing _Outcome	success_landings	failure_landings
Controlled (ocean)	0	0
Failure	0	3
Failure (drone ship)	0	5
Failure (parachute)	0	2
No attempt	0	0
No attempt	0	0
Precluded (drone ship)	0	0
Success	38	0
Success (drone ship)	14	0
Success (ground pad)	9	0
Uncontrolled (ocean)	0	0

Boosters Carried Maximum Payload

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

```
select DISTINCT "Booster_Version"
	from SPACEXTBL
	where (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

- F9 v1.0 B0003F9 v1.0 B0004F9 v1.0 B0005F9 v1.0 B0006F9 v1.0 B0007F9 v1.1 B1003F9 v1.1F9 v1.1 B1011F9 v1.1 B1010F9 v1.1 B1012F9 v1.1 B1013F9 v1.1 B1014F9 v1.1 B1015F9 v1.1 B1016F9 v1.1 B1018F9 FT B1019F9 v1.1 B1017F9 FT B1020F9 FT B1021.1F9 FT B1022F9 FT B1023.1F9 FT B1024F9 FT B1025.1F9 FT B1026F9 FT B1029.1F9 FT B1031.1F9 FT B1030F9 FT B1021.2F9 FT B1032.1F9 FT B1034F9 FT B1035.1F9 FT B1029.2F9 FT B1036.1F9 FT B1037F9 B4 B1039.1F9 FT B1038.1F9 B4 B1040.1F9 B4 B1041.1F9 FT B1031.2F9 B4 B1042.1F9 FT B1035.2F9 FT B1036.2F9 B4 B1043.1F9 FT B1032.2F9 FT B1038.2F9 B4 B1044F9 B4 B1041.2F9 B4 B1039.2F9 B4 B1045.1F9 B5 B1046.1F9 B4 B1043.2F9 B4 B1040.2F9 B4 B1045.2F9 B5B1047.1F9 B5B1048.1F9 B5 B1046.2F9 B5B1049.1F9 B5 B1048.2F9 B5 B1047.2F9 B5 B1046.3F9 B5B1050F9 B5B1054F9 B5 B1049.2F9 B5 B1048.3F9 B5B1051.1F9 B5B1056.1F9 B5 B1049.3F9 B5 B1051.2F9 B5 B1056.2F9 B5 B1047.3F9 B5 B1048.4F9 B5B1059.1F9 B5 B1056.3F9 B5 B1049.4F9 B5 B1046.4F9 B5 B1051.3F9 B5 B1056.4F9 B5 B1059.2F9 B5 B1048.5F9 B5 B1051.4F9 B5B1058.1F9 B5 B1049.5F9 B5 B1059.3F9 B5B1060.1F9 B5 B1058.2F9 B5 B1051.5F9 B5 B1049.6F9 B5 B1059.4F9 B5 B1060.2F9 B5 B1058.3F9 B5 B1051.6F9 B5 B1060.3F9 B5B1062.1F9 B5B1061.1F9 B5B1063.1F9 B5 B1049.7F9 B5 B1058.4

2015 Launch Records

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

select substr(Date, 4, 2) as month ,"Landing _Outcome", "Booster_Version", "Launch_Site" from SPACEXTBL where ((substr(Date,7,4) == '2015')& ("Landing _Outcome" == "Failure (drone ship)"))

Month	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:

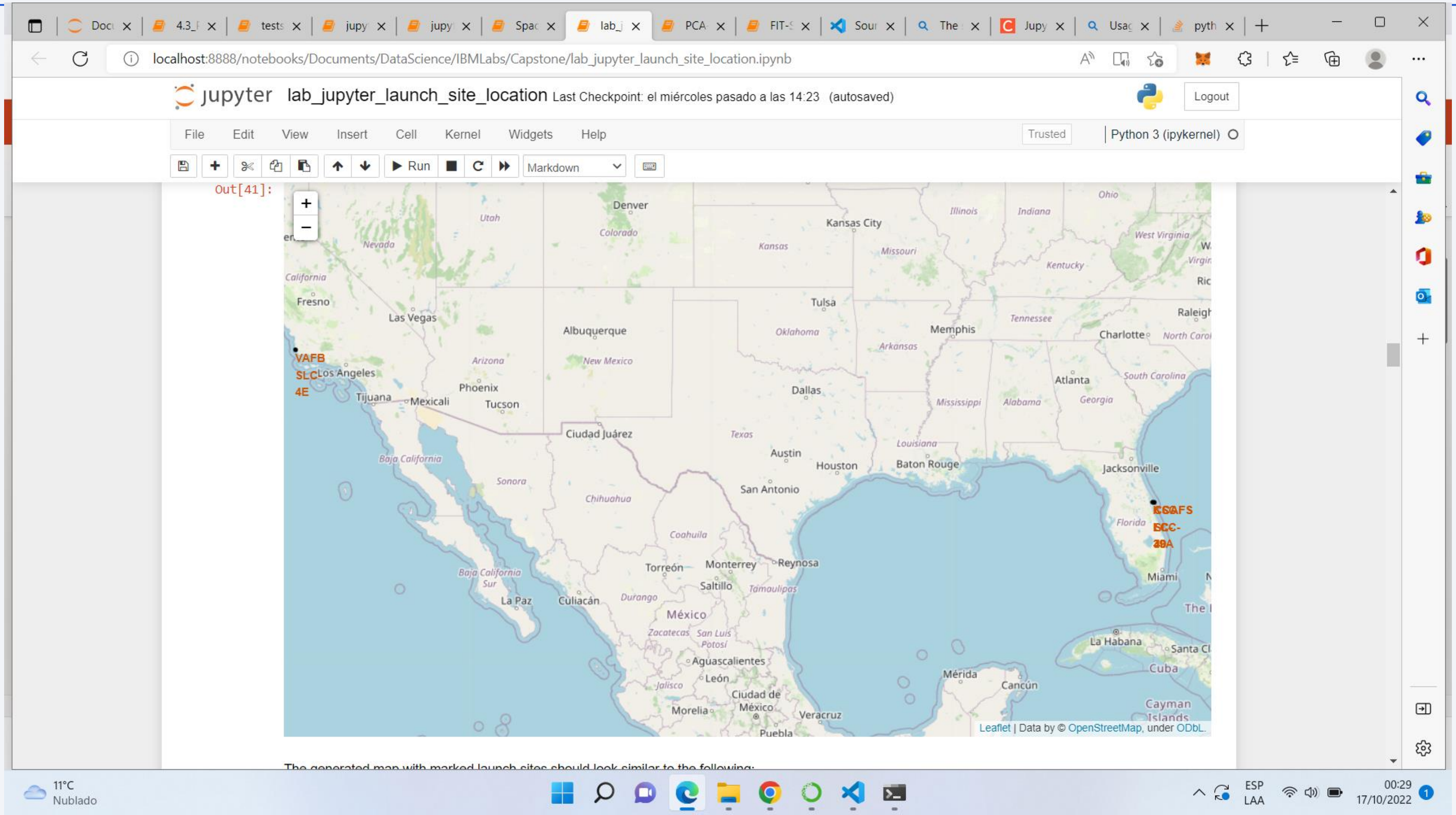
```
select "Landing _Outcome", count(*) from SPACEXTBL
where (("Date" > '04-06-2010') & ("Date" < '20-03-2017') & ("Landing _Outcome" like '%Success%'))
group by "Landing _Outcome"
order by (select count(*) from SPACEXTBL
        where (("Date" > '04-06-2010') & ("Date" < '20-03-2017') & ("Landing _Outcome" like '%Success%'))
        Group by "Landing _Outcome")
```

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

Section 3

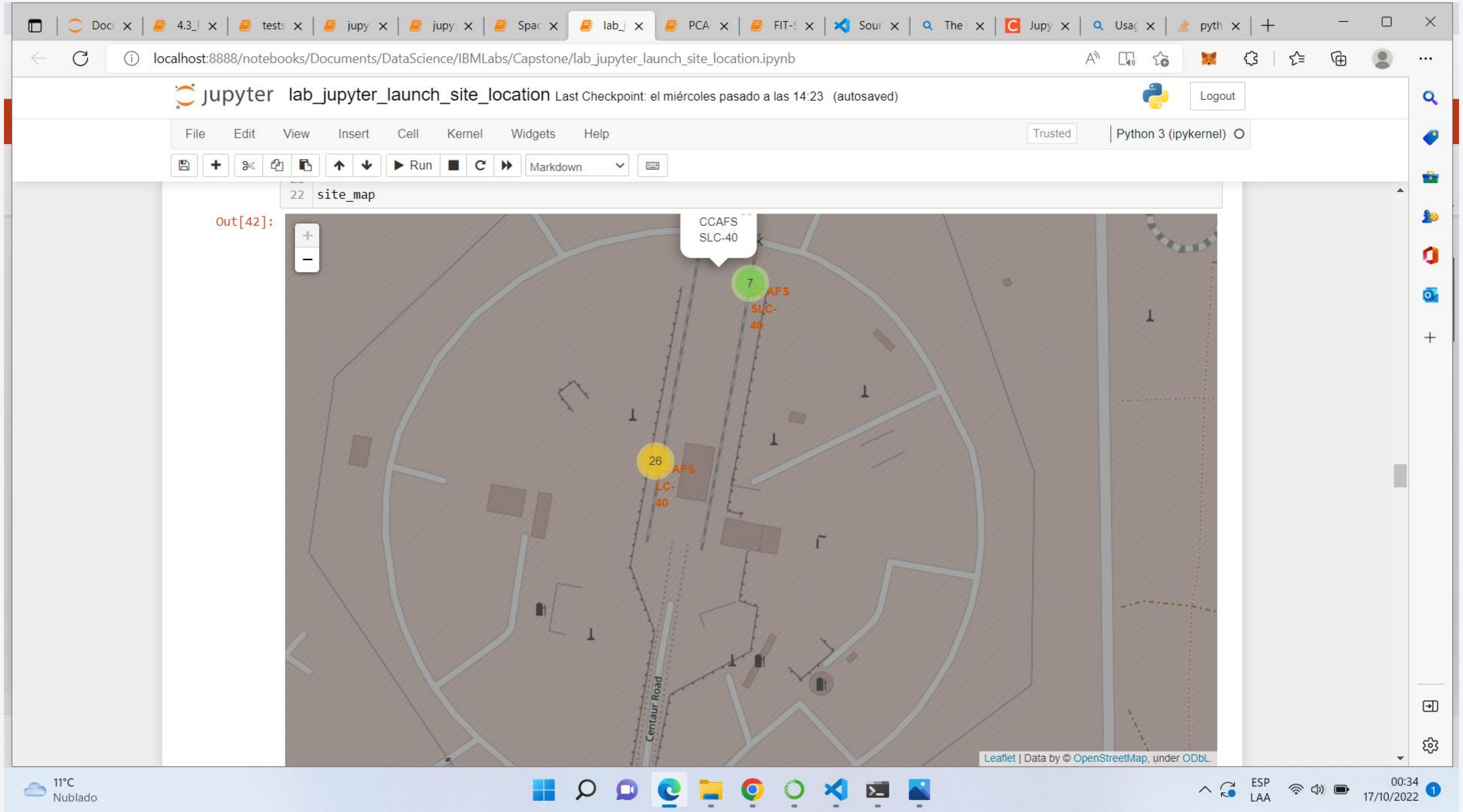
Launch Sites Proximities Analysis

Map of SpaceX Launch Sites



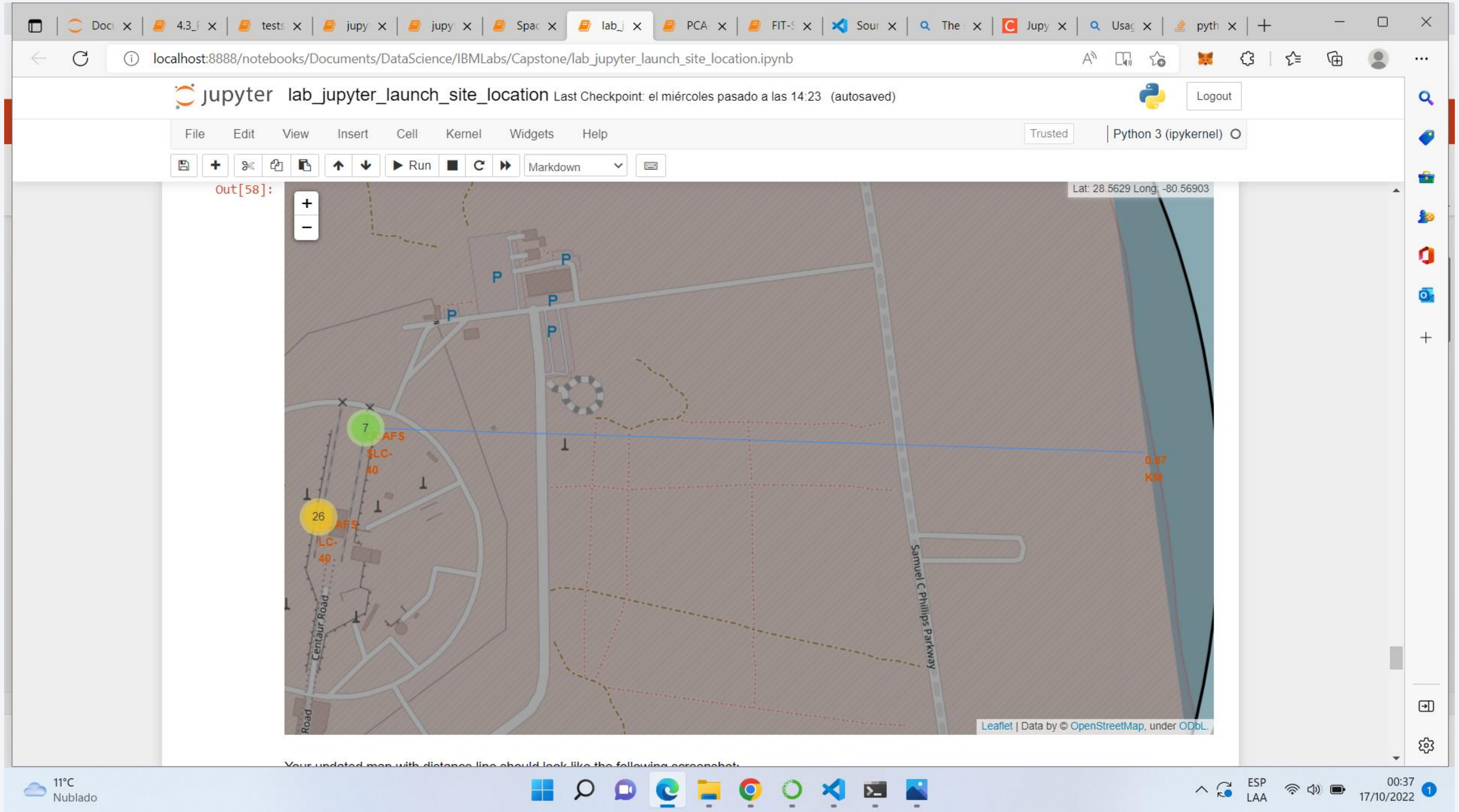
Two of the launch sites of Space X are in the East coast and one of them () is in the west coast.

Succesful and failed landings from Launch Sites



- Rockets launched from CCAFS SLC-40 shows a higher rate of succesful landings

Distance of Launch sites to the nearest coastline

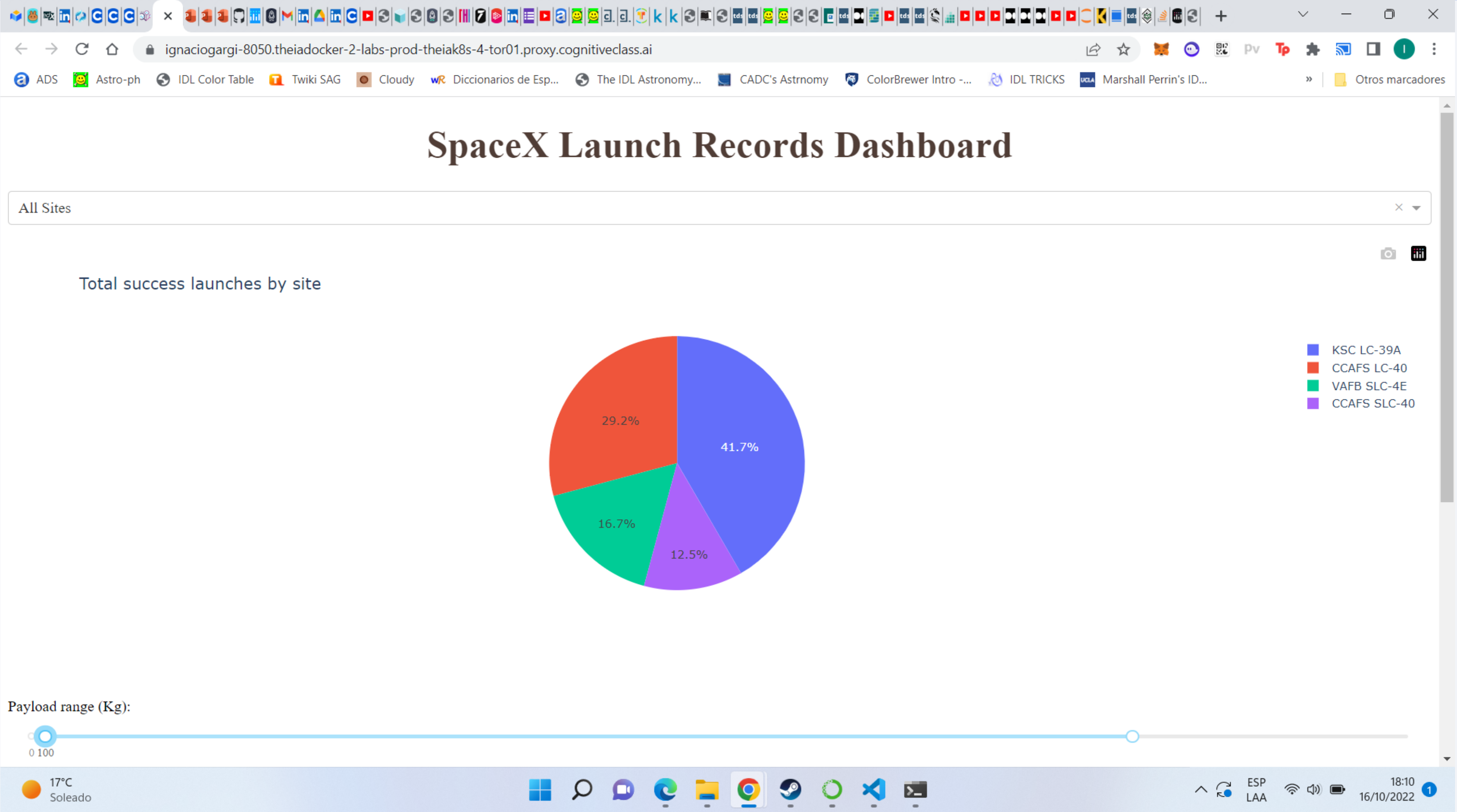


CCAFS SLC-40 is less than 1 km away from the coastline.

Section 4

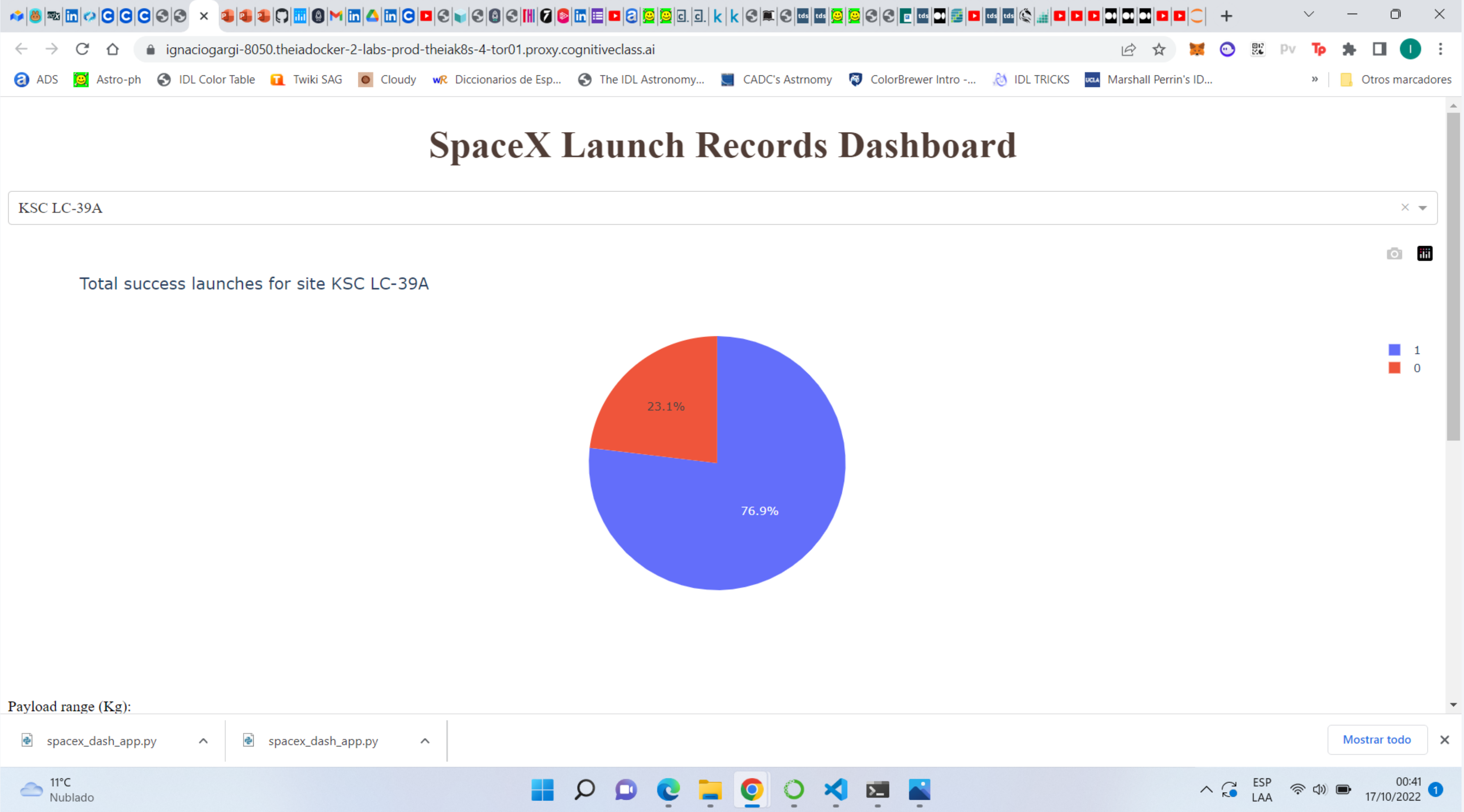
Build a Dashboard with Plotly Dash

Recovery Success per launch site



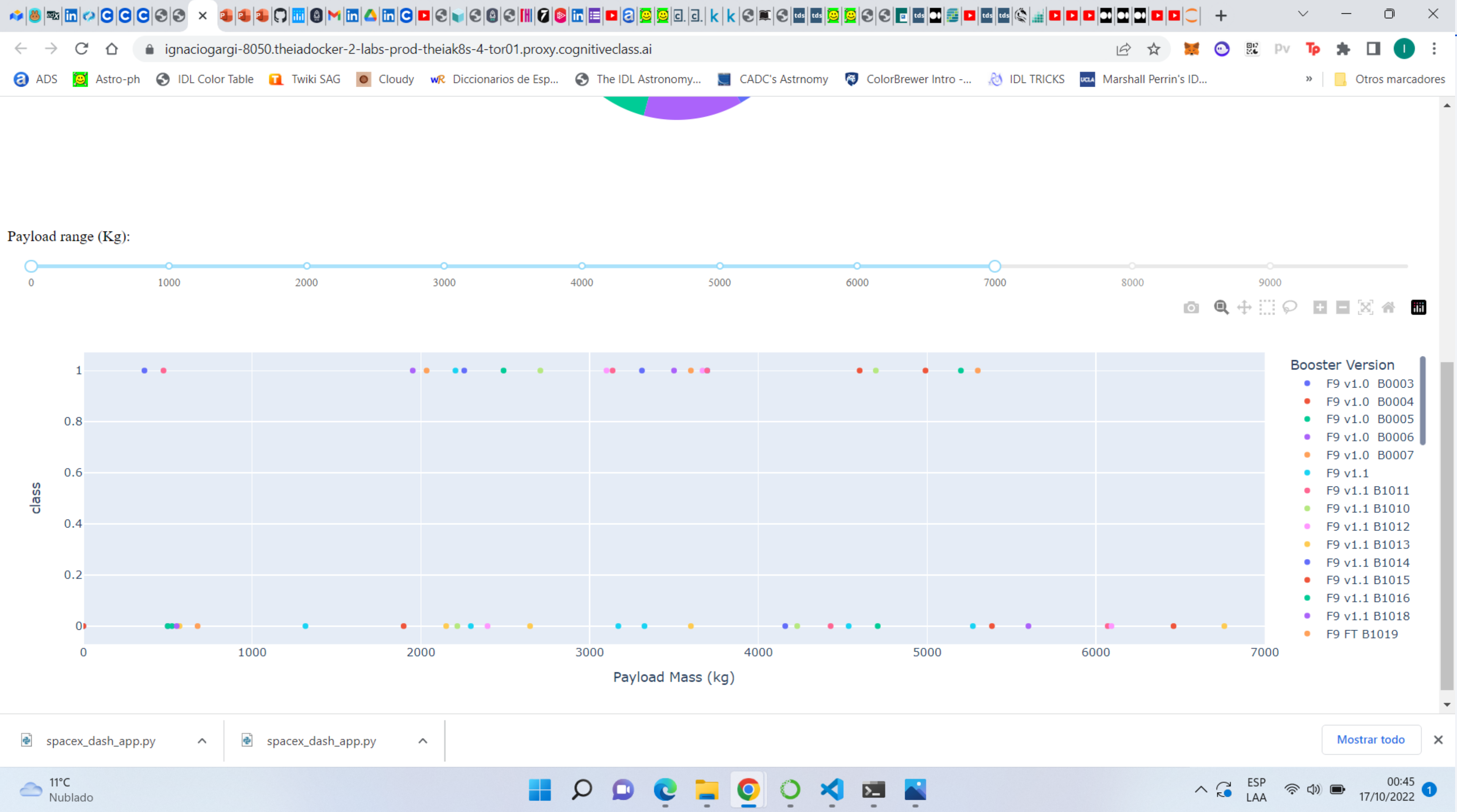
KSC KC 39A has the largest number of success launches, reaching almost 42 % of all the success launches.

The most effective Launch Site



- The Launch site with the highest launch success number has a 77% of successful launches.

Success as a function of Payload

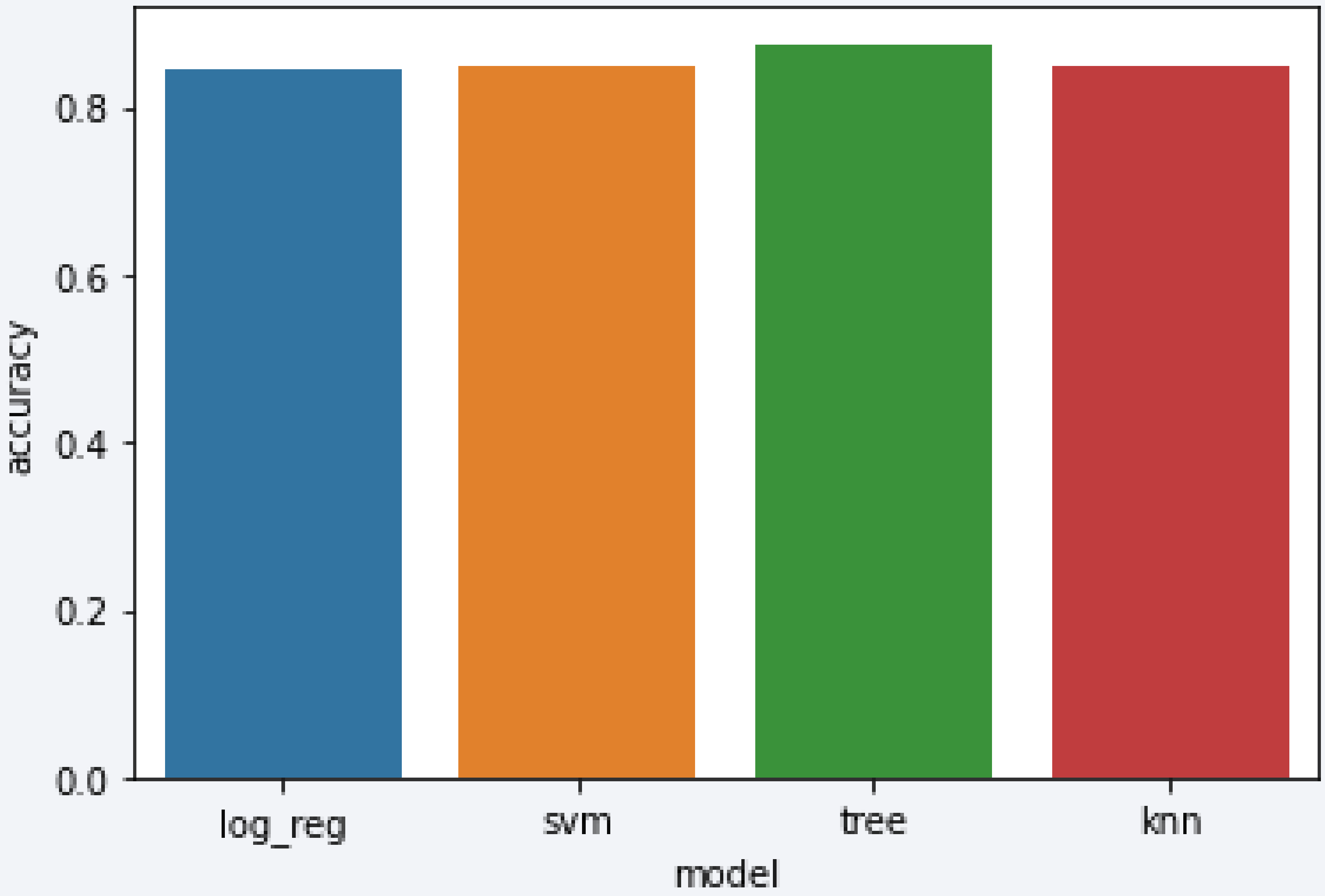


Rockets with payloads above 5500 are more commonly not successful, although 3 B4 boosters did it

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Of the 4 models we tried, the decission tree reached the highest accuracy.

Confusion Matrix

- The confusion matrix of our best model shows no errors in the prediction of succesfully landed stages. However , the model incorrectly predicts 3 cases as succesful, that weren't.

Conclusions

- Space X is recovering more stages as times goes by.
- Heavy loaded launches are more prone to fail, although this result might be due to the low number statistics.
- Launches in the KSC LC-39A Launch site are more succesful.

- We can predict with a high accuracy and low error the launches that did not land

Succesfully. We can predict with an accuracy of around 86% the succesful landings.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!