IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andre R S Pereira
June 11, 2024

Outline

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

#EXTERNO.CONFIDENCIAL

Executive Summary

- Space Exploration Technologies Corporation, commonly referred to as SpaceX, is an American spacecraft manufacturer, launch service provider and satellite communications company headquartered in Hawthorne, California. The company was founded in 2002 by Elon Musk with the goal of reducing space transportation costs and ultimately developing a sustainable colony on Mars. The company currently produces and operates the Falcon 9 and Falcon Heavy rockets along with the Dragon and Starship spacecraft.The company offers internet service via its Starlink subsidiary, which became the largest-ever satellite constellation in January 2020 and, as of April 2024, comprising more than 6,000 small satellites in orbit.
- I am a data scientist and SpaceX hired me to carry out an exploratory analysis of the company's database and, using artificial intelligence technology, machine learning and other tools, I could carry out cost forecasts and be prepared to face an imminent attack. competitors in the first stage rocket landing.

#EXTERNO.CONFIDENCIAL

Introduction

- It's not new today, more precisely since the 1950s, countries have been fighting for space in the universe.Launching satellites and rockets for space exploration and other activities require millions of dollars from companies and governments, one of the biggest expenses being the return performance of the first stage of the rocket structure, such as a successful landing.
- My job comes in at this point, collecting all the data and information that control the issue of the first landing such as the launch process, payload mass, launch site, orbit and reaching conclusions such as whether it is possible to predict the success of the landing in the first stage, which would drastically reduce the costs of this process.s

#EXTERNO.CONFIDENCIAL

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The SpaceX datasets were collected through an API provided by the company, and also from internet sources such as the Wikepedia website, using web scraping technology.

- Perform data wrangling

  Using the Python programming language and libraries such as Pandas and Numpy, the data was pre-processed, as well as cleaning, normalization and standardization, along with techniques for transforming categorical data into numeric data.

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

  Use of SQL queries (Structured Query Language) in databases and for visualization and stories, libraries such as Seabornand Matplotlib.

#EXTERNO.CONFIDENCIAL

Methodology

Executive Summary

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

  Use of SQL queries (Structured Query Language) in databases and for visualization and stories, libraries such as Seabornand Matplotlib.

- Perform interactive visual analytics using Folium and Plotly Dash

  Folium Map In Plotly Dash Dashboards were used

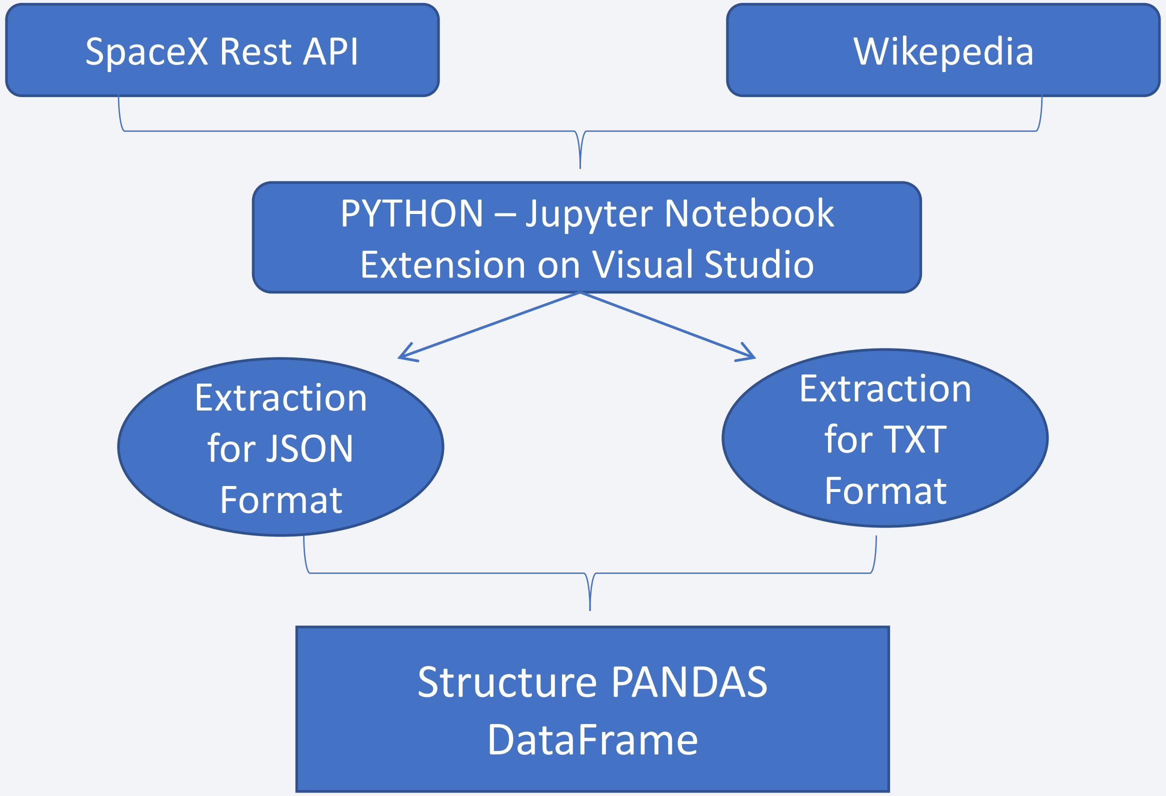
- Perform predictive analysis using classification models

  Using the Scikit Learn library, the data was divided into training and test sets, and after using numerous algorithms and making hyperparameter adjustments in each of them, it was possible to determine the best performance for the objective of the work.

#EXTERNO.CONFIDENCIAL

Data Collection

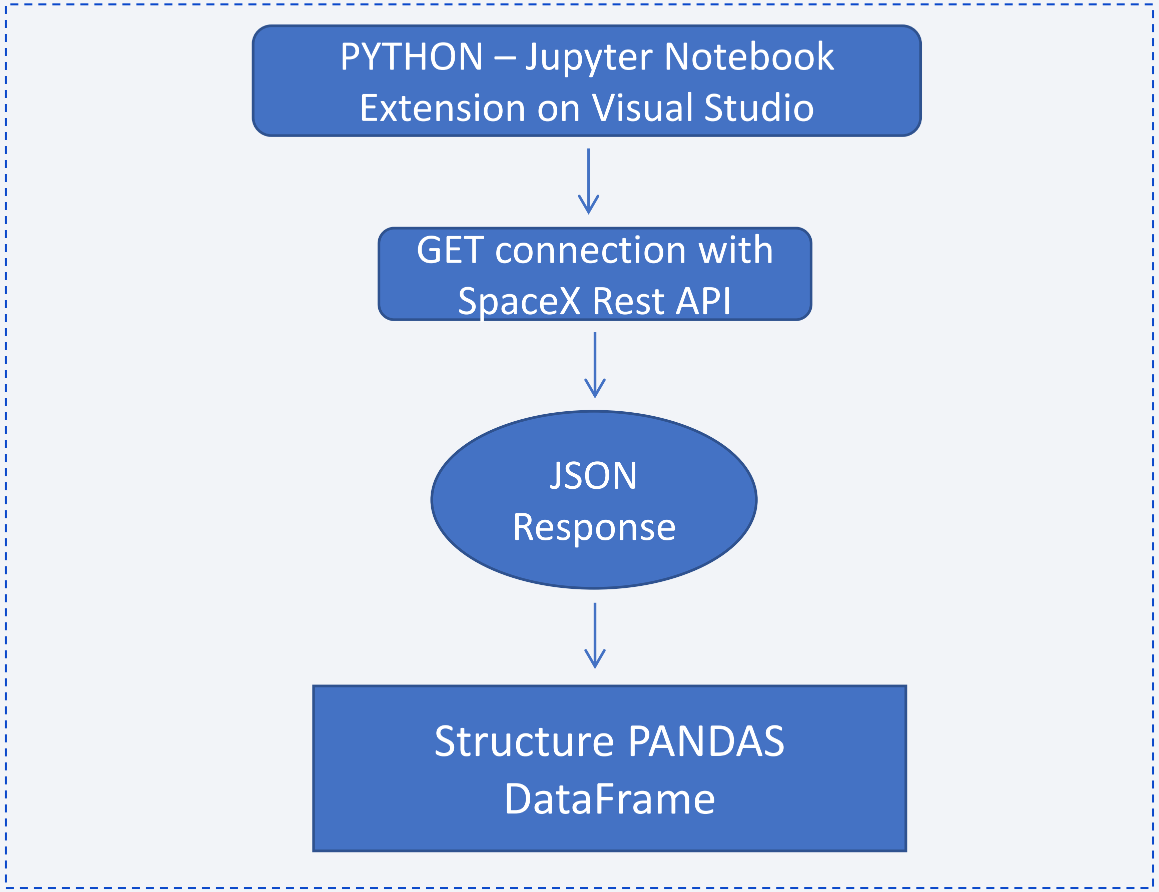
- API: Data obtained for launch, rocket, core, capsule, starlink, launch pad, and landing pad data from SpaceX's open source REST API.
- Internet sites: data collected from Wikipedia articles.


SpaceX Rest API
Wikepedia
PYTHON – Jupyter Notebook Extension on Visual Studio
Extraction for JSON Format
Extraction for TXT Format
Structure PANDAS DataFrame

Data Collection – SpaceX API

- A direct connection was made to the SpaceX API, and the URL GET method was used to capture the data in JSON file format (geo info, rocket type, orbit, flights, etc.). The treatment was carried out in PYTHON for subsequent transformation into a table (DataFrame PANDAS).

- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of SpaceX API.

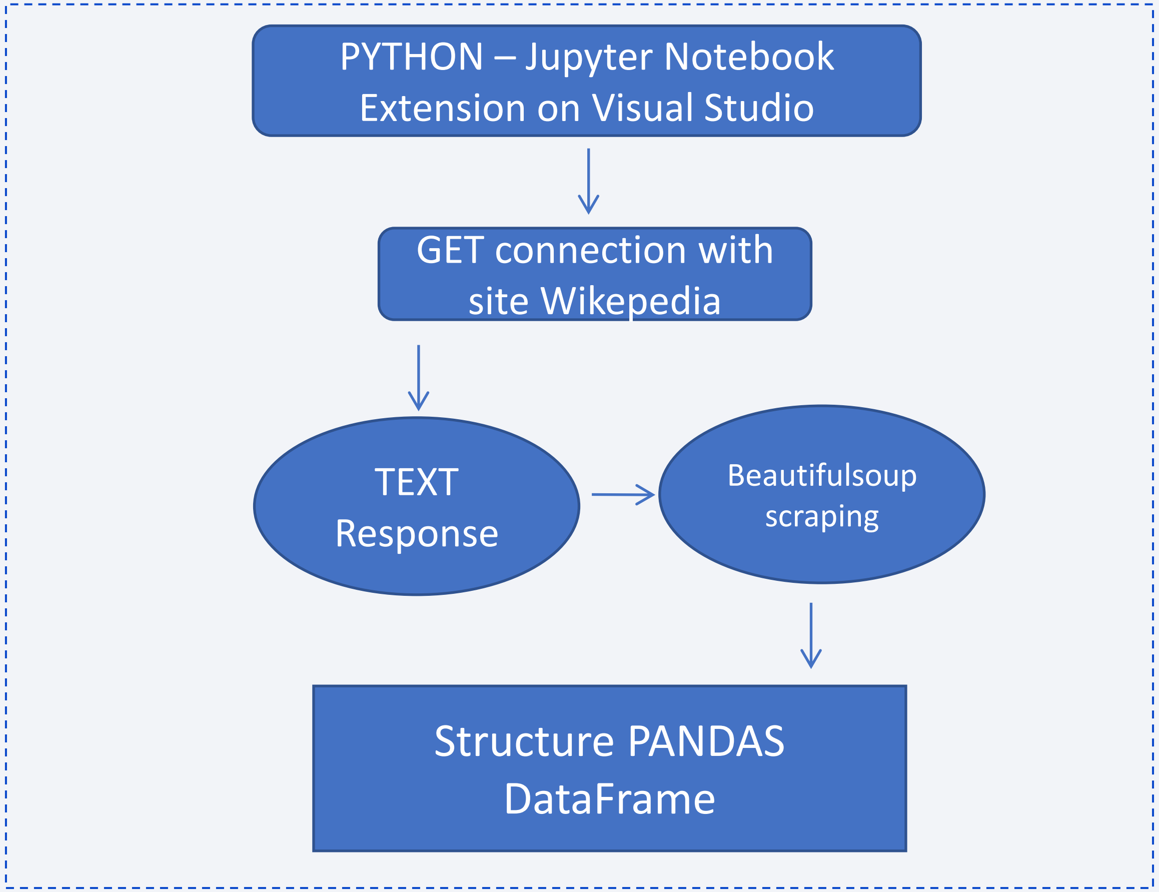

PYTHON – Jupyter Notebook Extension on Visual Studio
GET connection with SpaceX Rest API
JSON Response
Structure PANDAS DataFrame

#EXTERNO.CONFIDENCIAL

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of Scraping.


PYTHON – Jupyter Notebook Extension on Visual Studio
GET connection with site Wikepedia
TEXT Response
Beautifulsoup scraping
Structure PANDAS DataFrame

Data Wrangling

- At this stage of the process and with the databases extracted in table format, it was verified that there is diverse information on successful and unsuccessful landings in the first stage. It will convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- Steps: loading the collected dataset, identifying missing values, treating categorical columns as numerical, obtaining the number of launches on each site, number and occurrence of each orbit, creating a landing outcome label and determining the success rate of returning the first stage of the rocket.
- Then, select the best resources to train with a machine learning model.
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of Scraping.

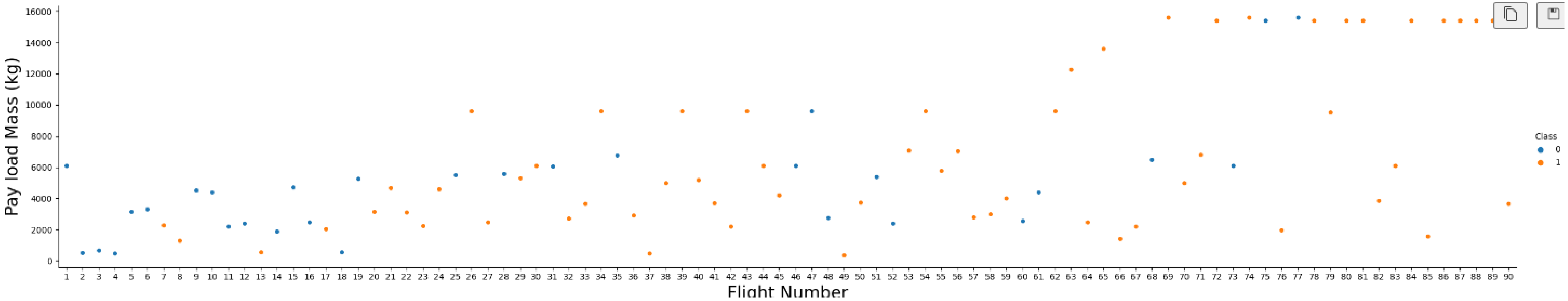
#EXTERNO.CONFIDENCIAL

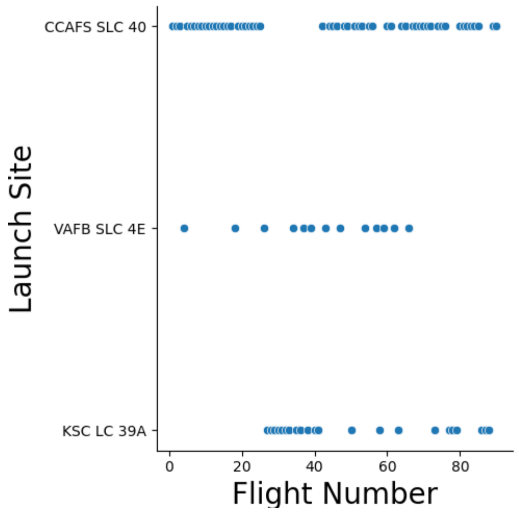
EDA with Data Visualization

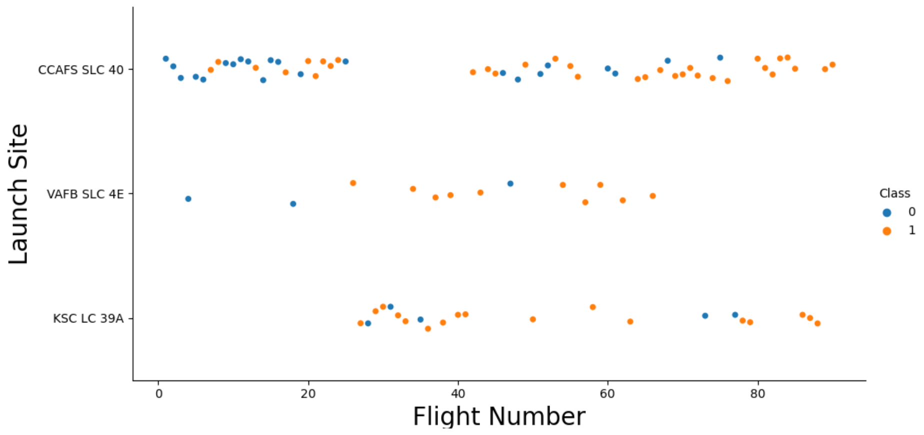
- At this stage of the work, EDA (Exploratory Data Analysis) is to complete the list of functionalities with the target, using various data science tools, telling stories with visualization with graphics and transforming data into information.
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of EDA.

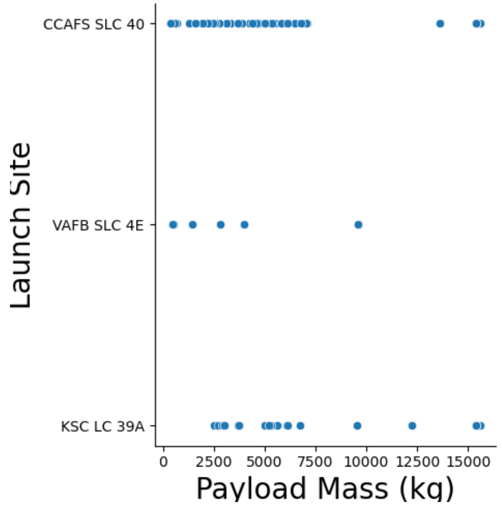
#EXTERNO.CONFIDENCIAL

EDA with Data Visualization



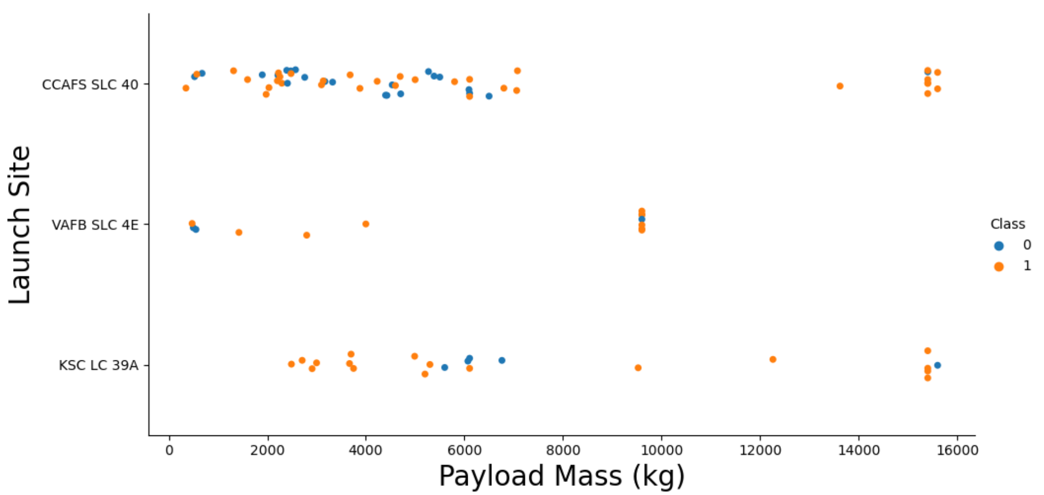


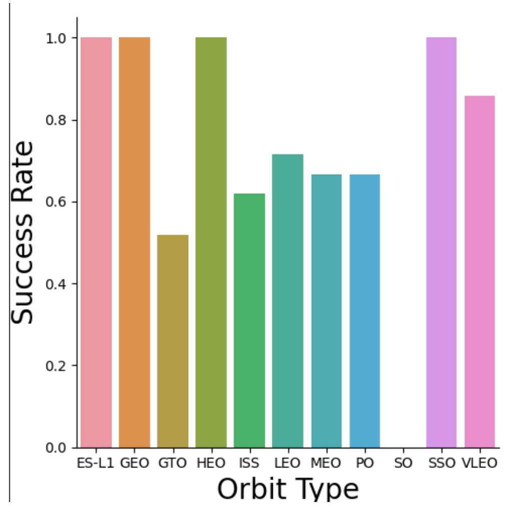


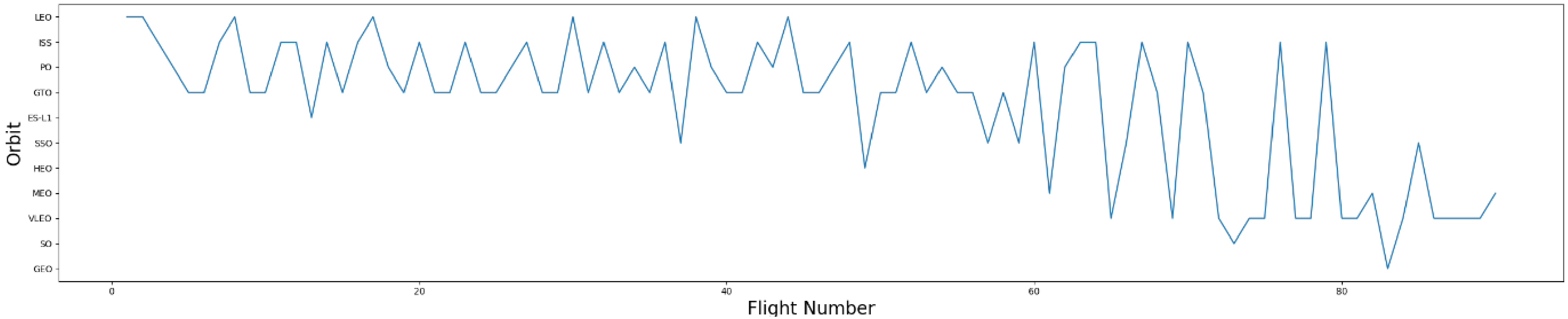


#EXTERNO.CONFIDENCIAL

EDA with Data Visualization

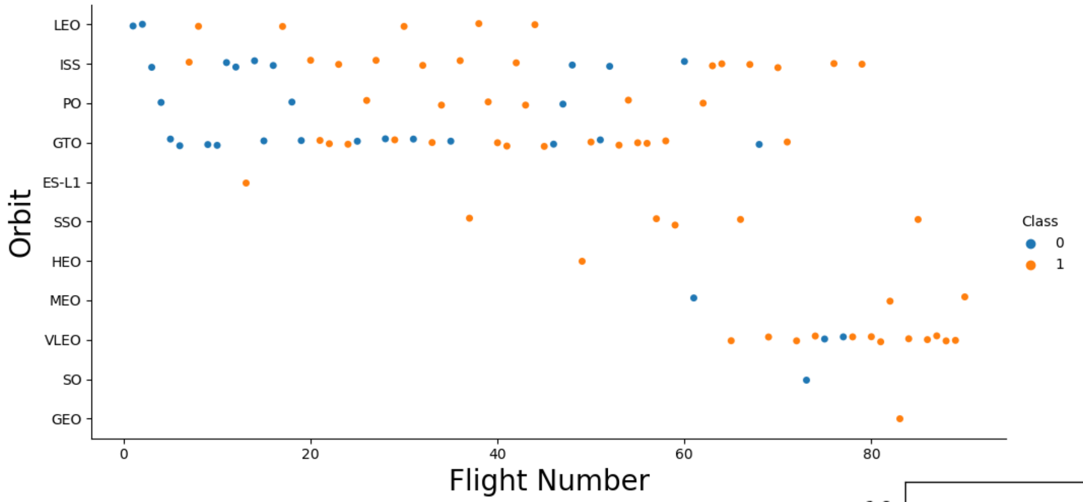


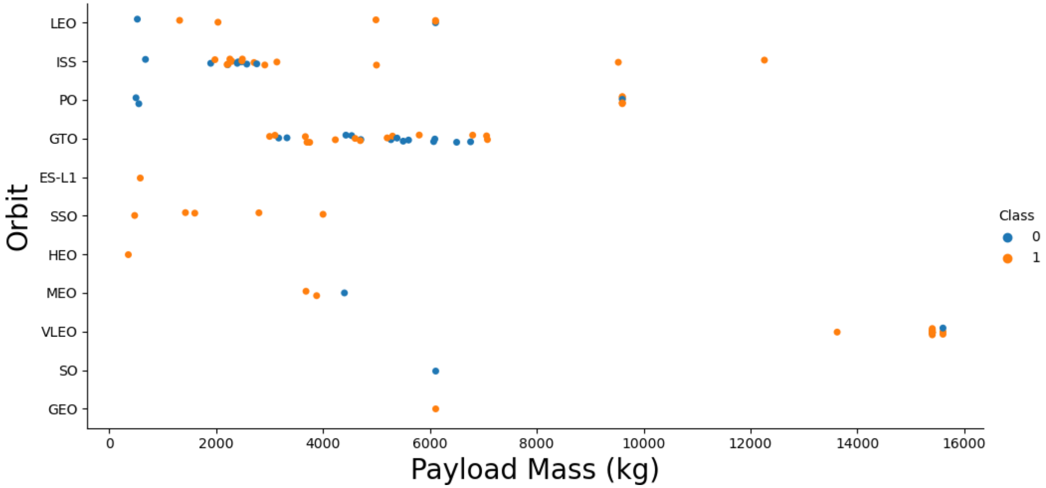


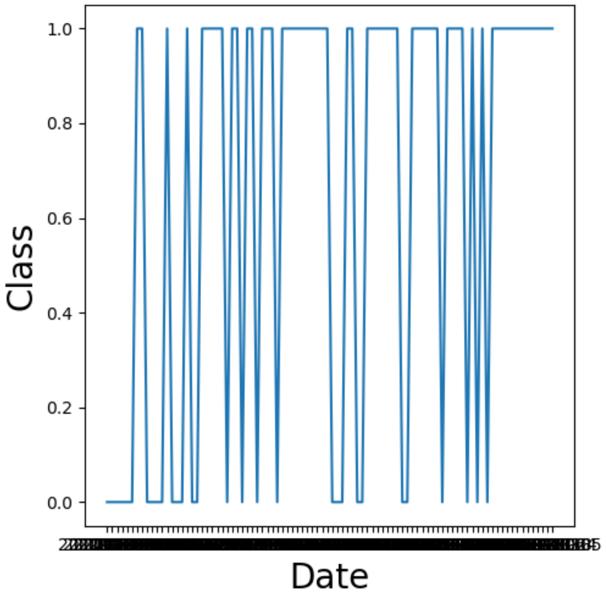


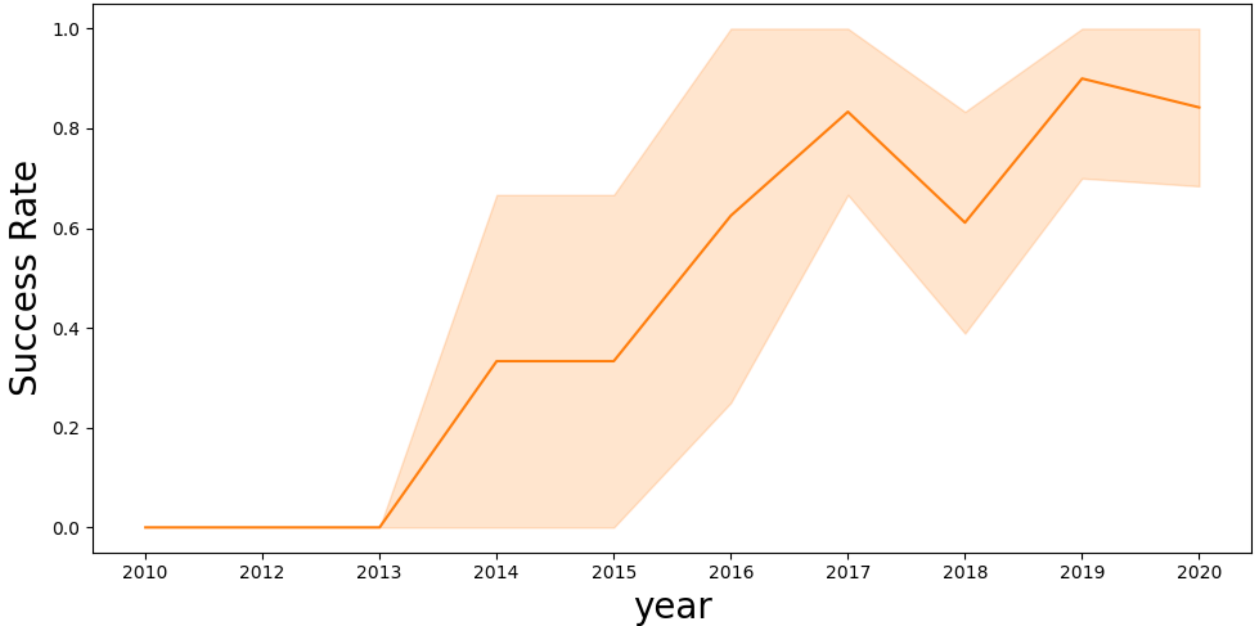
#EXTERNO.CONFIDENCIAL

EDA with Data Visualization









EDA with SQL

- To complete the data set, data was obtained through advanced SQL queries, which includes a record for each payload transported during a SpaceX mission to outer space.
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of SQL queries.

#EXTERNO.CONFIDENCIAL

Build an Interactive Map with Folium

- At this stage of the process, the Folium library is used to represent geospatial data, making drawing marks, circles and lines as markers on an interactive map.
- Calculated distances on interactive maps
- Plot coordinates and cluster marks
- Interactive map to analyze proximity to the launch site
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of Folium interactive maps.

#EXTERNO.CONFIDENCIAL

Build a Dashboard with Plotly Dash

- Interactive dashboard generation that contains pie charts and scatter plots to analyze data and stories.
- Created panel to analyze launch records interactively with Dash technology.
- Through the browser with the localhost address, it is possible to interact with the Dashboard and interactive graphs.
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of Folium interactive maps.

#EXTERNO.CONFIDENCIAL

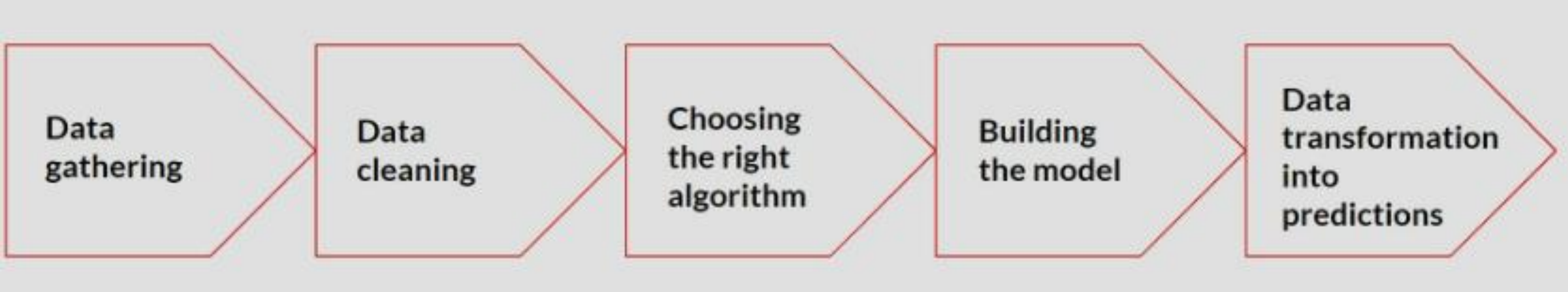
Predictive Analysis (Classification)

- With the data collected, processed and filtered with the appropriate features and target, it was decided that the machine learning model to be used will be classified, with 1 being success and 0 being failure in terms of landing the first stage of the rocket.
- Performing the division of data sets between test and training and 4 classification algorithms: LR (Logistic Regression), SVM (Support Vector Machine), DT (Decision Tree) and KNN (Nearest Neighbors).
- Used hyperparameter changes to find better performances and performance analysis techniques such as Confusion Matrix, F1 Score and Jaccard Score.
- Click here for GitHub URL to get the complete PYTHON file (.ypynb Jupyter Notebook format) of Predictive Analysis.

#EXTERNO.CONFIDENCIAL

Predictive Analysis (Classification)

Machine Learning Pipelines



Data gathering

Data cleaning

Choosing the right algorithm

Building the model

Data transformation into predictions

#EXTERNO.CONFIDENCIAL

Results

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

#EXTERNO.CONFIDENCIAL

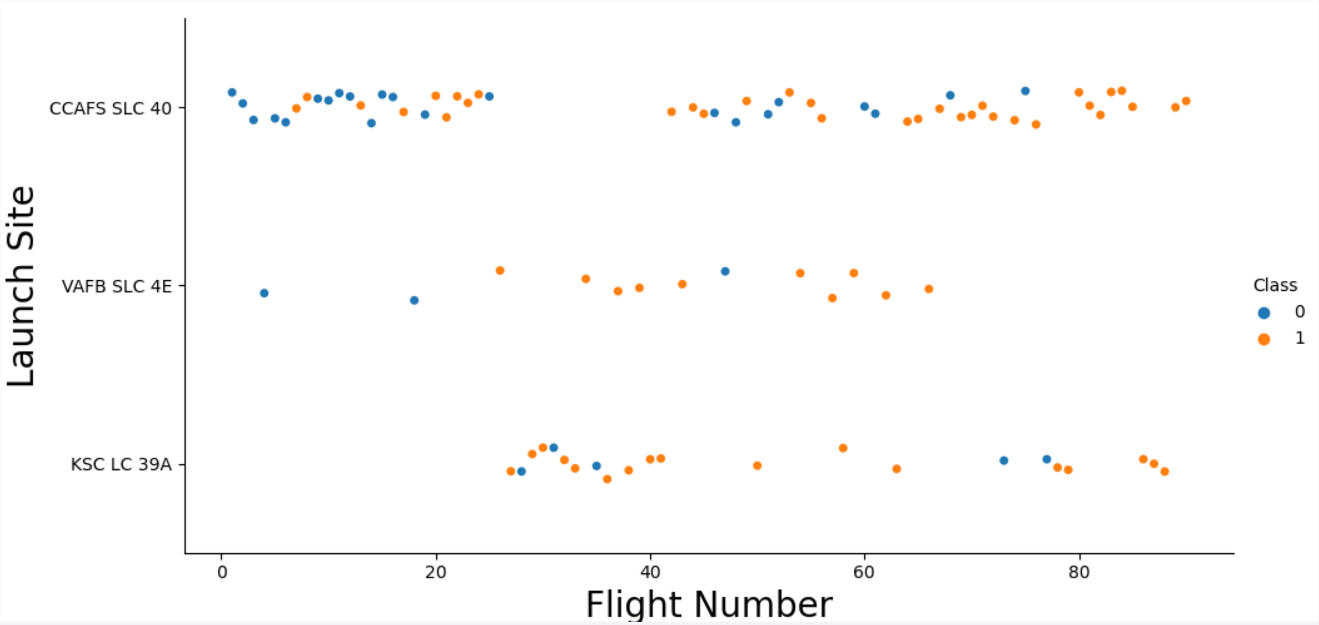
Section 2

Insights drawn from EDA

#EXTERNO.CONFIDENCIAL

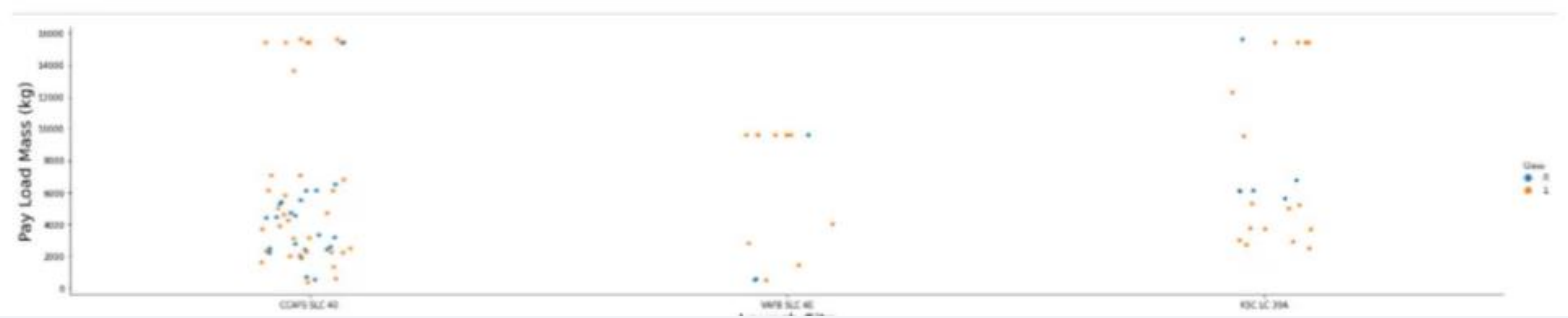
Flight Number vs. Launch Site

- It was found that most launches are launched from CCAFS-SLC-40
- It was found that fewer launches from the VAFB SLC 4E website
- A previous flight launches were from site CCAFS-SLC-40, followed by KSC-LC-39A



#EXTERNO.CONFIDENCIAL

Payload vs. Launch Site



- The VAFB SLC 4E was found to have low payload launches

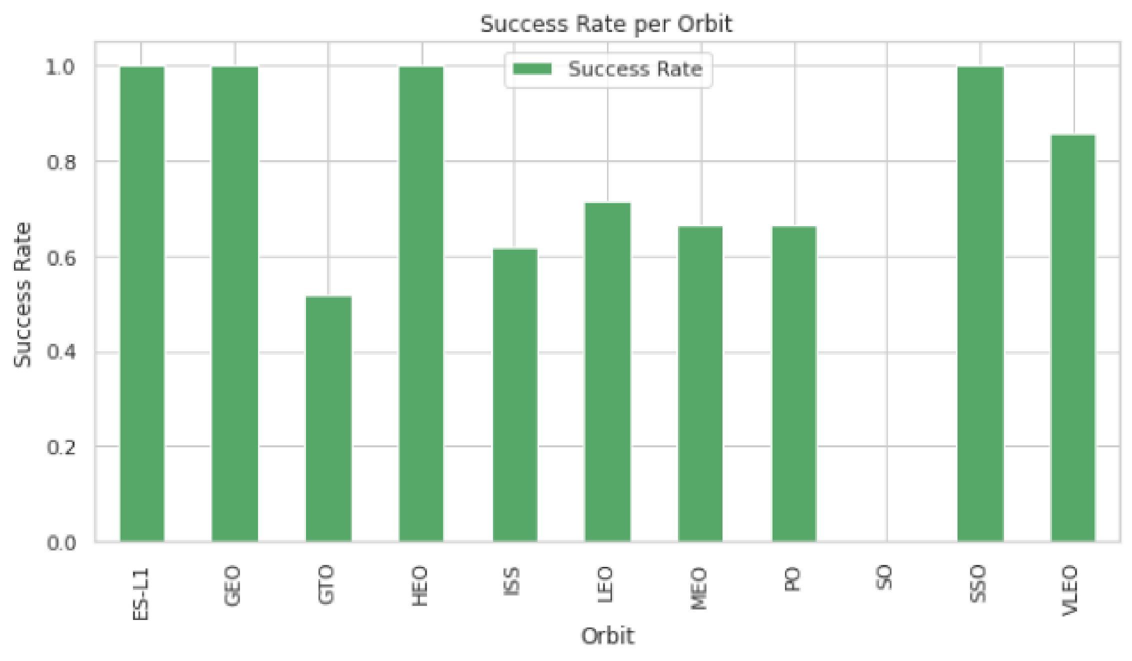
- CCAFS SLC 40 was found to have more higher payload launches and lower payload launches

#EXTERNO.CONFIDENCIAL

Success Rate vs. Orbit Type

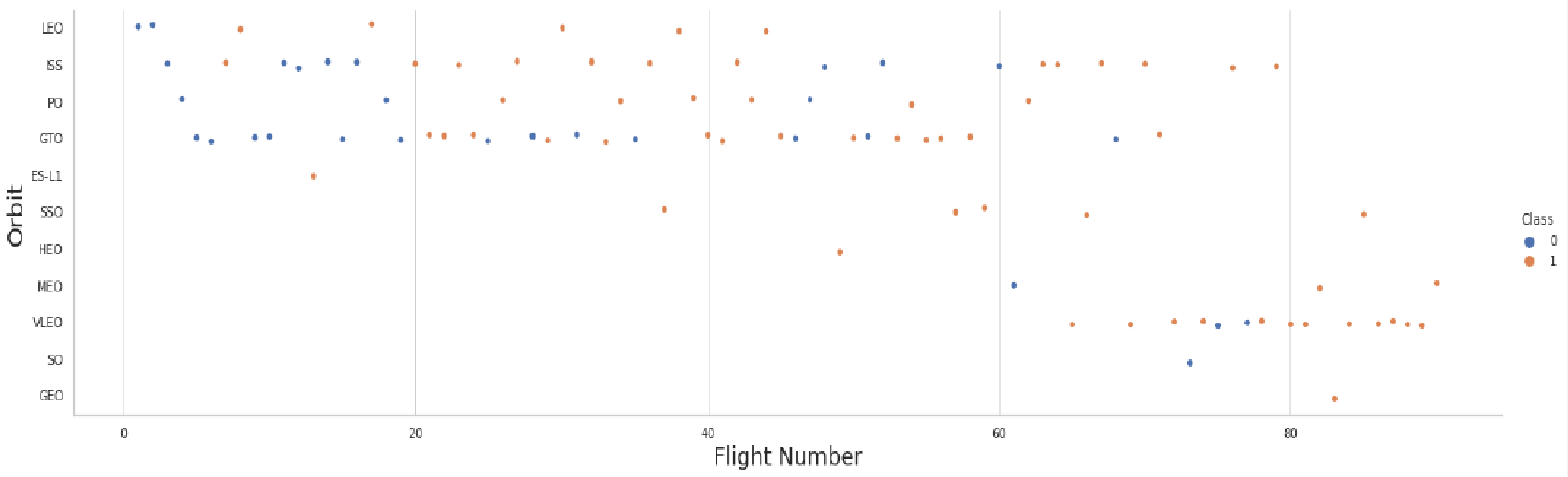
- GEO,HEO & ES-L1,SS) have high success rate.

- Now, when the orbit is GTO, there are many failures and it is worth delving into the reasons why.



#EXTERNO.CONFIDENCIAL

Payload vs. Orbit Type

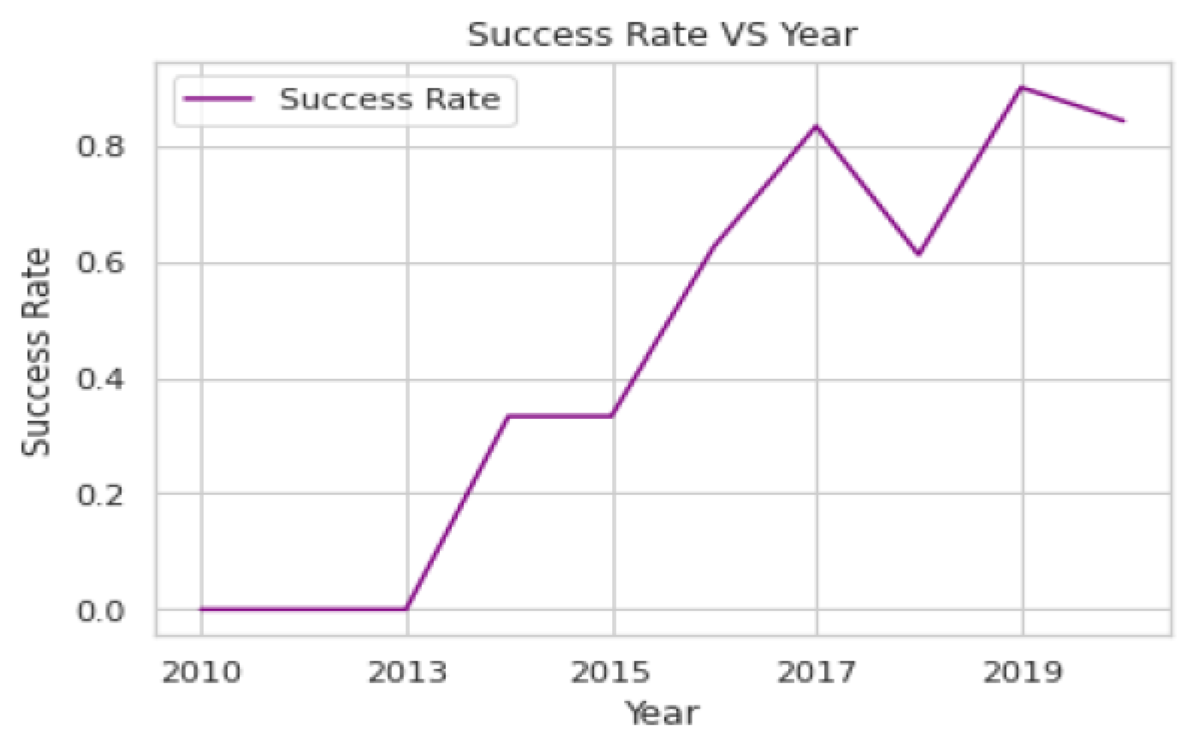


- It is clear to note that in LEO orbit, Success appears related to the number of flights. on the other hand, there appears to be no relationship between the flight number in the GTO orbit.

#EXTERNO.CONFIDENCIAL

Launch Success Yearly Trend

- It can be seen by analyzing the data and we saw its success rate which shows an increase in the probability of successful landing, an evolution from 2013 to 2020.

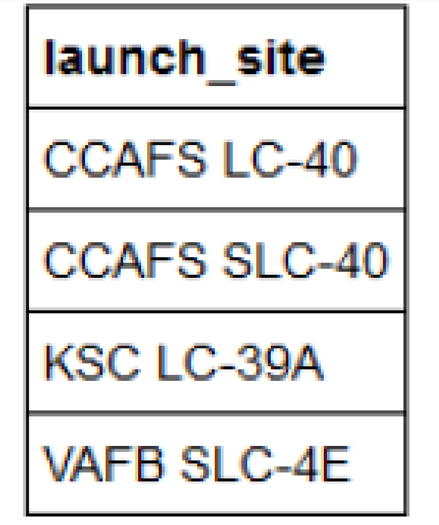


#EXTERNO.CONFIDENCIAL

All Launch Site Names

- An SQL query below shows 4 different sites for rocket launches.

%sql select launch_site distinct from SPACEXTB



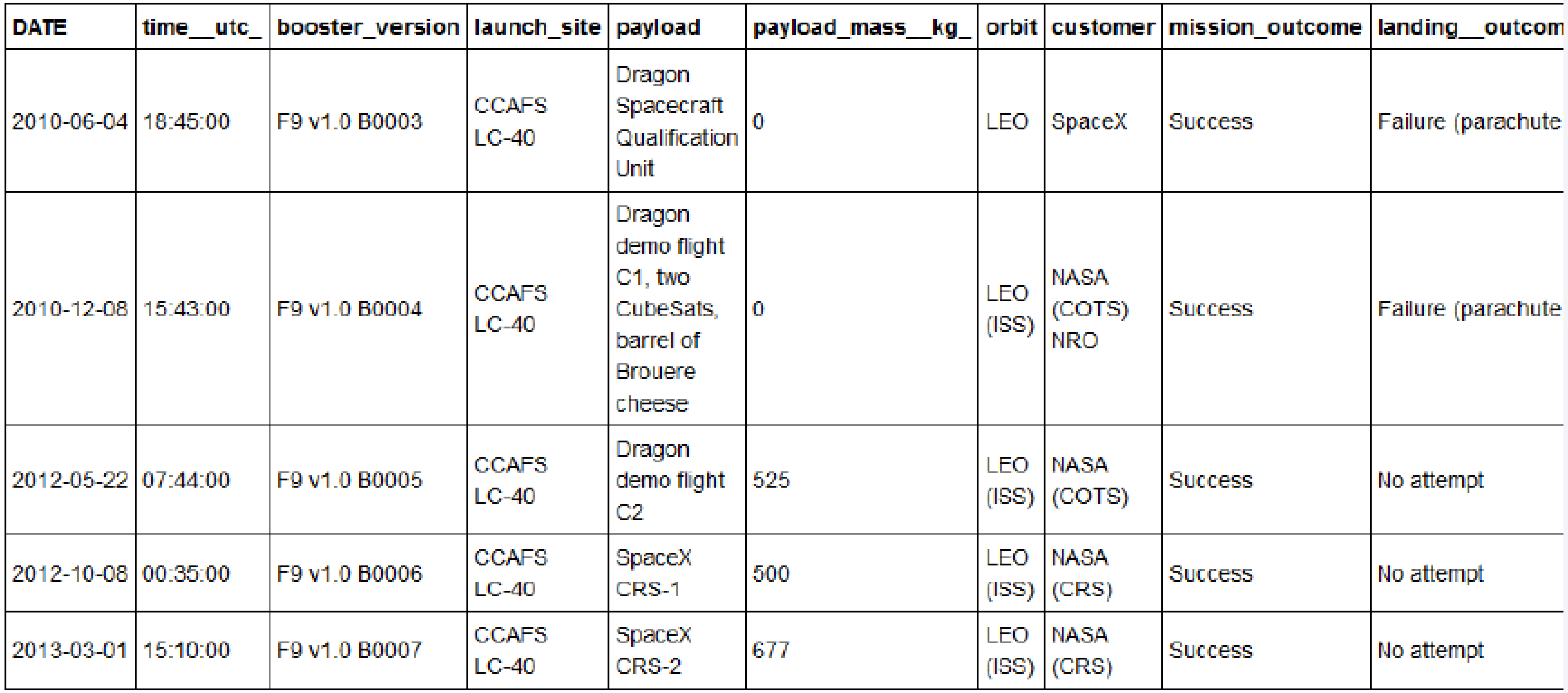
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

#EXTERNO.CONFIDENCIAL

Launch Site Names Begin with 'CCA'

- An SQL query below shows launches sites.

*%sql select * from SPACEXTBL where lanch_site like 'CCA%' limit 5*



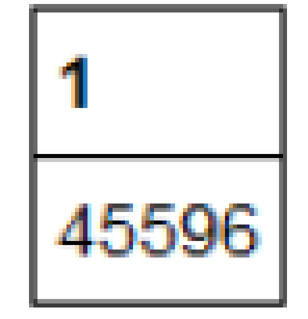
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

#EXTERNO.CONFIDENCIAL

Total Payload Mass

- An SQL query below shows the total payload carried by NASA boosters

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)’
```



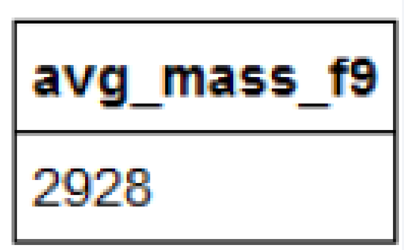
1
45596

- Note the total amount of payload that moved to space by NASA.

Average Payload Mass by F9 v1.1

- An SQL query below shows the calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as avg_mass_f9 from SPACEXTBL where booster_version = 'F9 v1.1’
```



avg_mass_f9
2928

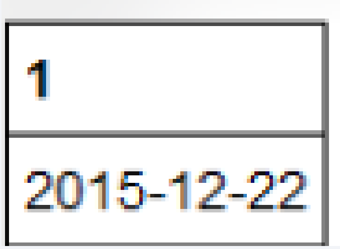
- Note that the average mass of the load transported by the F9 1.1 booster version is 2,928 kg.

#EXTERNO.CONFIDENCIAL

First Successful Ground Landing Date

- An SQL query below shows the first successful landing outcome on ground pad

- %sql select min(date) from SPACEXTBL where landing__outcome = 'Sucess (ground pad)’



1
2015-12-22

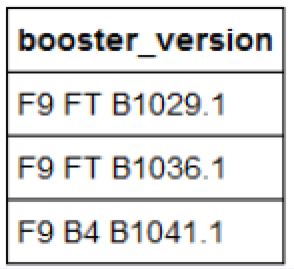
- Note that date of the first sucessful landing outcome on ground pad was 2015-12-22.

#EXTERNO.CONFIDENCIAL

Successful Drone Ship Landing with Payload between 4000 and 6000

- An SQL query below shows list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
  $sql select booster_version from SPACEXTBL whare (landing__outcome = ‘Success (drone ship)’ and
(payload_mass__kg > 4000 and payload_mass__kg > 6000))
```



booster_version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

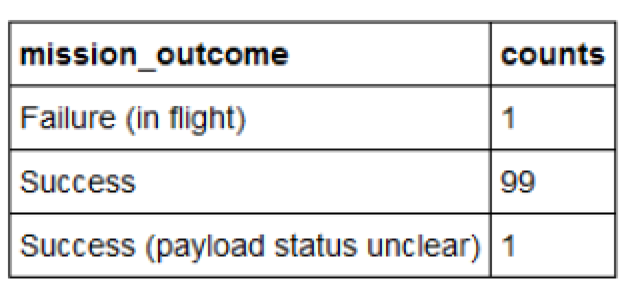
- Note that the success in the exposed range with payload between 4000 and 6000

#EXTERNO.CONFIDENCIAL

Total Number of Successful and Failure Mission Outcomes

- An SQL query below shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

$sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome



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

- Note that the success rate of mission results is the most dominant (99), and a single failed mission.

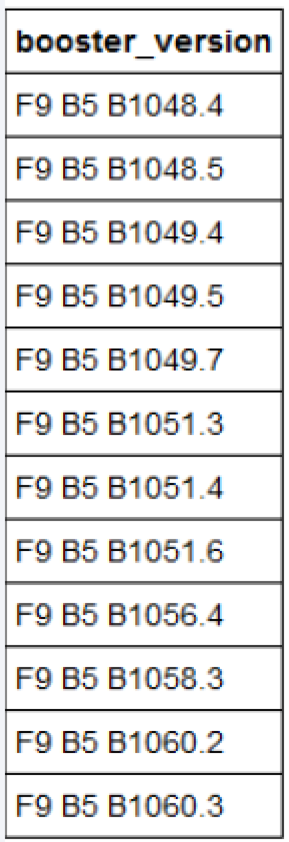
#EXTERNO.CONFIDENCIAL

Boosters Carried Maximum Payload

- An SQL query below shows the names of the booster which have carried the maximum payload mass

%sql select mission_outcome, ount(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome

- Note booster version that carry the maximum payload starts: F9 B5 and ranges B1048 up to B1060

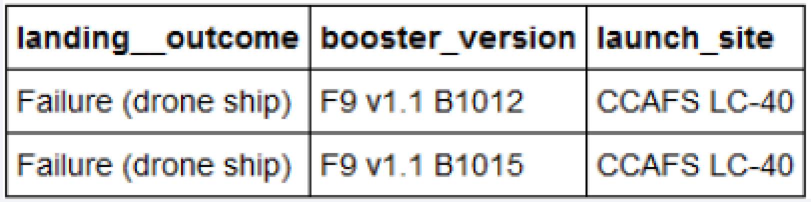


booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

#EXTERNO.CONFIDENCIAL

2015 Launch Records

- An SQL query below shows the list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



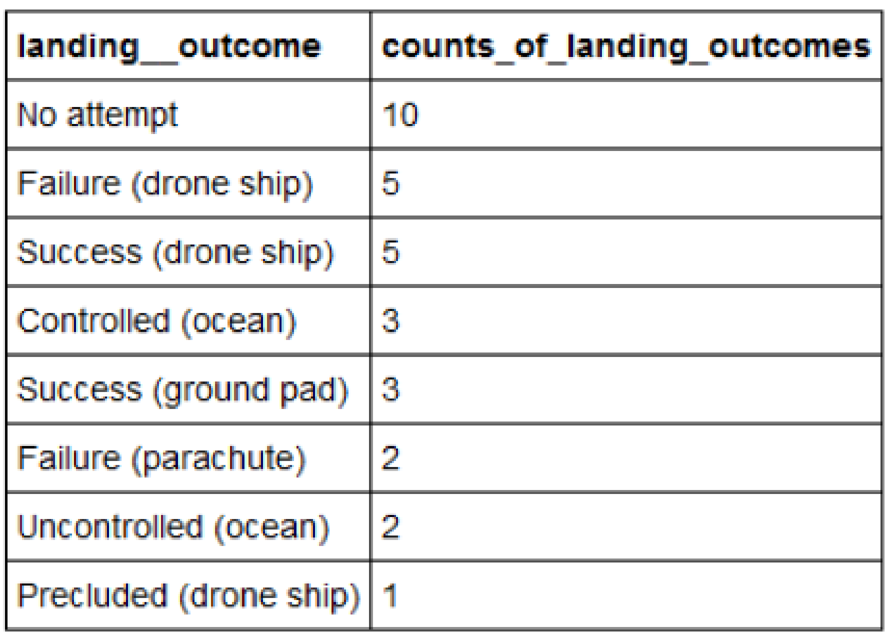
landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Noted that there are 2 landing failures in 2015 on a drone that both in the same location, CCFS LC-40 and with the same booster version F9 v1.1

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

- An SQL query below shows rank landing outcomes between 2010-06-04 and 2017-03-20

```
$sql select landing__outcome, count(*) as counts_of_landing_outcomes from  SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20' group by landing_outcome order by
count(landing_outcome) desc
```



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

#EXTERNO.CONFIDENCIAL

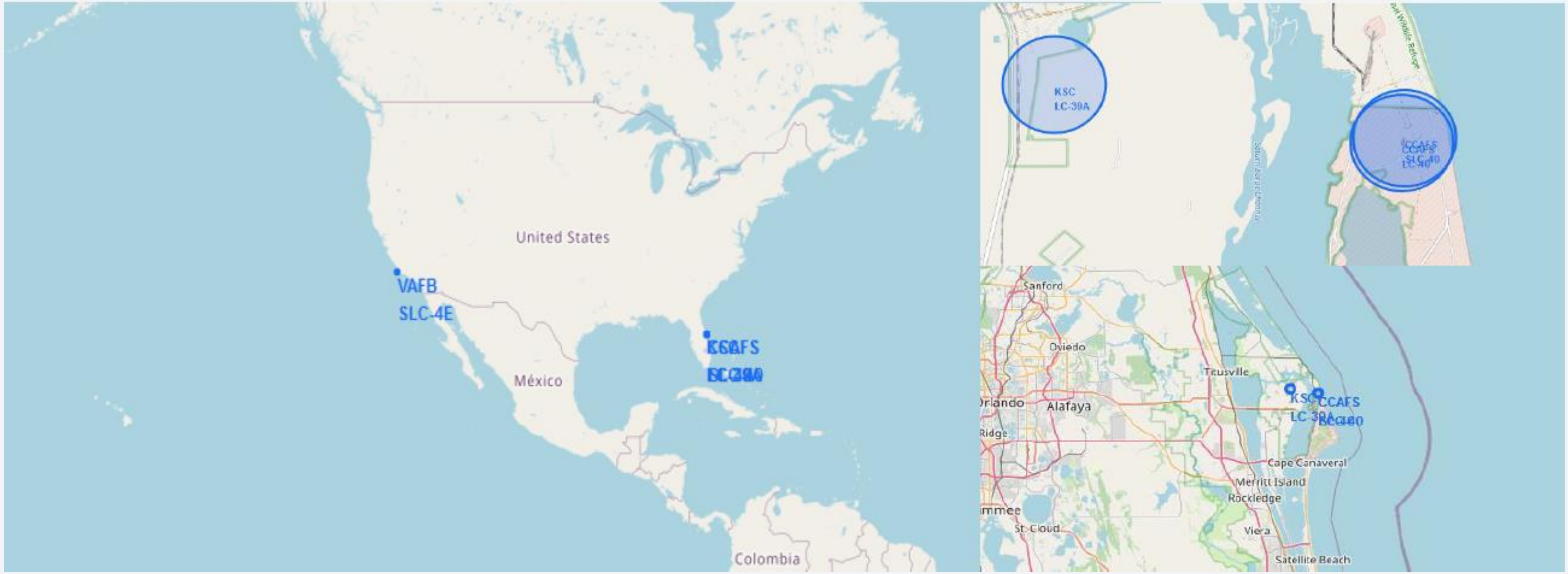
Section 3

Launch Sites Proximities Analysis

#EXTERNO.CONFIDENCIAL

Folium Map Screenshot Lauch Sites

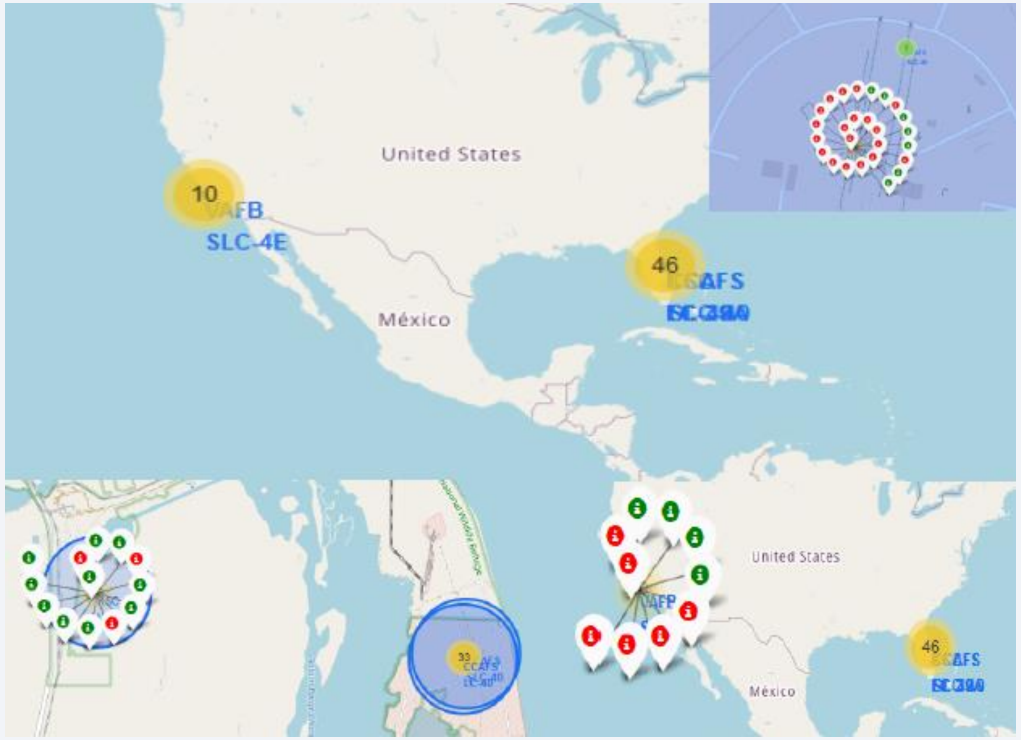
- Note on the map generated by Folim that all site locations are close to the coast and the equator, SPACEX focuses on locations close to water and at zero latitude to avoid unwanted accidents. The launch sites are distributed in the states Florida and California.



#EXTERNO.CONFIDENCIAL

Folium Map Screenshot Sucess x Lauch Local

- Note on the map generated by Folium the launch locations that have successful and unsuccessful launches, with success marked green and unsuccessful launches red.



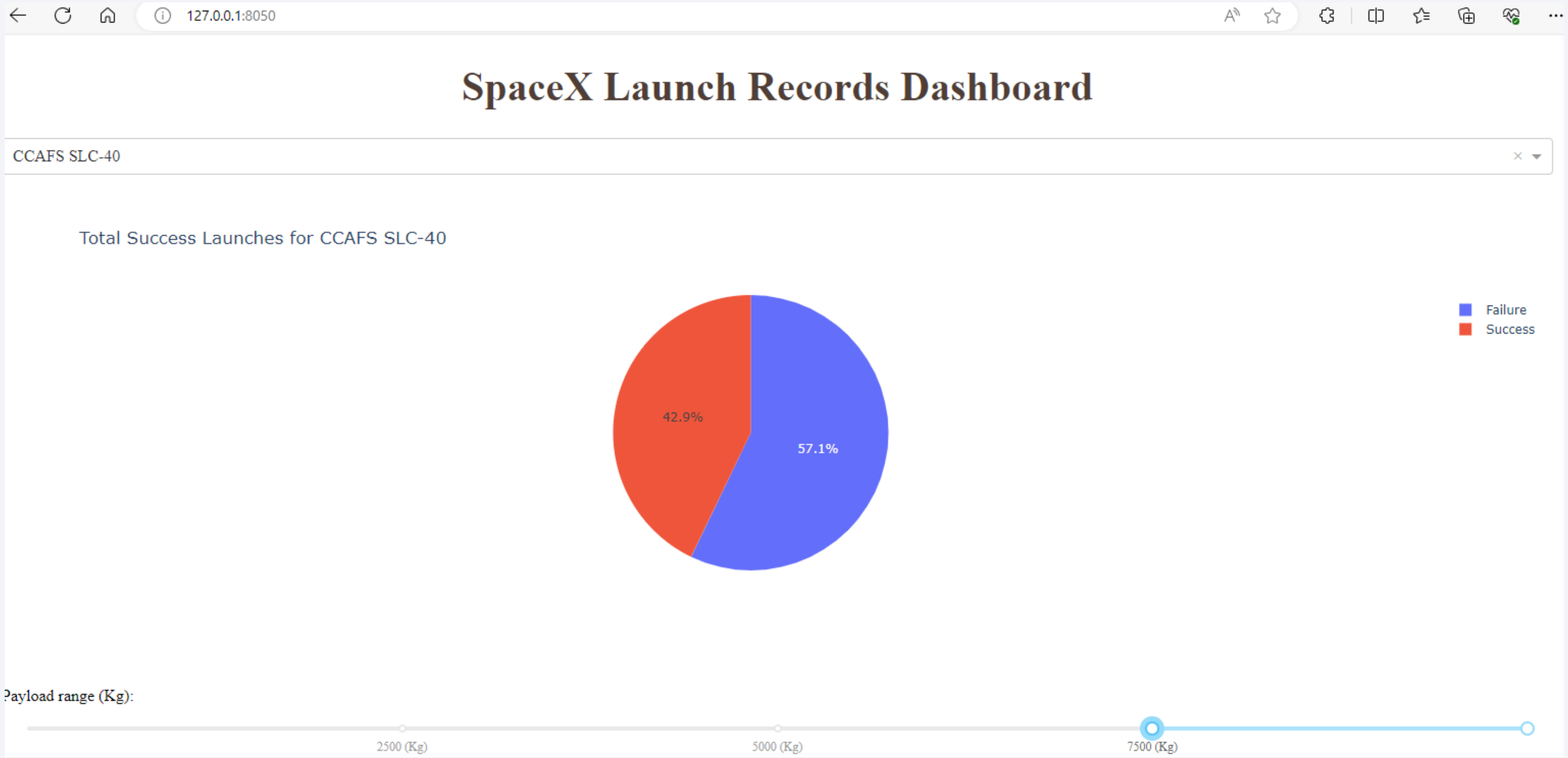
#EXTERNO.CONFIDENCIAL

Section 4

Build a Dashboard with Plotly Dash

Dashboard Screenshot Lauch Success for CCAFS SLC-40

- Note in the dash a total sucess launches for CCAFS SLC-40



127.0.0.1:8050

SpaceX Launch Records Dashboard

CCAFS SLC-40

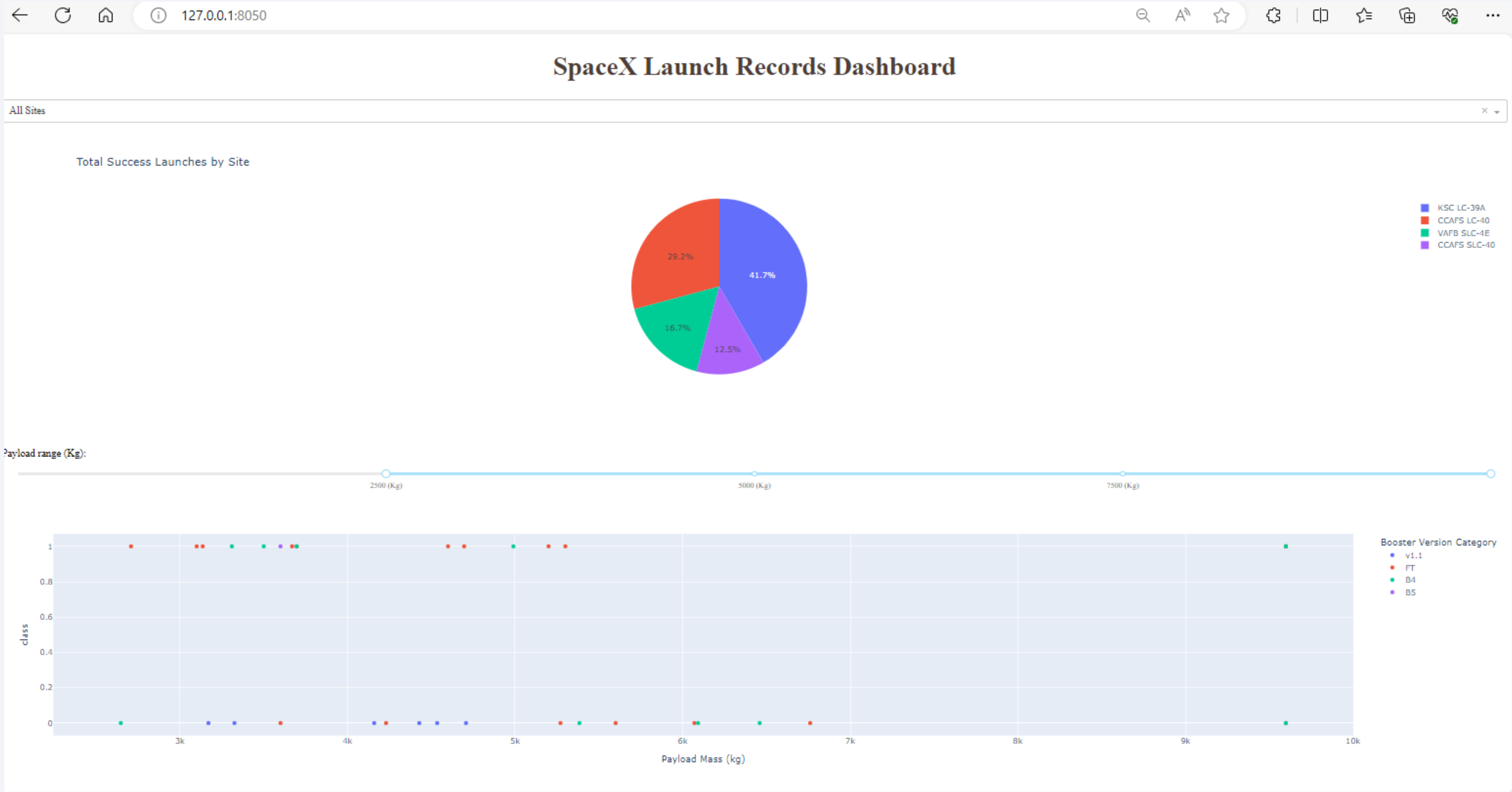
Total Success Launches for CCAFS SLC-40

Payload range (Kg):

#EXTERNO.CONFIDENCIAL

Dashboard Screenshot Lauch Success for All Sites

- Note in the dash a total sucess launches for all sites



127.0.0.1:8050

SpaceX Launch Records Dashboard

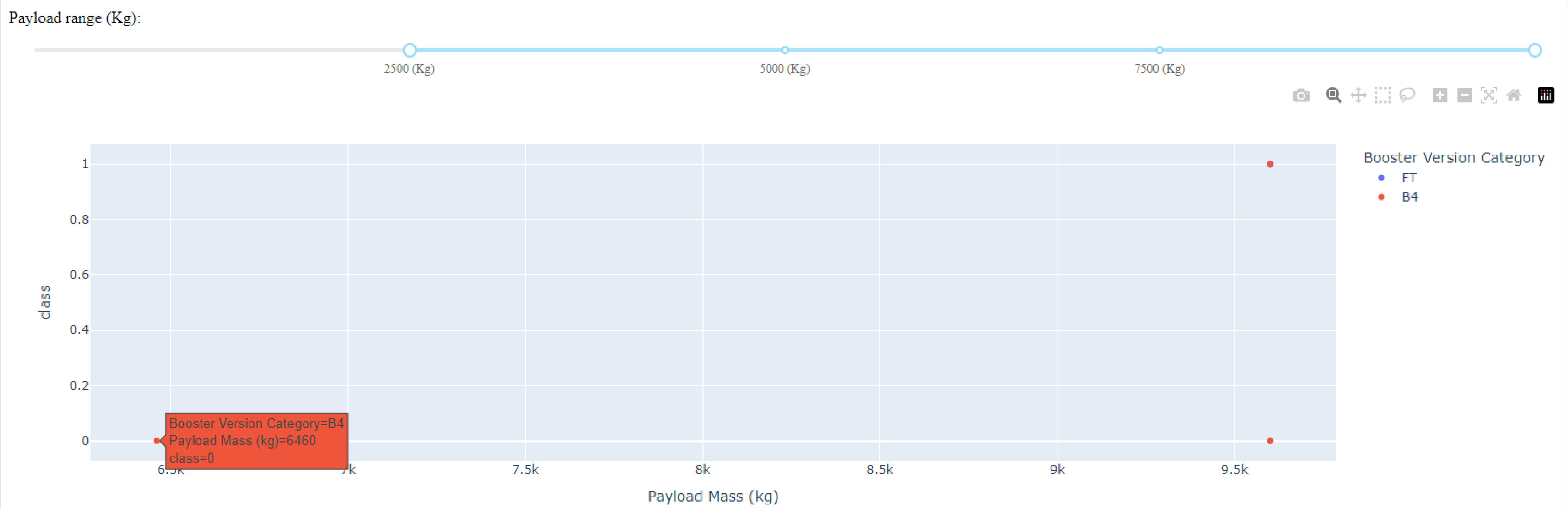
All Sites

Payload range (Kg):

#EXTERNO.CONFIDENCIAL

Dashboard Screenshot Scatter

- Note in the dash a VAFB SLC-4E booster version category



Payload range (Kg):

#EXTERNO.CONFIDENCIAL

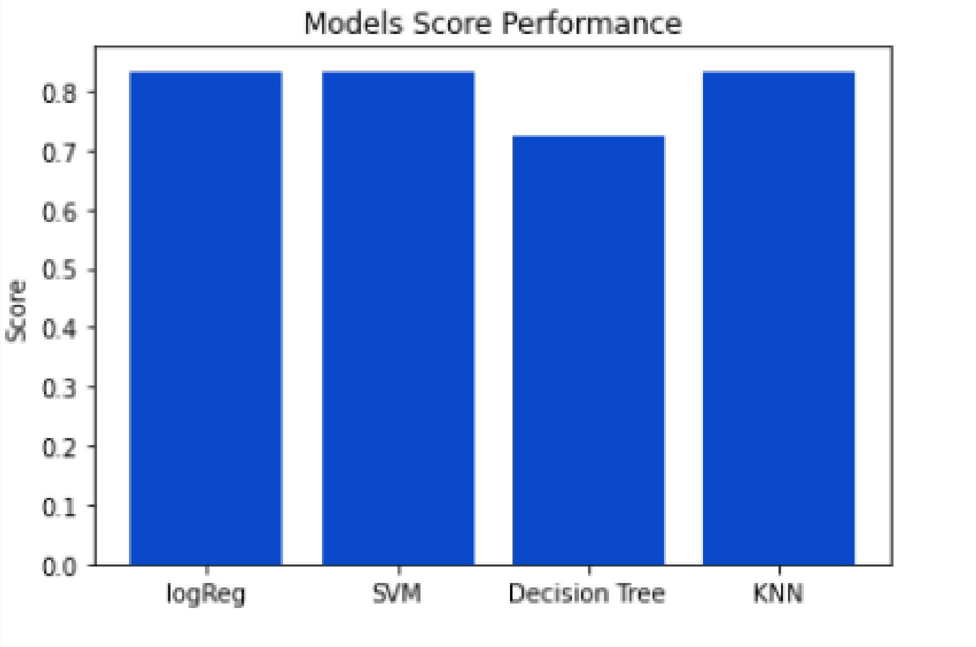
Section 5

Predictive Analysis (Classification)

#EXTERNO.CONFIDENCIAL

Classification Accuracy

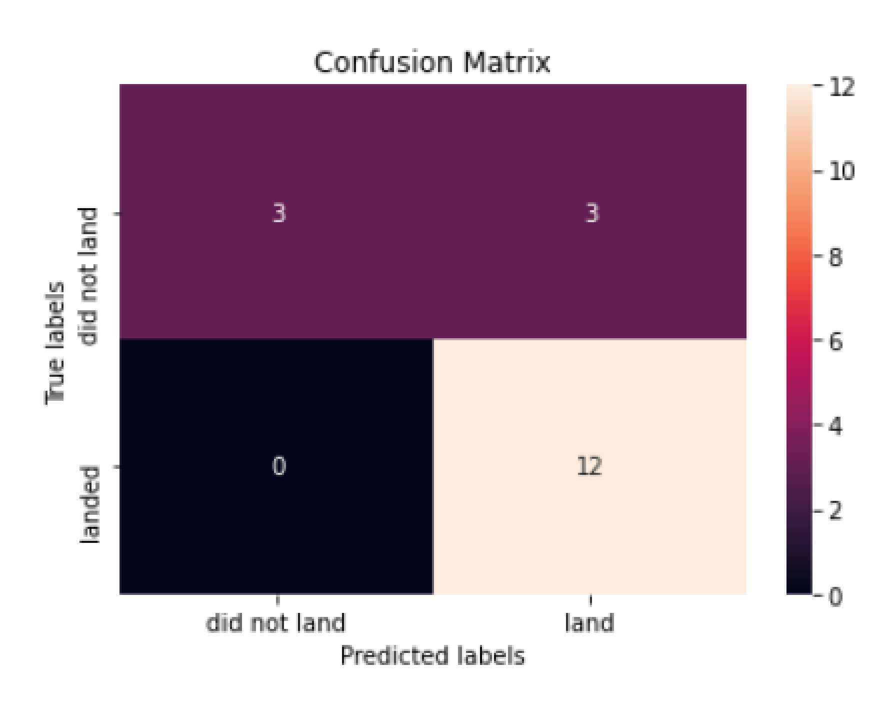
- Note in the graph of the performances of the algorithms used, with DT being the worst. Better performance for JC and SVM.



#EXTERNO.CONFIDENCIAL

Confusion Matrix

- Matrix Confusion KNN, SVM



True Positives = 12
False Positive = 0
True Negative = 3
False Negative = 3

#EXTERNO.CONFIDENCIAL

Conclusions

- When the organization properly stores its data, in a secure and scalable database, the work of the data scientist becomes possible to carry out the collection, cleaning and treatment according to the problem presented or the need for actions, such as SPACEX's work to predict success or failure in rocket landings in the first stage. In addition to several powerful Insights that begin to be seen when handling data correctly.
- The first stage is crucial for determining costs, which is a differentiator to beat the competition and in this case, as a defense against threats.
- Several variables were analyzed that could impact the first landing, where it is evident that launch sites are in strategic locations such as close to coastlines, orbits and also the use of reinforcements.
- This concludes the study and the predictive model prepared to accept new variables and predict success and failure of first stage landings, staying ahead of the competition and reducing operational costs.

Appendix

- Link GITHUB of project
- Link SPACEX API
- Link Official Site SPACEX

#EXTERNO.CONFIDENCIAL

Thank you!

#EXTERNO.CONFIDENCIAL