IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francis Pamela Tercero Varela
04.05.2023

Outline

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

Executive Summary

The methodology used consisted, on the one hand, of collecting data using the SpaceX API and Web scrap Falcon 9 launch records with BeautifulSoup and then performing data cleaning.

On the other hand, performing Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models to help us determine the probability of success of the Falcon 9 first stage will land successfully.

Obtaining that, the different supervised models used, the logistic regression, support vector machine and k nearest neighbors models have a higher accuracy of 0.83% therefore they better fit the data, compared to the decision tree classifier model, being the model with the lowest accuracy of 0.56%.

Introduction

We are a company that seeks to win the race to navigate through space, we seek to compete against other rocket companies by providing commercial flights into space at the best price making it accessible to more and more people on planet earth.

Therefore, through the present analysis we seek to determine an estimate of the price of a rocket launch to improve our offer through the reuse of the first stage and a better decision making of the most suitable place to launch our rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Requesting through the SpaceX API.
  - Performing web scraping.
- Perform data wrangling
  - Exploratory analysis of the data to later label them to identify the success or failure of the first stage landings.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing the data, Splitting into training and test set, finding best hyperparameter for the models using GridSearchCV and finally verifying the accuracy on the validation data using the data attribute best_score_

Data Collection

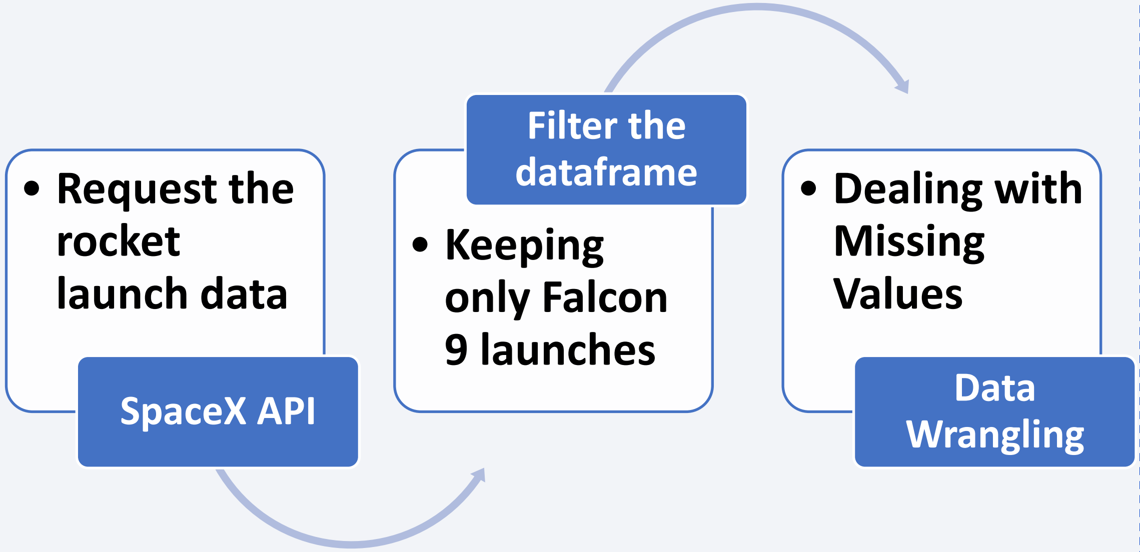
The data sets were collected by:

- Requesting through the SpaceX API to extract information using identification numbers in the launch data.
- Performing web scraping to collect Falcon 9 historical launch records from a Wikipedia.

Data Collection – SpaceX API

- The data collection was carried out through the following steps:

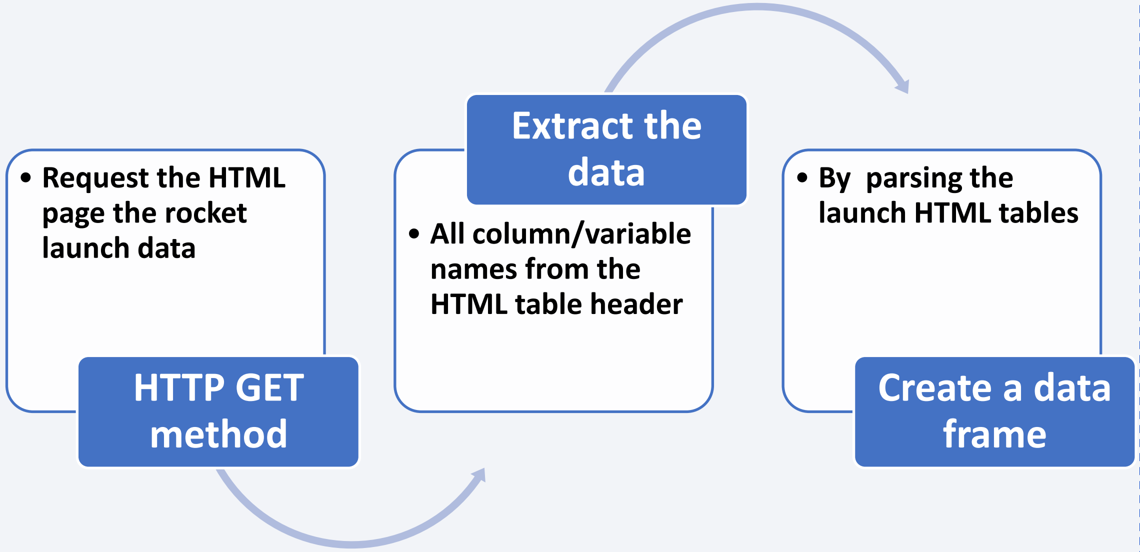
- GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/Lab%20SpaceX%20data%20collection%20API.ipynb



Data Collection - Scraping

- The web scraping was carried out through the following steps:

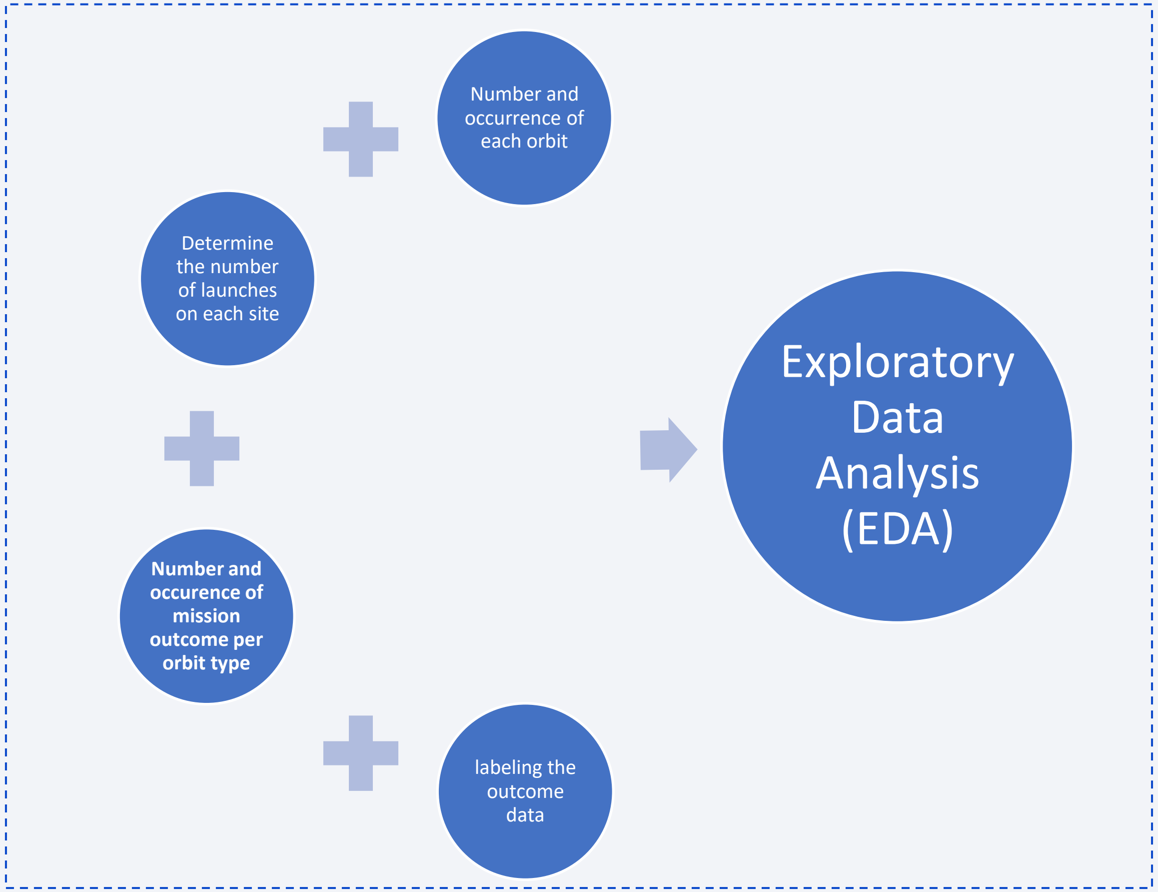
- GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/Lab%20SpaceX%20webscraping.ipynb



Data Wrangling

- The data wrangling was carried out through the following steps:

- GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/Lab%20SpaceX%20data%20wrangling.ipynb



EDA with Data Visualization

The exploratory analysis of the data was carried out on the following relationships of variables using bar graphs and scatterplot:

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload vs Launch
- Success rate of each orbit type
- Flight Number and Orbit type
- Payload and Orbit type
- Launch success yearly trend

GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The SQL queries were the following:

- 5 records where launch sites 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 succesful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass. Use a subquery

GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Interactive visual analytics was performed using Folium to identify the success/failed launches for each site on the map and the distances between a launch site to its proximities, identifying:

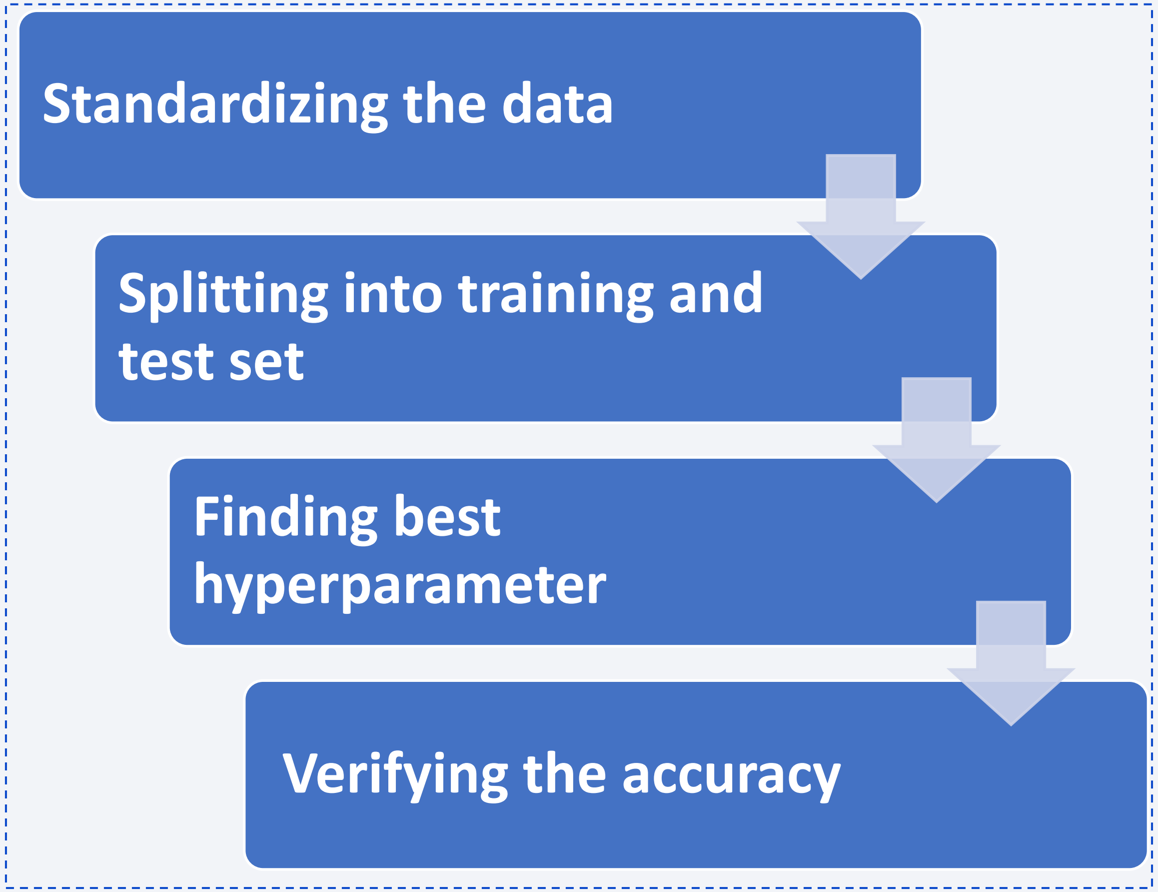
- Each launch site on the site map using folium.Circle and folium.Marker
- Adding a folium.Marker to marker_cluster for each launch result in the SpaceX data frame
- Using MousePosition and calculating the distance between the coastline point and the launch site.

GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/lab_jupyter_launch_site_location.ipynb

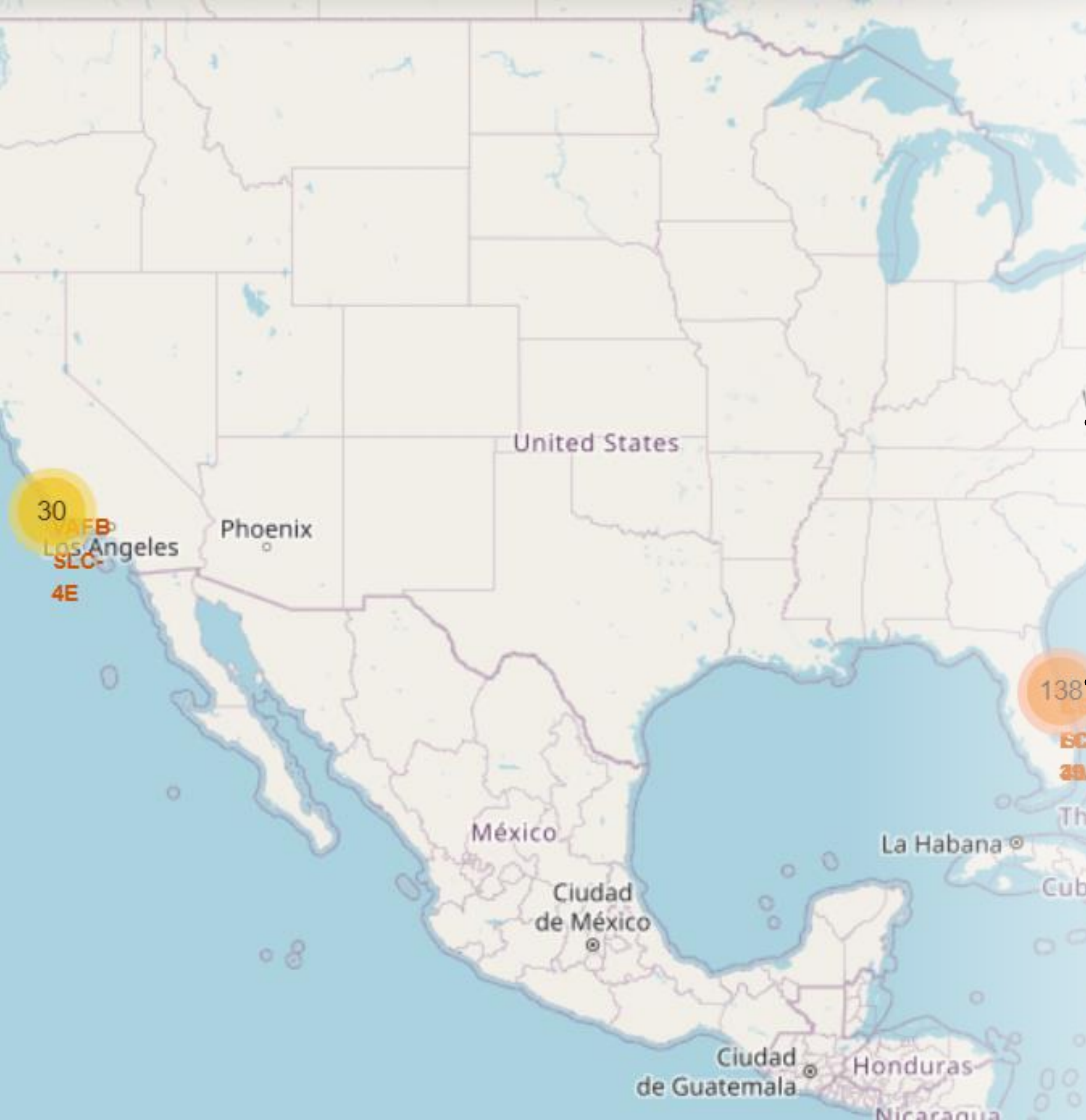
Predictive Analysis (Classification)

The classification models were implemented, with the purpose of verifying the one(s) that best fit the data for the objective prediction:

- Logistic regression,
- Support vector machine
- K nearest neighbors and
- Decision tree classifier



- GitHub URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

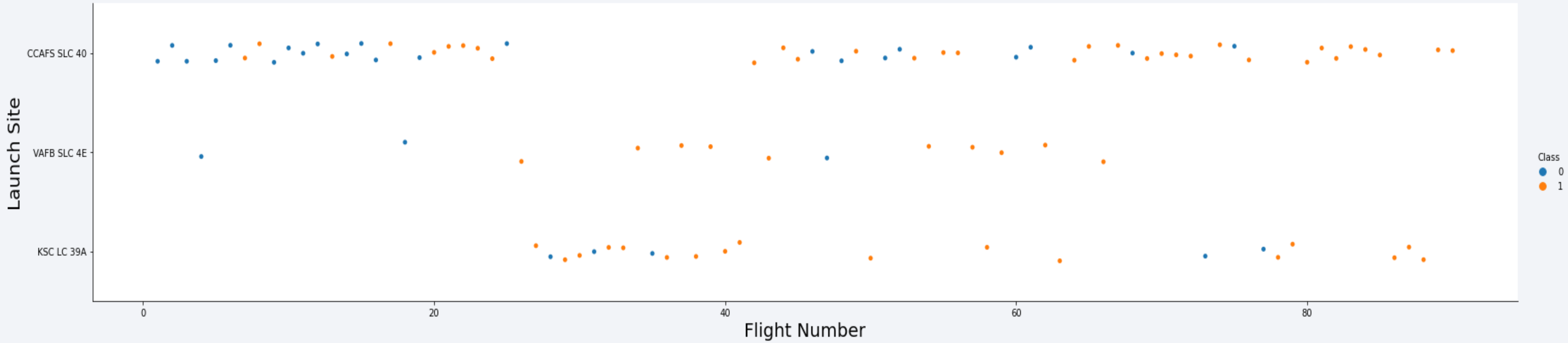
- Exploratory data analysis helped us to evaluate the completeness of the data, identify and label the main variables that would serve us throughout the study.

- The logistic regression, support vector machine and k nearest neighbors models they better fit the data (accuracy of 0.83%).

Section 2

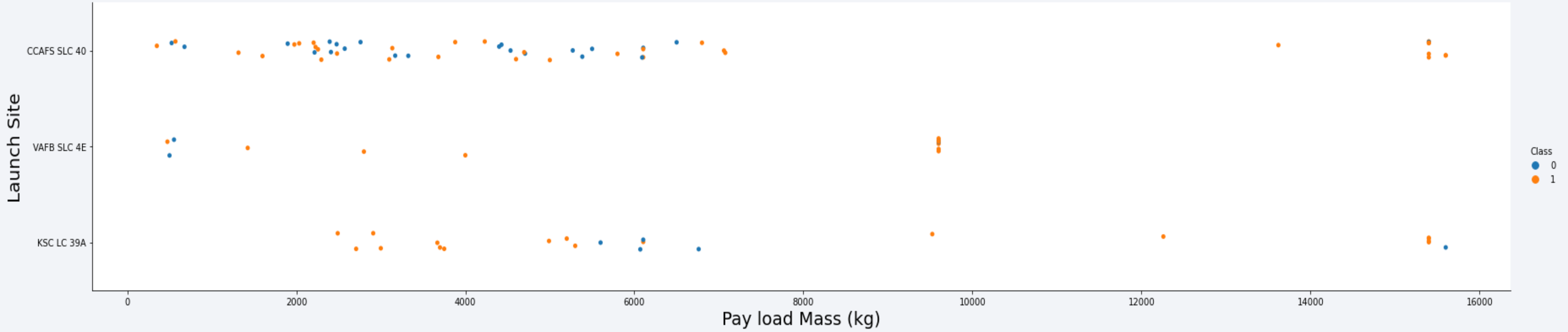
Insights drawn from EDA

Flight Number vs. Launch Site



The graph suggests that CCAFS SLC-40 is the most used launch site with the highest number of successful landings.

Payload vs. Launch Site

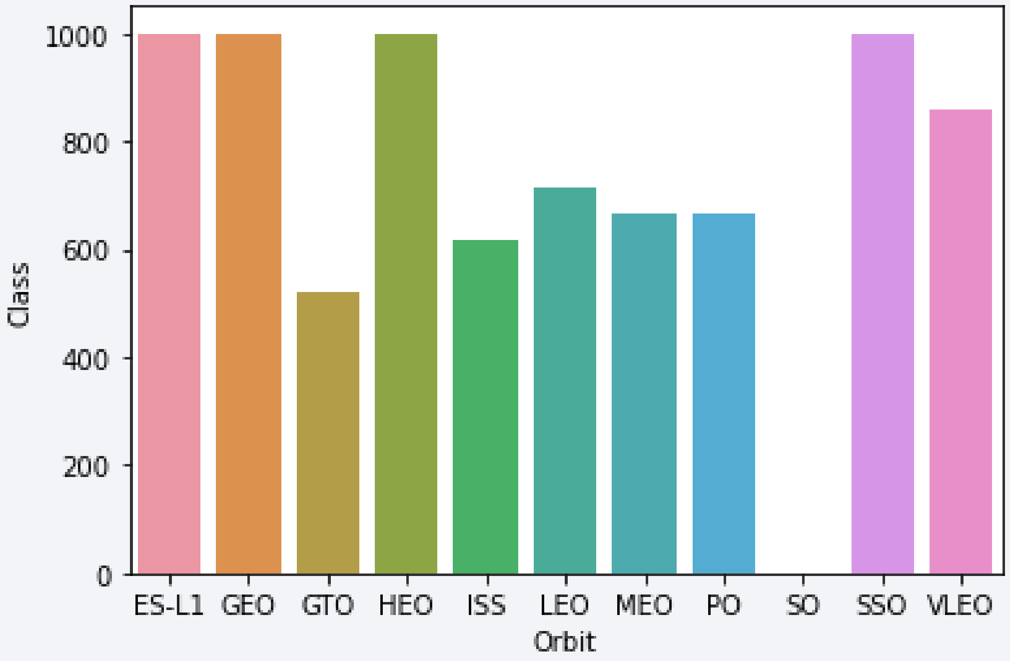


CCAFS SLC-40 and KSC LC 39A are the places where the most launches have been carried out and from which it is possible to launch rockets with a mass greater than 10,000 kg. Similarly, they have a high number of successful landings.

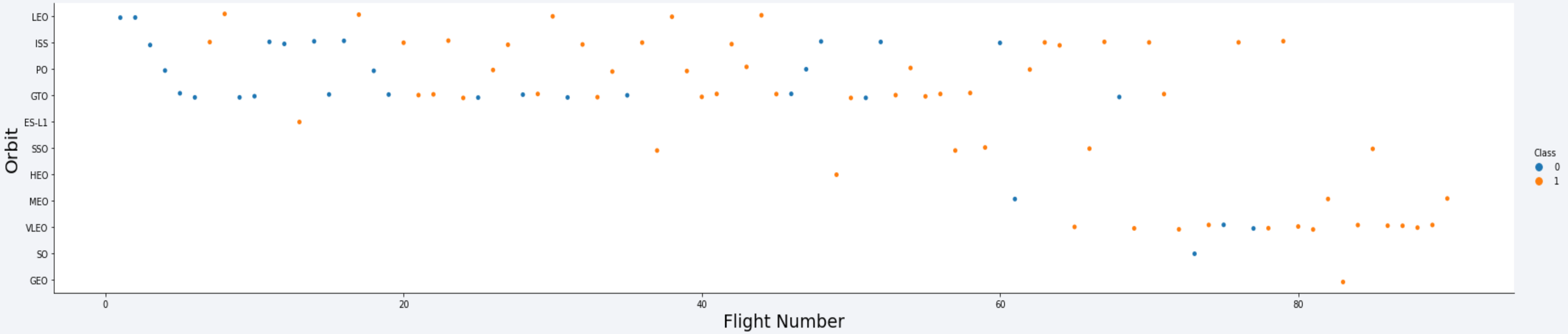
Success Rate vs. Orbit Type

The orbits with the highest success rate are:

- ES-L1 (100%)
- GEO (100%)
- HEO (100%)
- SSO (100%)
- VLEO (86%)
- LEO (71%)

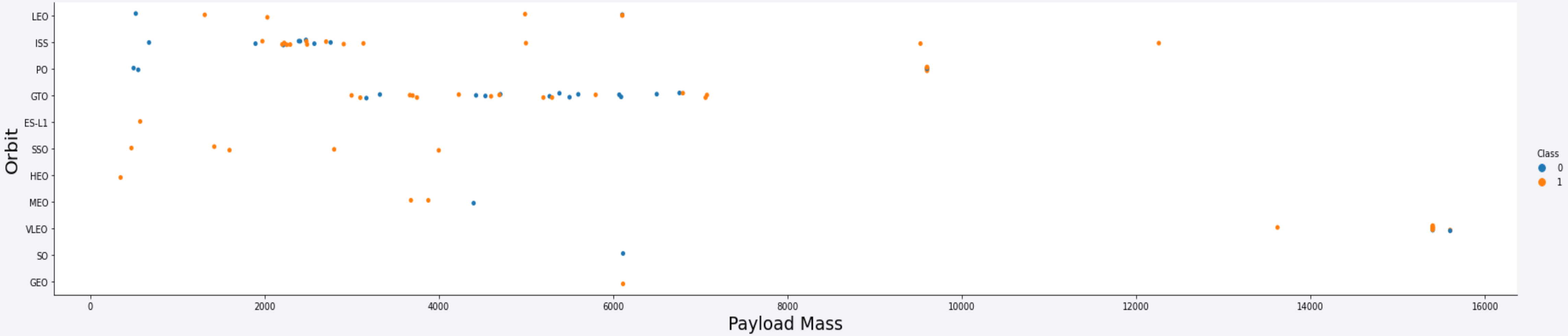


Flight Number vs. Orbit Type



- There is no clear relationship between the orbit and the number of flights in the first 4 orbits.
- In the 5th to 7th orbit there have been successful successful flights regardless of the orbit and the number of flights.
- The last 4 orbits have a greater number of flights and a high % of success.

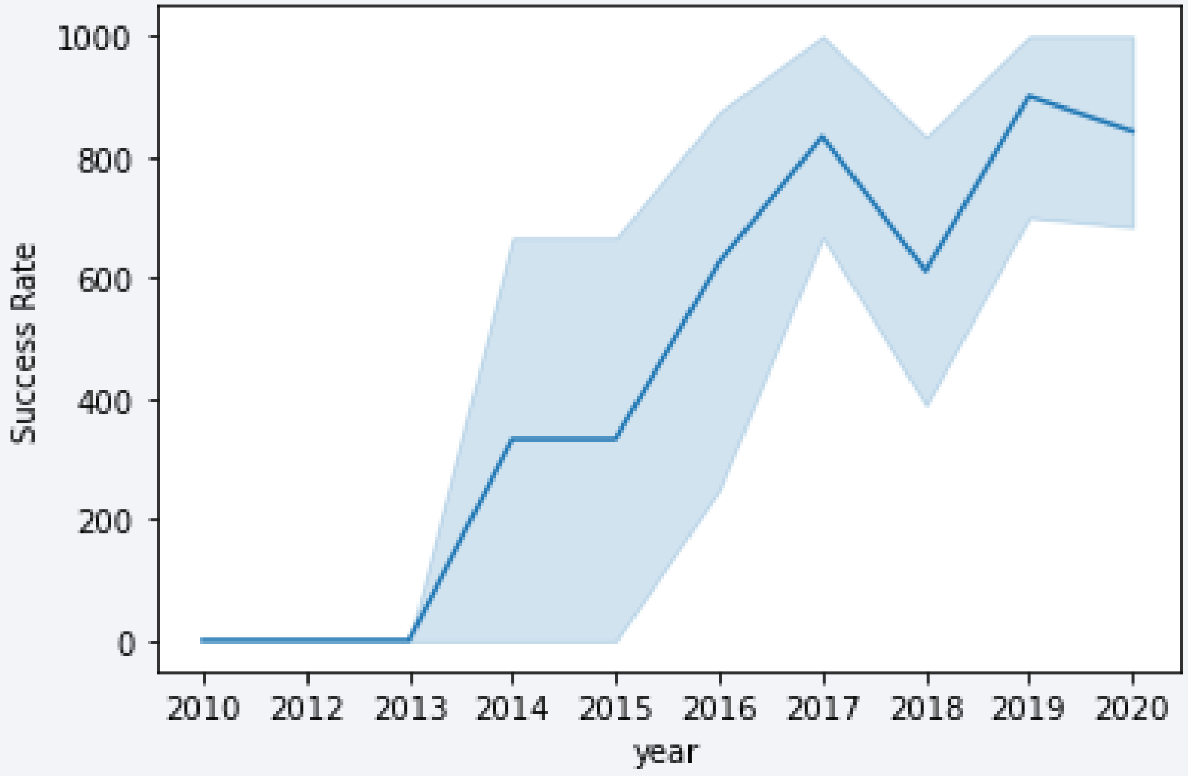
Payload vs. Orbit Type



- Launches in ES-LI, SSO y HEO orbits work very well with payload less than 4000 kg, contrary to ISS, PO y VLEO orbits for launches of more than 9000 kg

Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020.



All Launch Site Names

- The unique launch site names are as follows:

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

- Obtained by applying the following query: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL

Launch Site Names Begin with 'CCA'

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

- The 5 records where launch sites begin with `CCA` were obtained by applying the following query:

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

Total Payload Mass

total_payload_mass
45596

- To get the total payload carried by NASA's boosters, we run the following query:

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

Average Payload Mass by F9 v1.1

average_payload_mass
2928.4

- To get the the average payload mass carried by booster version F9 v1.1, we run the following query:

%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where Booster_Version like 'F9 v1.1'

First Successful Ground Landing Date

Date
01-03-2013

- The dates of the first successful landing outcome on ground pad was obtained by executing the following query:

%sql select min(Date) as Date from SPACEXTBL where Mission_Outcome like 'Success'

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were obtained by executing the following query:

%sql select Booster_Version from SPACEXTBL where (Mission_Outcome like 'Success') and (Landing_Outcome like 'Success (drone ship)') and (payload_mass__kg_ BETWEEN 4000 and 6000)

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Count
Failure (in flight)	3
Success	98

- The total number of successful and failure mission outcomes were obtained by executing the following query:

%sql select Mission_Outcome, count(*) as Count from SPACEXTBL group by Mission_Outcome like 'Success'

Boosters Carried Maximum Payload

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

- The List of the names of the booster which have carried the maximum payload mass were obtained by executing the following query:

%sql select Booster_Version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)

2015 Launch Records

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 mass were obtained by executing the following query:

%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site from SPACEXTBL where DATE like '2015%' AND Landing__outcome like 'Failure (drone ship)'

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

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

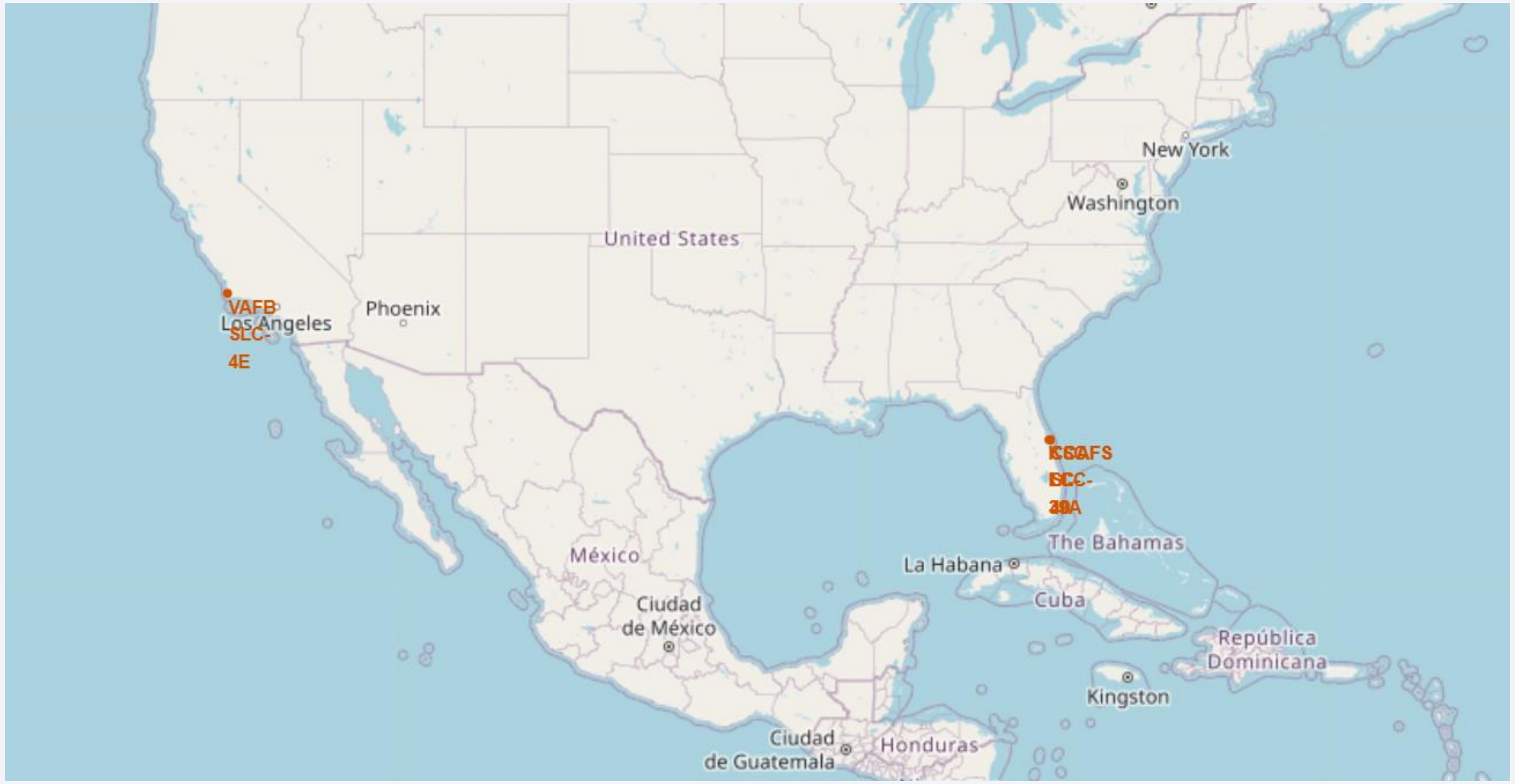
- The Rank of the landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order were obtained by executing the following query:

%sql select landing__outcome, count(*) as count from SPACEXTBL where Date >= '2010-06-04' and Date <= '2017-03-20' GROUP by landing__outcome ORDER BY count Desc

Section 3

Launch Sites Proximities Analysis

Launch Sites

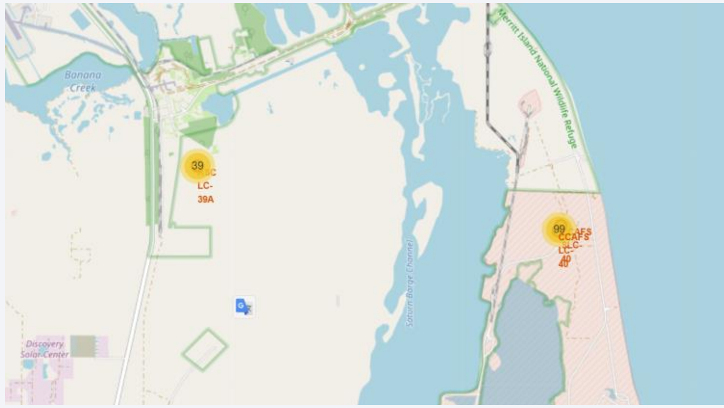


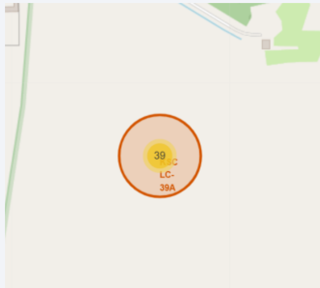
The location of the launch sites is very close to the coast

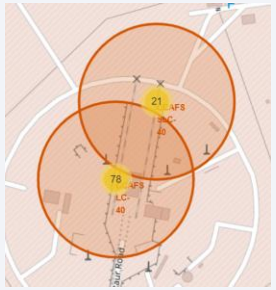
<Folium Map Screenshot 2>

West Coast

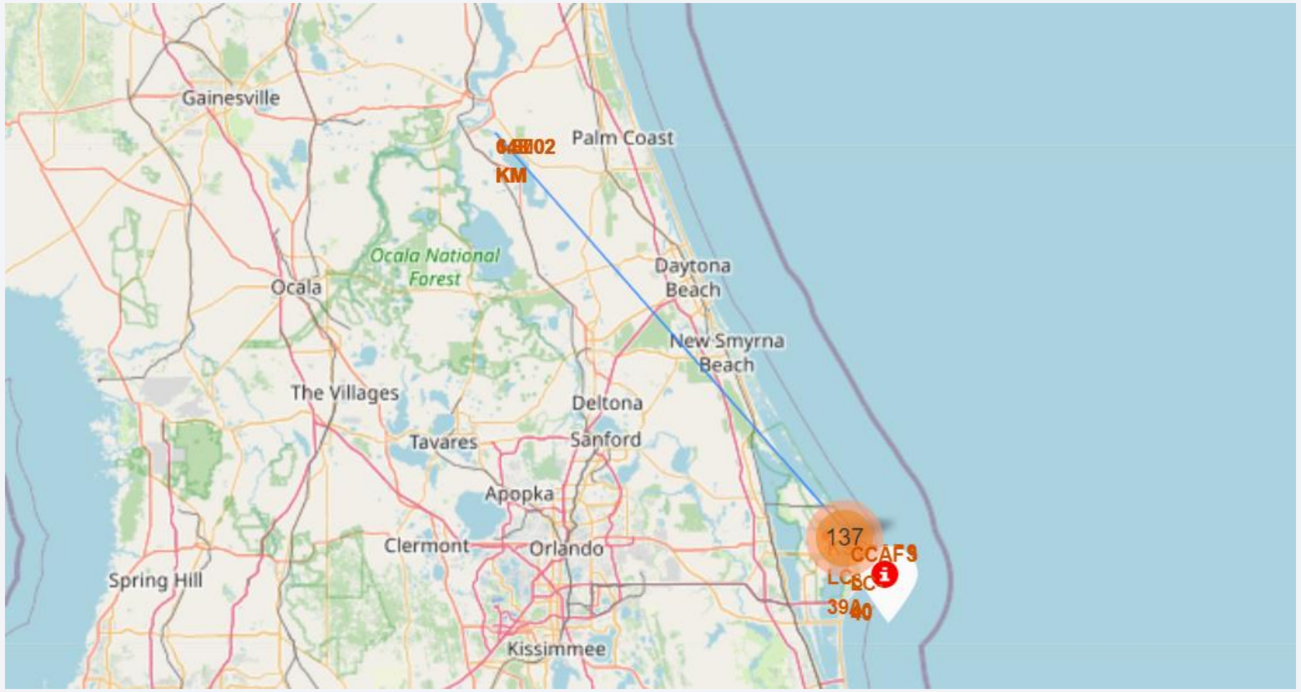
East Coast







<Folium Map Screenshot 3>



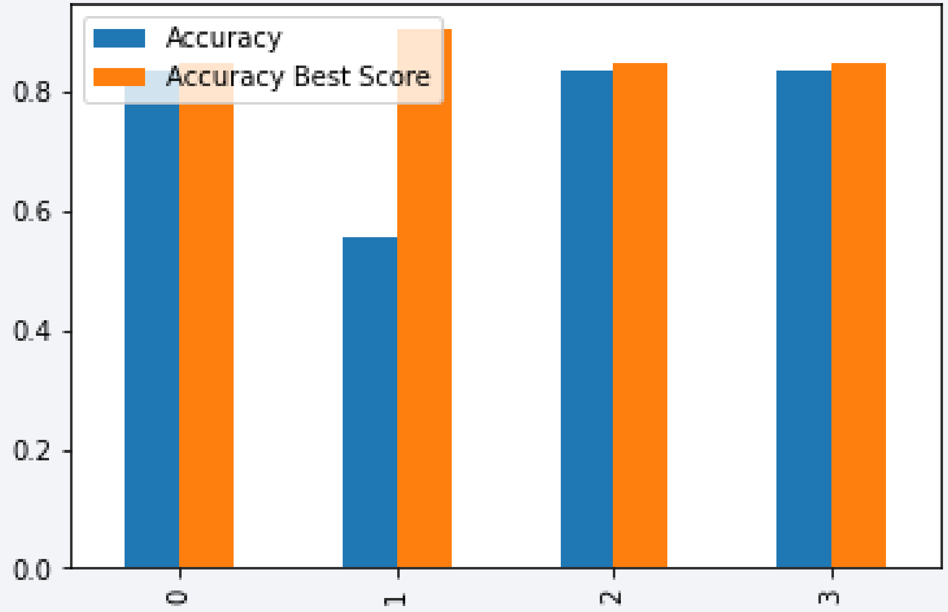
Each launch site seems to be a safe distance away from the city.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The logistic regression, support vector machine and k nearest neighbors models they better fit the data (accuracy of 0.83%).



Confusion Matrix

- For the three aforementioned models we obtain the following confision matrix.
- Predicting 12 landings in the true landing quadrant and 3 missed landings in the missed landing quadrant.

Conclusions

- Data is the best source for analysis like the current one, helping us make better decisions about future plans or investments.
- Flights from the KSC LC 39A launch site are less risky for a mass between 2,000-5,000 kg and 8,000-14,000 kg.
- The flights to the ES-L1, GEO, HEO and SSO orbits have a high probability of success, therefore there could be a better use of the rockets.
- With technological advances, the development of better rockets also improves by having more data on each failed and successful flight, making them safer for those who crew them.
- Machine learning techniques are an ally when estimating the future or trying to predict it.

Appendix

- The present result has been thanks to the guidance of the instructors responsible for the IBM Data Science certification. Big thanks to them.
- GitHub Repository URL: https://github.com/pamepath/IBM-Data-Data-Science-Certification-

Thank you!