IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Omar David Toledo Leguizamón
July 10th 2024

Outline

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

Executive Summary

- Summary of methodologies:

➢ It was done an extraction of from different data sources through two main techniques.

➢ It was done an exhaustive EDA to get variables for a classification task.

➢ Different models were designed, implemented and compared for the classification task

- Summary of all results

➢ It was gotten an 83% of accuracy at the moment of predicting the possible result of a SpaceX landing.

Introduction

- Project background and context

➢ As a company that is interested in SpaceX launches across the last decade, it has been a challenge to understand without the complex space engineering knowledge the different conditions how this company has achieved reusing their boosters in different missions saving millions of dollars.

- Problems you want to find answers

➢ As a Data scientist, the challenge is to understand how has been the evolution of SpaceX successful devices recovery due to many conditions such as payload mass, launching sites and more others; leading us to the following main question: **Is it possible to predict if the landing outcome will be successful or not using just the mission information?**

Section 1

Methodology

Methodology

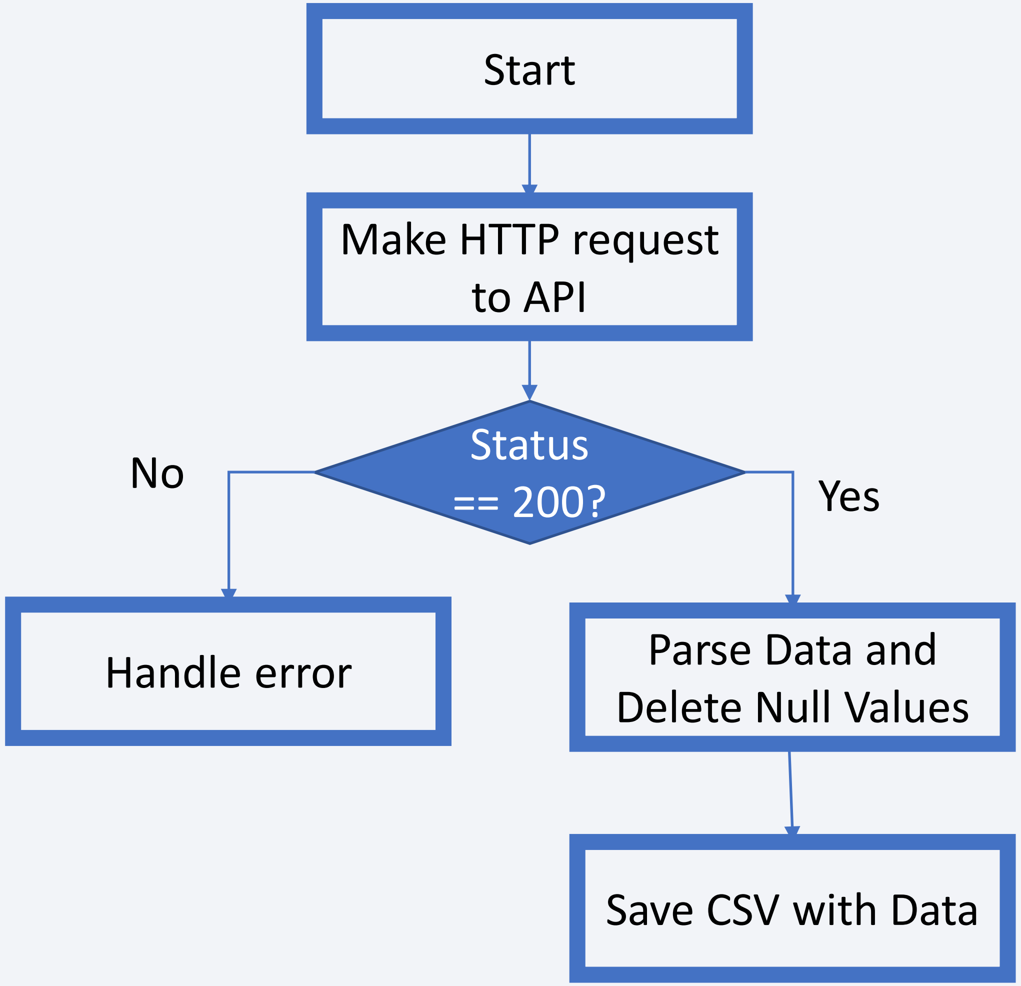
Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX API and Web Scraping
- Perform data wrangling
  - Describe was manipulated using Pandas functions
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Decission Tree, KNN and Logistic Regression Models were applied

Data Collection – SpaceX API

- We show the flowchart with the design of the Data Collection trough the SpaceX API process

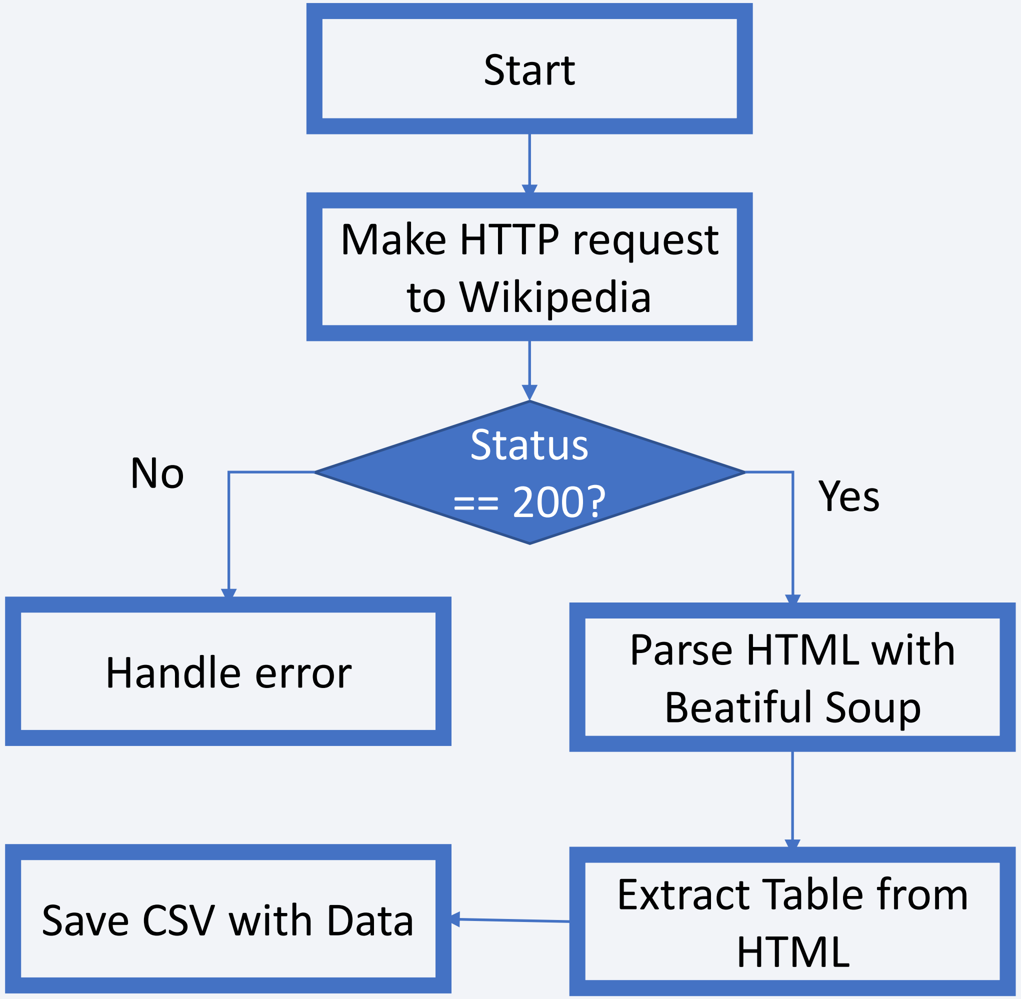
- All the data collection through the API can be found here: [Data Collection API Notebook]



Data Collection - Scraping

- We show the flowchart with the design of the Data Collection trough the Wikipedia web Scraping process

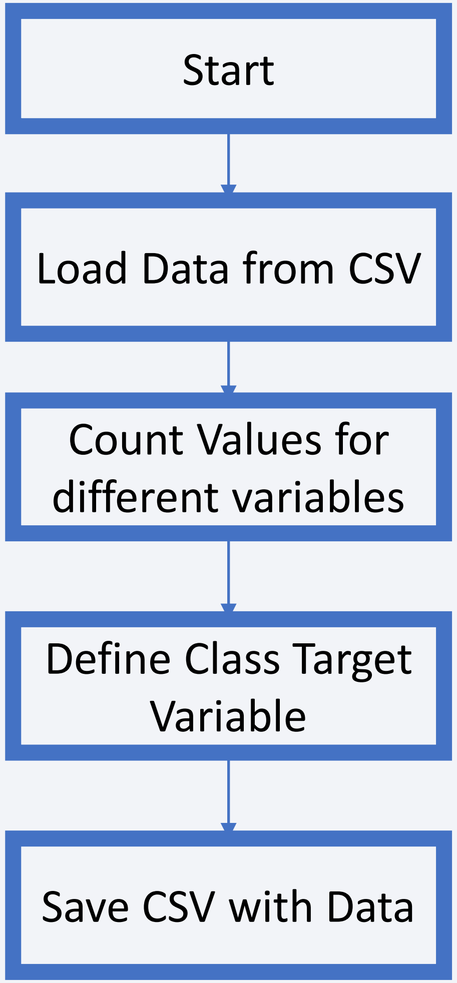
- All the data collection through the Web Scraping can be found here: [Data Collection Web Scraping Notebook]()


Start
Make HTTP request to Wikipedia
Status == 200?
No
Yes
Handle error
Parse HTML with Beatiful Soup
Extract Table from HTML
Save CSV with Data

Data Wrangling

- Data was loaded and the Mission Outcomes were counted and defined, at the end the Class target variable was created

- All the data collection through the Web Scraping can be found here: Data Wrangling Notebook



Start

Load Data from CSV

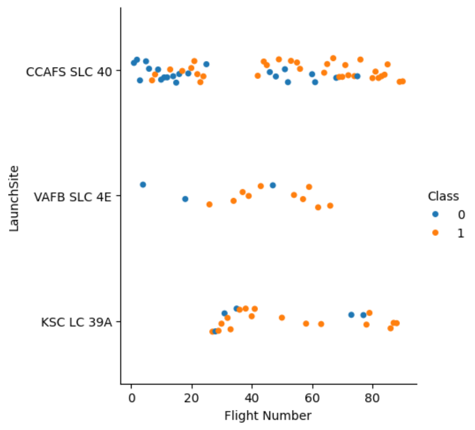
Count Values for different variables

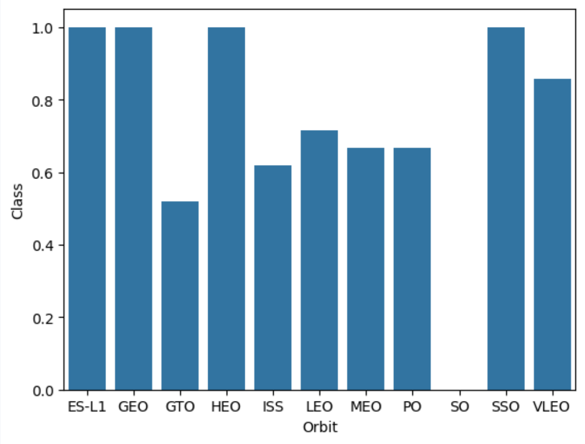
Define Class Target Variable

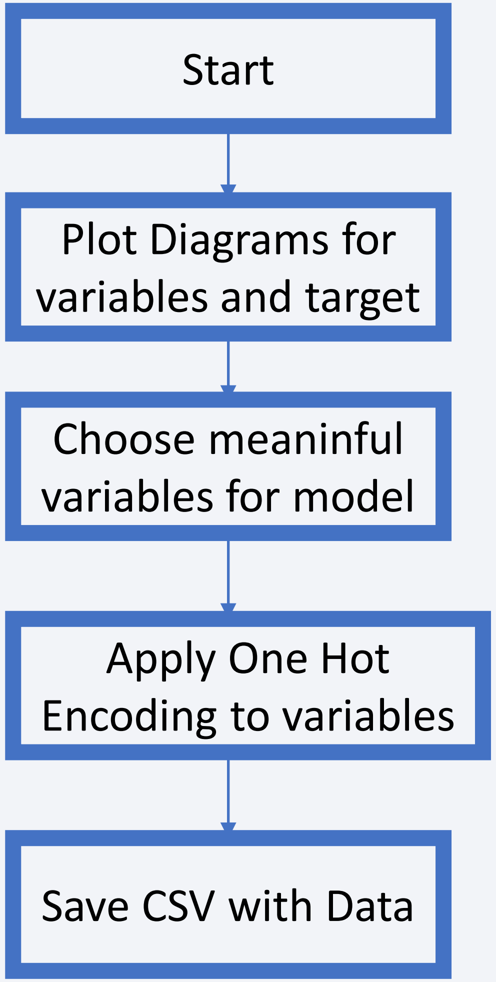
Save CSV with Data

EDA with Data Visualization

- For the EDA with Data Visualization, it has done a set of plots in order to check variable correlation
- The EDA with Data Visualization can be found in the following GitHub link: [EDA with Data Visualization Notebook]()







Start

Plot Diagrams for variables and target

Choose meaninful variables for model

Apply One Hot Encoding to variables

Save CSV with Data

EDA with SQL

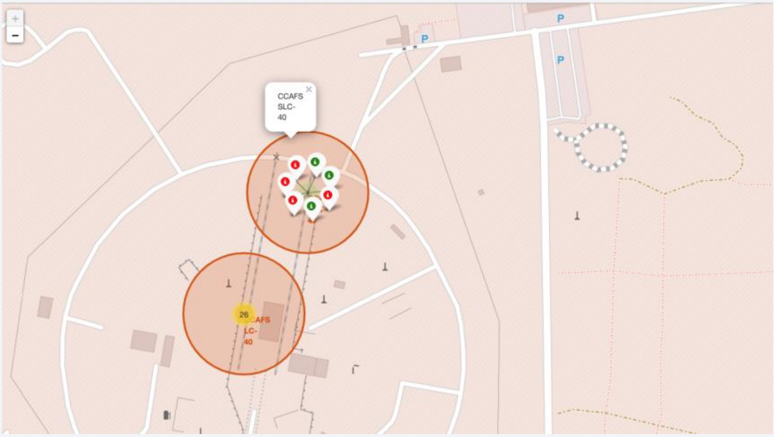
- Some of the following SQL queries were done to get the data for a further analysis:

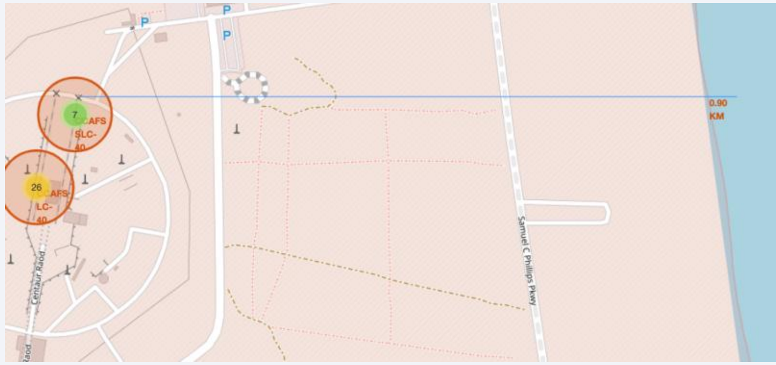
1. Get the total count for each different type of Mission Outcome.
2. Get the average payload mass for some specific Booster Version
3. Get the first date in which a successful landing was achieved.
4. Get the total count for each different type of Landing Outcome.
5. Get different mission data for a specific Landing site

- The complete analysis done with SQL can be found in the following GitHub link: EDA with SQL Notebook

Build an Interactive Map with Folium

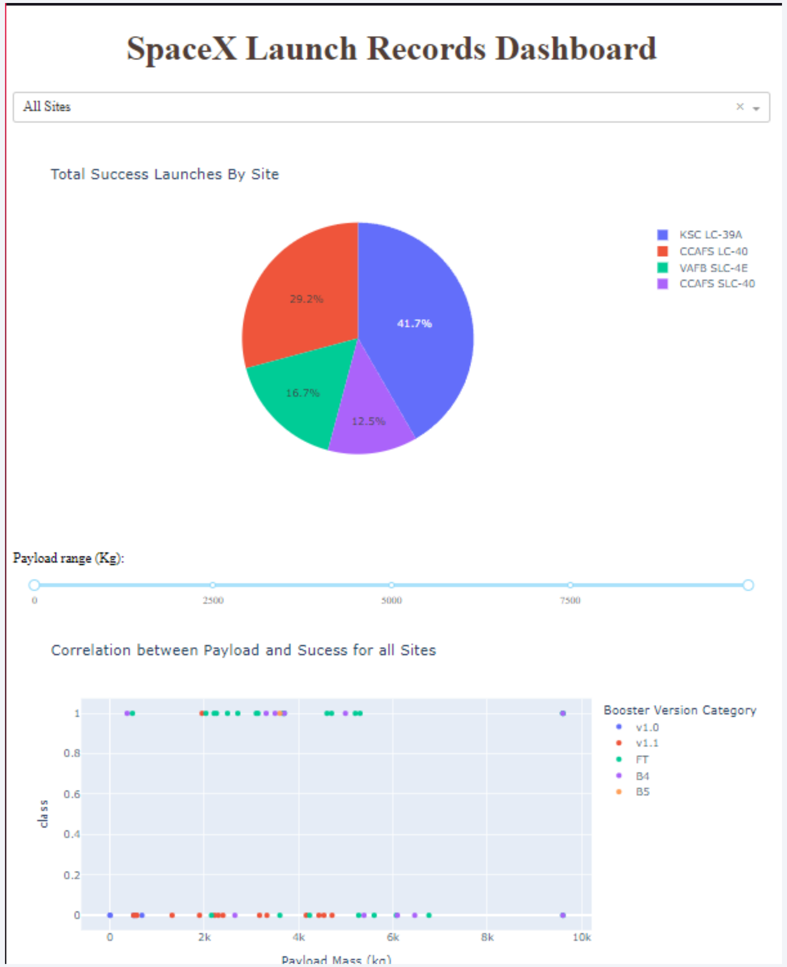
- Different maps were designed to show the different launch sites and the mission outcomes that we could find on each case
- These maps were done using Folium base maps, Markers and circles for each Launch Site and Mouse Pointers to get specific coordinates
- The develop of the different maps can be found in the following GitHub link: [Map with Folium Notebook]()





Build a Dashboard with Plotly Dash

- In the Dashbord it can be found to type of plots, the first one, a pie plot with the different mission outcomes, and in the second one, a scatter plot for the outcome by payload mass.
- These plots were done by using plotly objects, option selectors, range sliders and a callback function for each plot.
- The Dashborad development can be found in the following GitHub link: [Dashboard Development]()



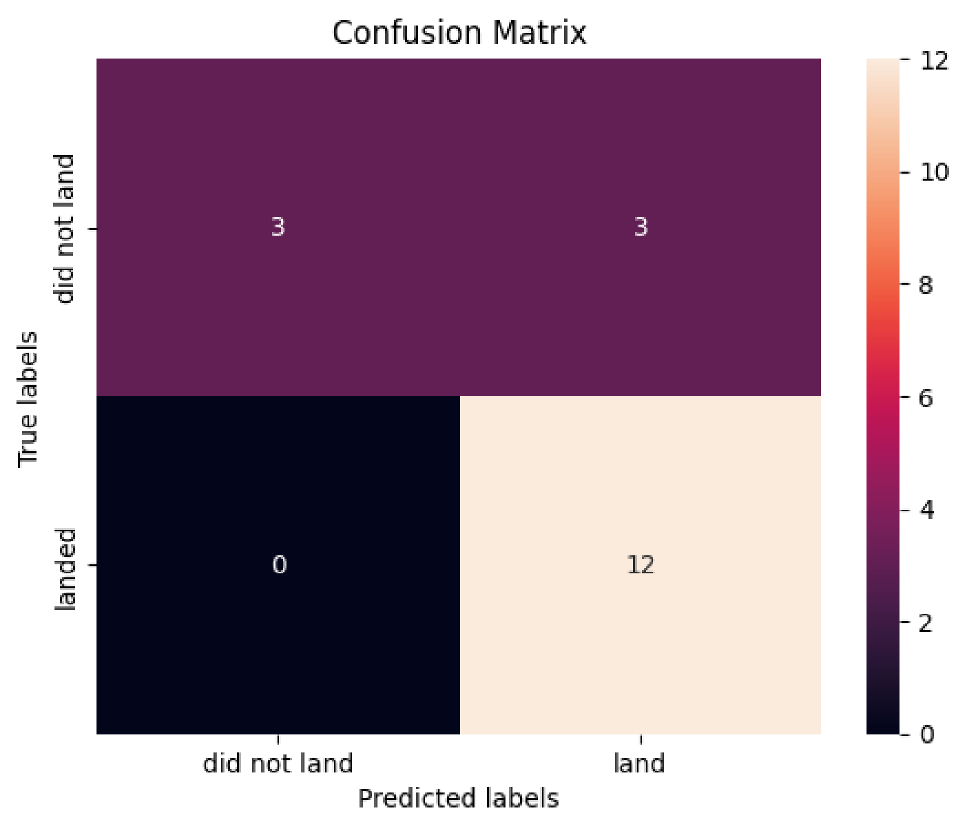
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

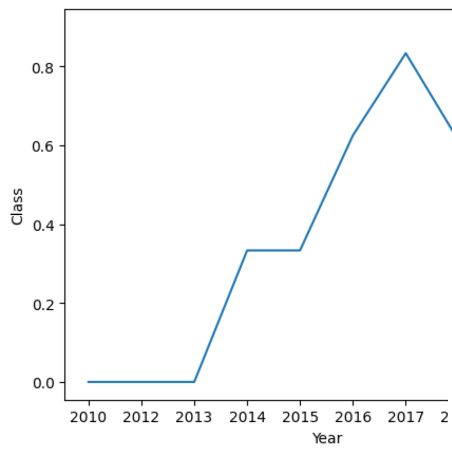
Predictive Analysis (Classification)

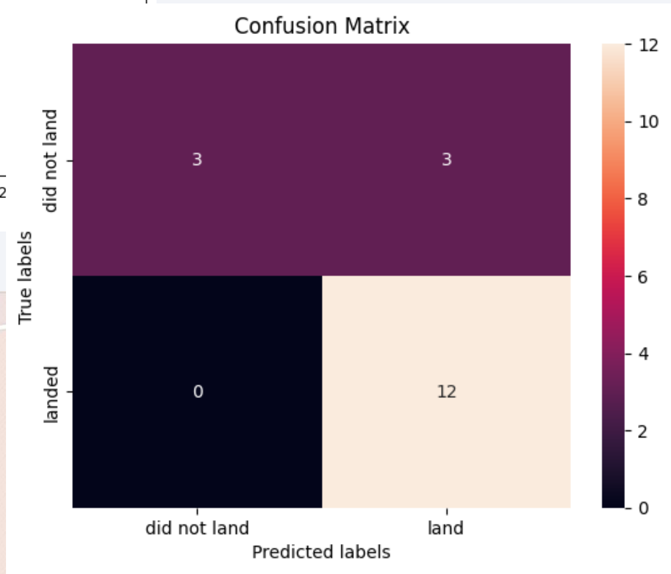
- For the model, it was defined as a classification task, so it was developed by using SVM, KNN, Decision Tree and Logistic Regression.
- The data was normalized and then split before passing it to the models; and the results were displayed by using a confusion matrix.
- The model development can be found in the following GitHub link: Predictive Analysis Notebook



Results

- Exploratory data analysis results: It was found a correlation between some launch sites, orbit and year with the success rate of the landings
- Interactive analytics demo in screenshots: The different map regions shown the correlation between the launch site and success rate
- Predictive analysis results: The predictive analysis shown that after choosing the best model, it could be predicted new landing results with an accuracy of 83%





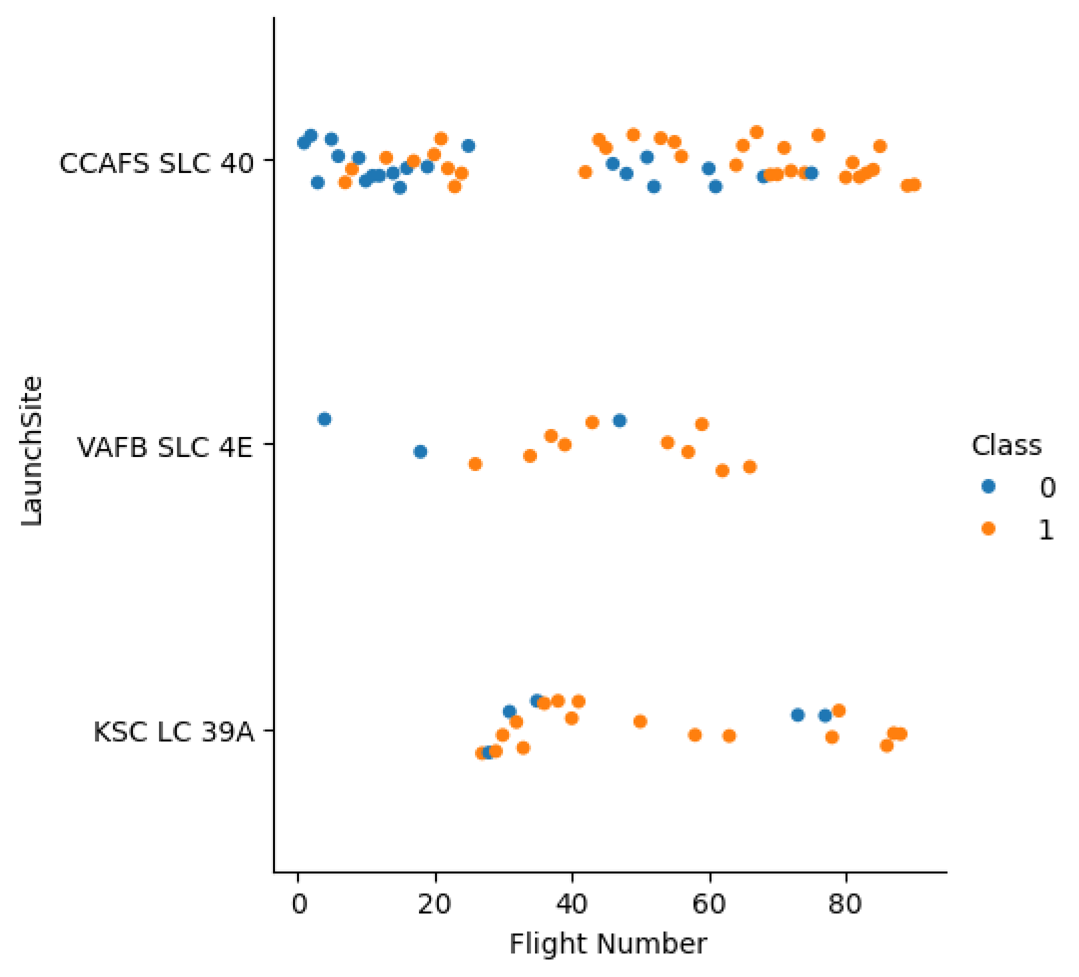
+
−
P
CCAFS SLC-40
26
AFS LC-40

Section 2

Insights drawn from EDA

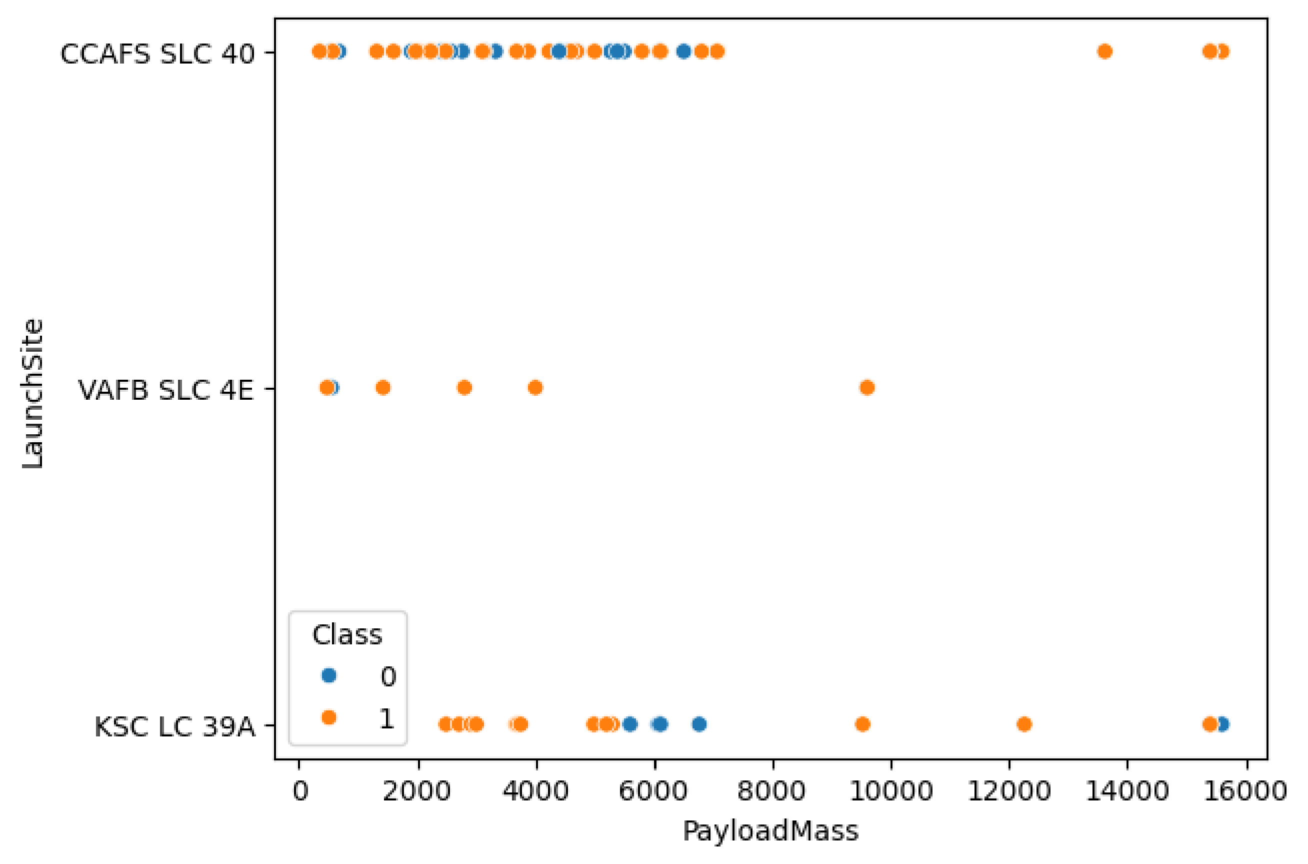
Flight Number vs. Launch Site

- As the flight number increased, it seemed that the successfull outcomes described by class 1 appeared more often, this could be extremely related to the time evolution process. As more launches were done, the chances of a successfull landing increased



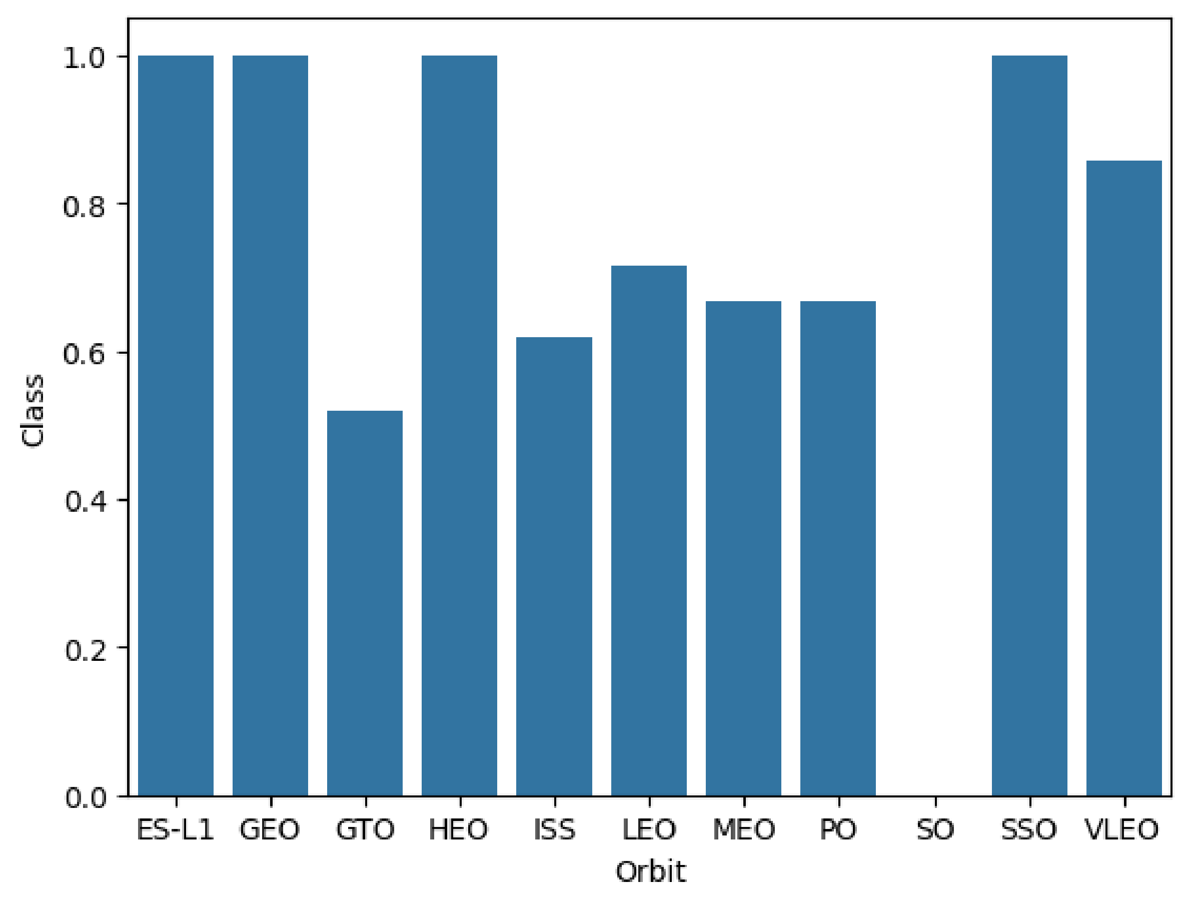
Payload vs. Launch Site

- As the payload increases, it is shown that there are less launches. However, in VFAB SLC 4E seems to have a better behaviour for success launches than other Launch Sites. Also as the number of launches are done in a Launch Site, it seems more difficult to see a correlation with the Class outcome.



Success Rate vs. Orbit Type

- We can see that ES-L1, GEO, HEO and SSO orbits have a perfect success rate (The 100% of the missions were succesfull). The others (except of SO) had a rate higher to 50%
- For the SO Orbit, all the launches had a failure outcome as its rate is 0%

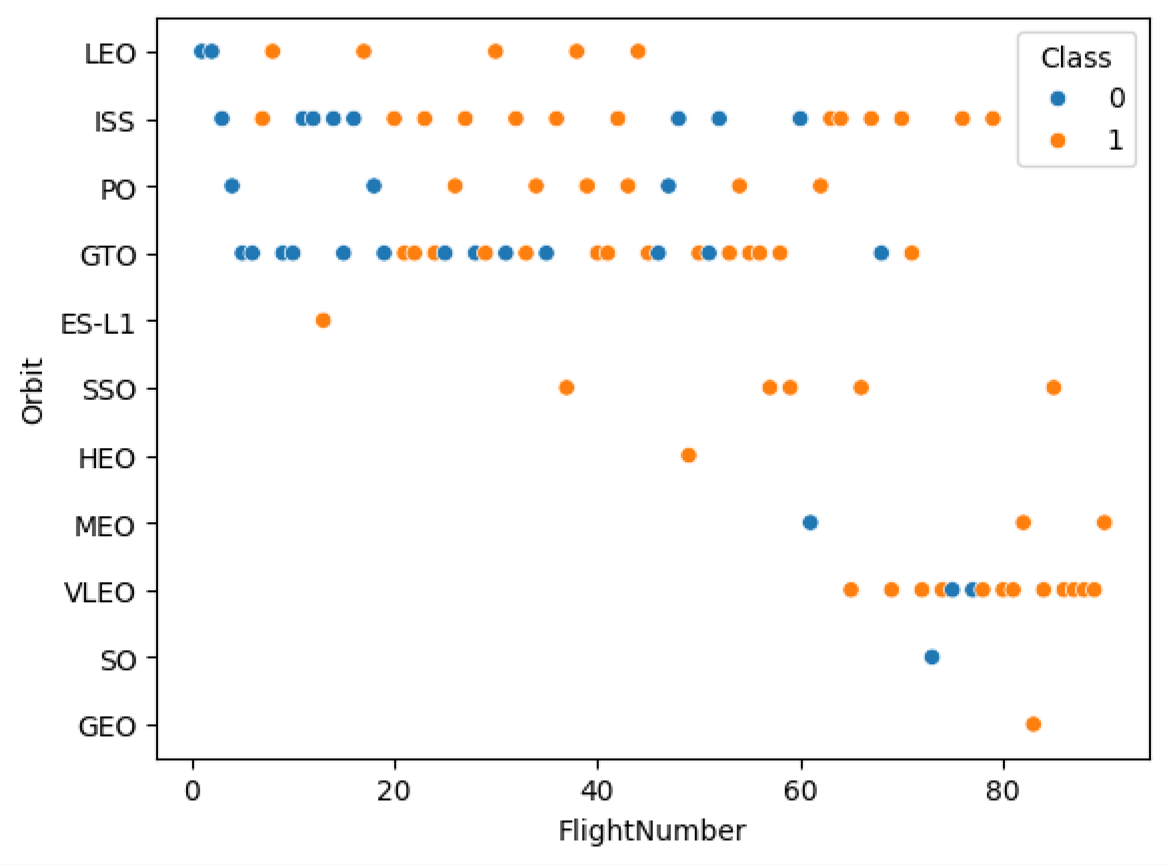


Flight Number vs. Orbit Type

- For LEO and MEO Orbit it seems that from an specific flight the outcomes started to be successful.

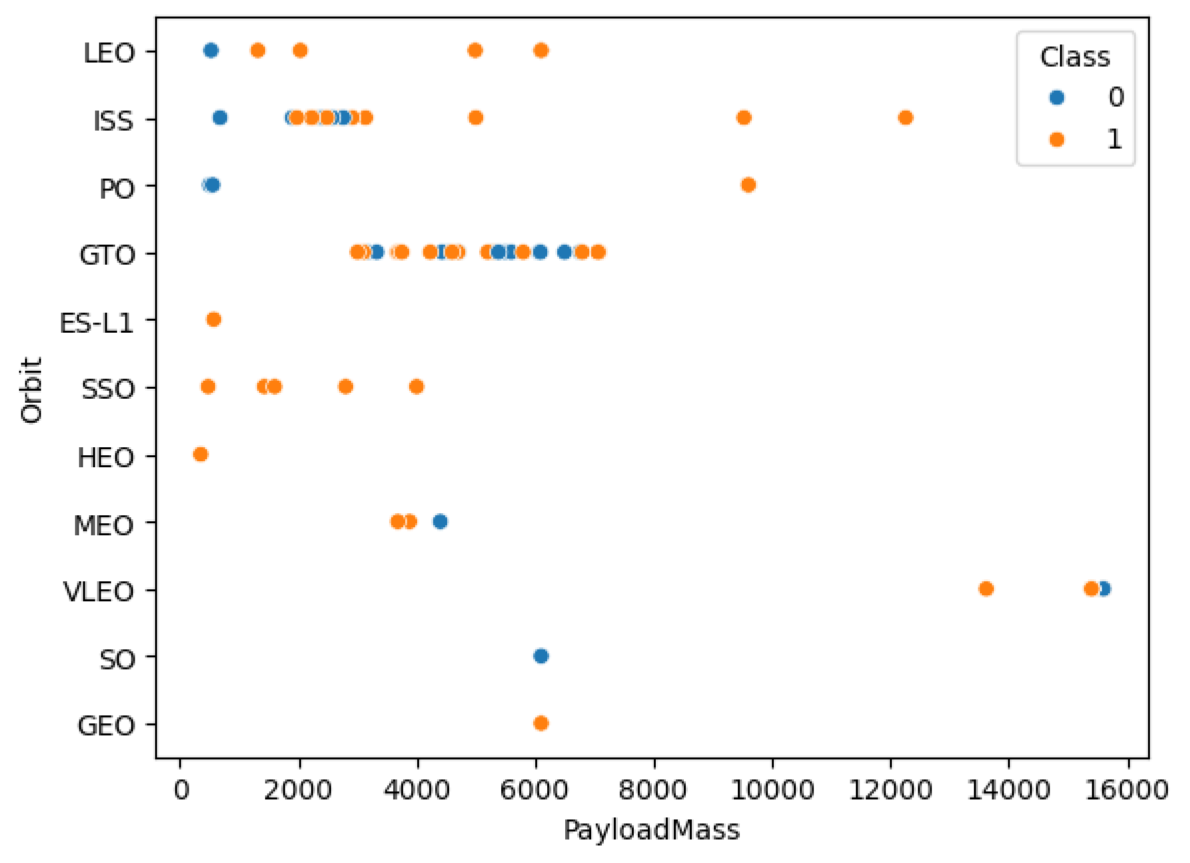
- For SSO, GEO, ES-L1 and HEO all the outcomes were successful no matter the flight number.

- For the other Orbits, it is not easy to distinguish a correlation between the variables.



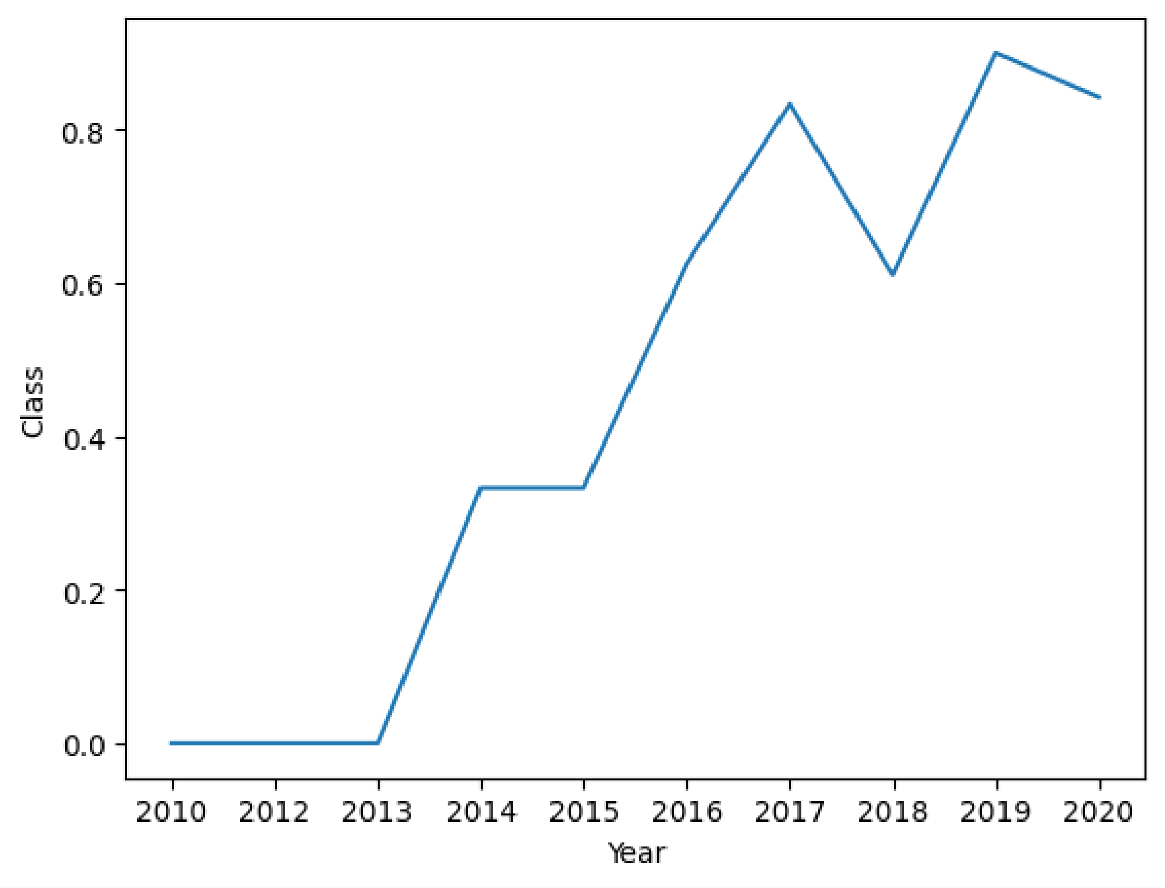
Payload vs. Orbit Type

- It is easy to predict the outcome as the payload increases in some Orbits like LEO, ESL-1, VLEO, SSO and HEO.
- With Orbits with more launches associated, it is not distinguished a correlation as with the flight number.



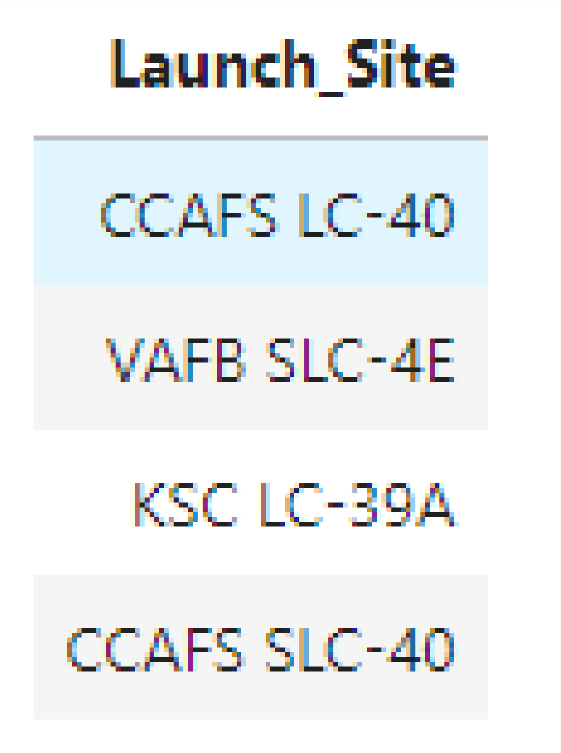
Launch Success Yearly Trend

- The result for this plot is interesting. It show us that the successfull cases have been increasing through the years as it seems that Space Engineers get more experience and insights to improve their models and misión tasks.



All Launch Site Names

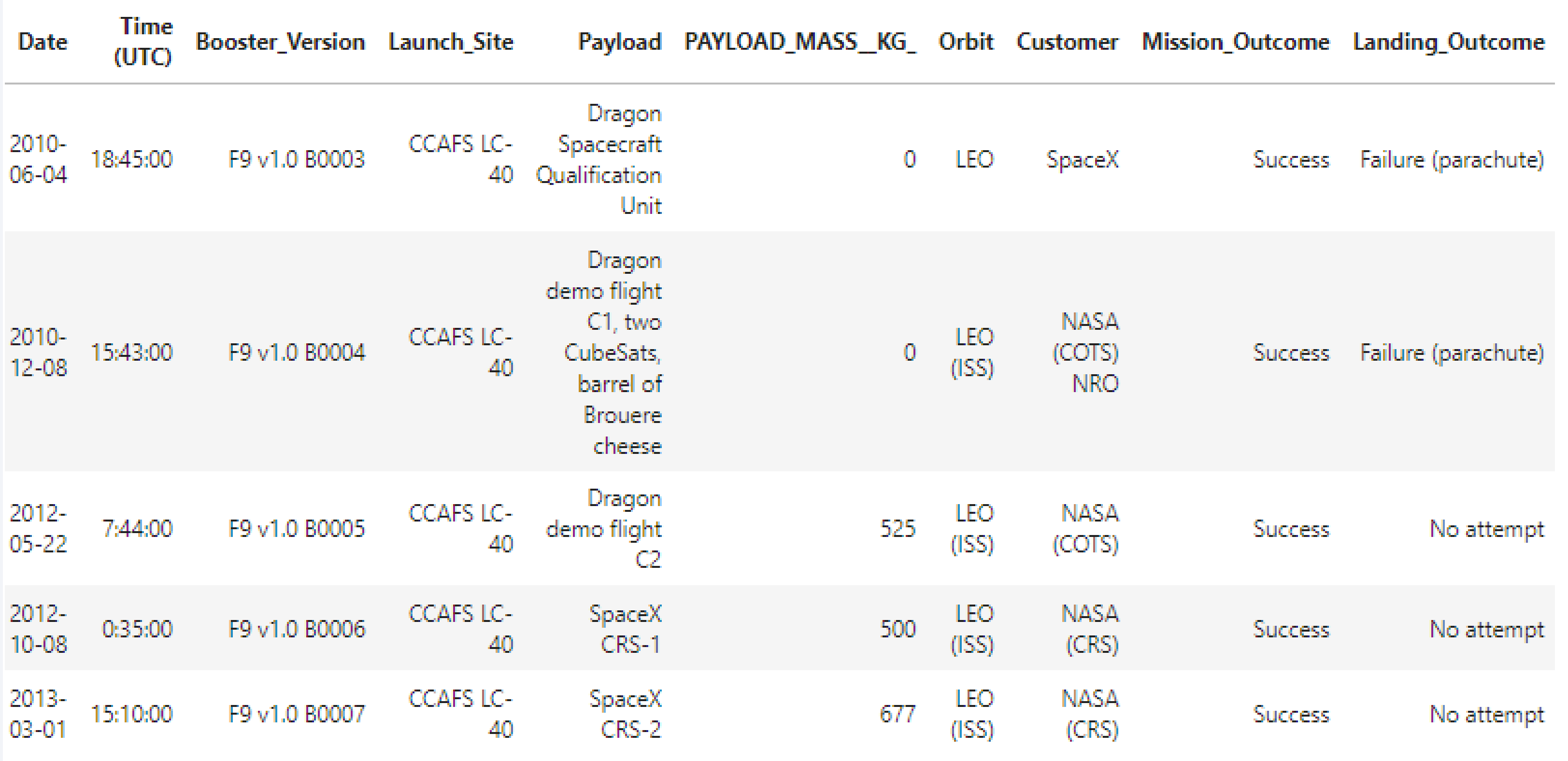
- There were just four different Launch Sites. This was a good new, so that it meant that a Exploratory Analysis could be possible through visualization as the plots done would not have a massive amount of information that could be difficult to interpret.



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

Launch Site Names Begin with 'CCA'

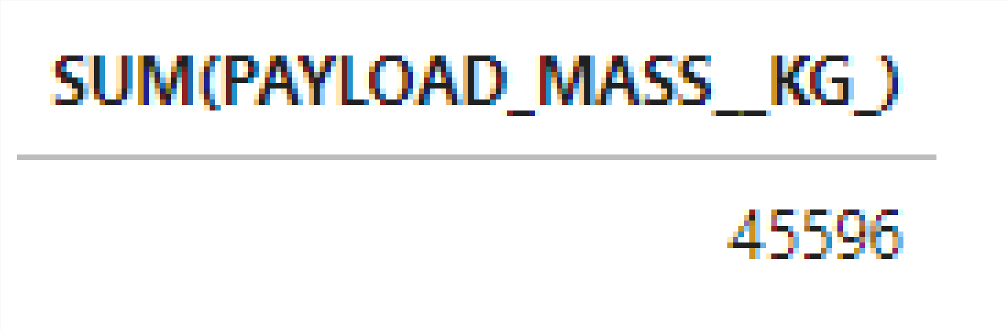
- As the Launch Site with this prefix was easy to get, it was shown that a filtering process done by the Launch Site was possible and feasible



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

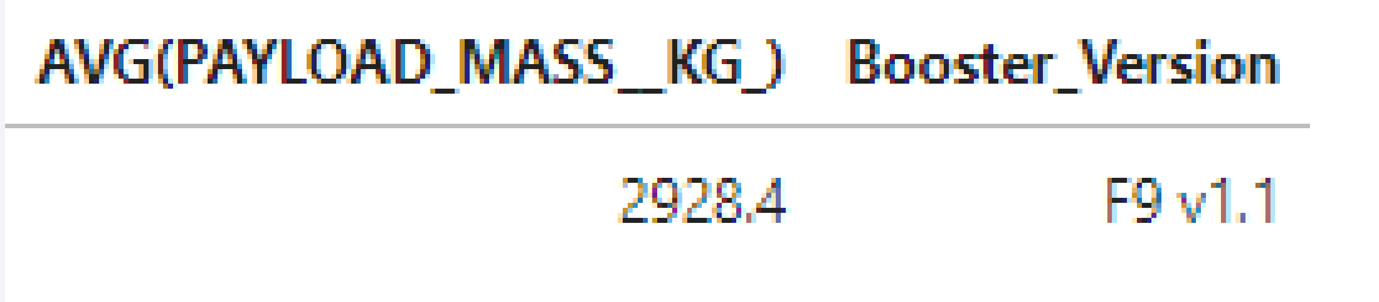
- As it was a single query, it could shown us that NASA missions carried a total Payload Mass higher to 45 tons.



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

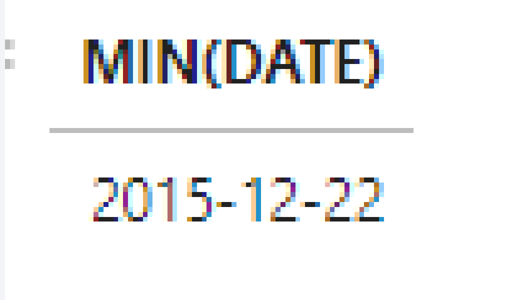
- As the query showed, the mentioned Booster version carried an average payload Mass of 2.9 tons



AVG(PAYLOAD_MASS__KG_)	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

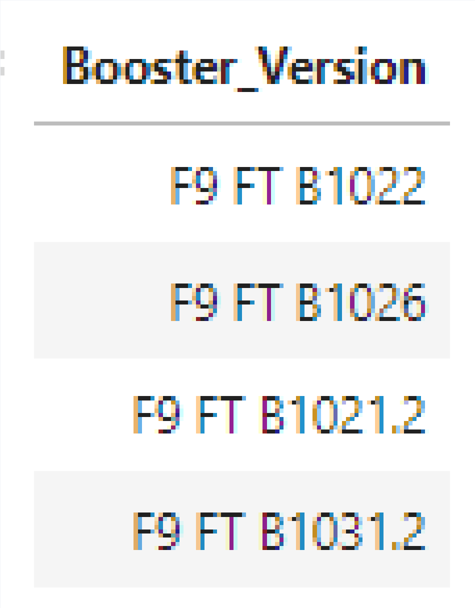
- With the data, we could know that the first successful ground landing date was in 2015



MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

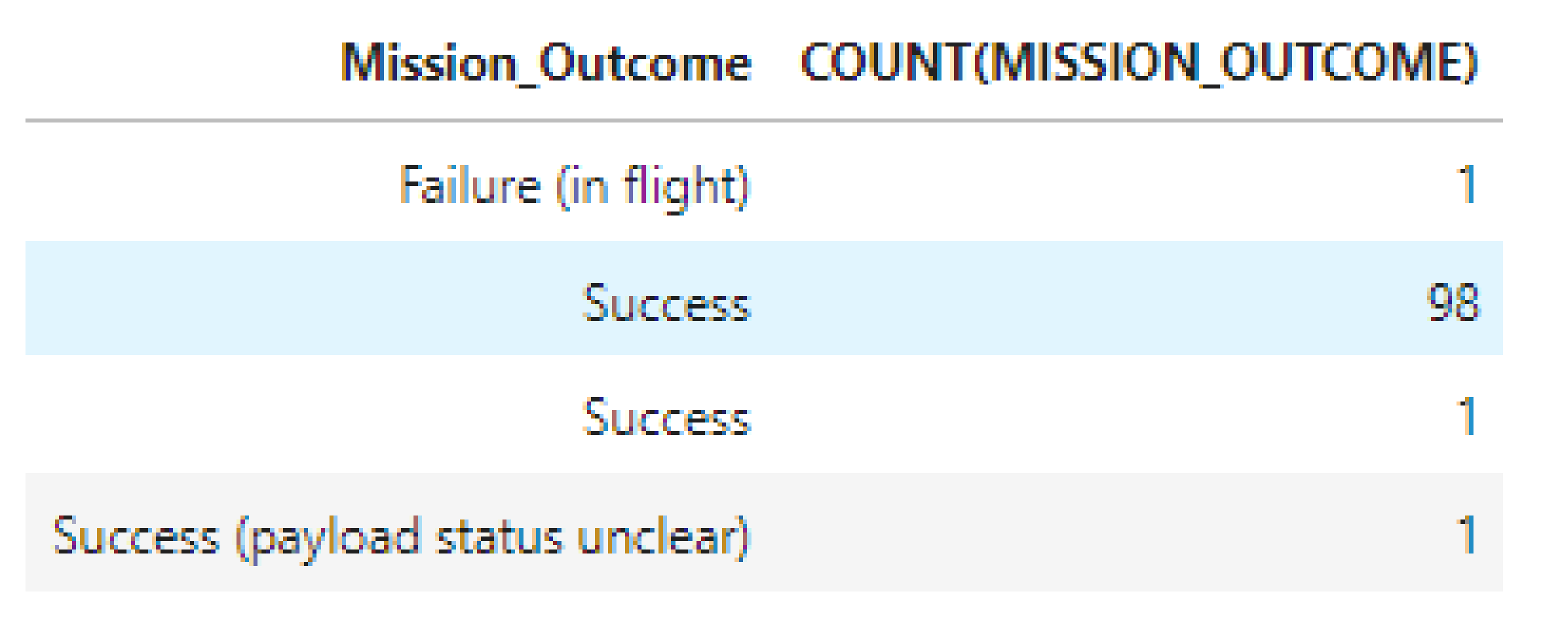
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 can be seen in the following query result:



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

Total Number of Successful and Failure Mission Outcomes

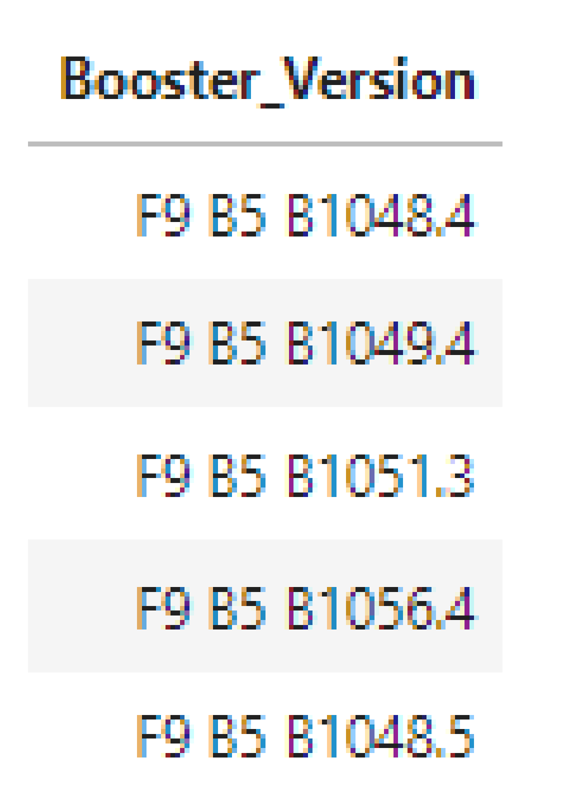
- The total number of successful and Failure outcomes can be seen in the following query result:

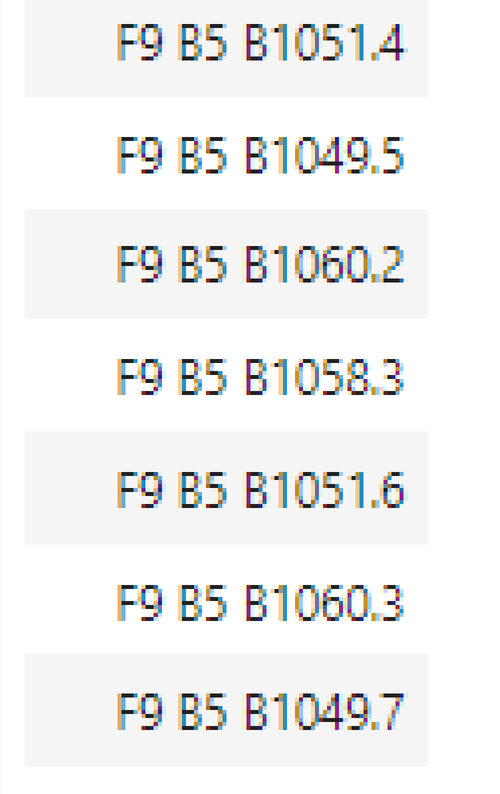


Mission_Outcome	COUNT(MISSION_OUTCOME)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The following query result shows which boosters carried the Maximum Payload

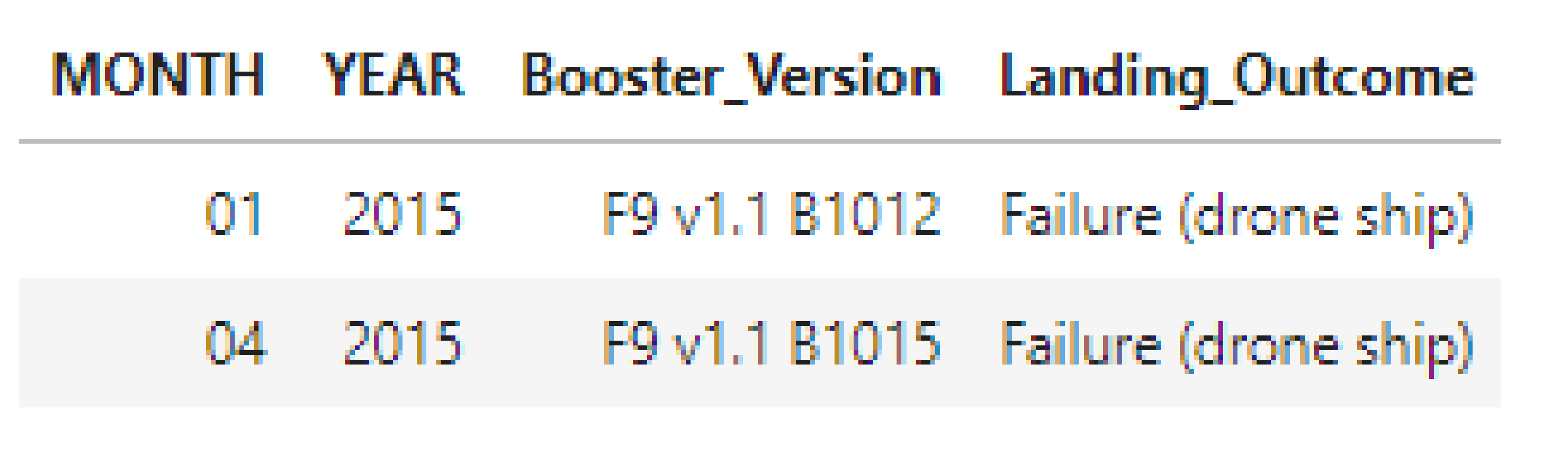




Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

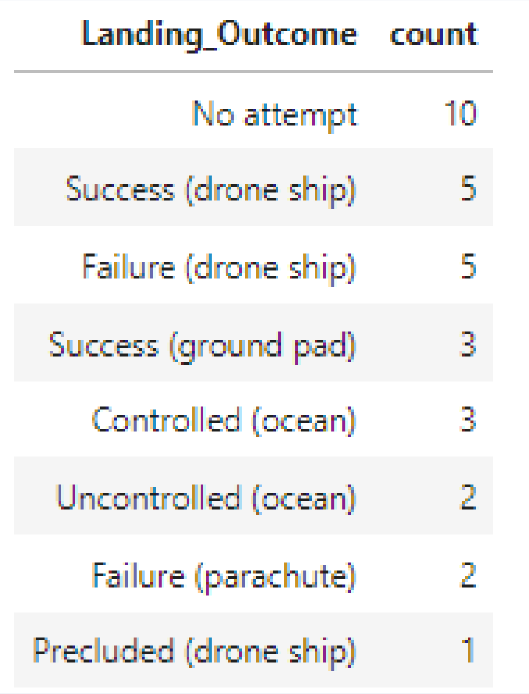
- The failed 2015 records can be seen in the following query result:



MONTH	YEAR	Booster_Version	Landing_Outcome
01	2015	F9 v1.1 B1012	Failure (drone ship)
04	2015	F9 v1.1 B1015	Failure (drone ship)

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

- The result of the query that ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order is the following:



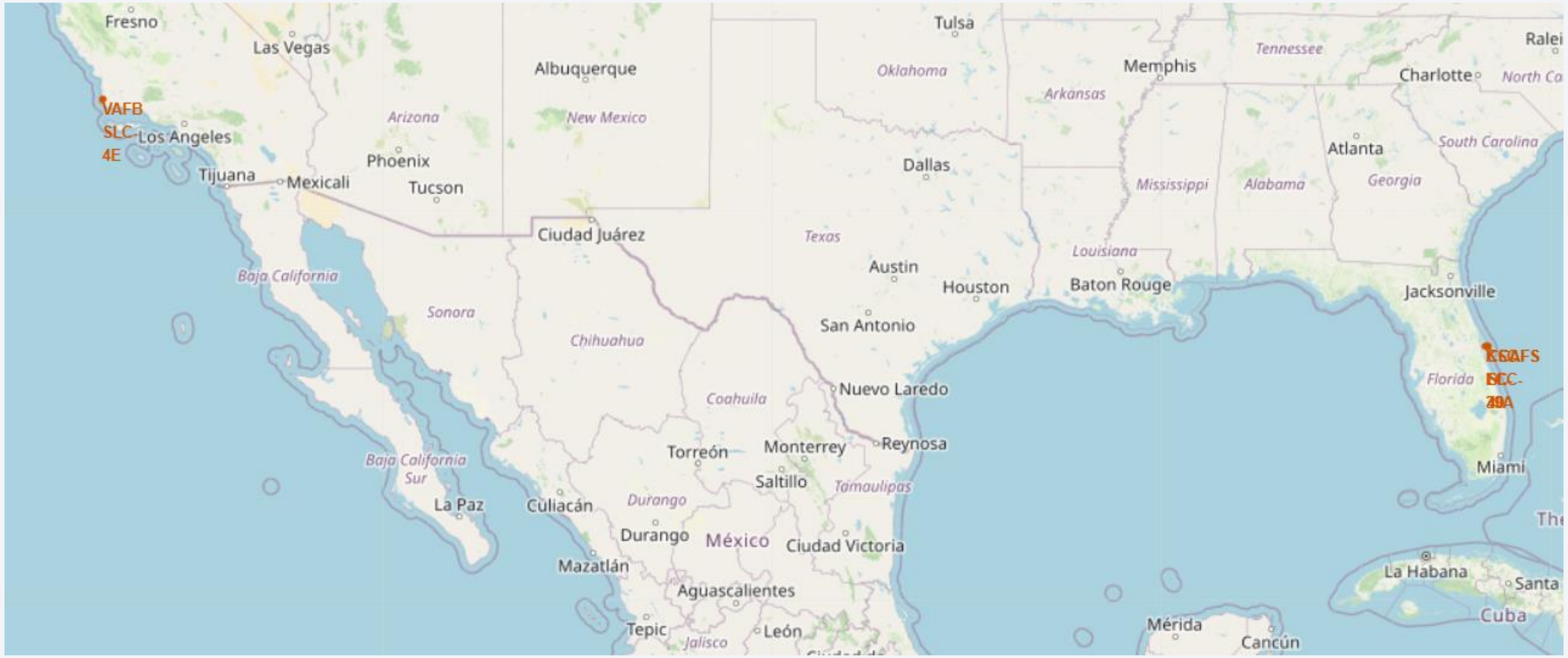
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

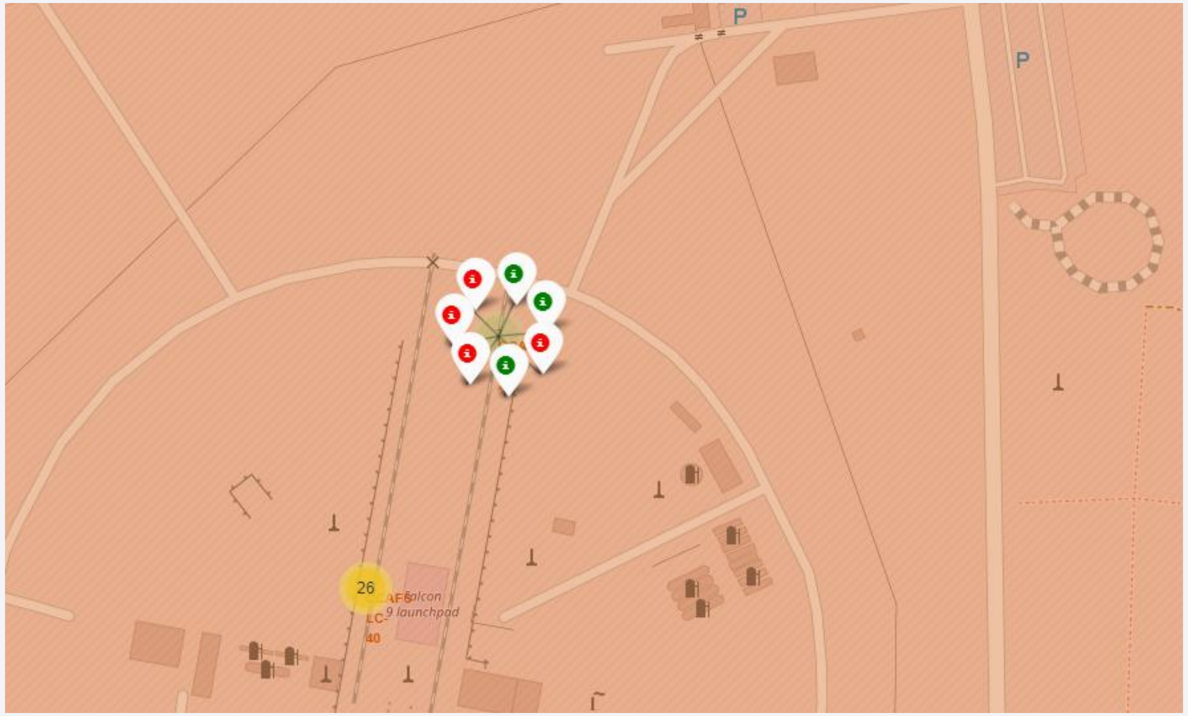
Launch Sites Global locations

- In the image, we can see that the location sites are grouped in two main clusters in USA. The first one is in Florida and the second one in California.



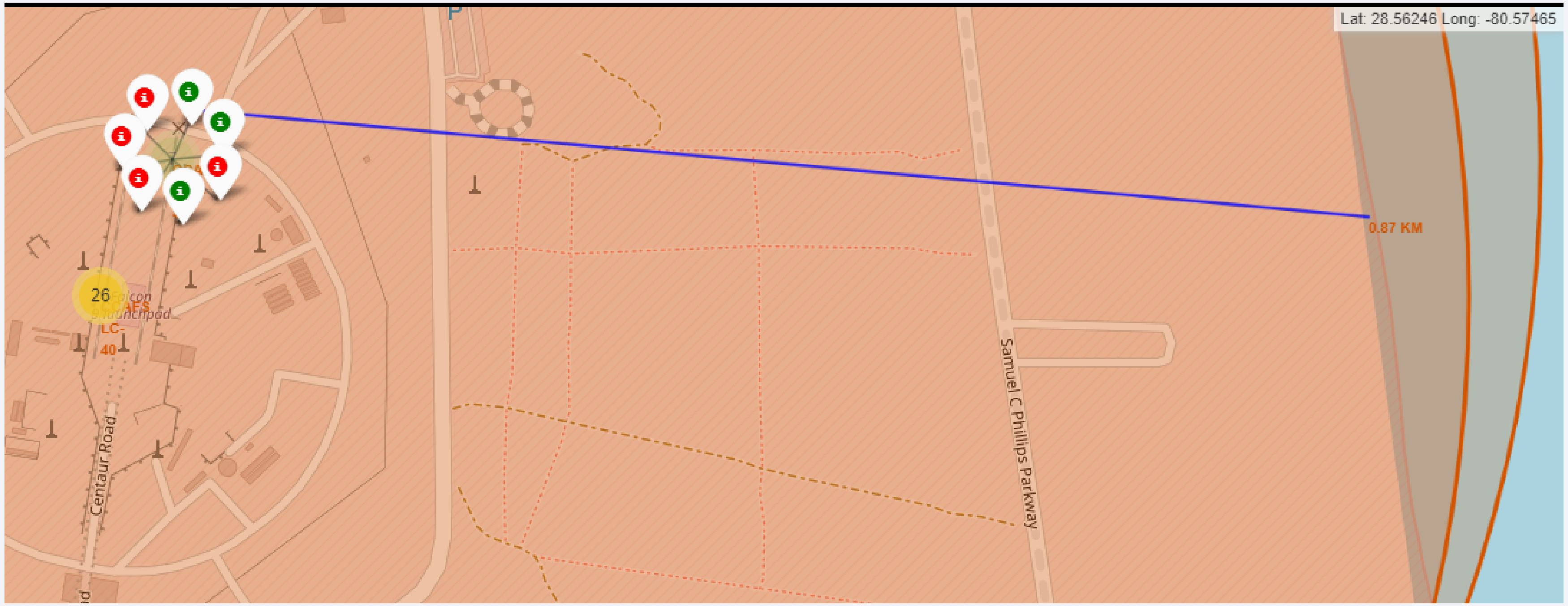
Launch Outcomes Map

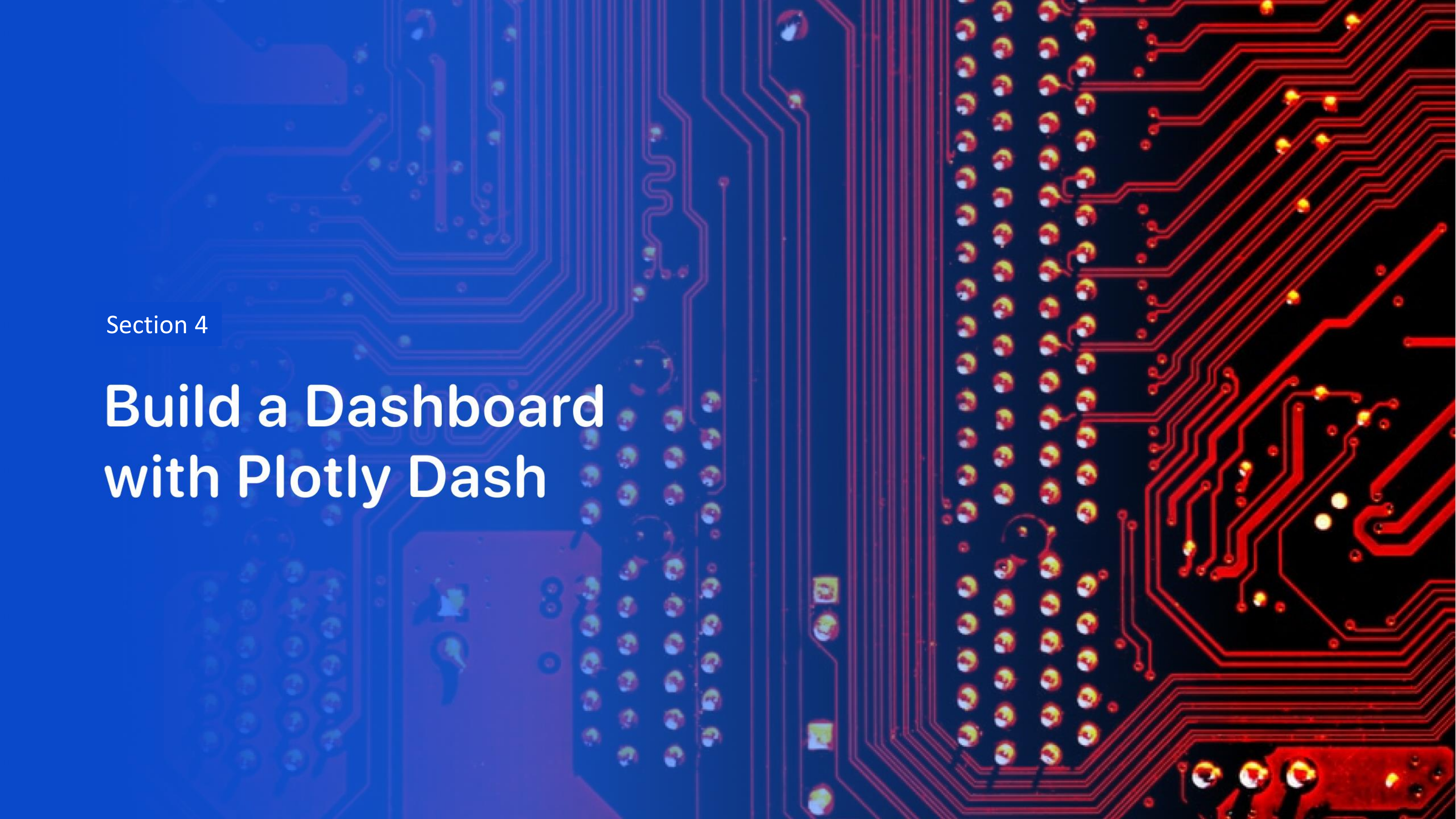
- As we can see in this map, the different Launch outcomes are showed in specific coordinates as they are successful (green) or failed (red)



Launch Sites Locations

- As we can see in this map, the different launch sites are near some interests points and some metrics to get its distance as the specific coordinates could de obtained from the map.



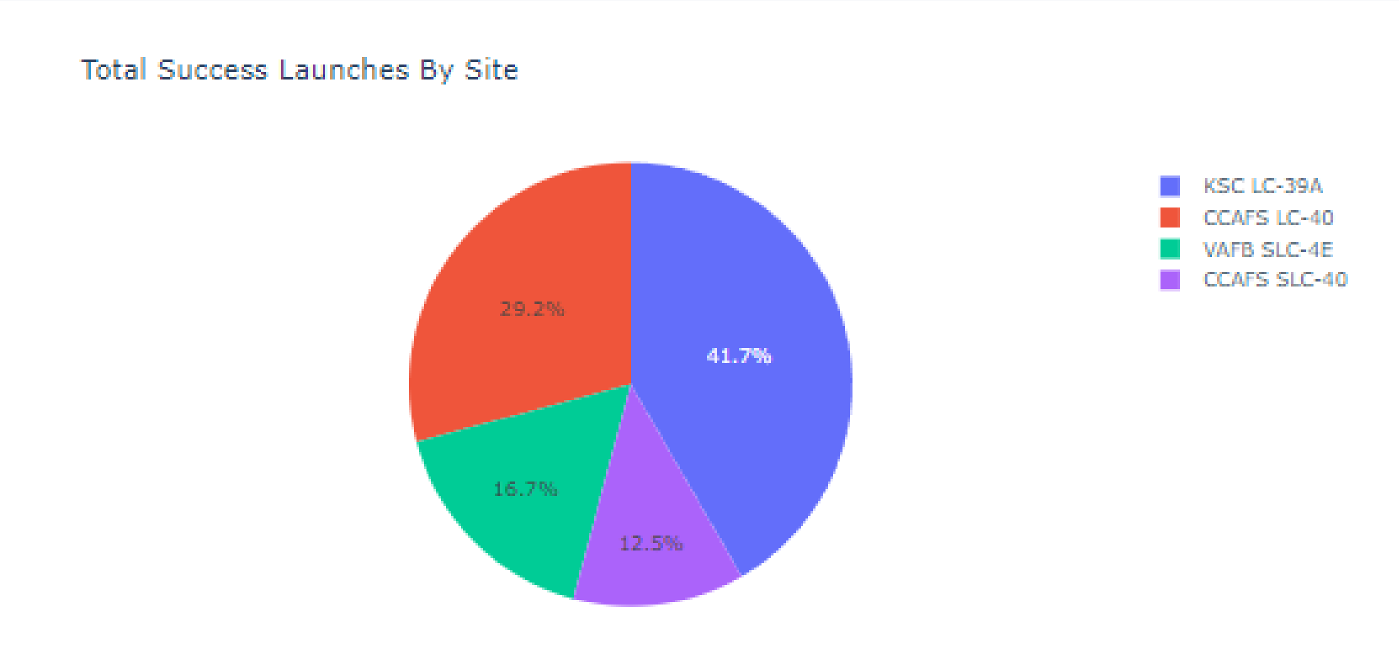


Section 4

Build a Dashboard with Plotly Dash

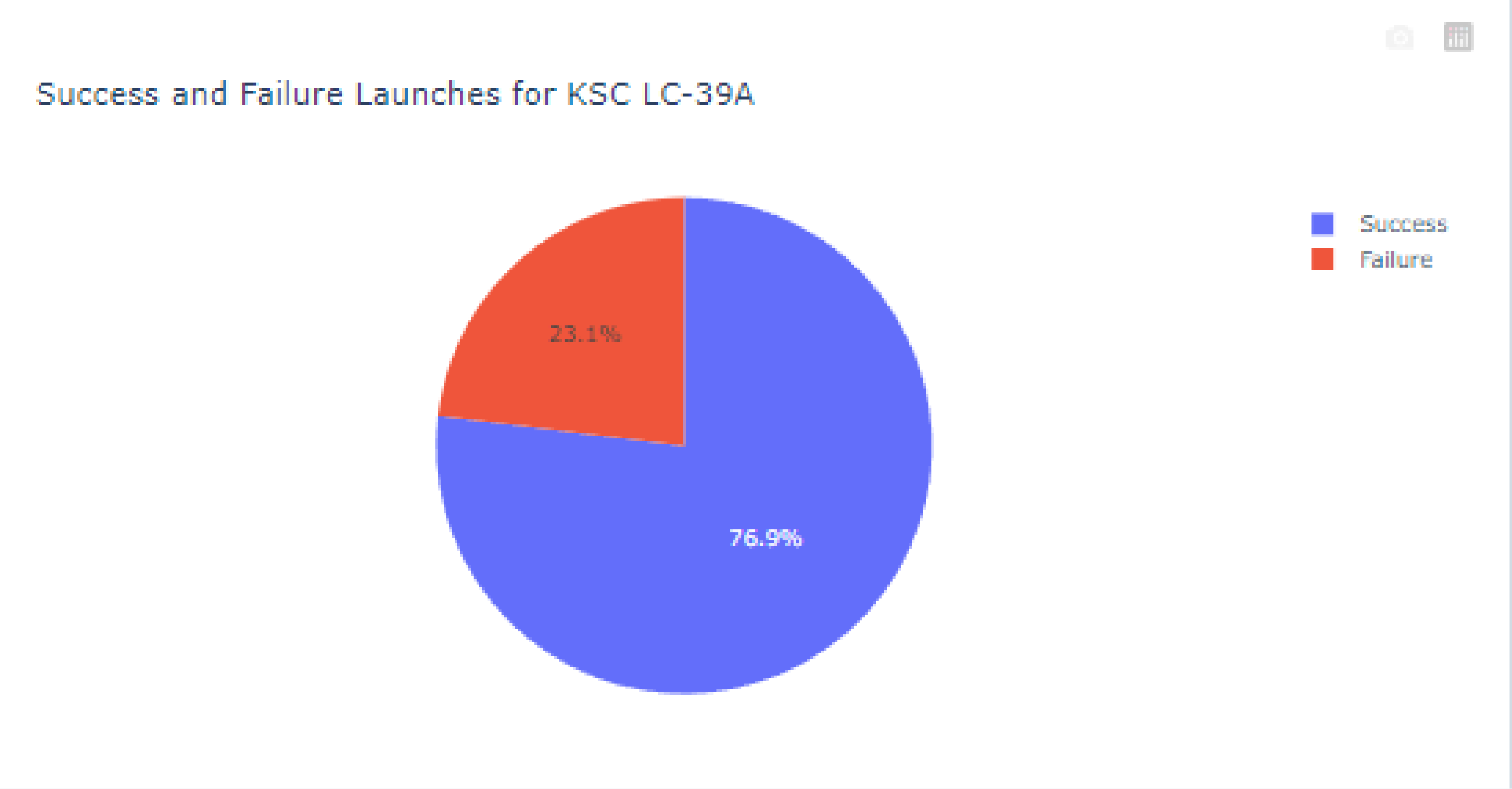
Launch Success for all sites

- The following dashboard screenshot shows the successful launches distributed by each Launch Site. It can be seen that KSC LC-39A has the largest count of successful outcomes



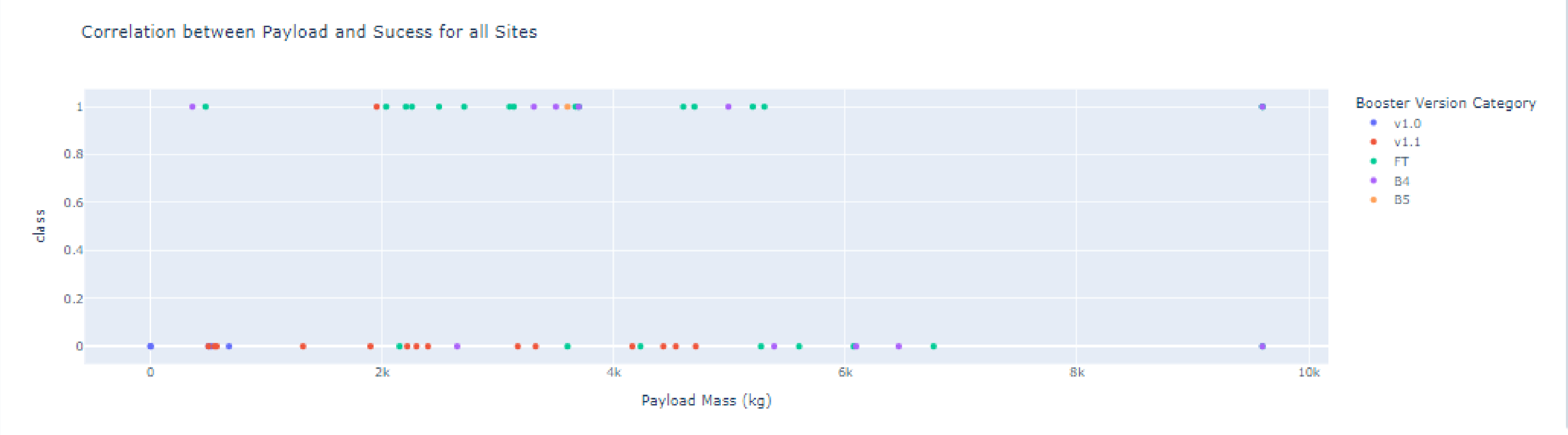
The impressive success rate for KSC LC-39A

- As we could see that KSC LC-39A had the largest count of success outcomes, it was not too difficult to think that it had the best success rate. The following chart confirms it



The Payload against the Launch Outcome

- As we can see, there is not a high remarkable tendency between payload mass and the launch outcome. However, we can see that depending on the Booster Version, the behavior seems more interesting.

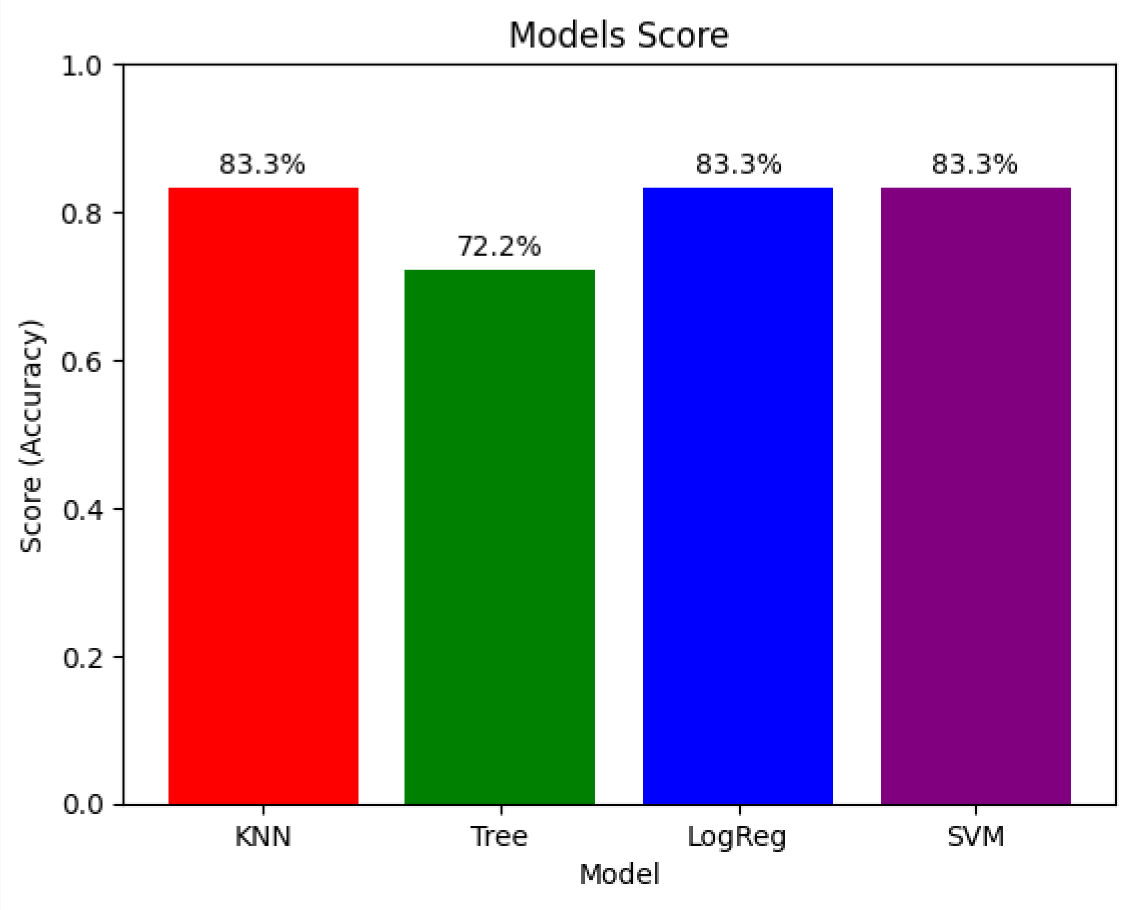


Section 5

Predictive Analysis (Classification)

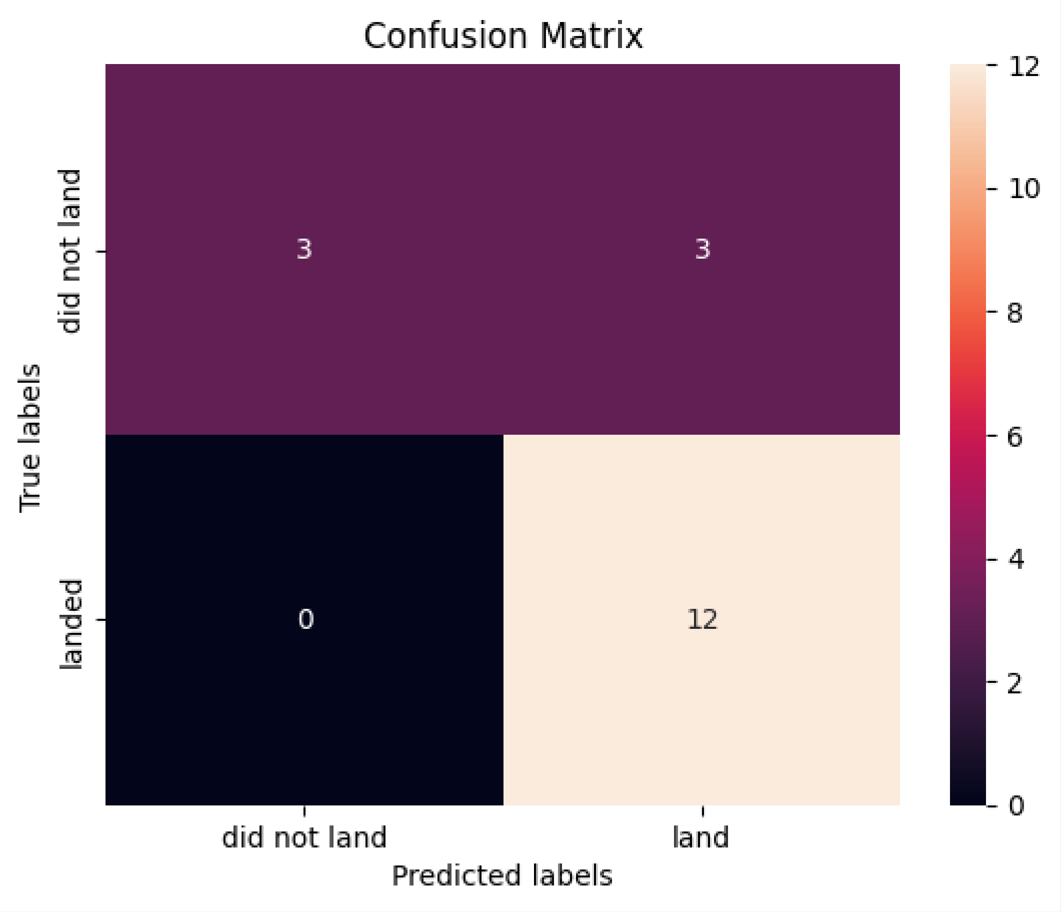
Classification Accuracy

- The bar chart helps us to visualize the accuracy for the 4 different classification models.
- If we detail the plot, we can see that the K-Nearest Neighbors, the Logistic Regression and The Support Vector Machine models have the same and highest accuracy with 83.3%



Confusion Matrix

- As the three models have exactly the best performance, we will take the confusion matrix of just one of them.
- As we can see in the results, we are getting just Type I error (False Positive, or in our case a landing predicted when it was bad).



Conclusions

- As time passes through we can see that the successful rate increases as more knowledge is obtained from previous success and failure situations
- In the predictive task, many models had a very similar performance; however, for time issues, it is better to use just the Logistic Regression model as it is the easiest to compute.
- After the EDA through SQL and Visualization, just some variables were taking into account and it could help to get a good performance in the predictive task
- After using the Folium map, it could be seen that the launch sites were not too different between them if they were in the same state (They share many similar near places and interest place distances)

Appendix

- As it is a good practice to include the complete material used for the visualizations and analysis. In the following GitHub link it can be found all the notebooks, codes and files generated during the project development:

  IBM SpaceX Project Capstone Full Development

Thank you!