IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Santiago Montenegro Quiroga
26/03/2024

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

- **Summary of methodologies:**

- The data collection was using web scraping and the SpaceX Api, also, exploratory data analysis including data wrangling, data visualization and interactive visual analytics, and the use of machine learning prediction.

- **Summary of all results:**

- From the project, it was possible to extract valuable data from the sources, also, the machine learning prediction showed the best model to predict which characteristics are fundamental to choose the best way, and the exploratory data analysis allowed to identify which features are the best to predict the success of the launching.

Introduction

Project background and context:

- The SpaceX is a rocket company that launches satellites at low price, about 70% less than their competitor because they land their satellites to reusing them.

Problems you want to find answers:

- We want to estimate the best way to calculate the total cost for the launching, by predicting successful landings of the first stage of rockets, and also, where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The data used from SpaceX was obtained from two principal sources:

  - The API from SpaceX: https://api.spacexdata.com/v4/rockets/
  - WebScraping from the list of falcon 9: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

- Perform data wrangling:

  The collected data was enriched by creating a landing outcome label based on resulting data after summarizing and analyzing features to achieve the objective.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL: This will be presented along the presentation.
- Perform interactive visual analytics using Folium and Plotly Dash: This will be presented along the presentation.
- Perform predictive analysis using classification models

  Here, the data is normalized, divided into training and test data sets, and then is evaluated by the different classification models, being the qualification parameter the accuracy of each one.

Data Collection – Falcon 9

We request the Falcon9 web page

Then we extract the column and variable names from the header

Finally, we create a data frame by parsing the launch HTML tables

Data source:
https://github.com/Santiago2301/Data_Science_Capstone/blob/8d661c2b76baf5ab06433d42347305d6ced59466/DataCollection_WebScraping.ipynb

Data Collection – SpaceX API

We request the API and parse the SpaceX launch data

Then we filter data to only include the Falcon9 information launches

Finally, we deal with the missing values

Data source:
https://github.com/Santiago2301/Data_Science_Capstone/blob/0a2395b821ee445b3de2540268f9de1857ed9dd4/Data_Collection_API.ipynb

Data Wrangling

We do the exploratory data analysis on the data set

Then the summaries launches per site, ocurrences of each orbit and occurrences of mission outcome per orbit type were calculated

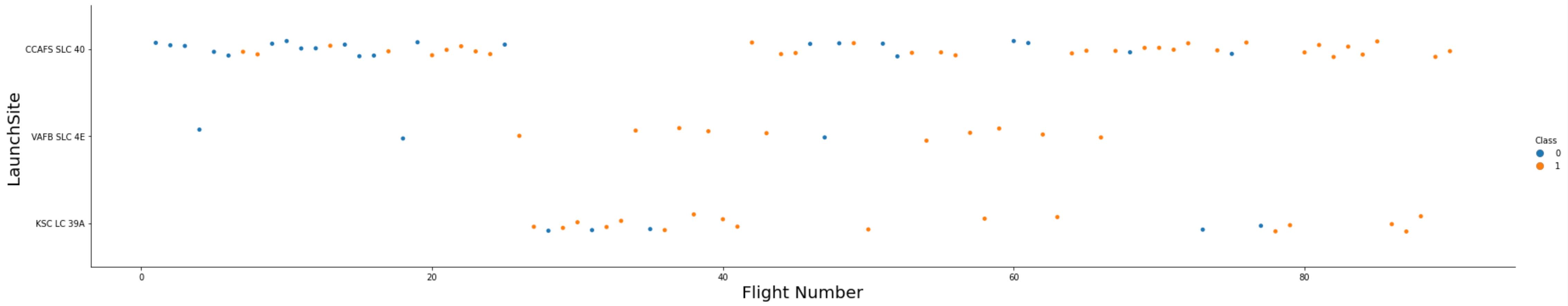
Finally, the landing outcome label was created from the outcome column

Data source:
https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/Data%20Wrangling.ipynb

EDA with Data Visualization

- To explore the data, we use barplots and scatterplots to visualize the relationship between the pair of features, where we analyze the launchsite, flight number, payload mass and orbit.

- Source code: https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/Data%20Wrangling.ipynb



EDA with SQL

- The following SQL queries were used:

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; and Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20.

- Source code: https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/EDA.ipynb

Build an Interactive Map with Folium

- We use markers, circles, lines and marker clusters in the Folium Maps:

Markers indicate points like launch sites.

Marker clusters indicates groups of events in each coordinate, like launches in a launch site.

Lines are used to indicate distances between two coordinates.

Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center.

- Source code: https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/Interactive_VisualAnalytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- For the Dashboard, we use graphs and plots for the percentage of launches by site, and payload range in order to visualize data.
- This allows use to quickly analyze the relation between payloads and launch sites, helping us to identify where is the best place to launch according to the payloads information.
- Source code: https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/spacex_dash_app.py

Predictive Analysis (Classification)

We select the four classification models, logistic regression, support vector machine, decision tree and k nearest neighbors.

We prepare the data and standardize them

We test each model with combinations of hyperparameters

We compare the results

Data source:
https://github.com/Santiago2301/Data_Science_Capstone/blob/c223e72b280949fd07a0fd674e2736d03e68b18b/Machine_Learning_Prediction.ipynb

Results

Exploratory data analysis results:

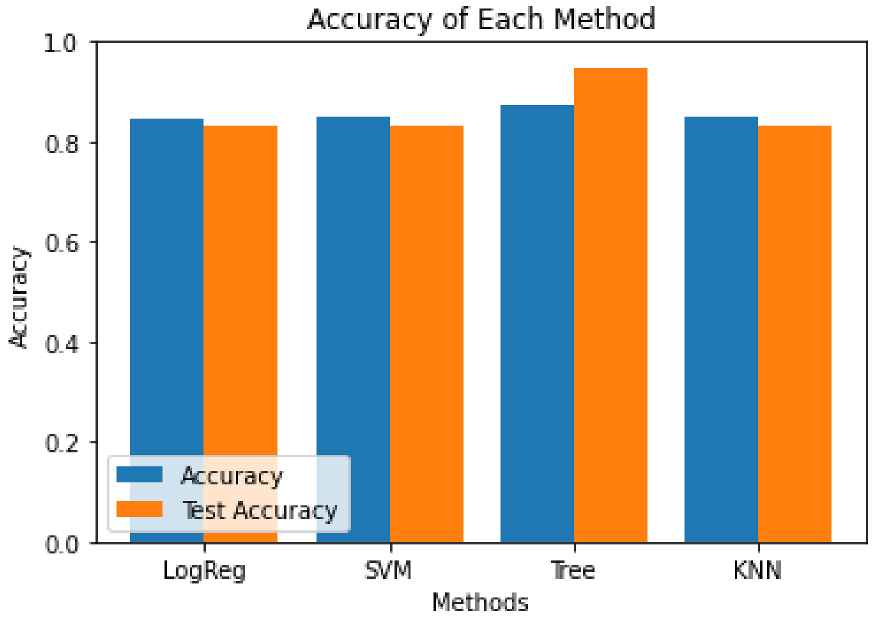
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 fiver year after the first launch
- Space X uses 4 different launch sites
- The first launches were done to Space X itself and NASA
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The number of landing outcomes became as better as years passed

Results

- Through the use of interactive analytics allowed us to identify the launch sites that used to be in safety places, like near the sea for example, and as we can see, most launches happens at east coast launch sites:

Results

- Also, the predictive analysis showed that the decision tree classifier is the best model to predict the correct landings, having accuracy over 87% and accuracy for the test data over 94%:

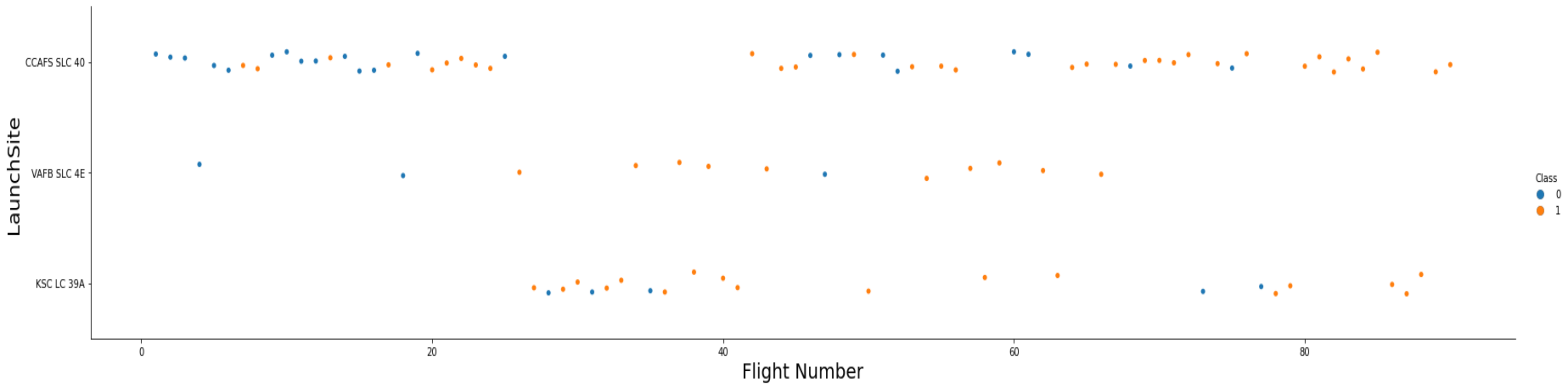


Section 2

Insights drawn from EDA

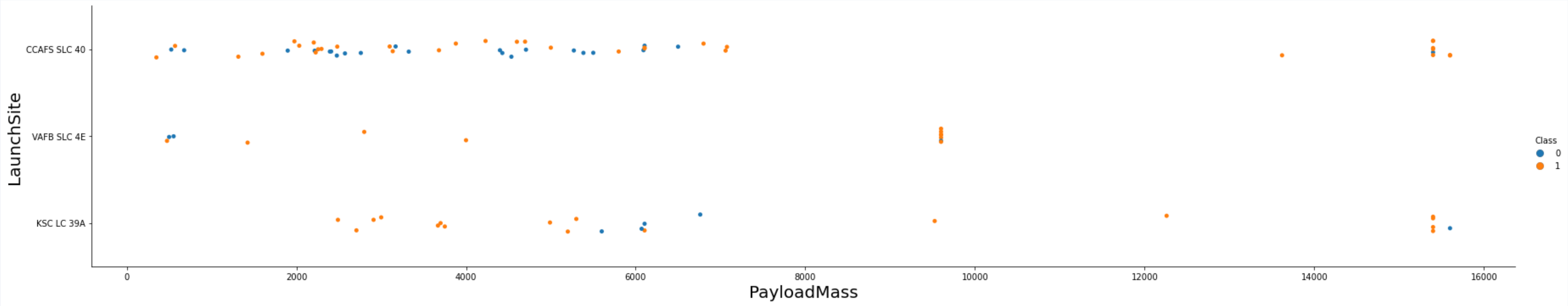
Flight Number vs. Launch Site

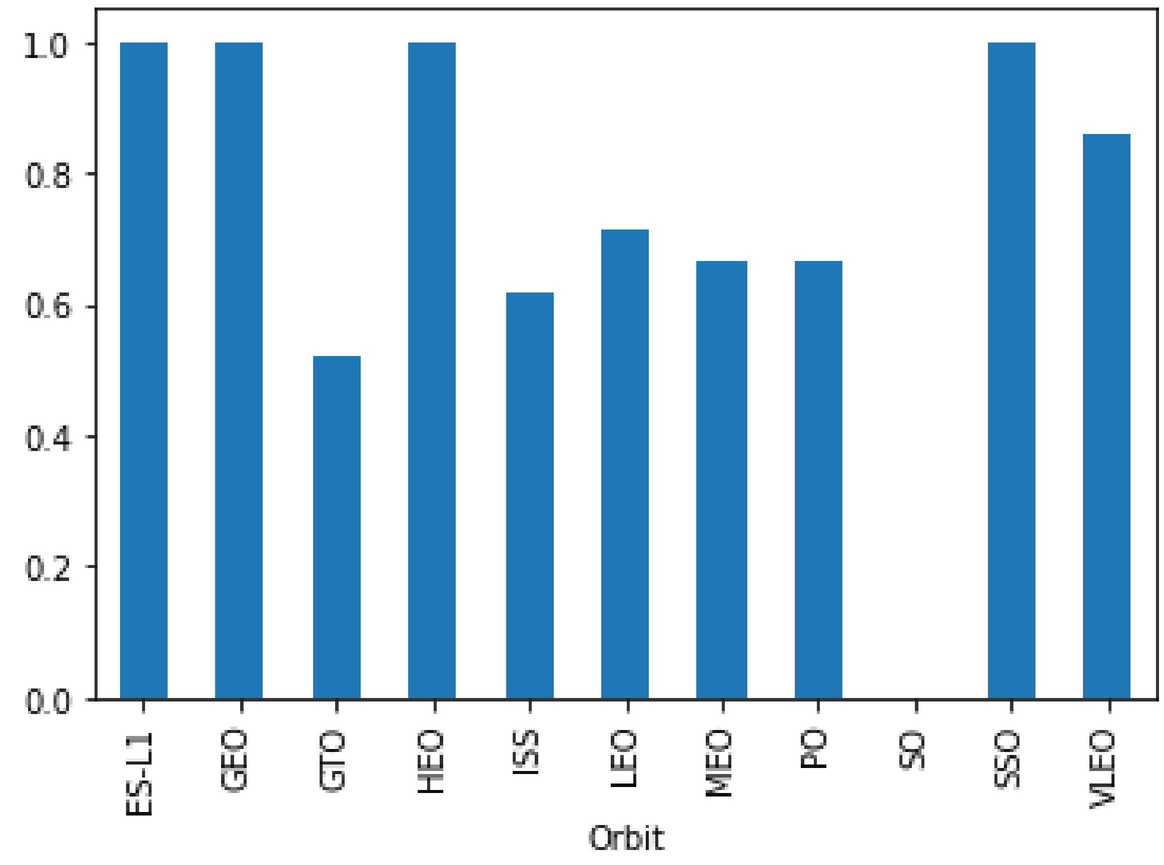
- According to the plot above, we can verify that the best launch site is CCAF5 SLC 40, where most of the recent launches were successful, then, there is the VAFB SLC 4E and finally the KSC LC 39 A, and this also shows that the general success rate has improved over time.



Payload vs. Launch Site

- From the graph we can say that payloads over 9,000kg (about the weight of a school bus) have excellent success rate, and payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



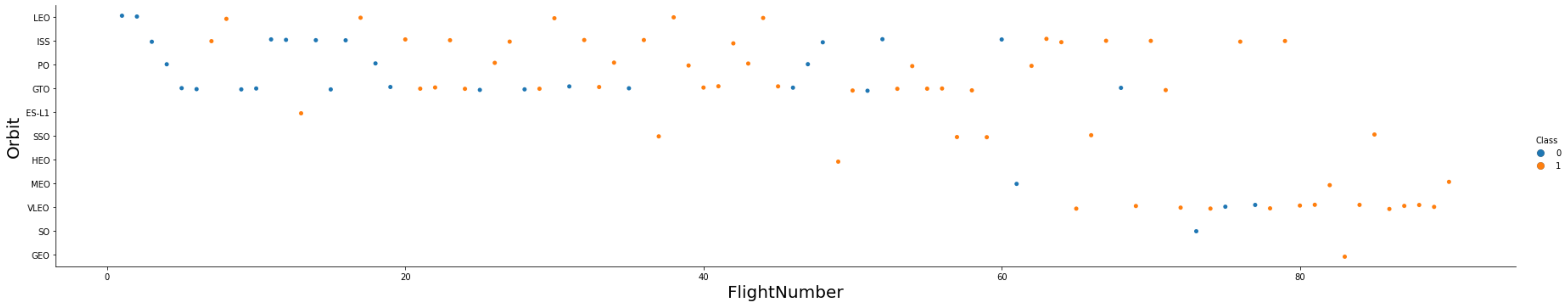


Success Rate vs. Orbit Type

The biggest success rates happens to the orbits ES-L1, GEO, HEO and SSO.

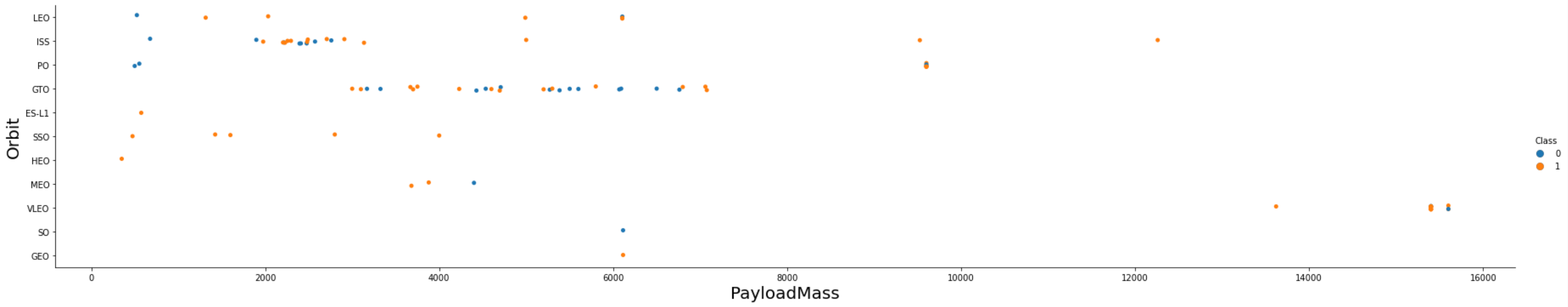
Flight Number vs. Orbit Type

- From the graph, we can say that the success rate has improved over time to all orbits, and that VLEO orbit seems like a big opportunity, due to recent increase of its frequency.



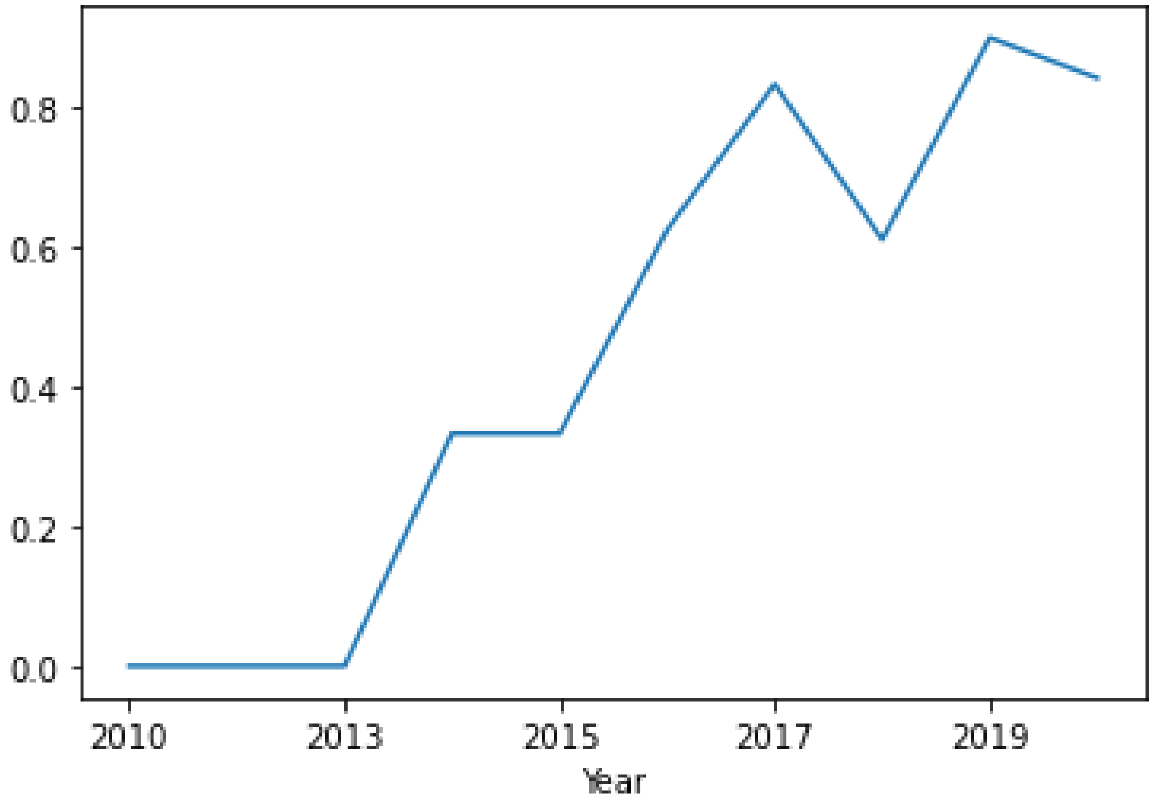
Payload vs. Orbit Type

- From this graph, we can say that ISS orbit has the widest range of payload and a good rate of success, and that there are few launches to the orbits SO and GEO, but, there is no relation between payload and success rate to orbit GTO.



Launch Success Yearly Trend

- The data shows that success rate started increasing in 2013 and kept until 2020, this seems to mean that the first three years were a period of adjusts and improvement in technology.



All Launch Site Names

- By selecting unique ocurrences of “launch_site” values from the data set, we can get four launch sites, these are:
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- By using some specific queries, we can find 5 samples of cape Canaveral launches, we can see them in the following graph:

Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
4/06/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
8/12/2010	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1/03/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass - Average Payload Mass by F9 v1.1

- We can calculate the total payload calculated above, by summing all payloads whose codes has the “CRS” key characters, which corresponds to NASA, having a total payload of 11628 KG.

- By filtering data by the booster version above and calculating the average payload mass we can obtain the value in KG, with a value of 2928.

First Successful Ground Landing Date

- By filtering data by successful landing outcome on the ground pad, and getting the minimum value for date it’s possible to identify the first occurrence, this happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

By selecting different booster versions according to the filters above, we can find 4 results, this are the boosters which have successfully landed on drone ship and that has a payload mass greater than 4000 but less than 6000, the booster versions are the following:

- F9 FT B1021,2
- F9 FT B1031,2
- F9 FT B1022
- F9 FT B1026

Mission Outcome	Occurrences
Success cases	99
Success (payload status unclear)	1
Failure (in flight)	1

Total Number of Successful and Failure Mission Outcomes

- By grouping mission outcomes and counting records for each group can led us to the summary above, and as a result, we can get a number of successful and failure mission outcomes:

Boosters Carried Maximum Payload

By using previous methods, we can find the boosters which have carried the maximum payload mass registered in the data set, this are the following:

- 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

2015 Launch Records

- Through the query search we can find only two occurrences of failing landing outcomes in drone ship, their booster version and launch site names in year 2015, this are the following:

Booster Version	Launch Site
F9 V1,1 B1012	CCAFS LC-40
F9 V1,1 B1015	CCAFS LC-40

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

- Here we can see the ranking of 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, meaning that this view of data alerts us that “no attempt” must be taken in account.

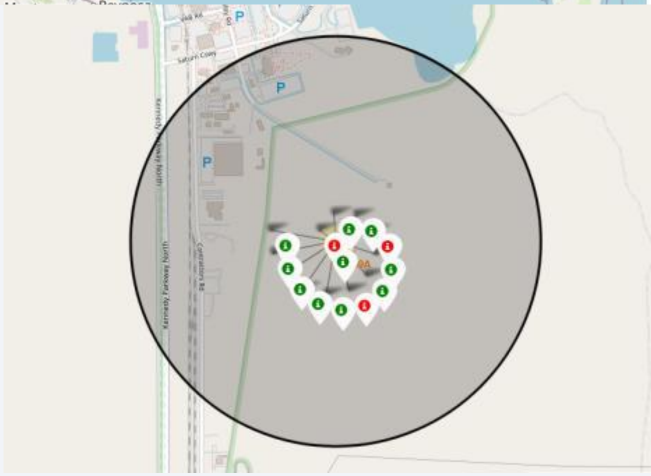
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

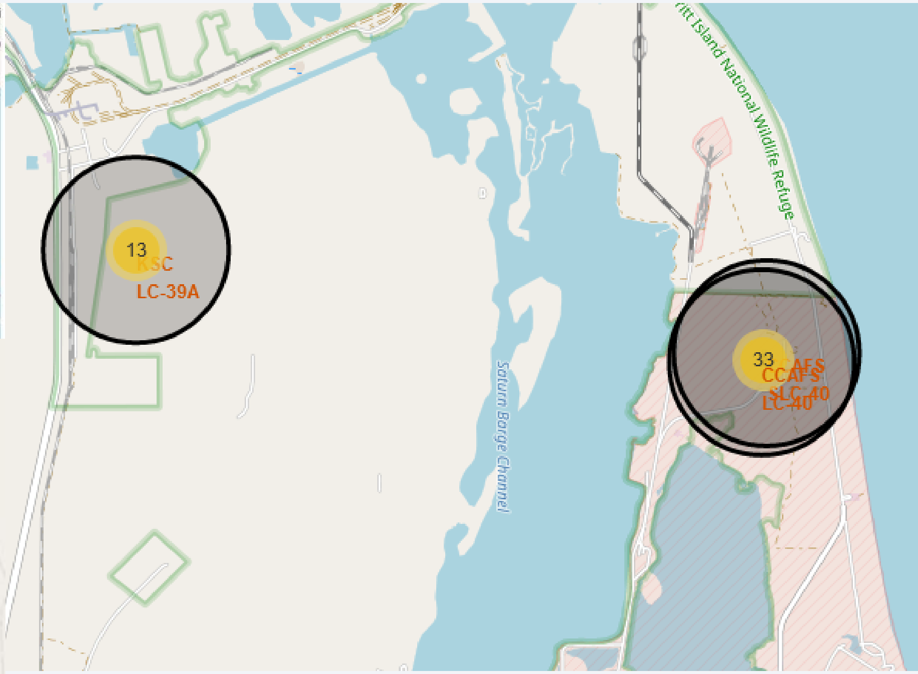
Section 3

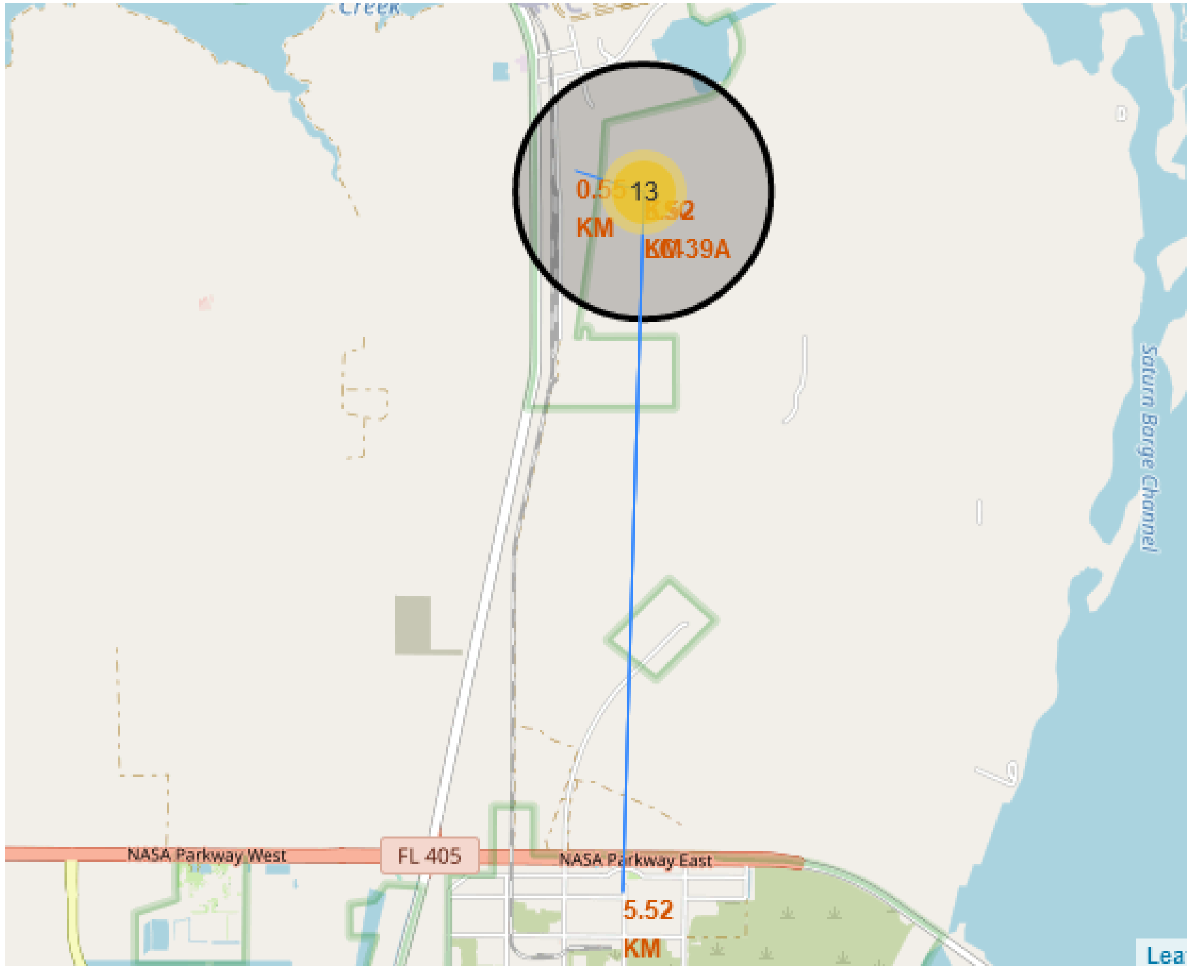
Launch Sites Proximities Analysis

Launch Outcomes by Site

- This is an example of KSC LC-39A launch site outcomes, where green markers indicate successful and red one indicate the failures.





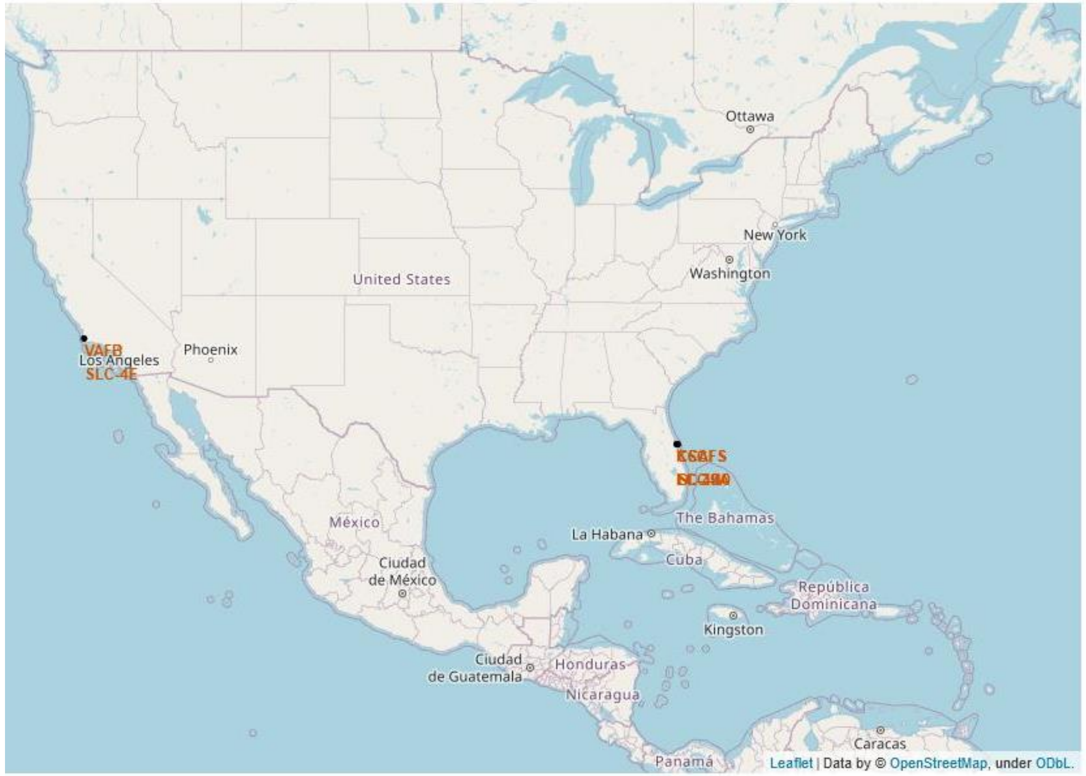


Logistics and Safety map

- Here we can see the launch site KSC LC-39A, which has good logistics aspects, being near railroad, and relatively far from inhabited areas.

Full map launch site

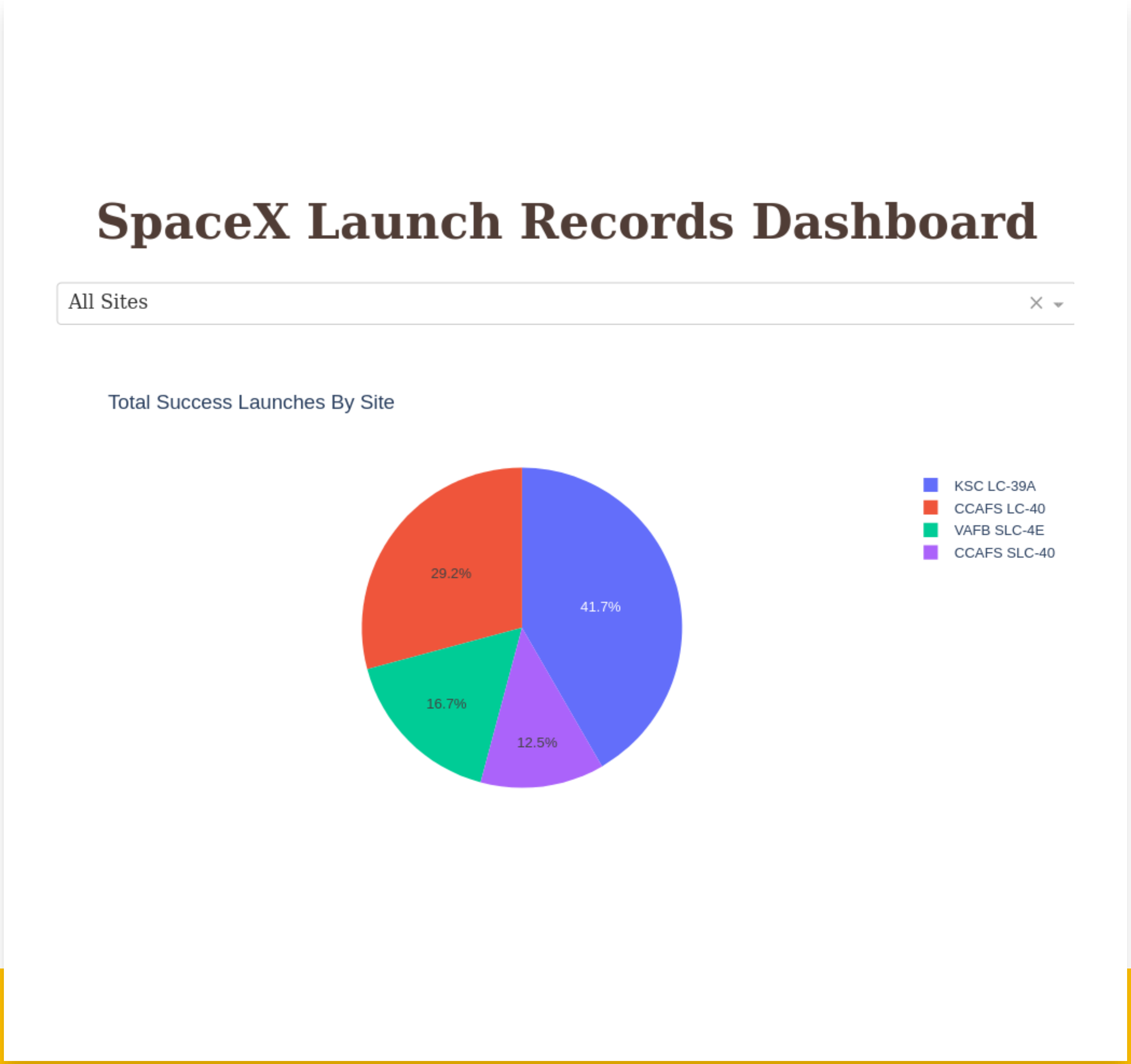
- As we can see here, the launch sites are mainly focused near the sea, but not too far from roads and railroads, this may be because of safety.



Section 4

Build a Dashboard with Plotly Dash

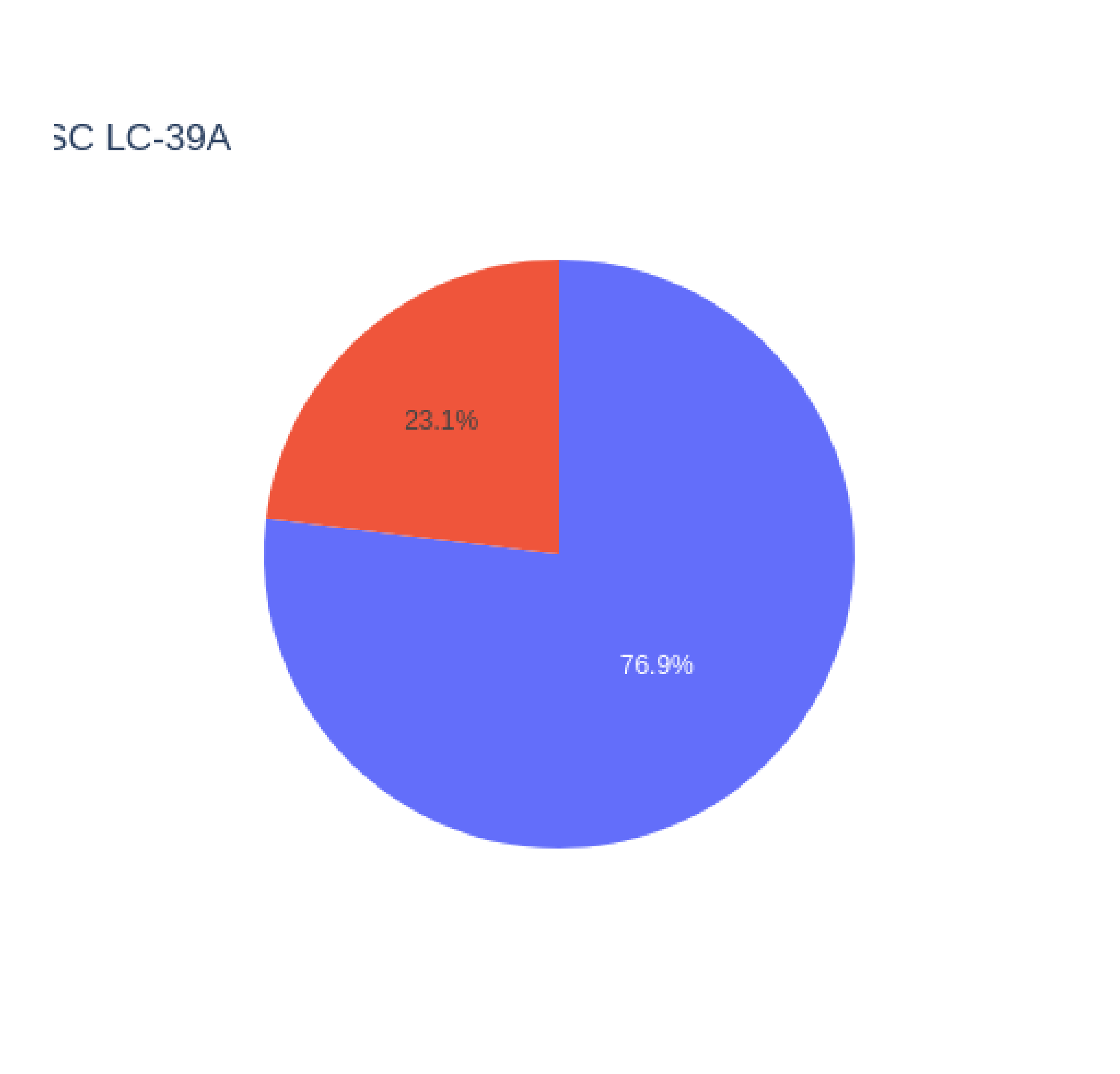
Successful launching sites



- This dashboard shows us the places where launches are done, as a result of the information analyzed, we can conclude that this is a substantial factor to the mission success.

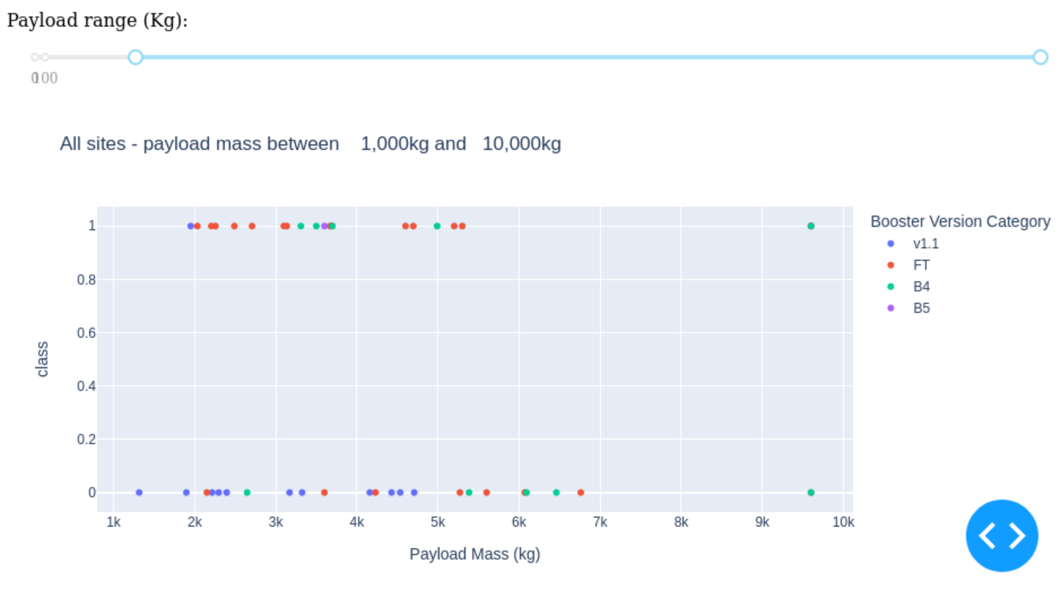
Highest launch success ratio

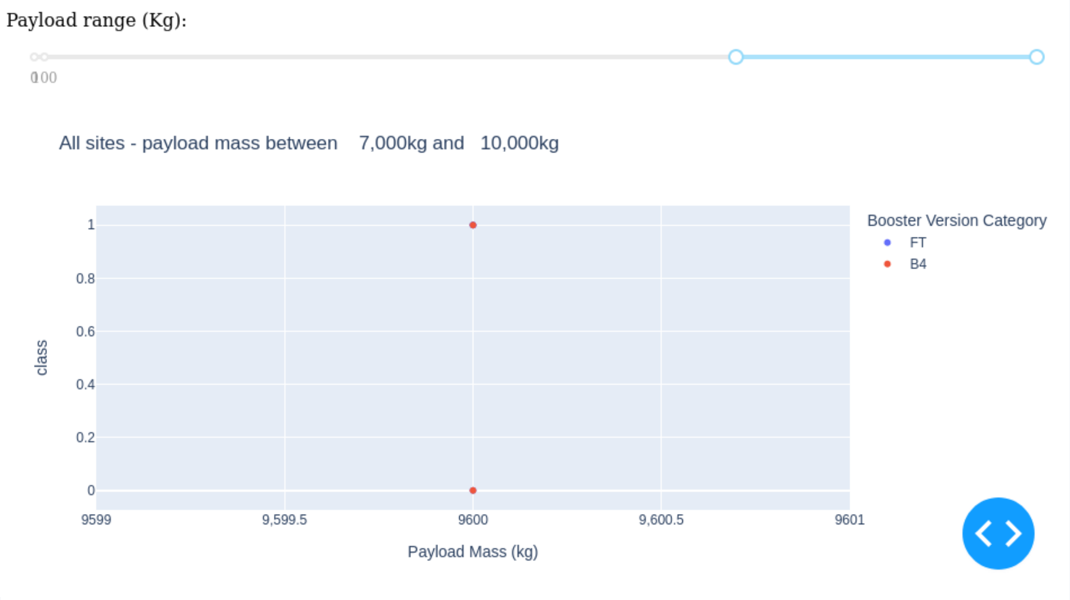
- Based on the data, we can say that 76.9% of launches are successful, as we can see in the following graph:



Payload vs Launch Outcome

- Based on the data, we can say that payloads under 6000 KG and FT boosters are the most successful combination, and we can't determine the risk and success of launches over 7000 KG, because there is no information to get a conclusion.



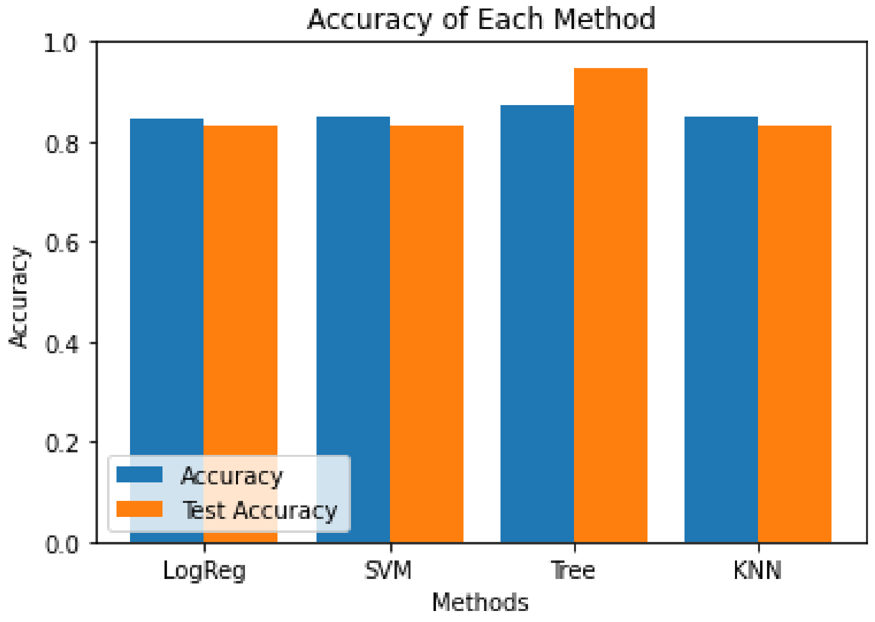


Section 5

Predictive Analysis (Classification)

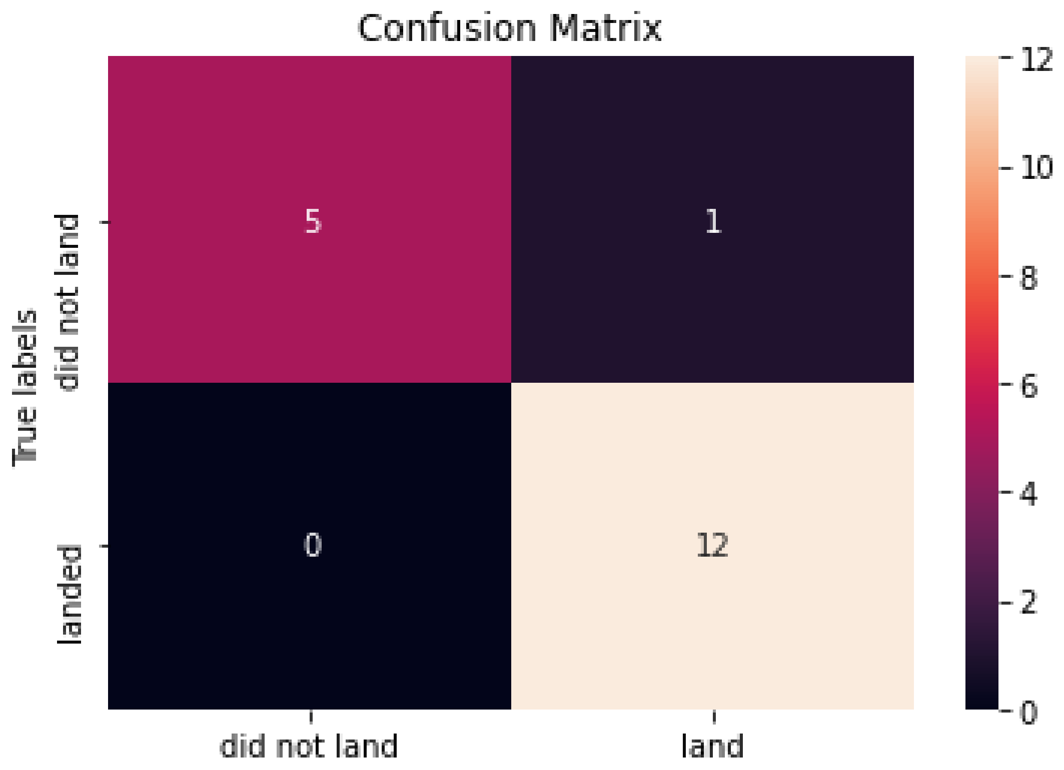
Classification Accuracy

- As we mentioned earlier, we tested four classification models, and the accuracy is shown in the following plot graph; based on that, we can conclude that the model with the highest classification accuracy is the decision tree classifier, which has accuracies over than 87%.



Confusion Matrix

- The confusion matrix of the decision tree classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones, this is shown in the following graph:



Conclusions

Based on all the process shown, we can conclude some important points, which were analyzed based on the different data sources, such as:

- The best launch site is KSC LC-39A
- Launches above 7000 KG are less risky
- The decision tree classifier can be used to predict successful landings and increase profits, being fundamental to future landing experiments.
- Besides most of the mission outcomes are successful, we can say that the successful landings seem to improve over the time, which makes sense according to the evolution of rockets design and continuous iteration.

Thank you!