IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cristian Hernán Rueda Meza

04-Oct-2021

Outline

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

Executive Summary

- For this project we started gathering the needed data from 2 sources, first we used the SpaceX API, then we used web scrapping to get the data located in Wikipedia. Then we wrangle the data and did a exploratory data analysis, in this stage we found the variables we would use to predict the success of the missions, and used some graphics to understand the data. We finally made a classification model.
- We found there were some variables that affect the success of a mission and we created a classification model upon this information to find whether a mission would be successful or not.

Introduction

- We are interest on entering in the industry of rockets and spacecraft, so we wanted to know the main competitors. In this project we wanted to analyze the falcon 9 rocket launches, and how SpaceX could keep the cost of the missions lower compare to others providers of this service. We could see SpaceX reuse the first stage of their rockets and keep low cost with this strategy.
- Due to the possibility that the first stage of the rockets fail their landing when coming back, and the high cost that these fails represent, we wanted to find if there was any variable that affect the landing of the first stage of the rocket and minimize the risk of fail.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We used the SpaceX API, and web scrapping to get the data about the SpaceX launches located in Wikipedia.
- Perform data wrangling
  - We classified the data to get just the information about the Falcon 9 rockets, the decided the variables that were more import for us. Also we replace the missing data of the Payload Mass variable with the mean.
- 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 created 4 classification models with the wrangled data, used the function GridSearchCV to find the best parameters for the 4 models and used the best_score_ method to find the best classifier.

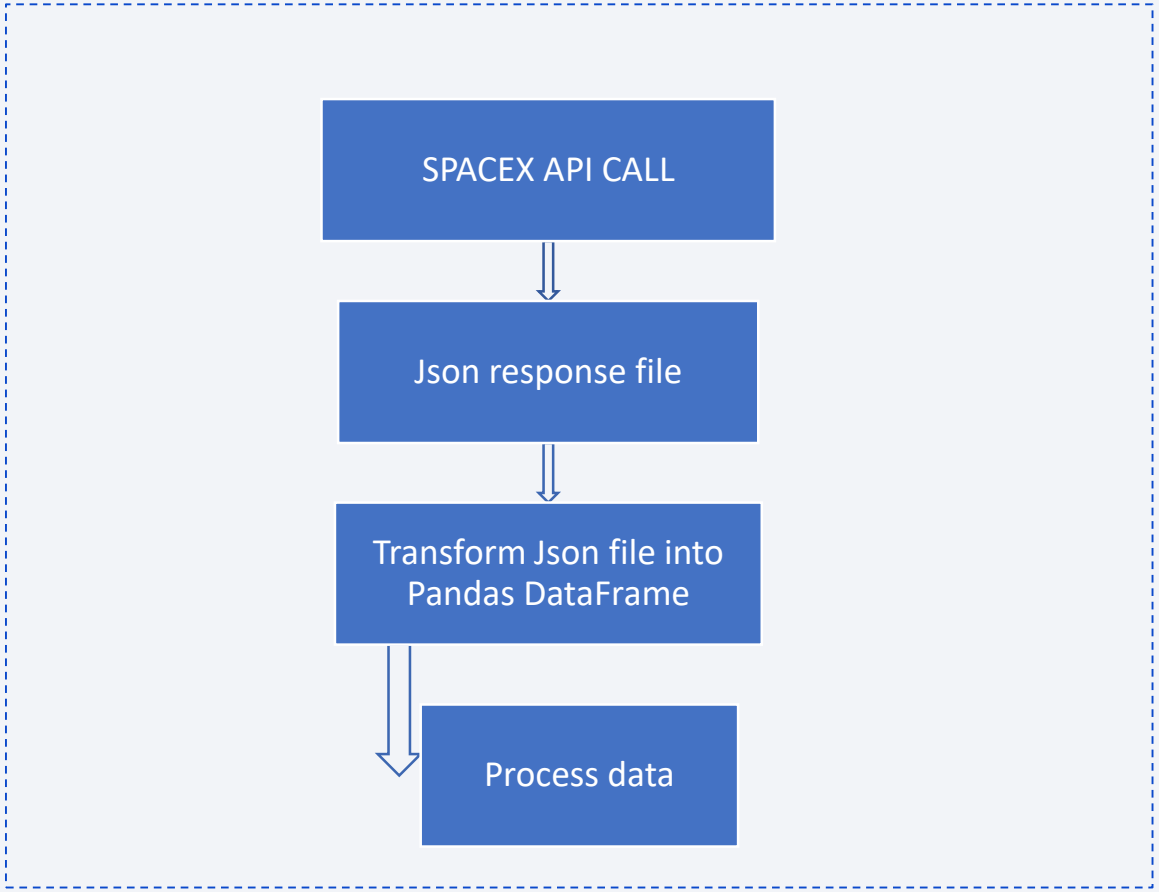
Data Collection

- For the data collection process we used SpaceX REST API, and web scrapping of information found in Wikipedia.

Data Collection - SpaceX API

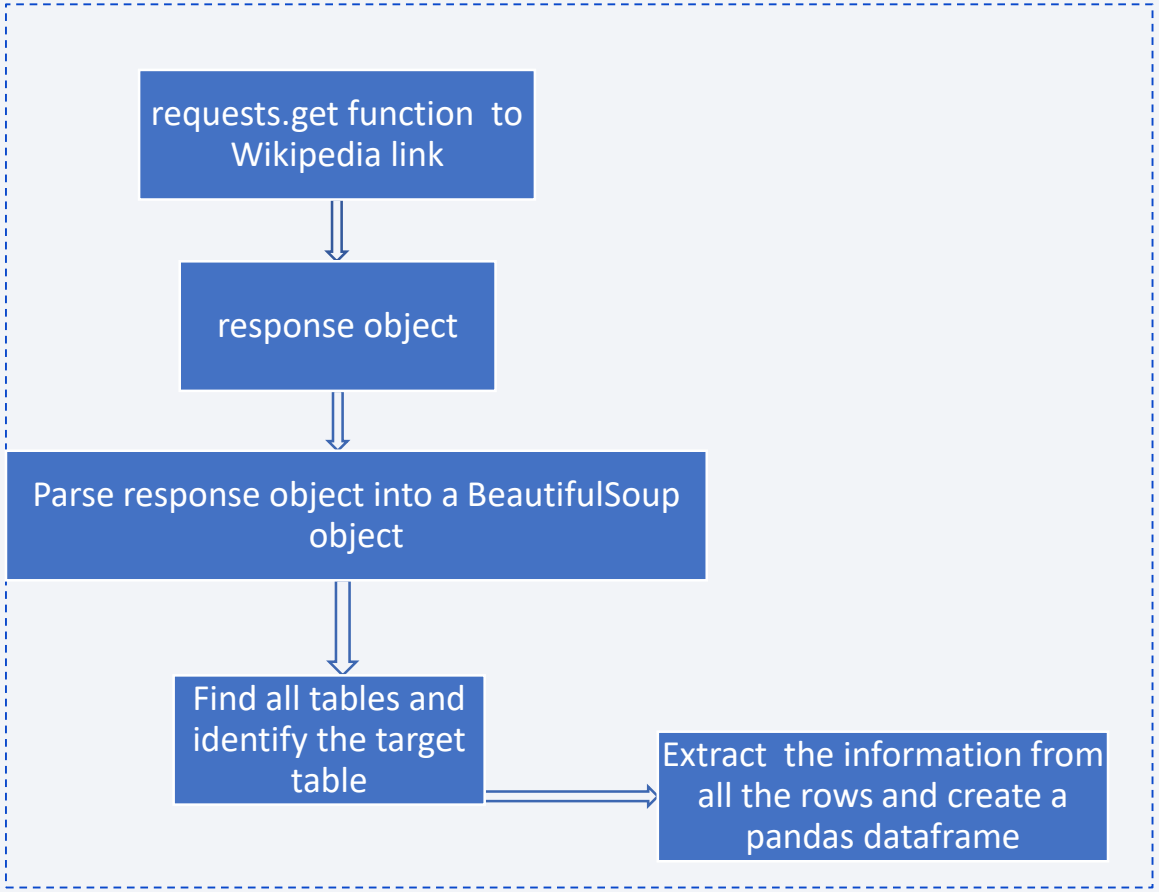
- We first made a SpaceX REST API call that returned a Json response file, we transformed the Json file into a pandas dataframe with the method json_normalize.

- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/Data%20Collection%20API.ipynb



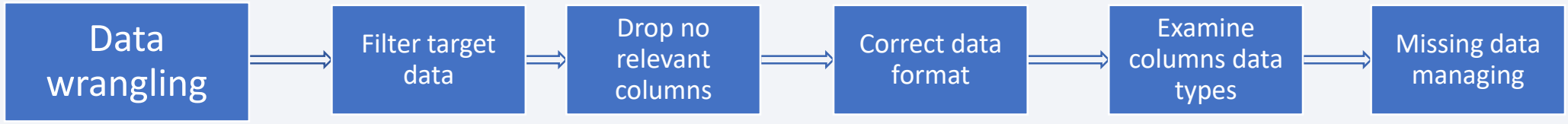
Data Collection - Scraping

- Present We used the request.get() method that returned a html file which we parse into text, then we created a beautifulsoup object, and with this object we find all the tables, identified our target table, then extract the information from all the rows and with this information created a pandas dataframe.
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We first filtered the dataframe to use just the falcon 9 data, decided the variables that were relevant, corrected the format of the data, examined the types of the variables, then dealed with the missing data, and replace the Payload mass missing data with the mean.
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/EDA.ipynb



EDA with Data Visualization

We made 7 plots to find correlations between the variables.

- Scatterplot of the FlightNumber vs. PayloadMass
- Catplot to plot FlightNumber vs LaunchSite
- Scatterplot of launch sites vs payload mass
- Bar chart for the sucess rate of each orbit
- Scatterplot of FlightNumber vs Orbit type
- Scatterplot of Payload vs. Orbit
- Lineplot of yearly average success rate
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

We made several SQL queries

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA‘
- The 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
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- https://github.com/ChruedaM/Coursera_Capstone/blob/e77453439439f9af064bc63741951672bb6be305/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

We added several objects to the interactive maps.

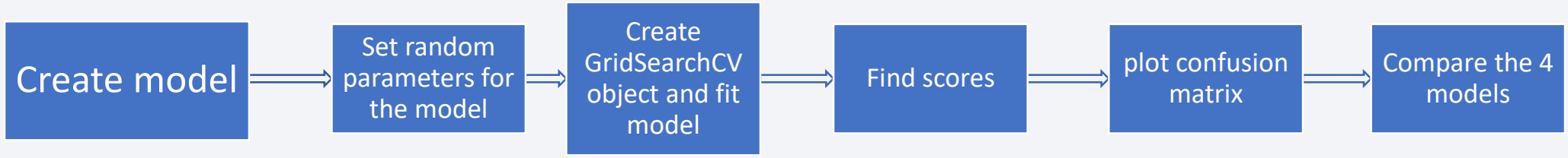
- Circles, to mark the Launch sites
- We created a marker cluster, where we gathered markers for each landing, coloring green the successful landings, and red the unsuccessful ones
- Markers of text, to signalize the distance to different locations like highways, railways and coast lines, then joined the distances with a line.
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- We added a pie plot, where we could choose to show the information for the 4 launch sites, or for each one using a drop down, the graph depict the success rate for the chosen launch site.
- We added a scatter plot which shows the relation between the payload mass and the success of the mission, using a dropdown we could choose to show the information for the 4 launch sites, or for each one, and with a slider we can filter the range of payload mass to show.
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/spacex_dash_app.py

Predictive Analysis (Classification)

- We created 4 classification models (logistic regression, support vector machine, decision tree classifier, k nearest neighbors), for each model we created aleatory parameters, then created a GridSearchCV object and fit the models to optimize them and find the best parameters for each one with the best scores, than compare the 4 models to find the model with the best performance. Also we plot the confusion matrix for each model.
- https://github.com/ChruedaM/Coursera_Capstone/blob/d24639cd7a412bcd6a23cd0666ac878a4d24d191/Machine%20Learning%20Prediction.ipynb

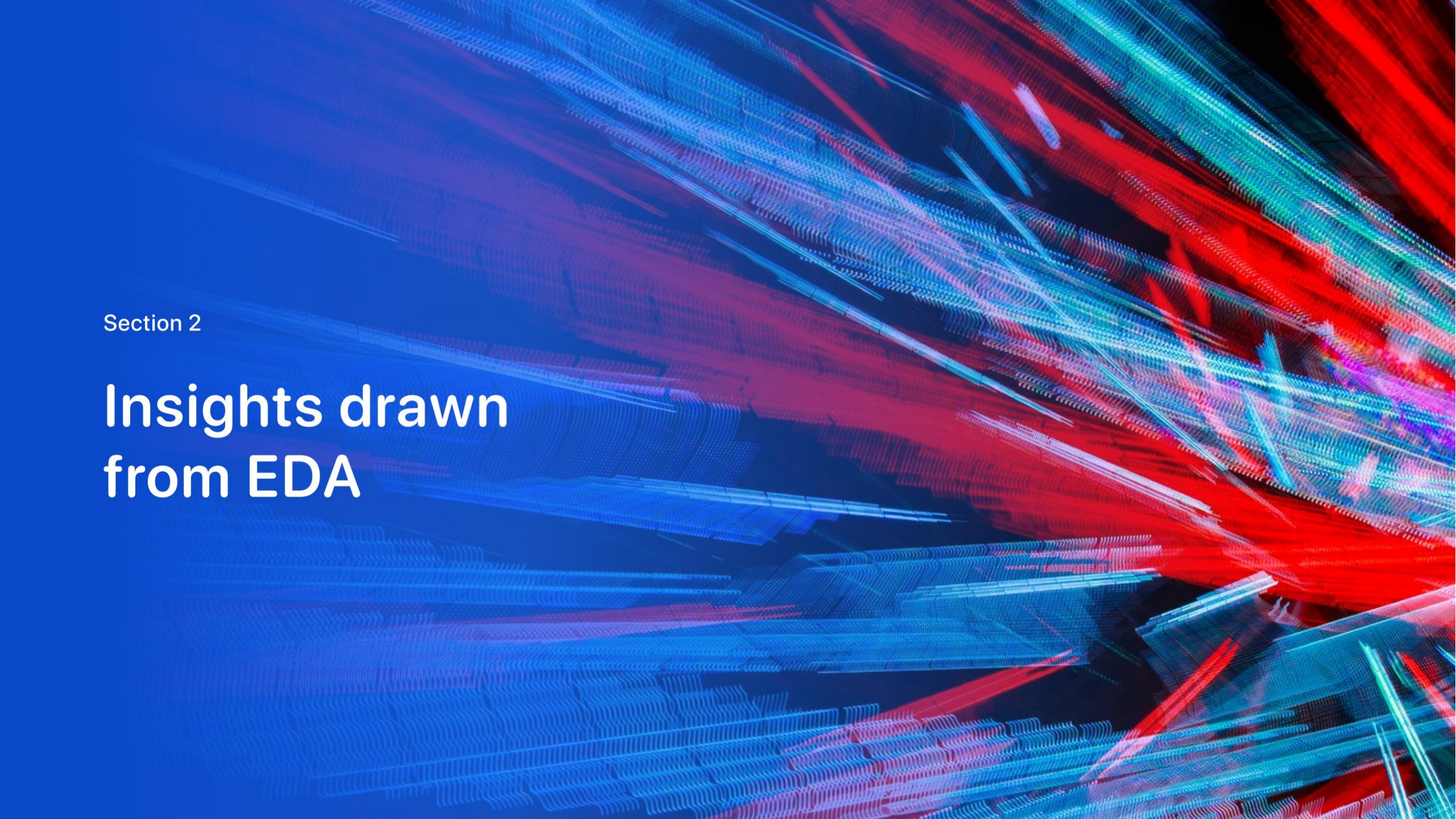


Results

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

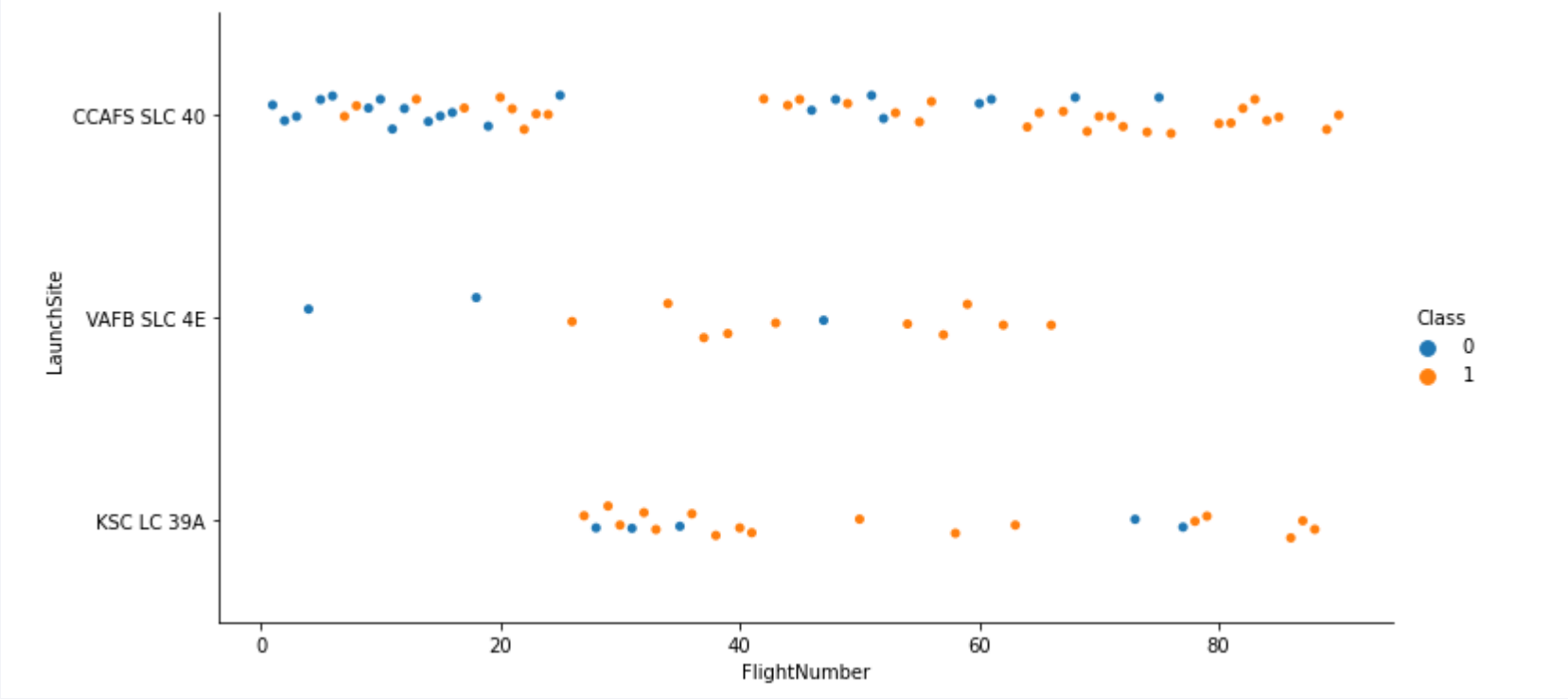
Section 2

Insights drawn from EDA



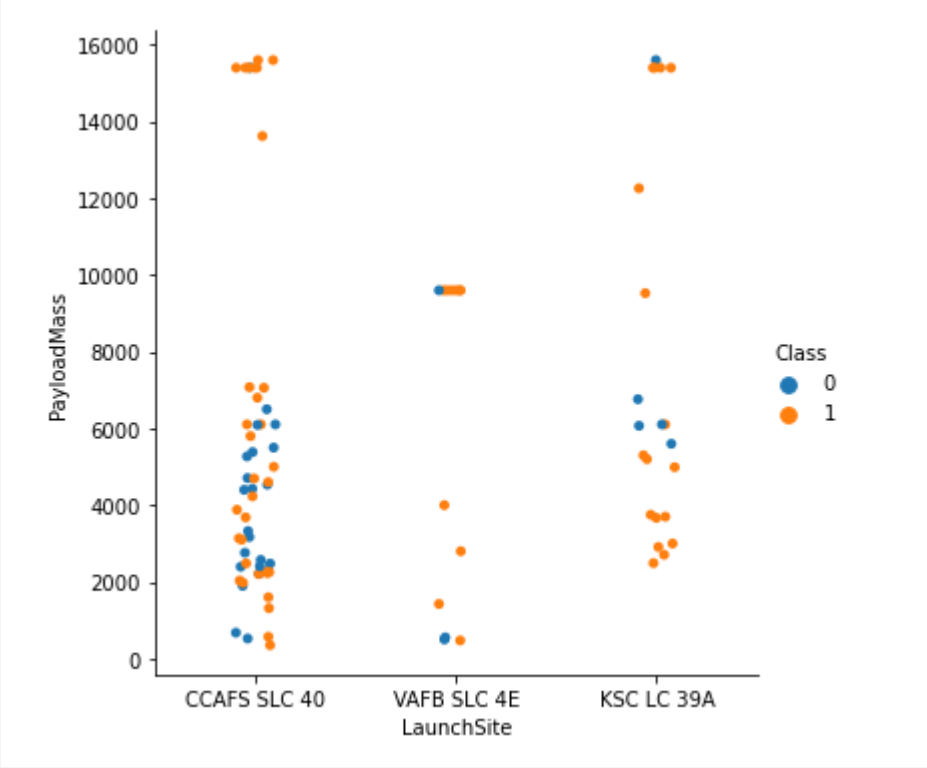
Flight Number vs. Launch Site

- In the scatter plot of Flight Number vs. Launch Site we can see that the CCAFS SLC 40 is the most common launch site, around the 25th and to the 42nd most flights launched from KSC LC-39A, after that CCAFS SLC 40 went back as most common launch site. When the number of launches has augmented the failures for CCAFS SLC 40 has diminish.



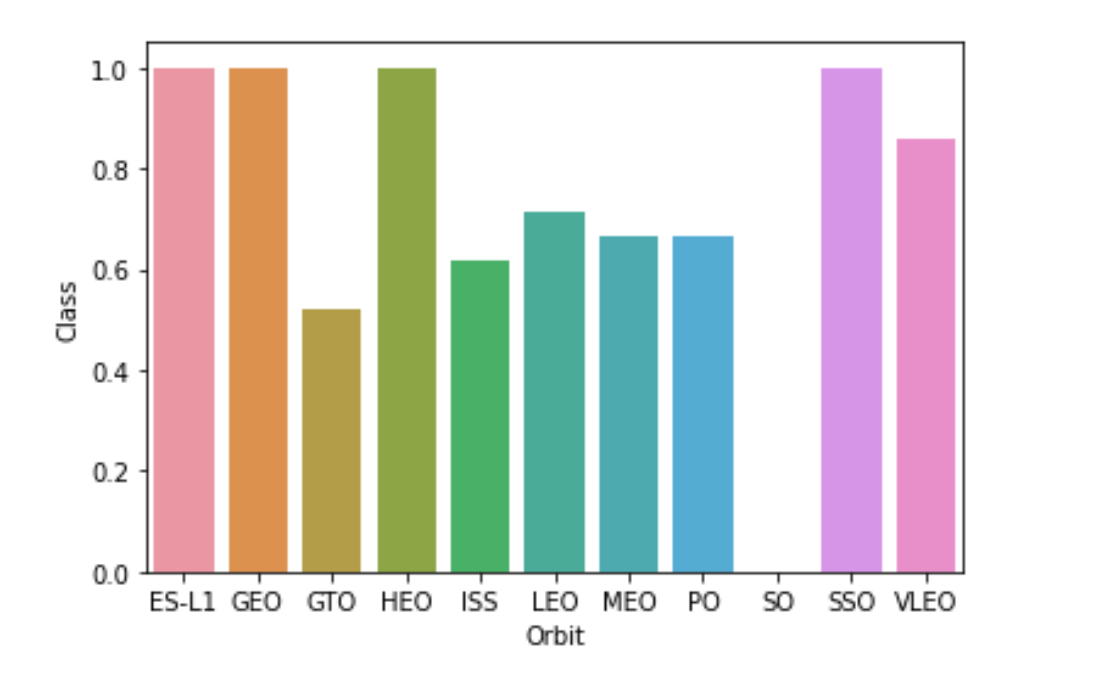
Payload vs. Launch Site

- The scatter plot of Payload vs. Launch Site shows that all the highest payloads are launch from CCAFS SLC 40 and KSC LC-39A, there is a midpoint, 10.000kg where almost all flights launched from VAFB SLC-4E, and the payloads lower than 8.000 are launched from the 3 Launch Sites with more frequency at CCAFS SLC 40.
- Most of the failures have happened in the CCAFS SLC 40, but with high payload mass the number of failures diminish



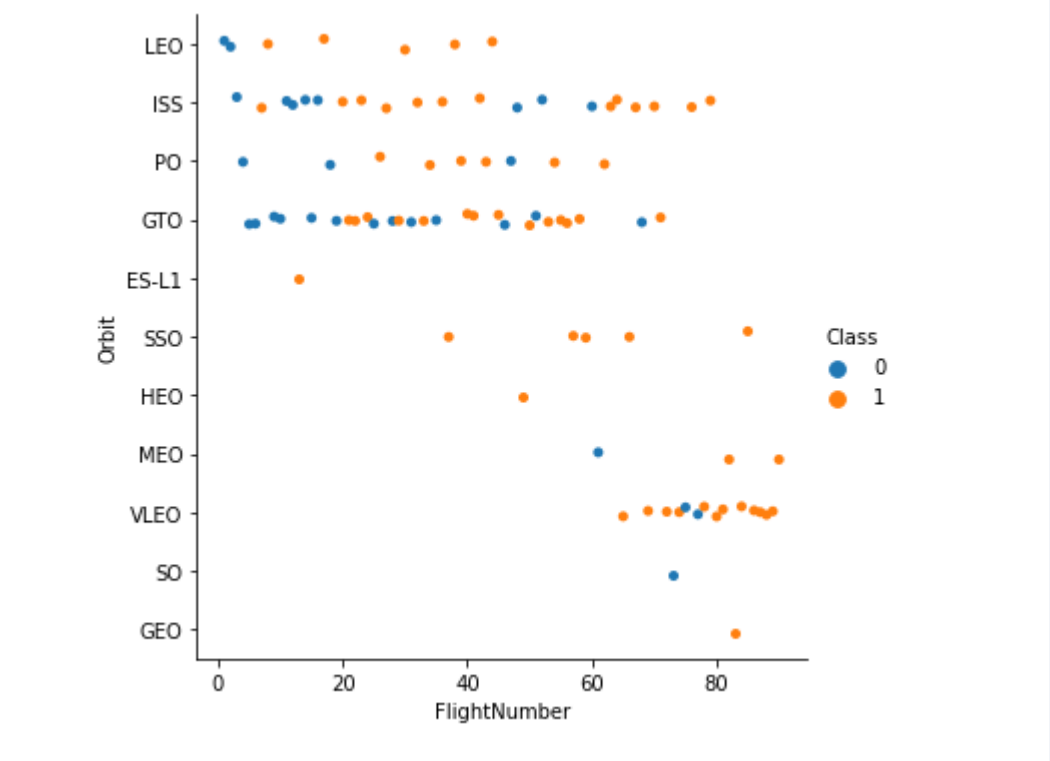
Success Rate vs. Orbit Type

- In the bar chart for the success rate of each orbit type we can see that ES-LI, GEO, HEO and SSO have a success rate of 100%, the SO orbit have a success rate of 0%, the others orbits have a success rate between 50% and 90%



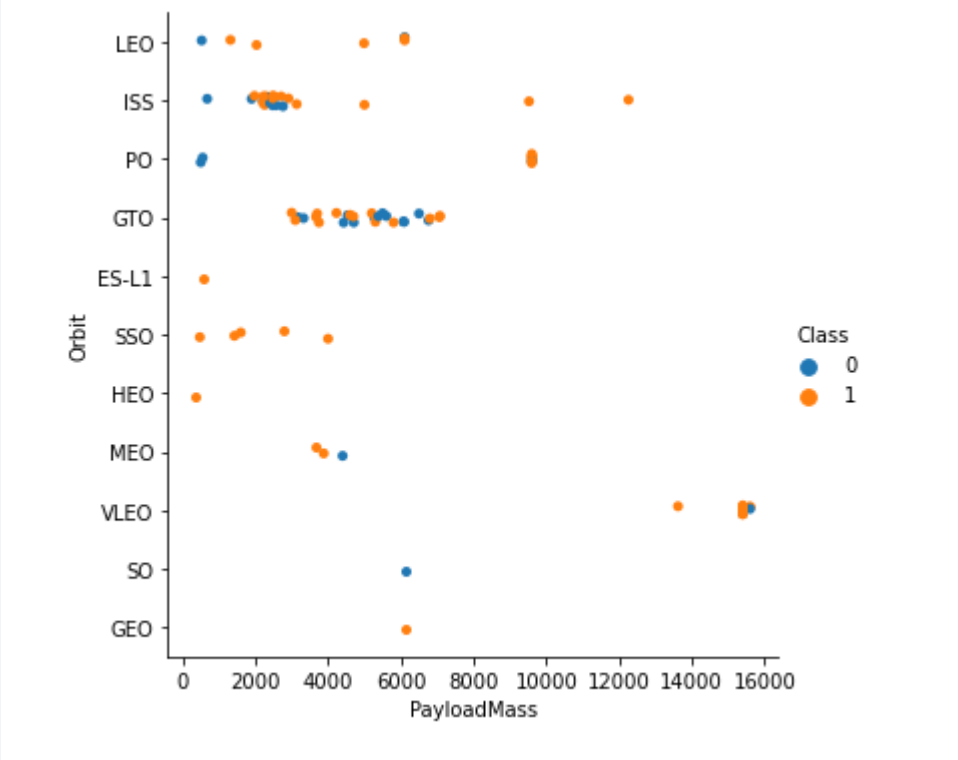
Flight Number vs. Orbit Type

- The scatter plot of Flight number vs. Orbit type we see that in the LEO orbit the 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.



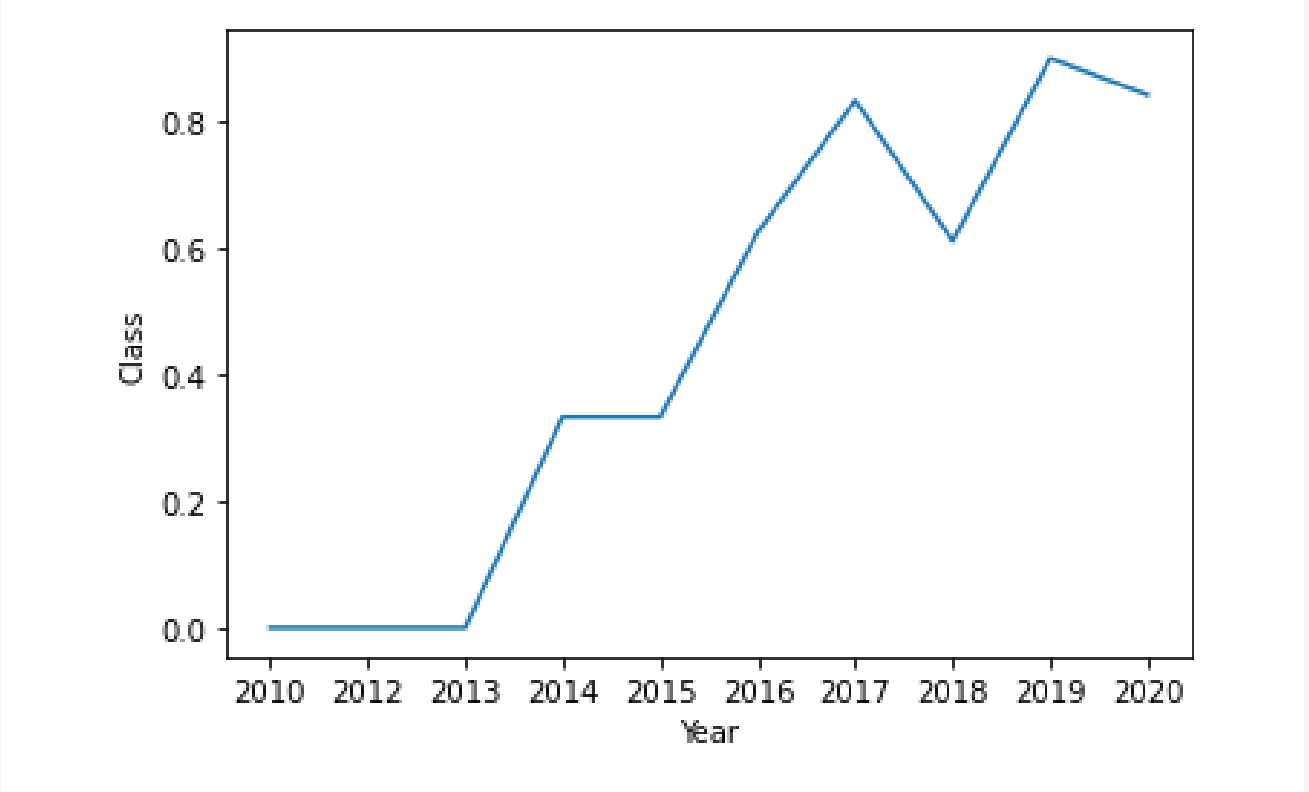
Payload vs. Orbit Type

- In the scatter plot of payload vs. orbit type we observe that Heavy payloads have a negative influence on GTO orbits and positive on Polar, LEO and (ISS) orbits.



Launch Success Yearly Trend

- In the line chart of yearly average success rate we can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- We found the names of the unique launch sites

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

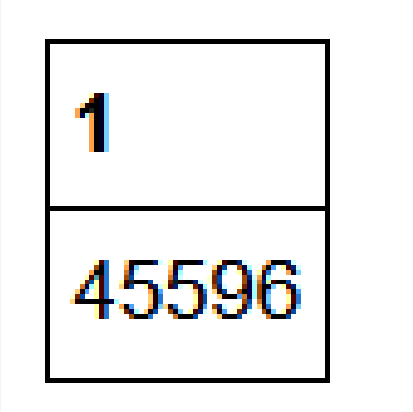
Launch Site Names Begin with 'CCA'

- We performed a query to find 5 records where launch sites begin with `CCA`

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	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

Total Payload Mass

- We calculated the total payload carried by boosters from NASA, with a total of 45596 Kg.



Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1, with an average payload mass of 2534Kg

1	
2534	

First Successful Ground Landing Date

- We found the dates of the first successful landing outcome on ground pad, the December 22nd of 2015.

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- We calculated the total number of successful and failure mission outcomes with 99 successes, 1 failure and 1 success with payload status unclear.

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- We Listed the names of the booster which have carried the maximum payload mass, a total of 12.

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

- We listed the failed landing_outcomes in drone ship, their booster versions, and launch site names 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

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

- We ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, starting in 16-03-2017 and finishing in 8-12-2010

DATE	landing__outcome
16/03/2017	No attempt
19/02/2017	Success (ground pad)
14/01/2017	Success (drone ship)
14/08/2016	Success (drone ship)
18/07/2016	Success (ground pad)
15/06/2016	Failure (drone ship)
27/05/2016	Success (drone ship)
6/05/2016	Success (drone ship)
8/04/2016	Success (drone ship)
4/03/2016	Failure (drone ship)
17/01/2016	Failure (drone ship)
22/12/2015	Success (ground pad)
28/06/2015	Precluded (drone ship)
27/04/2015	No attempt
14/04/2015	Failure (drone ship)

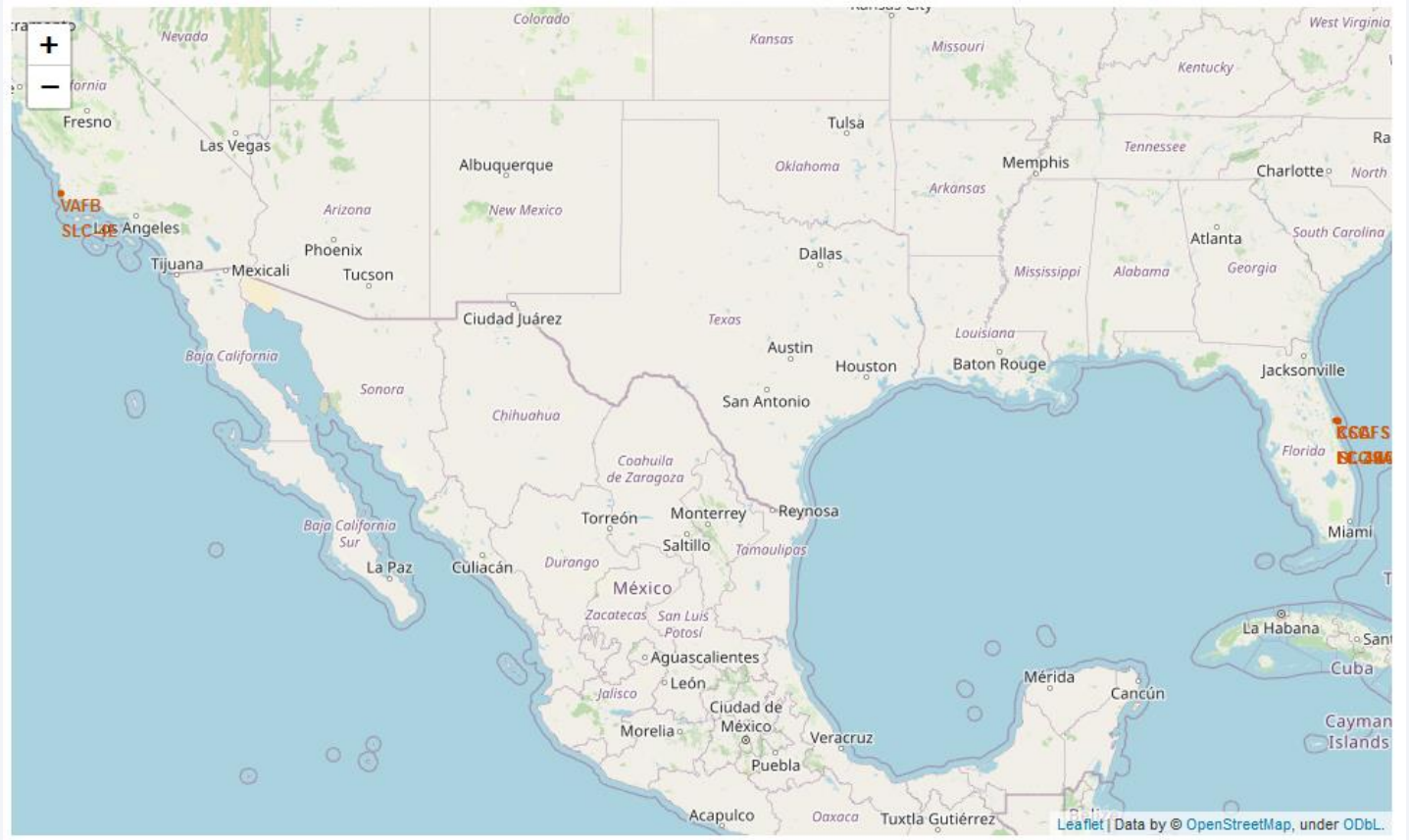
DATE	landing__outcome
2/03/2015	No attempt
11/02/2015	Controlled (ocean)
10/01/2015	Failure (drone ship)
21/09/2014	Uncontrolled (ocean)
7/09/2014	No attempt
5/08/2014	No attempt
14/07/2014	Controlled (ocean)
18/04/2014	Controlled (ocean)
6/01/2014	No attempt
3/12/2013	No attempt
29/09/2013	Uncontrolled (ocean)
1/03/2013	No attempt
8/10/2012	No attempt
22/05/2012	No attempt
8/12/2010	Failure (parachute)

Section 4

Launch Sites Proximities Analysis

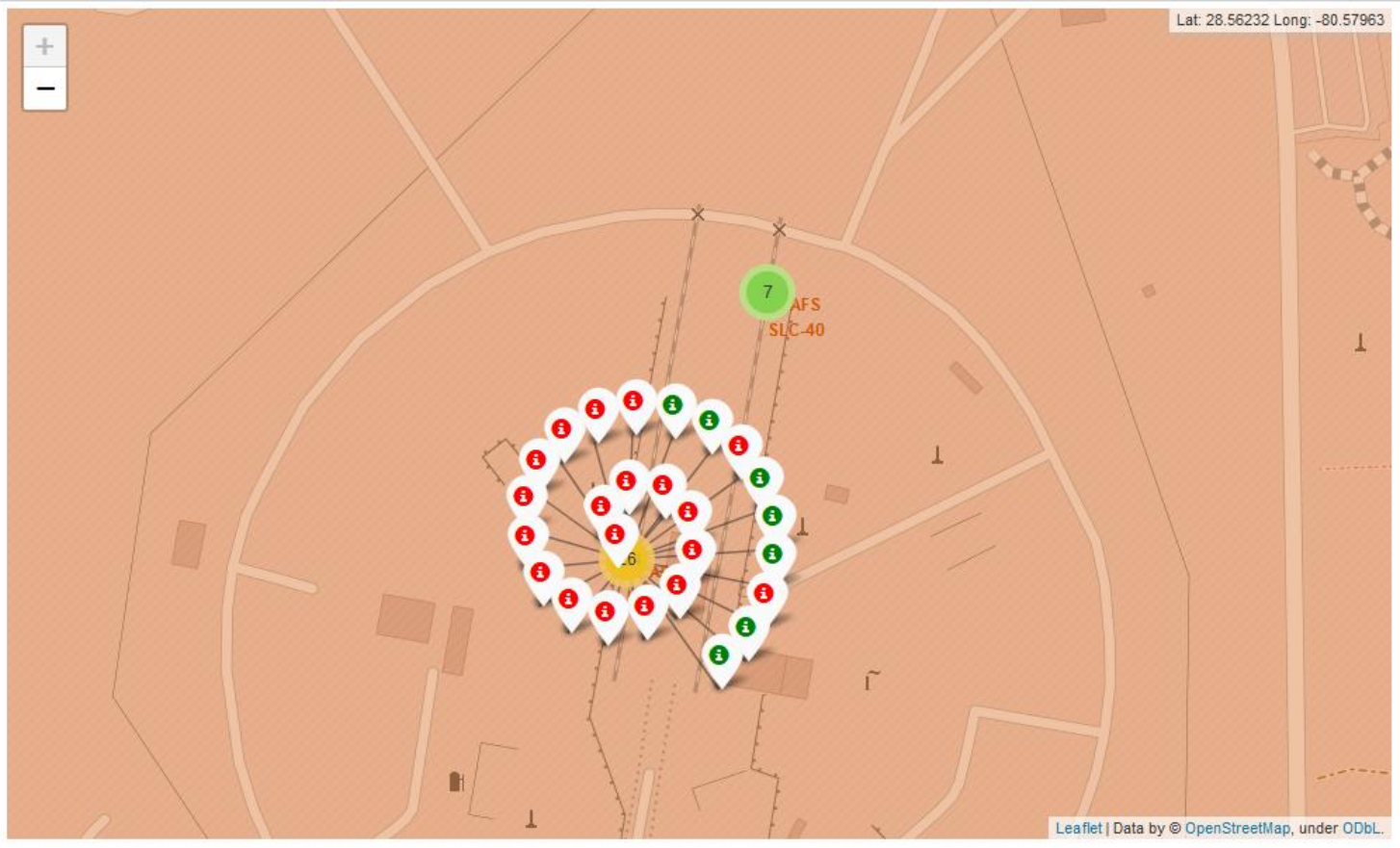
Map of launch sites

- We made a map that allows us to see where the launch sites are sited, all of them close to the coast line.



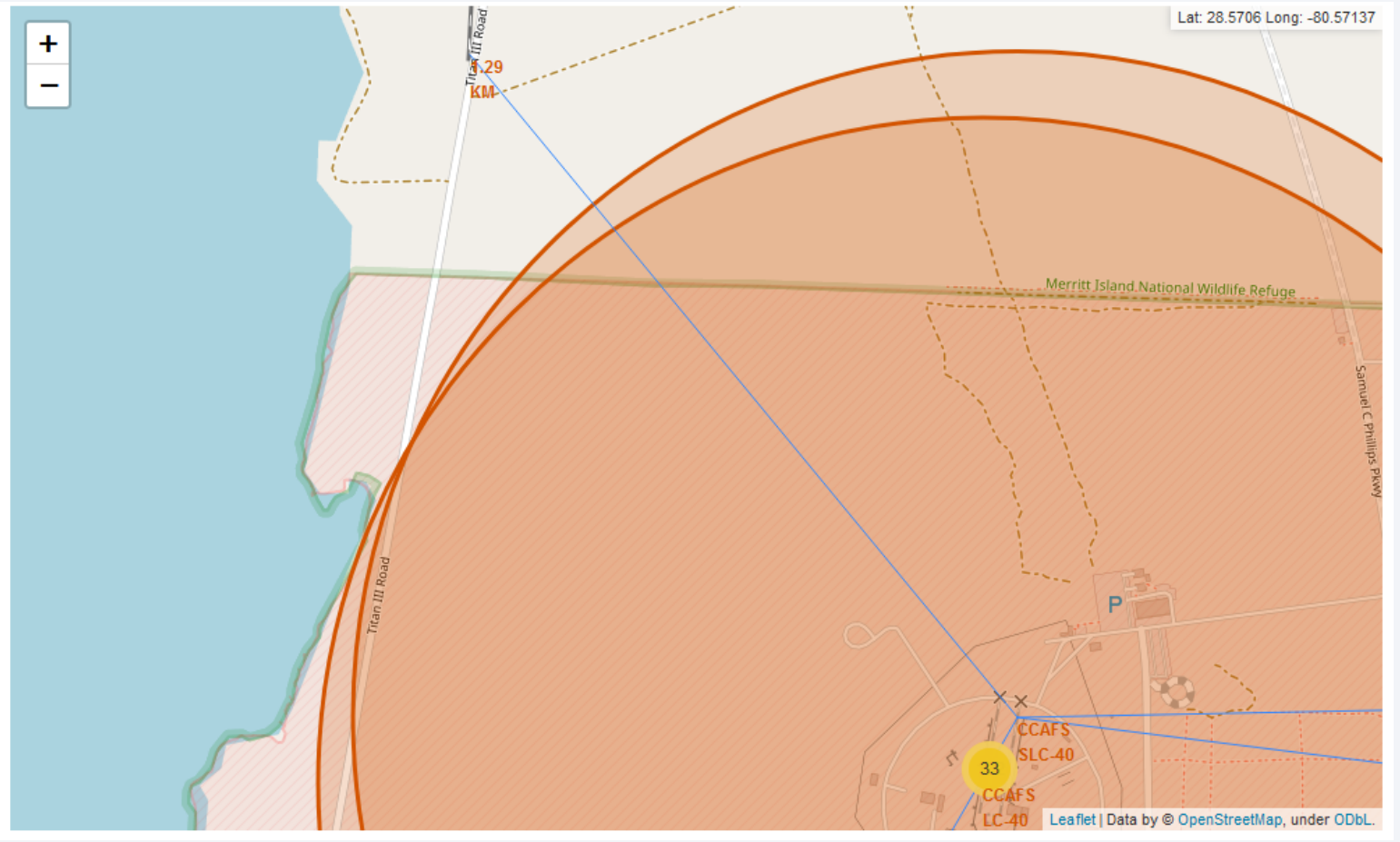
Color-labeled launch outcomes

- We created a cluster of markers, where we depicted the successful launch outcomes as green markers and the unsuccessful ones as red markers, for each launch site.



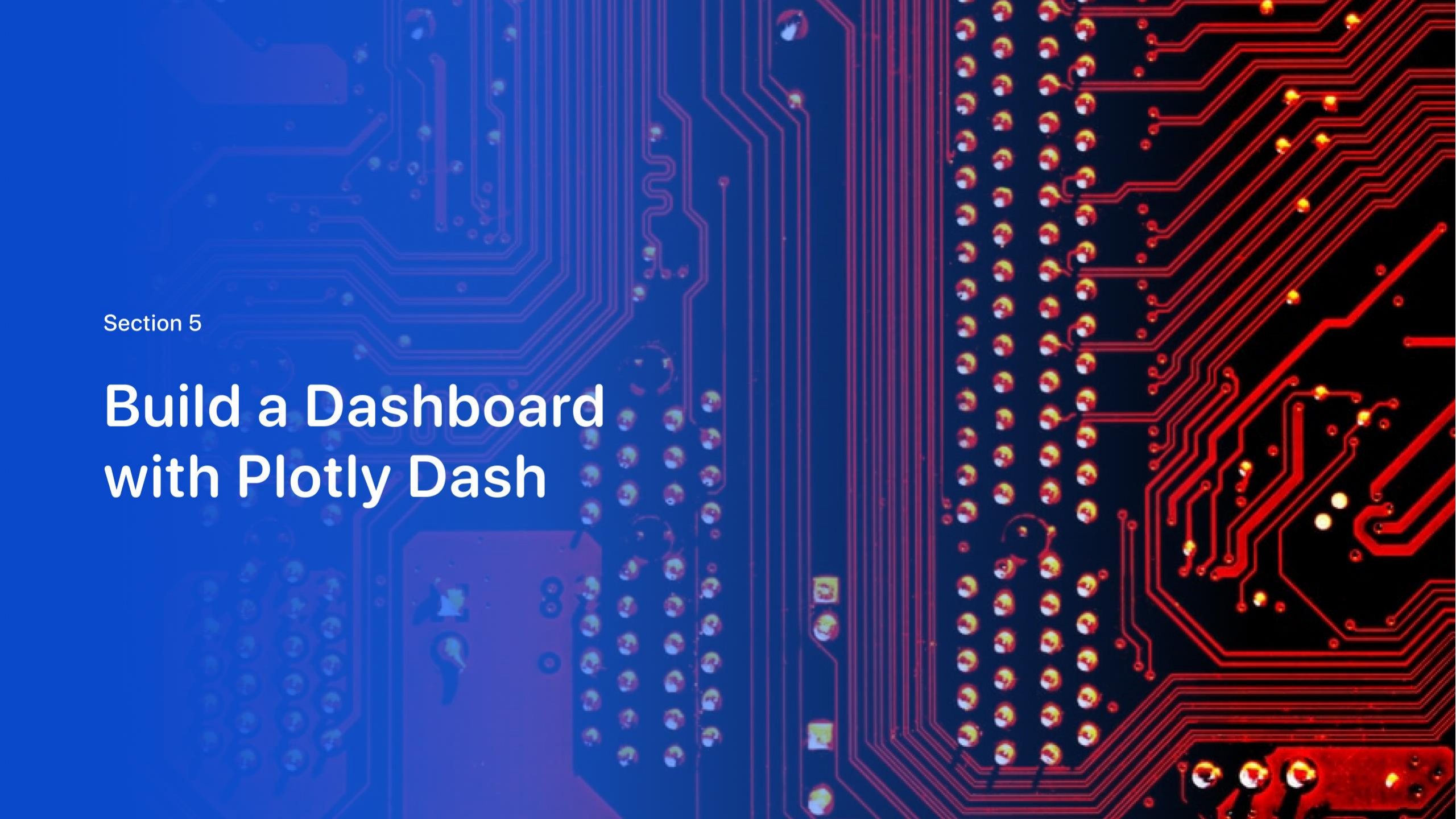
Proximities to launch sites

- We located different places like highways, railways and coast lines, found the distance to the launch sites and draw a line that shows the distance between these places.



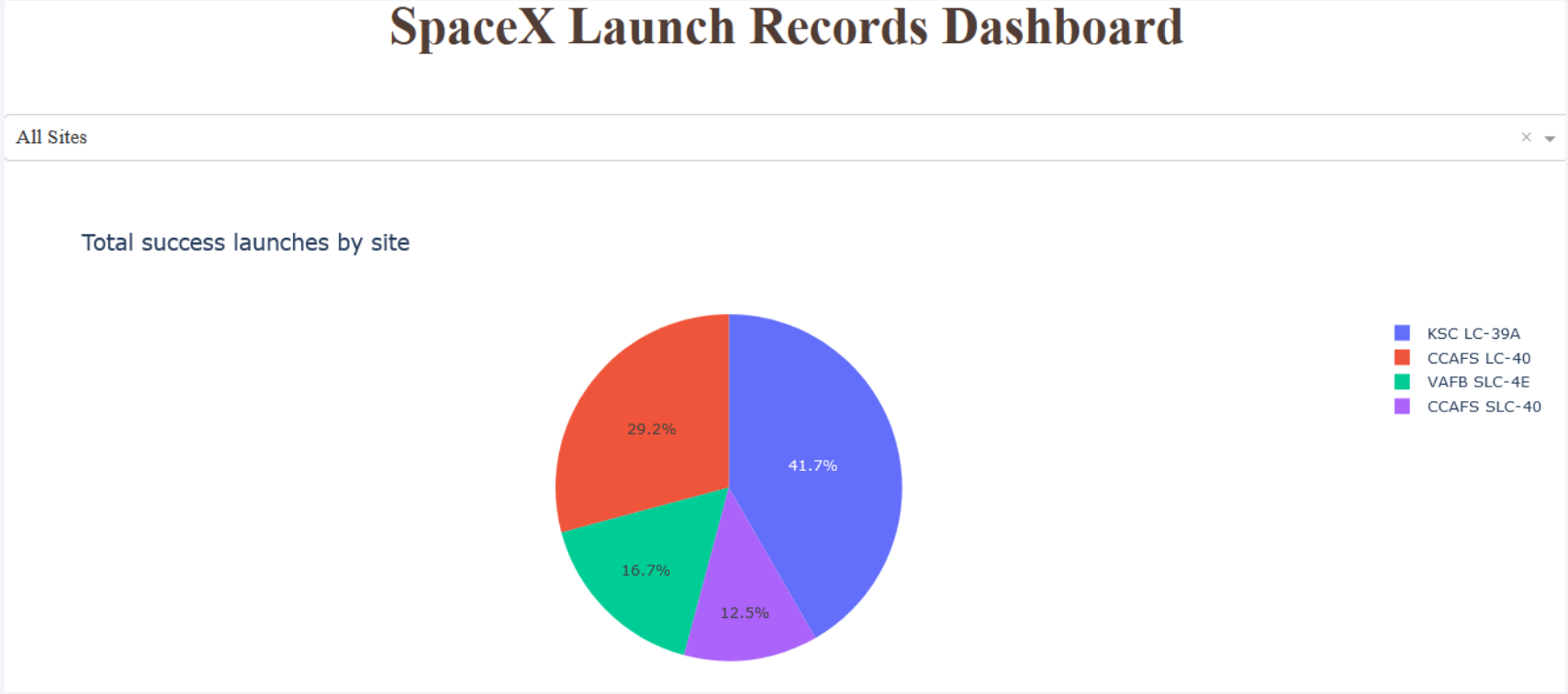
Section 5

Build a Dashboard with Plotly Dash



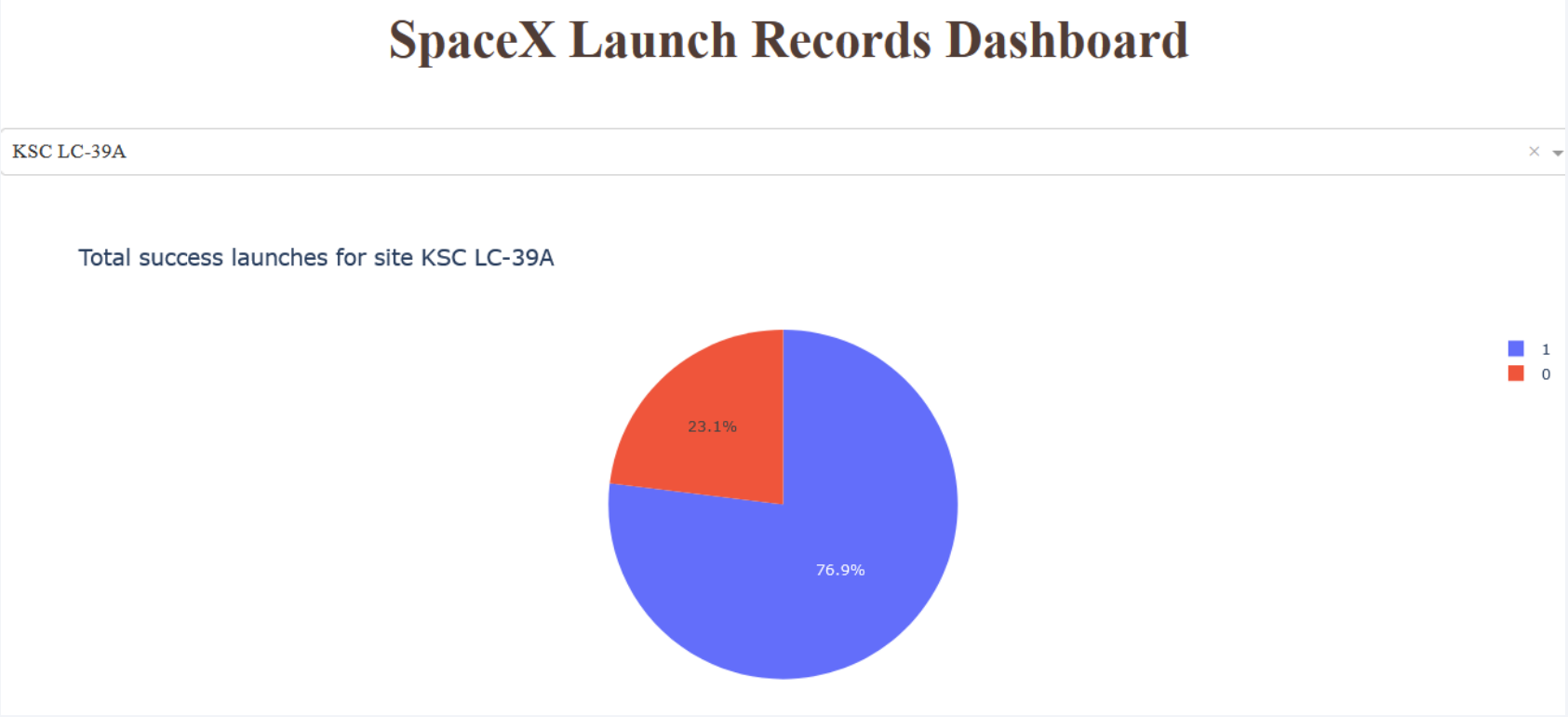
Piechart of launch success count for all sites

- We plotted a pie chart of launch success count for 4 all sites, where KSC LC-39A has the highest number of successful launches with a 41.7%



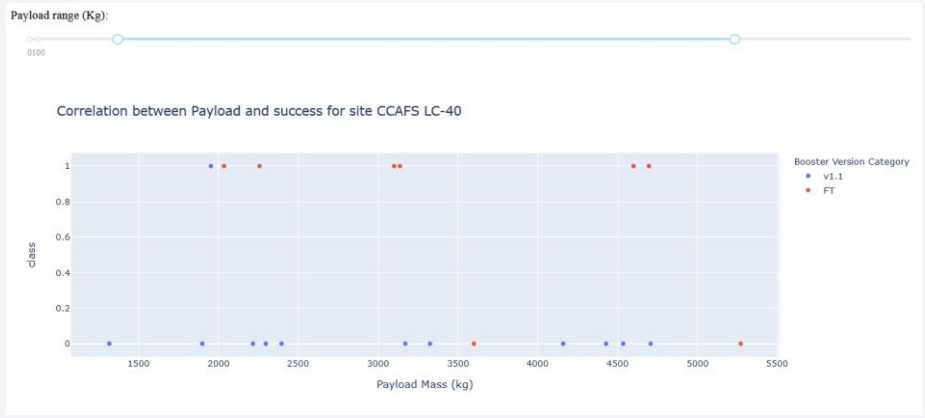
Piechart for the launch site with highest launch success ratio

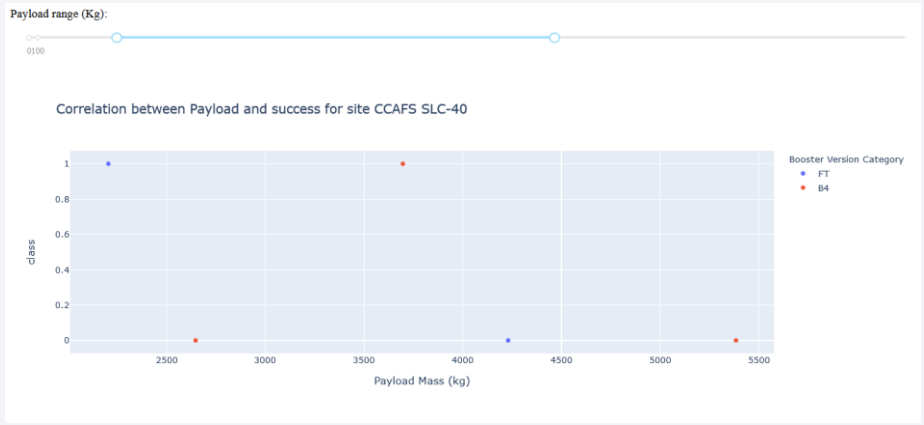
- We plotted a pie chart of launch success ratio for KSC LC-39A, with a launch success ratio of 76.9%

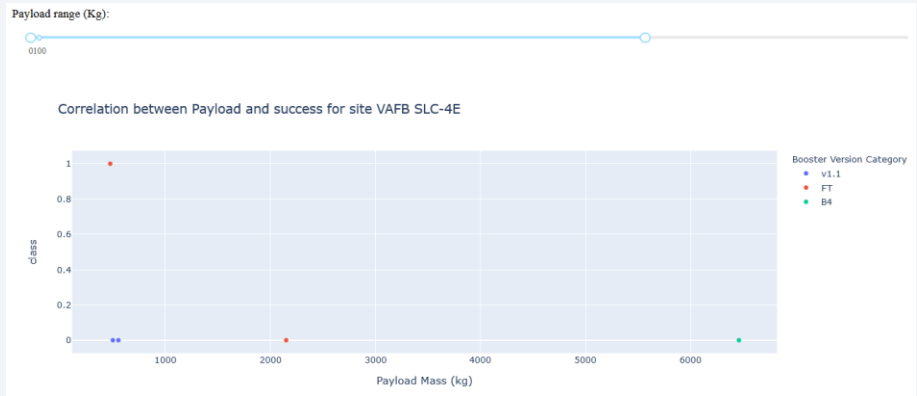


Payload vs. Launch Outcome scatter plot

- We made a plot for each launch site, with different payload mass ranges, showing how the payload could affect the success of the rocket launches, and how differs depending on the launch site.







Section 6

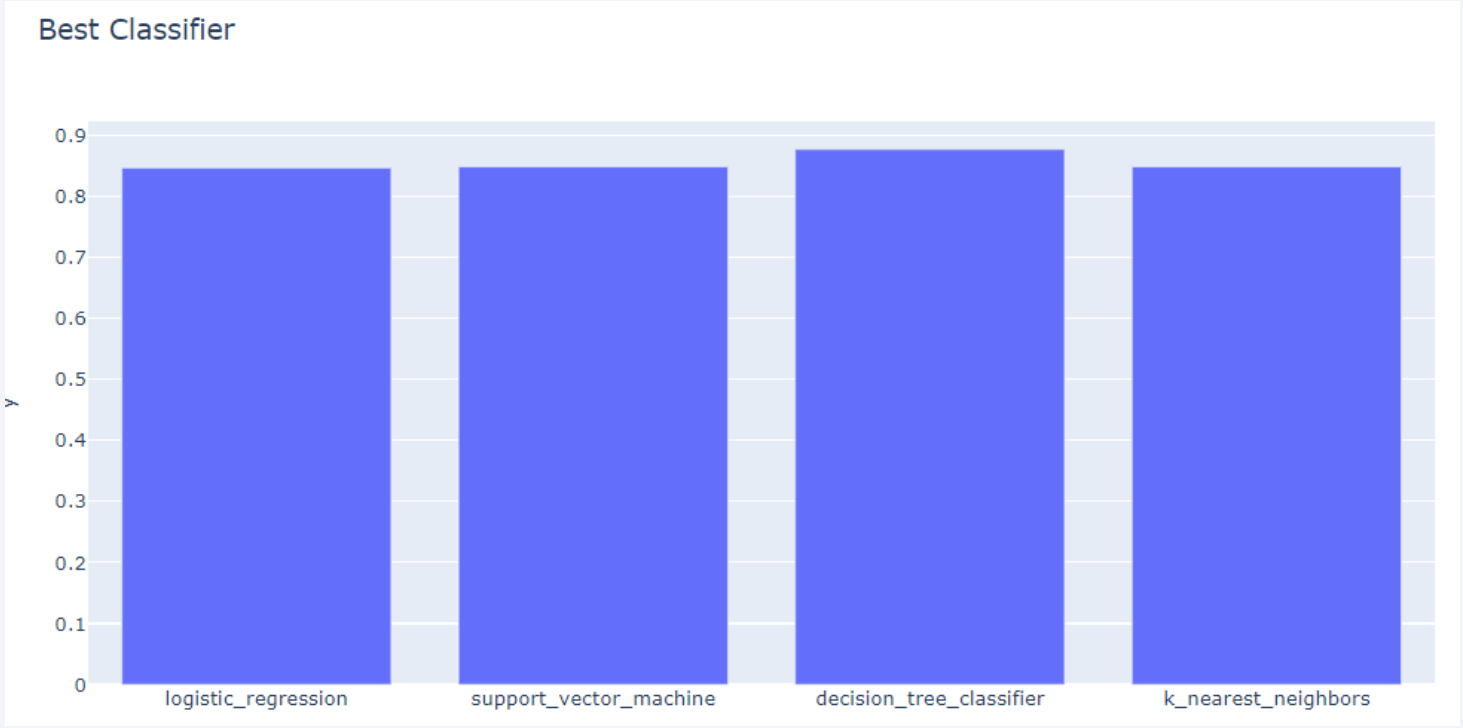
Predictive Analysis (Classification)

Classification Accuracy

- We can see the 4 models have a similar accuracy, being the decision tree classifier the one with the highest accuracy, with 87.67%

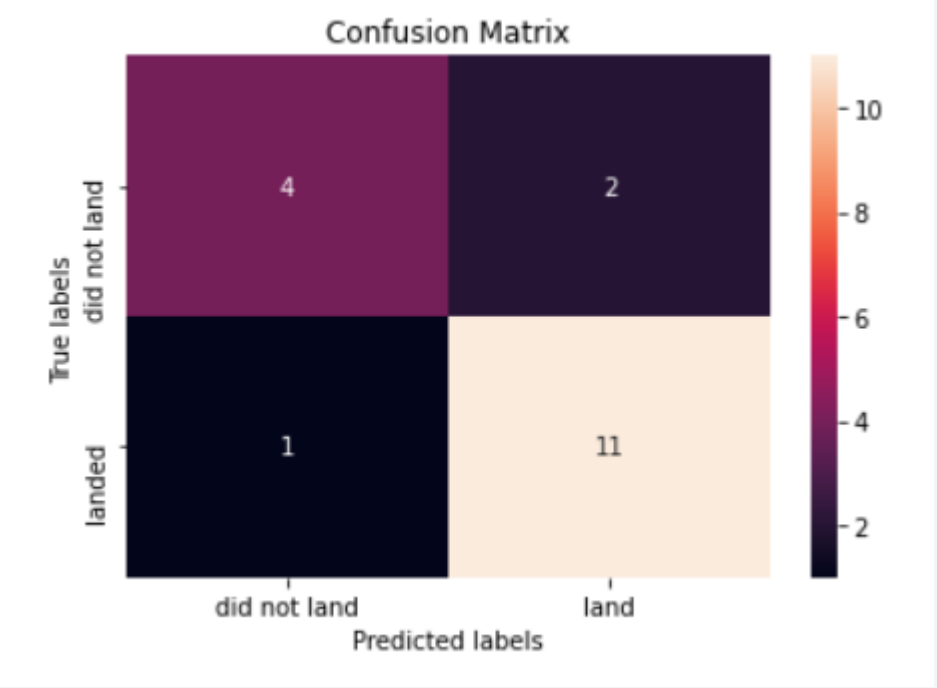
- The decision tree with the best performance has the next parameter :

- tuned hyperparameters :(best parameters) {'criterion': 'gini', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}



Confusion Matrix

- We can see the confusion matrix of the decision tree classifier. It shows that 11 out of 12 times the model said the first stage of the rocket landed it actually landed, and just 1 time the model failed.
- It shows as well that 4 times out of 6 times the model said that the first stage of the rocket didn't landed the model was right, and just 2 times failed.



Conclusions

- It is possible to create a classification model to predict whether the first stage of a rocket will land successfully or not.
- There is a correlation between PayloadMass, Orbit, LaunchSite and the success of the landing of the first stage of a rocket launched by SpaceX.
- With the available data, we found the best predictive model to find if the landing of the first stage off a rocket launched by SpaceX will be successful is a decision tree classifier with parameters: tuned hyperparameters :(best parameters) {'criterion': 'gini', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
- This project can help to guide future investigations about the industry of rockets and spacecraft like ¿How the payload mass of a rocket affects the success rate of landing for each orbit? Or ¿What is the best model of classification for the falcon 1 rocket?

Appendix

- All the notebooks and python files used in this project can be found in the link: https://github.com/ChruedaM/Coursera_Capstone.git

Thank you!