IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Paulo Albuquerque

05/03/2022

Executive Summary

For this study, a series of data science methodologies was carried out so that it was possible to understand, analyze and make predictions in the face of the behavior of the launches of the Space X company's rockets, between the years 2010 to 2020, as a work of conclusion of the IBM Data Science course. After retrieving the data, it was possible to list the fundamental parameters that ensured a successful launch and landing, such as launch location, rocket load, year of launch, booster model, etc. To this end, graphs were created to better visualize these data, showing the relationship between each parameter mentioned above. In this way, after understanding the database and processing it, it was possible to carry out tests of prediction models of the success rate for future releases through the use of machine learning algorithms.

Introduction

As mentioned, the work aims to understand and make predictions, through machine learning, to list whether a launch, and landing, of the Space X company's rockets will be successful. The work serves as the final project of the IBM Data Science Professional Certificate, in this way it proposes the application of the knowledge obtained during the 10 months of the course.

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Section 1

Methodology

Methodology

The synthesis of the methodology is given:

- Data collection methodology:
  - Collected data from an API (Space X REST API)
  - Data Collection - Scraping
- Data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

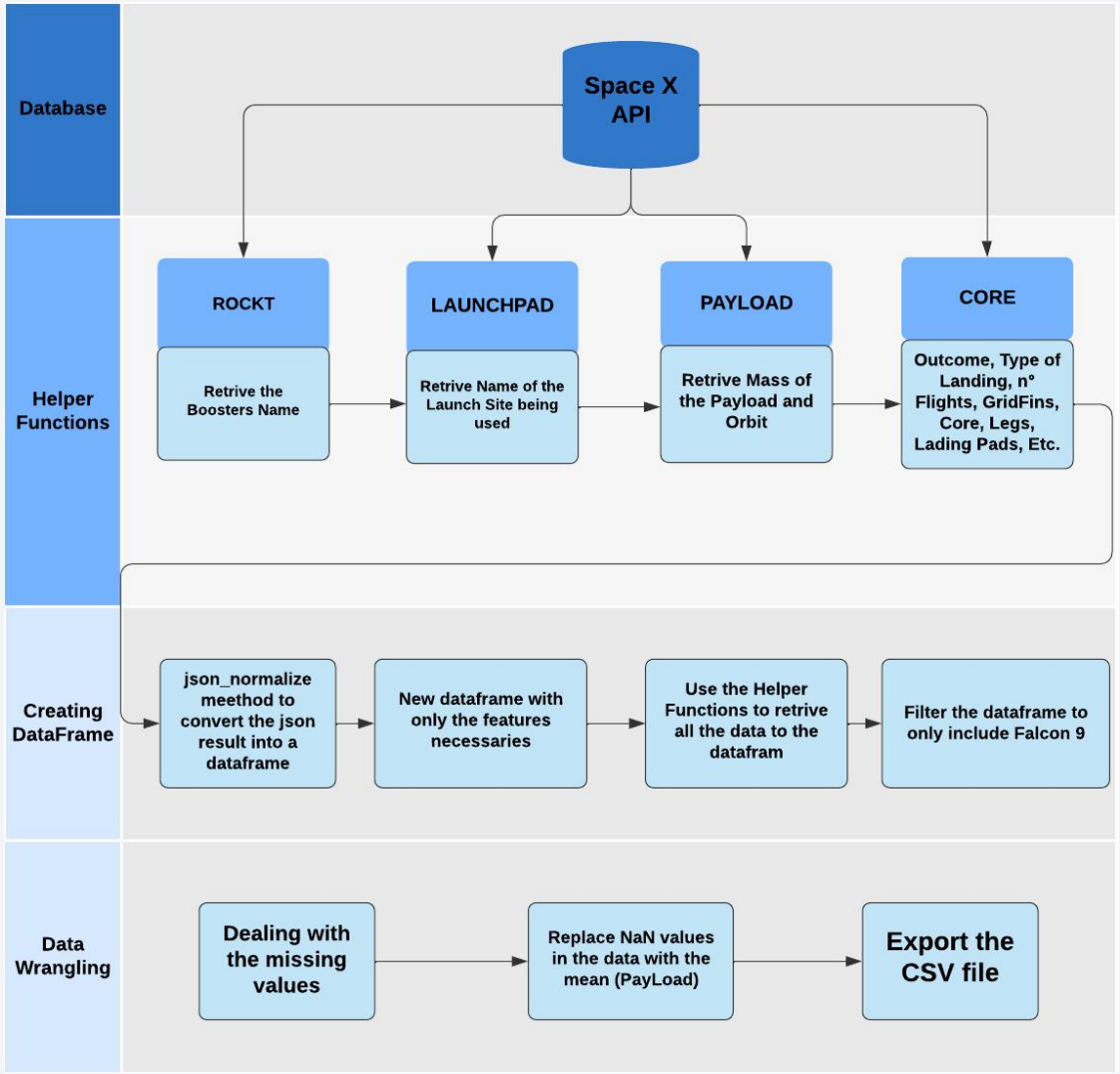
Data Collection

For the data retrieve section, two main methods were used, which are data collection through a REST API and also through Web Scraping.

For both, flowcharts were created to better understand how the data recovery from the database took place.

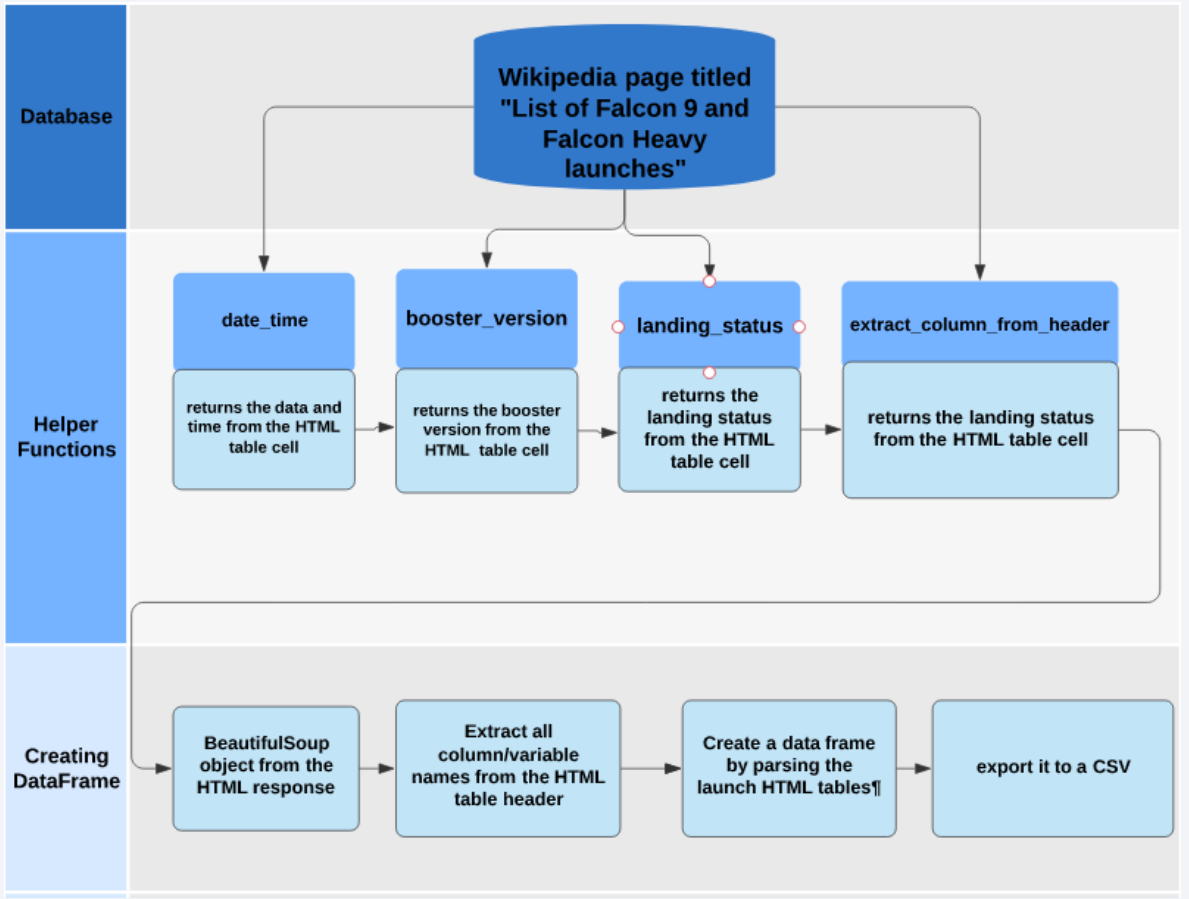
Data Collection – SpaceX API

- As an external reference and peer-review purpose here is the GitHub URL of the completed SpaceX API calls notebook :
- https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API%20Lab.ipynb



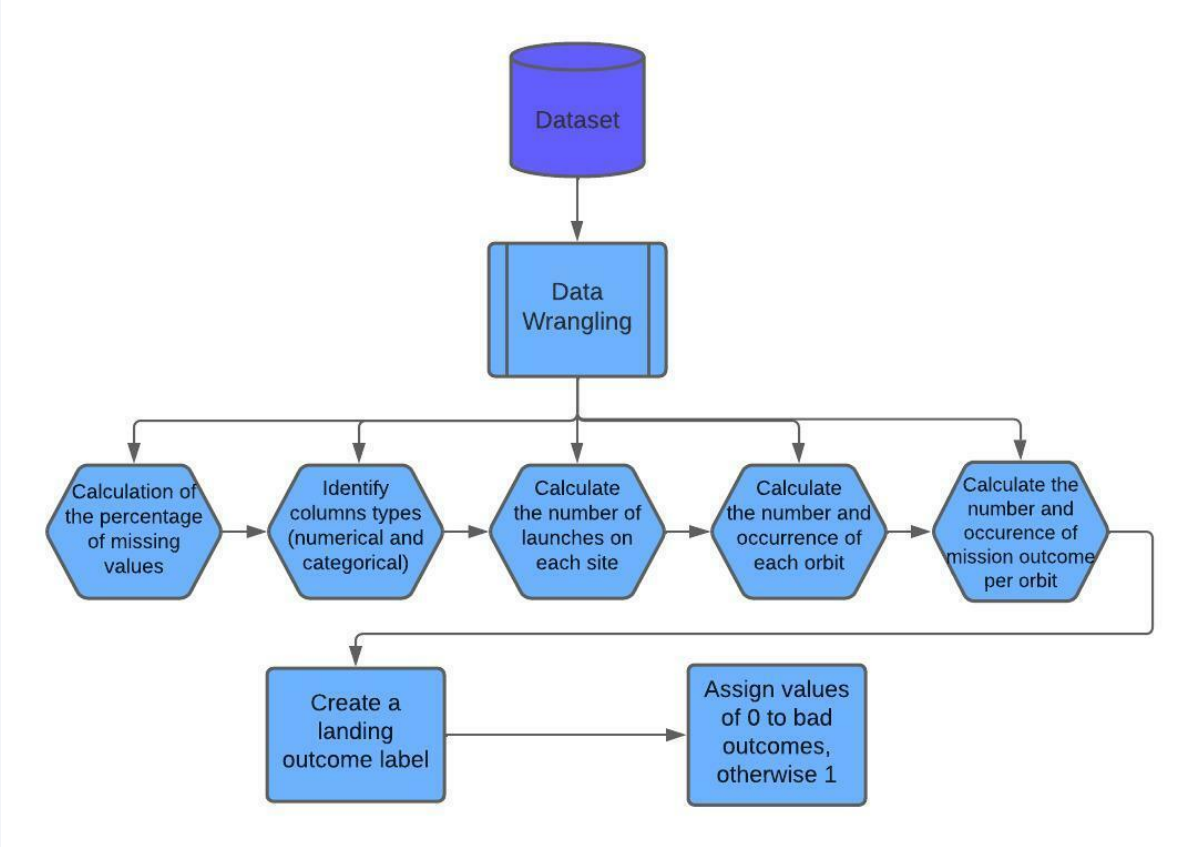
Data Collection - Scraping

- As an external reference and peer-review purpose here is the GitHub URL of the completed SpaceX API calls notebook :
- https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident;
- Data wrangling we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- GitHub URL of your completed data wrangling related notebooks: https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

A fundamental part of data analysis is the possibility of visualizing the relationships between numerous variables (categorical or numerical) and, from there, making decisions regarding the creation of models and understanding which factors will interfere in the result we aim for. Therefore, throughout the project, several graphs were created to explain, in the best possible way, these relationships and trends. The graphs created during the study were:

- Scatterplots :
    - FlightNumber vs. PayloadMass;
    - FlightNumber vs LaunchSite;
    - Payload vs Launch Site;
    - FlightNumber vs Orbit type;
    - Payload vs Orbit type.
- Bar Chart;
    - Sucess rate of each orbit.
- Line Chart;
    - Year vs Success Rate.

GitHub URL:https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries were also performed to simulate the recovery of datasets stored in a database managed by SQL sequences. The commands performed (which you can check in the attached link) were:

- Display the names of the unique launch sites in the space mission;
- Display 5 records where launch sites begin with the string 'CCA';
- Display the total payload mass carried by boosters launched by NASA (CRS);
- List the date when the first successful landing outcome in ground pad was achieved;
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
- List the total number of successful and failure mission outcomes;
- List the names of the booster_versions which have carried the maximum payload mass;
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Rank 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.

GitHub URL: https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- For the spatial understanding of the launches and their outcomes, maps were created with markers of the launch sites, successful/failed launches and, also, the distances between the launch sites and the points of importance (lanes, trails, sea coast) were calculated. among its proximities;
- The addition of these aforementioned objects was necessary to analyze and establish any parameter that would positively or negatively influence a launch. Given the technical complexity of a rocket launch, something like the launch site's location on the Earth globe can influence the mission's chances of success.
- Here is the link to the codes used to create the maps and their respective markers : https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- For a possible presentation to stakeholders, the use of the Plotly Dash tool is very useful, due to the fact of the easy absorption of the contents displayed by the created dashboard. In this way, interested parties can navigate through several ways of visualizing the data previously treated and implemented in graphs;
- For the created dashboard, it is possible to verify the ratio of the success ratio when compared to the Payload mass (in kg) for each Launchsite and their respective Boosters. In addition, it is clear which launchsites are more successful.
- GitHub URL : https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/Dash_SpaceX_Capstone%20(1).ipynb

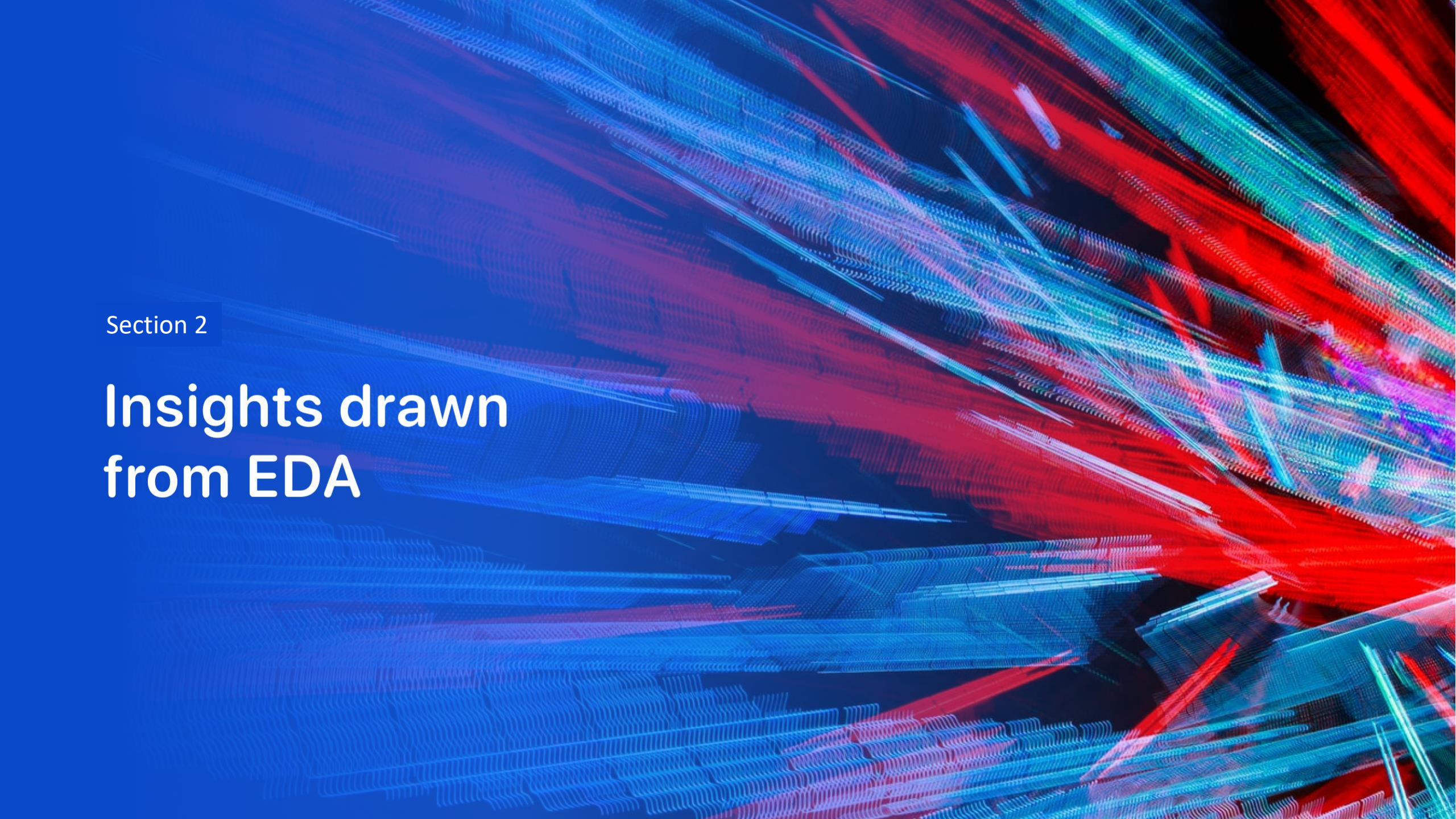
Predictive Analysis (Classification)

- Using the scikit-learn, seaborn matplotlib libraries it was possible to evaluate three prediction models and find their best Hyperparameter, namely:
  - Support Vector Machine;
  - Classification Trees;
  - Logistic Regression.
- The evaluation methods of the models were:
  - Accuracy to the test set (score method);
  - Confusion Matrix;
  - GridSearchCV.
- GitHub URL: https://github.com/albuquerquejp/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

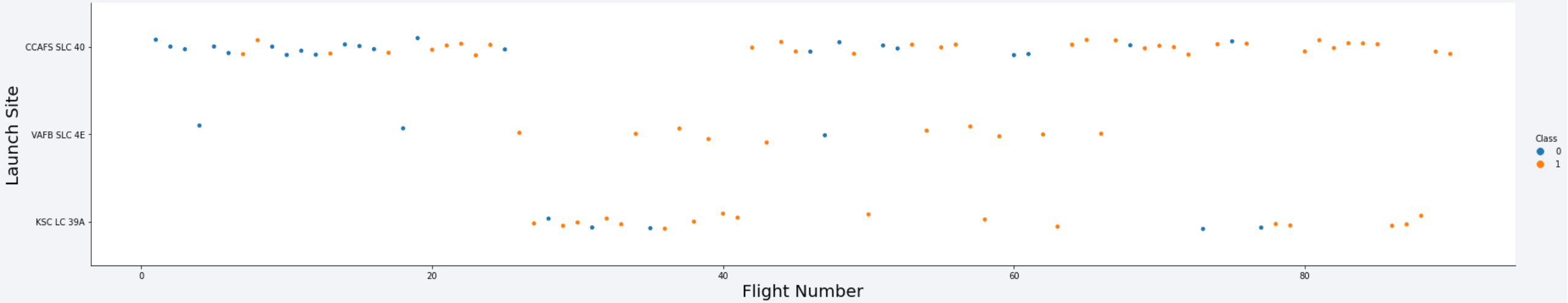
Results

Exploratory data analysis results

- The use of SQL proved to be fundamental for fast and effective data recovery;
- The creation of graphs, using several libraries, provided a fundamental insight for comparing results between different variables and quantified success reasons, successful booster trends as well as analytical visualization of launch sites;
- The predictive analysis returned good results, however, through accuracy assessment methods, it was possible to list the best ones to be applied in the prediction of future outcomes.

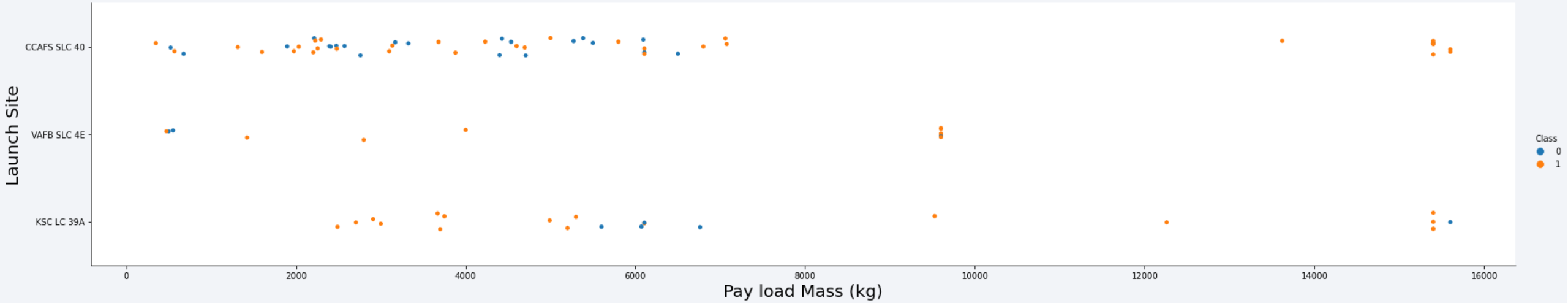

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Some patterns were found, such as:
  - Success rate (indicated by “Class 1”) significantly increased after the first 20 launches;
  - There was a gap between the launches carried out by the website “CCAFS SLC 40” between the 20-40 flights;
  - The same gap is seen for the VAFB SLC 4E, when it has more significance in terms of the number of flights between the numbers 20-60, after which it was no longer used.

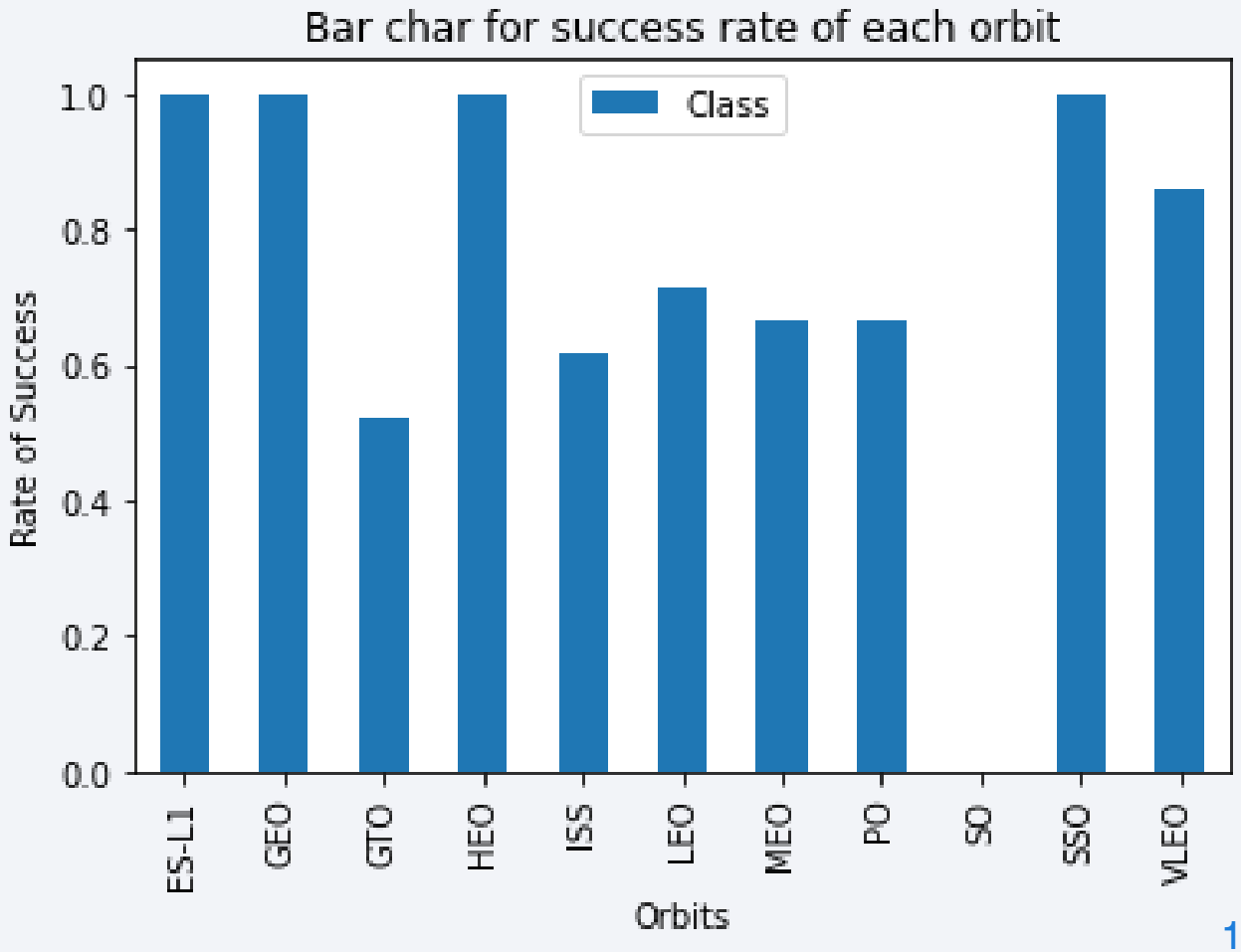
Payload vs. Launch Site



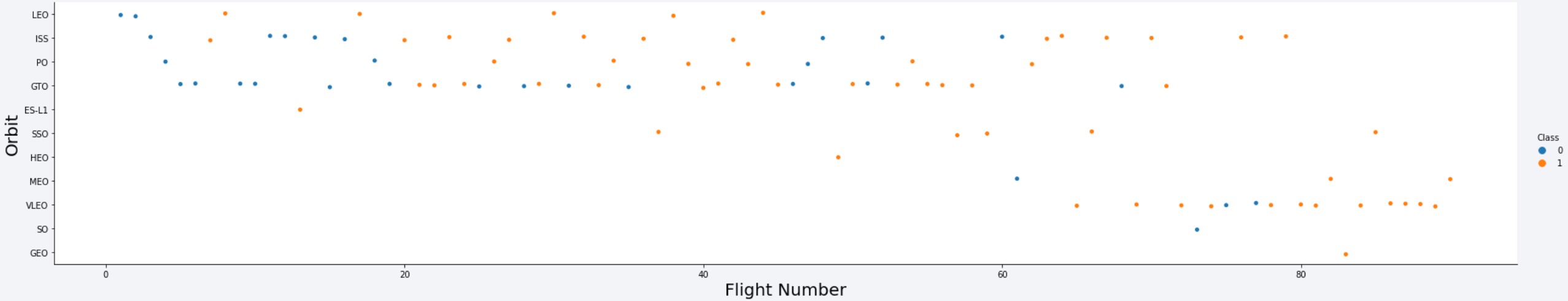
- No launches for heavy rockets (Pay load Mass >10000kg) on Launch Site VAFB SLC 4E;
- Heavy payload mass rockets are more successful than lighter ones.

Success Rate vs. Orbit Type

- Viewing the graph, it is possible to list the orbits that had the highest level of success, namely:
  - ESL-L1
  - GEO
  - HEO
  - SSO
- All of the above were 100% successful.;
- The SO orbit can be considered an outlier, due to the rate of success being equal to 0.

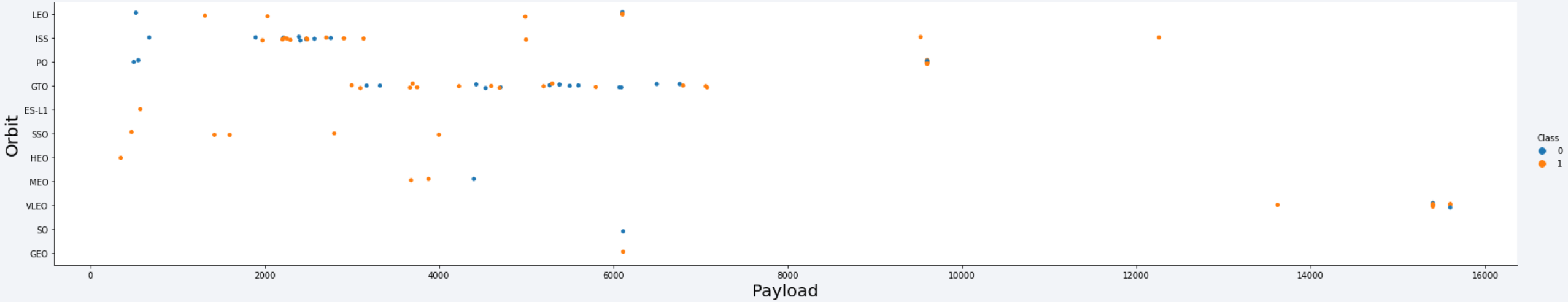


Flight Number vs. Orbit Type



- 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

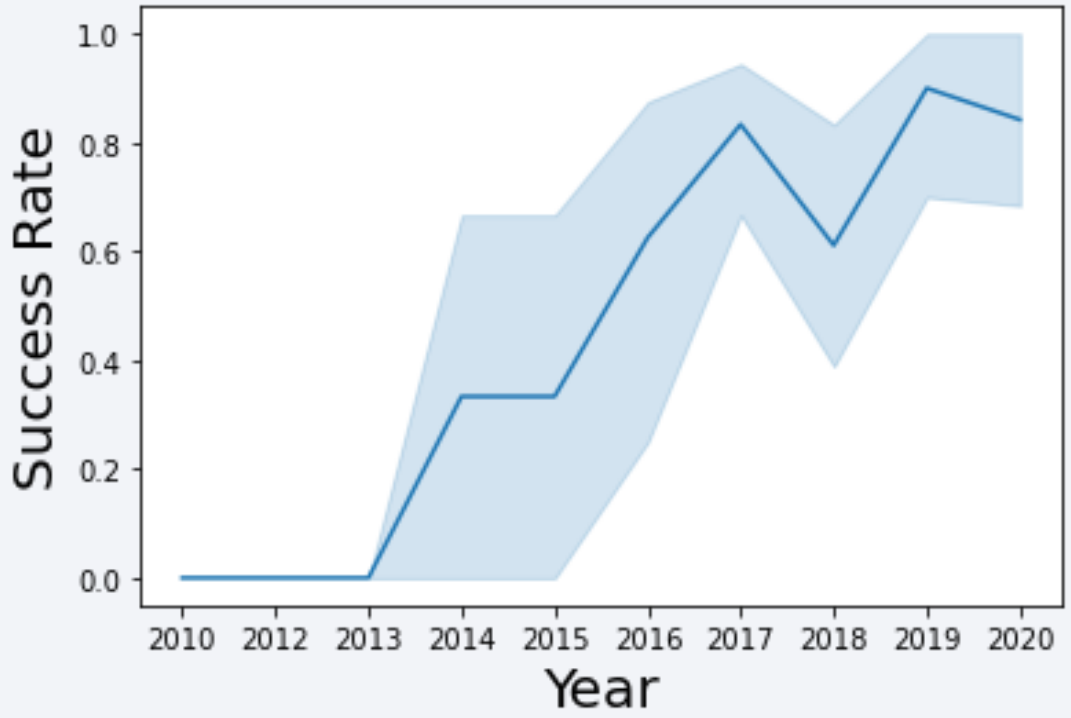


The VLEO (Very Low Earth Orbits) Orbit has the highest Payload mass in kg, even being the closest to the earth (altitude less than 450 km), going from finding as a logical thought that the higher the altitude the higher the Payload.

Orbits such as GTO, ISS and PO were the ones that had the greatest variation in terms of Payload, ranging from values less than 4000kb, to greater than 12000kg.

Launch Success Yearly Trend

- By viewing the graph, it is possible to confirm a constant growth in terms of the success rate;
- The success rate has greatly increased from the years 2015 and 2016.



All Launch Site Names

Names of the unique launch sites:

- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-43

Launch Site Names Begin with 'CCA'

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

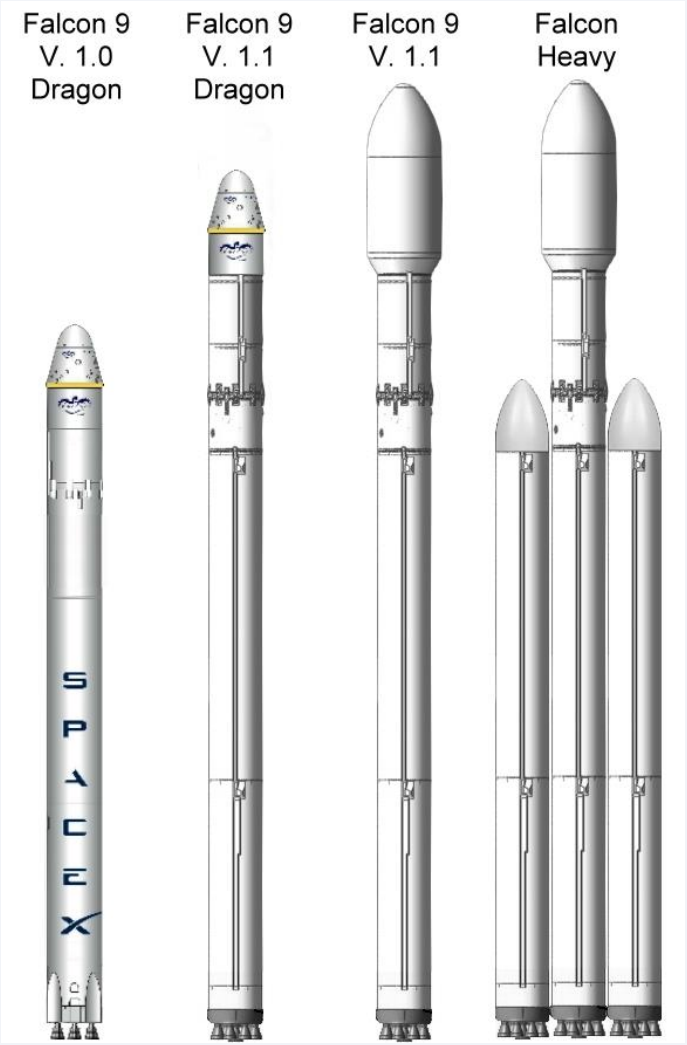
Total payload carried by boosters from NASA:

111268,00 kg

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1:

2534,00 kg



First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad:

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

- JCSAT-14
- JCSAT-16
- SES-10
- SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes:

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

Boosters Carried Maximum Payload

Booster which have carried the maximum payload mass:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

DATE	landing_outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank 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:

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

Section 3

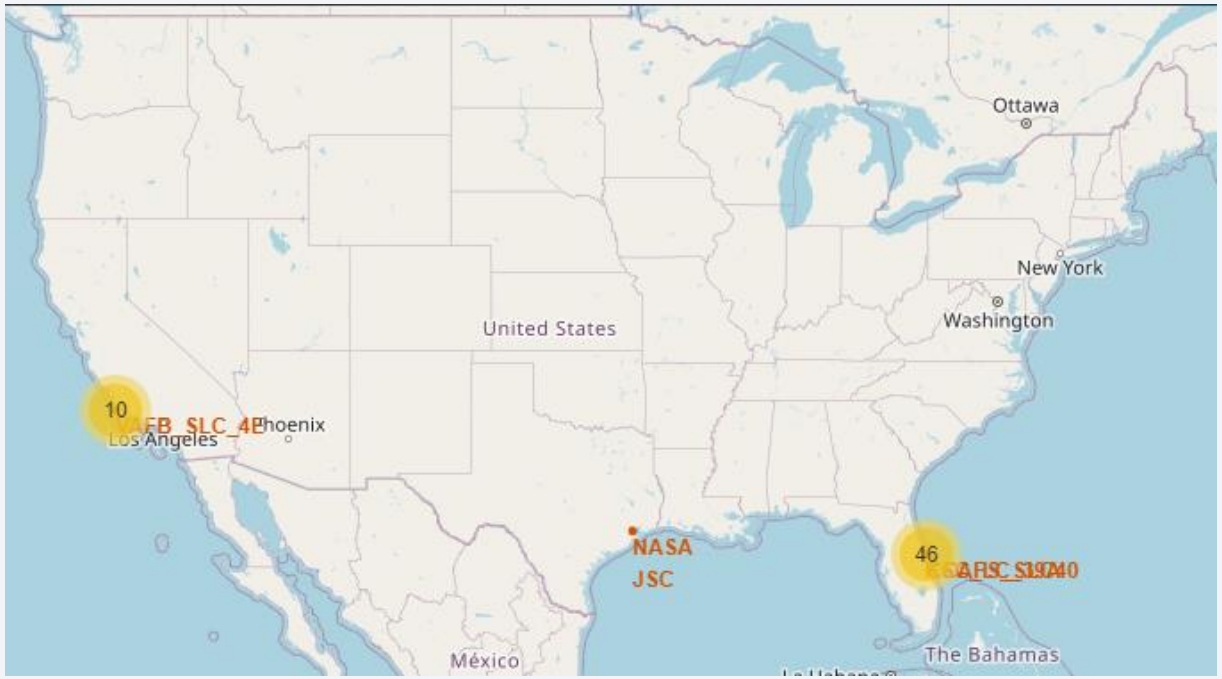
Launch Sites Proximities Analysis

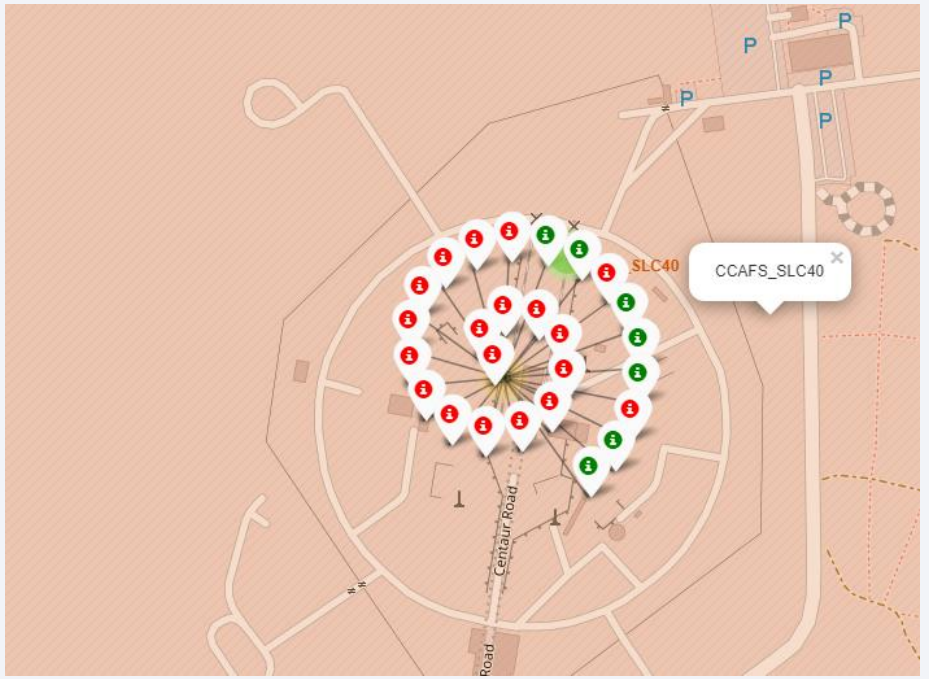
Map from all the launch sites

On this map the locations for the following Launch Sites were marked (all in the United States):

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

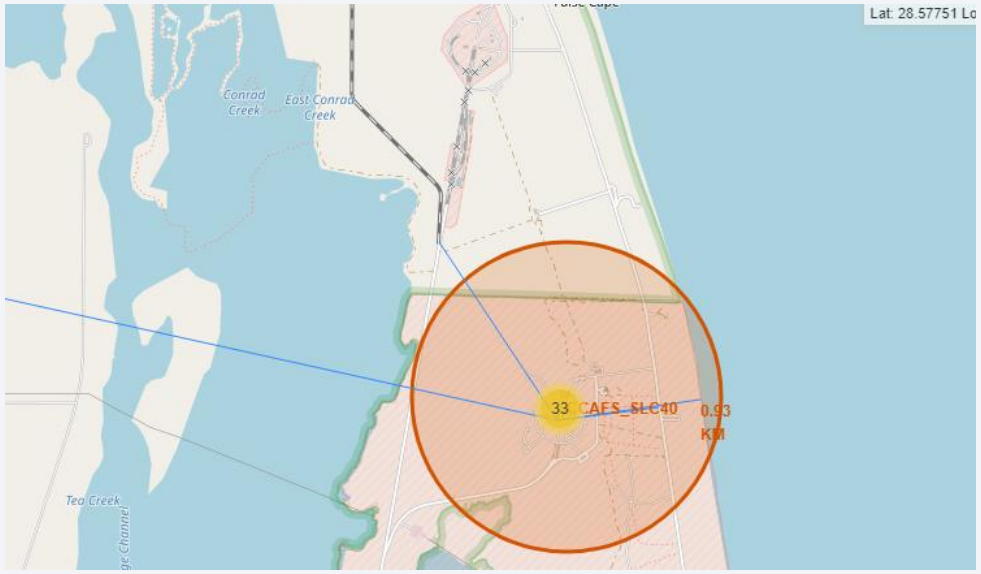
Map of the Locations of each Launch and their Outcome

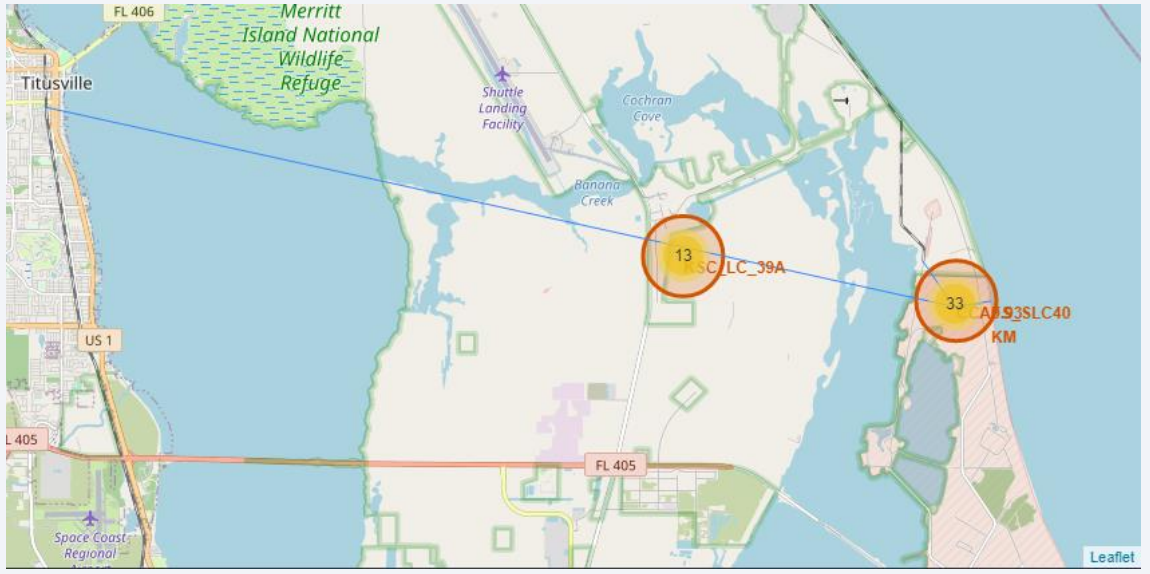




After demarcating the locations of the Launche Sites, it was also possible to mark the location for each launch carried out on them. The green color markers indicates those that were successful, while the red markers are for those that failed.

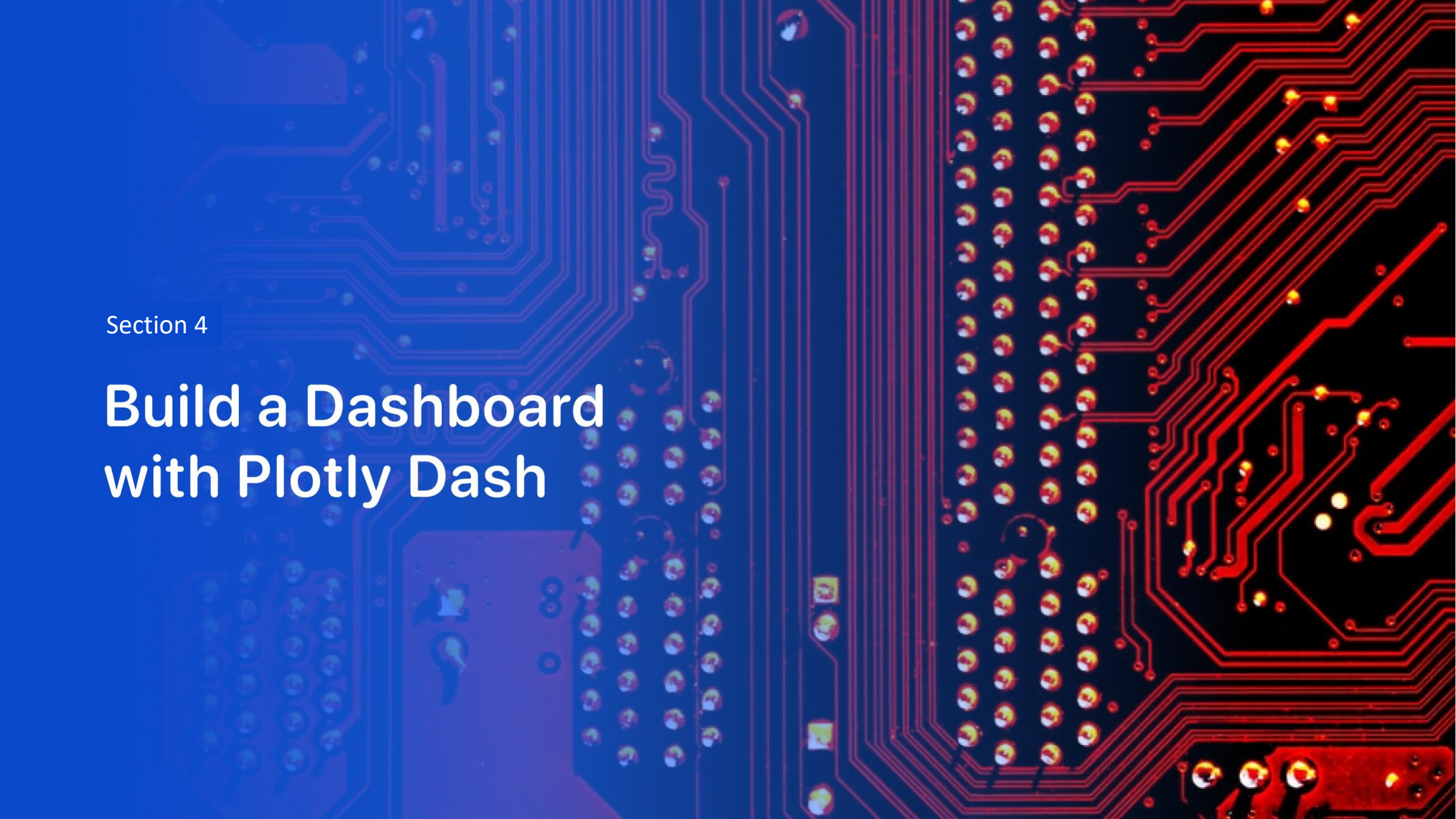
Distance from closest, city, railway and highway and cost



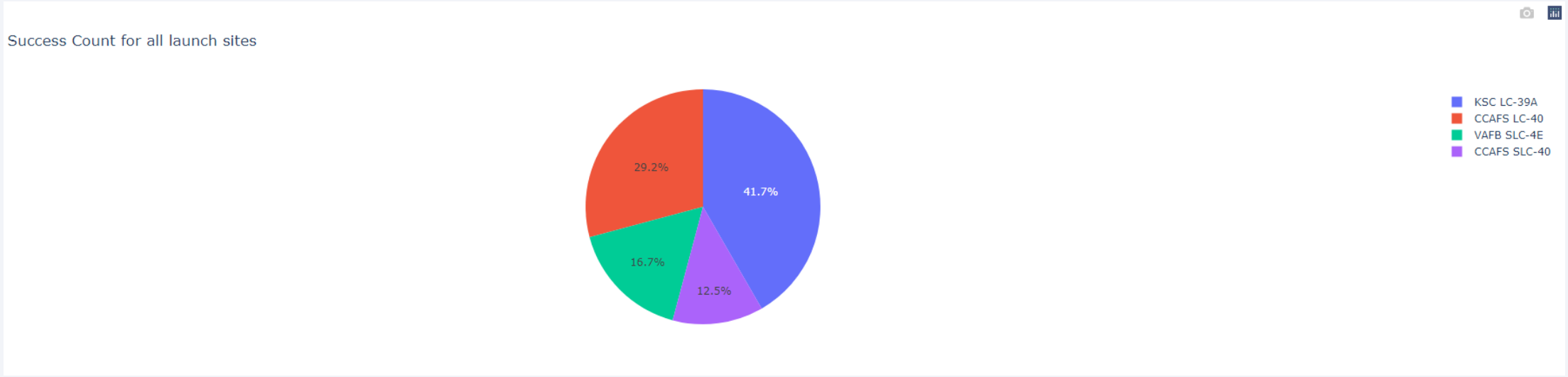


For the map above, the distance (using the blue line) from the Launch Site CCAFS SLC40 to points of importance was marked, which are:

- Distance from a closest city: 23.32 km
- Distance from a closest railway: 1.38 km
- Distance from a closest highway: 0.65 km
- Distance from a closest Coastline: 0.93 km


Section 4
Build a Dashboard with Plotly Dash

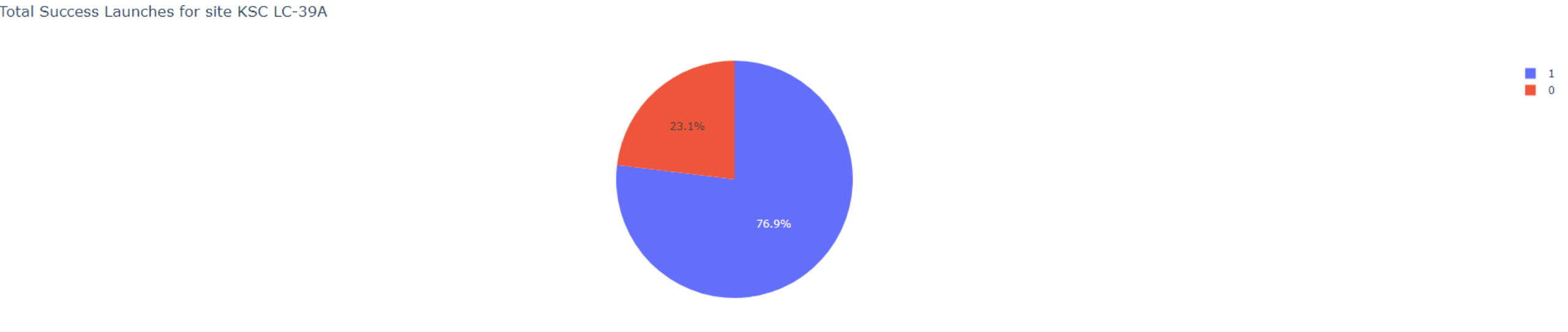
SpaceX Launch Records Dashboard



The chart above shows the large number of successful launches for the Launch Site KSC LC-39A, with 41.7% of all successes. On the other hand, VAFB SLC-4E and CCAF SLC-40, combined, amount to no more than 30% of all successful launches.

This ensures, for future releases, a greater perception of the probability of success taking into account the Launch Site.

Total Success Launches for site KSC LC-39A



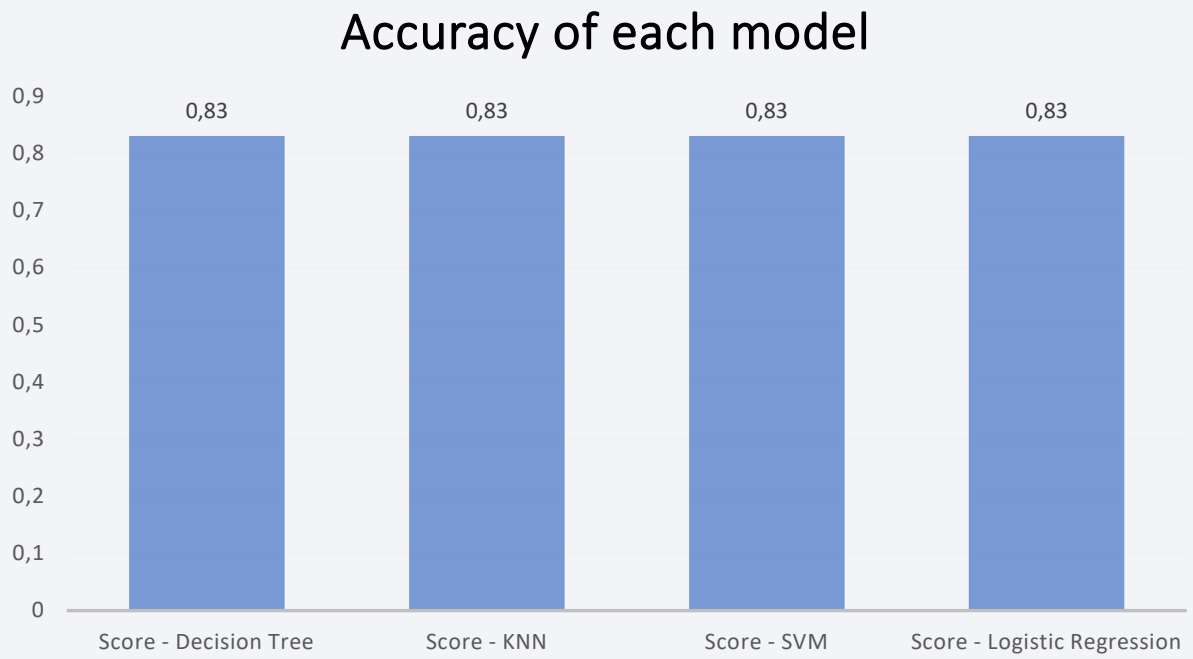
Now, taking into account only the Launch Site with the highest number of successes (KSC LC-39A), it was possible to confirm a great total success rate for all their launches. For all launches already carried out on this Launch Site, it obtained a total of 76.9% success.

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Section 5

Predictive Analysis (Classification)

Classification Accuracy



In view of all the models used in the project, for all of them, the accuracy rate for the set test, using the score method, was 0.83. This value confirms the excellent performance of the chosen prediction algorithms.

Confusion Matrix

Checking the values analyzed by the confusion matrix (for logistic regression model), it is possible to list the following facts :

- The model was very efficient to predict the launches that landed (12/12);
- However, it had an accuracy of only 50% for the landings that did not land (3/6);
- Giving a relatively high value for false negatives.

Conclusions

- The visualization of data through scatter plots proved to be fundamental to establish the best and most efficient variables to be used in predictive models.;
- A large increase in successful launches from the year 2015 shows the great technological advancement of the Space X company, confirmed by the graph shown previously;
- The relationship between the success rate and the target orbit proved to be very strong, however, on the other hand, some relationship between the number of flighs and orbits only showed for some, not all, orbits;
- All predictive classification methods showed good performance in terms of prediction for new launches, since for both the training set and the test set it was possible to obtain accuracy rates above 80%.

Appendix

- https://www.python.org/
- https://github.com/albuquerquejp/Applied-Data-Science-Capstone
- https://pandas.pydata.org/
- https://www.spacex.com/
- https://numpy.org/

Thank you!