IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<SHU PENG>
<Oct 2021>

Outline

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

Executive Summary

◆ Summary of methodologies

✓ First, Collect the Falcon 9 launch data via SpaceX API, and do the web scrapping, then store it into IBM Db2 for further analytics.

✓ Second, do data cleaning, replace all missing values with either mean of the column or 0.

✓ The visualization process of the data

✓ Finally, perform a predictive analysis (several supervised learning methods) to train and determine the best model for prediction and then give the prediction over the success of future landings (success or fail).

Introduction

◆Project background and context

- Since SpaceX can reuse the first stage, the Falcon 9 rocket launch was announced at a cost of 62 million dollars (relatively cheap), while other vendors cost more than 165 million dollars each. Therefore, from the processing, analysis and modeling of data related to Falcon 9, we can predict whether the first stage will land successfully or not and hence predict the cost of each launch (Analytic Approach).

- To do this, some data about Falcon 9 should be firstly scrapped and cleaned. (Data Requirements).

Section 1

Methodology

Methodology

- First, data were collected through web scraping from Wikipedia and SpaceX API. Then, the data should be cleaned and filtered.
- After that, perform the data wrangling
- In this step, analyzing the data from the Payload Mass column some missing values

Therefore, we completed these missing values with the arithmetic mean of this column.

Methodology

Then, we get the insights about the data using some tools for visualization

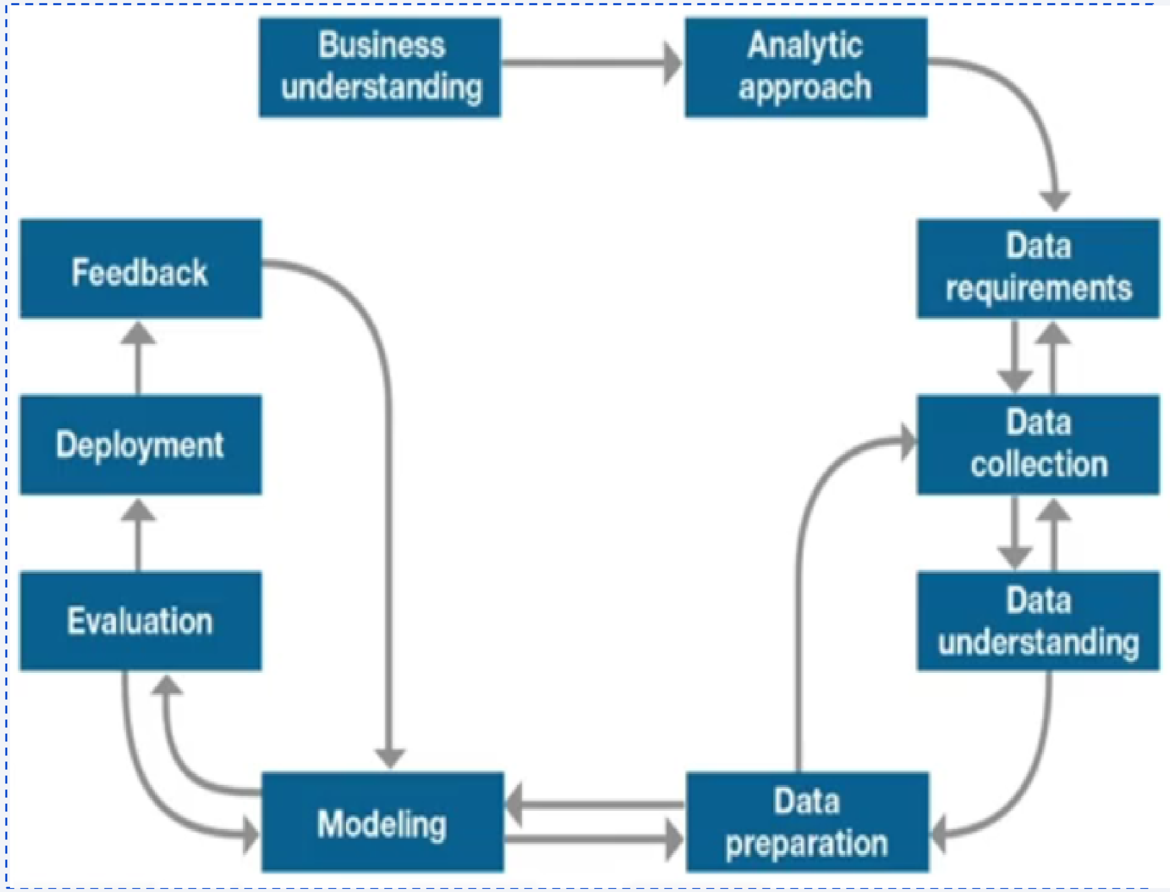
and SQL queries.

- Here, we can see data distribution into a map of launch sites available in the dataset to visualize successful and unsuccessful landings at each launch site.
- Next, we analyze the relation between the features of dataset using Seaborn library.
- Predictive analysis using classification models
- Finally, after all these steps, we prepare the data for modeling. Then, some

supervised machine learning algorithms are applied and their results compared.

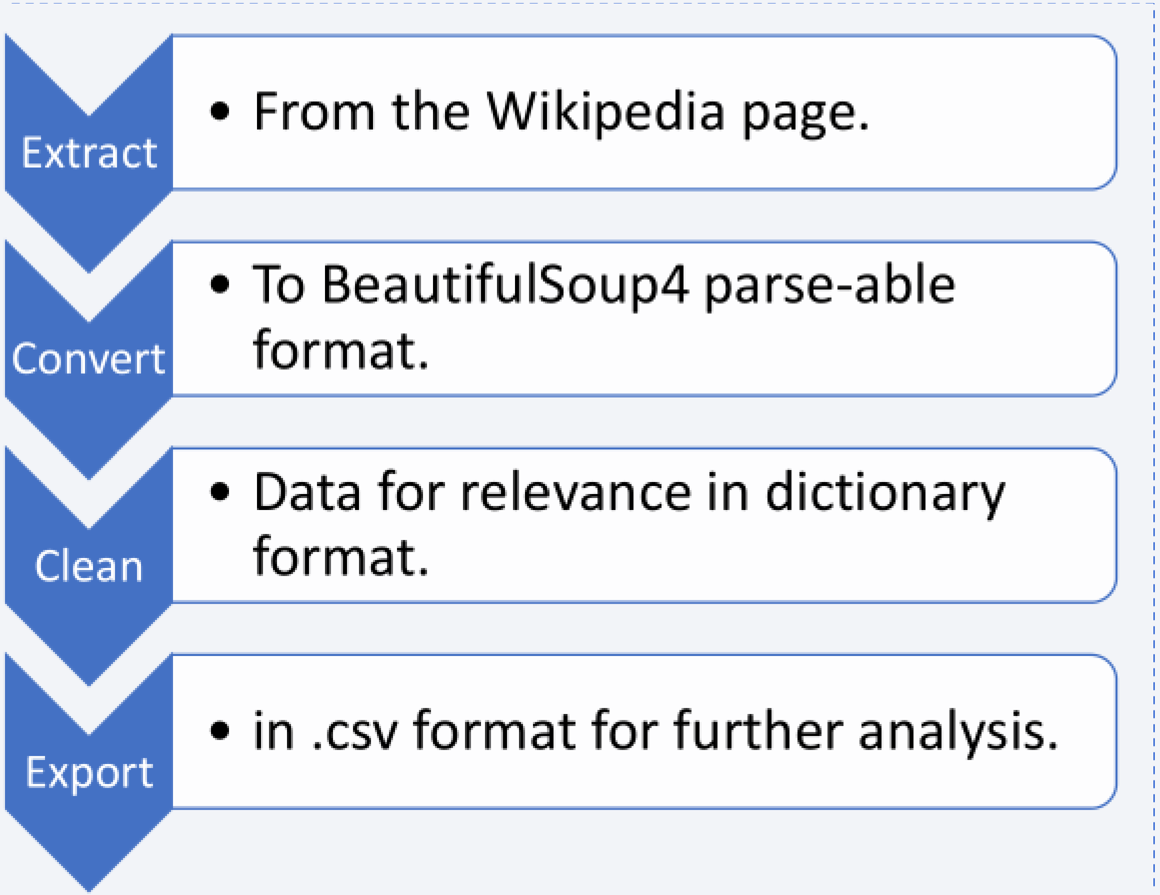
Data Collection – SpaceX API

- The flowchart on the right demonstrates the complete cycle of a Data Science project. In final section of the project, we reach the stage of evaluating machine learning models, and then applied to our prediction problem.



Data Collection - Scraping

- Request for the Wikipedia page for data collection
- Parse the table data from html text (SpaceX API)
- use beautifulSoup4 library.
- Create pandas data frame from table data.
- Construct the data and store it in CSV.



Extract
- From the Wikipedia page.

Convert
- To BeautifulSoup4 parse-able format.

Clean
- Data for relevance in dictionary format.

Export
- in .csv format for further analysis.

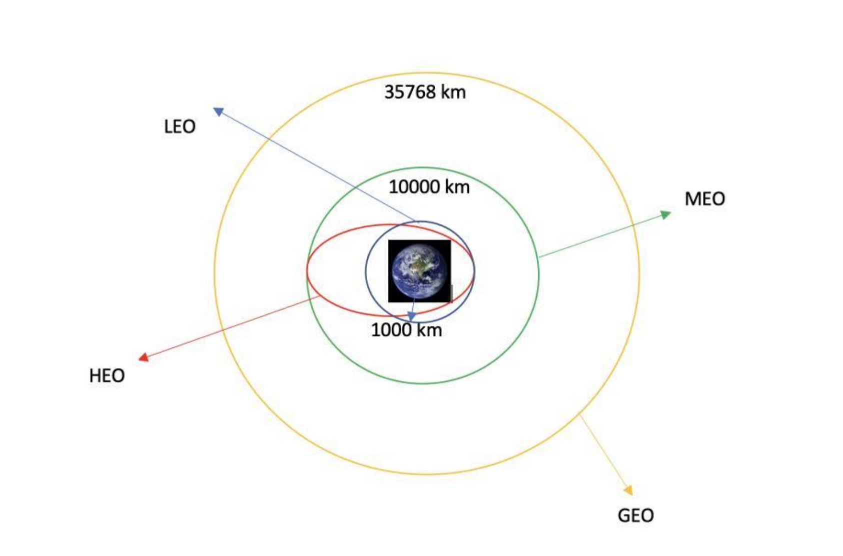
Data Wrangling

- The first step is to calculate the number of launches per site.

- Then, the diagram shows the number of occurrence of each orbit.

- Also, it presents the number of occurrences of outcome per orbit.

- Finally, create landing outcome label.



https://github.com/SHU2022/testrepo2/tree/master

EDA with Data Visualization

- Data visualization is a crucial step to know more about the data,
- Besides allowing us to visualize through graphs the relationship between two
- or more resources, and this analysis helps us to carry out a good modeling of
- the data later on.

- https://github.com/SHU2022/testrepo2/tree/master

EDA with SQL

- In most cases, the dataset should be analyzed and save as a .csv (comma
- separated values) file, perhaps on the internet, or even kept in a specific
- database.

- Therefore, as a data scientist, knowing how to query a database is often become necessary.

https://github.com/SHU2022/testrepo2/tree/master

Build an Interactive Map with Folium

- Map objects which are created and added to the folium map are given below:
- Markers: Added to mark a specific area with a text label on a specific coordinate.
- Circles: Added to highlight circle areas with a text label on a specific coordinate.
- Marker Cluster: Marker clusters were used to simplify the containing many markers having the same coordinates.
- Mouse Position: Used to get coordinate for a mouse over a point on the map (proximities). It helps to find the coordinates easily of any points of interests while exploring the map.

https://github.com/SHU2022/testrepo2/tree/master

Build a Dashboard with Plotly Dash

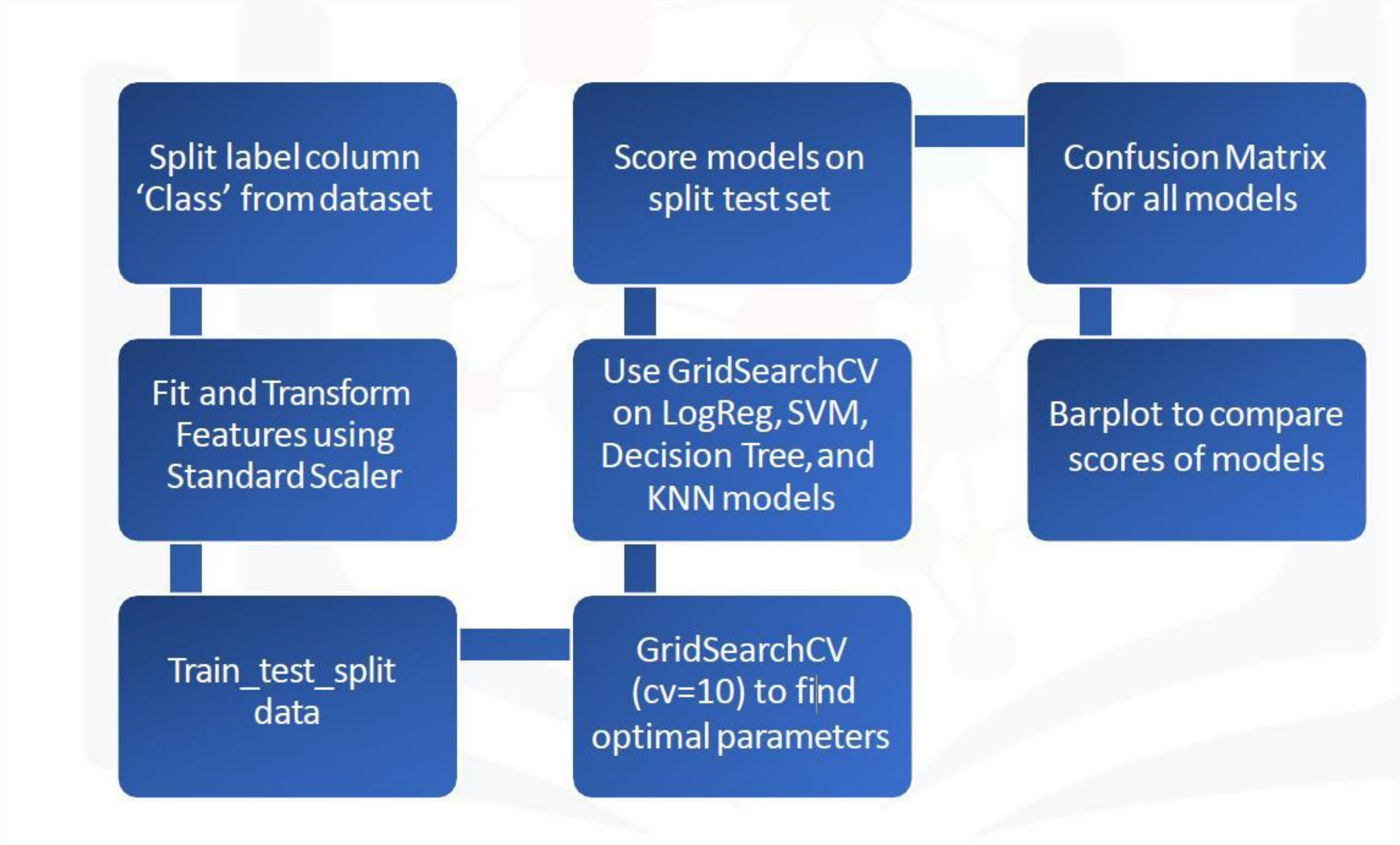
- Dash is a Python framework created by Plotly to create interactive web applications.
- The plots and interactions have been added then, users can perform interactive visual analysis on SpaceX launch data in real-time.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter plot chart.

- https://github.com/SHU2022/testrepo2/tree/master

Predictive Analysis (Classification)

- Firstly, we have to create a column called ‘class’ which is our target. The data in this column
- is set to '0' (zero) if the first stage landing was not successful or it is set to 1 (one) if is a
- successful landing.
- So, we must to standardize all data values until they have the same weight for modeling.
- And lastly, the data is divided into training data (80%) and testing data (20%). Training
- data is used to train the model and testing data is used to evaluate the model.
- Then, the models are trained and hyperparameters are selected using the function
- GridSearchCV. This function make exhaustive search over specified parameter values for an estimator.

Predictive Analysis (Classification)



Split label column 'Class' from dataset

Fit and Transform Features using Standard Scaler

Train_test_split data

GridSearchCV (cv=10) to find optimal parameters

Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

Barplot to compare scores of models

Results

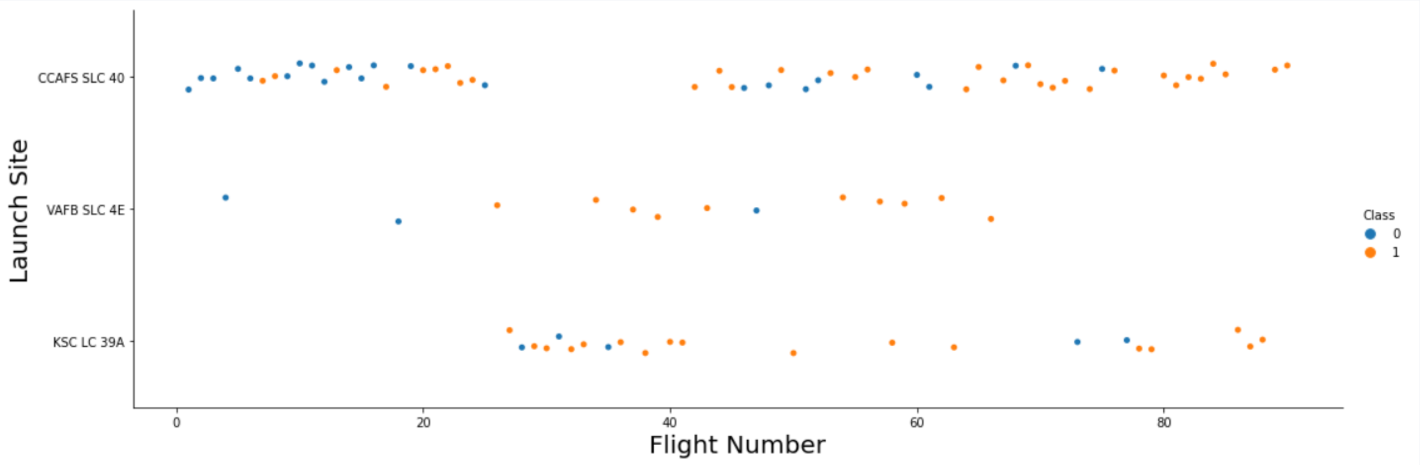
- The next slides show images from each phase of this project. All these results were generated in the notebooks provided in the previous slides (links).

- Models were built using Scikit-Learn, data were previously normalized and models hyper parameters were found using a Grid Search with a 10 fold cross validation, in the end the best performing model has been selected based on accuracy.

Section 2

Insights drawn from EDA

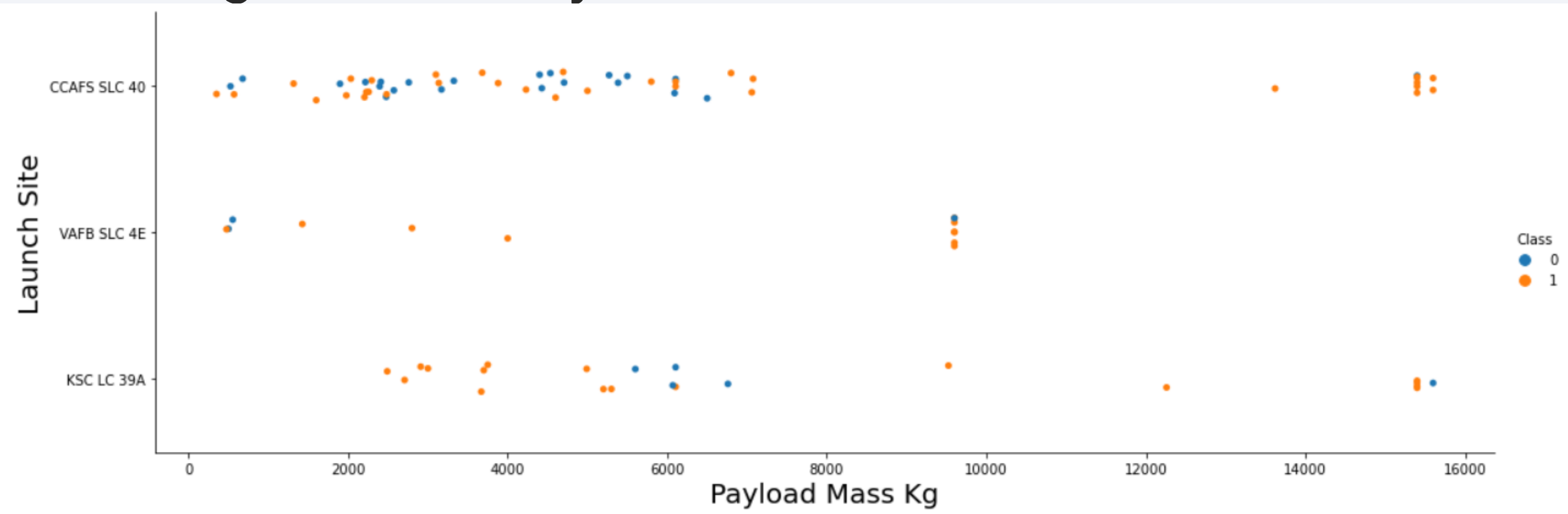
Flight Number vs. Launch Site



- In this graph, it is possible to visualize the relationship between Launch Sites and the success rate over the number of flights. It is clear that the success cases represented by the class 1 increase in same ratio as the number of flights. It can represent improvements made on Launch Sites or in another feature.

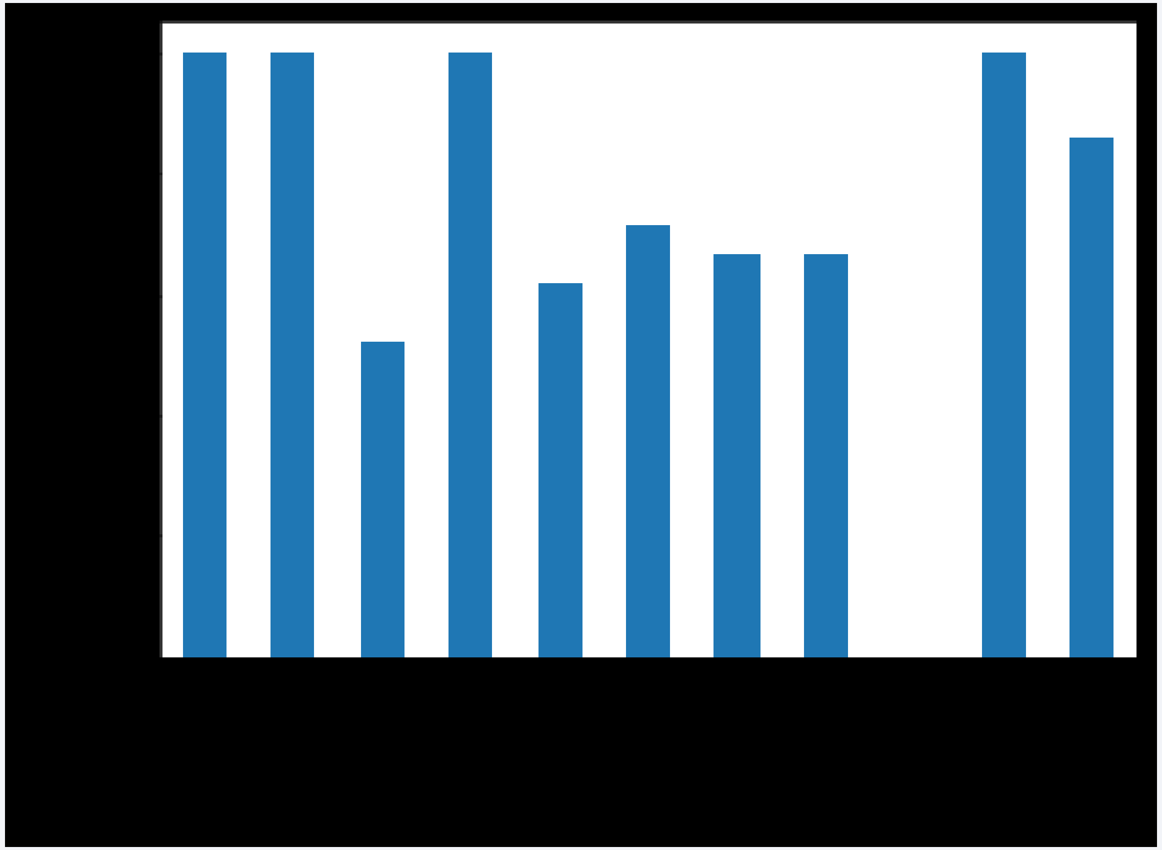
Payload vs. Launch Site

- In the CCAFS SLC 40 the distribution of successful or unsuccessful landing is balanced between payload mass from 0 to 8000, but is more successful in higher payload mass. In the VAFB SLC 4E Launch Site, most of landing was successful regardless of payload mass. As well as on the VAFB SLC 4E launch site, KSC LC 39A also had the most successful landings, but there is a detail when the payload mass is close to 6000 kg, because in this range the landings were mostly unsuccessful.

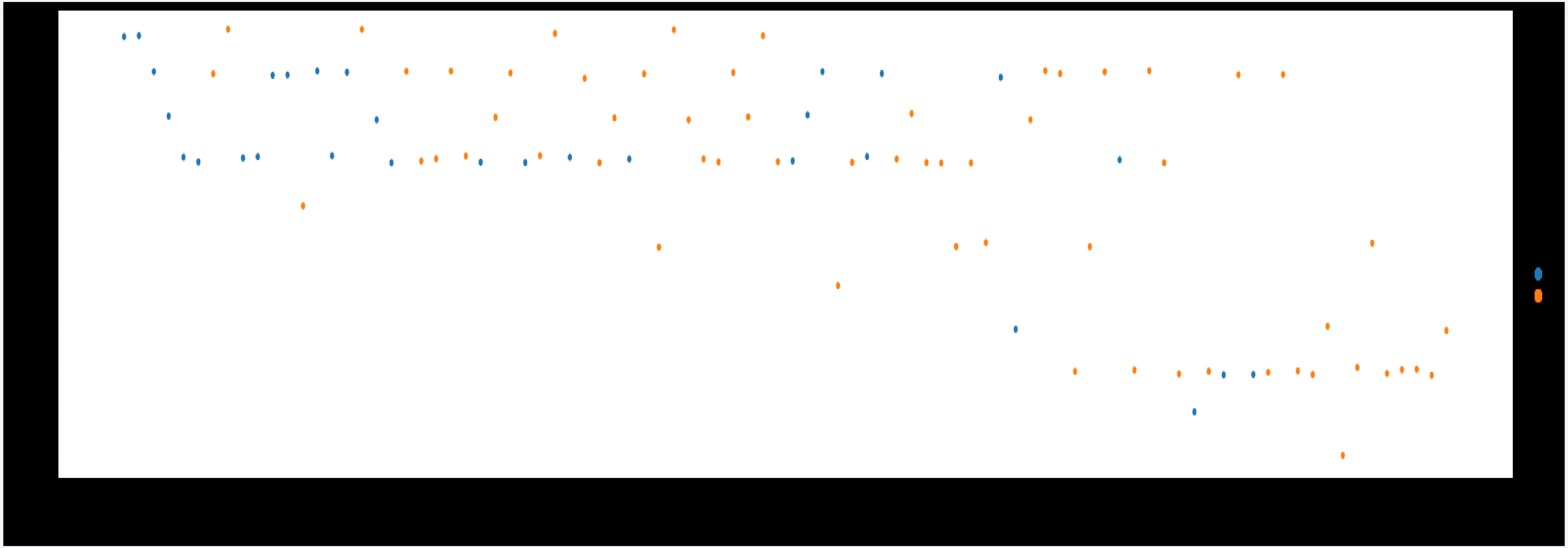


Success Rate vs. Orbit Type

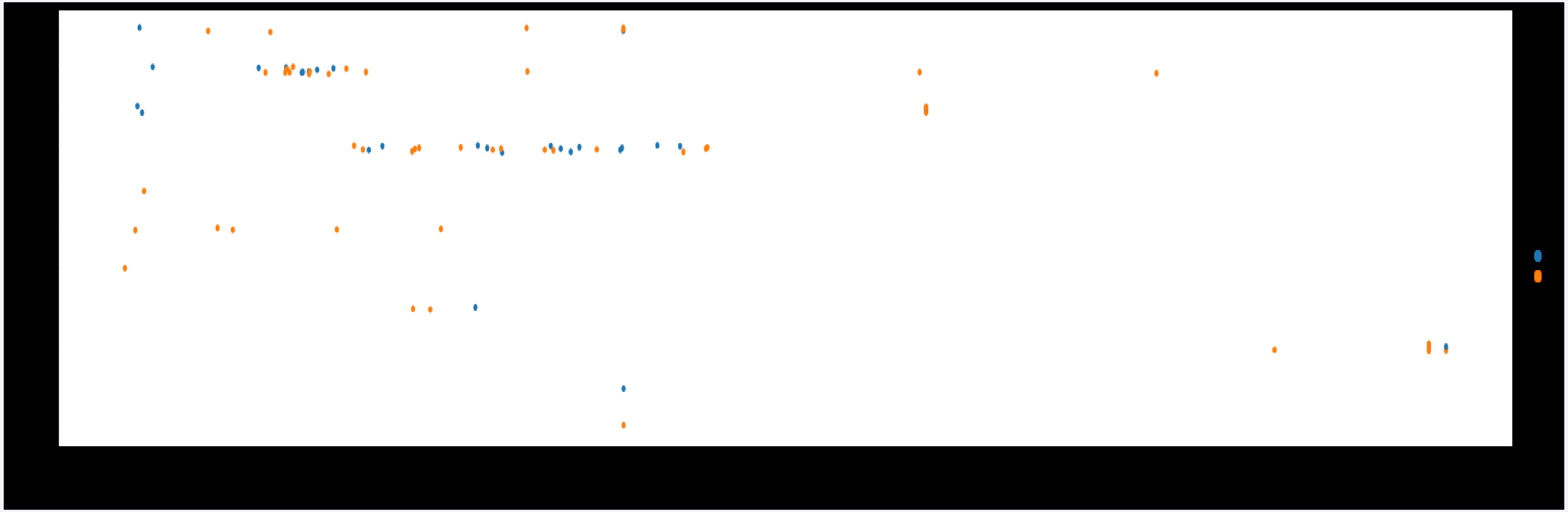
- There is a strong correlation between these two indeed as we can observe the SO or GTO Orbit type are quite risky
- as the success rate is below 0.6. However, some orbit type provide a 1.0 success rate which is perfect but can hide
- suspicious data. Indeed if for this orbit type only one rocket has been launched the reliability of this hypothesis is null.



Flight Number vs. Orbit Type

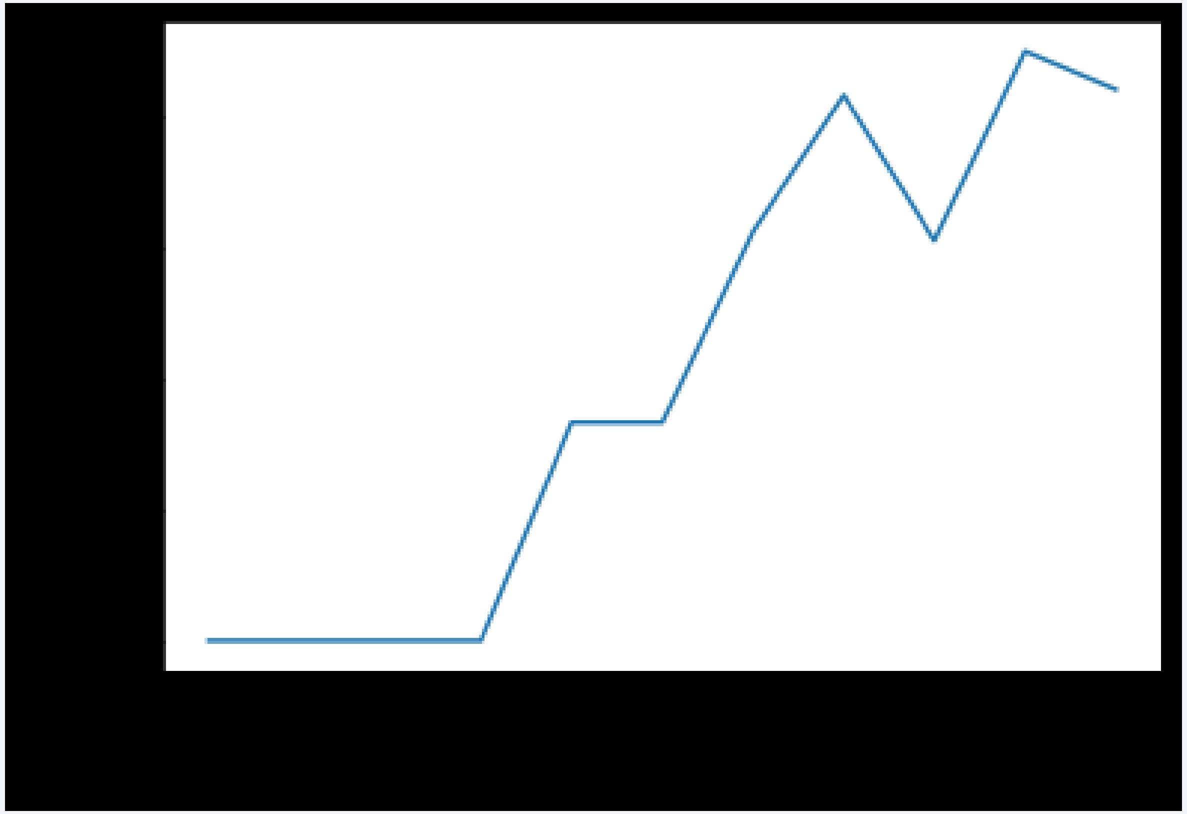


Payload vs. Orbit Type



Launch Success Yearly Trend

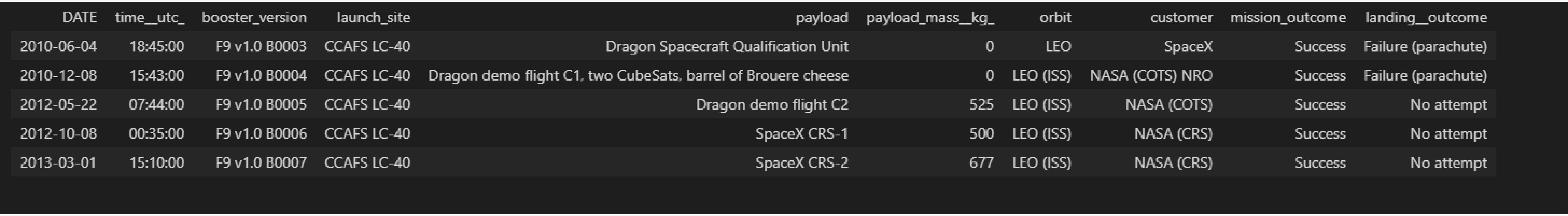
- Here the chart demonstrates that as Humans learn more and more through the years thanks of Sciences, it results in a significant rocket launches success rate increasing.



All Launch Site Names

- Launch sites are :
- • CCAFS LC-40
- • CCAFS SLC-40
- • KSC LC-39A
- • VAFB SLC-4E
- • SQL QUERY: 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
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

- The total payload carried by boosters from NASA is 45596 kg.
- • SQL QUERY: select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 4 2534 kg.
- • SQL QUERY: select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%';

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was 2015- - 12- - 22.
- • SQL QUERY: select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

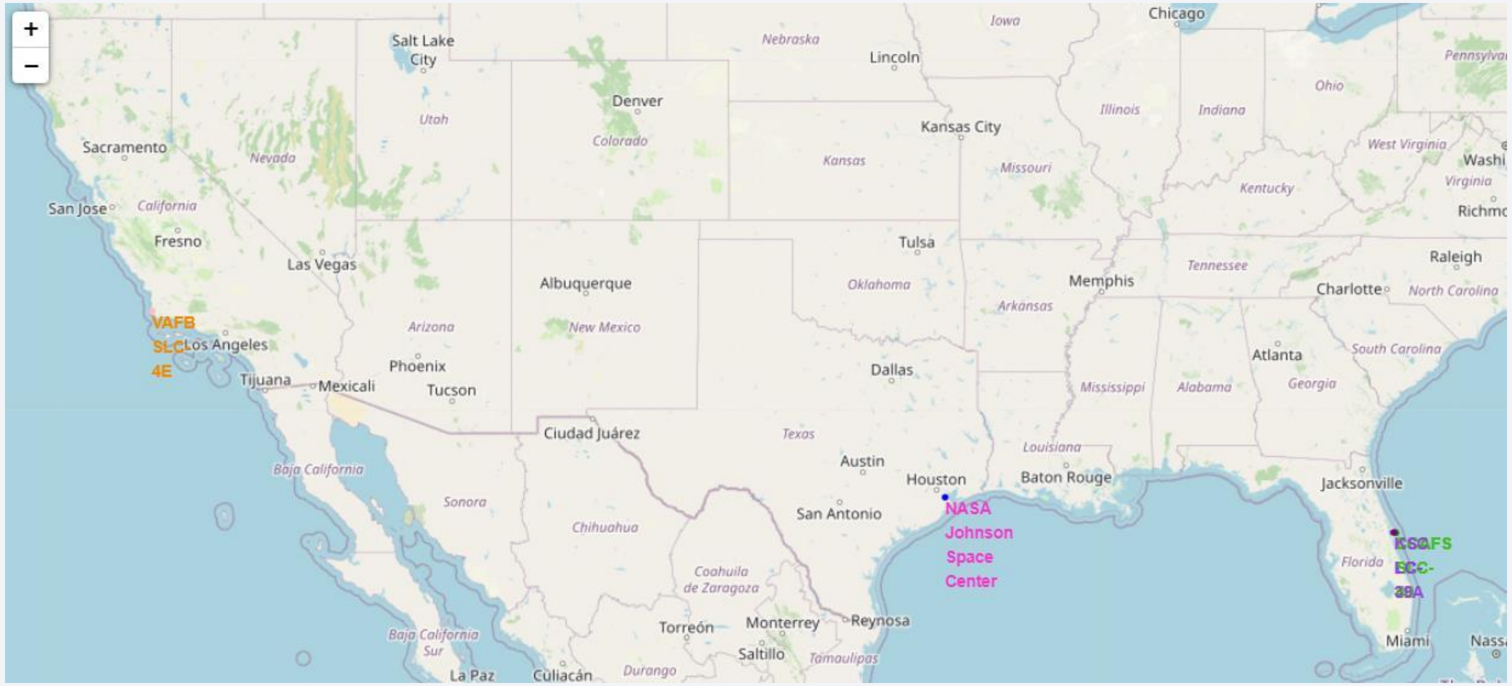
- List the names of boosters which have successfully landed on drone ship and had payload
- mass greater than 4000 but less than 6000:
- • F9 FT B1032.1
- • F9 B4 B1040.1
- • F9 B4 B1043.1
- • SQL QUERY: select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)'
- AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;

Section 4

Launch Sites Proximities Analysis

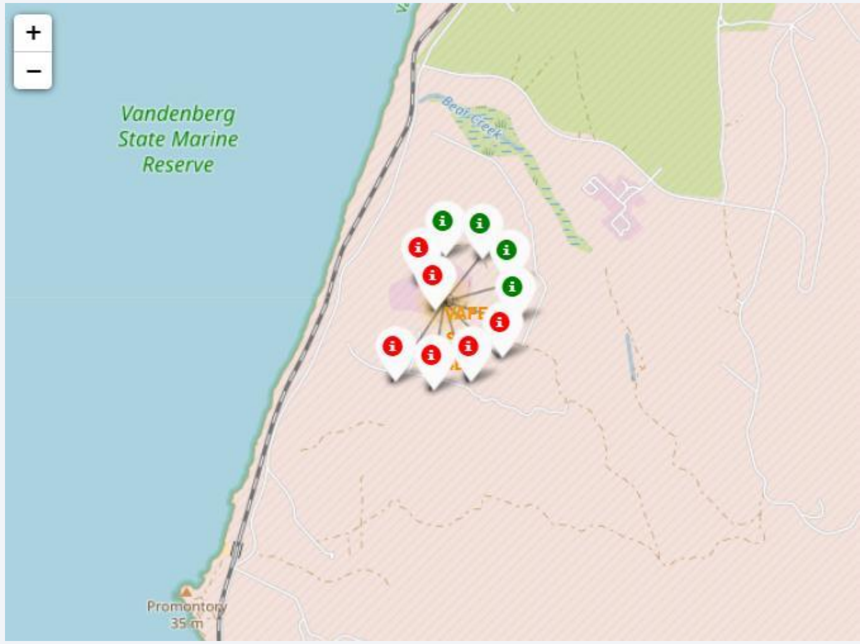
Launch Sites on the Map

- Geographic coordinates are just numbers that can not provide any intuitive insight into where the launch locations are, unless you are very good at geography. Therefore, creating a map we can visualize these places through their coordinates. On the map below, the launch sites are marked.

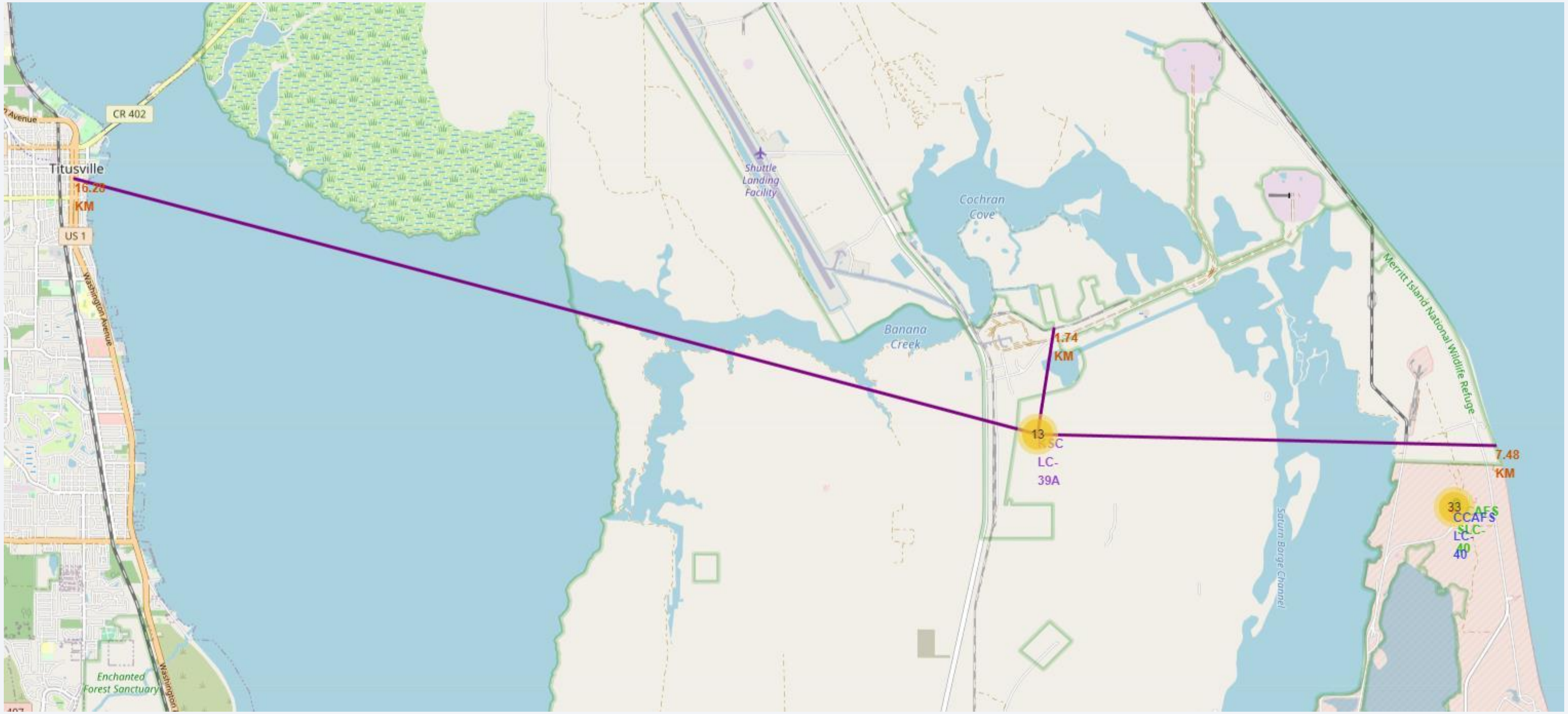


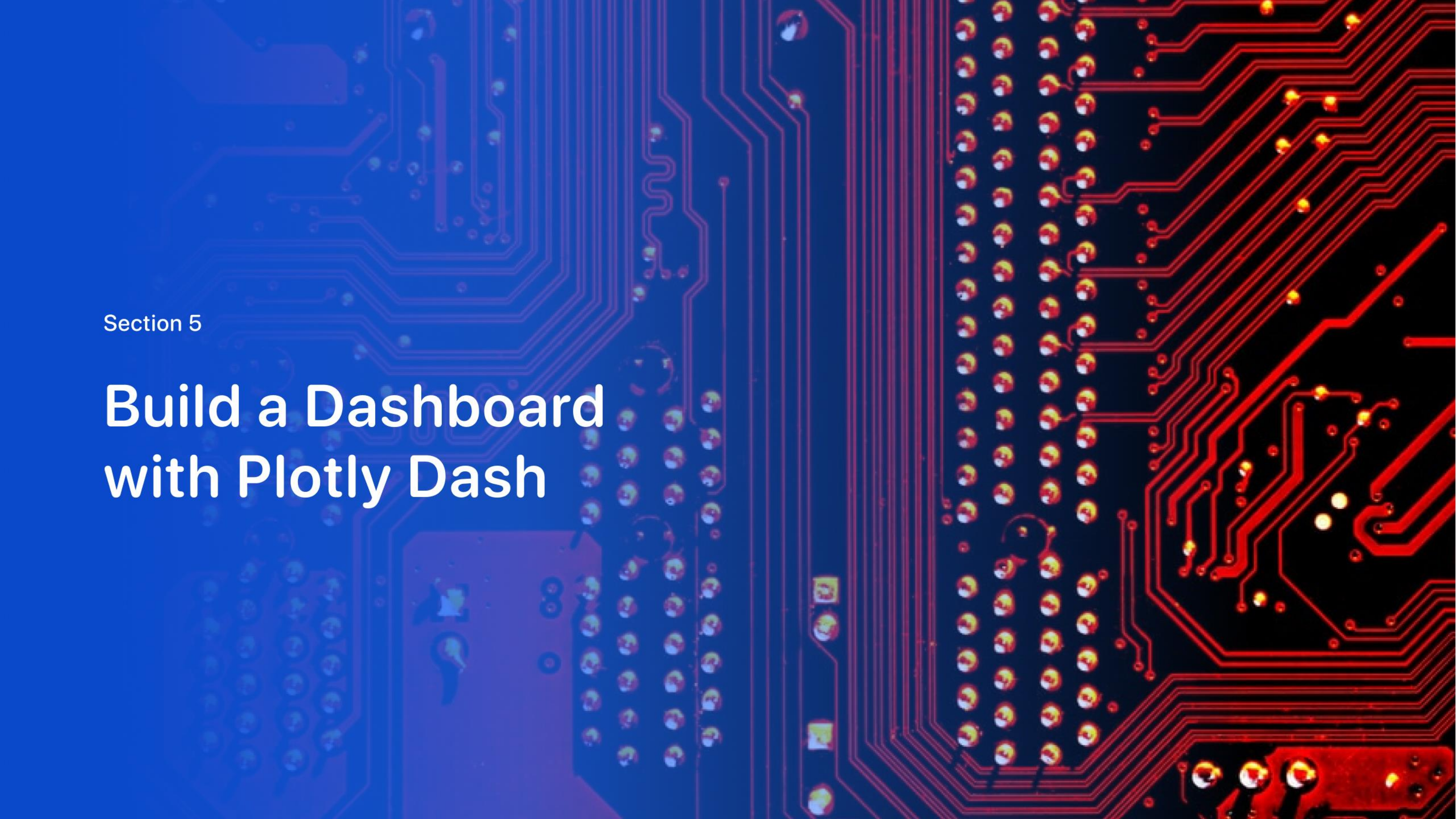
Launch clusters

- In this map, we can visualize the successful and unsuccessful launches through the color of pins in the cluster of each launch site. The green color is successful launches and the red ones are unsuccessful.



Measuring Distances

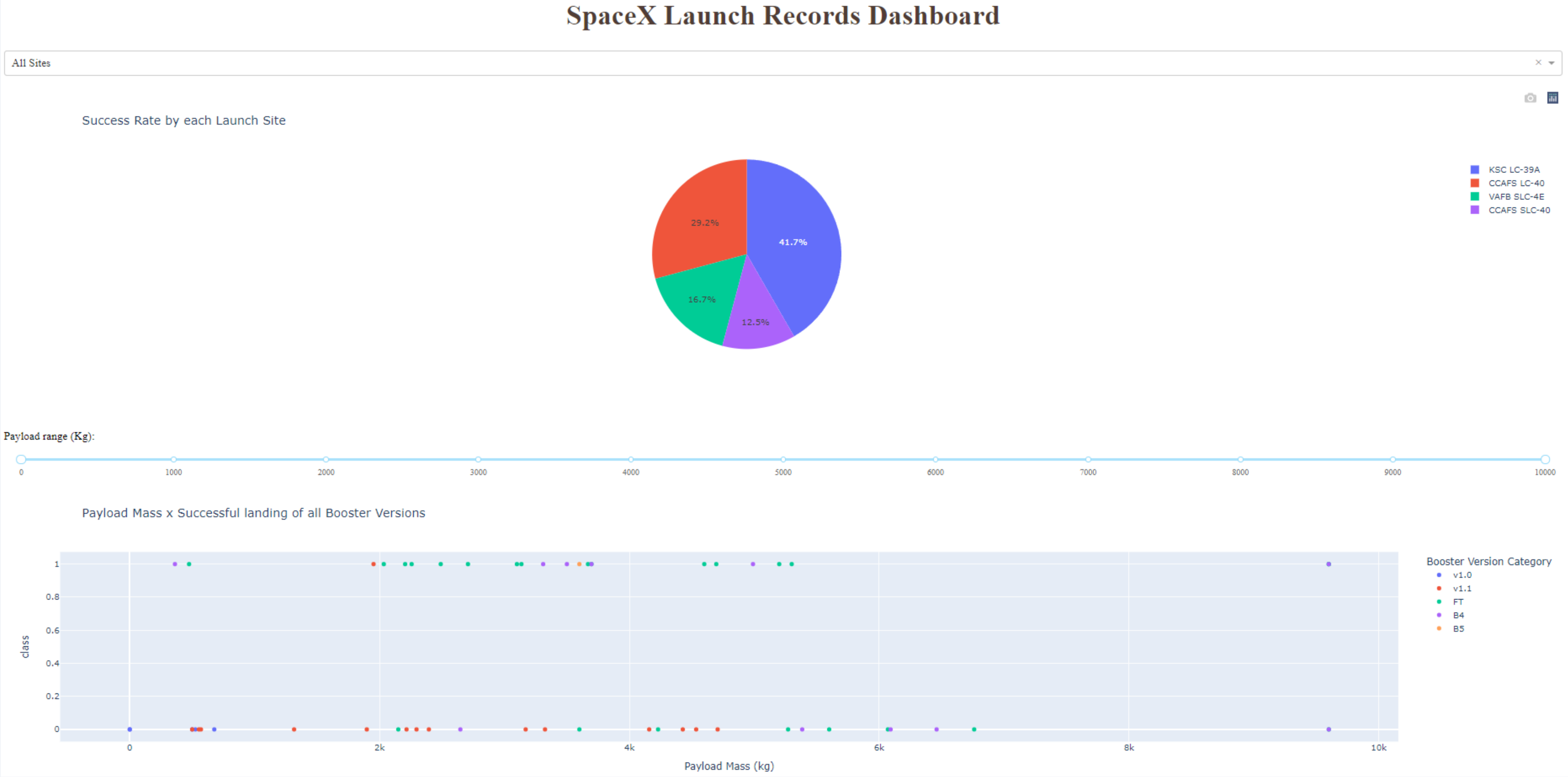




Section 5

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard – All Launch Sites

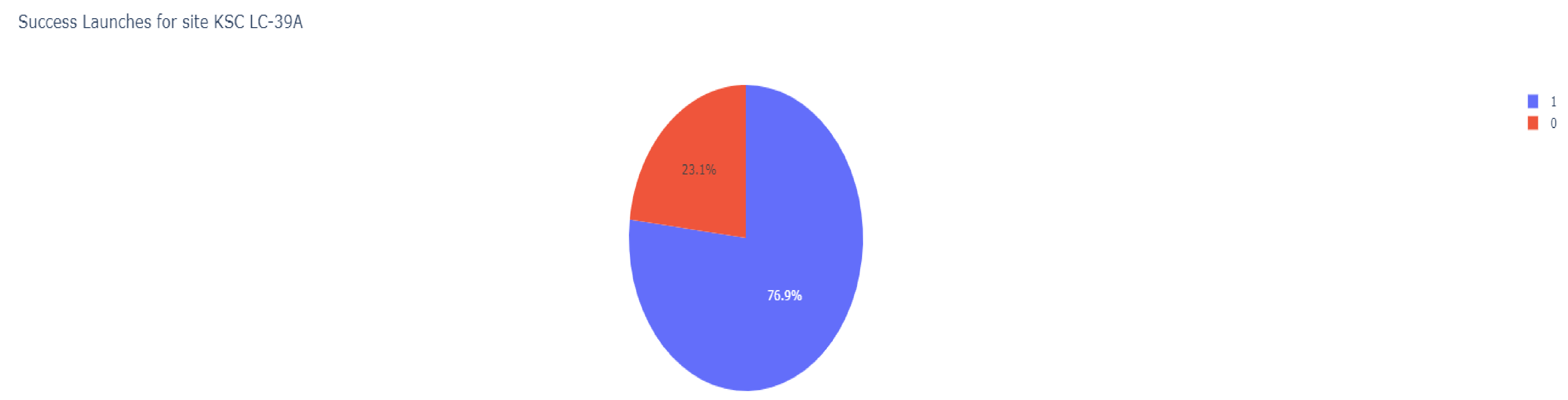


SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

SITE WITH HIGHEST LAUNCH SUCCESS RATIO



Success Launches for site KSC LC-39A

Section 6

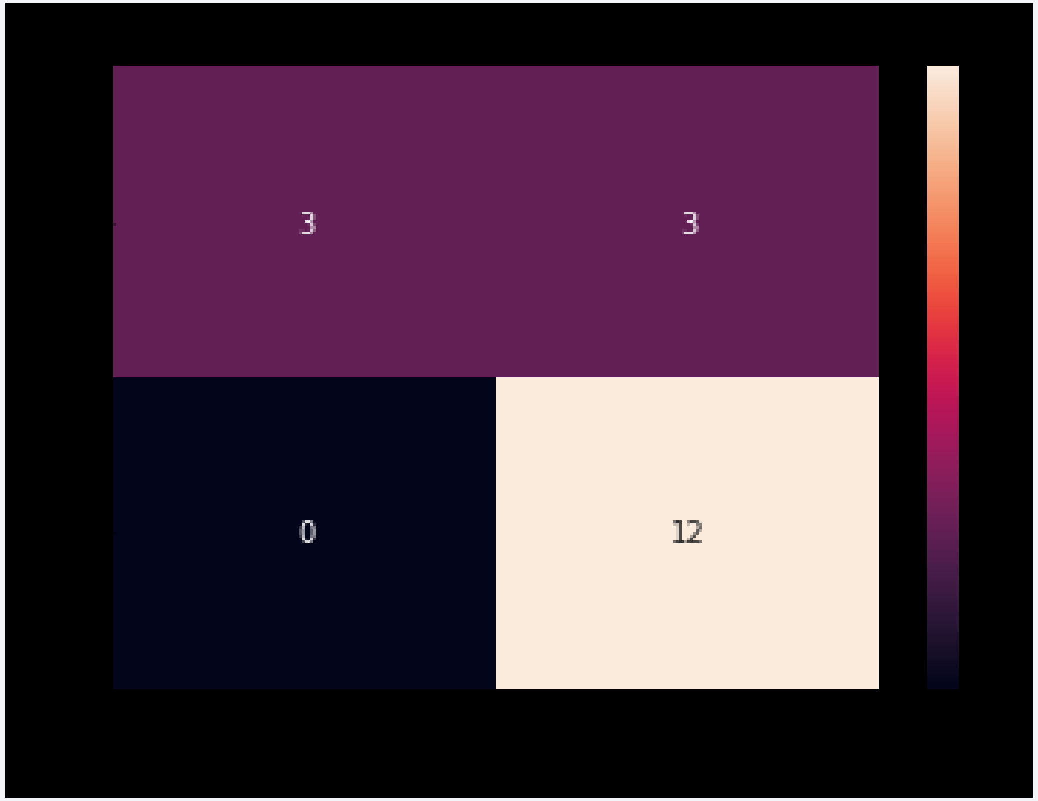
Predictive Analysis (Classification)

Classification Accuracy

- The best model based on the accuracy is a n Decision e Tree Classifier with a score of 0.884.

Confusion Matrix

- The Confusion Matrix of the Decision Tree Classifier
- True Positive : 12
- False Negative : 0
- True Negative : 3
- False Positive : 3
- The model is quite interesting as it predicts a lot of times the
- good labels, however 3 times it predicted the success of the
- mission and the mission failed. Reducing the amount of False
- Positive would be a good idea to avoid spending Millions and
- years of work.
- It could be done using Boosting or maybe look at a model with a
- lower accuracy but a better precision.



Conclusions

- It is important to follow the sequence of steps to achieve the main objective of the project, for example ETL, EDA, modeling and implementation;
- There are many ways to get a dataset, such as scraping websites, downloading from the Internet, querying a database or even doing it from scratch;
- Doing proper the data wrangling and exploratory data analysis steps, using visualizations such as charts, maps, or arranging a more complete visualization on a dashboard is essential to better understand the data;
- The four algorithms applied in this project to predict the successful launching of Falcon 9 scored well in this case.

Thank you!