IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mouhcine Ouchen
November 18, 2023

Outline

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

Executive Summary

- This works aims to study the SpaceX Falcon 9 first stage landing outcome in relation to different other variables such as launch site, payload mass, orbit type…
- To do so, SpaceX launches data was collected from different sources and then formatted and cleaned to be suitable for the analysis.
- The cleaned data was subject to an exploratory descriptive analysis was done to assess the correlation between different variables of the data and to retrieve more information.
- Later, we have done a predictive analysis to predict Landing outcome from the other variables using multiple trained classifiers.
- The descriptive analysis led to discover a link between different variables while the predictive analysis allowed us to confirm that Landing outcome can be predicted with a high (up to 89%) from the other variables.

Introduction

- This project a SpaceY project in the context of gathering information about SpaceX Falcon 9 to estimate the price of each launch.
- As the price of a launch depends on the possibility of reuse of the rocket's first stage, our objective is to predict if SpaceX will reuse the first stage: or more precisely if the first stage will land successfully or not (landing outcome).
- The landing success may depend on different variables such as payload mass, orbit type, launch site and multiple other parameters.
- The project aims to:
  - Check if there is any correlation between the Landing outcome and the different other parameters of the rocket launch.
  - Try to make a predictor with a high accuracy of the landing outcome based on the other launch parameters.

Section 1

Methodology

Methodology

Executive Summary

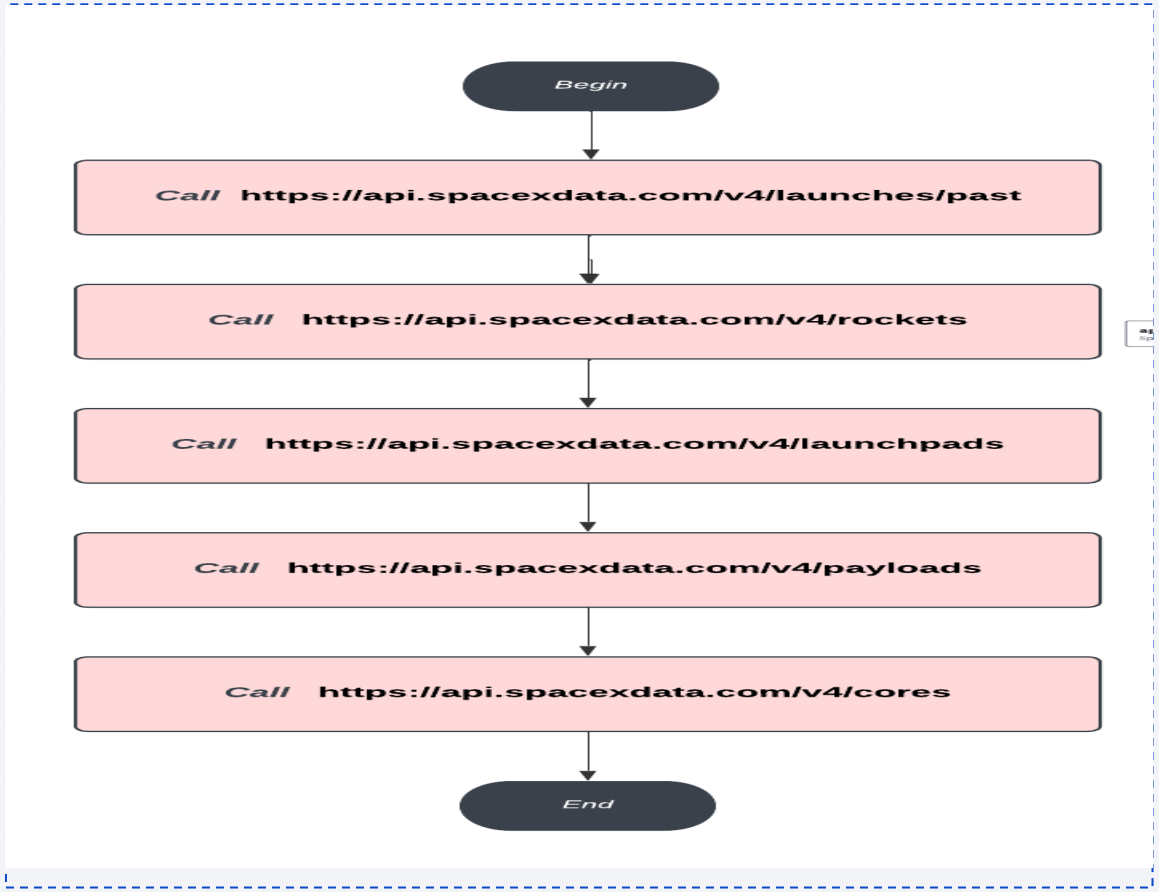
- Data collection methodology:
  - The data was collected from the SpaceX Rest API and from SpaceX Falcon F9 Wikipedia page.
- Perform data wrangling
  - The missing were dealt with and all the categorial variables were converted to numerical ones.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple models, such as logistic regression, were created and were trained using training data then tested using test data to get accuracy score to assess the best model.

Data Collection

- The data was collected using two main sources:
  - The first one is the SpaceX Rest API https://api.spacexdata.com/v4 which provides multiple informations about the SpaceX launches, the rockets used, the launchpads and a bunch of other informations about SpaceX. Only data related to Falcon 9 was kept.
  - The second one is a Wikipedia page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches providing Falcon 9 launches history.

Data Collection – SpaceX API

- We start by collecting data using https://api.spacexdata.com/v4/launches/past. It includes data about past launches made by SpaceX. However, some fields are just codes, so we need to retrieve real values from by calls to:
  - **rockets** to get booster names.
  - **payloads** to get the payloads mass and the orbit.
  - **launchpads** to get the names of the launch sites with their latitudes and longitudes.
  - **cores** to get type of the landing, its outcome and other related data.
- The data is then filtered to keep only Falcon 9 data.
- For more details, see Data Collection Notebook on Github.



Data Collection - Scraping

- Data is collected using webscraping library BeautifulSoup from the List_of_Falcon_9_and_Falcon_Heavy_launches Wikipedia page.
- The first phase is to retrieve the all-page content.
- The second step is to extract the tables columns names and their corresponding data.
- The last phase is to save the retrieved data into a data frame to be ready to process.
- For more details, check the Web Scraping Notebook on Github.

Data Wrangling

- In the collected data, the target variable, which is the landing outcome, is a categorial variable with values in the form: **False Ocean** which means unsuccessful landing on the ocean, or **True RTLS** which means successful landing to a ground pad, and so on.
- So, from this variable, we create a new variable **Class** which will be the new target. The new variable is a numerical variable with 2 values: 0 for unsuccessful, and 1 for successful landing.
- To see more, check the Data Wrangling Notebook on Github.

EDA with Data Visualization

- Multiple plots were generated to assess the relationship between different variables in the data set:
  - **Launch Site by Flight Number (differentiated by success or fail of the landing)**
  - **Launch Site by Payload (differentiated by success or fail of the landing)**
  - **Success rate by Orbit type**
  - **Orbit type by Flight Number (differentiated by success or fail of the landing)**
  - **Orbit type by Payload (differentiated by success or fail of the landing)**
  - **Launch success by Year**

- Check the Visualization Notebook for more details.

EDA with SQL

- Multiple queries were performed here to obtain:
  - **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)**
  - **the average payload mass carried by booster version F9 v1.1**
  - **the date when the first succesful landing outcome in ground pad was acheived**
  - **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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015**
  - **the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order**
- Reference the Descriptive Analysis SQL Note book on Github.

Build an Interactive Map with Folium

- With folium we have generated multiple maps for different purposes:
  - The first one contains the launching sites locations marked by red circles.
  - The second one marks for each site the successful and unsuccessful landings. This allows us to get the landing success rate of each site directly on the map.
  - The last one adds a line with a calculated distance between one site and the some proximity like coastline, railway…
- Details are on Site Location Notebook

Build a Dashboard with Plotly Dash

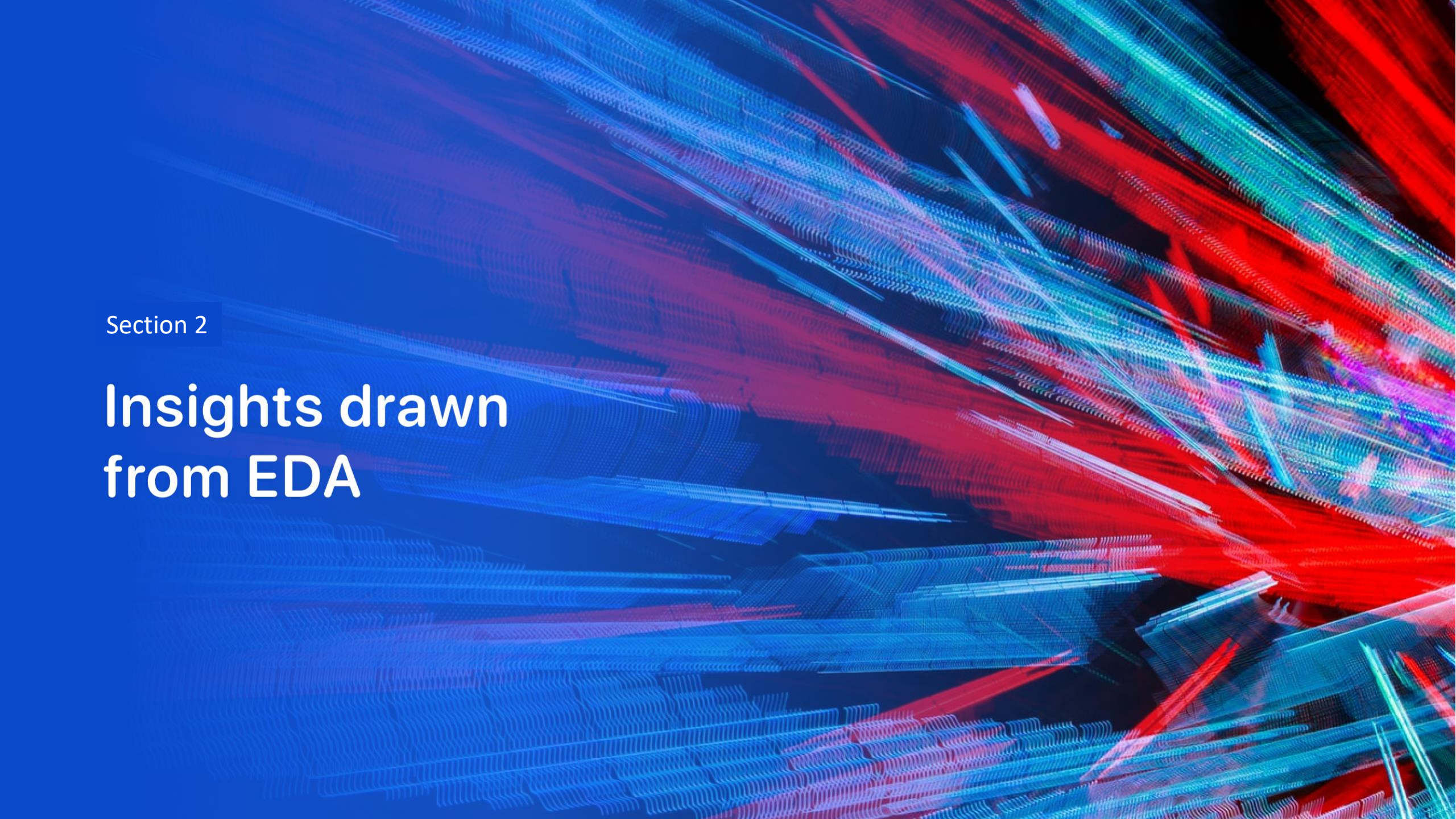
- In this part, we have created a dashboard using Plotly Dash. The dashboard allowed us to generate pie charts for:
  - All sites: the chart represents the participation of each site in the total successful landings.
  - Each site: the chart shows the number of successful and unsuccessful landings for this site.
- The dashboard contains also interactive scatter charts representing Success/failure vs Payload mass for different booster versions. The payload mass min and max value in the chart are controlled by a Slider to visualize the relation for different mass ranges. These scatter plots are also available by site or globally.
- The GitHub URL of this part is Dashboard Python File .

Predictive Analysis (Classification)

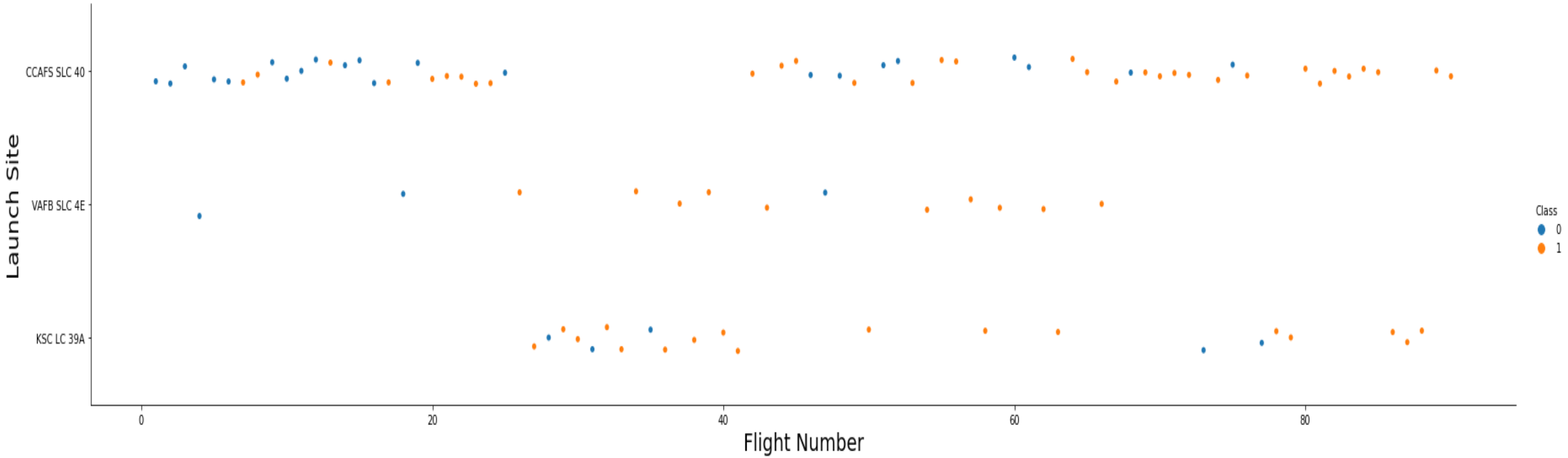
- At the beginning, the data was split into 2 parts, the training part and the test part which accounts for 20% of the data.
- We have then deployed 4 types of classifiers to predict the landing outcome based on different features such as payload mass, launch site, booster version…. The types of classifiers are: Logistic Regression, SVM, Decision Tree and K-nearest neighbors.
- For each type of classifier, we created a GridSearch with different hyperparameters combinations. The training process takes the training data features and target and trains the classifier with the different combinations to get the best one.
- For each type, the classifier with the best parameters is tested on test data and evaluated to assess its accuracy.
- For more check Predictive Analysis Notebook on Github

Results

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

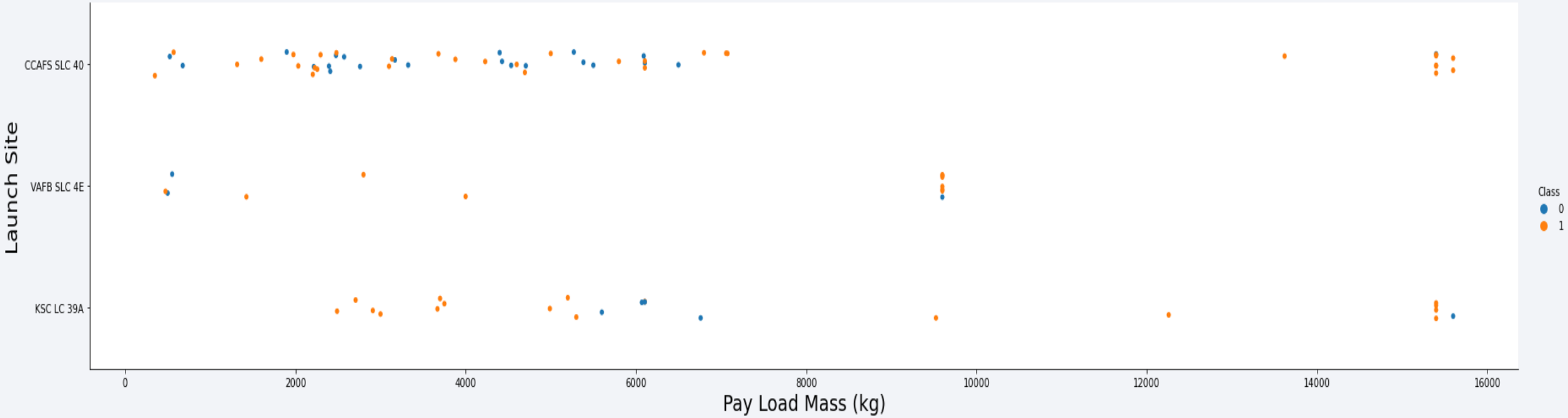

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- As we can see from the chart, most rocket launches were at **CCAFS SLC 40** site. This site was used for the early missions which have high failure rate but also for the last missions which are almost all successful.

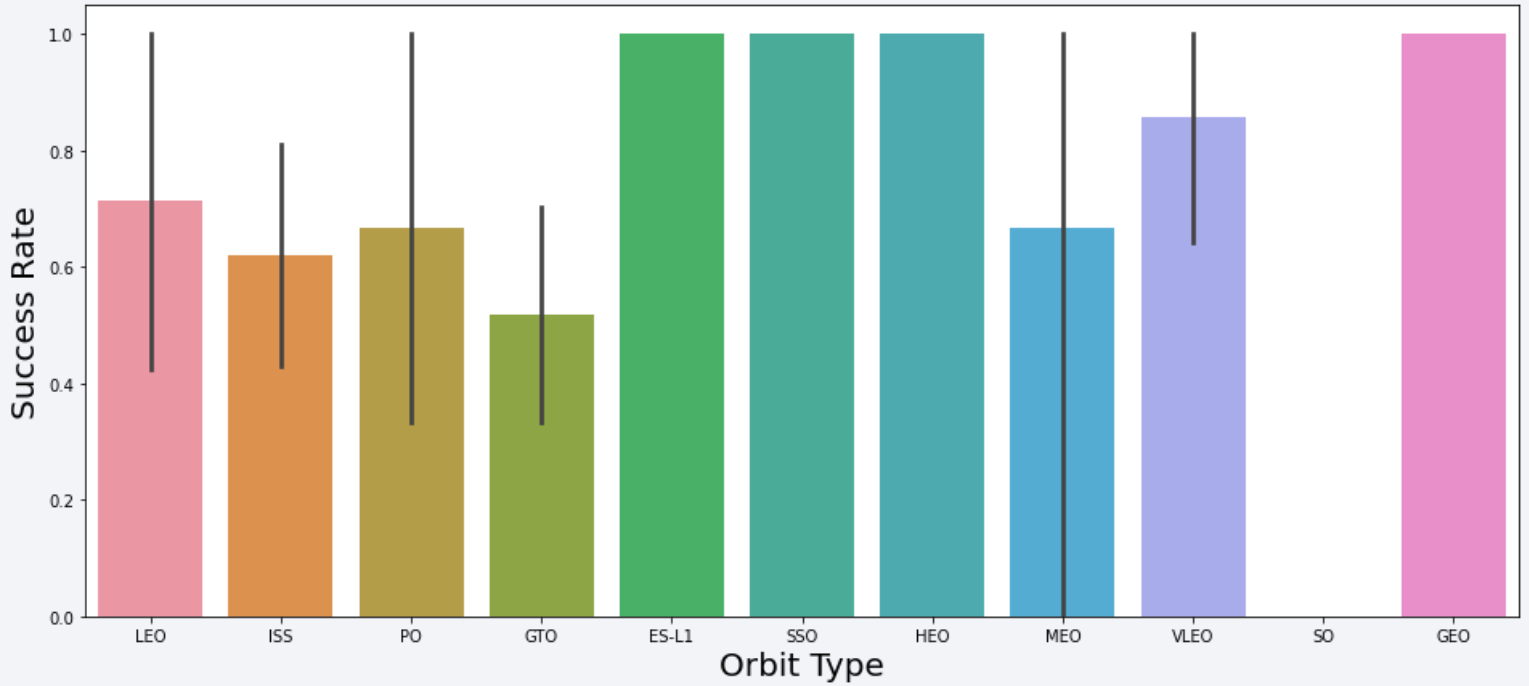
Payload vs. Launch Site



- This scatter point chart tells us that the VAFB-SLC site is not used for heavy payload rockets launch (mass greater than 10000). The other sites, in the other hand, are used for light and heavy payloads indifferently.

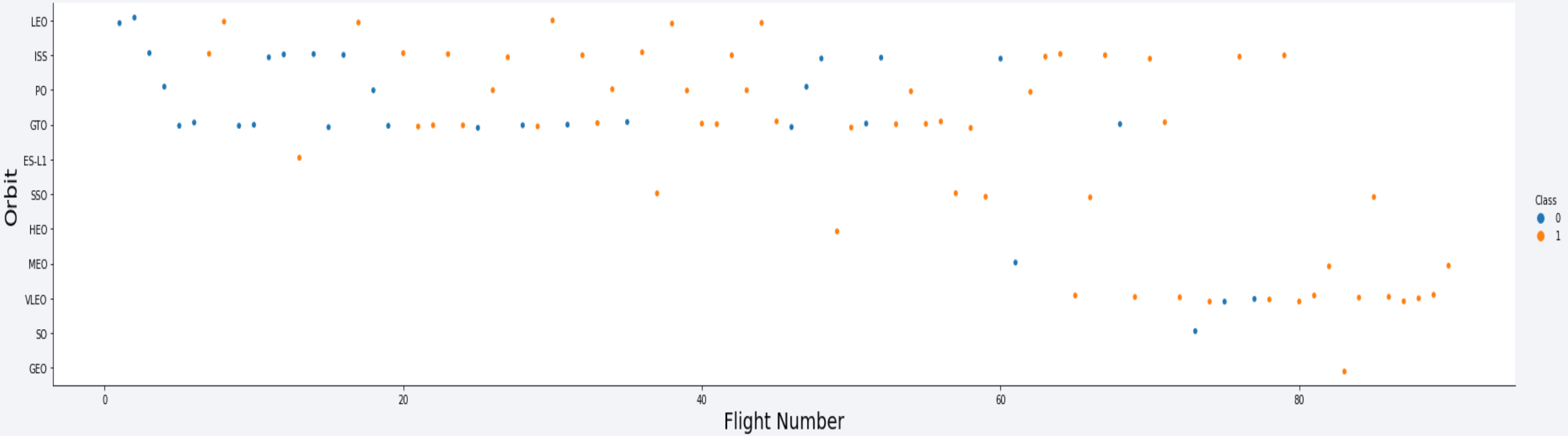
Success Rate vs. Orbit Type

- We can see here that Geo stationary orbit, Highly elliptical orbit, Sun-synchronous orbit and L1 lagrange point launches are 100% successful.



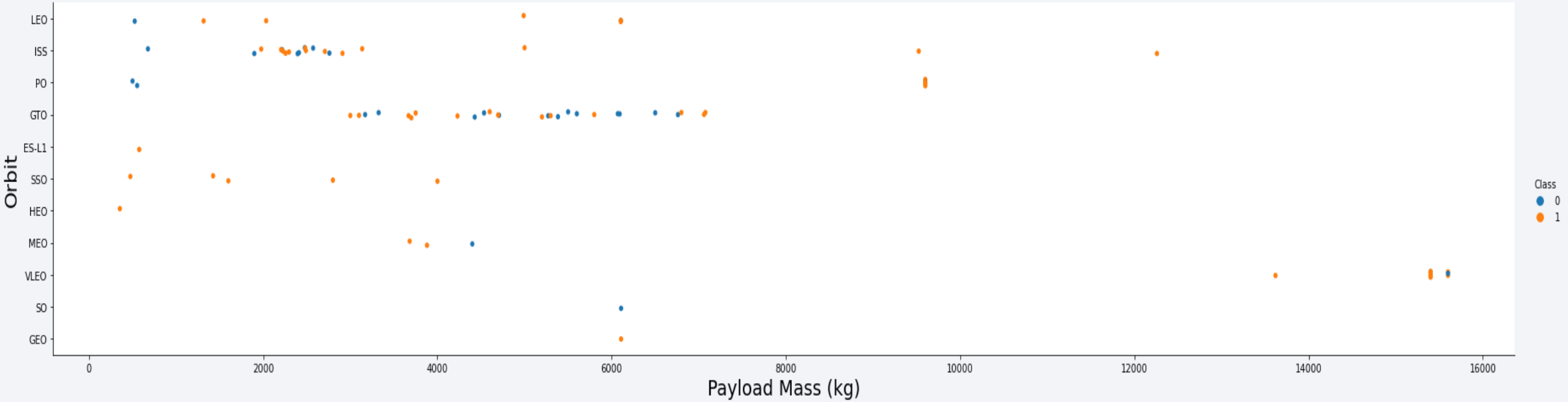
- Low altitude orbits have a lower success rate.

Flight Number vs. Orbit Type



- We can notice that, for Low earth orbit, the first launches failed to recover the first step but with time the recovery became totally successful. We cannot say the same thing for the other orbits.

Payload vs. Orbit Type

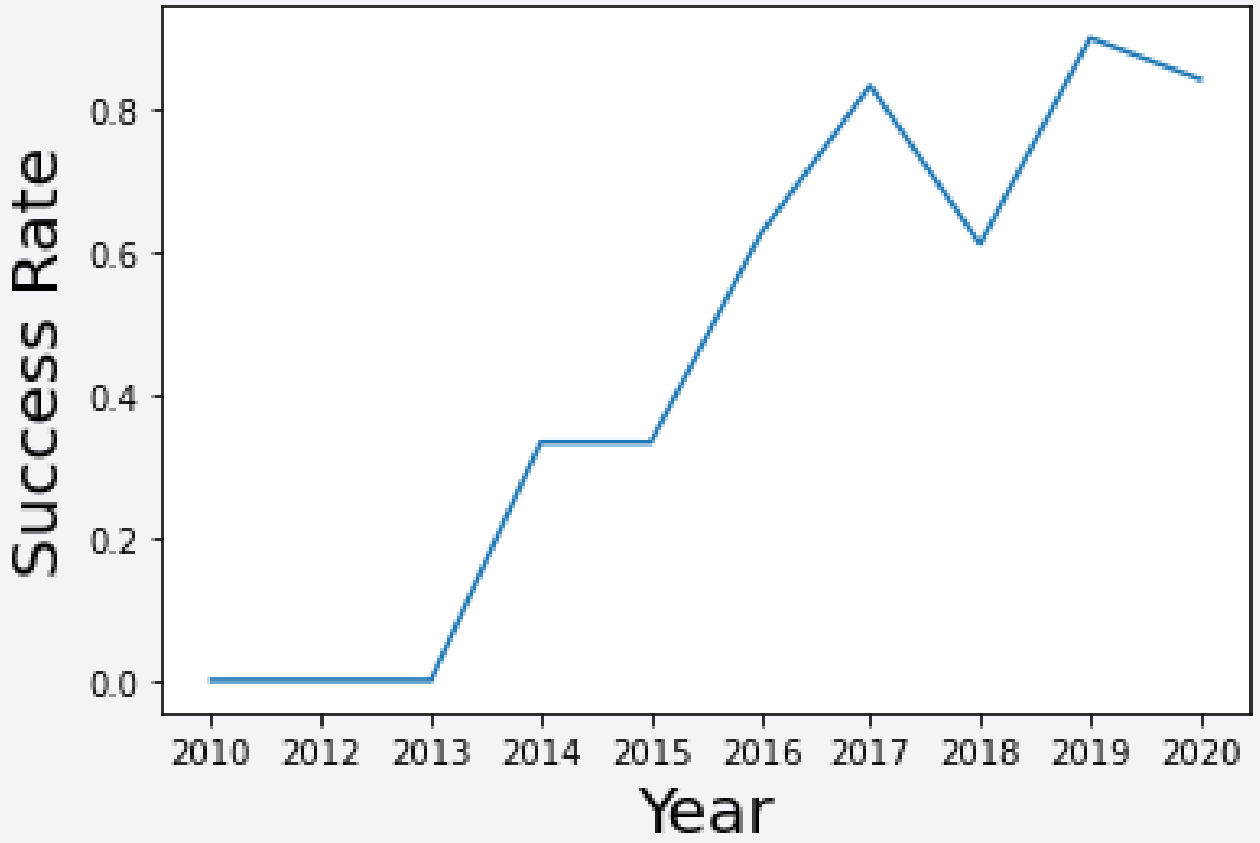


- With heavy payloads, the successful landing are more for Polar, LEO and ISS orbits.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- The landing success rate has been increasing during time. From 2010 to 2013 it was near to 0. In 2019 it was greater than 80%.

- There were some downturns, but the trend is ascending.



All Launch Site Names

- This table shows the different launch sites retrieved from **SPACEXTABLE.** We got 4 distinct sites.

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

Launch Site Names Begin with 'CCA'

- Here we have 5 launch missions where launch sites begin with `CCA`.
- We can see that those missions date from 2010 to 2013 and all of them fail to land evn if the missions were successful.

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

- The total payload carried by boosters from NASA is **45596 kg**.

- Nasa has a lot of space missions and need to carry a lot of material to space and to the international space station.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **14642** kg.

- This means that F9 v1.1 boosters can carry a heavy payload without any problems.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is **2015-12-22**.

- It took a long time for SpaceX to achieve a successful landing on ground pad as the first tries were in 2010.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The table below shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Only 4 boosters have made this exploit as drone ship landing is not so easy to achieve

Booster Name
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The table below shows the total number of successful and failure mission outcomes.
- As we can see, almost all the missions were successful except one.

Mission outcome	Count
Failure	1
Success	100

Boosters Carried Maximum Payload

- The table lists the names of the booster which have carried the maximum payload mass
- A lot of boosters have carried a max payload mass which SpaceX is developing multiple performant boosters.

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

- Below the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Two missions failed to land in drone ship, one in January and the other in April, both were launched at **CCAFS LC-40**.

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The table shows the ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- The missions with no landing attempt are the most frequent. This means that, during its early age, SpaceX was not trying to land the first step for a lot of missions.

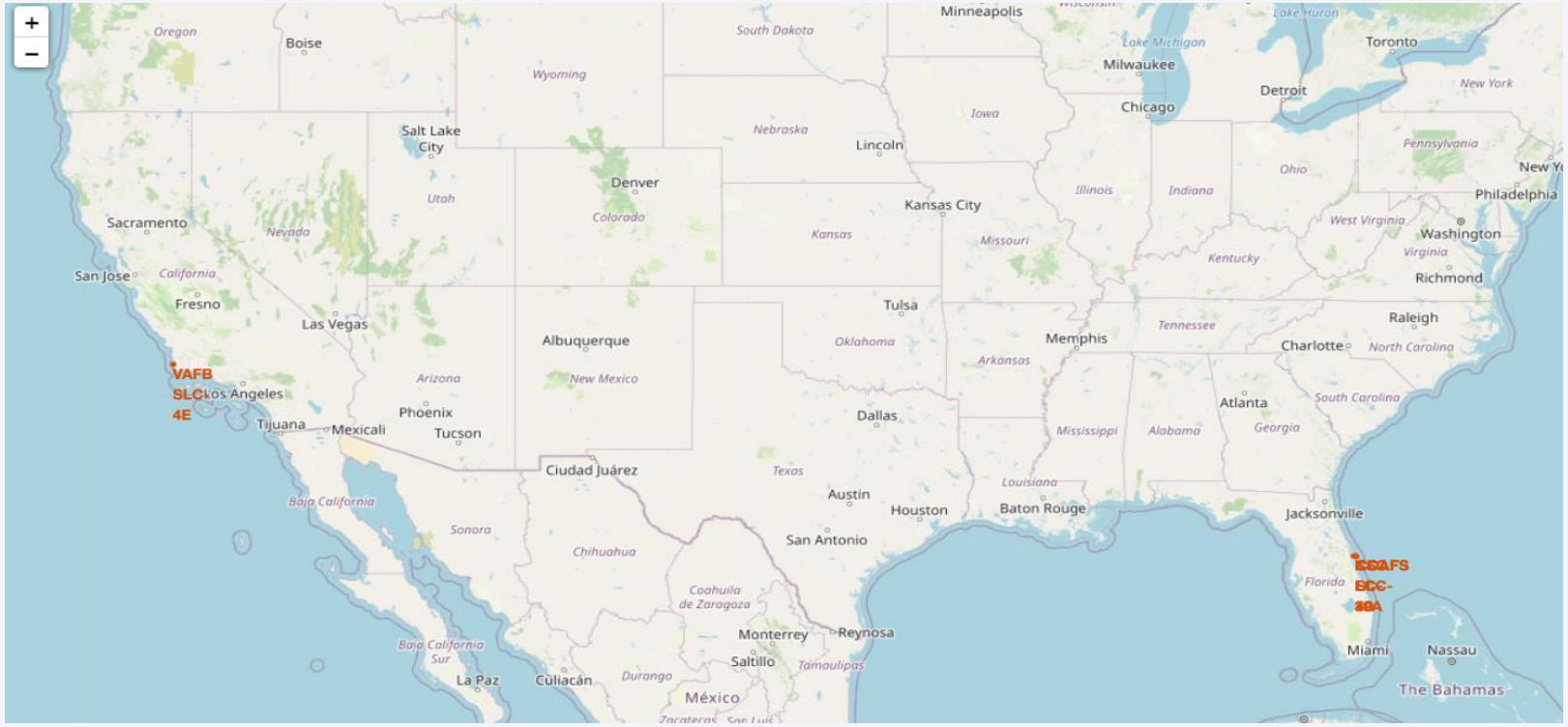
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

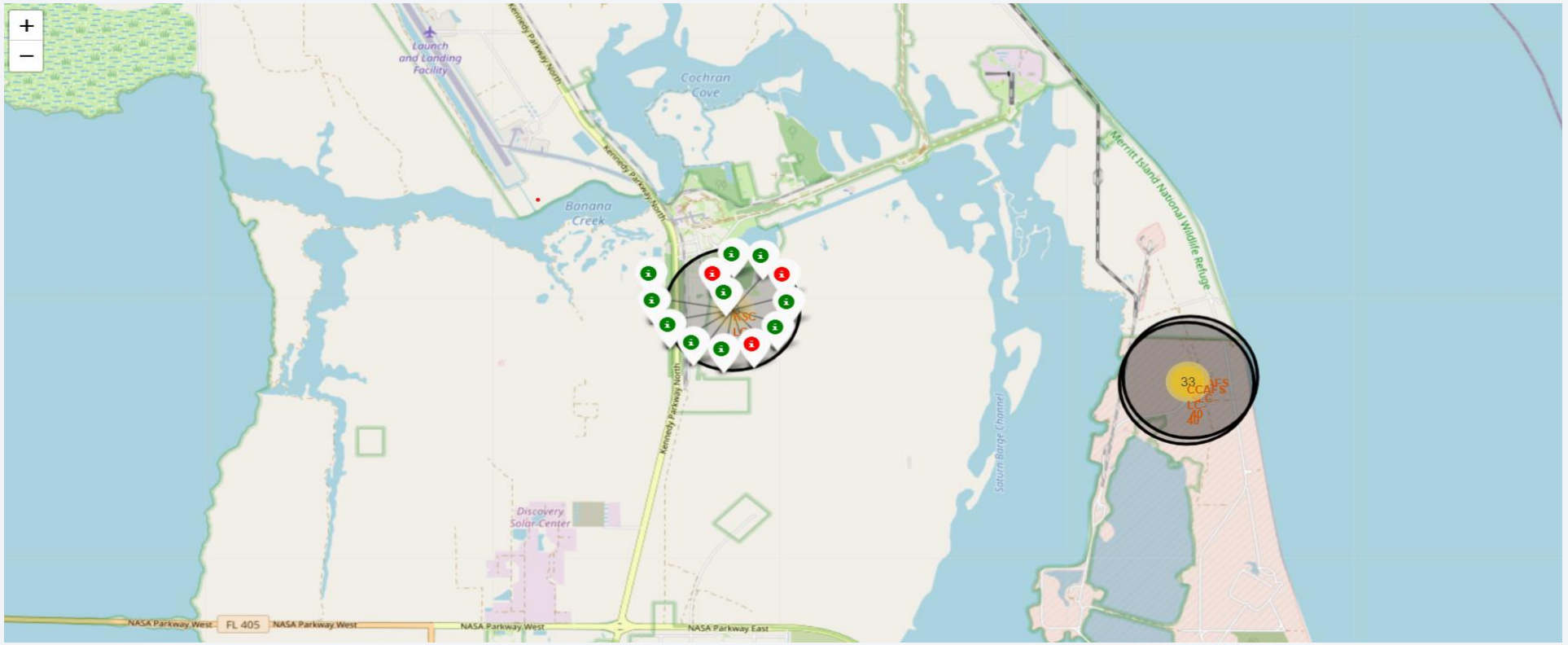
Launch sites location in the map

- The map below shows the location of SpaceX launch sites marked by red circles.
- We can notice that all sites are close to the sea and are the most possible close to the equator to exploit the earth rotation speed efficiently.



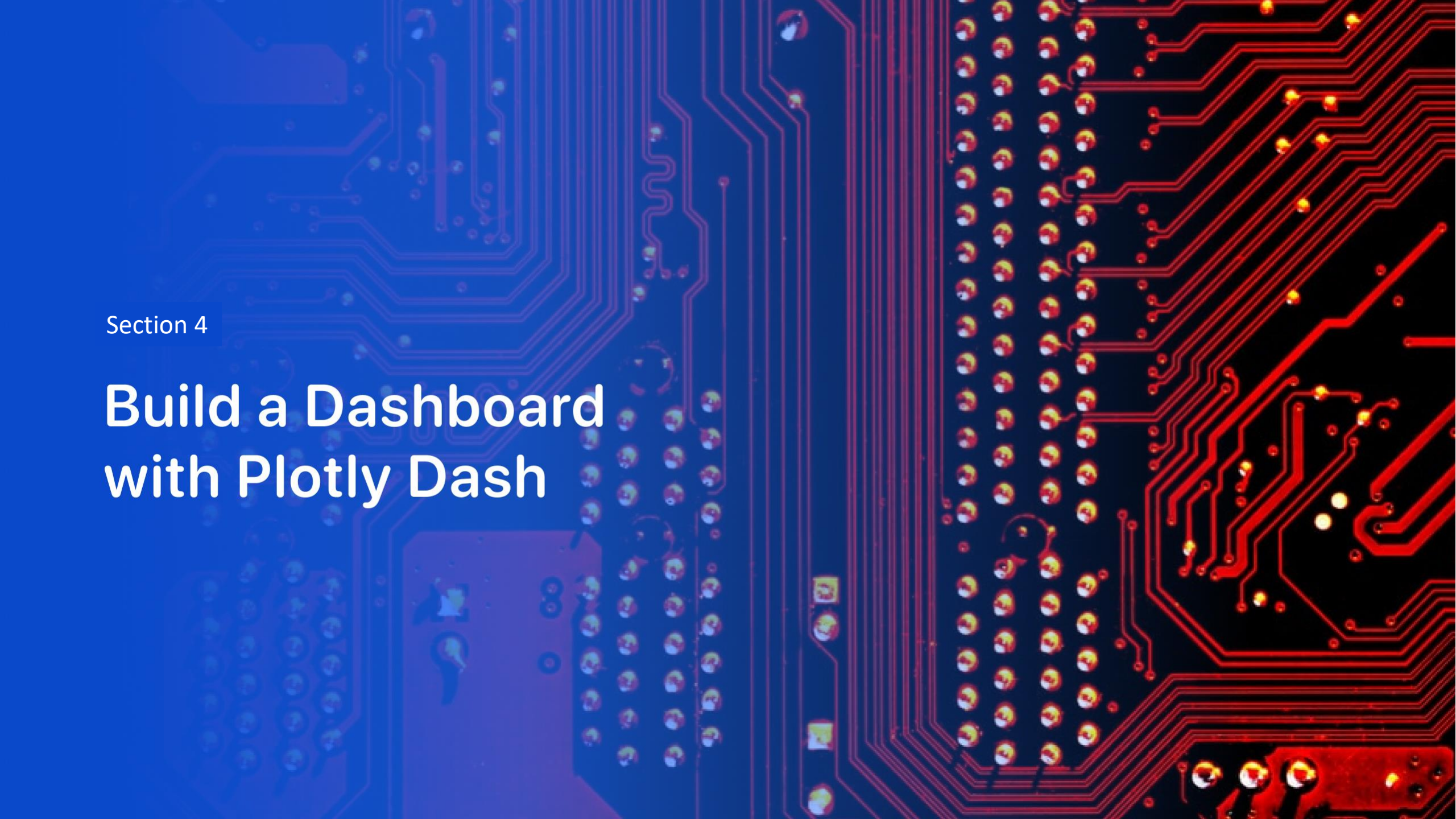
Successful and failed at KSC LC-39A site

- The map below shows the successfull (green) and failed (red) landings at the **KSC LC-39A** site.
- From the 13 launches made at this site, we can see that 10 succeded to land. This make a landing success rate of 77%.



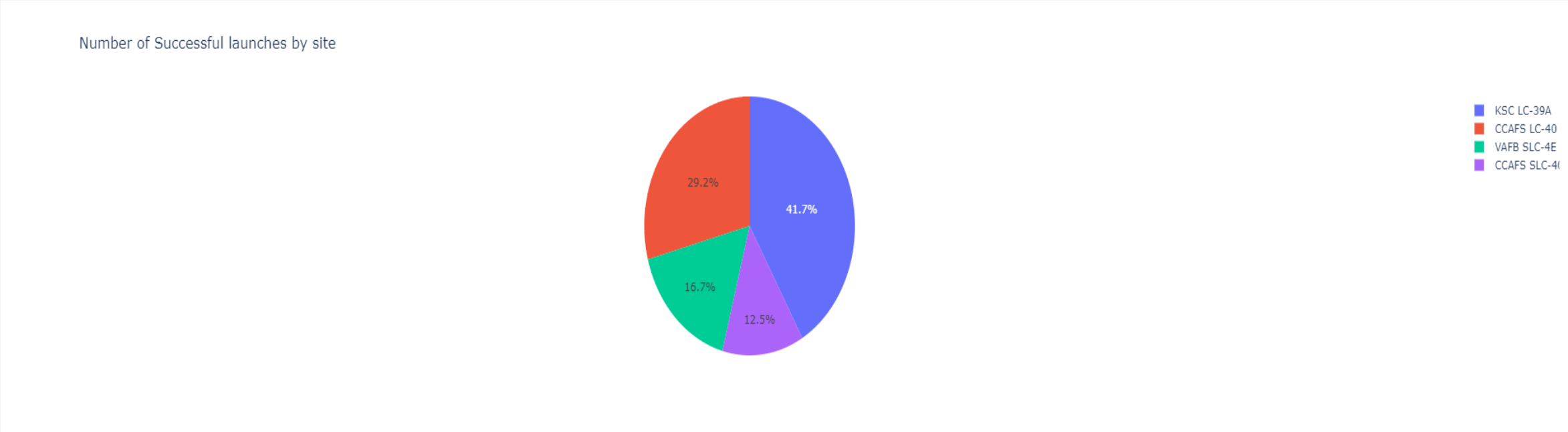
KSC LC-39A Site proximity to coastline

- The map below shows the distance between the site KSC LC-39A and coastline.
- This distance is approximately 6.29 Km.


Section 4
Build a Dashboard with Plotly Dash

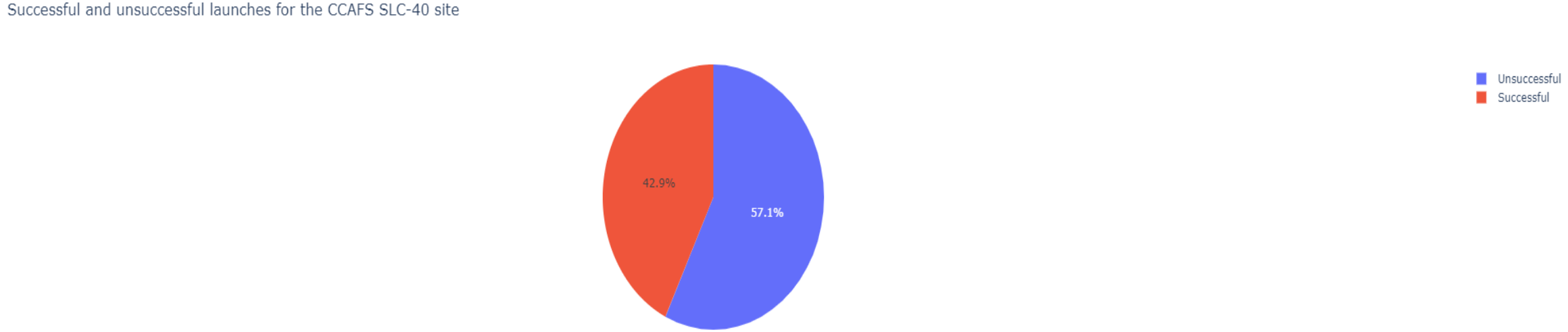
Distribution of successful landings by launch site

- The chart below shows the share of each launch site in total successful landings.
- We can see that the two sites KSC LC-39A and CCAFS LC-40 participate at more than 70% of successful landings.



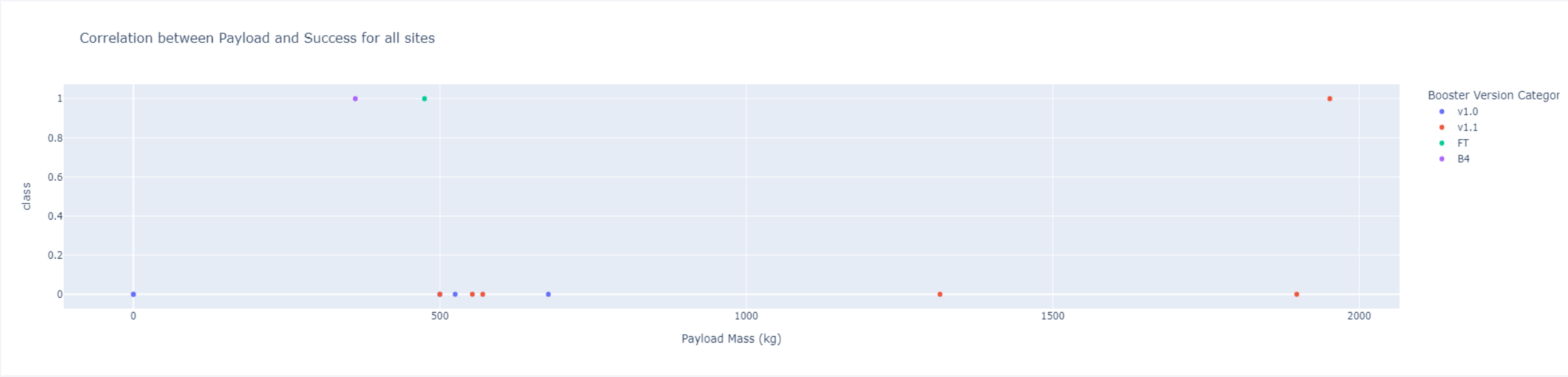
The highest landing success rate

- The pie chart below represents the success and fail rate in the most successful site: CCAFS SLC-40.
- The site shows a landing success rate of 42.9% which is higher than all the other sites.



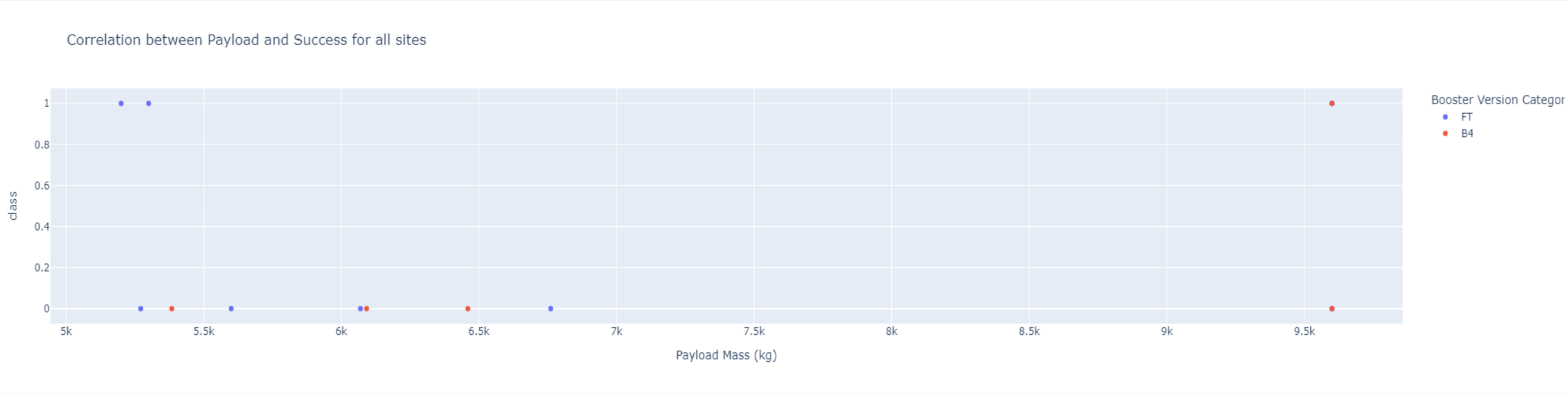
Landing success rate by payload mass and booster version

- The following charts represent the relation between the landing success rate, the payload mass and the booster version.
- As shown below, for light payloads (mass less than 2000 kg), the B4 and the FT boosters show the best landing success rate of 100%.



Landing success rate by payload mass and booster version

- While, for heaviest payloads (greater 5000 kg), only the FT booster achieves a landing success rate of 33% compared to 20% for the B4 booster. The other types do not carry heavy loads.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The table provides the accuracy of the different models used to predict the landing outcome based on the different features such as payload mass, orbit type…

- We can see that Decision tree classifier leaded to the best score of 88.89%

Model	Test Score
Logistic regression	0.833333
SVM	0.833333
Decision tree	0.888889
KNN	0.611111

Confusion Matrix

- The image below shows the confusion matrix of the decision tree classifier on test data.
- We can see that the classifier predicted correctly all the successful landings. It only had 2 false positives by predicting 2 not landed missions as landed.

Conclusions

- We can see that there are some relation between the multiple launch parameters and the landing outcome and between each others.
- We can assess that landing outcome can be predicted with a high accuracy using classifiers such as Decision Tree Classifier.
- Therefore, Launch price can be estimated based on the parameters of the launch itself.

Thank you!