IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Clement SOUFFEZ
18/03/2023

Outline

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

Executive Summary

- **Summary of methodologies:**
  - Data collection through web scraping and SpaceX API scraping.
  - Data wrangling ,EDA. Also data visualization and interactive visual analytics.
  - Use of four Machine Learning algorithms to make predictions.

- **Summary of all results:**
  - Data of great quality were collected from the SpaceX API and their website.
  - EDA and data wrangling helped us to identify the best features that help predict successful launch outcomes.
  - The best classification model was determined, the one that predicts the features to be taken into account for successful launches.

Introduction

- **Background and context**
- Our company SpaceY is competing against SpaceX and in this project we want to predict if the Falcon9 first stage will land successfully.
- **Problems we want to find answers:**
- We will have to find the best launch sites that are likely to allow a successful landing.
- Find the price of each launch by gathering data about SpaceX: We will train a machine learning model and use public information and see if SpaceX will reuse first stage.
- Will SpaceX reuse the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data is collected through web scraping.
  - Data collection through SpaceX API.
- Perform data wrangling:
  - Find and dealing with Missing values.
  - Find Categorical and numerical columns.
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Influence of Flight number and Payload on launch outcome.
  - Out plotting of the Flight number against the payload mass.
  - Yearly trend of the launch outcome.

Methodology

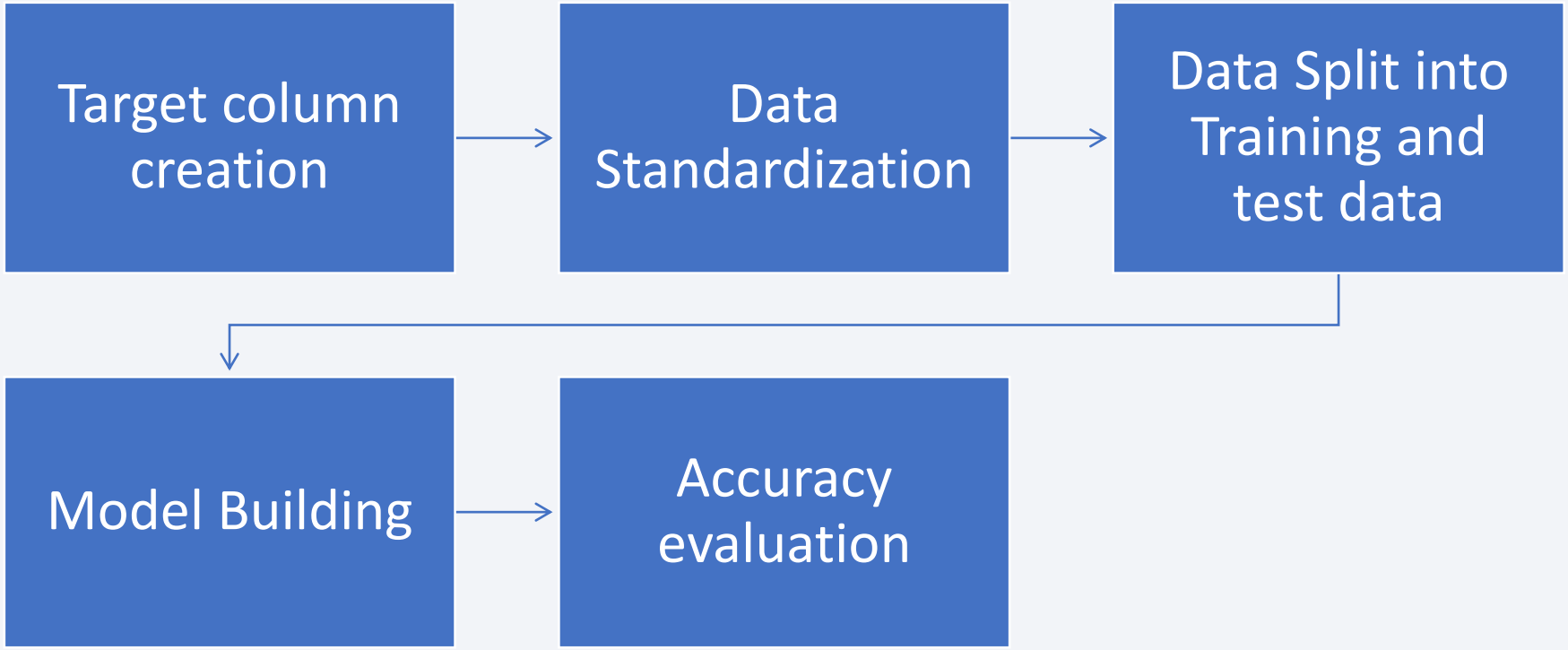
Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - How to build, tune, evaluate classification models

Methodology

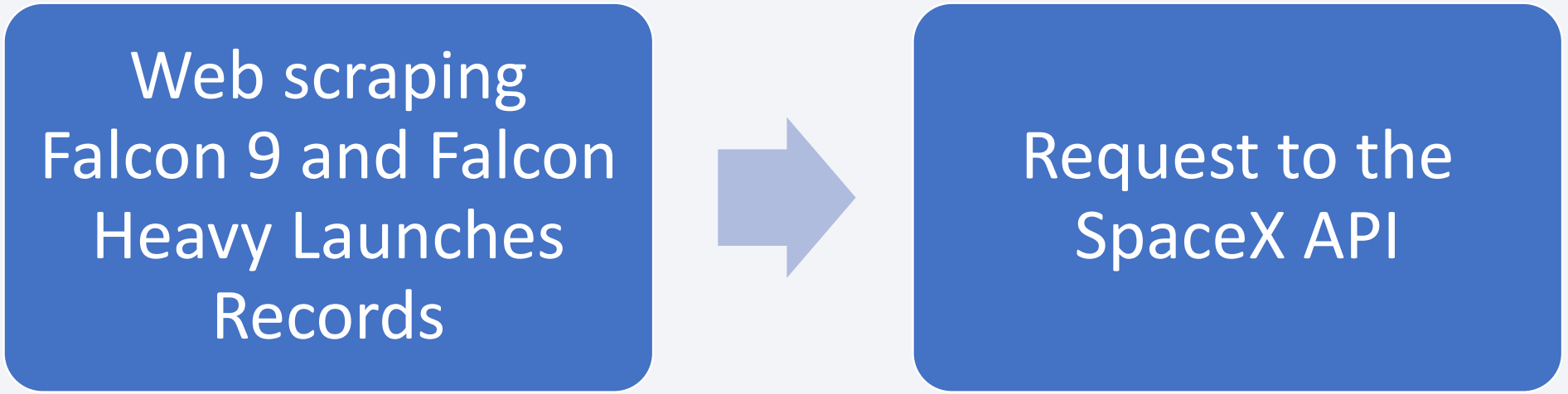
Executive Summary

- **Process of performing the predictive analysis using classification models:**



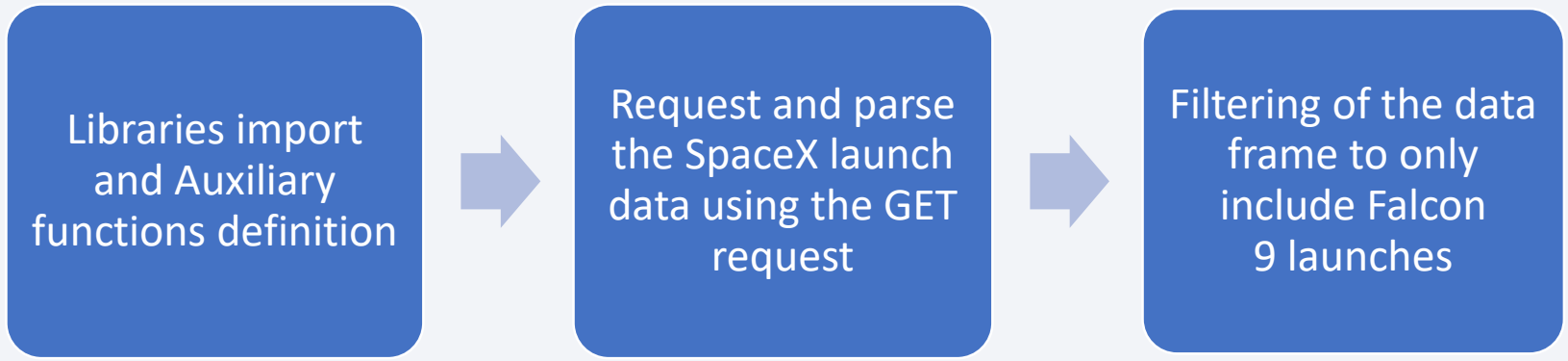
Data Collection

- **Process of collecting data**:



Data Collection - SpaceX API

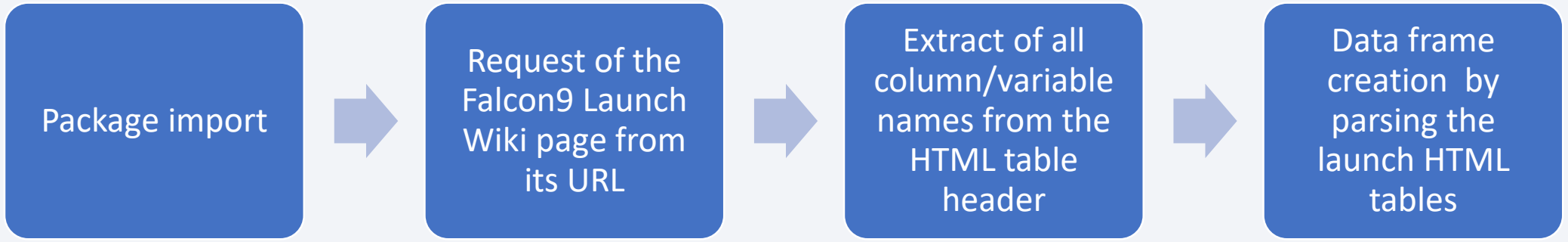
- **Process of collecting data through SpaceX API:**



Source:
https://rb.gy/qmq3dj

Data Collection - Scraping

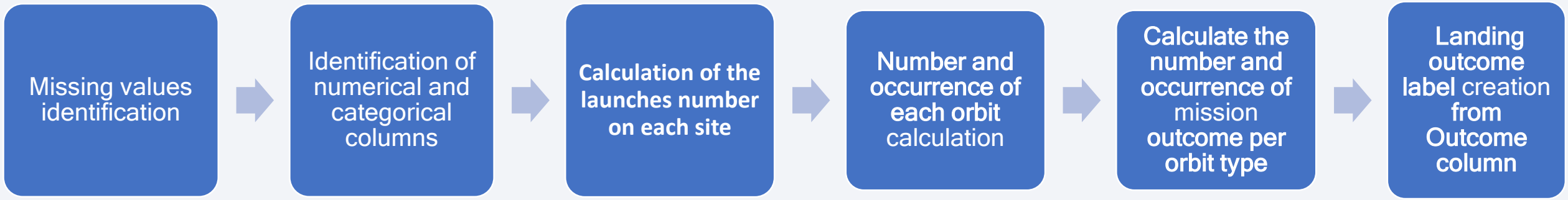
- **Process of collecting data through web scraping**:



Source :https://rb.gy/h4dkux

Data Wrangling

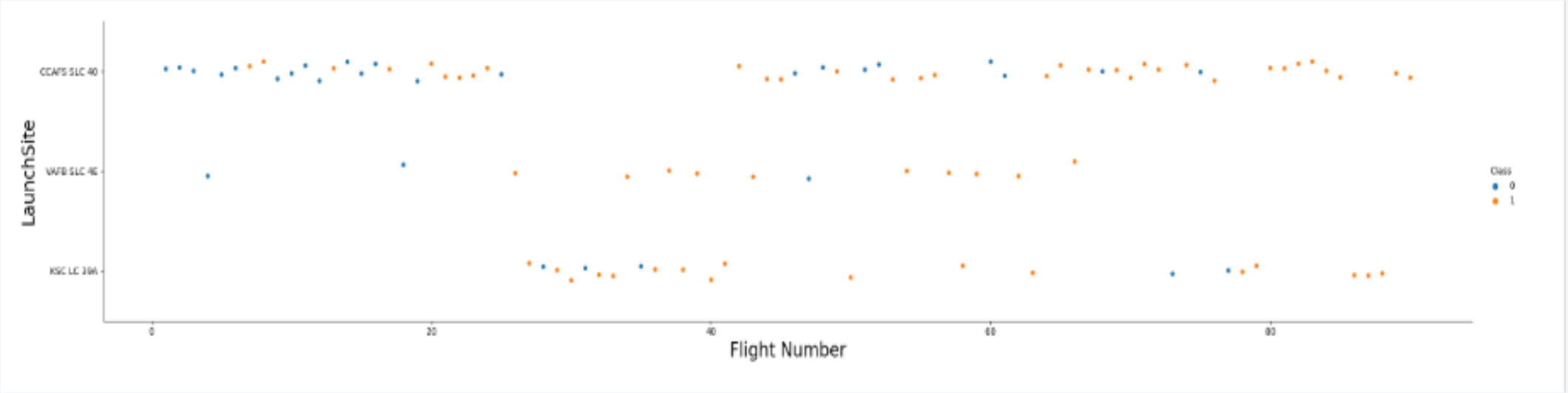
- **Data wrangling process**:



Source: https://rb.gy/09mgng

EDA with Data Visualization

- **Flight Number vs Launch site**:



- This scatter plot was used to plot out the flight number against the launch site . Here we see that the different launch sites have a different success rate.

Source URL :

https://rb.gy/lqoedk

EDA with SQL

- The SQL queries that were performed:
- **Names of the unique launch sites in the space mission**
- %sql SELECT DISTINCT(Launch_Site)FROM SPACEXTBL;
- **Display 5 records where launch sites begin with the string 'CCA'**

  %sql SELECT* FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- **Total payload mass carried by boosters launched by NASA (CRS)**
- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
- **Average payload mass carried by booster version F9 v1.1**
- %sql SELECT ROUND(AVG(PAYLOAD_MASS__KG_)) FROM SPACEXTBL
- **Date when the first successful landing outcome in ground pad was acheived.**

  %sql SELECT min(DATE) FROM SPACEXTBL
- **Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- %sql SELECT Booster_Version FROM SPACEXTBL WHERE `landing _outcome` = 'Success (drone ship)'AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

EDA with SQL

- **Total number of successful and failure mission outcomes**
- %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE '%success%' OR Mission_Outcome LIKE '%failure%'
- **Names of the booster_versions which have carried the maximum payload mass. Use a subquery**
- %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_)FROM SPACEXTBL);
- **Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
- %sql SELECT substr(Date, 4, 2) as month,substr(Date,7,4) as year,`landing _outcome`,Booster_Version,Launch_Site FROM SPACEXTBL WHERE su bstr(Date,7,4)='2015' AND `landing _outcome` LIKE 'Failure%';
- **Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.**
- %sql SELECT `landing _outcome`, COUNT(`landing _outcome`) as nr_of_Successful_landings FROM SPACEXTBL WHERE (`landing _outcome` LIK E 'Success%') AND `DATE` BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY `landing _outcome` ORDER BY 'nr_of_Successful_landings' DESC

Source:

https://rb.gy/zndpy0

Build an Interactive Map with Folium

- **Markers** are objects helping to make the names of the launch sites visible on a map.
- **Circles** are objects that serve as delimiters of the launch site area.
- **The lines** represent the distance between a point A and a point B on the map.

- Source: https://rb.gy/c4oehp

Build a Dashboard with Plotly Dash

- The plots ,graphs and interactions added to the Dashboard help us to visualize the influence the payload mass can have on the launch success rate of a site.
- Also the dashboard can help see which launch site has a better launch success rate over the years. The goal was to identify what launch site is the best.

Source :https://rb.gy/audlfi

Predictive Analysis (Classification)

- **Process of finding the best classification model:**

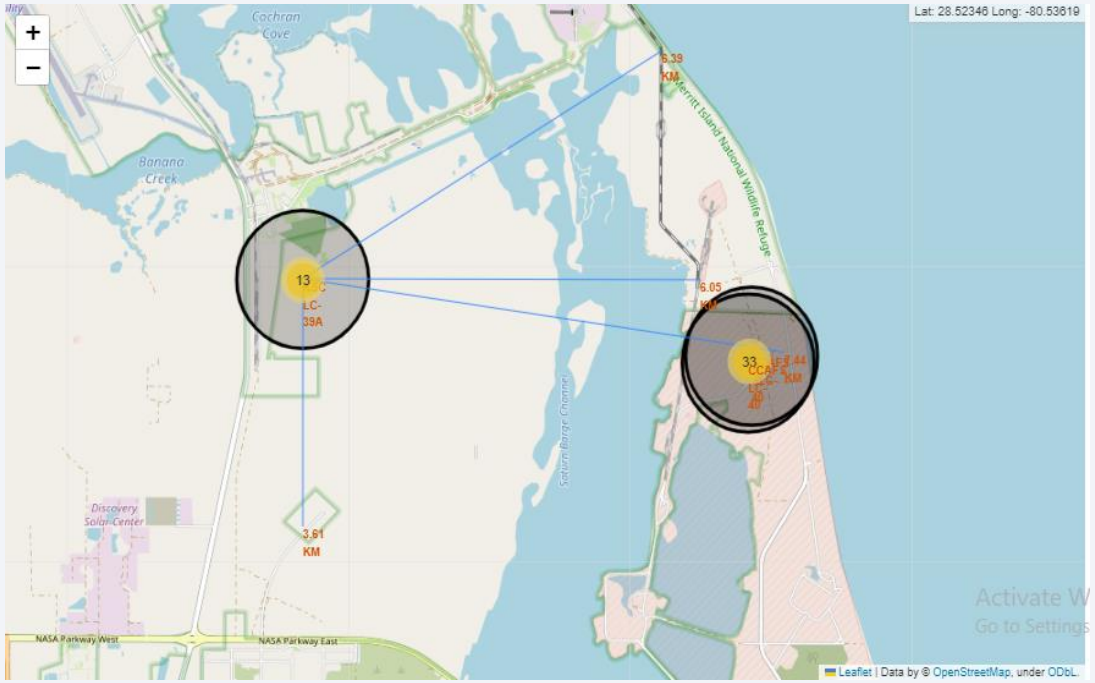
- Source: https://rb.gy/oahkz9

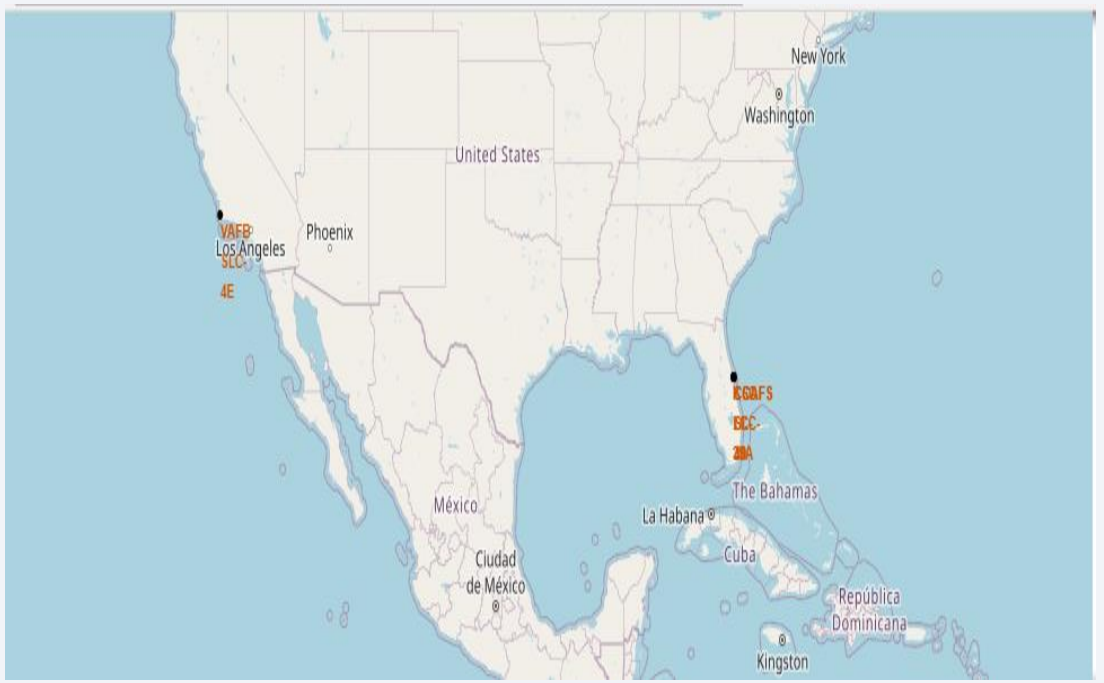
Exploratory Analysis Results

- **Exploratory data analysis results**:
- SpaceX has four different launch sites;
- The total payload mass carried by boosters launched by NASA (CRS) amounts to 619967 kg;
- The average payload mass carried by booster version F9 v1.1 is 6138 kg;
- The boosters versions F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2 have success in drone ship and have payload mass greater than 4000 Kg but less than 6000 Kg;
- Two boosters versions F9 v1.1 B1012 and F9 v1.1 B1015 had failure in landing outcomes in drone ship in year 2015;
- Success rate keeps increasing with time.

Exploratory Analysis Results

- **Exploratory data analysis results**:

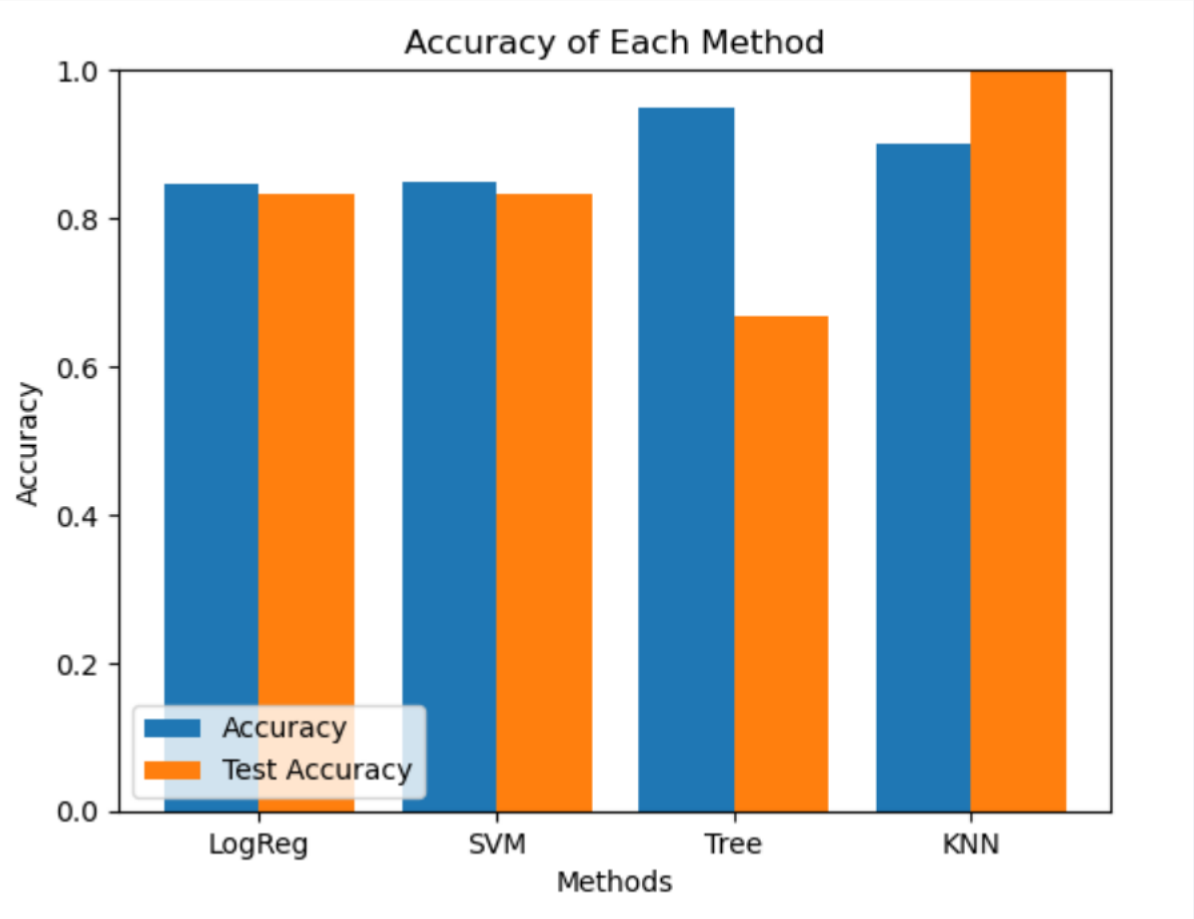


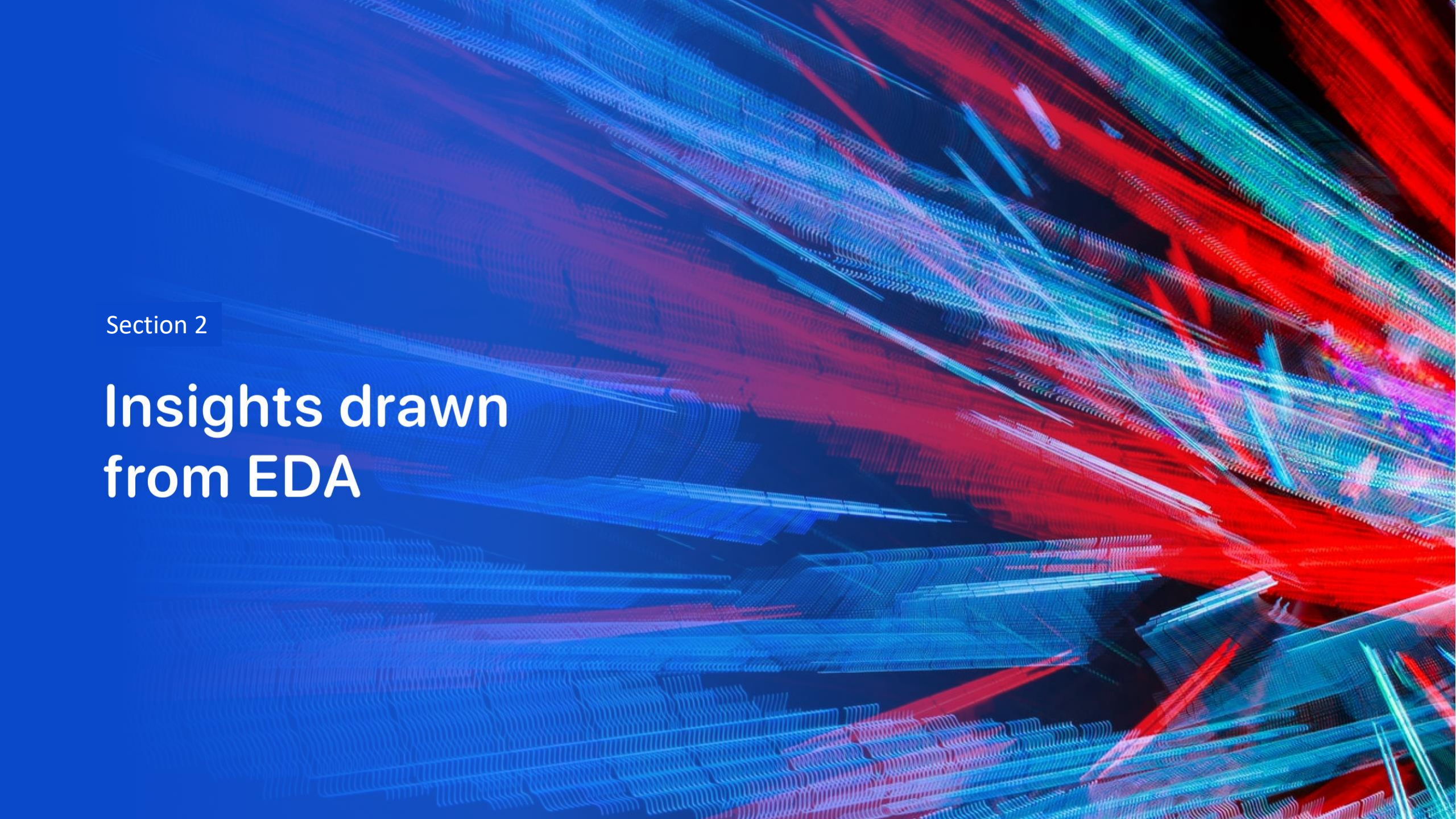


- Launch sites are situated at the east cost and particularly at the coastlines. There is a proximity of the launch sites with the railways and highways to facilitate the transportation of the materials from one point to another one.

Predictive Analysis Results

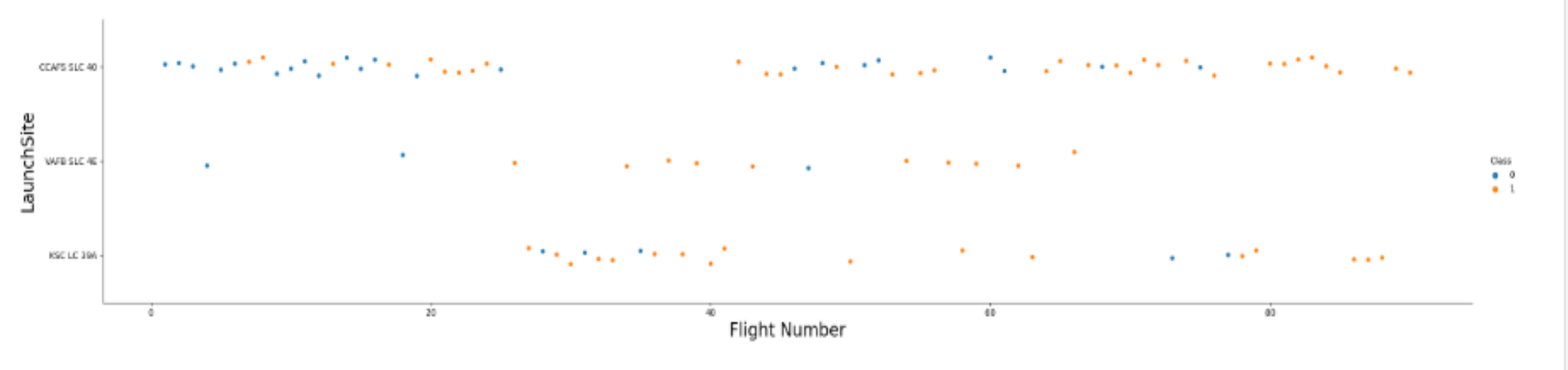
- **Predictive Analysis results**:
- The accuracy on the test data is at 1 and the one on the train is around 0.8 when looking at the KNN model. The best model is then the K-nearest Neighbor.




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

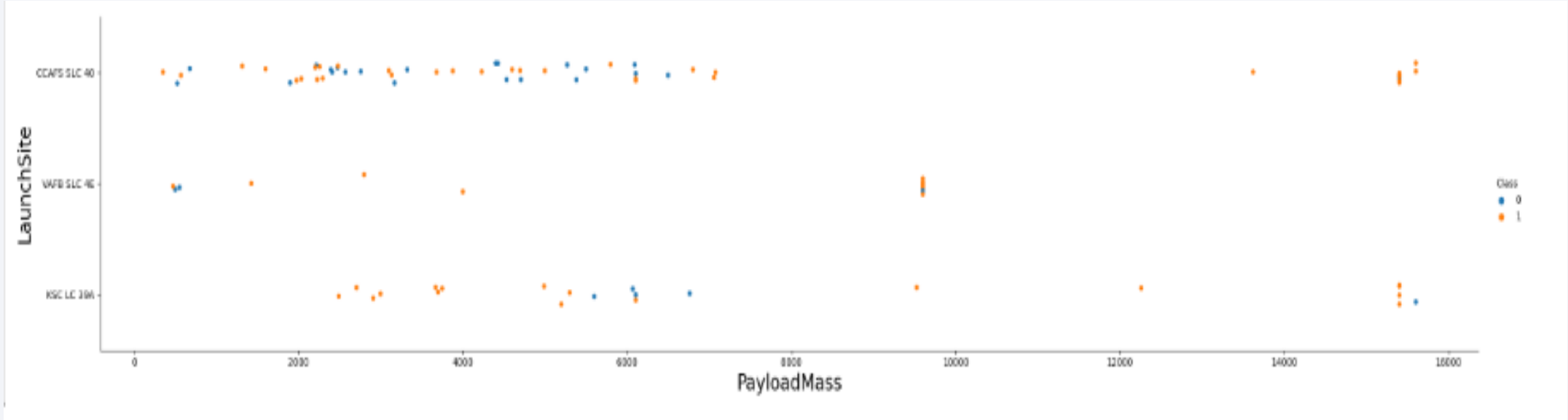
- **Flight Number vs. Launch Site:**



- The launch site CCAF5 SLC 40 has more launches and more success along the way and is therefore the best one.
- Second one is VAFB SLC 4E and in third place we have KSC LC 39A.

Payload vs. Launch Site

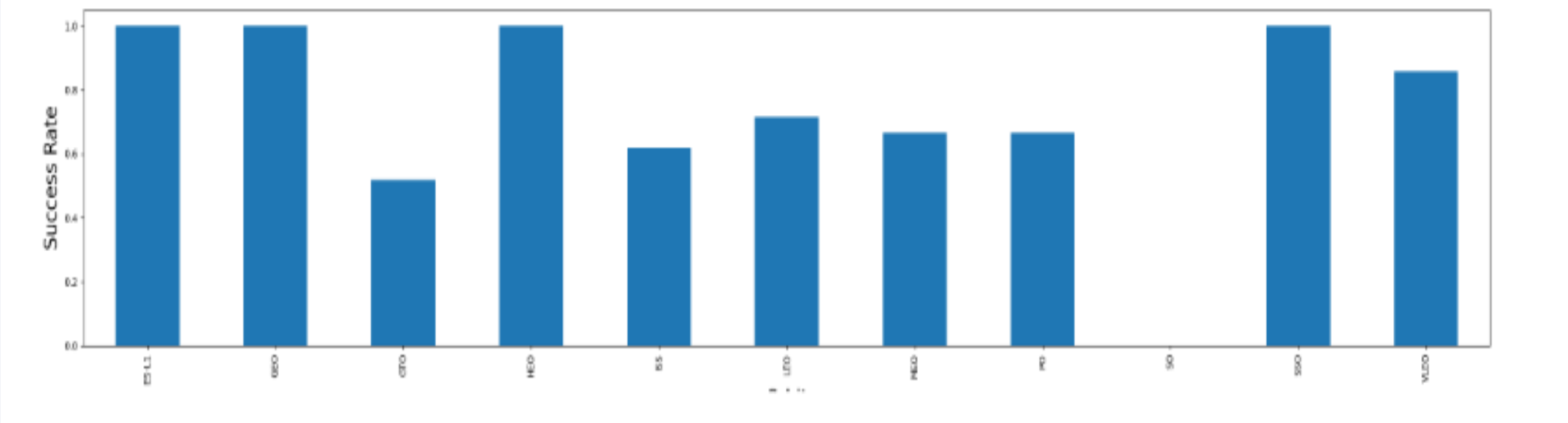
- **Payload vs. Launch Site:**



- There are less launches when the payload gets more heavy(8000kg).
- Starting at around 9000kg the success rate is very good specially for the CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type

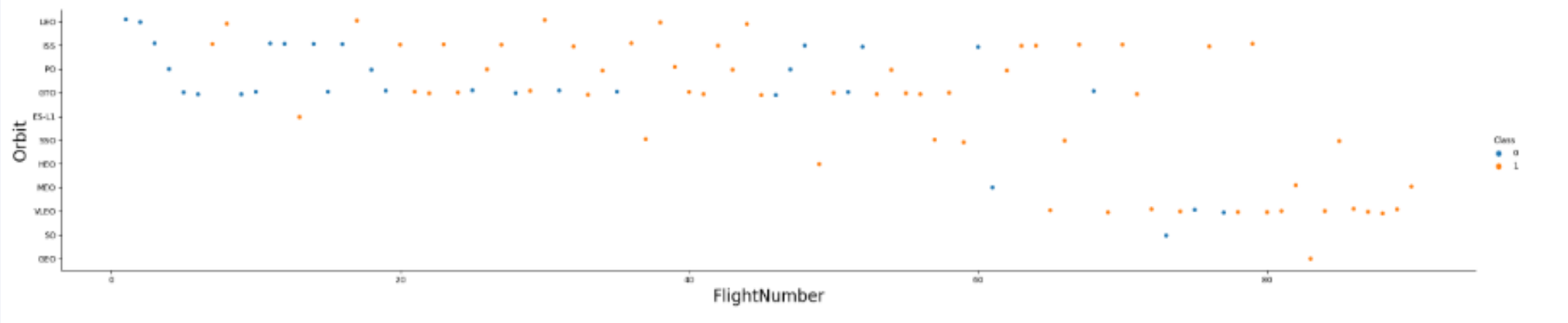
- **Success Rate vs. Orbit Type:**



- The best orbits for successful launches are SSO,VLEO,HEO,GEO,ES-L1.

Flight Number vs. Orbit Type

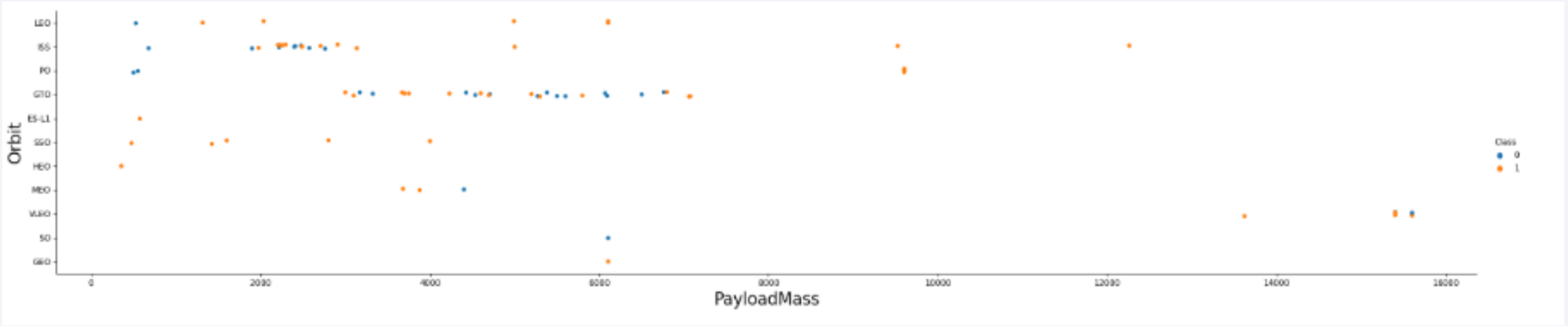
- **Flight Number vs. Orbit Type**



- In the LEO orbit the Success is related to the number of flights.
- there seems to be no relationship between flight number when in GTO orbit for example.

Payload vs. Orbit Type

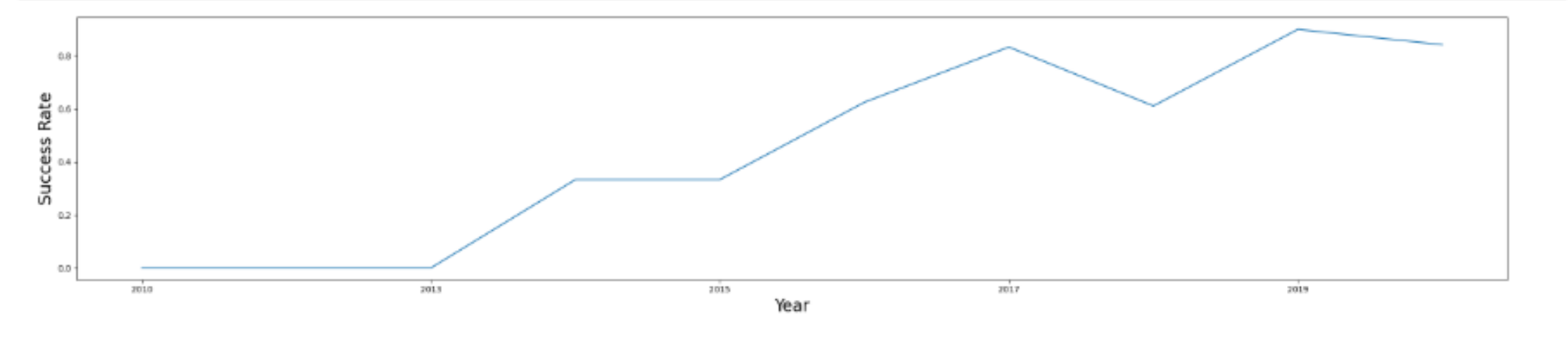
- **Payload vs Orbit Type:**



- The heavy the payload mass is the more successful the orbits Polar and ISS are.
- For GTO for example there is a mix of success and failure.

Launch Success Yearly Trend

- **Launch Success Yearly Trend:**



- Starting from the year 2013 the success rate seems to be increasing until the year 2020.
- It seems that with the experience and improvement of materials the trend will go higher.

All Launch Site Names

- According to the data there are four unique launch sites:

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

- The names are found in the database through the use of the unique occurences “Launch_site“ values.

Launch Site Names that start with 'CCA'

- There are five Launch site names beginning with ‘CCA’:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The data was retrieved by selecting all the data and applying the filters ‘CCA’ in the database.

Total Payload Mass

- Total payload carried by boosters from NASA:

SUM(PAYLOAD_MASS__KG_)
619967

- This payload mass is calculated by summing all the payload containing the code (CRS) which refers to NASA.

Average Payload Mass by F9 v1.1

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

ROUND(AVG(PAYLOAD_MASS__KG_))
6138.0

- The result is obtained by filtering payload mass to the booster version F9 v1.1 and then finding the average of that payload mass in the database.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

min(DATE)
01-03-2013

- The date was found by filtering the minimum value possible among dates in the database.

Successful Drone Ship Landing with Payload between 4000 and 6000

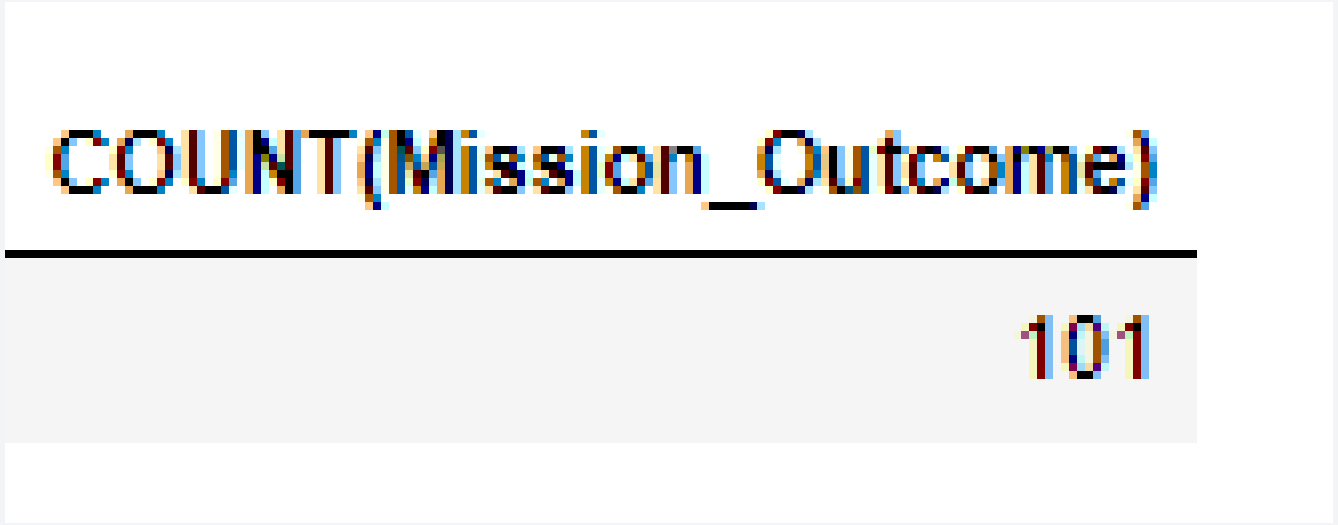
- **Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg:**

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- To find the list ,we filtered the landing outcomes and also the payload mass between 4000kg and 6000 kg in the database.

Total Number of Successful and Failure Mission Outcomes

- **Total number of successful and Failure Mission Outcomes:**



COUNT(Mission_Outcome)
101

- We counted the number of mission outcomes in the database and that brought the above result.

Boosters That Carried Maximum Payload

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

- The list was found by looking for all the booster versions and filtering by the maximum payload mass.

2015 Launch Records

- **2015 Launch Records:**

month	year	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- We found the above results by looking for the failed landing Outcome in the database and by applying a filter by year 2015, the Booster version and the launch site.

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

- **Landing Outcomes between 2010-06-04 and 2017-03-20**:

Landing _Outcome	nr_of_Successful_landings
Success (ground pad)	6
Success (drone ship)	8
Success	20

We retrieved the list by counting the number of Landing outcomes and we applied filter values like “Success“ and “Year“.

Section 3

Launch Sites Proximities Analysis

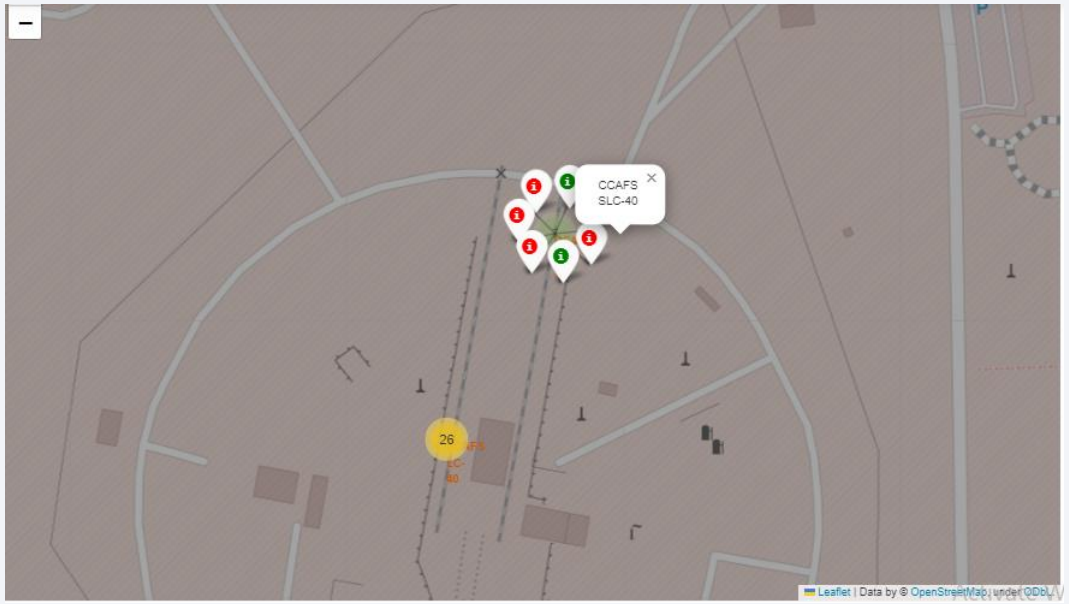
Sites Locations

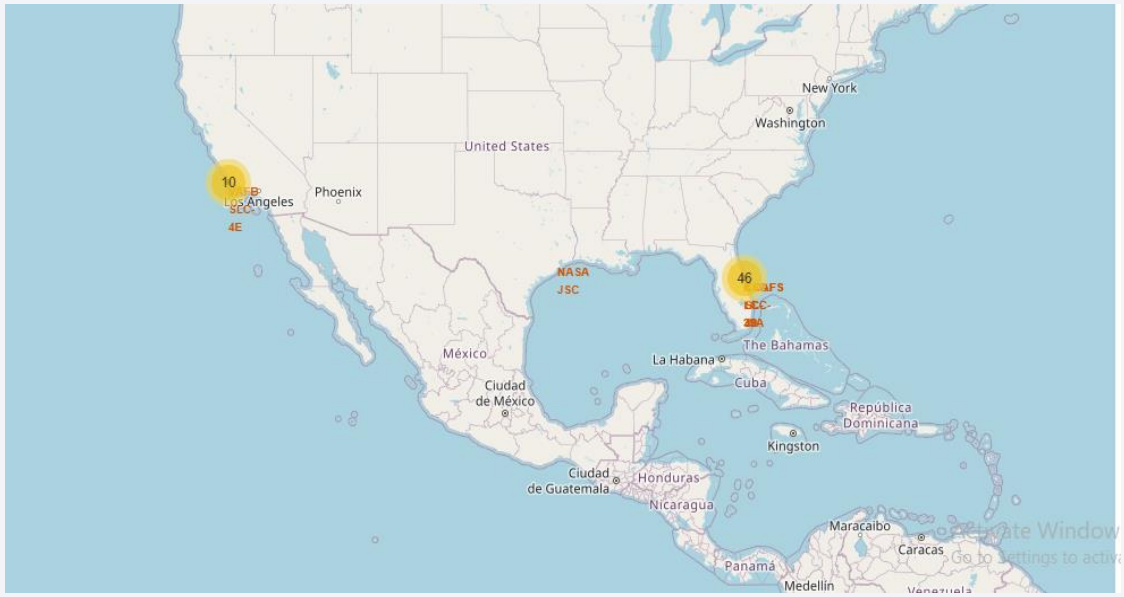
- **Sites locations on the map**:

- We can see on the map that the sites locations are near the water.There is no site to see in the middle of a city.

Launch Outcomes

- **Launch Outcomes**:

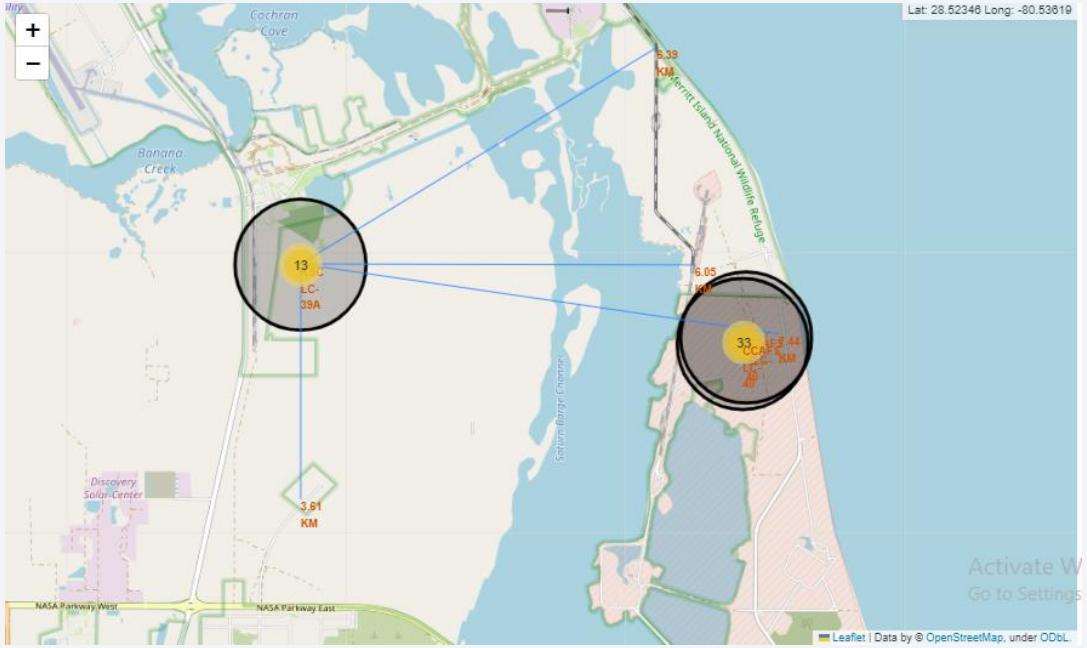




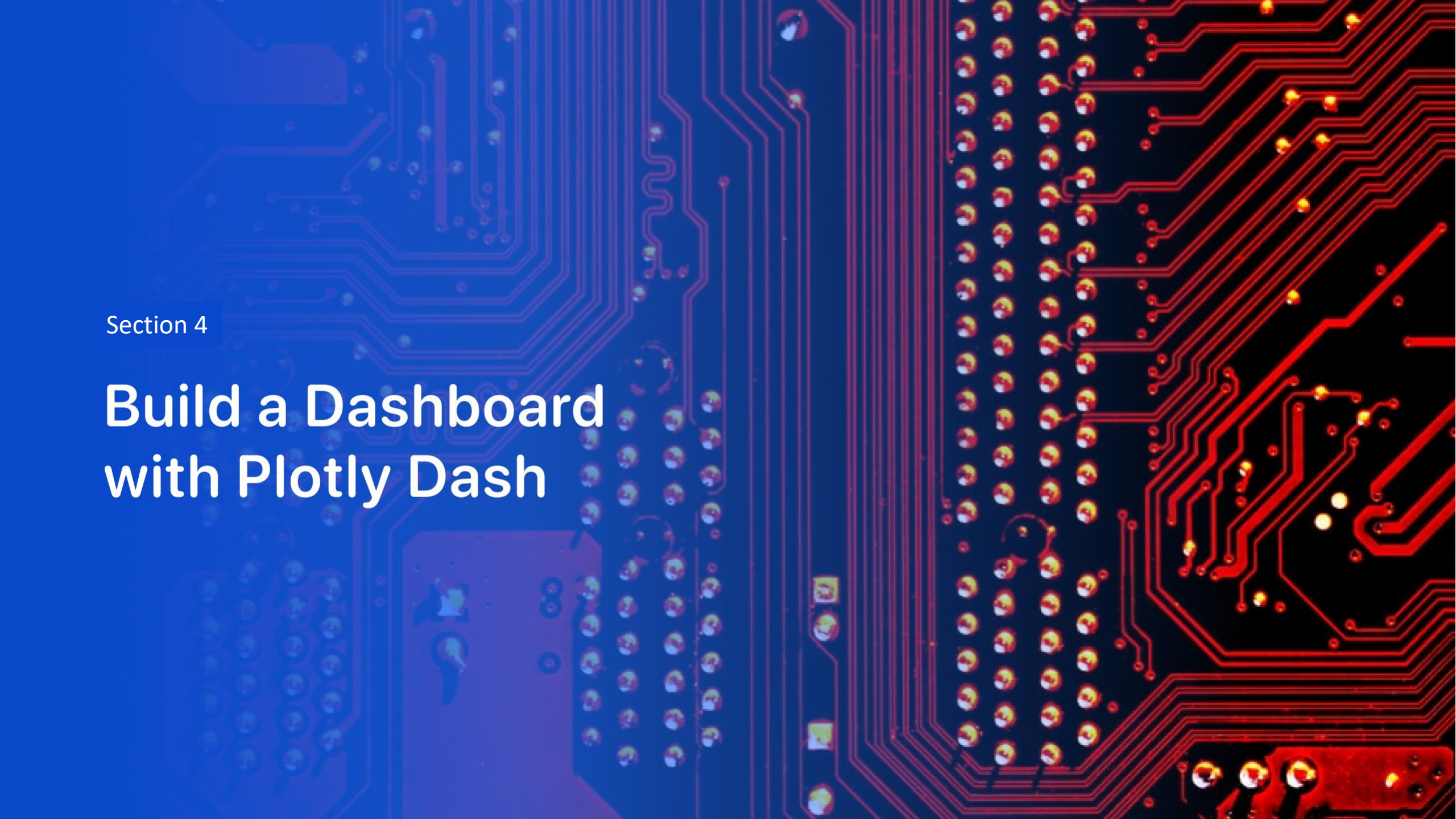
- On the right we have the yellow marked sites and on the left we zoomed in and we can see that the launch site CCAFS SLC 40 seems to have a very low success rate.

Launch Site

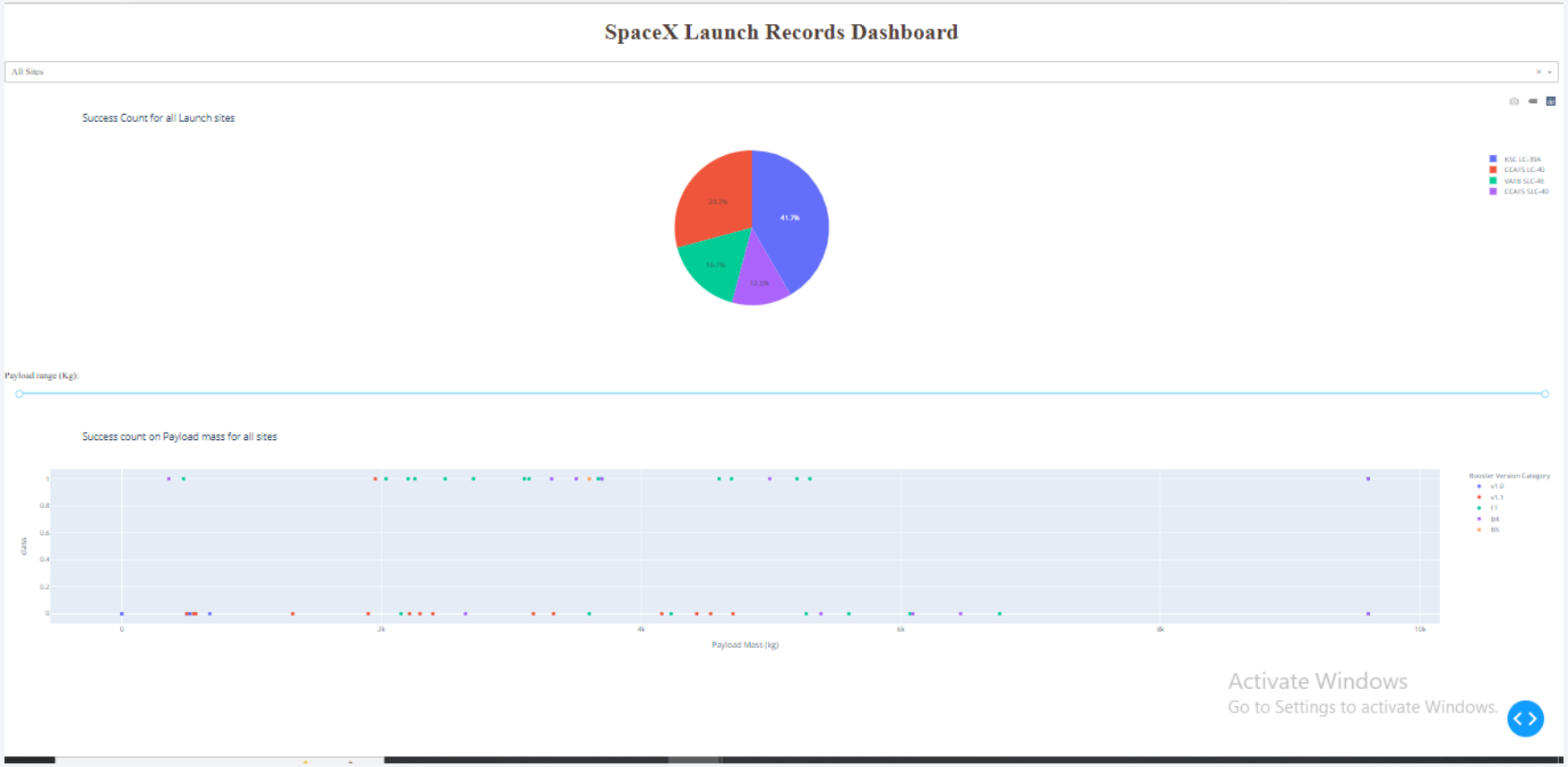
- **Launch site proximity**:



- The maps display a proximity of the sites to the coastline. Also we can observe that there are railways and highways near the site.The goal of such proximity is surely to facilitate the transportation of the materials to sites.

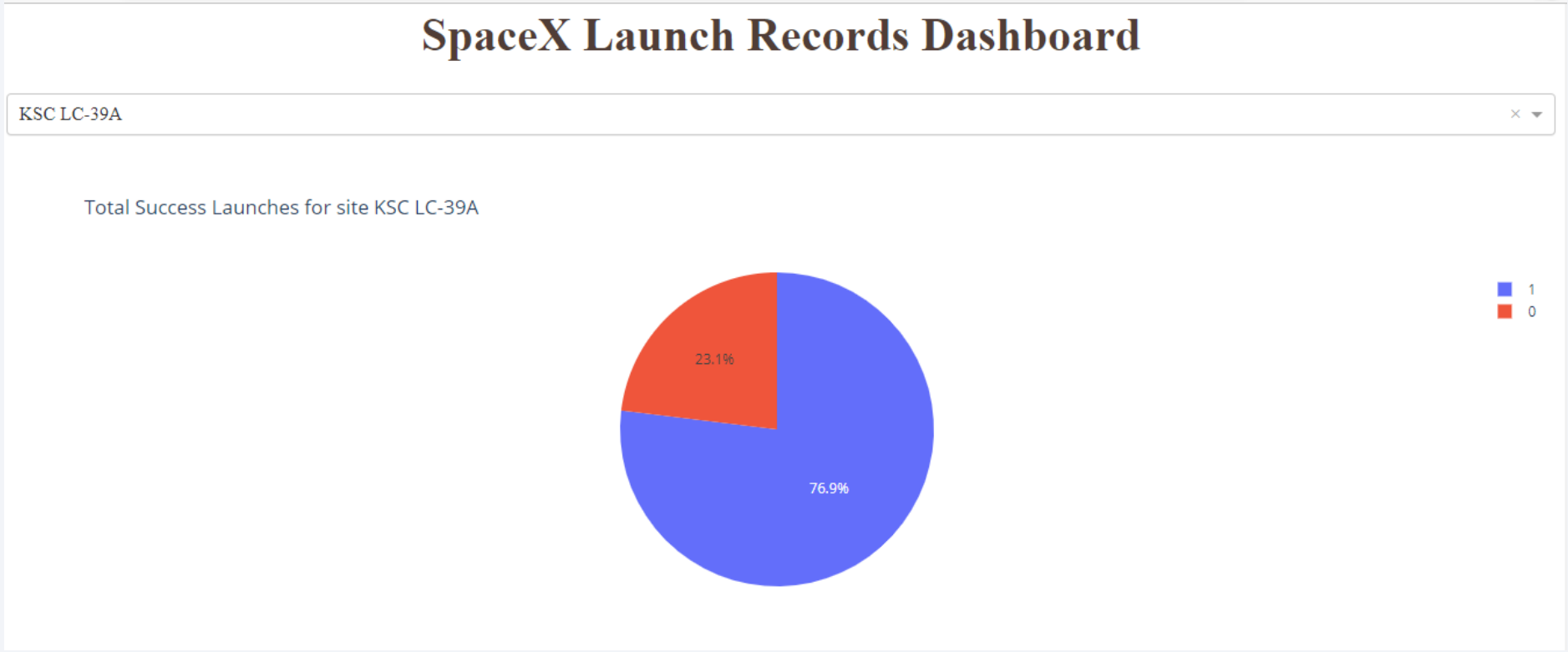

Section 4
Build a Dashboard with Plotly Dash

Success count For All Launch sites



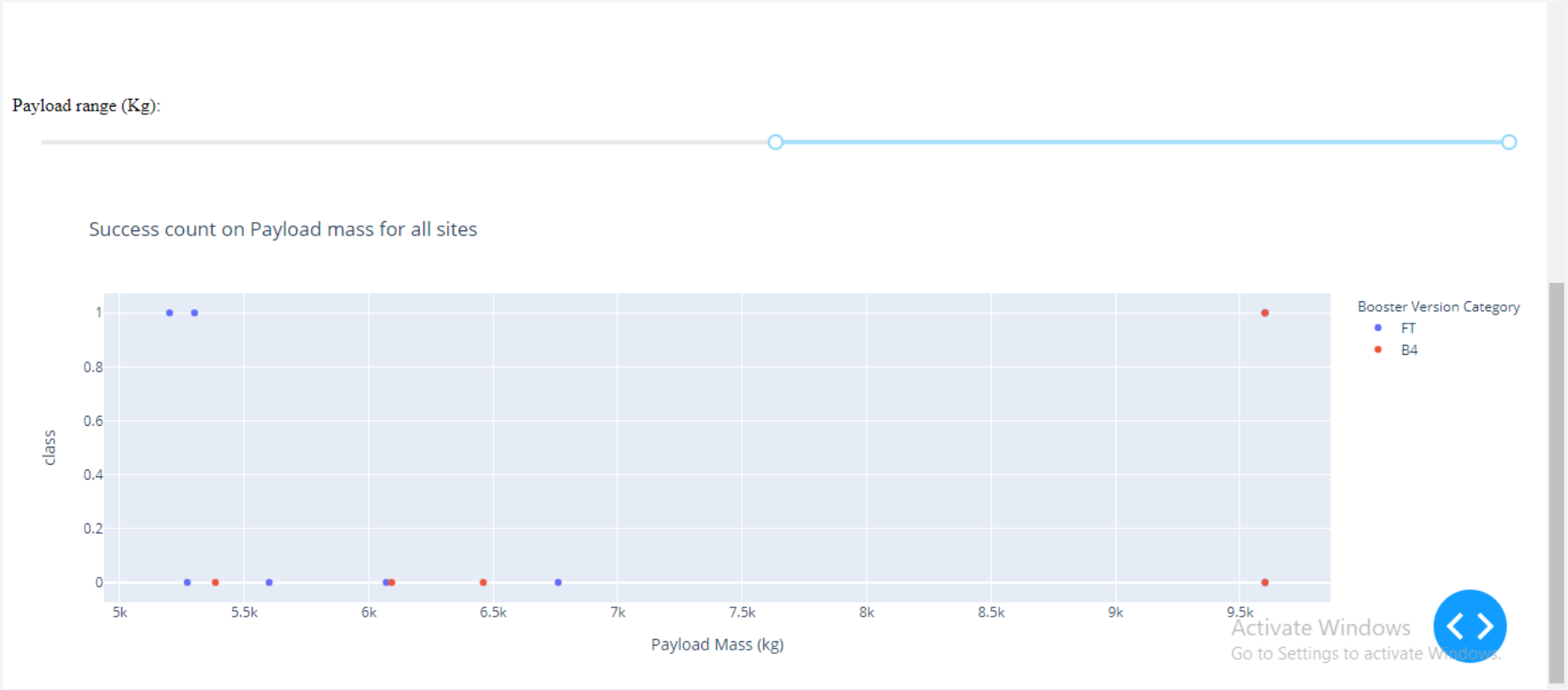
- We can see that when the payload mass is over 6000 kg there are less launches. So the heavier the payload mass the less launches there are.

Success ratio for site KSC LC-39A



- The site KSC LC-39A is the one having the most successful launches with a percentage of 76.9% compared with the other sites.

Payload vs Launch Outcome



- With the payload Mass starting at 6000 Kg there are nearly no launches. The Booster version FT has a larger success rate.

Payload vs Launch Outcome

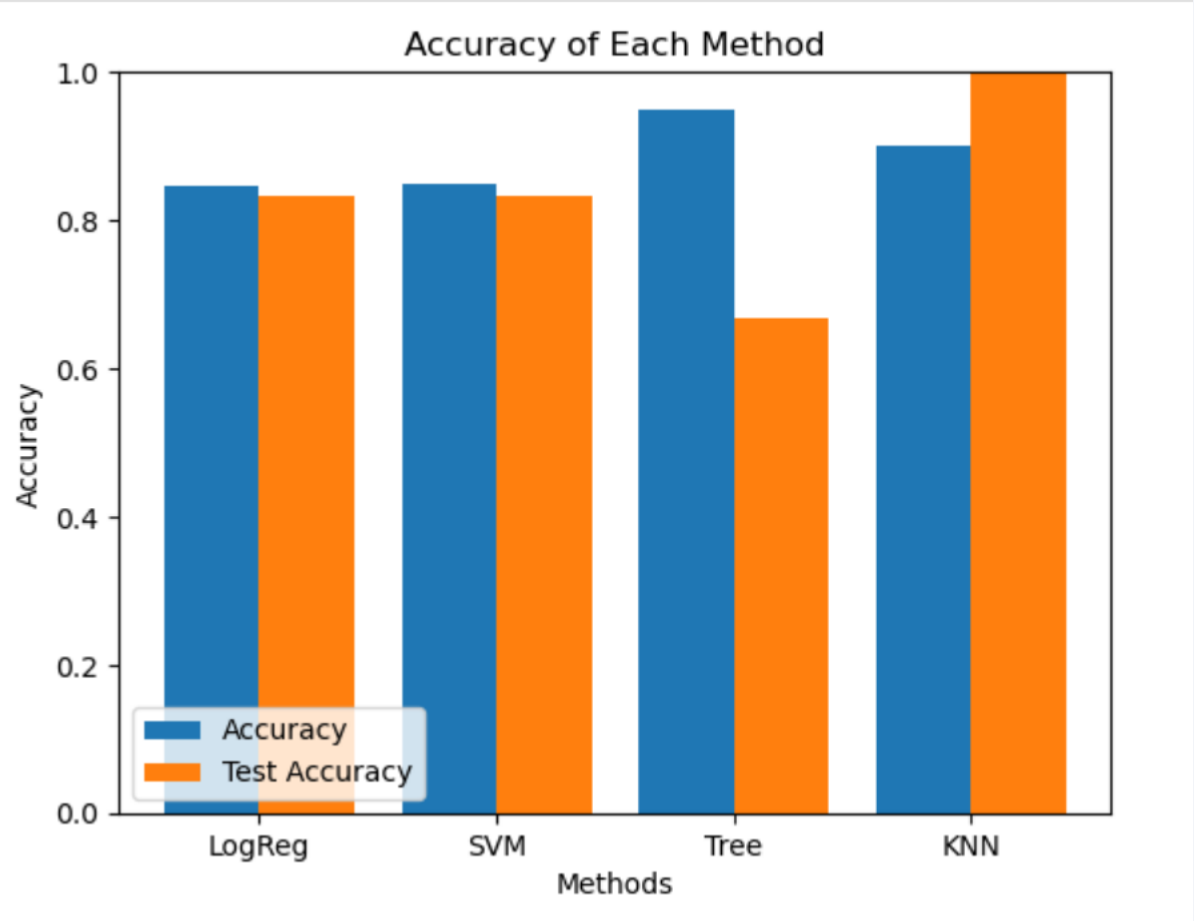
- Starting at 3000 kg there are less and less launches. The Booster version FT is also the one with a greater success.

Section 5

Predictive Analysis (Classification)

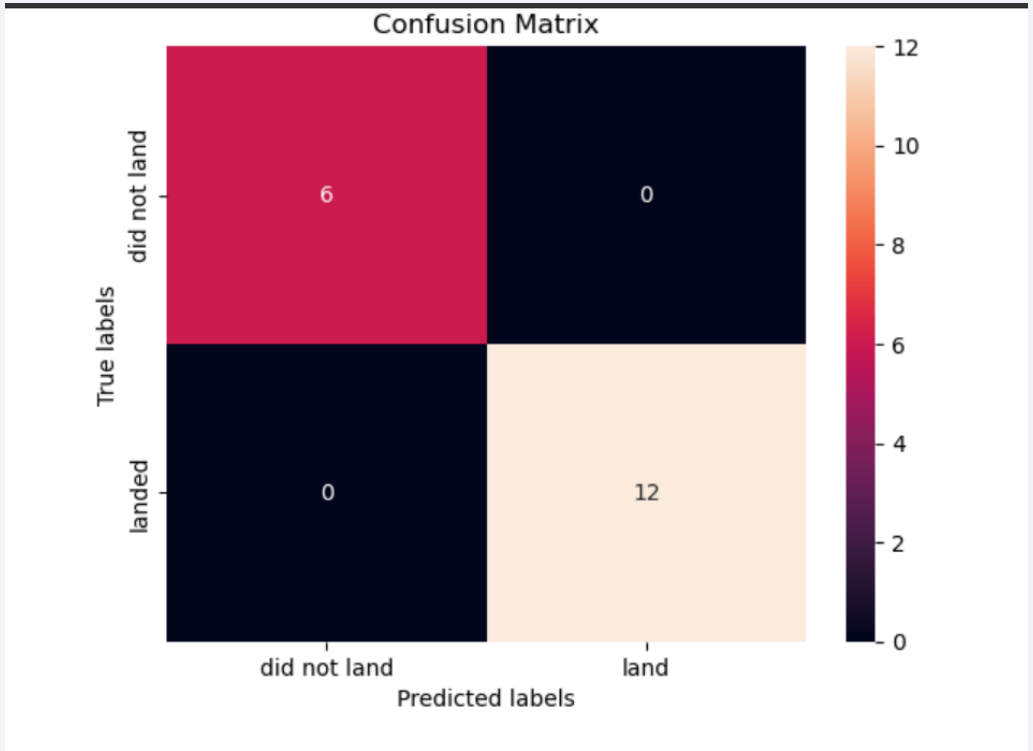
Classification Accuracy

- The K-nearest Neighbor has the highest classification accuracy



K-nearest Neighbor Confusion Matrix

- Confusion matrix of the K-nearest Neighbor which is our best performing model:



- The confusion matrix proves the accuracy of the K-nearest Neighbor by showing big numbers of true positive and true negative compared to the false ones.

Conclusions

- The best launch site is CCAFS SLC 40.
- Mission outcomes are generally successful and landing outcome is improving with time. That improvement comes surely from the improvement of past mistakes.
- The best classification model is the K-nearest Neighbor that helps us to predict successful landings therefore make more money on the long run.

Appendix

- I did not mention the number of successful and failure mission outcomes. Here is the sql query ➔ Select count (Mission_Outcome)FROM SPACEXTBL WHERE Mission_Outcome LIKE '%success%' OR Mission_Outcome LIKE '%failure%'
- Link to the project: https://github.com/ClementSouffez/IBM-Applied-Data-Science-Capstone-Project

Thank you!