IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Higor Mazza e Silva
May 1, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: API and Web Scrapping
  - Data Wrangling
  - Exploratory Data Analysis (EDA): Pandas and SQL
  - Interactive Maps with Follium and Dashboards with Plotly Dash
  - Predictive Analysis: Classification Machine Learning (ML) Algorithms
- Summary of all results
  - Insights and graphs from EDA
  - Data Presentation with Dashboard and Maps
  - Predictive analysis – Landing Success

Introduction

- Project background and context
  - Space travel is booming at a rate not seen since the Apollo Missions. Great efforts are being made to make space tourism accessible, and the exploration of the moon, Mars (even colonization) are being seriously discussed and are, in fact, closer than ever.
  - One of the major leaps in the space race (that contribute for the booming) was the design of reusable, reliable, and safe rockets. Reusability allows companies like SpaceX to re-fly the most expensive parts of the rocket, which helps to reduce the launches cost. The key component of this business model is the success of landing the core boosters.

- Problems you want to find answers
  - There are any patterns on the previous launch's data that we can explore?
  - Can we used ML models over previous launches to predict the success of a new mission?

Section 1

Methodology

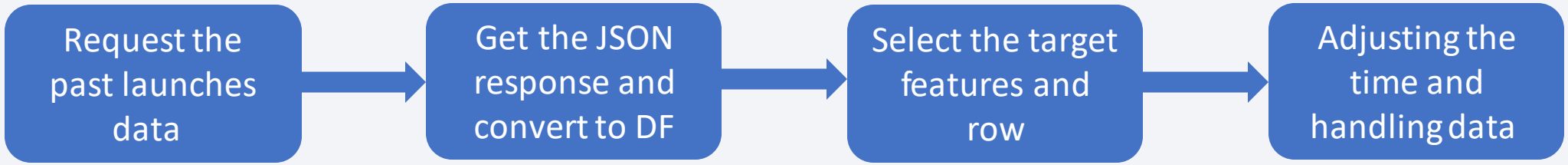
Methodology

Executive Summary

- Data collection methodology:
  - To methods were used to collect date: SpaceX-API and Web Scrapping HTML Tables
- Perform data wrangling
  - Determine Training Labels and Encode Data
- 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

- We used the SpaceX API to request the target data as follow – Raw DataFrame:



Using the get method from the requests lib.
Get hold of the data from past launches.

Converting the JSON response to a pandas DataFrame(DF) through the json_normalize() method

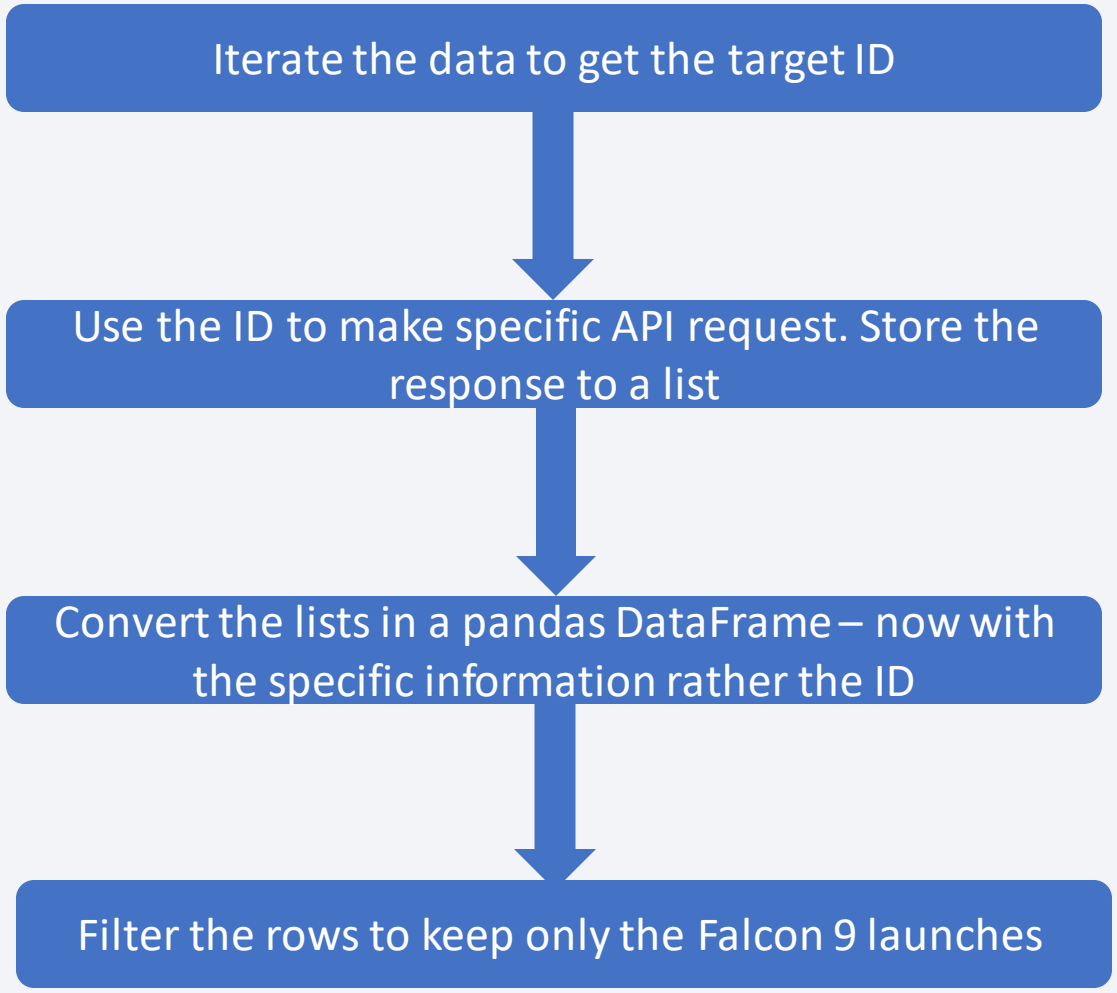
Selecting all the target columns and removing the rows with multiple cores.

Restrict the dates from the Launches in the DF and access the data stored as a list.

GitHub Link - Data Colletion API

Data Collection – SpaceX API

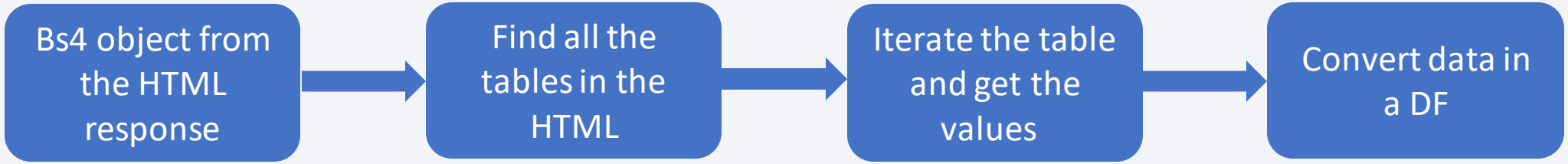
- Some data was encoded as an ID
- Another API was conduct using the ID's obtained fom each launch.
- We iterate over the data and made specific API requests - e.g. getBoosterVersion uses the ID to request the Booster 'name' and append it to a list



GitHub Link - Data Colletion API

Data Collection - Scraping

- BeautifulSoup(bs4) to web scrap HTML tables about Falcon 9 launch records from Wikipedia:



Get the HTML from the Wikipedia page through a get.request

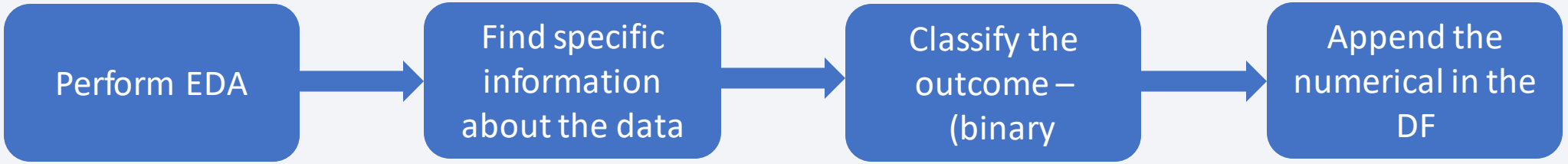
Parse the HTML to get all the tables and assing the result to list – identify the target table

Extracted the values from each row.

GitHub Link - Data Collection Web Scrapping

Data Wrangling

- The Goal of the Data Wrangling was to identify pattern in the data and select the training labels



The general information was analyzed in order to find the datatypes, proportion of missing values...

Identify the Launch Sites and it's occurrence. The same was performed for the Orbit and Outcome feautures

The content of the Outcome column was classified based on a binary system. Good and bad outcomes (1,0). Transforming it in numerical data.

Now, the numerical equivalent of the Outcome can be used as training data.

GitHub Link – Data Wrangling

EDA with Data Visualization

- The following plots were used to help with the Data Viz:
- Scatter :Payload Mass (kg) vs Flight Number – Get the relation between mass and the flight number, and use color as an indicator of successful landings.
- Scatter: Flight Number vs Launch Site - Success rate for each site can be observed as the flight number increases.
- Scatter: Payload Mass vs Launch Site - The success rate for site, given a mass, can be visualized using different colors to represent the "class" feature.
- Bar plot: Obrit vs Success Rate - By calculating the mean value ("class") for each orbit the success rate can be observed.

GitHub Link - EDA with Data Viz

EDA with Data Visualization

- The following plots were used to help with the Data Viz:

- Scatter: Flight Number vs Orbit - The success rate for each orbit can be observed as the flight number progresses

- Scatter: Flight Number vs Orbit - The success rate for each orbit can be observed for the different mass ranges.

- Line plot: Date vs Success Rate - Success rate over the years.

GitHub Link - EDA with Data Viz

EDA with SQL

SQL magic was used to query the following pieces of data:

- Names of the unique launch sites;
- 5 records where launch sites begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- The date of the first successful landing outcome in a ground pad;
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
- Total number of successful and failure mission outcomes;
- The names of the booster_versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship in the year 2015;
- Ranking of landing outcomes (count)

GitHub Link - EDA SQL

Build an Interactive Map with Folium

- We used tree Follium Objects:
  - Circle to indicate location of Launch Sites;
  - Marker to get info of landing success for each site (cluster);
  - Marker was used to indicate the distance in km;
  - MousePosition to get [lat, long] of a specific map point (use to calculate the distance);
  - Polyline as a qualitative indication of the distance between Launch Sites and its proximities;

GitHub Link - Map Visualization with Folium

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard;

- Two filter were applied: Launch Sites and Payload Mass Range
  - The Launch site was presented as a Dropdown menu;
  - The mass range as a RangeSlider;

- There were two graphs displayed:
  - Pie chart that received as a input the Launch Site;
    - If ALL a chart showing the success rate for all the site was generated;
    - For a specific site, the proportion of success and failure of that site was rendered;
  - Scatter plot showing the success/ failure in relation with the Payload Mass – it takes two inputs: Launch Site (from dropdown) and a mass range (slider);
    - The color was used as a dimension to show the Booster Version;

GitHub Link - Dashboard

Predictive Analysis (Classification)

The primary goal of predictive analysis was to develop a classification model to determine if the first stage will land.

- We standardize the data and determined the Training Labels (X). Then assign the dependent variable (Y) to the class feature.
- After standardizing the dataset, it was split into a Train, Test data, with test_size of 0.2 (20%);
- 4 models were evaluated:
  - Logistic regression;
  - Decision Tree;
  - Support Vector Machine;
  - KNN;
- The best hyperparameters were find with GridSearch analysis;

GitHub Link - Predictive Analysis

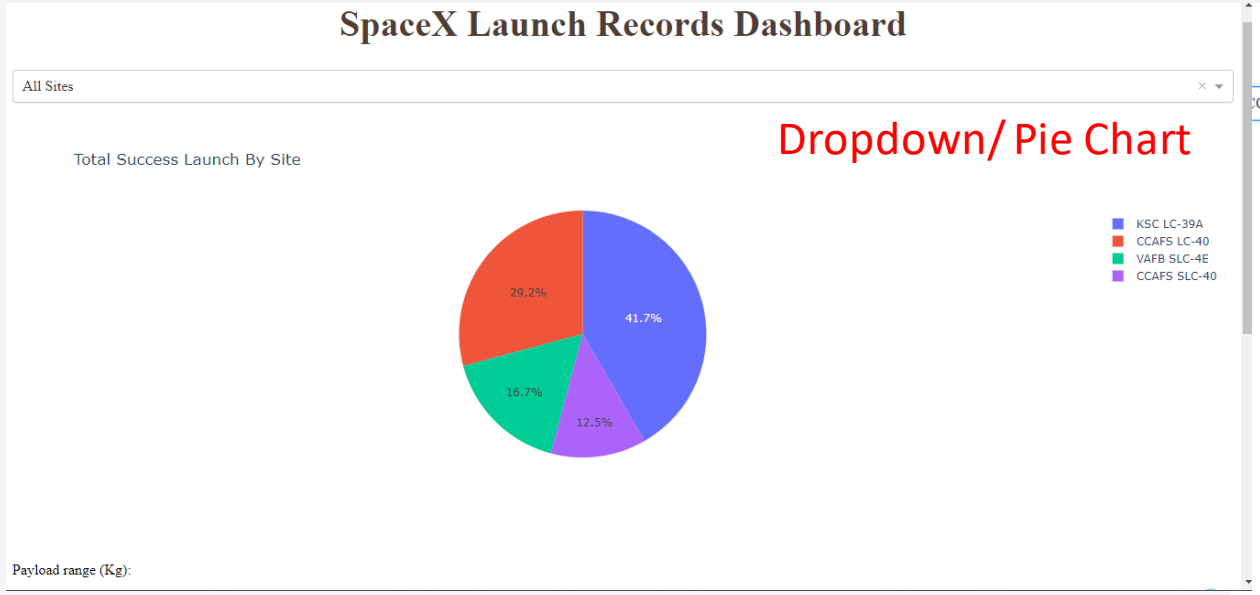
Results

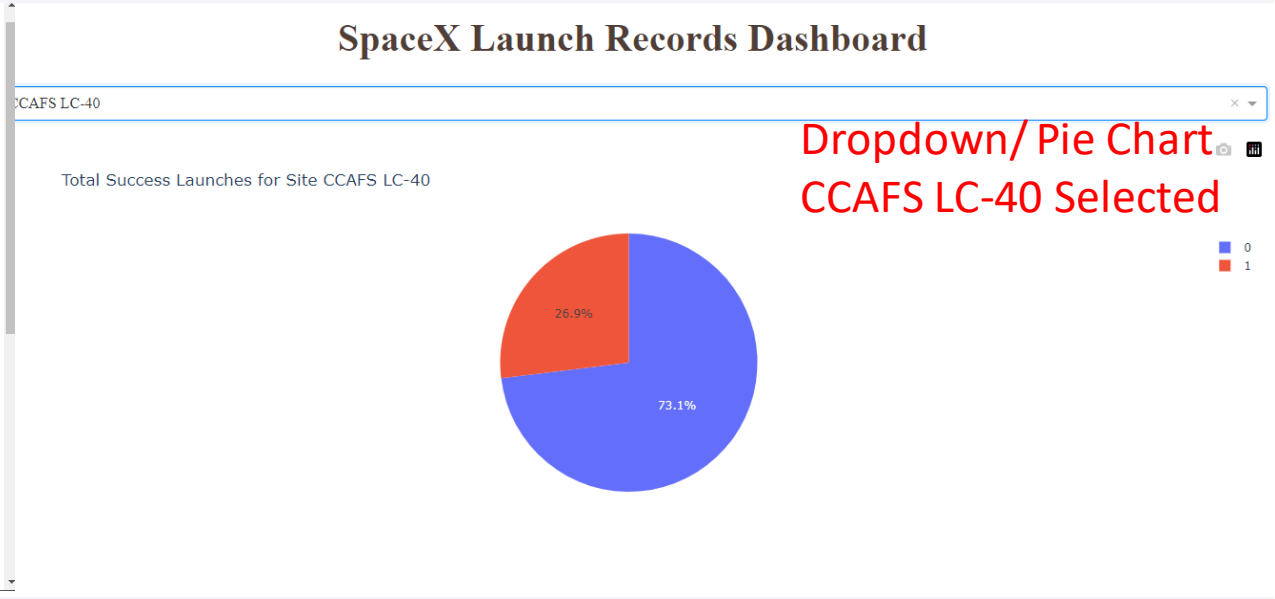
- Exploratory data analysis results - Summary

Data Visualization	SQL
• As the flight number increases, the landing is more likely to be successful • Different Sites have different success rate • The orbits ES-L1, GEO, HEO and SSO have a Success Rate of 1 (100%), but less launches; • The orbit is related to the flight number, e.g. all data points for VLEO are after flight number 60. The relation of the orbit occur with payload mass too. • And the Success Rate kept increasing since 2013	• There are 4 unique Launch Sites • NASA (CRS) as a customer has sum of almost 50000 kg of payload • The first successful landing in a ground was in 2015 • The mission outcome have 100 flights classifies as success to 1 failure • For the landing outcome, most records were assign to "No attempts", follow by a same number of "Success" and "Failure" at a Drone Ship.(2010 – 2017)

Results

- The first pie chart give the total number of success launches per site. We can see that KSC is responsible for more than 40%, while CCAFS less than 13%
- The second pie chart from the second one reports the proportion between successful and failed launches. The CCAFS had a success rate of 73%

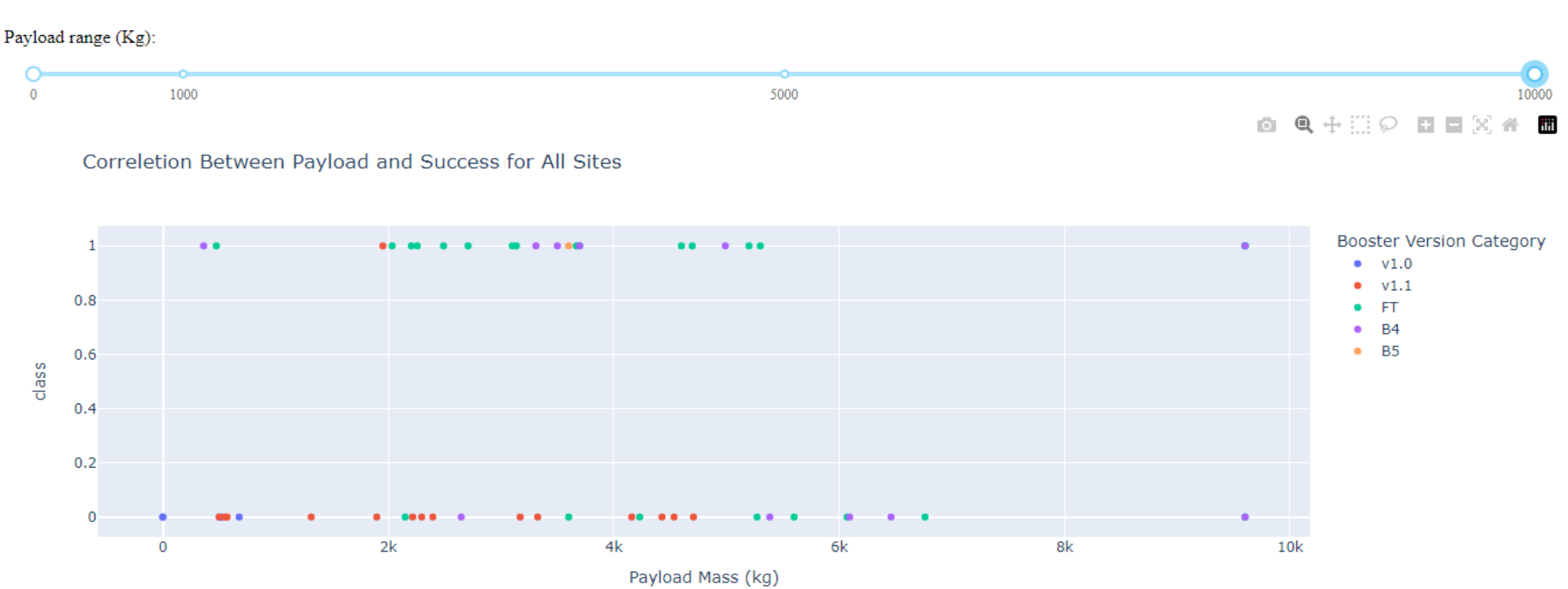




Results

- This graphs in the dashboard takes 2 inputs: Payload Mass Range and Launch Site – the color aspect points to the
  - You can filter ALL or a specific launch site and see what is the success frequency for the given payload mass.

Slider/ Scatter Plot



Results

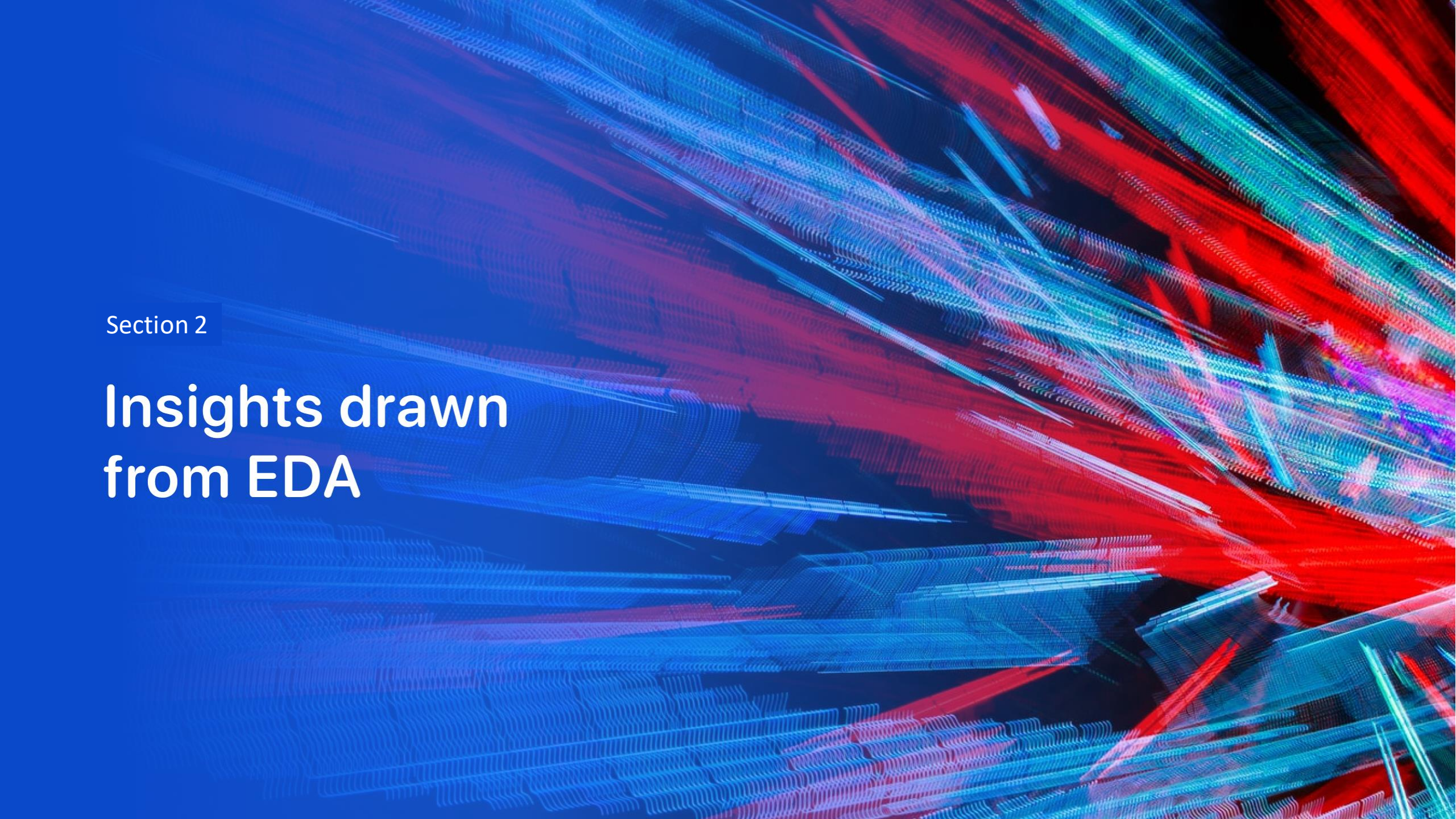
- Predictive analysis results

Model	Hyperparameters	Accuracy (Training Data)	Accuracy (Test Data)
Log. Regression	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	0.846	0.833
Dec. Tree	'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}	0.889	0.833
SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid' }	0.848	0.833
KNN	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}	0.848	0.833

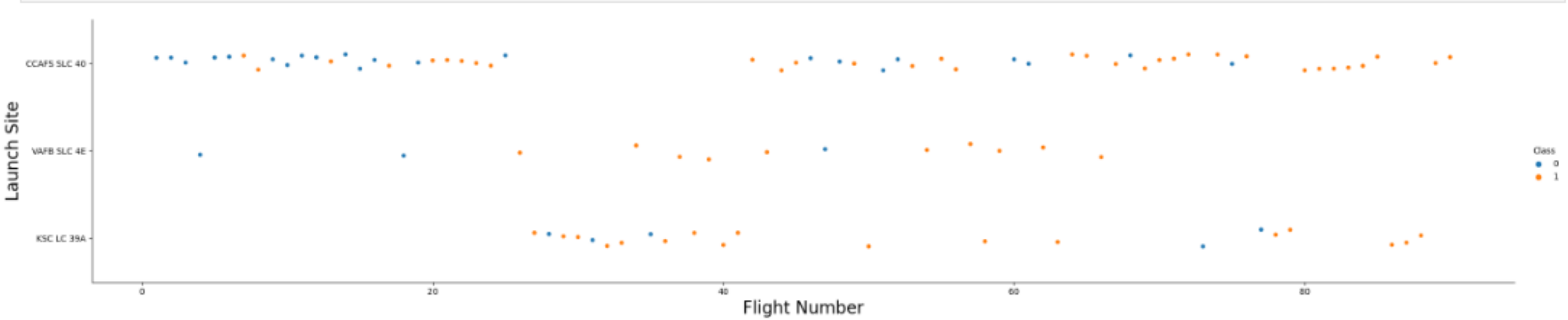
- Grid Search was perform with cv = 10;

Section 2

Insights drawn from EDA

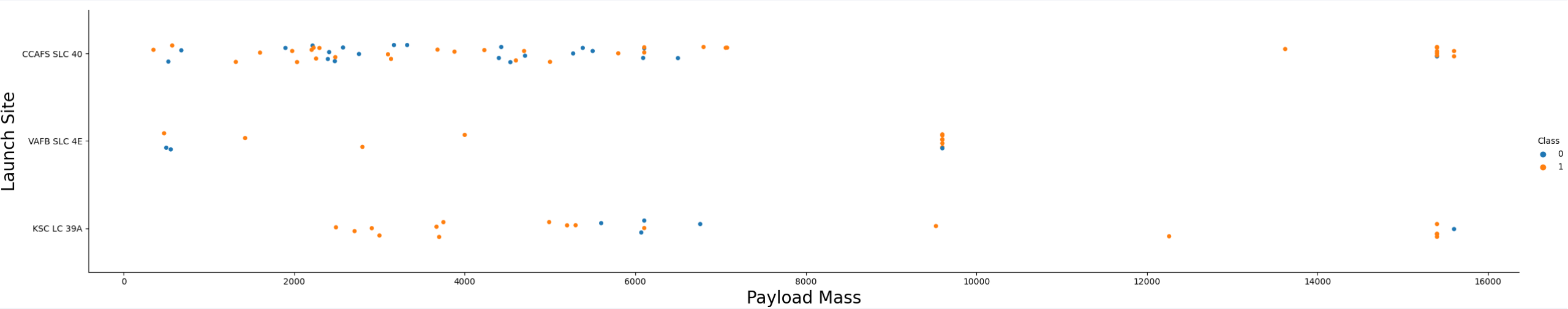


Flight Number vs. Launch Site



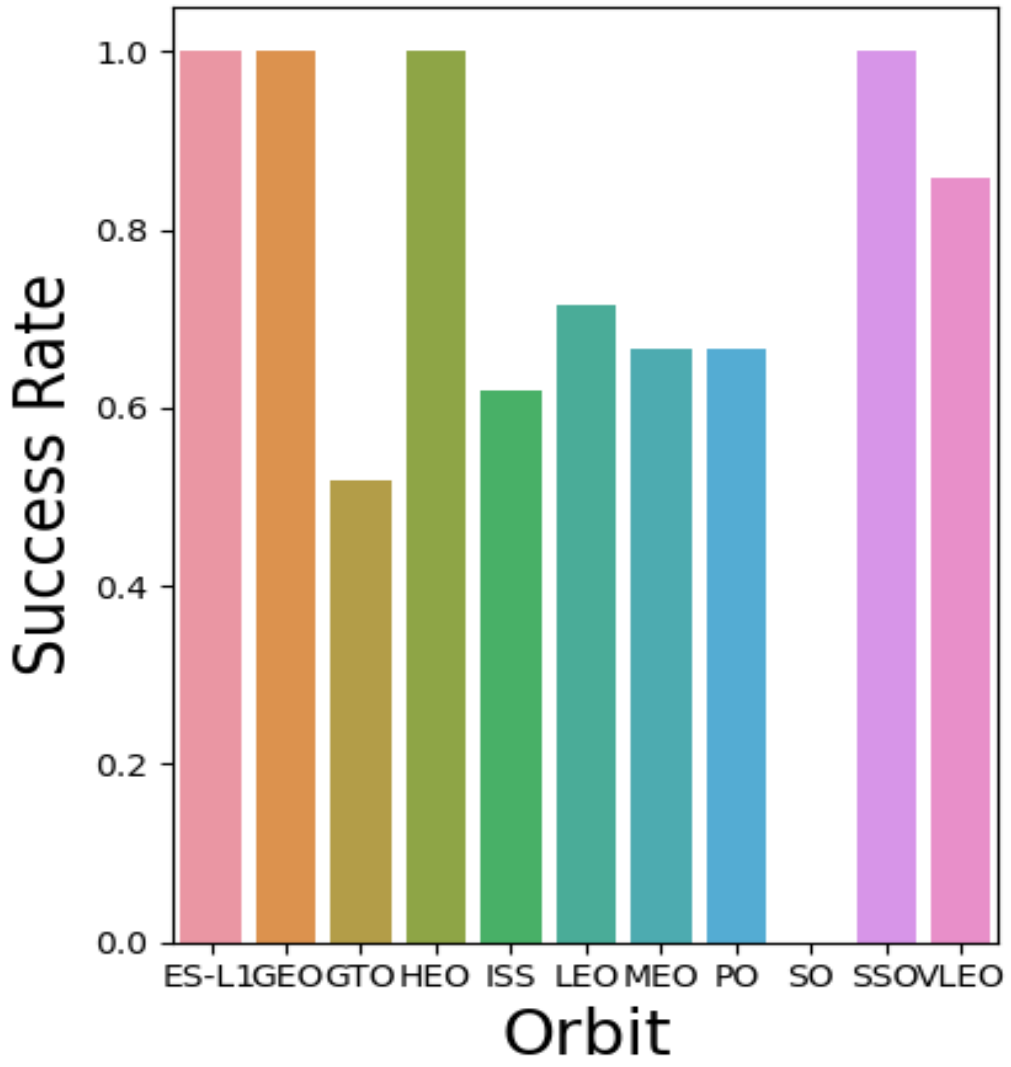
- The first observation that can be made is the higher success rate for larger flight numbers
- The success rate and number of flights vary by site

Payload vs. Launch Site



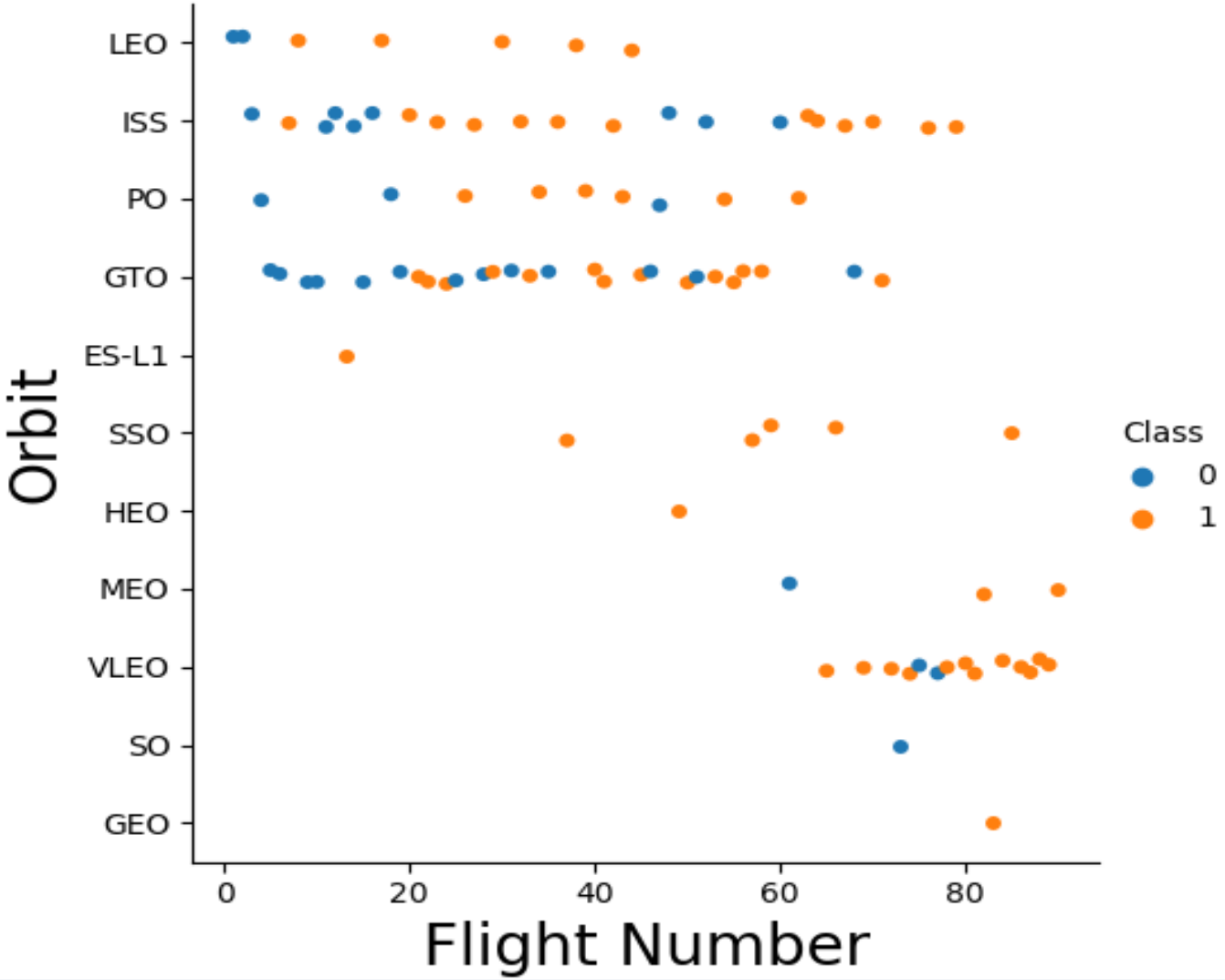
- VAFB SLC 4E does not show any launch with Payload Mass greater than 10000 kg;
- Higher Payloads (>8000 kg) appear to have had a higher rate (CCAFS, KSF);
- The majority of Flights takes place in the lower Payload Mass values;

Success Rate vs. Orbit Type



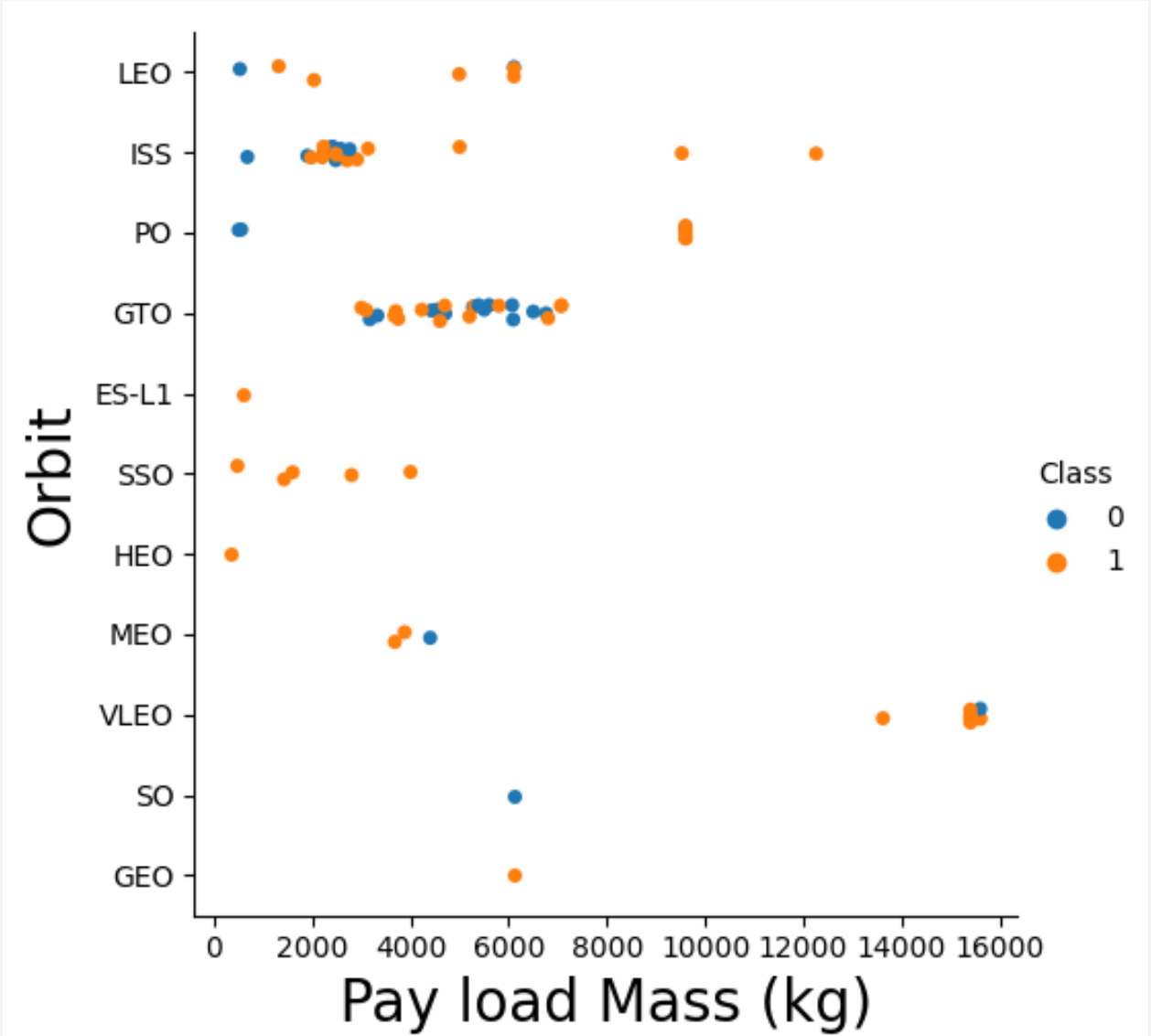
- Bar chart of the orbit type and their success rate (calculated by the mean of the "class" feauture
- ES-LI, GEO and SSO orbits have a rate of 100%
- SO presents a 0 as it's rate, but we will see in the next graph that only one launch reach this orbit;

Flight Number vs. Orbit Type



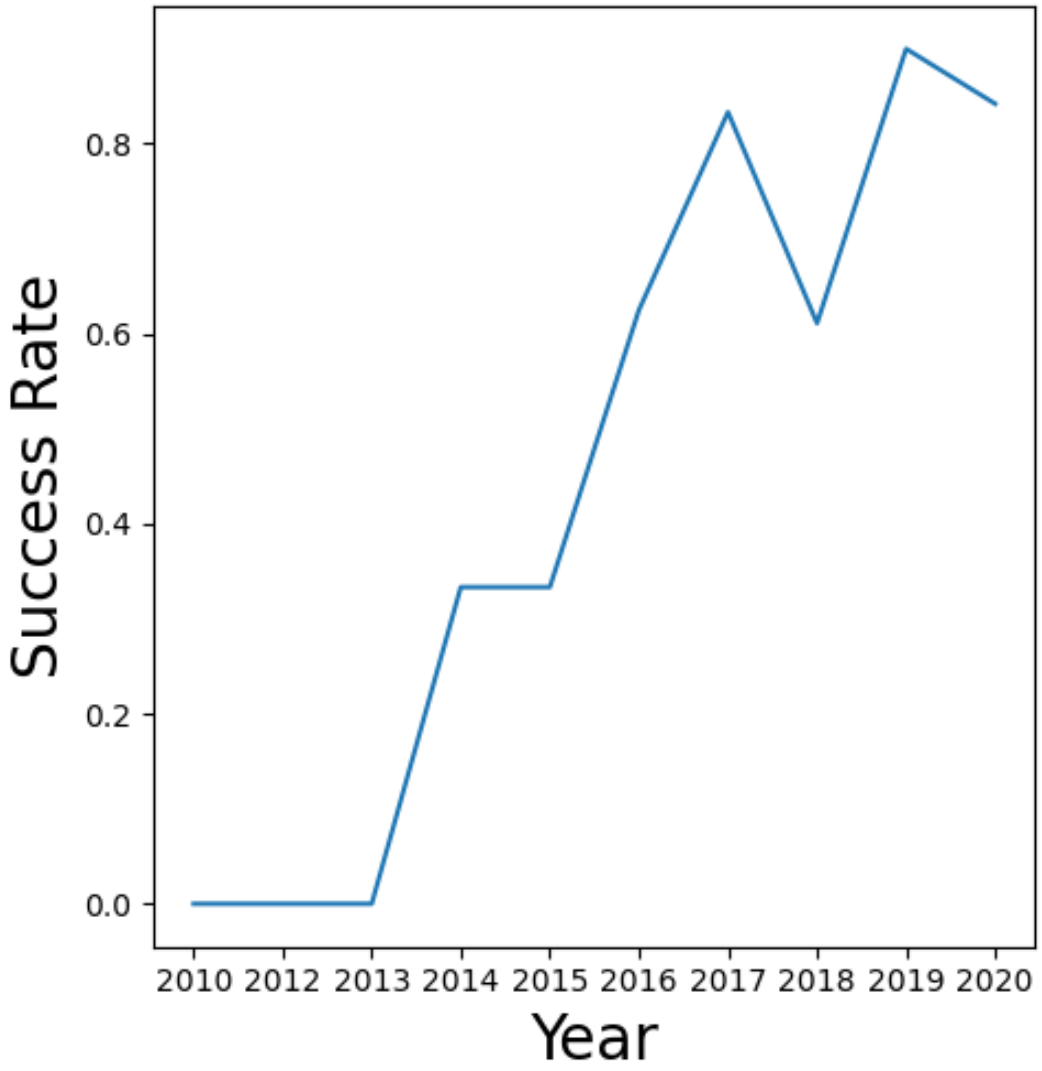
- Earlier flights mostly occured in the LEO, ISS, PO, GTO orbits;
- Later ones took place in VLEO – showed a higher success rate;
- In the previous bar chart, we saw that ES-L1, HEO and GEO had a 100% success rate. However, here we can see that are fewer launches to these orbits;

Payload vs. Orbit Type



- Payloads of 10000 or more appear only at ISS and VLEO – with great success ratio;
- To GTO happened a lot of launches in the 3000 to 7000 kg range (bad success rate);
- ES-L1, SSO and HEO only carry low payloads;

Launch Success Yearly Trend



- The success rate was compared to the years in which launches took place;
- We can see that landing success has improved gradually since 2013;

All Launch Site Names

- The launch site names in the dataset were retrieved using SQL;
- The statement uses the distinct method to get only the unique names;

Out[4]:

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

Launch Site Names Begin with 'CCA'

Out[5]:

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

- Here are presented 5 records with the name of the site begging with 'CCA';
- To retrieve these records, the like command was used;
- We can see that most of them went to LEO/ LEO(ISS)

Total Payload Mass

- The total payload carried by boosters from NASA (CRS)

Out[14]:

SUM
48213

- Values in the figure was retrieved as the sum of all the payloads mass were NASA (CRS) was the customer (There other customer starting with NASA);

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

Out[7]:

AVG
2534

- We calculate the average of the mass carried by F9 v 1.1 the booster version;

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad;

Out[8]:

DATE
2015-12-22

- The first successful landing on a ground pad happened only in 2015;

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;

Out[9]:

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

- Only the boosters F9 BT have successfully landed on the drone ship in the presented conditions;

Total Number of Successful and Failure Mission Outcomes

Out[10]:

outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- We can see that almost all mission completed their primary outcome;

- 100 successful to 1 failure (in flight);

Boosters Carried Maximum Payload

Out[11]:

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

- All the boosters that carry the max payload were the F9 B5 version;

2015 Launch Records

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

Out[12]:

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

- There were 2 failures in landing outcome in the year of 2015. Both took place at CCAFS LC-40 and used F9 v1.1 booster;

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,

- The most reported record was 'No attempt', followed by drone ship landing (Failure and Success - 5 appearances each);

- The least common landing outcome for the period was 'Precluded (drone ship)';

Out[16]:

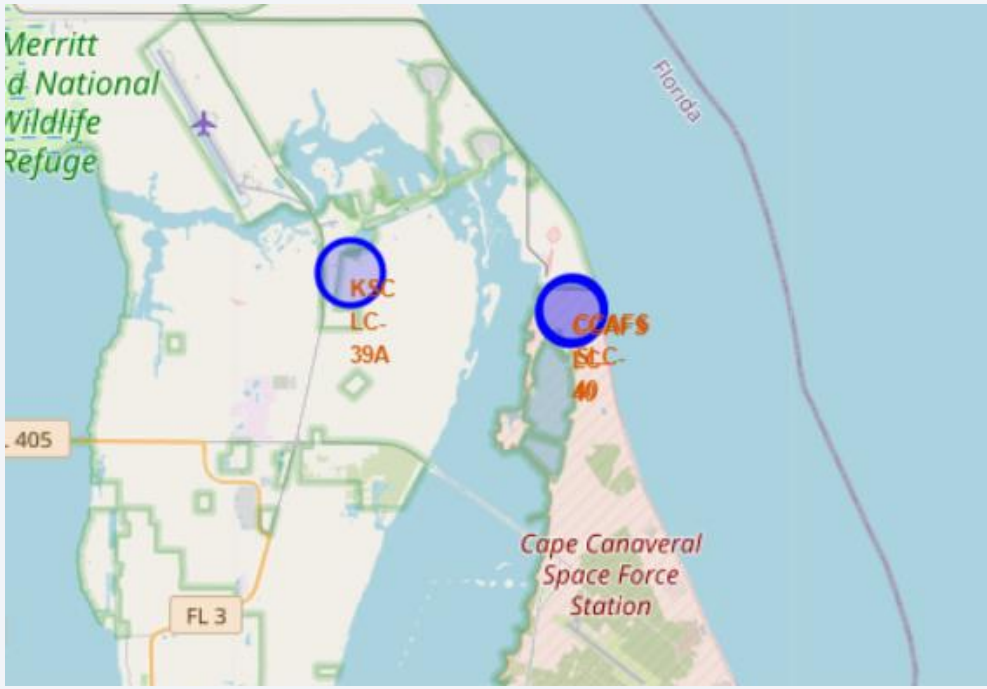
landing__outcome	count_outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Map – Launch Sites Location

- Next slide present the zoomed map;

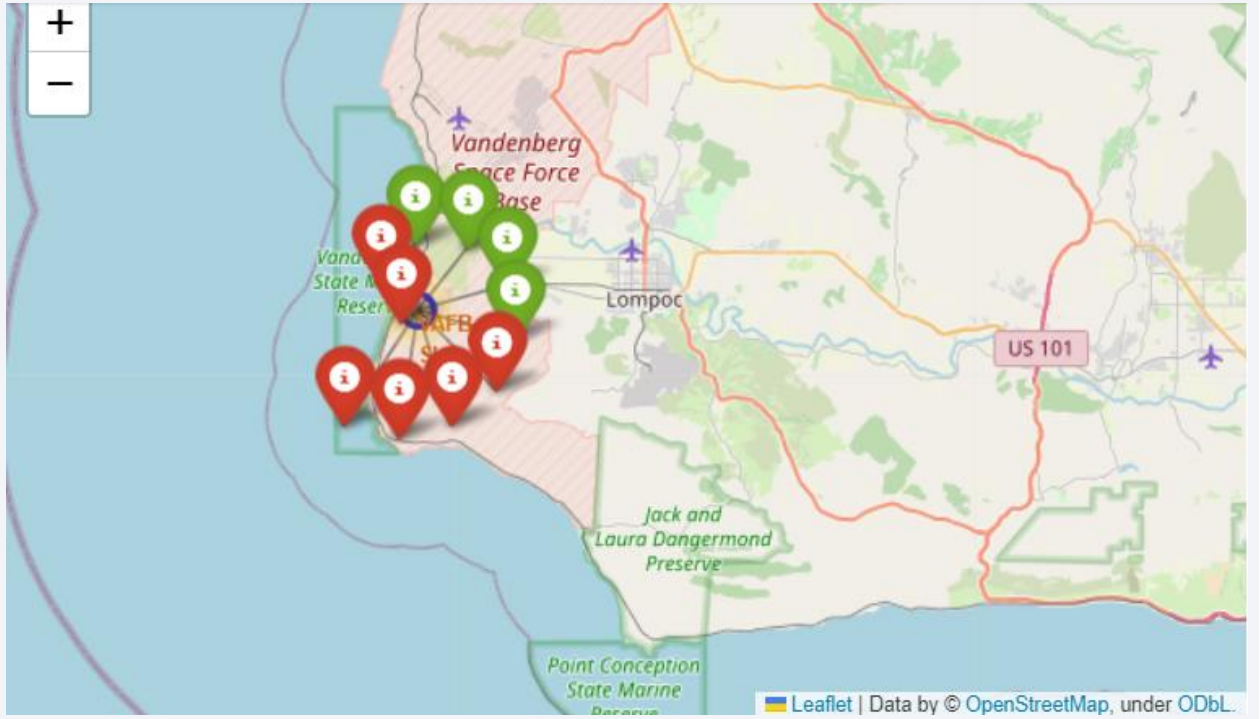
Map – Launch Sites Location



- The launch sites are located closer to the equator and in the vicinity of a coastline;

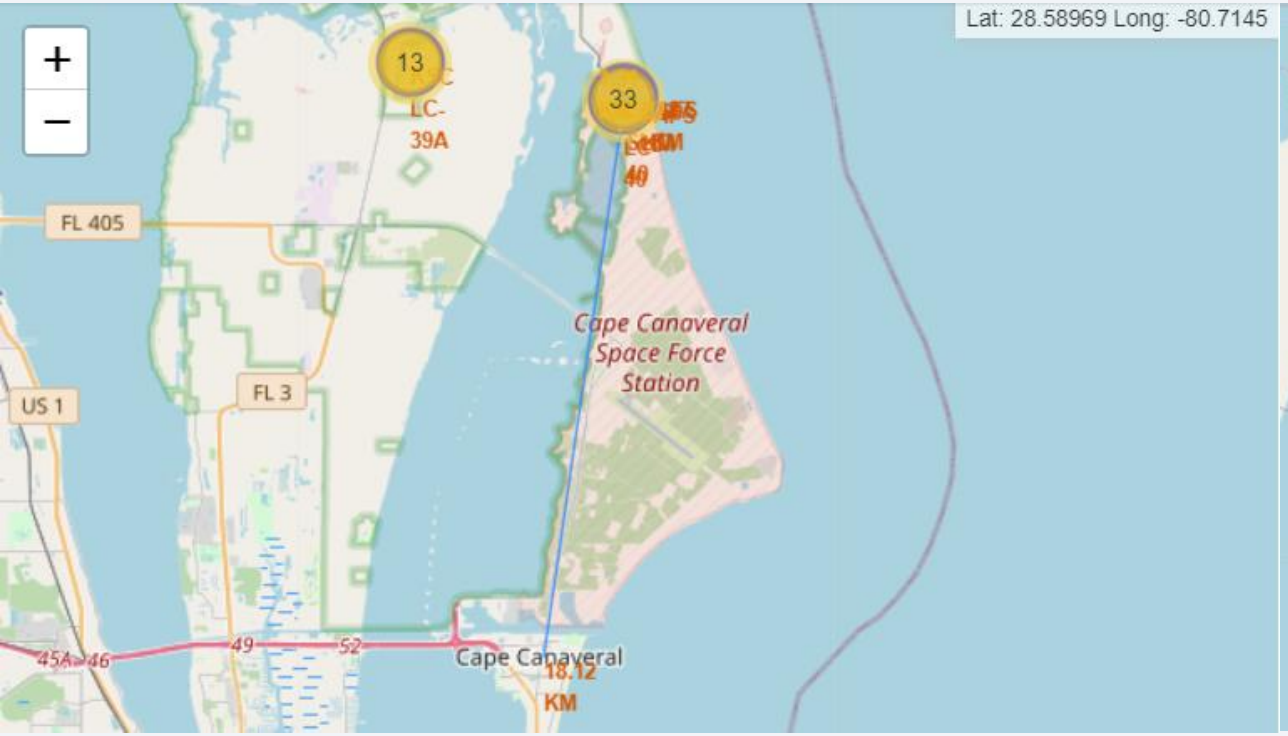
- Spacecrafts are launched from the equator to take optimum advantage of the earth's rotational speed and are launched near coastlines to ensure safety and facilitate landing;

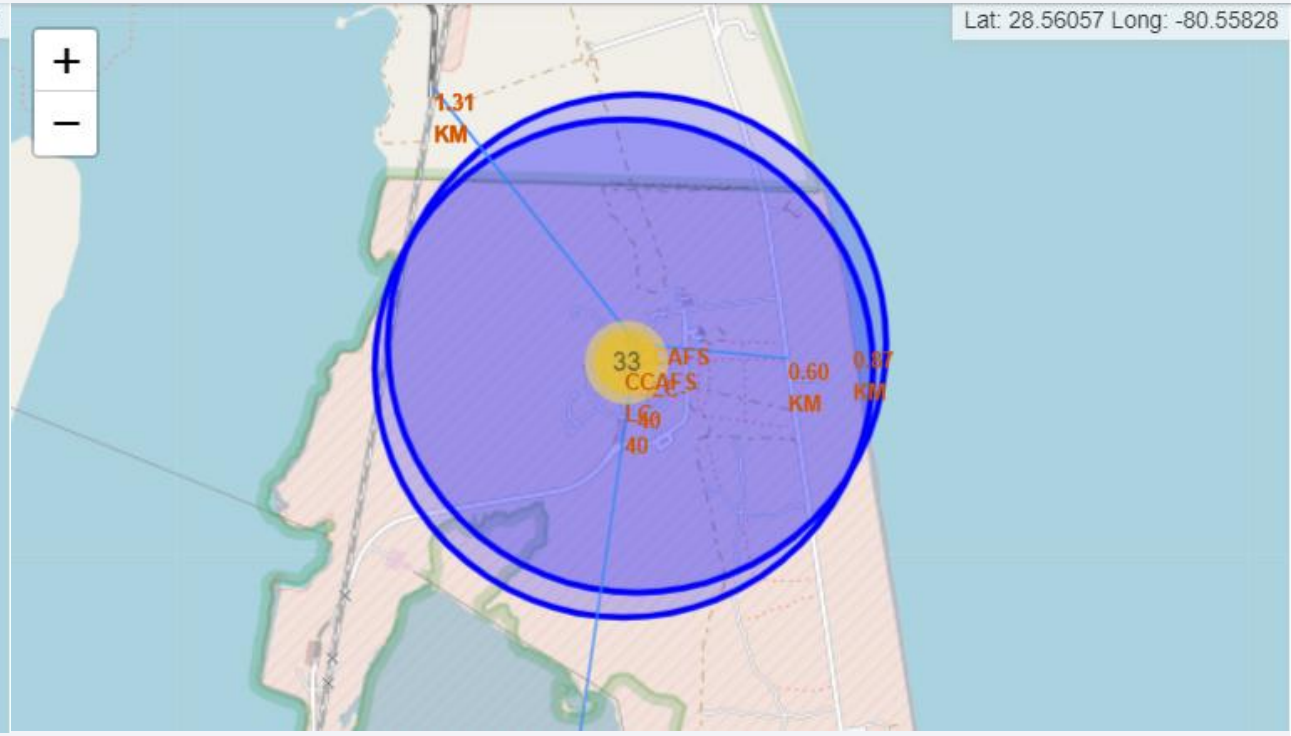
Map – Launches Outcome



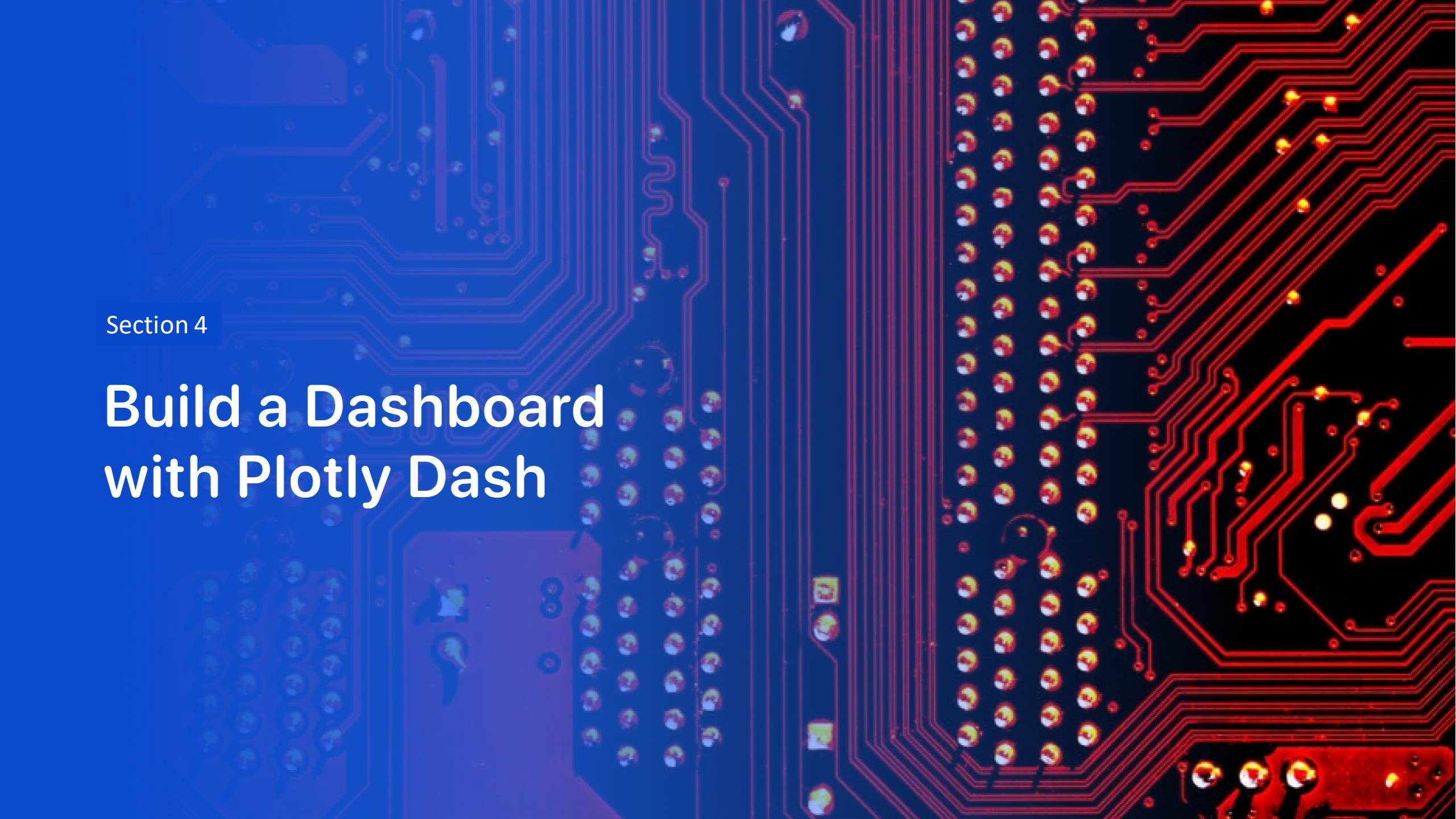
- We made a cluster indicating all the launches for the site (10 and 46) in the first fig;
- On click the cluster expands to show the landing outcomes (red=failure, green= success), presented in the second fig;

Map – Distance Between Sites Proximities

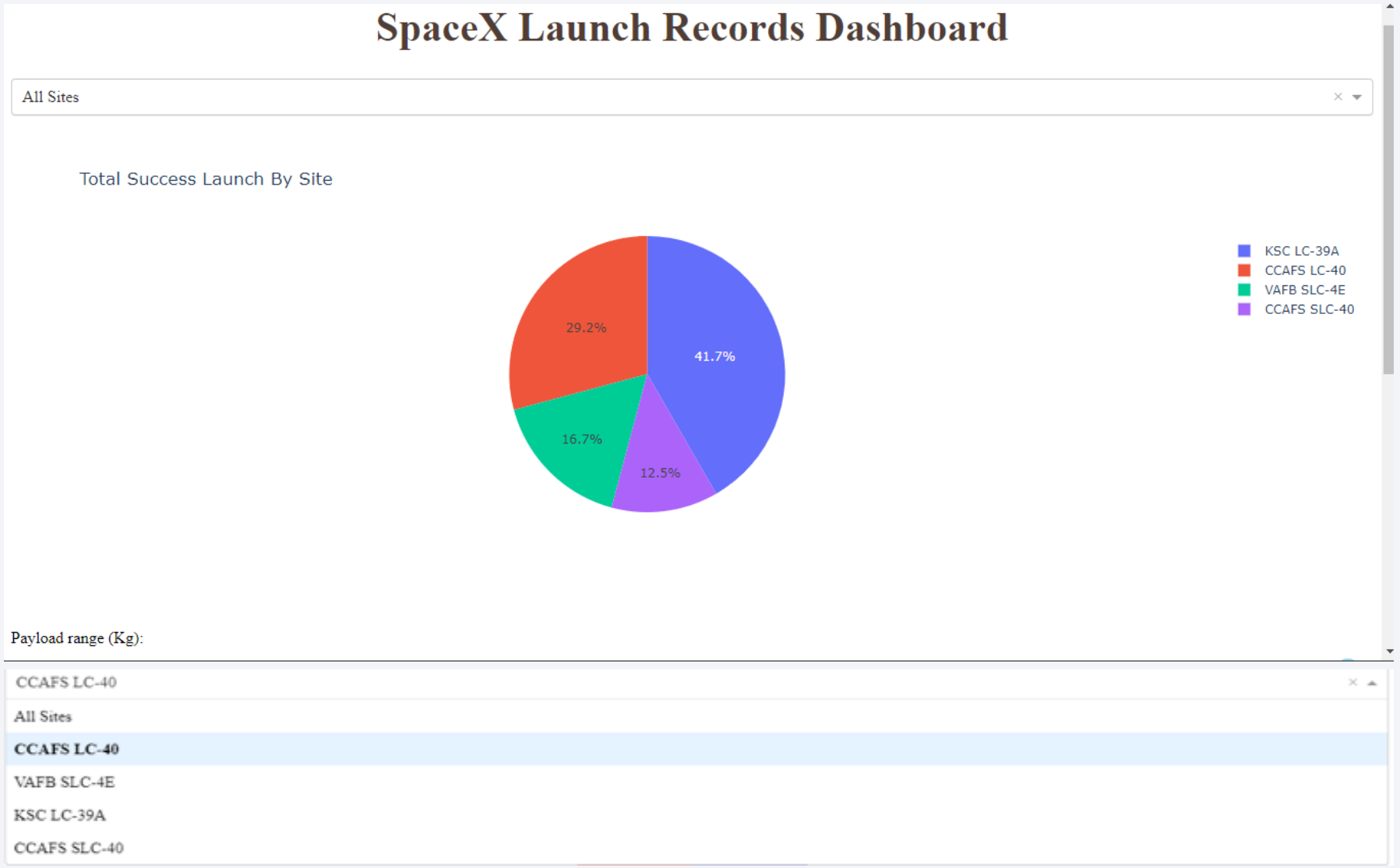




- The light-blue line indicates a qualitative measure of the site and it's proximities, the number in km is showed in a marker in orange;
- It's positioned near railways and highways to facilitate logistic, but are located far from city centers to safety reasons


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

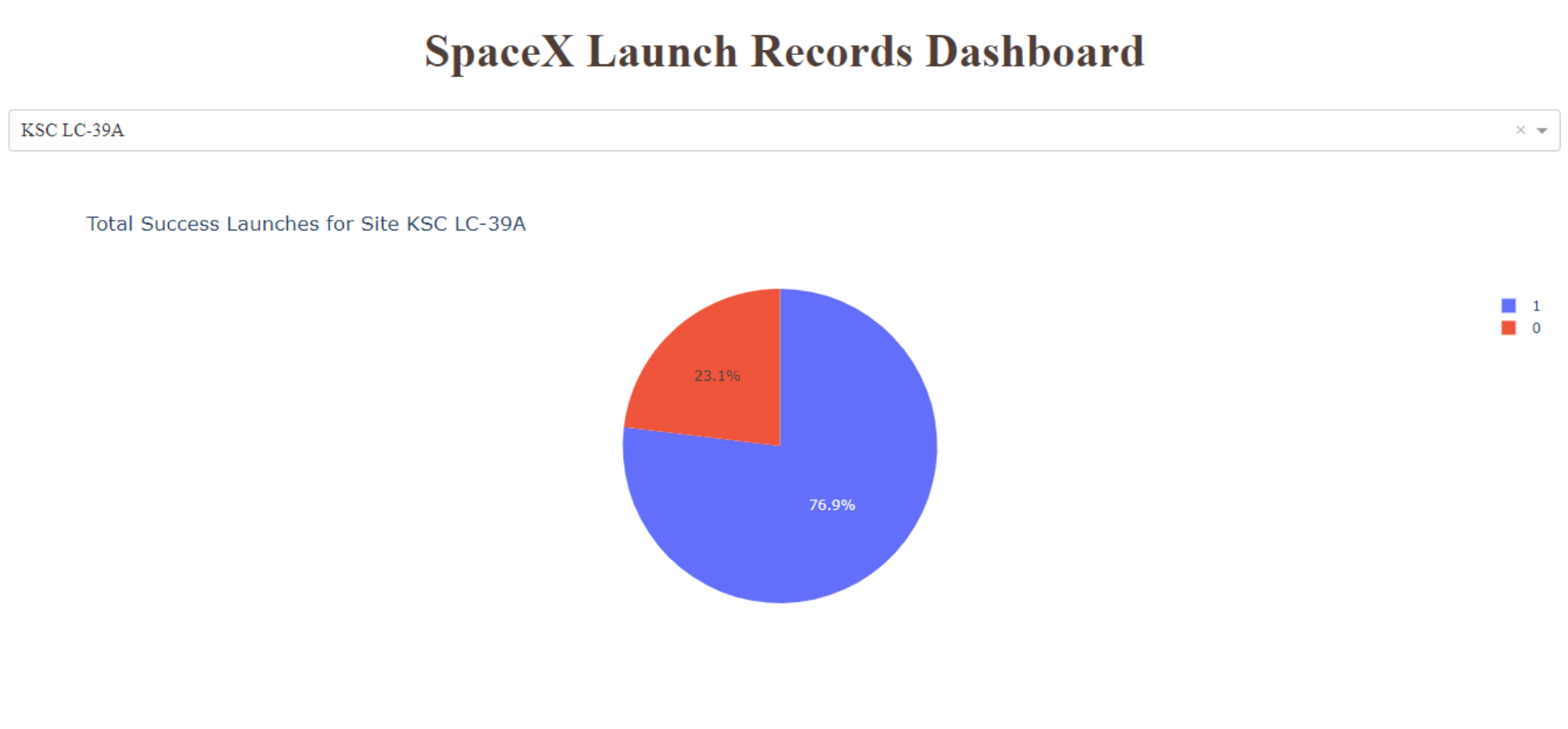


The pie chart display the proportion of Success Launches between different sites Sites, when the dropdown selection is "All Sites";

KSC-LC-39A correspond to 41% of the total successful landing (blue);

We can use the dropdown to retrieve specific data;

SpaceX Launch Records Dashboard - Filtered



- KSD LC-39A have the higher Success Rate which 76,9%;

SpaceX Launch Records Dashboard – Payload vs

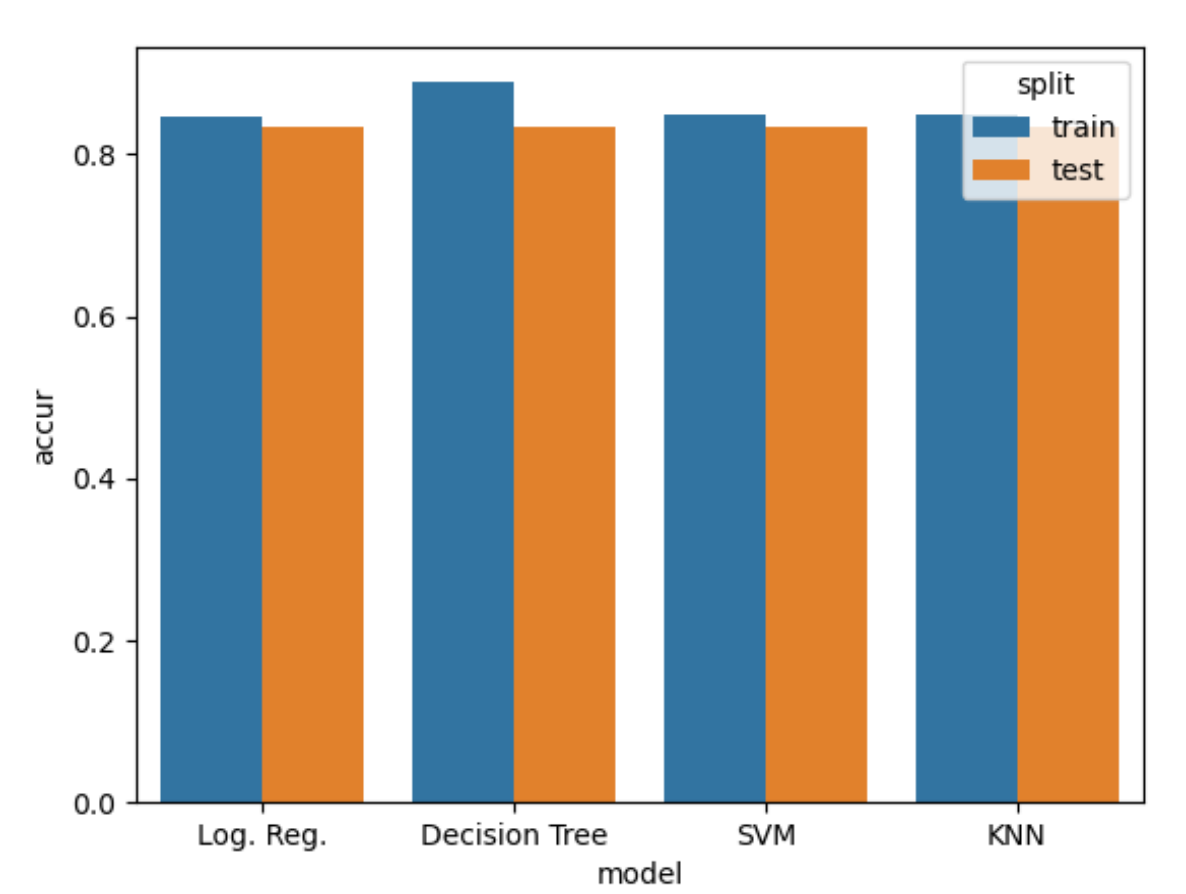
- Using the Slider, the range between 1000 and 5000 kg was selected. The scatter plot displayed show the relation between payload mass and class (landing success);
- The color indicate the booster version category for each launch;

Section 5

Predictive Analysis (Classification)

Classification Accuracy

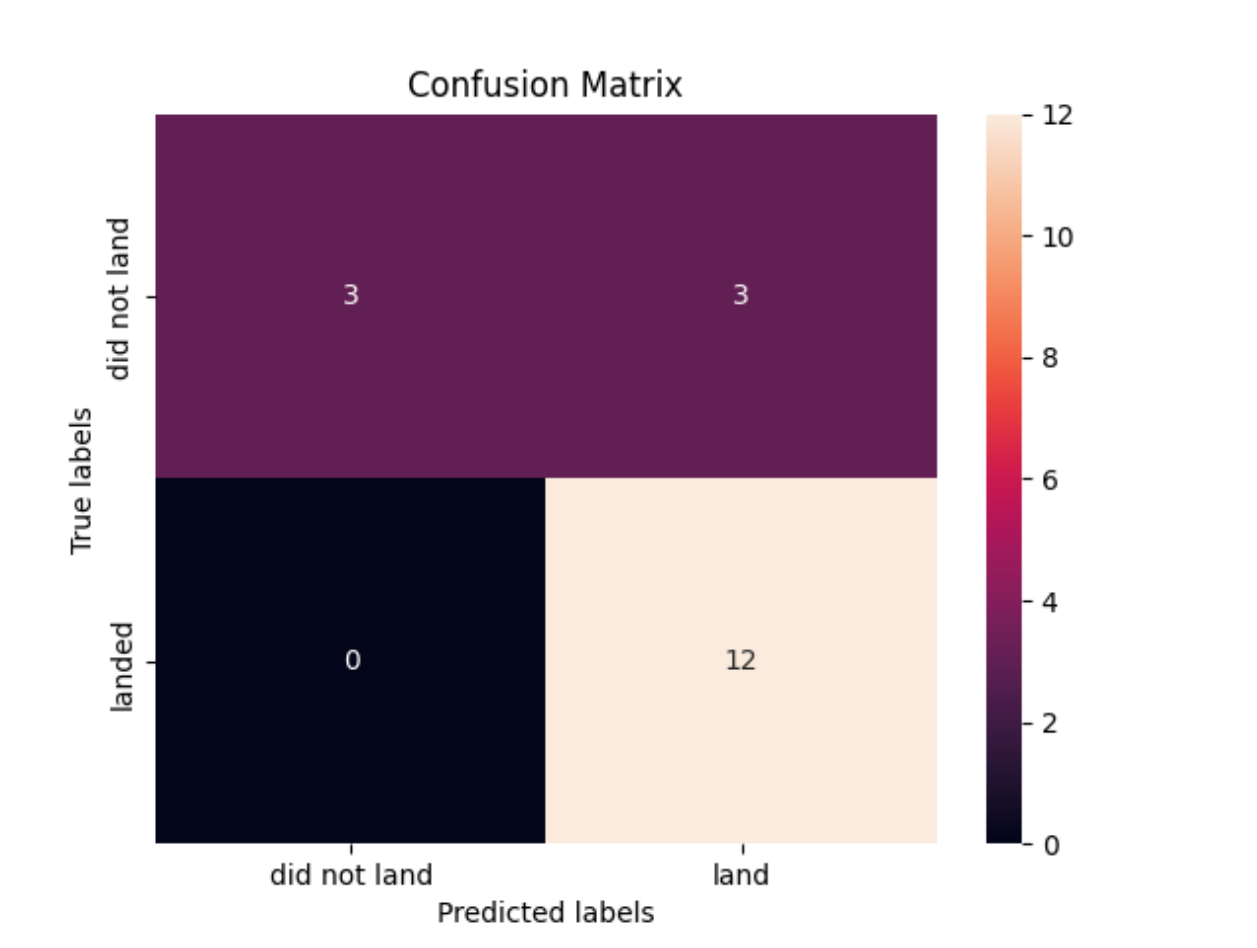
- Decision Tree Model – Best Accuracy



- The Decision Tree Method perform the best for the dataset;
- But, as already introduced in the results section, the difference was noticed only at the train set;
- This is likely due to the small test set size (small number of instances);
- It is important to remember that decision trees are prone to overfitting. See the Appendix.

Confusion Matrix – Decision Tree

- The x-axis present the predict label, the y-axis the true label for the test data. The color scale indicate the number of records in each quadrant.



- We can see that the model predict 12 success landings correct (right inferior corner)

- But predict 3 as "land" when the correct outcome was "did not land" – this is a false positive that can be ajusted in the model

Conclusions

- We could notice how the Flight Number, Payload Mass and Orbit influence the outcome:
  - Higher Flight Number (more recent) and higher mass presented better success rate;
- Through the maps, we can see that launch sites are located in proximity to the equator and near key transportation locations such as highways and railways.;
- A interactive view of the data allow us to look at the relation between Success Rate and Launch Site
  - Being KSD LC-39A the higher number of successes and better Success Rate
- Decision Tree model presented the best accuracy for the dataset;

Appendix

- If the hypothesis of overfitting due to a small number of instances is correct, we should expect to see high variance in the results for the accuracy and even for the confusion matrix as we rerun the model;
- Running the whole process again (Splip, Train, Test)

Model/ Rerun	Accuracy (Training Data)	Accuracy (Test Data)
Log. Regression	0.846	0.833
Dec. Tree	0.901	0.888
SVM	0.848	0.833
KNN	0.848	0.833

The values changed.... We can run it more times(next slide)

Appendix

1		2		3	
Accuracy (Test Data)	Accuracy (Training Data)	Accuracy (Test Data)	Accuracy (Training Data)	Accuracy (Test Data)	Accuracy (Training Data)
0.833	0.873	0.944	0.88	0.875	0.833

In the run 2, the test accuracy was greater the training one!

This is a strong indication of the effect of the small sample size provided for the test set.

- In most runs, the Decision Tree model performs better; however, it is important to note the effect that a small number of instances can have on the results.

Appendix

- All the code can be access in the follow repository : GitHub Repo

Thank you!