IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Camila A. Pareja H.

Outline

- Executive Summary
- Introduction
- Methodology
- EDA & Visualizations
- Results & Conclusion

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA With SQL
  - Interactive map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analytics (Classification)
- **Summary of all results**
  - EDA to understand results on visual form
  - Predictive Analytics to maximize launch success

Introduction

- In this project, our objective is to predict and analyze whether the Falcon 9 from SpaceX will land successfully or not based on previous results. A successful land not only means ”job well done”; but also, a huge advantage among competitors. SpaceX cost per launch estimates at 62 million dollars compared to 165 million dollars. Because SpaceX can reuse the first stage, the cost per launch is a lot less than its competitors. Predicting success, can therefore, provide the correct information for bidding against SpaceX rocket launches.

- Because a successful launch is our goal, understanding our data correctly is key because it contains answers to all of our problems. Therefore, we must ask ourselves: What variables relationships influence a successful launch? Are there any patterns we should be aware of? Is our dataset efficient or do we need more information/cleaning? Among others.

Section 1

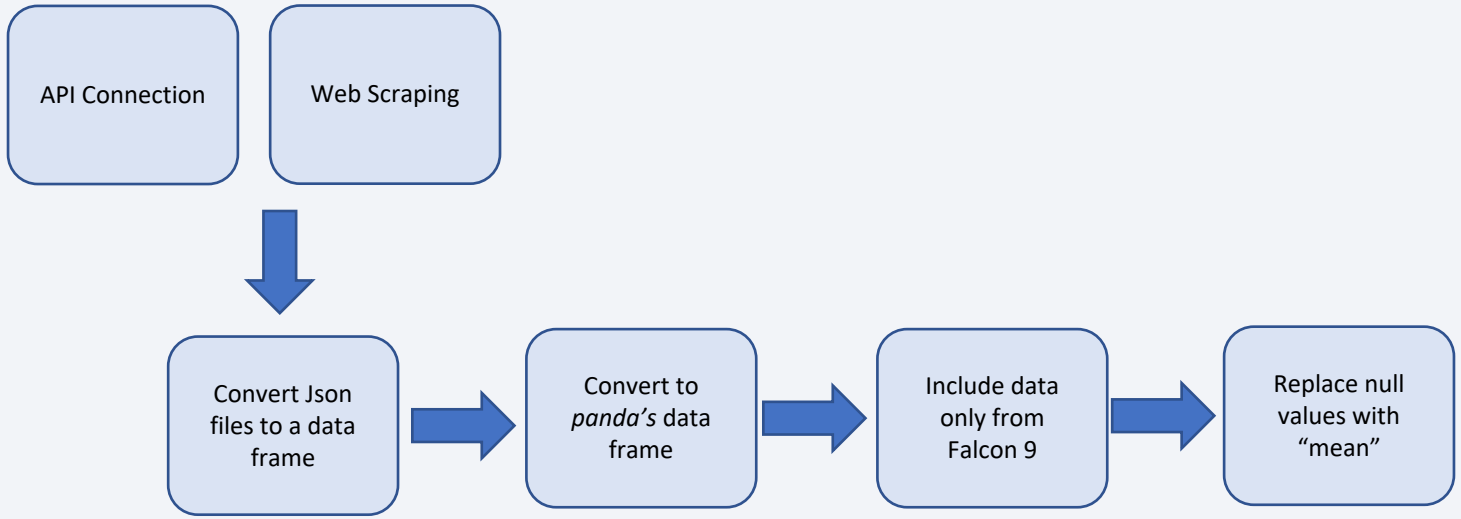
Methodology

Methodology

- Data collection:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Data wrangling:
  - One Hot Encoding for Machine Learning
  - Drop irrelevant columns
  - Replace Null Values
  - Filter 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

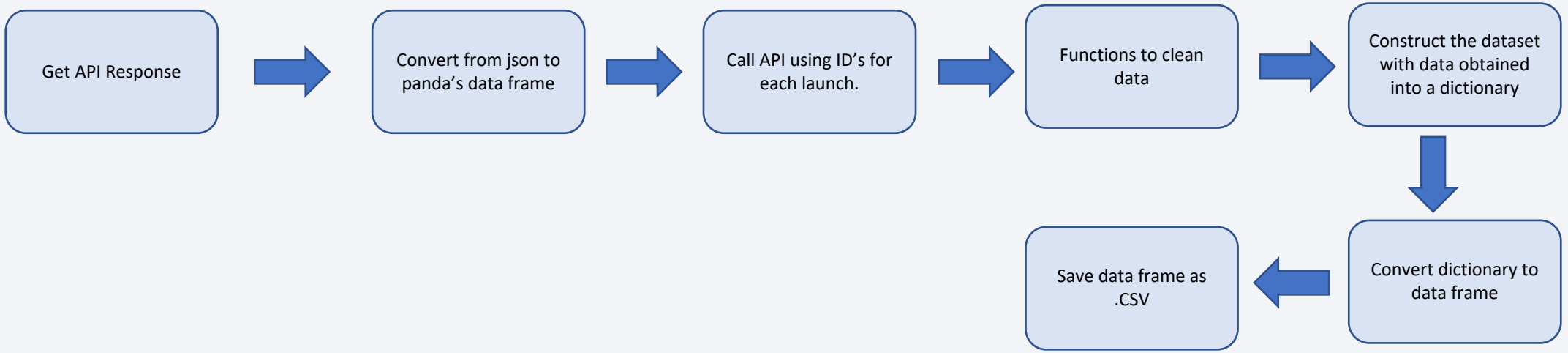
Data Collection

- We worked with two external datasets:
  - SpaceX REST API (data about launches in general, payload, landing specifications, outcome, etc.)
  - Web Scraping from Wikipedia (data about launches in general, launch site, Orbit, Customer, Outcome, etc.)



Data Collection – SpaceX API

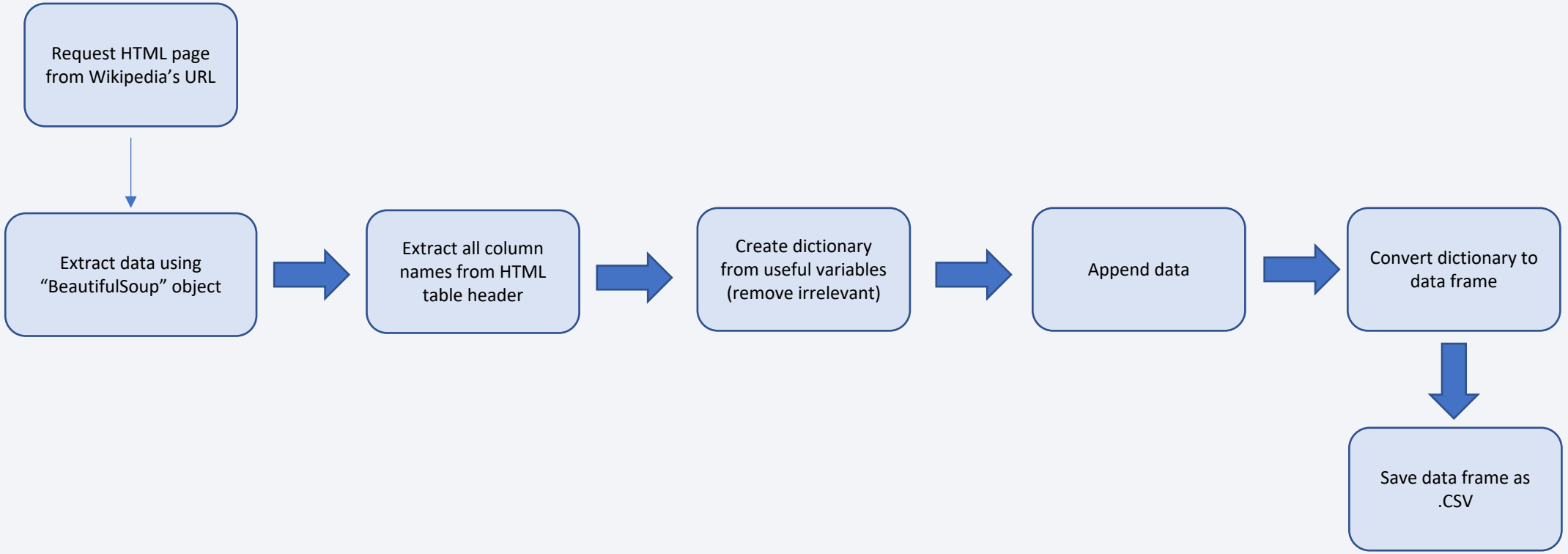
- Since we are requesting data from an API, we use the function “requests.get” to pull specific information from SpaceX into our dataset. The process is the following:



GITHUB URL

Data Collection - Scraping

- Additional information is needed; therefore, we pull specific information from Wikipedia by scraping their data.



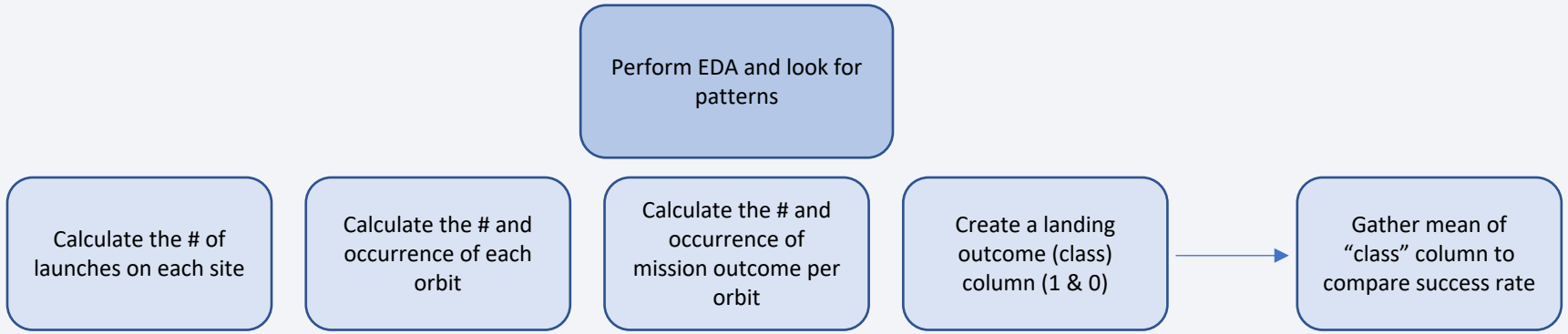
GITHUB URL

Data Wrangling

- During this process, we use EDA to find patterns in our data that will provide valuable information to later train our supervised models. This data contains several different cases about unsuccessful and successful landings and the reasons of these. In order to understand the landing outcomes more clearly, we need to understand certain categories (Found in table 1) and covert these outcomes for better analysis into 1 if the outcome was successful and 0 if it was unsuccessful.

Table 1

True Ocean	False Ocean	True RTLS	False RTLS	True ASDS	False ASDS
Successfully landed to a specific region in the ocean	Unsuccessfully landed to a specific region in the ocean	Successfully landed to a ground pad	Unsuccessfully landed to a ground pad	Successfully landed on a drone ship	Unsuccessfully landed on a drone ship

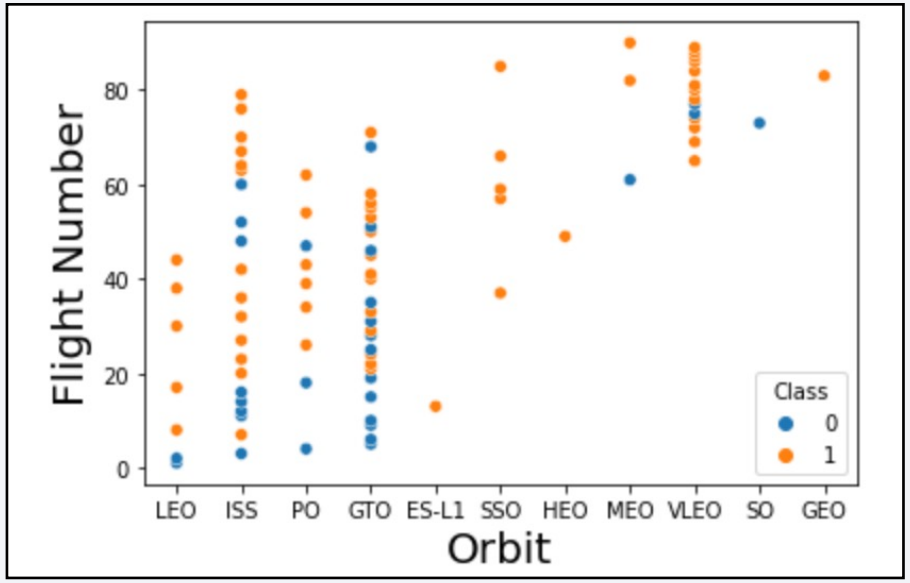


GITHUB URL

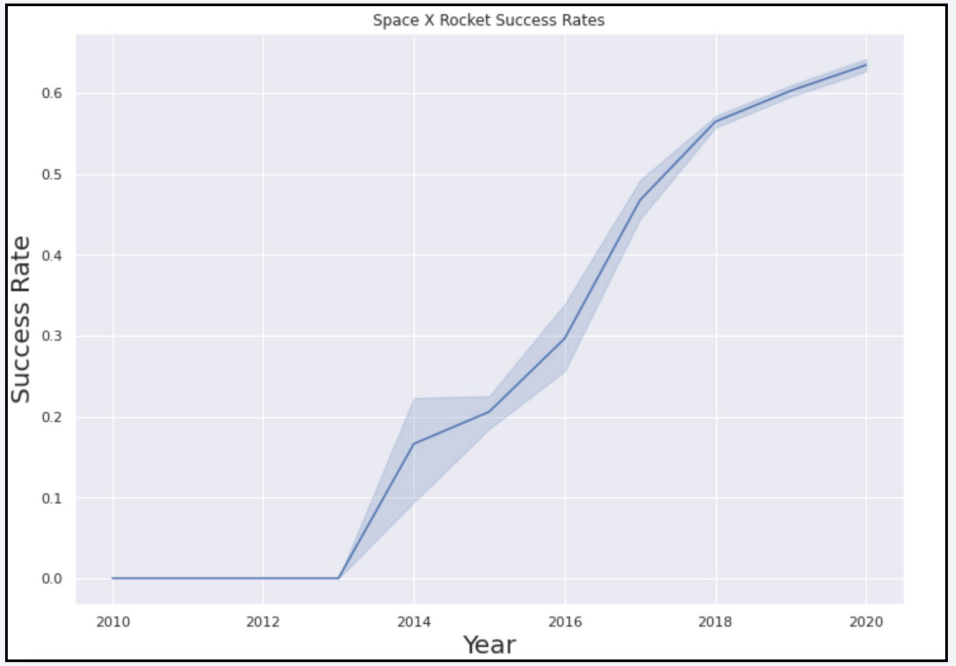
EDA with Data Visualization

- There were three types of visualizations used for EDA: Scatter plots, bar plot, and a Line Chart.

Scatter Plots: By picking any two variables, we can see their correlation and how they are affected by each other. In the example below, we can see the relationship between variable "Orbit" and variable "Flight #" from which we can notice a good relationship between Orbit LEO and flight # and poor relationship with Orbit GTO and flight #.

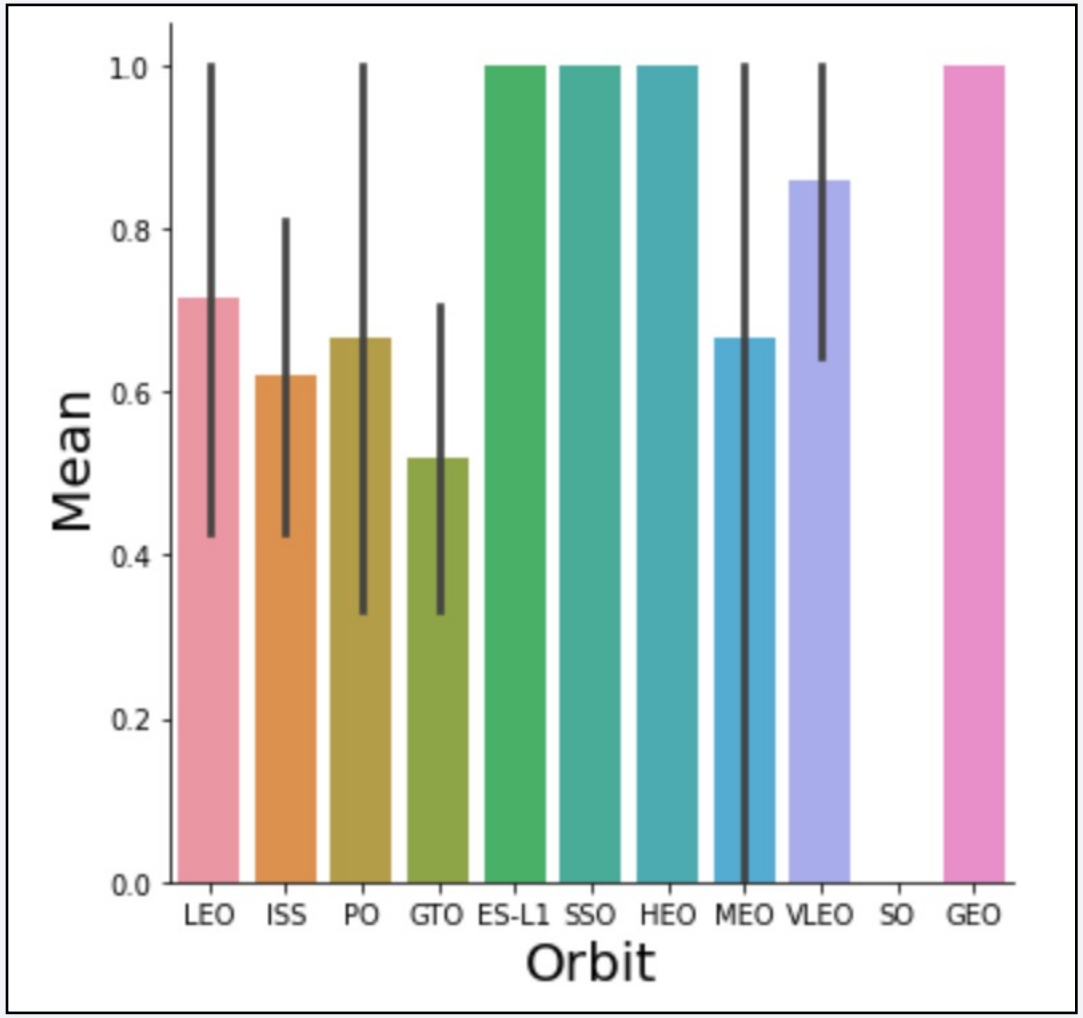


Line Chart: By using a line chart, we can easily understand trends and form predictions for future decisions. As seen below, we can determine that the success rate has been improving every year consistently since 2013.



EDA with Data Visualization (Cont.)

Bar Chart: Similar to previous visuals, a bar chart can help us understand the relationship that exists between two variables. In this case, we can conclude that Orbits GEO, HEO, SSO, and ES-L1 have the highest success rates. Whereas Orbit GTO has the lowest success rate.



GITHUB URL

EDA with SQL

- Examples below are some of the results gathered by using SQL queries to pull information from our data table:
  - Unique launch sites
    - CCAFS LC-40, CCAFS SLC-40, CCAFSSLC-40, KSC LC-39A, and VAFB SLC – 4E
  - Total payload mass carried by boosters launched by NASA
    - 45,596 Kg
  - Average payload mass by booster version F9 v1.1
    - 2,928Kg
  - Date from successful landing outcome in drone ship
    - June 5, 2016
  - Names of boosters with successful ground pad landing with a payload mass >4000 but >6000
    - F9 FT B1032.1, F9 B4 B1040.1, F9 B4 B1043.1
  - Total number of successful and failure outcomes
    - Success -> 100
    - Failures -> 1
  - Count of successful landing outcomes between June 2010 and March 2017
    - 34

Interactive Map with Folium (Cont.)

- In order to show the locations of each launch site in our map, we use the longitude and latitude coordinates of each site and receive a map with three yellow circles that show the exact locations.

Interactive Map with Folium (Cont.)

- We can go a bit further into detail and use these locations to measure distances between these launch sites. For example, we can measure the distances to highways, railways, coastlines, among others, and see which are at a closer proximity from these launch sites. On this map, we can see that there is a closer distance to a coastline at 0.90km from CCAFS SLC-40 launch site.

GITHUB URL

Dashboard with Plotly Dash

- Dashboards are an easy and effective way to show results. They are great sources of information that can be understood by tech and not tech professionals. Python is a great tool to build easy and interactive dashboards. In this case, we can provide our results with two of the most used visuals within a dashboard:
  - **Pie Charts:** Since we are interested in analyzing successful and unsuccessful launches, pie charts provide us with a simple visual to either compare all sites at once or focus individually on each site and analyze their successful and unsuccessful outcomes. Because we only have a few sites, having a pie chart does a wonderful job providing results.
  - **Scatter Plots:** With this type of visual, we can compare the relationship/correlation between two variables. We can easily understand certain patterns and it is a straightforward source of information. Because we have several booster versions, having a scatter plot helps us visualize all of them at once and compare their outcomes to see which ones are doing better than others.

GITHUB URL

Predictive Analysis (Classification)

- Building the model:

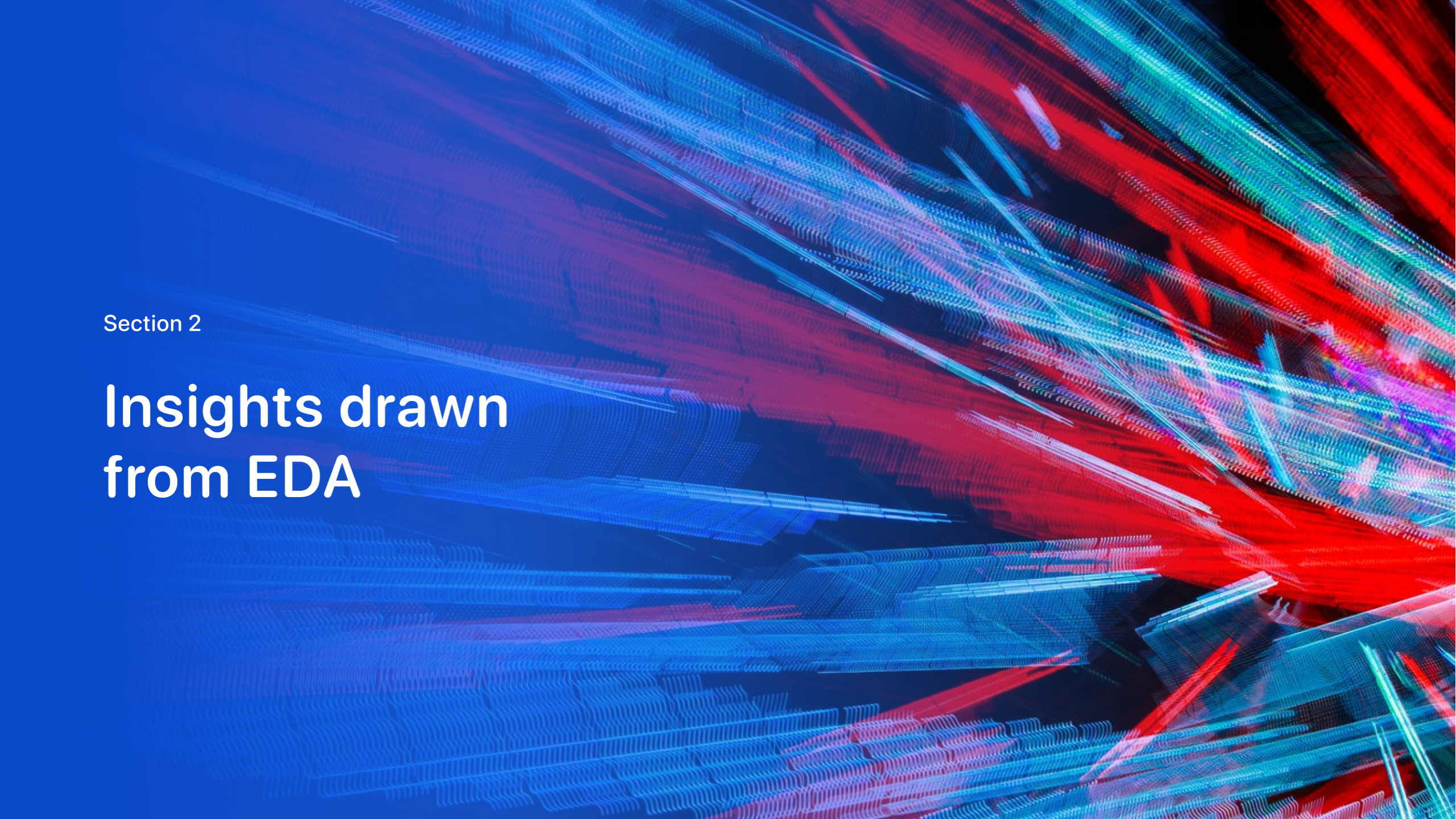
Load the data and import packages	Transform the data	Split data into training and testing sets	Test Machine learning algorithms	Set parameters and algorithms to GridSearchCV	Train our datasets

- Evaluating the model:

Verify accuracy scores	Tune hyperparameters for each type of algorithm	Plot Confusion Matrix

- Improving the model:
  - Use feature engineering & tuning
- Finding the best performing classification model:
  - Models were pretty similar; they all provided a good accuracy score. Therefore, it would be easy to choose as either would do the job just fine. However, if we had to go with one, the model with best accuracy score from all would be the chosen one.

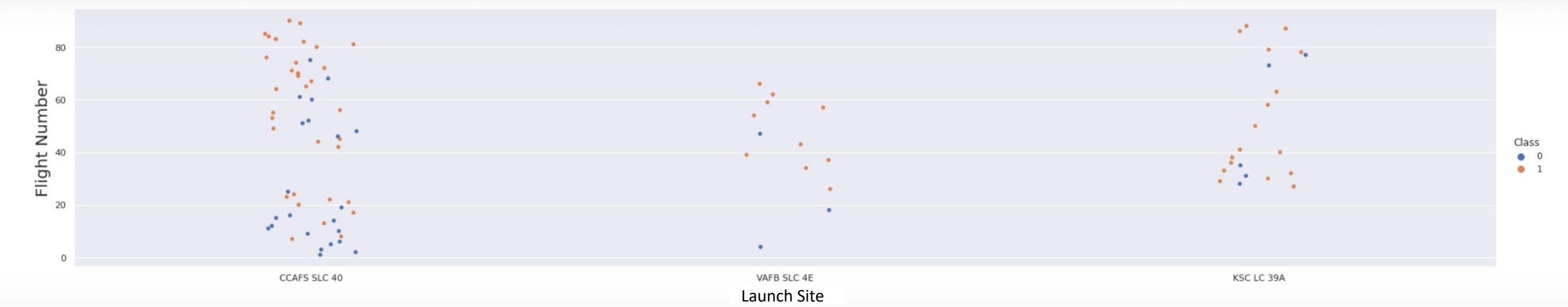
GITHUB URL



Section 2

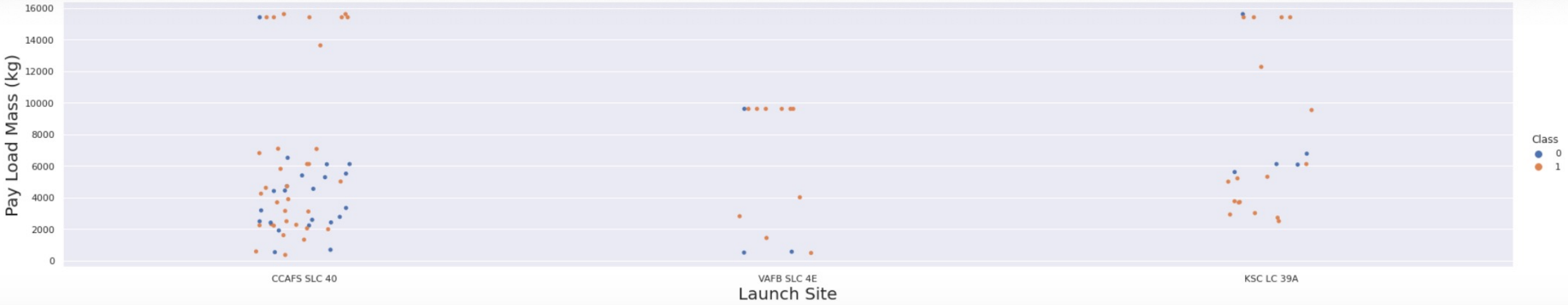
Insights drawn from EDA

Flight Number vs. Launch Site



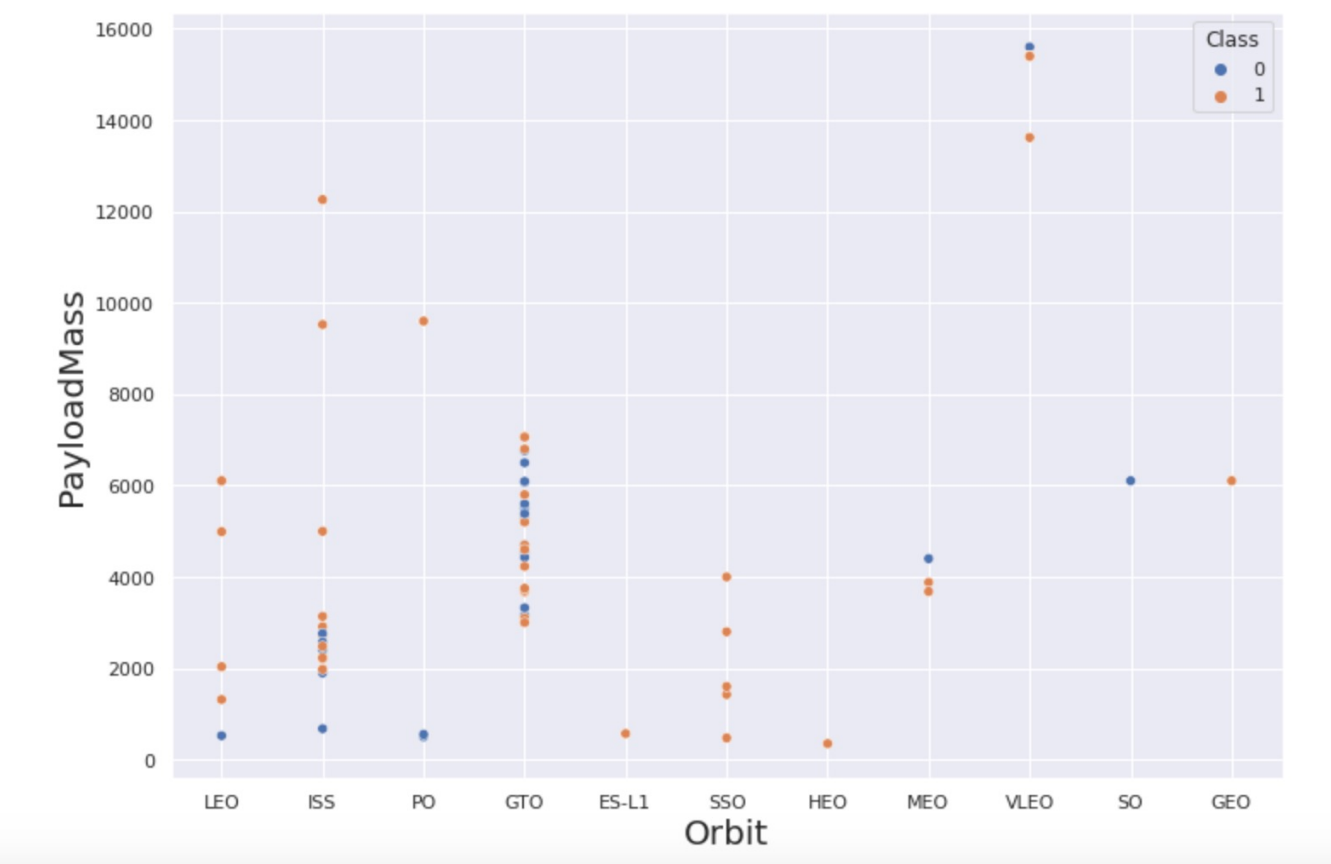
Although launch site CCAFS SLC 40 has the majority of the flights, KSC LC - 39A launch site has
Received a better success rate at 77% total. However, it is clear that the more flights launching from a site, the
Better chances of a good outcome.

Payload vs. Launch Site



For launch site CCAFS SLC-40, we notice that the higher the payload mass the higher the chances of a successful outcome. However, for launch site KSC LC-39A, it seems to be the opposite.

Payload vs. Orbit Type



We observe that heavy loads have a negative effect on GTO Orbit. However, we see a positive effect on SSO and ISS Orbits. On the other hand, Orbit SSO seems to have a positive effect on lower pay loads instead.

All Launch Site Names

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

By using “DISTINCT” in our SQL query, we pull only unique Names within a specific column (Launch_Site). In this case, we receive The information shown in the left table as a result.

Total Payload Mass

	Total Payload Mass
0	45596

By using the function ”SUM” within our column named “Payload_Mass_KG” (Which contains the weight from each booster) we can receive the total Number of kilograms or total payload mass. If we want to pull this information from a specific Customer, we would use the “WHERE” clause to call only the information from a specific source/customer.

Average Payload Mass by F9 v1.1

	Average Payload Mass
0	2928

By using the function ”AVG” within our column named “Payload_Mass_KG” and A “WHERE” clause filtering only F9 v1.1 “booster_version”, we receive the average weight of 2,928kg.

First Successful Drone Ship Landing Date

Date which first Successful landing outcome in drone ship was acheived.
06-05-2016

By using the function ”MIN” within our “Date” column and the “WHERE” clause to filter only a Successful outcome from ground drone ship we receive the result of June 5th, 2016, as the first Successful landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

By selecting the booster version, along with a “WHERE” clause filtering to a successful landing outcome from “drone ship” And weight between 4000kg and 6000kg (Payload_Mass_KG”) we receive the three entries above.

Total Number of Successful and Failure Mission Outcomes

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

Because we need to go a bit in more detail within our data, we use subqueries to pull the correct information as follows:
SELECT(SELECT Count(Mission_Outcome) from SpaceX where Mission_Outcome LIKE '%Success%' as Successful_Mission_Outcomes,
(SELECT Count(Mission_Outcome) from SpaceX where Mission_Outcome LIKE '%Failure%' as Failure_Mission_Outcomes.

To understand this query a bit better, we break it down as follows:

SpaceX -> our table with data
Mission_Outcome -> 1 if it was successful and 0 if it was unsuccessful
LIKE '%Success%' -> Pull data only containing "success" outcomes from Mission_Outcome column (1)
LIKE '%Failure%' -> Pull data only containing "failure" outcomes from Mission_Outcome column (0)
Successful_Mission_Outcomes -> New column name for positive outcomes
Failure_Mission_Outcomes -> New column name for negative outcomes

Boosters Carried Maximum Payload

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows × 2 columns

By using the “DISTINCT” function we pull unique booster versions from our data. In order to receive only those that have the highest weight, we include the function “MAX” from ”Payload_mass_kg” and create a column named “Maximum Payload Mass” to show our results. Because we also want to have a list that goes from heavier to lighter weight, we include “ORDER BY” and “DESC” so our table starts with the heavier boosters first as seen on the Left side table.

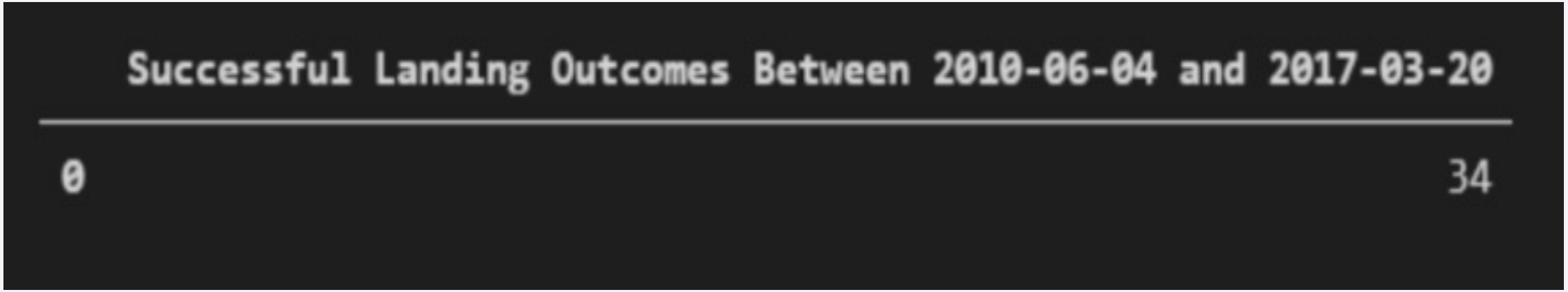
2015 Launch Records

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

The table on the left-hand side is the result of a combination query that pulls data only from a specific year, In this case 2015. We achieve this by firstly verifying if our date field is saved as a date format and continue to by calling the date, booster version, launch site, and landing outcome.

As see on this table, we only filter to show successful outcomes from this specific year in order starting from January to December.

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



Successful Landing Outcomes Between 2010-06-04 and 2017-03-20	
0	34

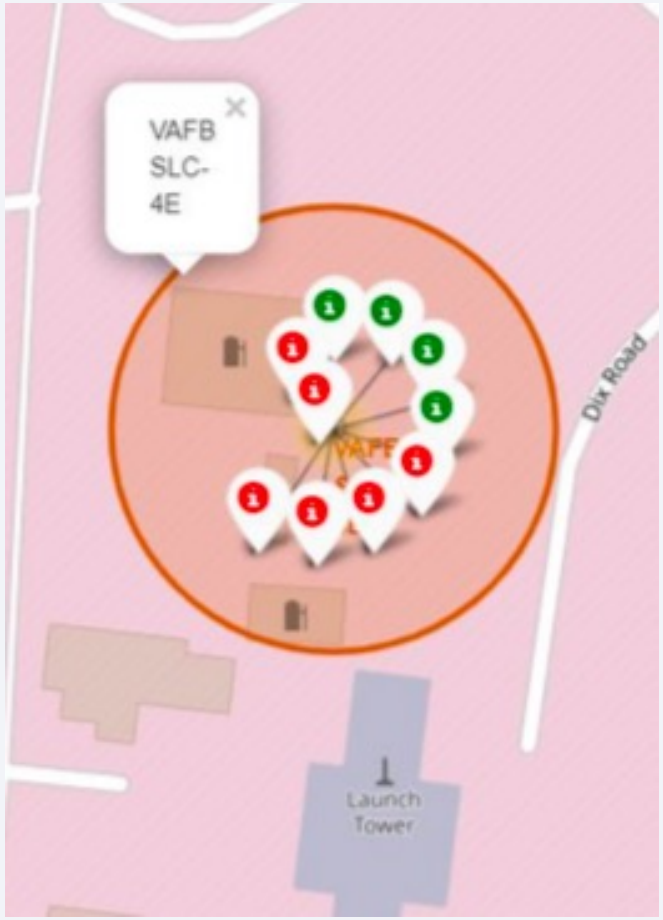
We receive the results above, by counting the successful outcomes within a specific time frame from June 4th, 2010 To March 20, 2020, from which we receive 34 successful landing outcomes.

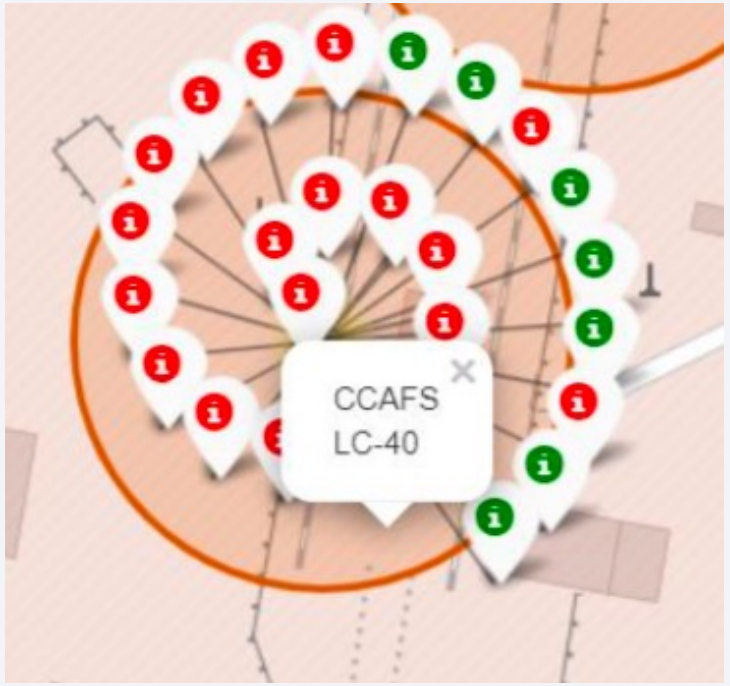
Section 4

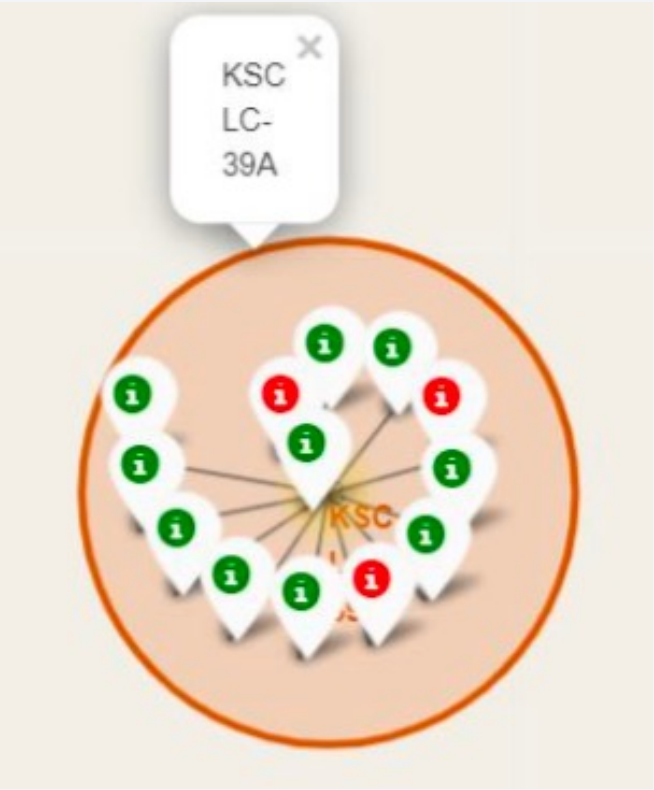
Launch Sites Proximities Analysis

Successful & Unsuccessful launches

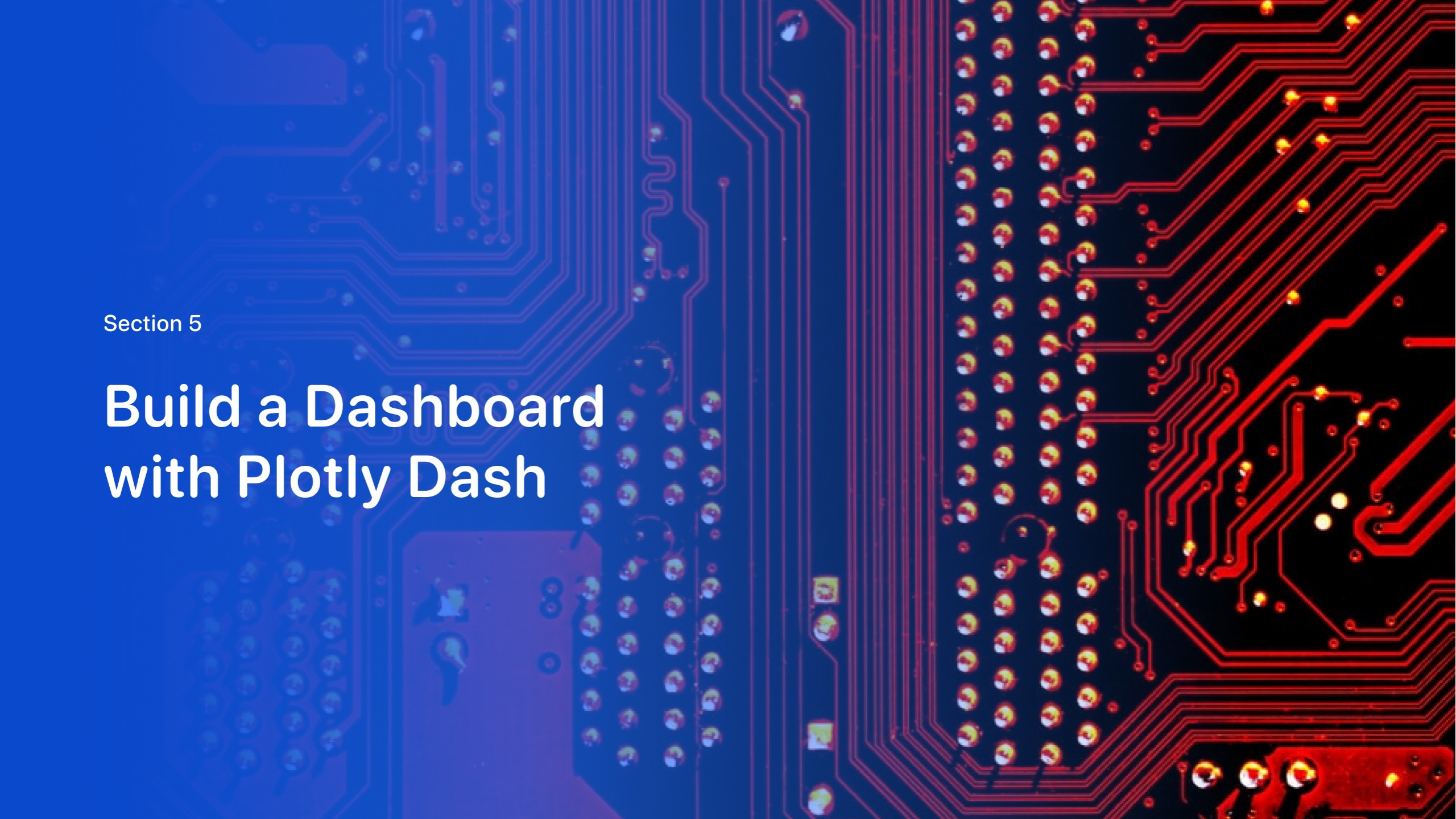
Green dots show successful launches; Whereas red dots show unsuccessful ones.







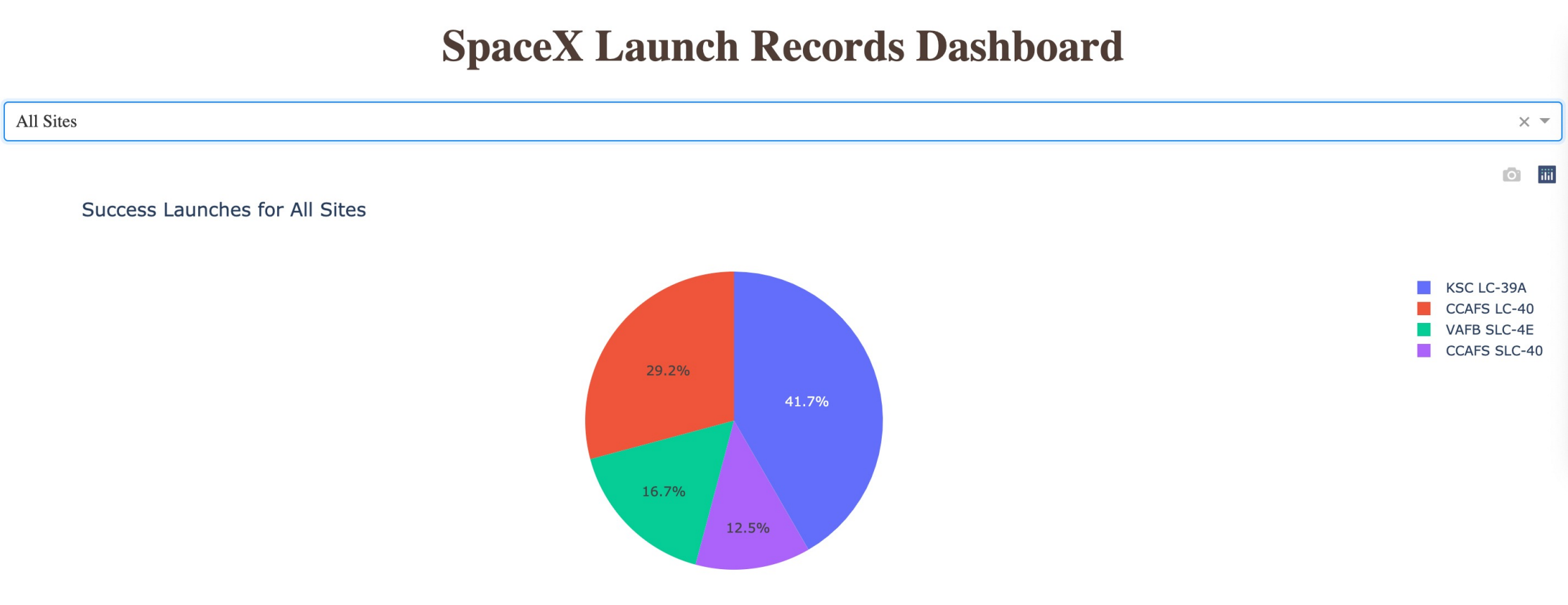
*More on Slide 14 & 15



Section 5

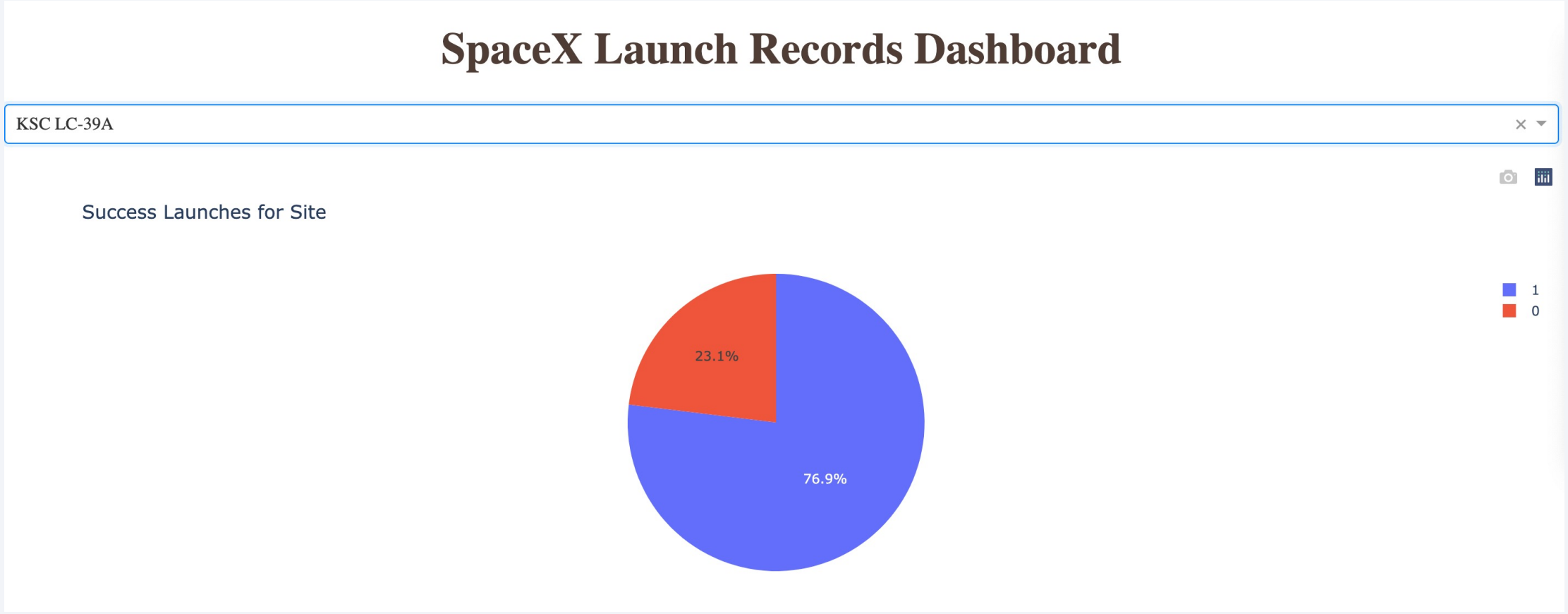
Build a Dashboard with Plotly Dash

Total Successful Launches by Sites



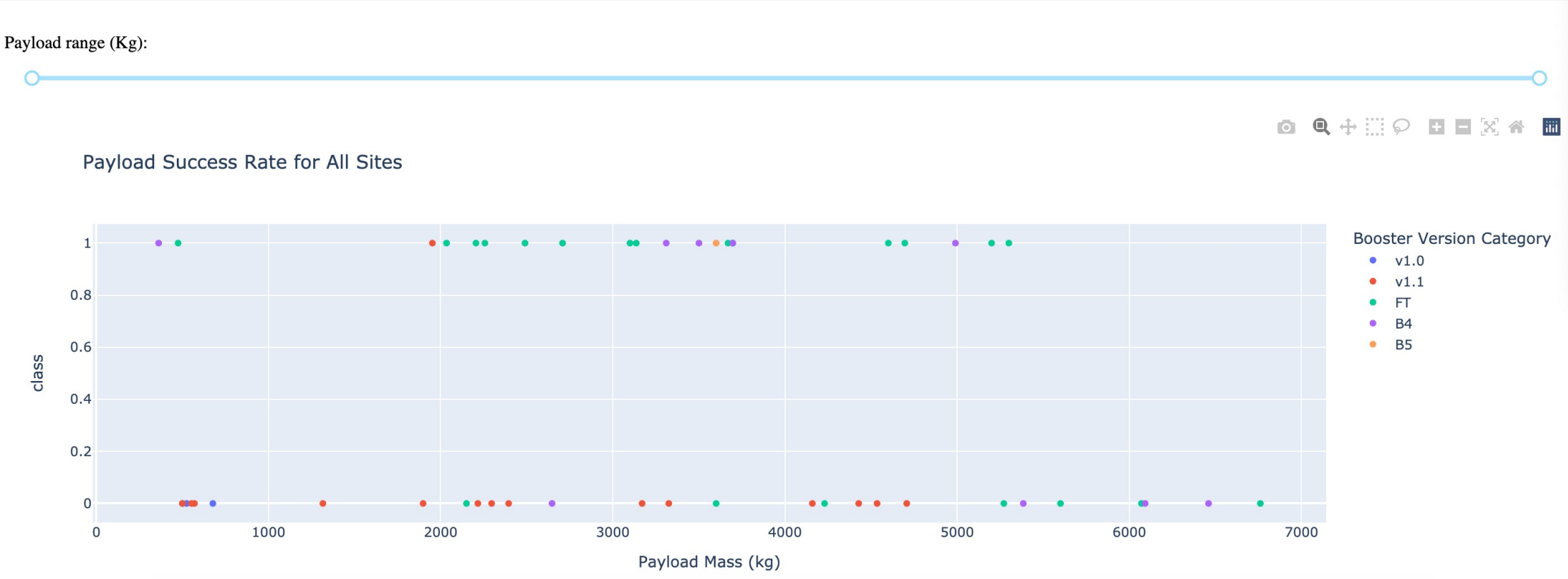
Site KSC LC-39A had the most successful launches and CCAFS SLC-40 had the least.

Site KSC LC-39A



Site KSC LC-39A achieved a 76.9% success rate and a 23.1% failure rate

Payload Mass vs Launch Outcome for all sites



Payload vs Launch Outcome for all sites with different weight ranges

Weight from 0kg to 5,000kg

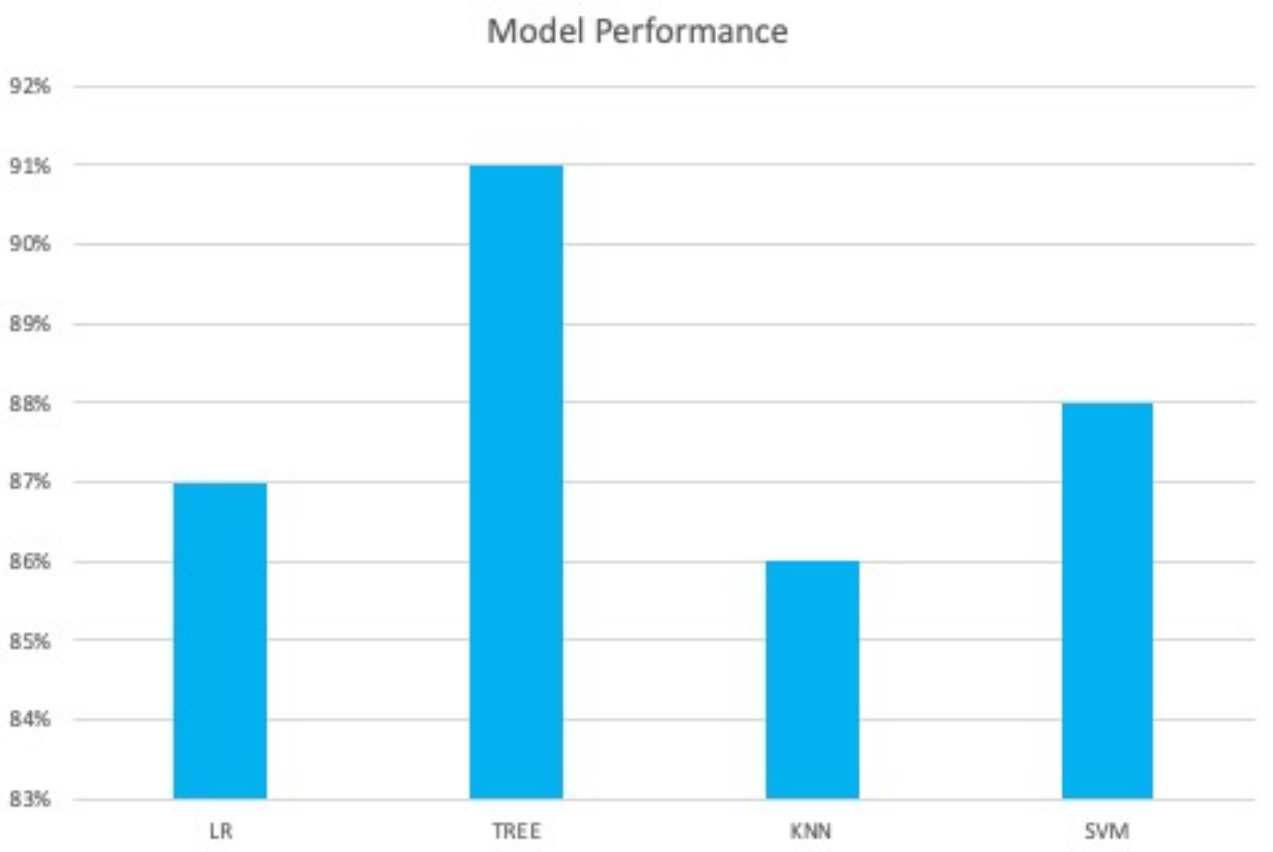
Weight from 5,000kg to 10,000

We can clearly observe comparing these two visuals, that boosters that are lighter have a higher success rate than those that are heavier.

Section 6

Predictive Analysis (Classification)

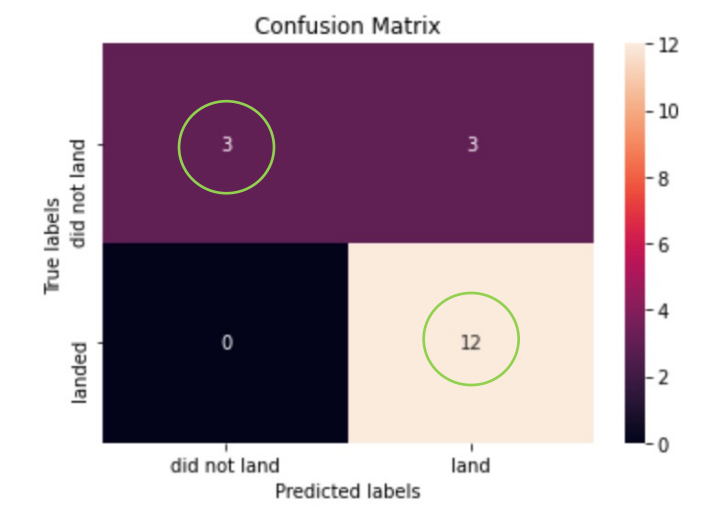
Classification Accuracy



All models performed really well. However, as it is clear in our bar chart the model that performed best was the tree classification model with an accuracy of 91%

Confusion Matrix

Confusion Matrix results from tree classification model



Example to understand the confusion matrix

(Analytics Steps by Neelam Tyagi)

From 18 predicted lands, the confusion matrix from our winner model shows that from 18, 12 of those that were predicted as landed actually landed. However, from 6 that did not land, 3 were predicted correctly and the other half were not.

Results & Conclusion

- Exploratory data analysis & Visualizations
  - After performing a deep EDA and visualizing our results, we can conclude the following:
    - Orbits with best success rates were GEO, HEO, SSO, and ES-L1
    - The launch site with the most successful outcomes was KSC LC-39A with a success rate of 77%
    - Since 2013, outcomes have been improving consistently
    - Low weighted boosters have a higher success rate
- Predictive analysis:
  - All models performed well, with accuracy scores above 86%. However, the model that performed best is the Tree Classifier as shown below:
    - **Model's performance:** SVM -> 88%, **Tree -> 91%,** knn -> 86% , lr -> 87%

Thank you!