IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Ravi>
<Sept 25 2022>

Outline

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

Executive Summary

Summary of methodologies

Data collection steps include REST API's and Web Scraping

Data Wrangling methods include converting classification variables into binary values or unit vectors via one-hot-encoding

Null values were dealt with by replacing them with either the mean or 0

SQL queries and scatter plots were used to initially investigate the data to get the important features for predictive analysis

Classification models were used to make predictions in the end

Summary of all results

Features such as Orbit, Launch Site, Flight Number, and more seemed to be the most relevant

All classification models performed at about 83%, with the error mostly lying in false positive detection

Introduction

The project is about looking at SpaceX data to predict and reduce cost of rockets.

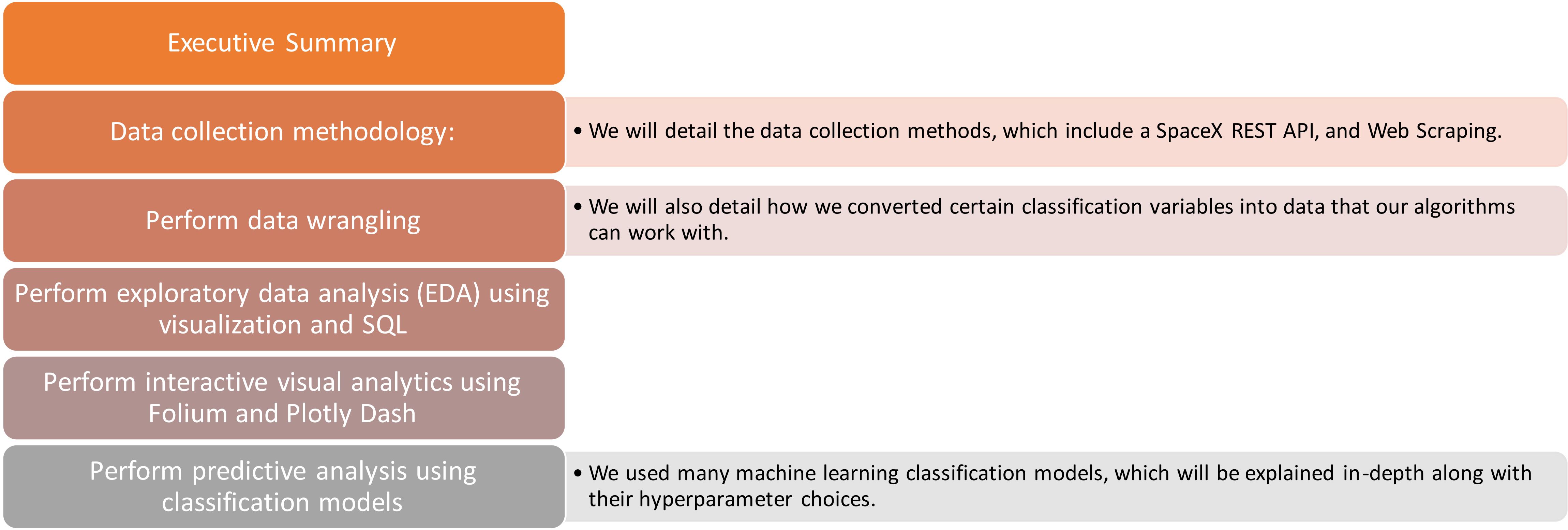
Rockets that can land safely in the first stage of launch significantly reduce cost, so this is an important metric that will be considered and predicted.

But ultimately, the goal is to determine the price of a rocket launch using machine learning methodologies rather than rocket science.

Section 1

Methodology

Methodology



Executive Summary

Data collection methodology:

- We will detail the data collection methods, which include a SpaceX REST API, and Web Scraping.

Perform data wrangling

- We will also detail how we converted certain classification variables into data that our algorithms can work with.

Perform exploratory data analysis (EDA) using visualization and SQL

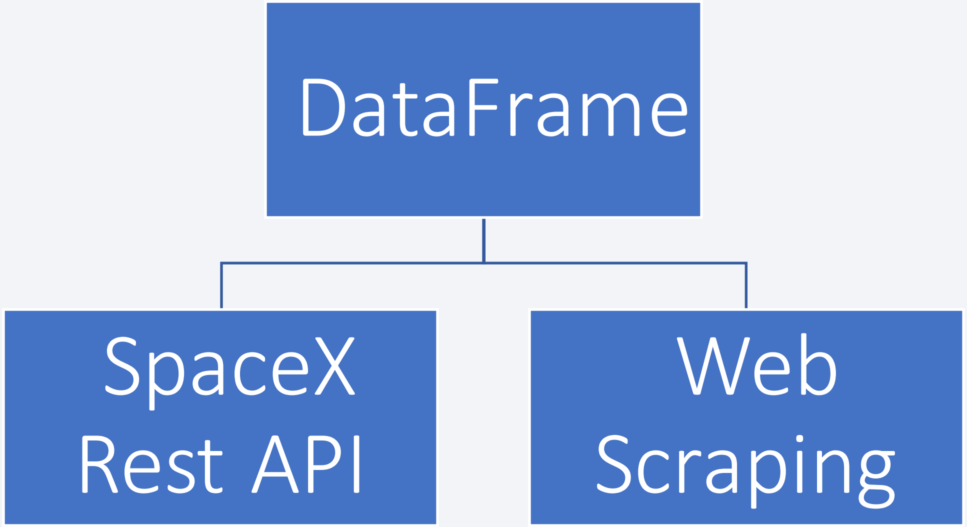
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- We used many machine learning classification models, which will be explained in-depth along with their hyperparameter choices.

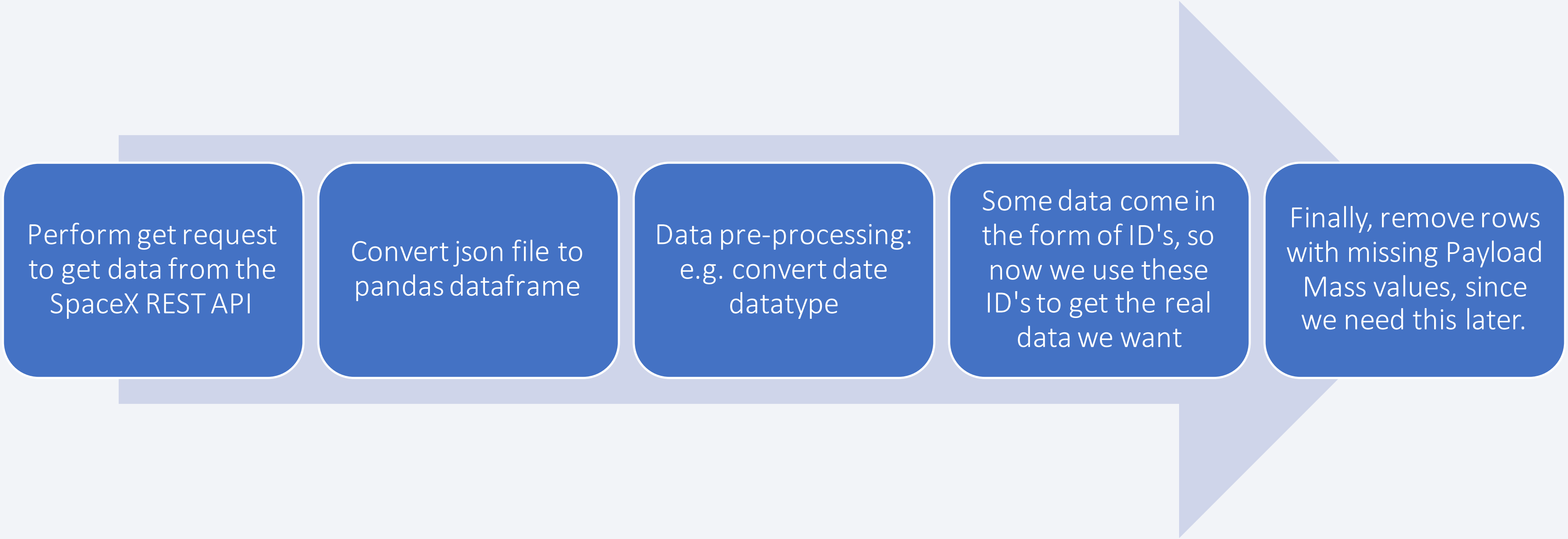
Data Collection

- We use two different methods to collect the data, namely a REST API and Web Scraping, both explained in more detail later.
- After the data is collected we then move onto data wrangling.



Data Collection – SpaceX API

- Data collection with SpaceX REST API. Here is a GitHub link to the code: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/jupyter-labs-spacex-data-collection-api.ipynb.



Perform get request to get data from the SpaceX REST API

Convert json file to pandas dataframe

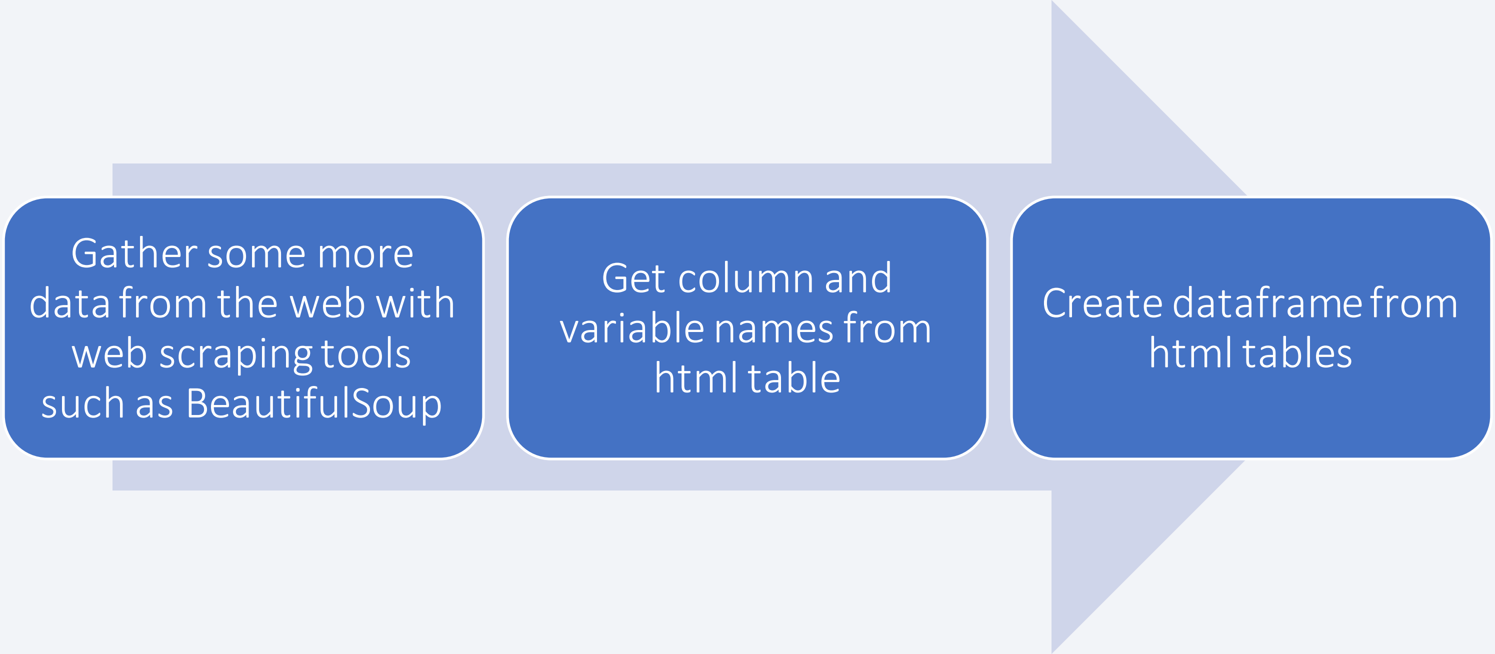
Data pre-processing: e.g. convert date datatype

Some data come in the form of ID's, so now we use these ID's to get the real data we want

Finally, remove rows with missing Payload Mass values, since we need this later.

Data Collection - Scraping

- Web scraping process. Link to code here via GitHub: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/jupyter-labs-webscraping.ipynb.



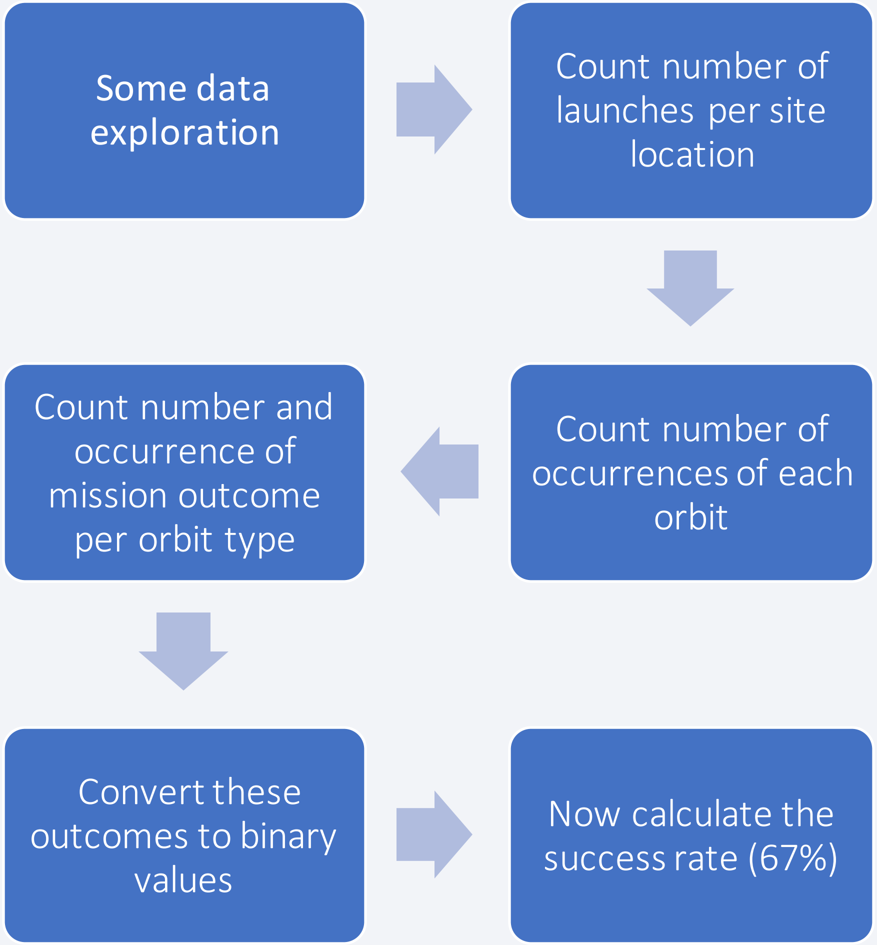
Gather some more data from the web with web scraping tools such as BeautifulSoup

Get column and variable names from html table

Create dataframe from html tables

Data Wrangling

- Data was processed with some basic data exploration and then converting an important class to binary values.
- GitHub link to code: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb.



Some data exploration

Count number of launches per site location

Count number of occurrences of each orbit

Count number and occurrence of mission outcome per orbit type

Convert these outcomes to binary values

Now calculate the success rate (67%)

EDA with Data Visualization

- Data was for the most part explored via scatter plots. The data points are also colored to indicate if the outcome was successful or not.
- We looked at various plots, comparing the features: FlightNumber, Orbit, Payload Mass, and Launch Site.
- The plots vary from the x axis and y axis being any two of these features, to help see if there was any correlation or information in these various combinations.
- For example, there is a Launch Site vs Payload Mass plot that shows that rockets with Payload Mass about 14000 with launch site CCAFS SLC 40 have mostly successful outcomes.
- See this GitHub link for details: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/jupyter-labs-eda-dataviz.ipynb.

EDA with SQL

Some SQL queries were performed to explore the data:

- Displaying names of unique launch sites
- Displaying first 5 records beginning with CCA in Launch Site name
- Getting total Payload Mass from NASA (CRS)
- Average Payload Mass from booster version F9 v1.1
- Date of first successful landing outcome in ground pad
- Booster Version with successful drone ship landing and payload mass between 4000 and 6000
- Total number of successful and failed mission outcomes
- Booster Version with max payload mass
- Failure landing outcomes in drone ship from year 2015
- Ranking the number of succesful landing outcomes from april 2010 to march 2017

GitHub link: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb.

Build an Interactive Map with Folium

Used Folium to ad circles and markers to show successful and failed launch sites.

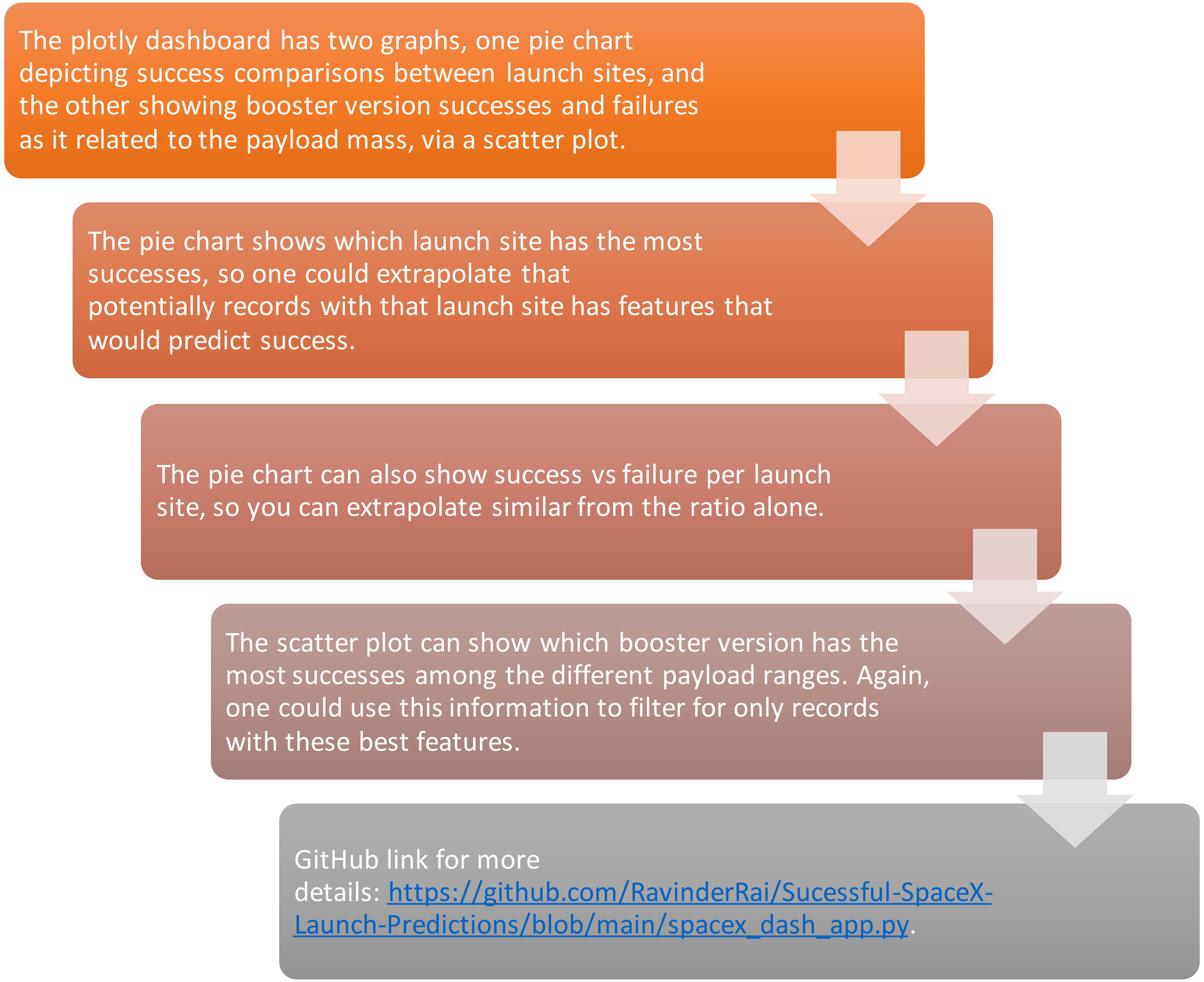
This gives a clear view of where they are, revealing potential information as to what type location launching sites frequent.

Since most of these launches were done in the same place, a marker cluster was added to click on to view all launches in a small radius.

Furthermore, some lines were added to show the distance from the launch site to the water or a city. The difference in length of these lines should indicate where launch sites typically take place on a map.

GitHub link to code: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/lab_jupyter_launch_site_location.ipynb.

Build a Dashboard with Plotly Dash



The plotly dashboard has two graphs, one pie chart depicting success comparisons between launch sites, and the other showing booster version successes and failures as it related to the payload mass, via a scatter plot.

The pie chart shows which launch site has the most successes, so one could extrapolate that potentially records with that launch site has features that would predict success.

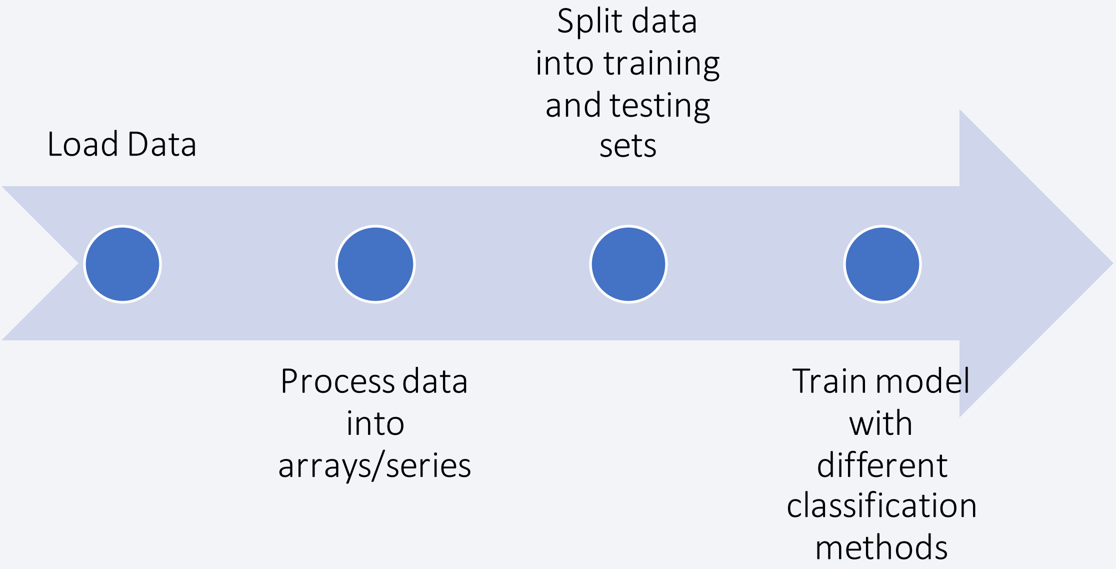
The pie chart can also show success vs failure per launch site, so you can extrapolate similar from the ratio alone.

The scatter plot can show which booster version has the most successes among the different payload ranges. Again, one could use this information to filter for only records with these best features.

GitHub link for more details: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions/blob/main/spacex_dash_app.py.

Predictive Analysis (Classification)

- Tested various classification models and used accuracy metric to evaluate them
- GitHub link for more details.

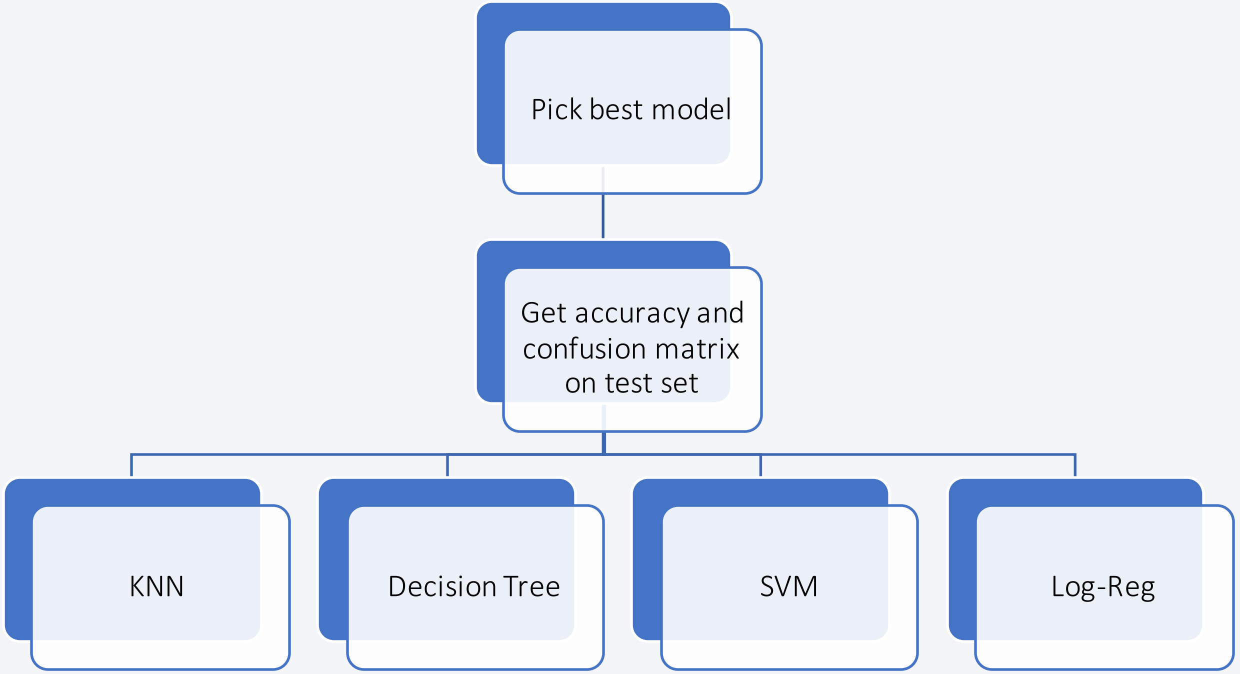


Load Data

Process data into arrays/series

Split data into training and testing sets

Train model with different classification methods



Pick best model

Get accuracy and confusion matrix on test set

KNN

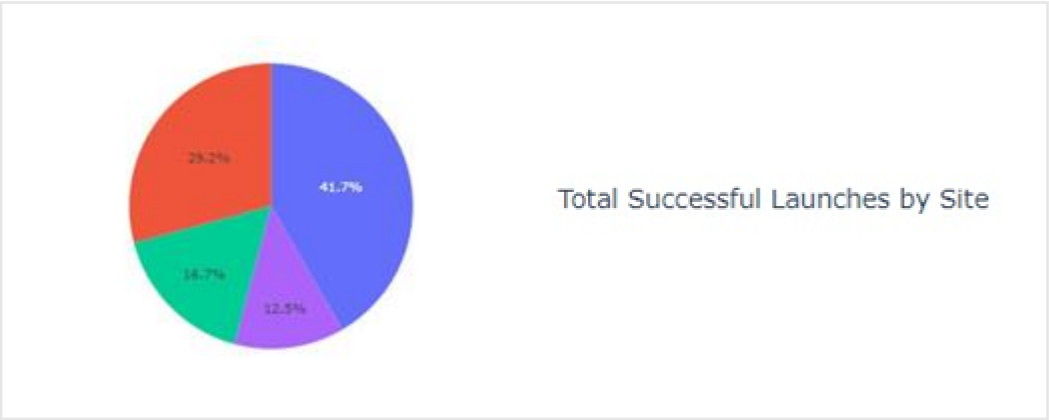
Decision Tree

SVM

Log-Reg

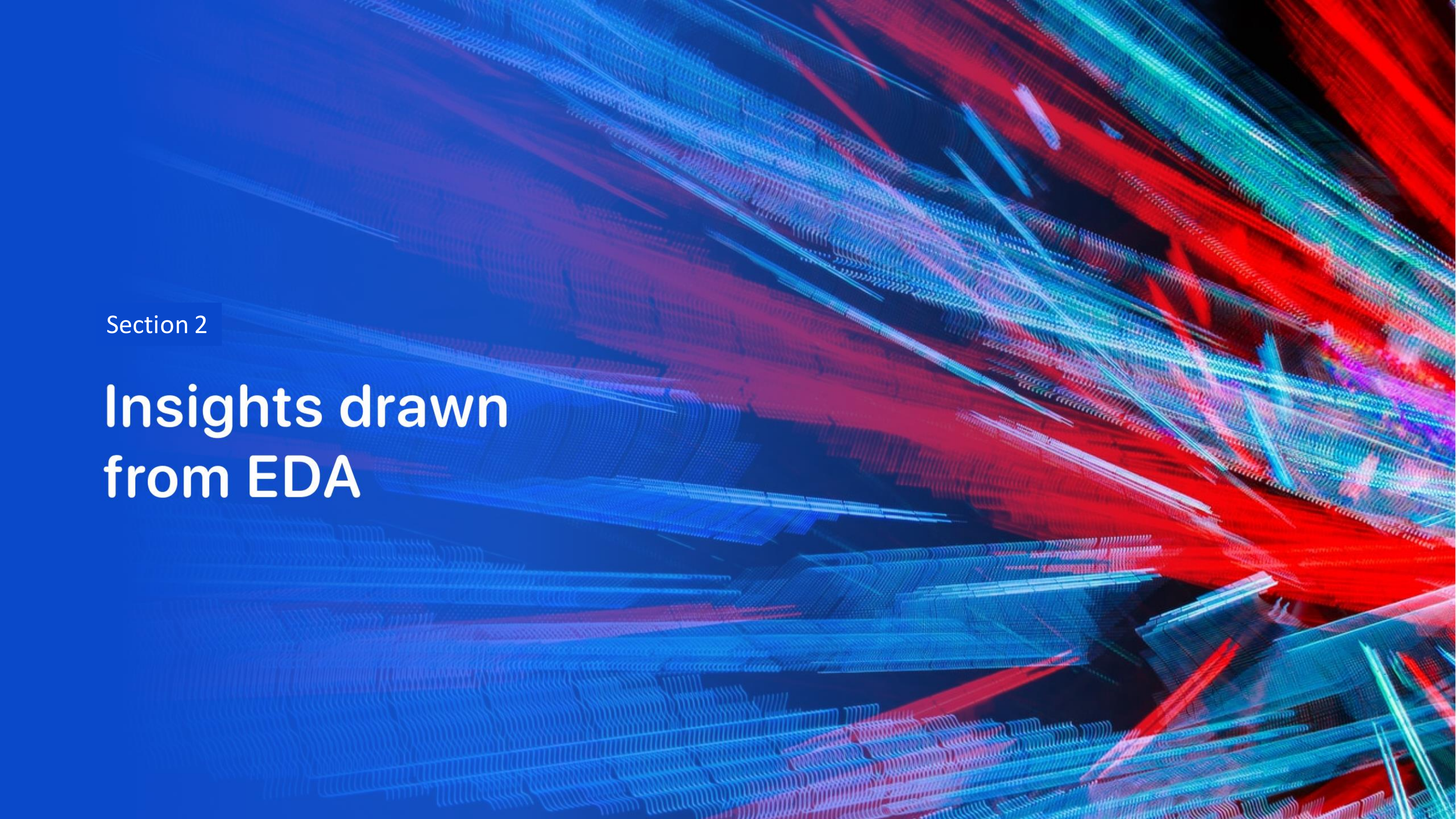
Results

- The exploratory data analysis results show that the important features to consider for predictive analysis later on are indeed the Payload Mass, Orbit type, FlightNumber, and Serial, among a few others.
- Similar results can be seen from the interactive analytics demo (screenshot below).
- The predictive analysis results show that all classification methods have the same accuracy (83%), so any of them are fine.



Section 2

Insights drawn from EDA



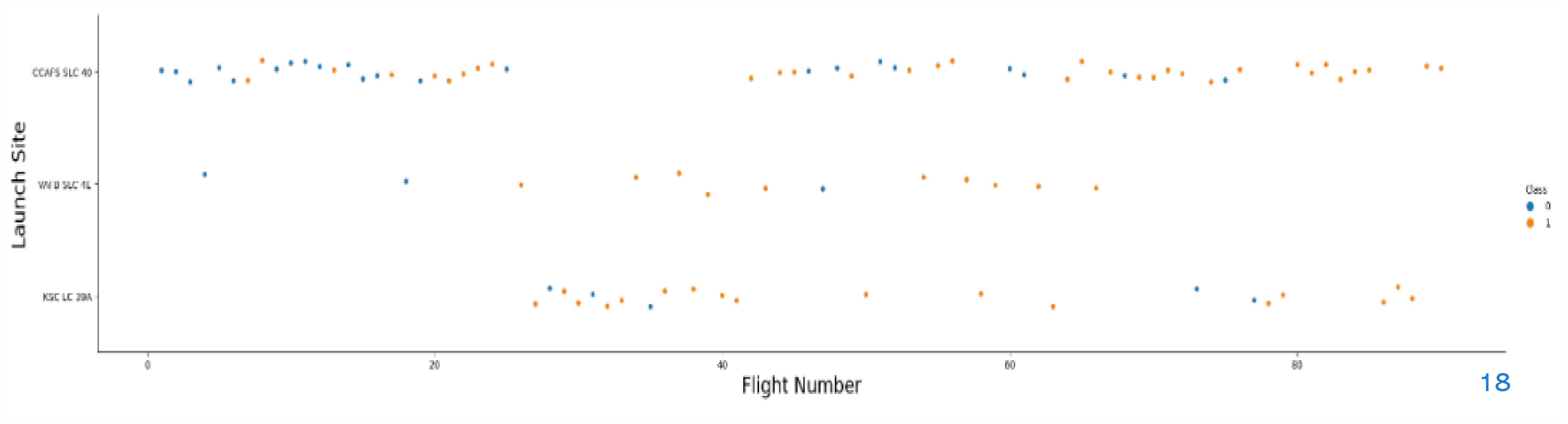
Flight Number vs. Launch Site

This plot show Launch Site vs Flight Number. Orange dots represent successful landing (of first stage) and blue the opposite.

This plot has a few insights, like for higher flight numbers with launch site CCAFS SLC 40, launches are largely successful

On the other hand, for lower flight numbers the success rate is rather mixed.

Similarly, for the VAFB SLC 4E launch site, launches are largely successful for mid-range flight numbers, but not for lower flight numbers.



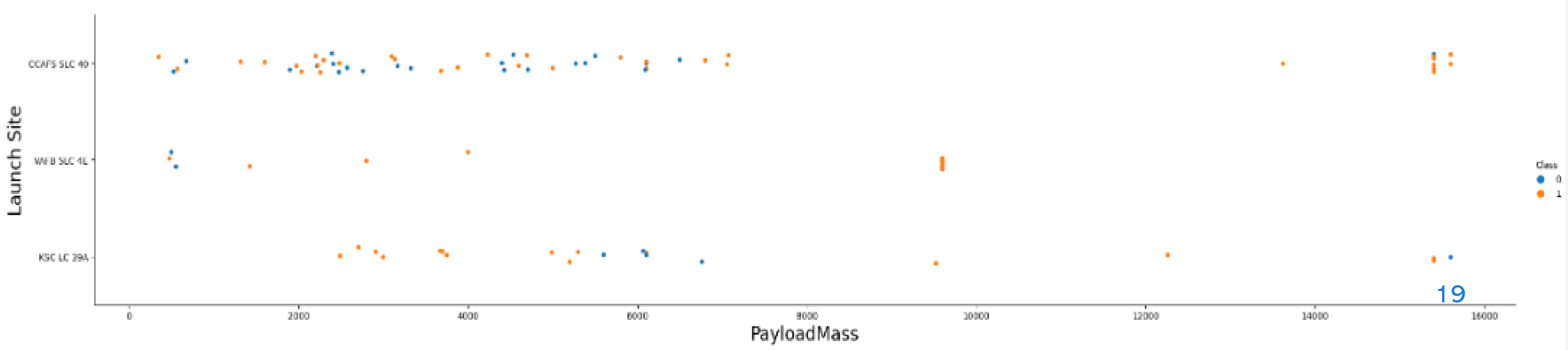
Payload vs. Launch Site

This plot show Launch Site vs Payload Mass**. Orange dots represent** successful landing (of first stage) **and blue** the opposite.

One insight here is that **for higher** Payload Mass **with launch site CCAFS SLC 40, launches are largely successful**.

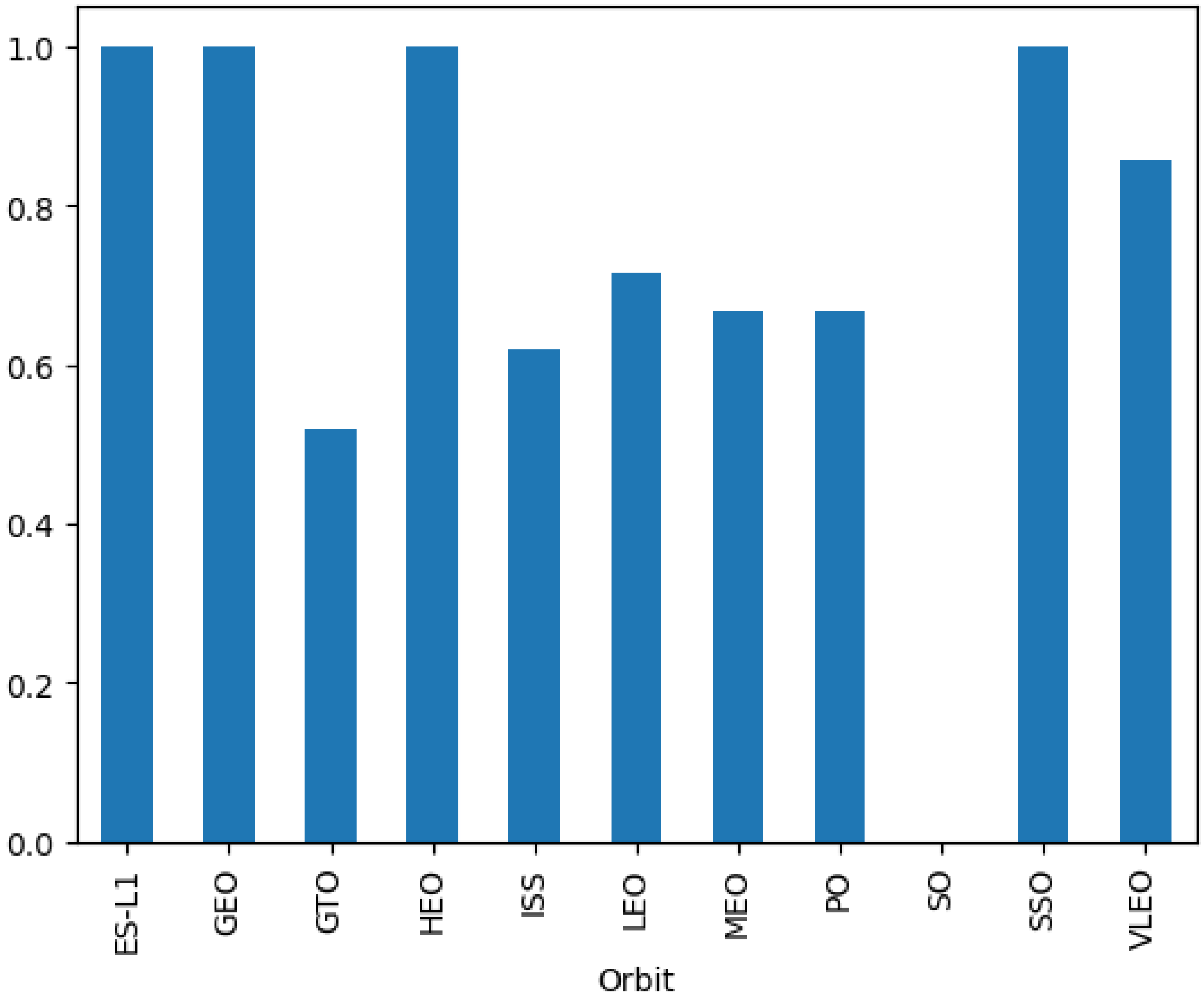
On the other hand, it **is** quite mixed otherwise.

But the other launch sites seem to be mostly successful for any Payload Mass.



Success Rate vs. Orbit Type

- This bar chart displays the success rates of the different orbit types
- As you can see, there are 4 orbits that have perfect success rates.
- Aside from VLEO, the others are not quite so great, hovering around 60%, and even SO with 0%.



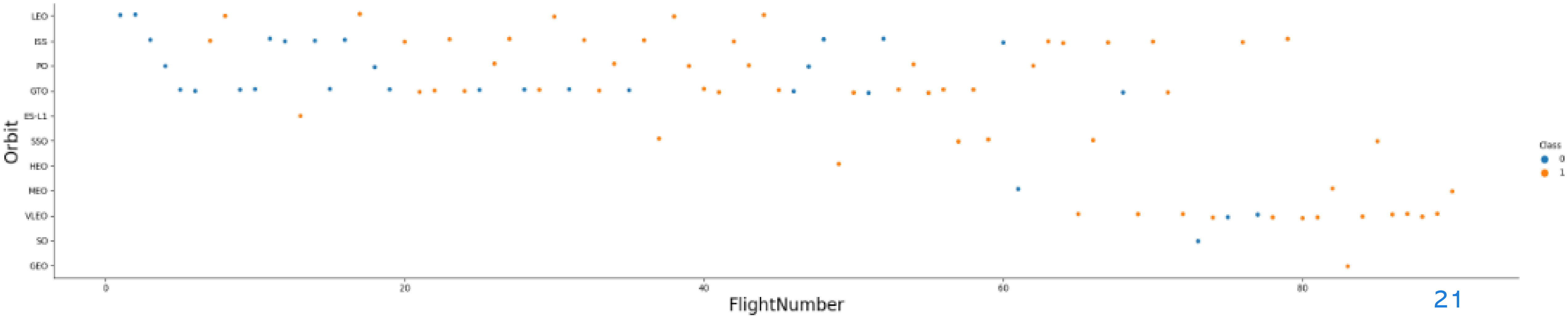
Flight Number vs. Orbit Type

This plot show Flight Number vs Payload Mass. **Orange dots represent** successful landing (of first stage) **and blue** the opposite.

One insight here is that the VLEO orbit higher flight numbers is **largely successful**.

On the other hand, it **is** quite mixed for the GTO orbit, hinting at no relationship.

But the ISS orbit might be more successful with higher flight numbers.



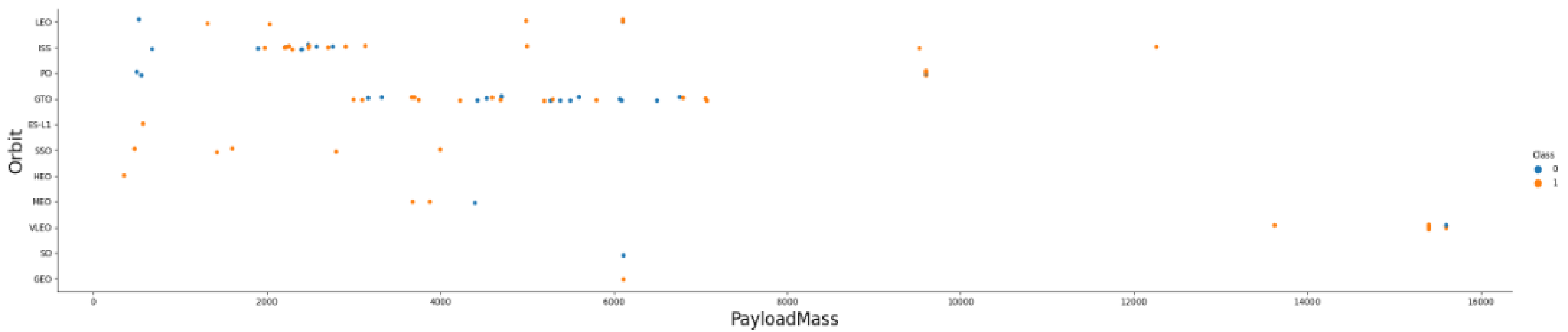
Payload vs. Orbit Type

This plot show Orbit type vs Payload Mass**. Orange dots represent** successful landing (of first stage) **and blue** the opposite**.**

One insight here is that for heavier payloads Polar, LEO, and ISS are more **successful**.

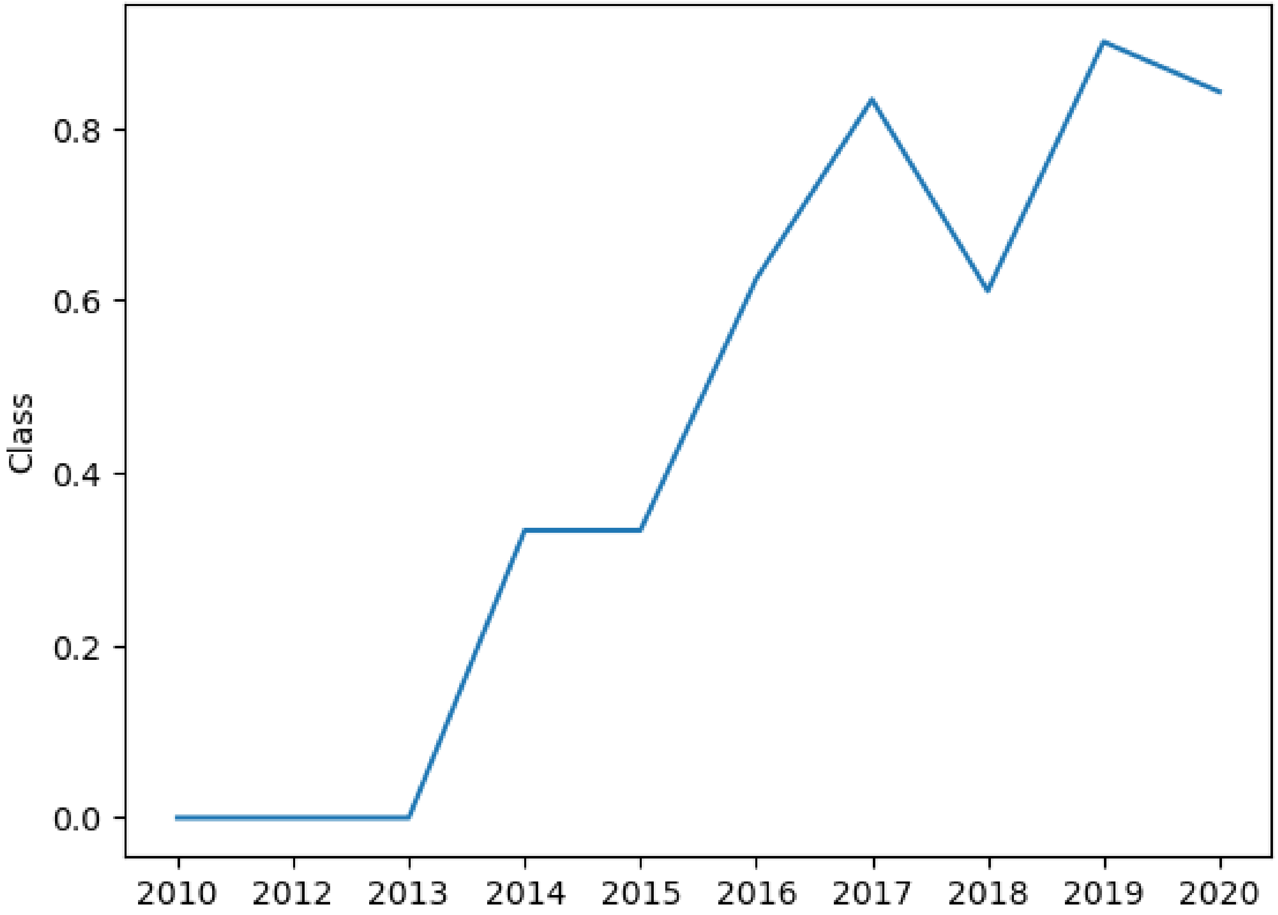
On the other hand, it **is** quite mixed for the GTO orbit again, hinting at no relationship**.**

Similarly for ISS with light Payloads.



Launch Success Yearly Trend

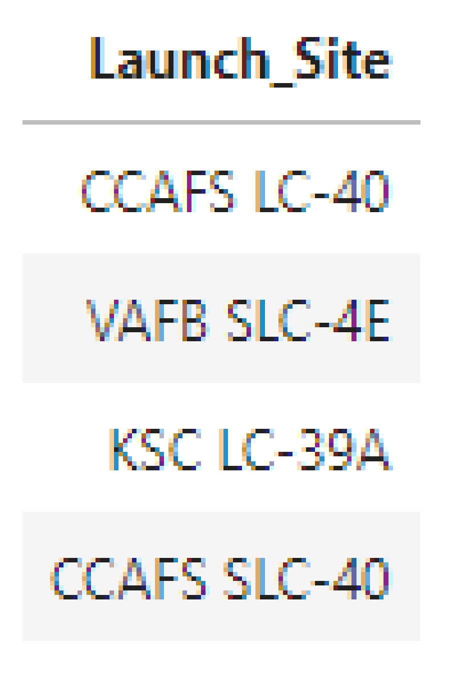
- This plot shows the yearly average success rate.
- You can see a significant improvement, especially as we move towards 2017.
- Note the minor dip in 2018, which seems to have corrected itself since 2019 has the highest success rate.



All Launch Site Names

A SQL query depicting unique launch site names

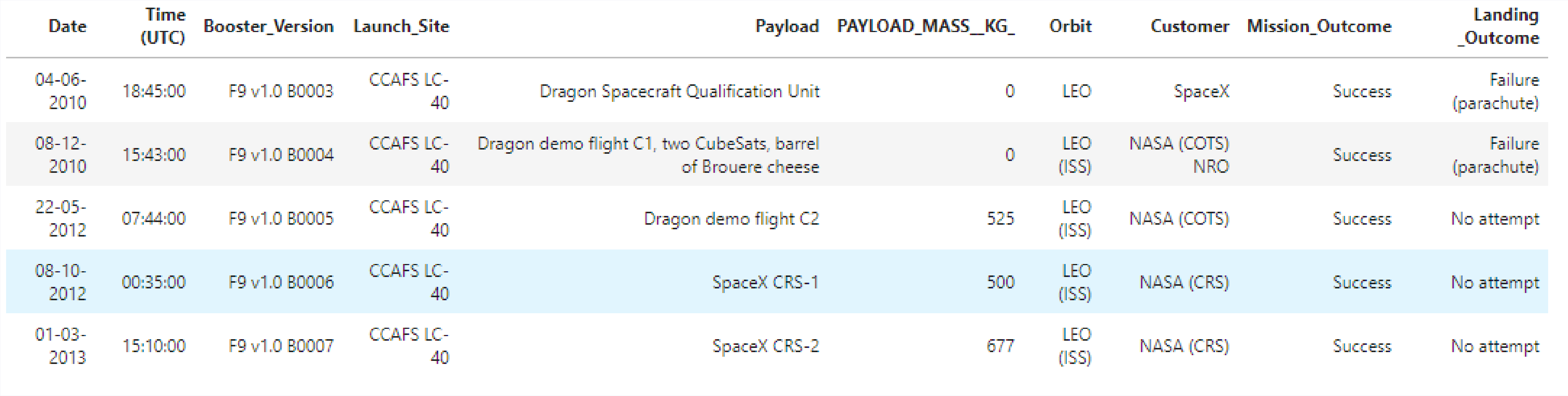
As you can see, there are only 4, which you should already be familiar with from previous the slides.



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

Launch Site Names Begin with 'CCA'

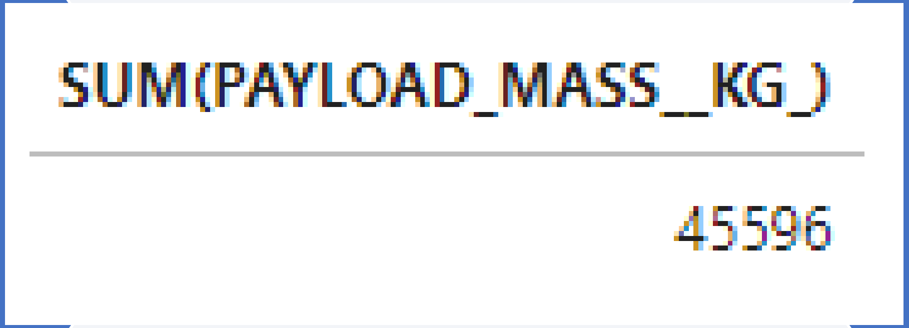
- A SQL query with 5 records where launch sites begin with `CCA`
- The reason for doing this is that there are two launch sites with names that start with 'CCA', which you saw in the last slide.



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

Total Payload Mass

The total payload carried by boosters from NASA

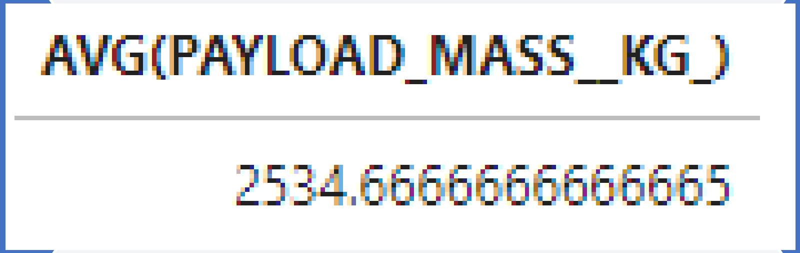


SUM(PAYLOAD_MASS__KG_)
45596

Note this is not the absolute total since we filtered for NASA customers only

Average Payload Mass by F9 v1.1

The average payload mass carried **by boosters** version F9 v1.1

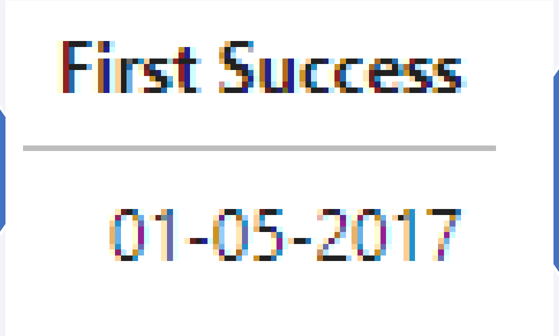


AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Note the actual average may vary since we **filtered for** the specific booster version that we care about here

First Successful Ground Landing Date

The first successful landing outcome on ground pad was in 2017.



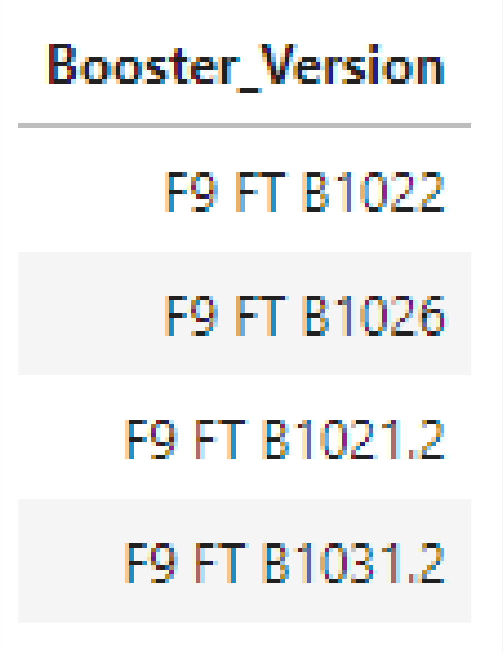
First Success
01-05-2017

Note from the line chart a few slides ago, 2017 was also the year where average success rate went up.

Successful Drone Ship Landing with Payload between 4000 and 6000

A SQL query depicting the names of boosters which had successful landings on drone ship with payload mass between 4000 and 6000

As you can see, there are only 4, signifying either a lack of data or poor success rate for this case

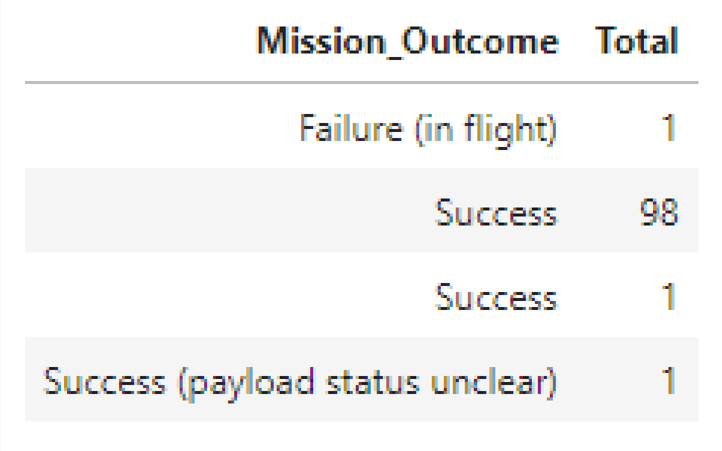


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

A SQL query depicting the number of successful and failure mission outcomes

As you can **see,** when we disregard landing, success rate is quite high.

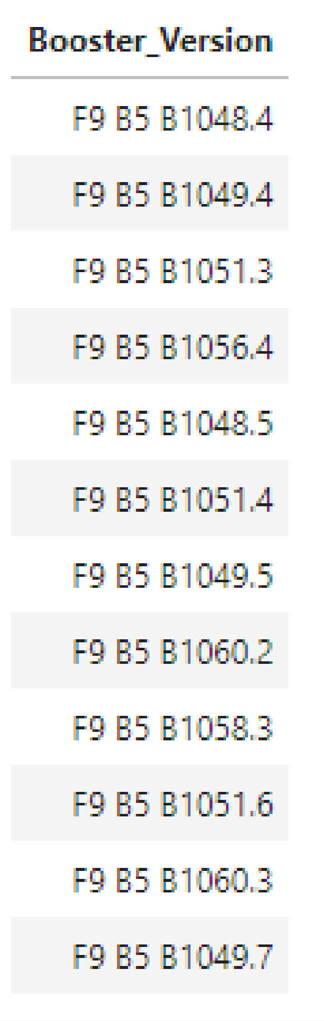


Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

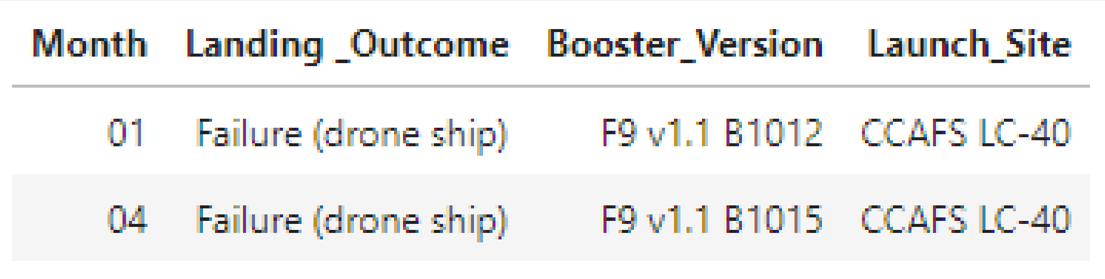
A SQL query depicting the booster versions which carries the maximum payload mass

As you can **see,** there are various versions in the list, indicating the F9 B5 booster version can handle high payload mass



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



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

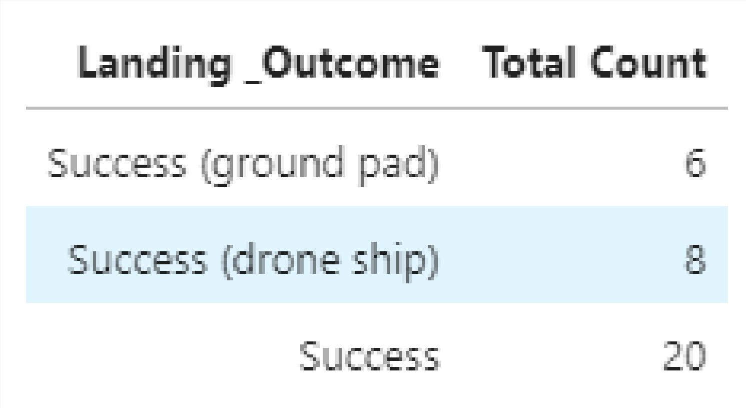
The **failed** landing outcomes **in drone ship, their booster versions, and launch site names for in year 2015**

As you can see, since there are only two values, this may possibly hint at these feature values being favoured as they have a small number of failures

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

A SQL query depicting the number of successful landing outcomes from 2010-06-04 to 2017-03-20

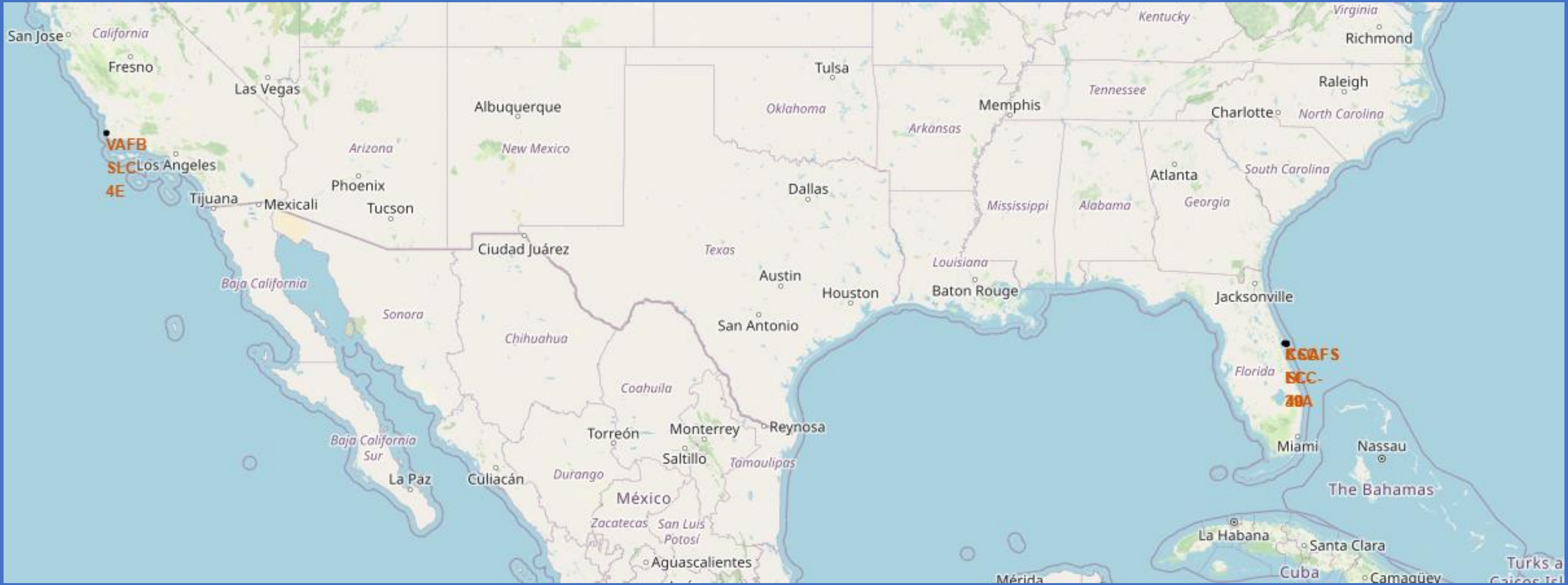
As you can **see,** the ground pad and drone ship success contribute to less than half the total success each



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

Section 3

Launch Sites Proximities Analysis

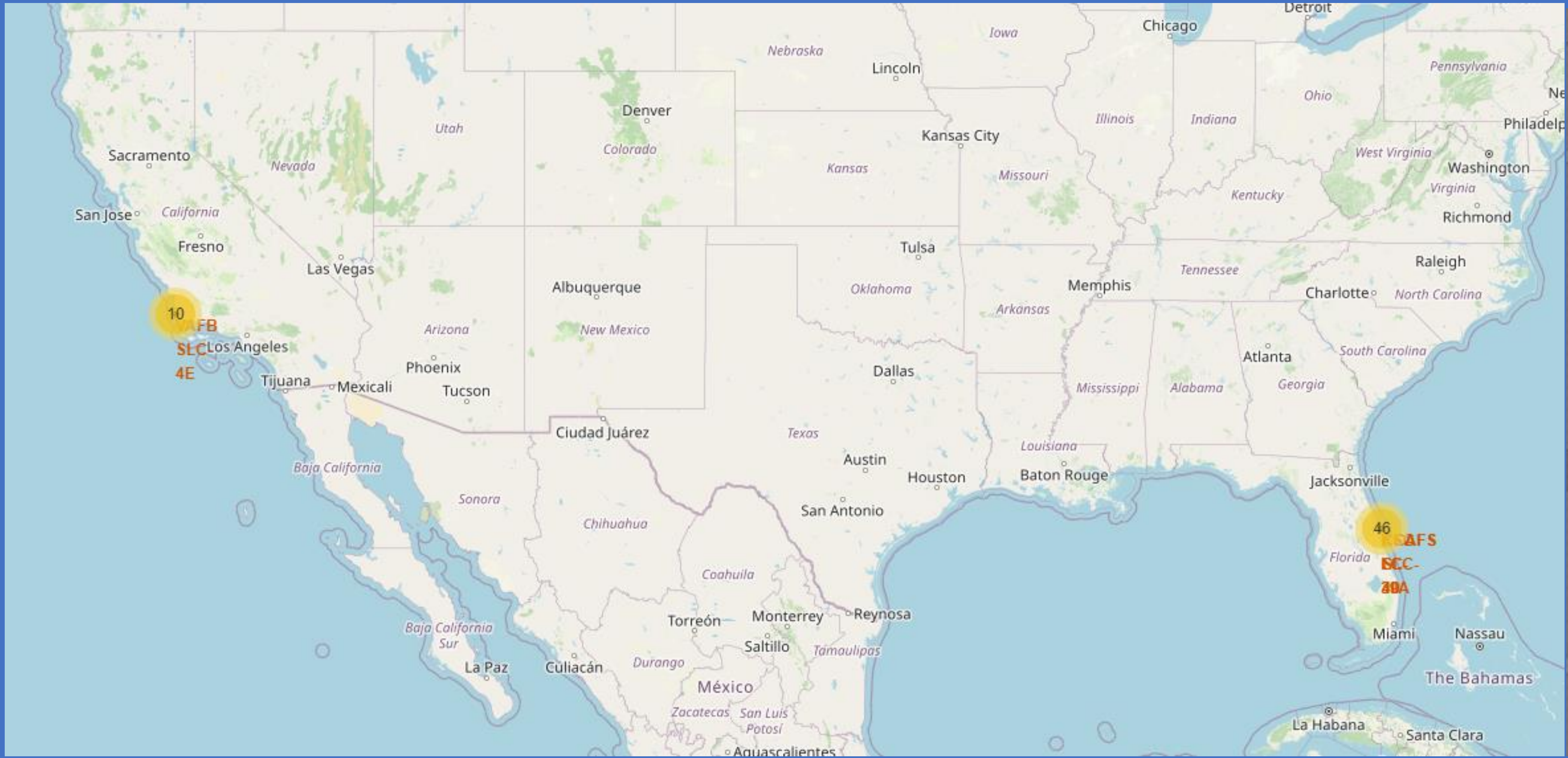


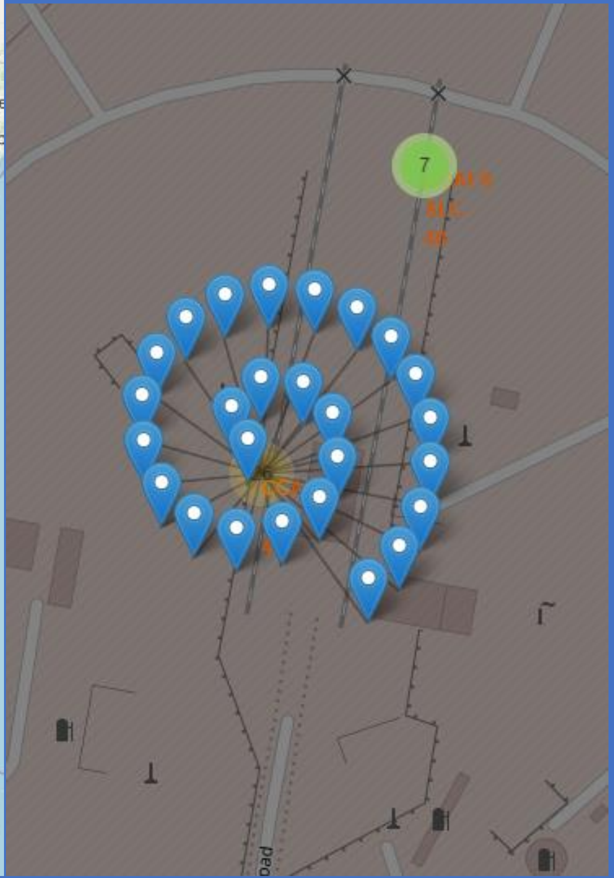
Marking Launch Sites on a Map (Folium 1)

In the below map you can see the locations of the launch sites. They are marked by black dots.

There are only two visible because three of the launch sites are actually right next to each other

Something to note about these launch sites is that they are all near the water and away from cities



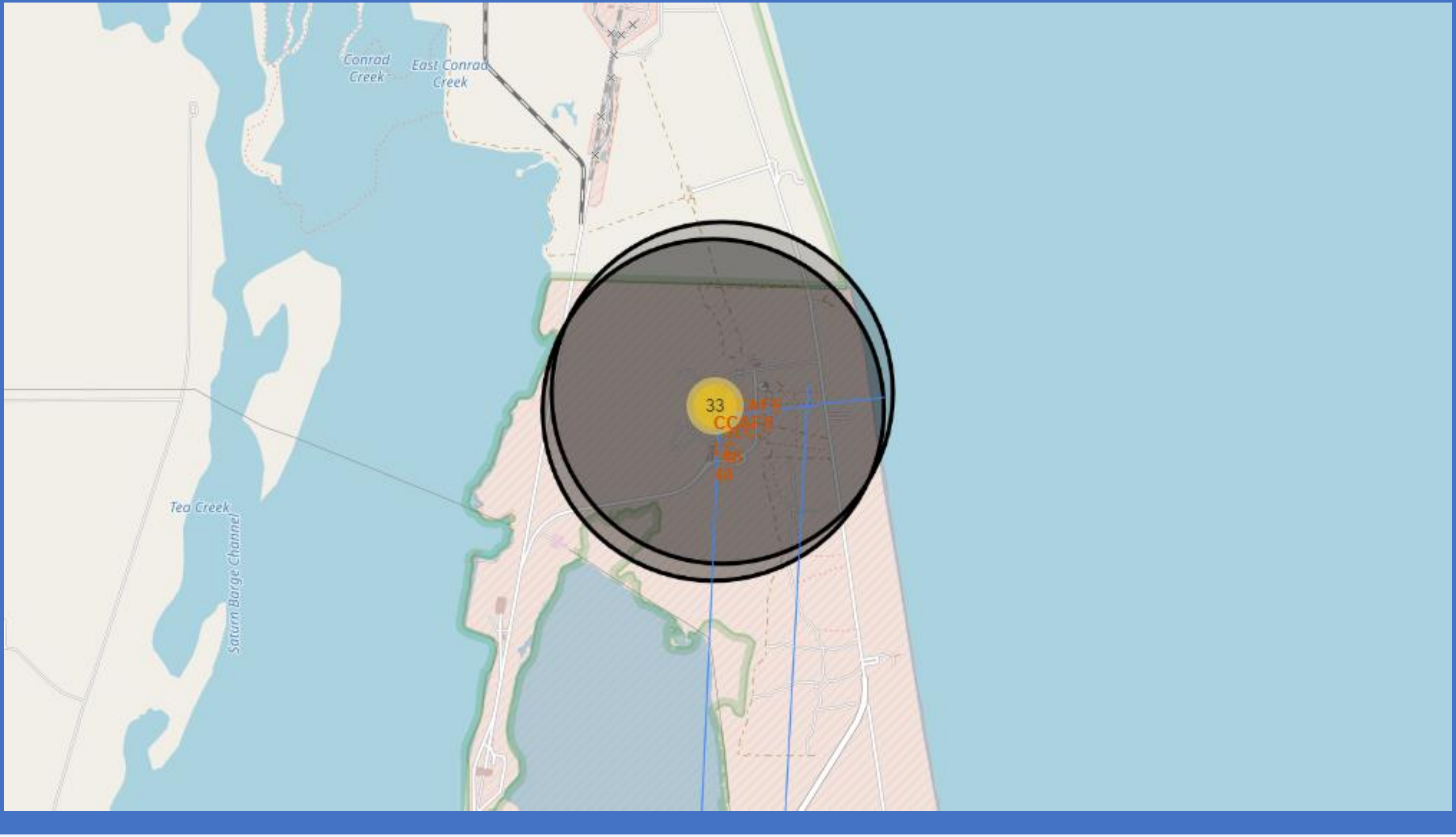


Marking Launch Outcomes with a Cluster (Folium 2)

Since the launch outcomes occur in the same site, this map has marker clusters to show them

By clicking on a launch site, you will see the launch outcome record in that launch site

For example, if you click on the CCAFS LC 40 site, you will see the image on the right above

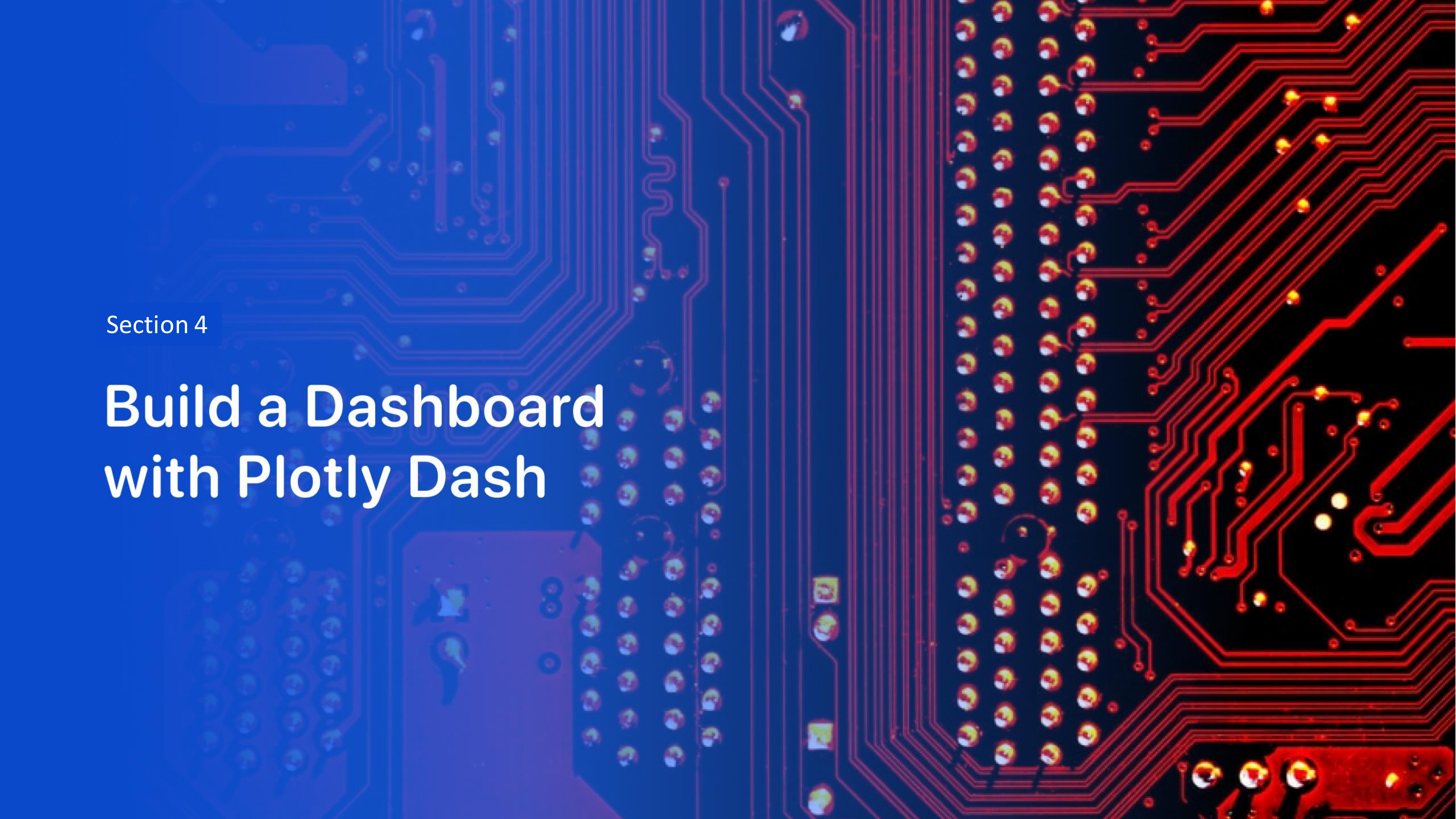


Marking Distances from Launch Sites to Notable Locations (Folium 3)

This map marks with a blue line how far a launch site is from a notable proximity

You can see a horizontal blue line in the image above, which shows how far the launch site is from the water

Alternatively, the lines going down extend to cities. This contrast shows that launch sites are always near low populated areas

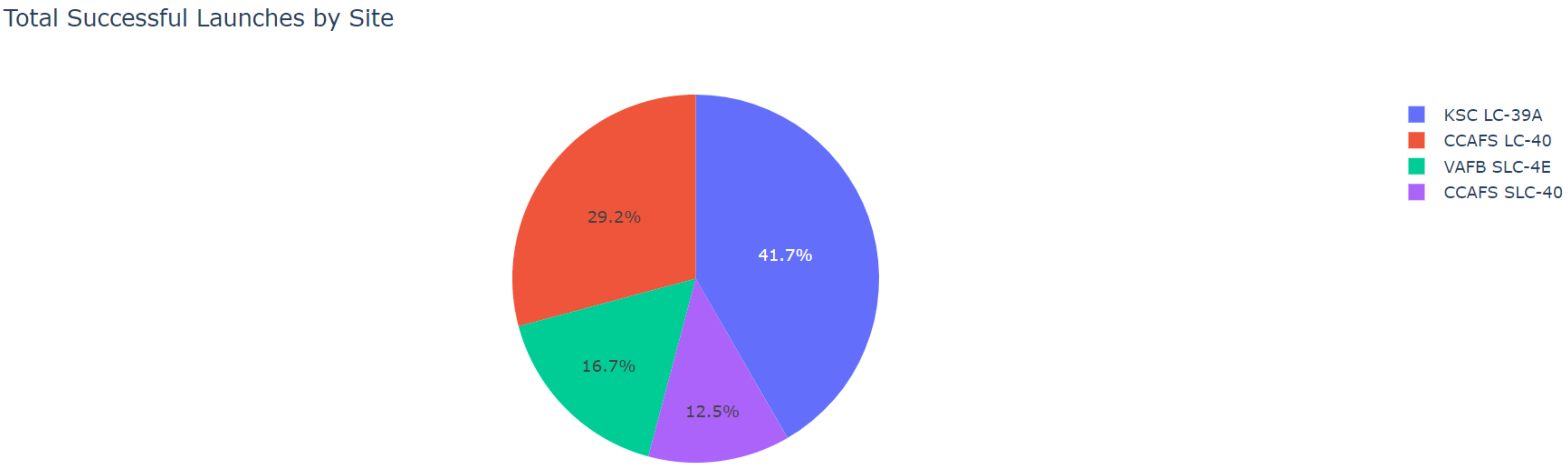


Section 4

Build a Dashboard with Plotly Dash

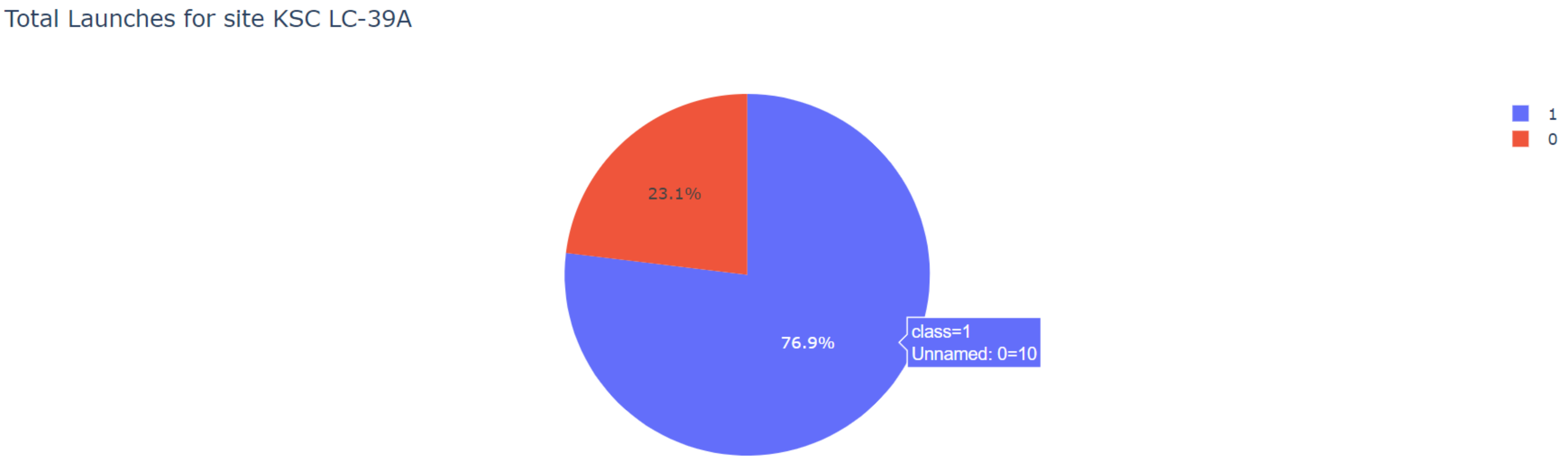
Comparing Launch Sites by their Successes

- This pie chart clearly shows that the KSC LC-39A Launch Site has the most successes.
- Though this may not be the whole story, as their individual launch rate is also important, it is a good sign for it at least.
- Naturally, we can make the opposite assumption for the CCAFS SLC-40 Launch Site



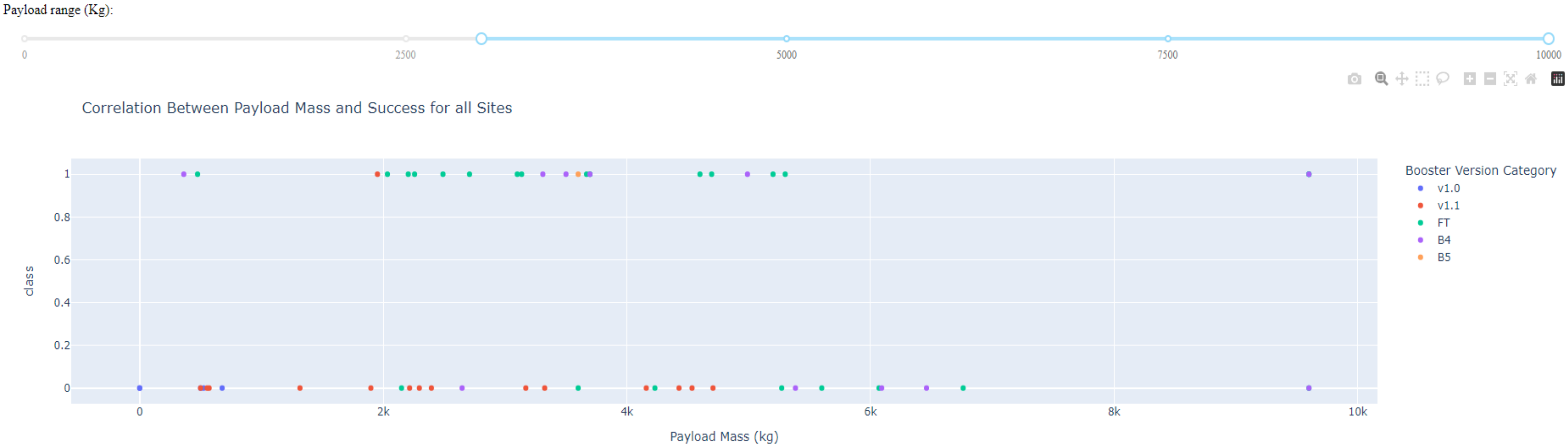
Launch Site with Highest Success Rate

- This pie chart shows that the KSC LC-39A Launch Site has a success rate of almost 77% (blue is success and red is failure).
- This is in fact the highest success rate, so our assumption from the last slide about this launch site being the best seems to be true.
- Turns out the lowest success rate belongs to the launch site with the least amount of success from the last slide too.



Comparing All Launch Site Successes for their Booster Version against Payload

- From the plot you can see that the FT Booster Version has the highest success rate.
- Moreover, the v1.1 seems to have the least success rate, while the others are rather mixed.
- Furthermore, you can clearly see that higher payloads generally do not lead to many successes

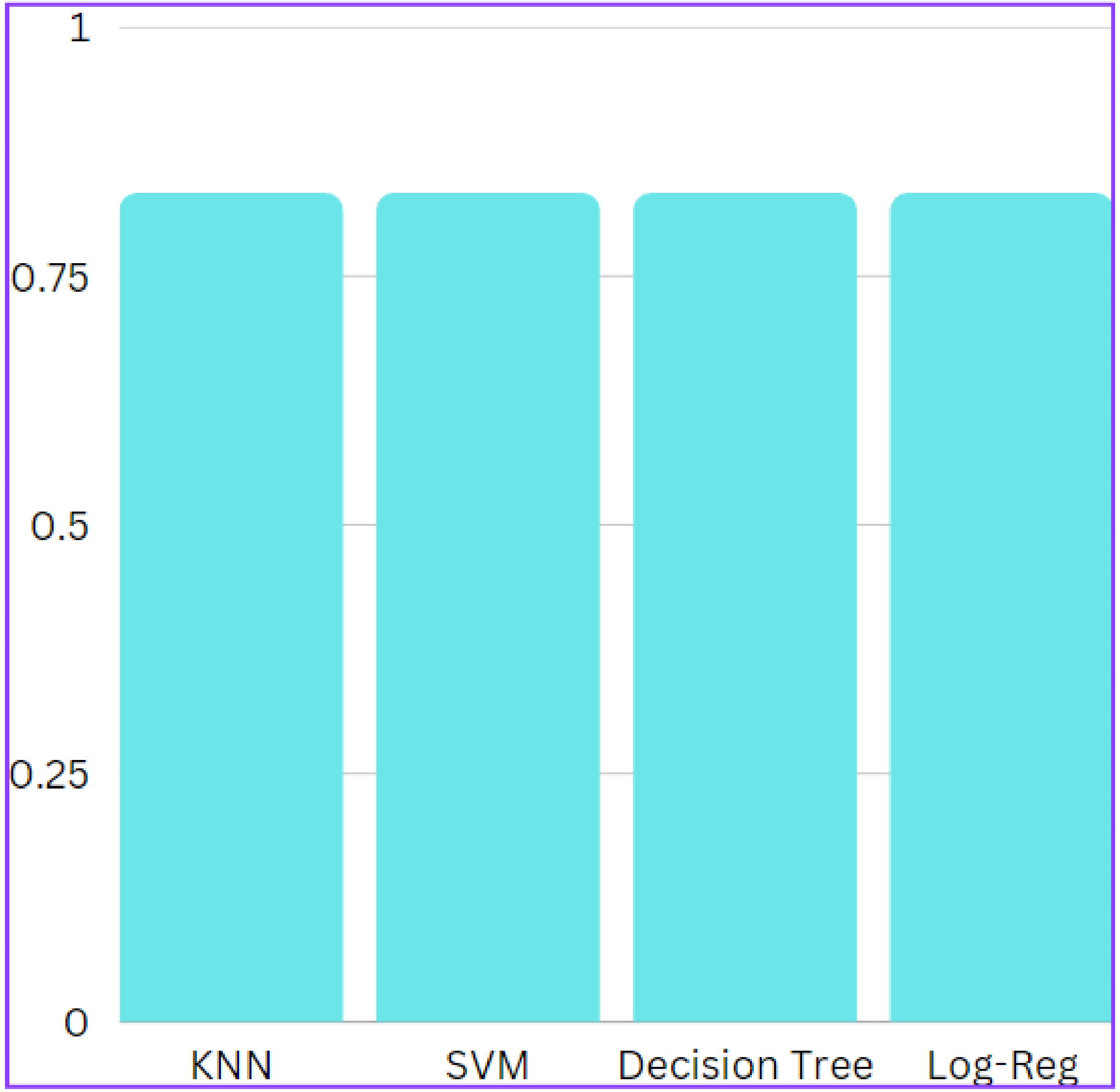

Payload range (Kg):

Section 5

Predictive Analysis (Classification)

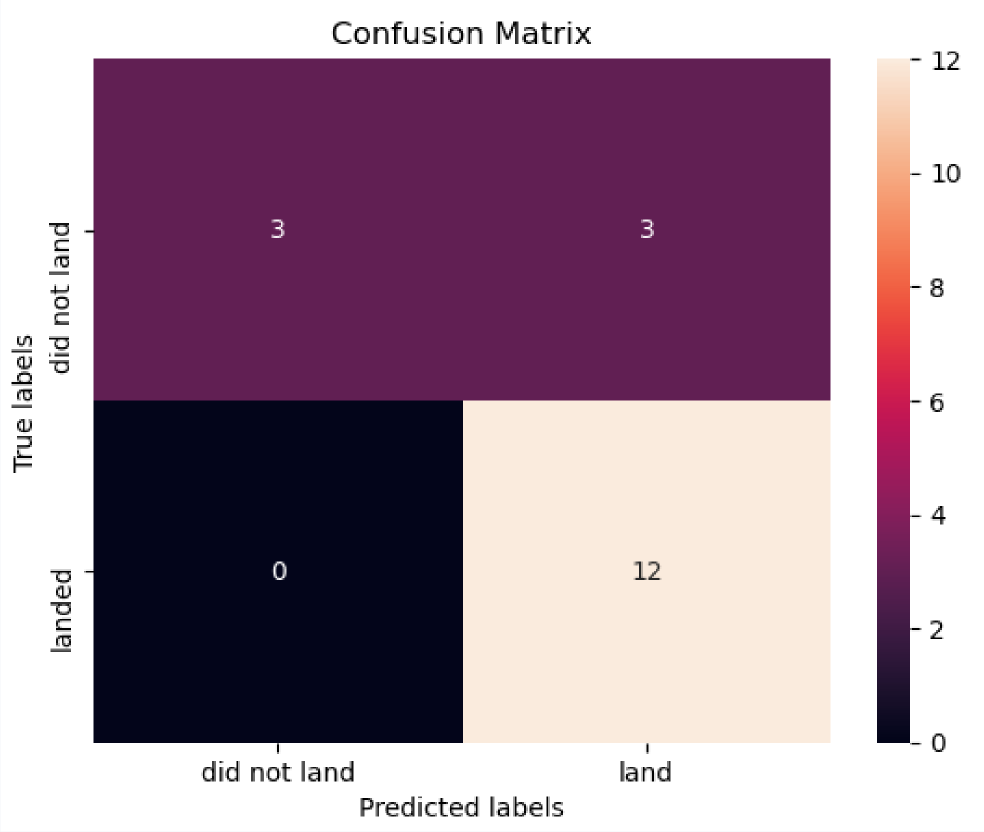
Classification Accuracy

- The bar chart on the right is displays the accuracy on the test set of each model
- As you can see, all models miraculously have the same accuracy, so there is no best one, or they all are!



Confusion Matrix

- Since each model had the same performance, they all have identical confusion matrices
- Nonetheless, the confusion matrix on the right is from the decision tree model
- Naturally from the 83% accuracy, the confusion matrix shows that the model did perform well
- However, with zero false negatives, and a few false positives, we know the model could still be improved to catch these false positives better



Conclusions

From Data Collection and Wrangling steps, we know this data is a bit messy to begin with, so if starting from scratch be sure to perform the necessary pre-processing steps

Data Analysis shows some key highlights about the data, regarding certain relationships between different features and which ones may be irrelevant

It is important to note that dealing with null values is always tricky, and had they been handled differently than the way they were handled here, the data analysis results could possibly have been different

All classification models performed well but also the same. This is suspicious and may need further investigation

In any case, the models all missed some false positive cases, and no negative, so there is a clear weakness in the model to be further improved on

Appendix

- A GitHub link to all files shared before and any other relevant files for this project: https://github.com/RavinderRai/Sucessful-SpaceX-Launch-Predictions

Thank you!