Data Science Capstone project

<Gloria Burengengwa>

<30/08/2021>

Outline

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

Executive Summary

- At first, we collected the Space X data from an API
- We extracted the required data using Beautiful Soup python library
- We operated preprocessing steps where we replaced all missing values and classified the landing outcome into 2 classes whether it failed or not
- We conducted a thorough analysis using SQL and
- We conducted a spatial analysis using folium python library, which allowed us to locate where the launch sites are located on a map and to identify the successful and failed flight on each site
- We visualized the relationship between different variables to help us understand the landing outcome

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. In this capstone, as a data scientist I am interested to determine whether the first stage will land successfully or not in such a way that if the first land will land I can determine the cost of a lunch. With this information, an alternate company can bid against SpaceX for a rocket launch.

- Problems you want to find answers

Firstly, we want to predict if the Falcon 9 first stage will land successfully.

Alongside that, we will explore the reason why it lands successfully or not:

It might depends on the payload each launch carries or on the date it was launched or maybe from the site it was launched.

Methodology

- Data collection methodology:
  - Data is collected from an API
  - A get request is made to the SpaceX API
  - To make sure the request was successful, we check if we have the 200 status response code
  - Then a pandas dataframe is created using the json_normalize function after decoding the response
- Perform data wrangling
  - Our interest is only in Falcon 9 data, therefore we remove all other SpaceX launches
  - Then we check for the missing values and replace them with their mean values
- 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

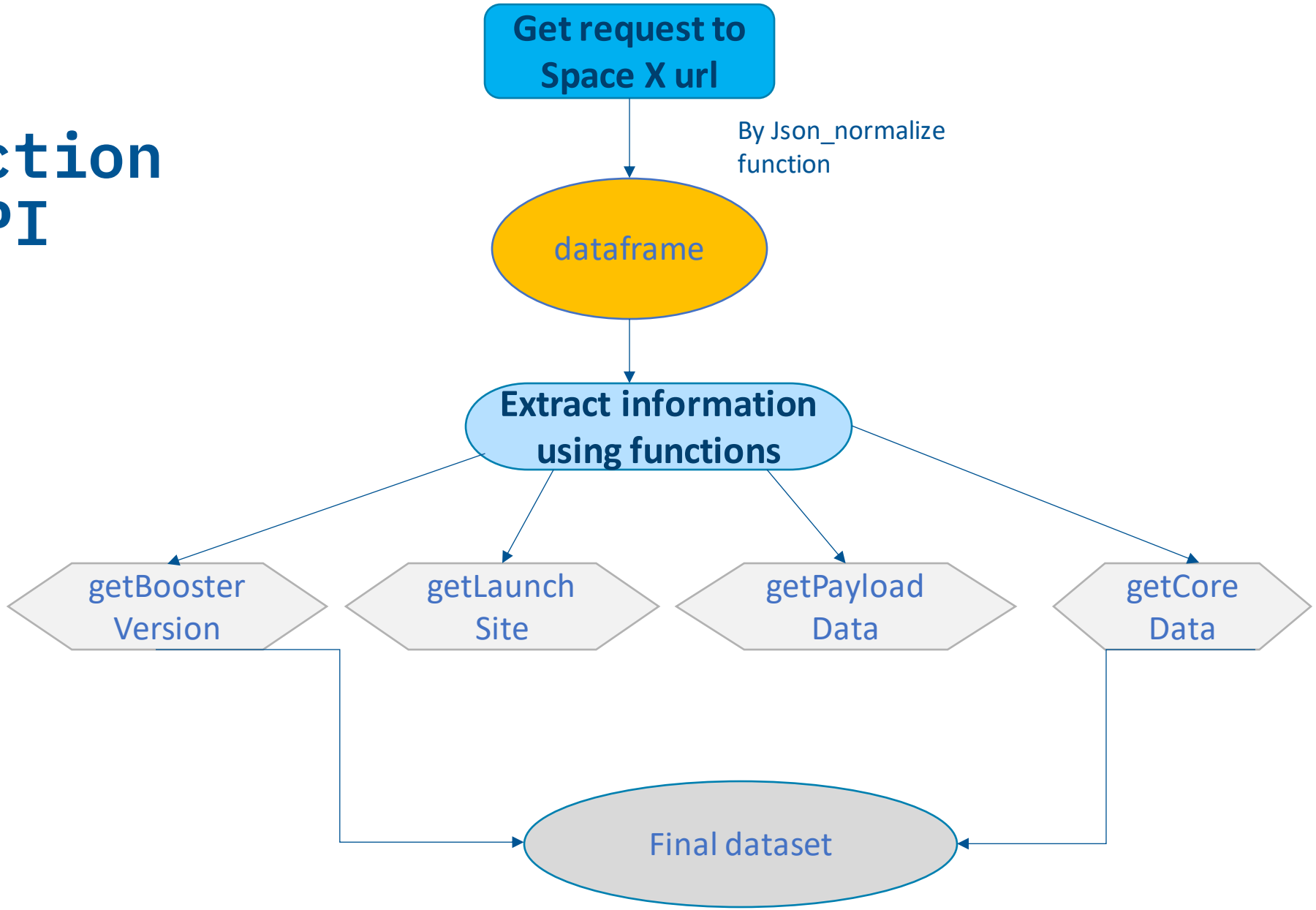
Methodology

The next sections will contain the process of the analysis

Data collection

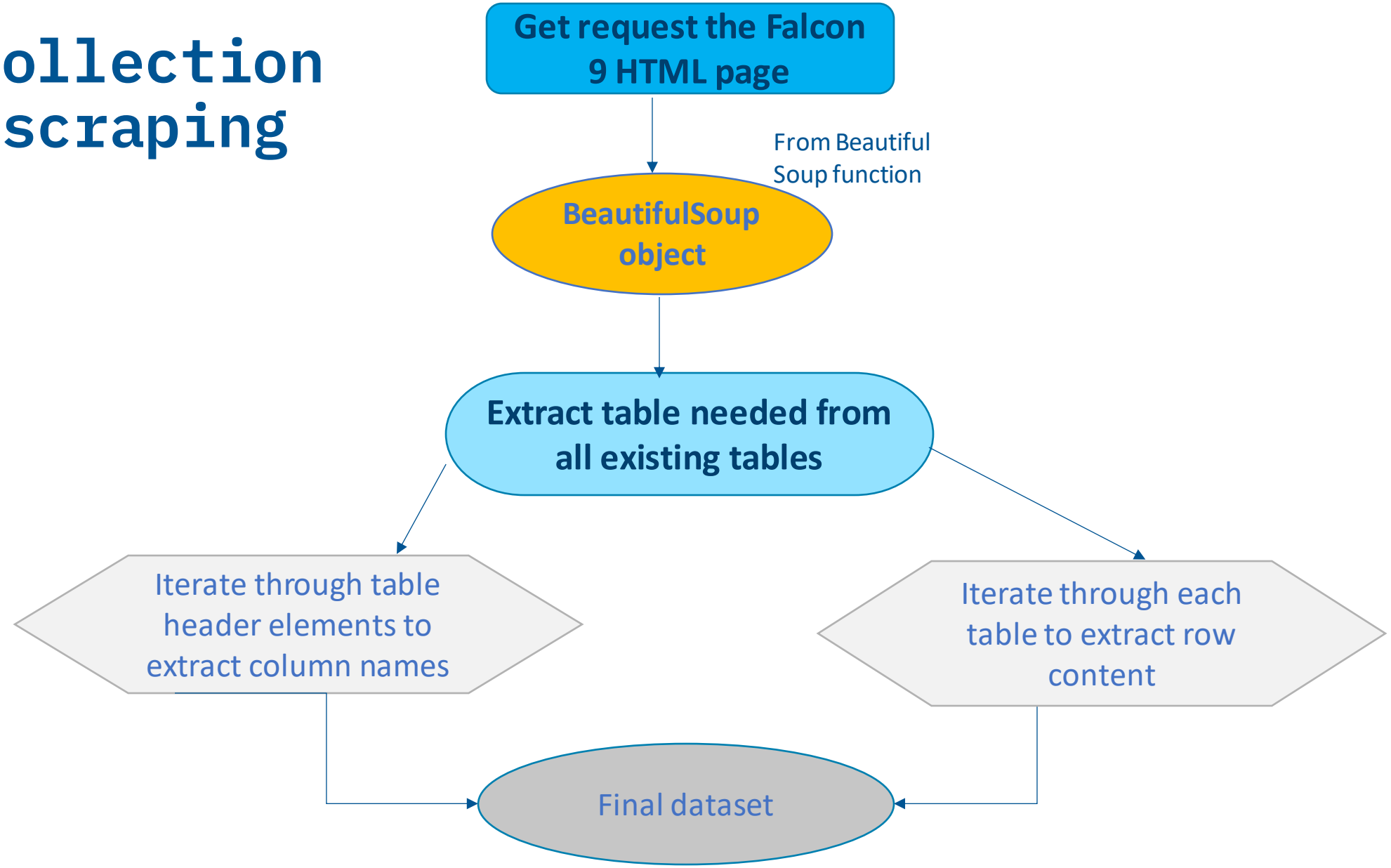
- Description of data sets collection:
  - Data was collected from an API
  - A get request was made to the SpaceX API
  - To make sure the request was successful, we checked if we have the 200 status response code
  - Then a pandas dataframe is created using the json_normalize function after decoding the response
  - The API is used again to get information about the launches using different IDs given for each launch. We use IDs because a lot of the data is written in the form of identification number. This is the case of the rocket, payloads, launchpad and cores columns in the dataset.

Data collection – SpaceX API



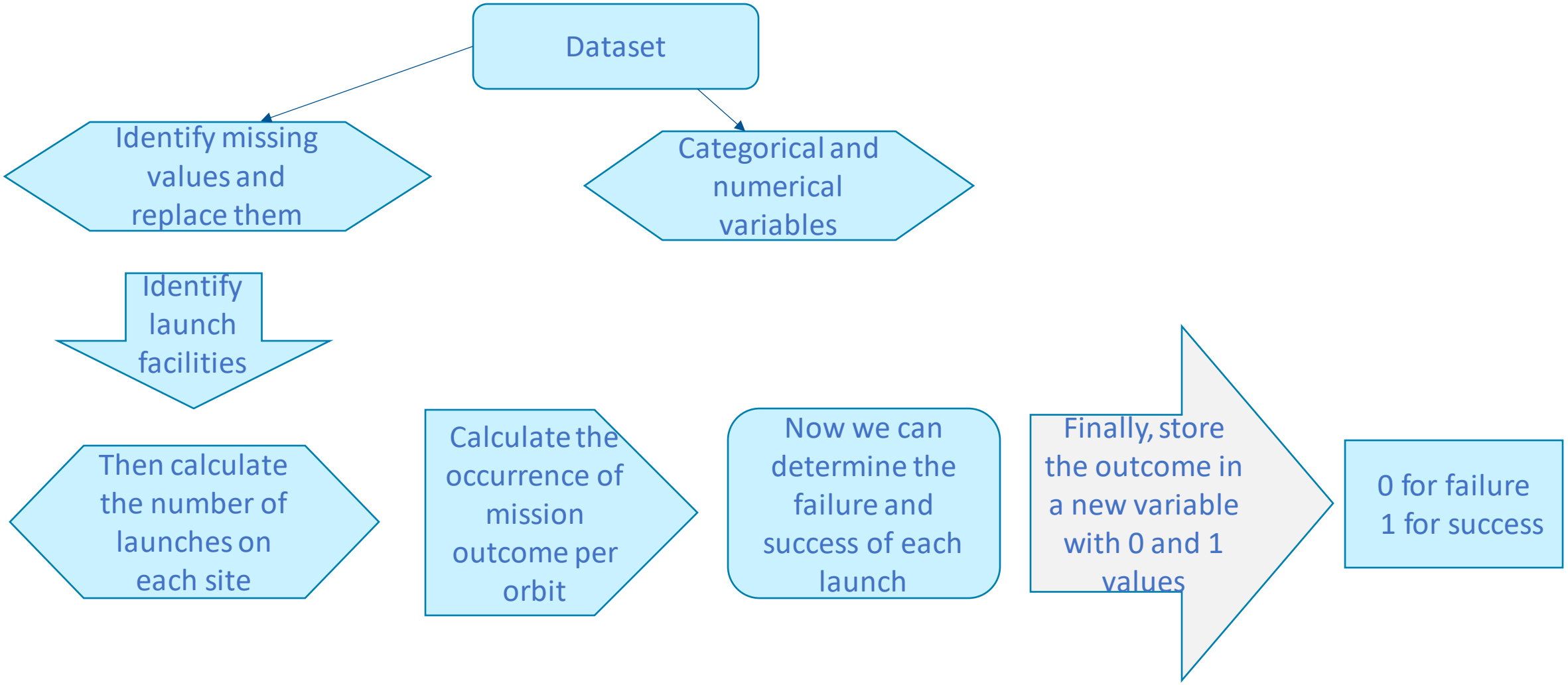
https://github.com/burenG1/Capstone-DS/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data collection – Web scraping



https://github.com/burenG1/Capstone-DS/blob/main/jupyter-labs-webscraping.ipynb

Data wrangling



https://github.com/burenG1/Capstone-DS/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with data visualization

- We visualize the relationship between flight numbers and different other variables like(payloads, launch sites and orbit type)
- For the above visualizations we used the scatter point plot because it shows us whether there exist a relationship between 2 variables
- We also visualize the relationship between payload and launch site to see if successful landing might vary for the different launch site depending on the payloads
- We examine also successful rate per orbit type
- Then we used the line chart to visualize the launch success yearly trend, why the line chart? Because it helps us capture the trend

https://github.com/burenG1/Capstone-DS/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We performed distinct function to display the unique name of the launch sites in the space mission
- Using ‘limit’ we displayed the first five records where launch sites begin with the string ‘CCA’
- We used 2 functions, distinct on customer and sum for only NASA on payload variable using like in the where clause to display the total payload mass carried by boosters launched by NASA
- We used the avg function to display the average payload mass carried by booster version F9 v1.1
- Use of min function to find the date when the first successful landing outcome in ground pad was achieved
- Used between in the where clause to list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Use of count and group by to list the total number of successful and failure mission outcomes
- Use a subquery to list the names of the booster versions which have carried the maximum payload mass
- Use of two conditions with ‘and’ to list the failed landing outcomes in drone ship, their booster versions, and launch site names for the in year 2015
- Use of or, and, group by and desc functions to rank the count of landing outcomes (such as Failure on drone ship or Success on ground pad) between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/burenG1/Capstone-DS/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an interactive map with Folium

- We used the markers objects to add locations on a map using their latitude and longitude coordinates
- We used the circle object to add a highlighted circle area and we added a text label on specific coordinate using the popup label
- We added a mouse position on the map to get coordinate for a mouse over a point on the map so that we could get the coordinates of railways in each launch site proximity
- We then used the new coordinates collected to calculate the distance between the point and the site
- Then the folium.PolyLine object with using the calculated distance

https://github.com/burenG1/Capstone-DS/blob/main/lab_jupyter_launch_site_location(2).ipynb

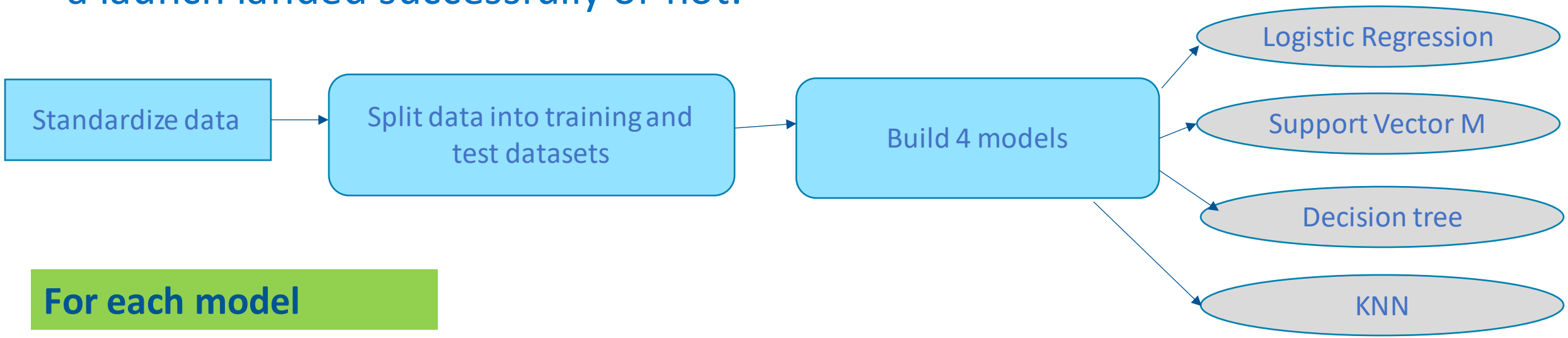
Build a Dashboard with Plotly Dash

We have 3 plots showing dashboards created using this dataset

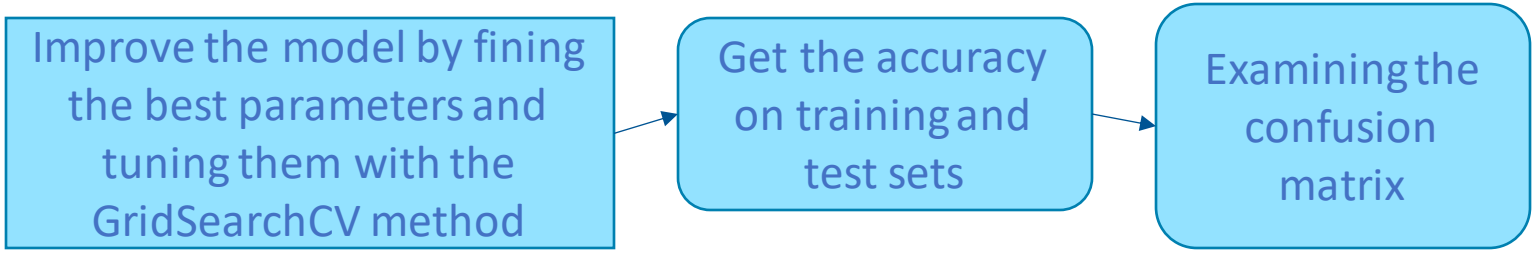
- The first plot, a pie chart, shows a dashboard of counts of all launches for all the sites.
- The second plot also a pie chart, shows the number of launches on the KSC LC-39 A site which is found to be the launch site with highest launch success ratio
- The last plot shows a dashboard of the landing outcome vs payload scatter plot for all sites.

Predictive analysis (Classification)

We built four different machine learning algorithms to classify whether a launch landed successfully or not.



For each model



The result is that they all perform the same however there is a minor difference when evaluating the training sets

https://github.com/burenG1/Capstone-DS/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

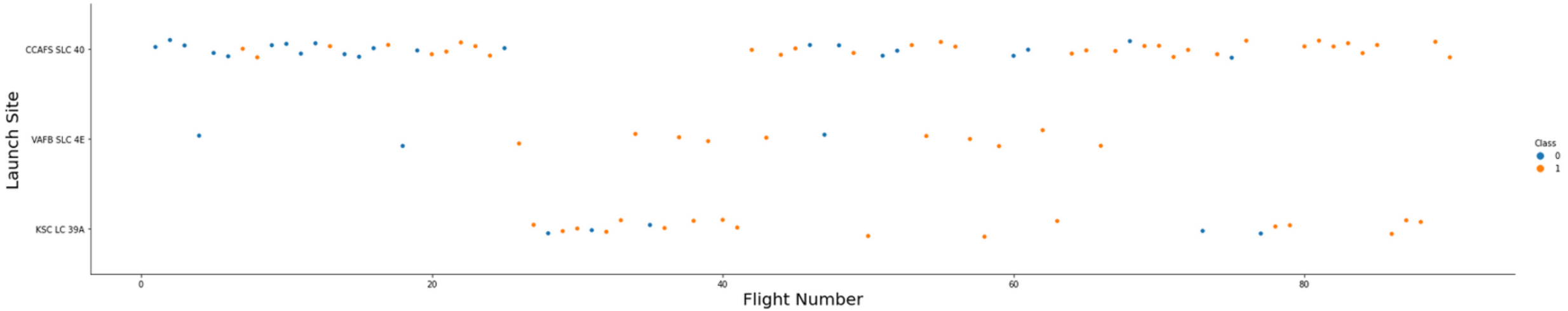
Results

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

EDA with Visualization

The next sections will contain the exploratory data analysis with visualization

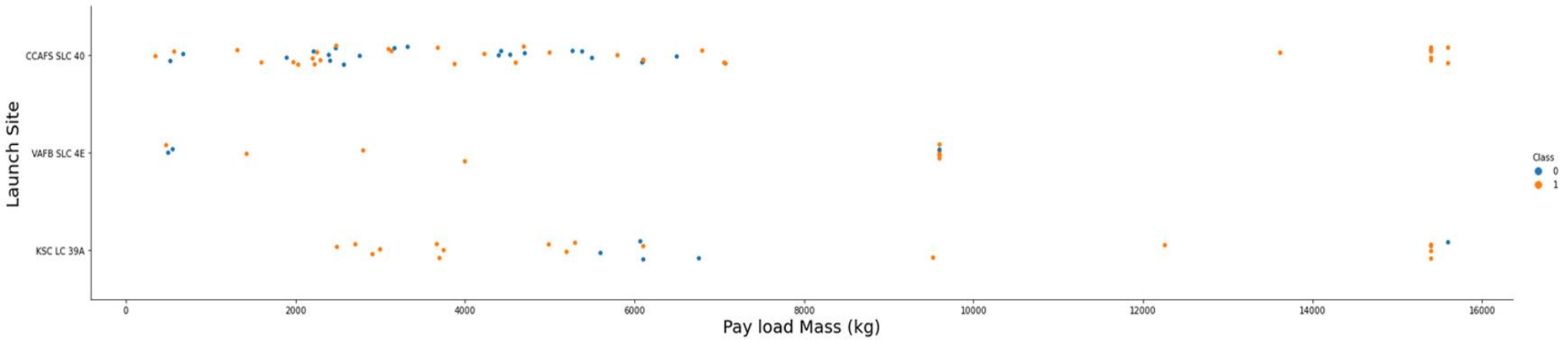
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

We see that at the begining of the launch, first stages were most likely to fail but the more the number of flights increases the better it gets for the three different sites.

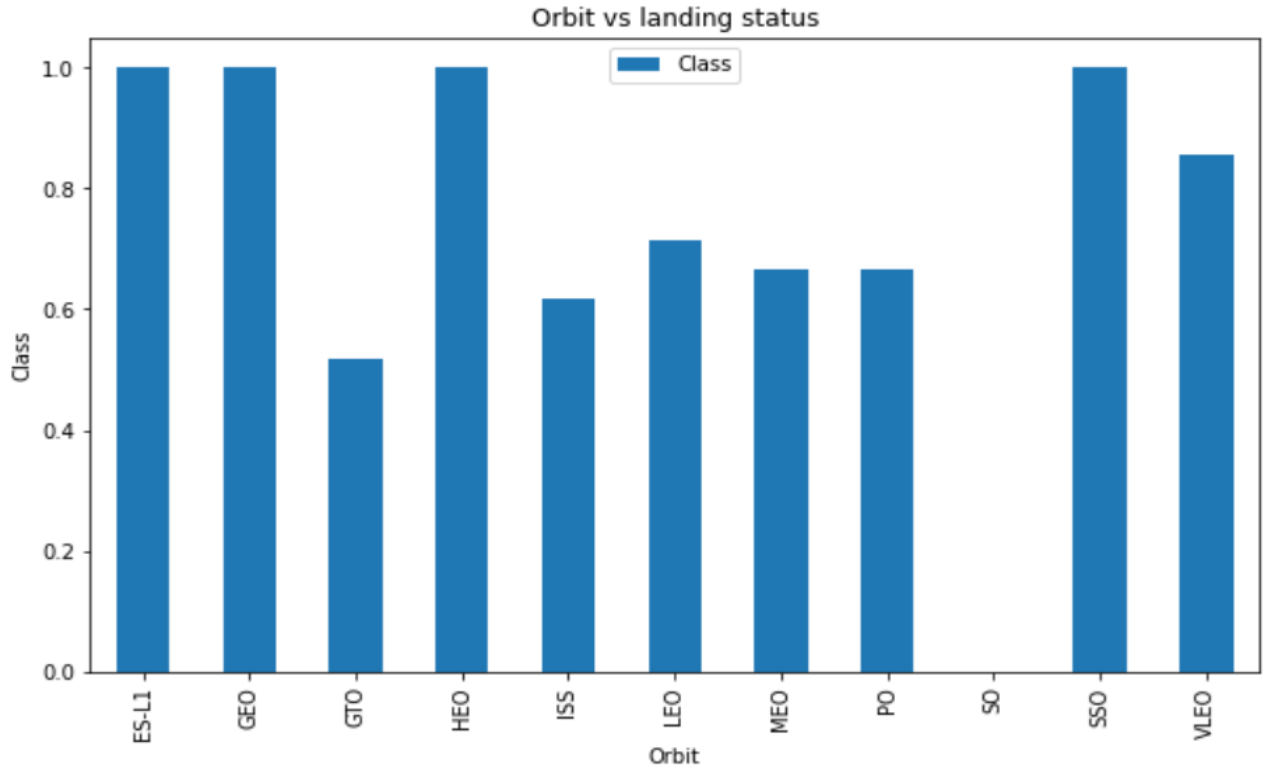
Payload vs. Launch Site



Now try to explain any patterns you found in the Payload Vs. Launch Site scatter point chart.

It is shown that all the three launch sites mostly cary load below 8000kg and that is where we see the highest probability for the first stage to fail. Likewise, we see that the more load the flight carries the more likely the first stage is to land successfully.

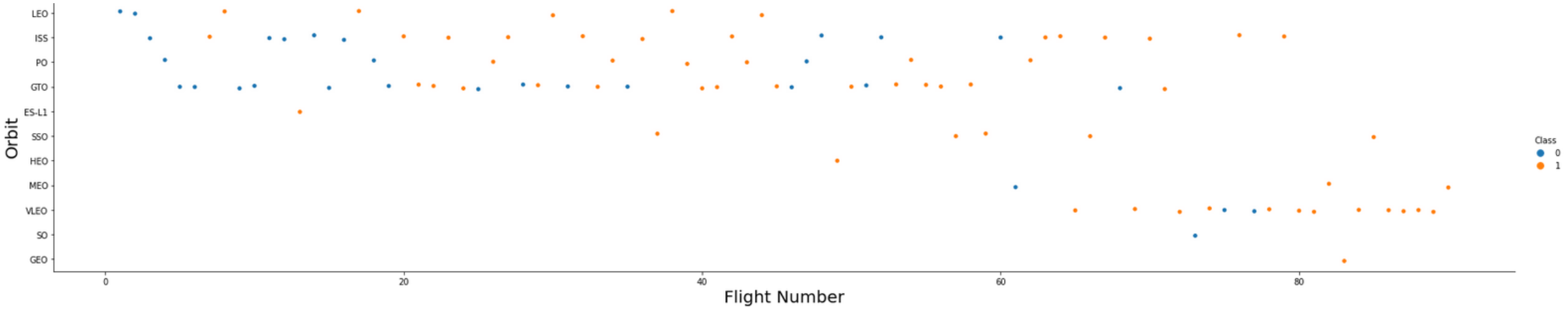
Success rate vs. Orbit type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

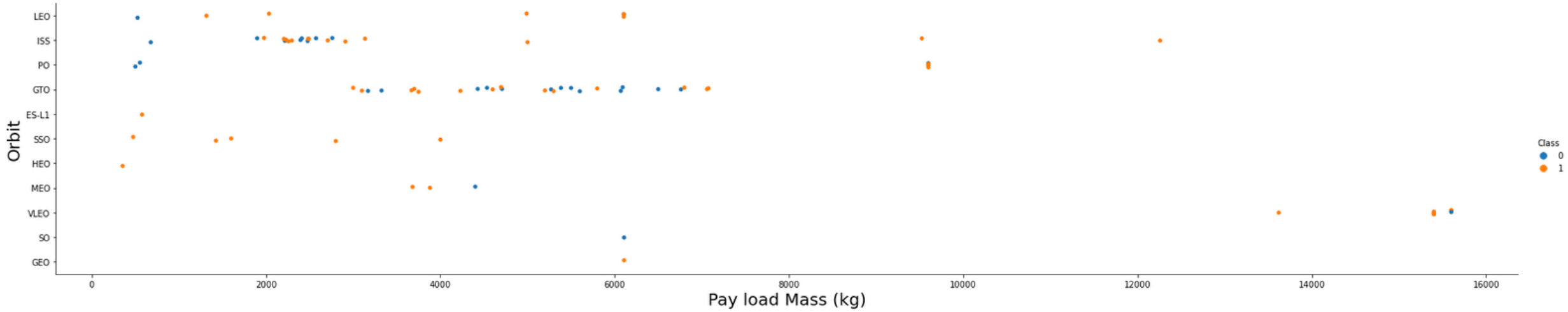
The orbits with the highest success rate are 4 orbits:ES-11, GEO, HEO, SSO. After these 4 comes the VLEO orbit that has a success rate above 80% and the least successful is SO

Flight Number vs. Orbit type



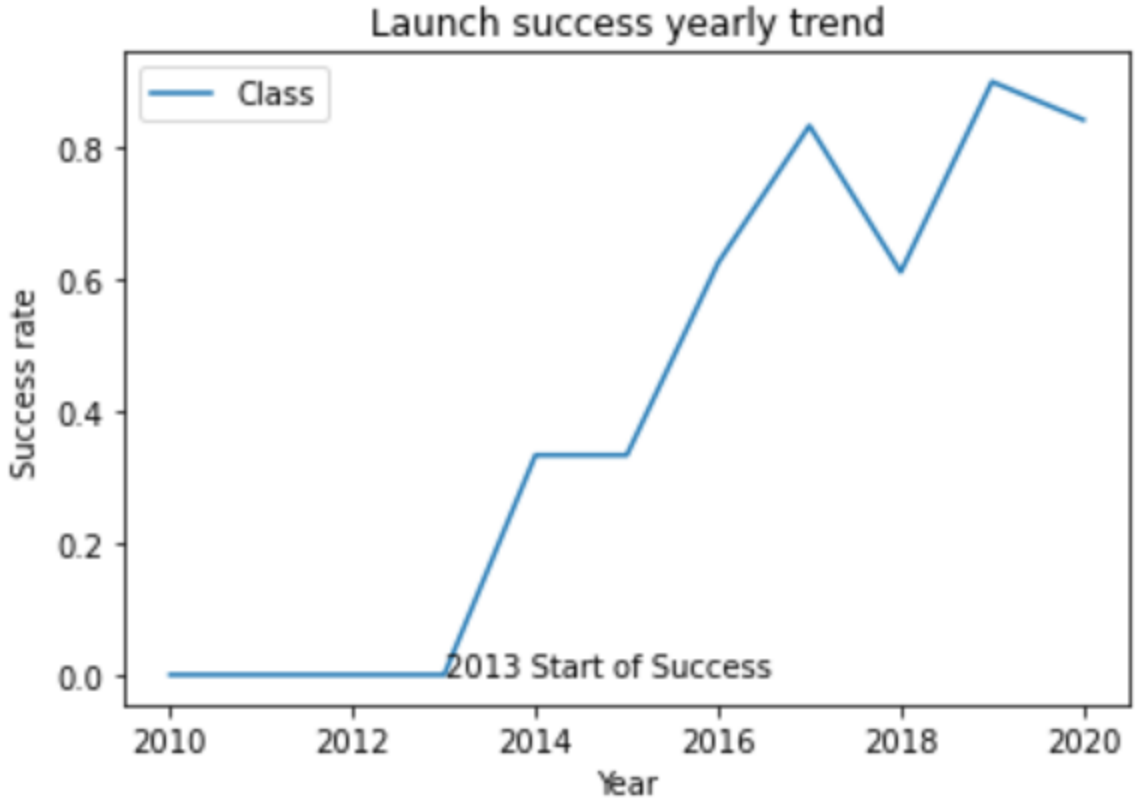
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit type



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch success yearly trend



you can observe that the sucess rate since 2013 kept increasing till 2020

EDA with SQL

The next sections will contain the exploratory data analysis done using SQL

All launch site names

We have 4 unique launch sites and they are the following:

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Launch site names begin with `CCA`

We have 2 launch sites that begin with CCA

CCAFS LC-40 and CCAFS SLC-40

But if we limit the query to the first five records we only end up with one site CCAFS LC-40

Total payload mass

NASA alone carried a total of 107010kg.

This is found by taking the sum of all payload mass when the customer is NASA

Average payload mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.400000

First successful ground landing date

The first successful landing outcome in ground pad was achieved on 2015-12-22

Successful drone ship landing with payload between 4000 and 6000

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

Total number of successful and failure mission outcomes

Mission outcome	Total number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

We observe that the majority of the mission succeeded

Boosters 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

2015 launch records

In the year 2015, the mission failed in two months, January and April as found below:

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

In the interval of 7 years, we observed a higher number of successful mission on the drone ship compare to the ground pad

landing__outcome	total_number
Success (drone ship)	5
Success (ground pad)	3

Interactive map with Folium

The next sections will contain some maps created using Folium

Marking all launch sites on a map

On this map we observe all launch sites' location markers on a global map.

Two launches are almost shown overlapped but if zoomed in they are not, it is only because we wanted to show all launches in one map.

Marking the success/failed launches for each site on the map

[23]:

On this map, we observe three launch sites and we created a marker cluster where we added the number of flights that succeeded and failed on each site

Distance between a launch site to its proximities

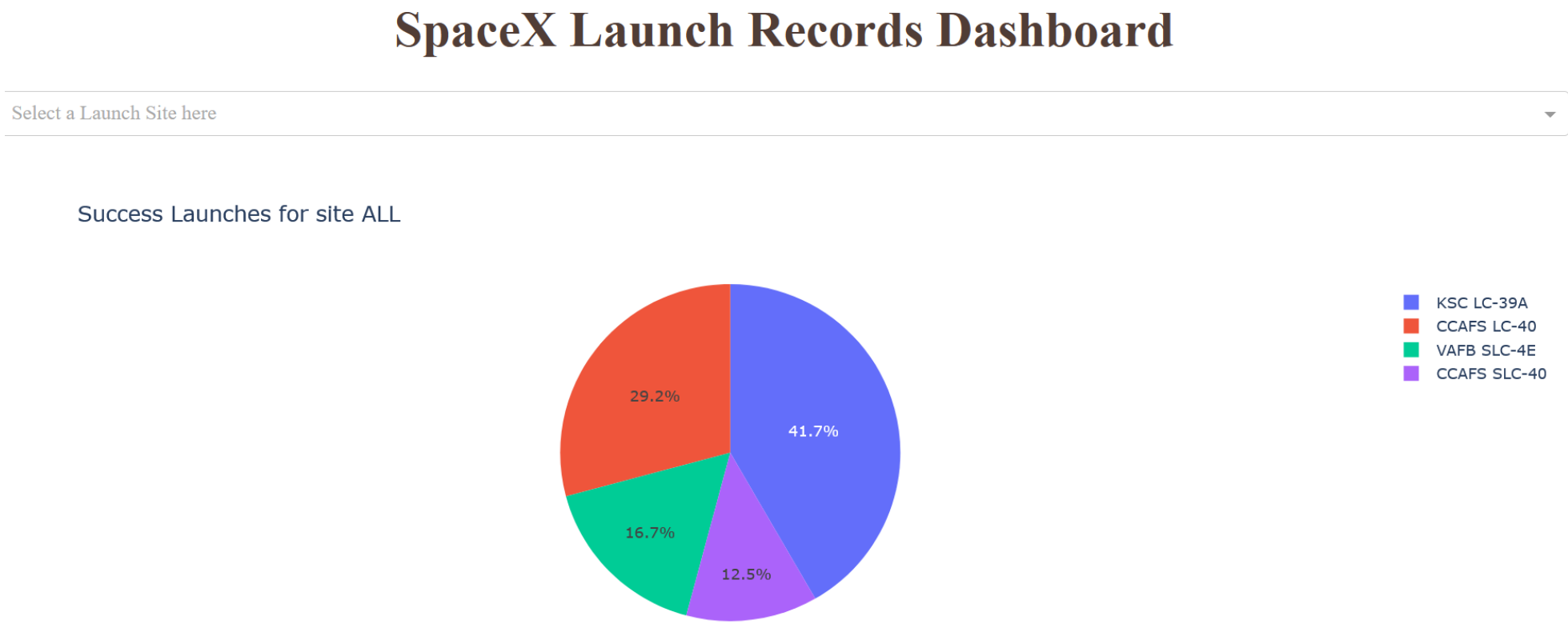
Here, the map shows the distance from a launch site to the closest railway with a calculated distance of 0.90 km.

Build a Dashboard with Plotly Dash

The next sections will contain the dashboard made with Plotly dash

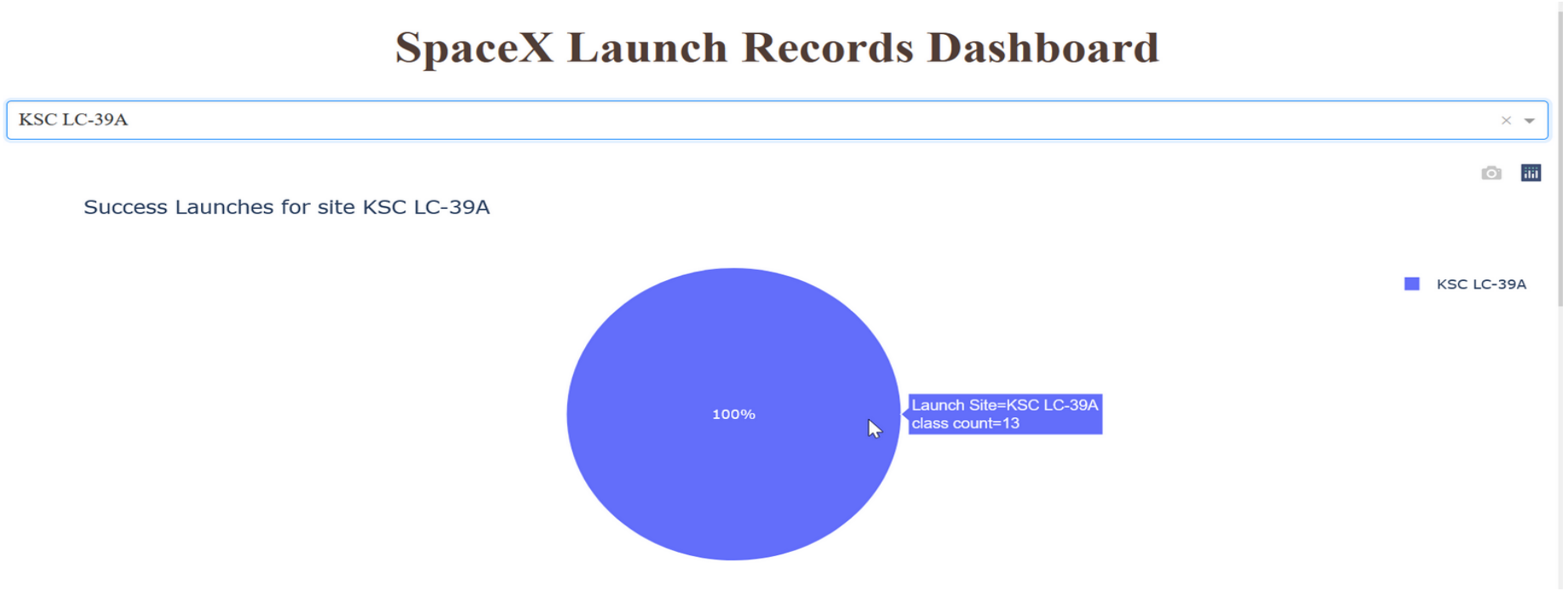
Launch success count for all sites

- It is observed that KSC LC-39A has the highest success launch
- Also the least successful launch site is shown to be CCAFS SLC-40



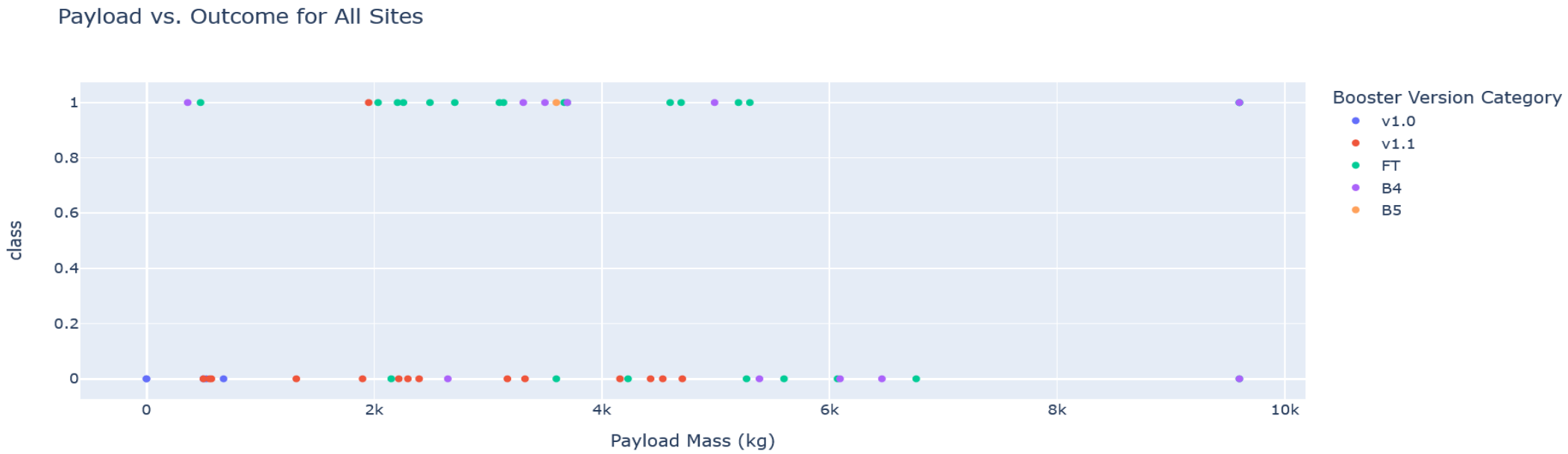
Piechart for the launch site with highest launch success ratio

The plot here shows that there are a maximum of 13 launches on the KSC LC-39 A site.



Payload vs. Launch Outcome scatter plot

Payload range (Kg):

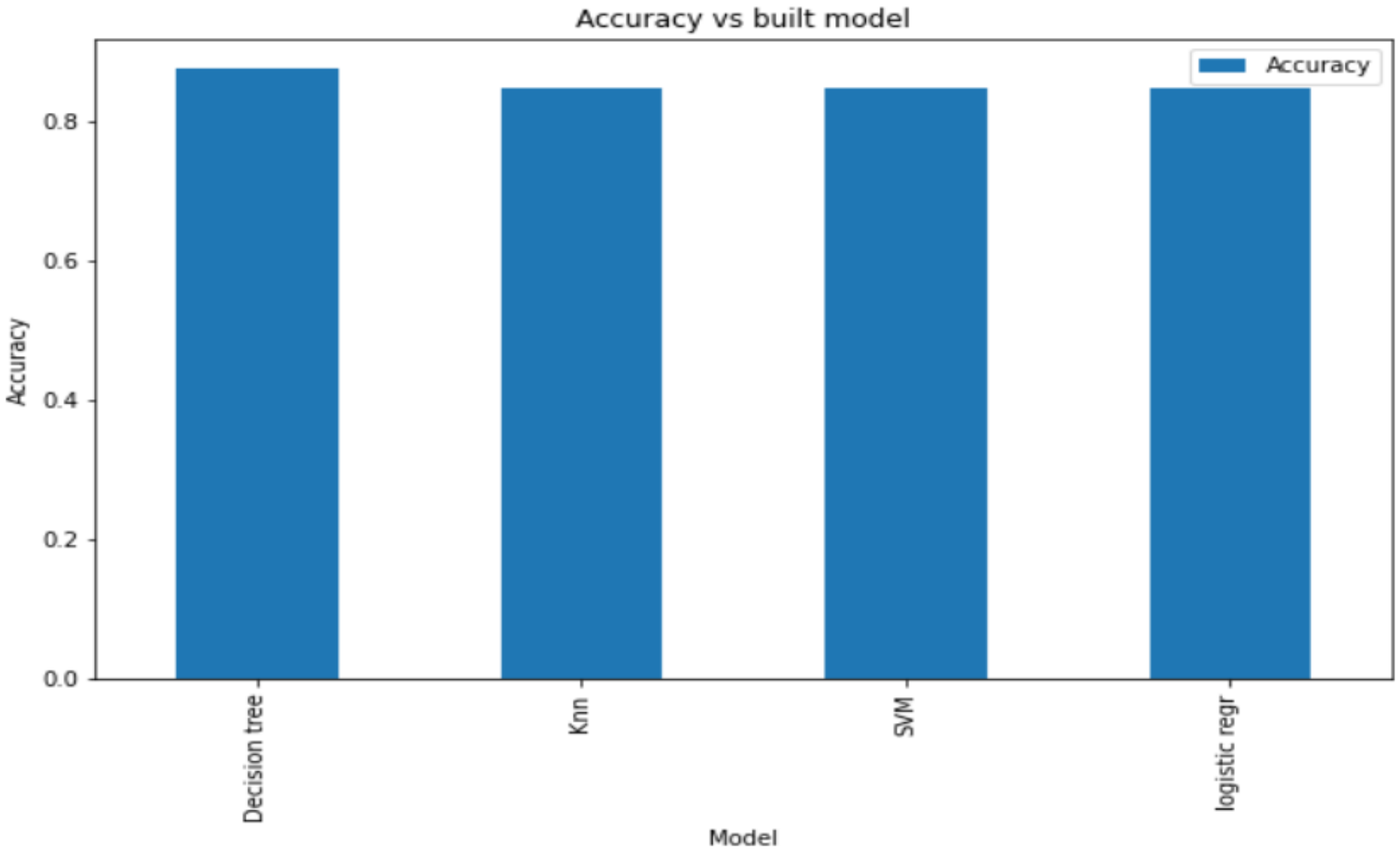


The plot shows the landing outcome for all sites with respect to the payload each launch carries. The colors of the dots showed on the plots denote the different booster versions.

Predictive analysis (Classification)

The next sections contain the predictive analysis conducted

Classification Accuracy

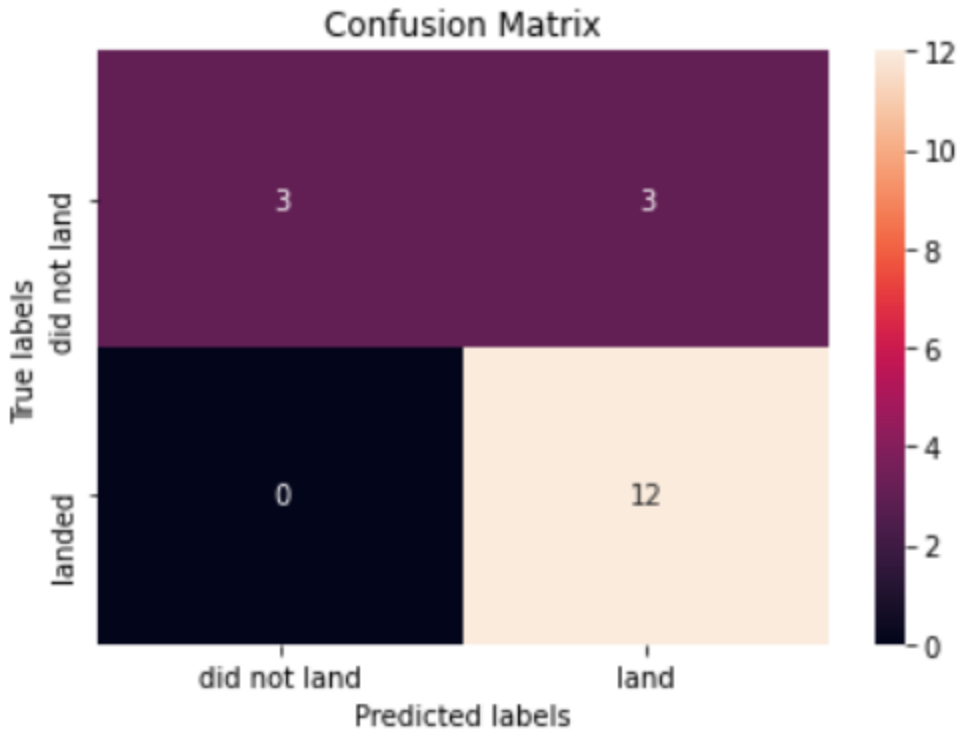


Decision tree model has the highest best score among the four models

Confusion Matrix

The confusion matrix shows a very good prediction for the landed flights however it does not perform well for the one that did not land.

There are 3 false negatives for the launches that did not land successfully. Which means they have been wrongly predicted.



CONCLUSION

- Point 1: We have shown that the Falcon 9 first stage landed successfully
- Point 2: It is clear that there exist a relationship between the number of flights each site had launched and its success rate. At the beginning, first stages were most likely to fail but the more the number of flights increases the better it gets.
- Point 3: To date it is shown that the maximum load that NASA has carried is 107010kg
- Point 4: It is observed that before 2013 there were no successful missions
- Point 5: The first successful landing on the ground was observed on 2015-12-22
- Point 6: The orbit SO is the least successful one with nearly 0% of success landing
- Point 7: The orbits that showed the highest success rate are 4: ES-11, GEO, HEO and SSO
- Point 8: Given that the dataset is small, it was not clear which of the four models perform the best.
- Point 9: The same analysis should be done on a larger dataset to determine which model to use for this type of problem.