IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Niare Doyom
16-07-2022

# Outline

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

# Executive Summary

- We used the spacex api to collect the data.
- And did one hot encoding to convert categorical data into numerical data
- We also used Beautifulsoup to scrape the spacex wikipedia but did not end up using that data
- We noticed that most of the launches where from near costal region and far away from cities
- We also noticed that as payload increases success rate decreases also
- Most of later landings were successful, although there are multiple factor that determines the success rate
- Most of the successful landings were done using the FT booster version
- Also landing site KSC LC-39A had the highest success rate, which mostly used the FT booster version
- Also using the booster version V 1.1 had the lowest success rate.
- Other than that Launch site CCAFS SLC-40 had the lowest success rate whose, whose favorite booster version was B4

# Introduction

- In this project we tried our best to predict whether the Falcon 9 first stage will land or not.
- Determining whether a first stage lands or not will help us determine the cost of the launch
- Our goal is to see what are the factors that affect the success rate of landing and how exactly do they affect the success rate, once we determine what are the factors which will most likely lead to a successful landing we will test it against our prediction models to see how exactly does it work.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

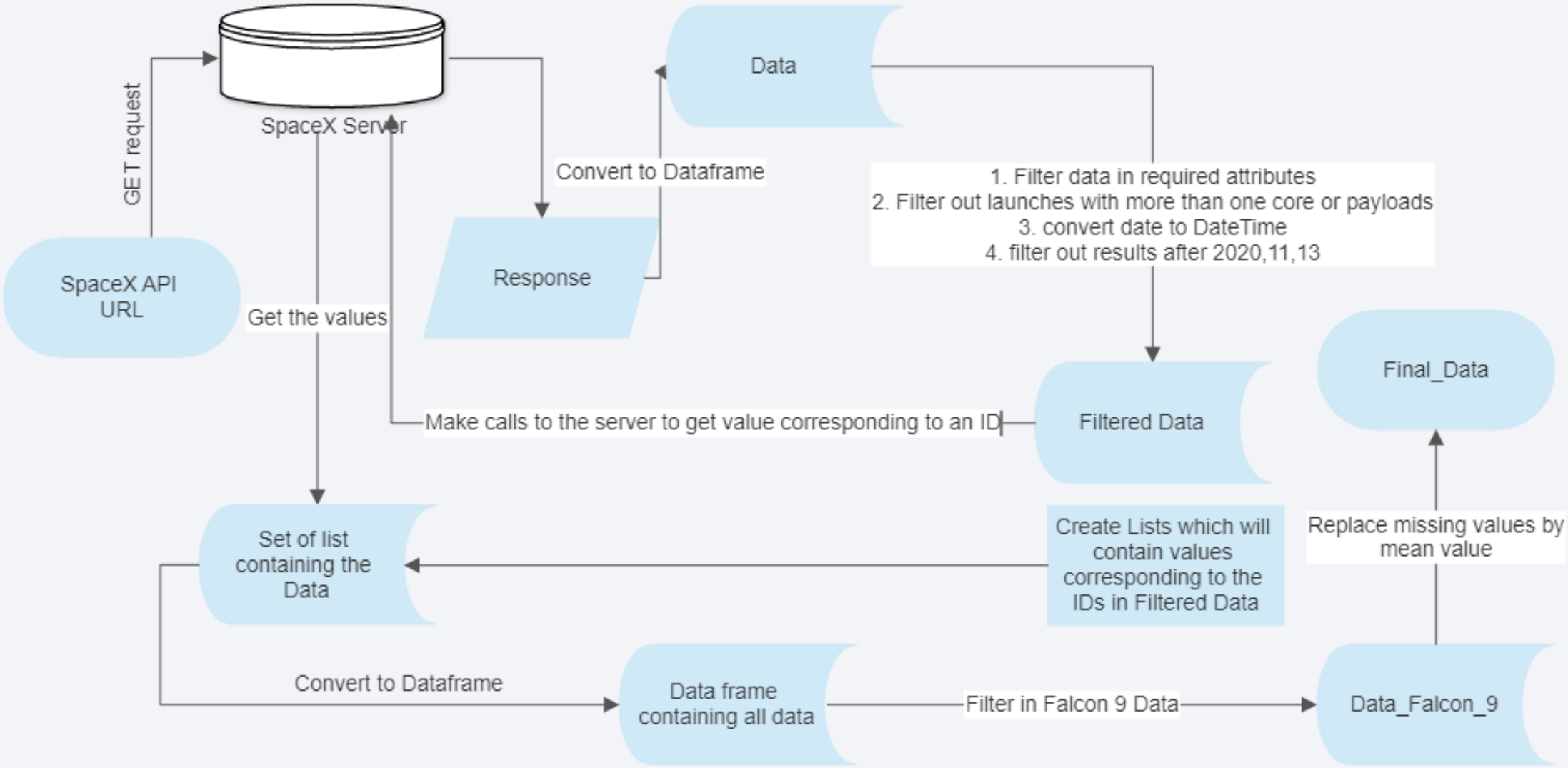
# Data Collection

- We Used space X REST API to collect the data.
- We also scrapped the SpaceX Wikipedia page to collect the data
- The data collected from the space X REST API ended up being much cleaner than the data scraped from SpaceX web page.

# Data Collection – SpaceX API

- We use the SpaceX API URL to perform a GET request to the SpaceX sever.
- Which returns a response which gets converted to a JSON file and gets converted to a Data Frame.
- Then from out data we filter in the attributes that we need.
- We filter in only the launches with a single core and payload we convert the date into datetime
- In our data some of the columns still don't have the actual value but IDs to the actual value
- First we create a bunch of list which will later serve as column value to our final filtered data frame
- Then we use the function that we have created to request the value of the data from the API using the IDs
- Then we store them into the list, once we have all the relevant data in the list we form a dictionary with it which is then converted to a data frame.
- Finally we filter out the falcon 9 data and replace the missing value by the mean of the column value.
- Github Link

# Data Collection – SpaceX API

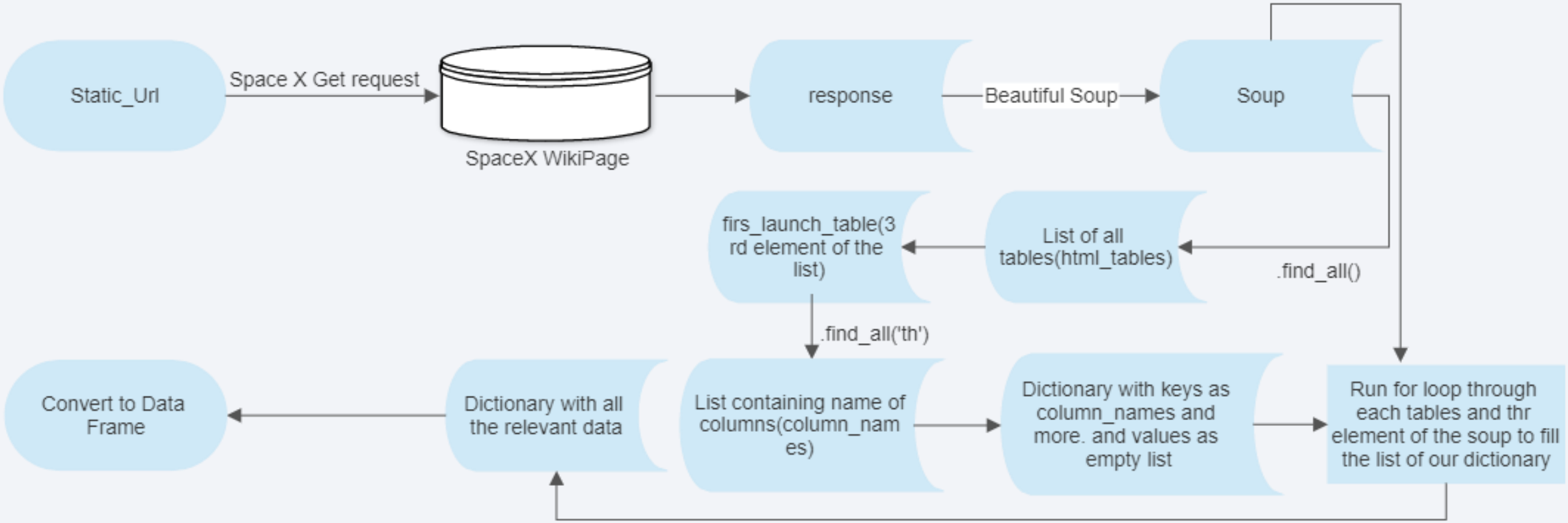


# Data Collection - Scraping

- First like before we have the static URL
- Then we make a GET request to SpaceX wikipage and get html code in response form as a response
- Then we convert it into a soup object using the beautiful soup library
- We then use the fin_dall method to get all the tables from the soup object
- We see that the 3rd table contains column names and use that to acquire all the column names,
- We then create a dictionary with those column name as the key for the dictionary and value as  empty lists
- We again go through the soup object to secure all the value for the keys of our dictionary
- Then once we have the dictionary with all the relevant data, we convert the dictionary to a data frame.

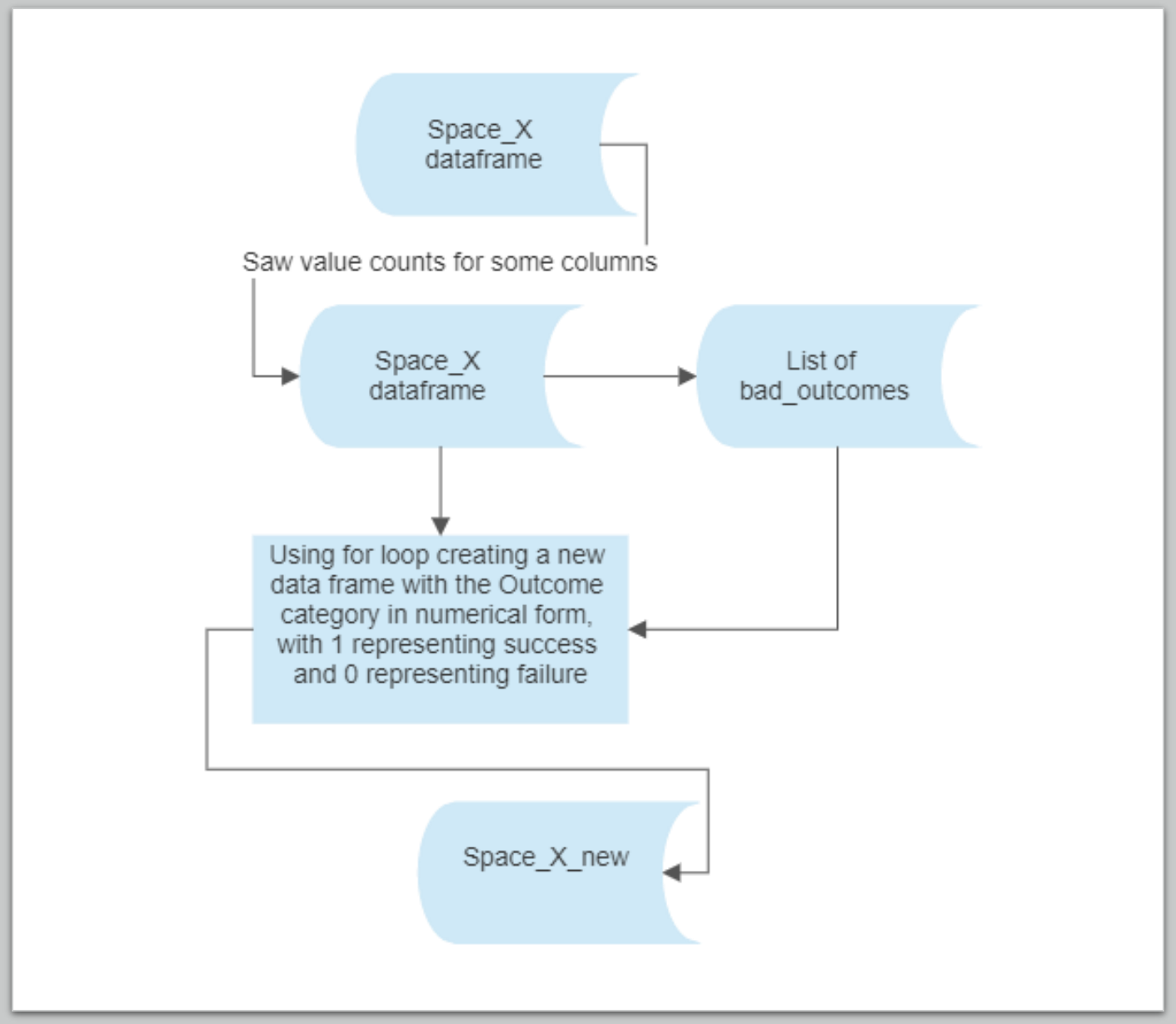
- Github Link

# Data Collection - Scraping



# Data Wrangling

- We have the Space_X Data Frame
- We first start out  by doing some exploratory data analysis and finding value_counts for columns in our dataframe
- We did the value counts for launch sites, Orbit and Outcomes
- Then we converted the categorical outcome column to numerical outcome column
- [Github Link]



# EDA with Data Visualization

- We made a scatter plot of Flight number vs Payload mass, Flight number vs launch site, Flight number vs orbit and Orbit vs Payload mass to see how those values were related to each other and the success.
- We made bar graph between Success rate vs Orbit, Years vs success rate, to see in which orbit did they find most successes and to see how the landing rate improved as we go through the years.
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- Github Link

# EDA with SQL

- Listed distinct Launch Sites
- Listed launches which was from the launch site whose name starts with CCA
- Listed total payload mass carried by the boosters from NASA
- Listed average payload mass carried by booster version F9 v1.1
- Listed the date for which the first stage landed successfully for the first time
- Listed Total number of mission successes and failures
- Listed booster version which carried the maximum payload
- Listed month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Listed the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- Github Link

# Build an Interactive Map with Folium

- Marked all the launch sites in the map to see how far are they from each other and how they are located in the map of US
- Marked success and failed launches for each site on the map to get a visual indicator of success rate vs site
- Performed marker clustering for the above since some of the launch site are very close to each other
- Calculated distance between launch sites to its proximities and created a line between them, to see if the launch sites are close to various landmarks
- Github Link

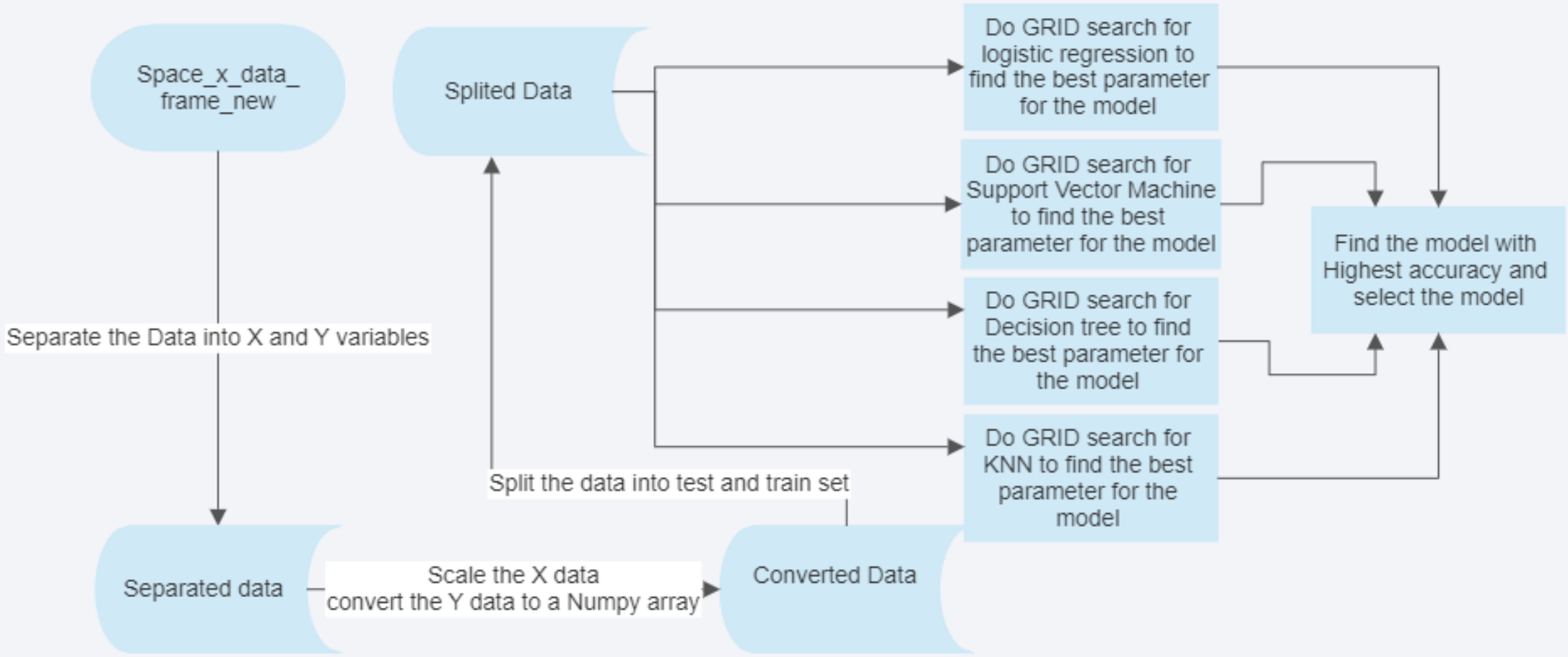
# Build a Dashboard with Plotly Dash

- Added a drop down menu to select locations on dash web app.
- Added a pie chart displaying number of successful launches and number of unsuccessfull launches for a specific location, or if one selects the all option It will show the number of successes in each location
- Added Scatter plot for, Payload mass vs success for either a specific location or all the location, with hue being booster version
- The pie chart gives us an idea of which location had the highest successes
- The scatter plot gives us an idea of which, which booster version had the highest success and what is the ideal payload to take in flight.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- Github Link

# Predictive Analysis (Classification)

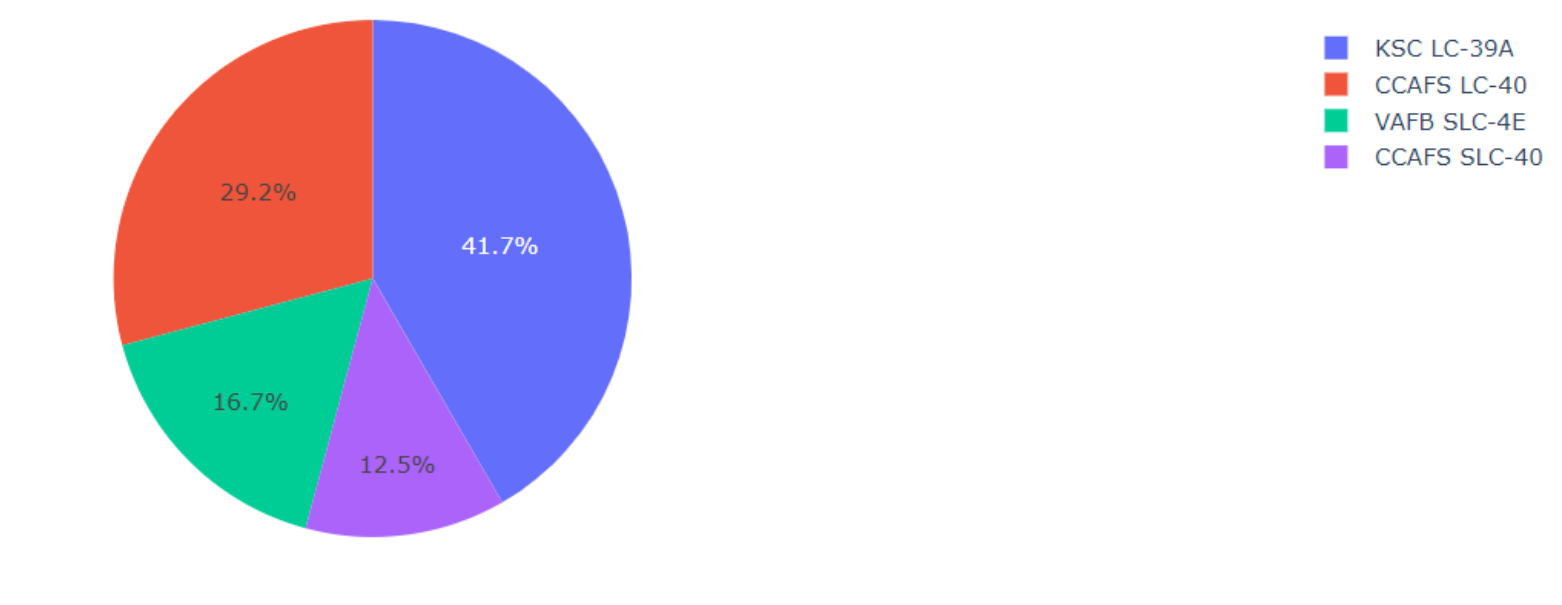
- Space X was separated into X and Y variables,
- Then scaled using the Standard Scaler method.
- Y data was converted to NumPy array,
- Data was splitted into testing and training set,
- Data was fed into GRID search for logistic regression, support vector machines, decision tree and KNN to find the best parameters for these models, then we selected the model with highest accuracy for prediction
- Github Link

# Predictive Analysis (Classification)



# Results

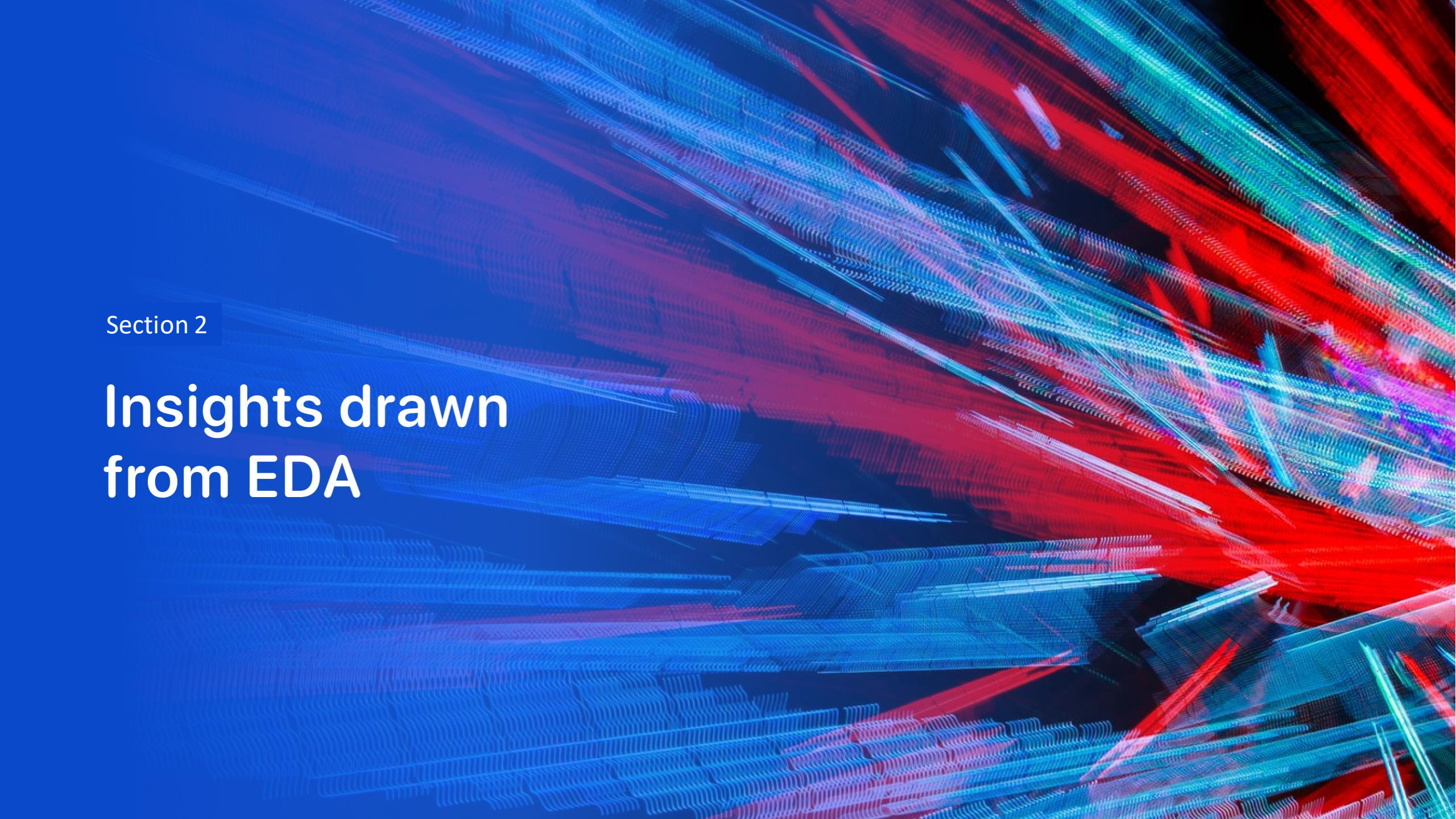
- It seemed success rate of the data increased with flight no.



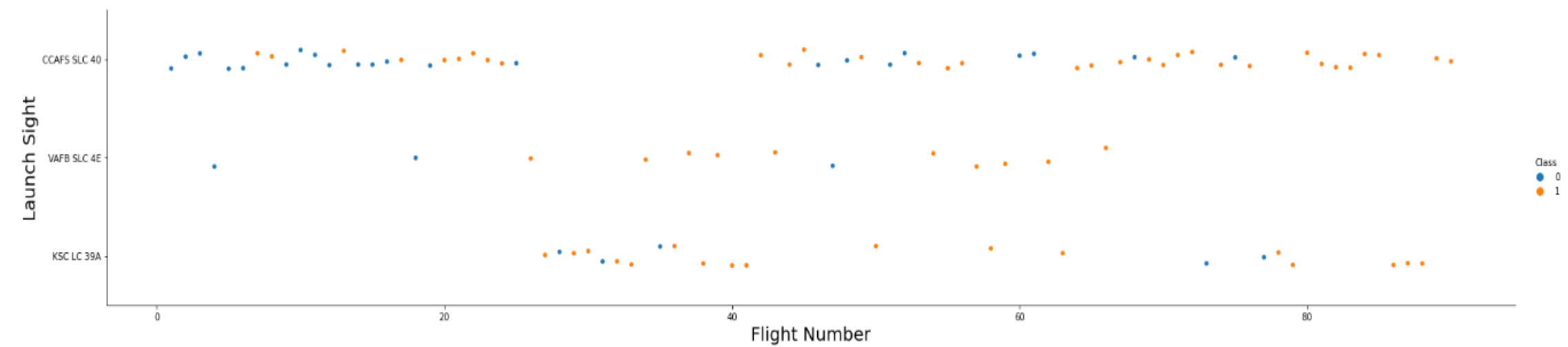

- It seems the maximum R score all the algorithm used were around 0.833

Section 2

# Insights drawn from EDA

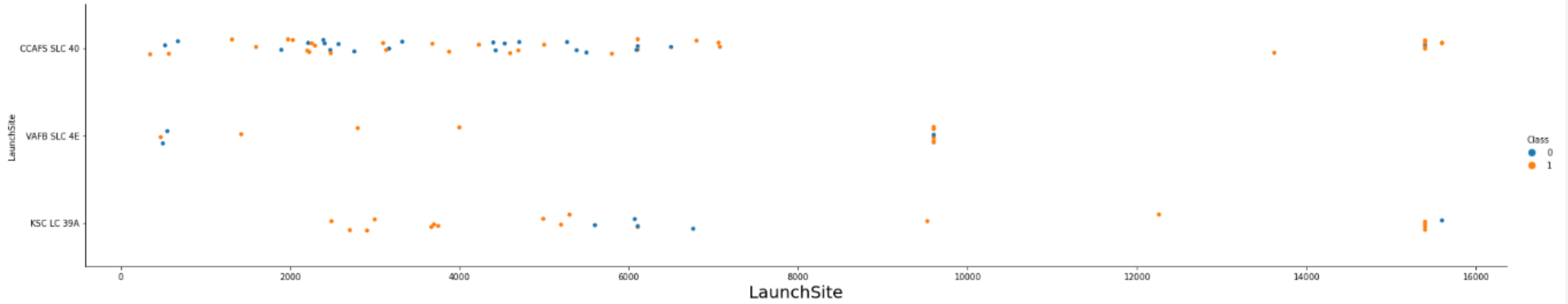


# Flight Number vs. Launch Site



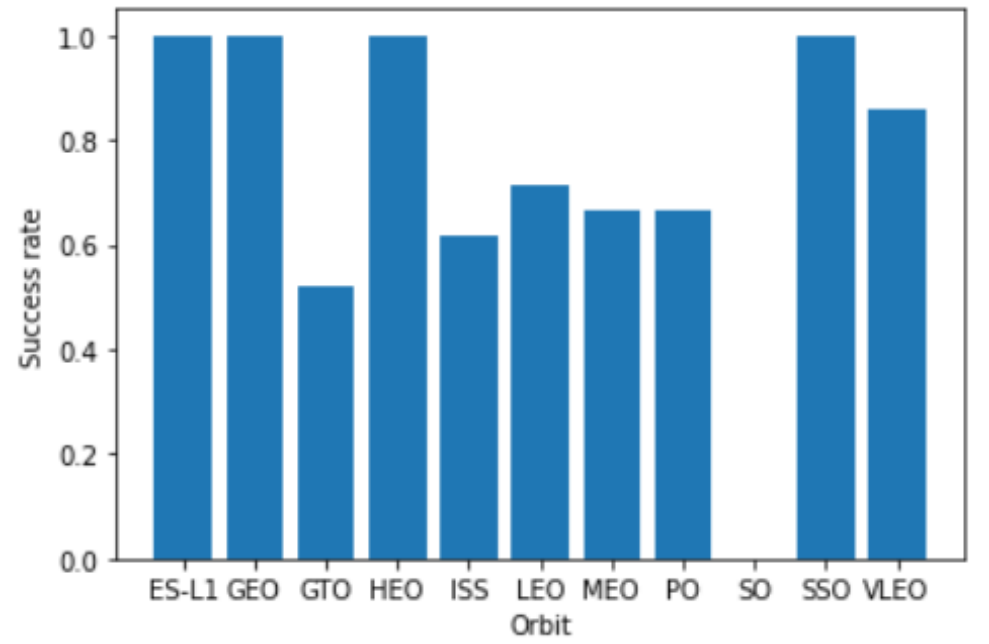

- It seems most of the earlier launches were done on CCAFS SLC 40, with most of them failing to land, most of the later launches were also done on the same location but most landed successfully this time, VAFB seems to have lowest no. Of launches scattered through out. KSC seems to have the second highest no. Of launches with most of the landings, being successful.

# Payload vs. Launch Site



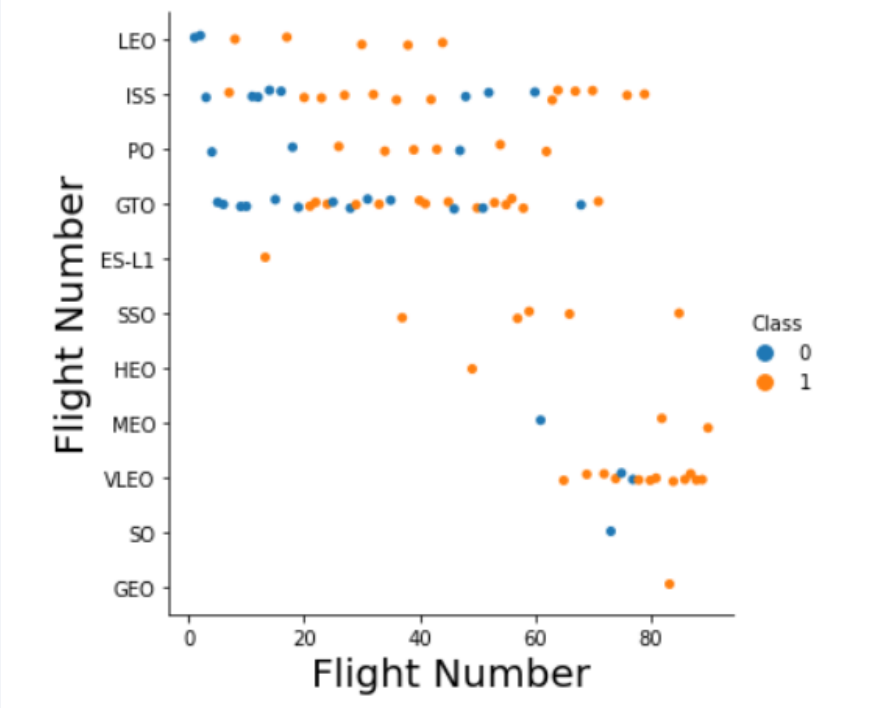

- Interestingly it seems most of the launches with payload around 9650 were done from the launch site VAFB SLC 4E
- Most of the Launches with very heavy payload were done from CCAFS SLC 40 and KSC LC 39A

# Success Rate vs. Orbit Type



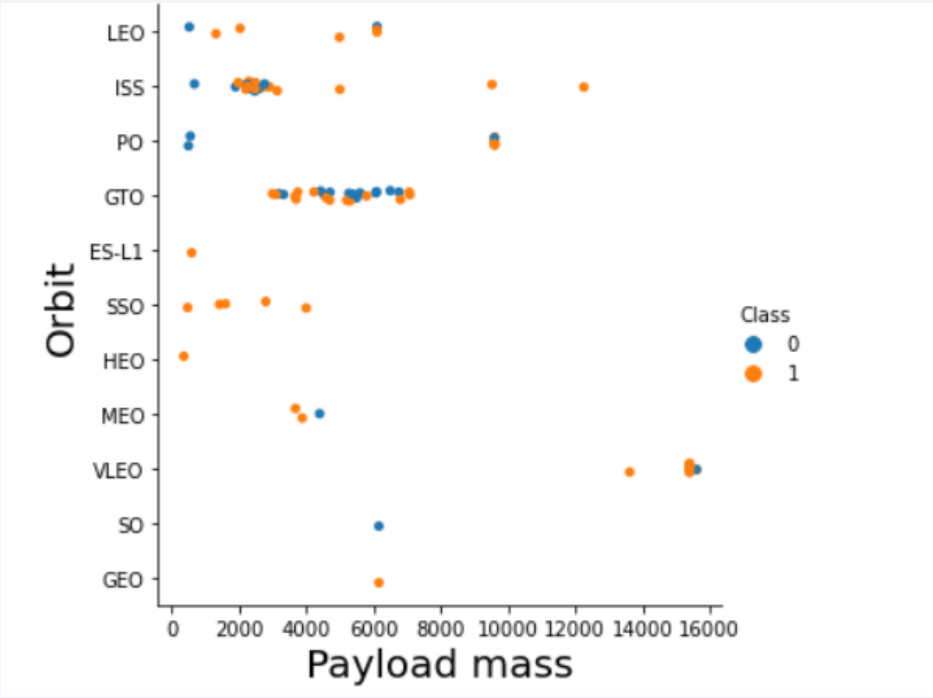

- It seems most of the successes were or HEO, SSO, GEO, ES-L1 launches

# Flight Number vs. Orbit Type



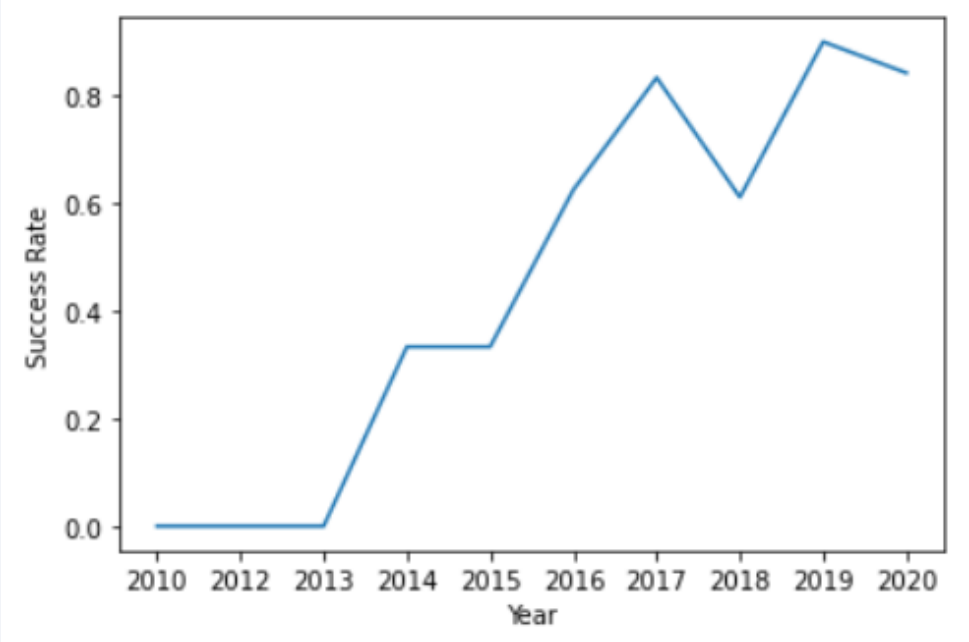

- Seems most of the later flights were for VLEO, and we also see a trend that as flight number increases number of successes increase for each orbit, but not for GTO orbit

# Payload vs. Orbit Type




- As seen in the graph most of the heavy payloads are for VLEO, medium payload for GEO and lightest payload for ISS, as seen before GTO's success rate is pretty low.

# Launch Success Yearly Trend




- As seen in the graph above success rate started growing from 2013 it had a dip and 2018 but went back up again starting 2019

# All Launch Site Names

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

As you can see there are 4 distinct launch sites in our data

# Launch Site Names Begin with 'CCA'

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

- The above are the records for launches whose launch site name begins with 'CCA'

# Total Payload Mass

| SUM("PAYLOAD_MASS__KG_") |
| --- |
| 99980 |

- The above is the total payload mass carried by boosters launches by NASA

# Average Payload Mass by F9 v1.1

| AVG("PAYLOAD_MASS__KG_") |
| --- |
| 2534.6666666666665 |

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

# First Successful Ground Landing Date

| MIN(Date) |
| --- |
| 22-12-2015 |

- The above is the date for first successful landing

# Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The above are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

# Total Number of Successful and Failure Mission Outcomes

- The Below gives us an Idea of the total number of successful and failure mission outcome

| Mission_Outcome | COUNT("Mission_Outcome") |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# 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 |

- The above gives us the names of the booster which have carried the maximum payload mass

# 2015 Launch Records

| substr(Date,4,2) | Landing _Outcome | "Booster_Versions" | Launch_Site |
|---|---|---|---|
| 01 | Failure (drone ship) | Booster_Versions | CCAFS LC-40 |
| 04 | Failure (drone ship) | Booster_Versions | CCAFS LC-40 |

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

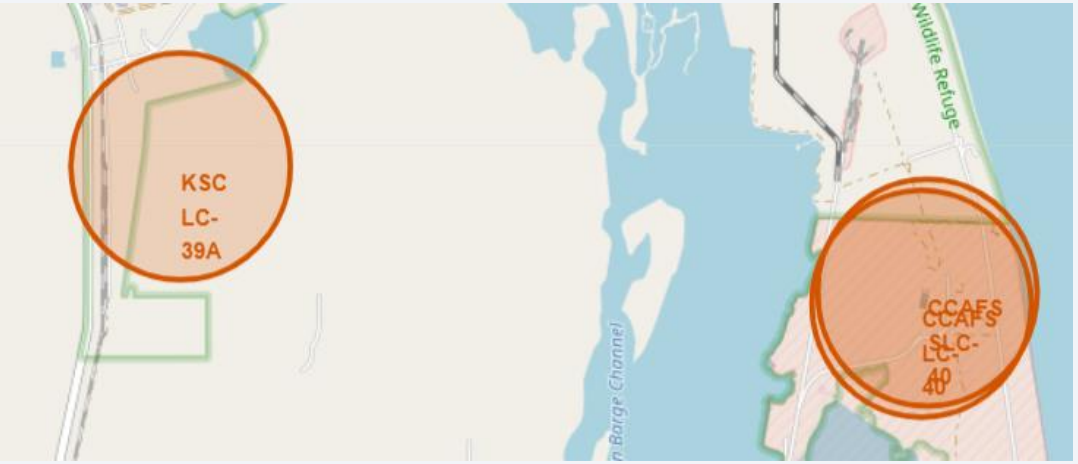
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 08-04-2016 | 20:43:00 | F9 FT B1021.1 | CCAFS LC-40 | SpaceX CRS-8 | 3136 | LEO (ISS) | NASA (CRS) | Success | Success (drone ship) |
| 06-05-2016 | 05:21:00 | F9 FT B1022 | CCAFS LC-40 | JCSAT-14 | 4696 | GTO | SKY Perfect JSAT Group | Success | Success (drone ship) |

- The above ranks 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


Section 3
Launch Sites
Proximities Analysis

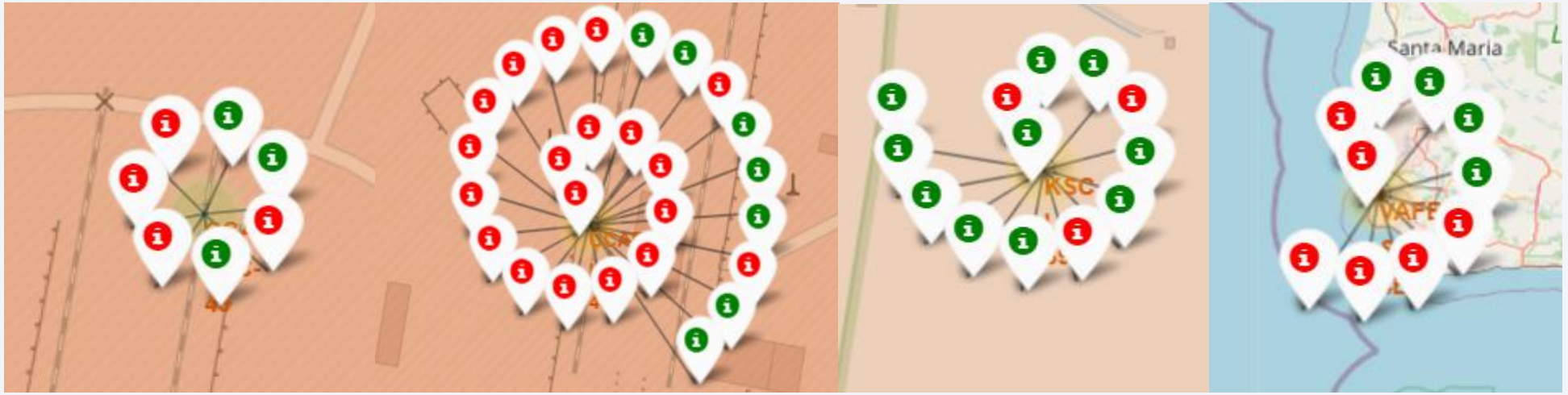
# Geography Analysis


- As evident in the above picture most of the launches were from the coastal region
- Most of the Launches were from near the equator region
- One of the launch site is on the opposite side of the country from the other 3 launch site

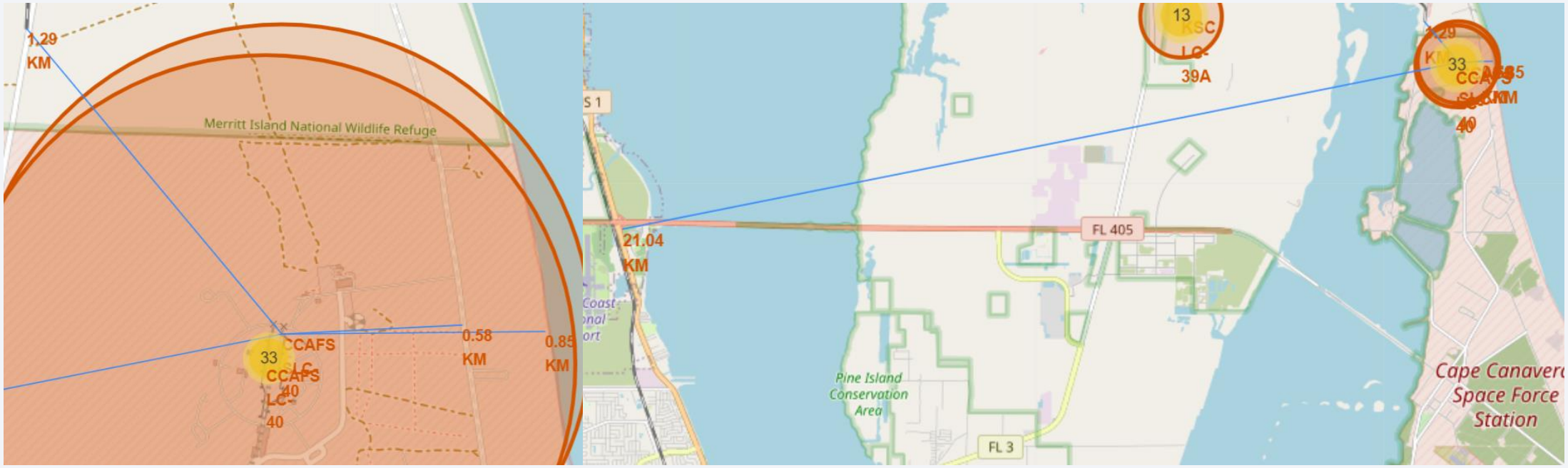
# Geography success analysis

CCAFS SLC 49 CCAFS SLC 40 KSC LC-39A VAFB SLC-4E

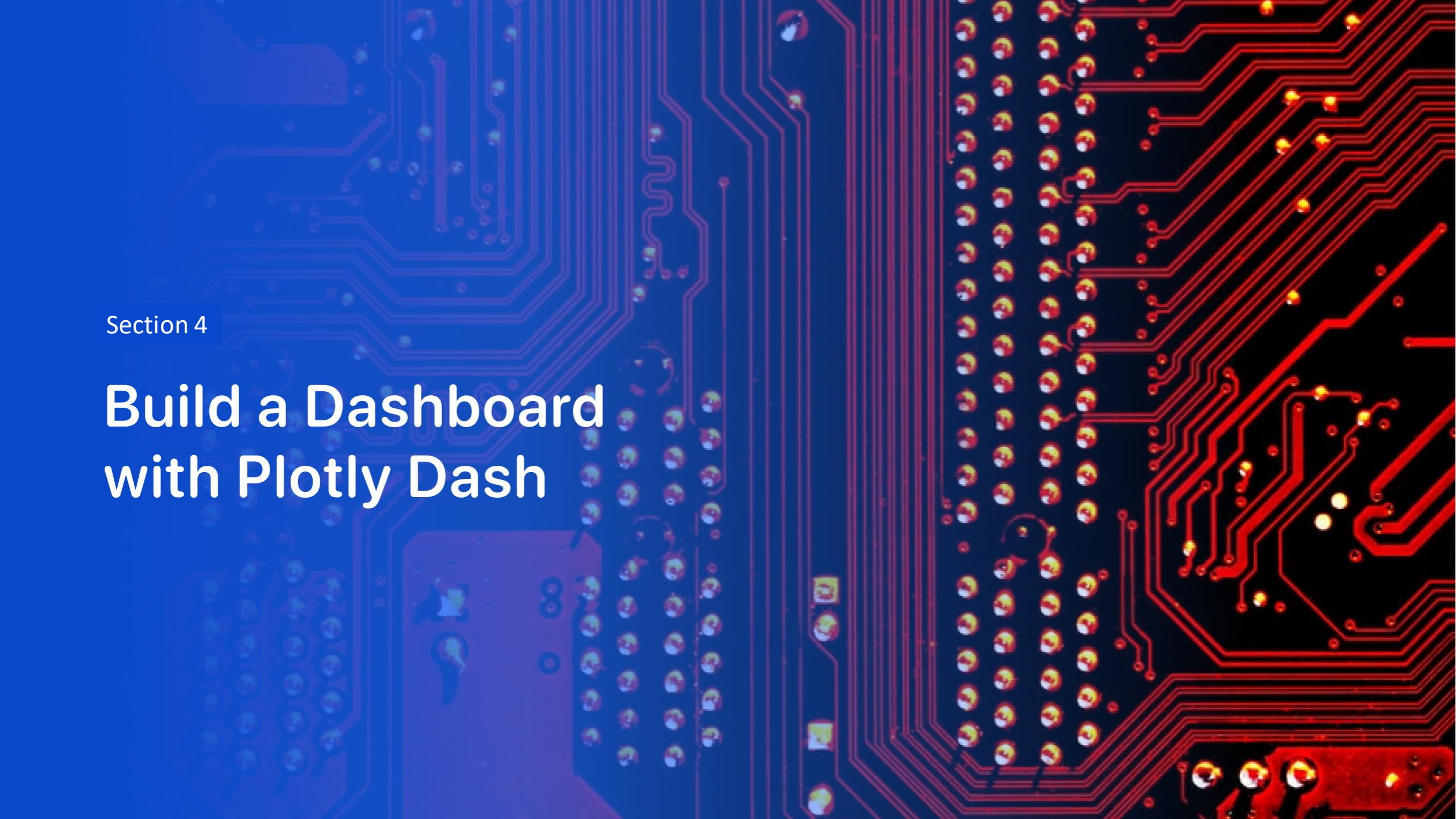



- So evidently KSC LC-39A had the highest rate of success, and CCAFS SLC 40 was the location from where most launches were done.

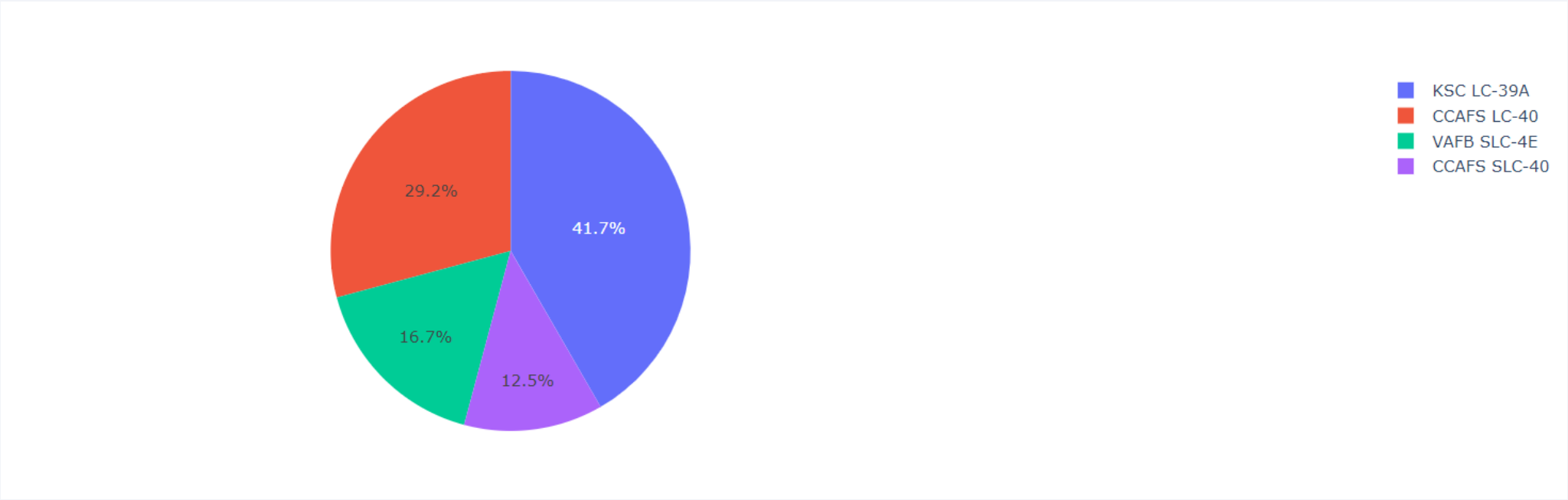
# Launch site proximity analysis




- As evident from the graph above nearest railway station and highway are pretty close to launch site, most likely because it helps in transportation of heavy launch materials
- The cities are located pretty far away from the launch site, mainly for safety reasons is my guess

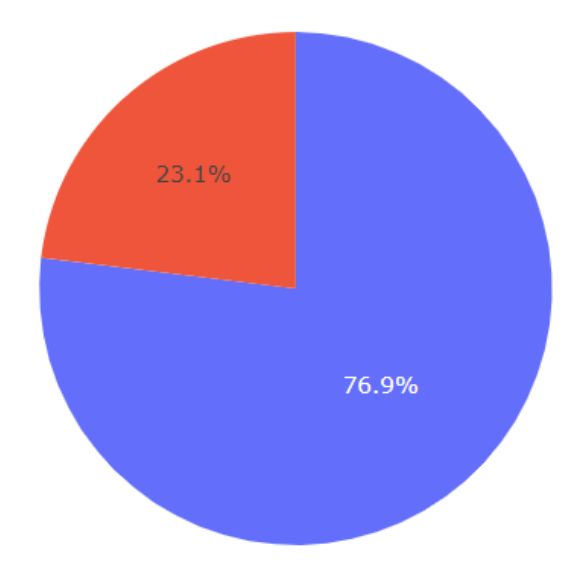

Section 4
Build a Dashboard with Plotly Dash

# Pie of Success



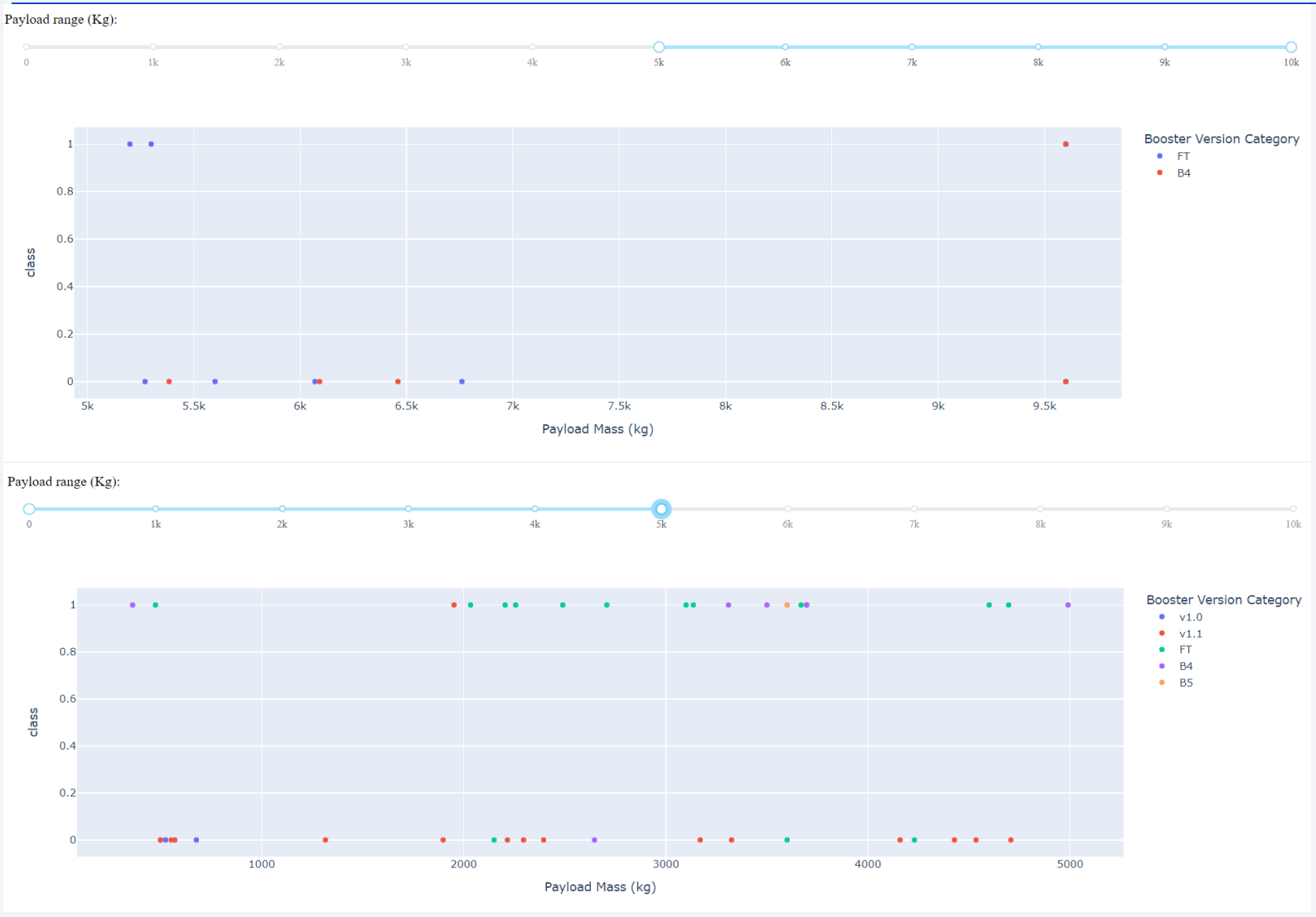

- As seen in the chart above KSC LC-39A had the most amount of successes
- CCAFS SLC-40 has the least amount of success

# How successful is most successful?




- It seems the most successful launch site has a success rate of 76.9%

# Payload success analysis



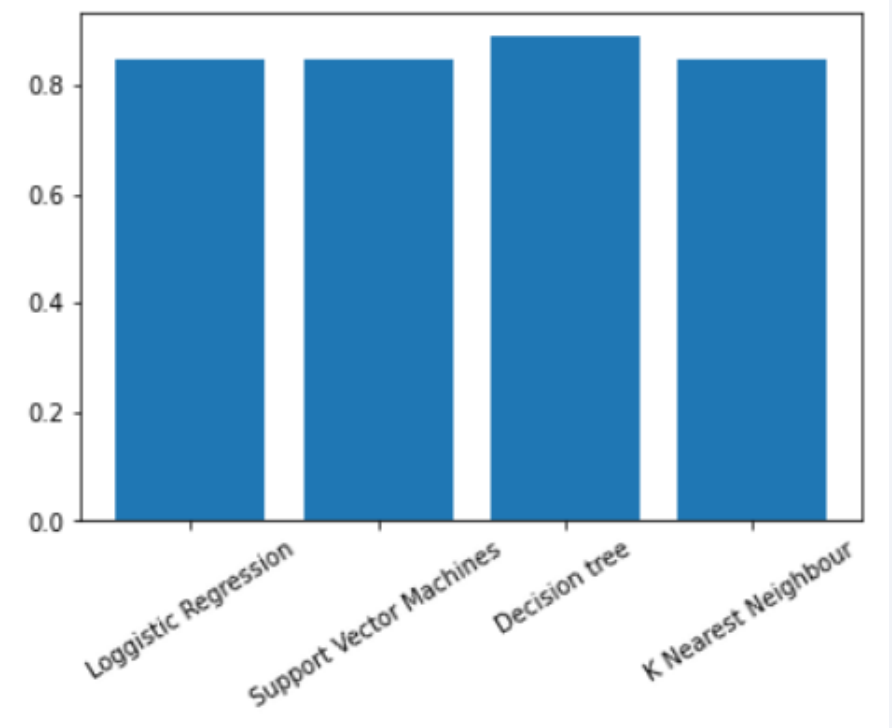
# Payload success analysis

- As seen the graphs there are not a lot of successes for high payload range, but for low payload range the successes seem to be equally distributed
- Most of the success were from the FT booster version

Section 5

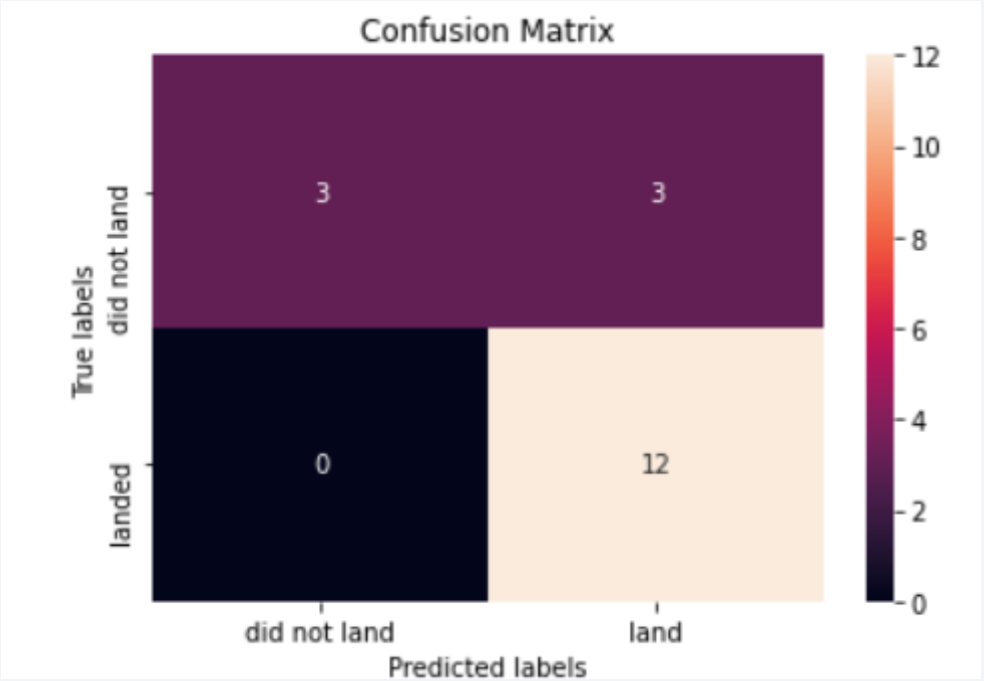
# Predictive Analysis (Classification)

# Classification Accuracy




- All the model has the same testing data accuracy, using the training data accuracy to judge the model. Based on that it seems decision tree has the highest accuracy

# Confusion Matrix




- All the model gave same confusion matrix, it seems the models have difficulty identifying false positive.

# Conclusions

- In this analysis we have saw dependence of success rate on various factors, with the most trivially impactful factor being date of launch, and most non trivially impactful factor being booster version, orbit also played an important role
- We analyzed the geography of launch site realized they are near equator region and coastal region, we saw most of the successful launches were from KSC LC39A
- We made dash app to see relation between success rate vs location and payload vs success, realized most of the success were for low payloads.
- Finally we saw the most successful machine learning model was decision tree which gave a accuracy of 8.8+

# Appendix

- Github Link
- My Notion notes

# Thank you!