IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sanket Anand
10-03-2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The following method were used to collect and analyze data
  - Data Collection - Data was collected from Space X API and through webscraping.
  - Data Wrangling - Data was cleaned and prepared for analysis using pandas
  - Data Analysis was done using visualization approach. Libraries such as seaborn, folium and plotly were used.
  - Machine Learning models were used for prediction on unseen data.
- Summary of all results
  - Past launch data from SPACE X was collected
  - Null values in the data were dealt during data wrangling
  - Exploratory Data Analysis helped to determine best features for analysis
  - Machine Learning prediction helped to find the best model having highest accuracy

Introduction

- Past launch data by space X will be used throughout the analysis to determine the best ways which can be used to determine the success of the mission

- Our objective here is to find answers for the following questions
  - The best ways to determine the success chances of a launch.
  - Best ways to reduce the overall cost for the launch
  - Best ways to pick a site for the launch

Section 1

Methodology

Methodology

Executive Summary

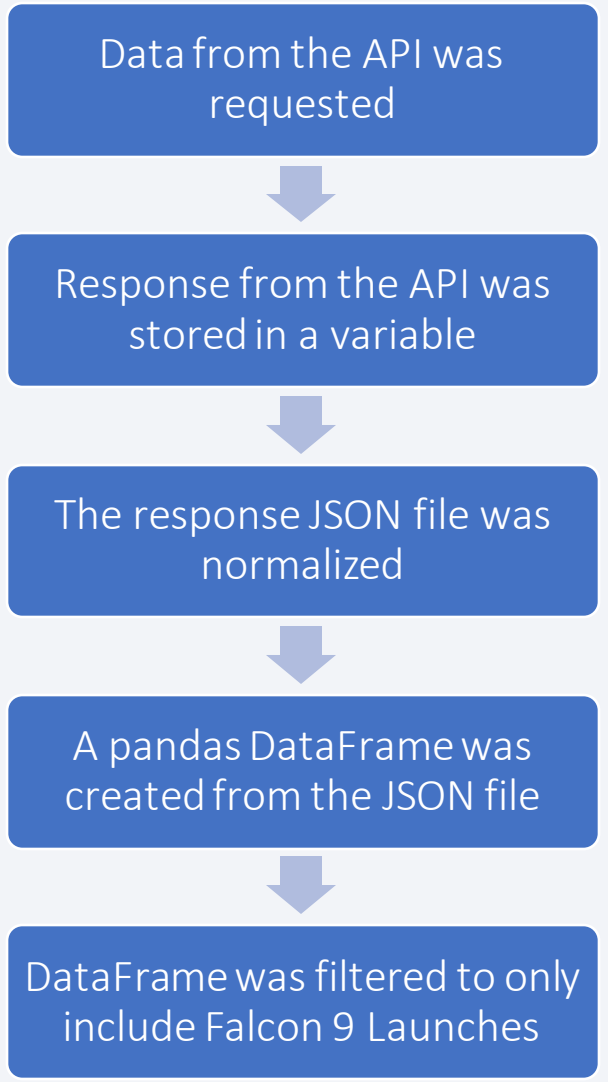
- Data collection methodology:
  - Data was collected using two sources
    - SPACE X API - https://api.spacexdata.com/v4/rockets/
    - Webscraping - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Missing data was replaced with the mean of all the data instances
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Using SPACE X API
  - Using the ‘requests’ python library data from SPACE X API was requested.
  - The result JSON file was normalised and converted into a pandas dataframe.
  - Unnecessary columns were removed from the data frame
  - Data was then filtered to include only Falcon 9 Launches
- Using web scraping
  - Using the ‘requests’ library Falcon 9 and Falcon Heavy Launches record from wikipedia was requested
  - Beautiful Soup (python library) object was created from the HTML response
  - All the columns name and data were extracted and converted into a pandas dataframe

Data Collection – SpaceX API

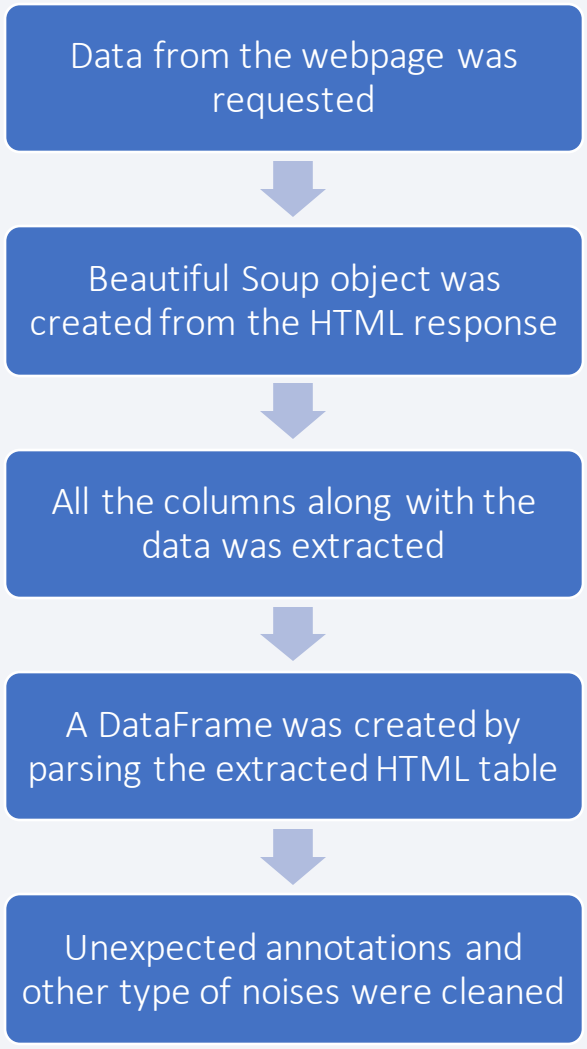
- The SPACE X API was used according to the flowchart to extract data



- Source code - Data collection using API

Data Collection - Scraping

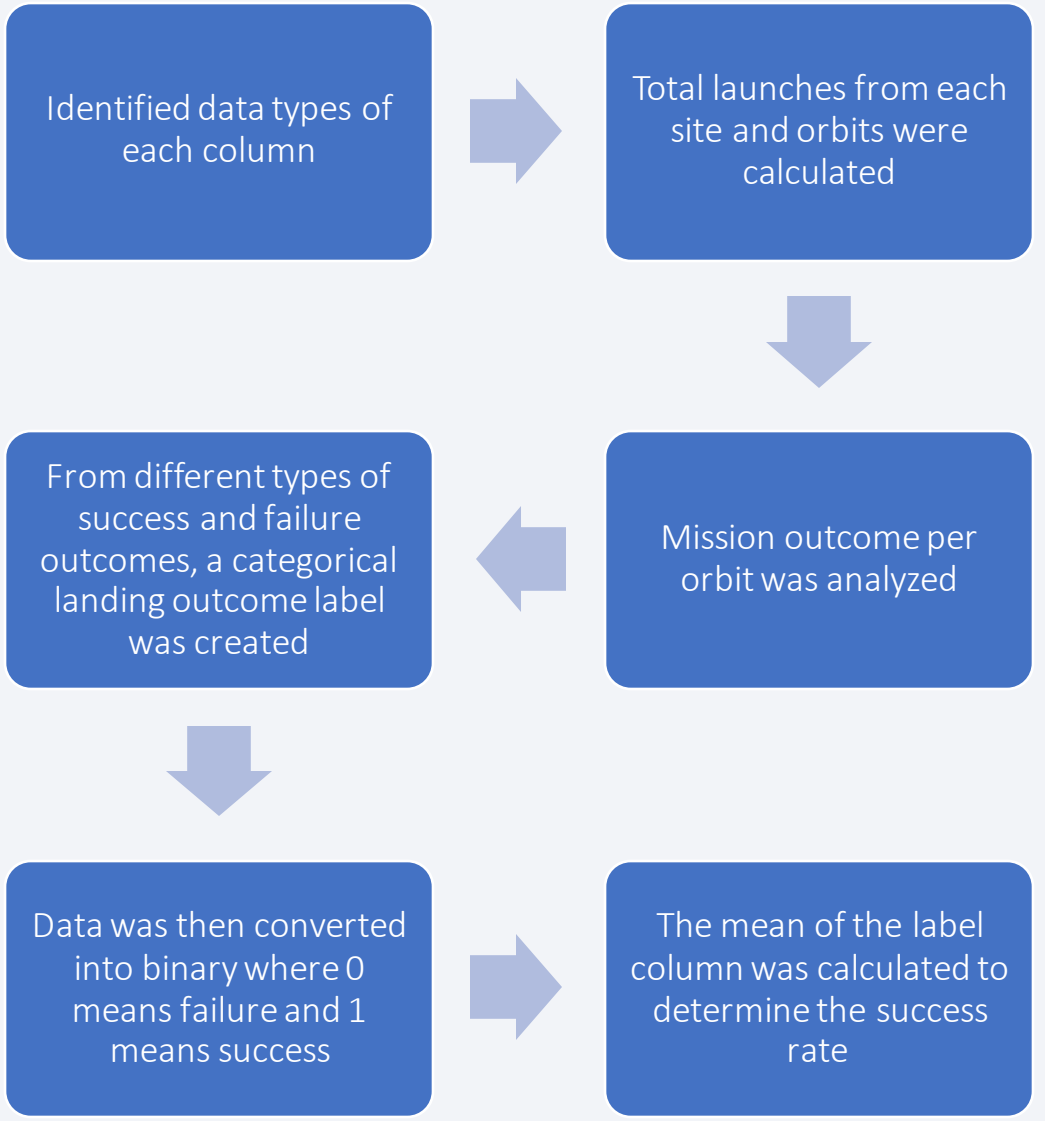
- Data from the website was scraped according to the following flowchart.



- Source code - Data collection with webscraping

Data Wrangling

- Data Wrangling was done using the pandas dataframe we created.
- Data was cleaned for analysis according to the following flowchart
- Source code -[Data Wrangling]



EDA with Data Visualization

- Scatter plots, Lineplot and Barplots were made using Seaborn python library.
  - Scatter plots were used to determine relation between two features and their impact on the label
  - Bar plot were used to determine success rate for various orbits launches
  - Line plot was used to understand the growth in success rate over the years

- Source code - EDA with Visualization

EDA with SQL

- Following SQL queries were performed on the data

  - Names of the unique launch sites used for the launches
  - **5 records where launch sites begin with the string 'CCA'**
  - **The total payload mass carried by boosters launched by NASA (CRS)**
  - **Average payload mass carried by booster version F9 v1.1**
  - **Date when the first successful landing outcome in ground pad was achieved.**
  - **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**
  - **Names of the booster versions which have carried the maximum payload mass.**
  - **Failed landing outcomes in drone ship, their booster versions and launch sites in the year 2015**
  - **Ranked different successful landing outcomes and their count between the 04-06-2010 and 20-03-2017 in descending order**

Source code - EDA with SQL

Build an Interactive Map with Folium

- Markers, Circles, Lines and Marker clusters were used with Folium
  - Markers were used to mark launch instances
  - Circles were used to highlight areas around specific coordinates, for example – NASA Johnson space center
  - Marker clusters were used because same launch sites were used for various launch instances
  - Lines were used to show distance between two coordinates

- Source code - Interactive map with folium

Build a Dashboard with Plotly Dash

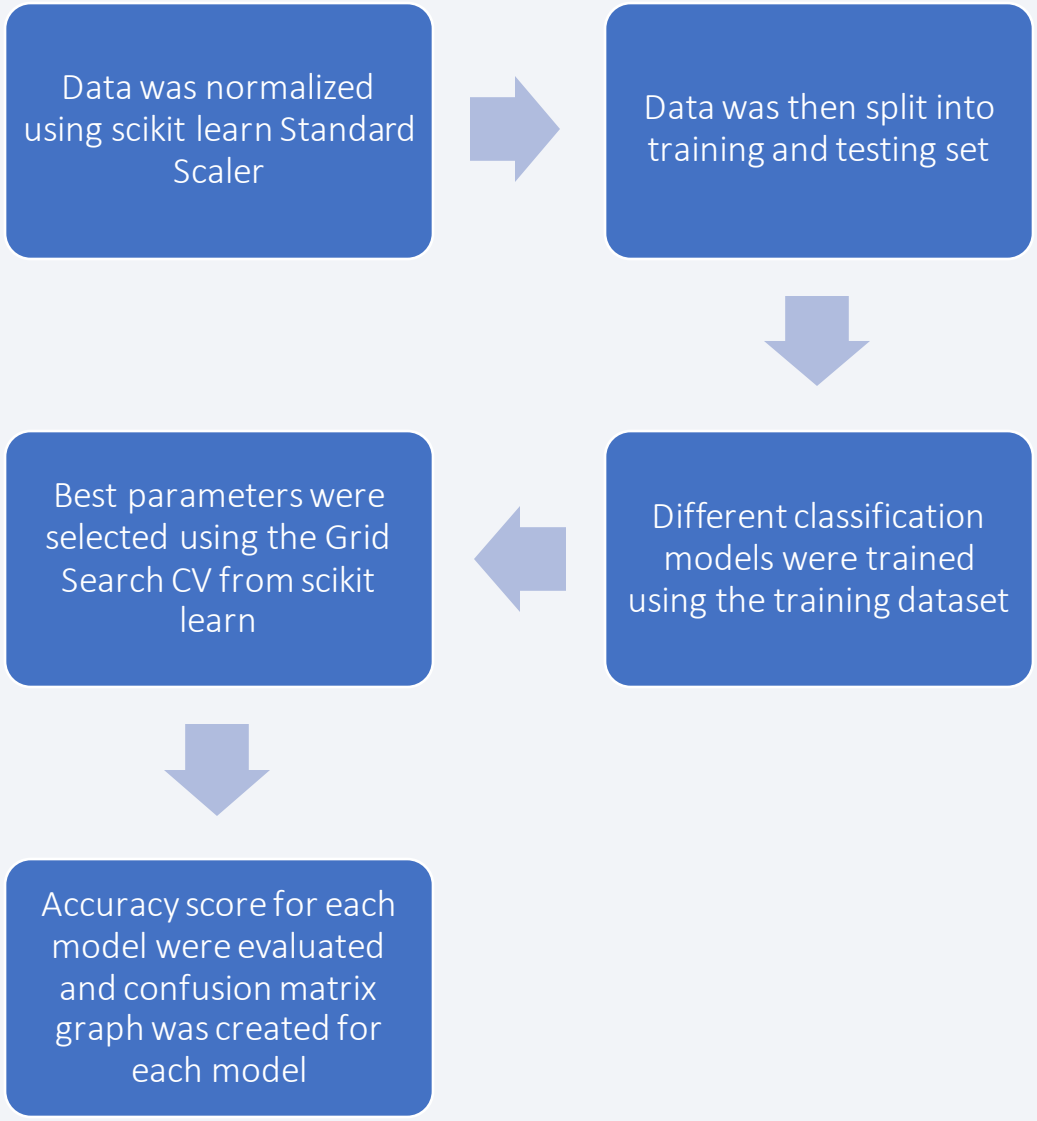
- Interactive pie chart and scatter plot were built in the dashboard.
- Pie chart
  - To show the total successful launches count for all sites
  - If a specific launch site was selected, shows the Success vs. Failed counts for the site
- Scatter plot
  - To show the correlation between payload mass and launch success

- With the help of these two graphs we can easily determine the payload mass having good success chances at a specific launch site

- Source code - Dashboard app

Predictive Analysis (Classification)

- After the data is prepared and ready, different machine leaning classification models were built and trained upon the data and the results were compared.



- Source code - Predictive analysis

Results

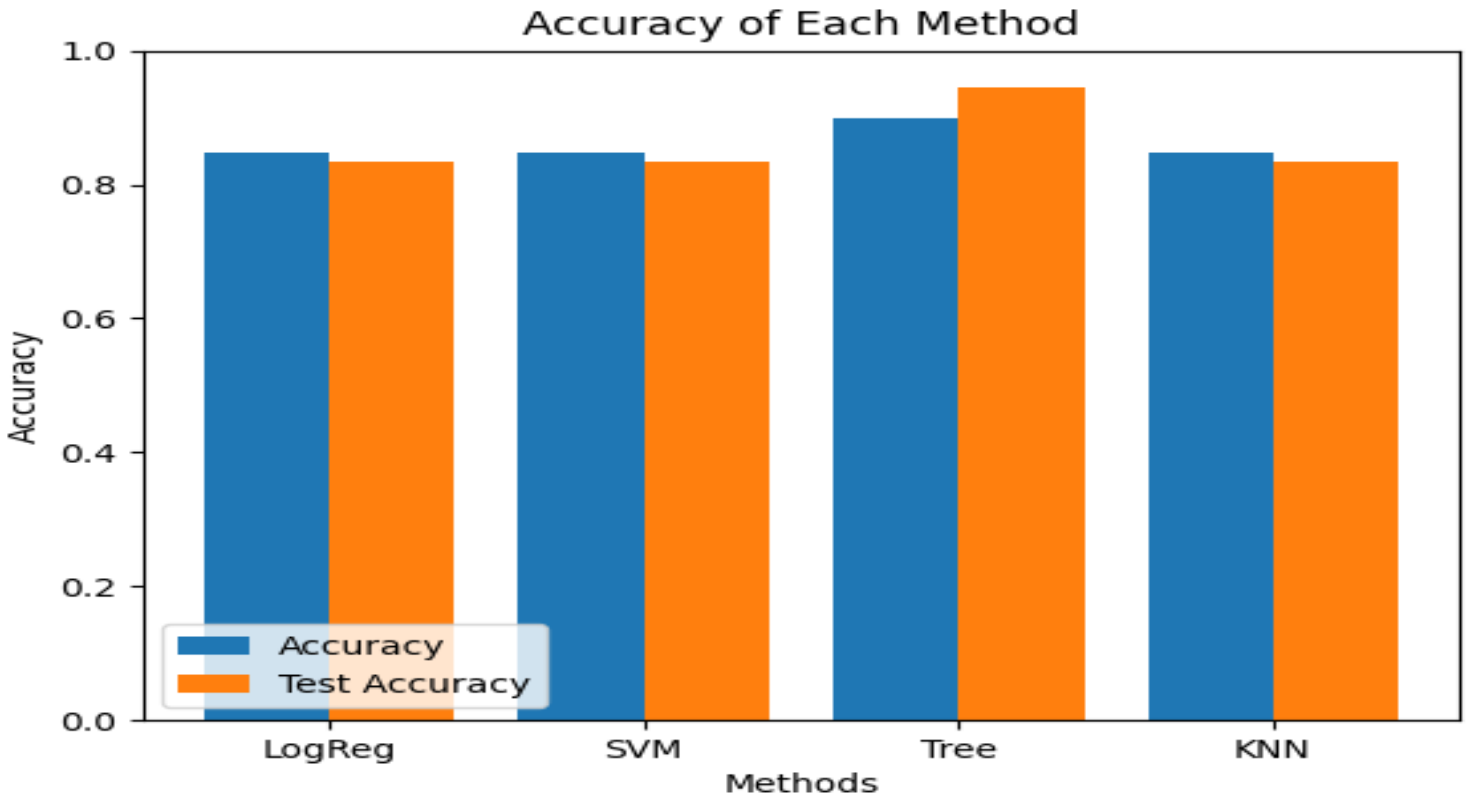
- Exploratory data analysis results
  - Four different launch sites are being used
  - Average payload mass carried by booster version F9 v1.1 is 2928.4 kg
  - The first successful landing at ground pad was in the year 2015
  - Over the years the success rate is growing rapidly
  - Many orbit types such as ES-L1, GEO have 100% success rate
  - Higher payload mass have relatively high success rate
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

- Visual Analytics
  - Most of the launches happened from the sites in the eastern coast
  - All the launch sites are situated near the sea and unpopulated areas

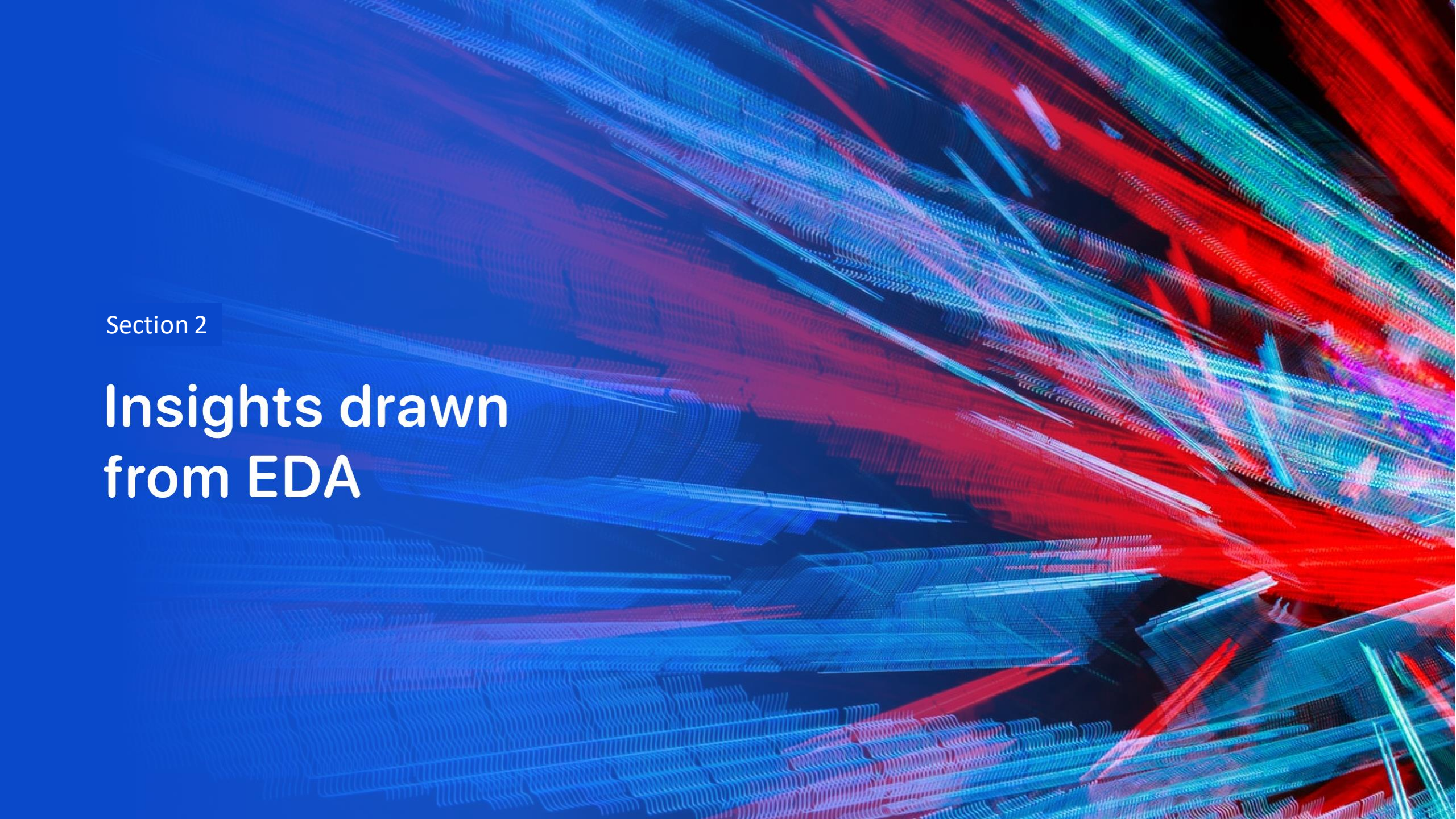
Results

- Predictive Analysis
  - From checking the accuracies of different models, we can observe that the decision tree model is giving the best result

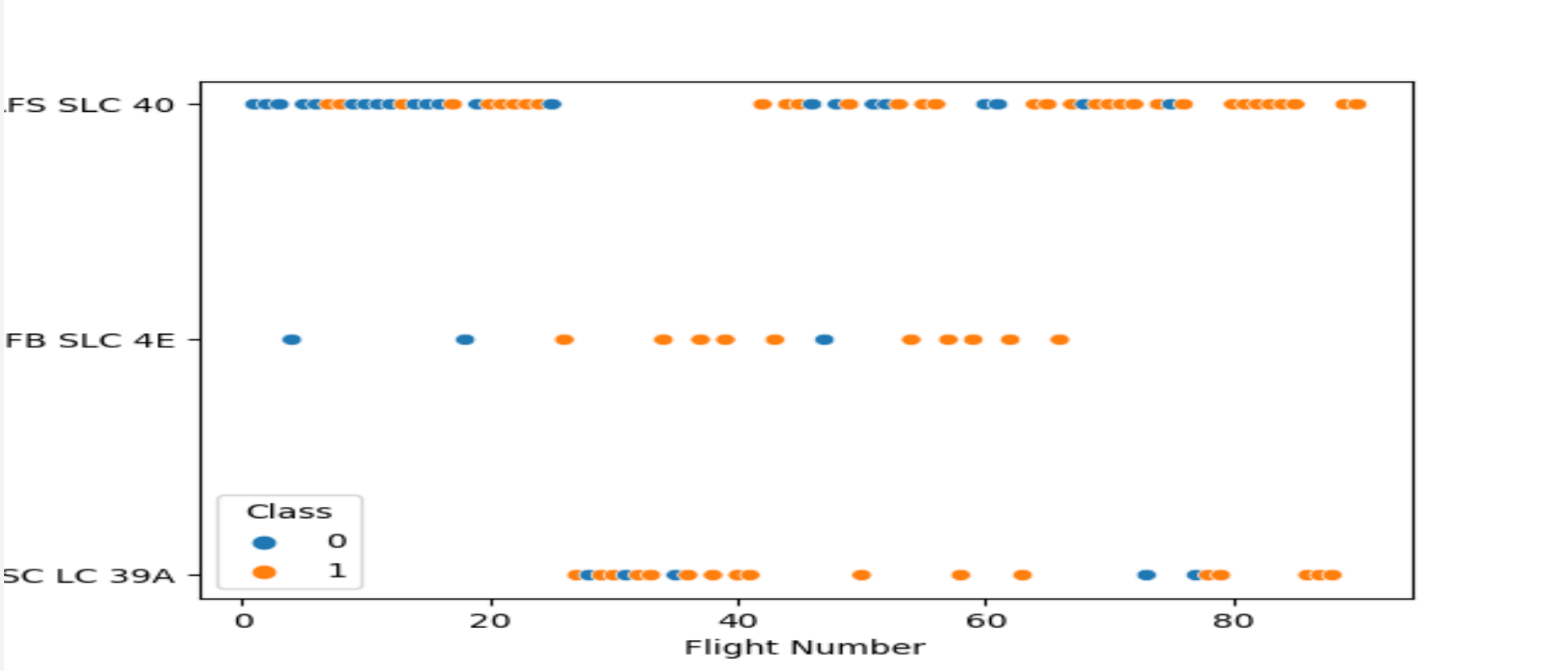


Section 2

Insights drawn from EDA

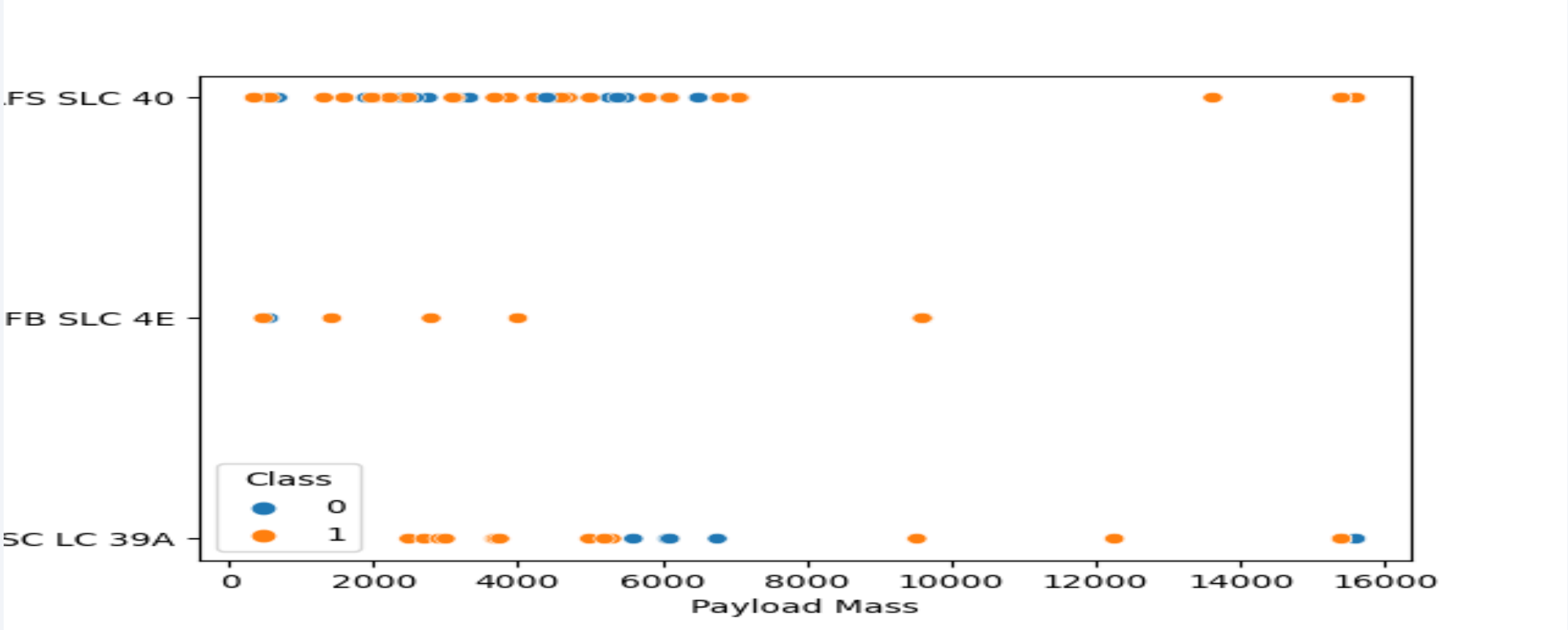


Flight Number vs. Launch Site



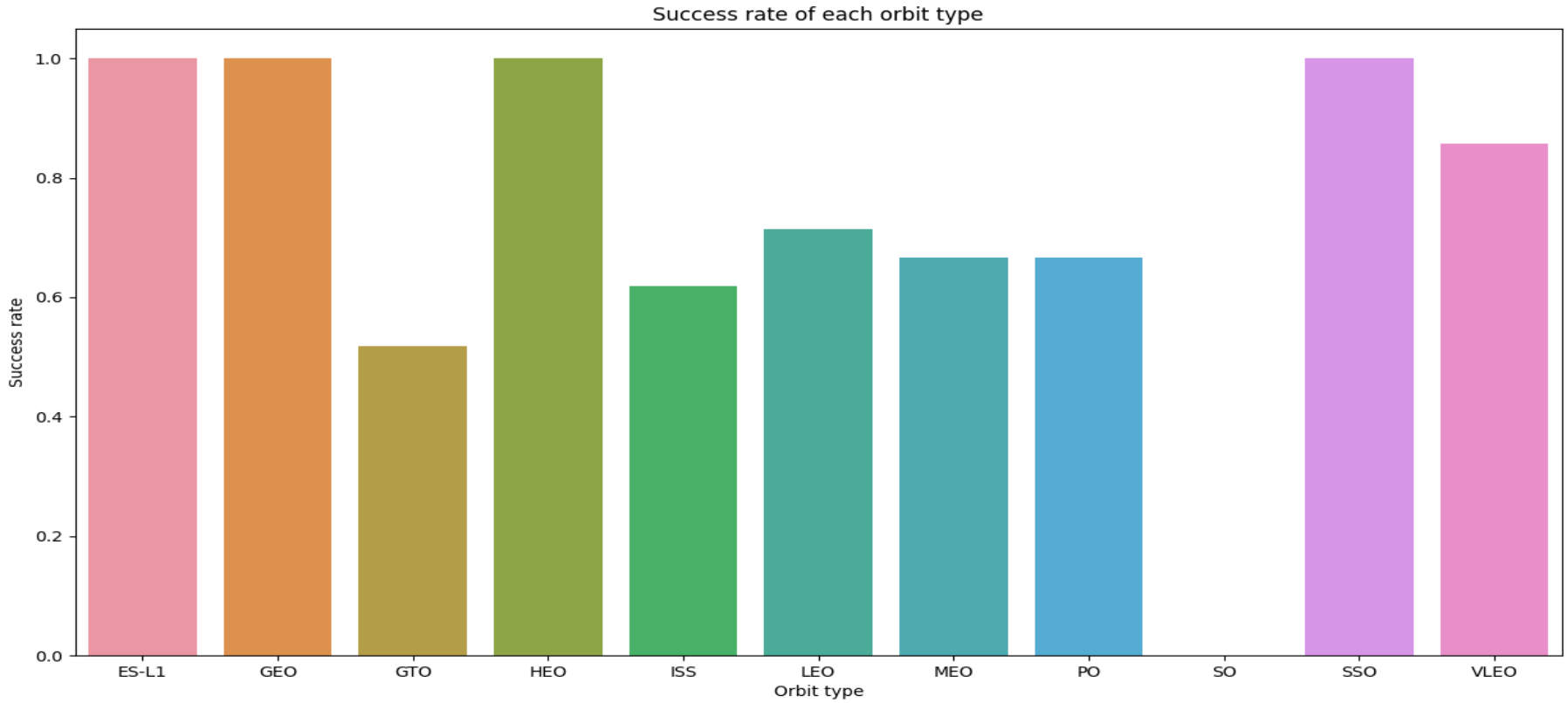
- We can observe that FS SLC 40 is the most used launch site and it has a high success rate

Payload vs. Launch Site



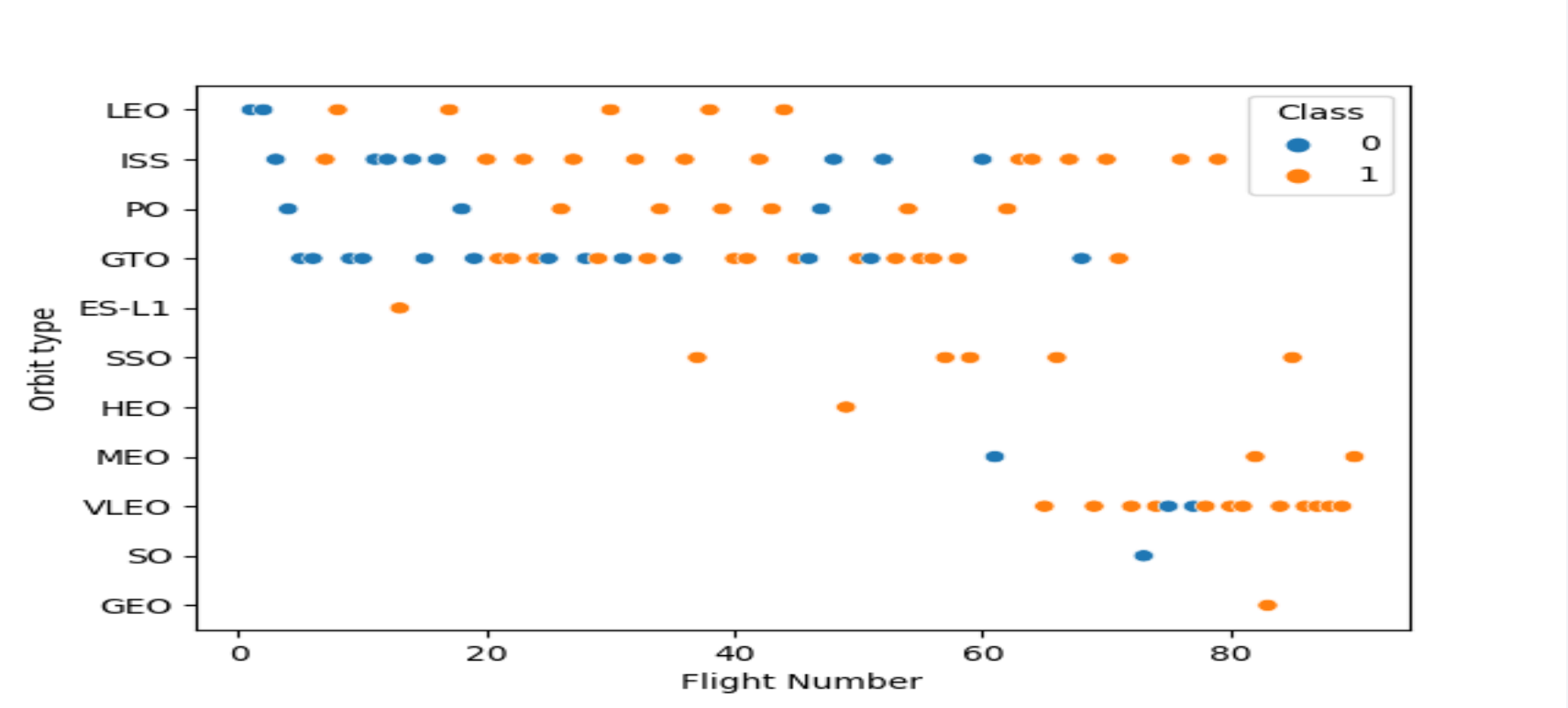
- Higher payload mass have high success rate
- Launch Site FB SLC 4E have very less launches compared to the others

Success Rate vs. Orbit Type



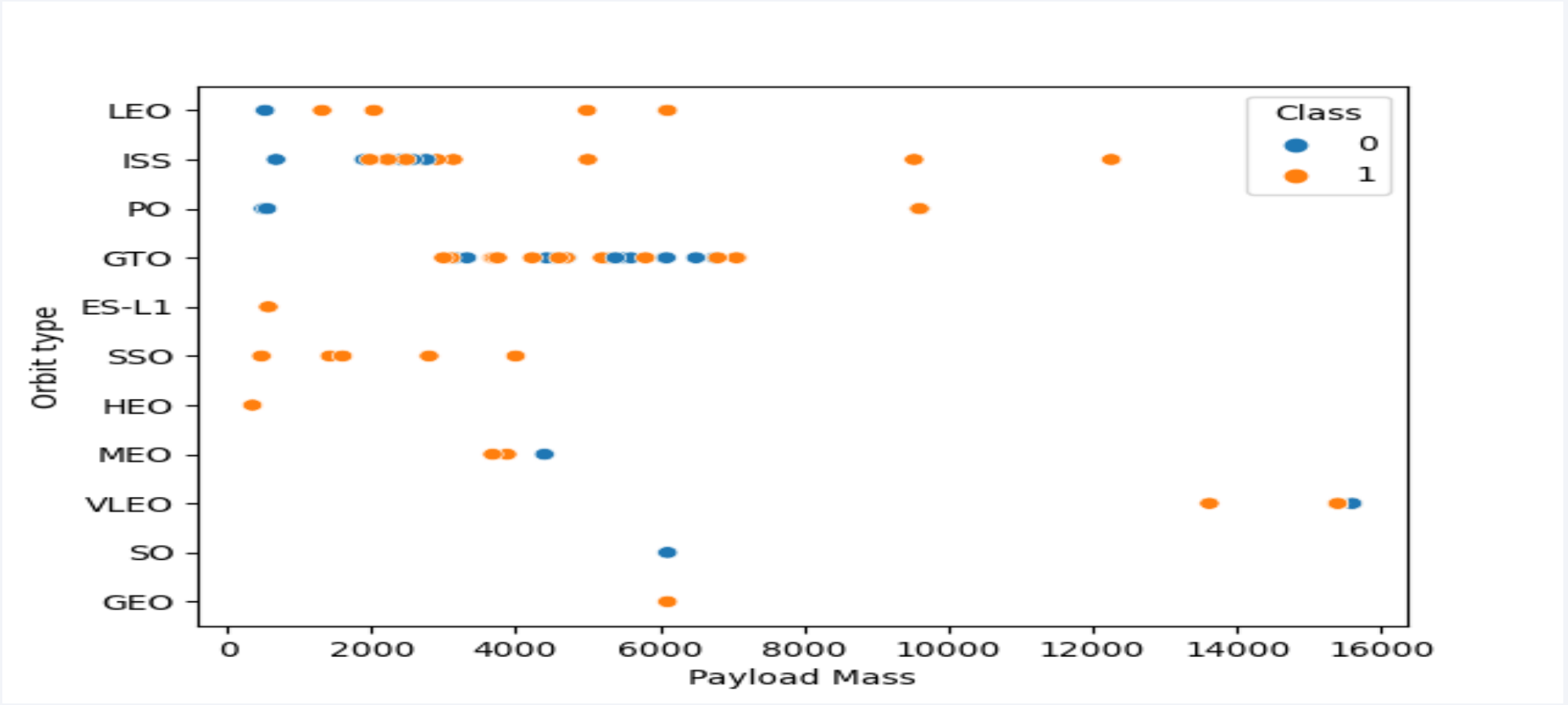
- Many orbit types have 100% success rate

Flight Number vs. Orbit Type



- Here we can observe that orbit types having 100% success rate in the previous slides have only a few launches
- Over time success rate for each orbit has increased
- Launches to VLEO Orbit have increased in recent launches

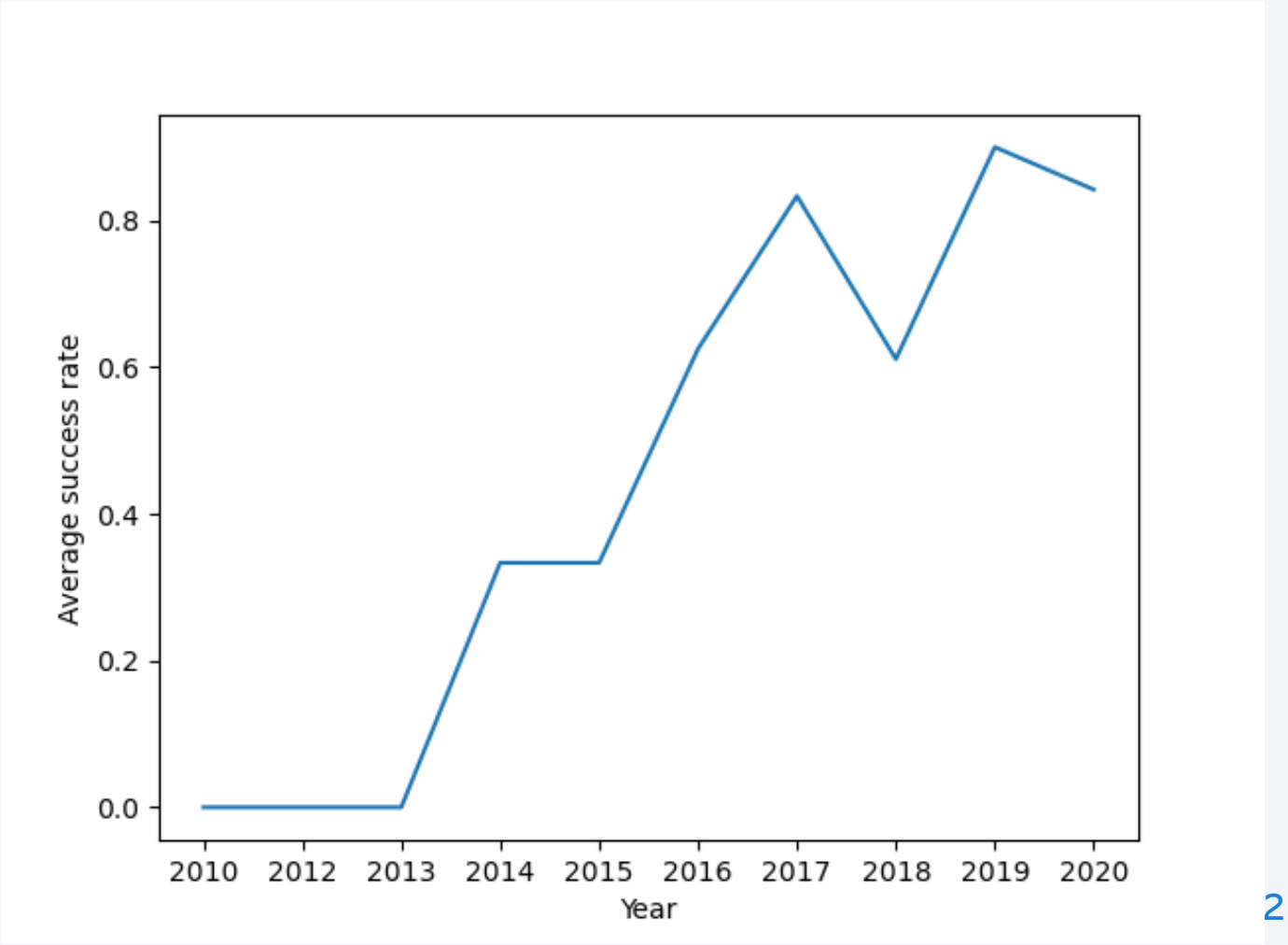
Payload vs. Orbit Type



- Payload Mass greater than 2500 have 100% success rate in LEO and ISS orbits
- SSO orbit have 100% success rate irrespective of payload mass
- Payload mass doesn't have much impact on success rate for GTO orbits

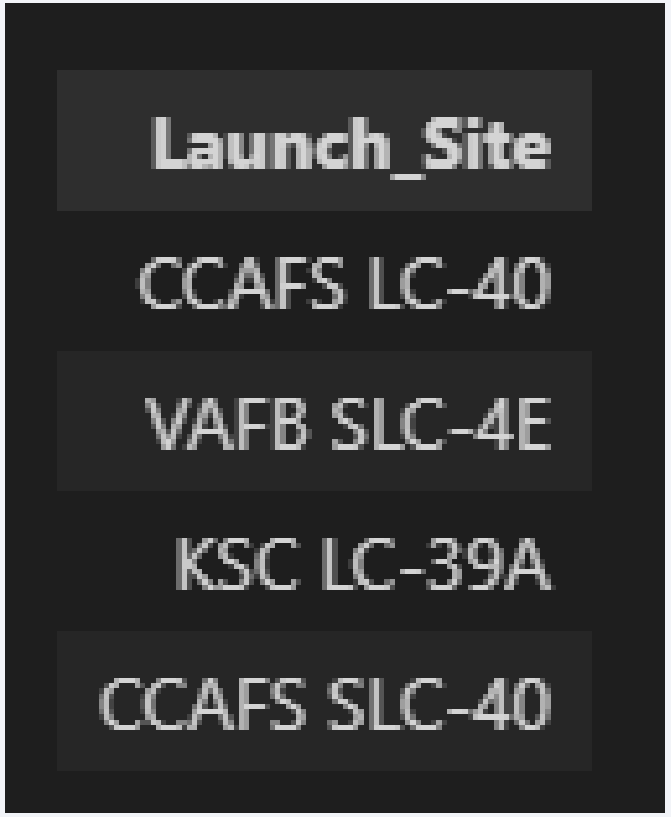
Launch Success Yearly Trend

- We can see a rapid growth in success rates over the years starting from 2013.
- Success rate dropped in the year 2018 but is growing again since then



All Launch Site Names

- There are the different Launch sites that SPACE X have used for their launches.
- These are the 4 different sites used for the launches.



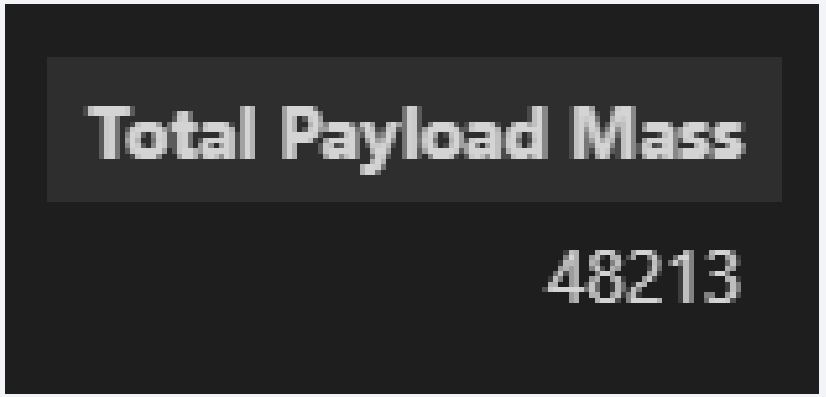
Launch Site Names Begin with 'CCA'

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

- These are first five results for the Launch site CCAFS LC-40
- We can observe that all launches have succeeded which shows that CCAFS LC-40 launch site have high success rate

Total Payload Mass

- Total payload mass carried by launchers from NASA (CRS) is approx. 48000 kgs
- This shows that there are many instances of launches done by NASA



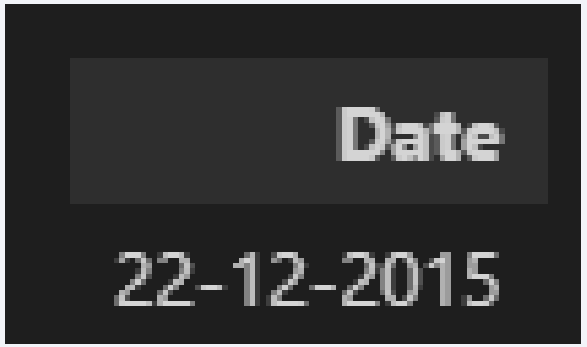
Average Payload Mass by F9 v1.1

- Average payload mass carried by F9 v1.1 is approx. 2900
- This shows that too much heavy payload weren't being used even though they have relatively higher success rate.
- We can conclude that higher payload mass carried by F9 v1.1 is either not possible or is costly

First Successful Ground Landing Date

- By sorting the data for successful ground landings date-wise, we found that the first ever successful ground pad landing was done on 22 December, 2015

- In the previous slides we have seen that success rate rapidly increased after 2013, still the first successful landing for ground pad was in the end of 2015



Successful Drone Ship Landing with Payload between 4000 and 6000

- There are four Falcon 9 boosters which have successful drone ship landing with a high payload mass between 4000kg-6000kg

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

Total Number of Successful and Failure Mission Outcomes

- On the data we observed using SQL we have 100 instances where the mission was success (for one instance payload status is unclear)
- We got this outcome by querying data grouped by mission outcome

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

Boosters Carried Maximum Payload

- In our data these booster versions carried the maximum weight for a single launch (15600 kg)

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

- We queried our data to observe the drone ship landing failure instances in the year 2015, booster versions used, launch sites for the launches and the month in which launches were done
- These are the results which shows 2 failure outcome in the month January and April respectively
- Both the launches were done from CCAFS LC-40 launch site

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

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

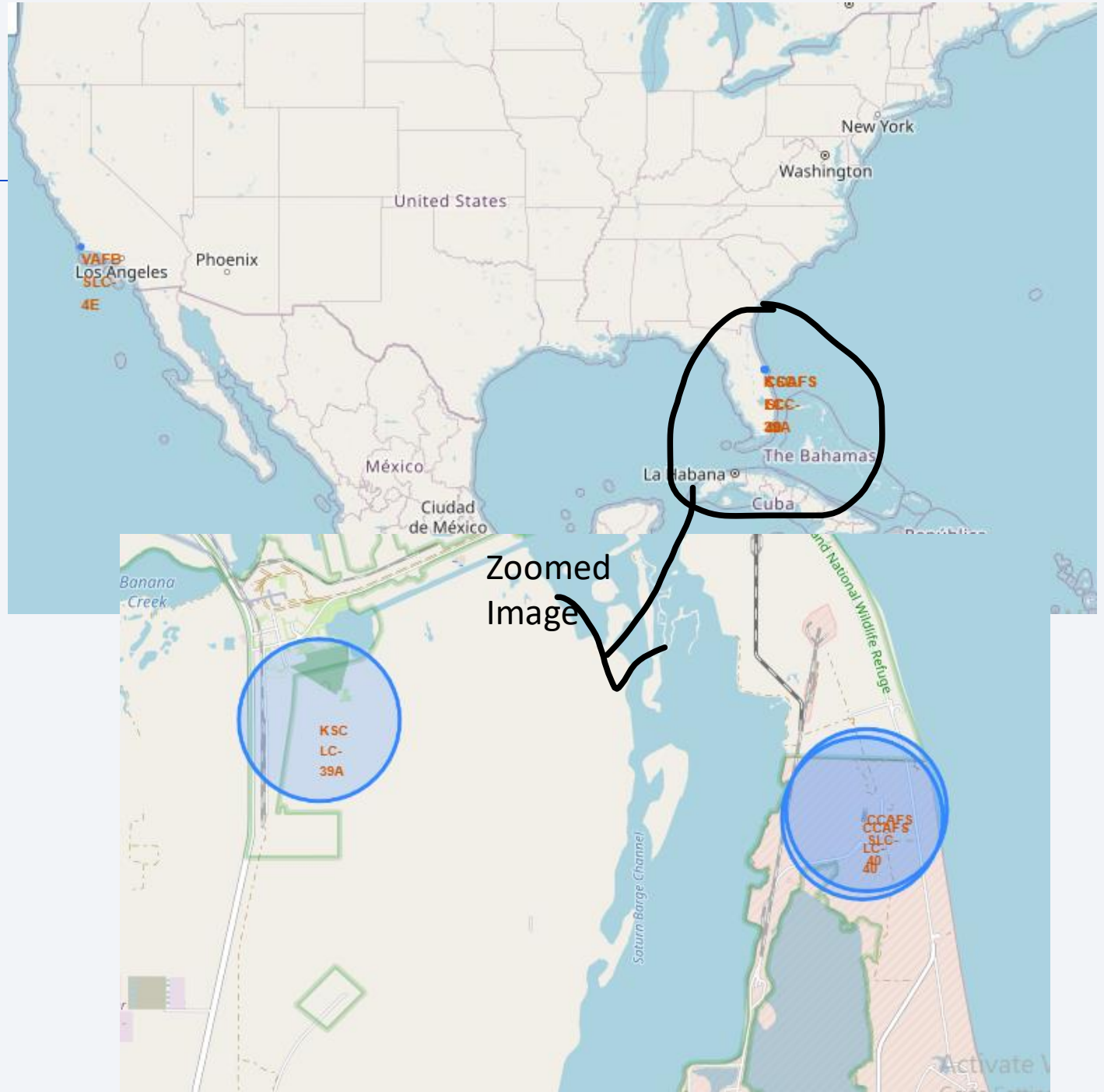
- We queried our data to observe successful landing outcomes between 04 June 2010 and 20 March 2017 and these are the results.

Landing _Outcome	COUNT(*)
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

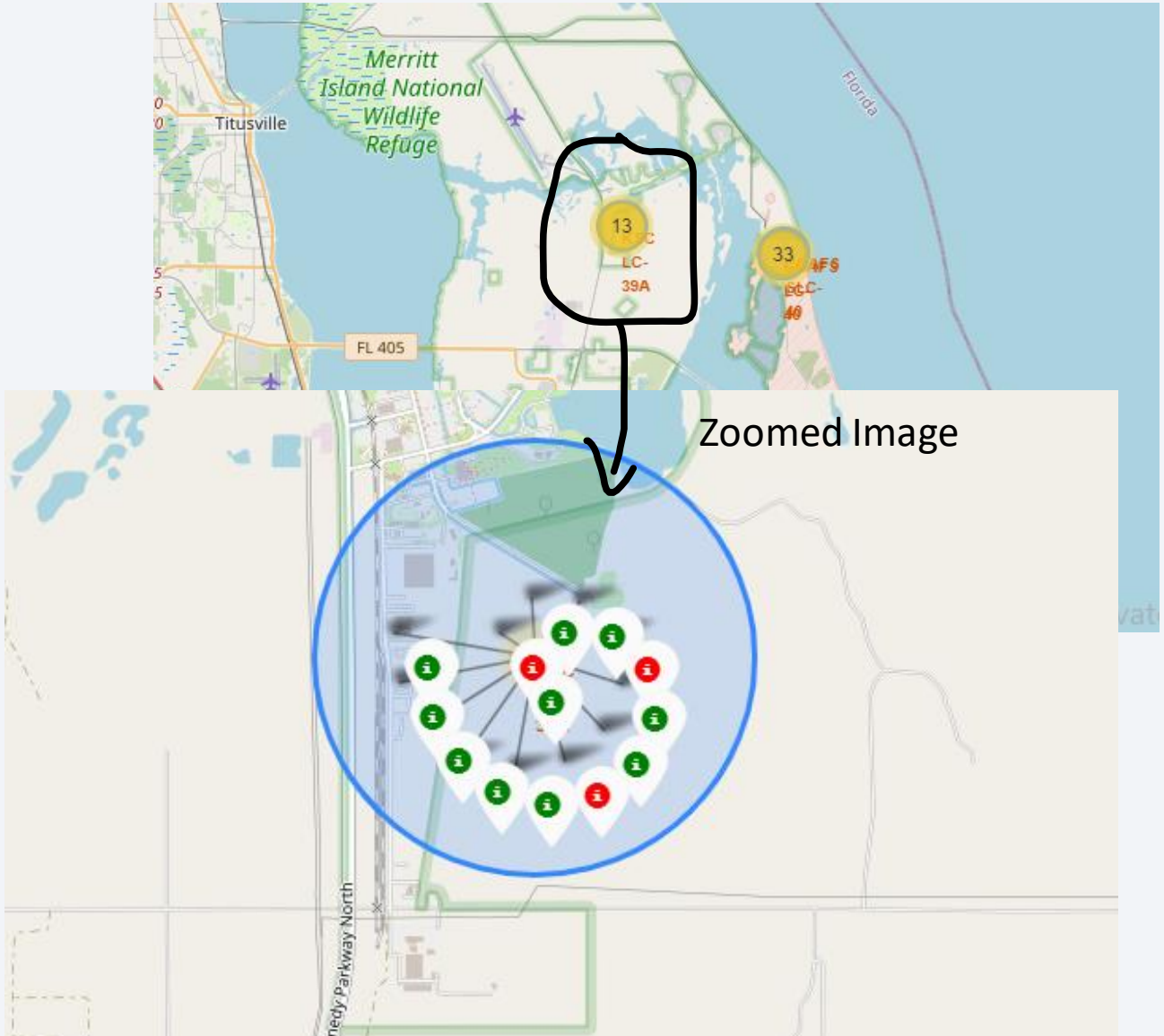
Launch Sites

- We can see all the four launch sites are close to water
- Three of the launch sites are close to each other
- Most of the launches in our observation was done from the eastern sites

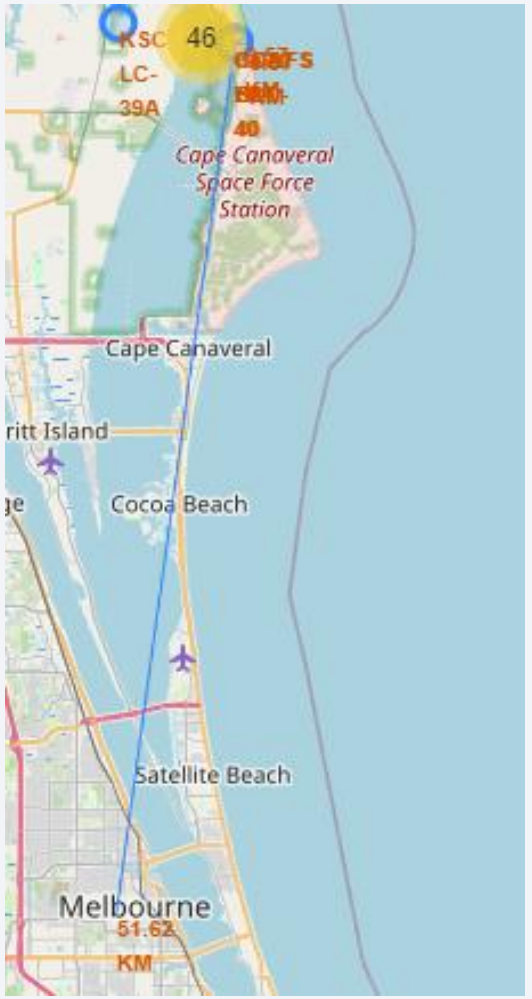


Launch Instances

- Many of the launches was done from the same site. To avoid confusion and overlapping in the map we used marker cluster.
- When we zoom into the cluster at the launch site KSC LC-39A we can observe various launch instances
- Here green means successful launch and red means failed outcome.
- We can observe here that the success rate is roughly more than 80%

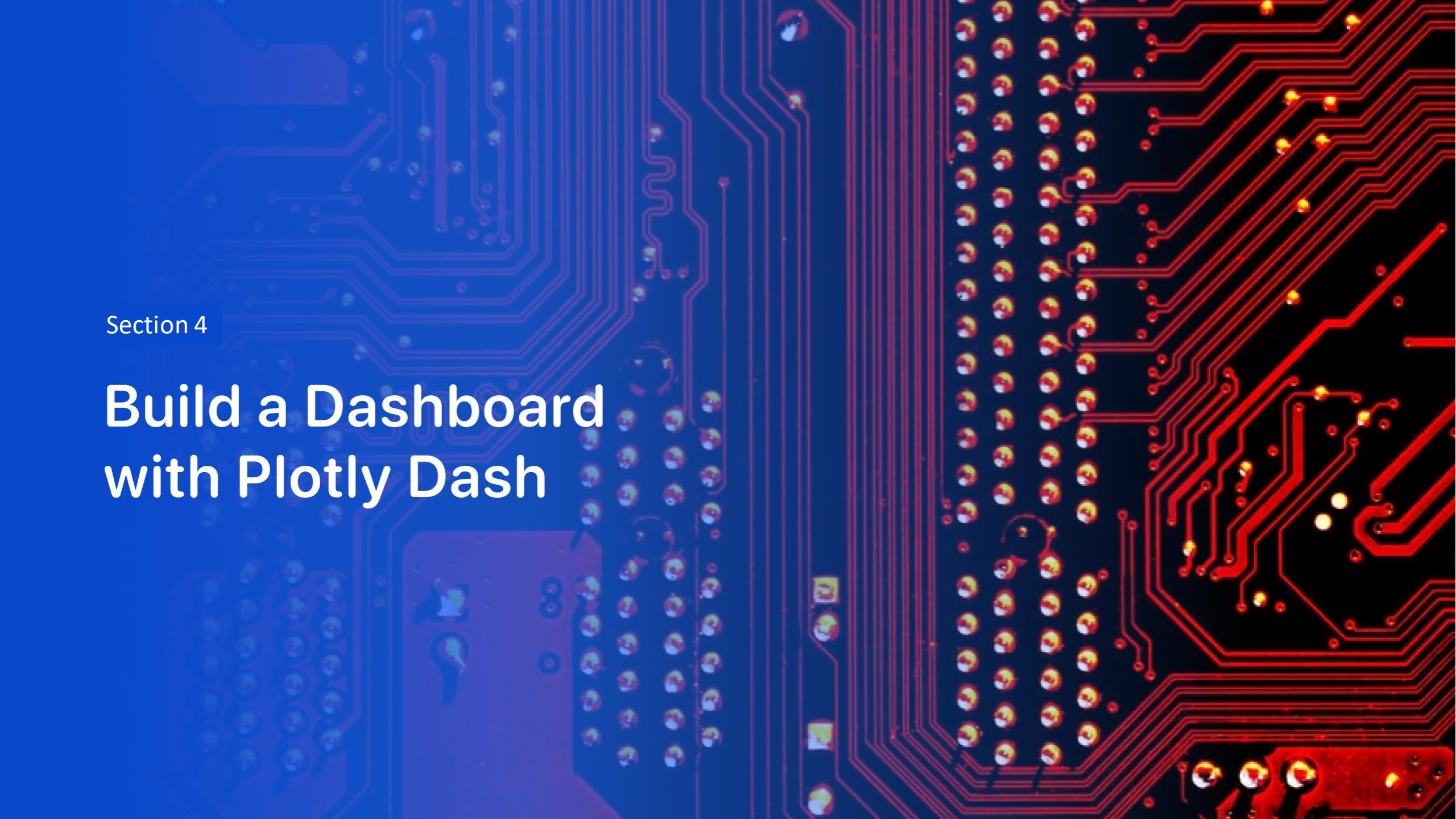


Ideal Launch Site location



Ideal Launch Site location

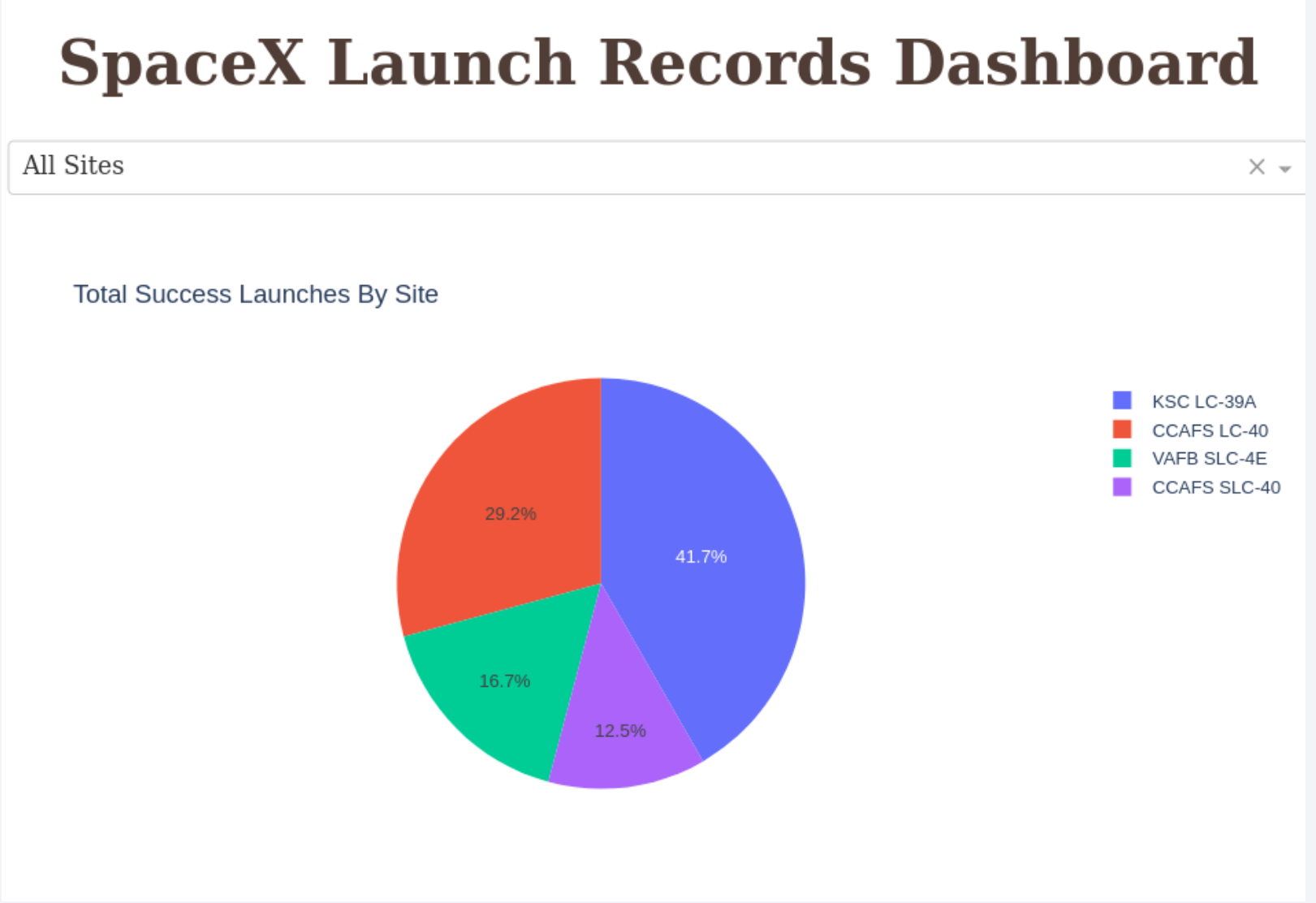
- In the first picture we can observe the location of launch site from three important logistical locations
  - Sea coast – We can observe sea coast is a distance of 0.87 km from the site
  - Train track – Train track is very close to the site at a distance of just 0.03 km (Note – It may be not visible in the attached screenshot in the last slide because of overlapping notations)
  - Highway – We can observe highway is also close to the launch site at a distance of 0.57 km
- In the second picture we can observe that the location of the launch site is far from populated cities like Melbourne in the picture which is at a distance of approx. 52 km
- So we can conclude that a good launch site has proper transportations close by, are near the coast and are far from populated locations.


Section 4
Build a Dashboard with Plotly Dash

Total successful launches

- The following graph represents distribution of successful launches distribution among different sites

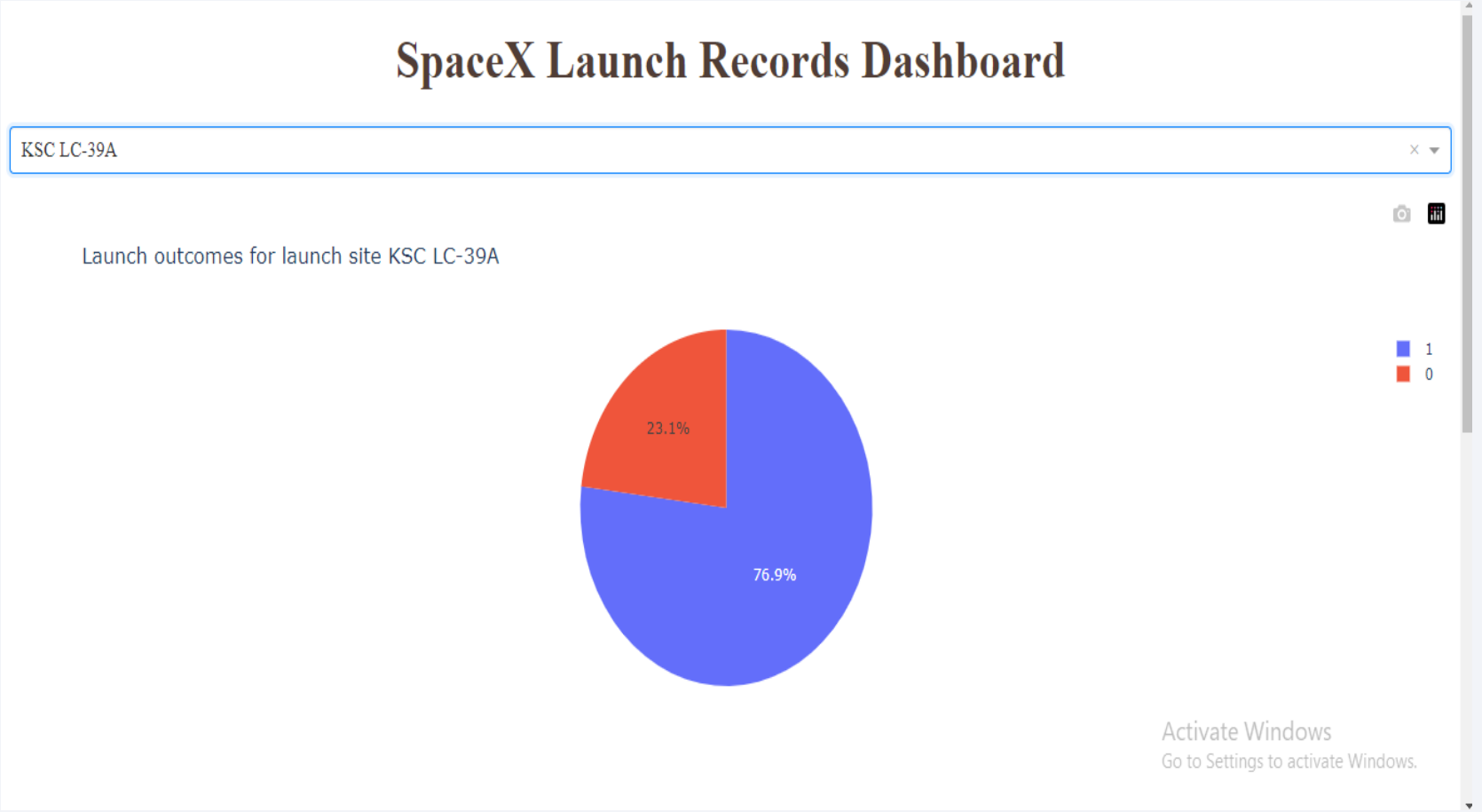
- We can observe that KSC LC-39A have highest success percentage of 41.7% followed by CCAFS LC-40 at 29.2% of total successful launches
- Note – This data represents each launch sites share in total successful launches and not the success vs failure percentage



KSC LC-39A success rate

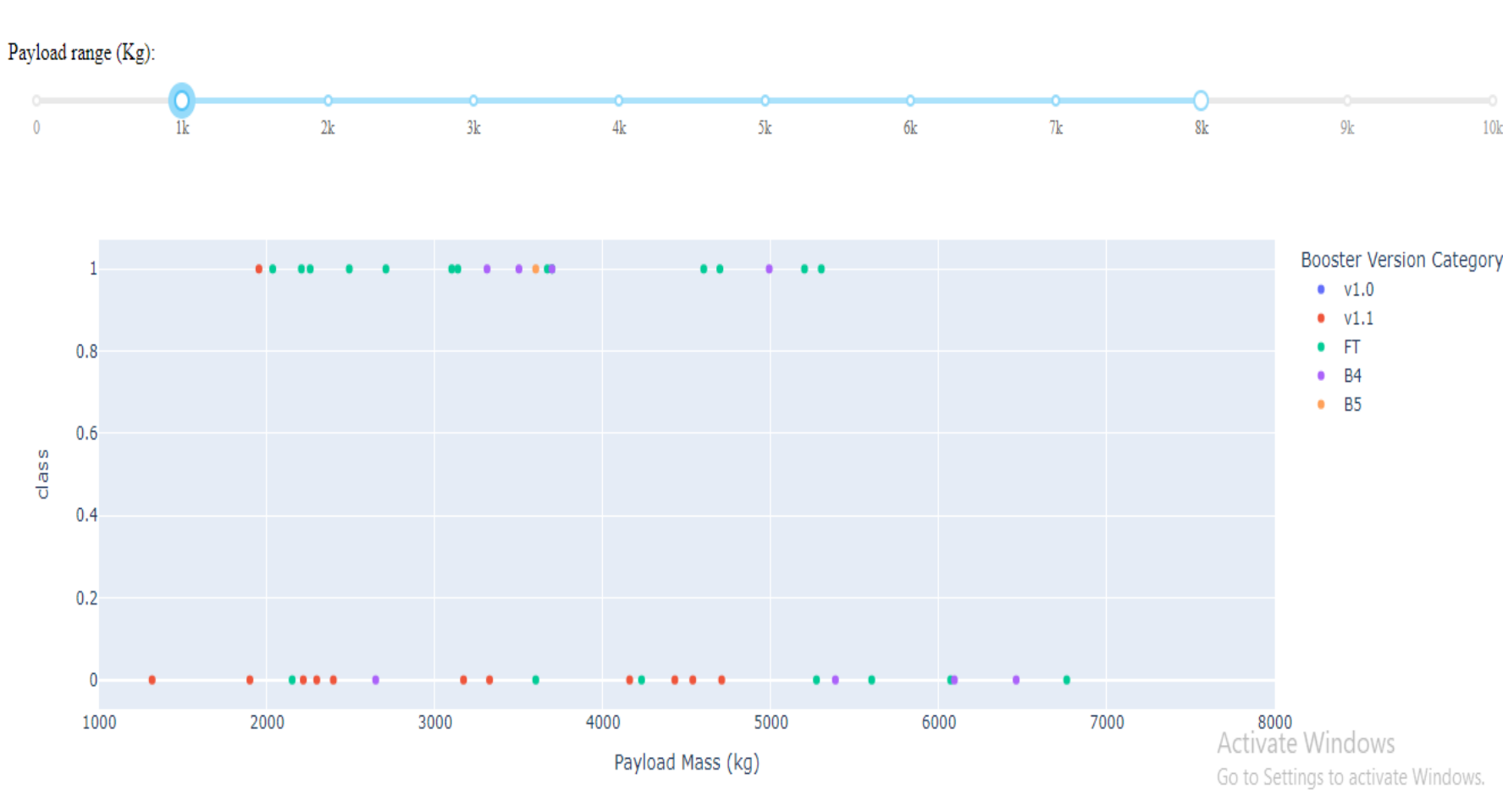
The following graph shows the success vs failure launches from KSC LC-39A launch site

We can observe a high success rate of approx. 77%



Success rate for different payload

- The following graph shows success rate for different range of payload mass.
- Different colors dots represents different booster version category
- We can observe a high success rate for the booster category FT compared to the others

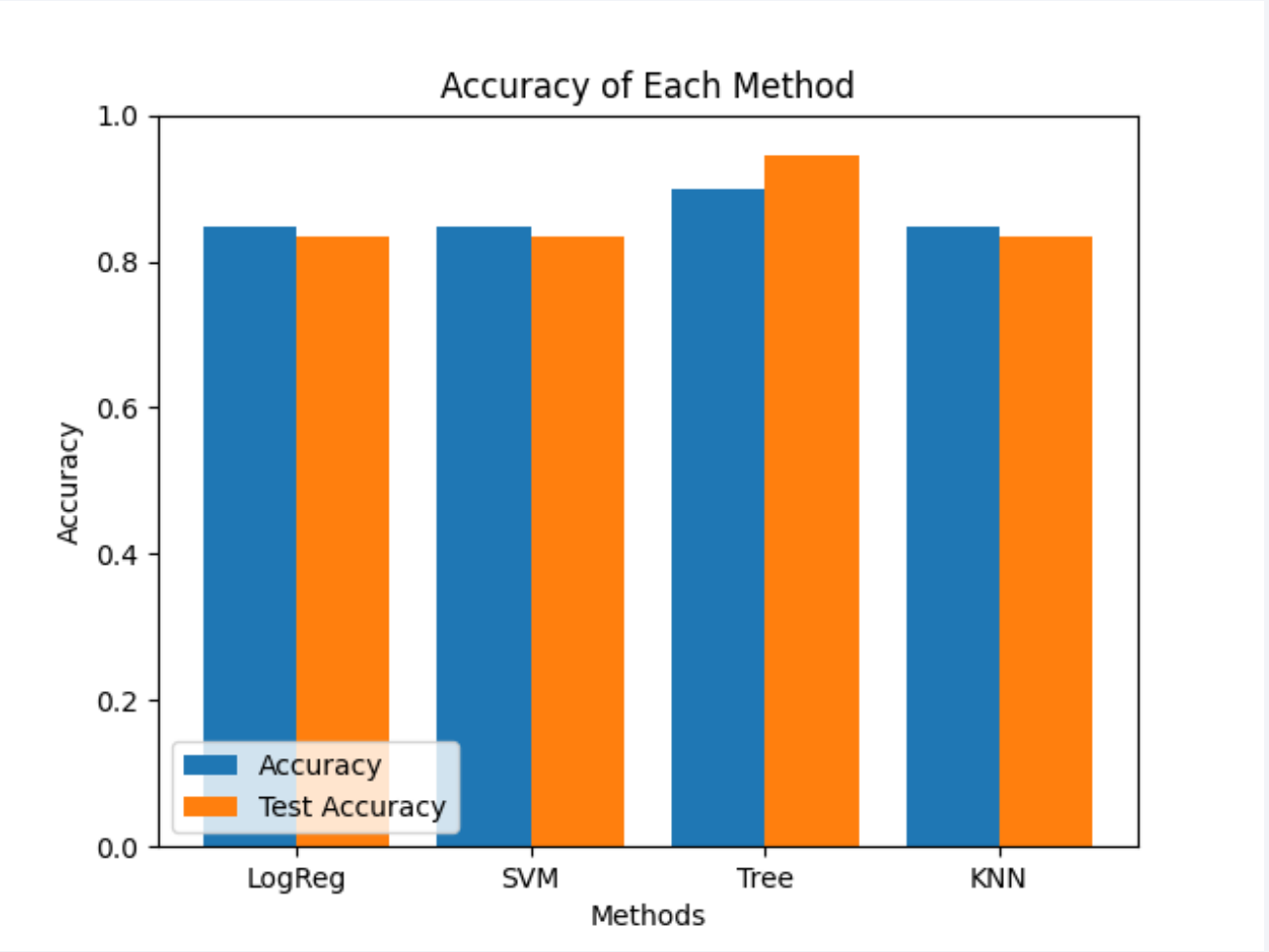


Section 5

Predictive Analysis (Classification)

Classification Accuracy

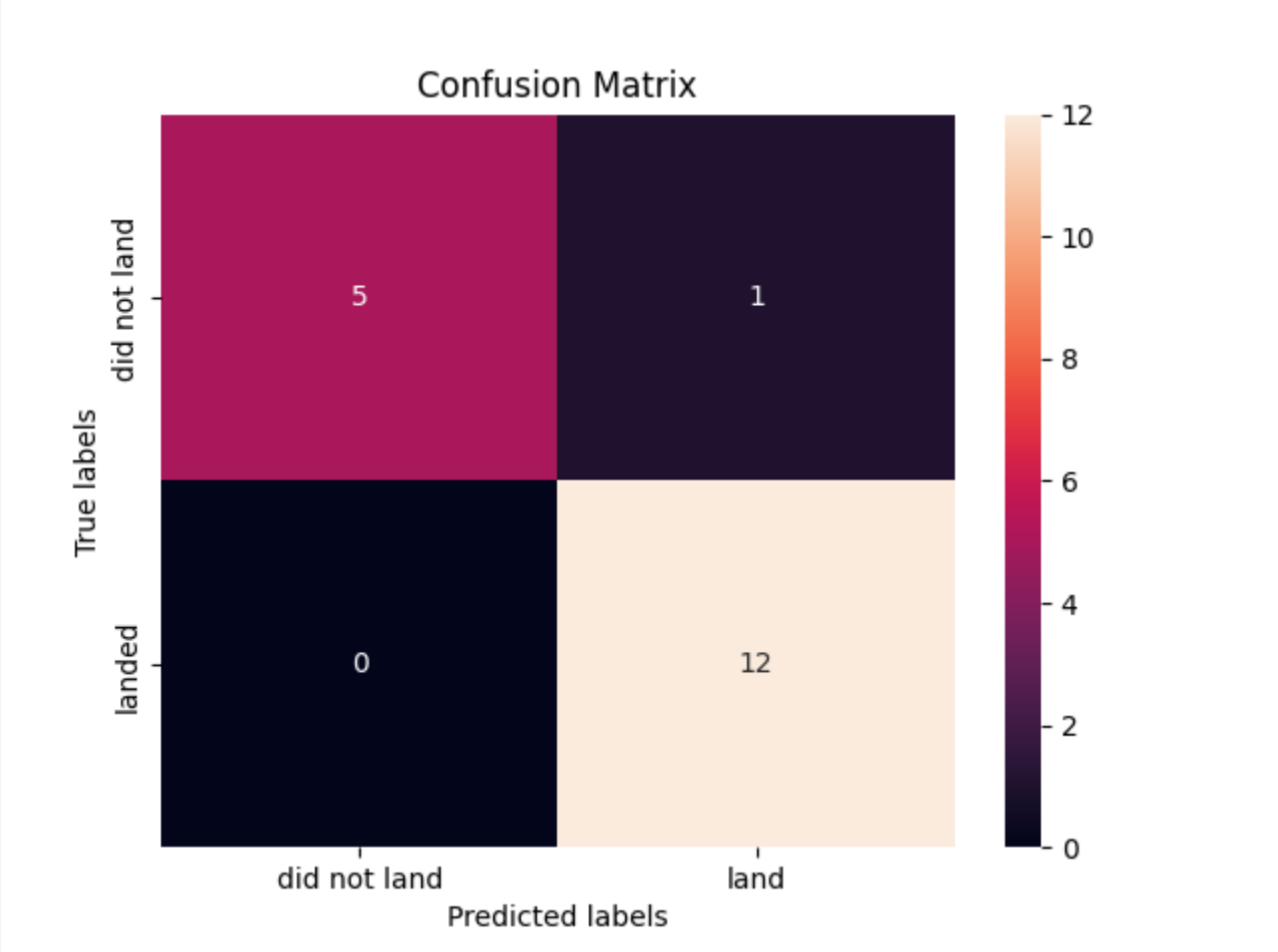
- We can clearly see that the decision tree classifier performed the best on both the training and testing data.



Confusion Matrix

From the following confusion matrix plot for decision tree classifier model we can observe two key points

- Big numbers of true positive and true negative
- There was only one instance of wrong prediction, a false positive



Conclusions

- The best launch site will be KSC LC 39A followed by CCAFS LC-40
- Launch sites are selected based upon their geographical location and their distance with developed logistic infrastructure at a distance from populated areas
- Most of the mission outcomes were successful however successful landing outcomes are improving rapidly over time
- Decision tree classifier gives the best result on our data and can be used to predict new unseen data

Thank you!