
# Winning Space Race with Data Science

Alberto Martin Visciglia
18/03/22

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)

- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

# Introduction

- Project background and context

We will predict if the Falcon 9 first stage will land successfully. 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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

  - What determine if the rocket will land successfully?
  - The effect each variable impact in determining the success rate of a successful landing.
  - What conditions does SpaceX have to achieve to get the best success landing rate.

Section 1

# Methodology

# Methodology

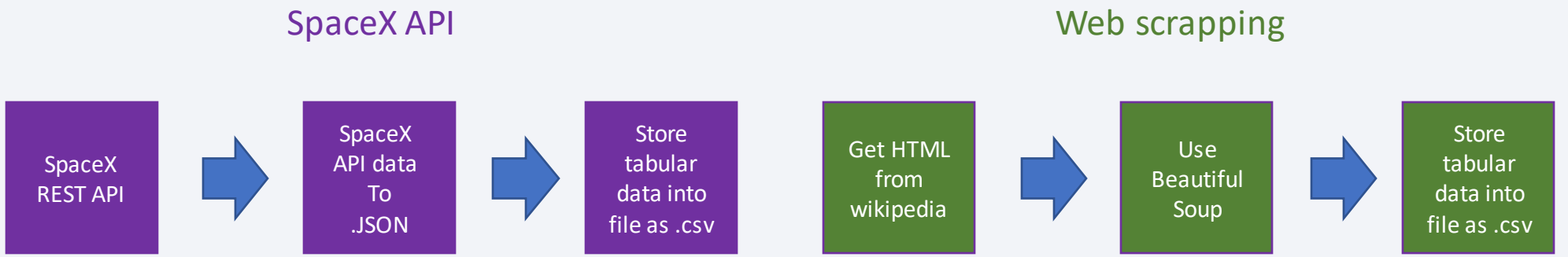
## Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - (Web Scrapping) from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- 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

- Describe how data sets were collected.

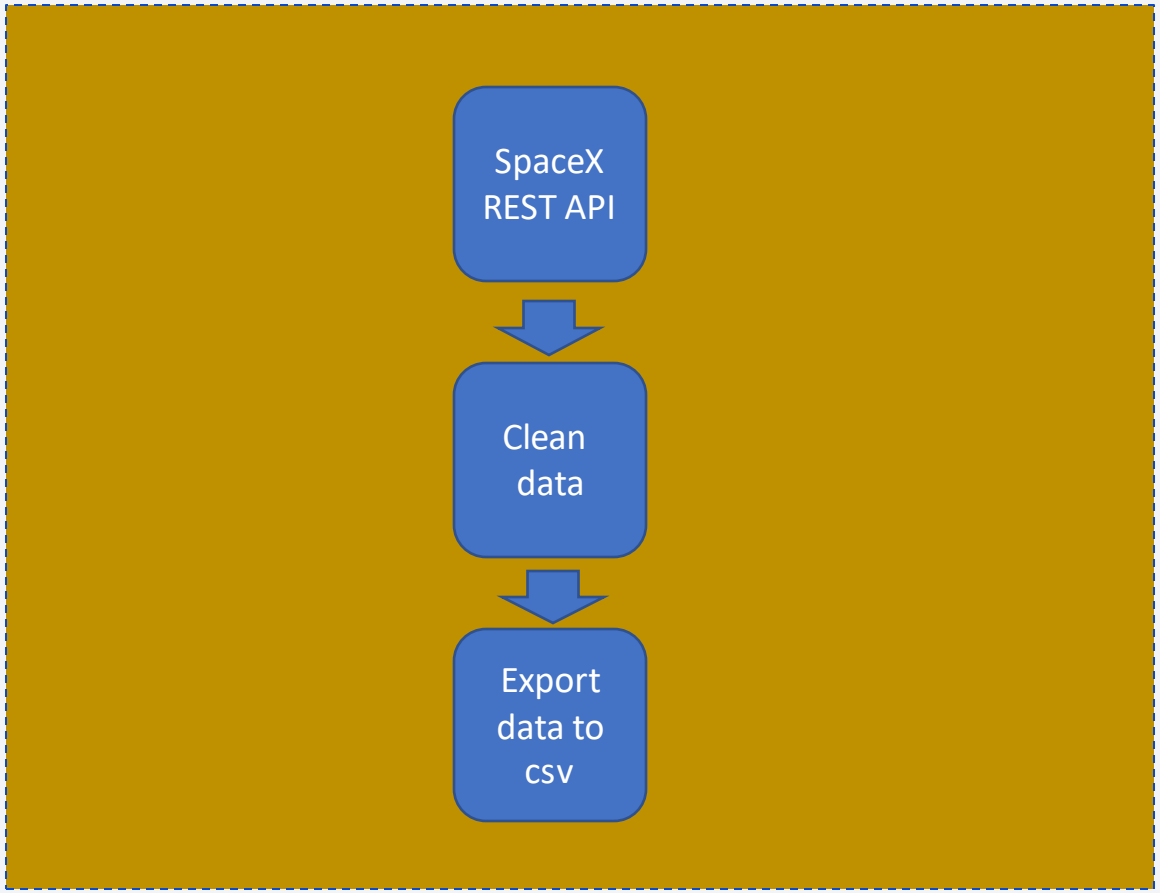
  - We worked with SpaceX launch data that is gathered from the SpaceX REST API.
  - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.



# Data Collection – SpaceX API

- Response from API
- Transform Response to a .JSON
- Clean data
- Export data to a .csv

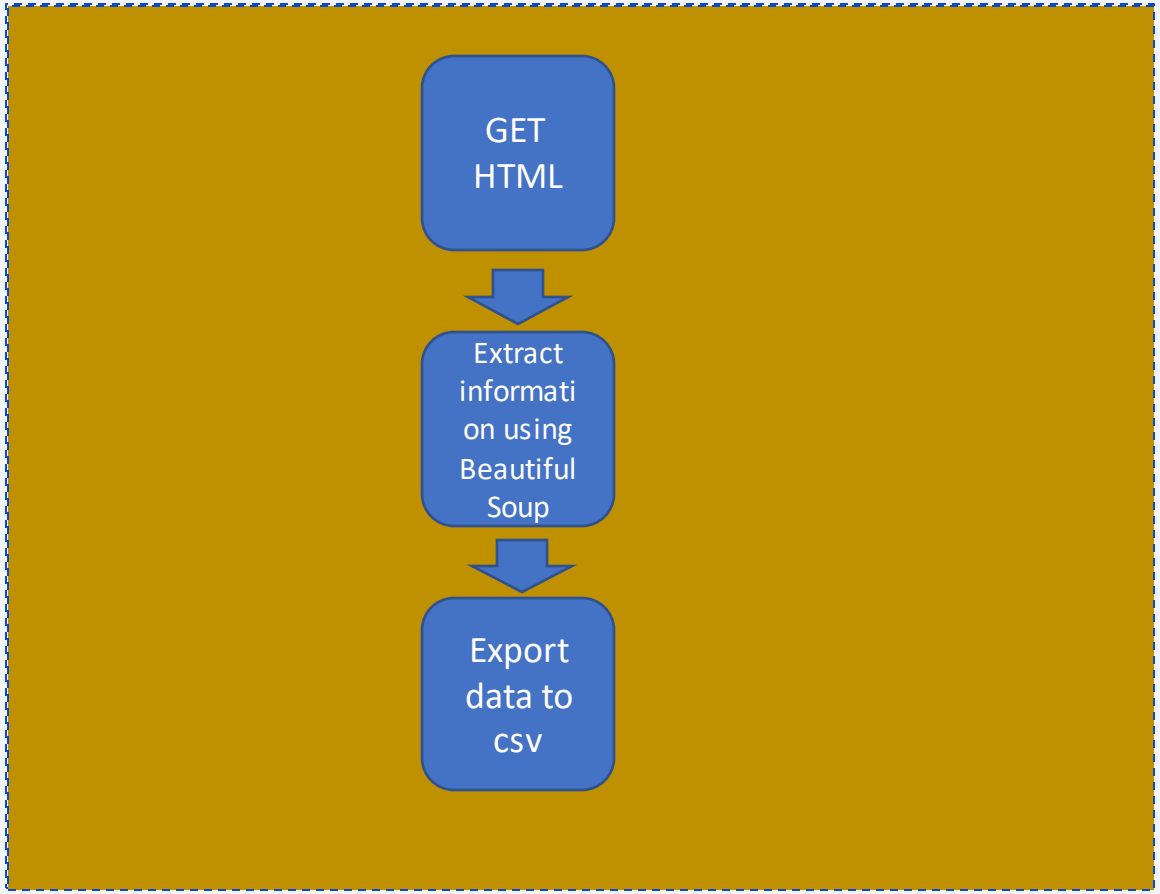
- GitHub url to Notebook



# Data Collection - Scraping

- Get HTML
- Use Beautiful Soup to obtain tables and column names to extract the required data
- Export data to a .csv

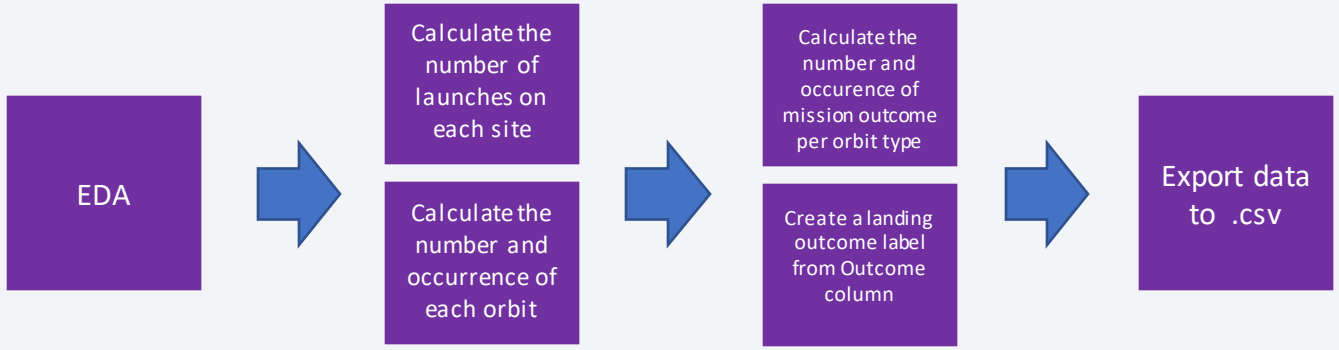
- GitHub url to Notebook



# Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, **`True`** Ocean means the mission outcome was successfully landed to a specific region of the ocean while **`False`** Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. **`True`** RTLS means the mission outcome was successfully landed to a ground pad **`False`** RTLS means the mission outcome was unsuccessfully landed to a ground pad. **`True`** ASDS means the mission outcome was successfully landed on a drone ship **`False`** ASDS means the mission outcome was unsuccessfully landed on a drone ship.

- GitHub url to Notebook



# EDA with Data Visualization

**Scatter Graphs being drawn:**

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass

**Bar Graph being drawn:**

- Mean VS. Orbit

**Line Graph being drawn:**

- Success Rate VS. Year

- GitHub url to Notebook

# EDA with SQL

**Performed SQL queries to gather information about the dataset**

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

- GitHub url to Notebook

# Build an Interactive Map with Folium

- We added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- We calculated the distance from the Launch Site to various landmarks to measure patterns. Lines are drawn on the map to measure distance to landmarks.

- GitHub url to Notebook

# Build a Dashboard with Plotly Dash

- The dashboard is built with Dash web framework.
- Graphs
  - Pie Chart showing the total launches by a certain site/all sites
  - relative proportions of multiple classes of data.
  - size of the circle can be made proportional to the total quantity it represents.
- Scatter Graph of Outcome vs Payload Mass (Kg) for different Booster Versions
  - relationship between two variables.
  - range of data flow, i.e. maximum and minimum value.

# Predictive Analysis (Classification)

**BUILDING MODEL**
- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

**EVALUATING MODEL**
- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

**IMPROVING MODEL**
- Feature Engineering
- Algorithm Tuning

**FINDING THE BEST PERFORMING CLASSIFICATION MODEL**
- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

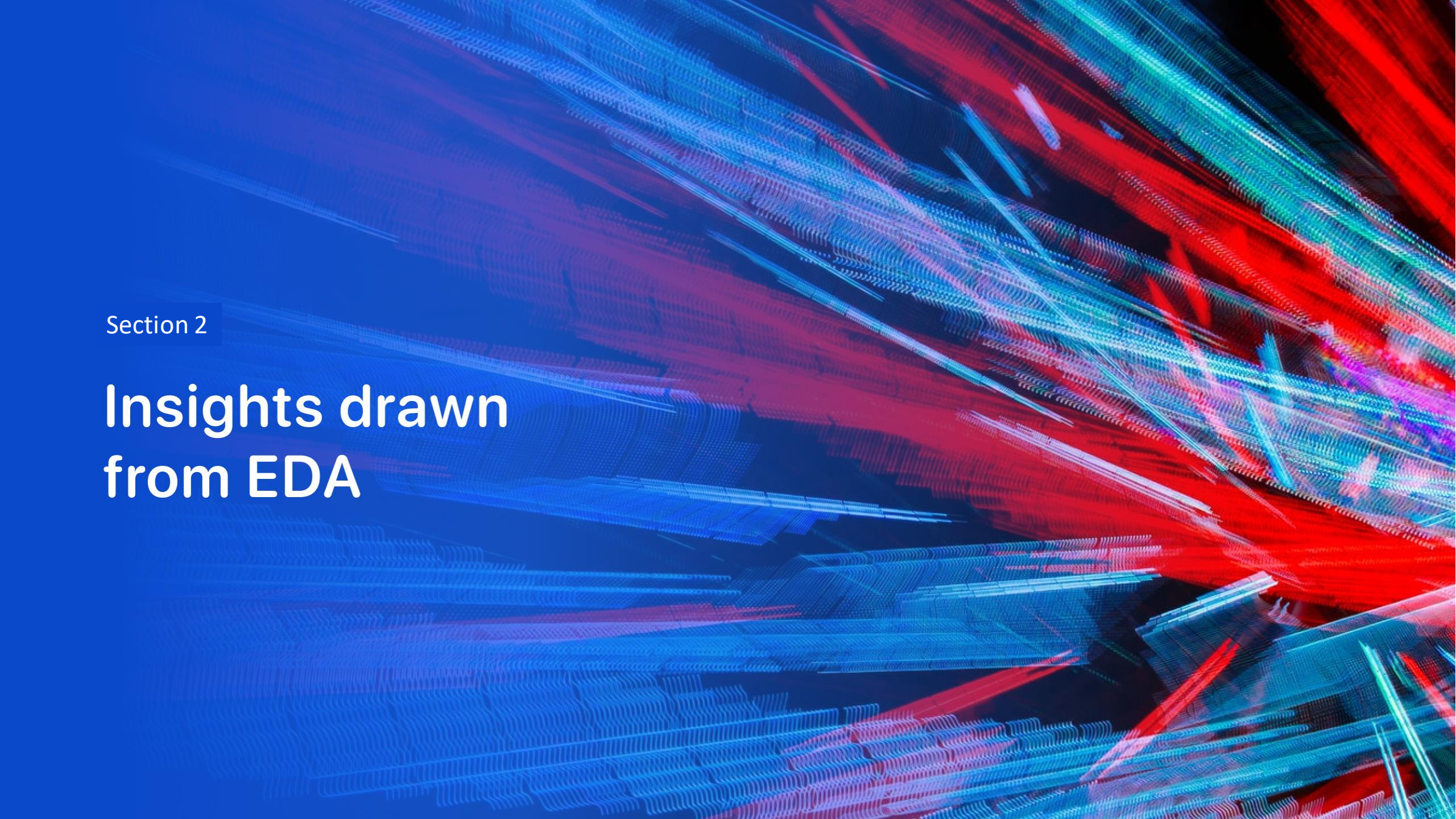
- GitHub url to Notebook

# Results

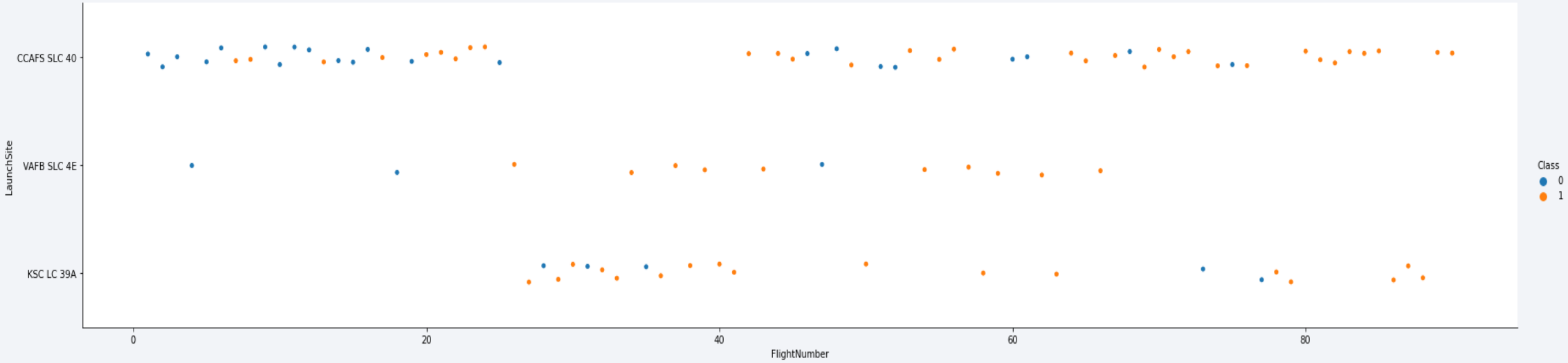
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

# Insights drawn from EDA

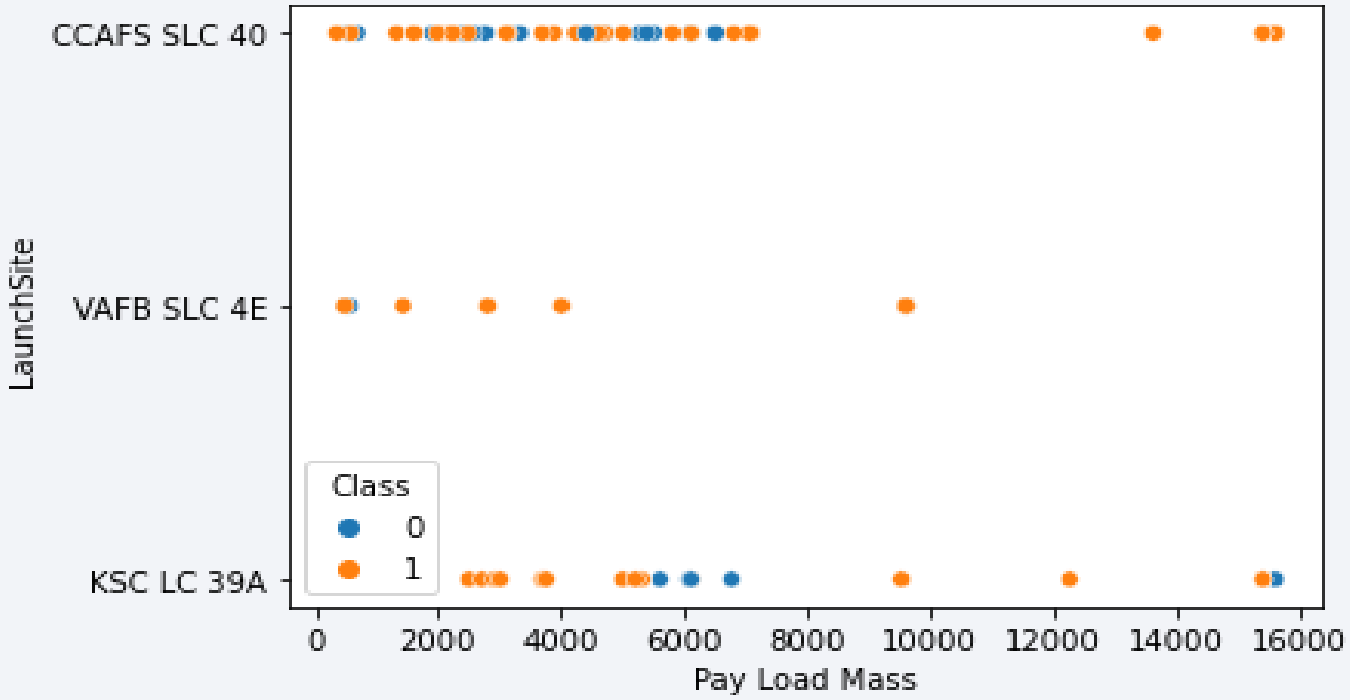


# Flight Number vs. Launch Site



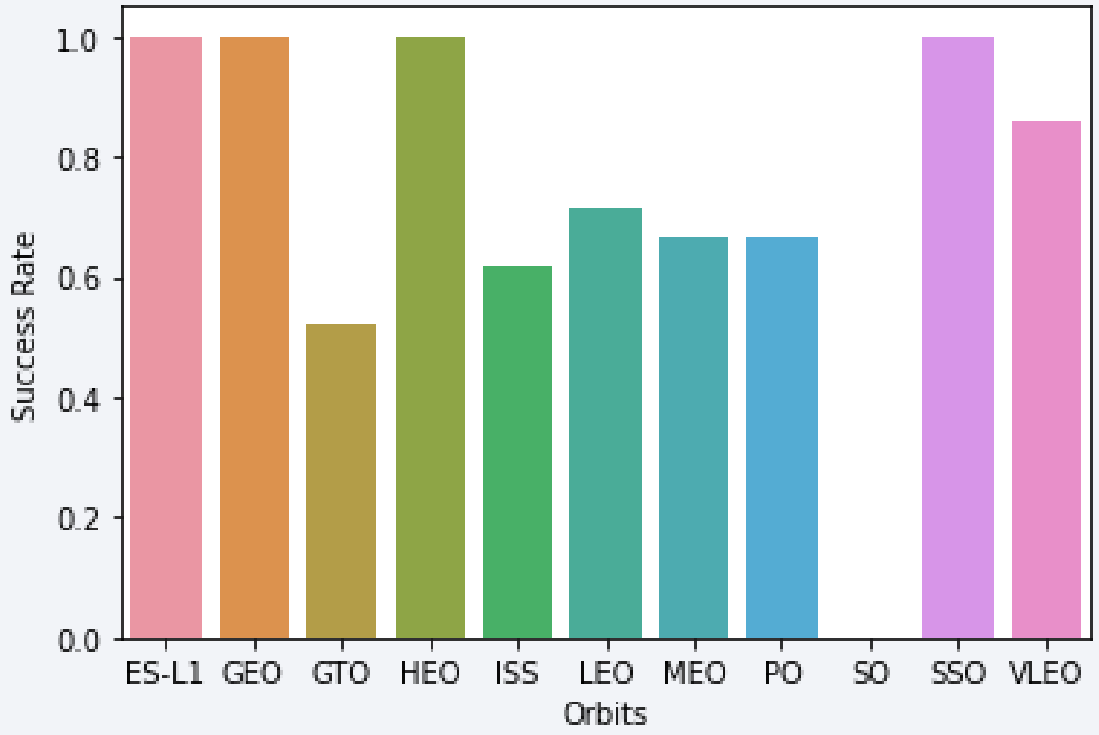

The trend seems to be, more amount of flights at a launch site, greater the success rate at a launch site.

# Payload vs. Launch Site



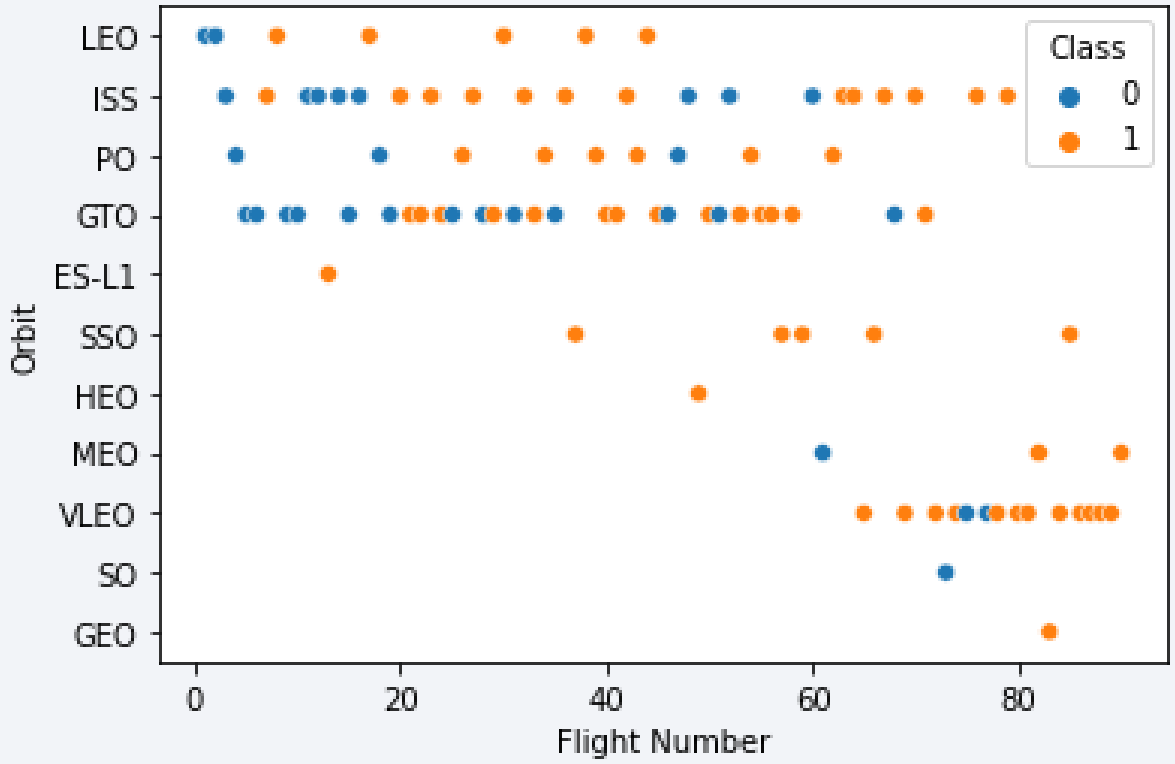

The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.

# Success Rate vs. Orbit Type



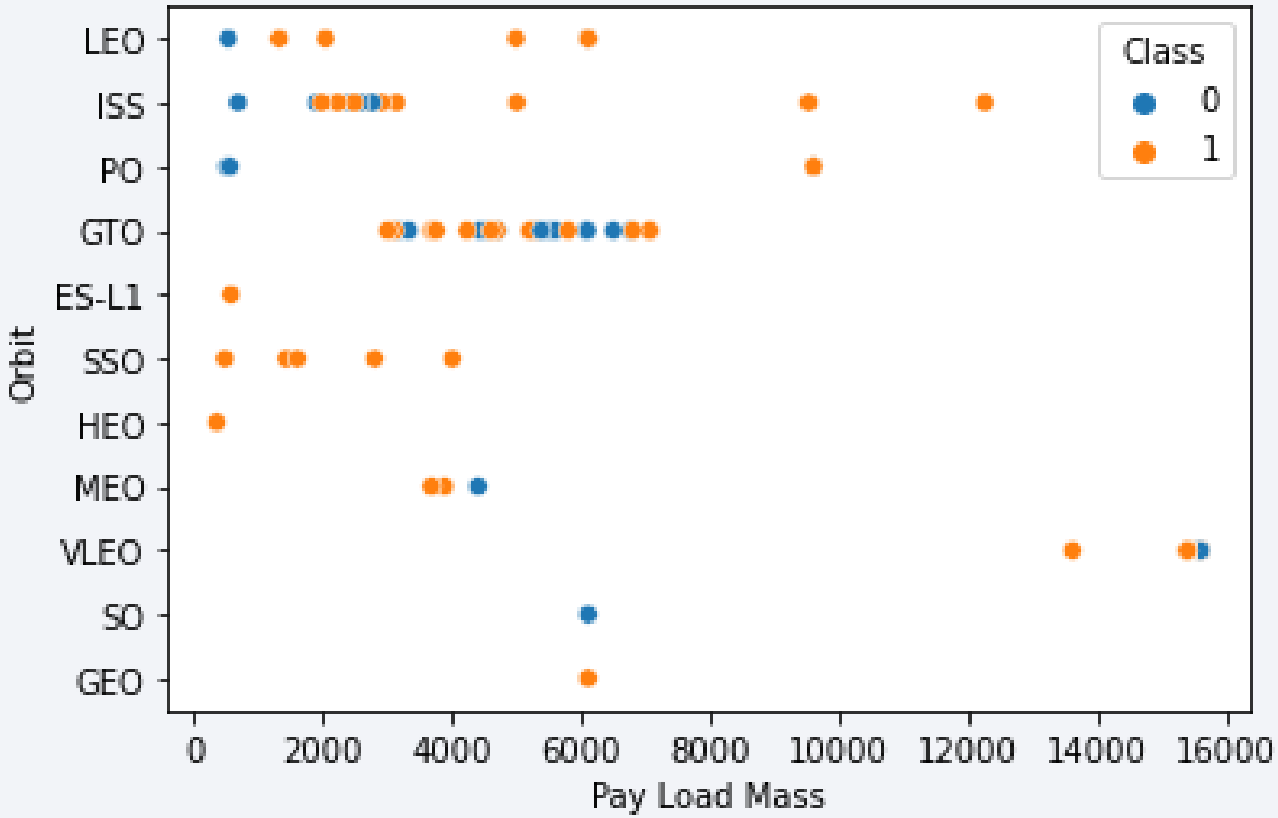

- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

# Flight Number vs. Orbit Type



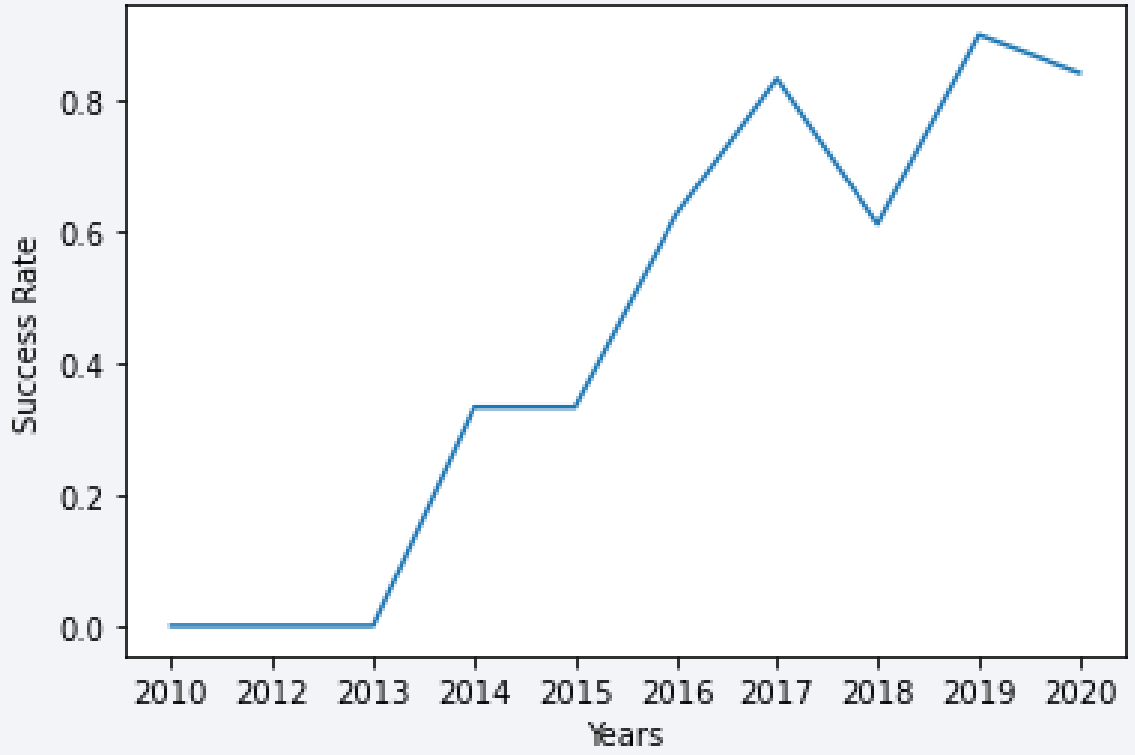

- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

# Payload vs. Orbit Type




- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

# Launch Success Yearly Trend




- We can observe that the sucess rate since 2013 kept increasing till 2020.

# All Launch Site Names

- CCAFS LC-40
  CCAFS SLC-40
  CCAFS SLC-40
  KSC LC-39A
  VAFB SLC-4E
- Using 'DISTINCT' in the query and it will only
  show Unique values in the Launch_Site column from tblSpaceX

# Launch Site Names Begin with 'CCA'

- select TOP 5 * from tblSpaceX
  WHERE Launch_Site LIKE 'KSC%
- TOP 5 means it will only show 5 records and LIKE keyword has a wild card with the words ‘KSC%’ the percentage in the end suggests that the Launch_Site name must start with KSC

# Total Payload Mass

- select SUM(PAYLOAD_MASS_KG_)  TotalPayloadMass from tblSpaceX where Customer = 'NASA (CRS)'",'TotalPayloadMass
- Total payload mass = 45596
- SUM ads each value in the column PAYLOAD_MASS_KG_
  WHERE filters the dataset to only perform calculations on Customer NASA (CRS)

# Average Payload Mass by F9 v1.1

- select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX where Booster_Version = 'F9 v1.1'
- Average payload mass = 2928
- AVG calculate the average in the column PAYLOAD_MASS_KG_
  WHERE filters the dataset to only perform calculations on Booster_version F9 v1.1

# First Successful Ground Landing Date

- select MIN(Date) SLO from tblSpaceX where Landing_Outcome = ”Success (drone ship)”
- First Successful Ground Landing Date : 06-05-2016
- MIN calculates the minimum date in the column Date
  WHERE filters the dataset to only perform
  calculations on Landing_Outcome Success (drone ship)

# Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B1032.1
- F9 B4 B1040.1
- F9 B4 B1043.1
- select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (ground pad)'
  AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000
- WHERE filters the dataset to Landing_Outcome = Success (drone ship)
  AND specifies additional filter conditions Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

# Total Number of Successful and Failure Mission Outcomes

- Total number of successful = 100
- Total number of failure = 1
- SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
  LIKE '%Success%') as Successful_Mission_Outcomes,
  (SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
  LIKE '%Failure%') as Failure_Mission_Coutcomes
- We use subqueries that are a composition of the previous ones.

# Boosters Carried Maximum Payload

- F9 B5 B1048.4
- F9 B5 B1048.5
- F9 B5 B1049.4
- F9 B5 B1049.5
- F9 B5 B1049.7
- SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS _KG_) AS [Maximum Payload Mass]
  FROM tblSpaceX GROUP BY Booster_Version
  ORDER BY [Maximum Payload Mass] DESC
- DISTINCT will only show Unique values in the Booster_Version column
  GROUP BY puts the list in order set to a certain condition.
  DESC arranges the dataset into descending order

# 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) |

- SELECT DATENAME(month, DATEADD(month, MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month,
  Booster_Version, Launch_Site, Landing_Outcome
  FROM tblSpaceX
  WHERE (Landing_Outcome LIKE N'%Success%')
  AND
  (YEAR(CONVERT(date, Date, 105)) = '2017')

- MONTH function returns name month. CONVERT converts NVARCHAR to Date.
  WHERE filters Year to be 2017

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

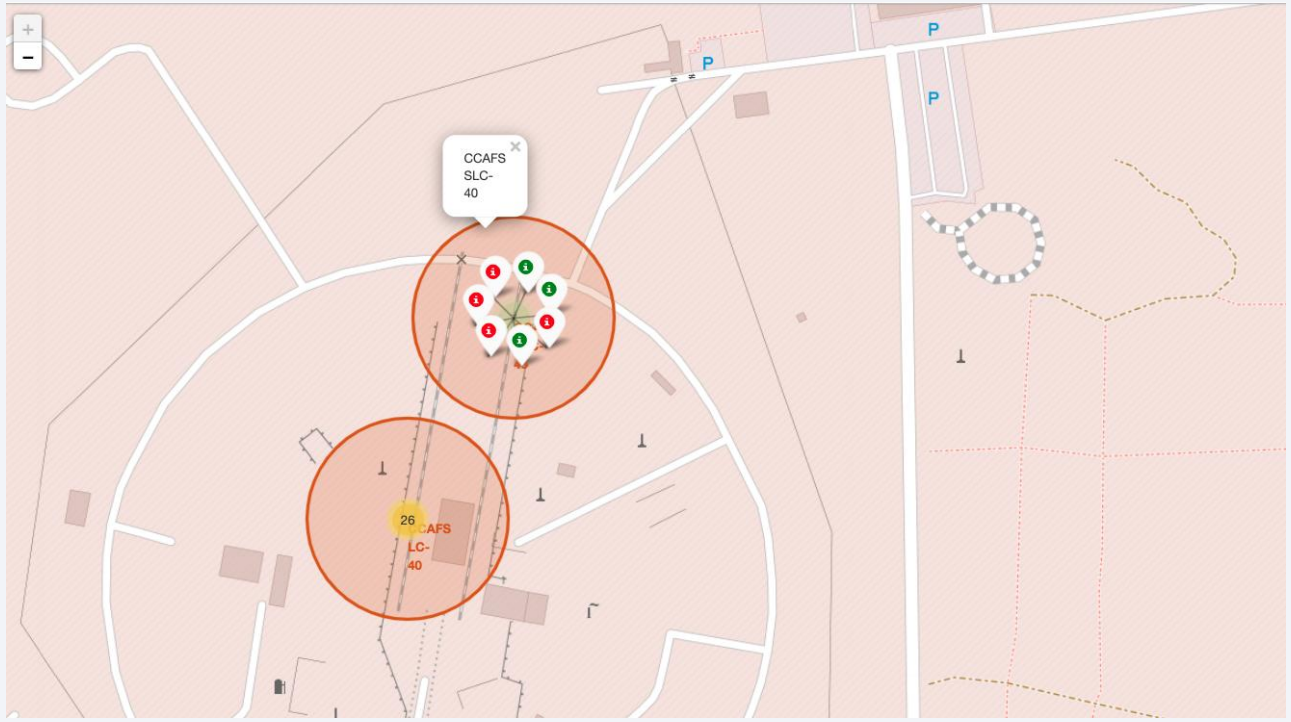
- Successful landing outcomes between the date 2010-06-04 and 2017-03-20 = 34
- SELECT COUNT(Landing_Outcome)
  FROM tblSpaceX
  WHERE (Landing_Outcome LIKE '%Success%')
  AND (Date > '04-06-2010')
  AND (Date < '20-03-2017')
- COUNT counts records in column
  WHERE filters data


Section 3
Launch Sites
Proximities Analysis

# Launch sites

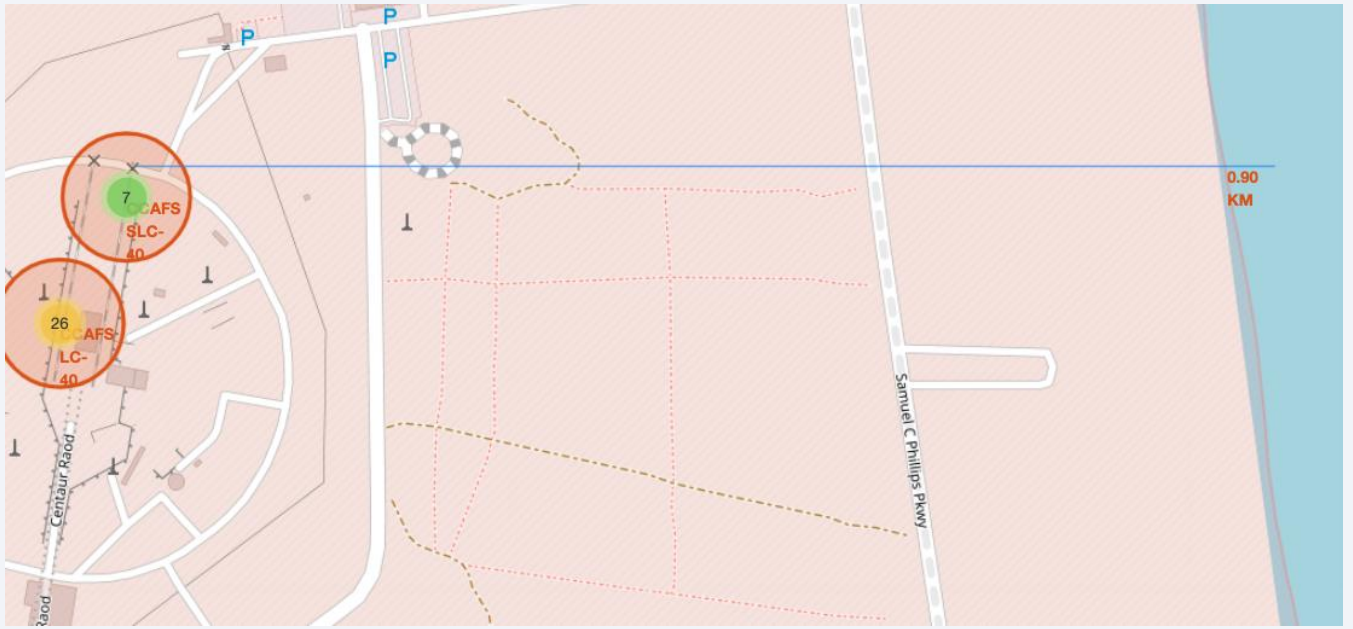

SpaceX launch sites are in Florida and California

# Colour Labelled Markers

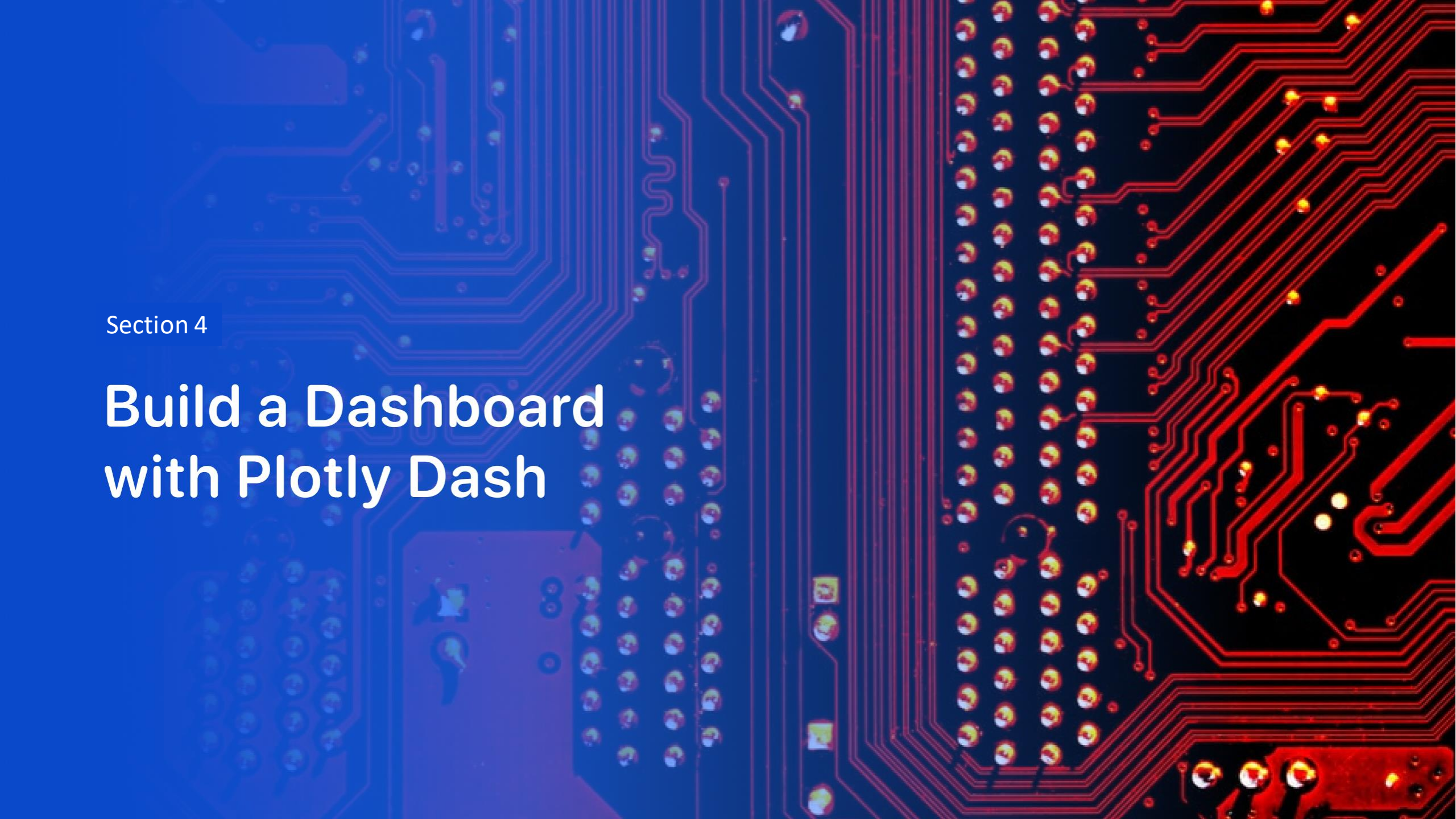



Green Marker shows successful Launches and Red Marker shows Failures

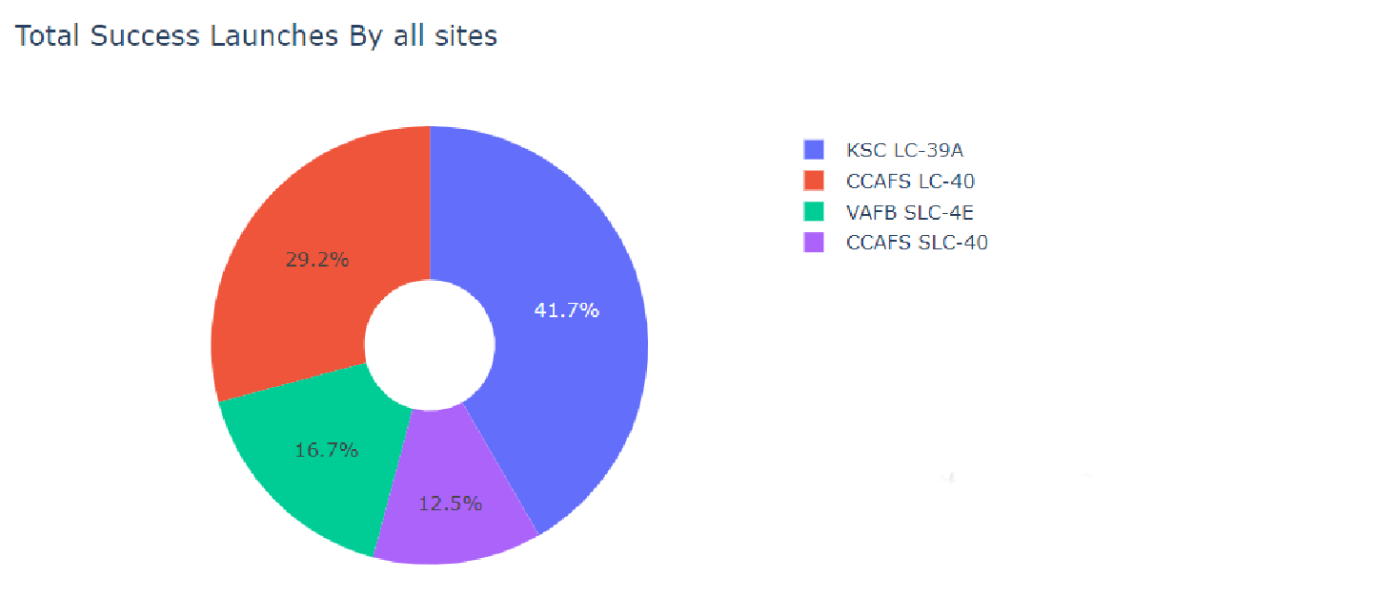
# Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference




CCSFS-SCL-40 Distance to coast

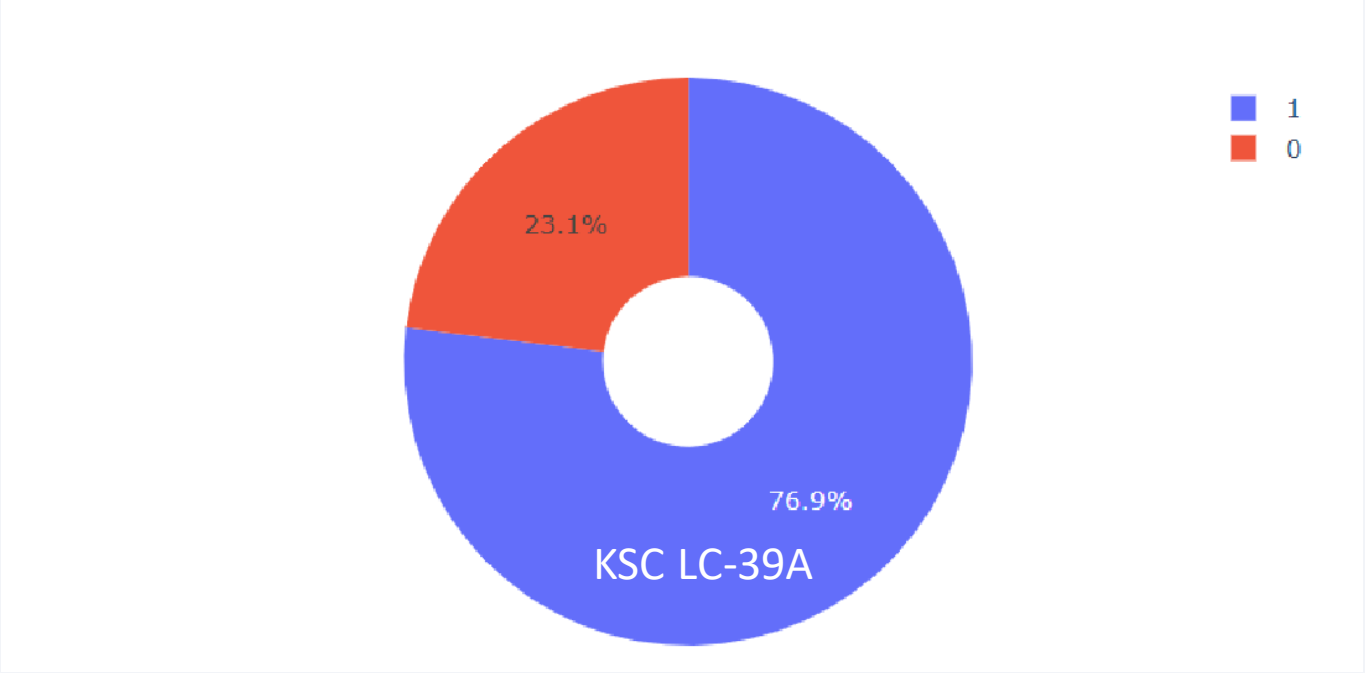

Section 4
Build a Dashboard with Plotly Dash

# DASHBOARD – Pie chart, success percentage achieved by each launch site



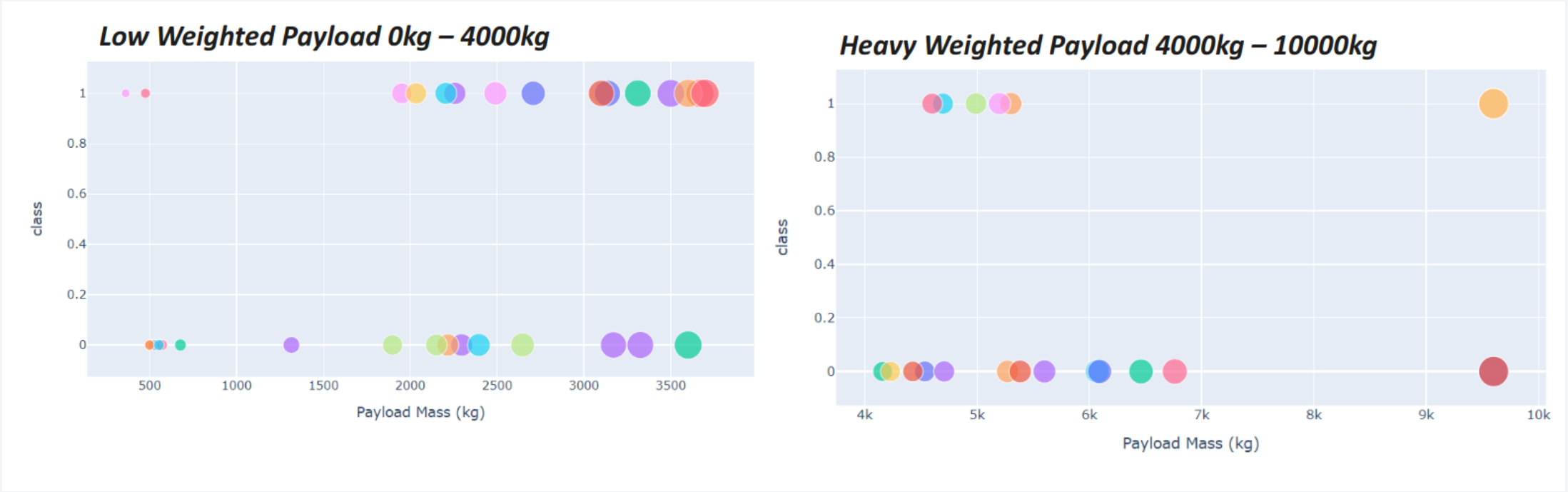

KSC LC-39A had the highest number of success launches

# DASHBOARD – Pie chart for the launch site with highest launch success ratio




KSC LC-39A achieved a 76.9% success rate

# DASHBOARD – Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

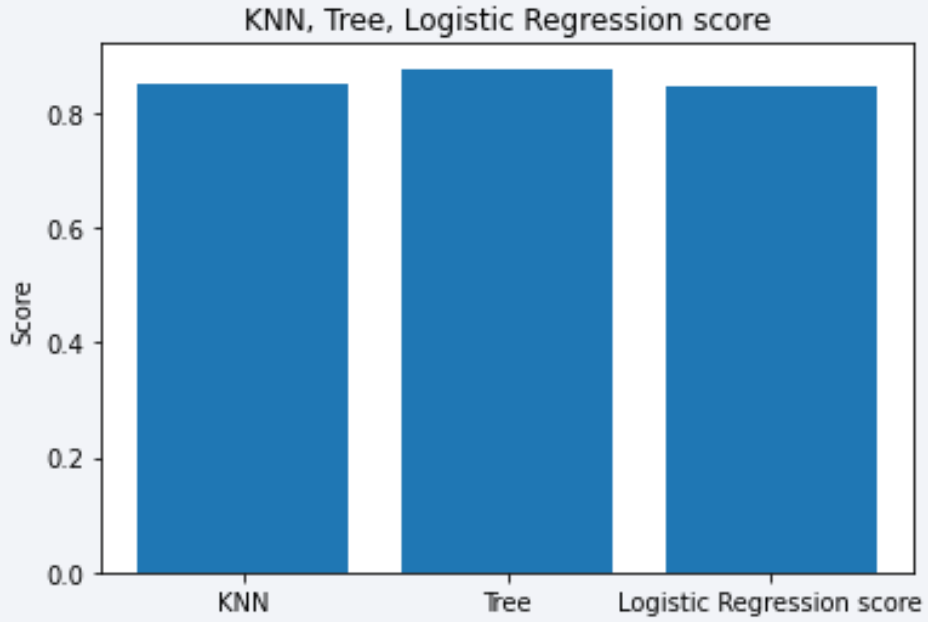



Success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy




- Tree Classification has the highest score using training data, but with the test data all performed the same, the reason for this is the low number of samples for training, 18.

# Confusion Matrix for tree

- We see that Tree can distinguish between the different classes. The major problem is false positives.

# Conclusions

- The Three Classifier Algorithms performed the same for this dataset
- KSC LC-39A had the most successful launches from all the sites
- The success rates for SpaceX launches is proportional to the time in years they start
- Low weighted payloads perform better than the heavier payloads
- Orbits GEO,HEO,SSO,ES-L1 has the best Success Rate

# Thank you!