IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

<Tony Trinh>
<24/06/2022>

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with S Q L
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result from Machine Learning Lab

# 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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - Determine the factors associated for successful landing of the rocket at first stage.
  - Relationship with each rocket variables for each outcome at the end.
  - Past history and its future prediction for successful landing using Machine Learning.

Section 1

# Methodology

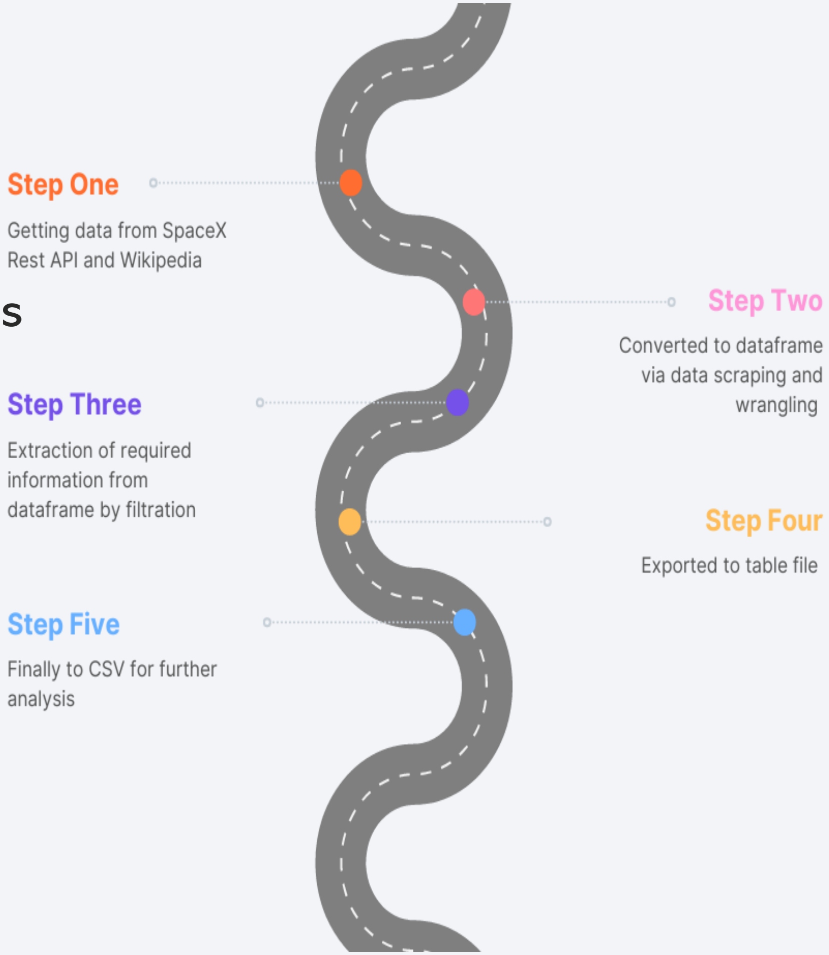
# Methodology

## Executive Summary

- Data collection methodology:
  - FromSpaceXRestAPI provided by course
  - Wikipedia via webscraping
- Perform data wrangling
  - Data was processed using one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - User build and evaluate classification models

# Data Collection

- Describe how data sets were collected.
    - Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. General steps being used for collection of data are as follows---
- You need to present your data collection process use key phrases and flowcharts



Step One

Getting data from SpaceX Rest API and Wikipedia

Step Two

Converted to dataframe via data scraping and wrangling

Step Three

Extraction of required information from dataframe by filtration

Step Four

Exported to table file

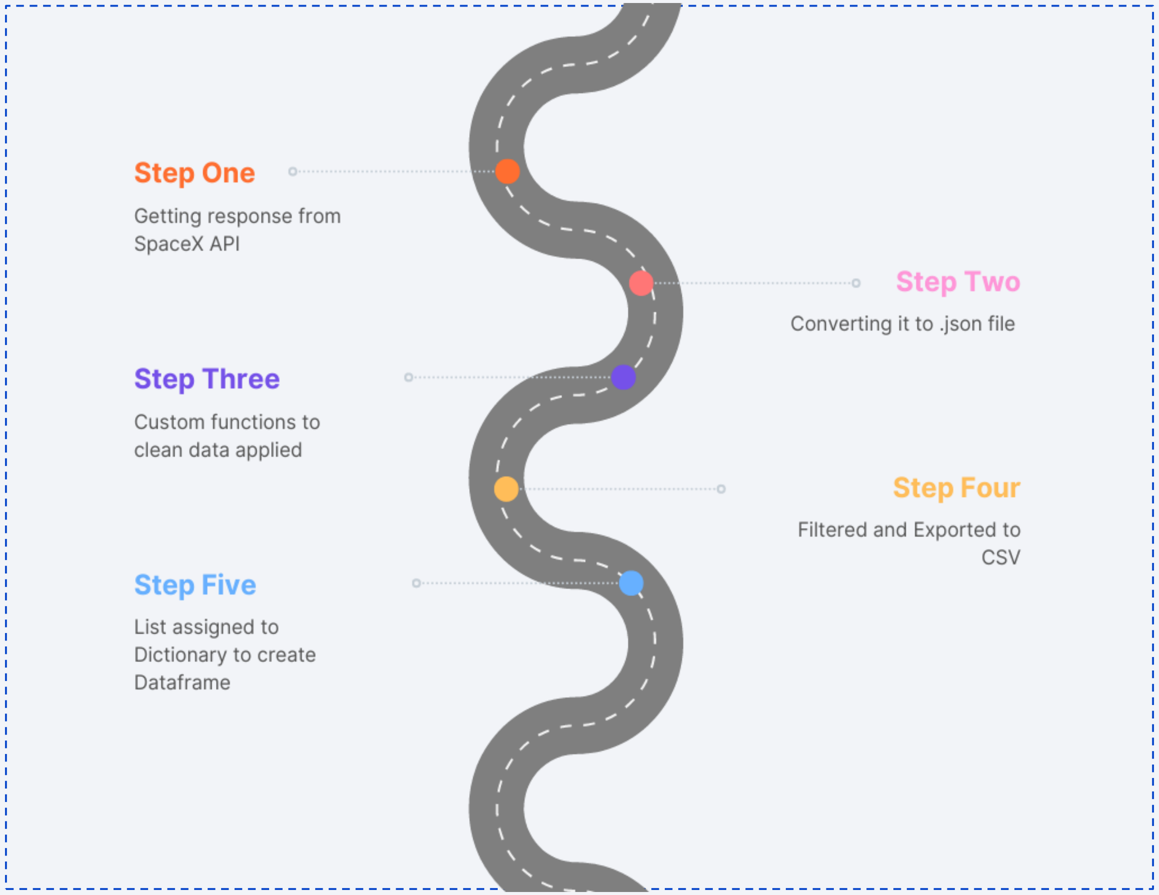
Step Five

Finally to CSV for further analysis

# Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose



Step One

Getting response from SpaceX API

Step Two

Converting it to .json file

Step Three

Custom functions to clean data applied

Step Four

Filtered and Exported to CSV

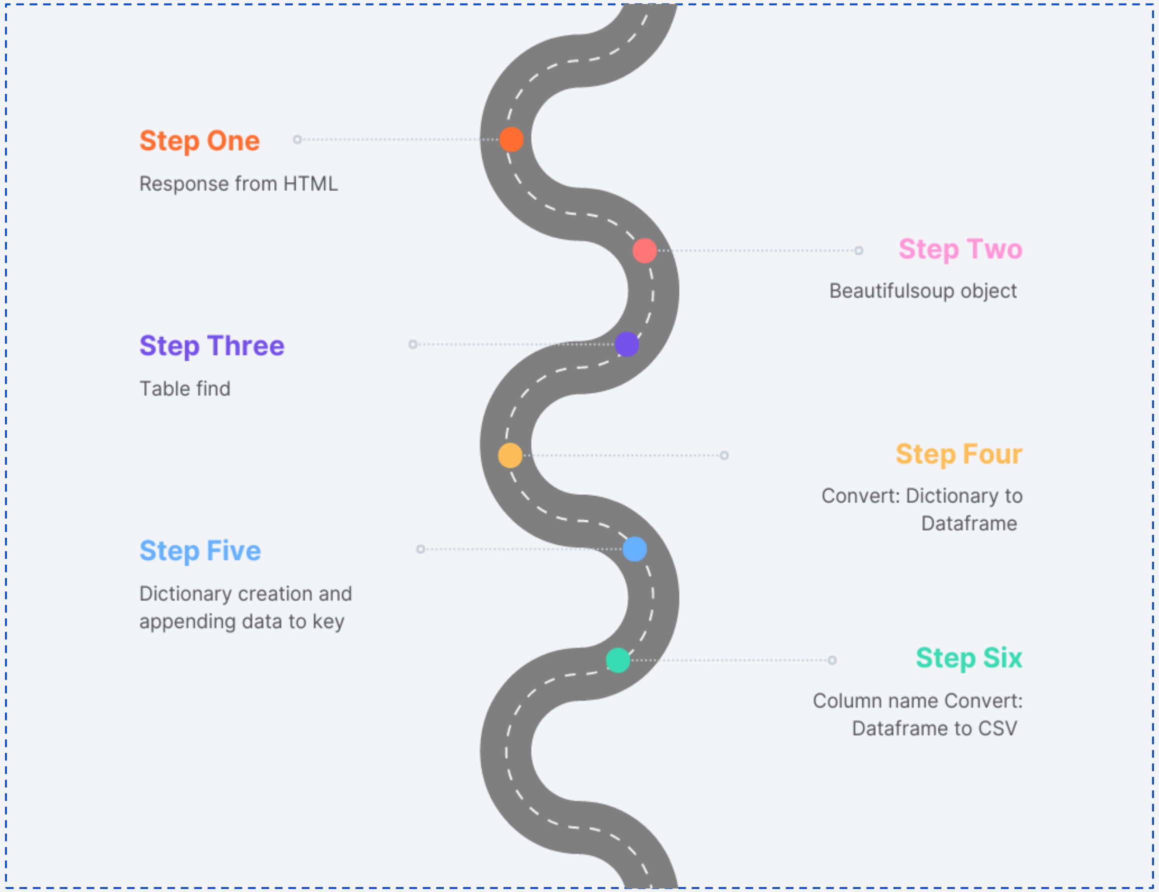
Step Five

List assigned to Dictionary to create Dataframe

# Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



**Step One**
Response from HTML

**Step Two**
Beautifulsoup object

**Step Three**
Table find

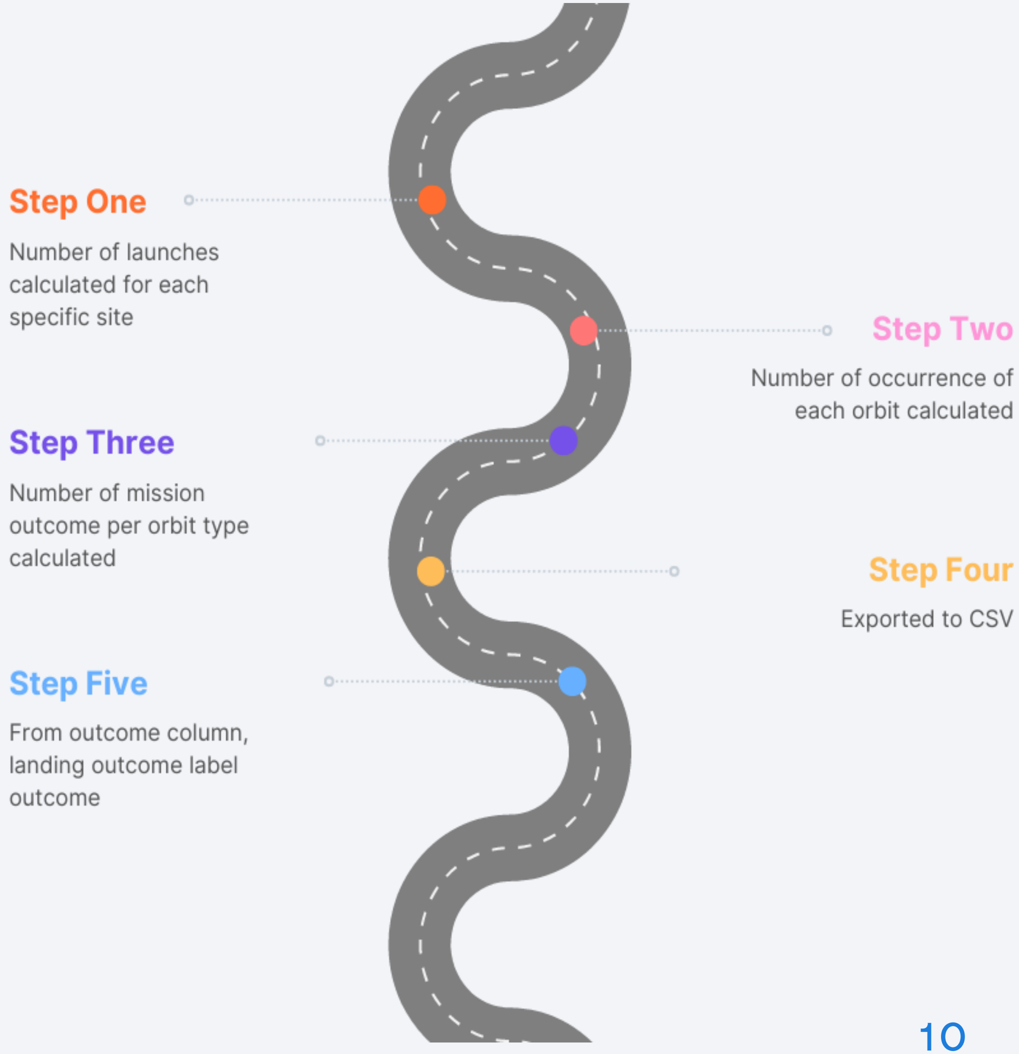
**Step Four**
Convert: Dictionary to Dataframe

**Step Five**
Dictionary creation and appending data to key

**Step Six**
Column name Convert: Dataframe to CSV

# Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.
- You need to present your data wrangling process using key phrases and flowcharts



**Step One**

Number of launches calculated for each specific site

**Step Two**

Number of occurrence of each orbit calculated

**Step Three**

Number of mission outcome per orbit type calculated

**Step Four**

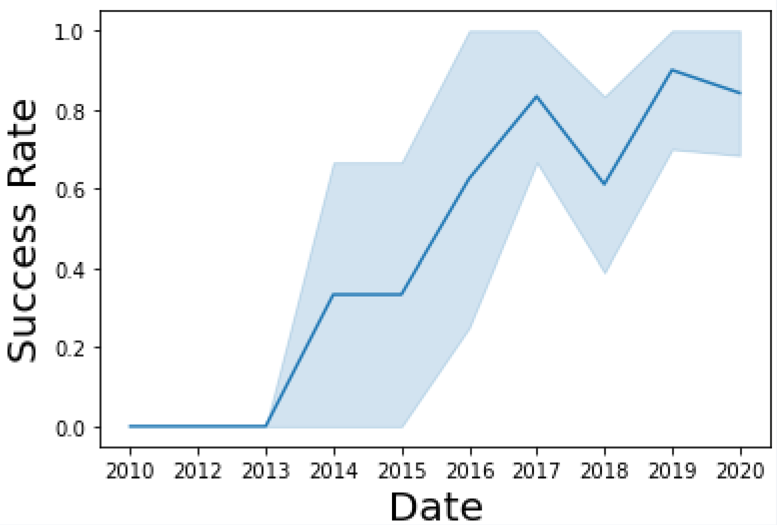
Exported to CSV

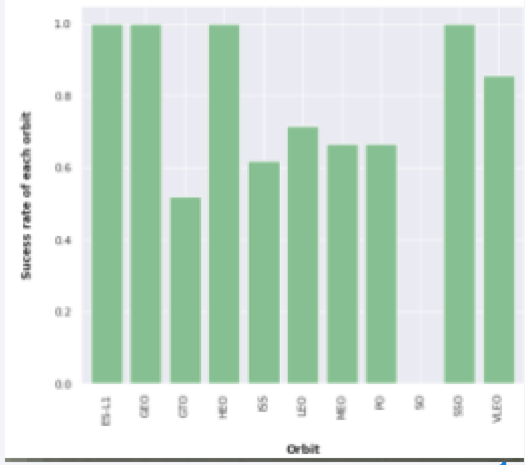
**Step Five**

From outcome column, landing outcome label outcome

# EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Following kinds of graphs has been prepared using matplotlib.pyplot.
  - Flight number vs PayloadMass
  - Flight numbervs Launchsite
  - Payload vs Launchsite
  - Flightnumber vs Orbittype
  - Payload vs Orbittype





# EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Displaying the names of the launch sites.
  - Displaying 5 records where launch sites begin with the string ‘CCA’.
  - Displaying the total payload mass carried by booster launched by NASA (CRS).
  - Displaying of average payload mass carried by booster version F9 v1.1.
  - Listing the date when the first successful landing outcome in ground pad was achieved.
  - Listing the names of the boosters which have success in drone ship 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 failed landing outcomes in drone ship, their booster versions, and launch sites names for in year 2015.
  - Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

# Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

  - Folium actually makes it very easy to understand and visualize data that has been processed using Python on

    interactiveleafletmap.This libraryuses coordinates(latitudeandlongitude)forlocatingtheeachspecificsiteandbeing

    circledwithlabeledname.Although,witheasy interactiveunderstandingtolocationonmap,launchsiteshas alsobeen

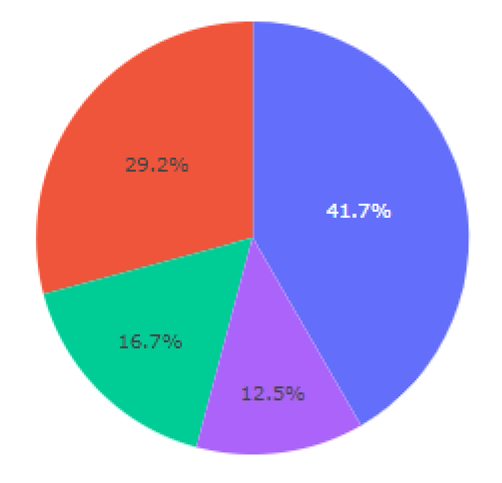
    demarcated,successfullauncheswithclass ‘1’asGreenandfailurelauncheswithclass ‘0’asRed.Andalsodistance

    fromthecoastlinealsobeencalculatedinkm.

- Explain why you added those objects

# Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - An interactive dashboard with Plotly dash is built which completely user friendly and fun to play for observing the changes happening with different components.
  - Pie chart is prepared showing total launches by a certain site and as a whole.
  - A scatter plot graph is also prepared showing relation without come and payload mass(kg) for different booster version– which is again an user interactive plot to play with.



# Predictive Analysis (Classification)

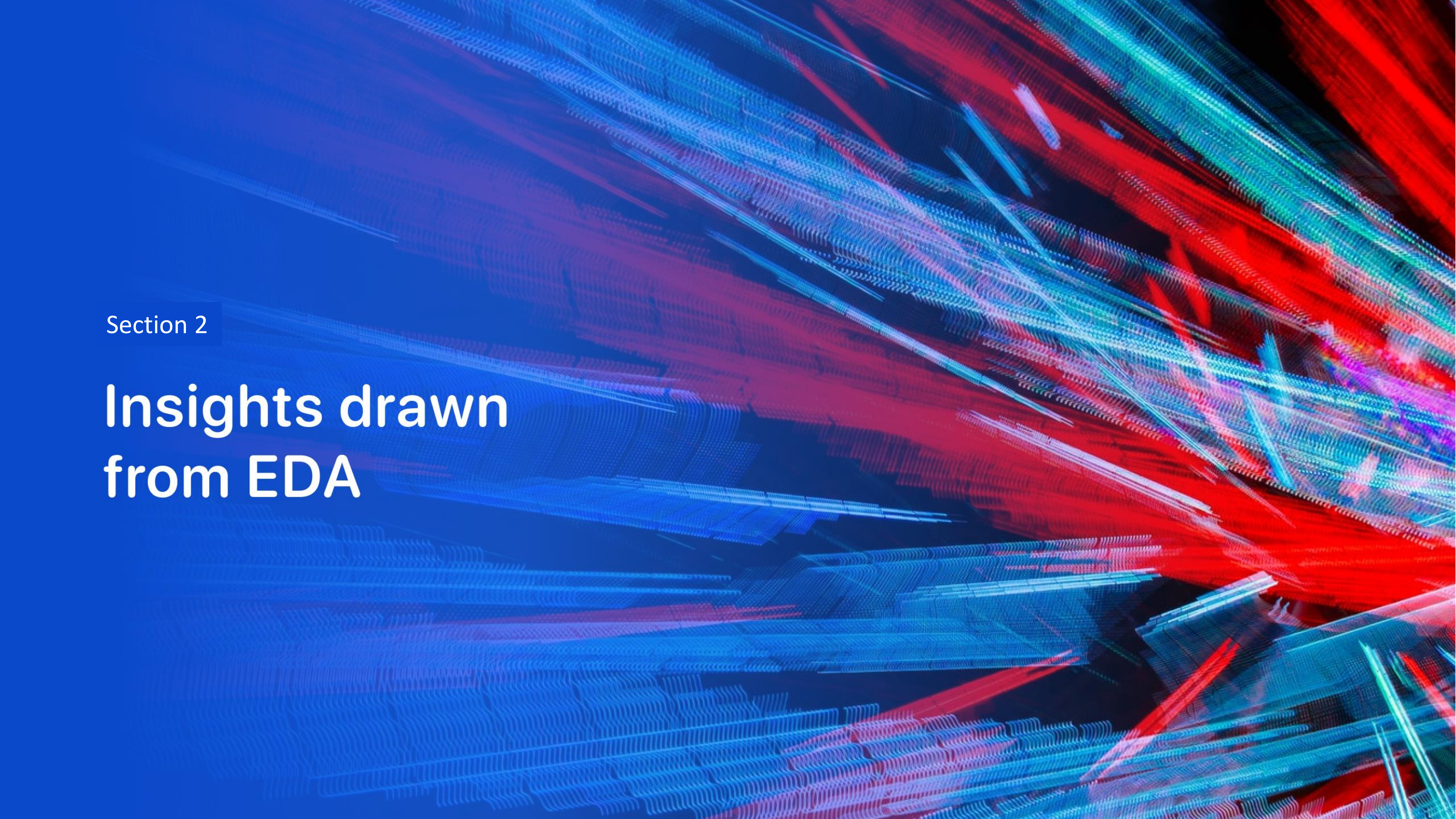
- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Load the data set into NumPy and Pandas
  - Transform the data and then split into training and test datasets.
  - Decide which type of ML to use.
  - Set the parameters and algorithms to Grid Search CV and fit it to dataset.
- You need present your model development process using key phrases and flowchart
  - Check the accuracy for each model
  - Get tuned hyperparameters for each type of algorithms.
  - Plot the confusion matrix.
  - Use Feature Engineering and Algorithm Tuning
  - The model with the best accuracy score will be the best performing model.

# Results

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

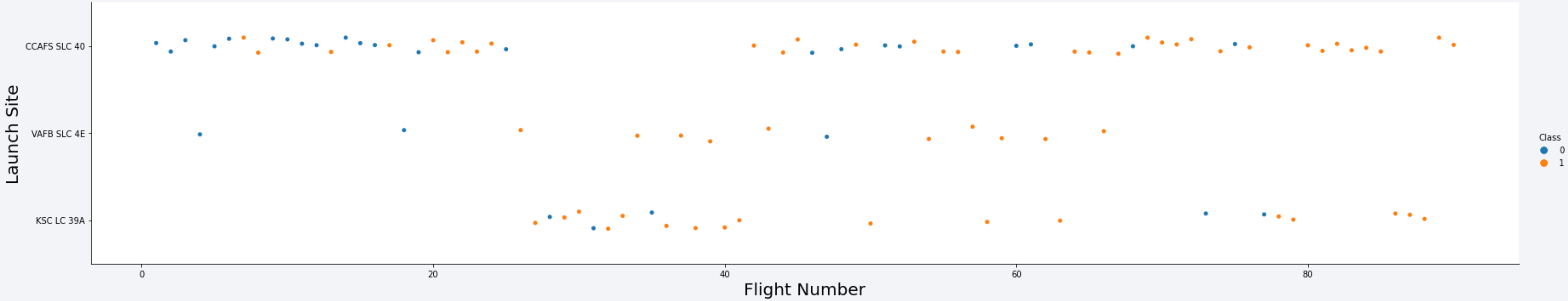
Section 2

# Insights drawn from EDA



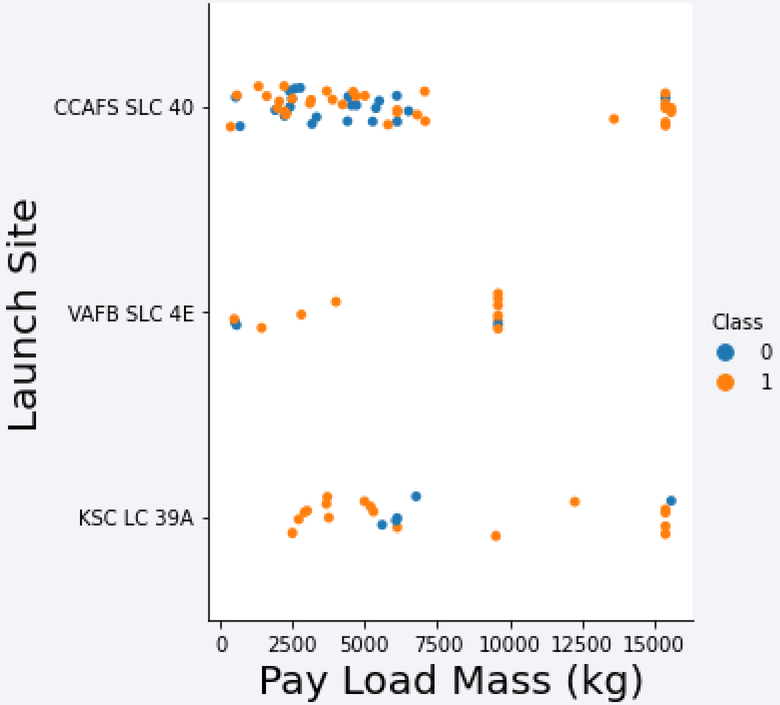
# Flight Number vs. Launch Site

Presented scatter plot shows the larger the flights amount in the launch site, the greater the success rate be. However, site CCAFS SLC40 shows the least pattern



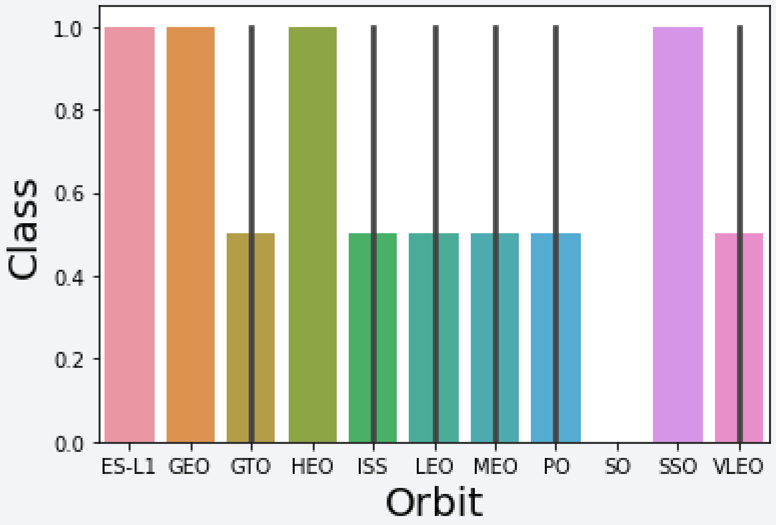
# Payload vs. Launch Site

Presented scatter plot shows the pay load mass is greater than 7000kg, and the probability of the success rate will be highly increased . However, particular there i s no clear and pattern for saying the launch site is completely dependent on the payload mass for the success rate.



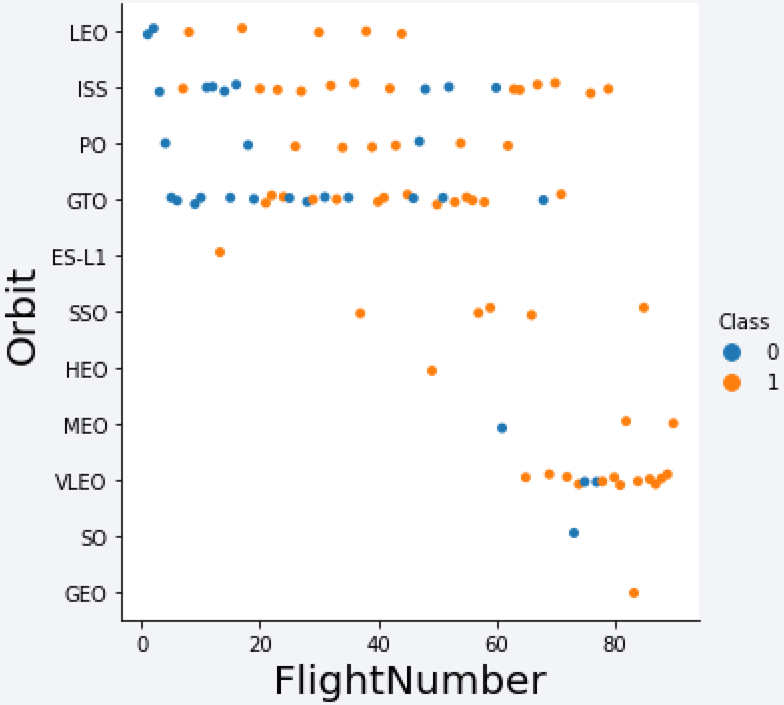
# Success Rate vs. Orbit Type

Present graph shows the possibility of the orbits to influences the landing outcomes which is for some orbits it is 100% success rate such as SSO, HEO, GEO AND ES - L1 while SO orbit produced 0% rate of success. However, deeper analysis show that 4 orbits has occurrence of 1suchas GEO, SO, HEO and ES - L 1 which mean this data need more dataset to seen pattern or trend before we draw any further conclusion .



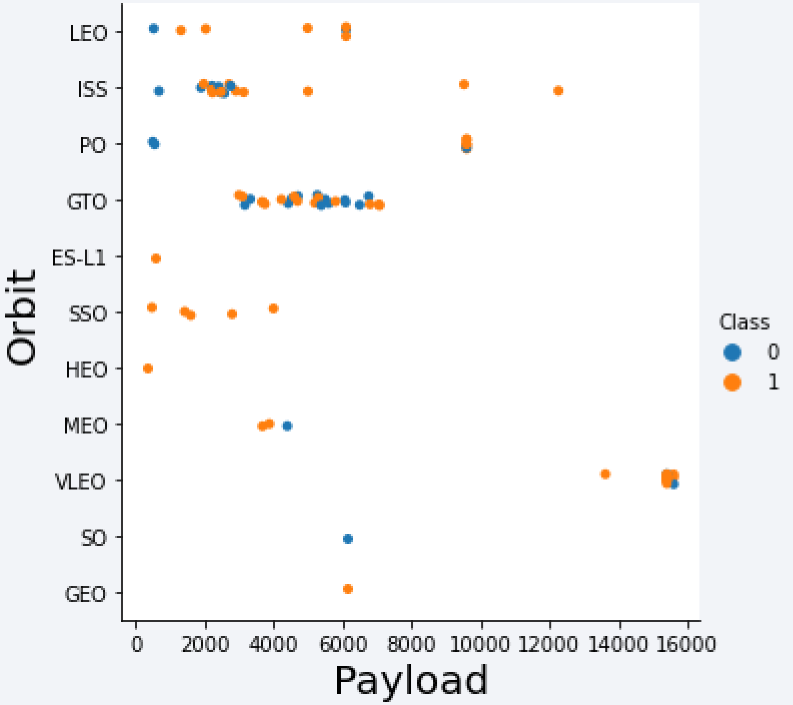
# Flight Number vs. Orbit Type

Present scatterplot shows that, the larger flight number on each orbits, the greater will be the success rate (especially LEO orbit) except for GTO orbit which depicts no relationship between both attributes. Orbit that only has 1 occurrence should also be excluded from above statement as it's needed more dataset.



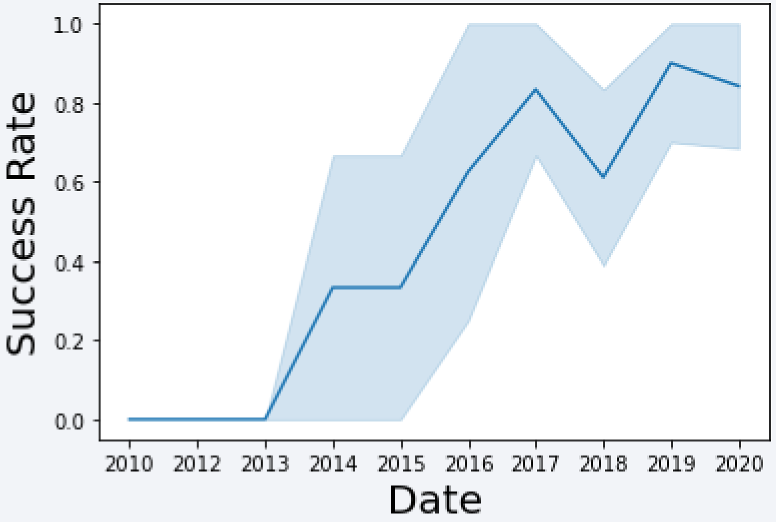
# Payload vs. Orbit Type

Heavier payload is showing the positive impact on LEO, ISS and PO orbit. However, it is also showing the negative impact on MEO and VLEO orbit. GTO orbit seem to depict no relation between the attributes. Meanwhile, again, SO, G andHI orbit need more dataset to see any pattern or trend.



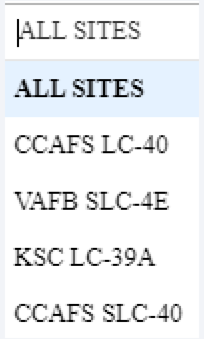
# Launch Success Yearly Trend

The present trend clearly showcases the increasing trend straight from year 2013 to 2020 but with some minor dips in 2015 ,2018 and 2020. If this trend continue for the next year onward. The success rate will steadily increase until reaching1/100% success rate.



# All Launch Site Names

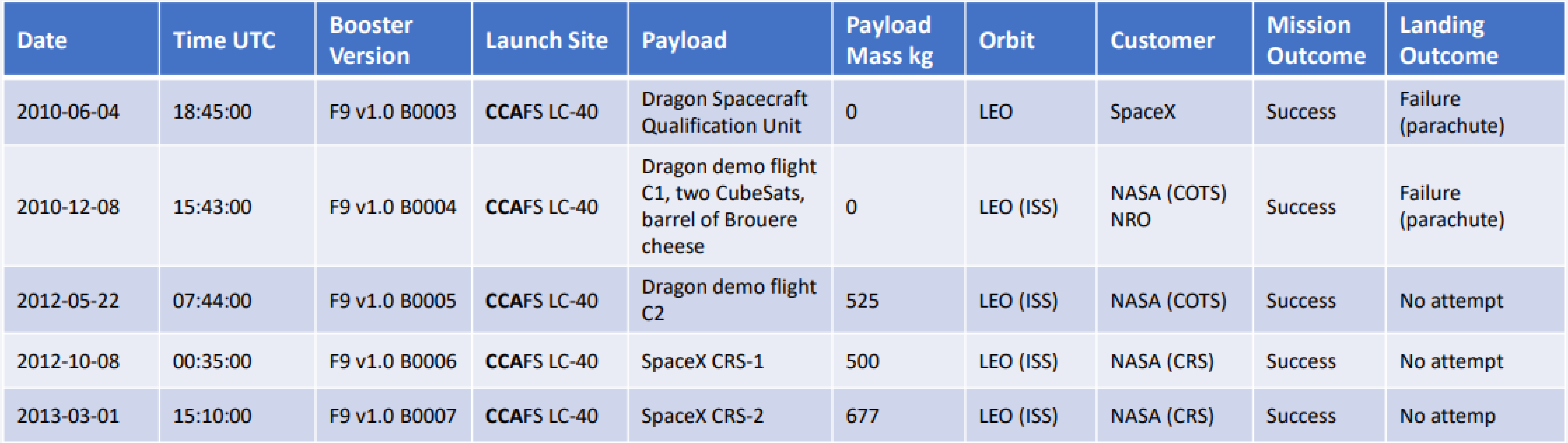
- 4 launch sites:



- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

# Launch Site Names Begin with 'CCA'

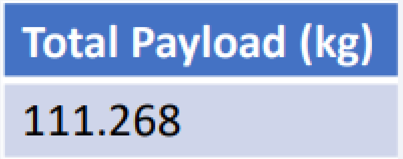
Here’s the ‘%sql’ is used to display the 5 records where launch sites begin with `CCA`



| Date | Time UTC | Booster Version | Launch Site | Payload | Payload Mass kg | Orbit | Customer | Mission Outcome | Landing Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 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) |
| 2012-05-22 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attemp |

# Total Payload Mass

Here the total payload mass is been calculated with ‘SUM’ query to carried by boosters from NASA which is 455%.

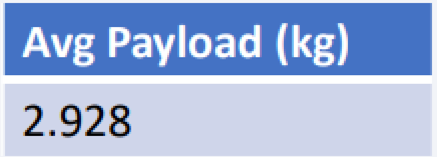


| Total Payload (kg) |
|---|
| 111.268 |

Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

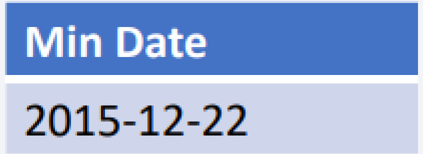


| Avg Payload (kg) |
|---|
| 2.928 |

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg

# First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

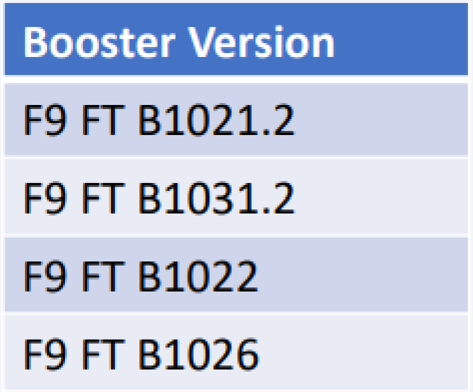


| Min Date |
|---|
| 2015-12-22 |

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.

## Successful Drone Ship Landing with Payload between 4000 and 6000

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

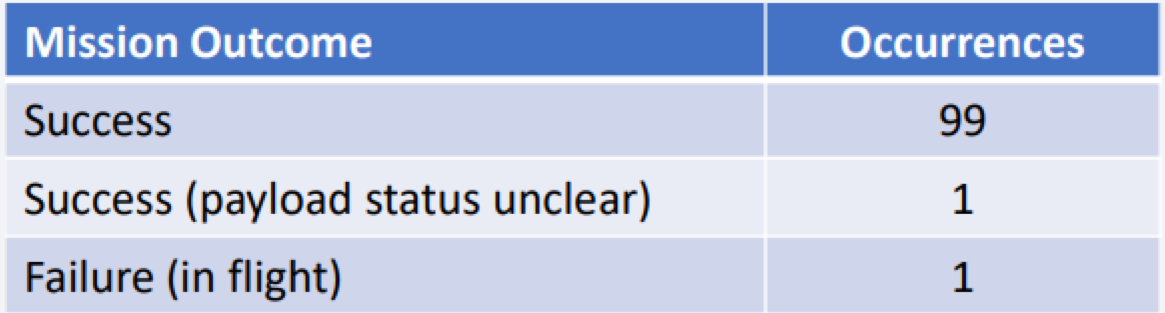


| Booster Version |
|---|
| F9 FT B1021.2 |
| F9 FT B1031.2 |
| F9 FT B1022 |
| F9 FT B1026 |

- Selecting distinct booster versions according to the filters above, these 4 are the result

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

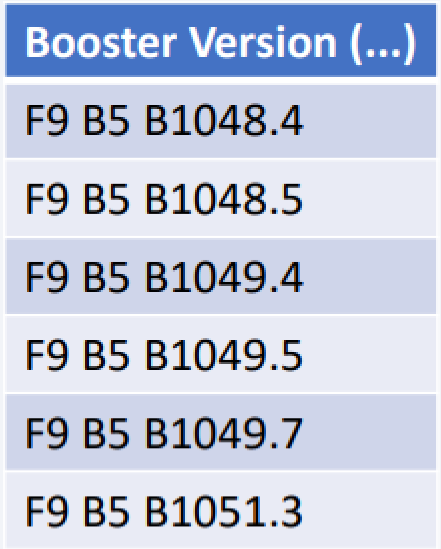


| Mission Outcome | Occurrences |
|---|---|
| Success | 99 |
| Success (payload status unclear) | 1 |
| Failure (in flight) | 1 |

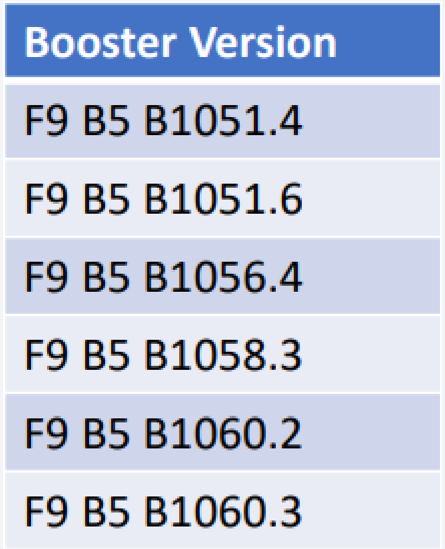
- Grouping mission outcomes and counting records for each group led us to the summary above.

# Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



| Booster Version (...) |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1048.5 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1049.7 |
| F9 B5 B1051.3 |

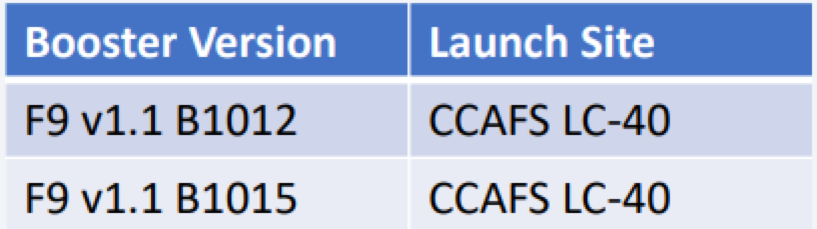


| Booster Version |
|---|
| F9 B5 B1051.4 |
| F9 B5 B1051.6 |
| F9 B5 B1056.4 |
| F9 B5 B1058.3 |
| F9 B5 B1060.2 |
| F9 B5 B1060.3 |

- These are the boosters which have carried the maximum payload mass registered in the dataset

# 2015 Launch Records

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

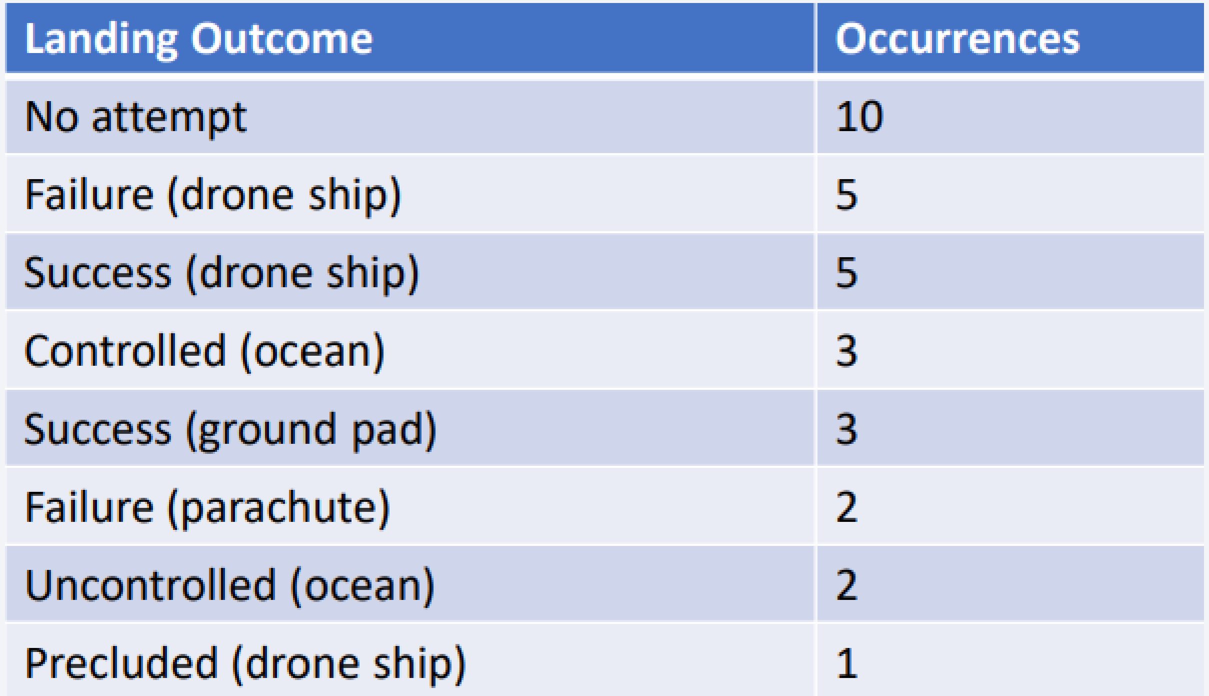


| Booster Version | Launch Site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

- • The list above has the only two occurrences.

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

- Rank 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



| Landing Outcome | Occurrences |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

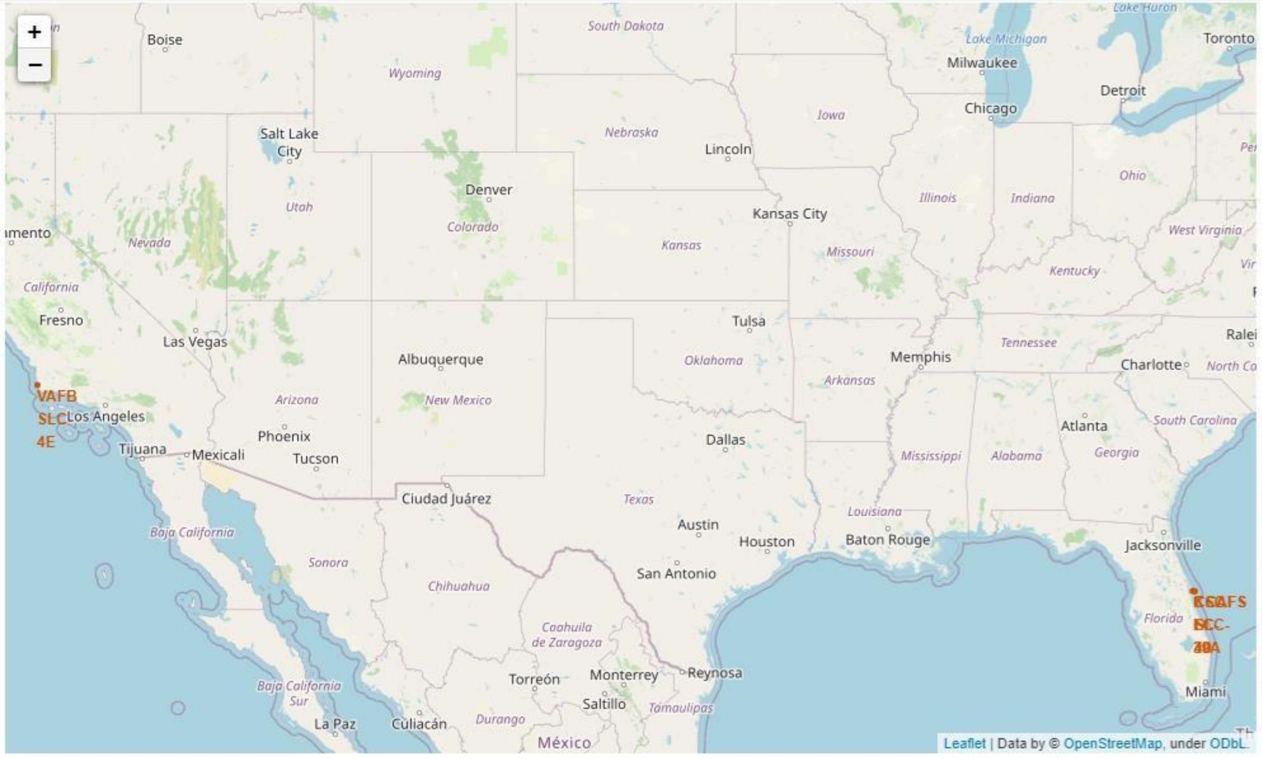
- • This view of data alerts us that “No attempt” must be taken in account.

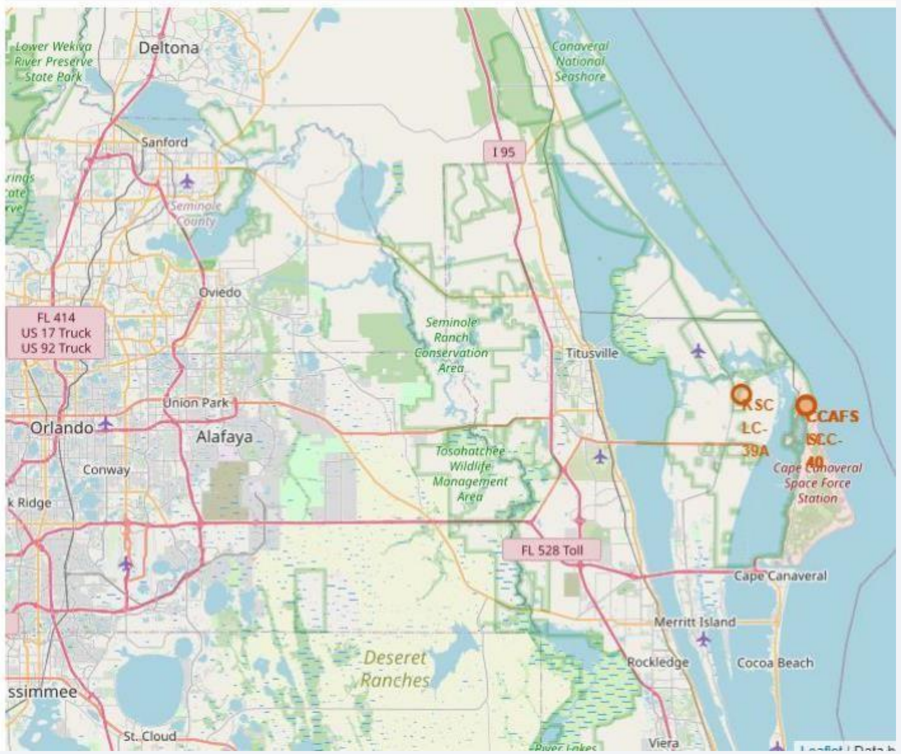
Section 3

# Launch Sites Proximities Analysis

# Launch Site Locations

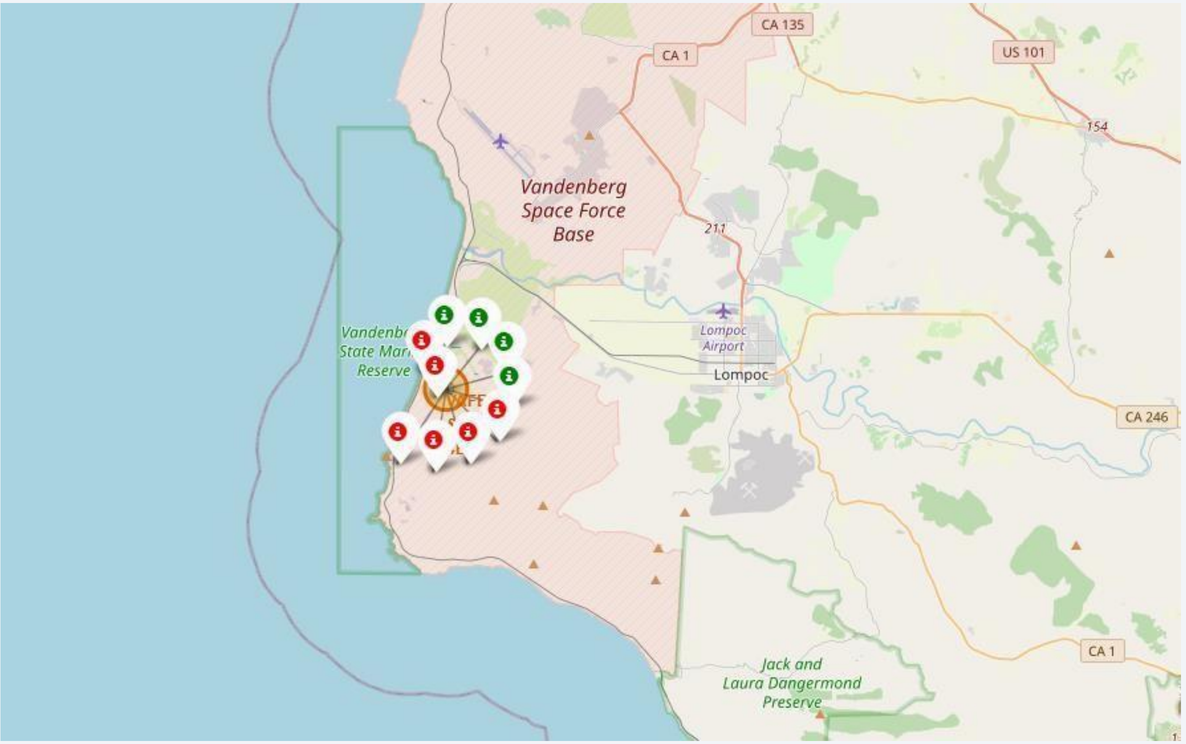
- The left map shows all launch sites relative US map. The right map shows the two Florida launch  sites since they are very close to each other. All launch sites are near the ocean.





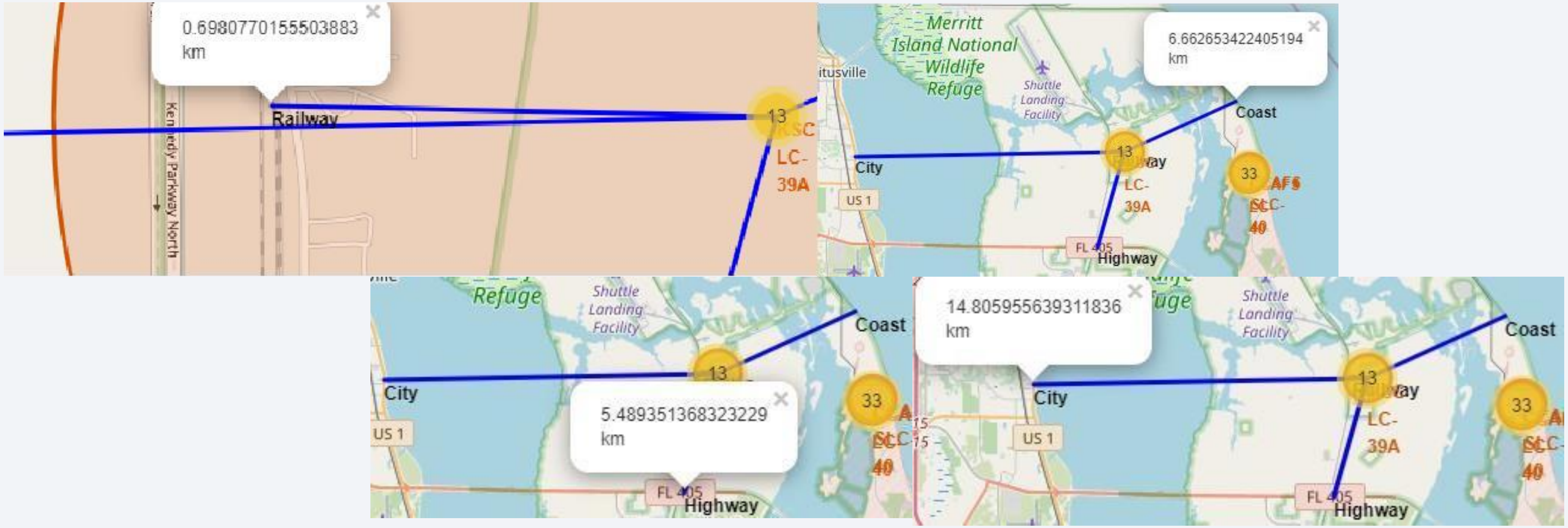
# Color-Coded Launch Markets

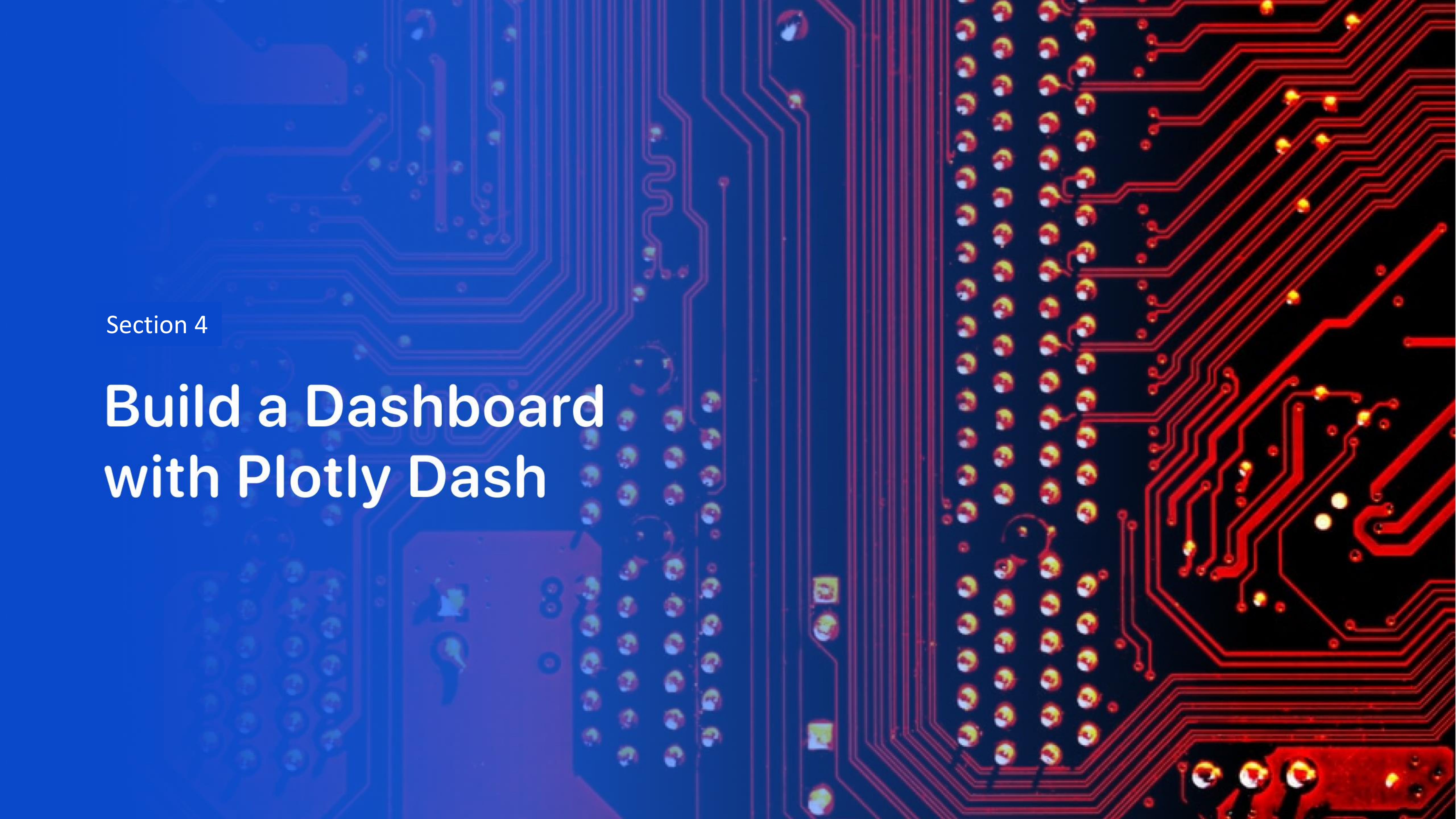
- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.


CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
Vandenb State Mari Reserve
CA 246
CA 1
Jack and Laura Dangermond Preserve

# Key Location Proximities

- Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.



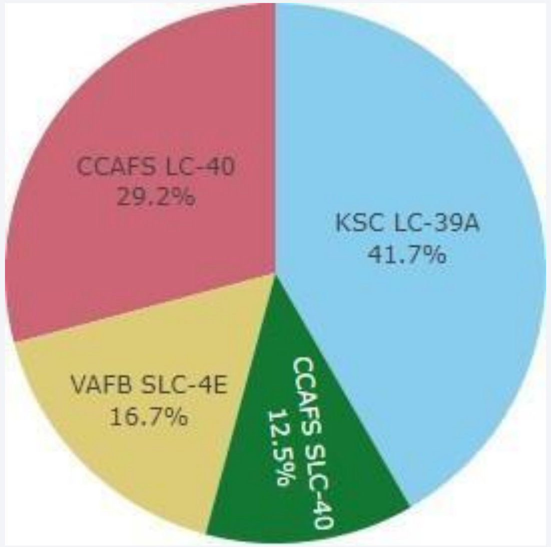


Section 4

# Build a Dashboard with Plotly Dash

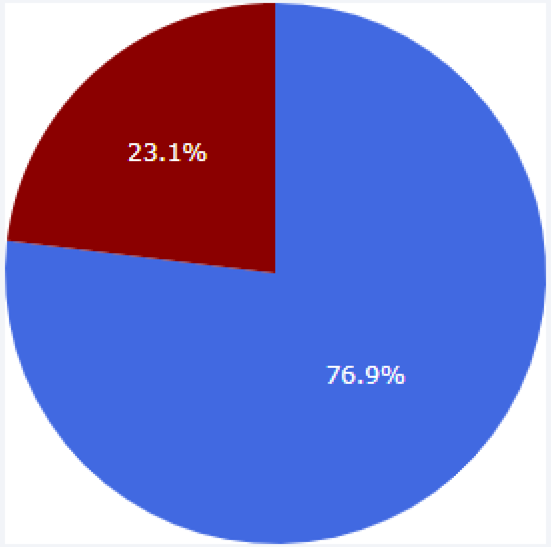
# Successful Launch Across Launch Sites

- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of  CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the  successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



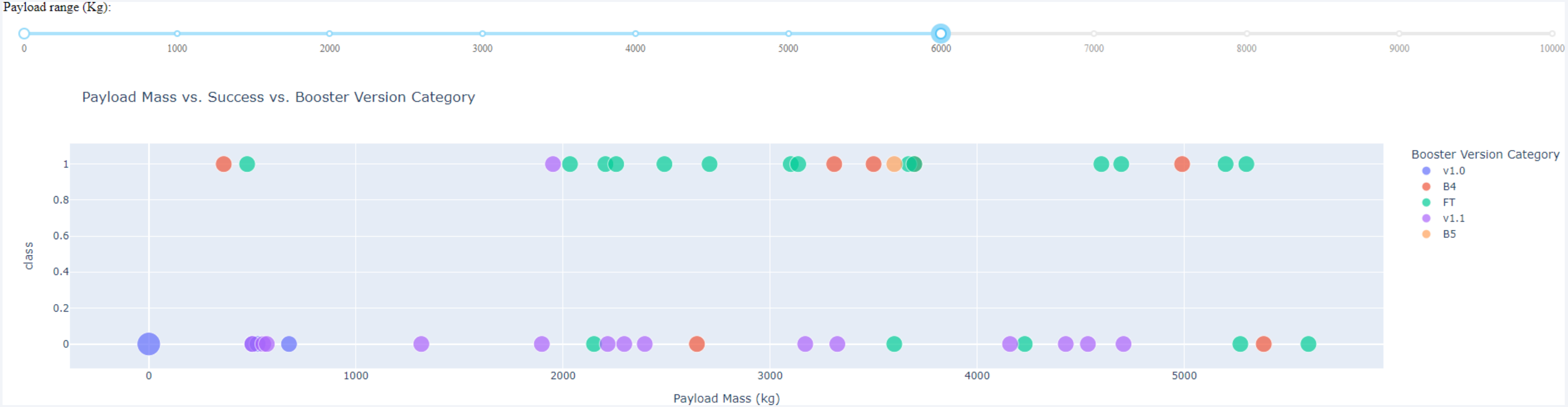
# Highest Success Rate Launch Site

- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.



# Payload Mass vs Success vs Booster

- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.



Payload range (Kg):

Section 5

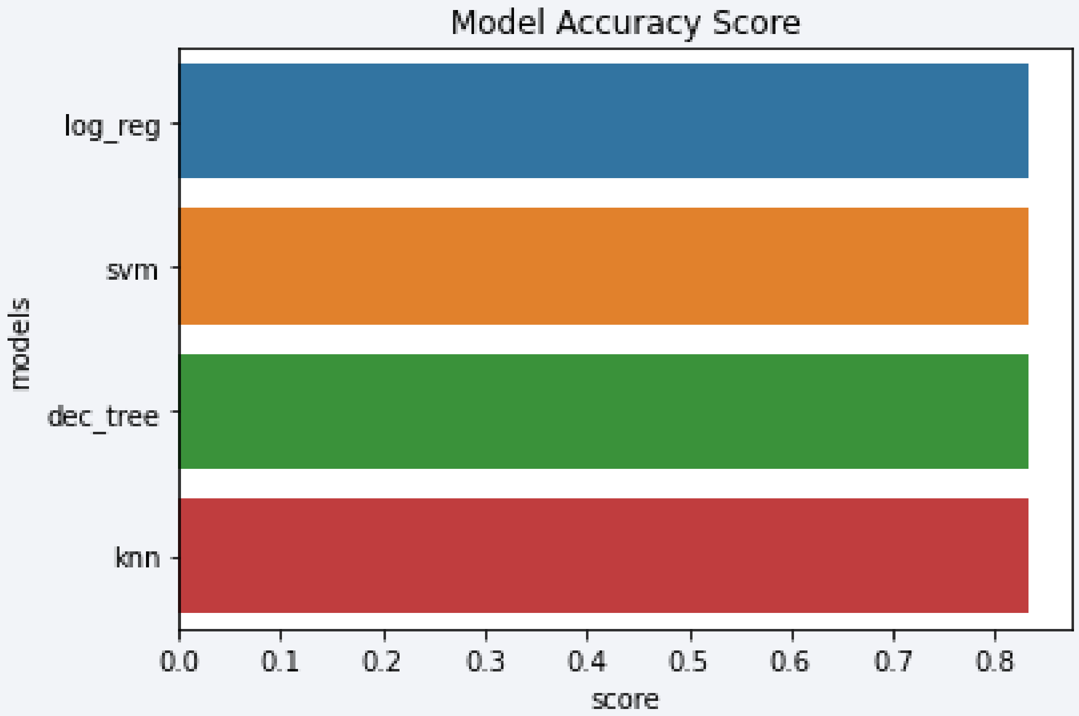
# Predictive Analysis (Classification)

# Classification Accuracy

- All models had virtually the same accuracy on the test set at 83.33%

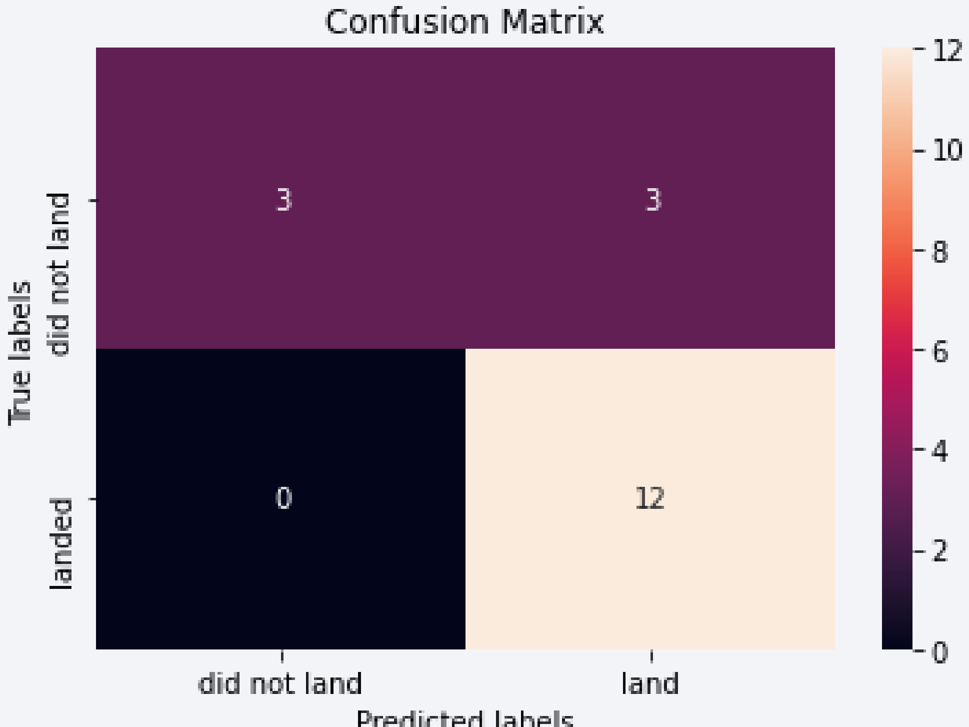
accuracy. It should be noted that test size is small at only sample size of 18.

- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.



# Confusion Matrix

- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



# Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- Github: https://github.com/TaKMooN/Capstone-Projects

# Thank you!