
# Winning Space Race with Data Science

Moein
2023-03-05

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix (for other relevant pictures and details)

# Executive Summary

3

- Summary of methodologies
  - Data Collection: I gathered data from SpaceX API and web scraping to collect the necessary data for my project.
  - Data Wrangling: I cleaned and pre-processed the collected data to ensure it was in a suitable format for further analysis.
  - Exploratory Data Analysis: I explored the data using various libraries like matplotlib, seaborn, and folium to gain insights and identify patterns or trends.
  - Data Visualization: I used data visualization techniques to present the results of your EDA in a clear and concise manner.
  - Model Development: I developed models using scikit-learn library and implemented different algorithms like SVM, Logistic Regression, Decision Tree, and KNN to predict the first stage of the SpaceX Falcon 9 rocket will land successfully or not.
  - Model Evaluation: I evaluated the performance of the developed models using appropriate metrics and techniques to select the best model.
  - Reporting Results: Finally, I reported my findings and results to the stakeholders in a clear and concise manner.

# Executive Summary

- Summary of all results
  - After exploring the data using various visualization techniques, it was observed that there were some significant correlations between different variables, such as the payload mass, legs, flight number and the final outcome of the landing. Additionally, it was found that the success rate of the first stage landing in launch sites was around 60-77%, indicating that the Falcon 9 rocket is relatively reliable in this regard.
  - Several machine learning models were trained on the data using various algorithms, including SVM, Logistic Regression, Decision Tree, and KNN. After model evaluation, the Decision Tree model was found to be the best performer, with an accuracy score of around 87-88%.

# Introduction

- Project background and context
  - This project aims to use data science methodologies to predict the success of the Falcon 9 first stage landing. The aerospace industry is a highly complex and challenging field that requires a high degree of precision and reliability in all aspects of its operations. The landing of the first stage of a rocket is a critical part of the launch process, and any failure can result in significant financial and reputational costs for the launch provider.
  - SpaceX has made significant advancements in reusable rocket technology, and accurate predictions of first stage landing success can provide valuable information for space industry stakeholders. Furthermore, new startups entering the industry need to develop innovative approaches and technologies to compete with established players like SpaceX. Data science methodologies can provide a valuable toolset for these startups to gain insights and develop predictive capabilities to inform their decision-making.
  - To address these challenges, we collected data from the SpaceX API and web scraping and performed exploratory data analysis using Matplotlib, Seaborn, Folium, and other visualization tools. We developed and evaluated machine learning models using the scikit-learn library, including SVM, logistic regression, decision trees, and KNN, to find the best performing model.
  - Our goal is to demonstrate the potential of data science methodologies in the space industry and provide valuable insights for stakeholders. Through our work, we hope to contribute to the development of innovative approaches and technologies and help drive growth and innovation in the aerospace industry.

# Introduction

Problems you want to find answers:

The goal of this project is to predict the success of the Falcon 9 first stage landing using data science methodologies. The landing of the first stage of a rocket is a critical part of the launch process, and accurate predictions can provide valuable information for space industry stakeholders. By predicting the success of the landing, we can determine the potential cost of a launch, which is a significant factor in the industry.

Through our work, we aim to answer the following questions:

1. Can we predict the success of the Falcon 9 first stage landing using data science methodologies?
2. What factors have the most significant impact on the first stage landing success?
3. How accurate are our predictive models, and how can we improve their performance?
4. How can our predictions help inform decision-making and drive growth and innovation in the aerospace industry?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - We collected data for this project from two sources: the SpaceX API and web scraping. We used the API to collect data on Falcon 9 launches from 2010 to 2021, including information such as the launch site, payload mass, and landing outcomes and etc.
- Perform data wrangling
  - We wrangled data using pandas and matplotlib library of python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We build, tune, evaluate classification models using scikit-learn library in python

# Data Collection

- We collected data from Spacex API and WebScraping. our API records contained data from 2010-04-06 to 2020-05-11 and webscraping records contained data from 2010-04-05 to 2021-06-06.
- There is also a few difference between data that collected from API and WebScraping and because of some limitation we use API data records for train and feature engineering
- For collecting data we use some of python library such as requests, BeautifulSoup and pandas

# Data Collection – SpaceX API

- A flowchart about data collection with SpaceX REST calls
- GitHub URL

1. Request and parse the SpaceX launch data using the GET request.
2. Keep some features that we want to use
3. For each launch in the list, extract the launch date, booster name, mass of the payload and the orbit.
4. Make a separate GET request to the SpaceX API for each flight number to retrieve more detailed information on the launch, including the rocket components and landing outcome.
5. Extract the relevant information from the response, including the status of the first stage landing (success or failure).
6. Store the extracted data in a structured format for analysis.

# Data Collection - Scraping

- A flowchart about data collection with Scraping Wikipedia web page
- GitHub URL

1. Make a GET request to Wikipedia address
2. Finding the header of table
3. Collect data of each row of table
4. Cleaning the data of each row and store them in pandas Dataframe

# Data Wrangling

We create a new column name "class" to demonstrate failure and success for each launch in numeric way

FlowChart

- We import dataset that we have been collected before "dataset_part_1.csv" to a pandas dataframe
- Count value of "Outcome" column using "value_counts()"
- We create a "bad_outcomes" set of outcomes where the second stage did not land successfully
- We Create a list of [0, 1] that each row contained "bad_outcomes" elements we assign it to 0 in class column and vice versa

GitHub URL

# EDA with Data Visualization

- There are 5 plot that I listed below:
  - Payload mass(Kg) vs Flight number: we want to see how the Flight Number (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome.
  - Flight number and Launch site: we want to see relation between flight number and launch site and how would affect the launch outcome.
  - Payload mass(Kg) and Launch sites: to check there is any relationship between launch sites and their payload mass.
  - Success rate of each Orbit: to check there are any relationship between success rate and orbit type.
  - Flight number and Orbit type: to see there is any relationship between Flight Number and Orbit type.
  - Payload Mass(Kg) and Orbit type: we want to research the relationship between Payload and Orbit type
  - Yearly trend of Success rate: we create it to get the average launch success trend.
- GitHub URL

# EDA with SQL

- summarize of the SQL queries
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - Count successful landing
  - Date of first successful landing outcome
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the "booster versions" which have carried the maximum payload mass
  - List the records which will display the month names, failure "landing outcomes" in drone ship ,"booster versions", "launch site" for the months in year 2015.
  - Rank the count of "landing outcomes" between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL

# Build an Interactive Map with Folium

- Objects that have been added to map
  - Circle, Marker, Line, Mouse Position
- Circle: we have been added circle for demonstrate the launch sites
- Marker: we have been added Markers for labeling stuffs
- Line: Lines used for connect proximities around of each launch sites
- Mouse Position: this thing used for finding Latitude and Longitude of each location we want to calculate the distance of proximities location from launch site
- GitHub URL

# Build a Dashboard with Plotly Dash

- Objects that we have added to our dashboard
  - Pie Chart, Scatter Plot, Slide range, Dropdown menu
- Pie Chart: for finding relation between each launch sites and success rate
- Scatter Plot: for finding relation between payload mass and success rate of each booster version category
- GitHub URL
- Note: we used default Spacex dataset due to some limitation such as booster version column and therefore we did not use our wrangled dataset but we can find interesting insights from this datasets and you can find the results in pictures folder in related GitHub URL I mentioned above.

# Predictive Analysis (Classification)

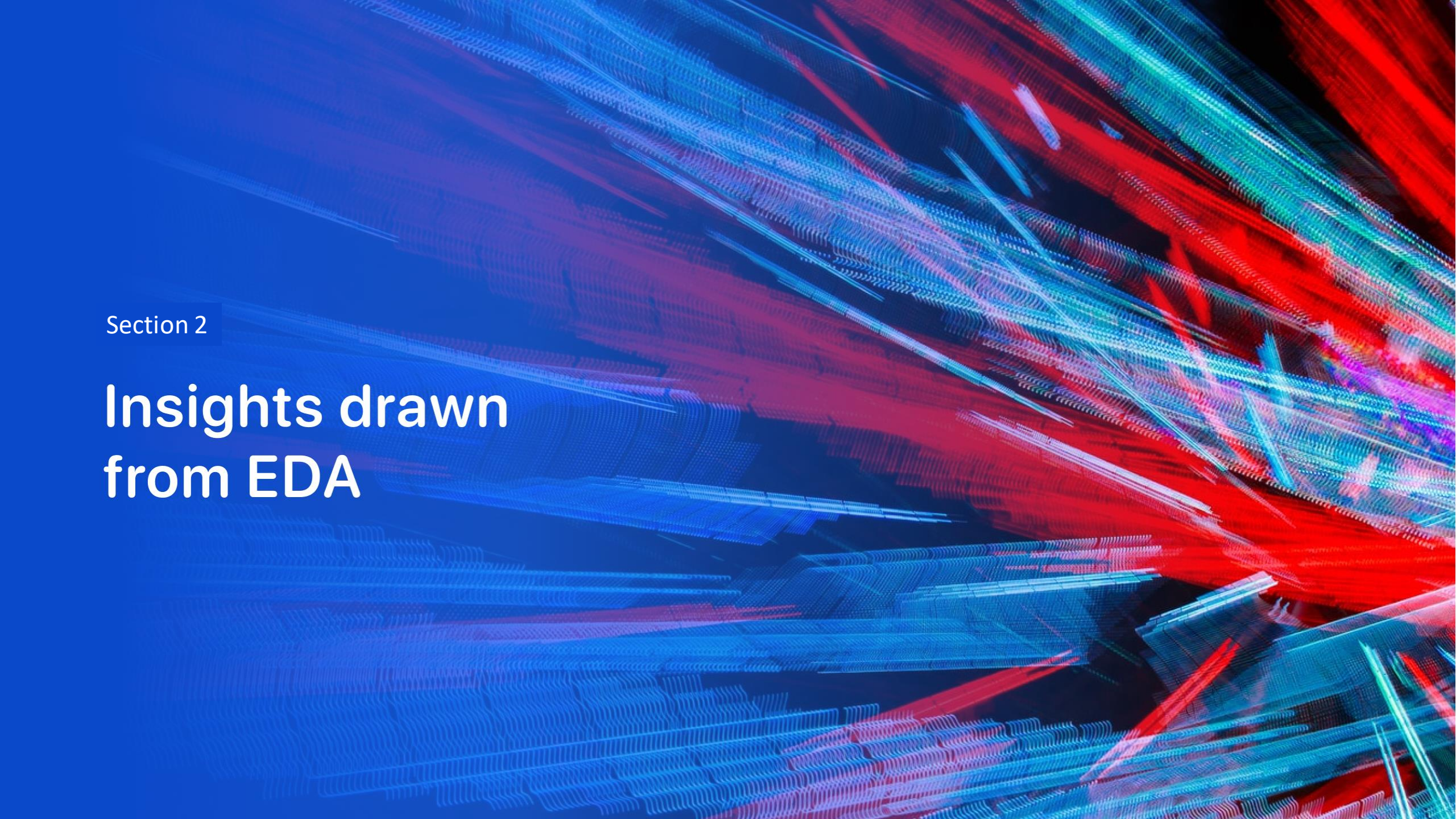
- Here the steps of predictive analysis
  - We built model with our pre-processed and wrangled dataset "dataset_part_3.csv".
  - We evaluated our model with GridSearchCV method that use a part of train dataset for validation and also we used test dataset for evaluation
  - We also used GridSearchCV method and pass multi-parameters that used for tuning model and improve it during training
  - We found the best model using comparing models performance on validation dataset and test dataset
- Flow Chart of model building
  - Use "dataset_part_2.csv" for retrieving the class column
  - Use "train_test_split" method for create train and test dataset
  - Create model from related object in scikit-learn library
  - Train and evaluate models using GridSearchCV method
  - Test models on test dataset and Comparing models using score method, Confusion matrix and Bar plot
- GitHub URL

# Results

- Exploratory data analysis results
  - Based on exploratory data analysis There are relation between several factors and success rate of landing first stage of our rocket and we can use them for build and train a machine learning model
- Interactive analytics
  - With this we found that there are relation between payload mass, booster version category and success of first stage landing.
  - There is also relation between proximities of each launch site to cities, railroad, highway, coastline and success of launch
- Predictive analysis results
  - As we saw, we predicted the outcome with 87-88% of correct prediction

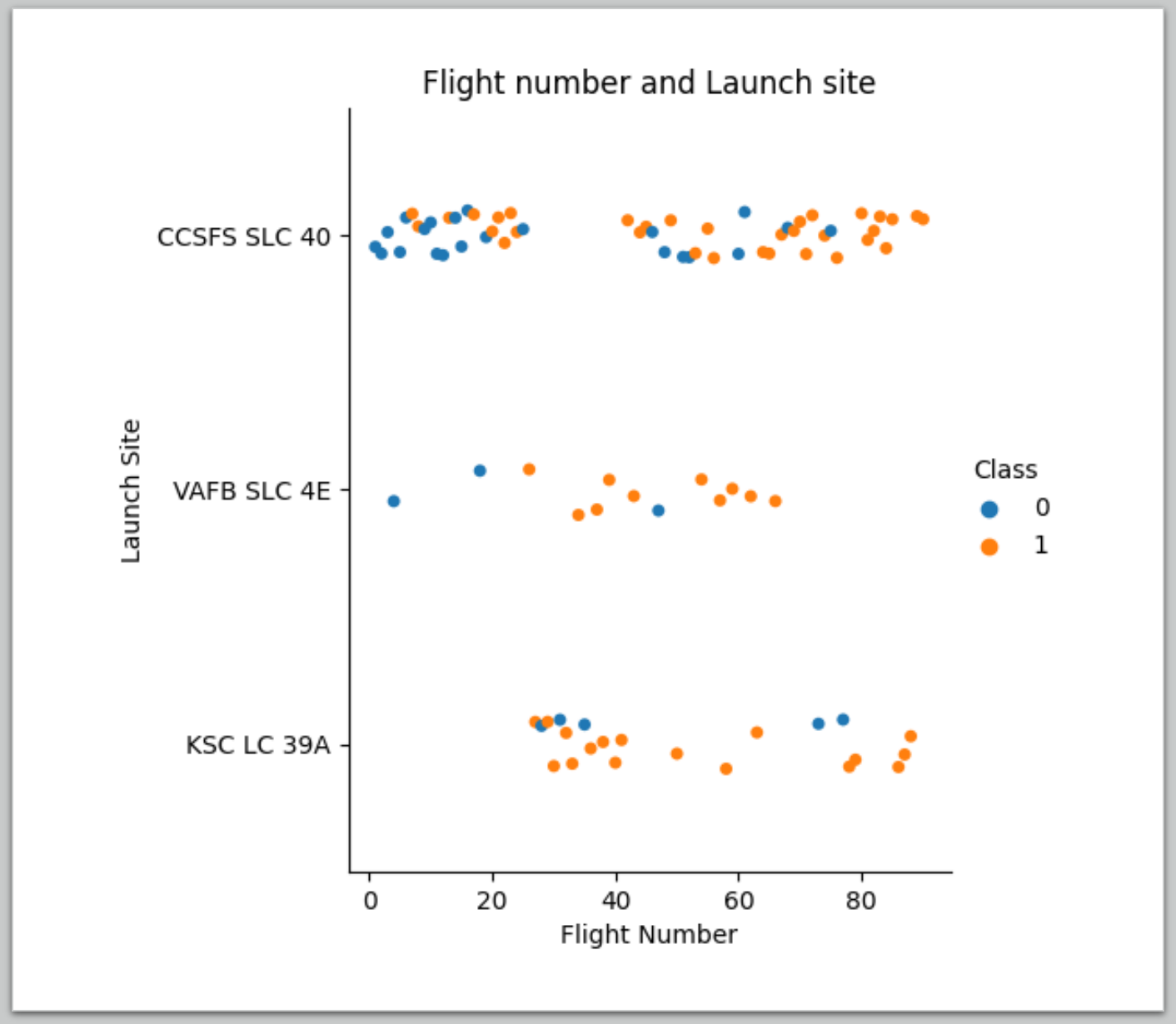
Section 2

# Insights drawn from EDA



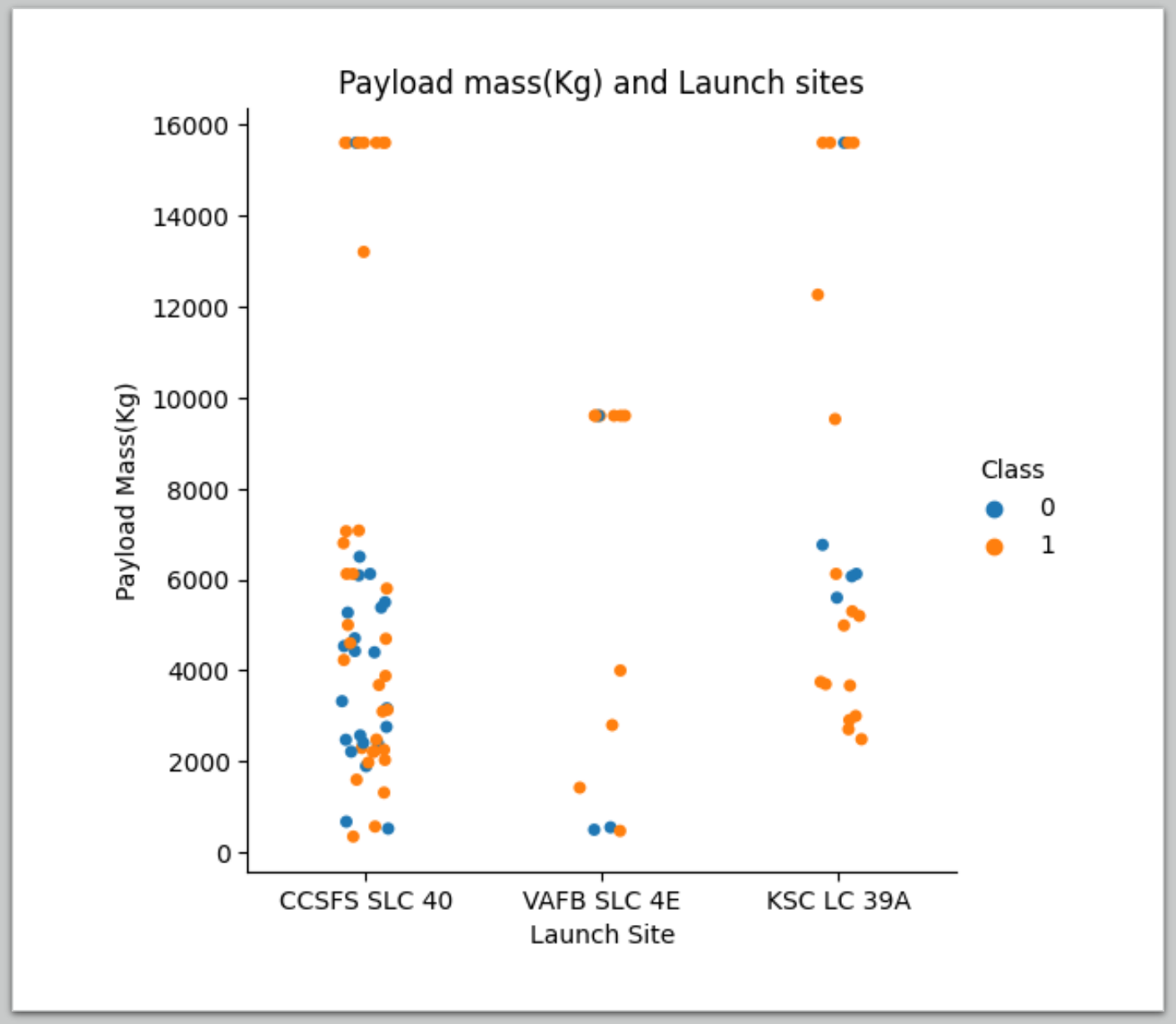
# Flight Number vs. Launch Site

- We can see as well as flight number increase success rate in each launch site also increase.



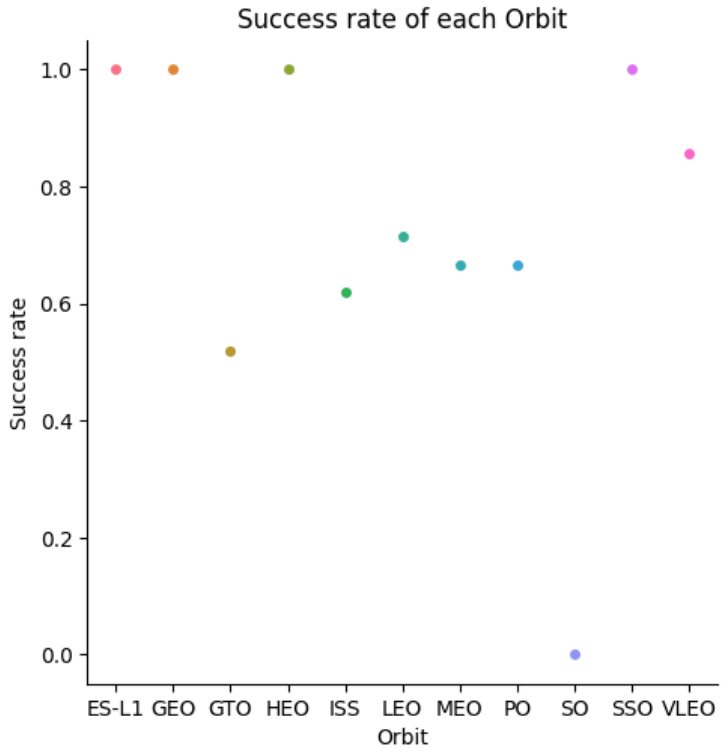
# Payload vs. Launch Site

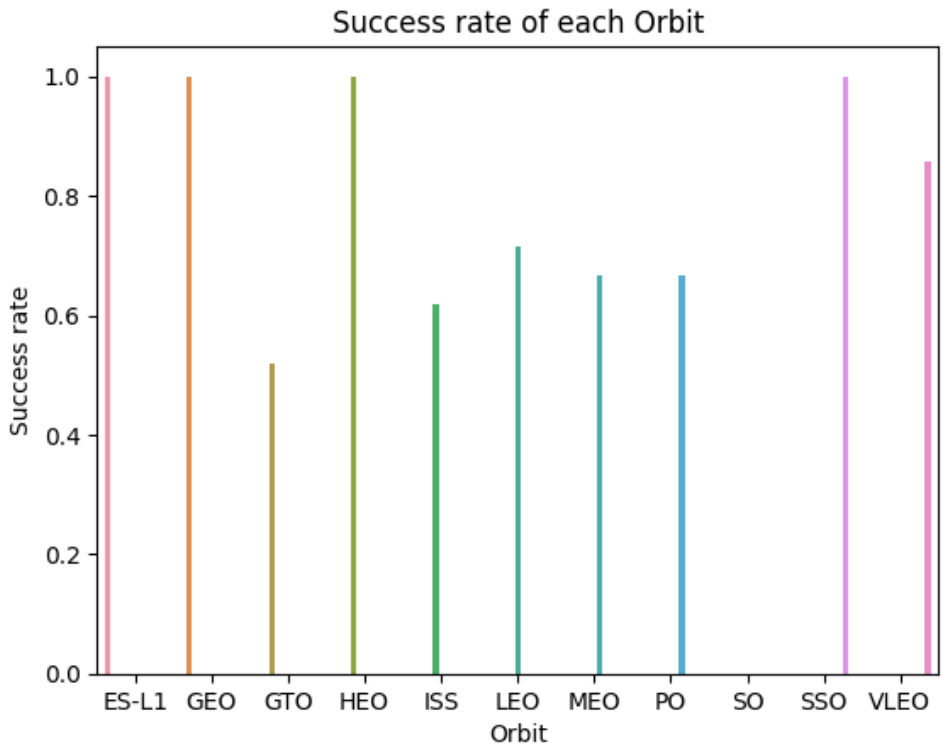
- if we observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000) and one of launch site was good an low payload mass and other one was good at heavy payload mass.



# Success Rate vs. Orbit Type

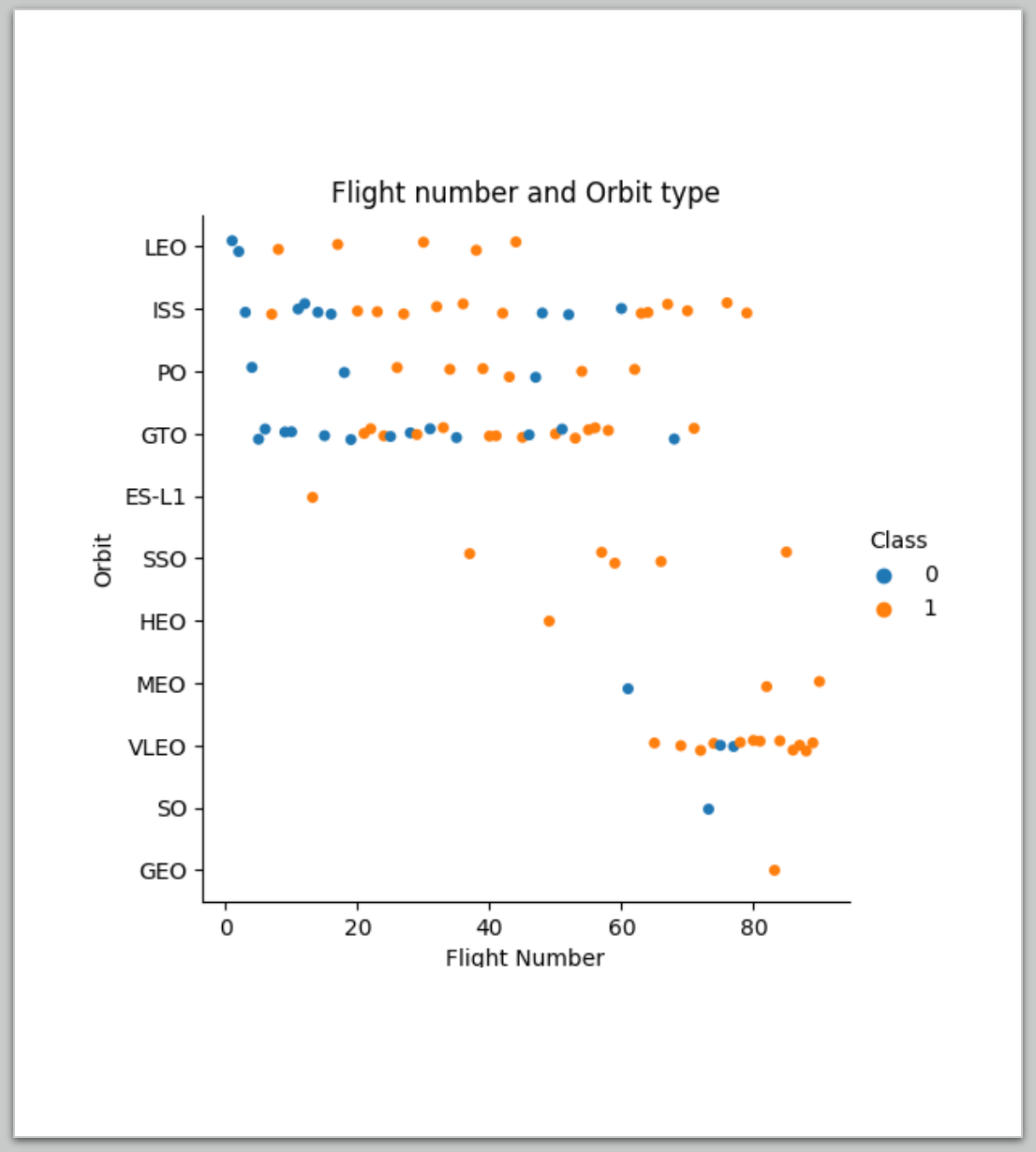
- As we can there are a few orbits with high success rate, so we can find out that there is relation between success rate and Orbit type



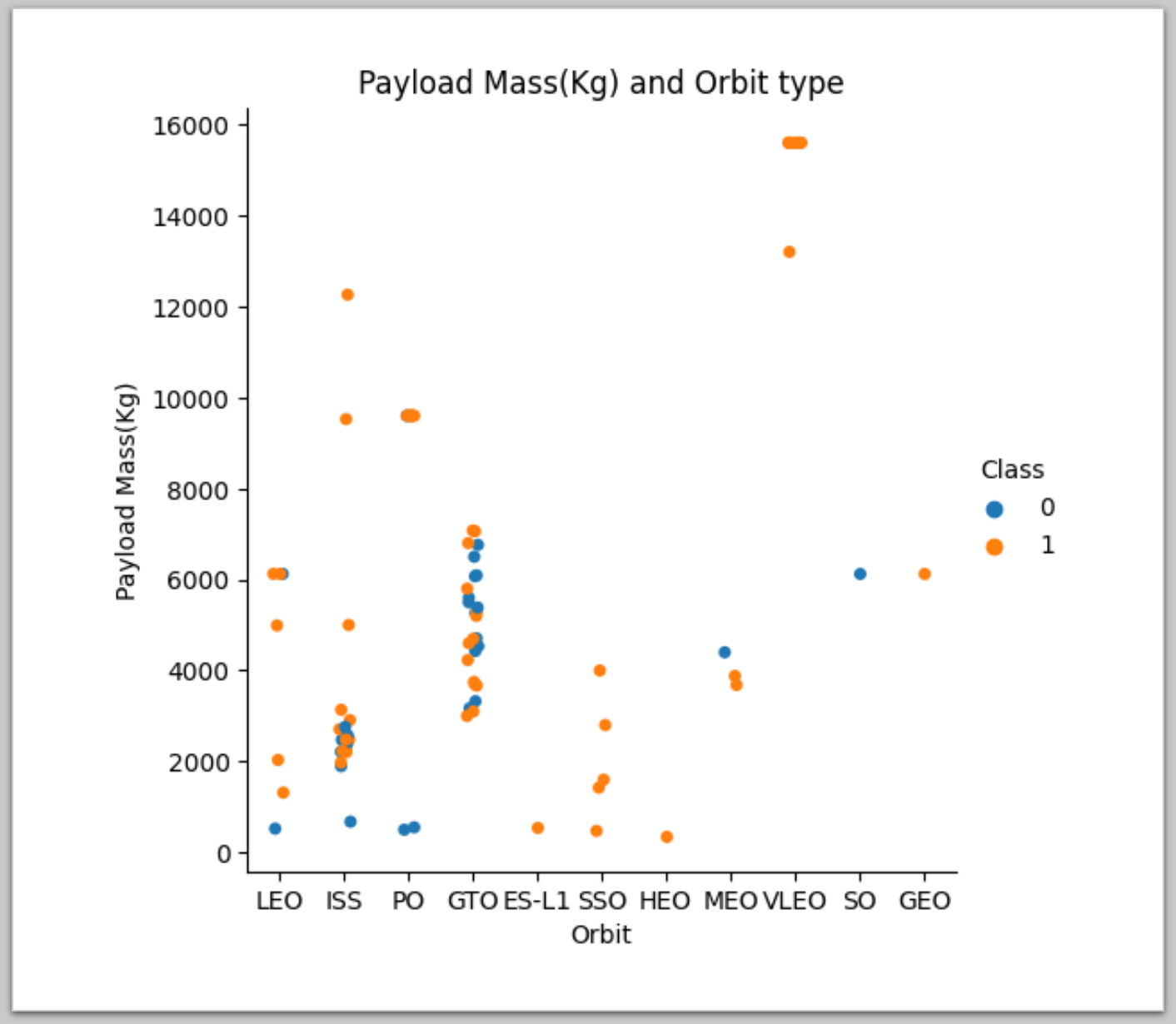
# Flight Number vs. Orbit Type

- 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.



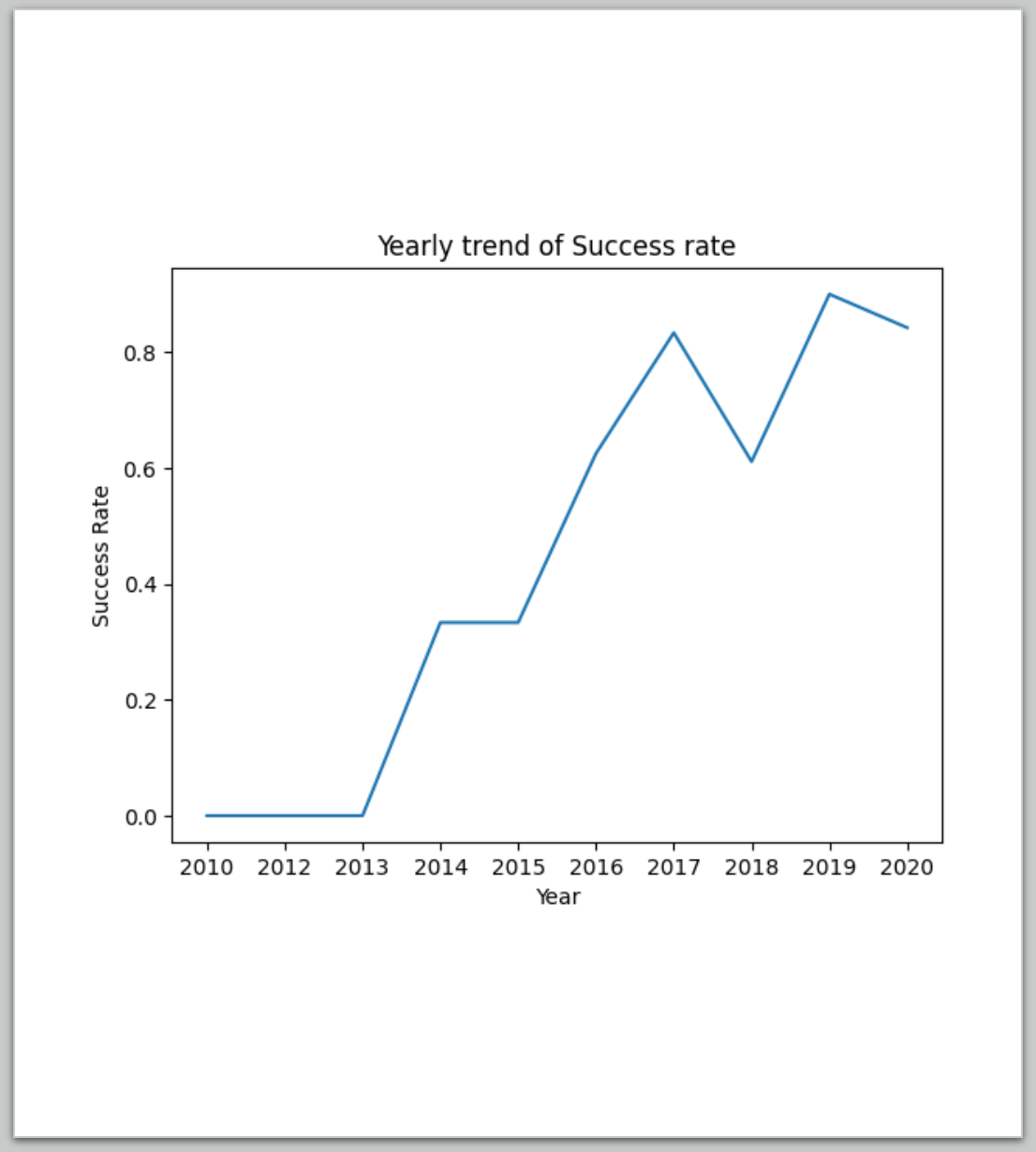
# Payload vs. Orbit Type

As we can see with heavy payloads the successful landing or positive landing rate are more for Polar,LEO,VLEO 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

As we can observe that the success rate since 2013 kept increasing till 2020



# All Launch Site Names

- Launch site names:
  - CCSFS SLC 40: success rate in this launch site is 60%
  - KSC LC 39A: success rate in this launch site is 77%
  - VAFB SLC 4E: success rate in this launch site is 76%
- We get these launch site names from "dataset_part_2.csv" that is our data we collected form Spacex API and not collected from prepared dataset that we used in Dash interactive dashboard.
- Note: for consistency of our project we used prepared dataset in Dash interactive dashboard and SQL exploratory data analysis
- More details in appendix

# Launch Site Names Begin with 'CCA'

- We find first 5 records where launch sites begin with `CCA` and that launch site was "CCAFS LC-40" which show us that first 5 launch at began use this launch site where mission outcome all of them was success and almost NASA was customer of them

# Total Payload Mass

- the total payload carried by boosters from NASA = 45596 Kg
- This show us that NASA boosters can't carried heavy payload mass

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 = 2928.4 Kg
- This show us That this booster version carried payload with this mass in average

# First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad = 22-12-2015
- After 5 years we can see we have first successful landing outcome on ground pad

# Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters and payload mass are shown in table
- As it is obvious we can use this boosters between this payload range for get better chance to have success landing outcome in "drone ship"

| BOOSTER VERSION | PAYLOAD MASS (KG) |
|---|---|
| F9 FT B1022 | 4696 |
| F9 FT B1026 | 4600 |
| F9 FT B1021.2 | 5300 |
| F9 FT B1031.2 | 5200 |

# Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes shown in table
- In 101 mission outcome, Spacex could got an optimal situation for launching rockets and get better result.

| Count |
| --- |
| 101 |

| 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 |

# Boosters Carried Maximum Payload

- The booster which carried maximum payload shown in the table
- We can see that "B5 B10..." Used for carried heavy payload mass.

# 2015 Launch Records

- Table show us list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- From all launch in 2015 we have 2 failure landing outcome, and we can see in 2015 they used V1.1 booster version and CCAFS LC-40 launch site

| month | year | Landing Outcome | Booster Version | Launch Site | Mission Outcome |
|---|---|---|---|---|---|
| 01 | 2015 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 | Success |
| 04 | 2015 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 | Success |

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

- You can see rank of 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

- we can see at the end of 2015 and above we have more successful landing, and also we can see in 2012 we have 10 not attempt landing outcome

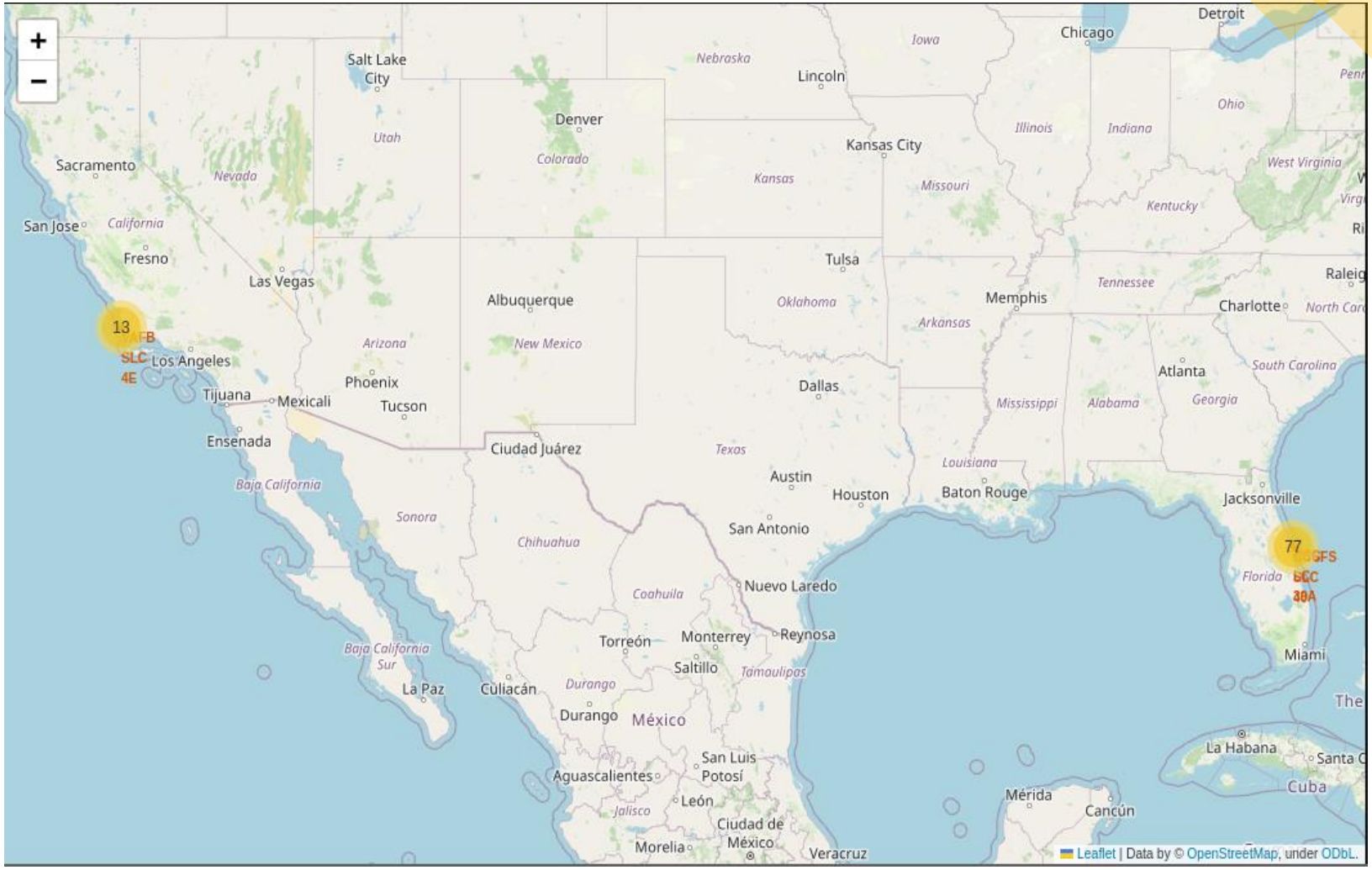
| Date | Landing Outcome | Count |
|---|---|---|
| 22-05-2012 | No attempt | 10 |
| 08-04-2016 | Success (drone ship) | 5 |
| 10-01-2015 | Failure (drone ship) | 5 |
| 22-12-2015 | Success (ground pad) | 3 |
| 18-04-2014 | Controlled (ocean) | 3 |
| 29-09-2013 | Uncontrolled (ocean) | 2 |
| 04-06-2010 | Failure (parachute) | 2 |
| 28-06-2015 | Precluded (drone ship) | 1 |


Section 3
Launch Sites
Proximities Analysis

# Launch sites Location

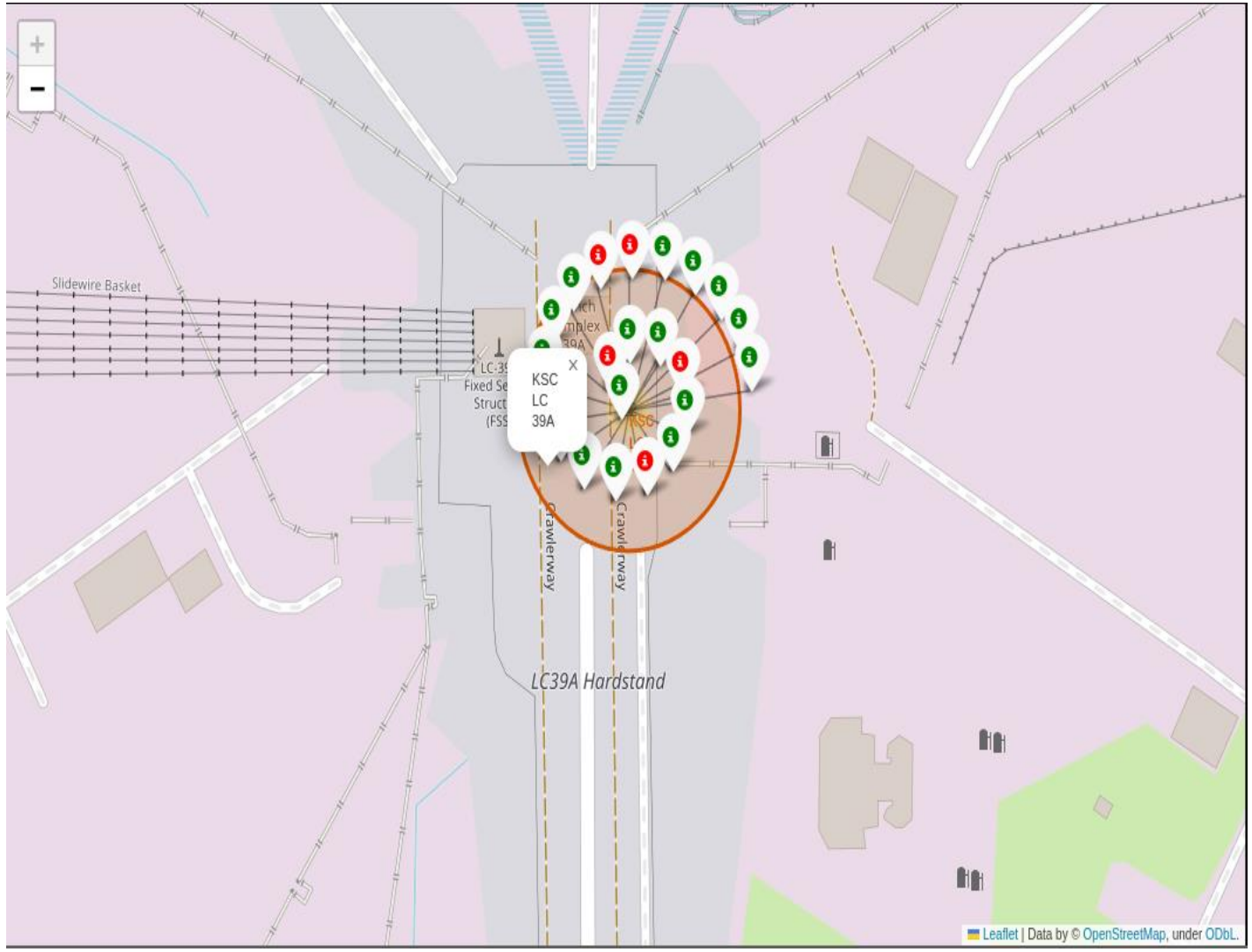
We can see:

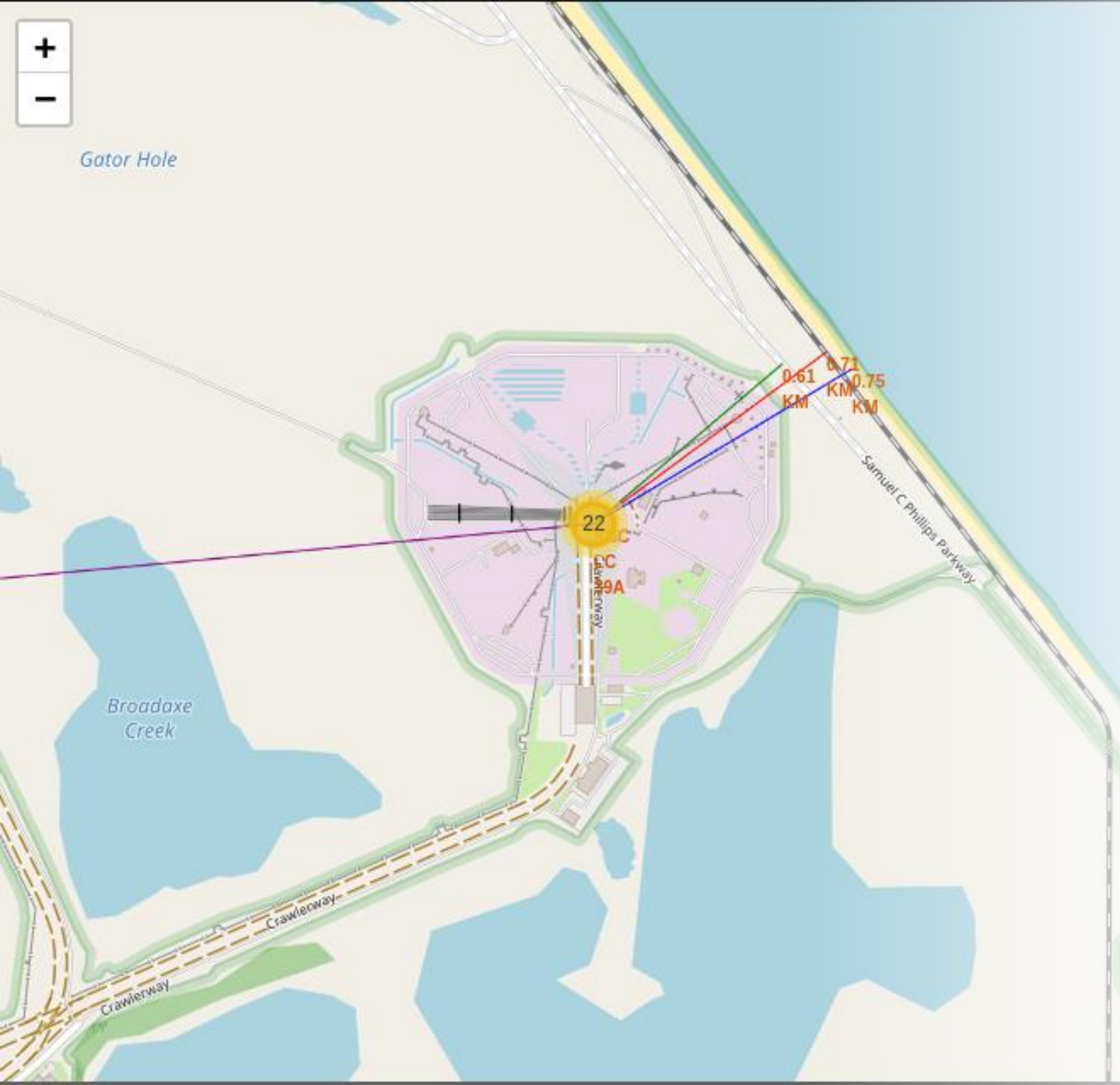
- all launch sites were beside coastline
- Below the equator line
- Most of launches were from east side of US



# Launch Sites Outcomes

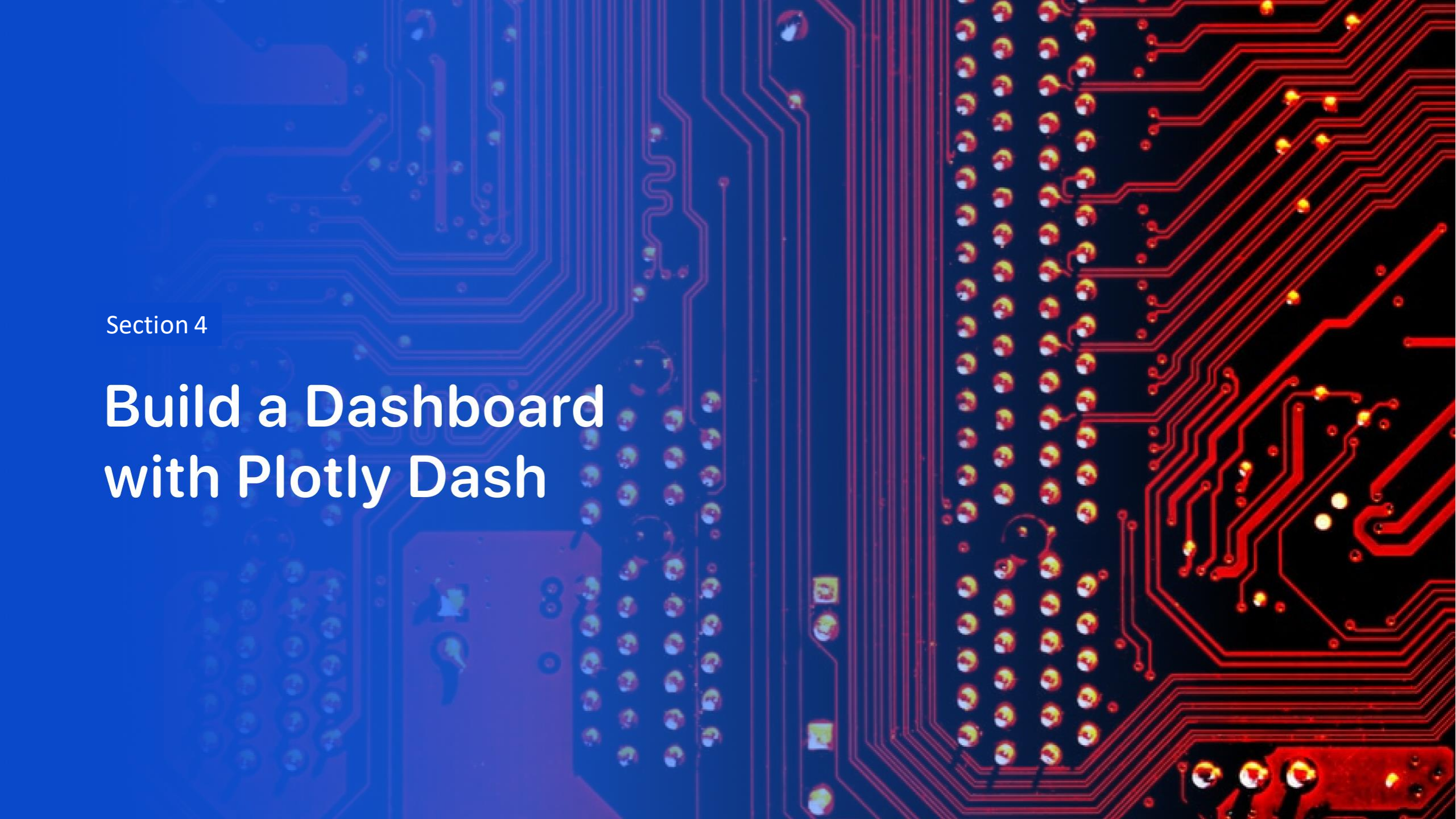
- We labeled outcomes of all launch sites with "green" for successful outcome and "red" for failed.
- We can see in "KSC LC 39A" we have had 5 failed and 17 successful launches.
- With this we can show which launch sites have greater success rate






# Launch site proximities

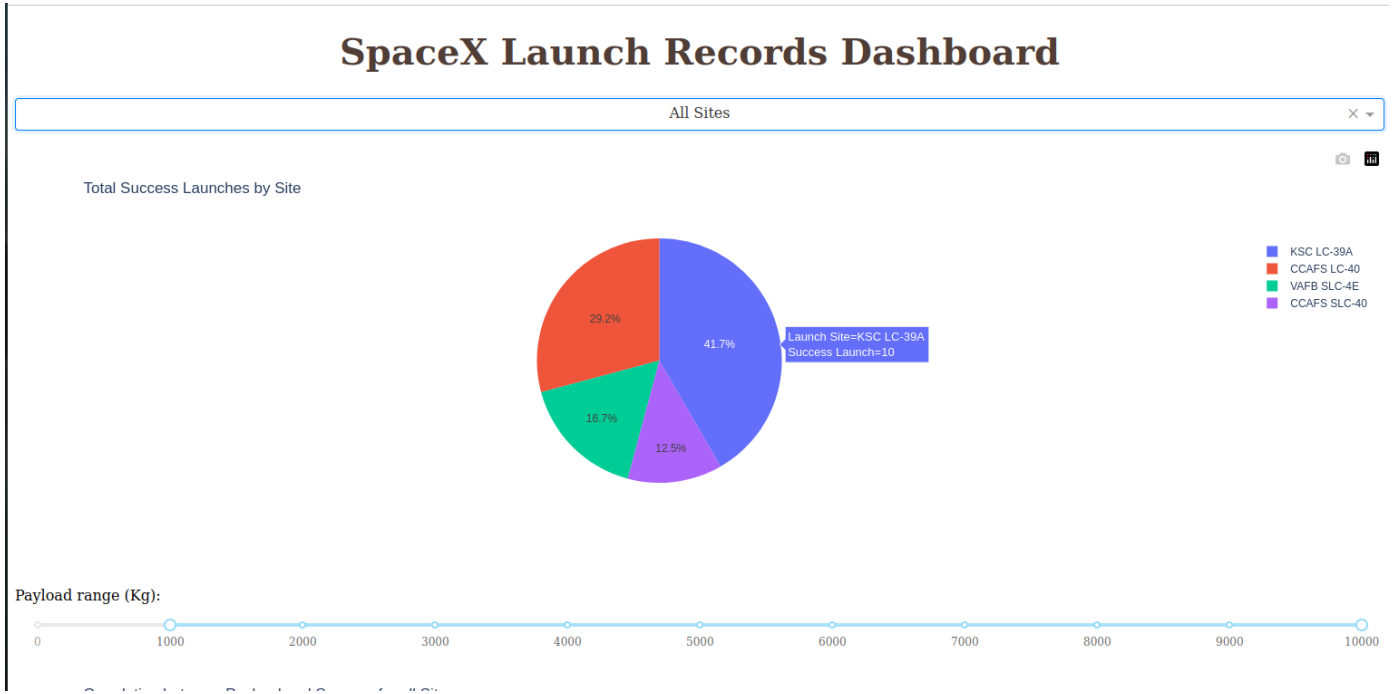
- We used lines for connect each launch site with closest city, railway, coastline and highway. With this we can find that how this distance affect success rate
  - We compared launch sites and figured it out that best distance for closest city, railway, coastline and highway.
    - City = 15-22 km
    - Highway = 0.61-1.8 Km
    - Railway = 0.71-1.30 Km
    - Coastline = 0.75-1.37 Km
- It is better to have Coastline, Railway, Highway in one side and city in another side of launch site


Section 4
Build a Dashboard with Plotly Dash

# Distribution of success rate

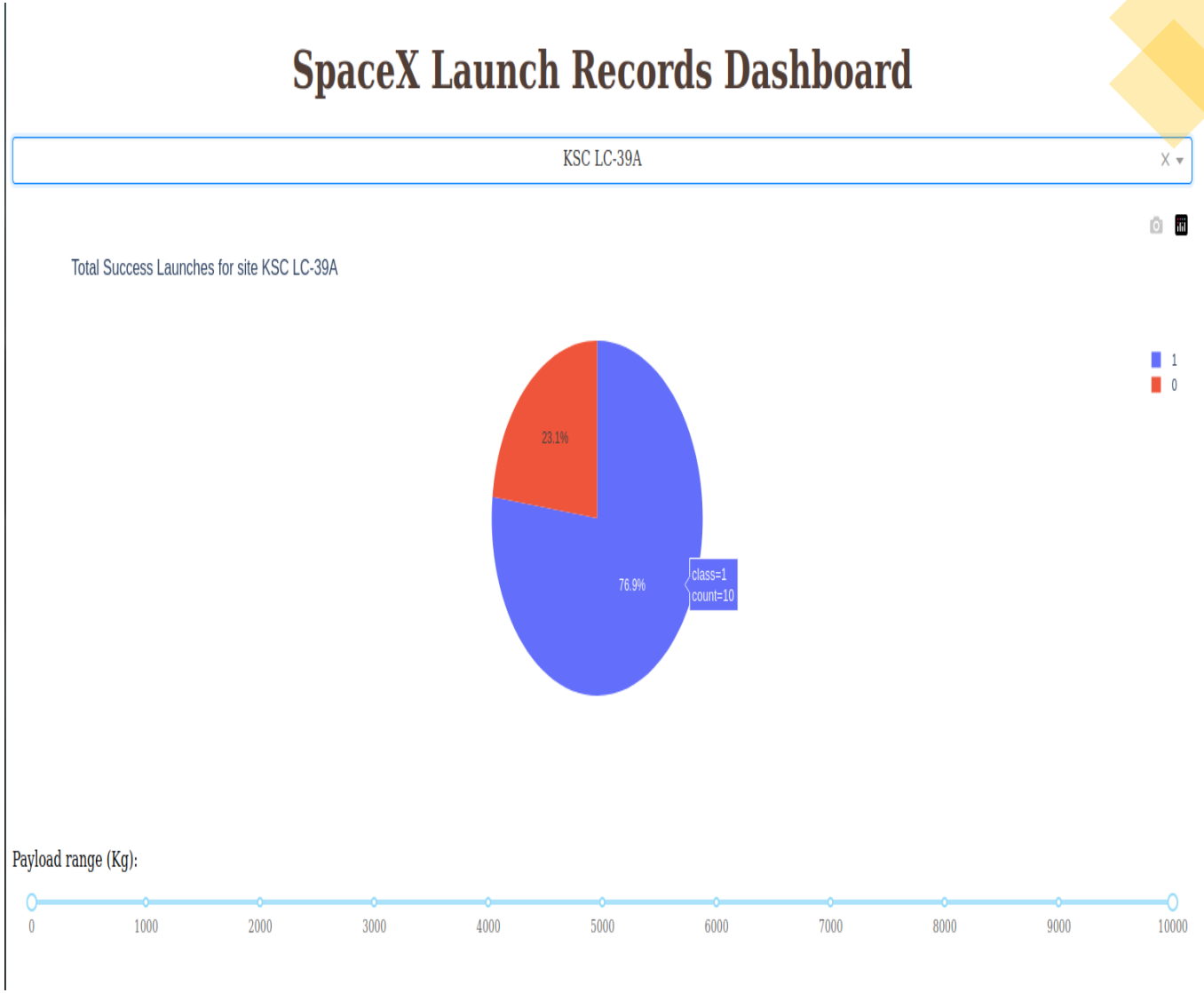
In this Pie chart we can see that "KSC LC-39A" launch site has highest success rate than other launch sites with 41.7%.

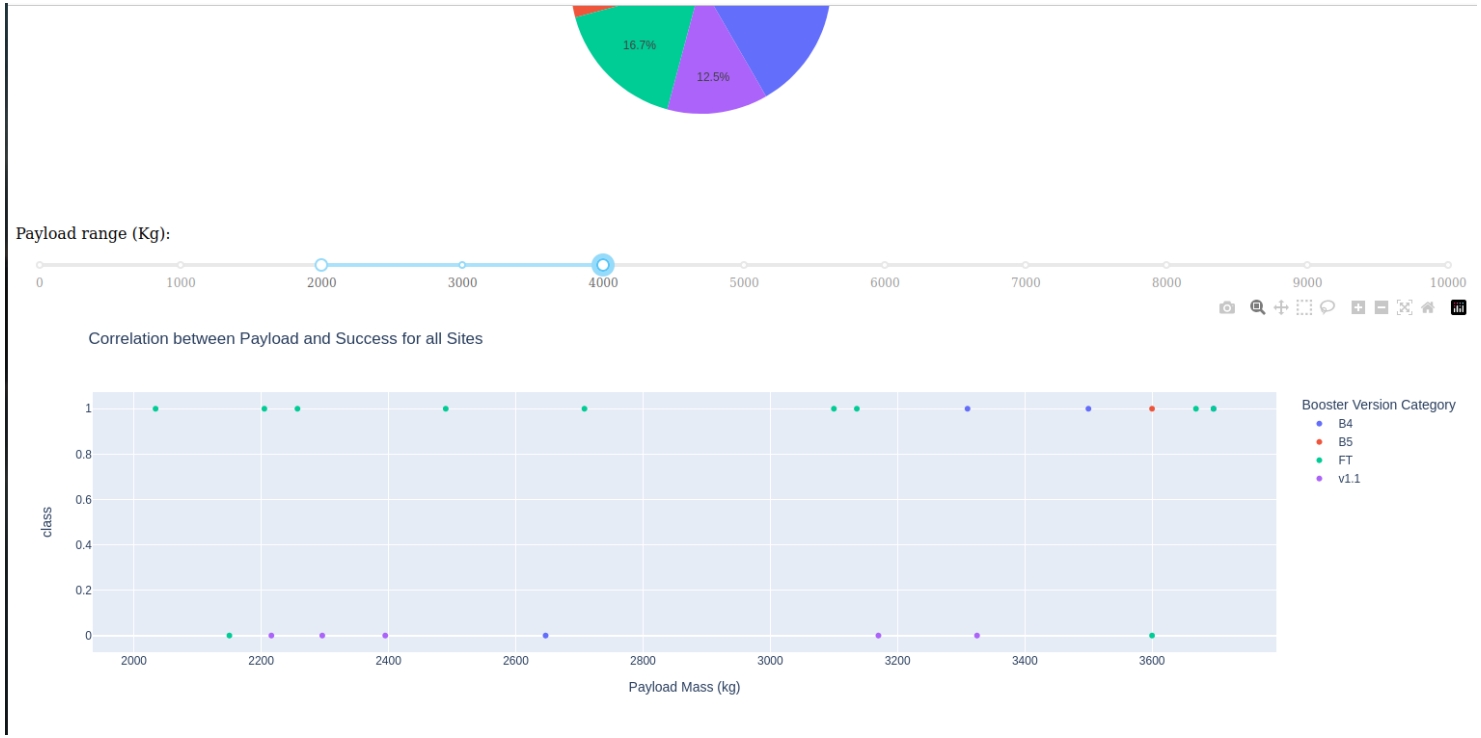
- Note: we use prepared dataset for this analysis.



# highest success rate

We can see "KSC LC-39A"  has the highest success rate ratio with 76.9% of success from all launches in this site.



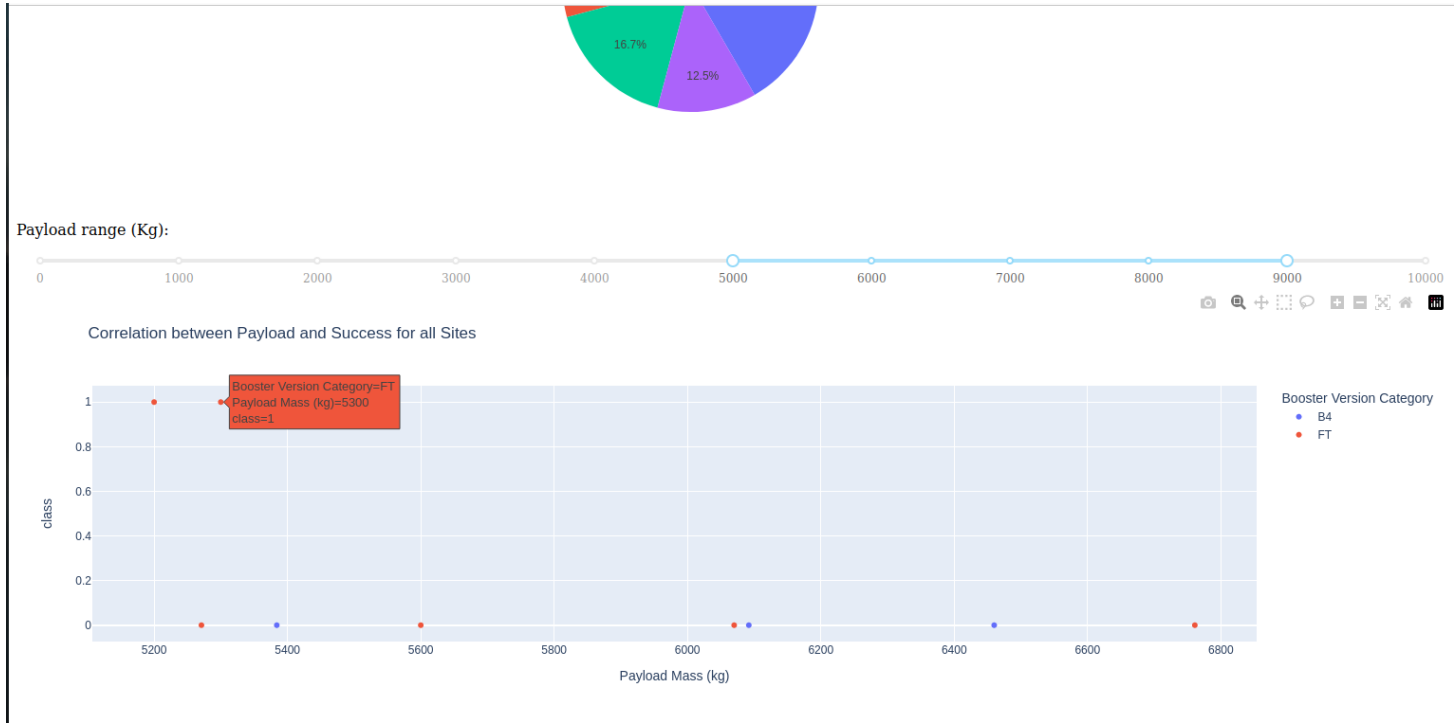



# Payload ranges with the highest launch success rate

- We can see we have largest number of highest launch success rate in range 2000-4000 Kg

# Payload ranges with the lowest launch success rate

- We can see we have largest number of lowest launch success rate in range 5000-9000 Kg
- Other ranges listed below:
  - 0-1000 Kg
  - 4000-6000 Kg
- Note: we added other pictures in appendix



# F9 booster version

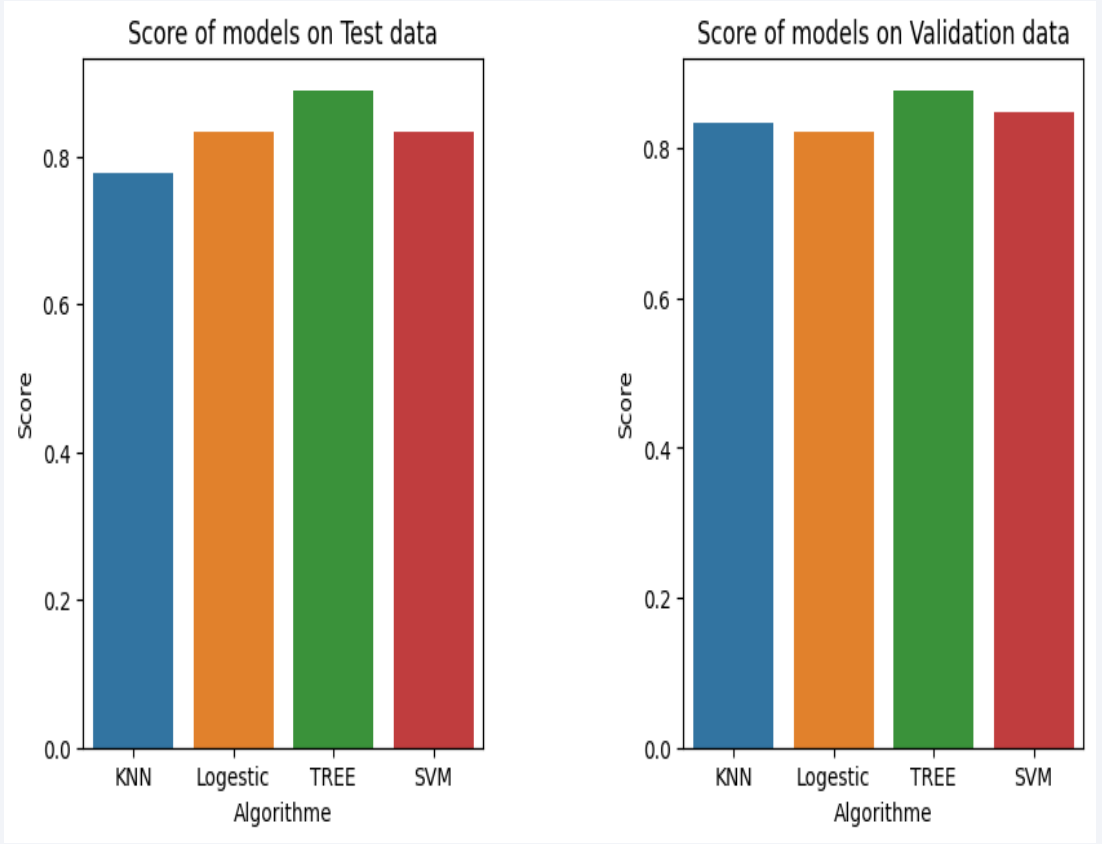
- In picture we can see "FT" version have largest number of highest launch success rate in all range of payload mass
- There are booster versions with highest and lowest success rate that we listed below:
  - Highest: *1- B5 2- FT
  - Lowest: 1- V1.0 2- V1.1
- *B5 only had one launch that was successful, so we can't say that this booster version can be in highest success rate group.
- Note: we added other pictures in appendix

Section 5

# Predictive Analysis (Classification)

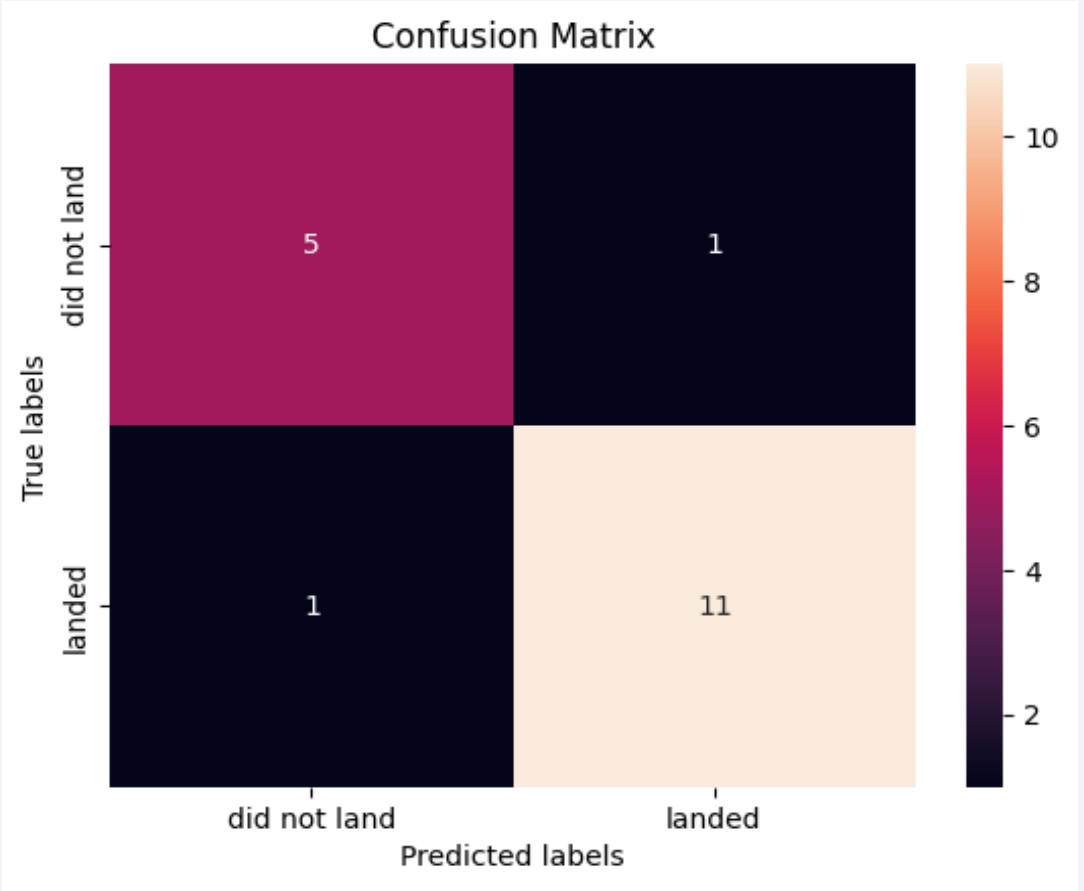
# Classification Accuracy

- In this bar chart we can see "Decision Tree classification" algorithm had best performance on validation and test dataset.



# Confusion Matrix of Decision Tree Classifier

- Here we show result of tree classifier in test data:
  - 11 of 12 landed records predicted correctly
  - 5 of 6 did not land records predicted correctly



# Conclusions

- Our analysis demonstrates that it is possible to predict with reasonable accuracy whether the first stage of the SpaceX Falcon 9 rocket will land successfully. This information is valuable for companies bidding against SpaceX for a rocket launch, as they can estimate the cost of their own launches more accurately.
- Our predictive models were developed using a range of machine learning algorithms, including SVM, logistic regression, decision trees, and KNN. After evaluating the models based on various metrics, we found that the decision tree model performed best for our specific problem.
- Our analysis highlights several factors that contribute to the success of the Falcon 9 first stage landing, including Orbit, launch site location, Payload Mass and the specific configuration of the rocket. These findings could inform future improvements to the Falcon 9 rocket design or launch procedures.
- There are also some factors that make noise for our prediction such as the weather conditions at the launch site may not be fully captured in our dataset, or there may be discrepancies in the reported landing outcome.
- Overall, our project demonstrates the value of applying the data science methodology to a real-world problem, using a combination of data collection, data cleaning and preprocessing, exploratory data analysis, machine learning, and data visualization techniques. The results of our analysis can help inform decision-making and improve efficiency in the commercial space industry.
- Finally, we acknowledge that our analysis has limitations, including the small sample size of Falcon 9 launches and the potential for bias in the data. Further research could expand on our findings and explore additional factors that influence the success of rocket launches.

# Appendix

- You can find pictures of our analysis in Interactive dashboard for relation between payload mass range and booster version with success rate in github url
- Exploratory data analysis with SQL in Github
- Click on Github to go to project url on github

# Thank you!