IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Almahdi Mohmed Ali
2024-02-25

Outline

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

Executive Summary

Methodology:

- Data collection with web scraping and SpaceX API
- Exploratory data analysis including data wrangling and visualization
- Building predictive models and testing it

Introduction

- 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.
- The objective is to determine whether the first stage will land successfully or not and determine how much it will cost depending on historical data.

Section 1

Methodology

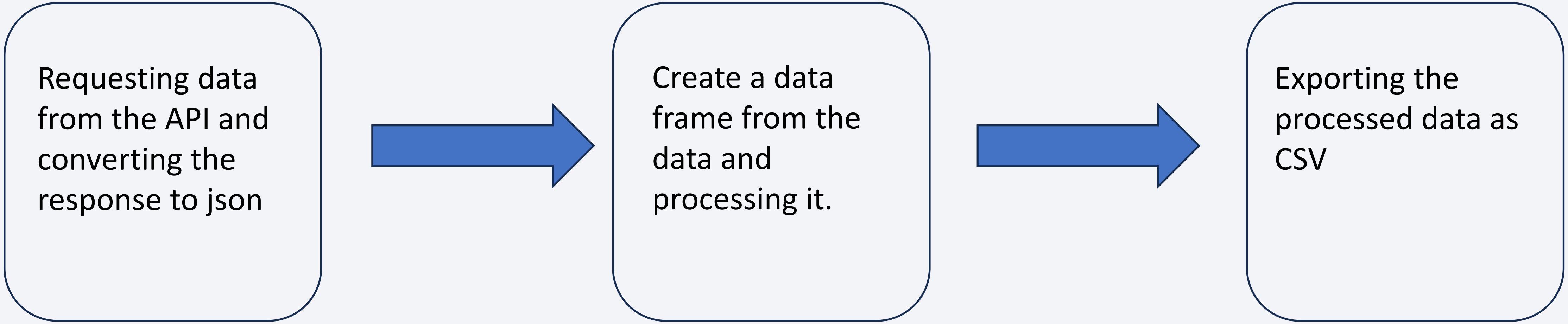
Methodology

Executive Summary

- Data collection methodology:
  - Data was obtained from SpaceX API via web scraping
- Perform data wrangling
  - During data exploration, rows and columns with too many missing values were removed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected data was standardized ,split into training and testing sets and evaluated by four different classification models. Best hyperparameters were taken

Data Collection

- Data was collected from SpaceX API(https://api.spacexdata.com/v4/rockets/) and from a table in wikipedia



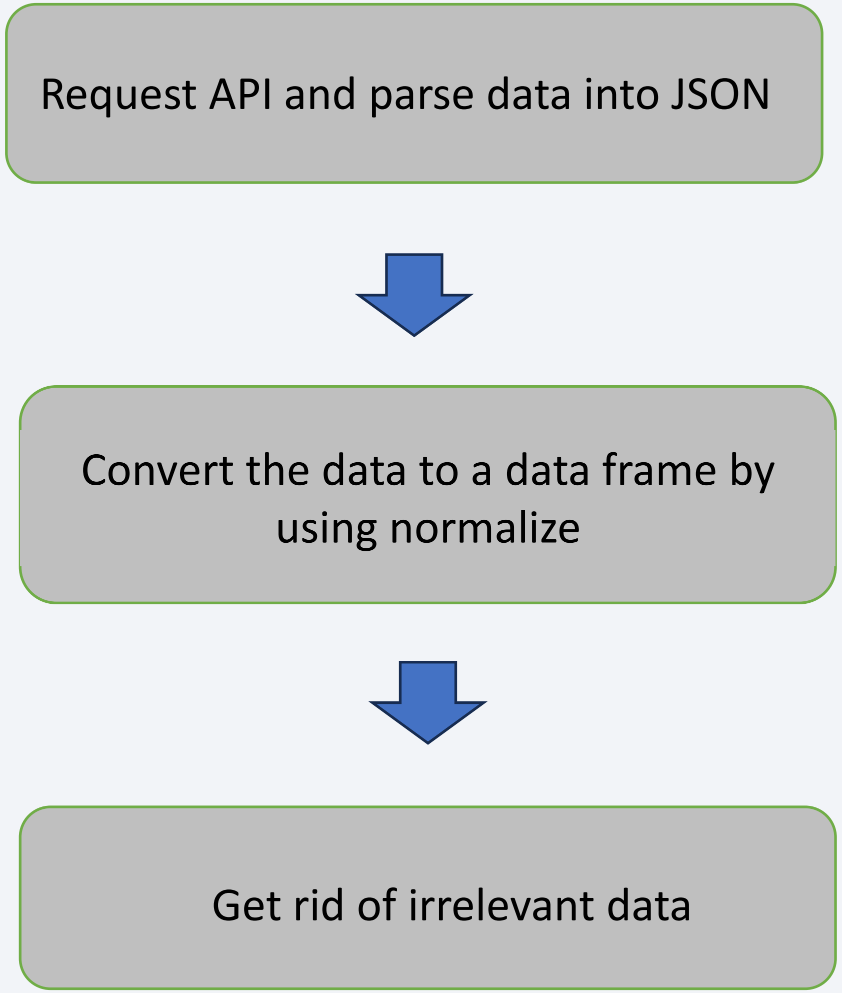
Requesting data from the API and converting the response to json

Create a data frame from the data and processing it.

Exporting the processed data as CSV

Data Collection – SpaceX API

- Request the data from the API, covert the data requested into a data frame using (json_normalize) so it is more eligible.
- Many columns are IDs which are not useful in our prediction model but we can use it to make other API calls to get more relevant data such as rockets, payload, launchpad and cores
- Code source: https://github.com/AlmahdiAhmed/spaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



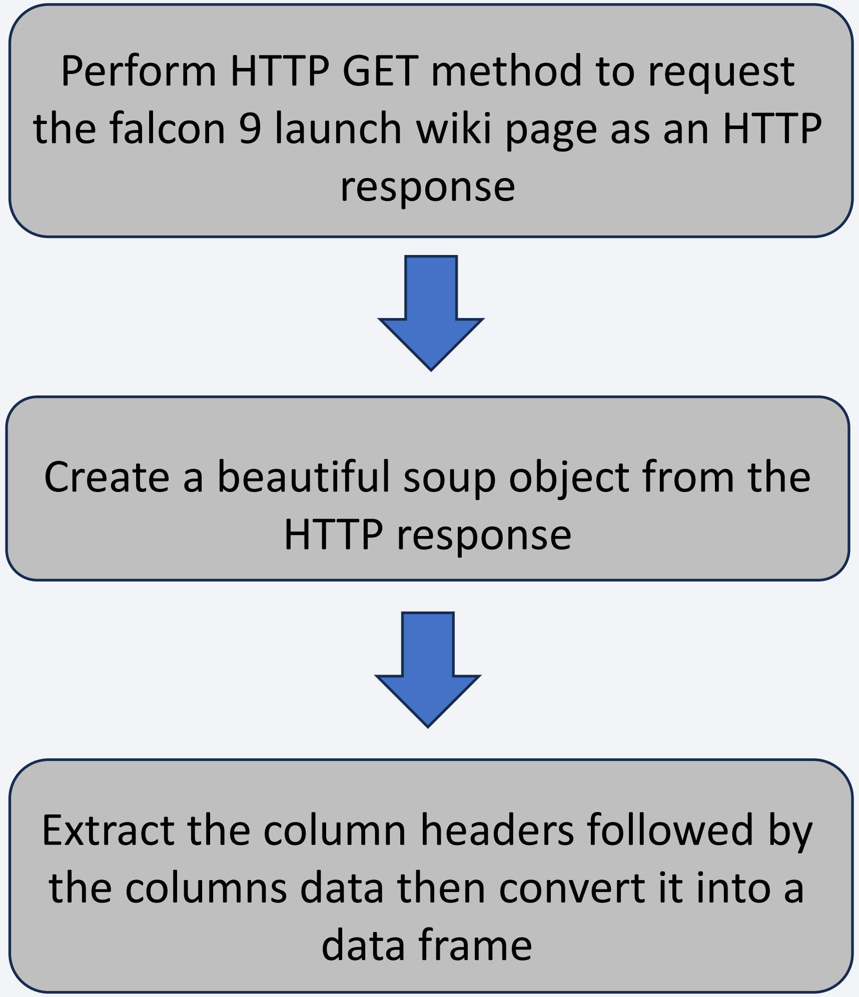
Request API and parse data into JSON

Convert the data to a data frame by using normalize

Get rid of irrelevant data

Data Collection - Scraping

- Extract Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup library
- Parse the table and covert it into a pandas data frame.
- Code source: https://github.com/AlmahdiAhmed/spaceX/blob/main/jupyter-labs-webscraping.ipynb



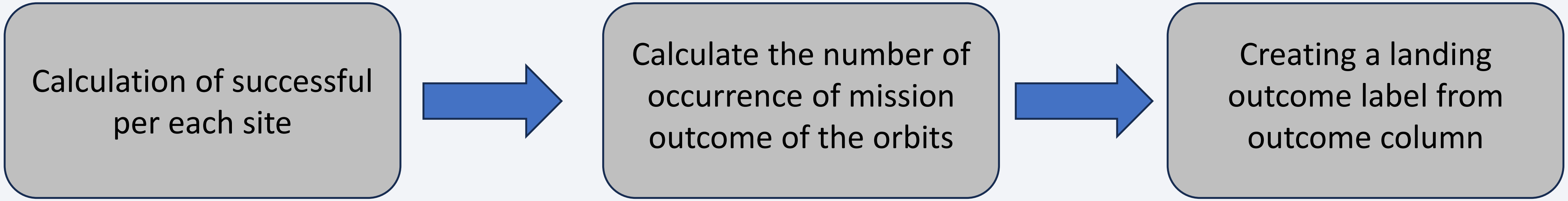
Perform HTTP GET method to request the falcon 9 launch wiki page as an HTTP response

Create a beautiful soup object from the HTTP response

Extract the column headers followed by the columns data then convert it into a data frame

Data Wrangling

- The objective is to perform exploratory data analysis and determine training labels
- Identify the number of missing values in each attribute
- Launches per site, occurrence of each orbit and occurrence of mission outcome per orbit type were calculated



Calculation of successful per each site

Calculate the number of occurrence of mission outcome of the orbits

Creating a landing outcome label from outcome column

- Source code: https://github.com/AlmahdiAhmed/spaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatter plots and bar charts were used to visualize relationship between pair of features
- A scatter plot was used to see the effect of flight numbers and payload on the launch outcome
- Another scatter plot was used to see the effect of launching site on the flight number and payload
- A bar chart was used to check if there is any relationship between success rate and orbit type
- A line chart was used to see the success rate in each year
- Source code: https://github.com/AlmahdiAhmed/spaceX/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- Query to display the names of the unique launch sites in the space mission
- Query to display five records where launch sites begin with the string 'CCA'
- Query to display the total payload mass carried by boosters launched by NASA (CRS)
- Query to display average payload mass carried by booster version F9 v1.1
- Query to list total number of successful and failure mission outcomes
- Query the name of booster version which have carried the maximum payload mass

Code source:

https://github.com/AlmahdiAhmed/spaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The objective is to mark all launch sites on a map, mark successful/failed launches for each site on the map and calculate the distance between a launch site to its proximities
- A folium map object with initial center NASA was created
- Circle was added for each launch site based on its coordinates
- Green marker was used for successful launch and red for failed launch
- Launch site will have multiple markers that's why marker clusters were used
- Polyline was drawn between a launch site to the selected coastline point and to the closest city, railway or highway

- Code source: https://github.com/AlmahdiAhmed/spaceX/blob/main/lab_jupyter_launch_site_location.ipynb

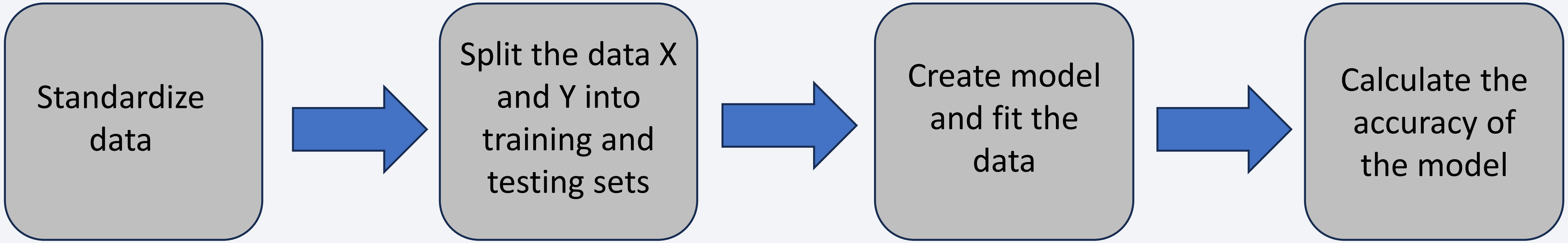
Build a Dashboard with Plotly Dash

- The dashboard application contains inputs components such as a dropdown list and a range slider to interact with pie chart and scatter point chart
- Pie chart was added to visualize launch success counts from all launch site and you can chose a specific launch site from the dropdown menu
- Scatter plot was used to with X axis to be the payload and the Y axis to be launch outcome. Each booster version has its own color

- Code source: https://github.com/AlmahdiAhmed/spaceX/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine SVM, decision tree and K nearest neighbors
- Data was split into training and testing data, fitted in the models. Model and hyperparameters with best outcomes were selected


Standardize data
Split the data X and Y into training and testing sets
Create model and fit the data
Calculate the accuracy of the model

- Code source:

https://github.com/AlmahdiAhmed/spaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

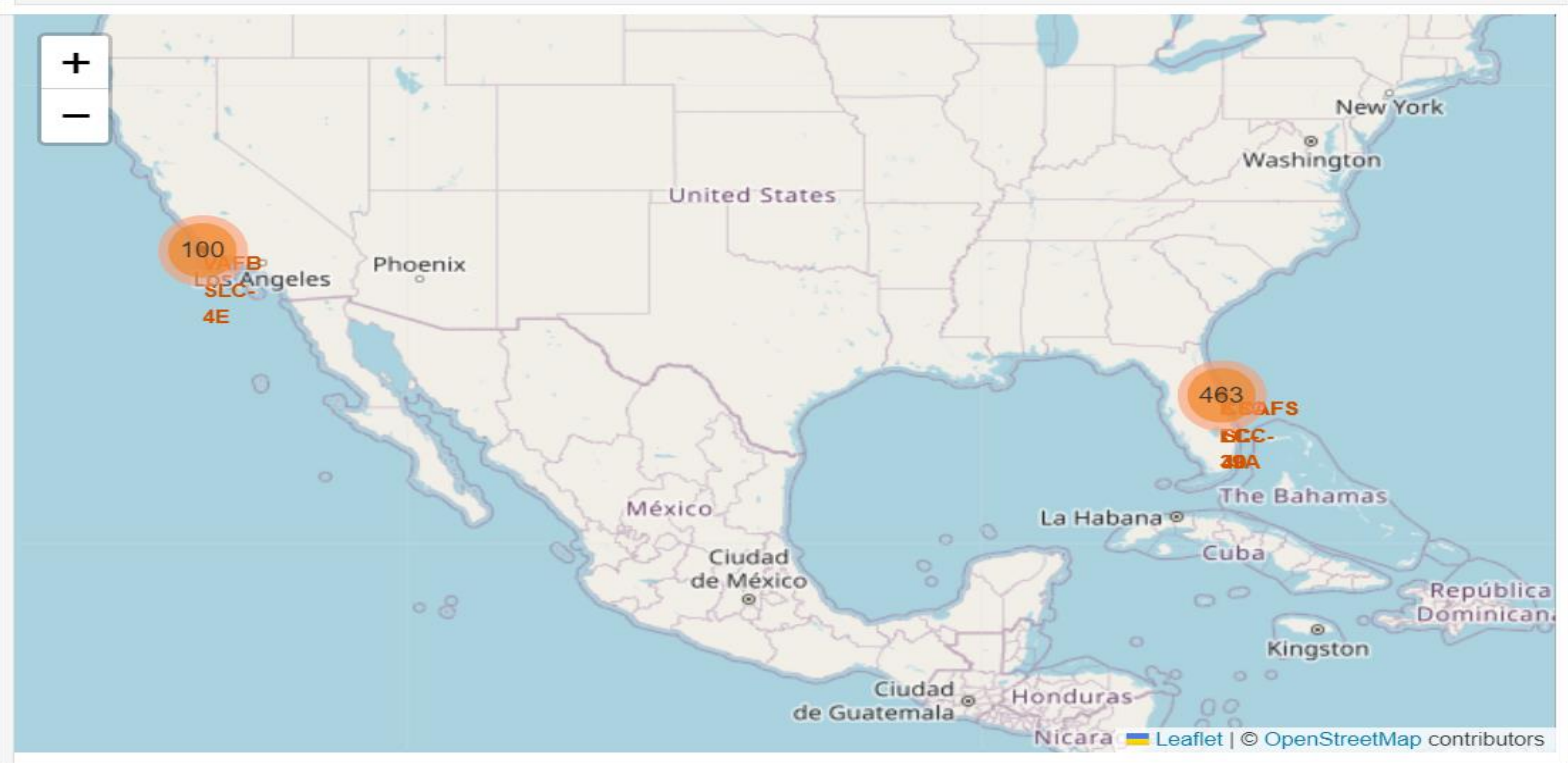
Results

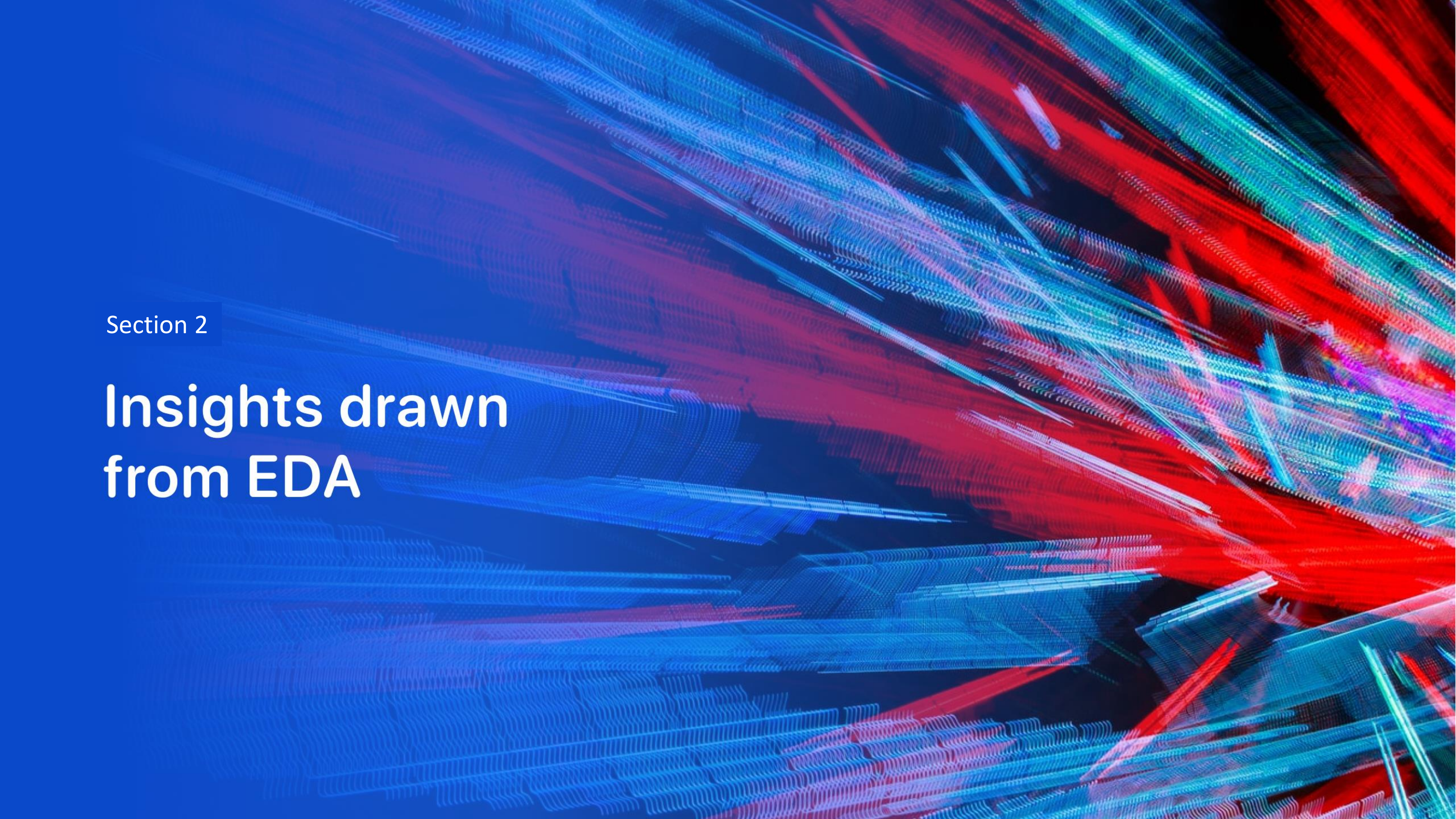
- Exploratory data analysis results:
  - Space X uses 4 different launch sites
  - Average payload of F9 v1.1 booster is 2,928 kg
  - First success landing outcome happened is 2015 five years after the first landing
  - The number of landing outcomes become as better as years passed

- Predictive analysis showed that Decision tree classifier is the best model to predict successful landing having 87%accuray

Results

- All launches happened in safe zones near coastline





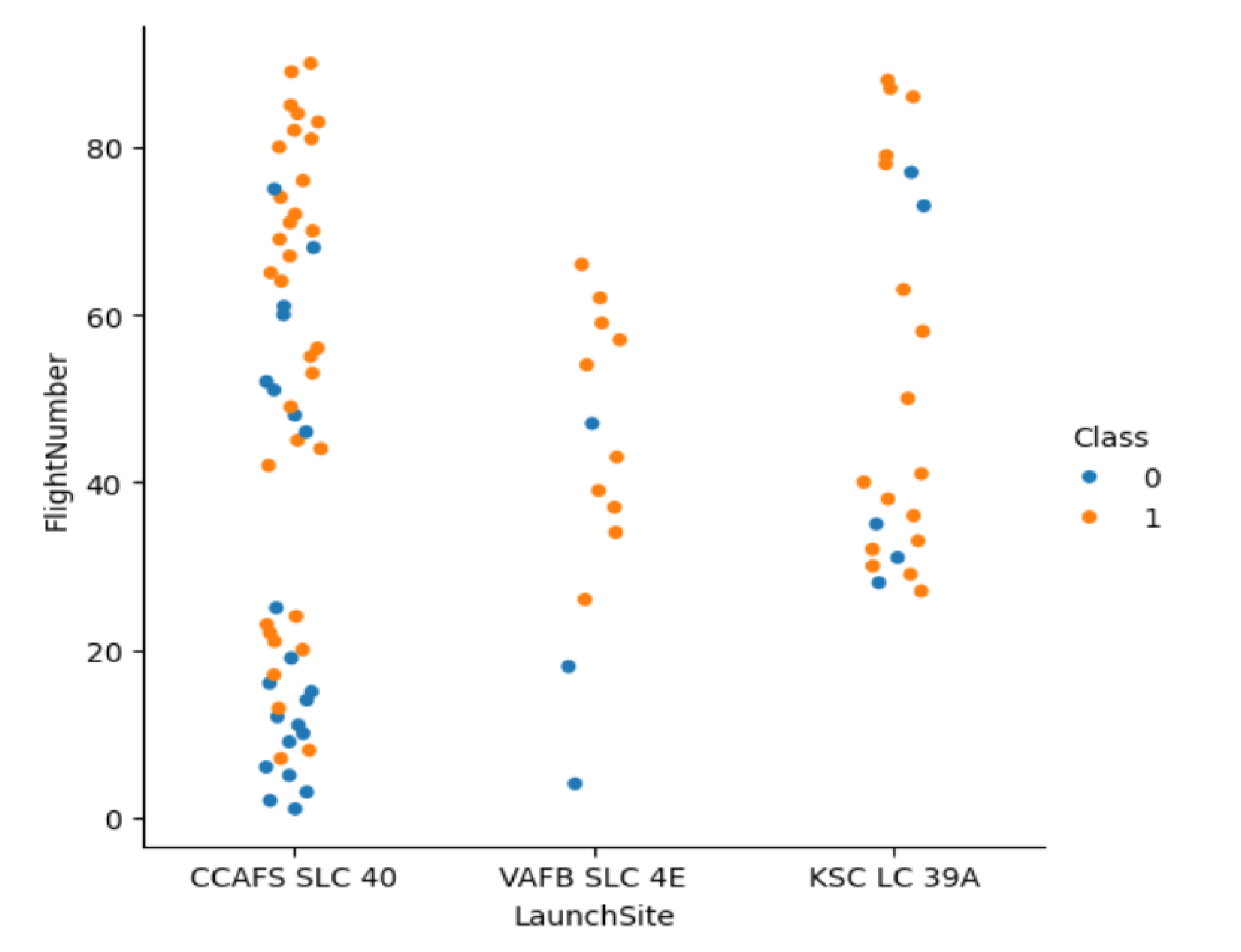
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

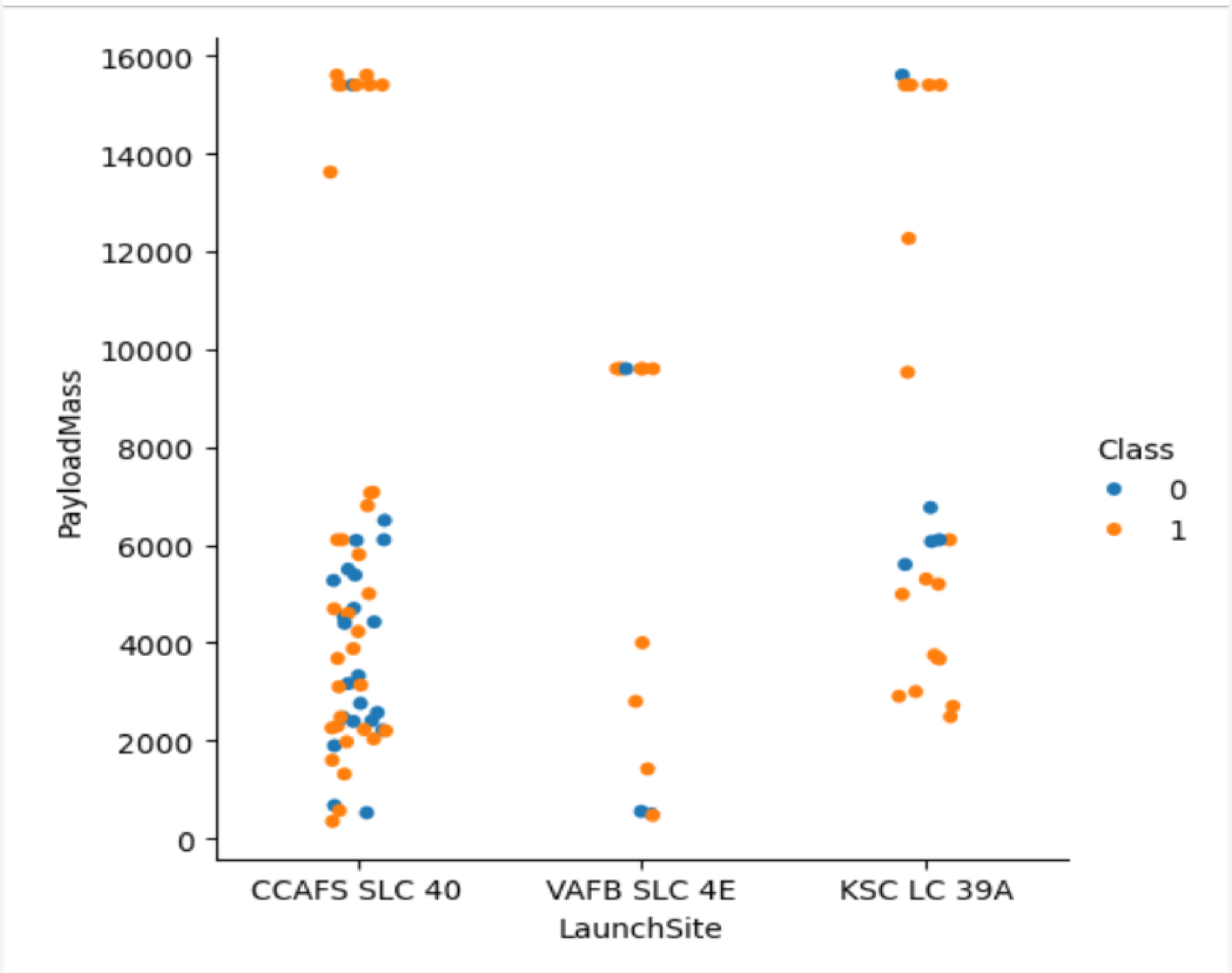
- Number vs. Launch Site

- This scatter plot shows the number of successful and failed flights on each launch site



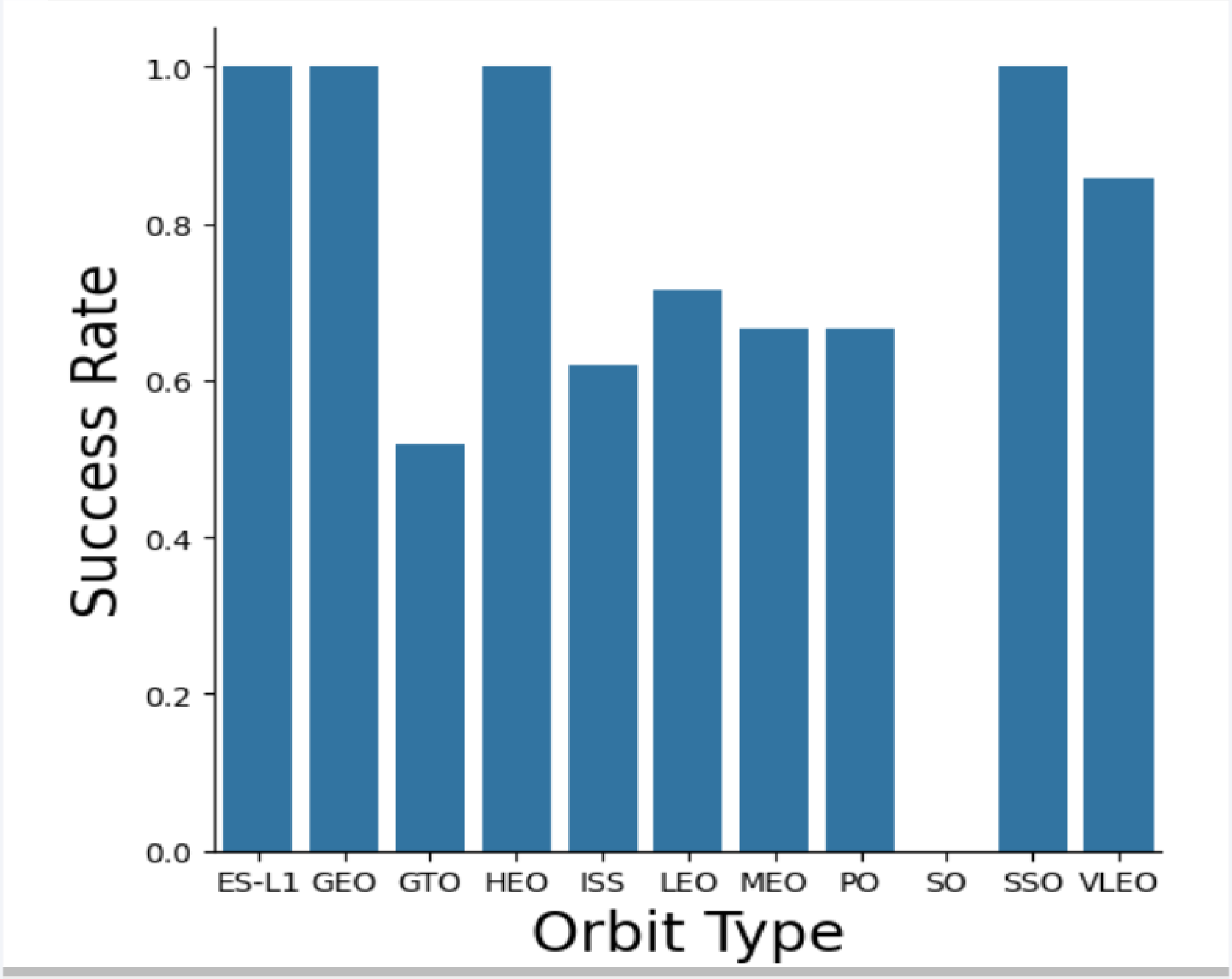
Payload vs. Launch Site

- This plot shows the change in payload mass of each launch and it landing outcome
- Payload between 8000 and 10000 has excellent success rate



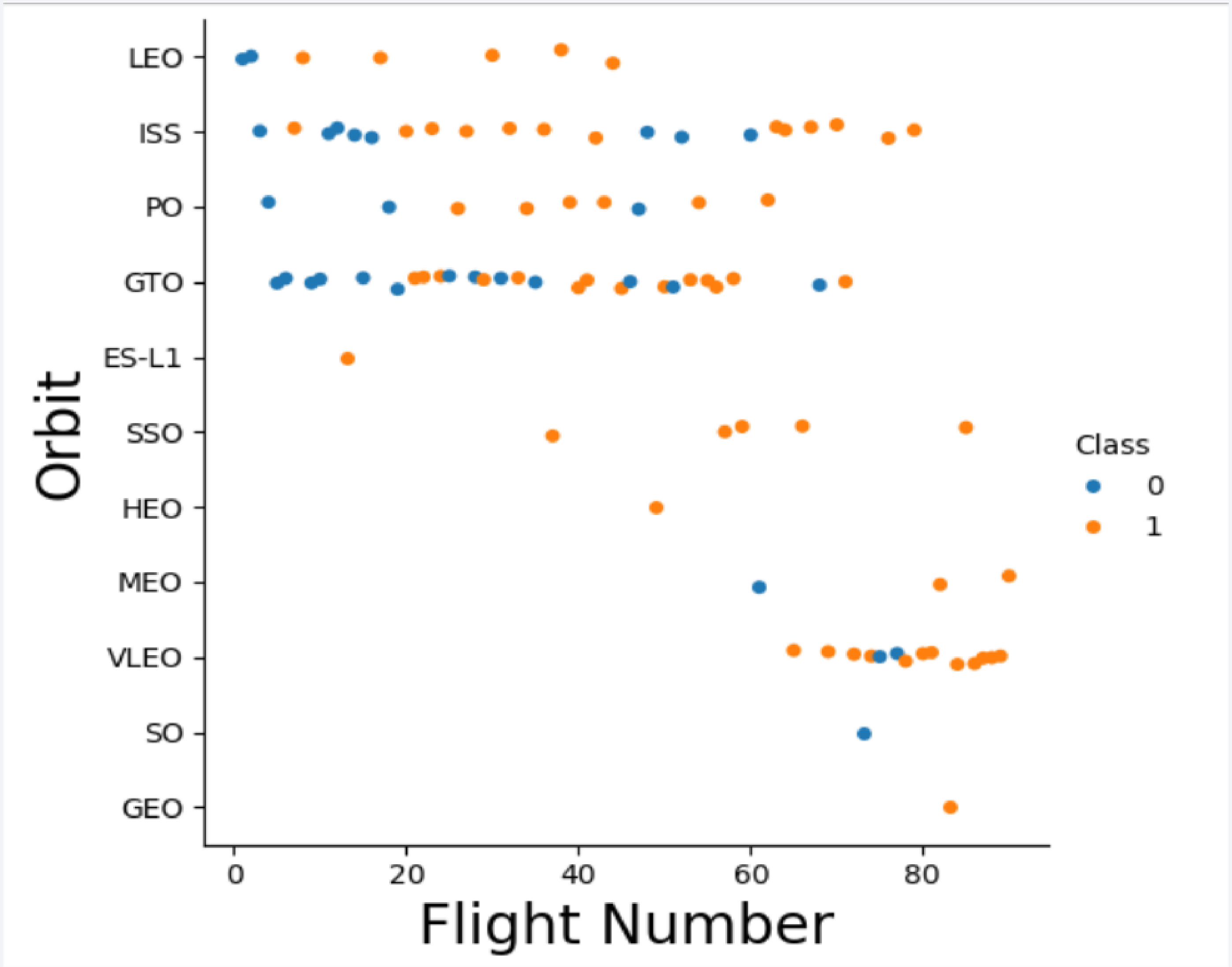
Success Rate vs. Orbit Type

- Bar chart comparing success rate of each orbit type
- It's clear the SO have zero success rate



Flight Number vs. Orbit Type

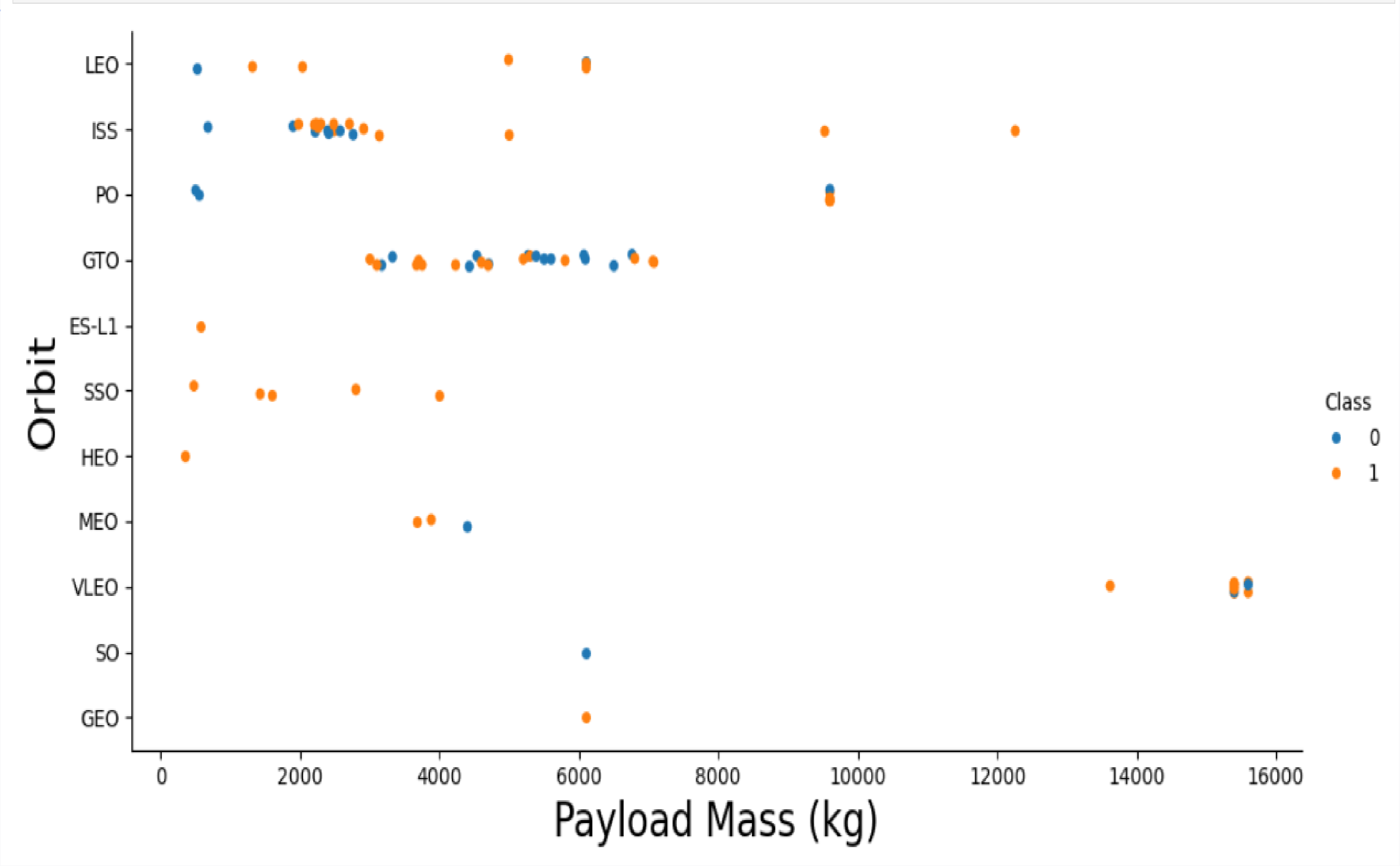
- This scatter plot the number of flights on each orbit



Payload vs. Orbit Type

- This plot shows the relation between orbits and payload mass

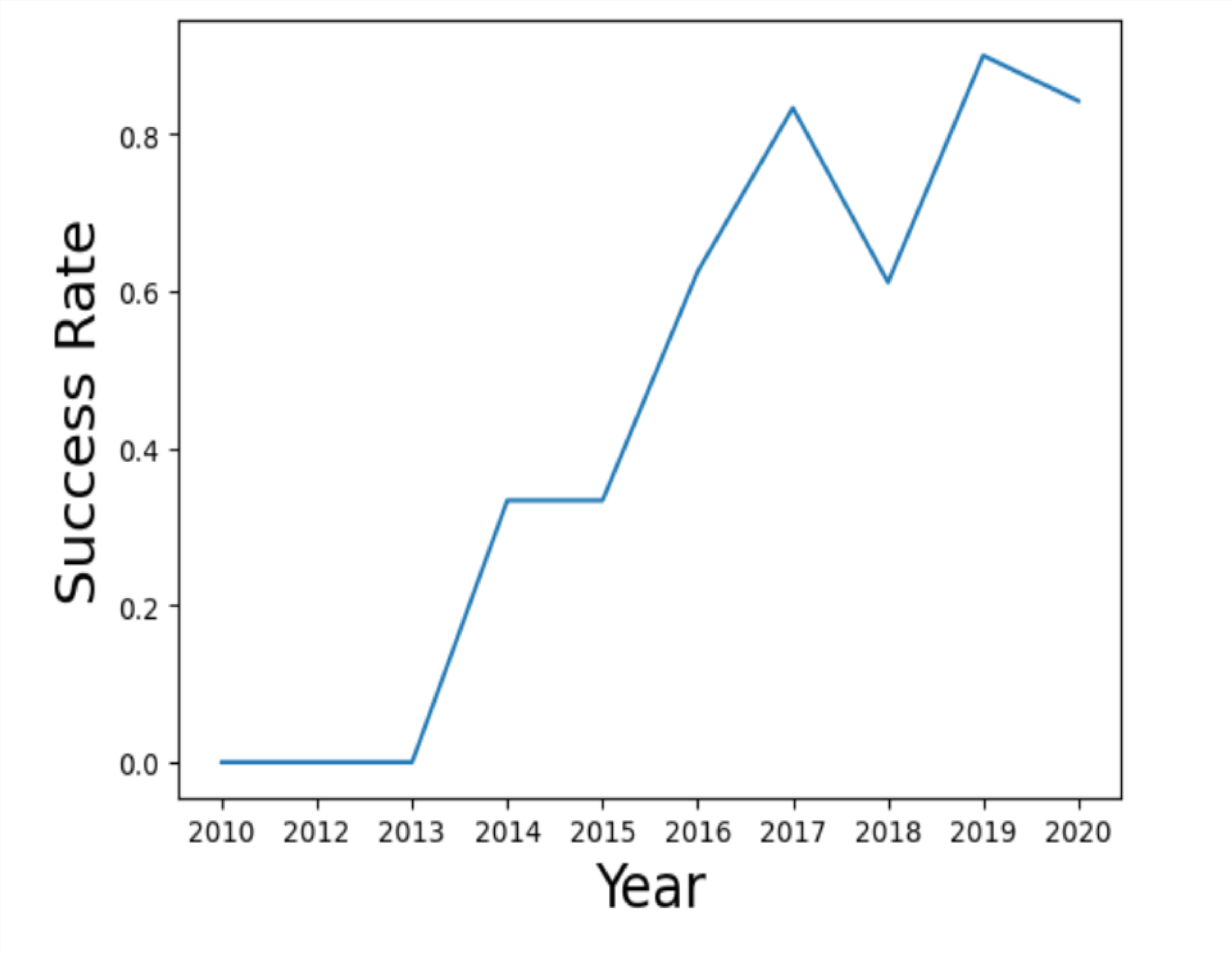
- There is no relation between payload and success rate to orbit GTO



Launch Success Yearly Trend

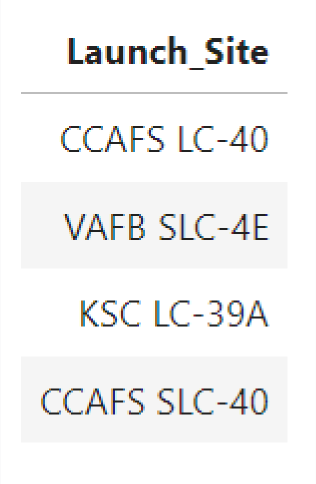
- This line chart shows the relationship between success rate and year

- Starting from 2013, success rate has increased



All Launch Site Names

- A SQL query was performed to get all launch sites

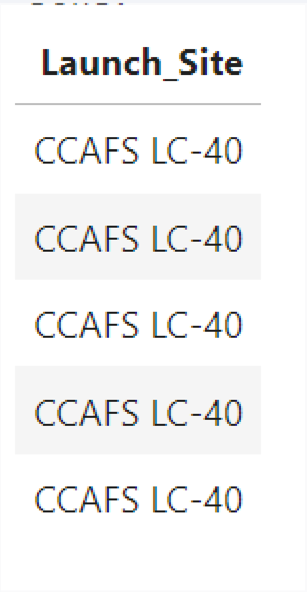


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- This was done by querying launch site from the database

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

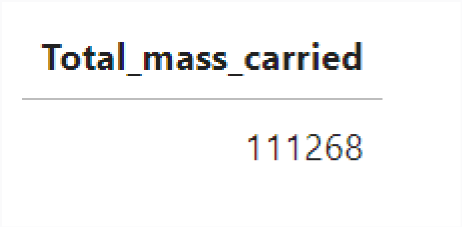


Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- This was done by querying the database then using regex and lastly limiting the results to five

Total Payload Mass

- Total payload carried by boosters from NASA

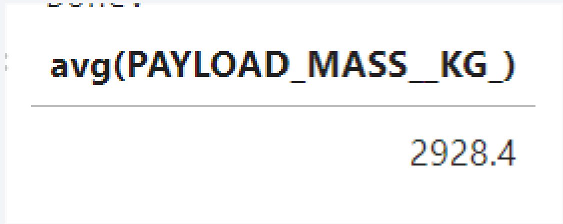


Total_mass_carried
111268

- This result was obtained by summing all payload values where it was launched by NASA ‘CRS’

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

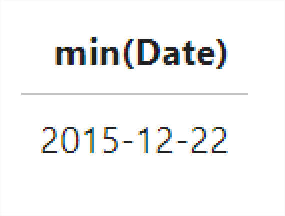


avg(PAYLOAD_MASS__KG_)
2928.4

- This result was obtained by using the average function on the payloads where booster version is F9 v1.1

First Successful Ground Landing Date

- First successful landing outcome on ground pad

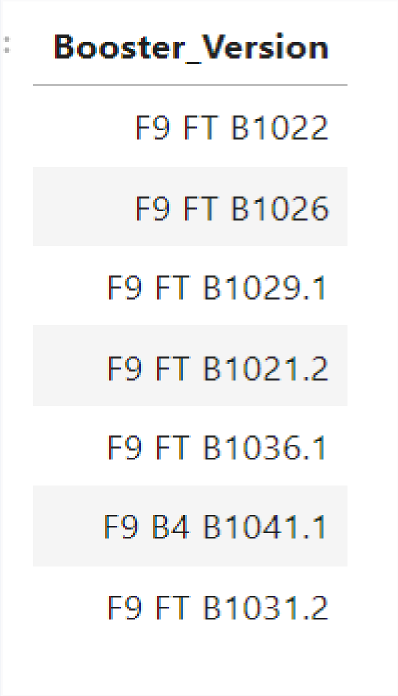


min(Date)
2015-12-22

- This was obtained by getting the minimum date where the landing outcome was successsful

Successful Drone Ship Landing with Payload between 4000 and 6000

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

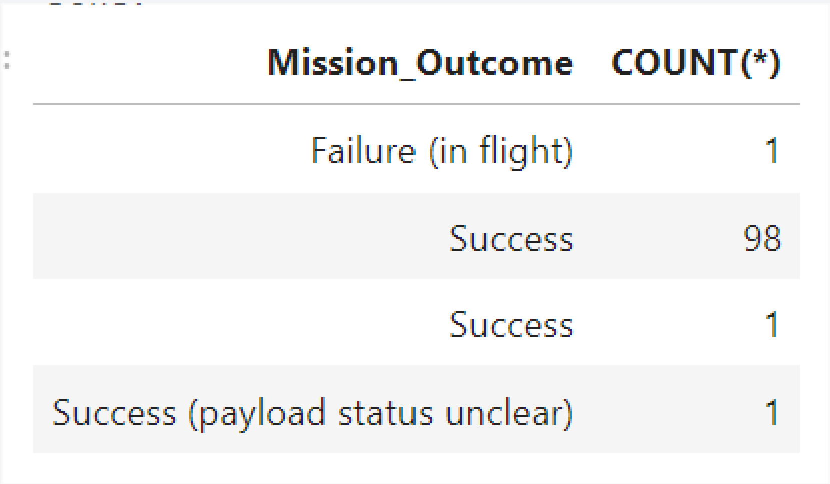


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1036.1
F9 B4 B1041.1
F9 FT B1031.2

- First we query booster version with successful outcome then we limit the payload to be greater than or equal to 4000 and less than 6000

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes



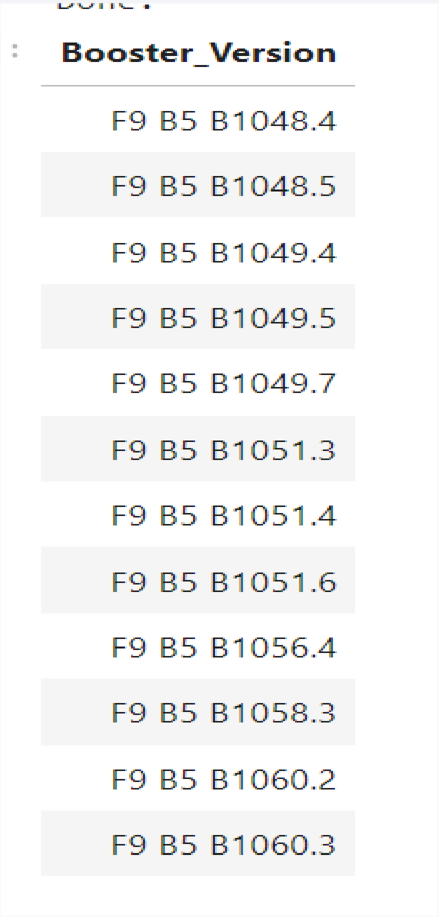
Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- This result was obtained by grouping our data by the mission outcome then selecting to display mission outcome and its count

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

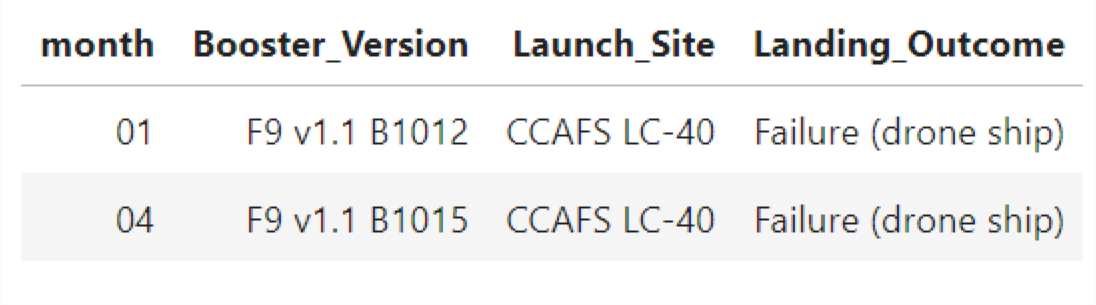
- A subquery was used, in the sub query we obtained the max pay load and order it by booster version
- In the main query, only distinct booster version were obtained to avoid duplication



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
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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



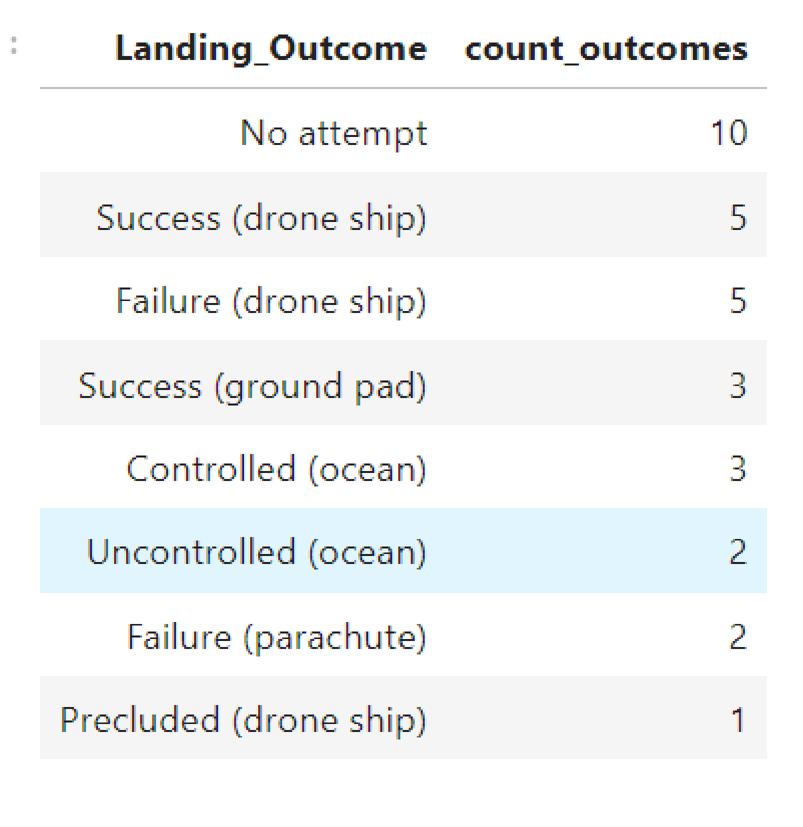
month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- This table was obtained by querying month, booster version, launch site and landing outcome where landing outcome was a fail at the year 2015

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

- Ranks 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

- This table was obtained by getting the landing outcome and its count outcome where data between 2010-06-04 and 2017-03-20 then grouping by landing outcome and ordering by count outcome discendingly

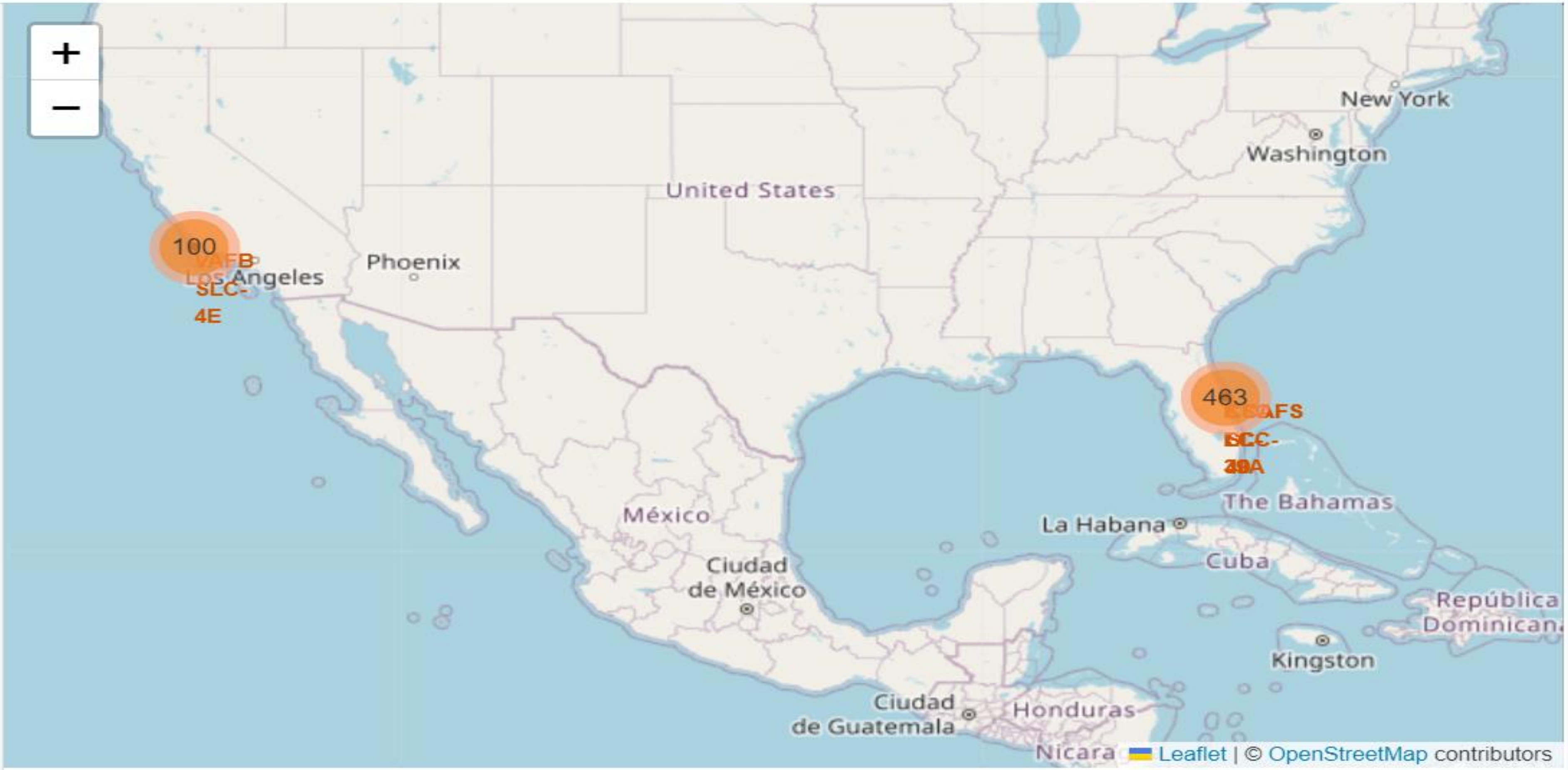


Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

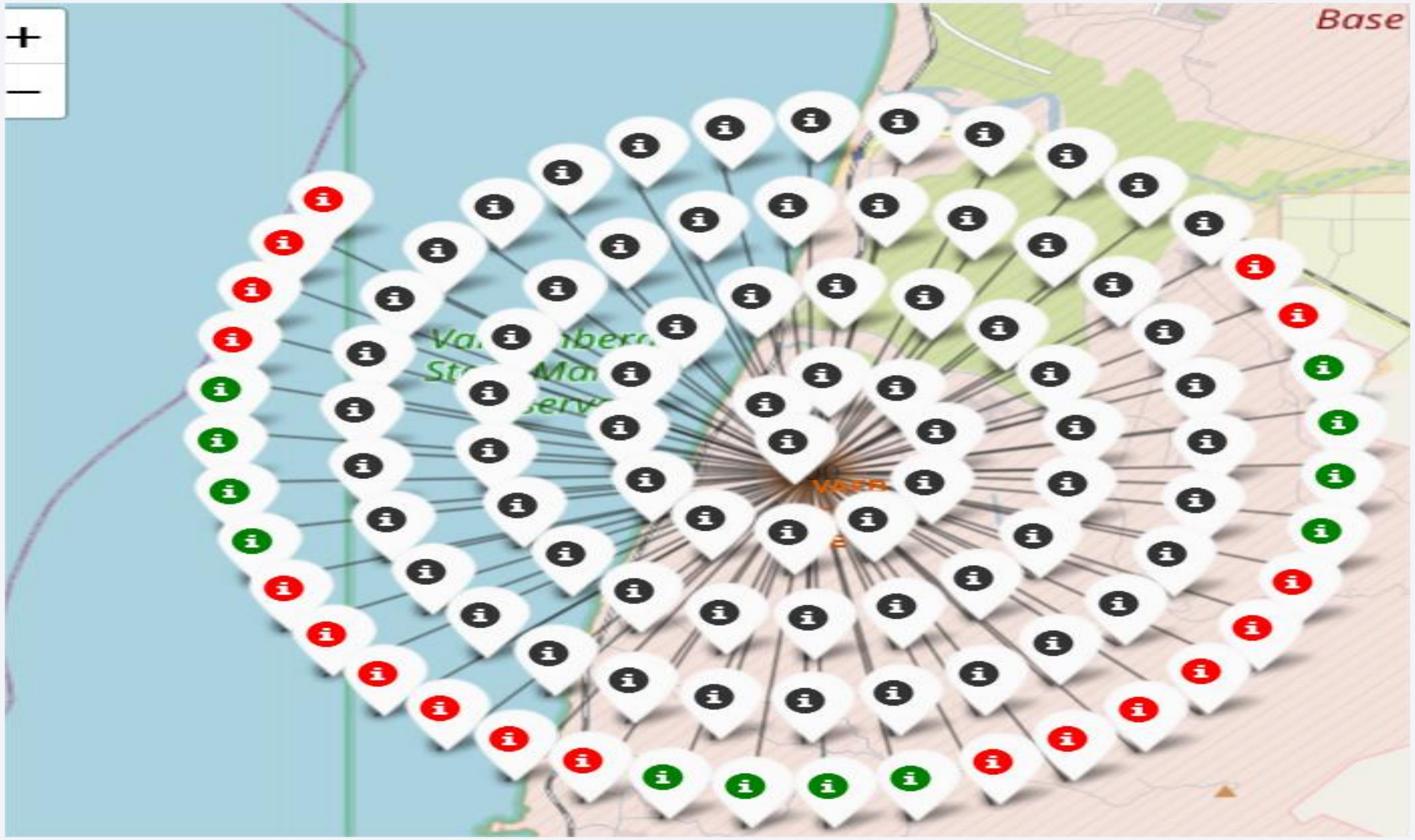
Launch Sites Proximities Analysis

All launch sites



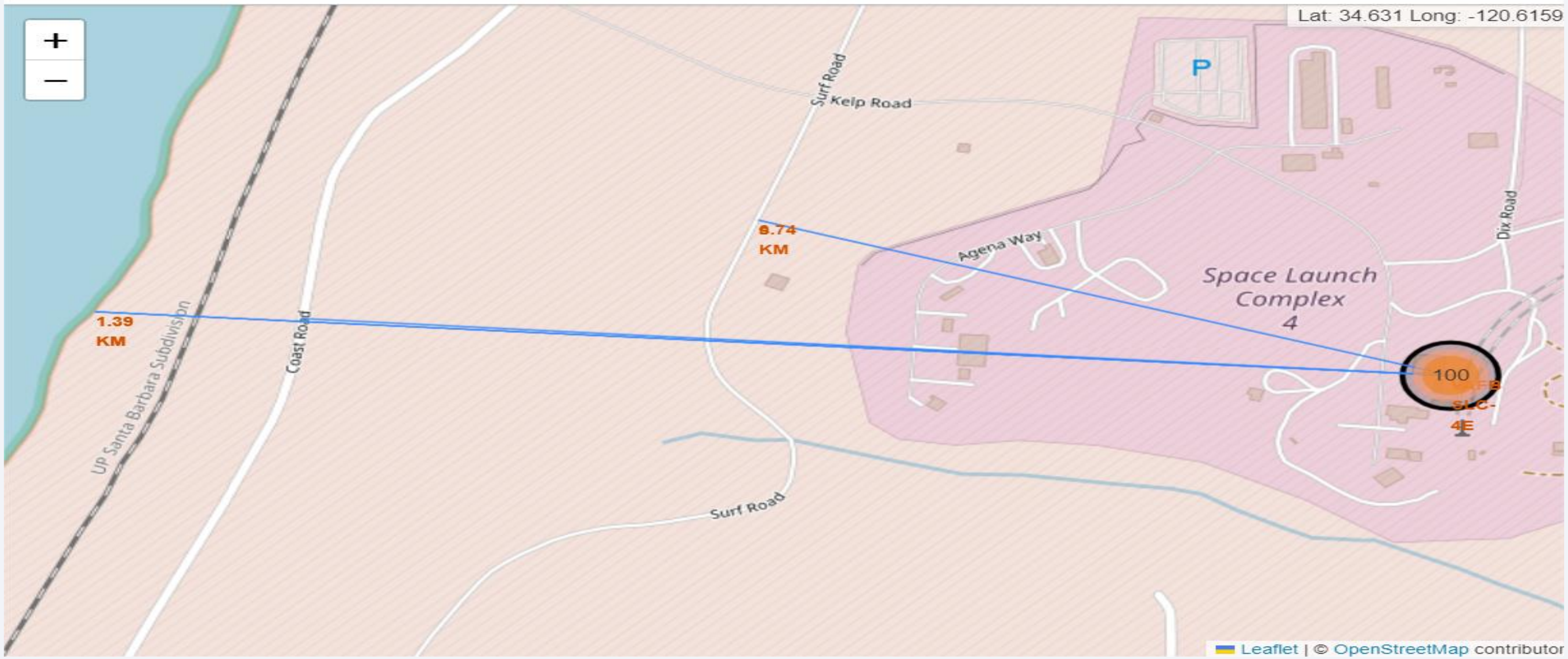
This map shows that are launch sites are in safe zone which is near coastline

Launch outcome by site

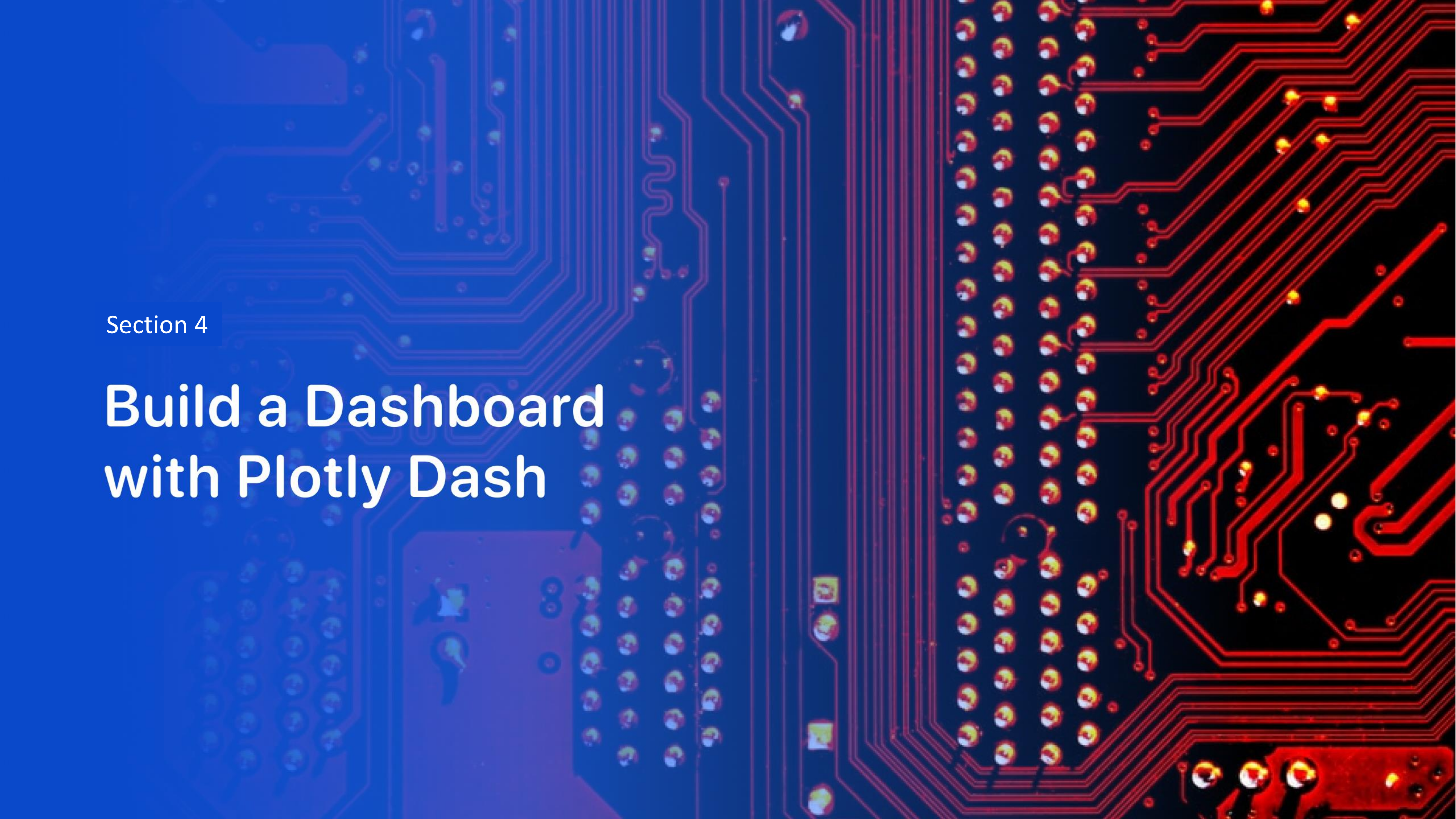


- Green marker represent successful launch while red marker represent failed

Proximities to launch site



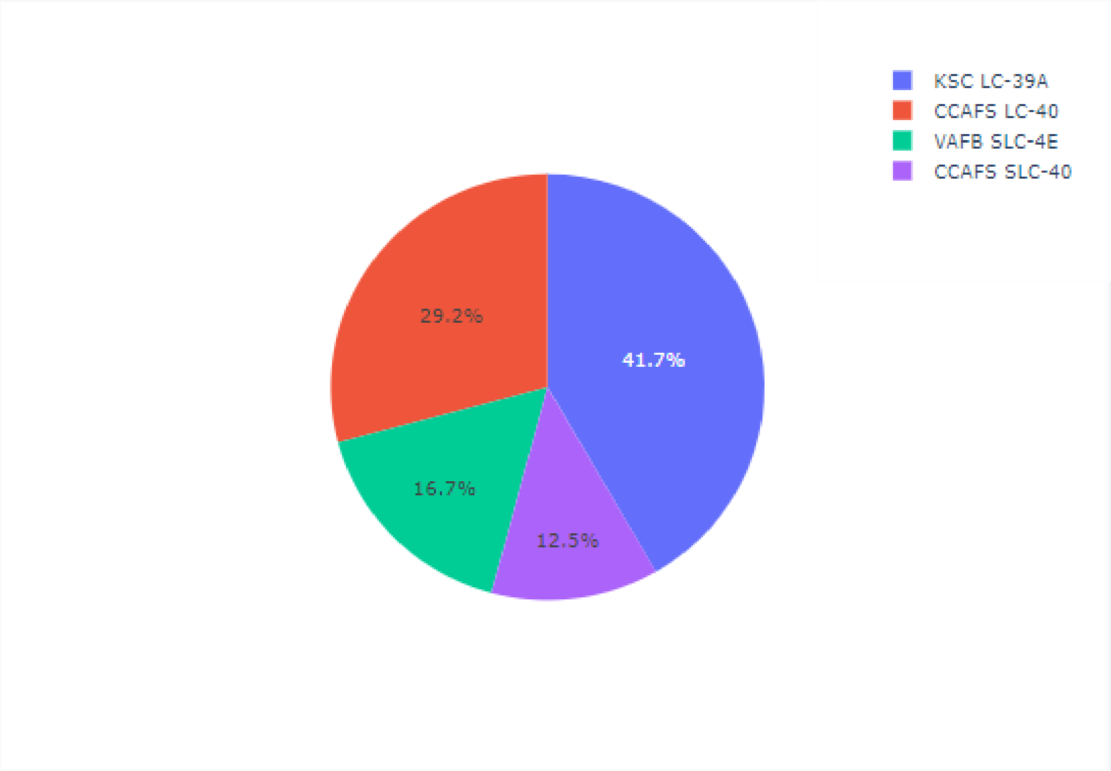
- This map shows that the launch site is 1.39 km from the coast and .74 km from Surf Road



Section 4

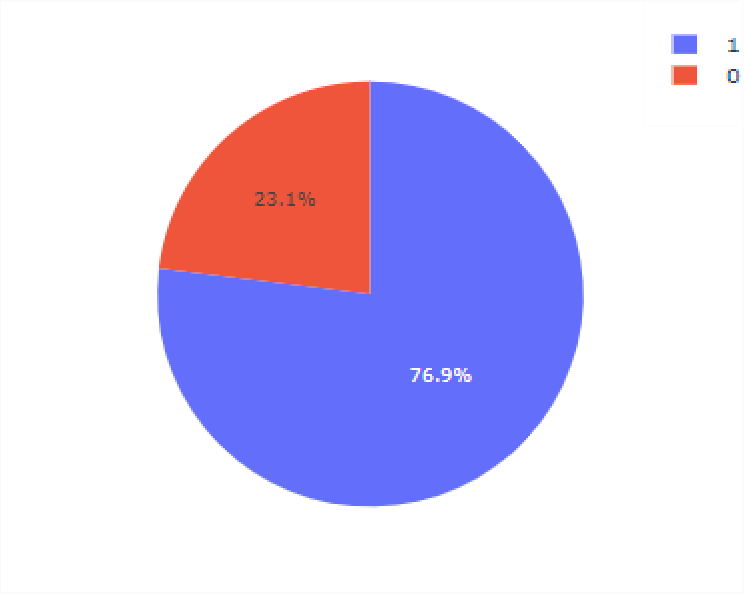
Build a Dashboard with Plotly Dash

All sites success count



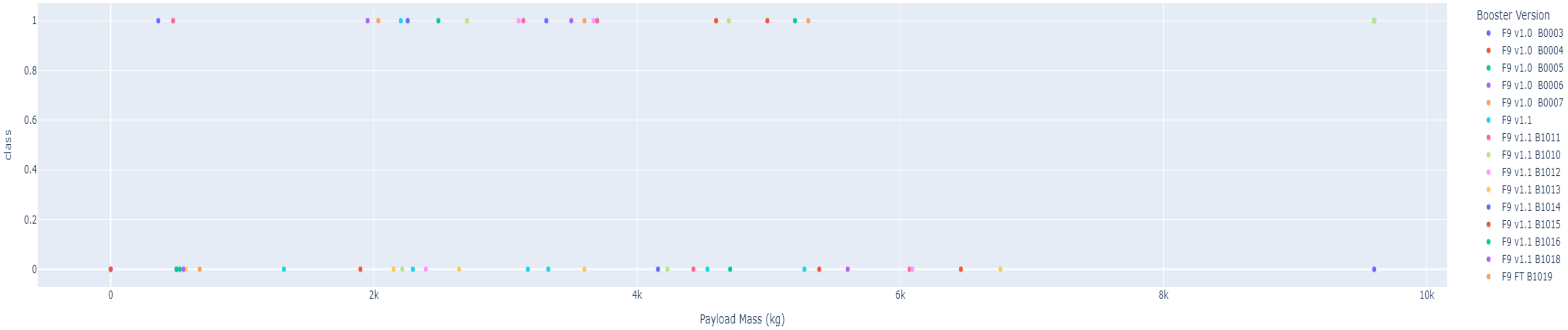
- This pie chart shows the KSC LC-39A has the highest success rate

Highest success ratio



- 76.9% were successful wile 22.1% failed

Effect of payload mass on outcome



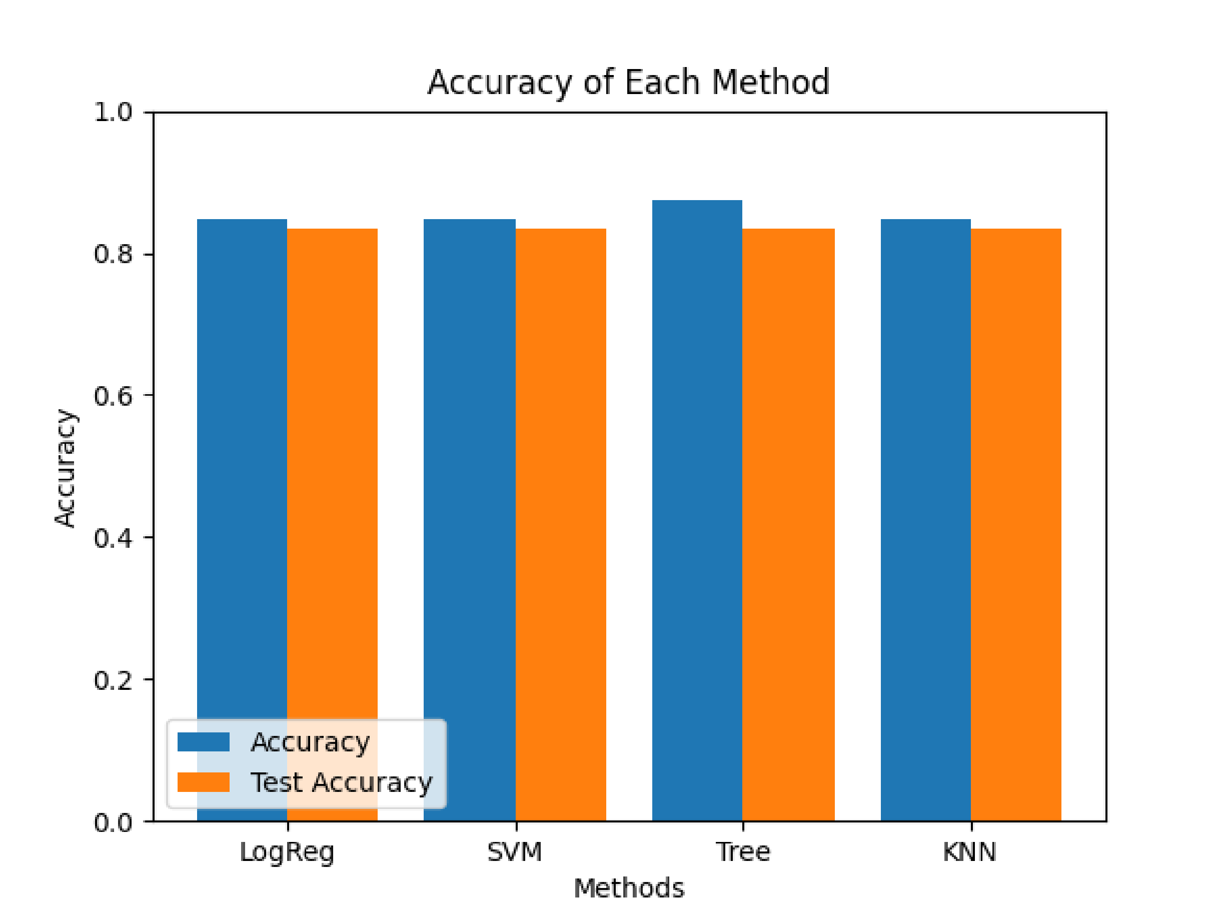
- Payload under 6,000 and FT boosters are the most successful combination

Section 5

Predictive Analysis (Classification)

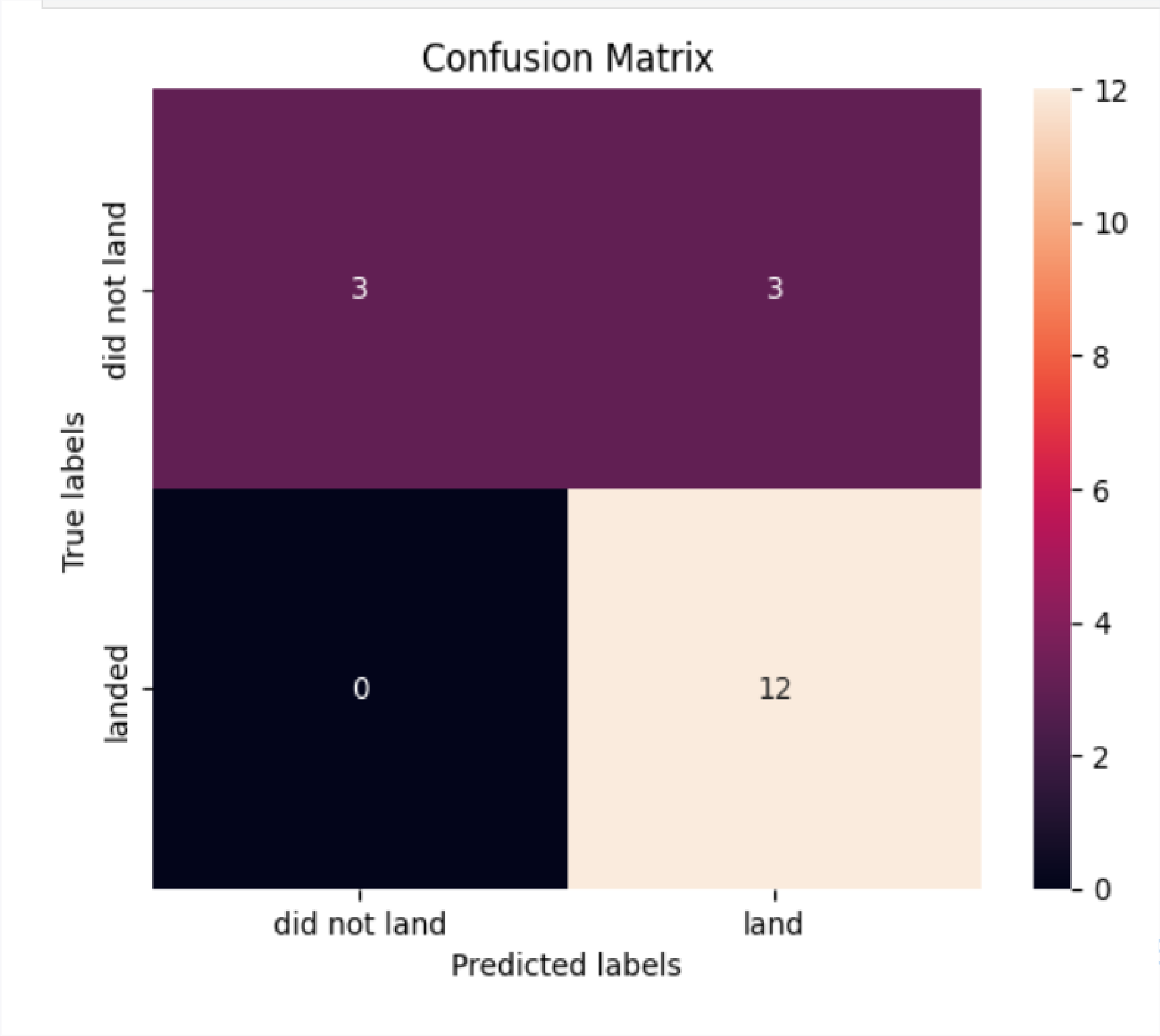
Classification Accuracy

- Four classifications were tested
- Decision tree classifier has the highest accuracy



Confusion Matrix

- Confusion matrix shows the accuracy by demonstrating the true positive, false positive, true negative and false negative of each value



Conclusions

- The best launching site is KSC LC-39A
- Most suitable payload is between 8000 and 10000
- First success landing outcome happened in 2015 five years after the first launch
- The number of successful landing outcomes increase with years passing
- …

Thank you!