IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Rully Yulian MF
11-November-2025

# Outline

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

# Executive Summary

- Methodologies:
  - Collecting data from various data sources
    - API
    - Wiki Page: Processed by web scraping technique
  - Data Wrangling
    - Checking missing values
    - Find some patterns in data with some data analysis
  - Exploratory Data Analysis (EDA)
    - Loading data into database
    - Exploring and analyzing data with sql aggregate function

# Executive Summary

- Methodologies:
  - Exploratory Data Analysis (EDA)
    - Visualize relationship between features using Matplotlib and Seaborn
    - Feature Engineering
  - Interactive Data Visualization
    - Mapping data into map using Folium
    - Creating dashboard using Plotly Dash
  - Predictive Analysis
    - Data classification using some algorithm: Logistic Regression, Support Vector Machine, Decision Tree, and KNN

# Executive Summary

- Results:
  - CCAFS SLC 40 is the launch site which has highest number launch around 55 launch
  - ISS is the orbit type which has the highest accessed orbit
  - First date of successful landing with ground pad is 2015-12-22
  - There are around 100 success outcomes and 2 failed outcomes
  - All launch sites are at the coastline, far away from city, railway, and highway
  - CCAFS LC-40 has the highest successful rate
  - Decision tree is the best model for predicting successful landing. It has highest accuracy value from the other models.

# Introduction

- We are in the commercial space age
- Some companies are competing for making space travel
- SpaceX spending less cost than others
- Gathering information and data about SpaceX
- How can we compete with SpaceX?
- What aspects influence the low cost by SpaceX?
- Can we build the model to predict the successful landing?

Section 1

# Methodology

# Methodology

## Executive Summary

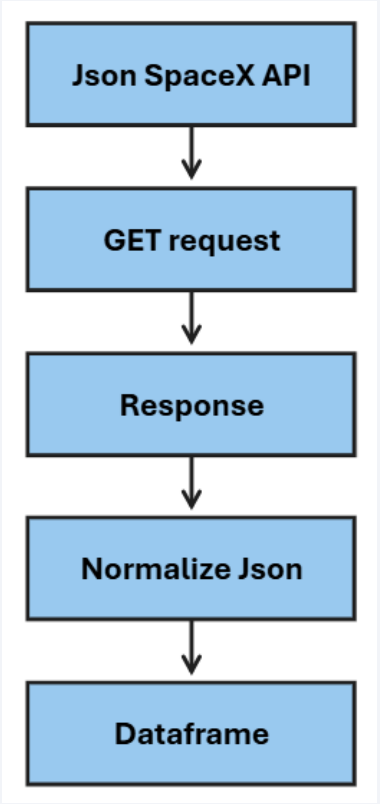
- Data collection methodology:
  - Collecting data from API using GET request and Extracting HTML table from Wiki Page using web scraping
- Perform data wrangling
  - Cleaning data, analyzing data, and visualize relationship between features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split train test data, standardizing data, build models using Logistic Regression, SVM, Decision Tree and KNN. Find the best hyperparameter using GridSearchCV. Visualizing the output using Confusion Matrix

# Data Collection

- Data collected from various resources
- Json data from SpaceX API: processed by GET request
  - Json data converted to dataframe using normalize json method
- SpaceX wiki page: processed by web scraping technique
  - Extracting and Filtering HTML table data using BeautifulSoup for Falcon9 Launch

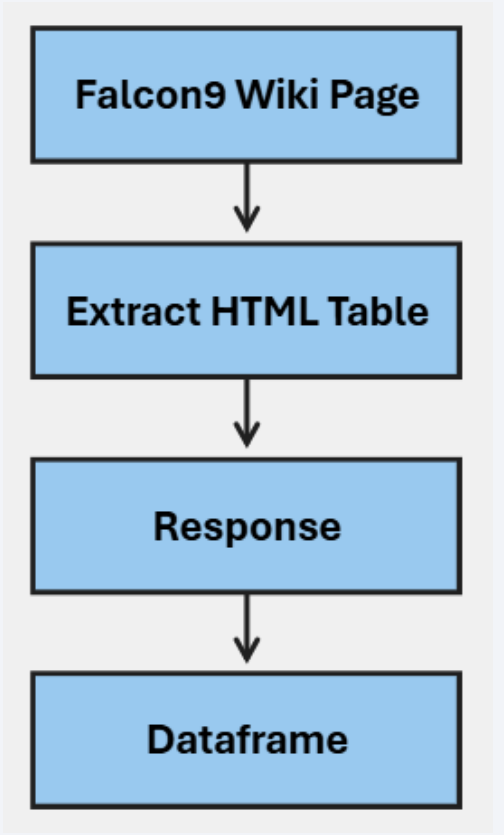
# Data Collection – SpaceX API

- Read json data from API using GET request method
- Convert json data from response into dataframe using normalize method
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%201-Introduction/jupyter-labs-spacex-data-collection-api.ipynb



# Data Collection - Scraping

- Read wiki page url for Falcon9 using GET request
- Extract HTML table from response using BeautifulSoup
- Load extracted data into dataframe
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%201-Introduction/jupyter-labs-webscraping.ipynb



# Data Wrangling

- Find data pattern to determine the label for building supervised model
- Load SpaceX dataset from csv file
- Calculating:
  - Number of launches on each site
  - Number and occurrence of each orbit
  - Number and occurrence of mission outcome of the orbits
- Creating landing outcome label from outcome
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%201-Introduction/labs-jupyter-spacex-Data%20wrangling.ipynb

# EDA with Data Visualization

- Creating scatter plot using searborn to visualize the relationship between feature
- Visualizing the relationship for features:
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - Success rate of each orbit type
  - Flight Number and Orbit Type
  - Payload Mass and Orbit Type
- Creating line plot to visualize the launch success yearly tren
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%202-Exploratory%20Data%20Analysis/edadataviz.ipynb

# EDA with SQL

- Distinct query to find unique launch site
- Where clause to find launch site name begin with ‘CCA’
- Sum function to calculate payload mass carried by NASA (CRS)
- Avg function for payload mass carried by booster version F9 v11
- Min function to find date for first successful landing in ground pad
- Between predicate to find booster version which have payload mass greater than 4000 and less than 6000
- Group by clause to list the number of successful and failure mission
- Max function with subquery to list all booster version that have carried the maximum payload mass
- Date and substr function to extract month from date to find the failure landing drone ship in 2015
- Count function and order by clause to rank the landing outcome
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%202-Exploratory%20Data%20Analysis/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

- Marker object used to mark specific location and customize visualization with icon and can contain text or popup
- Circles object create a highlighted circle area with text label around the point location with a certain radius
- Mouse position object to get the Lat and Long coordinate when mouse over on the map
- Polyline object to connect between two coordinates with line and can be used to visualize the distance between coordinates
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%203-Interactive%20Visual%20Analytics%20and%20Dashboard/lab_jupyter_launch_site_location.ipynb

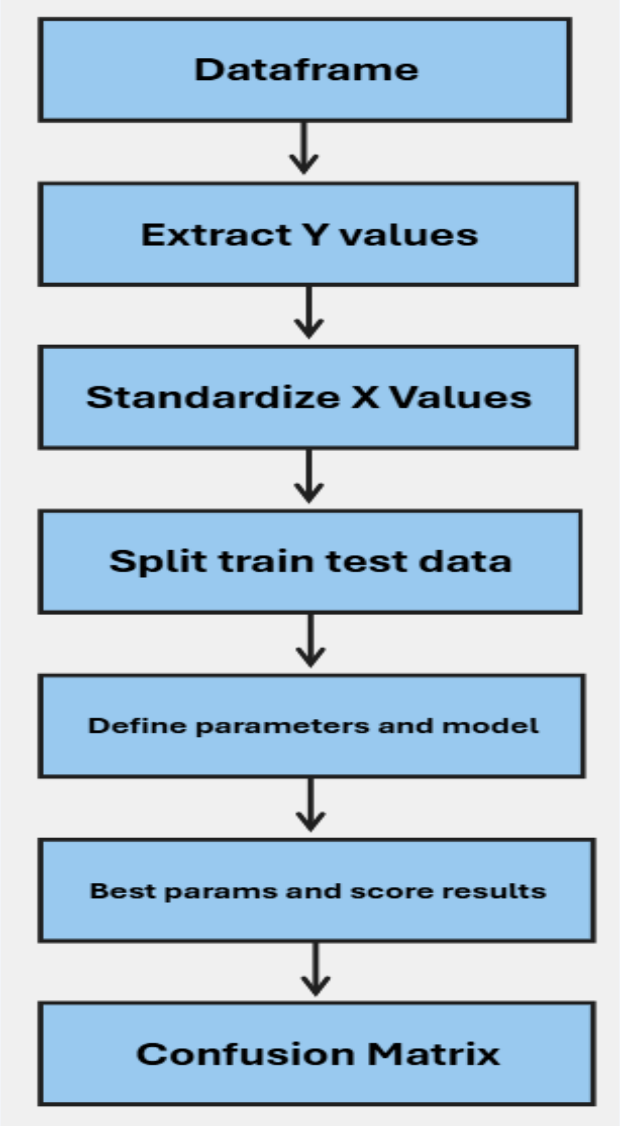
# Build a Dashboard with Plotly Dash

- Pie chart used to visualize summary value or for categorical data
  - Success rate for all sites
  - Success rate for certain site
- Scatter plot to visualize the relationship between features
  - Payload mass versus successful landing
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%203-Interactive%20Visual%20Analytics%20and%20Dashboard/spacex-dash-app.py

# Predictive Analysis (Classification)

- Classification model using algorithm: Logistic Regression, Support Vector Machine, Decision Tree, and KNN
- Split train test data and standardizing data
- Find the best hyperparameter using GridSearchCV for each model
- Display the best parameter using best_params_ attribute
- Displat the accuration model using best_score_ attribute
- Visualizing the output using Confusion Matrix
- https://github.com/yulianmf/coursera-ads-capstone/blob/main/Module%204-Predictive%20Analysis%20(Classification)/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

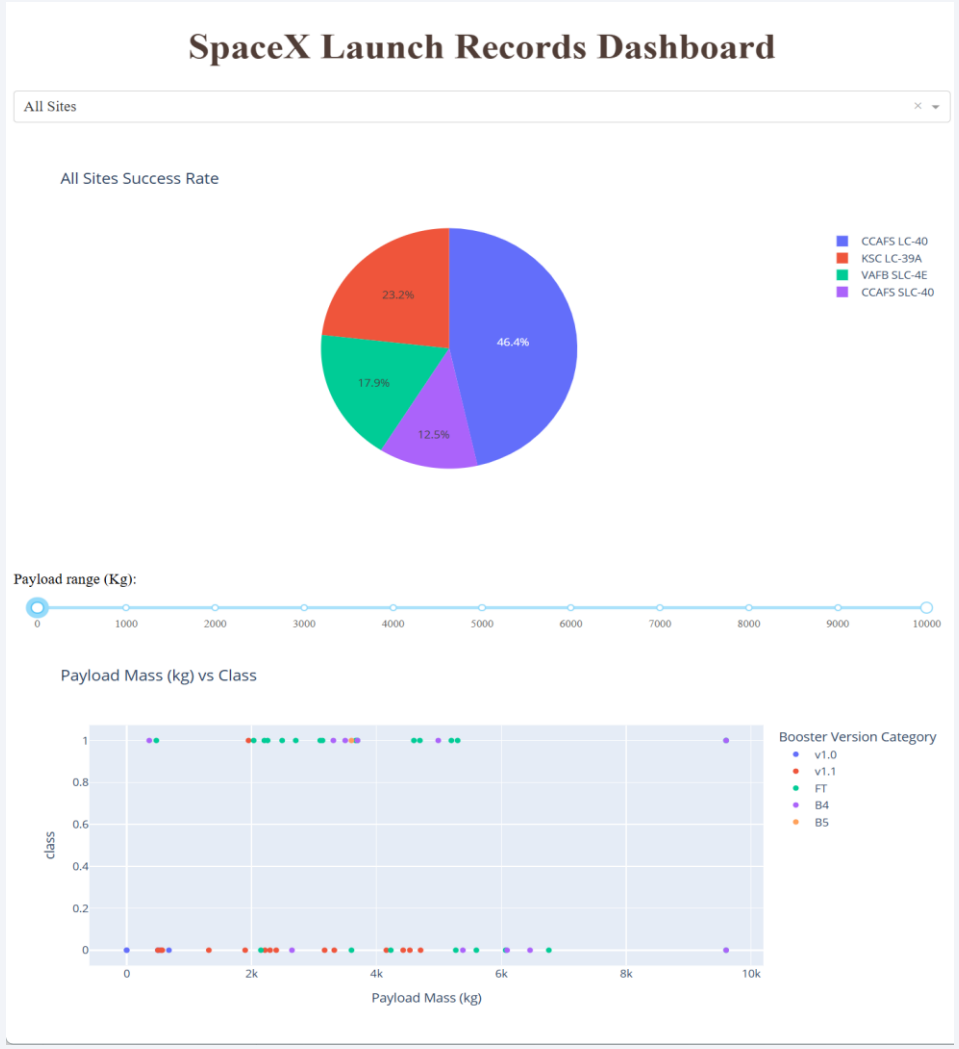
# Predictive Analysis (Classification) – Flow Chart

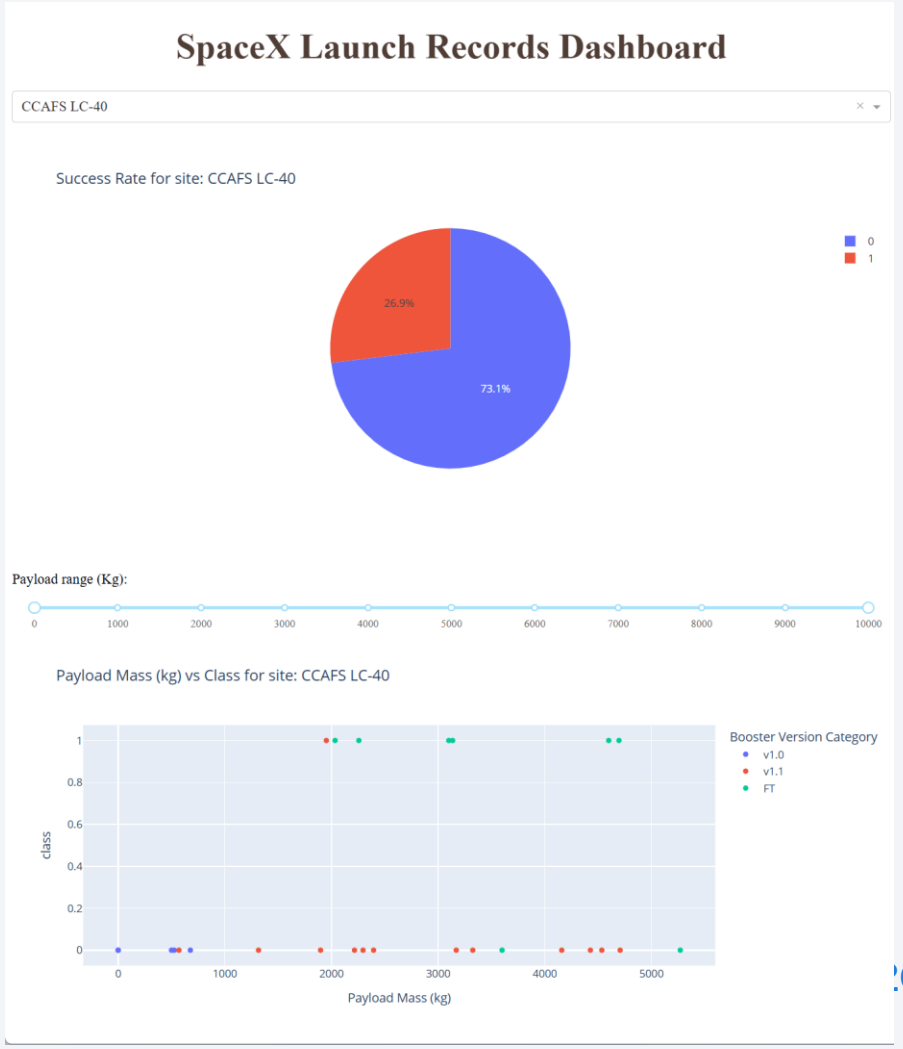


# Exploratory Data Analysis Results

- There are relationships between these features from scatter plot:
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - Success rate of each orbit type
  - Flight Number and Orbit Type
  - Payload Mass and Orbit Type

# Interactive Analytics Results



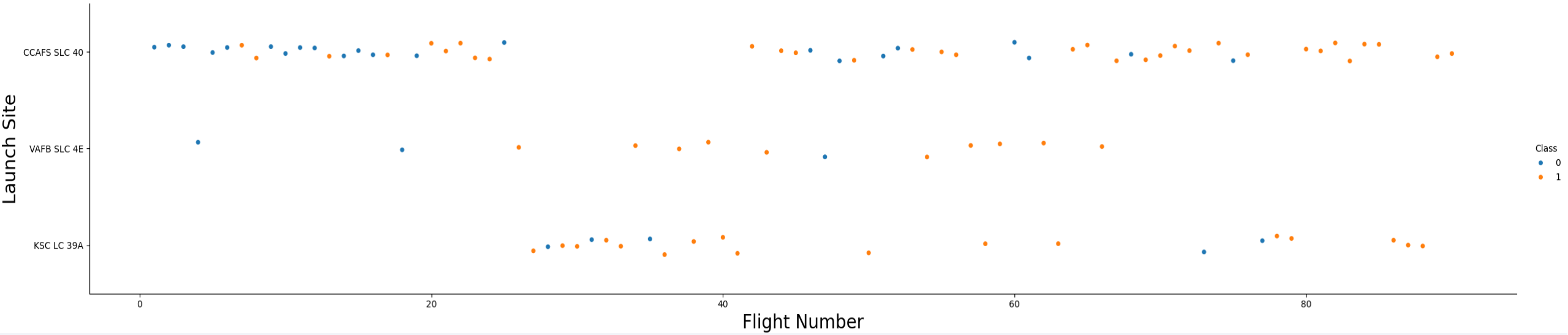
# Predictive Analysis Results

- Decision tree produces the highest accuracy value around 0.8625 comparing with the others
- Score method produces the same value for all algorithm around 0.8333

Section 2

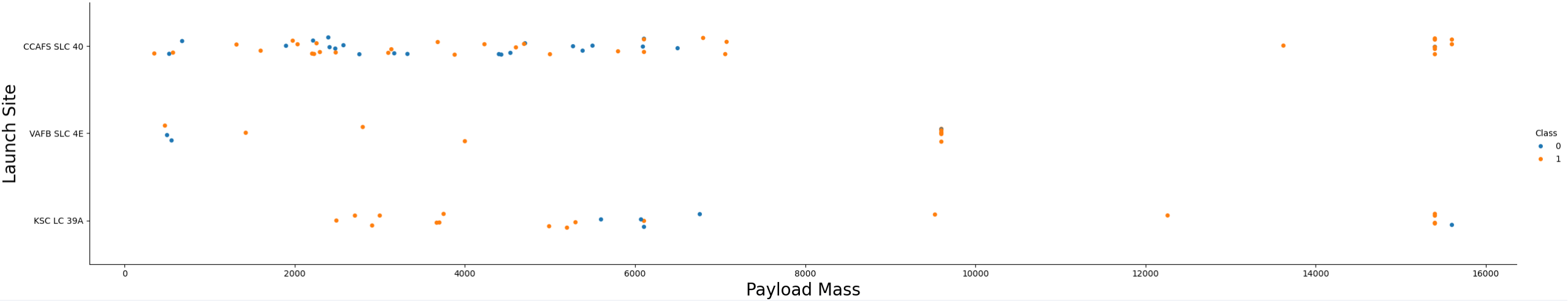
# Insights drawn from EDA

# Flight Number vs. Launch Site



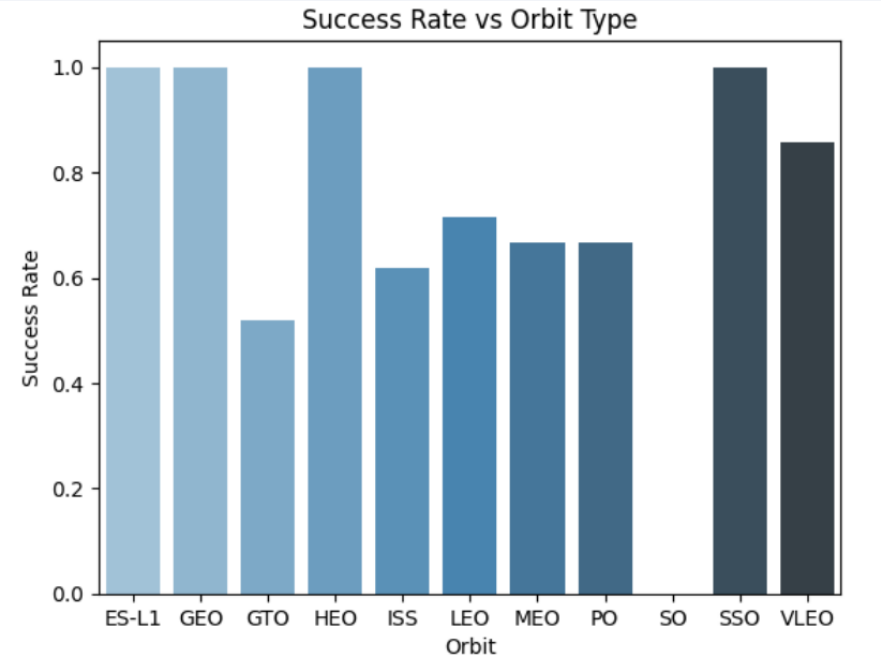

- CCAFS SLC-40 site has successful rate variation in early launch and it is the most used launch site
- VAFB SLC 4E is the site with the most minimum launch however it has the higher success ratio than failed from its flight
- KSC LC 39A is the site used after several other site launched and has higher success ratio than failed from its flight

# Payload vs. Launch Site



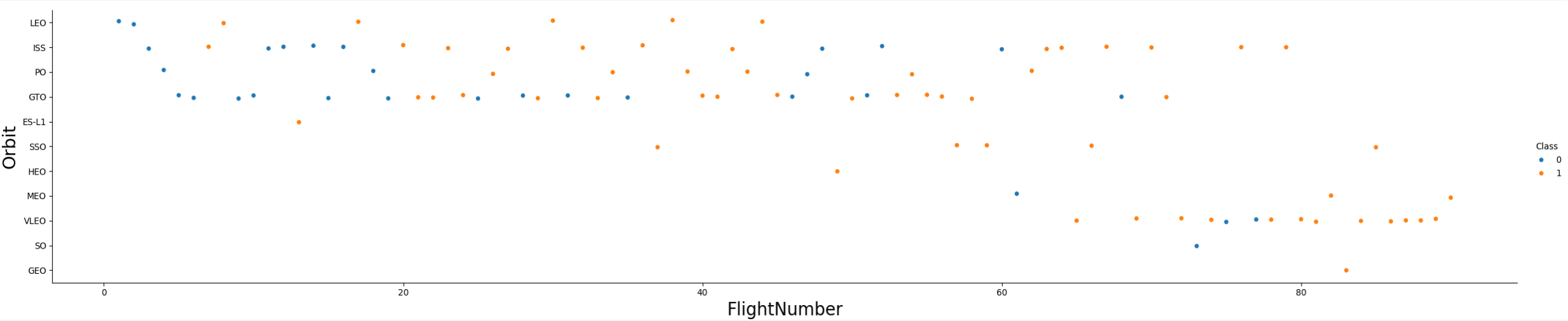

- There are no rockets launched from VAFB-SLC for heavypayload mass greater than 10000
- For payload above 8500 the three sites choosen for the flight except VAFB SLC 4E is used only for around 9500
- CCAFS SLC 40 is the site most used for payload between around 500 and 7000

# Success Rate vs. Orbit Type



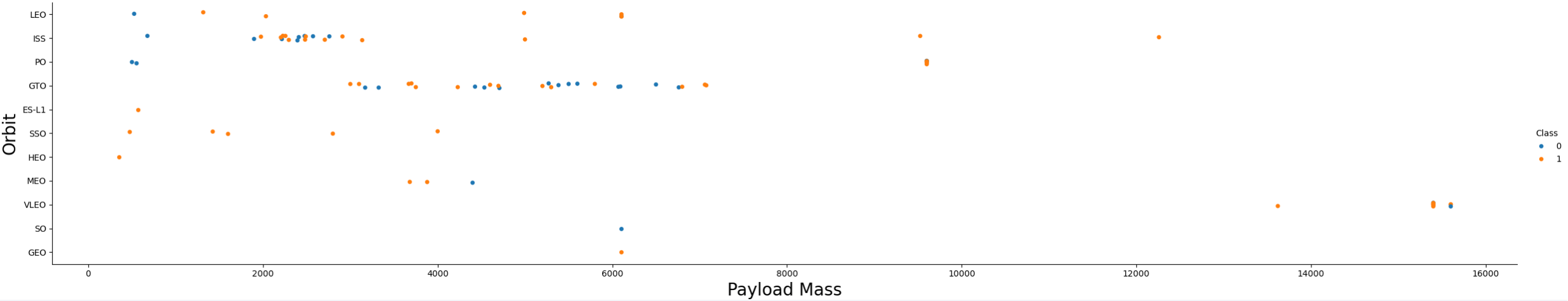

- ES-L1, GEO, HEO, and SSO are the orbit type which has most high successful rate than others
- GTO is the orbit type which has most lower successful rate from the others

# Flight Number vs. Orbit Type




- In the LEO orbit success seems to be related to the number of flights
- There are no relationship between flight number and success rate in the GTO orbit

# Payload vs. Orbit Type




- ISS, LEO, and PO has successful landing for heavy payloads
- For GTO the successful ratio is around 50:50

# Launch Success Yearly Trend


- Zero successful rate between 2010 and 2013
- Increasing successful rate from 2013 till 2020

# All Launch Site Names

- The unique launch sites are:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC 39A
  - CCAFS SLC 40
- The results above generate from sql distinct function
- Query: select distinct(Launch_Site) from SPACEXTABLE

# Launch Site Names Begin with 'CCA'

- Launch sites name begin with ‘CCA’ generated by where clause using like predicate and % after ‘CCA’ (‘CCA%’)
- Limit 5 used as predicate to limit the record numbers produced
- The results are 5 records from date in (2010-06-04, 2010-12-08, 2012-05-22, 2012-10-08, 2013-03-10)
- Query: select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5

# Total Payload Mass

- The total payload carried by boosters from NASA is 45596
- Sum function is used to calculate the total payload
- The record filtered by where clause for NASA only
- Query: select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)'

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4
- Avg function is used to calculate the average payload
- The record filtered by where clause for F9 v1.1 only
- Query: select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1’

# First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 2015-12-22
- Min function is used to find the successful first date
- The record filtered by where clause for Success ground pad only
- Query: select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 generated by where clause with between predicate for 4000 and 6000 range
- There are 24 booster version with successful drone ship landing
- Query: select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ between 4000 and 6000

# Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes generated by using count function on mission outcome and the record grouped by mission outcome
- There are 4 mission outcome results:
  - Failure (in flight): 1 mission
  - Success: 98 mission
  - Success: 1 mission
  - Success (payload status unclear): 1 mission
- Query: select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome

# Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass generated by subquery technique.
- Payload mass from outer query assigned by inner query which produces max value for the payload
- Inner query row filtered for only both rows had the same booster version
- There are 12 booster version with the highest payload max around 15600 kg
- Query: select S1.Booster_Version, S1.PAYLOAD_MASS__KG_ from SPACEXTABLE as S1 where S1.PAYLOAD_MASS__KG_ = (select max(S2.PAYLOAD_MASS__KG_) from SPACEXTABLE S2 where S2.Booster_Version = S1.Booster_Version) \order by S1.PAYLOAD_MASS__KG_ desc

# 2015 Launch Records

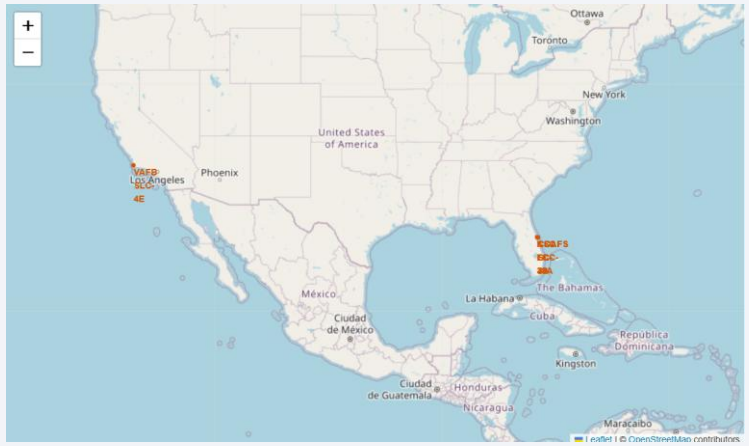
- To Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 produced by extracting the month and year from the date using substr function.
- The records filtered by where clause for 2015 year only and landing outcome is failure drone ship
- There are 2 failed landing in 2015:
  - 2015-01-10: Failure (drone ship) with F9v1.1 B1012 booster version
  - 2015-04-14: Failure (drone ship) with F9v1.1 B1015 booster version
- Query:select Date, substr(Date, 6,2) as month, Landing_Outcome, Booster_Version from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome='Failure (drone ship)'

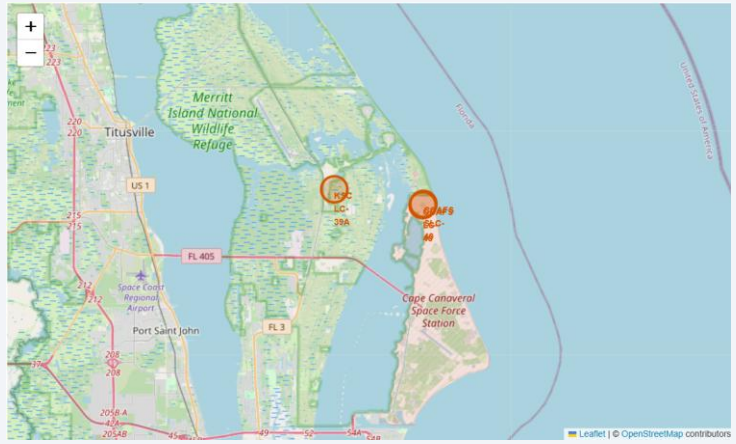
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- To 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 generated by:
  - Count function for landing outcome for the date between 2010-06-04 and 2017-03-20. The records grouped by landing outcome
  - Results ordered by count landing outcome using desc command
- The highest landing outcome is No attempt with 10 outcome
- The lowest landing outcome is Precluded (drone ship) with 1 outcome
- Query: select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count desc


Section 3
Launch Sites
Proximities Analysis

# All Launch Sites on Map

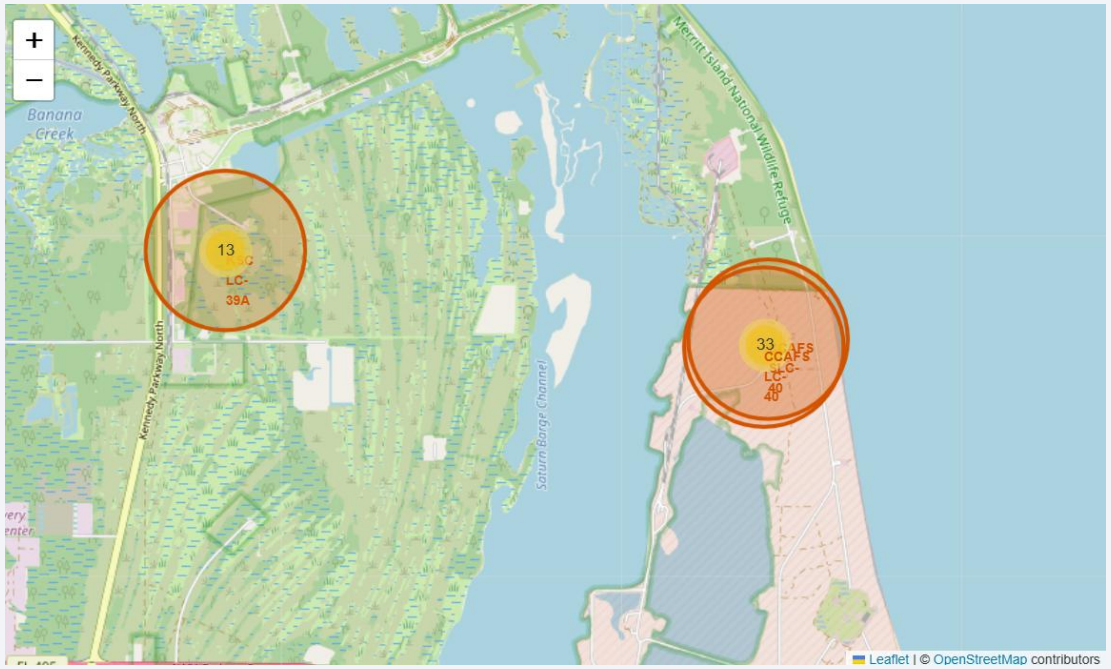


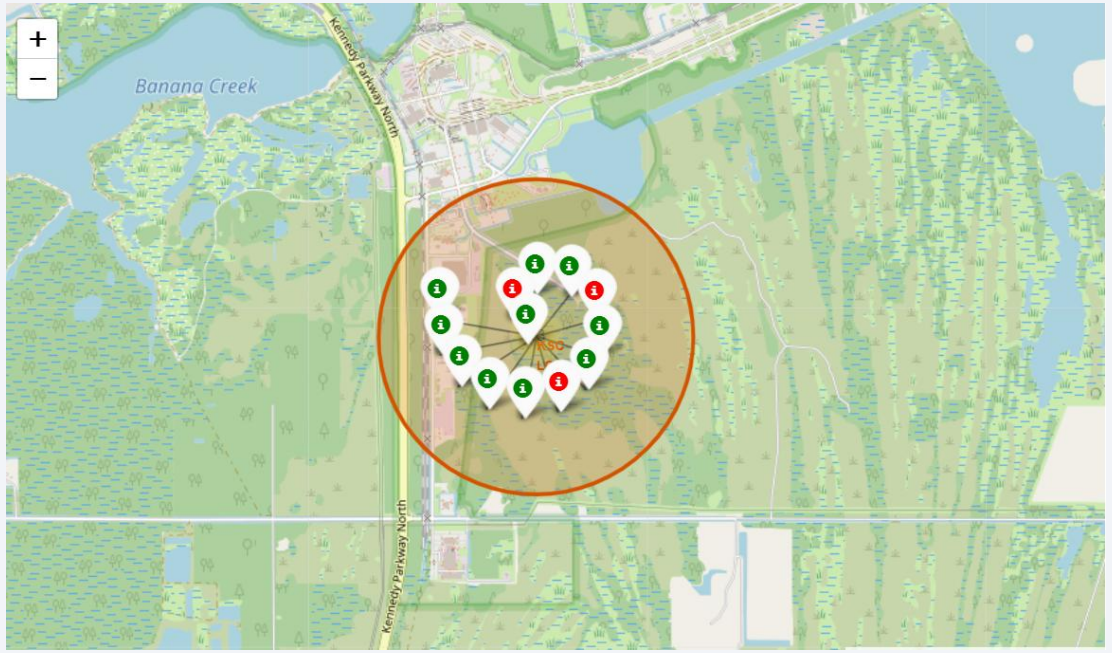





- Four launch sites mapped into map.
- Three sites on the right side map, the other one in the left side map
- All sites placed at the coast, far from highway, railway, and city

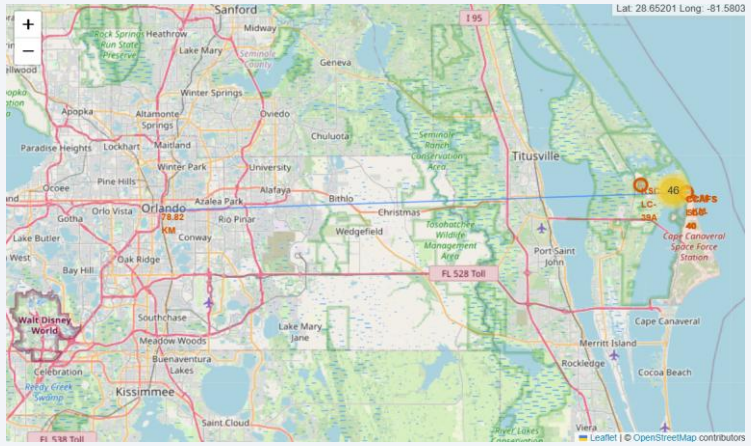
# Launch Outcomes on Map

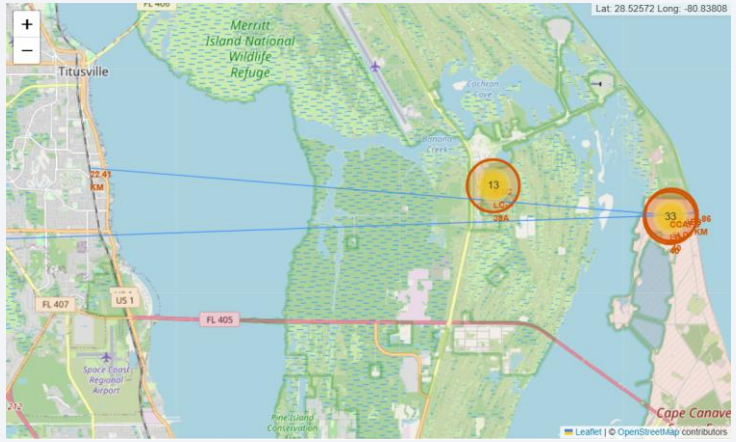


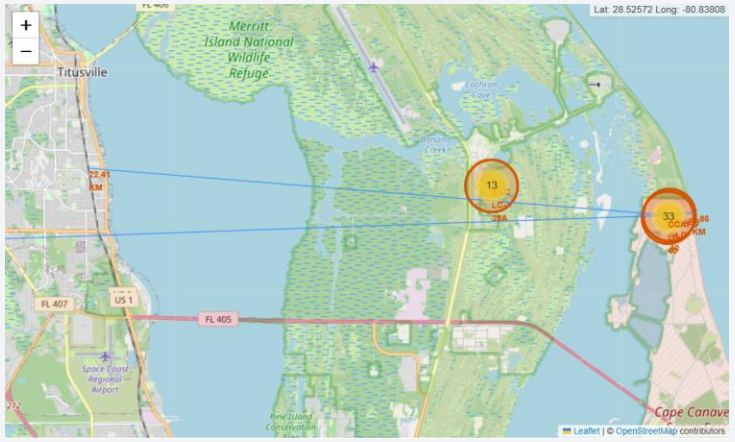




- All launch outcomes are mapped inside the circle marker
- Each circle marker shows the numbers of the launch outcomes
- Success outcome show by green circle and red circle for failure outcome

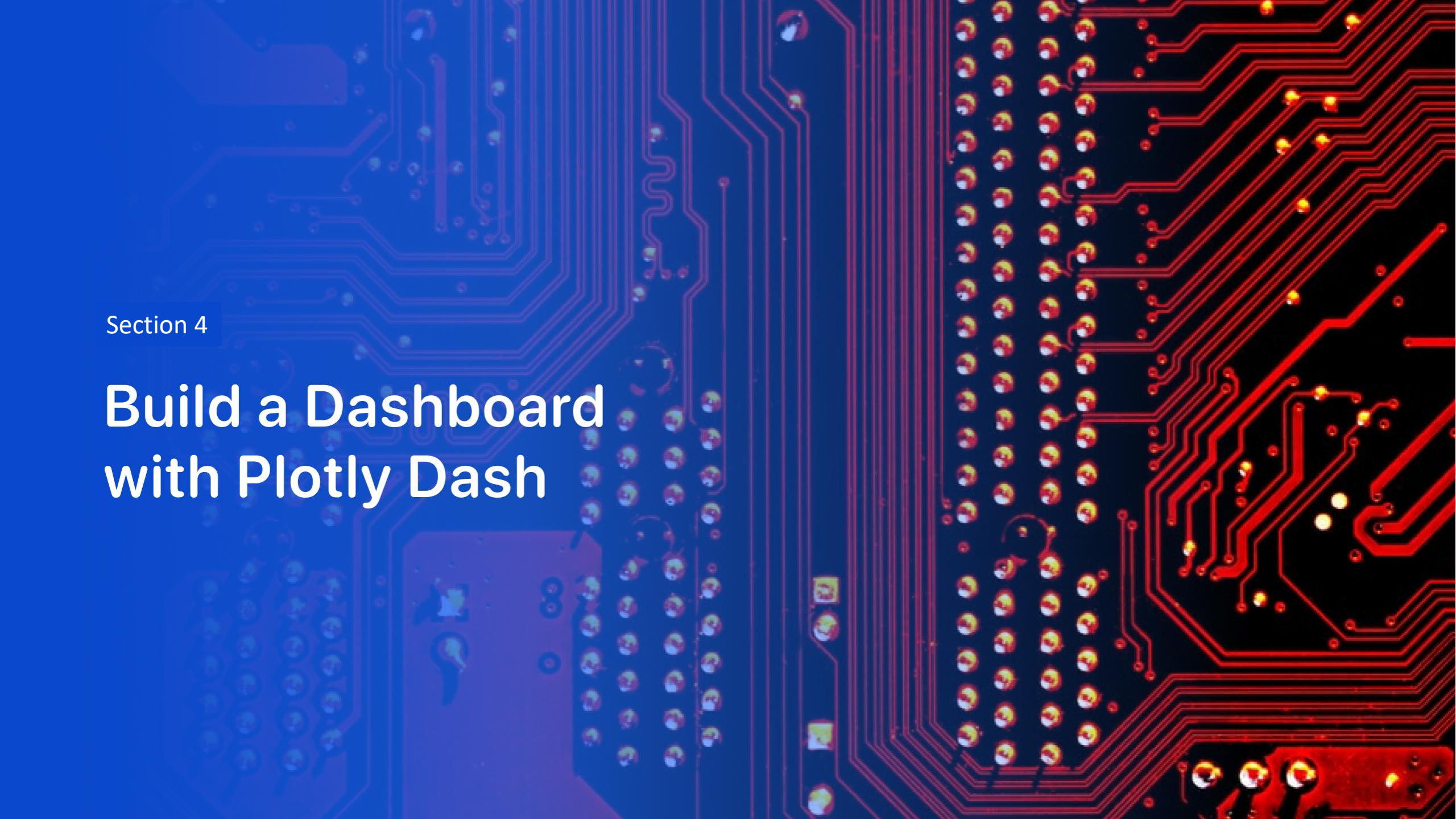
# A Launch Site Proximities



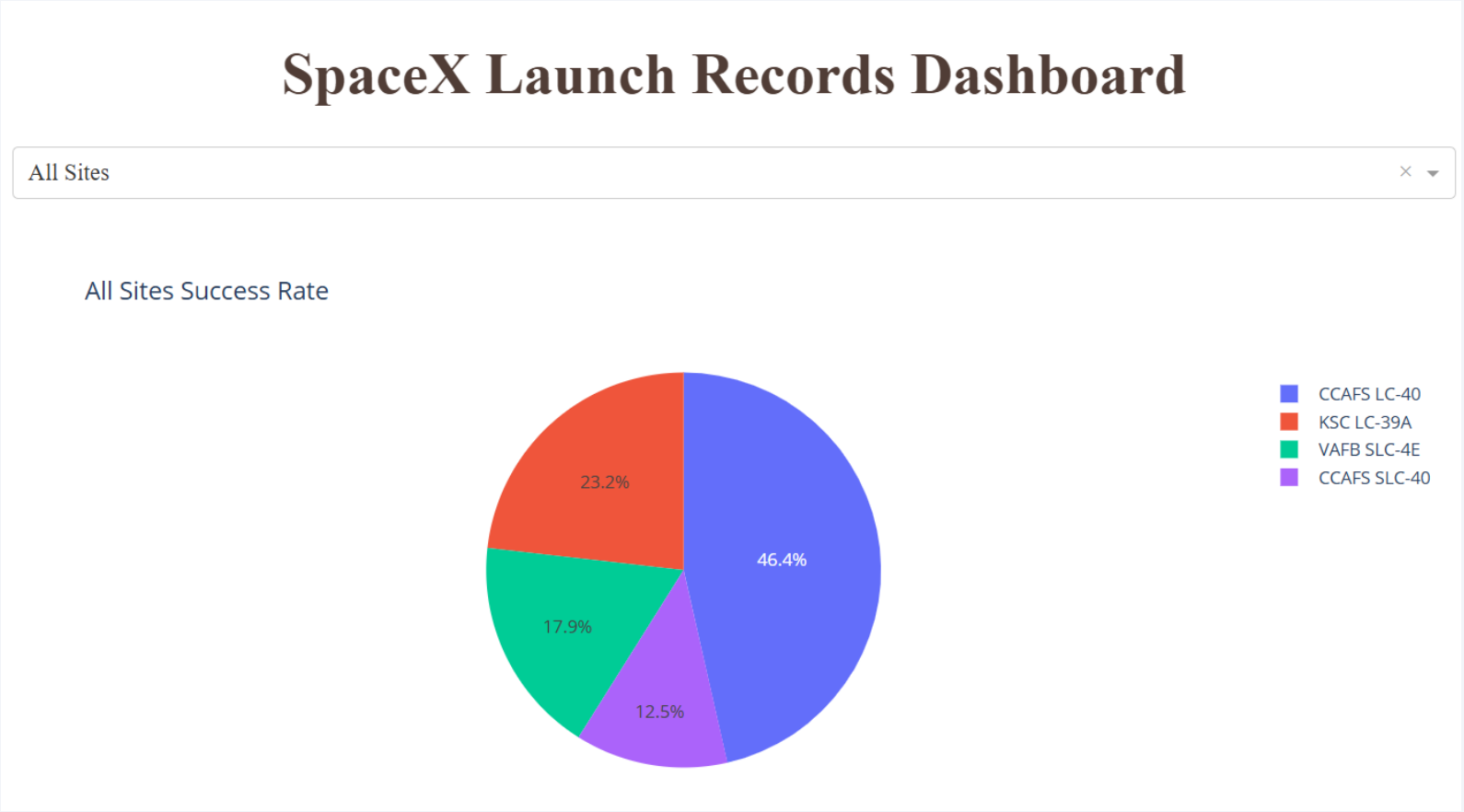




- Left map shows the proximities between selected site and Orlando city. The distances around 78.81 km
- Map at the middle shows the proximities between selected site and railway. The distances around 22.4 km
- Right map shows the proximities between selected site and highway. The distances around 22.29 km


Section 4
Build a Dashboard with Plotly Dash

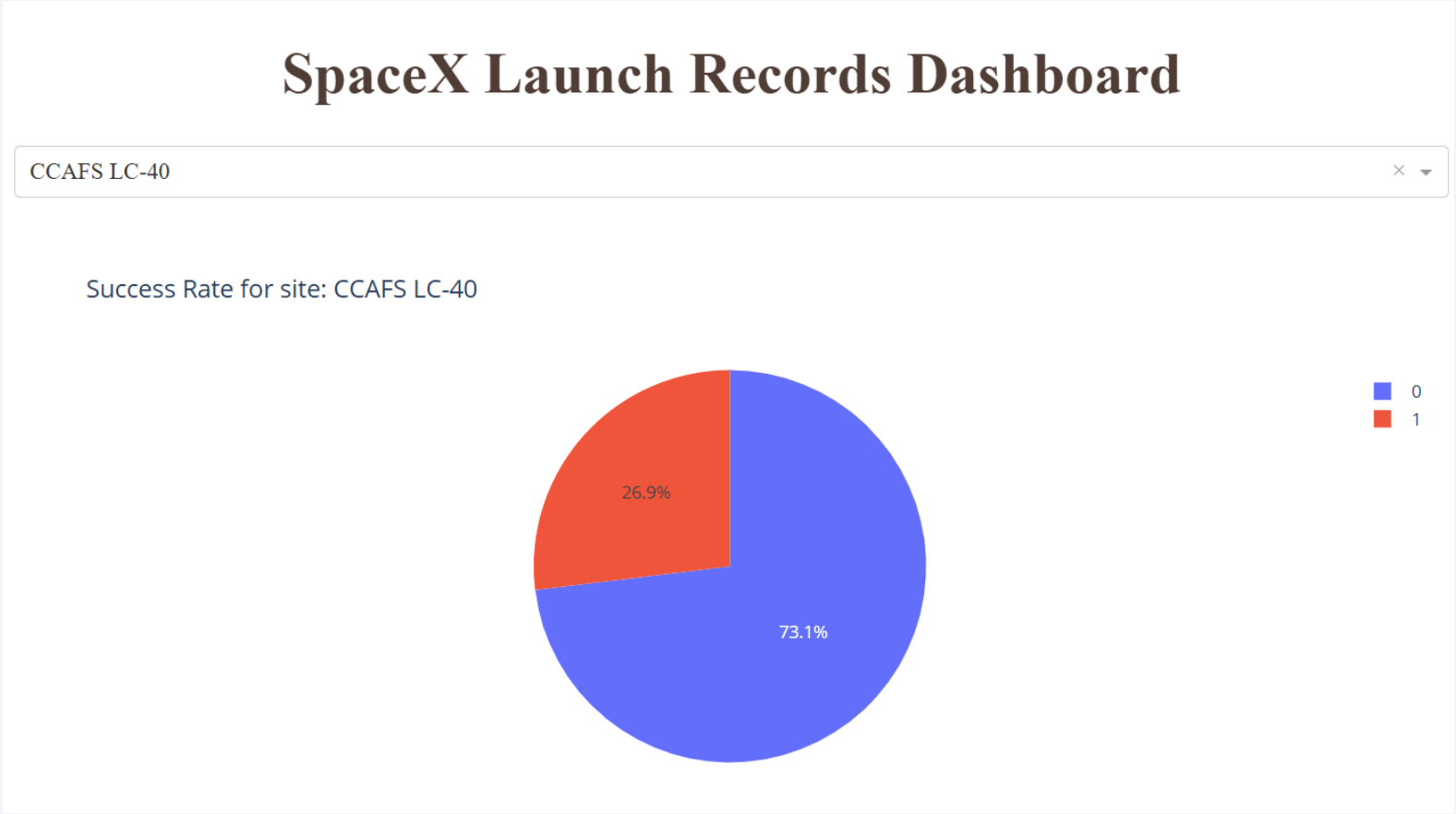
# Pie Chart (All Sites Success)

- CCAFS LC-40 site has the highest success rate
- CCAFS SLC-40 site has the lowest success rate



# Pie Chart (Highest Success Ratio Site)

- CCAFS LC-40 site has the ratio 73.1% for success and 26.9% for failure



# Scatter Plot (Payload vs Success Rate)

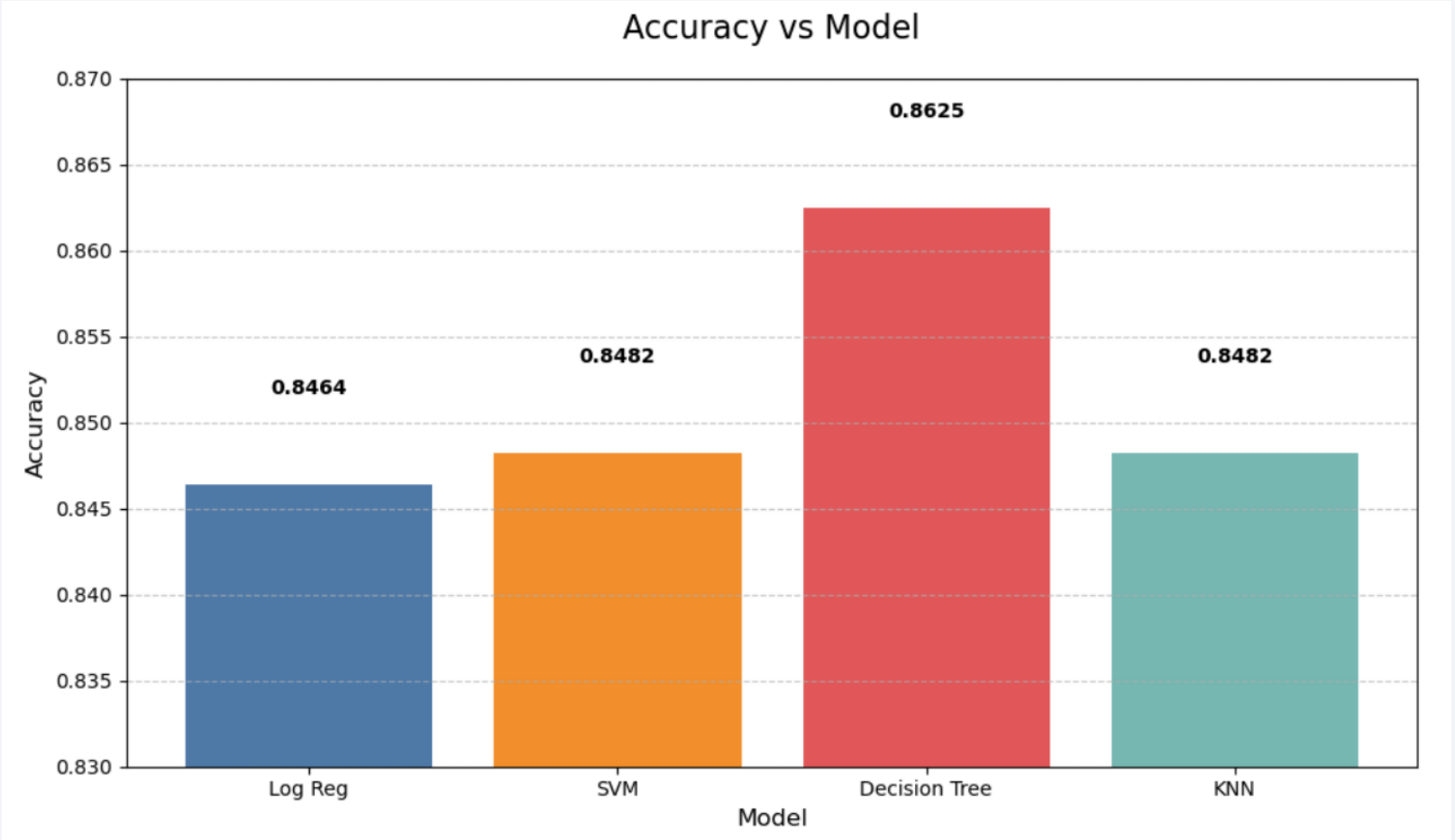


- Left Scatter plot shows the success rate for each booster version with all payload values
- Right scatter plot shows the success rate for each booster version within 0-5000 range payload

Section 5

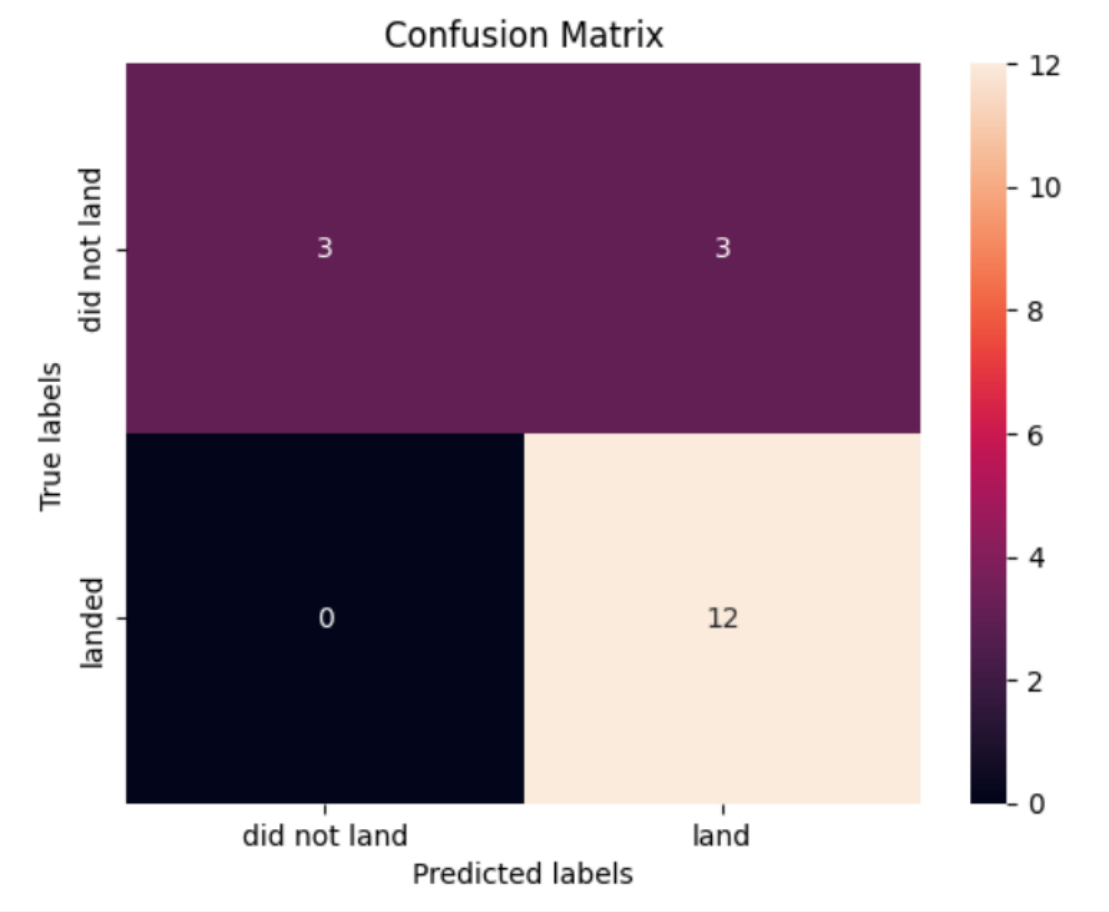
# Predictive Analysis (Classification)

# Classification Accuracy




- Decision tree model has the highest accuracy around 0.8625

# Confusion Matrix




- Confusion matrix from decision tree model has 3 false positive with 3 land fall into did not land

# Conclusions

- CCAFS SLC 40 is the launch site which has highest number launch around 55 launch
- ISS is the orbit type which has the highest accessed orbit
- First date of successful landing with ground pad is 2015-12-22
- There are around 100 success outcomes and 2 failed outcomes
- All launch sites are at the coastline, far away from city, railway, and highway
- CCAFS LC-40 has the highest successful rate
- Decision tree is the best model for predicting successful landing. It has highest accuracy value from the other models
- Increasing success rate over the time

# Appendix

- SQL Queries:
  - Task 1: select distinct(Launch_Site) from SPACEXTABLE
  - Task 2: select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
  - Task 3: select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)'
  - Task 4: select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1'
  - Task 5: select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
  - Task 6: select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ between 4000 and 6000
  - Task 7: select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome

# Appendix

- SQL Queries:
  - Task 8: select S1.Booster_Version, S1.PAYLOAD_MASS__KG_ from SPACEXTABLE as S1 where S1.PAYLOAD_MASS__KG_ = (select max(S2.PAYLOAD_MASS__KG_) from SPACEXTABLE S2 where S2.Booster_Version = S1.Booster_Version) order by S1.PAYLOAD_MASS__KG_ desc
  - Task 9: select Date, substr(Date, 6,2) as month, Landing_Outcome, Booster_Version from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome='Failure (drone ship)'
  - Task 10: select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count desc

# Thank you!