DATA SCIENCE JOURNEY

Xuefei Meng

2023/07/21

OUTLINE

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

EXECUTIVE SUMMARY

- In this project, I learned and review a lot of Data Science knowledge and coding skills
  - For Coding, learn and review to using Python and SQL on data
  - Learn and review Knowledge of data collection/data wrangling/Exploratory Data Analysis/Visual Analytics/Machine Learning(Predictive Analysis)
- This project use the SpaceX landing data
- In this project,
  - Using python and SQL as the basic coding lanauge
  - Using data Science knowledge on SpaceX landing data to finish the data collection/wrangling/Analysis
- After data analysis finish, EDA with visualization results/EDA with SQL results/predictive analysis (classification) results will be giving base on SpaceX landing data and its analysis.

INTRODUCTION

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. This project use the SpaceX landing dataTherefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- In this project, we are trying to answer was whether for a given set of characteristics (including payload mass, orbit type, launch location, etc.) for a Falcon 9 launch, would the first stage of the rocket land successfully?

METHODOLOGY

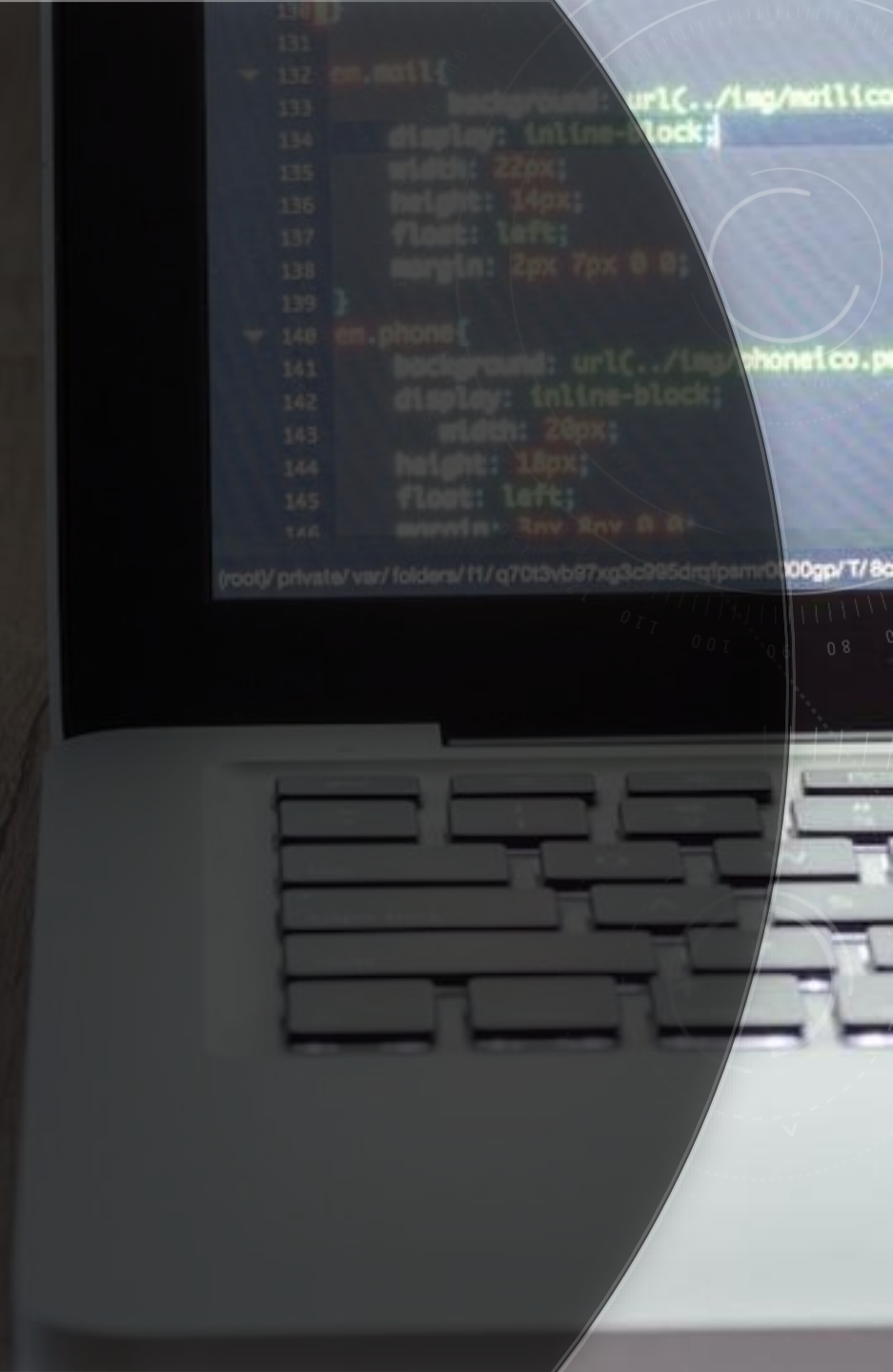
- 1 .Data collection
  - SpaceX API
  - Web scraping
- 2. Data wrangling
- 3. Exploratory data analysis (EDA) with SQL
  - Pandas and NumPy
  - SQL
- 4. Exploratory data analysis (EDA) with Data visualization
  - Matplotlib and Seaborn
  - Folium
- 5. Machine learning prediction
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

METHODOLOGY--DATA COLLECTION

- Data collection is the process of gathering raw data from various sources in order to analyze and derive insights from it. In this project, we make a get request to the SpaceX API and performing web scraping to collect Falcon 9 historical launch records
- Identify data sources:
  - We get SpaceX API from "https://api.spacexdata.com/v4/launches/past“
  - We get HTML table from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Gather data
  - SpaceX API columns: rocket; success; failures; payloads; launchpad; flight_number; name; date_utc; date_unix; date_local; date_precision upcoming; cores; auto_update; tbd; launch_library_id; id; fairings.reused; fairings.recovery_attempt; fairings.recovered; fairings.ships
  - HTML columns: Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome’
- Data quality assessment: Evaluate the quality of the collected data.

METHODOLOGY--DATA COLLECTION

- SpaceX API:
- GitHub Link: Spacex data collection API
- Steps:
  - Request (Space X APIs)
  - Check the content of the response
  - Make the requested JSON results
  - Json_normalize to DataFrame data from JSON
  - Get information about the launches using the IDs given for each launch
  - Combine the columns into a dictionary
  - Create a Pandas data frame from the dictionary
  - Dealing with Missing Values
- Result: We end up with 90 rows or instances and 17 columns or features.



METHODOLOGY--DATA COLLECTION

•Web Scraping:

•GitHub Link: Spacex Webscraping

- Steps:
  - Request the Falcon9 Launch HTML page
  - Create a BeautifulSoup object from the HTML response
  - Extract all info from the HTML table
  - Create dictionary
  - Iterate through table cells to extract data to dictionary
  - Converted dictionary into a Pandas dataframe
- Result: We end up with 90 rows or instances and 17 columns or features.

METHODOLOGY--DATA WRANGLING

•Data wrangling, also known as data cleaning or data preprocessing, refers to the process of transforming and preparing raw data for analysis.

•In project, we will Identify and calculate the percentage of the missing values in each attribute, calculate the number of launches for each site, calculate the number and occurence of mission outcome per orbit type, Create a landing outcome label from Outcome column

•GitHub Link: Data wrangling

METHODOLOGY--EXPLORATORY DATA ANALYSIS(EDA) WITH SQL

•Tool

- Github:EDA with sql
- Pandas and NumPy

•Pandas and NumPy are used for data manipulation, analysis, and processing, which includes:

- The number of launches on each launch site
- The number of occurrence of each orbit
- The number and occurrence of each mission outcome

- SQL

•The data is queried using SQL to get better understand the database:

- Load the SQL extension and establish a connection with the database
- Using SQL Python integration
- Display names of the unique launch sites in the space mission/total payload mass carried by boosters launched by NASA (CRS)/Display the average payload mass carried by booster version F9 v1.1

METHODOLOGY--EXPLORATORY DATA ANALYSIS(EDA) WITH DATA VISUALIZATION

- Tool
- GitHub Link:
  - Matplotlib:EDA with Visualization
- Matplotlib and Seaborn
- Matplotlib and Seaborn are visualization libraries in Python that provide powerful tools for creating visually appealing and informative plots and charts.
- The plots and charts are used to understand more about the relationships between several features, such as:
  - The relationship between flight number vs launch site
  - The relationship between payload mass vs launch site
  - The relationship between success rate vs orbit type

METHODOLOGY--INTERACTIVE MAP WITH FOLIUM

•Tool

•GitHub Link:

•Folium:Interactive Visual Analytics with Folium lab

- Folium

•Folium is a Python library used for creating interactive maps and visualizations.

•The Folium library is used to:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

METHODOLOGY--PLOTLY DASH DASHBOARD

•Tool

•GitHub Link:

- Plot Dashboard :spacex_dash_app.py
- Dashboard

•Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.

•Using a pie chart and a scatterplot, the interactive site shows:

- •The total success launches from each launch site
- •The correlation between payload mass and mission outcome (success or failure) for each launch site

METHODOLOGY--MACHINE LEARNING PREDICTION

- Using the Scikit-learn library to create our machine learning models.
- Github link:Machine Learning Prediction
- Steps:
  - Load data and split label column/Class from data
  - Standardize the data
  - Split the data into training and testing data
  - Create a Machine Learning Models:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree Classifier
    - K Nearest Neighbors (KNN)
  - Calculate the accuracy on the test data
  - Compare and Find the method performs best:

RESULTS

- EDA with SQL
- EDA with visualization
- interactive map with Folium
- Plot Dash dashboard
- predictive analysis (classification)

RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH SQL

- The names of the unique launch sites in the space mission

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- 5 records where launch sites begin with the string 'CCA'

Date	Time_(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_C
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (pa
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (pa
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No

RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH SQL

- The total payload mass carried by boosters launched by NASA (CRS)

total_payload_mass
45596.0

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

average_payload_mass
2534.6666666666665

- The date when the first successful landing outcome in ground pad was achieved.

First landing successful in group pad
01/08/2018

RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH SQL

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The total number of successful and failure mission outcomes

Outcomes	Total Number
Failed	40
Success	61

RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH SQL

- The names of the booster_versions which have carried the maximum payload mass

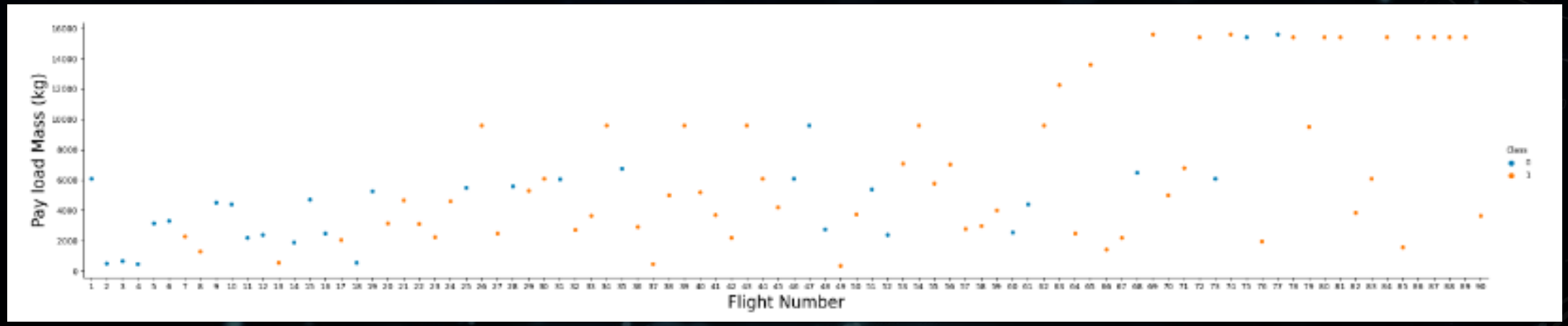
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

substr(Date, 4, 2)	Landing_Outcome	Booster_Version
10	Failure (drone ship)	F9 v1.1 B1012
04	Failure (drone ship)	F9 v1.1 B1015

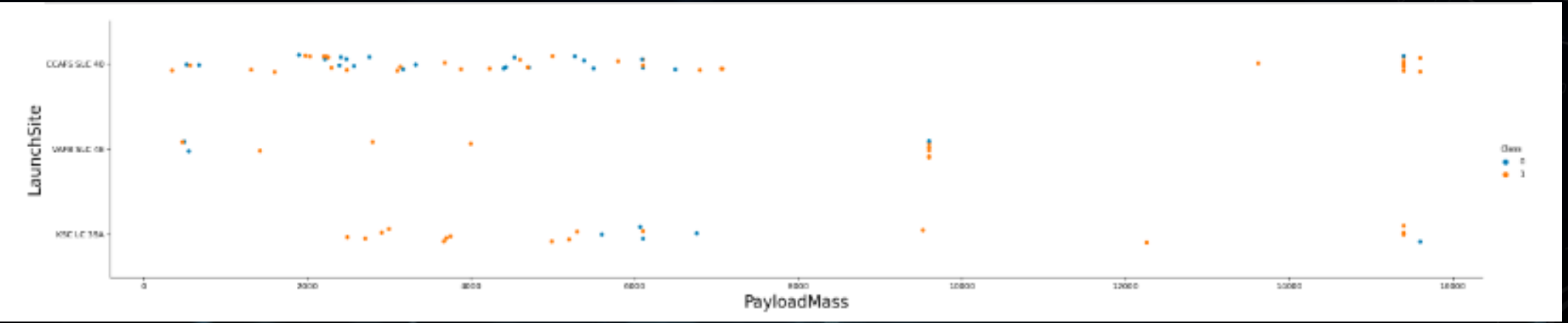
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- FlightNumber vs LaunchSite



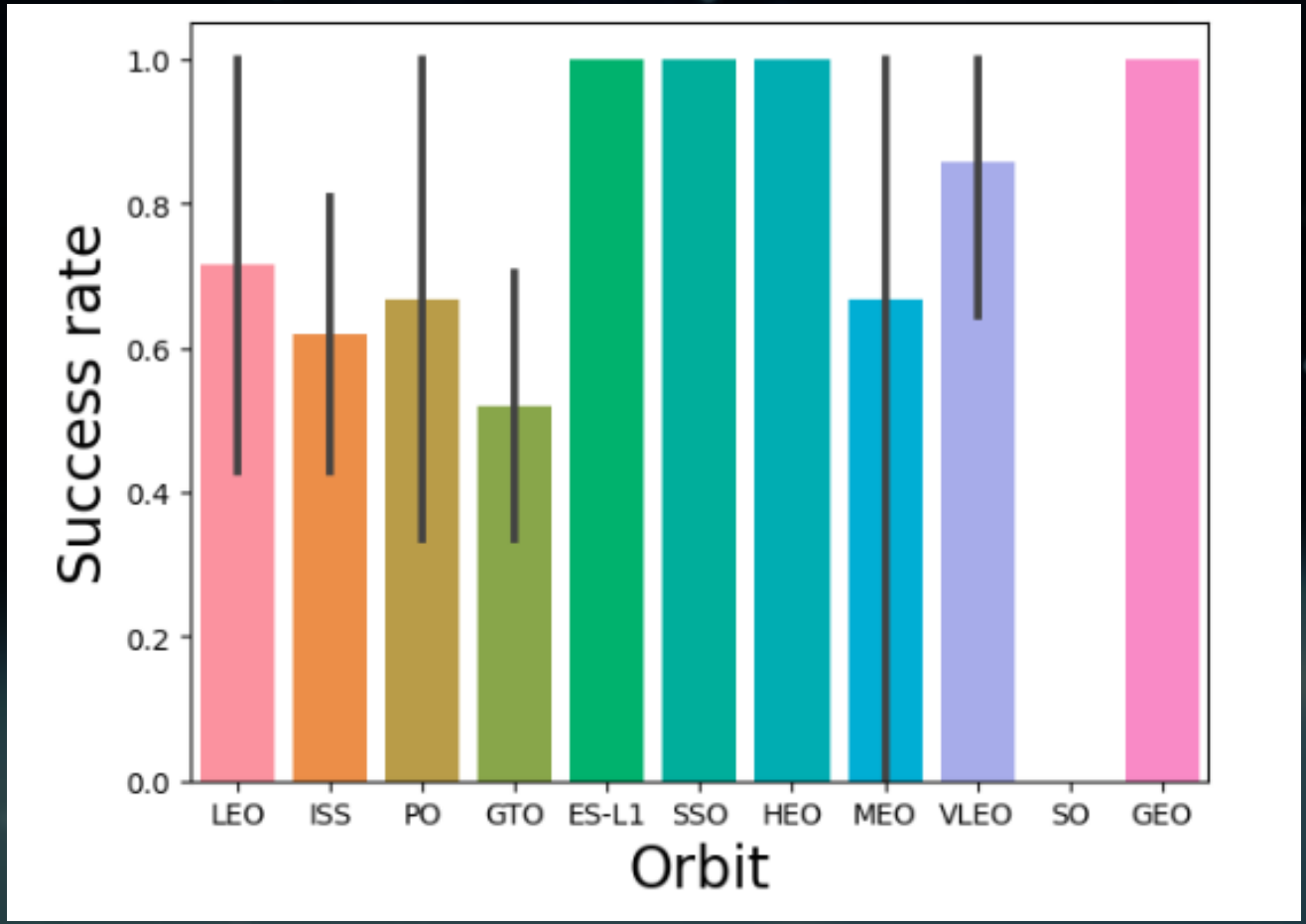
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- Payload vs Launch Site



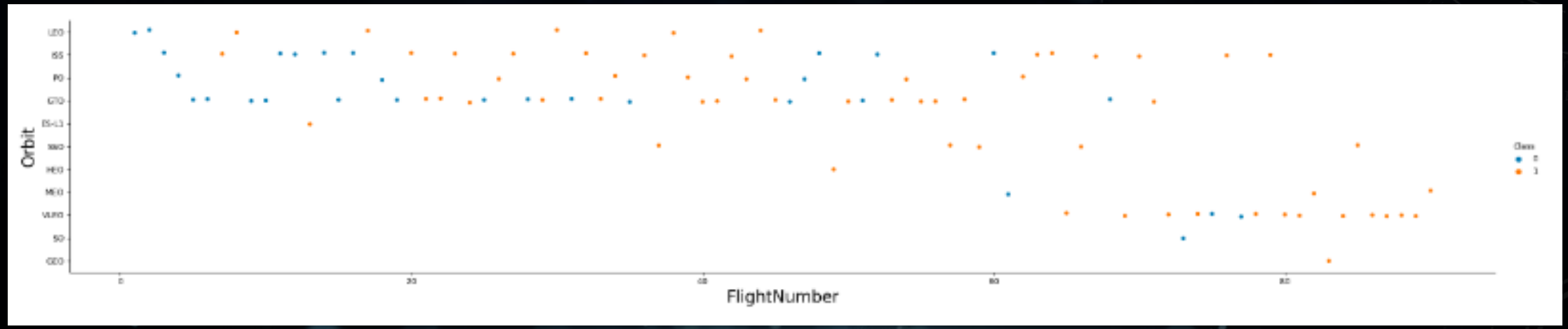
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- SuccessRate vs Orbit Type



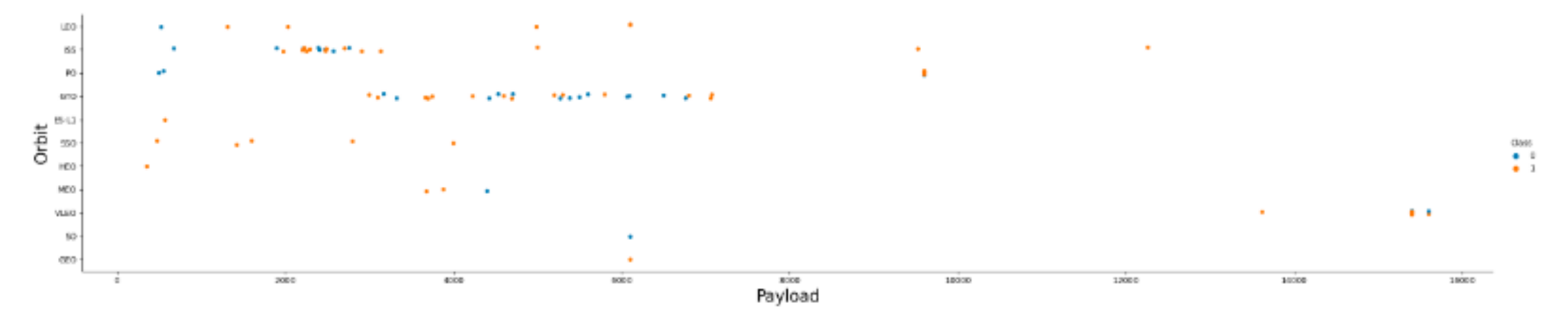
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- FlightNumber vs Orbit Type



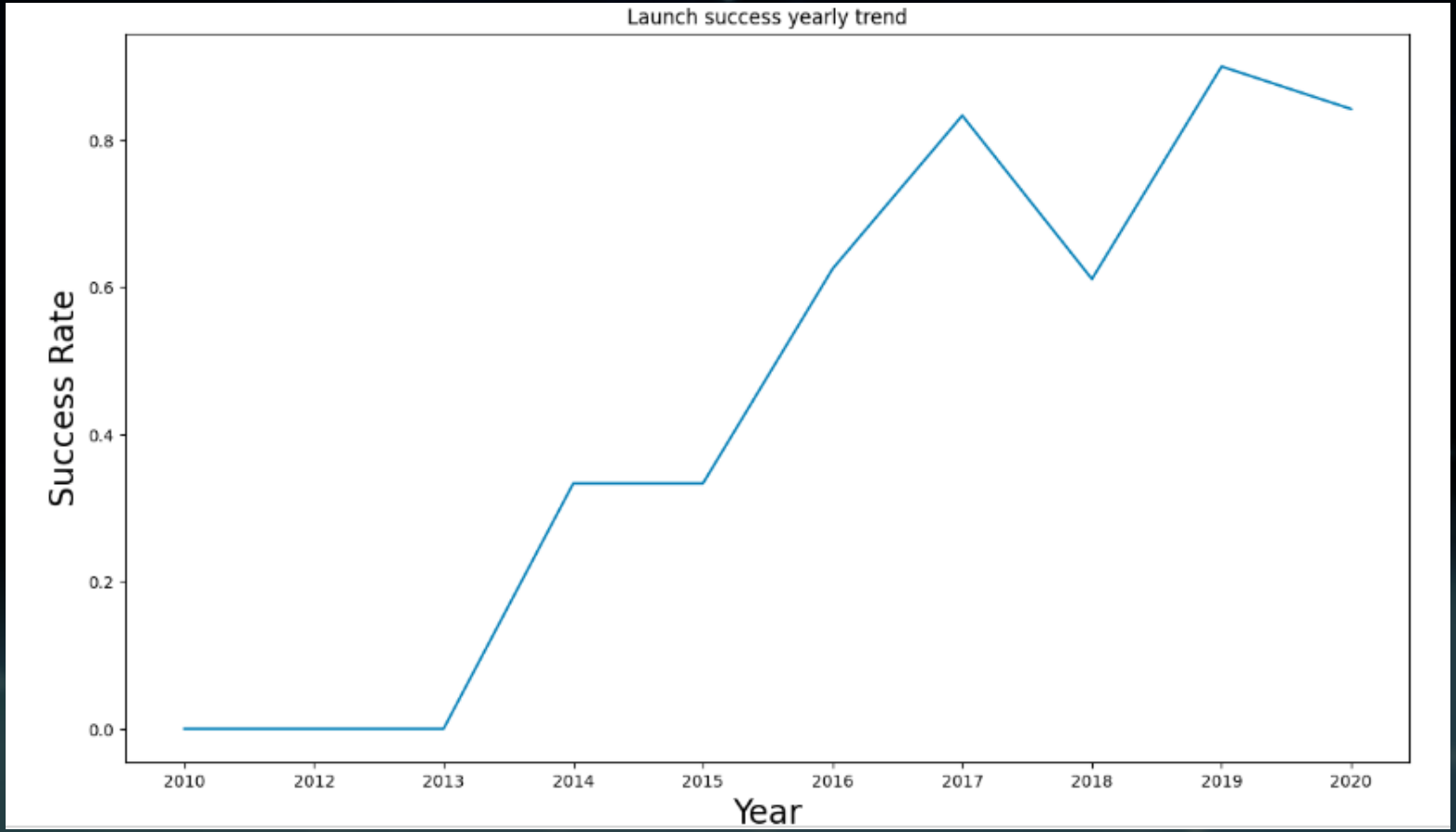
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- Payload vs Orbit Type



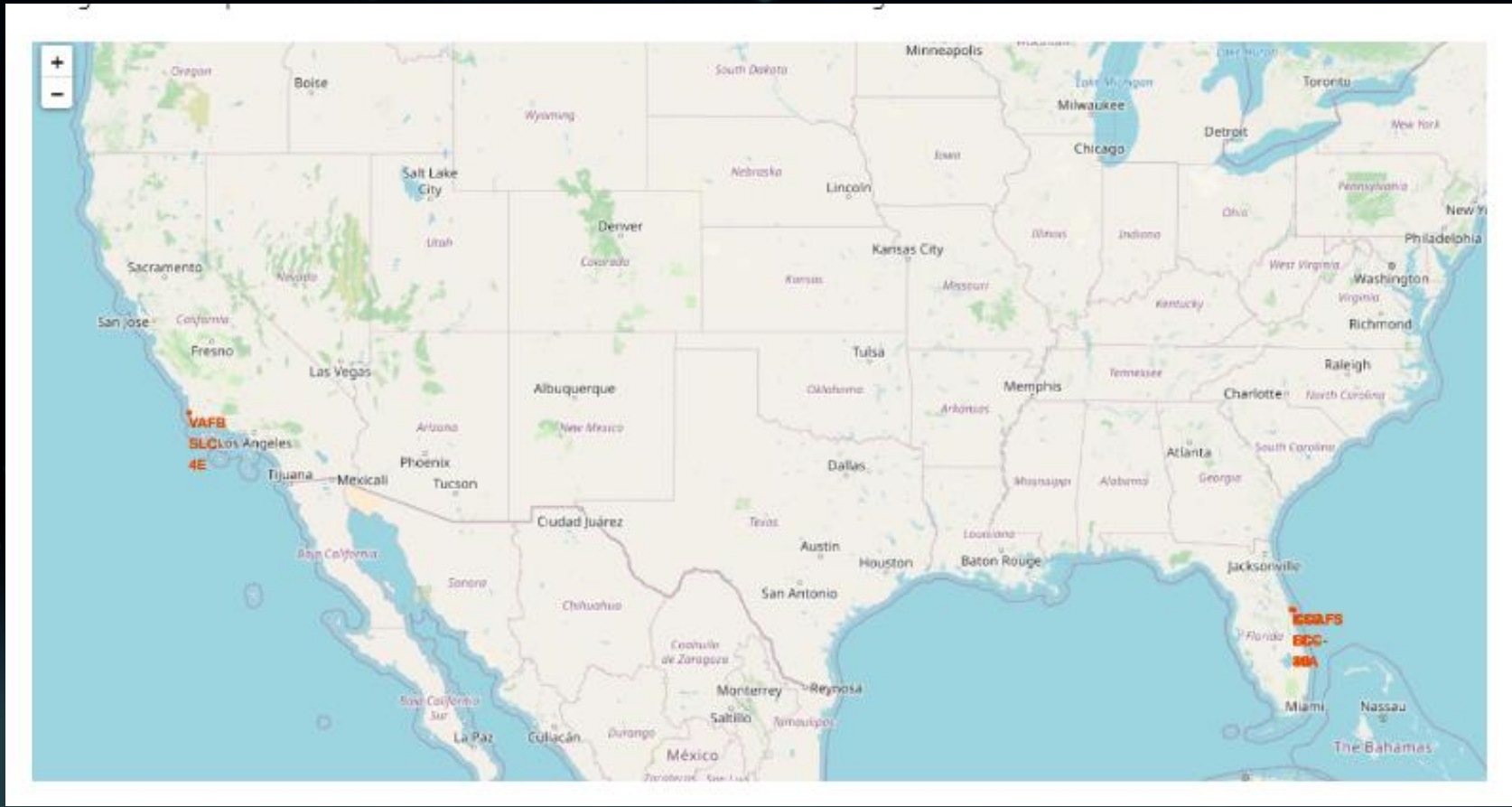
RESULT--EXPLORATORY DATA ANALYSIS(EDA) WITH VISUALIZATION

- The average launch success trend



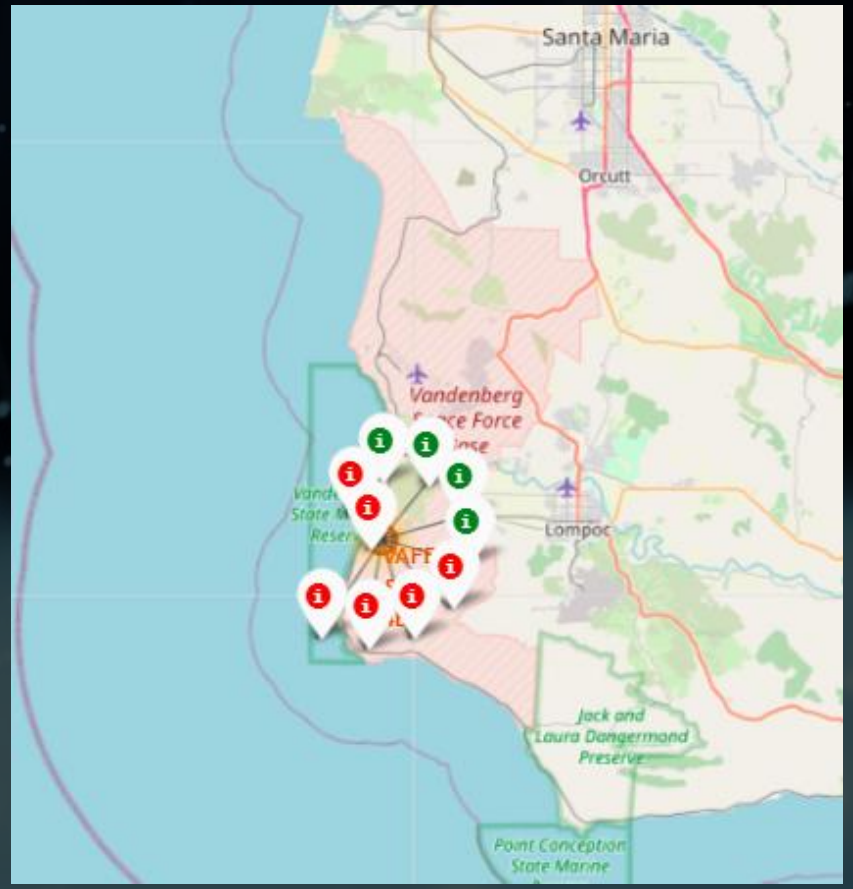
RESULT--INTERACTIVE MAP WITH FOLIUM

- Mark all launch sites on a map



RESULT--INTERACTIVE MAP WITH FOLIUM

- Mark the success/failed launches for each site on the map
  - When zoom in, the 10 points will show in green and red points

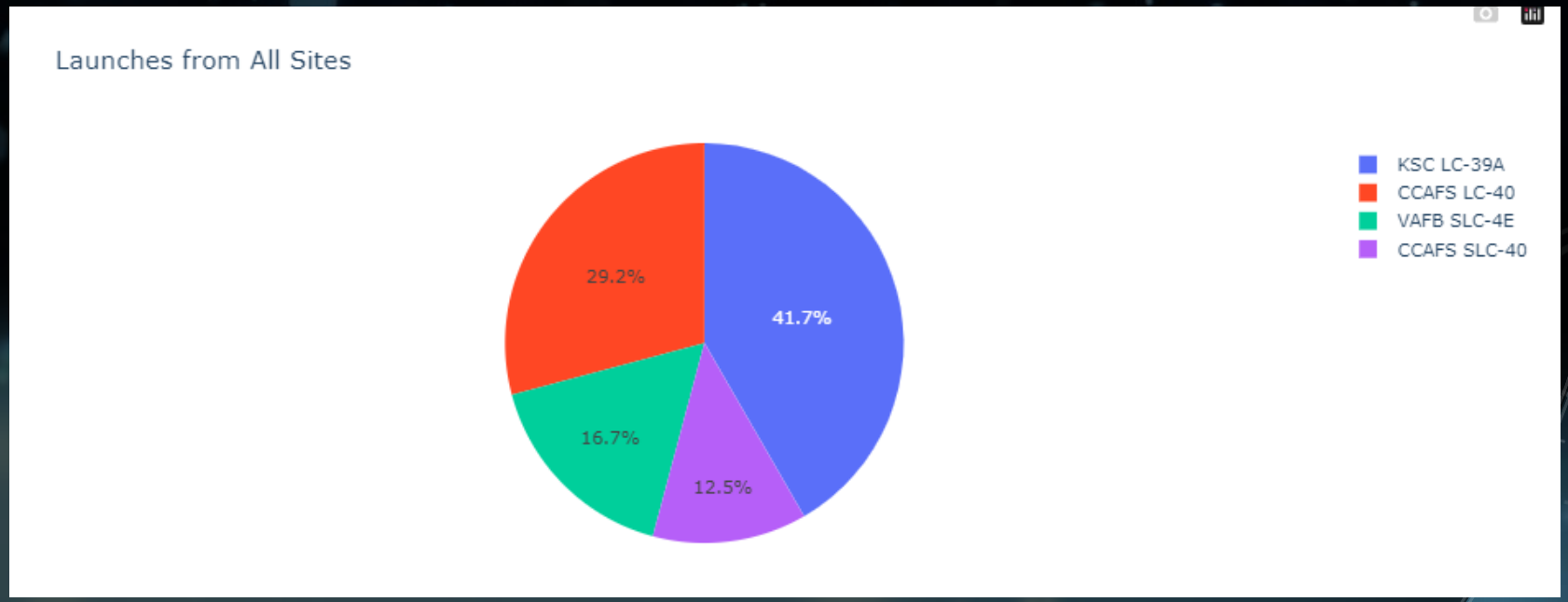


RESULT--INTERACTIVE MAP WITH FOLIUM

- Calculate the distances between a launch site to its proximities

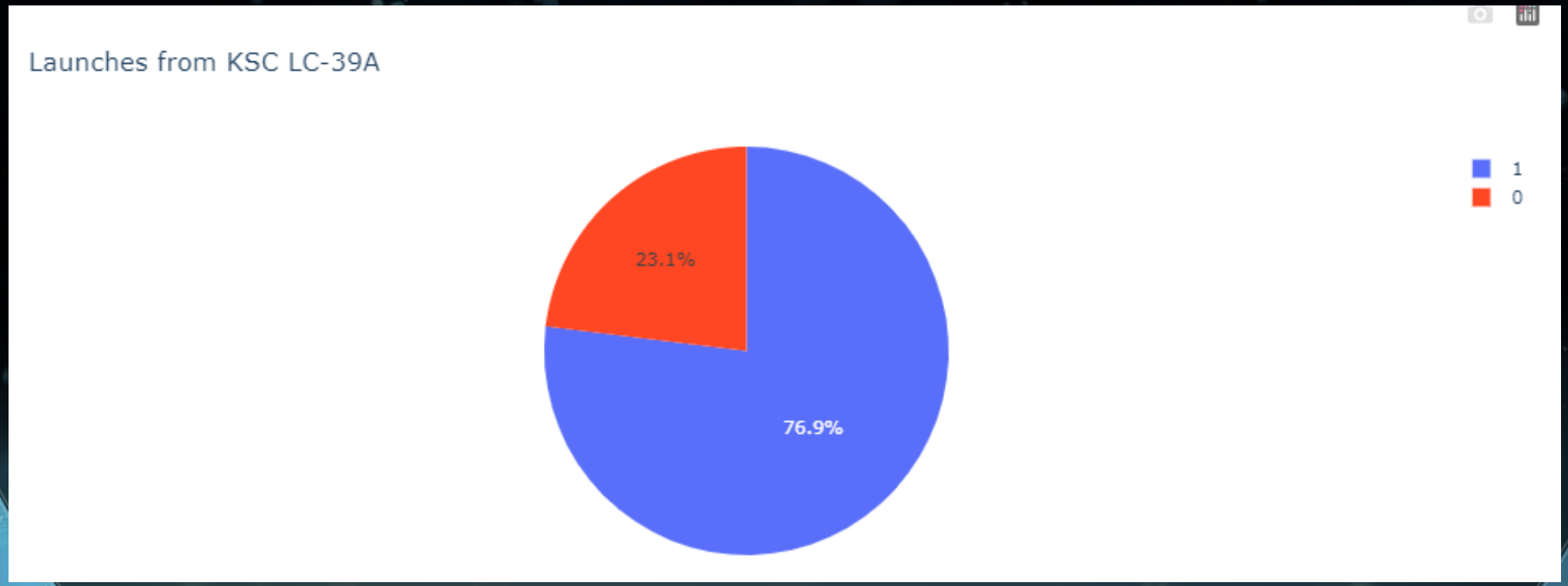
RESULT--PLOTLY DASH DASHBOARD

- Which site has the largest successful launches?
  - From the all site pie chart, we can know CCAFS LC 40 is the old name of CCAFS SLC 40 so CCAFS and KSC have the same amount of successful landings



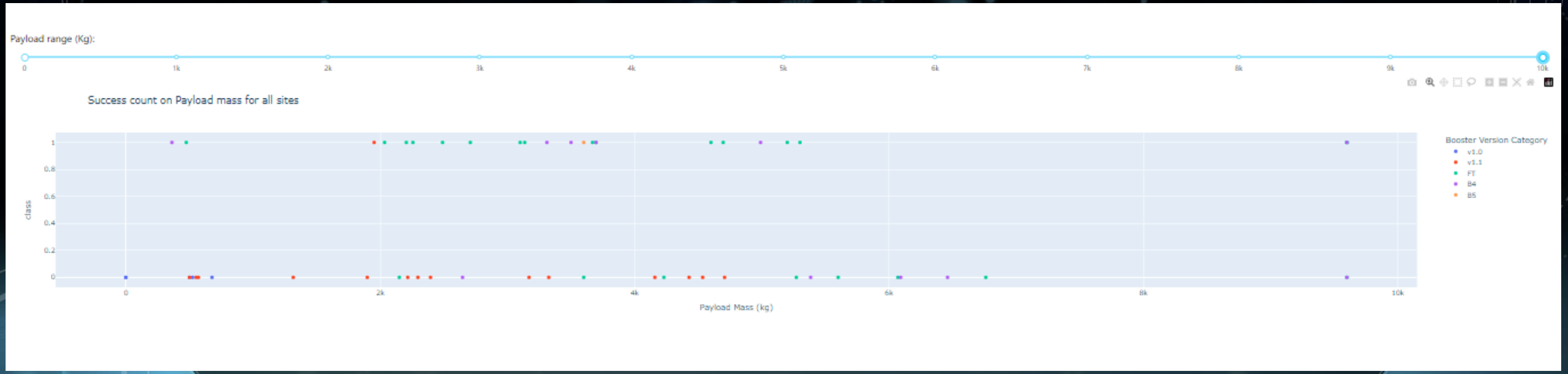
RESULT--PLOTLY DASH DASHBOARD

- Which site has the highest launch success rate?
  - Choosing theKS LC-39A, we got the largest successful launches whose successful rate is 76.9%(1 means successful)



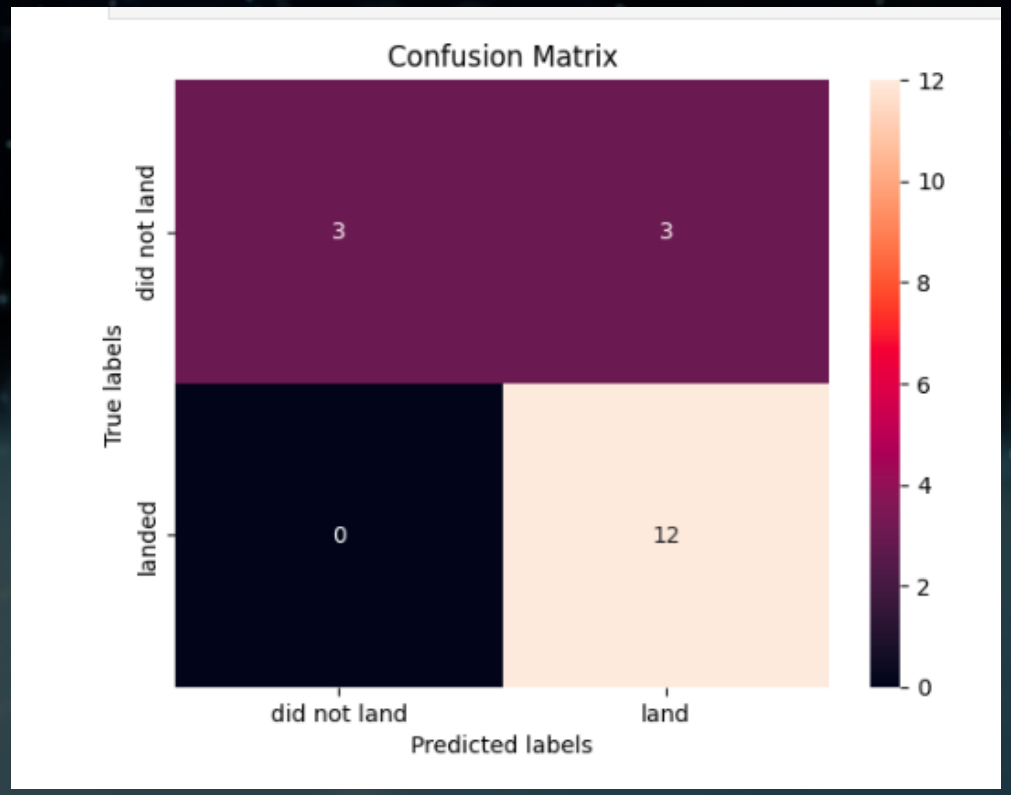
RESULT--PLOTLY DASH DASHBOARD

- payload range(s) vs success rate
  - Select the range is from 2k – 6k, we can got the highest launch success rate
  - Select the range is from 6k – 10k, we can got the lowest launch success rate



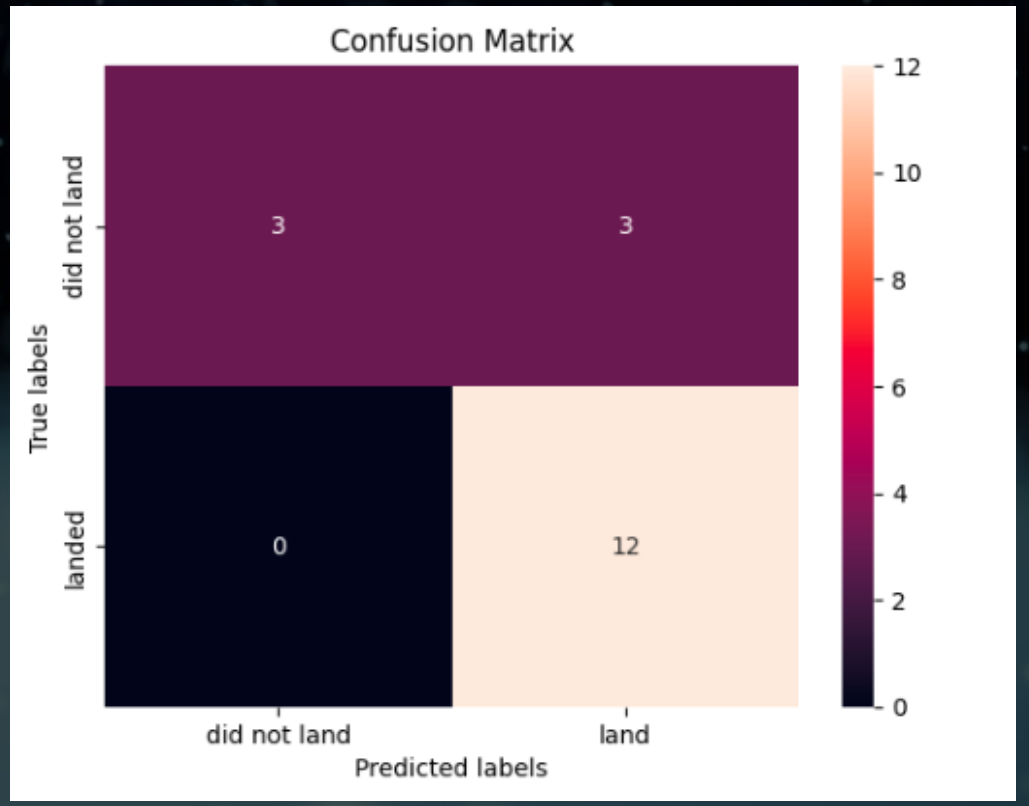
RESULT--PREDICTIVE ANALYSIS (CLASSIFICATION)

- Logistic Regression
  - The accuracy on the validation data: `0.8464285714285713`
  - The accuracy on the test data: `0.8333333333333334`
  - Confusion matrix:



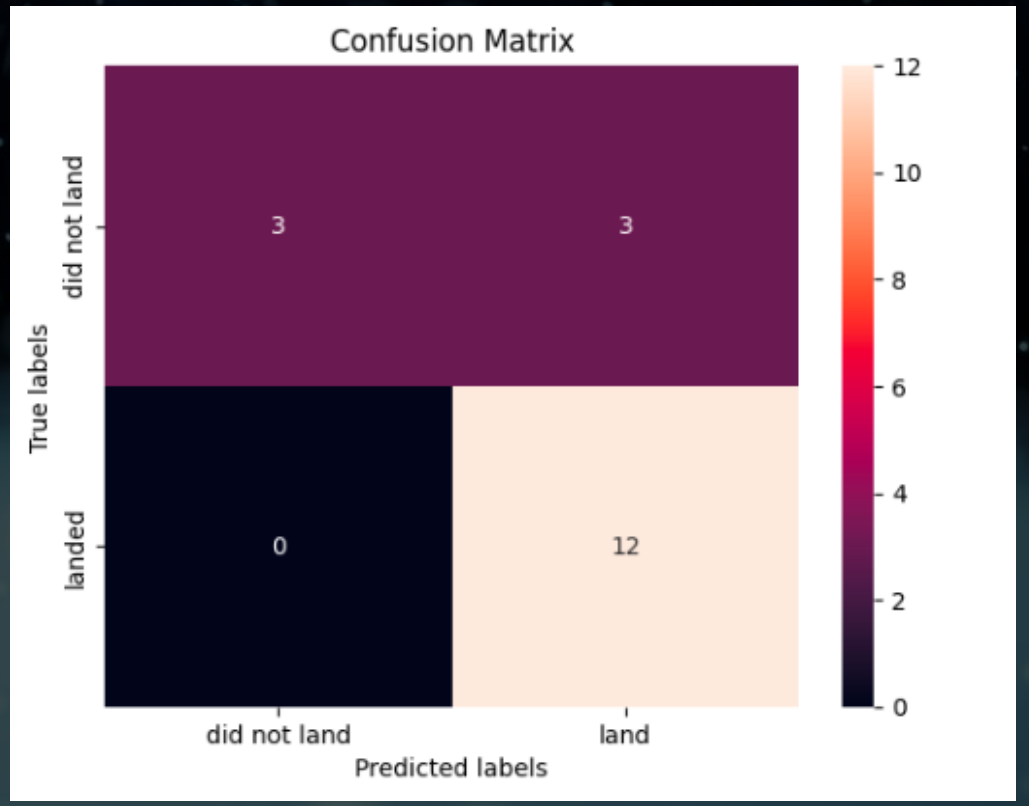
RESULT--PREDICTIVE ANALYSIS (CLASSIFICATION)

- Decision Tree Classifier
  - The accuracy on the validation data: 0.8892857142857142
  - The accuracy on the test data: 0.8333333333333334
  - Confusion matrix:



RESULT--PREDICTIVE ANALYSIS (CLASSIFICATION)

- K Nearest Neighbors(KNN)
  - The accuracy on the validation data: `0.8482142857142858`
  - The accuracy on the test data: `0.8333333333333334`
  - Confusion matrix:



RESULT--PREDICTIVE ANALYSIS (CLASSIFICATION)

- After compare the four different methods, we know that
  - When tested on the test set, they all share the same accuracy score and confusion matrix their GridSearchCV best scores are used to rank them instead.
  - On the GridSearchCV best scores, there are a little difference.
  - The models are ranked in the following order with the best to the worst:
    - 1. Decision tree (GridSearchCV best score: 0.8892857142857142)
    - 2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
    - 3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
    - 4. Logistic regression (GridSearchCV best score: 0.8464285714285713)

DISCUSSION

- In this project, we most discuss and using the data are launches and landings, such as its payload mass or orbit type. In the future, we also can build machine learning models to develop some other account features, such as weather conditions, engine performance and so on

CONCLUSION

- Predicting the success of the Falcon 9 first stage landing is a complex task due to the many factors influencing the outcome.
- In this project, we using historical data and machine learning models can provide insights and probabilities, such as its payload mass or orbit type.
- It's important to recognize that the future success of landings may be affected by unforeseen events or changes in the rocket's design or procedures.
- This project can be a valuable exercise in exploring data analysis and machine learning techniques, but it should not be seen as a definitive predictor of future outcomes.
- The primary objective of such a project would be to gain insights into the factors that contribute to successful landings and to potentially identify areas for improvement in the landing process.

APPENDIX

- SpaceX data: "https://api.spacexdata.com/v4/launches/past“
- Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- GitHub Link:Applied-Data-Science-Capstone