IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Yuki Nakayama>
<1st,MAR,2024>

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

SUMMARY OF METHODOLOGIES

01 Data Collection – SpaceX API

02 Data Collection - Scraping

03 Data Wrangling

04 EDA with Data Visualization

05 EDA with SQL

06 Build an Interactive Map with Folium

07 Build a Dashboard with Plotly Dash

08 Predictive Analysis (Classification)

Summary of all results

- Developed predictive models to forecast the success or failure of future SpaceX launches based on historical data, achieving a high level of accuracy and reliability.

Utilized machine learning

- algorithms such as logistic regression to predict launch outcomes and assess the likelihood of mission success.

Introduction

Project background and context

Space X 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 Space X can reuse the first stage. Therefore 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 space X for a rocket launch.

Problems you want to find answers

What factors contribute to the success of landing?

SUCCESSFUL LANDING

UNSUCCESSFUL LANDING

Yuki Nakayama 1st,MAR,2024

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Accessing SpaceX data using the SpaceX API.
  - Retrieving information about launches, rockets, payloads, and other relevant data.
- Perform data wrangling
  - Cleaning and formatting the retrieved data. Handling missing values, duplicates, and outliers. Transforming data into a suitable format for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create classification models with algorithms like logistic regression, decision trees, SVM and KNN fine-tune parameters, and evaluate accuracy using metrics such as accuracy.

Data Collection

Requesting Data from SpaceX API

Utilizing the SpaceX API, relevant information such as flight details, booster versions, payload data, and launch site details were retrieved through HTTP requests.

Data Preprocessing and Cleaning

Upon obtaining the raw data, preprocessing steps were undertaken, including handling missing values, converting data types, and filtering out irrelevant information. This ensured that the data was structured properly and ready for analysis.

Web Scraping Wikipedia for Historical Launch Records

Web scraping techniques were employed to extract historical launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." This process involved parsing HTML tables to extract launch details such as flight numbers, dates, times, booster versions, launch sites, payloads, payload masses, orbits, customers, launch outcomes, and booster landing statuses.

Creating Pandas DataFrames for Analysis and Model Training

The collected data was organized into Pandas DataFrames, facilitating exploratory data analysis (EDA) and predictive modeling. These DataFrames served as the foundation for subsequent analysis, feature engineering, and model training phases.

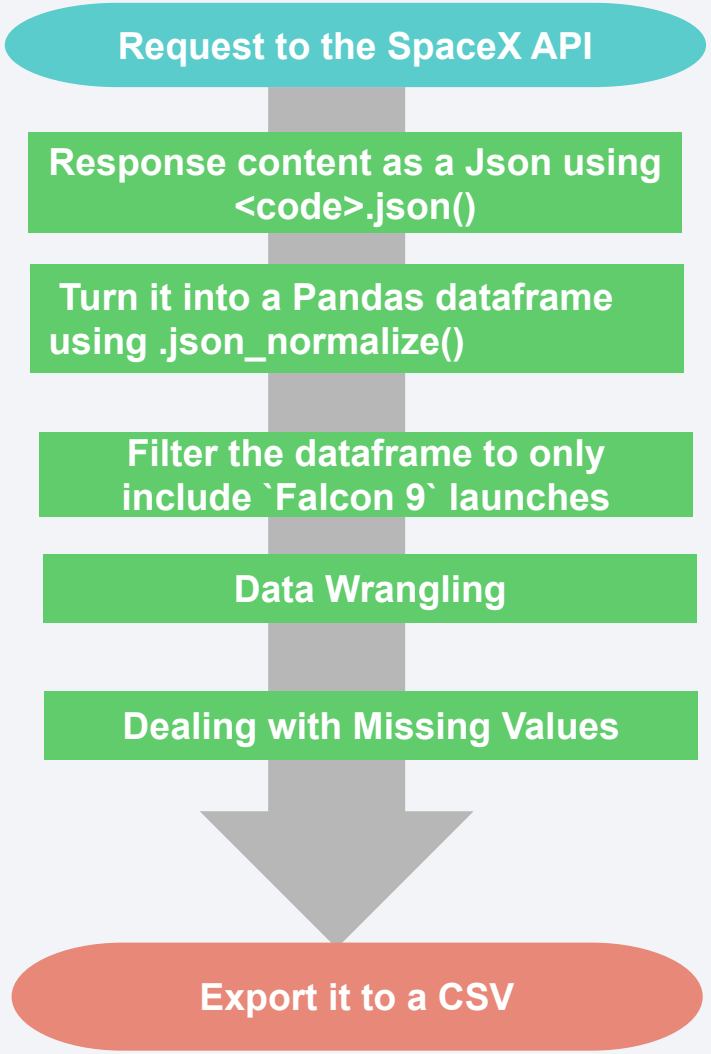
Data Collection – SpaceX API

Now let's start requesting rocket launch data from SpaceX API with the following URL:

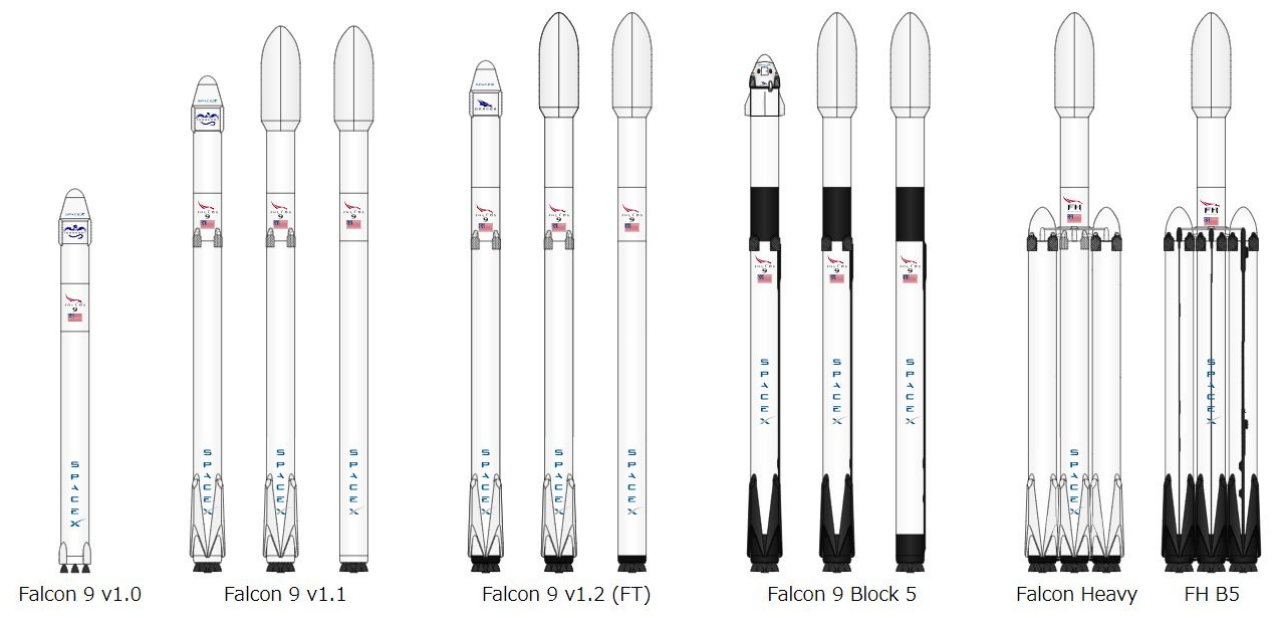
```
[6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

[7]: response = requests.get(spacex_url)
```

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-1_jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping



Request the Falcon9 Launch Wiki page from its URL

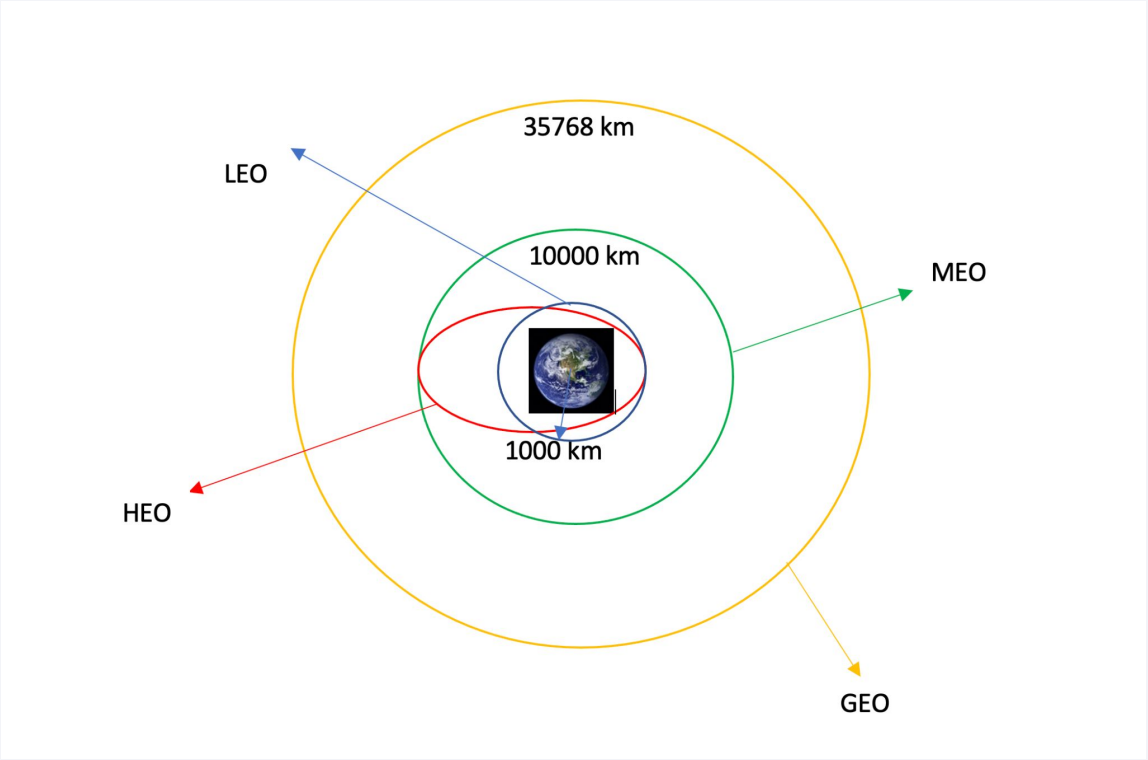
Create a `BeautifulSoup` object from the HTML `response`

Extract all column/variable names from the HTML table heade

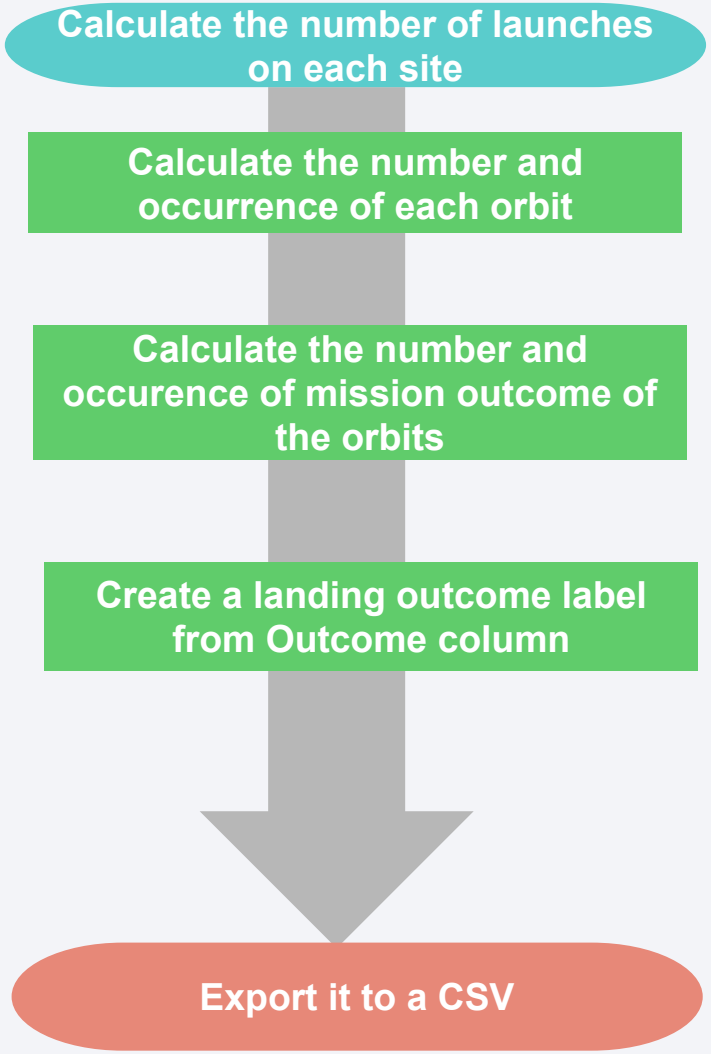
Export it to a CSV

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-2_jupyter-labs-webscraping.ipynb

Data Wrangling



Each launch aims to an dedicated orbit



https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week1-3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Launch Success Yearly Trend

The success rate has been increasing over the years, indicating improvements in launch technology and processes.

Payload vs. Orbit Type

Payload mass may impact the success rate differently for different orbit types.

Flight Number vs. Orbit Type

The success rate may vary for different orbit types, and it may also be influenced by the number of flights.

Flight Number vs. Launch Site

Different launch sites have different success rates, with some sites having higher success rates than others.

Payload vs. Launch Site

Some launch sites have limitations on the payload mass they can handle, leading to different success rates for different payload masses at each launch site.

Success Rate vs. Orbit Type

Some orbit types have higher success rates than others, providing insights into which orbits are more favorable for successful launches.

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week2-1_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Yuki Nakayama 1st,MAR,2024

EDA with SQL

- 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:
- List the date when the first successful landing outcome in ground pad was achieved:
- 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 using a subquery
- List the records displaying the month names, failure landing_outcomes in drone ship, booster versions, and launch_site for the months in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week2-1_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

1 Markers

Markers provide a clear visual indication of the location of each launch site, allowing users to easily identify them.

2 Circle

Circles can be used to indicate specific areas or zones (e.g., NASA Johnson Space Center) on the map, providing additional context or emphasis to certain locations.

3 Marker Clusters

hen there are many markers in close proximity, marker clusters help prevent overcrowding and improve the map's readability by dynamically grouping nearby markers.

4 PolyLine

PolyLines visually represent the distance or route between two locations, providing insight into the proximity or connectivity between the launch site and important landmarks(e.g., coastline, railway, highway).

5 Distance Markers

These markers provide information on the distance between the launch site and key landmarks, helping to assess the accessibility or proximity of the launch site to critical infrastructure.

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week3-1_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Correlation between payload and Success for all Sites

The chart enables users to visually explore any potential correlation between payload mass and launch success across all sites.

Total Success Launches By Site

This pie chart visualizes the total number of successful launches for each launch site.

It provides an overview of the distribution of successful launches across different launch sites.

Total Success Launches for site CCAFS LC-40

This pie chart specifically focuses on the total number of successful launches for the CCAFS LC-40 launch site.

It allows users to see the success rate of launches specifically from this site

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week3-2_dash_completed.png

Yuki Nakayama 1st,MAR,2024

Predictive Analysis (Classification)

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Four classification algorithms were considered: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K Nearest Neighbors (KNN).

For each algorithm, a GridSearchCV object was created to search for the best hyperparameters using cross-validation.

Model Training and Hyperparameter Tuning

The accuracy scores of all models were compared.

The model with the highest accuracy score was identified as the best performing model.

Selection of Best Performing Model

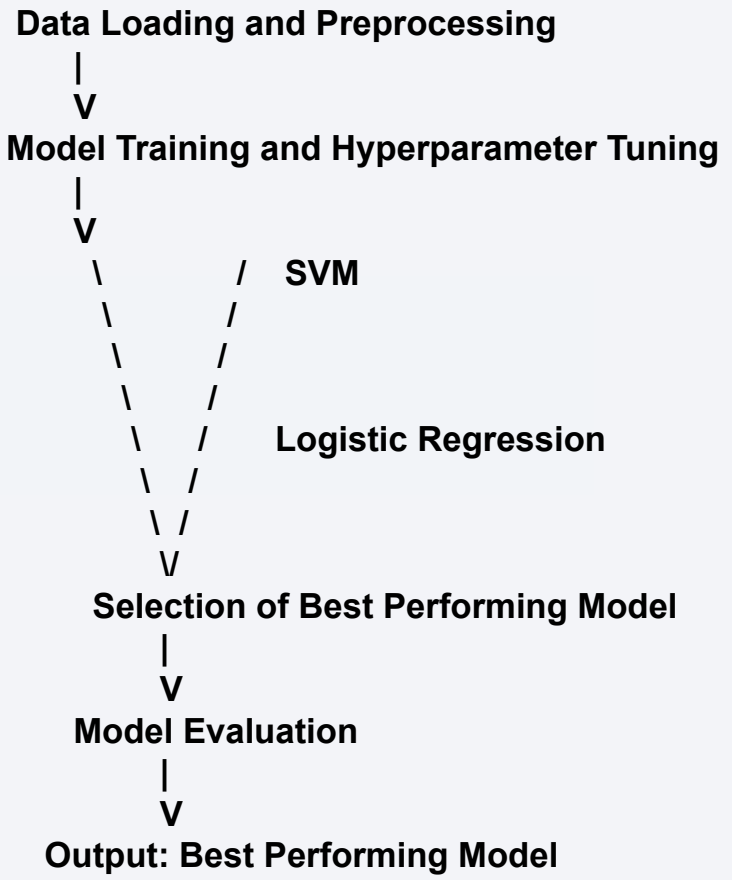
Data Loading and Preprocessing

The dataset containing information about SpaceX Falcon 9 rocket launches was loaded.
・The target variable (class) was extracted and standardized, and the data was split into training and testing sets.

Model Evaluation

The accuracy of each model was evaluated using the testing data.

Confusion matrices were generated to visualize the performance of each model in distinguishing between successful and unsuccessful landings.



https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/week4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Yuki Nakayama 1st,MAR,2024

Results

Predictive analysis results

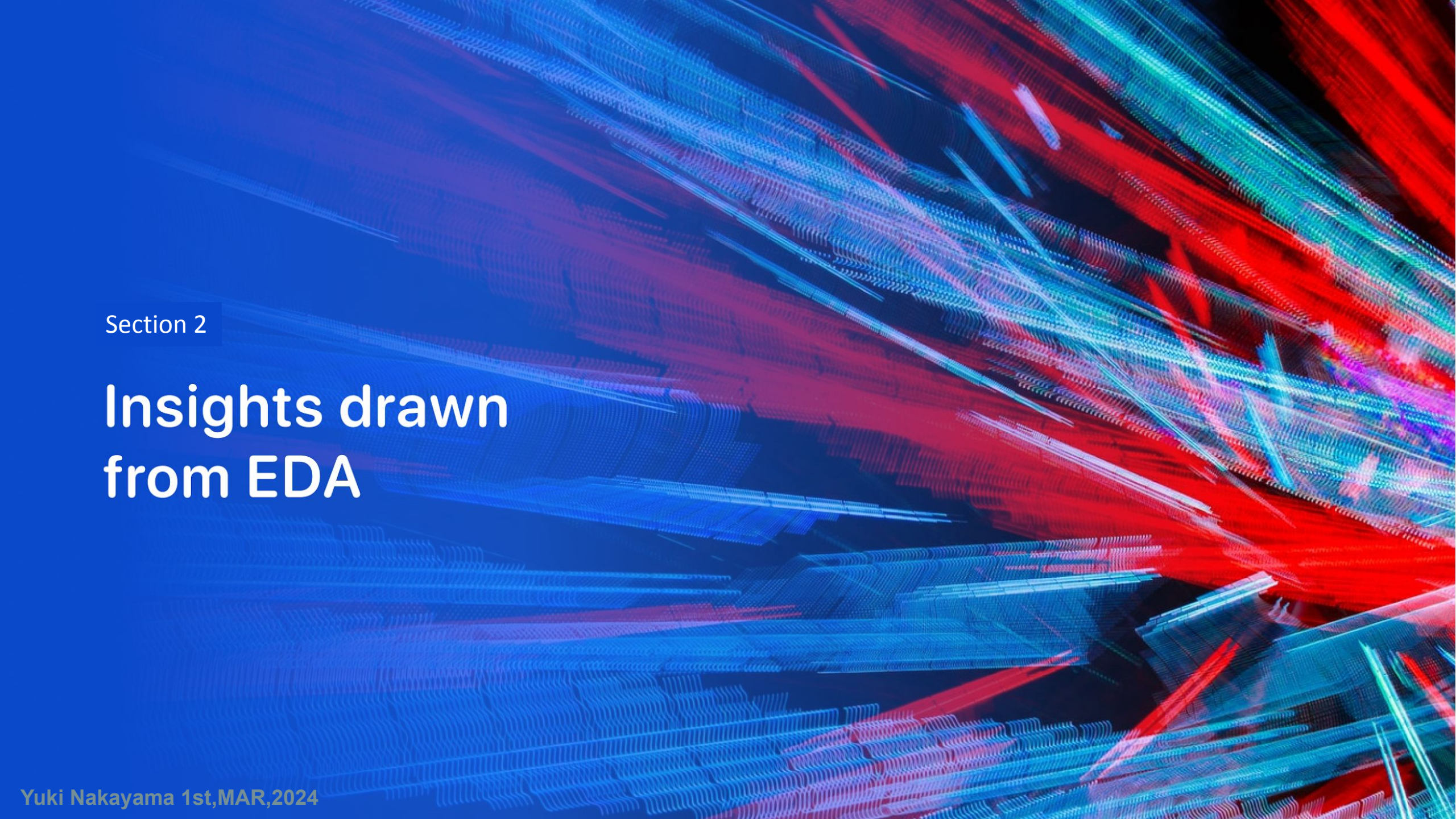
The classification model development process involved data loading, preprocessing, training, and evaluation of Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K Nearest Neighbors (KNN) models, with the best performing model selected based on accuracy scores.

Interactive analytics demo in screenshots

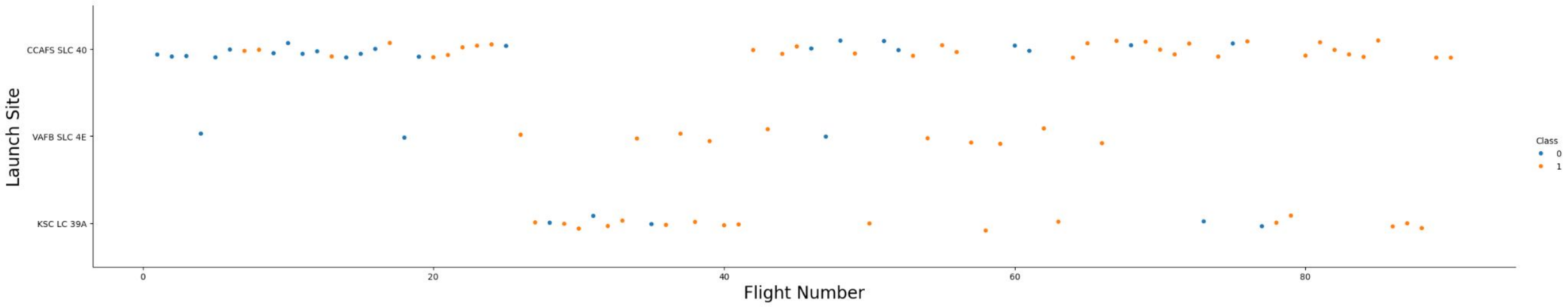
The interactive analytics demo includes markers for all launch sites, an overview of launch outcomes at each site, and displays distances between launch sites and their proximities.

Exploratory data analysis results

Exploratory data analysis reveals site-specific variations in launch success rates, diverse payload masses, potential correlations between payload mass and success, insights into SpaceX's technological advancements through booster versions, and trends in launch success over time.


Section 2
Insights drawn from EDA
Yuki Nakayama 1st,MAR,2024

Flight Number vs. Launch Site

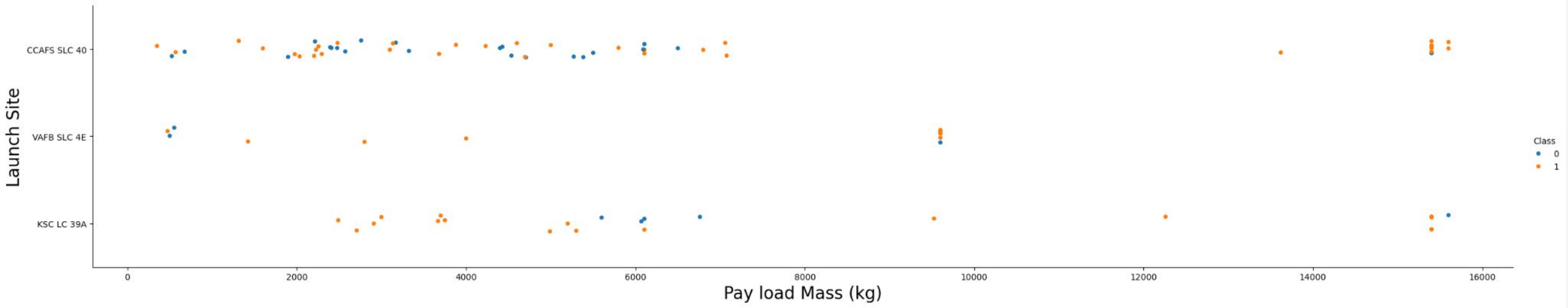


●:Success

●:failed

As the flight number increases, the first stage is more likely to land successfully.
This trend seems to be present across all launch sites.

Payload vs. Launch Site

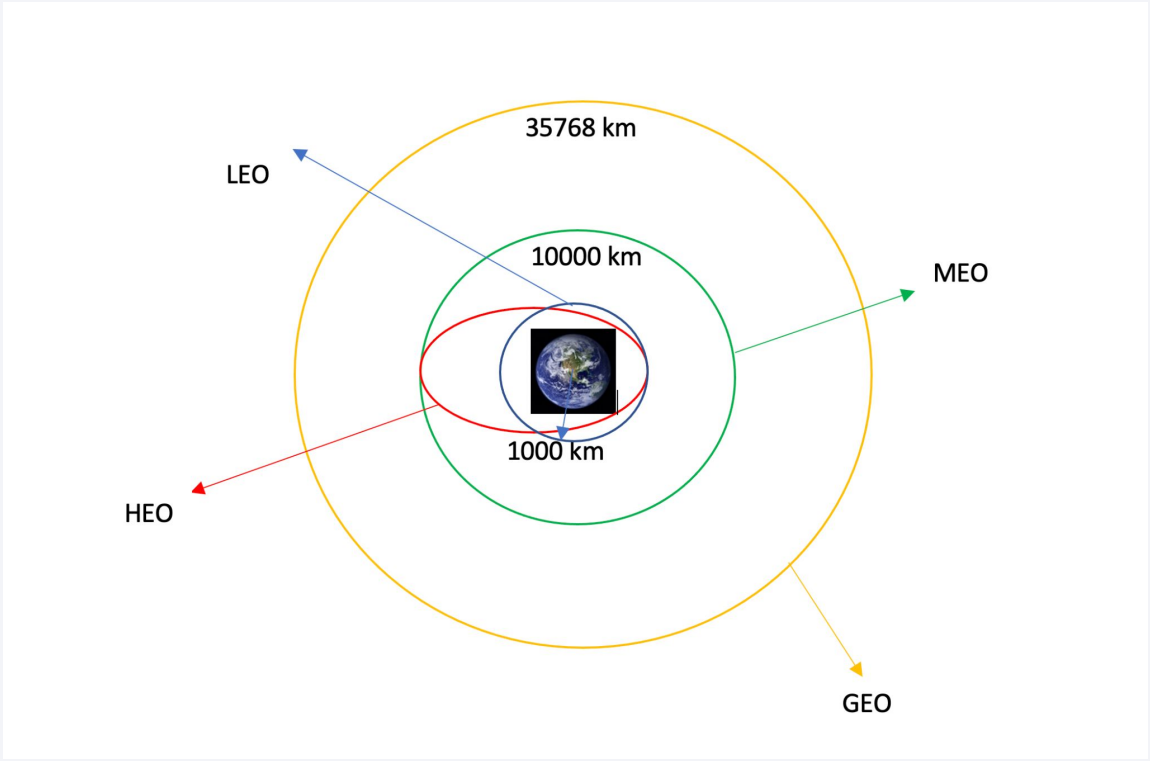


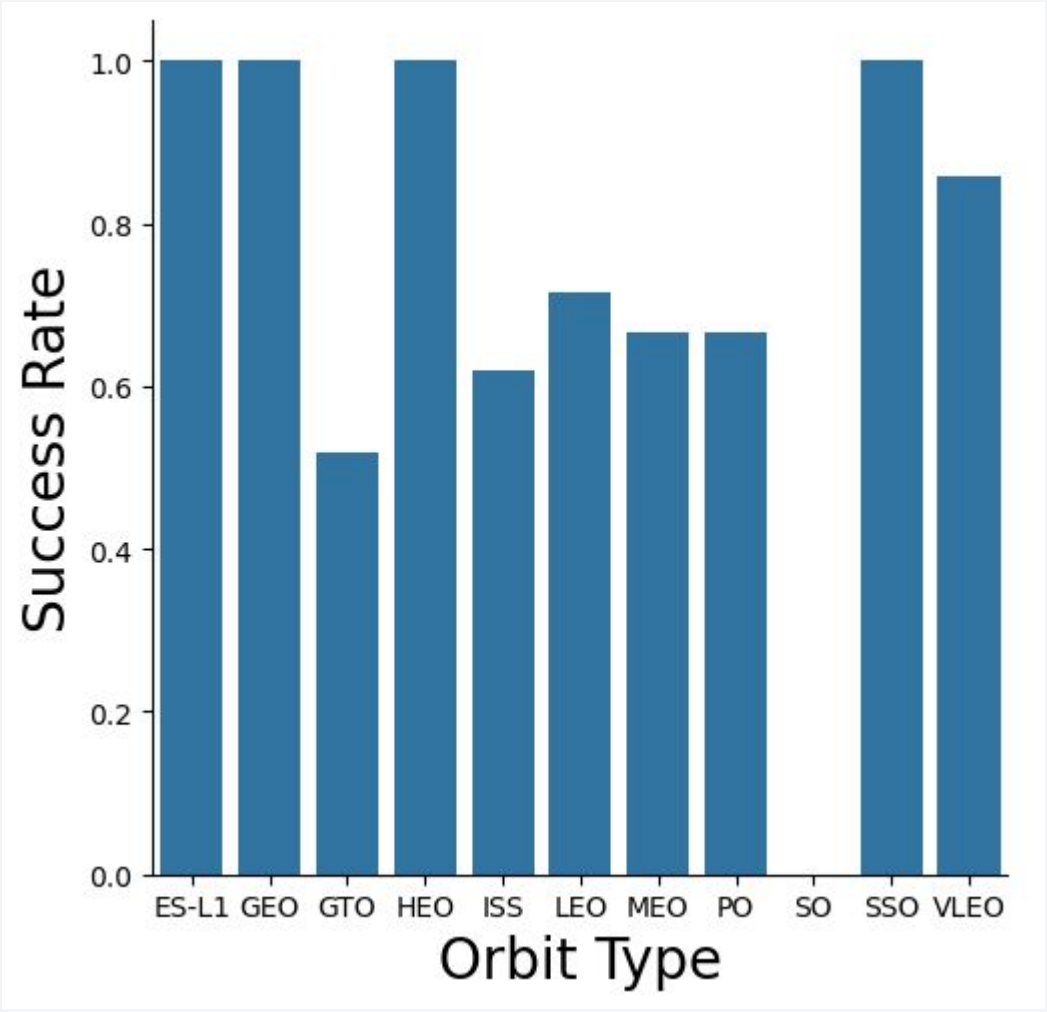
●:Success

●:failed

It appears that as the payload increases, so does the flight number.

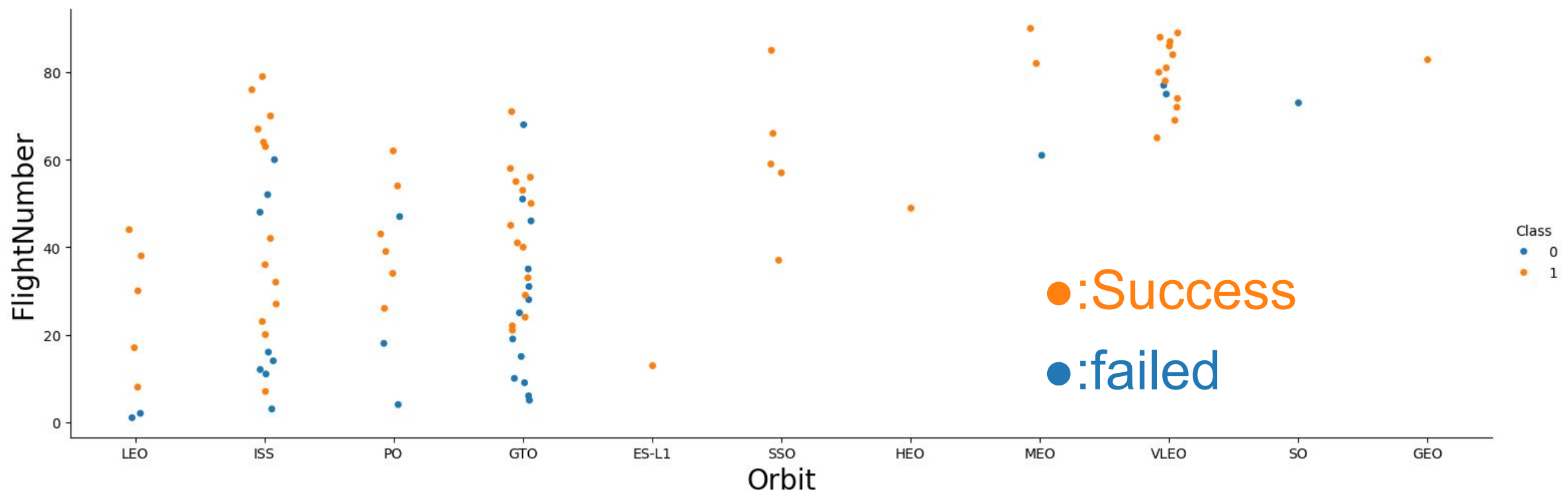
Success Rate vs. Orbit Type





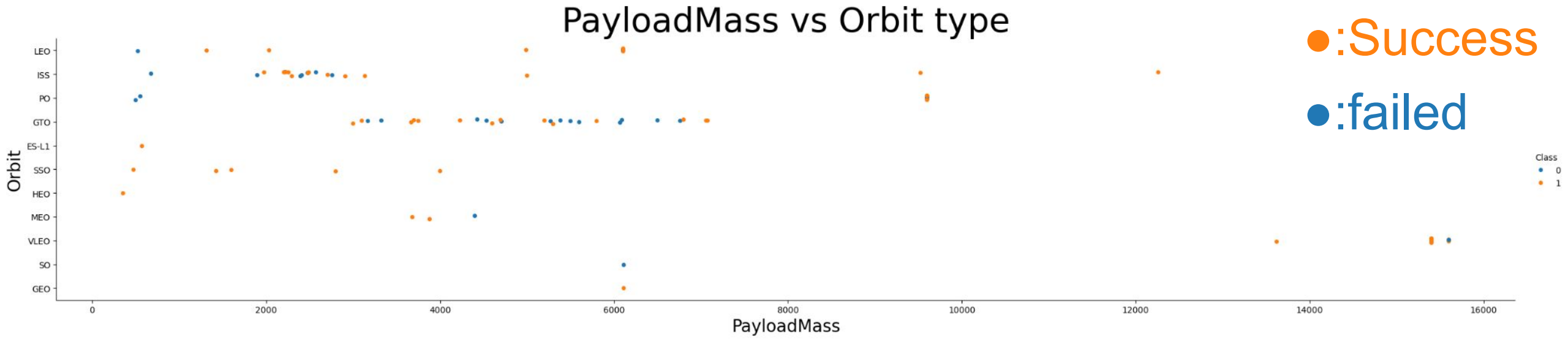
ES-L1, GEO, HEO, and SSO have a high success rate.

Flight Number vs. Orbit Type



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

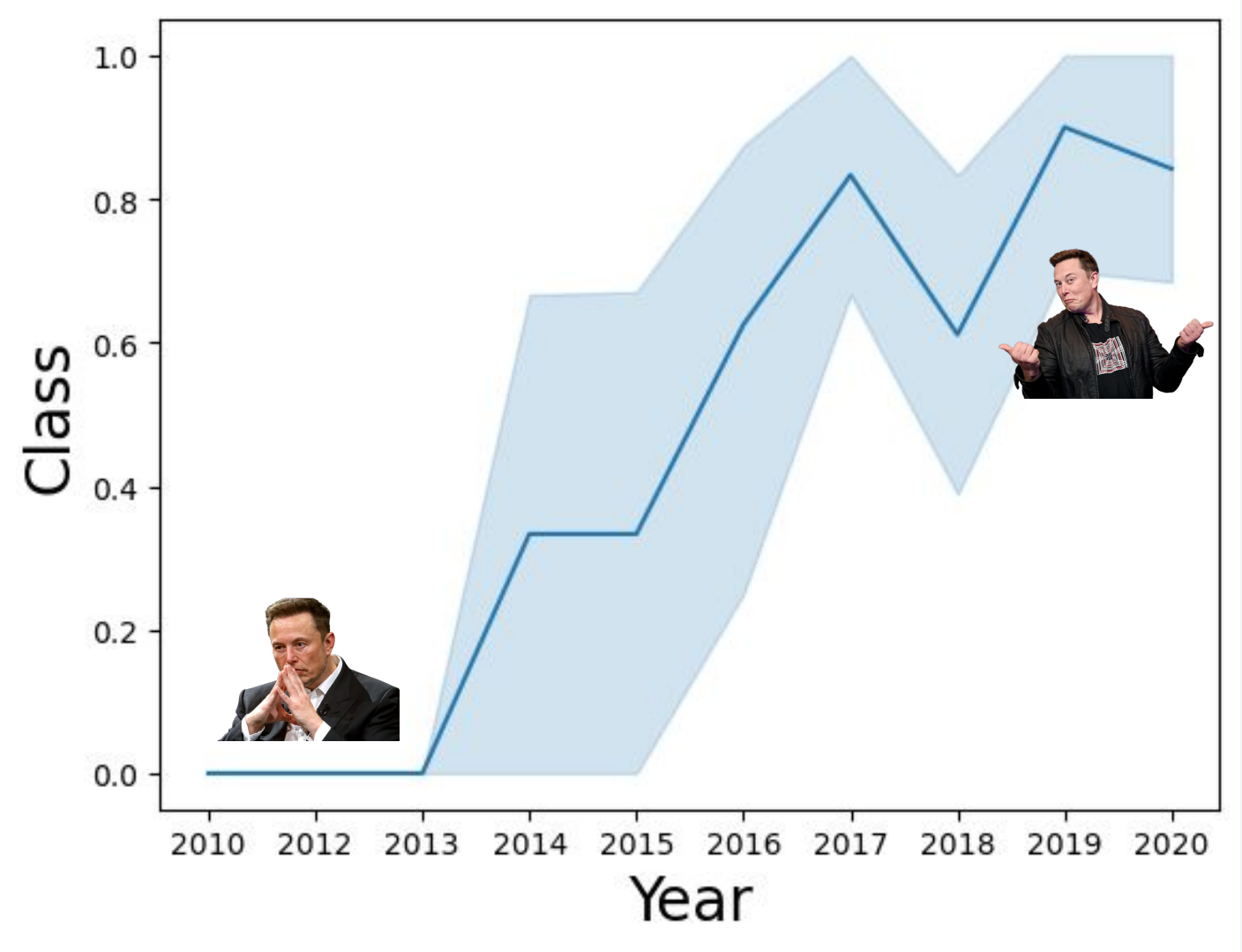


For heavy payloads, it appears that the successful landing rate or positive landing rate is higher for Polar, LEO, and ISS.

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

It can be noted that the success rate has shown a consistent increase from 2013 to 2020.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

SQL QUERY WITH A SHORT EXPLANATION

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

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTBL where LAUNCH_SITE like "CCA%" limit 5
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 attempt

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

SQL QUERY WITH A SHORT EXPLANATION

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

SQL QUERY WITH A SHORT EXPLANATION

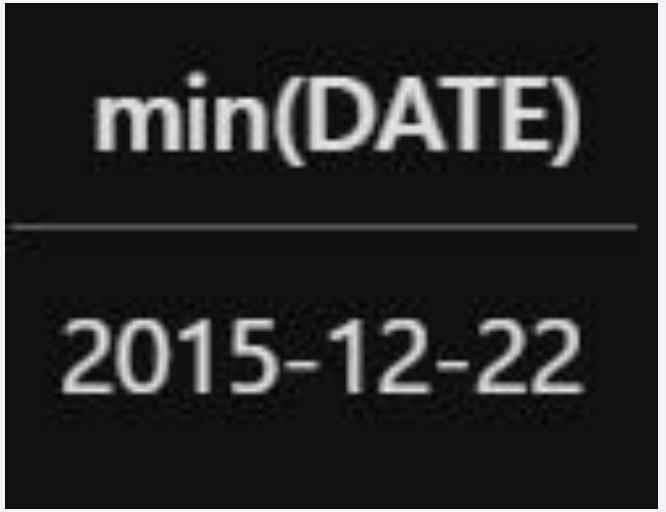
avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

```
%sql select min(DATE) from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)'
```

SQL QUERY WITH A SHORT EXPLANATION



Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

SQL QUERY WITH A SHORT EXPLANATION

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql \
SELECT \
    Landing_Outcome, \
    COUNT(*) AS Outcome_Count \
FROM \
    SPACEXTBL \
WHERE \
    Landing_Outcome IN ('Success', 'Failure') \
GROUP BY \
    Landing_Outcome;
```

SQL QUERY WITH A SHORT EXPLANATION

Landing_Outcome	Outcome_Count
Failure	3
Success	38

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT Booster_Version \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL LIMIT 1);
```

SQL QUERY WITH A SHORT EXPLANATION

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

Falcon 9 Block 5

2015 Launch Records

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)

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql \
SELECT \
    substr(Date, 6, 2) AS Month, \
    Booster_Version, \
    Launch_Site, \
    Landing_Outcome \
FROM \
    SPACEXTBL \
WHERE \
    substr(Date, 0, 5) = '2015' \
    AND Landing_Outcome LIKE 'Failure (drone ship)%';
```

SQL QUERY WITH A SHORT EXPLANATION

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.

```
%sql \
SELECT \
    Landing_Outcome, \
    COUNT(*) AS Outcome_Count \
FROM \
    SPACEXTBL \
WHERE \
    Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY \
    Landing_Outcome \
ORDER BY \
    Outcome_Count DESC;
```

SQL QUERY WITH A SHORT EXPLANATION

Landing_Outcome	Outcome_Count
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

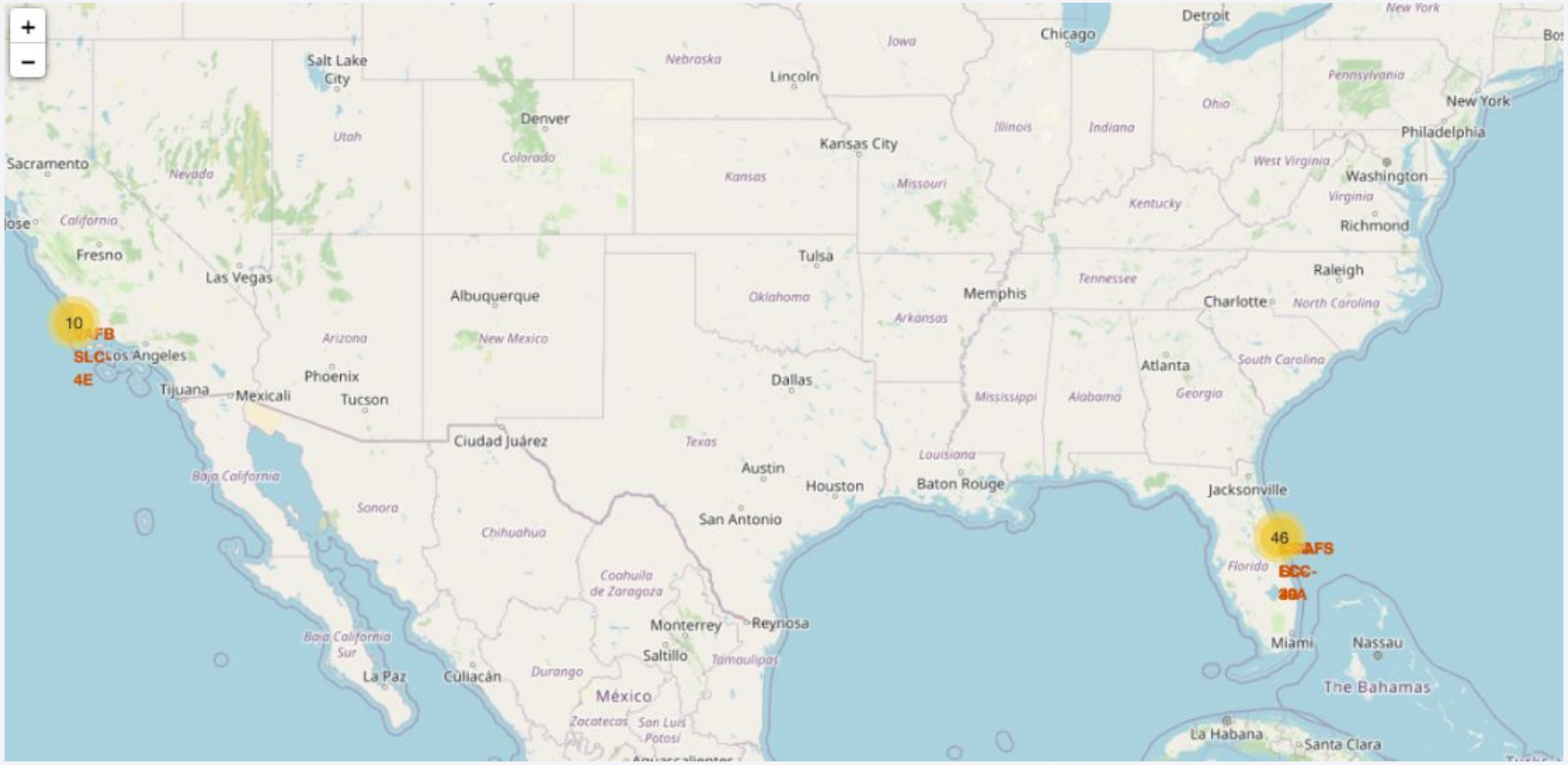
Launch Sites Proximities Analysis

Yuki Nakayama 1st,MAR,2024

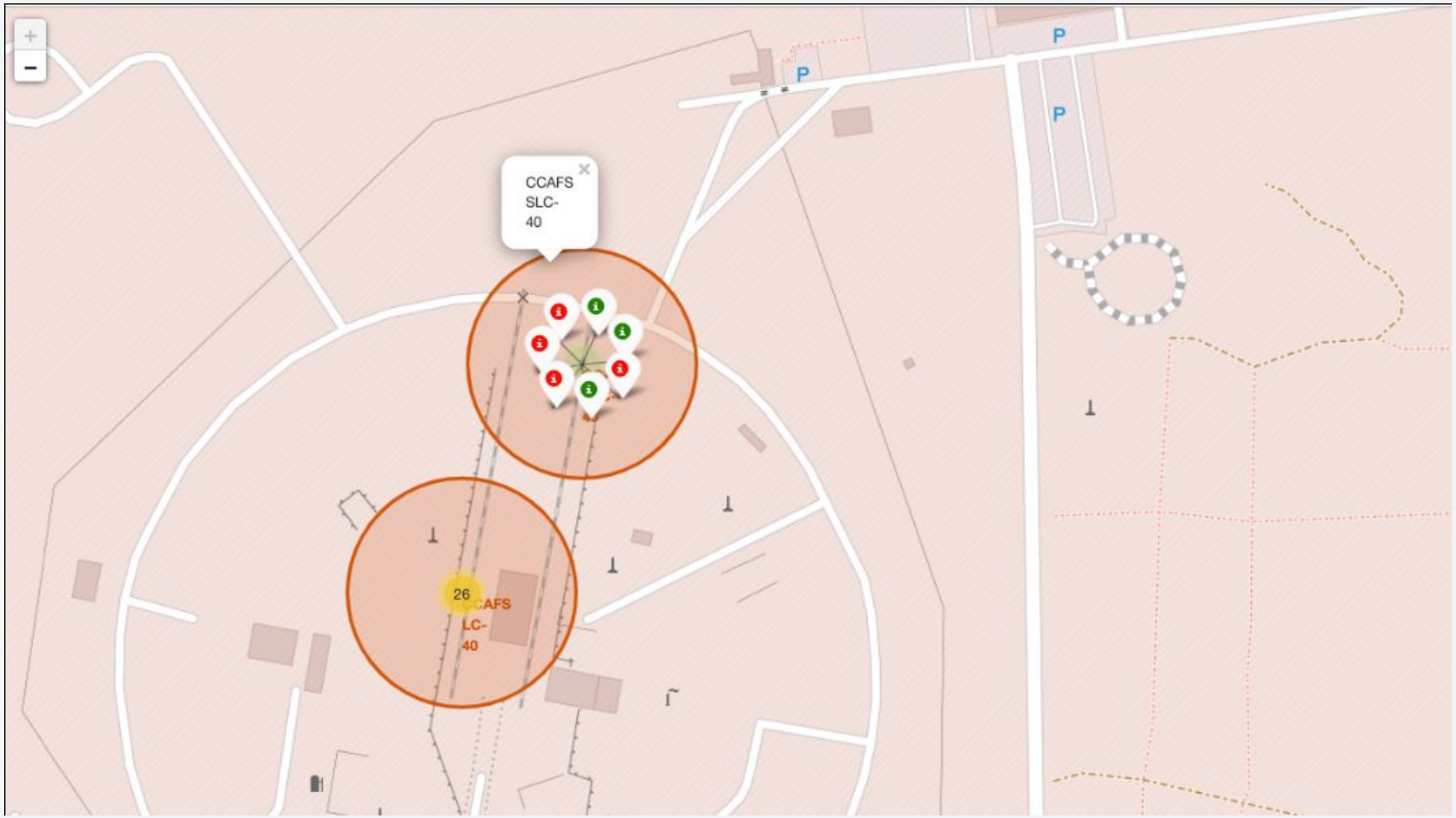
Markers all launch sites on a map

All launch sites are typically located near the Equator due to the benefits of Earth's maximum rotational speed, which results in reduced fuel needs and lower costs for reaching orbit.

All launch sites are strategically situated in close proximity to the coast to ensure safe paths over water, thereby minimizing risks to populated areas during launch failures.
In addition, it is worth noting that these systems can aid in the effective removal of rocket debris, while also providing logistical benefits for both operations and support.



An overview of the outcomes of launches at each site on the map



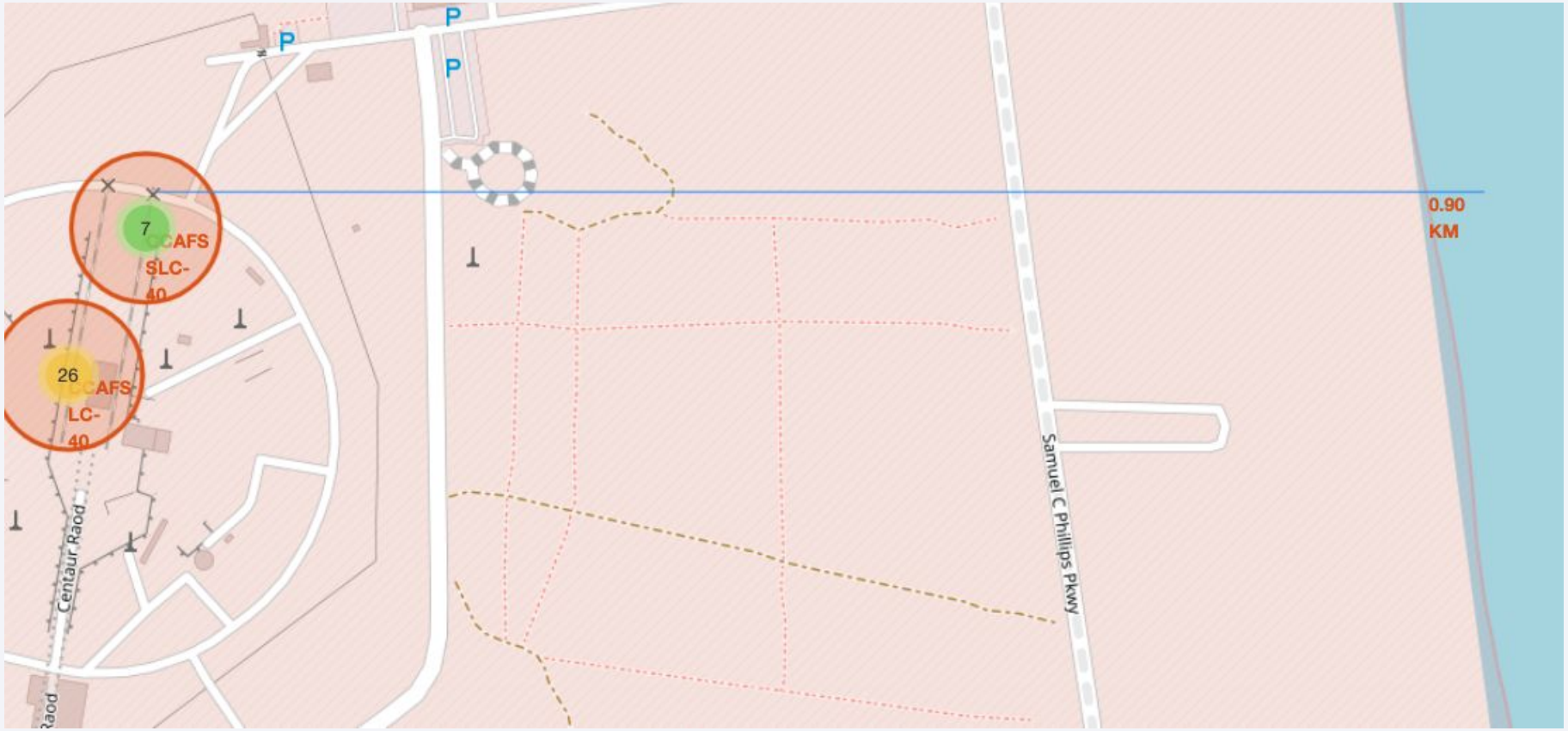
Distances between a launch site to its proximities

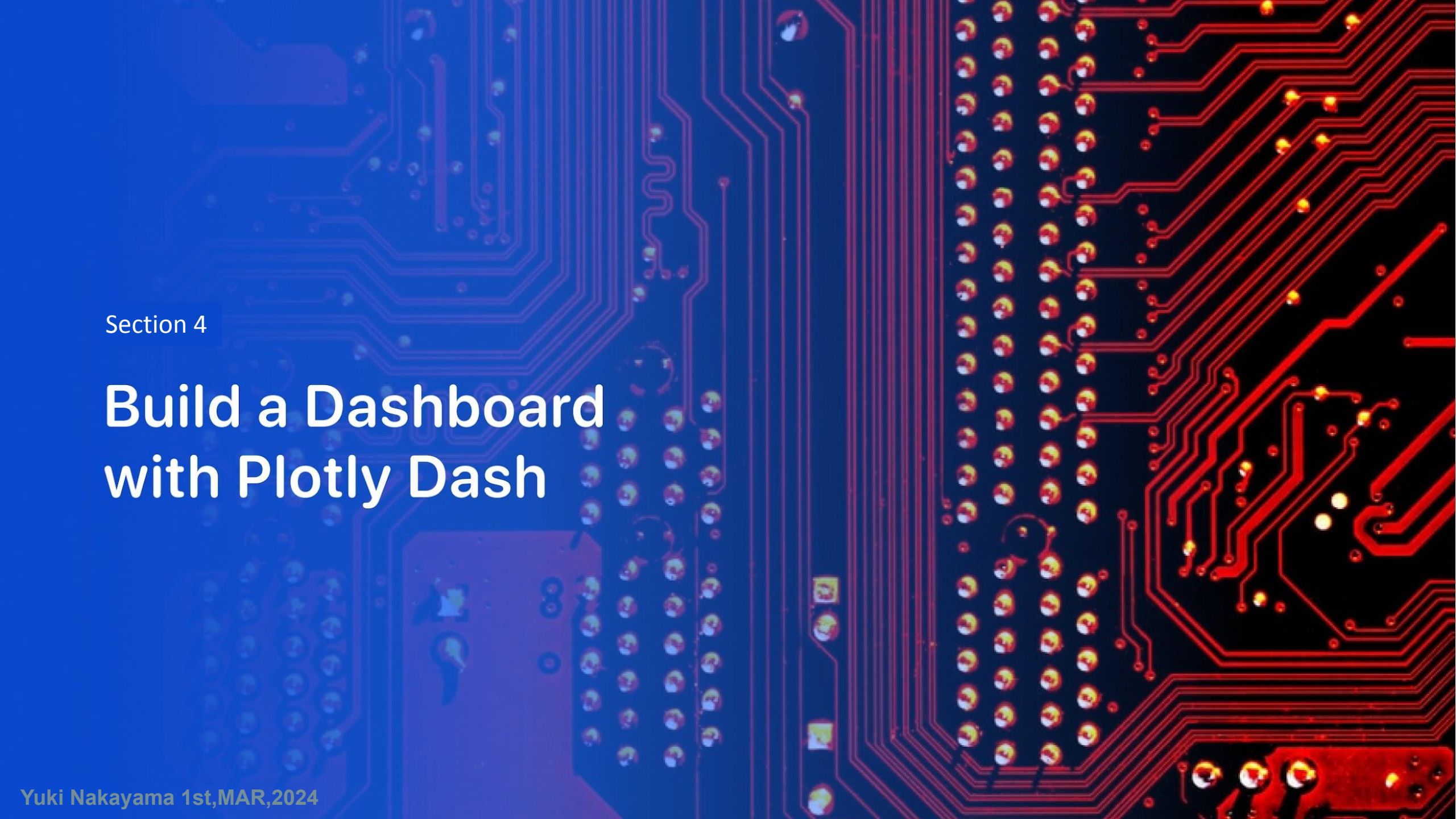
Launch sites are in close proximity to railways : Efficient transportation of heavy equipment.

Launch sites are in close proximity to highways: Easy access for personnel and support vehicles.

Launch sites are in close proximity to coastline: Clear trajectory over water, minimal risks to populated areas.

launch sites keep certain distance away from cities: Minimize risks to human life and property.



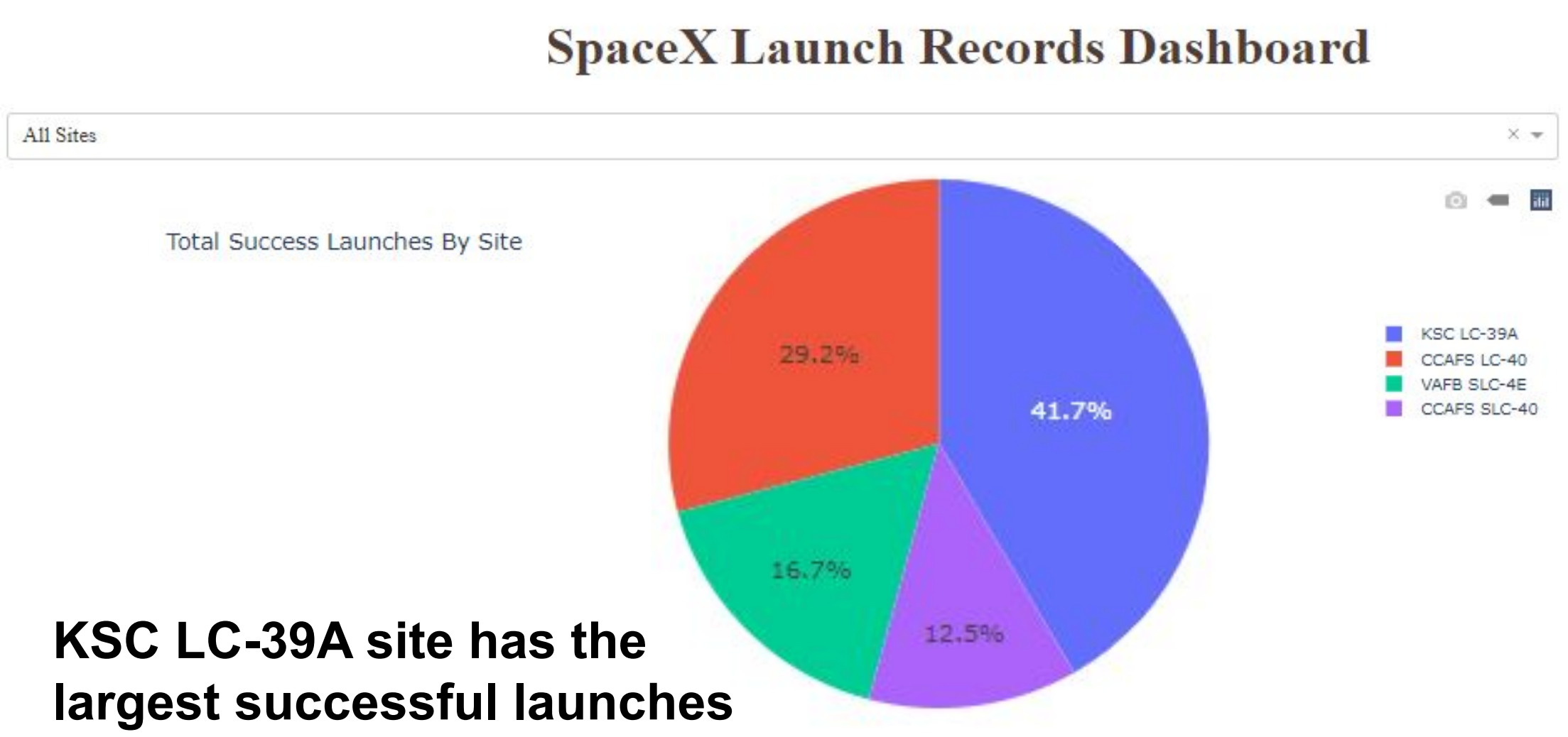


Section 4

Build a Dashboard with Plotly Dash

Yuki Nakayama 1st,MAR,2024

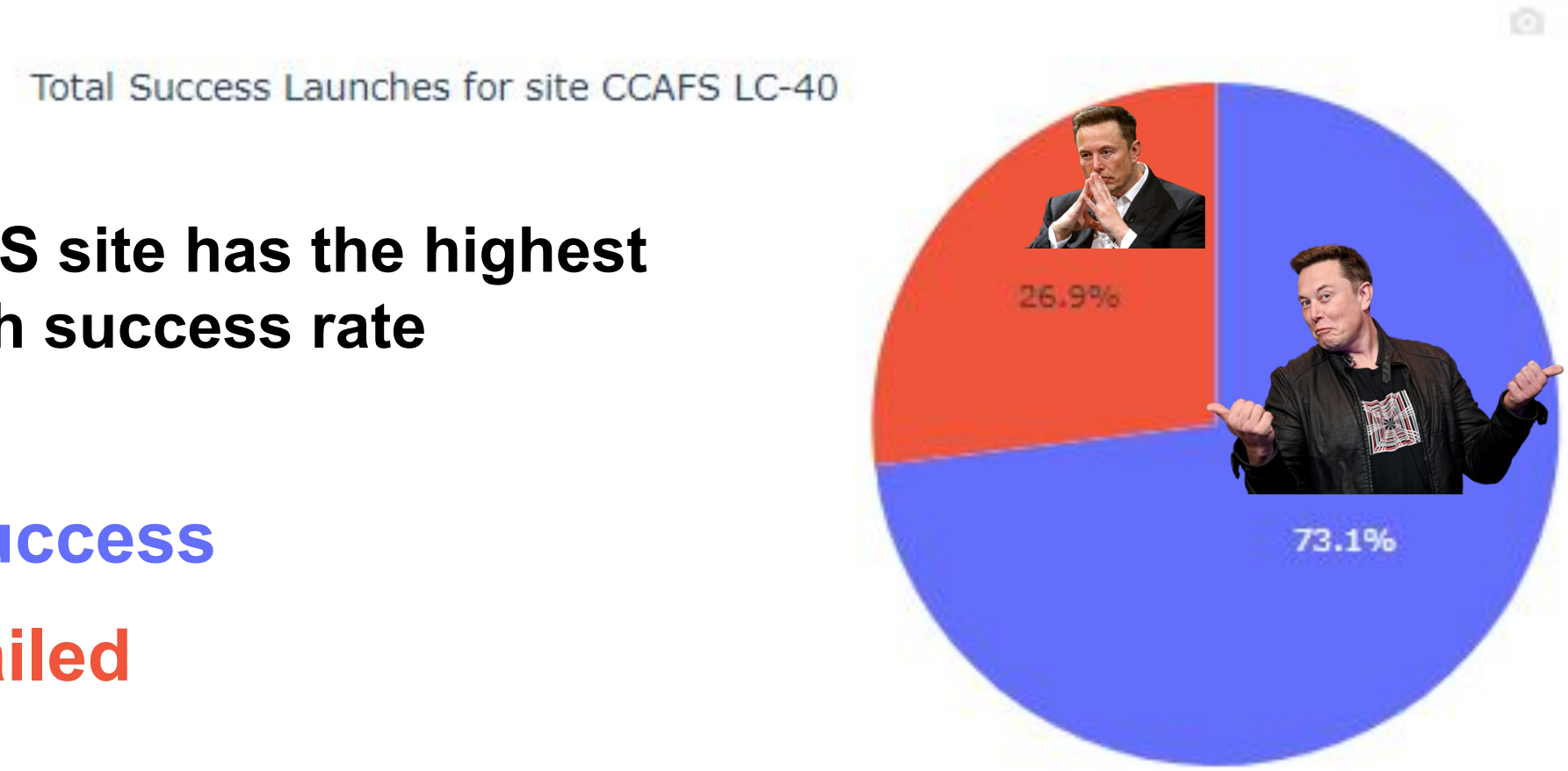
Total Success Launches By Site



KSC LC-39A site has the largest successful launches

Total Success Launches for site CCAFS LC-40

CCAFS LC-40



CCAFS site has the highest launch success rate

- **Success**
- **Failed**

Correlation between payload and Success for all Sites

- Payload from 2000kg to 5000kg has the highest launch success rate.
- Payload more than 6000kg has the lowest launch success rate.
- F9 Booster version FT has the highest launch success rate.

Section 5

Predictive Analysis (Classification)

Yuki Nakayama 1st,MAR,2024

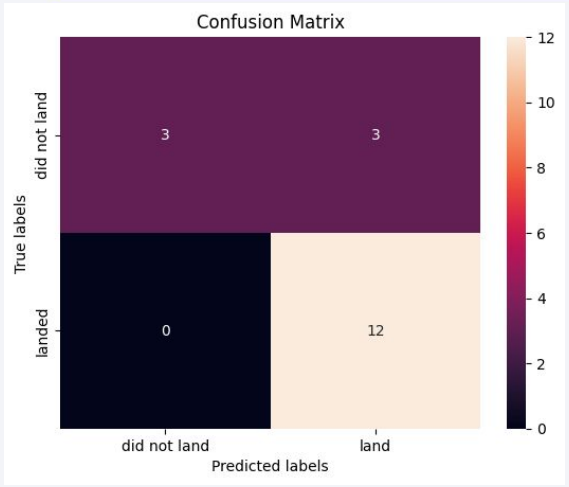
Classification Accuracy

Logistic regression decision tree,

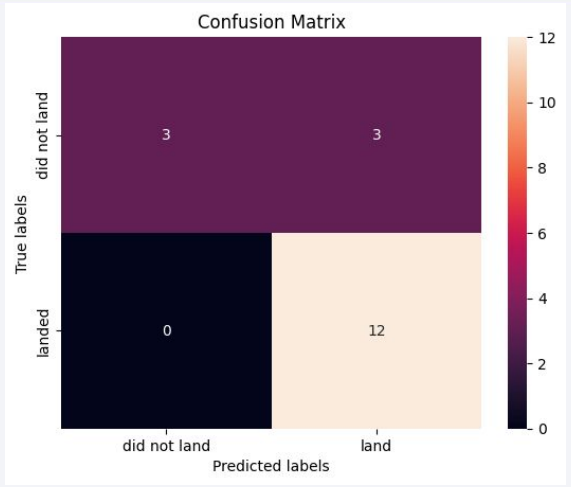
SVM and KNN model has the same classification accuracy

ML Method	Accuracy Score (%)
Support Vector Machine	83.333333
Logistic Regression	83.333333
K Nearest Neighbour	83.333333
Decision Tree	83.333333

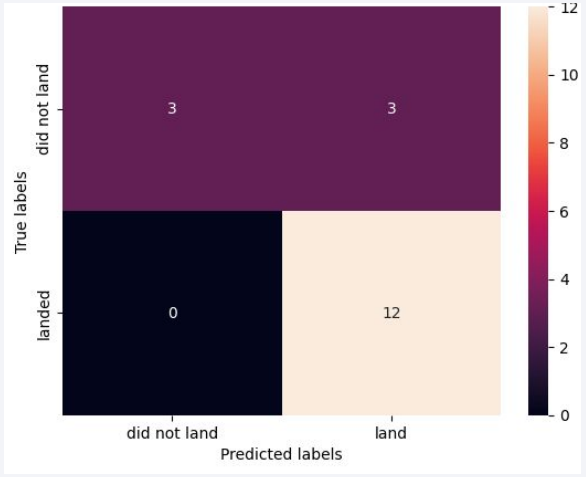
Confusion Matrix



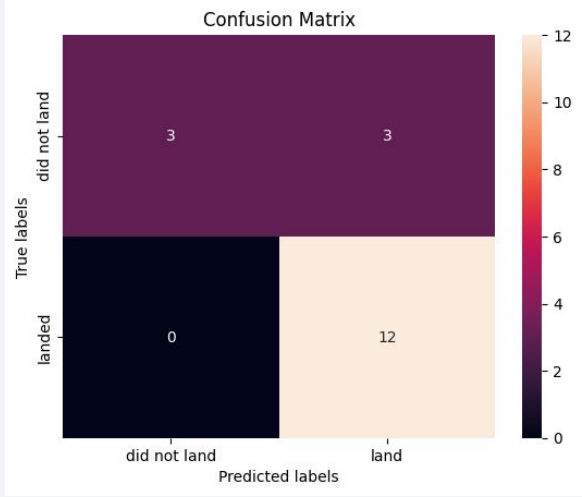
LOGISTIC REGRESSION



DECISION TREE

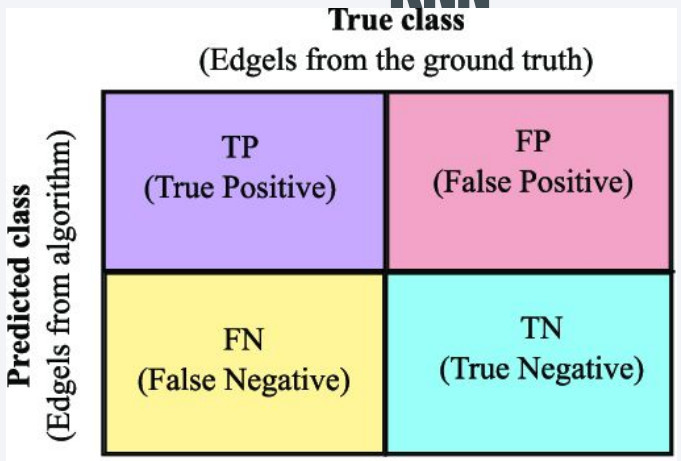


SVM



KNN

Upon examining the confusion matrix, it is clear that logistic regression decision tree, SVM and KNN are capable of distinguishing between the different classes. However, false positives remain a significant issue.



Tariq, N., Hamzah, R. A., Ng, T. F., Wang, S. L., & Ibrahim, H. (2021). Quality assessment methods to evaluate the performance of edge detection algorithms for digital image: A systematic literature review. IEEE Access, 9, 87763-87776.

Conclusions

1 Flight Number

As the flight number increases, the first stage is more likely to land successfully.

2 Orbit Type

ES-L1, GEO, HEO, and SSO have a high success rate.

3 Payload

- The payload increases, so does the flight number.
- F9 Booster version FT has the highest launch success rate.

4 Surroundings

- All launch sites are typically located near the Equator
- All launch sites are strategically situated in close proximity to the coast

5 Launch site

- sKSC LC-39A site has the largest successful launches
- CCAFS site has the highest launch success rate

6 Classification

Upon examining the confusion matrix, it is clear that logistic regression decision tree, SVM and KNN are capable of distinguishing between the different classes

Appendix

https://github.com/YukiG16/IBM-Data-Science-Certificate/blob/main/Course10_applied-data-science-capstone/Data-science-using-SpaceX-API/

- week1-1_jupyter-labs-spacex-data-collection-api.ipynb
- week1-2_jupyter-labs-webscraping.ipynb
- week1-3_labs-jupyter-spacex-Data wrangling.ipynb
- week2-1_jupyter-labs-eda-sql-coursera_sqllite.ipynb
- week2-2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
- week3-1_lab_jupyter_launch_site_location.jupyterlite.ipynb
- week3-2_dash_completed.png
- week4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite

Thank you!

Yuki Nakayama 1st,MAR,2024