IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AMRUHA AHMED

15th October, 2024.

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-

By Amruha Ahmed

Outline

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

Executive Summary

Summary of methodologies

- The following methods were applied to gain a comprehensive picture of the data available:
- Data Collection
- Data Wrangling
- EDA
- Interactive Analytical Dashboard
- Predictive Analysis

Summary of all results

- EDA and visualizations using dashboards helped in understanding how each parameter is affected by other and gain valuable insights
- Accuracy of predicting whether the first stage of Falcon 9 rocket will land or not is 83.33%

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.

Problems you want to find answers

- determine the price of each launch for **Space Y , that is competing with SpaceX.**
- Gather information about SpaceX
- Creating dashboards for better insights
- Whether SpaceX will reuse the first stage of Falcon 9 or not using machine learning

Section 1

Methodology

By Amruha Ahmed

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scraping and API's
- Perform data wrangling
  - Data was processed using value_counts( ) and functions of descriptive statistics
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were built using Logistic Regression, KNN, SVM, Decision Tree. Grid Search was applied. Models were evaluated using accuracy, precision, recall, F1 Score and confusion matrix

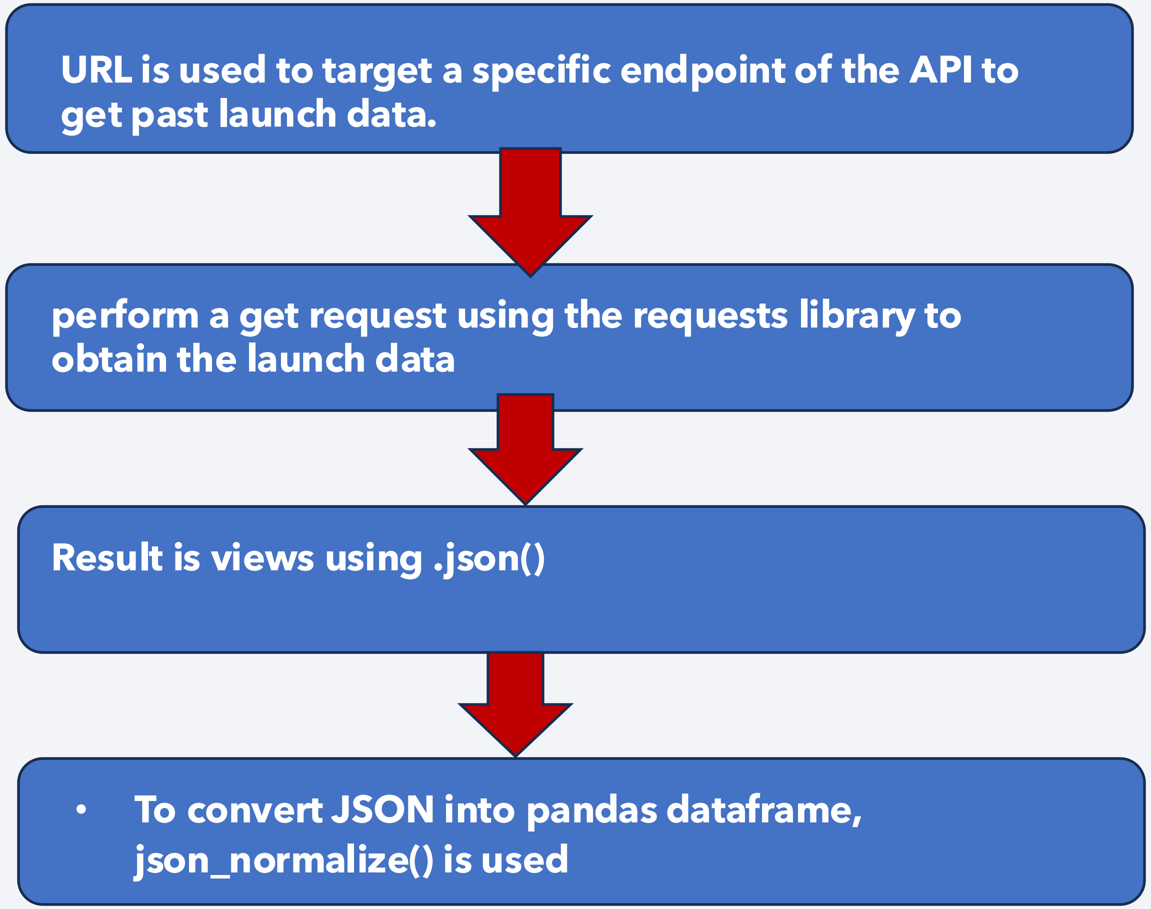
Data Collection

How the datasets were collected?

- API :https://api.spacexdata.com/v4/launches/past
- Wikipedia : https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The data is obtained from Space X API : https://api.spacexdata.com/v4/launches/past
- GitHub URL to my .ipynb notebook containing Data Collection using Space X API Code: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20collection%20using%20api.ipynb



URL is used to target a specific endpoint of the API to get past launch data.

perform a get request using the requests library to obtain the launch data

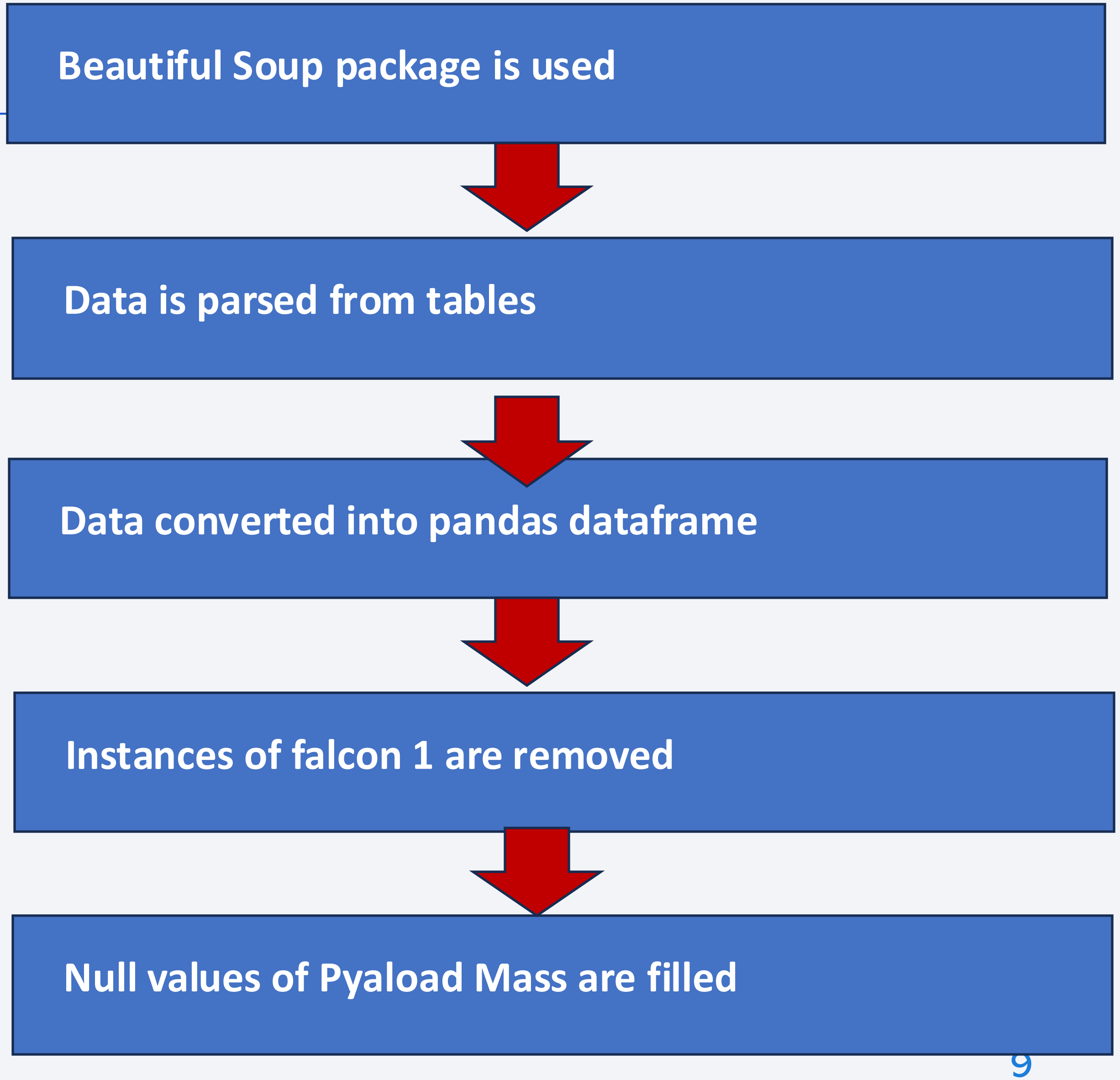
Result is views using .json()

- To convert JSON into pandas dataframe, json_normalize() is used

Data Collection - Web Scraping

- The data is obtained from Wikipedia : https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub URL to my .ipynb notebook containing Data Collection using Web Scraping Code:

https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20collection%20using%20web%20scraping.ipynb



Beautiful Soup package is used

Data is parsed from tables

Data converted into pandas dataframe

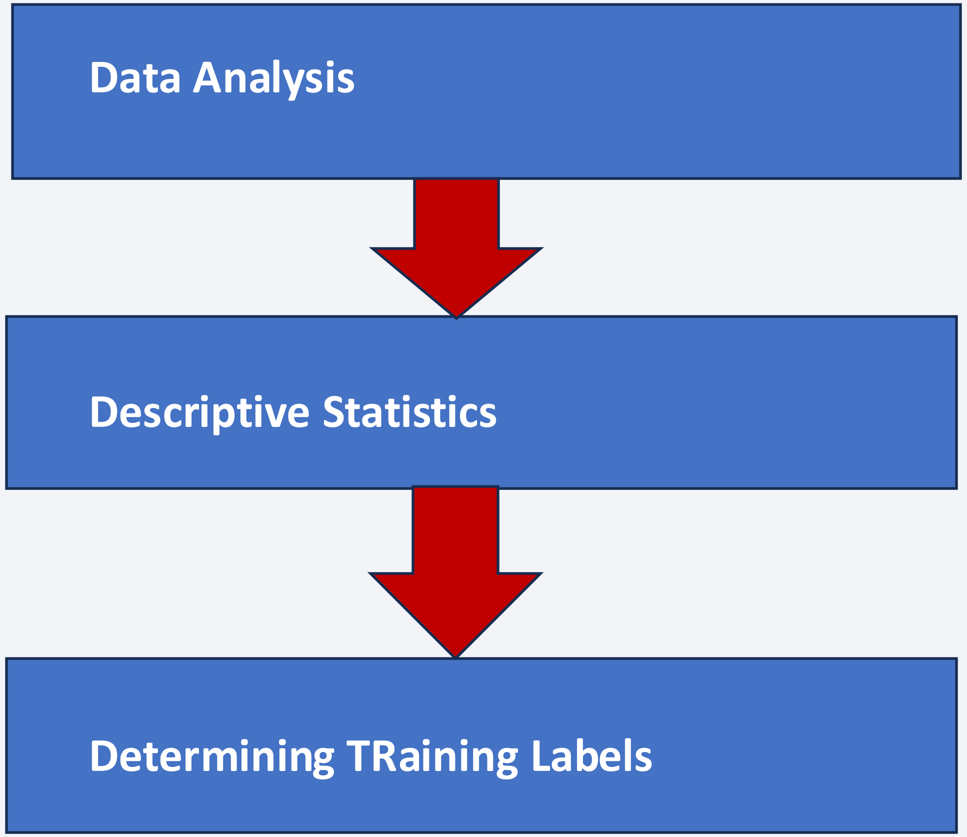
Instances of falcon 1 are removed

Null values of Pyaload Mass are filled

Data Wrangling

Data Wrangling of the dataset involved the following steps:

- Calculating the number of launches on each site
- Calculating the number and instance of each orbit
- Calculating the number and ocurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- Exporting the resultant dataframe into dataset_part_2.csv


Data Analysis

Descriptive Statistics

Determining TRaining Labels

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20wrangling.ipynb

EDA with Data Visualization

Exploratory Data Analysis of the dataset using Pandas and Matplotlib libraries in Python involved the following tasks :

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload Mass and Launch Site
- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload Mass and Orbit type
- Visualize the launch success yearly trend

All these graphs were made using **scatter plot** as it easily helps in finding correlation among two variables. Hues can be added accordingly

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/eda%20with%20pandas%20and%20matplotlib.ipynb

EDA with SQL

Exploratory Data Analysis of the dataset using SQL involved the following tasks :

- Installing SQL alchemy
- Connecting to a database
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first succesful landing outcome in ground pad was acheived.

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/eda%20with%20sql.ipynb

EDA with SQL

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranking 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.

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/eda%20with%20sql.ipynb

Build an Interactive Map with Folium

Markers and Circles:

- Marker and Circle are used to indicate the NASA Johnson Space Center at Houston, Texas.
- They are also used to indicate the Launch Sites
- If a launch was successful , then we use a green marker and if a launch was failed, we use a red marker

Procedure:

- create a folium `Map` object,
- Create and add `folium.Circle` and `folium.Marker` for each launch site on the site map
- Create a new column in `spacex_df` dataframe called `marker_color` to store the marker colors based on the `class` value
- Mark the success/failed launches for each site on the map
- For each launch result in `spacex_df` data frame, add a `folium.Marker` to `marker_cluster`
- Calculate the distances between a launch site to its proximities

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/interactive%20visualizations%20using%20folium.ipynb

Build a Dashboard with Plotly Dash

Summary of Graphs Used

- A pie chart to depict the Lauch sites and a scatter plot to show the Correlation between Payload and Success that are made interactive using Payload slider and selection of Launch Sites.

Procedure:

- Reading the airline data into pandas dataframe
- Creating a dash application
- Creating an app layout
- Adding a dropdown list to enable Launch Site selection
- Adding a callback function for `site-dropdown` as input, `success-pie-chart` as output. And a function decorator to specify function input and output

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/interactive%20dashboard%20using%20dash.py

Predictive Analysis (Classification)

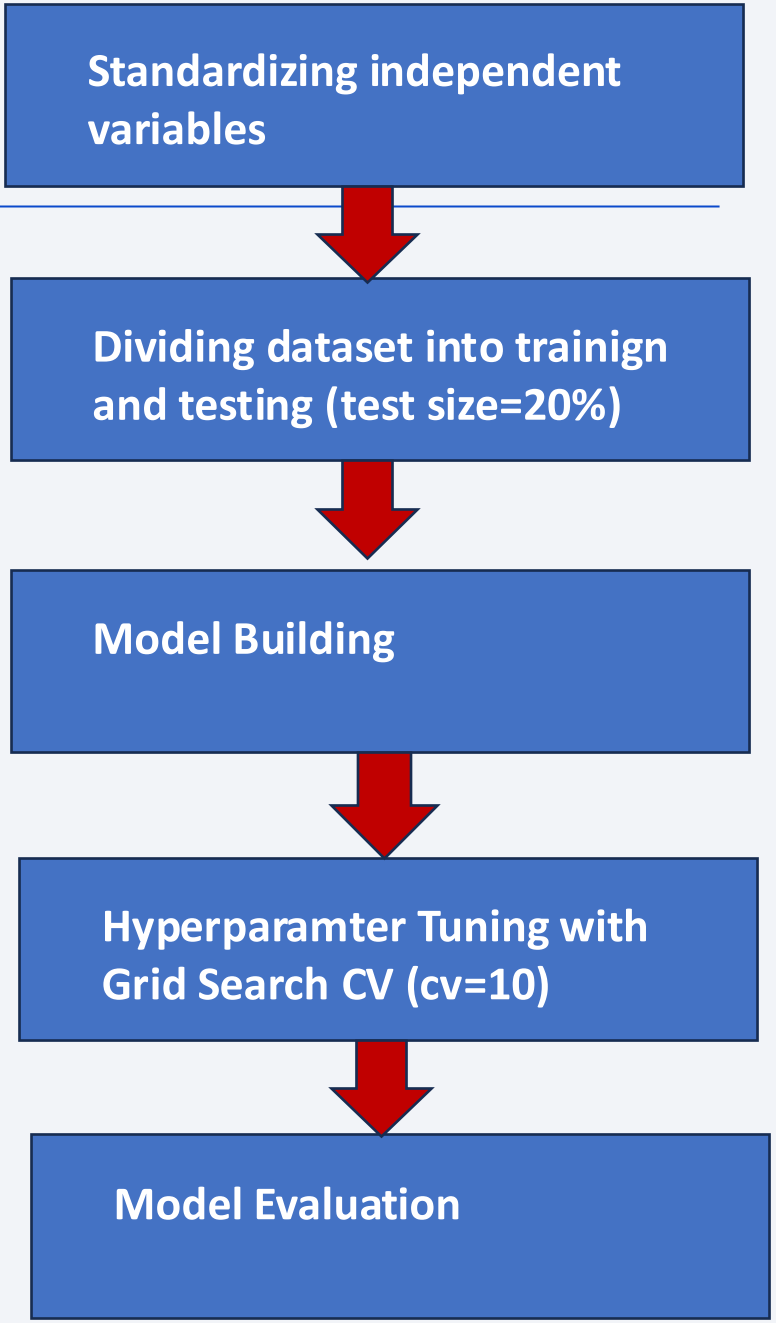
4 different classification models were built :

- Decision Tree Classifier
- Support vEctor Machine
- K Neaest Neighbors
- Logistic Regression

Model is evaluated using :

- Accuracy
- Precision
- recall
- f1 score

GitHub Link:https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/predictive%20analysis(classification).ipynb



Standardizing independent variables

Dividing dataset into trainign and testing (test size=20%)

Model Building

Hyperparamter Tuning with Grid Search CV (cv=10)

Model Evaluation

Model Evaluation

Model	accuracy	precision	recall	F1 score
Logistic Regression	0.8333333333333	0.8	1	0.88888888
Support Vector Machine	0.83333333333333	0.8	1	0.88888888
Decision Tree	0.72222222222222	0.81818	0.75	0.7826086
K Nearest Neighbors	0.833333333333	0.8	1	0.88888888

Results

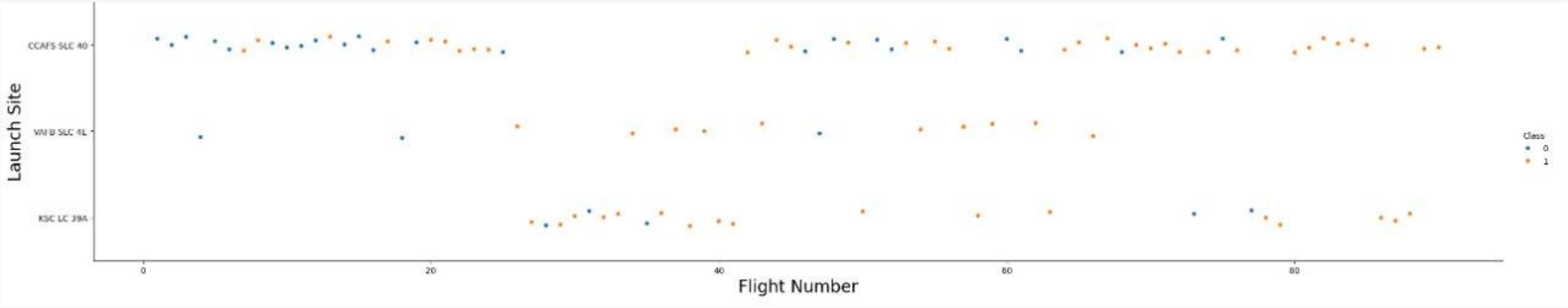
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

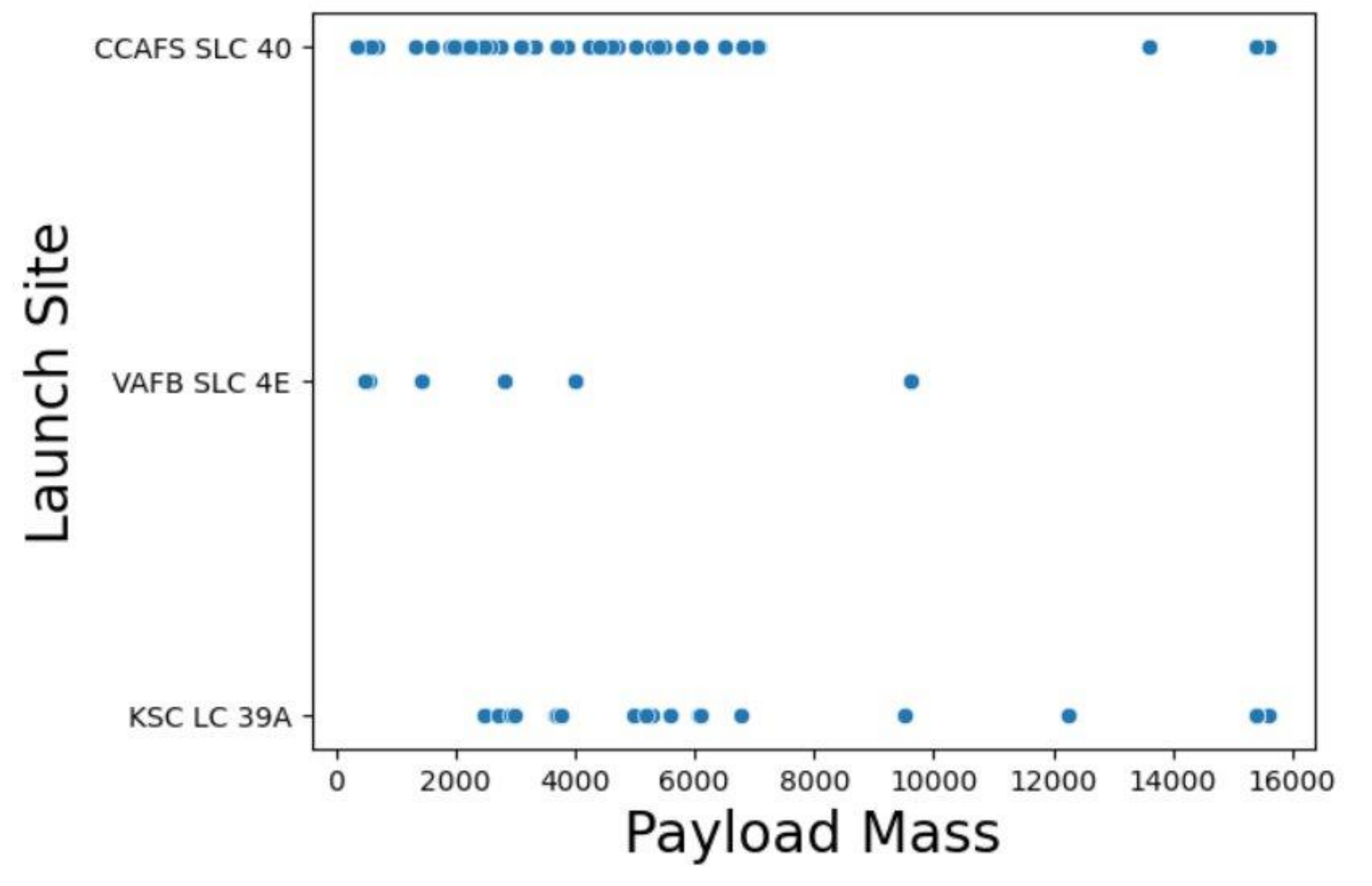
- Results of Task 1: Visualize the relationship between Flight Number and Launch Site
- Insights Gathered : Flight Numbers are higher in CCAFS SLC 40 , with most of them being successful



Payload vs. Launch Site

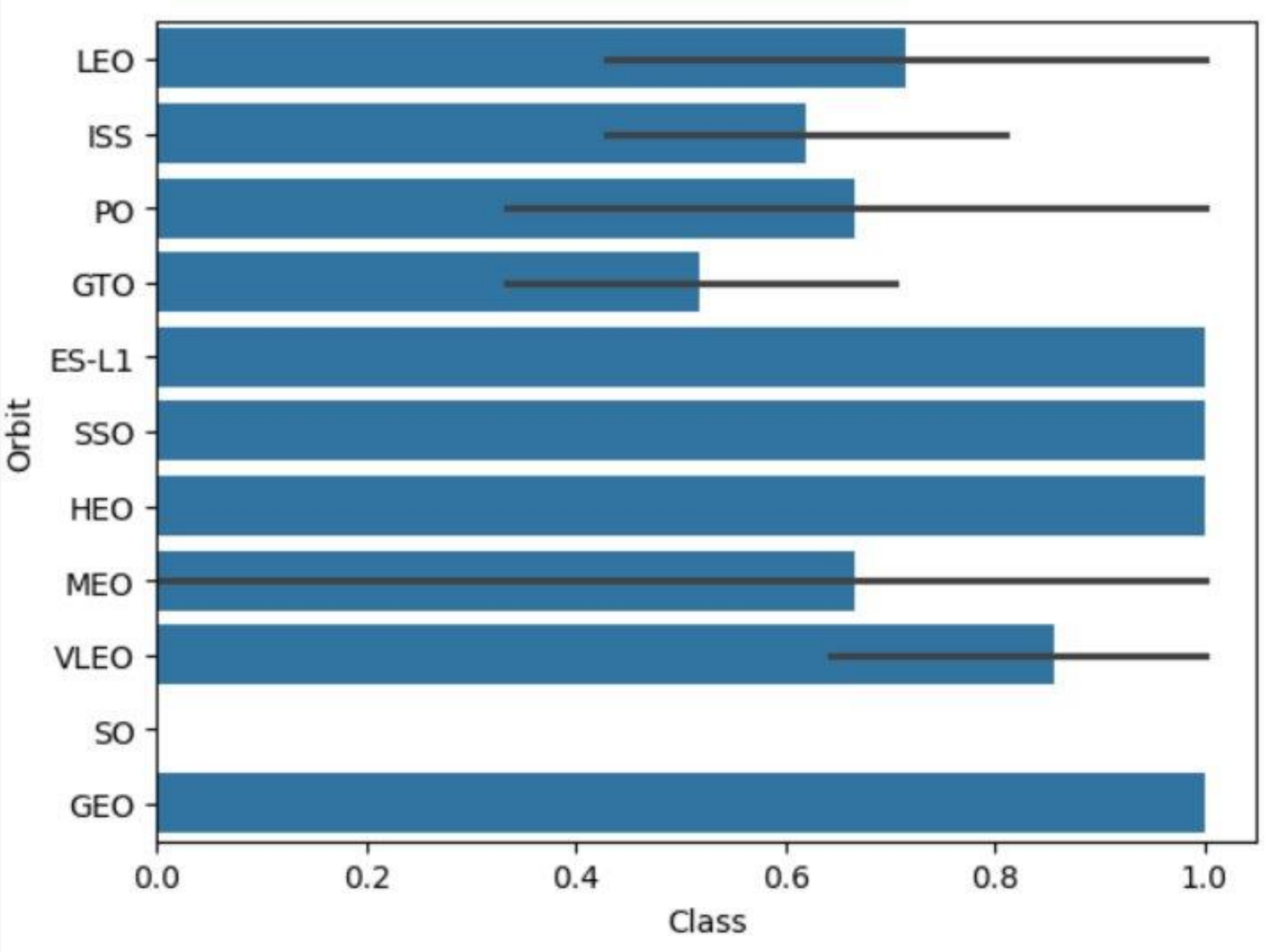
- Results of Task 2: Visualize the relationship between Payload Mass and Launch Site

- Insights Gathered : CCFAS SLC 40 AND KSC LC 39 A have the highest Payload Masses recorded



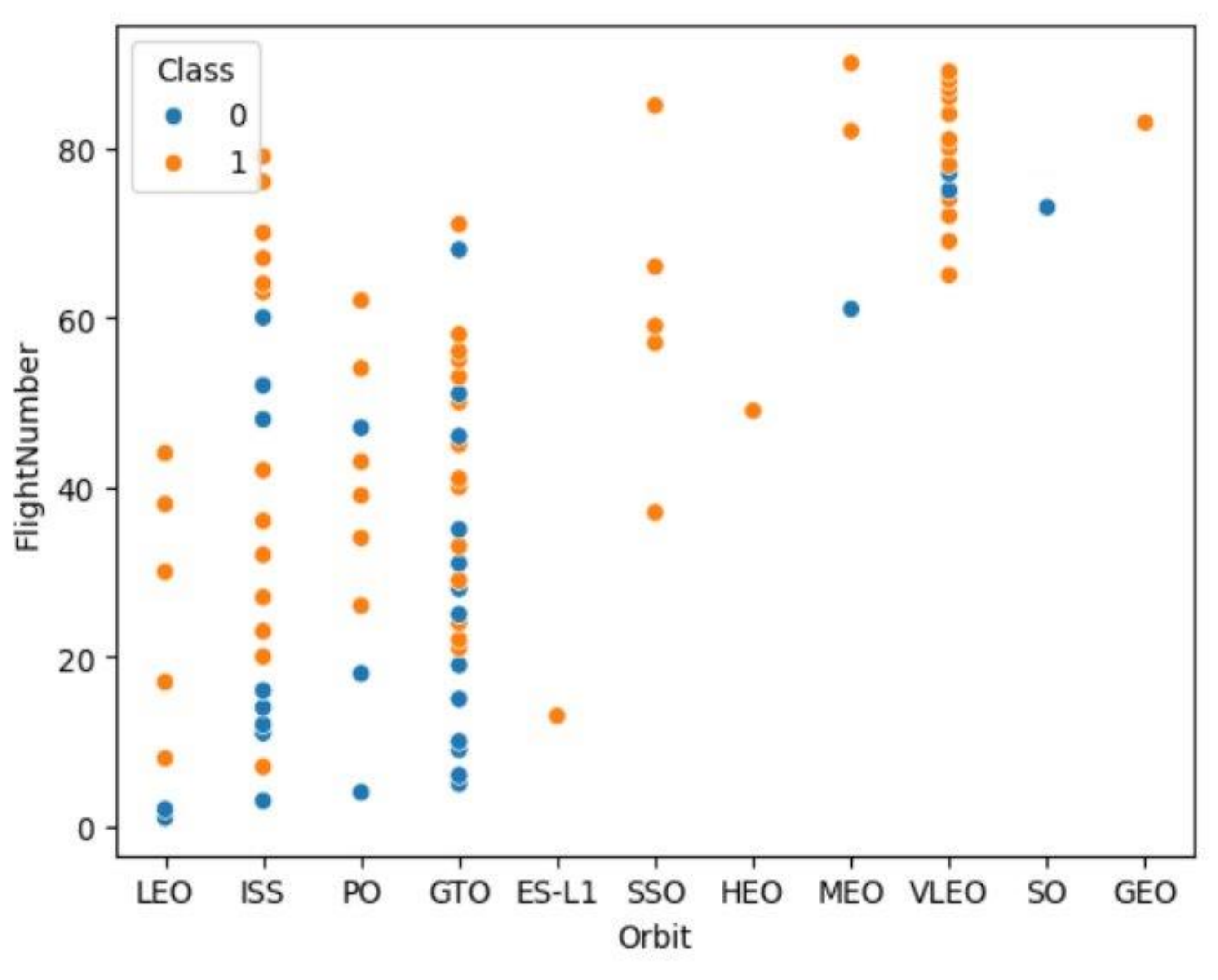
Success Rate vs. Orbit Type

- Results of Task 3: Visualize the relationship between success rate of each orbit type
- Insights Gathered : ES –L1 , SSO,HEO and GEO have the highest success rate whereas SO has the lowest success rate



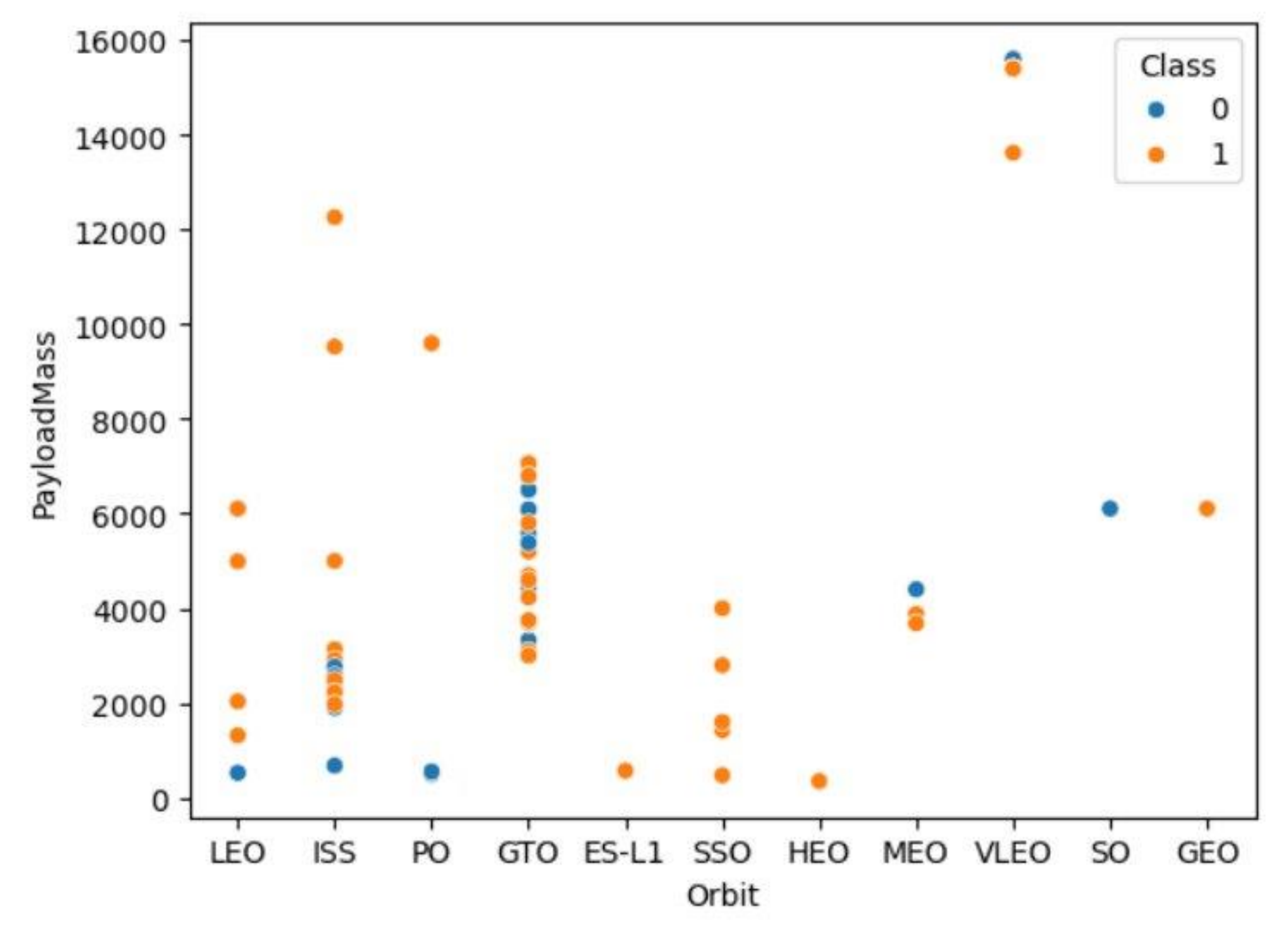
Flight Number vs. Orbit Type

- Results of Task 4 : Visualize the relationship between FlightNumber and Orbit type
- Insights Gathered : In LEO orbit, higher number of flights has higher success rate. IN SSO, each flight has high success rate but there is no clear pattern fr the rest of the orbits



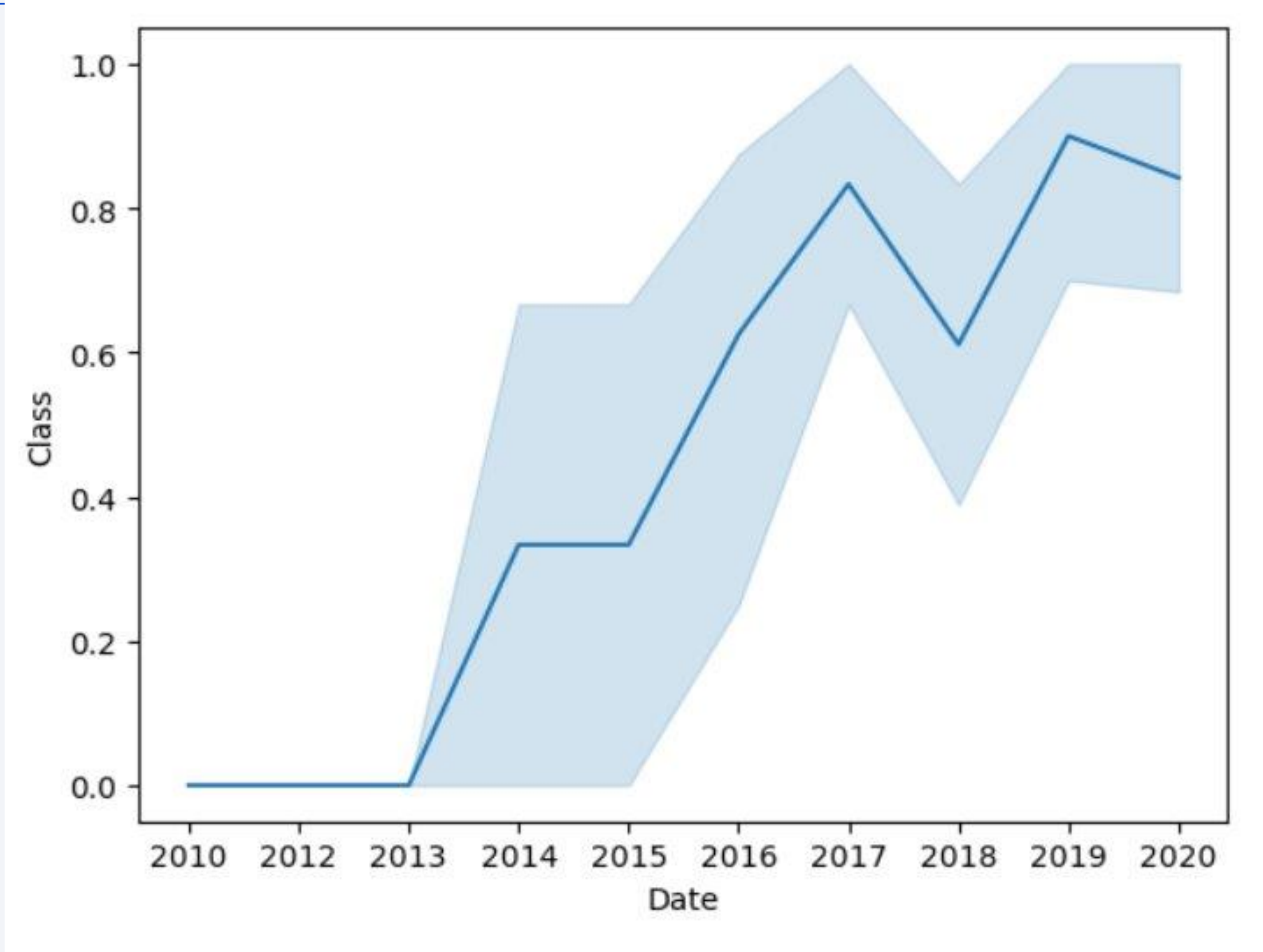
Payload vs. Orbit Type

- Results of Task 5: Visualize the relationship between Payload Mass and Orbit type
- Insights Gathered : for orbit types LEO, SSO, higher payload mass guarentees success rate



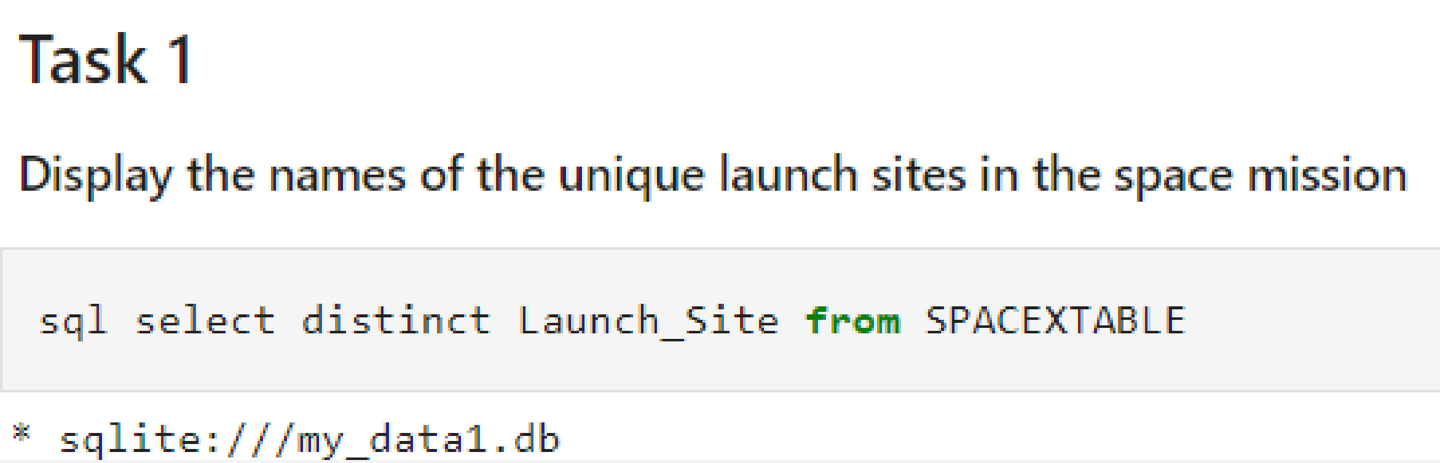
Launch Success Yearly Trend

- Results of Task 6: Visualize the launch success yearly trend
- Insights Gathered : there is a steady increase in success rate form 2010 to 2020 , with a slight dip in 2018



All Launch Site Names

- Results of Task 1:Display the names of the unique launch sites in the space mission
- Query



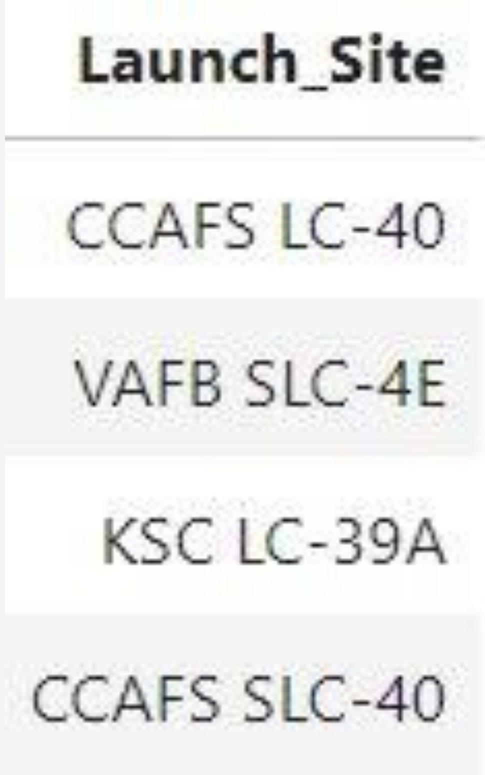
Task 1

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

```
sql select distinct Launch_Site from SPACEXTABLE
```

```
* sqlite:///my_data1.db
```

Result

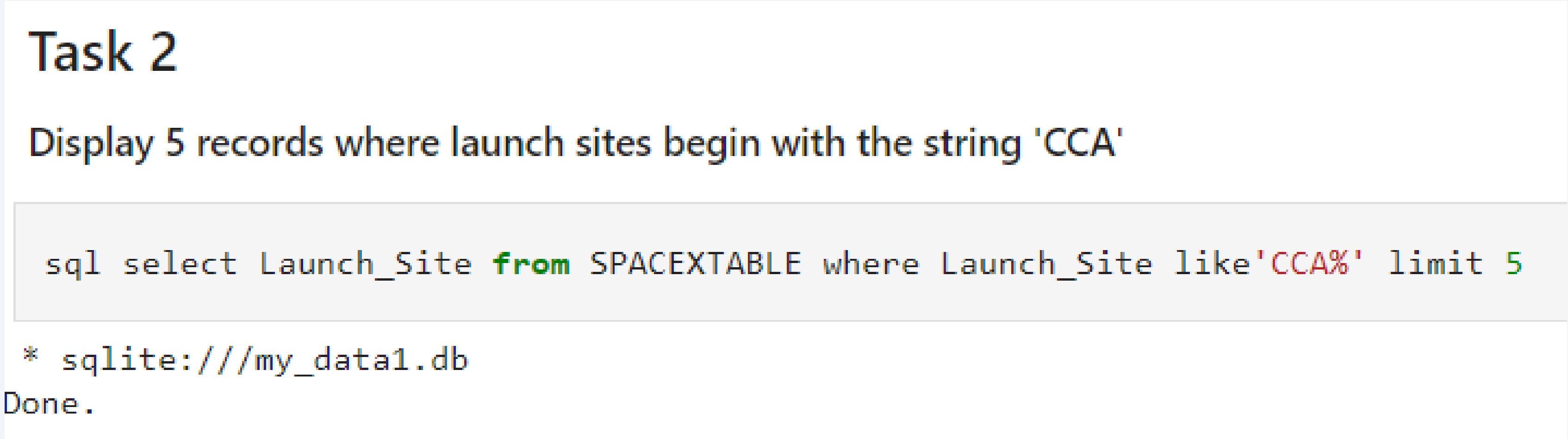


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Results of Task 2:Display 5 records where launch sites begin with the string 'CCA'
- Query

Result

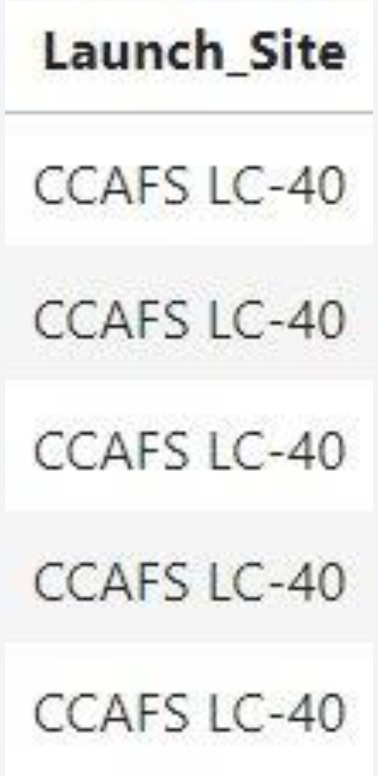


Task 2

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

```
sql select Launch_Site from SPACEXTABLE where Launch_Site like'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

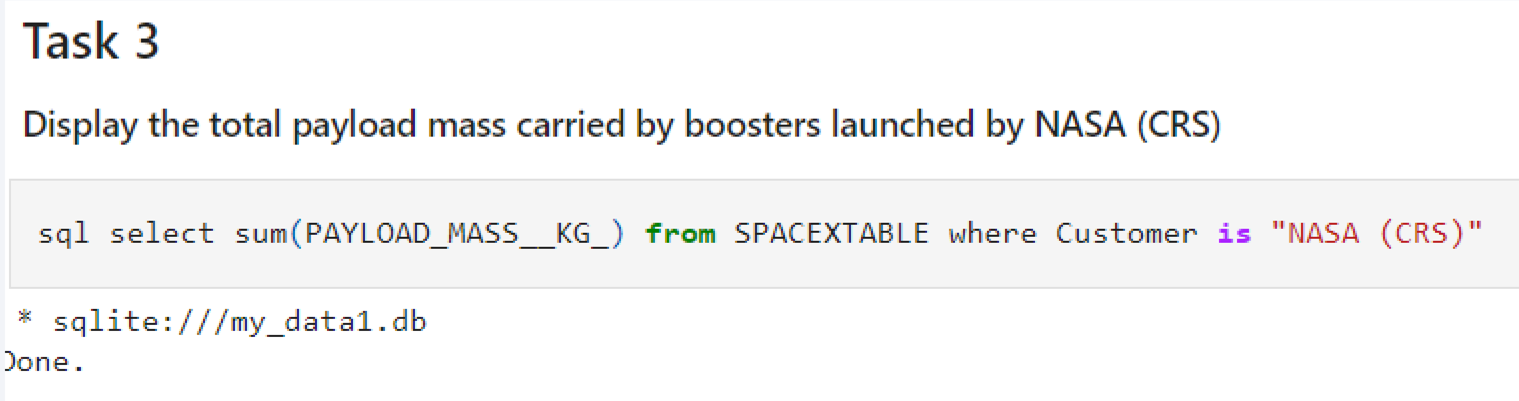


Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Results of Task 3:Display the total payload mass carried by boosters launched by NASA (CRS)
- Query

Result

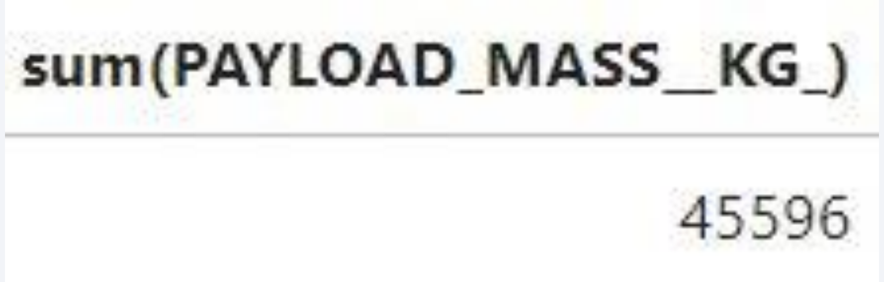


Task 3

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

```
sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer is "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```



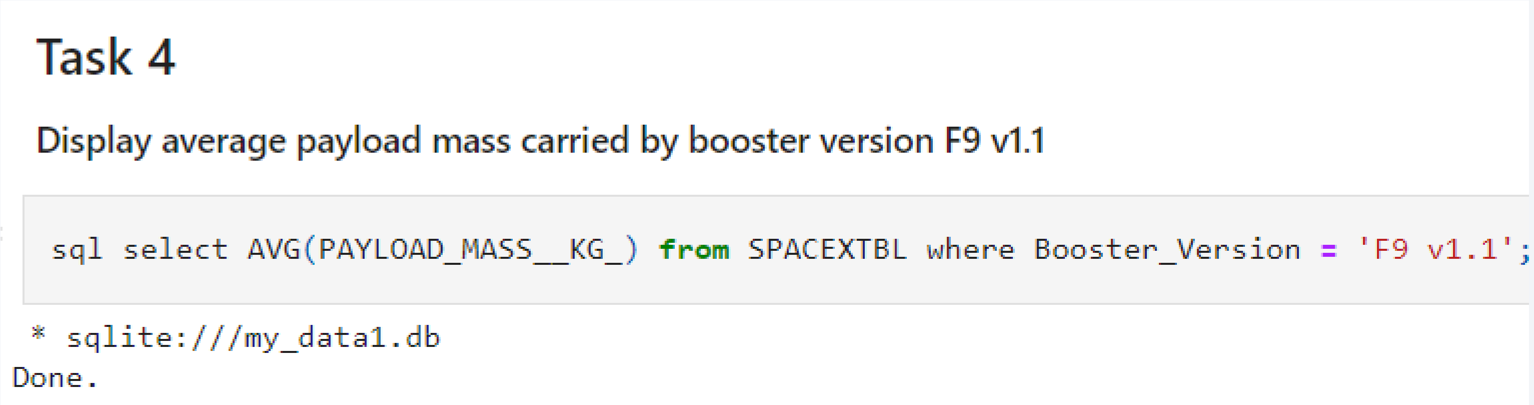
sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Results of Task 4:Display average payload mass carried by booster version F9 v1.1

- Query

Result

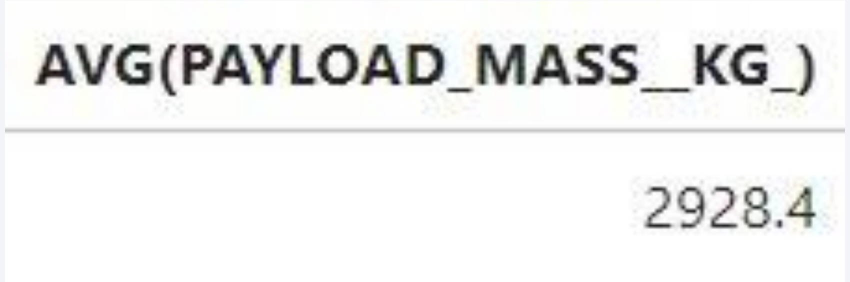


Task 4

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';
```

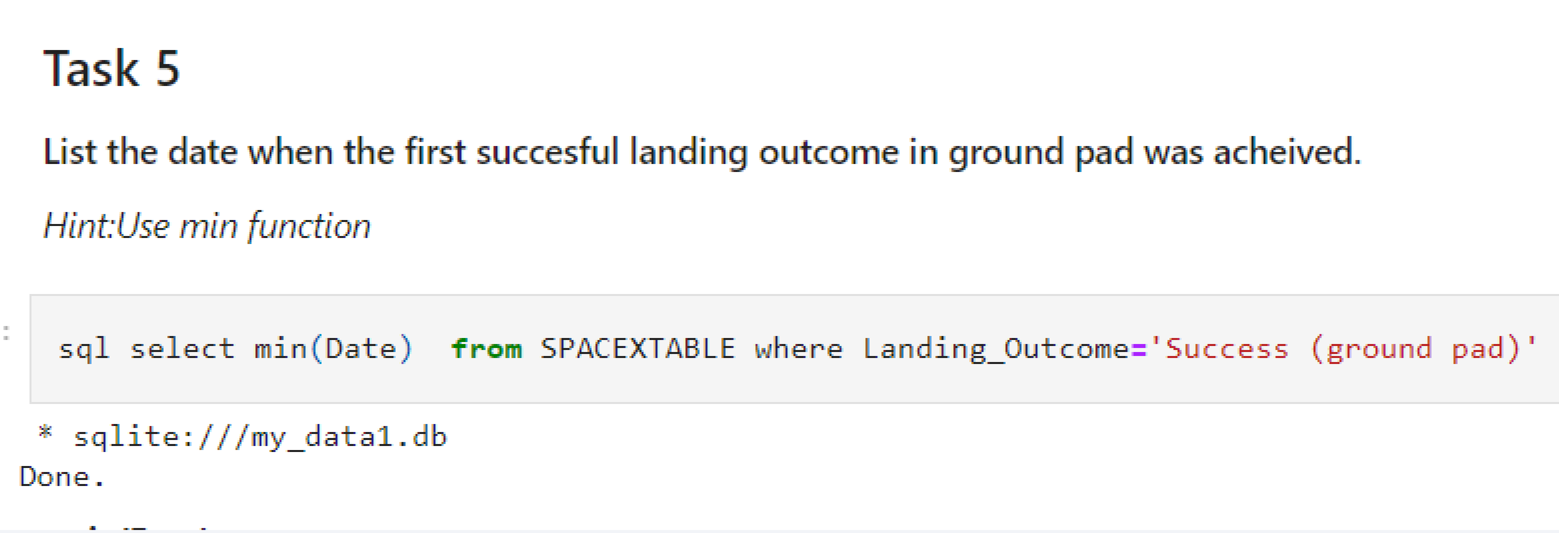
```
 * sqlite:///my_data1.db
Done.
```



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Results of Task 5:List the date when the first succesful landing outcome in ground pad was acheived.



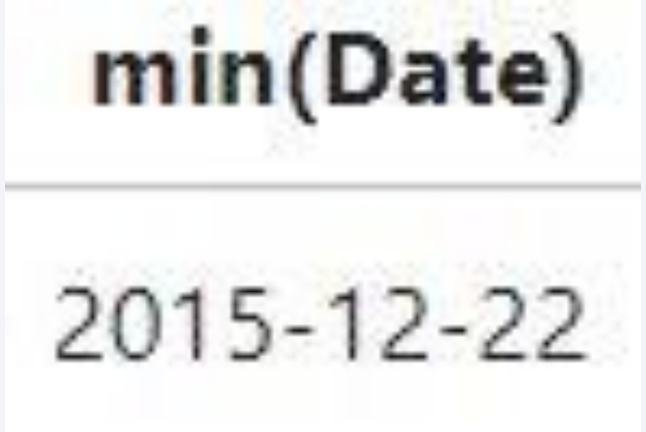
Task 5

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

Hint:Use min function

```
sql select min(Date)  from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

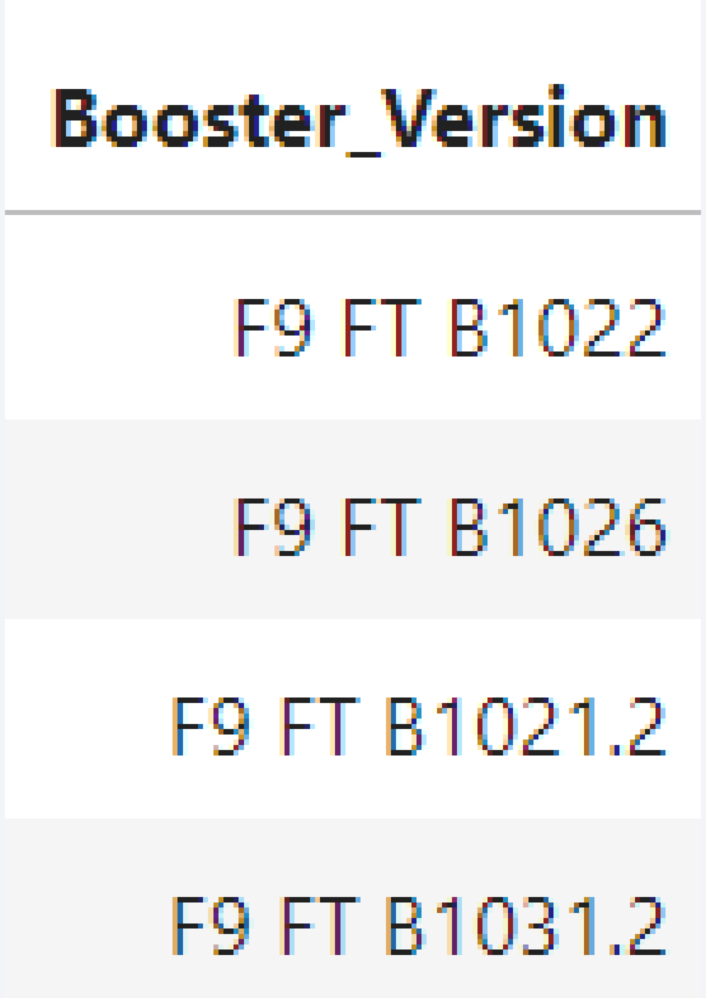


min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Results of Task 6:List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

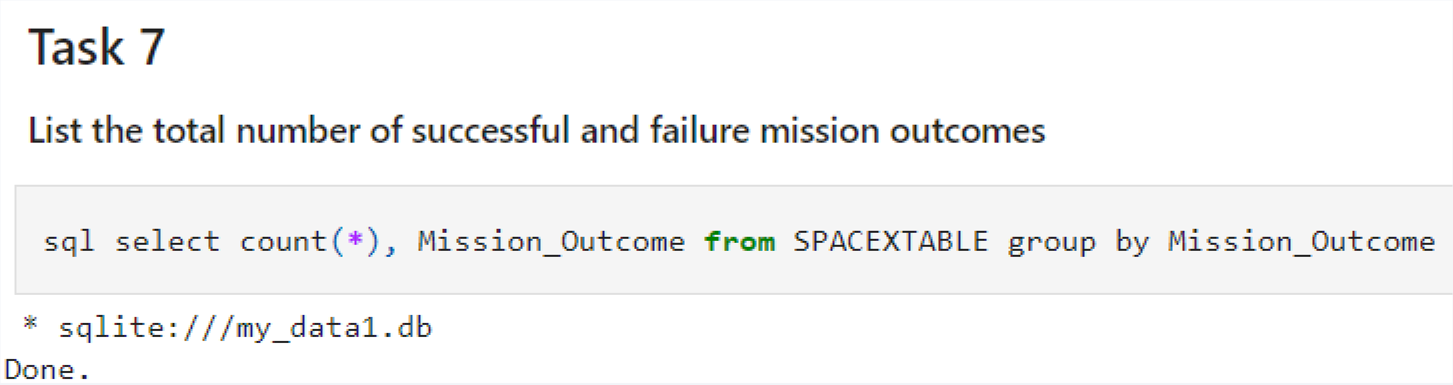
sql select distinct Booster_Version **from** SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 **and** 6000 **and** Landing_Outcome = 'Success (drone ship)'



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

Total Number of Successful and Failure Mission Outcomes

- Results of Task 7: List the total number of successful and failure mission outcomes

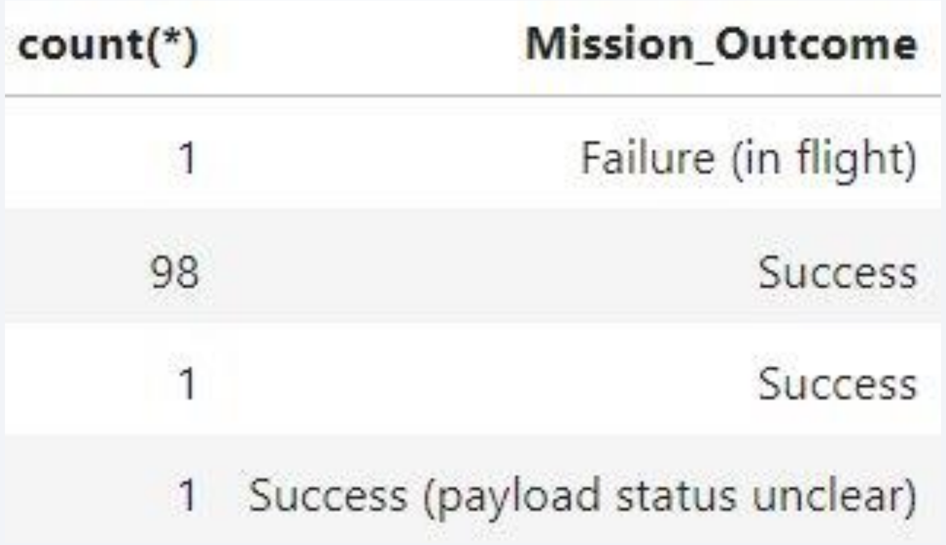


Task 7

List the total number of successful and failure mission outcomes

```
sql select count(*), Mission_Outcome from SPACEXTABLE group by Mission_Outcome
```

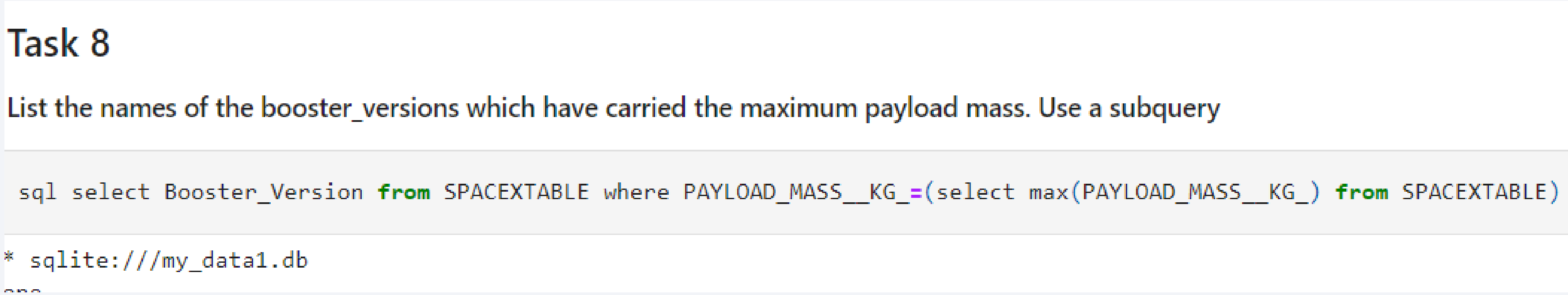
```
 * sqlite:///my_data1.db
Done.
```



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

Boosters Carried Maximum Payload

- Results of Task 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

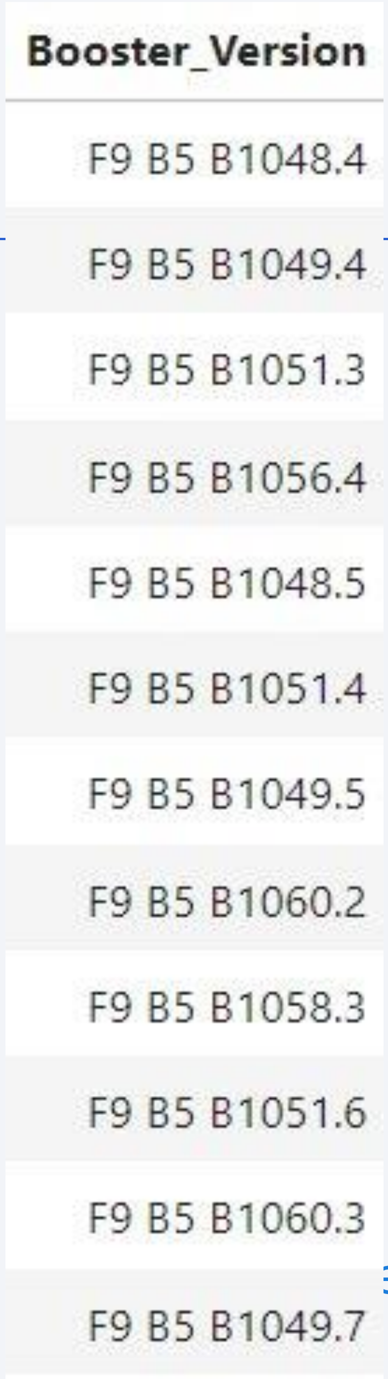


Task 8

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

```
sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
* sqlite:///my_data1.db
```

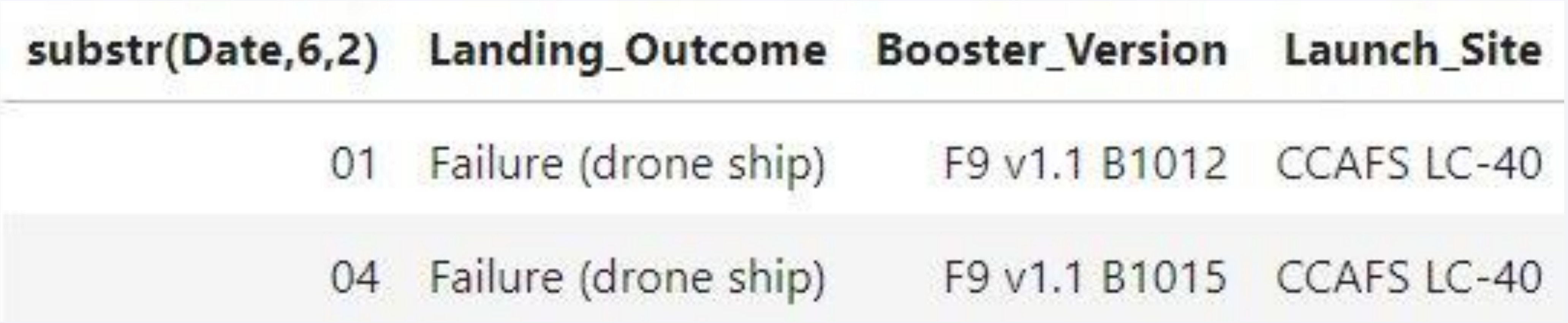


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Results of Task 9: 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.

```
sql select substr(Date,6,2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE
where substr(Date,0,5)=='2015' and Landing_Outcome='Failure (drone ship)'
```

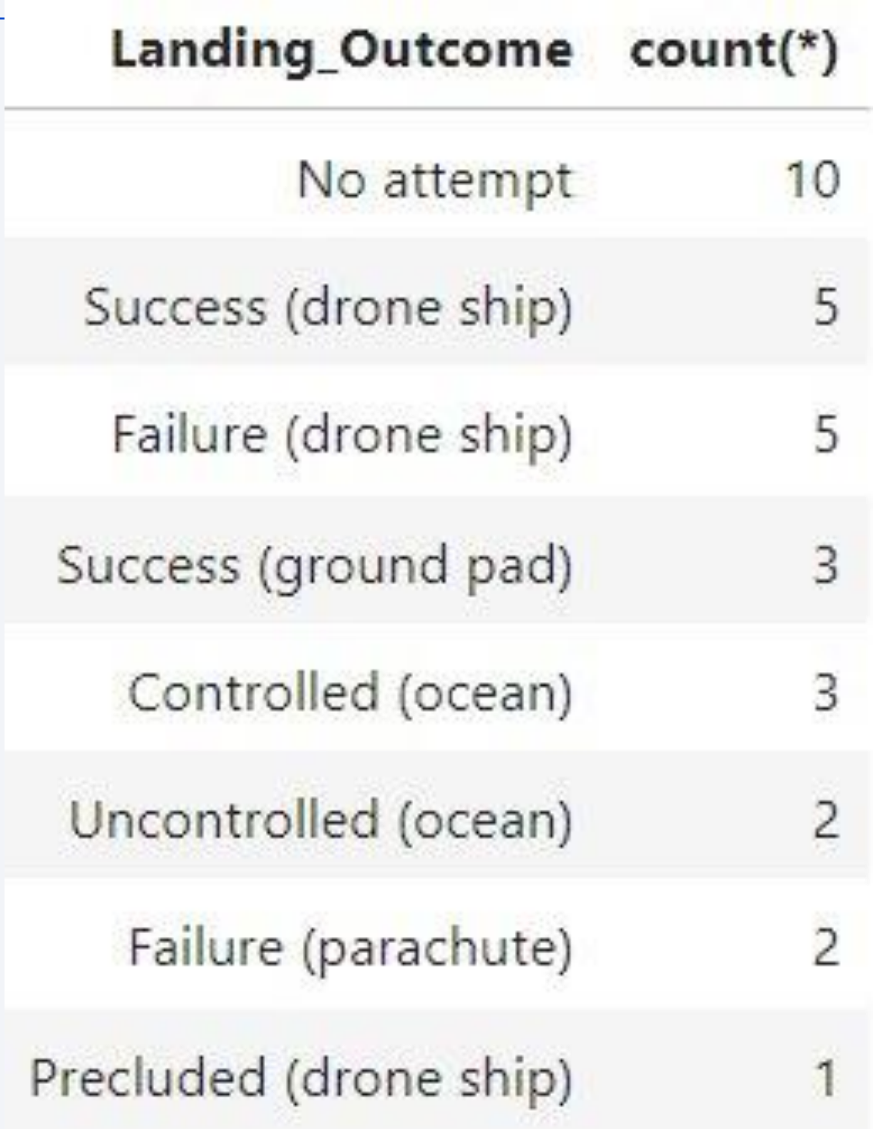


substr(Date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Results of Task 10: 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(*) **from** SPACEXTABLE where date between '2010-06-04' **and** '2017-03-20' group by Landing_Outcome order by count(*) desc

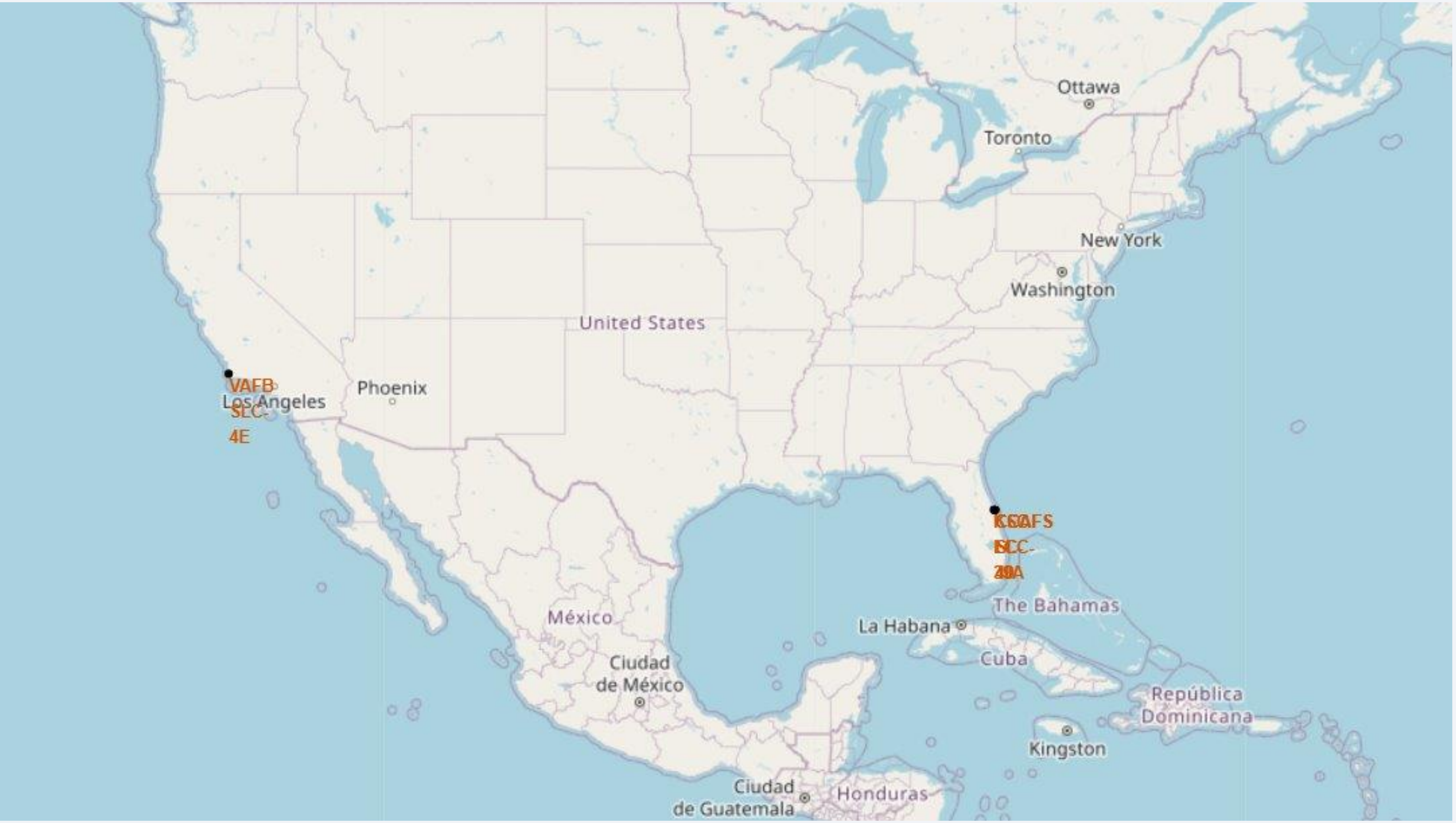


Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

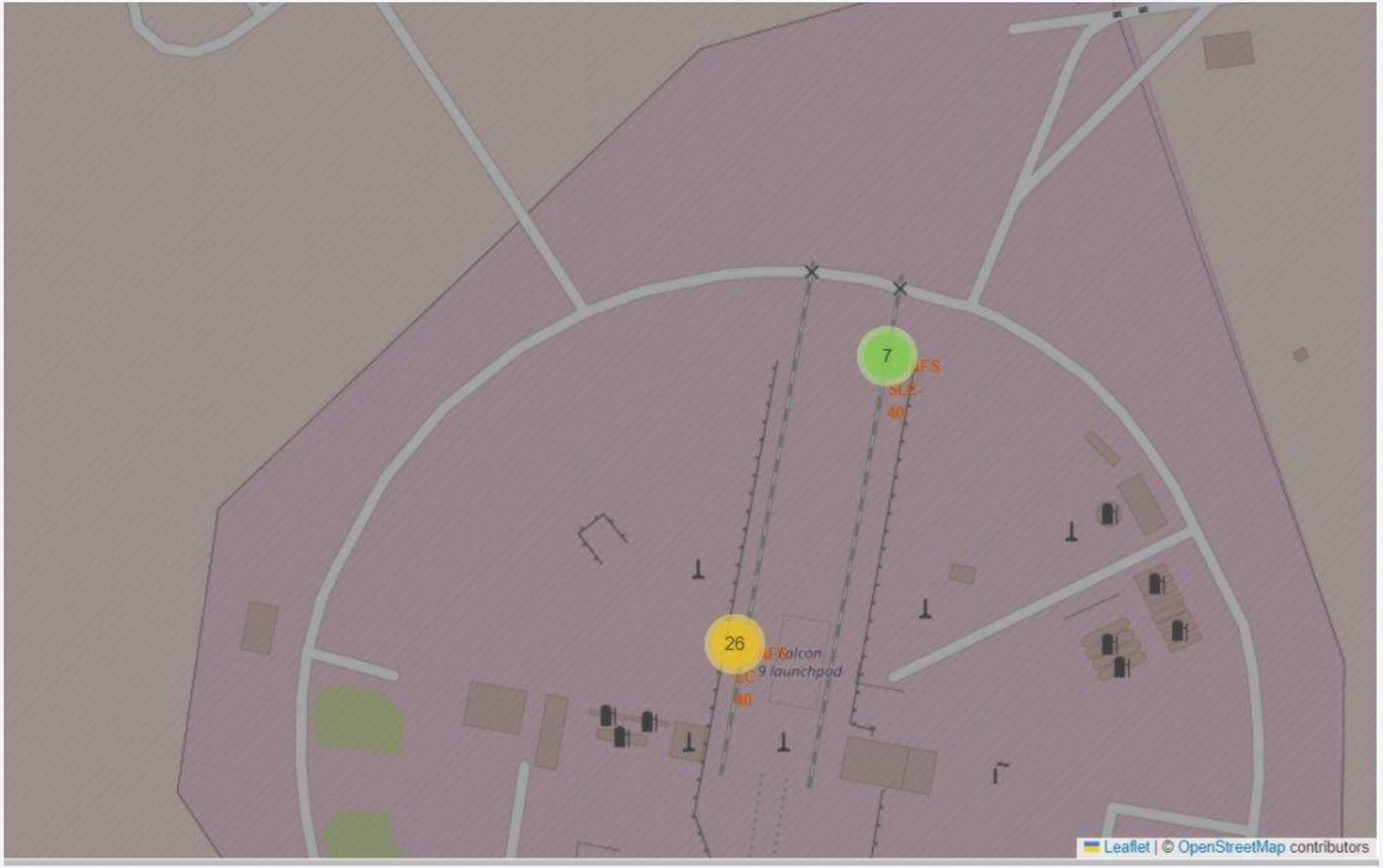
Section 3

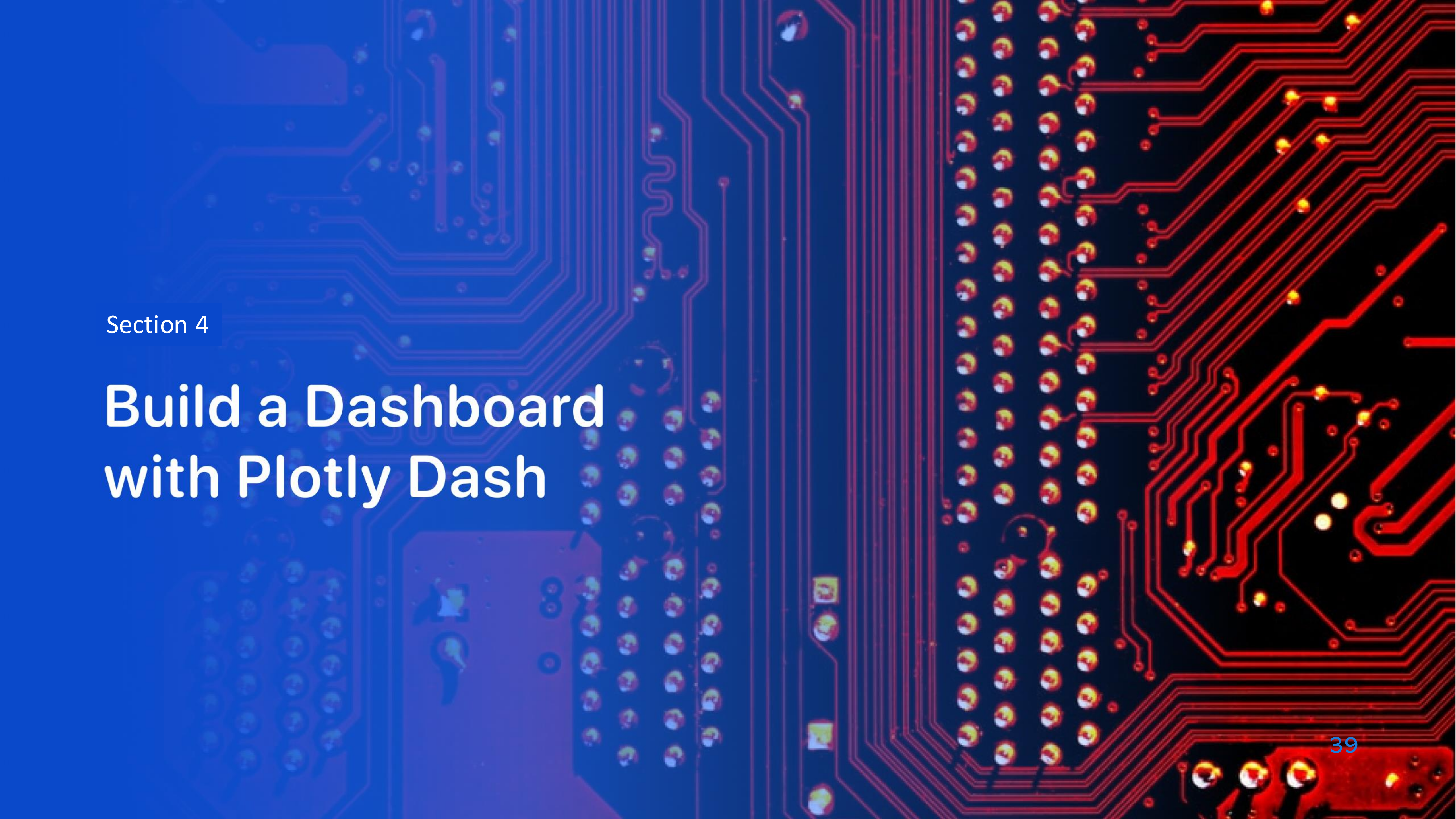
Launch Sites Proximities Analysis

Marking Launch Sites on the Map



Color-labeled launch outcomes





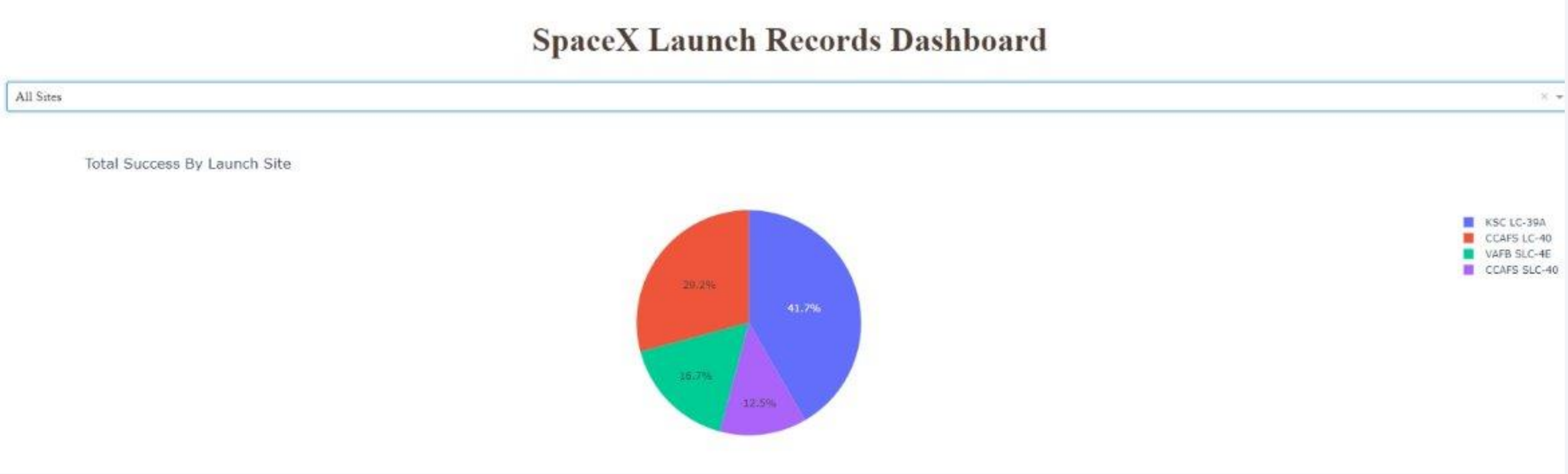
Section 4

Build a Dashboard with Plotly Dash

Total Success By Launch Site

+ KSC LC 39 A has the highest total success by launch site

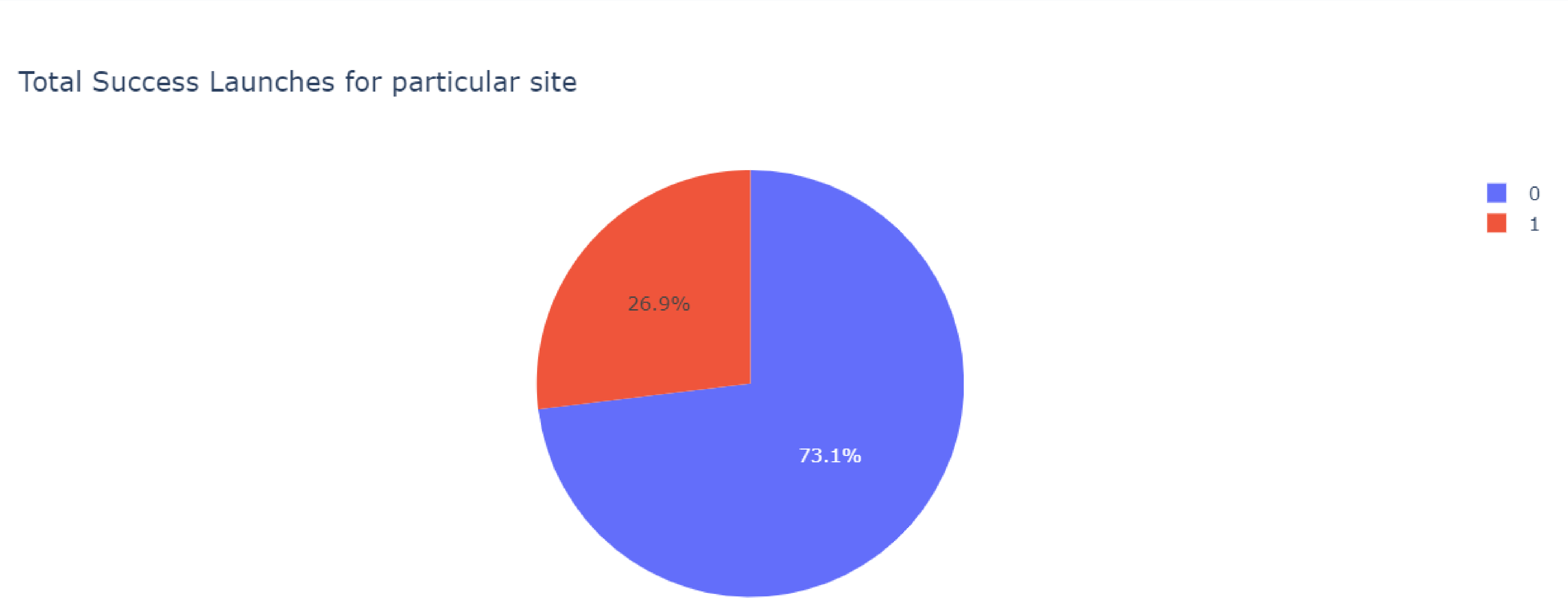
+ CCAFS SLC 40 has the least success by launch site



SpaceX Launch Records Dashboard

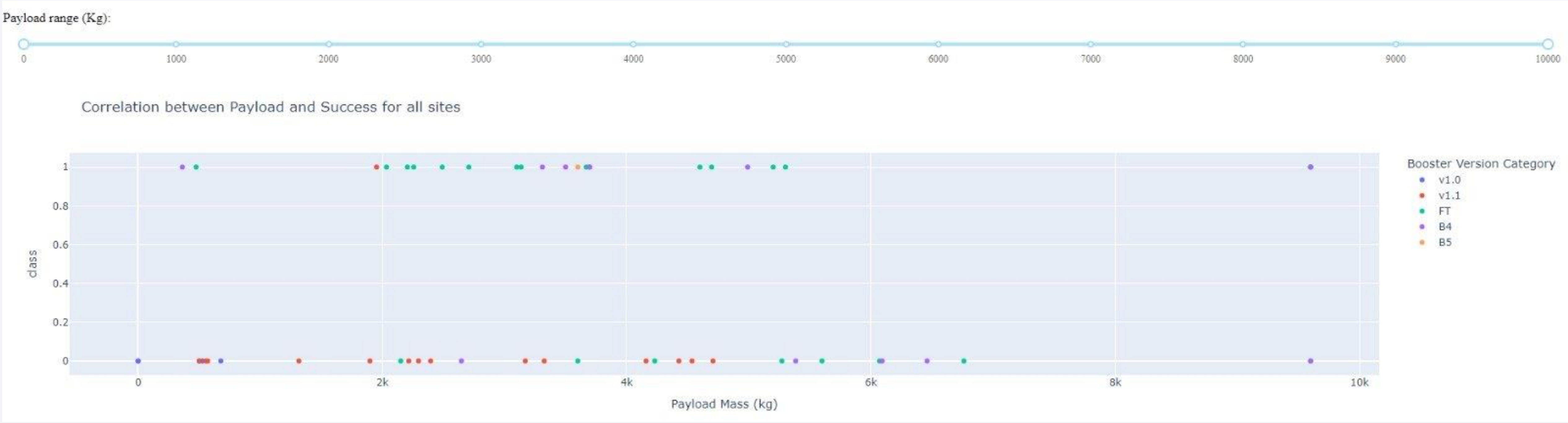
All Sites

Total Success Launches for a particular site



Correlation between Payload and Launch Outcome

- Graph for Correlation between Payload and Success for all sites
- Made interactive using the filters of payload range and Launch sites drop down
- 2k to 6k payload range has the highest amount of success rate
- FT Booster Version has the highest amount of success rate



Payload range (Kg):

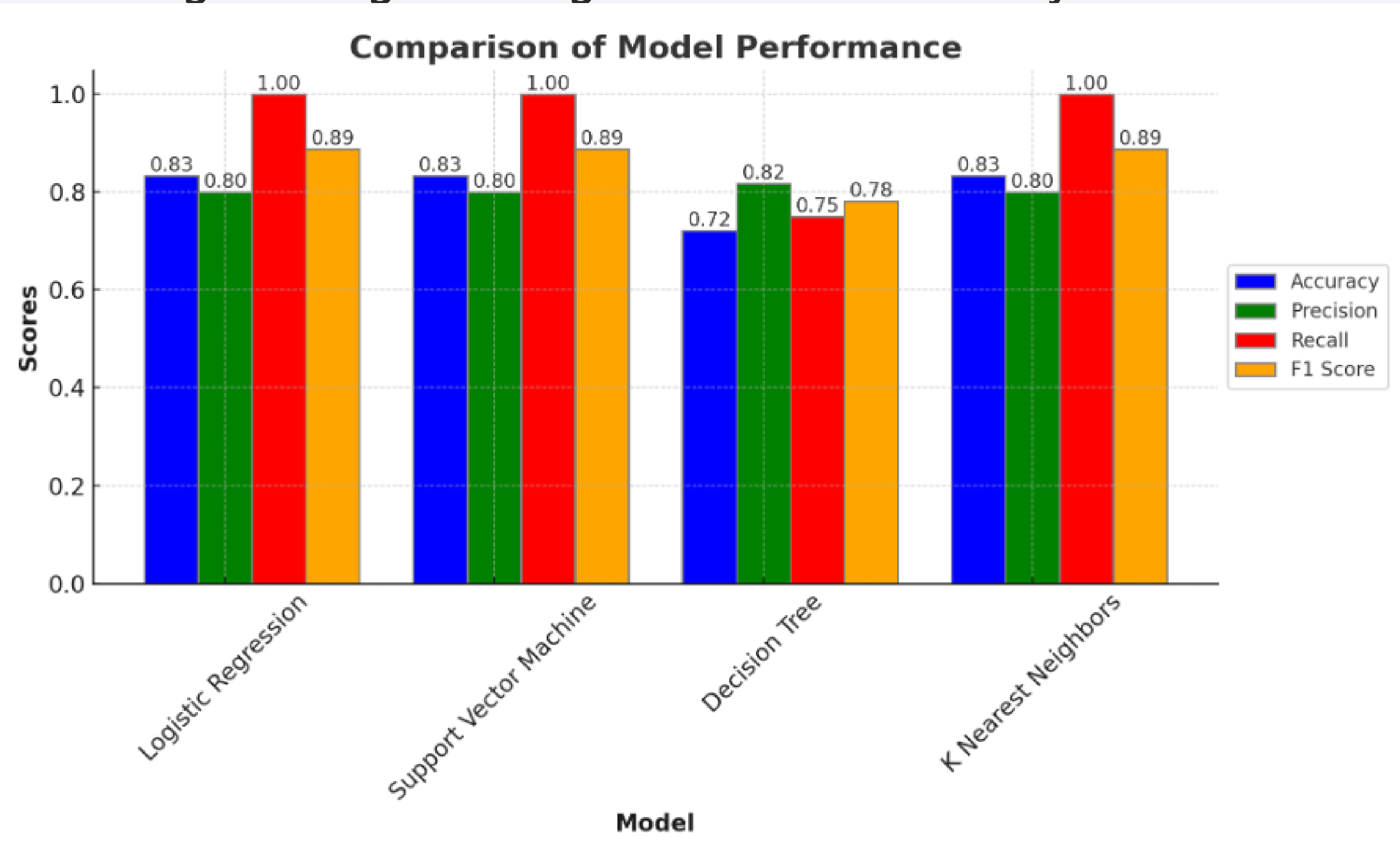
Correlation between Payload and Success for all sites

Section 5

Predictive Analysis (Classification)

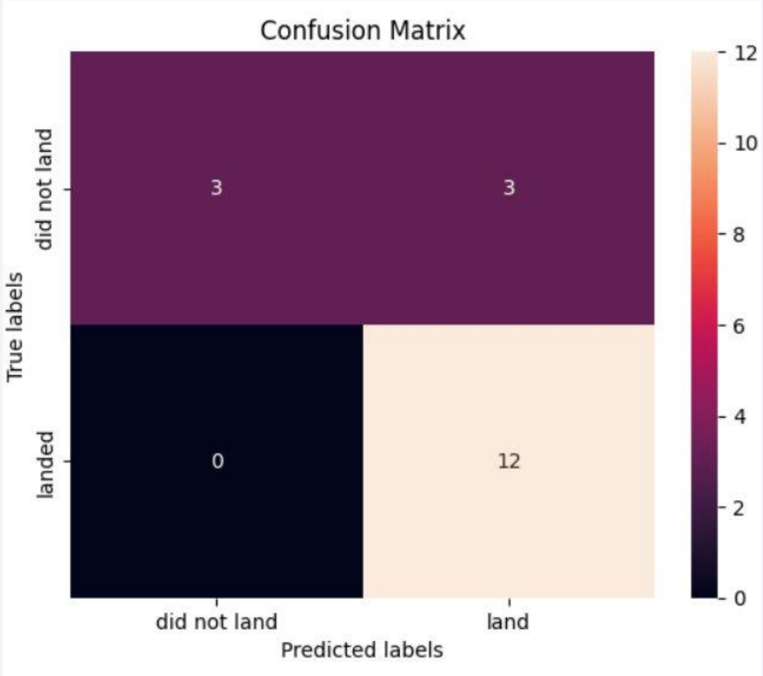
Classification Accuracy

- KNN, SVM and Logistic Regression give the same accuracy of 83.33%

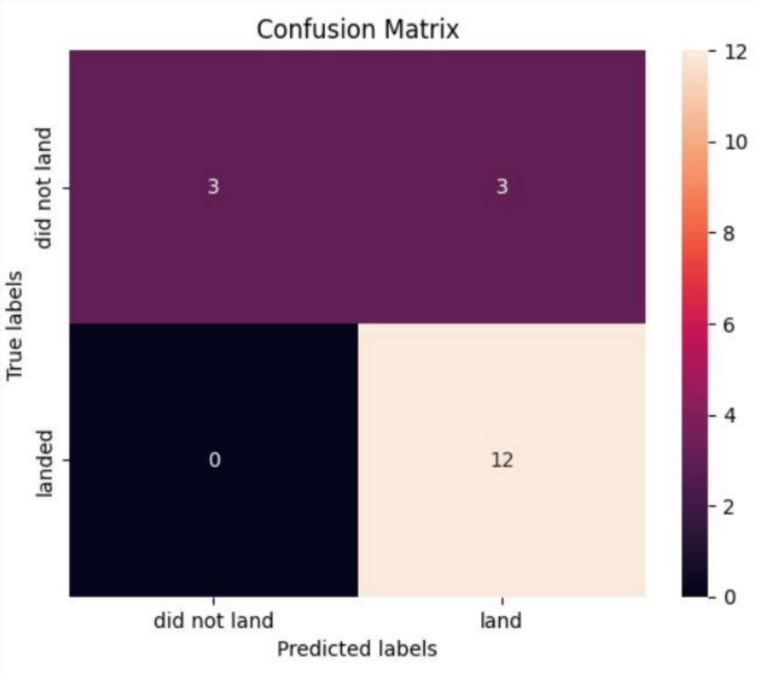


Confusion Matrix

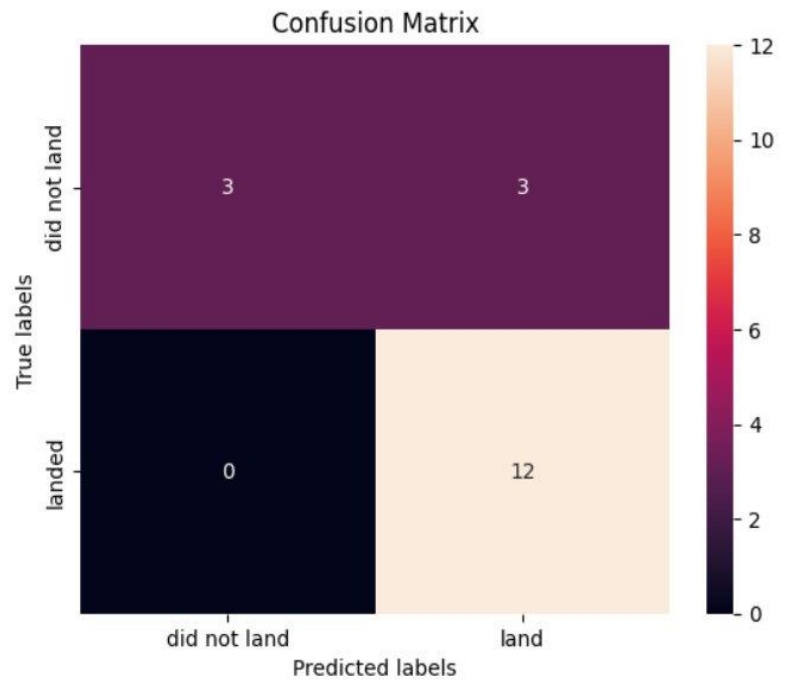
- Confusion Matrix for Logistic Regression



- Confusion Matrix for SVM



- Confusion Matrix for KNN



Conclusions

- KNN, SVM and Logistic Regression are the best performing models
- Flight Numbers are higher in CCAFS SLC 40 , with most of them being successful
- CCFAS SLC 40 AND KSC LC 39 A have the highest Payload Masses recorded
- KSC LC 39 A has the highest total success by launch site
- CCAFS SLC 40 has the least success by launch site
- 2k to 6k payload range has the highest amount of success rate
- FT Booster Version has the highest amount of success rate

Appendix

Table of Comparison for Different Models and their Best Parameters Chosen using Grid Search CV

Model	Best Parameters Chosen
Logistic Regression	'C'= 0.01, 'penalty'='l2', 'solver'= 'lbfgs'
Support Vector Machine	'C'= 1.0, 'gamma'= 0.03162277660168379, 'kernel'= 'sigmoid'
Decision Tree	'criterion'='gini', 'max_depth'= 2, 'max_features'= 'sqrt', 'min_samples_leaf'= 1, 'min_samples_split'= 5, 'splitter'= 'random'
K Nearest Neighbors	'algorithm'= 'auto', 'n_neighbors'= 10, 'p'= 1

Thank you!

By Amruha Ahmed