IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sachin Agarwal

July 6, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis and Data visualization
  - Exploratory Data Analysis with SQL
  - Data visualization with Folium and Plotly Dash
  - Model building for predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Analysis screenshots
  - Model results

Introduction

- Project background and context
  - In the commercial space age, SpaceX is the most successful company, inter-alia other competitors, making space travel affordable.
  - Its Falcon 9 rocket launches costs of $ 62 million as compared to competitors costs upwards of $ 165 million dollars each.
    - Mainly due to reuse of first stage of rocket launch
  - ***Objective:*** As data scientist for new rocket company, SpaceY, if we can determine if the first stage will land, we can determine the cost of a launch.
    - To determine the price of each launch.
    - Gather information about Space X and create dashboards for the team.
    - To determine if SpaceX will reuse the first stage, through machine learning model and use of public information.

Section 1

Methodology

Methodology

Executive Summary

- Data collection strategy:
  - Using dual-source data collection approach through:
    - SpaceX REST API
    - Wikipedia Web Scraping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model

Data Collection

- ❑ Data collection strategy:
  - ❑ To ensure a comprehensive dataset for SpaceX launch analysis, dual-source data collection approach was employed:

a) **SpaceX REST API**

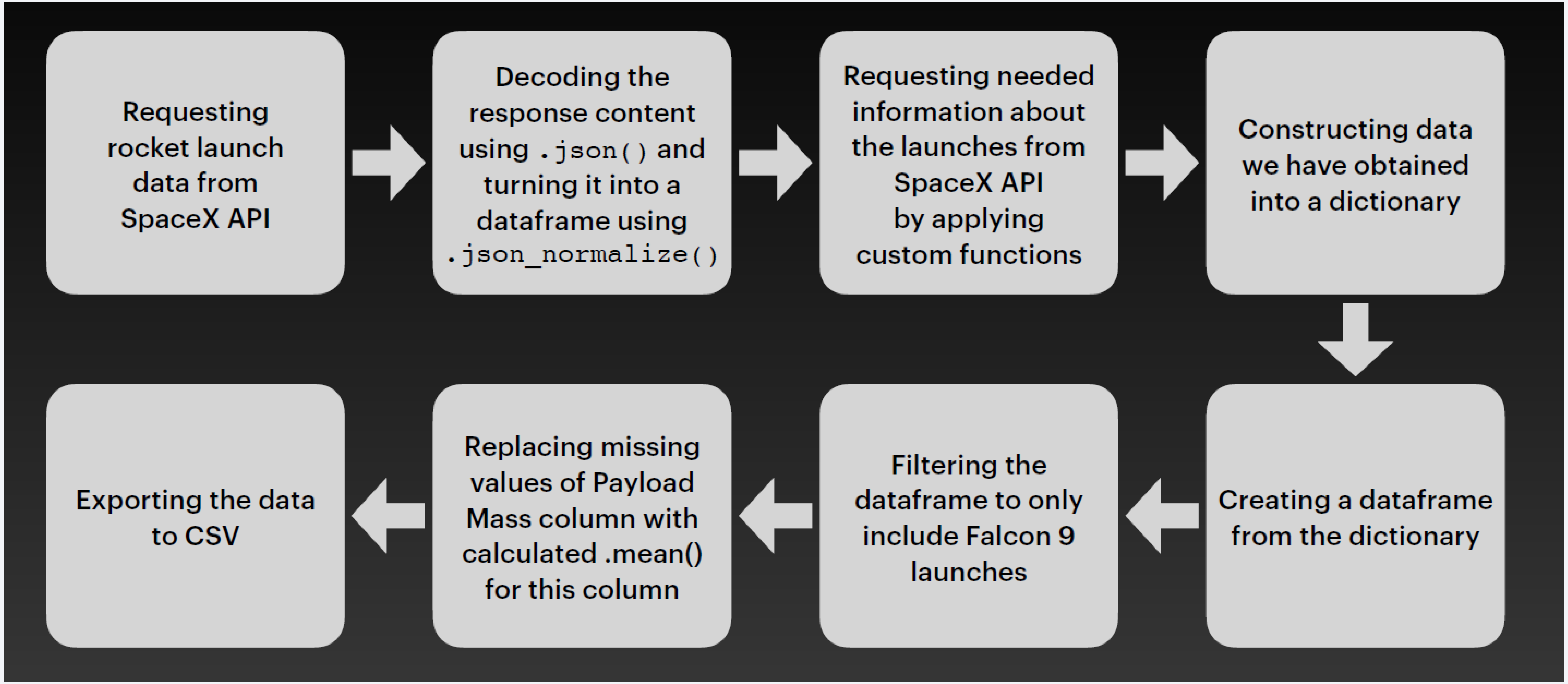
- ❑ Structured and detailed mission data was retrieved directly via API calls, capturing key technical & flight attributes such as Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Landing Pad, Longitude, Latitude, etc.

b) **Wikipedia Web Scraping**

- ❑ To enrich the dataset with additional context and fill in gaps, tabular launch data from SpaceX's Wikipedia page was extracted such as Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster version and Landing, etc.

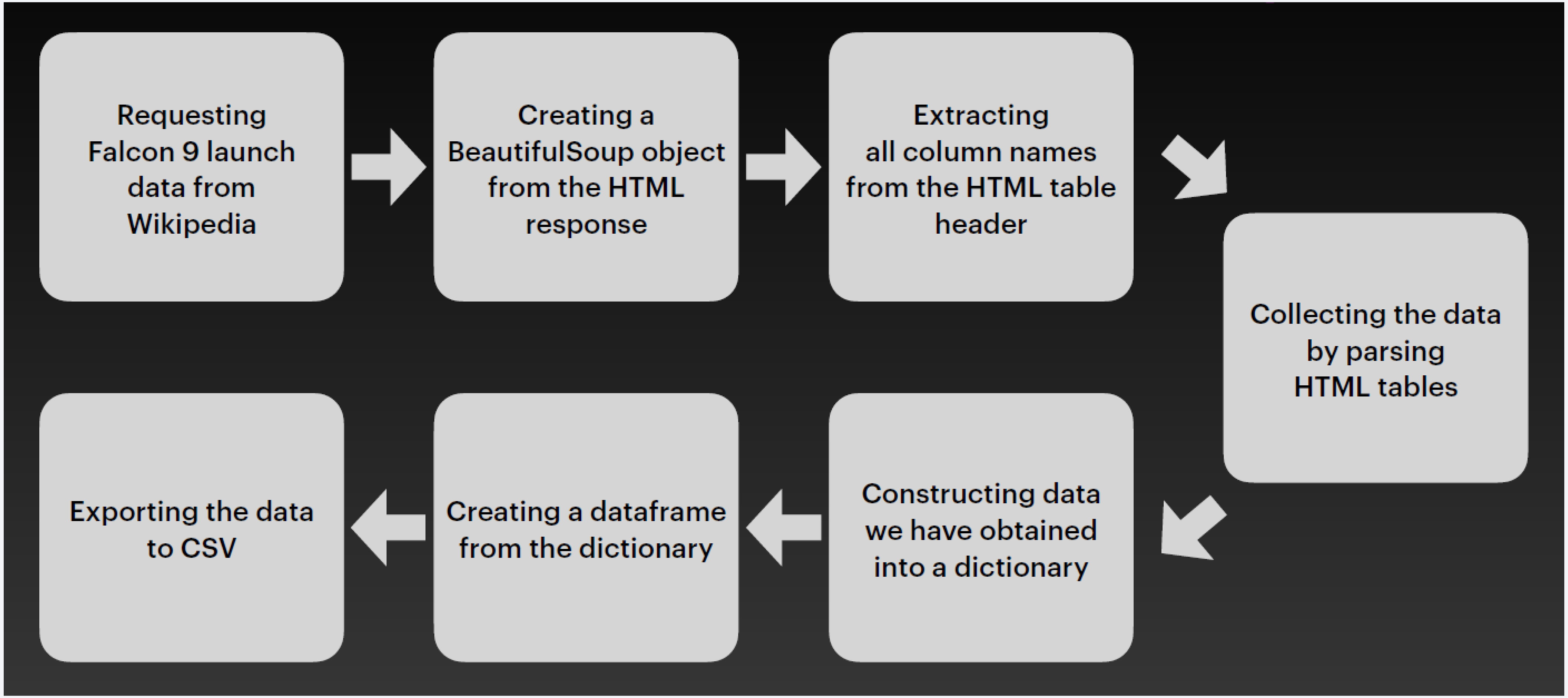
- ❑ By integrating both sources, a richer and more complete dataset suitable for in-depth analysis and insights into SpaceX launch operations was compiled.

Data Collection – SpaceX API



https://github.com/a1sachin/SpaceX-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping



https://github.com/a1sachin/SpaceX-project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

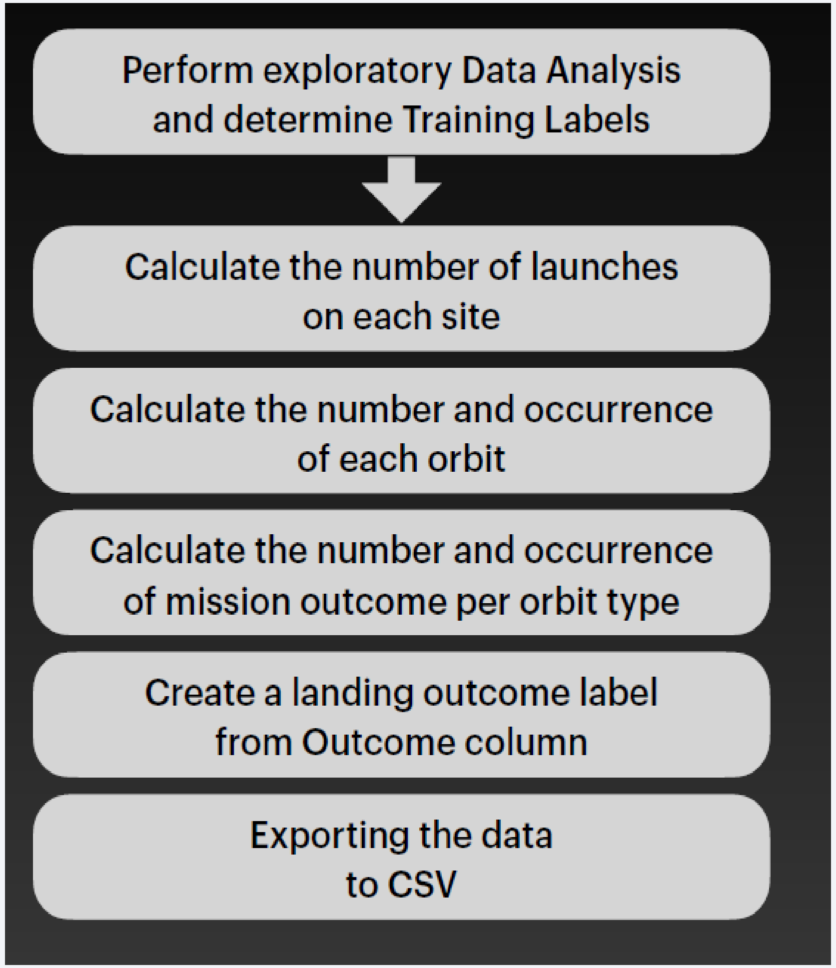
Booster Landing Outcome – Label Encoding Logic used for various types of landing zones and landing outcomes:

❑ **Types of Landing Zones**

- RTLS (Return to Launch Site): Landing attempt on a ground pad near the launch site
- ASDS (Autonomous Spaceport Drone Ship): Landing attempt on a floating drone ship in the ocean
- Ocean: Landing directly in the sea - planned or unplanned

❑ **Landing Outcome Variants**

- True [Zone] → Booster landed successfully (e.g., True RTLS) → 1
- False [Zone] → Booster attempted landing but failed (e.g., False ASDS) → 0



https://github.com/a1sachin/SpaceX-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **EDA was carried out using various data visualization charts:**
  - Flight Number vs Payload Mass : how they affect launch outcome
  - Flight Number vs Launch Site : to observe relationship between the two
  - Payload Mass vs Launch Site : to observe relationship between the two
  - Orbit Type vs Success Rate : to check any relationship between the two
  - Flight Number vs Orbit Type : to observe relationship between the two
  - Payload Mass vs Orbit Type : to observe relationship between the two
  - Success Rate Yearly Trend : to visualize yearly trend

- **If a relationship exists, they could be used in machine learning model**

https://github.com/a1sachin/SpaceX-project/blob/main/edadataviz.ipynb

EDA with SQL

- **Following SQL queries were performed:**
  - Names of the unique launch sites in the space mission
  - First 5 records of launch sites beginning with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Boosters' names having success in drone ship & have payload mass between 4000 to 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Details of failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the count of landing outcomes between 04-06-2010 and 20-03-2017

https://github.com/a1sachin/SpaceX-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

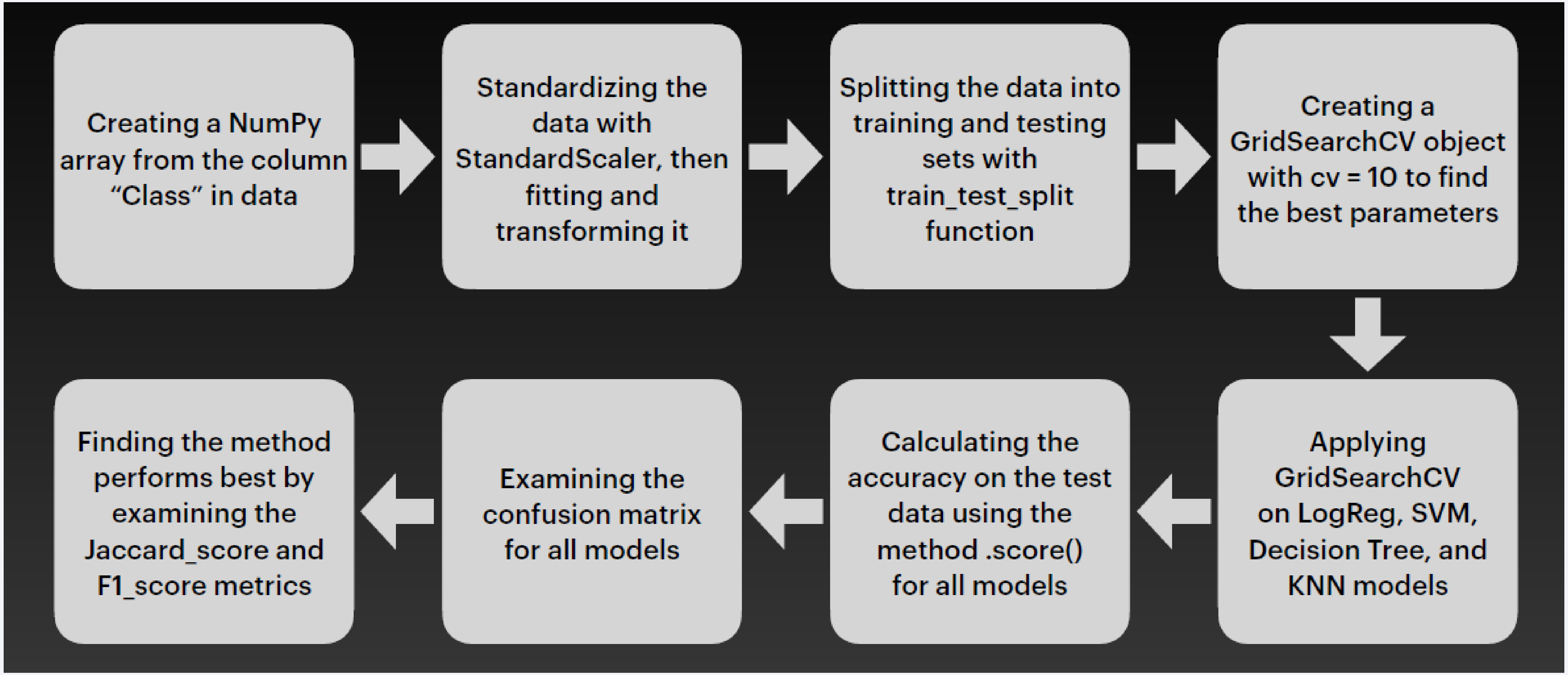
- **Markers of all Launch Sites:**
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- **Colored Markers of the launch outcomes for each Launch Site:**
  - Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- **Distances between a Launch Site to its proximities:**
  - Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

https://github.com/a1sachin/SpaceX-project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
  - Added a dropdown list to enable Launch Site selection.
- **Pie Chart showing Success Launches (All Sites/Certain Site):**
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- **Slider of Payload Mass Range:**
  - Added a slider to select Payload range.
- **Scatter Chart of Payload Mass vs. Success Rate for different Booster Versions:**
  - Added a scatter chart for correlation between Payload and Launch Success.

Predictive Analysis (Classification)



https://github.com/a1sachin/SpaceX-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

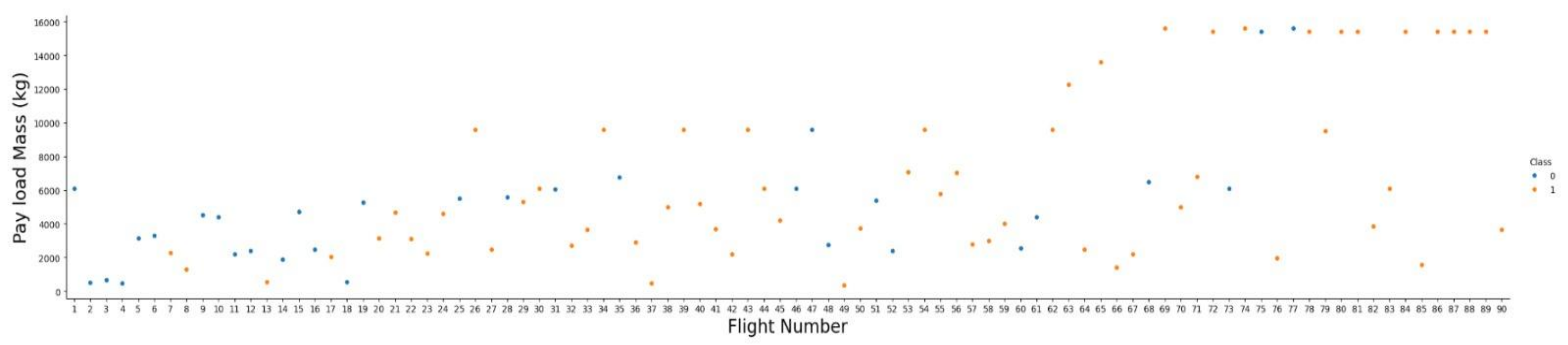
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



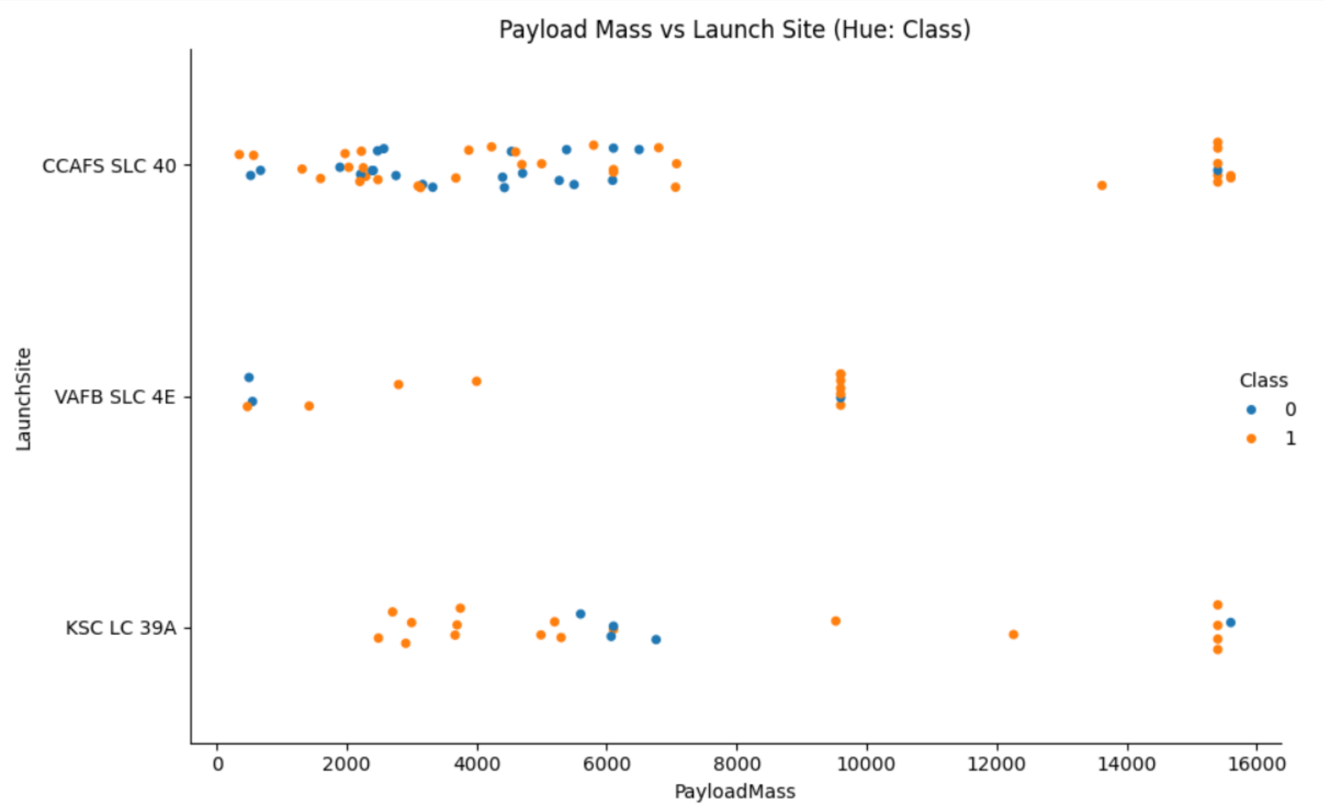
Explanation:

- ❑ Launch success increased with more number of flights. Thus, it can be assumed that every new launch has a higher success rate.
- ❑ Even with higher payload mass, the success rate was high.

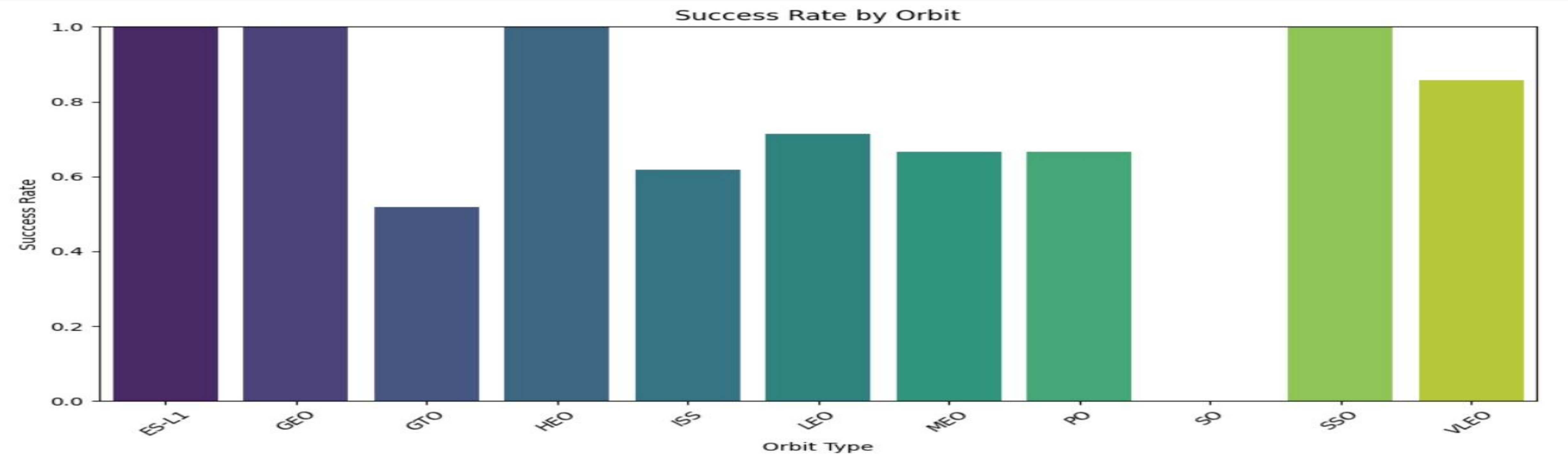
Payload vs. Launch Site

Explanation:

- Launches with payload mass over 7000 kg were highly successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kgs.



Success Rate vs. Orbit Type



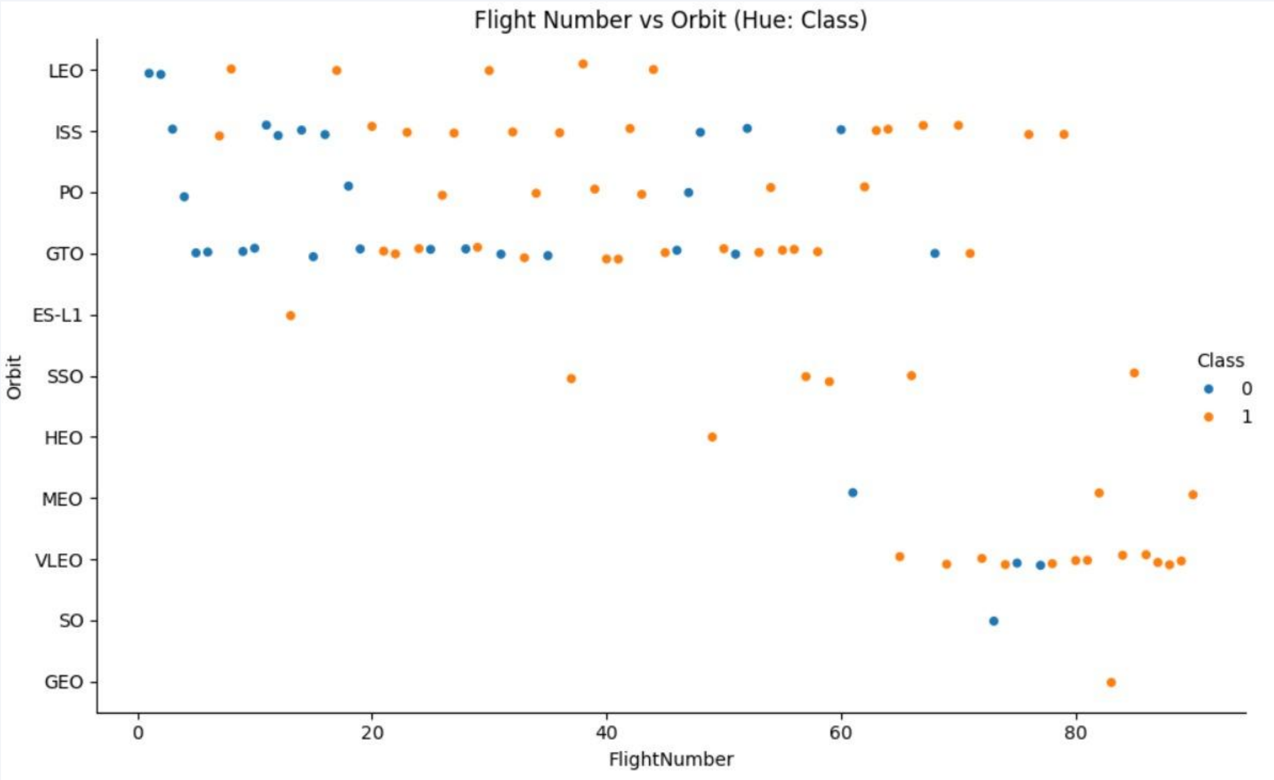
Explanation:

- ES-L1, GEO, HEO, SSO Orbits had 100% success rate
- SO orbit had 0% success rate

Flight Number vs. Orbit Type

Explanation:

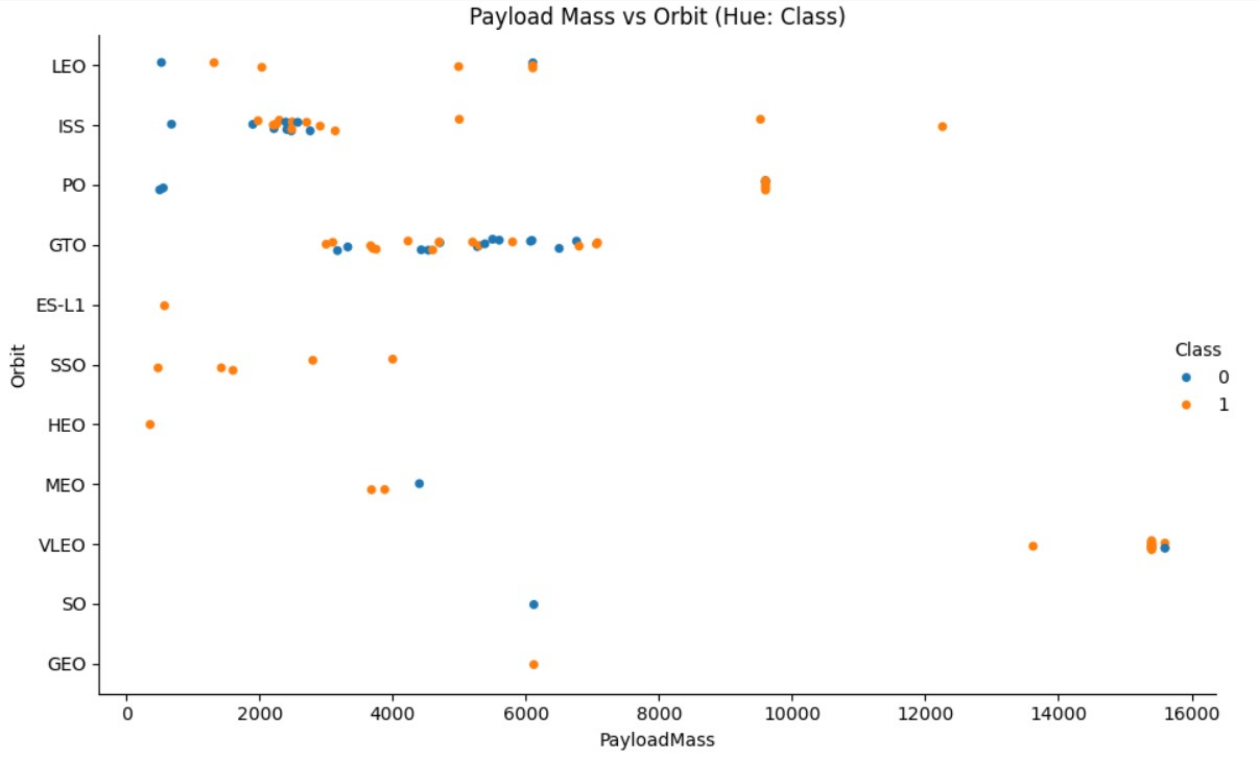
- As seen earlier, the success rate is higher with increased number of flights.
- There were higher successes observed for LEO, ISS, PO and GTO, where the number of flights ranged between 20-50.



Payload vs. Orbit Type

Explanation:

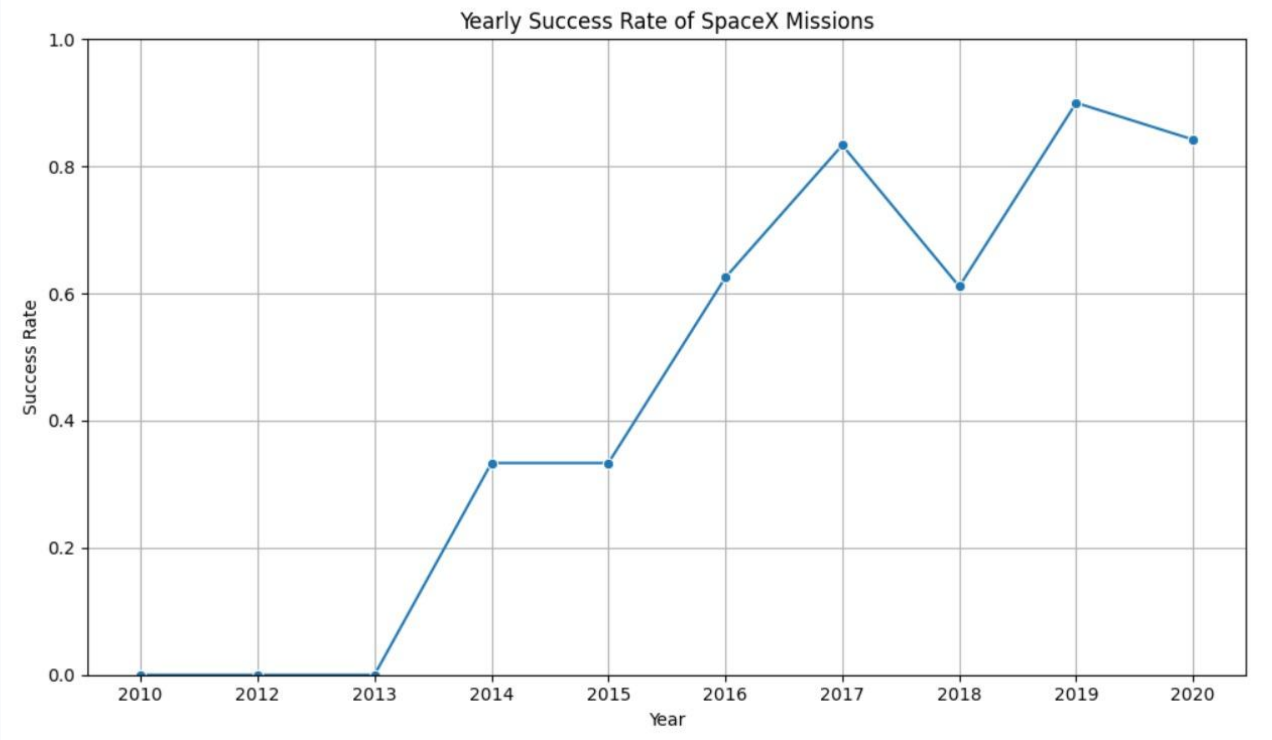
- The success ratio was lesser for payloads less than 7000 kgs, mainly for LEO, ISS, PO and GTO orbits.



Launch Success Yearly Trend

Explanation:

- Barring 2018 and 2020, the success rate increased significantly post 2013.
- The success rate was 0% till 2013.



All Launch Site Names

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

[11]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

[11]:

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'

```
[12]: %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

[12]:

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)

[13]:
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

[13]:

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

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

[14]:
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") as Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'
```

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

[14]:

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql SELECT MIN(Date) AS First_Successful_Ground_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

First_Successful_Ground_Landing
2015-12-22

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 SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAY
```

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

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 COUNT(CASE WHEN "Mission_Outcome" LIKE 'Success%' THEN 1 END) AS Successful_Missions, COUNT(CASE WHEN "Mission_Outcome" LIK
```

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

Successful_Missions	Failed_Missions
100	1

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
)
```

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

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

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
  strftime('%m', Date) AS Month_Number,
  strftime('%Y', Date) AS Year,
  CASE strftime('%m', Date)
    WHEN '01' THEN 'January'
    WHEN '02' THEN 'February'
    WHEN '03' THEN 'March'
    WHEN '04' THEN 'April'
    WHEN '05' THEN 'May'
    WHEN '06' THEN 'June'
    WHEN '07' THEN 'July'
    WHEN '08' THEN 'August'
    WHEN '09' THEN 'September'
    WHEN '10' THEN 'October'
    WHEN '11' THEN 'November'
    WHEN '12' THEN 'December'
  END AS Month_Name,
  "Landing_Outcome",
  "Booster_Version",
  "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND strftime('%Y', Date) = '2015'
```

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

Month_Number	Year	Month_Name	Landing_Outcome	Booster_Version	Launch_Site
01	2015	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

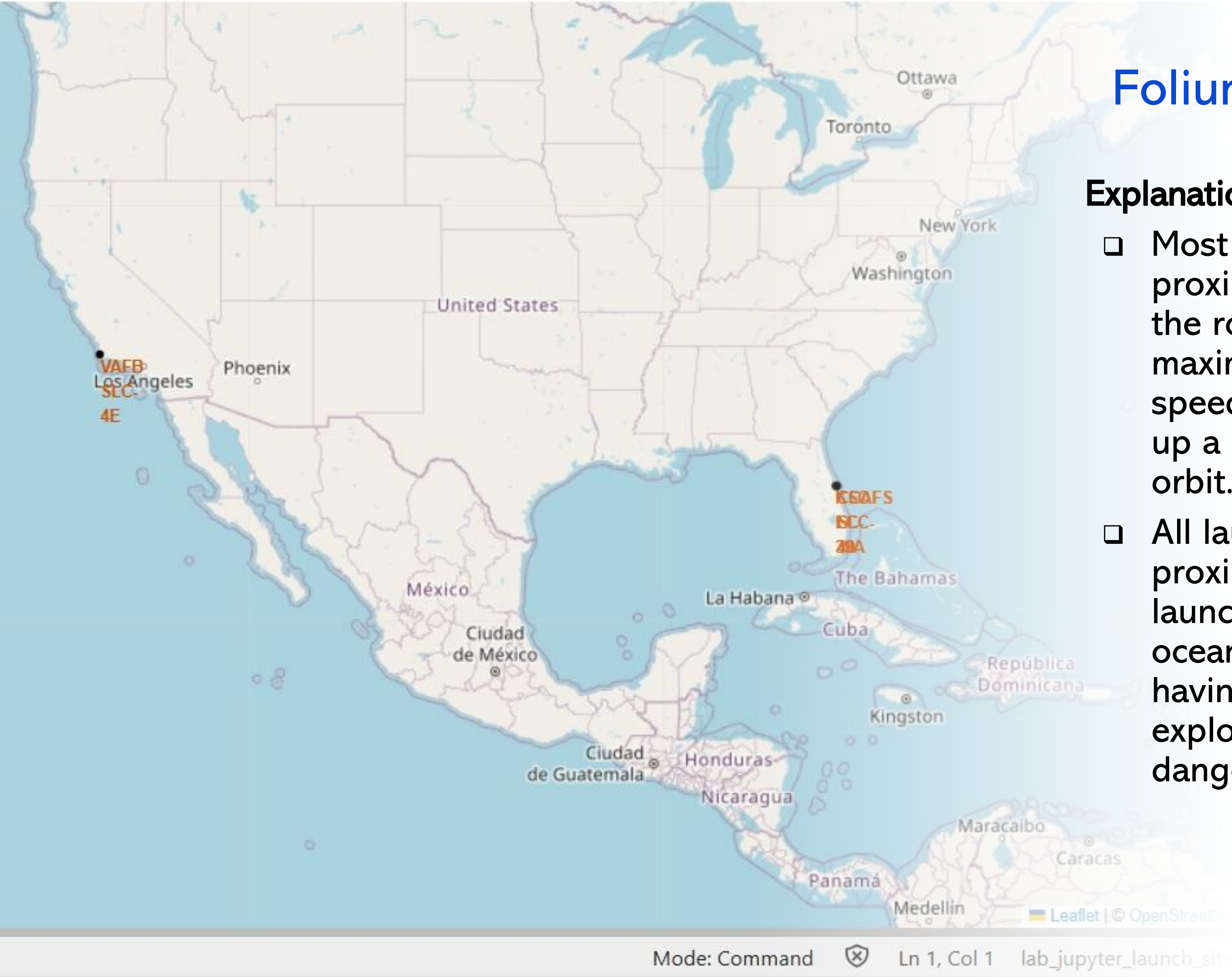
```
%%sql
SELECT
   "Landing_Outcome",
   COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Outcome_Count DESC
```

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

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

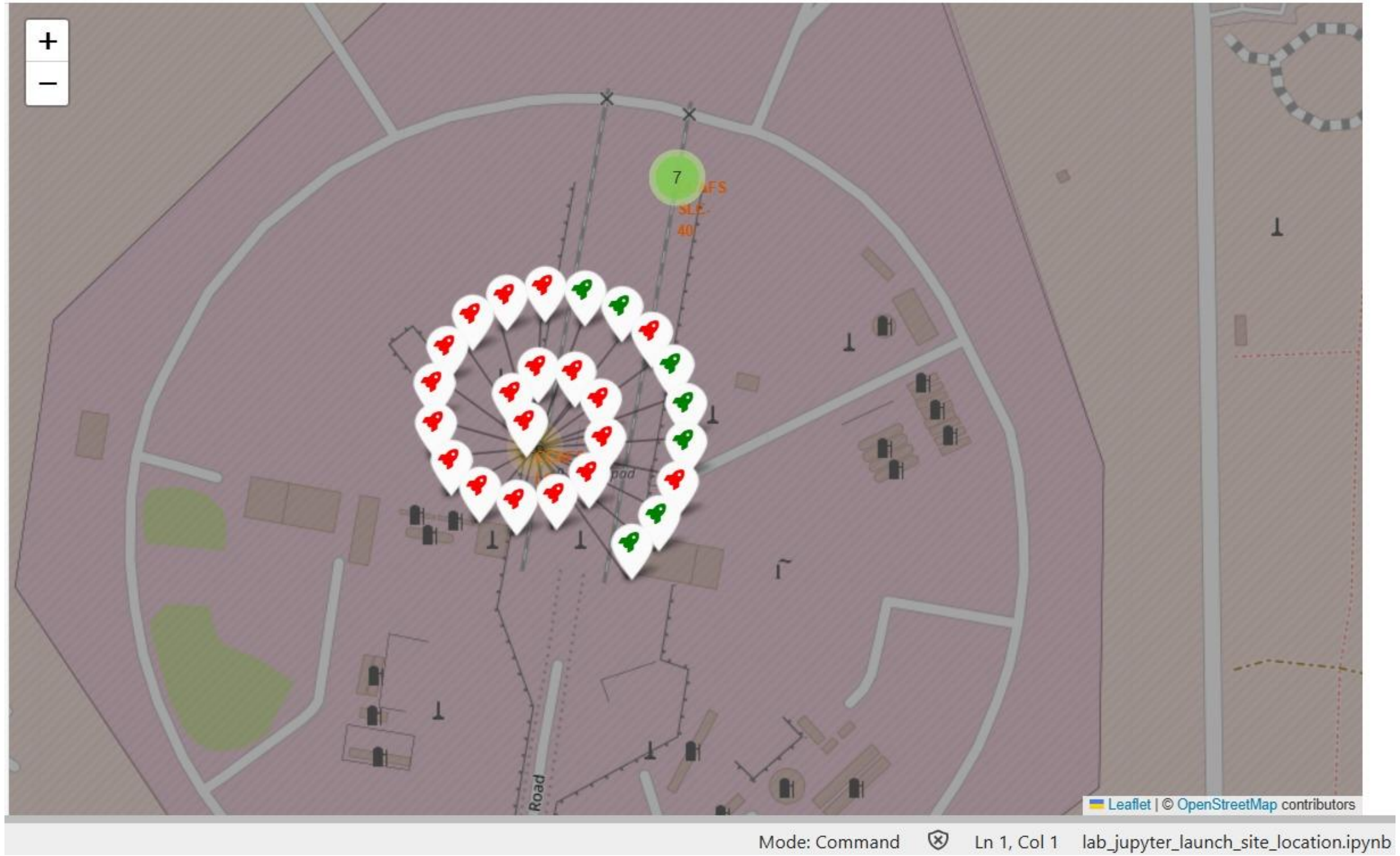


Folium Map - Screenshot 1

Explanation:

- Most of Launch sites are in proximity to the Equator line since, the rotational speed of the Earth is maximum at the equator. The high speed helps the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding on land, thereby dangering lives.

Folium Map - Screenshot 2



Explanation:

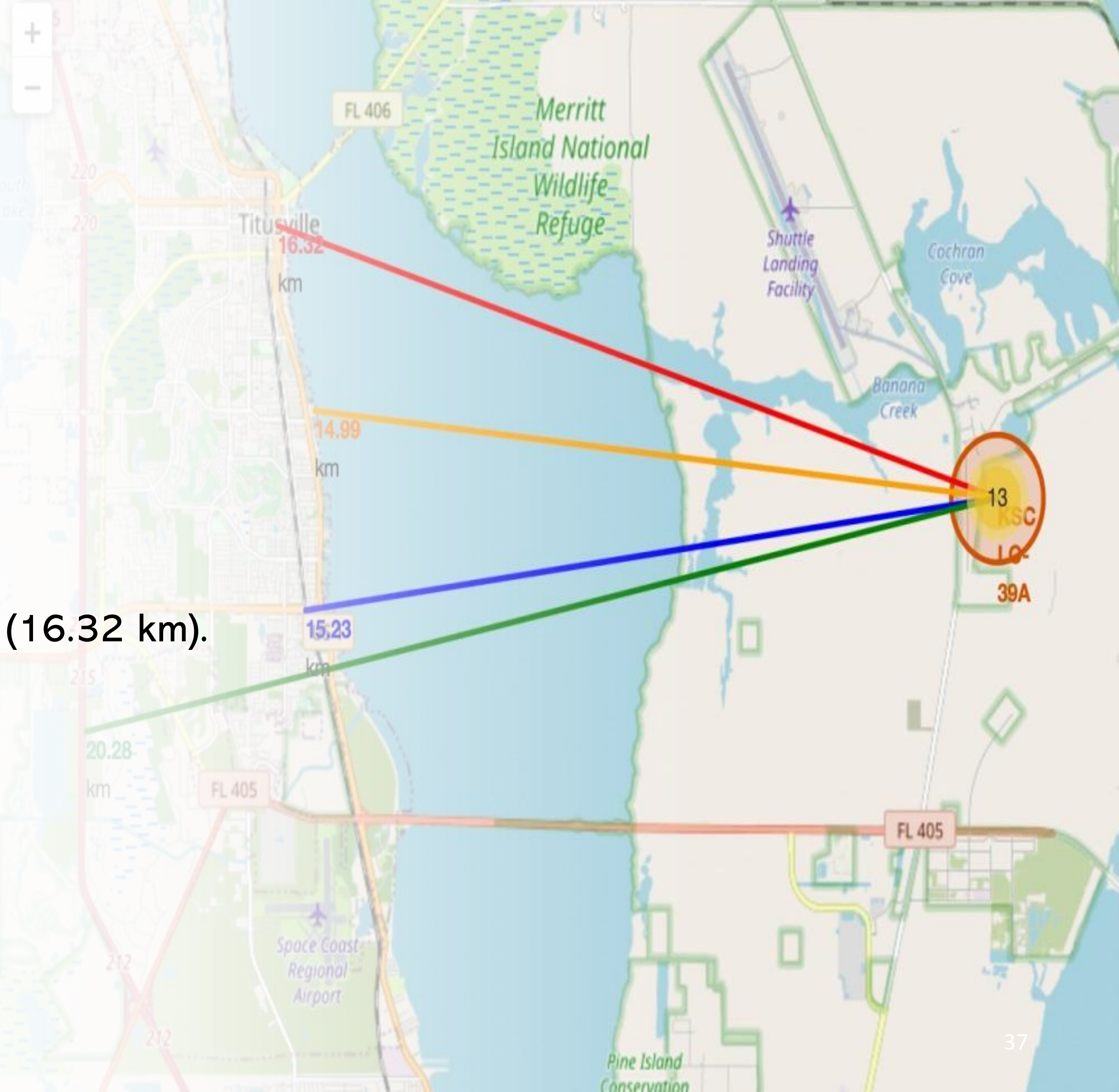
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

Folium Map - Screenshot 3

Explanation:

Launch site KSC LC-39A:

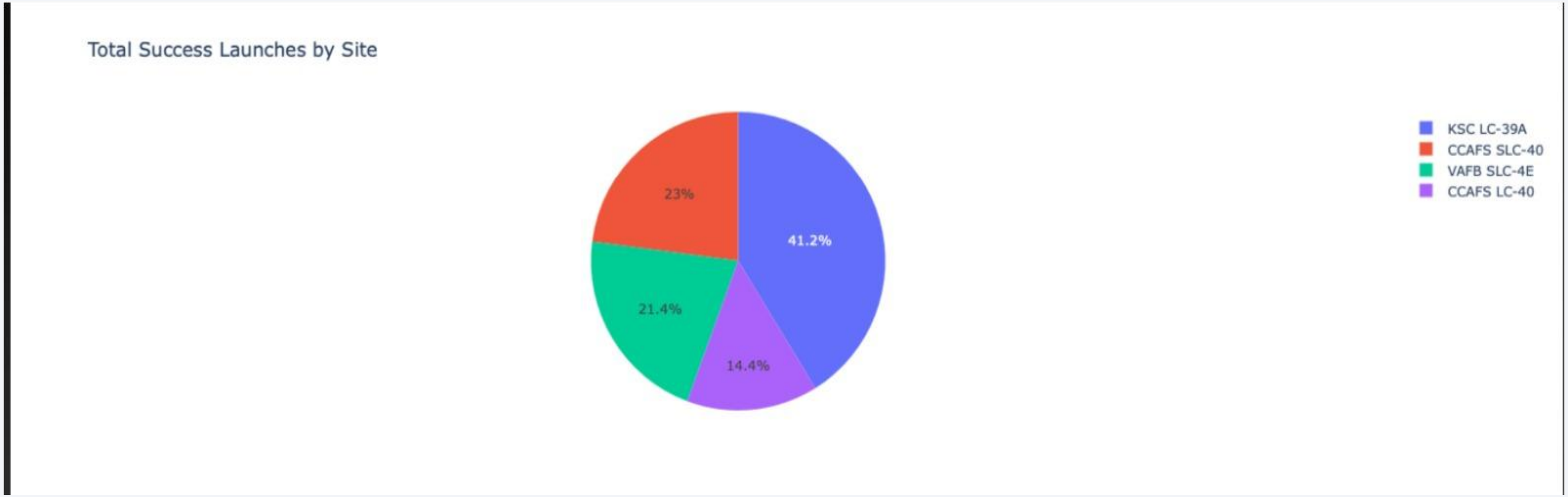
- Relatively close to railway (15.23 km)
- Relatively close to highway (20.28 km)
- Relatively close to coastline (14.99 km)
- Relatively close to closest city Titusville (16.32 km).



Section 4

Build a Dashboard with Plotly Dash

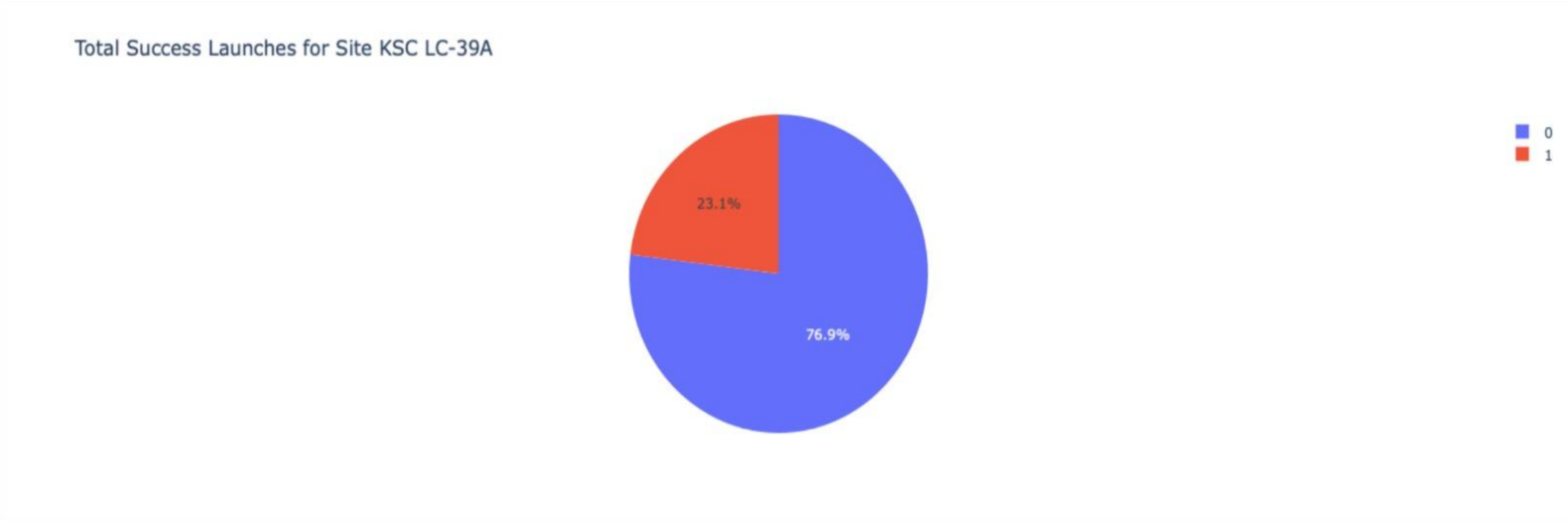
Dashboard - Screenshot 1



Explanation:

- Launch site KSC LC-39A has maximum success rate of about 41.2%.

Dashboard - Screenshot 2



Explanation:

- ❑ Launch site KSC LC-39A has success rate of 76.9% with 10 successful and only 3 failed landings.

Dashboard - Screenshot 3

Explanation:

- ❑ The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Find the method performs best:

```
[32]
print("Logistic Regression Accuracy:", logreg_cv.score(X_test, Y_test))
print("SVM Accuracy:", svm_cv.score(X_test, Y_test))
print("Decision Tree Accuracy:", tree_cv.score(X_test, Y_test))
print("KNN Accuracy:", knn_cv.score(X_test, Y_test))
```

```
Logistic Regression Accuracy: 0.8333333333333334
SVM Accuracy: 0.8333333333333334
Decision Tree Accuracy: 0.8888888888888888
KNN Accuracy: 0.6111111111111112
```

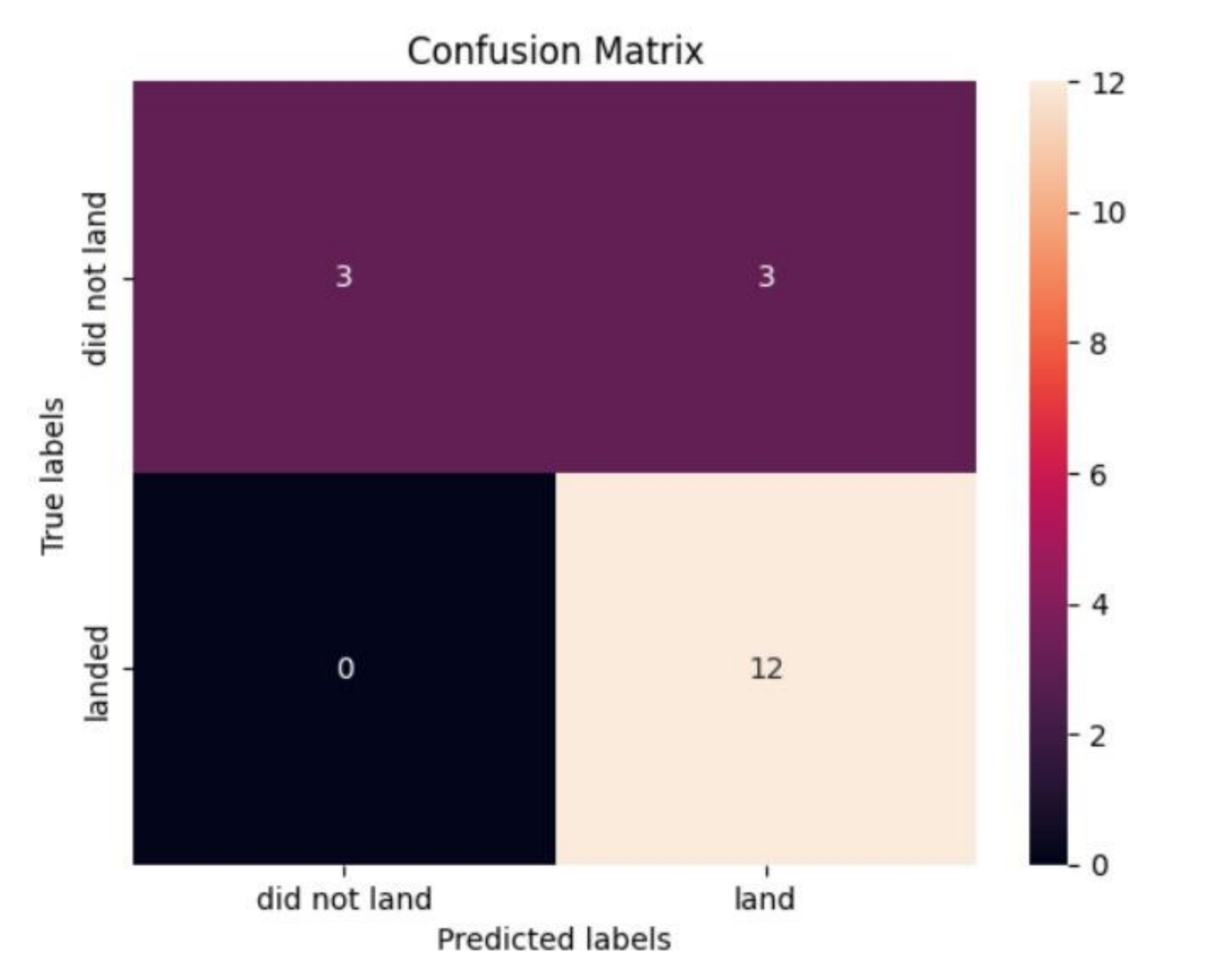
Explanation:

- Decision Tree Model resulted in best accuracy rate of 88.9%
- Other models like Logistic Regression and SVM also had higher accuracy rate of about 83%.

Confusion Matrix

Explanation:

- Examining the confusion matrix of Decision Tree model, we observe that it can distinguish between the different classes.
- We see that the major problem is false positives.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass showed better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all are in very close proximity to the coast.
- The success rate of launches increased over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!