IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Karnaz Obaidullah
November 18th 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

Goal of Capstone: Predict if the Falcon 9 first stage will land successfully

Why care about this problem?

- According to SpaceX, Falcon 9 rocket launches cost 62 million dollars, while other providers cost >165 million dollars.
- SpaceX saves millions by reusing the first stage.
- Therefore, if we can predict if the first stage will land we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and Webscraping Wikipedia
- 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 models

Data Collection

- Describe how data sets were collected.
  - Data was collected using SpaceX API and Webscraping Wikipedia
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/karnazko27/IBMCapstone/blob/main/1%20Data%20Collection%20Using%20API.ipynb

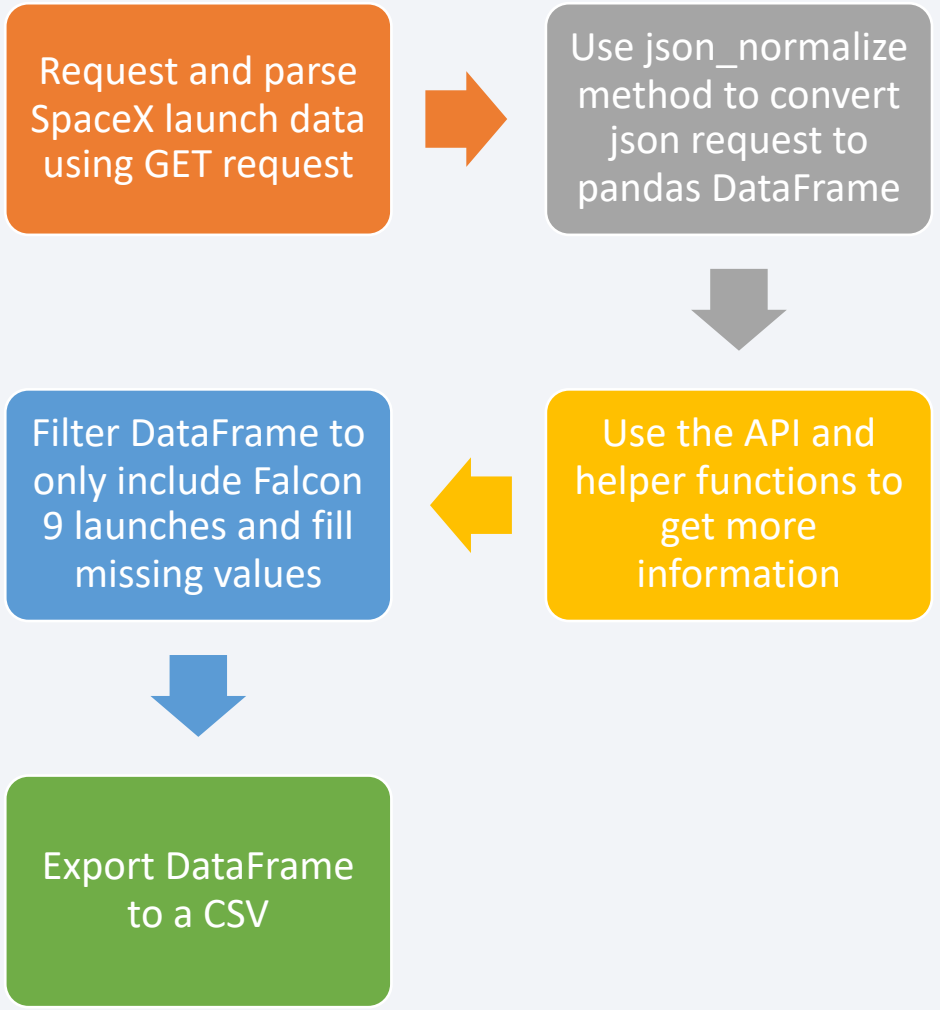


Fig 1: flowchart of SpaceX API calls

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- https://github.com/karnazko27/IBMCapstone/blob/main/2%20Data%20Collection%20with%20Web%20Scraping.ipynb

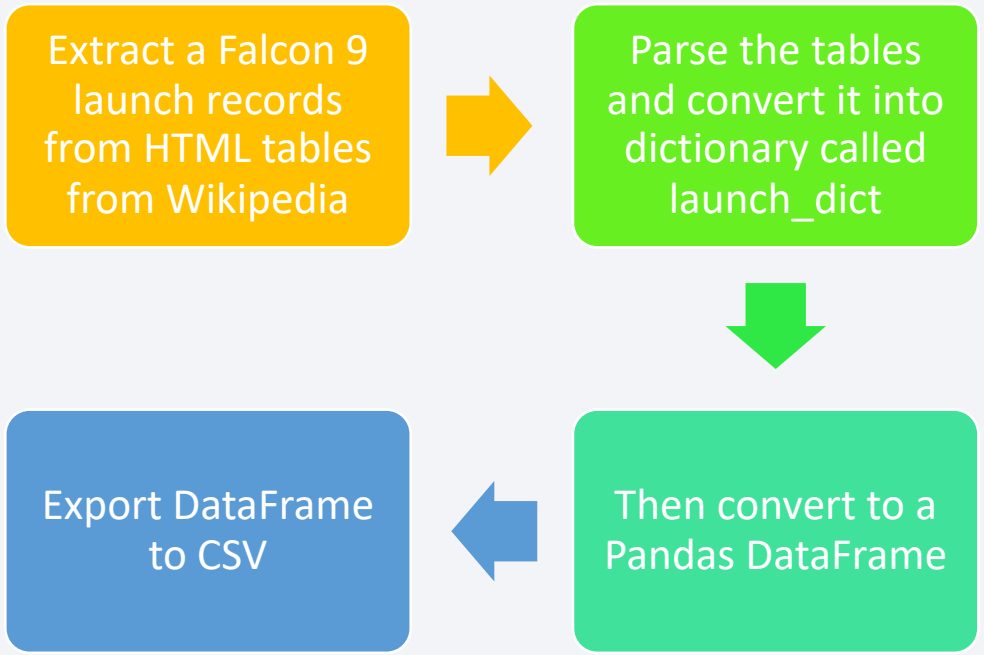


Fig 2: flowchart of web scraping

Data Wrangling

- Describe how data were processed
  - Calculate number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbits
  - Create a landing outcome label from Outcome column
- You need to present your data wrangling process using key phrases and flowcharts
- https://github.com/karnazko27/IBMCapstone/blob/main/3%20Data%20Wrangling.ipynb

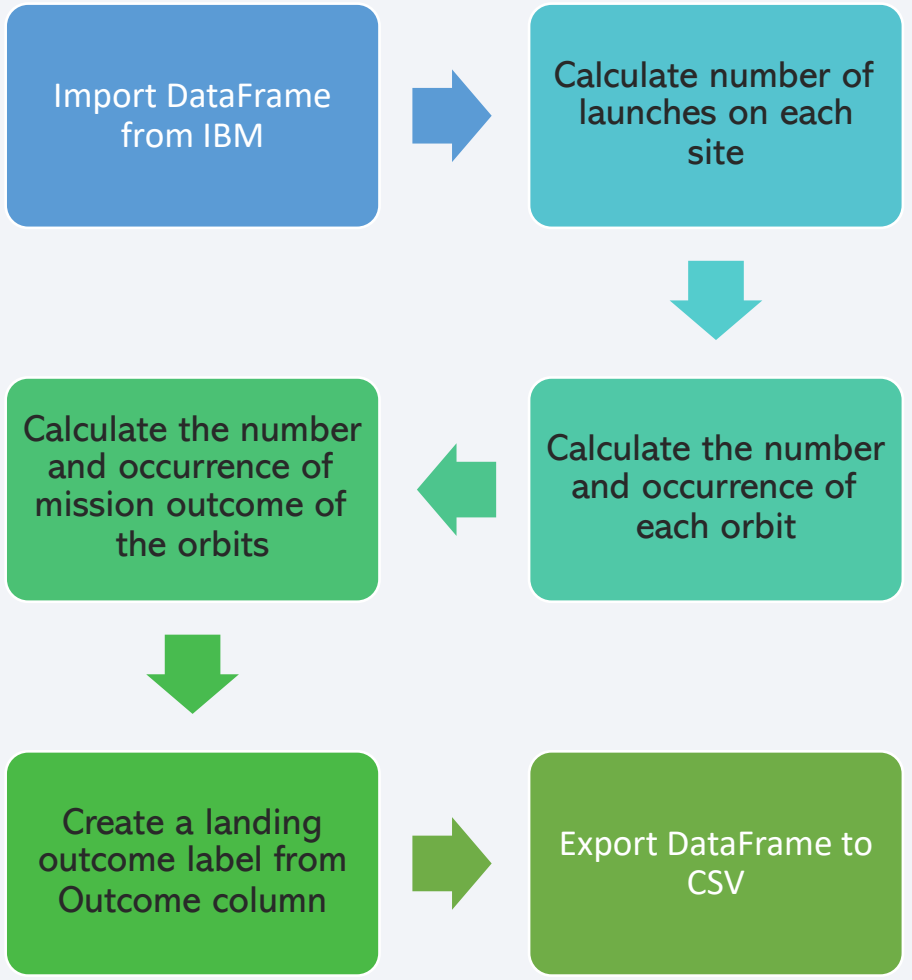


Fig 3: flowchart of data wrangling

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- https://github.com/karnazko27/IBMCapstone/blob/main/5%20Visualization.ipynb

EDA with SQL

- SQL queries performed:
  - 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 succesful landing outcome in ground pad was acheived.
  - 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. Use a subquery

EDA with SQL

- SQL queries performed:
  - 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.
  - 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.
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
- https://github.com/karnazko27/IBMCapstone/blob/main/4%20SQL_eda.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- https://github.com/karnazko27/IBMCapstone/blob/main/6%20folium_notebook.ipynb

Build a Dashboard with Plotly Dash

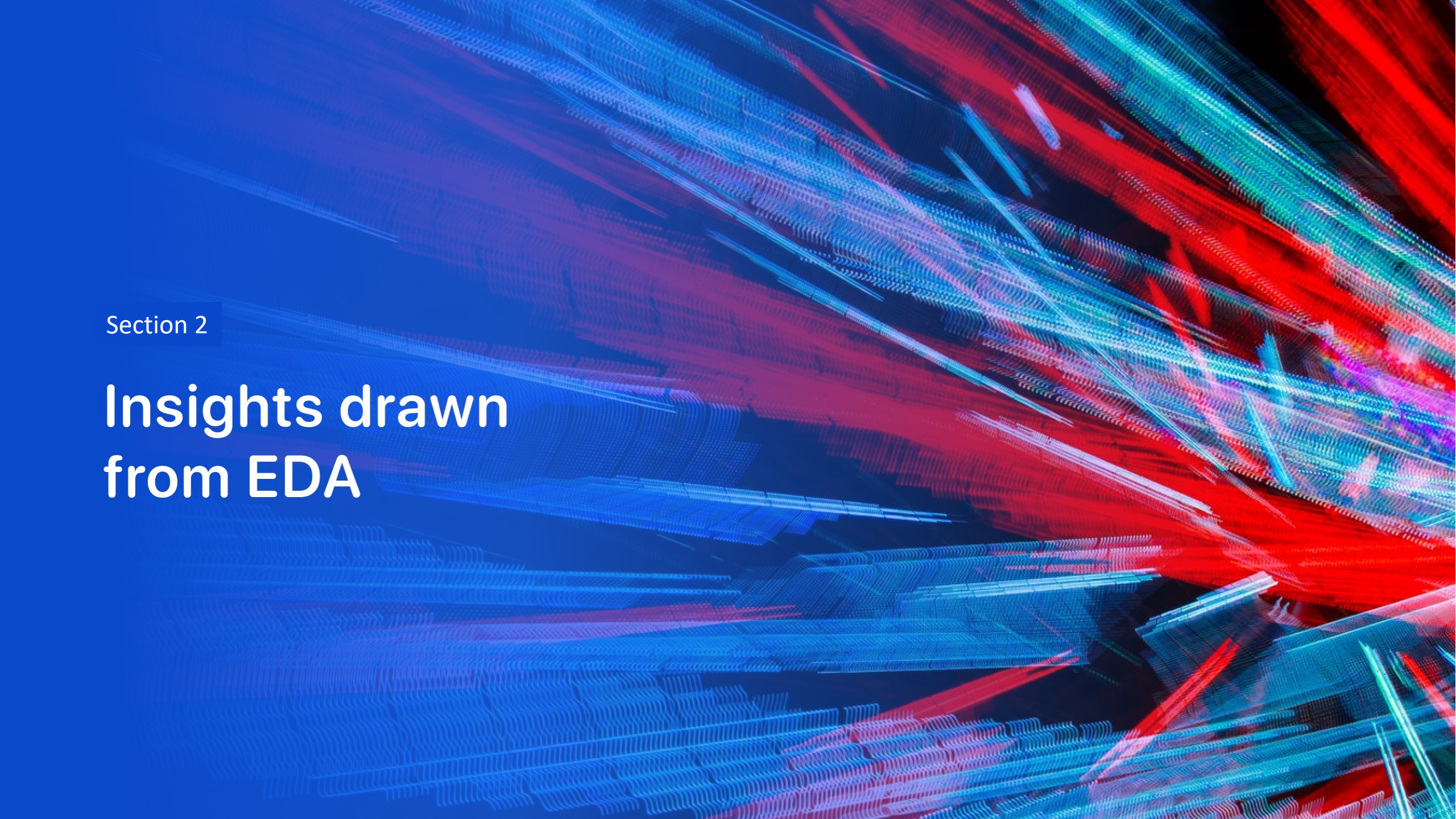
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- https://github.com/karnazko27/IBMCapstone/blob/main/7%20InteractiveDashboard.ipynb

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://github.com/karnazko27/IBMCapstone/blob/main/8%20machine_learning.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

- From the visualization we can see that launch success is higher the greater the number of flights.

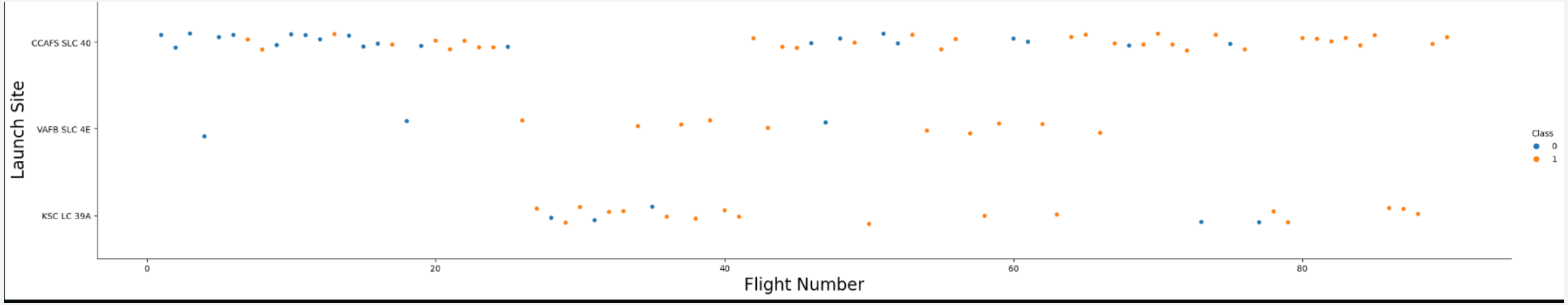


Fig 4: scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site

- There are less launches made the higher the payload mass.

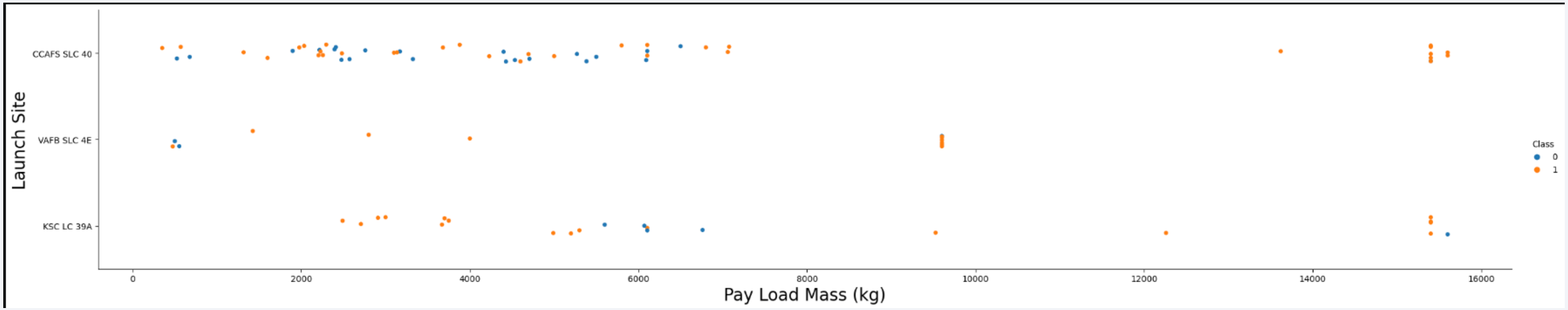


Fig 5: scatter plot of Payload vs. Launch Site

Success Rate vs. Orbit Type

- The highest success rates were for orbit types: ES-L1, GEO, HEO, and SSO.

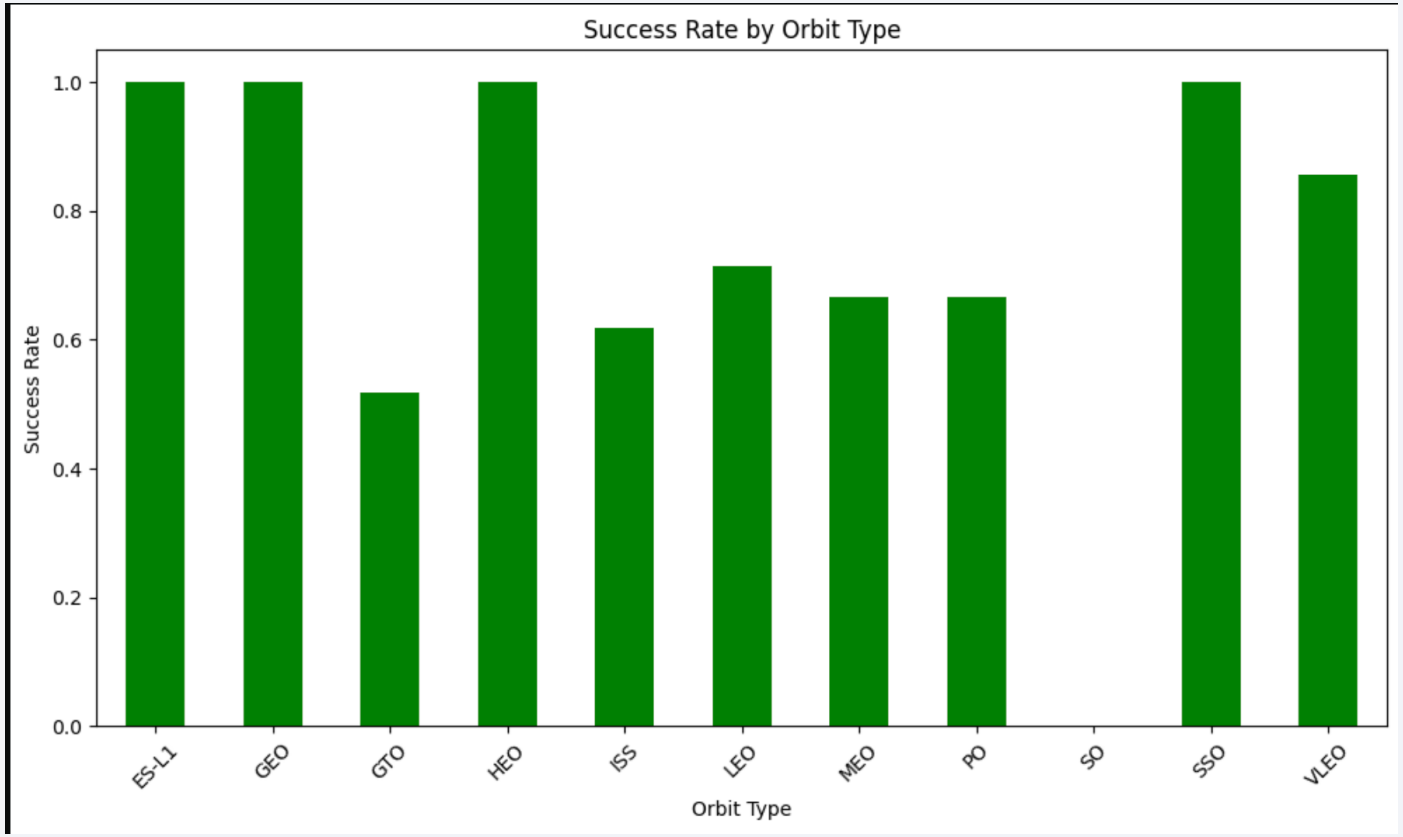


Fig 6: bar chart for the success rate of each orbit type

Flight Number vs. Orbit Type

- The GTO, ISS, PO, VLEO, SSO and LEO Orbits had the highest proportionate number of successes.

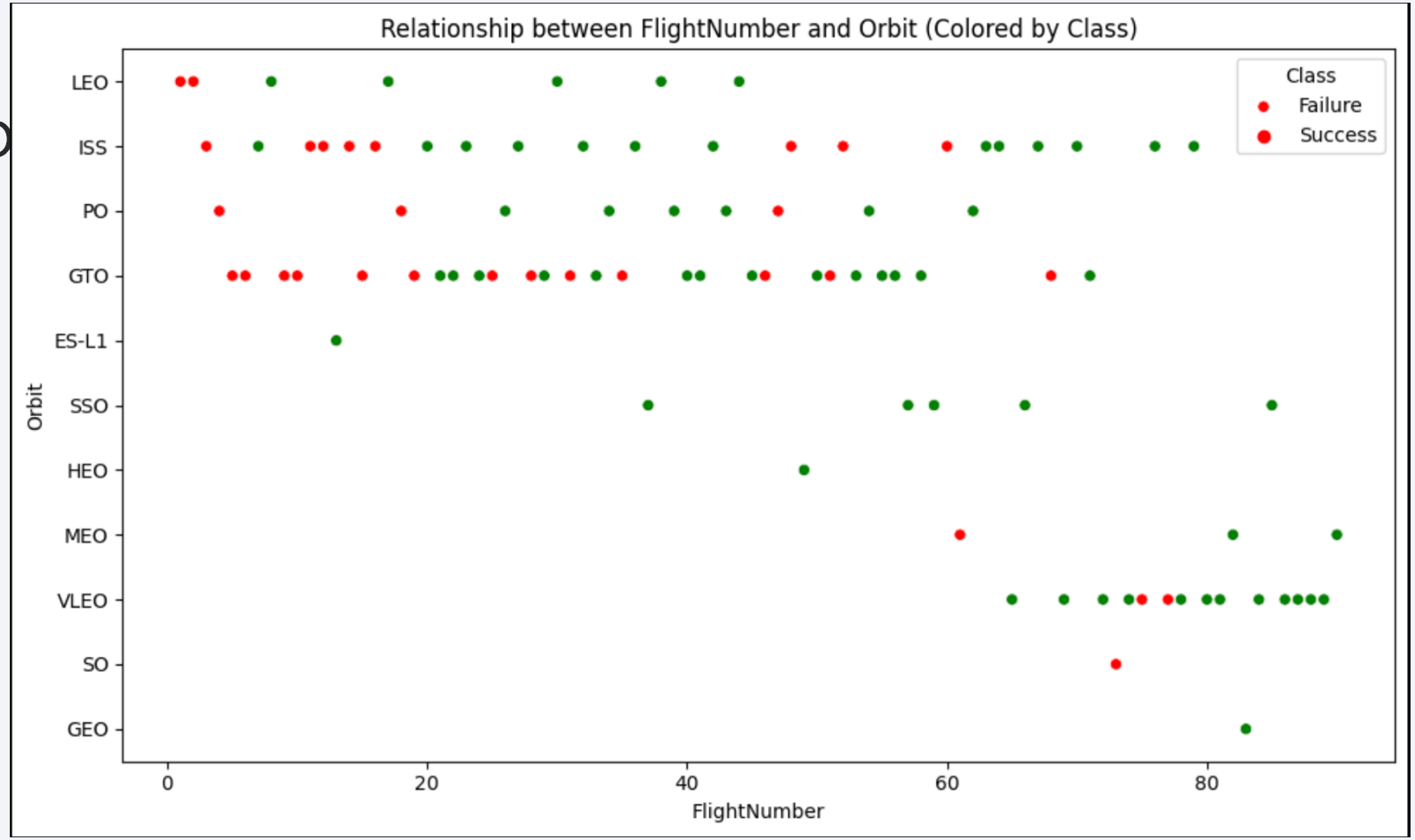


Fig 7: a scatter point of Flight number vs. Orbit type

Payload vs. Orbit Type

- With heavier payloads the successful landing or positive landing rate are more for PO, LEO and ISS.

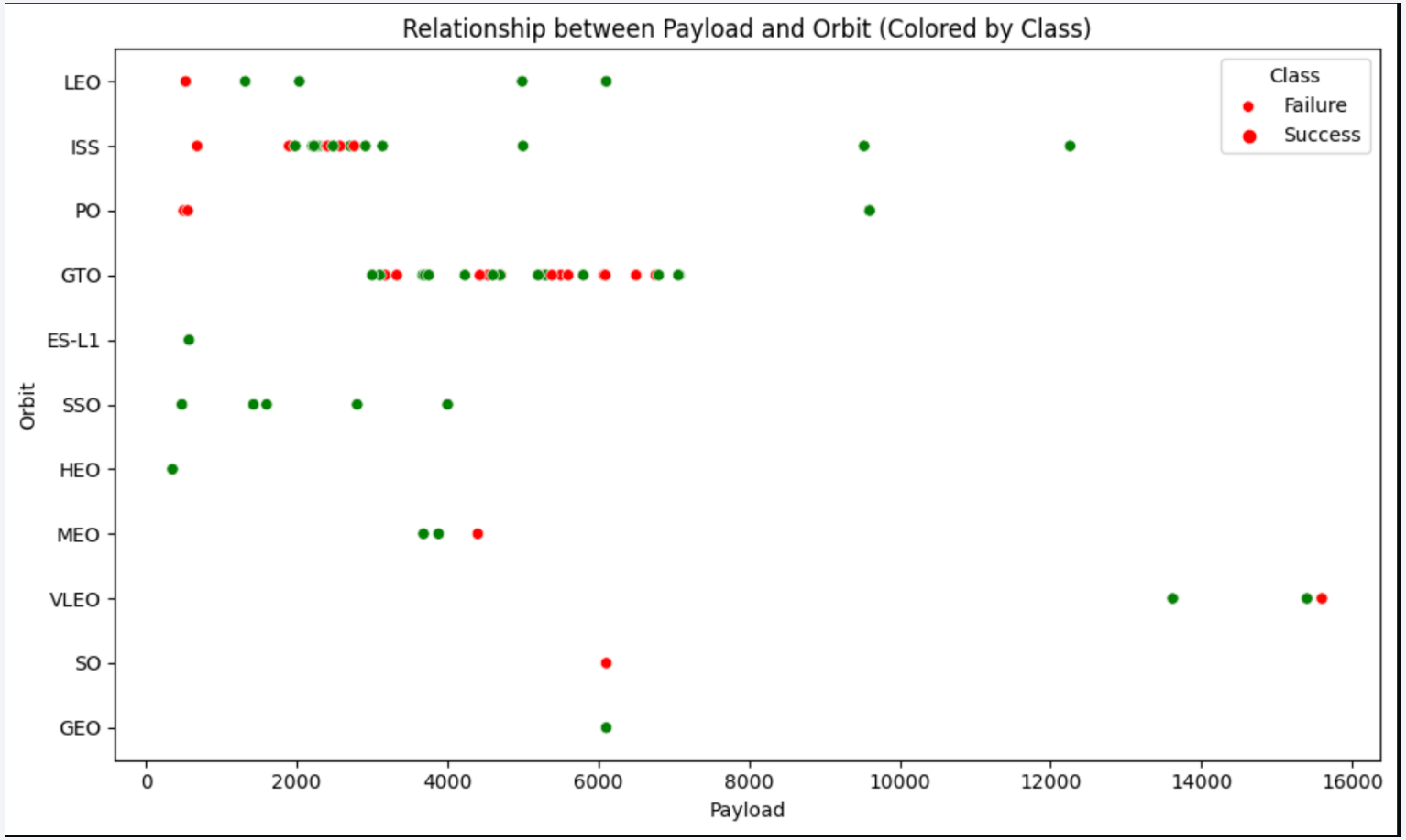


Fig 8: a scatter point of payload vs. orbit type

Launch Success Yearly Trend

- Success rate is higher with more recent launches than older ones.

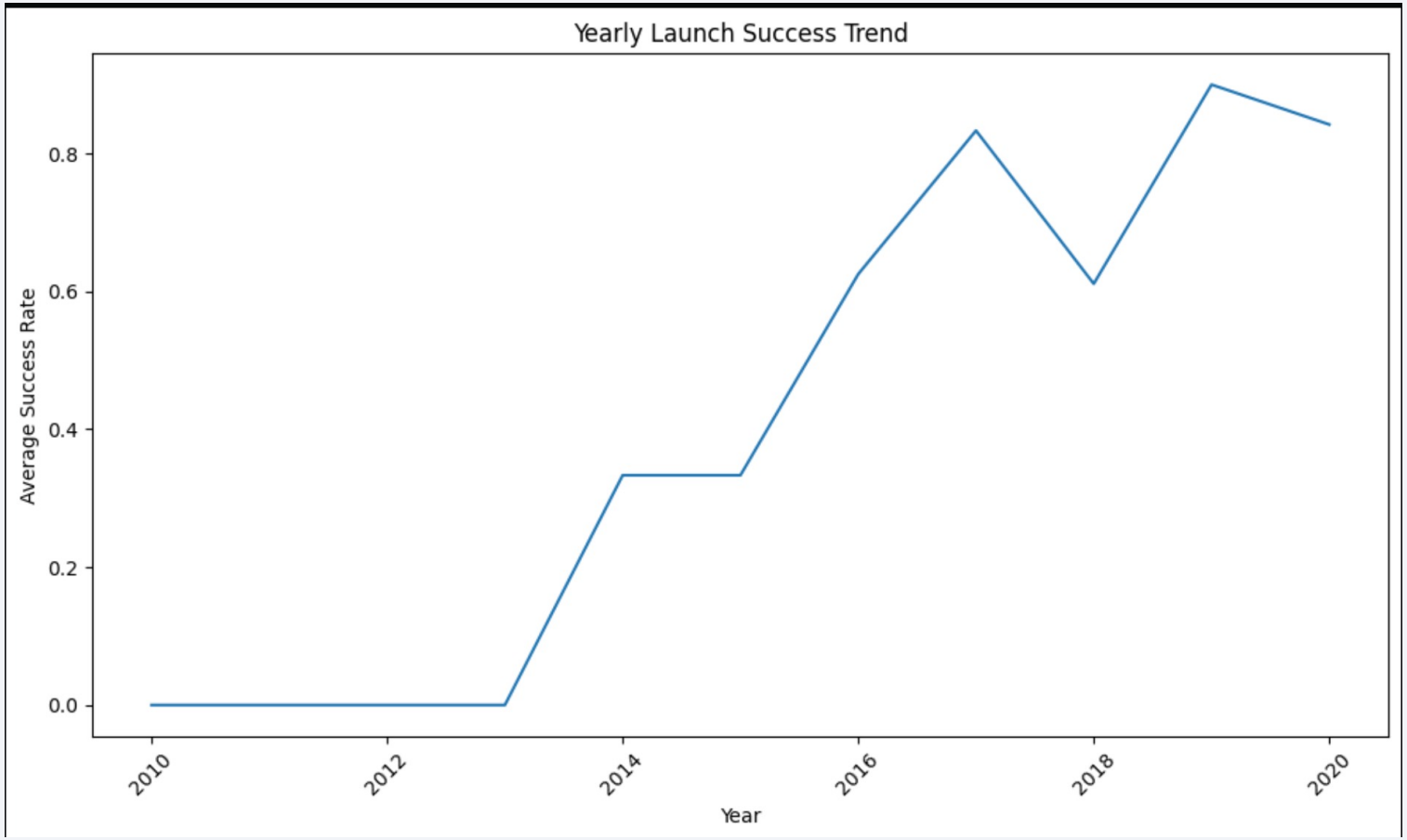


Fig 9: line chart of yearly average success rate

All Launch Site Names

- Find the names of the unique launch sites
- Use DISTINCT to show unique launch sites

```
[10]: %%sql
      SELECT DISTINCT("Launch_Site")
      FROM SPACEXTABLE
      LIMIT 5

       * sqlite:///my_data1.db
      Done.
[10]:   Launch_Site
      CCAFS LC-40
       VAFB SLC-4E
        KSC LC-39A
      CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- LIMIT 5 to show the first 5 records

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

[11]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5
```

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

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Use SUM() function to calculate total payload mass

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

       * sqlite:///my_data1.db
      Done.
[13]: SUM(PAYLOAD_MASS__KG_)
      45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Use AVG function for calculation

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Use MIN() function to find first date

```
[41]: %%sql
      SELECT MIN(Date)
      FROM SPACEXTABLE
      WHERE Landing_Outcome = 'Success (ground pad)'

       * sqlite:///my_data1.db
      Done.
[41]: MIN(Date)
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Use WHERE and <> to find the range of values

In [42]:

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ > 4000
  AND PAYLOAD_MASS__KG_ < 6000
```

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

Out[42]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Use GROUPBY and COUNT

In [43]:
```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY Mission_Outcome
```

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

Out[43]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Use SELECT and MAX() for this query

In [45]:

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

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

Out[45]:

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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- No records found for this query

In [47]:

```
%%sql
SELECT
    CASE substr(Date, 4, 2)
        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'
        ELSE 'Unknown'
    END AS Month,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTABLE
WHERE substr(Date, 7, 4) = '2015'
    AND Landing_Outcome LIKE 'Failure (drone ship)%'
```

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

Out[47]:

Month	Booster_Version	Launch_Site	Landing_Outcome

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
- Use GROUP BY, ORDER BY and WHERE and COUNT() functions for this query

In [49]:

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

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

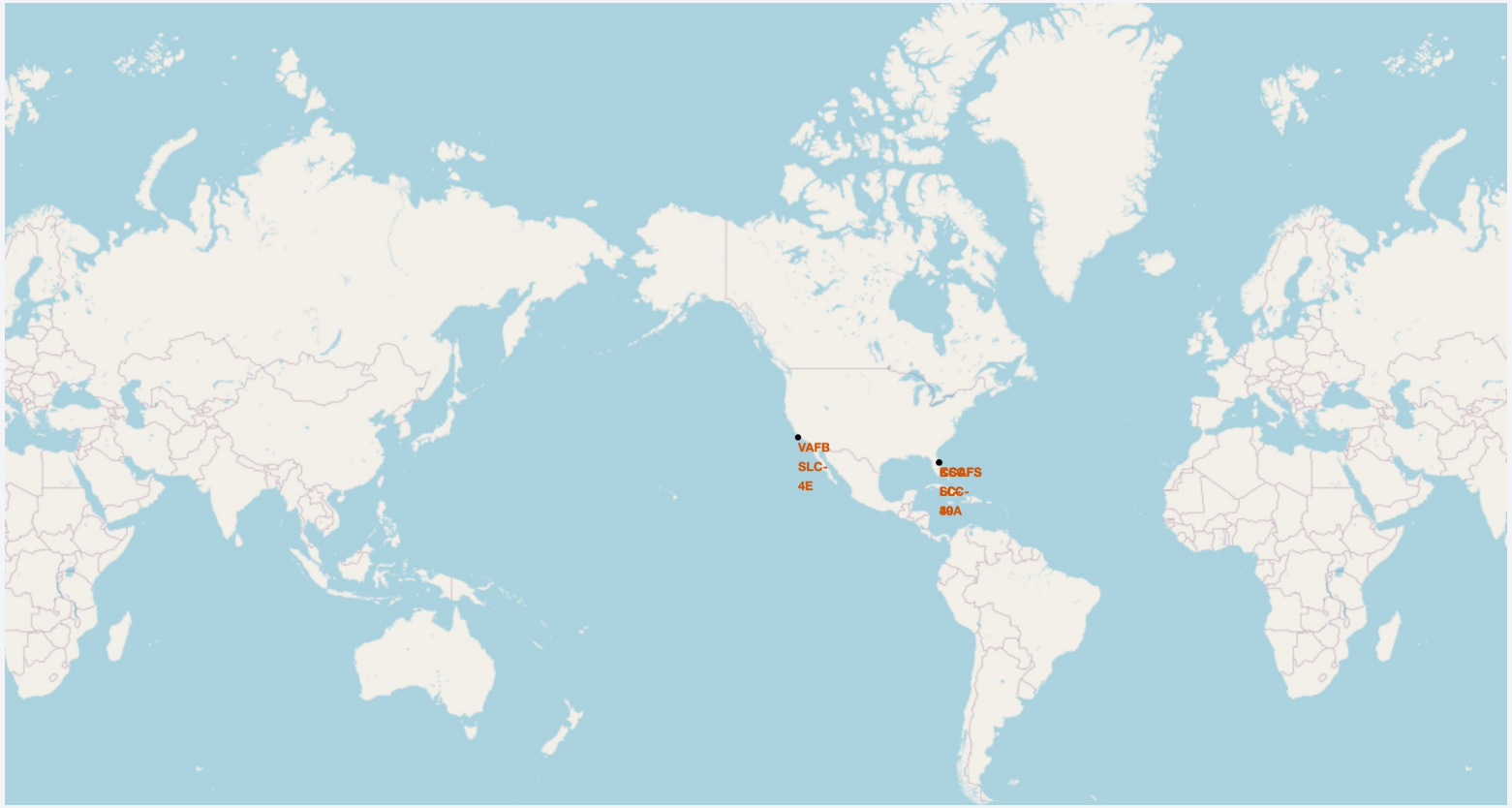
Out[49]:

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


Section 3
Launch Sites
Proximities Analysis

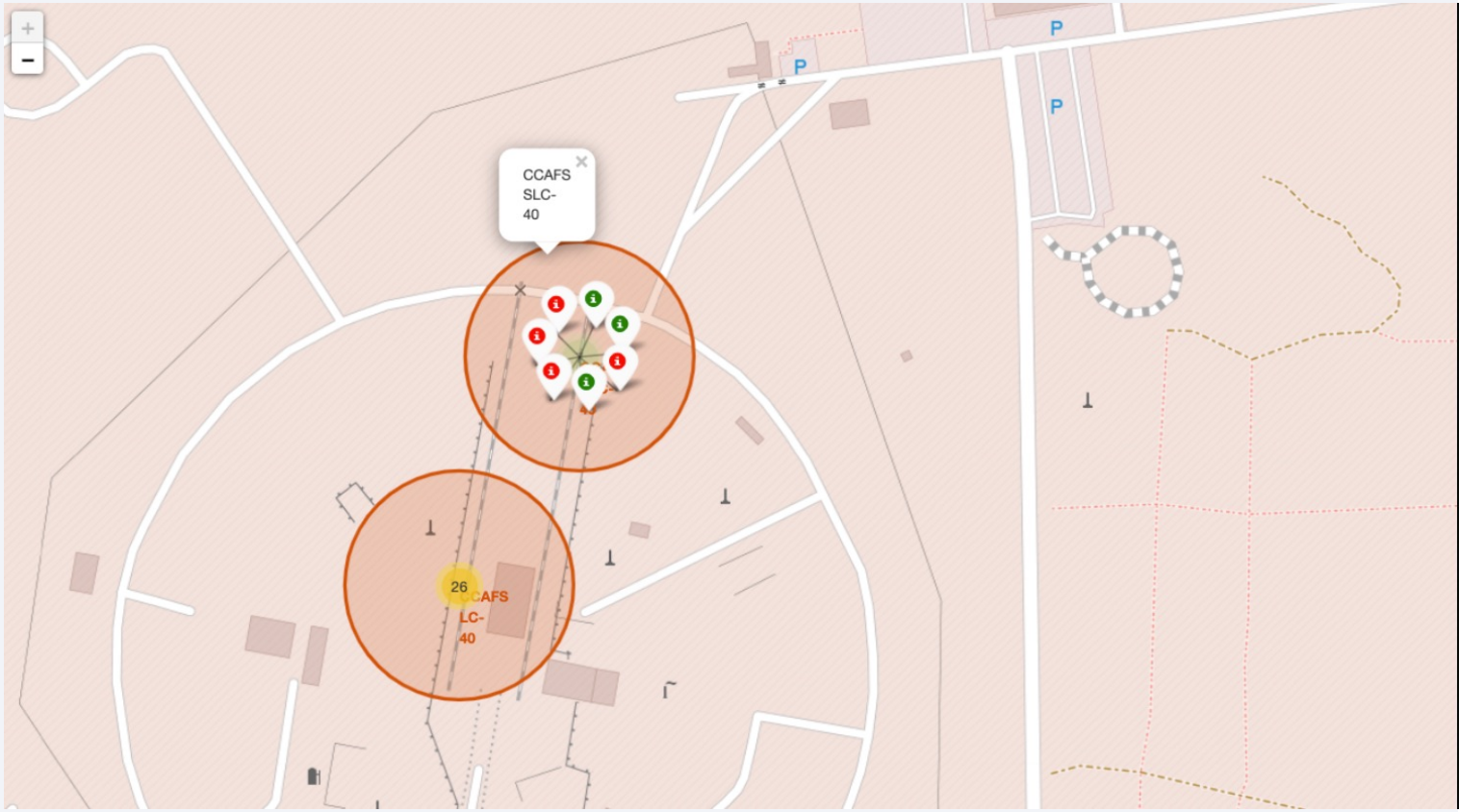
Global Map with SpaceX Launch Sites

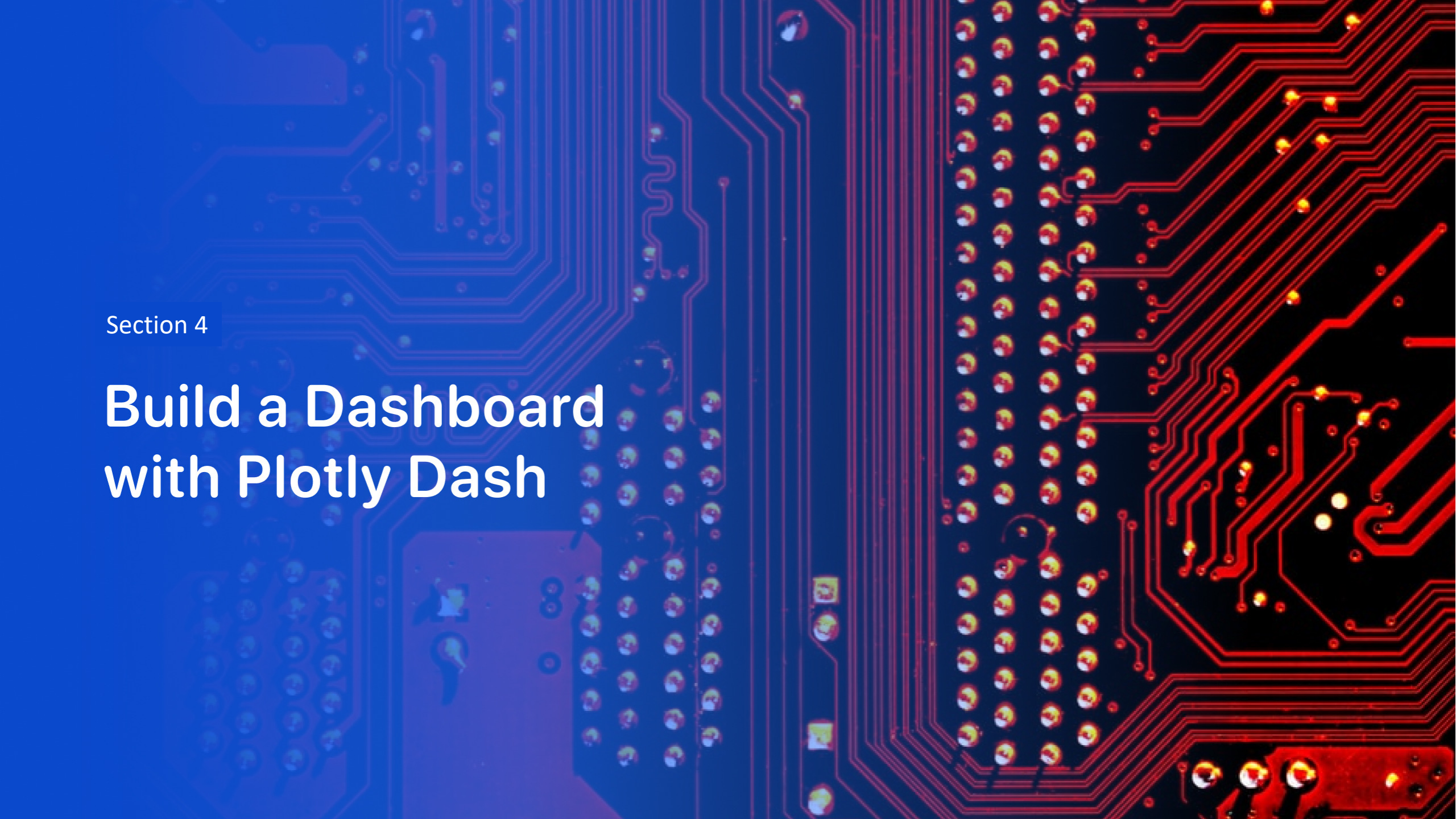
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- SpaceX Launch Sites are in California and Florida



Markers Showing Launch sites

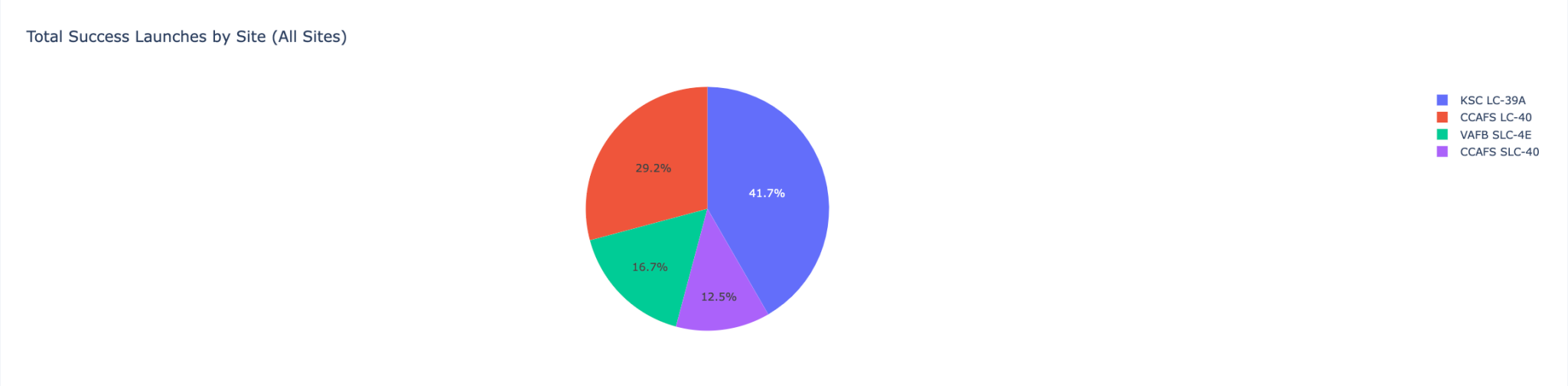
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- The launch sites are all in one area.




Section 4
Build a Dashboard with Plotly Dash

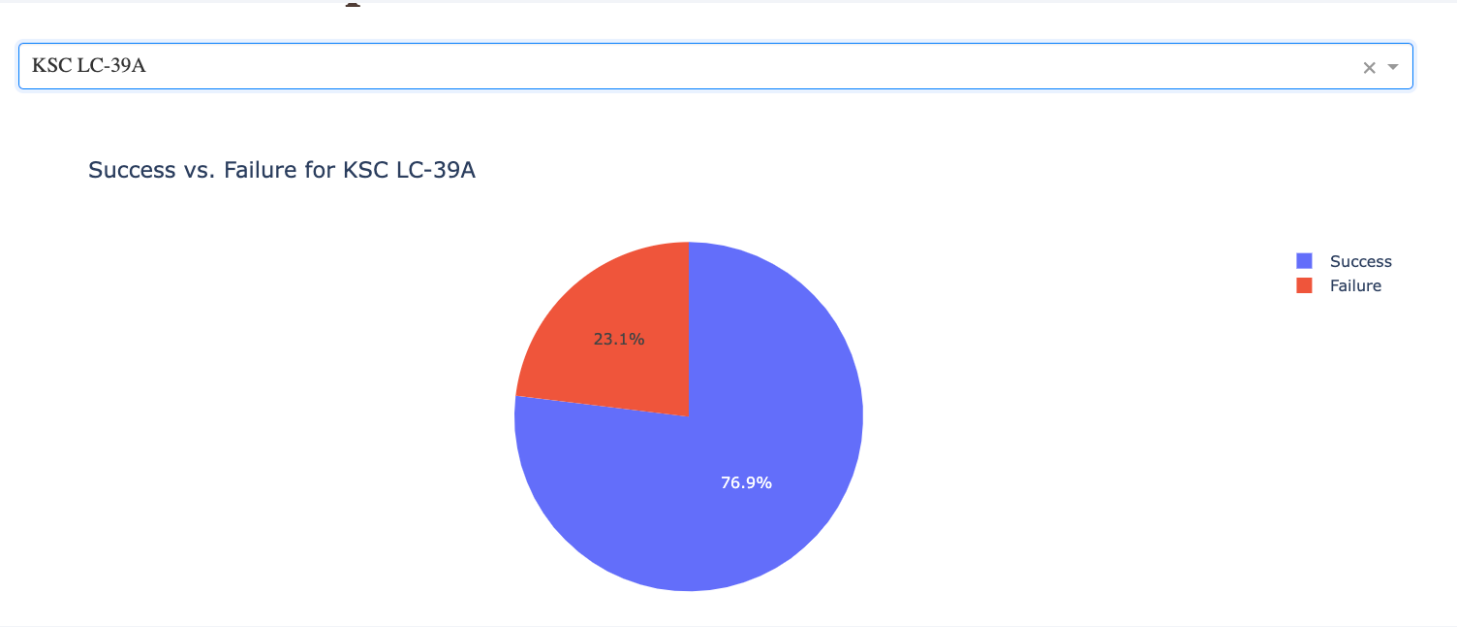
Pie chart of Launch Success For All Sites

- screenshot of launch success count for all sites, in a piechart
- KSC LC-39A has 41.7% launch success which is the highest for all sites.



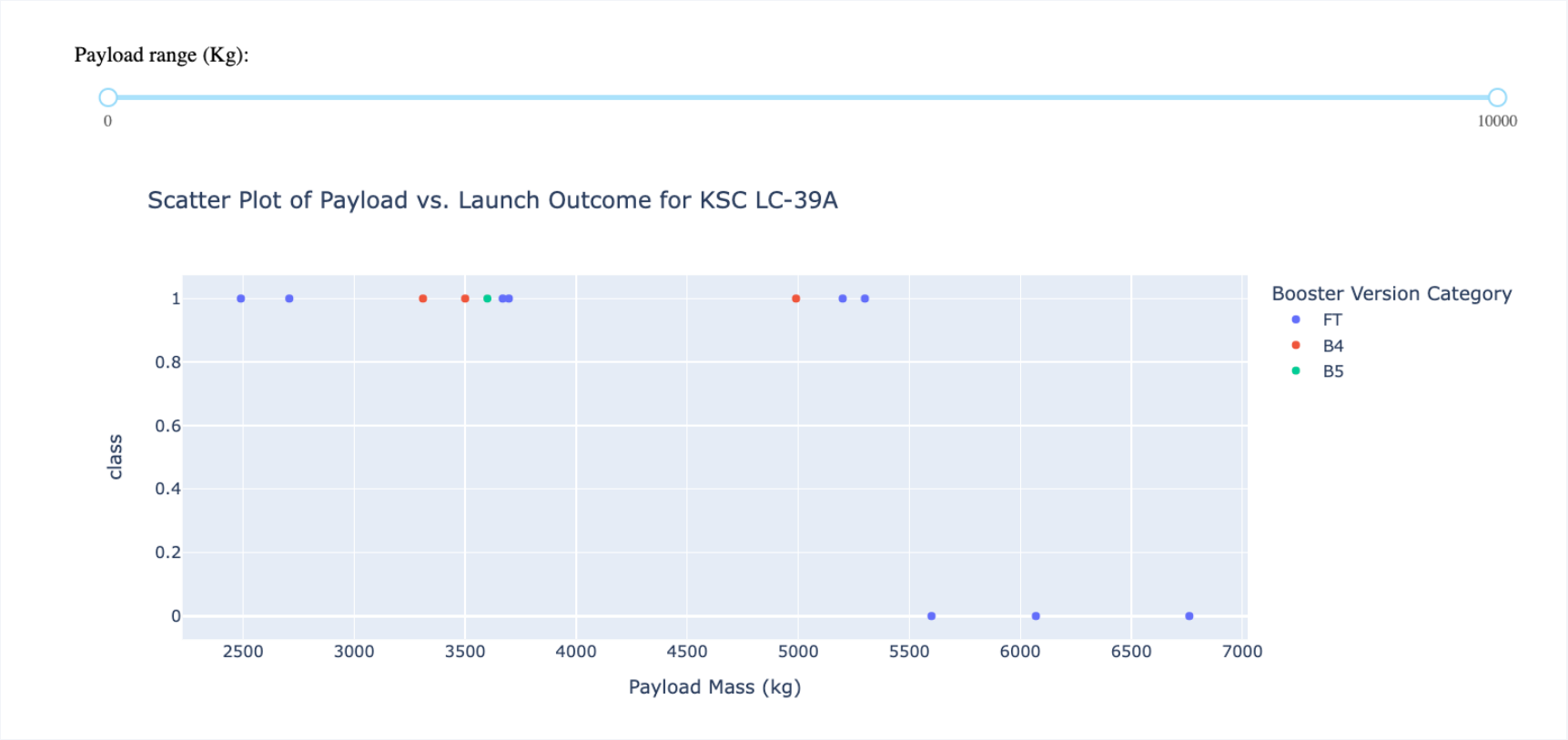
Launch Site with Highest Success Ratio

- Highest success ratio is for KSC LC-39A



Payload vs. Launch Outcome scatter plot for all sites

- The higher the payload the worse the outcome



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Logistic Regression model has the highest accuracy

```
logreg_cv = GridSearchCV(lr, parameters, cv=10)

# Fit the GridSearchCV object to find the best parameters
logreg_cv.fit(X_train, Y_train)

# Get the best parameters and best estimator
best_params = logreg_cv.best_params_
best_lr_model = logreg_cv.best_estimator_
```

We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params_` and the accuracy on the validation data using the data attribute `best_score_`.

[14]:
```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

TASK 5 ¶

Calculate the accuracy on the test data using the method `score`:

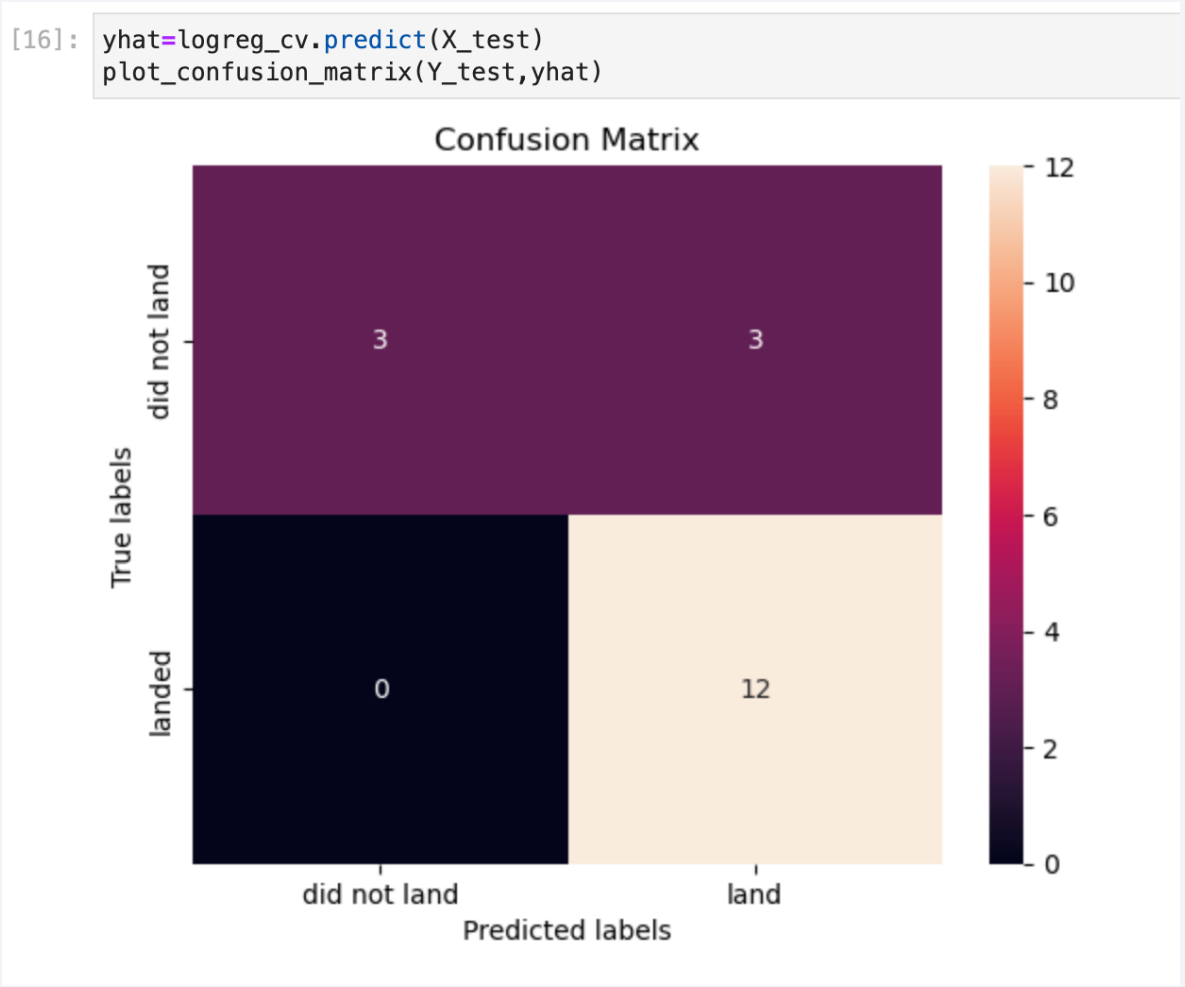
[15]:
```
test_accuracy = logreg_cv.score(X_test, Y_test)
print(test_accuracy)
```

```
0.8333333333333334
```

Confusion Matrix

- confusion matrix of the best performing model with an explanation

```
[16]: yhat=logreg_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The higher number of flights at a launch site had more successes.
- Payload mass being higher had worse outcomes.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the highest success rate.
- KSC LC-39A had the most successful launches.
- The Logistic Regression model has the highest training and testing accuracy overall.

Thank you!