IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Augustine Chukwumezie
November 30, 2023

Outline

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

Executive Summary

Summary of methodologies

- Different data collection approaches was employed such web scraping and using SpaceX API. The data was cleaned and transformed into a suitable format for further analysis. Then Exploratory Data Analysis was performed to understand the data properly and visualizations was also carried out before performing a predictive analysis using classification models.

Summary of all results

- Based on the obtained results, every model utilized in this project successfully predicted the outcome of the first stage rocket booster landing with an approximate 83% accuracy on the test data. Consequently, this outcome will enable SpaceY to make informed, data-driven decisions regarding their pursuit of an alternative to SpaceX and their exploration of whether SpaceX will proceed with reusing the first stage.

Introduction

Project background and context

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers like Virgin Galactic, Blue Origin, Rocket Lab, etc cost upward of 165 million dollars each and much of the savings is because SpaceX 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 SpaceX for a rocket launch.

Problems I want to find answers

SpaceY is a new commercial rocket launch provider and wants to compete with SpaceX in finding a cheaper alternative to travel to space and return and I was consulted to analyze the data from SpaceX and build a predictive system that can predict if the Falcon 9 first stage will land successfully or not while understanding the correlation between each rocket variables and successful landing rate.

Section 1

Methodology

Methodology

Executive Summary

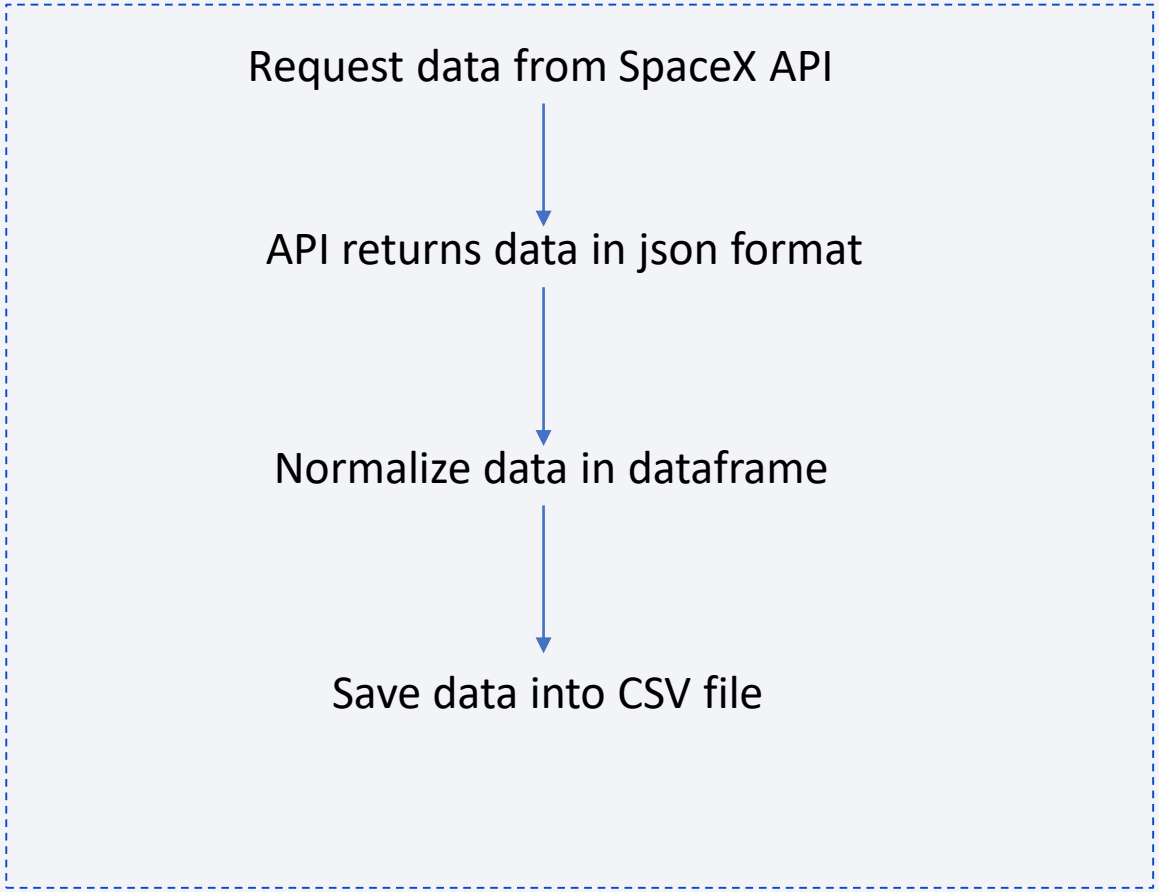
- Data collection methodology:
  - Using SpaceX API and web scraping of List of Falcon 9 and Falcon Heavy launches page
- Perform data wrangling
  - Data was cleaned and missing data was handled, and training labels was determined
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find the best hyperparameters for the classification models used(SVM, Decision Tree, Logistic Regression and KNN)

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

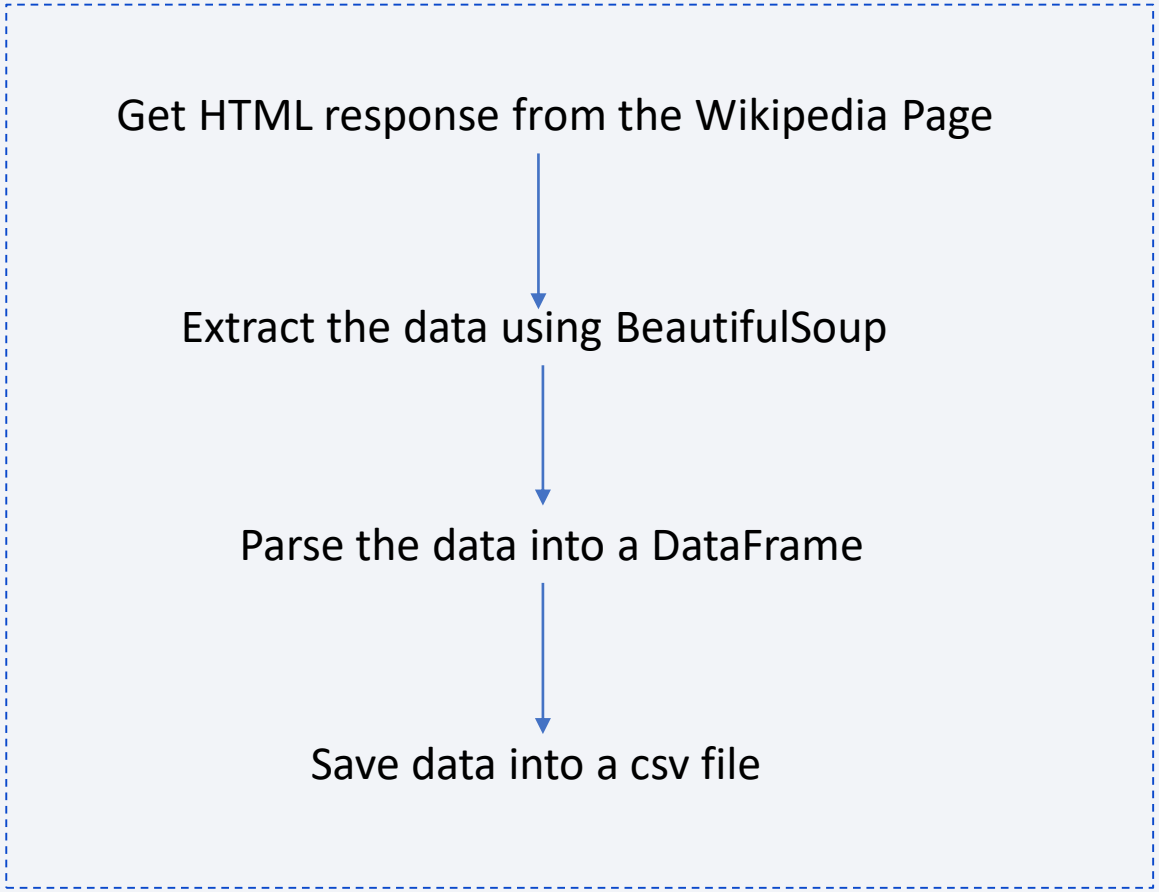
Data Collection – SpaceX API

- GitHub Link



Data Collection - Scraping

- GitHub Link



Data Wrangling

- Calculated the number of launches on each site
- Calculated the number and occurrence of each orbit
- Calculated the number and occurrence of mission outcome of the orbits
- Created a landing outcome label df[[“Class”]] from Outcome column
  - Class = 0; first stage booster did not land successfully
    - None None; not attempted
    - None ASDS; unable to be attempted due to launch failure
    - False ASDS; drone ship landing failed
    - False Ocean; ocean landing failed
    - False RTLS; ground pad landing failed
  - Class = 1; first stage booster landed successfully
    - True ASDS
    - True RTLS
    - True Ocean

- GitHub Link

EDA with Data Visualization

GitHub Link

- Scatter Chart: This shows the correlation between two variables as it shows how much one variable is affected by another variable
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Flight Number and Orbit type
  - Payload and Orbit type
- Bar Chart: It makes it easy to compare datasets across groups at a glance. One axis represents a category while the other axis represent a discreet value.
  - Success rate Orbit type
- Line Chart: This shows the trends of the data as regards time and can also forecast a future occurrence
  - Success Rate vs Year

EDA with SQL

GitHub Link

- The data was loaded, and SQL queries was executed to answer the following question:
  - 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. Use a subquery
  - 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

Build an Interactive Map with Folium

GitHub Link

- The objects that were created and added to the folium map are:
  - Markers that show all the launch sites on a map
  - Markers that show all the success/failed launches for each site on the map
  - Lines that show the distances between a launch site to its proximities
- These objects were added to answer the following questions:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?
  - Are all launch sites in proximity to the Equator line?
  - Are all launch sites in very close proximity to the coast?

Build a Dashboard with Plotly Dash

GitHub Link

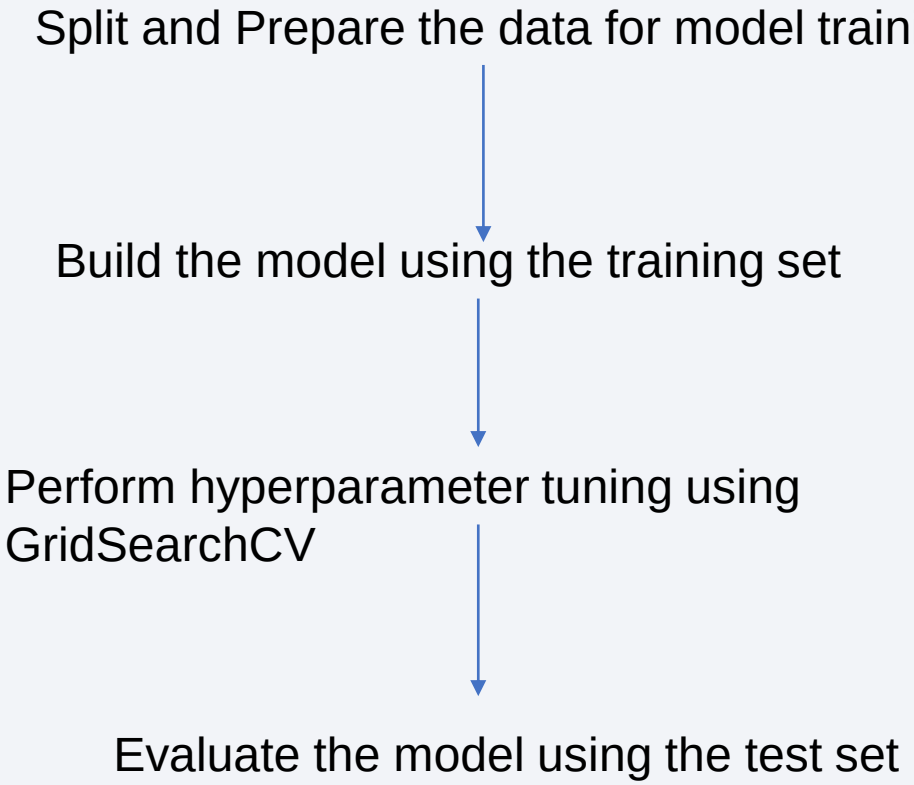
- The charts created were:
- Pie chart
  - This was added to show total success launches by sites to indicate a successful landing distribution across all launch sites or to indicate the success rate of individual launch sites.
- Scatter chart
  - This was added to show the relationship between Outcomes and Payload mass(Kg) by different boosters. It helps determine how success depends on the launch point, payload mass, and booster version categories.
  - Has 2 inputs: All sites/individual site & Payload mass on a slider between 0 and 10000kg

Predictive Analysis (Classification)

GitHub Link

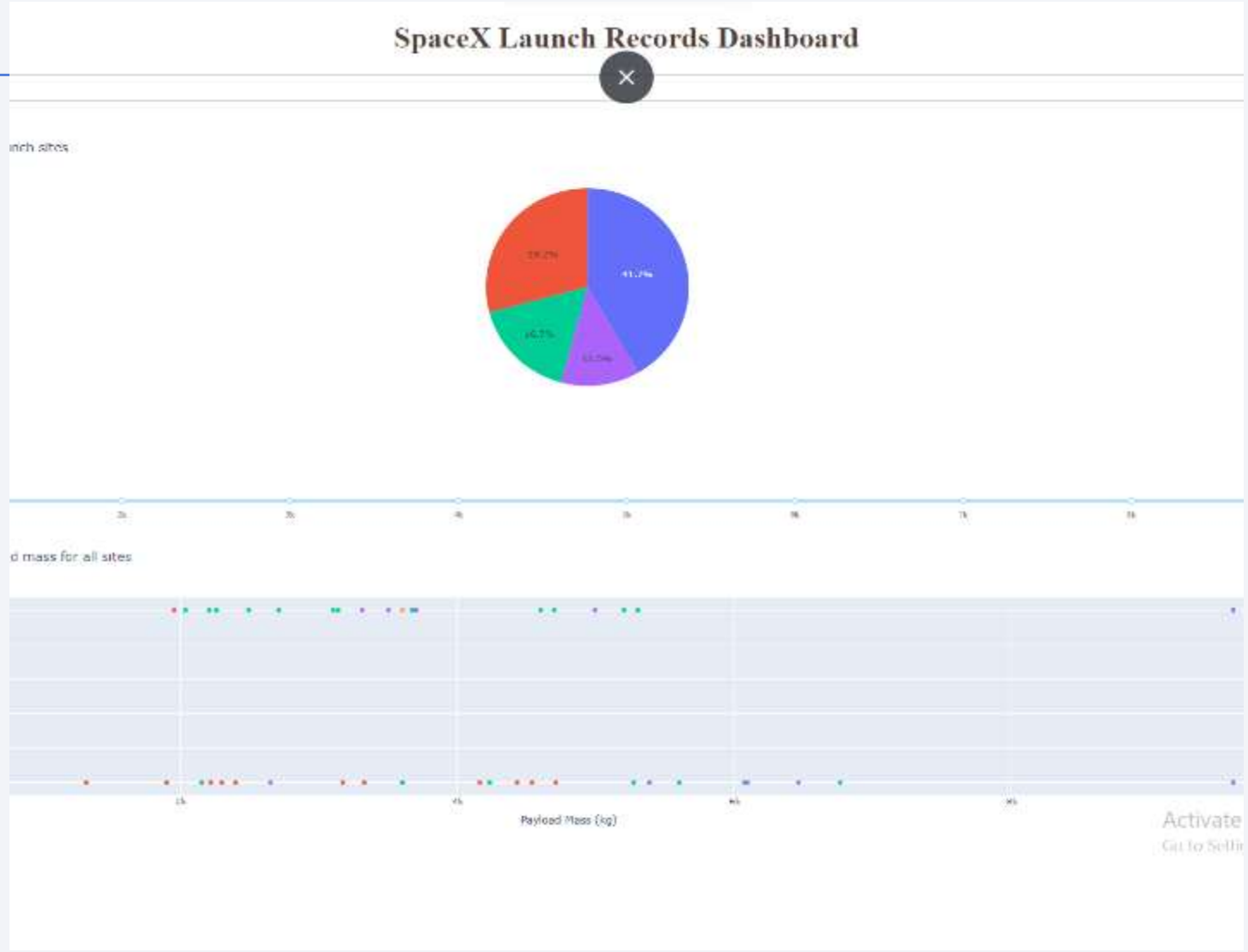
- Performed Exploratory Data Analysis and determined training labels
  - created a column for the class label
  - Standardize the data
  - Split into training data and test data
- Found the best hyperparameter for SVM, KNN, Classification Trees and Logistic Regression with Grid Search CV
- Found the model that performed best using the test data

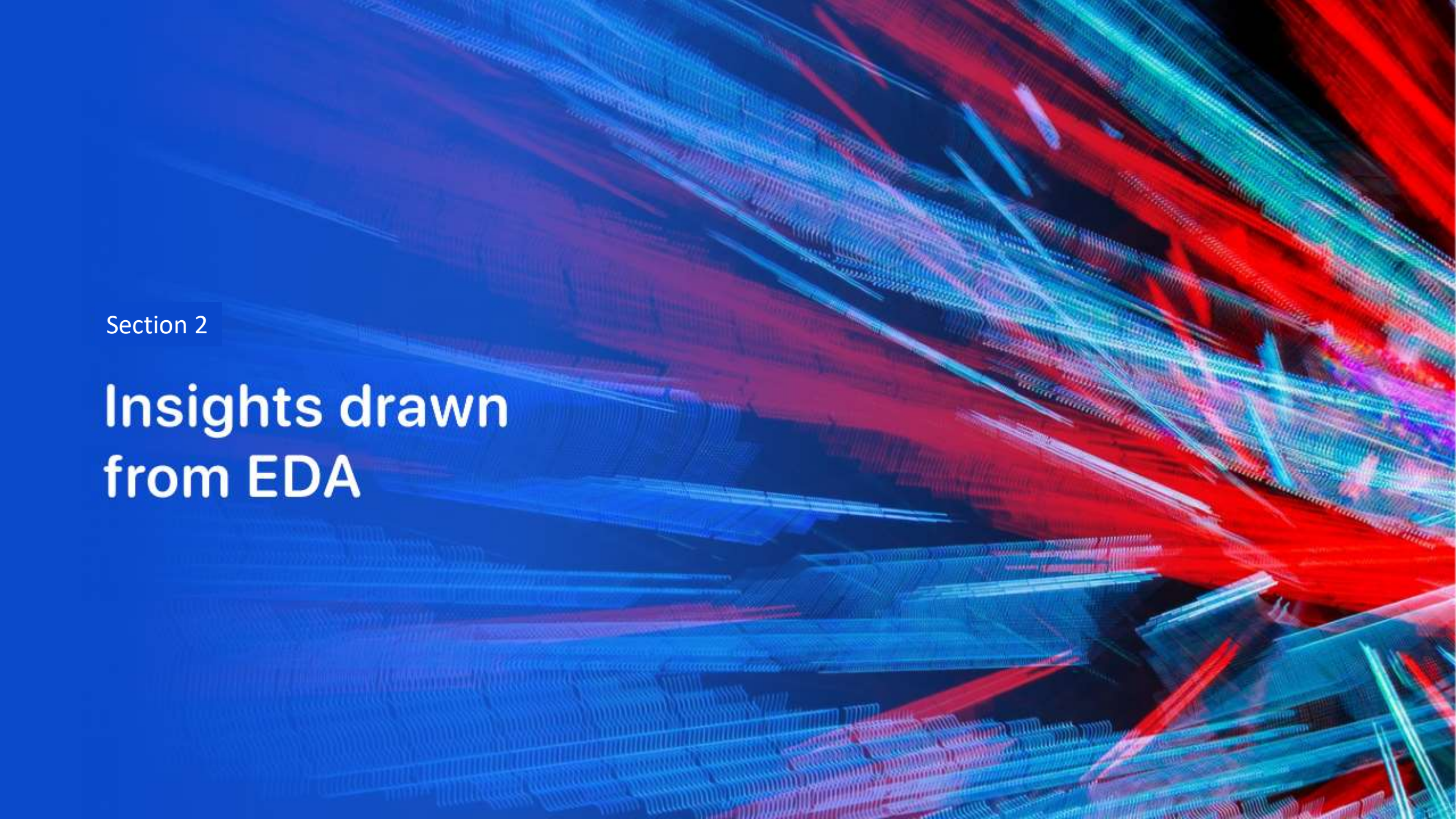
Predictive Analysis (Classification) Contd.



Results

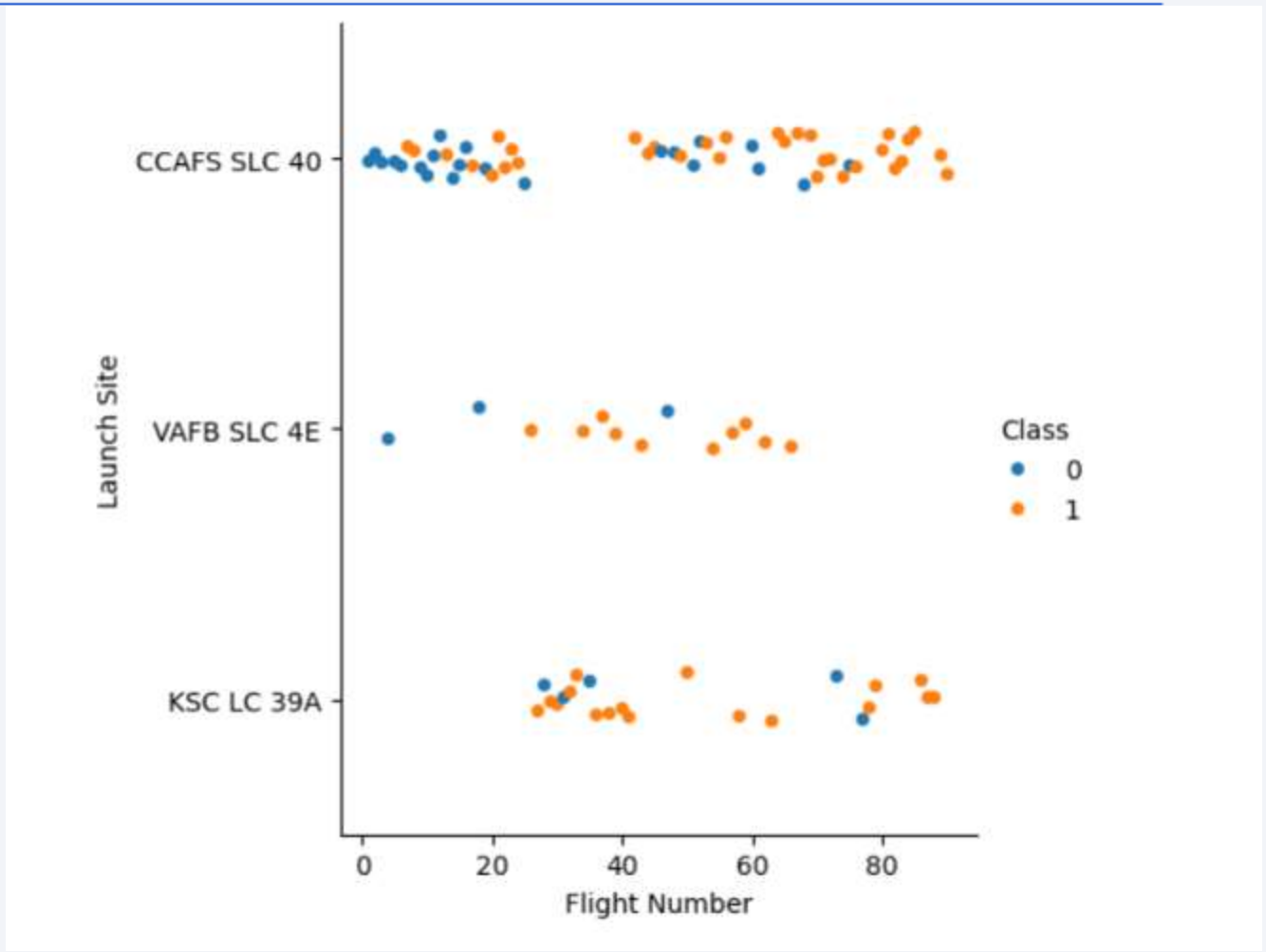
- The picture on the right is preview of the Dashboard with Plotly Dash.
- The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and Interactive Dashboard will be shown in the next slides.
- Comparing the accuracy of the four models used for this project; all return the same accuracy of about 83% for test data.




Section 2
Insights drawn
from EDA

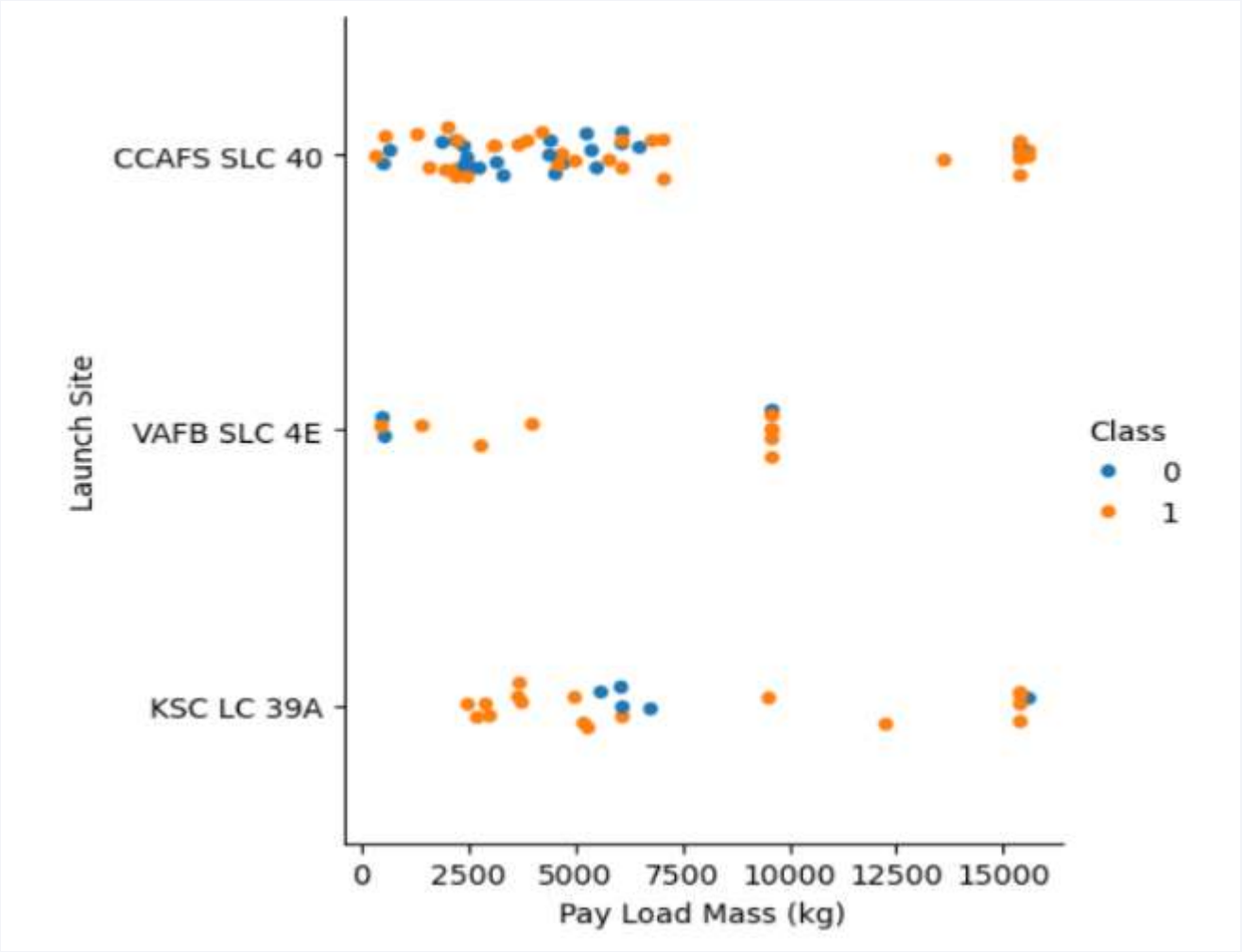
Flight Number vs. Launch Site

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- This chart shows that the success rate for each launch site increased as the number of flights increased.



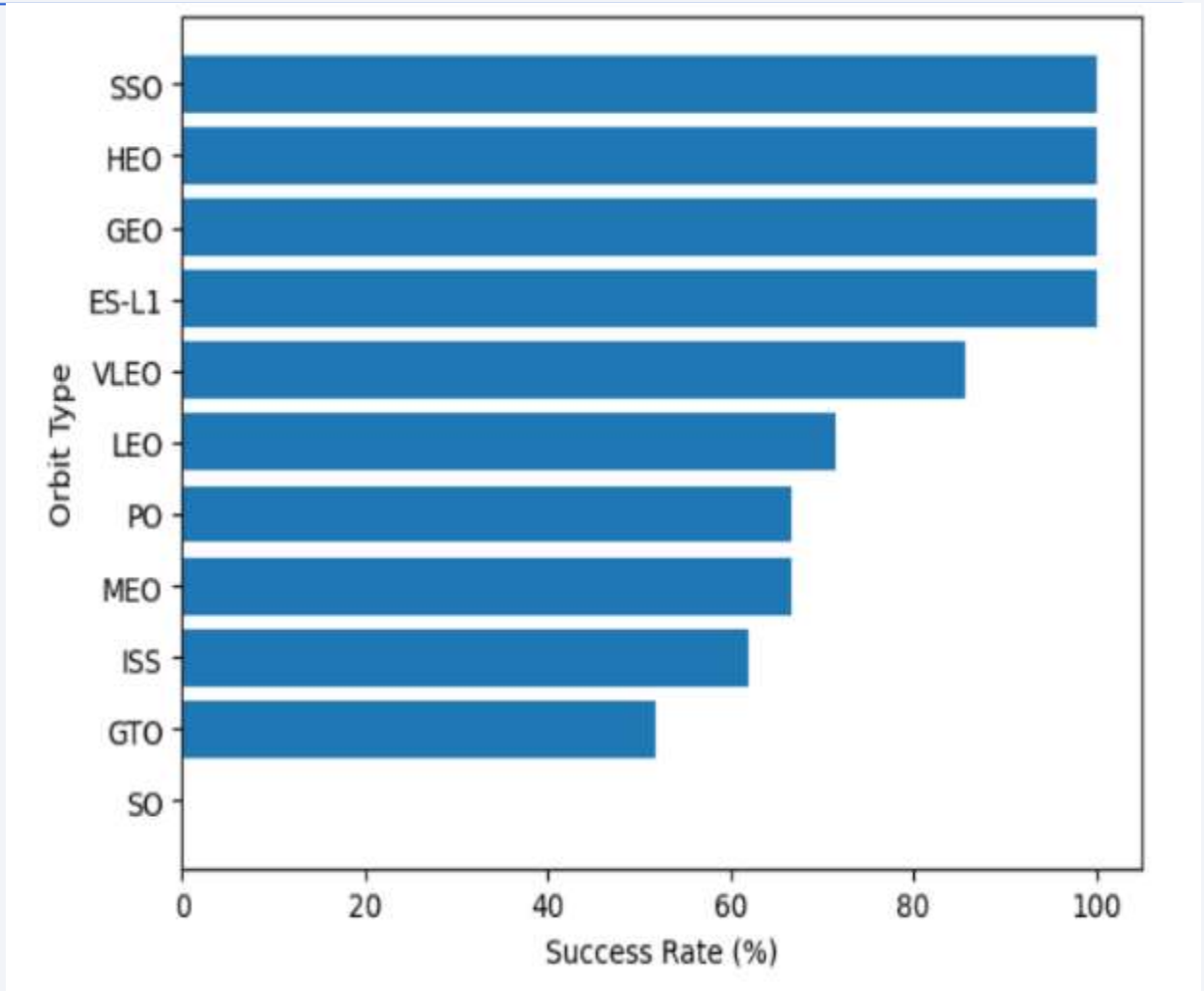
Payload vs. Launch Site

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- There is no correlation between the Launch site and the Pay Load Mass.
- The VAFB-SLC Launch site has no rockets launched for heavy payload mass(greater than 10000)



Success Rate vs. Orbit Type

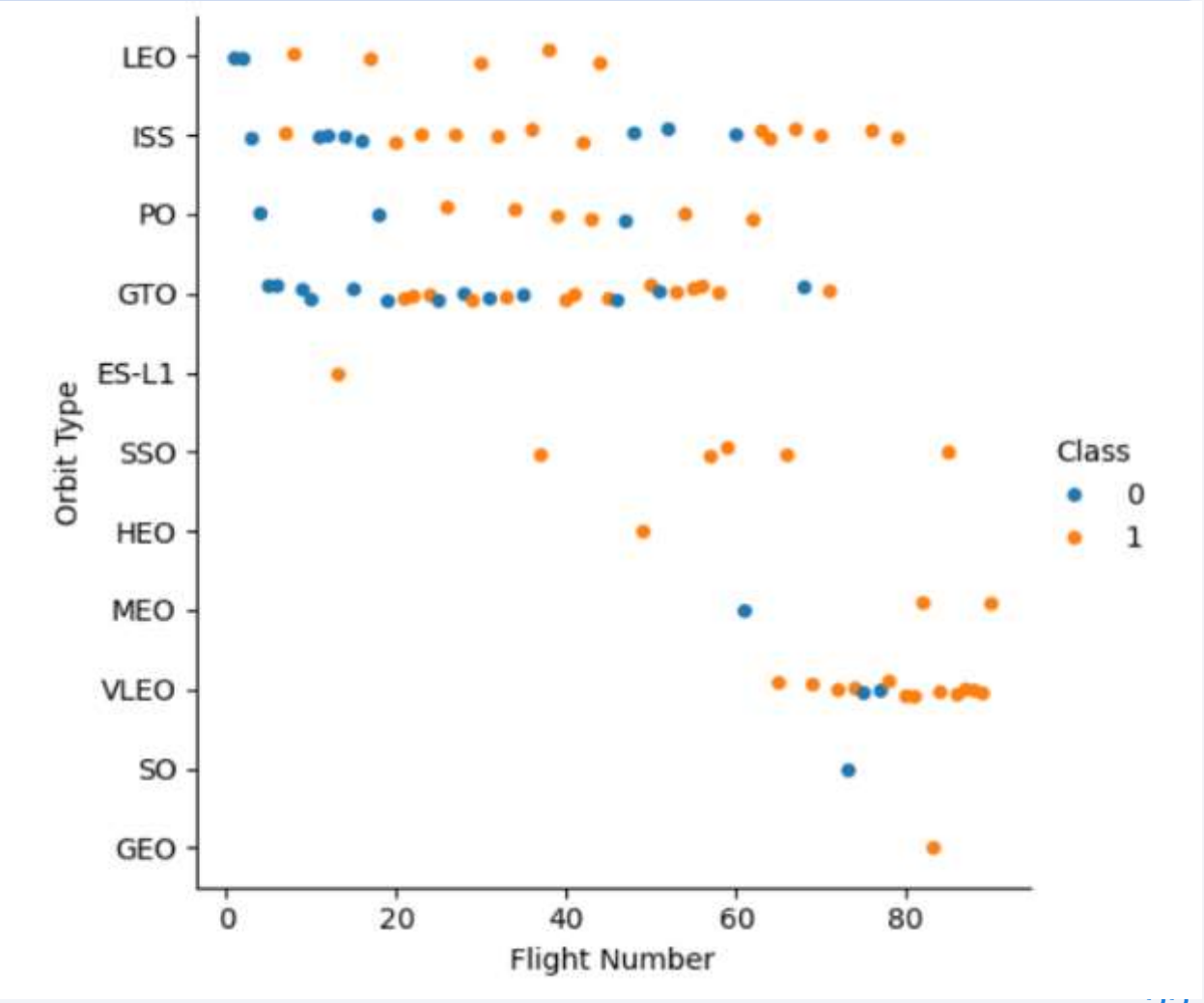
- Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%).
- However, the success rate of orbit type GTO is only 50%, and type SO, which doesn’t have any success rate



Flight Number vs. Orbit Type

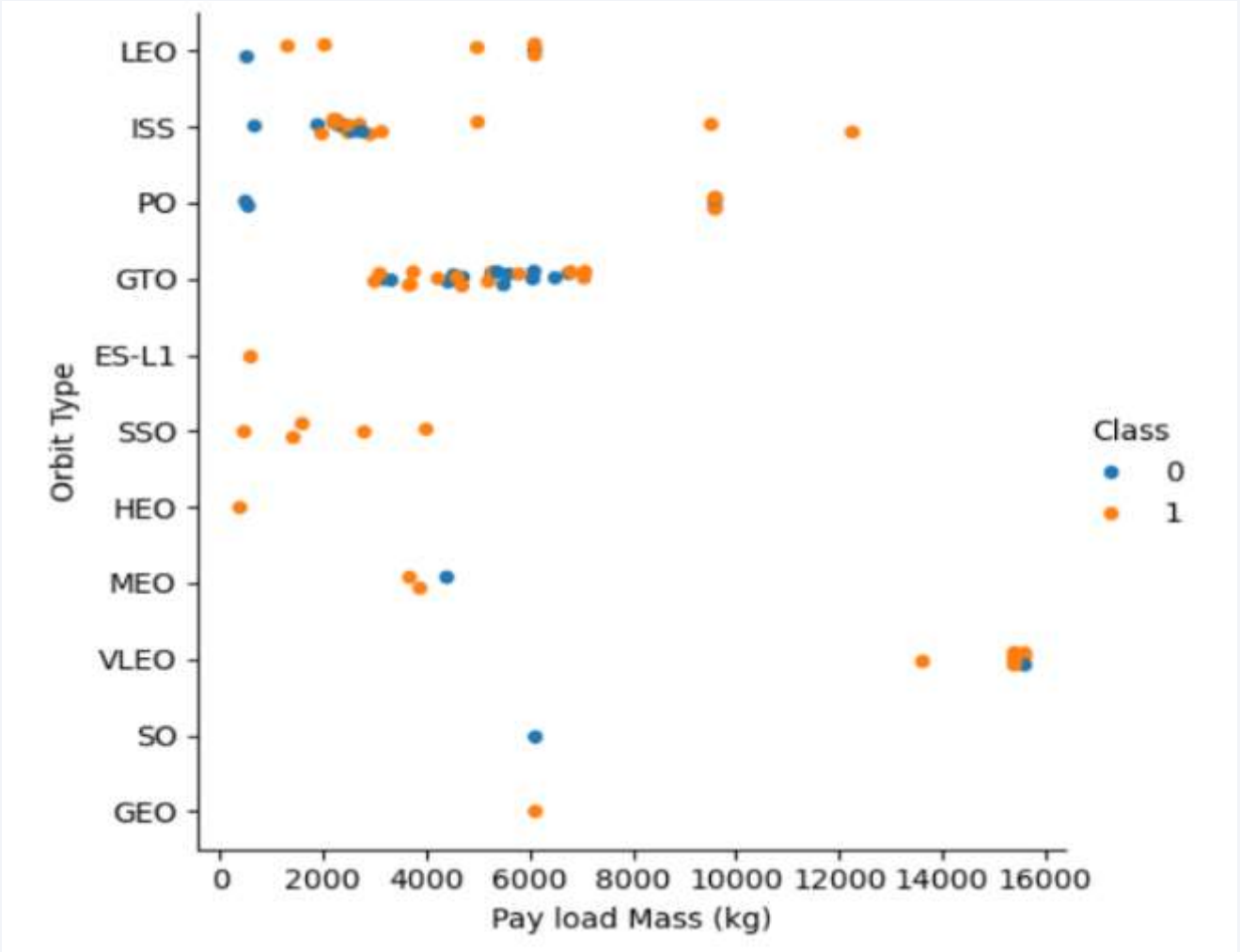
- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.

- From the chart, the LEO orbit success launches appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



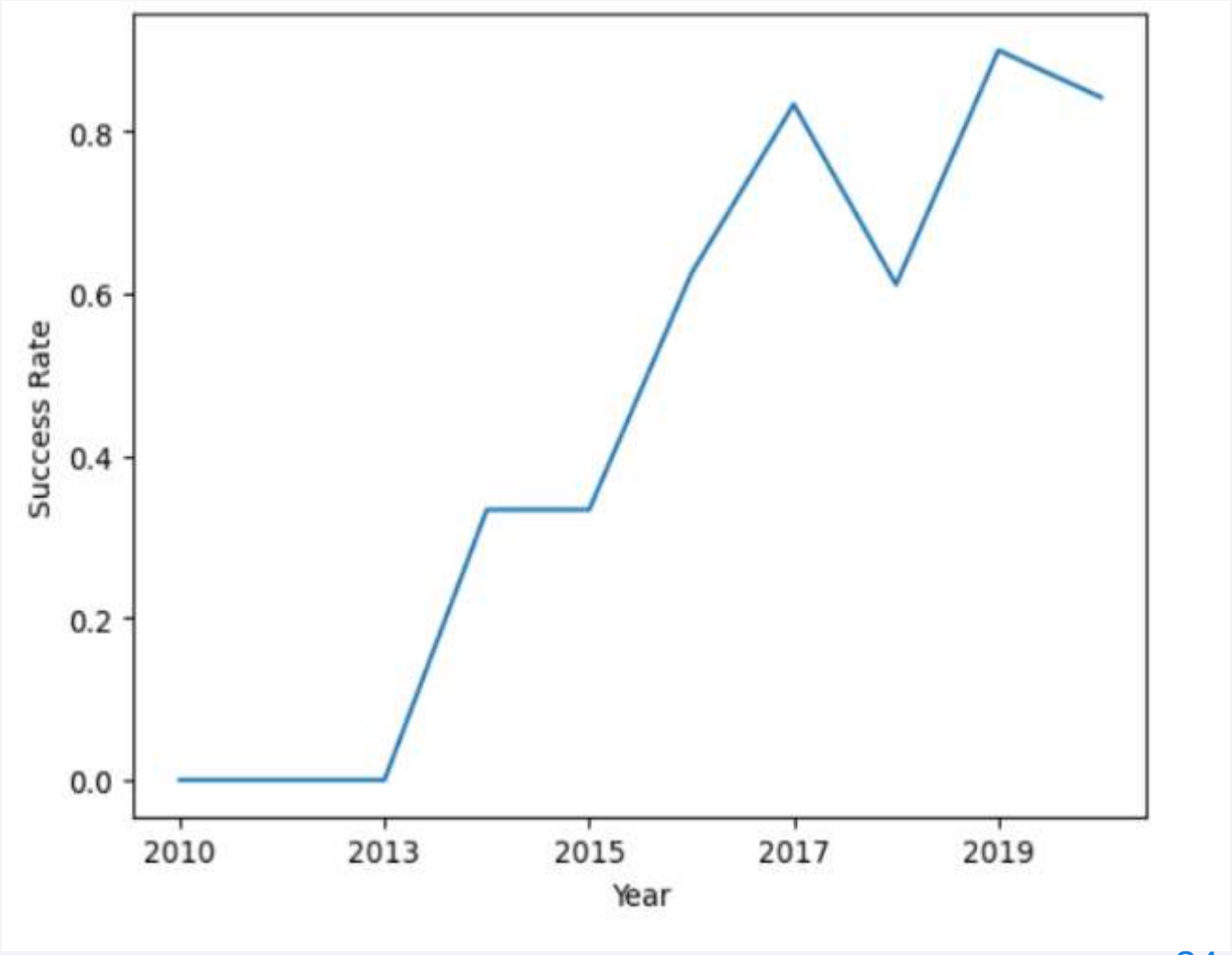
Payload vs. Orbit Type

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both gathered.



Launch Success Yearly Trend

- From the chart, you will observe that the success rate since 2013 kept increasing till 2020 although there was a slight dip in 2018



All Launch Site Names

- With the SQL DISTINCT clause, only unique values are displayed in the Launch_Site column from the SpaceX table.
- The result is four unique launch sites: CCAFS LC-40, VAFB SLC-4E, CCAFS SLC-40, KSC LC-39A

[9]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

[9]:

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

Launch Site Names Begin with 'CCA'

- Only five records of the SpaceX table were displayed using LIMIT 5 clause in the query.
- Using the wild card (LIKE) and the percent sign (%) together, the Launch_Site name starting with CAA was queried.

```
[10]: %%sql
      SELECT LAUNCH_SITE
      FROM SPACEXTBL
      WHERE LAUNCH_SITE LIKE 'CCA%'
      LIMIT 5;
```

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

[10]:

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

Total Payload Mass

- The SUM() function was used to calculate the sum of column PAYLOAD_MASS__KG_.
- The WHERE clause, filtered the dataset to perform calculations only if Customer is “NASA (CRS)”

```
[11]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Customer = 'NASA (CRS)';

       * sqlite:///my_data1.db
      Done.

[11]: SUM(PAYLOAD_MASS__KG_)
      ----------------------
                       45596
```

Average Payload Mass by F9 v1.1

- The AVG() function was used to calculate the sum of column PAYLOAD_MASS__KG_.
- The WHERE clause, filtered the dataset to perform calculations only if Booster_Version is “F9 v1.0”

```
[14]: %%sql
      SELECT AVG(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Booster_Version LIKE 'F9 v1.0%';

       * sqlite:///my_data1.db
      Done.

[14]: AVG(PAYLOAD_MASS__KG_)
      ----------------------
                       340.4
```

First Successful Ground Landing Date

- The MIN() function was used to find out the earliest date in the column DATE.
- The WHERE clause, filtered the dataset to perform a search only if Landing__outcome is “Success (ground pad)”

```
[16]: %%sql
      SELECT MIN(Date)
      FROM SPACEXTBL
      WHERE Landing_Outcome = 'Success (ground pad)';
```

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

[16]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The WHERE clause filtered the dataset to perform a search if Landing__outcome is Success (drone ship) while using the AND operator to display a record if additional condition PAYLOAD_MASS__KG_ is between 4000 and 6000.

```
[17]: %%sql
      SELECT BOOSTER_VERSION
      FROM SPACEXTBL
      WHERE LANDING_OUTCOME = 'Success (drone ship)'
          AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

[17]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- The COUNT() function was used to calculate the total number of columns.
- While the GROUP BY statement groups rows that have the same values into summary rows to find the total number in each Mission_outcome.

[18]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

[18]:

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

Boosters Carried Maximum Payload

- The subquery was first used to find the maximum value of the payload by using MAX() function, and then the result was filtered to perform a search if PAYLOAD_MASS__KG_ is the maximum value of the payload.

```
[19]: %%sql
      SELECT DISTINCT BOOSTER_VERSION
      FROM SPACEXTBL
      WHERE PAYLOAD_MASS__KG_ = (
          SELECT MAX(PAYLOAD_MASS__KG_)
          FROM SPACEXTBL);
```

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

[19]:

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

```
%%sql
SELECT substr(Date, 6, 2) AS month, Landing_Outcome AS failure_landing_outcomes, Booster_Version AS booster_version, Launch_Site AS launch_site
FROM SPACEXTBL
WHERE substr(Date, 1, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

month	failure_landing_outcomes	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

- The WHERE clause was used to filter the dataset to perform a search if Landing__outcome is Failure (drone ship). While using the AND operator to display a record with YEAR is 2015.
- The result shows that in 2015, there were two landing failures on drone ships.

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

- The WHERE clause was used to filter the dataset to perform a search if the date is between 2010-06-04 and 2017-03-20.
- While using the ORDER BY to sort the records by total number of landing and using DESC keyword to sort the records in descending order.
- According to the results, the number of successes and failures between 2010-06-04 and 2017-03-20 was similar.

[36]:
```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

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

[36]:

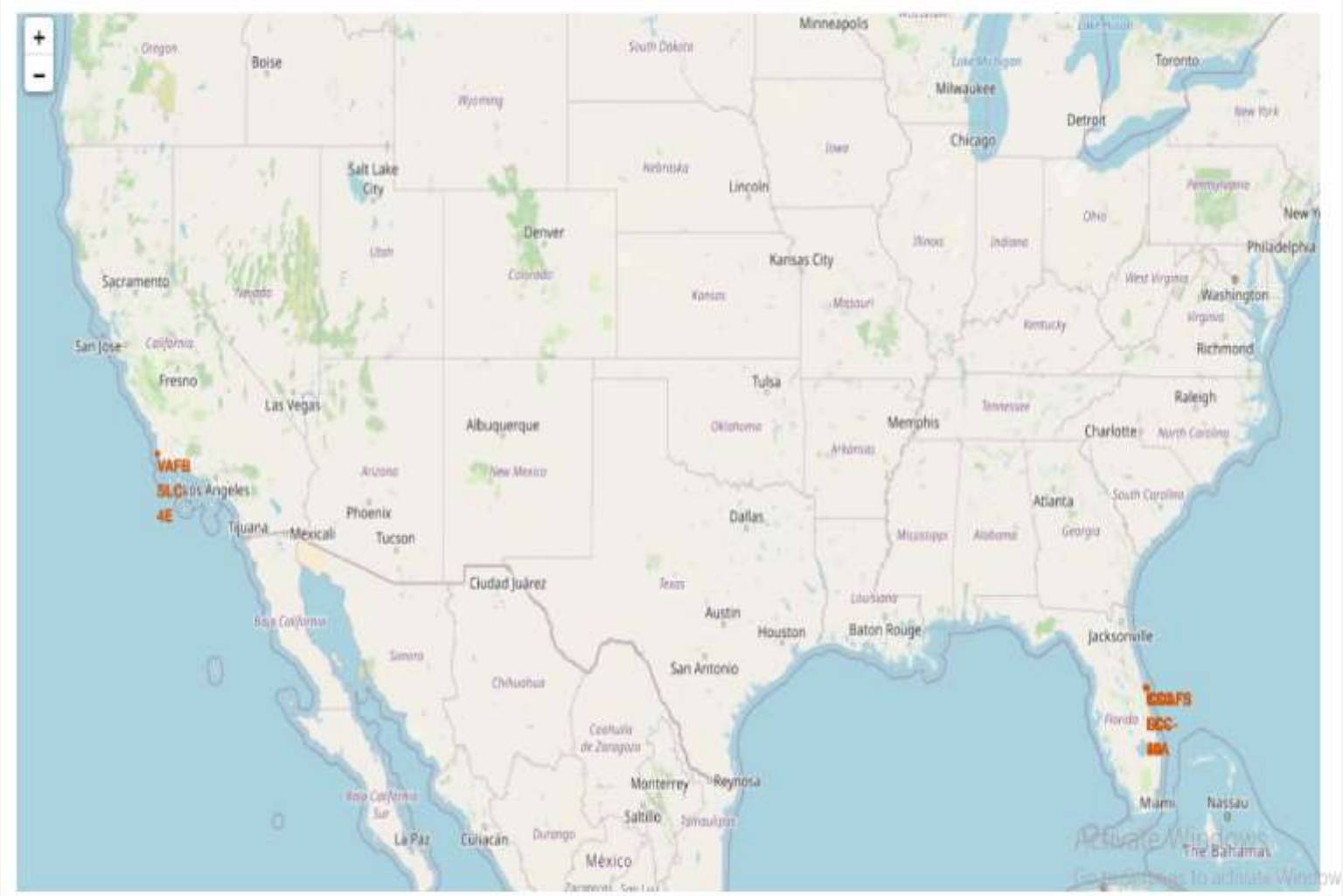
Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

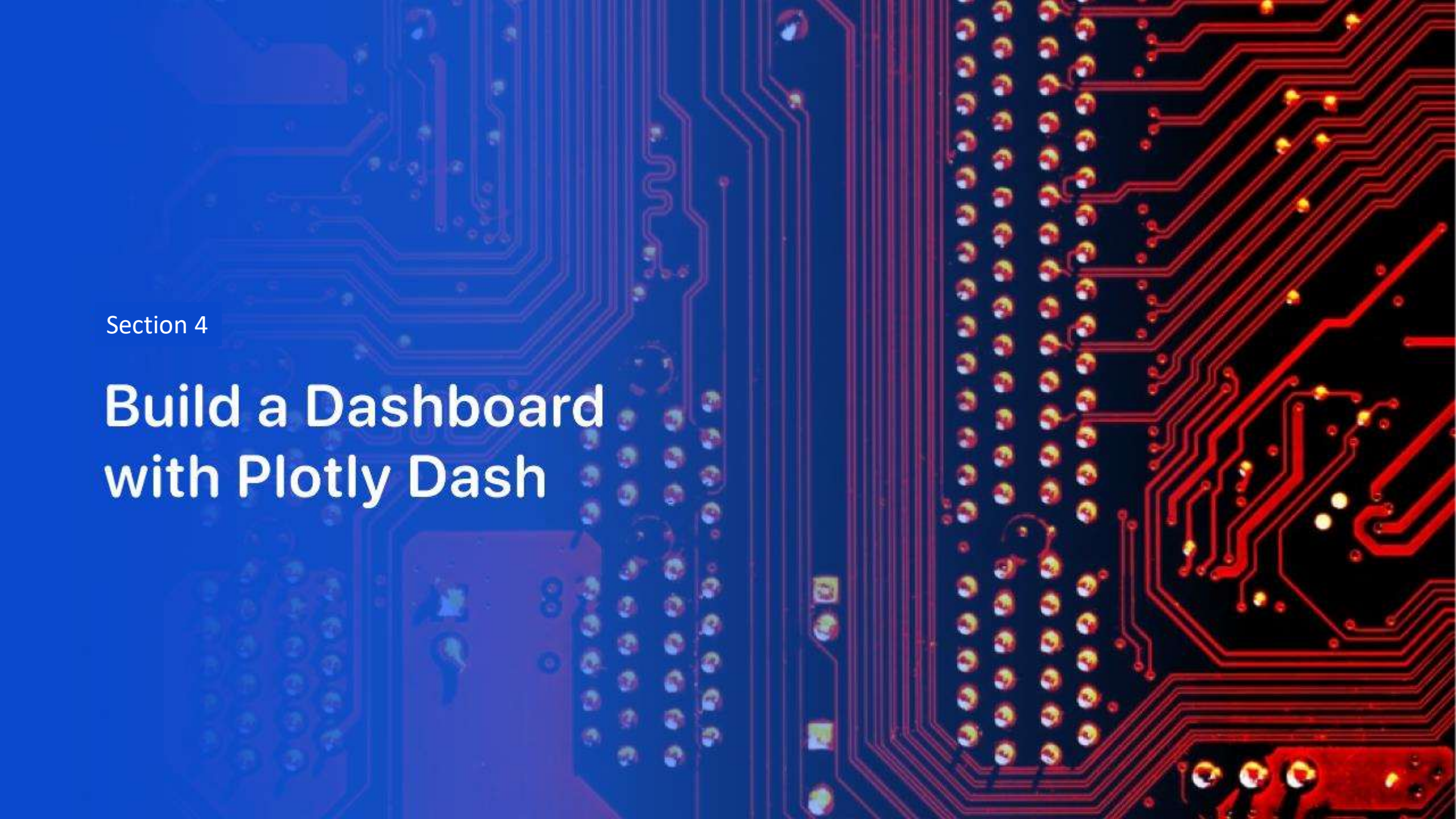
Section 3

Launch Sites Proximities Analysis

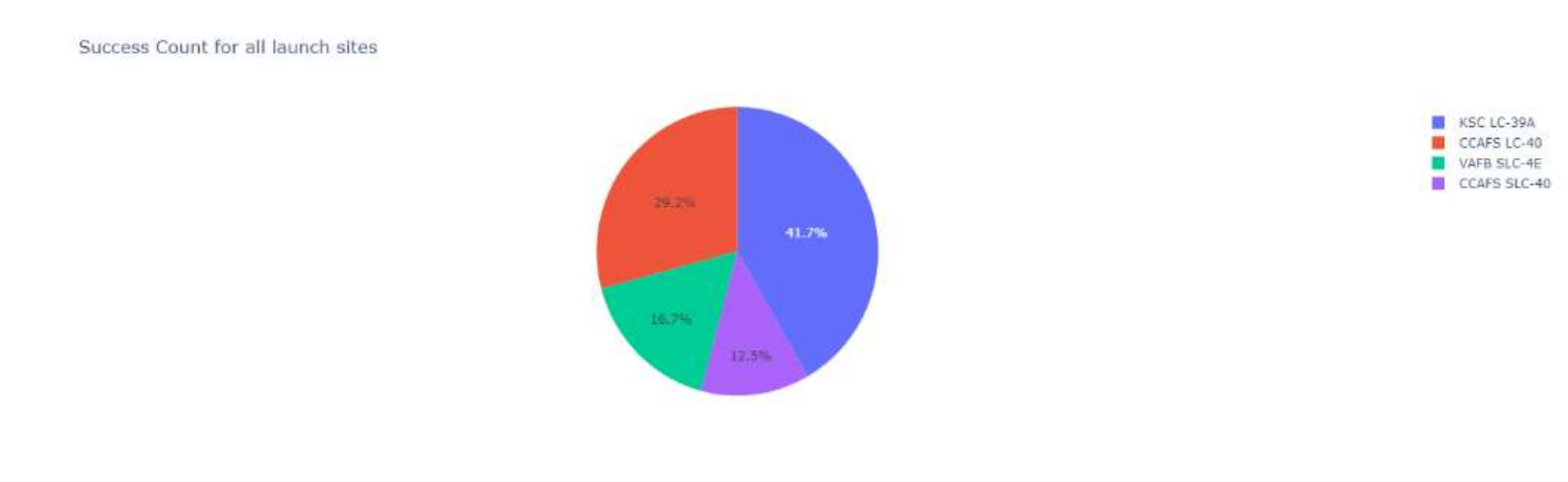
All Launch Sites' Locations

- The map shows all SpaceX launch sites, and it shows that they are all in the United States.
- Also, it can be seen on the map that all launch sites are near the coast.



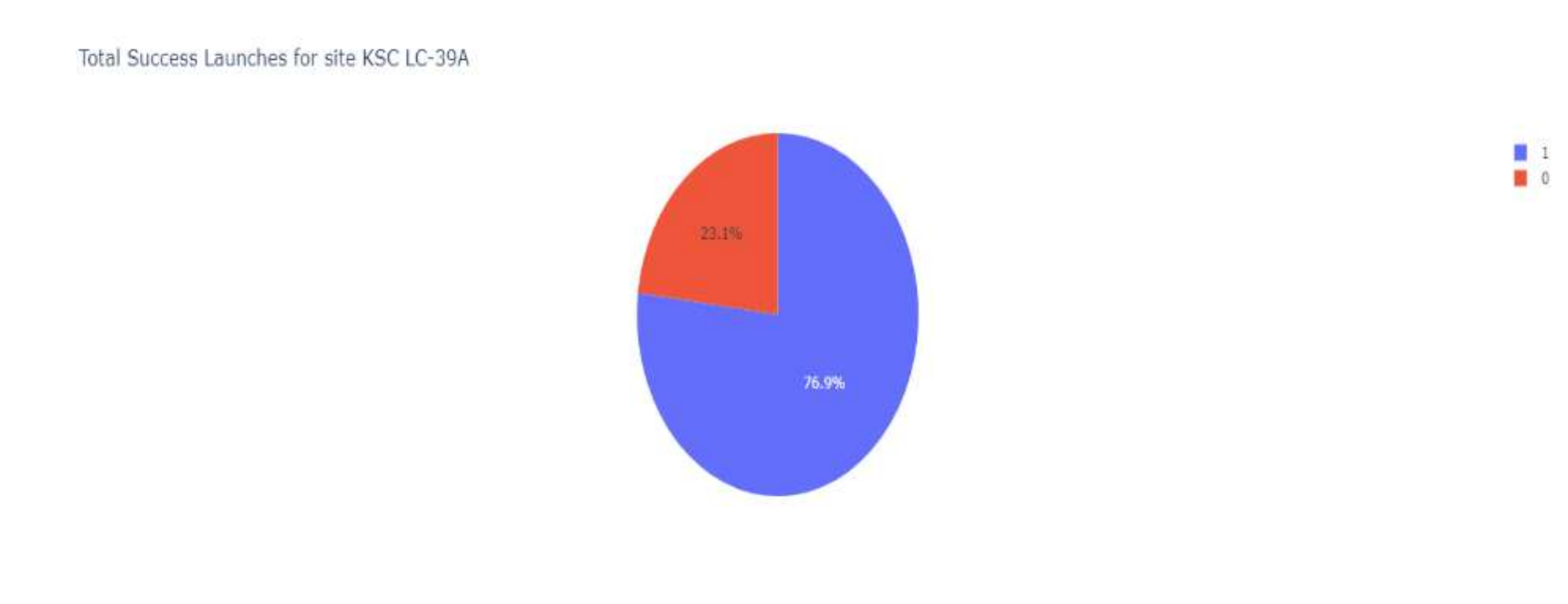

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launch by All Sites



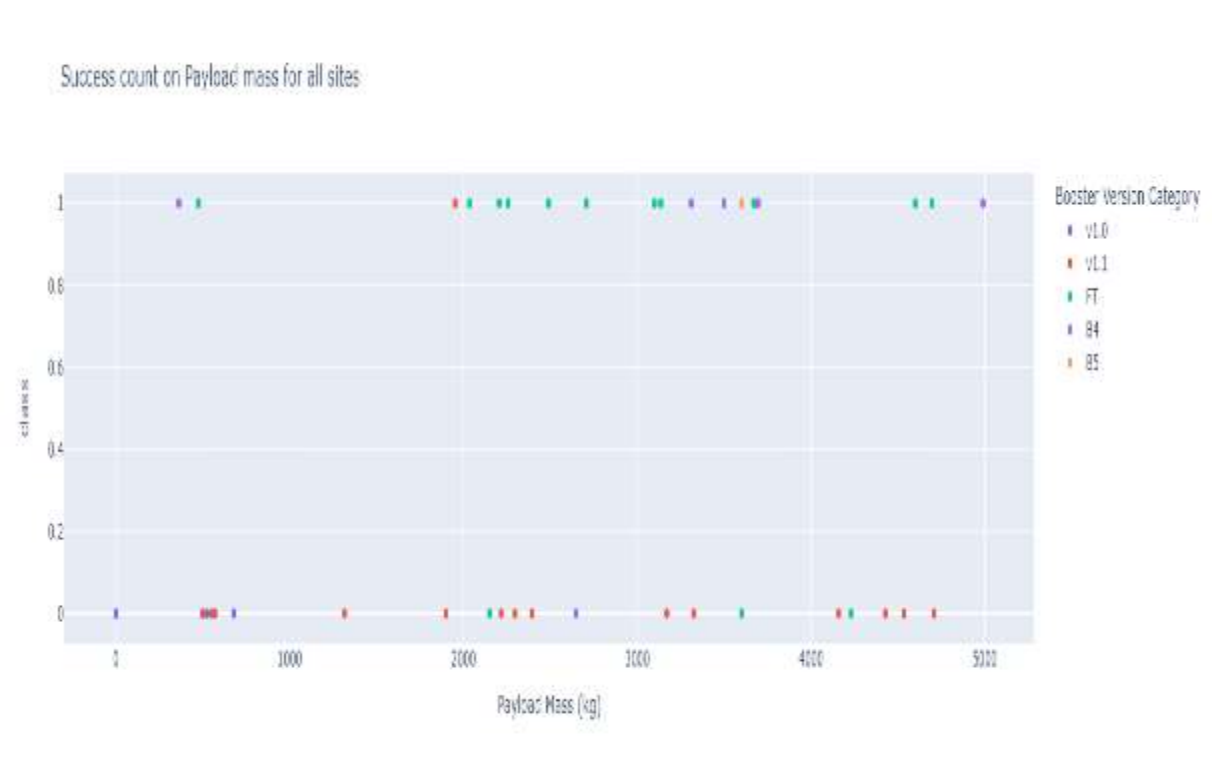
- The KSLC-39A records the most launch success among all sites while the CCAFS SLC-40 site has the fewest launch success.

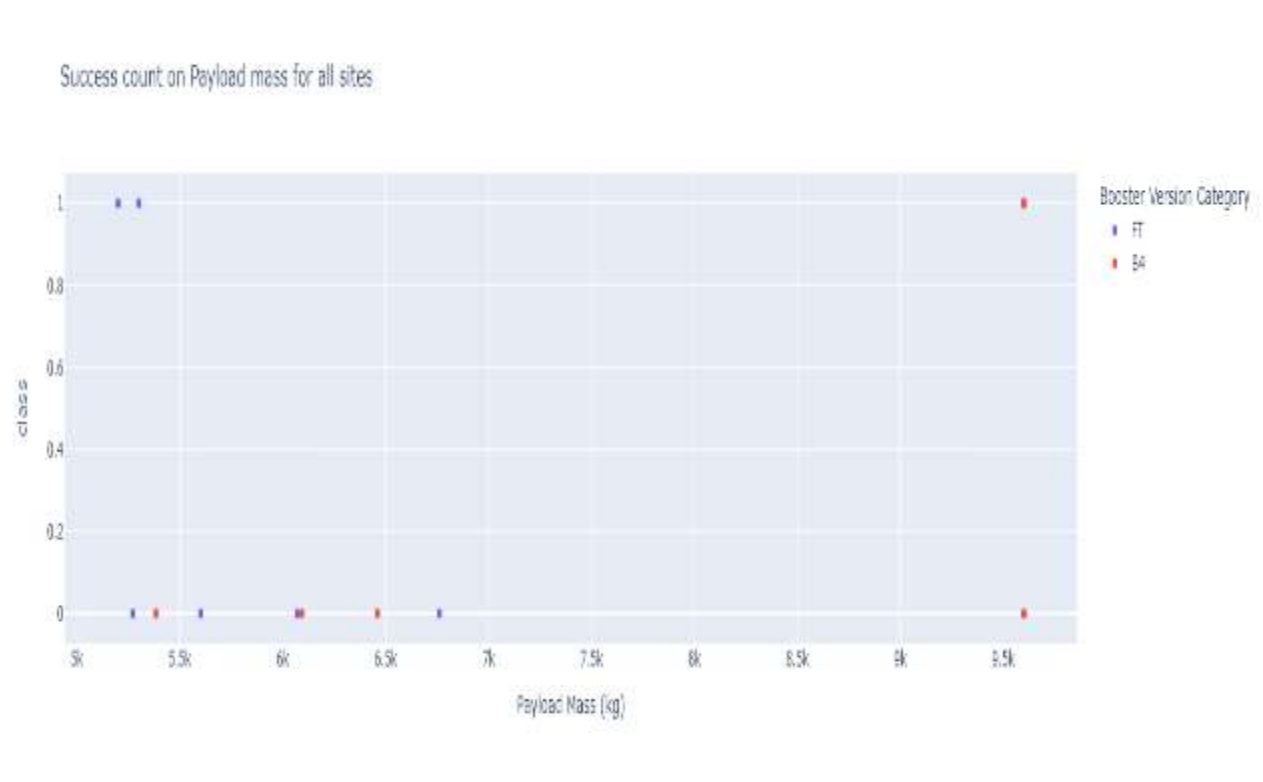
The Launch Site with Highest Launch Success Ratio



- The KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%).

Payload vs. Launch Outcome Scatter Plot for All Sites





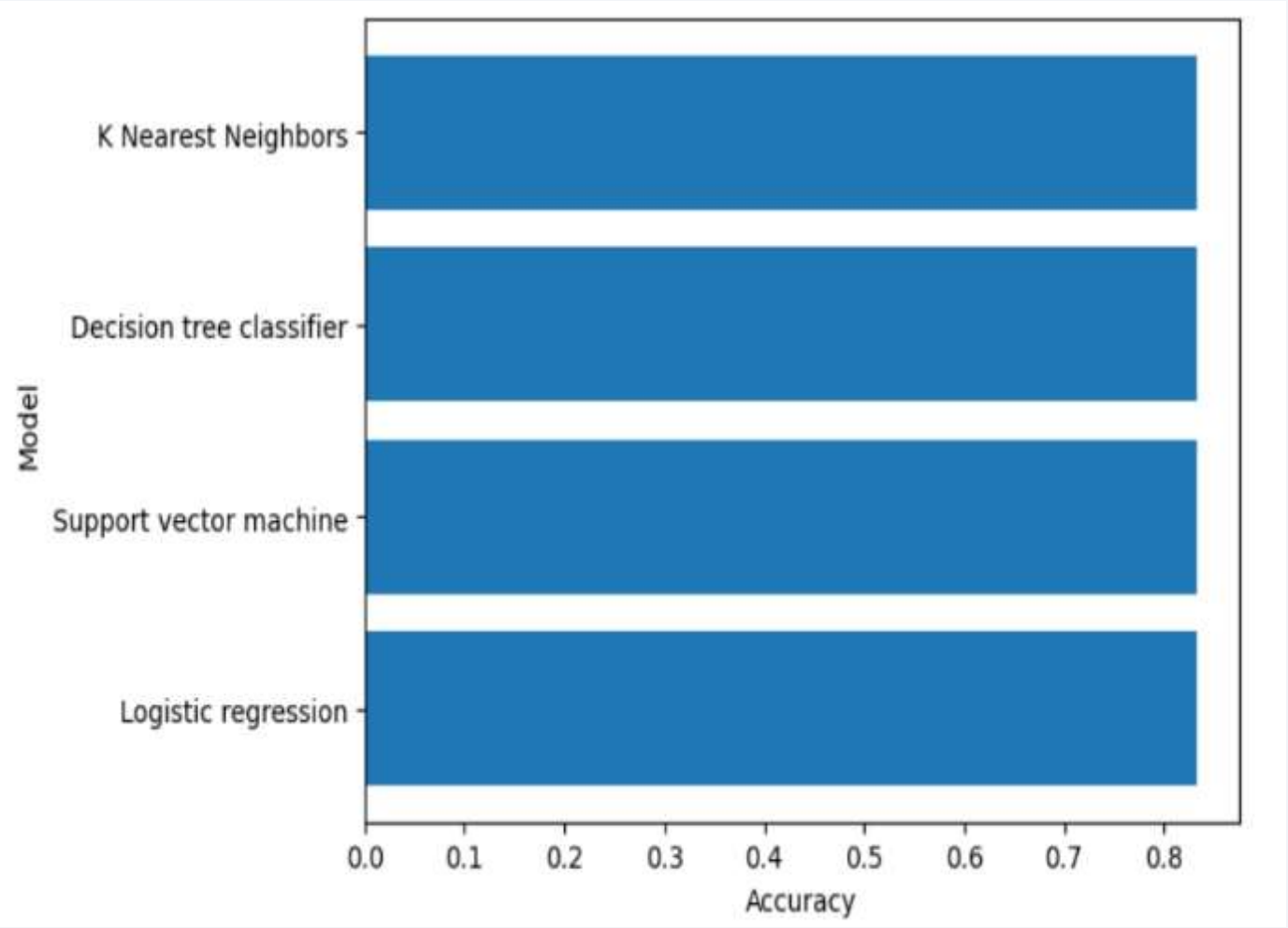
- These charts show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is higher than that of heavy weighted payloads(5000-10000 kg)

Section 5

Predictive Analysis (Classification)

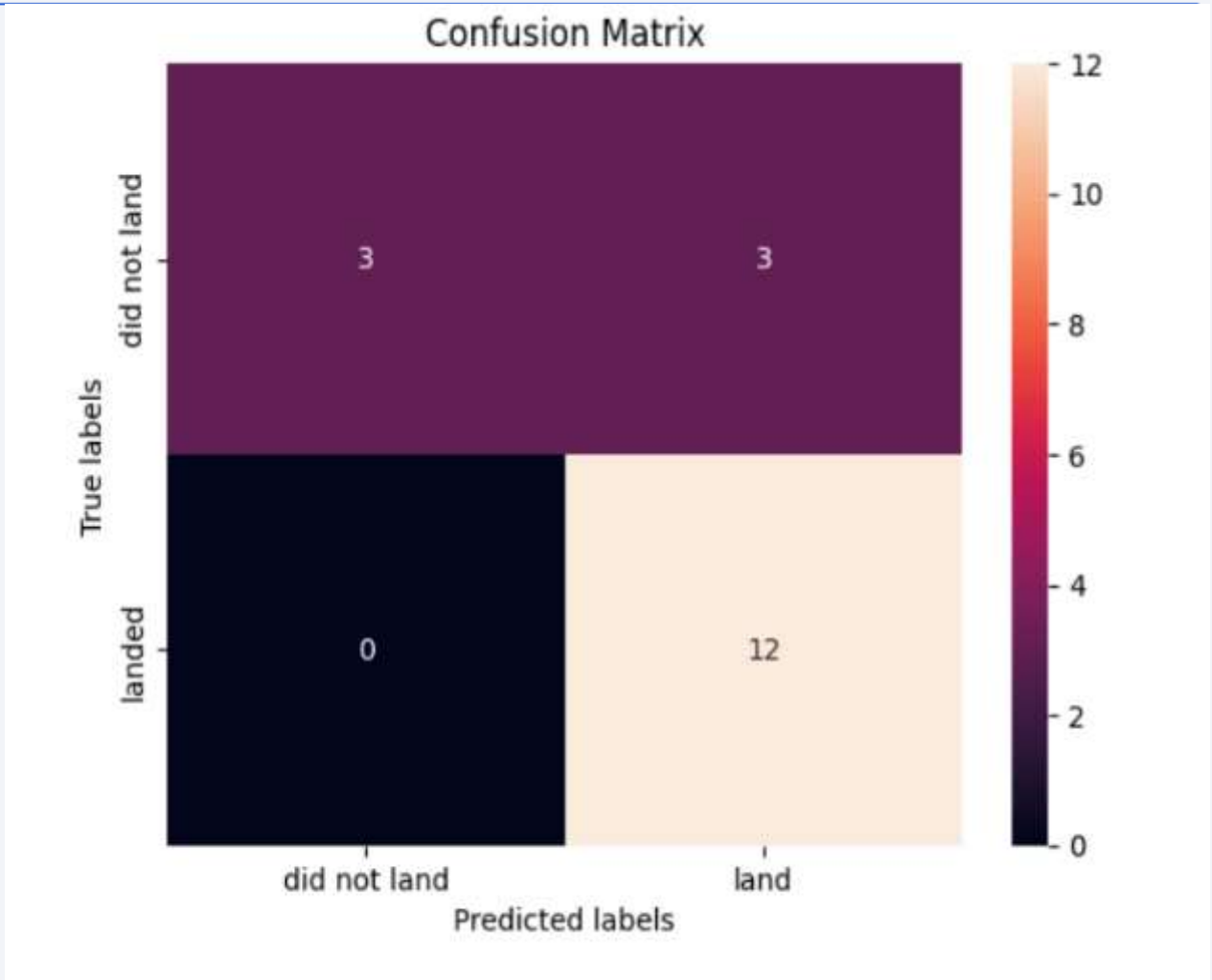
Classification Accuracy

- After comparing the accuracy of built models, they all preformed practically the same (83%) on the test data, except for the tree model which fit train data slightly better than the rest.
- It is good to note that these accuracy might increase with increase in the training data as small amount of data was used for training.



Confusion Matrix

- Coincidentally, all the confusion matrix of the four model are all the same because all models performed the same for the test set.
- The models predicted correctly 12 successful landings when the true label shows successful landing and 3 failed landings when the true label was failure.
- However, there were also 3 predictions that said successful landings when the true label shows failure (false positive).



Conclusions

In conclusion, the following can be deduced from the project:

- As the number of flights increased, the success rate also increased, and this shows that they got better successful landing with more trails.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%) and this shows that more focus should be on these orbital types.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.
- Finally, all the models have the same accuracy (83.33%), but it seems that more training data is needed to help the models perform better and improve their accuracy.

Appendix

- GitHub Link

Thank you!