IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Michal Ciemiega>
<August 13th, 2024>

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected from official SpaceX website and Wikipedia
  - Visual analysis were implemented to identify parameters which have impact on outcome
  - Interactive map was created to visualize launch sites and their sourrandings
- Summary of all results
  - Consecutive launches have positive impact on outcome – visible learning
  - There's no connection between payload mass and outcome
  - Some launch sites were used only for lighter payloads
  - Some orbits, like GEO, where targeted only during later launches
  - Heaviest payloads were sent to GEO orbit
  - Success rate improved with time
  - The best model to predict future outcomes would be Decision Tree model

Introduction

SpaceX 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 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two different sources – SpaceX website and Wikipedia
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A few different classification models were built and compared for their accuracy

Data Collection

- Data sets where collected with two different methods:
  - From SpaceX website, via available REST API
  - From Wikipedia, via web scraping tools

Data Collection – SpaceX API

- Following steps were performed:
  - Data download via API with GET request
  - Data normalization
  - Choosing columns that may have impact on Outcome
  - Data nested in cells was separated into different collumns
  - Data was filtered to include only Falcon9
  - Missing values in Payload column were replaced with average payload
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Following steps were performed
  - HTML page was downloaded with get method
  - Beautiful Soup object was created and html file was parsed
  - Table with Falcon9 launches was found and extracted
  - Data from the table was extracted and imported to Data Frame
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- In the first step following information was collected form the data
  - Number of launches from different launch sites
  - Number of launches to different orbits
  - Number of different outcomes
- Since outcomes could be separated in Success and Failure category, additional column – class was created, which contained numbers 1 for success and 0 for failure
- At the end average outcome was calculated
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Following charts where created
  - Scatter plot visualizing flight no., payload and outcome – to check if payload as well as learning with following launches have impact on the outcome
  - Scatter plot visualizing flight no., launch site and outcome – to check if launch site as well as learning with following launches have impact on the outcome
  - Scatter plot visualizing flight no., payload and launch site – to check if payload and launch site have impact on the outcome and identify which payloads were launched from which sites
  - Bar chart visualizing mean outcome for launches to different orbits – to identify if orbit have impact on the outcome
  - Scatter plot visualizing flight no., orbit and outcome – to check if orbit as well as learning with following launches have impact on the outcome
  - Scatter plot visualizing payload mass, orbit and outcome – to check if payload mass as well as orbit have impact on the outcome
  - Line plot, to visualize changes in outcome with time (years)

- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Following queries were performed
  - %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
  - %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 20;
  - %sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total_Payload_Mass" FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';
  - %sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Avarage_Payload_Mass" FROM SPACEXTABLE WHERE "Booster_Version" LIKE '%F9 v1.1%';
  - %sql SELECT MIN("Date") AS "First_successful_landing_date" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
  - %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
  - %sql SELECT "Mission_Outcome", COUNT(*) AS "Total_count" FROM SPACEXTABLE GROUP BY "Mission_Outcome";
  - %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
  - %sql SELECT CASE substr(Date, 6, 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' END AS Month_Name, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome="Failure (drone ship)"
  - %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;

- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Following objects were added to the map
  - Markers – to mark launch sites
  - Circles – to make launch sites more visible
  - Marker Clusters – to mark successful and failed landing
  - Lines – to show ditances to neares objetcs like highways, costlines or cities
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

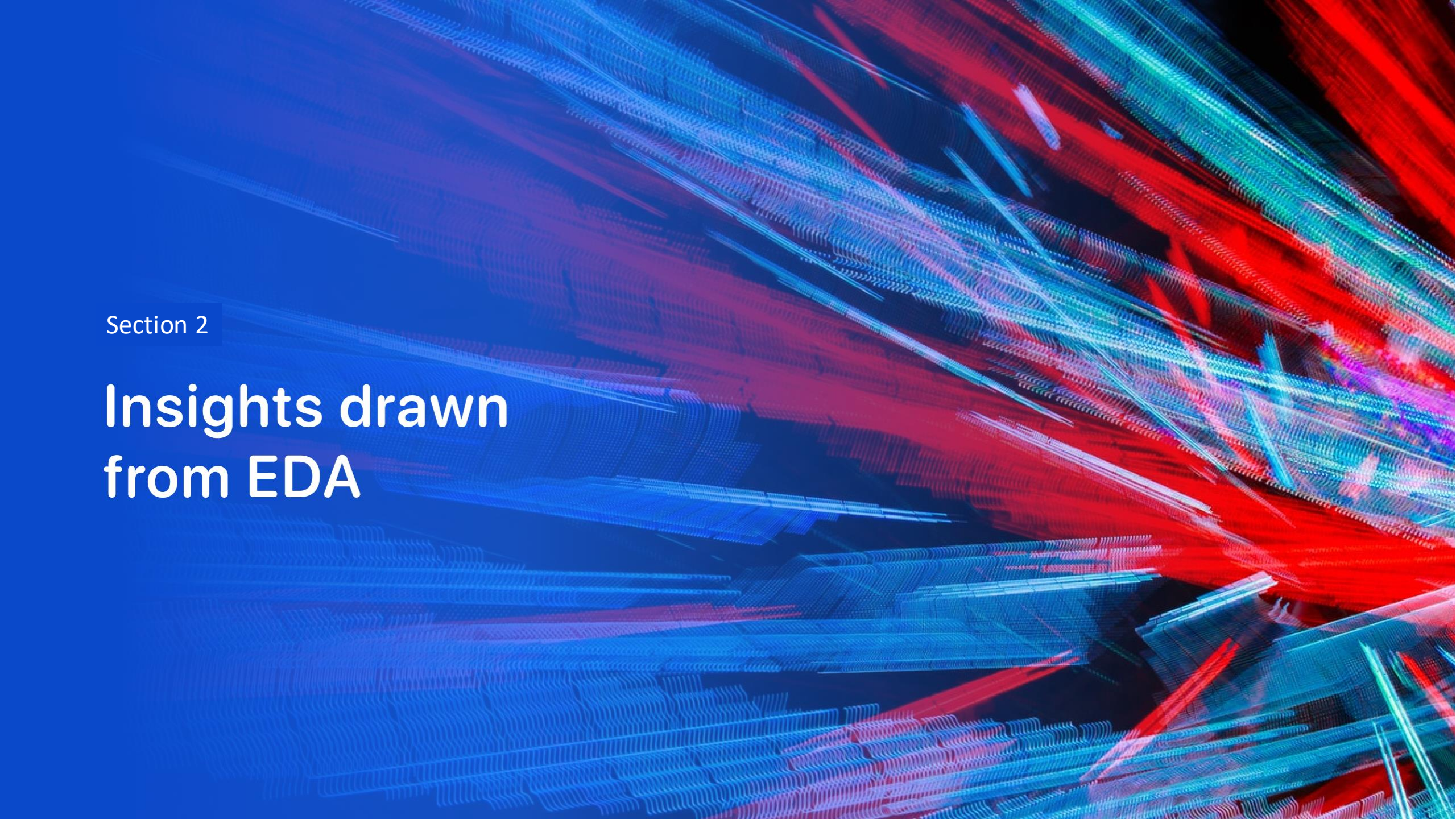
- Following plots were added to the dashboard
  - Pie Chart – to present successful landing from different launch sites
  - Scatter Pot – to present outcome for different payloads for different sites
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

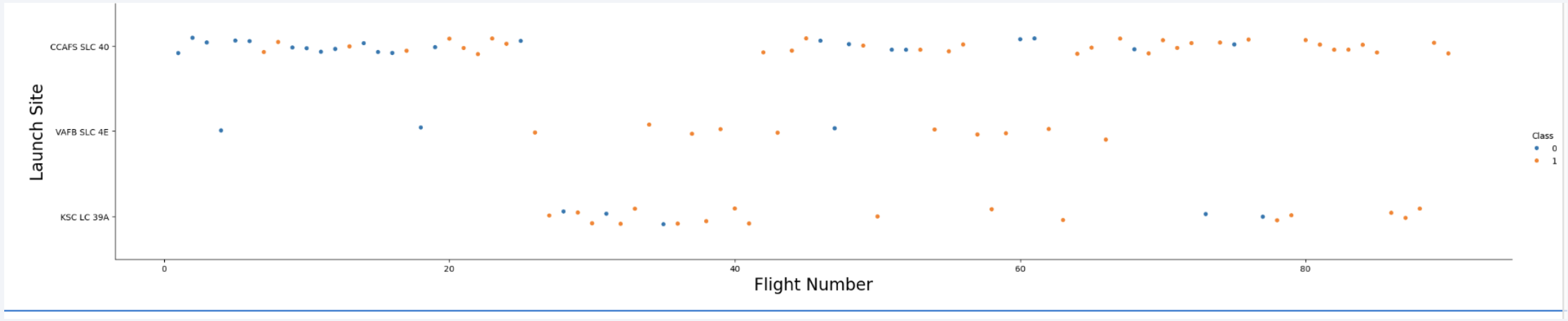
- Data was transformed with Standard Scaller and split into train and test data
- Different models were built – Logistic Regressin, Decission Tree, SVM, KNN
- Best parameters were chosen with GridSearch, confussion matrixes were plotted and accuracy of different models were compared
- Decision Tree scored the best reasult
- https://github.com/michalciemiega/Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

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

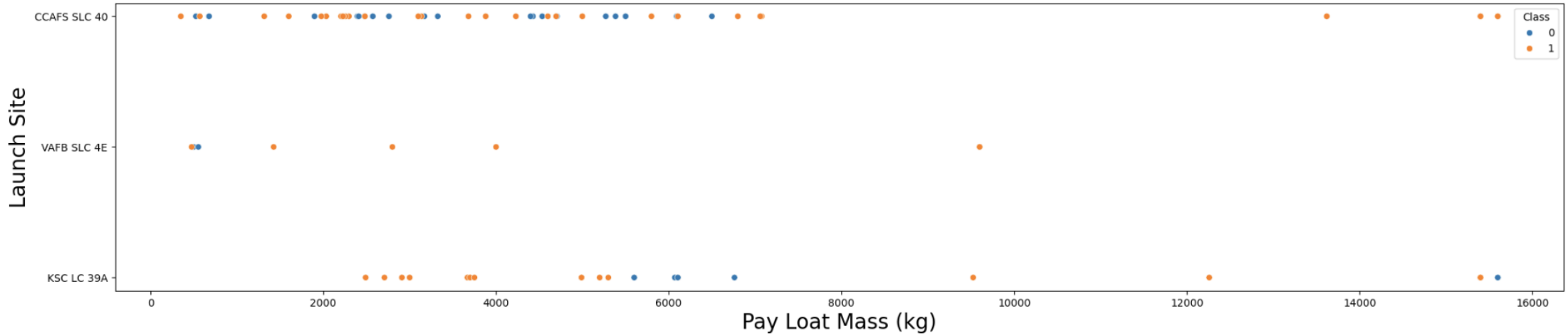

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Launch site doesn't seam to have impact on the outcome, but flight numer does

Payload vs. Launch Site

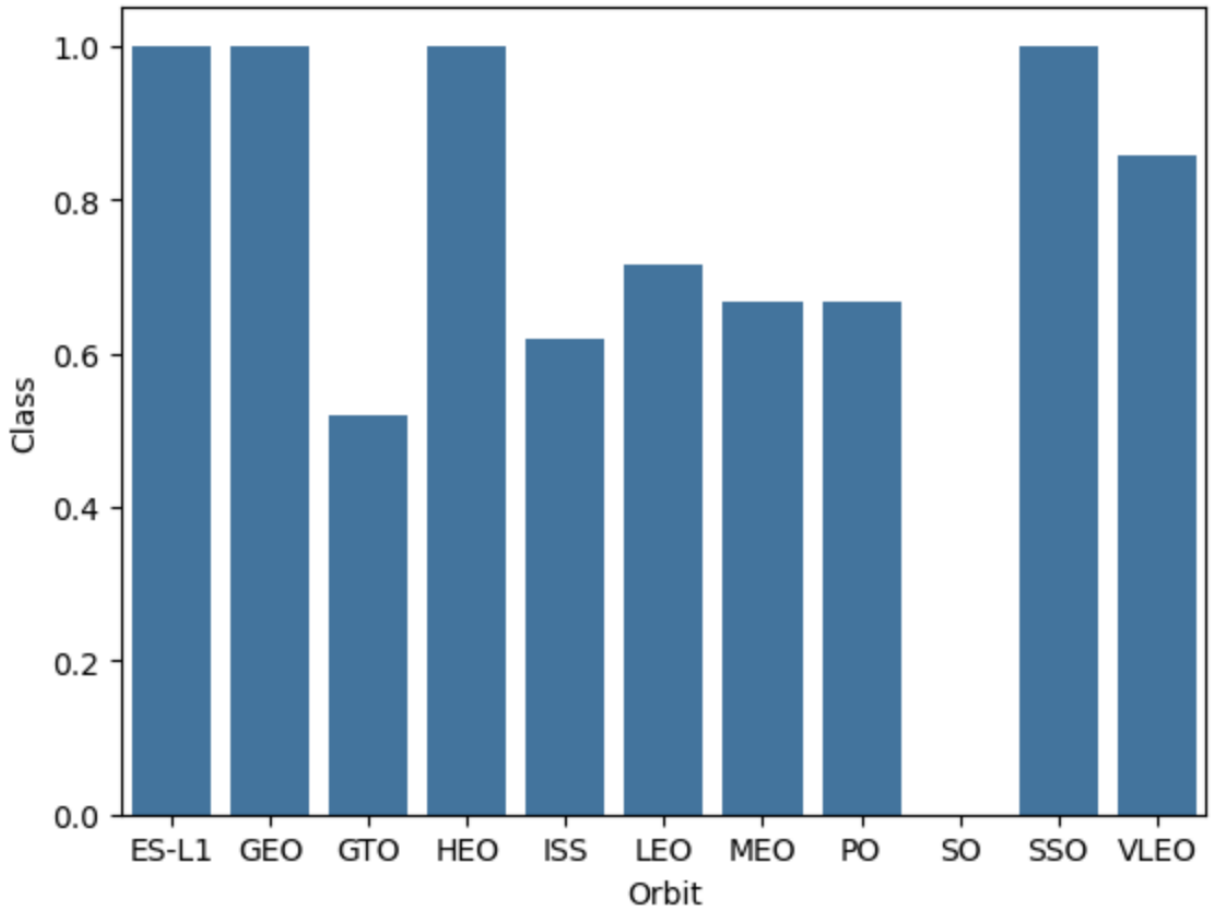


Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

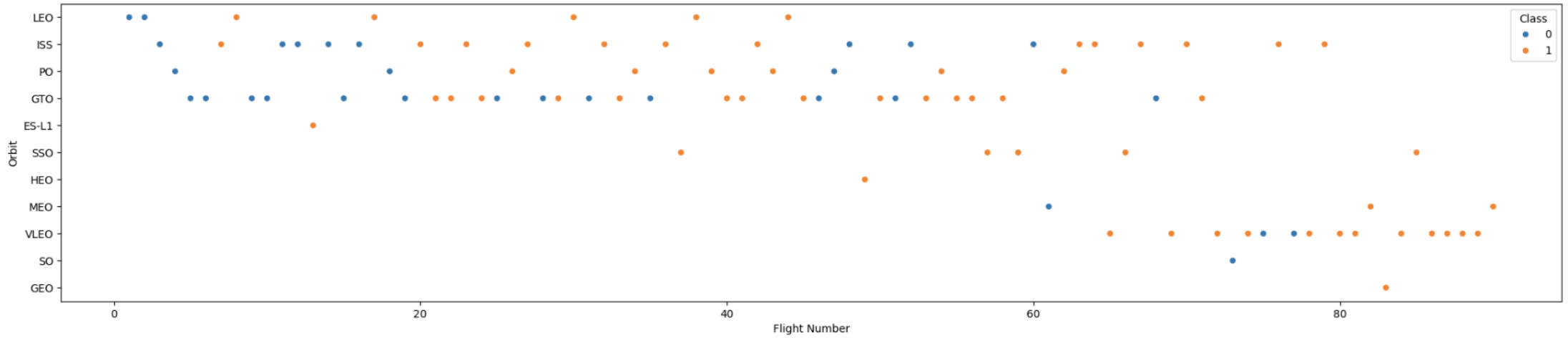
- Show a bar c
  success rate
  type

- Show the scr
  scatter plot v



Analyze the plotted bar chart to identify which orbits have the highest success rates.

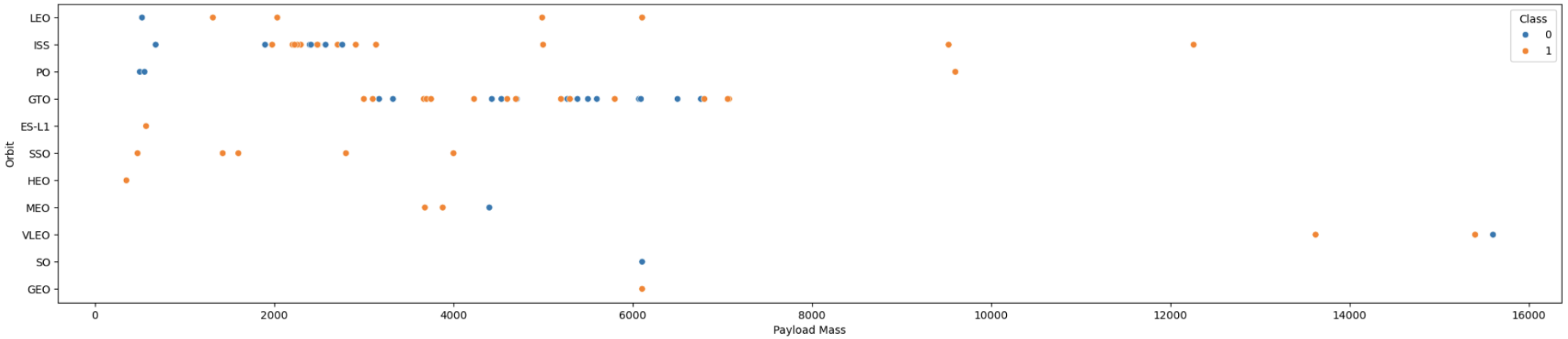
Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

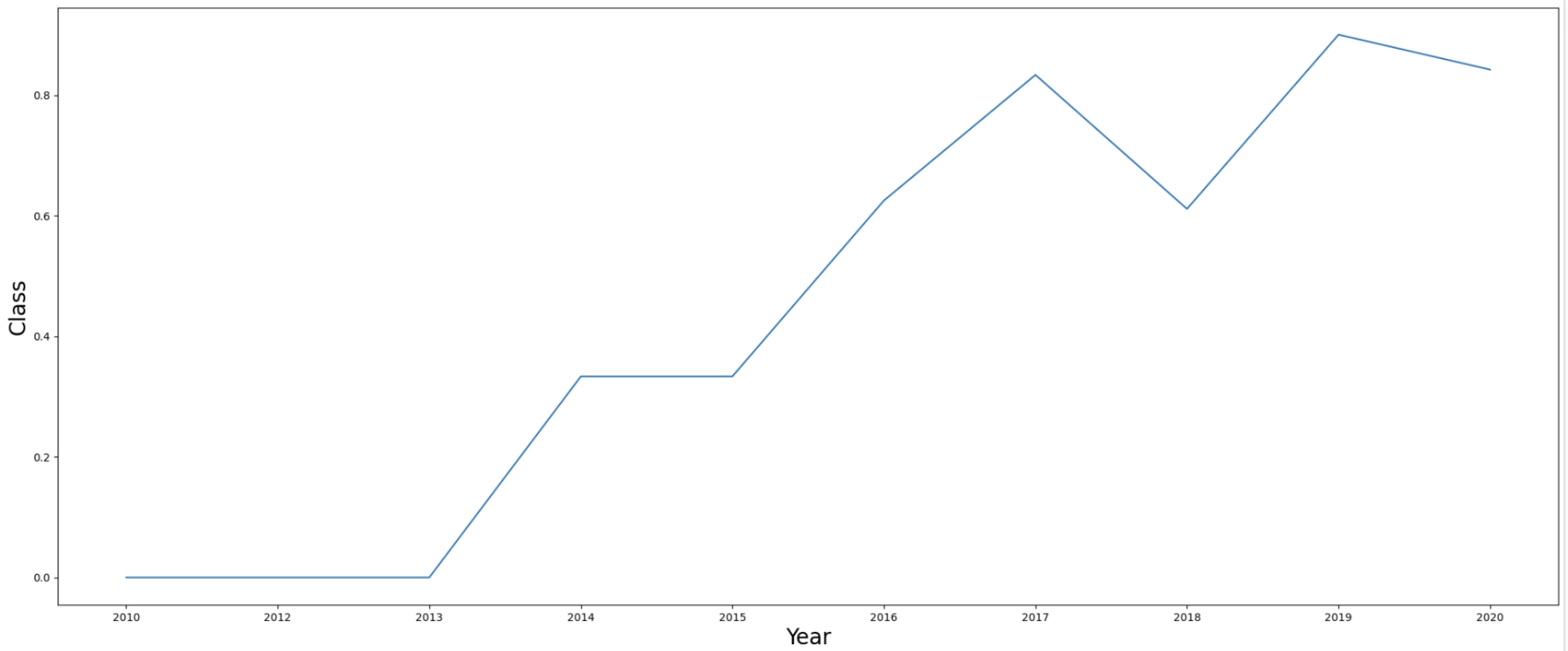
[32]:
```
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the class value
plt.figure(figsize=(25, 5))
sns.scatterplot(data=df, x="PayloadMass", y="Orbit", hue="Class")
plt.xlabel("Payload Mass")
plt.ylabel("Orbit")
plt.show()
```



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- 

- you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 20;
```

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

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total_Payload_Mass" FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Avarage_Payload_Mass" FROM SPACEXTABLE WHERE "Booster_Version" LIKE '%F9 v1.1%'
```

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

Avarage_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

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

First_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT "Mission_Outcome", COUNT(*) AS "Total_count" FROM SPACEXTABLE GROUP BY "Mission_Outcome";

3]:

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

Boosters Carried Maximum Payload

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

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 CASE substr(Date, 6, 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' END AS Month_Name, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome="Failure (drone ship)"

]:

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

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

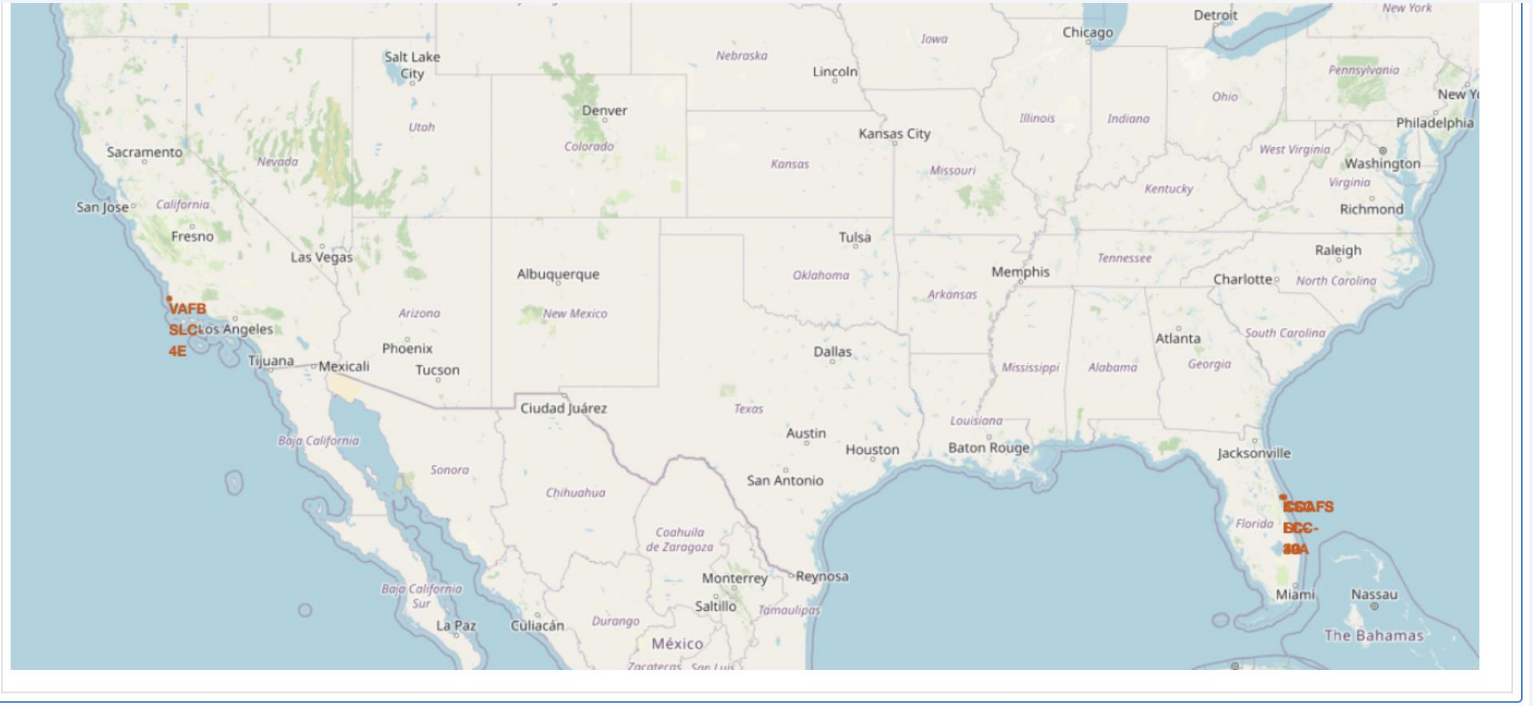
- %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

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

Launch Sites



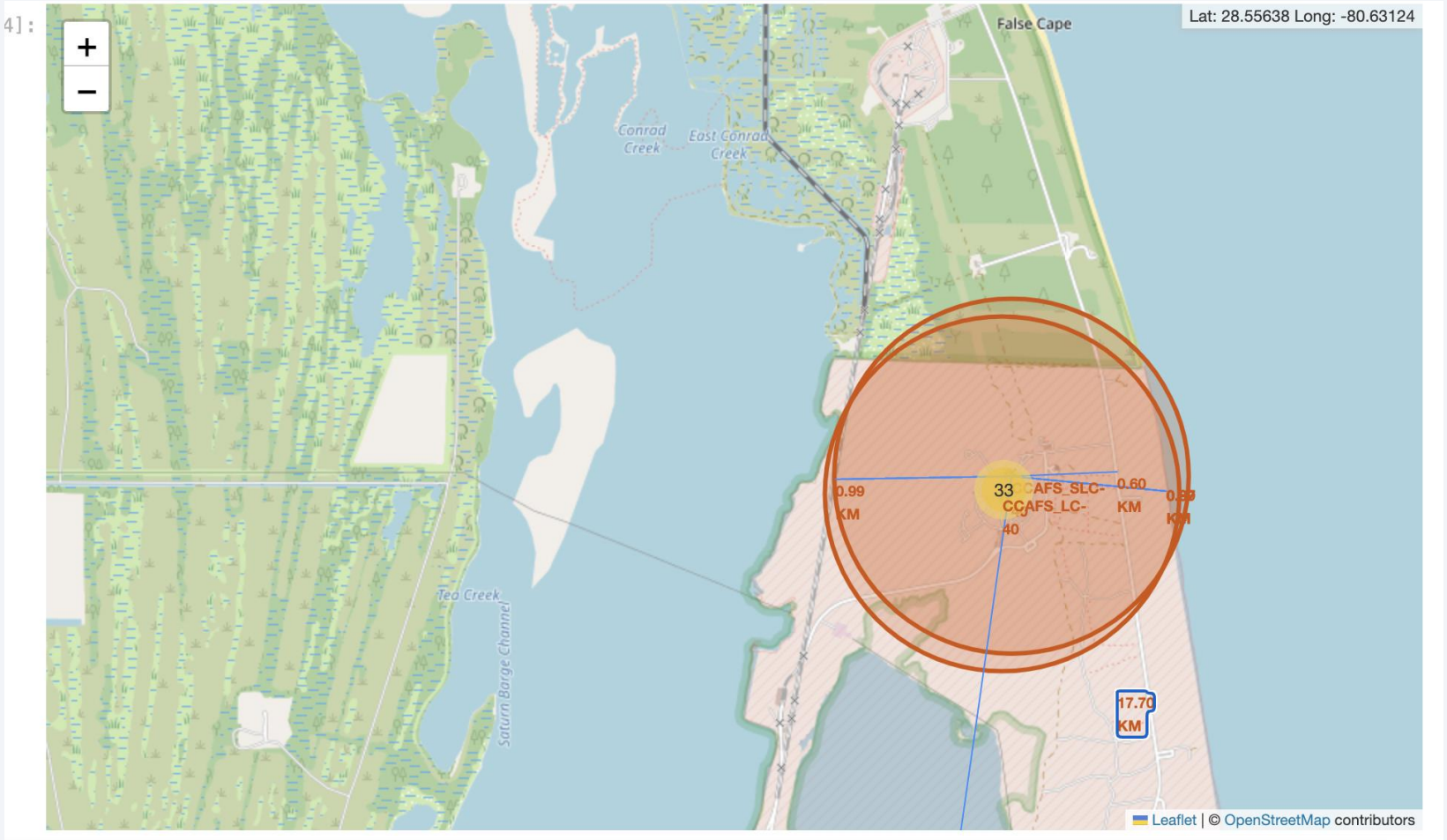
Now, you can explore the map by zoom-in/out the marked areas , and try to answer the following questions:

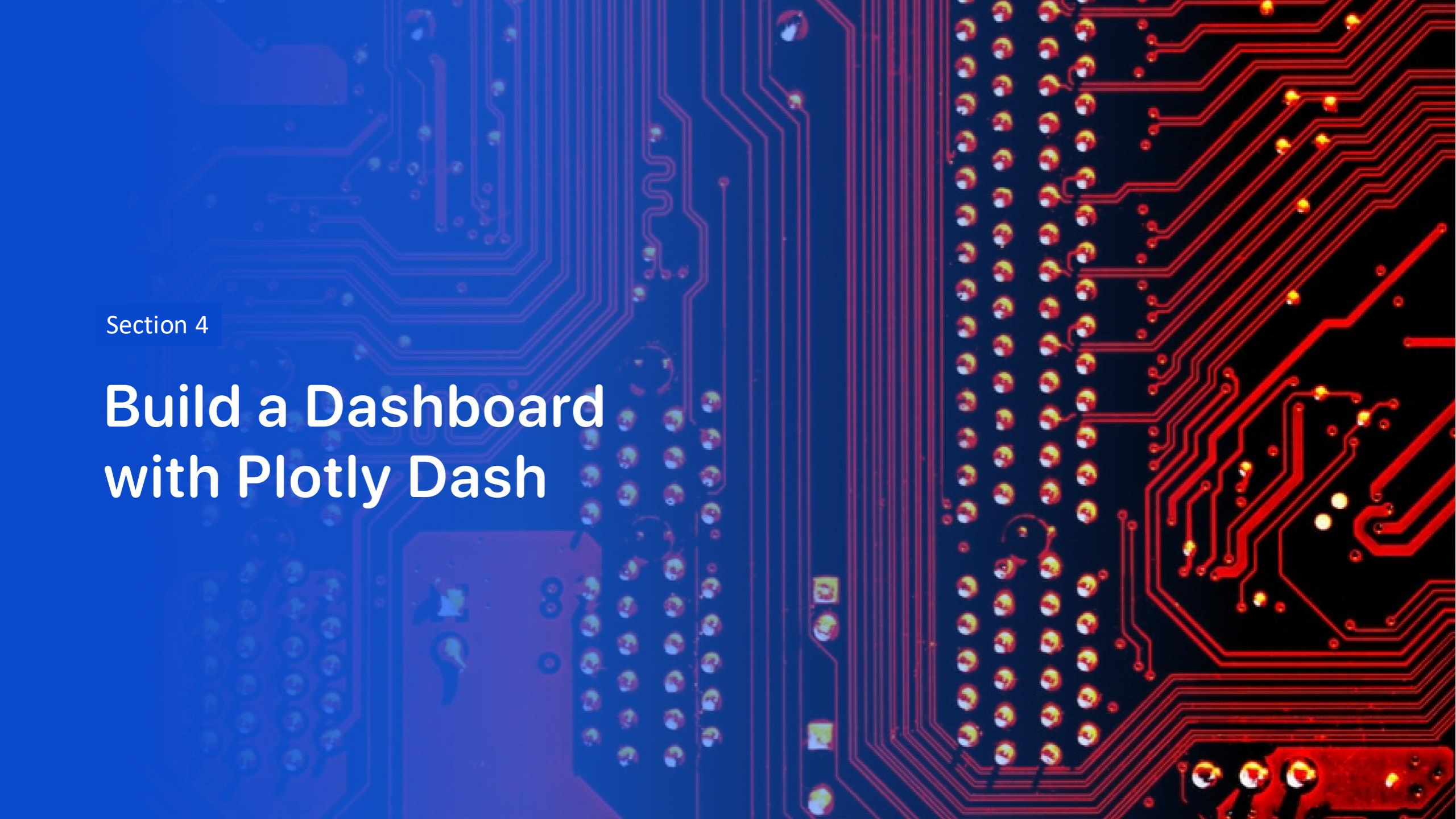
- Are all launch sites in proximity to the Equator line?
- Are all launch sites in very close proximity to the coast?

Also please try to explain your findings.

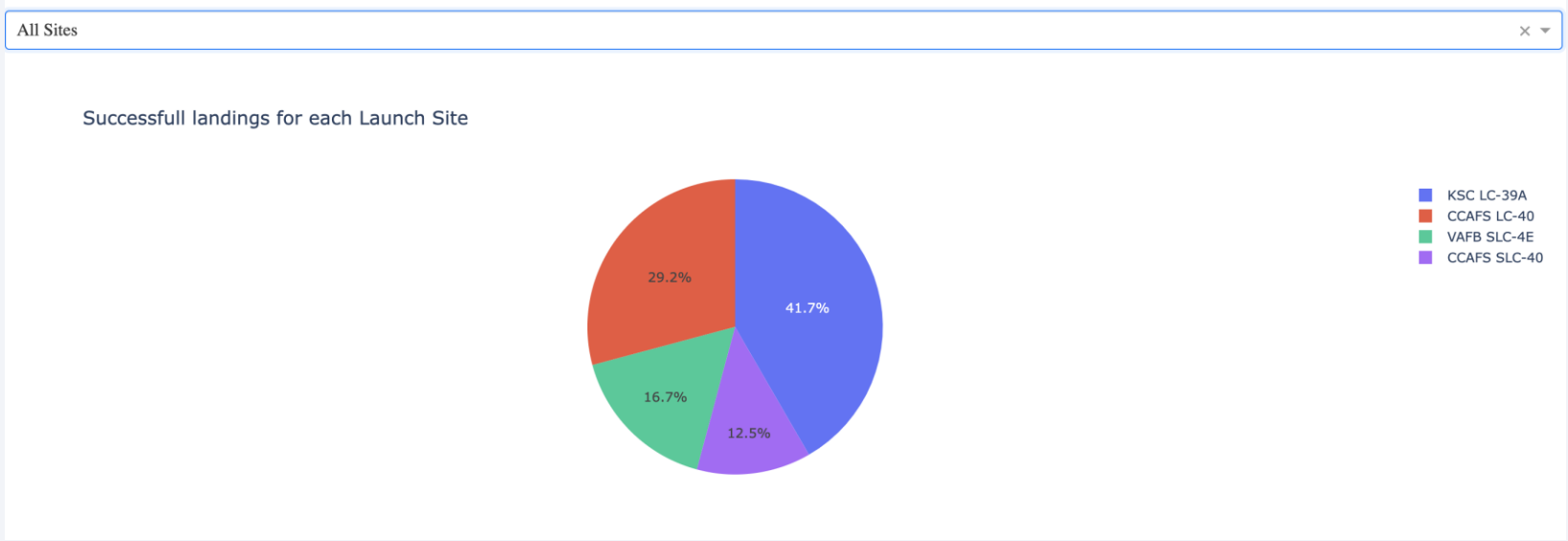
Coor-coded outcomes for CCAFS SALC-40

Proximities

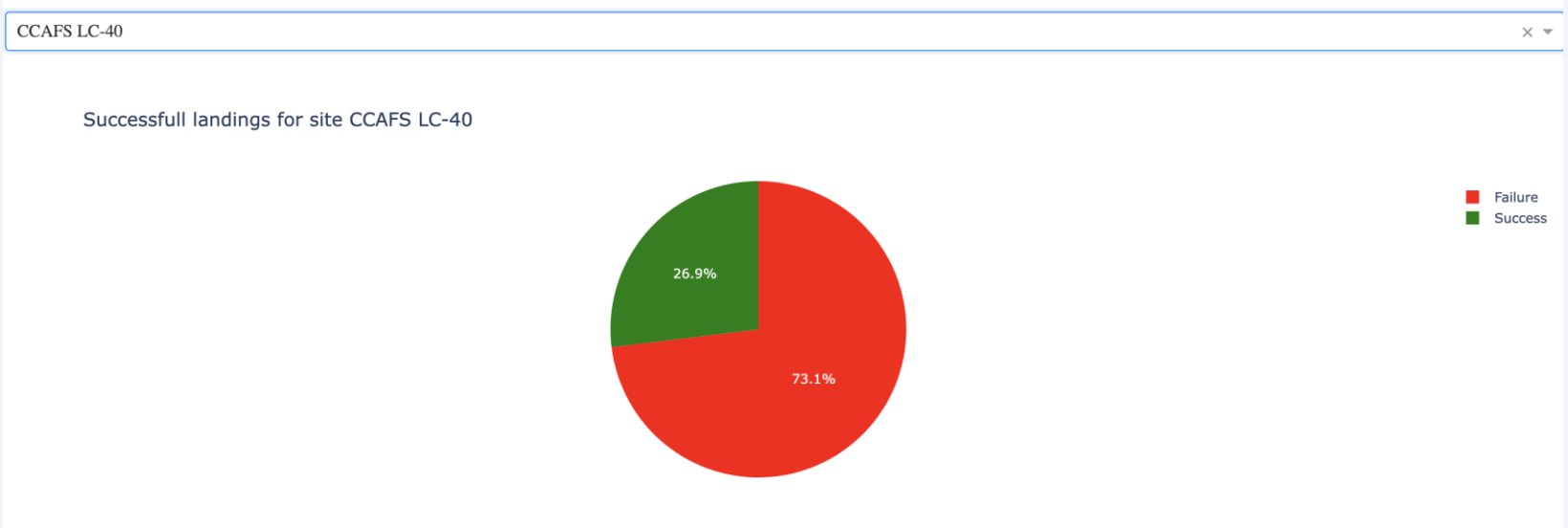



Section 4
Build a Dashboard with Plotly Dash

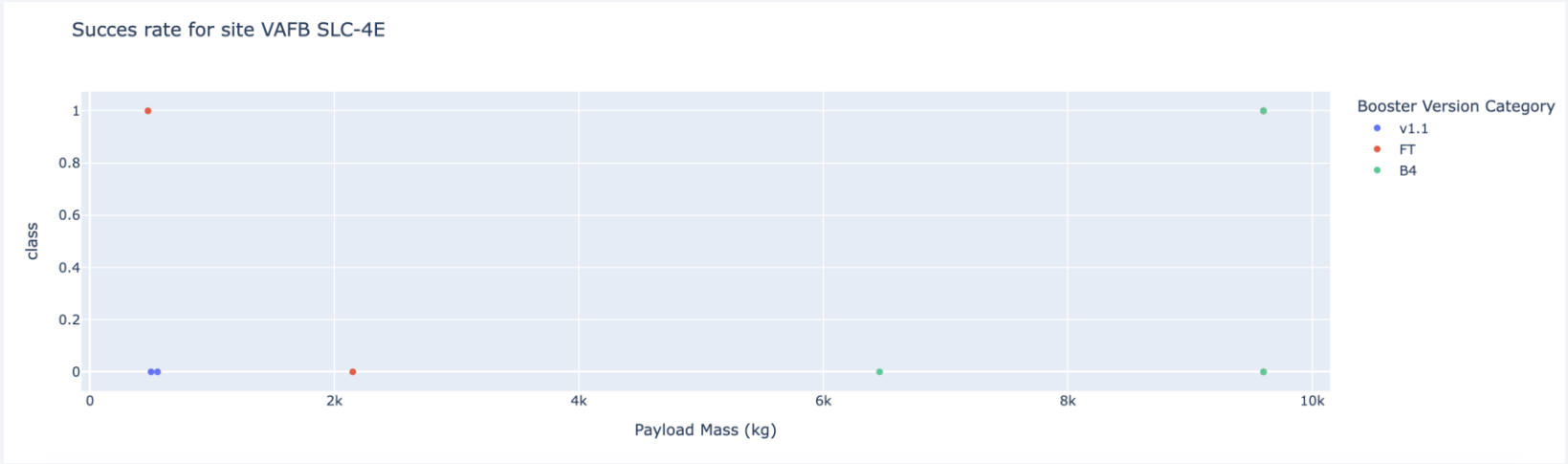
Launch success count for all sites



Launch site success and failed ratio



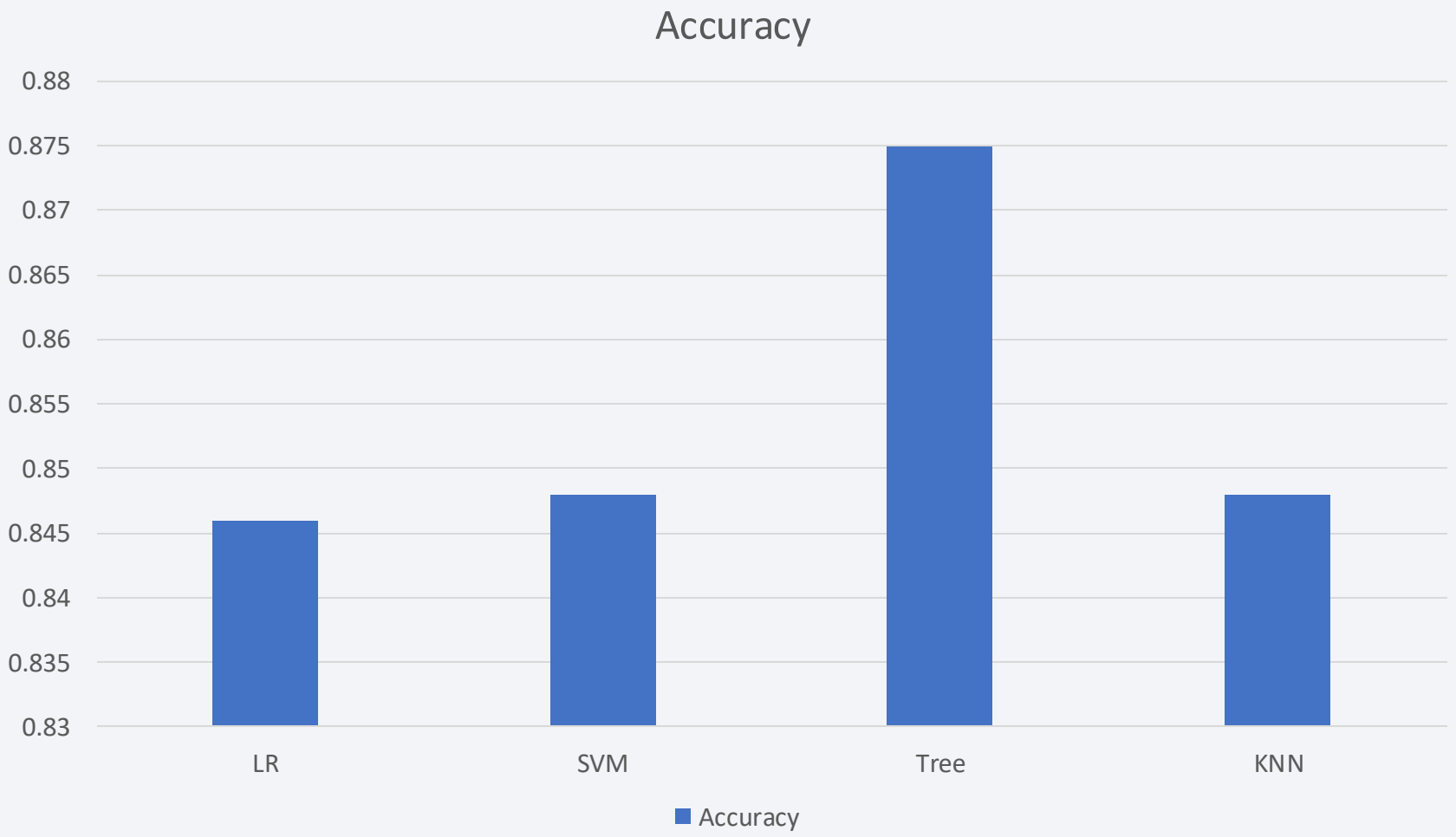
Payload vs. Launch Outcome



Section 5

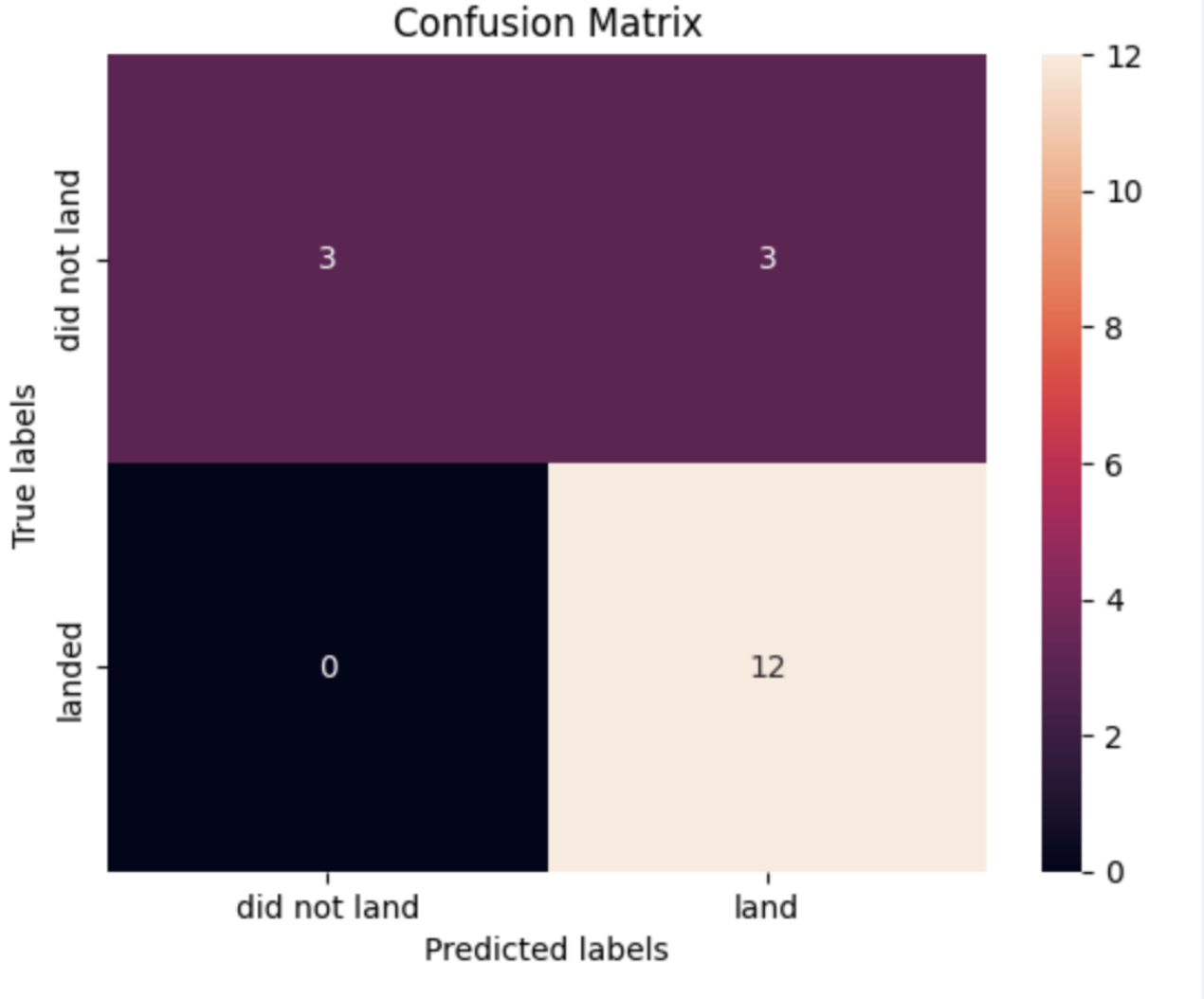
Predictive Analysis (Classification)

Classification Accuracy



The best performing model was decision tree

Confusion Matrix



Model was accurate predicting sucesfull landings, however it’s accuracy for failed lanings was law

Conclusions

- Time and learning curve had positive impact on outcome
- Launch sites and payload didn’t have impact on outcome
- Some launch sites were used only for lighter payloads
- GEO launches were limited
- More data would be needed for more accurate predictions

Thank you!