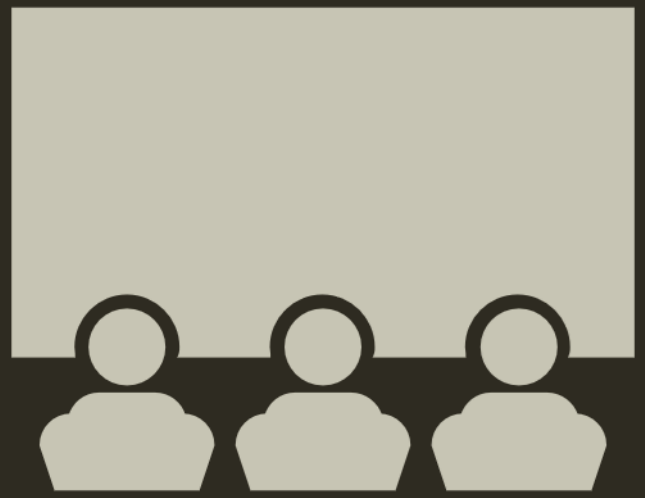
DATA SCIENCE CAPSTONE PROJECT

<NGIAM TEE JOHN>
<August 2021>

OUTLINE



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

EXECUTIVE SUMMARY

- This capstone project uses data from previous SpaceX launches and predicts future success launches from the data with machine learning models and visual aid.
- Visual aid such as graphs charts and maps are used for better illustration and communication to stakeholders.

INTRODUCTION

- Calculating the cost of a falcon 9 launch against its historical launches and its success rate using machine learning and charts to better predict future launches outcome
- How to determine whether future launches will result in a successful outcome

METHODOLOGY

- Data collection methodology:
- **Data were collected from SpaceX and Wikipedia**
- Perform data wrangling
- **Replacing missing data, one – hot encoding data and changing column variables**
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

METHODOLOGY

DATA COLLECTION – SPACEX API

Added a flowchart of SpaceX API calls here

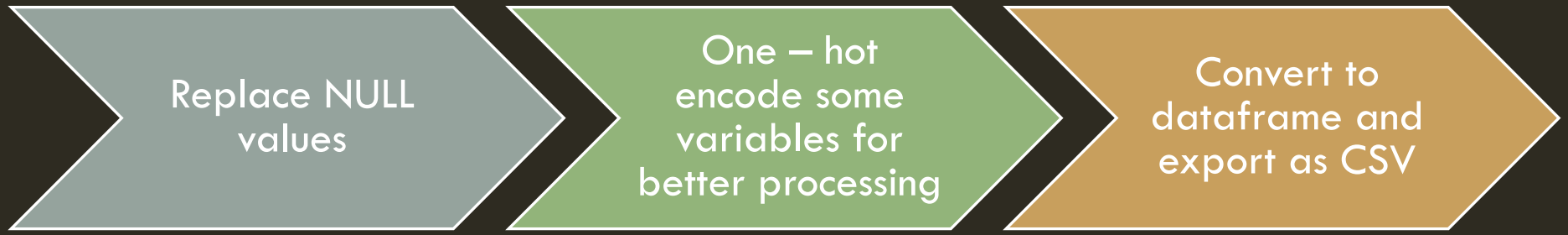
https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Data%20Collection.ipynb

DATA COLLECTION – WEB SCRAPING

Add a flowchart of web scraping here

https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Web%20Scrapping.ipynb

DATA WRANGLING



https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Data%20Wrangling.ipynb

EDA WITH DATA VISUALIZATION

Scatterplot is used to better visualize the correlation between two variables

Line chart is used to better visualize success rate with time

Bar chart is used to better visualize variables affecting success rate

https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Data%20Visualization.ipynb

EDA WITH SQL

SQL is used to better understand queries with specific conditions and to better understand the data used before using it on machine learning

https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20SQL.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

Circle marker is used to denote the launch site

Cluster marker is used to denote successful and unsuccessful launches on each launch site

Marker was added to visualize the distance between railroads, highways, and coastlines

https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Web%20Scrapping.ipynb

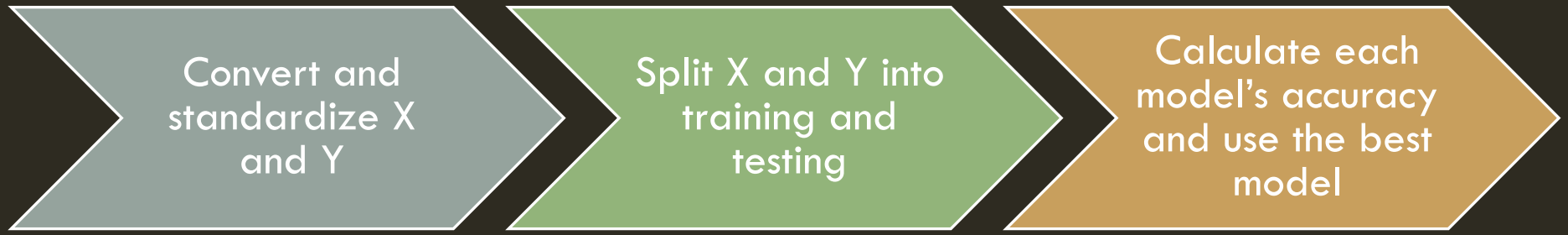
BUILD A DASHBOARD WITH PLOTLY DASH

Pie chart displays the successful outcomes and the unsuccessful outcomes for each and all launch sites

Scatter plot displays the correlation between payload mass and success rate for each and all launch sites

https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/spacex_dash_app.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

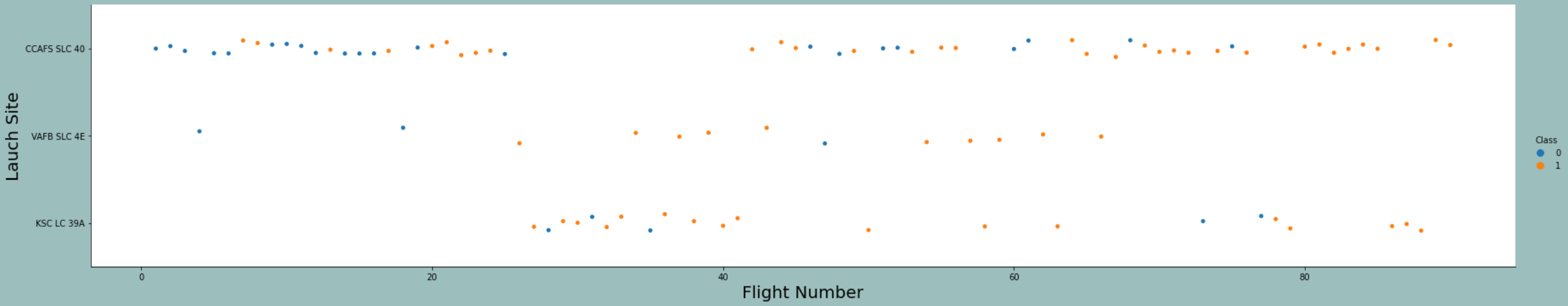


https://github.com/johnpro18/Coursera-Capstone-Project/blob/main/SpaceX%20_%20Machine%20Learning.ipynb

RESULTS

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

EDA WITH VISUALIZATION

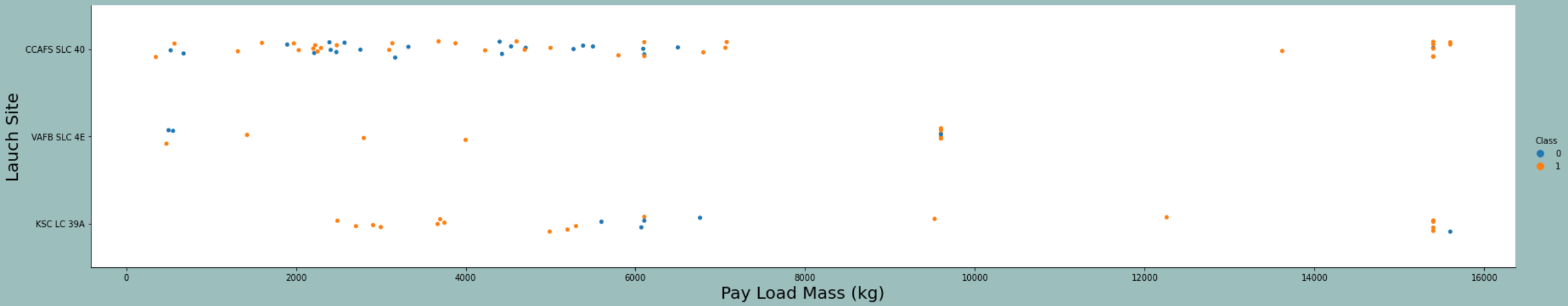


The most used launch site is CCAFS SLC-40 and the most recent launch is from the same launch site

FLIGHT NUMBER VS. LAUNCH SITE

Show a scatter plot of Flight Number vs. Launch Site

Show the screenshot of the scatter plot with explanations

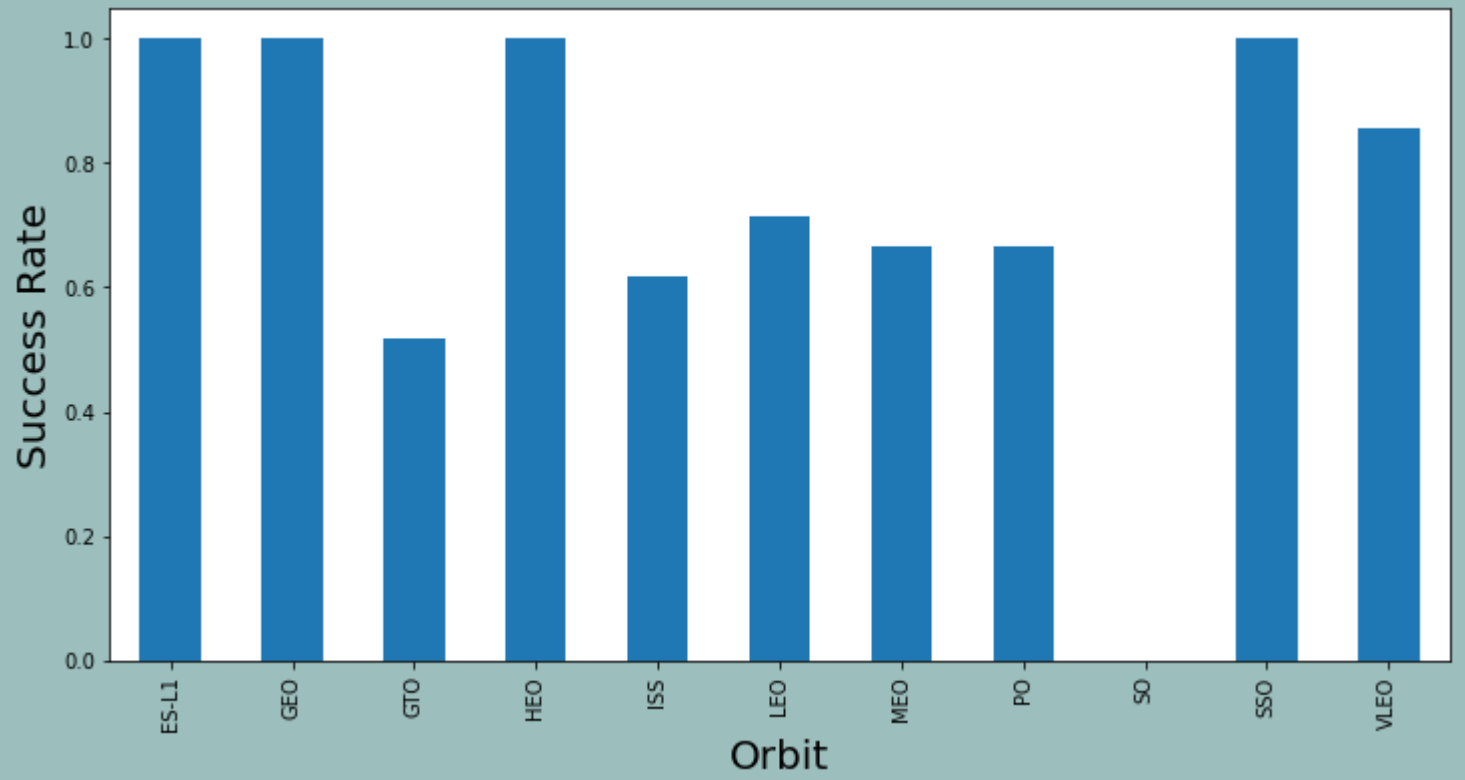


Launch site CCAFS SLC-40 is most used for normal payload and launch site CCAFS SLC-40 and KSC LC-39A are used for higher payload

PAYLOAD VS. LAUNCH SITE

Show a scatter plot of Payload vs. Launch Site

Show the screenshot of the scatter plot with explanations

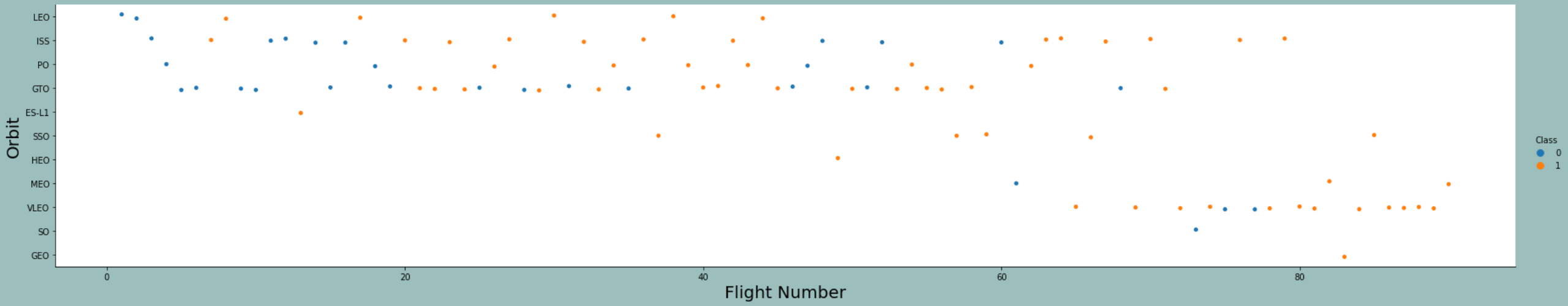


ES-L1, GEO, HEO, and SSO have the highest success rate

SUCCESS RATE VS. ORBIT TYPE

Show a barchart for the success rate of each orbit type

Show the screenshot of the scatter plot with explanations

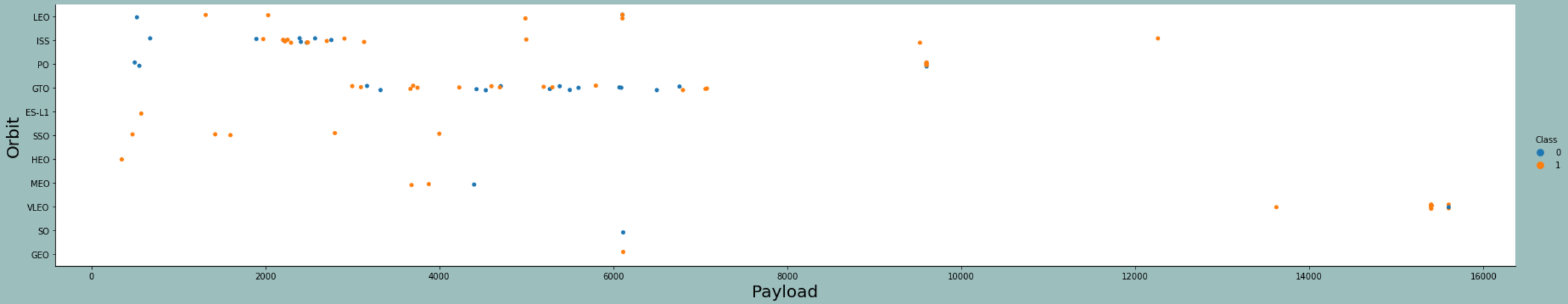


Earliest launches are above the GTO whereas recent launches are around VLEO

FLIGHT NUMBER VS. ORBIT TYPE

Show a scatter point of Flight number vs. Orbit type

Show the screenshot of the scatter plot with explanations

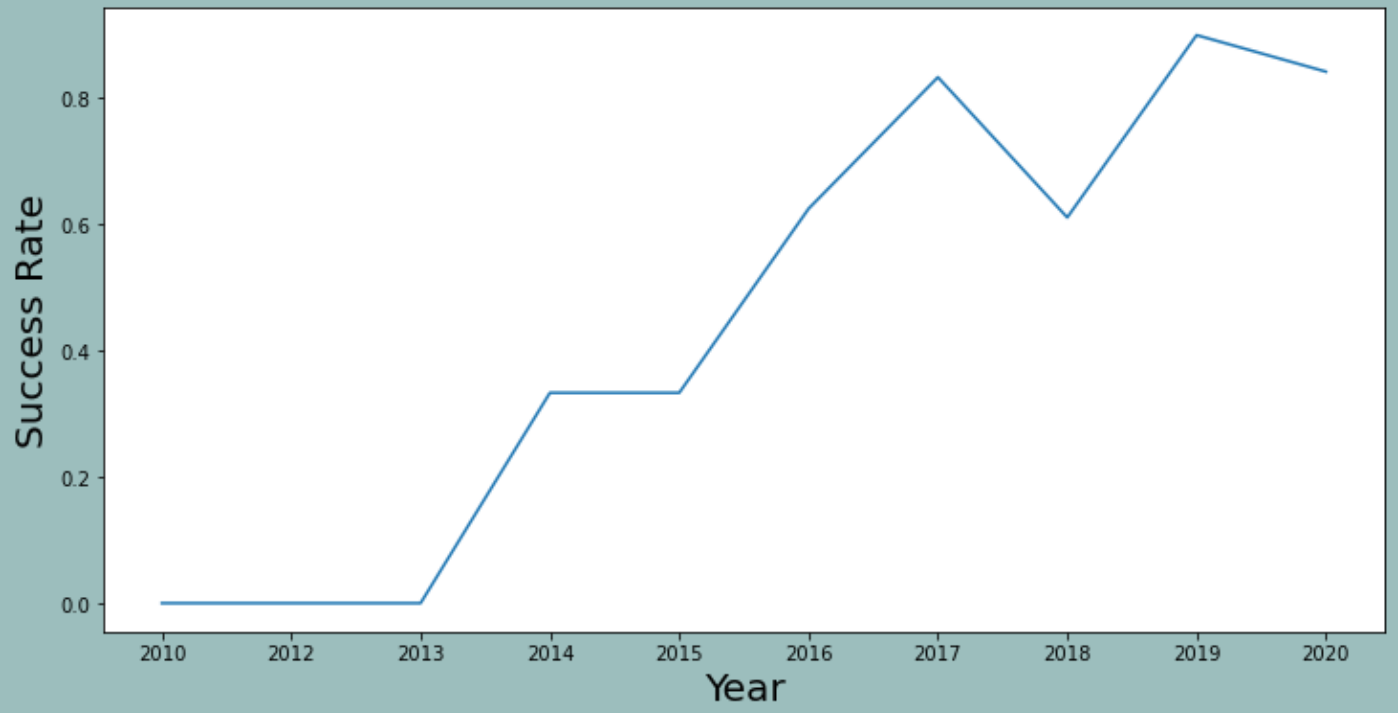


Lighter payload are launched towards GTO and ISS whereas heavier payload are launched towards VLEO

PAYLOAD VS. ORBIT TYPE

Show a scatter point of payload vs. orbit type

Show the screenshot of the scatter plot with explanations



Success rate increases from 2013 to 2017 and dipped slightly and continue to increase towards 2020

LAUNCH SUCCESS YEARLY TREND

Show a line chart of yearly average success rate

Show the screenshot of the scatter plot with explanations

EDA WITH SQL

ALL LAUNCH SITE NAMES

Find the names of the unique launch sites

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

Present your query result with a short explanation here

Using SELECT Statement and Distinct Statement, the launch site names were displayed

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

LAUNCH SITE NAMES BEGIN WITH `CCA`

Find all launch sites begin with `CCA`

CCAFS LC-40

Present your query result with a short explanation here

Using SELECT Statement with WHERE Condition '%CCA%' results were displayed

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL WHERE LAUNCH_SITE LIKE '%CCA%'
```

TOTAL PAYLOAD MASS

Calculate the total payload carried by boosters from NASA

45596

Present your query result with a short explanation here

Using SELECT and SUM STATEMENT the total payload mass was displayed

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

AVERAGE PAYLOAD MASS BY F9 V1.1

Calculate the average payload mass carried by booster version F9 v1.1

2928

Present your query result with a short explanation here

Using SELECT and AVG STATEMENT the average payload mass by F9 v1.1 is displayed

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

FIRST SUCCESSFUL GROUND LANDING DATE

Find the date when the first successful landing outcome in ground pad

2010-06-04

Present your query result with a short explanation here

Using SELECT and MIN STATEMENT the first successful ground landing date was displayed

```
%sql SELECT MIN(DATE) AS DATE FROM SPACEXTBL WHERE MISSION_OUTCOME = 'Success'
```

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

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

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

Present your query result with a short explanation here

Using SELECT and BETWEEN STATEMENT the booster versions with successful drone ship landing were displayed

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE '%Success (drone ship)%' AND PAYLOAD_MASS__KG_ BETWEEN 40
00 AND 6000
```

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

Calculate the total number of successful and failure mission outcomes

100 Successful Missions and 1 Failure Mission

Present your query result with a short explanation here

Using SELECT and COUNT STATEMENT to display total number of successful and failure missions

```
%sql SELECT COUNT(MISSION_OUTCOME) AS SUCCESS FROM SPACEXTBL WHERE MISSION_OUTCOME = 'Success'
```

```
%sql SELECT COUNT(MISSION_OUTCOME) AS FAILURE FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Failure%'
```

BOOSTERS CARRIED MAXIMUM PAYLOAD

List the names of the booster which have carried the maximum payload mass

Present your query result with a short explanation here

```
%sql SELECT DISTINCT(BOOSTER_VERSION)FROM SPACEXTBL WHERE (SELECT DISTINCT(MAX(PAYLOAD_MASS__KG_)) FROM SPACEXTBL)
```

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

1	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)

Present your query result with a short explanation here

Using SELEECT and WHERE STATEMENT to display landing outcome with failure (drone ship)

```
%%sql SELECT MONTHNAME(DATE), BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXTBL
WHERE YEAR(DATE) = 2015 AND LANDING__OUTCOME LIKE 'Failure (drone ship)'
```

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

Rank the count of successful landing_outcomes

between the date 2010-06-04 and 2017-03-20

in descending order.

Present your query result with a short explanation here

landing__outcome	DATE
Success (ground pad)	2017-02-19
Success (drone ship)	2017-01-14
Success (drone ship)	2016-08-14
Success (ground pad)	2016-07-18
Success (drone ship)	2016-05-27
Success (drone ship)	2016-05-06
Success (drone ship)	2016-04-08
Success (ground pad)	2015-12-22

Using SELECT and ORDER BY DESC STATEMENT to display successful landing outcome in descending order between 2010-06-04 and 2017-03-20

```
%%sql SELECT LANDING__OUTCOME, DATE FROM SPACEXTBL
WHERE LANDING__OUTCOME LIKE '%Success%' AND (DATE BETWEEN '2010-06-04' AND '2017-03-20') ORDER BY DATE DESC
```

INTERACTIVE MAP WITH FOLIUM

LAUNCH SITES LOCATION

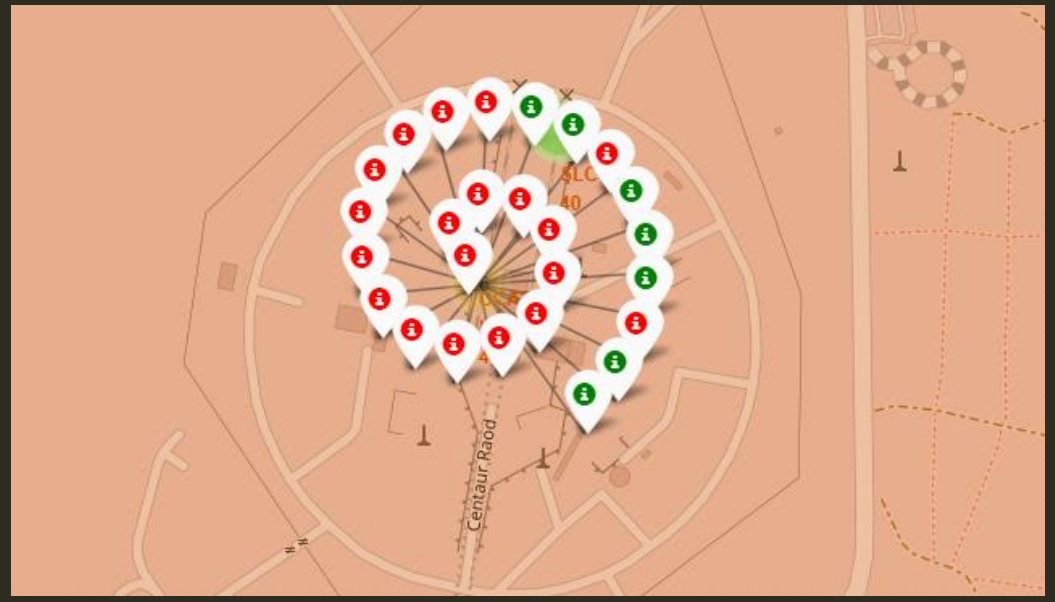
Explain the important elements and findings on the screenshot

Launch sites are positioned near the equator and the coast line

LAUNCH SITE OUTCOMES

Explain the important elements and findings on the screenshot

The folium map shows all the successful and unsuccessful launches at each launch site



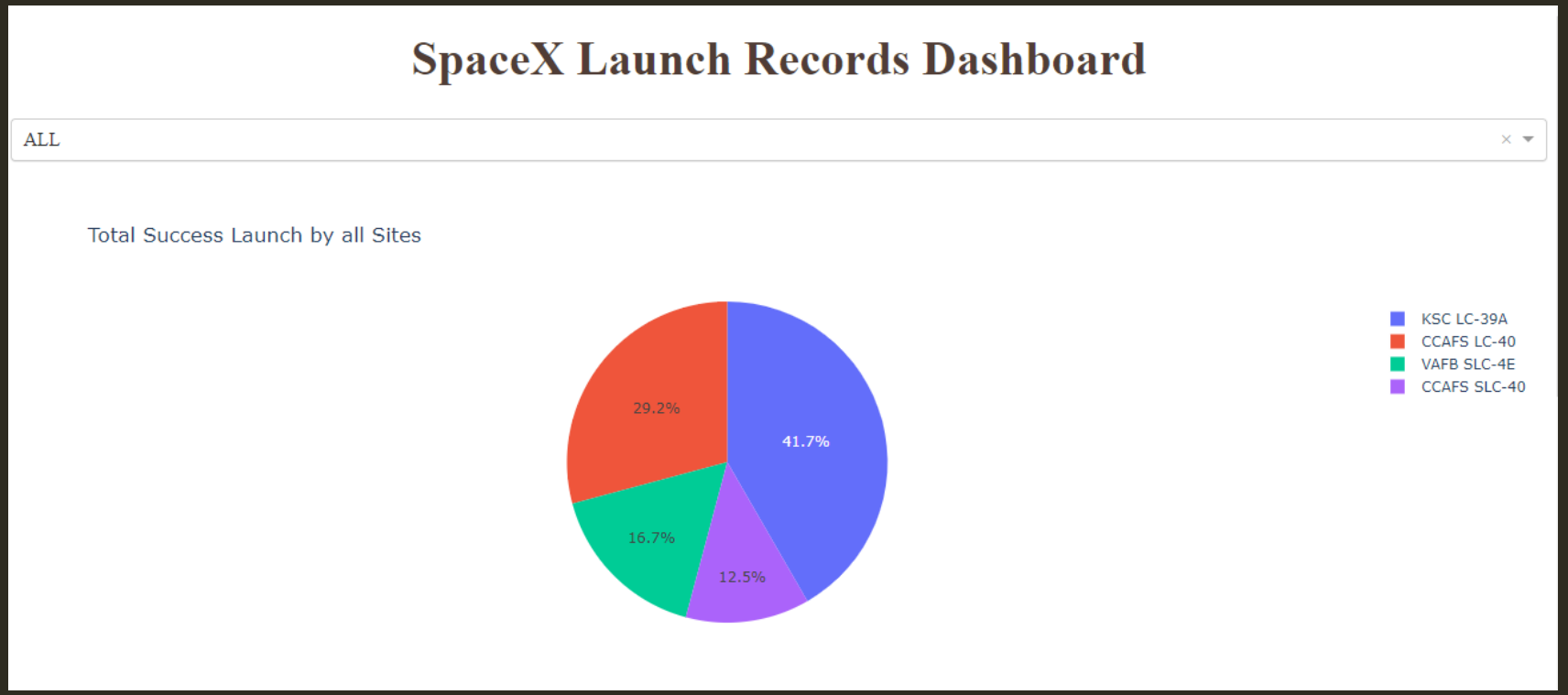
CLOSEST RAILWAY TO LAUNCH SITE

Explain the important elements and findings on the screenshot

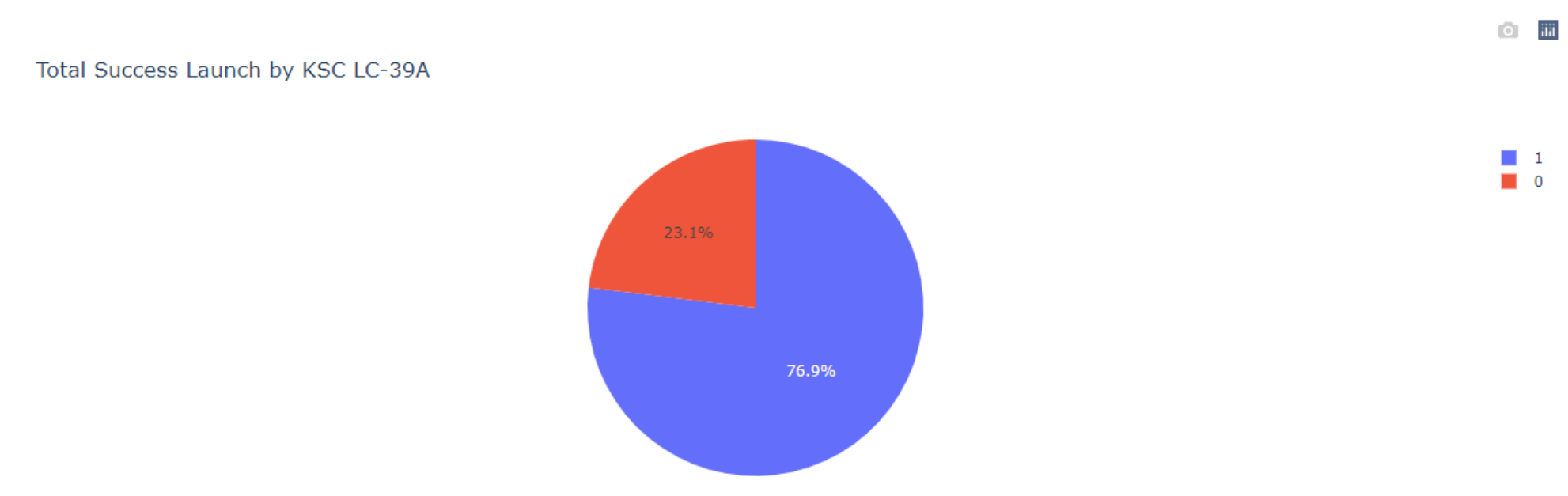
The folium map shows the closest railway for transportation of rocket parts and personnel

BUILD A DASHBOARD WITH PLOTLY DASH

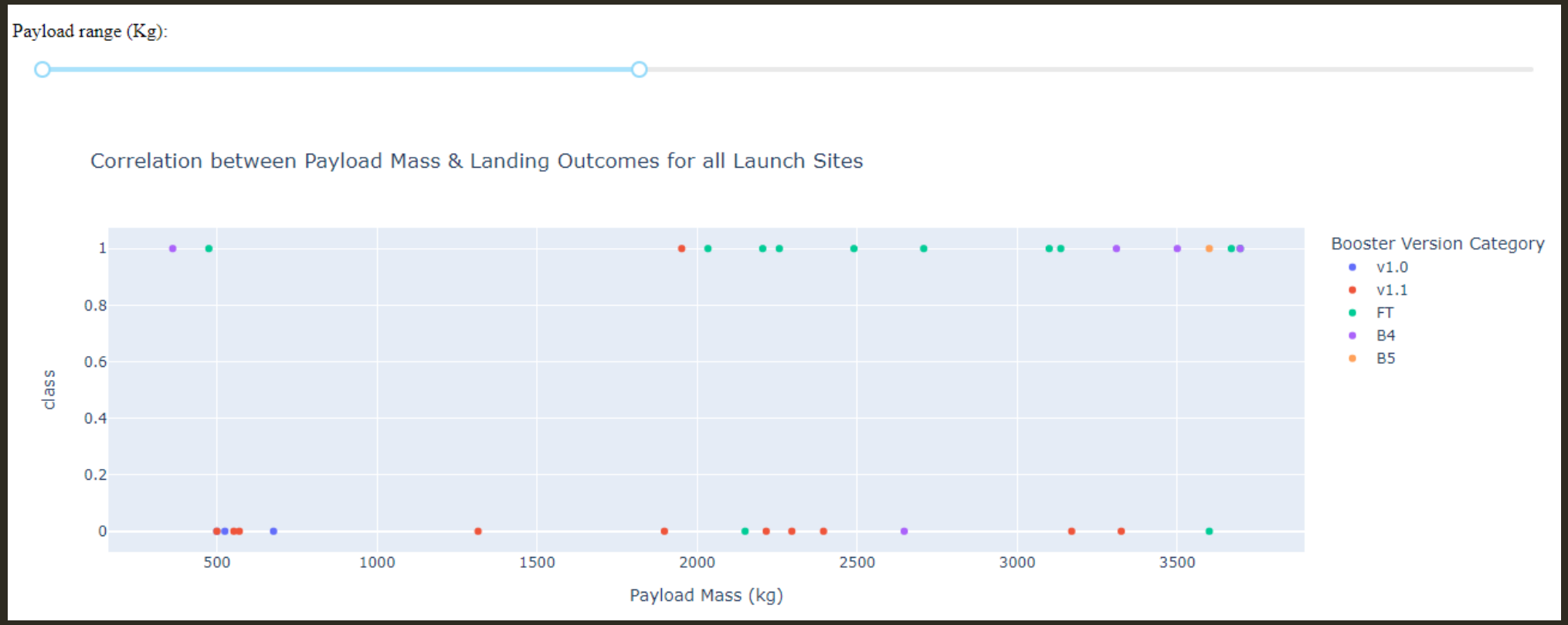
DASHBOARD LANDING SITE

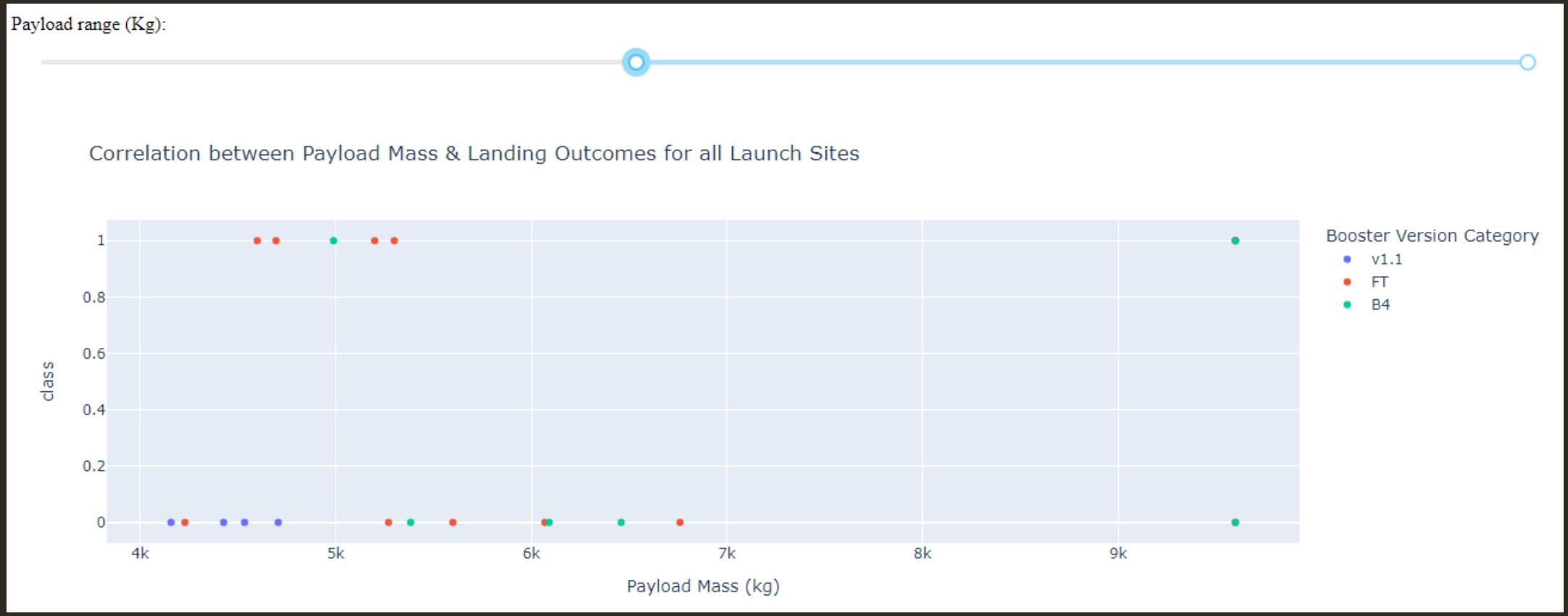


HIGHEST SUCCESS RATE FOR A LAUNCH SITE



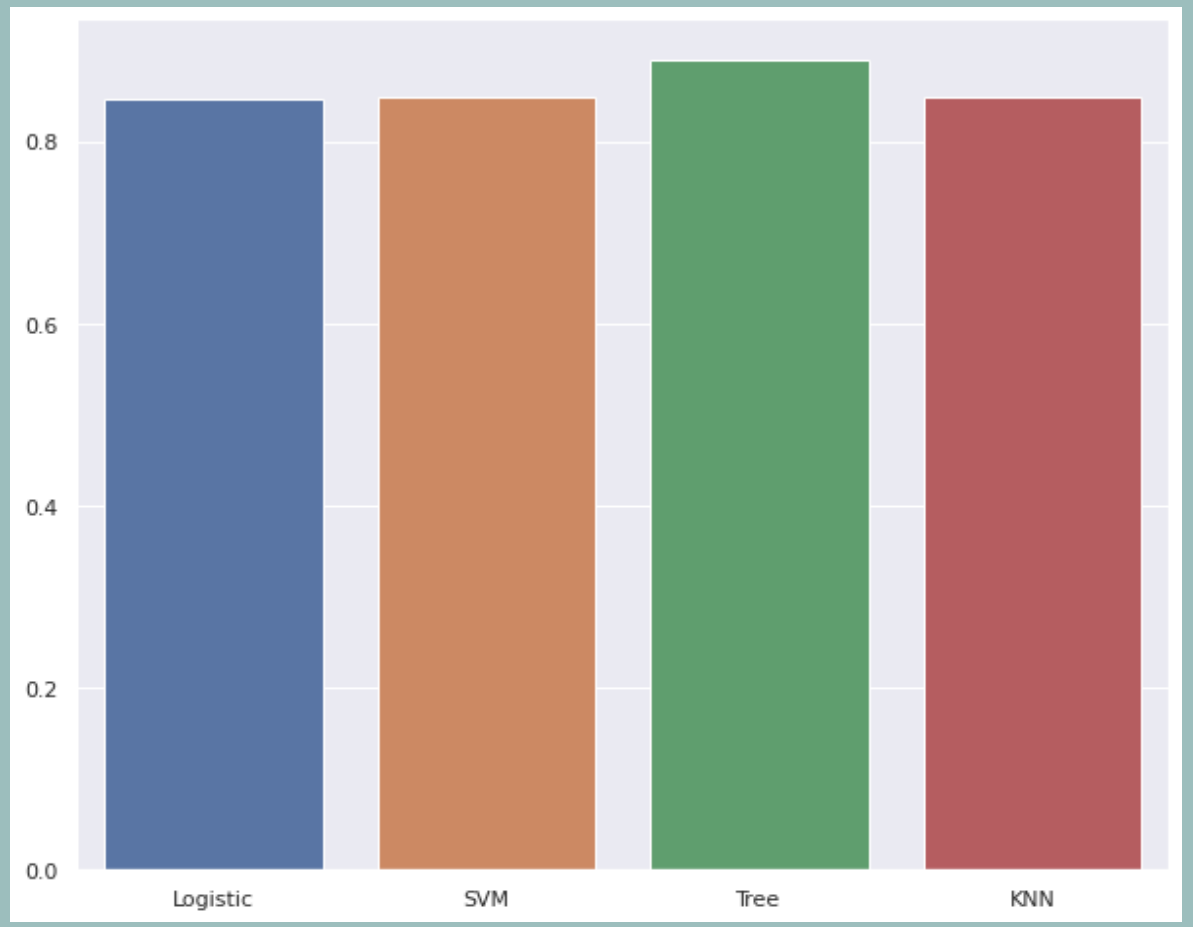
CORRELATION PAYLOAD & OUTCOME FOR ALL SITES





B4 boosters are used more often for higher payload mass

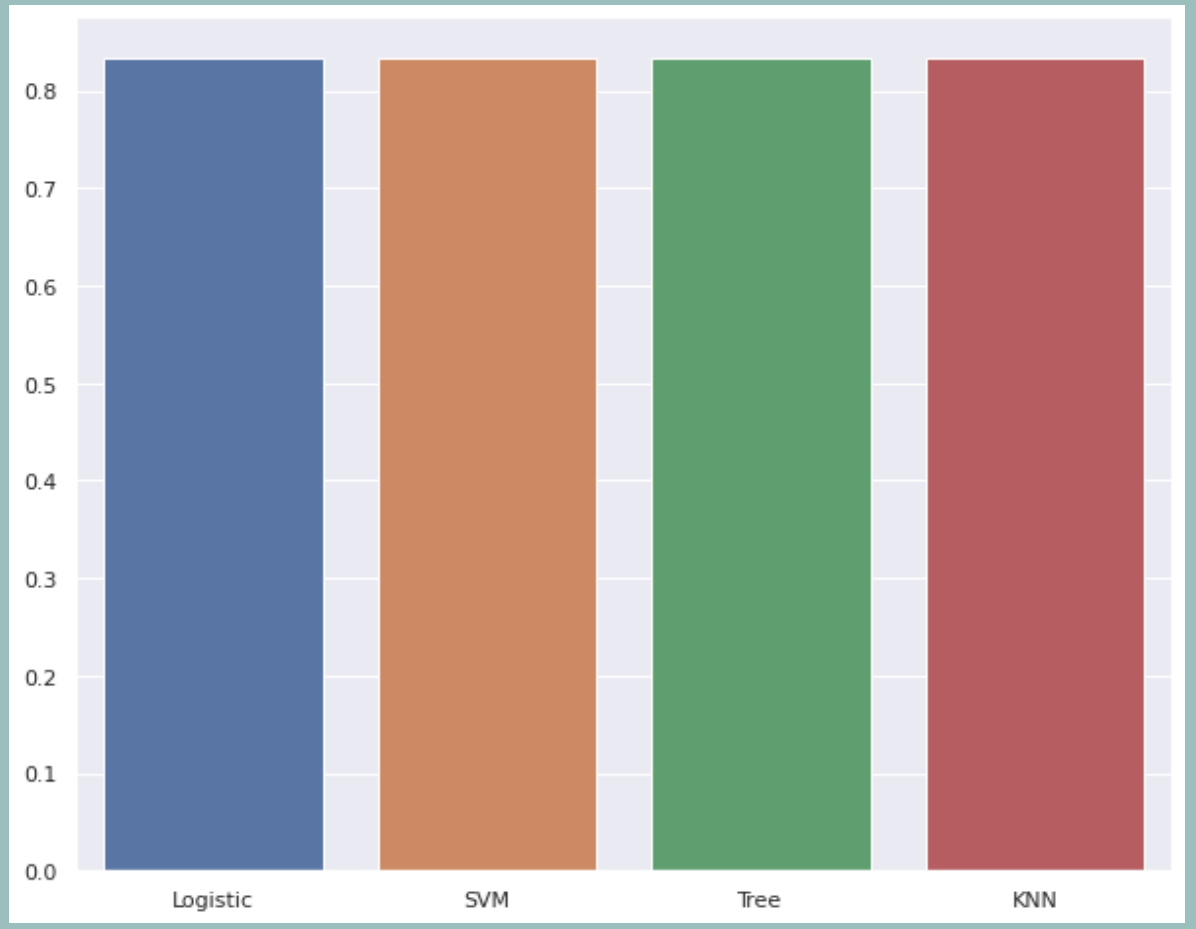
PREDICTIVE ANALYSIS (CLASSIFICATION)



TRAIN CLASSIFICATION ACCURACY

Visualize all the built model accuracy for all built models, in a barchart

Find which model has the highest classification accuracy



TEST CLASSIFICATION ACCURACY

SINCE TEST CLASSIFICATION HAS EQUAL ACCURACY TREE TRAIN CLASSIFICATION HAS HIGHER ACCURACY WHICH WILL BE USED

CONFUSION MATRIX

There were 3 correct predictions of did not land outcomes, 3 incorrect predictions of landed outcomes and 12 correct predictions of landed outcomes

CONCLUSION

All launch sites have equally close success rate however the increase of success rate may be due to the increase of payload mass.

Tree model is used as the best model as it has a accuracy of 83.33% which will be more accurate as more data is collected and tested.

APPENDIX

Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project