IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

LACENE Dihia
2/1/2024

Outline

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

Executive Summary

- Summary of methodologies
    1. Data Collection through API and through Web Scrapping
    2. Data Wrangling
    3. Exploratory Data Analysis (EDA) with SQL
    4. Exploratory Data Analysis (EDA) with Data Visualization
    5. Visual Analytics with Folium
    6. Plotly Dash Dashboard
    7. Predictive Analysis (Classification)
- Summary of all results

    Conclusion

Introduction

- Background:

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

- The Challanges :

1.How does Playoad mass,Launch site,Number of flights,and orbits affect first-stage landing success

2.Rate of successful landing over time

3.Best predictive model for succesfful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data Collection done through the use of SpaceX REST API and Web Scrapping techniques
- Perform data wrangling
  - By Filtering the data ,handling missing values and applying one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - By building 4 different classification models to predict landing outcome
  - Tune and evaluate the accuracy of each model to find the best model and parameters combination

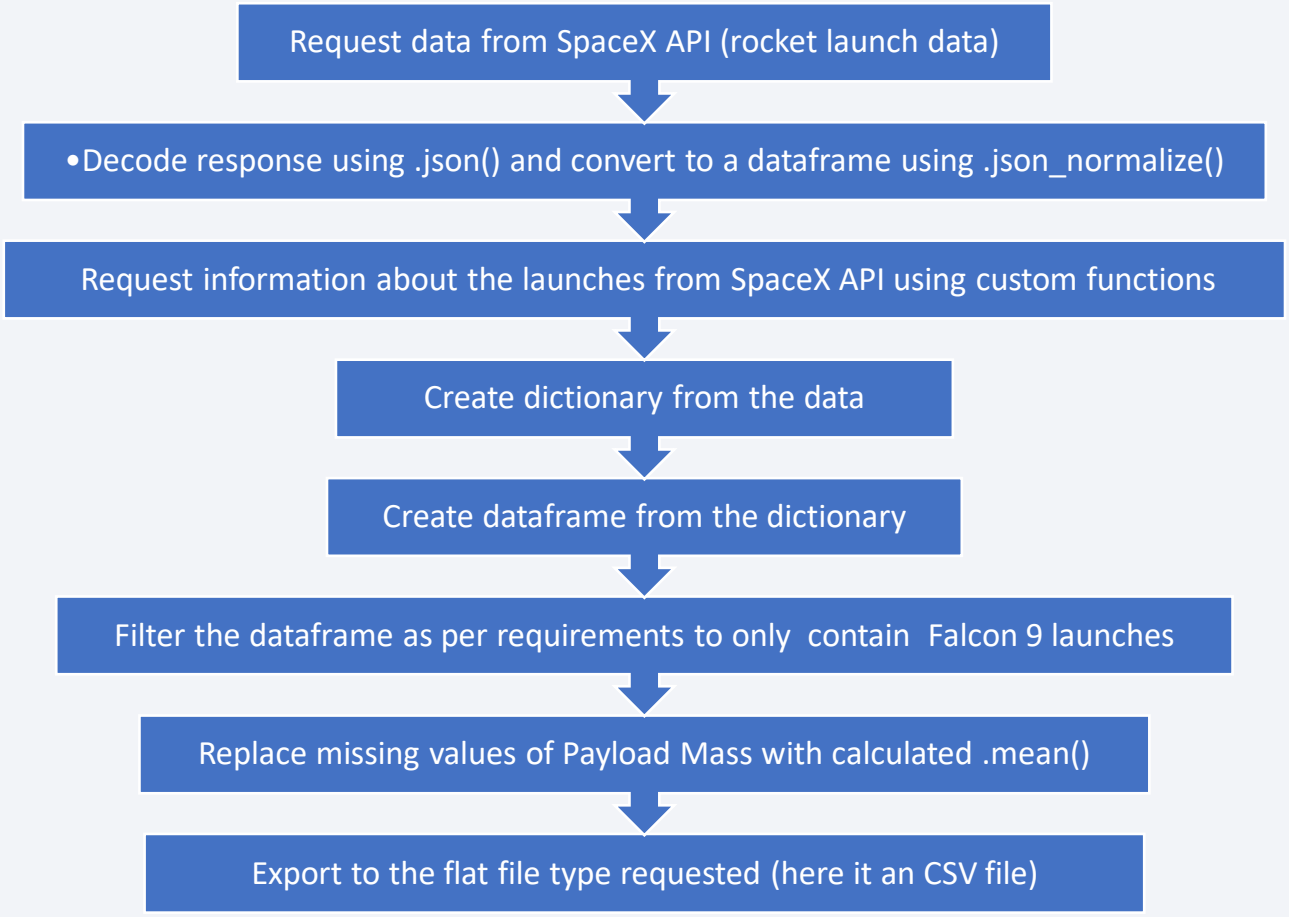
Data Collection

1. API SpaceX API (rocket launch data)

2. **Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia** « List of Flacon 9 and Falcon Heavy Launches » https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

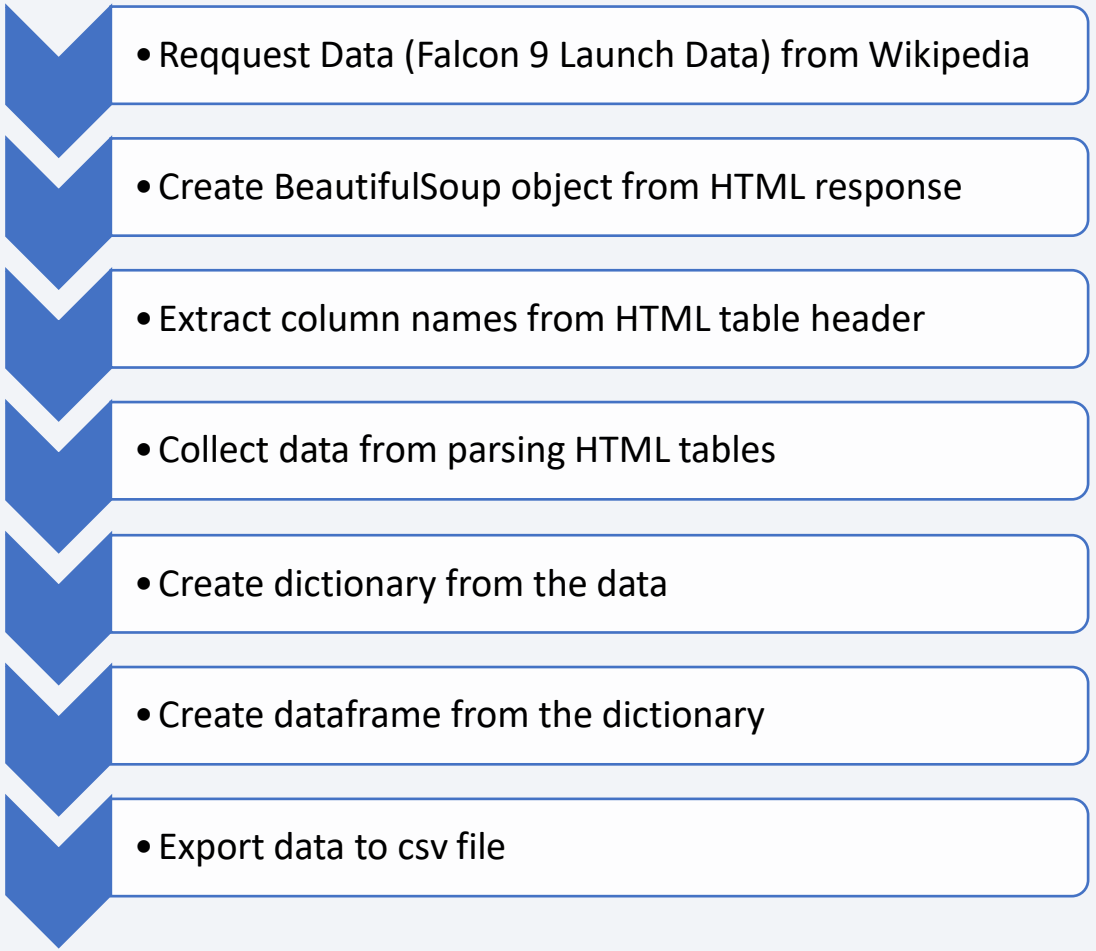
https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/01%20Data%20Collection%20spacex-data%20api.ipynb



Data Collection - Scraping

- **Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia**
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

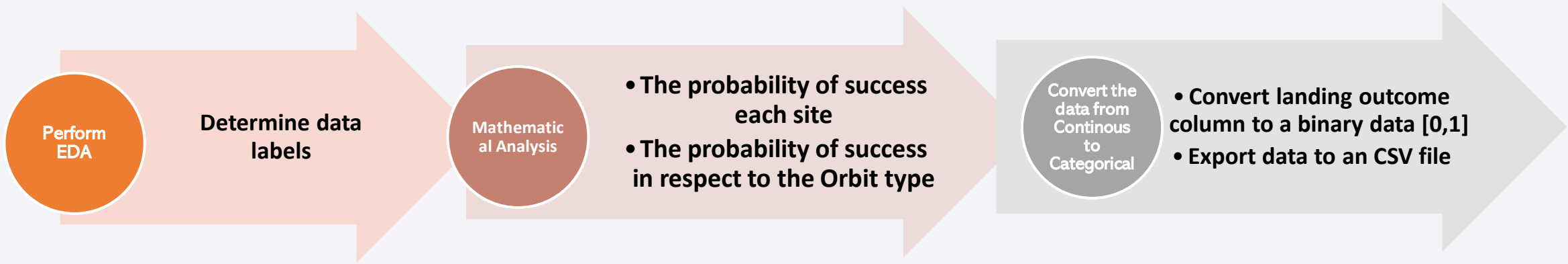
https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/02%20Webscraping.ipynb



Data Wrangling

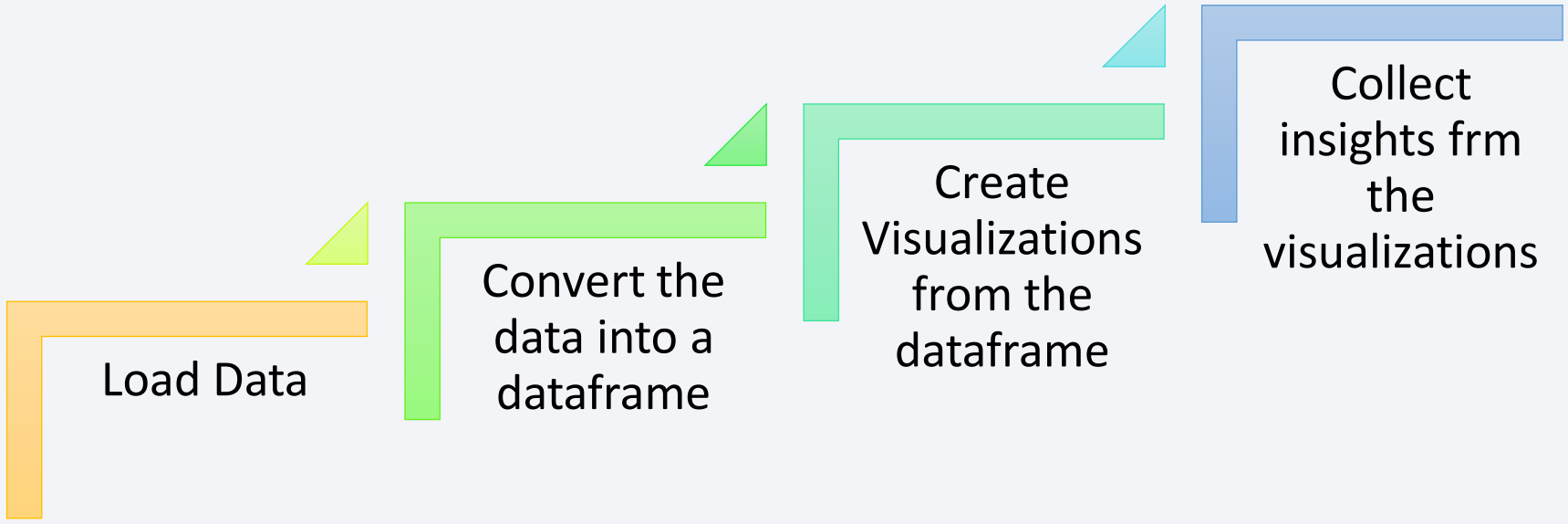
- **It is the process of cleaning and unifying messy and complex data sets for access and analysis.**

https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/03%20Data%20wrangling%20Spacex.ipynb



EDA with Data Visualization

- Exploratory Data Analysis is an approach of analyzing data sets to summarize their main characteristics ,using statistical graphics and other data visualization methods



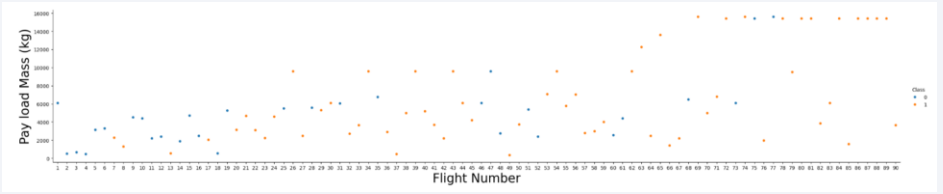
https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/05%20EDA%20Data%20Visualization%20.ipynb

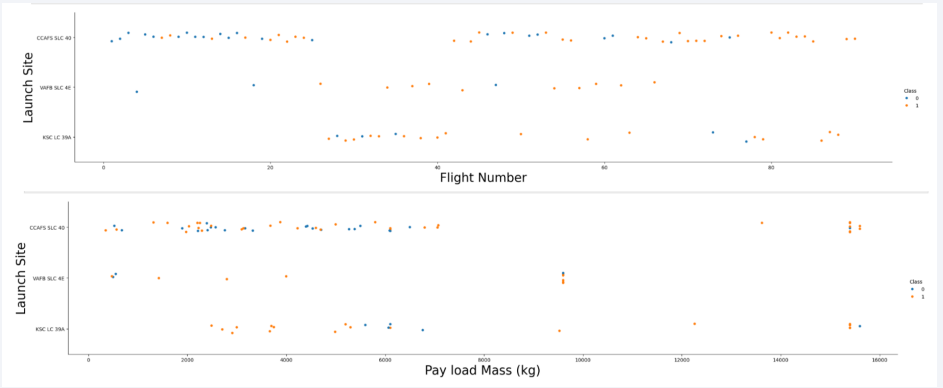
EDA with Data Visualization

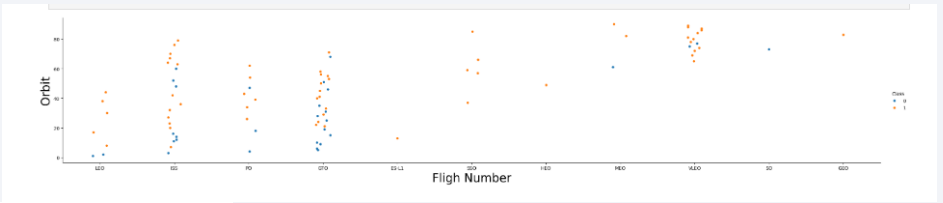
- Analysis

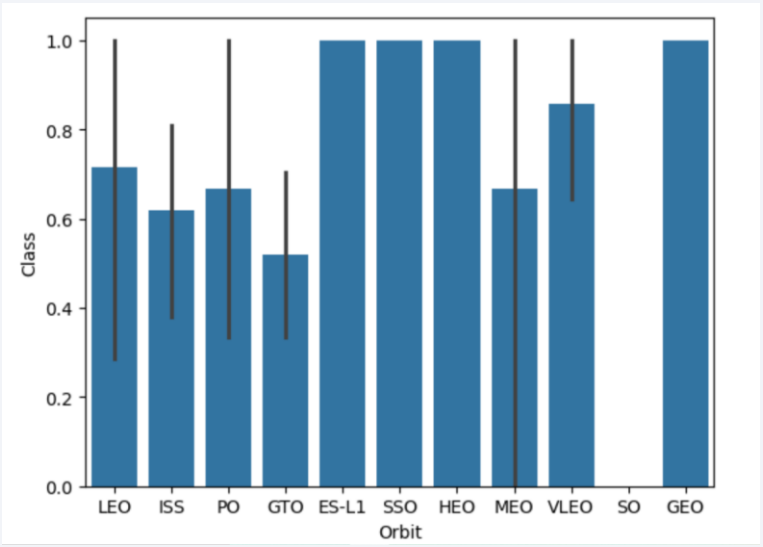
1.View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists

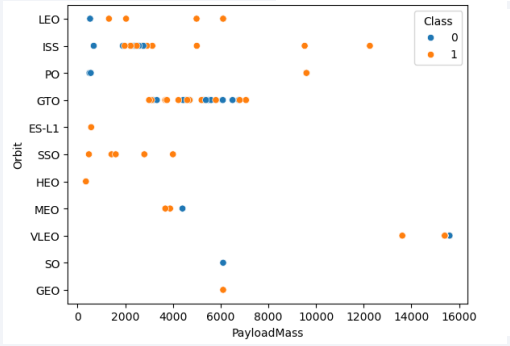
2. Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

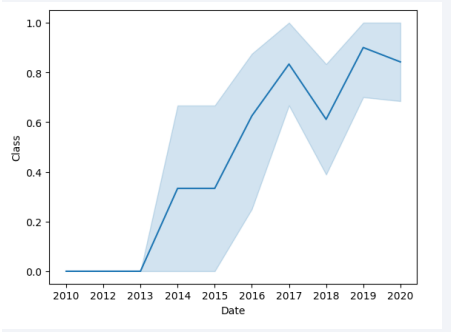












EDA with SQL

Queries Display:

1. Names of unique launch sites
2. 5 records where launch site begins with ‘CCA’
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1.

https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/04%20%20EDA%20-sql-coursera_sqllite.ipynb

List:

1. Date of first successful landing on ground pad
2. Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
3. Total number of successful and failed missions
4. Names of booster versions which have carried the max payload
5. Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
6. Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added red circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates

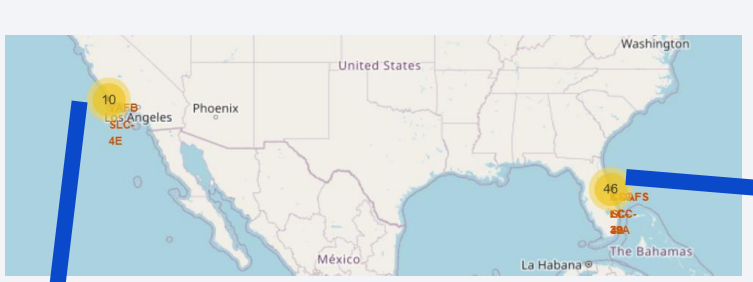
- Added black circles at all launch sites coordinates with a popup label showing its name

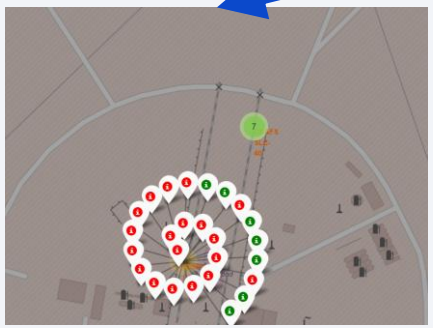
https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/06%20Folium%20Map%20SpaceX%20Launch%20Site.ipynb

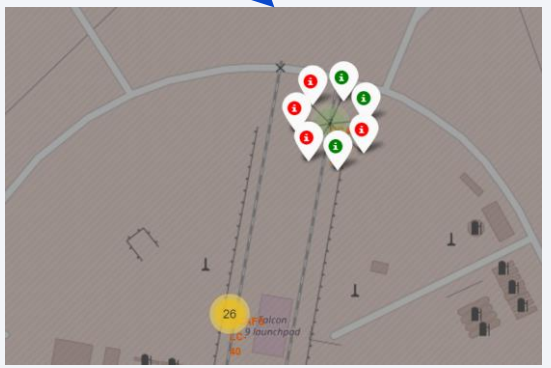
Build an Interactive Map with Folium

Colored Markers of Launch Outcomes

• Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates





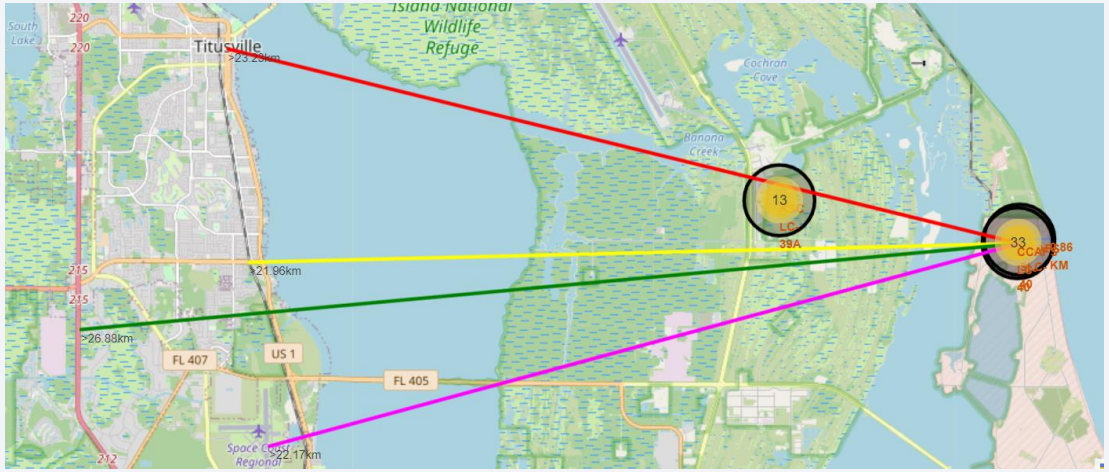


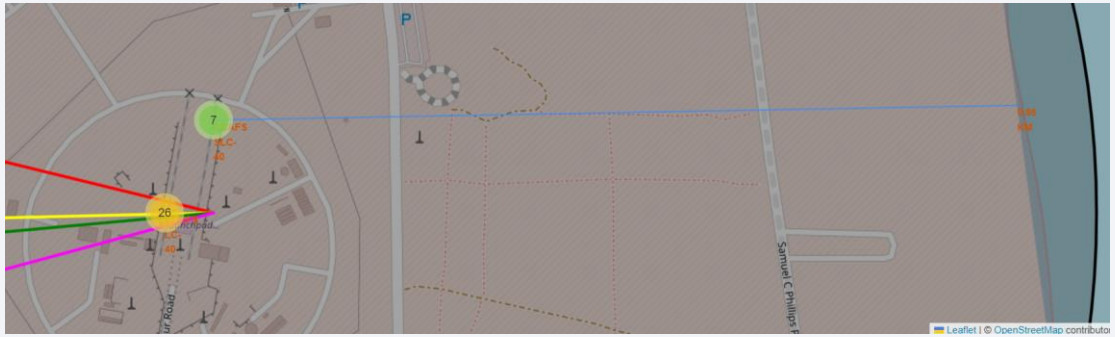
https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/06%20Folium%20Map%20SpaceX%20Launch%20Site.ipynb

Build an Interactive Map with Folium

Distances Between a Launch Site to Proximities

• Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city





https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/06%20Folium%20Map%20SpaceX%20Launch%20Site.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

• Allow user to select all launch sites or a certain launch site Dashboard with Plotly Dash Slider of Payload Mass Range

Pie Chart Showing Successful Launches

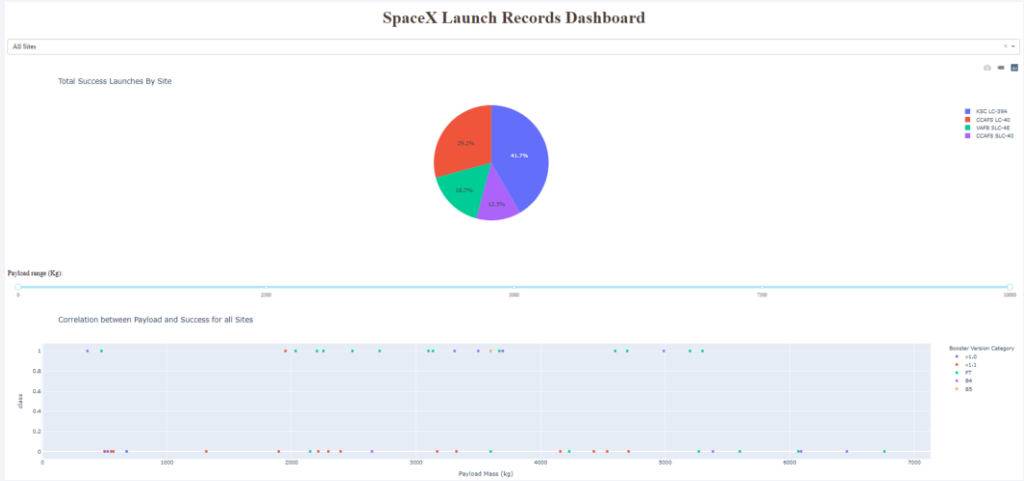
• Allow user to select payload mass range Pie Chart Showing Successful Launches

Slider of Payload Mass Range

• Allow user to see successful and unsuccessful launches as a percent of the total Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

• Allow user to see the correlation between Payload and Launch Success



https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Create NumPy array from the Class column
- Standardize the data with StandardScaler.
- Fit and transform the data
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Test Accuracy

https://github.com/Lacenedihia/Apllied-Data-Science-Submission/blob/main/07%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory Data Analysis

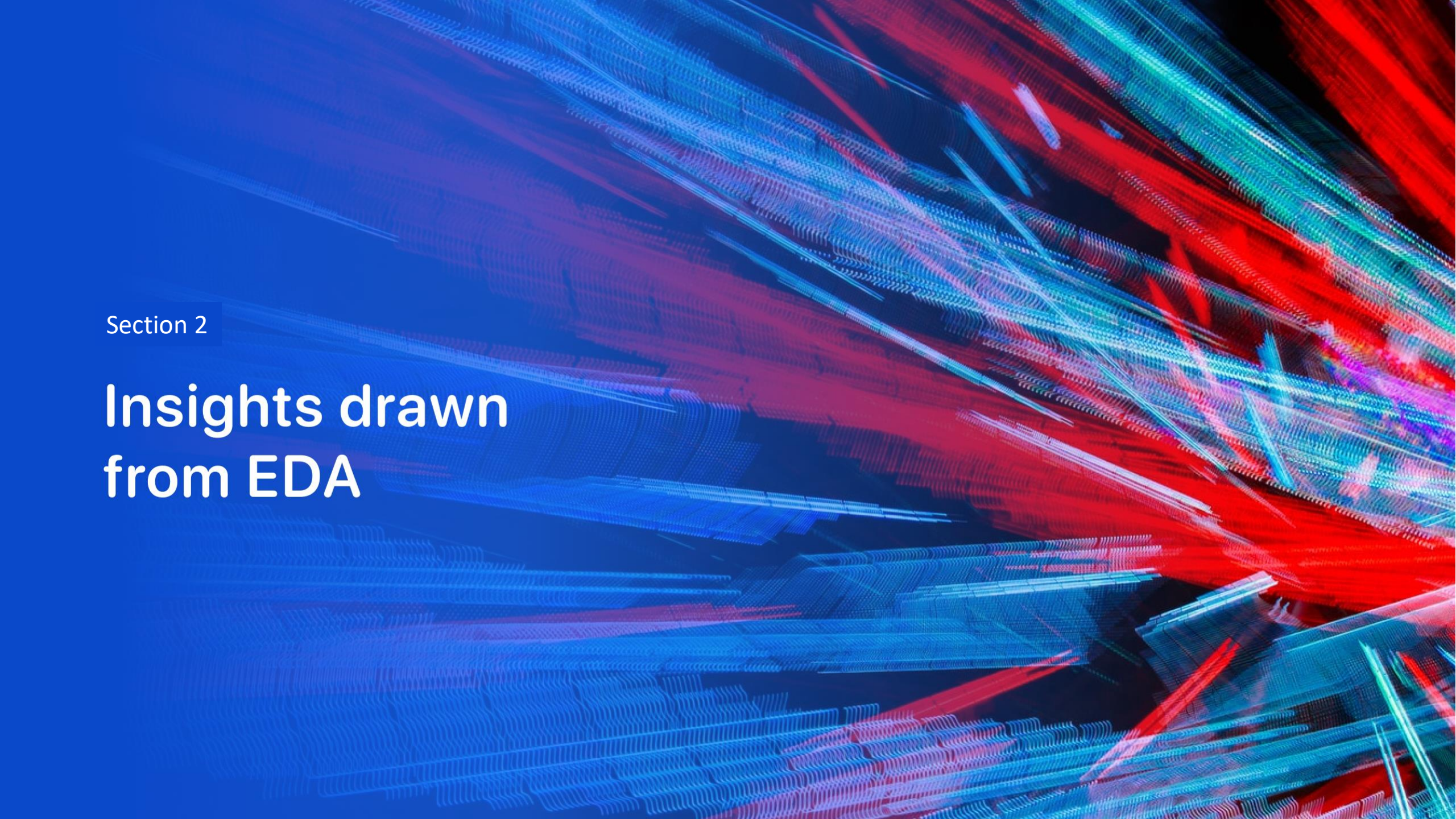
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

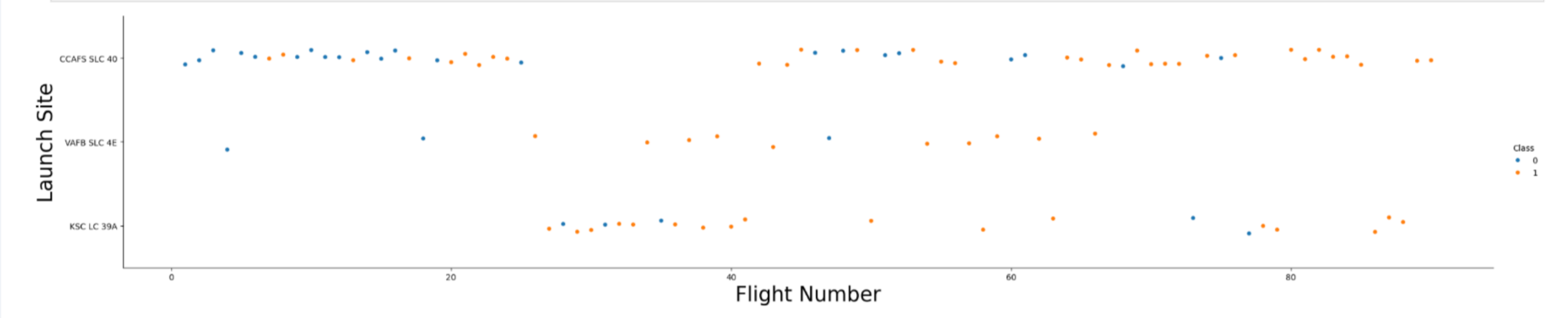
- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

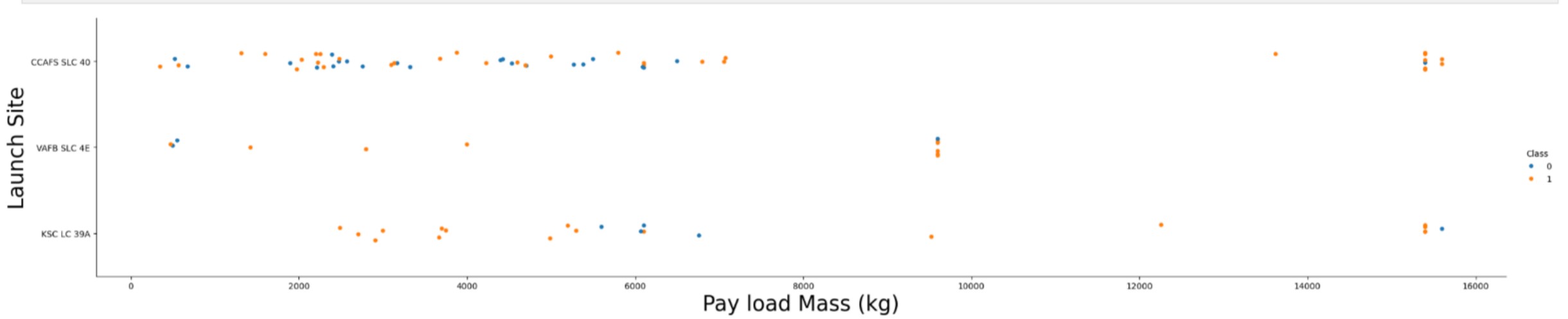
- • Earlier flights had a lower success rate (blue = fail)
- • Later flights had a higher success rate (orange = success)
- • Around half of launches were from **CCAFS SLC 40** launch site
- • VAFB SLC 4E and KSC LC 39A have higher success rates
- • We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



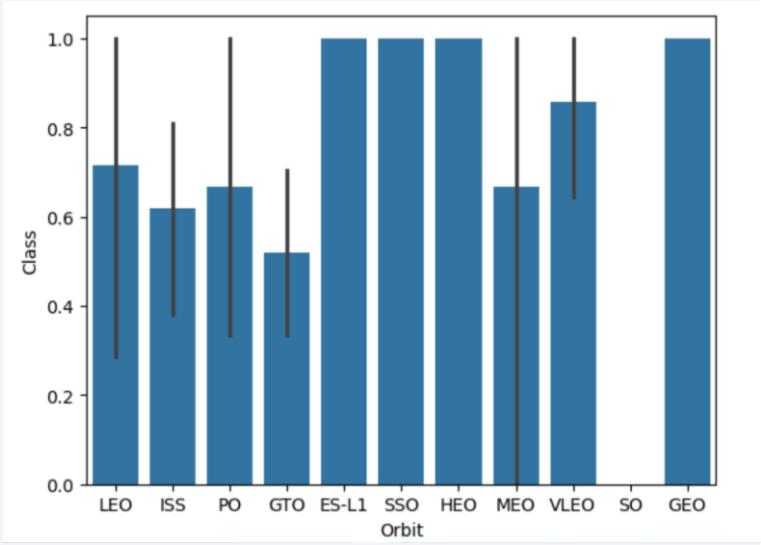
Success Rate vs. Orbit Type

Exploratory Data Analysis

100% **Success Rate : ES-L1, GEO,HEO and Sso**

50%-80% **Success Rate : GTO , ISS,LEO,MEO,PO**

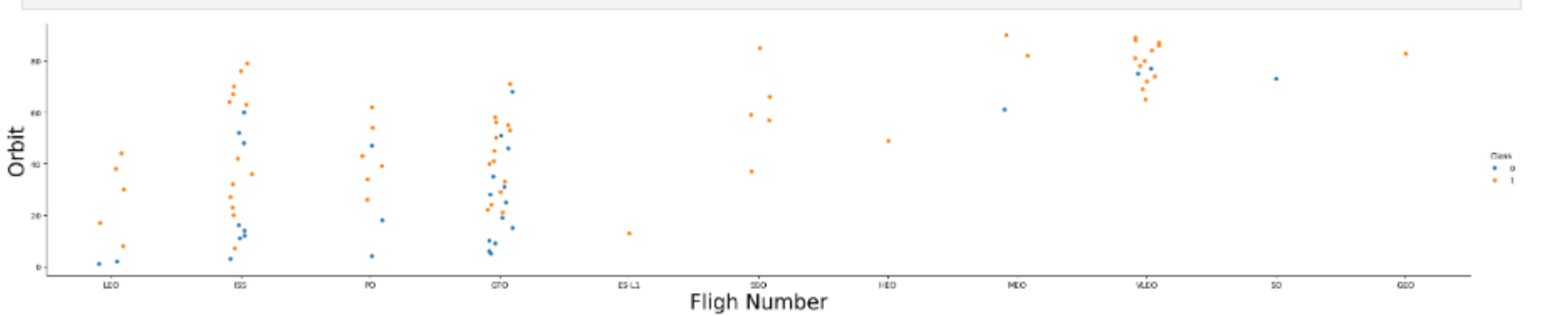
0% **Success Rate: SO**



Flight Number vs. Orbit Type

Exploratory Data Analysis

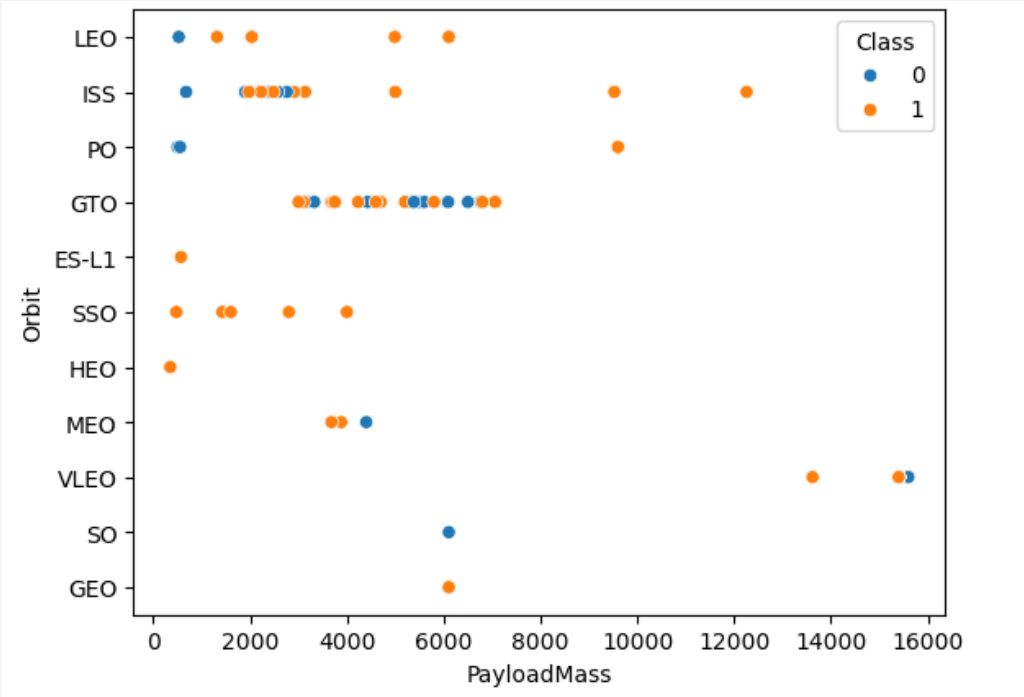
- ❑ The success Rate typically increases with the number of flights per each orbit especially for LEO orbit



Payload vs. Orbit Type

Exploratory Data Analysis

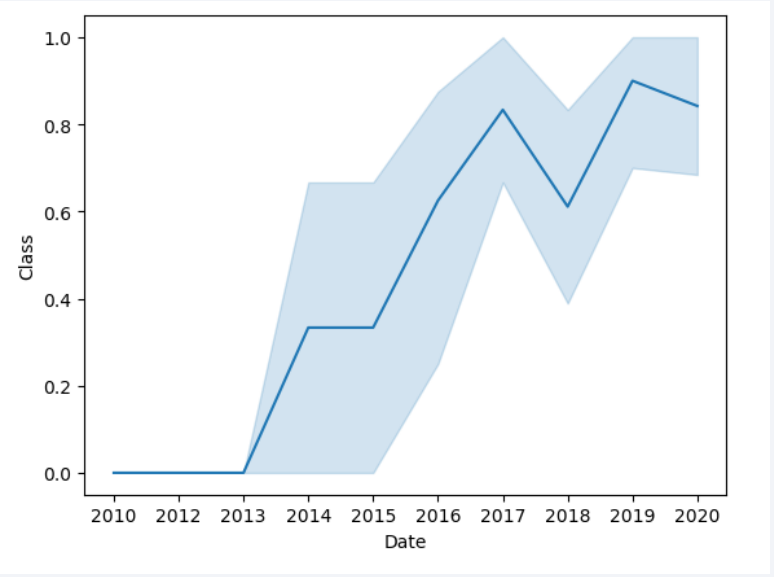
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

Exploratory Data Analysis

Since 2013 a steady improvement in the success rate can be witnessed



All Launch Site Names

Launch Site Names:

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

Records with Launch Site Starting with CCA

```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 vs Average Payload Mass

- **Total Payload Mass**

45.596 Kg (total) carried by boosters launches by NASAS (CRS)

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4l
   sqlite:///my_data1.db
Done.
```

1
45596

- **Average Payload Mass**

2.928 Kg (average) carried by booster version F9 v1.1

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was recorded the **12/22/2015**

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

1. JSCAT-14
2. JSCAT-16
3. SES-10
4. SES-11
5. EchoStar 105

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- **The total number of successful and failure mission outcomes**
  - 99 Success
  - 1 Success(with a payload status unclear)
  - 1 Failure in Flight

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

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

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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 Failed Landings on Drone Ship

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

❑ F9 v1.1 B1012 : CCAFS LC-40 : 10-01-2015

❑ F9 v1.1 B1015 : CCAFS LC-40 : 14-04-2015

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites with Markers

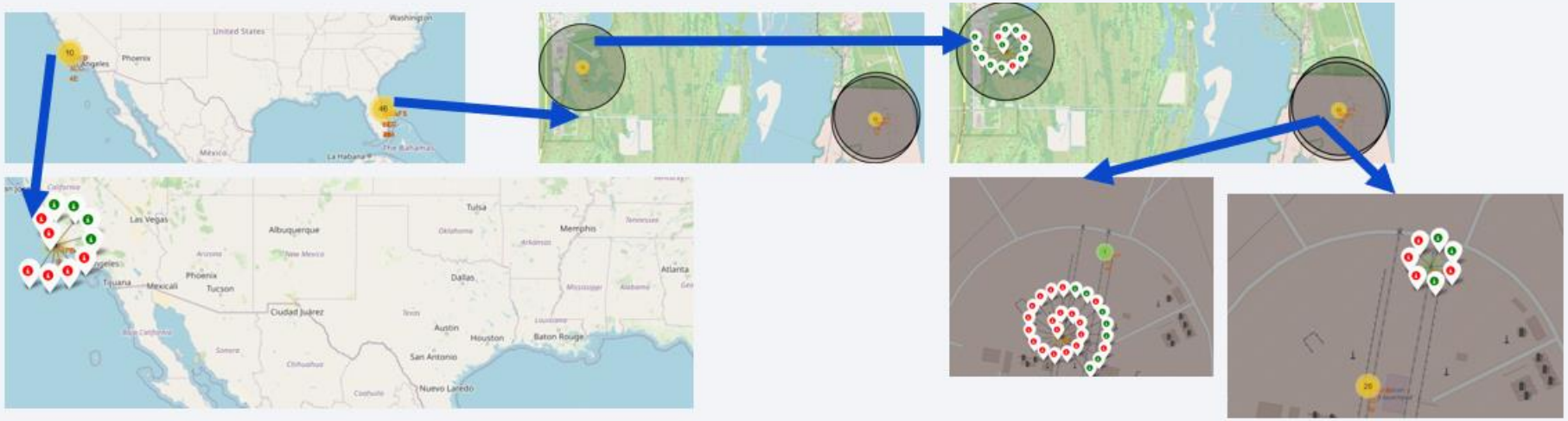
- **The Launch Sites are near the Equator**

As the closer the launch site to the equator , the easier it is to launch to equatorial orbit,and more help you get from Earth’s rotation for a prograde orbit due to the rotation speed of earth , that helps save the cost of putting in extra fuel and boosters

Launch Outcomes

At Each Launch Site

- Green markers for successful launches
- Red markers for unsuccessful launches

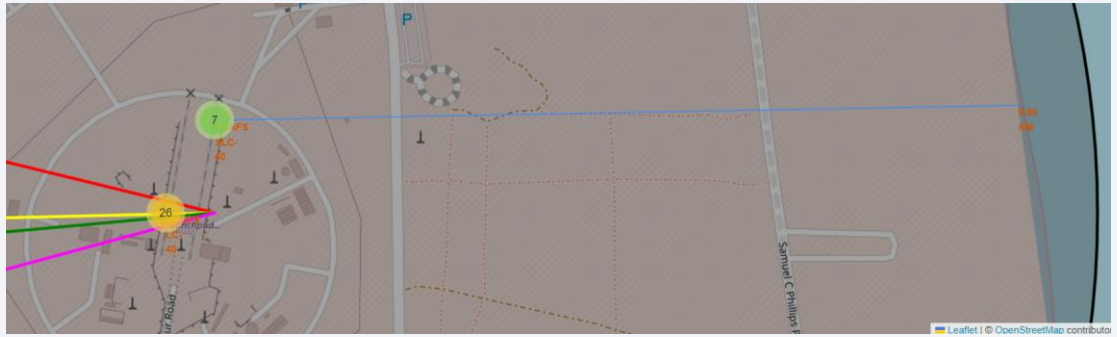


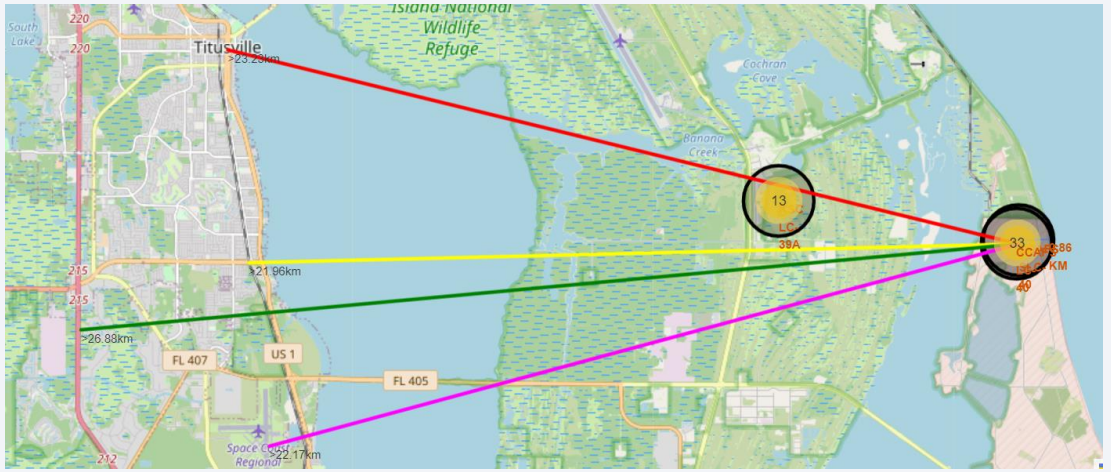
Distance to Proximities

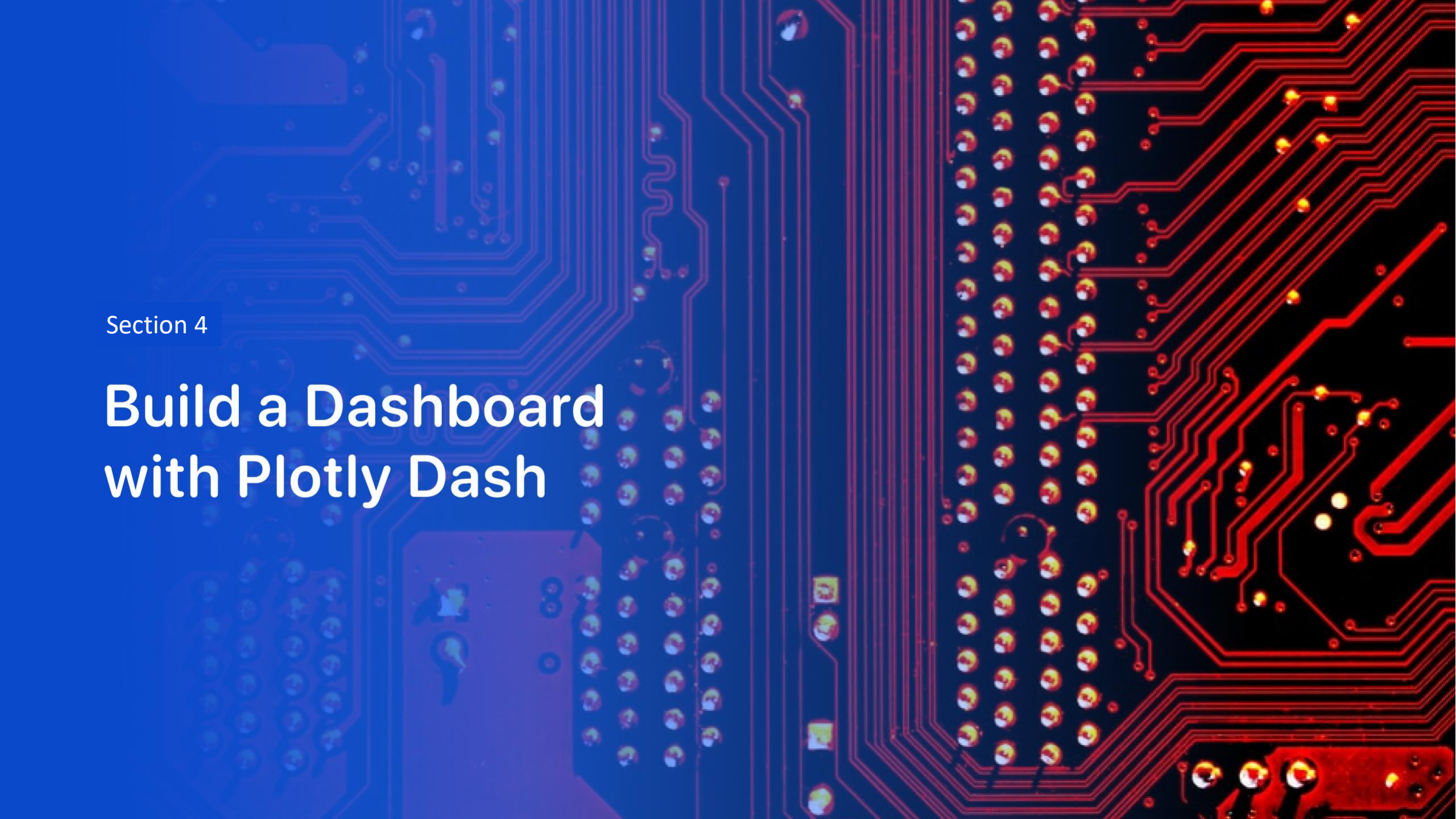
CCAFS SLC-40

- 23.23 Km from the nearest city
- 22.17 Km from the nearest airport
- 21.96 Km from the nearest railway
- 26.88 Km from the nearest highways
- 0.86 Km from the nearest coastline

There is a need for an exclusion zone around the launch site to keep unauthorized people away and keep people safe



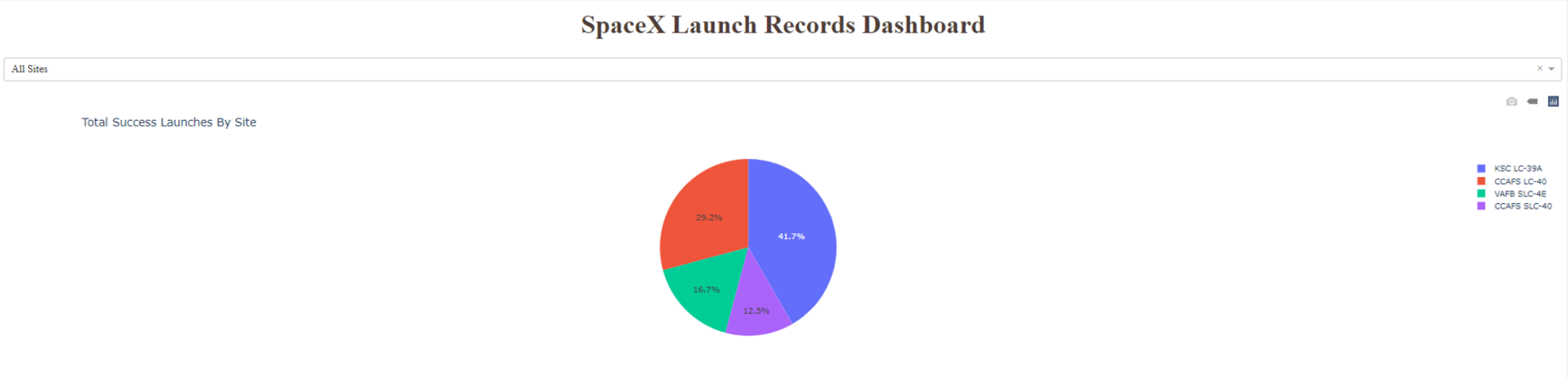



Section 4
Build a Dashboard with Plotly Dash

Launch Success by Sites

Success as Percent of Total

KSC LC-39A has the most successful launches amongst launches sites (41.2%)



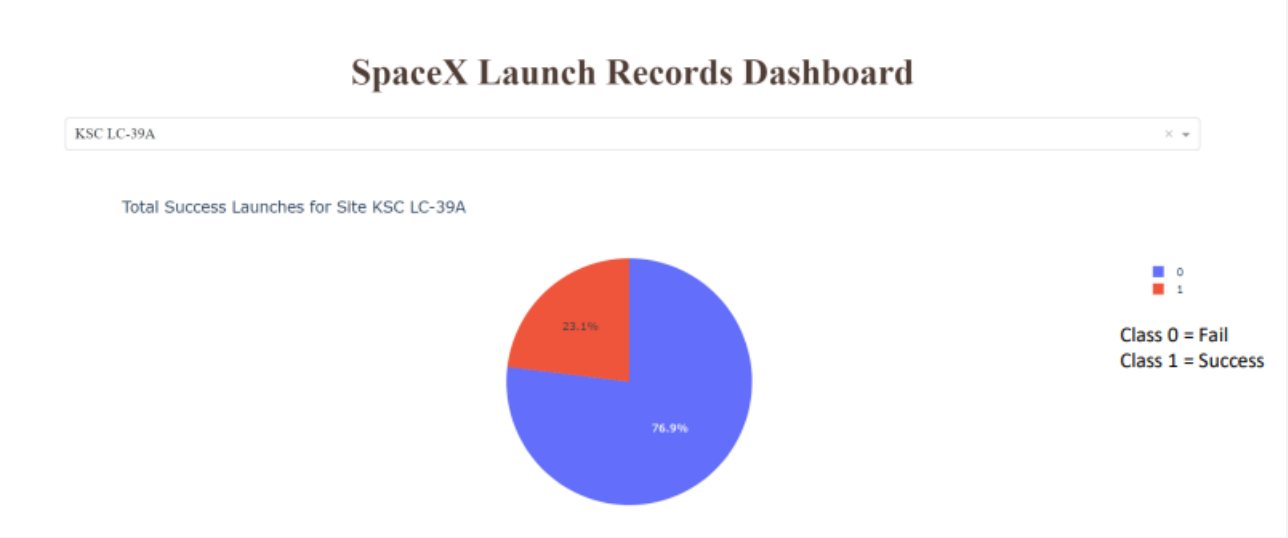
Launch Success(KSC LC-29A)

- Success as Percent of Total

KSC LC-39A has the highest success rate amongst launch sites(76.9%)

10 Successful Launches

3 Failed Launches



Payload Mass and Success

- By Booster Version

Payload between 2,000Kg and 5,000Kg have the highest success rate

1 Means Successful Outcome ----- 0 Means Unsuccessful Outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

All the models performed at the same lebel and had the same scores and accuracies.This is likely due to the small dataset

	KNN	Tree	Logistic Regression	SVM
Test Data Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

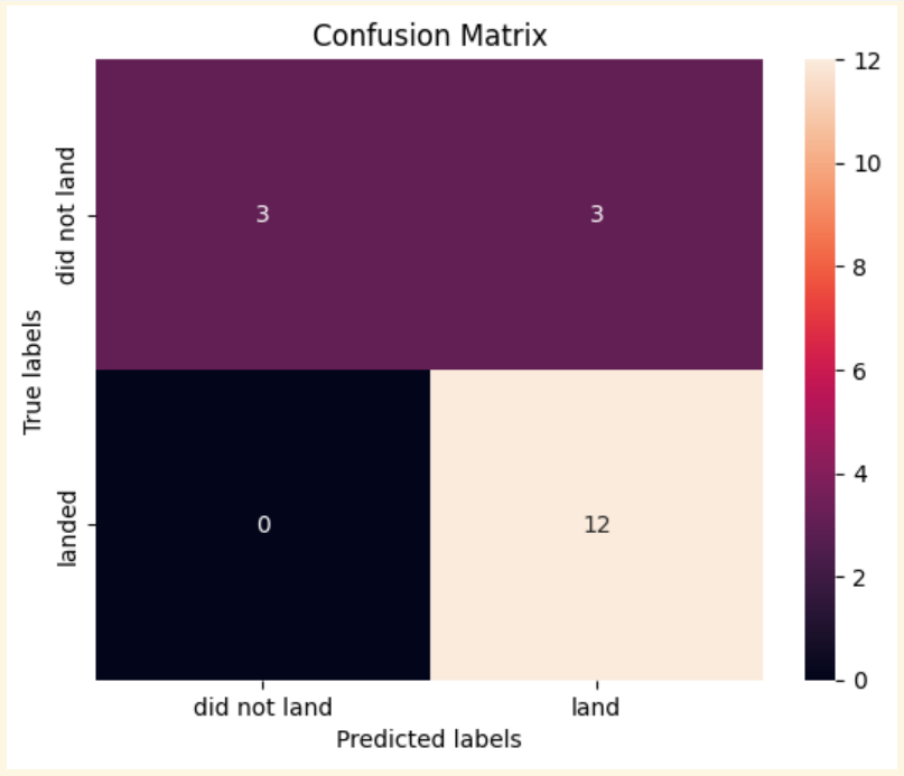
- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrics were identical
- The outputs

❑12 TP

❑3 TN

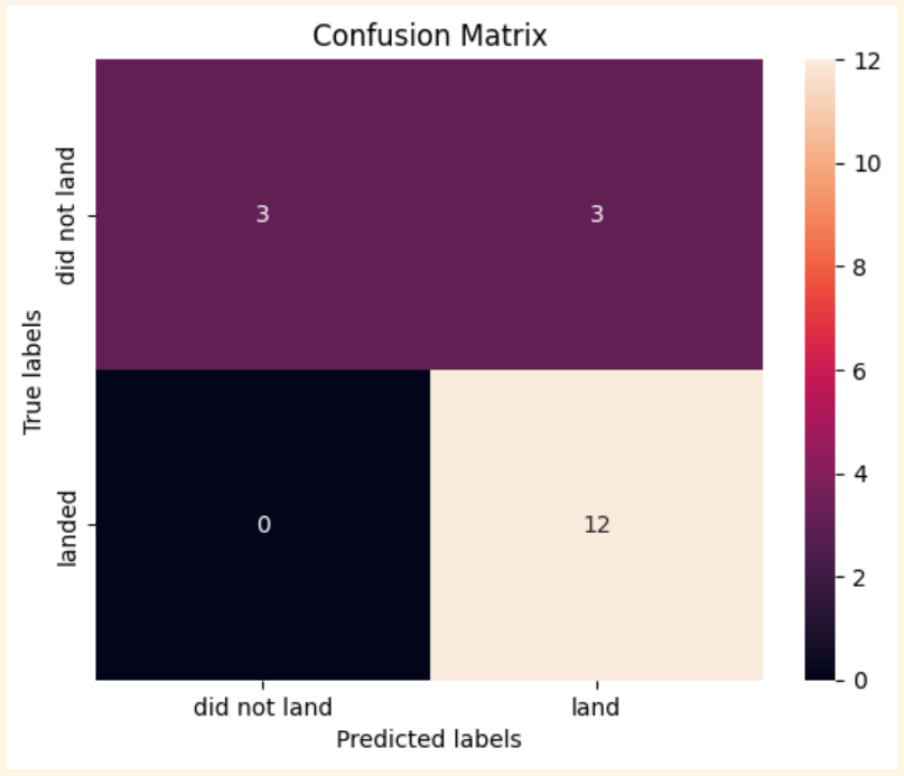
❑3 FP (not good)

❑0FN



Confusion Matrix

- **Precision** $\frac{TP}{TP+FP}=80\%$
- **Recall** $\frac{TP}{TP+FN}=1$
- **F1 Score** $2\times\frac{Precision\times Recall}{Precision+Recall}=89\%$
- **Accuracy** $\frac{TP+TN}{TP+TN+FP+FN}=83.3\%$



Conclusions

- **Model Performance**:The models performed similarly on the test
- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast • Launch Success: Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

- **Dataset:** A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- **Feature Analysis / PCA:** Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- **XGBoost:** Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

Conclusions

- Research

Model Performance:The models performed similarly on the test set with the decision tree model slightly outperforming

Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters

Coast: All the launch sites are close to the coast • Launch Success: Increases over time

KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg

Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate •

Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

- The Github Link :

https://github.com/Lacenedihia/Apllied-Data-Science-Submission

Thank you!