Winning Space Race with Data Science

Percival Mahwaya

05 September 2024

OUTLINE

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

EXECUTIVE SUMMARY

The goal of this research project is to determine the conditions that lead to a successful rocket landing. To forecast the following, we employed a number of machine learning classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

The primary actions taken to achieve this consist of:

- *Data Collection and Wrangling*
- *Exploratory Data Analysis*
- *Interactive Data Visualizations*
- *Machine Learning Prediction*

After thorough data collection, wrangling, and analysis, I performed extensive model tuning to find the best model. The main goal was to identify the model that provided the highest accuracy on our test data.

The results showed that Decision Tree Model outperformed the others, providing the highest accuracy after hyperparameter tuning. I will walk you through the methodologies and results that led us to this conclusion.

INTRODUCTION

- *Leading space company SpaceX works to lower the cost of space travel for all people. Its accomplishments include sending manned space missions, satellite constellation launches that enable internet access, and spacecraft to the International Space Station. The fact that SpaceX can reuse the first stage of the Falcon 9 rocket saves a significant amount of money; other providers charge upwards of 165 million dollars per launch. SpaceX advertises Falcon 9 rocket launches on its website. Thus, we can calculate the launch cost if we can ascertain whether the first stage will land.*
- *The majority of failed landings are planned, and occasionally SpaceX will carry out a controlled landing in the ocean. The primary question we will address is whether the first stage of the rocket landing will be successful for a given set of launch parameters (payload mass, orbit type, launch site, etc.) in order to achieve our reuse.*

Section 1

Methodology

METHODOLOGY

The overall methodology consists of:

1. ***Data collection and Wrangling***

- *SpaceX API* *Web Scraping*

2. ***Exploratory Data Analysis (EDA)***

- *Pandas* *Numpy* *SQL*

3. ***Data Visualization Techiniques***

- *Matplotlib* *Seaborn* *Folium* *Plotly Dash*

4. ***Machine Learning Prediction***

- *Logistic Regression* *Support Vector Machine (SVM)*
- *Decision Tree* *K-Nearest Neighbors (KNN)*

Data Collection and Wrangling

- *I collected the data from SpaceX API: https://api.spacexdata.com/v4/rockets/ . It was then imported into Python using pandas for easier manipulation and cleaning. The data required some preprocessing, including handling missing values.*

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	102	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	103	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058

- *We replaced the 26 missing values of the LaunchingPad column of our dataframe and ended up with 90 rows and 17 columns.*

Data Collection and Wrangling

- *For Web Scraping we used the Falcon 9 launch data from Wikipedia | Source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922*
- *After the preprocessing was completed the dataframe that we remained with had 121 rows and 11 columns.*

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

DATA COLLECTION – SPACEX API

The GitHub URL:

https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/Lab%201%20Data%20Collection%20SpaceX%20API.ipynb

Steps to be undertaken:

- *Request data from SpaceX API (rocket launch data)*
- *Decode response using .json() and convert to a dataframe using .json_normalize()*
- *Request information about the launches from SpaceX API using custom functions*
- *Create dictionary from the data*
- *Create dataframe from the dictionary*
- *Filter dataframe to contain only Falcon 9 launches*
- *Replace missing values of Payload Mass with calculated .mean()*
- *Export data to csv file*

DATA COLLECTION - SCRAPING

The GitHub URL:

https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/Lab%202%20Data%20Collection%20with%20Web%20Scraping.ipynb

Steps to be undertaken:

- *Request data (Falcon 9 launch data) from Wikipedia*
- *Create BeautifulSoup object from HTML response*
- *Extract column names from HTML table header*
- *Collect data from parsing HTML tables*
- *Create dictionary from the data*
- *Create dataframe from the dictionary*
- *Export data to csv file*

DATA WRANGLING

The data was subsequently processed to handle any missing entries and categorically encoded using one-hot encoding. Additionally, a new column, 'Class,' was added to the dataframe. This column identifies the success or failure of a launch, with '0' indicating a failed launch and '1' indicating a successful launch.

Steps undertake:

- *Performing EDA*
- *Creating binary landing outcome column (dependent variable)*
- *Export data to a csv file*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/Data%20Wrangling%20SpaceX.ipynb

EDA WITH DATA VISUALIZATION

Charts Plotted:

- *Relationship between Flight Number and Launch Site*
- *Relationship between Payload Mass and Launch Site*
- *Relationship between success rate of each orbit type*
- *Relationship between FlightNumber and Orbit type*
- *Relationship between Payload Mass and Orbit type*
- *Visualization of the launch success yearly trend*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/EDA%20Data%20Visualisation.ipynb

EDA WITH SQL

Queries performed were to:

- *Display the names of the unique launch sites in the space mission*
- *Display the total payload mass carried by boosters launched by NASA (CRS)*
- *List the date when the first succesful landing outcome in ground pad was achieved*
- *List the total number of successful and failure mission outcomes*
- *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.*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/EDA%20SQL.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- *We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.*
- *We added blue circle at NASA Johnson Space Center's coordinate with a popup label*
- *We added red circles at all launch sites coordinates with a popup label showcasing its name.*
- *Colored markers of launch outcomes were added, green for successful ones and red for unsuccessful launches*
- *We added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway and city*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

BUILD A DASHBOARD WITH PLOTLY DASH

We build a dashboard with the following Components and Features:

- *Dropdown menu for launch site selection for the purpose to filter the visualizations based on the selected launch site.*
- *Plotted pie chart for launch success rates to display the proportion of successful vs. Failed launches.*
- *We had a payload mass range slider to allow users to filter the data based on the payload mass range.*
- *Plotted a scatter plot for payload mass and launch successto visualize the correlation between payload mass and launch success, with points color-coded by the booster version.*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/tree/main/Ploty%20Dash%20Visualizations

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*
- *Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())*
- *Calculate accuracy on the test data using .score() for all models*
- *Assess the confusion matrix for all models*
- *Identify the best model using the Score method for accuracy*

GitHub URL | https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

RESULTS

Exploratory data analysis summary

- *Launch success has improved over time*
- *Orbits ES-L1, GEO, HEO, and SSO have 100% success rate*

Interactive analytics summary

- *Launch sites are far enough away from anything so that a failed launch could damge key points or vital points such as (city, highway, railway) while in the vicinity enough to bring resources easier.*

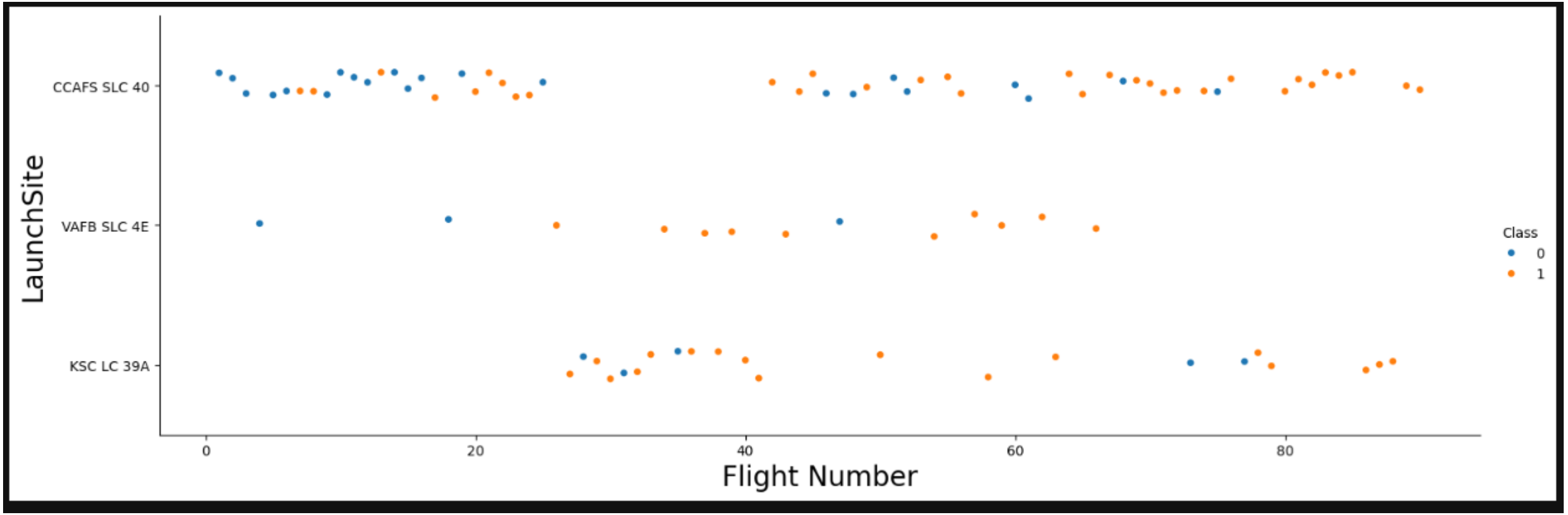
Predictive analysis summary

- *The decision tree model is the best predictive model for the dataset.*

Section 2

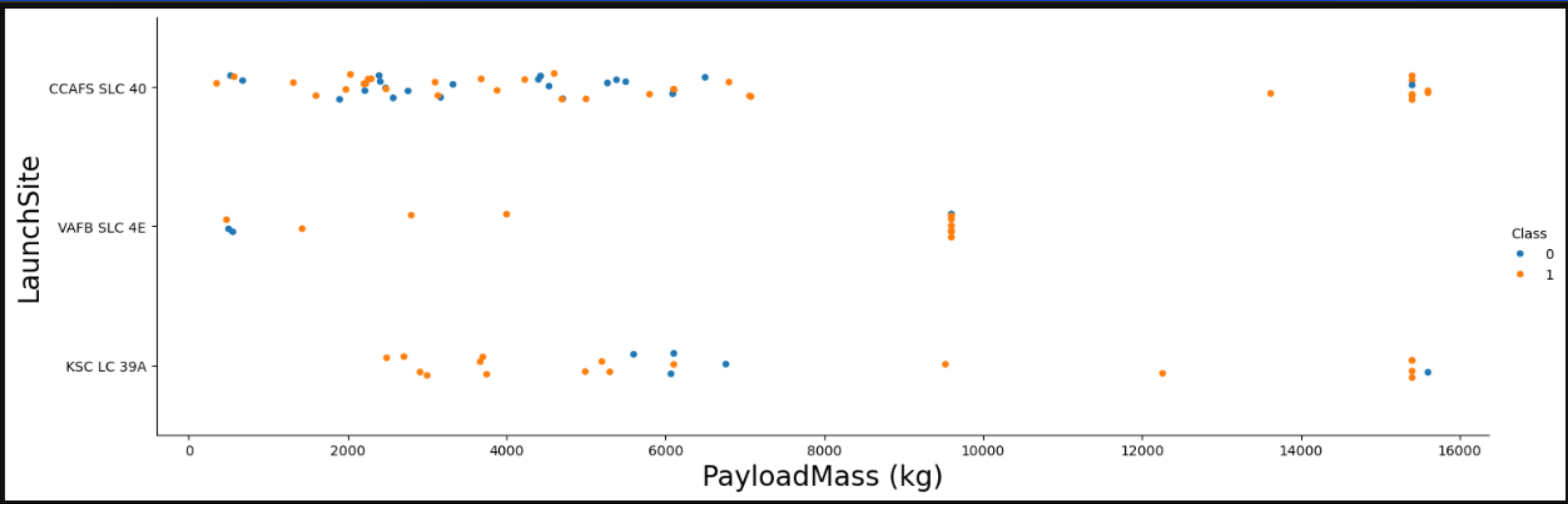
Insights drawn from EDA

FLIGHT NUMBER VS. LAUNCH SITE



- *Earlier launches had low success rate (blue=fail) but later launches had had an increase in success launches (orange=success)*
- *Almost 50% of launches were from CCAFS SLC40.*

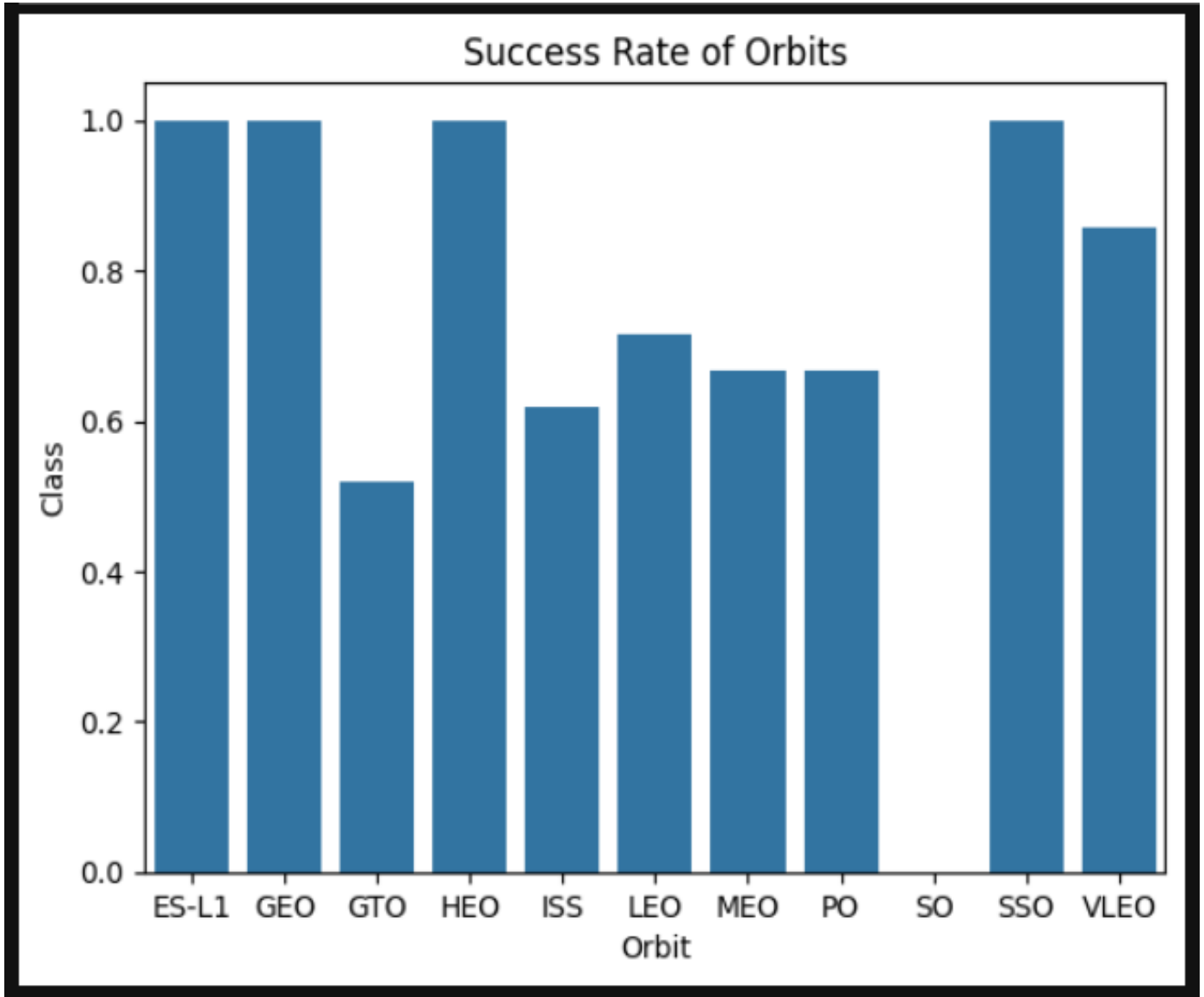
PAYLOAD VS. LAUNCH SITE



- *The higher the payload mass (kg), the higher the launch success rate*
- *Greater number of success launches for mass greater than 7000kg*

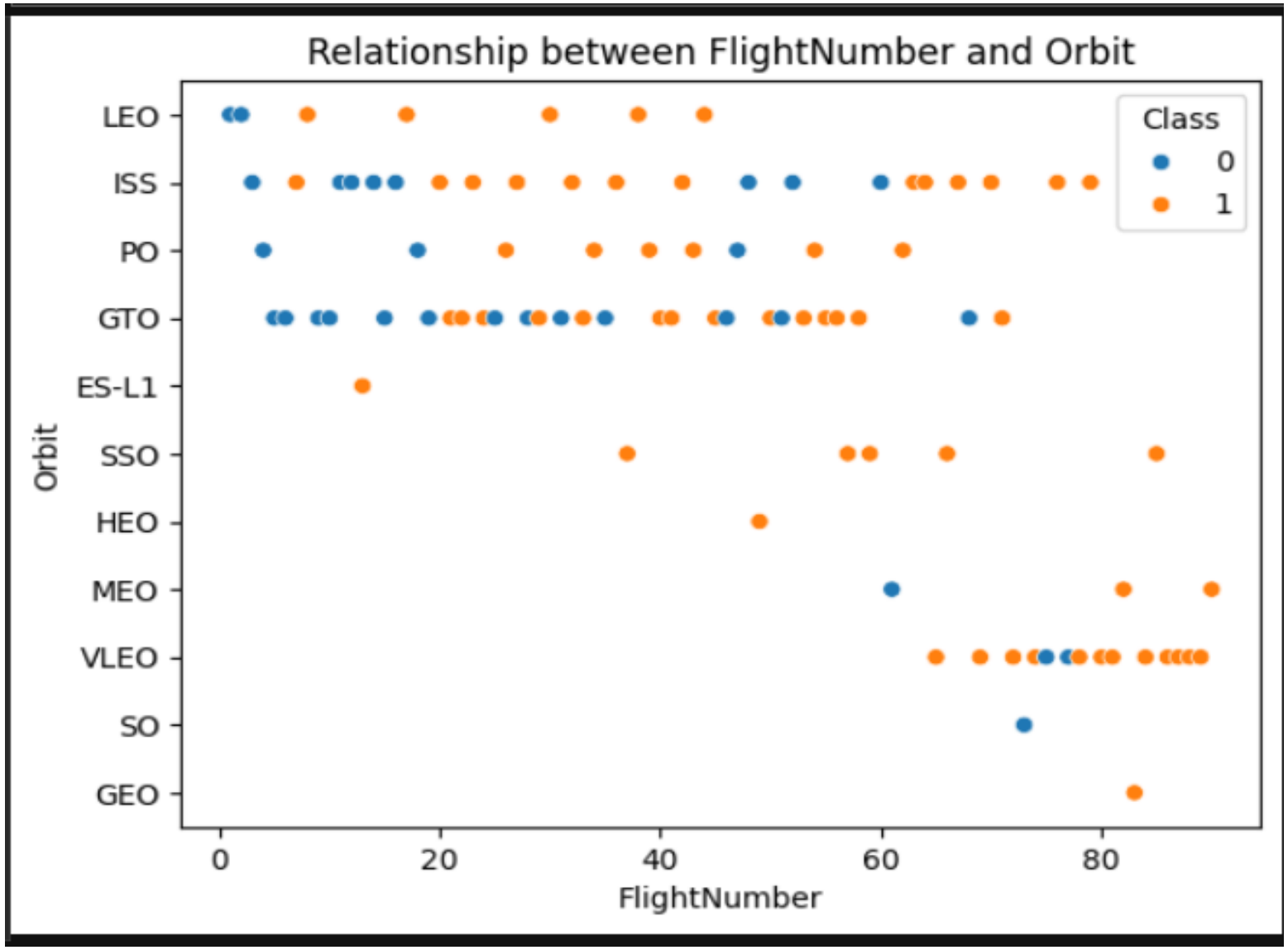
SUCCESS RATE VS. ORBIT TYPE

- *100% success rate for ES-L1, GEO, HEO, AND SSO*
- *50-80% success rate for GTO, ISS, LEO, MEO, PO*
- *No success rate for orbit SO*



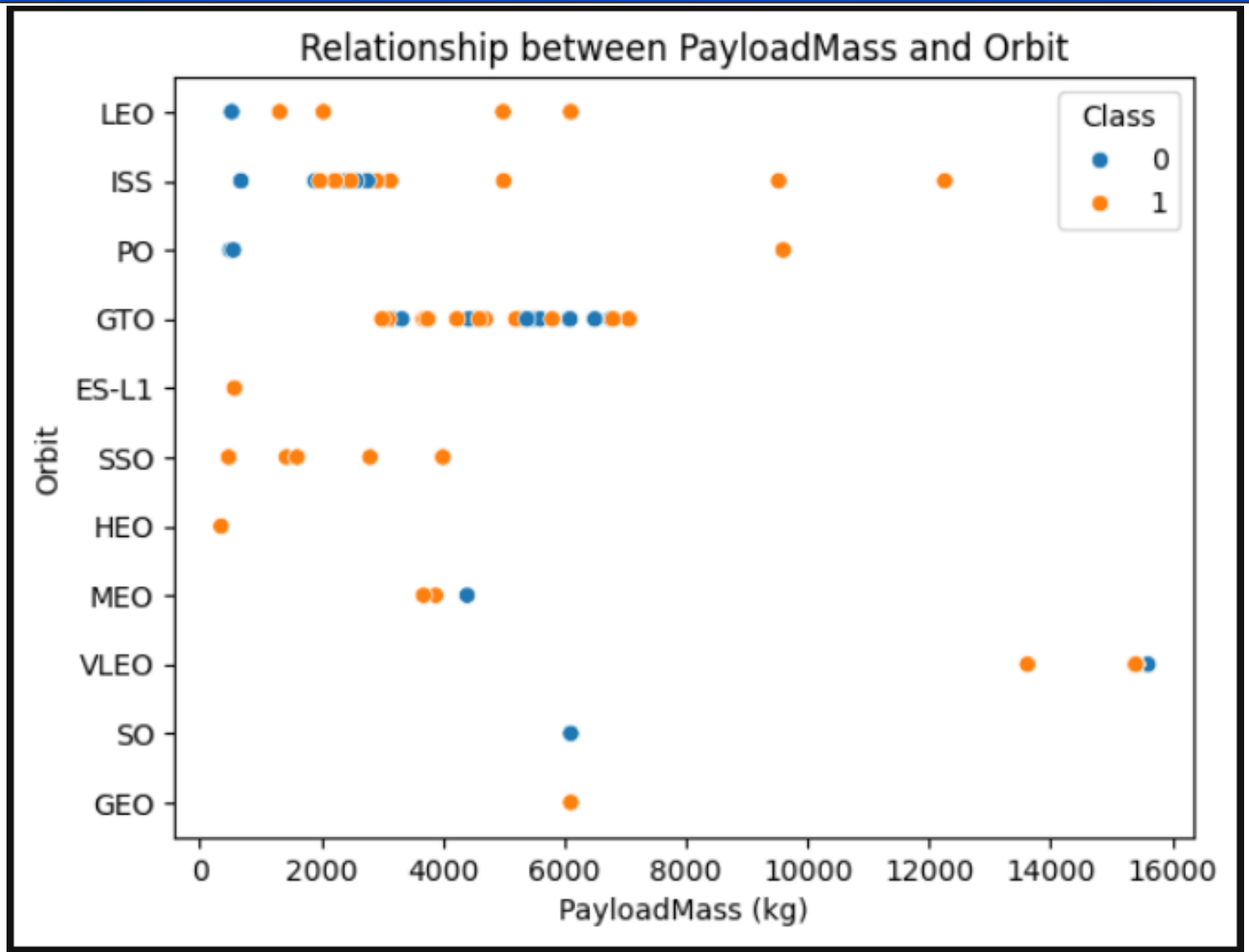
FLIGHT NUMBER VS. ORBIT TYPE

- *We 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.*



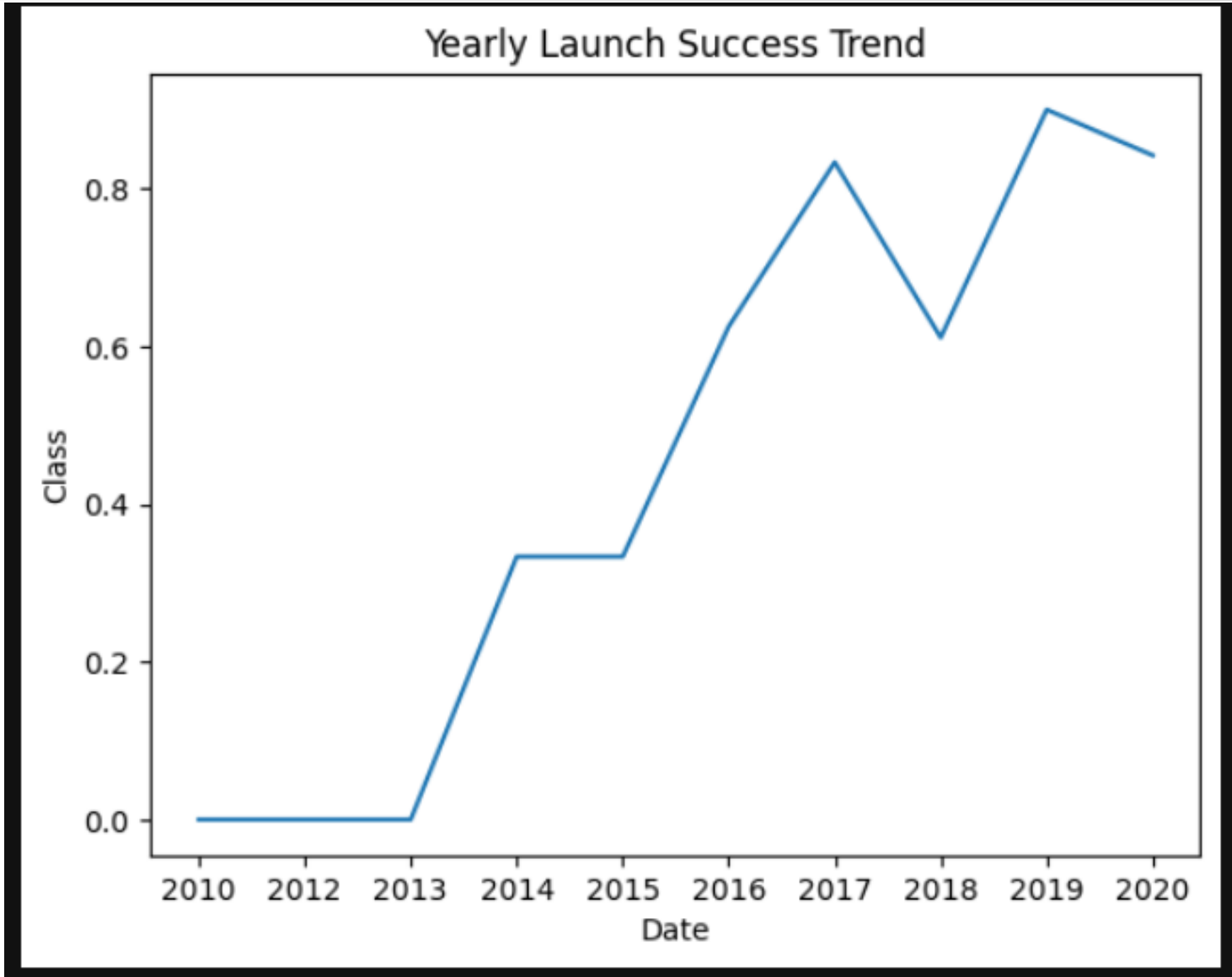
PAYLOAD VS. ORBIT TYPE

- *With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.*
- *For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present*



LAUNCH SUCCESS YEARLY TREND

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



ALL LAUNCH SITE NAMES

- *The names of the unique launch sites in the space mission*

```
%sql select distinct Launch_Site  from SPACEXTBL
 * 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'

Records where launch sites begin with `CCA`

```
%sql select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit(5)
```

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

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

TOTAL PAYLOAD MASS

The total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
sum(PAYLOAD_MASS__KG_)
45596
```

AVERAGE PAYLOAD MASS BY F9 V1.1

The average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

avg(PAYLOAD_MASS__KG_)
----------------------
                2928.4
```

FIRST SUCCESSFUL GROUND LANDING DATE

The dates of the first successful landing outcome on ground pad

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing With Payload Between 4000 And 6000

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

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

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

Total Number Of Successful And Failure Mission Outcomes

The total number of successful and failure mission outcomes

```
%sql select count(Mission_Outcome) from SPACEXTBL
```

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

count(Mission_Outcome)
101

Boosters Carried Maximum Payload

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

```
%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 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,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTBL
WHERE
    substr(Date, 1, 4) = '2015'
    AND Landing_Outcome LIKE '%Failure%'
    AND Landing_Outcome LIKE '%drone ship%';
```

Section 3

Launch Sites Proximities Analysis

Launch Sites

Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost due to the rotational speed of earth that helps save the cost of putting in extra fuel and boosters.

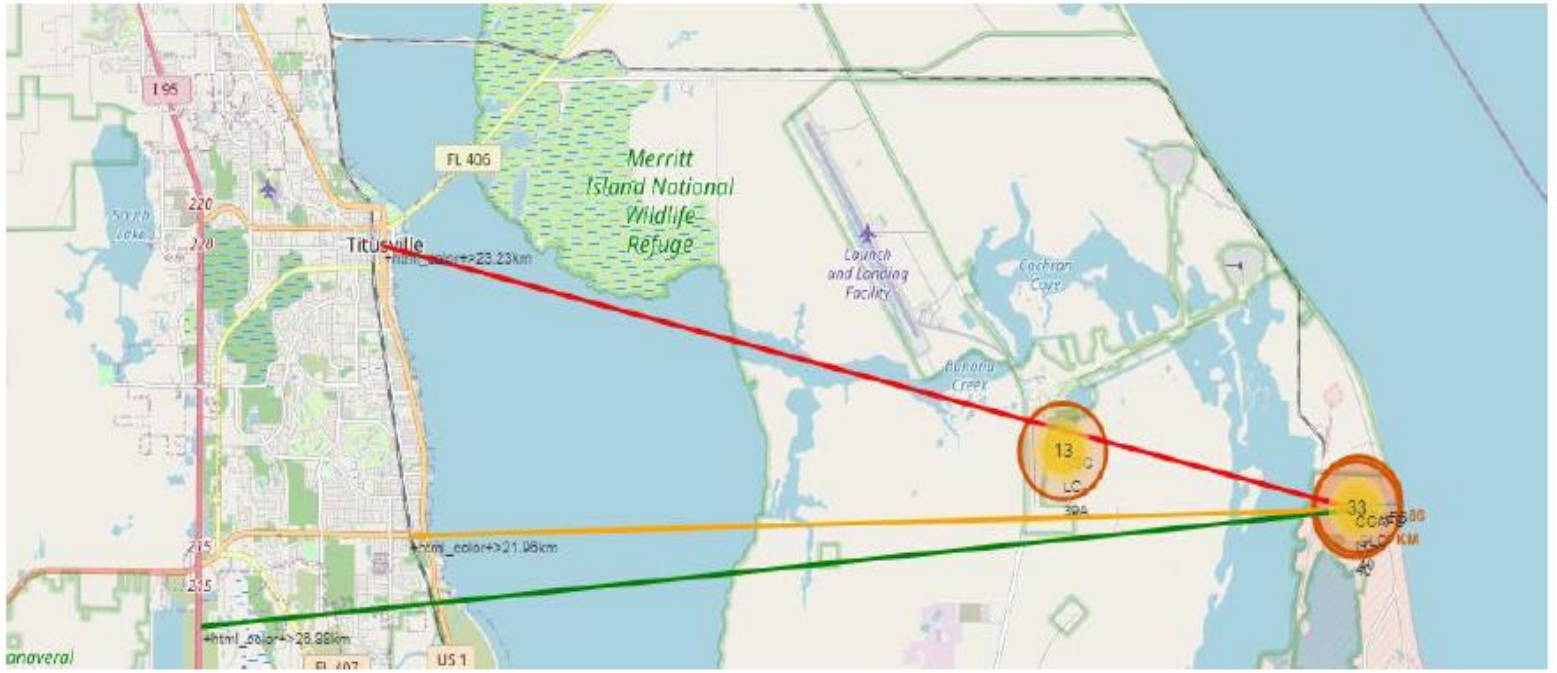
Launch Outcomes

Green markers for successful launches, Red markers for unsuccessful launches

Launch site CCAFS SLC 40 has a 3/7 success rate 42.9%)

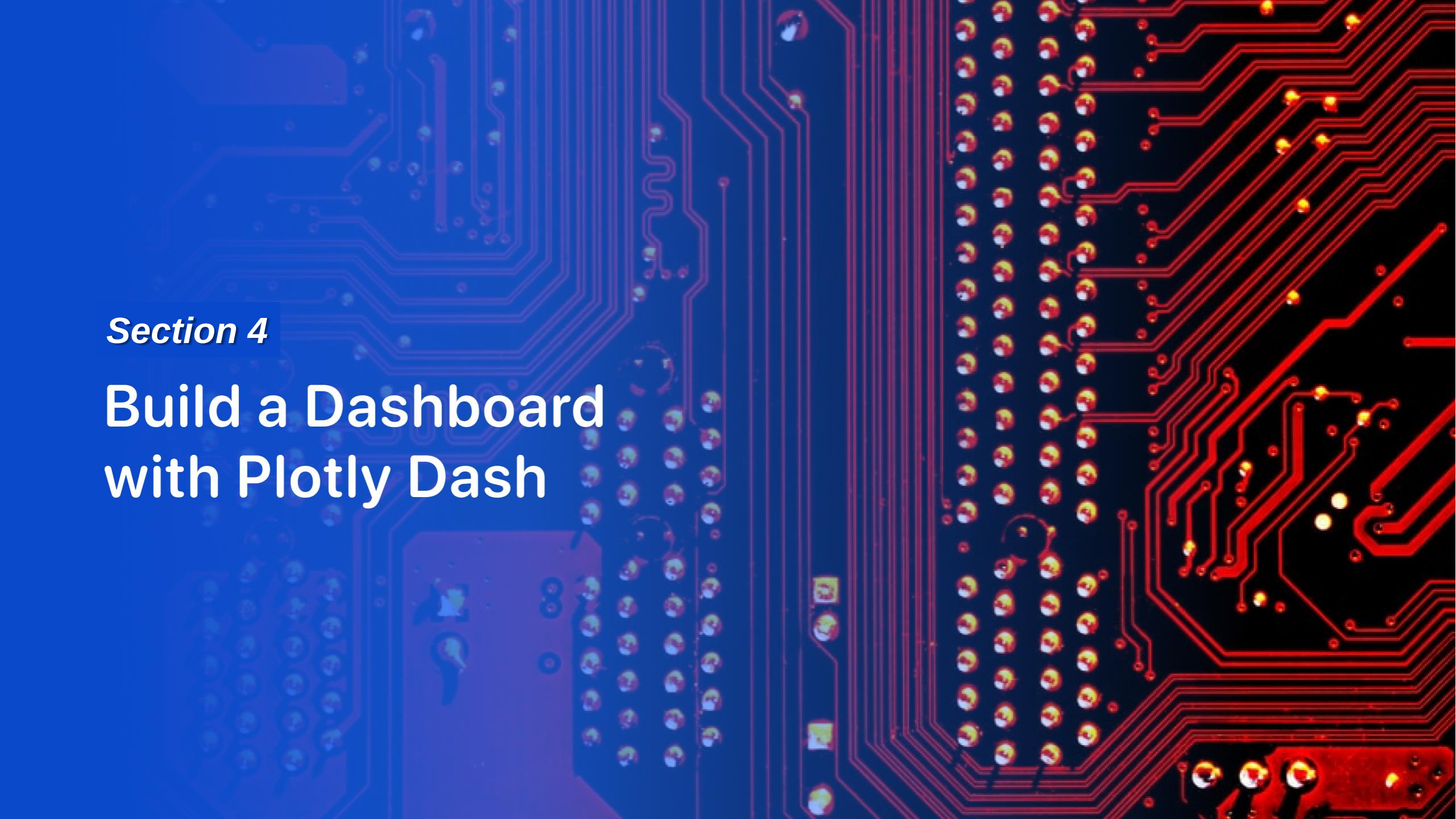
Distance to Proximities

- ***Coasts**: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.*
- ***Safety / Security:**needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.*
- ***Transportation/Infrastructure and Cities**: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.*

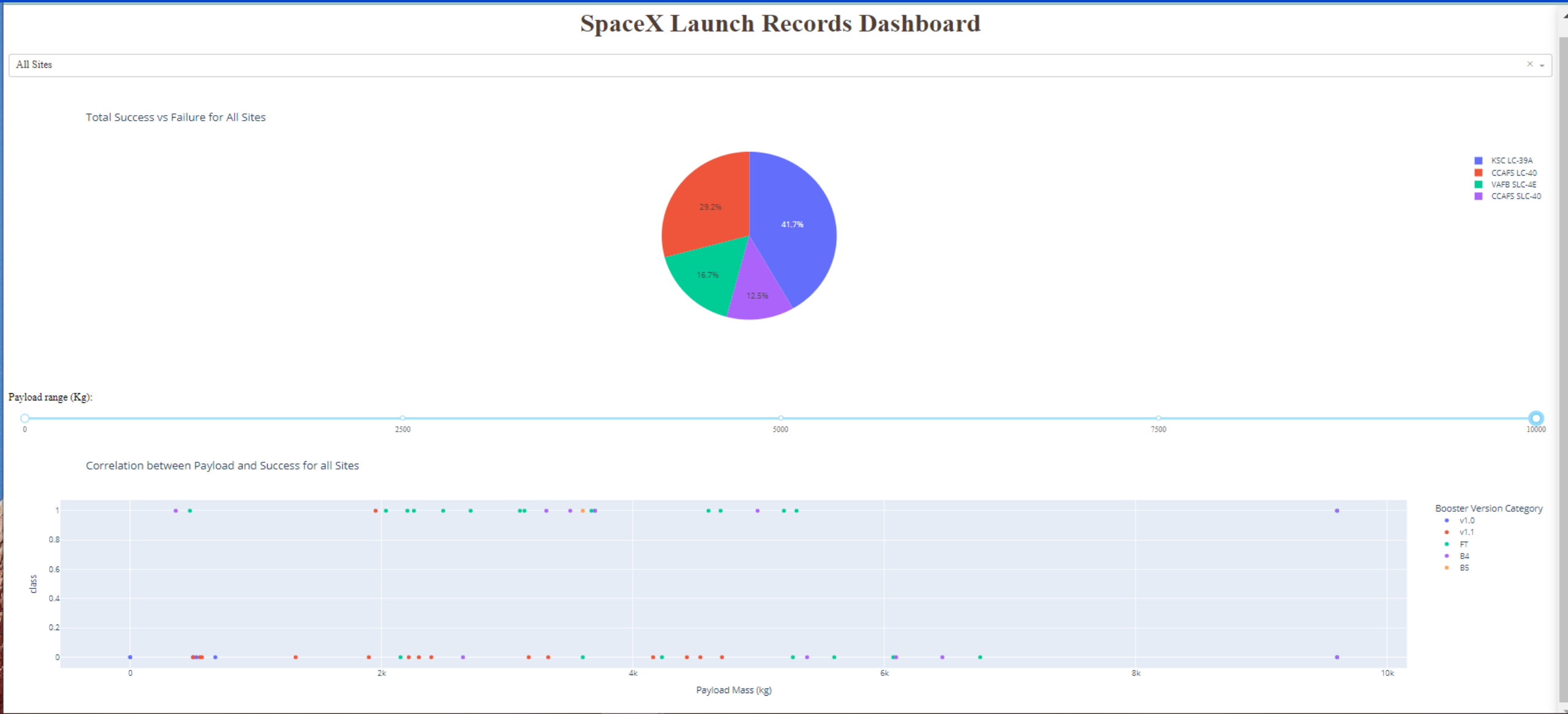


Section 4

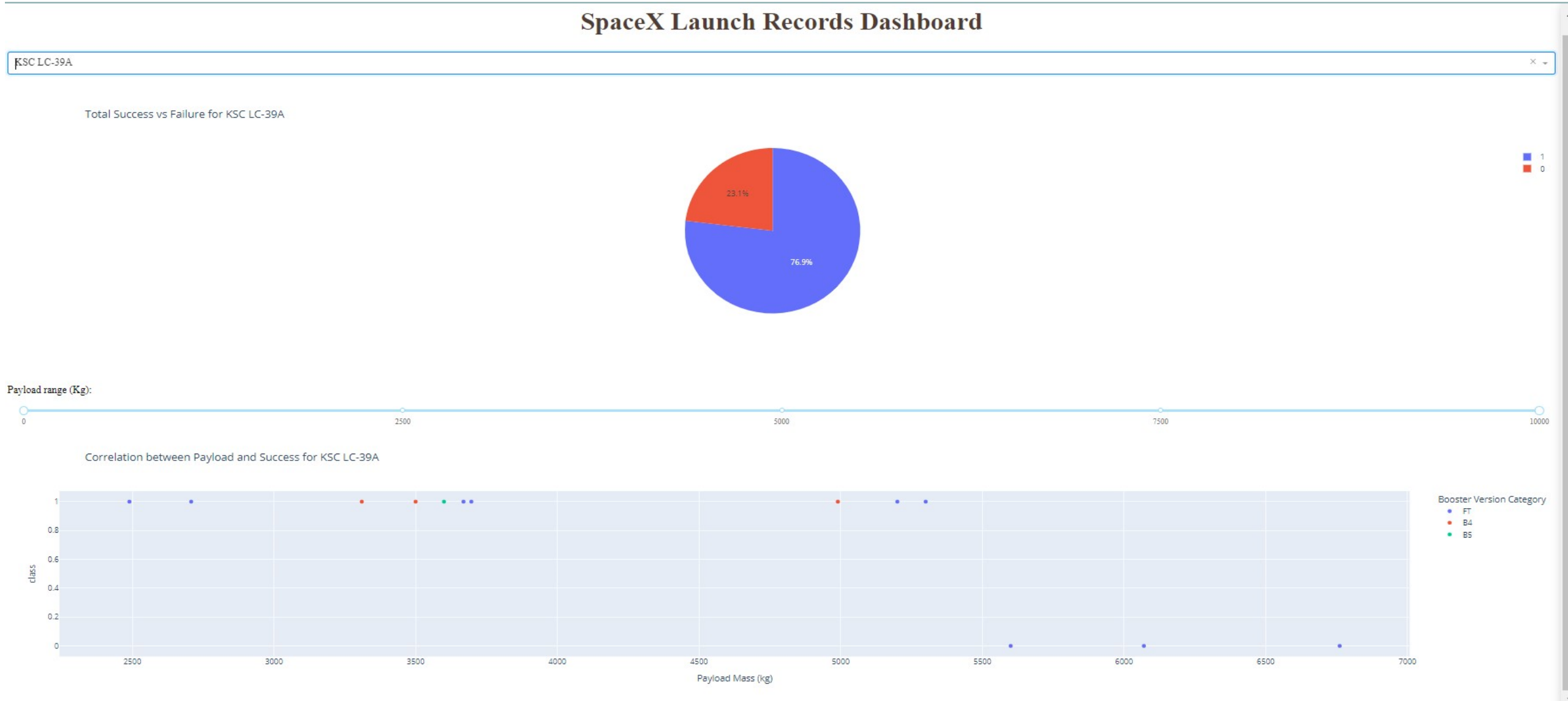
Build a Dashboard with Plotly Dash



Launch Success by Site



Launch Success (KSC LC-299A)

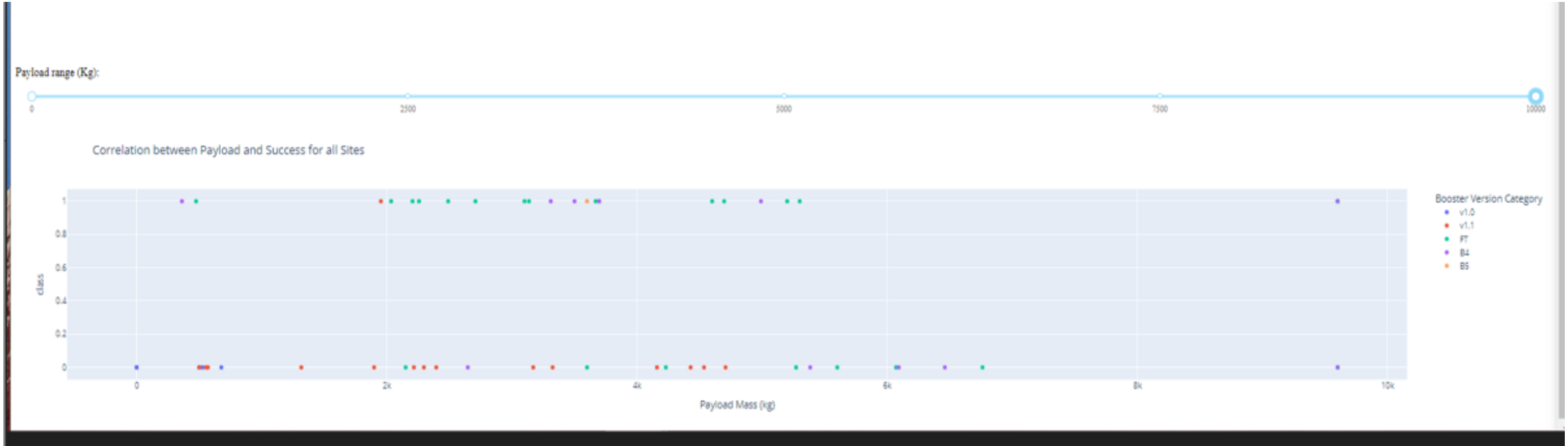


Payload Mass and Success

By Booster Version

Payloads between
2,000 kg and 5,000 kg have the highest success rate

1 indicating successful outcome and 0 indicating an unsuccessful outcome



Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
''' The decision tree classifier is the best method among the tested ones.'''
```

```
' The decision tree classifier is the best method among the tested ones.'
```

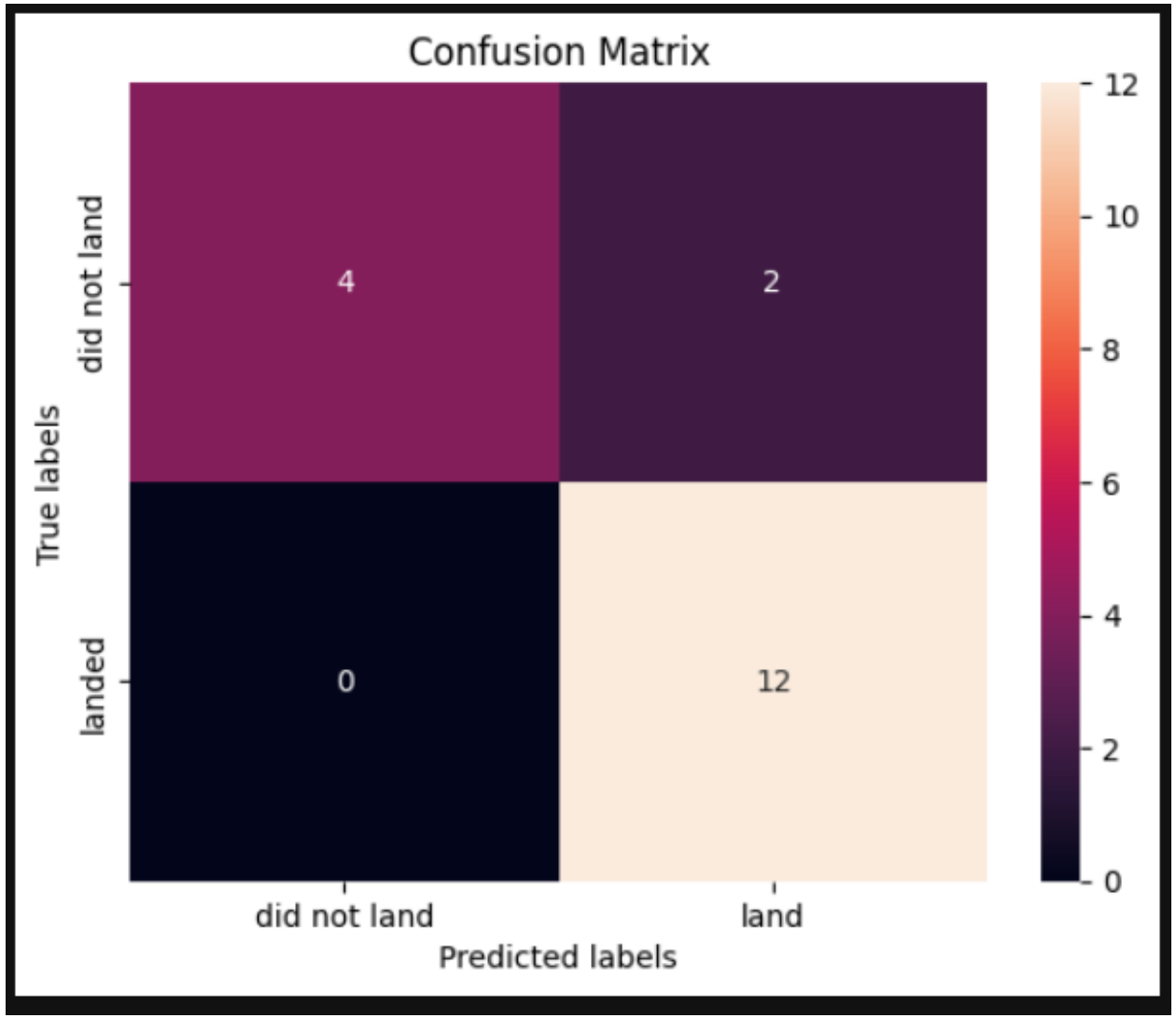
```
# Determine which method performs best
accuracies = {
    'Logistic Regression': logreg_cv.score(X_test, Y_test),
    'SVM': svm_cv.score(X_test, Y_test),
    'Decision Tree': tree_cv.score(X_test, Y_test),
    'K-Nearest Neighbors': knn_cv.score(X_test, Y_test)
}

best_method = max(accuracies, key=accuracies.get)
print(f"The best performing model is: {best_method} with an accuracy of {accuracies[best_method]:.4f}")
```

```
The best performing model is: Decision Tree with an accuracy of 0.8889
```

Confusion Matrix

The best performing model is: Decision Tree with an accuracy of 0.8889



Conclusions

- *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*
- *In conclusion, after testing and tuning four different classification models, the Decision Tree model emerged as the best performer with an accuracy of 88.89%. The exploratory data analysis provided valuable insights into feature relationships, and the interactive tools such as Folium and Plotly Dash helped visualize data in new ways. Moving forward, we can use this model to predict outcomes and refine it further if additional data becomes available.*

Appendix

- https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/tree/main/Exploratory%20Data%20Analysis
- https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py
- https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/blob/main/spaceXdash_app.txt
- https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/tree/main/Exploratory%20Data%20Analysis
- https://github.com/Mahwaya/IBM-Applied-Data-Science-Capstone-Project/tree/main/Ploty%20Dash%20Visualizations

Thank you!