IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sarah Siedlik
March 31, 2022

Outline

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

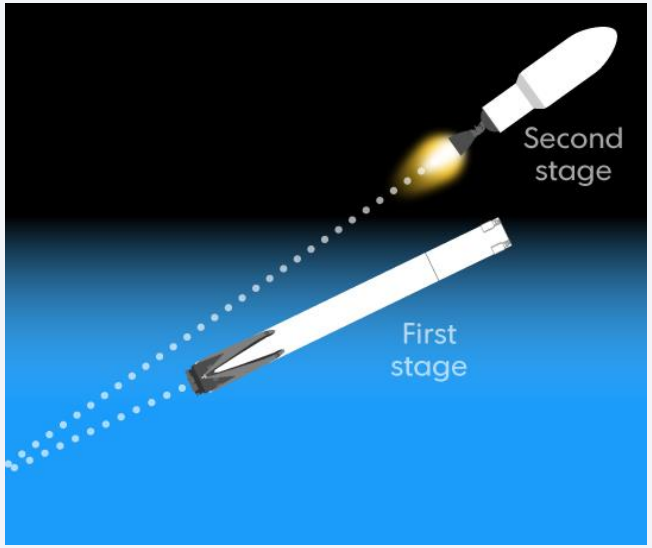
Executive Summary

Summary of Methodology

- Data Collecting using an API
- Web Scraping a Wikipedia page
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Predictive Analysis

Introduction

- In this capstone I took the role as a data scientist working for a new rocket company, Space X
- If Space X is able to reuse the first stage, it will save the company millions of dollars. With this knowledge my goal was to determine the price of each launch.
- I did this by gathering information about Space X, analyzing the data using SQL and data visualization tools, and created an interactive Dashboard
- I trained a machine learning model, using publicly available information, to predict if Space X will reuse the first stage.



Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Get requests to the Space X API and web scraping data from Wikipedia
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create machine learning model based on the training inputs

Data Collection

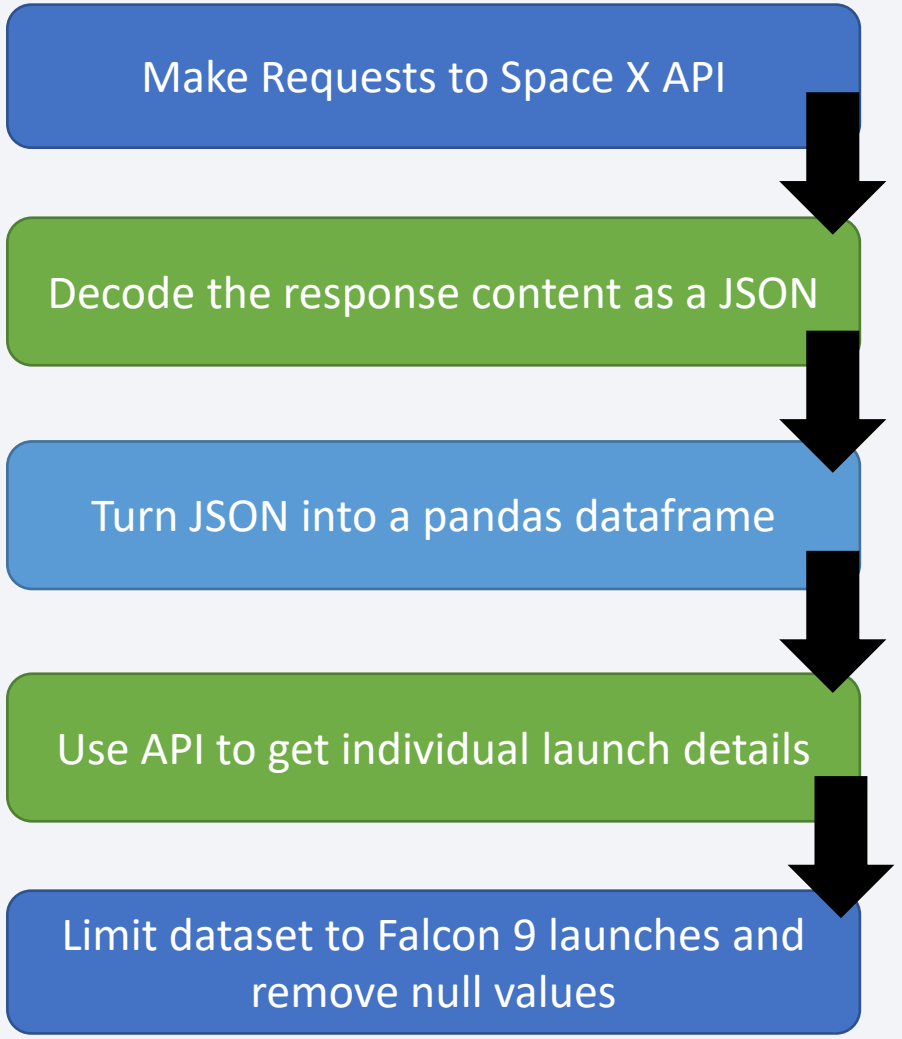
The data sets were collected by:

1. Get requests to the Space X API (https://api.spacexdata.com/v4/launches/past)
2. Decode the response content as a JSON and turned it into a Pandas dataframe
3. Use the API again to get information about each launch
4. Limit down the dataframe to include only the features we want and to only include Falcon 9 launches
5. Null values were replaced
6. Web scape the Space X Wikipedia page

Data Collection – SpaceX API

- Space X Call Flowchart

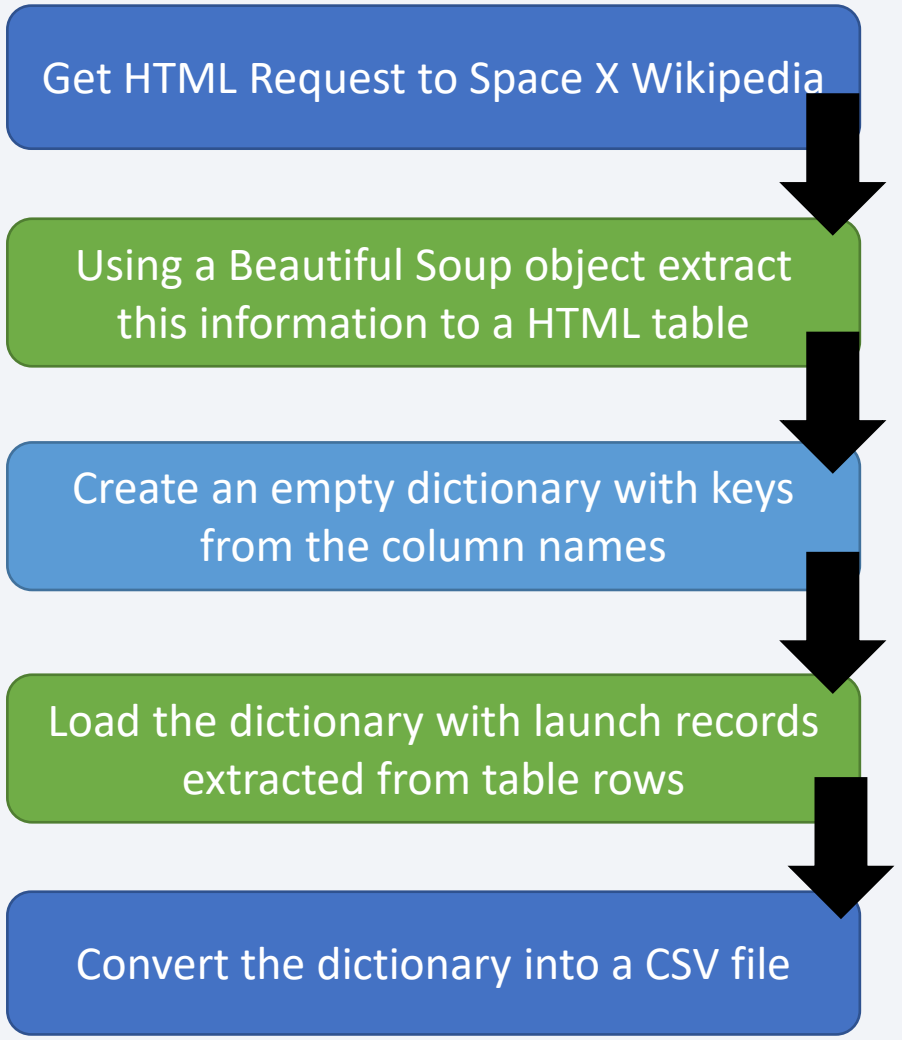
- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/CollectingData.ipynb



Data Collection - Scraping

- Web scraping from Wikipedia page for Falcon 9 historical launch information

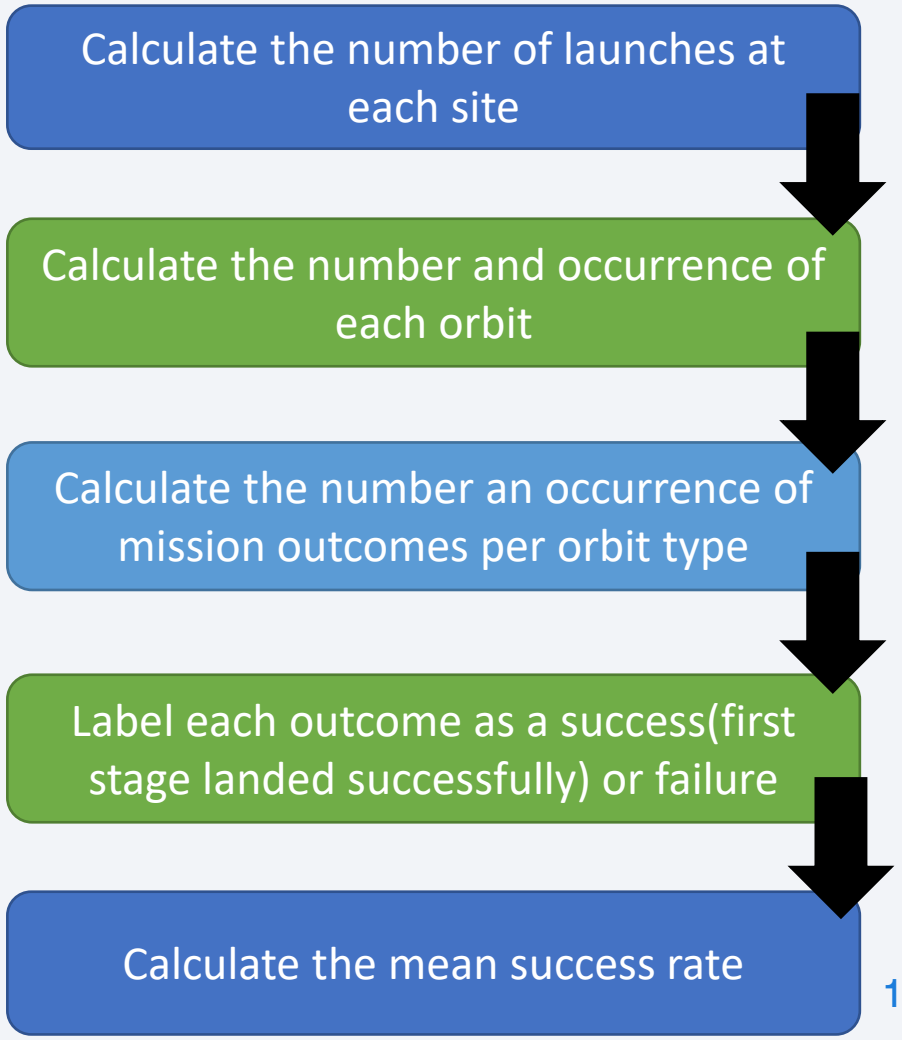
- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/Webscaping.ipynb



Data Wrangling

- Exploratory Data Analysis performed to look for patterns in the data set

- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/DataWrangling.ipynb



EDA with Data Visualization

- Exploratory Data Analysis performed using visualization
  - Scatter plot to analyze relationships between independent and dependent variables – Flight number vs Payload Mass, Flight Number vs Launch Sites, Payload and Launch Sites, Flight Number and Orbit Type, Payload and Orbit Type
  - Bar Chart for categorical data success rate of each orbit
  - Line plot for success rate over time (date)

- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/EDAwithVisualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where the launch site begins with ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass varied by booster version F9 v1.1
- List the date when the first successful landing outcome on the ground pay was achieved
- List the names of the boosters which have success in drop ship and have payload mass between 4,000 and 6,000 kg
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drop ship, their booster versions, and launch site names for 2015
- Rank the count of landing outcomes (such as failure (drone ship) or success (ground pad)) between 6/4/2010 and 3/20/2017 in descending order
- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/EDAwithSQL.ipynb

Build an Interactive Map with Folium

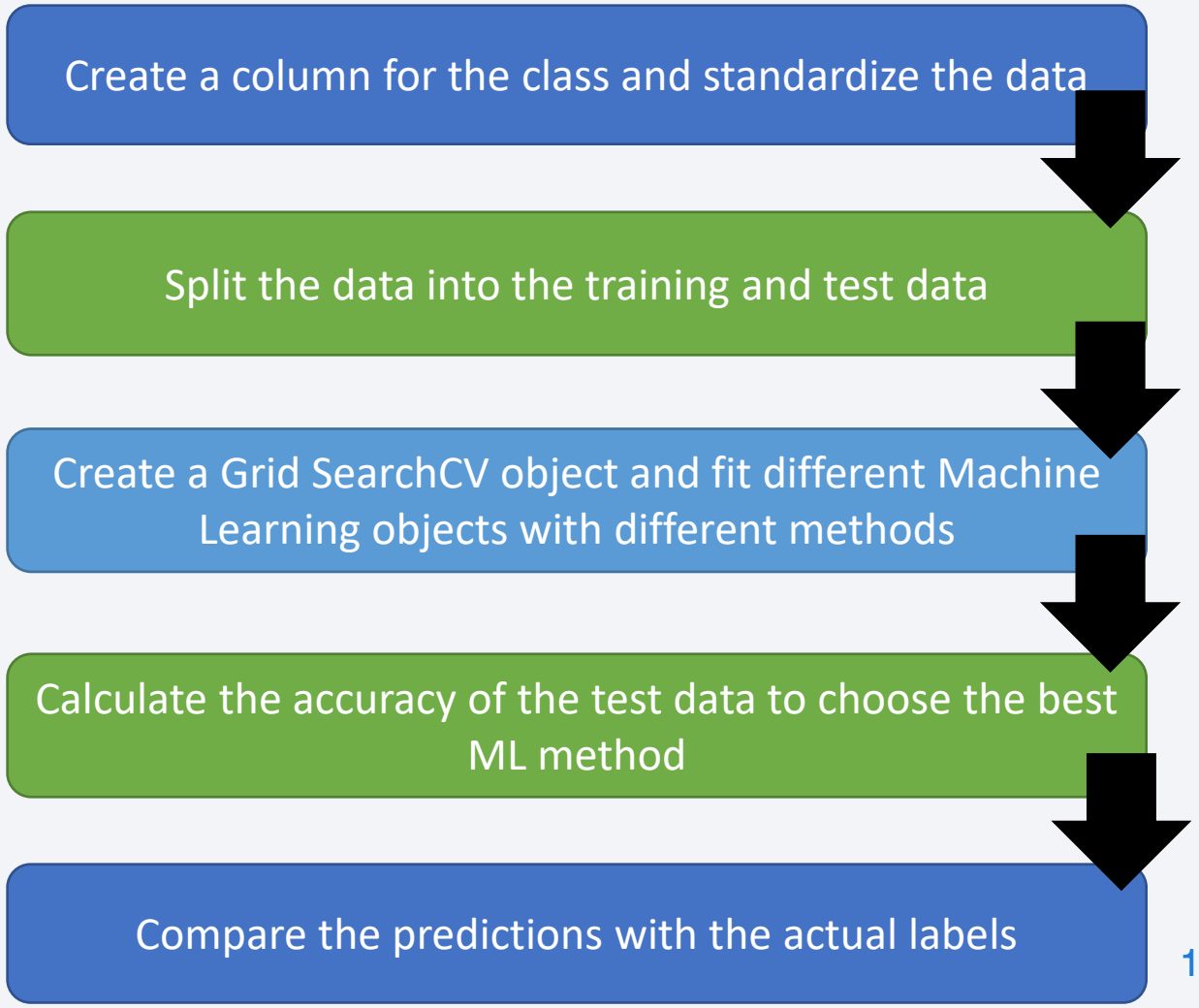
- Folium markers were used to show the Space X launch sites and their nearest important landmarks like railways, highways, cities, and coastlines.
- Polylines were used to connect the launch sites to their nearest land marks.
  - Red represents the locations of rocket launch failures
  - Green represents the locations of rocket launch successes
- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/Visual AnalyticswithFolium.ipynb

Build a Dashboard with Plotly Dash

- Pie charts and scatter plots were used to visualize the launch records of Space X on a Dashboard
- The pie chart displayed the rocket launch success rate per launch site and were interactive.
- Scatter plots were used to analyze certain features. These plots help us to understand and visualize what may have lead to the success rate at each site based on payload mass and booster versions.

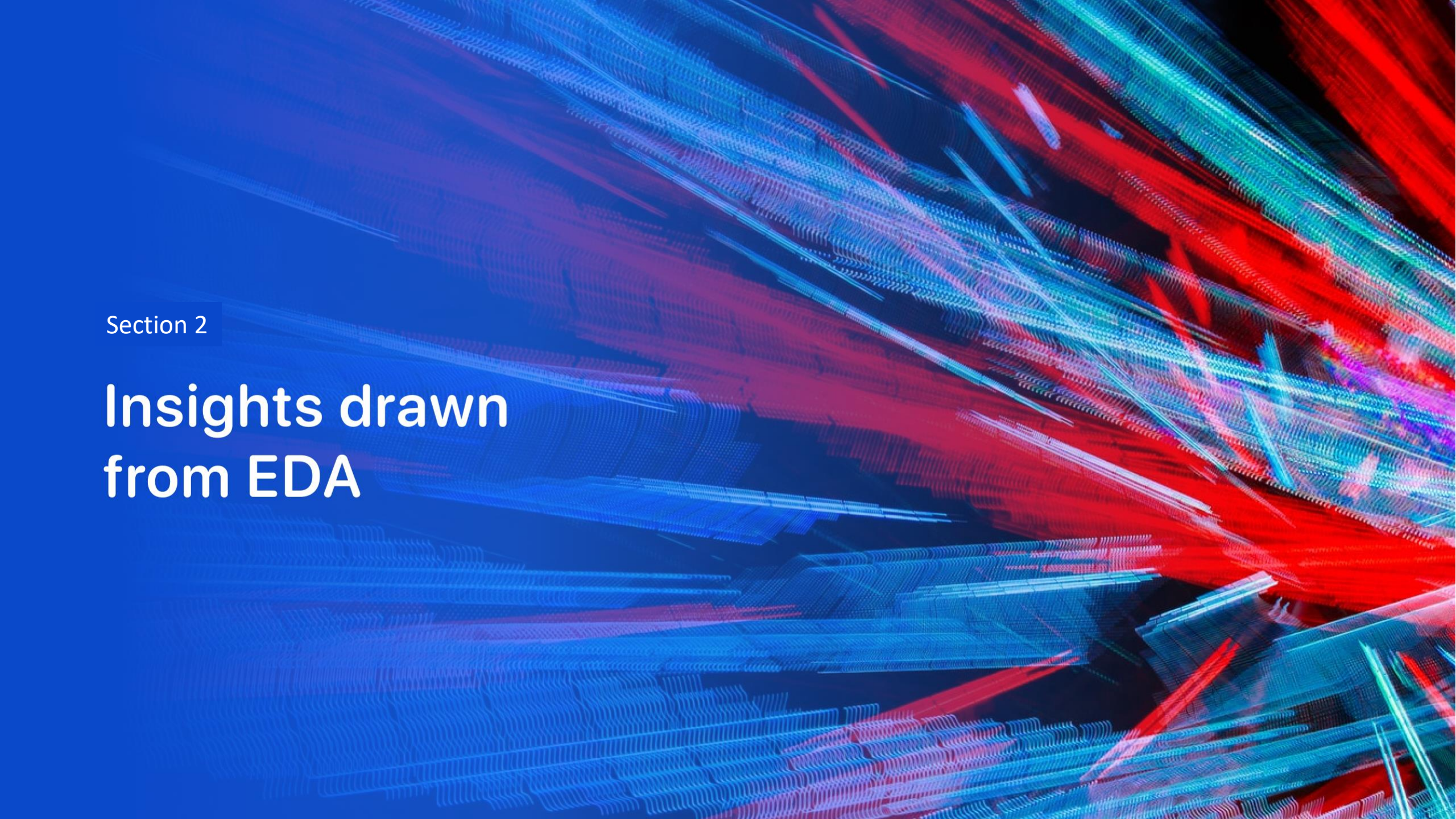
Predictive Analysis (Classification)

- Scikit-learn is a Machine Learning library that was used for predictive analysis.
- A predictive analysis was run to see if the first stage would land given the data
- Github completed notebook: https://github.com/ssiedlik/FinalCapstoneSpaceYProject/blob/main/MachineLearningPrediction.ipynb

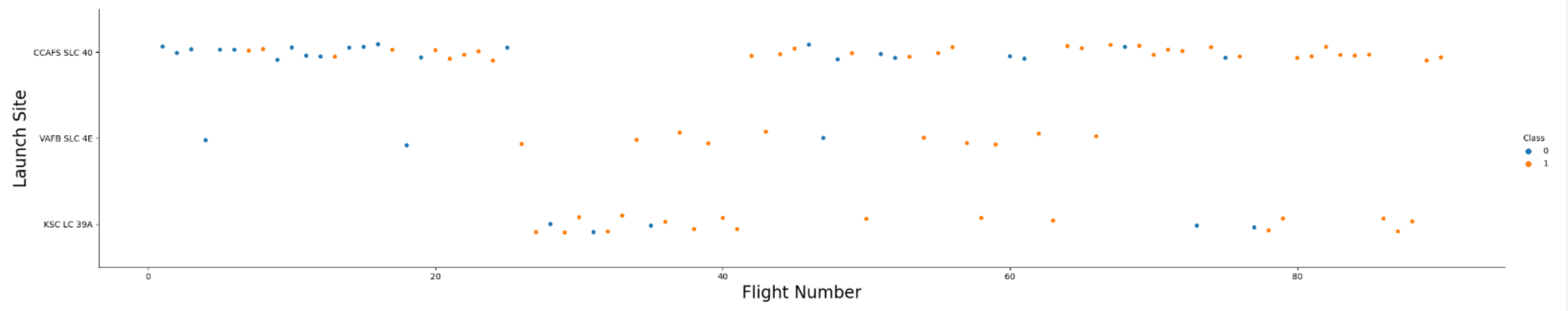


Results

- The exploratory data analysis has show that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line perhaps making it easier to test rocket landings in the water.
- Sites are also located near highways and railways. This may help with the transport of equipment.
- The machine learning algorithm was able to predict the landing success of rockets with an accuracy of 83.33%

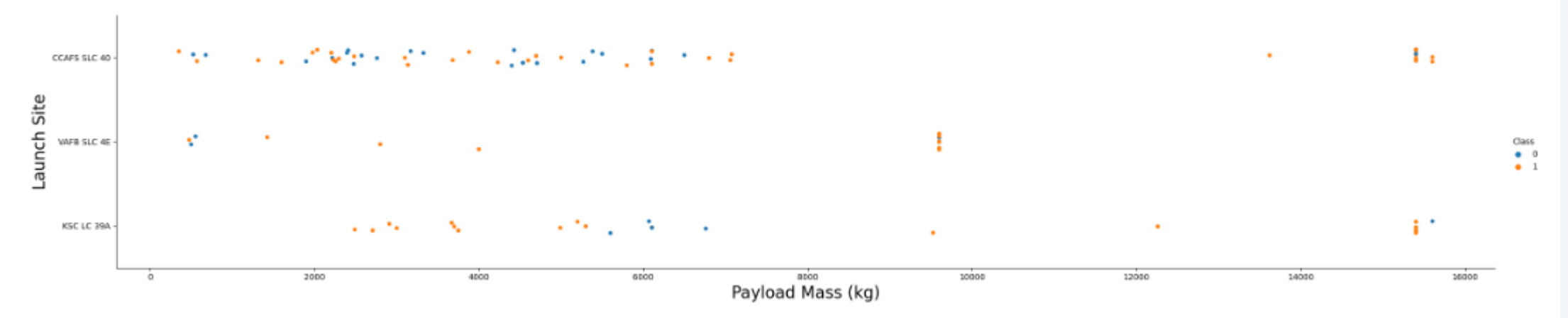

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



As flight numbers increased the flights were more successful. Successful landings are shown in orange while unsuccessful landings are shown in blue. Launch site CCAFS SLC 40 had the most number of landings.

Payload vs. Launch Site

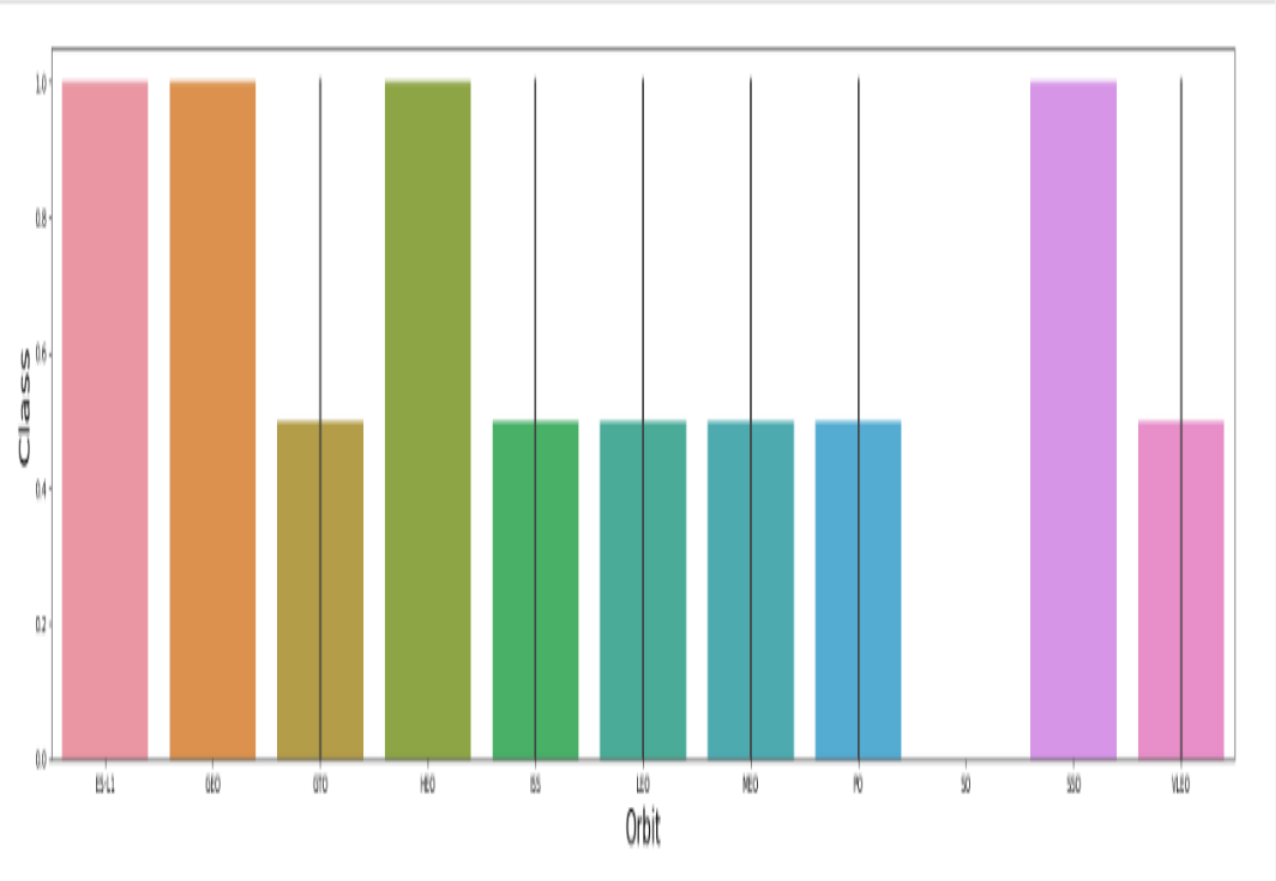


As you can see the launch site VAFB-SLC had no launches above a payload mass of 10,000 kg. Successful landings are shown in orange while unsuccessful landings are shown in blue.

Success Rate vs. Orbit Type

Orbits that had the highest success rate:

- ES-L1
- GEO
- HEO
- SSO

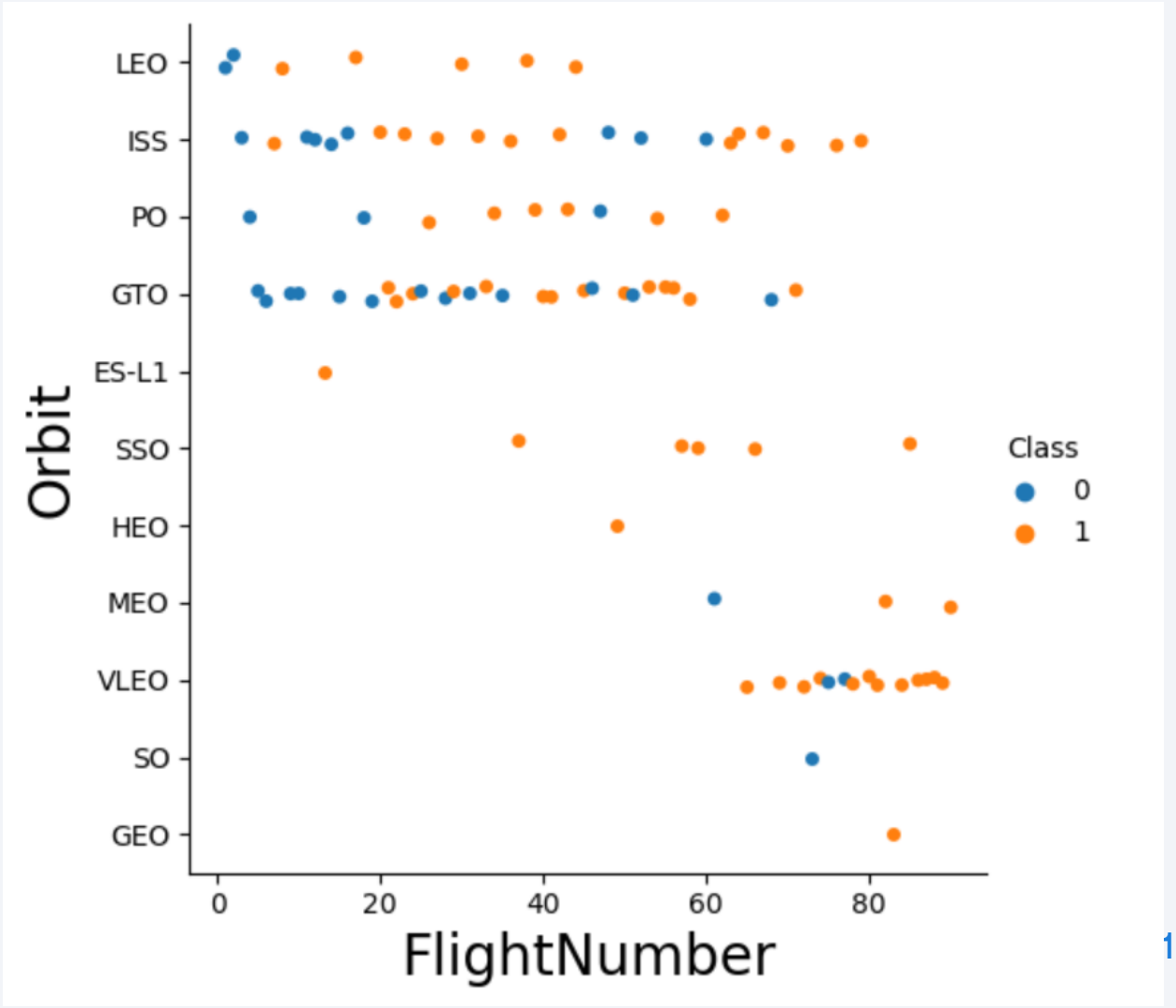


Flight Number vs. Orbit Type

A scatter plot of flight number vs. orbit type is shown.

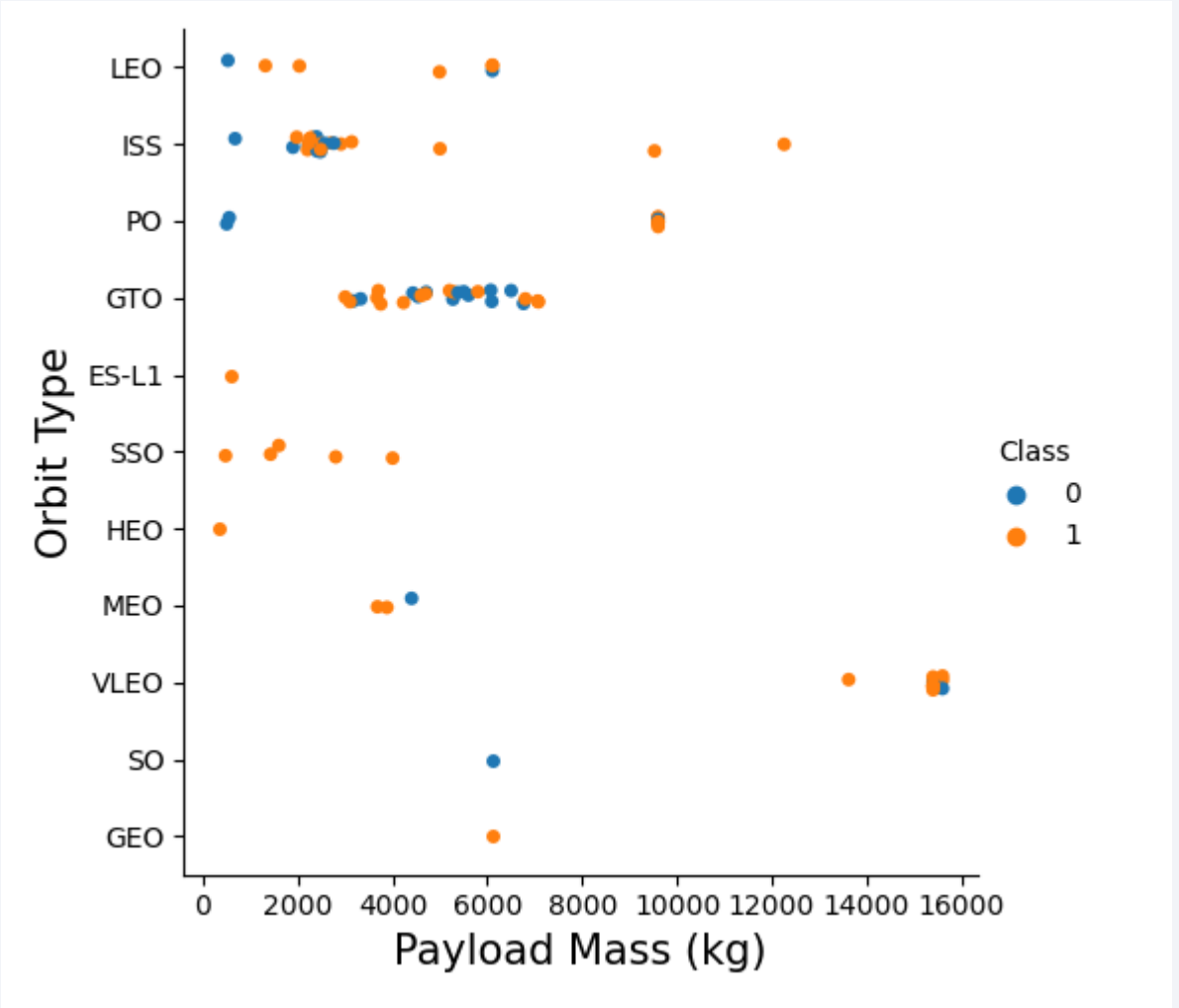
The LEO orbit success appears replated to the number of flights.

There seems to be no relationship between flight number and the GTO orbit.



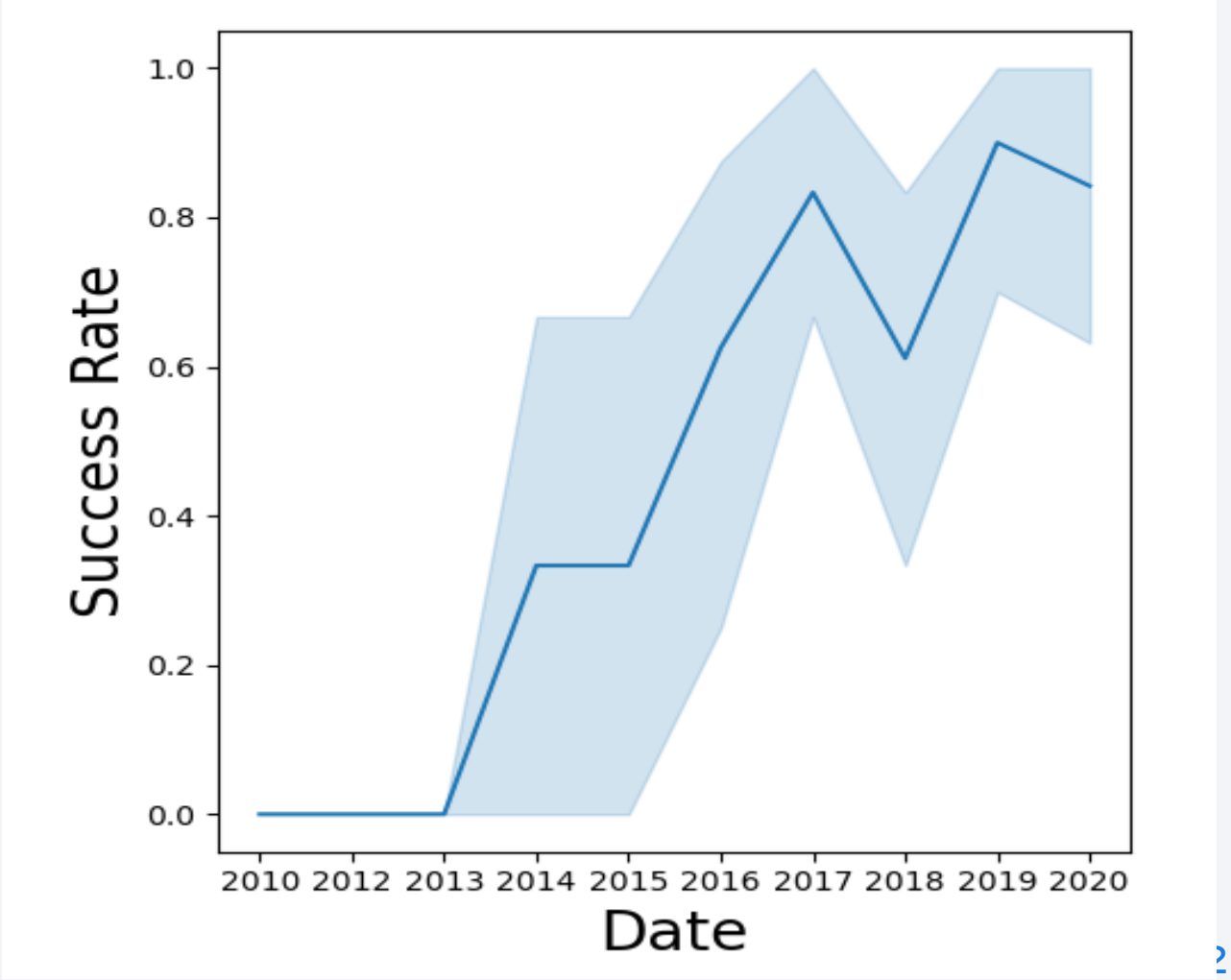
Payload vs. Orbit Type

- With heavier payloads the successful landings are more for Polar, LEO, and ISS
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The line plot shows the success rate since 2013 kept increasing until 2020



All Launch Site Names

- The unique launch sites where rocket launches were attempted were:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4e

```
%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'

The first 5 records where launch sites begin with `CCA` were all from CCAFS LC-40.
As you can see other companies other than Space X were testing their rockets

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) was 45,596 kg

```
%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 was 2,928.4

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The first successful ground landing date was December 22, 2015

```
%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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 between 4,000 and 6,000 kg were:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2

```
%sql SELECT BOOSTER_VERSION from SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' and  PAYLOAD_MASS__KG_  >4000 and PAYLOAD_MASS__KG_ <6000;
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful missions was 98 and the total number of failures was 1.

```
]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Failure (in flight)';
 * sqlite:///my_data1.db
Done.
]: count(MISSION_OUTCOME)
                        1

]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success';
 * sqlite:///my_data1.db
Done.
]: count(MISSION_OUTCOME)
                       98
```

Boosters Carried Maximum Payload

- Twelve boosters carried the maximum payload from the payload mass data.

```
%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

- Two boosters F9 v1.1B1012_CCAFS LC-40 and F9v1.1B105 CCAFS LC-40 failed to land in 2015

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The number of successful landings have increased since 2015

```
%sql select * from SPACEXTBL where Landing__Outcome = 'Success (ground pad)' or and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 3

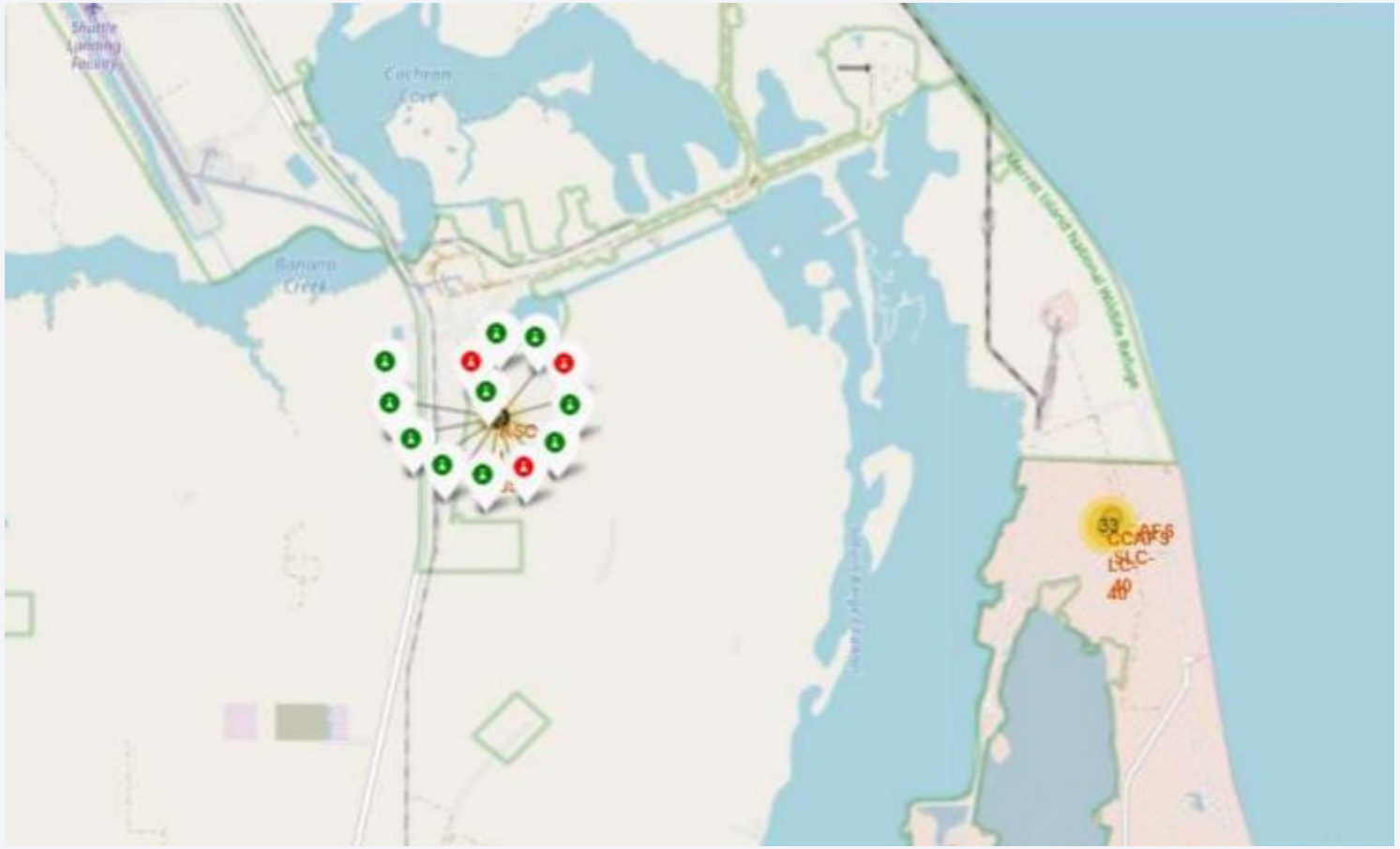
Launch Sites Proximities Analysis

Launch Site Locations

Launch sites are near the coast and are a couple thousand kilometers away from the equator

Success Rates of Rocket Launches

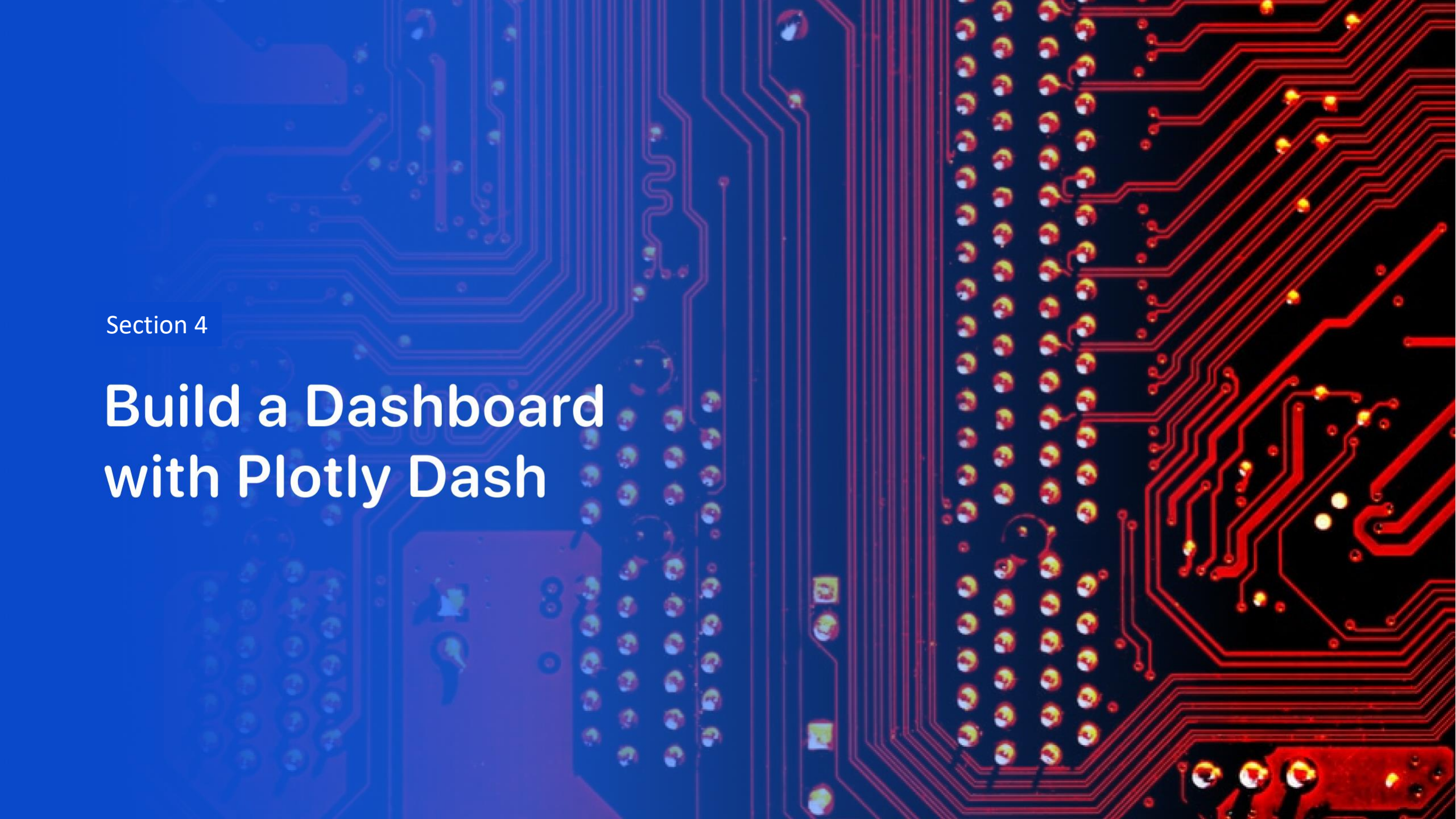
The successful launches are represented in green while red represents the failed rocket launches.



Surrounding Landmarks

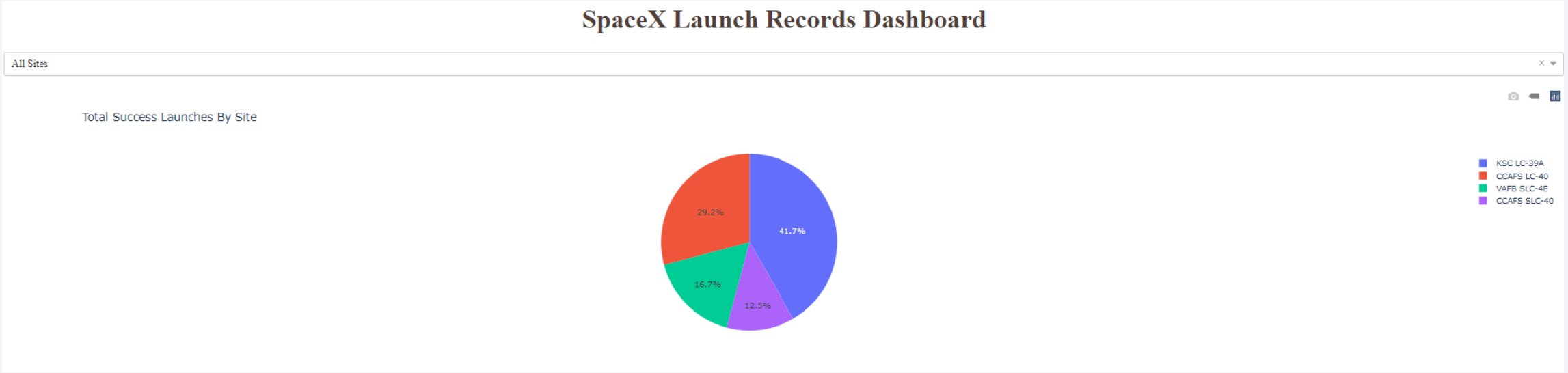
- It appears that launch sites are at least 18 km away from cities in order to prevent crashes near heavily populated areas.
- Launch sites are close to railways and highways perhaps due to easy transport of materials.
- Close to the ocean.

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue


Section 4
Build a Dashboard with Plotly Dash

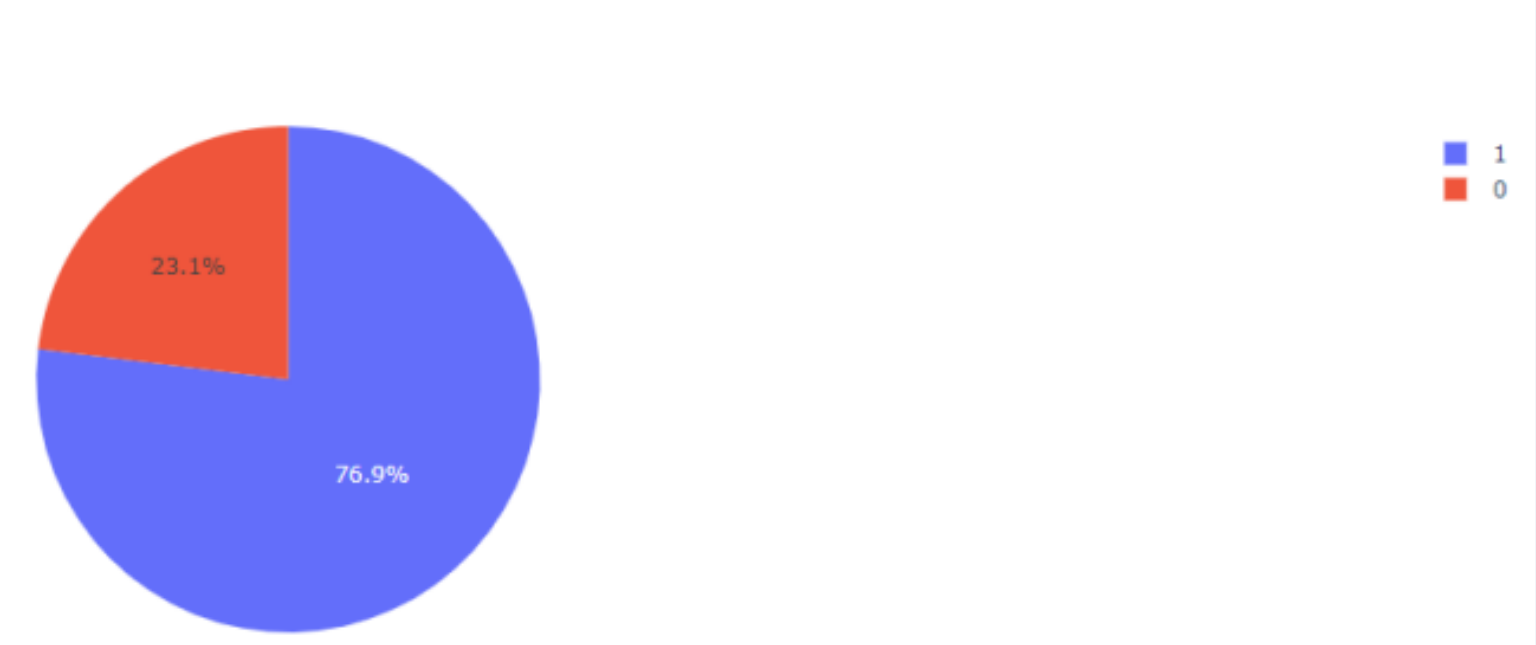
Launch Success by Site

The most successful launch site was KSC LC-39A



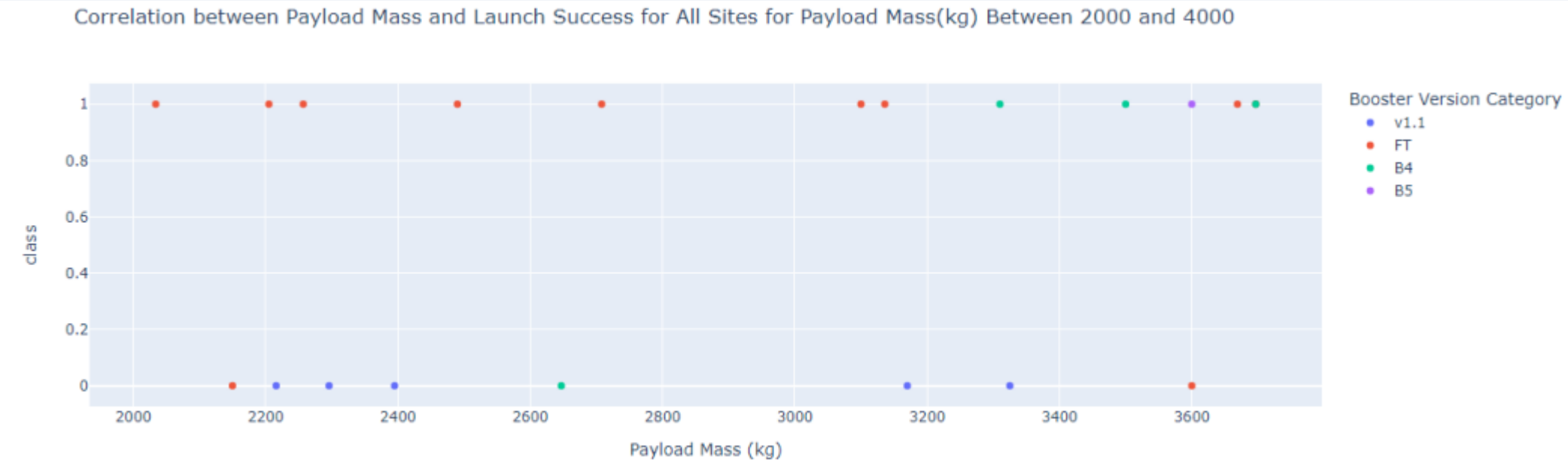
Success Rate for KSC LC-39A

The success rate for KSC LC-39A was 76.9%



Payload vs. Launch Outcome Scatter Plot

- This plot is controlled by a slider. It appears that the payload between 2000 kg and 4000 kg has the highest success rate.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

All machine learning methods have the same accuracy score of 83.33%. Therefore a Logistic Regression was used for classification.

Find the method performs best:

In [28]:

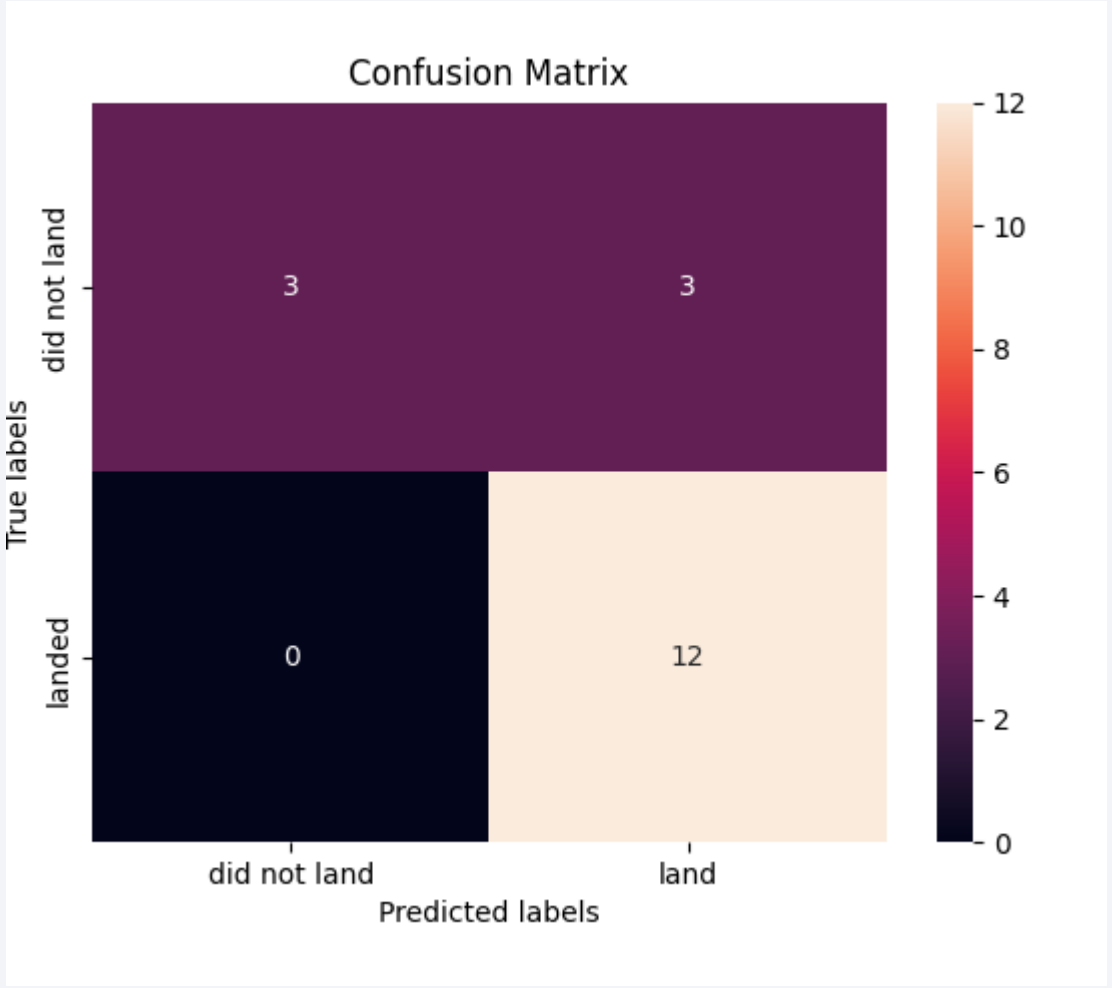
```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

Out[28]:

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

Confusion Matrix

The confusion matrix of Logistic Regression reveals that the major problem is false positives failing to accurately predict 3 labels.



Conclusions

To be competitive with Space X with a new company Space Y the following must be considered:

- All launch sites are located near the coast and transportation options like railways. The launch sites are also a good distance away from cities.
- Site KSC LC-39A had the highest launch success rate.
- Since 2015 the success rate has significantly increased which was why it was correlated to flight number.
- This data was used to train a machine learning model that was able to predict the success rate of landing outcomes with an 83.3% accuracy.

Thank you!