DATA SCIENCE CAPSTONE

Lawrence Mak

OUTLINE

EXECUTIVE SUMMARY

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

INTRODUCTION

Background and Context

SpaceX offers Falcon 9 rocket launches at a cost of $62 million, much lower than competitors due to their ability to reuse the first stage. Predicting the success of the first stage landing is crucial for estimating launch costs and competing bids. This project aims to create a machine learning pipeline for predicting successful first stage landings.

Questions to be answered

What variables affect a successful landing?

How does the interaction amongst these variables determine the success rate

What is the best algorithm for binary classification

METHODOLOGY

THE OVERALL METHODOLOGY INCLUDES:

1. DATA COLLECTION, WRANGLING, AND FORMATTING, USING:

- SPACEX API
- WEB SCRAPING

2. EXPLORATORY DATA ANALYSIS (EDA), USING:

- PANDAS AND NUMPY
- SQL

3. DATA VISUALIZATION, USING:

- MATPLOTLIB AND SEABORN
- FOLIUM
- DASH

4. MACHINE LEARNING PREDICTION, USING

- LOGISTIC REGRESSION
- SUPPORT VECTOR MACHINE (SVM)
- DECISION TREE
- K-NEAREST NEIGHBORS (KNN)

DATA COLLECTION – SPACEX API

1. Request rocket launch data from SpaceXAPI

2. Decode the response content using.json() and turning it into a dataframe using .json_normalize()

3. Request needed information about the launches from SpaceXAPI by applying custom functions

4. Construct obtained data into a dictionary

5. Create a dataframe from the dictionary

6. Filter the data frame to only include Falcon 9 launches

7. Replace missing values of Payload Mass column with calculated.mean() for this column

8. Export the data to CSV

GITHUB

DATA COLLECTION – WEB SCRAPING

1. Request Falcon 9 launch data from Wikipedia

2. Create a BeautifulSoup object from the HTML response

3. Extract all column names from the HTML table header

4. Collect the data by parsing HTML tables

5. Construct data obtained data into a dictionary

6. Create a dataframe from the dictionary

7. Export the data to CSV

GITHUB

DATA WRANGLING

Data Wrangling:

- The data is processed to handle missing entries and encode categorical features using one-hot encoding.
- An additional column named 'Class' is introduced to the dataframe, where:
  - 'Class' contains 0 if a launch is failed and 1 if it is successful.
- The resulting dataset comprises 90 rows (instances) and 83 columns (features).

In the dataset, various scenarios depict the outcome of booster landings:

- True Ocean: Successful landing in a specific region of the ocean.
- False Ocean: Unsuccessful landing in a specific region of the ocean.
- True RTLS: Successful landing on a ground pad.
- False RTLS: Unsuccessful landing on a ground pad.
- True ASDS: Successful landing on a drone ship.
- False ASDS: Unsuccessful landing on a drone ship.
- These outcomes are primarily transformed into Training Labels:
  - "1" indicates a successful booster landing.
  - "0" indicates an unsuccessful landing.

EDA WITH PANDAS AND NUMPY

Charts were plotted to visualize various relationships:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Success Rate Yearly Trend

Scatterplots illustrate the relationship between variables, which could be utilized in a machine learning model if a relationship exists.

Bar charts provide comparisons among discrete categories, aiming to demonstrate the relationship between specific categories and a measured value.

Line charts depict trends in data over time (time series).

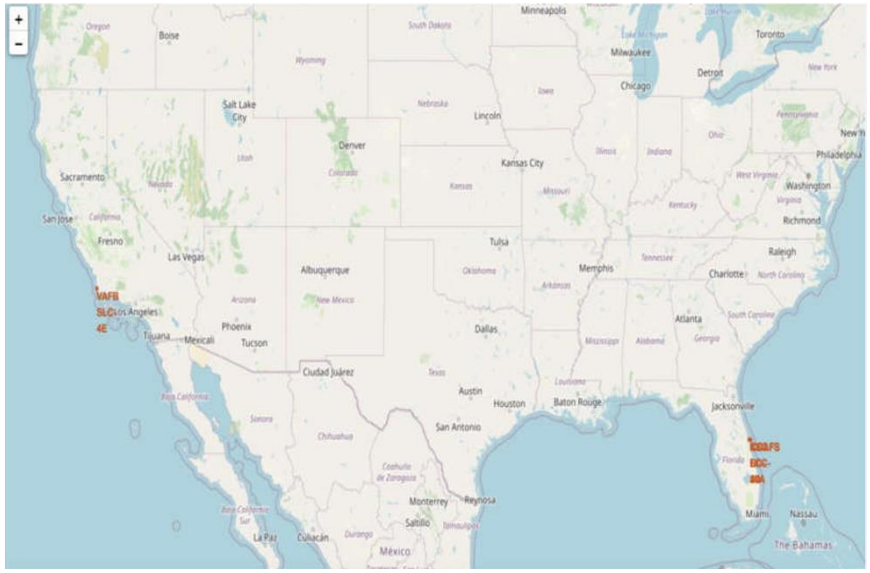
EDA WITH SQL

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in droneship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass
- Listed the failed landing outcomes in droneship, their booster versions, and launch site names for the months in year 2015
- Ranked the count of landing outcomes (such as failure (droneship) or success (groundpad)) between the date 2010-06-04 and 2017-03-20 in descending order

INTERACTIVE MAP WITH FOLIUM

- All launch sites were marked, and map objects like markers, circles, and lines were added to indicate launch success or failure for each site on the folium map.
- Launch outcomes (failure or success) were assigned to class 0 and 1, with 0 denoting failure and 1 denoting success.
- Utilizing color-labeled marker clusters, launch sites with relatively high success rates were identified.
- Distances between launch sites and nearby features were calculated to address questions such as:
  - Proximity to railways, highways, and coastlines.
  - Distance maintained from cities.

ALL LAUNCH SITES ON A MAP

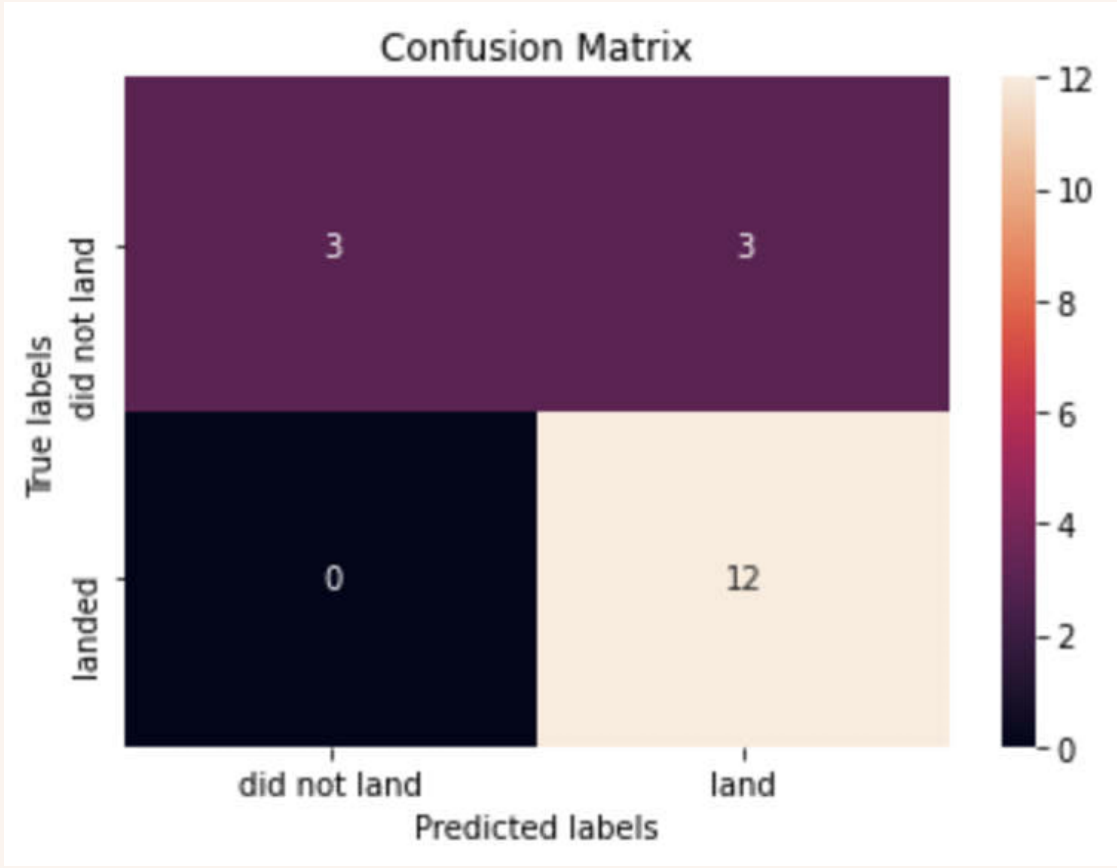


BUILDING A DASHBOARD WITH PLOTLY

- Launch Sites Dropdown List:
  - A dropdown list has been added to enable selection of launch sites.
- Pie Chart Showing Success Launches (All Sites/Certain Site):
  - A pie chart has been included to display the total count of successful launches for all sites and the success versus failed counts for the selected site, if a specific launch site was chosen.
- Slider of Payload Mass Range:
  - A slider has been implemented to select the payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the Different Booster Versions:
  - A scatter chart has been added to illustrate the correlation between payload and launch success for various booster versions.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- The data was loaded using numpy and pandas, transformed, and split into training and testing sets.
- Various machine learning models were constructed, and different hyperparameters were tuned using GridSearchCV.
- Accuracy served as the metric for the model, which was enhanced through feature engineering and algorithm tuning.
- The best-performing classification model was identified.

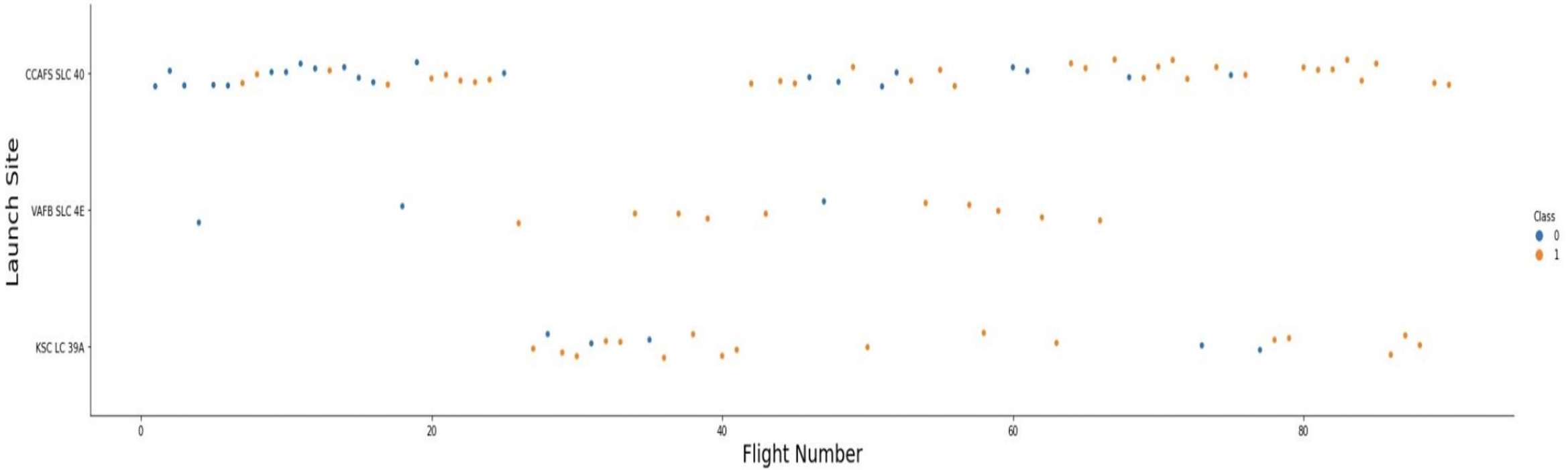


Decision Tree:

GridSearchCV best score: 0.8892857142857142
Accuracy score on test set: 0.8333333333333334

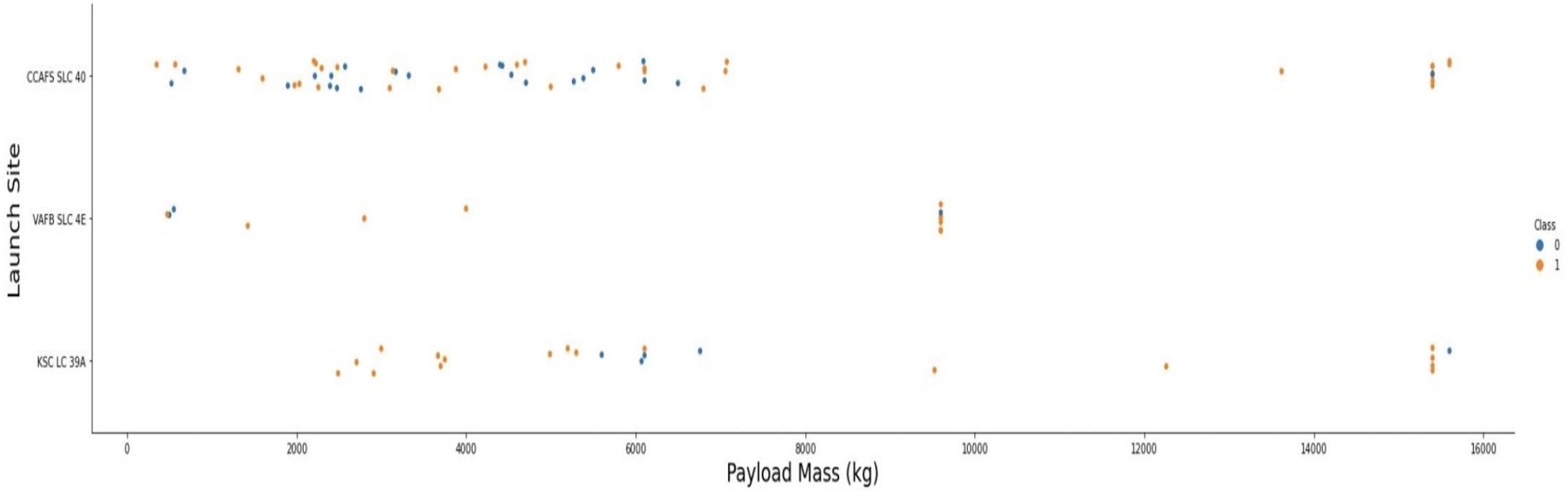
SECTION 2: INSIGHTS DRAWN FROM EDA VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE



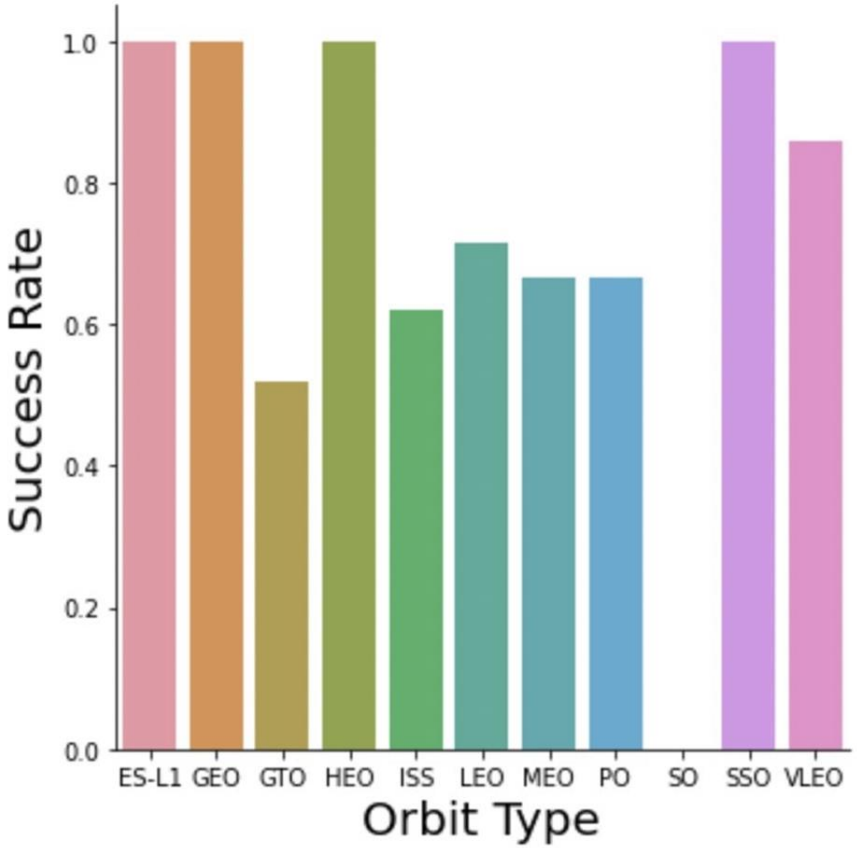
- Initial flights experienced failures, whereas recent flights achieved success.
- Approximately half of all launches occur at the CCAFS SLC-40 launch site.
- VAFB SLC-4E and KSC LC-39A demonstrate higher success rates.
- It is reasonable to expect that each new launch exhibits a higher success rate.

PAYLOAD VS. LAUNCH SITE



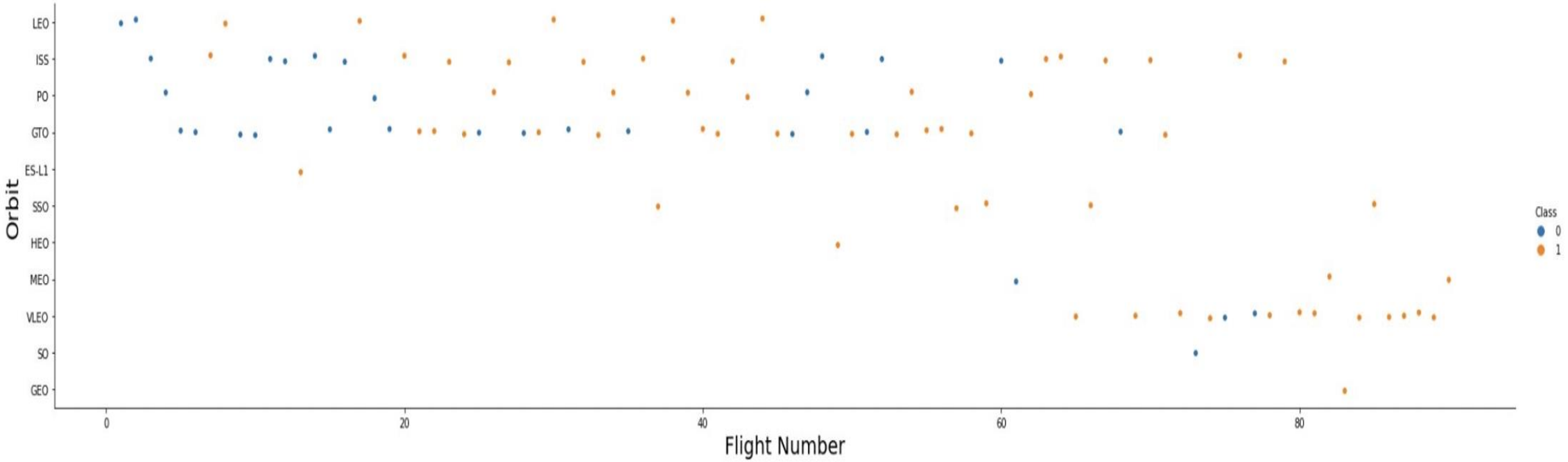
- For each launch site, success rates increase with higher payload mass.
- The majority of launches with payload mass over 7000kg resulted in success.
- KSC LC-39A achieved a 100% success rate for payload mass under 5500kg as well.

SUCCESS RATE VS. ORBIT TYPE



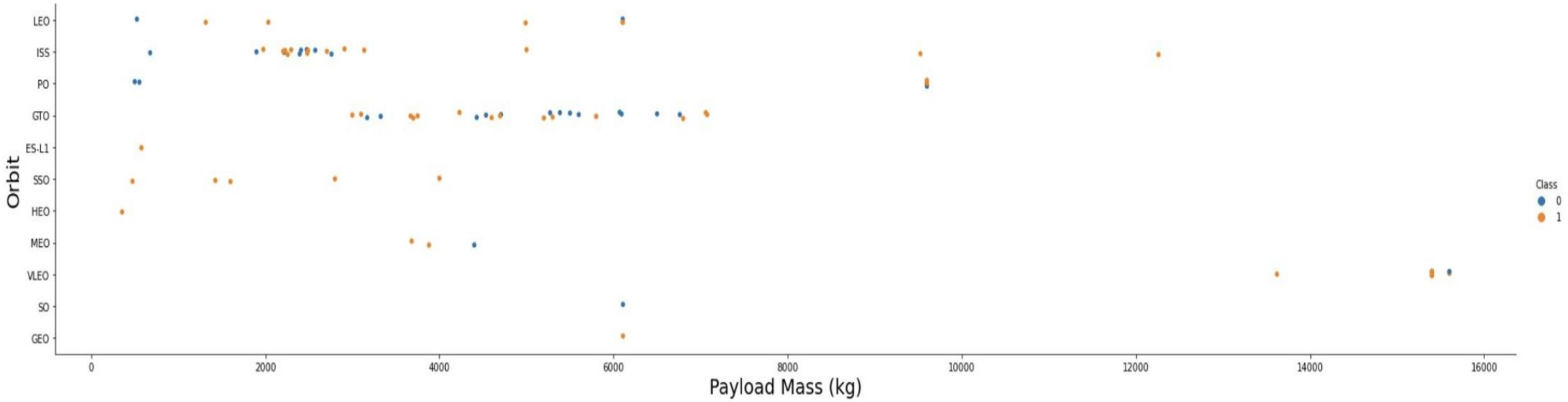
From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the highest success rate.

FLIGHT NUMBER VS. ORBIT TYPE



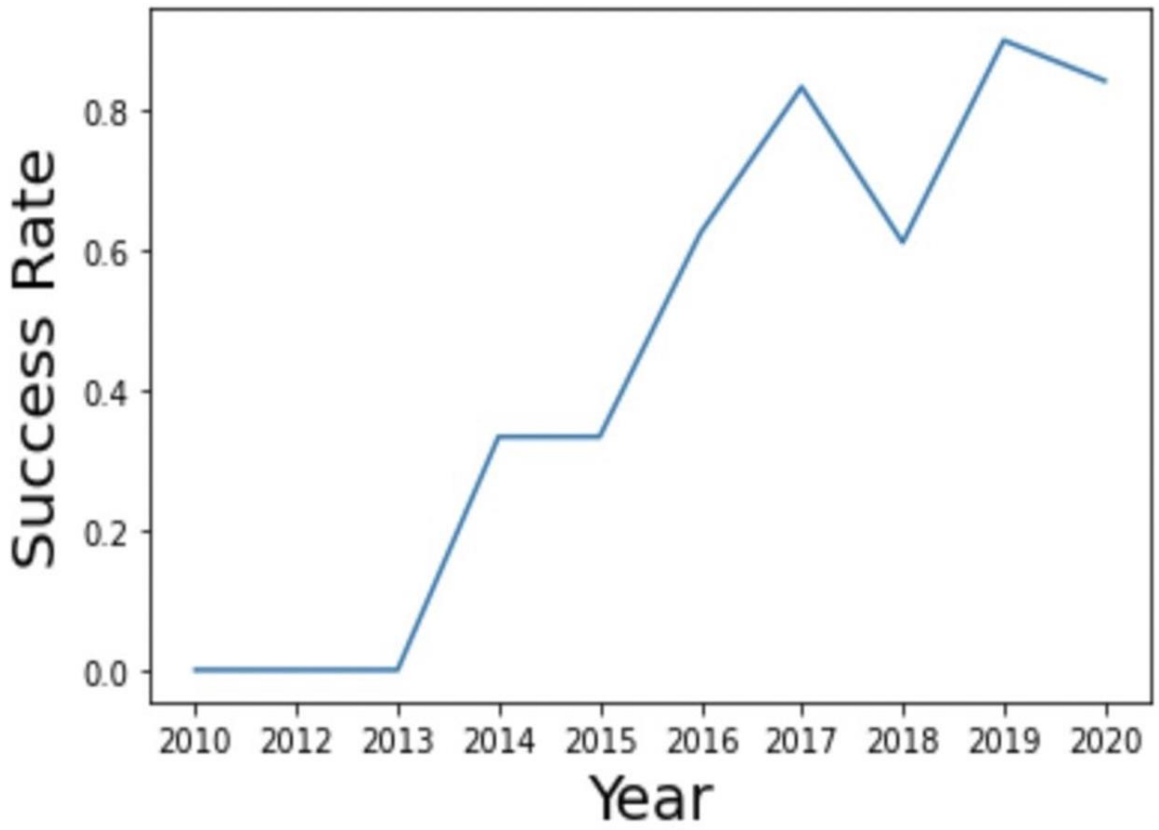
The diagram illustrates the relationship between Flight Number and Orbit type. It is evident that in the LEO orbit, success tends to correlate with the number of flights. However, in the GTO orbit, there is no discernible relationship between flight number and orbit.

PAYLOAD MASS VS. ORBIT TYPE



Heavy payloads exert a negative influence on GTO orbits but have a positive impact on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND



The success rate rises overtime as more improvements on the technology are made

SECTION 3: INSIGHTS DRAWN FROM EDA SQL

LAUNCH SITE NAMES THAT BEGIN WITH 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.data
98/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Use the query to display 5 records where the launch site names begins with 'CCA'

GITHUB

TOTAL PAYLOAD MASS

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_payload_mass
45596

Use the query to display total payload mass by boosters launched by NASA

GITHUB

AVERAGE PAYLOAD MASS BY F9 V1.1

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

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

average_payload_mass
2534

Use the query to display average payload mass by Booster version F9 V1.1

GITHUB

FIRST SUCCESSFUL GROUND LANDING DATE

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pa
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

first_successful_landing
2015-12-22

Use the query to display the first successful ground landing date

GITHUB

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

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

Use the query to display the names of boosters that had a successful landing with a payload mass between 4000 and 6000

GITHUB

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Use the query to display the amount of successful and non-successful missions

GITHUB

BOOSTERS CARRIED MAXIMUM PAYLOAD

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPAC
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
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

Use the query to display the names of boosters which have successfully carried the maximum payload

GITHUB

2015 LAUNCH RECORDS

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
08/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Use the query to display a list of the failed landings in drone ship, their booster versions, and launch site names for the year 2015

GITHUB

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

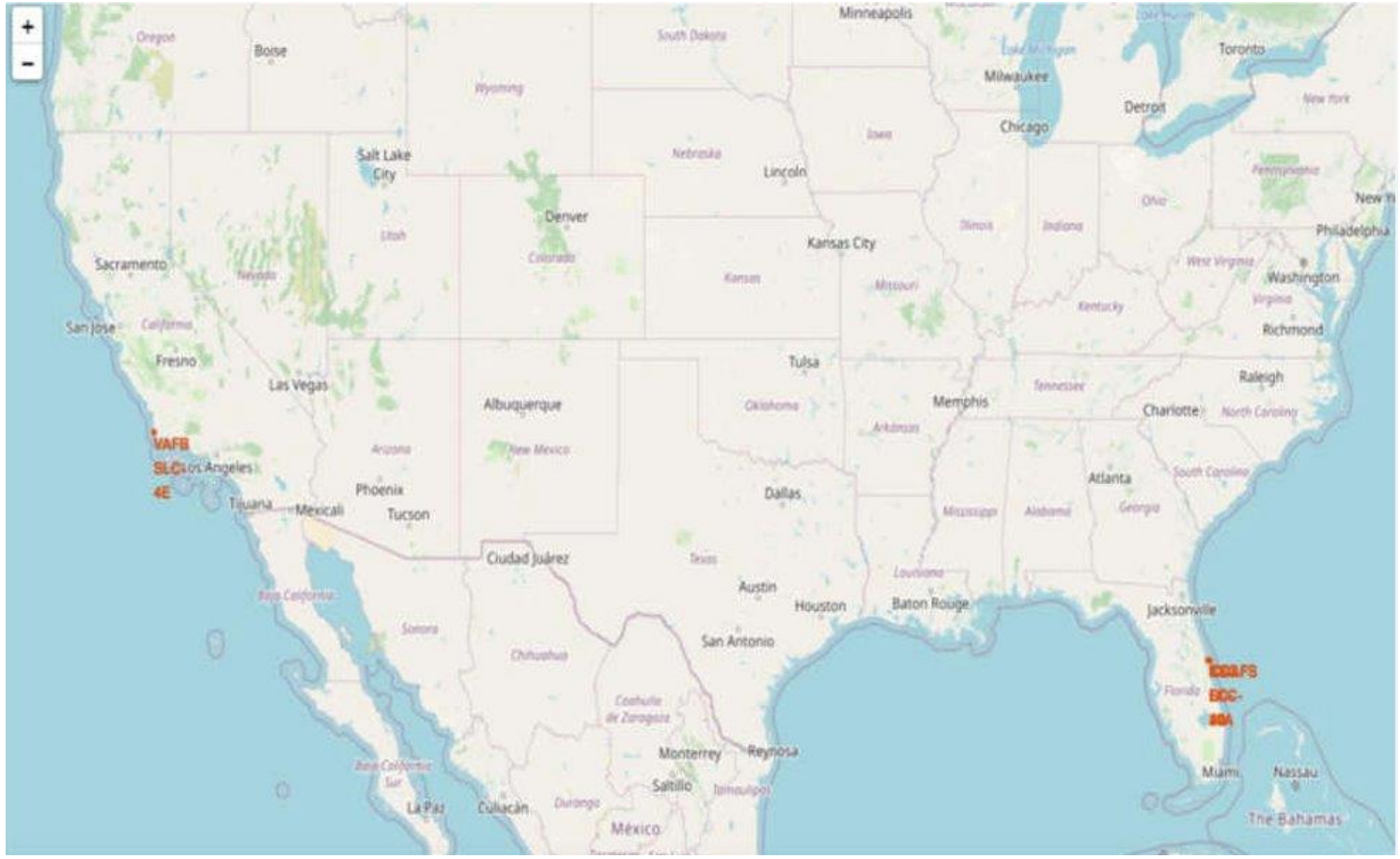
Landing outcomes and the count of landing outcomes were selected from the data, and the WHERE clause was utilized to filter for landing outcomes between June 4, 2010, and March 20, 2010.

The GROUP BY clause was applied to group the landing outcomes, and the ORDER BY clause was used to arrange the grouped landing outcome in descending order.

GITHUB

SECTION 4: INTERACTIVE MAP WITH FOLIUM

ALL LAUNCH SITE LOCATIONS ON A GLOBAL MAP

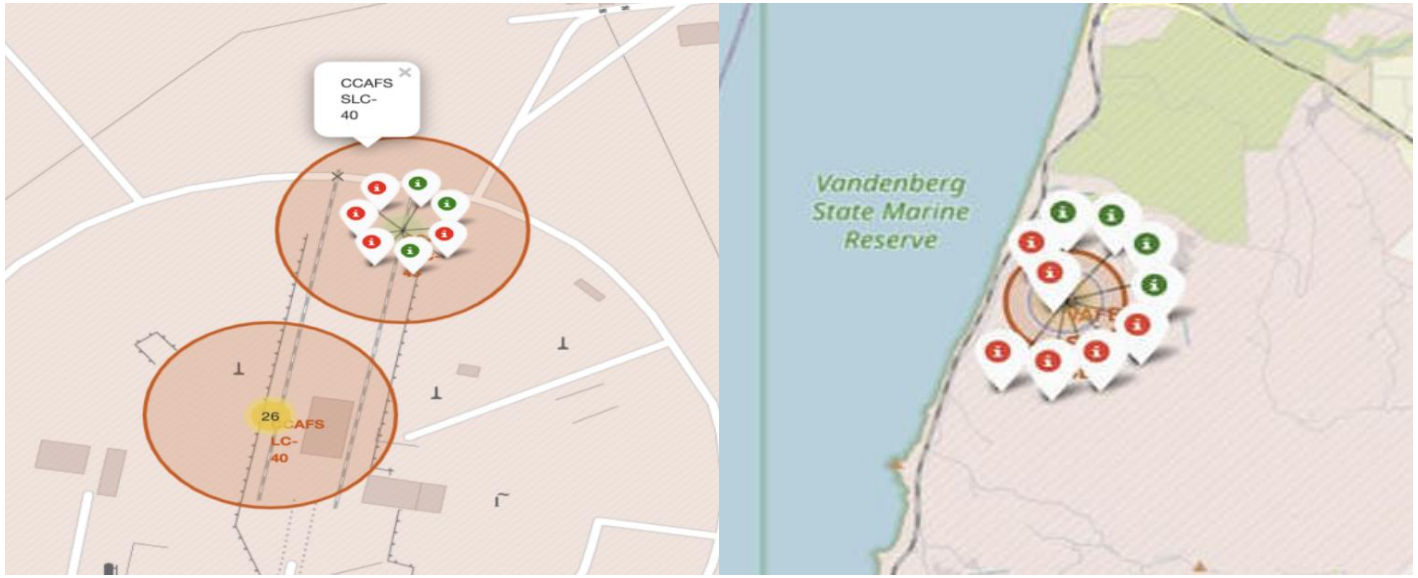


Most launch sites are situated in proximity to the Equator line. This location is advantageous because the land moves faster at the equator than at any other point on the Earth's surface, with a speed of approximately 1670 km/hour. When a spacecraft is launched from the equator, it inherits this velocity, aiding it in achieving and maintaining orbit due to inertia.

Additionally, all launch sites are positioned very close to the coast. Launching rockets towards the ocean minimizes the risk of debris falling or exploding near populated areas.

COLOR LABELLED LAUNCH RECORDS ON THE MAP

SUCCESSFUL AND FAILED LAUNCH SITES



From the color-labeled markers, we can readily identify launch sites with relatively high success rates:

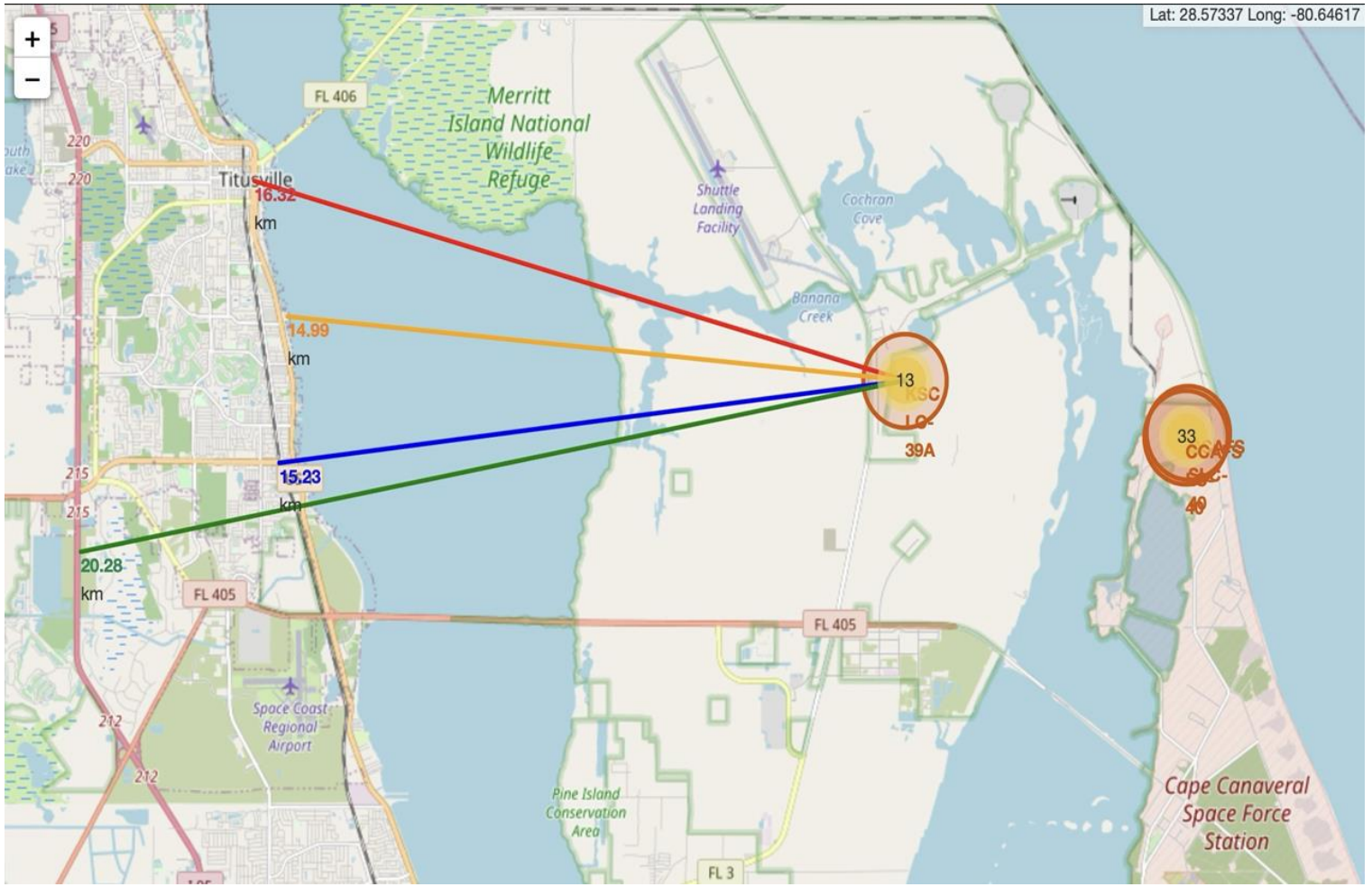
- Green Marker = Successful Launch
- Red Marker = Failed Launch Launch Site KSC LC-39A exhibits a very high success rate.

THE DISTANCE BETWEEN KSC LC-39A TO ITS PROXIMITIES

Visual analysis of launch site KSC LC-39A reveals its proximity to various landmarks:

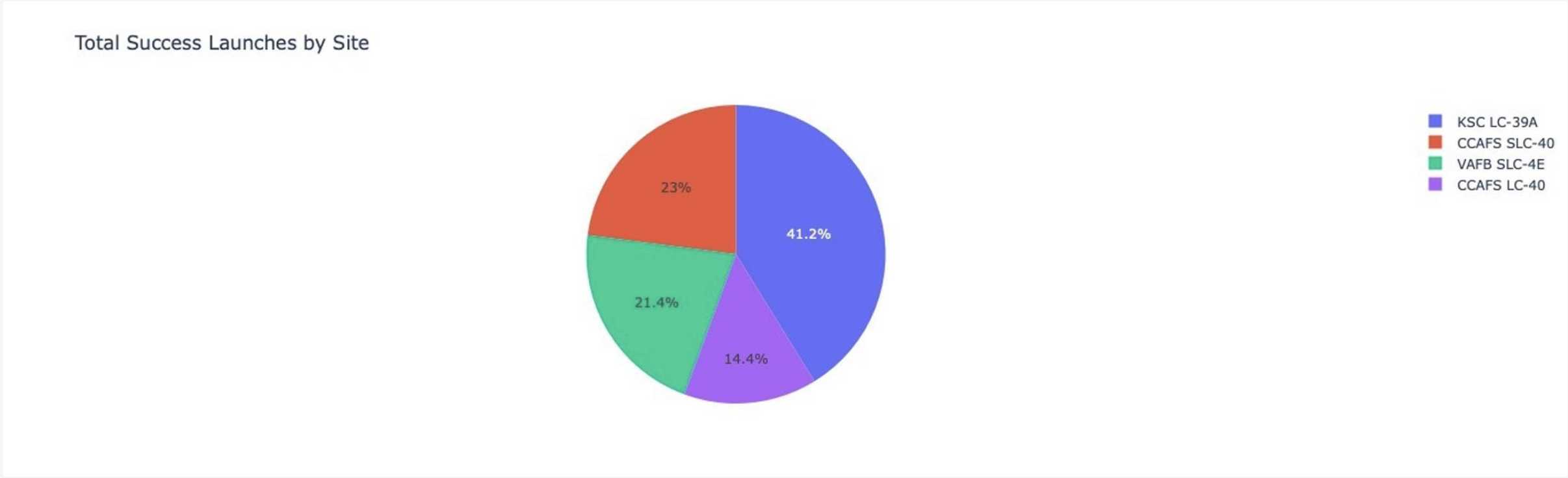
- It is relatively close to a railway, approximately 15.23 km away.
- It is also near a highway, about 20.28 km away.
- The coastline is nearby, with a distance of approximately 14.99 km.
- Additionally, the launch site is relatively close to its nearest city, Titusville, at a distance of around 16.32 km.

Considering the high speed of a failed rocket, which can cover distances of 15-20 km in just a few seconds, there is potential danger to populated areas in the vicinity.



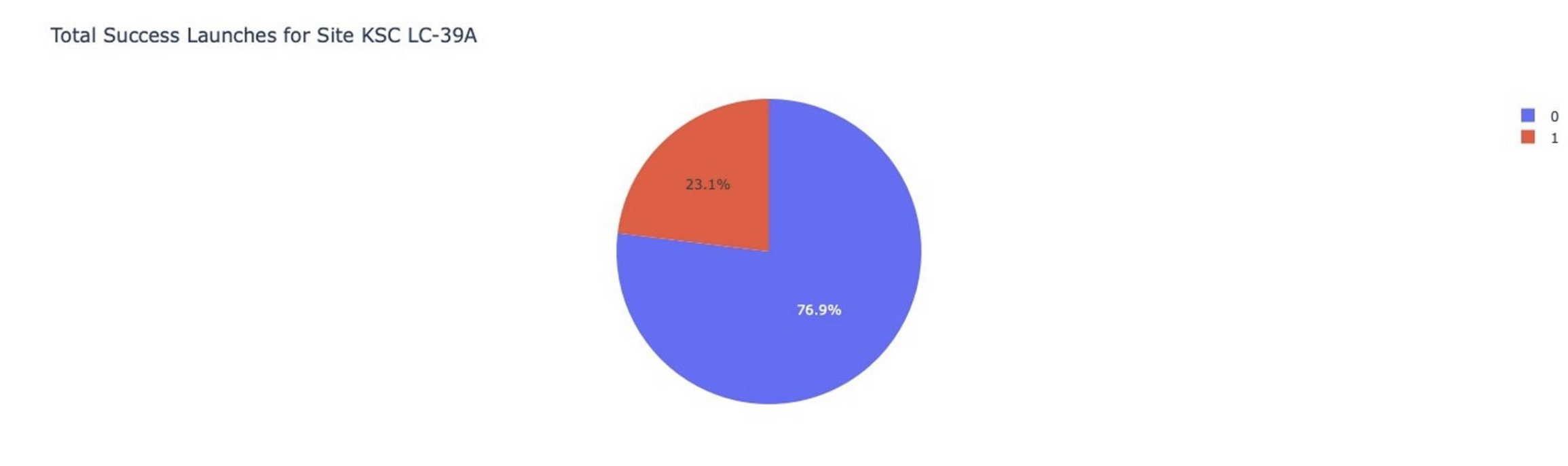
SECTION 5: BUILDING A DASHBOARD WITH PLOTLY

LAUNCH SUCCESS BY SITE



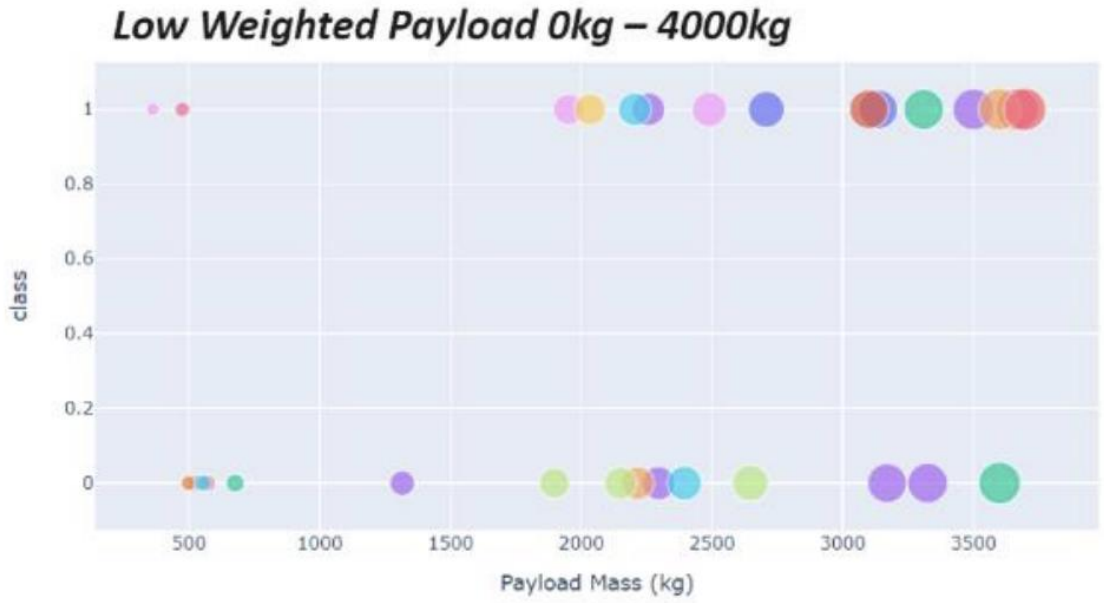
The diagram shows that from all sites, KSC LC-39A has the most successful launches

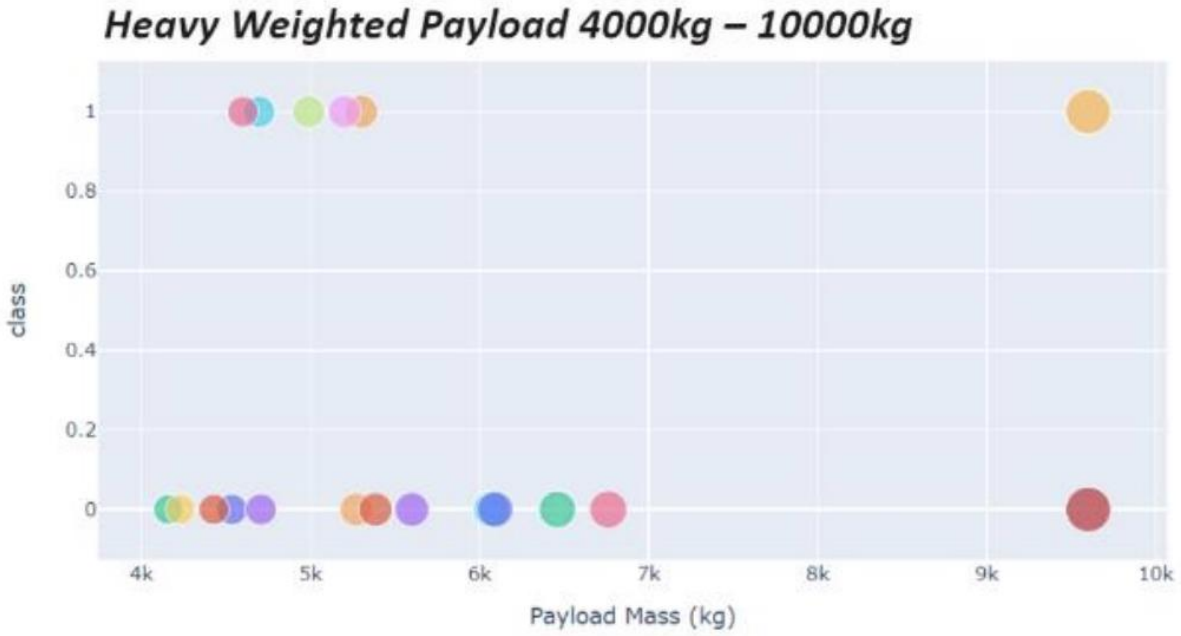
LAUNCH SITE WITH HIGHEST SUCCESS RATIO



The diagram shows that KSC LC-39A has the highest successful landing ratio, standing at 10:3

PAYLOAD MASS VS LAUNCH OUTCOMES FOR ALL SITES





The diagram shows that payloads between 2000 and 5500kg have the highest success rate

SECTION 6: PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

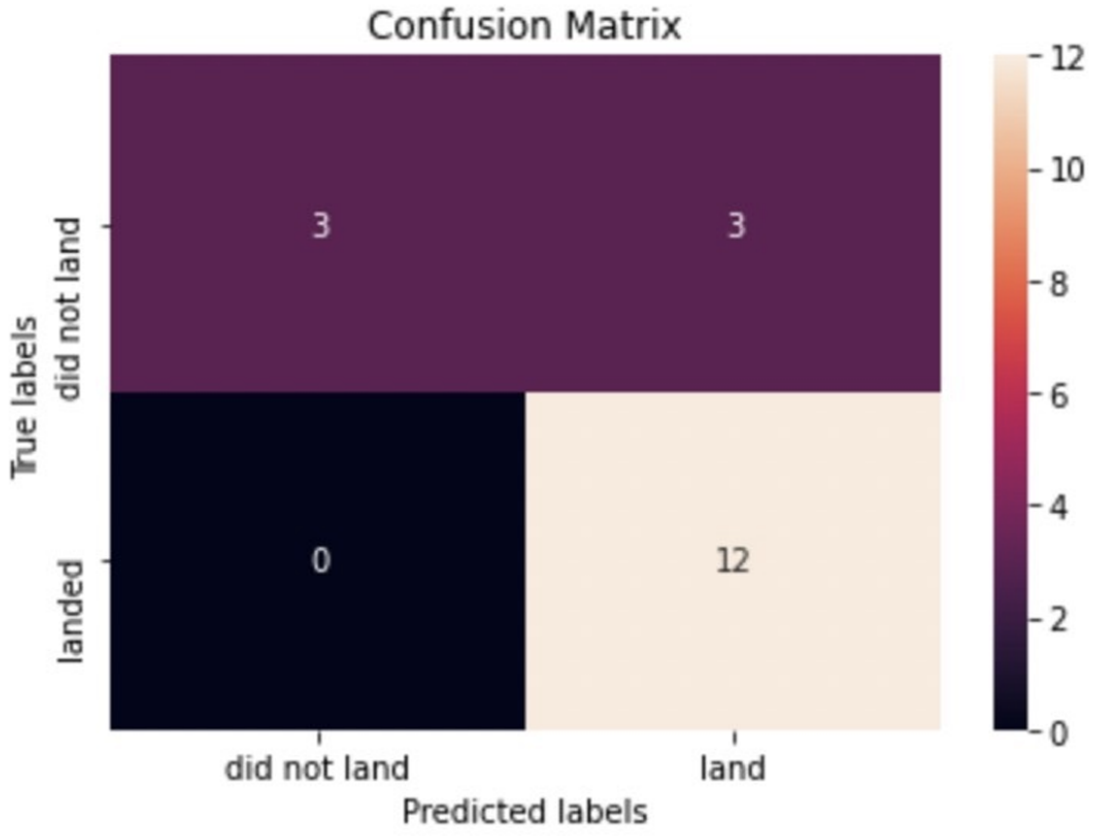
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

The decision tree classifier is the model with the highest classification accuracy

CONFUSION MATRIX

The confusion matrix for the decision tree classifier indicates that the classifier is capable of distinguishing between different classes. However, the primary issue lies in false positives, where unsuccessful landings are incorrectly identified as successful landings by the classifier.



CONCLUSIONS

- The Decision Tree Model is identified as the best algorithm for this dataset.
- Launches with lower payload masses exhibit better results compared to launches with larger payload masses.
- Most launch sites are situated in proximity to the Equator line, and all sites are very close to the coast.
- The success rate of launches demonstrates an increasing trend over the years.
- Launch site KSCLC-39A boasts the highest success rate among all sites.
- The larger the number of flights at a launch site, the higher the success rate.
- Launch success rate began to rise in 2013 and continued until 2020.
- Orbits ES-L1, GEO, HEO, SSO, and VLEO exhibit the highest success rates.
- KSC LC-39A registers the highest number of successful launches among all sites.

THANK YOU

Lawrence Mak

647-884-7651

maklawrence1110@gmail.com