Predicting Space X Falcon 9's First Stage Landing Success

By: Benedict Z. Castro

02/10/2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis (EDA)
  - Interactive Visual Analytics
  - Machine Learning Models
- Results
- Conclusion

EXECUTIVE SUMMARY

This research was conducted to study and identify the different data features for a rocket's successful first stage landing, and to predict the success rate. In order to achieve, these methodologies where employed:

- **Data collection** was done using the SpaceX REST API and through web scraping
- **Data wrangling** was employed to transform raw data to desired form
- **Exploratory data analysis** was performed to visualize key relationships between features which include *payload mass*, *launch site*, *flight number*, and *yearly trend*
- **Analysis of data using SQL** to calculate metrics like *total payload mass*, *payload ranges* for successful launches, and *total number of successful and failed outcomes*
- **Visualize** launch site success conditions and environments by utilizing geographical markers
- **Build** machine learning models to predict the first stage landing success using *logistic regression*, *support vector machine (SVM)*, *decision tree* and *K-nearest neighbor*

EXECUTIVE SUMMARY

It was observed that:

- *KSC LC-39A* has the highest success rate compared to all landing sites
- The orbits *ES-L1, GEO, HEO, and SSO* exhibited 100% success rates
- From the yearly trend, launch success rates have improved
- Launch sites were chosen to be close to coastlines

The **decision tree** model showed the best results, but only slightly outperforming the other predictive models within 5% difference.

INTRODUCTION

Background

Space Exploration Technologies Corp., or better known as SpaceX, is a global aerospace industry leader with the purpose of revolutionizing space technology so that it would be accessible for everyone. One of its greatest contributions to humanity include sending astronauts and spacecraft to the International Space Station (ISS).

SpaceX's Falcon 9 rocket launches are being advertised with inexpensive costs (around 62M dollars) as compared to its other competitors with launch costs roughly around 162M dollars. This was made possible because of their genuine idea of reusing the rocket's first stage. Thus, it is essential to determine the success rate of the first stage landing in order to estimate the cost of a launch.

Study:

- factors that affect the rocket launch's first stage landing success
- relationship between these features with the success rate
- predictive models that can be utilized to determine the landing success rate

Methodology

METHODOLOGY

Executive Summary

- Data collection methodology:
  - data was collected using the SpaceX REST API and through web scraping
- Perform data wrangling
  - transforming raw data to usable form through filtering, 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
  - How to build, tune, and evaluate classification models

METHODOLOGY

Data Collection - API

- Requested rocket launch data from SpaceX API
- Decoded response content using json functions from pandas
- Requested information about the launches using the API and auxiliary functions
- Created dataframe from requested data
- Cleaned the dataframe
  - filtered Falcon 9 launches
  - replaced missing values with calculated mean statistic
  - exported to a csv file

- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-spacex-data-collection-api.ipynb

METHODOLOGY

Data Collection – Web Scraping

- Requested rocket launch data for Falcon 9 from Wikipedia
- Instantiated BeautifulSoup object from HTML response
- Collected data by parsing HTML tables
- Created dataframe from parsed data
- Cleaned the dataframe
- Exported to a csv file

- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-webscraping.ipynb

METHODOLOGY

Data Wrangling

- Performed EDA and determined training data labels
- Calculated:
  - number of launches for each site
  - number and occurrence of orbit, and mission outcome for each orbit type
- Created binary labels for landing outcomes
- Exported data results to csv
- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-data-wrangling.ipynb

METHODOLOGY

EDA with Data Visualization

- Plotted charts to visualize patterns and identify relationships
  - Flight Number vs. Payload
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Payload Mass vs. Orbit Type
  - Landing Success Rate vs. Orbit Type
- The relationship between the said features were analyzed using scatter plots which would later be essential in predicting the landing success
- Landing success rates for each orbit were showcased with the use of bar charts for better visualization.
- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-eda-dataviz.ipynb

METHODOLOGY

EDA with SQL

Determined:

- names of unique launch sites
- launch sites that start with “CCA”
- first successful landing on ground pad
- boosters which landed successfully on a drone ship and have 4000 kg < payload mass < 6000 kg.
- booster versions which have carried the max payload mass
- failed landing outcomes on a drone ship (including booster version and launch site) in 2015

Calculated:

- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-eda-sql-coursera_sqllite.ipynb

METHODOLOGY

Interactive Map with Folium

- Launch sites were marked with circle map objects – blue circle at NASA Johnson Space Center and red circles at all launch sites – with popup labels indicating their coordinates or location
- Launch outcomes were marked – green if successful and red if unsuccessful – using clusters to identify launch sites with high success rates
- Marker lines were used to show the distances between launch site *CCAFS SLC-40* and establishments like coastlines, railways, highways, and cities to assess location factors
- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-launch-site-location.ipynb

METHODOLOGY

Dashboard with Plotly Dash

- Added dropdown list of launch sites to allow users to access data for specific sites or for all sites
- Created a pie chart to depict successful and unsuccessful launches as part of a whole
- Added a slider for payload mass range to easily navigate between different ranges
- Created a scatter plot showing the relationship between the outcome result and payload mass for the different booster versions
- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-interactive-dashboard-with-plotly-dash.ipynb

METHODOLOGY

Predictive Analysis (Classification)

- Created a NumPy array from the class data column
- Standardized and transformed the data using StandardScaler and transform
- Split the data into training and testing using train_test_split
- Created a GridSearchCV object to perform parameter optimization
- Performed GridSearchCV for different algorithms
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- Calculated accuracy scores for all models with test data
- Created confusion matrices for all models
- Assessed the best performing classification model through feature engineering and algorithm tuning and using accuracies scores as metric
- The link to the notebook flowchart is available here:
  - https://github.com/benicastro/data-portfolio/blob/main/ibm-data-science/jupyter-labs-predictive-analysis.ipynb

RESULTS

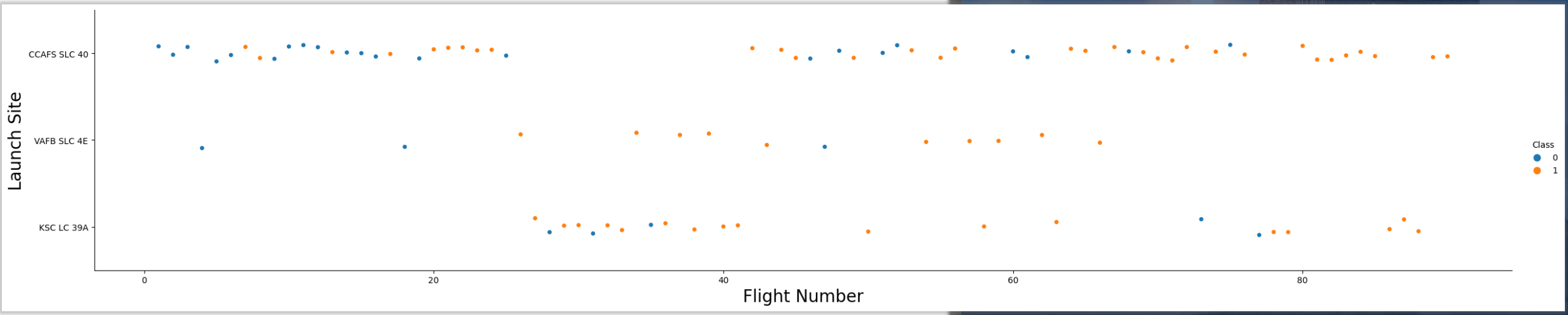
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Insights drawn from EDA

RESULTS

Flight Number vs. Launch Site

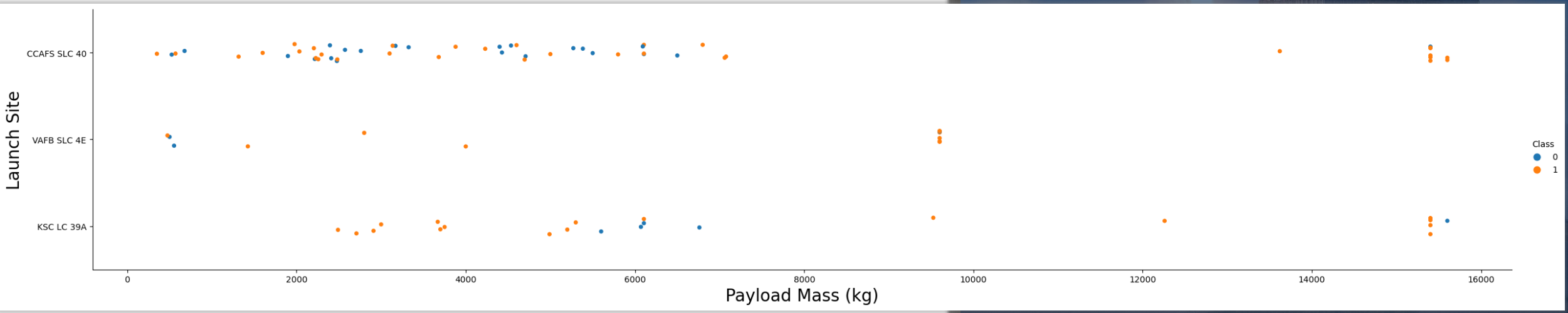
- It can be seen that with increasing flight number for a specific launch site, the greater the success rate.
  - earlier flights had a lower success rate
  - later flights had a higher success rate
- It can be inferred that new launches can lead to higher success rates



RESULTS

Payload vs. Launch Site

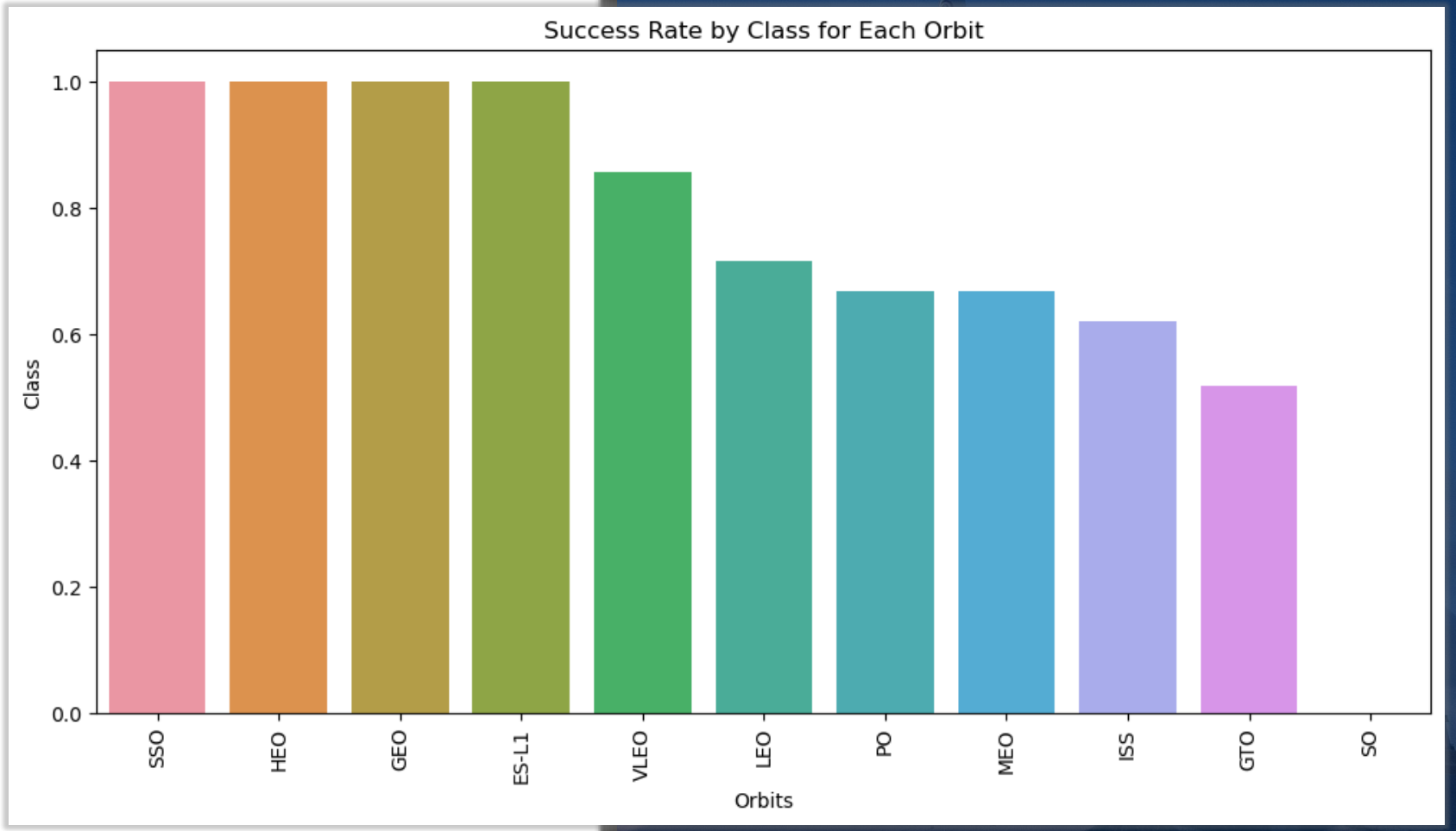
- The greater the payload mass, the higher the success rate
- Launches with payload masses less than 5,000 kg for KSC LC 39A have 100% success rate
- It can be observed that majority of launches with payload masses greater than 7,000 kg have relatively high chances of success



RESULTS

Success Rate vs. Orbit Type

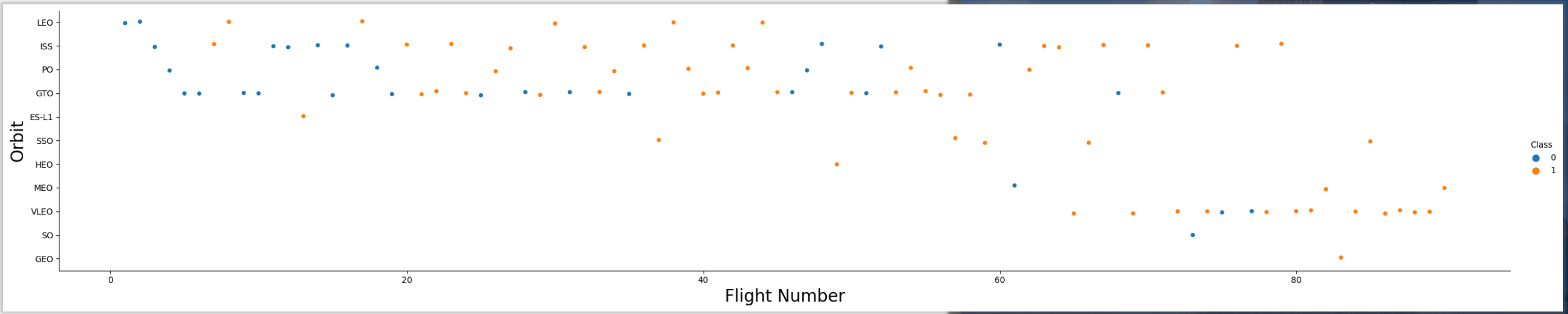
- Four orbit types have 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Only the SO orbit type have 0% success rate



RESULTS

Flight Number vs. Orbit Type

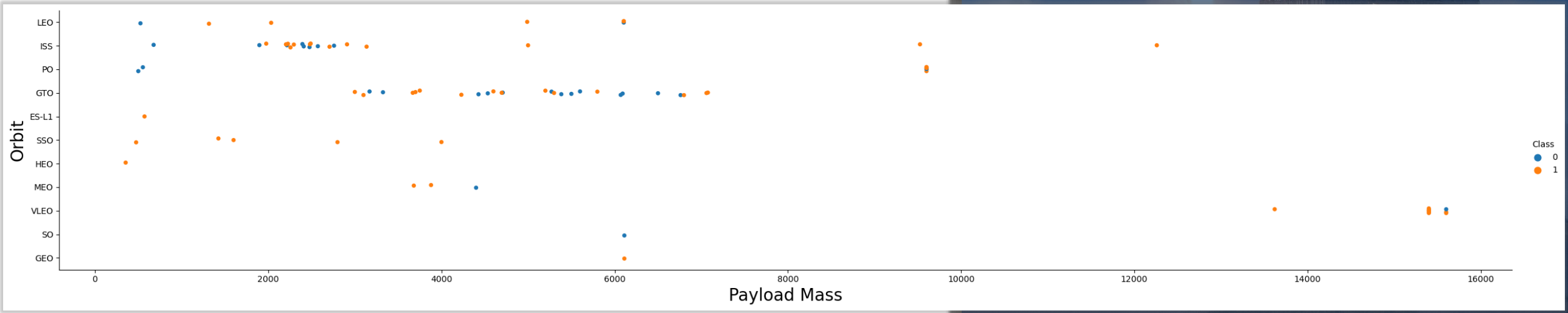
- Generally, with increasing flight number, the success rate also increases as evident in the LEO orbit.
- The GTO orbit is an exception to this as there is no apparent relationship between the two variables.



RESULTS

Payload vs. Orbit Type

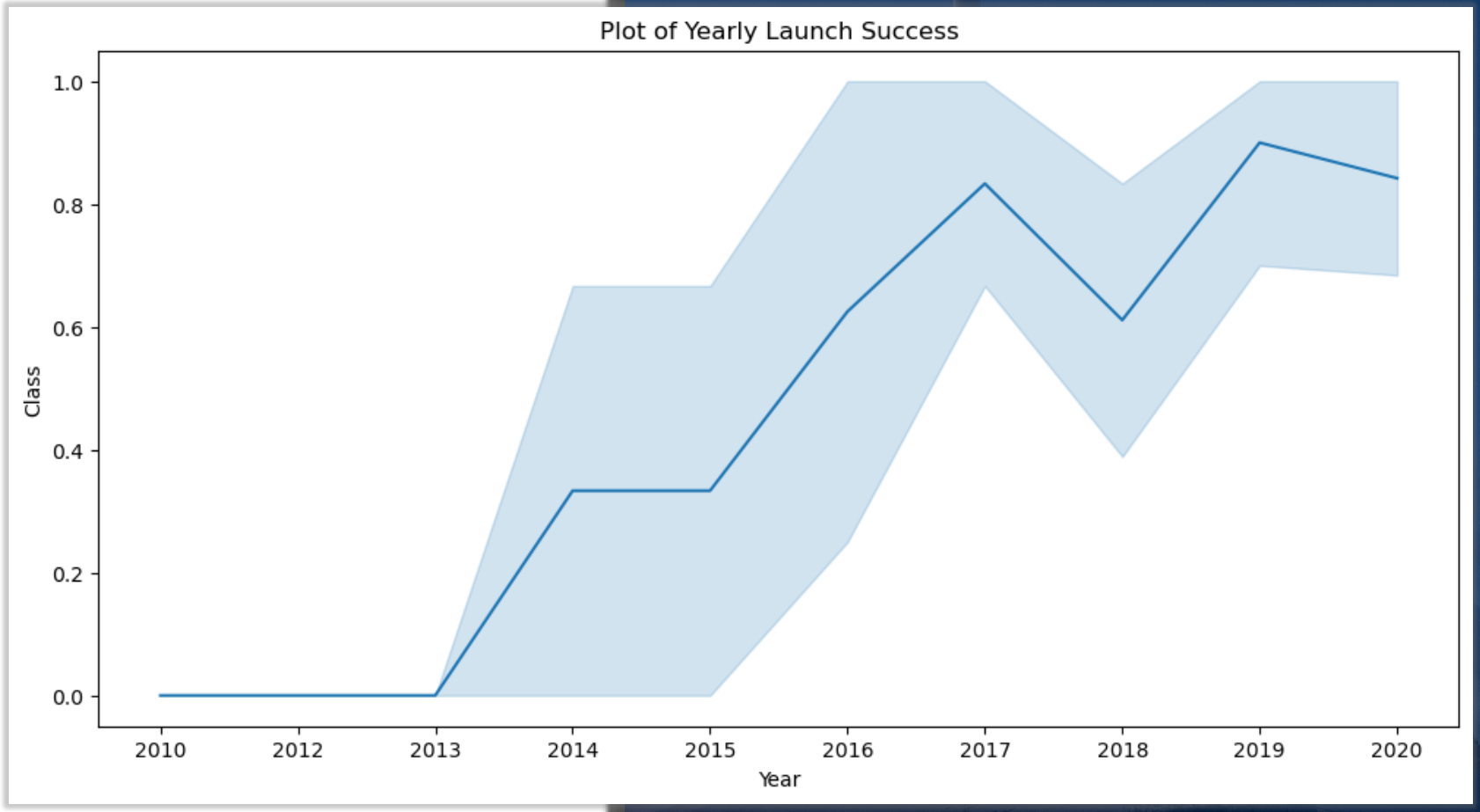
- It can be seen that for majority of the launches in the orbits ISS, LEO, and PO, the heavier the payloads, the higher the chances of success



RESULTS

Launch Success Yearly Trend

- It can be observed that starting from 2013, the success rate steadily increases until 2020, with a dip during 2017-2018



RESULTS

All Launch Site Names

- The DISTINCT keyword was used to show only the unique launch sites from the data

Display the names of the unique launch sites in the space mission

In [12]:
```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTABLE ORDER BY 1
```

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

Out[12]:

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

RESULTS

Launch Site Names Beginning with “CCA”

- The LIKE keyword was used together with the condition keyword WHERE to query the data

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

```
In [13]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[13]:

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

RESULTS

Total Payload Mass

- The total payload mass is calculated to be 45,596 kg.

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

In [14]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

Out[14]:

SUM(PAYLOAD_MASS__KG_)
45596

RESULTS

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated to be ~2534.67.

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

```
In [15]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
```

Out[15]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

RESULTS

First Successful Ground Landing Date

- The first successful landing on a ground pad happened on the 22nd of December 2015.

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
In [17]: %sql SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

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

Out[17]:

first_successful_landing
2015-12-22

RESULTS

Successful Drone Ship Landing with Payload between 4000 and 6000

- The conditional keyword WHERE was used, together with the logical operator AND to filter the booster versions needed.

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

In [18]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYL
```

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

Out[18]:

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

RESULTS

Total Number of Successful and Failure Mission Outcomes

- The conditional keyword WHERE was again utilized with the keyword LIKE and wildcard % to retrieve the number of mission outcomes.

List the total number of successful and failure mission outcomes

In [21]:
```
# Successful Mission Outcomes
%sql SELECT COUNT(Mission_Outcome) As success_outcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%';
```

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

Out[21]:

success_outcome
100

In [22]:
```
# Failure Mission Outcomes
%sql SELECT COUNT(Mission_Outcome) As failure_outcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%';
```

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

Out[22]:

failure_outcome
1

RESULTS

Boosters Carrying Maximum Payload

- This query was done with the help of the MAX() function inside a subquery.
- The maximum payload mass was 15,600 kg.

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

```
In [23]: %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPA
```

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

Out[23]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

RESULTS

2015 Launch Records

- For this query, the keywords WHERE, LIKE, AND, and BETWEEN were used to filter the needed data.

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [24]:
```
%sql SELECT Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND
```

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

Out[24]:

Booster_Version	Launch_Site	Landing_Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

RESULTS

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

- The GROUP BY and ORDER BY keywords were utilized in this query to retrieve the ranking of landing 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.

In [25]:
```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY La
```

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

Out[25]:

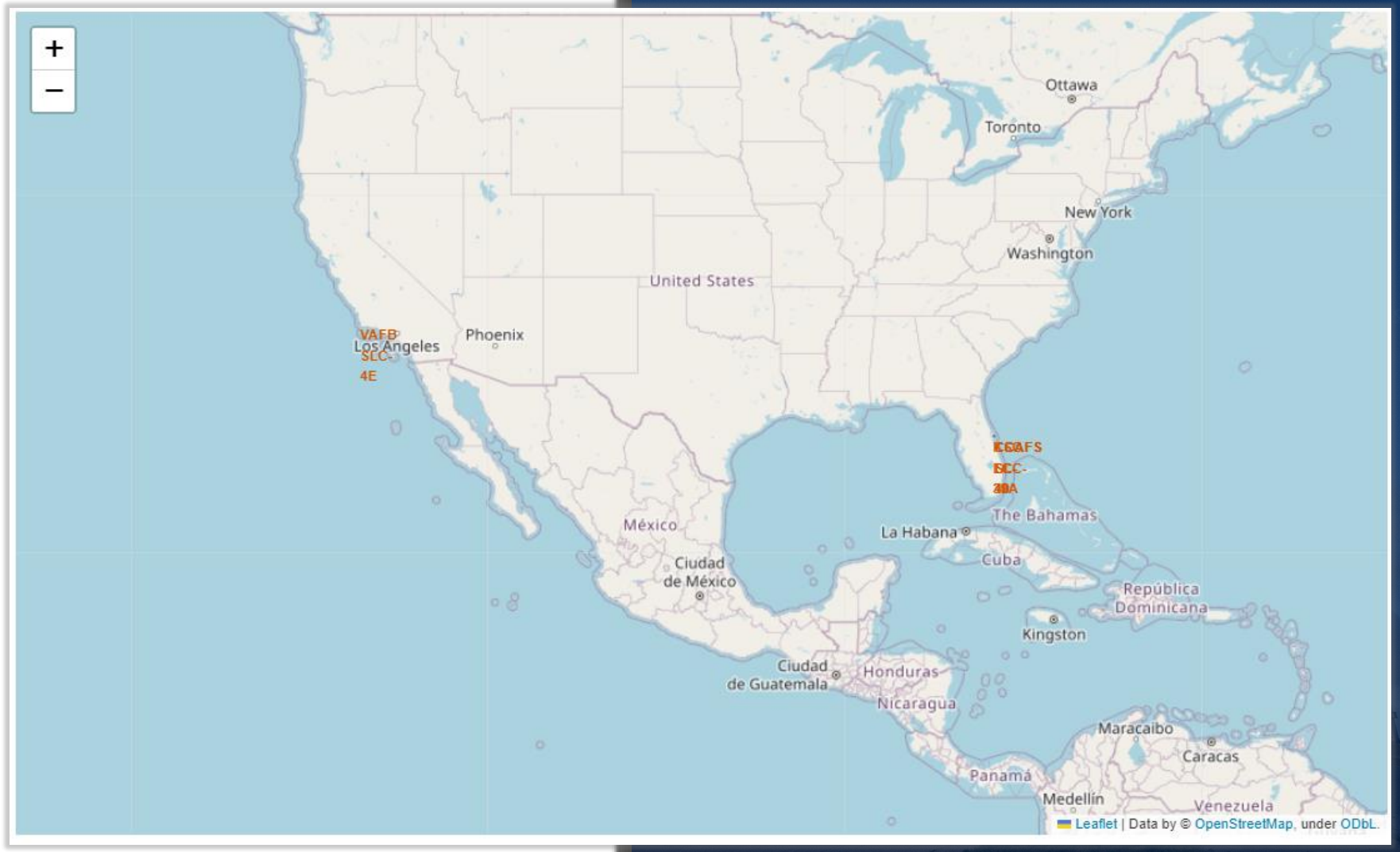
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Launch Sites
Proximities Analysis

RESULTS

All Launch Sites Global Map Markers

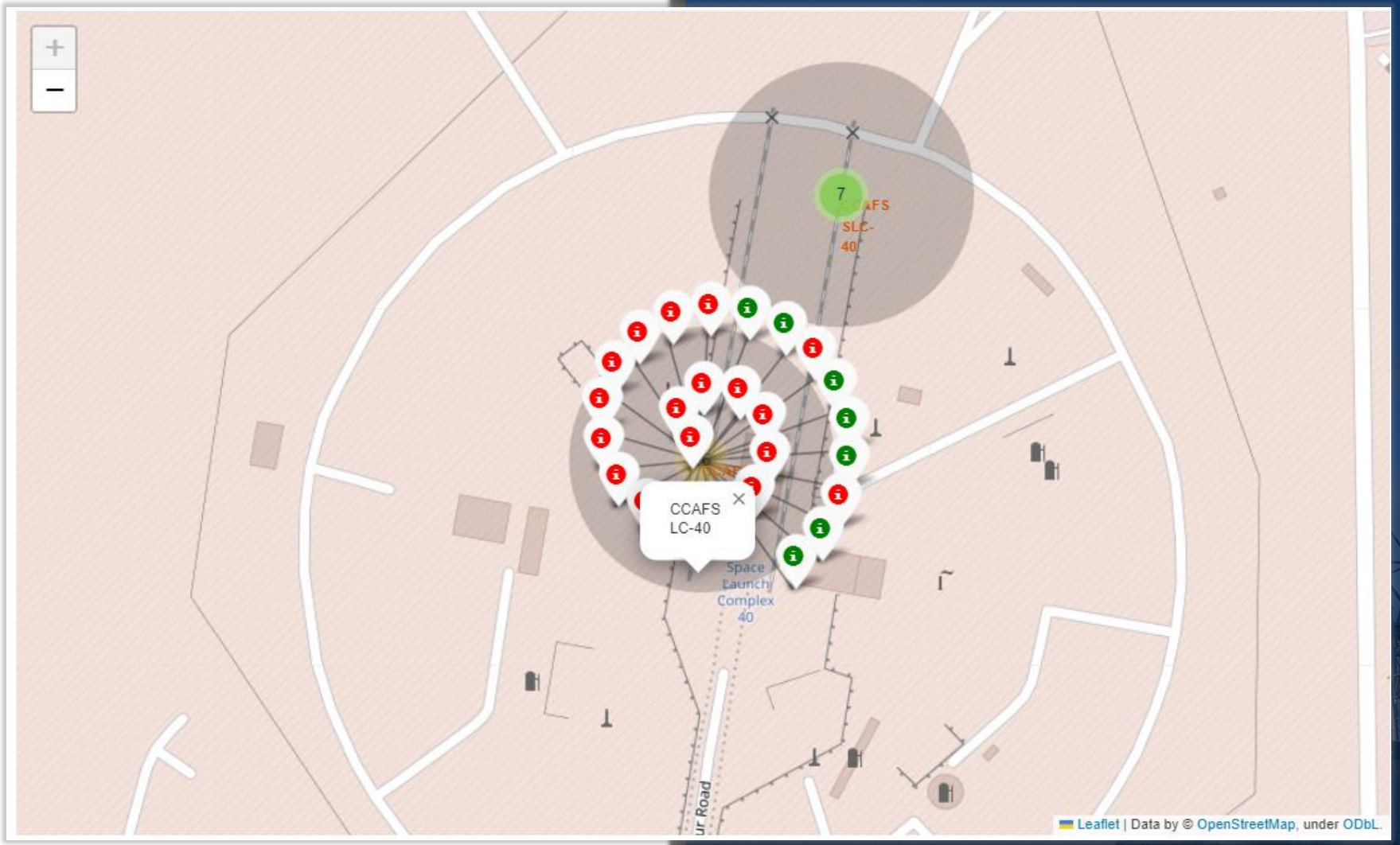
- The launch sites for SpaceX are located near coastlines as seen in the figure.
- Markers were used in order to identify every launch site for better understanding.



RESULTS

Launch Site Markers

- CCAFS SLC-40 Launch Site
- Green Markers show successful launches while Red Markers show failed launches

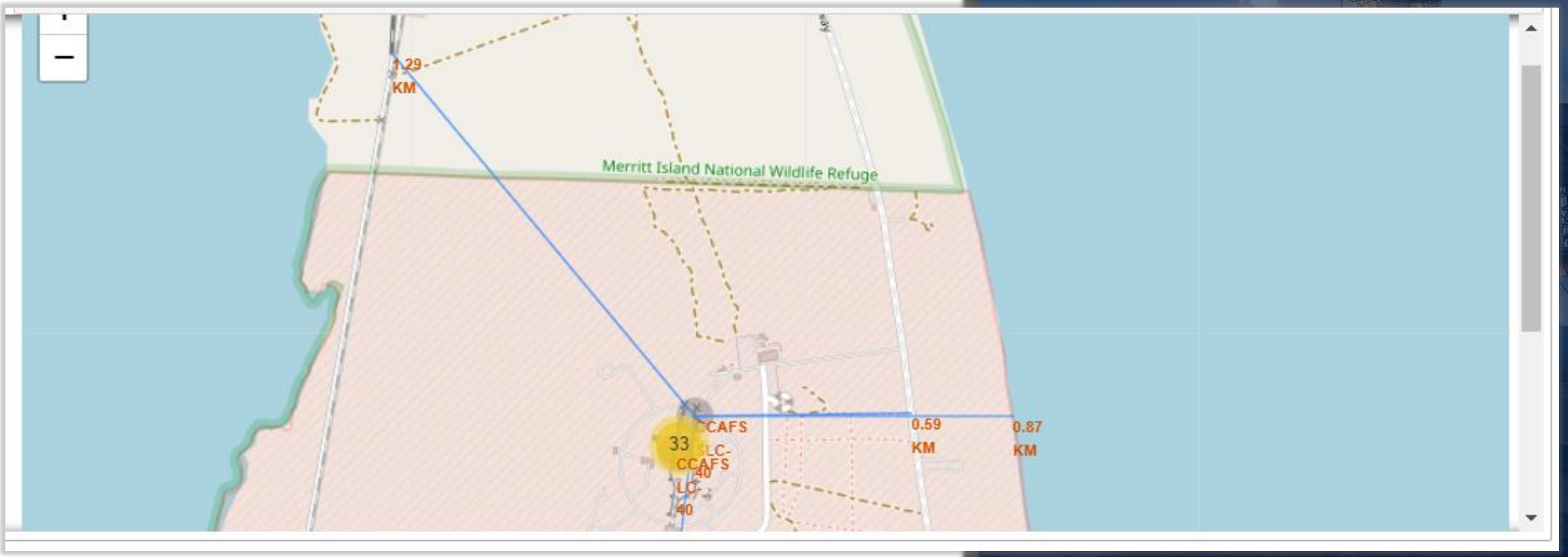


RESULTS

Launch Site Distance to Proximities

- Launch sites are:
  - not in close proximity to railways
  - not in close proximity to highways
  - in close proximity to coastlines
  - kept in certain distance away from cities

Coastline Distance: 0.87 km
City Distance: 51.74 km
Railway Distance: 1.29 km
Highway Distance: 0.59 km

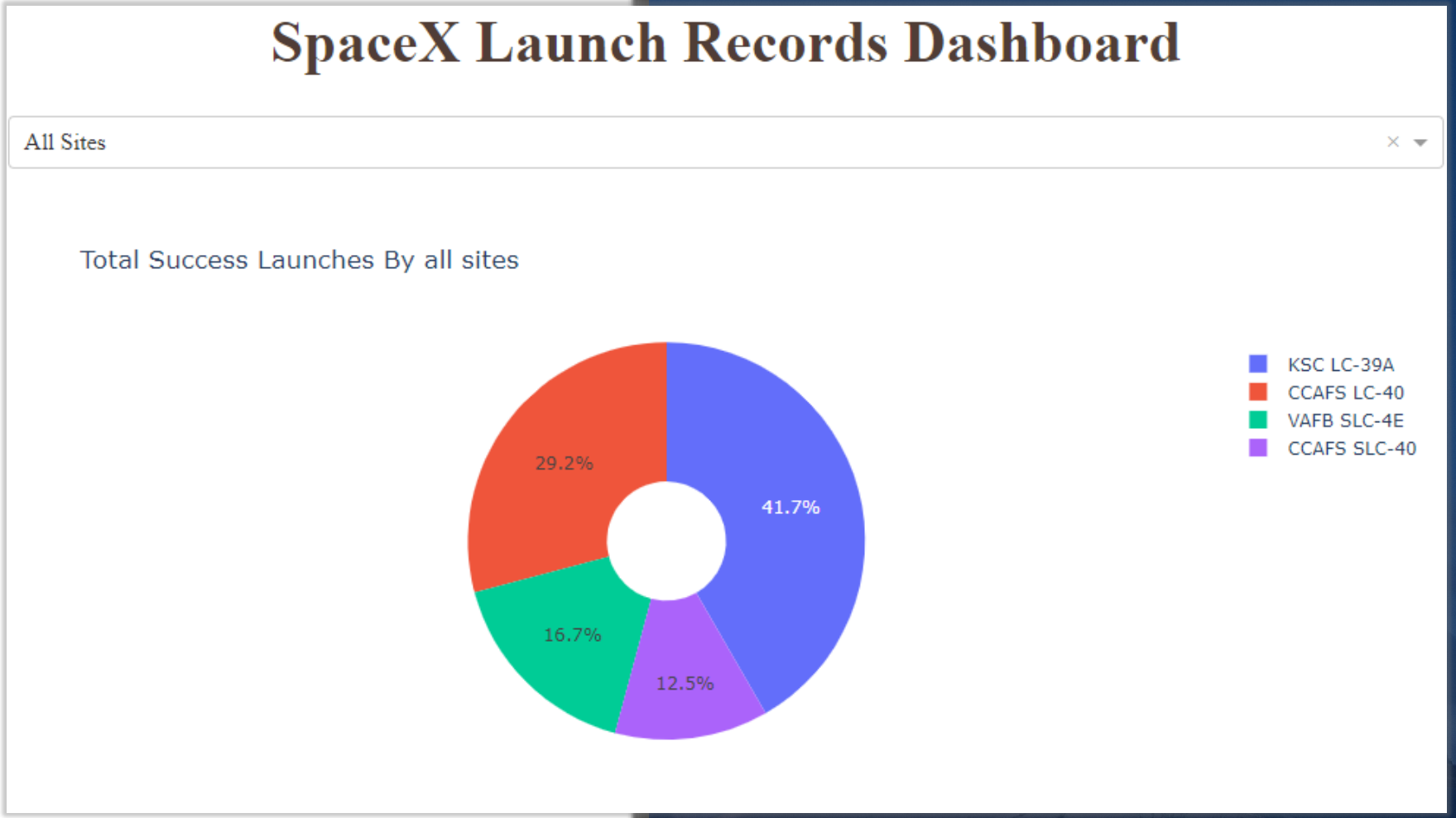



Build a Dashboard
with Plotly Dash

RESULTS

Launch Success by Site

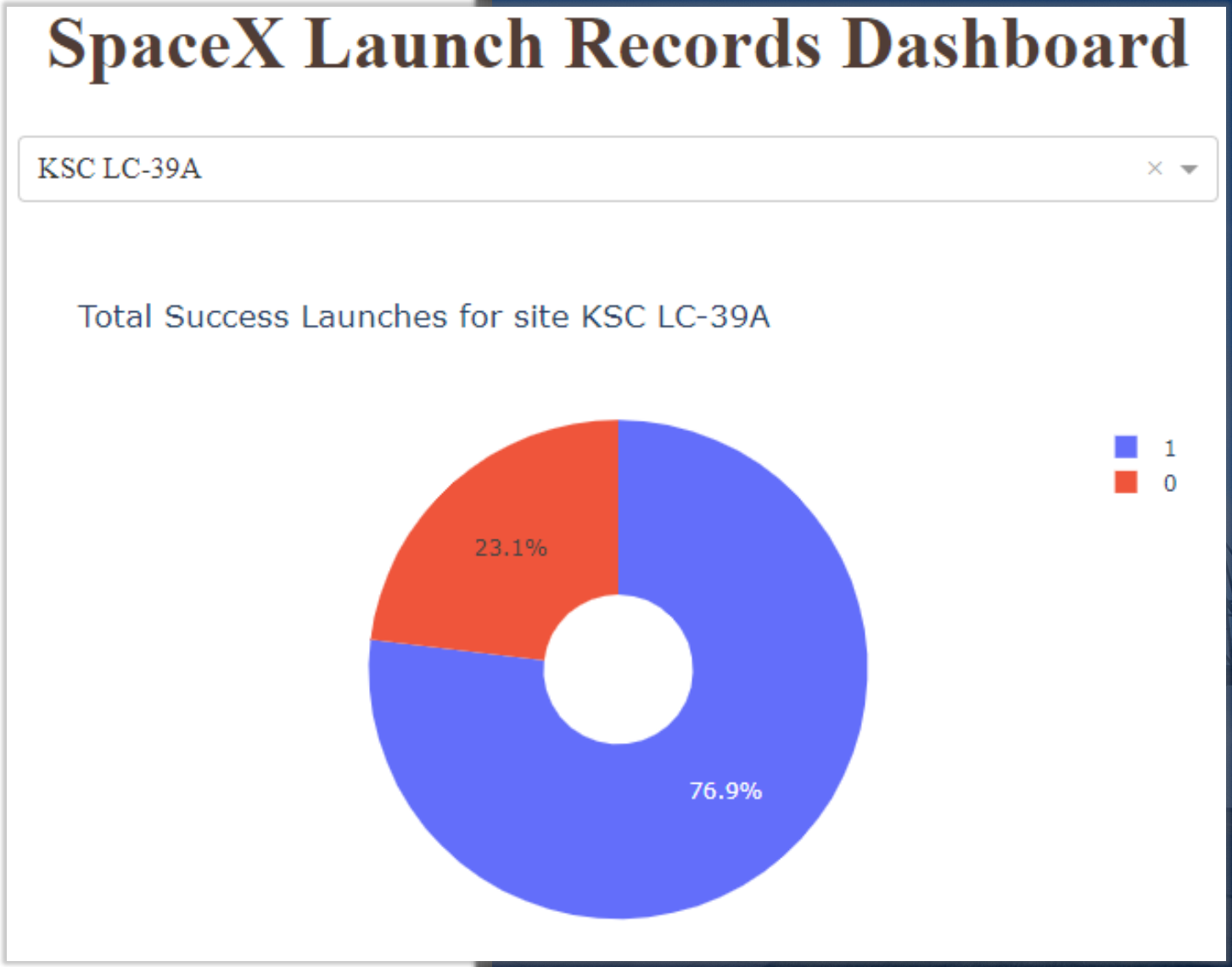
- It can be seen that KSC LC-39A has the most number of successful launches compared to other launch sites, with around 41.7% of the total.



RESULTS

Launch Success for KSC LC-39A

- Upon inspection, it was found out that KSC LC-39A had the highest success rate with around 76.9%.



RESULTS

Payload Mass vs. Launch Outcome for All Sites

by Booster Version

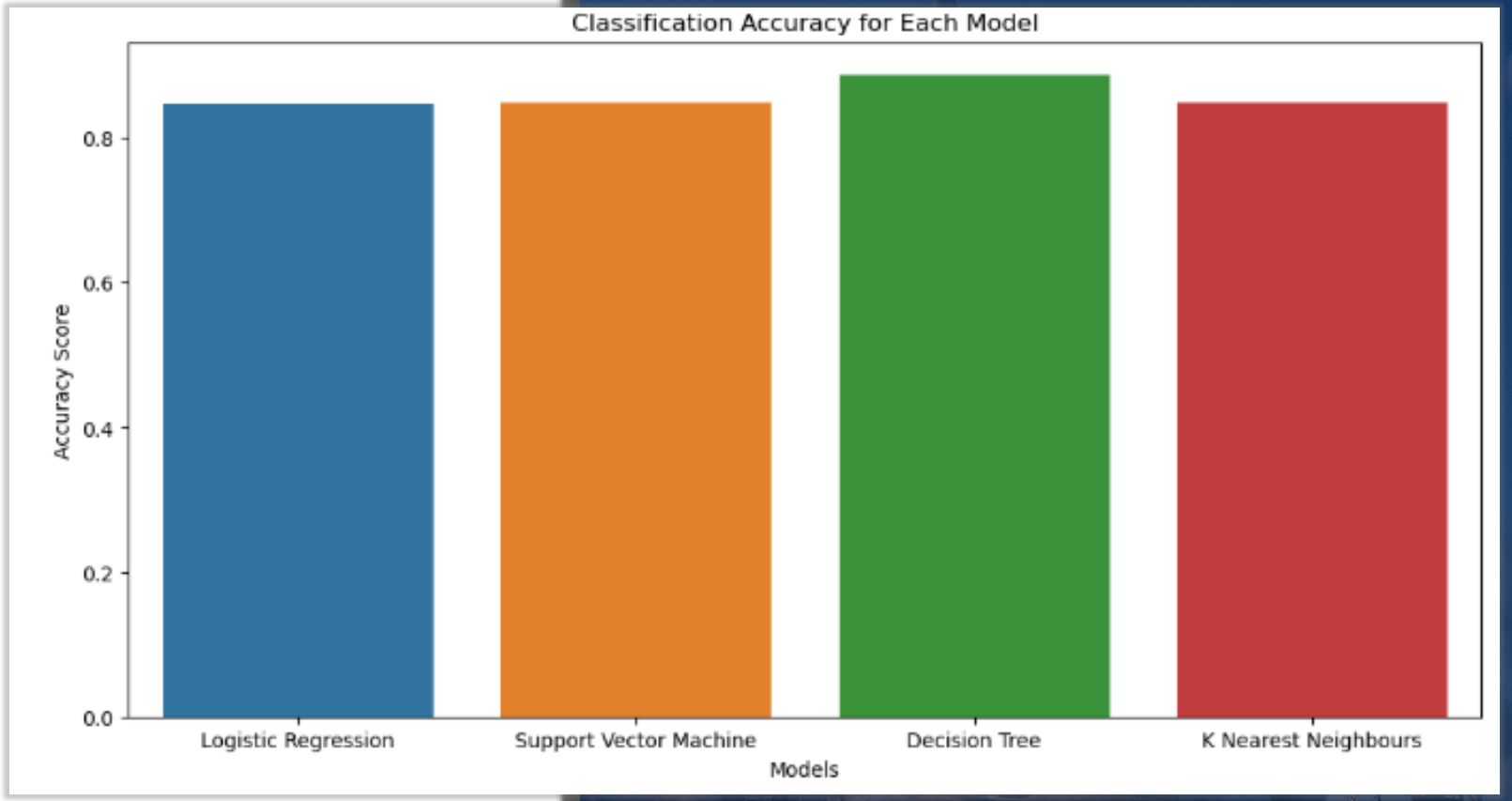
- The success rates for launches with small payload masses are relatively higher than the launches with heavy payload masses.
- Specifically, payload masses between 2,000 kg and 5,000 kg have the highest values for success rate


Predictive Analysis
(Classification)

RESULTS

Classification Accuracy

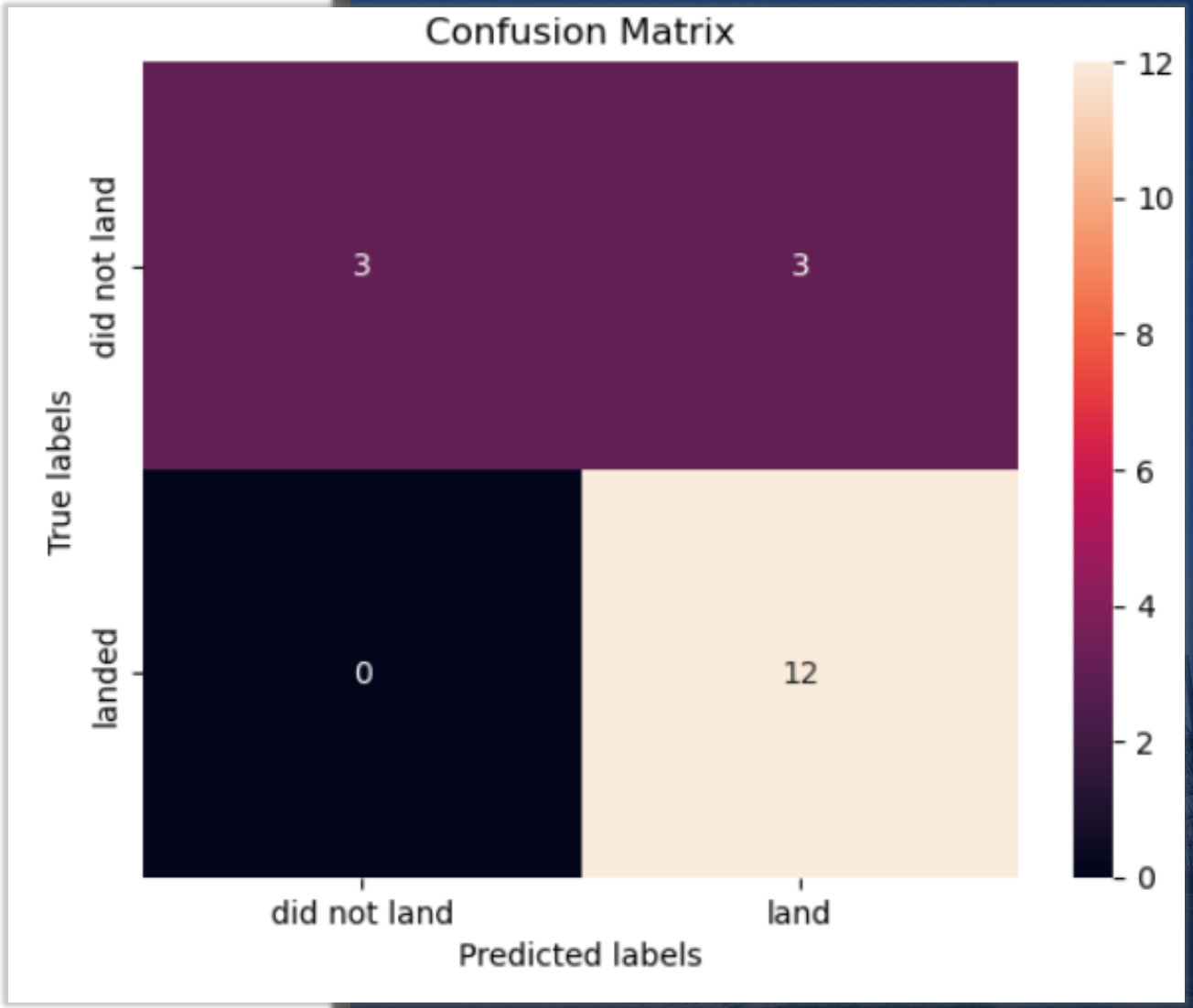
- It can be observed that all the predictive models performed relatively at the same level as based on their accuracy scores (.best_score_) with differences of only about 4%.
- It can be recalled that the dataset used was small, and thus, may account for these findings.
- The Decision Tree Model had the highest classification accuracy, slightly outperforming the other models.



RESULTS

Confusion Matrix

- It was noted that all the confusion matrices from the different models were the same
- One major problem with the results is the presence of false positives (Type I error) where unsuccessful landing outcomes were marked as successful by the classifier.



CONCLUSIONS

- As the flight number increases, the success rate also improves across all sites.
- Launch success rates increase overtime, starting from 2013 till 2020.
- Launch sites are strategically placed close to coastlines.
- KSC LC-39A had the most number of successful launches.
- The orbits ES-L1, GEO, HEO, and SSO boast a 100% success rate.
- Launches with higher payload mass have relatively higher success rates.
- All the predictive models yielded almost the same results with the Decision Tree classifier slightly outperforming the rest with 4% difference


Thank You!