IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tanaka Courage Mawonedzo
14 October 2024

Outline

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

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

Project background

This capstone project aims to predict the successful landing of SpaceX's Falcon 9 first stage, a crucial factor in determining the cost efficiency of rocket launches. SpaceX offers launches at $62 million, significantly lower than other providers, primarily due to its ability to reuse the first stage of its rockets. By accurately predicting the likelihood of a successful landing, potential competitors could use this analysis to bid more competitively against SpaceX for launch contracts.

Business Objective

Develop a predictive model to determine if the Falcon 9 first stage will land successfully, enabling cost estimation for launches and informing strategic decisions for companies aiming to compete with SpaceX.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best
  - results

Data Collection

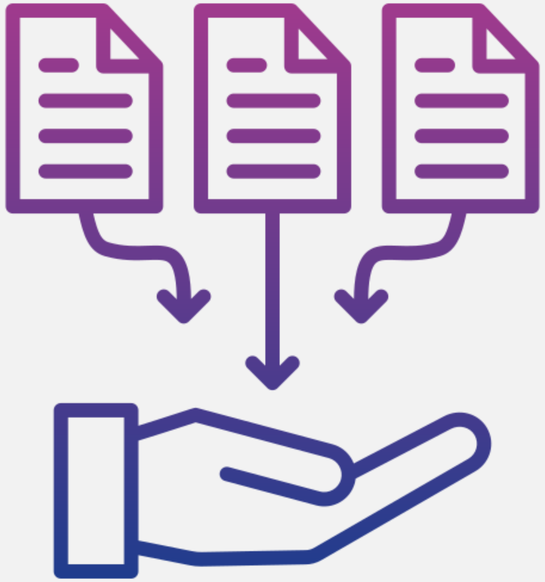
Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. Both data collection methods was used to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time



Data Collection – SpaceX API

Steps

- Request data from SpaceX API (rocket launch data)
- Decode response content using .json() and turning it into a dataframe using .json_normalize()
- Request information about the launches from SpaceX API by applying custom functions
- Construct data that was obtained into a dictionary
- Create a dataframe from the dictionary
- Filter the dataframe to only include Falcon 9 launches
- Replace missing values of Payload Mass column with calculated .mean() for this column
- Export the data to CSV

GitHub URL: Data Collection Space X API

Data Collection - Scraping

Steps

- Requesting Falcon 9 launch data from Wikipedia
- Creating a BeautifulSoup object from the HTML response
- Extracting all column names from the HTML table header
- Collecting the data by parsing HTML tables
- Constructing data we have obtained into a dictionary
- Creating a dataframe from the dictionary
- Exporting the data to CSV

[GitHub URL: Data Collection with Web Scraping]()

Data Wrangling

- The cleaned data was first imported and checked what percentage of missing values was in the LaunchingPad since it was the only column still containing missing value indicating when “no LaunchingPad” was used. The datatype of each column was then checked and there were 4 different datatypes(int64, object, float64 and bool). Further analysis like value count of LaunchSite was also examined for the various facilities Cape Canaveral Space Launch Complex 40 VAFB SLC 4E seem to have the highest count of 55.
- A new feature called “class” from the outcome's column was created were all outcome containing the name “False” and “None” were regarded as bad, therefore a value of zero(0) was assigned for the bad outcome and one(1) for good outcome.
- The success rate of all the good outcome was calculated which made up to 66.67% of the class feature.

GiHub URL: Data wrangling

Data Wrangling

Steps

- Perform exploratory Data Analysis and determine Training Labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

GiHub URL: Data wrangling

EDA with Data Visualization

Scatter Graphs

- Flight Number VS. Paload Mass
- Flight Number VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS.Payload Mass

Scatter plots show how much one variable is affected by another. The relationship beatween two variables is called their correlation . Scatter plots usually consist of a large body of data.

Bar Graphs

- Mean VS. Orbit

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

Line Graphs

- Success Rate VS. Year

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

GiHub URL: EDA with Data Visualization

EDA with SQL

Queries

Display:

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

List:

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

GiHub URL: EDA with SQL

Build an Interactive Map with Folium

Markers Indicating Launch Sites

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

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

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

[GitHub URL: Interactive Map with Folium]

Build a Dashboard with Plotly Dash

The dashboard is built with Plotly Dash web framework and contains the following components:

Dropdown List with Launch Sites : The user can select all launch sites or a certain launch site

Pie Chart Showing Successful Launches : The user can see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range : The user can select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version : The user can see the correlation between Payload and Launch Success

GitHub URL: Dashboard with Plotly Dash

Predictive Analysis (Classification)

BUILDING MODEL

- Load the dataset into NumPy and Pandas
- Transform Data
- Split the data into training and test data sets
- Check how many test samples available
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

MODEL

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

GiHub URL: Predictive Analysis

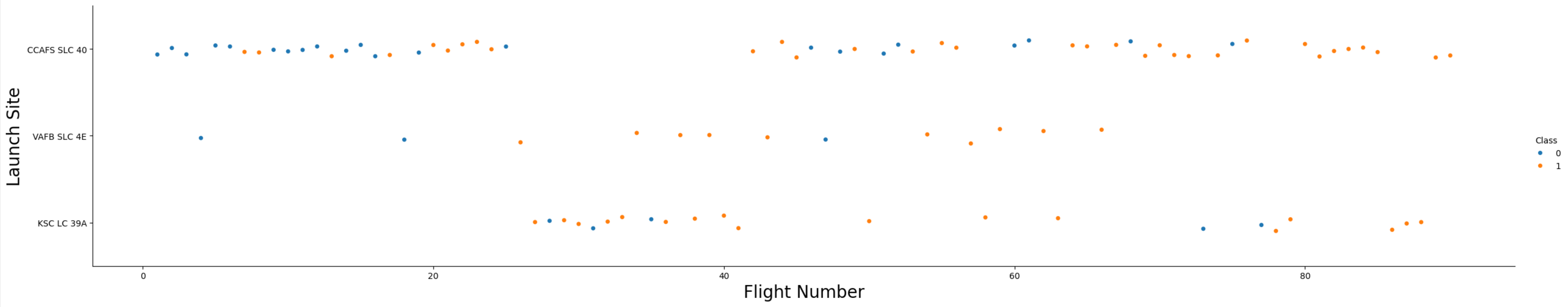
Results

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

Section 2

Insights drawn from EDA

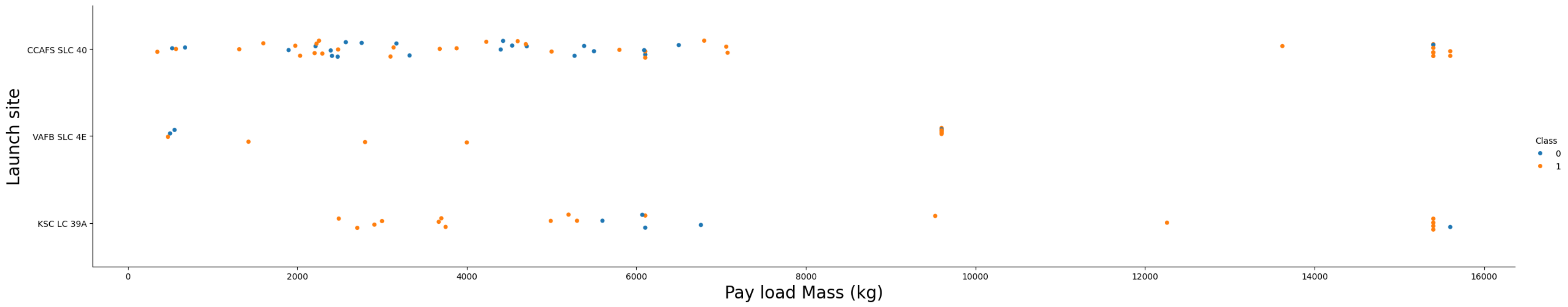
Flight Number vs. Launch Site



Exploratory data analysis results

- 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

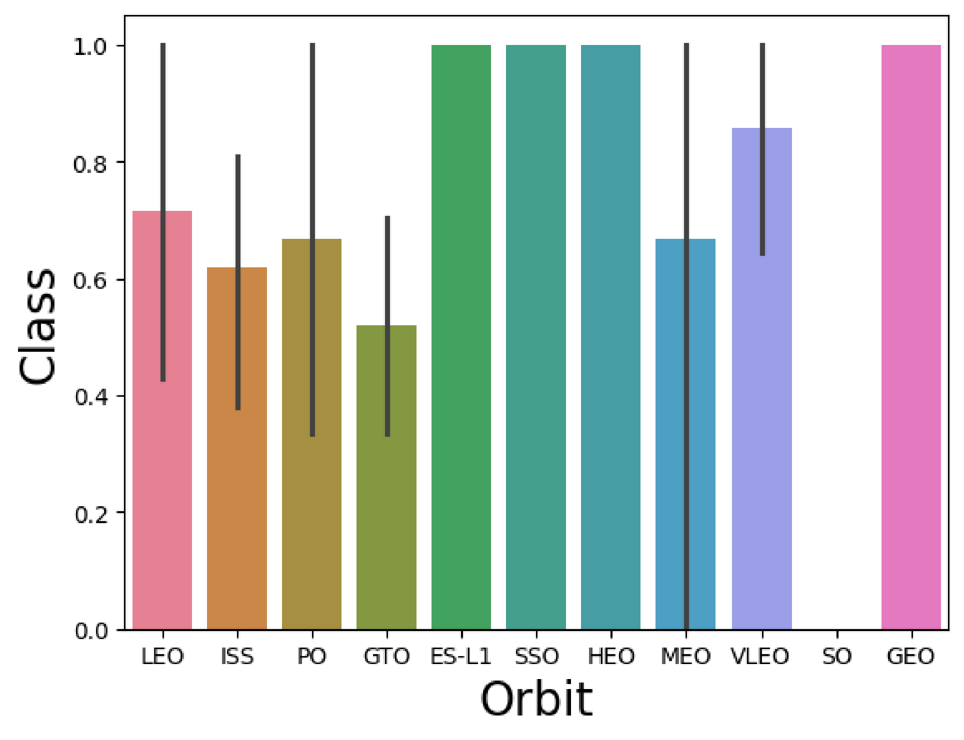
Payload vs. Launch Site



Exploratory data analysis results

- 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

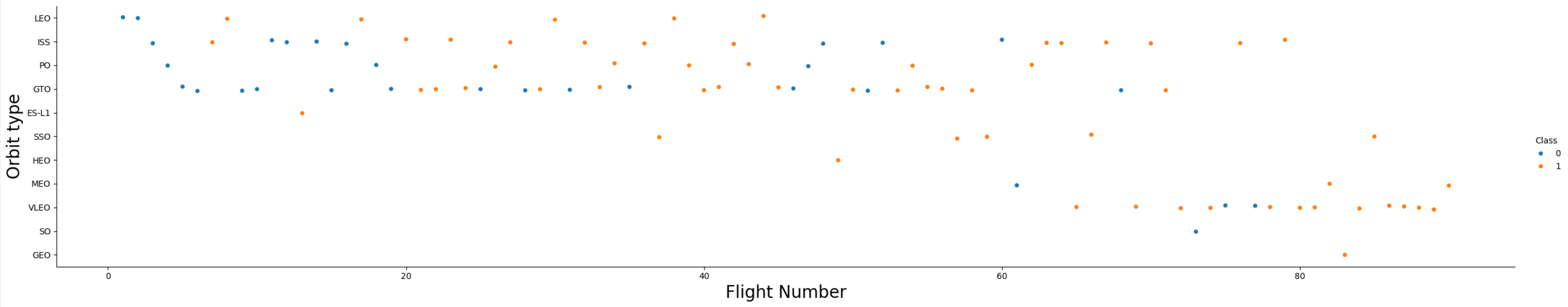
Success Rate vs. Orbit Type



Exploratory data analysis results

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO

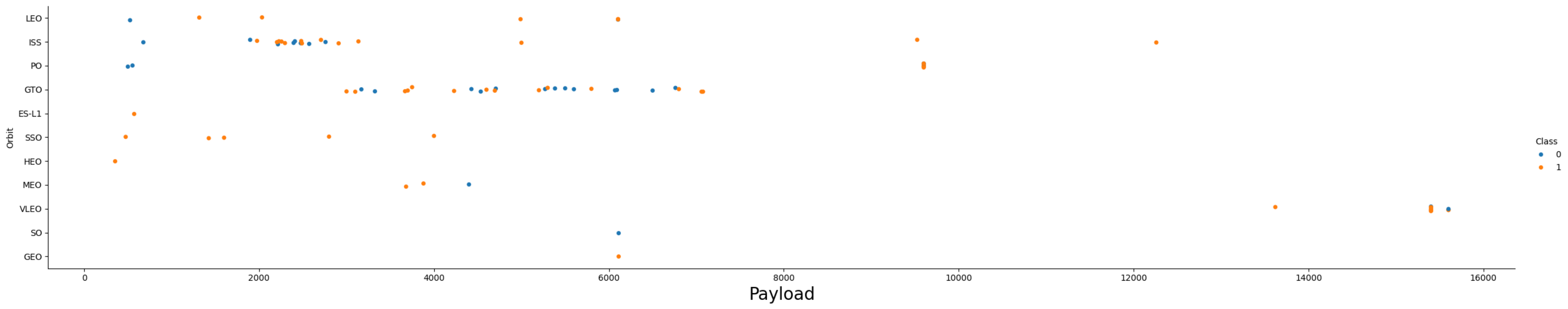
Flight Number vs. Orbit Type



Exploratory data analysis results

- It is observed that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

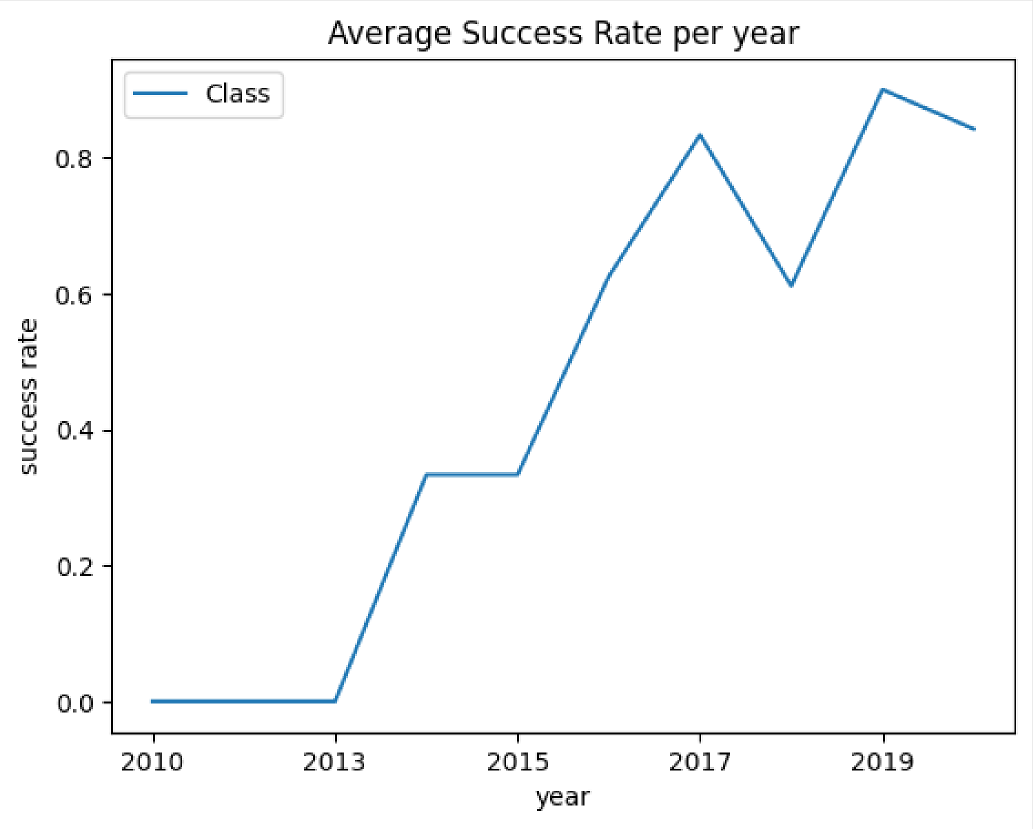
Payload vs. Orbit Type



Exploratory data analysis results

- It is observed that heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar, LEO, (ISS) orbits.

Launch Success Yearly Trend



Exploratory data analysis results

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013

All Launch Site Names

```
In [10]: %sql SELECT DISTINCT "launch_site" FROM SPACEXTABLE;

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

Out[10]:

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

- Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from SPACEXTABLE

Launch Site Names Begin with 'CCA'

In [11]:

```
%%sql
SELECT *
FROM SPACEXTABLE
where "LAUNCH_SITE"LIKE "CCA%"
limit 5;
```

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

Out[11]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Attached above is a list of Top 5 names of a Launch Site which starts with ‘CCA’

Total Payload Mass

In [12]:
```
%%sql
select Customer, sum(PAYLOAD_MASS__KG_) as Total_NASA_CRS_mass
from SPACEXTABLE
where Customer = "NASA (CRS)";
```

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

Out[12]:

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

- total payload mass carried by boosters launched by NASA (CRS)

Average Total Payload Mass by F9 1.1

```
In [13]: %%sql
         select Booster_Version, avg(PAYLOAD_MASS__KG_) as avg_Booster_versionF9_v1_1
         from SPACEXTABLE
         where Booster_Version = "F9 v1.1";

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

Out[13]:

Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

- Average payload mass carried by booster version F9 v 1.1

First Successful Ground Landing Date

```
In [14]:
%%sql
select  Mission_Outcome, min(Date) as Date_First_Succ_Land
from SPACEXTABLE
where Landing_Outcome ='Success (ground pad)';

 * sqlite:///my_data1.db
Done.
Out[14]:
```

Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

- The date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [15]:

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

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

Out[15]:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

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

Total Number of Successful and Failure Mission Outcomes

In [16]:
```
%%sql
select Mission_Outcome, count(Mission_Outcome) as "Total (Success or failure)"
from SPACEXTABLE
GROUP BY MISSION_OUTCOME;
```

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

Out[16]:

Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [17]:

```
%%sql
select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_
from SPACEXTABLE
where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_)
                            from SPACEXTABLE);
```

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

Out[17]:

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

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

2015 Launch Records

```
In [18]: %%sql
SELECT  Date, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
where Landing_Outcome= 'Failure (drone ship)' and Date <= "2015-12-31";

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

Out[18]:

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

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

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

In [19]:

```
%%sql
select Landing_Outcome, count(Landing_Outcome) as "Total Count"
from SPACEXTABLE
where Landing_Outcome = "Failure (drone ship)" or Landing_Outcome = "Success (ground pad)" and
Date between "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome
order by Landing_Outcome desc;
```

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

Out[19]:

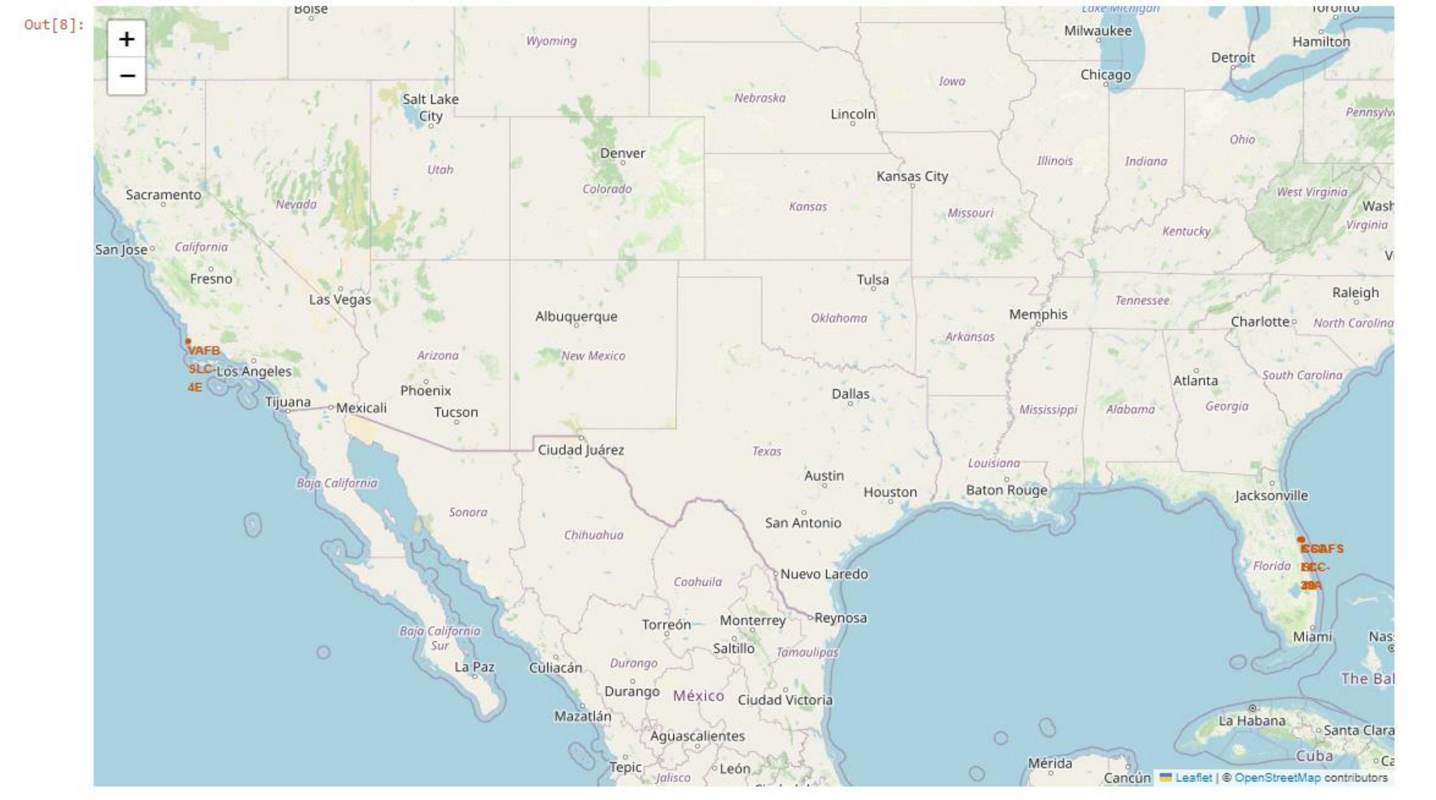
Landing_Outcome	Total Count
Success (ground pad)	3
Failure (drone ship)	5

- Rank landing outcome between the date 2010-06-04 and 2017-03-20, in descending order.

Section 3

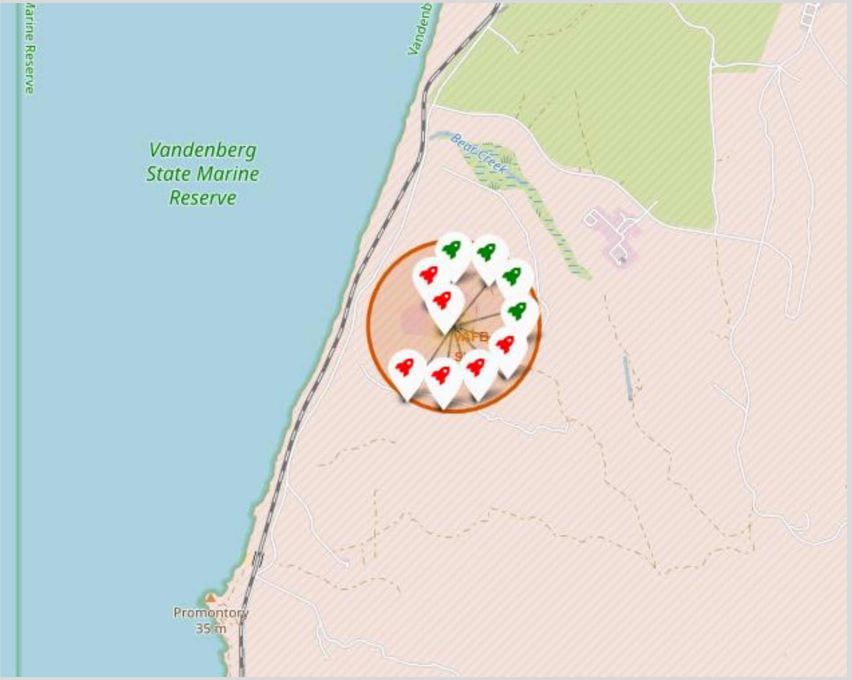
Launch Sites Proximities Analysis

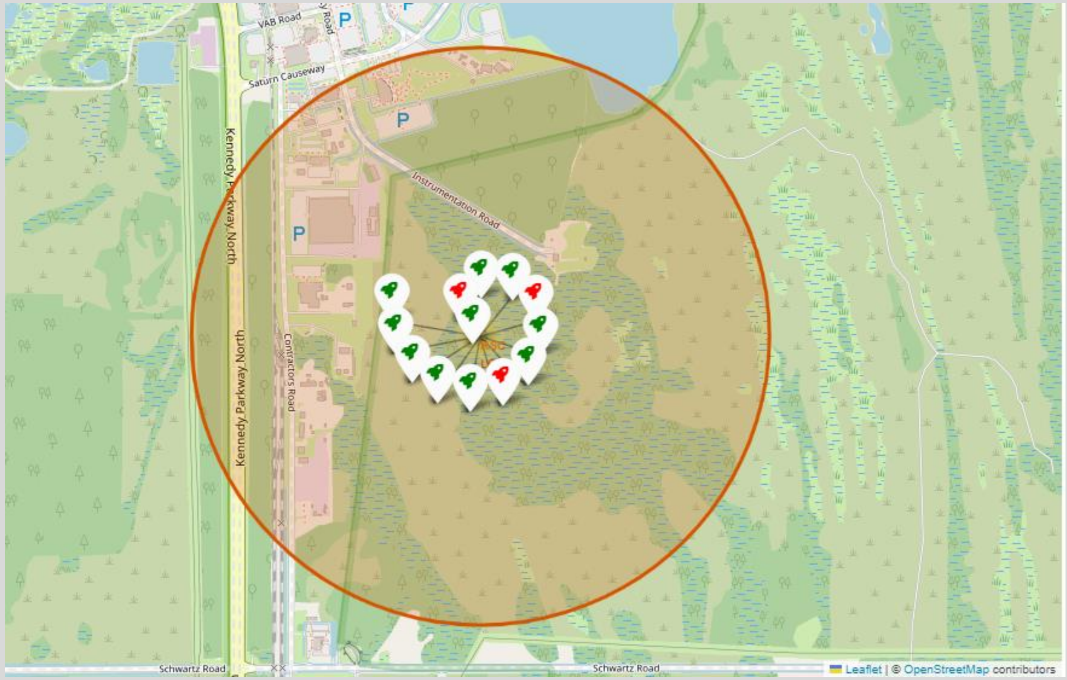
Marked Launch Sites



- It is observed that all the marked launch sites are in very close proximity to the coast.
- 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.

Marked Launch Outcomes

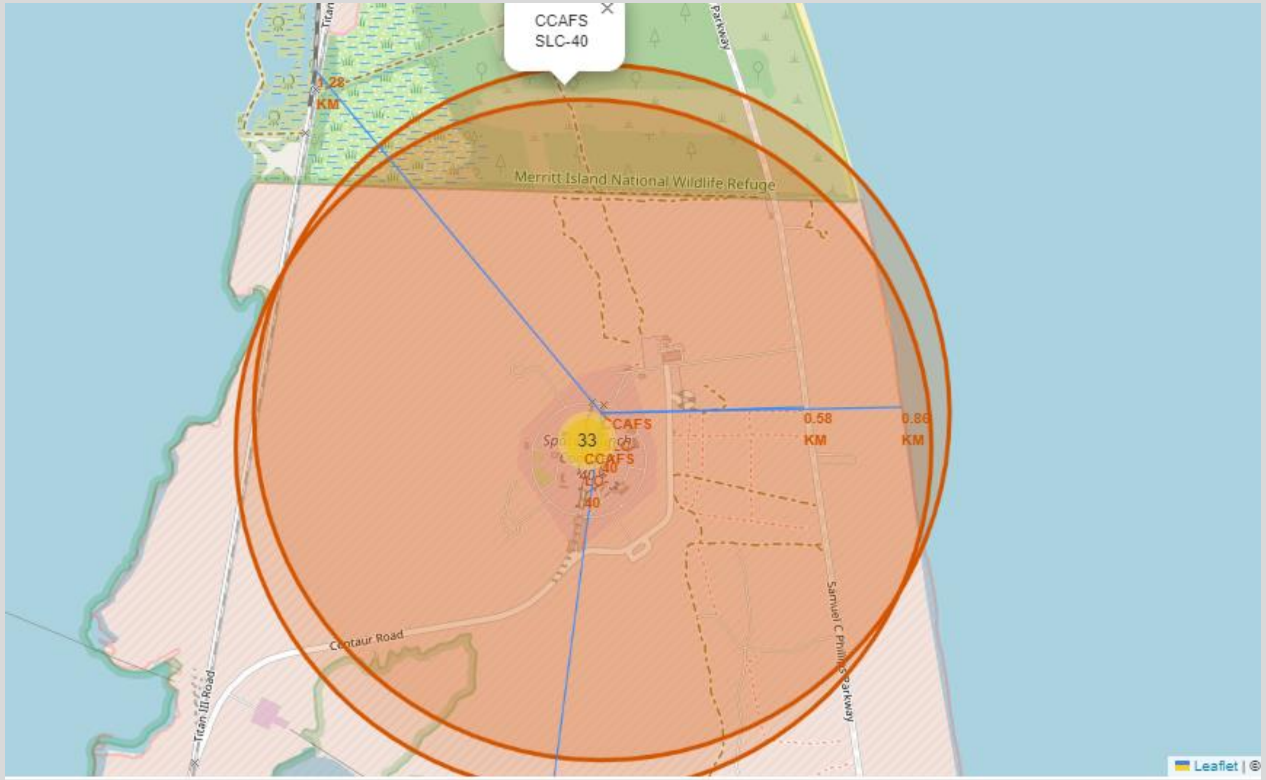




All the launch sites was added:

- Greenmarkers for successful launches
- Redmarkers for unsuccessful launches

Launch Site Proximities

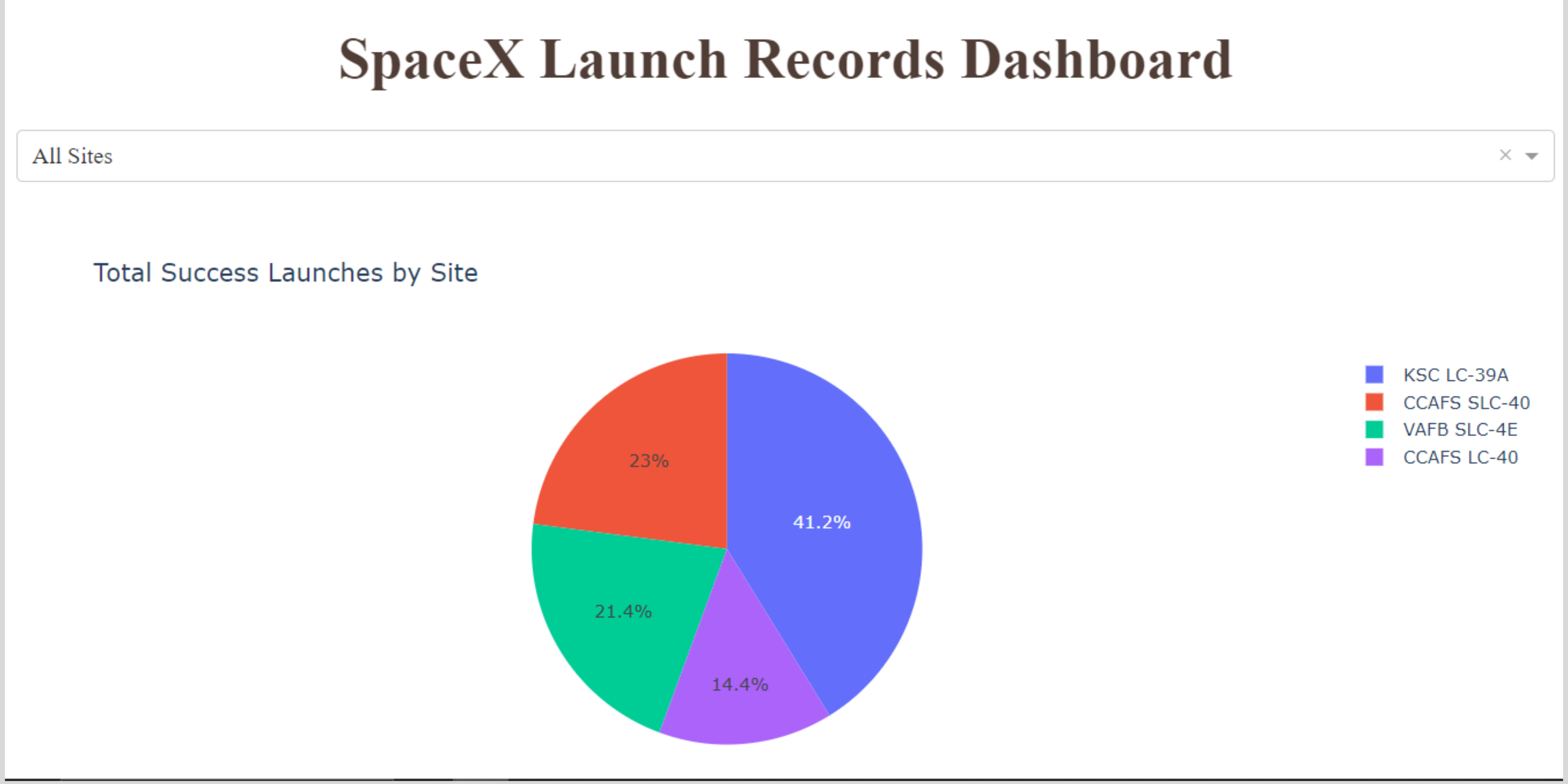


- It is observed that Launch Site are proximity to railways ,highways and the coastline to allow easy transport of heavy rocket components.
- launch sites keep certain distance away from cities which minimizes risks to people and property in case of launch failures and reduces noise impact.

Section 4

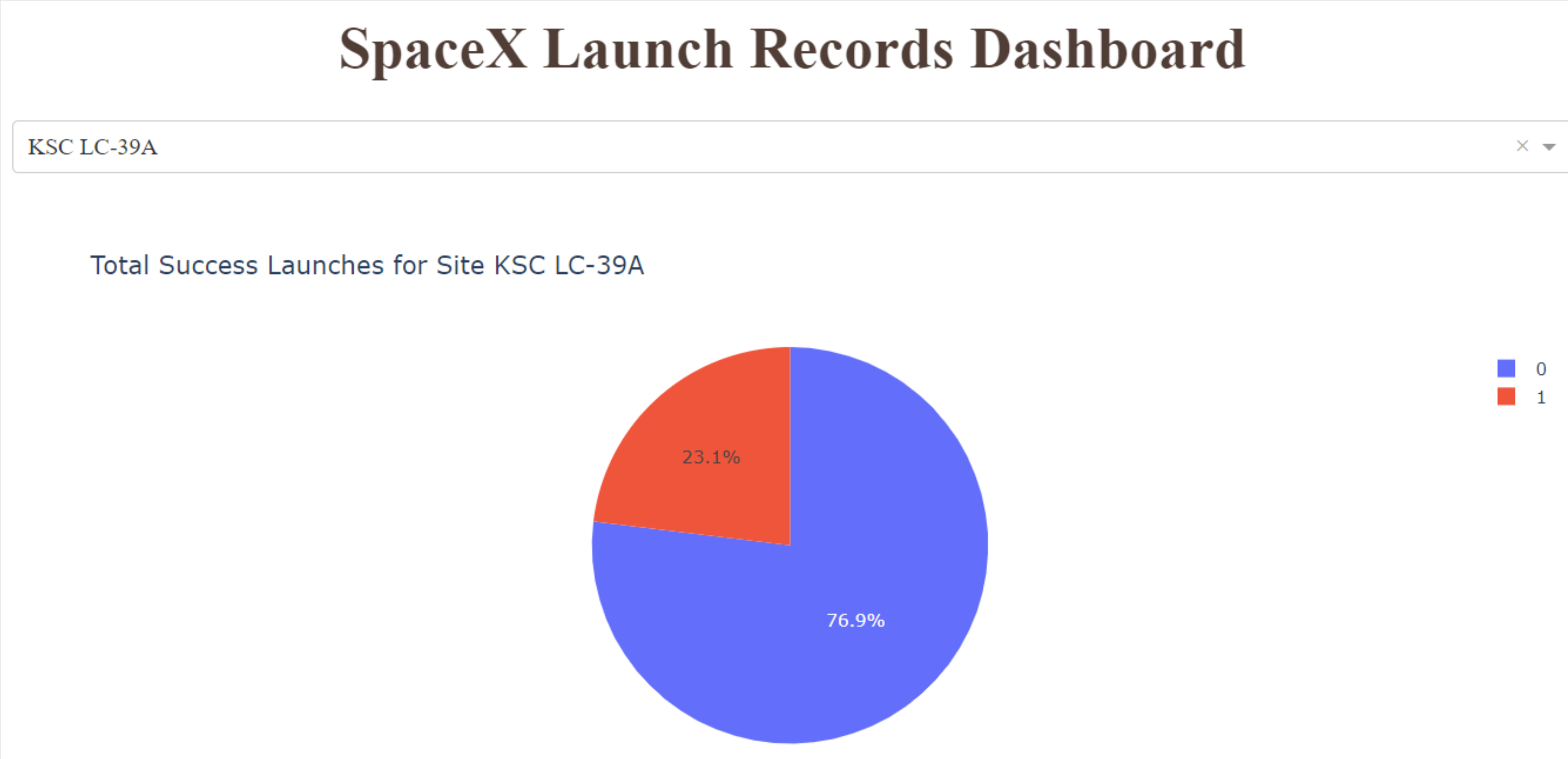
Build a Dashboard with Plotly Dash

Dashboad -Total Success Launches Pie Chart



- Pie Chart showing the succusses percentage achieved by each launch site
- KSC lC-39A had the most successful launches from all the sites

Dashboad - Pie Chart highest success Ratio

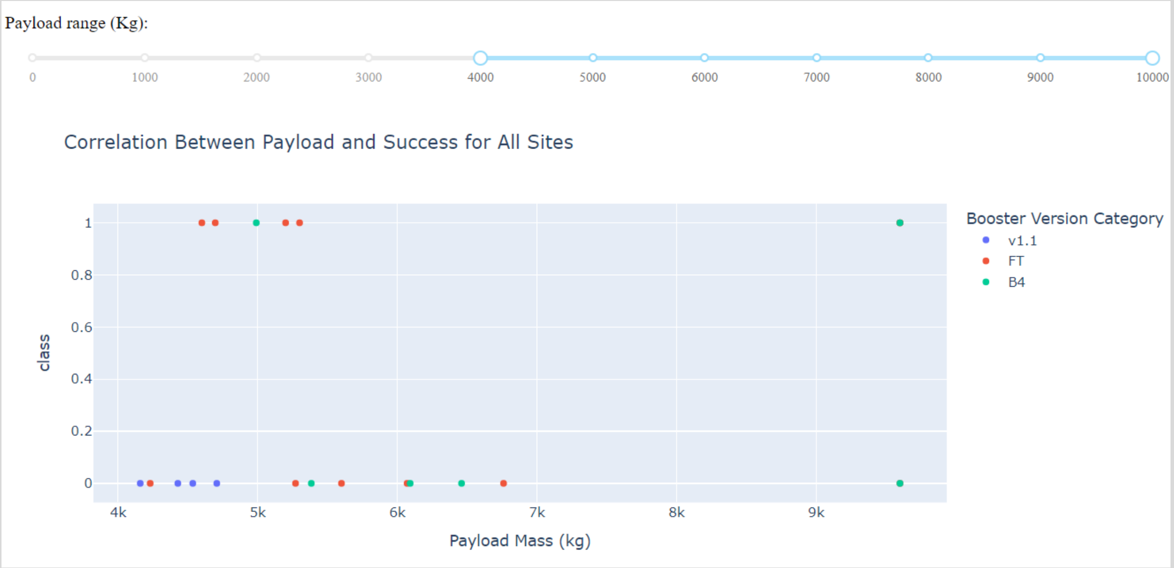


- KSC lC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Dashboad – Payload vs. Launch outcome

Low payload 0-4000kg

heavy payload 4000-10000kg



- It is observed that the success rate for a low weighed payload is higher than the heavy weighed payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Out[46]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

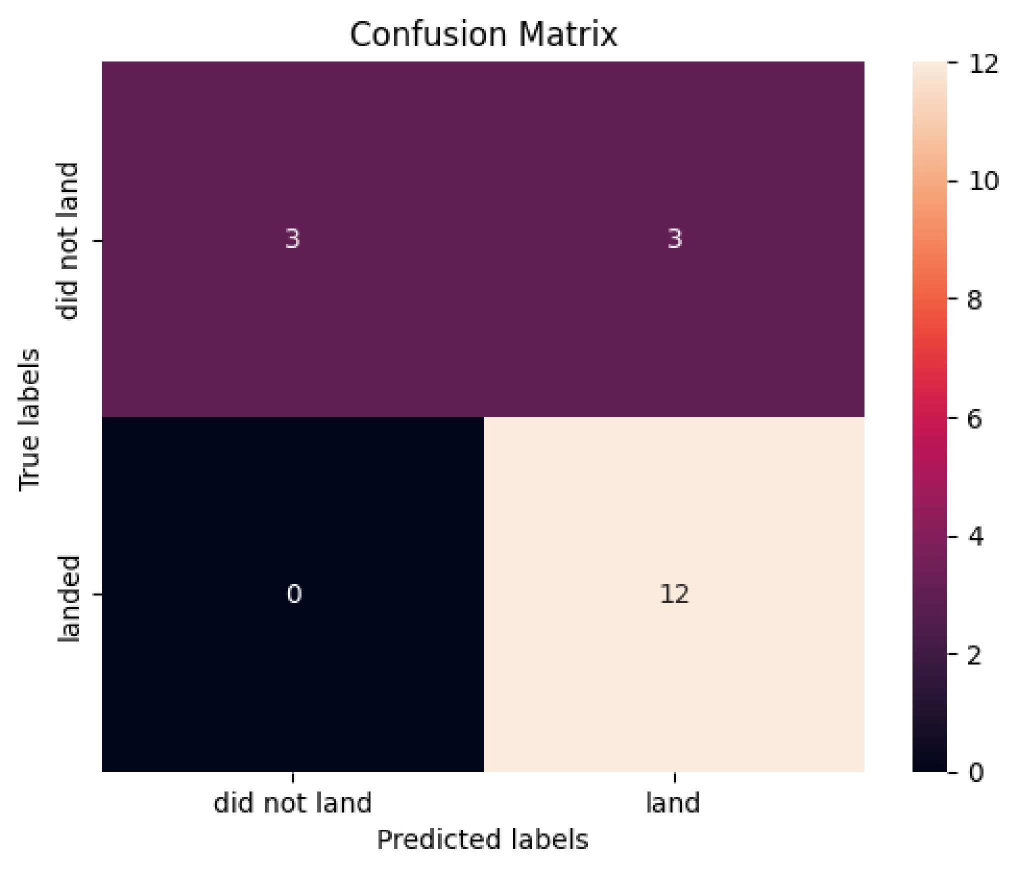
In [52]:

```
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_)
```

```
{'KNeighbors': np.float64(0.8482142857142858), 'DecisionTree': np.float64(0.875), 'LogisticRegression': np.float64(0.84642857
14285713), 'SupportVector': np.float64(0.8482142857142856)}
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 0.5, 'min_samples_leaf': 4, 'min_samples_split': 5,
'splitter': 'random'}
```

- All the models achieved similar performance levels, with comparable scores and accuracy, likely due to the limited size of the dataset. The Decision Tree model had a slight edge over the others, as indicated by its .best_score_, which represents the average score across all cross-validation folds for a particular set of parameters.

Confusion Matrix



- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good

Confusion Matrix Outputs:

- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative

Precision = TP / (TP + FP)

- 12 / 15 = .80

Recall= TP / (TP + FN)

- 12 / 12 = 1

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- 2 * (.8 * 1) / (.8 + 1) = .89

Accuracy= (TP + TN) / (TP + TN + FP + FN) = .833

Conclusion

- **Model Performance:** The models demonstrated similar outcomes on the test set, with the Decision Tree model showing a slight edge over the others.
- **Equator:** Many launch sites are positioned near the equator to take advantage of Earth's rotational speed, providing a natural boost that reduces the need for additional fuel and boosters.
- **Coast:** All launch sites are situated near coastal areas.
- **Launch Success:** The rate of successful launches has increased over time.
- **KSC LC-39A:** This launch site boasts the highest success rate, achieving a 100% success rate for payloads under 5,500 kg.
- **Orbits:** Orbits such as ES-L1, GEO, HEO, and SSO have seen a 100% success rate.
- **Payload Mass:** Generally, across all launch sites, a higher payload mass (kg) corresponds to a greater success rate.

Conclusion

Considerations:

Dataset: Expanding the dataset could enhance the predictive analysis, allowing for a better understanding of whether the results are applicable to a broader set of data.

Feature Analysis / PCA: Performing additional feature analysis or principal component analysis (PCA) could potentially improve the model's accuracy.

XGBoost: This study did not include XGBoost, a highly effective model. It would be worthwhile to explore whether it could outperform the other classification models used.

Appendix

All relevant assets like Python code snippets, queries , charts, Notebook outputs SQL , and data sets included in this presentation can be found on my GitHub .

Thank you!