IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prathibha Harish
21/08/2024

Table of Contents

Executive Summary

- **Summary of Methodologies**

  The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:
  - **Collect** data using SpaceX REST API and web scraping techniques
  - **Wrangle** data to create success/fail outcome variable
  - **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
  - **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
  - **Explore** launch site success rates and proximity to geographical markers
  - **Visualize** the launch sites with the most success and successful payload ranges
  - **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

- **Summary of all Results**

Exploratory Data Analysis:

*Launch success has improved over time
*KSC LC-39A has the highest success rate among landing sites
*Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:

*Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:

*All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX –or a competing company –can reuse the first stage.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API & Web Scraping techniques
- Perform data wrangling
  - Data wrangling was done by filtering data, handling missing values
  - Web scrap Falcon 9 launch records with `BeautifulSoup` - Extract a Falcon 9 launch records HTML table from Wikipedia and Parse the table and convert it into a Pandas data frame
- Exploratory data analysis (EDA) was done using visualization and SQL
- Interactive visual analytics was done using Folium and Plotly Dash
- Perform predictive analysis using classification models

Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection

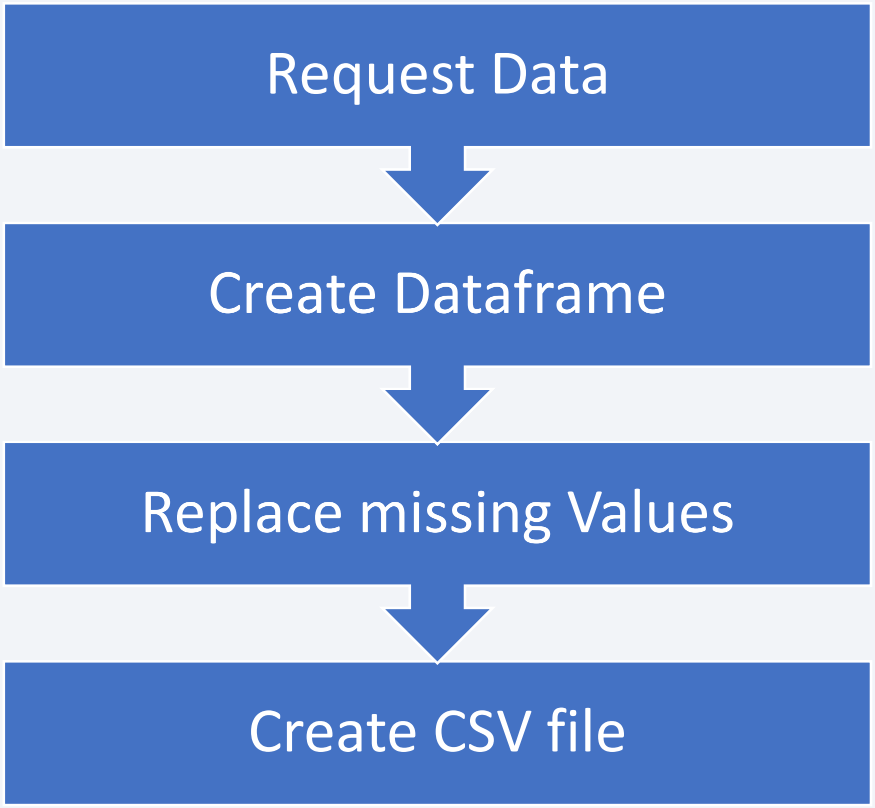
Steps

- **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 – SpaceX API

GitHub URL

- [Applied-DataScience-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project]



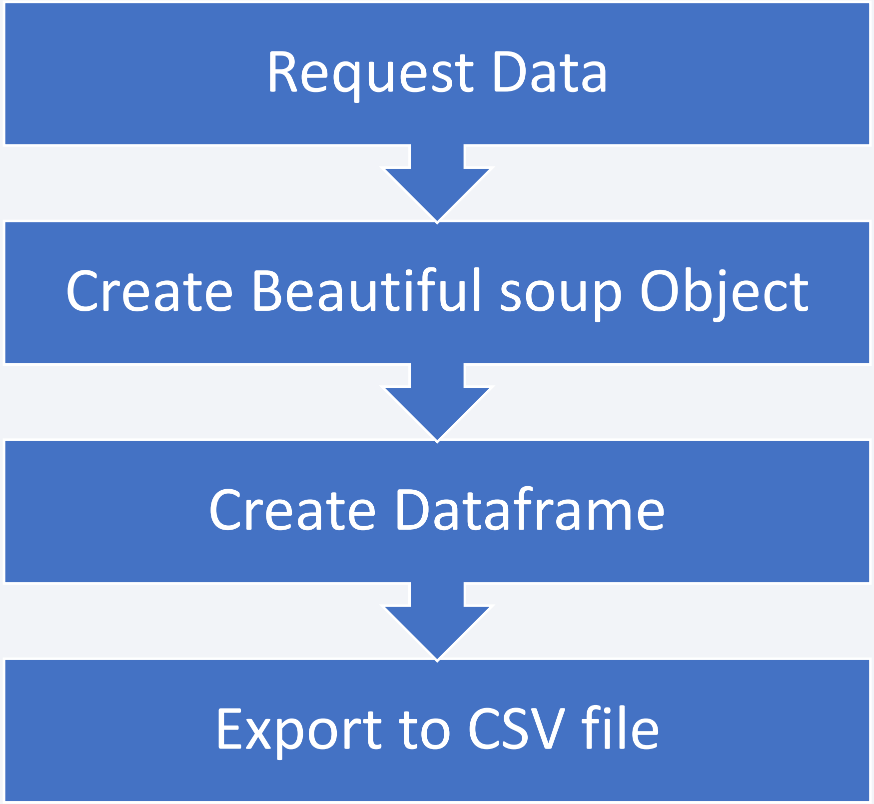
Data Collection - Scraping

Steps

- **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

GitHub URL of the completed web scraping notebook

- [Applied-DataScience-Capstone-Project/jupyter-labs-webscraping.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project]()



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

Steps

- **Perform EDA** and determine data labels
- **Calculate:**# of launches for each site
- # and occurrence of orbit
- # and occurrence of mission outcome per orbit type]
- **Create binary** landing outcome column (dependent variable)
- **Export data** to csv file

GitHub URL

[Applied-DataScience-Capstone-Project/labs-jupyter-spacex-Data wrangling.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project]()

Landing Outcome

Landing was not always successful

- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean
- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS:** meant the mission had a successful landing on a ground pad
- **False RTLS:** represented an unsuccessful landing on a ground pad
- **True ASDS:** meant the mission outcome had a successful landing on a drone ship
- **False ASDS:** represented an unsuccessful landing on drone ship
- **Outcomes converted** into 1 for a successful landing and 0 for an unsuccessful landing

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- **View relationship** by using **scatter plots**. The variables could be useful for machine learning if a relationship exists
- **Show comparisons**among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value.

GitHub URL

Applied-DataScience-Capstone-Project/edadataviz.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project

EDA with SQL

Queries

- Display the Names of unique launch sites.
- Display 5 records where launch site begins with 'CCA'.
- Display Total payload mass carried by boosters launched by NASA (CRS).
- Display the Average payload mass carried by booster version F9 v1.1.
- List the Date of first successful landing on ground pad was achieved.
- List the Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000.
- List the Total number of successful and failed mission outcomes.
- List the Names of booster versions which have carried the maximum payload mass.
- 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.
- 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

Applied-DataScience-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite (1).ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project

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

Map with Folium

- **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 SLC-40** and its proximity to the nearest coastline, railway, highway, and city

GitHub URL

Applied-DataScience-Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Add a callback function to render ’success-pie-chart’ based on selected site dropdown
- Allow user to see successful and unsuccessful launches as a percent of the total

Range Slider of Payload Mass

- Allow user to select payload mass range

Scatter Plot Showing Payload Mass vs. Success Rate by Booster Version

- Add a callback function to render’ success-payload-scatter-chart’
- Allow user to see the correlation between Payload and Launch Success

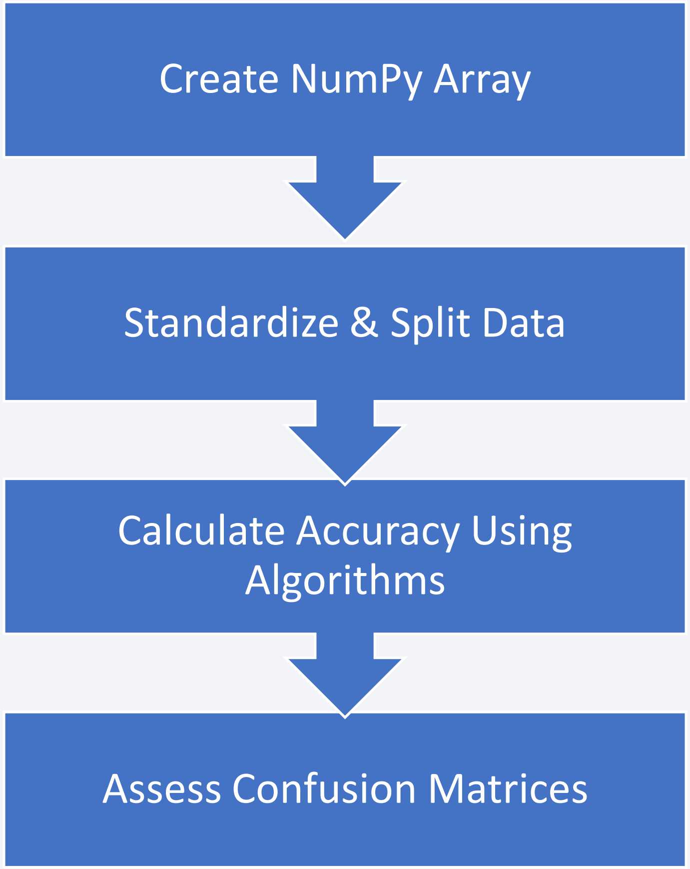
Predictive Analysis (Classification)

Charts

- **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 Jaccard_Score, F1_Score and Accuracy

GitHub URL

- [Applied-DataScience-Capstone-Project/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · prathibhaGithub2023/Applied-DataScience-Capstone-Project]



Results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

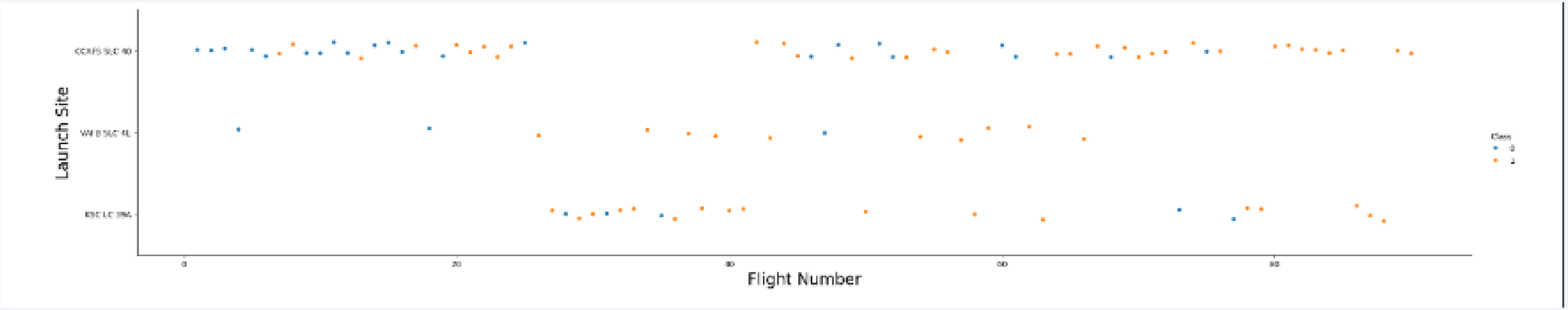
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

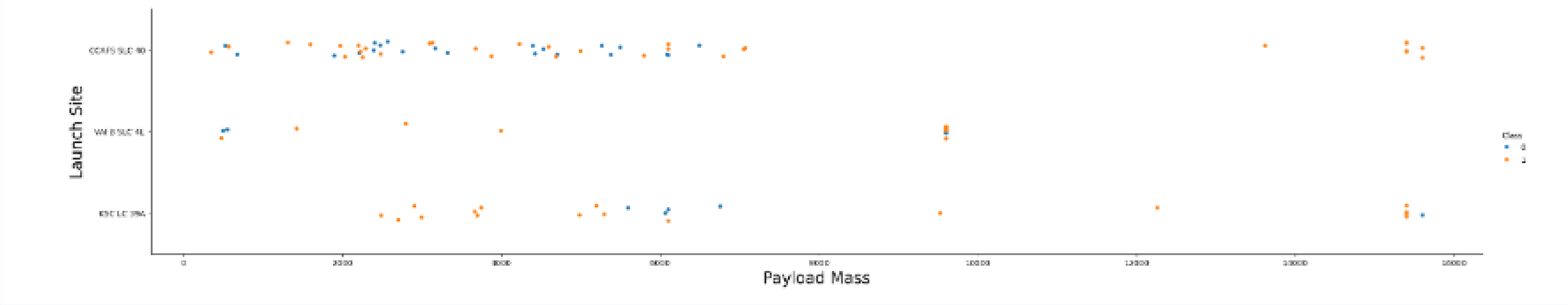
- **Earlier flights** had a **lower success rate** (**blue = fail**)
- **Later flights** had a **higher success rate** (**orange = success**)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

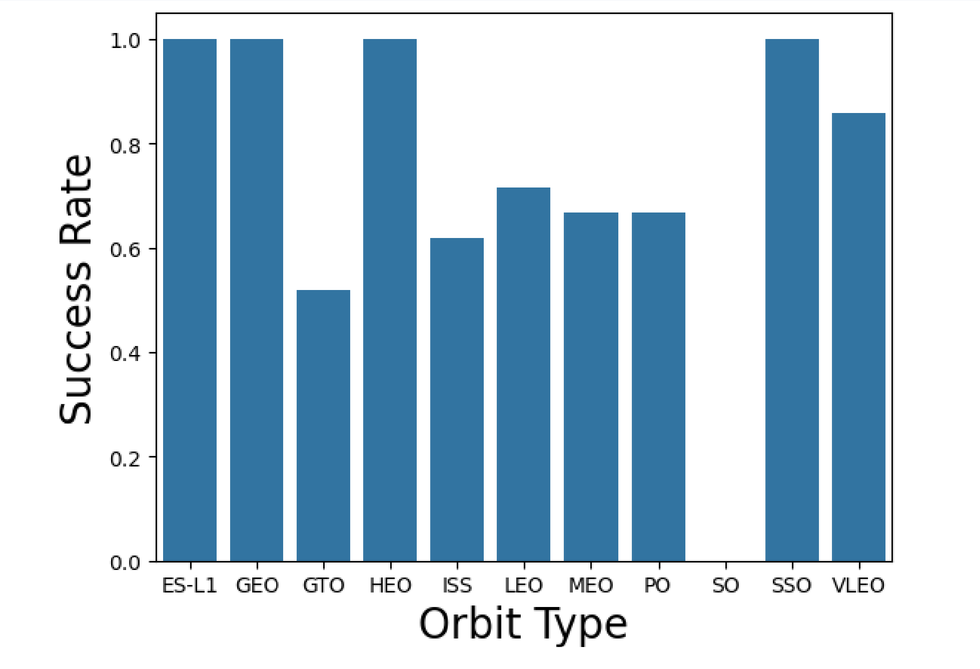
- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate**
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

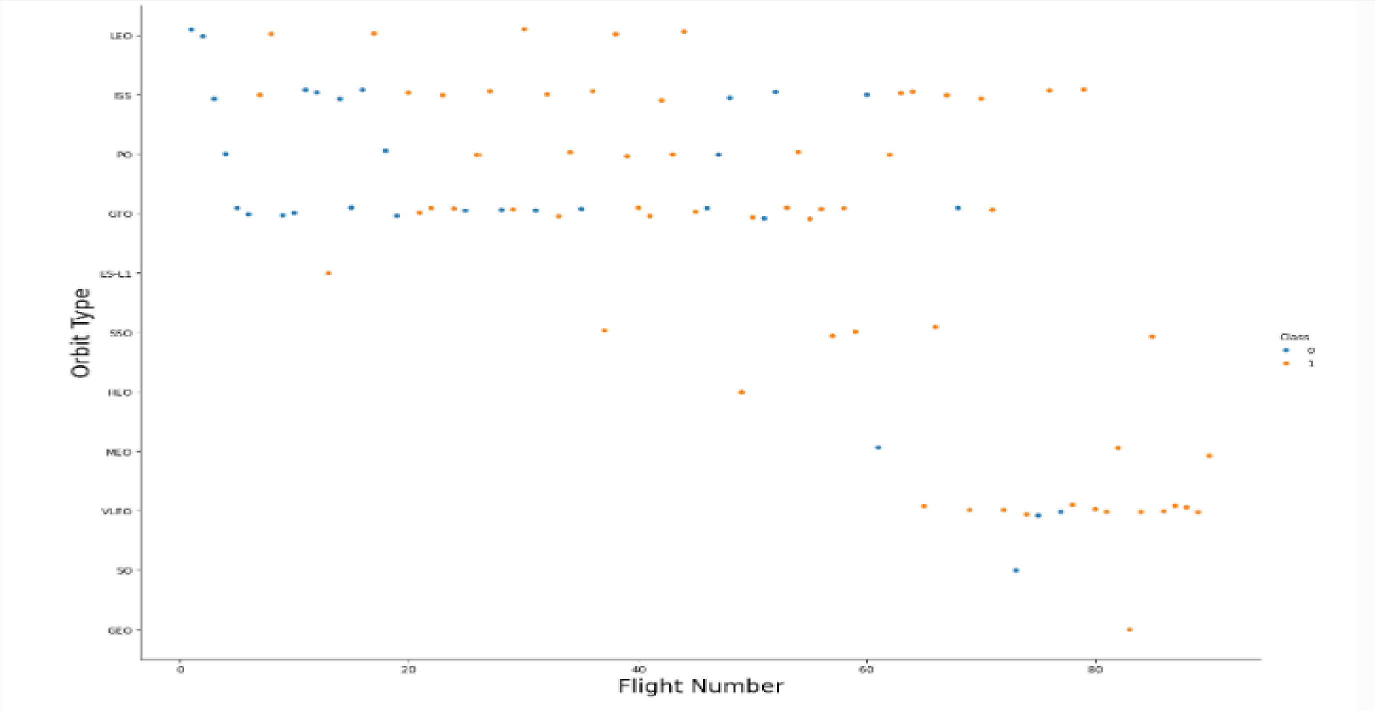
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

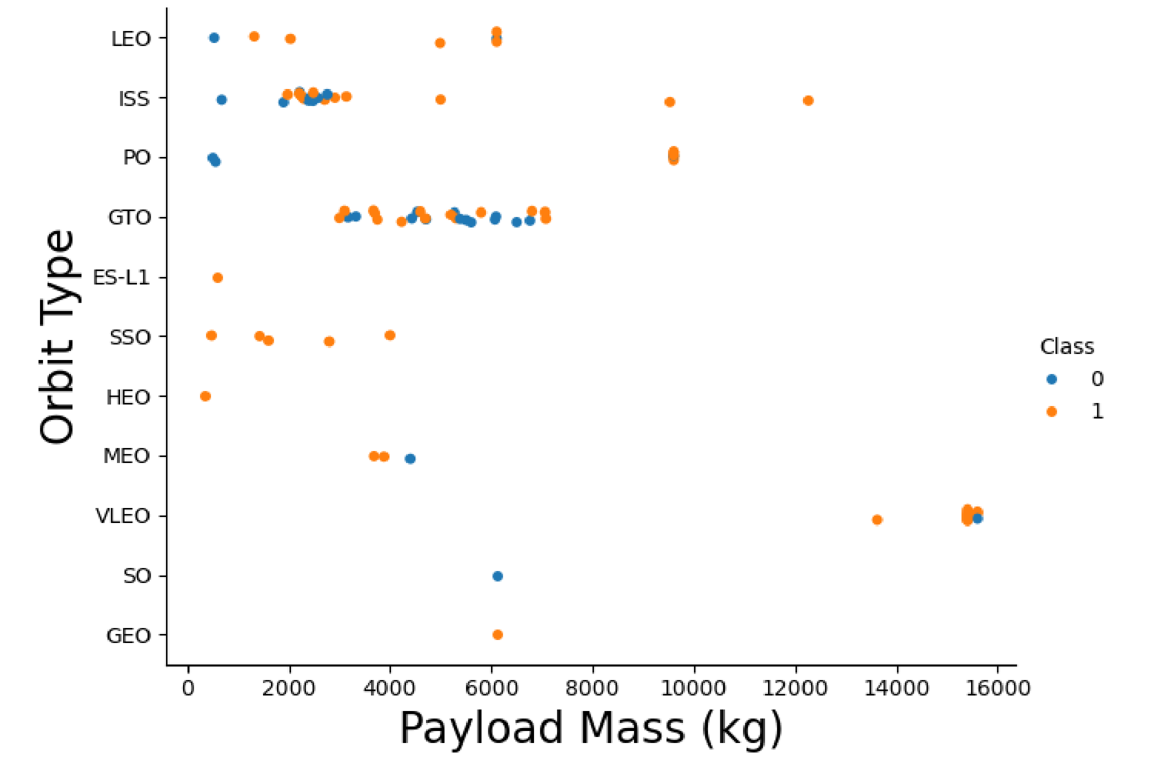
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

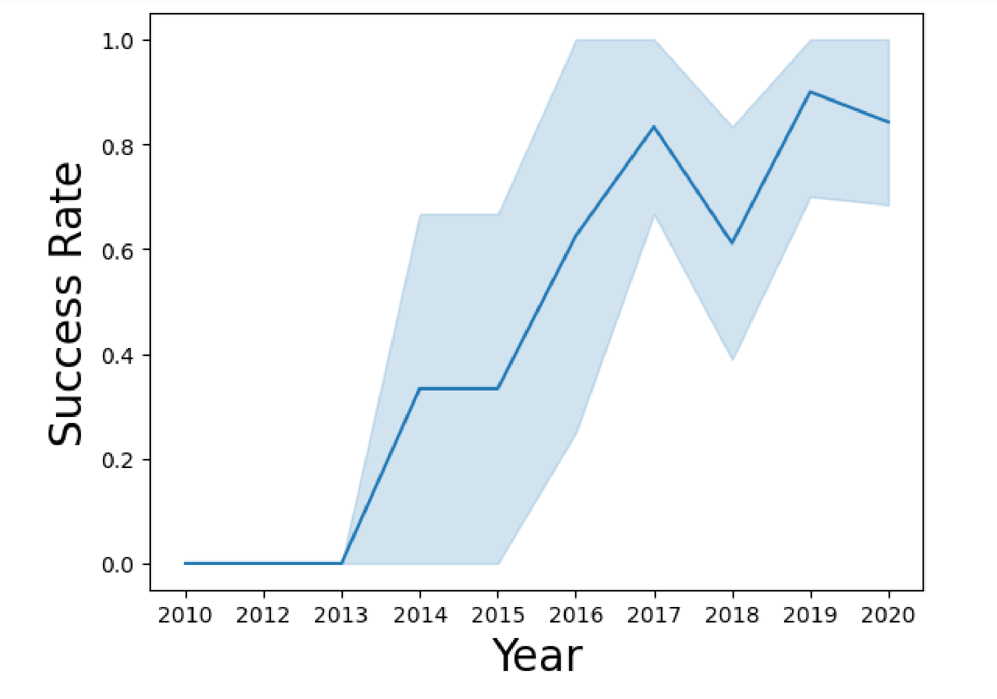
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

Exploratory Data Analysis

- 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

Launch Site Names

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

Task 1

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

In [32]:

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

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

Out[32]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

In [33]:
```
%sql select distinct * from SPACEXTABLE where Launch_Site LIKE '%CCA%' limit 5
```

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

Out[33]:

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

Total Payload Mass

Total Payload Mass

- **45,596 kg** (total) carried by boosters launched by NASA (CRS)

Task 3

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

In [34]:
```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

Out[34]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average Payload Mass

- **2,928.4 kg** (average) payload mass carried by booster version F9 v1.1

Task 4

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

In [35]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

Out[35]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The first successful landing outcome in ground pad was achieved on 2010-06-04

Task 5

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

Hint:Use min function

In [36]:
```
%sql select min(Date) from SPACEXTABLE where Mission_Outcome like '%Success%'
```

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

Out[36]:

min(Date)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Task 6

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

In [37]:
```
%%sql select distinct(Booster_Version) from SPACEXTABLE
where Mission_Outcome = 'Success' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

Out[37]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 100 Success
- 1 Failure

Task 7

List the total number of successful and failure mission outcomes

In [38]:

```
%%sql select count (Mission_Outcome) as sucessful_outcome from SPACEXTABLE
where Mission_Outcome like '%success%'
```

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

Out[38]:

sucessful_outcome
100

In [39]:

```
%%sql select count (Mission_Outcome) as failure_outcome from SPACEXTABLE
where Mission_Outcome like '%fail%'
```

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

Out[39]:

failure_outcome
1

Boosters Carried Maximum Payload

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

In [40]:
```
%%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE
where(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
order by PAYLOAD_MASS__KG_ desc
```

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

Out[40]:

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

2015 Launch Records

Displaying the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

Task 9

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [41]:

```
%%sql select substr(Date,6,2), Booster_version, Launch_Site, Mission_Outcome
from SPACEXTABLE
where Mission_Outcome like '%failure%' AND substr(Date,0,5) = '2015'
```

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

Out[41]:

substr(Date,6,2)	Booster_Version	Launch_Site	Mission_Outcome
06	F9 v1.1 B1018	CCAFS LC-40	Failure (in flight)

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

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

Task 10

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 [42]:

```
%%sql select count (Mission_Outcome) as counts, Mission_outcome from SPACEXTABLE
where Mission_Outcome like '%Success%' AND Date BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY counts DESC
```

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

Out[42]:

counts	Mission_Outcome
30	Success

In [43]:

```
%%sql select count (Mission_Outcome) as counts, Mission_outcome from SPACEXTABLE
where Mission_Outcome like '%fail%' AND Date BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY counts DESC
```

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

Out[43]:

counts	Mission_Outcome
1	Failure (in flight)

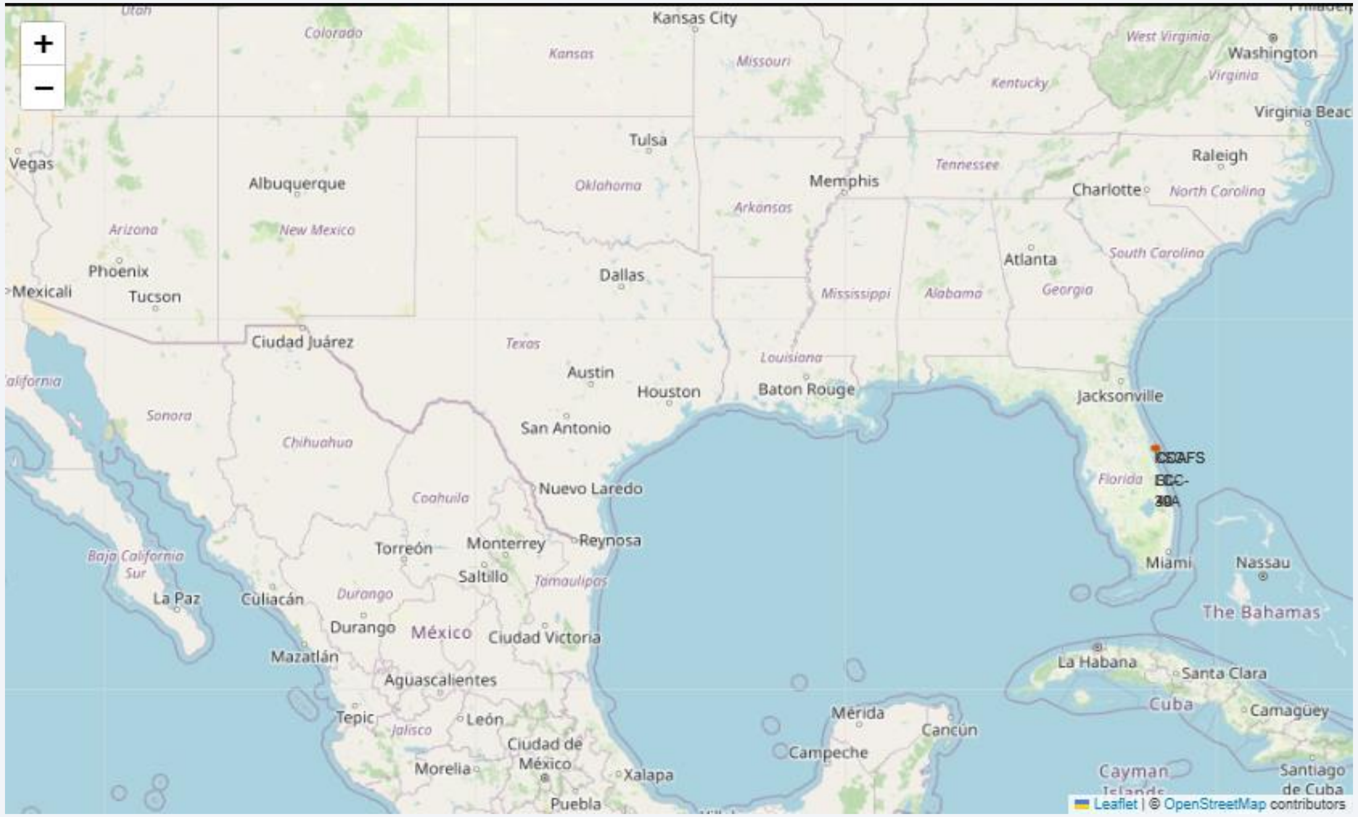
Section 3

Launch Sites Proximities Analysis

Launch Sites

Launch Sites With Markers

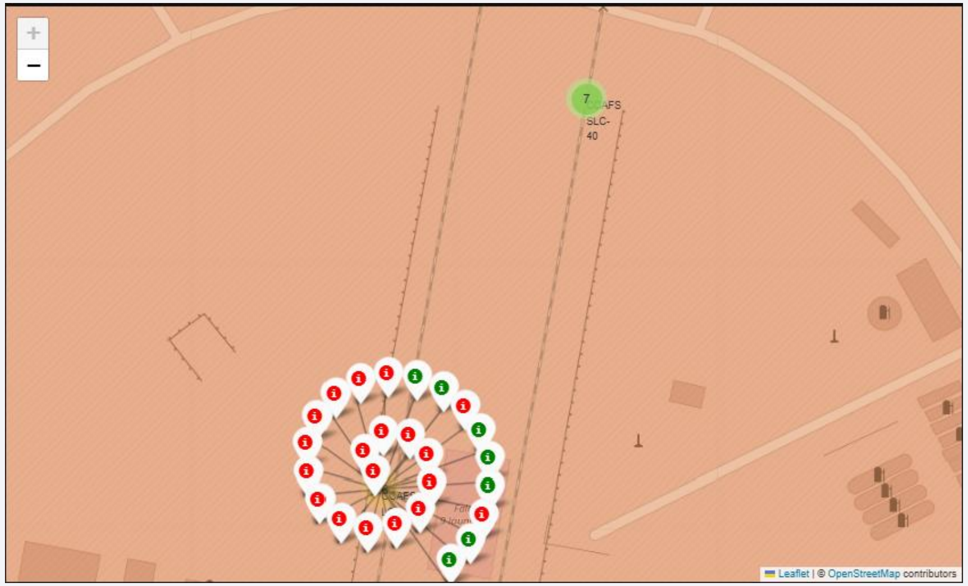
- **Near Equator**: the closer the launch site to the equator, the **easier** it is **to launch** to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an **additional natural boost**-due to the rotational speed of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

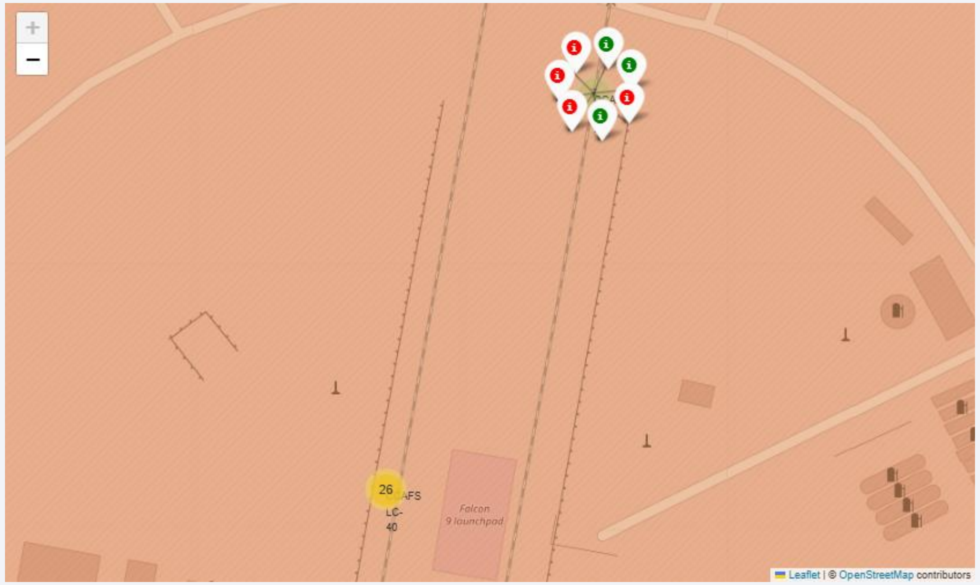


Launch Outcomes

Launch Outcomes At Each Launch Site

- **Green** markers for successful launches
- **Red** markers for unsuccessful launches
- Launch site **CCAFS SLC-40** has a **3/7 success rate** (**42.9%**)

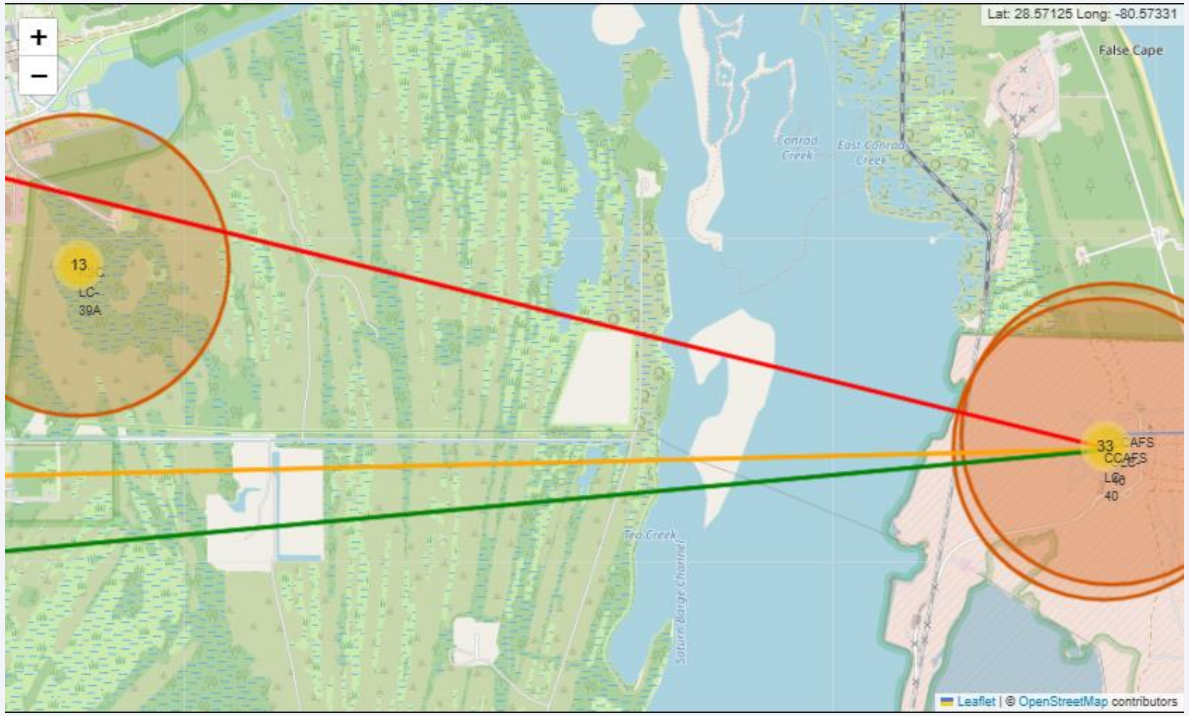




Distance to Proximities

Distances between the launch site CCAFS SLC-40 to its proximities

- **.86 km** from nearest coastline
- **21.96 km** from nearest railway
- **23.23 km** from nearest city
- **26.88 km** from nearest highway
- Regarding Kennedy Space Center: Distance to railways, highways and cities are good enough, but the distance to coastline is very less. Also regarding the fact that rockets will head towards the east.
- Regarding Vandenberg AFB: The only larger place to the East of the AFB is Lompoc, CA



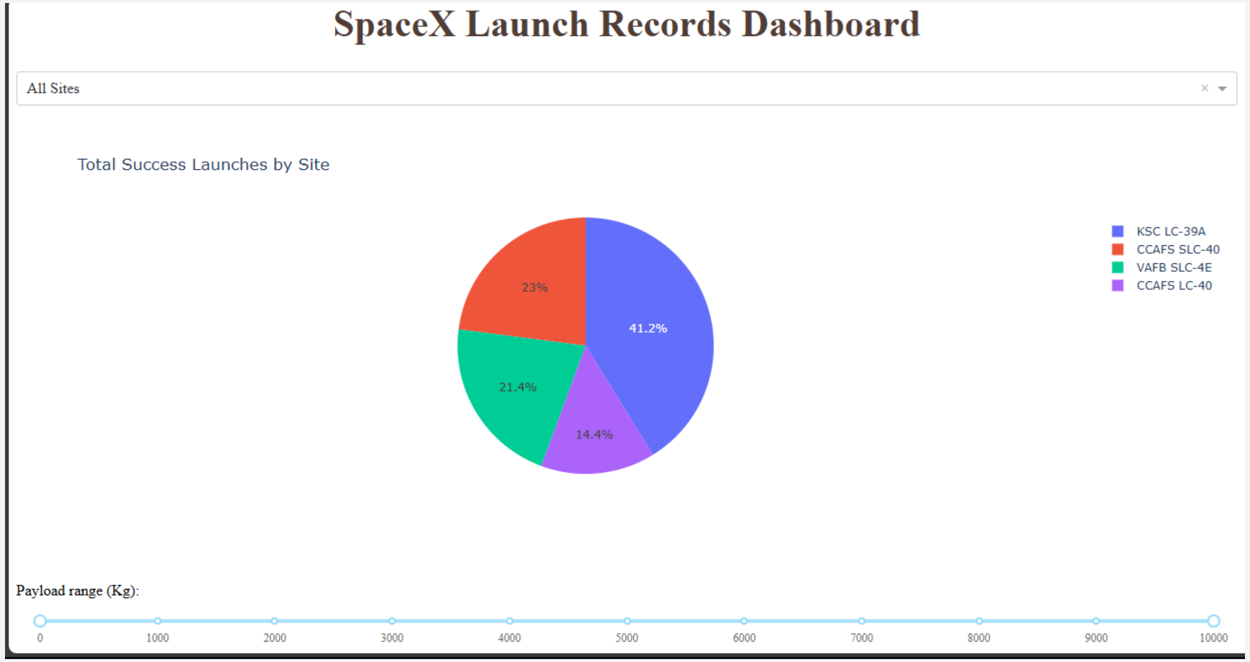
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count by Sites

Total Success Launches by sites as Percentage

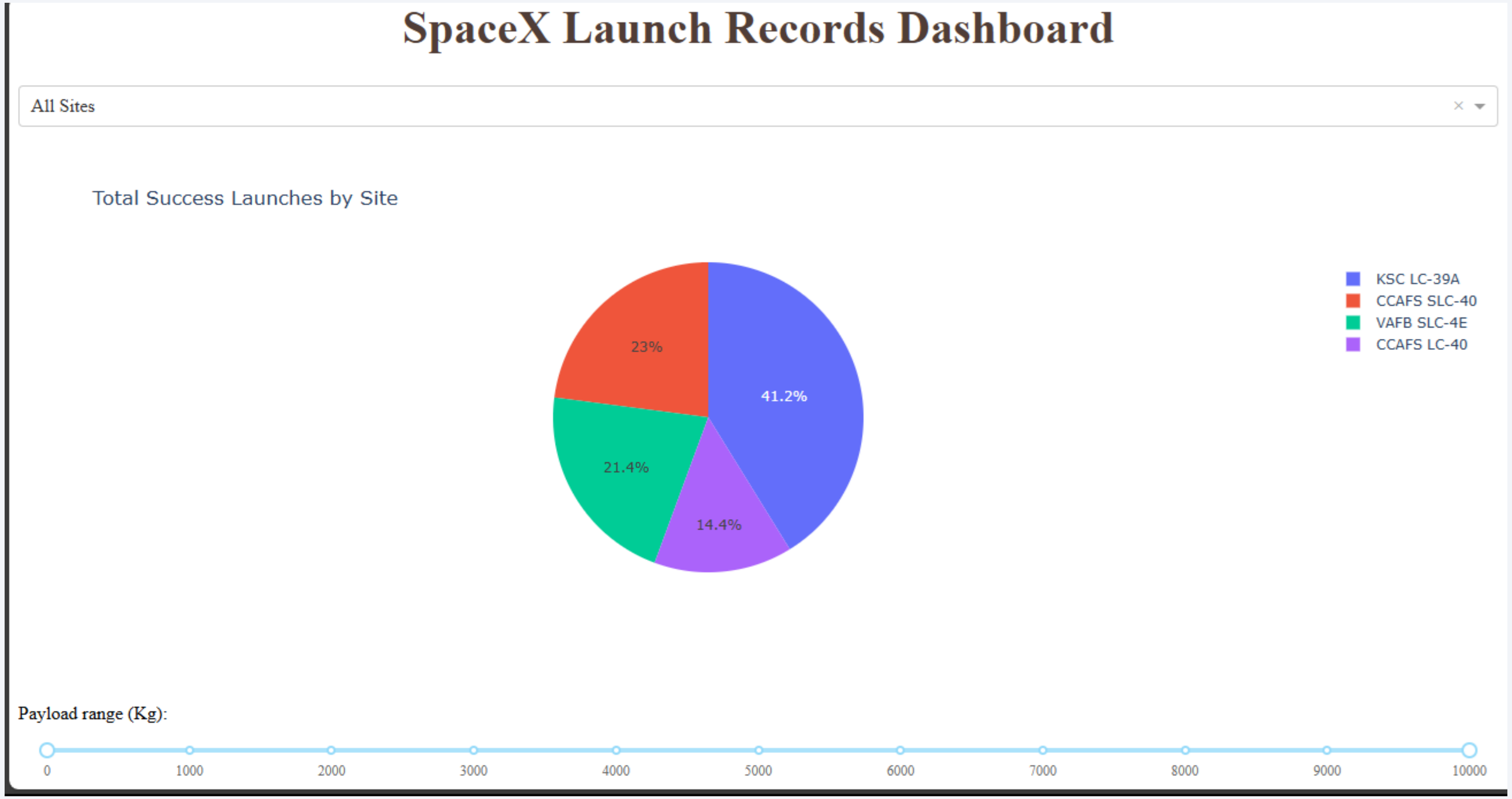
- **KSC LC-39A** has the **success** rate of 41.2%
- **CCAFS SLC-40** has the **success** rate of 23%
- **VAFB SLC-4E** has the **success** rate of 21.4%
- **CCAFS LC-40** has the **success** rate of 14.4%



Highest Launch Success Ratio

Highest launch Success as Percent of Total

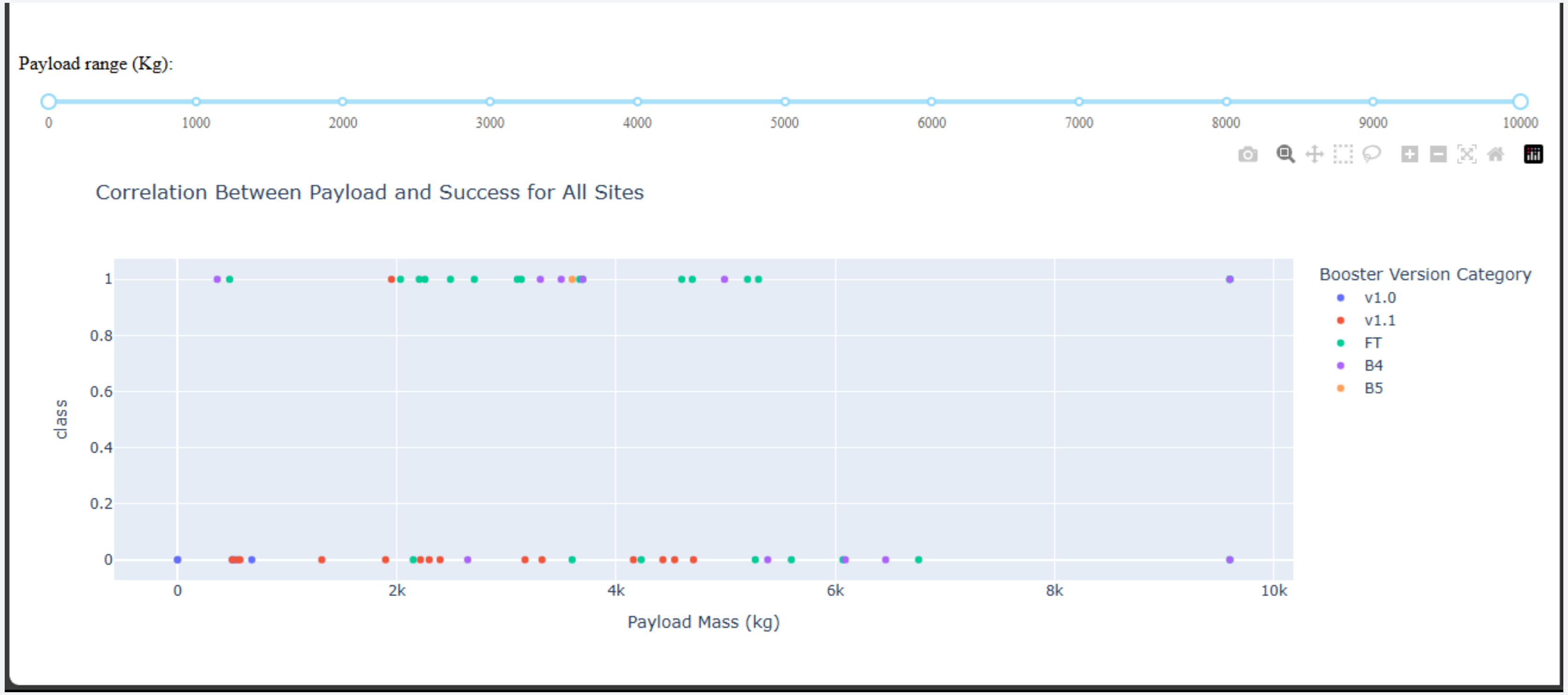
- **KSC LC-39A** has the **highest success rate** amongst launch sites (**41.2%**)



Payload vs. Launch Outcome

Correlation between Payload and success for all sites with different payload 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

- All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The **Decision Tree model slightly outperformed**
- .best_score_ is the average of all cv folds for a single combination of the parameters

Out[48]:

	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	0.888889	0.833333

In [46]:

```
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 0.8892857142857145
Best params is : {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_sampl
es_split': 5, 'splitter': 'best'}
```

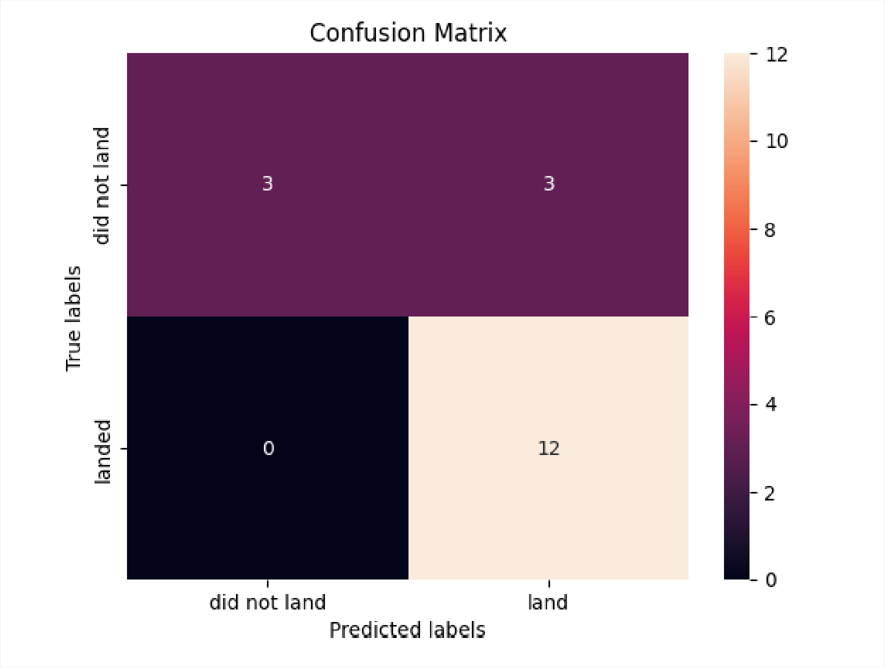
Confusion Matrix

Performance Summary

- 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

Conclusions

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator**: Most of the launch sites are near the equator for an additional natural boost due to the rotational speed of earth which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!