IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jackson C Northrup

9/20/2024

Outline

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

Executive Summary

- Summary of Methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of Results
  - Exploratory Data Analysis Results
  - Interactive Dashboards Screenshots
  - Results of Predictive Analystics

Introduction

- SpaceX is a leader in the Space Industry, They seek a way to make space travel more affordable for everyone. One way they are able to do this is the relatively inexpensive rocket launches they preform due to the reusable first stage of its Falcon 9.
- Using data that is available to the public we can use machine learning techniques that will help predict if the first stage of a launch will land, which can be used to predict the price of said launch.

Section 1

Methodology

Executive Summary

- Data collection methodology:
  - Data was collecting from SpaceX API, we then Normalized the .json into a pandas DataFrame.
  - Webscraping was also preformed
- Perform data wrangling
  - Successful landings were listed as a 1 and unsuccessful was set as 0
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - we created a training and test set and checked multiple methods to see which would preform the best

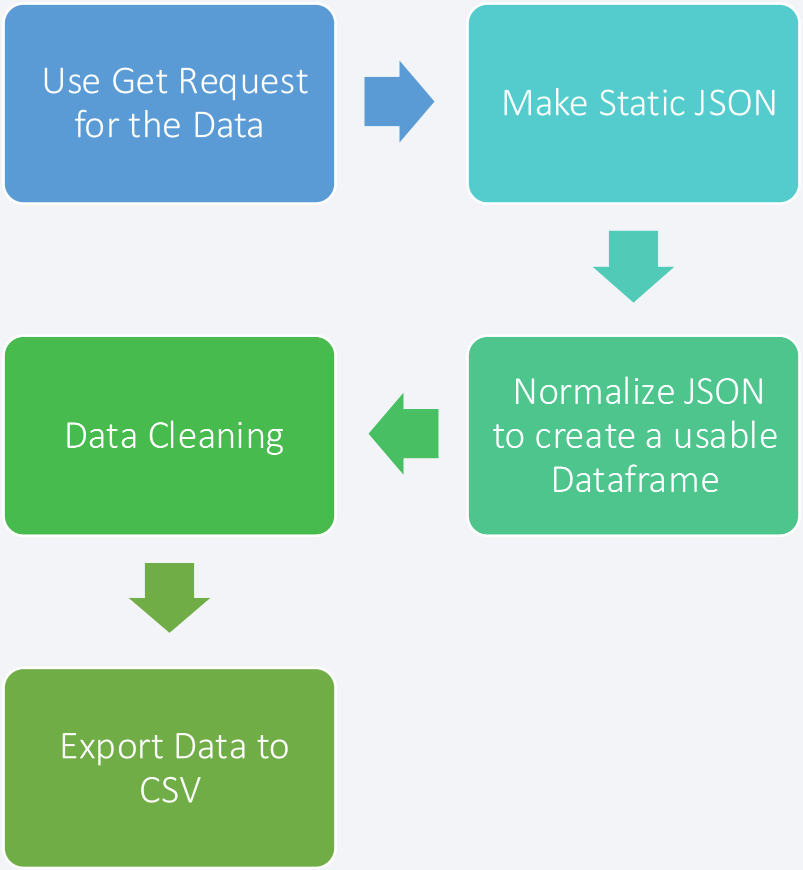
Data Collection

- Data was Collected with many methods
- Using a get request we got data from SpaceX REST API, we then normalized the .json() response into a pandas DataFrame.
- Data was then cleaned and missing values were addressed
- Web scraping was then preformed to get Falcon 9 launch records from Wikipedia using Beautiful Soup to get the data into a usable format
- This was done so that we could parse that data into a usable format for the analysis.

Data Collection – SpaceX API

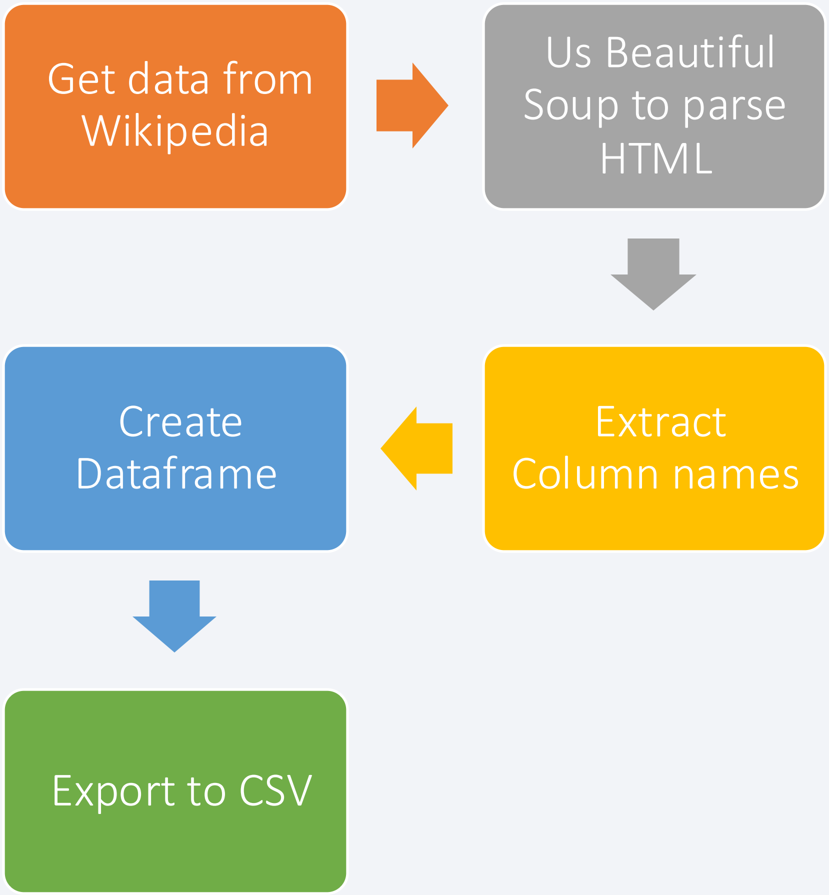
- Use a get request to grab the data from the SpaceX API, get it into a Pandas DataFrame, and finally do some minimal data cleaning.

- The notebook is located here [Github notebook for Data collection from the API]



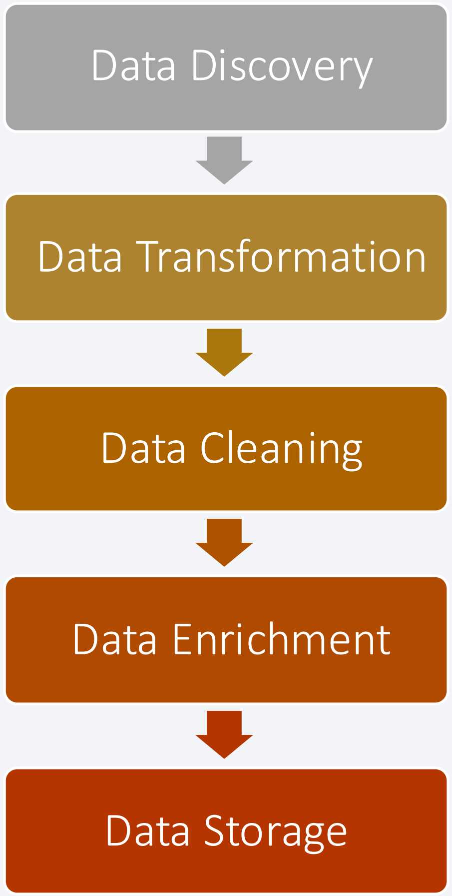
Data Collection - Scraping

- Using Beautiful Soup, we Webscrapped a table from Wikipedia for Falcon 9 Launch Records
- The Notebook for this process is here



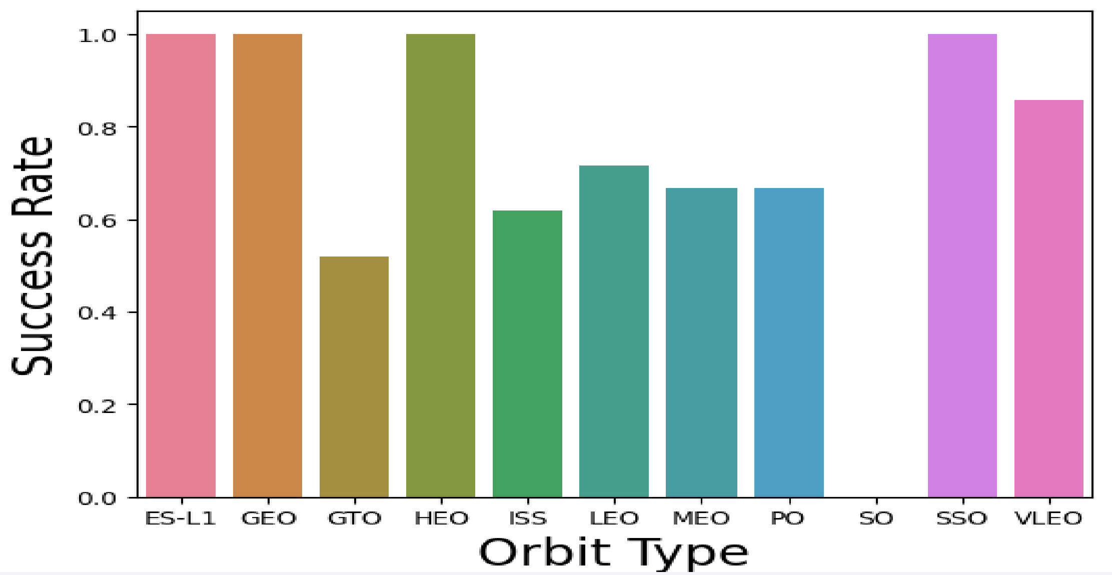
Data Wrangling

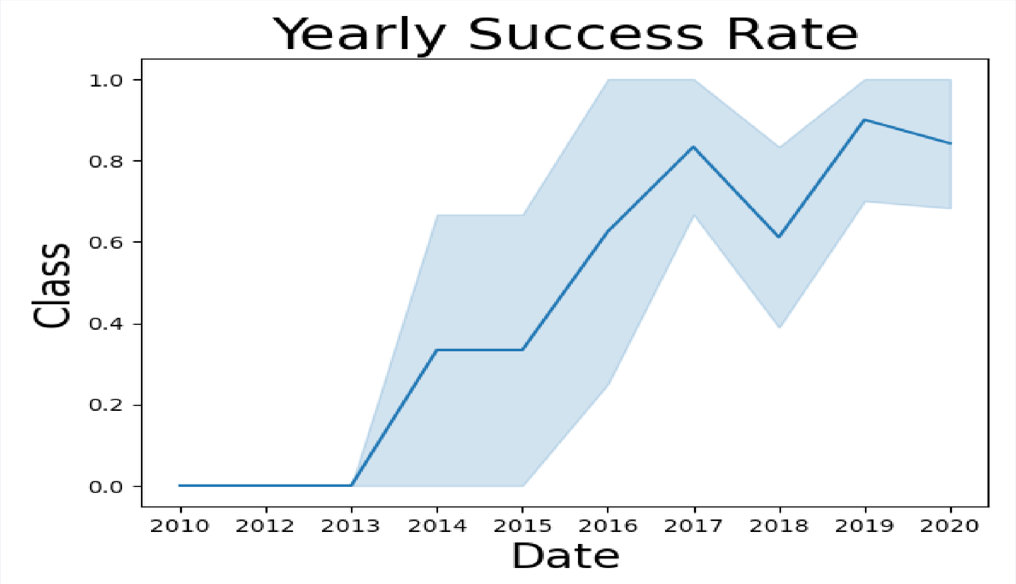
- We determined the number of launches from each site, as well as the number of different types of orbits, and the outcomes of those launches.
- [The Github Repository link]



EDA with Data Visualization

- We examined the relationship Flight Number and Launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend





[Github Repository link]

EDA with SQL

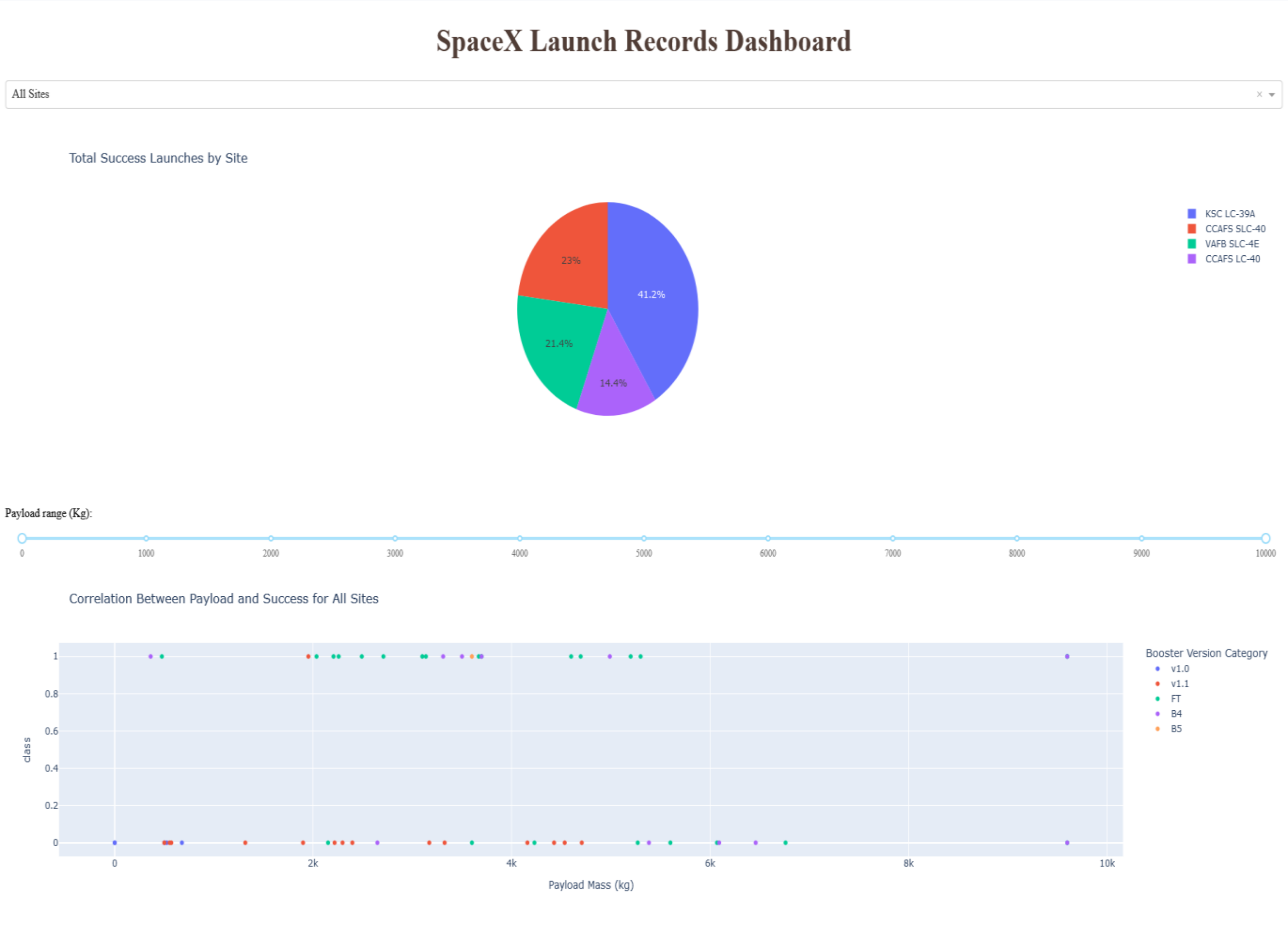
- Using SQLlite we were able to use SQL commands without leaving the Jupyter notebook
- Using Queries we did some basic EDA and found
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- GitHub notebook link

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.
- the Github URL

Build a Dashboard with Plotly Dash

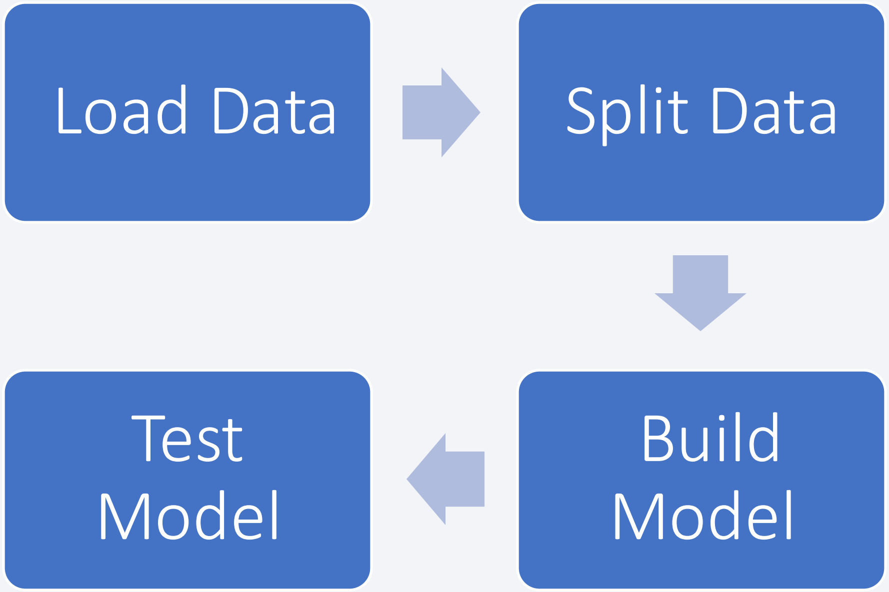
- we have a pie chart showing successful launches, a number line for payload range scatter plot showing payload vs successfulness class. There is a button to change between all sites and specific sites
- It is good to see the proportion of successful launches, as well as knowing the relationship between payload(kg) and launches outcome
- [The GitHub URL]



Predictive Analysis (Classification)

- The Data was Loaded using numpy and pandas, the data was transformed and split into training/testing sets
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- Accuracy on the testing set was the metric used for feature engineering and tuning the algorythm

- GitHub URL



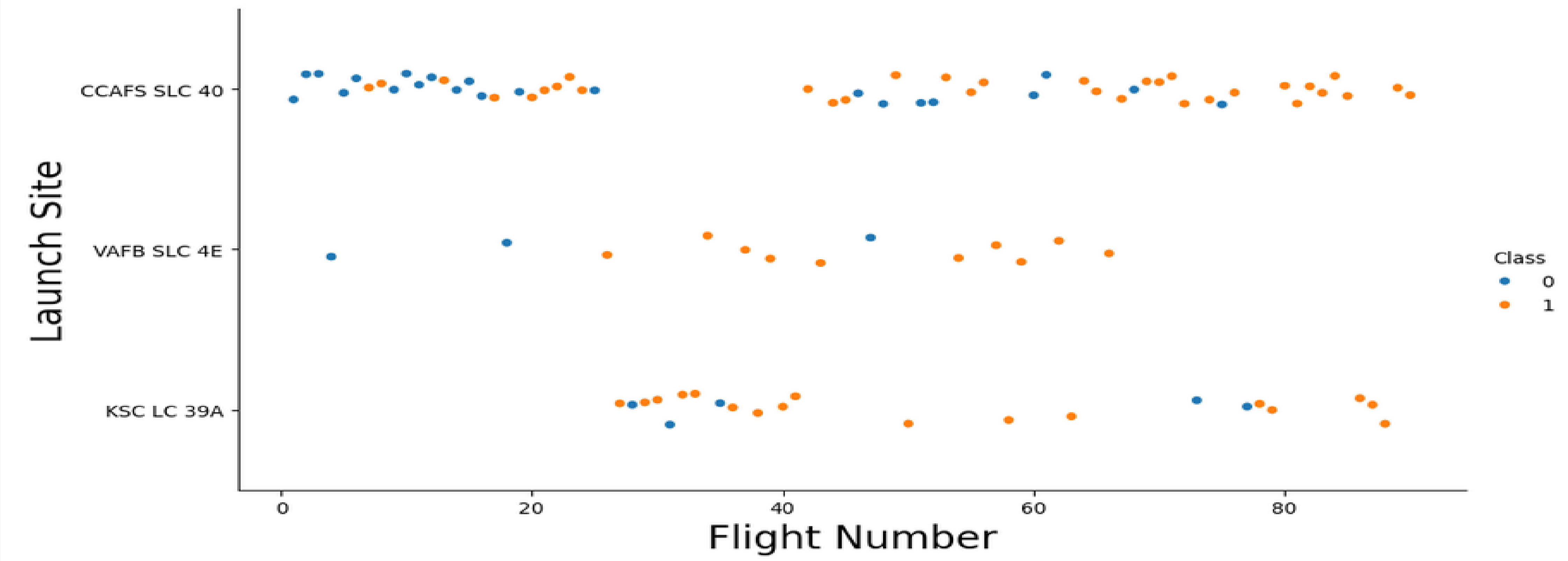
Results

- 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.

Section 2

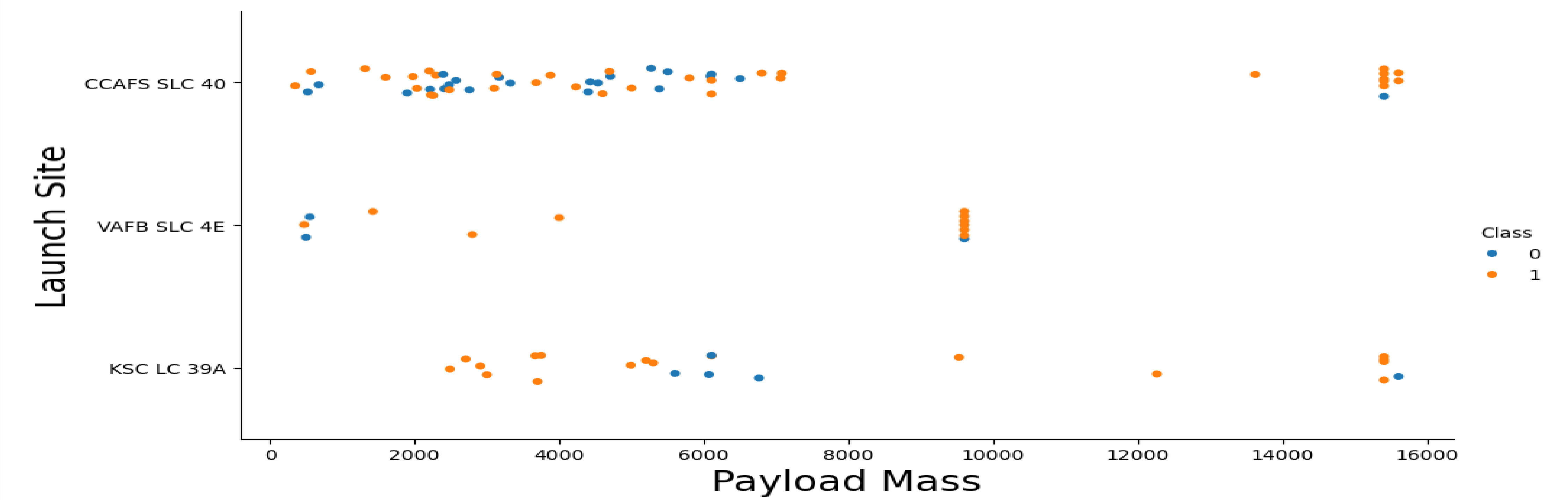
Insights drawn from EDA

Flight Number vs. Launch Site



We see that most unsuccessful launches happened at lower flight numbers.

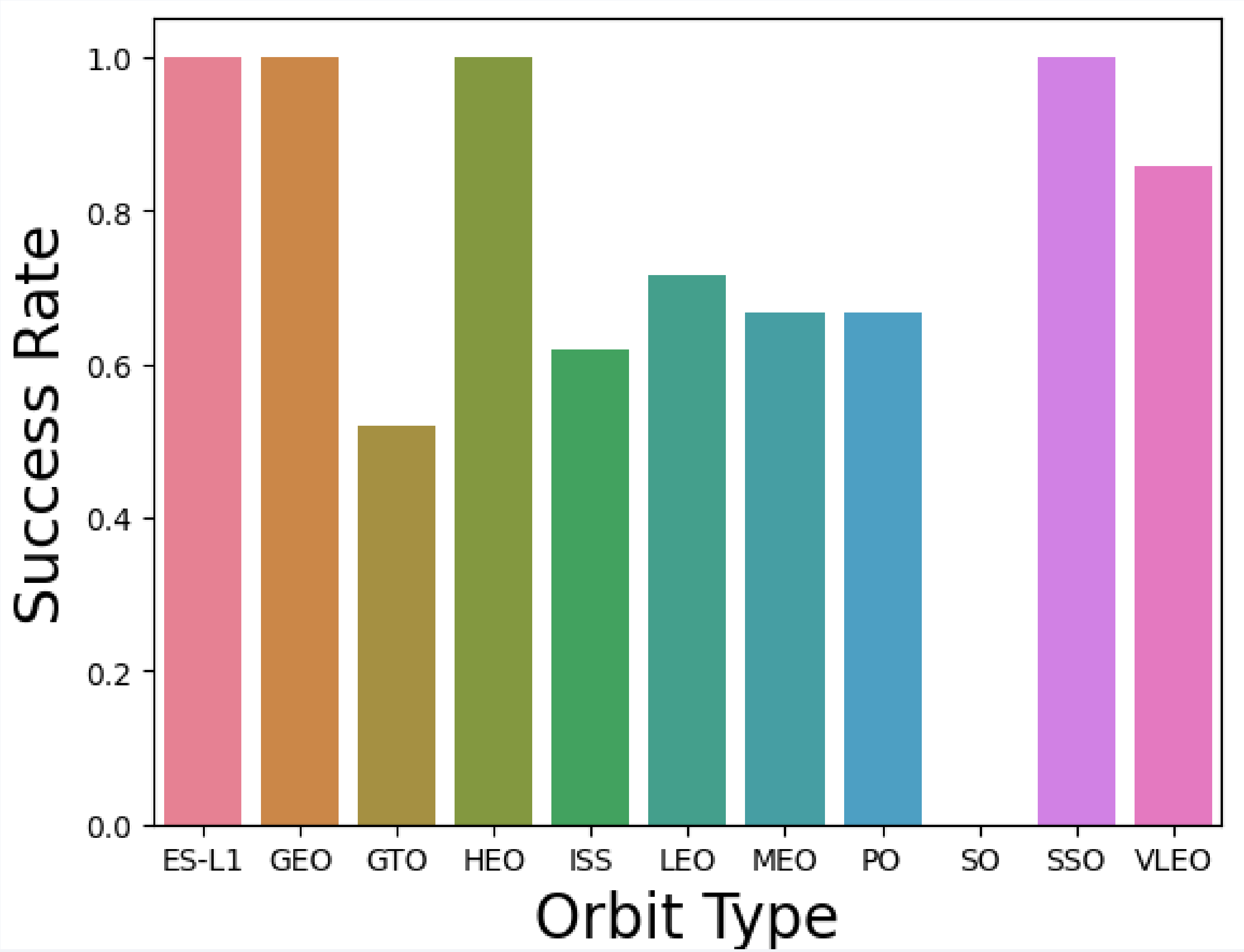
Payload vs. Launch Site



It seems that there is a positive correlation between payload mass and launch success at the CCAFS SLC 40 location

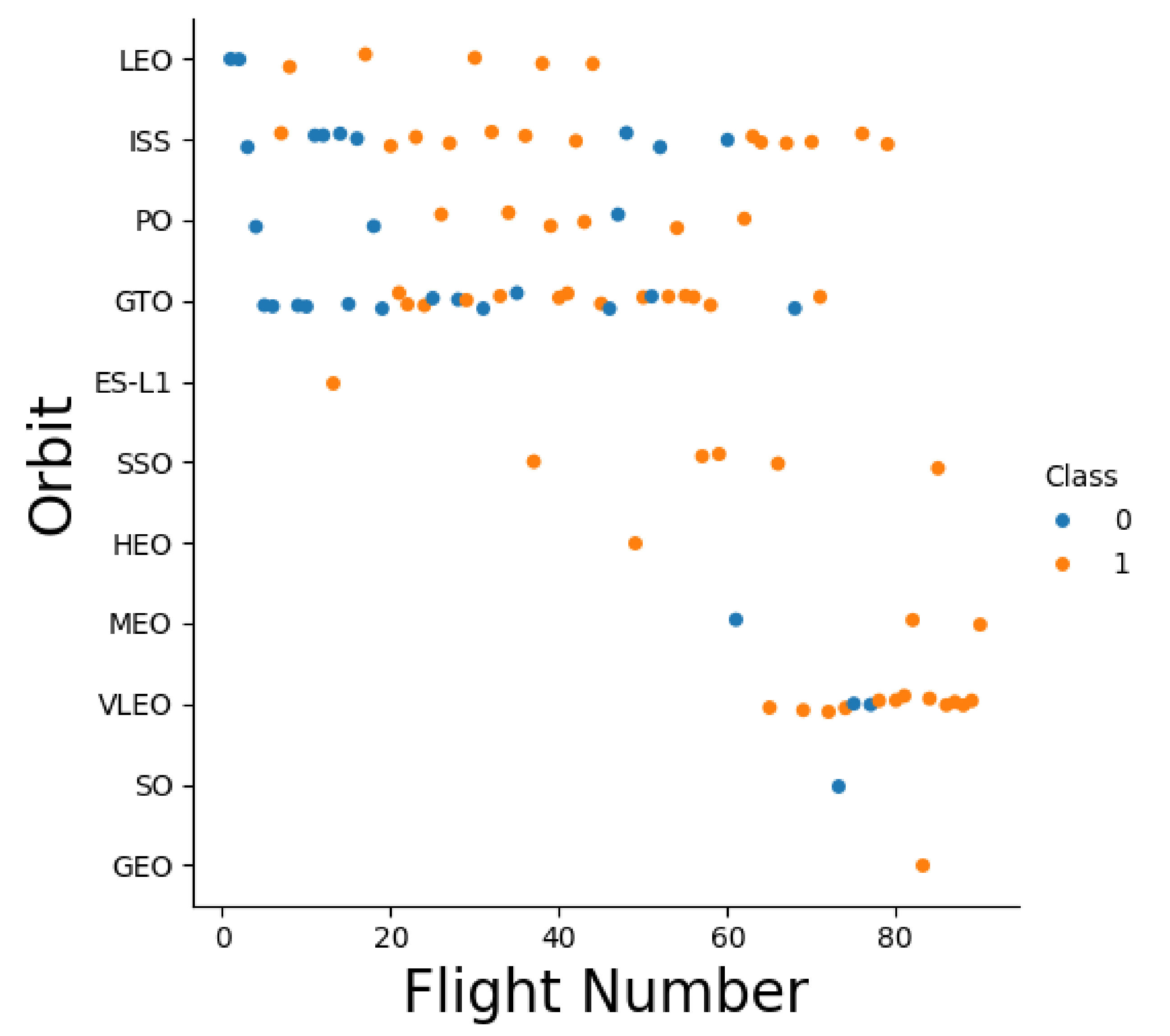
Success Rate vs. Orbit Type

Some orbit types seem to have a full success rate with Vleo orbits following those with an .80+ success rate the rest sit below that.

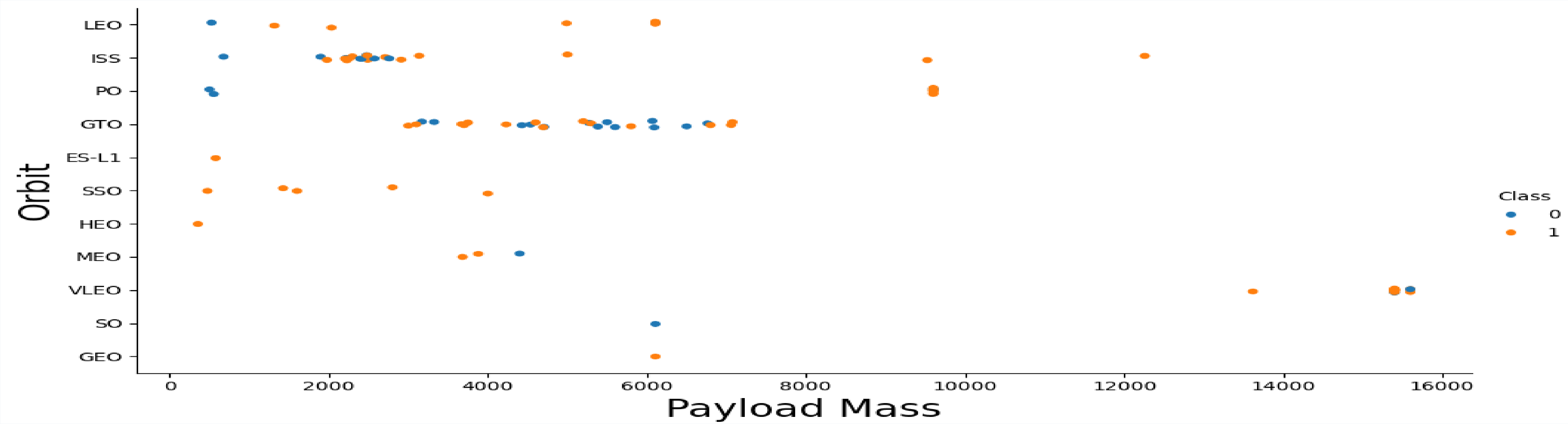


Flight Number vs. Orbit Type

The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

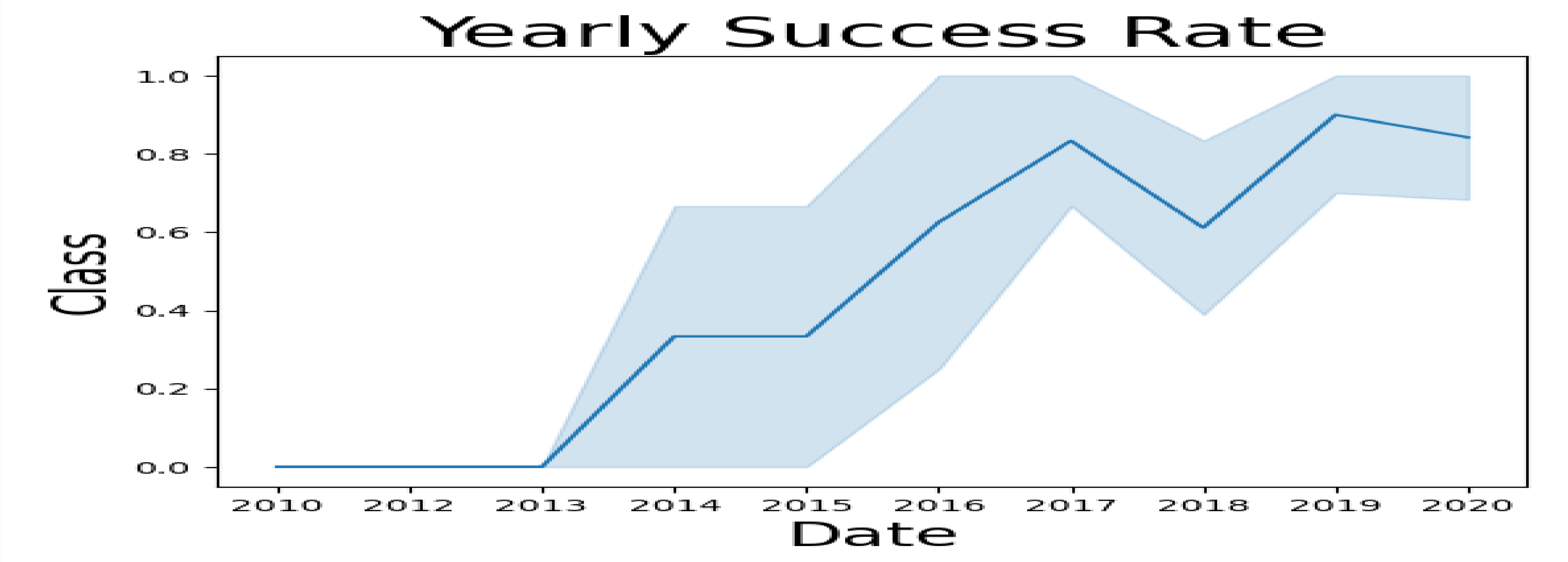


Payload vs. Orbit Type



- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend



- We can see that over time the yearly success rate increases until 2020 where the data ends.

All Launch Site Names

- We can find unique launch site names using the distinct keyword in sql

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- we can specify the number of records with the limit keyword

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parac
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 (parac
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- Getting a summation the payload lets us get the total mass, specifically with boosters for Nasa

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- here we find the average payload mass of the F9 v1.1 booster

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Finding the minimum date, and specifying a Success outcome lets us find the first successful Landing date.

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

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

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Using commands we get the boosters of successful drone ship landings where the payload mass is between 4000 and 6000 kg

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The number of missions

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

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

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- The booster versions that have carried the max payload mass in the data set.

```
%sql select distinct(BOOSTER_VERSION) from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

2015 Launch Records

- The Launch records for launches that happened in 2015

```
%sql SELECT substr(DATE,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [LANDING_OUTCOME] FROM SPACEXTBL where [LANDING_OUTCOME] = 'Failure (drone ship)' and substr(DATE,0,5)='2015';
```

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

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

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

- Ranking the landing outcomes from the above dates

```
%sql select LANDING_OUTCOME, COUNT(LANDING_OUTCOME) from SPACEXTBL where (DATE between '2010-06-04' and '2017-03-20') GROUP BY LANDING_OUTCOME ORDER BY COUNT(LANDING_OUTCOME) DESC
```

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

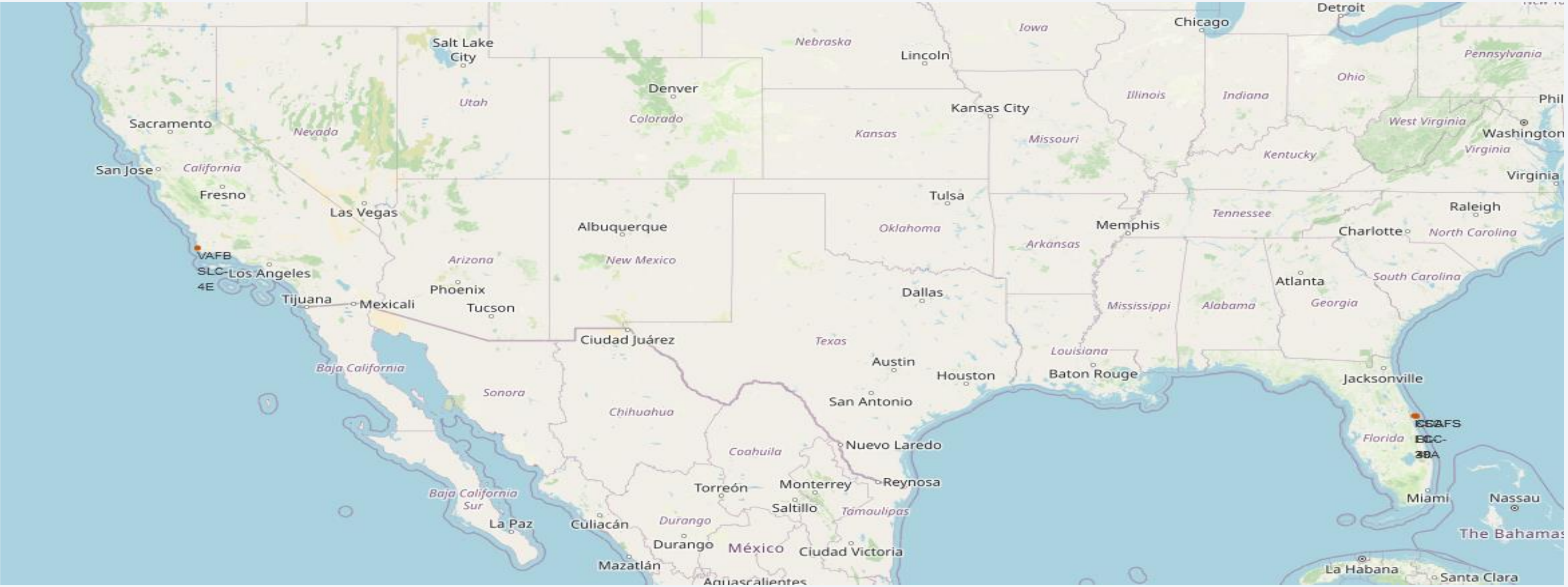
Landing_Outcome	COUNT(LANDING_OUTCOME)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Global Launch Site Locations

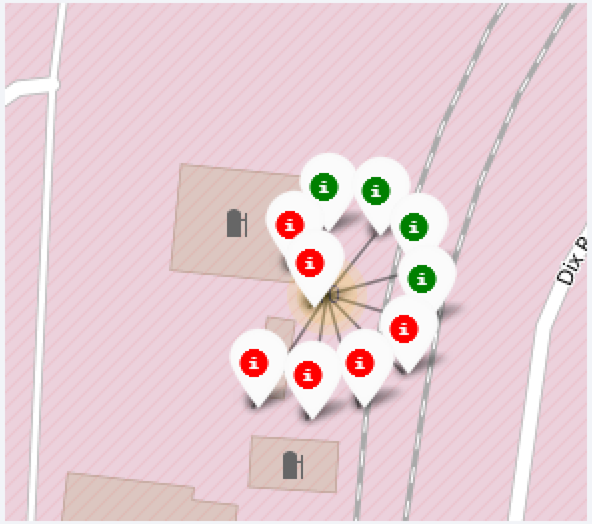
- This first map shows the two launch sites in California and Florida

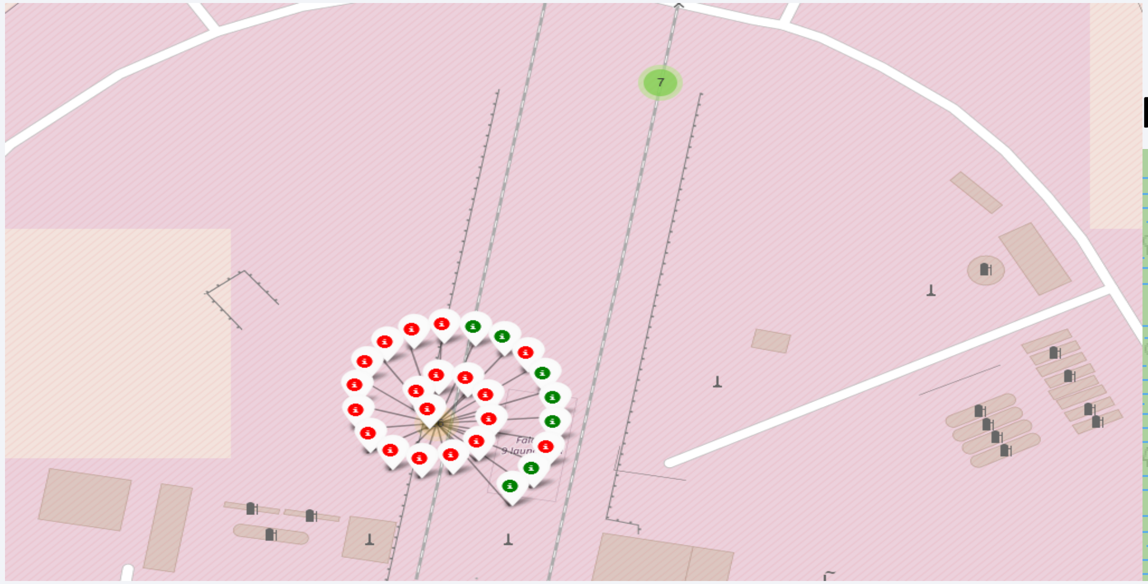


<Folium Map Screenshot 2>

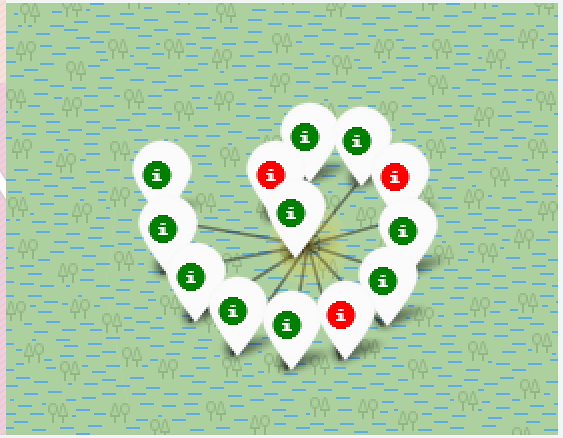
- Adding color to specific launch sites to denote green successes and red failures

California Launch site





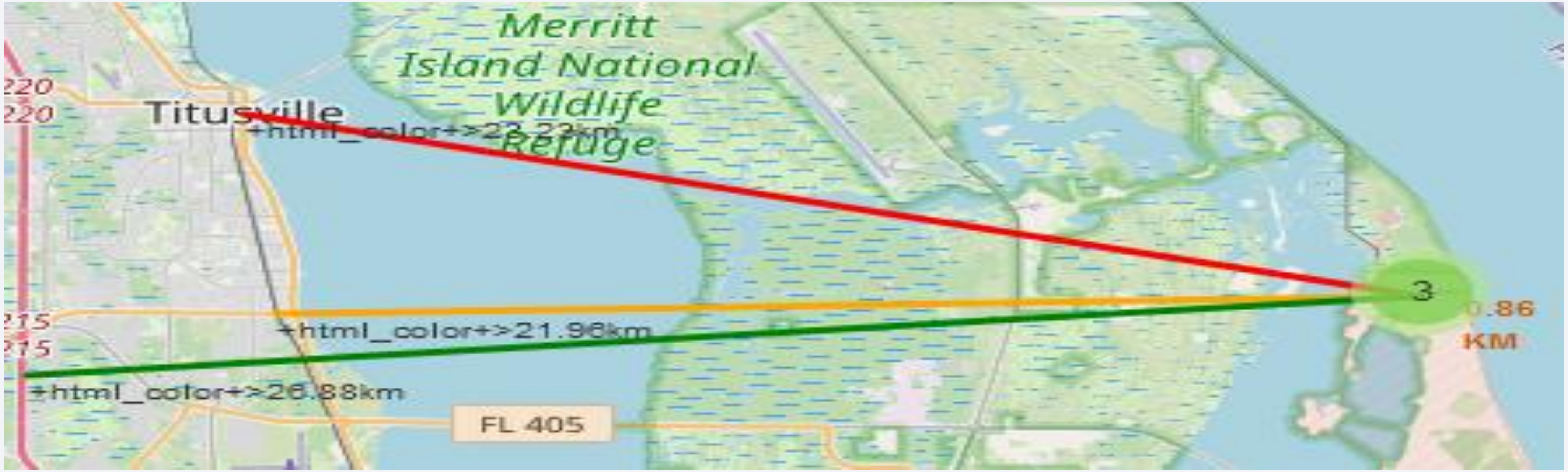
Florida Launch sites



Launch Site Distance

Looking at this we can see the distance to nearest objects

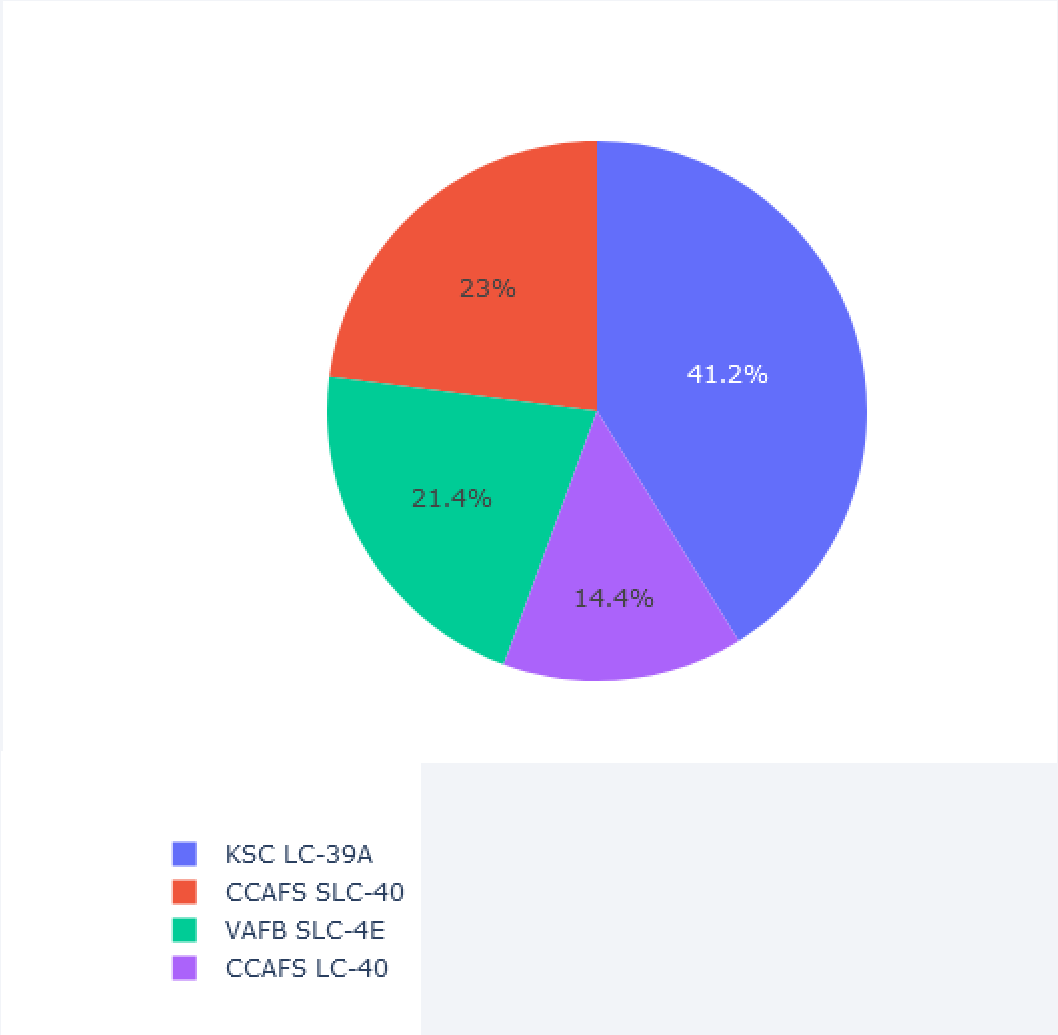
- We can see the launch site is not close to railways
- We can see the launch site is not close to any cities
- We can see the launch site is not near any highways
- The launch site is close to the coastline.



Section 4

Build a Dashboard with Plotly Dash

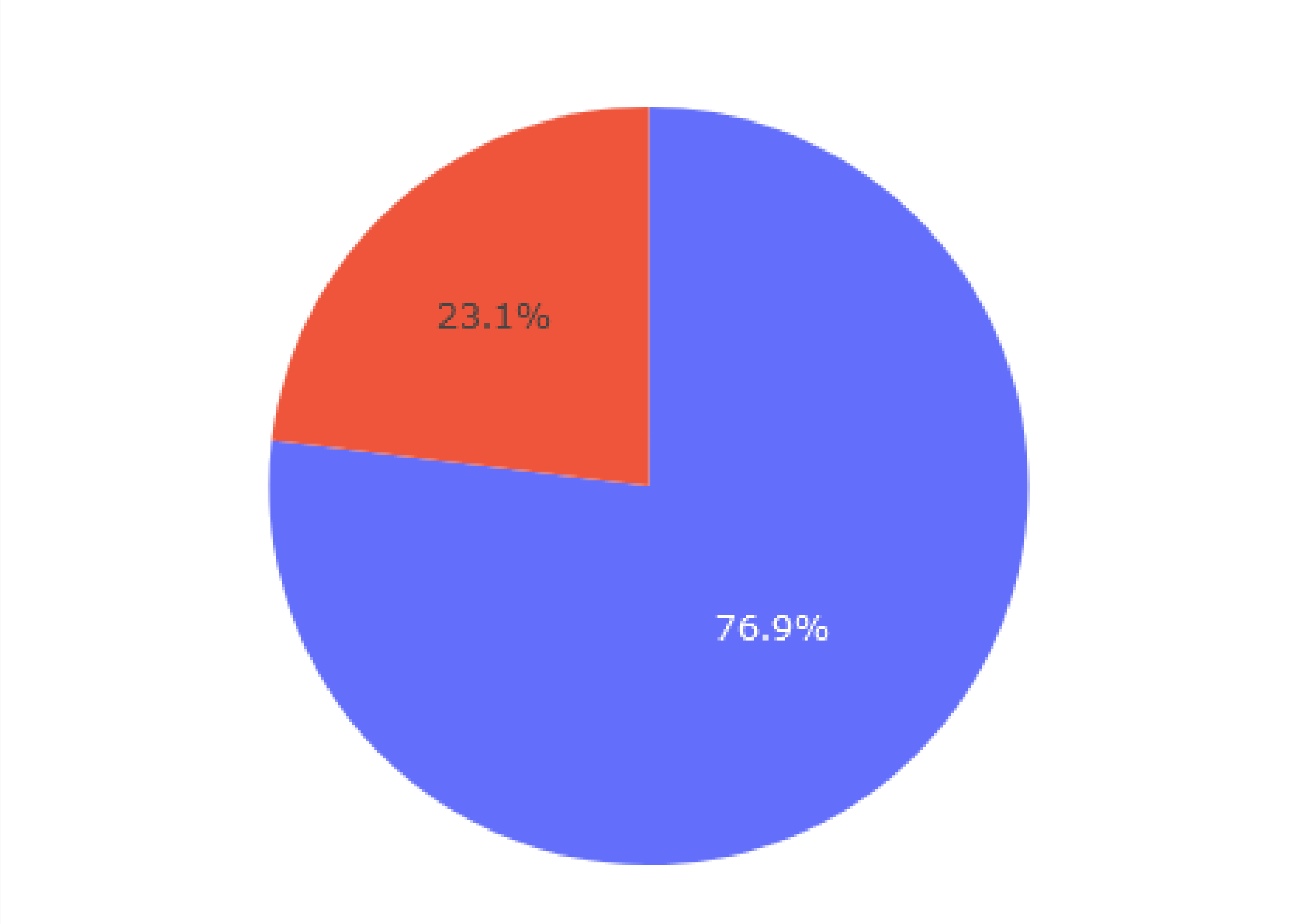
All Sites Launch Success Proportions



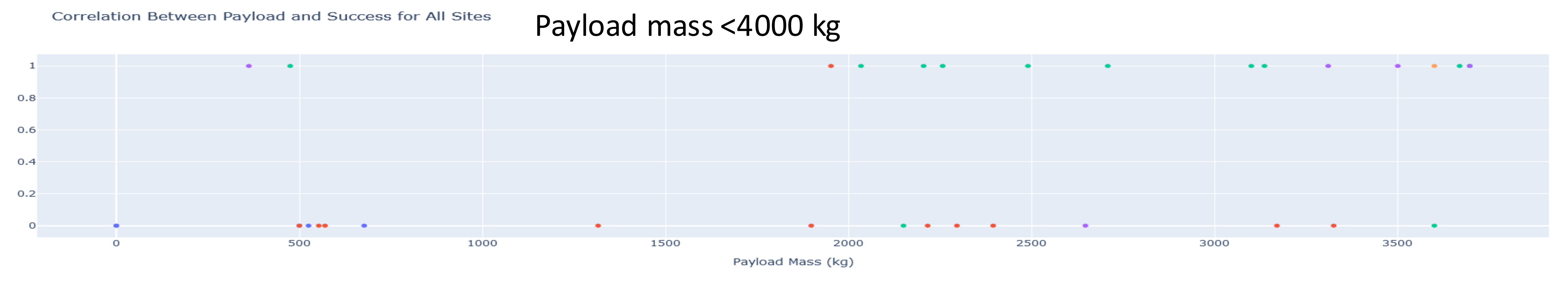
Most successful launches happened at the KSC LC 39a launch site.

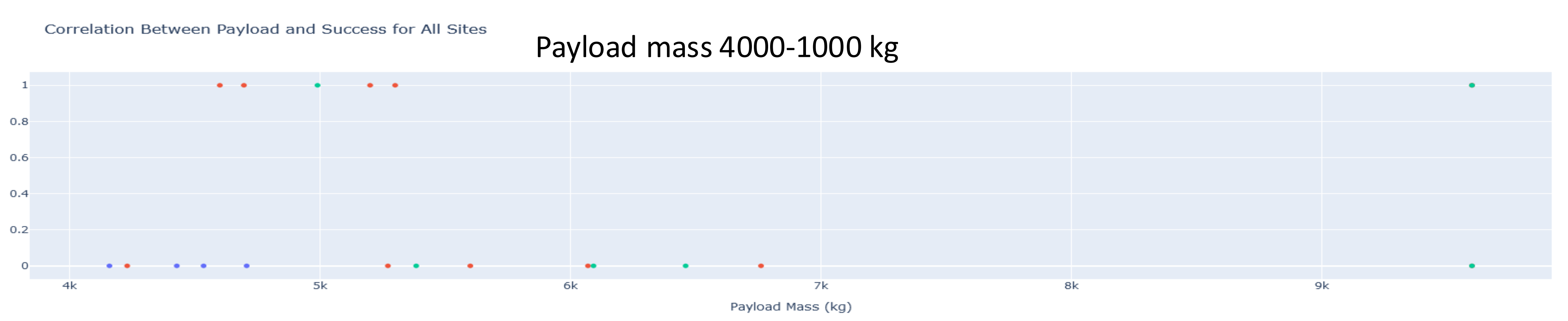
KSC LC-39A Launch Records

The Success Rate of this launch site is 76.9%
Leaving the remaining 23.1% as the rate of failure at this Site.



Payload slider and scatterplot





Here we can see the payload slider, we can see that lower weighted payloads had a higher success rate than higher rated payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

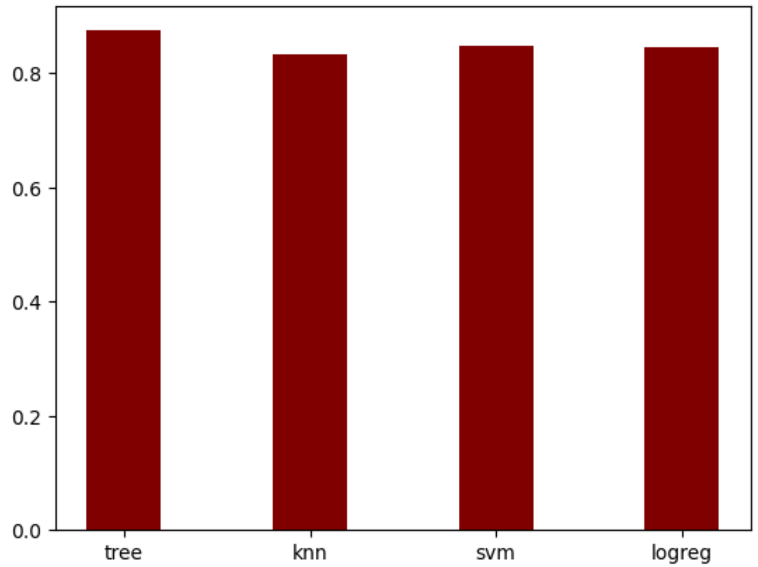
- Visualize the built model accuracy for all built classification models, in a bar chart

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

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

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

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

```
scores = {'tree':tree_cv.best_score_,'knn':knn_cv_score,'svm':svm_cv.best_score_,'logreg':logreg_cv.best_score_}
algorithm= list(scores.keys())
values = list(scores.values())
plt.bar(algorithm, values, color ='maroon',
        width = 0.4)
```

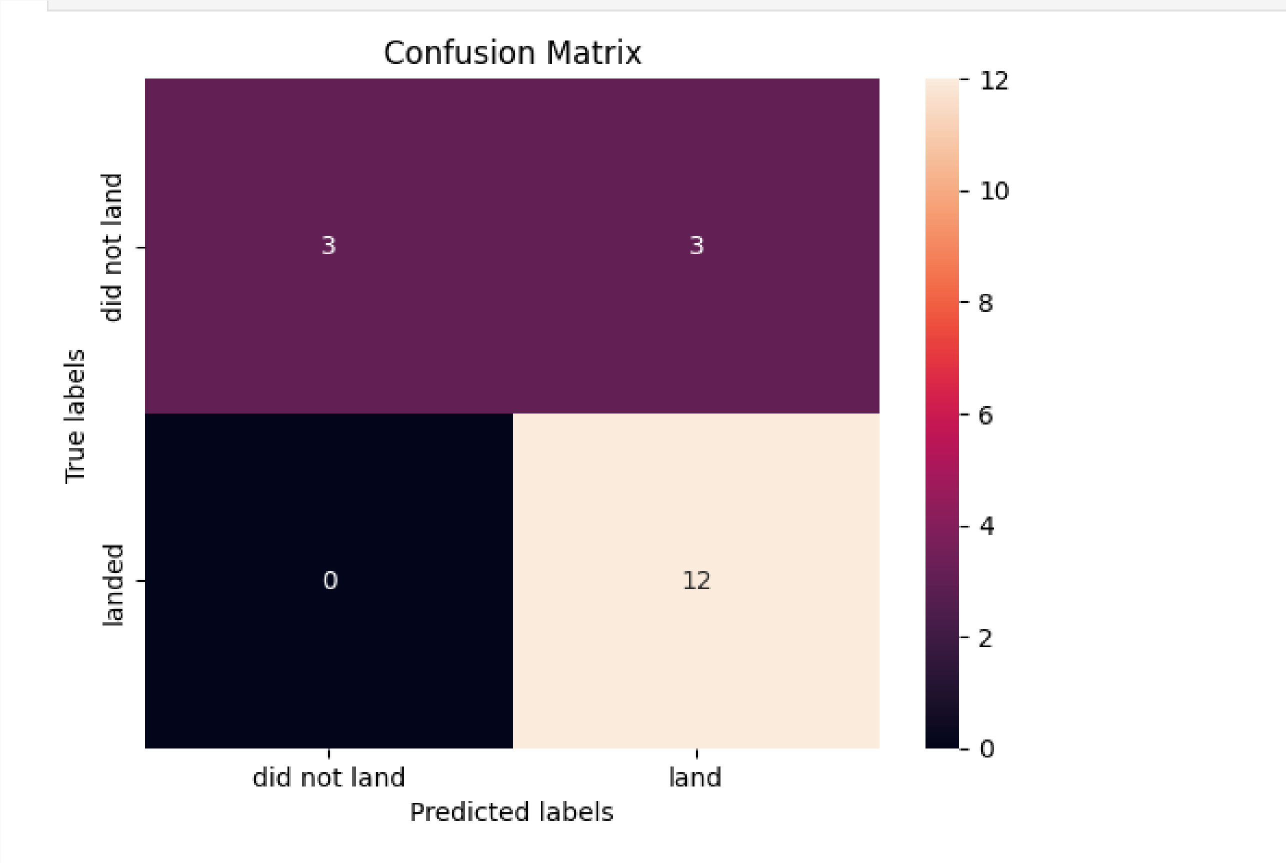
```
<BarContainer object of 4 artists>
```



While every method had the same accuracy on the test data, the tree data ended with the best score

Confusion Matrix for the Decision Trees Algorithm

- This is the confusion matrix for the Decision Trees Algorithm.



Conclusions

- Higher Flight numbers have a Higher chance of Success Outcomes
- There was an increase in success from 2013 to 2020
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier preformed machine learning the best in this task

Thank you!