CAPSTONE PROJECT

MELANIE GIN

9/17/2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Methodologies
  - Data Collection and Wraggling (API and Web Scraping)
  - Exploratory Data Analysis (EDA) SQL
  - EDA with Data Visualization
  - Interactive Visual Map Analysis with Folium and Dashboard
  - Machine Learning Predictive Analysis
- Results
  - Results from each of the methodologies

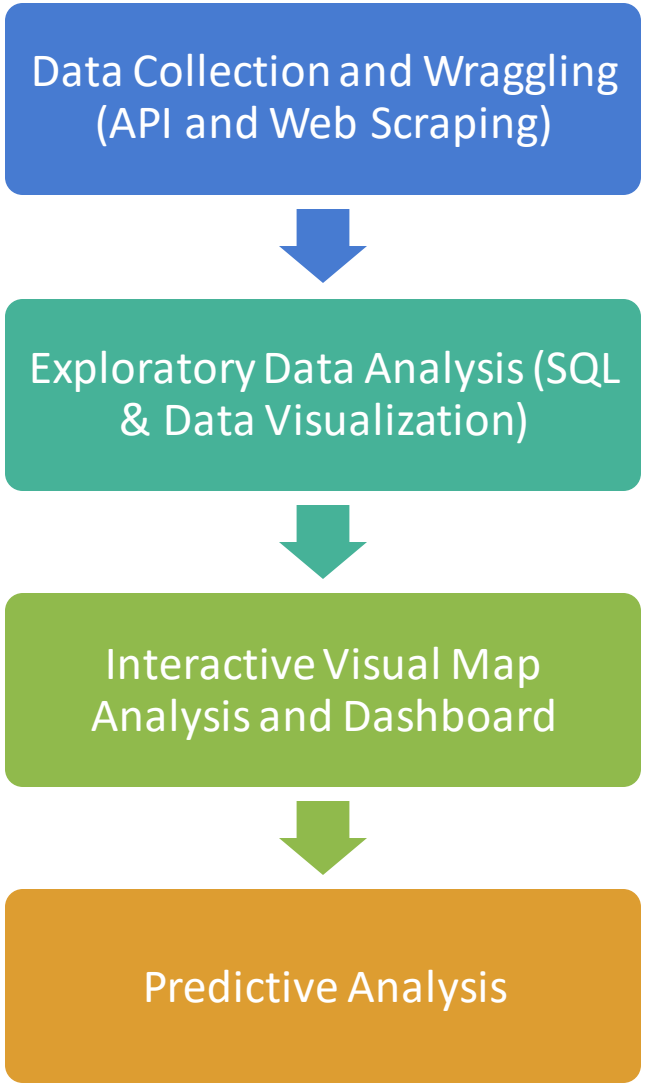
INTRODUCTION

Background: The data being explored is from the rocket company SpaceX, who has advertised that their Falcon 9 rocket launches only cost $65 million. This is very competitive from other rocket companies whose launches cost upwards of $162 million each. The incredibly savings is due to reusing rockets after successfully landing on the first stage. Therefore, the success of landing can determine the cost, which competitive companies can use to bid against SpaceX.

Problem:

- What are the factors for a failed or successful landing?

METHODOLOGY



METHODOLOGY: DATA COLLECTION

- Data Collection source:
  - Space X API: https://api.spacexdata.com/v4/rockets/
    - Request information from the API > convert to JSON file > create a dataframe with all the gathered data
  - Web Scrapping Falcon 9 historical launch records from a Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Converting the tables on the page to a dataframe

METHODOLOGY: DATA WRAGGLING

- From the wiki site, we scraped and parsed the tables of all the rockets to convert them into a pandas dataframe.
- From these outcomes, we converted them to binary factors, 1 means the booster successfully landed 0 means it was unsuccessful, which we assign to the column "Class".
- Then we found patterns with orbits and success rates

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

We can use the following line of code to determine the success rate:

```
df["Class"].mean()
```

```
0.6666666666666666
```

METHODOLOGIES SUMMARIZED

Exploratory Data Analysis (EDA) SQL

Learning about the data such as the sum and average of the payload mass

Success counts, success types, and booster versions

EDA with Data Visualization

Graphing features to find any patterns/relationships (flight number vs launch site, Payload vs Launch Site, success rate of each orbit type, FlightNumber and Orbit type, Payload and Orbit type, success yearly trend)

Interactive Visual Map Analysis with Folium and Dashboard

Labelled color markers depending on success or failure of launch at each locations (LA and Florida)

Built interactive dashboard to find results of success rate/counts faster and payload mass contributions.

Machine Learning Predictive Analysis

Found the accuracy of KNN, SVM, decision tree, and logistic regression with train-test sets

RESULTS

- Time was a key factor in success, as we saw the most recent launches had higher success rates (more successful outcomes) than earlier launches, especially after 2013.
- KFC site was more successful with its launches than the other four sites.
- Most successful launches and landings were from rockets in lower earth orbits.
- The accuracy in machine learning predictions for KNN, logistic regress, and support vector is 83%.

EXPLORATORY DATA ANALYSIS SQL

- Using SQL we found out the following:
  - Launch locations: CCAFS LC-40; VAFB SLC-4E; KSC LC-39A; CCAFS SLC-40
  - Boosters with successful drop ship landings: F9 FT B1022; F9 FT B1026; F9 FT B1021.2; F9 FT B1031.2
  - We learned a lot about the dats but the most important was the successes:
    - Mission outcome counts
    - Number of successful landings count based on outcome in descening order between 04-06-2010 and 20-03-2017.

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Outcome FROM SPACEXTBL GROUP BY Mission_Outcome
```

Mission_Outcome	Outcome
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Landing _Outcome	Success_Count
Success (ground pad)	6
Success (drone ship)	8
Success	20

```
%sql SELECT "Landing _Outcome", COUNT(*) as Success_Count FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE '%Success%' and Date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY Date DESC;
```

EXPLORATORY DATA ANALYSIS SQL (CONTINUED)

- Other results that might be of interest

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

```
%sql SELECT 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

```
%sql SELECT sum(PAYLOAD_MASS__KG_) as payload_sum from SPACEXTBL WHERE customer == 'NASA (CRS)'
```

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

payload_sum
45596

```
%sql SELECT avg(PAYLOAD_MASS__KG_) as payload_avg from SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%'
```

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

payload_avg
2534.6666666666665

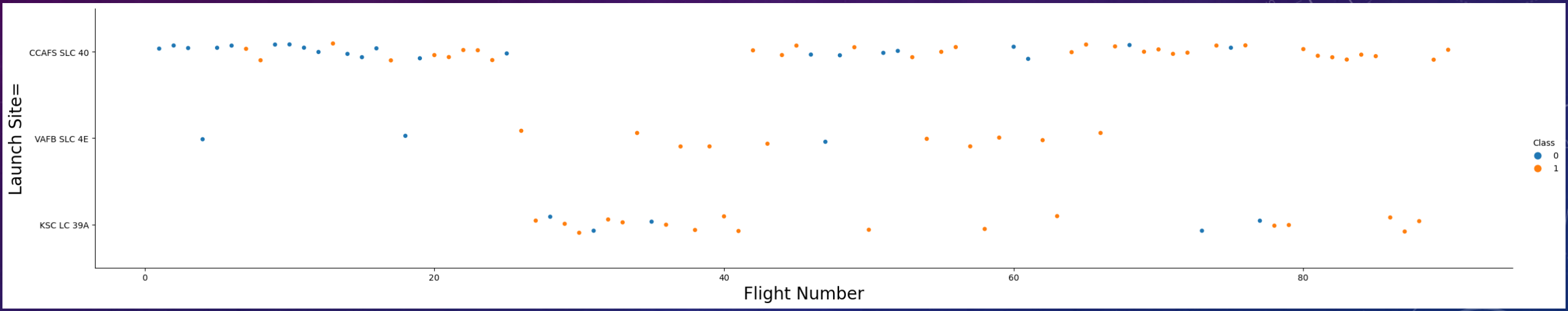
```
%sql select min(Date) from SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%';
```

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

min(Date)
01-05-2017

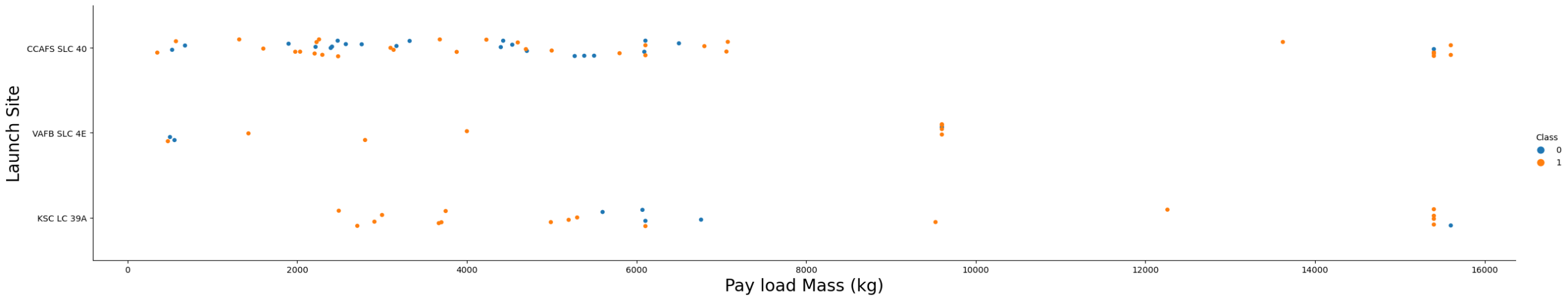
EDA WITH DATA VISUALIZATION (1/6)

- Graphs make it easy to see the relationship factors that affects the outcome. Here are the features we focused on (take a look at the x and y-axes)



We see that there have been more launches with CCAFS SLC 40, but it looks like from the pattern it was mostly because they were trying to find more of a successful outcome as you can see from earlier it's a tie with failures, while the others have more sequential successes and fewer launches.

EDA WITH DATA VISUALIZATION (2/6)



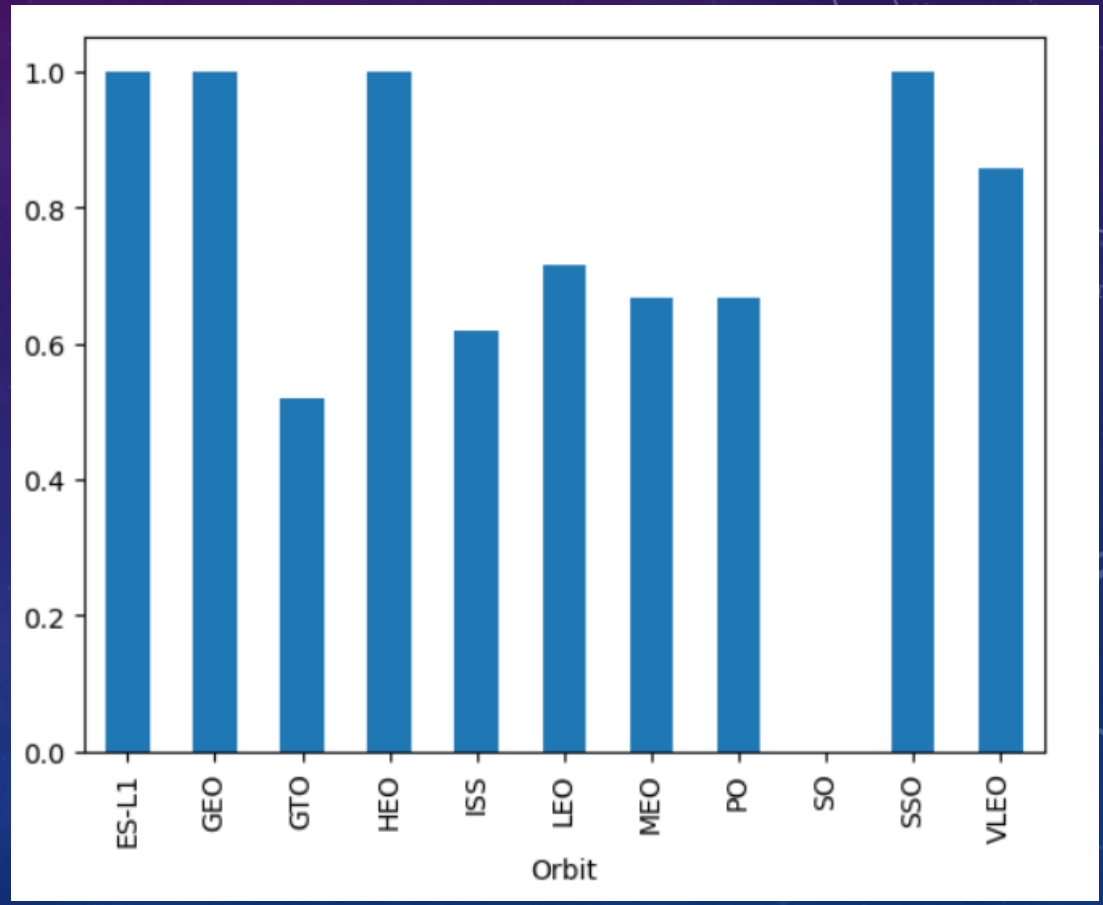
Note that VAF SLC 4E did not test lanuches with payloads masses more than 10000 kg

Again more launches were at the CCAF SLC 40 site and they tested more higher payload masses than KSC LC 39A.

And even though there were more launches at CCAF SLC 40, there were more of a success rate for KSC LC 39A.

EDA WITH DATA VISUALIZATION (3/6)

- The y axis is the mean of success rates
- High success rates in orbits: ES-L1, GEO, HEO, and SSO
- **SSO (or SO)**: It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time
- **ES-L1** :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth
- **HEO** A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth
- **HEO** Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi)
- **GEO** It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation



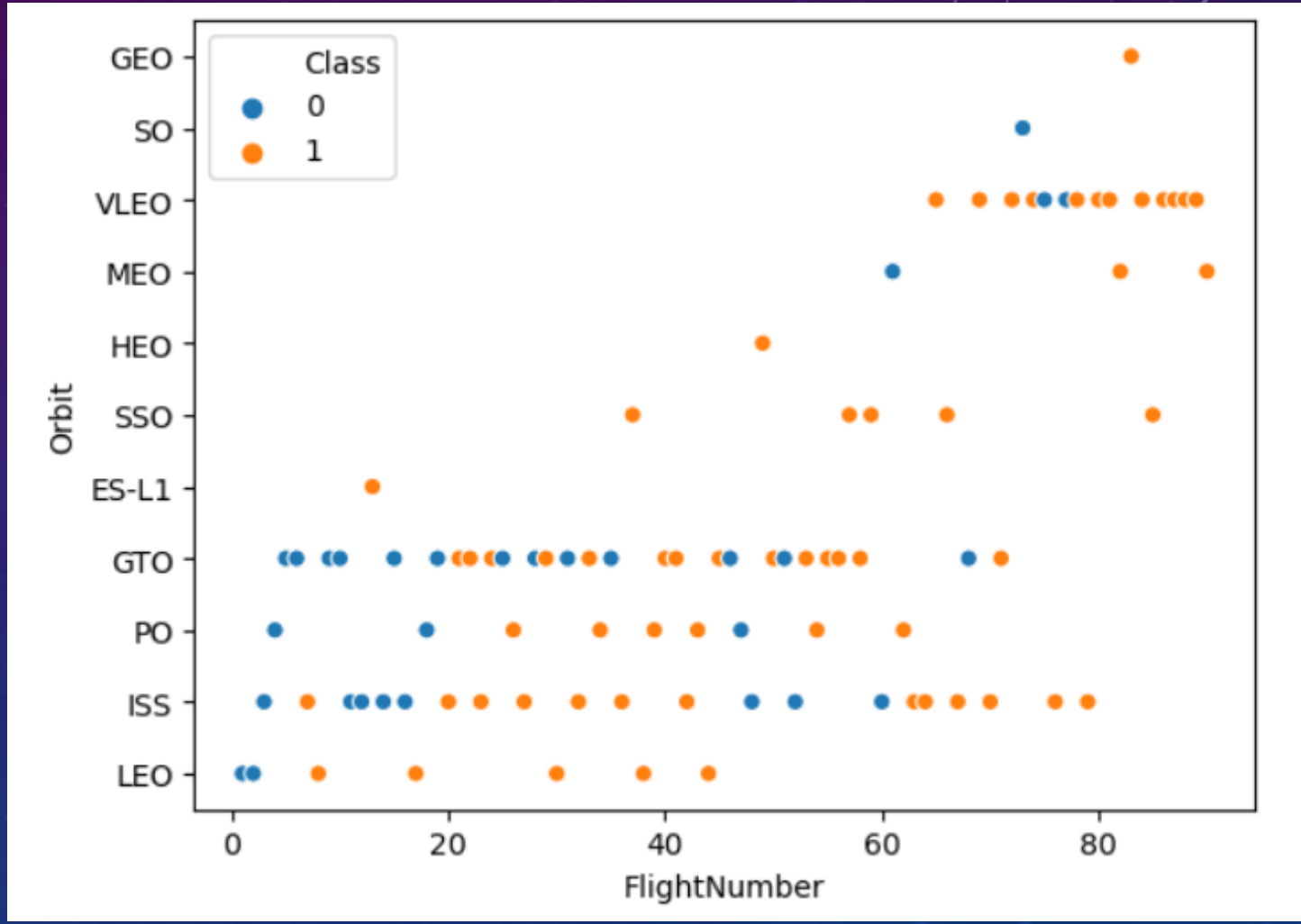
EDA WITH DATA VISUALIZATION (4/6)

Showing another graph focusing on the orbits, we see more detail about their high success rates.

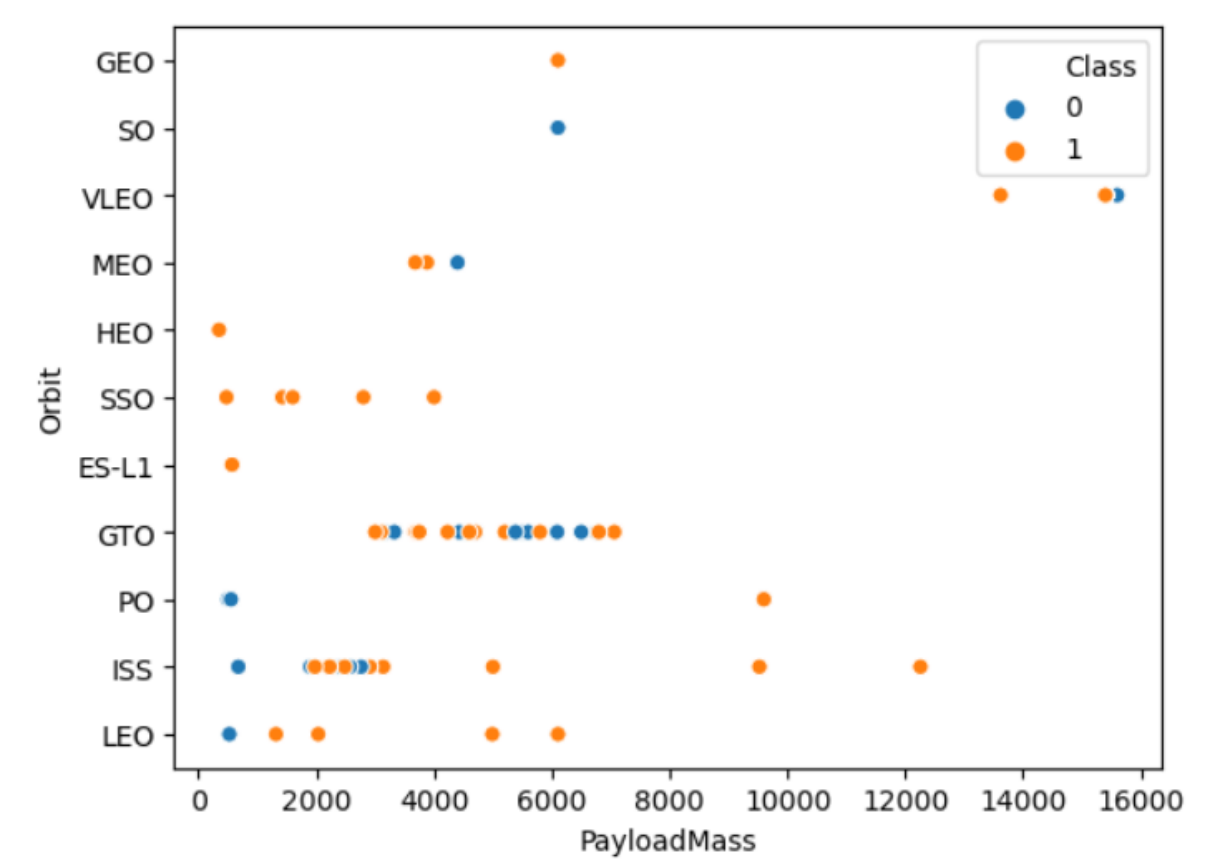
GEO looks to only have one rocket that was successful

No relationship between flight number when in GTO orbit

Overall, there are more success rates in later/recent launches than in earlier ones (makes sense)

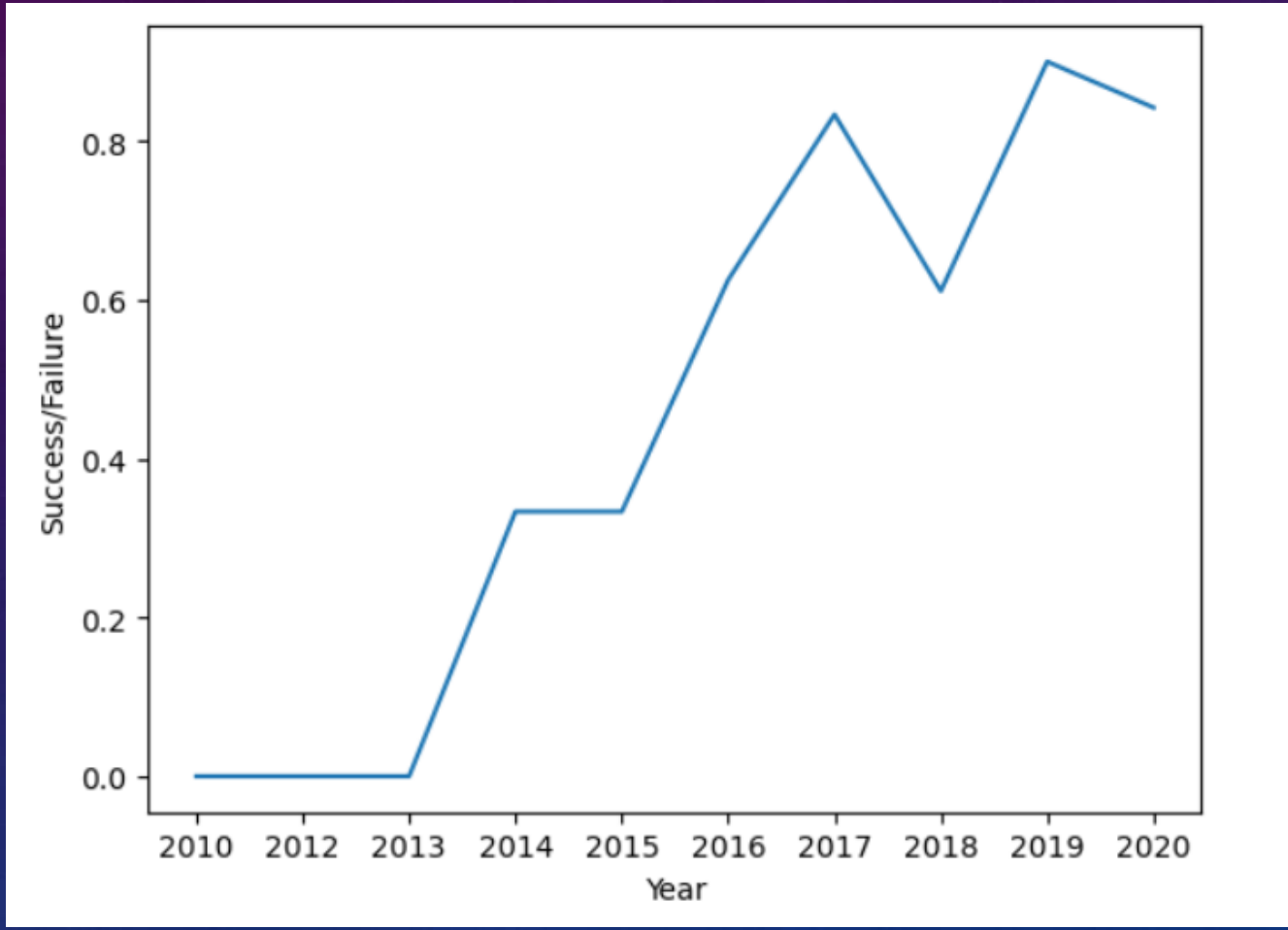


EDA WITH DATA VISUALIZATION (5/6)

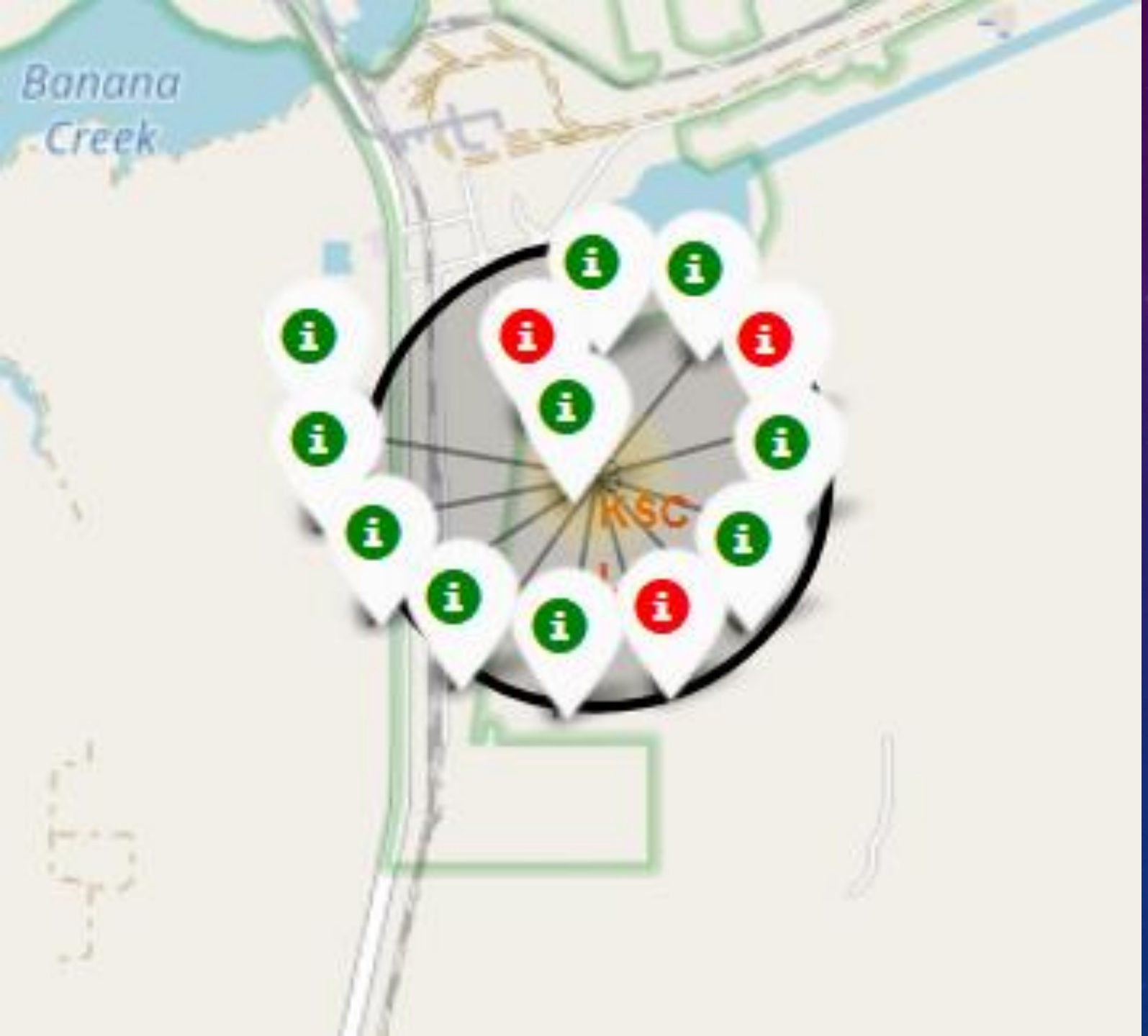


- Higher payload masses were towards ISS, PO, and VLEO
- **VLEO**: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation
- **ISS** A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada)
- **PO** It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth

EDA WITH DATA VISUALIZATION (6/6)



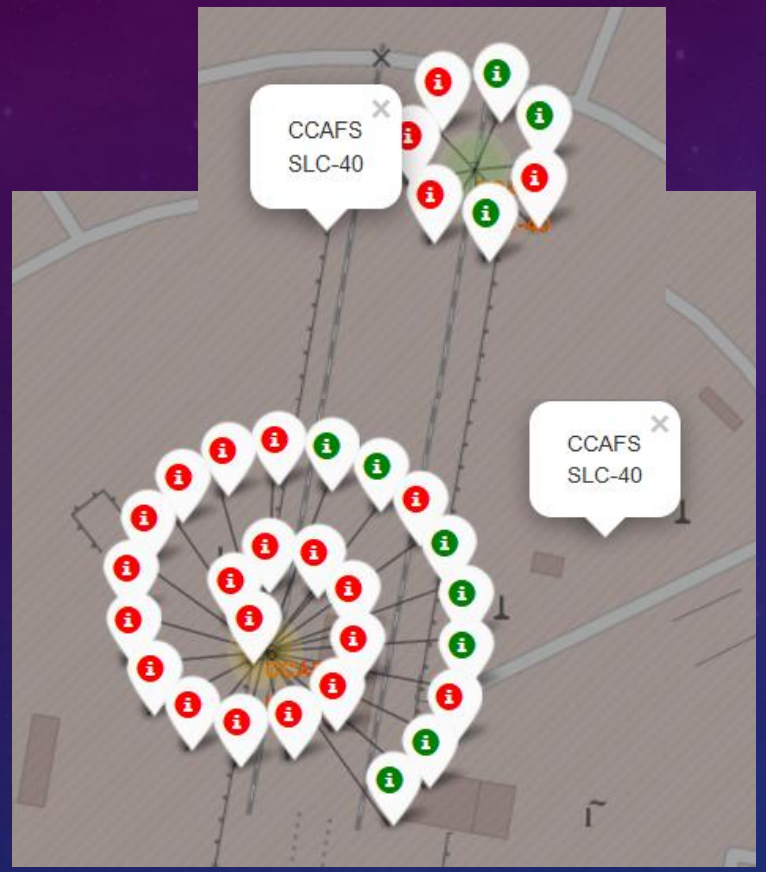
Overall, the success rate has increased since 2013

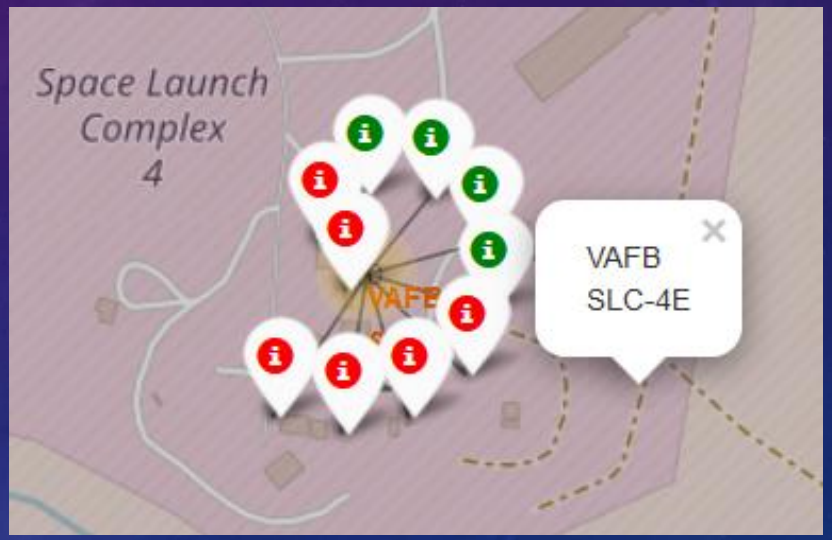


INTERACTIVE MAP WITH FOLIUM

- Out of all the sites, the site with the most successes was KSC LC-39A.
- I believe it's due to it being futher away from coastlines unlike the other two.

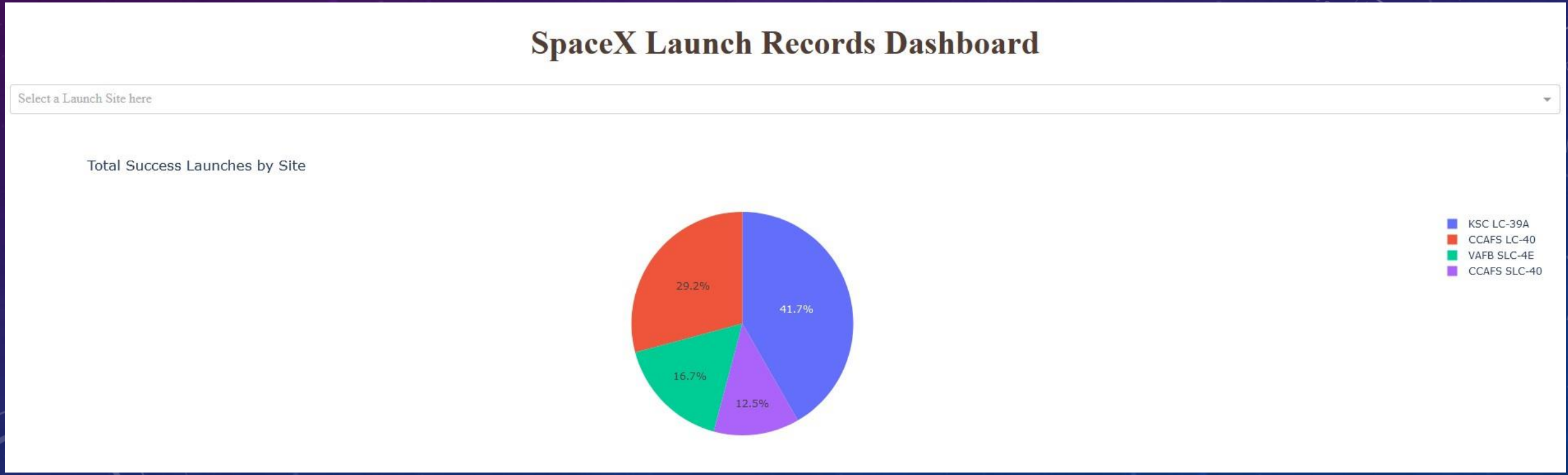
FOLIUM SCREENSHOTS OF OTHER MARKS



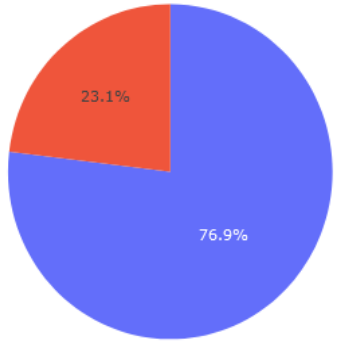


INTERACTIVE DASHBOARD

- A dashboard allows for a quicker and easier way to find answers to successes and failures. Which we see KSC LC-39A has a 41% success rate while the others are 20 or less.
- https://mgin14-8050.theiadocker-3-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/

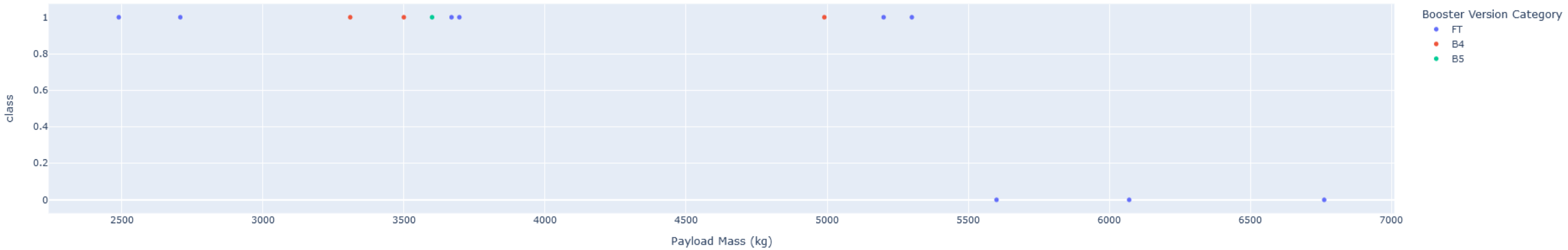


Total Success Launches for site KSC LC-39A



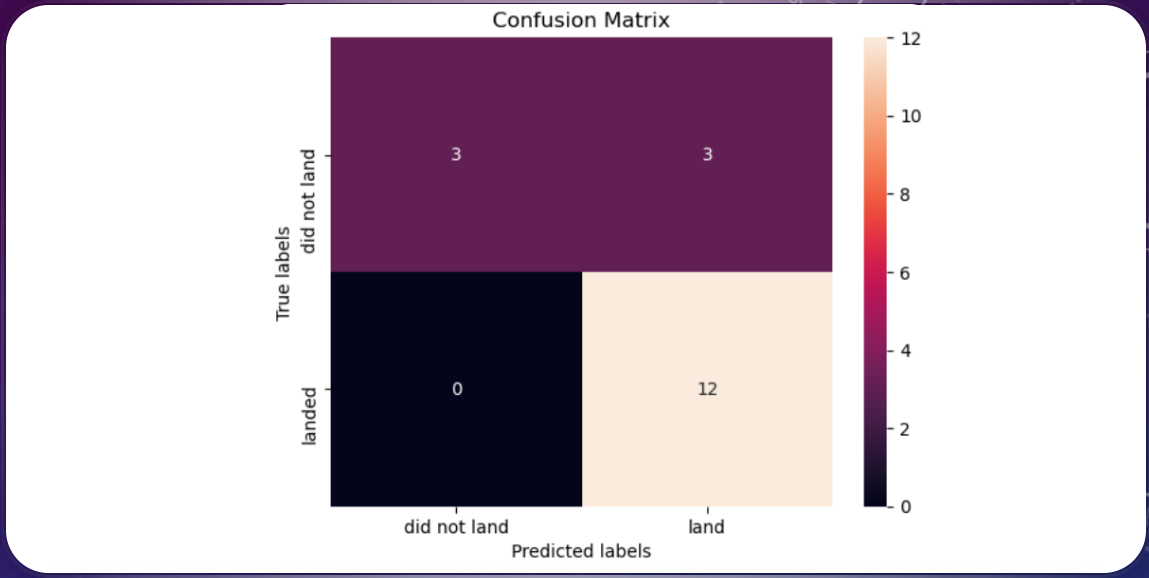
Payload range (Kg):

Correlation between Payload and Success for KSC LC-39A



PREDICTIVE ANALYSIS

- With machine learning, we did a train-test set for K nearest neighbors (KNN), descision tree, logistic regression, and support vector Machine.
- Confusion matrices helped to show true positives and negatives. All showed the same.
- We found that all but the decision tree worked the same and well with an 83% accuracy on the test sets. However, for best_score_ on the training set, decision tree did better with **% while the others had 84%.




```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 5,
'splitter': 'random'}
accuracy : 0.8888888888888888
```

CONCLUSION

- There weren't enough data on heavier payload masses to conclude which performed better. CCAF location had slightly more but not enough to overrule the other two. There were significantly more low weight payloads which we see KSC performed better.
- When it comes to predicting the success of a launch, the KNN, logistic regression, or support vector model should be used since they have a 83% accuracy outcome.
- What was most clear about the contributing factor to successful launches was time. Though the KSC site seems to have more successful launches overall. CCAF had more but it looked to be from trying to improve performance and increase success rate.
- Launches in lower earth orbits were more successful as well with flight numbers and payload masses: LEO and VLEO.