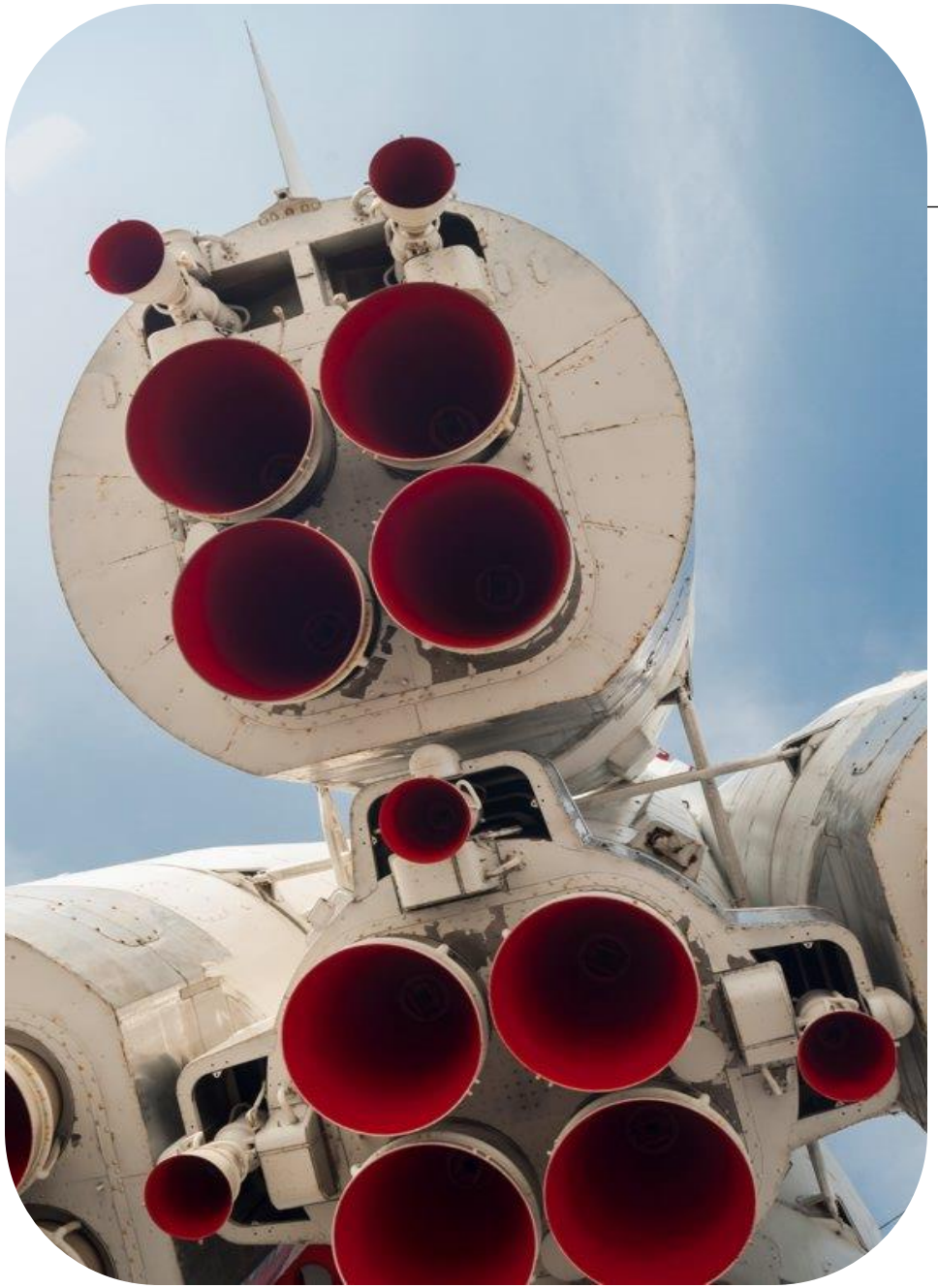


SpaceX Falcon 9 First Stage Landing Prediction

Wyatt Webster

November 29, 2023

OUTLINE

- **Executive Summary**
- **Introduction**
- **Methodology**
  - Data Collection & Wrangling
  - EDA
  - Predictive Analysis
- **Results**
  - Visualizations
  - Dashboard
  - Predictive Analysis
- **Conclusion**

EXECUTIVE SUMMARY

Methodology Summary

- **Data Collection:**
  - Using SpaceX REST API and web scraping.
- **Exploratory Data Analysis (EDA):**
  - Created training labels and explored data with SQL and visualizations.
- **Interactive Visualizations:**
  - Folium and Plotly Dash.
- **Predictive Analysis:**
  - Machine learning models utilized for predicting a successful landing.

Results Summary

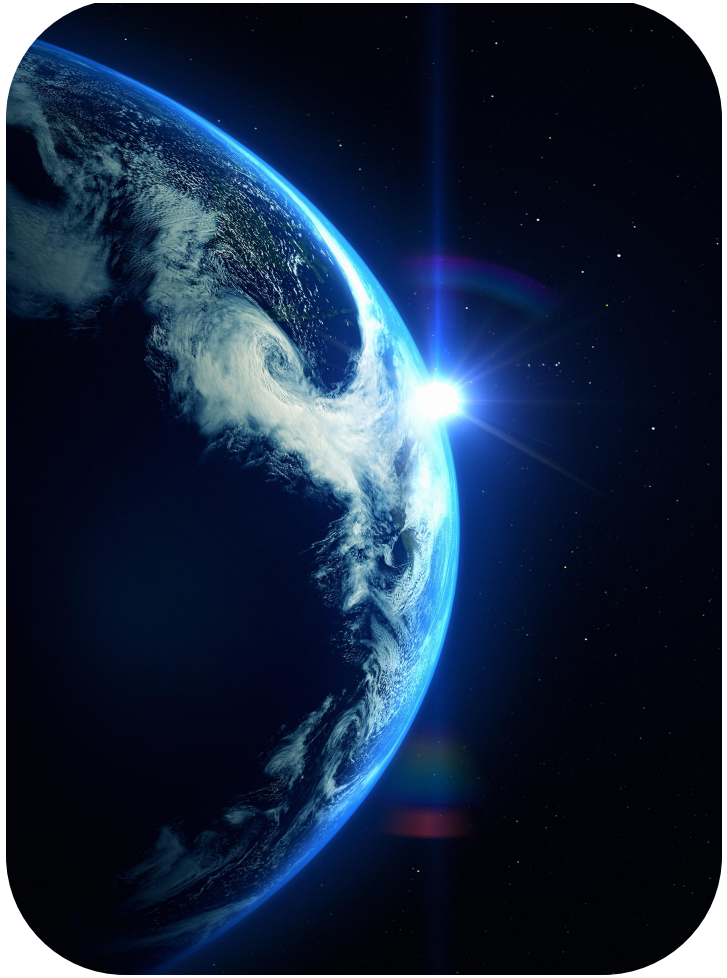
- **EDA:**
  - Launch success improved over time and correlated to payload mass.
  - Launches into specific orbits had much higher success rates than others.
  - Launch sites are located near coastlines.
  - KSC LC-39A had the highest overall launch success rate.
- **Predictive Analysis:**
  - All models performed roughly equal when compared used the accuracy metric.

INTRODUCTION

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The goal is to leverage machine learning models to predict whether the first stage of the rocket can be reused.

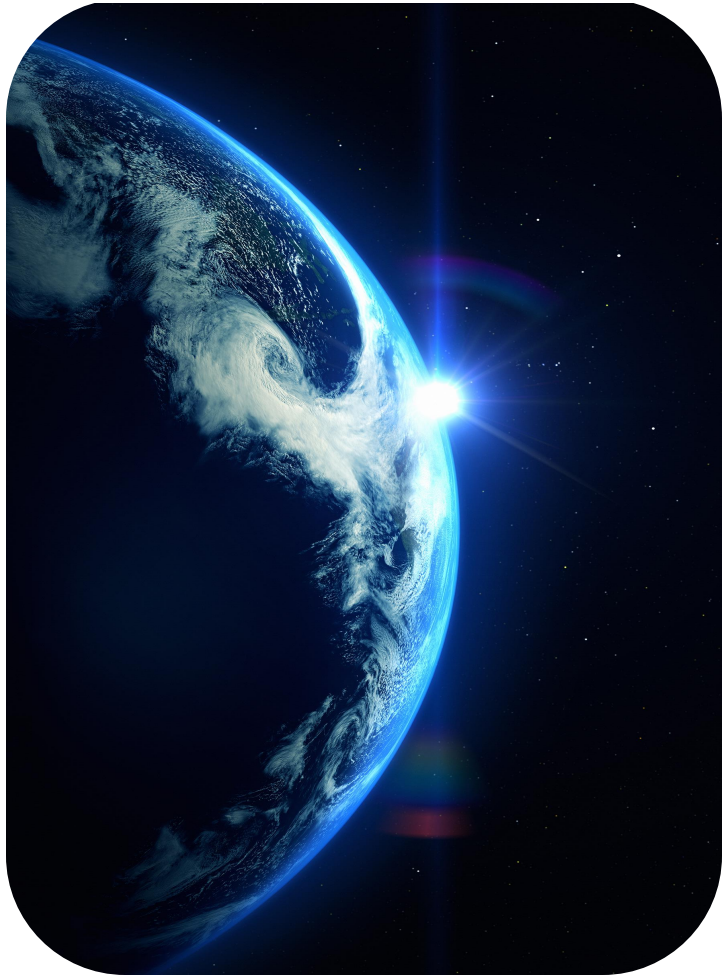
METHODOLOGY

METHODOLOGY – DATA COLLECTION



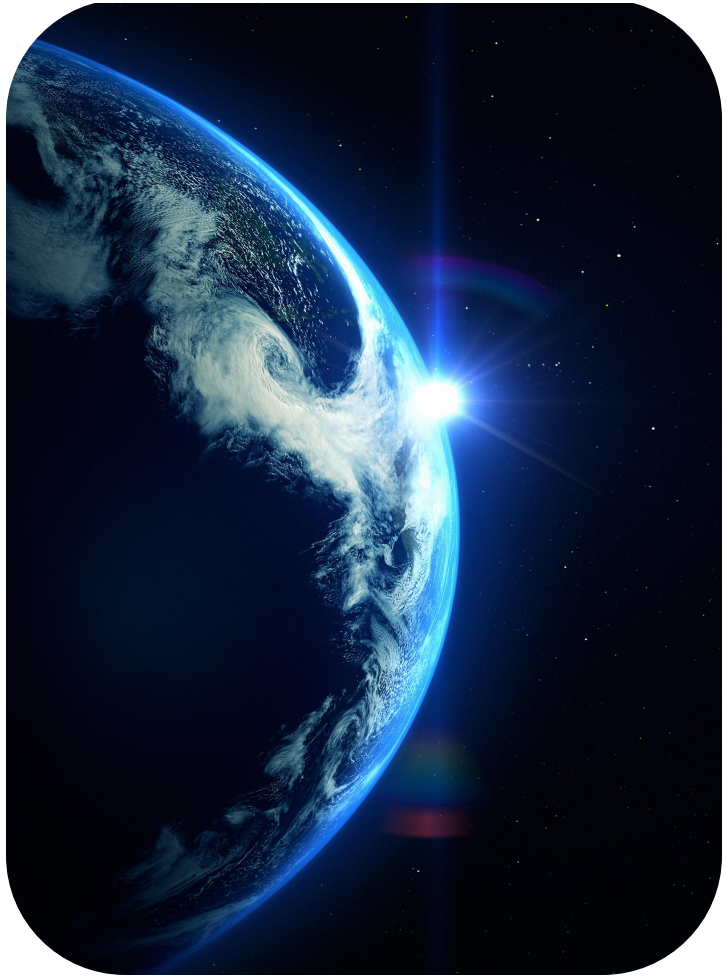
- **SpaceX RESTful API:**
  - Data collected using the SpaceX RESTful API.
  - Request and parse the data using the the GET request.
  - Convert collected data into a pandas DataFrame.
  - Perform initial data review, collect relevant features.
  - Filter the data to only include Falcon 9 launches.
  - Deal with missing values.
- **Web Scraping:**
  - Data scraped from launch records on Wikipedia using BeautifulSoup.
  - Use GET method to request Falcon 9 Launch page.
  - Extract all column names from HTML table.
  - Create a DataFrame by parsing the HTML launch table.

METHODOLOGY – DATA WRANGLING



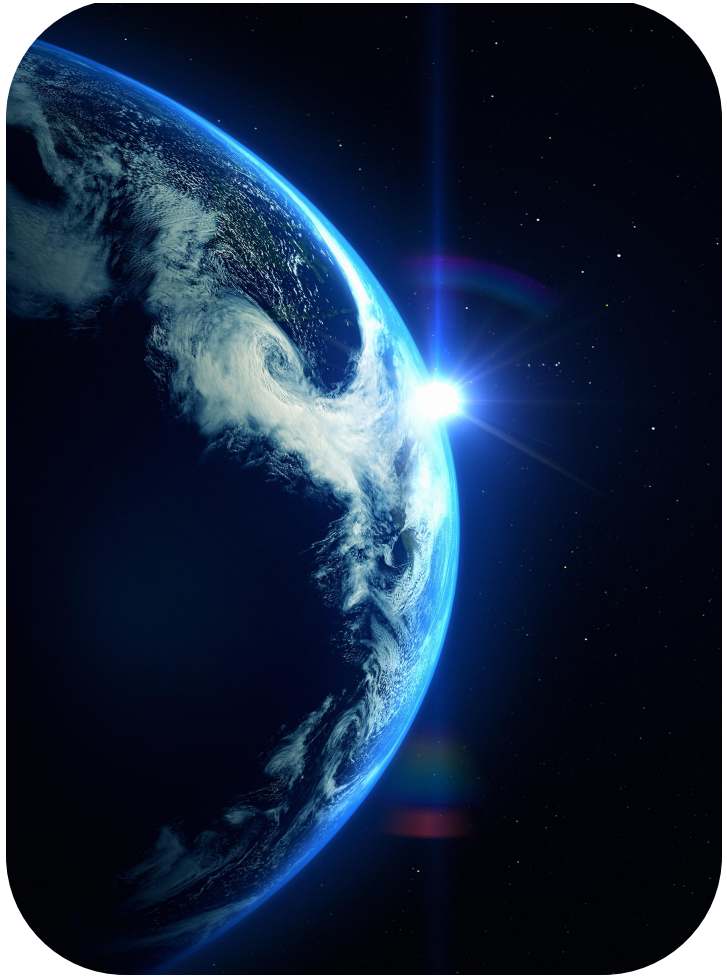
- **Perform Exploratory Data Analysis (EDA) and determine the training labels:**
  - Explore the number of launches from each site.
  - Determine the number and occurrence of each orbit.
  - Use the Outcome column to generate the training labels column (Class).
    - **Outcome**:
      - True Ocean -> 1
      - False Ocean -> 0
      - True RTLS -> 1
      - False RTLS -> 0
      - True ASDS -> 1
      - False ASDS -> 0
    - **Class**:
      - 0 -> The first stage did not land successfully.
      - 1 -> The first stage landed successfully.

METHODOLOGY – EDA WITH SQL



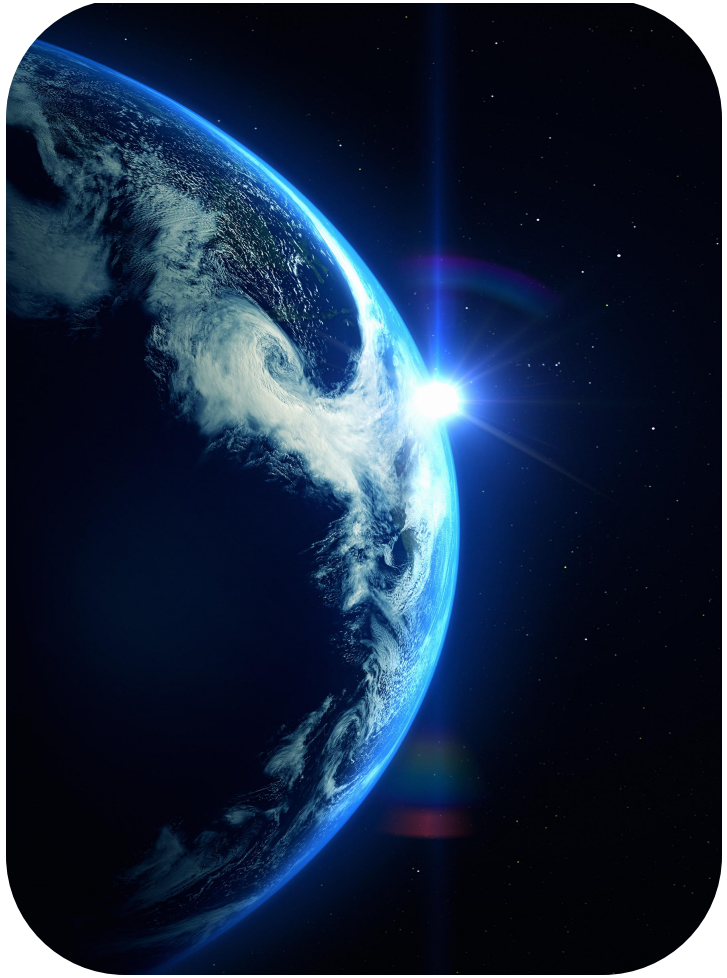
- Create and execute SQL statements to better understand the data.
- **Queries**:
  - Determine unique launch site names.
  - Explore launches from the same sites.
  - Average payload carried by F9 boosters.
  - Earliest successful landing outcome.
  - Explore which boosters have carried the largest payloads.
  - Review landing outcomes for various date ranges.

METHODOLOGY – EDA WITH VISUALIZATION



- **Visualizations using Matplotlib & Seaborn:**
  - Launch Site vs Flight Number.
  - Launch Site vs Payload Mass.
  - Bar plot to visualize success rate of launches into various orbits.
  - Orbit vs Flight Number.
  - Orbit vs Payload Mass.
  - Launch success rate yearly trend.
  - Select features to be used in the prediction of successful launches.
  - Convert categorical columns to dummy variables.
- Map of launch sites with Folium.
- Interactive dashboard with Plotly Dash.

METHODOLOGY – PREDICTIVE ANALYSIS



- Separate X (feature values) and Y (target values).
- Standardize the X data.
- Split into train and test sets.
  - 80% of samples for training, 20% for testing.
- **Machine learning models tested:**
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbours
- Use GridSearchCV for hyperparameter tuning.
- Rank the ML models based on accuracy score.

RESULTS

RESULTS – EDA WITH SQL

- Unique 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

- Earliest successful ground pad landing

```
In [19]: %sql select min(Date) as Date from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Date
2015-12-22

- Total # of successful and failed missions

```
In [20]: %sql select distinct(Mission_Outcome), count() as Count from SPACEXTABLE group by Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

Out[20]:

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

RESULTS – EDA WITH SQL

- Information from 5 launches at Cape Canaveral launch site

Display 5 records where launch sites begin with the string '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

RESULTS – EDA WITH SQL

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

```
In [17]: %sql select customer, sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like '%CRS%'
 * sqlite:///my_data1.db
Done.
```

Out[17]:

Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	48213

- Average payload mass for the F9 v1.1 booster version

```
In [18]: %sql select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Out[18]:

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

RESULTS – EDA WITH SQL

- Boosters with success in drone ship and payload between 4000-6000kg

In [22]:
```
%sql select Booster_Version, PAYLOAD_MASS__KG_ \
from SPACEXTABLE \
where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[22]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

RESULTS – EDA WITH SQL

- Booster versions that have carried the max payload mass

In [17]:
```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Out[17]:

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

RESULTS – EDA WITH SQL

- Launches in 2015

```
%sql select substr(Date, 6, 2) as Month, Date, Booster_Version, Launch_Site, Landing_Outcome \
FROM SPACEXTABLE 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)

- Landing outcomes between 2010-06-04 and 2017-03-20

In [19]:
```
%sql select Landing_Outcome, count(*) as count FROM SPACEXTABLE \
where Date between '2010-06-04' and '2017-03-20' \
group by Landing_Outcome order by count DESC
```

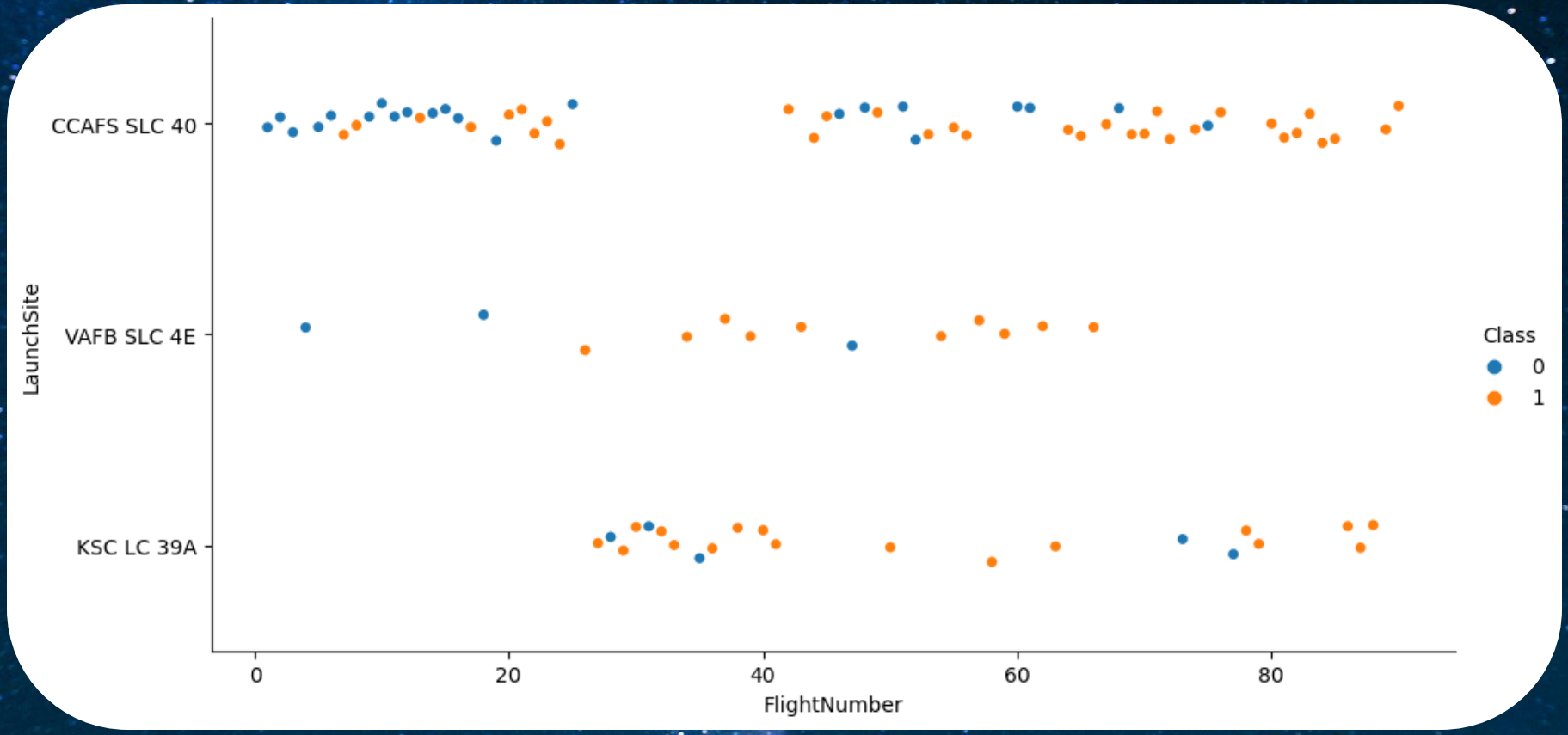
```
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Landing_Outcome	count
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

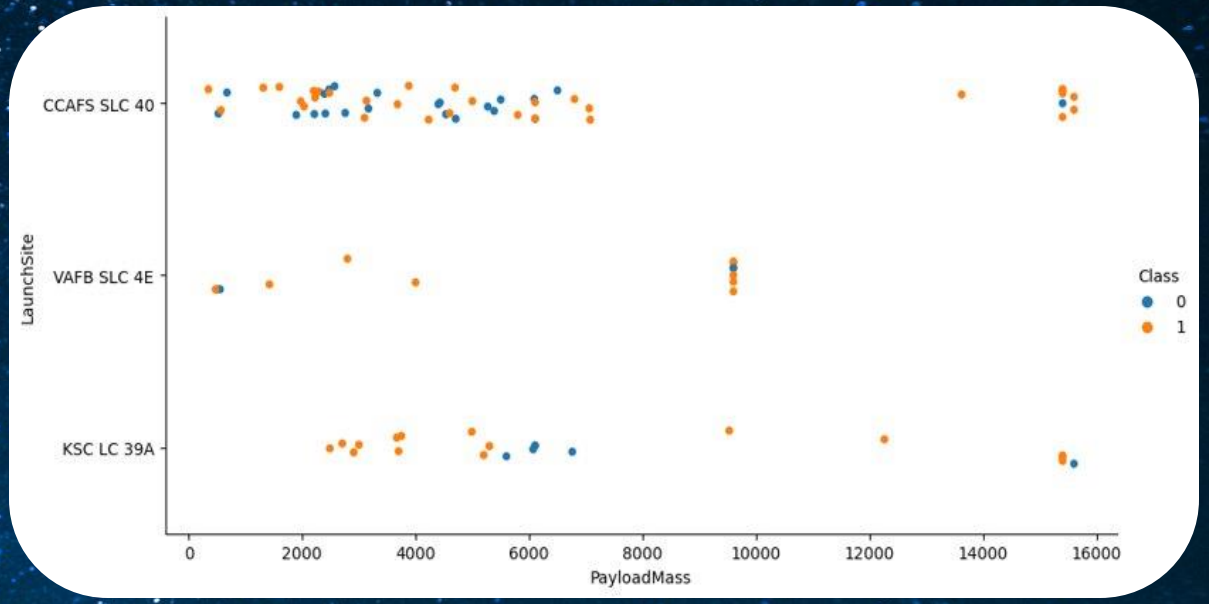
LAUNCH SITE VS FLIGHT NUMBER

Higher rate of failure (failed launches in **blue**) during earlier launches whereas the success rate notably increases after roughly the 20th launch.

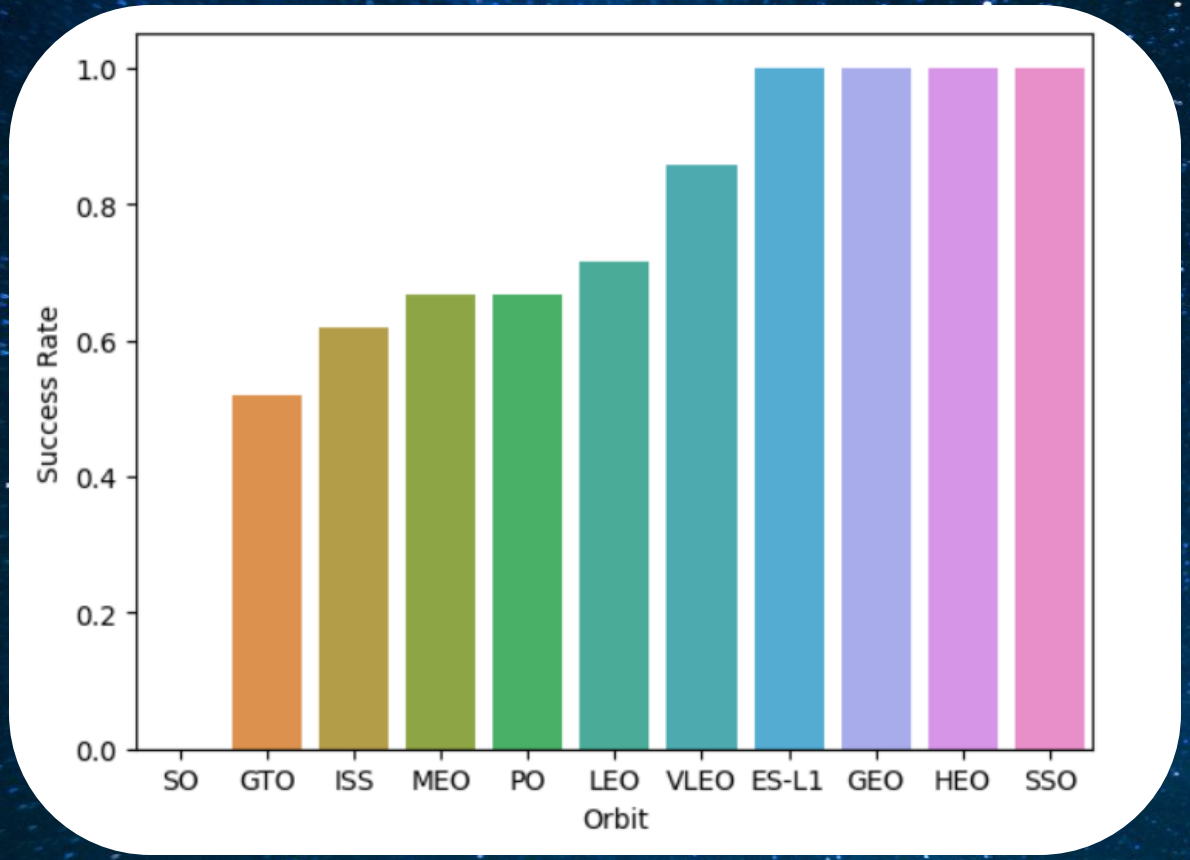


LAUNCH SITE VS PAYLOAD MASS

- No payloads larger than 10000kg launched from VAFB SLC 4E.
- Launches from KSC LC 39A had 100% success rate below a payload mass of ~5600kg.
- Fewer higher payload mass launches from CCAFS SLC 40 but higher success rate.



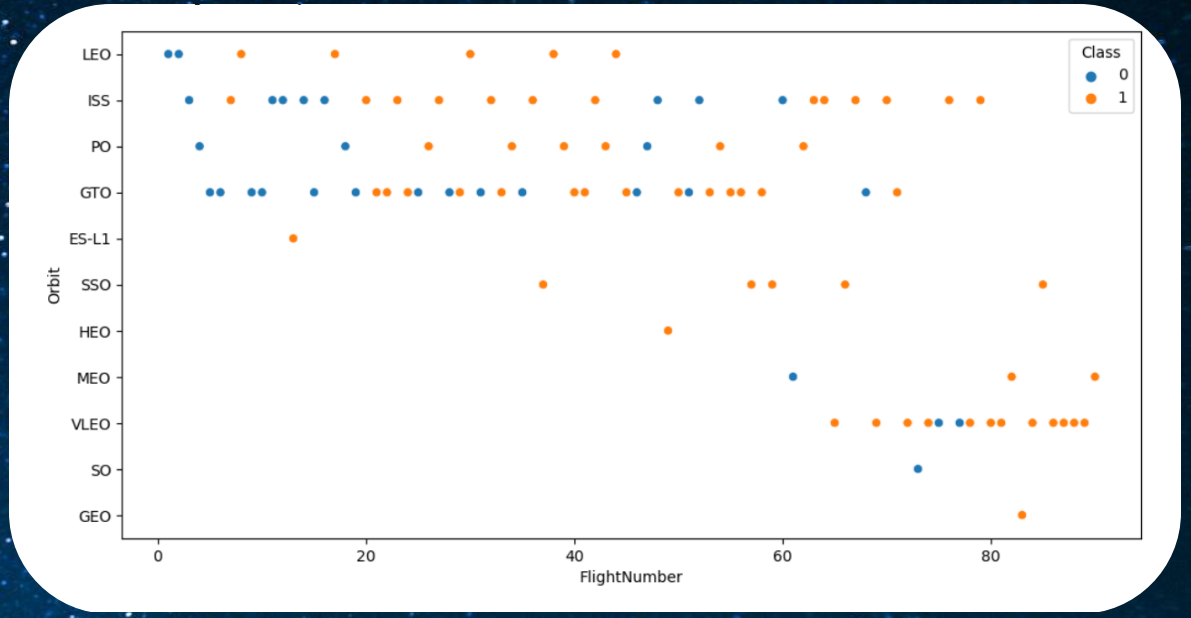
SUCCESS RATE OF EACH LAUNCH ORBIT



- **100% success rate:**
  - ES-L1, GEO, HEO, and SSO
- **50%-99%:**
  - GTO, ISS, MEO, PO, LEO, and VLEO
- **0%:**
  - SO

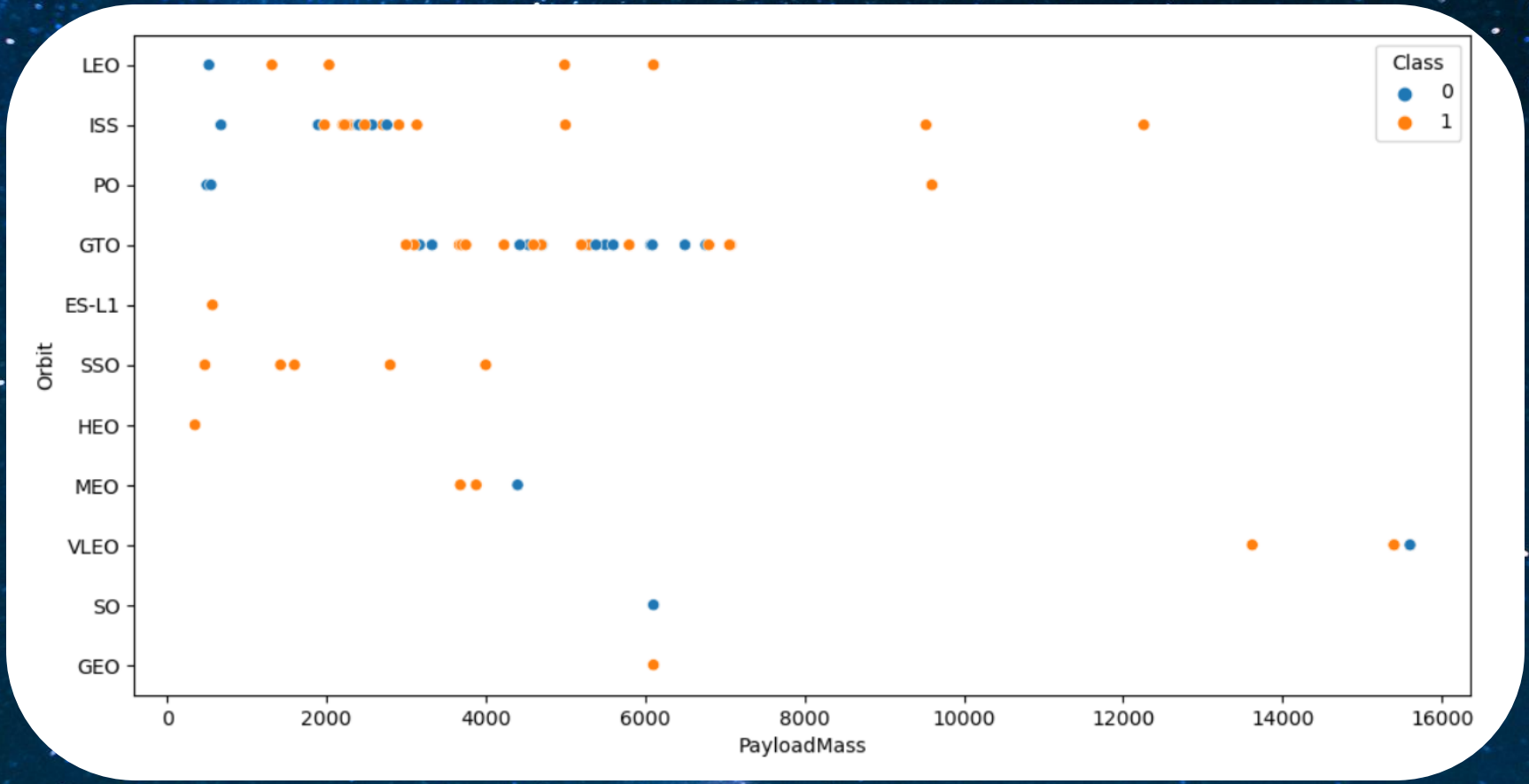
ORBIT VS FLIGHT NUMBER

- Still see a large number of failed launches in the first ~20 launches.
- No failed launches after initial two failures for LEO.
- 100% success rate for ES-L1, SSO, HEO, AND GEO (as demonstrated in previous chart also).

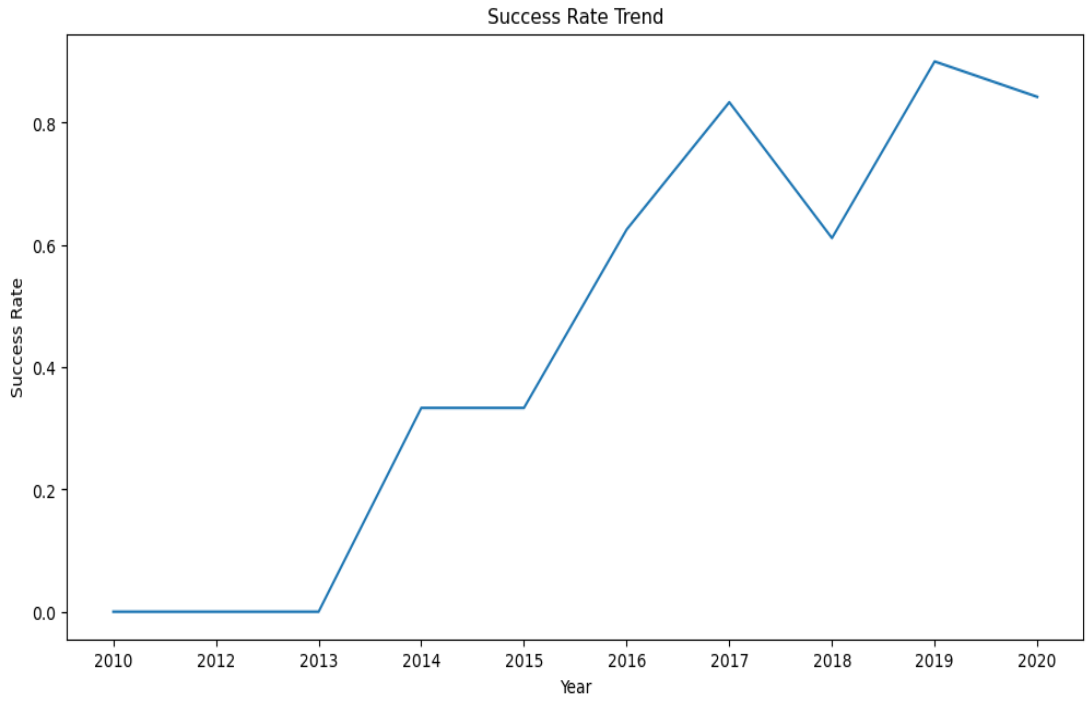


ORBIT VS PAYLOAD MASS

- High success rates with heavier payloads for LEO, ISS, and PO.



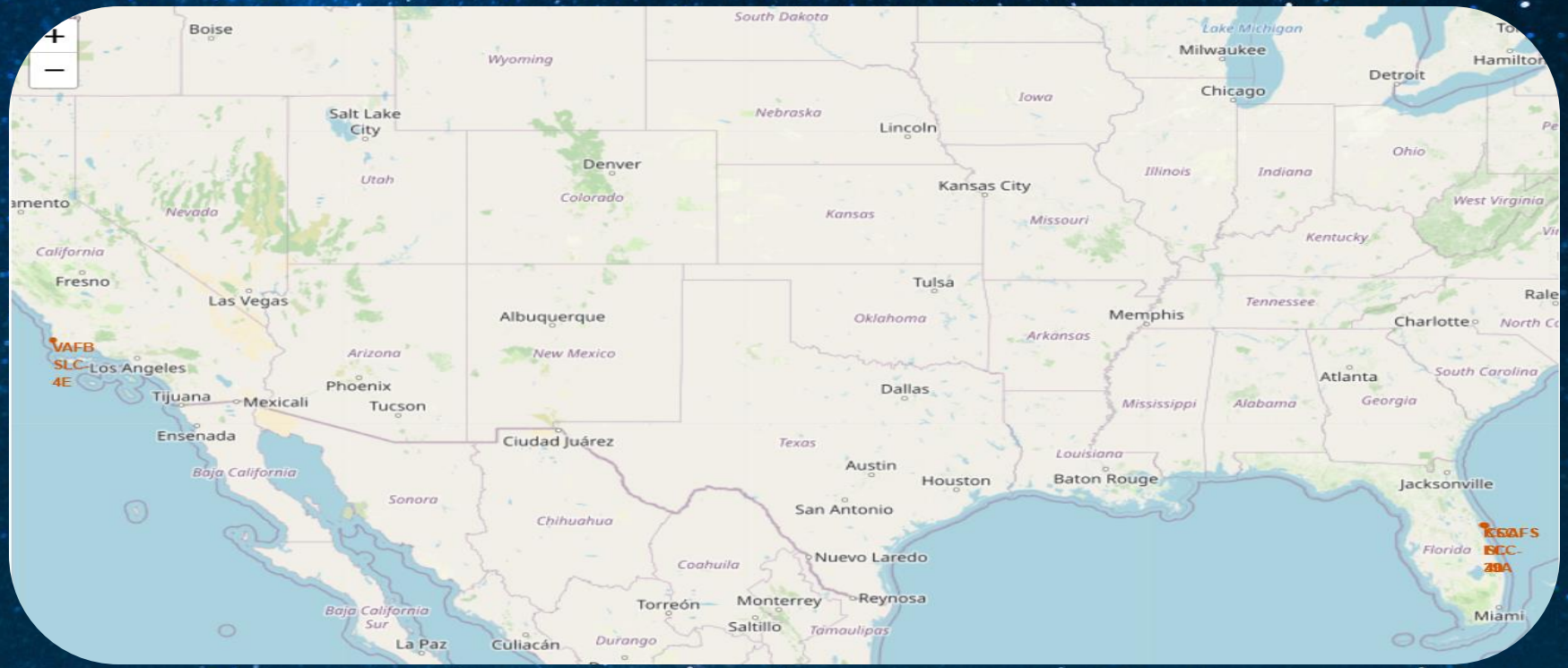
SUCCESS RATE TREND (2010-2020)



- Notably improved success rate over time, improving from 0% to >80%.
- Success rate plateaued in 2014.
- Dip in 2017, before improving again in 2018 to >80%.

LAUNCH SITE LOCATIONS - FOLIUM

- Launch sites are in close proximity to coast lines on either side of the US – California and Florida.

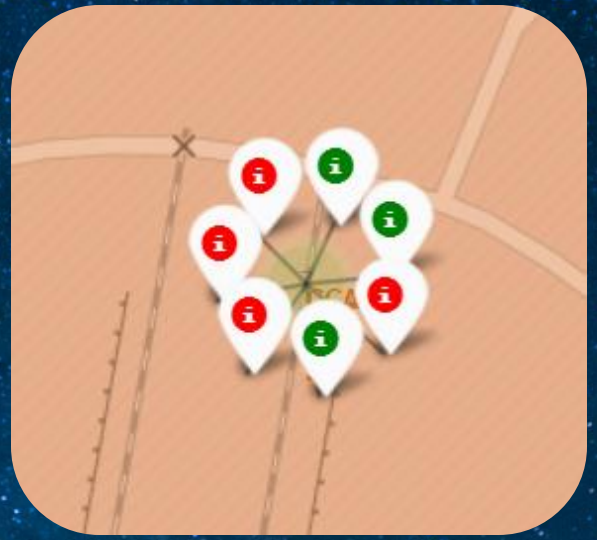


LAUNCH SITE LOCATIONS - FOLIUM

- Florida Launch Sites
  - Successful launch -> **Green**
  - Unsuccessful launch -> **Red**

CCAFS LC-40

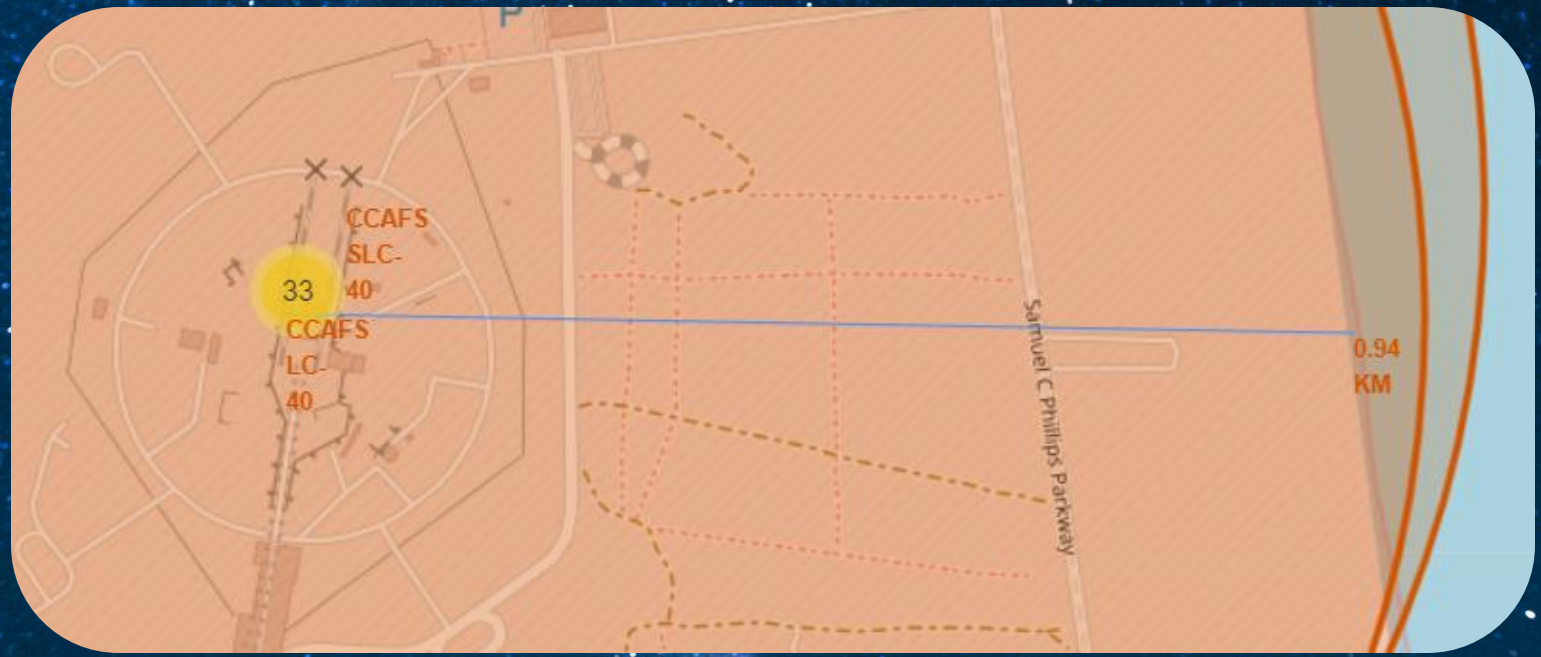
CCAFS SLC-40



KSC LC-39A

LAUNCH SITE LOCATIONS - FOLIUM

- Launch sites are situated very close to coast lines. The Cape Canaveral launch sites are ~1km from the nearest coastline. Having the booster fly over the ocean minimizes the risk of debris falling onto populated areas.

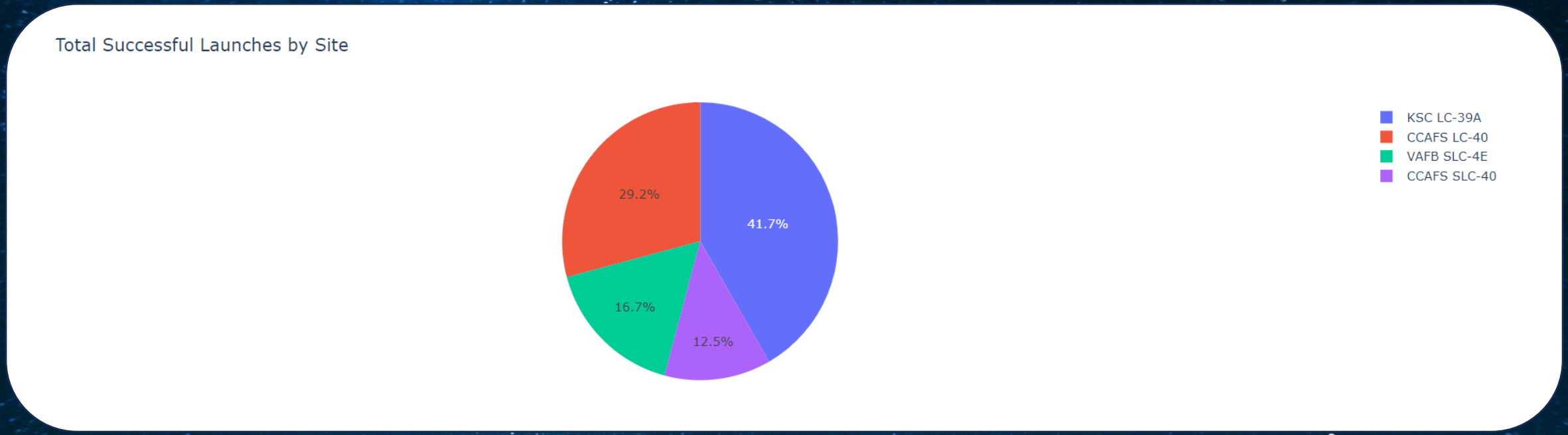


LAUNCH SITE LOCATIONS - FOLIUM

Launch sites are far from the nearest major highways and populated areas as a safety consideration in the event of an unsuccessful launch to prevent loss of life and damage to infrastructure.

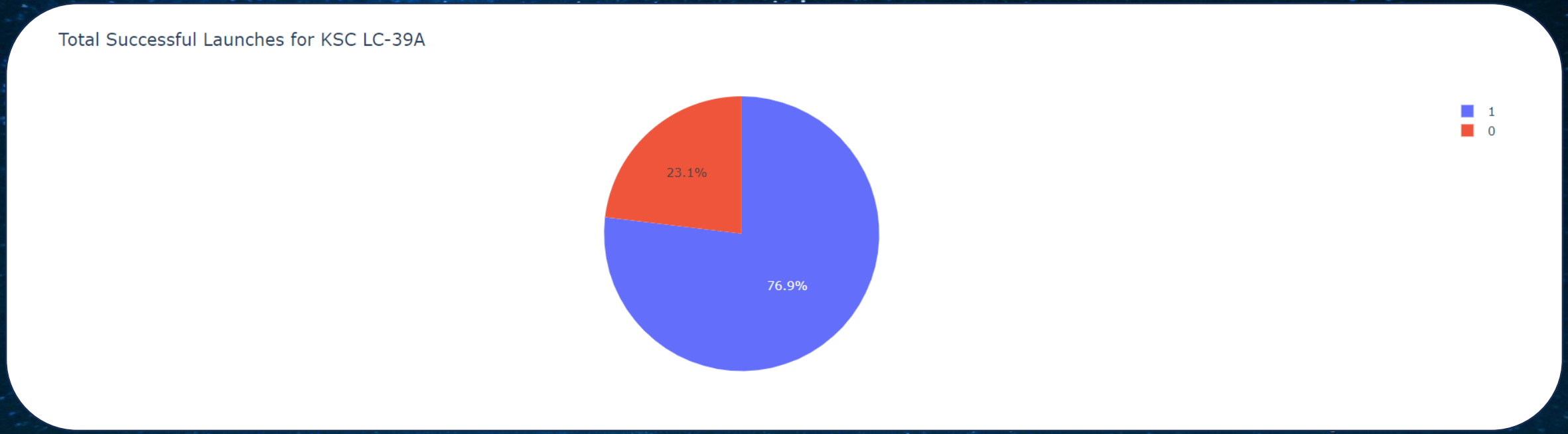
SPACEX LAUNCH RECORDS DASHBOARD

Highest number of successful launches occurred at KSC LC-39A (~42%).



SPACEX LAUNCH RECORDS DASHBOARD

KSC LC-39A launch site had a total success rate of ~77%.



SPACEX LAUNCH RECORDS DASHBOARD

Payloads between 2000kg and 5250kg have the highest success rate.

PREDICTIVE ANALYSIS - ACCURACY

- **Models tested:** Logistic Regression, SVM, Decision Tree, and KNN.
- Each model performed equally when using the accuracy (correct predictions / total predictions) as a metric of evaluation.

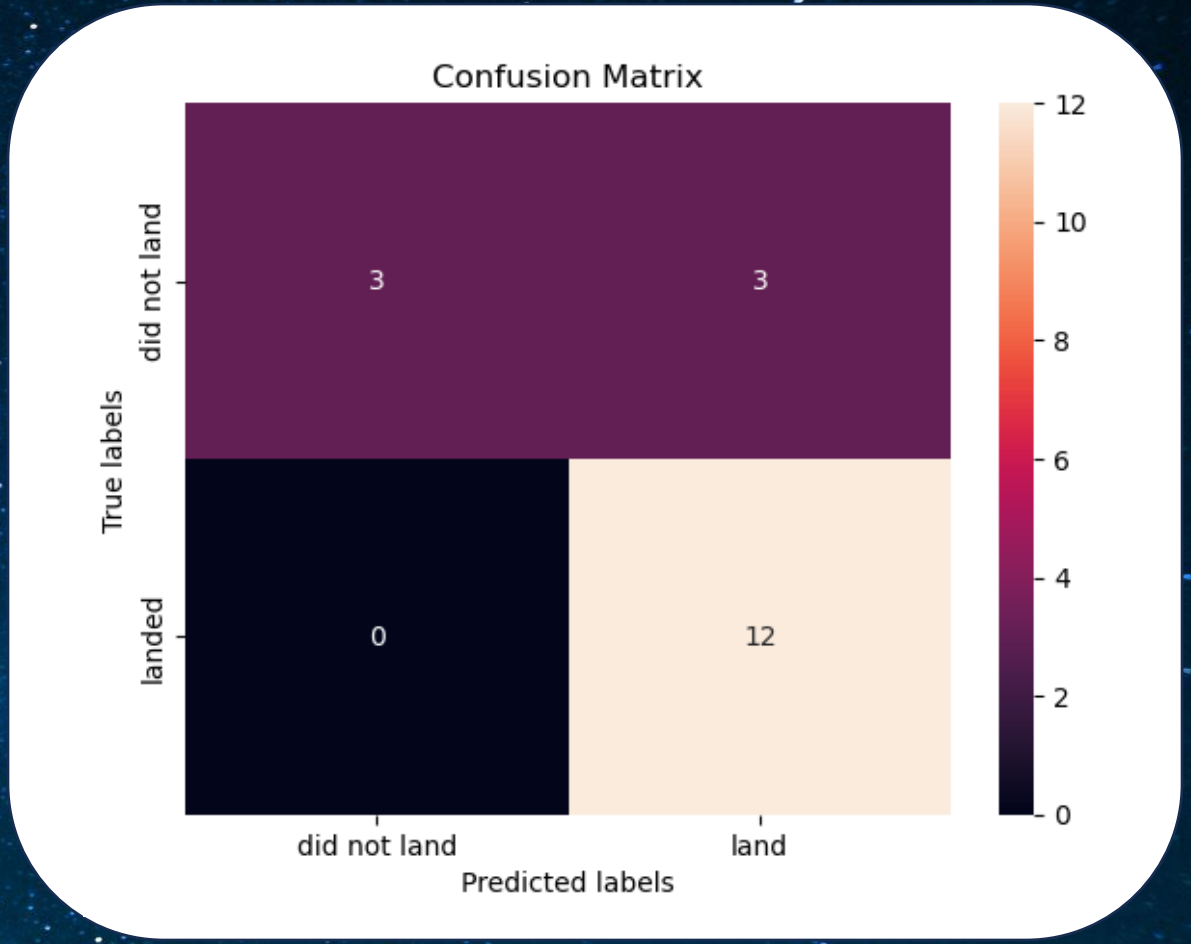
```
logreg_score = logreg_cv.score(X_test, Y_test)
svm_score = svm_cv.score(X_test, Y_test)
tree_score = tree_cv.score(X_test, Y_test)
knn_score = knn_cv.score(X_test, Y_test)

results = pd.DataFrame({'Estimator': ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN'],
                        'Score': [logreg_score, svm_score, tree_score, knn_score]}).set_index('Estimator')
results.index.name = None
results.sort_values('Score', ascending=False)
```

	Score
Logistic Regression	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

PREDICTIVE ANALYSIS – CONFUSION MATRICES

- All four ML models produced the same confusion matrix to the right.
- Values on the left to right diagonal represent correct predictions.
- 3 incorrect predictions and 15 correct predictions gives an accuracy of 83%.



CONCLUSION

CONCLUSION

Research:

- **Model Performance:**
  - All models performed equally when compared using the accuracy metric. Additional models could be tested for better performance.
- **Launch Sites:**
  - Located near coastlines and away from cities, highways, railways, and any other major infrastructure.
- **Success Rate:**
  - Improved over time. Much higher at some launch sites vs others and when launched into specific orbits.
- **Payload Mass:**
  - Generally, the higher the payload mass the better the success rate.

CONCLUSION

Further Improvements:

- **Dataset:**
  - A larger dataset would help train the models with more data and provide a larger test set (test set only had 18 samples).
- **ML Models:**
  - Tested several models, but there are many more that can be tested to ensure the best results are obtained.
- **RandomizedSearchCV:**
  - A grid search works well for a small range of parameters, but a larger range of parameters can be searched with shorter training time using RandomizedSearchCV for hyperparameter tuning.