Rocket Launch First Stage Reusage

Aaron Garcia

28-07-2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - EDA with SQL
  - EDA with Visualization
  - Interactive Maps with Folium
  - Plotly Dashboard
  - Predictive Analysis
- Conclusion

EXECUTIVE SUMMARY

Summary:

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 SQL and Data Visualization techniques to analyze main factors and calculating relevant statistics
- 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)

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)

METHODOLOGY

Steps

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Explore data via EDA with SQL and data visualization techniques
- Visualize the data using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection - API

- 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 – Web Scraping

- 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

Data Wrangling

- 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

Exploratory Data Analysis

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 (desc)

Exploratory Data Analysis

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.

Map with Folium

- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates
- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

Dashboard with Plotly

Dropdown List with Launch Sites

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

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

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

- Allow user to see the correlation between Payload and Launch Success

Predictive Analytics

- 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, support vector machine, decision tree, K-Nearest Neighbor
- 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

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

Launch Site Information

Launch Sites Names

```
%sql SELECT Launch_site FROM SPACEXTBL GROUP BY Launch_site;
```

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

Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

5 Records from launches starting with CCA

```
%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_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Payload Mass

Total Payload Mass carried by Boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_), Customer from SPACEXTBL where Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)	Customer
45596.0	NASA (CRS)

Average Payload Mass carried by F9 v1.1 Boosters

```
%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

Landing & Mission Info

1st Succesful Landing in Ground pad

```
%sql select MIN(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)
01/08/2018

Total of Every Outcome

```
%sql select Mission_outcome,count(Mission_outcome) \
from SPACEXTBL group by Mission_outcome;
```

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

Mission_Outcome	count(Mission_outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Booster Drone Ship Landing with mass between 4000 and 6000

```
%sql select Booster_Version, Landing_outcome, PAYLOAD_MASS__KG_ \
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	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696.0
F9 FT B1026	Success (drone ship)	4600.0
F9 FT B1021.2	Success (drone ship)	5300.0
F9 FT B1031.2	Success (drone ship)	5200.0

Boosters & Outcomes

Boosters with max Payload Mass

```
%sql select Booster_version, PAYLOAD_MASS__KG_ \
from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) \
                            from SPACEXTBL);
```

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

]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

Failed landings in Drone ship in 2015

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

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

Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

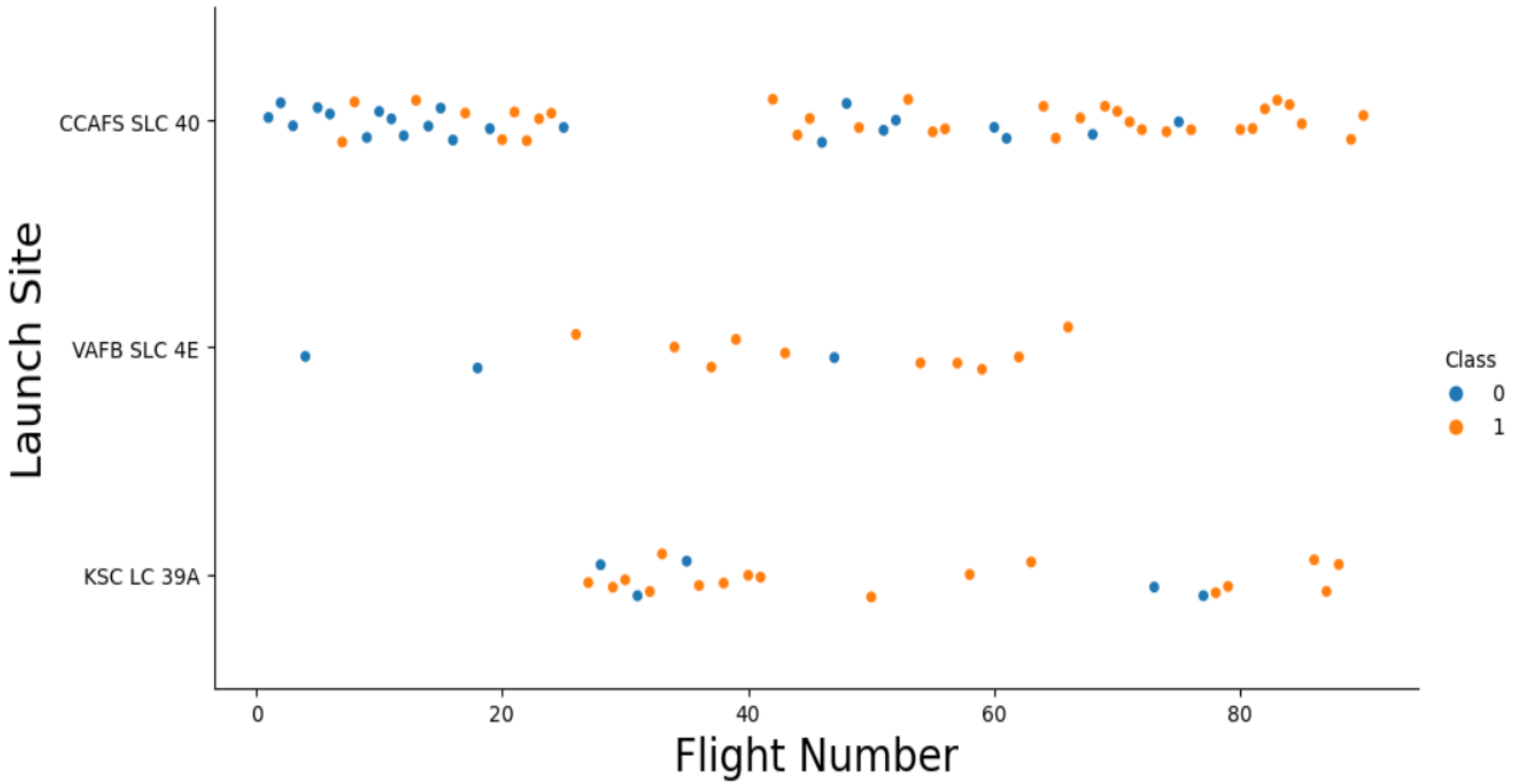
```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' \
group by [Landing_Outcome] \
order by count_outcomes DESC;
```

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

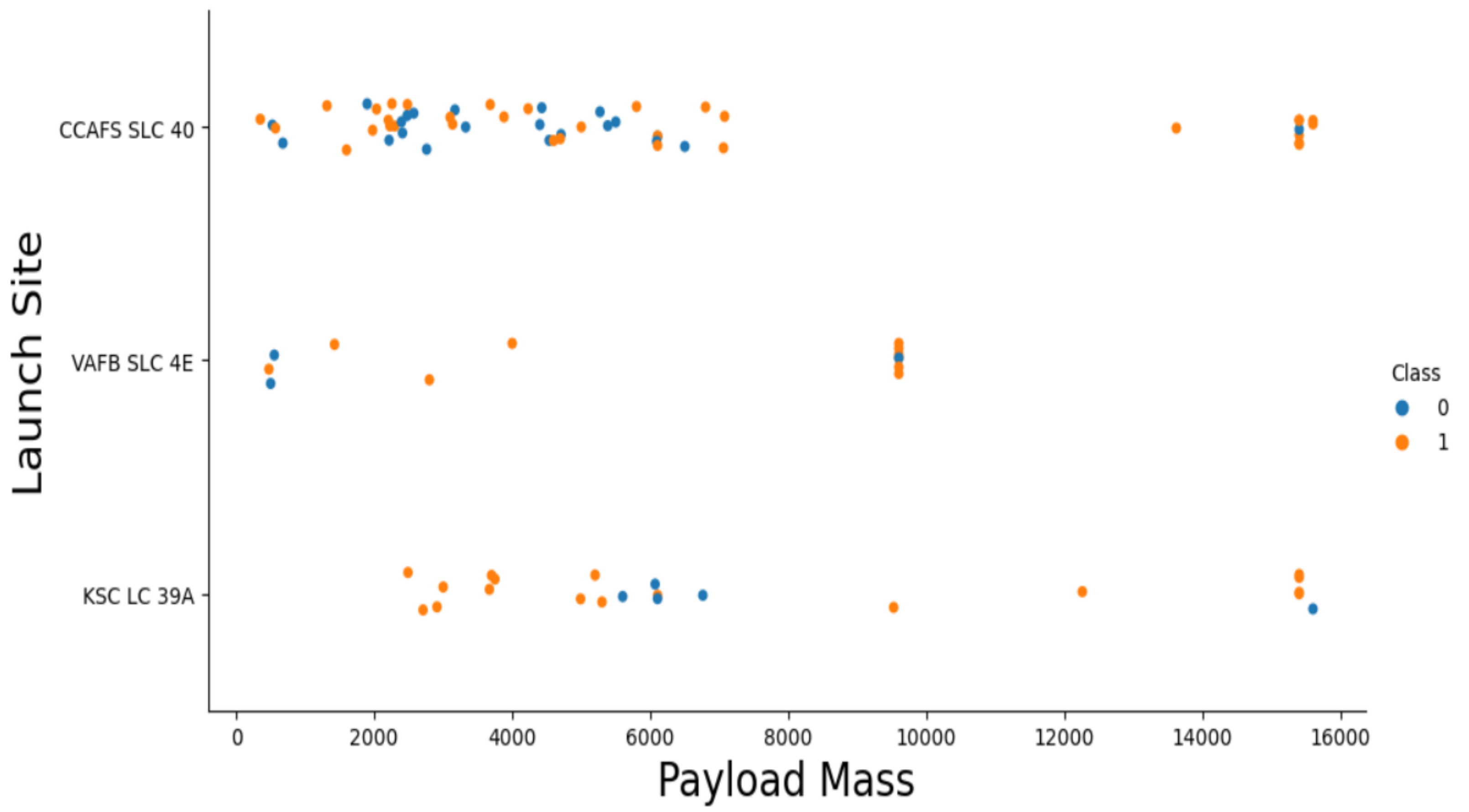
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Flight Number vs Launch Site

- 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



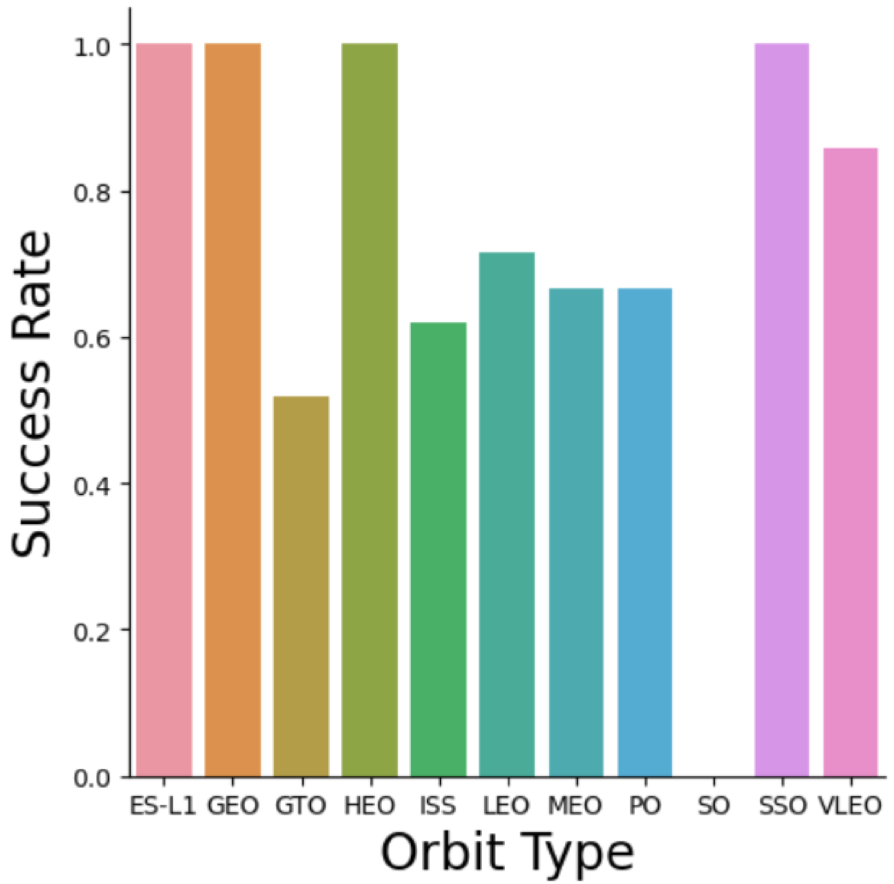
Payload Mass vs Launch Site



- 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

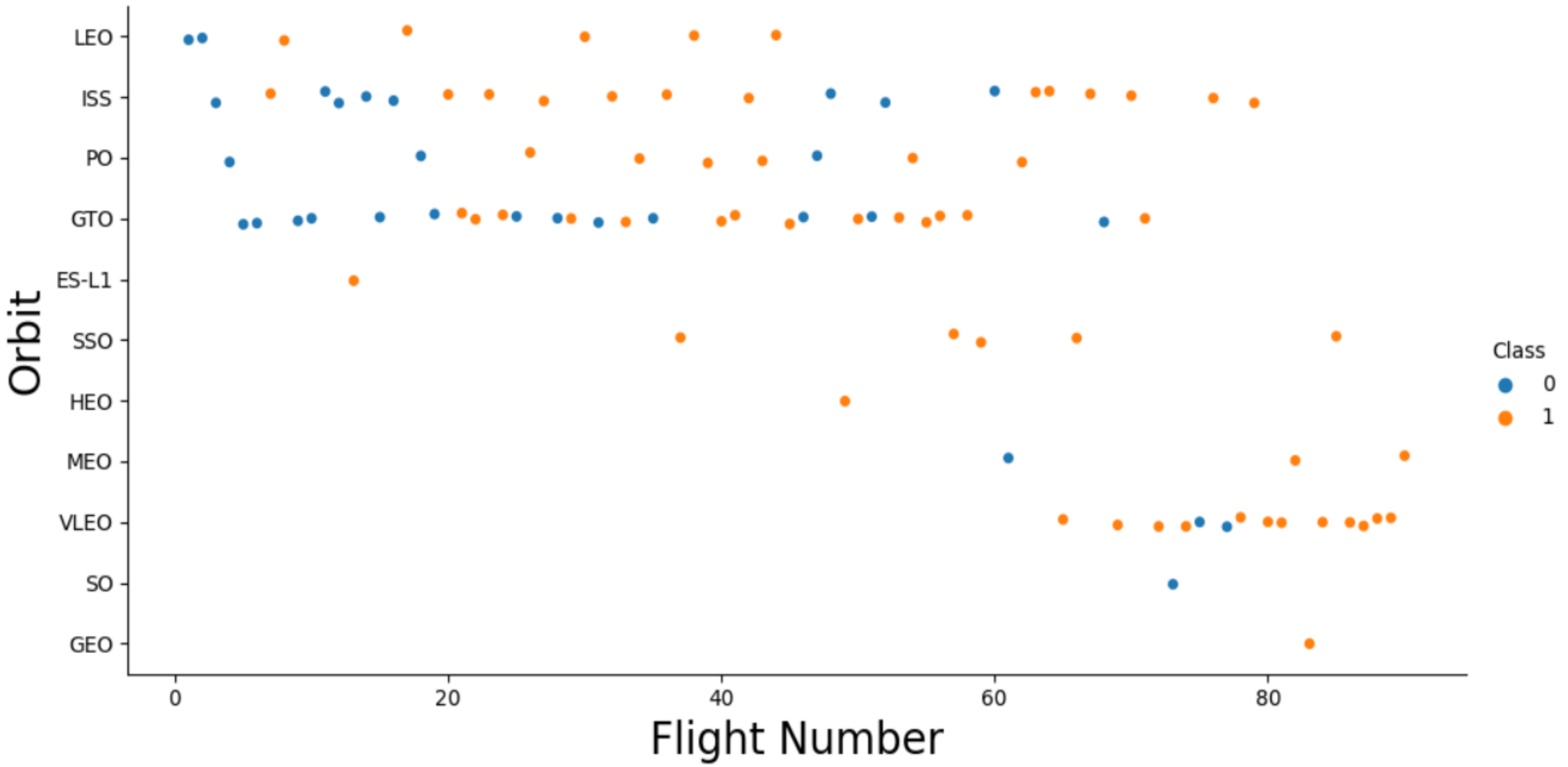
Success Rate by Orbit

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



Flight Number and Orbit

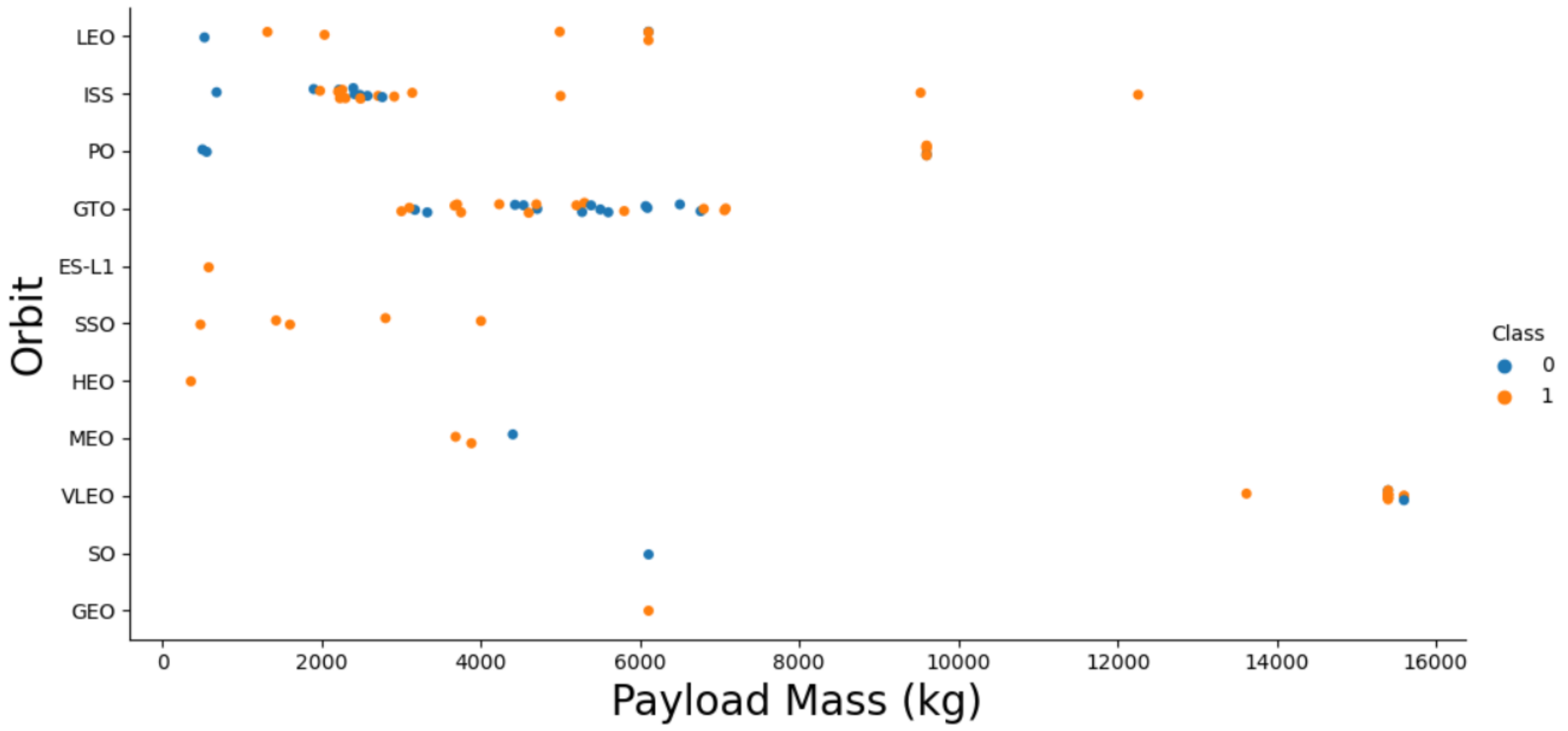
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload Mass and Orbit

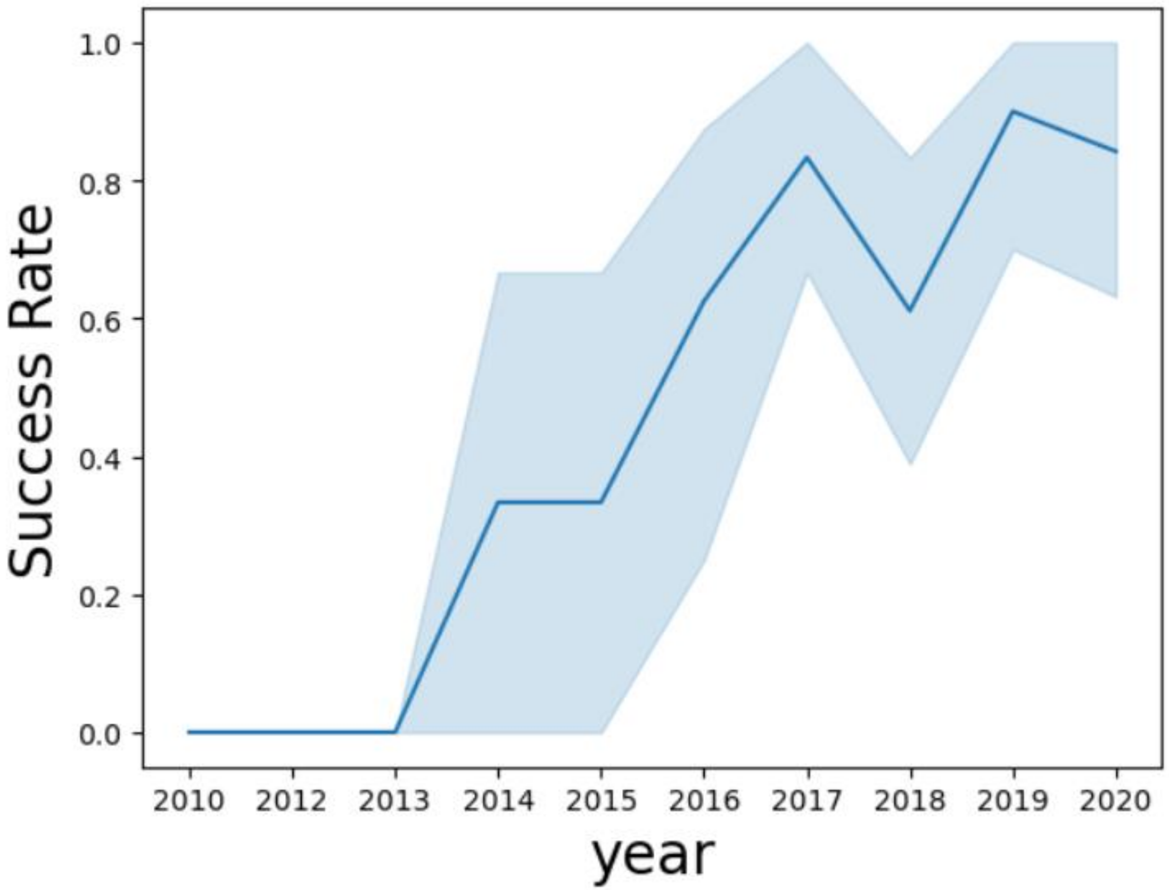
- Heavy payloads are better with LEO, ISS and PO orbits

- The GTO orbit has mixed success with heavier payloads



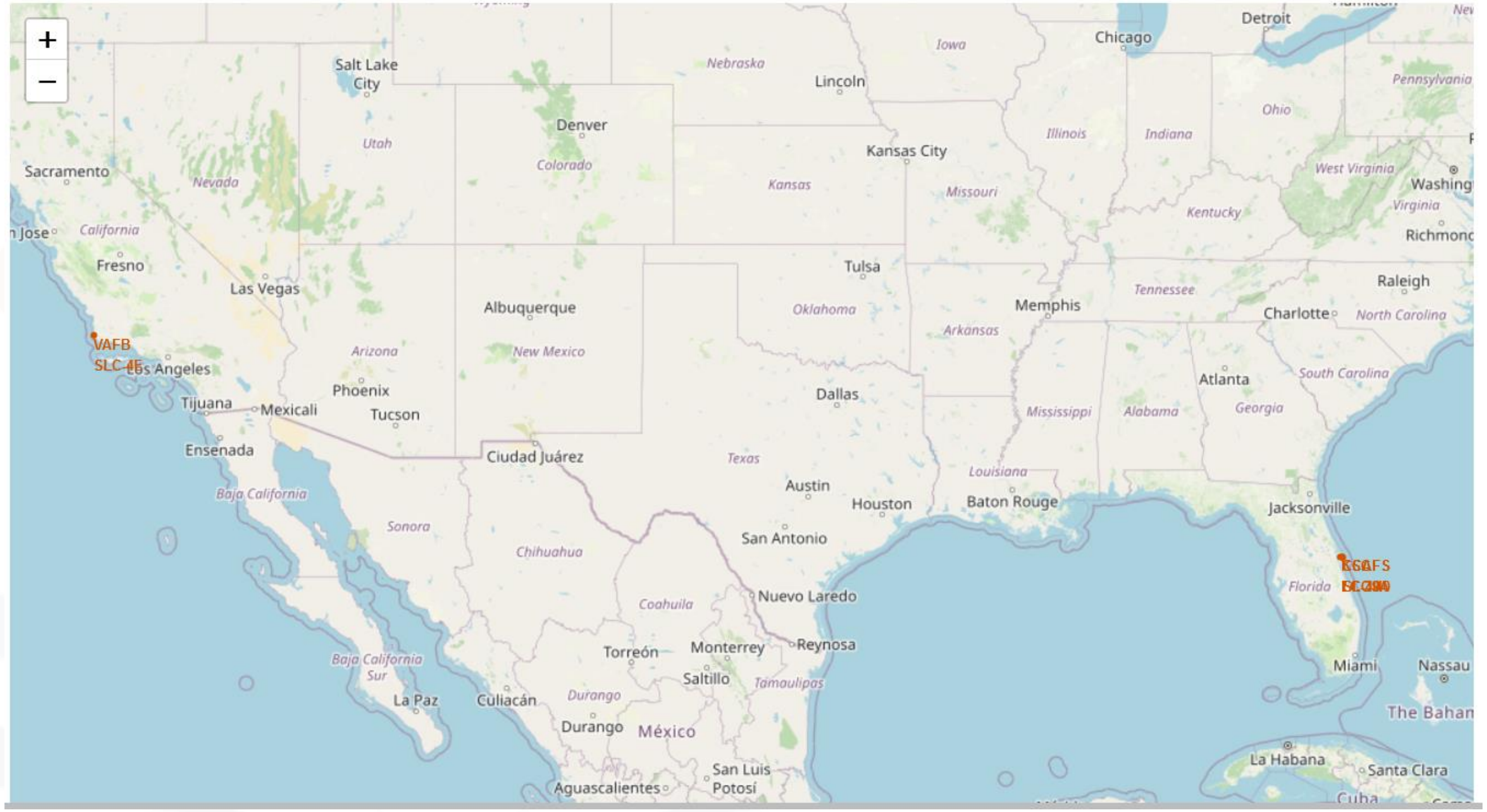
Launch Success Trend

- 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

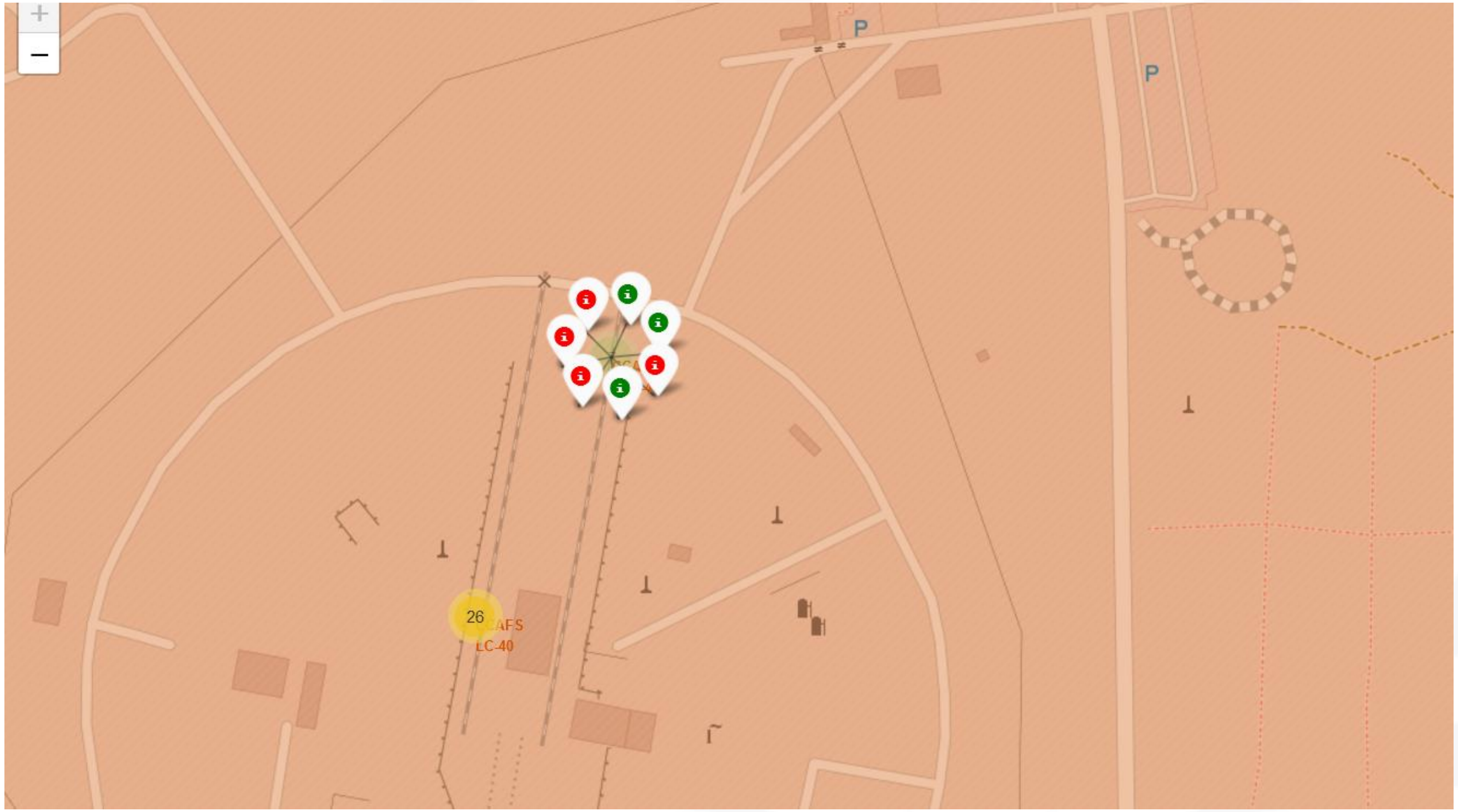


Launch Sites

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



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

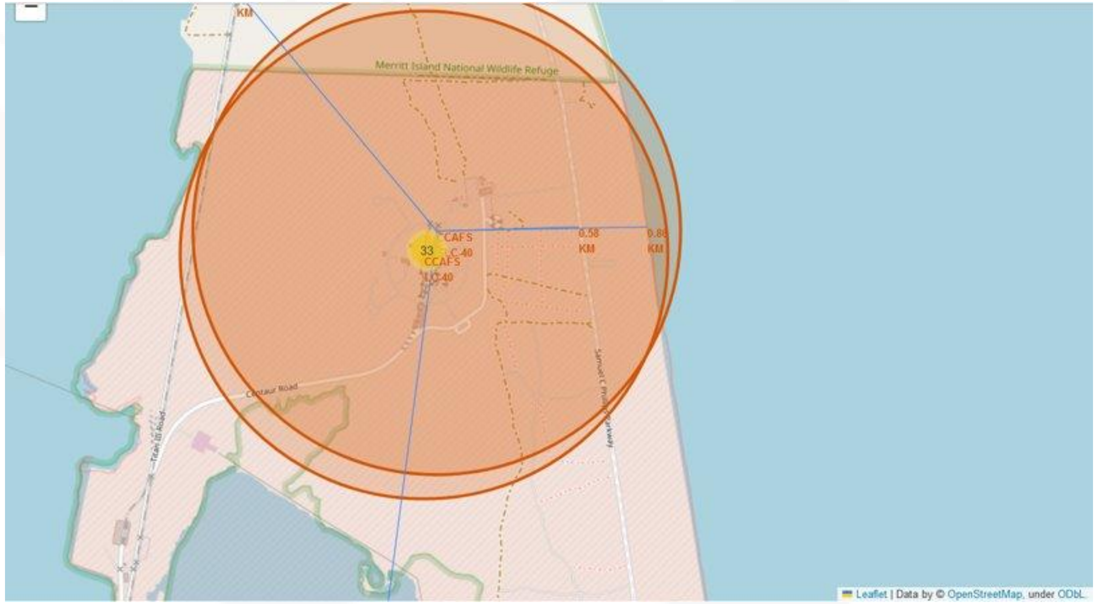
Distance to Proximities

- 0.86 km to nearest coastline
- 0.58 km to nearest road
- 1.28 km to nearest railway

Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.

Safety / Security: needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.

Transportation: need to be close enough to roads/rails/docks to be able to transport people and materials to support of launch activities.



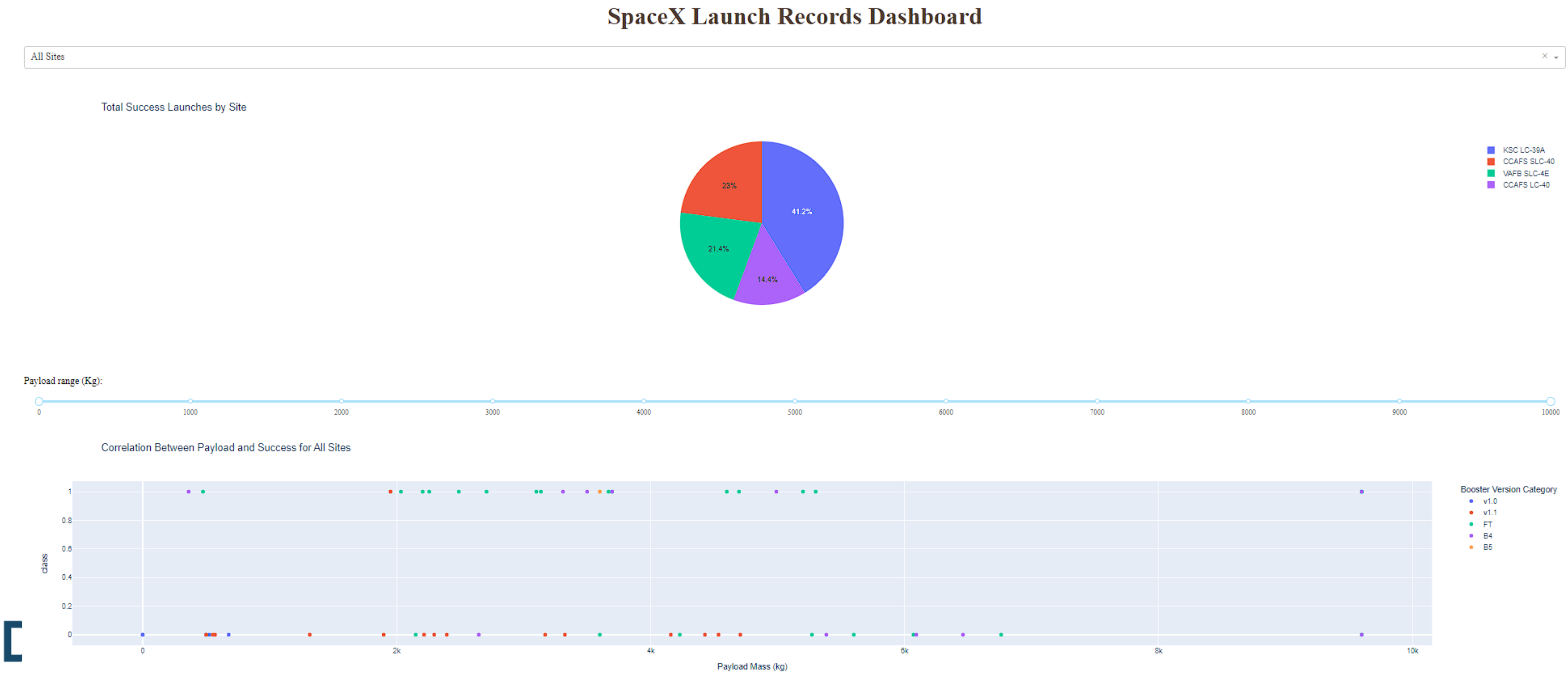
Distance to Proximities

- 51.43 km to nearest city (Melbourne)

Cities: need to be away from any city for any damage caused by a failed launch

Dashboard with Plotly

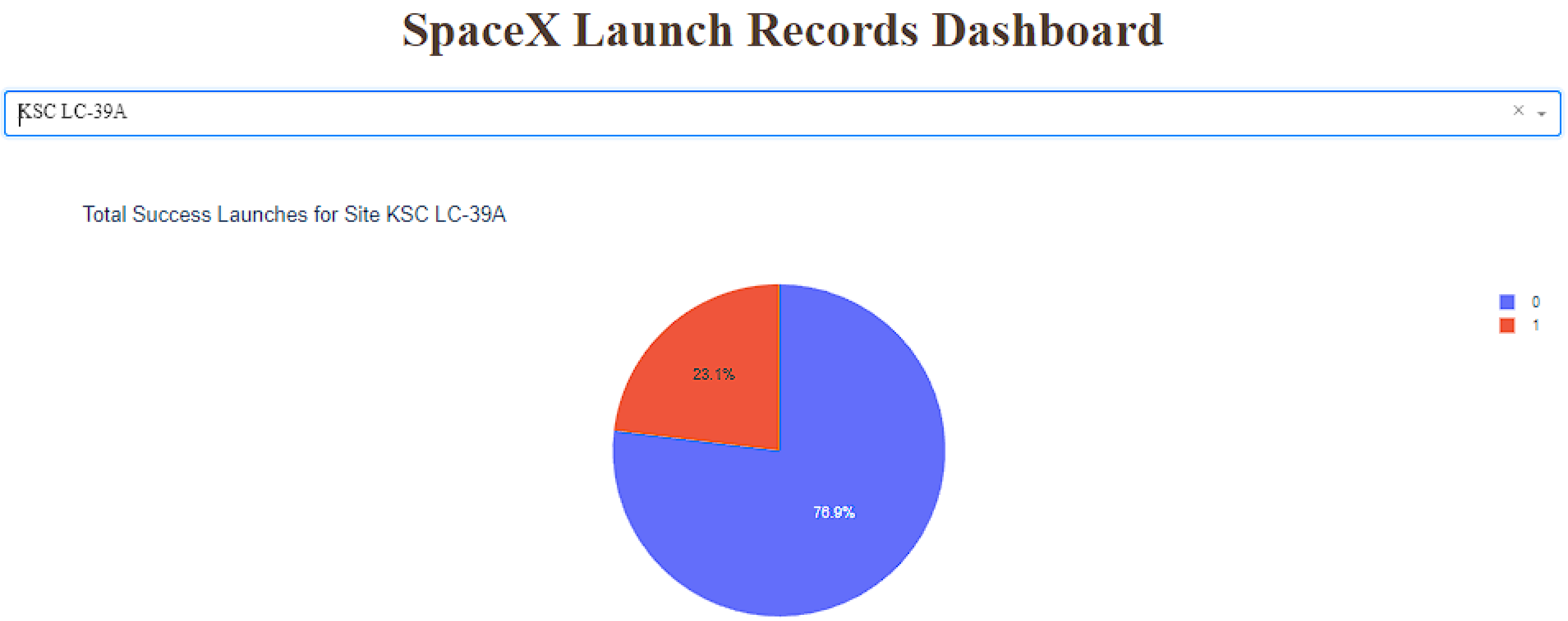
KSC LC-39A has the most successful launches amongst launch sites (41.2%)



IBM

Launch Success

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches (1=Failure, 0=Success)



Payload Mass and Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

Predictive Analysis

Accuracies

```python
print("Logistic Regression test set accuracy :",logreg_cv.score(X_test, Y_test))
print("Support Vector Machine (SVM) test set accuracy :",svm_cv.score(X_test, Y_test))
print("Decission Tree test set accuracy :",tree_cv.score(X_test, Y_test))
print("KN Neighbors test set accuracy :",knn_cv.score(X_test, Y_test))

# Decission Tree is the best method
```

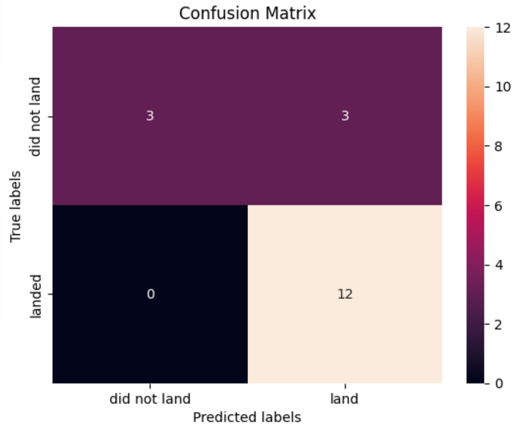
```
Logistic Regression test set accuracy : 0.8333333333333334
Support Vector Machine (SVM) test set accuracy : 0.8333333333333334
Decission Tree test set accuracy : 0.8888888888888888
KN Neighbors test set accuracy : 0.8333333333333334
```

From the 4 evaluated models, the Decision Tree model performed the best by having the highest accuracy and the lowest false positives.

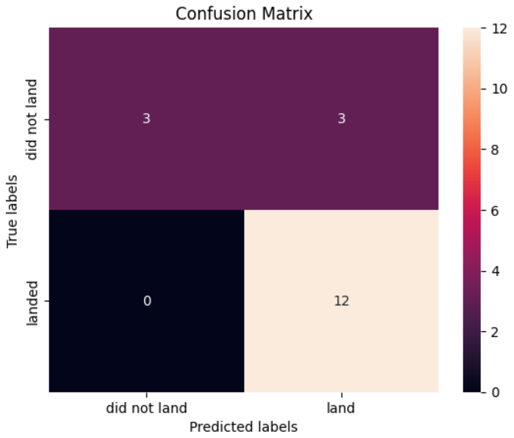
Logistic Regression, Support Vector Machine and K-Nearest Neighbors confusion matrices were identical.

Confusion Matrices

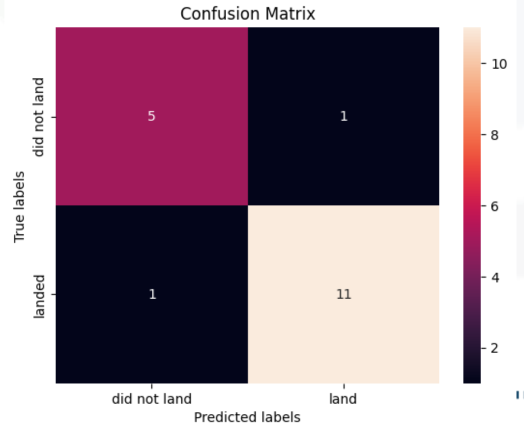
Logistic Regression



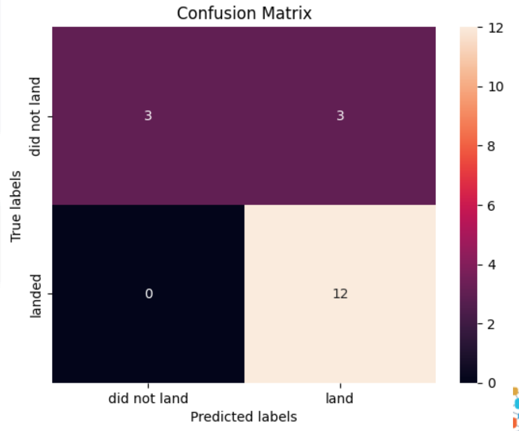
Support Vector Machine



Decission Tree



K-Nearest Neighbors



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
- Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- XGBoost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models