Winning Space Race with Data Science

Udaykumar gampa
02-09-2023

Outline

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

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 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

Project background and context

- 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.

Problems you want to find answers

- 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

- **Data collection methodology:**
  - Collect data using SpaceX REST API and web scraping technique.
- **Perform data wrangling:**
  - By filtering the data, handling missing values and applying one hot encoding to prepare the data for analysis and modeling.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters.

Data Collection – API

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 Describe how data sets were collected.

Github url :

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

Data Wrangling

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

Landing Outcome

- Landing was not always successful
- True Ocean: mission outcome had a successful landing to a specific region of the ocean

Landing Outcome Cont.

- **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 L

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 EDA with Visualization

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.

EDA with SQL

Queries

- 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)

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 SLC40 and its proximity to the nearest coastline, railway, highway, and city

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

- 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 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

Results

Exploratory data analysis results

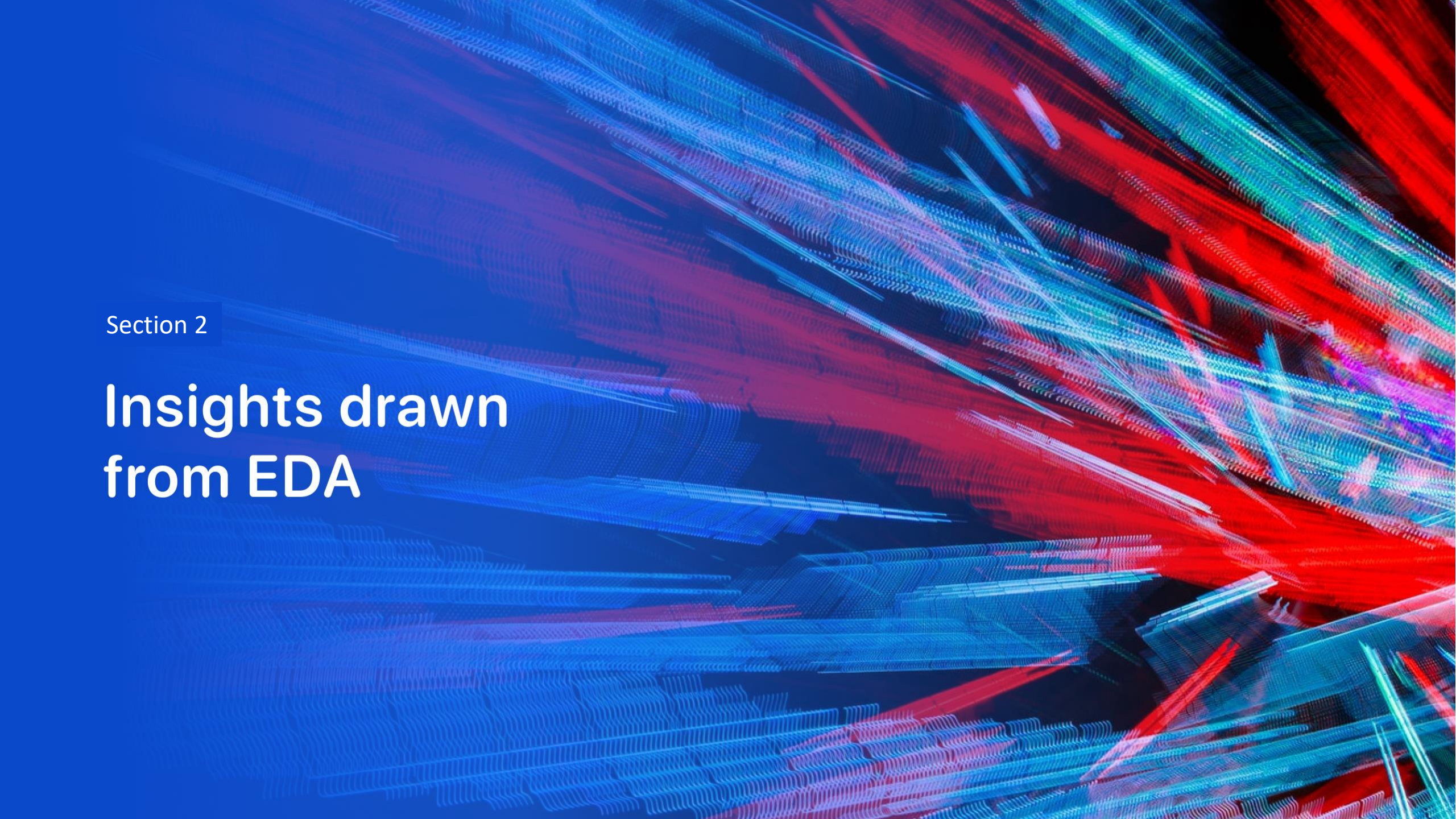
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Interactive analytics dashboard

- 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 analysis results

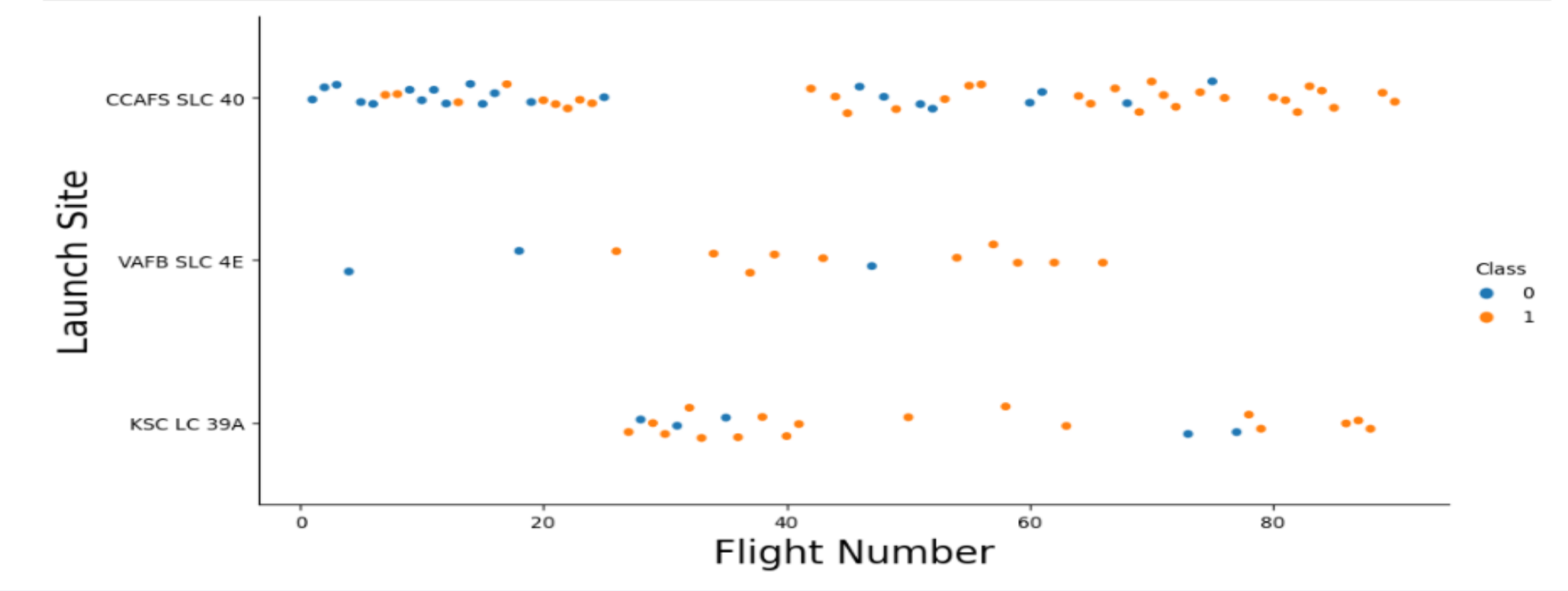
- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

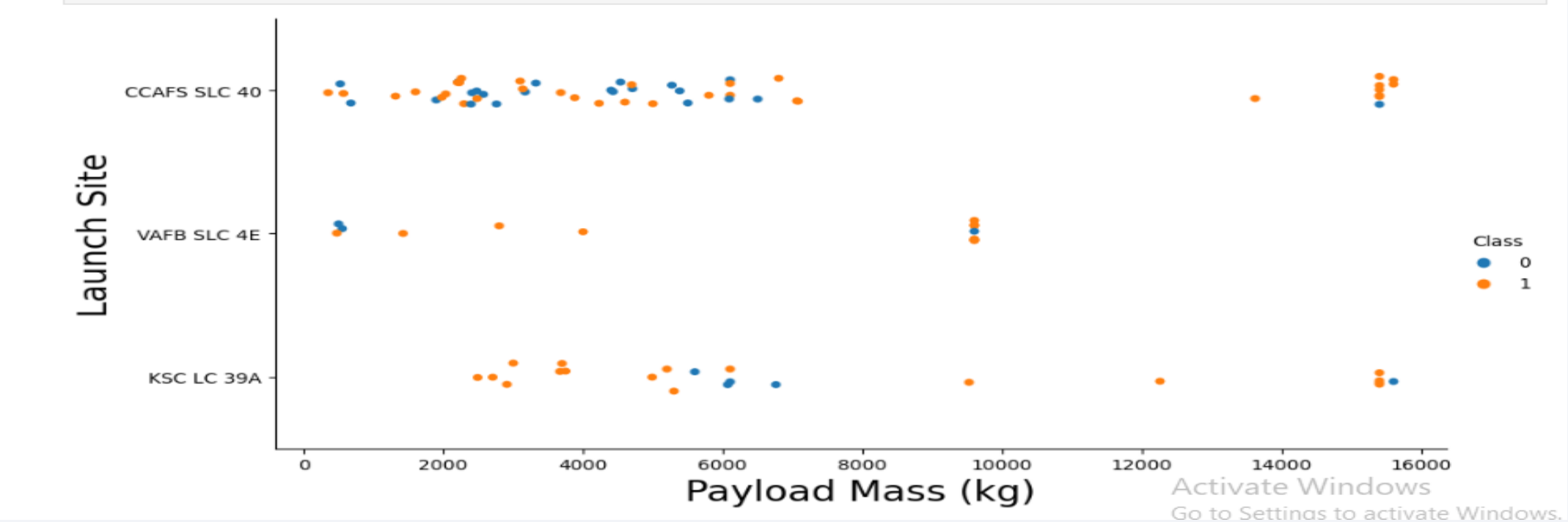
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

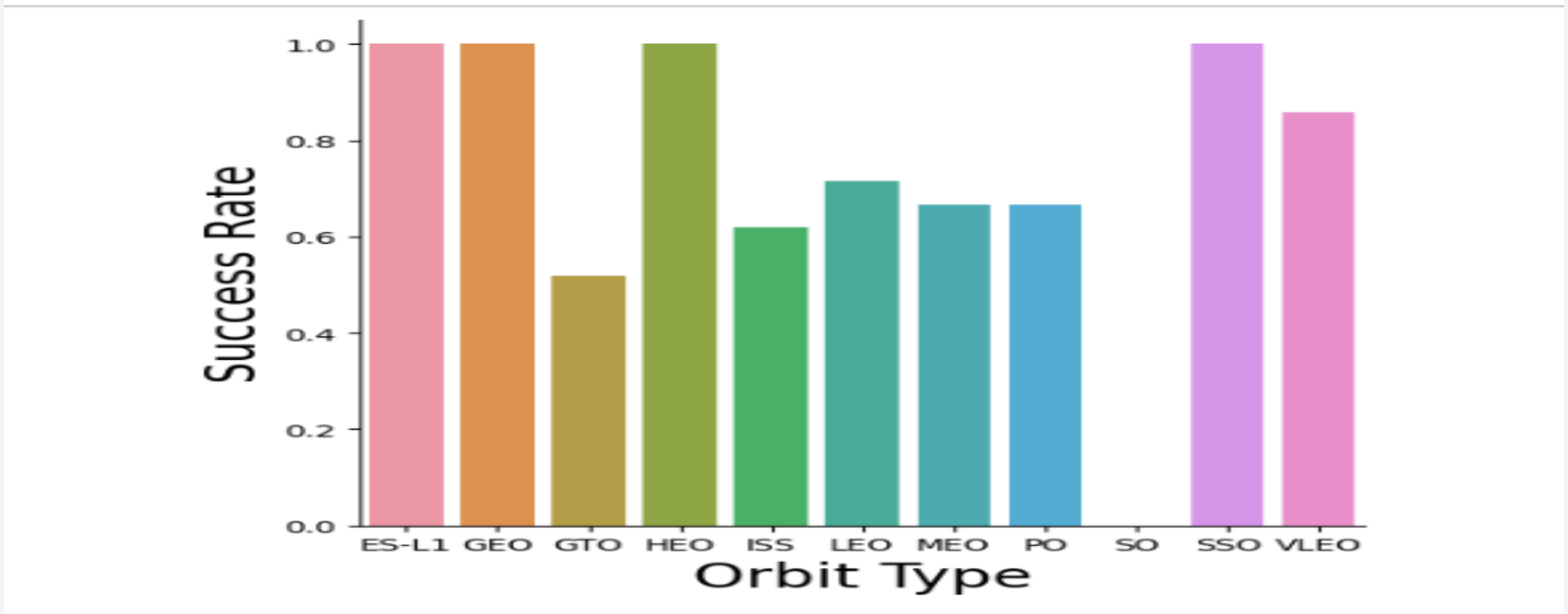
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

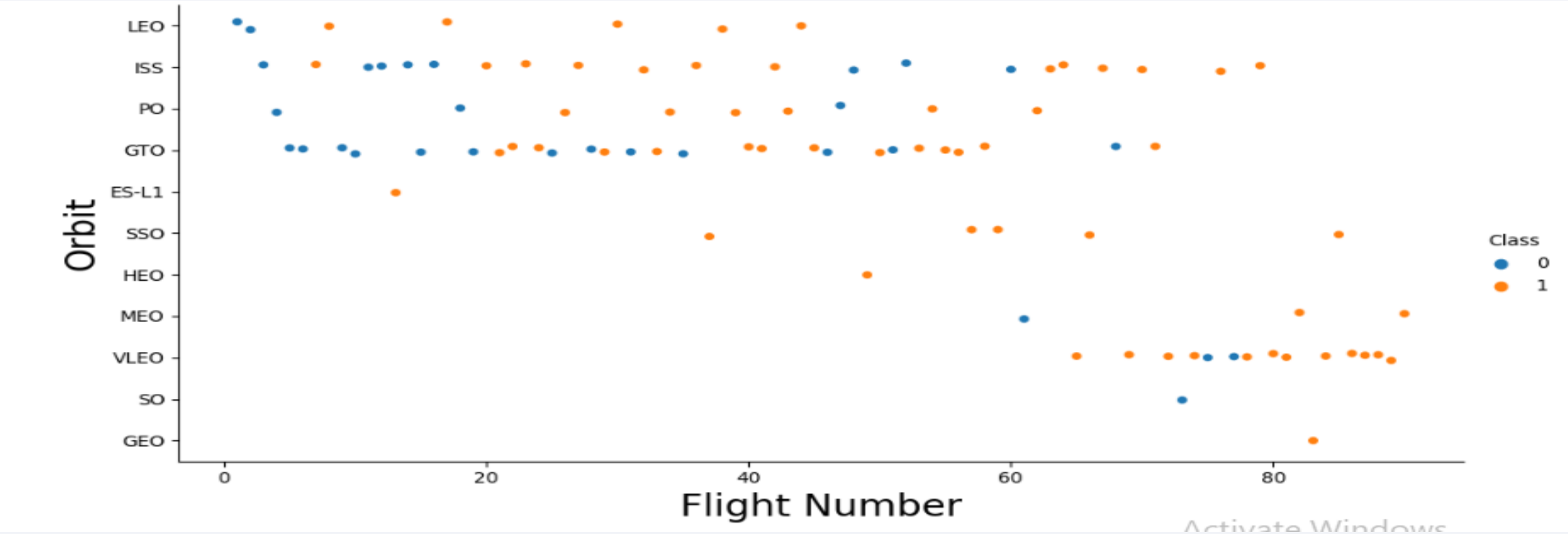
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

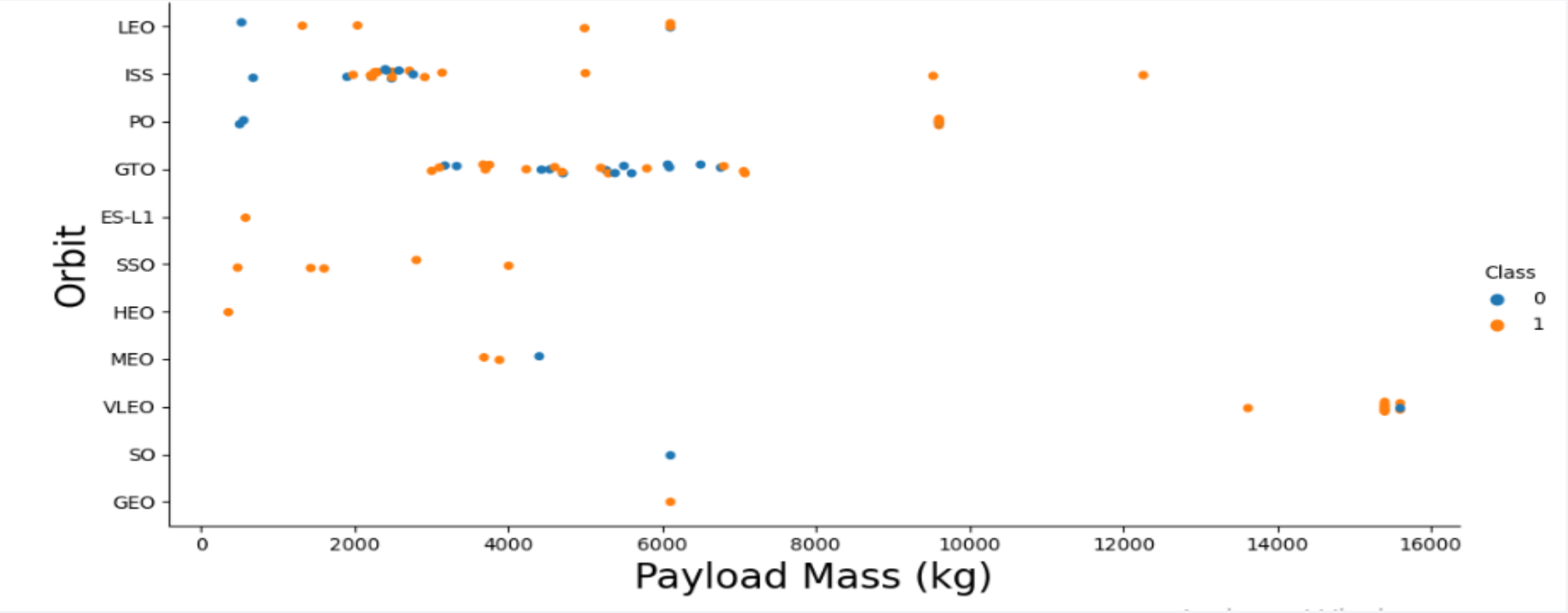
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

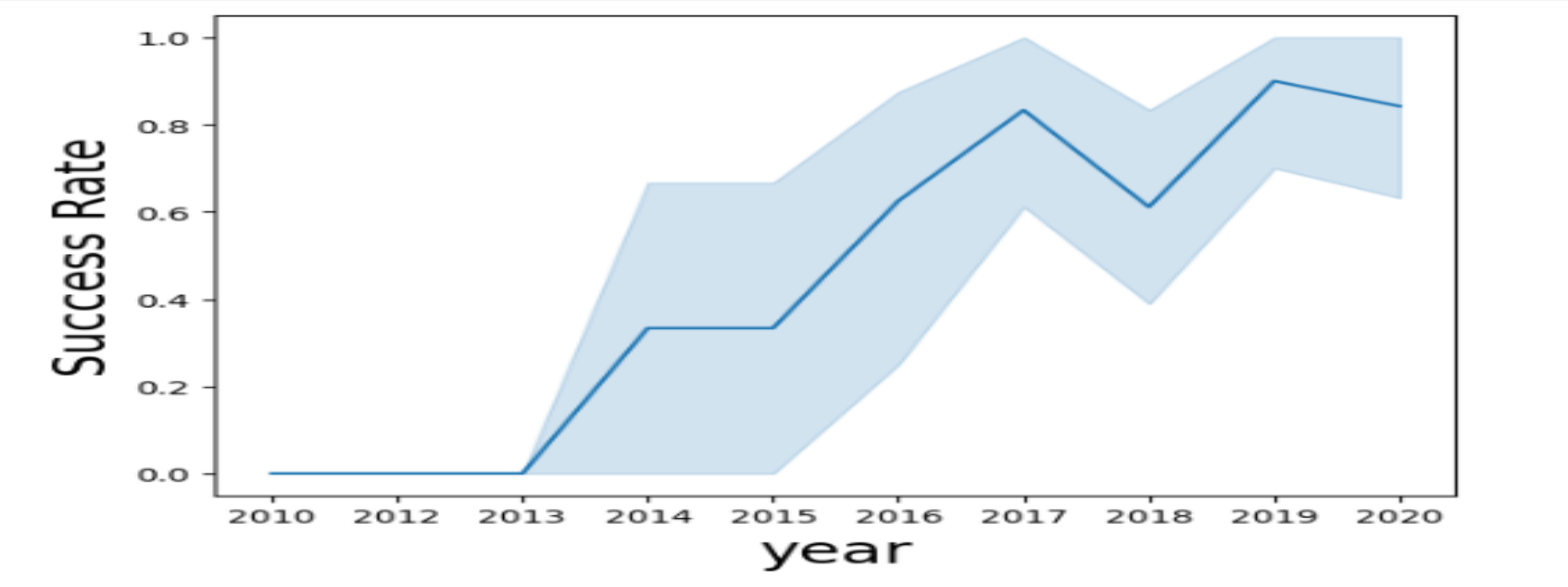
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



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

Unique launch sites names

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

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

```
[8]: %sql select distinct("LAUNCH_SITE") from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[8]:

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`

Task 2

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

```
[18]: %sql select * from SPACEXTABLE where LAUNCH_SITE like 'CCA%' limit 5
```

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

[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
2010-08-12	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)	Success	Failure (para
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight	525	LEO (ISS)			

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)

```
[14]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
```

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

[14]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average Payload Mass

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

Task 4

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

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

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

[16]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

1st Successful Landing in Ground Pad

- 12/22/2015

Task 5

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

Hint:Use min function

```
[19]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome= "Success (ground pad)"
```

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

[19]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

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

```
[26]: %sql select Payload from SPACEXTBL where (Landing_Outcome = 'Success (drone ship)') \
      and (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

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

[26]:

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

Task 7

List the total number of successful and failure mission outcomes

```
27]: %sql select count(Mission_Outcome) ,Mission_Outcome from SPACEXTBL group by Mission_Outcome
```

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

27]:

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

Boosters Carried Maximum Payload

Carrying Max Payload

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

```
[30]: %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ =(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

[30]:

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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql select substr(Date, 6, 2) as month, Date, Booster_Version, Launch_Site from SPACEXTBL \
where Landing_Outcome = 'Failure (drone ship)'and substr(Date,1,4)= '2015';
```

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

month	Date	Booster_Version	Launch_Site
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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.

[75]:
```
%sql select Landing_Outcome, count(*) as count_outcomes \
from SPACEXTBL \
where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome \
order by count_outcomes desc
```

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

[75]:

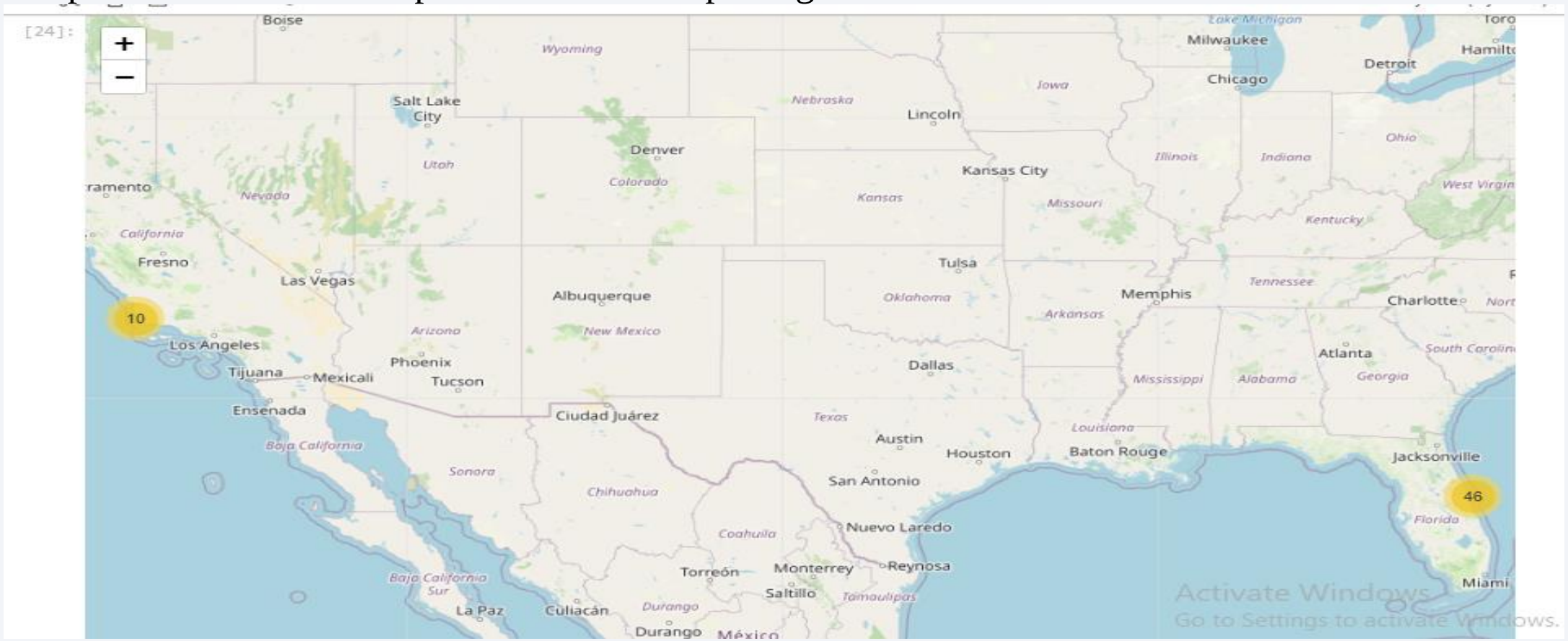
Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites with Markers

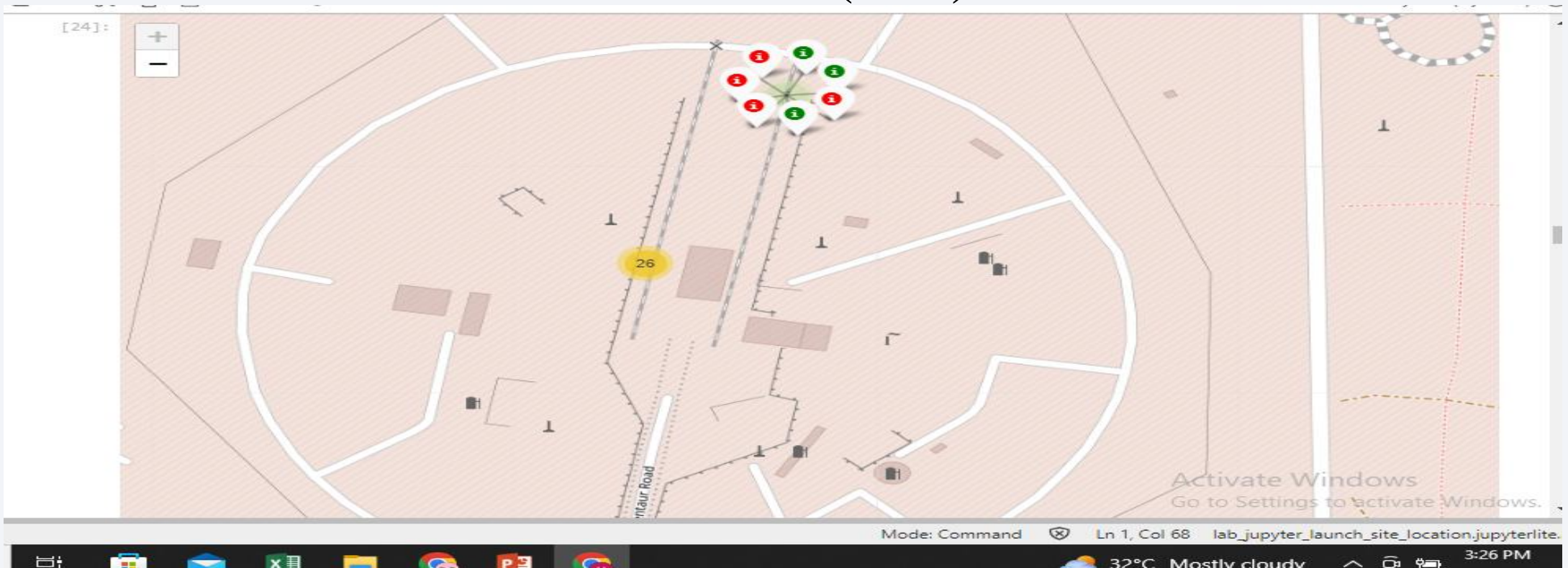
- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps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



Mark the success/failed launches for each site on the map

Outcomes:

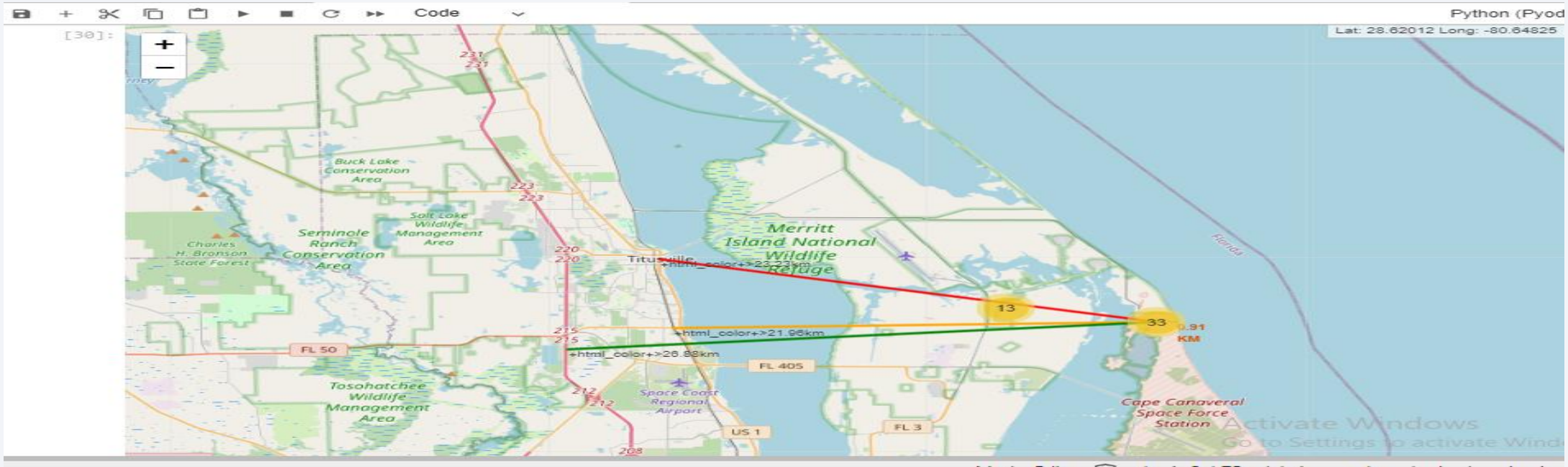
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

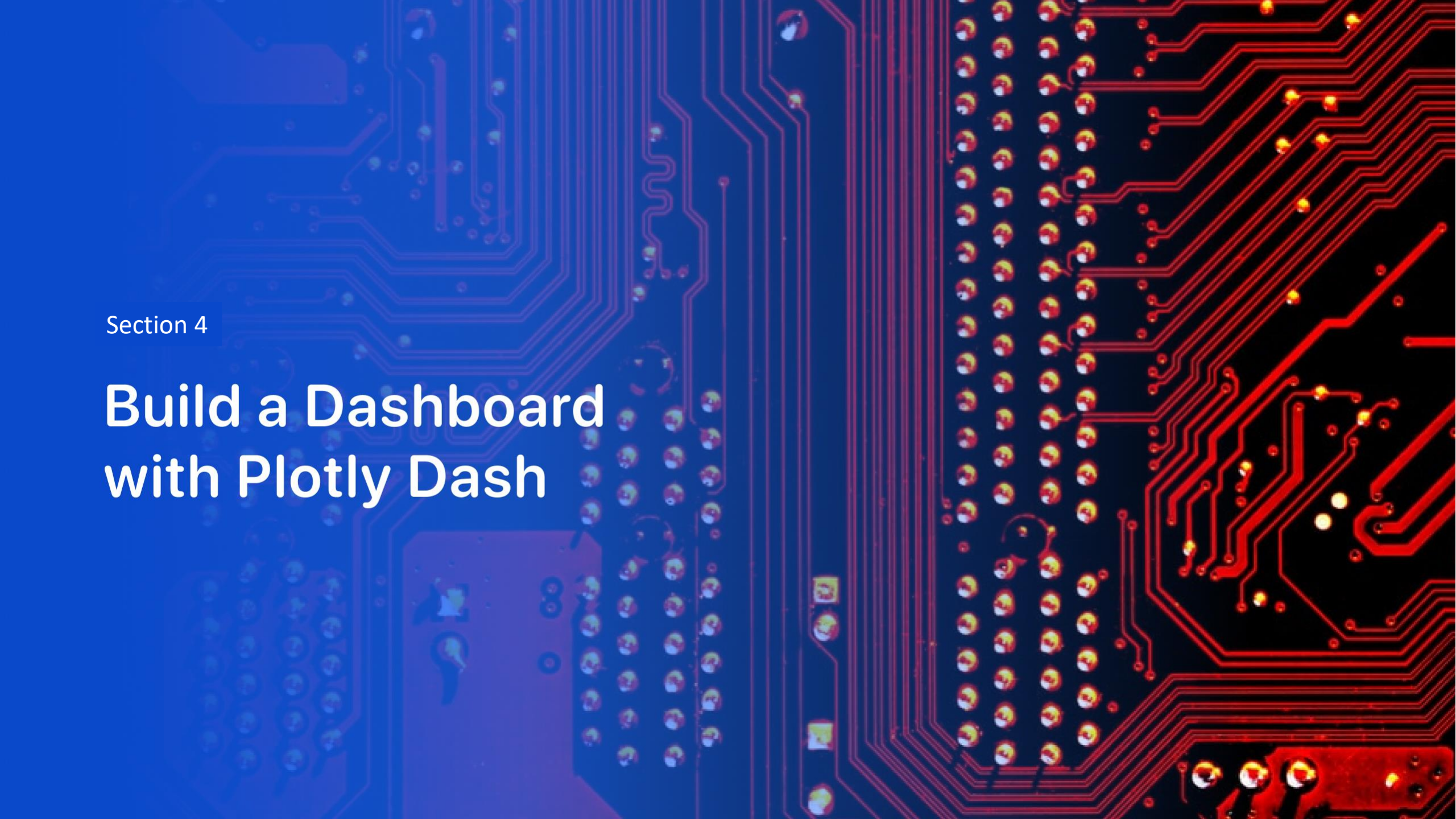


Distances between a launch site to its proximities

CCAFS SLC-40

- 91 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway

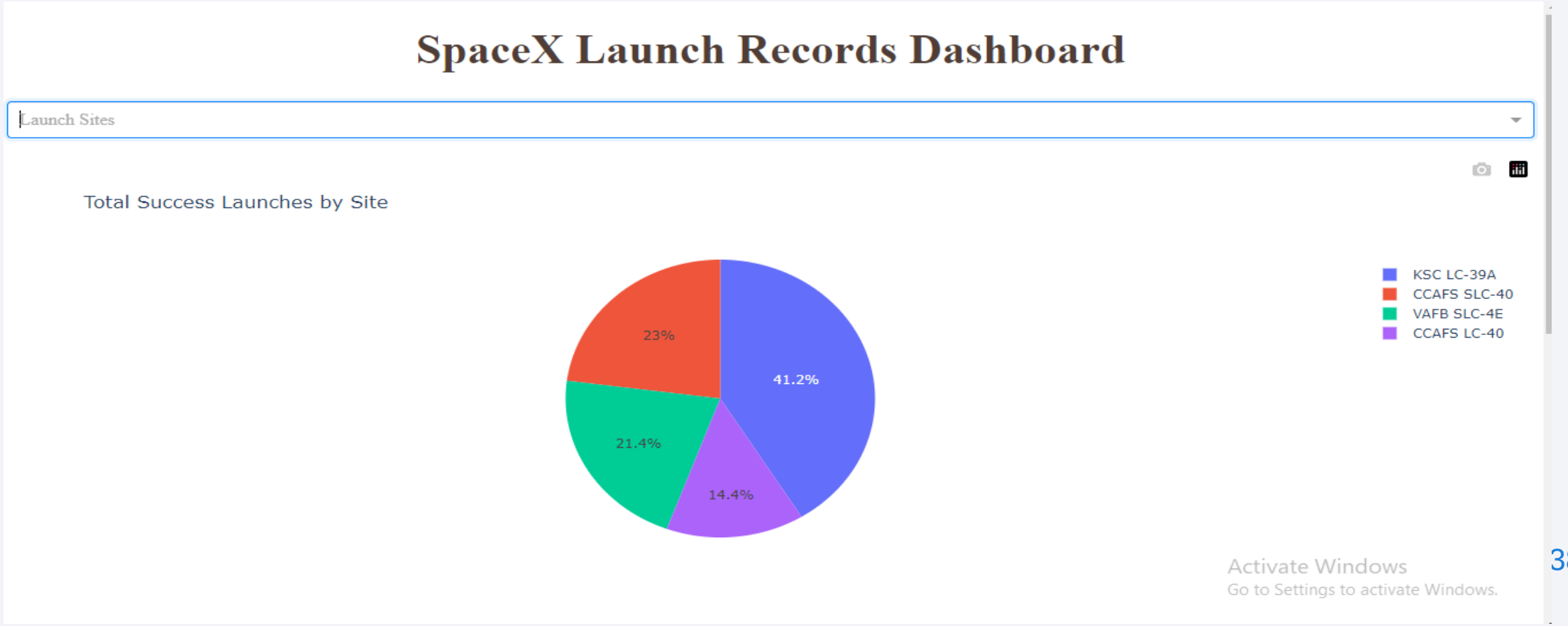



Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site

Success as Percent of Total

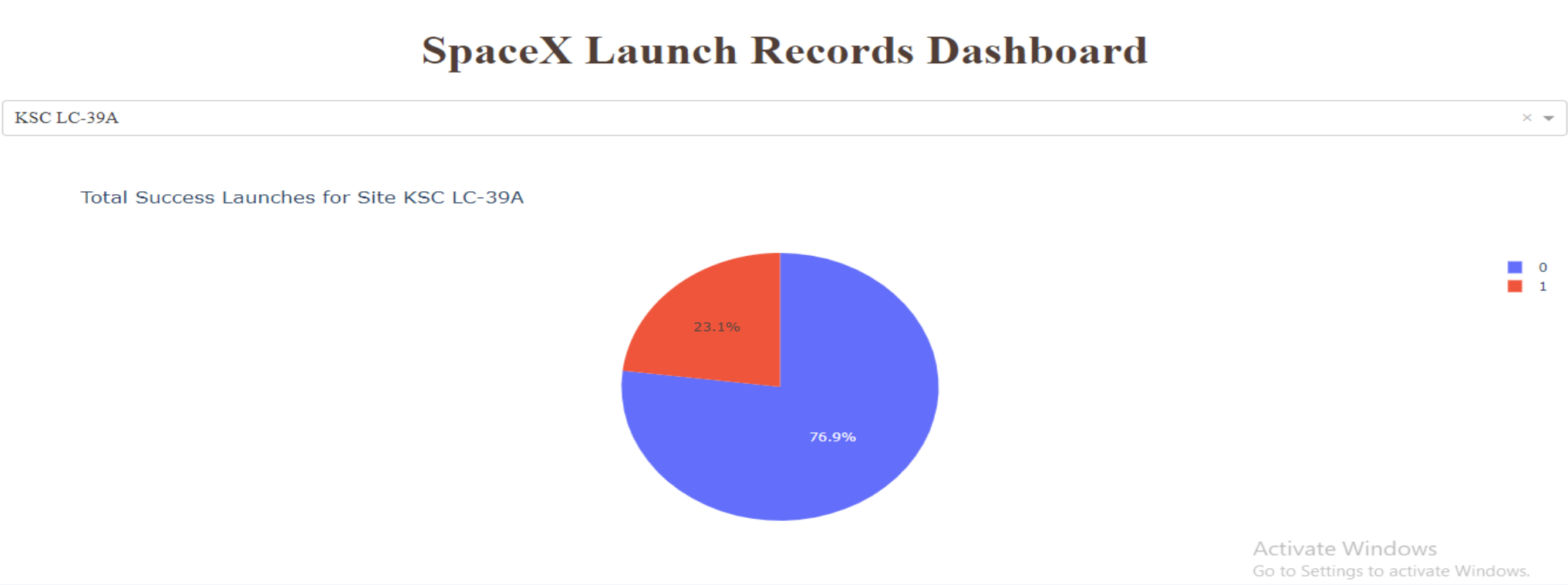
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-39A)

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



Payload Mass and Success

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

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 the rest when looking at best_score

• best_score_ is the average of all cv folds for a single combination of the parameters

TASK 12

Find the method performs best:

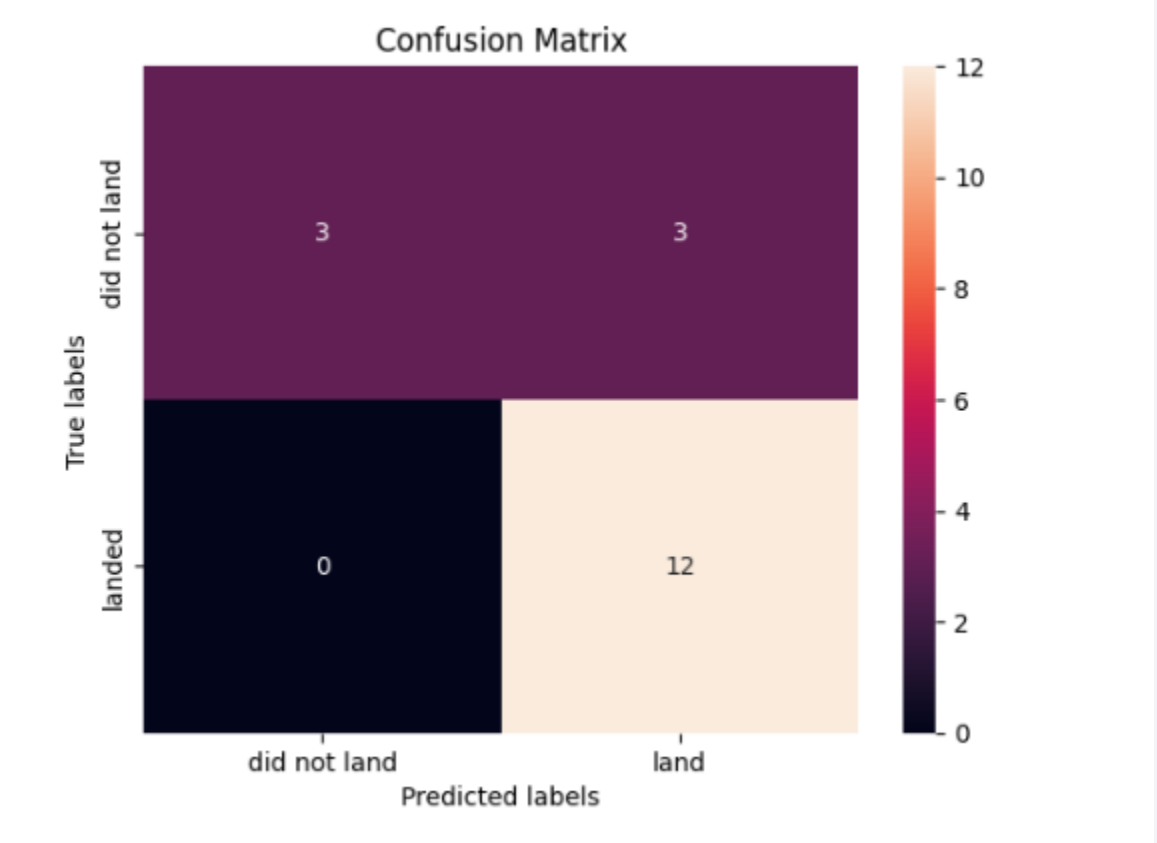
46]:
```
all_model_scores= {'KNeighbors':knn_cv.best_score_,
                   'DecisionTree':tree_cv.best_score_,
                   'LogisticRegression':logreg_cv.best_score_,
                   'SupportVector': svm_cv.best_score_}

algorithm = max(all_model_scores, key=all_model_scores.get)
print('Best model is', bestalgorithm,'with a score of', all_model_scores[algorithm])
if algorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if algorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if algorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if algorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 12, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter':
'random'}
```

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) = .83



Conclusions

Research

- **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

Conclusion

Things to Consider

• **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

Thank you!