IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Milad Gorgani

November 7th, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

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:

1. **Collect** data using SpaceX REST API and web scraping techniques
2. **Wrangle** data to create success/fail outcome variable
3. **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
4. **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
5. **Explore** launch site success rates and proximity to geographical markers
6. **Visualize** the launch sites with the most success and successful payload ranges
7. **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Executive Summary

Summary of all results

- Exploratory Data Analysis:

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

- Predictive Analytics:

All models performed similarly on the test set. The decision tree model slightly outperformed

- Visualization/Analytics:

Most launch sites are near the equator, and all are close to the coast

Introduction

- 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.
- Objectives:
- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings overtime
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

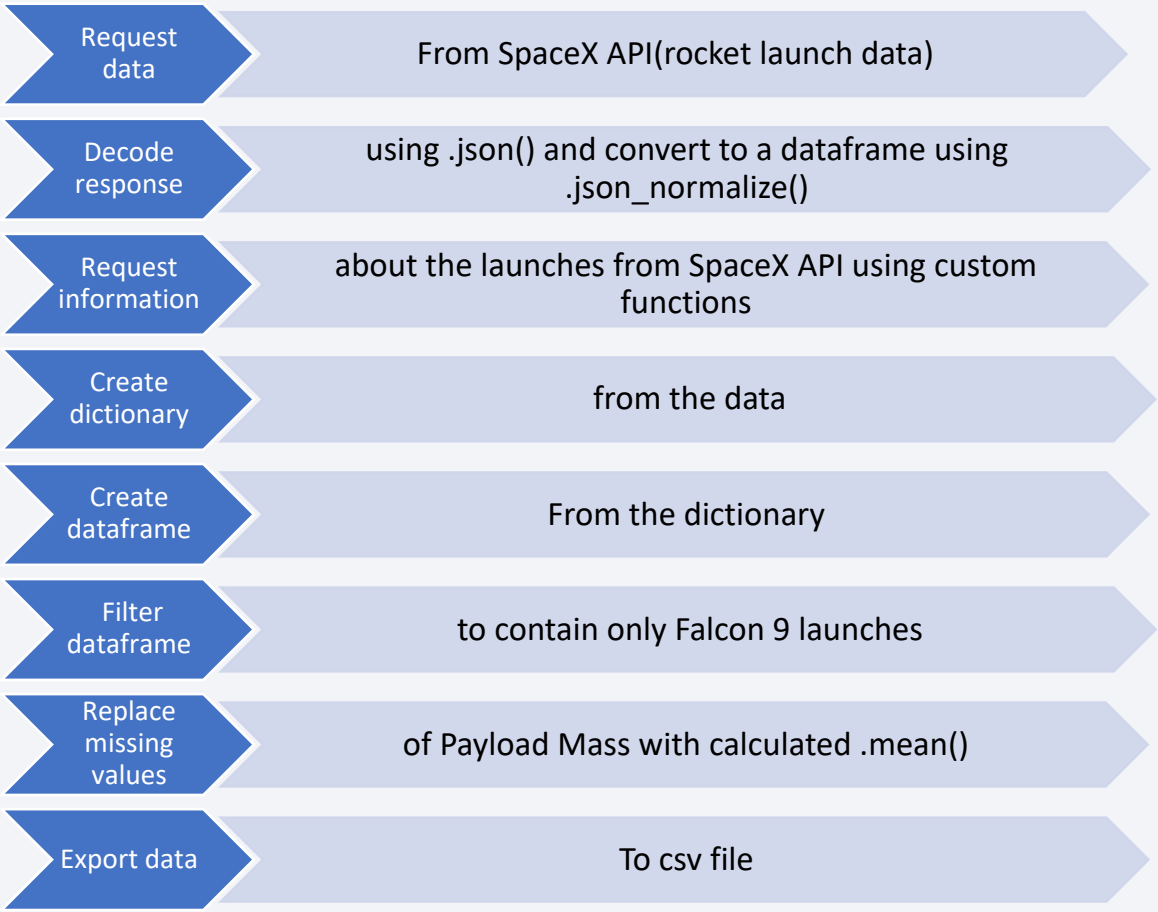
The data was collected using various methods

- Data collection was done using get request to the SpaceX API.
- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- We then cleaned the data, checked for missing values and fill in missing values where necessary.
- In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

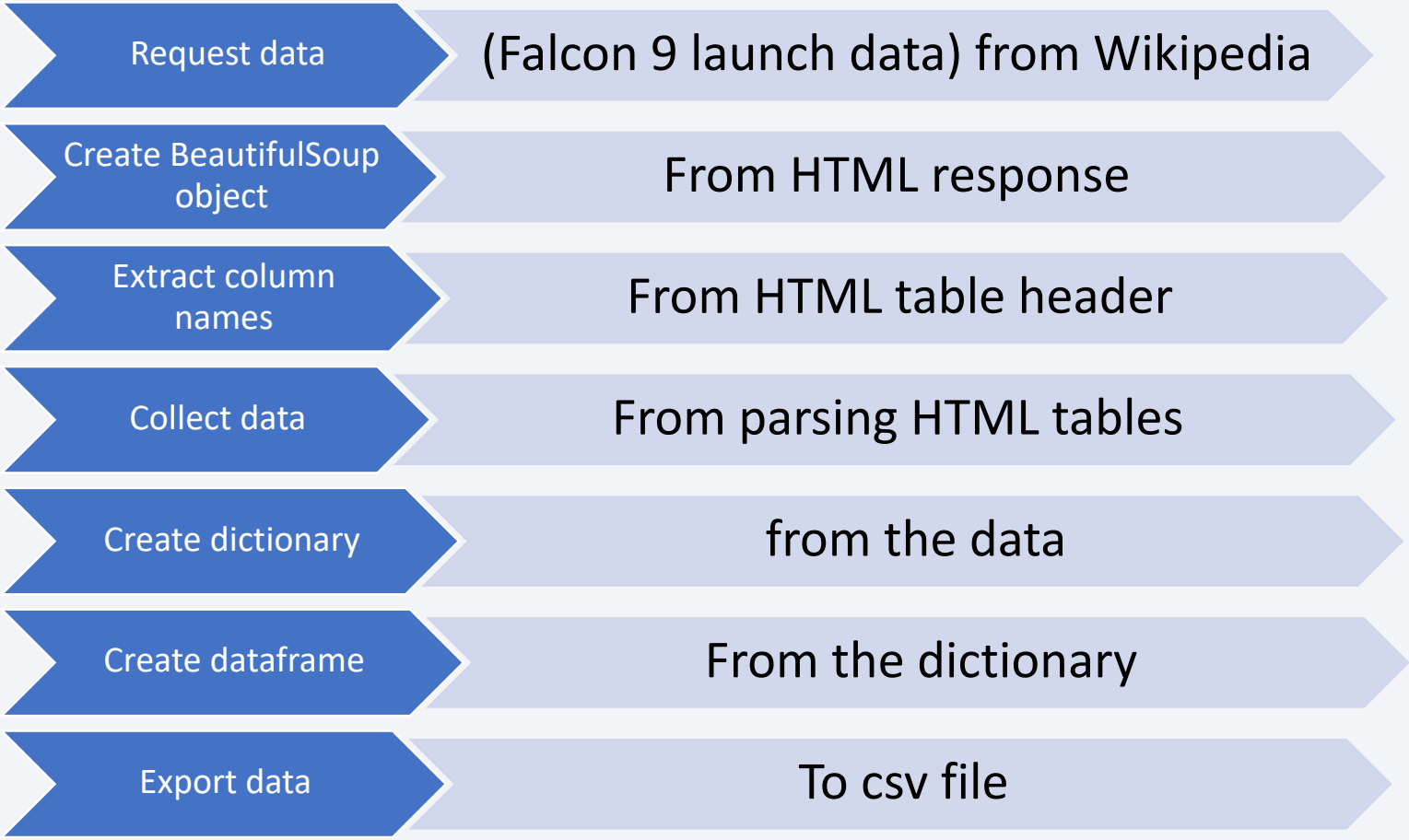
GitHub link



Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.

GitHub Link



Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.

GitHub Link

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. LaunchSite
- 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.

GitHub Link

EDA with SQL

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 F9v1.1.

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 between2010-06-04 and 2017-03-20 (desc)

GitHub Link

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 latitude and longitude coordinates

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 SLC-40 and** its proximity to the **nearest coastline, railway, highway, and city**

GitHub Link

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

GitHub Link

Predictive Analysis (Classification)

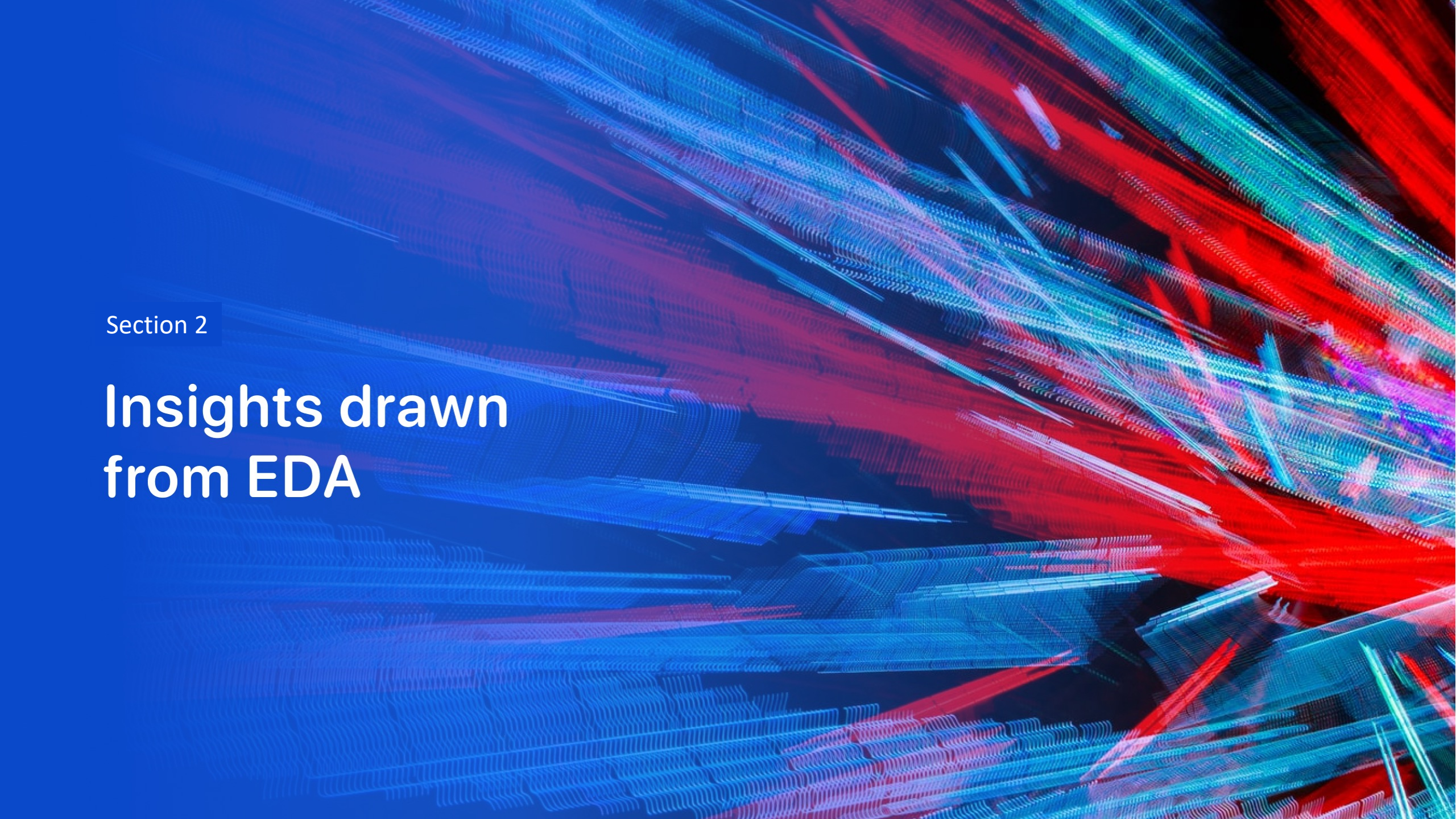
Charts

- **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

GitHub Link

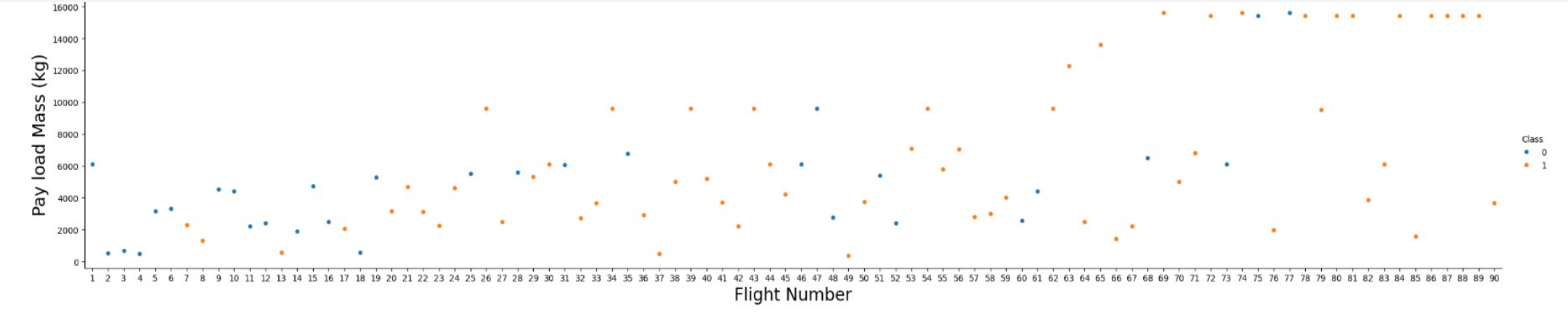
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

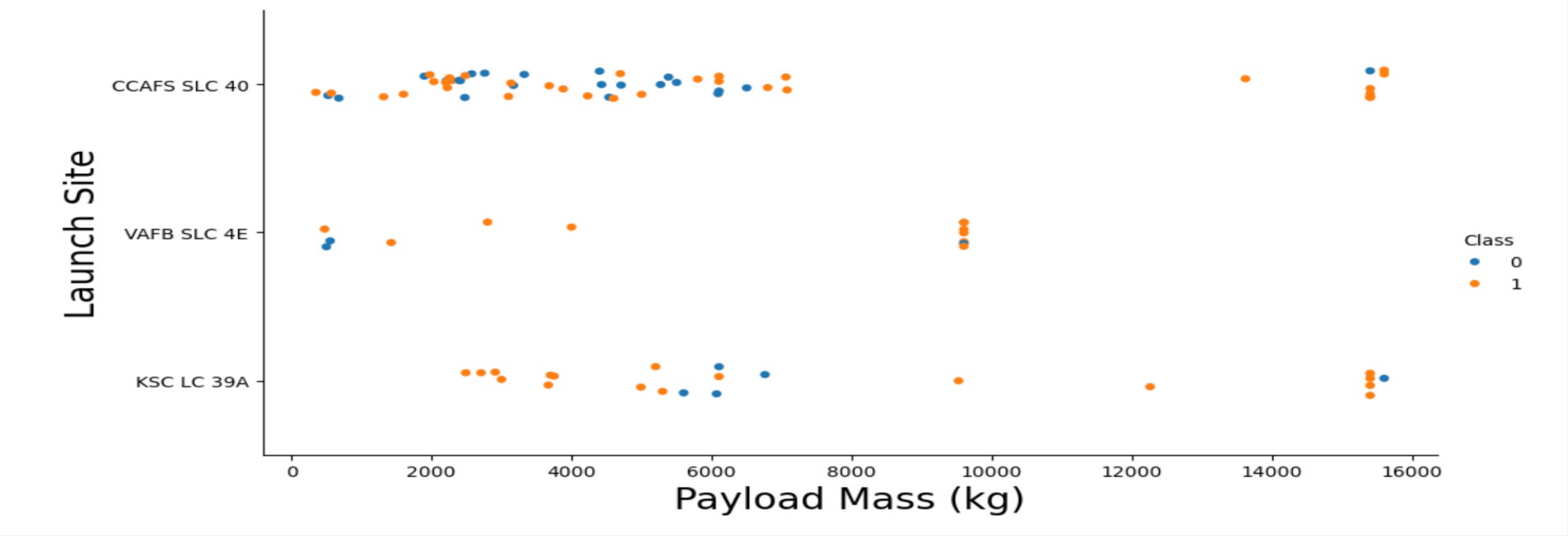
Flight Number vs. Launch Site

- **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



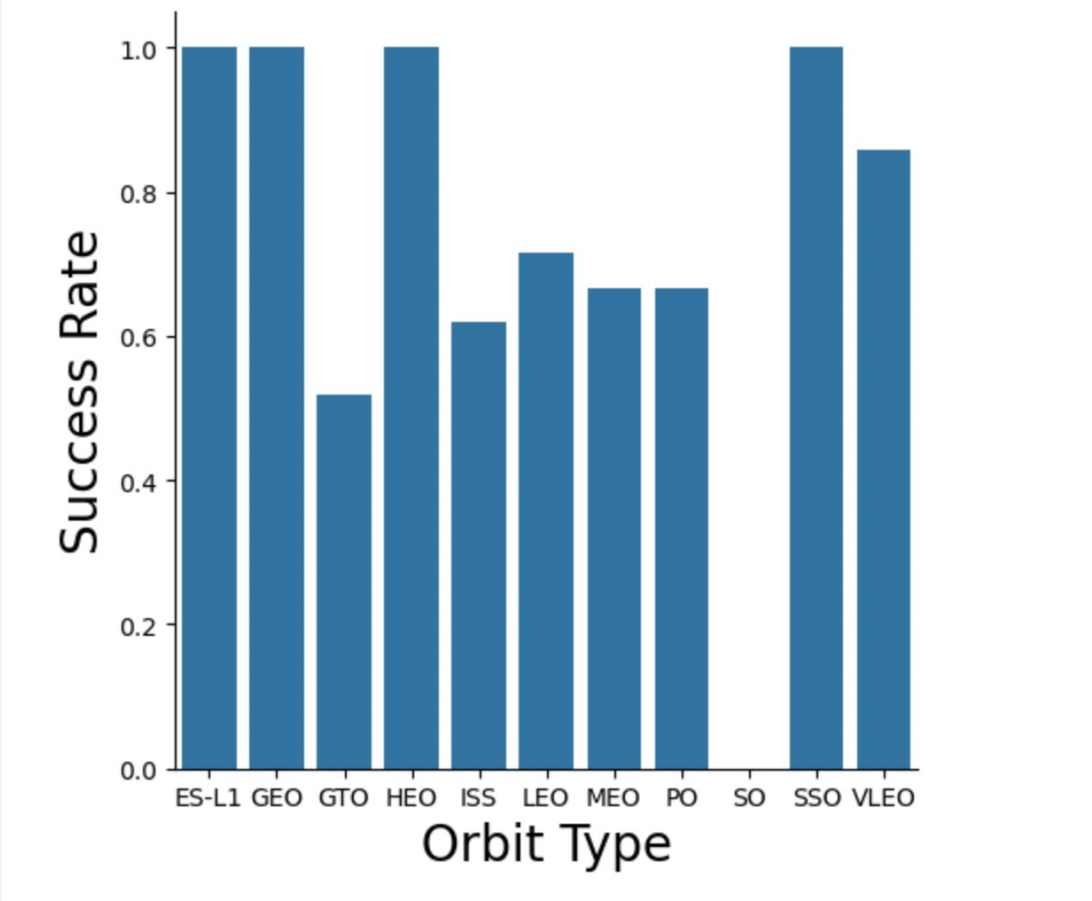
Payload vs. Launch Site

- 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



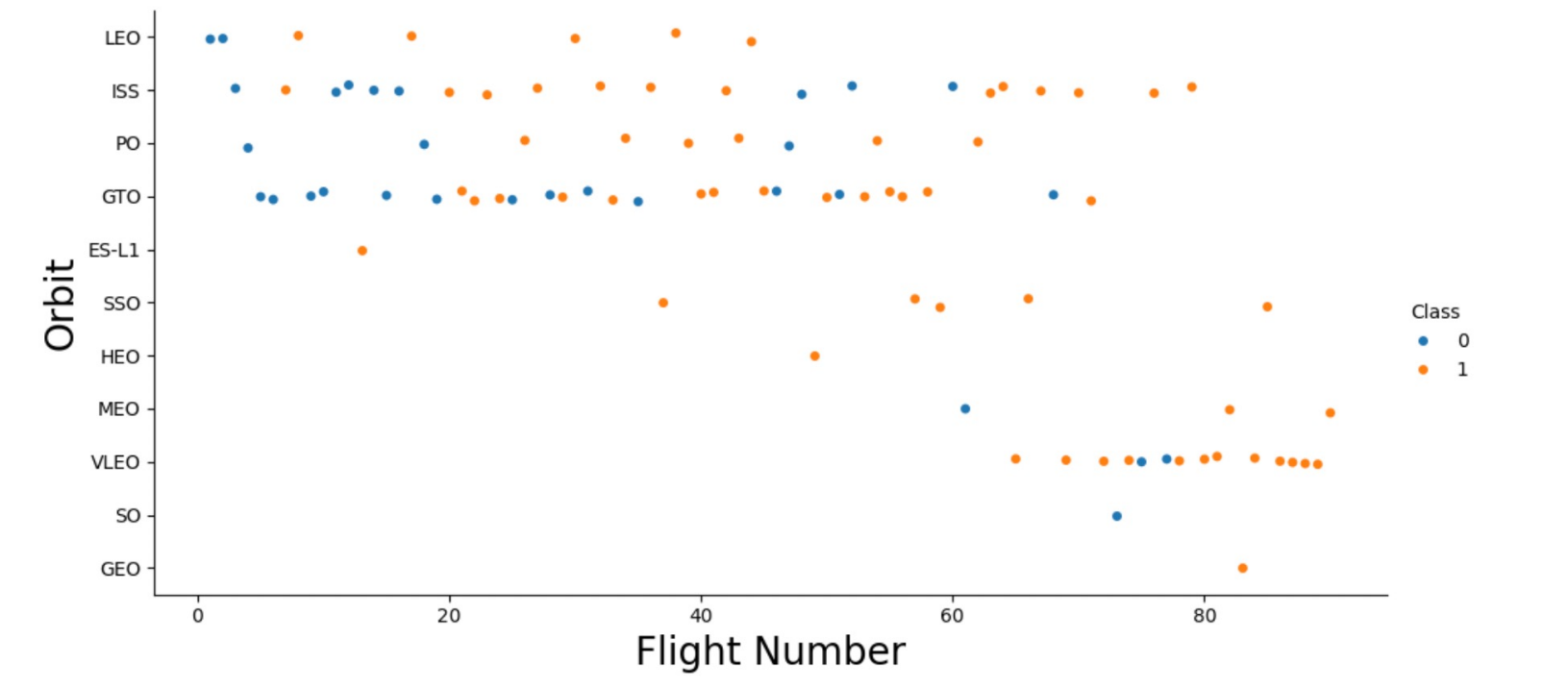
Success Rate vs. Orbit Type

- **100%SuccessRate**:ES-L1,GEO,HEOandSSO
- **50%-80%SuccessRate**:GTO,ISS,LEO,MEO,PO
- **0%SuccessRate**:SO



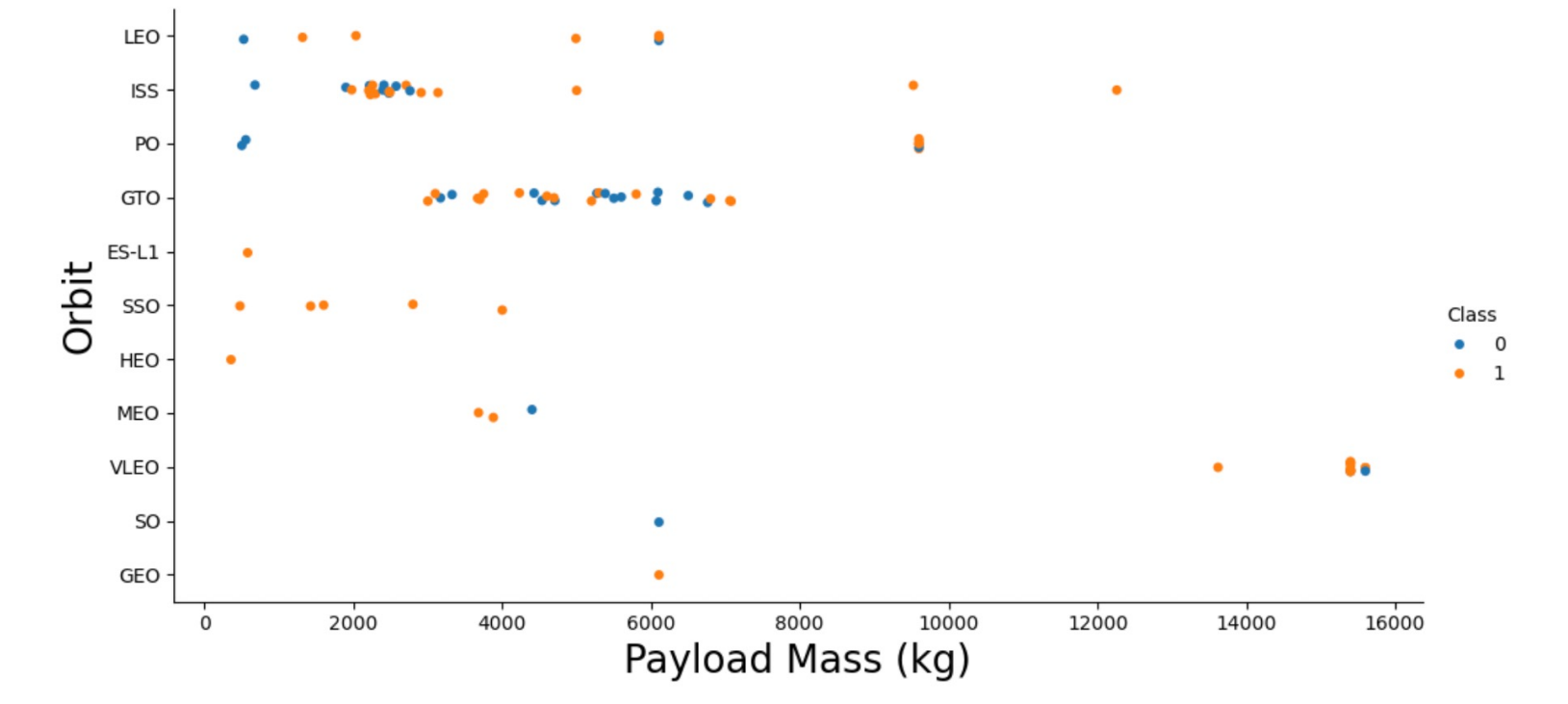
Flight Number vs. Orbit Type

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



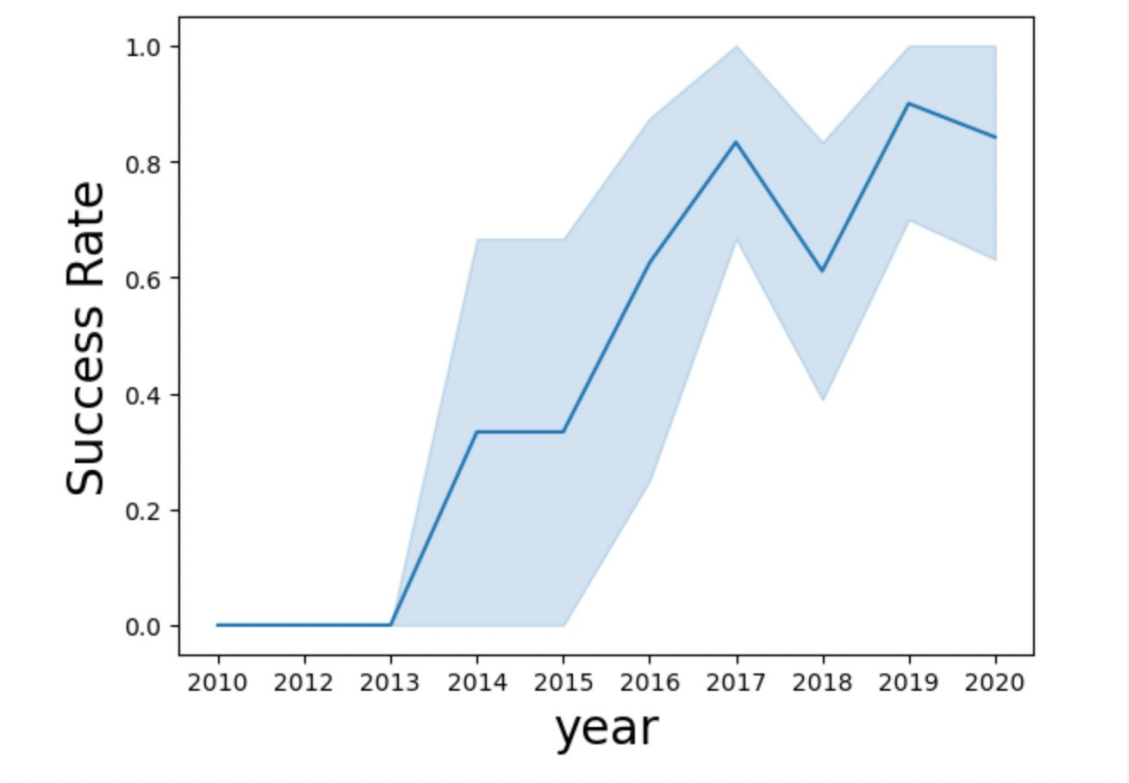
Payload vs. Orbit Type

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



Launch Success Yearly 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



All Launch Site Names

We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

```
8]: %sql SELECT distinct LAUNCH_SITE FROM SPACEXTBL;
 * 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'

We used the query below to display 5 records where launch sites begin 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	Lan
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fail
12/8/2010	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	Fail
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
10/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
%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

We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like '%' to filter for WHERE Mission Outcome was a success or a failure.

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
%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

2015 Launch Records

- We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 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
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome 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;
```

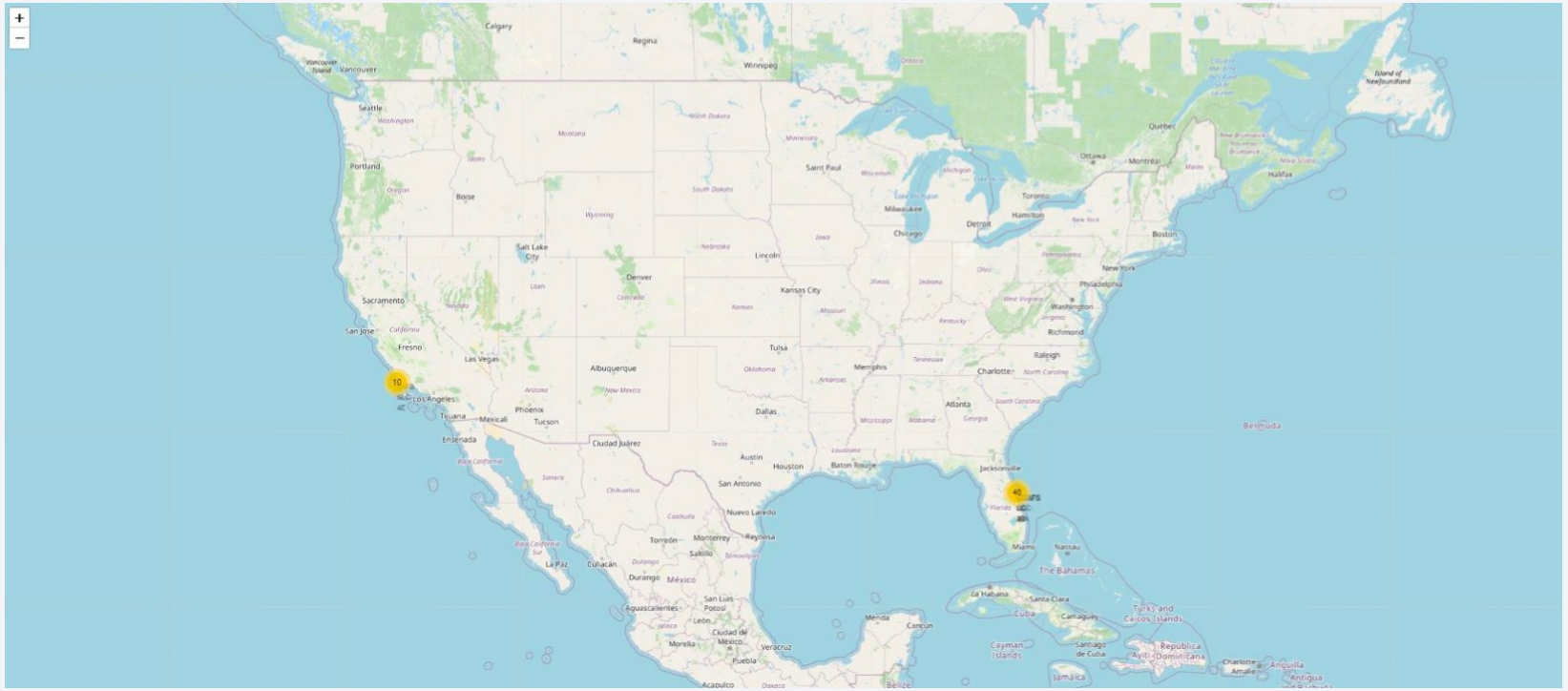
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

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.



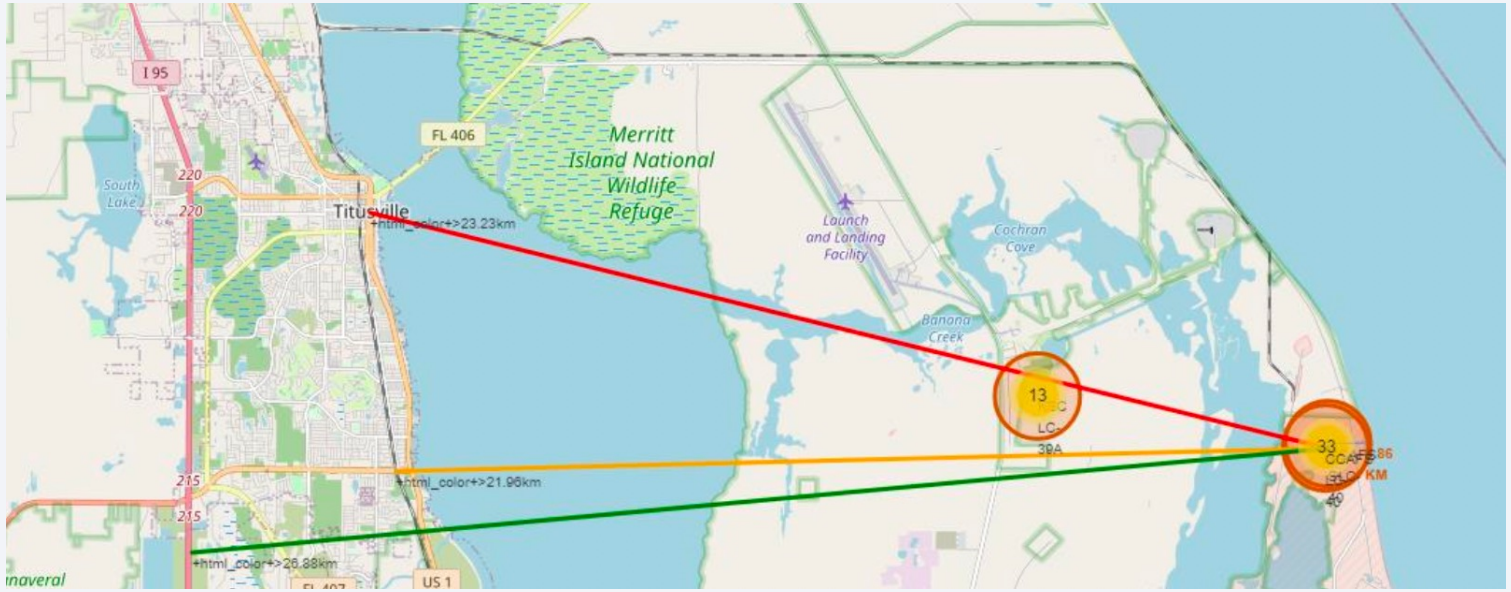
Launch Outcomes

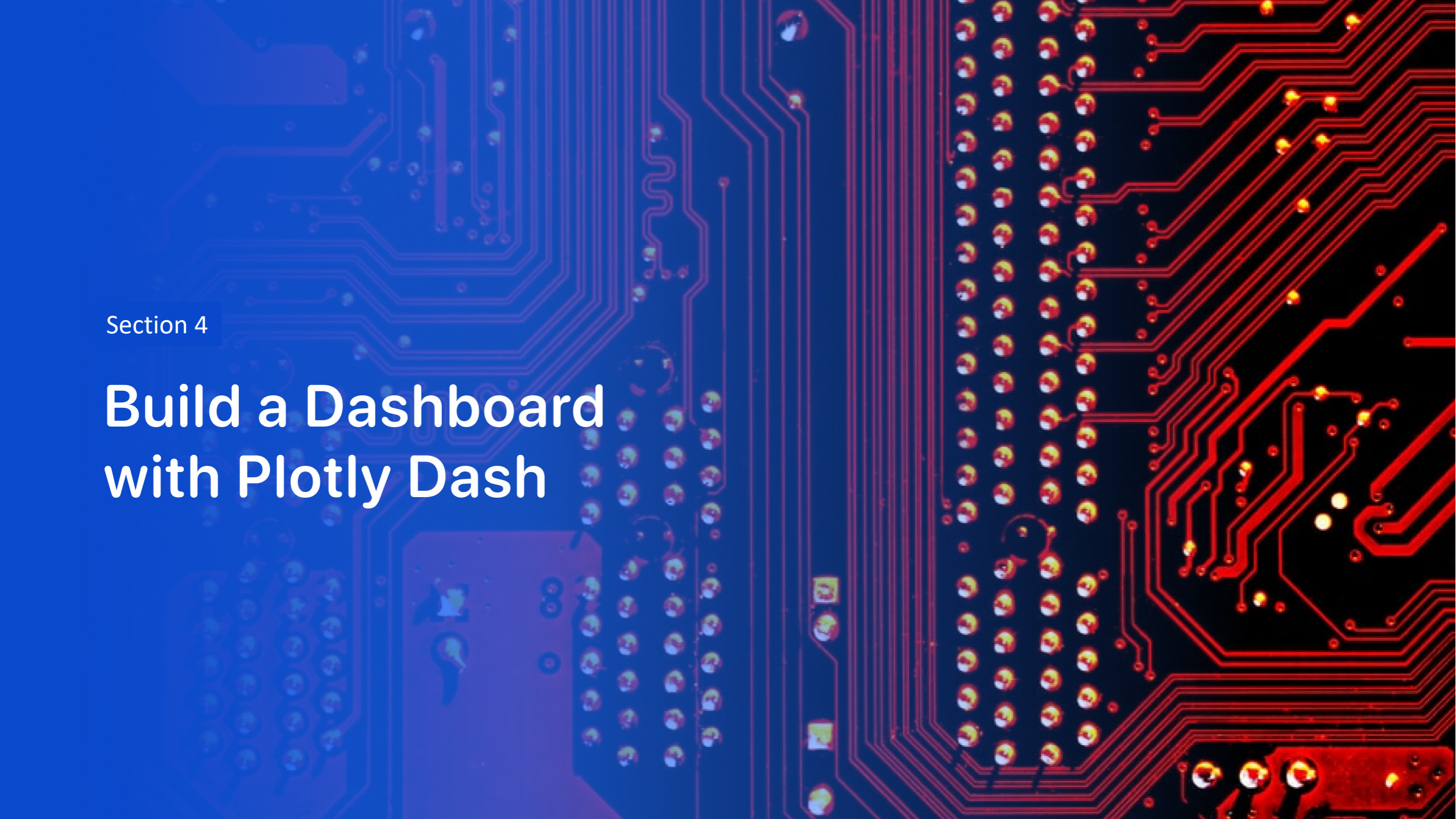
Outcomes:
- **Green** markers for successful launches
- **Red** markers for unsuccessful launches
- Launch site **CCAFSSLC-40** has a **3/7 successrate** (**42.9%**)

Distance to Proximities

CCAFS SLC-40

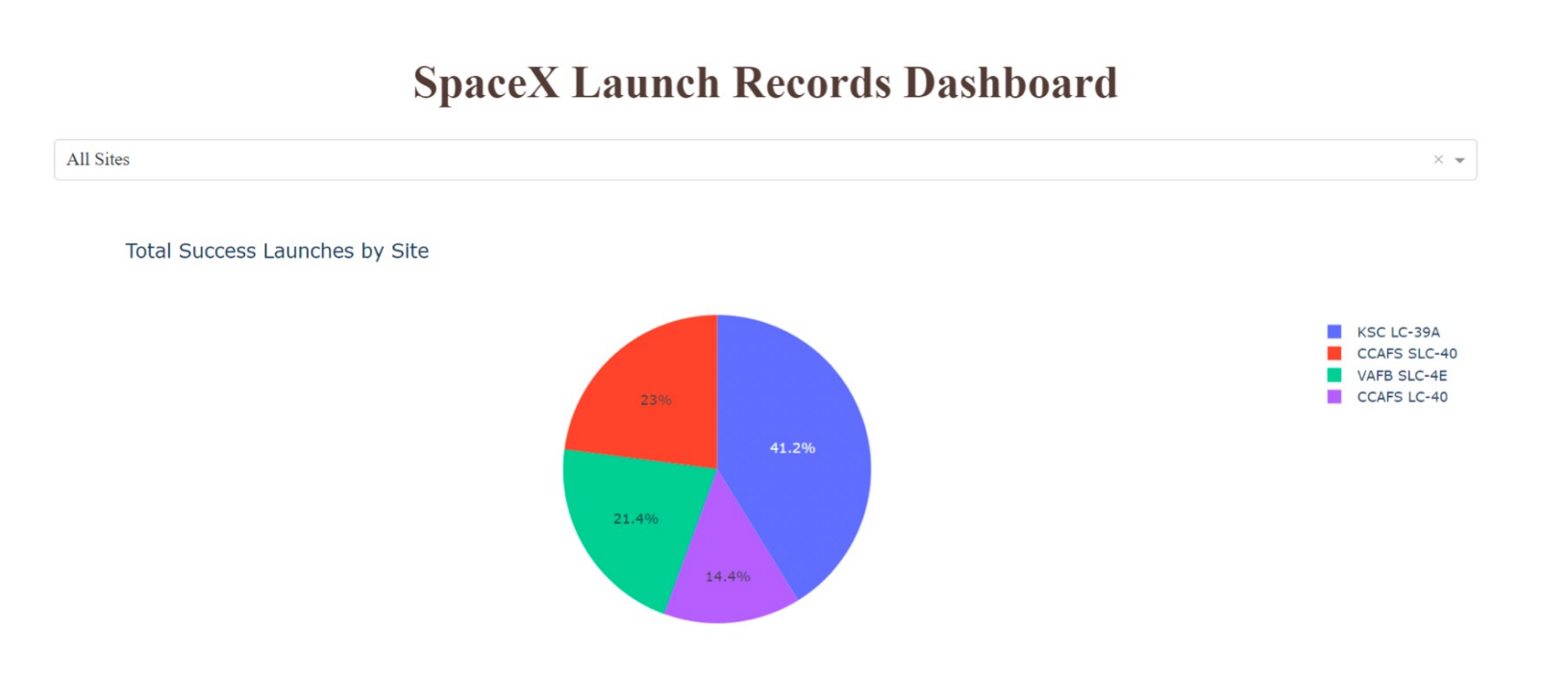
- **.86km** from nearest coastline
- **21.96km** from nearest railway
- **23.23km** from nearest city
- **26.88km** from nearest highway




Section 4
Build a Dashboard with Plotly Dash

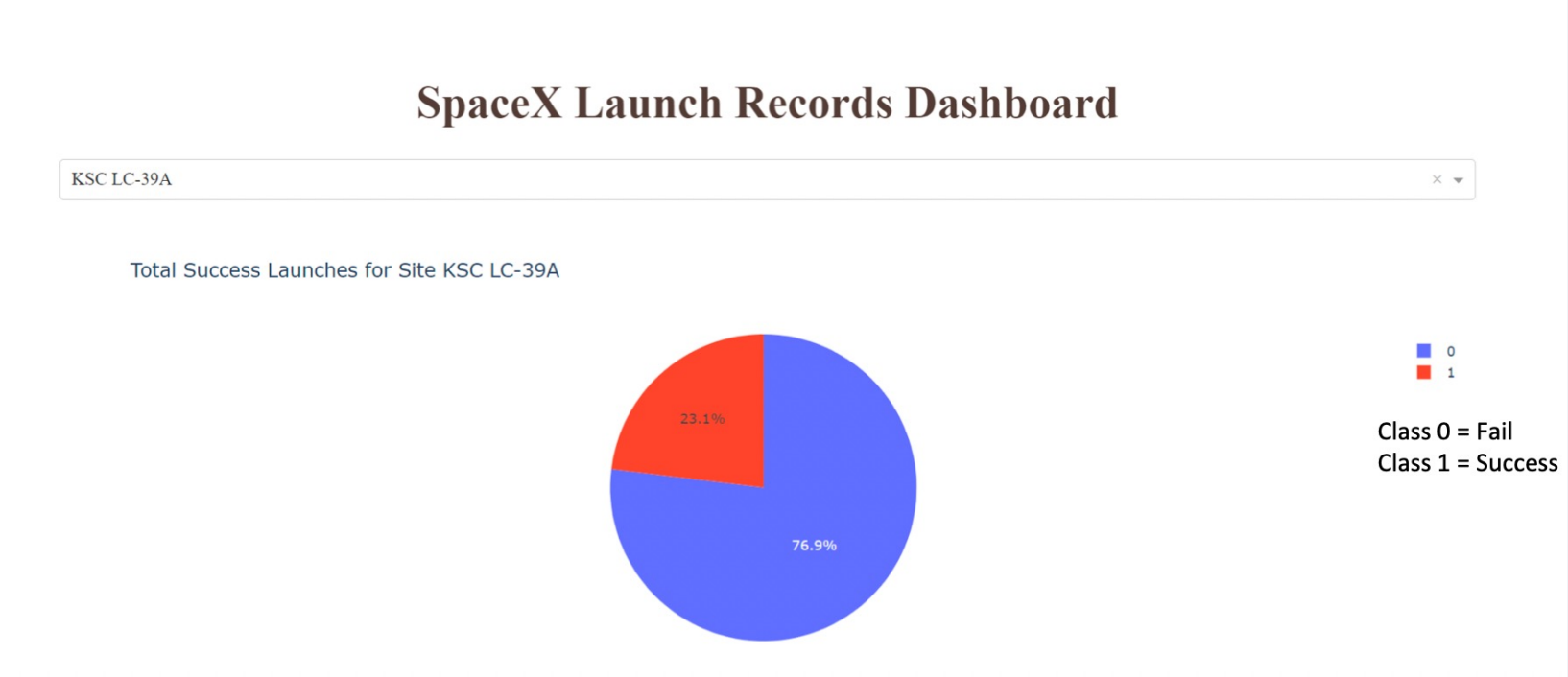
Launch Success by Site

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



Launch Success (KSC LC-29A)

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



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

Section 5

Predictive Analysis (Classification)

Classification 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

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

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

```
Best model is DecisionTree with a score of 0.9017857142857144
Best params is : {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples
_split': 2, 'splitter': 'best'}
```

Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm

ConfusionMatrixOutputs:

- 12 True positive
- 3 True negative
- **3 False positive**
- 0 False Negative

Precision = TP / (TP + FP)

• 12 / 15 = .80

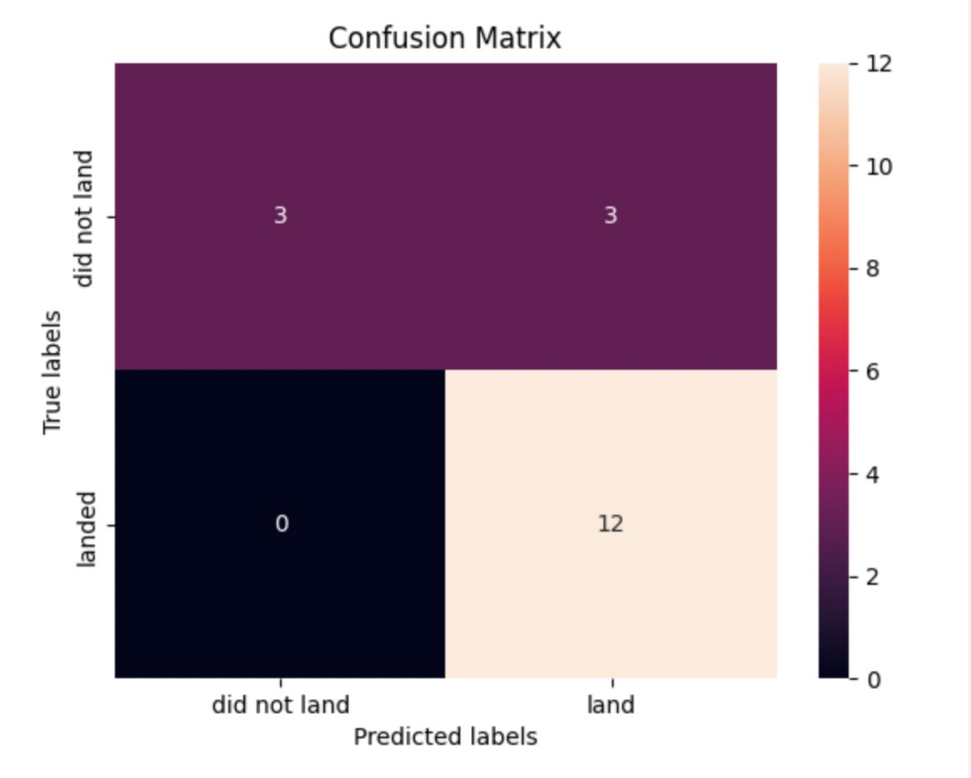
Recall = TP / (TP + FN)

• 12 / 12 = 1

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

• 2 * (.8 * 1) / (.8 + 1) = .89

Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



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,andSSOhavea100%successrate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!