IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahamadou Barthe
March 21st 2024

Table of contents

Executive Summary

- Methodology:
  - Data Collection through SpaceX REST API and Web Scarping
  - Data Wrangling to create success/fail outcome variables
  - Exploratory Data Analysis with Visualization techniques factors such as: Payload, Launch Site, Flight Number and Yearly trend.
  - Exploratory Data Analysis with SQL to calculate Total Payload, Payload range for successful launch, and total number of successful and failed outcomes.
  - Interactive Visual Analytics with Folium for Launch Site success rates and proximity to geographical markers
  - Interactive Dashboard with Plotly Dash for Launch Site with the most success and successful Payload ranges
  - Machine Learning Predictive Analysis using Logistic Regression, SVM, Decision Tree and KNN

Executive Summary

- Results:
  - Exploratory Data Analysis:
    - Launch success rates have 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/Analysis:
    - Most Launch Sites are near the equator and are also close to the coast
  - Predictive Analytics:
    - All the models performed similarly on the test set. However, the Decision Tree model provided a more accurate result.

Introduction

- Project background:

  Space X 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 Space X can reuse the first stage. Therefore, if we can determine whether the launch will be a successful one or if the launch will fail, whether the the first stage will land or not, we can determine the cost of a launch. The goal of the project predict if the SpaceX can reuse the first stage by using public data and machine learning models.

- Problems to solve:

  - How payload mass, launch site, number of flights, and orbits affect first-stage landing outcome.
  - What is the rate of successful landings over time.
  - What is the best predictive model for successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection:
  - Data was collected using SpaceX REST API and web scraping.
- Data wrangling
  - Filter the data, handle missing values and apply One-hot encoding to categorical features.
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive Visual Analytics using Folium and Plotly Dash
- Classification model building to predict landing outcomes.
  - Tune and evaluate models to find best model and parameters

Data Collection – REST API

- Steps:
  - Request data from SpaceX REST API (rocket launch data)
  - Decode response using .json() and convert to a dataframe using .json_normalize()
  - Resquest 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 the Payload Mass with calculated .mean()
  - Export data to csv file

Data Collection – Web Scraping

- Steps:
  - Request data (Falcon 9 launch data) from Wikipedia
  - Create BeautifulSoup object from HTML response
  - Extract columns names from HTML table header
  - Collect data from parsing HTML tables
  - Create dictionary from data
  - Create dataframe from dictionary
  - Export data to csv file

Data Wrangling

- Steps:
  - Perform EDA and determine data labels
  - Calculate:
    - Number of launches for each site
    - Number and occurrence of orbit
    - Number and occurrence of mission outcome per orbit type
  - Create binary landing outcome column (dependent variable)
  - Export data to csv file

EDA with 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 measured value.

EDA with 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 the first successful landing on ground pad
  - Names of boosters which had success landing on drone ship and have 4000 < payload mass < 6000
  - 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 the months in the year 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-04 (desc)

Maps with Folium

- Markers Indicating Launch Sites
  - Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using the latitude and longitude coordinates
  - Added red circles at all launch sites coordinates with a popup label showing its name using the latitude and longitude coordinates
- Colored markers of launch outcomes
  - Added colored markers of successful (green) and failure (red) launches at each launch site to show which launch site 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.

Dashboard with Plotly Dash

- Dropdown list with launch sites

  Allow users to select all launch sites or a certain launch site

- Pie chart showing successful launches

  Allow users to see successful and failed launches percentages

- Slider of payload mass range

  Allow users to select payload mass range

Scatter chart showing payload mass vs success rate by booster version

  Allow users to see the correlation between payload and launch success

Predictive Analysis

- 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 GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (log_reg()), support vector machine (SVM()), decision tree (tree()) and K-nearest neighbor (knn()).
- Calculate accuracy on the data using .score() for all models
- Assess the confusion matrix for all models
- Idendify the best model using Jaccard_Score, F1_Score and Accuracy

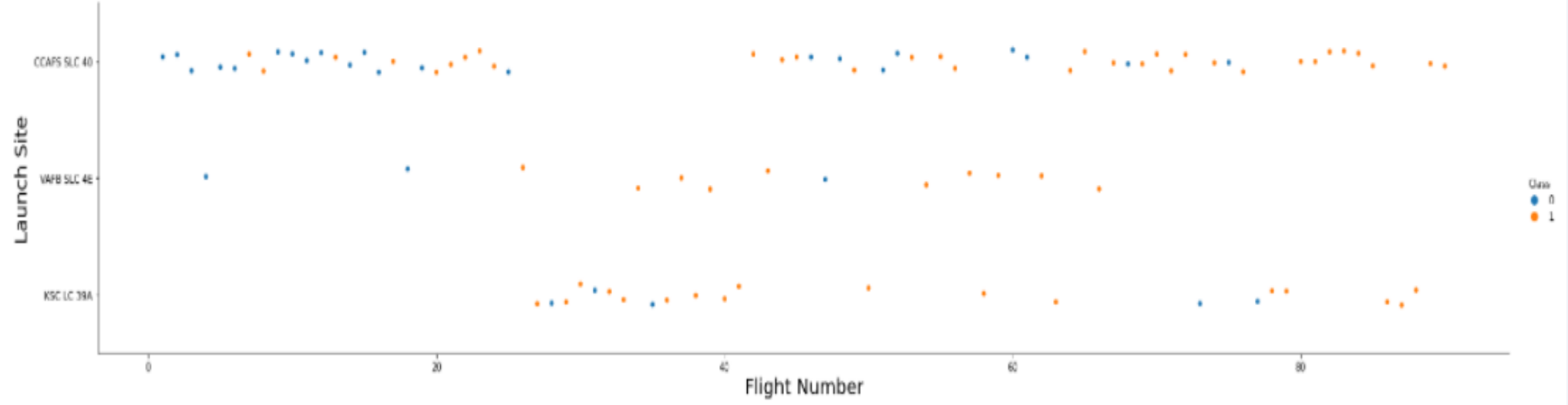
Results

Section 2

Insights drawn from EDA

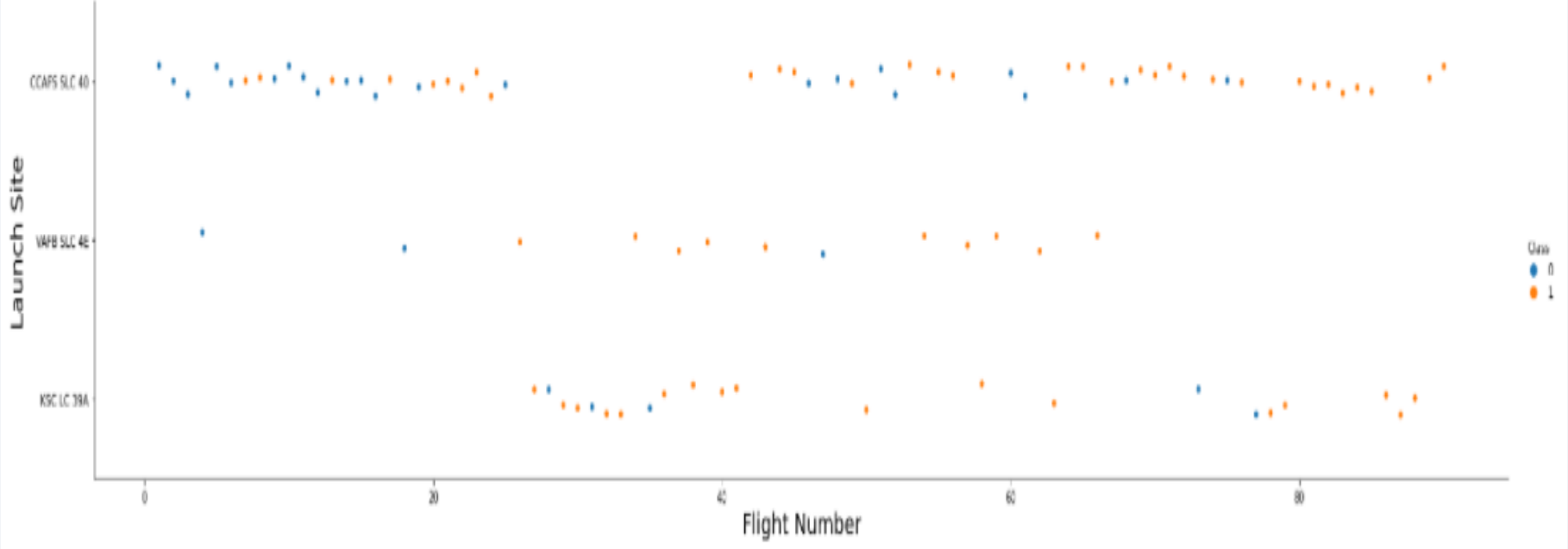
Flight Number vs. Launch Site

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



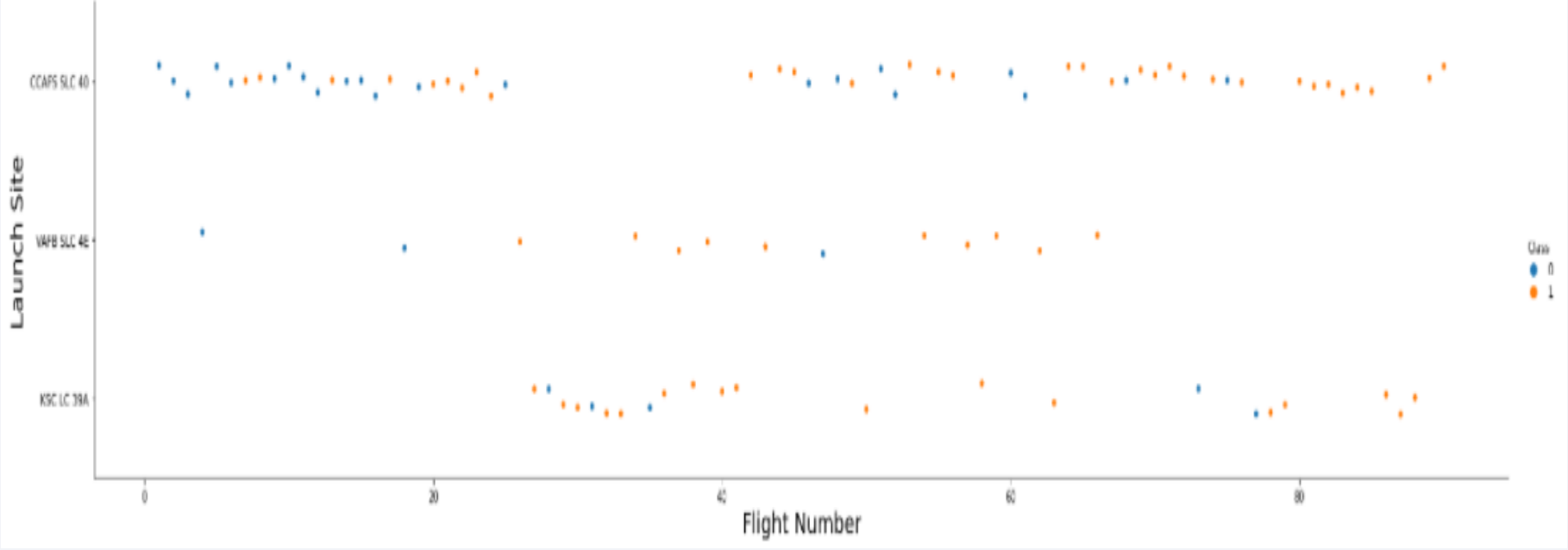
Payload vs. Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches 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



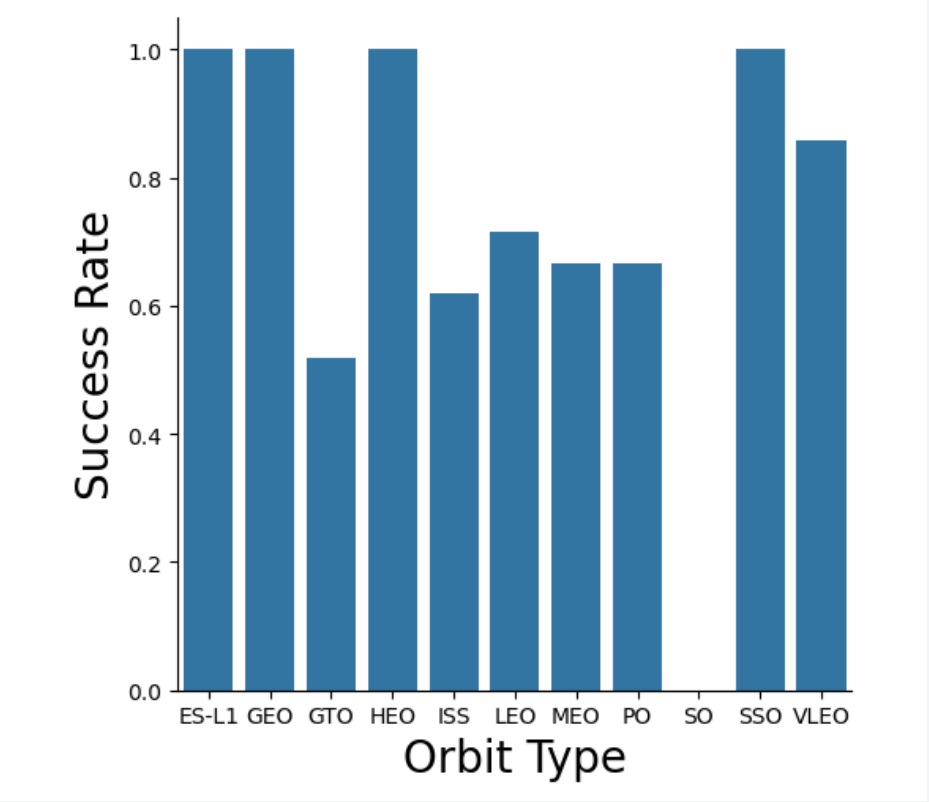
Success Rate vs. Orbit Type

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches 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



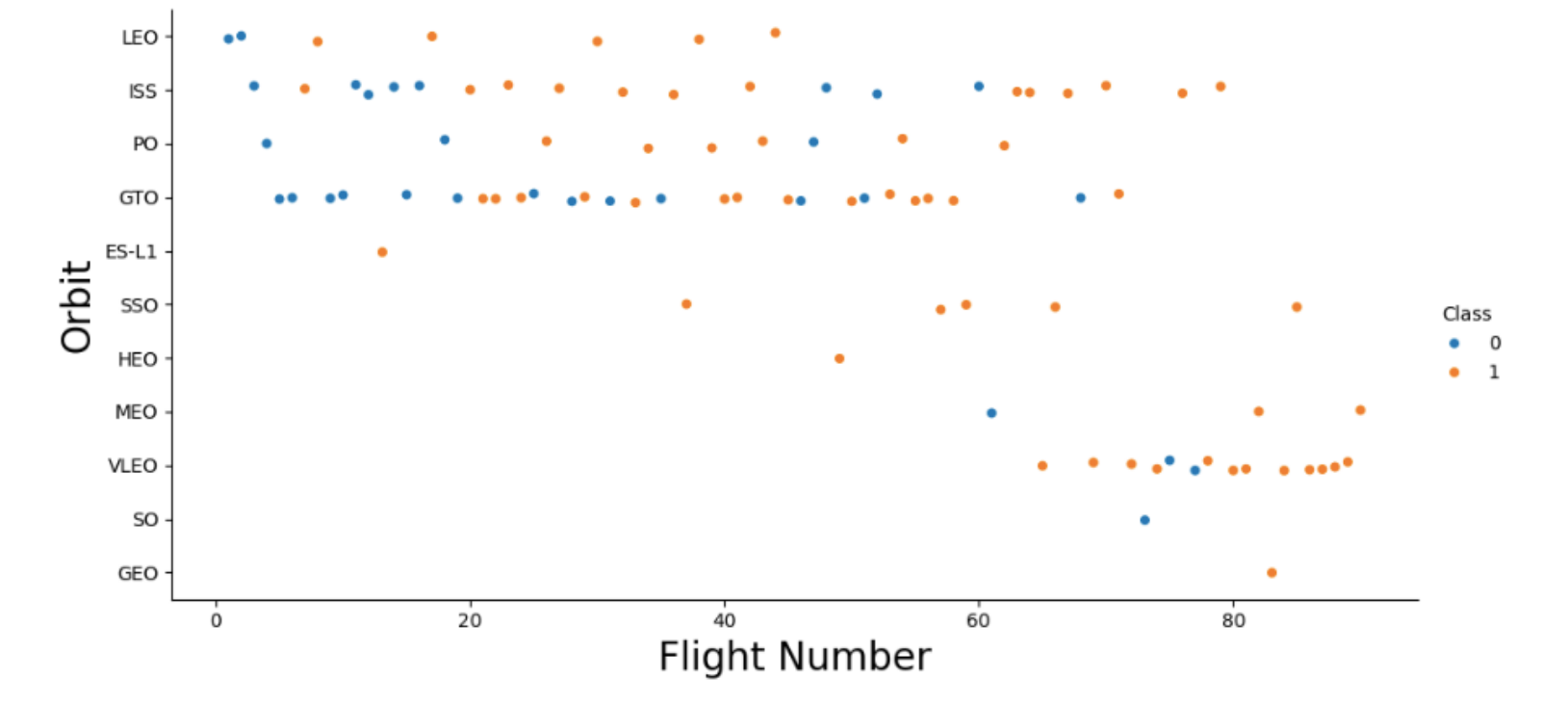
Flight Number vs. Orbit Type

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



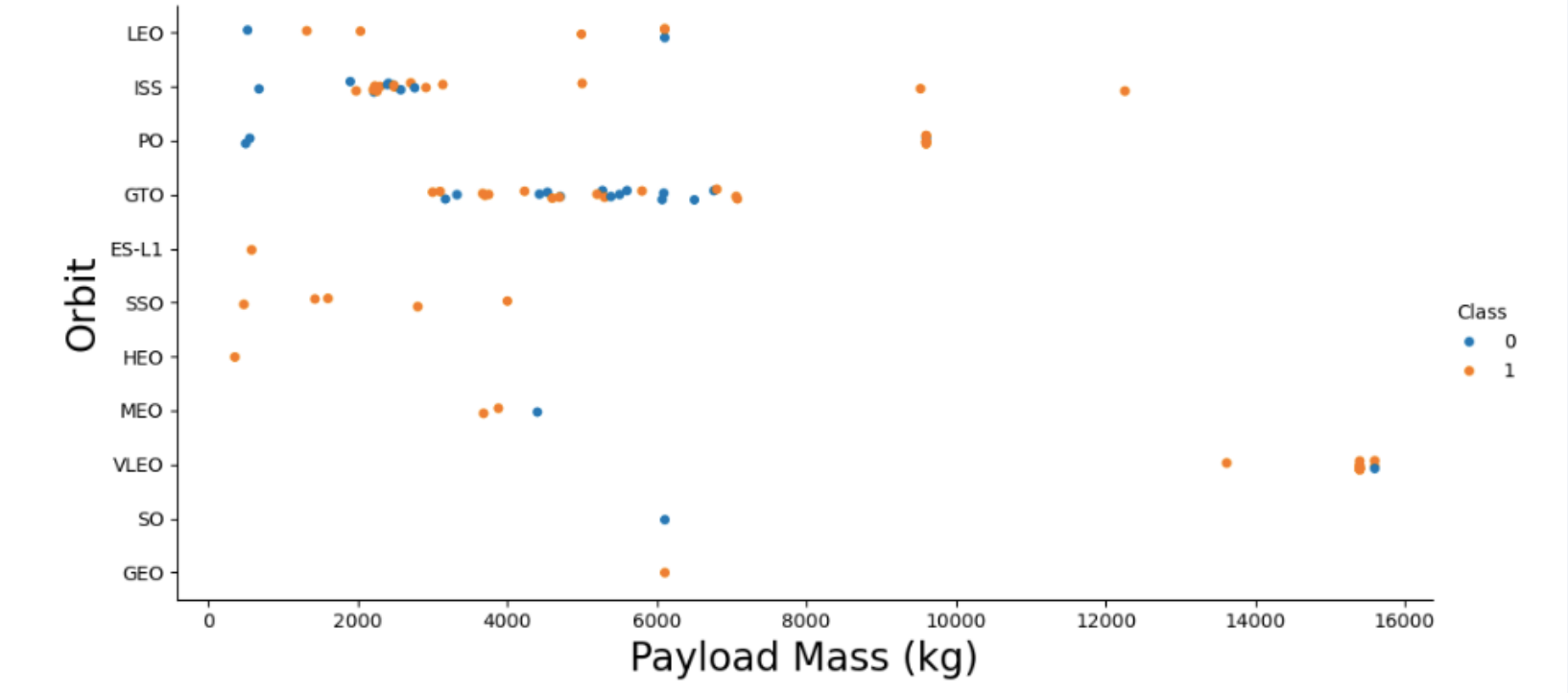
Flight Number vs. Orbit

- 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



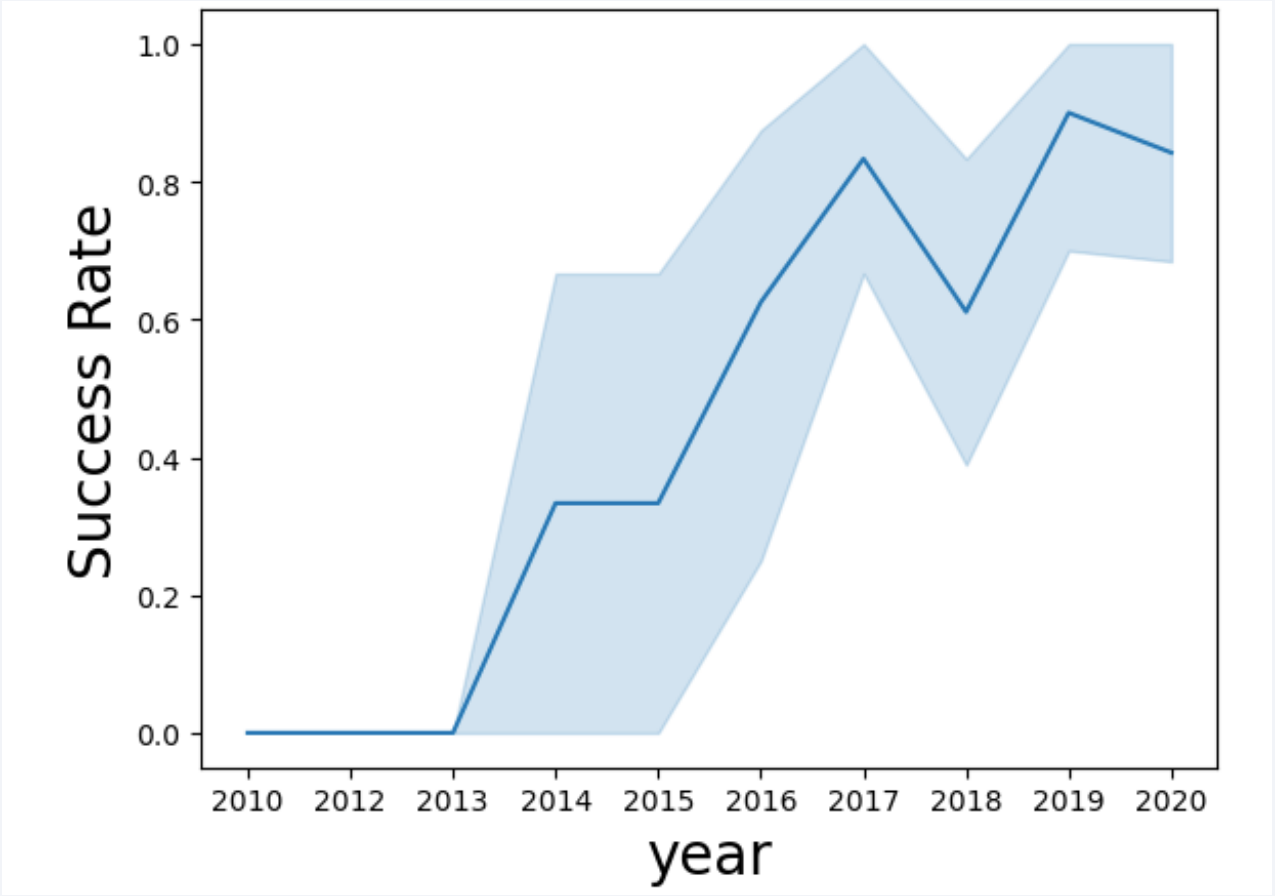
Payload vs. Orbit

- 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



Launch Site Information

Launch Site Names:

```
In [9]:
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Out[9]:

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

Records with Launch Sites starting with 'CCA' (displaying 5)

```
In [14]:
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Out[14]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Payload Mass

Total Payload Mass: 45, 596 kg (total) carried by boosters launched by NASA (CRS)

```
In [16]: %%sql
         SELECT SUM(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Customer = 'NASA (CRS)';

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

Out[16]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass: 340 kg (average) carried by booster version F9 v1.1

```
In [17]: %%sql
         SELECT AVG(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Booster_Version LIKE 'F9 v1.0%';

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

Out[17]:

AVG(PAYLOAD_MASS__KG_)
340.4

Landing & Mission Info

First Successful Landing on Ground Pad:12-22-2015

```
In [19]: %%sql
         SELECT MIN(Date)
         FROM SPACEXTBL
         WHERE Landing_Outcome = 'Success (ground pad)';

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

Out[19]:

MIN(Date)
2015-12-22

Booster Landed on Drone Ship

```
In [22]: %%sql
         SELECT BOOSTER_VERSION
         FROM SPACEXTBL
         WHERE LANDING_OUTCOME = 'Success (drone ship)' AND 4000 < PAYLOAD_MASS__KG_ <6000;

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

Out[22]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Landing & Mission Info

Total Number of Successful and Failed Mission Outcomes:

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

In [24]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Out[24]:

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

Boosters

Boosters carrying max payload

In [31]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

Out[31]:

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

Failed Landings on Drone Ship

In 2015 (Showing month, date, booster version, launch site and landing outcome)

In [46]:
```
%%sql
SELECT substr(DATE,6,2) AS Month, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015';
```

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

Out[46]:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Count of Successful Landings

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

In [44]:

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

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

Out[44]:

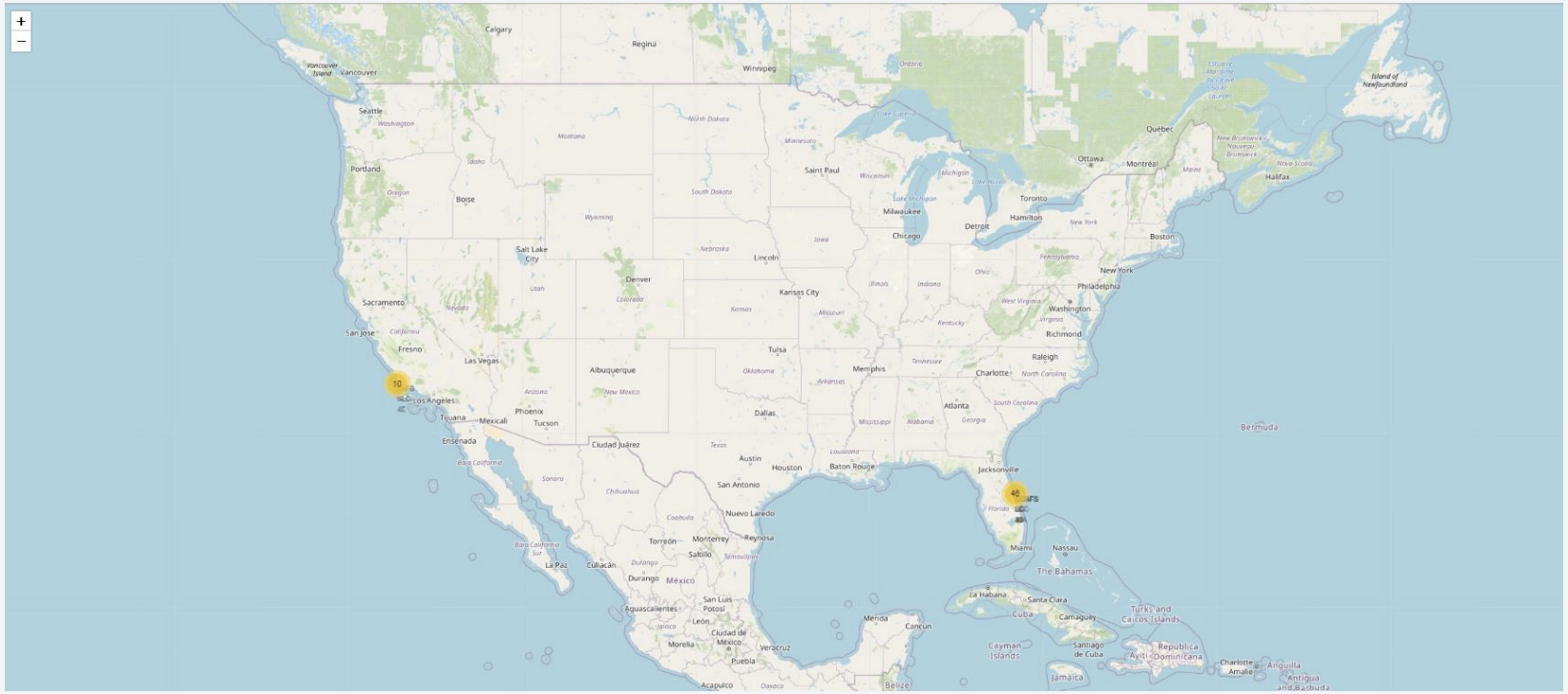
Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

Launch Sites

Near Equator: the closer the launch site is to the equator, the easier it is to launch to equatorial orbit, and the more help we get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boot due to the rotational speed of the earth. That helps in reducing the cost of the launch



Launch Outcomes

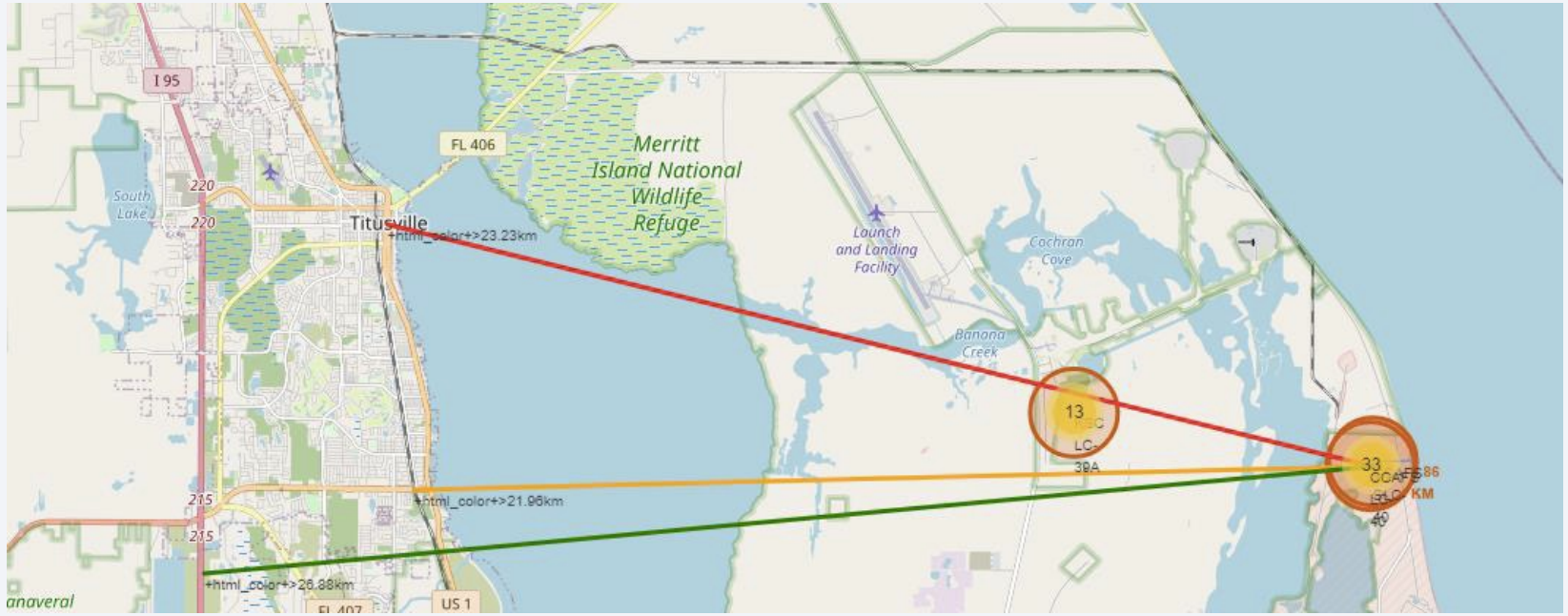
Outcomes:

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

Distance to Proximities

CCAFS SLC-40:

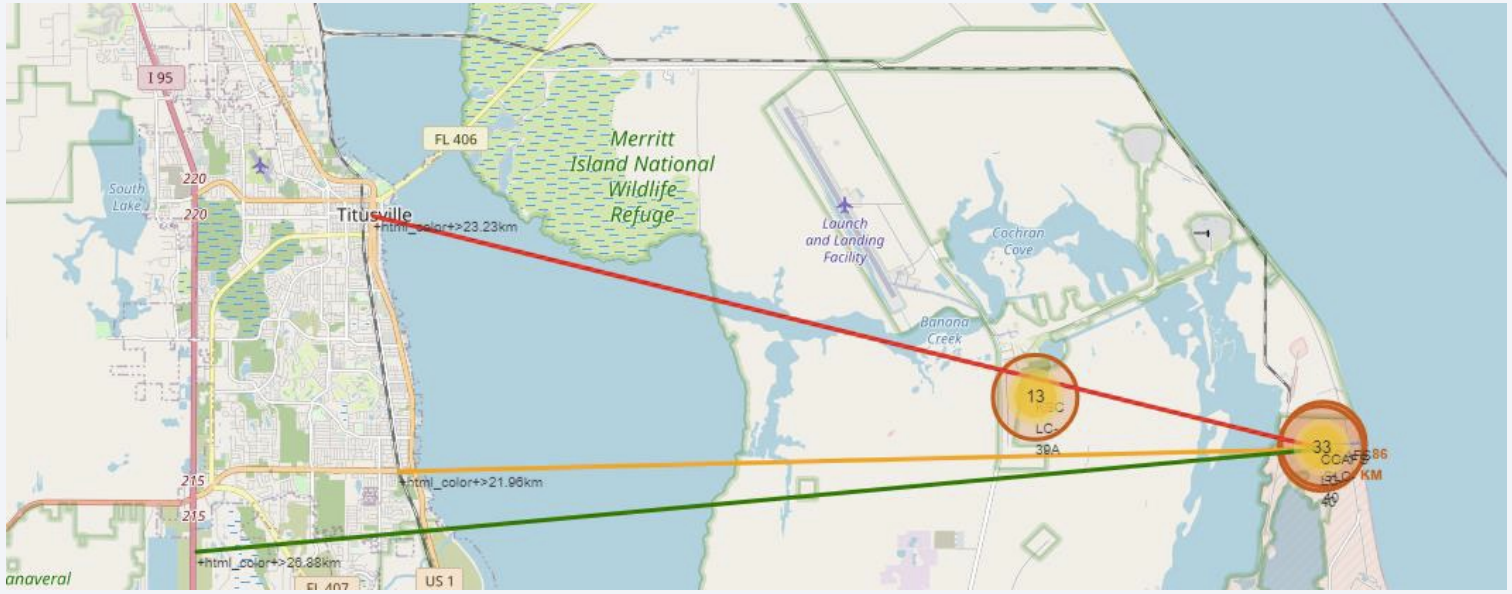
- .86 km from the nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway

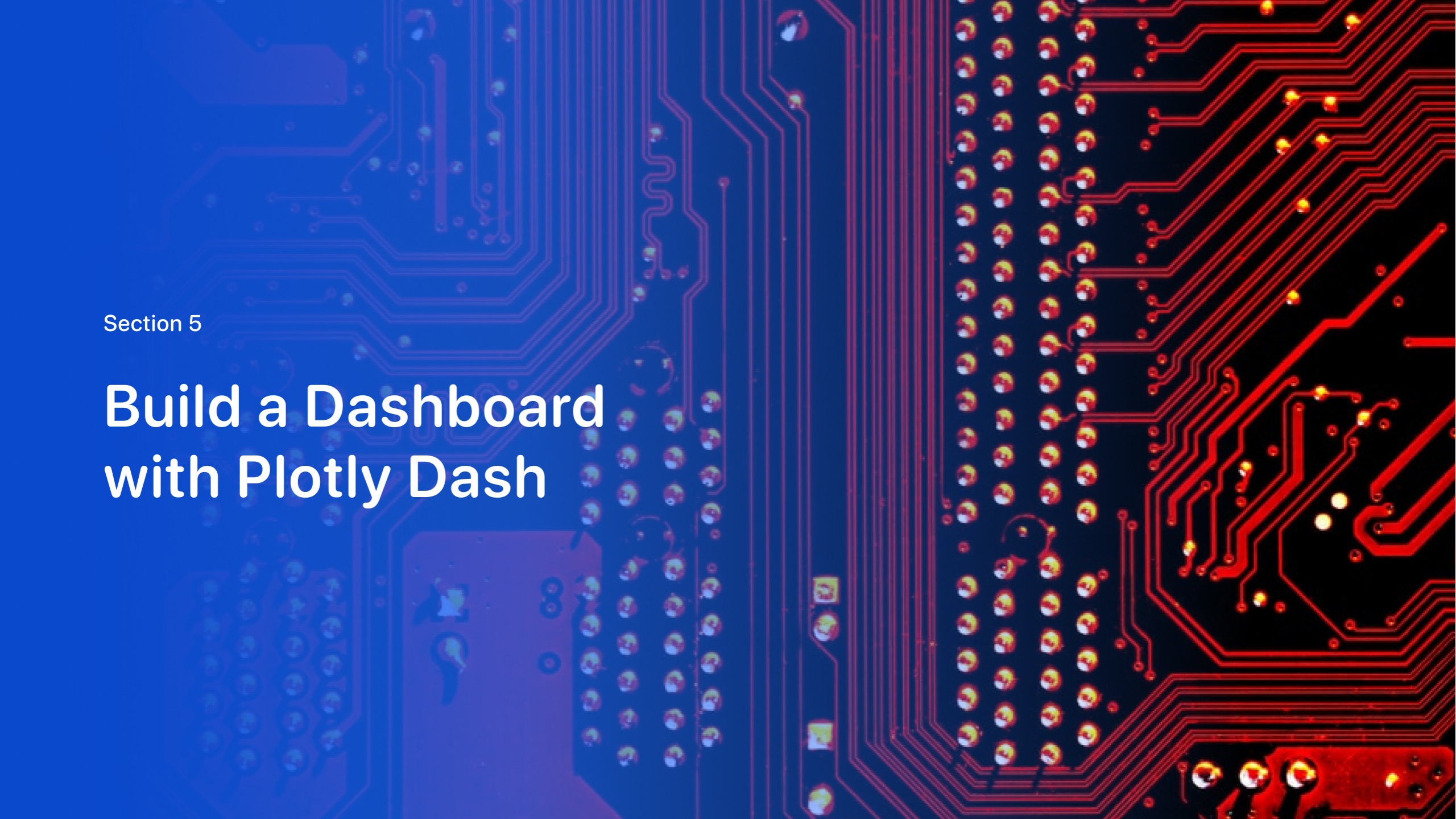


Distance to proximities

CCAFS SLC-40:

- 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/Infrastructure and Cities: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and equipment to or from launch sites.





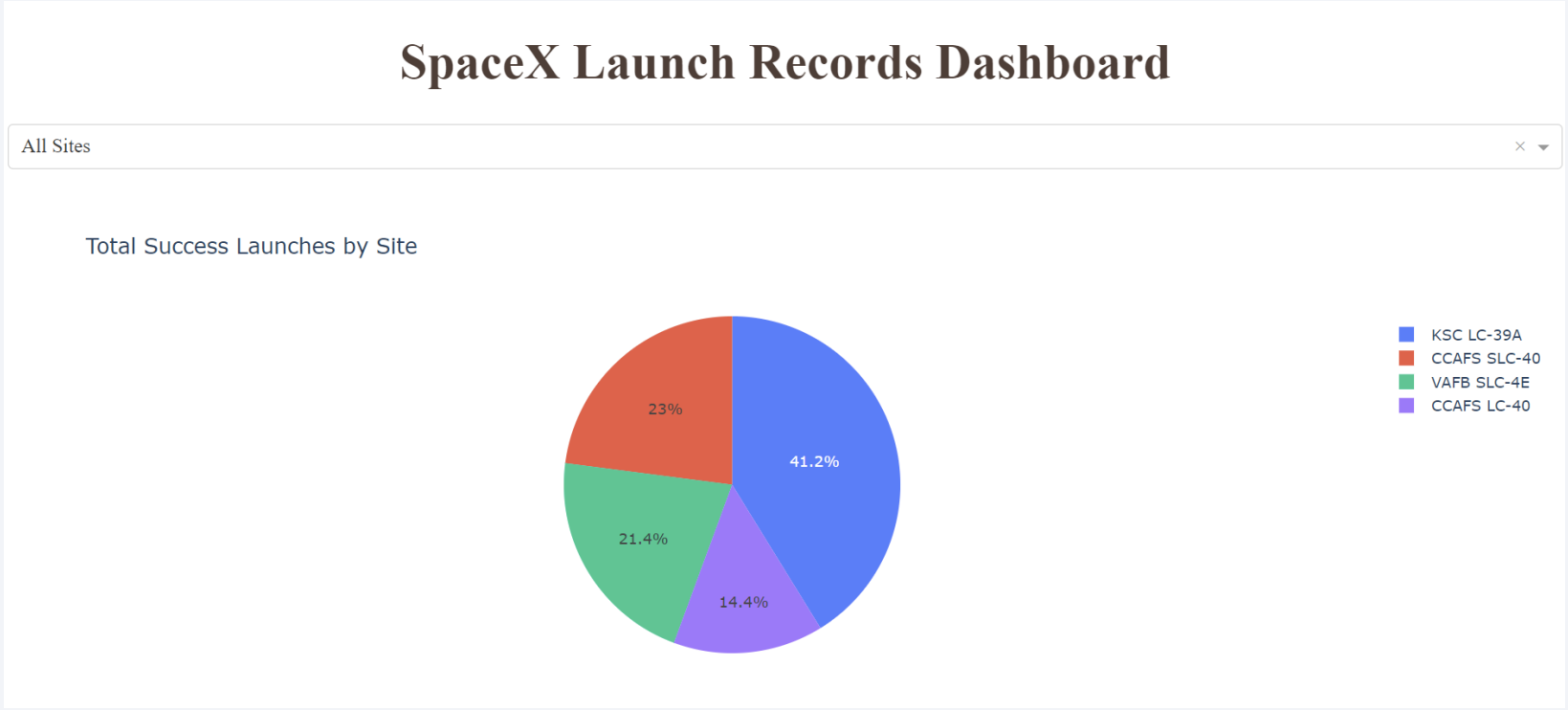
Section 5

Build a Dashboard with Plotly Dash

Launch Success (All Sites)

Success as Percent of Total:

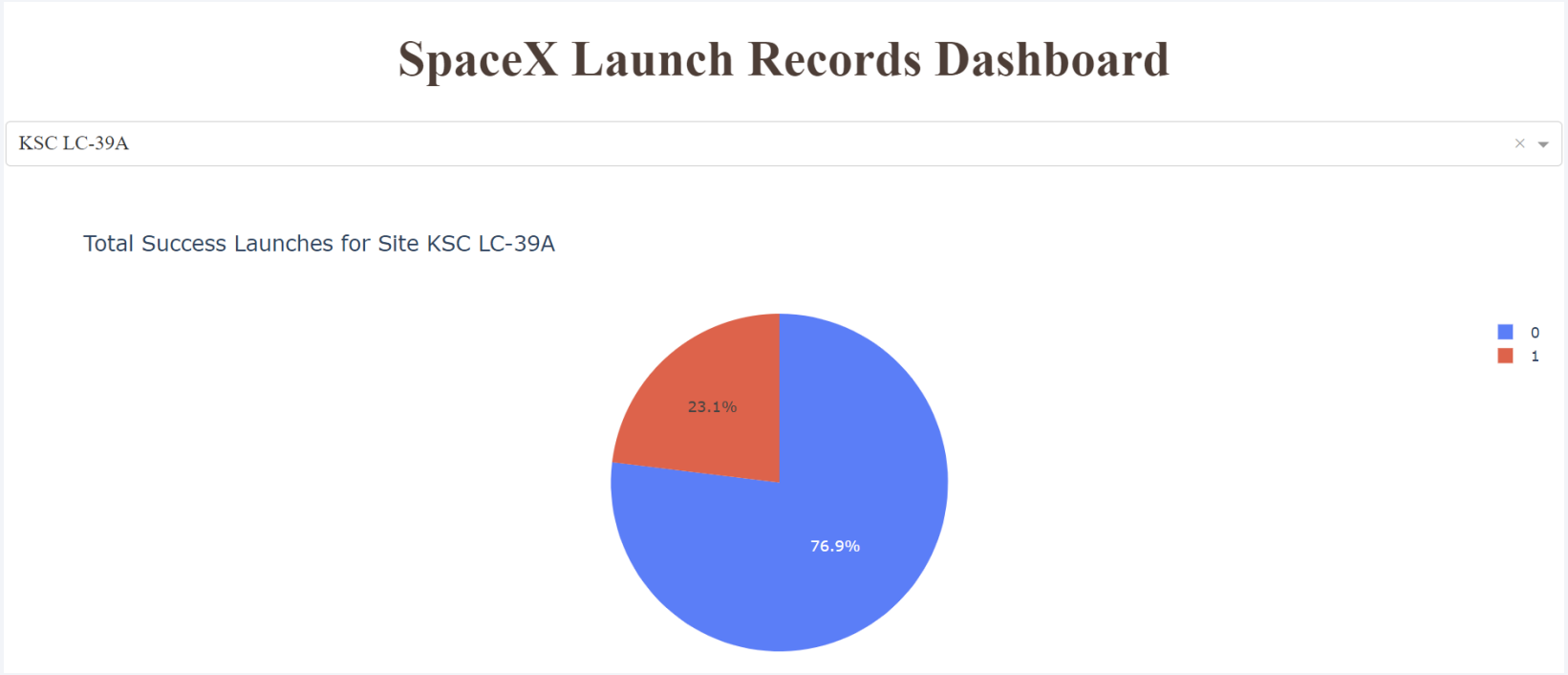
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

Success as Percent of Total:

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



Payloads 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 a failure

Section 6

Predictive Analysis (Classification)

Classification Accuracy

All the models performed about the same level and had the same scores and accuracy . This is likely due to the small dataset used in this case. The Decision Tree model slightly outperformed the rest when looking at .best_score_

Out[39]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.923077	0.800000
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

In [40]:

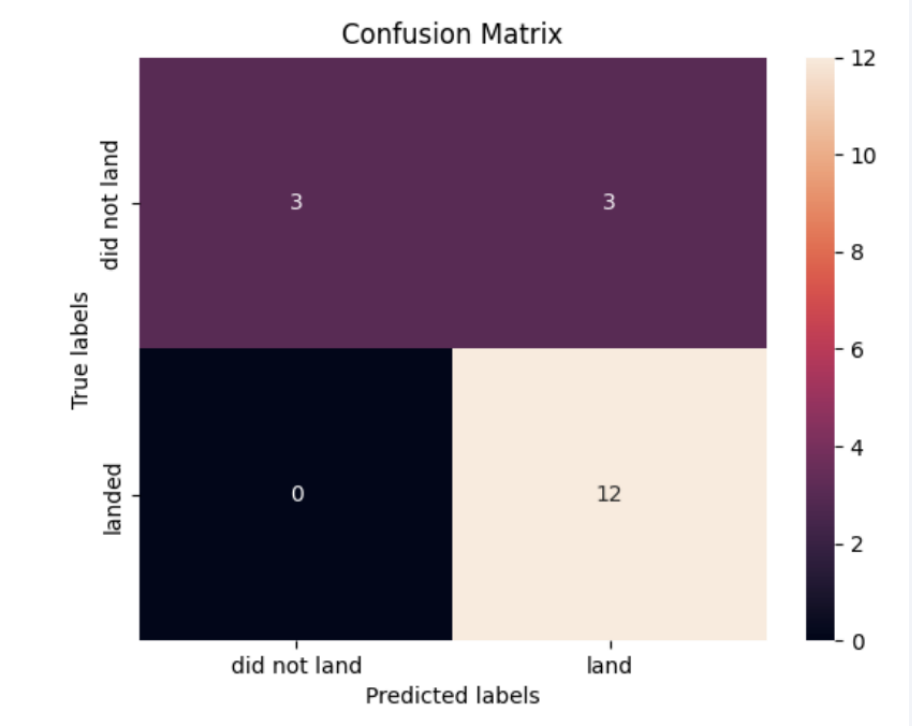
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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 0.8875
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, '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



Conclusion

Research:

- Model Performance: The models performed similarly on the test set with the decision tree slightly outperforming the rest.
- Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of the earth – which helps reduce the overall cost of the launch
- Coast: All 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

Thank you!