Winning Space Race with Data Science

Alice C. Long
April 17, 2023

Presentation Outline

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

Executive Summary - 1

Summary of methodologies

This research attempts to explore and identify factors for successful SpaceX rocket landings. In exploring this problem, the following methods were utilized:

- **Collect** data using SpaceX REST 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)

Executive Summary - 2

Summary of all results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-3A has the highest success rate of all SpaceX landing sites observed
- Orbits ESD-L1, GEO, HEO, and SSO all have a 100% success rate for landings

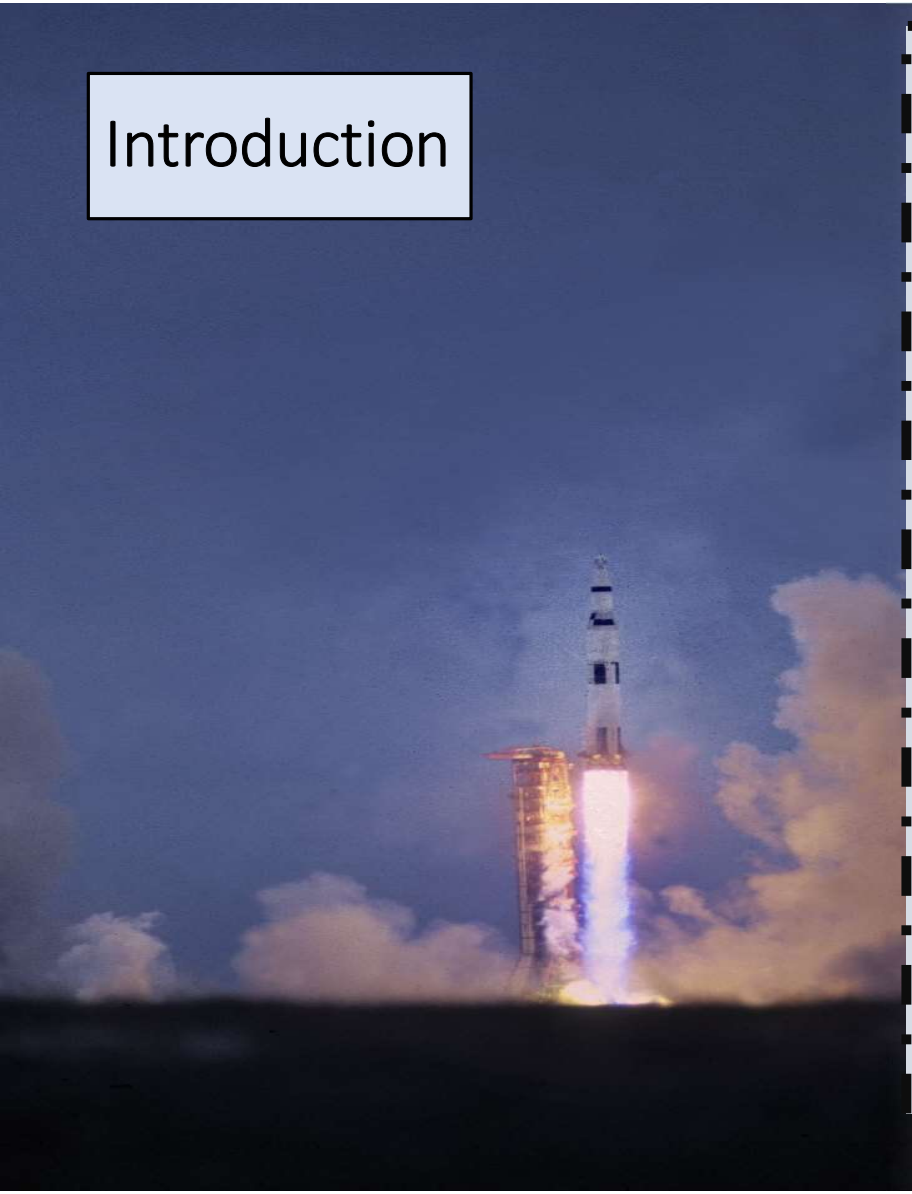
Visualization/Analysis

- Most of the launch sites are close to the equator, all are near a coastline

Predictive Analysis

- All models performed similarly using the test data set.
- The decision-tree model slightly out-performed the other models.

Introduction



Project background and context:

The SpaceX Falcon 9 program launches rockets at a cost of 62 million dollars - the company saves money by reusing the first stage of rockets. 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.

The problem under investigation:

- This project's primary task is to predict the frequency and accuracy of successful SpaceX Falcon 9 rocket landings. Specifically, this project explores:
  - How payload mass, launch site, number of flights, and orbits affects the first-stage landing process
  - Successful landings rate over time
  - Find the best predictive model for successful rocket landings

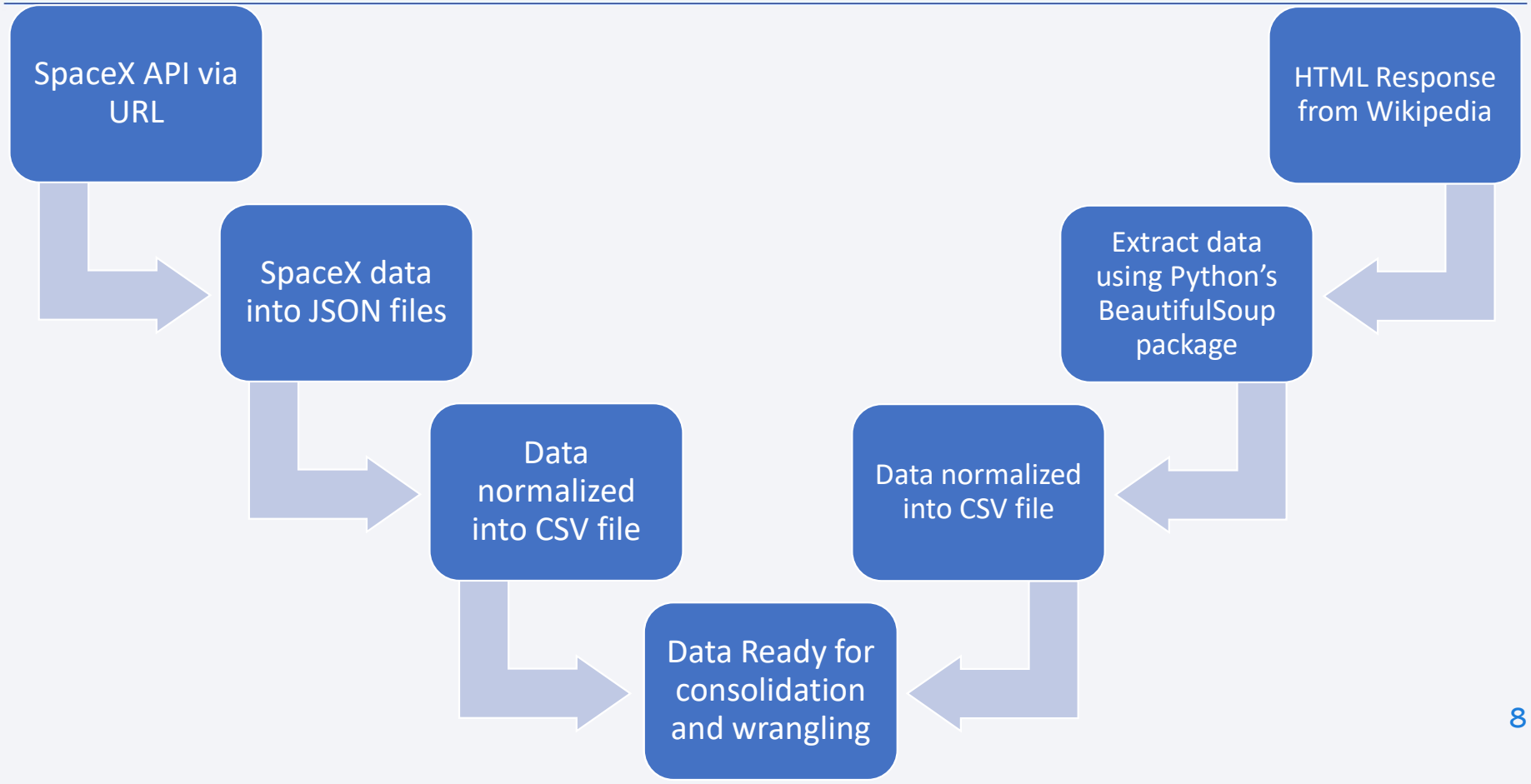
Section 1

Methodology

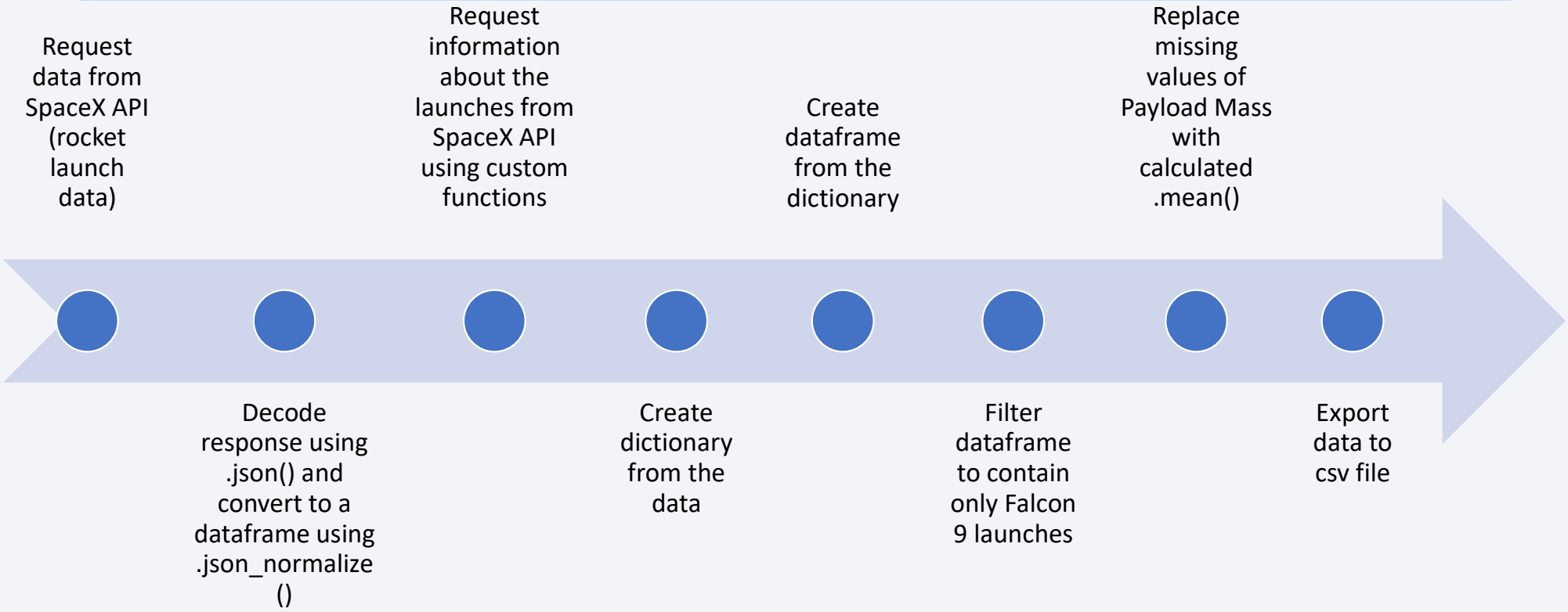
Methodology - Steps for Analysis

- **Collect Data:** SpaceX REST API; Web scraping from Wikipedia
- **Data Wrangling:** Filtering the data for missing values and applying One Hot Encoding data fields for Machine Learning and data cleaning for null values and irrelevant data
- **Explore**: Perform exploratory data analysis (EDA) using visualization and SQL
- **Visualize:** Perform interactive visual analytics using Folium and Plotly Dash
- **Build Models:** Perform predictive analysis using classification models; Logistic Regression, KNN, SVM, DT models used to evaluate the best predictor

Data Collection - Flowchart

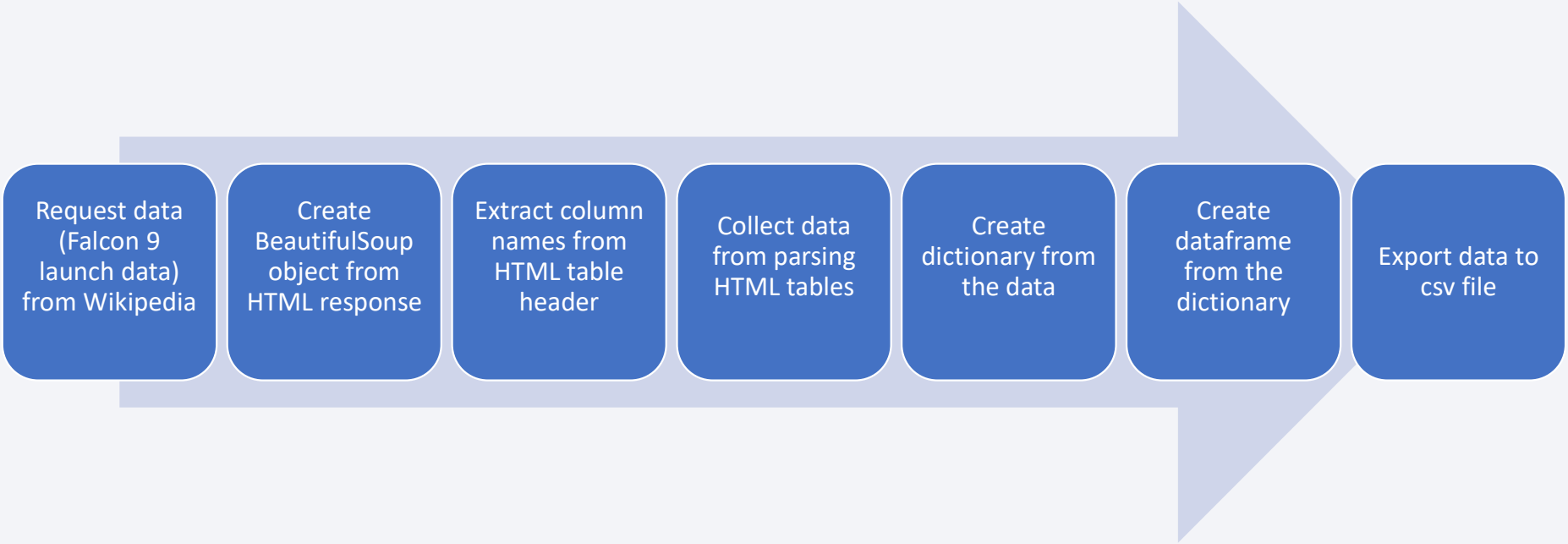


Data Collection – SpaceX API



- GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/1e7d687dacb93247236ea72ed0ecb8a78e369ca2/01%20Spacex-data-collection-api.ipynb

Data Collection – Web Scraping



- GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/a6bbe0502648663d0403e7a7a74a8ec5720ed472/02%20spacex-webscraping.ipynb

Data Wrangling

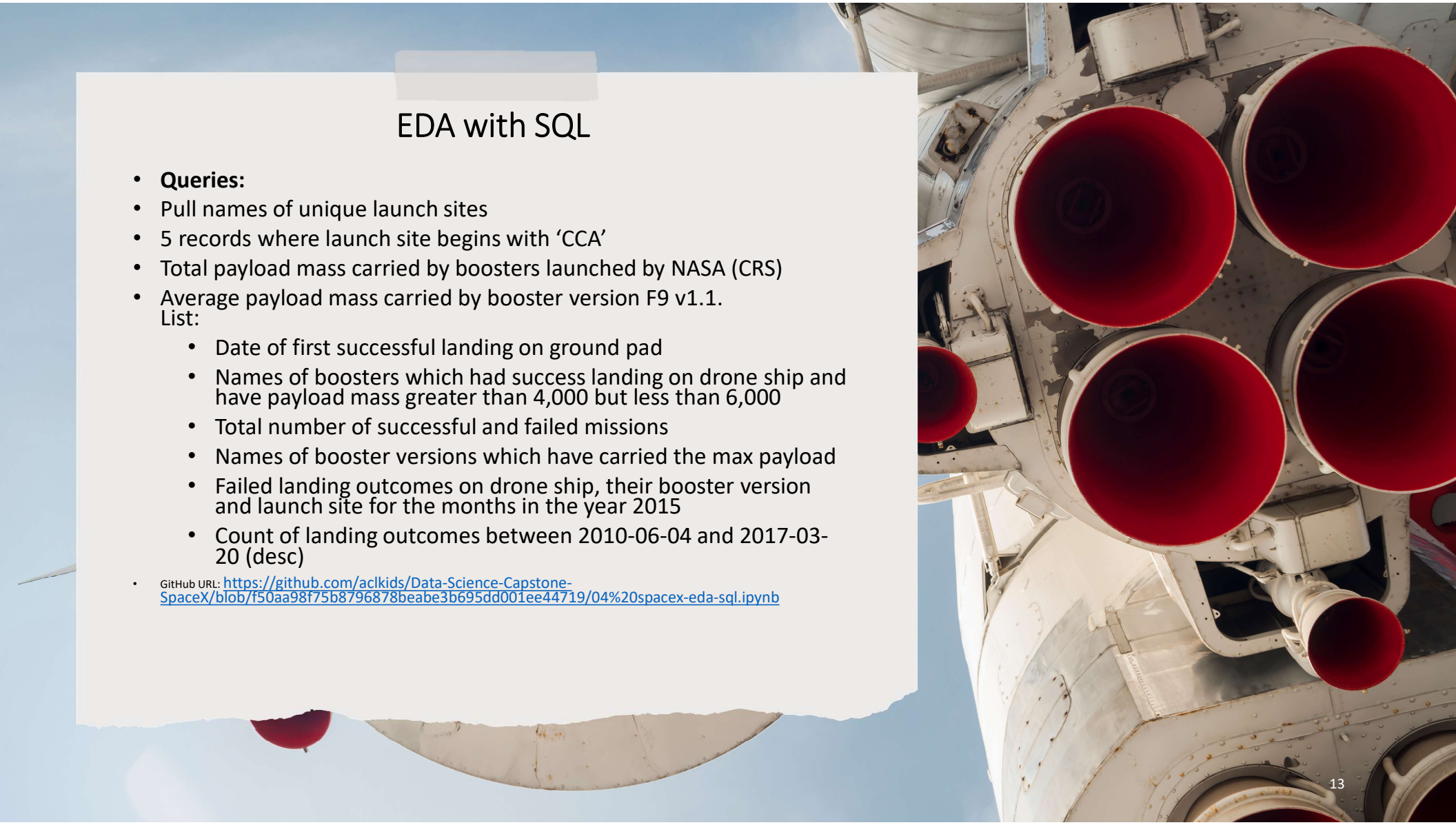
- Perform EDA and determine data labels
- Calculate:
  - # of launches per site
  - # and occurrences of orbit
  - # and occurrences of mission outcome (per orbit type)
- Create Binary
  - Landing outcome column (dependent variable)
- Export data to CSV file

Data Wrangling continued

- **Landing Outcome**
  - Landing not always successful
  - **True Ocean:** Mission outcome had a successful landing to a specific region of the ocean
  - **False Ocean:** Mission outcome was not successful to a specific region of the ocean
  - **True RLTS:** Mission successful landing on ground landing pad
  - **False RLTS:** Mission landing on ground landing pad not successful
  - **True ASDS:** Mission outcome had successful landing on drone ship
  - **False ASDS:** Mission outcome not successful landing on drone ship
- **Outcomes Converted:** into binary where 1 is successful landing and 0 is unsuccessful landing
- Github: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/bf57c3c9591020b7c6bfc81546a44571fb4444f8/03%20spacex-datawrangling.ipynb

EDA with SQL

- **Queries:**
- Pull 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)
- GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/f50aa98f75b8796878beabe3b695dd001ee44719/04%20spacex-eda-sql.ipynb



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

- **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 URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/700da6a889ce9bd54bafe873099d662a41d9df3b/05%20spacex-eda-dataviz.ipynb

Interactive Map with Folium

Markers Indicating Launch Sites

- Added yellow circle at NASA Johnson Space Center's coordinate with a popup label showing its name in blue using its latitude and longitude coordinates
- Added blue circles at all launch sites coordinates with a popup label showing the 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

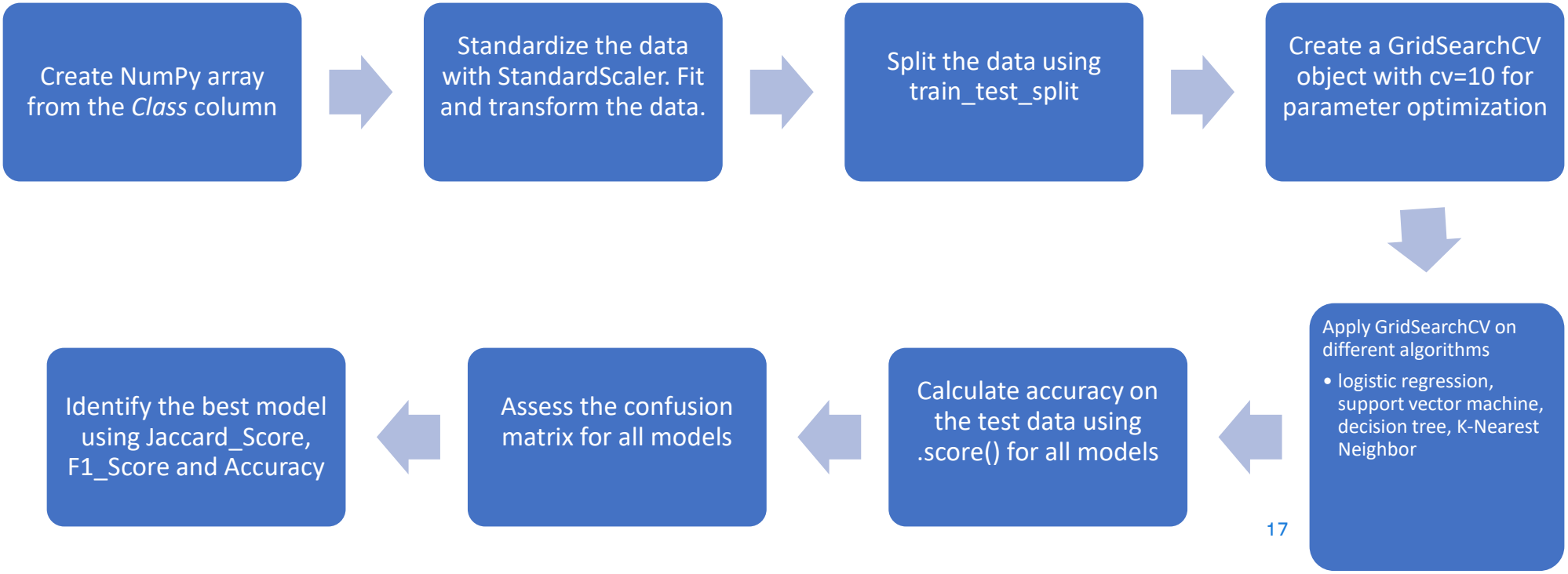
GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/d452472fd4dd8ac4ff3a2e69f77520ced7502ec5/06-viz%20with%20folium%20launch_site.ipynb

Dashboard with Plotly Dash

Dropdown Menu	Dropdown and App generates graphs	Slider menu for selecting payload range in kg
• Allows user to select specific launch sites and all launch sites	• Pie Chart to show successful and unsuccessful launches as percentage of total launches	• Dropdown menu with Slider generates scatter chart • payload mass v. success rate by booster version • Allows visualization of correlation between payload and launch success (labeled 'class')

GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/09b9f358930d6b1563827aaabdee62de8207fab2/07%20spacex_dash_app.py

Predictive Analysis (Classification)



GitHub URL: https://github.com/aclkids/Data-Science-Capstone-SpaceX/blob/09b9f358930d6b1563827aaabdee62de8207fab2/08_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

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

City Distance 23.234752126023245 Railway Distance 21.961465676043673 Highway Distance 26.88038569681492 Coastline Distance 0.8762983388668404

Results

Summary of 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

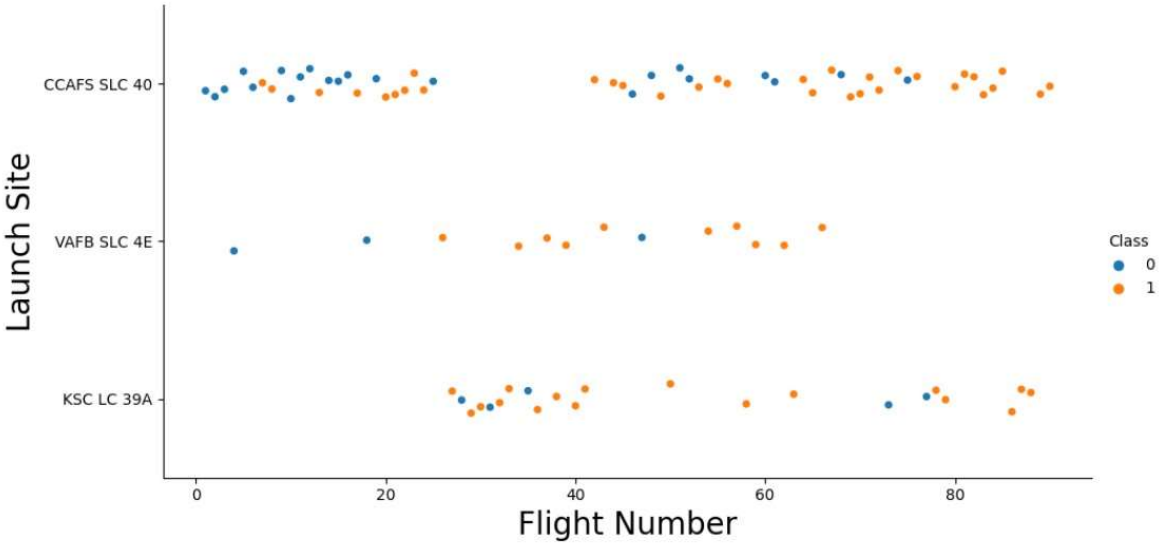
Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from areas susceptible to damage from failed launches (e.g., 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

Flight Number vs. Launch Site

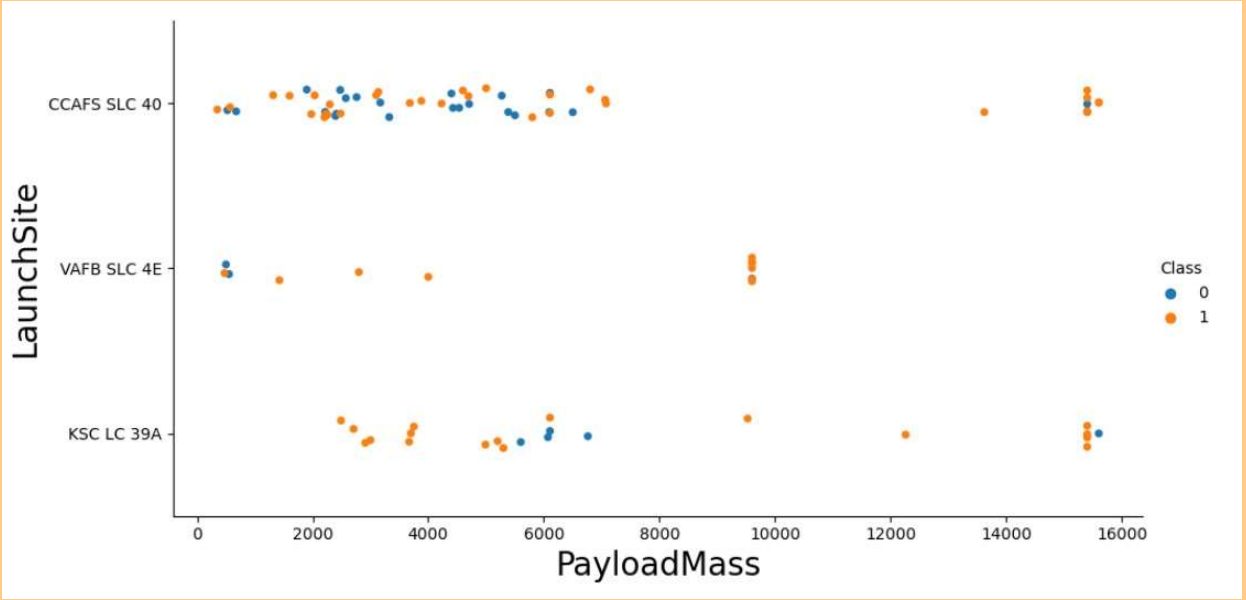


Exploratory Data Analysis

- Earlier flights had a lower success rate (marked as class=0, blue dots)
- Later flights had higher success rates (marked as class=1, orange dots)
- 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 rat

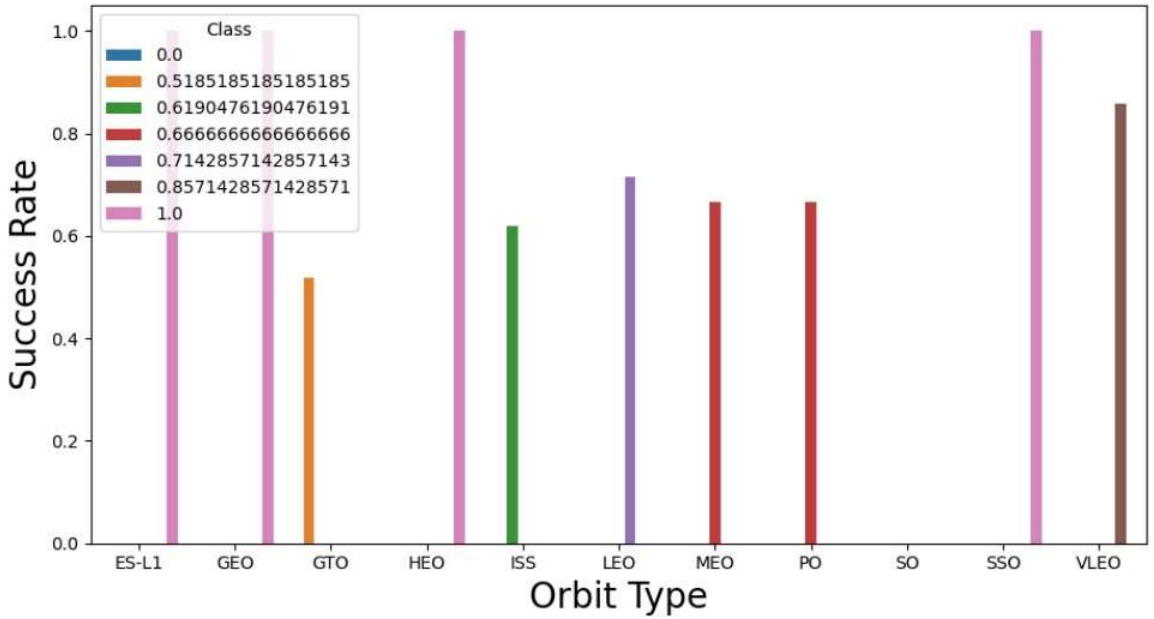
Payload vs. Launch Site

- 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



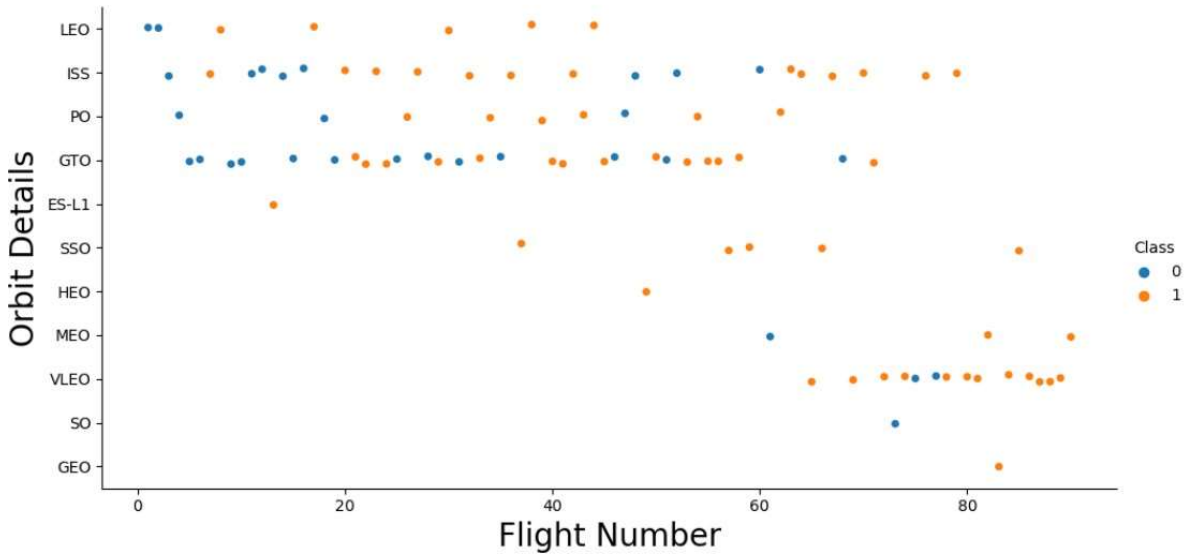
Success Rate vs. Orbit Type

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



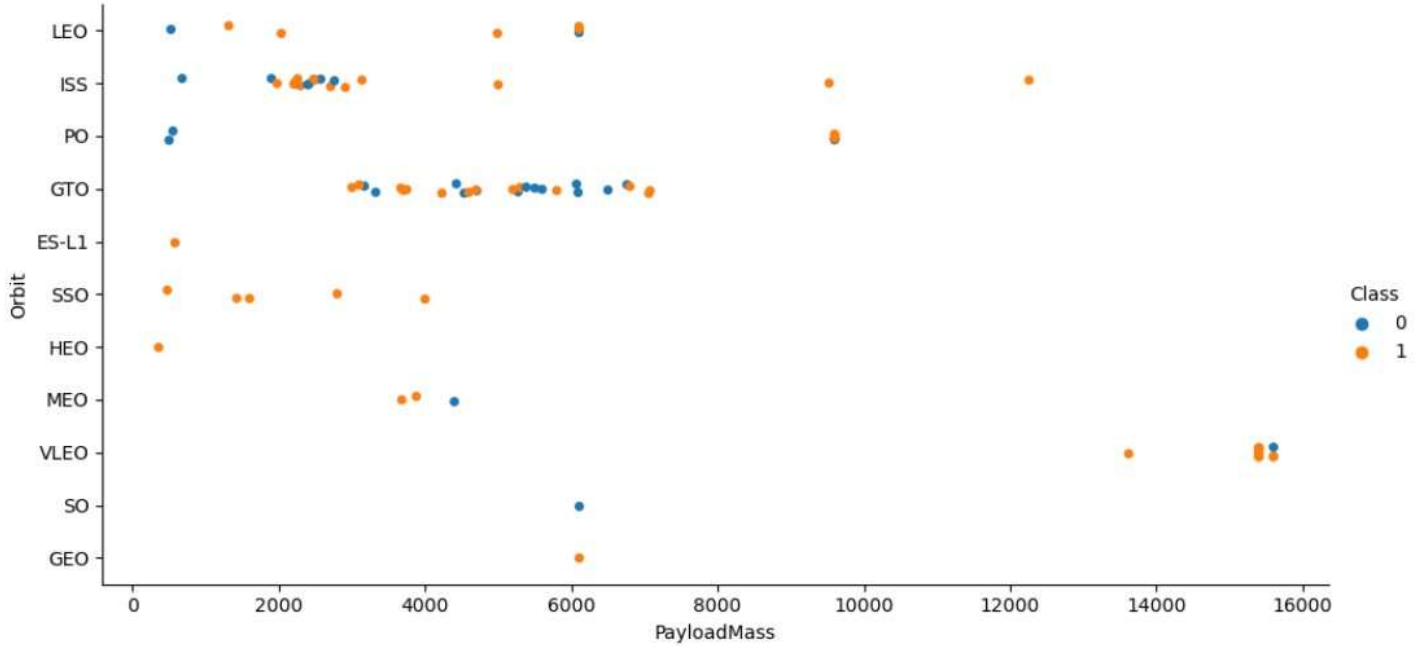
Flight Number vs. Orbit Type

- 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



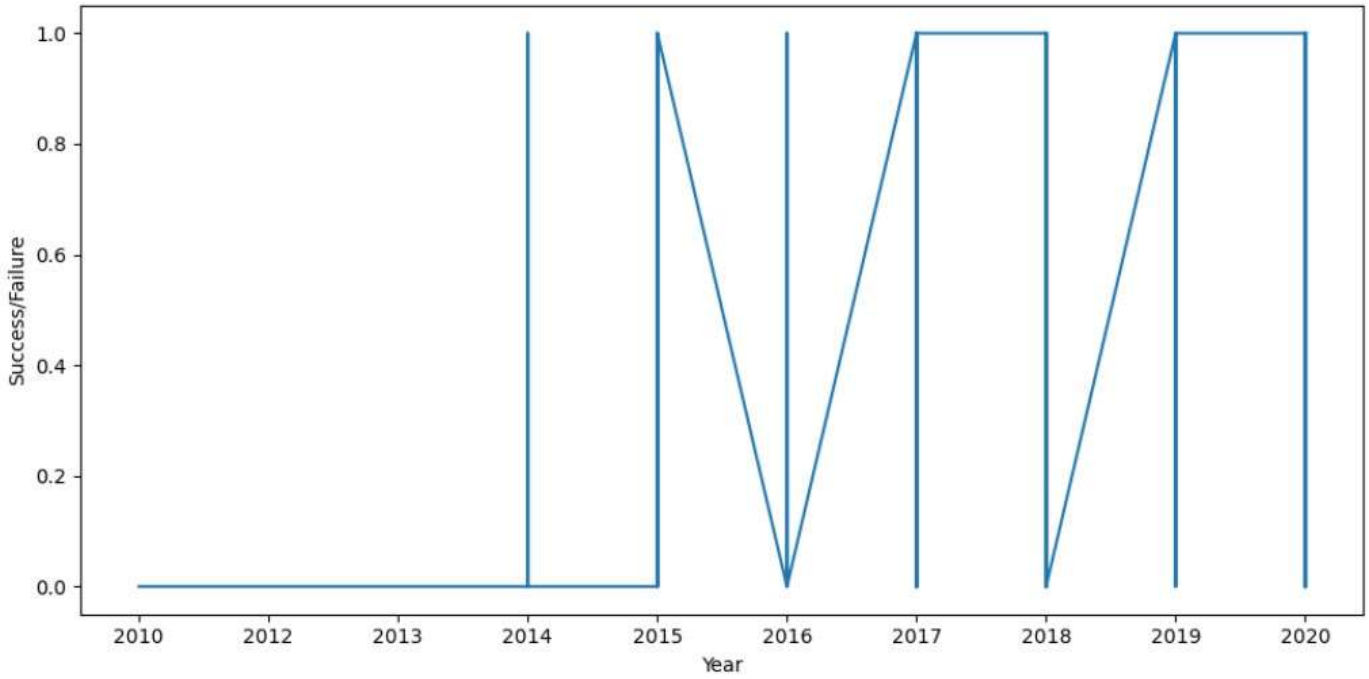
Payload vs. Orbit Type

- 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 Information

Landing Outcome

- Launch Site Names
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

```
# Select relevant sub-columns: `Launch Site`, `Lat(Latitude)`, `Long(Longitude)`, `class`
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
launch_sites_df
```

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

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

In [19]:
```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

Out[19]:

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

- This output shows 5 launch sites that begin with ‘CCA’

Total Payload Mass

61,9967 kg (total) carried by boosters launched by NASA (CRS)

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

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

payloadmass
619967.0

Average Payload Mass by F9 v1.1

Average payload mass by booster version F9 v1.1 is 6,138.3

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL;
```

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

min(DATE)
01/06/2014

- The first successful landing on a ground pad was on 1/6/2014

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

of the boosters which have success in drone ship and have payload mass greater than 4000 bu

```
elect BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME= 'Success (drone ship)' and PAY
```

```
'/my_data1.db
```

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

missionoutcomes
0
1
98
1
1

- Successful missions = 99
- Failure in flight = 1
- Success (payload unclear) = 1

Boosters Carrying the Maximum 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

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPA
```

```
 sqlite:///my_data1.db
one.
```

boosterversion
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

- Landing outcomes failed using a drone ship:
  - Booster version - F9 v1.1 B1012; Launch site – CCAFS LC-40; date: 10.1.2015
  - Booster version – F9 v1.1 B1015; Launch site – CCAFS LC-40; date: 4.14.2015

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,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)

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

- Success landing = 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

Rank the count of successful landing_outcomes between the date 04-06-

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

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

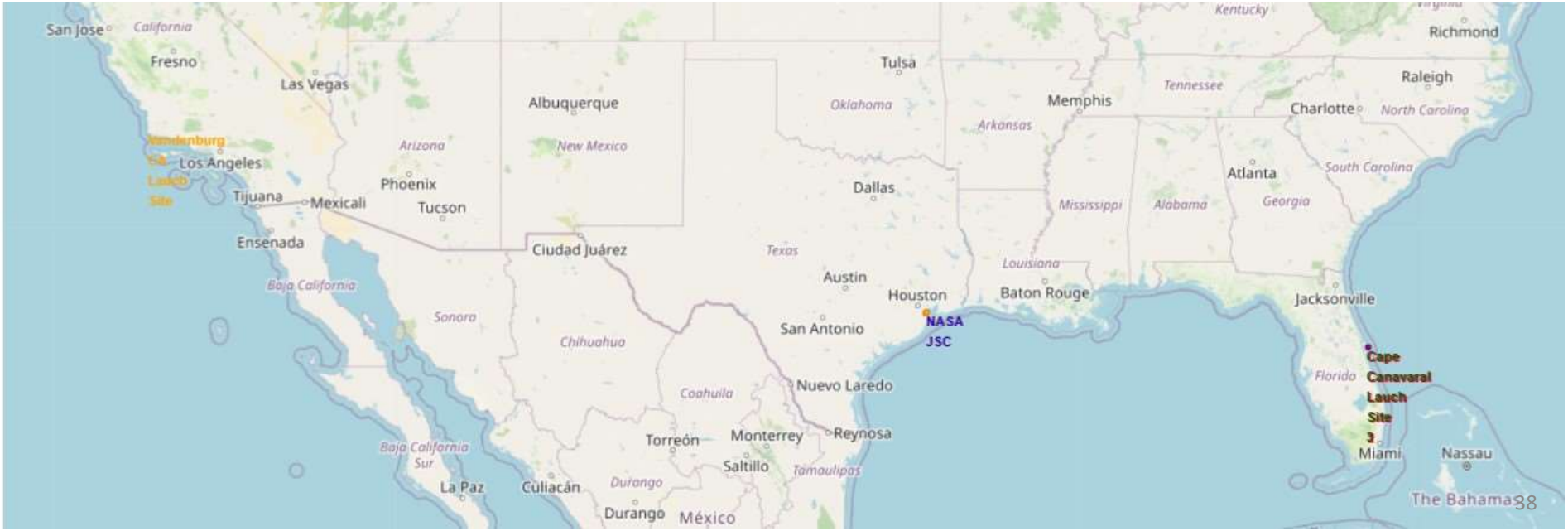
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Folium Map with Launch Sites

- The Folium map shows the launch sites generated from latitude and longitude.
  - Launch sites are near the equator and coast.
  - Being near the equator is beneficial because the Earth's equatorial orbit gives the rockets an additional natural boost.
  - This helps save the cost for each launch.



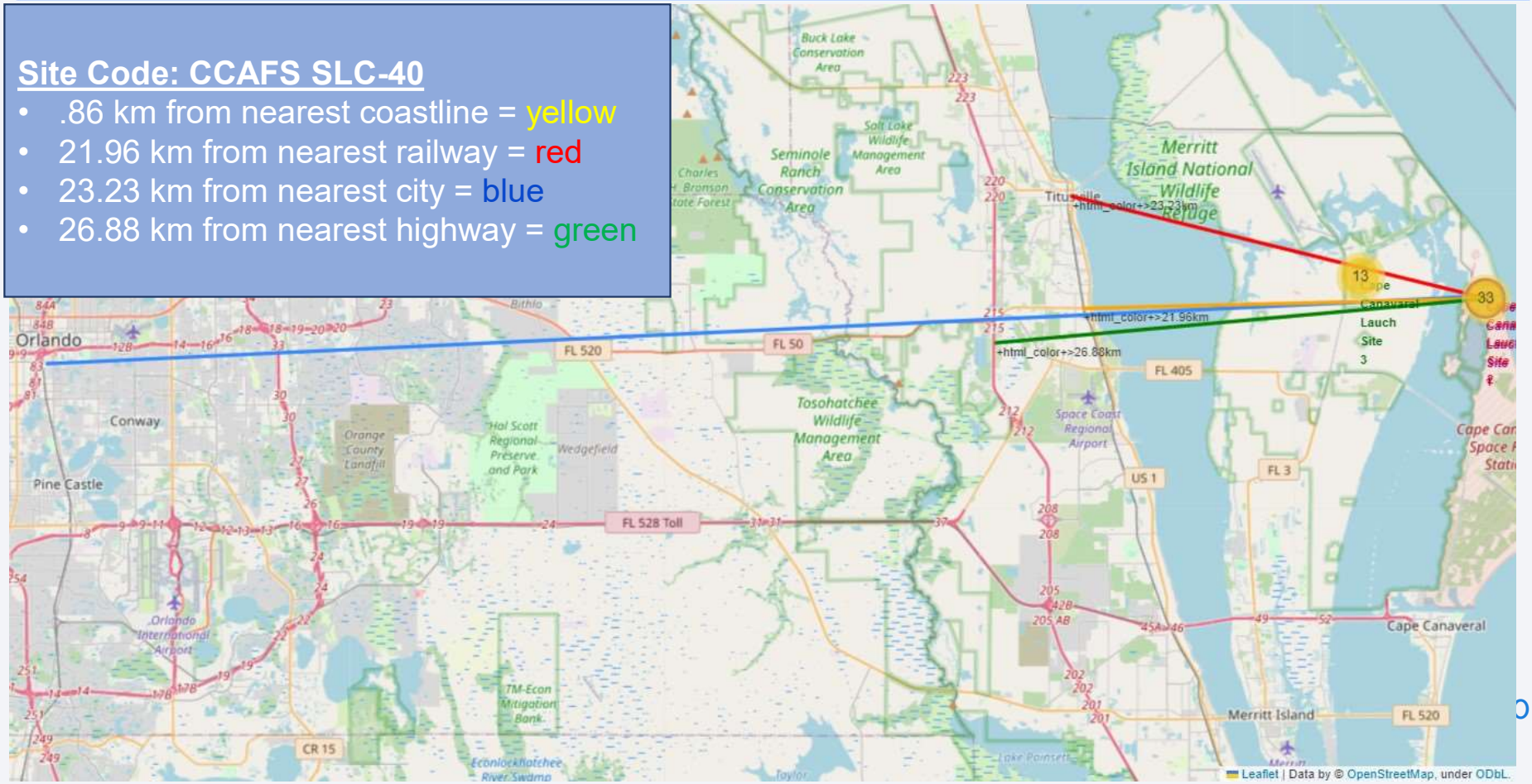
Folium Map Showing Success and Failure Outcomes for One Location

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

Folium Map Showing Distances from Cape Canavaral Launch Site

Site Code: CCAFS SLC-40

- .86 km from nearest coastline = yellow
- 21.96 km from nearest railway = red
- 23.23 km from nearest city = blue
- 26.88 km from nearest highway = green



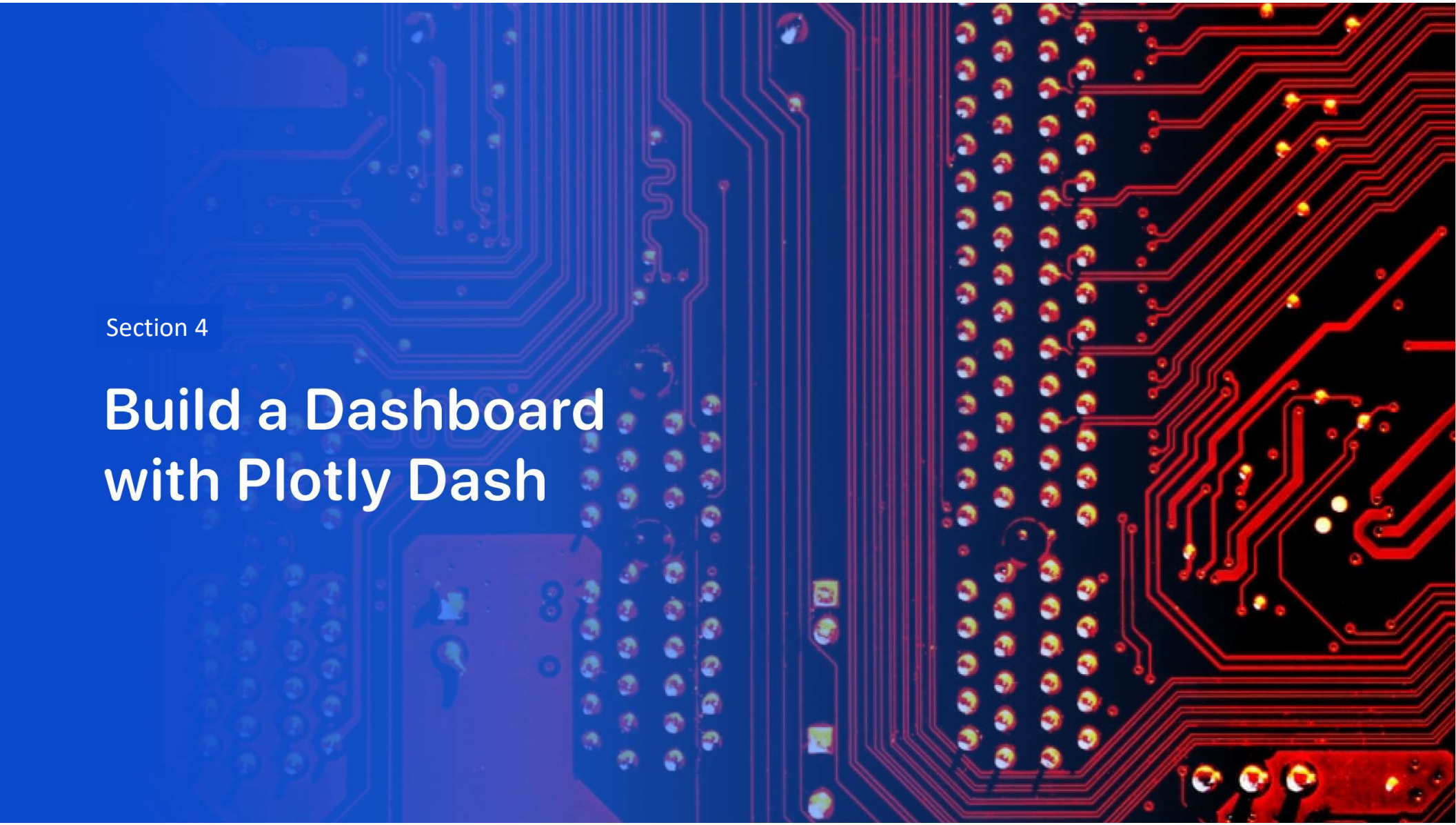
Distances to Proximities

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 material to or from it in support of launch activities.

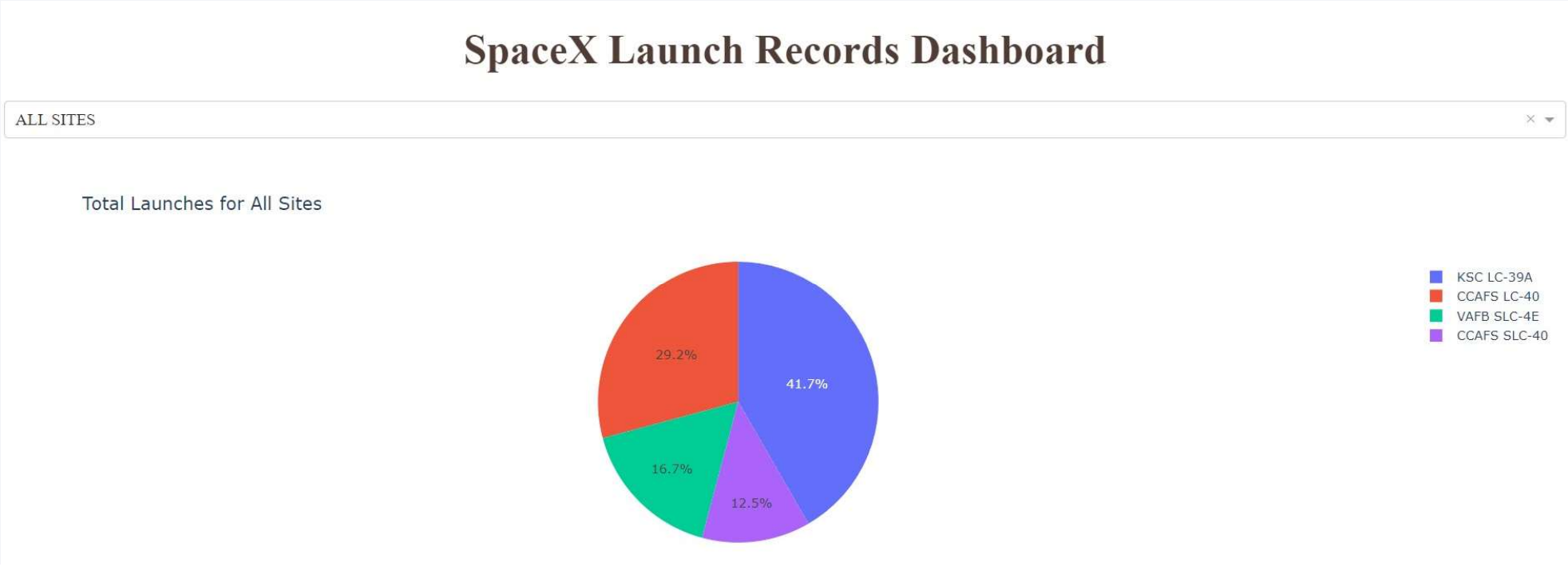


Section 4

Build a Dashboard with Plotly Dash

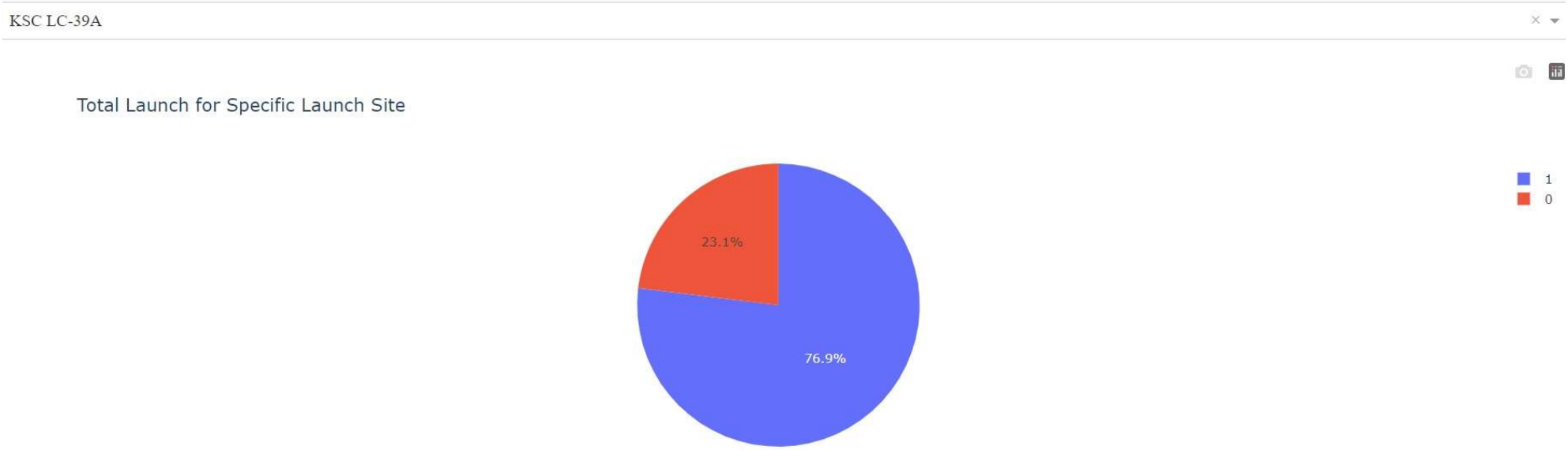
Launch Success by Site

- **Success as percent of total**
  - **KSC LC-39A** has the highest success rate at **41.7%**
  - **CCAFS SLC-40** has the lowest success rate **12.5%**



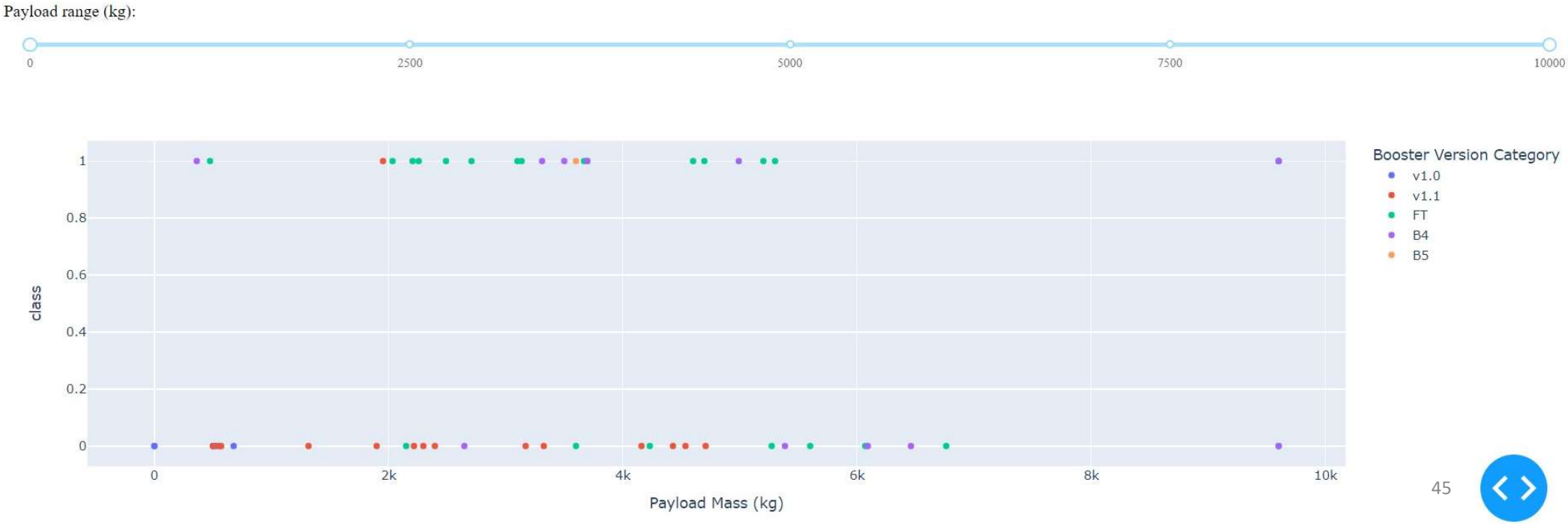
Launch Success (KSC LC-39A)

- **Success as Percent of Total**
- **KSC LC-39A** has a launch success rate of **76.9%**



Success and Payload Mass (kg)

- **By Booster Version**
- Payloads between **2,000 kg** and **5,000 kg** have the **highest success rate**
- 1 indicating successful outcome and 0 indicated failed outcome



Predictive Analysis

Classification Accuracy, Confusion Matrices

Classification Accuracy

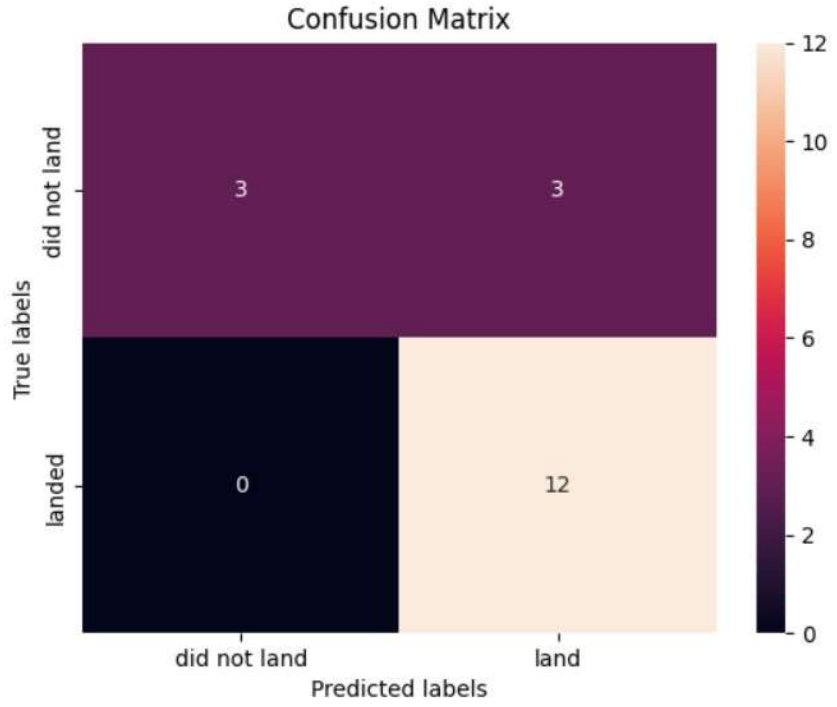
- All the models performed at about the same level and had the same scores and accuracy ratings.
  - This is likely due to the small size of the dataset.
- The Decision Tree model slightly outperformed the rest when looking at .best_score_
  - .best_score_ is the average of all cv folds for a single combination of the parameters

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


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

```
Best model is DecisionTree with a score of 0.888888888888889
Best params is : {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':
5, 'splitter': 'best'}
```

Confusion Matrix

```
yhat3 = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat3)

plt.show()
```



Purpose: A confusion matrix summarizes the performance of a classification algorithm. The confusion matrix for the decision tree method indicates good fit for the model.

- 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, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Commentary

- This relatively small dataset yielded conflicting results in some areas. A larger dataset will help improve the accuracy of the predictive analysis.
- Additional types of analysis (PCA) should be considered to improve interpretation and accuracy of data.

Thank you!