IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Riddhi Devalia
December 26, 2023

Outline

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

Executive Summary

Summary of methodologies

The project aims to identify the factors for a successful rocket landing. Following methodologies were used to get to the conclusion:

- Data Collection: Using SpaceX RestAPI and web scrapping methods
- Data Wrangling: To create success/failure outcome variable
- Data Exploration: Plotting graphs to visually explore the relationships between independent and dependent variables such as yearly trends, payload launch sites and mass.
- Data Analysis: Analyze the data with SQL queries to calculate following statistics: total payload, success/failure outcome attempts etc.
- Data Visualization: Build dashboards to explore relationship between launch sites and success outcomes
- Build Model: Predicting landing outcomes using a suitable classic classification Logistic regression , SVM, KNN & Decision tree algorithms

Summary of all results

- Exploratory Data Analysis: 1. Launch success had improved over time as noted in a yearly trend
  2. KSC LC 39A has the highest success rate amongst all landing sites
  3. Orbits GEO, HEO and SSO have 100% success rates
- Visualization Analysis: Most launch sites are close to the equator and near the coast
- Predictive Analytics: The decision tree model outperformed all the other models used to predict the outcomes. Although other models had achieved similar accuracy on all grounds.

Introduction

- ## Project background and context

- SpaceX 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 SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Our Goal in the project is to determine if the Falcon 9 first stage will land successfully or not.

- ## Problems you want to find answers

- How payload mass , launch sites, orbits and number of flights affect the first stage landing success
- Explore rates of successful landing over time
- Explore best predictive model for successful landing using binary classification models

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using 2 main methodologies: using SpaceX RestAPI and web scrapping
- Perform data wrangling
  - Data was wrangled using filtering methods, handling missing values and applying one hot encoding- to prepare data for analysis and modelling purposes.
- 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 – SpaceX API

- Steps:
  - Request Data from SpaceX API- Rocket launch data
  - Decode response using .json() and convert the data frame using .json_normalize()
  - Request information about the launches from SpaceX using custom functions
  - Create dictionary from data
  - Create dataframe from dictionary
  - Filter dataframe to only contain Falcon 9 launch data
  - Replace missing values of payload mass with calculated class mean
  - Export data to CSV
  - **GitHub URL** for Data Collection Using API: Click URL Here

Data Collection - Scraping

- Steps:

- Request falcon 9 launch data from Wikipedia
- Create Beautiful Soup object from HTML response
- Extract column names from HTML Table Header
- Collect data from parsing HTML tables
- Create dictionary from data
- Create dataframe from dictionary
- Export data to CSV files

- GitHub URL of the completed web scraping notebook: Click URL Here

Data Wrangling

- Steps:
- Perform EDA and determine data labels
- Calculate the below:
  - Number of launches per launch site
  - Number of flights per orbit
  - Number of success outcome per orbit type
- Create Binary landing outcome column (dependent column)
- Export data to CSV file

- GitHub URL of your completed data wrangling: Click URL here

EDA with Data Visualization

- Summarize charts plotted:
  - Flight number vs Payload
  - Flight number vs Launch Site
  - Payload Mass (kg) vs Launch Site
  - Payload Mass (kg) vs Orbit Type

Analysis: In the Payload Vs. Launch Site scatter point chart, you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

We observed that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

- GitHub URL of your completed EDA with data visualization- Click URL here

EDA with SQL

- Summarize the SQL queries:
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display average payload mass carried by booster version F9 v1.1
- Display the total payload mass carried by boosters launched by NASA (CRS)
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass.
- 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
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 (desc).

- GitHub URL of completed EDA with SQL notebook: Click URL here

Build an Interactive Map with Folium

- Markers indicating Launch Sites:
- Added Blue circle at NASA Johnson space center's coordinates with a popup label showing its name with longitude and latitude values
- Added Red circle at all launch sites coordinates with a popup label showing its name with longitude and latitude values
- Colored markers of launch outcomes:
- Added colored label markers of successful(Green) and unsuccessful(Red) launches at each launch site to show which launch sites have high success rates.
- Distance between launch site to proximities:
- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to nearest coastline, railway, highway and city
- GitHub URL of your completed interactive map with Folium map: Click URL here

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
  - Drop Down list with launch sites: Allowing users to select all or individual launch sites
  - Pie Chart showing successful launches: Allow users to see successful vs unsuccessful launches as a percent of the total
  - Slider for 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 correlation between payload and launch success
- GitHub URL of Plotly Dash lab- Click URL Here

Predictive Analysis (Classification)

- Model Summary:
  - Create a NumPy array from the column Class
  - Standardize the data
  - Split the data into training and test data.
  - Create a GridSearchCV object with cv = 10
  - Apply GridSearchCV on varied classification models such as SVM, KNN, Decision Tree and Logistic regression
  - Calculate the accuracy on the test data
  - Assess the confusion matrix for all models
  - Identify the best model using the Jaccard_Score, F1 Score and Accuracy

- GitHub URL of predictive analysis lab: Click URL here

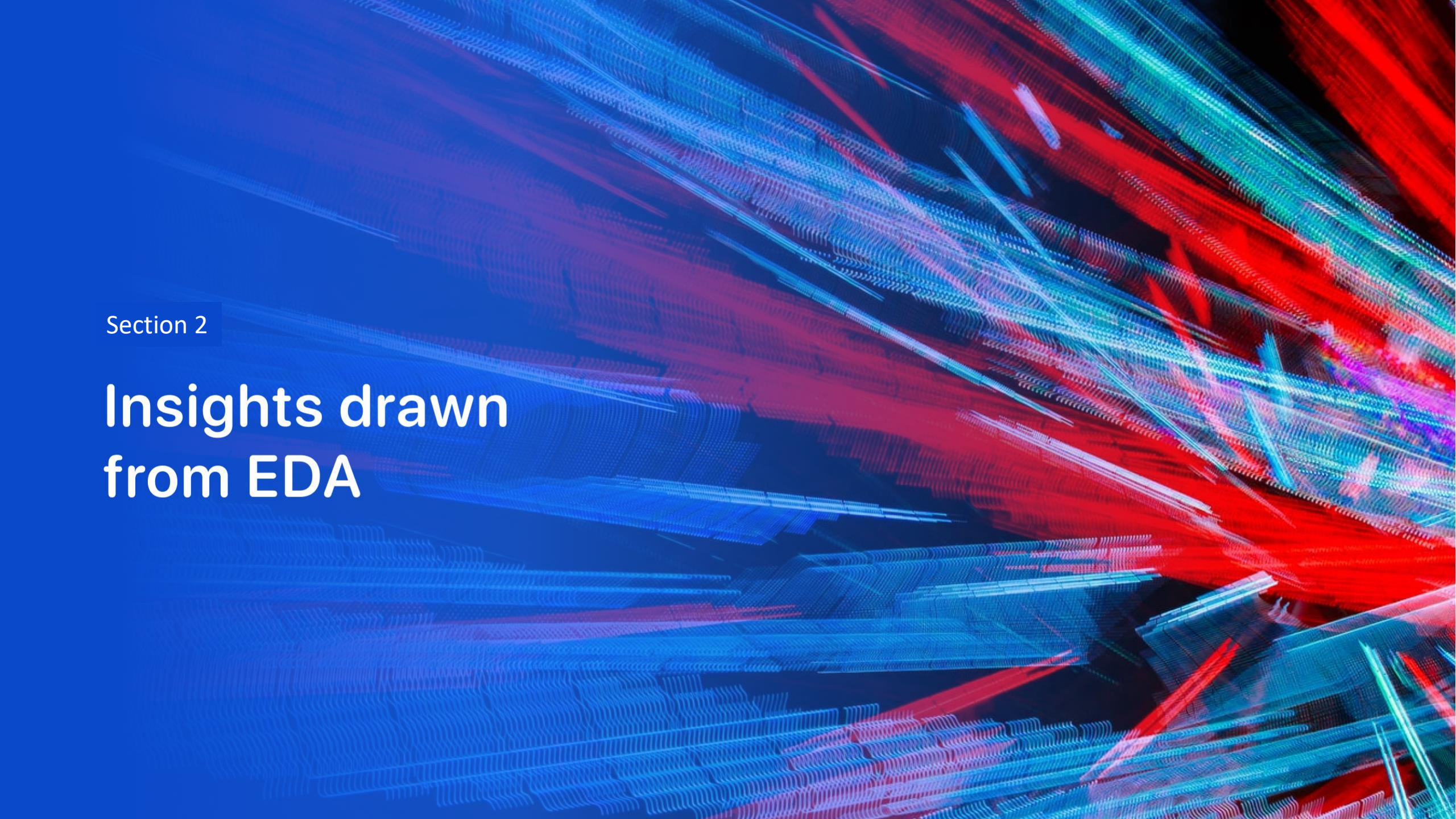
Results

- Exploratory data analysis:
- Launch success has improved over time
- KSC-LC 39A has the highest success rate among landing sites
- Orbits GEO, HEO, SSO have highest success rates of 100%

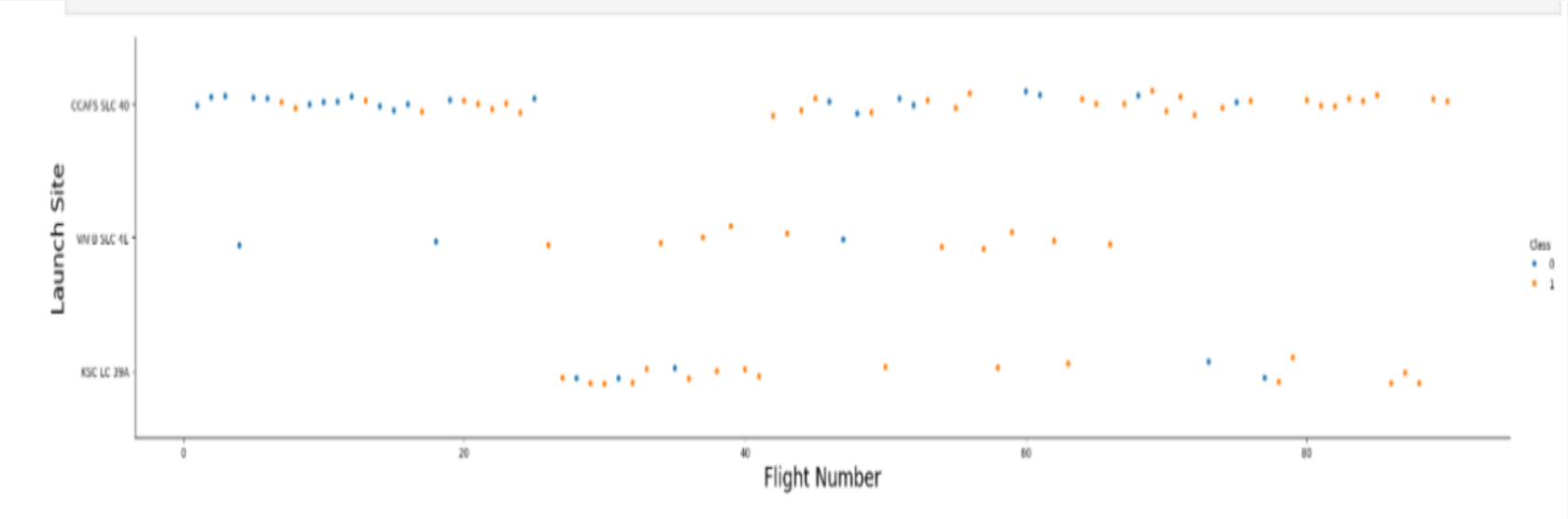
- Visual analysis:
- Most launch sites are near the equator and close to the coast.
- Launch sites are far enough from Highway railway and city to avoid any damage caused due to failing any launches

- Predictive Analysis:

Decision Tree Model is the best suited model for the dataset. It has outperformed all the other models taken into consideration for analysis.

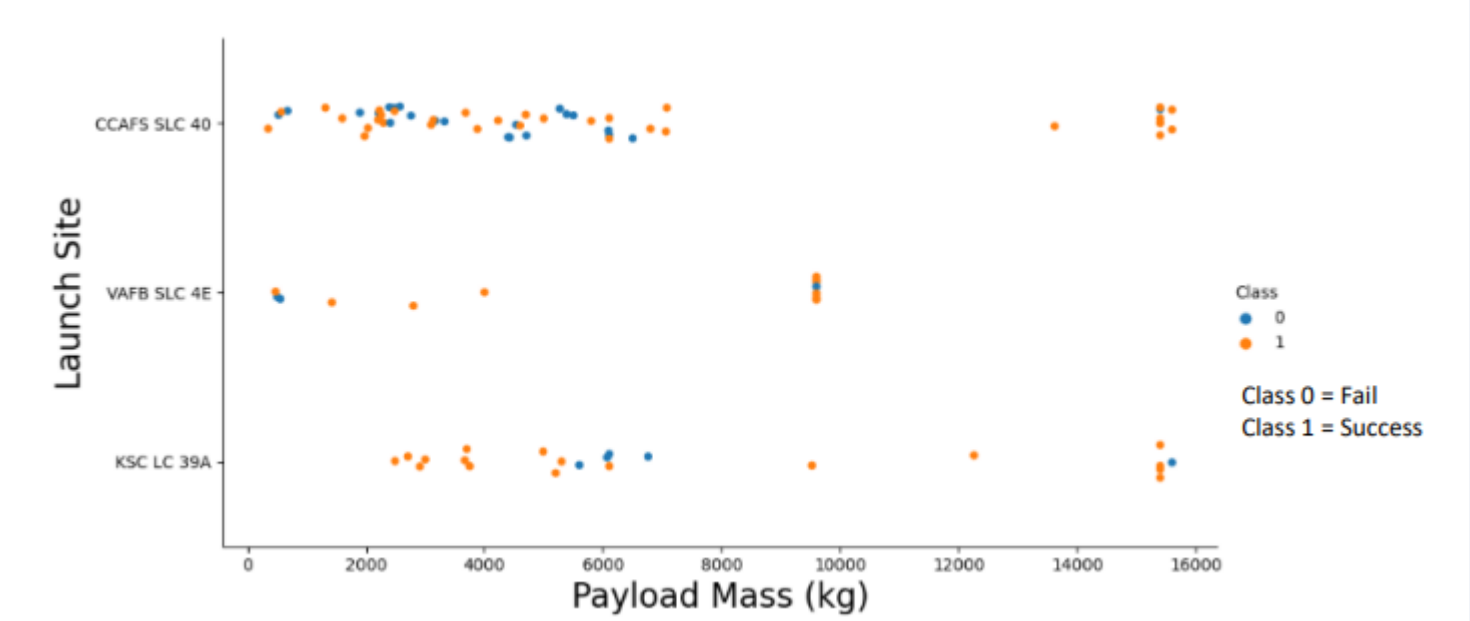

Section 2
Insights drawn from EDA

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

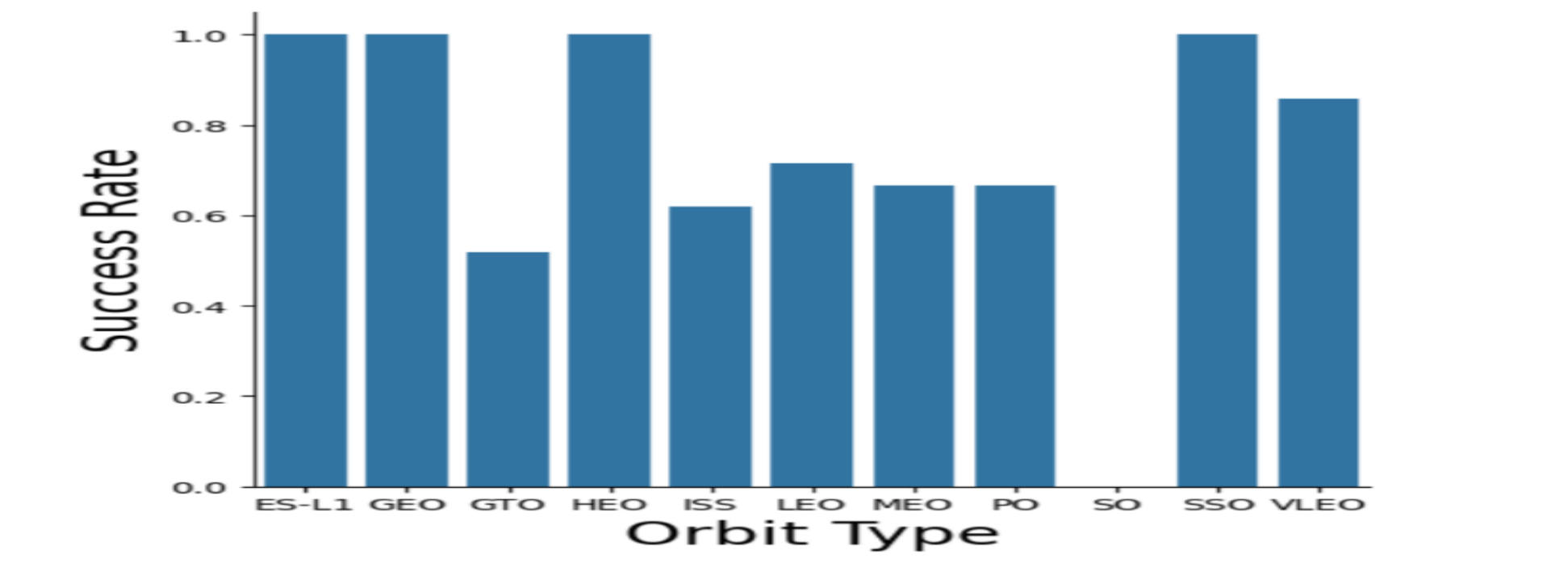
Payload vs. Launch Site



Typically, the higher the payload mass (kg), the higher the success rate

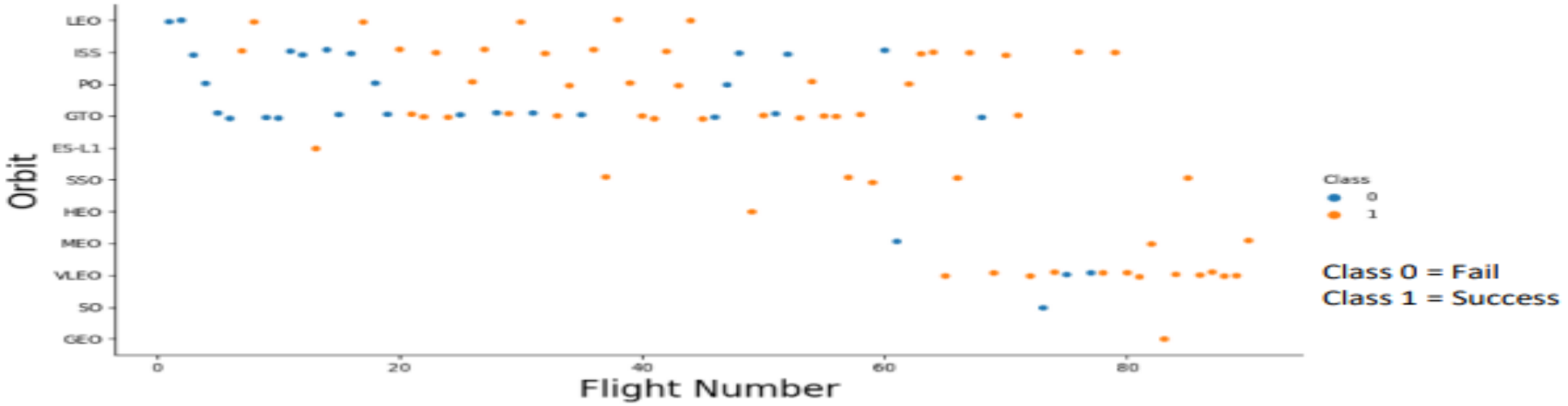
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg

Success Rate vs. Orbit Type



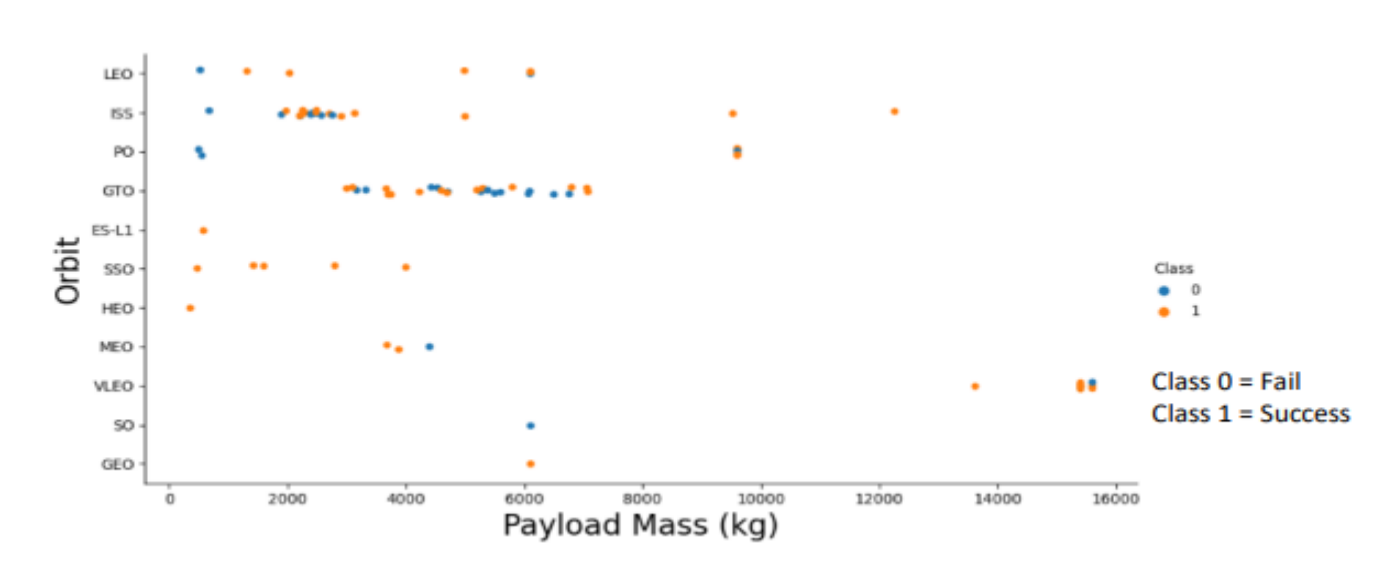
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO

Flight Number vs. Orbit Type



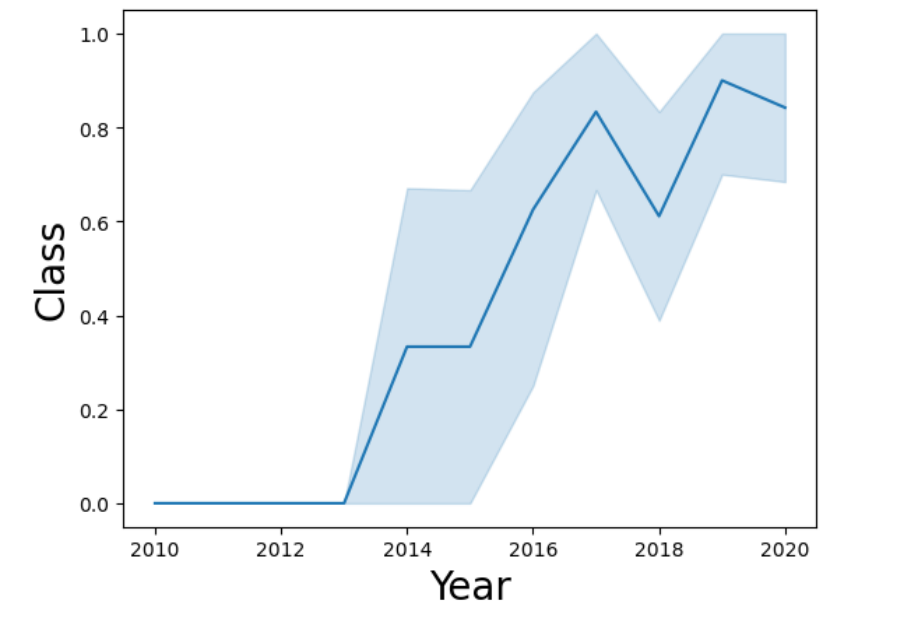
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend

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 Names

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

In [12]:
```
%sql select distinct "Launch_Site" from SPACEXTBL
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

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

In [13]:
```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

In [37]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

Out[37]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

In [39]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[39]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

Hint:Use min function

In [42]:
```
%sql SELECT MIN(DATE) \
    FROM SPACEXTBL \
    WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

Out[42]:

MIN(DATE)
2015-12-22

The first successful Ground landing occurred on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [43]:
```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Out[43]:

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

The above listed payloads are 4 boosters that have had success in drone ships with payload mass between 4k and 6k

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [44]:
```
%sql SELECT DISTINCT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Out[44]:

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

Boosters Carried Maximum Payload

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

In [45]:
```
%sql SELECT DISTINCT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[45]:

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 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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [48]:
```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

Out[48]:

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

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.

In [65]:
```
%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.
```

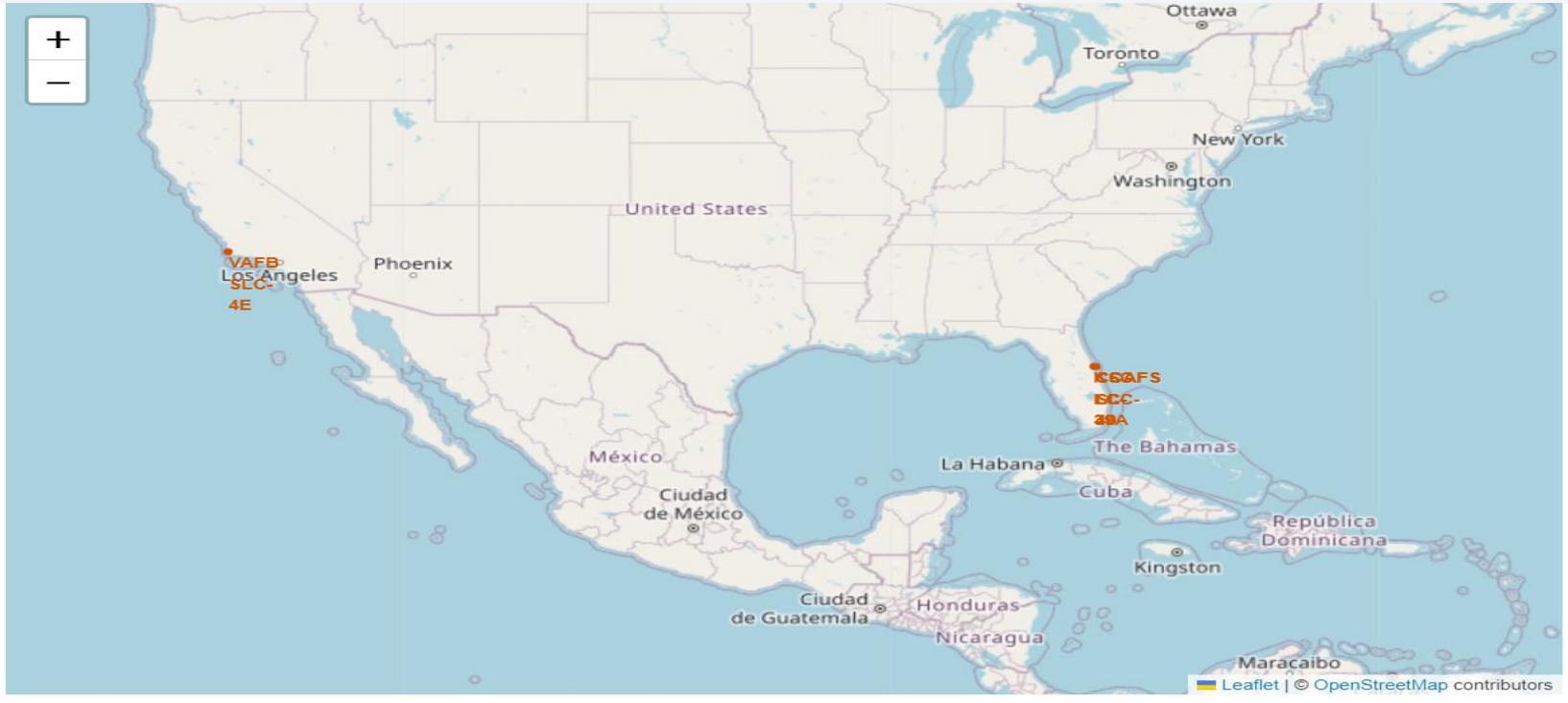
Out[65]:

Landing_Outcome	count_outcomes
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 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 Outcome

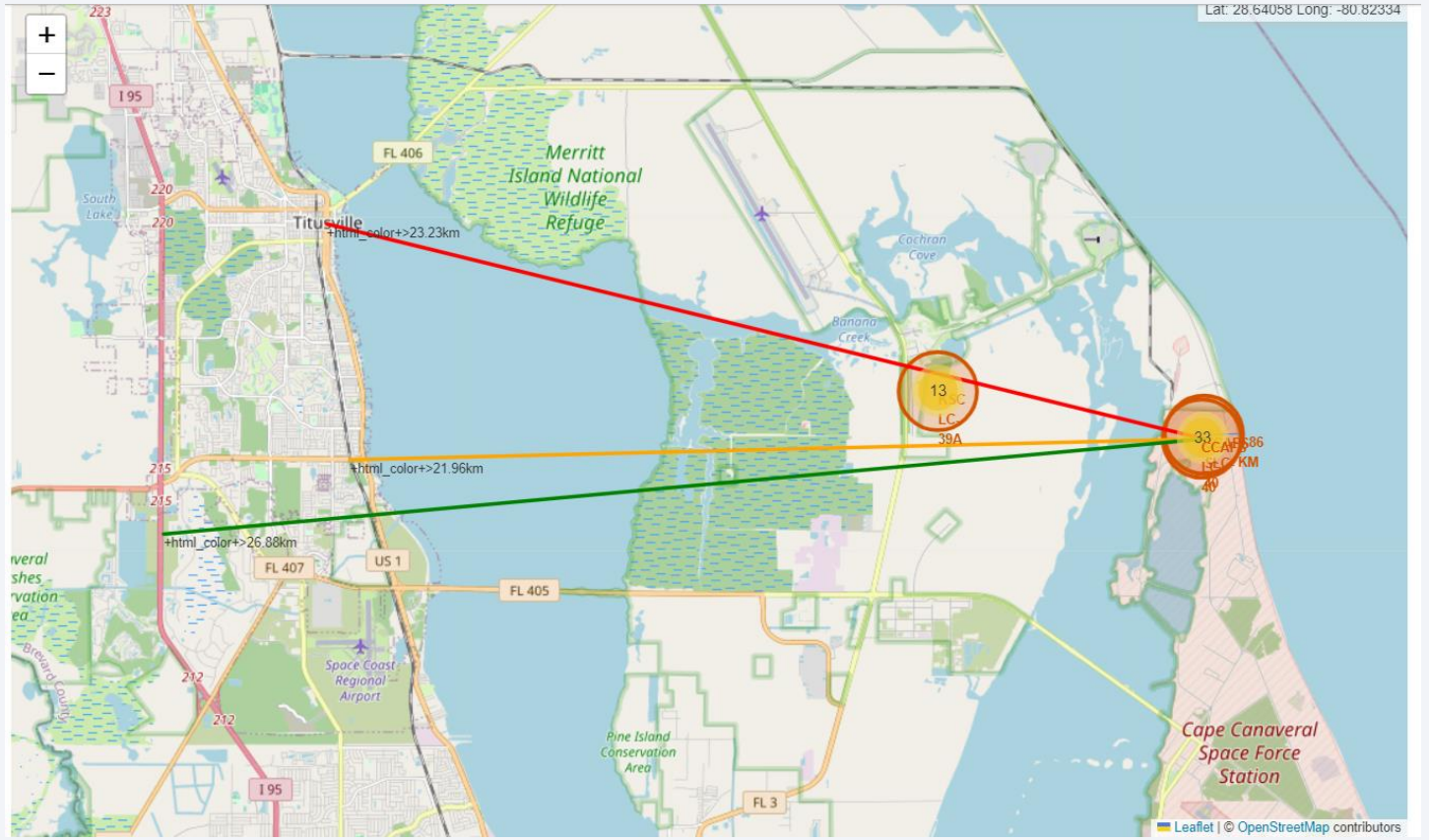
[15]:

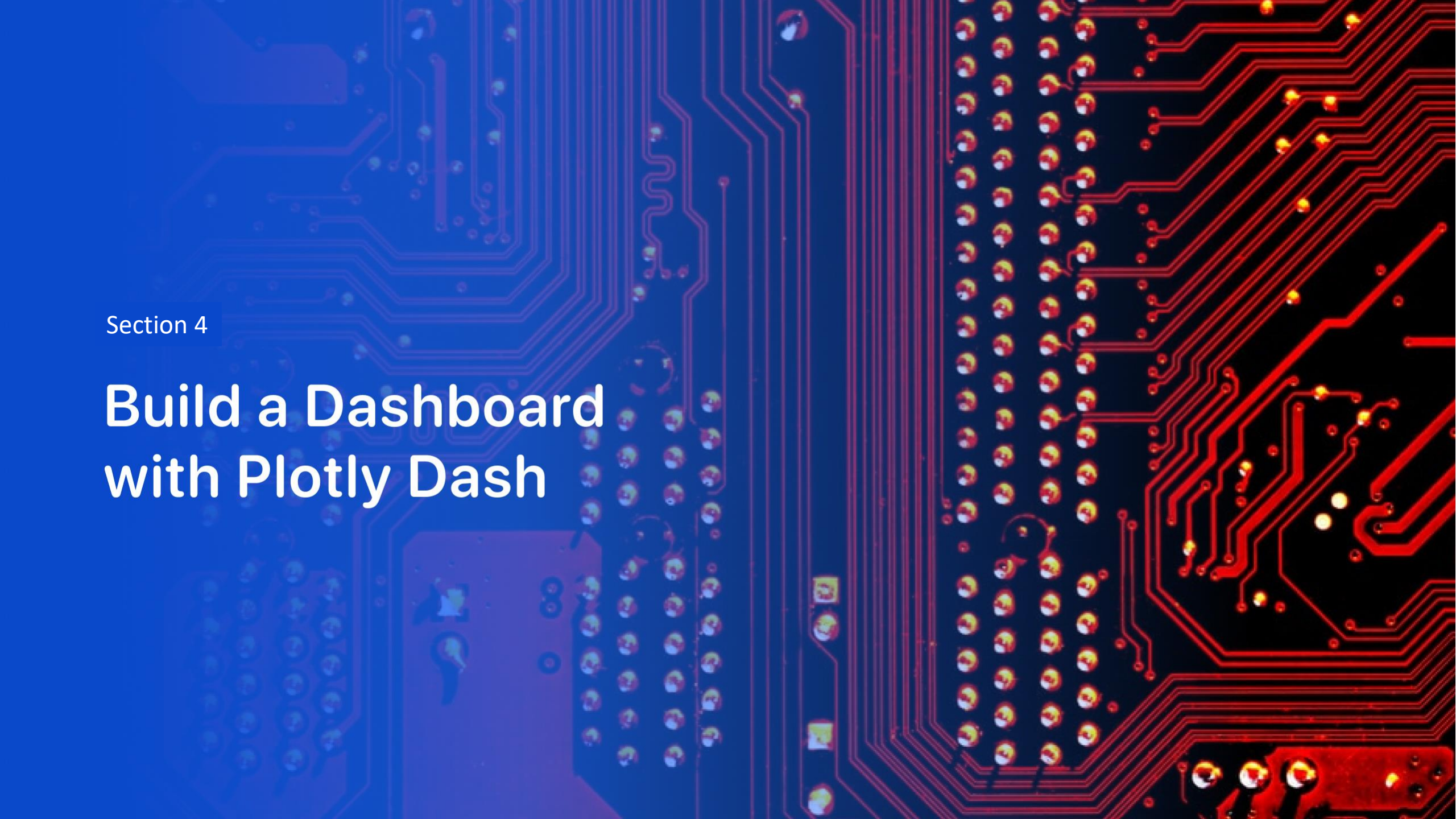
- Outcomes:
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

Distance to Proximities

- CCAFS SLC-40
- 86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway

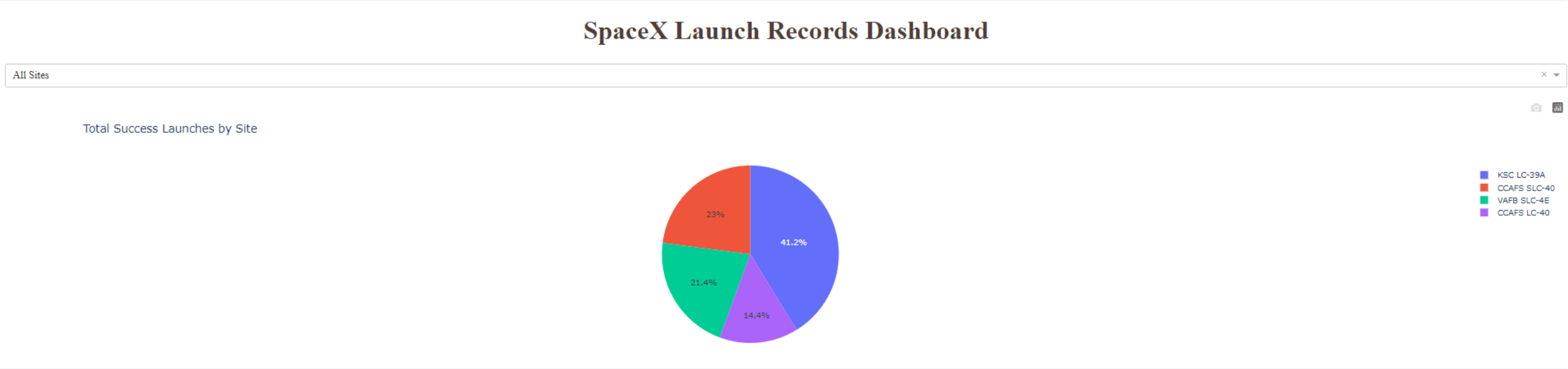
- Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- Safety / Security needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- Transportation/Infrastructure and Cities: need to 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

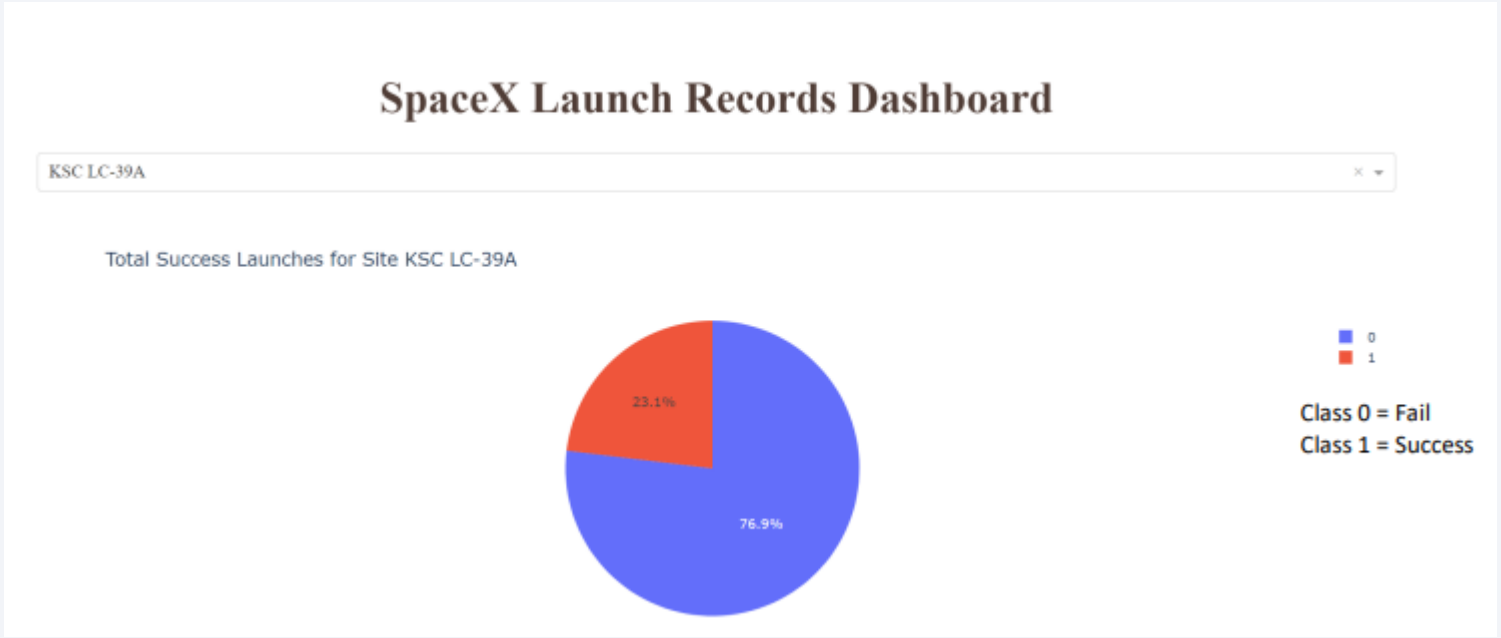
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

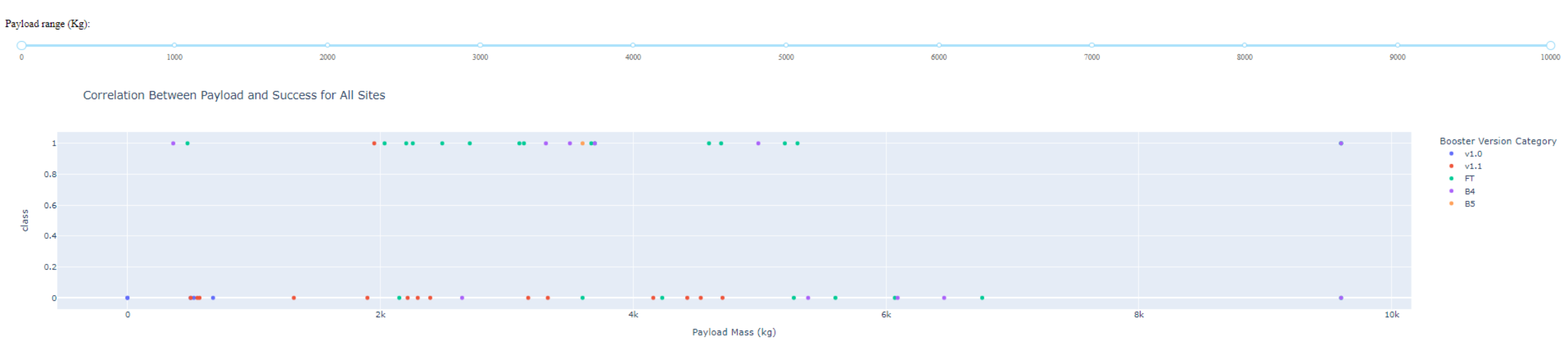
- KSC LC-39A has the highest success rate amongst launch sites (76.9%)

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

[90]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

[91]:

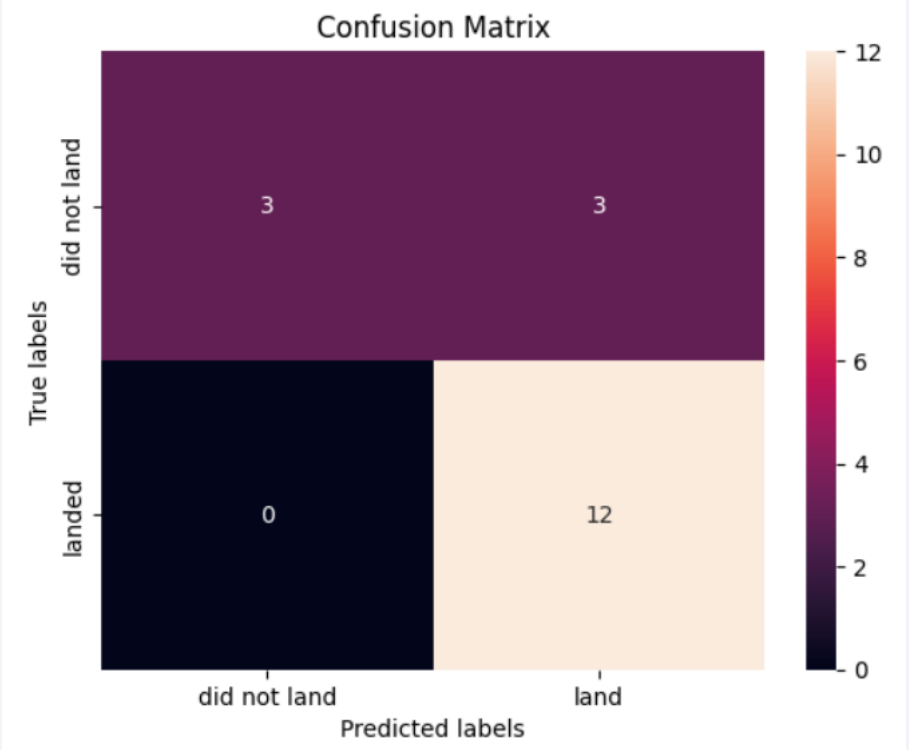
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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 0.875
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'best'}
```

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) = .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
- 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

Thank you!