Applied Data Science Capstone: SpaceY

Simeon Godwin

October 26, 2024

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Goal:
  - This research attempts to identify the features that contribute to a successful rocket landing.
- Methodologies
  - **Data Collection**: Data on launches was collected from SpaceX's REST API using web scraping techniques
  - **Data Wrangling**: Data was converted to provide meaningful feature values for analysis
  - **EDA**: Python data visualization and SQL techniques were used to explore the features and engineer them for launch success prediction
  - **Interactive Visualization**: Folium and Plotly Dash were leveraged to generate interactive, engaging visualizations about launches
  - **Predictive Analytics**: Multiple Machine Learning techniques were applied to key features in the data to predict the success of a launch
- Results
  - Overall, launch success improved as more launches were conducted
  - The ML models performed similarly on the test set, with relatively high accuracy
  - Certain launch sites had more success depending on the payload range

INTRODUCTION

Background

SpaceX is a leading innovator in the space travel industry that is able to perform many launches relatively inexpensively due to the reusability of their rocket, the Falcon 9. Our aim is to predict if the first stage of a launch of the Falcon 9 will land successfully, which is crucial to determining the cost of a launch.

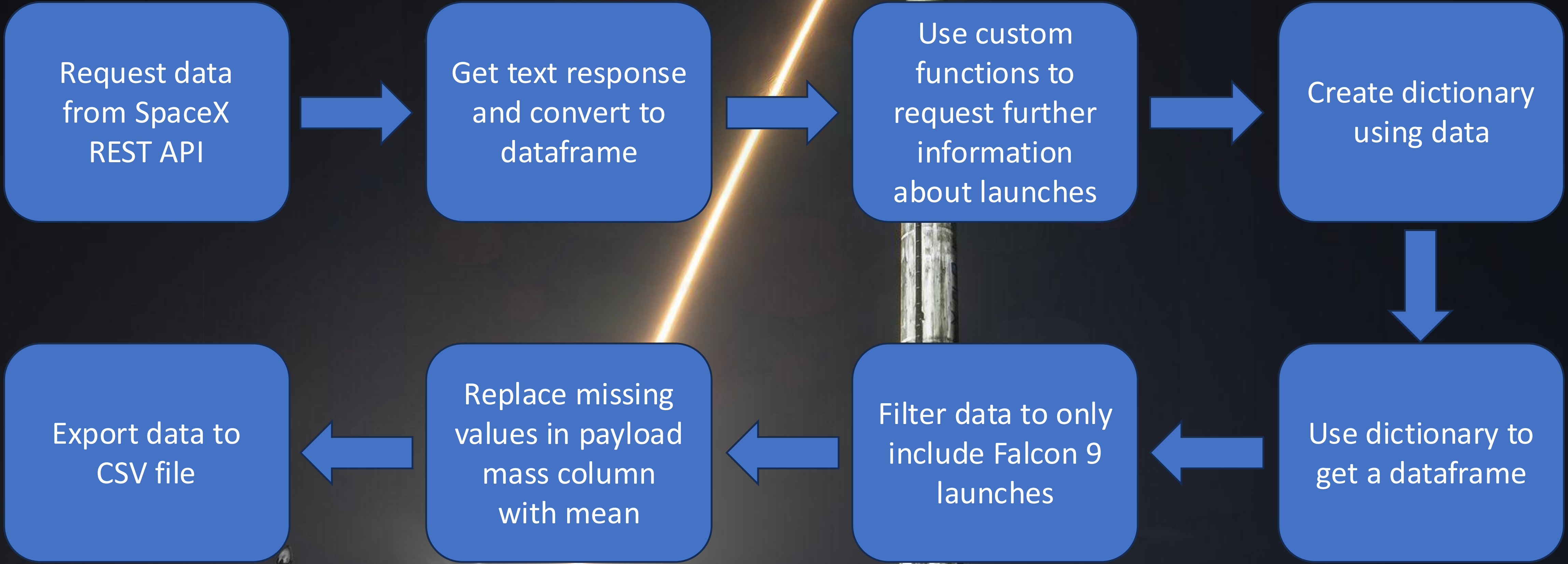
Topics of Exploration:

- How do payload mass, launch site, number of flights, and orbit affect the success of first-stage landing?
- How does the rate of successful landings change over time?
- What is the best predictive algorithm that can be used for the binary classification of a successful landing?

METHODOLOGY

- **Data Collection**: SpaceX has data on launches and their outcomes available through their REST API. Using web scraping, we pulled this data and filtered it down to data for the Falcon 9 rocket.
- **Data Wrangling**: The feature we are trying to predict is successful launches, but the data had multiple labels for launch outcomes. In order to conduct a binary classification, we converted this feature to binary labels of 0 or 1, indicating whether a landing was successful or not.
- **EDA**: First, we used SQL commands to get summaries of the features in the data. Then we also leveraged Python libraries to get informative visualizations and conduct feature engineering on the data.
- **Interactive Visualization**: We started by using Folium to interactively mark launch sites and launch outcomes on a map, also marking important locations around the sites. Then, we created an interactive dashboard using Plotly to show the launch outcomes for a chosen site or all sites in relation to payload mass.
- **Predictive Analytics**: We split the data into train and test sets and fit multiple models to the training portion: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors. We then evaluated these models on the test set to see which performed best for the binary classification.

Data Collection – REST API



Request data from SpaceX REST API

Get text response and convert to dataframe

Use custom functions to request further information about launches

Create dictionary using data

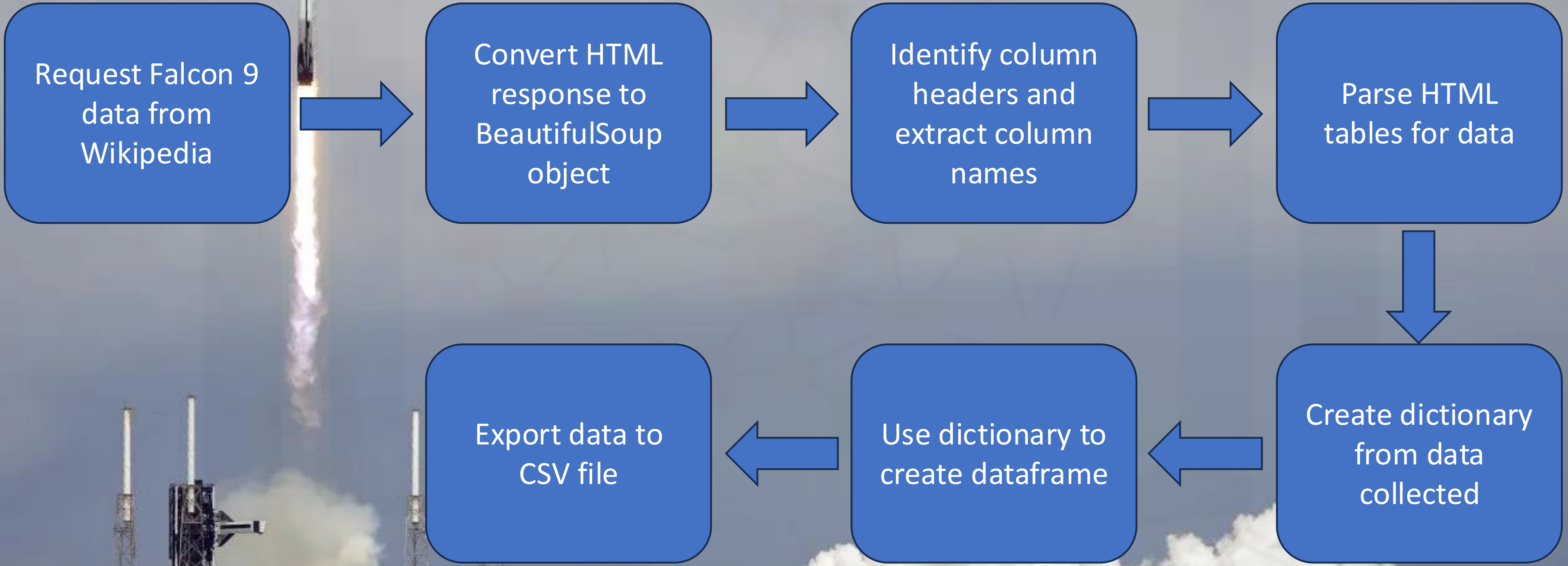
Use dictionary to get a dataframe

Filter data to only include Falcon 9 launches

Replace missing values in payload mass column with mean

Export data to CSV file

Data Collection – Web Scraping



Request Falcon 9 data from Wikipedia

Convert HTML response to BeautifulSoup object

Identify column headers and extract column names

Parse HTML tables for data

Create dictionary from data collected

Use dictionary to create dataframe

Export data to CSV file

Data Wrangling

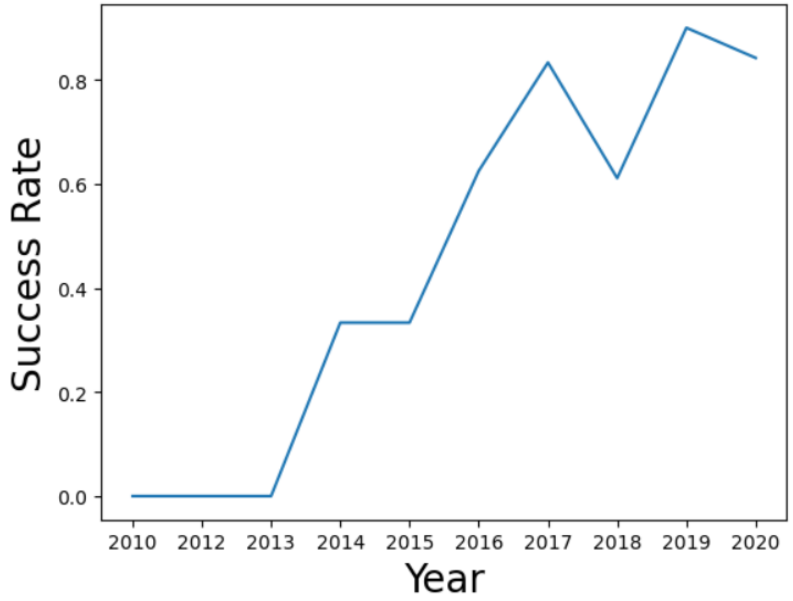
- Conduct preliminary EDA and identify feature labels
- Calculate frequencies of the following features:
  - Launches by site
  - Launches by orbit type
  - Launches and mission outcome by orbit type
- Add outcome column that is binary depending on successful landing
- Landing outcome processing:
  - 6 types of landing outcomes
  - True Ocean: class 1
  - False Ocean: class 0
  - True RTLS: class 1
  - False RTLS: class 0
  - True ASDS: class 1
  - False ASDS: class 0

EDA - SQL

- Execute queries to get the following results:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - 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. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015
  - 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

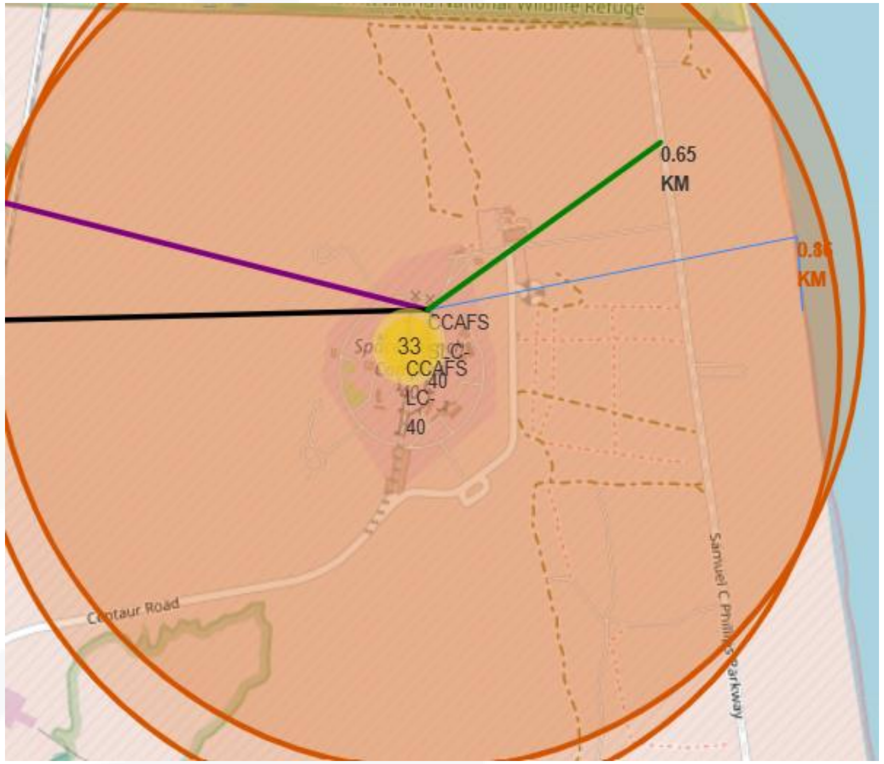
EDA – Data Visualization

- Create visualizations in Python to achieve the following:
  - Visualize the relationship between Flight Number and Launch Site
  - Visualize the relationship between Payload and Launch Site
  - Visualize the relationship between success rate of each orbit type
  - Visualize the relationship between FlightNumber and Orbit type
  - Visualize the relationship between Payload and Orbit type
  - Visualize the launch success yearly trend
- Feature Engineering
  - Create dummy variables to categorical columns
  - Cast all numeric columns to float64



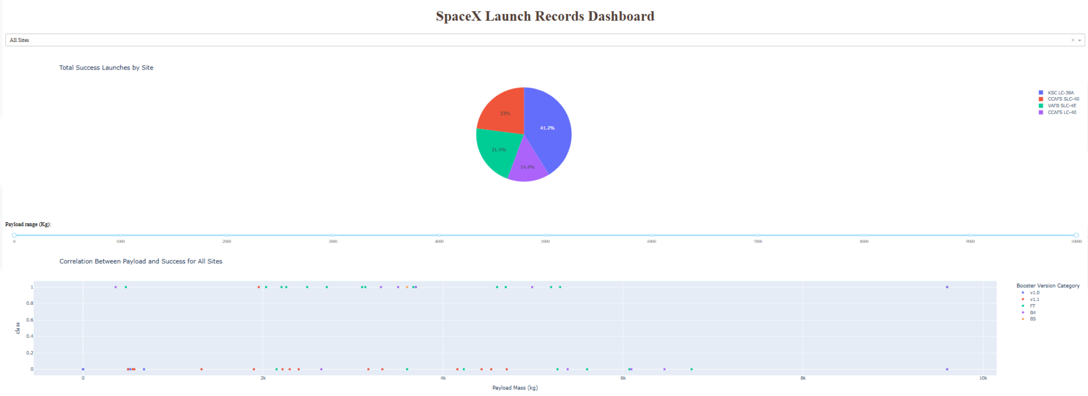
Interactive Visual Analytics– Folium

- Use Folium on a map to mark the following:
  - Mark all launch sites on a map
    - Add colored map object to indicate location of NASA JSC
    - Add more colored markers at all launch site coordinates with popups showing their names
  - Mark the success/failed launches for each site on the map
    - Add colored markers for each launch with successful launches colored green and unsuccessful launches colored red
  - Calculate the distances between a launch site to its proximities
    - Add a marker labeling the closest coastline to a launch site and draw a line between them with the distance as a label
    - Add a marker labeling the closest highway and draw a line between them with the distance as a label
    - Add a marker labeling the closest highway and draw a line between them with the distance as a label
    - Add a marker labeling the closest railway and draw a line between them with the distance as a label
    - Add a marker labeling the closest city and draw a line between them with the distance as a label

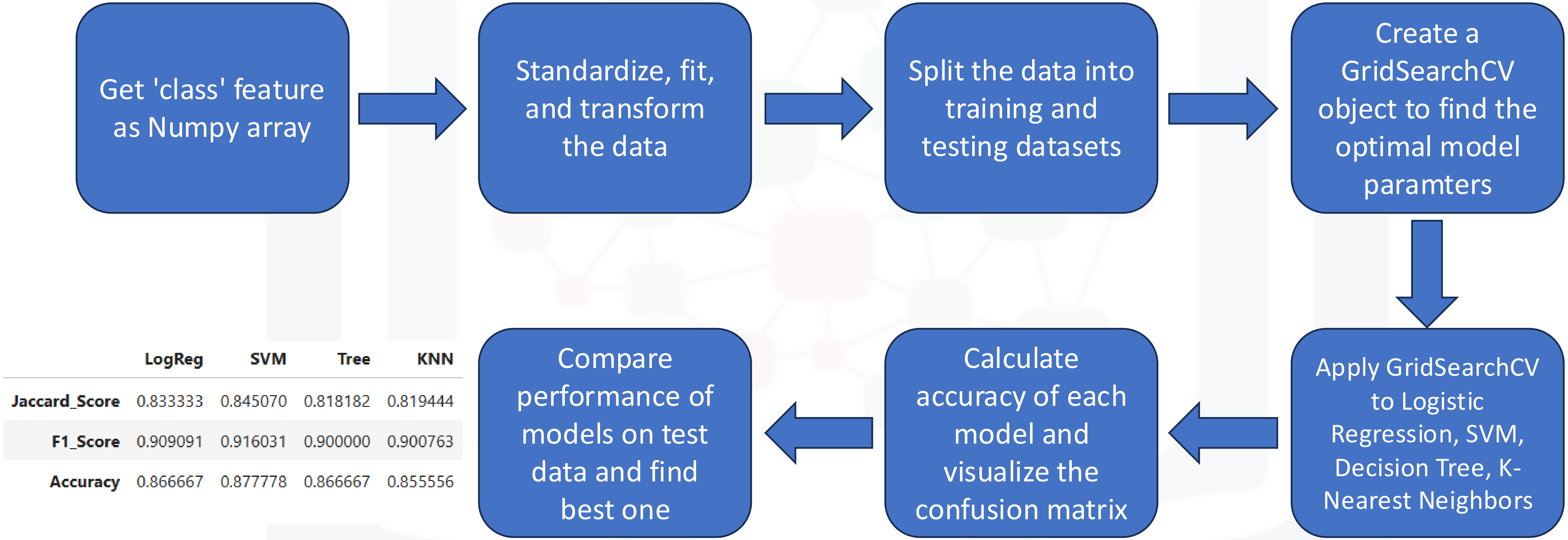


Interactive Dashboard – Plotly Dash

- Create an interactive dashboard summarizing the data with the following components:
  - Launch Site Dropdown
    - Allow users to select if they want to view data for all sites or a specific site
  - Pie Chart of Successful Launches
    - Display an interactive pie chart that shows the proportion of successful launches for the selected sites
  - Payload Mass Range Slider / Scatter Plot
    - Display a scatterplot with payload mass on the x axis and the value of the 'class' column color coded by booster version on the y axis (successful launch or not)
    - Allow users to adjust the range of payload mass on a slider



Predictive Analytics



Get 'class' feature as Numpy array

Standardize, fit, and transform the data

Split the data into training and testing datasets

Create a GridSearchCV object to find the optimal model paramters

Apply GridSearchCV to Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors

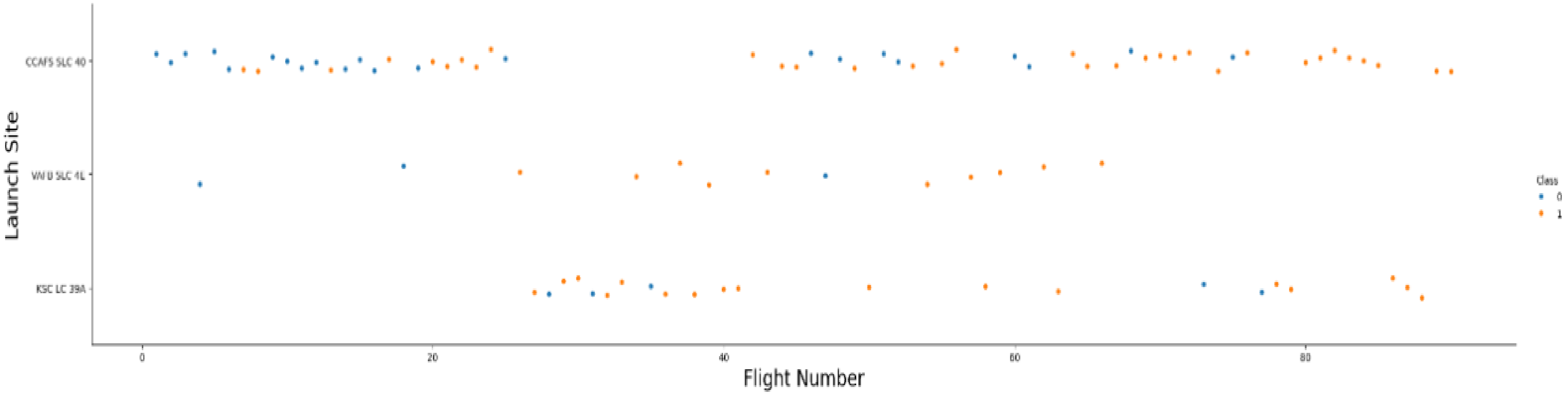
Calculate accuracy of each model and visualize the confusion matrix

Compare performance of models on test data and find best one

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.818182	0.819444
F1_Score	0.909091	0.916031	0.900000	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

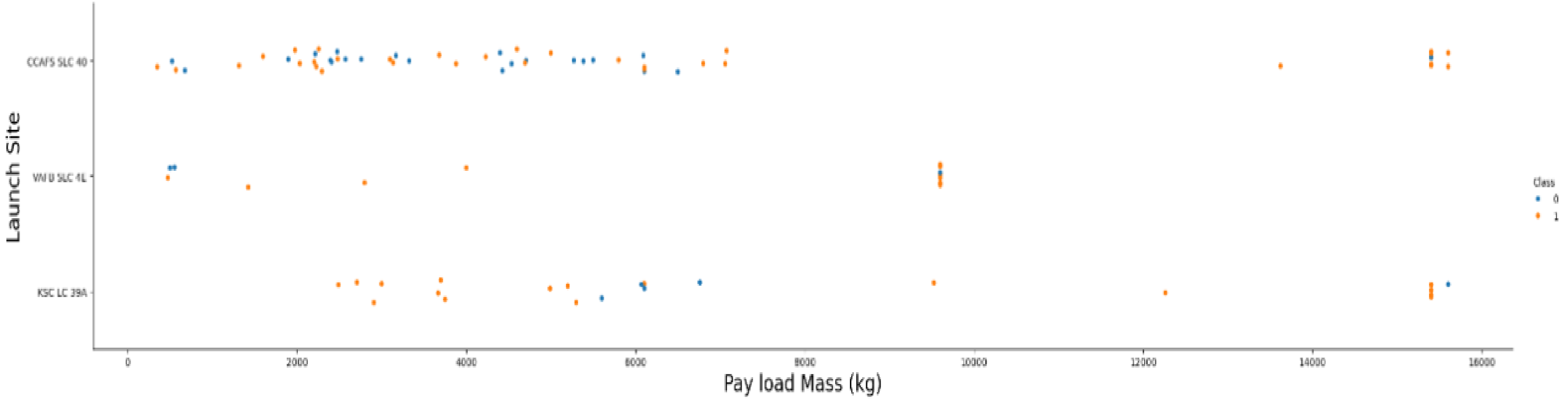
RESULTS

Flight Number vs. Launch Site



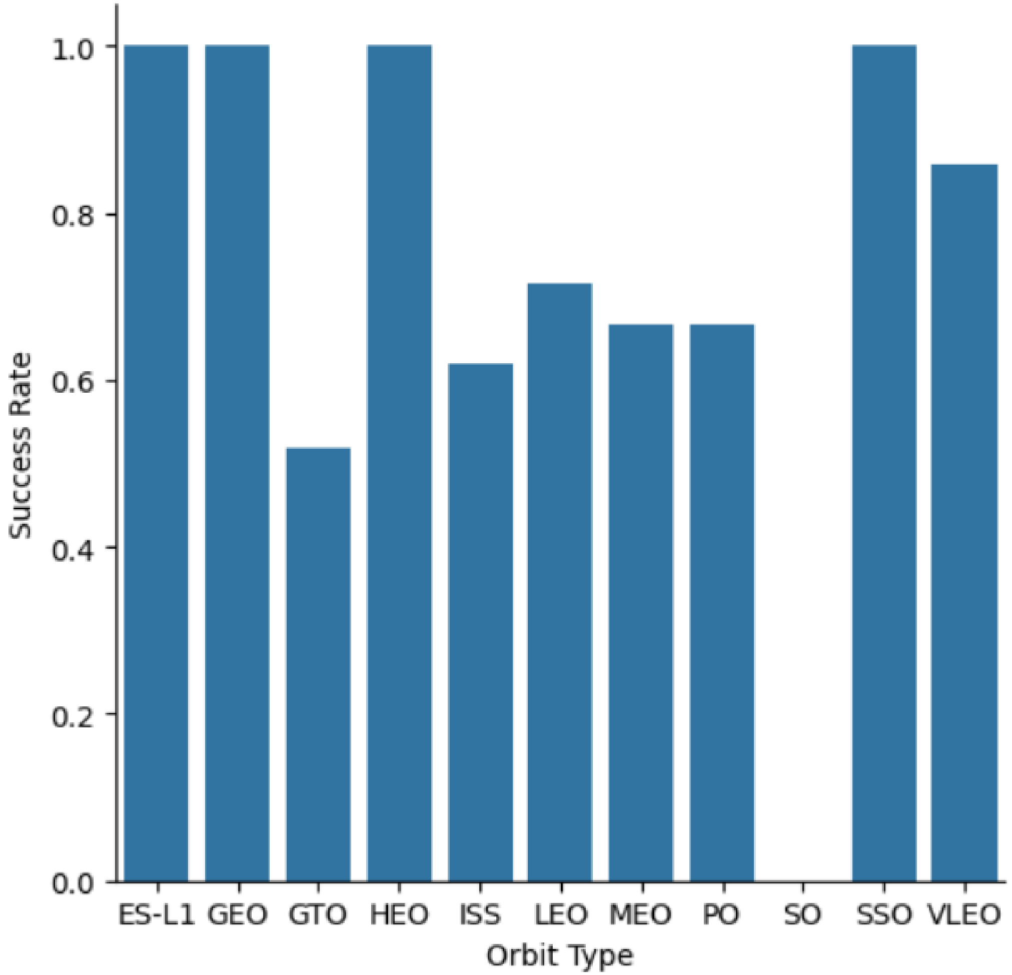
- Blue markers indicate failed launches while orange markers represents successful ones
- For all of these sites, the success rate increased as the number of flights increased
- This could mean that more flight experience and technical adjustments leads to greater launch success

Payload Mass vs. Launch Site



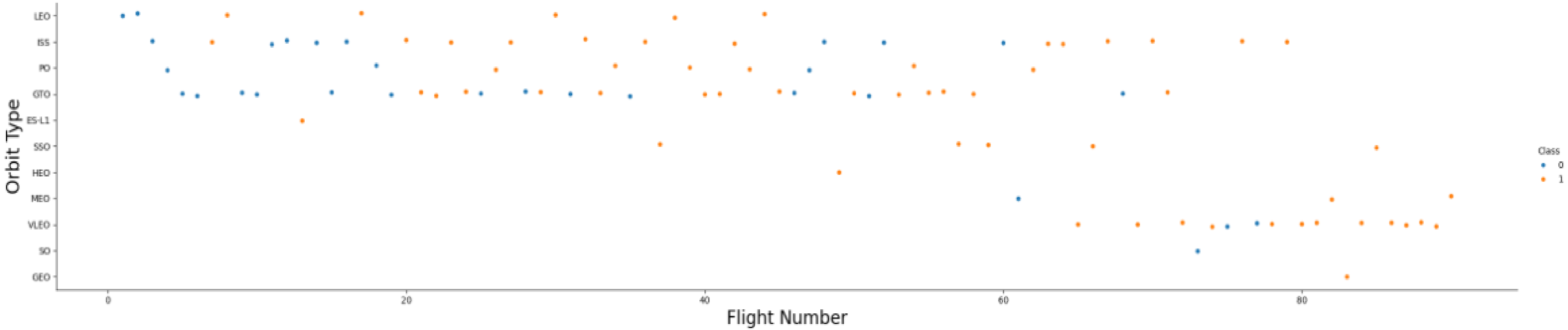
- Launch site CCAFS SLC 40 and KSC LC 39A had higher payload masses at the top end of their spectrum
  - Both these sites had very high success rates when the payload mass was above 12000kg
- Overall, the success rate at all sites was higher when the payload mass was above average
- This could mean that higher payload masses are conducive to more stable or successful landings

Success Rate by Orbit Type



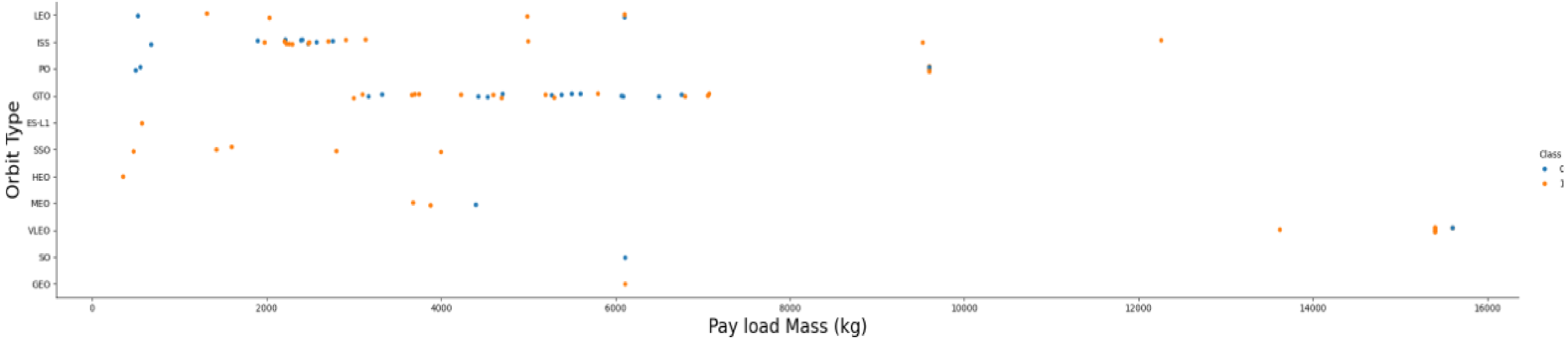
- The orbits ES-L1, GEO, HEO, and SSO had all successful launches
- This could either mean that these orbits are safer for launches by nature, or that there have not been enough launches in these orbits to see failures

Flight Number vs. Orbit Type



- The success rate is higher when there is a higher flight number, across all orbit types
- The orbits VLEO, MEO, SSO, and ISS seem to have significantly higher flight numbers than the other orbits, along with a very high rate of success
    - This could mean that these orbits are much more tested and stable because of their high number of launches

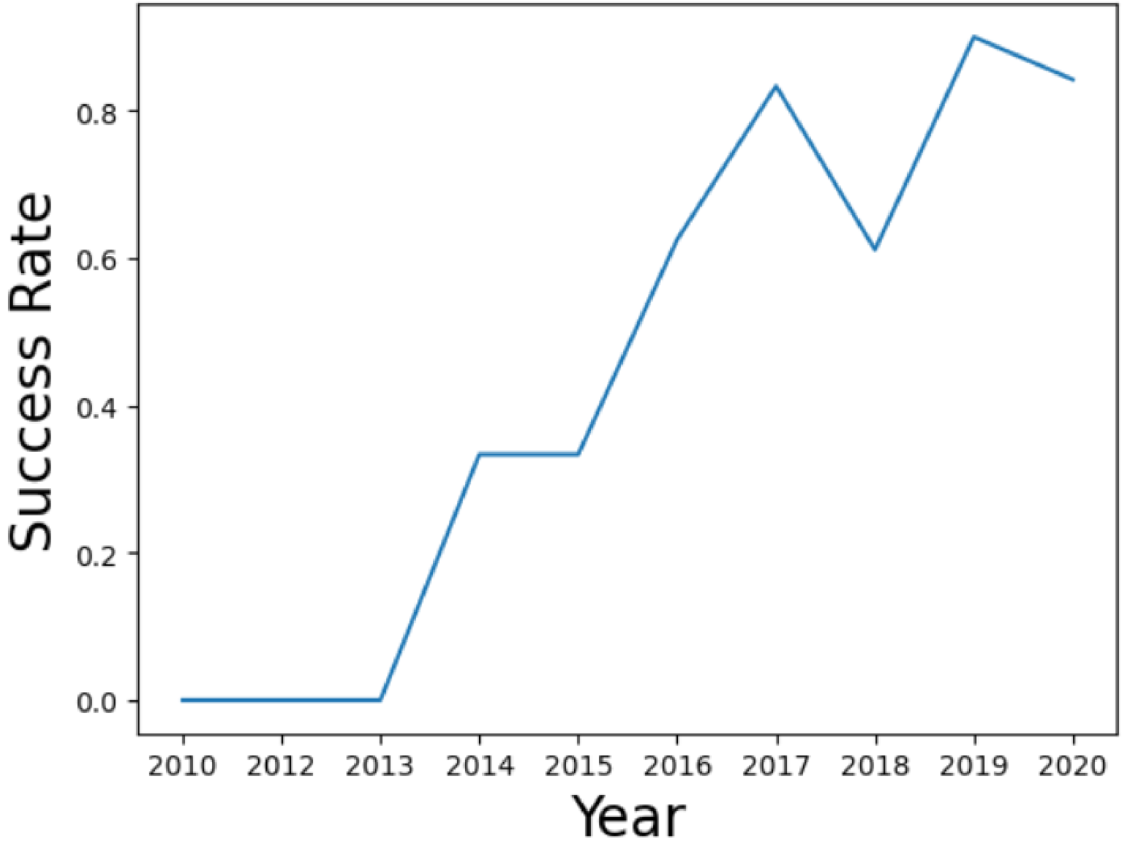
Payload Mass vs. Orbit Type



- The orbit VLEO has much higher pay load masses than the other orbits, and also a high success rate
- Similarly, ISS and PO have good success rates with high payloads
- However, the success rate is low for lighter payloads with LEO, ISS, and PO
- This could confirm that heaver payloads are conducive to more stable and successful launches

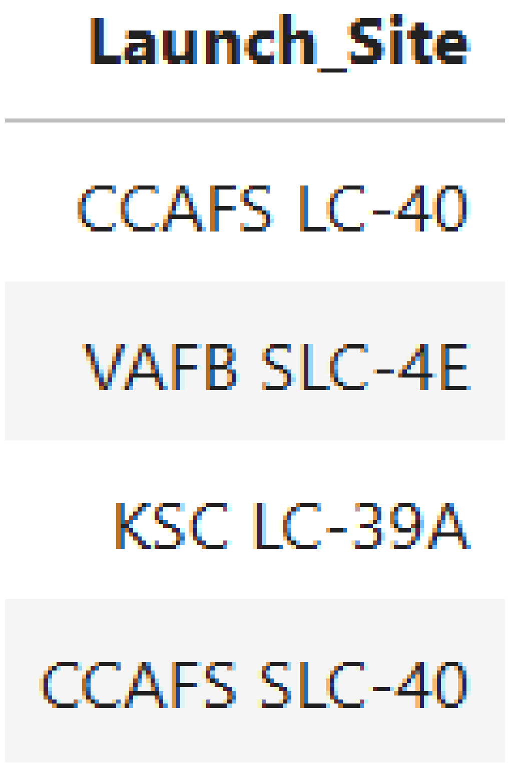
Launch Success Yearly Trend

- The success rate of launches increases overall during the time period from 2013 to 2020
- While there are a few dips in success rate, it seems that on the whole launches could be becoming more successful over time due to factors like improved technology, innovation, and launch experience



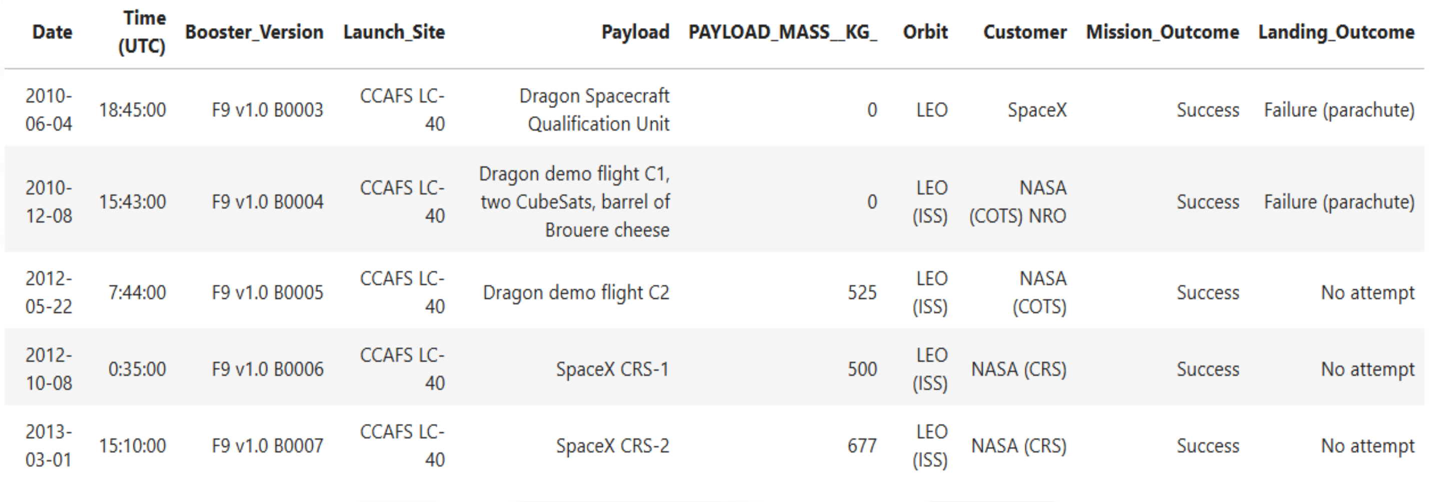
Launch Site Info

- List of Launch Sites:



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

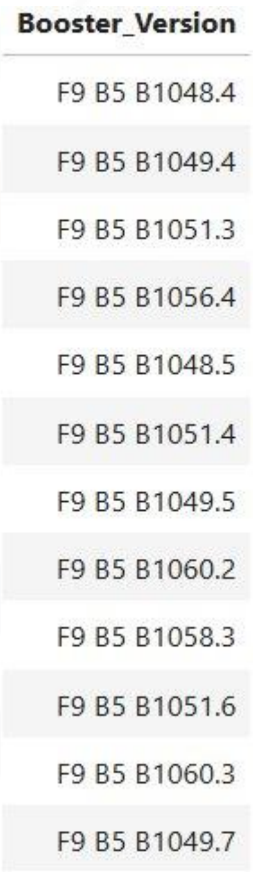
5 Records of Launch Sites Beginning with 'CCA':



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

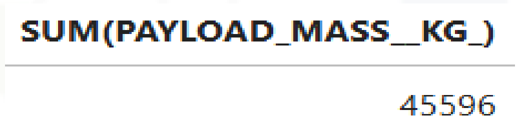
Payload Mass Info

A total of 12 booster versions have carried that maximum payload mass in the data



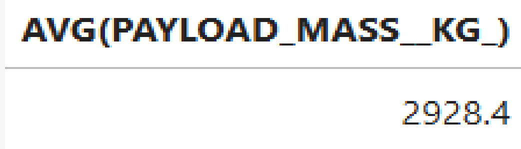
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

The total payload mass of all the launches is 45,596 kg



SUM(PAYLOAD_MASS__KG_)
45596

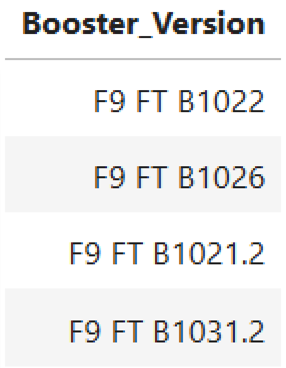
The average payload mass of a launch is 2,928.4 kg



AVG(PAYLOAD_MASS__KG_)
2928.4

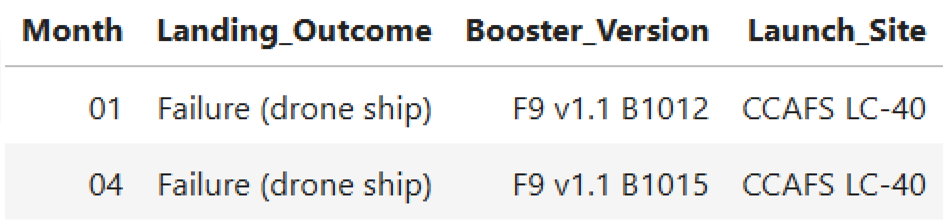
Landing Outcomes

4 boosters have success in drone ship and have payload mass greater than 4000 but less than 6000:



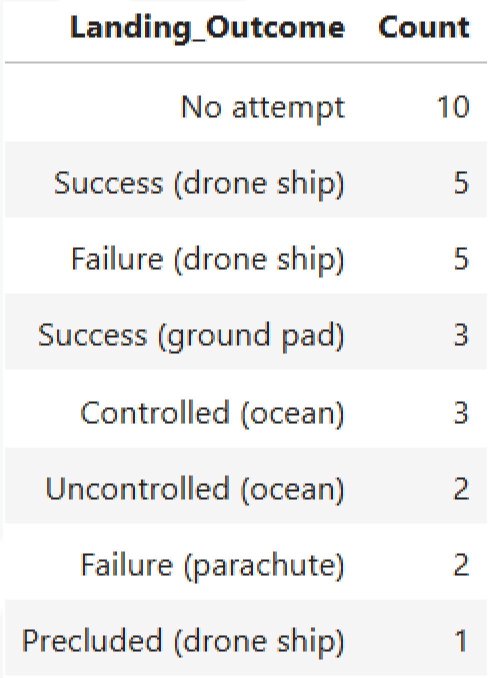
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015:



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

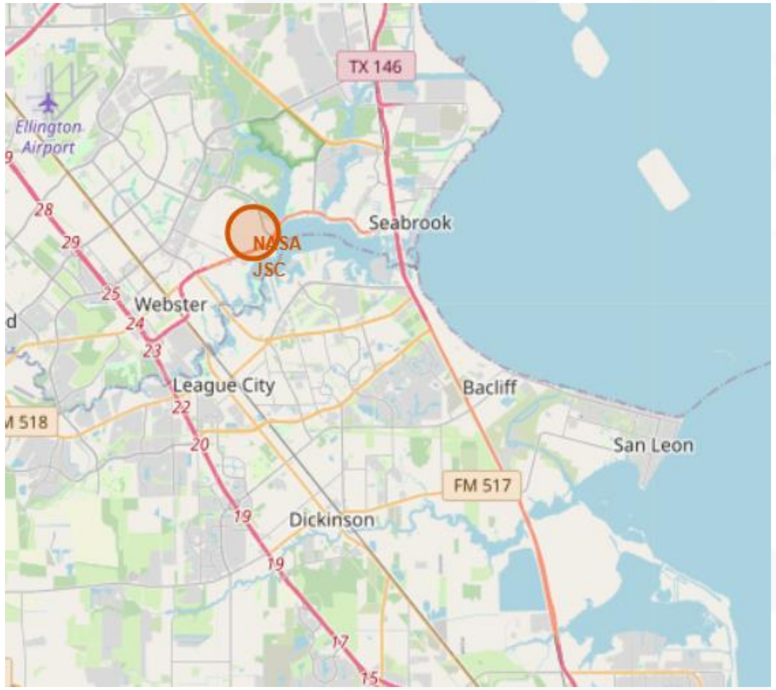
Count of landing outcomes between the date 2010-06-04 and 2017-03-20:



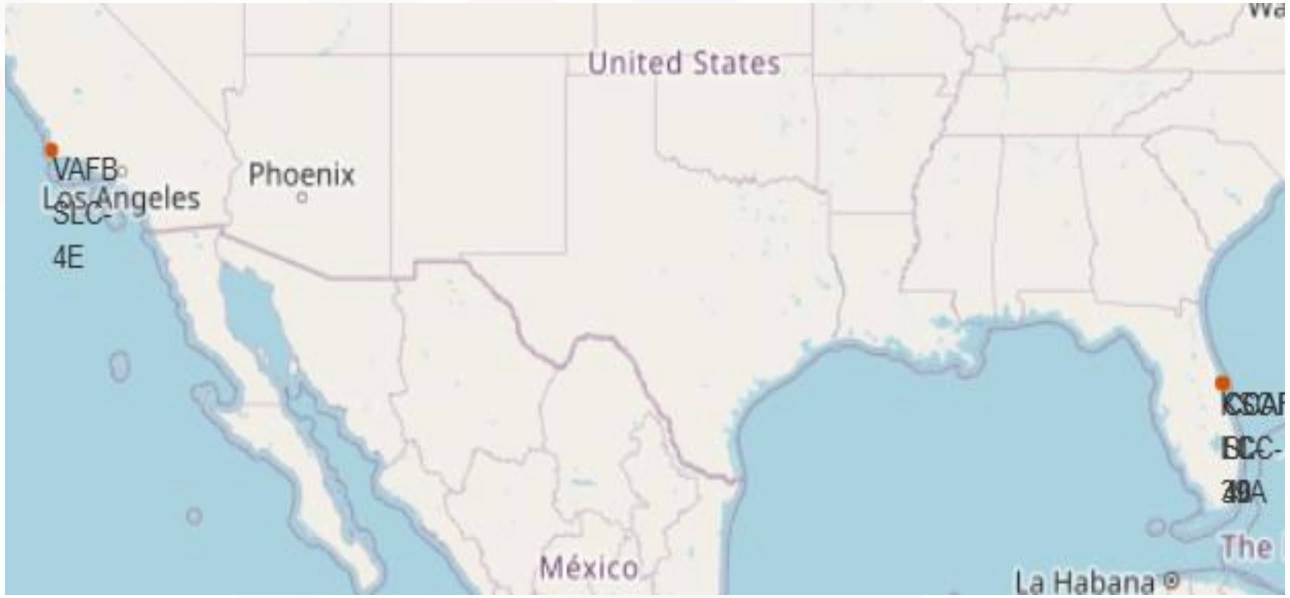
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Mapping Important Sites

Start our map marker at NASA JSC:

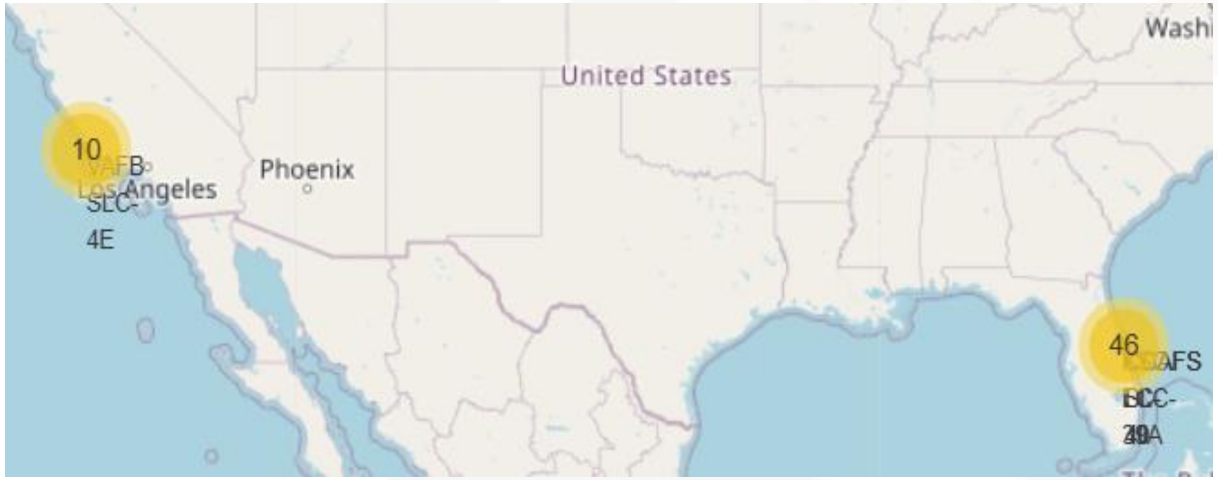


Mark and label each launch site from the data on out map:

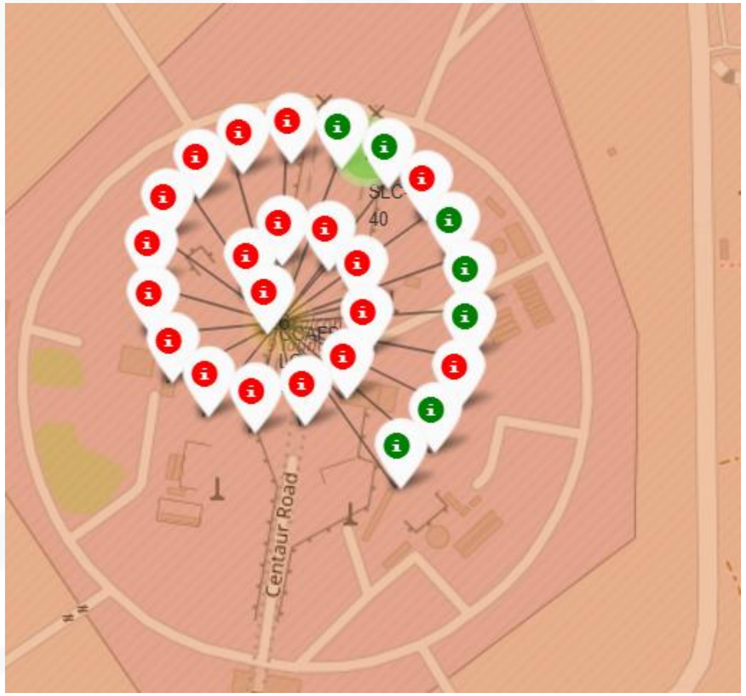


Marking Successful/Failed Launches

For each site, we can see the total number of launches and mark each as successful or unsuccessful:

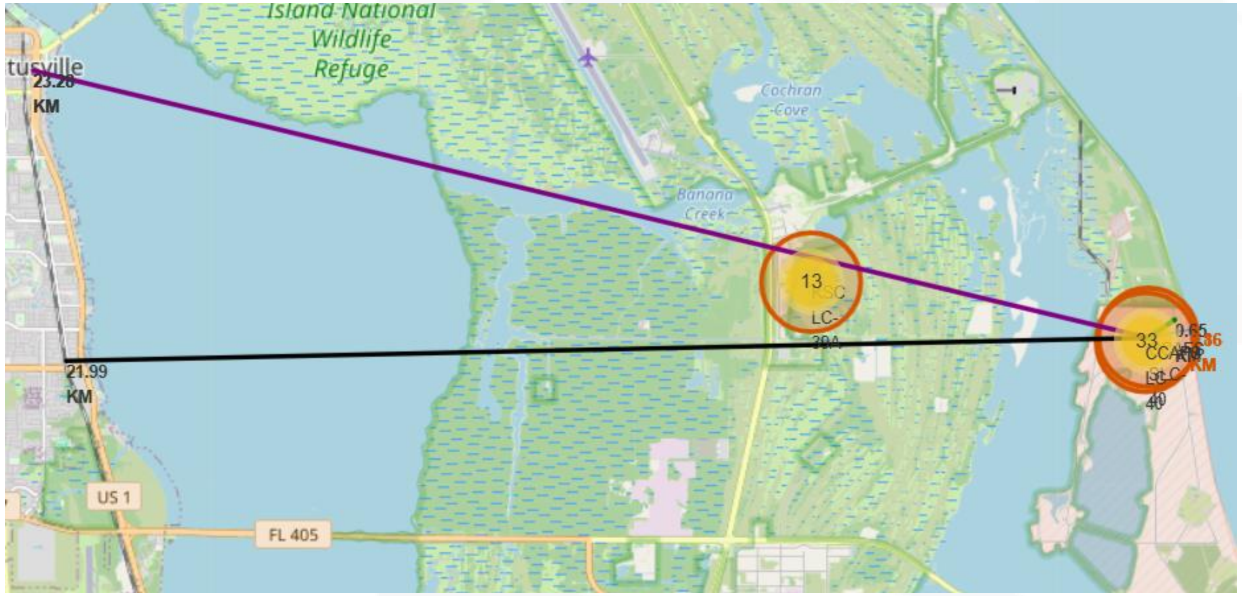


For example, here we can see that at CCAFS LC-40 there is a high proportion of unsuccessful launches. In contrast, at the close by site CCAFS SLC-40, the proportion of unsuccessful launches is much lower.

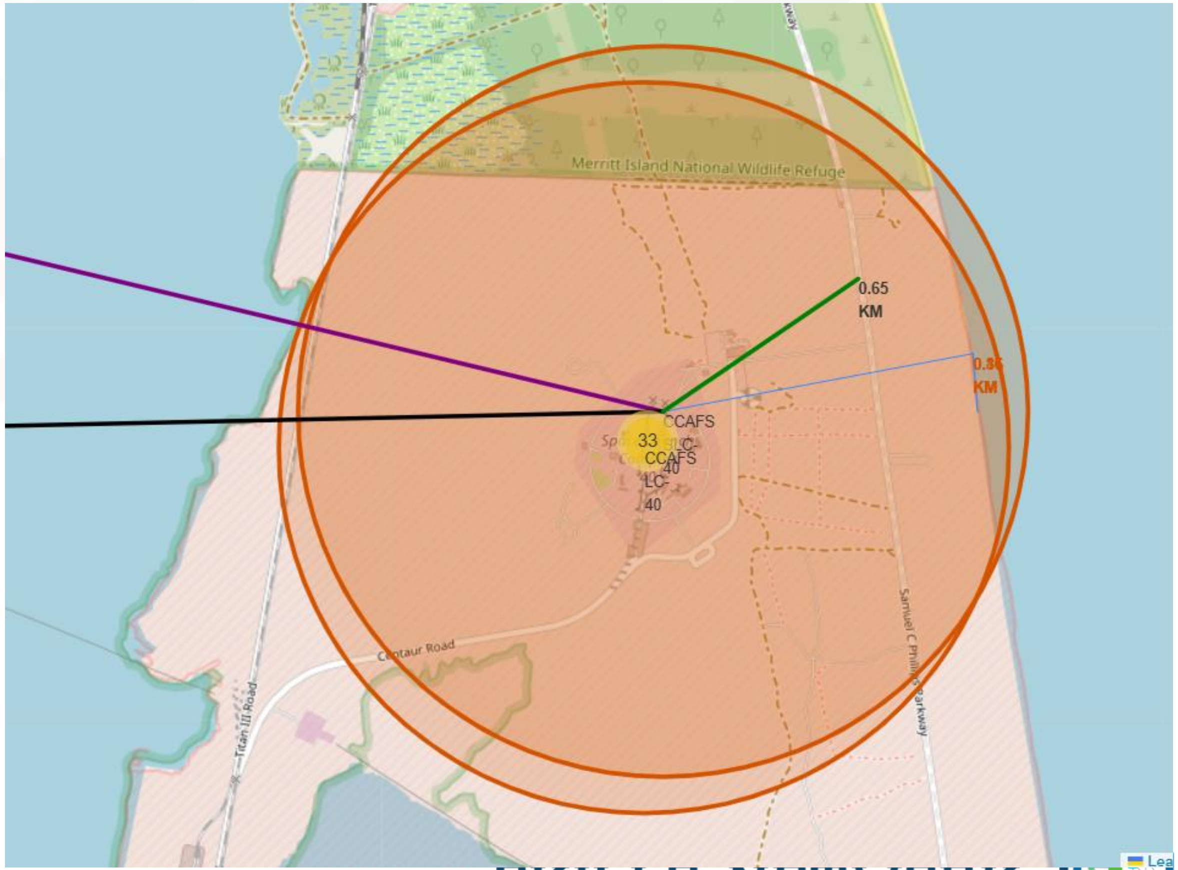


Launch Site Proximities

We can plot the distance from launch site CCAFS SLC-40 to the nearest coastline, railway, city, and highway, indicating them with lines:



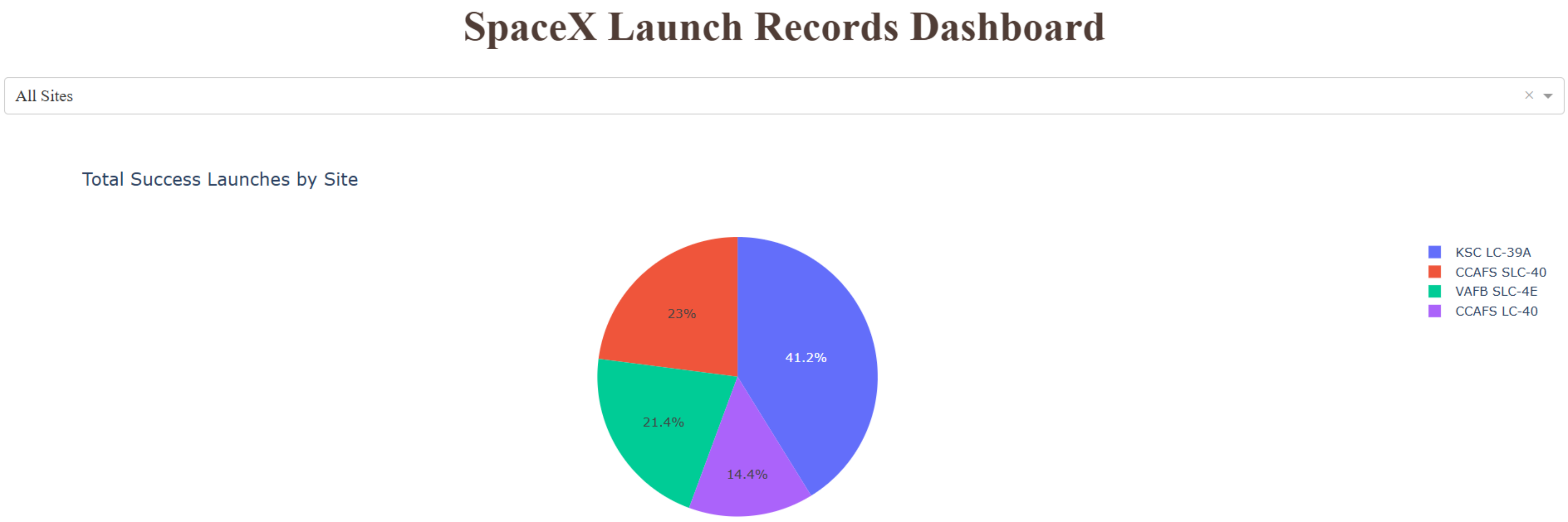
As we can see below, the launch site is relatively close to the coastline and highway, but it keeps a much greater distance from railways and cities.



Merritt Island National Wildlife Refuge
0.65
KM
0.86
KM
CCAFS
33
SLC-
40
CCAFS
LC-
40
Centaur Road
Titan III Road
Samuel C Phillips Parkway

Successful Launches for All Sites

This pie chart shows that the site KSC LC-39A is responsible for the highest number of successful launches out of all the sites, while site CCAFS LC-40 has the smallest number

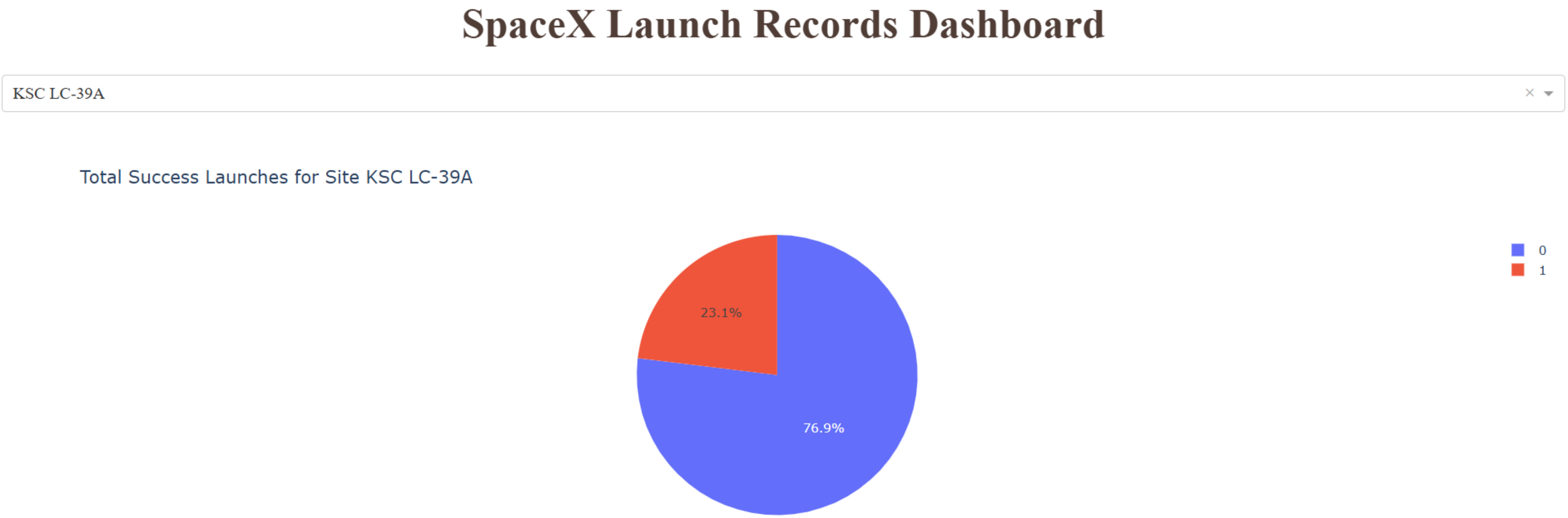


SpaceX Launch Records Dashboard

All Sites × ▾

Successful Launches for One Site

If we take a look at the site with the highest number of successful launches, we see that it had a very low failure rate of 23.1%, which means that it is a relatively reliable site for successful launches.



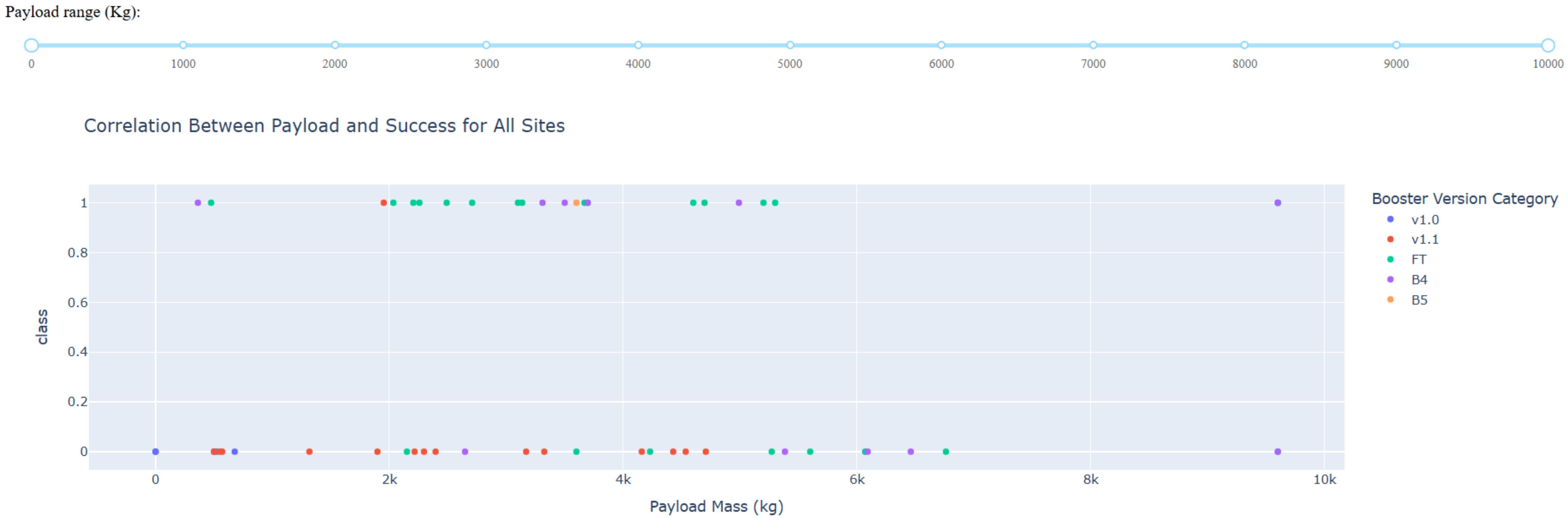
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

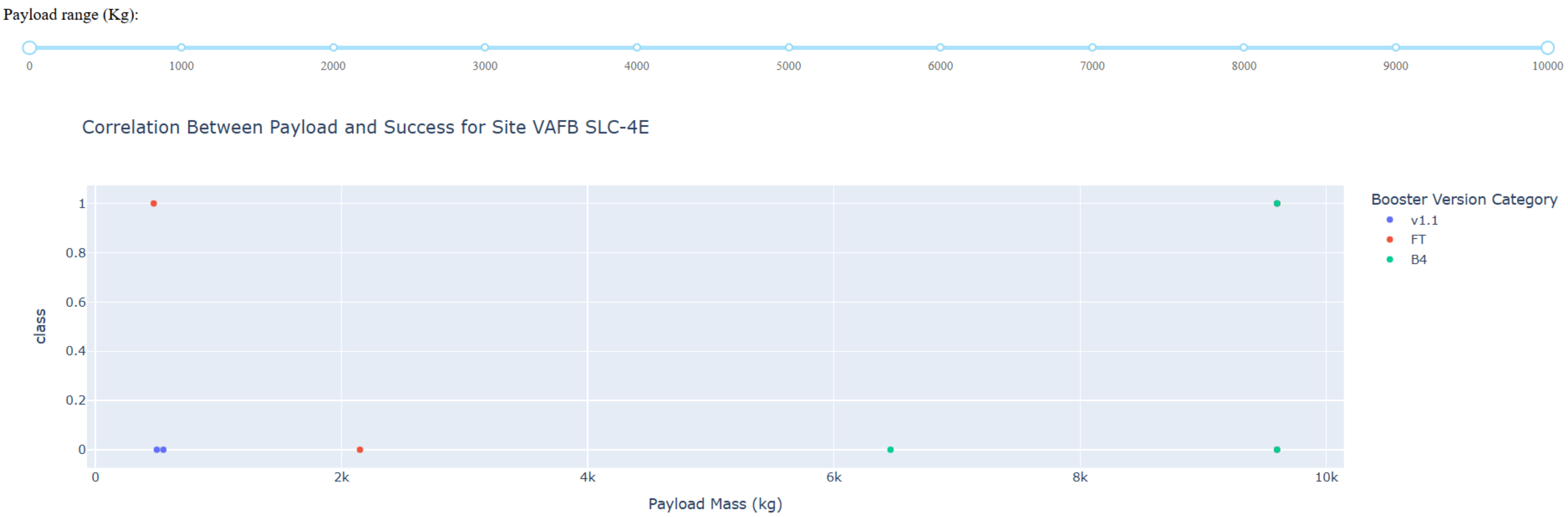
Successful Launches vs Payload Mass (All Sites)

The scatterplot shows that across all the sites, the FT booster version had a high rate of success for all payloads while the v1.1 booster had a very low rate of success for all payloads. Additionally, higher payload had lower success rates across the sites.



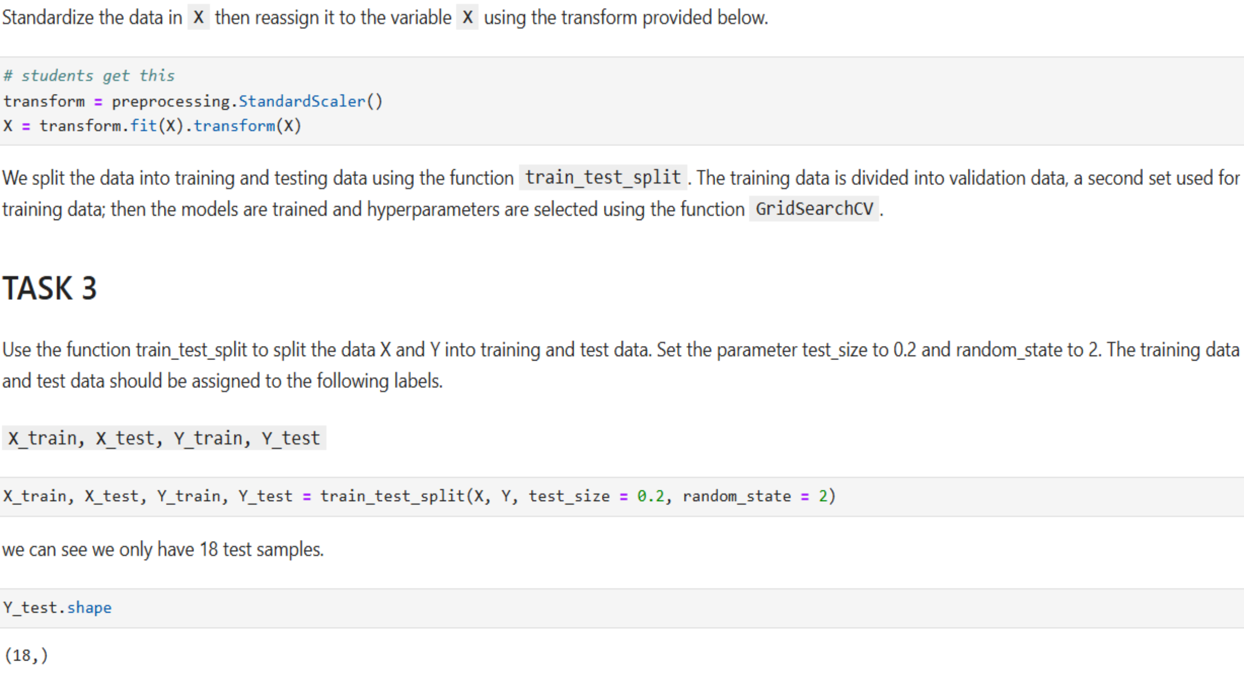
Successful Launches vs Payload Mass (One Site)

If we look at one of the sites with a lower rate of successful launches, we can see that VAFB SLC-4E had only success with a very low and a very high payload, but no successes in the middle of the payload range.



Predictive Analysis – Fitting and Testing Models

We can get our data into the proper format for our machine learning models by standardizing the features and transforming them. Then we can split the data into training and testing sets, with a test set size of 18 samples.



Standardize the data in `X` then reassign it to the variable `X` using the transform provided below.

```
# students get this
transform = preprocessing.StandardScaler()
X = transform.fit(X).transform(X)
```

We split the data into training and testing data using the function `train_test_split`. The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function `GridSearchCV`.

TASK 3

Use the function train_test_split to split the data X and Y into training and test data. Set the parameter test_size to 0.2 and random_state to 2. The training data and test data should be assigned to the following labels.

`X_train, X_test, Y_train, Y_test`

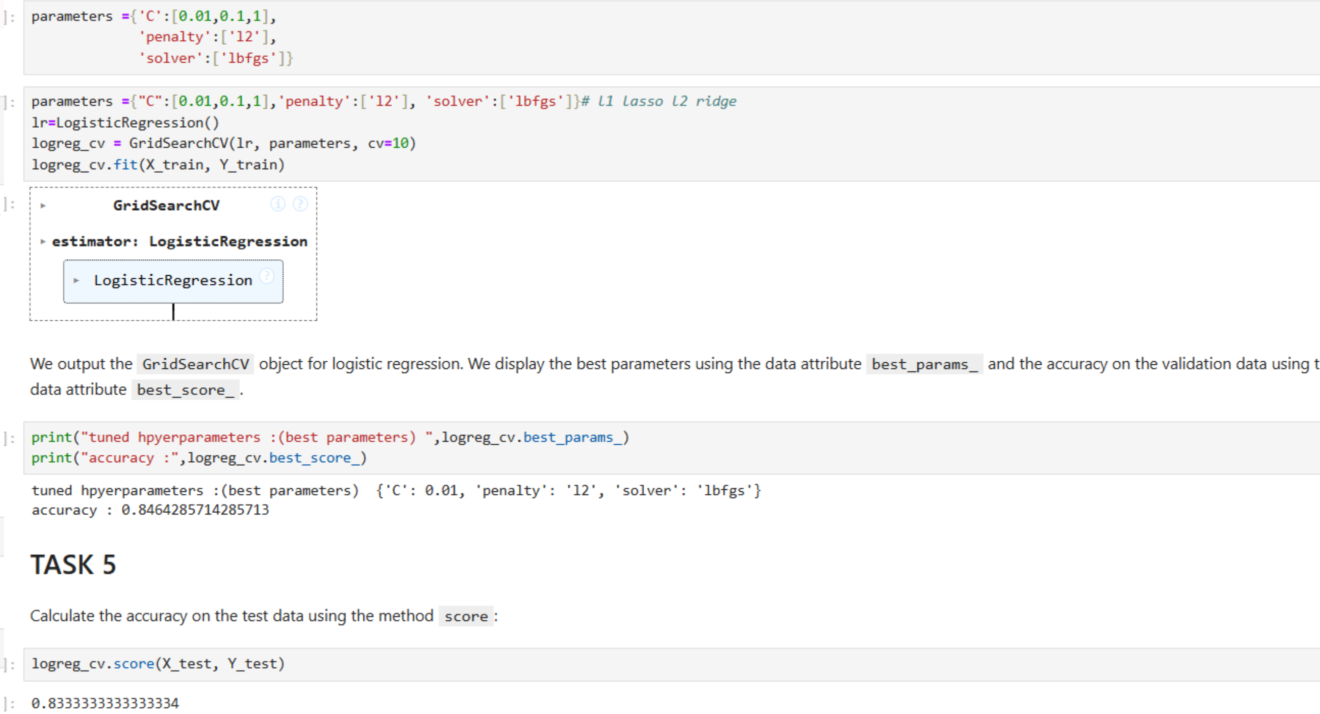
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 2)
```

we can see we only have 18 test samples.

```
Y_test.shape
```

```
(18,)
```

We can then use GridSearchCV to find the optimal parameters for a logistic regression model and fit this model to the training data. Then we can calculate its accuracy in classifying successful launches in the test set. Repeating this process for a SVM, Decision Tree, and K-nearest Neighbors shows that the outcomes of a launch can be predicted very accurately using multiple methods with the other features of the data that we have.



```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr, parameters, cv=10)
logreg_cv.fit(X_train, Y_train)
```

We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params_` and the accuracy on the validation data using t data attribute `best_score_`.

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

TASK 5

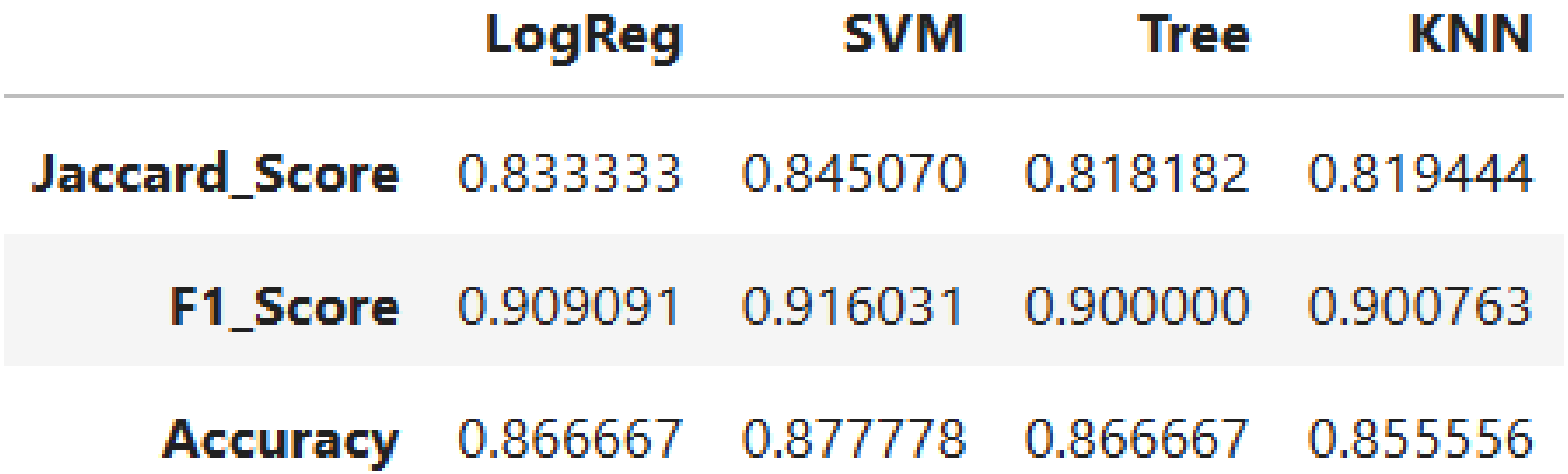
Calculate the accuracy on the test data using the method `score` :

```
logreg_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

Predictive Analysis – Comparing Models

This chart shows the performance of each model on the test data according to different metrics. SVM had the highest Jaccard score, F1 score, and accuracy by a very small margin. However, this does not give us enough confidence to say that SVM is the absolute best model for this classification task, since our test set is relatively small and small variations in models could make a big difference in performance.



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.818182	0.819444
F1_Score	0.909091	0.916031	0.900000	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

SpaceX Launches - FINDINGS & IMPLICATIONS

Findings

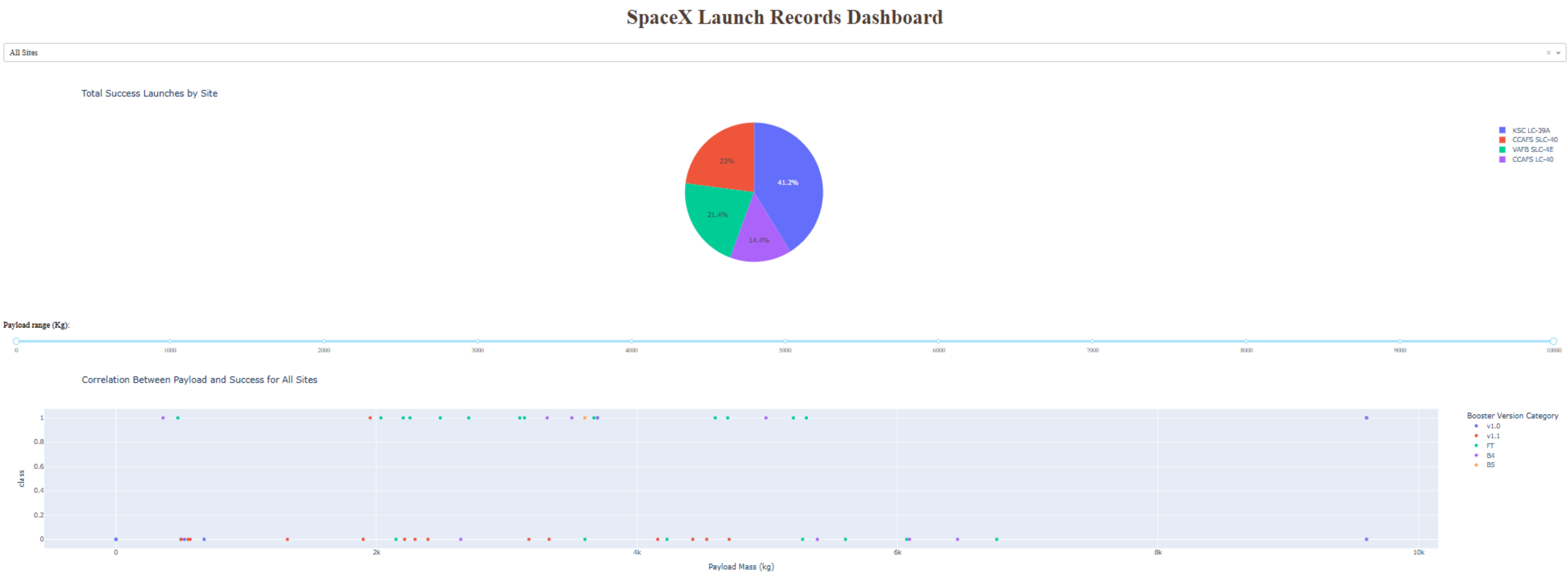
Predicting the success of a launch involves many factors, but some of the most important are the launch site, payload mass, number of flights, and orbit type. From our visualizations, we can see that certain sites such as KSC LC-39A had high overall successes, and was especially successful for lower payload masses, while other sites had more success with heavier payloads. However, overall across the sites higher payloads corresponded with more successful launches. The ES-L1, GEO, HEO, and SSO orbits had the highest success rates. When it comes to number of flights, it was clear that more number of flights led to greater success. Geographical considerations also came into play, as all of the launches were near coastlines and away from cities. Overall, the different models we applied performed very accurately and very similarly on the relatively small test data set using the features mentioned here as well as others.

Implications

In order to be cost efficient with launches, there are many factors we need to consider including launch site, payload mass, number of flights, and orbit type. By feeding all of these factors and more to a well trained machine learning model, we can have a prediction of whether or not a specific launch will be successful, and then consider what changes could make it more likely to be successful. This could be things like considering different sites to launch a rocket from, increasing the payload mass, or comparing with other booster versions.

DASHBOARD

Github Repository: https://github.com/simeonjg/IBM-Applied-Data-Science-Capstone.git



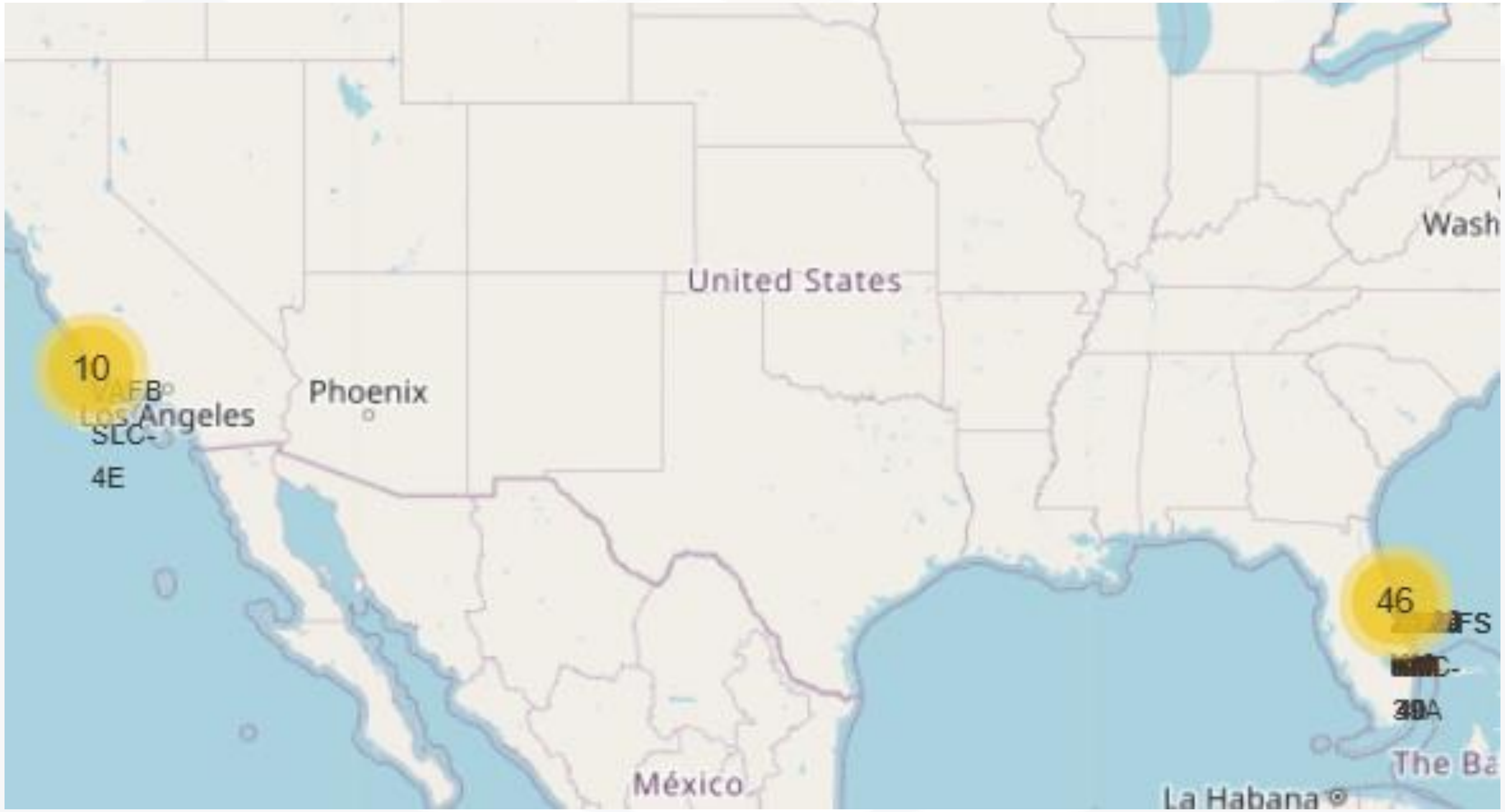
SpaceX Launch Records Dashboard

All Sites

Conclusion

By analyzing the features about a launch, we can train a machine learning model to accurately predict whether a launch will be successful or not, saving money for our company SpaceY. We have come to this conclusion by scraping data, exploring its features, engineering its components for analysis, visualizing important relationships and applying multiple machine learning models to the data.

APPENDIX



Full Folium Map

Sources

- https://spacecenter.org/spacex-dragon-capsule-hits-new-milestone/
- https://spaceflightnow.com/2020/11/29/photos-falcon-9-launches-and-lands-at-vandenberg-air-force-base/
- https://unifysolutions.net/supportedproduct/microsoft-sql-server/
- https://www.nasaspaceflight.com/2023/12/falcon-roundup/
- https://thehill.com/homenews/space/4905273-spacex-mission-boeing-astronauts/
- https://www.flyingmag.com/news/spacex-starship-launch-delay-necessary-faa-administrator-tells-congress/
- https://www.france24.com/en/americas/20240910-spacex-launches-rocket-with-civilian-crew-for-first-ever-private-spacewalk-elon-musk-falcon-9-jared-isaacman
- https://www.businessinsider.com/spacex-falcon-9-launches-updates-schedule