SpaceX Data Analysis

DTSA – 5841 IBM Capstone Project

Arshad Ullah

August 8, 2023

Agenda

01. Executive summary

02. Data Collection

03. Data Wrangling

04. Explorative Data analysis and Visualization (folium)

05. Exploratory Data analysis (SQL)

06. Data Visualization (Interactive dashboard)

07. Data Visualization (using seaborn)

08. Predicting successful stage-1 landings

09. Summary

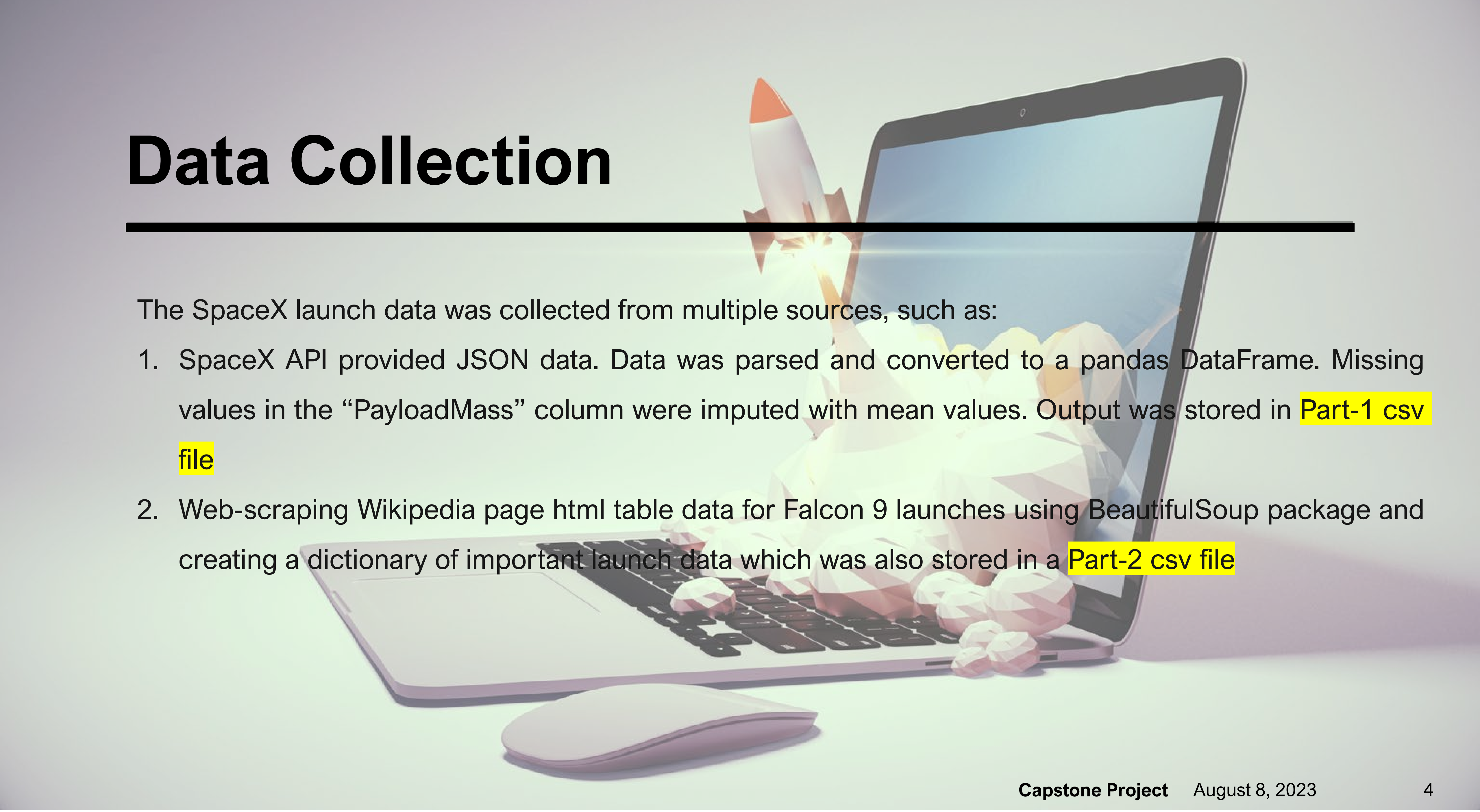
Executive Summary

SpaceX provides a unique capability, which is to reuse the Stage-1 of the launch vehicle, thus allowing it to have significantly less launch costs as opposed to its competitors using traditional launch vehicles.

The goal of this project is to analyze the launch data and predict if Falcon 9 Stage-1 will land successfully. This allows us to calculate the cost of a launch and predict factors that contribute to a successful launch.

We start by collecting the SpaceX launch data from multiple sources, perform exploratory data analysis and use machine learning methods to predict stage-1 landing outcomes.

Data Collection



The SpaceX launch data was collected from multiple sources, such as:

1. SpaceX API provided JSON data. Data was parsed and converted to a pandas DataFrame. Missing values in the “PayloadMass” column were imputed with mean values. Output was stored in Part-1 csv file
2. Web-scraping Wikipedia page html table data for Falcon 9 launches using BeautifulSoup package and creating a dictionary of important launch data which was also stored in a Part-2 csv file

Data Wrangling

Some of the data wrangling steps that were performed:

Part-1 csv file:

1. Check for NULL values in the data (only Landing Pad had 29 NULL values)
2. Number of launches per site and number of Orbits vehicle was launched for were calculated
3. The Landing outcome column was converted into a numeric Class variables with binary values and added as a column to the Part-2 csv file dataset. Percent of successful launches was also calculated (66.67%)

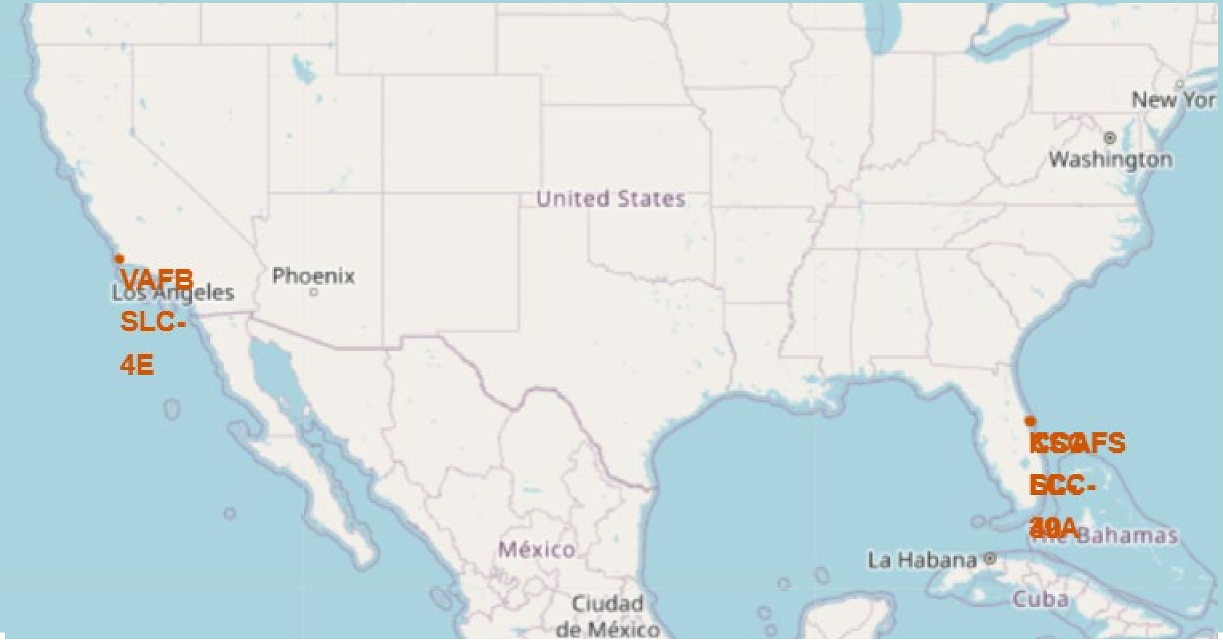
Part-2 csv file (Falcon-9 data):

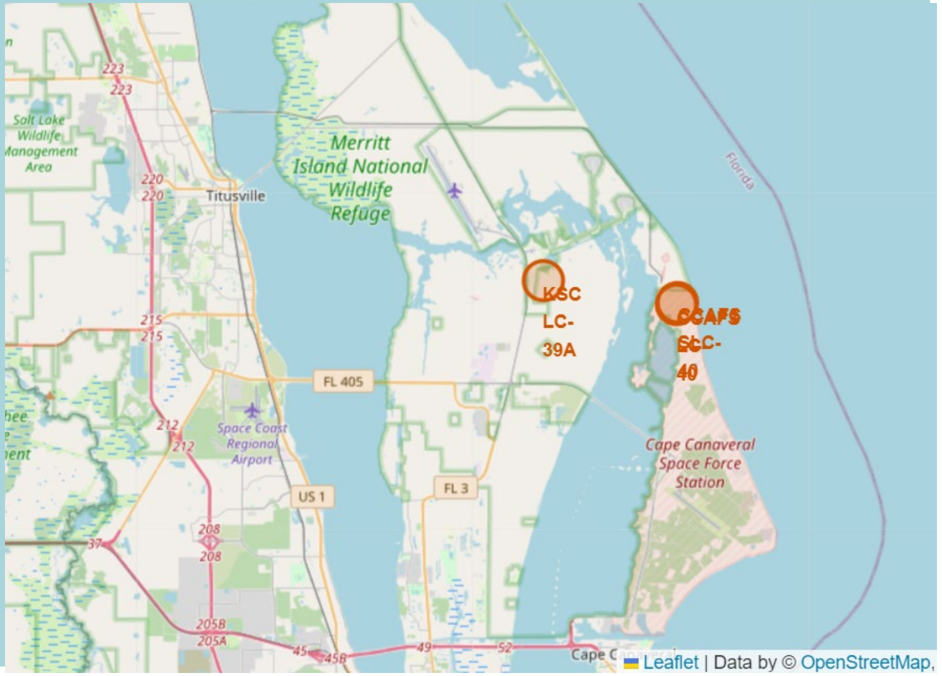
1. Feature engineering was performed by executing one-hot encoding for some categorical data columns such as Orbit, LaunchSite, LandingPad and Serial. Output was stored in a Part-3 csv file

Explorative Data analysis (Folium)

Using an augmented dataset with geospatial data appended for each launch site, each Launch site was mapped:

Note: Cape Canaveral, FL has 2 launch sites as shown on the right:

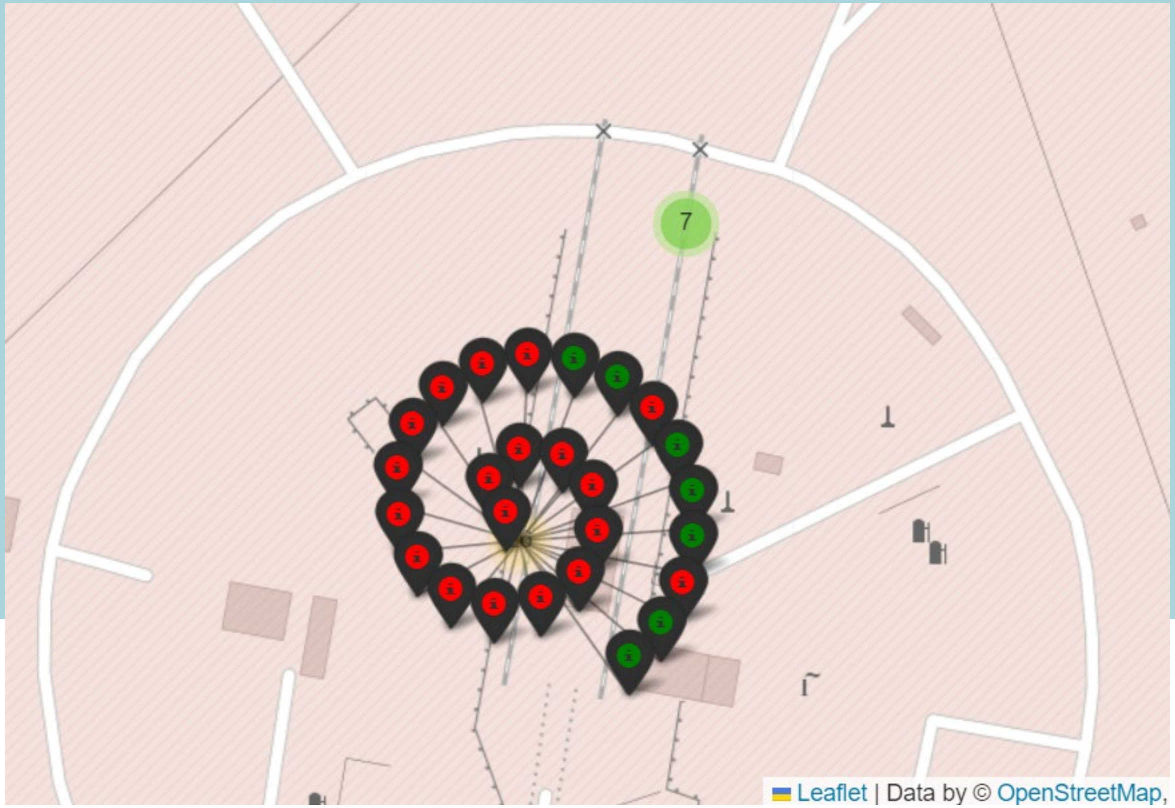


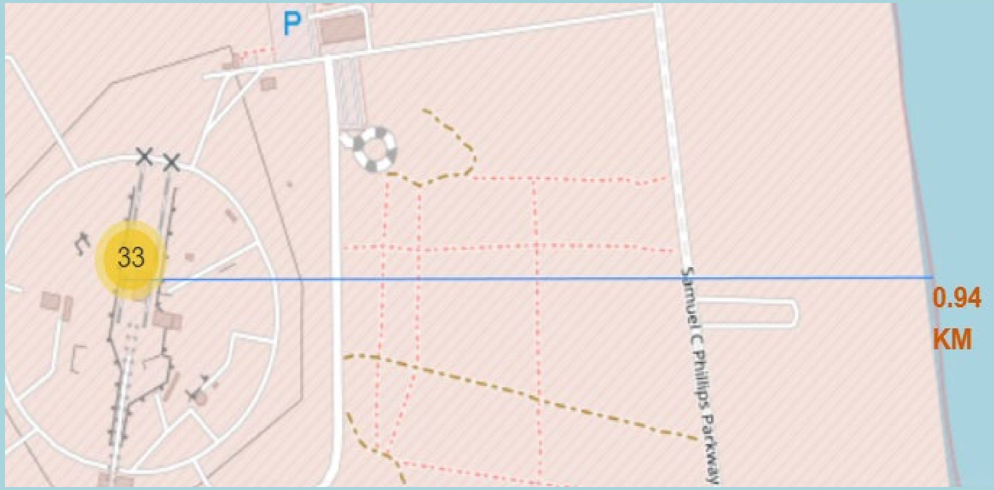


Explorative Data analysis (Folium) contd.

Continuing to work with the launch site success/failure data, markers were added showing these in green and red color.

Shown below are the markers for the launch site Cape Canaveral Launch Complex 40:

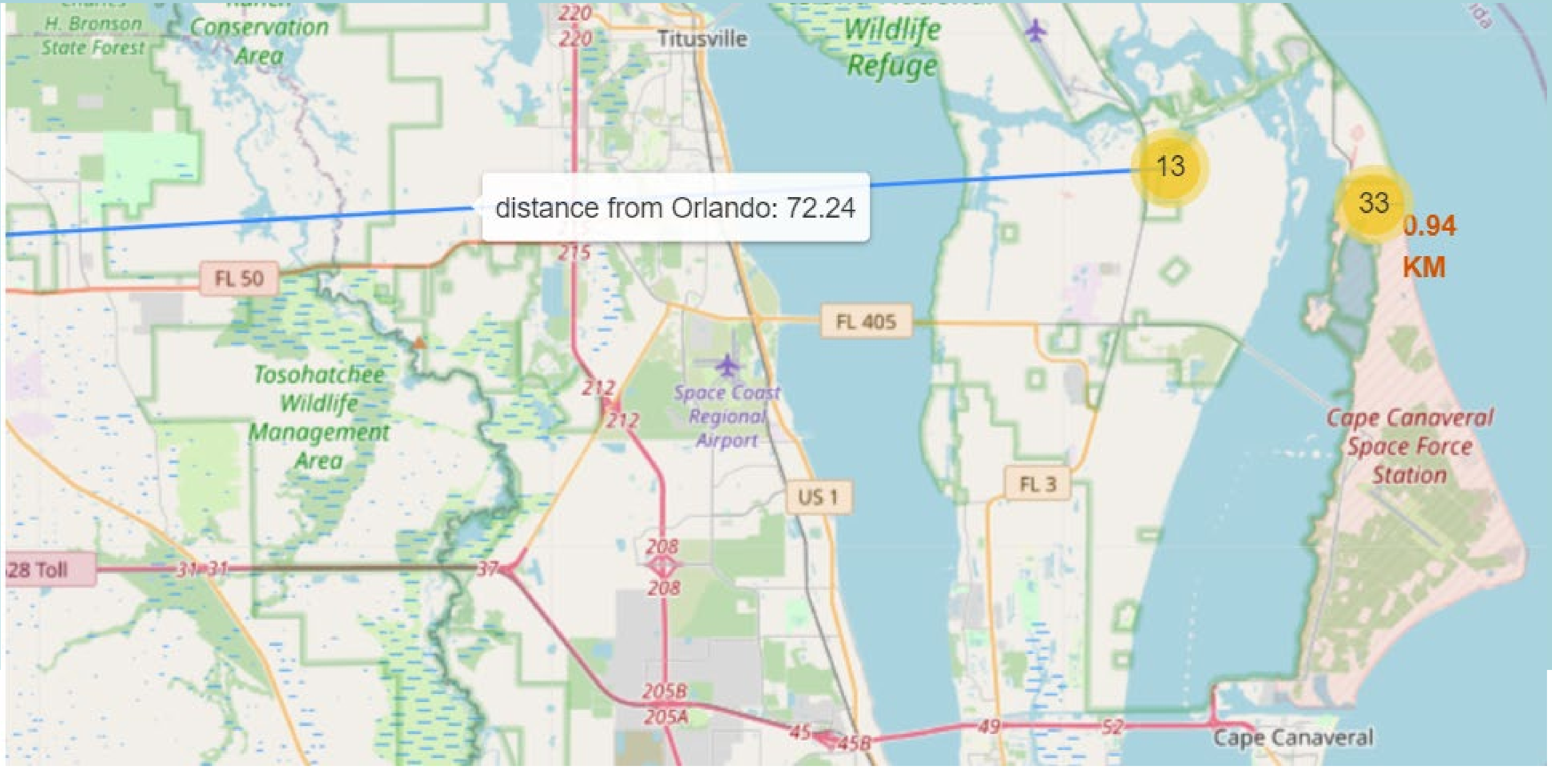




Distance to nearest coastline from Cape Canaveral launch site

Explorative Data analysis (Folium) contd.

Continuing to work with the launch site map data, distance to nearest large city (Orlando) from Kennedy space center is shown below:


distance from Orlando: 72.24
13
33
0.94
KM
Titusville
Wildlife Refuge
H. Bronson State Forest
Conservation Area
FL 50
FL 405
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
US 1
FL 3
Cape Canaveral Space Force Station
528 Toll
Cape Canaveral

Explorative Data analysis (using SQL)

Sqllite python package was used to create an in-memory database. SpaceX data csv file was read into a dataframe and uploaded into this database.

Columns in SPACEXTBL table:

```
Date
Time (UTC)
Booster_Version
Launch_Site
Payload
PAYLOAD_MASS__KG_
Orbit
Customer
Mission_Outcome
Landing_Outcome
```

Task 1: Unique Launch sites:

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

Explorative Data analysis (using SQL) contd.

Task 2: Display 5 rows for launch site starting with “CCA” (Cape Canaveral):

```
('06/04/2010', '18:45:00', 'F9 v1.0  B0003', 'CCAFS LC-40', 'Dragon Spacecraft Qualification Unit', 0.0, 'LEO',
'SpaceX', 'Success', 'Failure (parachute)')
('12/08/2010', '15:43:00', 'F9 v1.0  B0004', 'CCAFS LC-40', 'Dragon demo flight C1, two CubeSats, barrel of Brouere
cheese', 0.0, 'LEO (ISS)', 'NASA (COTS) NRO', 'Success', 'Failure (parachute)')
('22/05/2012', '7:44:00', 'F9 v1.0  B0005', 'CCAFS LC-40', 'Dragon demo flight C2', 525.0, 'LEO (ISS)', 'NASA
(COTS)', 'Success', 'No attempt')
('10/08/2012', '0:35:00', 'F9 v1.0  B0006', 'CCAFS LC-40', 'SpaceX CRS-1', 500.0, 'LEO (ISS)', 'NASA (CRS)',
'Success', 'No attempt')
('03/01/2013', '15:10:00', 'F9 v1.0  B0007', 'CCAFS LC-40', 'SpaceX CRS-2', 677.0, 'LEO (ISS)', 'NASA (CRS)',
'Success', 'No attempt')
['Date', 'Time (UTC)', 'Booster_Version', 'Launch_Site', 'Payload', 'PAYLOAD_MASS__KG_', 'Orbit', 'Customer',
'Mission_Outcome', 'Landing_Outcome’]
```

Task 3: Display total payload mass (in kg) launched by NASA (CRS):

```
      45596.0
```

Task 4: Display average payload mass carried by booster version F9 v1.1:

```
      2534.67
```

Task 5: List the date when the first successful landing outcome in ground pad was achieved:

```
      2015-12-22
```

Explorative Data analysis (using SQL) contd.

Task 6: names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg:

```
['Booster_Version']
('F9 FT B1022',)
('F9 FT B1026',)
('F9 FT  B1021.2',)
('F9 FT  B1031.2’,)
```

Task 7: List the total number of successful and failure mission outcomes

```
['outcome', 'count(*)']
(None, 898)
('Failure', 1)
('Success', 100)
```

Task 8: List the names of the booster versions which have carried the maximum payload mass

```
['Booster_Version']
('F9 B5 B1048.4',)
('F9 B5 B1048.5',)
('F9 B5 B1049.4',)
('F9 B5 B1049.5',)
('F9 B5 B1049.7 ',)
('F9 B5 B1051.3',)
('F9 B5 B1051.4',)
('F9 B5 B1051.6',)
('F9 B5 B1056.4',)
('F9 B5 B1058.3 ',)
('F9 B5 B1060.2 ',)
('F9 B5 B1060.3',)
```

Explorative Data analysis (using SQL) contd.

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

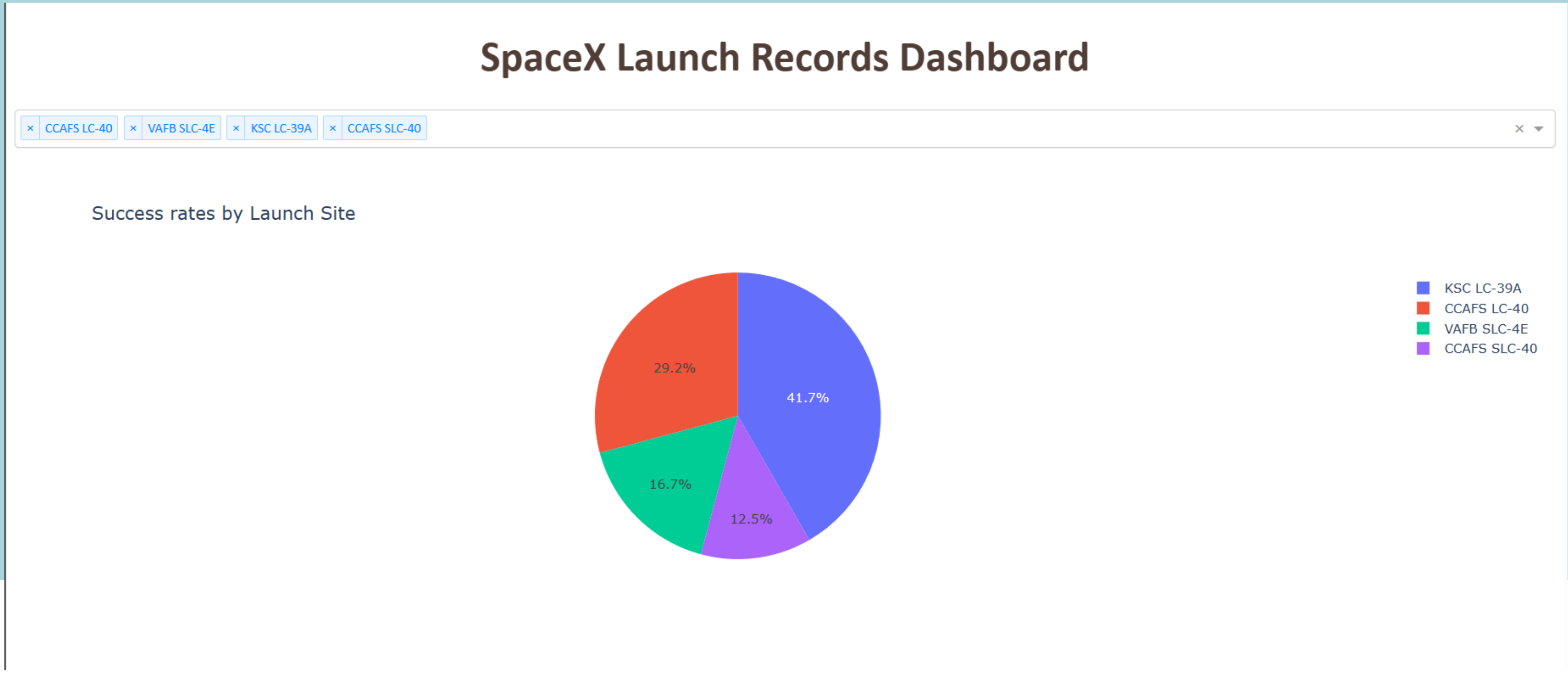
```
['month', 'Landing_Outcome', 'Booster_Version']
('04', 'Failure (drone ship)', 'F9 v1.1 B1015')
('10', 'Failure (drone ship)', 'F9 v1.1 B1012’)
```

Task 10: 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

```
['Landing_Outcome', 'cnt']
('Success', 38)
('No attempt', 21)
('Success (drone ship)', 14)
('Success (ground pad)', 9)
('Failure (drone ship)', 5)
('Controlled (ocean)', 5)
('Failure', 3)
('Uncontrolled (ocean)', 2)
('Failure (parachute)', 2)
('Precluded (drone ship)', 1)
('No attempt ', 1)
```

Data Visualization (Interactive dashboard)

An two part interactive dashboard was built using the **dash and plotly** packages. Part-1 was a pie chart of success rates by launch sites (selected by the user)

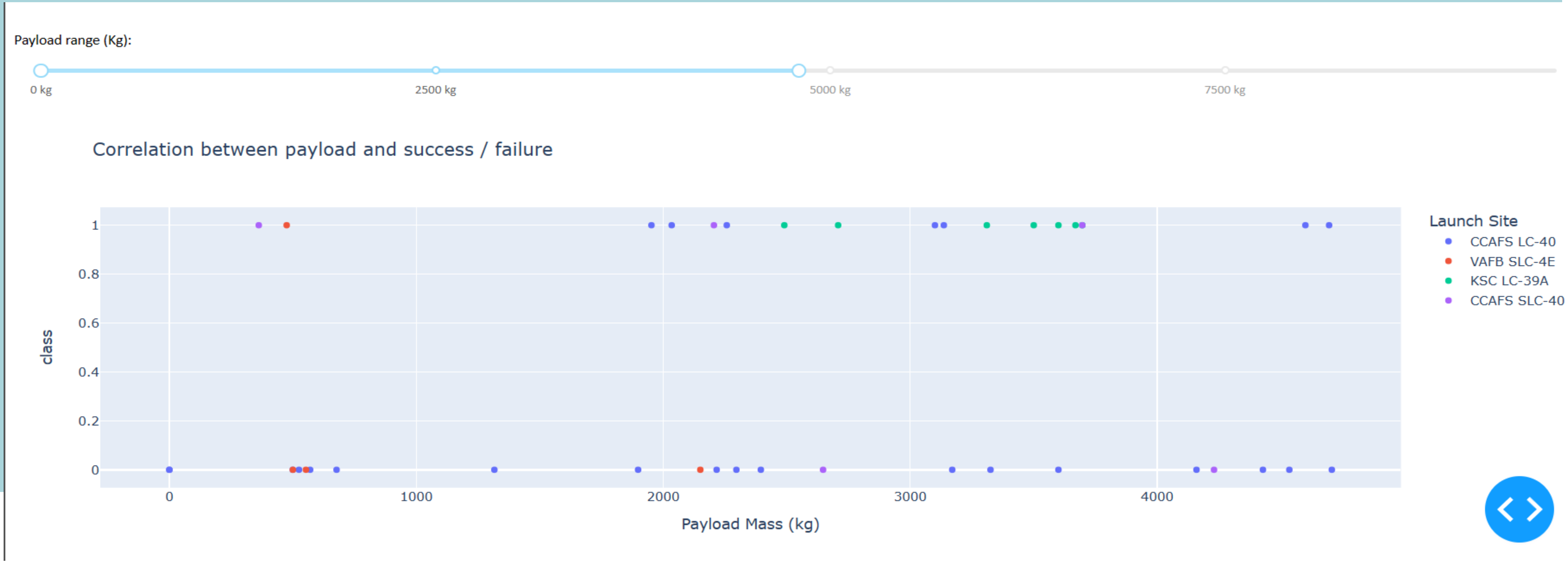


SpaceX Launch Records Dashboard

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

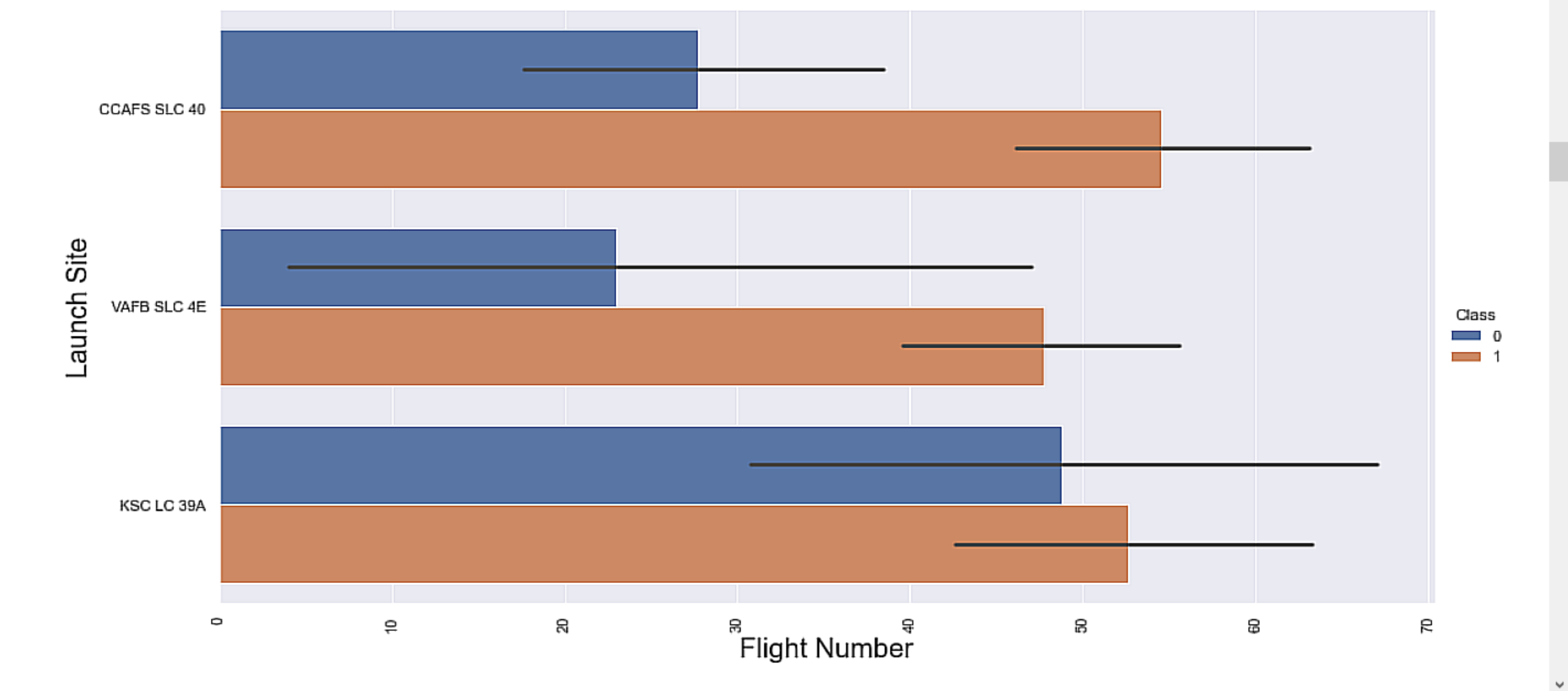
Data Visualization (Interactive dashboard) contd.

Part-2 showed success/failure for each site based on payload range that can be selected using a slide ruler. The number of failed launches increased for heavier payloads (above 5000 kg)



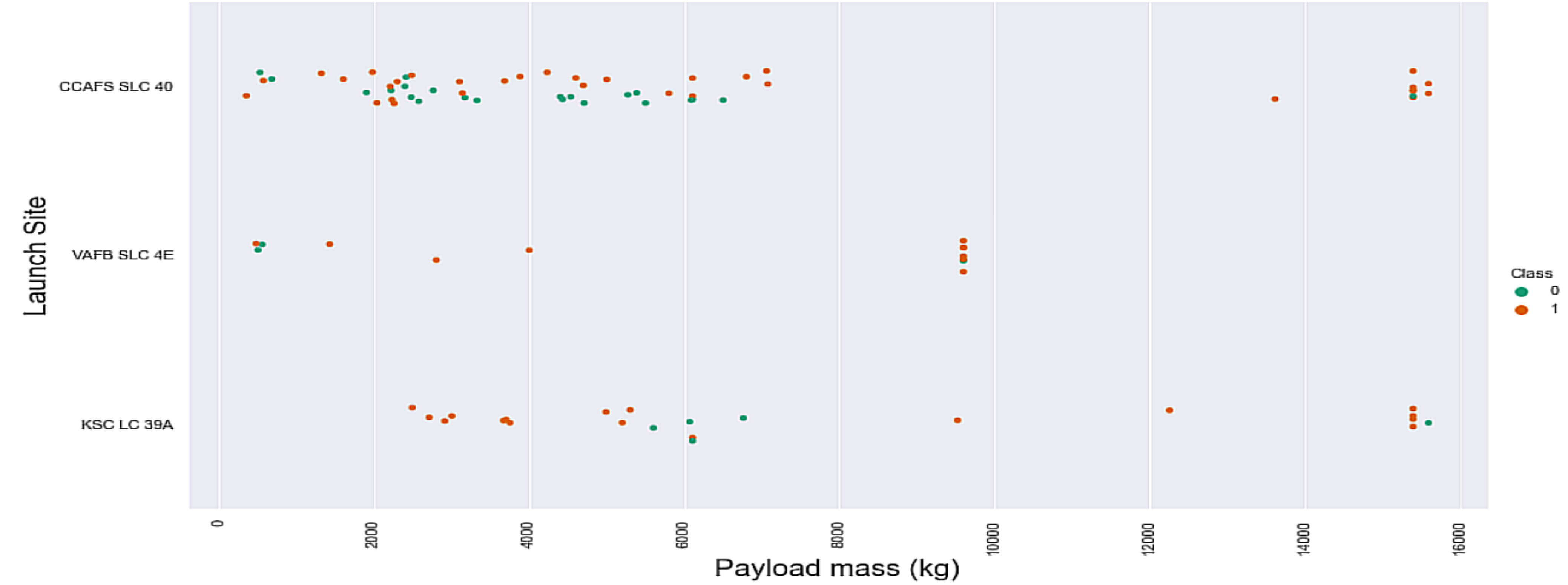
Data Visualization (using seaborn)

The Part-2 dataset with only Falcon-9 launch data, was used to do further data analysis and visualization. CCAFS SLC-40 launch site showed the best success to failure ratio.



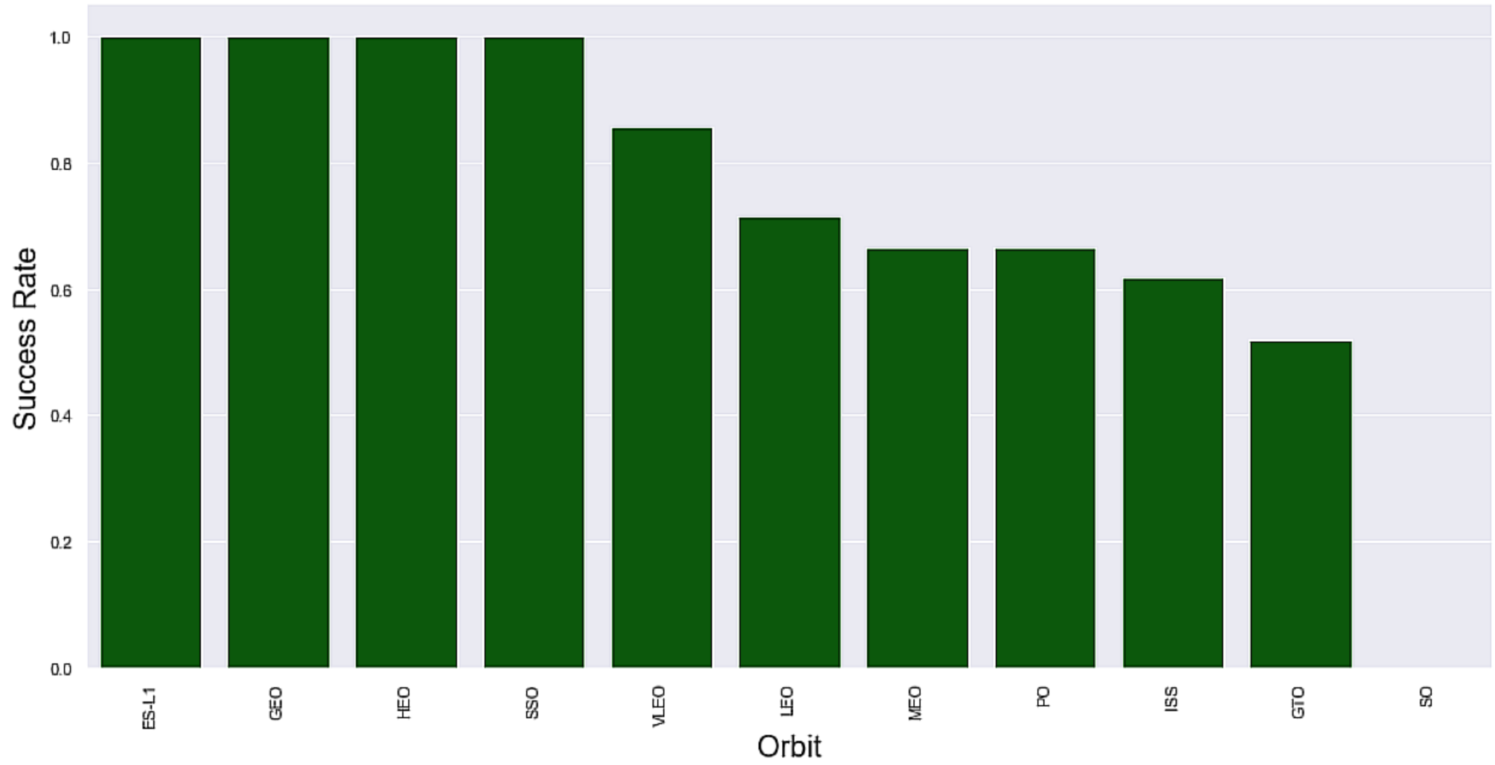
Data Visualization (using seaborn) contd.

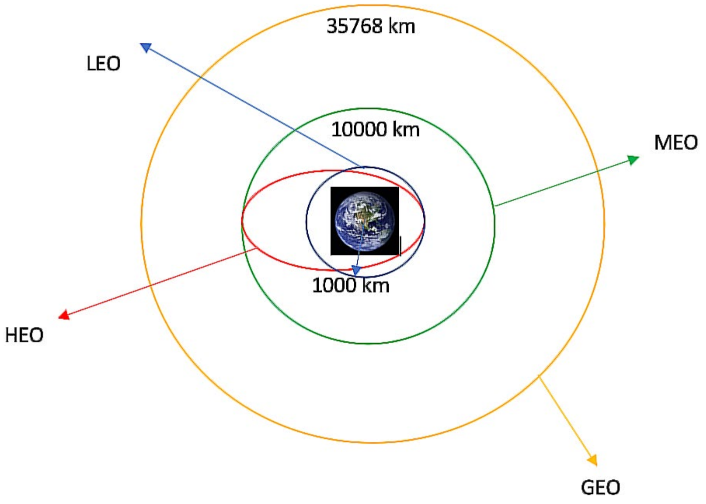
Continuing with the Falcon-9 launch data, failure and success launches by launch site was plotted. VAFB SLC-4E had no heavy payload launches (above 10,000 kg)



Data Visualization (using seaborn) contd.

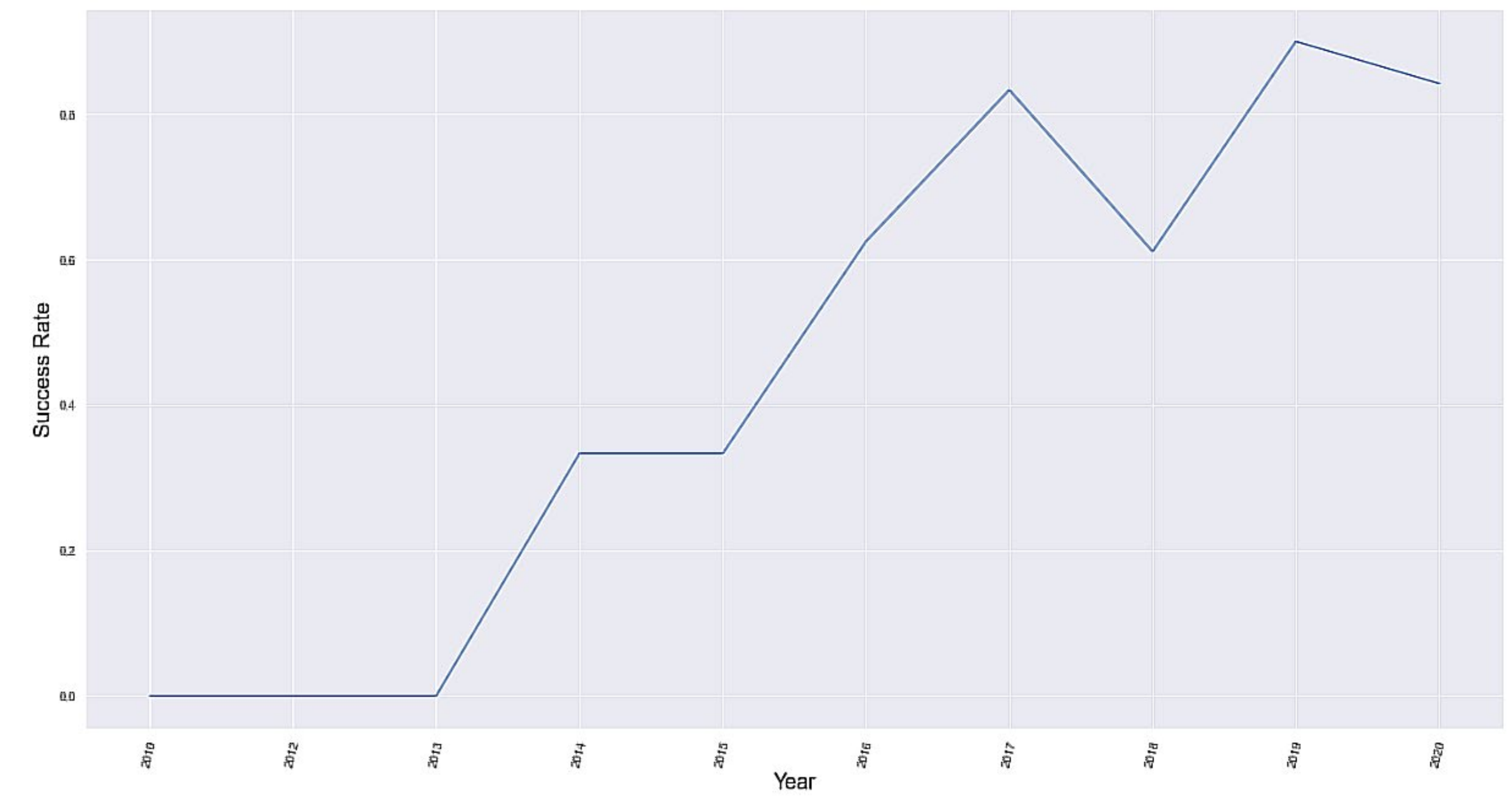
Continuing with the Falcon-9 launch data, failure and success launches by launch site was plotted. VAFB SLC-4E had no heavy payload launches (above 10,000 kg)





Data Visualization (using seaborn) contd.

Continuing with the Falcon-9 launch data, failure and success launches by Year was plotted.



Since 2013 success rates of stage-1 landing increased steadily with a slight dip in 2018 (maybe due to design changes)

Predicting successful stage-1 landings

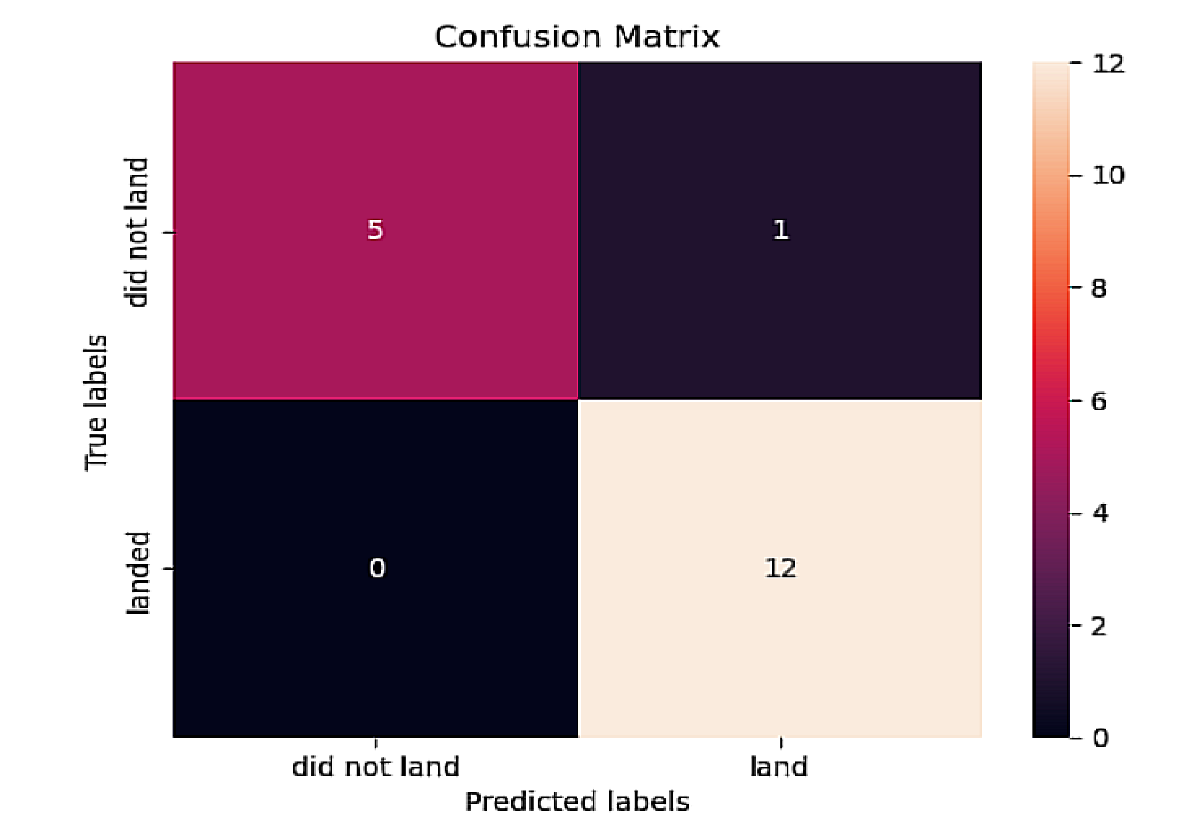
- We start with pulling the Class column from **Part-2 csv file** (that we transformed to binary values in the data-wrangling steps). This will be our Y variable for the machine learning methods for predicting stage-1 landing outcomes
- Next we take the entire **Part-3 csv file** (with the one-hot encoded data) and treat it as our **X** variable
- We next standardize the data in X using **StandardScaler** function from scikit-learn
- We next split the X and Y data into train and test datasets (80-20 split)
- We next apply multiple supervised and un-supervised machine learning methods listed below on the data and use **GridSearchCV** to find the best parameters for each method:
    1. Logistic Regression
    2. Support vector machines (SVM)
    3. Decision trees
    4. K-nearest neighbors

Predicting successful stage-1 landings – results

1. **Logistic Regression:**
   - best parameters: 'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs’
   - Train accuracy : 0.8464285714285713
   - Test accuracy : 0.8333333333333334
2. **Support vector machines (SVM):**
   - best parameters: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid’}
   - Train accuracy : 0.8482142857142856
   - Test accuracy : 0.8333333333333334
3. **Decision tree:**
   - best parameters: {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random’}
   - Train accuracy : 0.8875
   - Test Accuracy: 0.9444444444444444
4. **K-nearest neighbors:**
   - best parameters: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
   - Train accuracy : 0.8482142857142858
   - Test accuracy : 0.8333333333333334

Predicting successful stage-1 landings – Best results

Decision tree showed the best test accuracy with a score of 0.94
Confusion matrix:

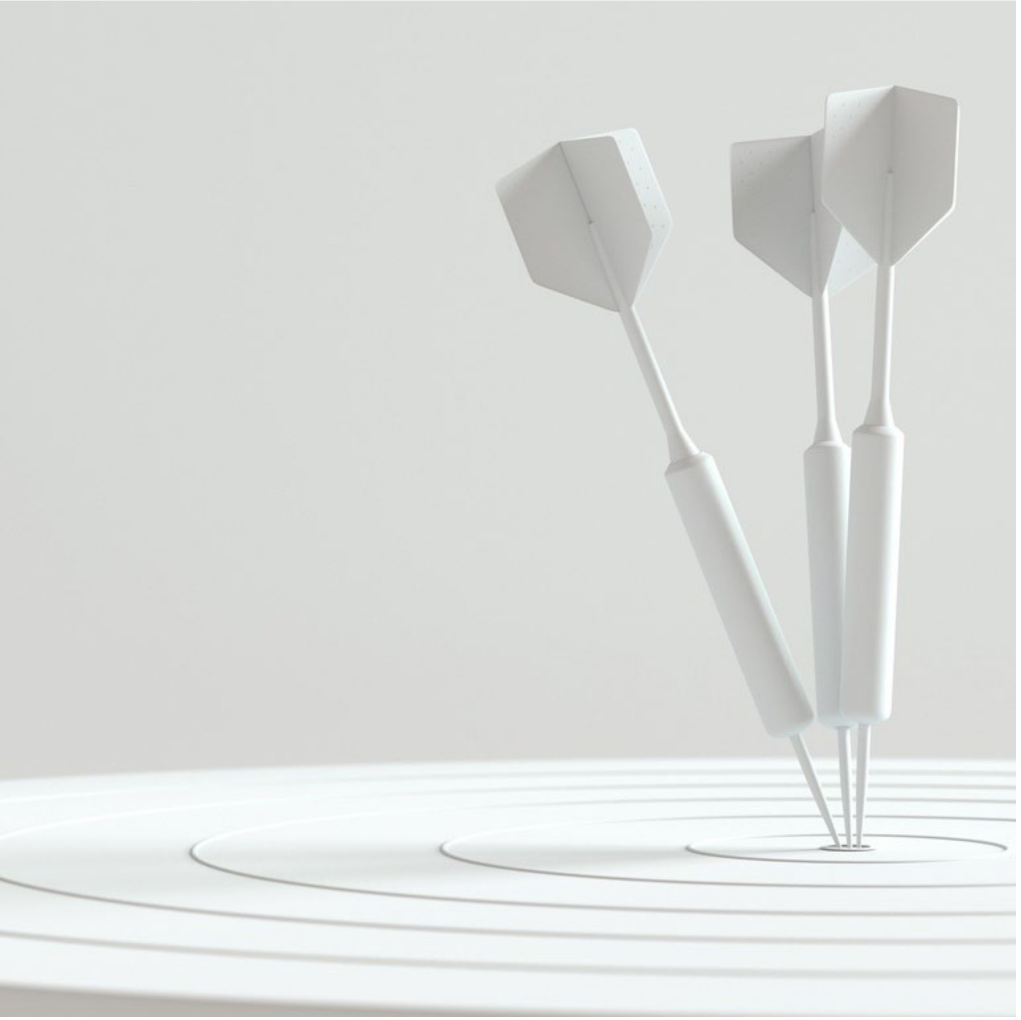


As is evident from the confusion matrix Decision tree shows the least number of false positives in the test dataset.

Note: *Decision trees should be used with care since they tend to overfit the train data. And also the test data was not out of sample and belonged to the same source dataset. More data and different sources should be used to achieve better models that would generalize with unseen data.*

Summary

- The chances of having a successful Stage-1 landing for the Falcon-9 is high (94%) and only 1 false-positive showing in the test results
- SpaceX has shown better success rates over the years as the technology and launch vehicle designs have improved
- Kennedy Space Center Launch Complex 39A in Florida has the highest success rate for stage-1 landings (42%)
- Launch sites are located near the coastline in the U.S. and away from large population areas



Thank you

Thank you to IBM Applied Data Science Capstone and Skills Network teams for providing the guidance, data and starter code to assist in accomplishing this project.

All project related code can be found here:

https://github.com/azullah/IBM-Capstone-Project-AU

Arshad Ullah

MS Data Science

University of Colorado, Boulder

arul6419@colorado.edu