Applied Data Science Capstone

Assignment Report

Kunihiro Takagi

July 26, 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Summary of Methodology

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data by filtering the data, handling missing values to prepare for analysis and modeling
- **Explore** data with SQL and data visualization techniques
- **Visualize** data using Folium and Ploty Dash
- **Build Models** to predict landing outcomes using classification models

Results

- **Exploratory Data Analysis:** Launch success rate has improved since 2013.
- **Visual Analysis:** KSC LC-39A has the most success outcomes among the launch sites.
- **Predictive Analysis:** Decision Tree Model slightly outperformed the other models.

INTRODUCTION

- In this capstone, I made a survey to determine if the Falcon 9 first stage would land successfully using data science techniques, because it would enable to estimate the launch cost well.
- 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.

METHODOLOGY

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data by transforming it into a desired format and handling missing values to prepare for further analysis and modeling
- **Explore** data via EDA with SQL and data visualization techniques
- **Visualize** the data using Folium and Plotly Dash
- **Build Models** to predict launch outcomes using classification models, and then evaluate models to find the best model and parameters

Data Collection

- **Data collection** is the process of gathering information from which you can create data sets to bring you a helpful insight for your business. In this capstone, the dataset was collected by **REST API** and **Web Scrapping**.

- For **REST API**, I used the get request to collect the data and then turned it into a pandas dataframe. The dataframe was cleaned and checked for missing values to be replaced with the mean values.
- For **Web Scrapping**, I used the Python BeautifulSoup library to extract the launch records as HTML table, parse and convert the table into a pandas dataframe for further process like exploratory data analysis.

Data Wrangling

- **Data Wrangling** is the process of transforming the data into a desired format, making it more useful for further analysis.
- I created a landing outcome label from the outcome column in order to make it easier for further analysis.
- I calculated the number of launches on each site, and then calculated the number and occurrence of mission outcome for each orbit type.

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Exploratory Data Analysis (EDA)

- **Exploratory Data Analysis (EDA)** is the process of analyzing data to summarize their main characteristics, creating a database to use **SQL** and **Data Visualization** techniques.
- I loaded the dataset into the corresponding table in a Db2 database and executed **SQL queries** for exploratory analysis.

```
[13]: sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[13]:

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

EDA with SQL

I performed queries to analyze the data as listed below:

- **Displaying the names of the launch sites.**
- **Displaying 5 records where launch sites begin with the string 'CCA'.**
- **Displaying the total payload mass carried by booster launched by NASA (CRS).**
- **Displaying the average payload mass carried by booster version F9 v1.1.**
- **Listing the date when the first successful landing outcome in ground pad was achieved.**
- **Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.**
- **Listing the total number of successful and failure mission outcomes.**
- **Listing the names of the booster_versions which have carried the maximum payload mass.**
- **Listing the failed landing_outcomes in drone ship, their booster versions, and launch sites names for in year 2015.**
- **Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.**

Examples of SQL Query

Displaying the names of the launch sites

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

```
sql SELECT LANDING_OUTCOME, COUNT(*) AS Total FROM SPACEXTBL
    WHERE DATE BETWEEN '04/06/2010' AND '20/03/2017' GROUP BY LANDING_OUTCOME ORDER BY Total DESC;
```

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

Landing_Outcome	Total
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

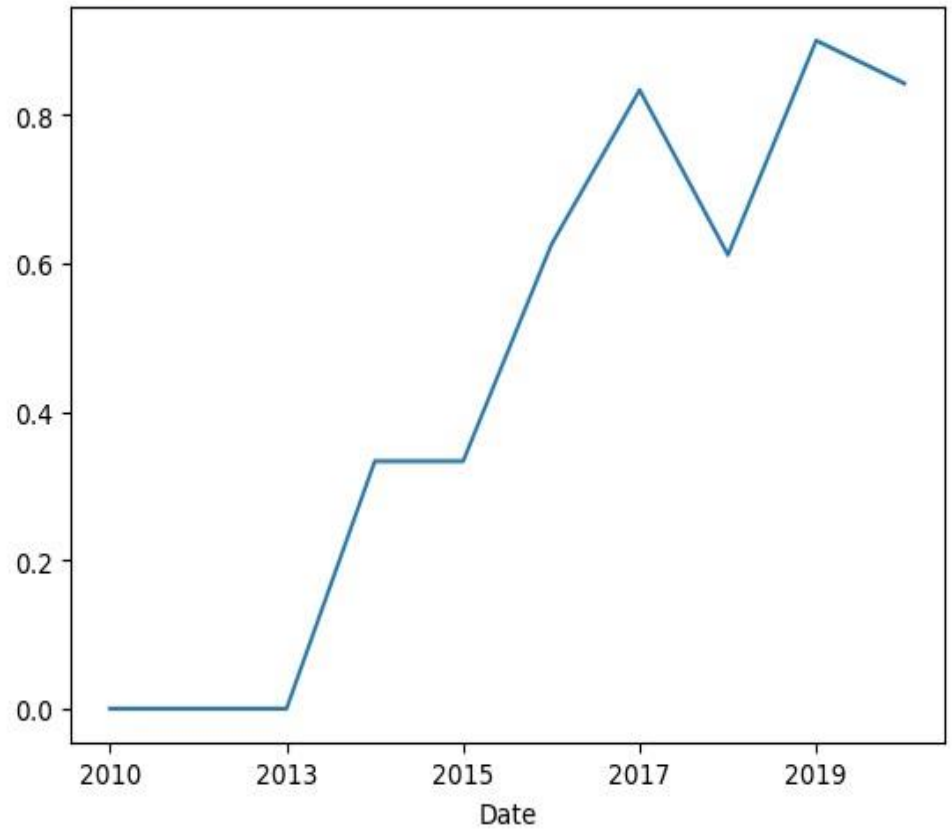
EDA with Data Visualization

- Using Data Visualization technique, I illustrated the characteristics of the data set and analyzed it.

- As shown in the right line graph, the landing success rate has improved since 2013.

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df.groupby('Date')['Class'].mean().plot.line()
```

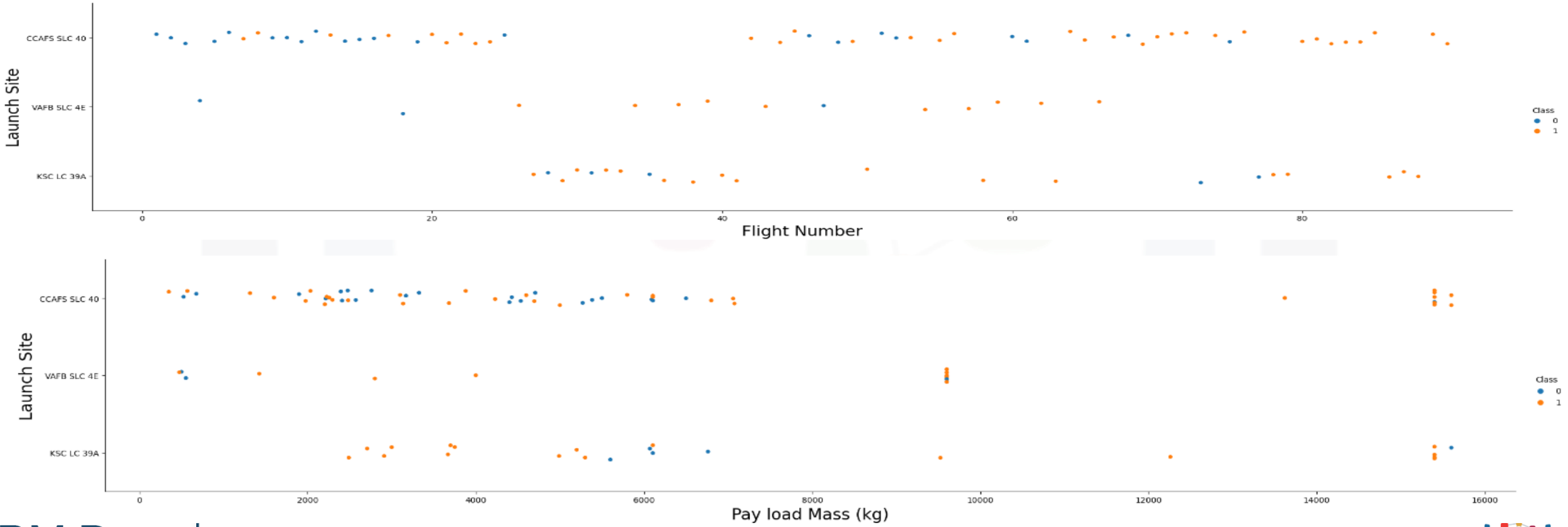
```
<AxesSubplot:xlabel='Date'>
```



EDA with Data Visualization

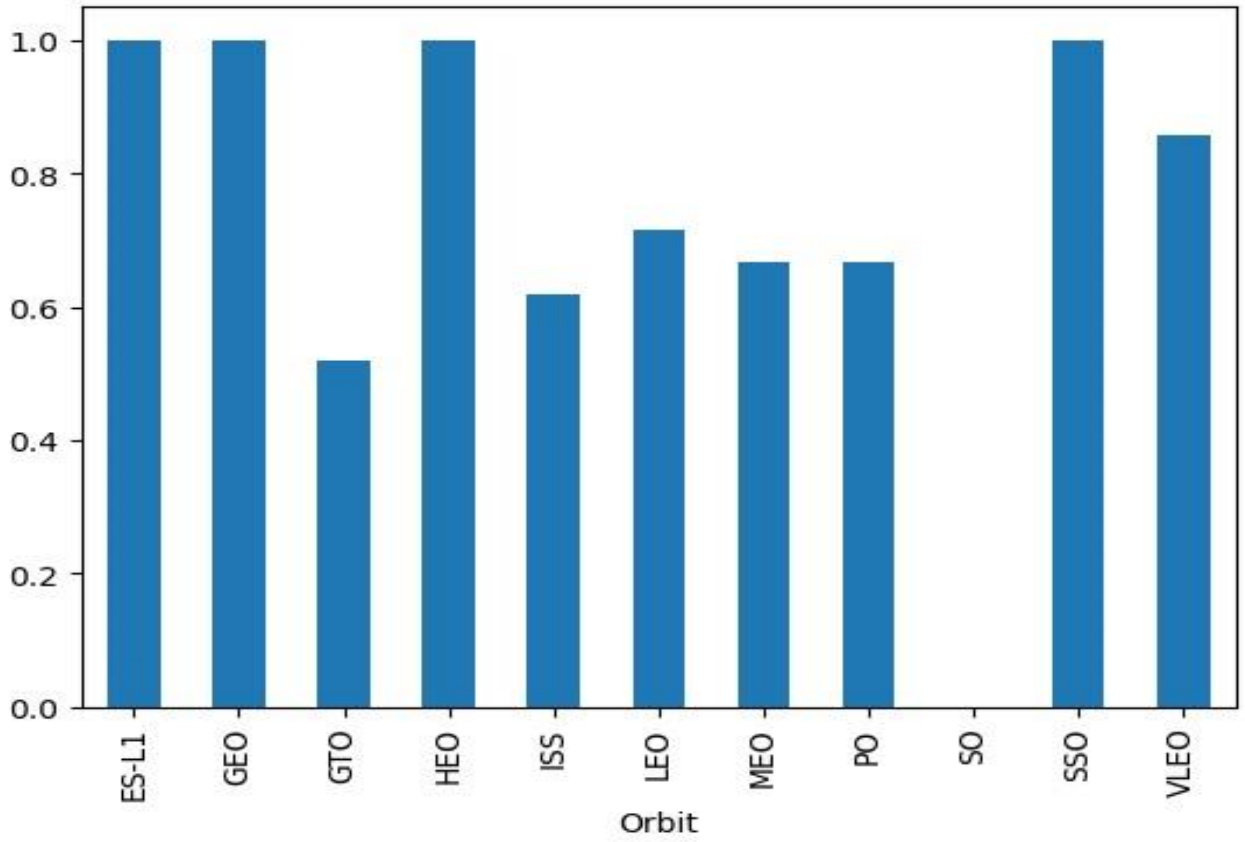
- I also visualized and observed the relationship in landing success among the features like Launch site, Flight number(Time) and Payload mass as shown below.

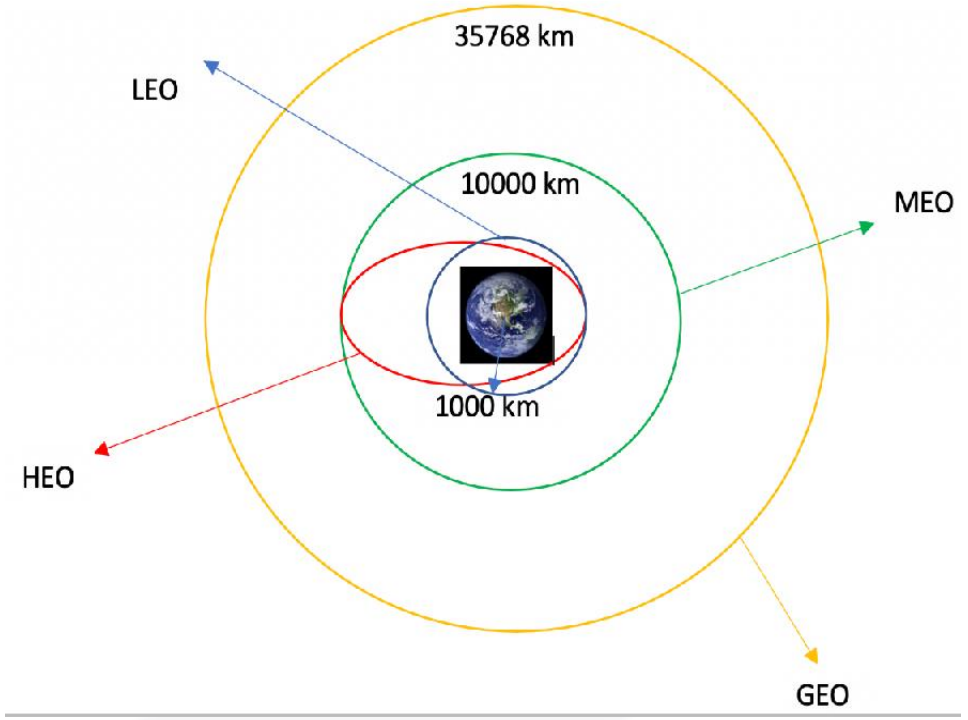
Blue markers(0): Unsuccessful launches, **Orange markers(1): Successful Launches**



EDA with Data Visualization

- I calculated and visualized the success rate of each orbit as shown below.
- ES-L1, GEO, HEO and SSO has 100% success rate.

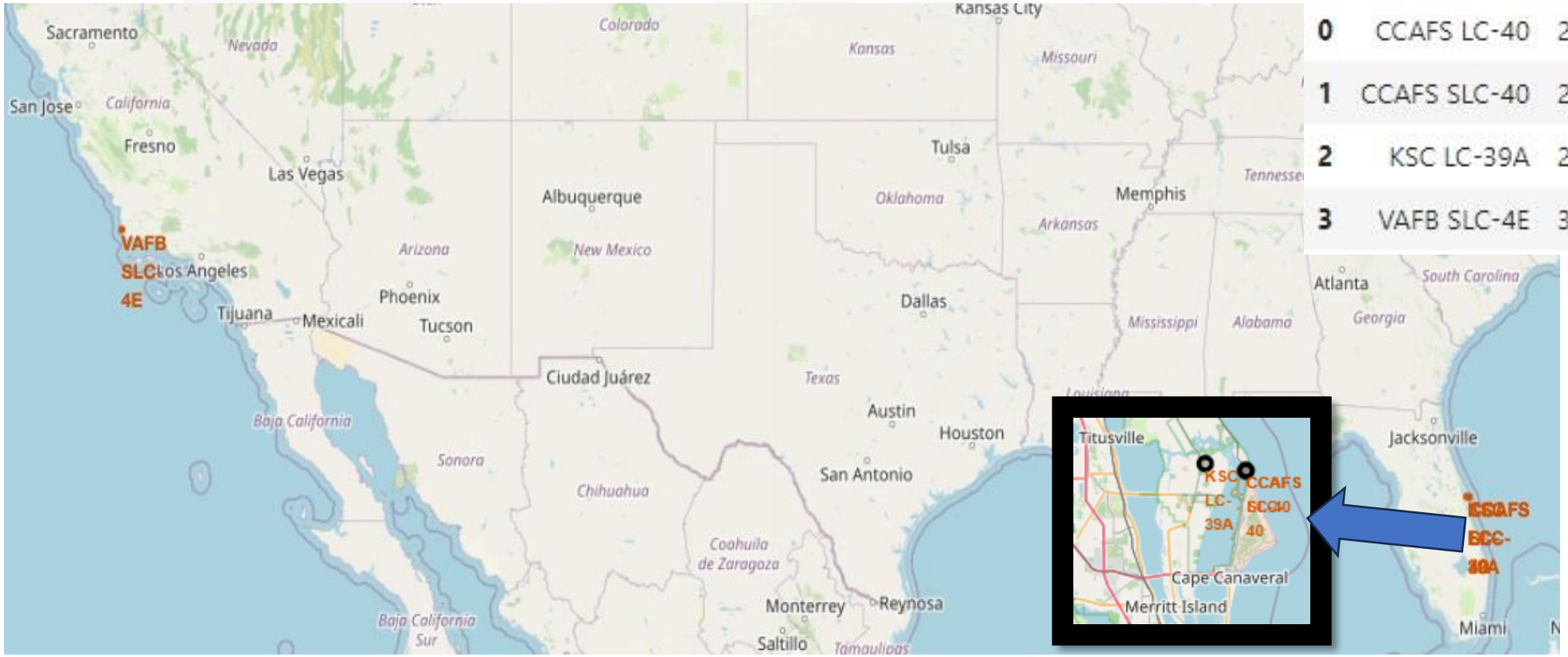




Interactive Visual Analysis(IVA) - Folium

- After collecting the coordinates for each launch site, I made a visual analysis on launch outcomes at each site using Folium.

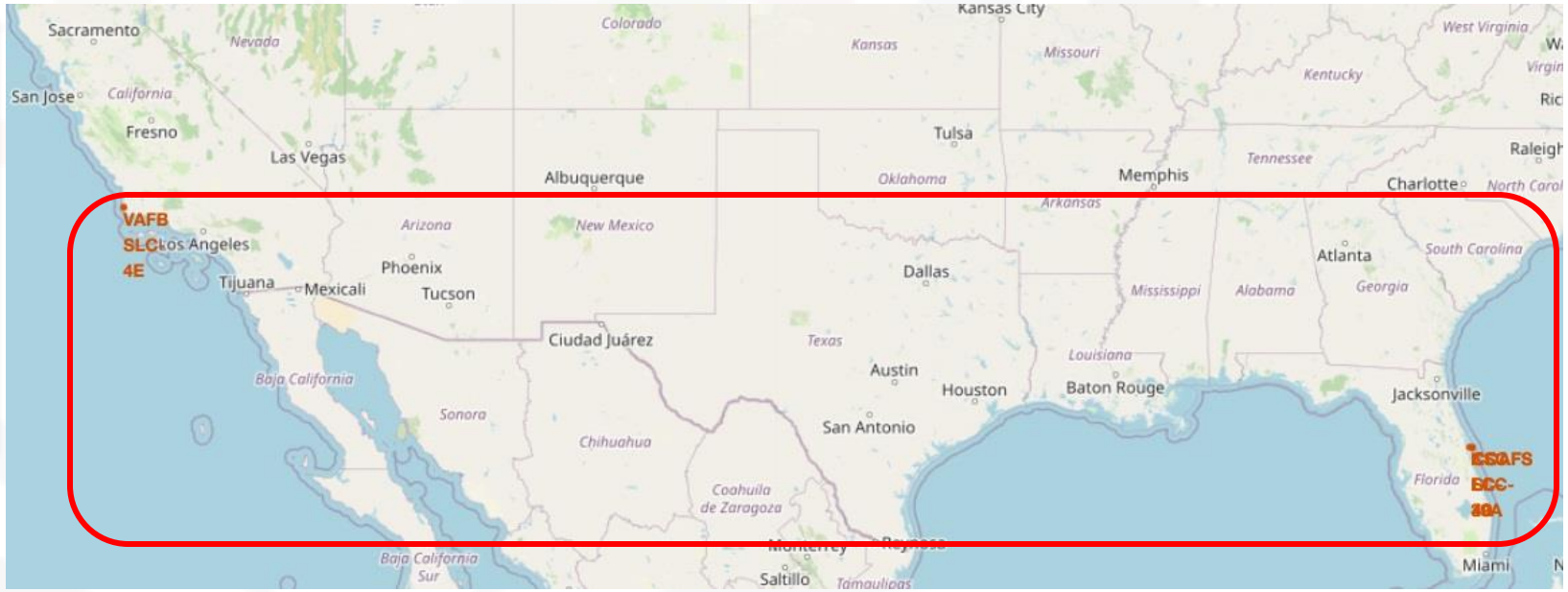
	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745



IVA-Folium Location of Launch sites

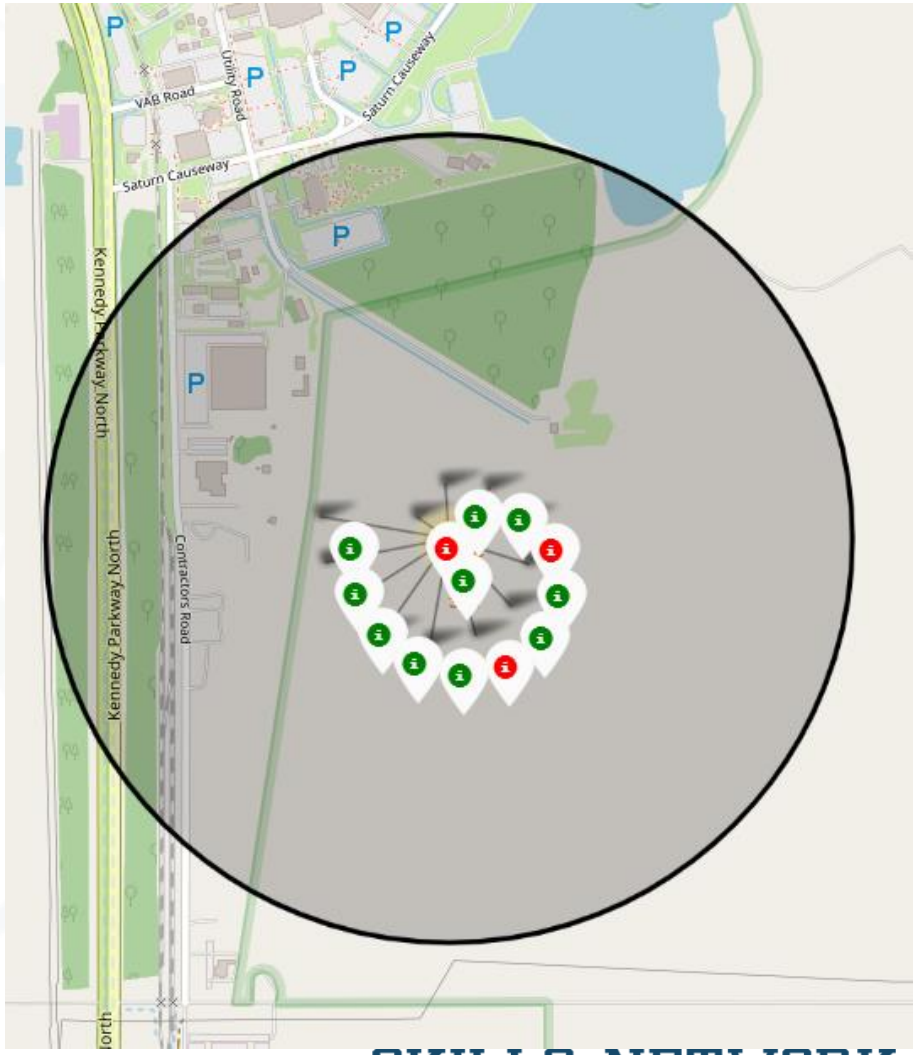
- Identifying the coordinates of the launch sites, I found they were **located nearby the equator**. It is supposed to make it easier for rockets to launch to equatorial orbit, and **save the cost of fuel and boosters**.

	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



IVA-Folium: Launch outcomes

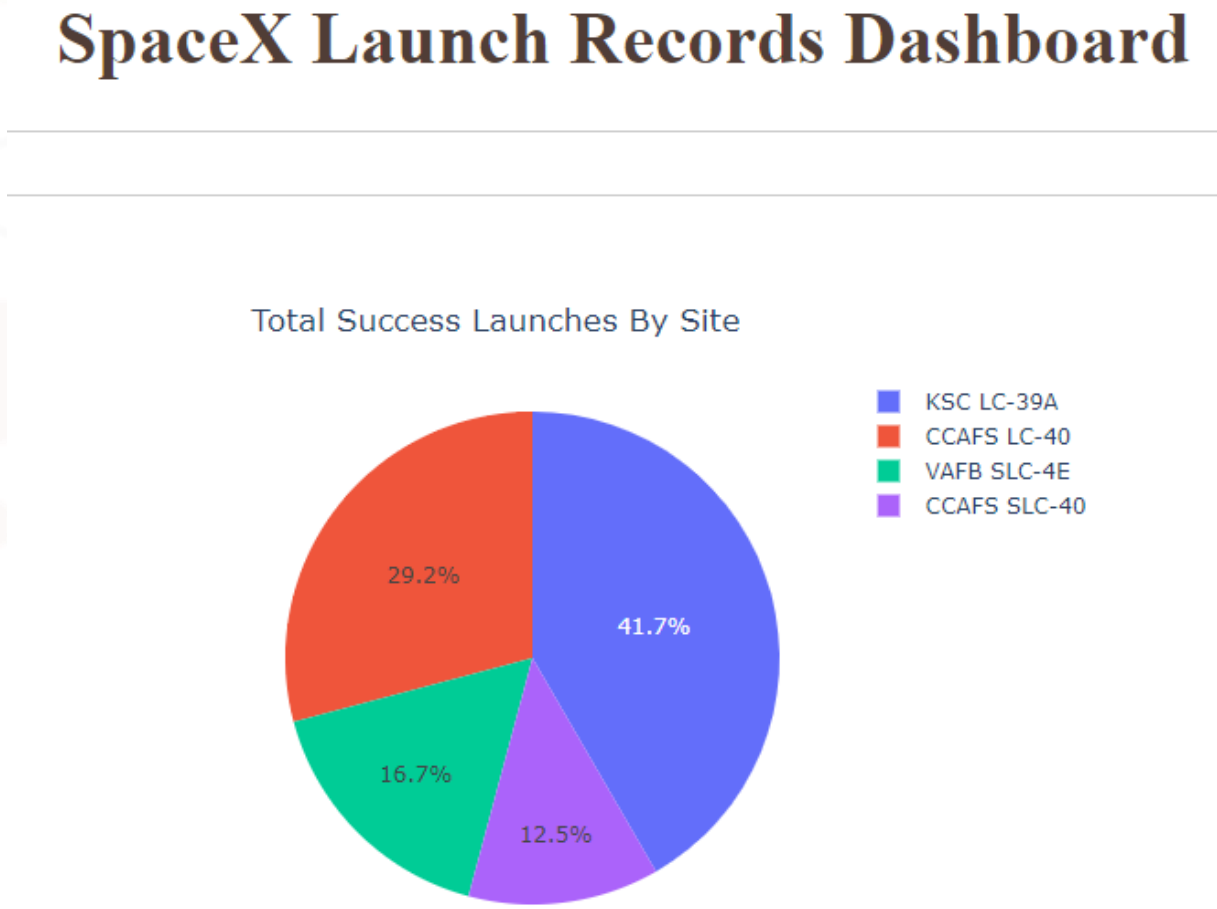
- At each launch site, I marked launch outcomes.
- Green markers illustrate **successful** outcomes.
- Red markers illustrate **unsuccessful** outcomes.
- Launch site, **KSC LC-39A** has a **10/13(76.9%) success rate**, as shown on the right map, and it is the **highest success rate** among those of all launch sites.



IVA-Plotly Dash

- Using Ploty Dash, I created an interactive dashboard for visual analysis.

- As shown in the right pie chart, **KSC LC-39A has the most successful launches** among all the launch sites.



Build Models

- I built models to predict launch outcomes using machine learning algorithms, including SVC, KNN, Logistic Regression and Decision Tree and evaluated them.
- The below evaluation shows the best algorithm is Decision Tree.

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'SVM':svm_cv.best_score_,'LogisticRegression':logr
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Parameter is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Parameter is :',knn_cv.best_params_)
if bestalgorithm == 'SVM':
    print('Best Parameter is :',svm_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Parameter is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.8625
Best Parameter is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_sampl
es_split': 2, 'splitter': 'random'}
```

Conclusion

Exploratory Data Analysis

- Launch success rate has improved since 2013.
- In terms of orbits, **ES-L1, GEO, HEO** and **SSO** has 100% success rate.

Visual Analysis

- Launch sites should be located nearby the equator to save launch cost.
- **KSC LC 39A** is the most appropriate launch site, because it has the most successful outcomes and the highest success rate among the launch sites.

Predictive Analysis

- **Decision Tree** is the best algorithm for prediction of launch outcomes, because it has shown the best scores in evaluation among all the models.

Appendix A

Distance from Launch site to Proximities

- In terms of safety and transportation, the distance from launch sites to the proximities should be taken into consideration.
- I marked down the coordinates of the proximities like city and railroad, and calculated the distances form the launch site, KSC LC-39A.

City(Titusville): 5.6km **Closest Railroad: 0.7km**

Appendix B

Confusion Matrix of Decision Tree Algorithm

- As shown below, False Positive should be improved.