APPLIED DATA SCIENCE CAPSTONE PROJECT

Navaneeth Krishna P

16 September 2024

EXECUTIVE SUMMARY

This project analyses the SpaceX's launch outcomes to assess factors influencing successful landings on drone ships. The analysis found that that payload mass and booster version are significant factors in determining the landing success rates. Specifically missions with payloads between 4000 kg and 6000 kg saw a higher success rate for drone ship landings in 2015. Data from SpaceX REST API and historical launch data stored in Wikipedia were used and processed using python and SQL. The study concludes that "payload mass and booster technology significantly improves the likelihood of successful landings which can further helps in determination of cost and reduction of cost".

TABLE OF CONTENTS

INTRODUCTION

SpaceX advertises falcon 9 rocket launches on its website, with a cost of 62 million dollars and 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 space X for rocket launch. Therefore, we try to answer these following questions such as

A. How do variables obtained from the spaceX launch data affect the success of first stage landing?

B. What is the accuracy and scores of algorithm applied?

C. Which algorithm can be best model fit for binary classification in this case?

METHODOLOGY

1. Data Collection

2. Data Wrangling

3. Exploratory Data Analysis with SQL

4. Exploratory Data Analysis with Visualisation

5. Interactive visual analytics using folium and plotly dash

6. Predictive Analysis

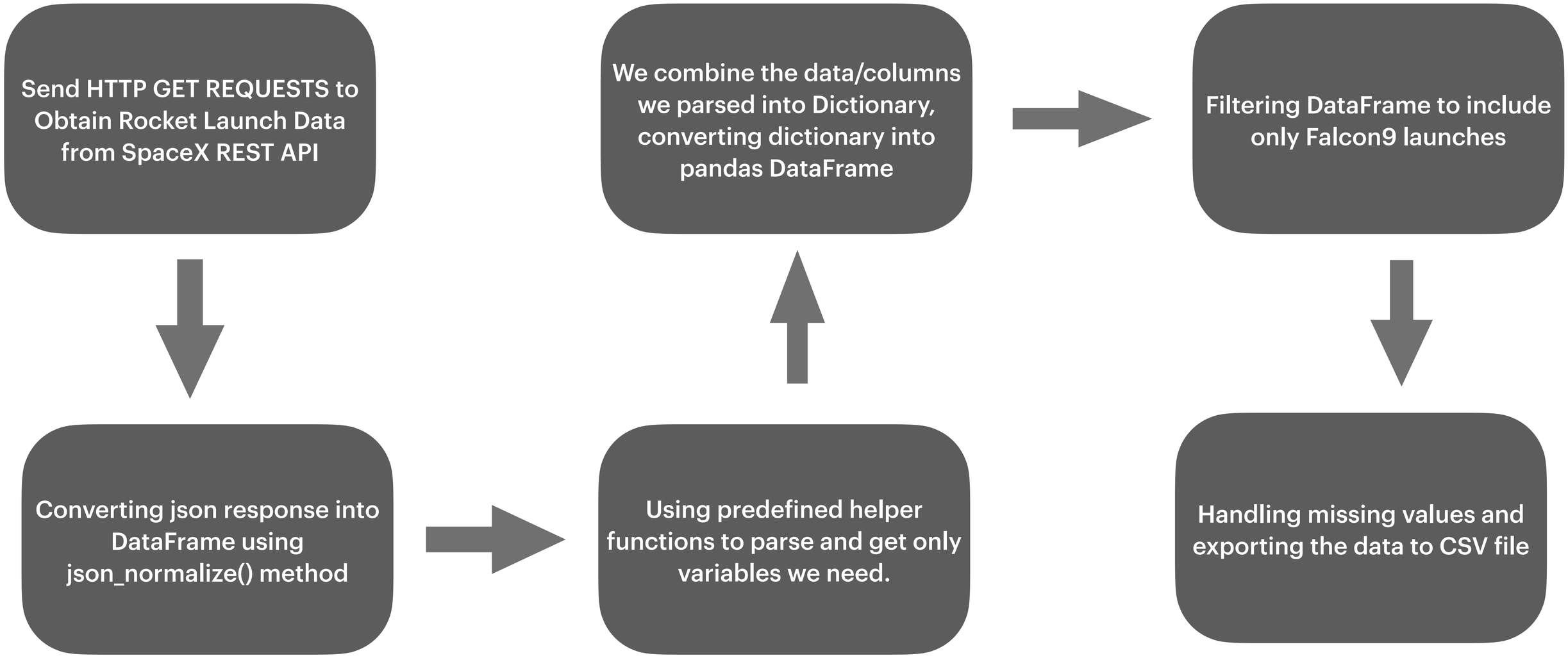
1. Data Collection

The data sources for this project are primarily from SpaceX REST API and then we scraped data from web particularly from Wikipedia. We used both of them for our data collection. We had to use both of these sources in order to to get complete information about the launches for a more detailed analysis.

Data Collection - API

- The spaceX REST API were used.
- This API is publicly available for anyone to use.
- We use the helper functions to help us use the API to extract info using identification numbers in the launch data
- We send HTTP GET REQUESTS to SpaceX REST API launch data and parse it
- Convert the json response into pandas DataFrame.
- Filter the DataFrame to include only Falcon9 launches
- Handle missing values
- Export to CSV file

Flow chart for API calls



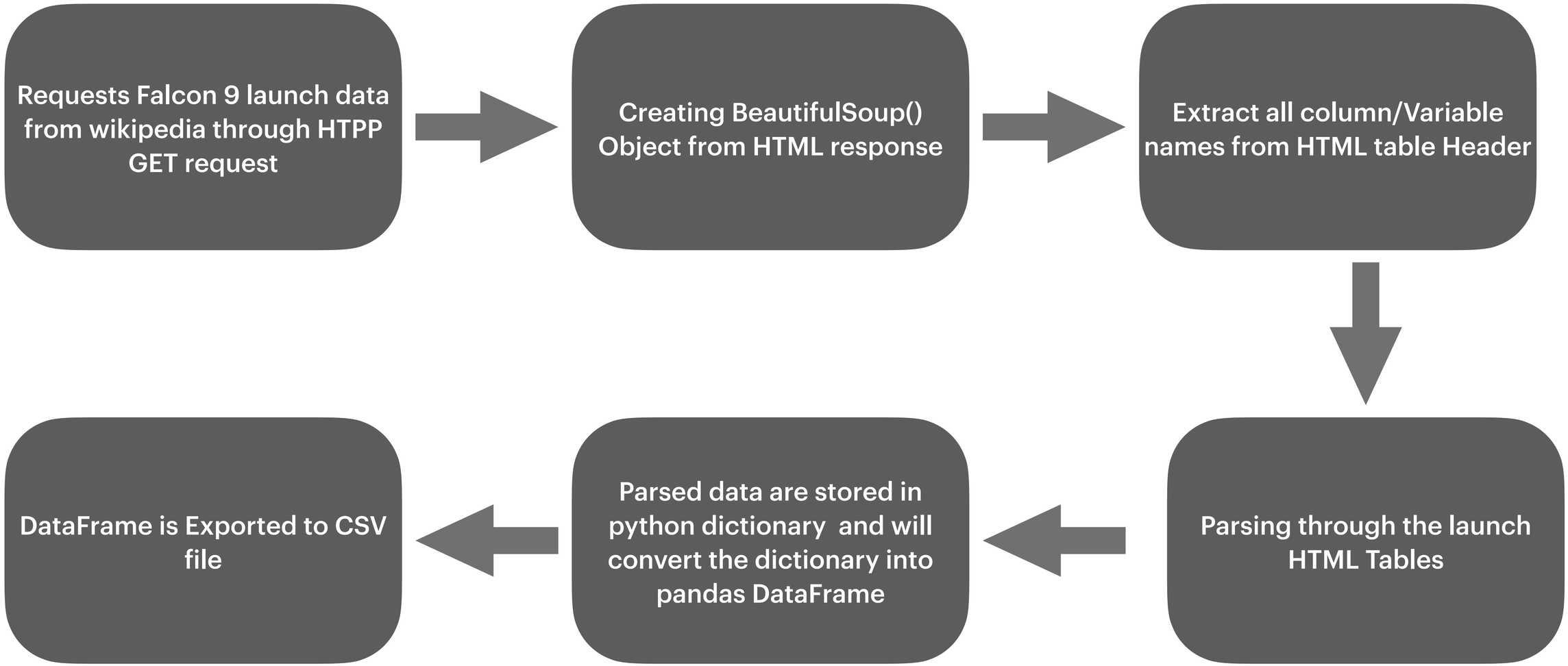
GitHub

Data Collection - Web

Performing web scraping to collect Falcon 9 historical records from Wikipedia page titled “List of Falcon 9 and Falcon Heavy Launches”

- Extract Falcon 9 launch records from HTML table from Wikipedia
- Parse the table
- Converting the table into pandas DataFrame

Flow chart for Web Scraping



GitHub

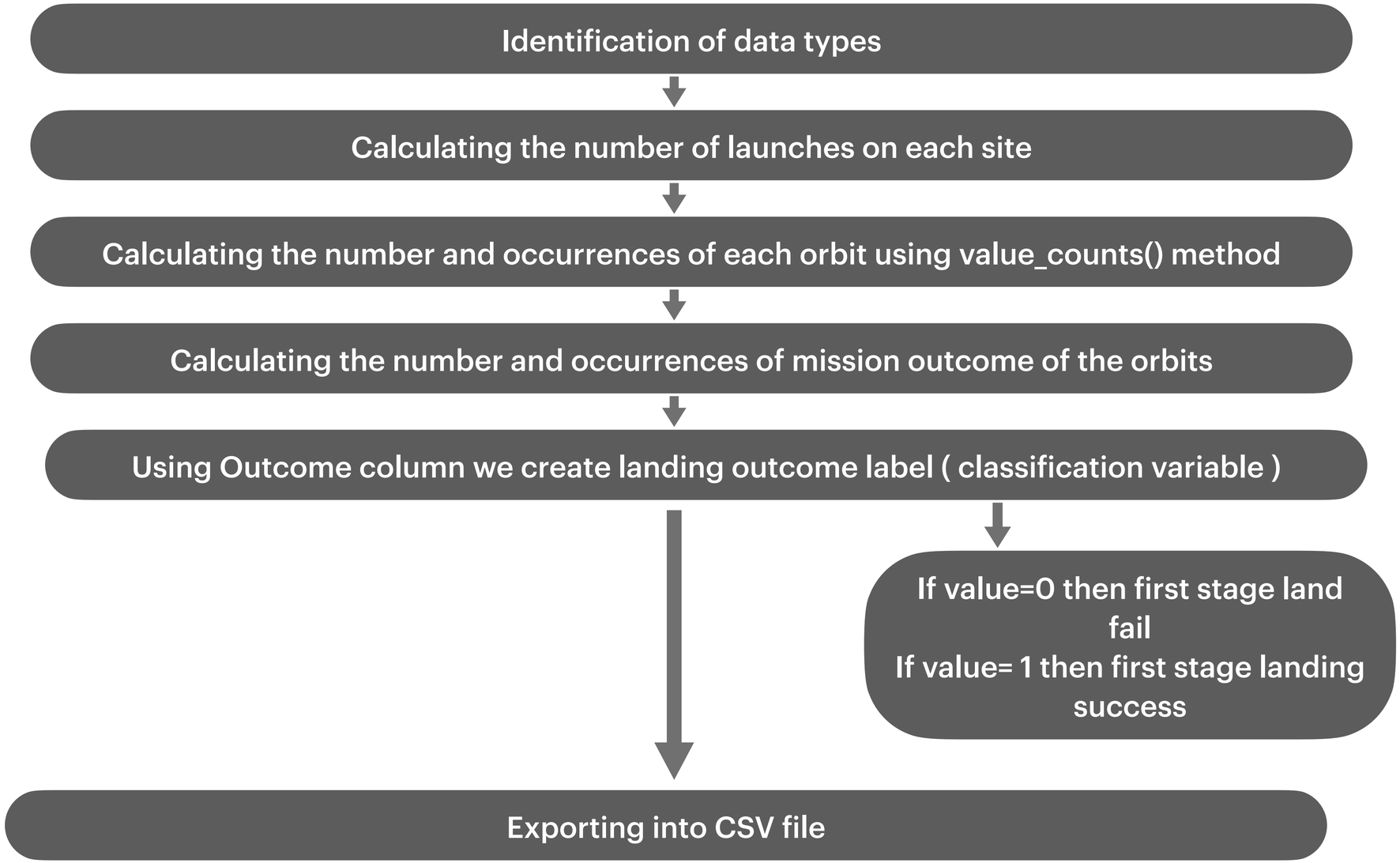
2. Data Wrangling

Data Wrangling refers to the process of transforming raw data into more usable format for analysis and other purposes. It is also known as data cleaning, data munging , data remediation, or data janitor work.

In this section we usually

- Perform Exploratory Data Analysis (EDA) to identify some patterns
- Determine labels for training supervised models

Flow chart for Data Wrangling



GitHub

3. Exploratory Data Analysis - SQL

We use SQLite3 which is a lightweight, self contained, serverless relational database management system. It implements SQL standard and is widely used because of its simplicity and minimal setup requirements.

Key characteristics of SQLite3 are:

- It is serverless : entire database is stored in a single file and the SQLite Library directly interacts with that file directly.
- Cross platform : can be used across various OS like Windows, Mac,Linux etc
- SQL compatibility
- Embedded in many software applications including browser like Firefox and google chrome
- Useful for small projects, prototypes and applications that doesn’t respire high concurrency and extensive scaling

EDA Tasks performed using SQL

1. Displaying names of unique launch sites in mission
2. Display 5 records where launch sites began with the string “CCA”
3. Display total payload mass carried by boosters launched by “NASA(CRS)”
4. Displaying Average payload mass carried by booster version F9 v1.1
5. Displaying first successful landing outcome in ground pad.
6. Listing the names of boosters which have success in drone ship and have payload mass between 4000 and 6000
7. Listing total number of successful and failure mission outcomes
8. Using sub-query listing names of booster versions which had carried maximum payload mass
9. Listing records of failure outcomes and its related attribute on year 2015
10. Rank the count of landing outcomes(I.e., success or failure) between the date 2010-06-04 and 2017-03-20

GitHub

Displaying Unique Names

Query and Output

```
%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

Explanation

- We got 4 unique launch sites

Displaying records of launch sites with “CCA”

Query and Output

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

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

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

Explanation

- Simple query to obtain 5 cases where we filtered only to include rocket launch site that begin with “CCA”

Sum total of payload mass carries by NASA(CRS)

Query and Output

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

```
[ ] %sql SELECT SUM(PAYLOAD_MASS__KG_) AS totalPayloadMassfromNASA FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
totalPayloadMassfromNASA
45596
```

Explanation

- Total Payload Mass carried by all the Boosters from NASA(CRS) is 45,596Kg

Average Payload Mass of Booster Version F9 v1.1

Query and Output

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS averagePayloadMass FROM SPACEXTBL WHERE Booster_Version like '%F9 v1.1%';
```

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

averagePayloadMass
2534.6666666666665

Explanation

- Average Payload Mass of Booster Version F9 v1.1

First Successful Landing in Ground Date

Query and Output

```
%sql SELECT MIN(date) AS First_successful_landing FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_successful_landing
2015-12-22

Explanation

- First successful ground landing date is 2015-12-22

Successful Drone Ship Landing with payload between 4000 and 6000

Query and Output

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome like "Success (drone ship)" AND (PAYLOAD_MASS__
```

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

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

Explanation

- We got only 4 Booster Version where the landing outcome is success

Total Number of Successful and Failure Mission Outcomes

Query and Output

```
%sql SELECT Mission_Outcome, count(*) AS total_number FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Explanation

- There is 100 success outcome and 1 is failure

Boosters Carried Maximum Payload

Query and Output

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXT
 * sqlite:///my_data1.db
Done.
```

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

Explanation

- We got 12 different boosters version that have maximum payload mass

2015 Launch Records

Query and Output

```
[ ] %%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.
```

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)

Explanation

- We have 2 failed landing outcomes in Drone ship in 2015.
- Both outcomes was launched from site CCAFS LC 40

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

Query and Output

```
%%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.
```

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

Explanation

- We have rank landing outcomes between given year

GitHub

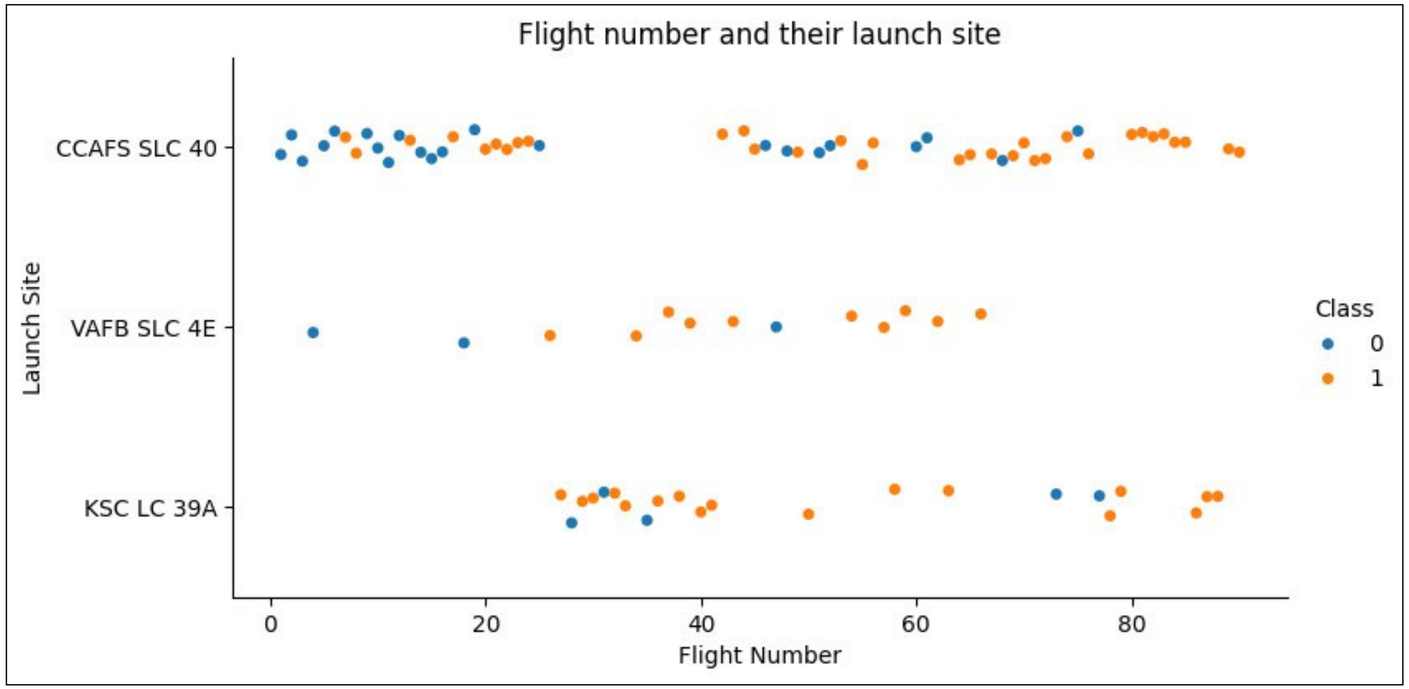
4. Exploratory Data Analysis - Visualisation

Perform EDA and prepare feature engineering using pandas and matplotlib library

Observations:

- Newly launched flights has higher rate of success
- VAFB SLC 4E and KSC LC 39A has higher success rate than CCAFS SLS 40
- Higher the payload mass the higher the success rate as we can see payload mass with >6500 kg has only 1 failure observed
- KSC LC 39A has recorded more success for <4000kg of payload mass
- Orbits with 100% rate of success are ES-L1,GEO,HEO and SSO
- SO is the only orbit with 0% success rate
- Orbits between 40 to 80% of success rate are GTO,ISS,MEO,PO and LEO
- Orbits between 80 to 90% of success rate are VLEO
- After year 2013 we can see the rise in success rate
- Created dummy variables to categorical columns and casted all numerical columns to float64 dtype

Flight Number vs Launch Site



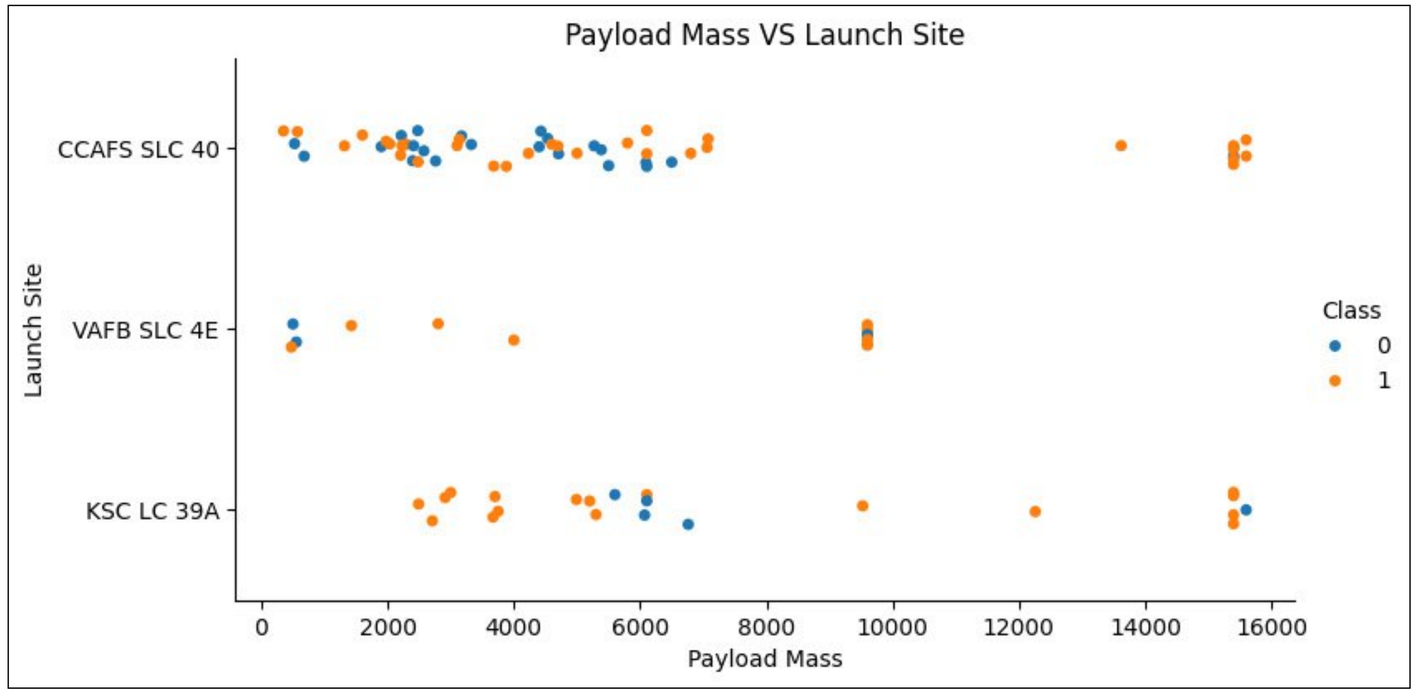
Blue markers = class 0 & failed first stage landing

Orange markers = class 1 & successful landing of first stage

Observations:

- KSC LC 39 has highest success rate
- Success rate are varying for different launch sites
- We can observe that as the number of flight increases the success rate increases.

Payload Mass vs Launch Site



Blue markers = class 0 & failed first stage landing

Orange markers = class 1 & successful landing of first stage

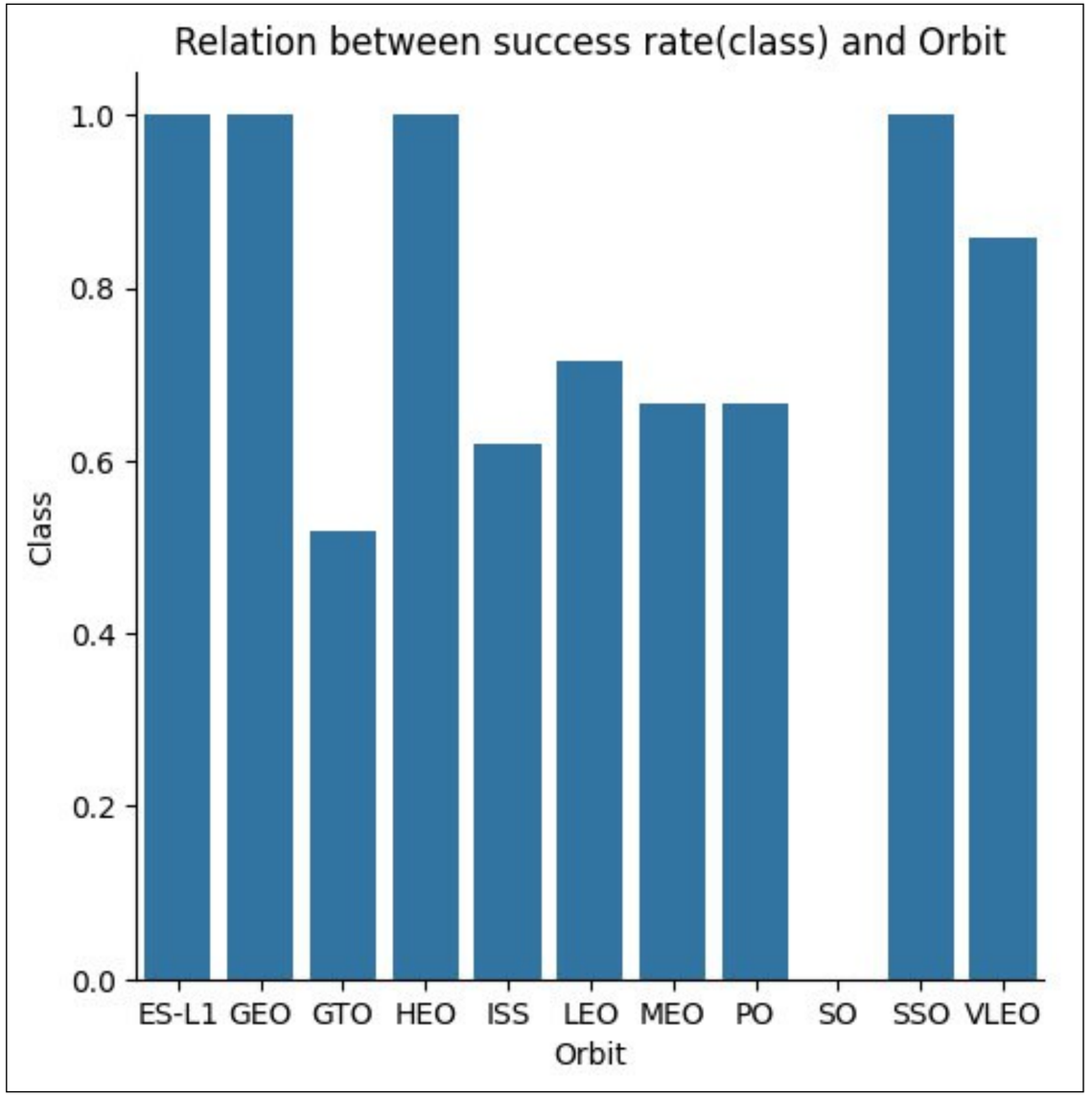
Observations:

- Higher the payload mass = higher the success rate
- Payload mass >6500 kg has only 1 failure case
- KSC LC 39A has recorded more success for <4000kg of payload mass

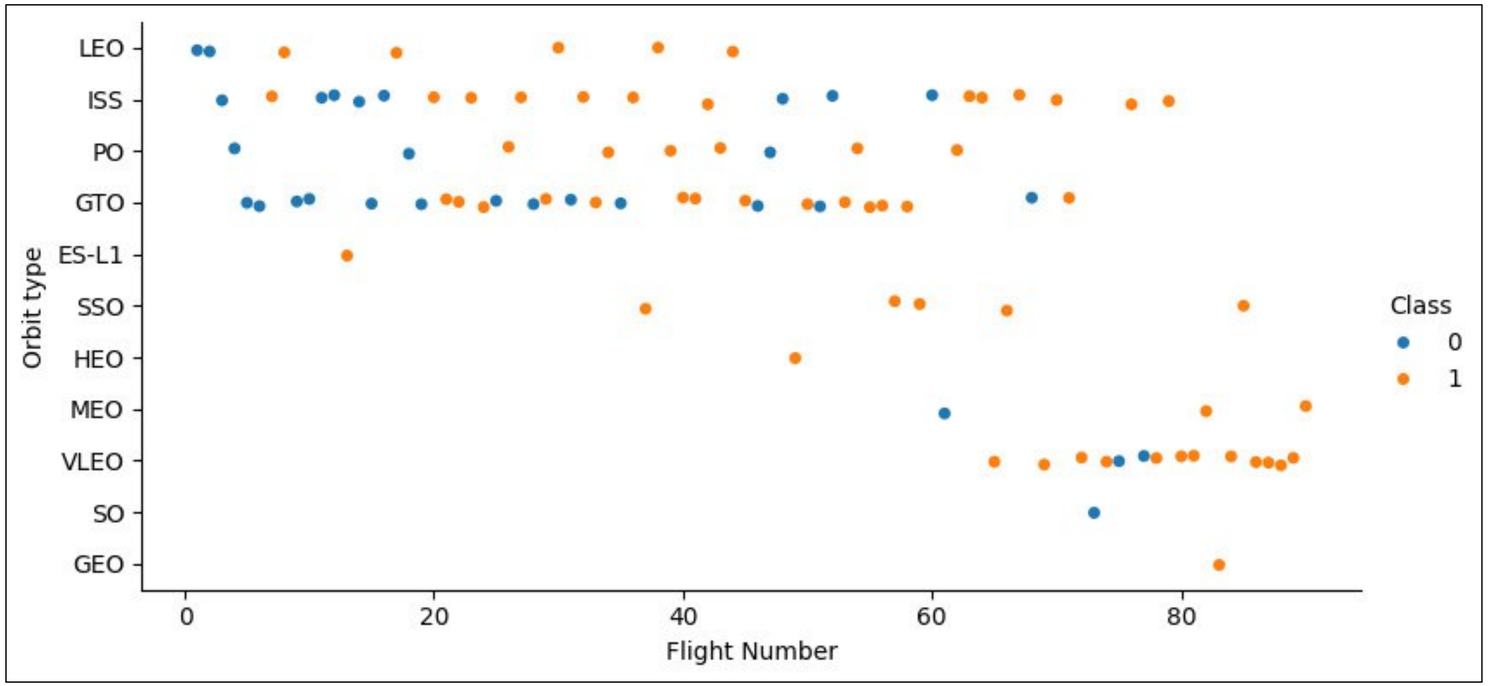
Success Rate vs Orbit Type

Observations:

- Orbits ESC L1, GEO,HEO and SSO has 100 percent success rate
- Orbit SO has 0% success rate
- Orbits GTO,ISS,MEO,PO and LEO has between 40 to 80%
- Orbit VLEO has probability of success between 80 to 90%



Flight Number vs Orbit Type



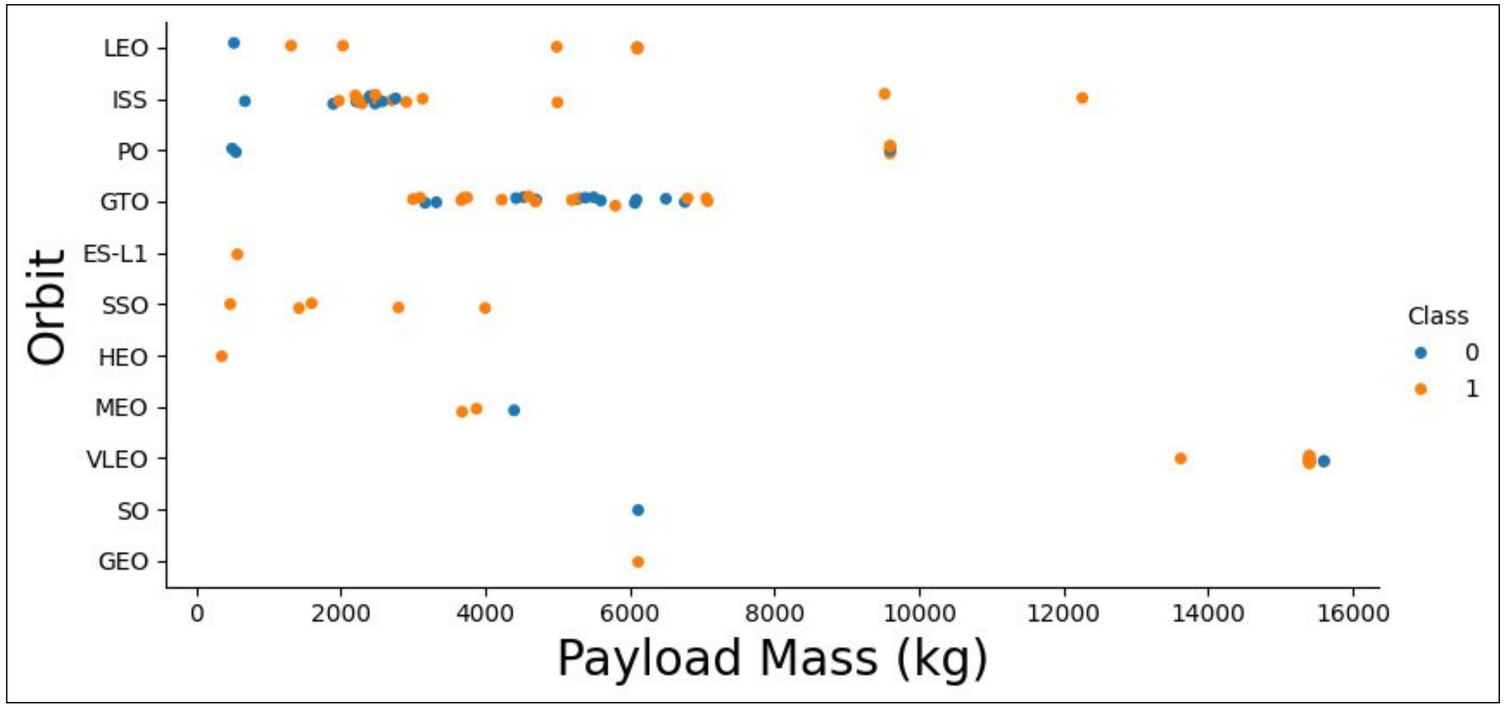
Blue markers = class 0 & failed first stage landing

Orange markers = class 1 & successful landing of first stage

Observations:

- We can see the relationship between the variables as the largest flight number has highest success rate
- We can also we the rate of failure decreases after flight number 60

Payload Mass vs Orbit Type



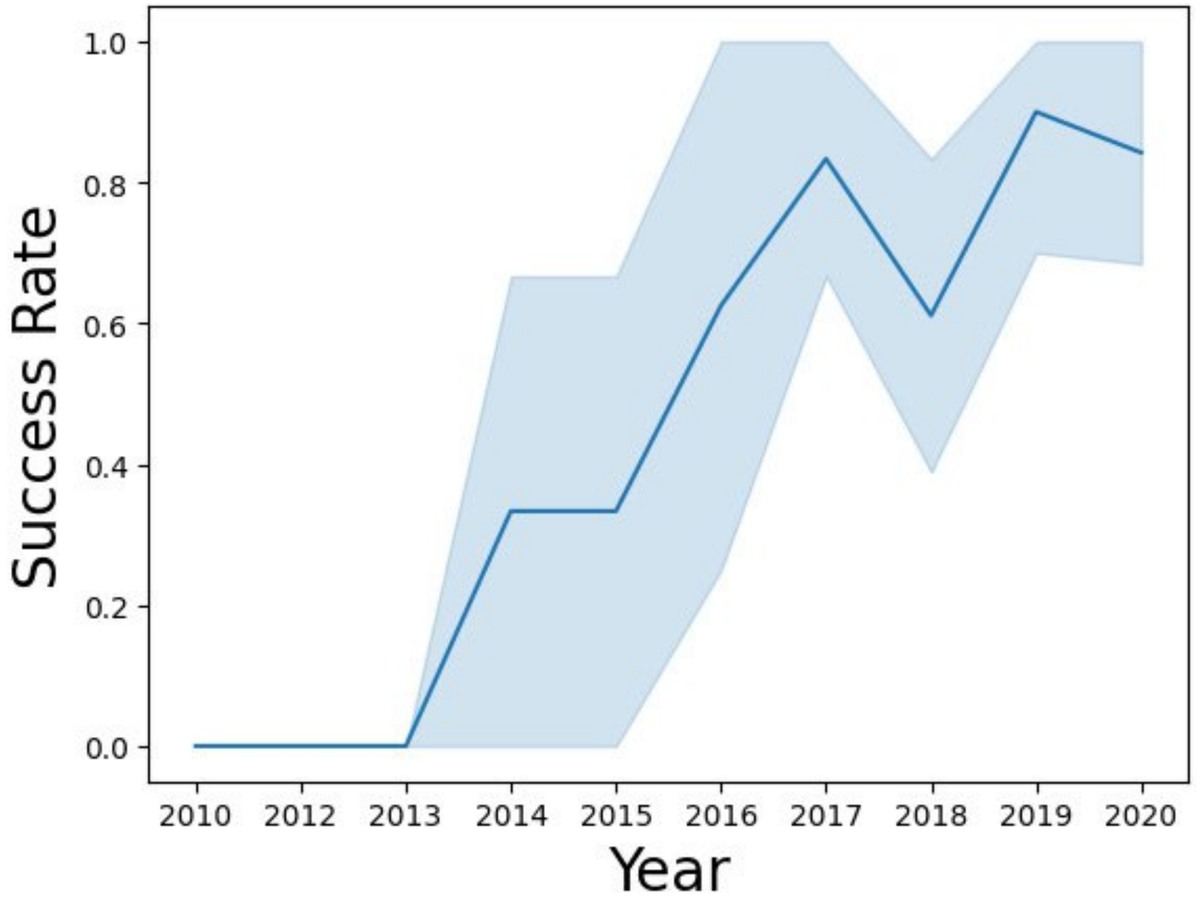
Blue markers = class 0 & failed first stage landing

Orange markers = class 1 & successful landing of first stage

Observations:

- Some orbits has better success rate than others such as SSO,HEO,LEO,GEO and others
- Some has 100% success some has 0% success but 0% success has only 1 launch happened
- Payload mass and orbit has obvious no correlation

Launch Success Yearly Trend

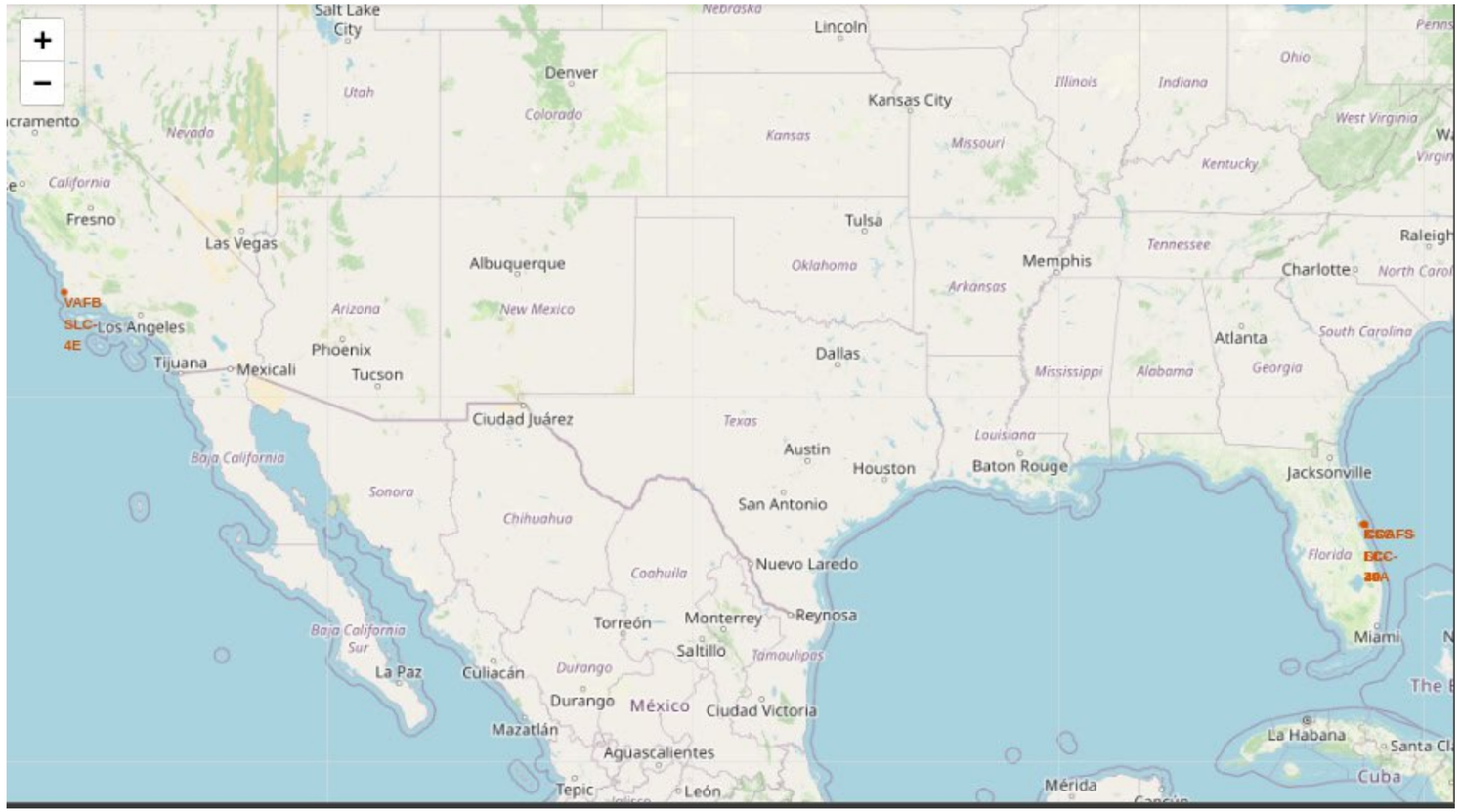


Observations:

- Success rate increases over year
- We can observe success rate was 0 but it increases significantly from the year 2013

5. Interactive visual analytics using Folium and Plotly dash

All Launch Sites Location

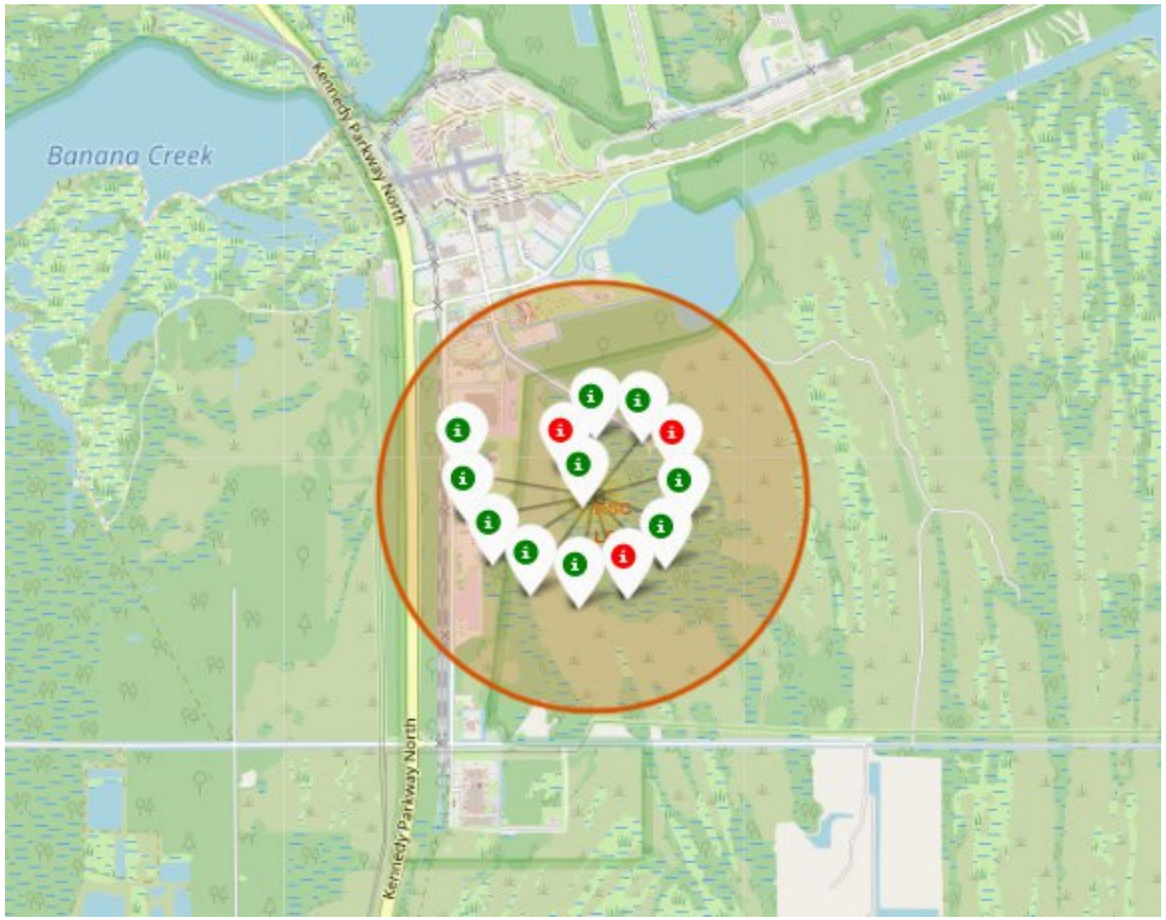


Launch site locations:

- VAFB SLC-4E(California)
- KSC LC-39A(Florida)
- CCAFS LC-40(Florida)
- CCAFS SLC-40(Florida)

Success and Failed Landings Markers

Launch site :
KSC LC-39A

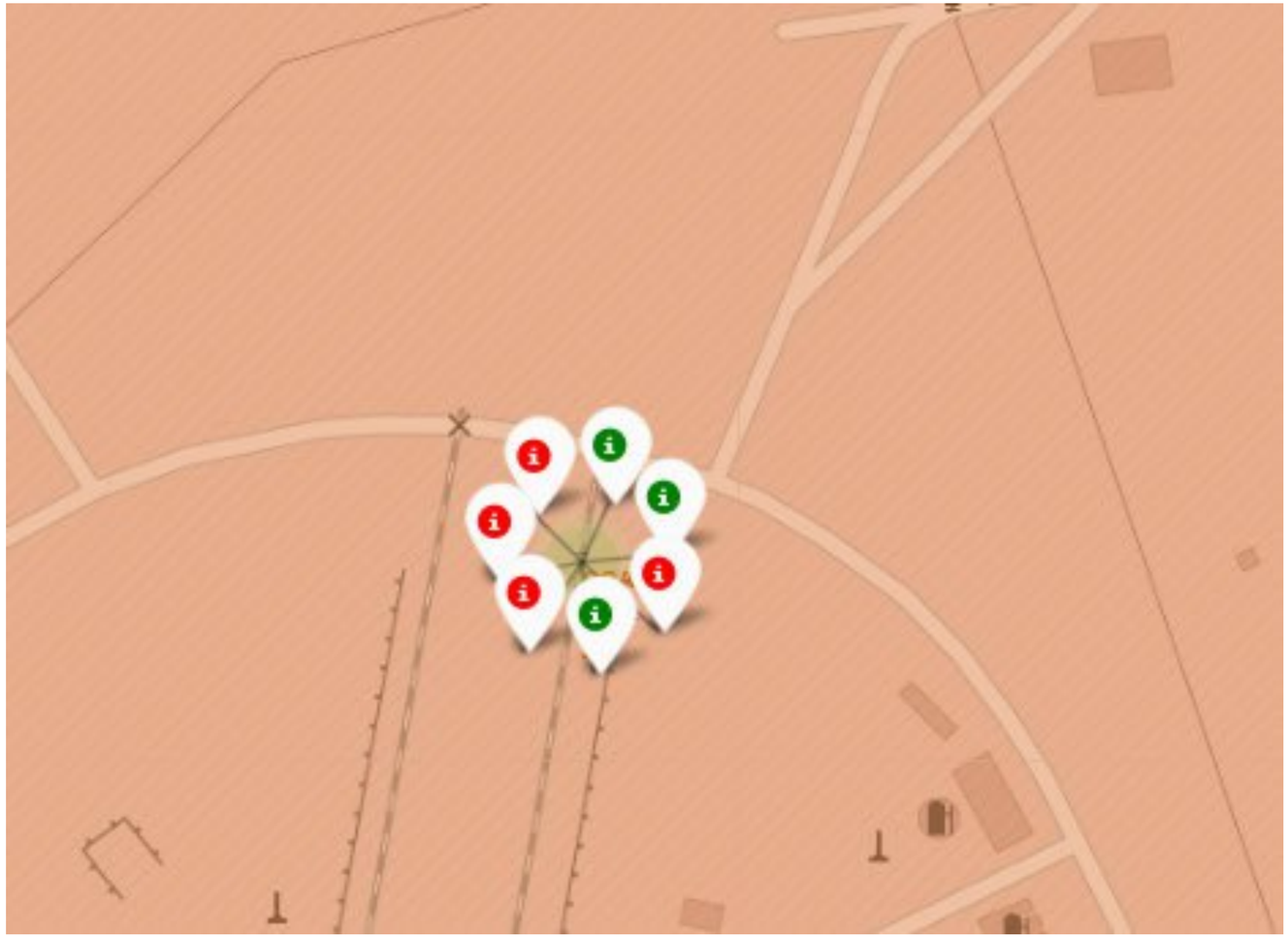


Note:

Red Marker = Failed Launch

Green Marker = Successful Launch

Launch site:
CCAFS SLC 40



Note:

Red Marker = Failed Launch

Green Marker = Successful Launch

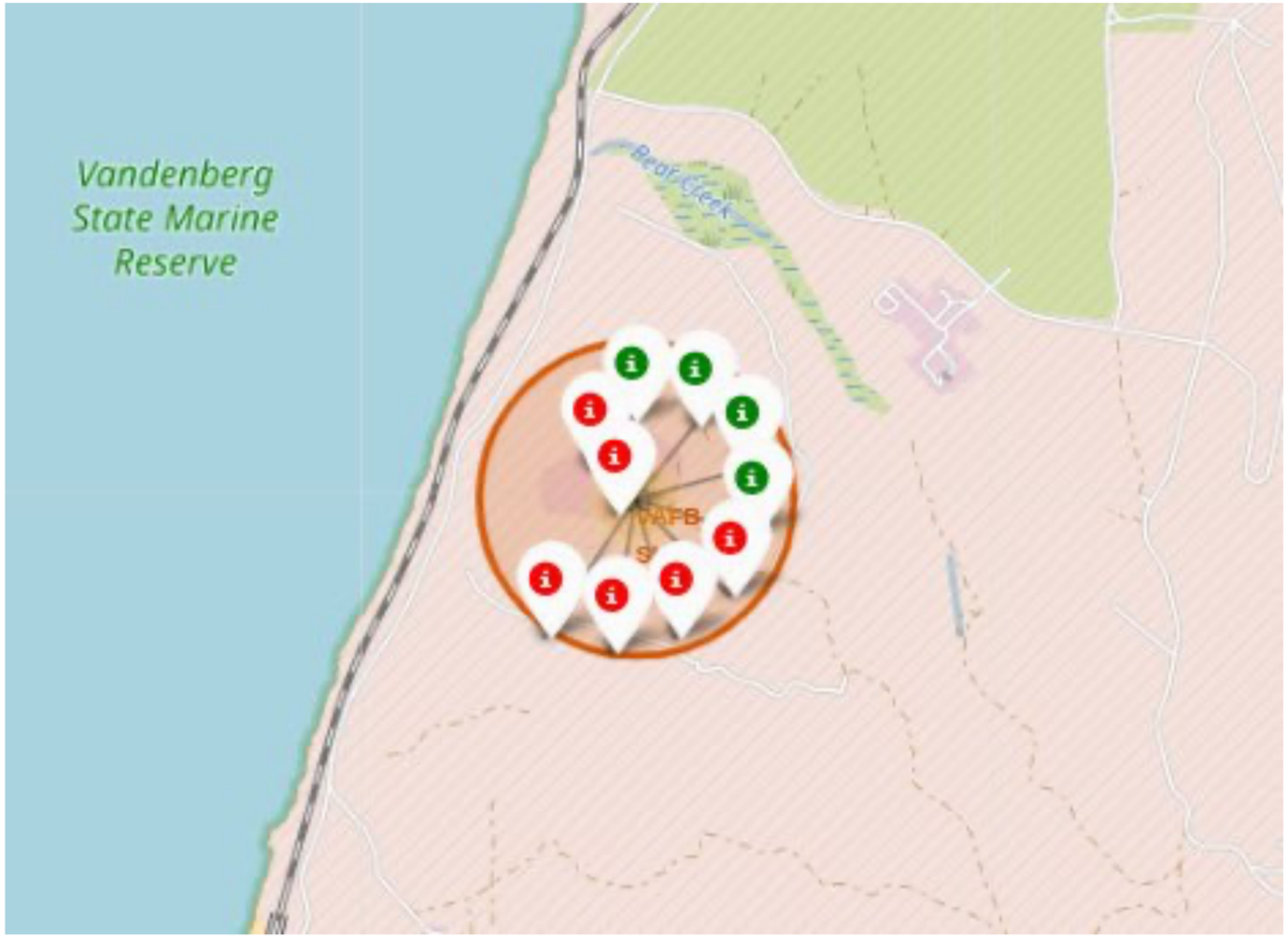
Launch site:
CCAFS LC 40

Note:

Red Marker = Failed Launch

Green Marker = Successful Launch

Launch site:
VAFB SLC 4E

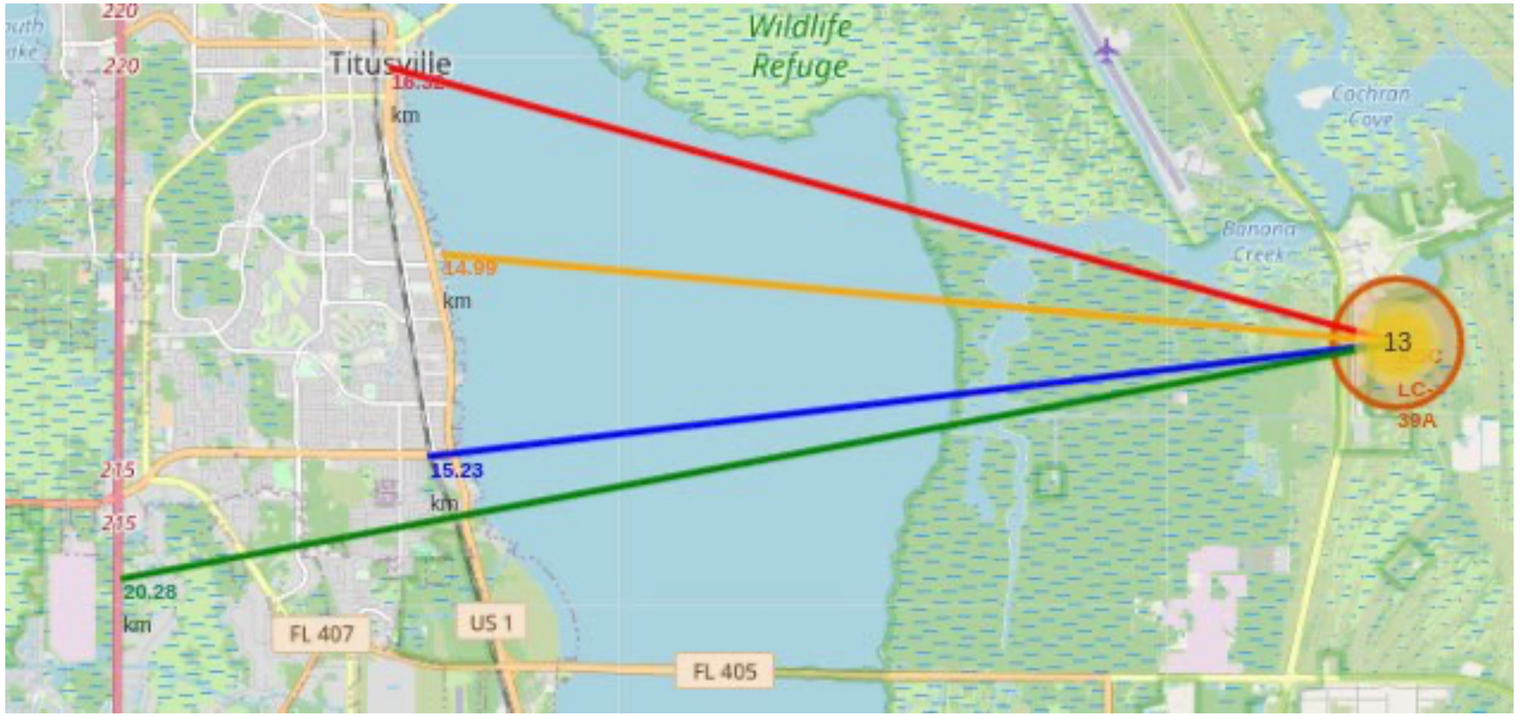


Note:

Red Marker = Failed Launch

Green Marker = Successful Launch

Distance from the launch site KSC LC-39A to its proximities



KSC LC-39A

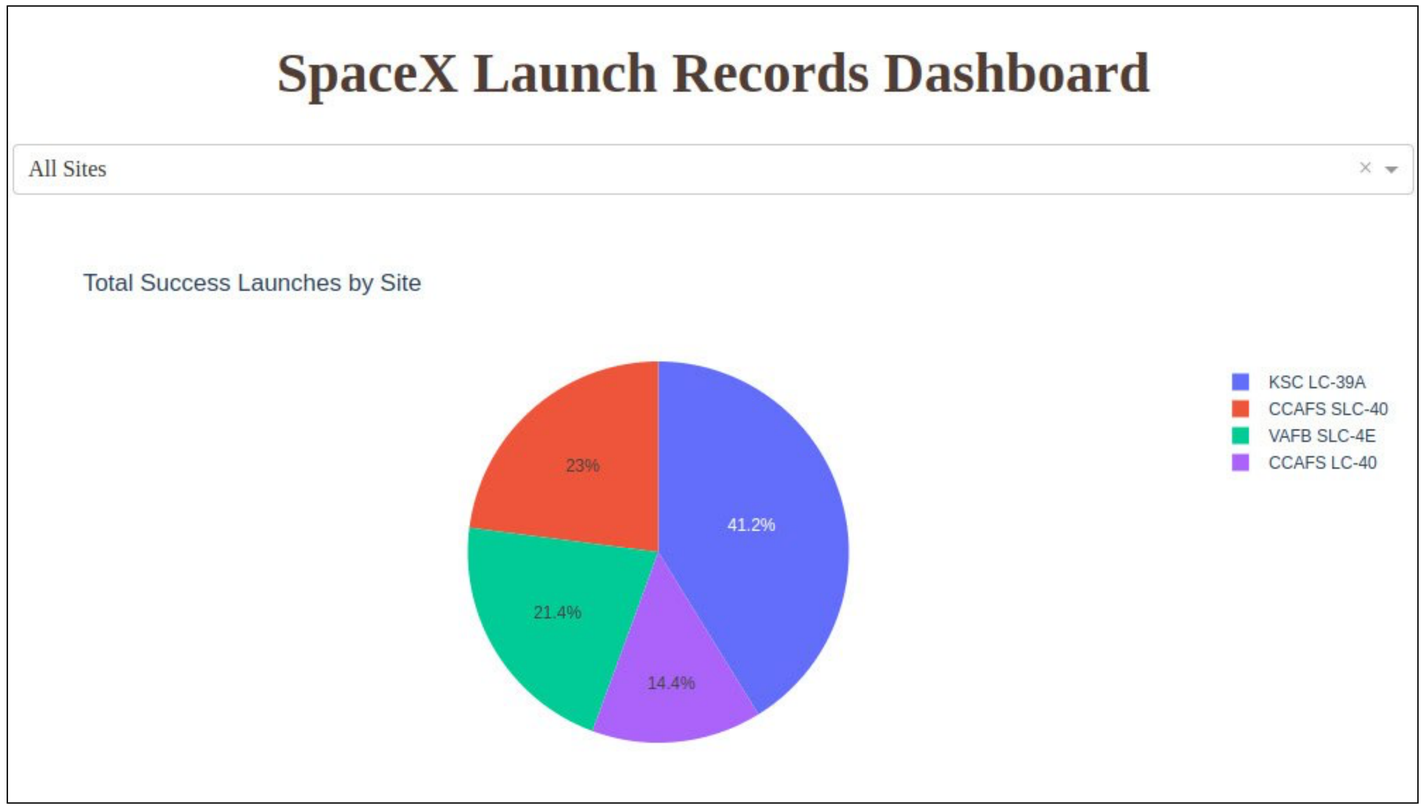
Explanation

- We can see from the visuals launch site KSC LC 39A is close to railway,Highway and titusvile street

GitHub

Build dashboard with Plotly Dash

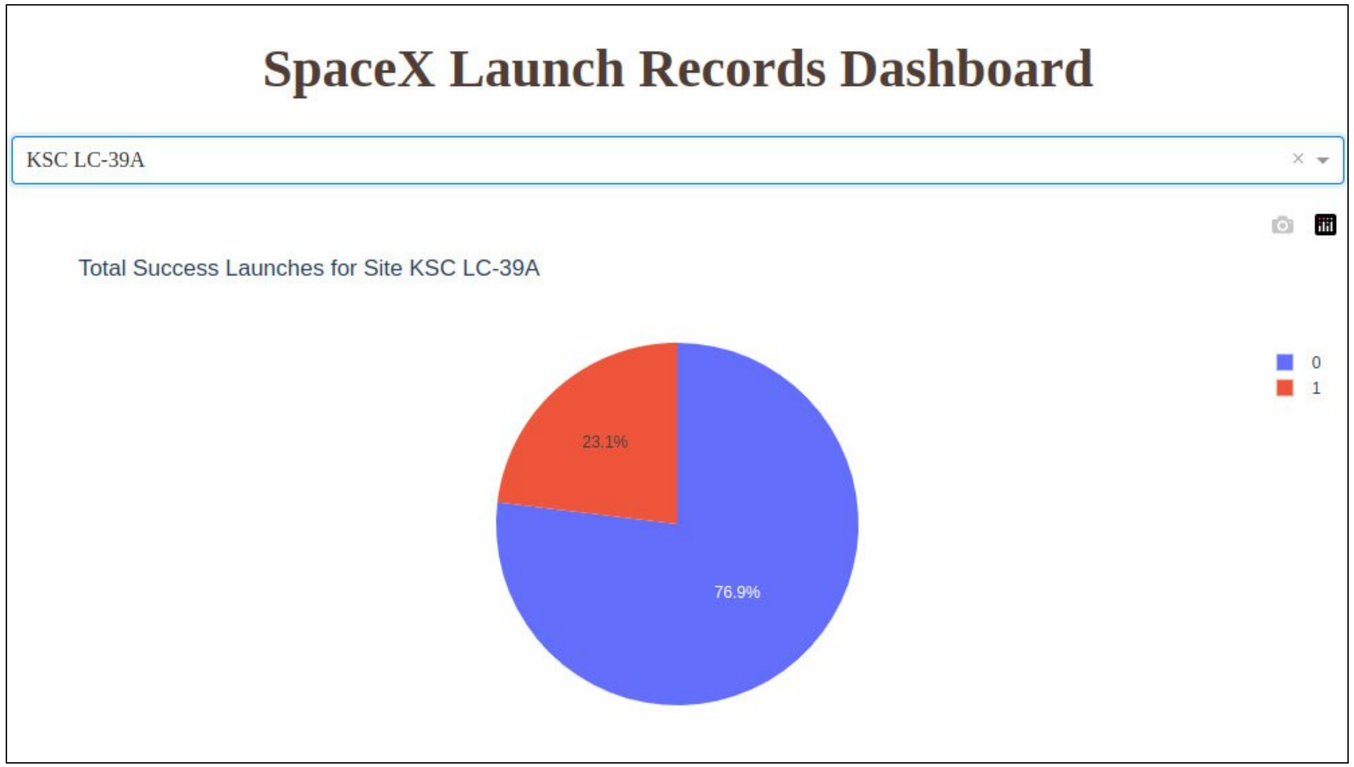
Launch Success Count for all Sites



Explanation:

- Drop down menu allows us to select each launch sites success/ failure
- The pie chart displays distribution of each launch sites outcome
- We can observe KSC LC 39A has highest successful launch

Launch Site with Highest Launch Success Ratio



Explanation:

- We can observe that launch site KSC LC 39A has 76.9% of success rate with 10 successful and 3 failed landings

Payload and Success for every site

Explanation:

- The chart shows success and payload mass for every launch site and every booster version category

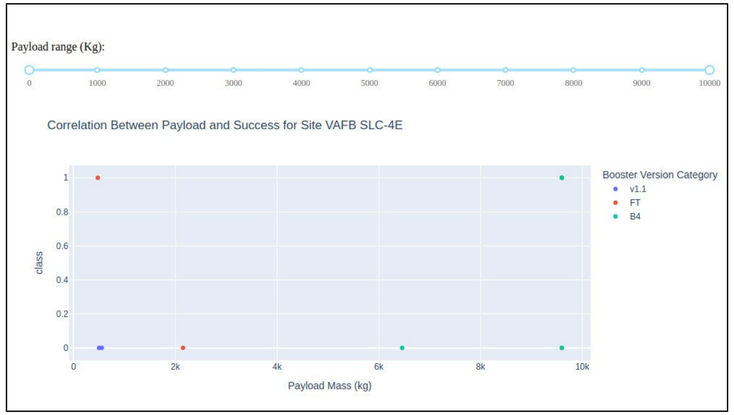
Payload and Success for each Launch Sites

KSC LC-39A

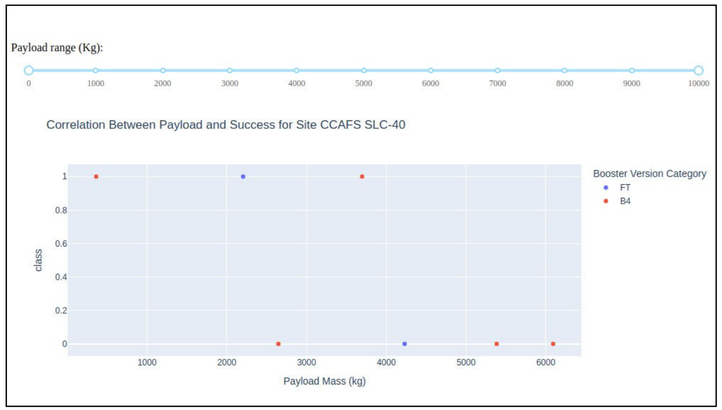
CCAFS LC 40

Explanation:

- This chart shows payload and success rate for every launch site with different booster version



VAFB SLC 4E



CCAFS SLC 40

GitHub

6. Predictive Analysis

Classification

Classification Accuracy

Examination of Test Set scores and accuracy

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

Observations:

- From the examination of test set scores and accuracy are Similar across every model
- It may be due to small test sample size(n=18 samples)
- Therefore we need to perform the application of model on total dataset

Examination of accuracy and score of entire data set

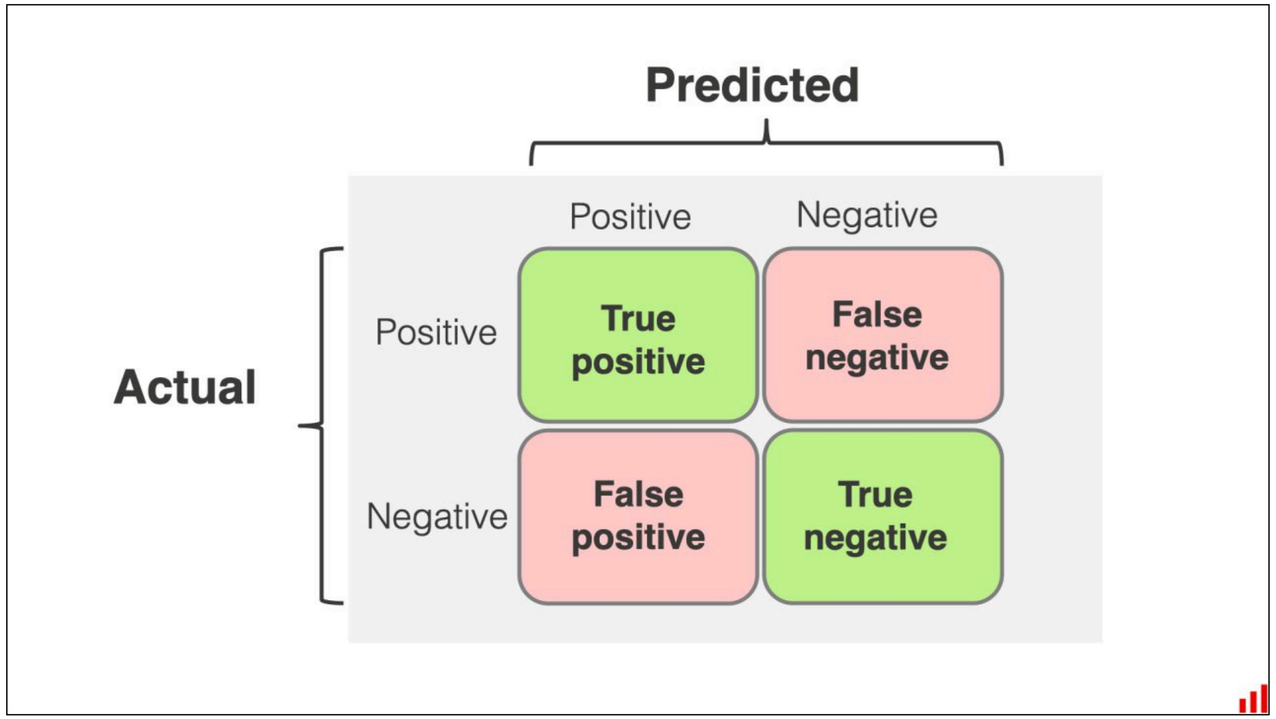
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.855072	0.819444
F1_Score	0.909091	0.916031	0.921875	0.900763
Accuracy	0.866667	0.877778	0.888889	0.855556

Observations:

- Upon performing model fitting and prediction on entire dataset we can see the differences of accuracy and scores
- The fluctuations we see is not in extreme values
- We can observe that Decision Tree algorithm performs best and shown highest score and accuracy comparatively other models

GitHub

Confusion Matrix

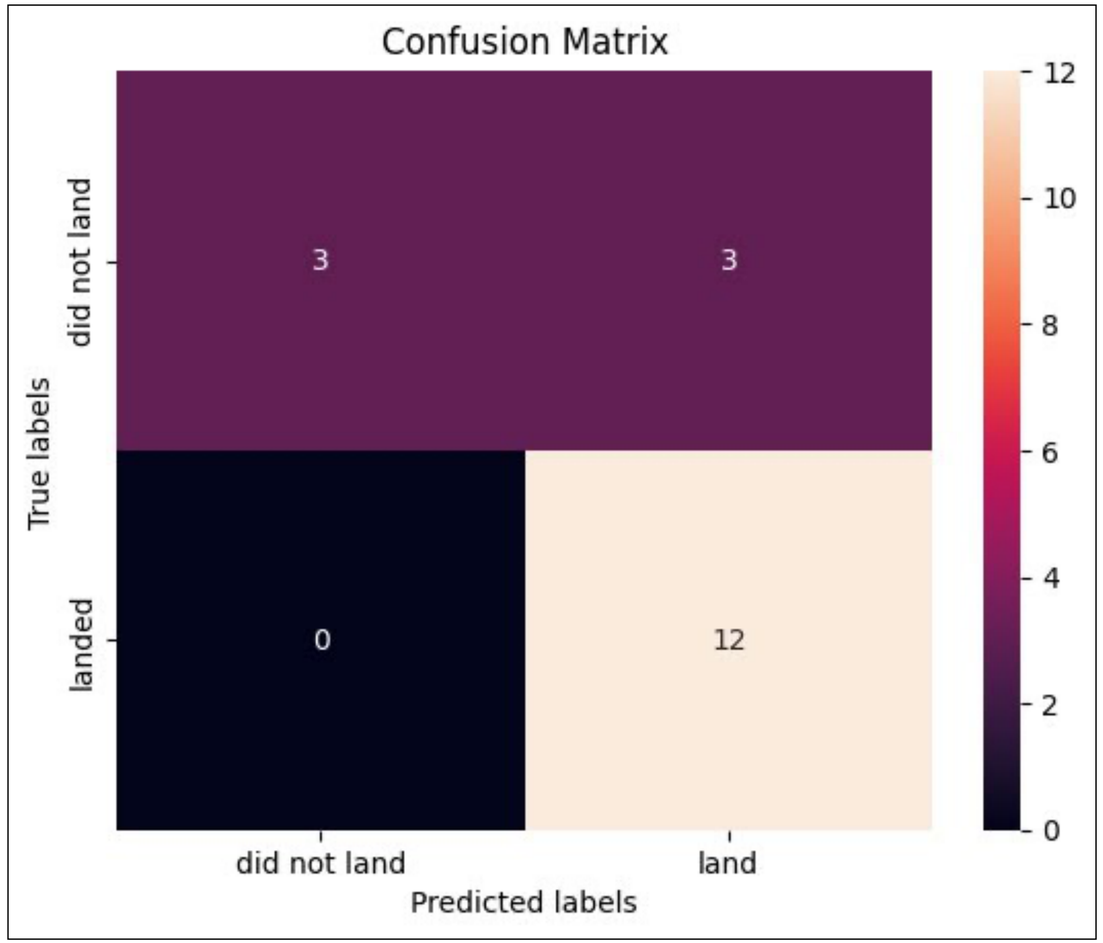


Confusion Matrix:

A table/2d matrix that compares a classifier models predicted values to the actual values in a data set to evaluate the models performance, it is also known as error matrix.

Confusion Matrix helps you understand how well your model is performing by showing true positives, true negatives, false positive false negatives

- True Positives = the number of predictions that are both correct and positive
- True Negatives = the number of predictions that are both correct and negative
- False Positives = the number of predictions that are incorrect and positive
- False Negatives = the number of predictions that are incorrect and negative



Observations:

- We have confusion matrix for the model Logistic Regression
- We have 12 True Positives and 3 True negatives
- We have 3 False positives and 0 False negatives
- Examining this confusion matrix we can say that logistic regression can distinguish between different classes

GitHub

CONCLUSION

- SpaceX Falcon 9 first stage landing outcomes is showing upward trend and if more launches are expected then more success are to be expected
- If more launches are documented (in future) the ML Algorithm models can surely predict with accuracy of >90% for first stage landing
- Decision Tree algorithm performs better for the dataset
- Larger the payload mass, the greater the success probability
- Launch site named “KSC KC-39A” has highest success rate out of all given launch sites
- Certain orbits like ESL1,GEO,HEO and SSO has 100% success rate.
- Only 1 orbit(I.e.,SO) has 0% success rate given we made only 1 launch in the site

APPENDIX

Data Source

REST API https://docs.spacexdata.com/

Wikipedia https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

GitHub Links

Data Collection API	GitHub	EDA Visualisation	GitHub
Data Collection Web	GitHub	Interactive charts Folium	GitHub
Data Wrangling	GitHub	Interactive charts Plotly	GitHub
EDA SQL	GitHub	Model prediction	GitHub

THANK YOU