IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Büşra Küden
21.07.2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- In this project we will predict if the Falcon 9 first stage will land successfully. 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- 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

- Data sets were collected as follows:
  - Request and parse the SpaceX launch data using the GET request
  - Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
  - Clean the data, checked for missing values and fill in missing values where necessary.
  - Perform web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

Data Collection – SpaceX API

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNe
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url)
data = response.json()
data = pd.json_normalize(data)
```

- Request and parse the SpaceX launch data using the GET request

- The GitHub URL of the completed SpaceX API calls notebook.

Data Collection - Scraping

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

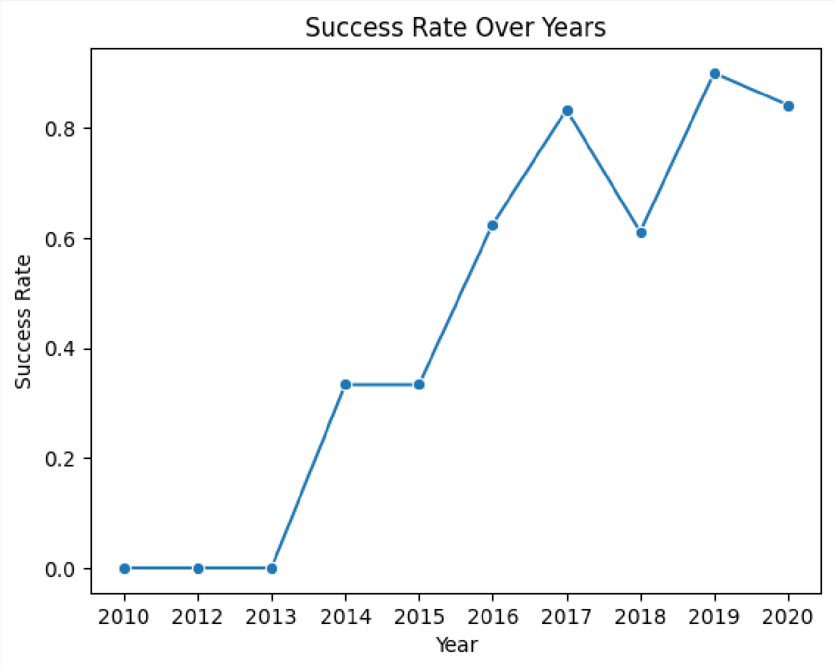
```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = launch_data[launch_data['BoosterVersion']=='Falcon 9']
```

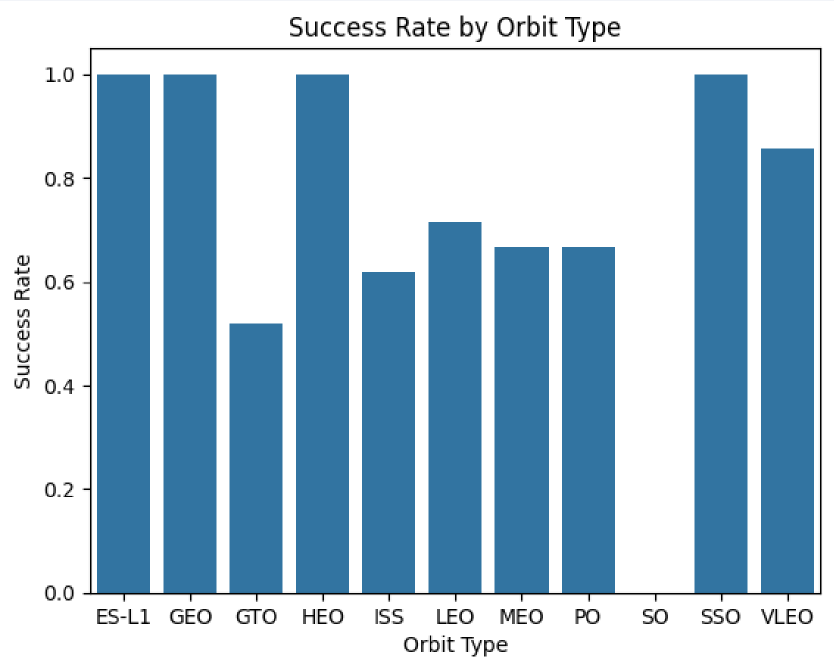
- Filter the dataframe to only include Falcon 9 launches
- The GitHub URL of the completed web scraping notebook

Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- The GitHub URL of your completed data wrangling related notebooks.

EDA with Data Visualization





- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.
- The GitHub URL of your completed EDA with data visualization notebook.

EDA with SQL

- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:

1. The names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA’
3. The total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. The date when the first succesful landing outcome in ground pad was acheived.
6. The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. The total number of successful and failure mission outcomes
8. The names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
10. 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

- The link to the notebook is: Github URL

Predictive Analysis (Classification)

- Models were trained and hyperparameters were selected using the function GridSearchCV.
- Accuracy were used as the metric for our model, improved the model using feature engineering and algorithm tuning.
- The link to the notebook is: Github URL

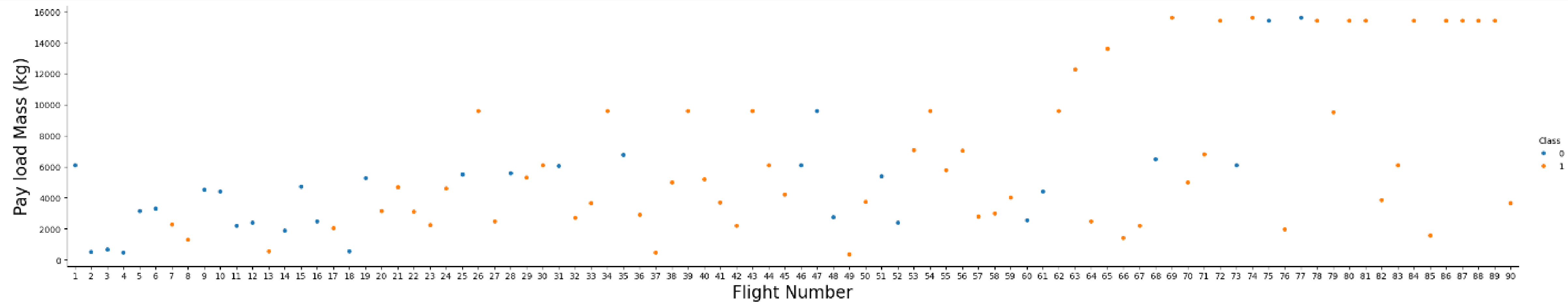
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

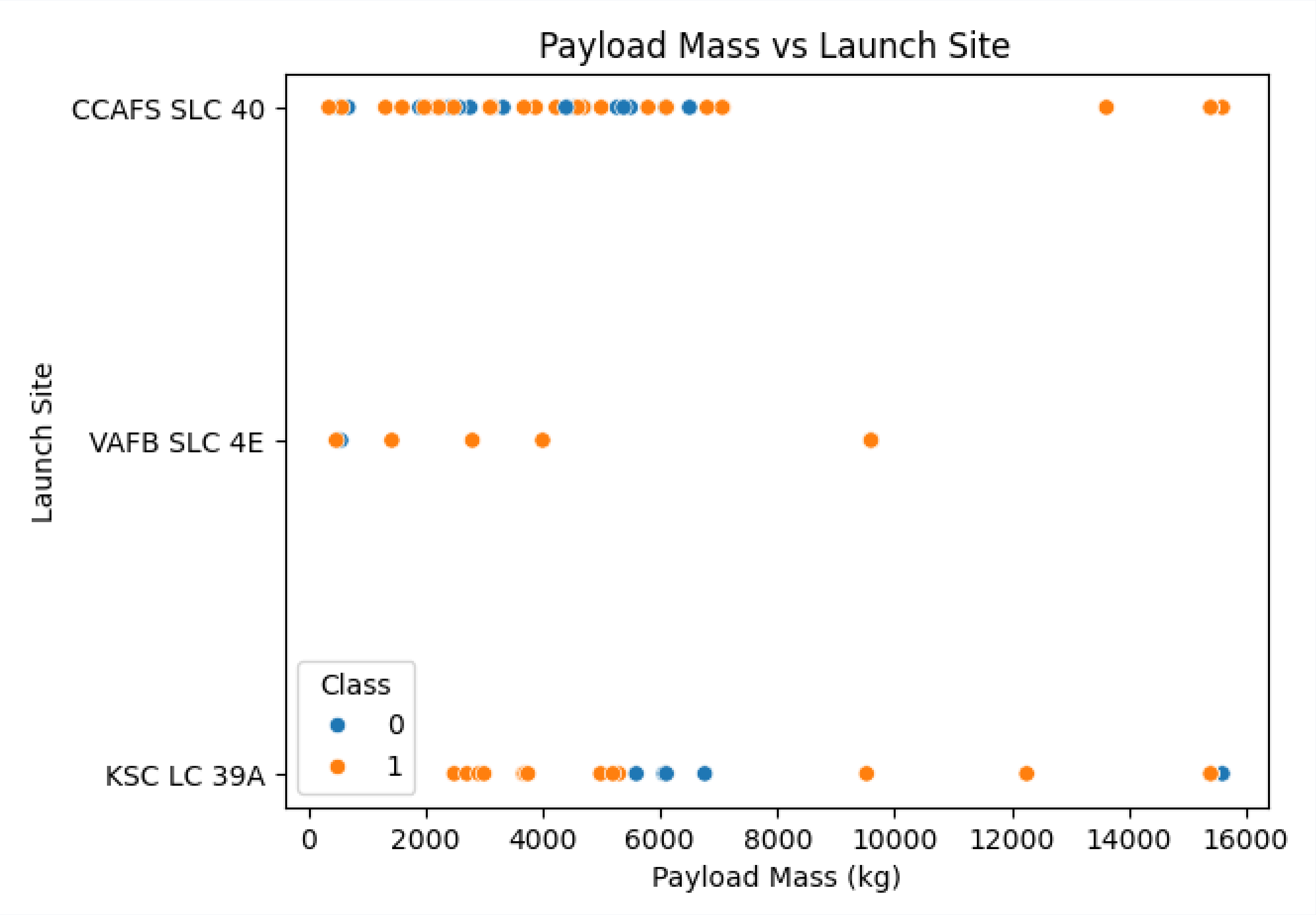
Flight Number vs. Launch Site



- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.

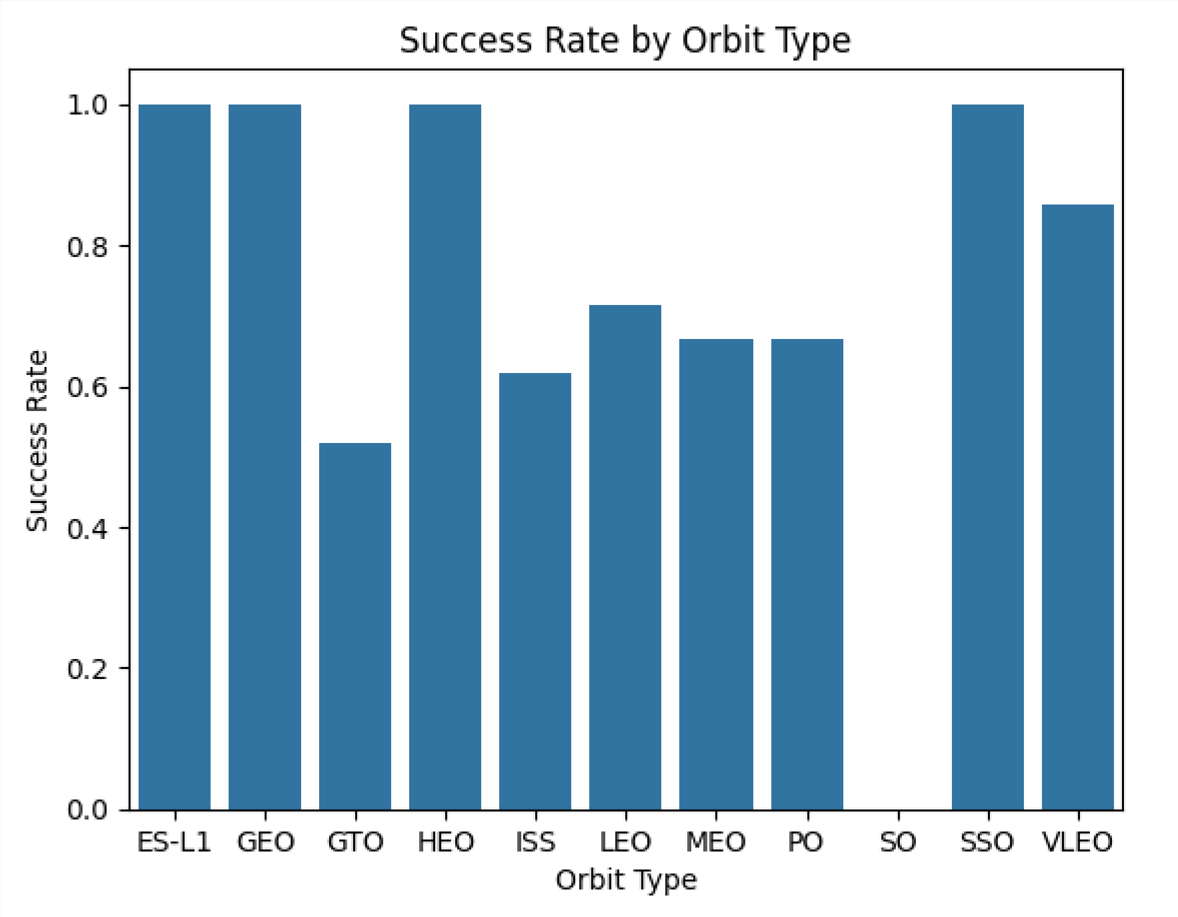
Payload vs. Launch Site

- We can observe that the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



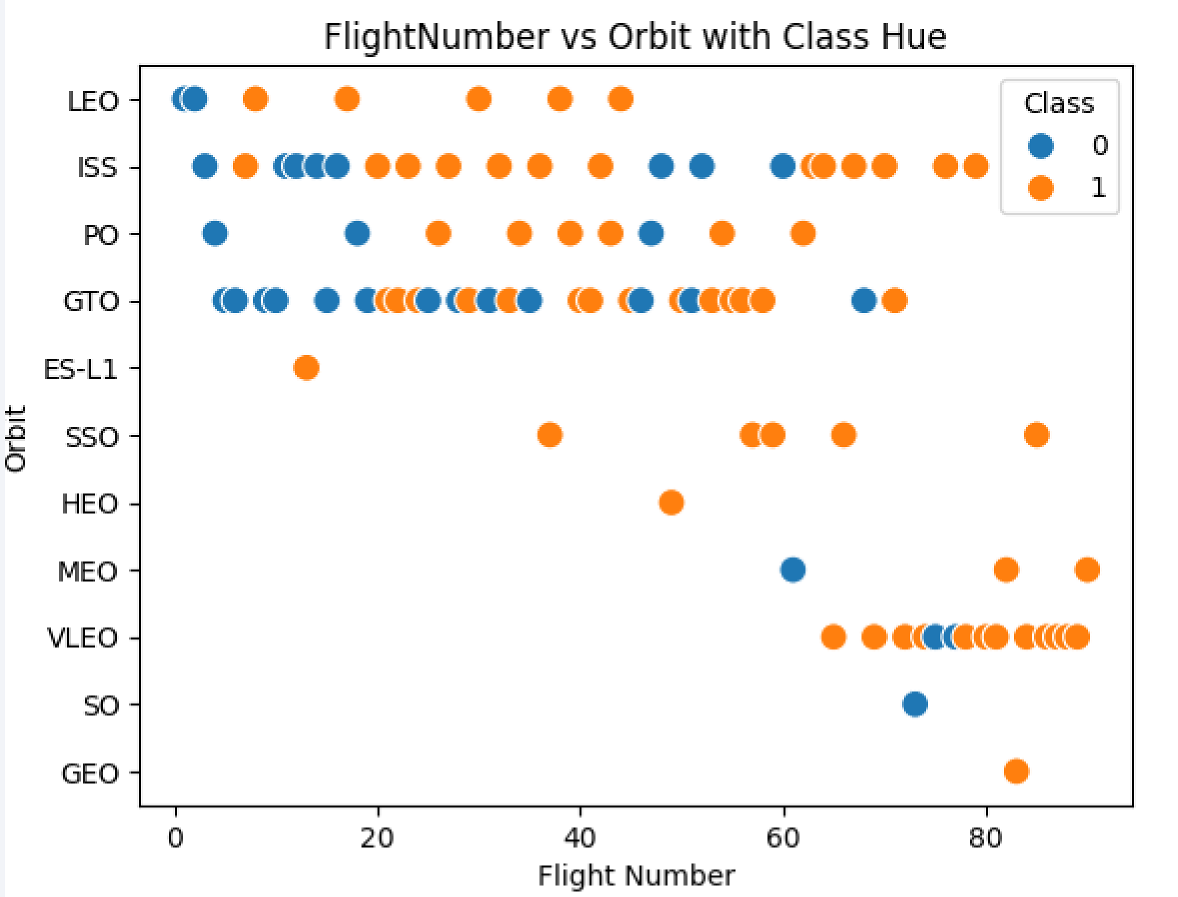
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO orbits had the most success rate.



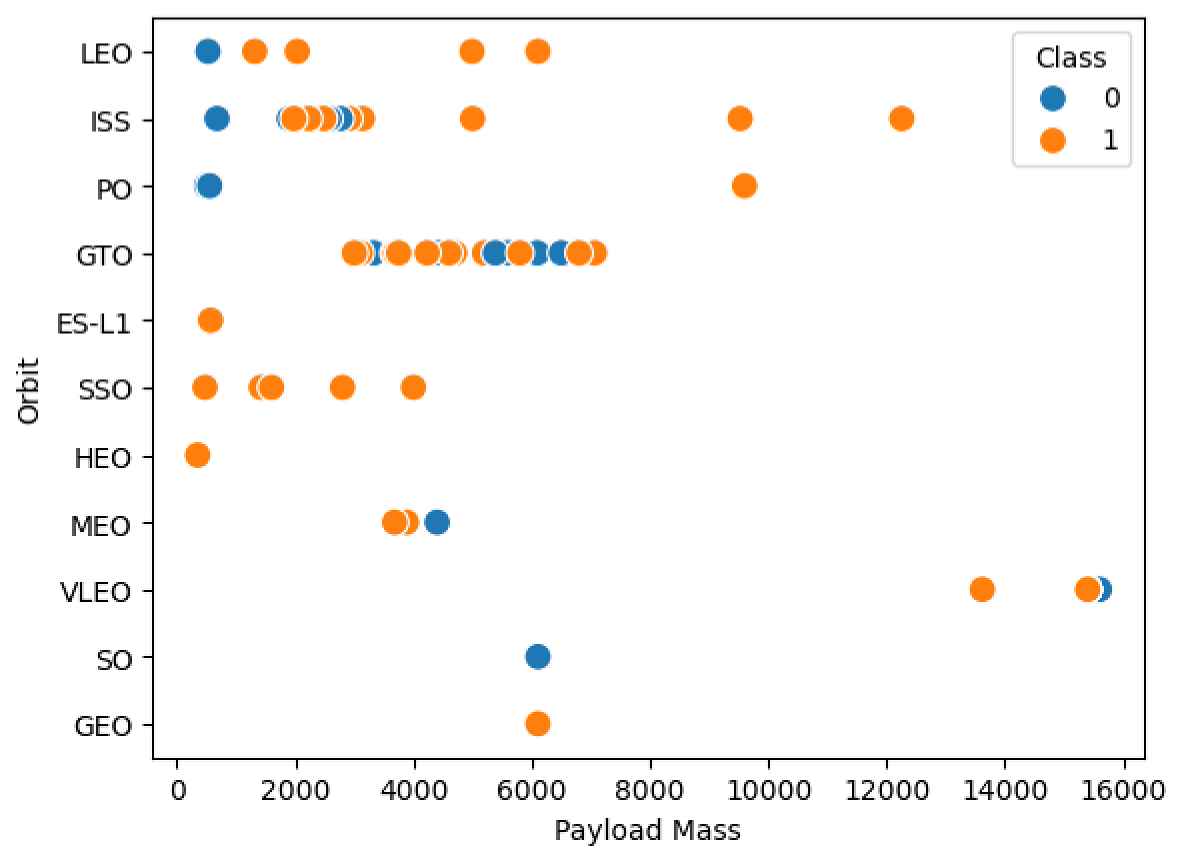
Flight Number vs. Orbit Type

- We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



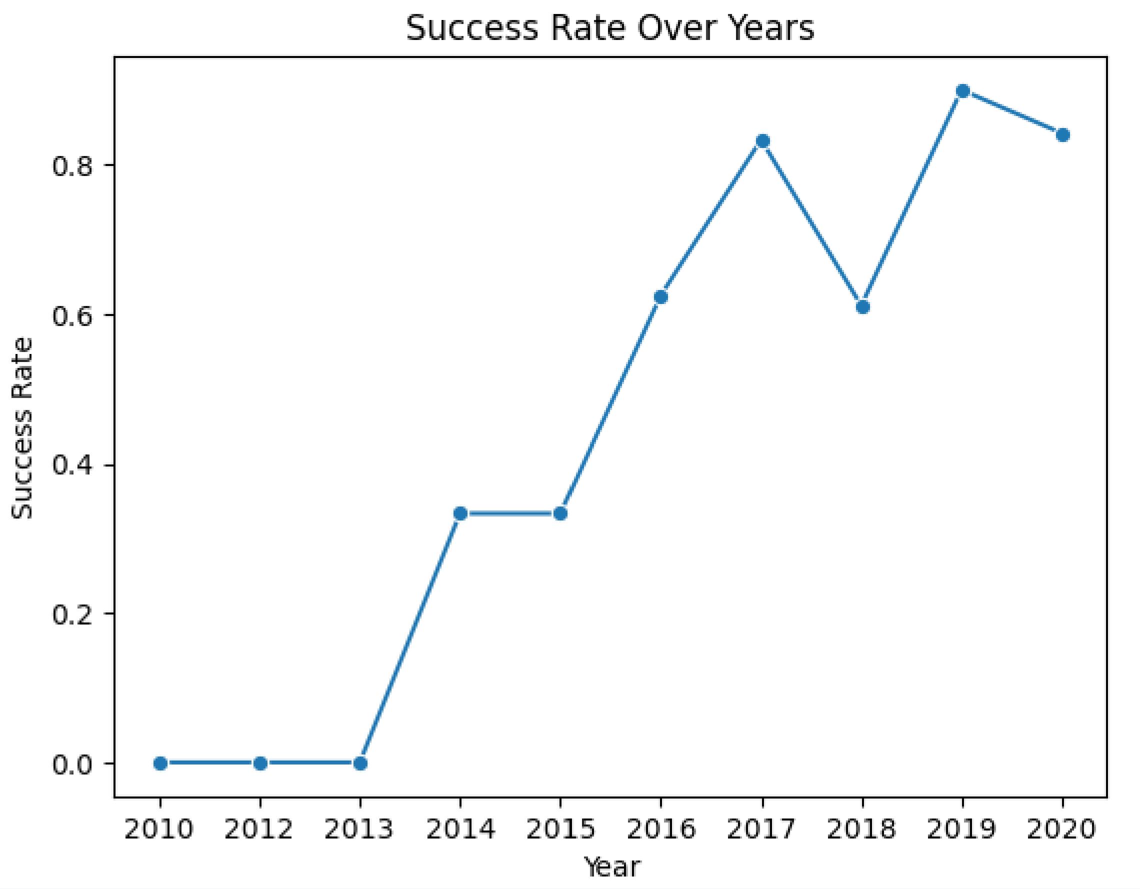
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2017. Eventhough it had dropped between the years 2017 and 2018, in 2019 it reached its peak.



All Launch Site Names

- Using DICTINCT key word unique launch sites were displayed.

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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

- The query above were used to display 5 records whose launch site names begin with ‘CCA’.

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer"="NASA (CRS)"
```

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

SUM("PAYLOAD_MASS__KG_")
45596

- We calculated the total payload carried by boosters from NASA as 45596 using the query below.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

average_payload_mass
2928.4

- Calculated the average payload mass carried by booster version F9 v1.1 and found as 2928.4.

First Successful Ground Landing Date

```
%%sql SELECT MIN("Date") AS first_successful_landing_date FROM SPACEXTABLE
WHERE "Landing_Pad" IS NOT NULL AND "Mission_Outcome" = 'Success';
```

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

first_successful_landing_date
2010-06-04

- To find the dates of the first successful landing outcome on ground pad «min» function were used.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Liste the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
%%sql SELECT Booster_Version
        FROM SPACEXTABLE
        WHERE Landing_Outcome = 'Success (drone ship)'
            AND PAYLOAD_MASS__KG_ > 4000
            AND PAYLOAD_MASS__KG_ < 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- While total number of successful missons is 100, total number of successful missons is 1.

```
%%sql SELECT
    CASE
        WHEN "Mission_Outcome" LIKE ('%Success%') THEN 'Success'
        WHEN "Mission_Outcome" LIKE('%Failure%') THEN 'Failure'
        ELSE 'Other'
    END AS Outcome,
    COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY Outcome
```

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

Outcome	Total
Failure	1
Success	100

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
%%sql SELECT "Booster_Version"
    FROM SPACEXTABLE
    WHERE "PAYLOAD_MASS__KG_" = (
        SELECT MAX("PAYLOAD_MASS__KG_")
        FROM SPACEXTABLE
    )
```

```
 * 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

2015 Launch Records

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%%sql SELECT Booster_Version, PAYLOAD_MASS__KG_
       FROM SpaceXTABLE
       WHERE PAYLOAD_MASS__KG_ = (
                                    SELECT MAX(PAYLOAD_MASS__KG_)
                                    FROM SpaceXTABLE
                                    )
       ORDER BY Booster_Version
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

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.

```
%%sql SELECT Booster_Version, Launch_Site, Landing_Outcome
        FROM SpaceXTABLE
        WHERE Landing_Outcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
```

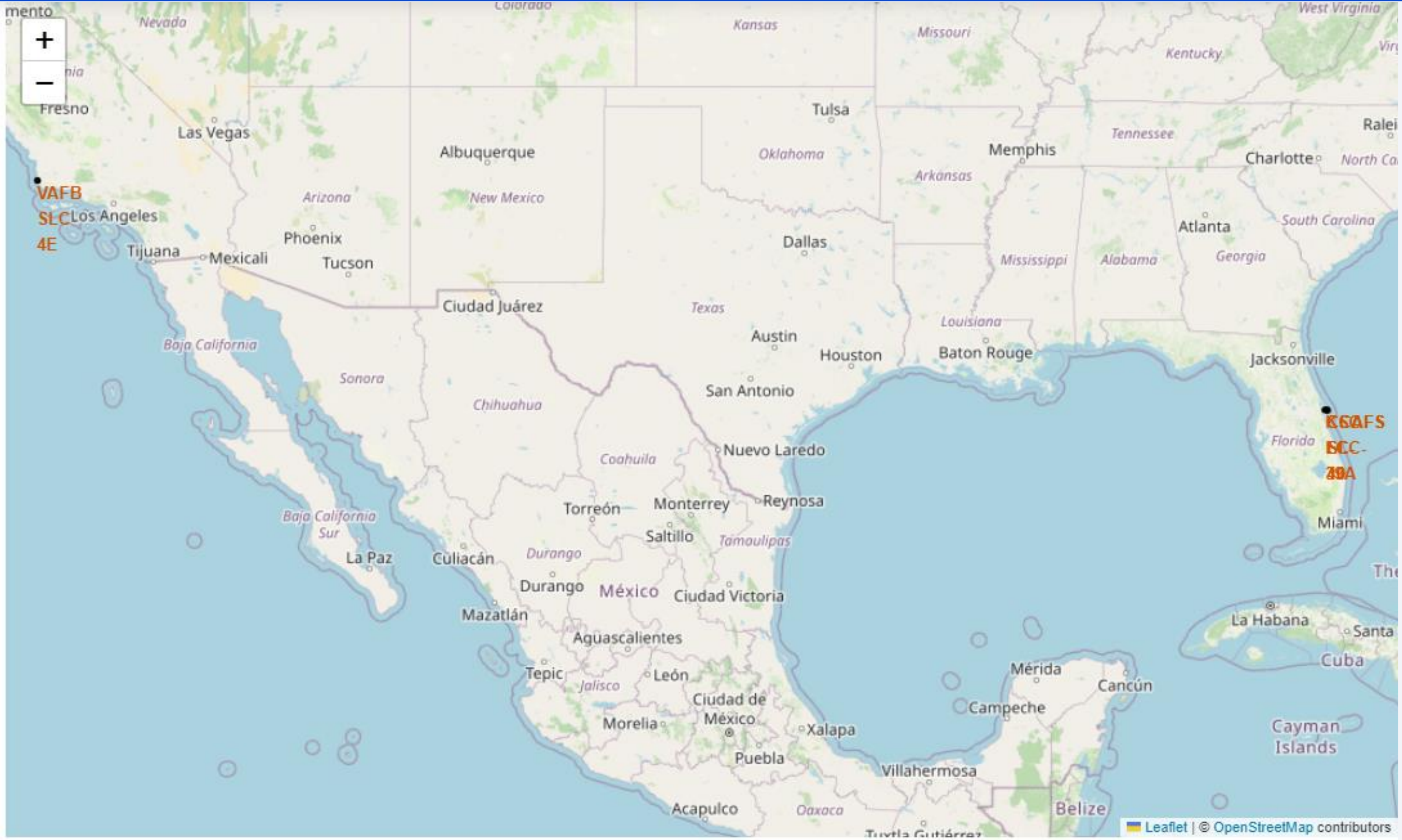
```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Launch_Site	Landing_Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

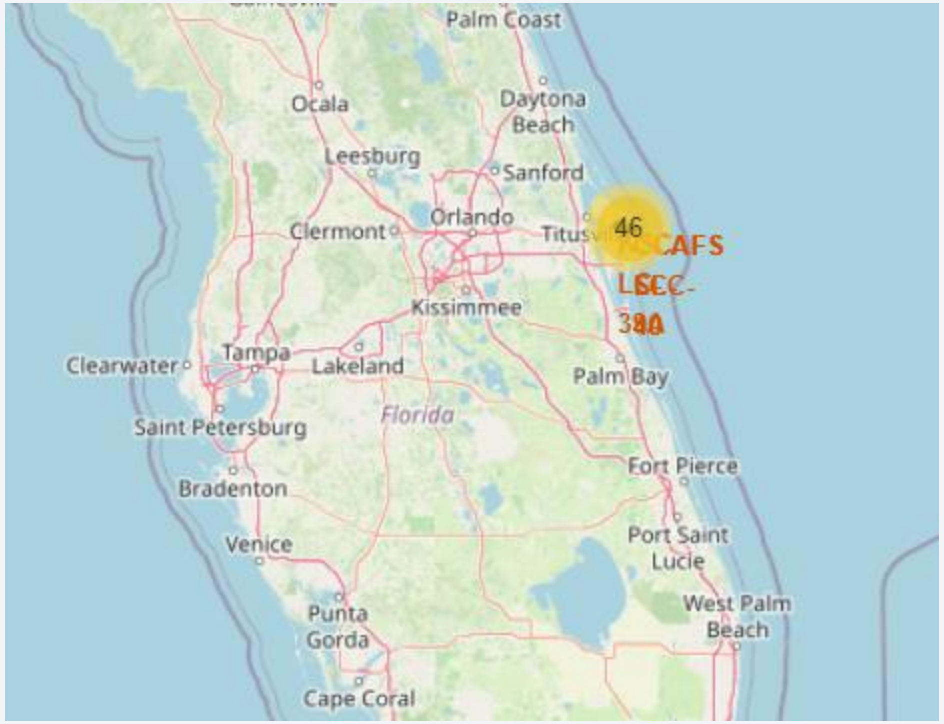
Section 3

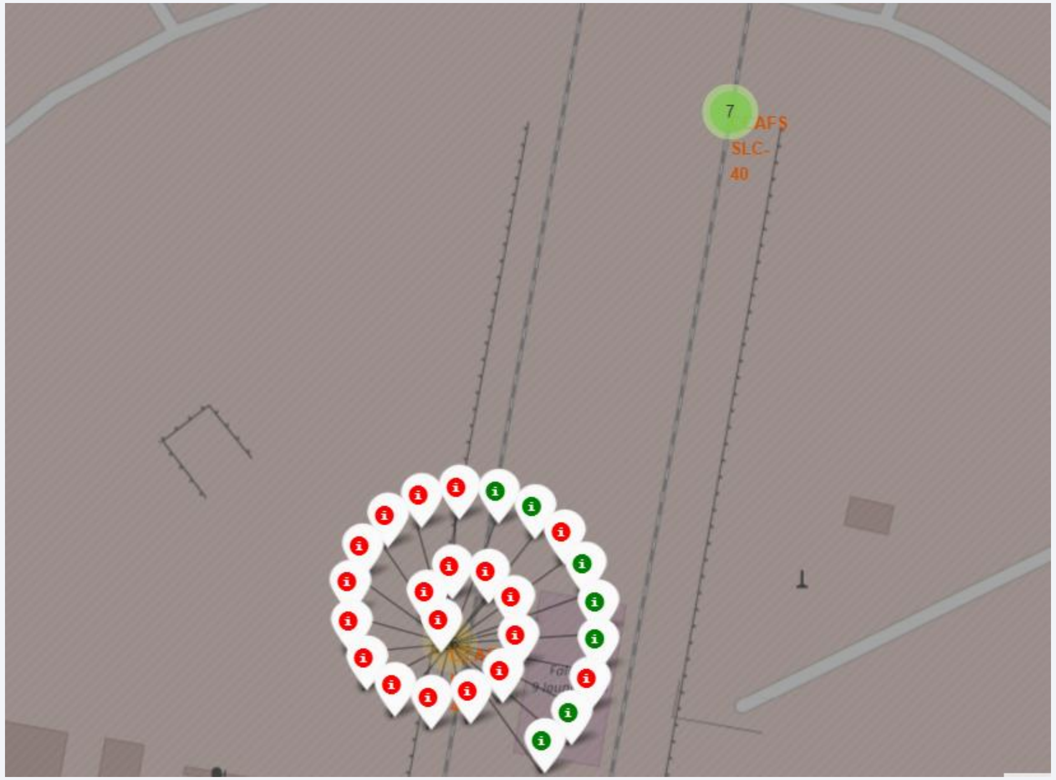
Launch Sites Proximities Analysis

All launch sites global map markers



Markers showing launch sites with color labels



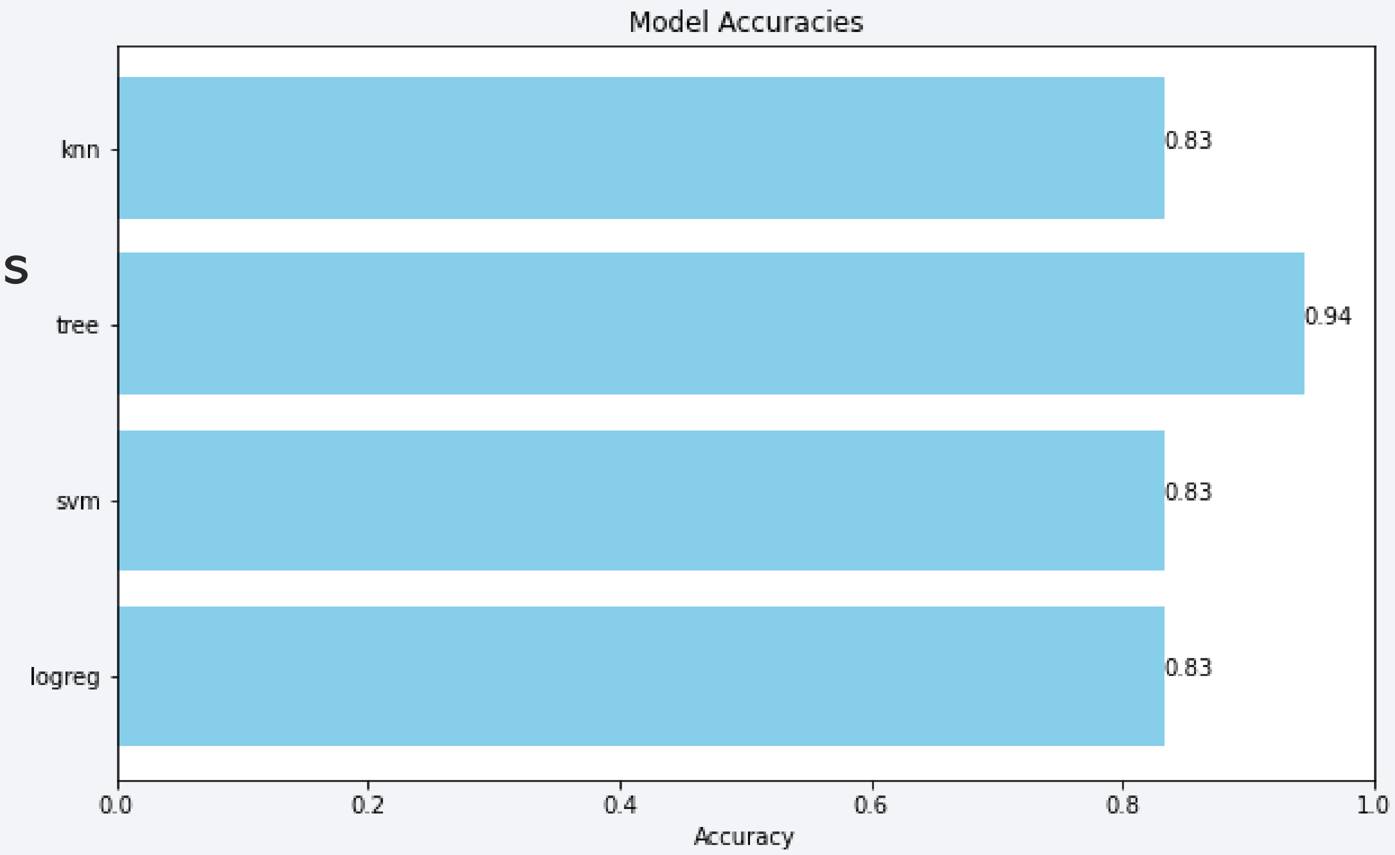


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions

- We can conclude that:
  - The larger the flight amount at a launch site, the greater the success rate at a launch site.
  - Launch success rate started to increase in 2013 till 2020.
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
  - KSC LC-39A had the most successful launches of any sites.
  - The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!