Applied Data Science Capstone Project

Shirish Senthil Kumar

25-12-2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Following methodologies were used to conduct the data analysis:

- Data Collection using web scrapping and Data Wrangling.
- Exploratory Data Analysis (EDA) with SQL, Data Visualization and interactive dashboard analytics.
- Machine Learning Predictions (Classification Analysis)

Brief Summary on the results

- EDA results to choose the best features of the data
- Interactive dashboard results
- Machine Learning Predictions based on standardized data.

Introduction

Objective: To conduct a comprehensive analysis and assess the success rate of first stage landing of a novel company SpaceY in comparison to SpaceX.

Desirable Outcomes:

- The effects of features such as Payload Mass, LaunchSite, Orbit, etc on the success of first stage landing
- Presentation of success rate results over the years.
- A highly accurate prediction model for the success rate based on Machine Learning

Methodology

Methodology

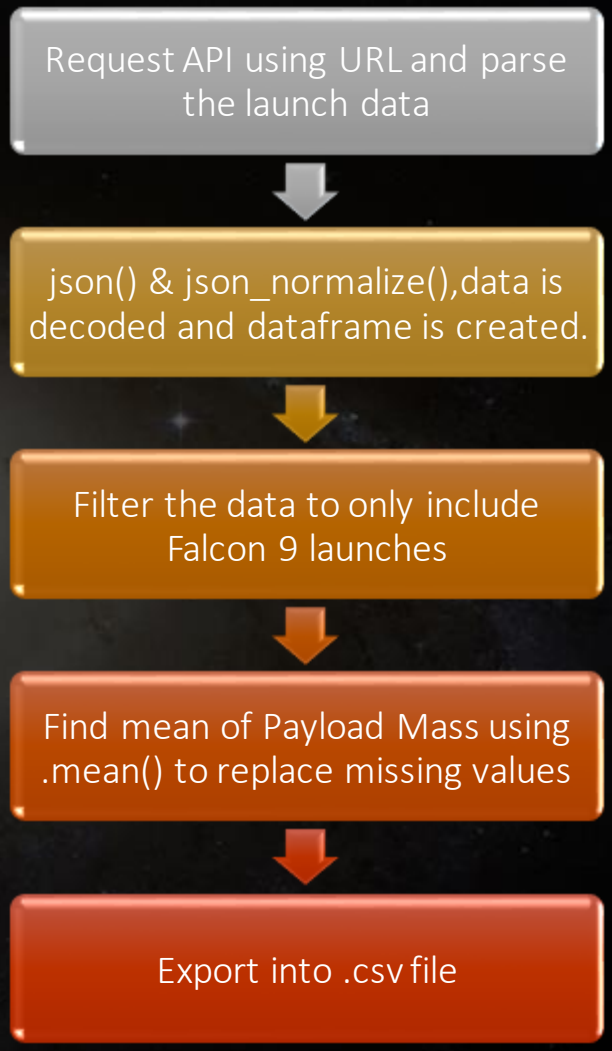
Summary

- Data Collection: Data was primarily collected from sources below
  - Sources:
    - SpaceX API (https://api.spacexdata.com/v4/launches/past)
    - Web scrapping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Data Wrangling
  - Data was labelled based on the outcome of the launch, making it easier to work on in EDA.

Data Collection - API

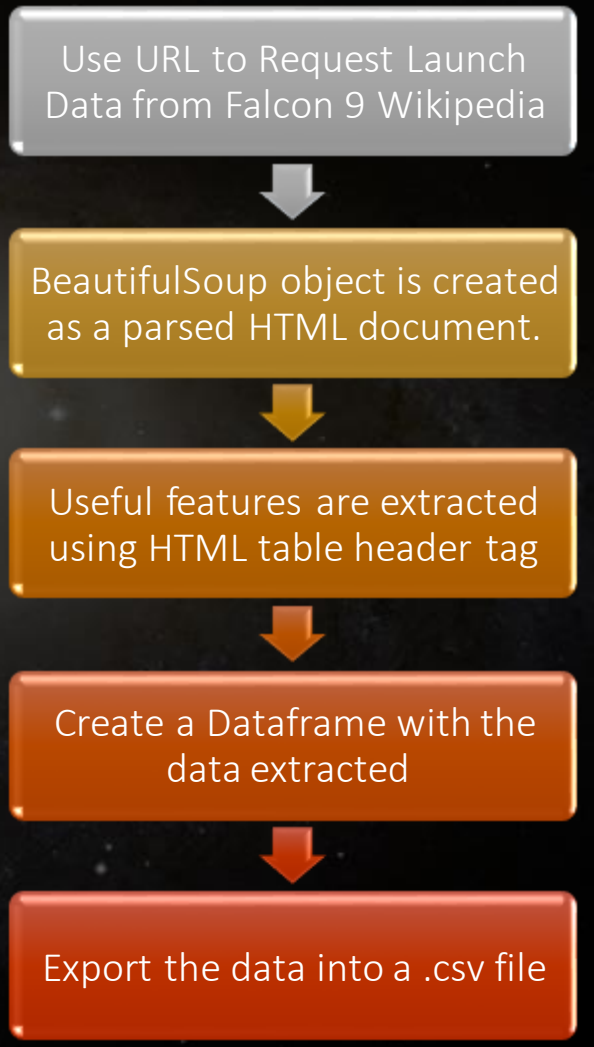
- Data Collection involved using SpaceX REST API data and web scrapping data from SpaceX Wikipedia webpage.
- The SpaceX REST API went through the process flowchart on the right to extract the raw data.

- Source:https://api.spacexdata.com/v4/launches/past
- Code:https://github.com/Shirish026/CapstoneProject/tree/main/Capstone%20Project



Data Collection – Web Scrapping

- Launch Data about SpaceX is also obtained from Wikipedia.
- The webpage URL is used to obtain the data. The process flowchart highlights the course of action.



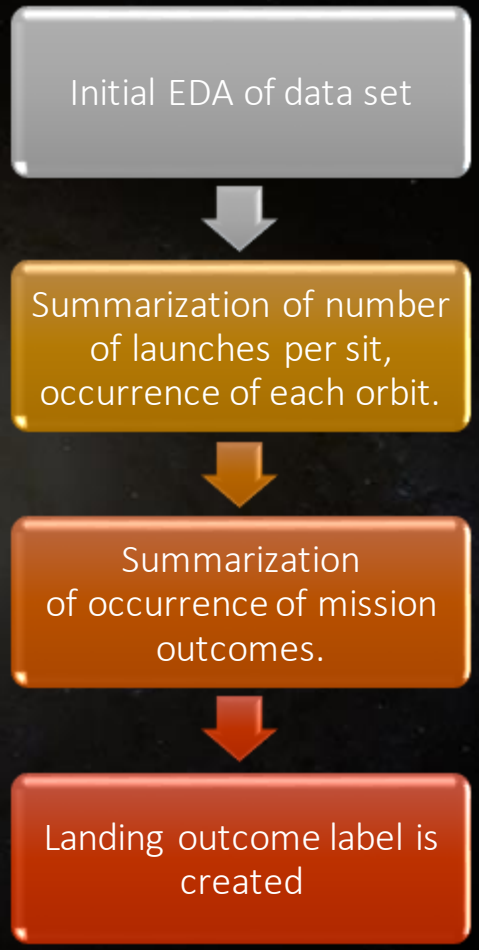
- Source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Code:https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/Data%20Web%20Scrapping.ipynb

Data Wrangling

Data Wrangling refers to the removal of errors and combining complex data sets to make them to perform exploratory data analysis (EDA).The dataset comprises of data pertaining to the success of landing as well the failures. The process flowchart represents the line of action in this methodology.

To filter out the outcomes of true successive landing from unsuccessful landings, a training label (0,1) is created along this column. 0 refers to unsuccessful landing and 1 refers to successful landings.

Code:https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/Data%20Wrangling.ipynb



Exploratory Data Analysis (EDA) with SQL

The following SQL queries were performed to extract valuable data:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List of names of boosters that succeeded and have payload mass between 4000-6000.
7. List the total number of successful and failure mission outcomes
8. List the names of booster versions that carried a max payload mass using Subquery.
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.
10. Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Code: https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/SQL%20EDA.ipynb

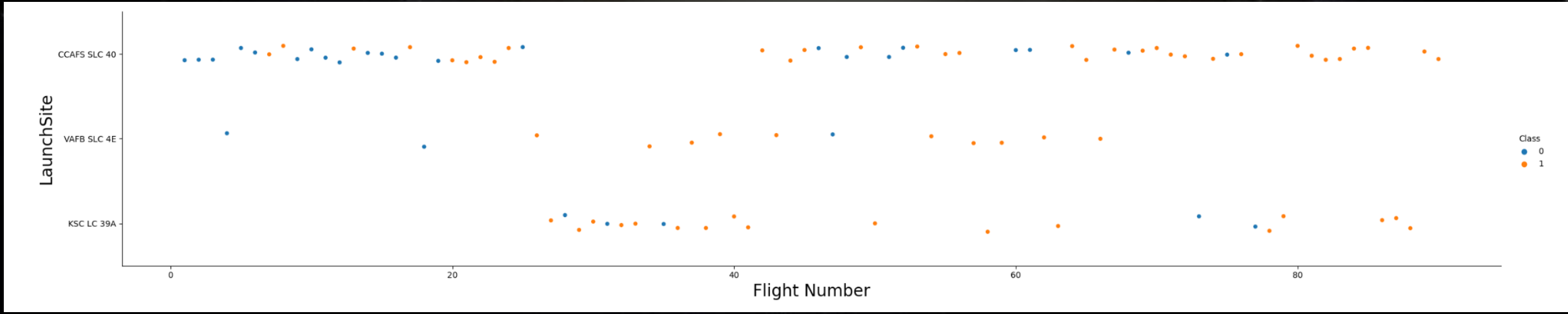
EDA with Data Visualization

Scatterplots, Bar plots and line plots were plotted to verify and visualize the correlation of features. These plots are:

1.Payload Mass vs Flight Number, 2. Launch Site vs Flight Number, 3. Launch Site vs Payload Mass, 4. Orbit and Flight Number, 5. Payload vs Orbit.

Code:https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/EDA%20Vizualization.ipynb

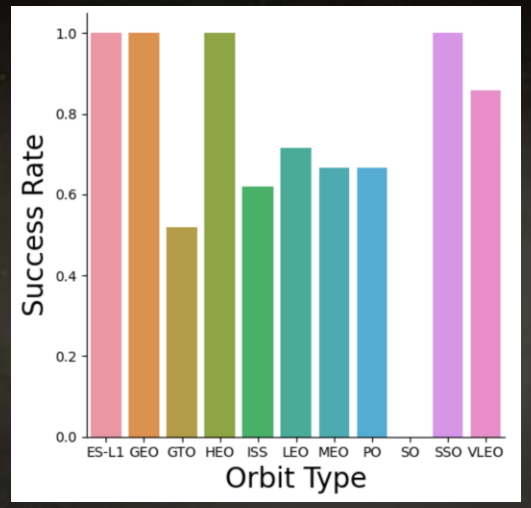
1. Scatter Plot: Launch Site vs Flight Number



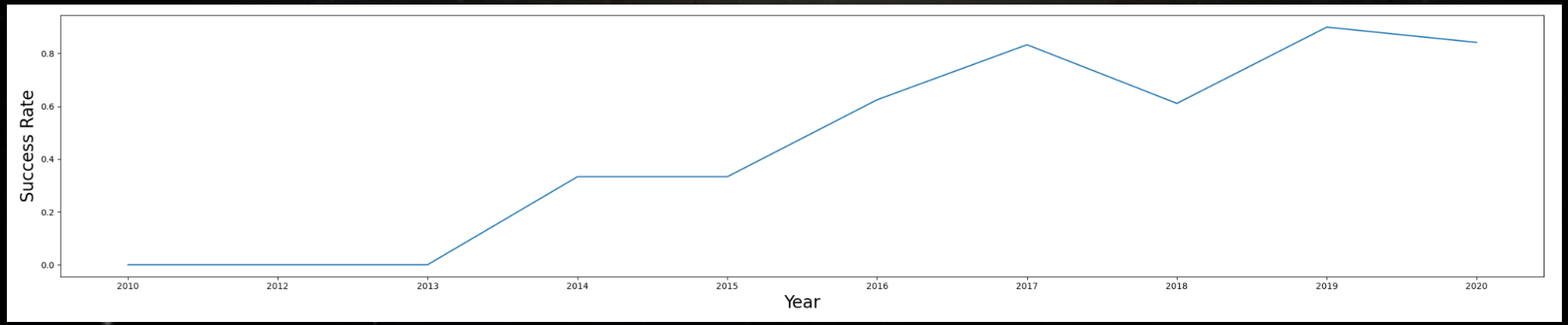
EDA with Data Visualization

Code:https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/EDA%20Vizualization.ipynb

2. Bar Chart Plot: Success Rate vs Orbit Type



3. Line Plot: Success Rate vs Year



Interactive Mapping with Folium

Folium enables interactive analysis enables the use of multiple leaflet maps and used in dashboarding.

The following functions of folium were used along with Folium Library:

1. Markers: Used to mark coordinates of the data in a real-world map (e.g. launch sites)
2. Circles: Provides a circular highlight to the markers' specific location (SpaceX launch site)
3. Marker Clusters provide the option to mark as a group of occurrences in each coordinate. (e.g., multiple launches at a launch site).
4. Plotting line between two points provides distance between those coordinates.

Code: https://github.com/Shirish026/CapstoneProject/tree/main/Capstone%20Project

Dashboarding with Plotly Dash

The following 2 plots were used to encapsulate and visualize data using an interactive dashboard:

1. Pie chart of Percentages of Launches by sites based on the choice of the dropdown menu.
2. A scatter plot to show relation between payload and Launch success. The payload range (Kg) is inputted using an interactive slider as shown in the source code.

The key idea is to understand the relation between payload and launch success to pick the best launch site to have a first stage success.

Code: https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/Dashboard.py

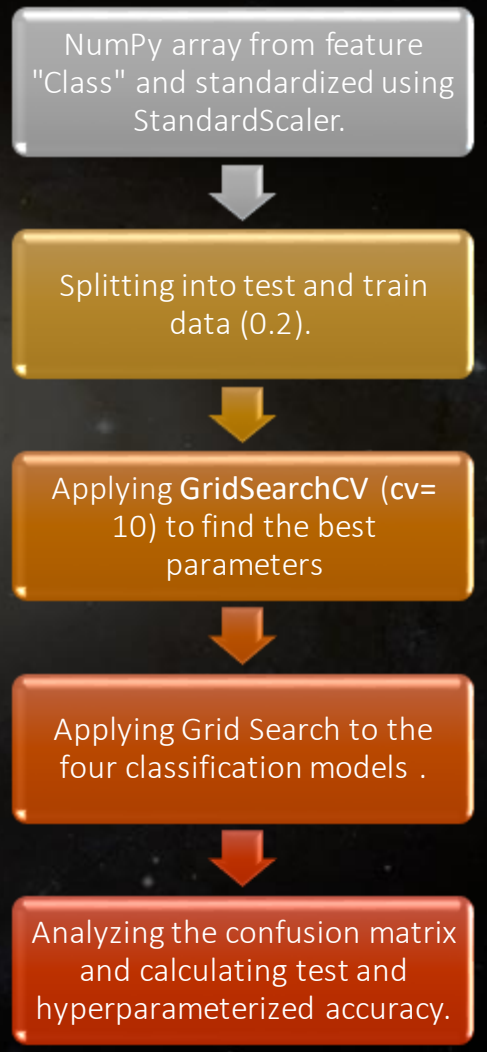
Predictive Analysis (Classification)

The four classification models used in this predictive analysis are stated:

1. Logisitic Regression
2. Support Vector Machines
3. Decision Tree
4. K nearest Neighbours

The process flowchart represents the steps of action taken in this analysis.

Code:https://github.com/Shirish026/CapstoneProject/blob/main/Capstone%20Project/ML%20Predictive%20Analysis.ipynb



Results of the Methodologies

- Consists of result snippets and an explanation:
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analytics (SQL & Viz)
  - Interactive Analytics (Folium & Plotly Dash)
  - Machine Learning Predictive Analysis (Classification)

Data Collection API

Figure below shows a few rows of the data collected from the SpaceX API that only contains information pertaining to Falcon 9 Booster Version.

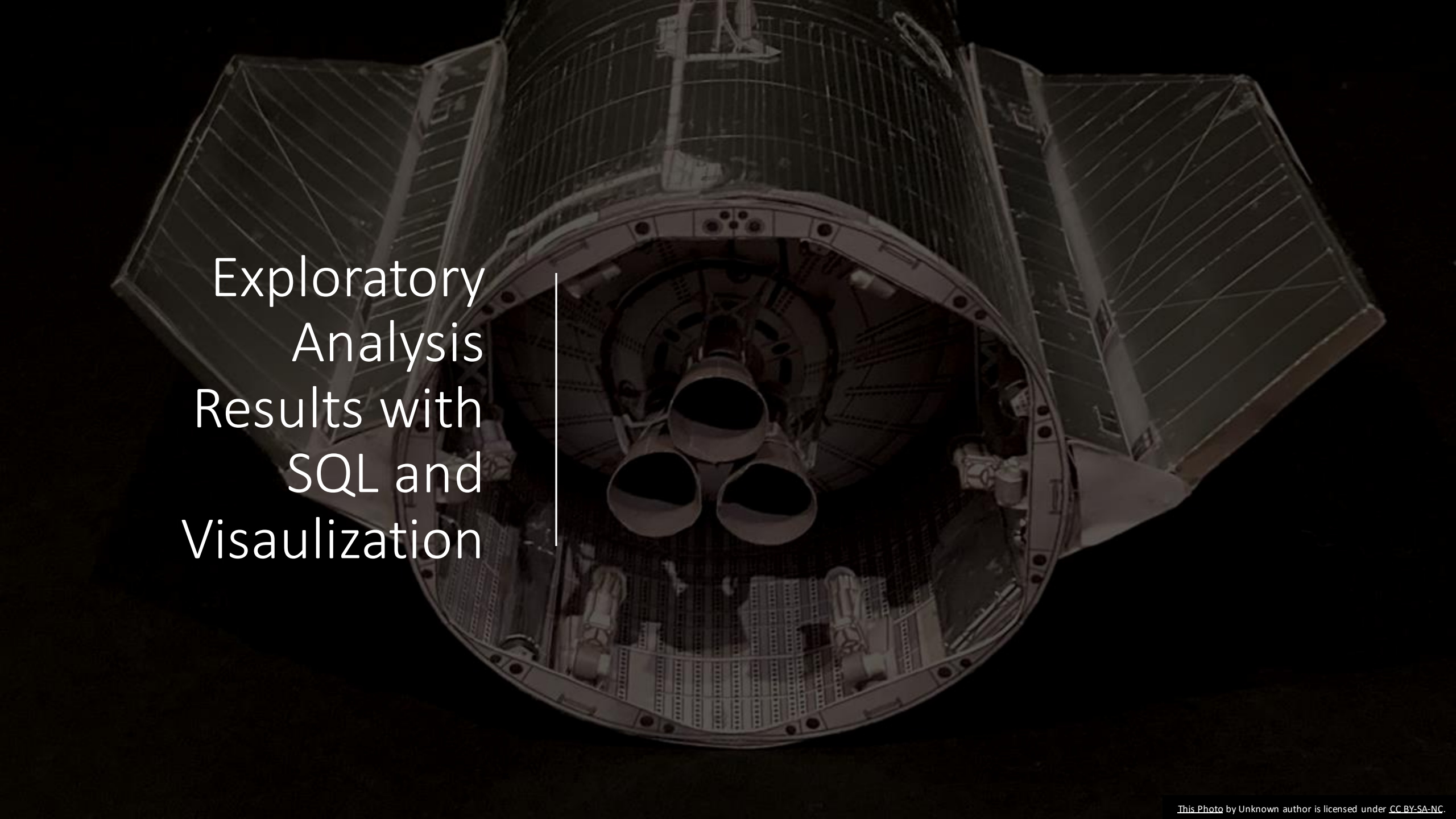
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False

Data Collection Web Scrapping

Figure below shows the data collected from Wikipedia in reference to SpaceX and entered as a data frame using Pandas.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
237	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
238	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56

Exploratory Analysis Results with SQL and Visaulization



EDA with SQL

1. Shows all the distinct launch site names obtained using SQL query:

```
In [9]: %sql select distinct launch_site from SPACEX;
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

2. Shows all the launch site names that being with 'CCA':

```
%sql select * from SPACEX where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

EDA with SQL

3. Displays the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_payload_mass
45596

4. Displays the average payload mass carried by booster version F9 v1.1

In [12]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

average_payload_mass
2534

EDA with SQL

5. Shows the date when the first successful landing outcome in ground pad was achieved.

```
In [13]: %sql select min(date) as first_successful_landing from SPACEX where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[13]:
```

first_successful_landing
2015-12-22

6. Shows the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [14]: %sql select booster_version from SPACEX where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[14]:
```

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

EDA with SQL

7. Displays the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEX group by mission_outcome;
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

8. Shows the names of the booster versions which have carried the maximum payload mass using a subquery

```
%sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX);
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

EDA with SQL

9. Shows the records which display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEX
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

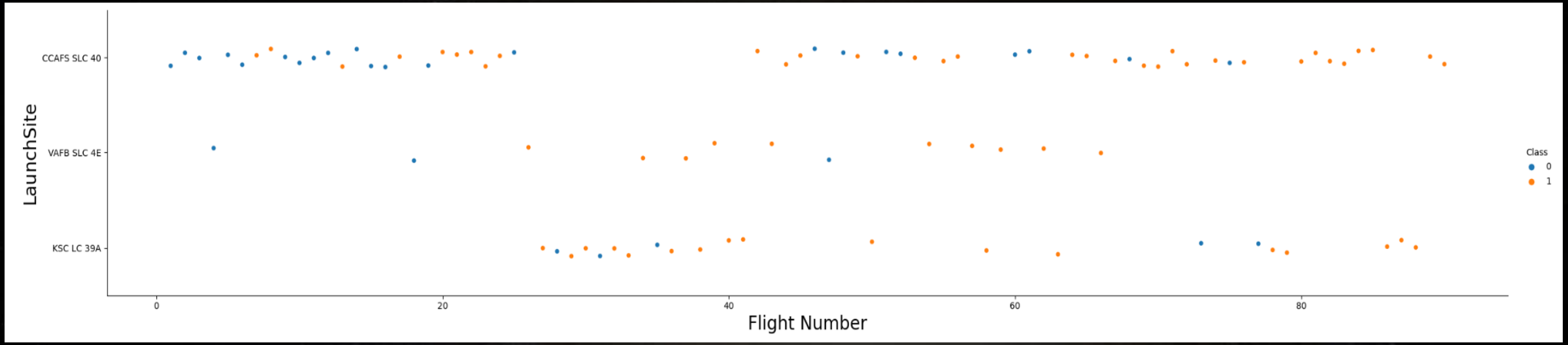
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEX
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wfd98799:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io
Done.
```

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

EDA with Visualization

1. Launch Site vs Flight Number



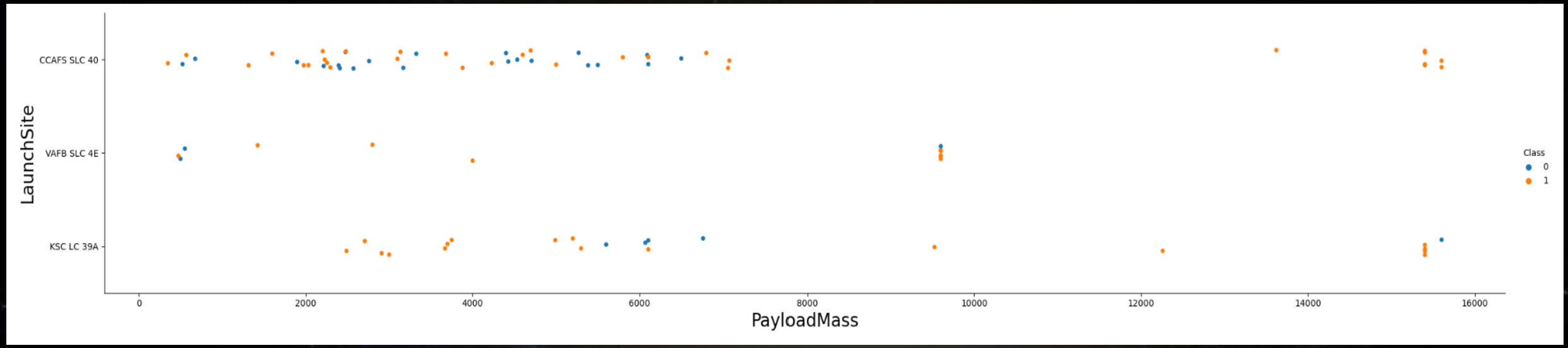
Analysis:

- The figure shows the landing outcomes (0 = Unsuccessful, 1 = Successful).

- Launch Site "CCAFS SLC 40" has had the most recent successful landings followed by "KSC LC 39A" and finally "VAFB SLC 4E".

EDA with Visualization

2. Launch Site vs Payload Mass (Kg)

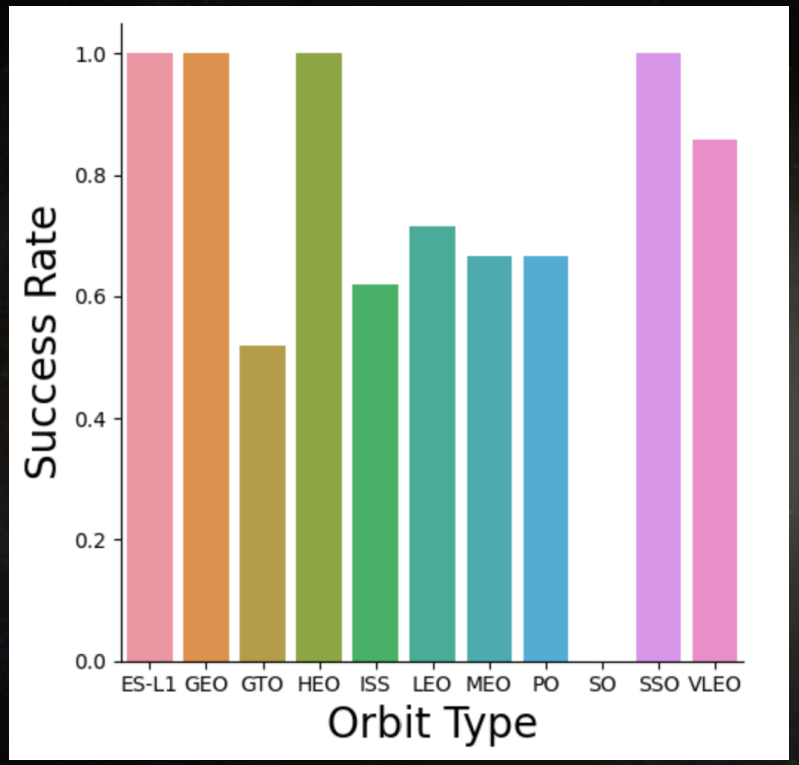


Analysis:

- The figure shows the landing outcomes (0 = Unsuccessful, 1 = Successful).
- Launch site "KSC LC 39A" has a favourable success rate as compared to the other launch sites.
- Lower Payload mass has primarily been launched from "CCAFS SLC 40"
- Lower Payload Mass has lesser success rate as compared to mid-high payload mass

EDA with Visualization

3. Success Rate vs Orbit Type

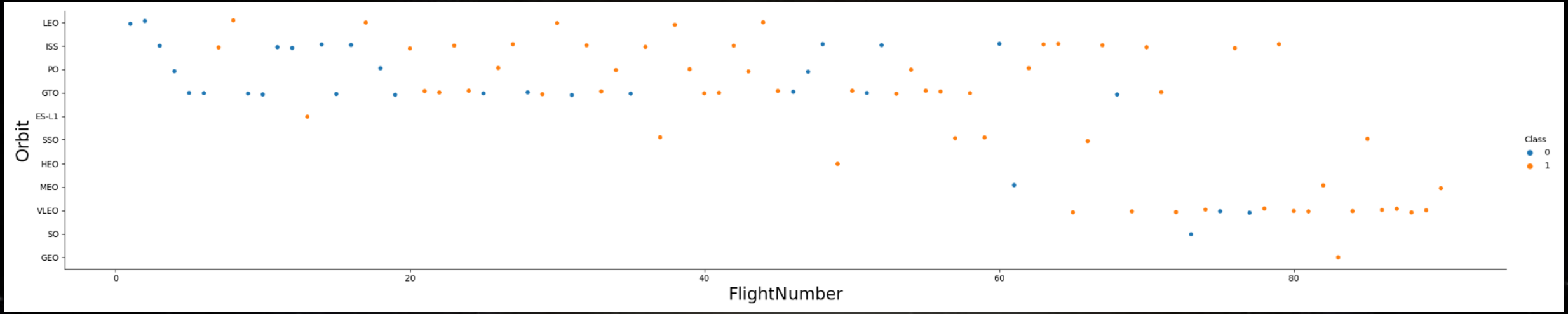


Analysis:

- Orbit Type "ES-L1", "GEO", "HEO" , "SSO" all have a success rate of 100% of landing the first stage rocket.
- Followed by orbit "VLEO" with 85%.
- Orbit Type "SO" has 0% success rate

EDA with Visualization

4. Orbit vs Flight Number

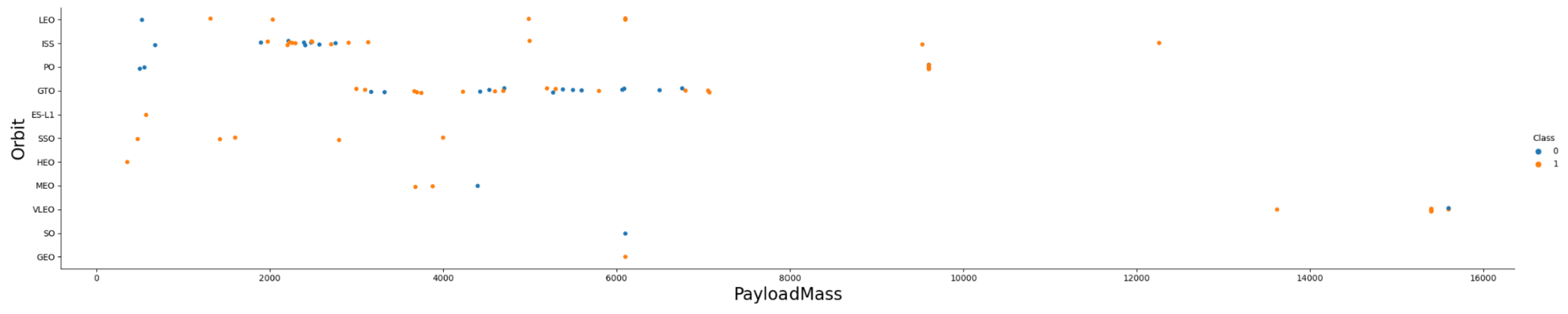


Analysis:

- The figure shows the landing outcomes (0 = Unsuccessful, 1 = Successful) and relation of the orbit for different flight numbers.
- Recent launches have been targeted at orbits "VLEO" & "MEO".
- The success rate has improved with the recent launches.

EDA with Visualization

5. Orbit vs Payload Mass

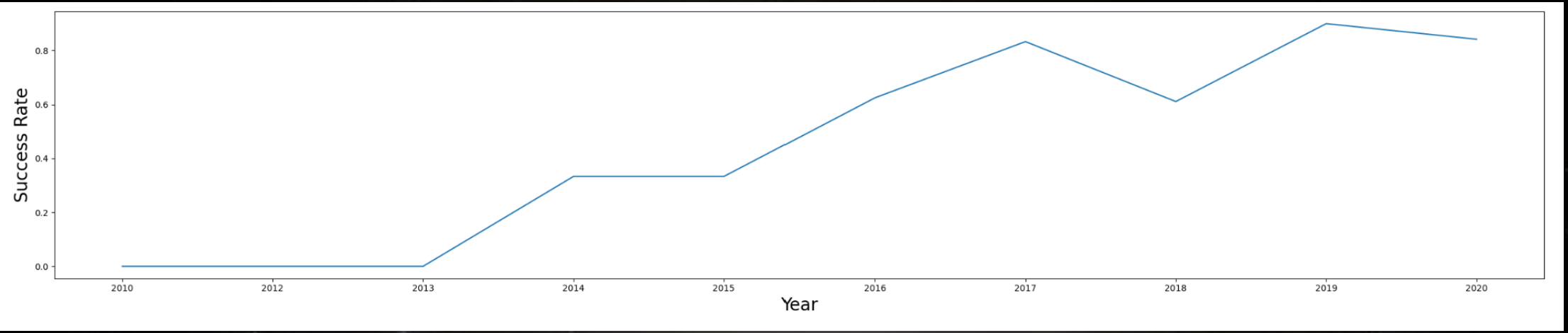


Analysis:

- The figure shows the landing outcomes (0 = Unsuccessful, 1 = Successful) and relation of the orbit for different flight numbers.
- For Payload Mass < 8000, the primary orbits used is "LEO", "ISS", "GTO" but fof higher payloads, the orbit used is "VLEO".
- Higher Payloads have had more success first stage landing than lower payloads.

EDA with Visualization

6. Success Rate vs Year

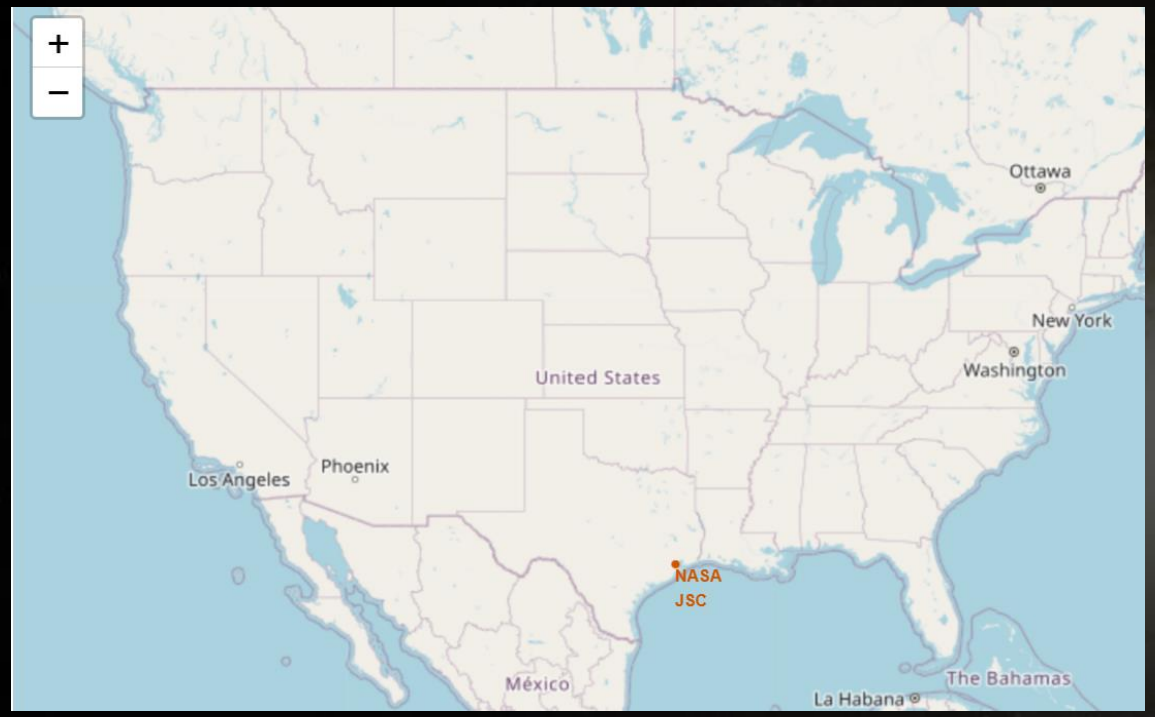


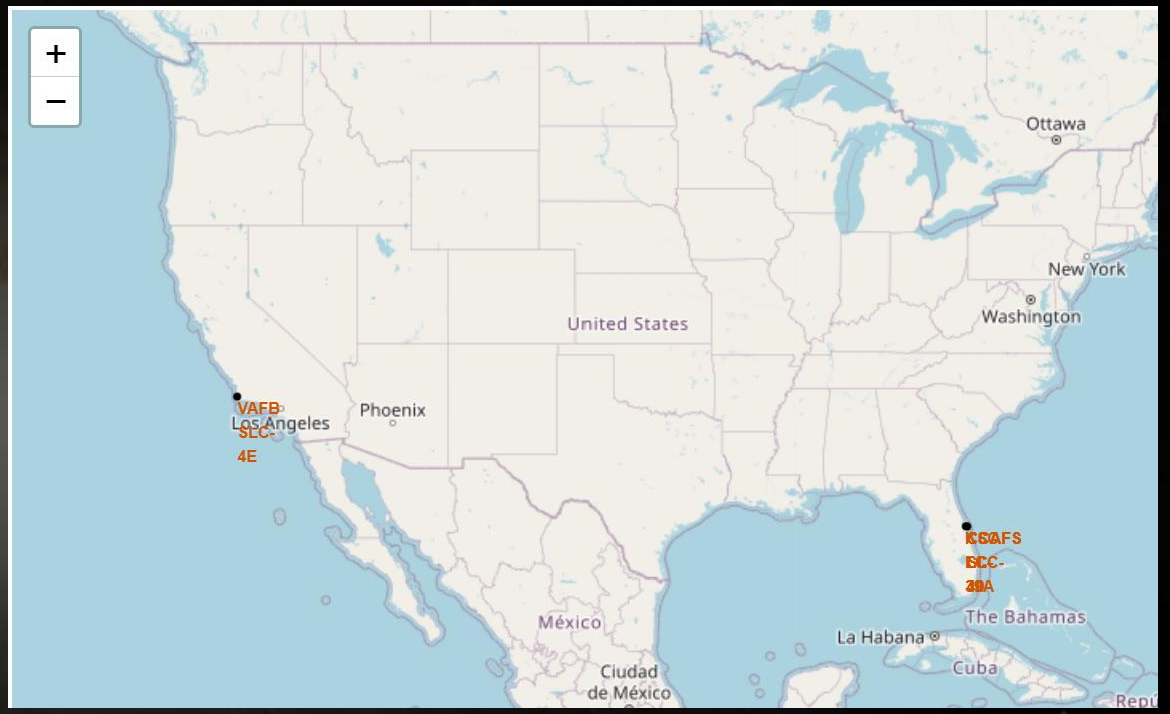
Analysis:

- It is evident that the success rate has improved drastically over the years.
- 2019 has the highest success rate of 90%.
- 2013-2014 has the highest increase in success rate from 0% to 40%.
- This states that there is still room for improvement to reach the effective 100% success rate.

Interactive Folium Map Analysis

Launch Sites





- The launch sites are primarily located at the coasts, to avoid any rocket failures landing on land. These launch sites are primarily located at the east and west coast.
- These launch markers are made with folium which allows for zooming & selection of the sites.

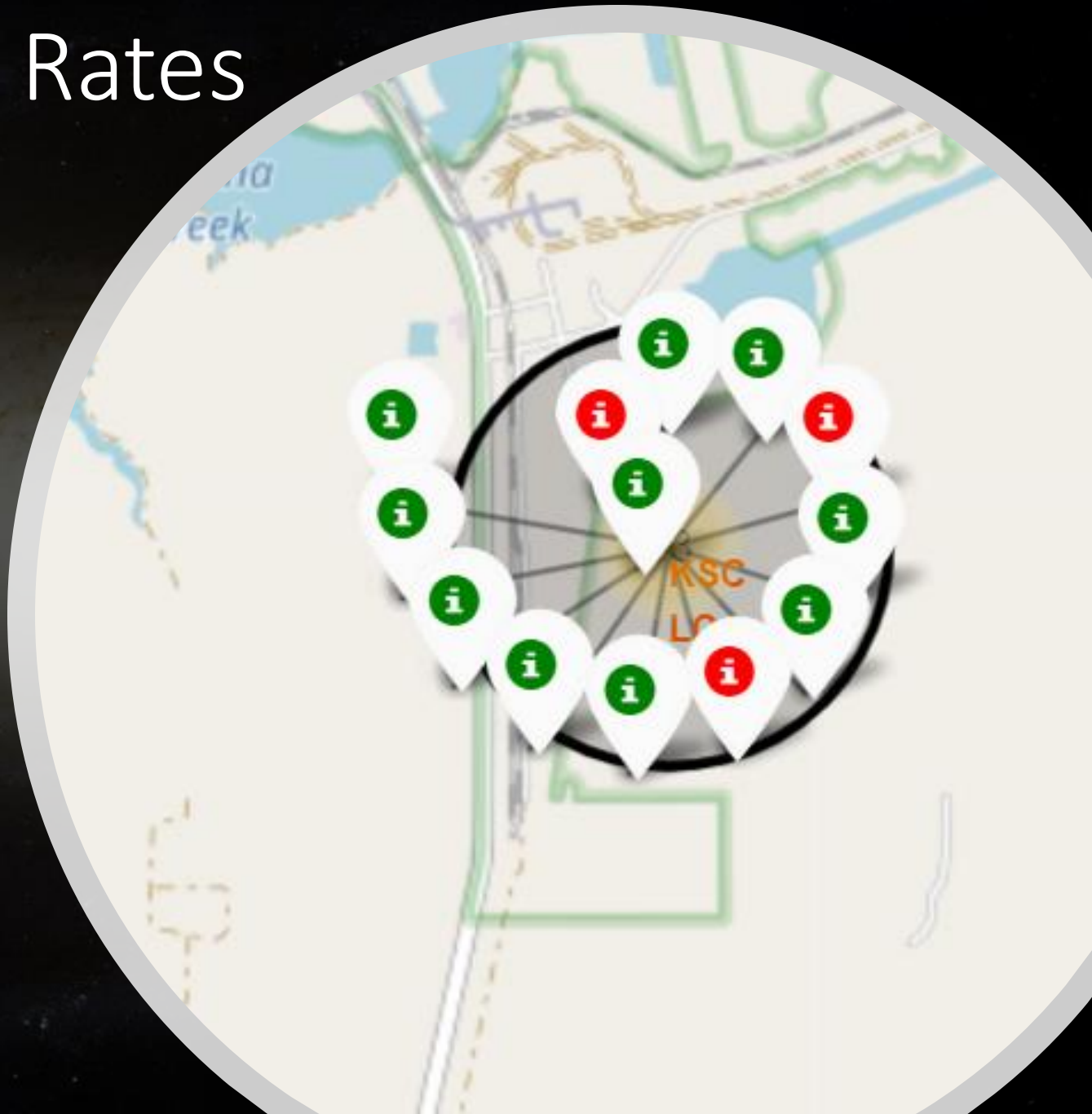
Colour Coded Success Rates

The colour coded launch sites represents the successful and unsuccessful launches. The figure on the right is made in Folium, for the launch site "KSC LC-39A".

GREEN MARKERS – Successful Launches

RED MARKERS – Unsuccessful Launches

This is further plotted for the other launch sites however, "KSC LC-39A" has the highest success rate.

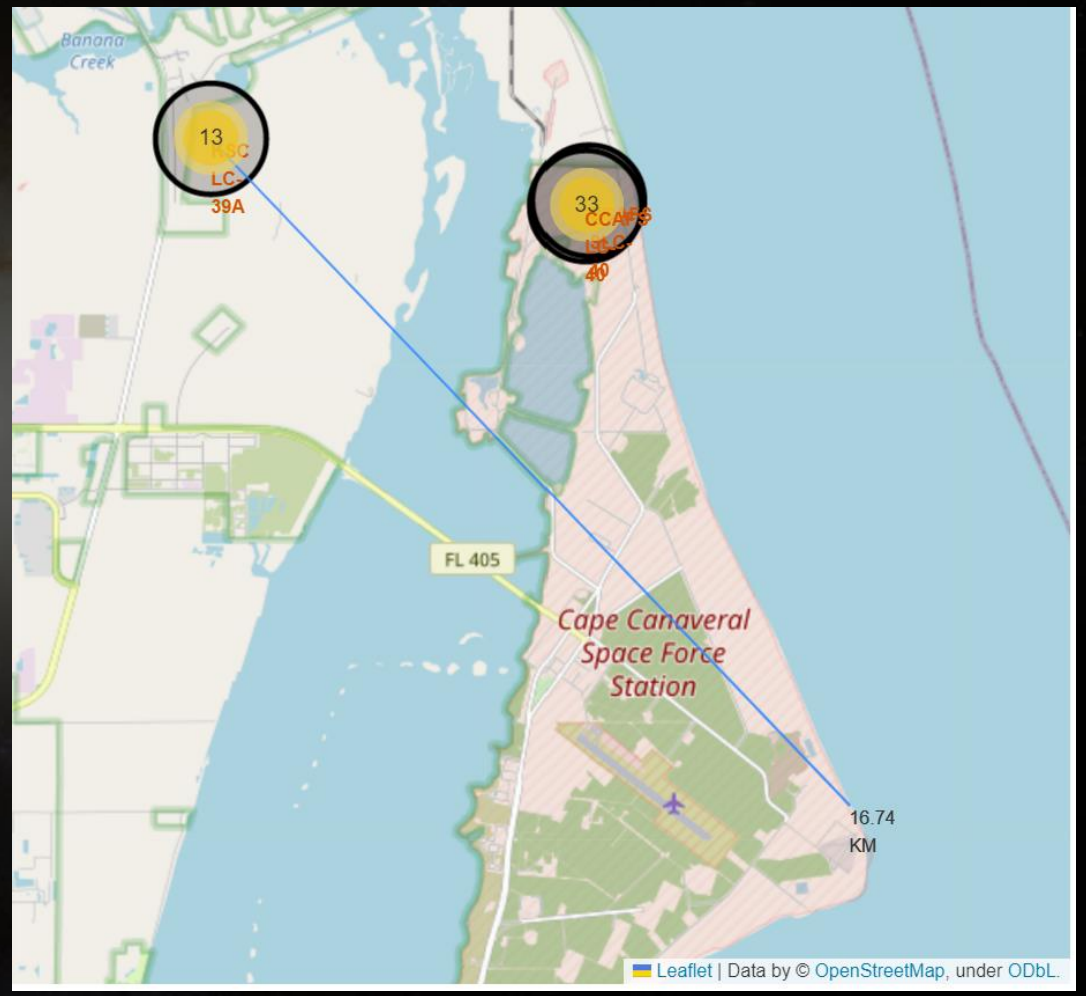


Distance between Launch site and Coastline

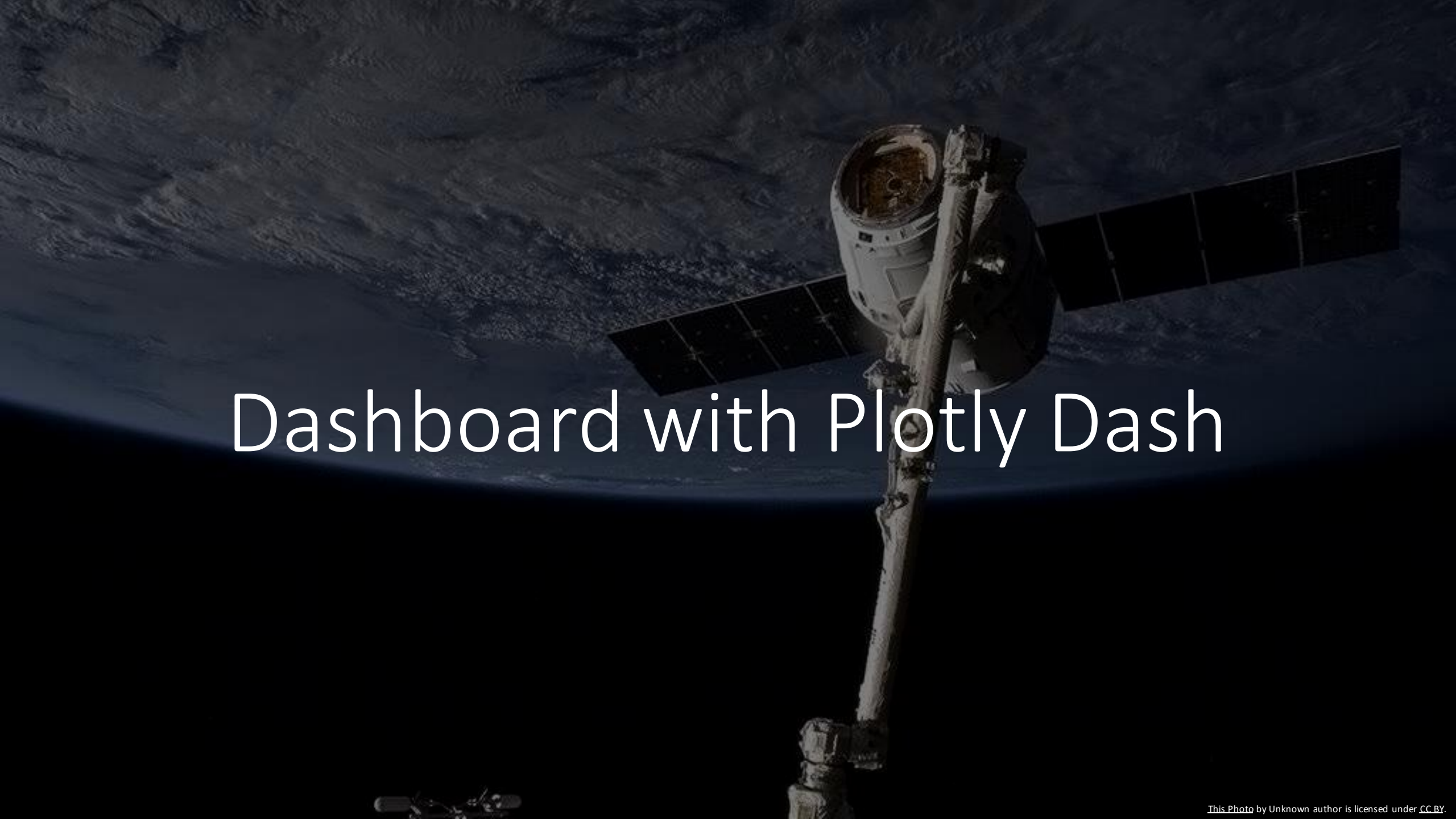
The figure shows the distance between the edge of Cape Canaveral space force station (treated as coastline) to the launch site "KSC LC-39A".

The coastline is 16.74Km away from the launchsite, represented by the line.

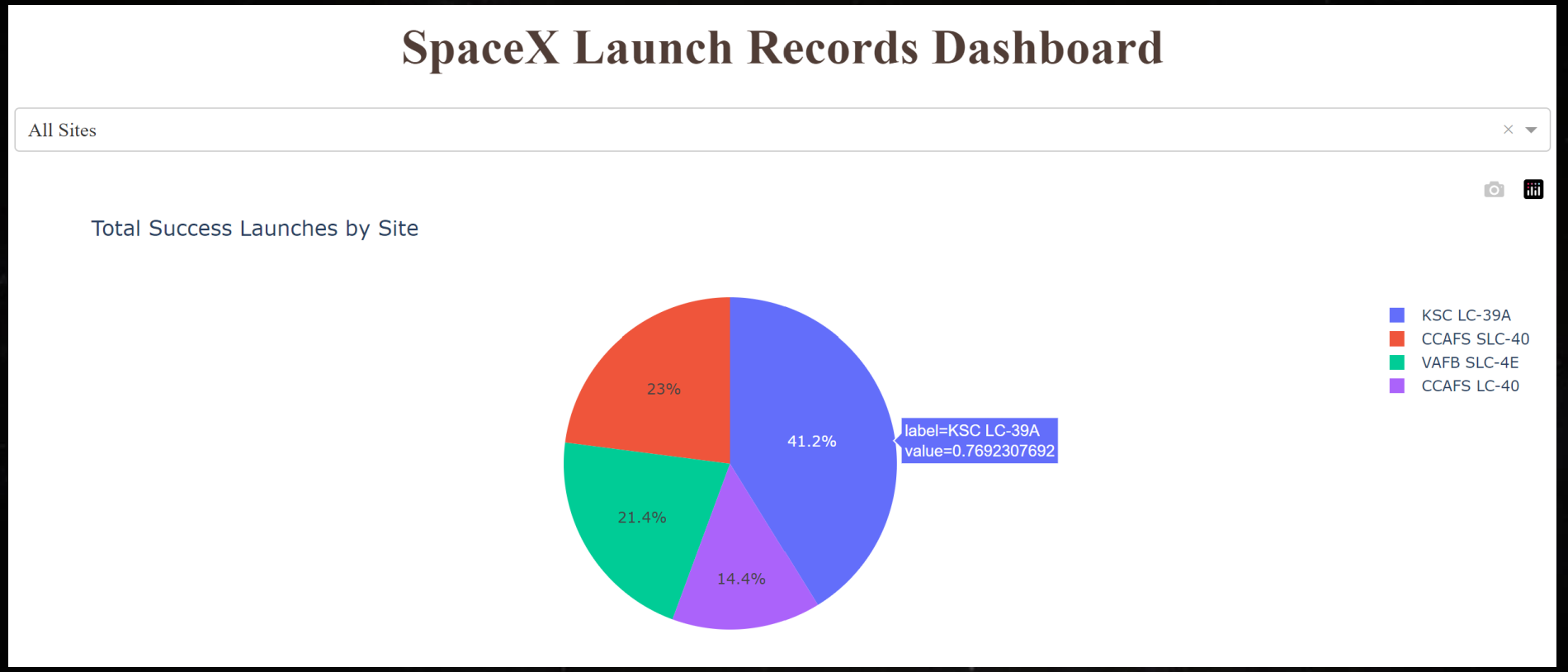
It is coded in Folium using the coordinates and its respective distance.



Dashboard with Plotly Dash

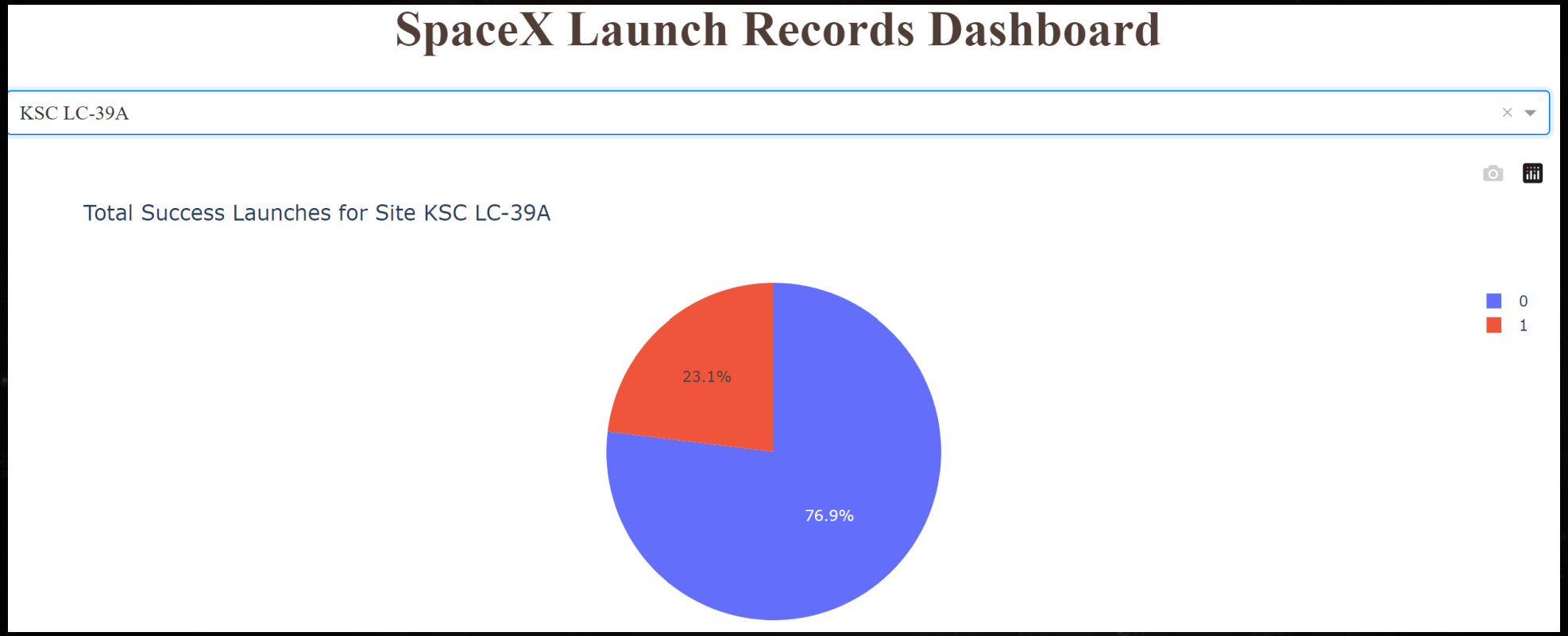


Pie chart indicating launch success of all sites



- From the pie chart, it is evident that KSC LC-39A provides highest success rate of 41.2%.
- For this presentation, launch site "KSC LA-39A" is investigated.

Pie chart indicating launch success of all sites



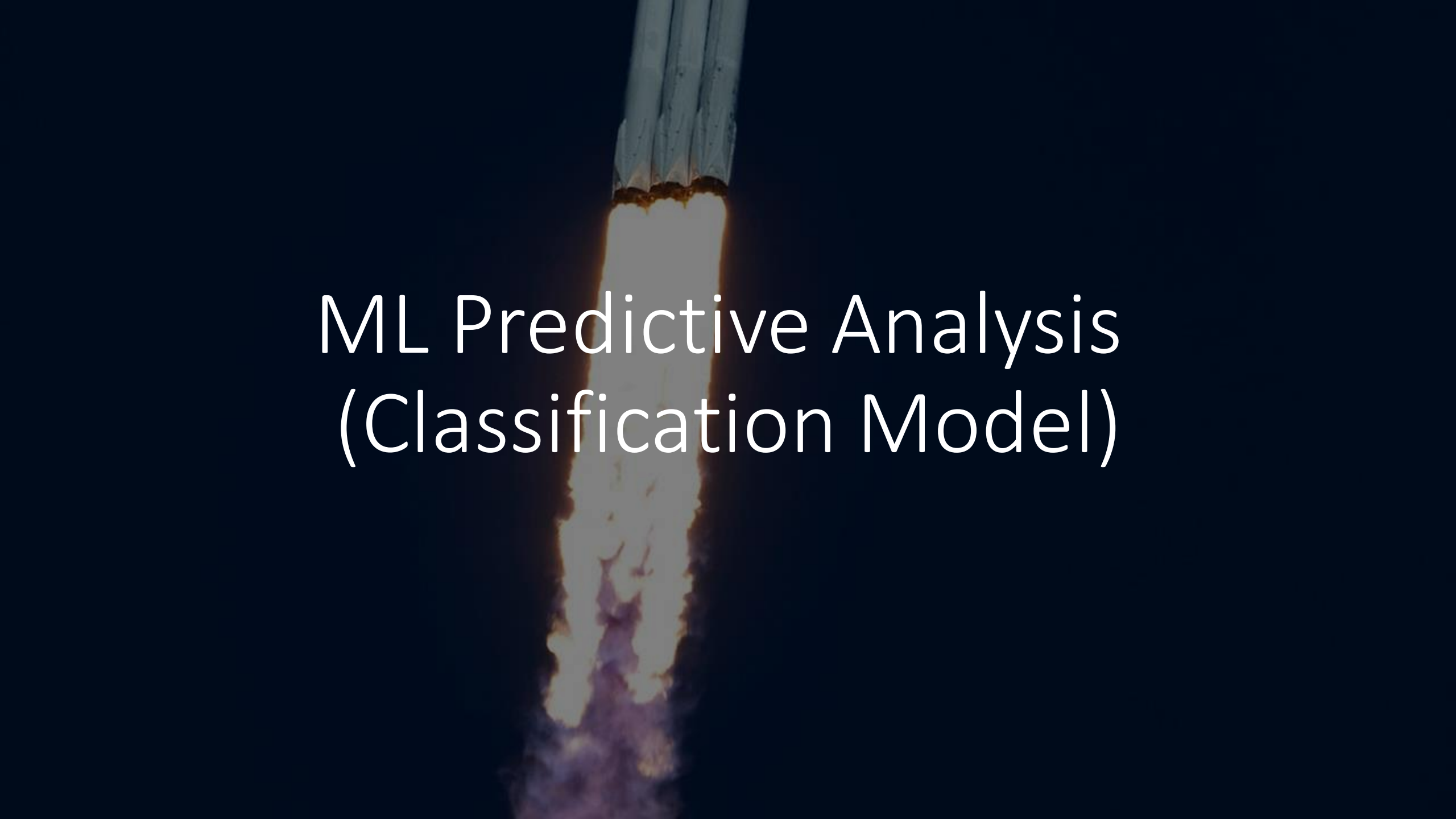
- Using the drop-down menu, "KSC LC-39A" launch site is chosen.
- "KSC LC-39A" launch site shows 76.9% successful launches (1) and 23.1% unsuccess launches (0).

Payload vs Success using Payload Slider

Using the payload range slider as shown in the screenshot, the optimum mass range is visualized.

Figure 1 shows the all the launches of the various booster version.

Figure 2 shows that in the payload range 2000-5000 Kg, with the selected booster versions, the rocket gives maximum launch success. There is not enough data for >7000Kg.

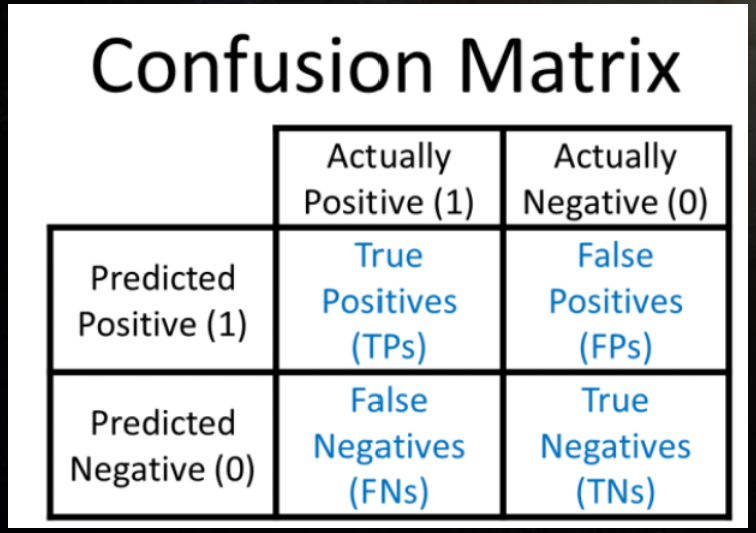


ML Predictive Analysis (Classification Model)

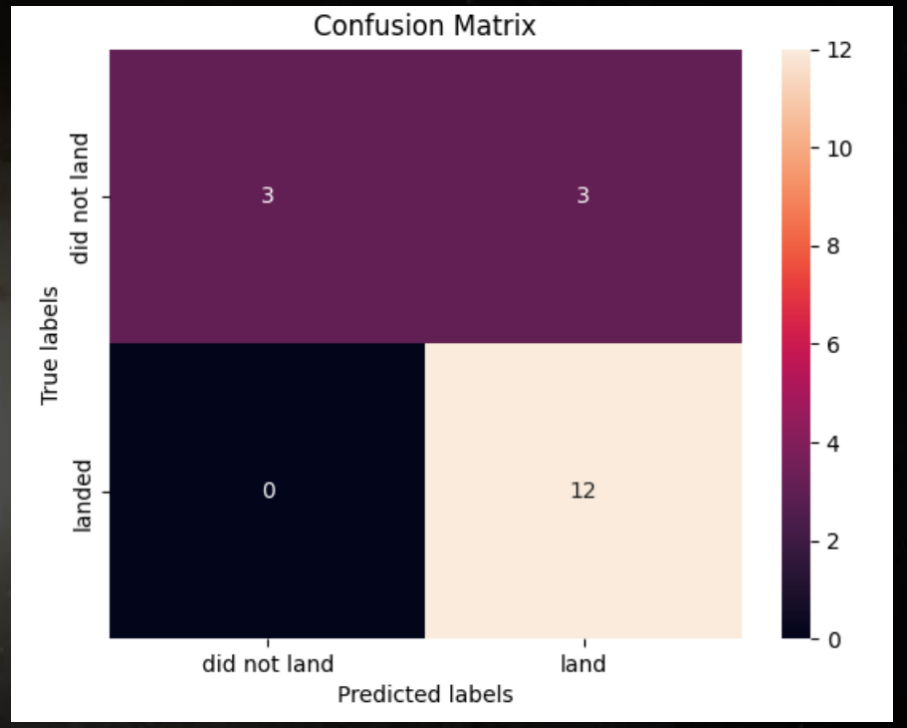
Confusion Matrix

The confusion matrix gives a detailed breakdown of the true vs the predicted labels.

The figure below can be compared to the confusion matrix to make conclusions about the model.



Ref:https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/



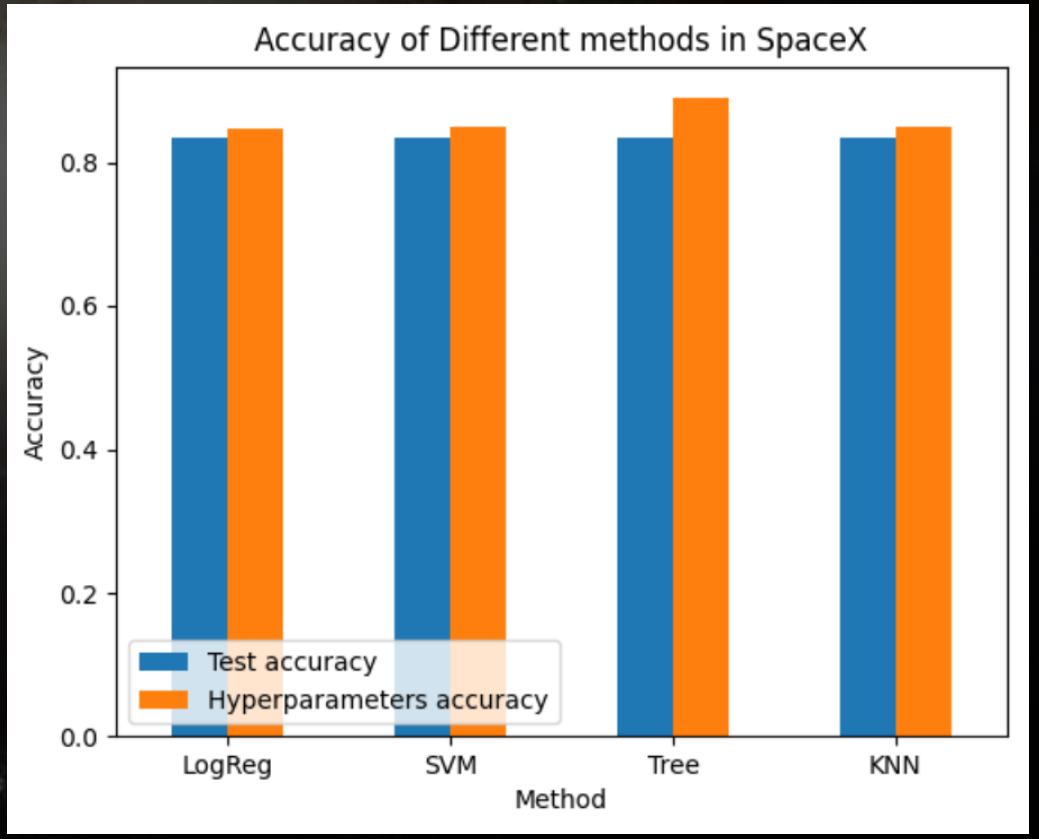
In this case, the model can predict with good accuracy of the true positive labels and true negative labels.

Accuracy of Data Model

From the machine learning models, it can be seen that DecisionTrees provides the highest test and hyper parameterized accuracy.

The models produces a high accuracy in predicting the next first stage landing success based on the variables.

	LogReg	SVM	Tree	KNN
Test Accuracy	0.833333	0.833333	0.888889	0.833333
Hyperparametric accuracy	0.846429	0.848214	0.914286	0.848214



Key Conclusions

- Orbit Type "ES-L1", "GEO", "HEO" , "SSO", "VLEO" all have a success rate of 100% of landing the first stage rocket.
- The success rates have increased over the years with 2019 having the highest success rate of 98%.
- The launch site with most success is "KSC LC-39A"
- All the launch sites are situated closed to coastlines, to avoid rockets landing on land.
- The ML prediction models provides high accuracy to the datasets with DecisionTrees methodology providing highest accuracy.
- The highest success rates come from payload mass range from 2000-5000 Kgs

Thank you

Github Link:https://github.com/Shirish026/CapstoneProject/tree/main/Capstone%20Project