IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Rusydi Al Affan
Agustus 2023

Outline

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

Executive Summary

- In this project, we harnessed the power of data science to analyze and optimize SpaceX's space launch outcomes. By leveraging various machine learning algorithms, we aimed to identify patterns and factors influencing launch success, ultimately enhancing mission reliability and efficiency.
- We gathered extensive datasets on SpaceX's space launch records. We employed multiple machine learning models, including Logistic Regression, and Support Vector Machine, Decision Tree. GridSearchCV was utilized to identify the best hyperparameters for each model, ensuring optimal performance. And Finally, we evaluated the models.
- Our data science approach has provided valuable insights into SpaceX's space launch outcomes. The Decision Tree model exhibited the highest accuracy on validation data, indicating its potential as the most robust and reliable method for predicting mission success. Leveraging data science in space exploration will continue to be instrumental in advancing humanity's reach into the cosmos.

Introduction

- Over the years, SpaceX has accumulated a vast amount of data on space launches, including various parameters such as booster versions, payload masses, and launch sites. With the increasing frequency of space missions, there is a need to analyze this data and optimize mission outcomes to ensure mission success and reliability.
- Problems you want to find answers

  1. Mission Success Prediction: Can we predict the success of space missions based on historical launch data and mission parameters?
  2. Influencing Factors: What are the key factors that influence the success or failure of space launches, such as payload mass or launch site location?
  3. Optimal Model Selection: Which machine learning model offers the most accurate predictions for mission success and can be used for future mission planning?
  4. Enhancing Mission Reliability: How can data science be leveraged to improve mission reliability and identify potential areas for optimization in SpaceX's space launches?

Section 1

Methodology

Methodology

Executive Summary

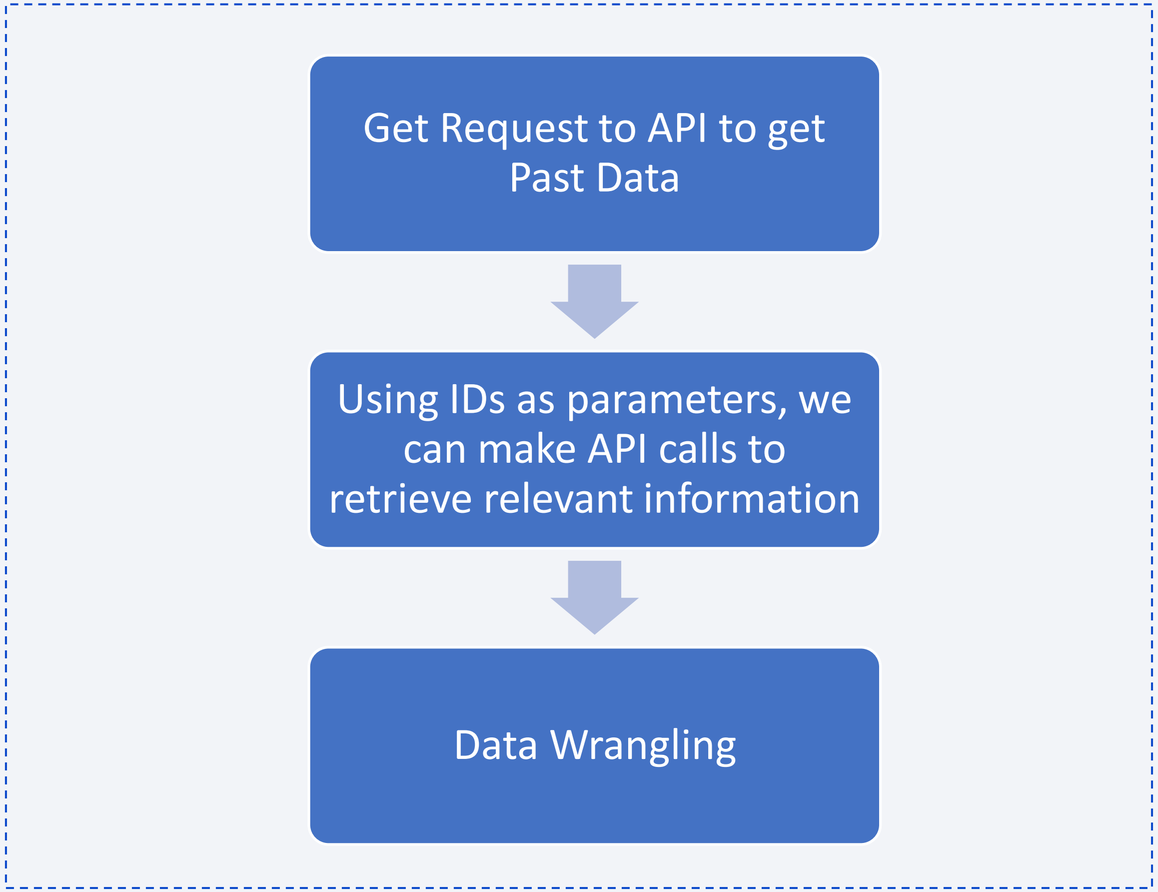
- Data collection methodology:
  - We gathered comprehensive space launch data from various sources, including booster versions, payload masses, launch sites, outcomes, and more. We get with API data Retrieval and Web Scraping
- Perform data wrangling
  - To ensure data quality, we cleaned, handled missing values, and standardized the dataset for further analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Utilizing machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM) and Decision Trees. We employed techniques like GridSearchCV to optimize model hyperparameters, enhancing their predictive capabilities. we visualized confusion matrices.

Data Collection

- Data From Space X was obtain from 2 source:
  1. SpaceX API (https://api.spacexdata.com/v4/launches/past)
  2. WebScraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- During the data collection process, we obtained several lists of features such as 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', and 'date_utc'. However, it is important to note that the data in these lists only consists of IDs. These IDs serve as parameters that need to be used to retrieve additional information from the provided API.

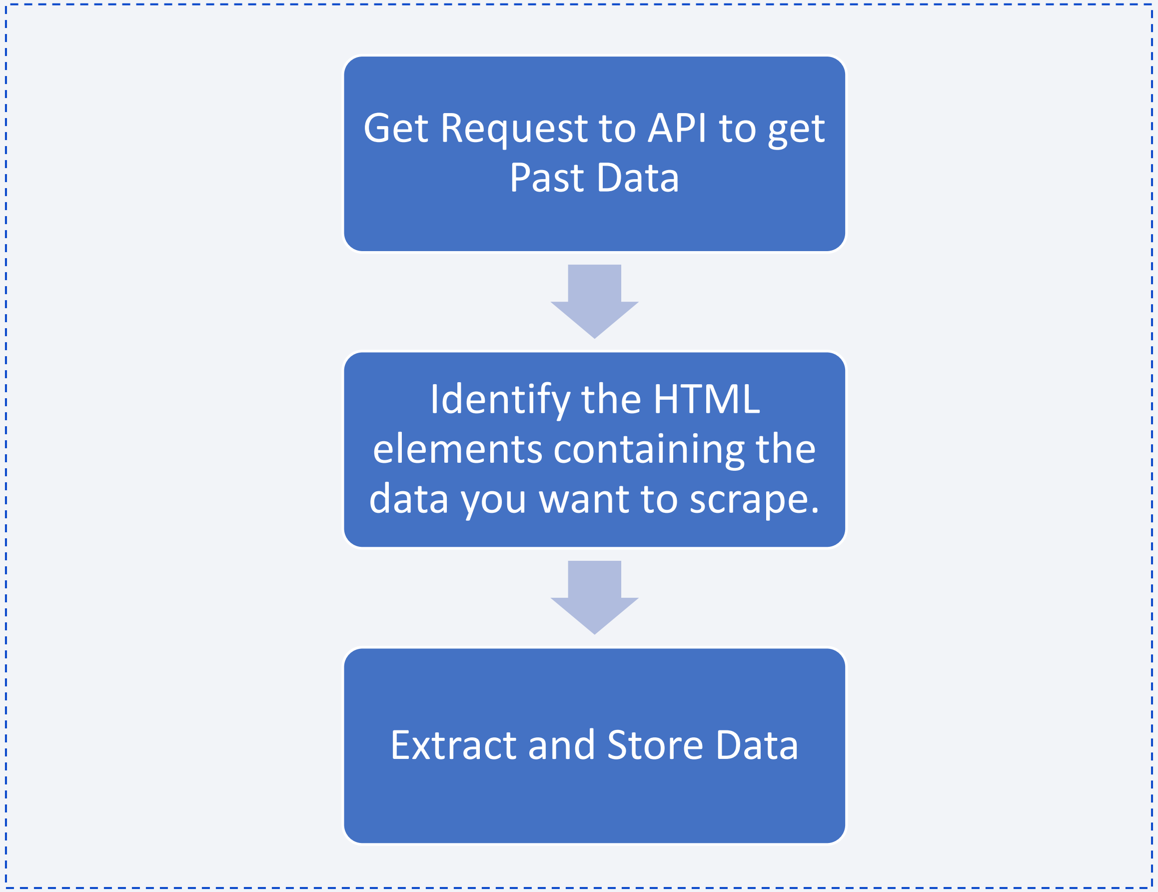
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- The GitHub URL of the completed SpaceX API calls notebook : https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/jupyter-labs-spacex-data-collection-api.ipynb



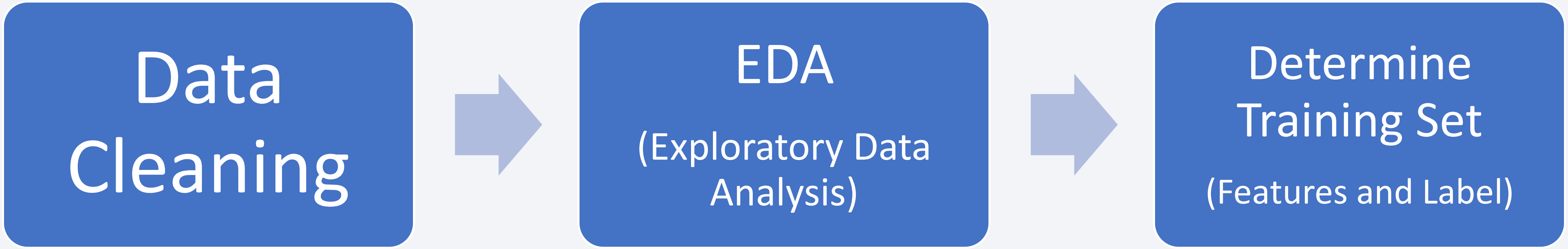
Data Collection - Scraping

- Data SpaceX can also be obtained from Wikipedia via WebScraping
- The GitHub URL of the completed web scraping notebook : https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/jupyter-labs-webscraping.ipynb



Data Wrangling

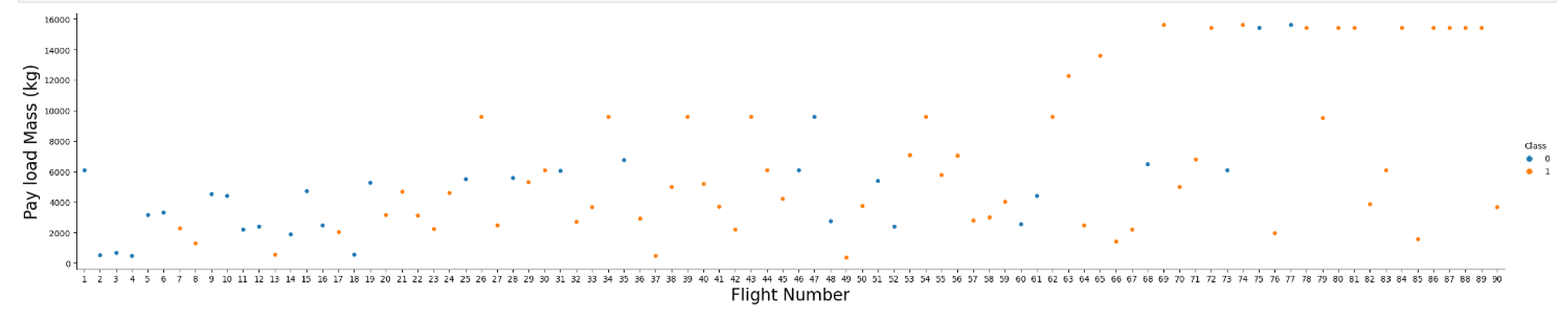
- The collected data underwent extensive cleaning to handle missing values, remove duplicates, and correct inconsistent data formats.
- We also perform EDA (Exploratory Data Analysis) to more understand the Information contains inside the data



- GitHub URL of your completed data wrangling related notebooks : https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of
- For example : This chart show relationship between flight number and payload Mass with Class Success Status of The Operation



- Add the GitHub URL of your completed EDA with data visualization notebook: https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(1).ipynb

EDA with SQL

- Summarize the SQL queries:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the   names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.
  - 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.

- GitHub URL of your completed EDA with SQL notebook : https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

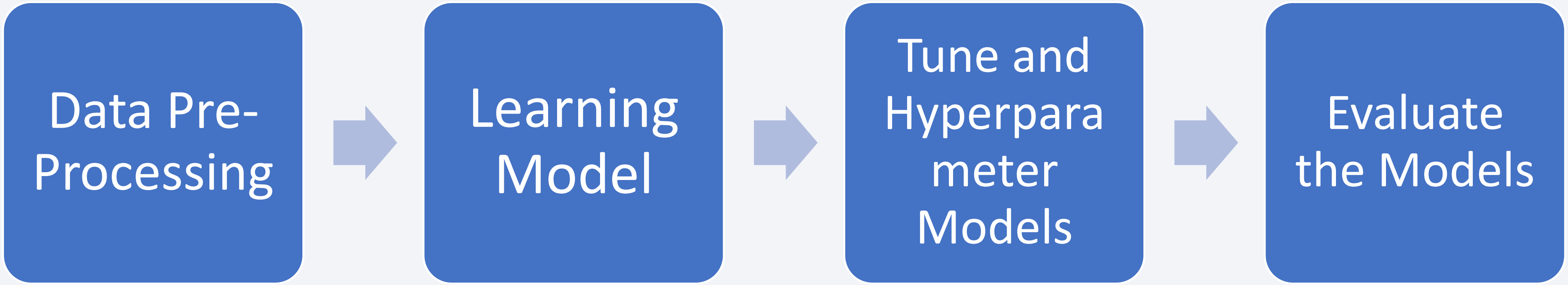
- Summarize what map objects :
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.
- To demonstrate the distance between critical objects and the radius of the launch site, providing an overview of safety and accessibility.
- GitHub URL of your completed interactive map with Folium map: https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions that added to a dashboard :
  - A Pie Chat that showing the total launches by a certain sites (Sites Can be Selected by a Dropdown Menu)
  - scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version. Payload Mass can be selected by range plot.
- Explain why you added those plots and interactions
- GitHub URL of your completed Plotly Dash lab : https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/spacex_dash_app.py

Predictive Analysis (Classification)

- loaded the data using numpy and pandas, transformed the data, split our data into training and testing, then built different machine learning models and tune different hyperparameters using GridSearchCV. And we used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning. Finally, We found the best performing classification model.



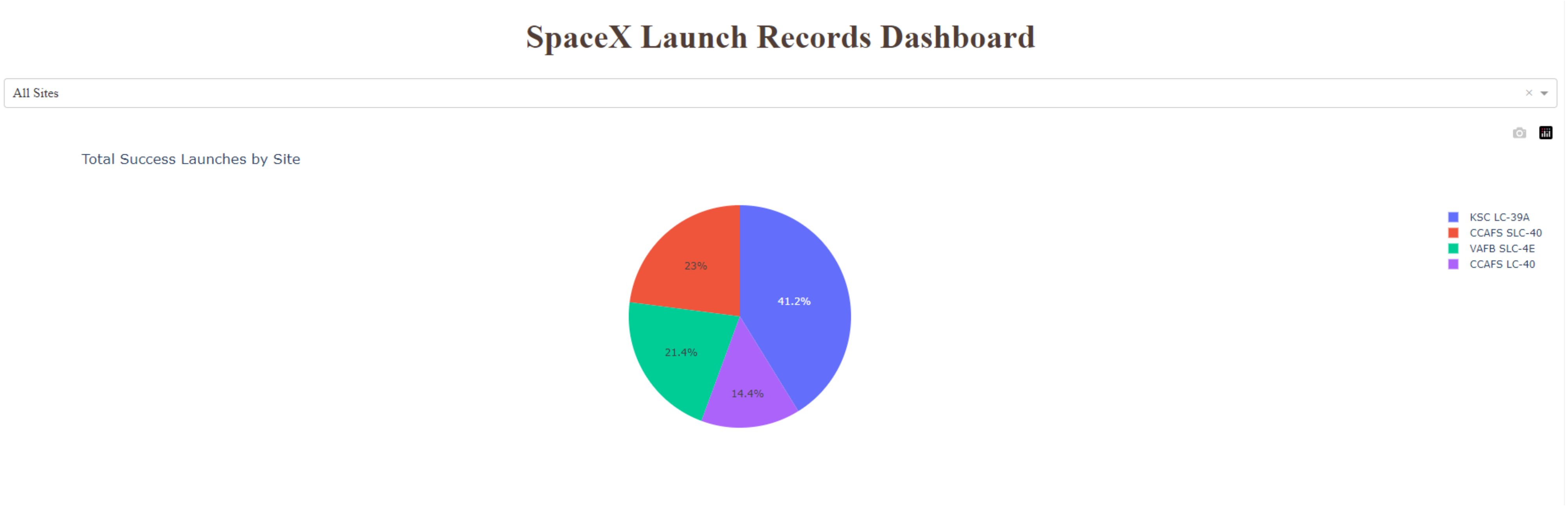
- Add the GitHub URL of your completed predictive analysis lab: https://github.com/rusydi16/Final-Assignment/blob/main/Course%2010/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

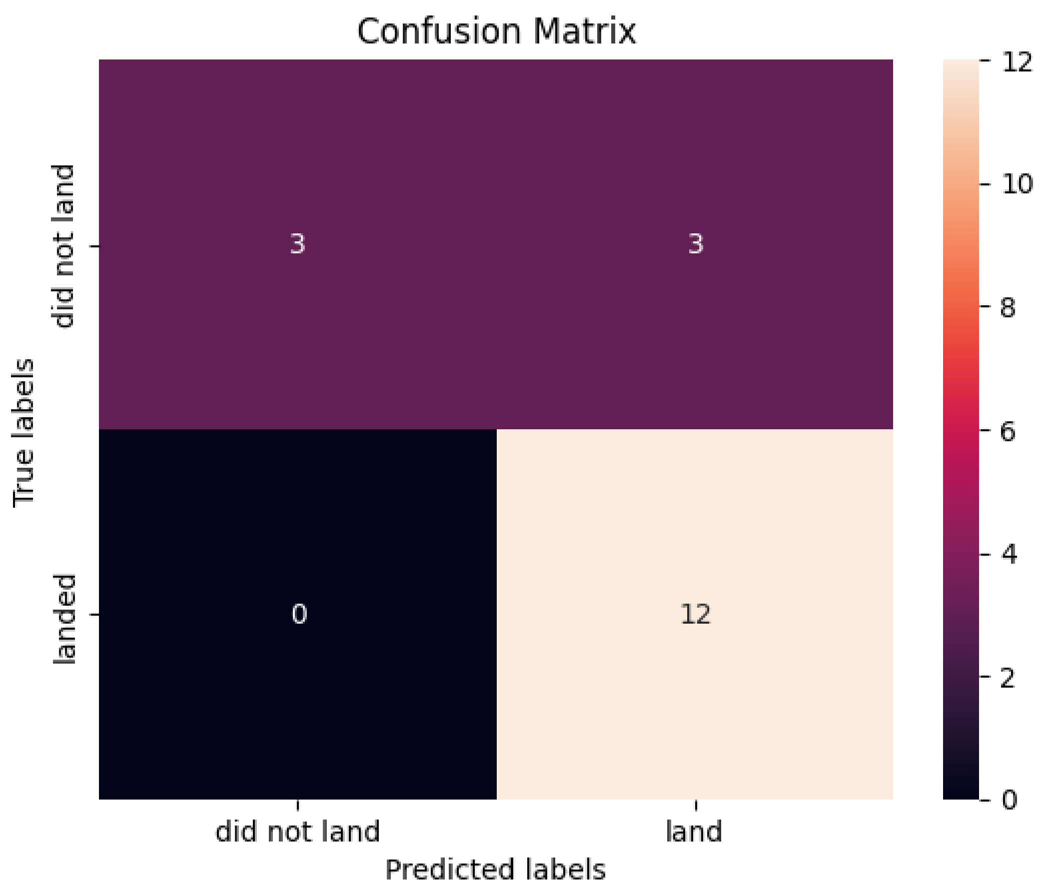
Results

- Interactive analytics demo in screenshots :



Results

- Predictive analysis results
  - With The best Model Method is **Decision Tree**
  - Accuracy on Validation Data : **~88%**

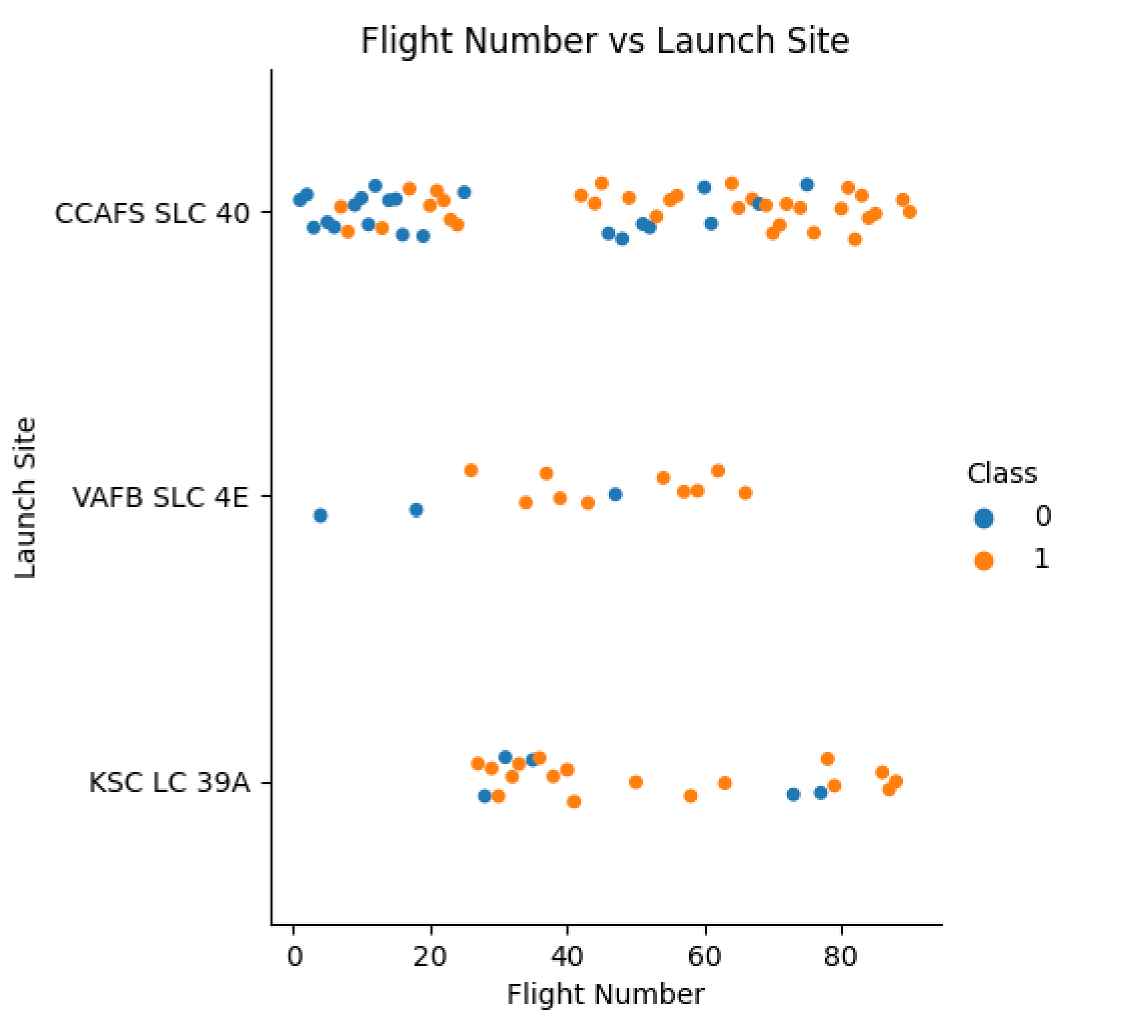


Section 2

Insights drawn from EDA

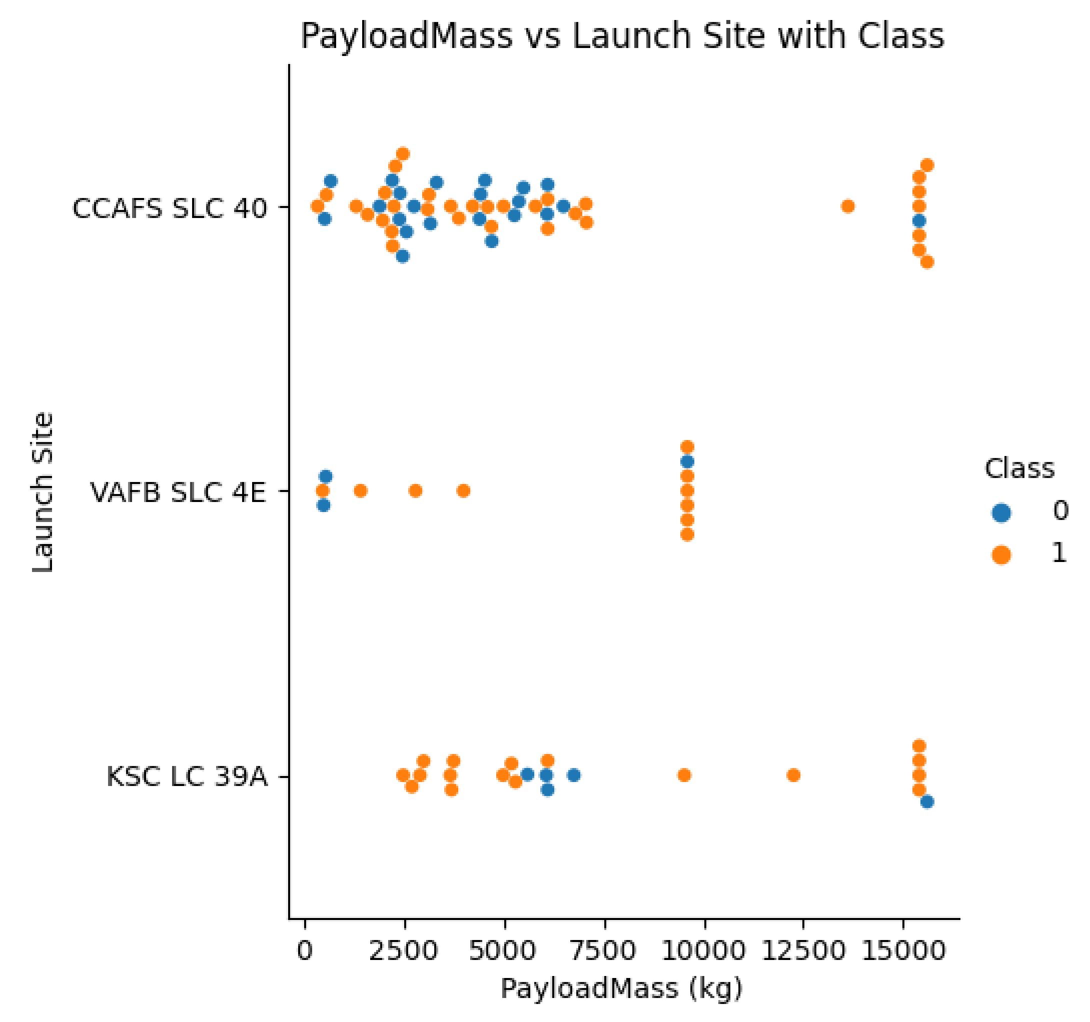
Flight Number vs. Launch Site

we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



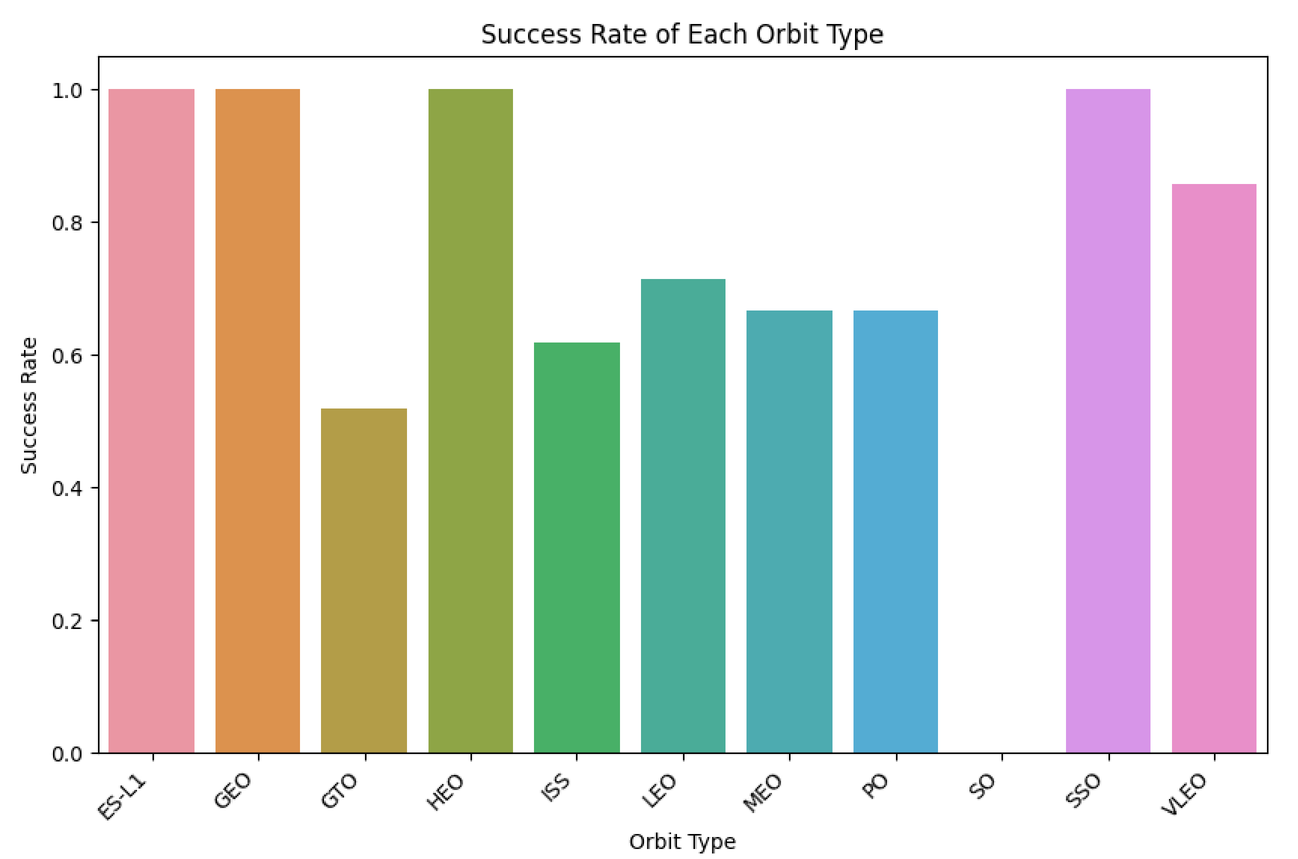
Payload vs. Launch Site

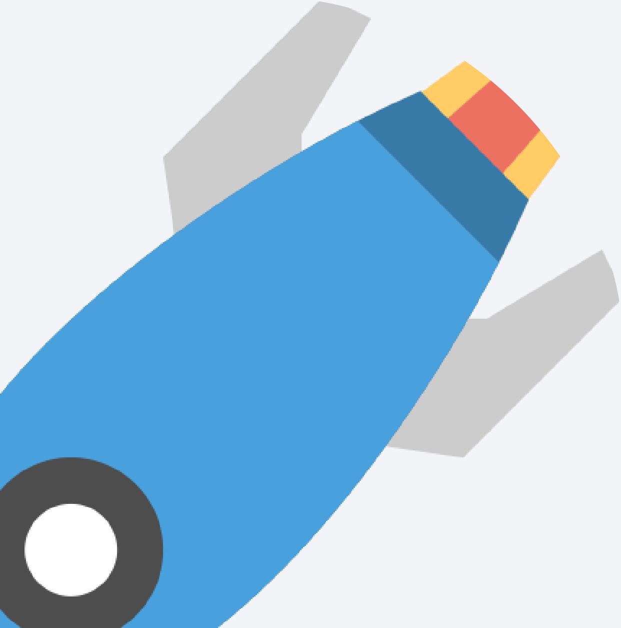
The Greater the payload mass for launce Site CCAFS SLC 40 the higher the success rate for the rocket



Success Rate vs. Orbit Type

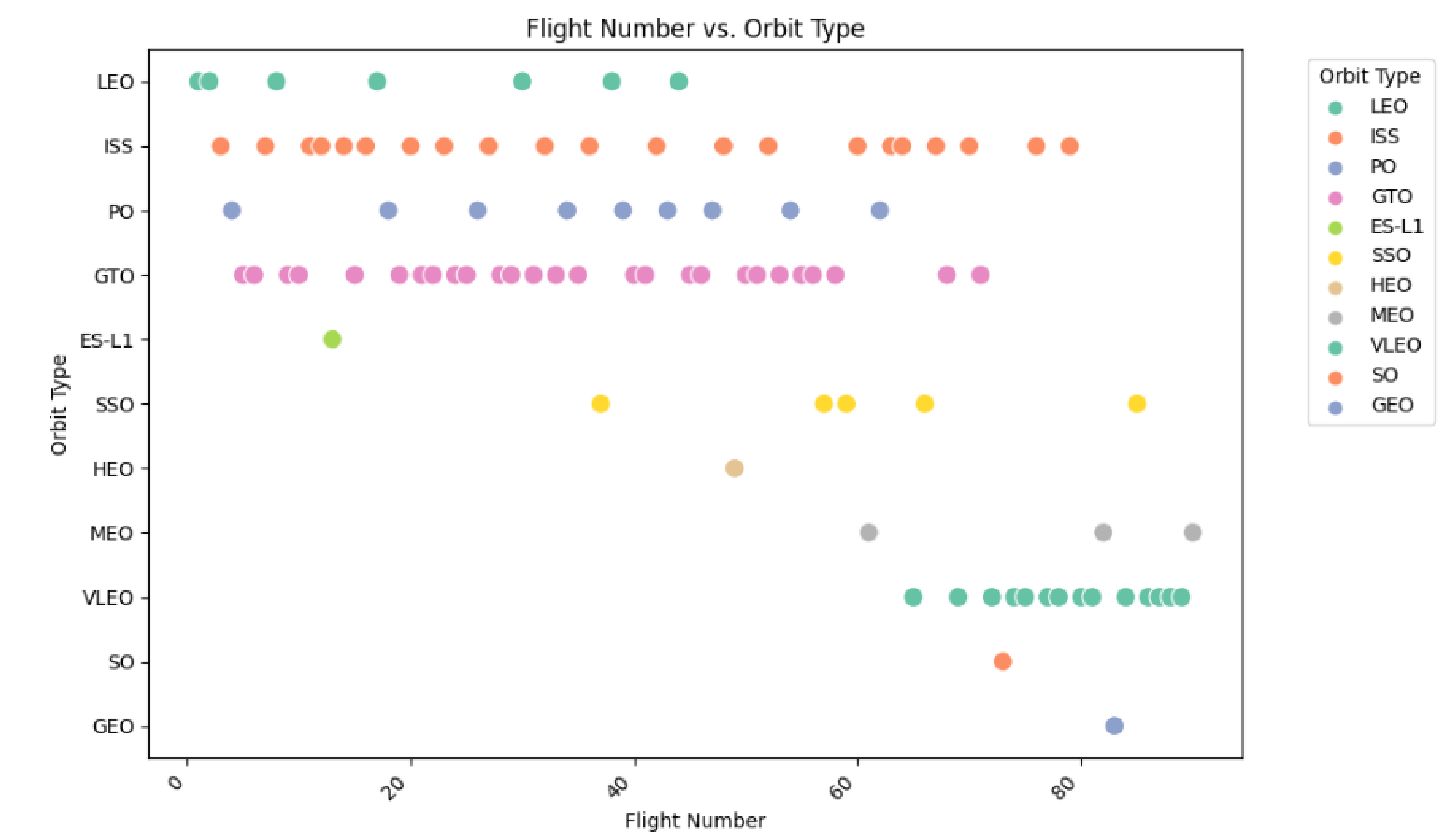
- From the plot, we can see ES-L1, GEO, HEO, SSO had the most success rate and SO had the less success rate





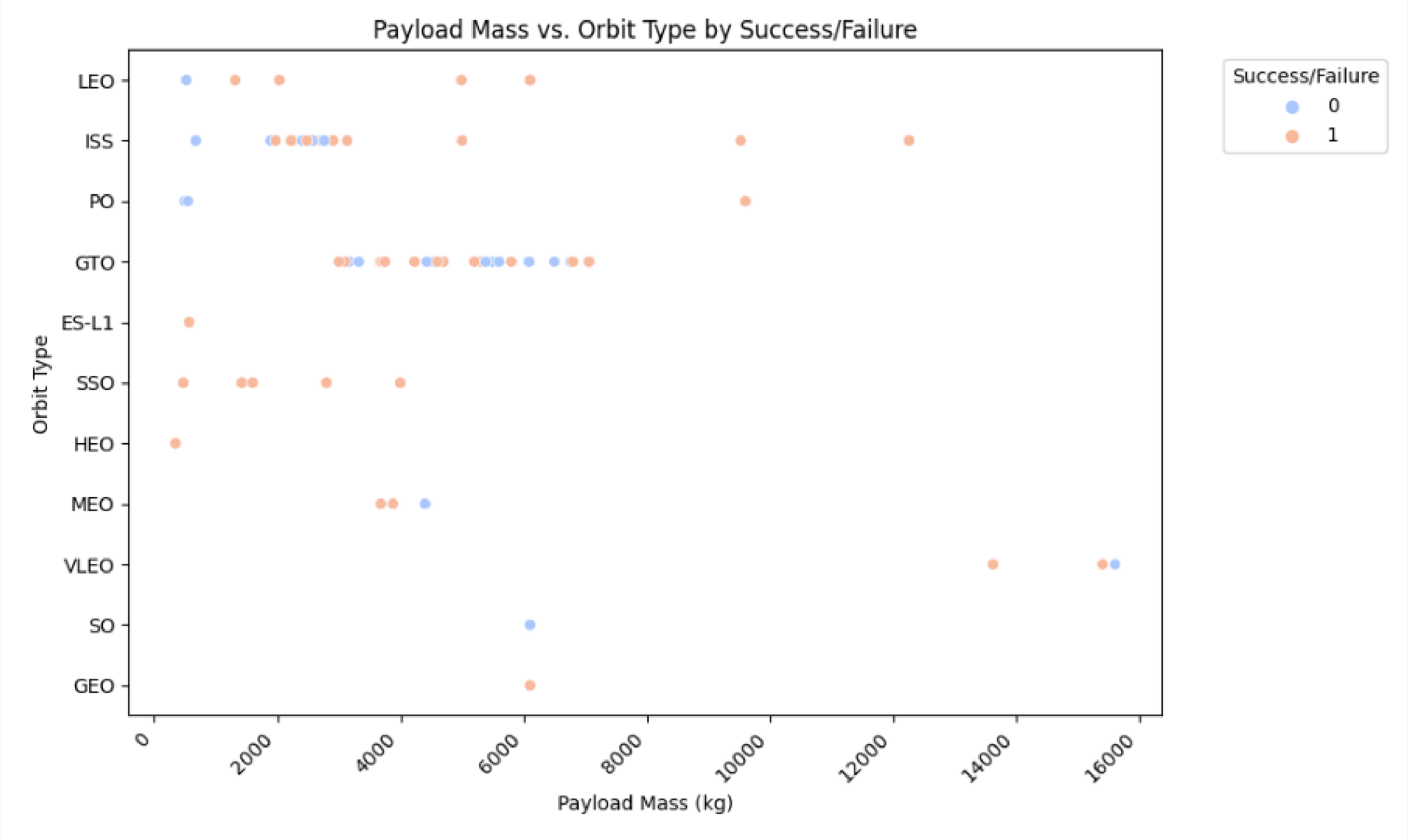
Flight Number vs. Orbit Type

- We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there **is no relationship** between flight number and the orbit



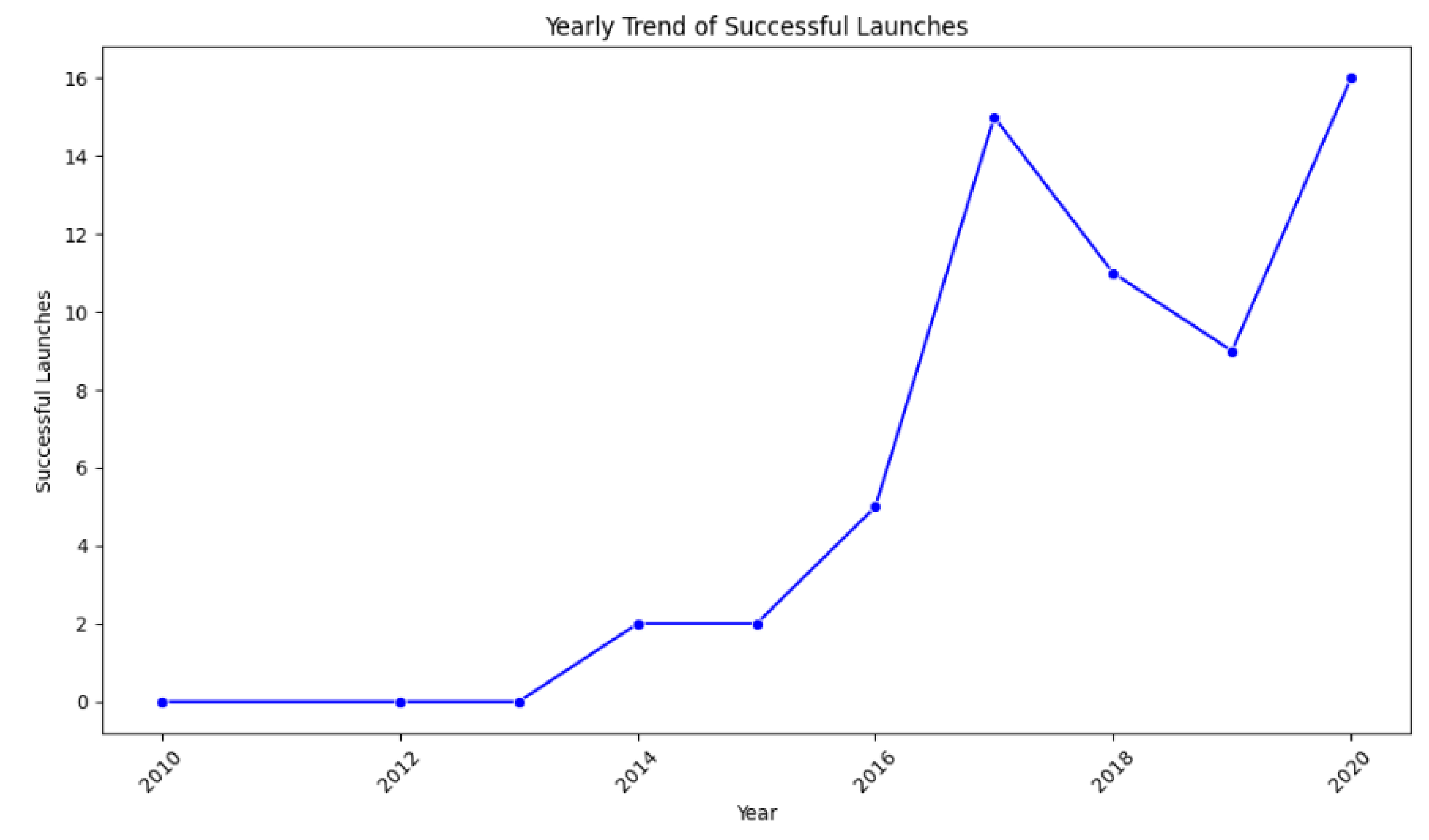
Payload vs. Orbit Type

We can observe that with heavy payloads, the successful landing are more for VLEO, LEO and ISS orbits



Launch Success Yearly Trend

From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

Display the names of the unique launch sites in the space mission

```
In [10]: %sql SELECT DISTINCT launch_site FROM SPACEXTBL

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

We used LIKE and Wildcard symbol for query the pattern of CCA launch Sites

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [12]:

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

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
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 atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

We calculated the total payload carried by boosters from NASA as 45596 using SUM() Function

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

```
In [25]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE lower("Customer") LIKE 'nasa (crs)%'

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

Out[25]:

SUM(PAYLOAD_MASS__KG_)
48213.0

Average Payload Mass by F9 v1.1

We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

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

In [24]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE lower("Booster_Version") LIKE 'f9 v1.1'
```

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

Out[24]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015. We used Substr() to ekstract DAY, MONTH, YEAR FROM Date Column because the data type from column is not perform well using MIN() Function.

```
[32]: %%sql SELECT MIN(substr("Date",7)||'-'||substr("Date",4,2)||'-'||substr("Date",1,2)) AS SUCCESS_LANDING_DROUND_PAD
              FROM SPACEXTBL
              WHERE TRIM(Landing_Outcome) LIKE 'Success (ground pad)'

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

[32]:

SUCCESS_LANDING_DROUND_PAD
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

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

```
[32]: %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_<6000)

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

[32]:

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

Total Number of Successful and Failure Mission Outcomes

We User Count() Function and GROUP BY to know each Mission Outcomes Flight Number

List the total number of successful and failure mission outcomes

```
[33]: %sql SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTBL GROUP BY 1 ORDER BY 2 DESC
```

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

[33]:

Mission_Outcome	COUNT(*)
None	898
Success	98
Success (payload status unclear)	1
Success	1
Failure (in flight)	1

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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[35]:
```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

[35]:

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

We used SUBSTRING to extract Months from Date Column and WHERE Clause to filter the output.

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[40]:
```
%%sql
SELECT substr("Date", 4, 2) as month, "Booster_Version", "Landing_Outcome", "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Failure (drone ship)' and substr("Date",7,4)='2015'
```

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

[40]:

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

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

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

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.

[25]:

```
%%sql
SELECT "Landing_Outcome", COUNT("Landing_Outcome")
FROM SPACEXTBL
where substr("Date",7)||substr("Date",4,2)||substr("Date",1,2)
      between '20100604' and '20170320'
GROUP BY 1
ORDER BY 2 DESC
```

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

[25]:

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

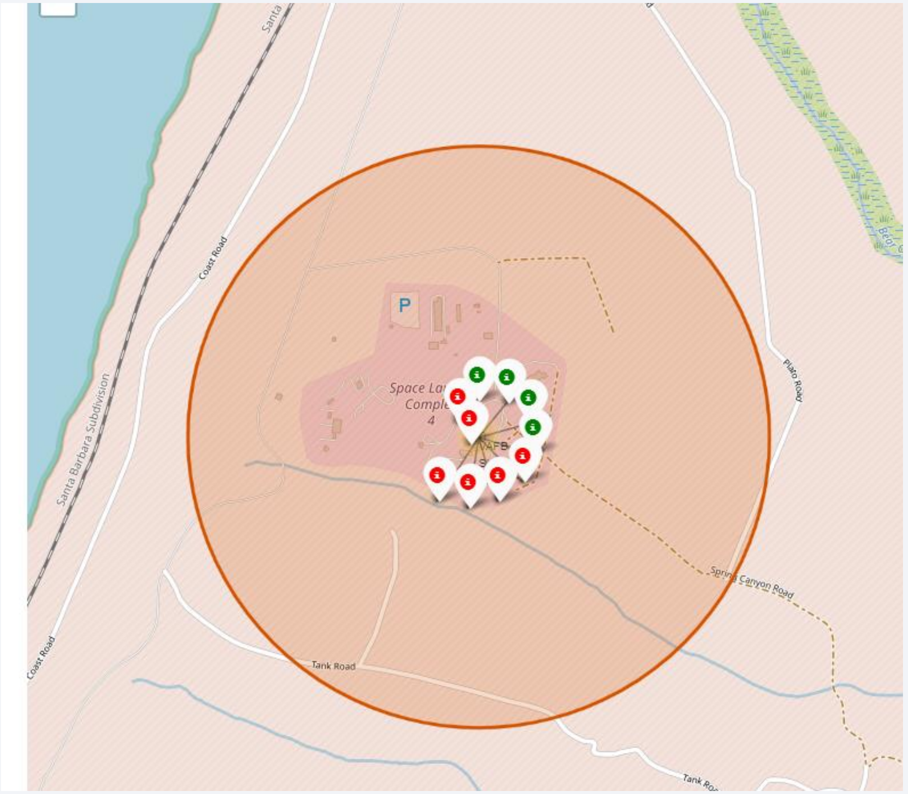
Section 3

Launch Sites Proximities Analysis

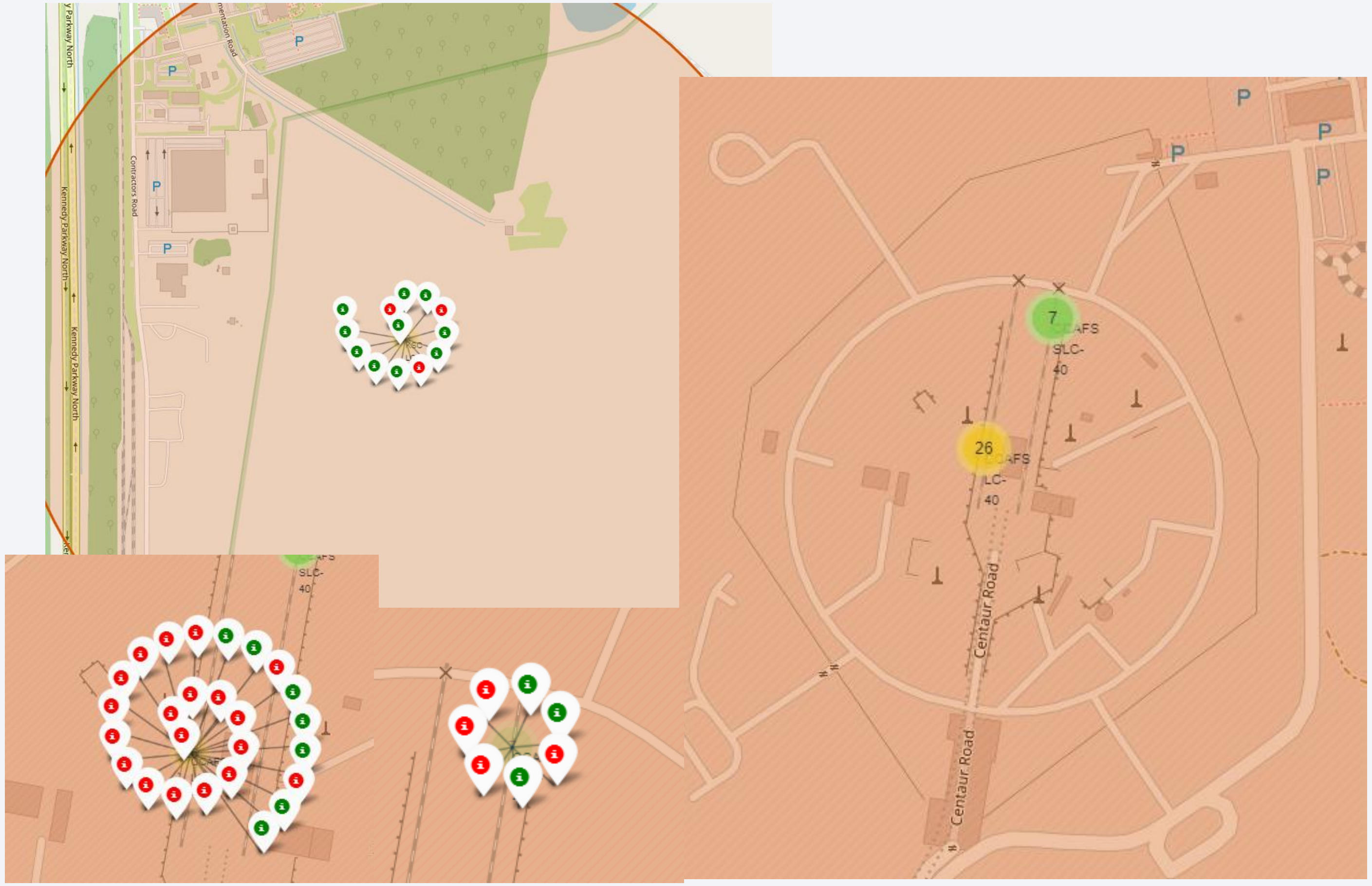
All launch sites global map markers

We Can see That the SpaceX Launch Sites are in United States of America Coasts. Florida and California

Markers showing launch sites with color labels



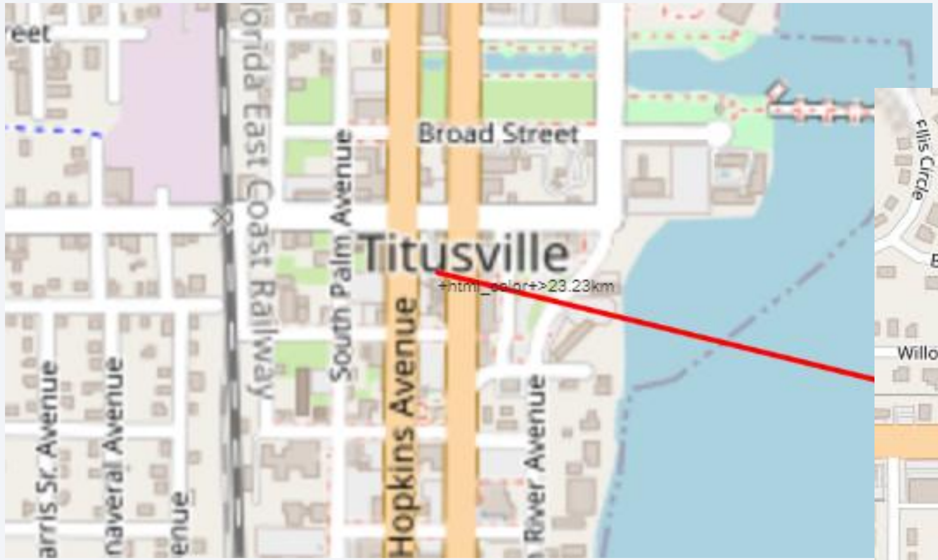
California Launch Sites

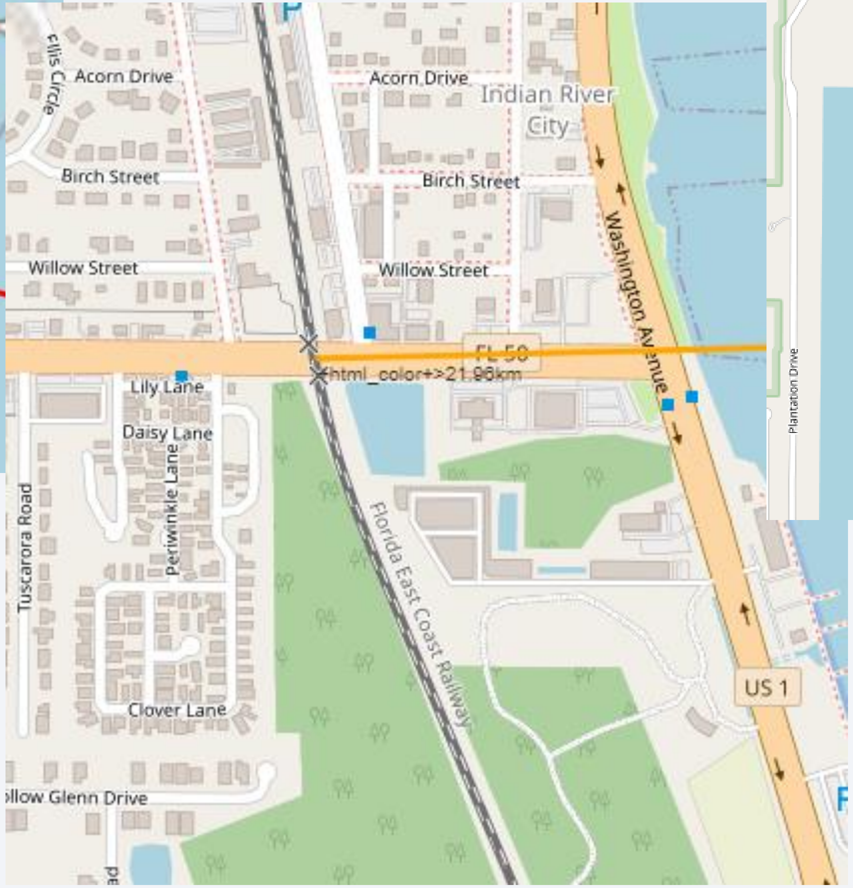


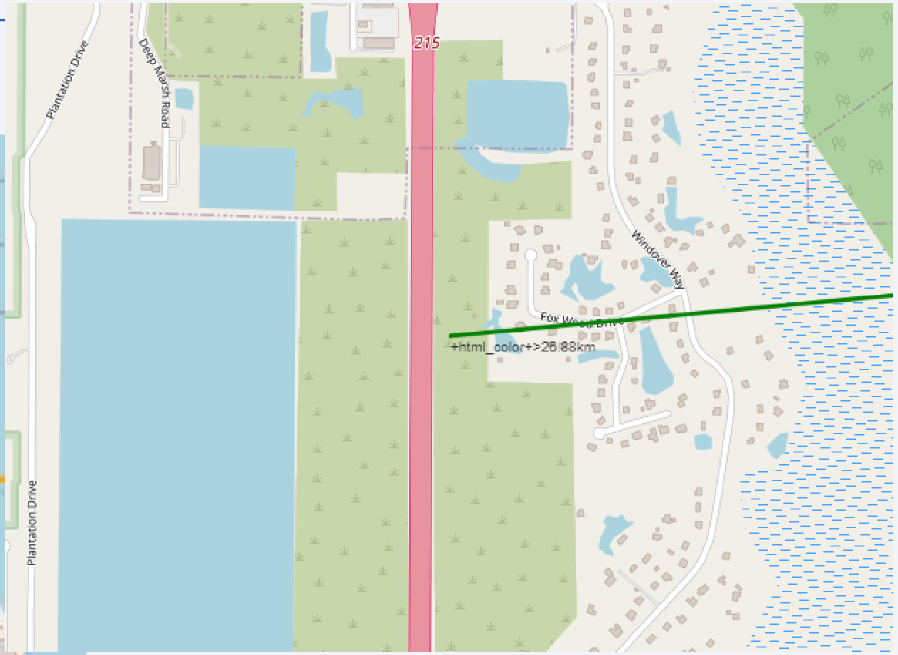
Florida Launch Sites

Green Marker shows successful Launches and *Red Marker* Show Failures

Launch Site distance to landmarks







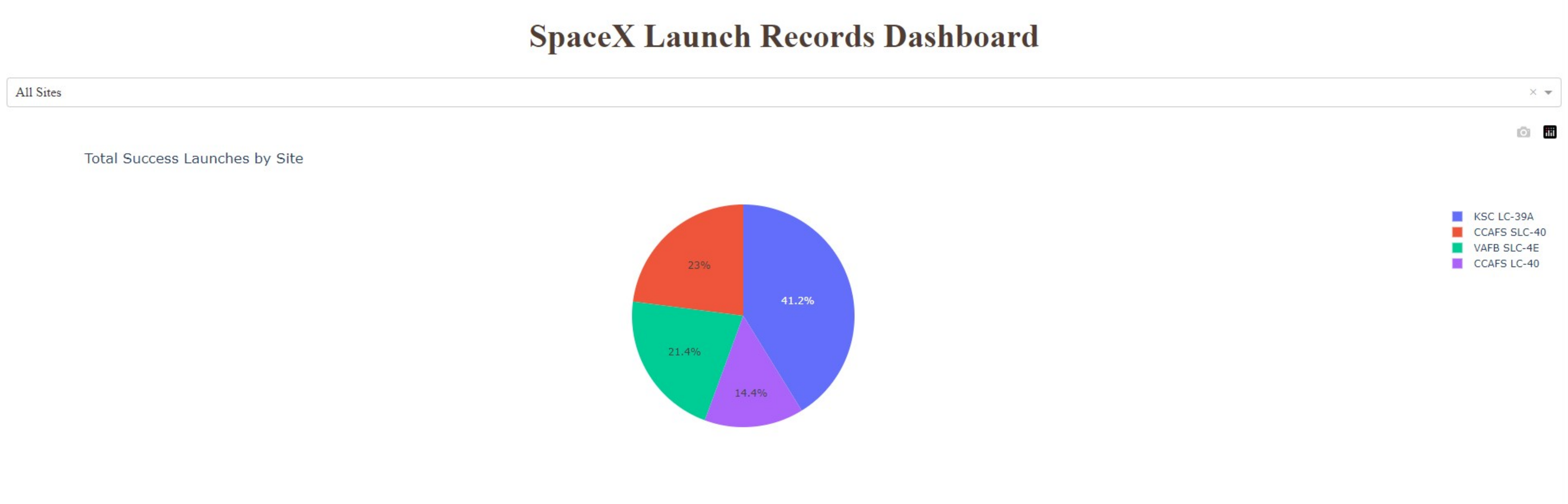
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

- City Distance 23.234752126023245
- Railway Distance 21.961465676043673
- Highway Distance 26.88038569681492
- Coastline Distance 0.8627671182499878

Section 4

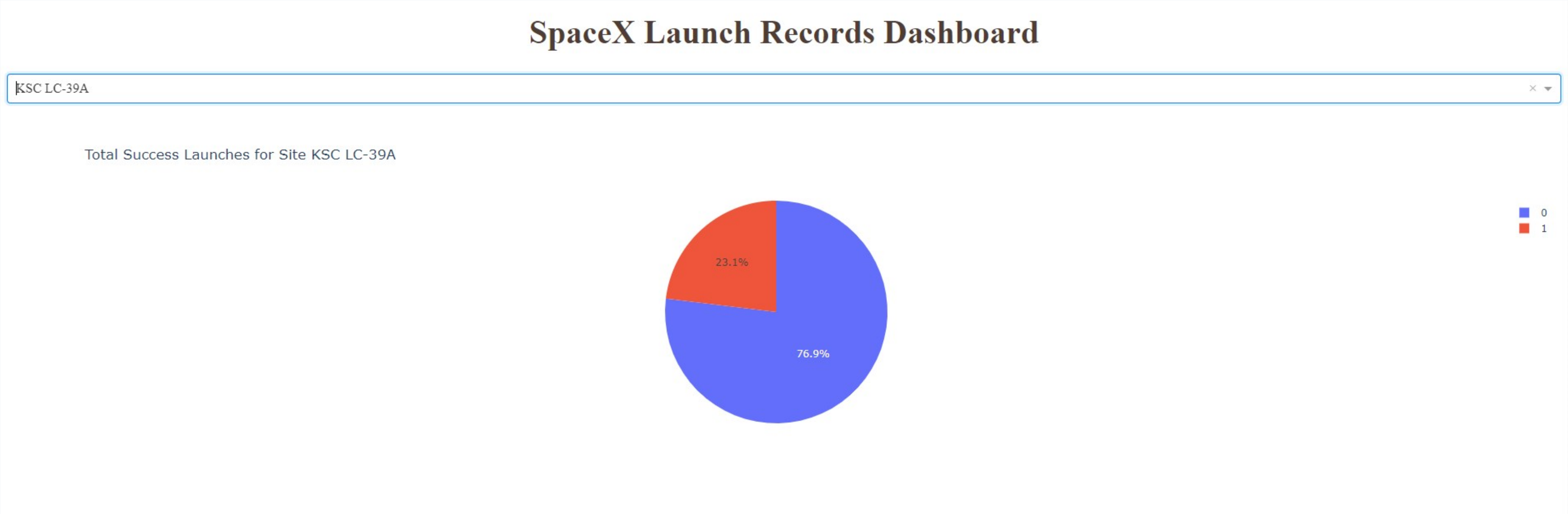
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



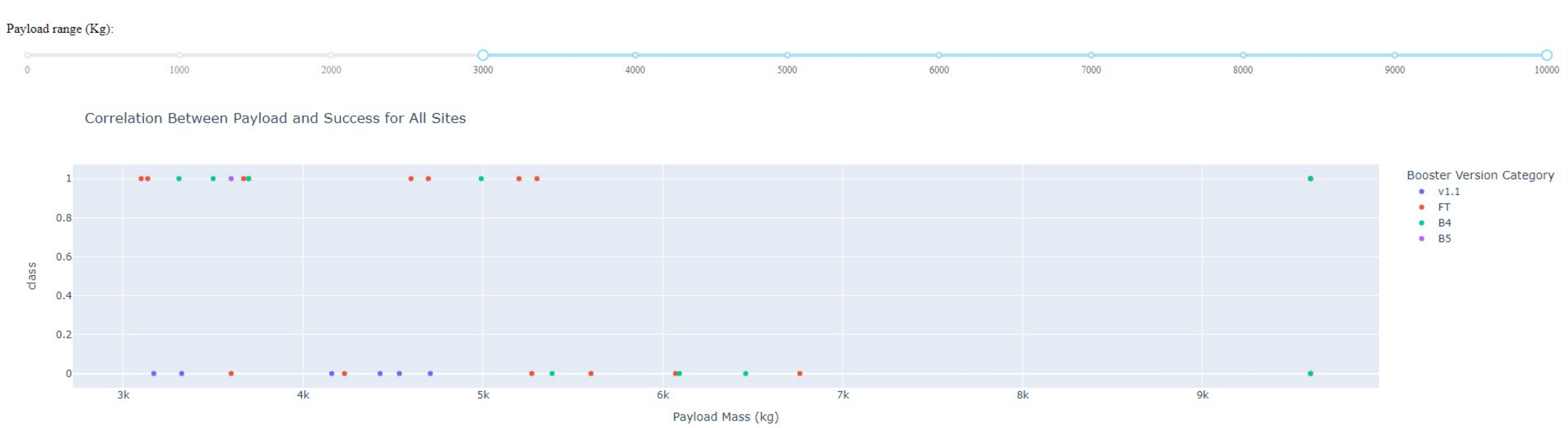
We Can see that KSC LC-39A had the most successful launches from all the sites

Pie chart showing the Launch site with the highest launch success ratio

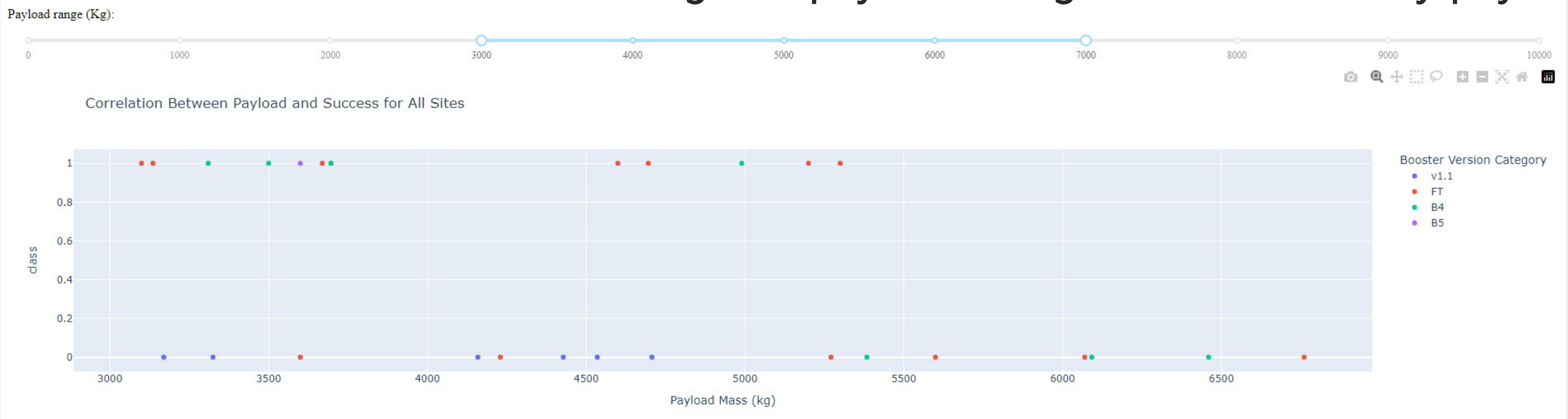


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



We can see that success rate low weighted payload is higher than the heavy payloads

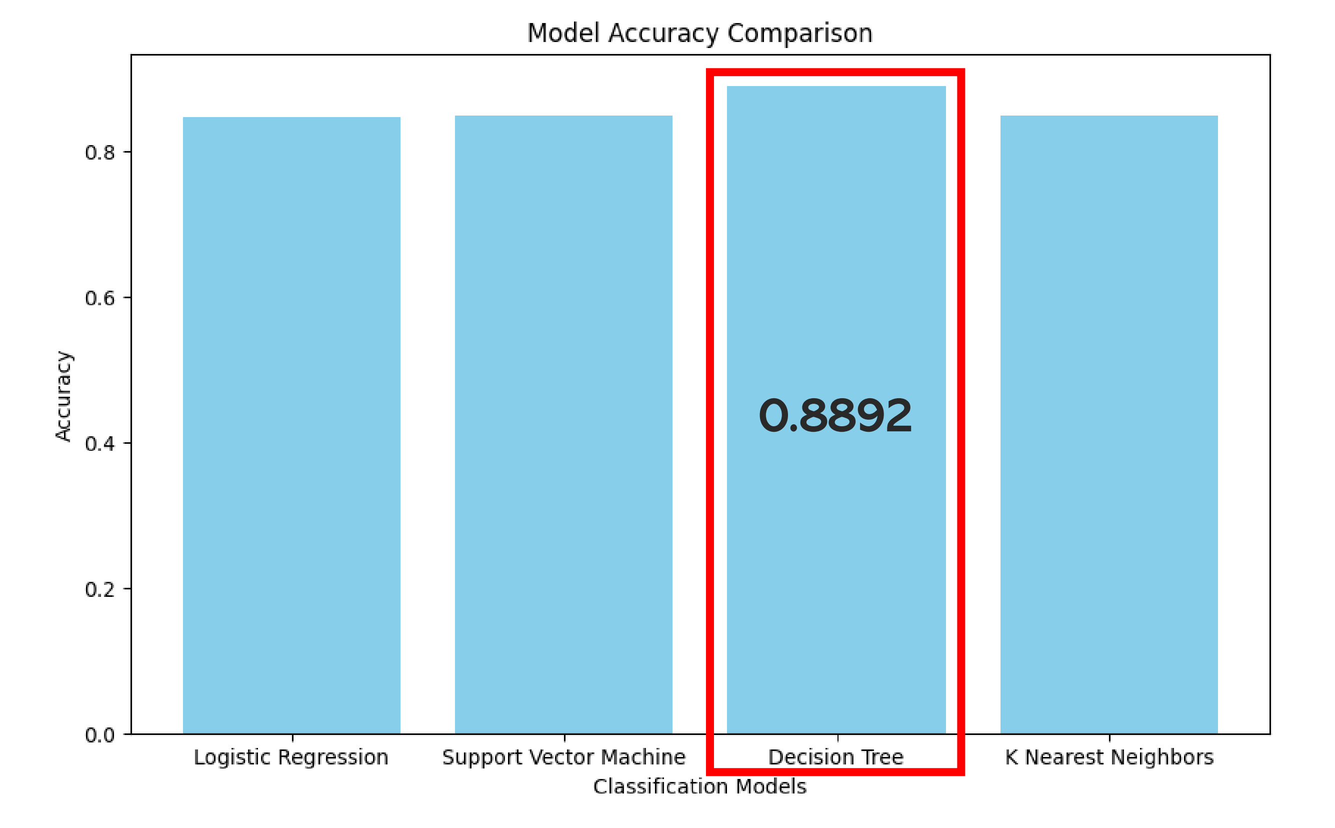


Section 5

Predictive Analysis (Classification)

Classification Accuracy

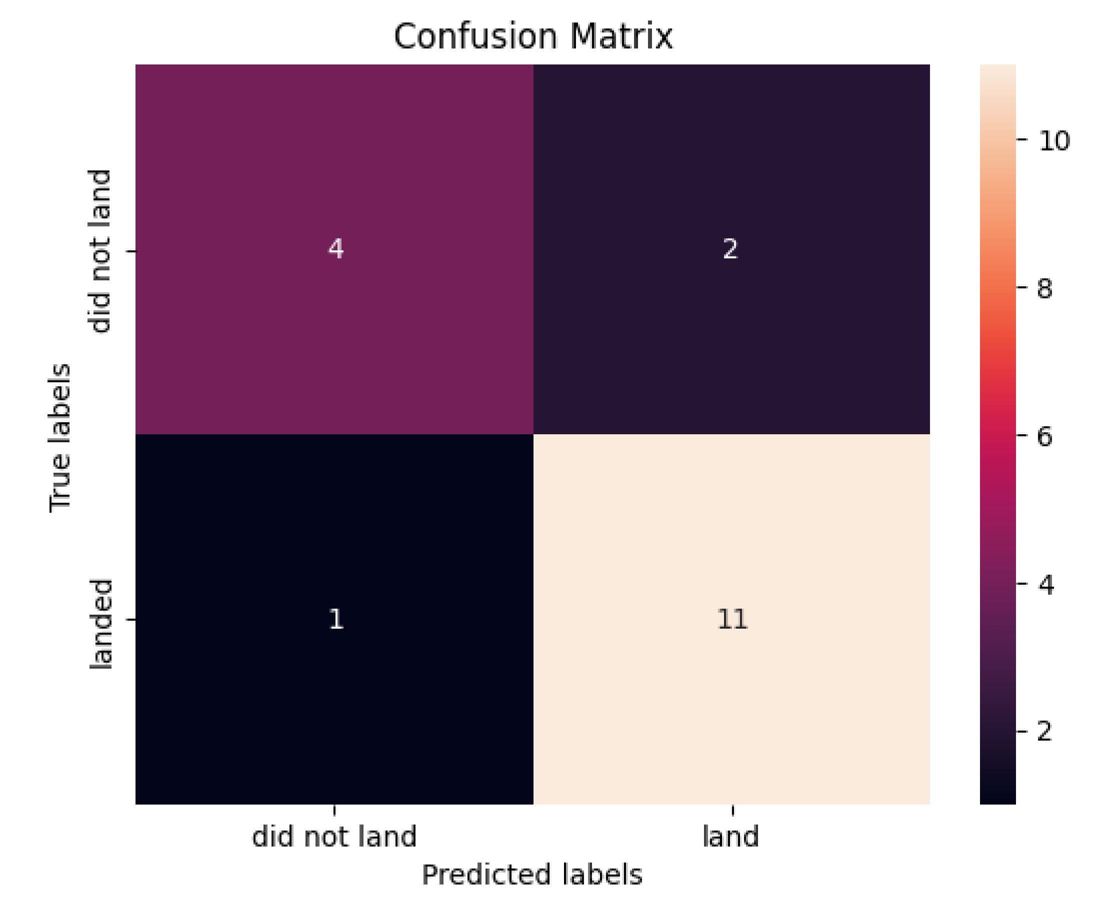
- Best performing method: **Decision Tree**
- Accuracy on validation data: **0.8892857142857142**



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.

Appendix

Github Repositori :

https://github.com/rusydi16/Final-Assignment/tree/main/Course%2010

Thank you!