Winning Space Race with Data Science

Lin Htet Win

09/11/2023

Outline

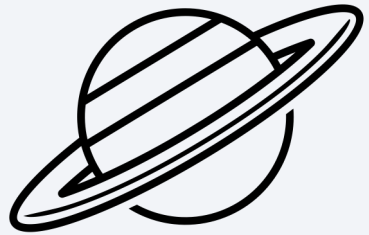
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data collection with api
  - Exploratory Data Analysis including data wrangling, data visualization and interactive dashboard to analyze the launch record.
  - Machine Learning models
- Summary of all results
  - Exploratory Data Analysis shows the result of meaningful patterns and features to model the machine learning.
  - Interactive Map to see the launch site and Interactive dashboard to see the launch record of all launch sites.
  - Machine Learning models show the best model to predict and which characteristics are affecting the success rate of the land state.

Introduction

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. This project is to act as SpaceY the company trying to compete and bid against SpaceX. The main purpose of this project is to estimate the launch price of each and predicting the successful land state with machine learning by gathering information about Space X and creating dashboards for the team.



Section 1

Methodology

Methodology

Executive Summary

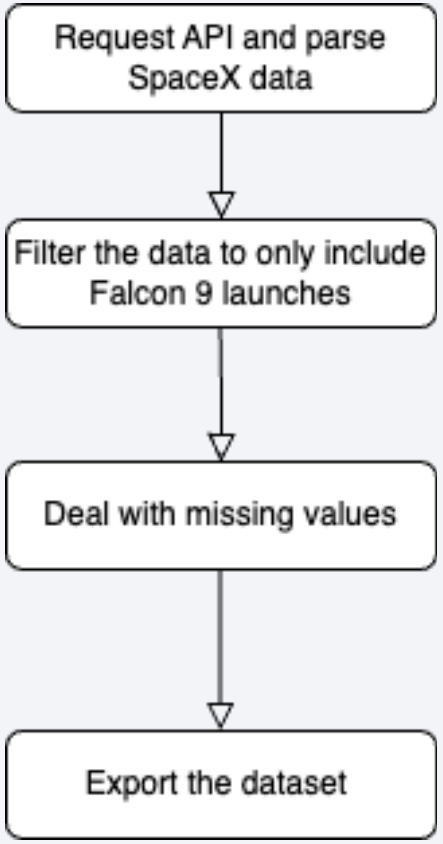
- Data collection methodology:
  - Data was collected from an API, specifically SpaceX Rest API.
- Perform data wrangling
  - Data was processed by creating a landing outcome as variable y (that contains 0,1) to represent the outcome of each launch.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until now is normalized and split the data into train and test tests. Four different machine learning models were used to evaluate and predict the data to see which model has better performance.

Data Collection

- First stage was collecting the required data.
- Datasets were collected through API and Web Scraping.
- You can find the dataset in the following links:
  - SpaceX Rest API: "https://api.spacexdata.com/v4/launches/past"
  - Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

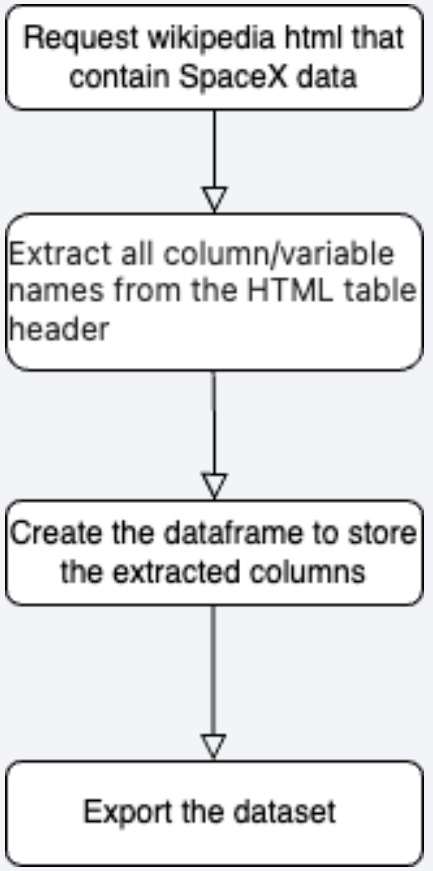
Data Collection – SpaceX API

- First, we request the data from SpaceX Rest API that offers all the information we need.
- We parse the data which means applying the custom functions to change the id value column to get the name we want.
- Next, we filter the data to only include Falcon 9 launches and deal with the missing values.
- Finally, we exported the collected data for further use case.
- GitHub URL for data collection juypter notebook to see the code



Data Collection - Scraping

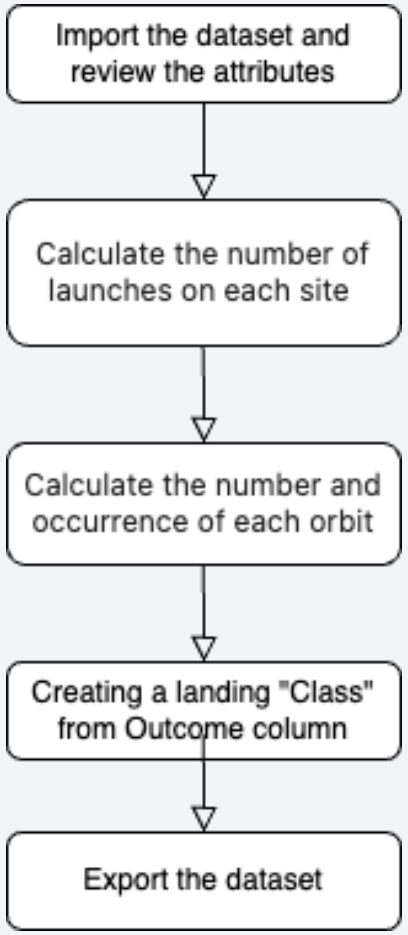
- In data collection through Web Scraping is not the same as data collection with API.
- First, we request the html page that contains SpaceX data and extract all column and variable from Html Table Header.
- Next, we create the dataframe to store the extracted columns and exported the dataframe as a csv file for further use cases.
- GitHub URL for data collection with web scraping juypter notebook to see the code



Data Wrangling

- We will do Exploratory Data Analysis to the dataset from the previous section and review the attributes.
- In this step, we will calculate the number of launches on each site and occurrence of each orbit types.
- The purpose of this is to prepare the data in a way that makes it accessible for effective use case so we create the landing Class label for training supervised models.
- Finally, we exported dataset as a csv file for further use cases.

- GitHub URL for data wrangling notebook



EDA with Data Visualization

- To explore the data, catplot were used to see how Flight Number and Payload , Flight Number and Launch Site would affect the launch outcome.
- Scatter plot were used to show the relationship between
  - Flight Numbers and Launch Sites
  - Payload and Launch Site
  - Flight Number and Orbit type
  - Payload and Orbit type
- Bar chart were used to show the success rate of each orbit type
- Line chart were drawn to show the average success rate across the year
- GitHub URL for data visualization notebook

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA;
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- GitHub URL for EDA with SQL notebook

Build an Interactive Map with Folium

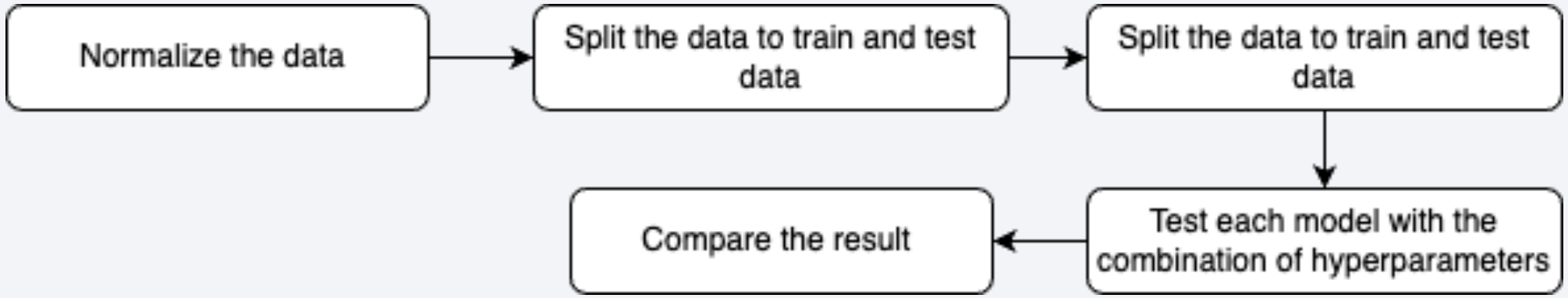
- I took the Latitude and Longitude of each sites and added Circle to mark around each site and Marker to add each launch site name.
- For each launch site, not only one launch occurs and we want to know each launch is success or fail. For these cases, I utilized the marker cluster to create the marker for all launch records.
- Also, we want to know the launch sites are how far from the coastline, city, highway and railway. For this problem, I utilized Haversine's formula to calculate the distance and used poly line to indicate the distance between two coordinates.
- GitHub URL of Folium map

Build a Dashboard with Plotly Dash

- In the interactive dashboard, I utilized these to visualize the data.
- The pie chart was used to visualize the percentage of success rate of launch sites.
- The scatter plot was used to visualize the correlation between launch site and payload mass.
- This combination allowed to quickly analyze which site has better success rate for payload mass and how much success for specifics payload mass kg.
- GitHub URL for dashboard app code

Predictive Analysis (Classification)

Four classification models, logistic regression, support vector machine, decision tree and k nearest neighbors were utilized to fit the data and compare the accuracy to determine what model is the best model to use .



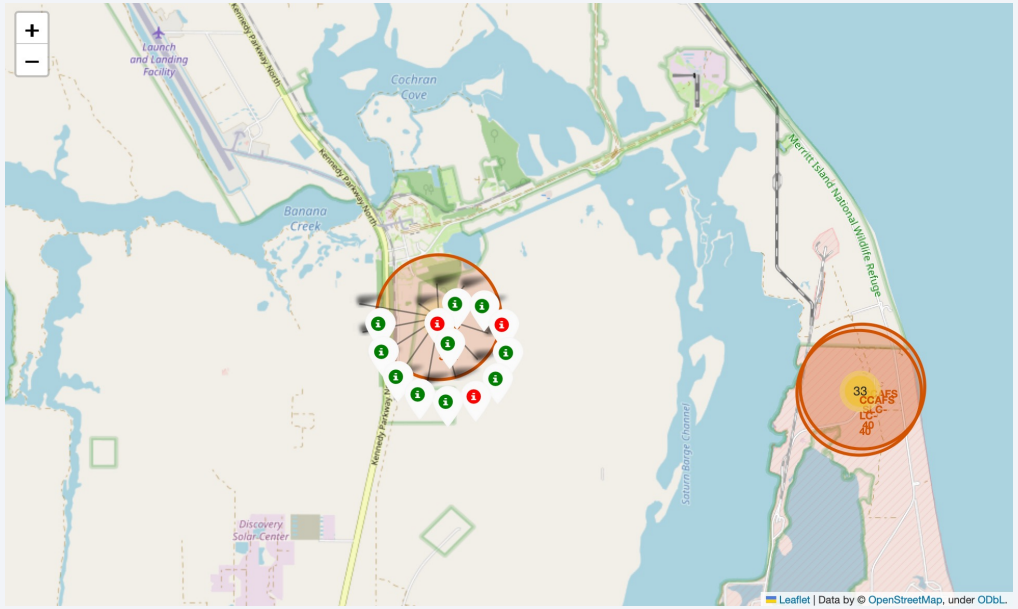
- GitHub URL for source code

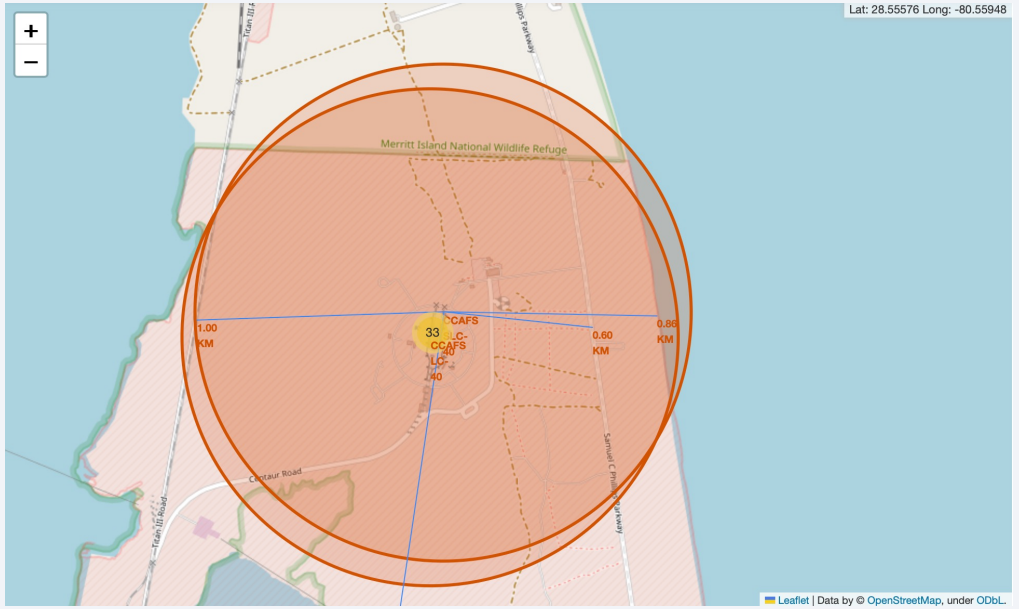
Results

- Exploratory data analysis results
  - SpaceX uses four different launch sites to launch their Falcon 9 rockets.
  - At the 2020, 45596 KG of payload mass has been carried by boosters launched at NASA (CRS)
  - First landing outcome in ground pad was achieved at 2015-12-22.
  - Almost 100% success mission outcome for booster Falcon 9 version.
  - There is only two failure landing in drone ship in 2015.
  - Success rate for launching has been increased steadily from 2013 and get the highest success rate at 2019, reaching 90%.

Results

Interactive analytics demo in screenshots

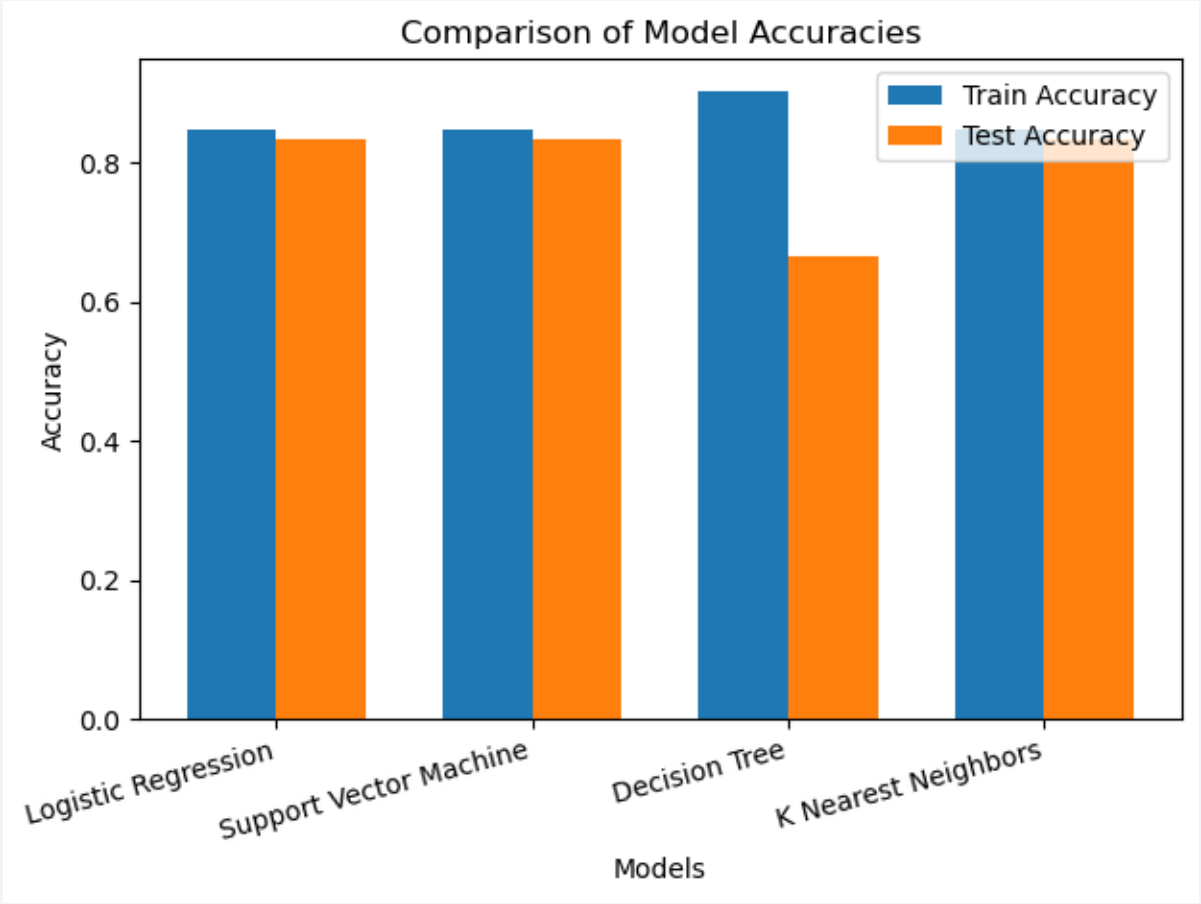


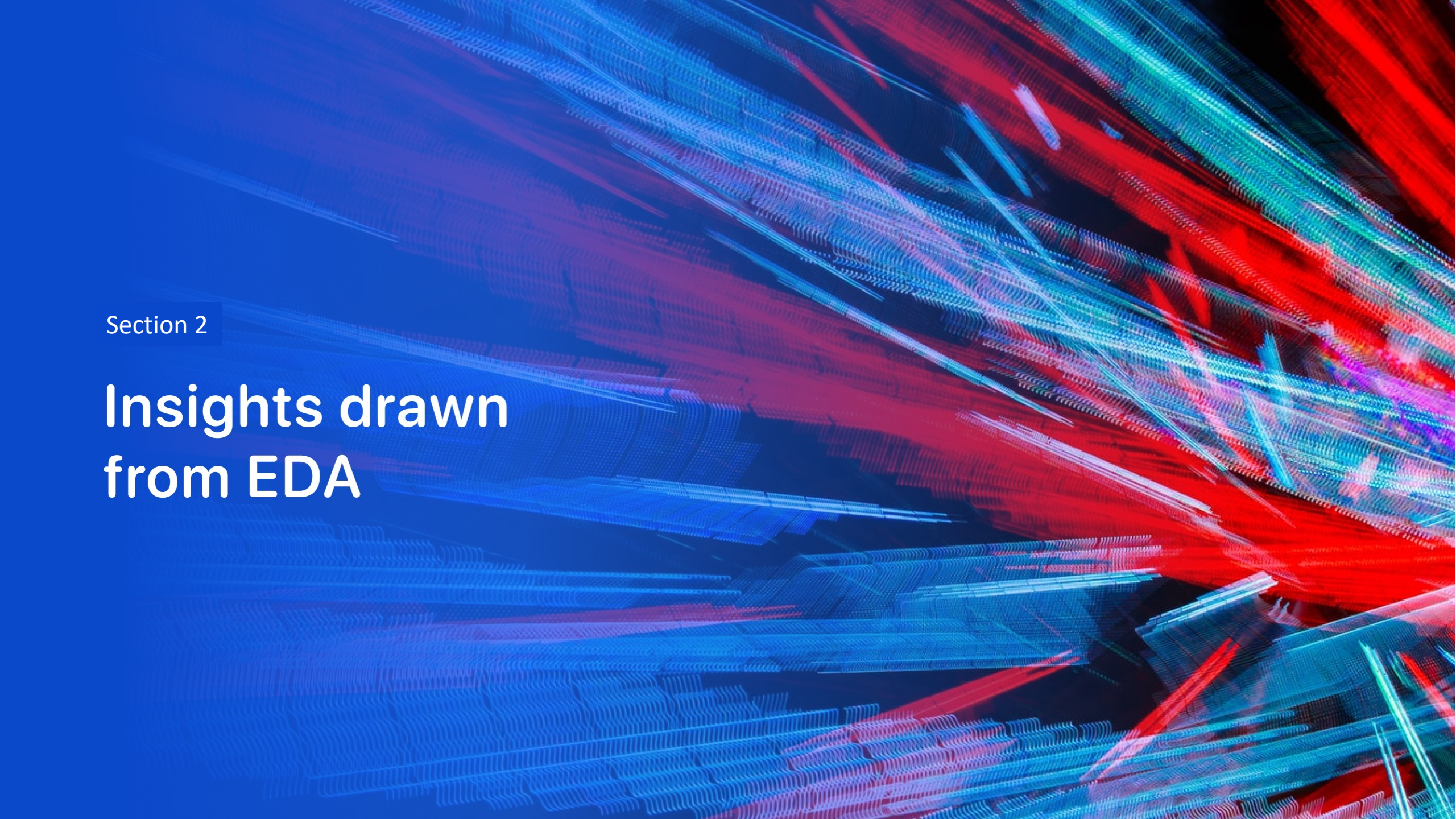


By utilizing interactive map can help identify which launch site has the best launch outcome, how far away from the coast, railway, highway and cities for safety measure.

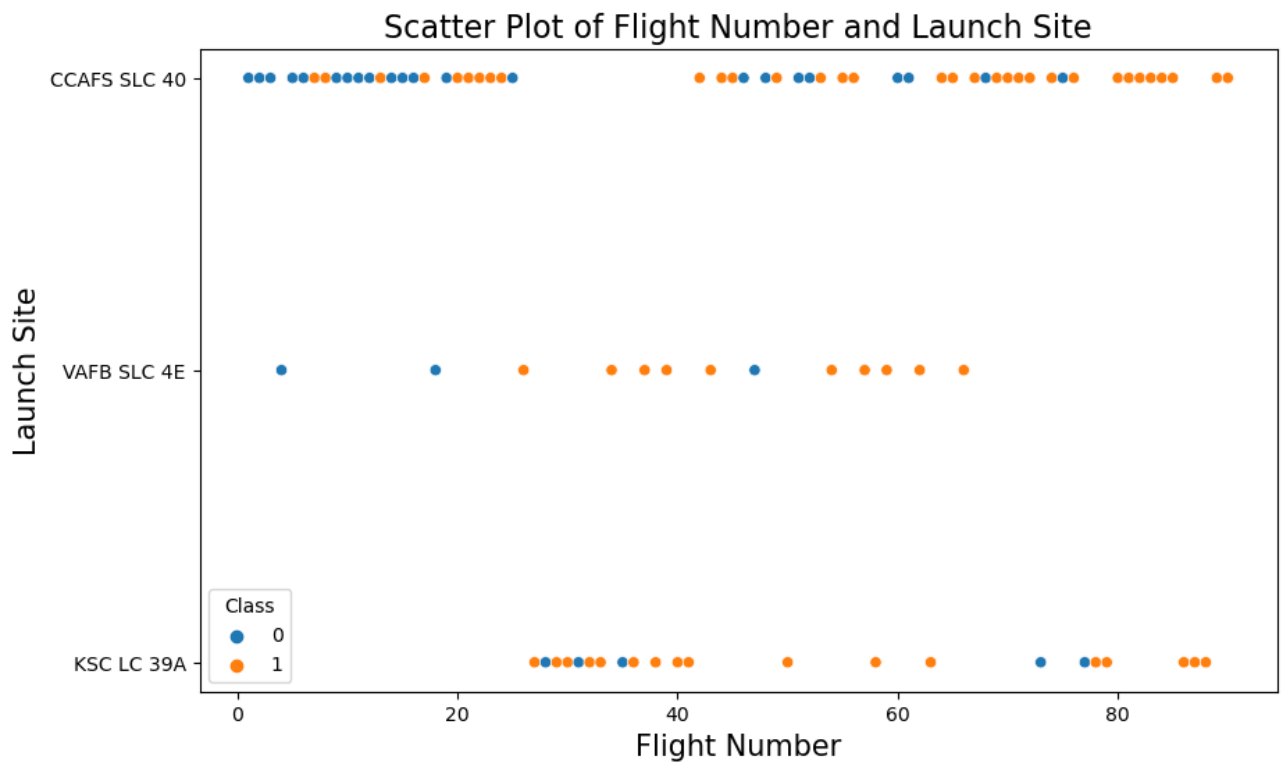
Results

- Predictive analysis results show that there is no a lot of different between the four classification models but decision tree test accuracy is significantly lower than other three models.
- The train and test accuracy for other three models looks the same because the dataset that used to train and predict is a small dataset.




Section 2
Insights drawn from EDA

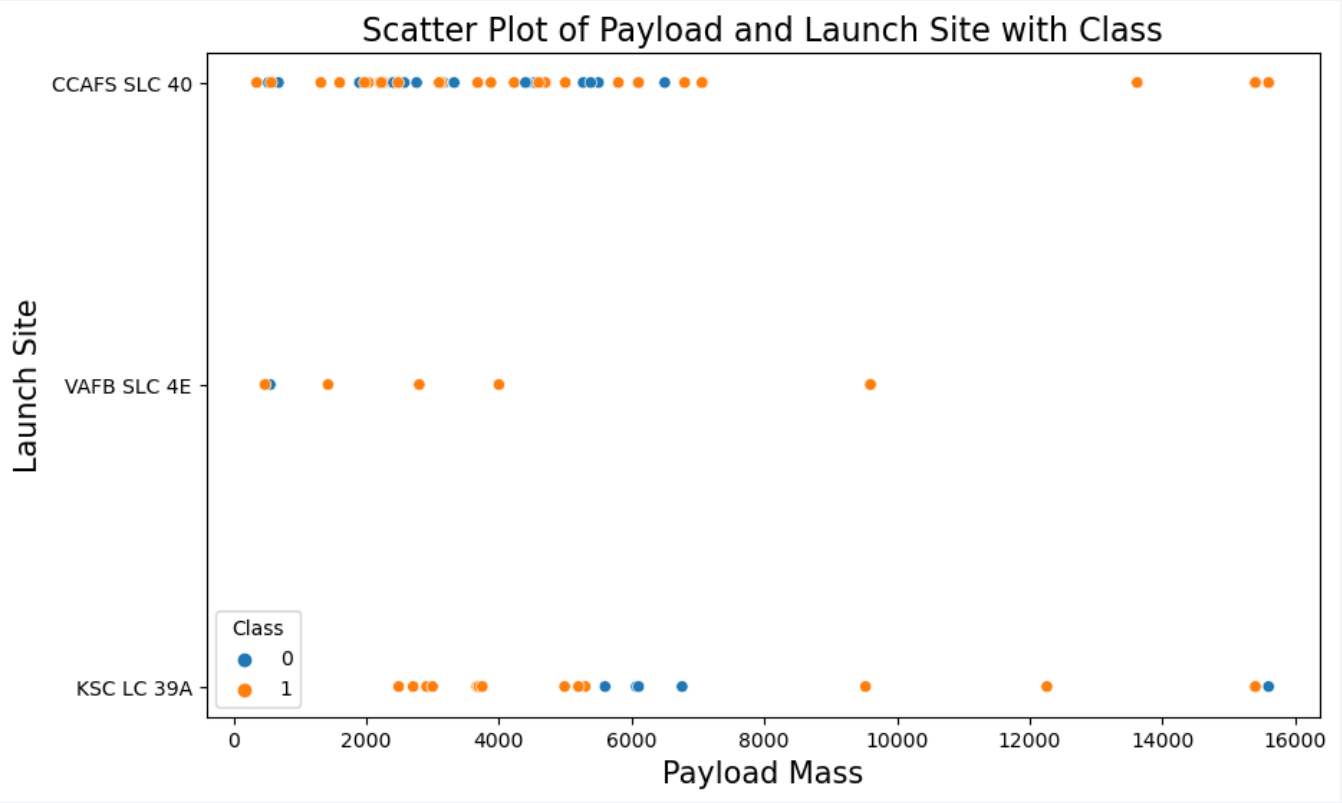
Flight Number vs. Launch Site



- CCAFS SLC 40 and KSC LC 39A has 100% success rate for above 80 Flight Number which means the more they launch the more their reliability in launching.
- Based on the scatterplot, CCAFS SLC 40 has more flight numbers than other two launch sites.

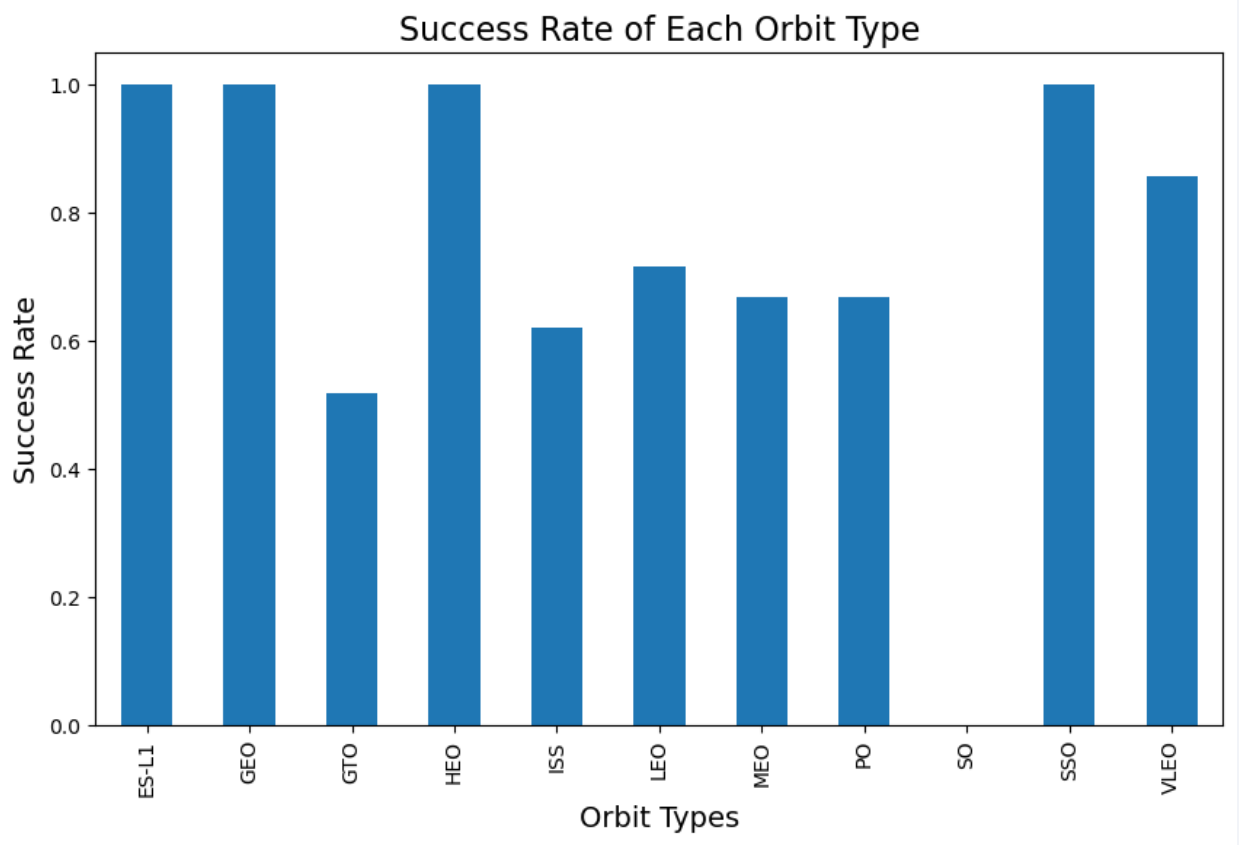
Payload vs. Launch Site

- Payload over 9000 KG has excellent success rate.
- For VAFB launch site, there is no data above 10000KG payload which means there is no rockets launched for heavy payload mass.



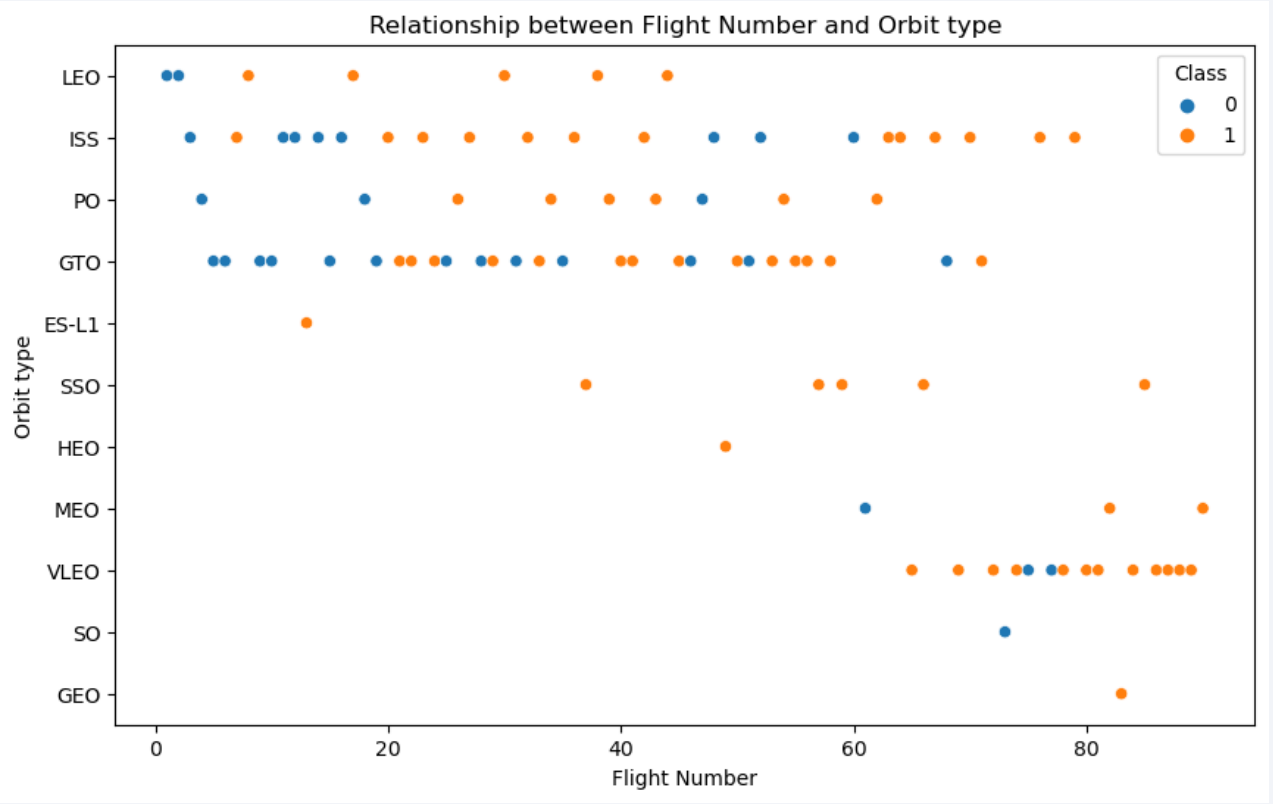
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbit types has perfect success rate.
- Interestingly, SO orbit has zero success rate.
- GTO has the lowest success rate beside SO with approximately 50%.

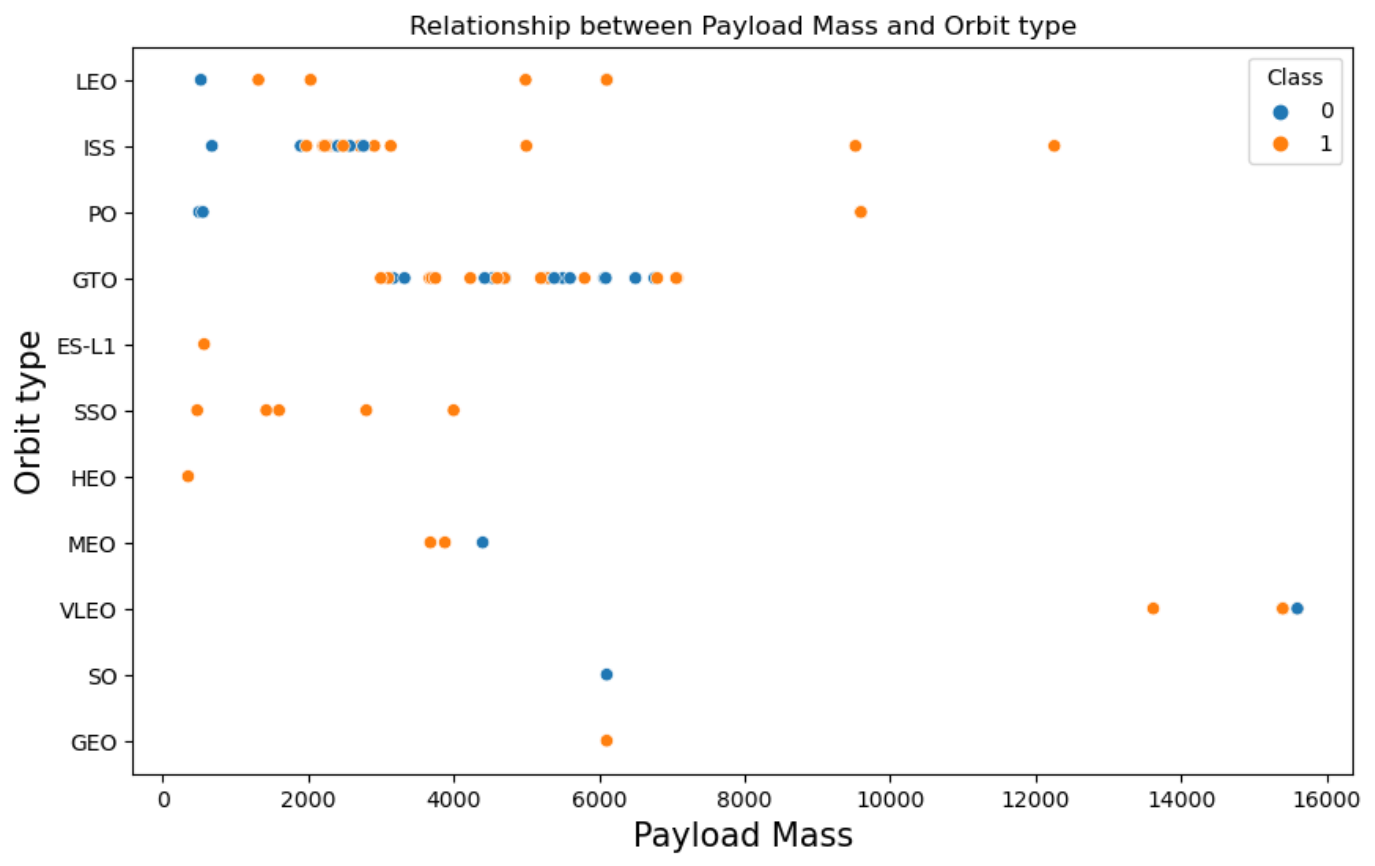


Flight Number vs. Orbit Type

- There is no specific relation between flight number and success rate in GTO orbit.
- In LEO orbit, as number of flight number increases the success rate also increase.
- VLEO orbit has the highest flight number in the last few years which means this orbit has new opportunity that we don't know.



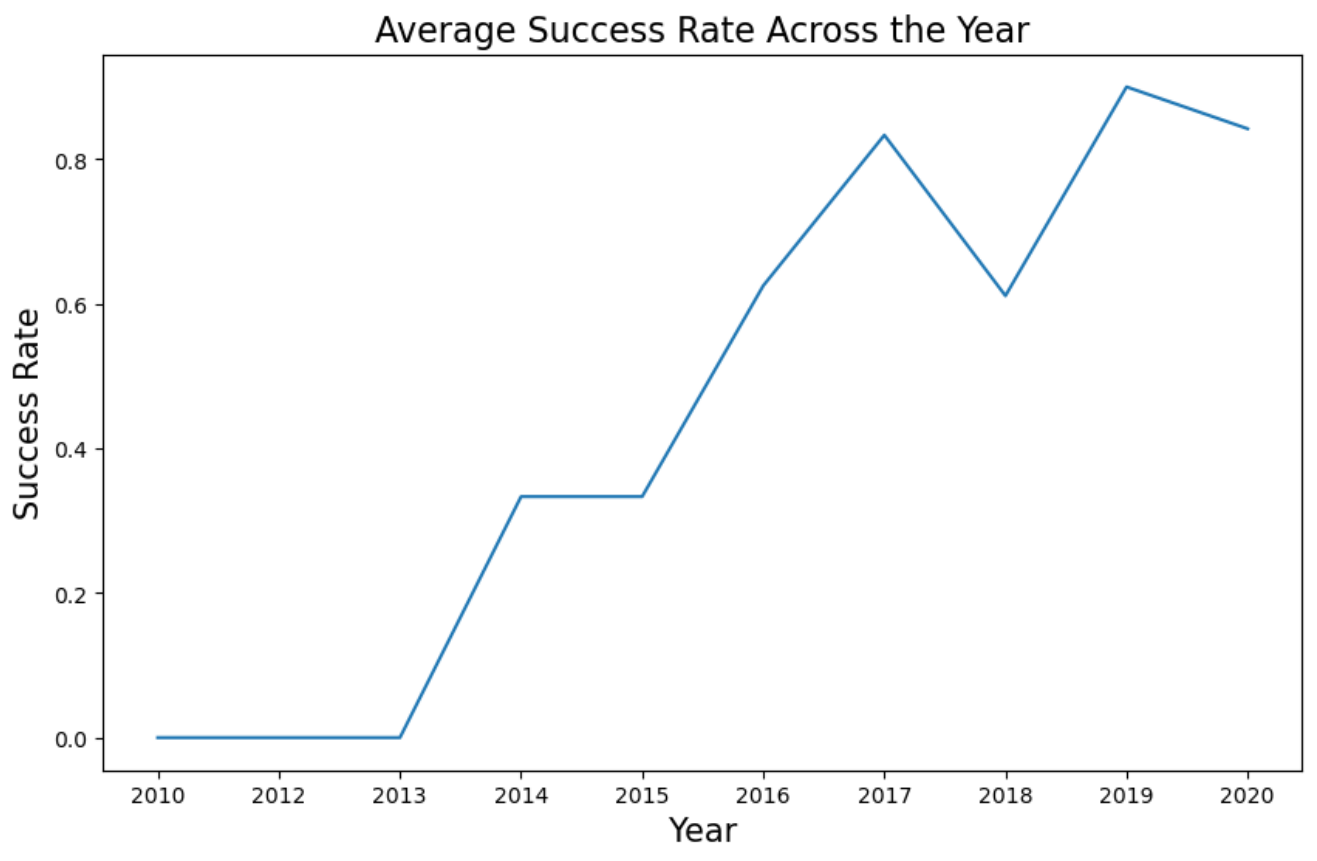
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, we can't distinguish as there is both success and fail landing are there.
- HEO orbit has the least payload mass carried launch.

Launch Success Yearly Trend

- From 2013 to 2020, the success rate exhibited a consistent upward trend.
- Notably, the year 2019 recorded the highest success rate within this period, reaching an impressive 90%



All Launch Site Names

- There are four different launch sites used by SpaceX.

SQL Query

```
%sql SELECT DISTINCT Launch_Site From SPACEXTABLE;
```

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- To get the unique launch sites, I utilized DISTINCT in the SQL query to the Launch Site column.

Launch Site Names Begin with 'CCA'

- FIVE records where launch sites begin with `CCA'

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

- For first five records, there is no success landing outcome as the data is in first three years where there is no success rate of landing outcomes.

Total Payload Mass

- Total payload mass carried by NASA at 2020 is 45596 KG. Total payload mass is calculated by summing all the payload mass where the customer is “NASA (CRS)”.

SQL Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

Result

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

SQL Query

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as AveragePayloadMass
FROM SPACEXTABLE
WHERE Booster_version = "F9 v1.1"
```

Result

AveragePayloadMass
2928.4

Average payload mass was calculated on the payload mass that was filtered by booster version is F9 v1.1 using WHERE clause in SQL query and get the average of the payload mass with AVG.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

SQL Query

```
%%sql
SELECT MIN(Date)
From SPACEXTABLE
Where Landing__Outcome like "%ground pad%"
AND Mission_Outcome = 'Success';
```

Result

MIN(Date)
2015-12-22

By filtering the data for landing outcome on ground pad and mission outcome as a success, we get the required data and utilizing min function to get the minimum value (which means first date) that give 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

SQL Query

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
AND LANDING__OUTCOME = 'Success (drone ship)';
```

Result

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

- Selecting only data with success landing on drone ship and had payload mass greater than 4000 but less than 6000 and utilizing DISTINCT to know the unique booster version.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcome:

SQL Query

```
%%sql
SELECT Mission_Outcome, COUNT(*) as Total
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

Result

MISSION_OUTCOME	Total
Failure (in flight)	1
Success	99
Success(payload status unclear)	1

Grouping the mission outcome and count the records get the required result. Interestingly, there is only one failure which means almost 100% success rate.

Boosters Carried Maximum Payload

- Booster version which have carried the maximum payload mass:

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

BOOSTER_VERSION
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7
F9 B5 B1049.5

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

MONTH	BOOSTER_VERSION	LAUNCH_SITE	LANDING__OUTCOME
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

I utilize substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year because SQLite doesn't support the month names. The outcome list has only two occurrences in 2015.

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

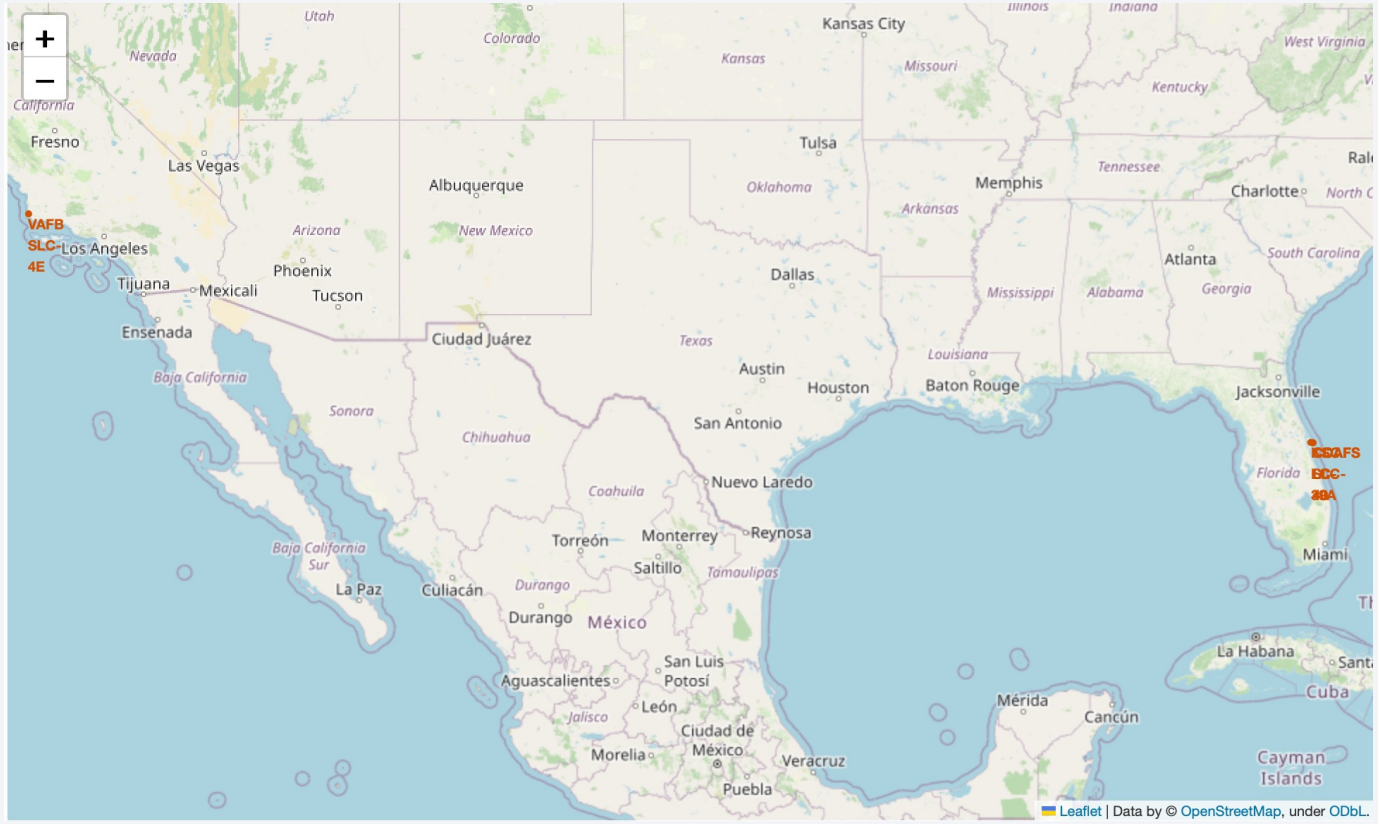
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

LANDING__OUTCOME	count
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

- This attempt shows us that the missing value in the Landing Outcome is important and need to taken in account.

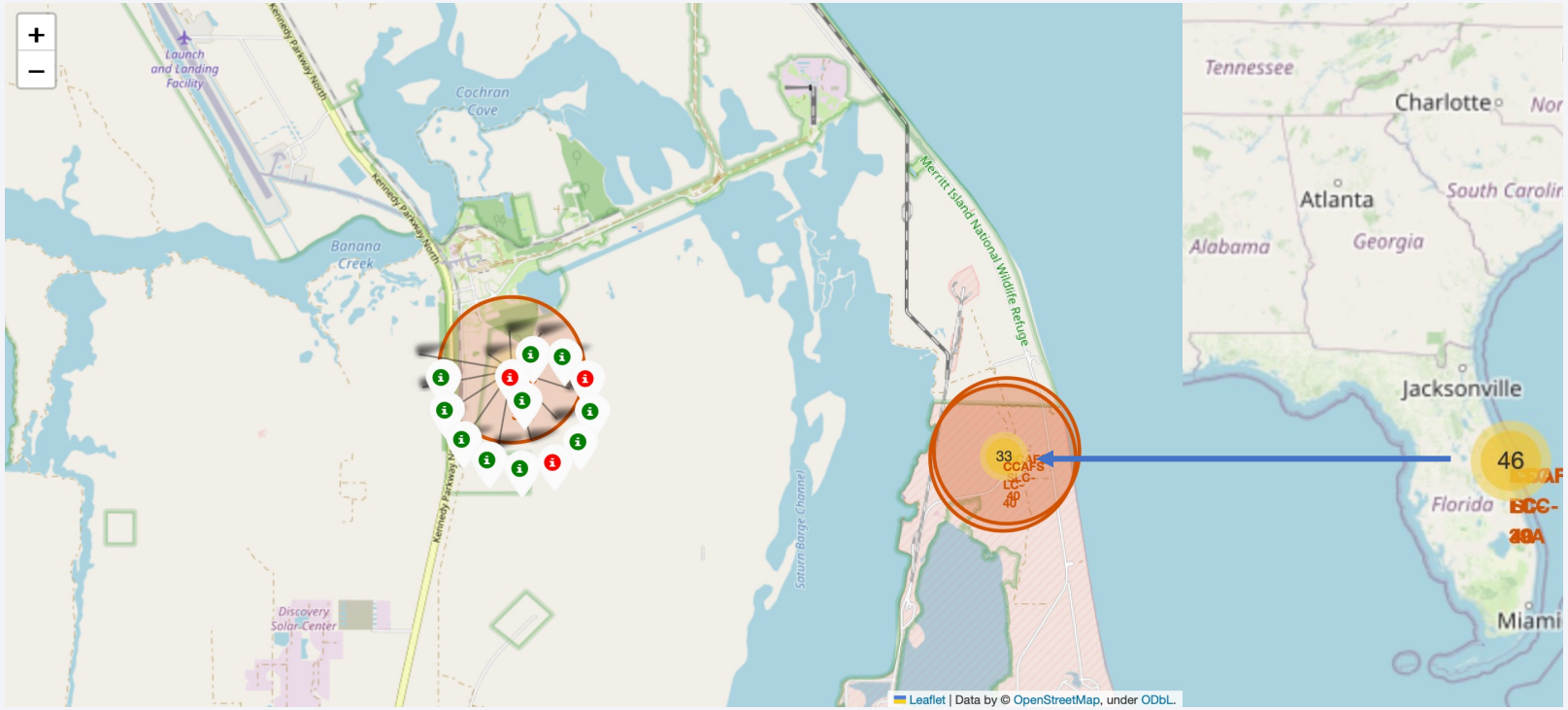

Section 3
Launch Sites
Proximities Analysis

ALL Launch Sites



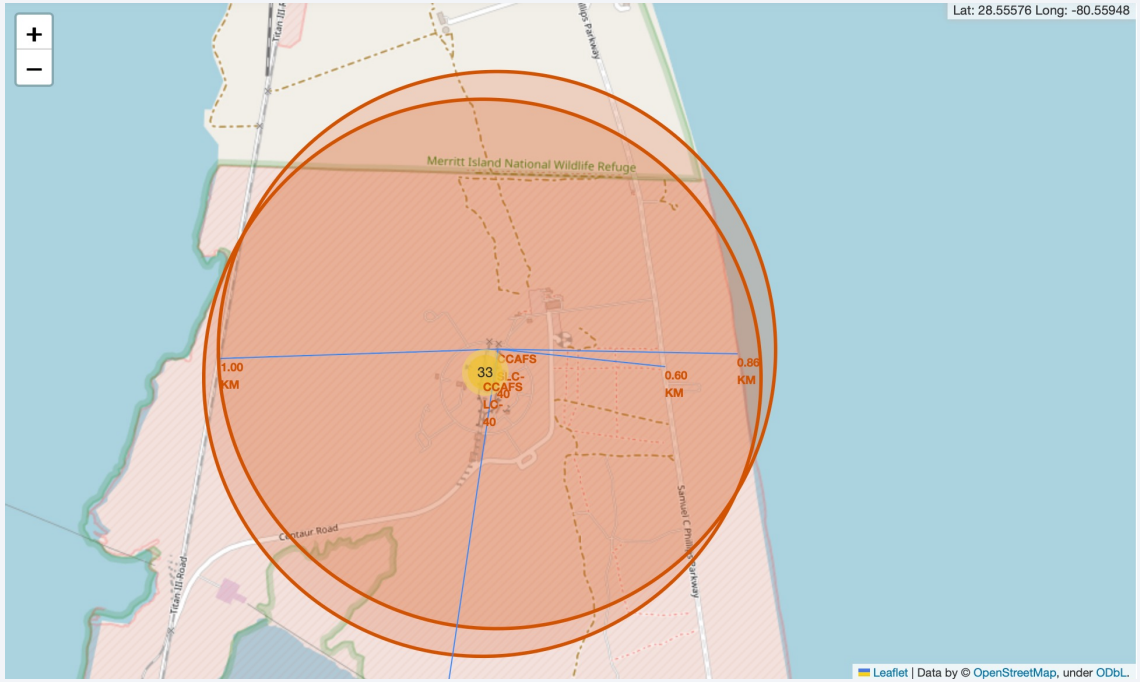
All Launch sites were marked by circle and marker in the folium map. By watching the map, all launch sites are near the coast and not too far away from the road.

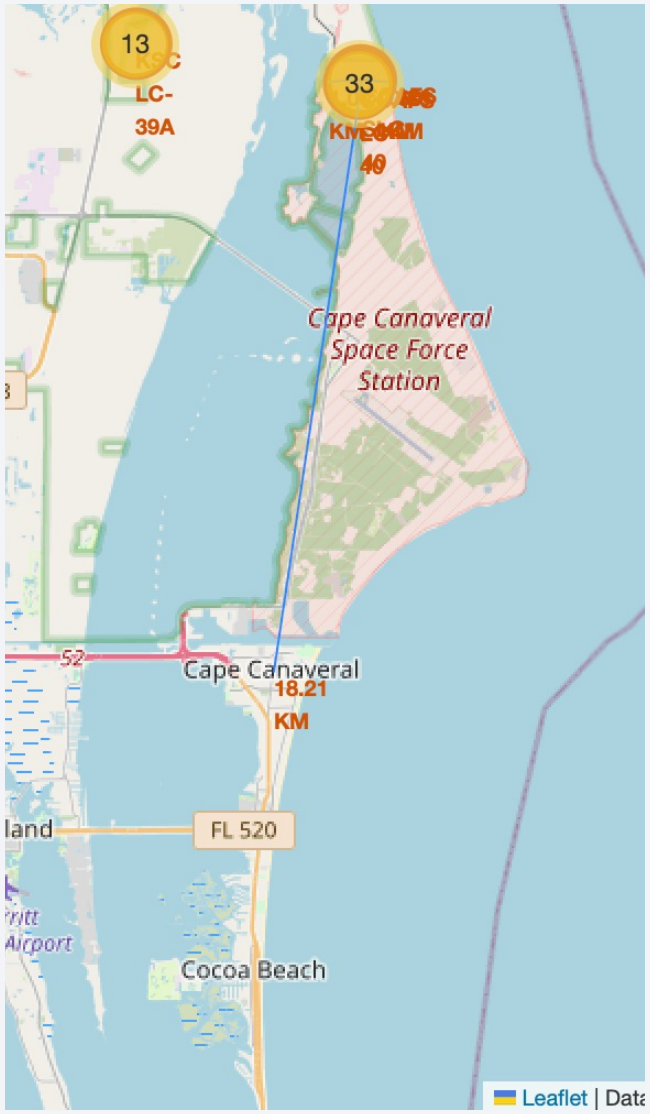
Colored Labeled Markers for each launch



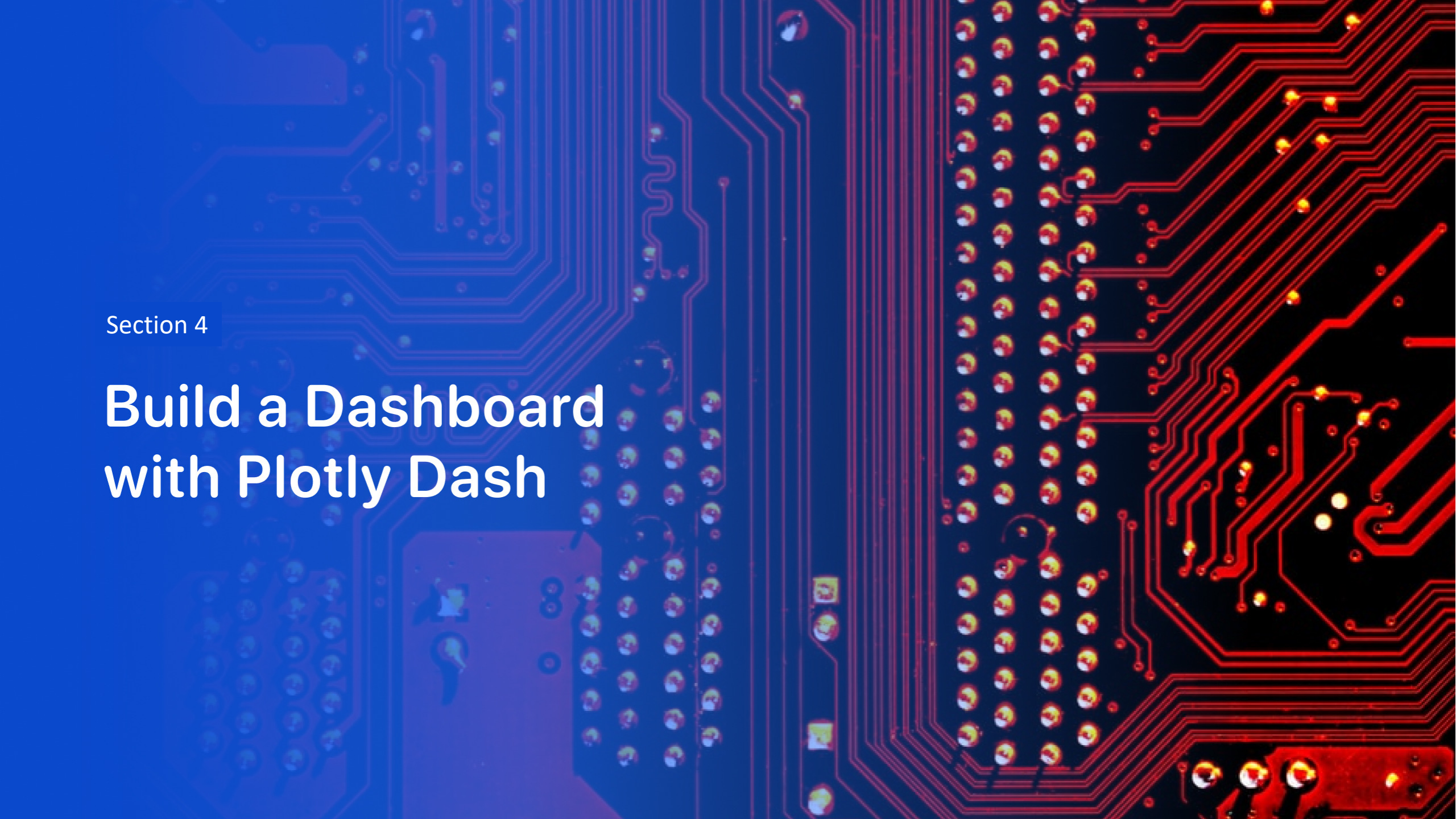
I create the markers for all launches. Green markers indicate successful and red ones indicate failure. By analyzing the colored markers, you can easily identify which sites has relatively high success rate.

Distance from launch site to landmarks

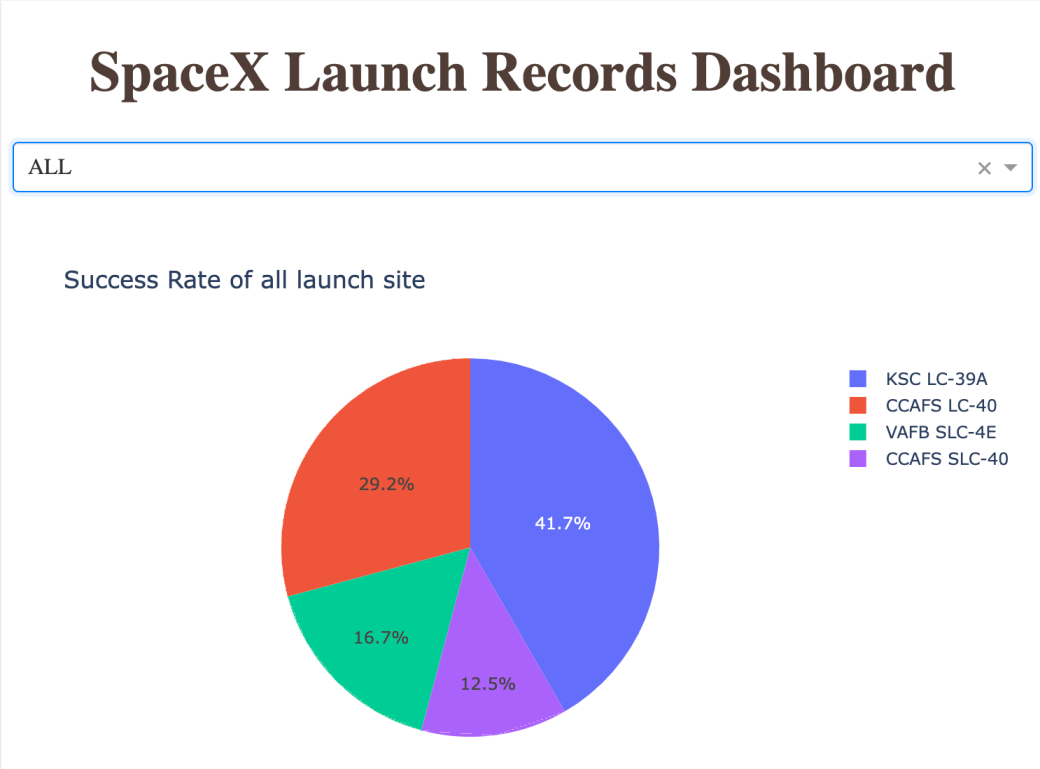




- Distance between launch site and landmarks are calculated with Haversine formula.
- After drawing distance lines to nearby landmarks, I found the city is far away from the launch site which makes the safe operational environment.

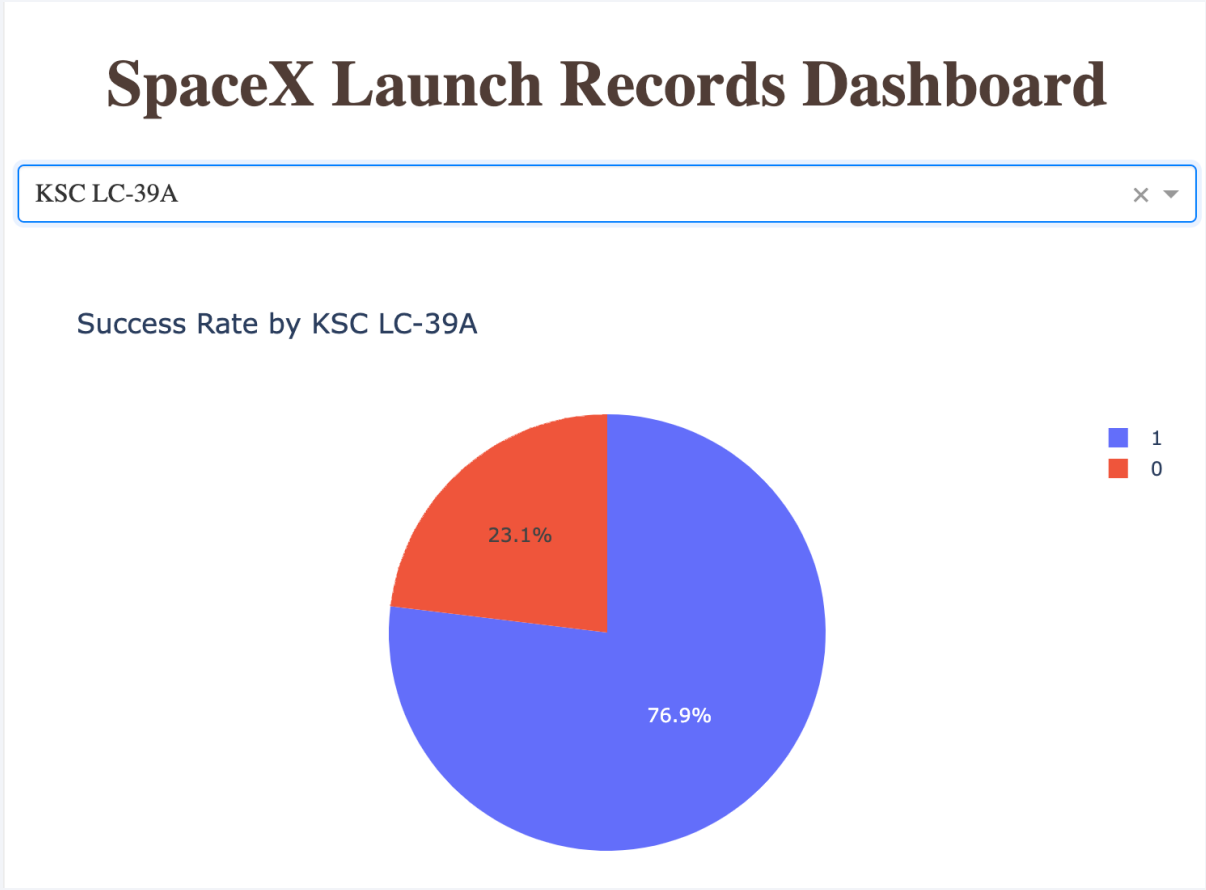

Section 4
Build a Dashboard with Plotly Dash

SpaceX Dashboard: Success rates for all launch sites



- We can see the success rate of launches by all sites.
- KSC LC-39A has the highest success rate with 41.7%.

SpaceX Dashboard: Success Rate for KSC LC-39A

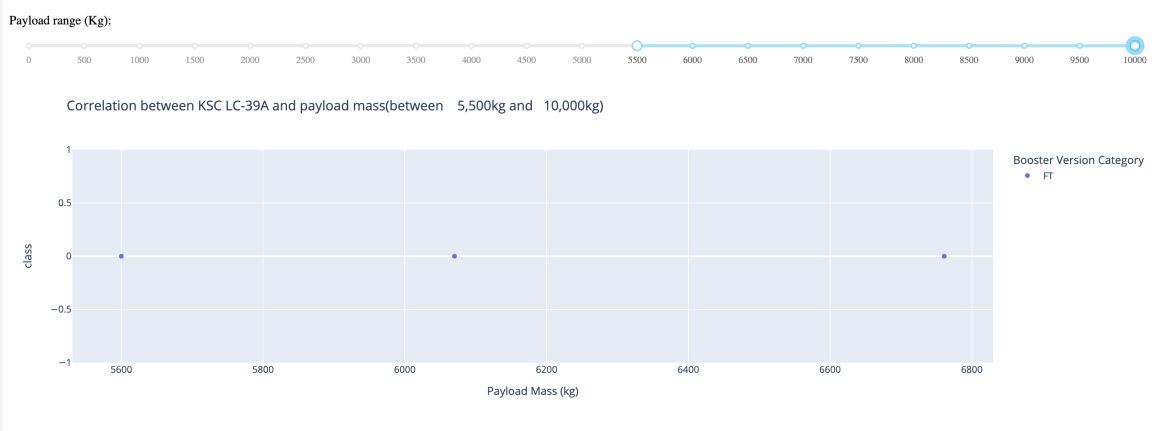


- KSC LC-39A has achieved a 76.9% success rate and 23.1% failure rate.

SpaceX Dashboard: Payload vs Launch Outcome

Low Payload Mass between 0kg and 5000 kg

High Payload Mass between 5000kg and 10000 kg



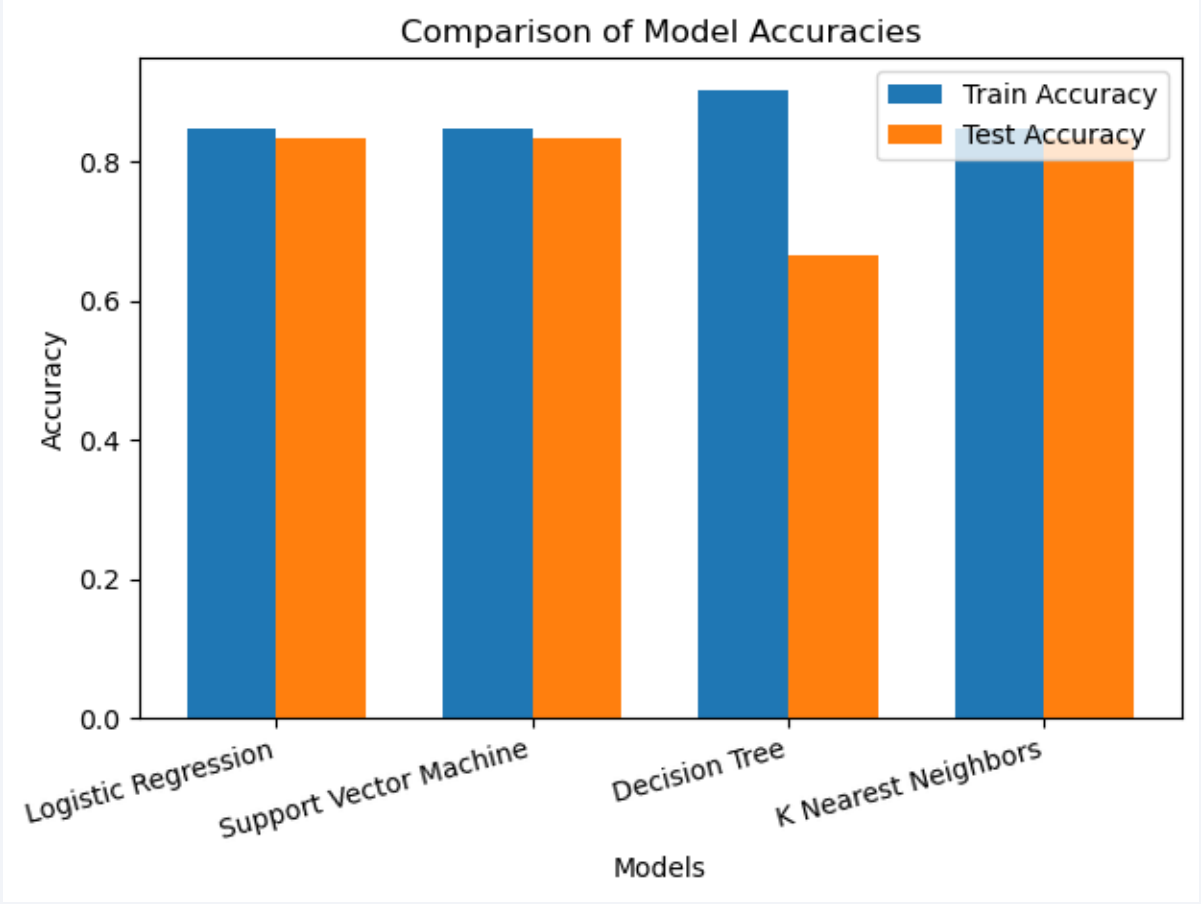
- KSC LC-39C has 100 percentage success rate for low payload mass between 0kg and 5000 kg.
- Interestingly, for high payload mass between 5000kg and 10000kg, it didn't have zero percentage success rate.

Section 5

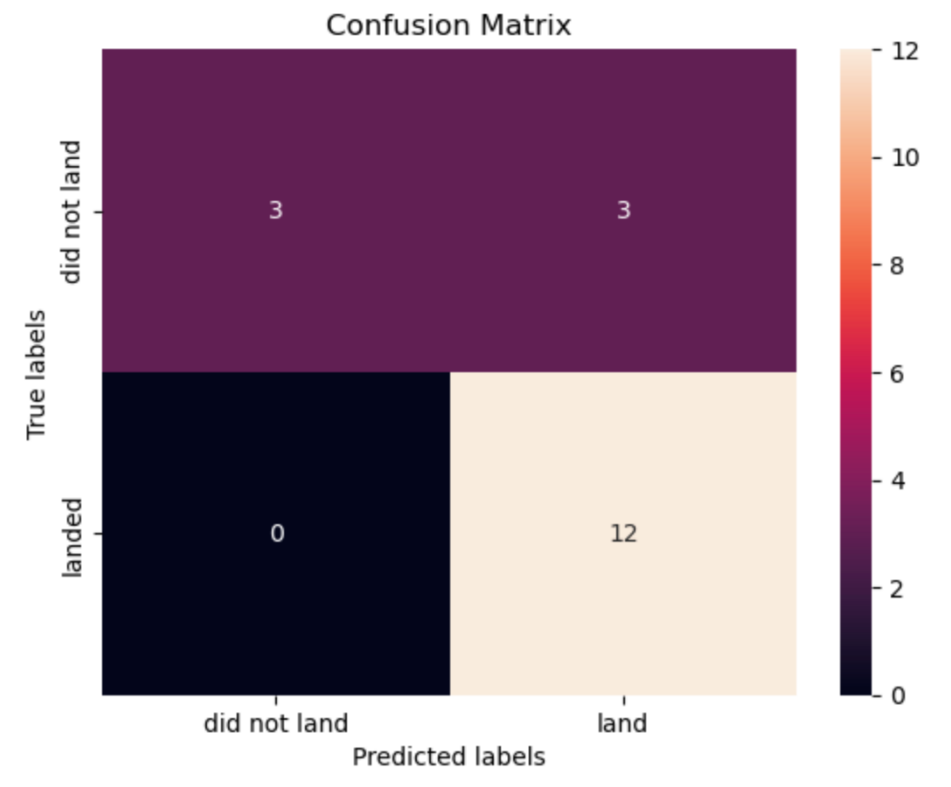
Predictive Analysis (Classification)

Classification Accuracy

- Four different classification were used to predict and their accuracies are shown besides.
- Decision Tree Test accuracy has significantly lower than other three models.
- Other three models produce same train and test accuracy because of relatively small dataset.



Confusion Matrix



- I choose the logistic regression model to show the confusion matrix.
- By the result of confusion matrix, the logistic regression can predict the true positive effectively but there is a problem for false positive.

Conclusions

- By manipulating and wrangling data, we got the information that we needed.
- To compete with SpaceX, the best launch site to launch is KSC LC-39A.
- Launches below low payload mass under 5000 kg has less risk than high payload mass above 5000kg.
- As success rate is directly proportional to the year, SpaceX will perfect the landing over the year.
- Logistic Regression can be used to predict the successful landings because it never miss the successful landing and can increase profit by reusing the first stage again.

Thank you!