IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daichi Azumi
June 2, 2023

Outline

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

Executive Summary

- Summary of methodologies

  The SpaceX launch data were collected with SpaceX API and web scraping.
  Next, data wrangling performed to it.
  After that, EDA is conducted with visualization tools such as pandas, matplotlib, and with SQL.
  Then, we explore and manipulate data in an interactive way with Folium and Plotly Dash.
  Finally, predictive analysis was performed using classification models such as logistic regression, SVM, decision tree classifier, and KNN.

- Summary of all results

  From EDA, the larger the payload mass and flight number, the higher the success rate. The success rate varies depending on orbit types. The types of booster versions are determined by the payload mass.
  Interactive analysis showed that flights at KSC LC-39A have the highest success rate. In the 2000-6000 kg payload range, most launches with booster version FT are successful.
  The results of the predictive analysis showed that the decision tree model performed with the highest accuracy.

Introduction

- Project background and context

  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.

- Problems you want to find answers

  What factors contribute to a successful launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API Calls, Web scraping
- Perform data wrangling
  - Perform a simple EDA and create a landing outcome column for predictive 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
  - Find hyperparameters by grid search and apply them to each classification models (Logistic Regression, SVM, Decision Tree, and KNN), then output confusion matrix

Data Collection

I performed data collection in two ways; SpaceX API and web scraping.

- SpaceX API

  SpaceX advertises Falcon 9 rocket launches on its website. Thus, the first way to obtain the dataset is to use the API.

- Web scraping

  We can see Falcon 9 historical launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches". By performing web scraping we can also obtain the data.

Data Collection – SpaceX API

It is possible to obtain the launch data by making a get request to the SpaceX API.

The flowchart on the right describes the SpaceX API calls.

Finally, the data were obtained in the form of data frames.

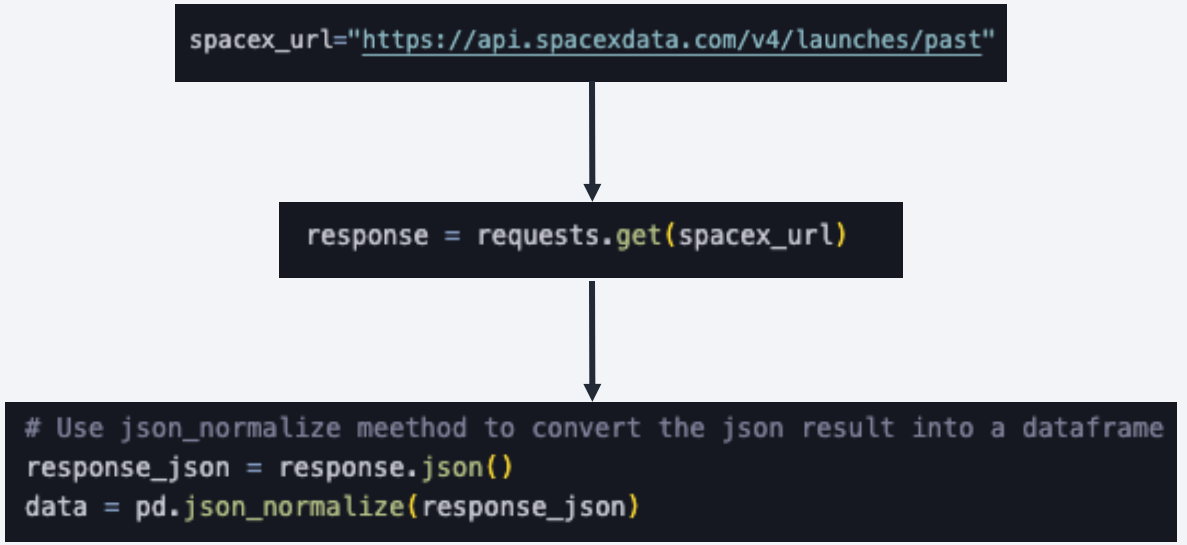


Figure: Flowchart of API calls

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_data_collection_api.ipynb

Data Collection - Scraping

Web scraping is another option for obtaining the launch data.

Here, Falcon 9 launch data in the form of HTML tables were extracted from a Wikipedia page by using the Python BeautifulSoup package.

Finally, the data in the form of HTML were converted into the form of data frames.

Figure: Flowchart of web scraping

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_webscraping.ipynb

Data Wrangling

Data wrangling was performed according to the process shown in the flowchart on the right.

First of all, to familiarize myself with the dataset I performed EDA.

Based on some results in EDA, labels were selected for use in predictive analysis.

In addition, target labels were converted into integer values for numerical computation.

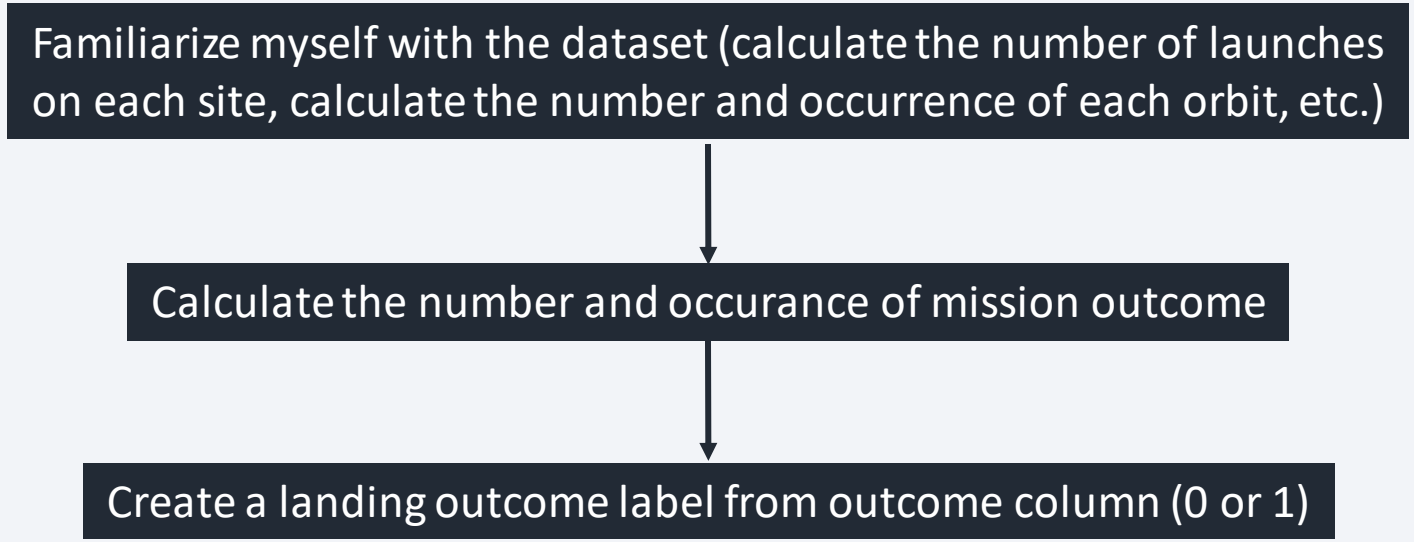


Figure: Flowchart of data wrangling

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_data_wrangling.ipynb

EDA with Data Visualization

Using Pandas and matplotlib libraries, EDA and feature engineering were performed.

Some charts were plotted in order to know the relationship between each variable and its impact on the target variable. For example;

- a scatter plot of Flight Number vs. Launch Site.
- a scatter plot of Payload vs. Launch Site
- a bar chart for the success rate of each orbit type
- a scatter plot of Flight Number vs. Orbit Type
- a scatter plot of Payload vs. Orbit Type
- a line chart of yearly average success rate

Then, influential features were selected and one-hot encoding was performed.

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_eda_data_visualization.ipynb

EDA with SQL

EDA with SQL was also performed.

Here are some examples of queries that were performed;

- 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)

There are still other queries that were executed. Those are introduced later.

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_eda_sql.ipynb

Build an Interactive Map with Folium

The successful launch may depend on the location and proximity of a launch site, so it can be valuable to analyze how influential the location or proximity is to the success rate.

By using the Python package folium, interactive visual analysis was performed.
For instance;

- mark all launch sites on a map
- Mark the success / failed launches for each site on the map
- Calculate the distances between a launch site and its proximities

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Interactive visual analytics enables users to explore and manipulate data in an interactive and real-time way.

The interactive dashboard was built with Plotly Dash to help users find visual patterns faster and more effectively.

The dashboard contains;

- input components such as a dropdown list and a range slider
- output components such as a pie chart and a scatter plot

The users can more easily see the relationship between flight sites, payload mass, booster versions, and success rate with it.

GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

The flowchart shows the process of predictive analysis.

In preprocessing, the data were standardized.
Then, split the data into a training set and a testing set.
After that, we trained the model and performed grid search to find hyperparameters.

Using hyperparameters we determined the model with the best accuracy, then output confusion matrix.

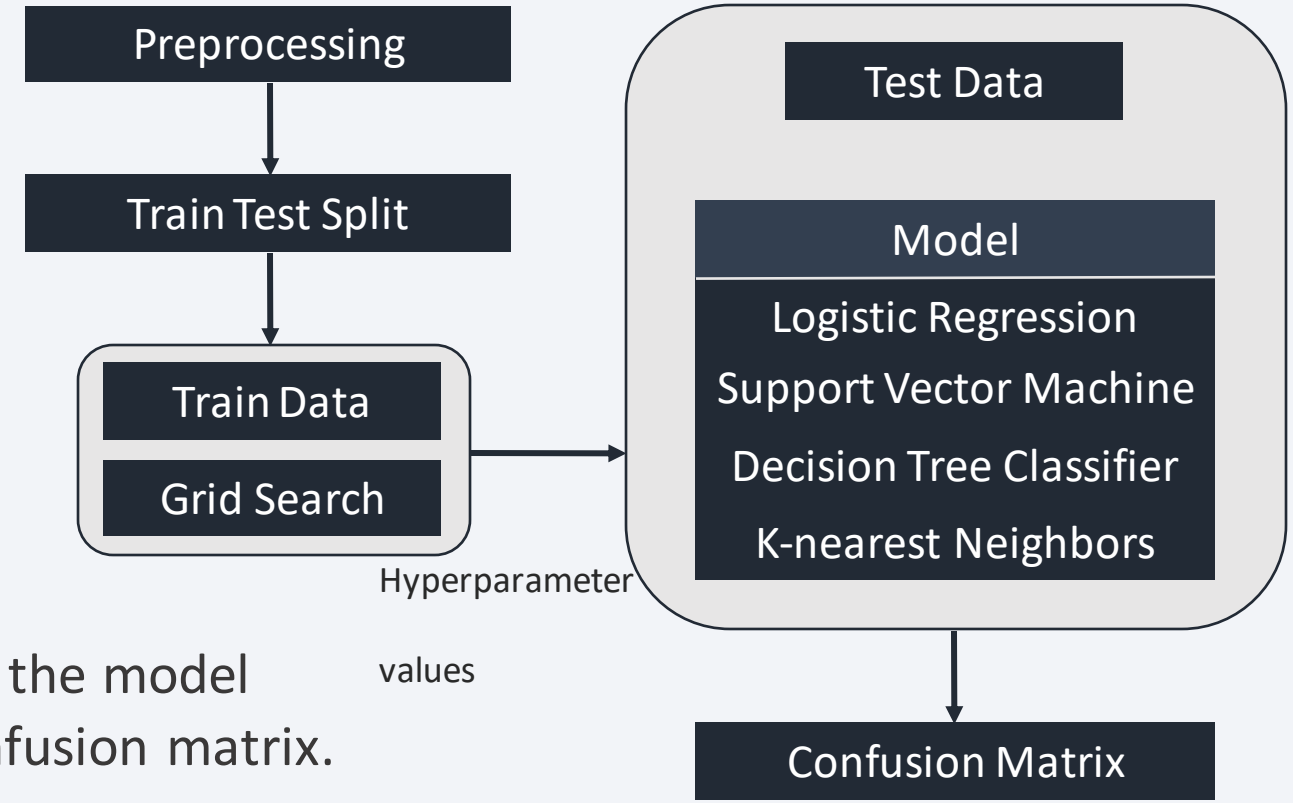


Figure: Flowchart of predictive analysis

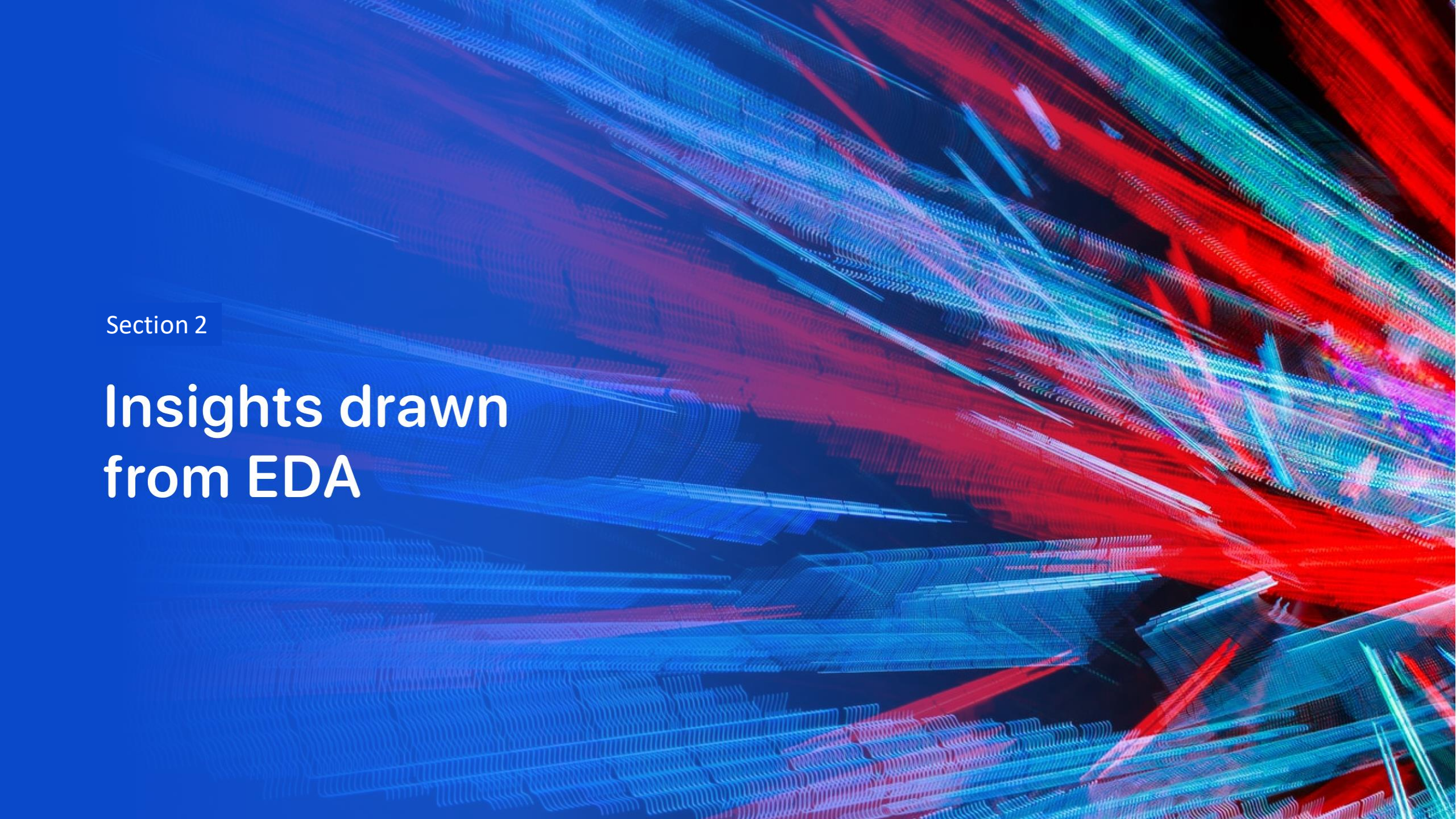
GitHub URL:
https://github.com/daichi-0909/coursera-capstone/blob/main/spacex_machine_learning_prediction.ipynb

Results

- EDA with data visualization
  The larger the payload mass, the higher the success rate.
  The larger the flight number, the higher the success rate. Related to this, the success rate continues to increase year after year.
  The success rate varies depending on orbit types.
- EDA with SQL
  The types of booster versions are determined by the payload mass.
- Interactive analytics with Folium and Plotly Dash
  Of the total number of successes, the success in KSC LC-39A is the largest at about 40%, and the success rate in KSC LC-39A is about 75%.
  Most successful launches were in the 2000-6000 kg payload range. The largest percentage of booster versions used in that range was FT.
- Predictive analysis
  The decision tree model performed with the highest accuracy.

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

The graph shows a scatter plot of Flight Number vs. Launch Site.

Overall, a success rate tends to increase as the number of flights increases.

This means that SpaceX excels in its ability to learn from its mistakes and succeed.

Flights in CCAFS SLC 40 have failed more frequently than flights at other sites, which can be attributed to the fact that most of the early flights were attempted there.

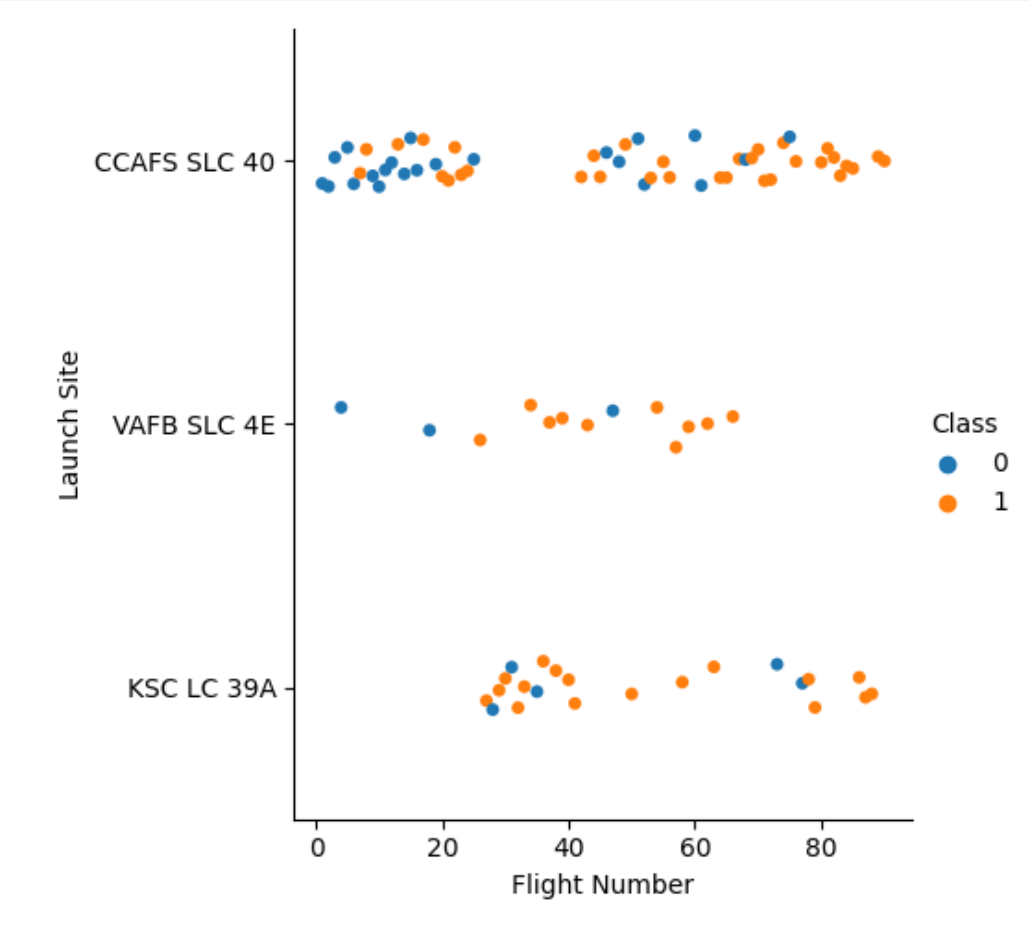


Figure: The scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site

The graph shows a scatter plot of Payload vs. Launch Site.

Overall, a success rate tends to increase as the pay load mass increases.

However, when it comes to flights in CCAFS SLC 40, it is difficult to say that there is a dominant relationship between payload and success rate, especially in the range of up to 7500 kg.

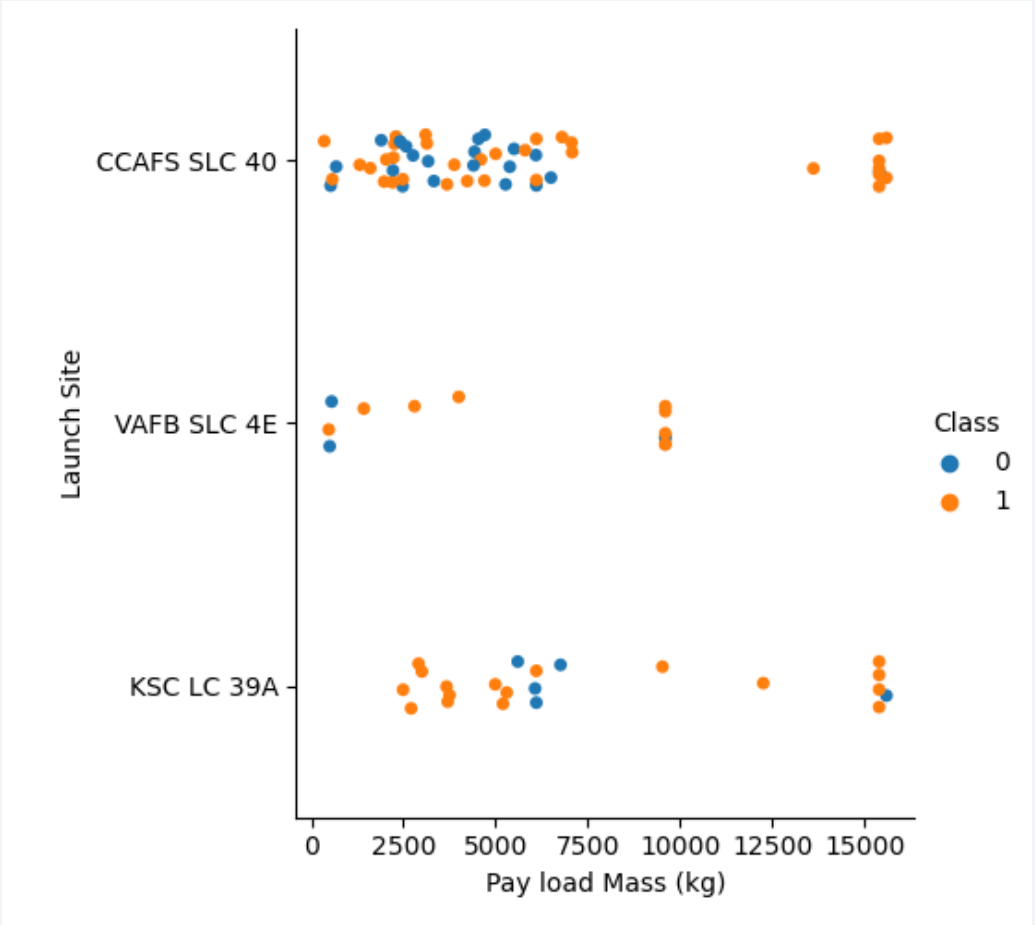


Figure: The scatter plot of Payload vs. Launch Site

Success Rate vs. Orbit Type

The graph shows a bar chart for the success rate of each orbit type.

The success rate differs depending on the orbit type, suggesting a relationship between two.

For orbit types with extreme success rates, such as 1 or 0, it is necessary to consider the background, for example, a small number of trials.

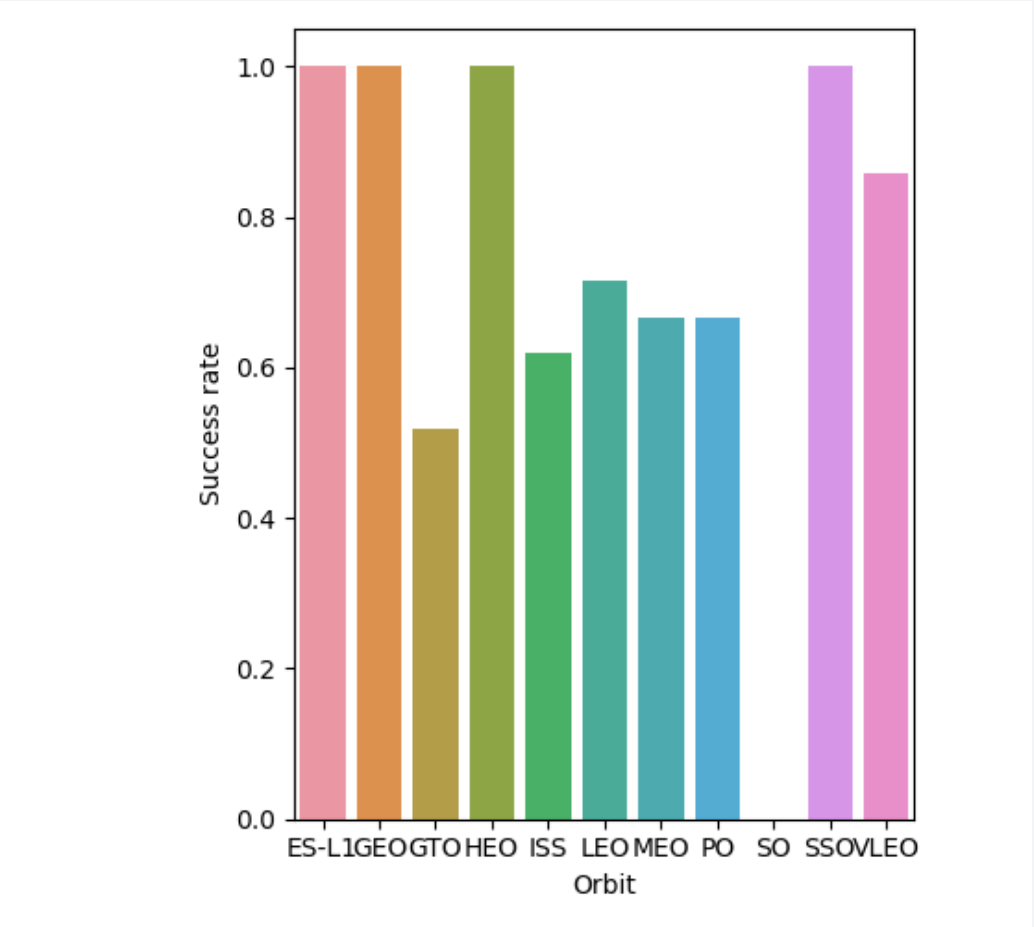


Figure: The bar chart of the success rate of each orbit type

Flight Number vs. Orbit Type

The graph shows a scatter plot of Flight Number vs. Orbit Type.

Overall, a success rate tends to increases as the number of flights increase.

For GTO, however, it is hard to find a relationship between the number of flights and the success rate.

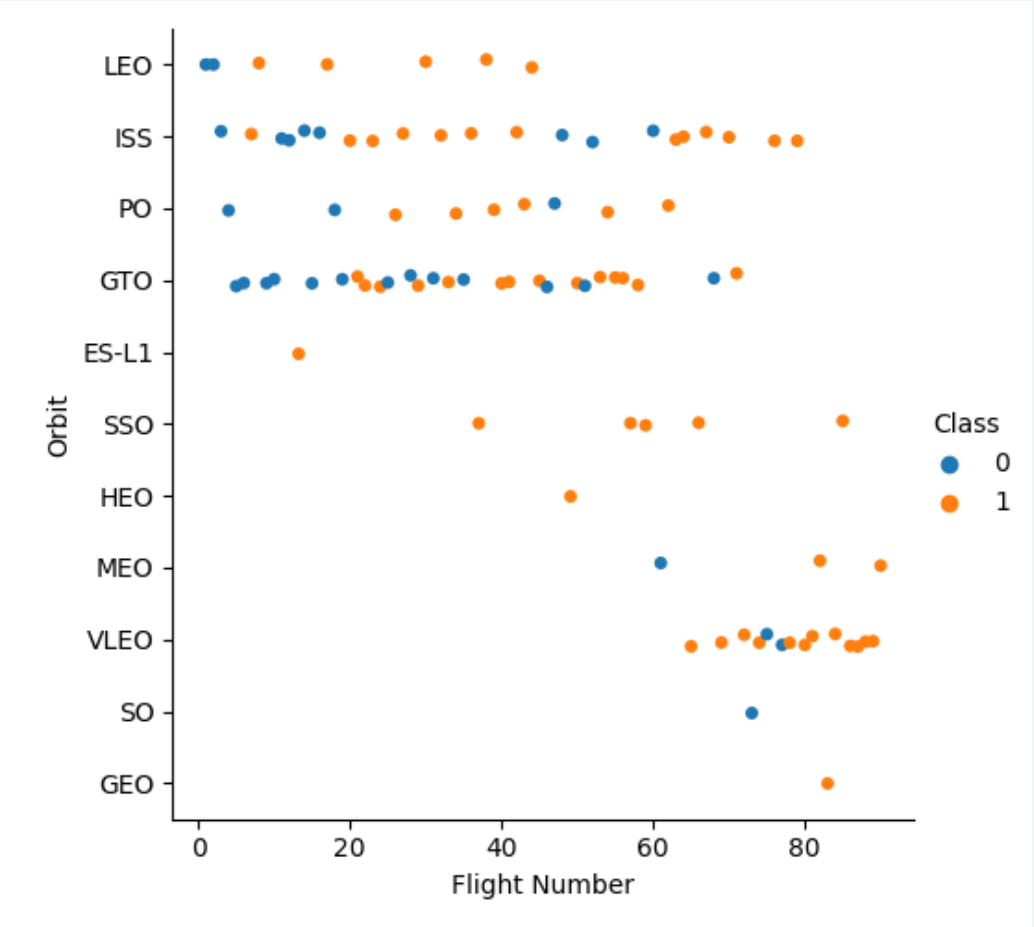


Figure: The scatter plot of Flight Number vs. Orbit type

Payload vs. Orbit Type

The graph shows a scatter plot of Payload vs. Orbit Type.

For LEO, ISS, and PO, it can be said that the success rate increases as the pay load mass increases.

For the rest, especially GTO, there seems to be no relationship between the two.

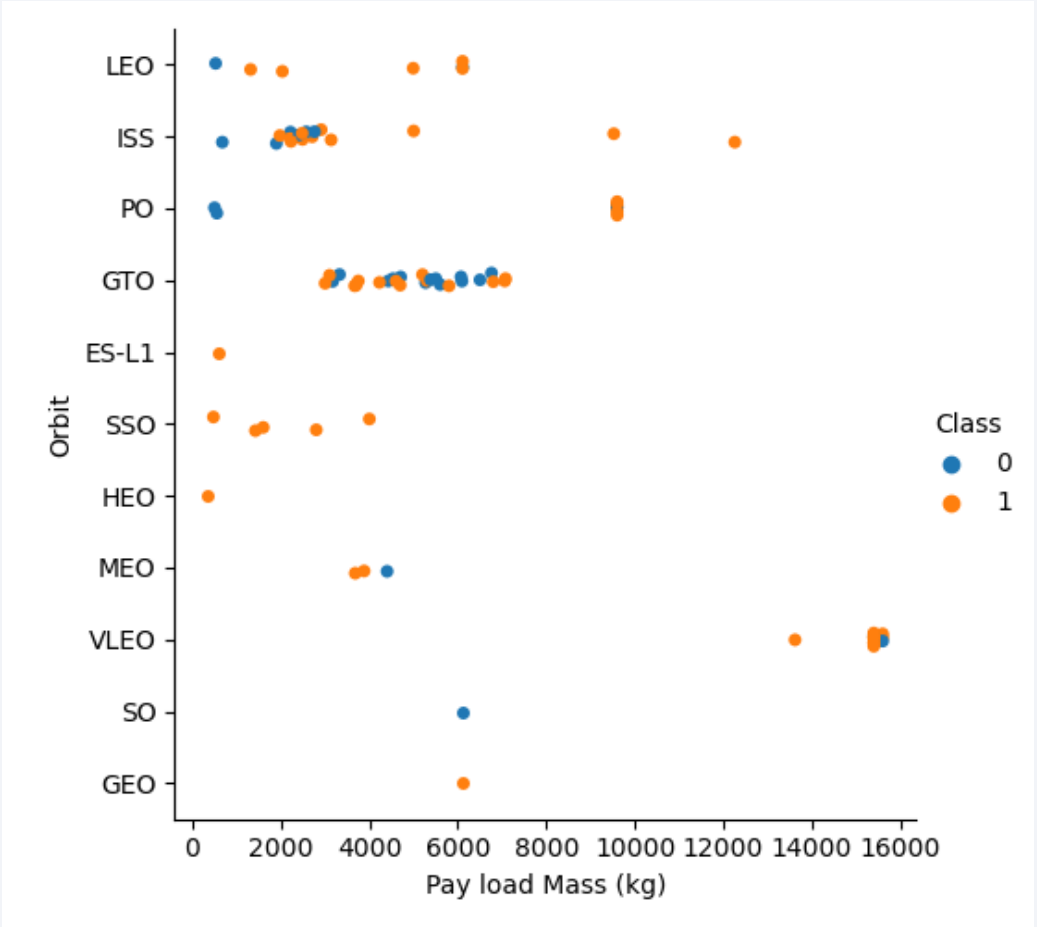


Figure: The scatter plot of Payload vs. Launch Site

Launch Success Yearly Trend

The graph shows a line chart of yearly average success rate.

From 2010 to 2013, the success rate was 0.0.

However, in 2014 the success rate rose to 0.3 and then continued to rise through 2020.

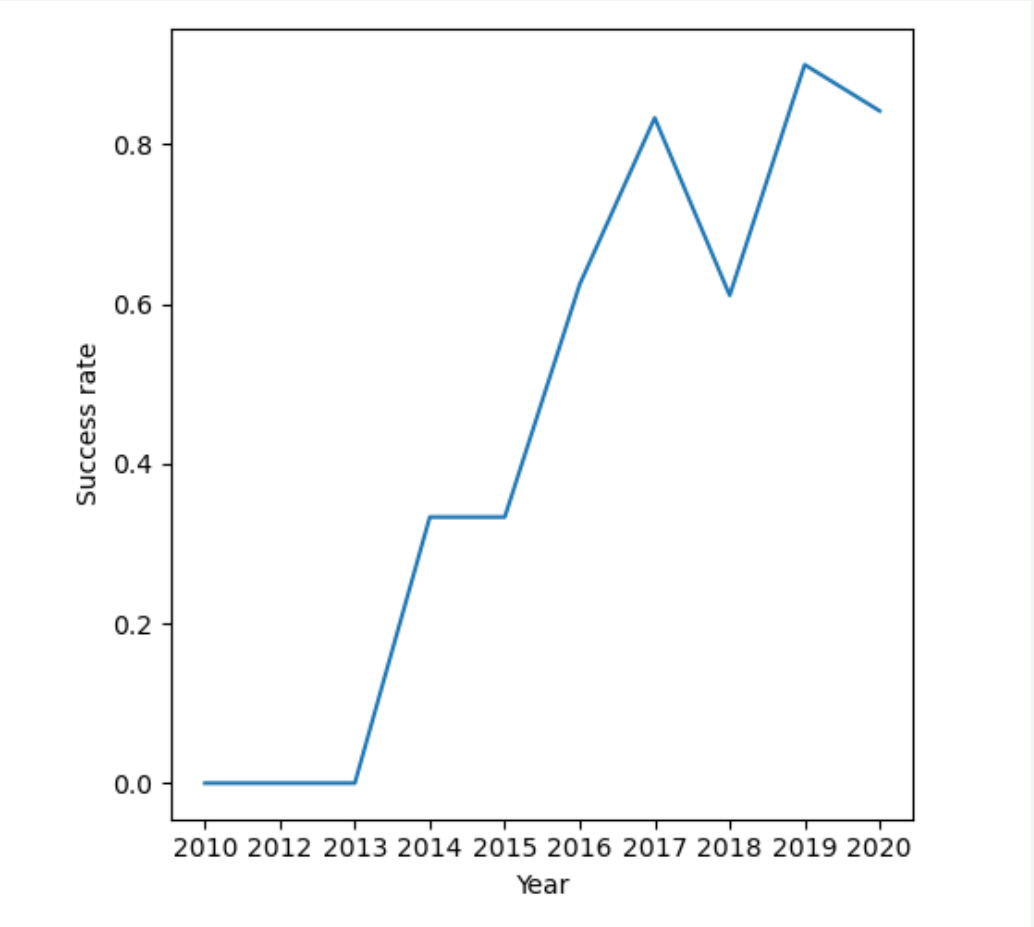


Figure: The line chart of yearly average success rate

All Launch Site Names

The two pictures below show the names of unique launch sites and the code used to extract it.

As you can see, SpaceX is launching from 4 different sites.

Figure: The names of unique launch sites and the SQL code used to extract it

Launch Site Names Begin with 'CCA'

The two pictures below show 5 records where launch sites begin with 'CCA' and the code used to extract it.

From the result, the common features are:

- Payloads are relatively small
- The target orbit type is LEO

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

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

Figure: 5 records where launch sites begin with "CCA" and the SQL code used to extract it

Total Payload Mass

The two pictures show total payload carried by boosters from NASA and the code used to extract it.

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

SUM(PAYLOAD_MASS__KG_)
45596.0

Figure: Total payload carried by boosters from NASA and the SQL code used to extract it.

Average Payload Mass by F9 v1.1

The two pictures show the average payload mass carried by booster version F9 v1.1 and the code used to extract it.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version = "F9 v1.1";
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Figure: The average payload mass carried by booster version F9 v1.1 and the SQL code used to extract it

First Successful Ground Landing Date

The two pictures below show the date of the first successful landing outcome on ground pad and the code used to extract it.

Considering that the launch has been underway since 2010, the difficulty of landing on ground pad can be glimpsed.

```
%%sql
SELECT "Date"
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)"
ORDER BY "Date" DESC
LIMIT 1;
```

Date
22/12/2015

Figure: The date of the first successful landing outcome on ground pad and the SQL code used to extract it

Successful Drone Ship Landing with Payload between 4000 and 6000

The two pictures below show the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, and the code used to extract it.

The result indicates that only F9 FT booster version has successfully landed on drone ship.

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE (Landing_Outcome = 'Success (drone ship)') AND
      (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

Figure: The names of boosters which have successfully landed on drone ship and had payload mass 4000 – 6000 kg, and the SQL code used to extract it

Total Number of Successful and Failure Mission Outcomes

The two pictures below show the total number of successful and failure mission outcomes and the code used to extract it.

As you can see, almost all flights are classified as successful.

```
%%sql
SELECT DISTINCT(Mission_Outcome), COUNT(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

Mission_Outcome	COUNT(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Figure: The total number of successful and failure mission outcomes and the SQL code used to extract it

Boosters Carried Maximum Payload

The two pictures show the names of the booster which have carried the maximum payload mass and the code used to extract it.

We can assume that the use of F9 B5 is suitable for carrying the maximum payload mass.

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM SPACEXTBL
                           );
```

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

Figure: The names of the booster which have carried the maximum payload mass and the SQL code used to extract it

2015 Launch Records

The two pictures show the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015. The code is used to extract the result.

From the result, it can be said that the booster version and the launch site are common.

```
%%sql
SELECT SUBSTR("Date", 4, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE SUBSTR("Date", 7, 4) = "2015" AND Landing_Outcome = "Failure (drone ship)";
```

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

Figure: The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, and the SQL code used to extract it

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

The two pictures show the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, and the code used to extract it.

It can be seen that the drone ship and the ground pad were successful about the same number of times.

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS LANDING_OUTCOME_COUNT
FROM SPACEXTBL
WHERE (Landing_Outcome LIKE 'Success%') AND ("Date" BETWEEN '04-06-2010' AND '20-03-2017')
GROUP BY Landing_Outcome
ORDER BY LANDING_OUTCOME_COUNT DESC;
```

Landing_Outcome	LANDING_OUTCOME_COUNT
Success	20
Success (drone ship)	8
Success (ground pad)	7

Figure: The count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, and the SQL code used to extract it


Section 3
Launch Sites
Proximities Analysis

Mark all launch sites

The picture shows a map with markers of all launch sites.

As you can see, all launch sites are in very close proximity to the coast.

Figure: The map with markers of all launch sites

Mark the success/failed launches for each site

The picture shows a map with markers of the launch success rate for each site.

The number of launches at the same site is indicated by cluster markers, with successful launches indicated by green markers and failures by red markers.

These pictures show that the success rate at KSC LC-39A is relatively high.

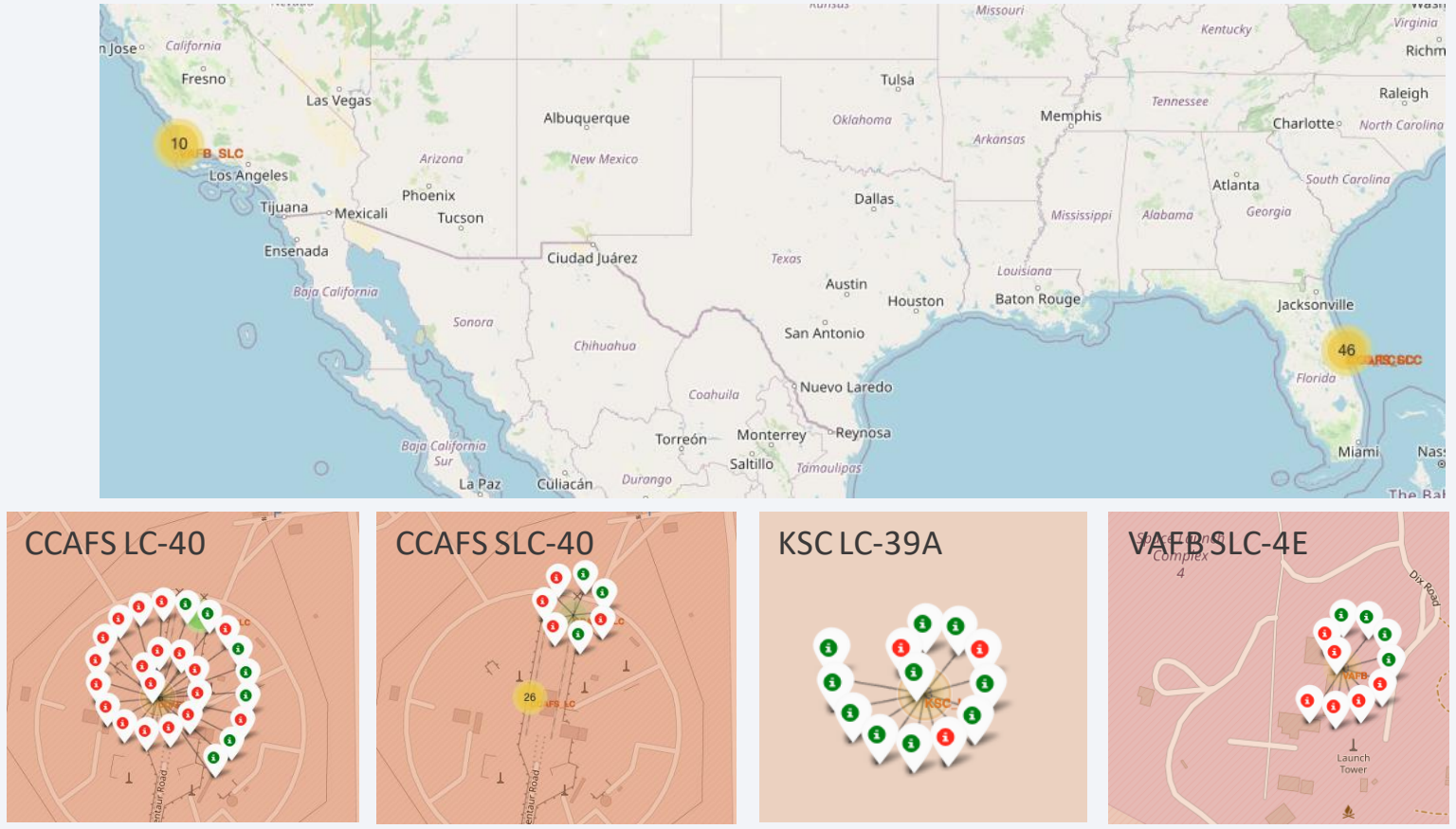


Figure: The map with markers of the launch success rate for each site

Calculate the distances between a launch site to its proximities

The picture shows the distance between VAFB SLC-4E and its proximities (railway, highway, coastline and city) with lines.

From the picture, it is obvious that the highway and city are far enough away from the launch site.

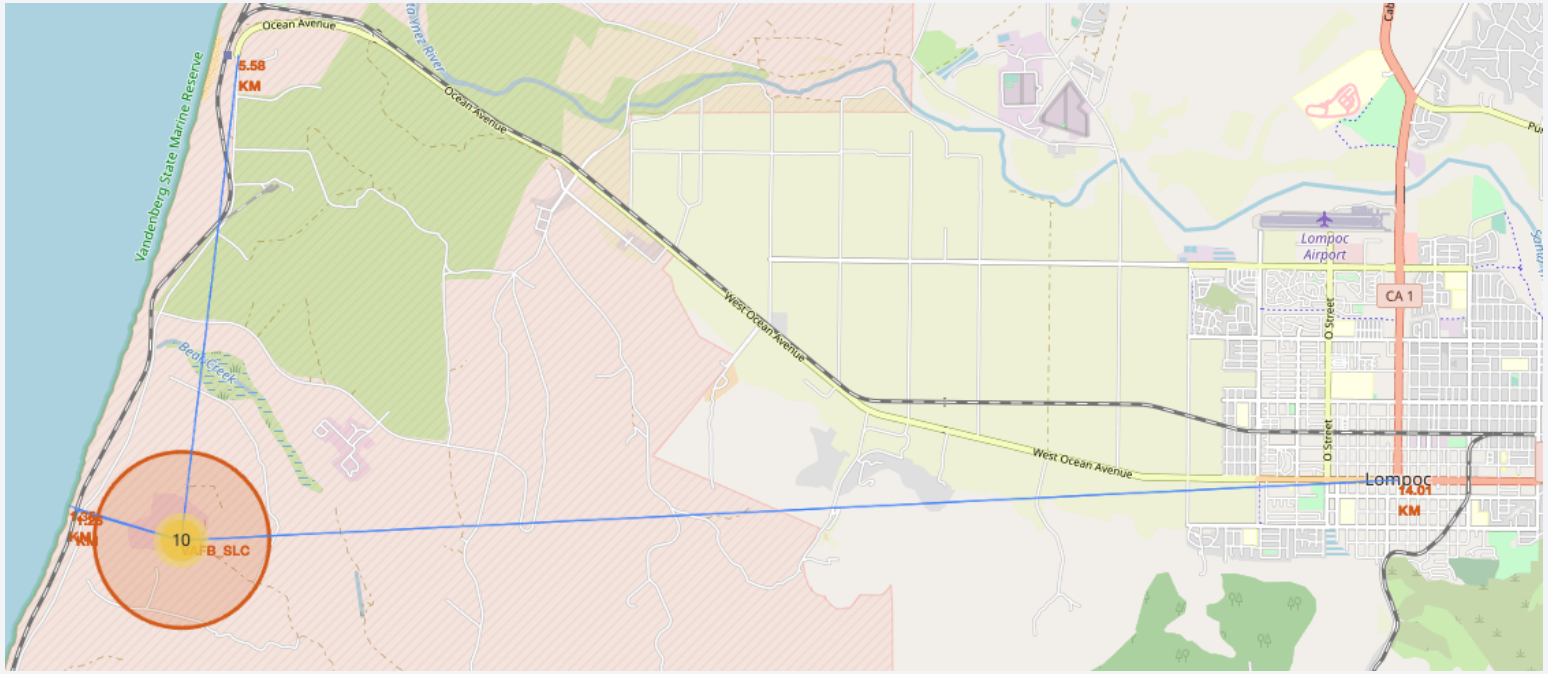
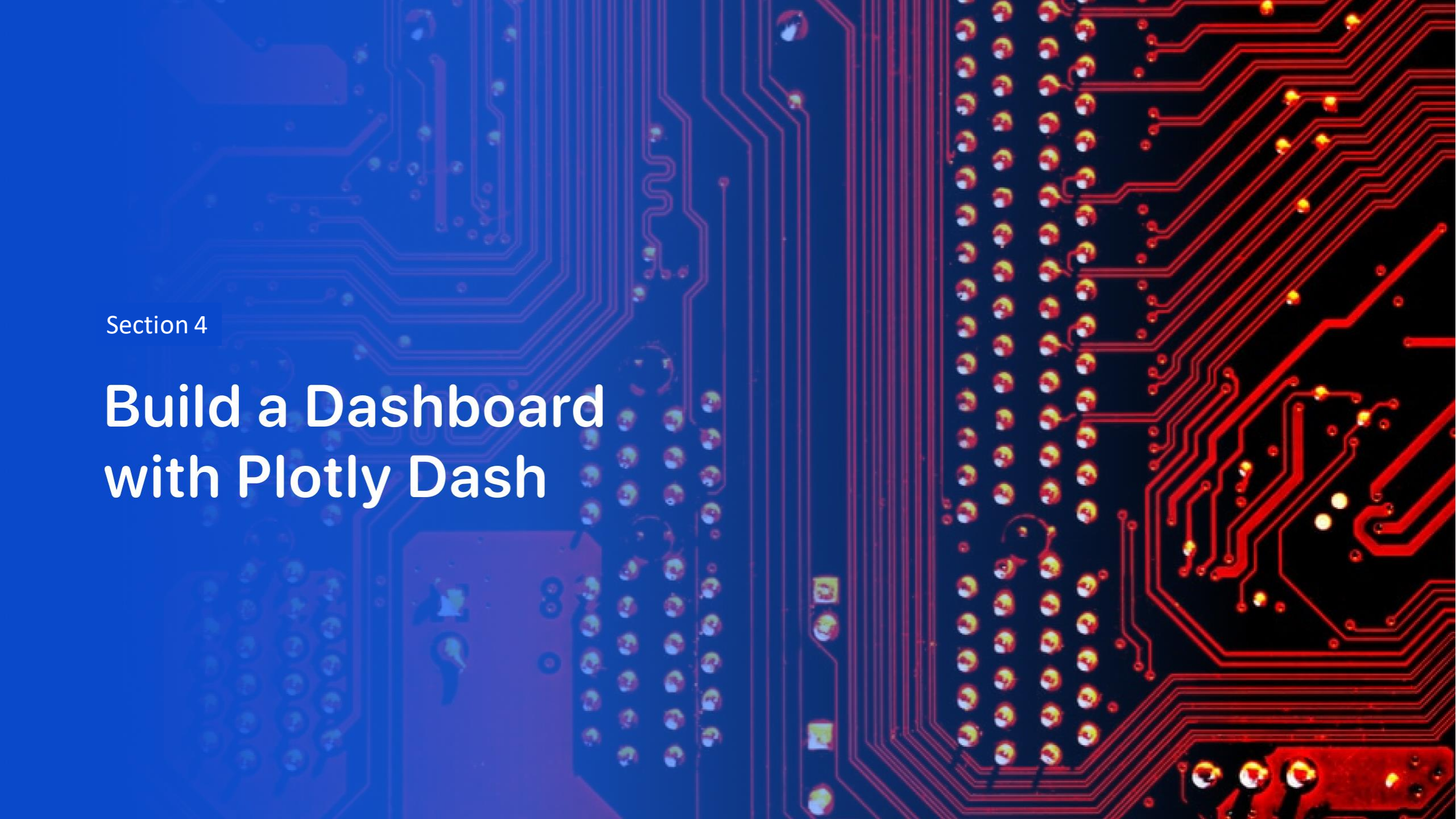


Figure: The map with lines showing the distances between VAFB SLC-4E and its proximities


Section 4
Build a Dashboard with Plotly Dash

Total success count for all sites in a pie chart

The picture shows the total success count of launch for all sites in a pie chart.

As you can see, launches at KSC LC-39A account for about half of all successful launches.

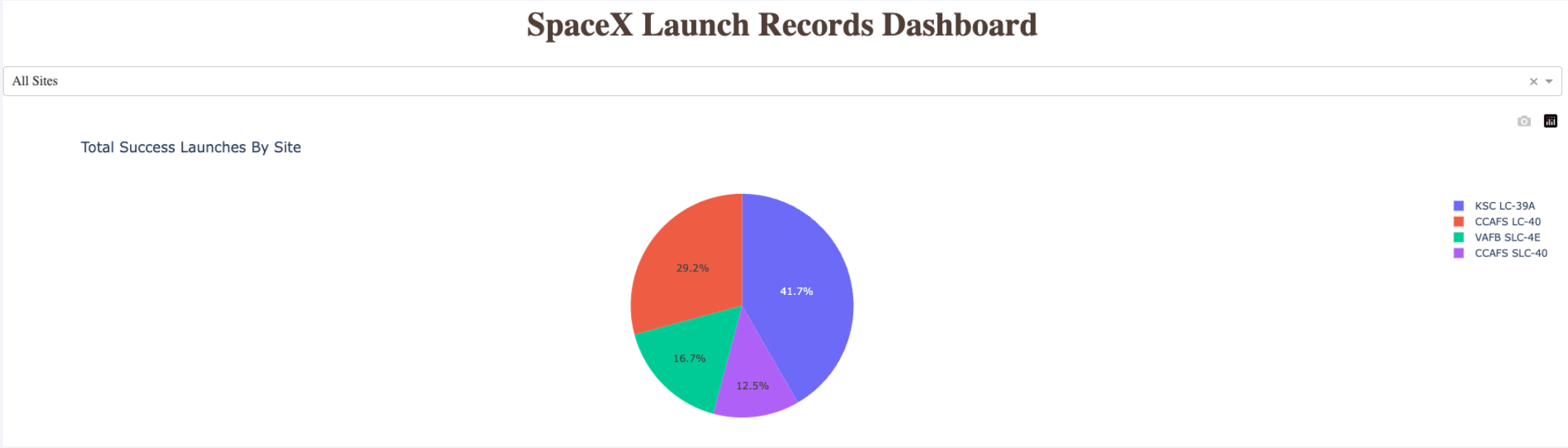


Figure: The total success count of launch for all sites in a pie chart

Total success count for KSC LC-39A

The picture shows the total success count of launch for KSC LC-39A in a pie chart.

The blue zone indicates successful launch, and the red zone indicates failure.

At KSC LC-39A, it can be said that three out of four launches are successful.

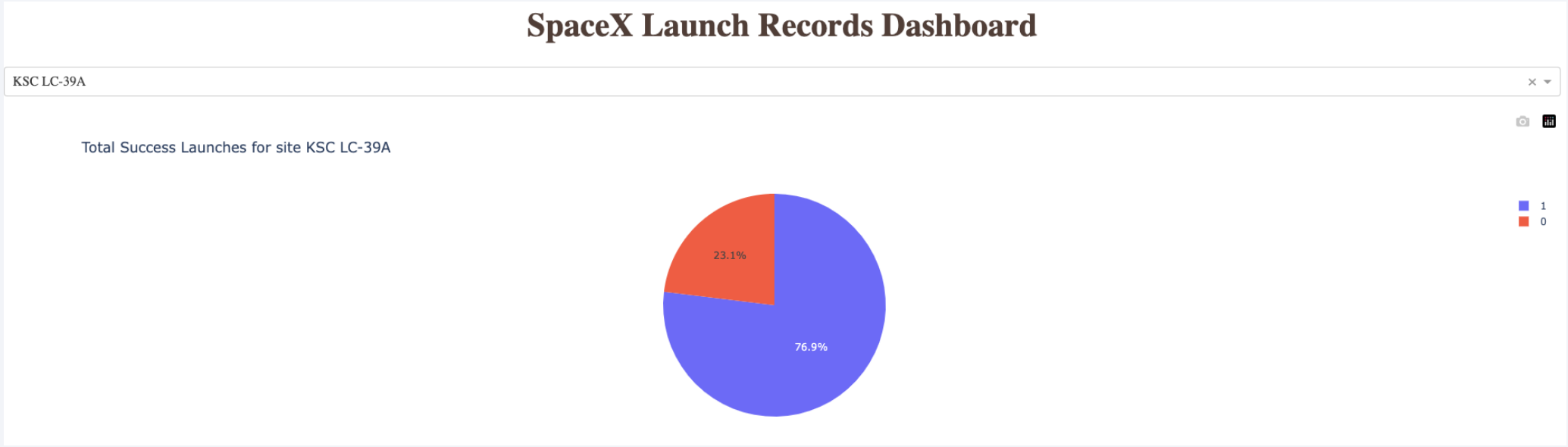


Figure: The total success count of launch for KSC LC-39A in a pie chart

Correlation between payload and success for all sites

The picture shows the scatter plot of Payload Mass vs. Launch Outcome for all sites, with different payload selected in the range slider.

The majority of successes are in 2000 kg to 6000 kg payload range, with a large percentage of launches which the booster version is FT.

Figure: The scatter plot of Payload mass vs. Launch outcome for all sites

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The graph shows a bar chart of model accuracy for all built classification models (logistic regression, support-vector machine, decision tree and k-nearest neighbor).

All models were measured to be equally accurate, with the decision tree model having the highest value.

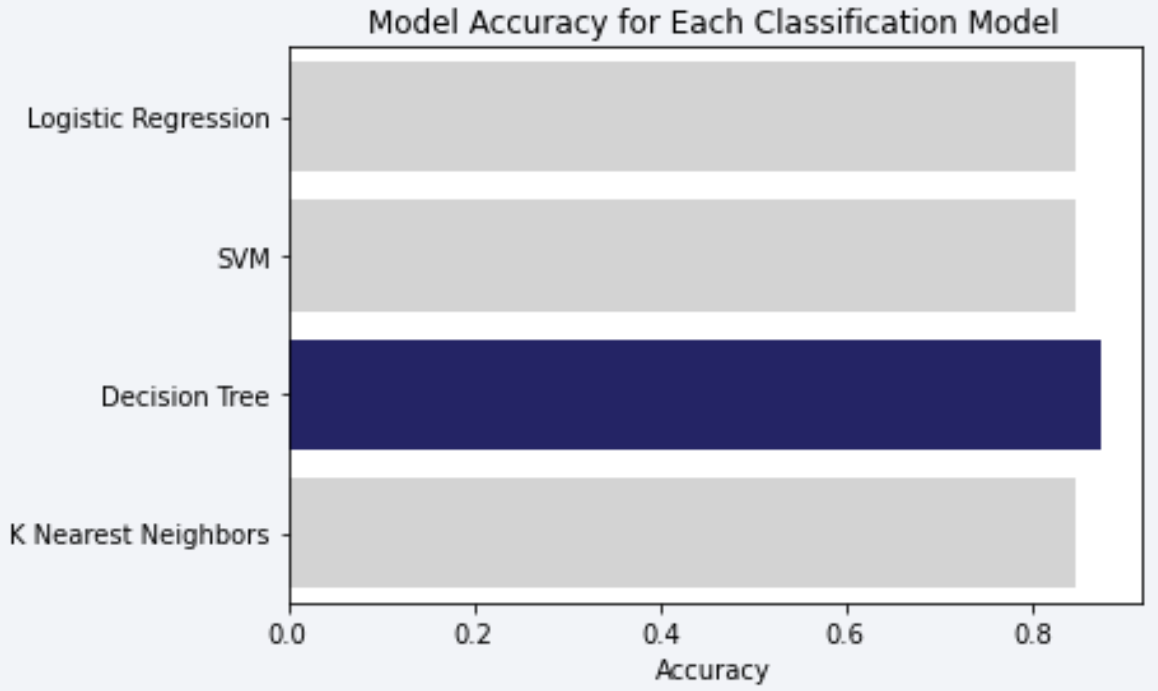


Figure: The bar chart of model accuracy for all built classification models

Confusion Matrix

The picture shows the confusion matrix of the best performing model (decision tree).

From the matrix, it can be said that accuracy improves when a value of false positives decreases.

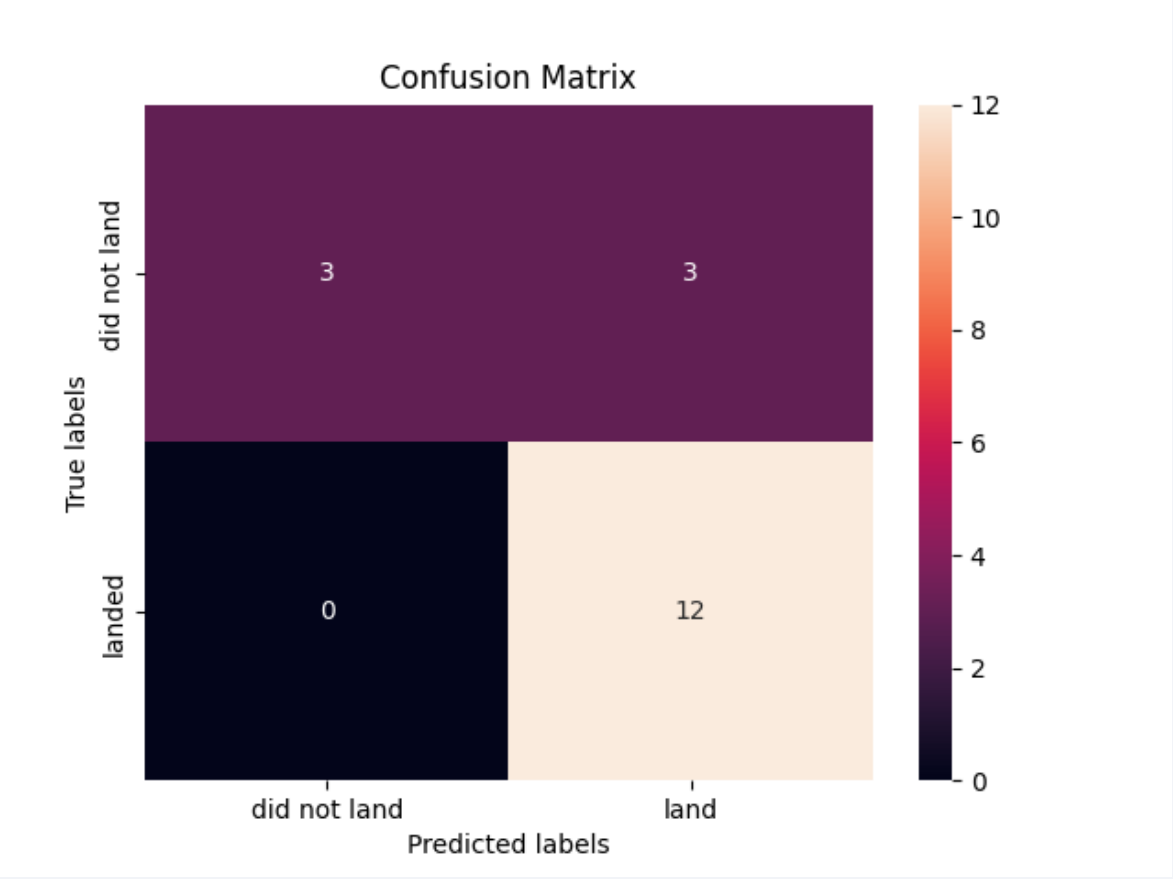


Figure: The confusion matrix of the decision tree classifier model

Conclusions

- The larger the payload, the higher the success rate.
- The success rate increases as the flight number increases. This suggests that SpaceX is able to apply the knowledge gained from each flight to subsequent flights.
- The success rate varies depending on orbit types. From this, it can be inferred that the difficulty of flight varies depending on the orbit aimed at.
- The number of successful flights on KSC-LC-39A was the highest, accounting for 40% of the total.
- The decision tree classifier model performed with the highest accuracy in the predictive analysis.

Appendix

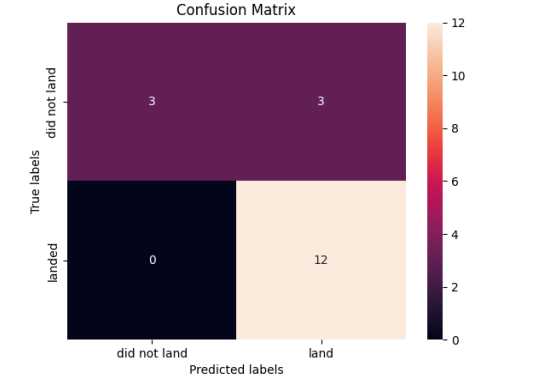
- Accuracy

Table: Accuracy of each model

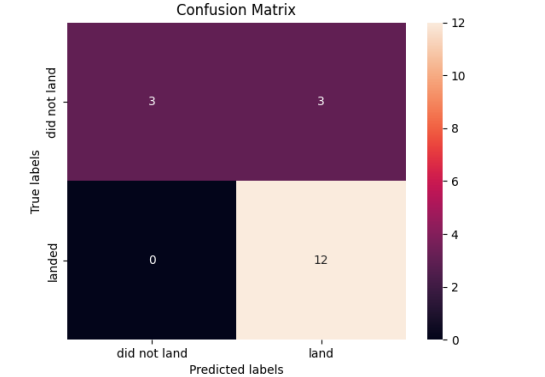
Model	Accuracy
Logistic Regression	0.8464
Support Vector Machine	0.8482
Decision Tree Classifier	0.8732
K-nearest Neighbors	0.8482

- Confusion matrix

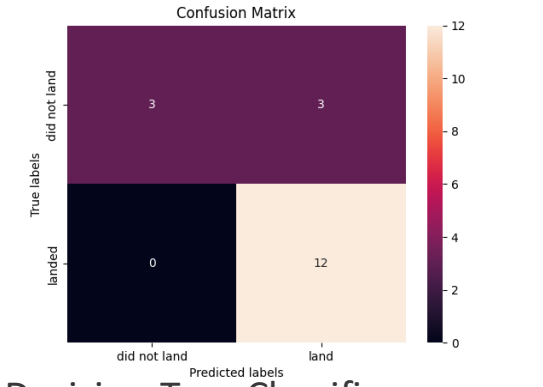
Logistic Regression



Support Vector Machine



Decision Tree Classifier



K-nearest Neighbors

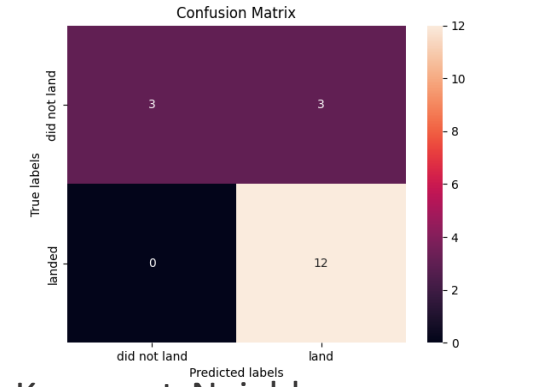


Figure: The confusion matrix of all classification models

Thank you!