IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Reygen Win Japar
26 March 2025

Outline

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

Executive Summary

- Summary of methodologies

  1) Data Collection: Retrieved launch data using the SpaceX API and web scraping techniques.

  2) EDA: Using SQL queries to extract insights on launch sites, payloads, and outcomes. After that, creates visualizations to identify patterns in launch success rates.

  3) Interactive Dashboard: Implemented dropdowns, sliders, and graphs to explore launch success rates based on payload, launch site, and booster version.

  4) Predictive Analysis: Built the models (Logistic Regression, SVM, Decision Tree, KNN) and did the tuning using GridSearchCV. And then evaluated the models using accuracy as the scoring.

Executive Summary

- Summary of all results

1) Launch Success Analysis: The KSC LC-39A launch site had the highest success rate (76.9%), while CCAFS SLC-40 had the lowest. The success rate increased over time from 2013 to 2020.

2) Orbit Type & Mission Outcomes: Some orbit types (ES-L1, GEO, HEO, SSO) had a 100% success rate, while SO orbit had 0% success due to only one launch attempt. The LEO orbit showed a strong correlation between flight number and success rate.

3) Machine Learning Predictions: All models performed similarly, suggesting that the dataset may be small or imbalanced. So, we might need to collect more data and improving the features by feature engineering.

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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

- Problems you want to find answers:

  Predict if the Falcon 9 first stage will land successfully based on the features provided.

Section 1

Methodology

Methodology

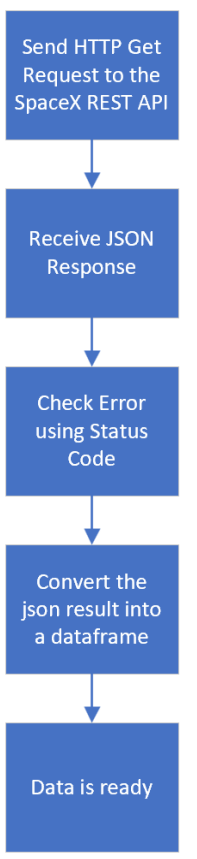
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

- First, we begin by querying the SpaceX REST API using HTTP GET Request. The API returns a JSON response. Then we check the error using status code. After that we convert the JSON to a Pandas DataFrame. Finally, the data is ready to be analyzed.

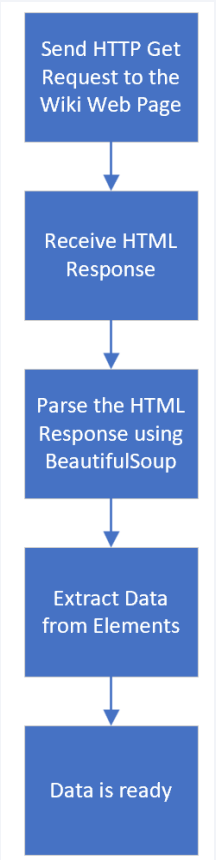
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%201%20-%20Data%20Collection%20and%20Wrangling/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

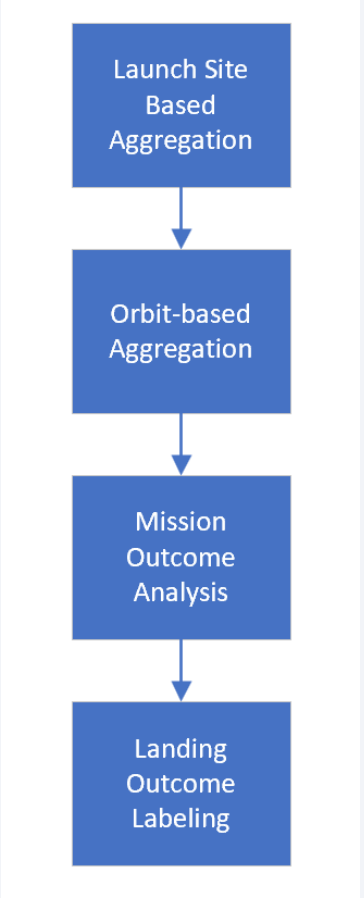
- First, we begin by sending a HTTP GET Request. The request returns a HTML response. Then we parse the HTML response using BeautifulSoup. After that we extract the column name and parse the HTML Tables to create a DataFrame. Finally, the data is ready to be analyzed.

- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%201%20-%20Data%20Collection%20and%20Wrangling/jupyter-labs-webscraping.ipynb



Data Wrangling

- First, we analyze the data collected before by calculating the number of launches on each site, calculating the number and occurrence of each orbit, and calculating the number and occurrence of mission outcome of the orbits. And then, we create a landing outcome label from Outcome column.
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%201%20-%20Data%20Collection%20and%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The chart I used:

1) Categorical Plot: to visualize the relationship between categorical variables and numerical variables, while also analyzing their relationship with the “Class” data.

2) Bar Plot: kind of catplot, to visualize the relationship between categorical variables and numerical variables. In this problem, we used barplot to visualize the relationship between success rate of each orbit type.

3) Line plot: to visualize a trendline over the year. In this problem, we used lineplot to visualize the trend of success rate each year.

- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%202%20-%20EDA/edadataviz.ipynb

EDA with SQL

- Displaying unique values of categories.
- Displaying records based on the condition needed.
- Displaying the total number of unique values in a category.
- Displaying the records with a maximum value of a certain category.
- Displaying time series records.
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%202%20-%20EDA/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

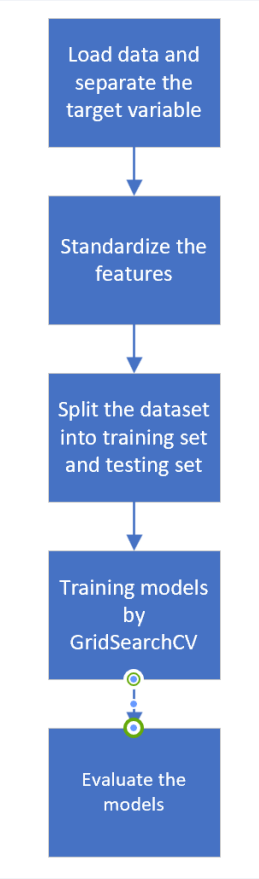
- Markers: to create a label marker to tell us the meaning of the marker (e.g: lauch site name).
- Circle: to create a circle mark on the coordinate we provide. In this problem, circle marker is used to mark the location of the launch sites.
- Line: to create a line mark along the coordinate we provide. In this problem, line marker is used to mark the distance between launch sites and some places (e.g: railway, highway, nearest city, nearest coastline).
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%203%20-%20Interactive%20Visualization/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown Menu for Launch Site Selection: to allow filtering out data to analyze individual launch sites separately, the default is ‘ALL’ launch sites.
- Pie Chart: to create a pie chart for analyzing the proportion of success rate of launching the SpaceX Falcon 9.
- Range Slider: to analyze how different payload ranges impact on the success rate of launches.
- Scatter Plot: to analyze the relationship between payload mass and launch success.
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%203%20-%20Interactive%20Visualization/spacex-dash-app.py

Predictive Analysis (Classification)

- First, I separate the target value and the features. Then, I standardize the data in X. After that, I split the dataset into training set and testing set (it should be done before standardizing the data).
- After the data is ready, I start to training the model using GridSearchCV. The model I used are: Logistic Regression, SVM, Decision Tree, and KNN. After training the model, I calculate the accuracy of the models and plot the confusion matrices.
- Link: https://github.com/trytry987/SpaceX-Falcon-9/blob/main/Module%204%20-%20Predictive%20Analysis/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

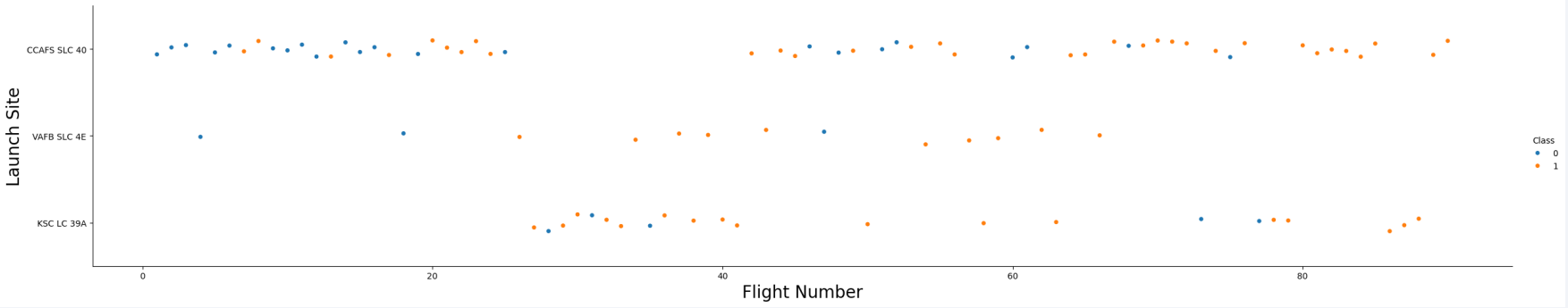


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

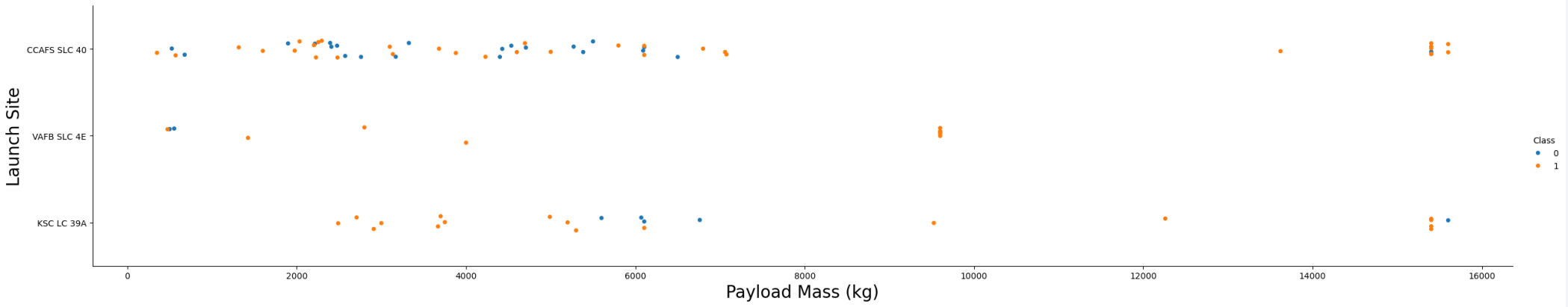
- Show a scatter plot of Flight Number vs. Launch Site



  - From the scatter plot, we could infer that the successful rate increases as the flight number increases at every launch sites.

Payload vs. Launch Site

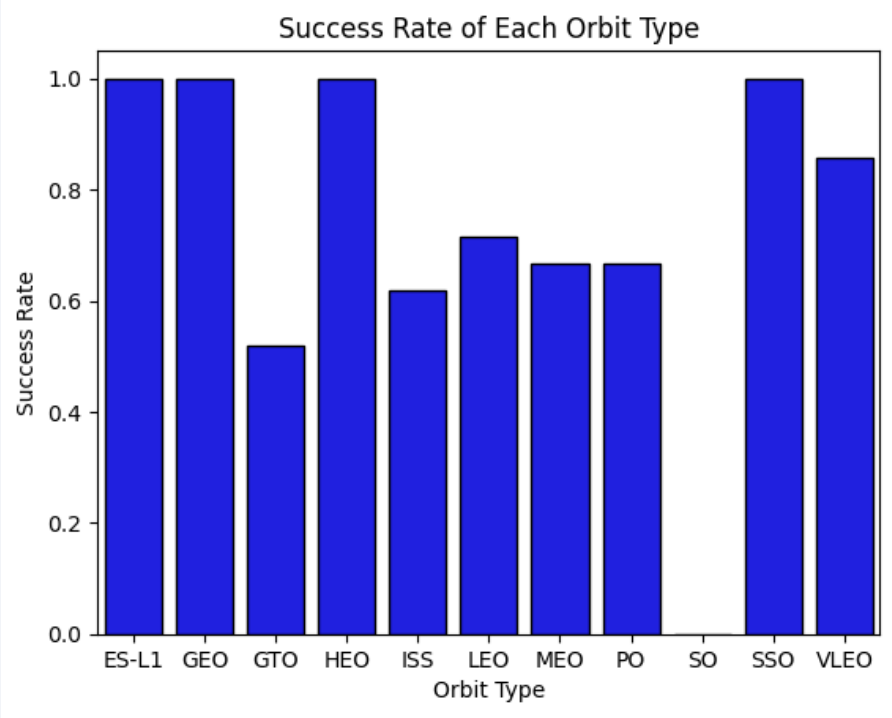
- Show a scatter plot of Payload vs. Launch Site



  - From the scatter plot, we could infer that there are no rockets launched for the payload mass more than 10000 kg at VAFB-SLC 4E. In VAFB-SLC 4E, as the payload mass increases, the success rate also increases.

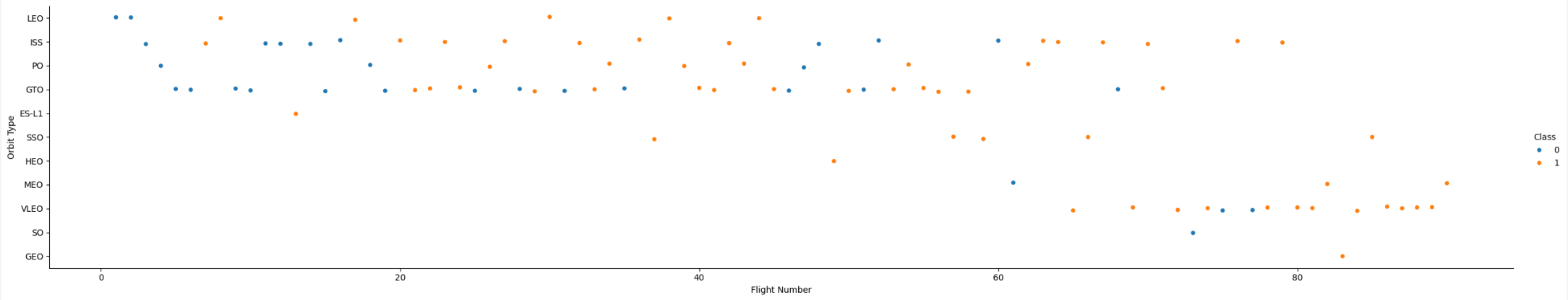
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- From the bar chart, we could infer that the orbits ES-L1, GEO, HEO, SSO have the highest success rate (100%), while the orbit SO has the lowest success rate (0%). This is because the rocket has only been launched in orbit SO once.



Flight Number vs. Orbit Type

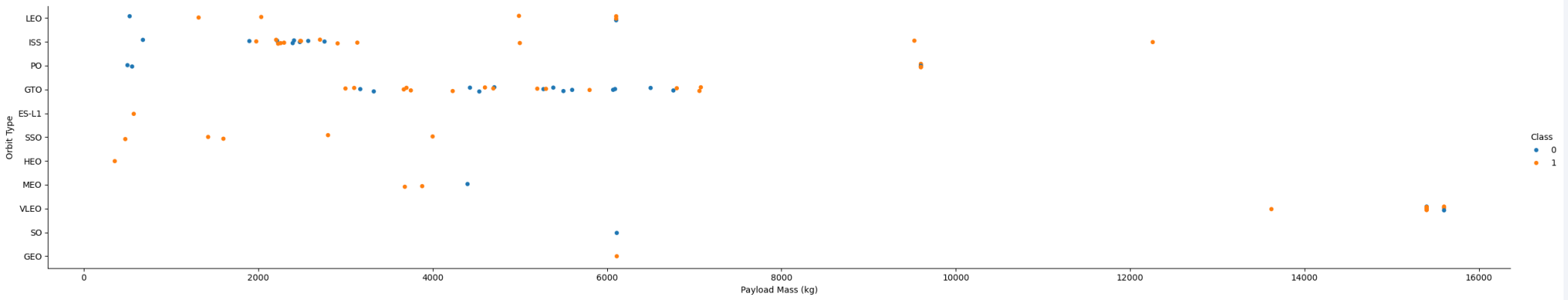
- Show a scatter point of
  Flight number vs. Orbit type



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

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

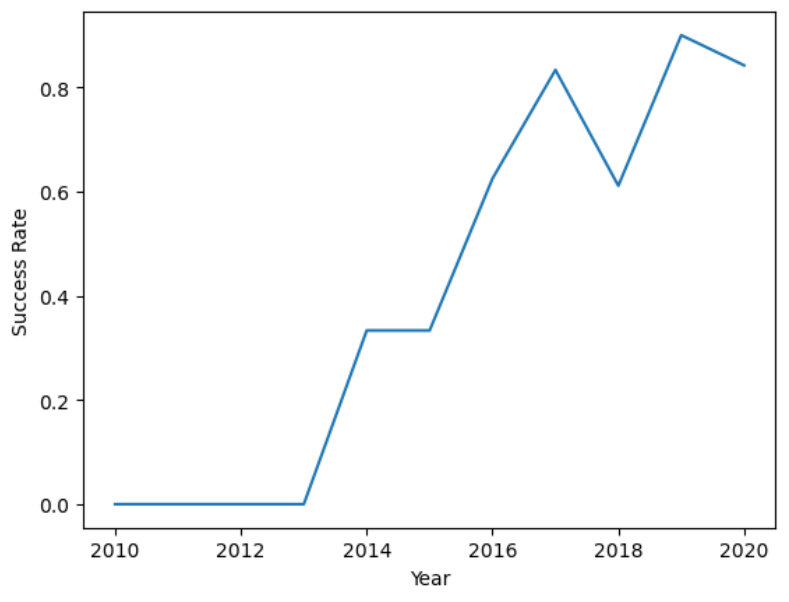


- From the scatter plot, we could infer that in the orbits LEO, ISS, and PO, as the payload mass increases, the success rate also increases. While for the orbit GTO, it's difficult to distinguish the relationship between payload mass and success rate, as both outcomes are present.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- From the line chart, we could infer that the success rate was increasing from 2013 till 2020.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here: There are 4 unique launch sites, that is, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40.

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

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here: The top 5 records where the launch sites begin with ‘CCA’ are in the year ranges from 210 to 2013, with 2 failure landing outcome and 3 no attempt.

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here: We could see that the total payload carried by boosters from NASA is 45596 kg.

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

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here: We could see that the average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

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

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

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here: We could see that the first successful ground landing date is at 22 December 2015.

```
%%sql
    SELECT MIN(Date)
    FROM SPACEXTABLE
    WHERE "Landing_Outcome" LIKE '%ground pad%';
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Present your query result with a short explanation here: We could see that there are 4 booster version with payload mass greater than 4000 but less than 6000.

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

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here: Here, there is 1 failure in flight, 99 success, and 1 success with the payload status is unclear.

```
%%sql
    SELECT "Mission_Outcome", COUNT(*) AS 'Total Number'
    FROM SPACEXTABLE
    GROUP BY TRIM("Mission_Outcome");
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here: We could see that 18 booster versions have the maximum payload mass, that is 15600 kg.

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

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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here: We could see that there are 2 failure in the drone ship in year 2015, exactly in January and April.

```
%%sql
    SELECT substr(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
    FROM SPACEXTABLE
    WHERE substr(Date, 0, 5) = '2015'
        AND "Landing_Outcome" = 'Failure (drone ship)';
```

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

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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Present your query result with a short explanation here: We could see that the most landing outcome is “No attempt”, this might be because the rockets are still in development phase.

```
%%sql
    SELECT "Landing_Outcome", COUNT(*)
    FROM SPACEXTABLE
    WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY "Landing_Outcome";
```

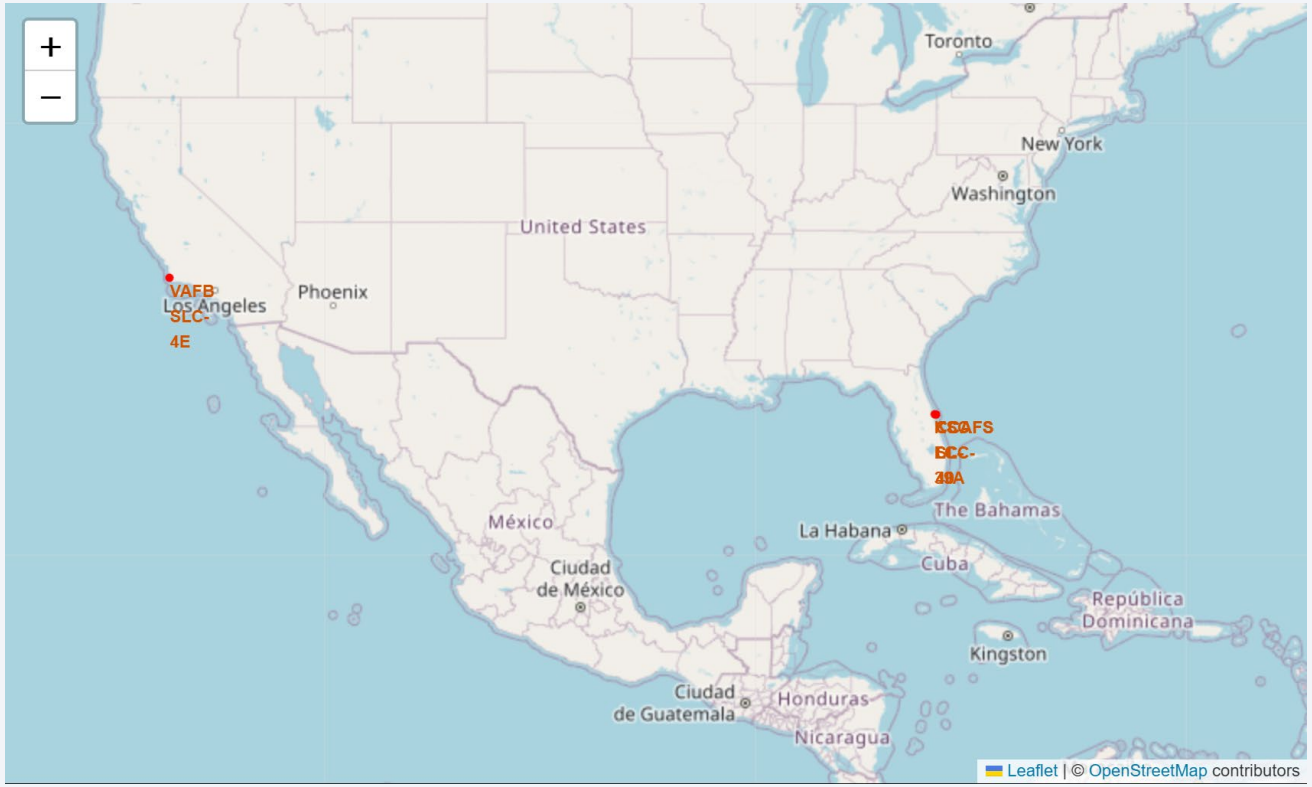
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

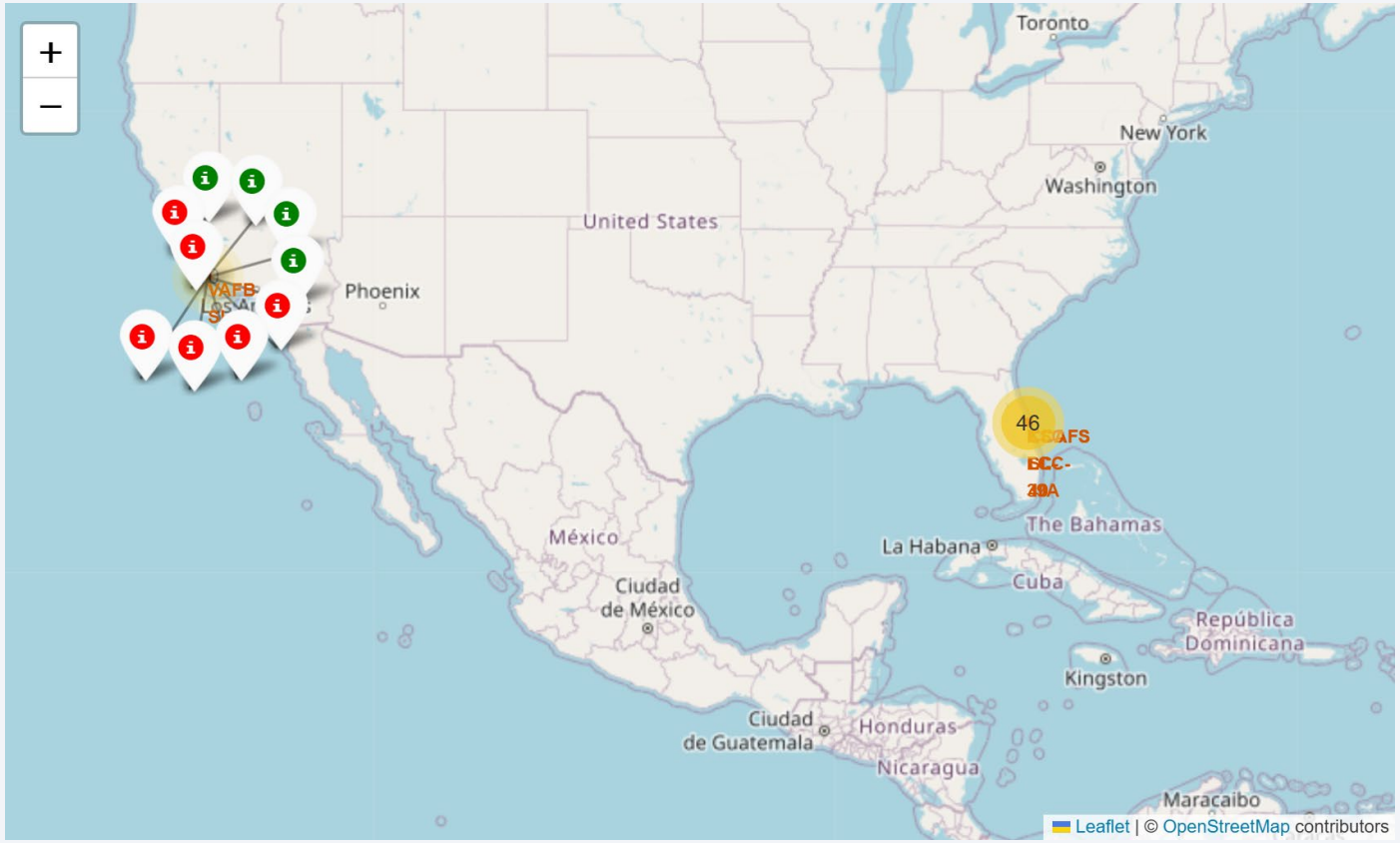
Launch Sites Proximities Analysis

Launch Sites Map



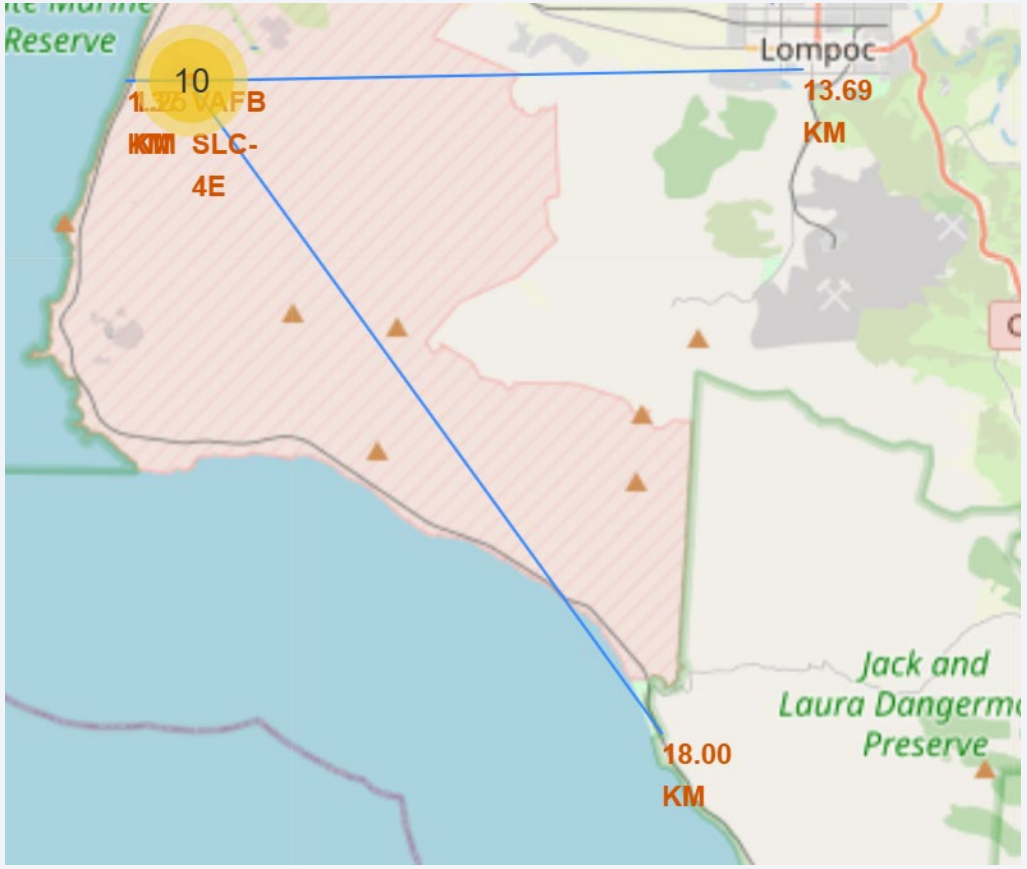
- From the map, we could observe that all of the launch sites are near the coastlines.

Color-Labeled Launch Sites Map



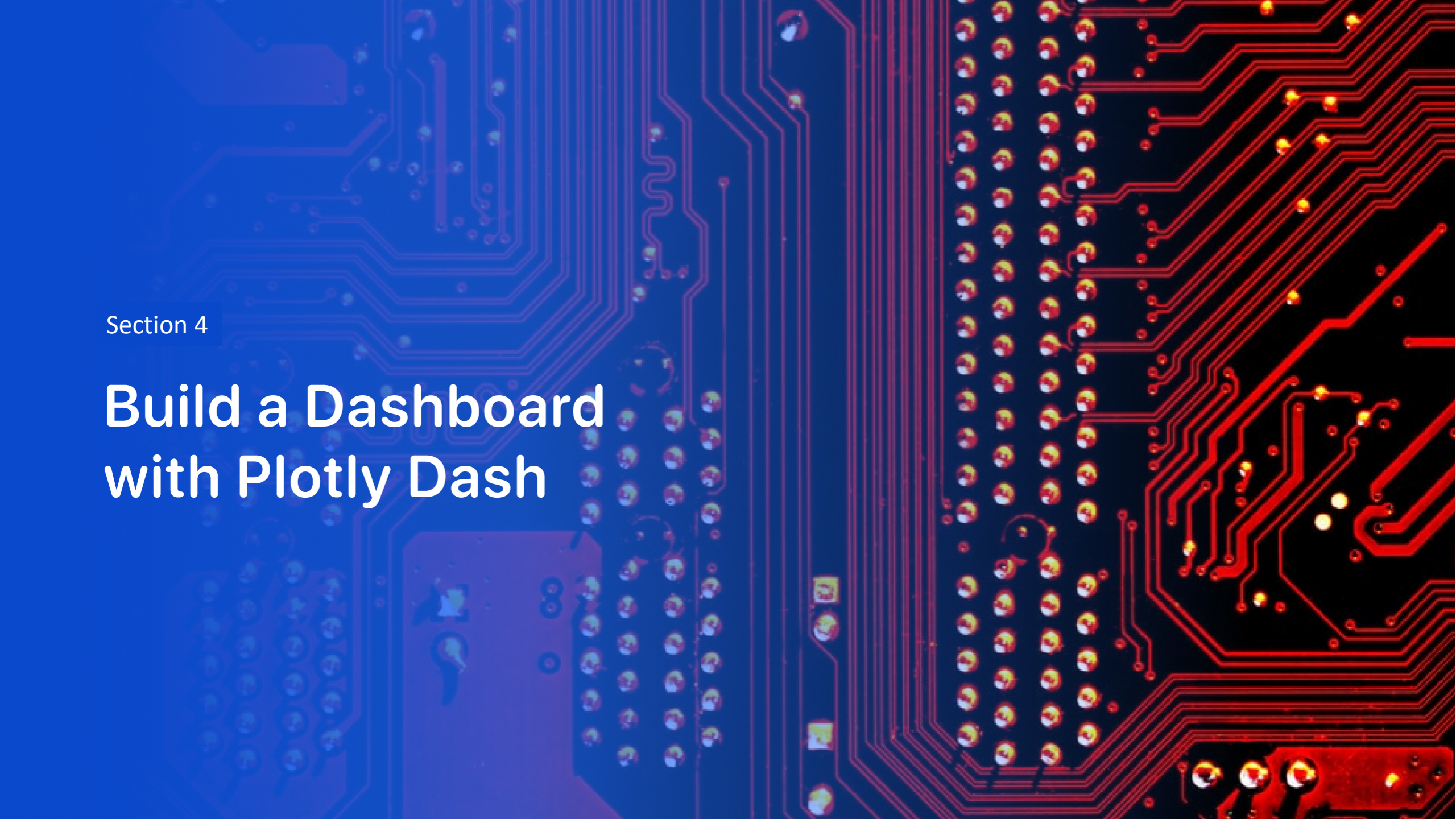
- 46 rockets are launched in the East Coast, while the other 10 rockets are launched in the West Coast. There were only 4 out of 10 are successfully landed.

Distance of a Launch Site to Its Proximities

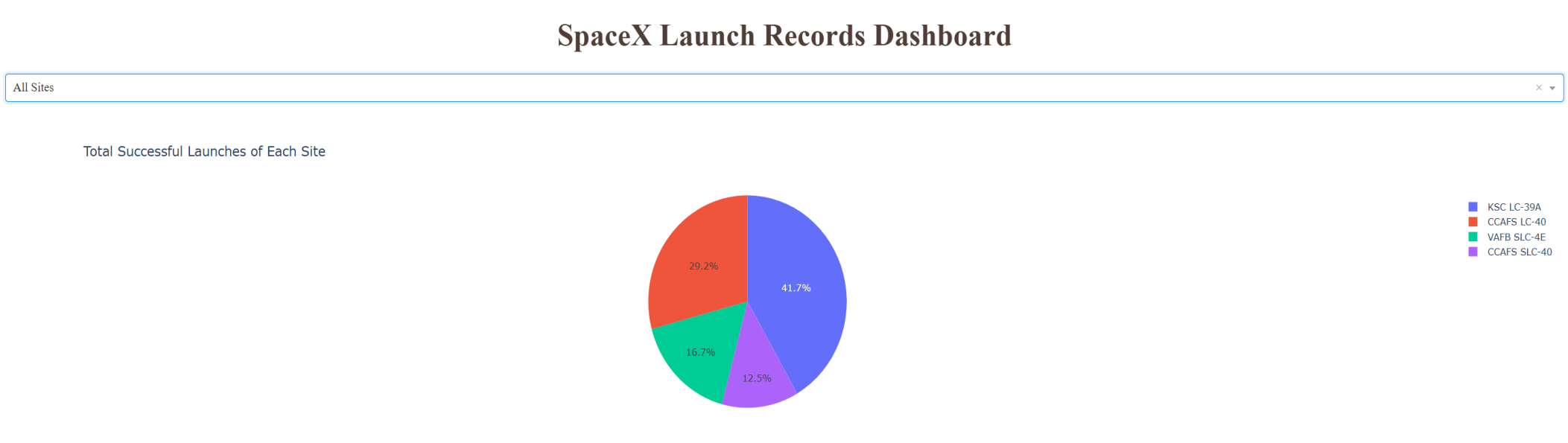


- We could see that the launch site is near by its proximities:

1) Coastline: 1.37 km
2) Railway: 1.25 km
3) Highway: 18 km
4) Nearest City: 13.69 km

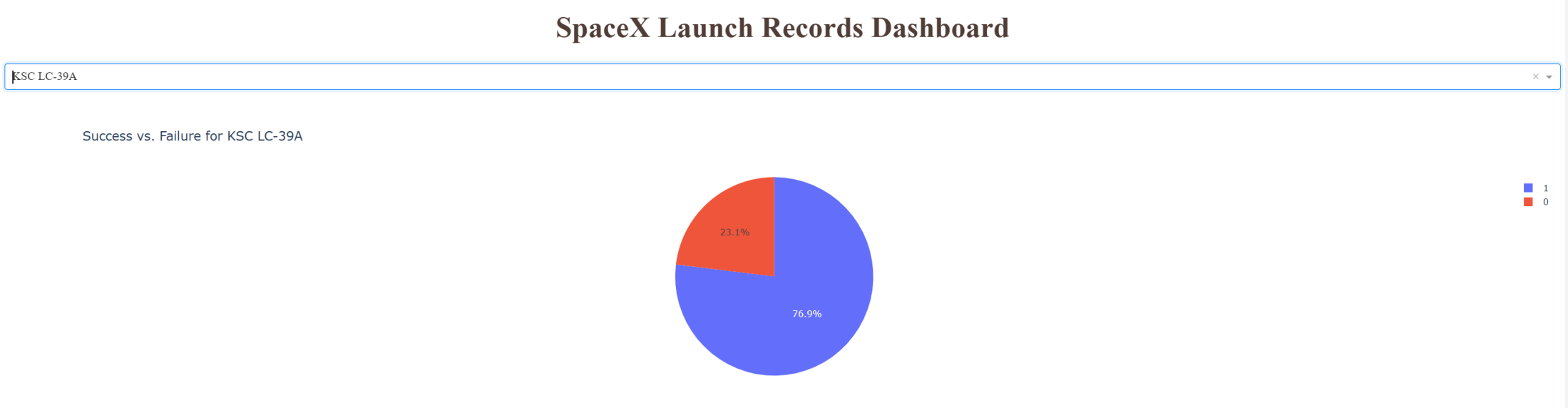

Section 4
Build a Dashboard with Plotly Dash

Launch Success Rate for All Sites



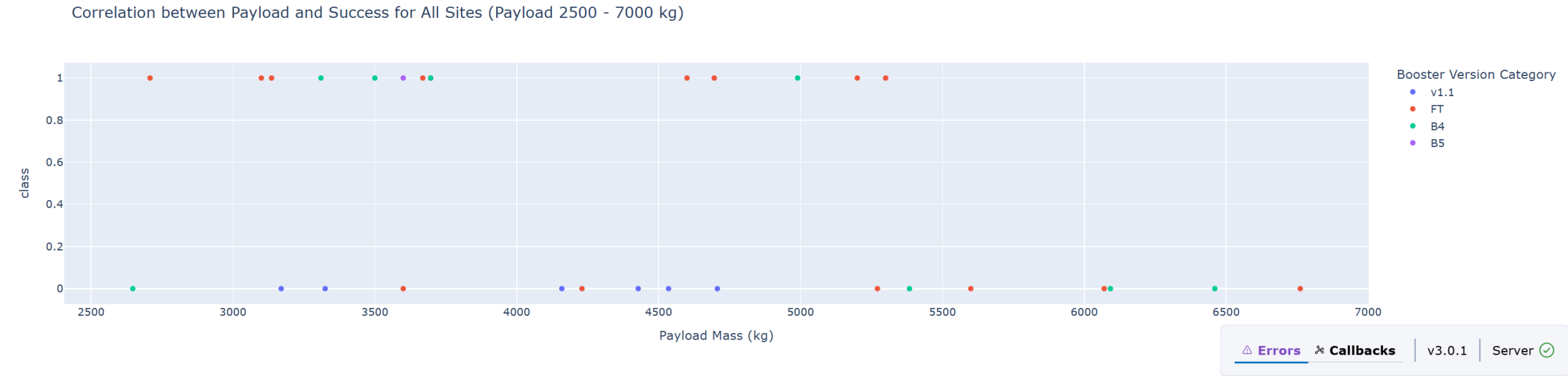
- From the pie chart, we could infer that KSC LC-39A has the highest launch success rate. While CCAFS SLC-40 has the lowest success rate.

The Pie Chart of KSC LC-39A



- Launch site KSC LC-39A has the highest launch success ratio.
- From the pie chart, we could observe that launch site KSC LC-39A has a successful rate of 76.9%.

Payload vs Launch Outcome for All Sites>



- From the scatter plot, we could observe that there are no booster version that is landed successfully at payload mass between 5300 kg and 7000 kg.

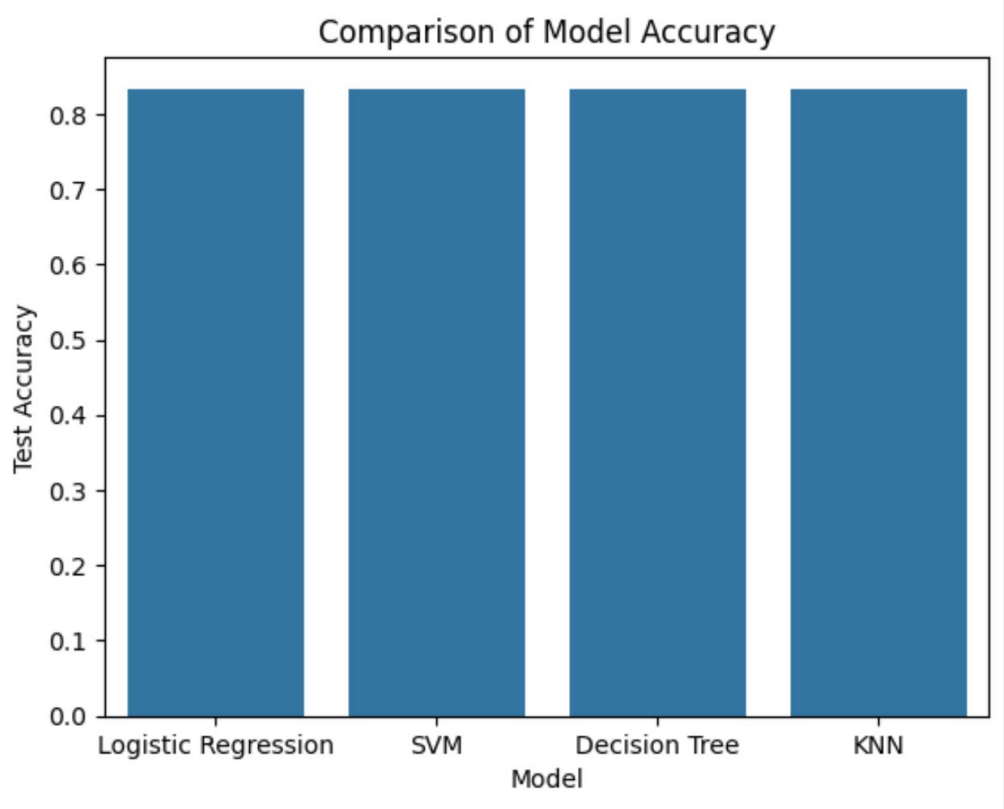
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- From the bar plot, we could observe that all of the models has the same accuracy, which is quite suspicious.

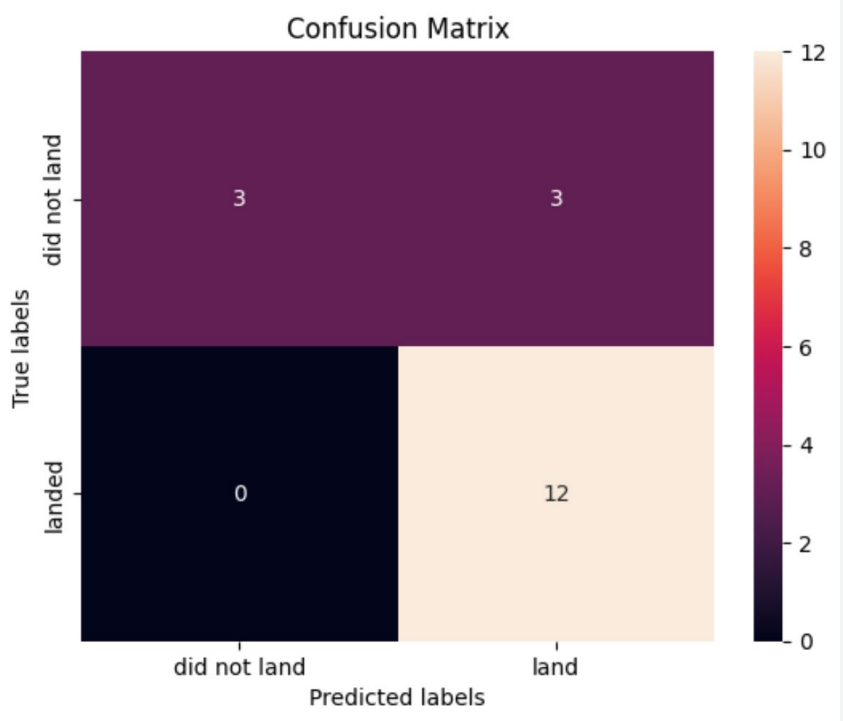
- This might be because the size of the testing set is so small, which is only 18.



Confusion Matrix

- Because the model has the same performance, so I took the confusion matrix of the KNN.
- From the table, we could observe:

1) True Positive = 12 (Bottom right)
2) False Positive = 3 (Upper right)
3) False Negative = 0 (Bottom left)
4) True Negative = 3 (Upper left)



Conclusions

- SpaceX's launch success rate improves with experience, as indicated by the correlation between flight number and success, and the trendline from 2013 to 2020 showed that the successful rate kept increasing.
- Certain launch sites (KSC LC-39A) perform better than others, suggesting that location factors may influence outcomes.
- The current classification models show similar accuracy, likely due to the limited dataset size. Future improvements could involve more data collection and feature engineering.

Appendix

- Link: https://github.com/trytry987/SpaceX-Falcon-9

Thank you!