Winning Space Race with Data Science

Peerapong Charoenkijwattanakul
28 June 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Analytics in screenshots
  - Predictive Analytics results

Introduction

- Project background and context

  Space X 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 Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

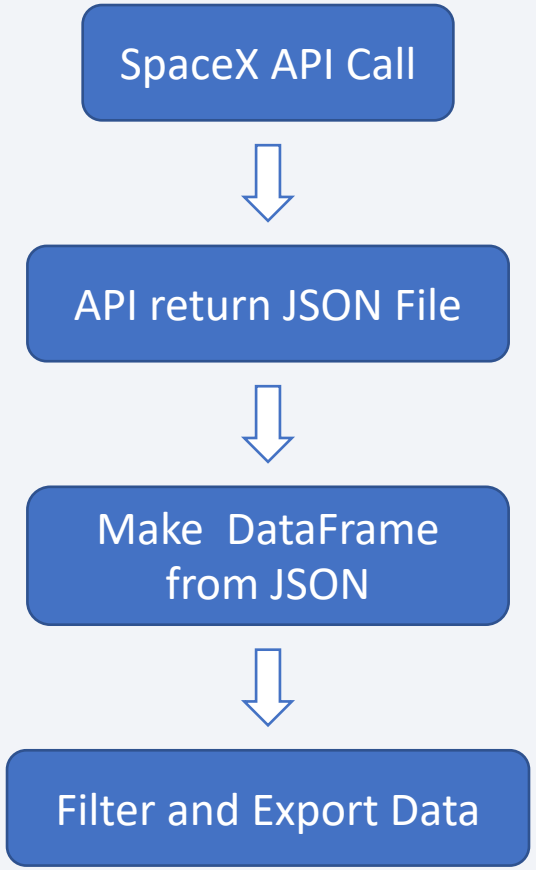
Methodology

- Data collection methodology:
  - Data collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Applied necessary transformations and cleaning techniques
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Explored data characteristics and relationships
- Perform interactive visual analytics using Folium and Plotly Dash
  - Created interactive visualizations for in-depth data exploration
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated classification models for outcome prediction

Data Collection

- Data collection involved multiple steps:
  - The first step was to retrieve the data by making a GET request to the SpaceX API.
  - Next, the response content was decoded as JSON using the .json() function, and then converted into a Pandas DataFrame using .json_normalize().
  - The data was then cleaned, with a thorough check for missing values and appropriate filling of those gaps when necessary.
  - Additionally, web scraping techniques were applied to extract Falcon 9 launch records from Wikipedia using BeautifulSoup.
  - The objective was to extract the launch records as an HTML table, parse the table, and convert it into a Pandas DataFrame for future analysis.

Data Collection – SpaceX API




```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json
data = pd.json_normalize(response.json())
```

```
# Lets take a subset of our dataframe keeping only the features we want and the flight
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']]

# We will remove rows with multiple cores because those are falcon rockets with 2 extra
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single value in
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then extracting the d
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

Data Collection – Web Scraping

Get HTML response from Wikipedia

⇩

Extract Data with BeautifulSoup

⇩

Filter table from soup and make DataFrame

⇩

Export Data

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_
```

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

```
# Use BeautifulSoup() to create a BeautifulSoup object
soup = BeautifulSoup(response.text, 'html.parser')
```

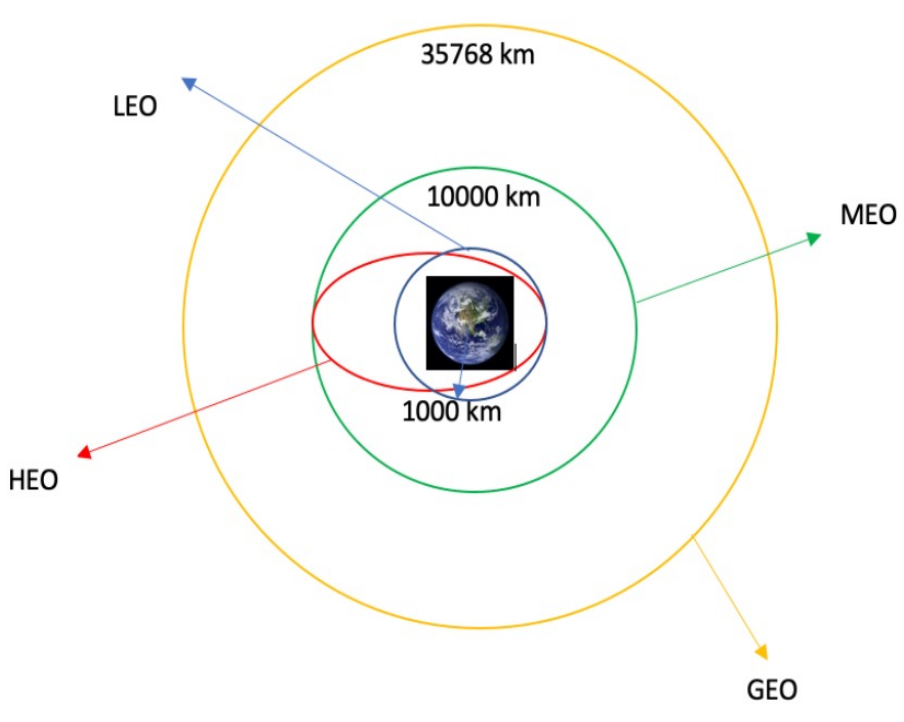
```
# Use the find_all function in the BeautifulSoup object
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

```
# Let's print the third table and check
first_launch_table = html_tables[2]
```

```
header_elements = first_launch_table.find_all('th')

for element in header_elements:
    name = extract_column_from_header(element)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

Data Wrangling

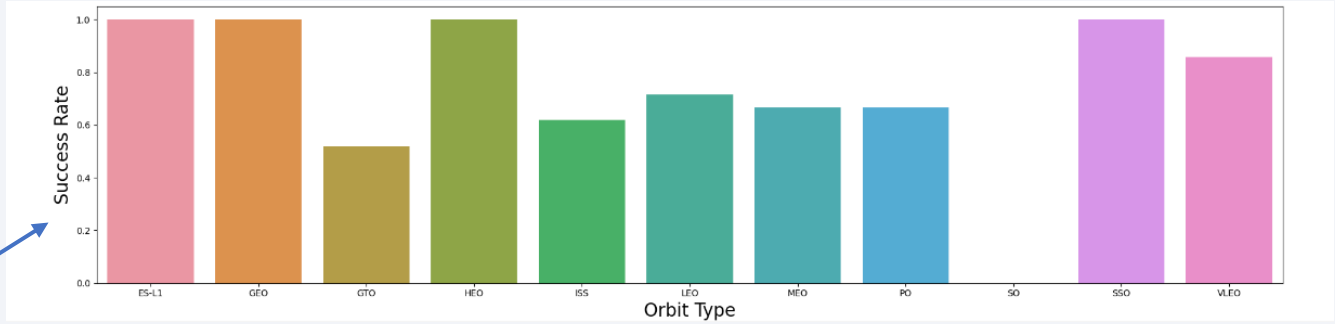


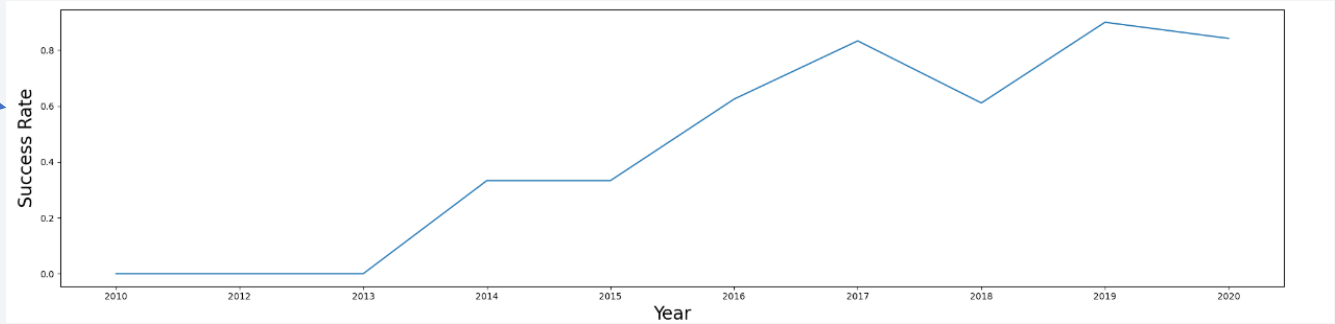
- Exploratory data analysis was conducted to determine the training labels.
- The calculation of the number of launches at each site and the occurrence of each orbit was performed.
- A landing outcome label was generated from the outcome column, *where 1 represents a successful rocket landing and 0 represents an unsuccessful landing.*
- The resulting data was exported to a CSV file.

EDA with Data Visualization

- The data was explored through visualization techniques to analyze various relationships.

  - Flight number vs. Launch site
  - Payload mass vs. Launch site
  - **Success rate of each orbit type**
  - Flight number vs. Orbit type
  - **Success rate yearly trend**





EDA with SQL

- The SpaceX dataset was seamlessly loaded into a PostgreSQL database within the Jupyter Notebook environment.
- By utilizing SQL for exploratory data analysis (EDA), insightful information was obtained from the dataset. Various queries were written to uncover specific details, such as:
  - Identifying the unique launch sites involved in space missions.
  - Calculating the total payload mass carried by boosters launched by NASA (CRS).
  - Determining the average payload mass carried by booster version F9 v1.1.
  - Counting the total number of successful and failed mission outcomes.
  - Retrieving details about failed landing outcomes on drone ships, including their associated booster versions and launch site names.

Build an Interactive Map with Folium

- Launch sites were visually marked on a folium map, and map objects such as markers, circles, and lines were added to indicate the success or failure of launches for each site.
- Launch outcomes (failure or success) were assigned to class 0 and 1, respectively, with 0 representing failure and 1 representing success.
- By utilizing color-labeled marker clusters, launch sites with relatively high success rates were identified.
- The distances between launch sites and their surrounding areas were calculated. This allowed for answering questions such as:
  - Are launch sites located near railways, highways and coastlines?
  - Do launch sites maintain a certain distance from cities?

Build a Dashboard with Plotly Dash

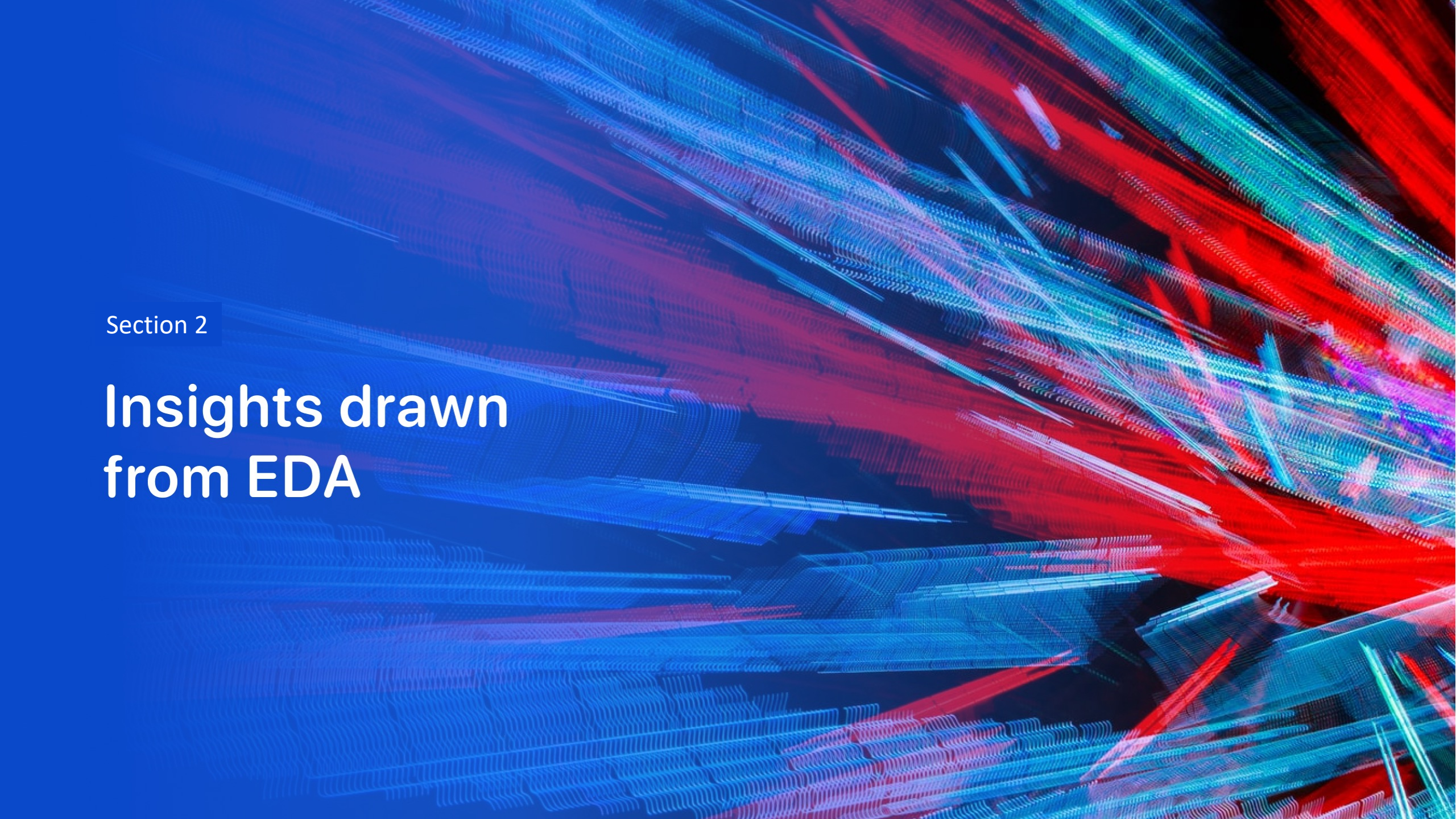
- An interactive dashboard was developed using Plotly Dash.
- Pie charts were utilized to visualize the total launches for specific sites.
- Scatter graphs were created to showcase the relationship between outcome and payload mass (in kilograms) for different booster versions.

Predictive Analysis (Classification)

- The dataset was split into training and testing sets using the train-test split method.
- Various machine learning models, including:
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Decision Trees
  - K-Nearest Neighbors (KNN)
- Hyperparameters of the models were fine-tuned using GridSearchCV.
- The models' performance was evaluated using accuracy as the chosen metric.

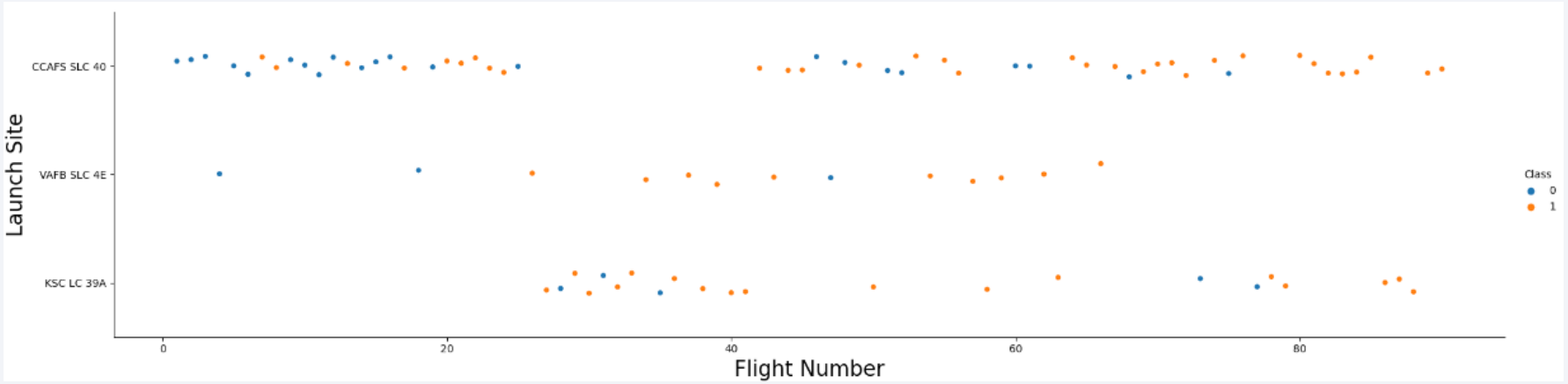
Results

- Successful landings have increased significantly since 2015, with a correlation found between landing outcomes and flight number.
- Launch sites are conveniently located near coastlines, aiding water-based rocket testing.
- Proximity to highways and railways enables efficient transportation of equipment and materials.
- Machine learning models achieved an 83.33% accuracy in predicting landing success, with potential for improvement by incorporating more data in future projects.


Section 2
Insights drawn from EDA

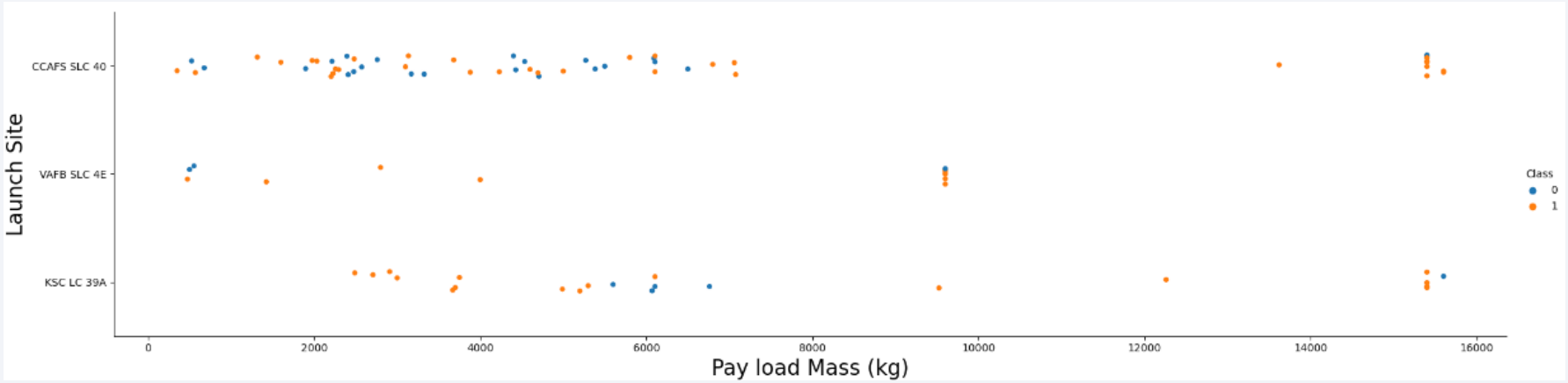
Flight Number vs. Launch Site

- The number of successful landings showed an upward trend as flight numbers increased. Additionally, the launch site CCAFS SLC 40 had the highest number of landing attempts, while VAFB SLC 4E had the lowest number of attempts.



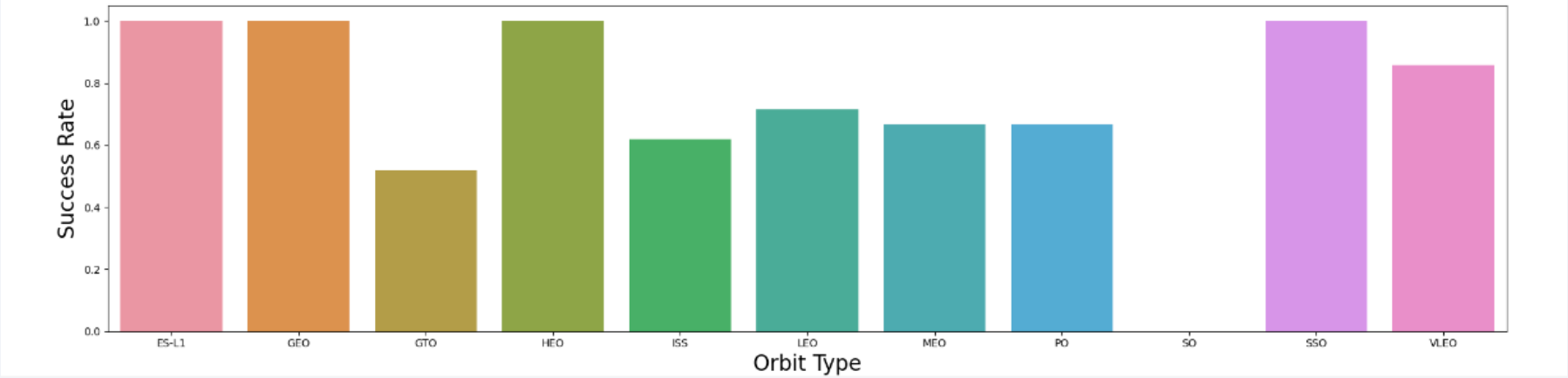
Payload vs. Launch Site

- At the VAFB-SLC launch site, no rockets were launched with a heavy payload mass exceeding 10,000. Furthermore, for the CCAFS SLC 40 launch site, a higher payload mass was associated with a greater success rate for the rockets.



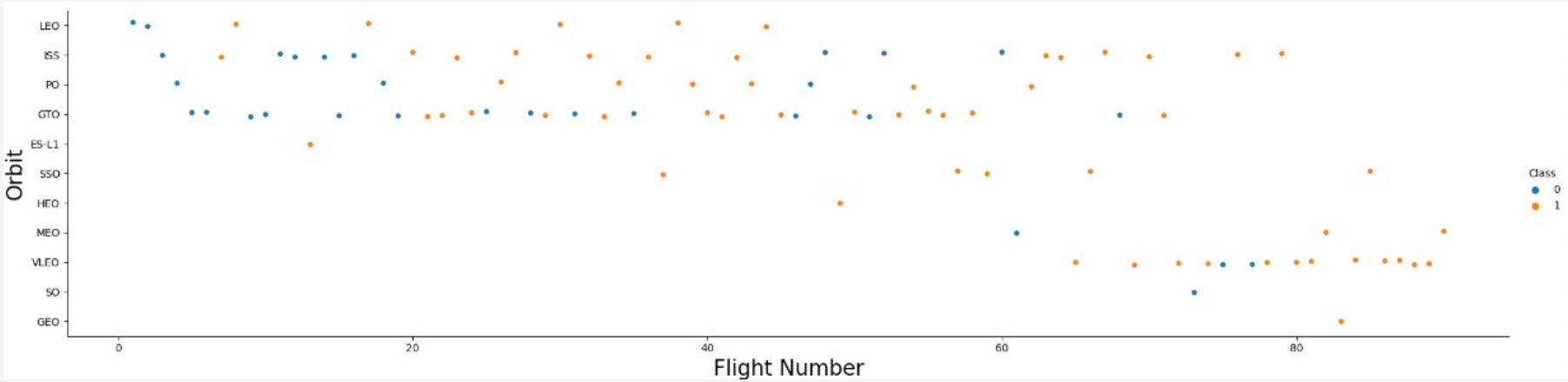
Success Rate vs. Orbit Type

- The bar chart of Success Rate vs. Orbit Type shows that the following orbits have the highest (100%) success rate: ES-L1, GEO, HEO and SSO.
- The orbit with the lowest (0%) success rate is: SO



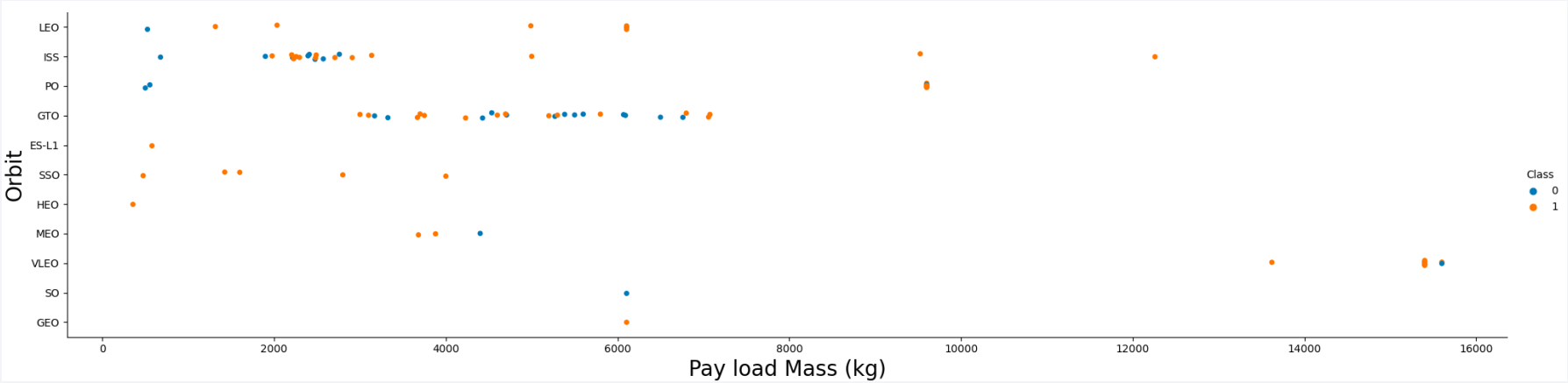
Flight Number vs. Orbit Type

- In the LEO orbit, there appears to be a relationship between the number of flights and the success rate. However, in the GTO orbit, no correlation is observed between the flight number and the success rate.



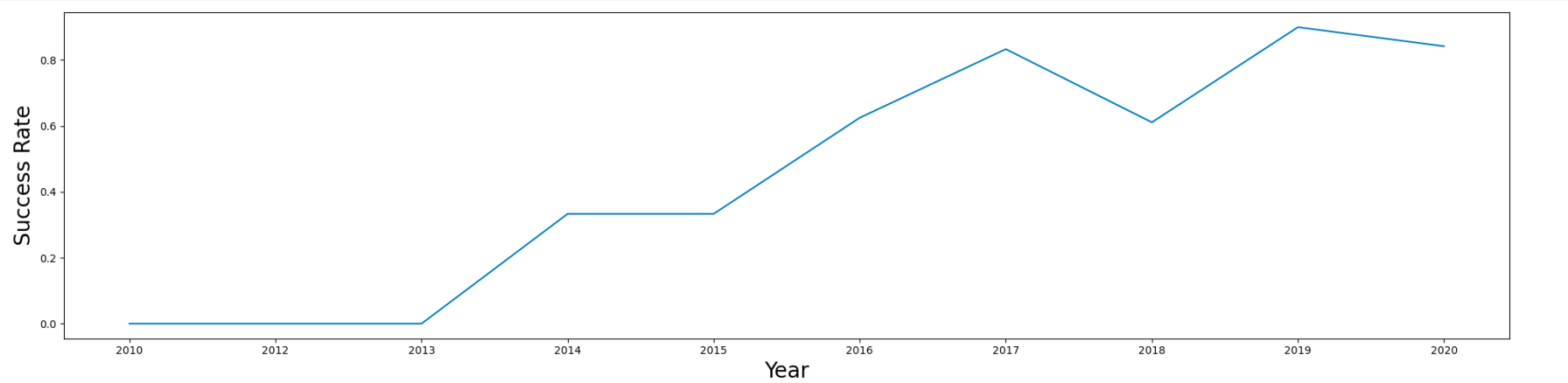
Payload vs. Orbit Type

- For heavy payloads, the success rate or positive landing rate is higher for launches in Polar, LEO, and ISS trajectories.
- However, in the case of GTO, it is difficult to differentiate the success rate clearly as both positive landings and negative landings (unsuccessful missions) are observed.



Launch Success Yearly Trend

- Over the span of seven years, from 2013 to 2020, the success rate exhibited a steady increase, reflecting significant advancements in mission execution and reliability.



All Launch Site Names

- The "UNIQUE" keyword retrieves only unique values from the Launch_Site column of the SPACEXTBL table, eliminating any duplicate entries.

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

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

Launch Site Names Begin with 'CCA'

- The "LIMIT 5" clause retrieves only 5 records from the query result, while the "LIKE" keyword, with the wildcard 'CCA%', is used to retrieve string values that begin with 'CCA' from the corresponding column.

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO
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)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)

Total Payload Mass

- The "SUM" keyword calculates the total of the LAUNCH column, while the "SUM" keyword with the associated condition filters the results to only include boosters from NASA (CRS).

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

SUM("PAYLOAD_MASS__KG_")
45596.0

Average Payload Mass by F9 v1.1

- The "AVG" keyword calculates the average of the PAYLOAD_MASS_KG column, while the "WHERE" keyword with the associated condition filters the results to only include the F9 v1.1 booster version.

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The "MIN" keyword calculates the minimum value of the DATE column, representing the earliest date. The "WHERE" keyword, along with the associated condition, filters the results to include only successful ground pad landings.

```
%%sql
SELECT MIN("Date") FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

MIN("Date")
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- The "WHERE" keyword is employed to filter the results based on the satisfaction of both conditions within the brackets, utilizing the "AND" keyword for conjunction. The "BETWEEN" keyword is utilized to select values within the range of 4000 to 6000, inclusive.

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
05/06/2016	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696.0	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14/08/2016	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600.0	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300.0	GTO	SES	Success	Success (drone ship)
10/11/2017	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200.0	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The "COUNT" keyword calculates the total number of mission outcomes, while the "GROUP BY" keyword is utilized to group these results based on the type of mission outcome.

```
%%sql
SELECT COUNT("Mission_Outcome") FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE 'Success%'
OR "Mission_Outcome" LIKE 'Failure%';
```

COUNT("Mission_Outcome")
101

Boosters Carried Maximum Payload

- A subquery is utilized in this scenario. The SELECT statement within the brackets identifies the maximum payload, and this value is employed in the subsequent WHERE condition.

```
%%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

2015 Launch Records

- The "WHERE" keyword is utilized to filter the results, specifically for failed landing outcomes, and further limited to the year 2015. This combination of conditions ensures that only the relevant data for failed landings in the specified year is included.

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

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

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

- The "WHERE" keyword, in conjunction with the "BETWEEN" keyword, is utilized to filter the results based on the specified date range. The filtered results are then grouped using the "GROUP BY" keyword and ordered using the "ORDER BY" keyword, with "DESC" specifying the descending order.

```
%%sql
SELECT
    CASE
        WHEN "Landing_Outcome" LIKE '%No attempt%' THEN 'No attempt'
        ELSE "Landing_Outcome"
    END AS "Modified_Landing_Outcome",
    COUNT(*) AS "Count"
FROM SPACEXTBL
WHERE "Date" BETWEEN '04/06/2010' AND '20/03/2017'
GROUP BY "Modified_Landing_Outcome"
ORDER BY "Count" DESC;
```

Modified_Landing_Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2

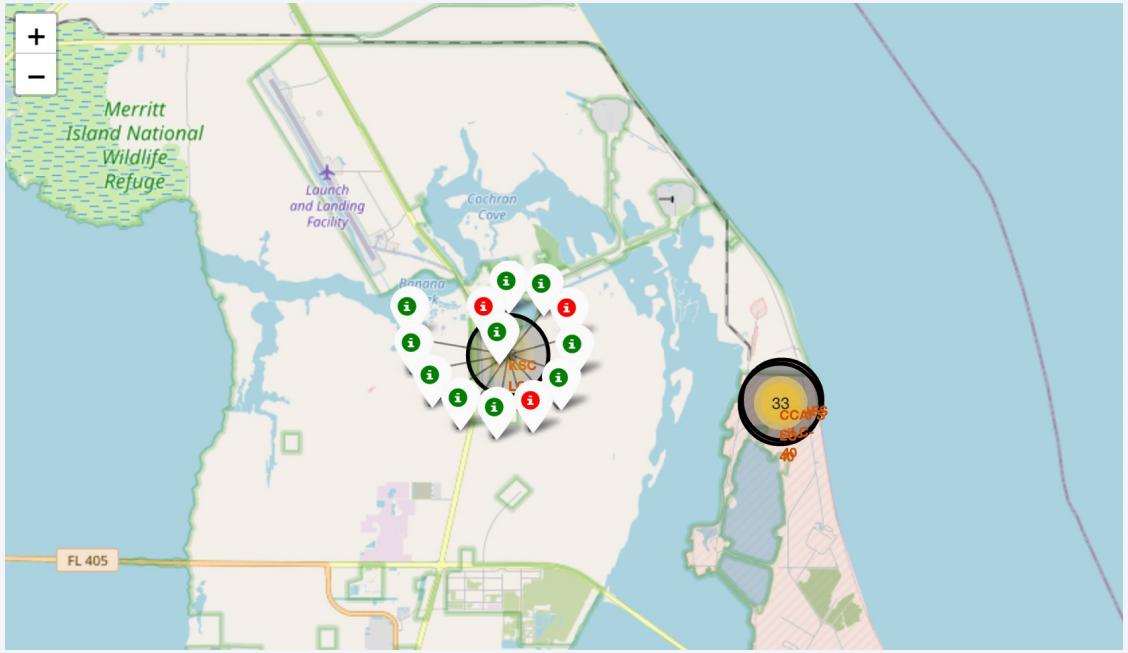

Section 3
Launch Sites
Proximities Analysis

All Launch Sites on a Map

- It is evident that all launch sites are located in close proximity to the coast and are several thousand kilometers away from the equator line.

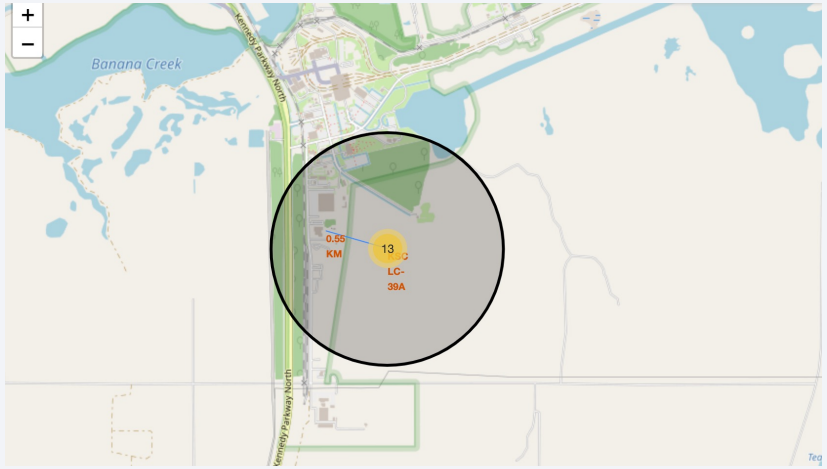
Success/Failed Launches

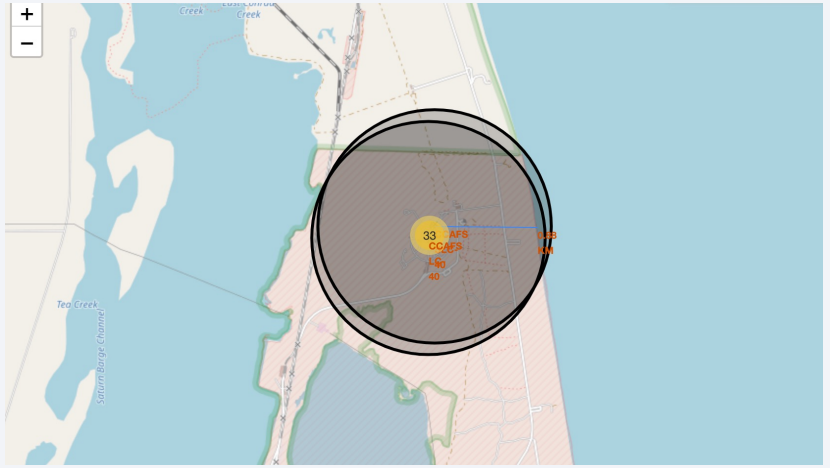
- Successful rocket launches are denoted by green markers, whereas failed launches are represented by red markers.
- KSC LC-39A exhibits the highest success rate among all launch sites

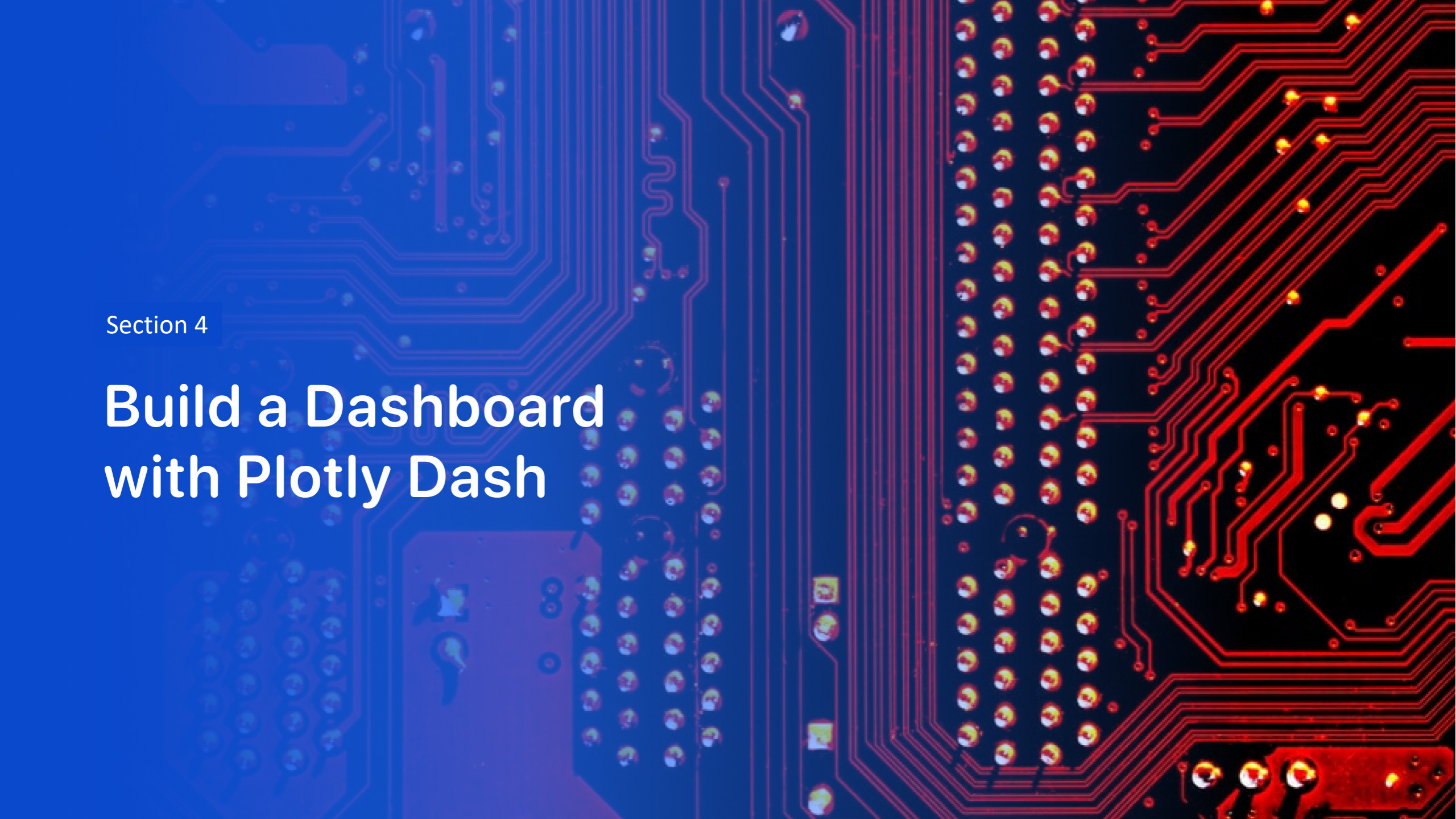


Proximity of Launch Sites to Landmarks

- The close proximity of launch sites to railways and highways is apparent, likely due to the necessary transportation requirements for rocket parts.
- The sites' proximity to coastlines is evident from the multitude of rocket landing tests conducted on water bodies like the ocean.

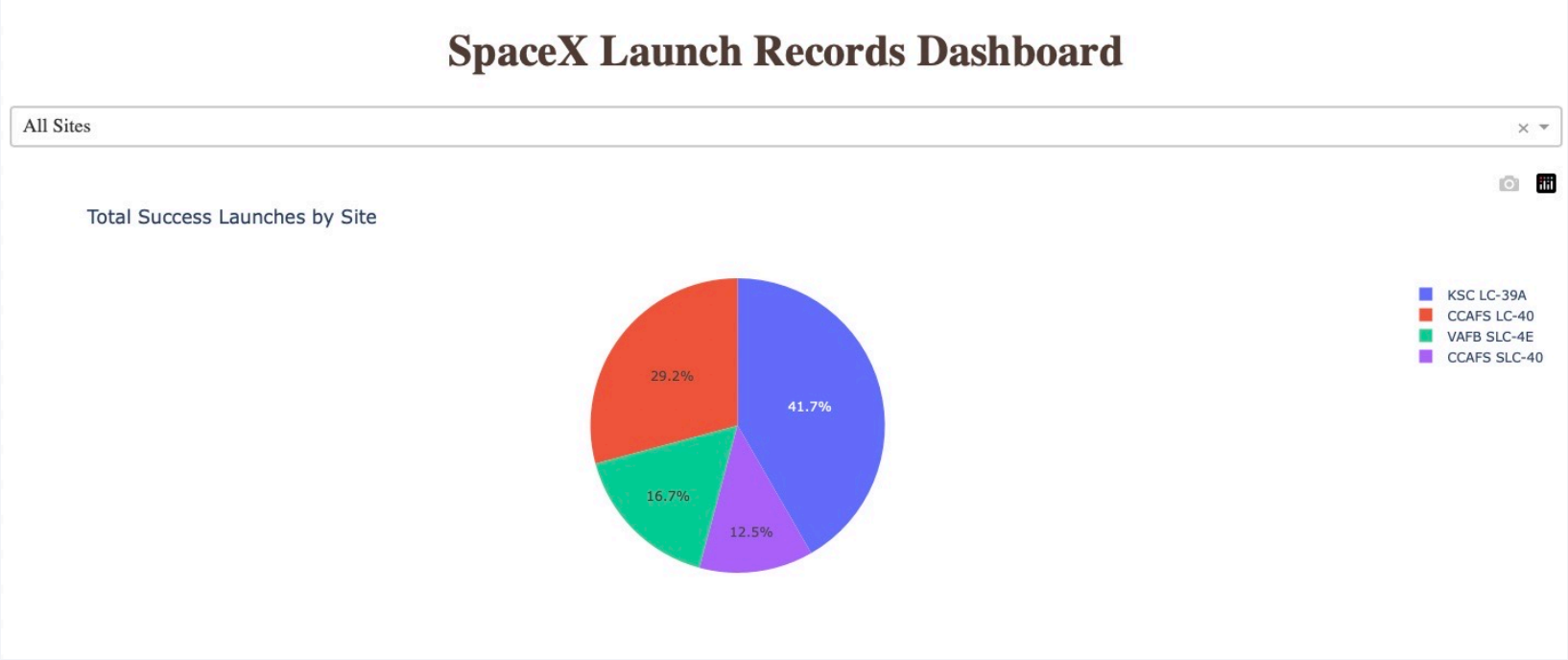





Section 4
Build a Dashboard with Plotly Dash

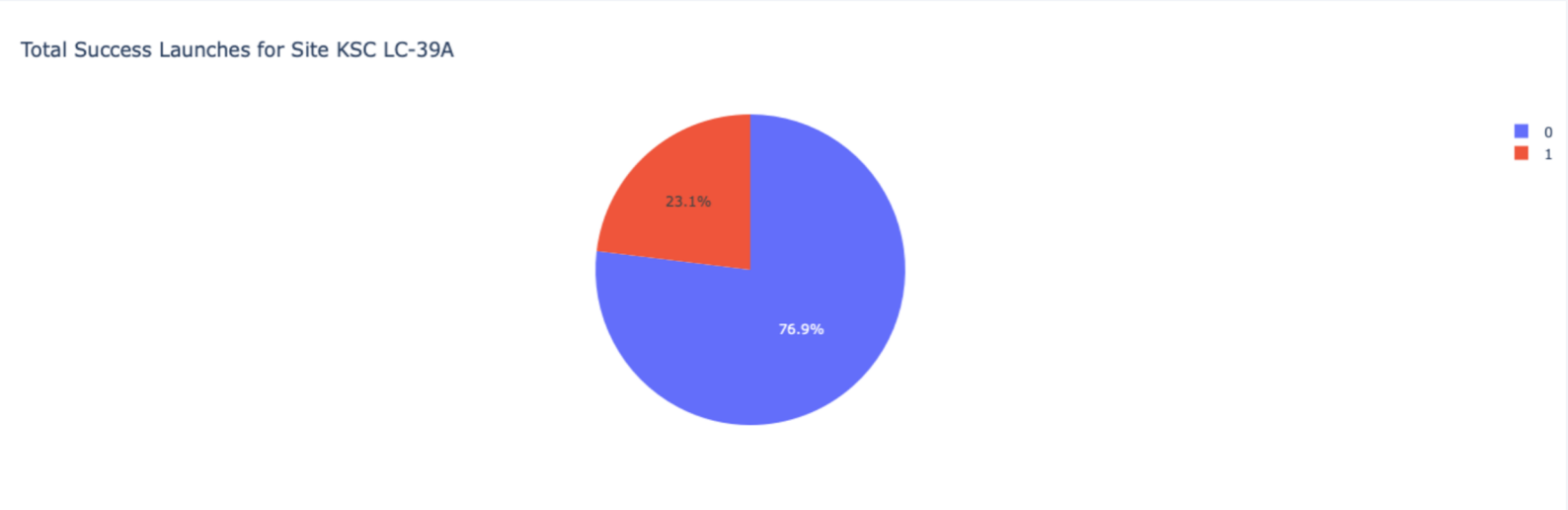
Total Successful Launches by Site

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.



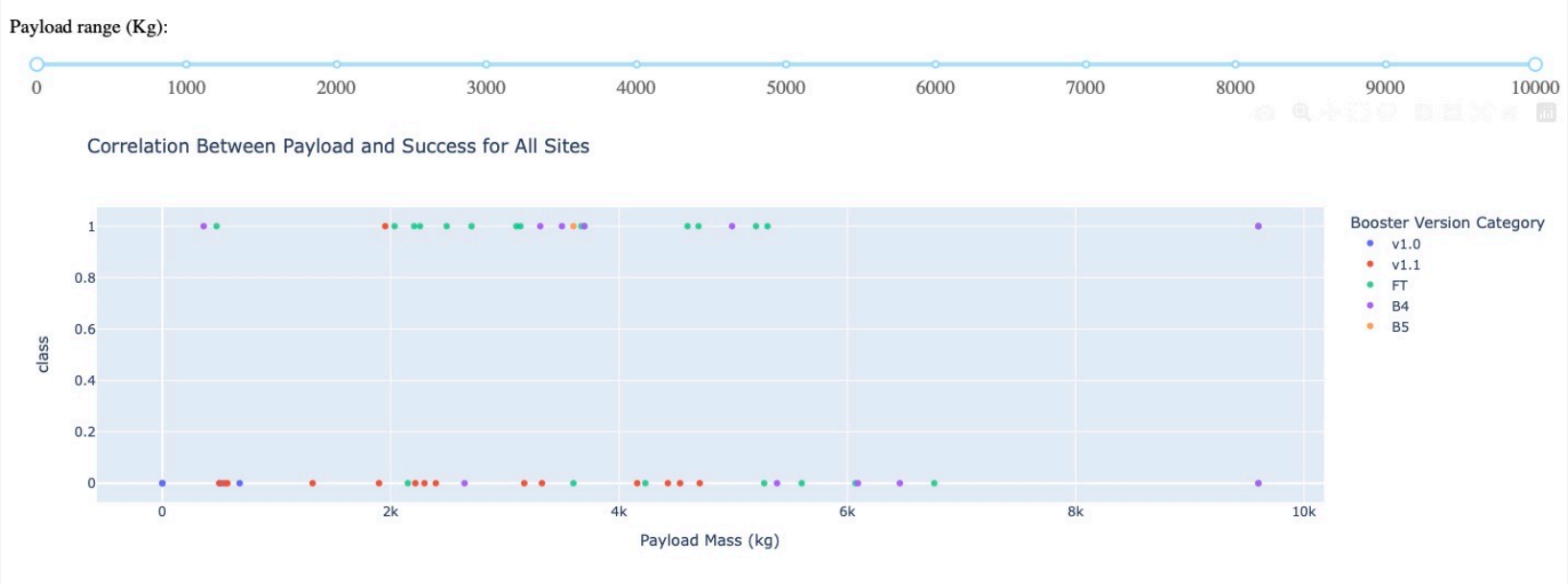
Total Successful Launches for Site KSC LC-39A

- The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.



<Dashboard Screenshot 3>

- The payload range between 2000 kg and 4000 kg shows the highest success rate.
- The booster version "FT" exhibits a higher success rate compared to other versions.

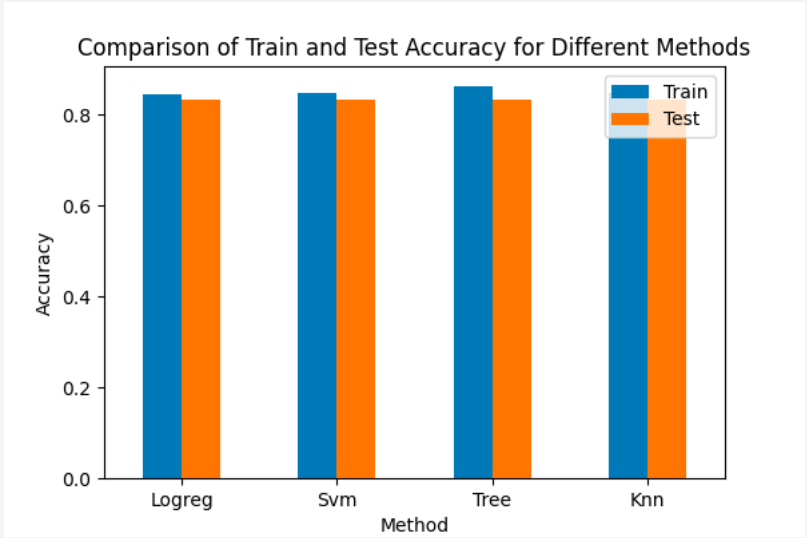


Section 5

Predictive Analysis (Classification)

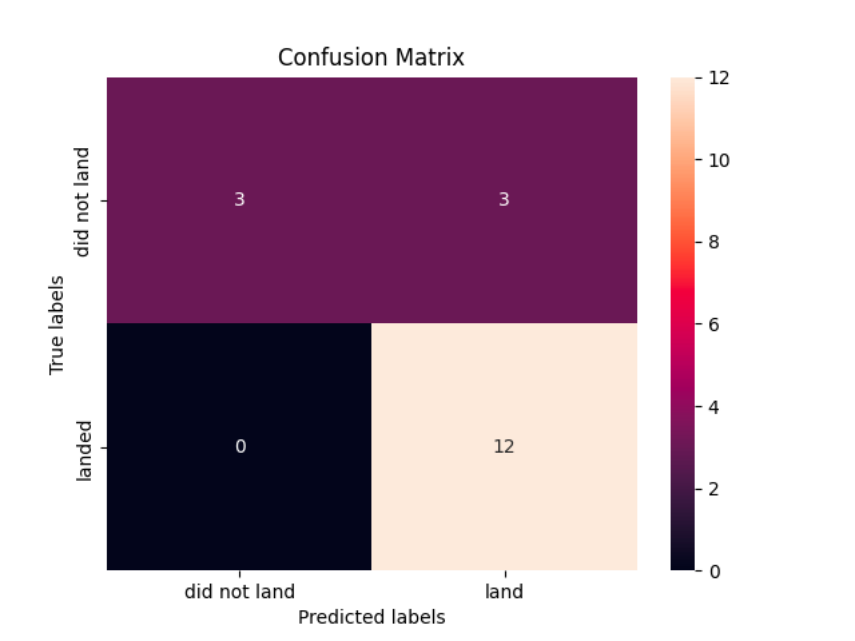
Classification Accuracy

- Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result:
  - The Decision Tree model has the highest classification accuracy
  - The Accuracy Score is 83.33%
  - The Best Score is 86.43%



Method	Train Accuracy	Test Accuracy
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.864286	0.833333
Knn	0.848214	0.833333

Confusion Matrix



- The confusion matrix indicates that out of the total 18 results, 3 are misclassified as false positives (represented in the top-right corner).
- However, the remaining 15 results are accurately classified, with 3 correctly identified as not landing and 12 correctly identified as landing.

Conclusions

- As the number of flights increases, the success rate at a launch site also increases, indicating that experience contributes to improved success rates over time.
- From 2013 to 2020, there was a consistent rise in the success rate, demonstrating notable progress in mission execution and reliability over a seven-year timeframe.
- Certain orbit types, such as ES-L1, GEO, HEO, and SSO, exhibit a 100% success rate, with the latter achieving 5 successful flights, indicating their reliability.
- The success rate for massive payloads (over 4000 kg) is lower compared to lower payloads.
- The launch site KSC LC-39A stands out with the highest number of successful launches (41.7% of the total) and the highest success rate (76.9%).
- The Decision Tree model outperforms other classification models, achieving an accuracy of 83.33%.

Thank you!