IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rohith Mohan
26-03-2024

https://github.com/ROHITHKM92

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of methodologies**

Data Collection via API, Web Scraping:	• Gathering data from various sources using APIs and web scraping techniques.
Data Wrangling:	• Cleaning and organizing collected data to prepare it for analysis.
Exploratory Data Analysis with Data Visualization:	• Using visualizations to uncover patterns and relationships within the data.
Exploratory Data Analysis with SQL:	• Delving deeper into the dataset using SQL queries to extract valuable insights.
Building Interactive Maps with Folium:	• Generating dynamic, interactive maps to visualize spatial data effectively.
Building Dashboards with Plotly Dash:	• Creating interactive dashboards for comprehensive data representation and exploration.
Predictive Analysis (Classification):	• Implementing classification algorithms to predict outcomes based on the data.

- **Summary of all results**

Exploratory Data Analysis Results:	• **Key findings and insights derived from exploratory data analysis.**
Interactive Maps and Dashboards:	• **Presentation of spatial data and insights through interactive maps and dashboards.**
Predictive Results:	• **Outcomes and accuracies of predictive models implemented for classification tasks.**

Introduction

Project Background and Context:

- The objective of this project is to forecast the successful landing of the Falcon 9 first stage. SpaceX asserts on its website that the Falcon 9 rocket launch costs $62 million, whereas other providers' costs exceed $165 million each. This significant price gap stems from SpaceX's ability to reuse the first stage. By determining the likelihood of the stage landing successfully, we can accurately estimate the cost of a launch. This insight holds value for competitors aiming to challenge SpaceX in the rocket launch market.

Issues to Address:

- What are the primary characteristics distinguishing successful and failed landings?
- How do various relationships among rocket variables affect the outcome of a landing?
- What conditions are conducive to maximizing SpaceX's landing success rate?

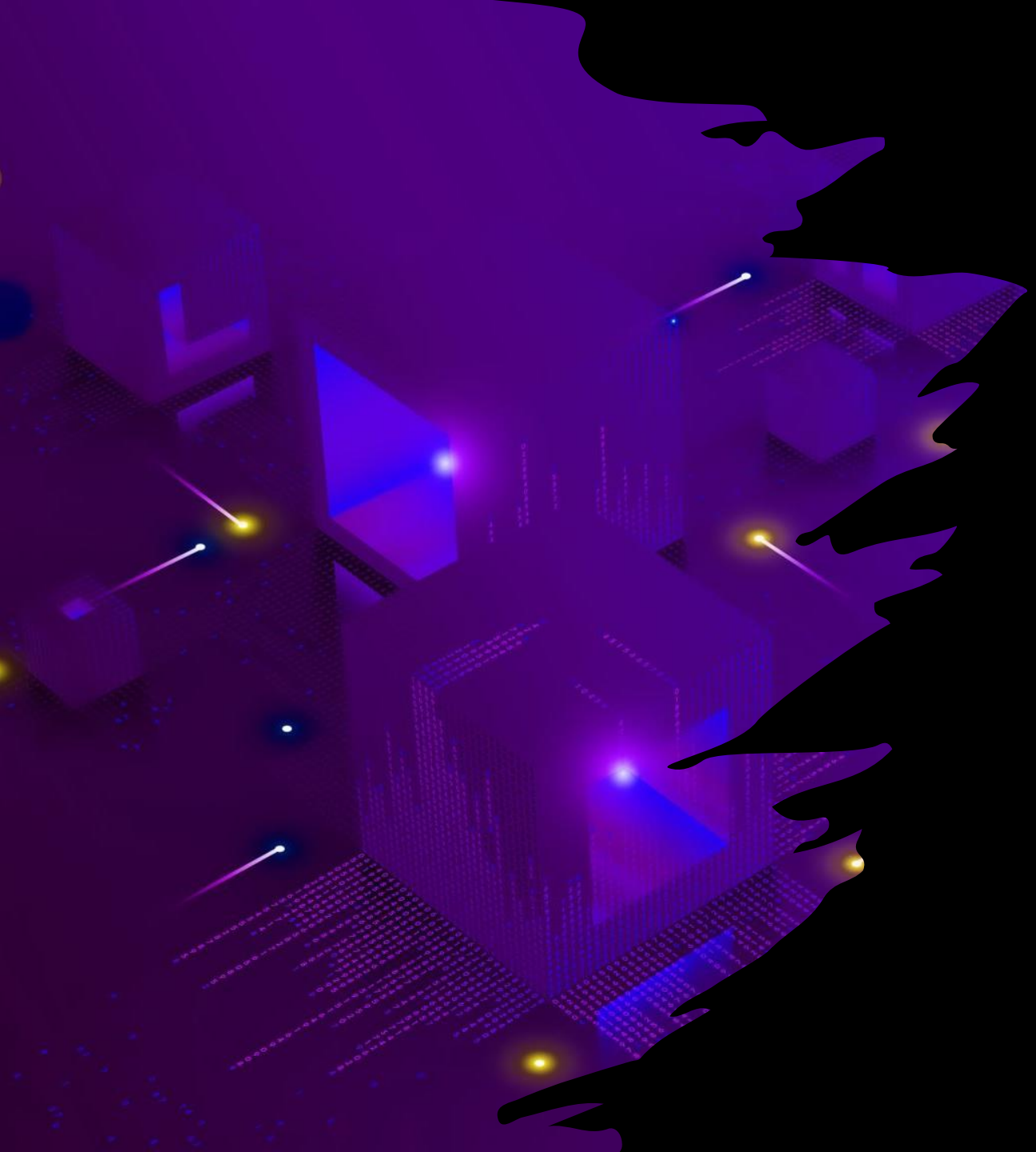
Section 1

Methodology

Methodology

Executive Summary

- ❑ Data collection methodology:
  - Utilization of SpaceX REST API
  - Web Scrapping from Wikipedia
- ❑ Perform data wrangling
  - Removal of Redundant Columns
  - Implementing One Hot Encoding for Classification Models
- ❑ 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

- **Dynamic Methods:** Utilized a fusion of API requests from SpaceX's REST API and Web Scraping techniques from the SpaceX Wikipedia entry.
- **Comprehensive Approach:** Employed both methodologies to ensure thorough data acquisition, enabling a robust and detailed analysis of SpaceX launches.

- The data collected via the SpaceX REST API included columns such as ***FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.***

- On the other hand, data obtained through Wikipedia web scraping encompassed columns like ***Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.***

Data Collection

- **Data sets are gathered from both the REST SpaceX API and web scraping Wikipedia.**

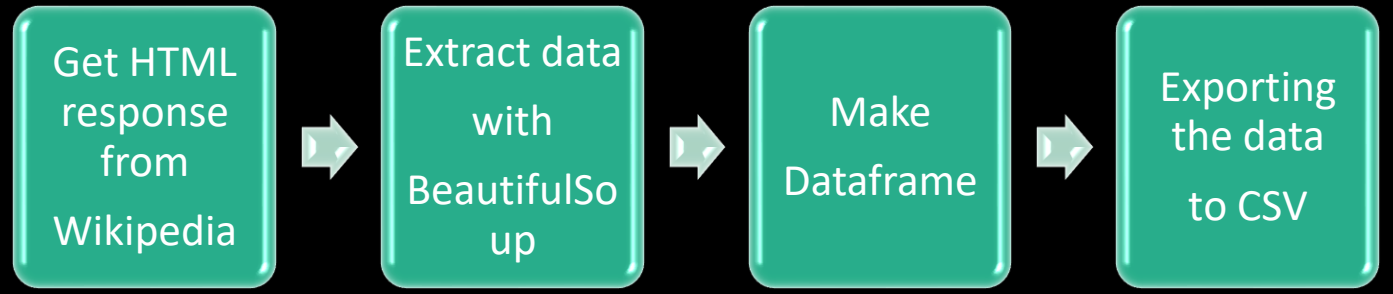
➢ The API provides information regarding rockets, launches, and payload details.

The Space XREST API URL is *api.spacexdata.com/v4/*

GitHub Link: SpaceX Data Collection API

➢ The data acquired through web scraping Wikipedia includes details on launches, landings, and payload information.

Wikipedia URL is *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*



GitHub Link: Data Collection - web scraping

Data Collection – SpaceX API

1. Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

2. convert the JSON result into a dataframe

```
data = pd.json_normalize(response.json())
```

3. Transform data

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

4. Create dictionary with data

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Create dataframe

```
data2 = pd.DataFrame(launch_dict)
```

6. Filter dataframe

```
data_falcon9 = data2[data2['BoosterVersion']=='Falcon 9']
```

7. Export to file

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

[GitHub Link: SpaceX Data Collection API]

Data Collection - Scraping

1. Getting Response from HTML

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

2. Create BeautifulSoup Object

```
soup = BeautifulSoup(response.text, 'html')
```

3. Find all tables

```
html_tables = soup.find_all('table')
```

4. Get column names

```
tc = first_launch_table.find_all('th')
for th in tc:
    name = extract_column_from_header(th)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

5. Create dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Add data to keys

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('ta
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as nu
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
```

See notebook for the rest of code

7. Create dataframe from dictionary

```
df=pd.DataFrame(launch_dict)
```

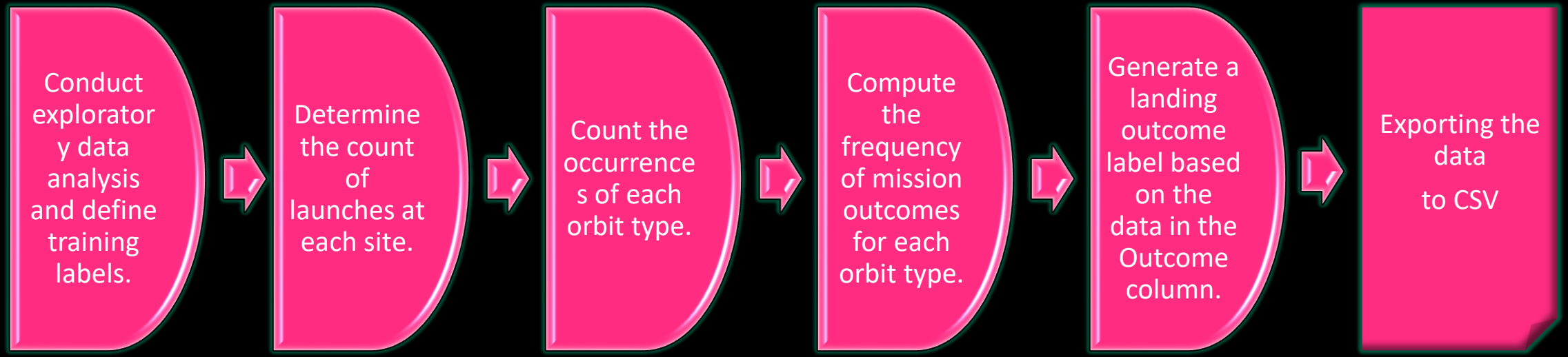
8. Export to file

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

[GitHub Link: Data Collection - web scraping]

Data Wrangling

The dataset records various instances of unsuccessful booster landings, categorized by landing location (ocean, ground pad, or drone ship). True Ocean indicates successful ocean landings, False Ocean signifies unsuccessful ones. Similarly, True RTLS denotes successful ground pad landings, while False RTLS indicates unsuccessful attempts. True ASDS signifies successful drone ship landings, whereas False ASDS represents failed attempts. These outcomes are converted into Training Labels: "1" denotes successful landings, while "0" denotes unsuccessful ones.

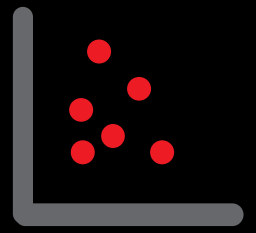


GitHub Link: Data Wrangling

EDA with Data Visualization

Scatter plots

Scatter plots depict the correlation between variables, such as the flight number versus payload mass, flight number versus launch site, payload versus launch site, orbit versus flight number, payload versus orbit type, and orbit versus payload mass. These plots illustrate the relationship between different factors.



Bar Graph

Bar graphs display the association between a numerical variable, such as success rate, and a categorical variable, such as orbit type.

Line Graph

Line graphs, like "Success rate vs. Year," visually represent data variables and their trends over time. They provide insights into the overall behavior of the data and facilitate predictions for unseen data points.

[GitHub Link: EDA with Data Visualization]

EDA with SQL

The following SQL queries were performed:

- Retrieving the names of the unique launch sites involved in the space missions.
- Displaying five records where launch sites start with the string 'CCA'.
- Showing the total payload mass carried by boosters launched by NASA (CRS).
- Presenting the average payload mass carried by booster version F9 v1.1.
- Identifying the date when the first successful landing outcome on a ground pad was achieved.
- Listing the names of boosters that successfully landed on a drone ship with a payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failed mission outcomes.
- Listing the names of booster versions that carried the maximum payload mass.
- Listing the failed landing outcomes on a drone ship, along with their booster versions and launch site names for the months in the year 2015.
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

GitHub Link: EDA with SQL

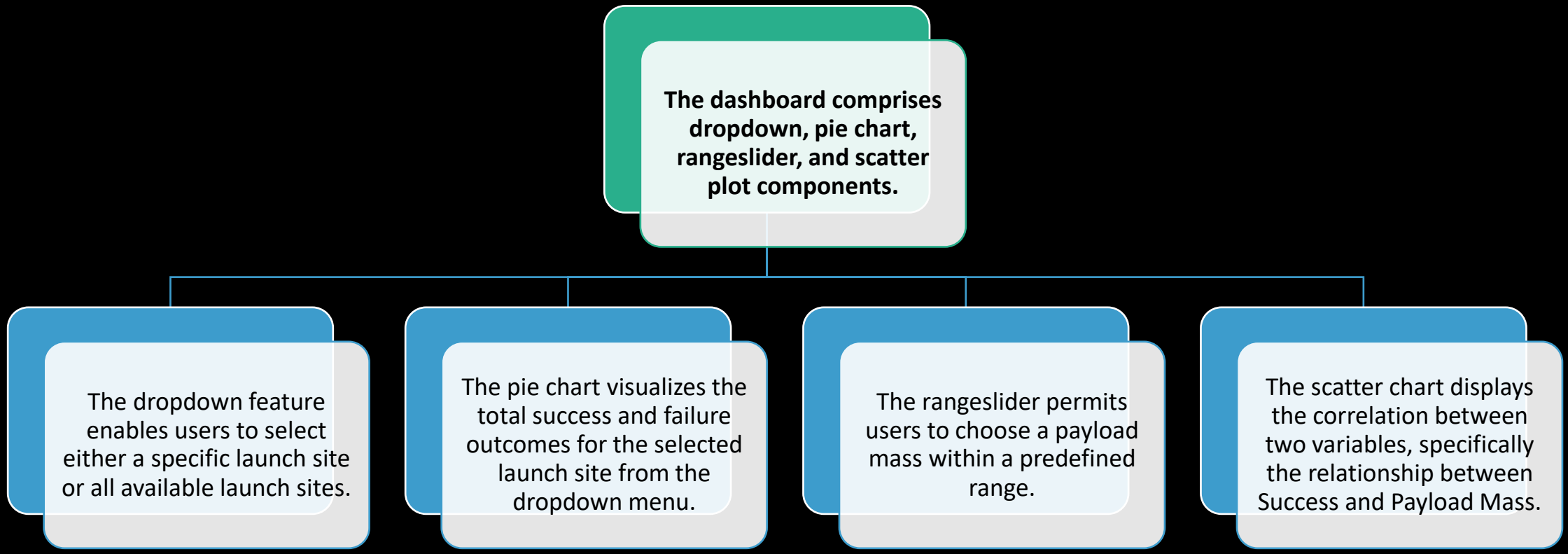
Build an Interactive Map with Folium

- **Markers indicating all Launch Sites:**
  - Integrated a Circle Marker along with Popup Label and Text Label representing NASA Johnson Space Center, utilizing its latitude and longitude coordinates as the starting location.
  - Implemented Markers with Circle, Popup Label, and Text Label for all Launch Sites, displaying their geographical positions and their proximity to the Equator and coastlines through their respective latitude and longitude coordinates.
- **Colored Markers depicting launch outcomes for each Launch Site:**
  - Introduced colored Markers to signify success (Green) and failure (Red) launches, utilizing Marker Cluster to discern Launch Sites with notably high success rates.
- **Distances between Launch Sites and their surroundings:**
  - Included colored Lines to illustrate distances between Launch Site KSC LC-39A (as an exemplar) and its proximities such as Railway, Highway, Coastline, and Closest City.

GitHub Link: Interactive Map with Folium

Build a Dashboard with Plotly Dash

GitHub Link: Dashboard with Plotly Dash



Predictive Analysis (Classification)

GitHub Link: Predictive Analysis

Data Preparation:	Load dataset. Normalize the data. Divide the data into training and test sets.
Model Preparation:	Choose machine learning algorithms. Define parameters for each algorithm using GridSearchCV. Train GridSearchModel models with the training dataset.
Model Evaluation:	Obtain the best hyperparameters for each type of model. Calculate accuracy for each model using the test dataset. Visualize the Confusion Matrix.
Model Comparison:	Compare models based on their accuracy. Select the model with the highest accuracy (refer to Notebook for results).

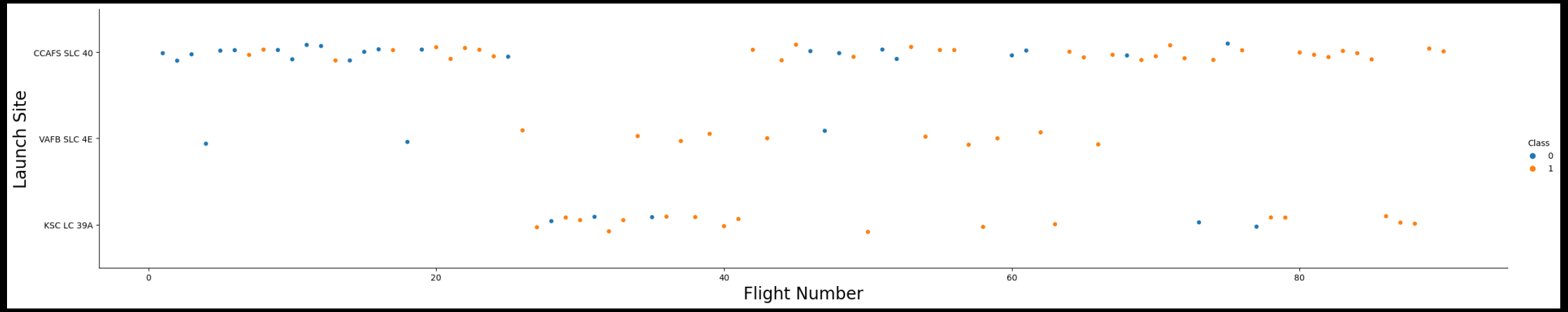
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

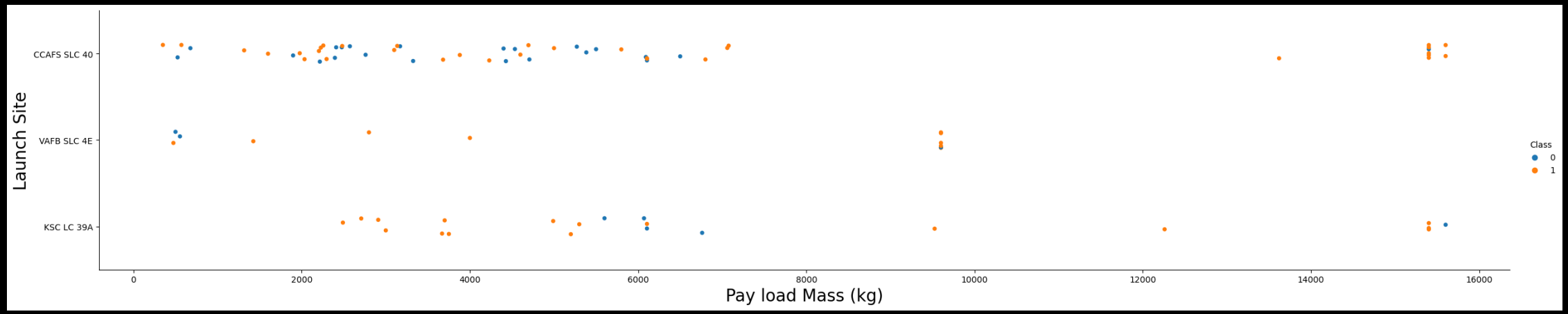
Flight Number vs. Launch Site



Explanation:

- Initial flights experienced failure, whereas recent ones achieved success consistently.
- Approximately half of all launches occurred at the CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A exhibit higher success rates.
- It is reasonable to infer that each successive launch demonstrates an improved success rate.

Payload vs. Launch Site



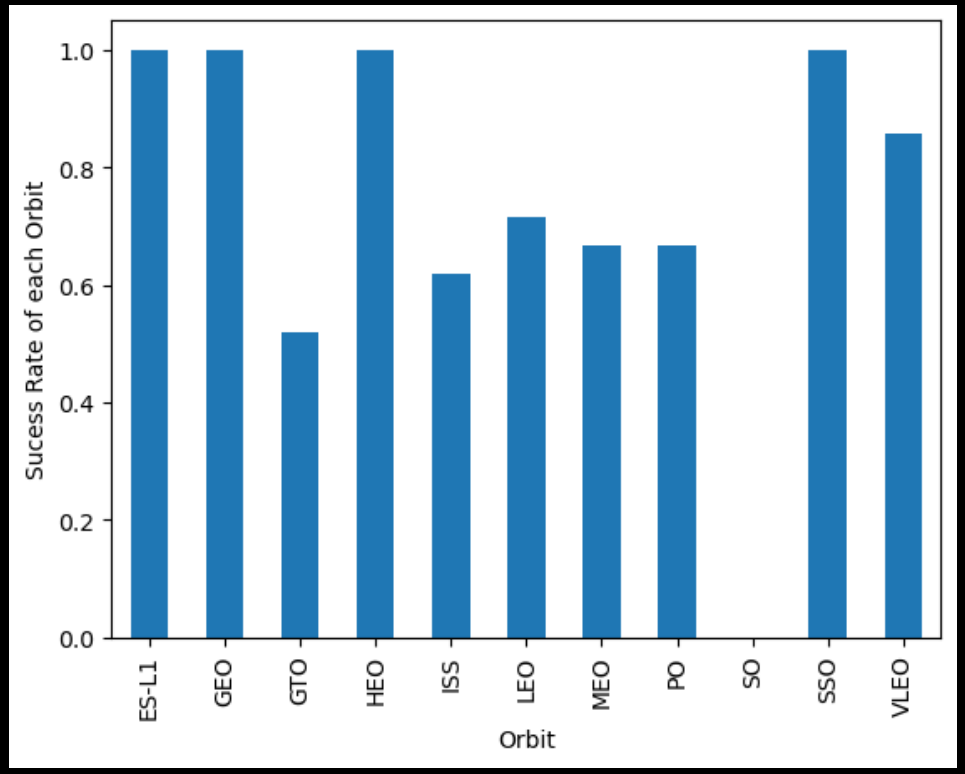
Explanation:

- Payloads exceeding 9,000kg, roughly equivalent to the weight of a school bus, demonstrate an exceptional success rate.
- Payloads surpassing 12,000kg appear feasible solely at the CCAFS SLC 40 and KSC LC 39A launch sites.

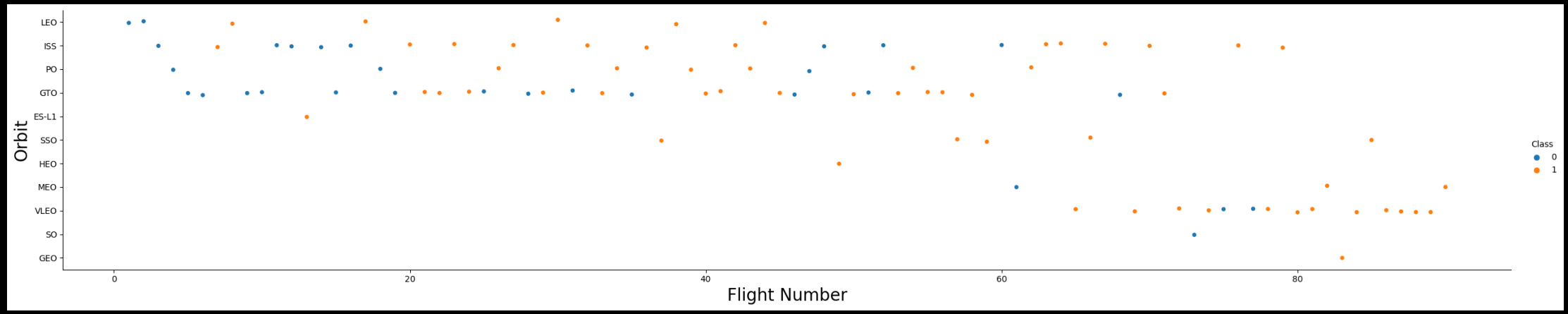
Success Rate vs. Orbit Type

Explanation:

- ❑ Orbits achieving a 100% success rate include ES-L1, GEO, HEO, and SSO.
- ❑ Orbits with a 0% success rate consist of SO.
- ❑ Orbits with success rates ranging between 50% and 85% encompass GTO, ISS, LEO, MEO, and PO.



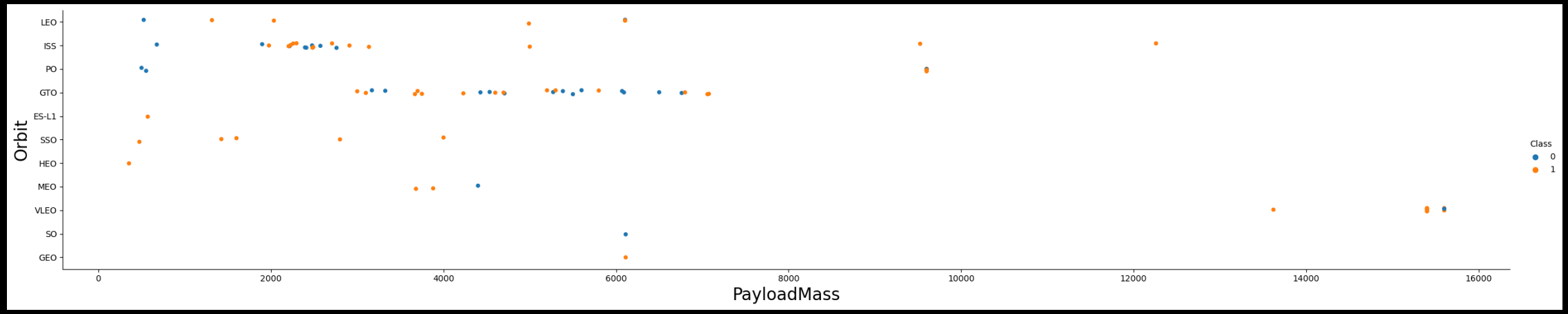
Flight Number vs. Orbit Type



Explanation:

- We observe a correlation between the success rate and the number of flights for the LEO orbit, indicating that the success rate tends to increase as the number of flights rises. For certain orbits like GTO, there appears to be no discernible relationship between the success rate and the number of flights. However, it is plausible to assume that the high success rate of orbits such as SSO or HEO is attributed to knowledge gained from previous launches for other orbits.

Payload vs. Orbit Type



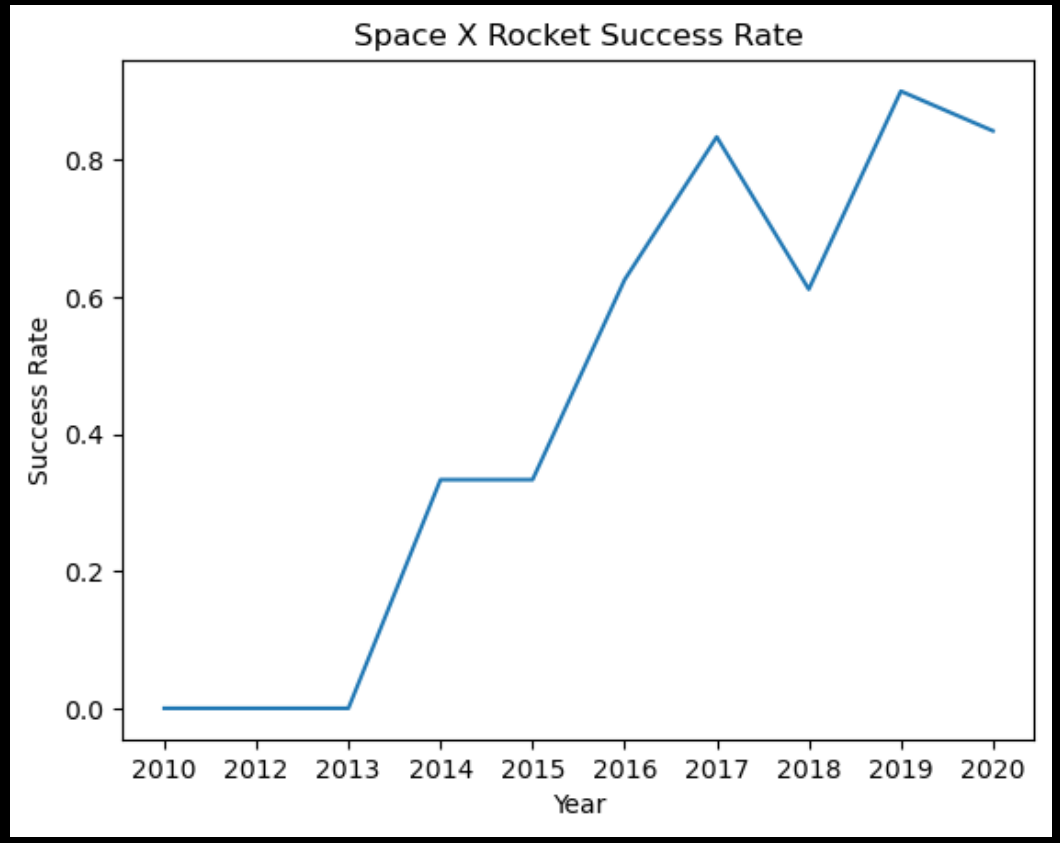
Explanation:

- The weight of payloads can significantly impact the success rate of launches in specific orbits. For instance, heavier payloads enhance the success rate for the LEO orbit. Conversely, reducing the payload weight for a GTO orbit enhances the likelihood of a successful launch.

Launch Success Yearly Trend

Explanation:

- The success rate began to increase in 2013 and continued to rise until 2020.
- It appears that the initial three years constituted a period of adjustment and technological improvement.



All Launch Site Names

```
%sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

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

Explanation:

- These values are derived by selecting distinct instances of the "launch_site" attribute from the dataset.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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

Explanation:

- Showing five records where launch sites start with the string 'CCA'.

Total Payload Mass

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

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

SUM("PAYLOAD_MASS__KG_")
45596

Explanation:

- Presenting the cumulative payload mass transported by boosters launched by NASA under the CRS program.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

Explanation:

- This query computes the average payload mass for all instances where the booster version includes the substring "F9 v1.1".

First Successful Ground Landing Date

```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success%'
```

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

MIN("DATE")
2015-12-22

Explanation:

- Filtering the data by successful landing outcomes on the ground pad and retrieving the minimum date value enables identification of the first occurrence, which took place on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

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

Explanation:

- Enumerating the names of boosters that achieved success in landing on a drone ship with a payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

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

SUCCESS	FAILURE
100	1

Explanation:

- The initial SELECT statement presents the subqueries that yield results. The first subquery calculates the count of successful missions, while the second subquery calculates the count of unsuccessful missions. The WHERE clause, followed by the LIKE clause, filters the mission outcomes. The COUNT function tallies the records filtered accordingly.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

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

Explanation:

- We employed a subquery to filter the data by selecting only the payload with the maximum weight using the MAX function. The main query utilizes the results of the subquery and returns the unique booster versions (SELECT DISTINCT) with the maximum payload mass.

2015 Launch Records

```
%sql SELECT substr("DATE", 6, 2) AS MONTH,"Booster_Version", "Launch_Site" FROM SPACEXTABLE\
WHERE "Landing_Outcome" = 'Failure (drone ship)' and substr("DATE",0,5) = '2015'
```

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

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Explanation:

- Enumerating the failed landing outcomes on drone ships, along with their booster versions and launch site names, for the months in the year 2015.

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

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS Occurrences
FROM SPACEXTABLE
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY Occurrences DESC
```

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

Landing_Outcome	Occurrences
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

Explanation:

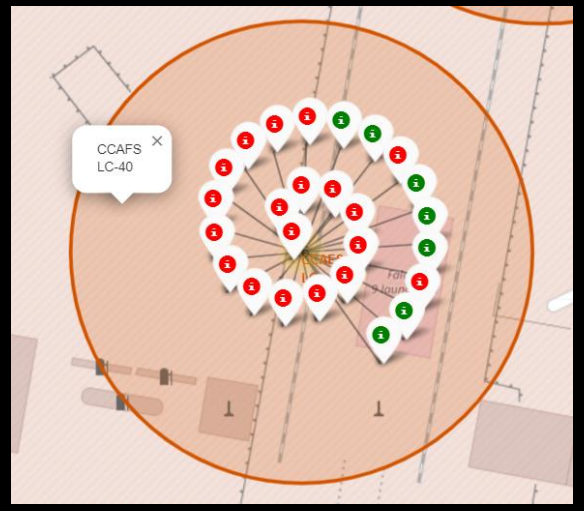
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates June 4, 2010, and March 20, 2017, in descending order.

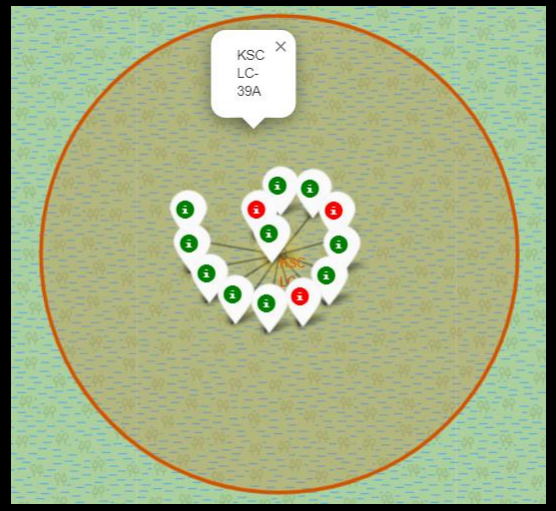

Section 3
Launch Sites
Proximities Analysis

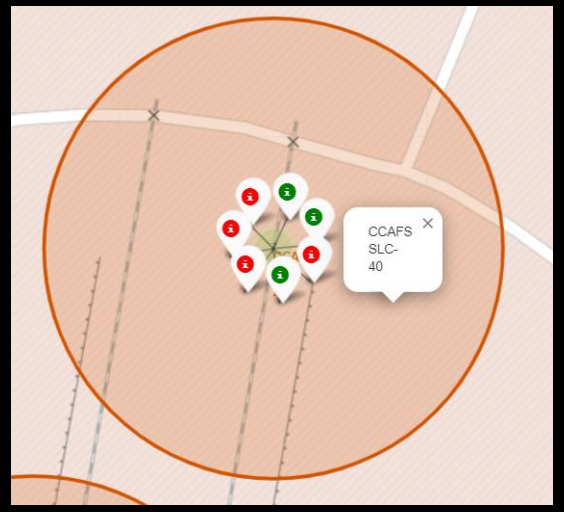
Markers displaying the locations of all launch sites on a global map.

Most launch sites are located near the Equator, where the Earth's rotation speed is highest, aiding spacecraft in achieving orbital velocity. Launching from the Equator enables spacecraft to maintain their pre-launch speed, enhancing their ability to sustain orbit due to inertia. Additionally, all launch sites are situated near coastlines, reducing the risk of debris falling or exploding near populated areas when rockets are launched towards the ocean.

Launch records on the map labeled with colors







The color-coded markers allow for quick identification of launch sites with notably high success rates: green markers signify successful launches, while red markers denote failed launches. Launch Site KSC LC-39A stands out with a notably high success rate.

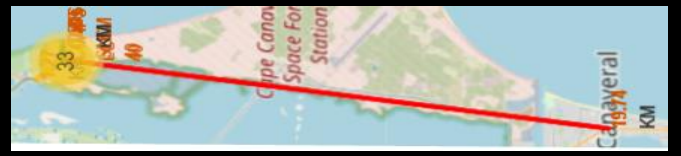
Distances between CCAFS SLC-40 and its proximities

Railway Proximity (1.27 km):
The launch site's relative closeness to a railway poses a potential danger as trains carrying hazardous materials could pass near the site. In the event of an accident or derailment, there is a risk of explosions or chemical leaks that could impact launch operations and personnel safety.

Highway Proximity (0.59 km):
The close proximity to a highway raises concerns about traffic congestion and accidents. In the event of a major incident such as a vehicle collision or hazardous material spill, emergency response efforts could be hindered, impacting both site operations and nearby communities.

Coastline Proximity (0.86 km):
Being near the coastline presents risks related to maritime activities such as shipping traffic and potential oil spills. Any incidents involving vessels in the vicinity could result in environmental contamination or logistical challenges for launch operations.

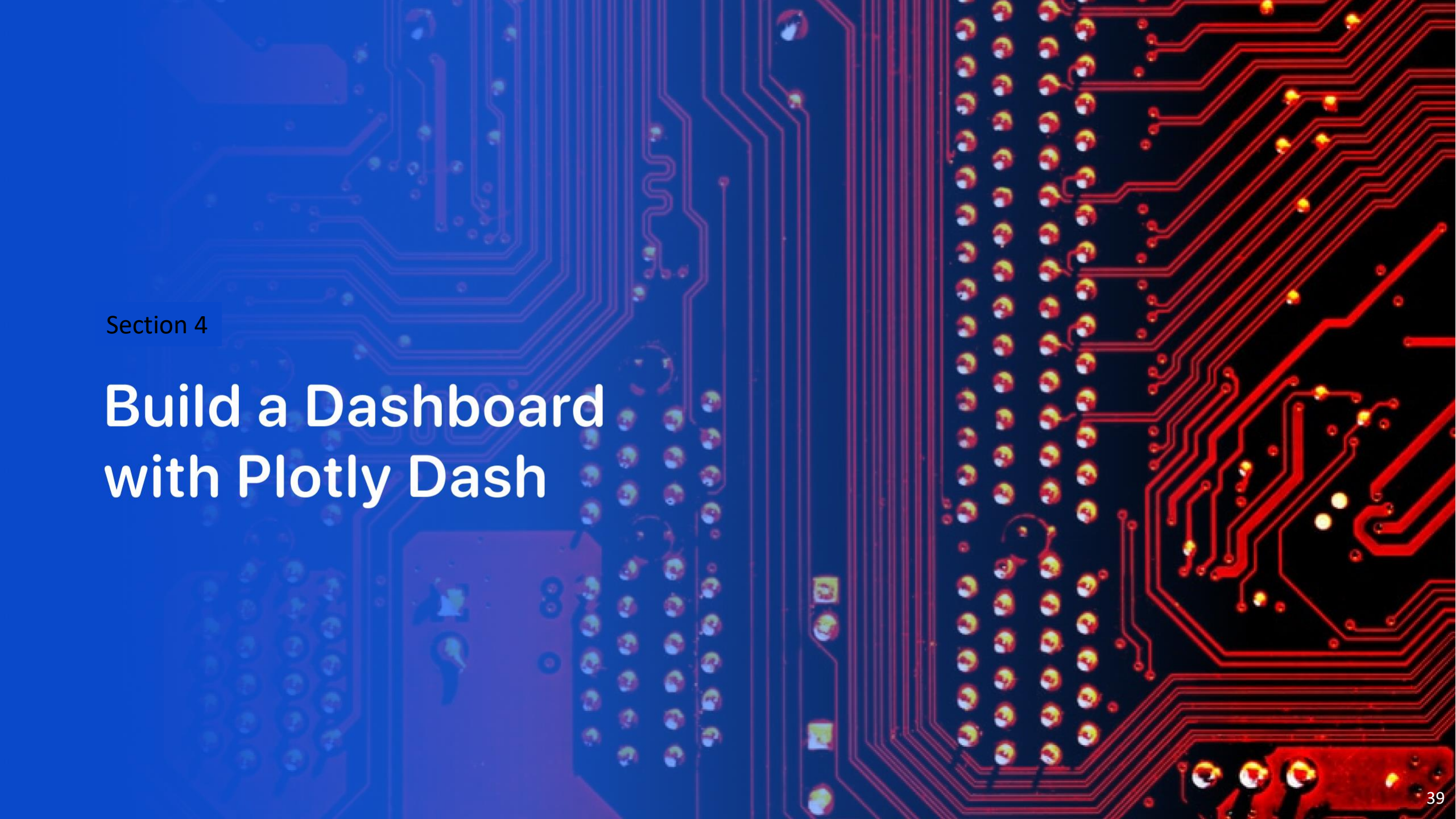
Closest City, Cape Canaveral (19.74 km):
While not immediately adjacent, the site's relative proximity to Cape Canaveral raises concerns about the potential impact on populated areas in the event of a launch failure or malfunction. Depending on the nature of the incident, there could be risks to public safety and property damage.



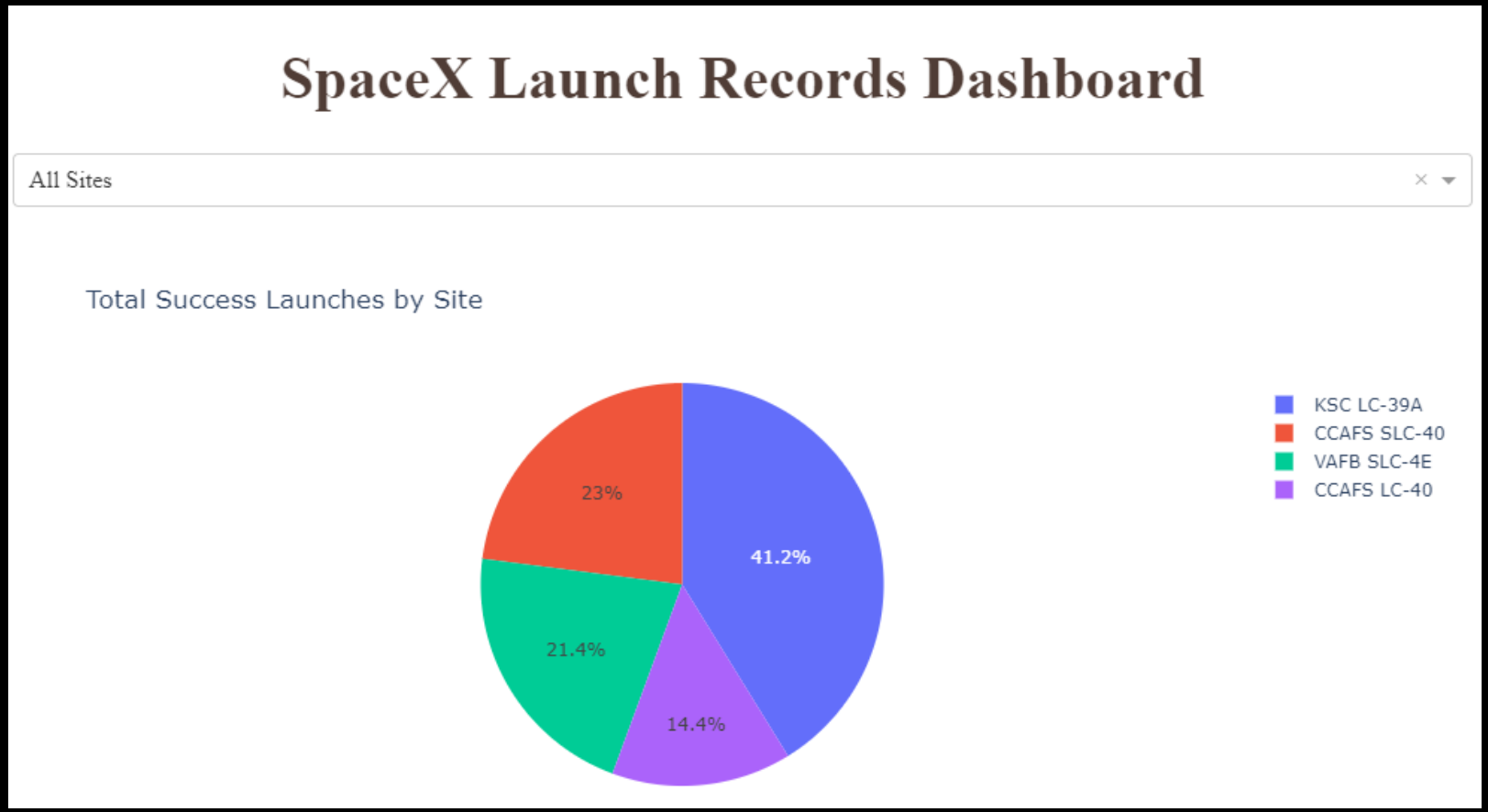
Overall, the analysis highlights the need for thorough risk assessment and mitigation strategies to address the potential dangers associated with the launch site's proximity to railways, highways, the coastline, and nearby populated areas. These measures are essential to ensure the safety of personnel, protect the environment, and minimize the impact of any unforeseen events on launch operations.

Section 4

Build a Dashboard with Plotly Dash

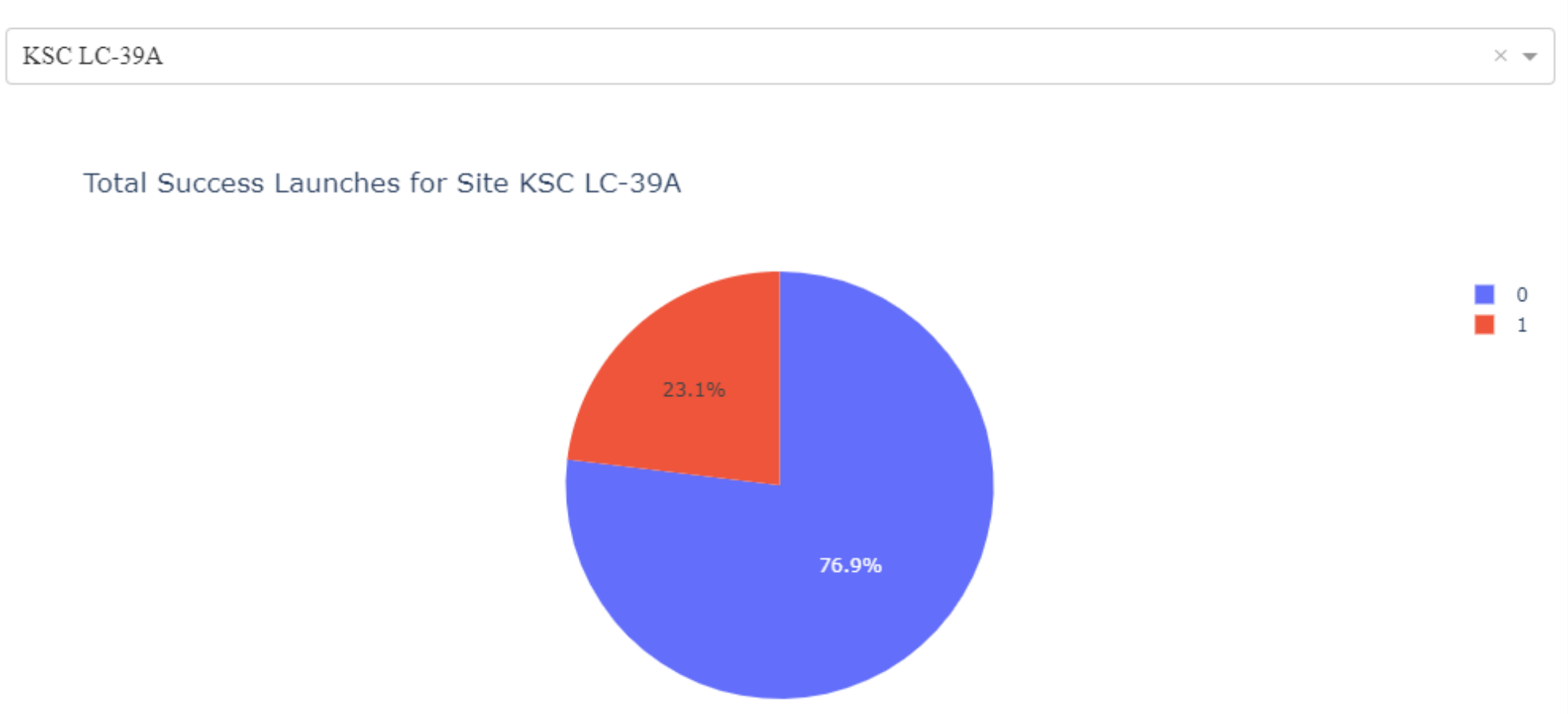


Launch success count for all sites



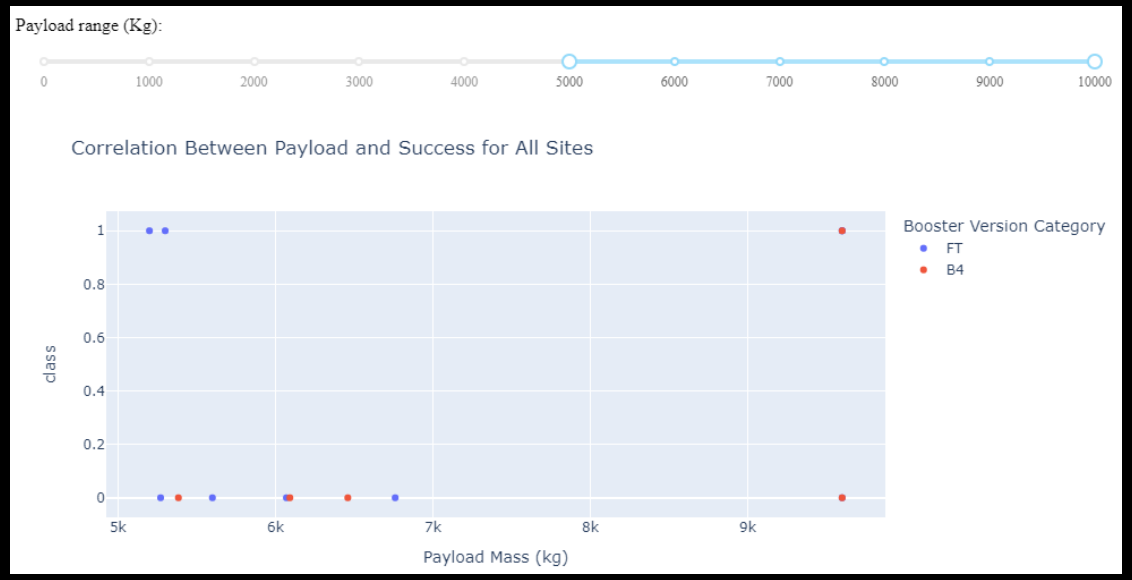
It's evident that KSC LC-39A exhibits the highest success rate for launches.

Total success launches for Site KSC LC 39A



- It's observed that KSC LC 39A has attained a success rate of 76.9%, accompanied by a failure rate of 23.1%.

Payload Mass vs. Launch Outcome Across All Sites



Payloads with lower weight demonstrate a higher success rate compared to heavier payloads.

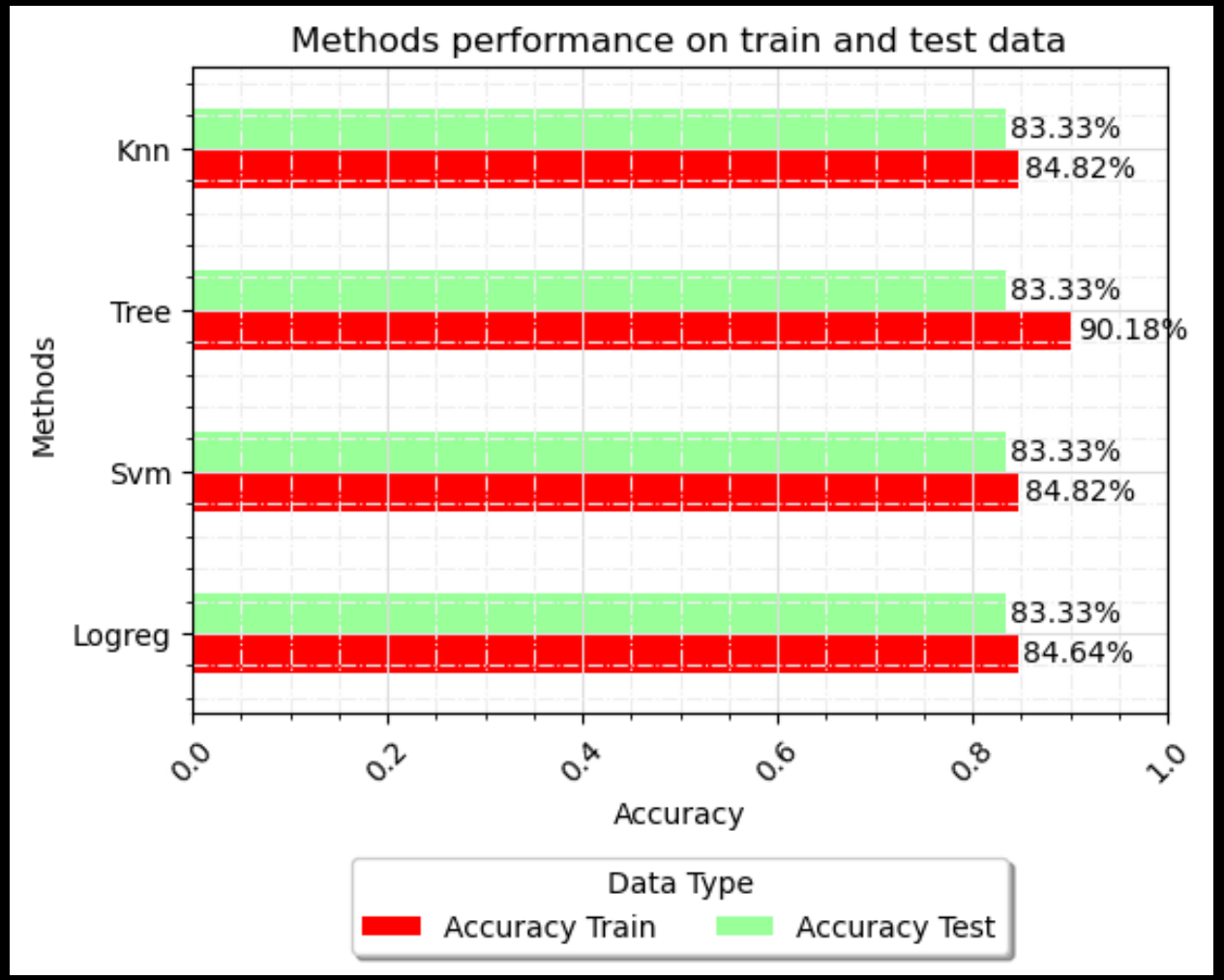
Section 5

Predictive Analysis (Classification)

Classification Accuracy

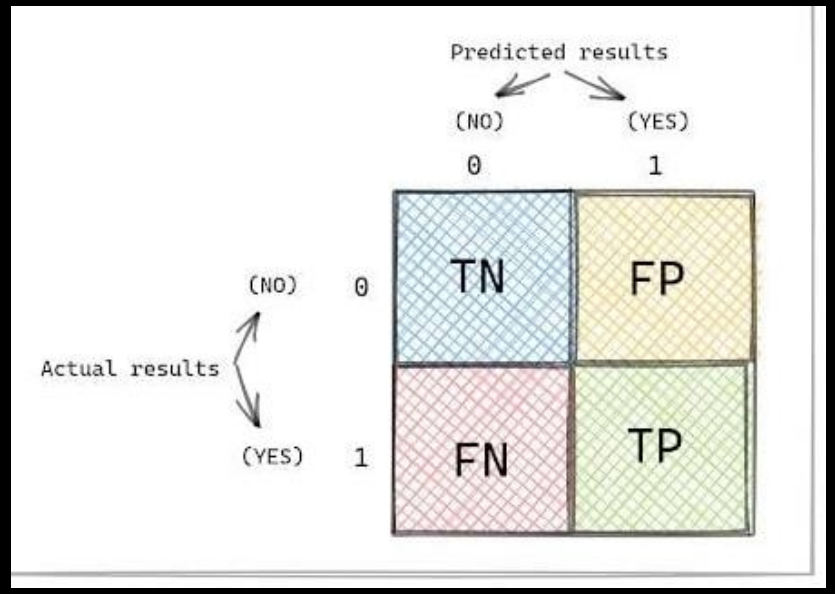
Model	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.901786	0.833333
Knn	0.848214	0.833333

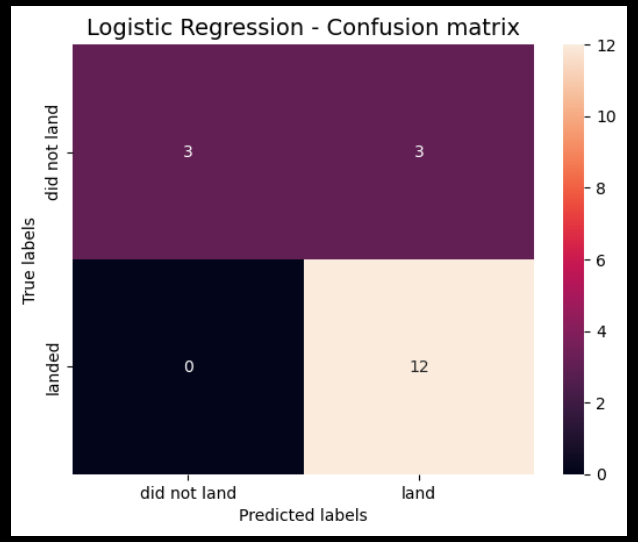
In the accuracy test, all methods exhibited comparable performance. To make a definitive choice between them, acquiring additional test data would be beneficial. However, if an immediate decision were necessary, we would opt for the decision tree method.

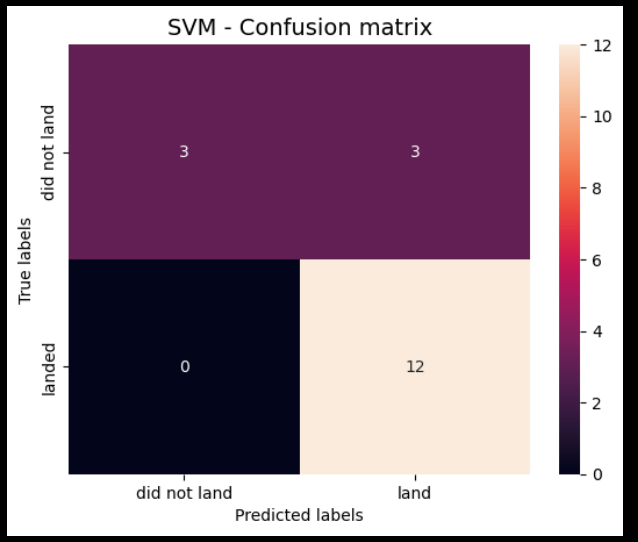


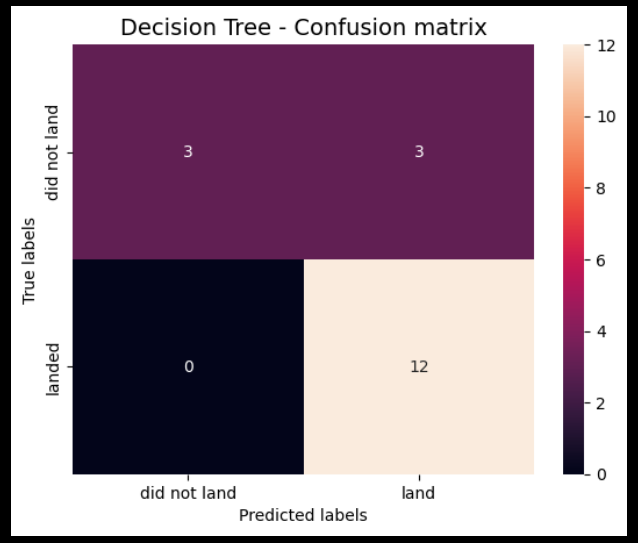
Confusion Matrix

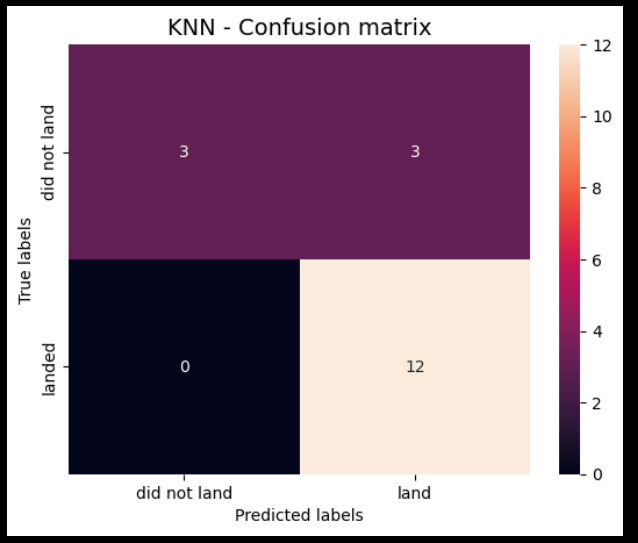
Since the test accuracies are identical, the confusion matrices also show identical results. The primary issue with these models lies in their tendency to produce false positives.











Conclusions

- Mission success hinges on various factors, including launch site, orbit, and accumulated experience from past launches.
- Orbits such as GEO, HEO, SSO, and ES L1 exhibit notably high success rates, highlighting their importance in mission planning.
- Payload mass is a critical consideration for mission success, with lighter payloads generally proving more effective across different orbits.
- While KSC LC 39A emerges as the top launch site in the dataset, the underlying reasons for its superiority require further investigation, possibly through the acquisition of atmospheric or other pertinent data.
- Despite comparable test accuracies, the Decision Tree Algorithm is chosen as the preferred model due to its superior train accuracy, underscoring its effectiveness in analyzing the dataset.

Thank you!