IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kamrul Hasan
25/06/24

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API and Web Scraping
  - Exploratory Data Analysis using SQL and Data Visualizations
  - Interactive Map using Folium
  - Interactive Dashboard using Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Map
  - Interactive Dashboard
  - Predictive Results

Introduction

- Project background and context
  - The goal of this project is to predict if the Falcon 9 first stage will land successfully. 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.
- Problems you want to find answers
  - What are the main characteristics of a failed or successful launch?
  - How does each of a rocket’s variable effect whether a launch is successful or a failure?
  - What are the best conditions for SpaceX to achieve the best landing success?

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

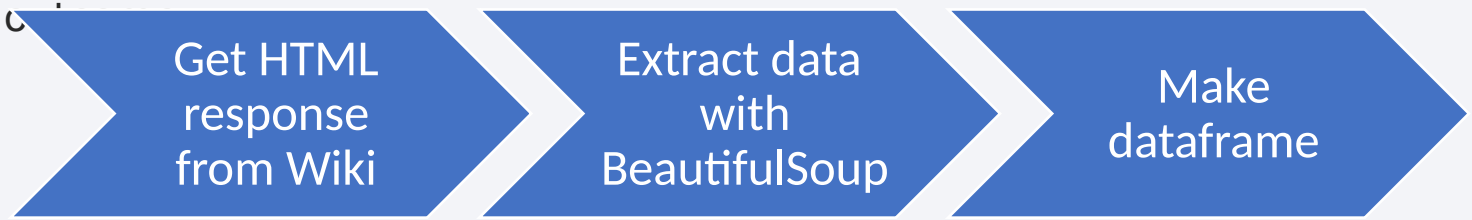
Data Collection

One way datasets are collected is from SpaceX API

The information obtained by the SpaceX API includes information on the rocket, the payload, and the outcome.

Another way datasets are collected is from Web Scrapping from Wikipedia

The information obtained by Wikipedia includes information on the launches, the payload, and the



Data Collection – SpaceX API

1. Getting Response from API

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

2. Get Json File from response

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

3. Transform Data

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

4. Create Dictionary

```
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 and Filter Dataframe

```
df = pd.DataFrame(launch_dict)
data_falcon9 = df[df['BoosterVersion'] == 'Falcon 9']
```

6. Export Data

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

Data Collection - Scraping

1. Get response from Wiki

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

2. Obtain all tables

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

3. Get column names

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

4. Create dictionary

```
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']=[]
```

5. Add data (See notebook for rest of code)

```
for table_number,table in
enumerate(soup.find_all('table',"wikitable
plainrowheaders collapsible")):
  # get table row
   for rows in table.find_all("tr"):
      if rows.th:
         if rows.th.string:
            flight_number=rows.th.string.strip()
            flag=flight_number.isdigit()
```

6. Create and export dataframe

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

- Needed to change categorical data of outcome to a binary variable: 1 means a success and 0 means a failure

1. Obtain counts and all the unique labels of outcome column

```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

2. Define all failed outcomes

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
```

3. Create Class column which contains the binary variable

```
landing_class = []
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
landing_class
```

EDA with Data Visualization

- **Scatter Graphs: Shows relationship between variables**
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit
  - Payload Mass vs Orbit
- **Bar Graphs: Shows relationship between categorical and numerical variables**
  - Success Rate vs Orbit
- **Line Graphs: Shows variable's trends**
  - Success Rate vs Year

EDA with SQL

- Display names of the unique launch sites
- Display the first five records where launch site begins with ‘CCA’
- Display total payload mass carried by those launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List date of first successful landing outcome in ground pad was achieved
- List names of boosters which have success in drone ship and have payload mass greater than 4000 and less than 6000
- List total number of successful and failure mission outcomes
- List names of booster versions which have carried the maximum payload mass
- List records which display month names, failure landing outcomes in drone ships, booster versions, and launch site for those in the year 2015
- Rank count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order

Build an Interactive Map with Folium

Folium Map object is centered on NASA Johnson Space Center

Red circle at NASA Johnson Space Center’s coordinates with label

Red circles at each launch site coordinates with labels

Clustering points to display diverse and multiple pieces of information for the same coordinates.

Markers to show successful and unsuccessful launches

Green = Successful and Red = Unsuccessful

Markers and lines to show distance between launch sites and key locations

This map is designed to enhance understanding of the problem and the data. It displays all launch sites, their surroundings, and the number of successful and unsuccessful launches.

Build a Dashboard with Plotly Dash

The dashboard contains a dropdown menu, a pie chart, a scatter plot, and a range slider

- Dropdown allows users to change launch site
- Pie chart shows total success and failures of the specified launch site
- Range slider allows for the user to change payload max in a fixed range
- Scatter plot shows the relationship between the success and payload mass

Predictive Analysis (Classification)

1. Data Preparation

Load and Normalize data. Split data into training and testing sets.

2. Model Training

- Selection of Machine Learning Algorithm
- Set parameters for GridSearchCV
- Train GridSearchCV model with training sets

3. Model Evaluation

- Obtain best hyperparameters
- Calculate accuracy of model with testing sets
- Plot Confusion Matrix

4. Model Comparison

Comparison models based on accuracy and choose the model with the best accuracy

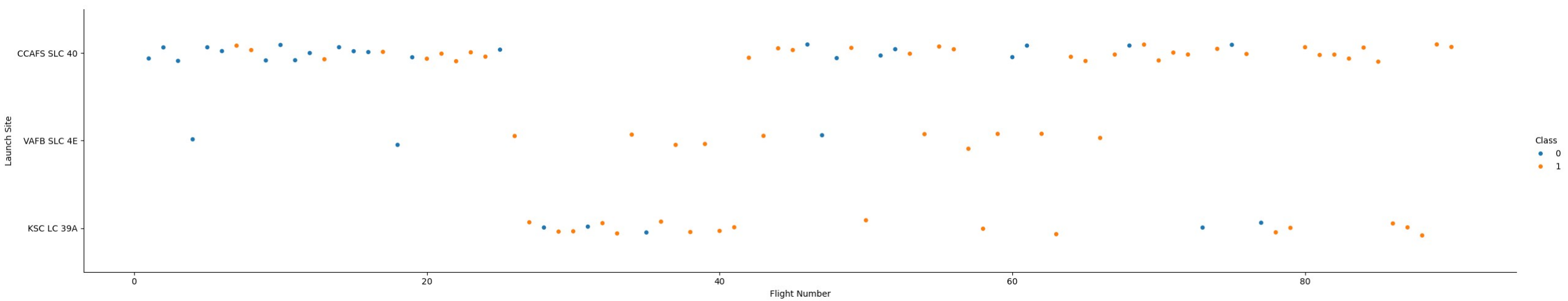
Results

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

Section 2

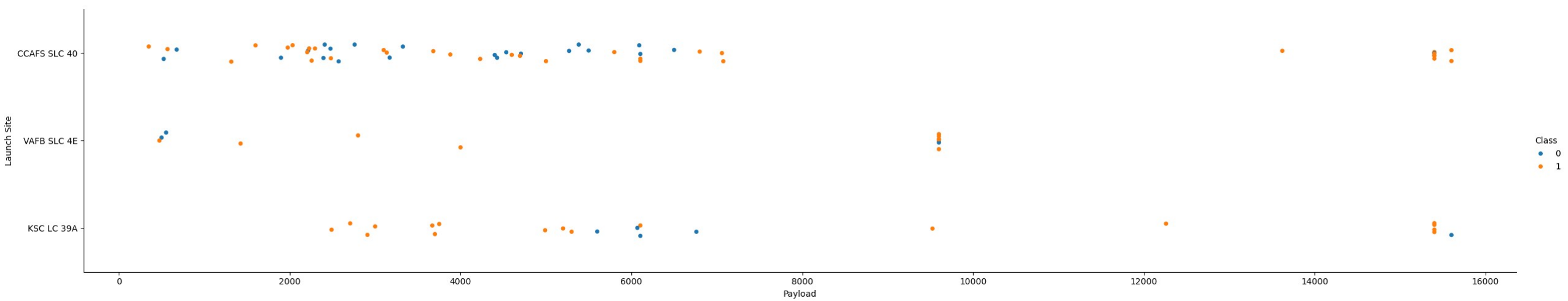
Insights drawn from EDA

Flight Number vs. Launch Site



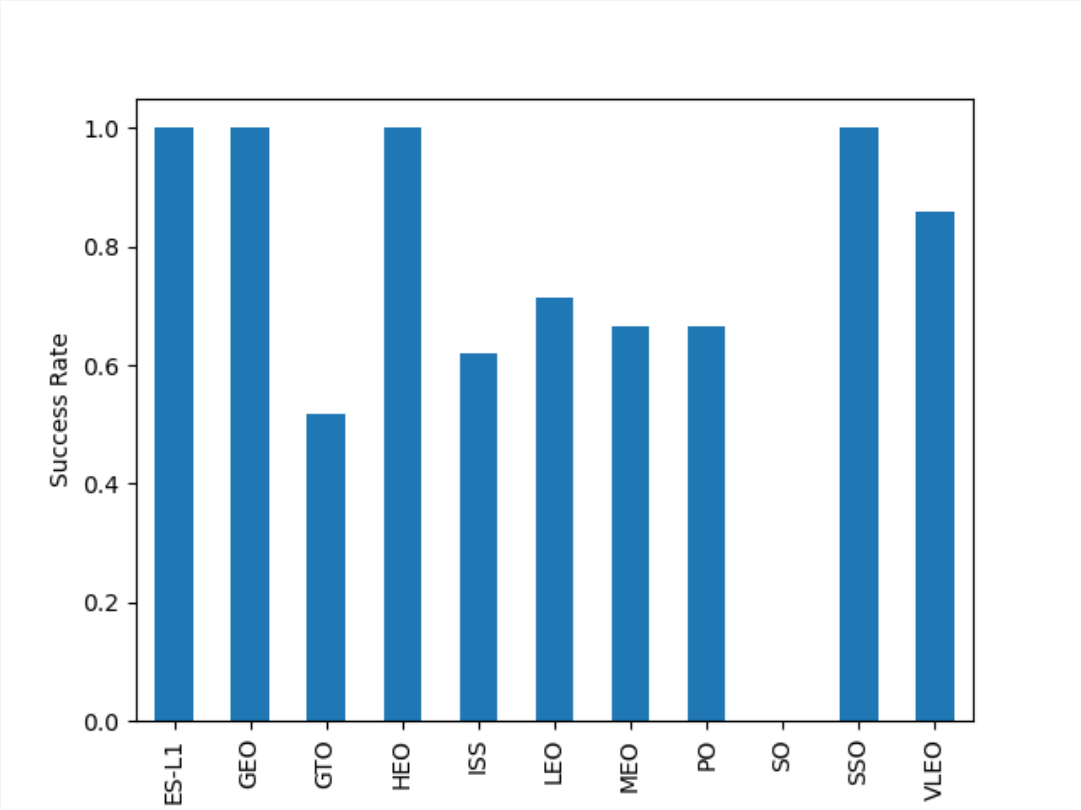
It is observed that KSC and VAFB launch sites have similar number of successes.

Payload vs. Launch Site



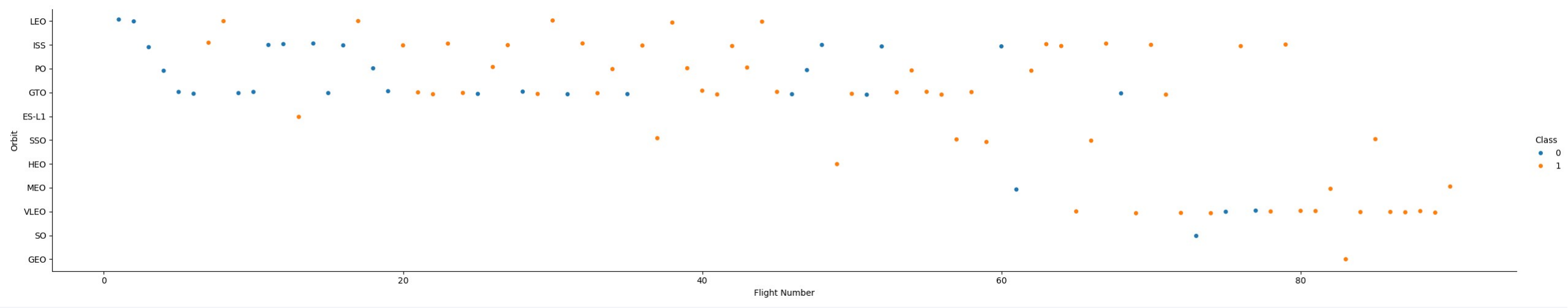
For all launch sites, it is observed that the number of successful launches increases with heavier payload masses. At the KSC launch site, there are also quite a few successes within the 2000-4000 kg payload range

Success Rate vs. Orbit Type



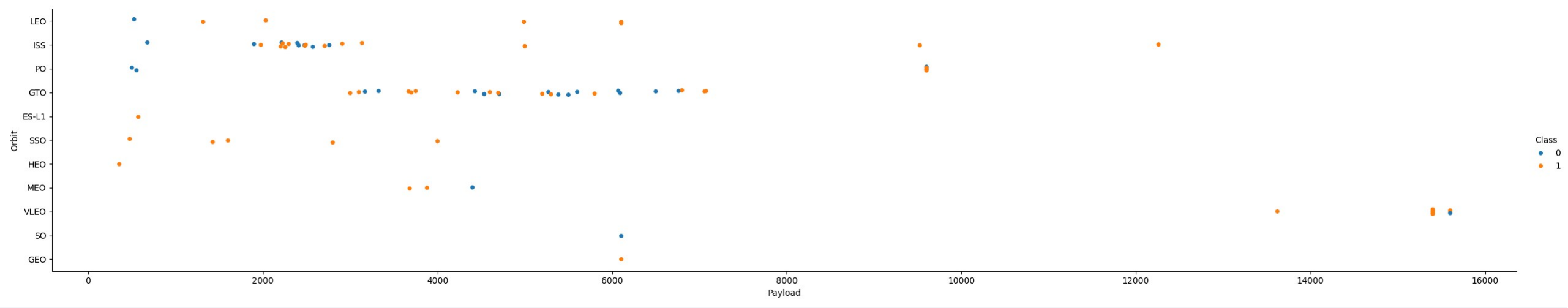
The ESL, GEO, HEO, SSO orbit types have 100 percent success rate, while the orbit type of SO as zero percent success rate

Flight Number vs. Orbit Type



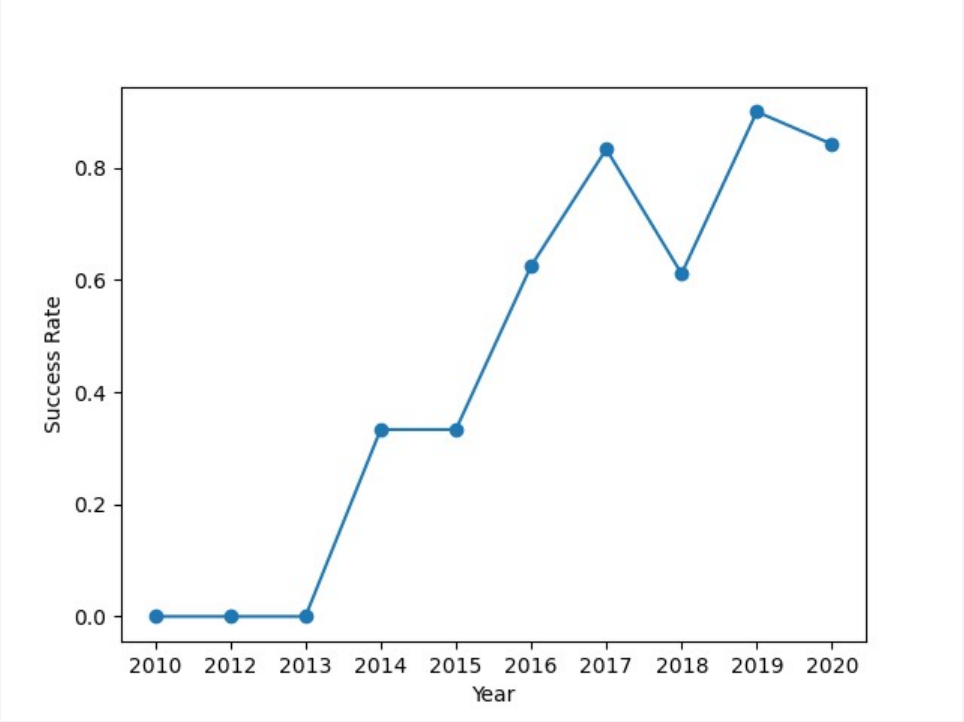
It is observed that those orbit types with 100 percent and zero percent success rates have small sample sizes.

Payload vs. Orbit Type



Payload mass significantly impacts certain orbit types. Specifically, ES-L1, SSO, and HEO orbits show higher success rates with smaller payload masses, while PO, LEO, and VLEO orbits tend to be more successful with larger payload masses.

Launch Success Yearly Trend



It can be observed that as the years have gone on the success rate has increased; however, it is also observed that there was a slight decrease success rate in 2018.

All Launch Site Names

%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL

Result: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40

Using the keyword distinct removes all duplicate values for Launch_Site

Launch Site Names Begin with 'CCA'

%sql SELECT * FROM SPACEXTBL WHERE “LAUNCH_SITE” LIKE ‘CCA%’

Date	Time (UTC)	Booster_ Version	Launch_Site	Payload	PAYLOAD _MASS__ KG_	Orbit	Customer	Mission_ Outcome	Landing_ Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	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

Using the keyword and like allows us to find only the rows where launch_site starts with CCA

Total Payload Mass

%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE “CUSTOMER” LIKE ‘NASA (CRS)’;

Result: 45,596 kg

Using the SUM keyword allows us to add up the value of PAYLOAD_MASS__KG_ where customer is NASA (CRS)

Average Payload Mass by F9 v1.1

%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE “BOOSTER_VERSION” LIKE ‘%F9 v1.1%’;

Result: 2534.67 kg

By using the keyword AVG allows us to find the mean value for PAYLOAD_MASS__KG_ for all the rows where the Booster Version is F9 v1.1

First Successful Ground Landing Date

%sql SELECT MIN(DATE), Landing_Outcome as Outcome, FROM SPACEXTBL WHERE Outcome like ‘%Success (ground pad)%’;

Result: 2015-12-22

By using the keyword MIN could find the first successful landing where the outcome is Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

%sql SELECT Booster_Version, Landing_Outcome as Outcome, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE Outcome like '%Success (drone ship) %' AND CAST(PAYLOAD_MASS__KG_ AS FLOAT) > 4000 AND CAST(PAYLOAD_MASS__KG_ AS FLOAT) < 6000;

Result: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

By using the keyword AND are able to query multiple things into one query

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;

Result:

- Success: 100
- Failure: 1

Retrieves the counts of mission outcomes labeled as 'Success' and 'Failure' from the table SPACEXTBL and presents them as columns named SUCCESS and FAILURE, respectively.

Boosters Carried Maximum Payload

%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE “PAYLOAD_MASS__KG_” = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

Result:

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

Retrieves unique booster versions from the table SPACEXTBL where the payload mass is equal to the maximum payload mass using a subquery.

2015 Launch Records

%sql SELECT substr(DATE, 6, 2) AS MONTH, Booster_Version, Launch_Site, Landing_Outcome as Outcome, DATE AS dt FROM SPACEXTBL WHERE Outcome like ‘Failure (drone ship)’ AND substr(dt, 0, 5) == ‘2015’;

Result:

MONTH	Booster_Version	Launch_Site	Outcome	dt
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

By using the keyword substr can find the month and year of a date to find the month and booster version of a launch from the year 2015 and resulted in a drone ship failure.

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

```
%sql SELECT COUNT(Landing_Outcome), Landing_Outcome as Outcome FROM SPACEXTBL
WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20' AND Outcome LIKE '%Success%'
GROUP BY Outcome ORDER BY COUNT(Outcome) DESC;
```

Result:

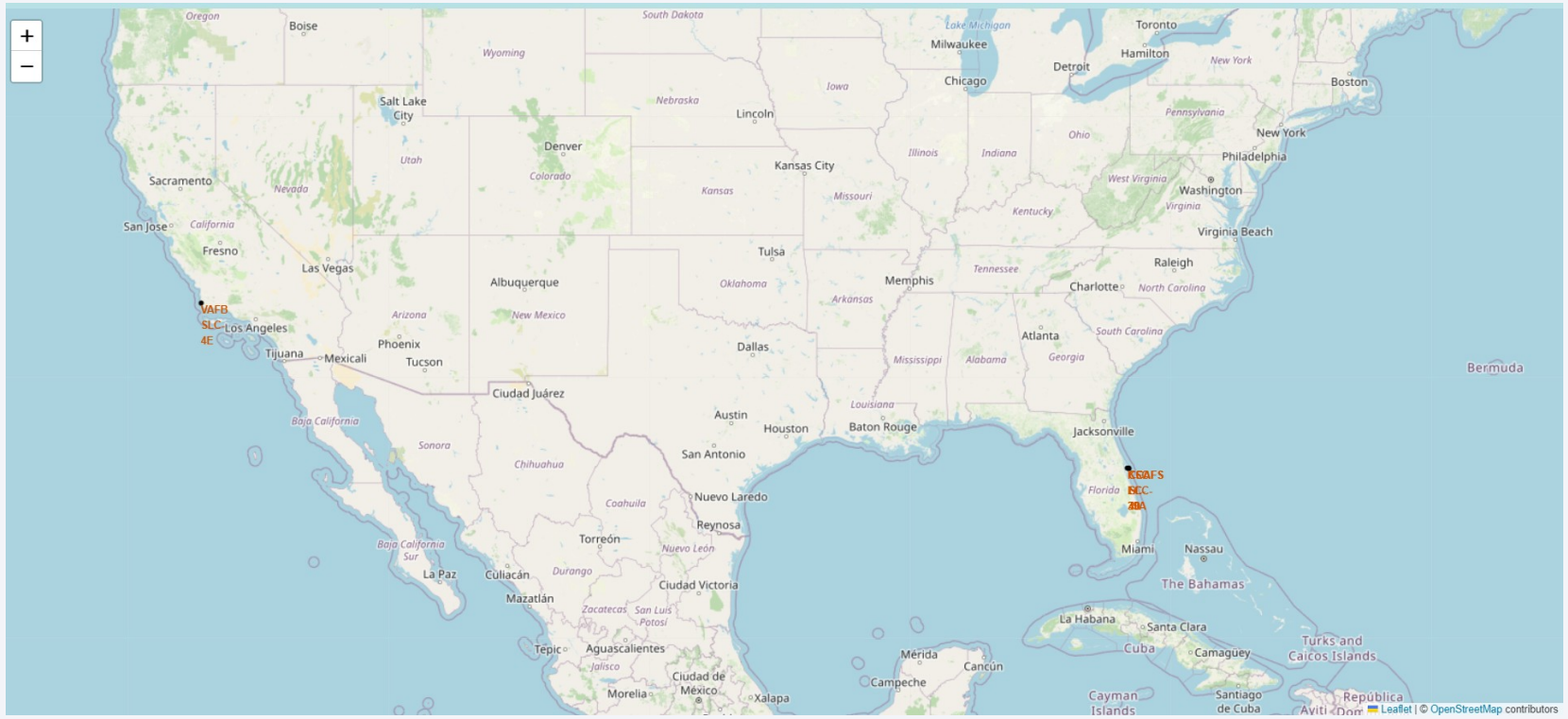
COUNT(Landing_Outcome)	Outcome
5	Success (drone ship)
3	Success (ground pad)

Use the keywords group by and order by to find the ranking of the counts of successful launches between the dates 2010-06-04 and 2017-03-20.

Section 3

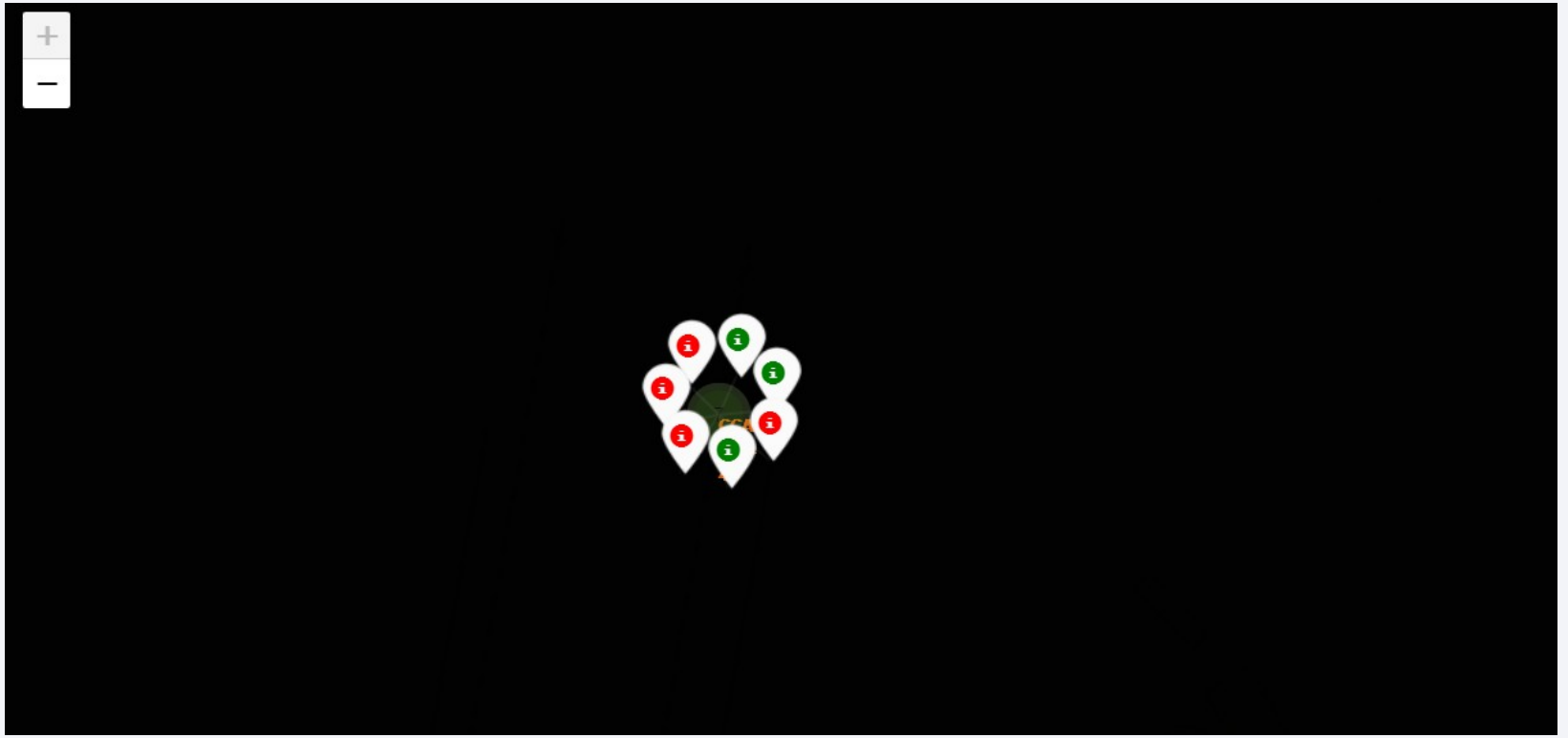
Launch Sites Proximities Analysis

Folium Map- Base Launch Sites



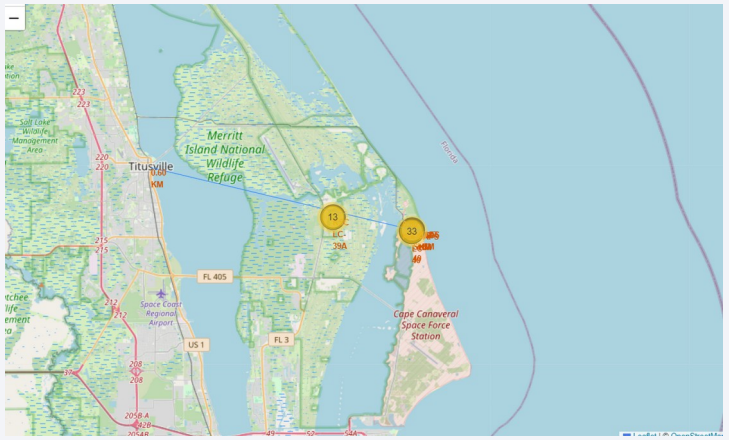
One can observe that the SpaceX launch sites are either in California or Florida

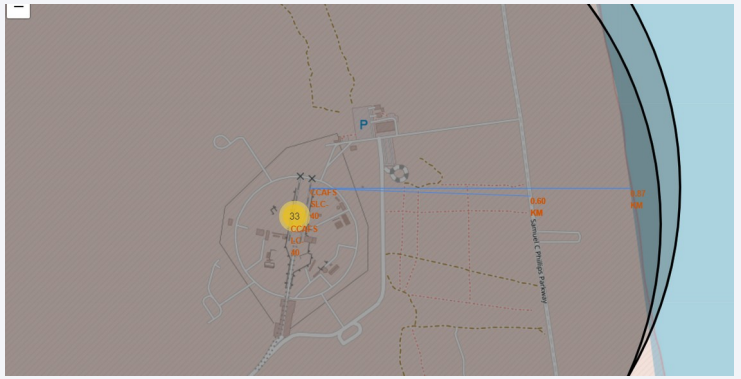
Folium Map- Colored Labeled Markers



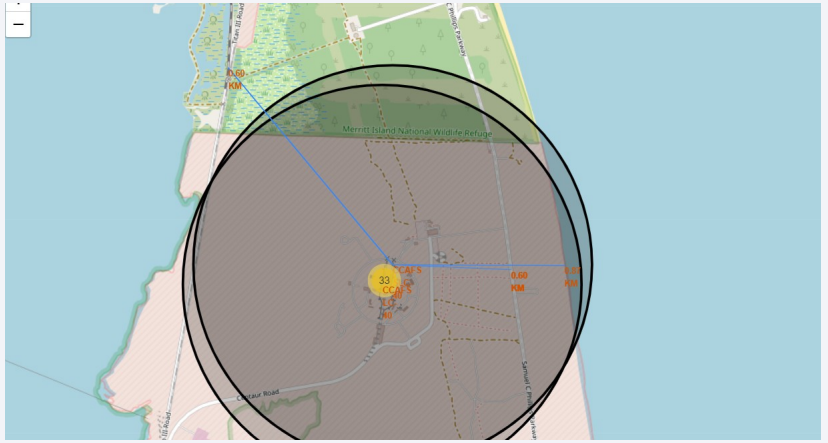
Green Marker represents a successful launch. Red Marker represents a failed launch

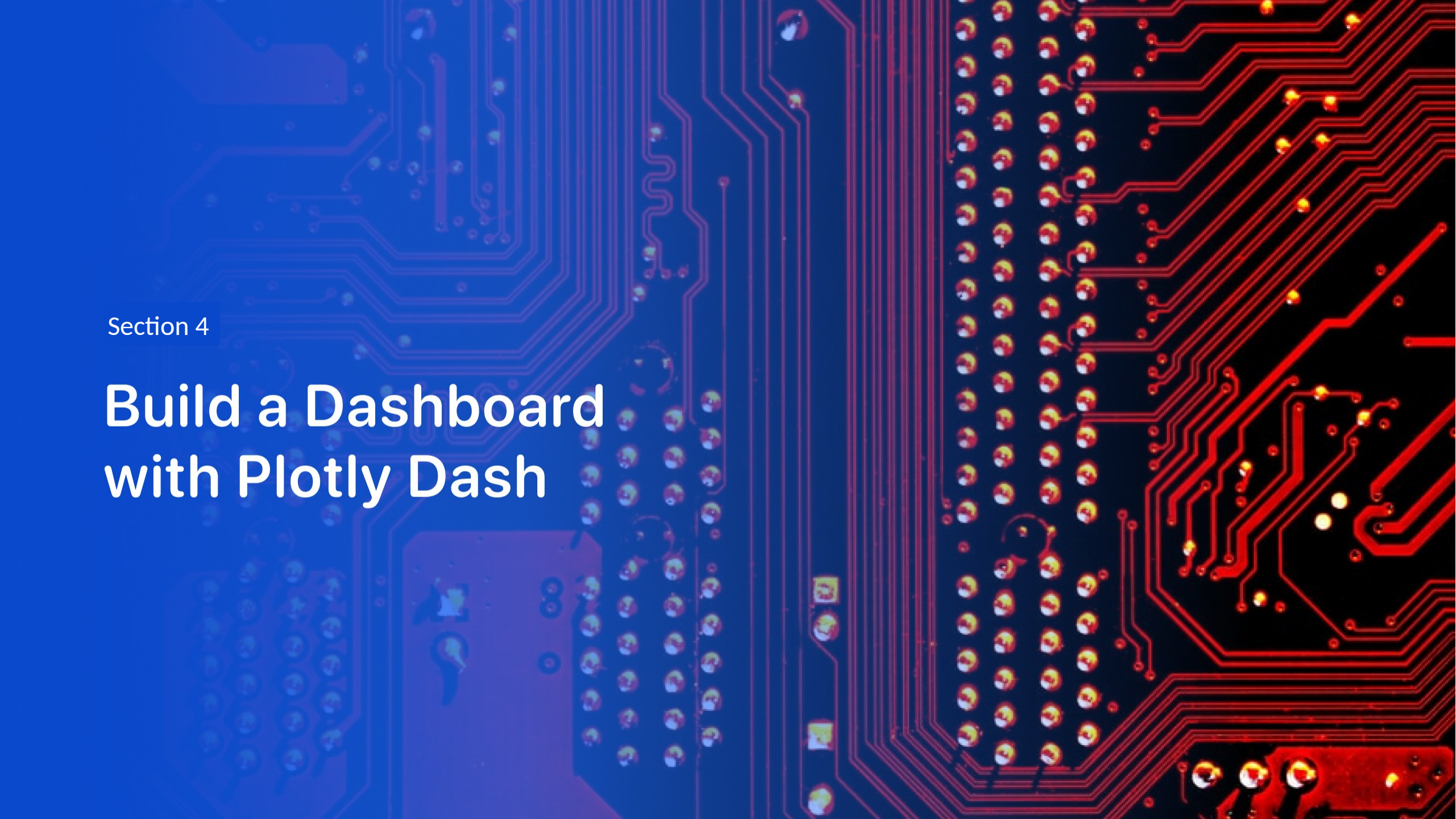
Folium Map- CCAFS SLC-40 and its Proximities



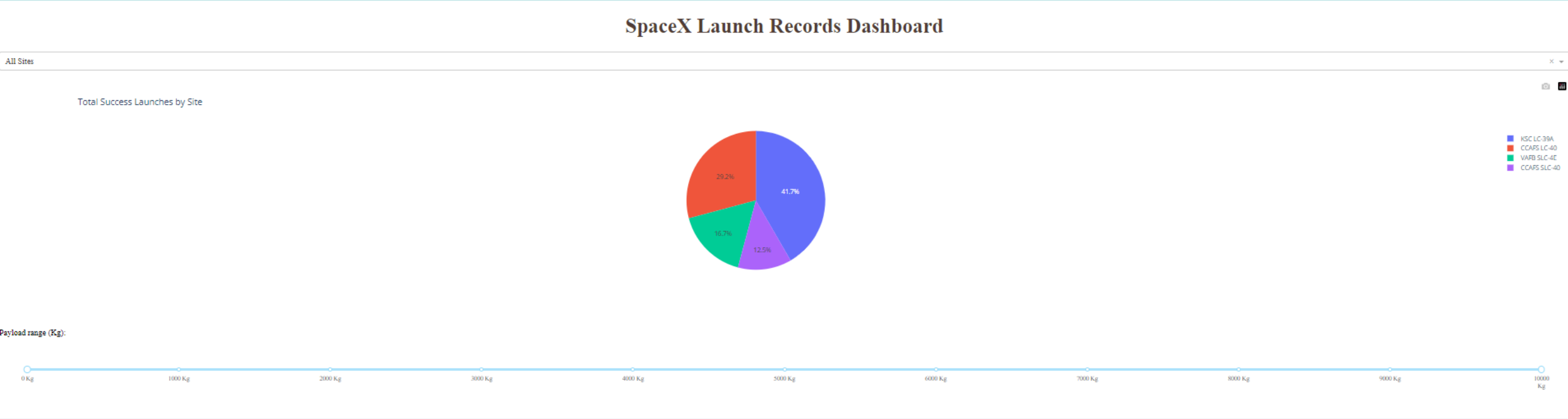


- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance from cities? No



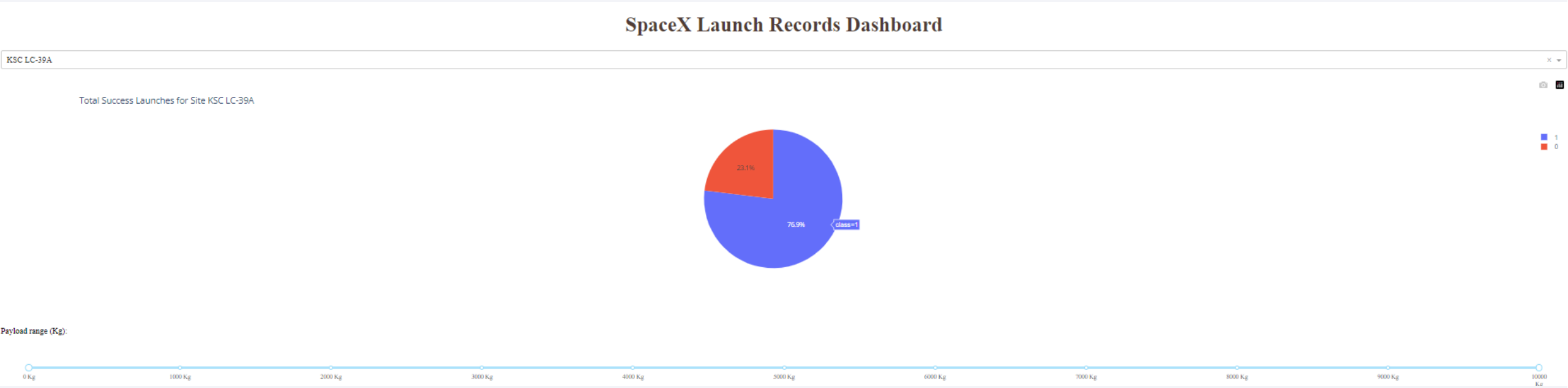

Section 4
Build a Dashboard with Plotly Dash

Dashboard- Total Successes By Site



It is observed that KSC has the most successful launches.

Dashboard- Success Percentage of KSC



It is observed that KSC has a success rate of 76.9%.

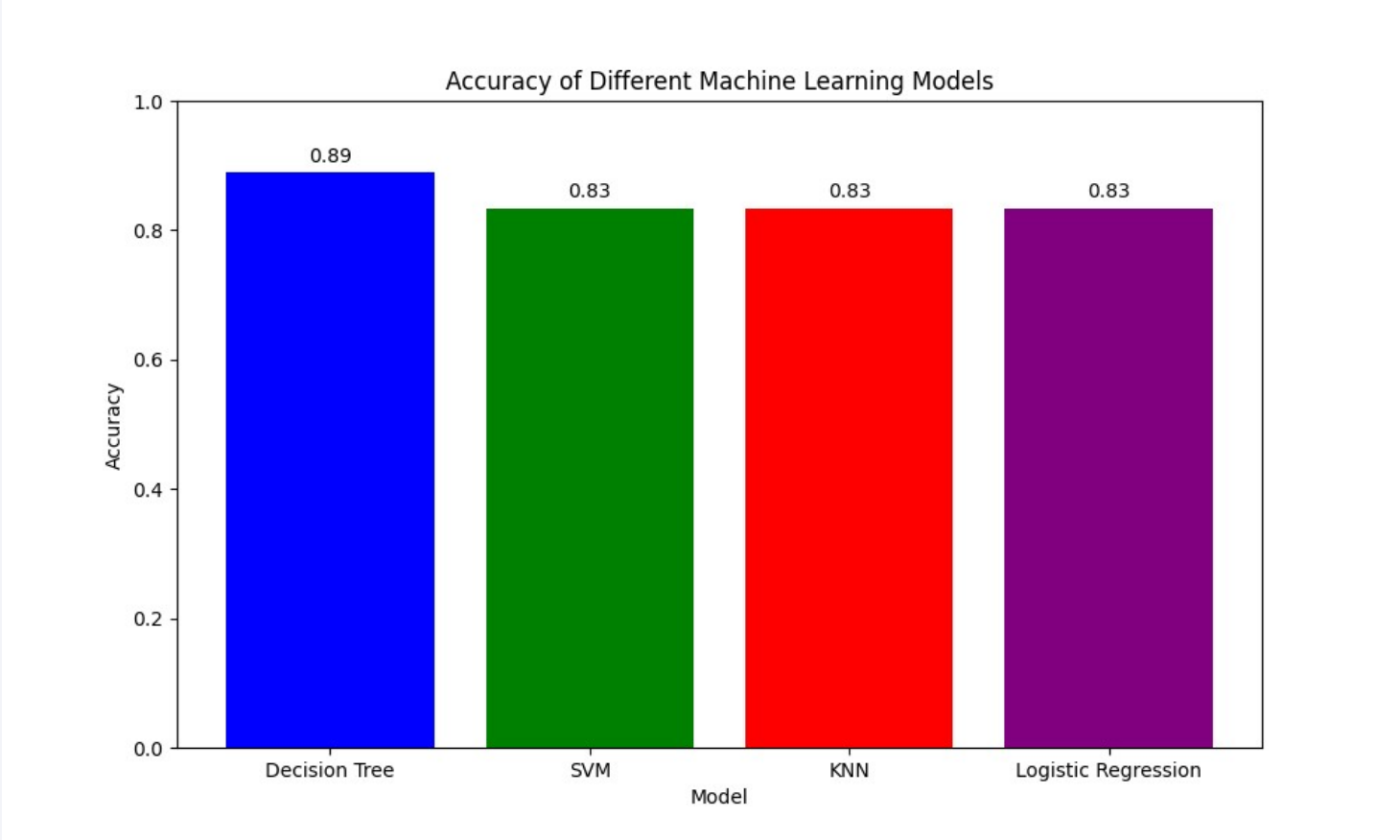
Dashboard- Success Rate of Payload Mass for All

It is observed that the highest success rates where mainly those in the 2000-4000 kg range.

Section 5

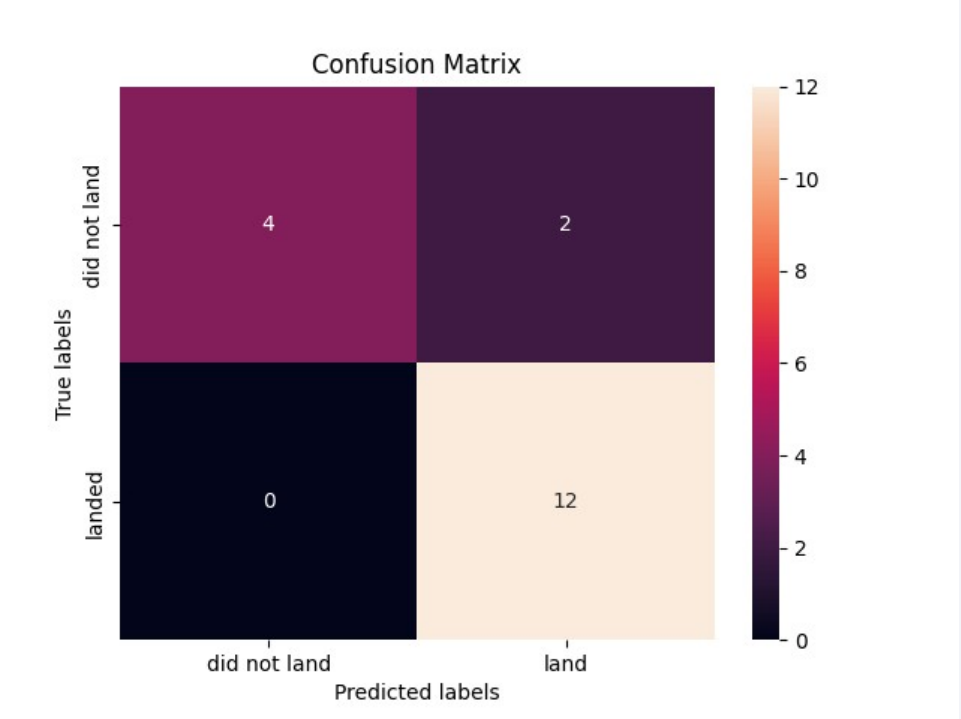
Predictive Analysis (Classification)

Classification Accuracy



The models are pretty much the same accuracy; however, the accuracy of the decision tree model is the best

Confusion Matrix



The model has a good accuracy; however, there are two false positive results.

Conclusions

- Landing outcome can be influenced by multiple factors such as launch site, orbit, and payload mass.
- Orbits that have the most successful landings are: GEO, HEO, SSO, ES-L1
- The payload mass significantly influences the outcome of a landing depending on the orbit, with some orbits requiring higher or lower payload mass for successful landings.
- The Decision Tree Model is the best at predicting the correct landing outcome.

Thank you!