# Winning Space Race with Data Science

Karlo Gabbriel Batto

June 15, 2023

IBM Developer SKILLS NETWORK

# Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- INSIGHTS AND CONCLUSIONS
- RECOMMENDATION
- APPENDIX

# Executive Summary

# Summary of Methodologies

- Data Collection API
- Data Collection API with Web Scraping
- Data Wrangling
- EDA with SQL
- EDA with Visualization
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis

# Summary of all Results

- Exploratory Data Analysis
- Interactive Analytics
- Predictive Analysis

# Introduction

# Project Background and Context

- Commercial Space age is now here
- One of the most successful companies that provide commercial space flight is Elon Musk's SpaceX.
- SpaceX's success comes from the affordability of its launches by reusing the first stage for future launches.
- How do we take advantage of the public information of these launches?
- The goal is to create a machine learning model that predicts a successful flight.

# Problems you want to find answers

- The problem is that we do not know the probability of a successful flight due to the number of variables that affect the flight of the first stage,

  1. What are the factors behind the success and failure of a flight?
  2. What is the probability of a successful landing based on the data gathered?

- Data must be collected from SpaceX's launches to determine the probability of landing.

# Methodology

# Methodology

- Data Collection Methodology
  - Using SpaceX API and Web Scraping
- Perform Data Wrangling
  - Determine launch outcomes
- Perform EDA w/ Visualization and SQL
- Perform Interactive Visual Analytics using Folium and Plotly Dash
- Perform Predictive Analysis using Classification Models
  - Build, tune, and evaluate classification models by scoring and using a confusion matrix

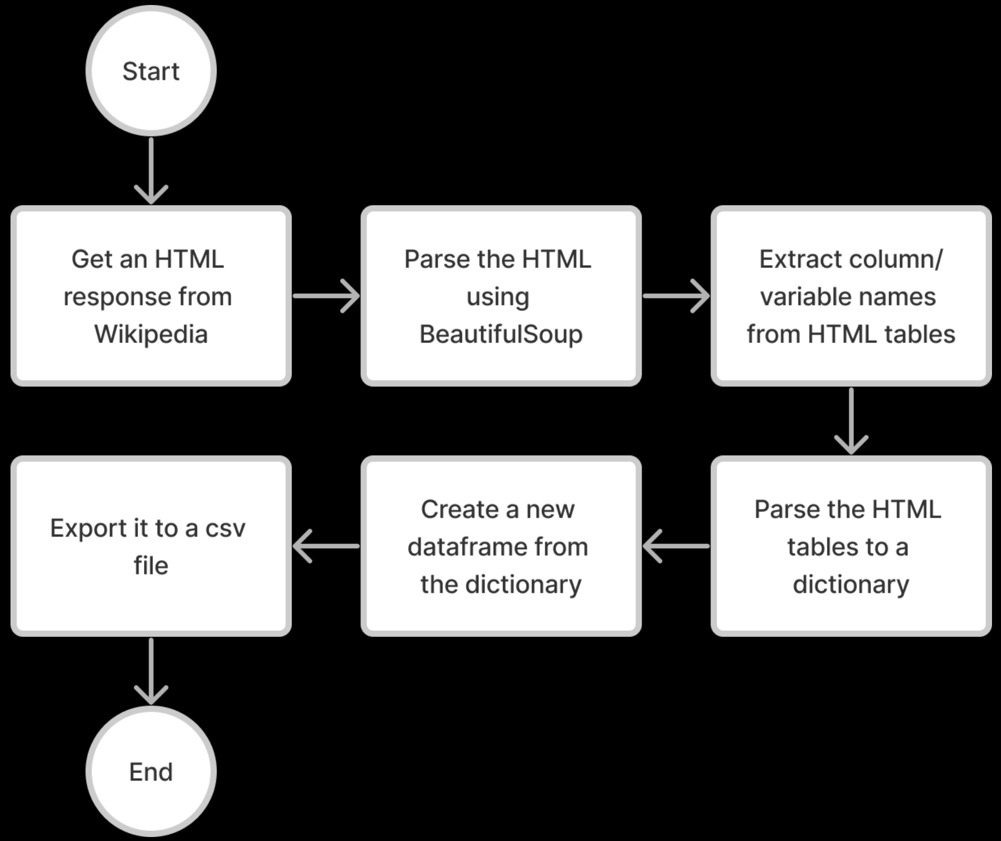
# Data Collection Methodology

There are two ways to collect SpaceX launch data.

- Accessing the SpaceX API with the REST API
  - https://api.spacexdata.com/v4
- Webscraping using Wikipedia as a source material using BeautifulSoup
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

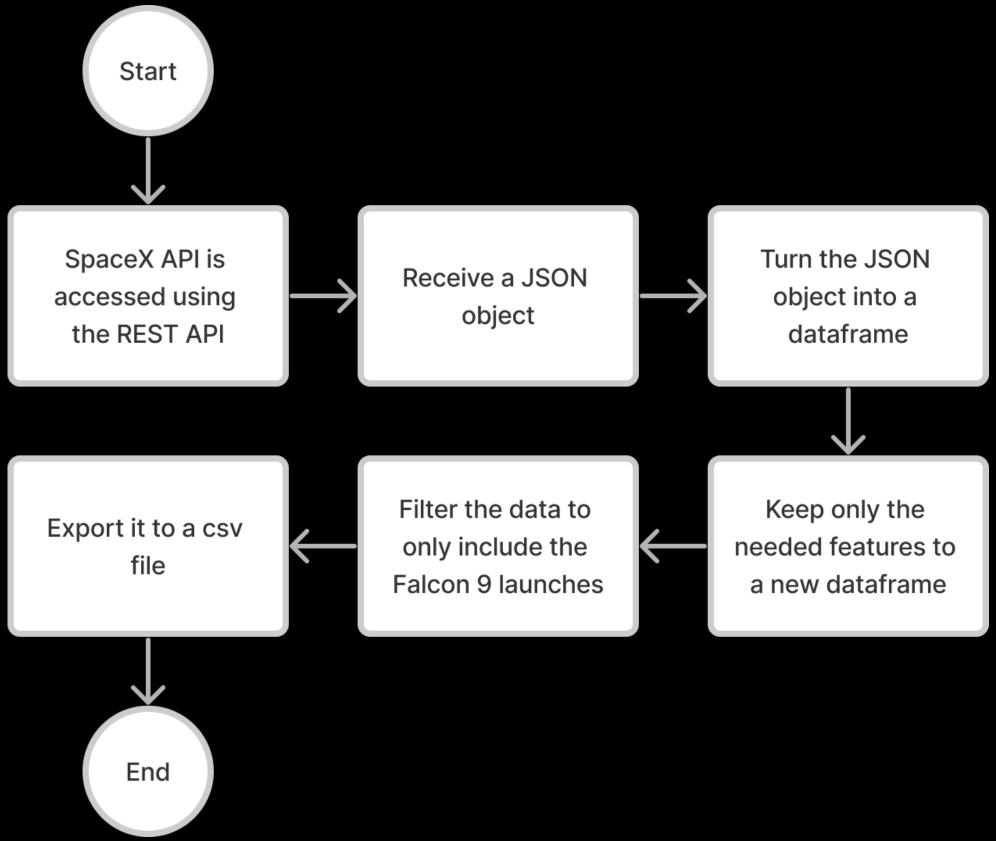
# Data Collection w/ SpaceX API

- SpaceX API is accessed
- Rocket, Launchpad, Payload, and Outcome data are used to create the dataset.
- Stored in a dataset for processing
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/1%20-%20Data%20Collection%20using%20an%20API.ipynb



# Data Collection w/ Web Scraping

- Get an HTML response from Wikipedia.
- BeautifulSoup is used
- Create a dataframe from the HTML tables.
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/2%20-%20Data%20Collection%20using%20Webscraping.ipynb



# Dealing with Missing Values

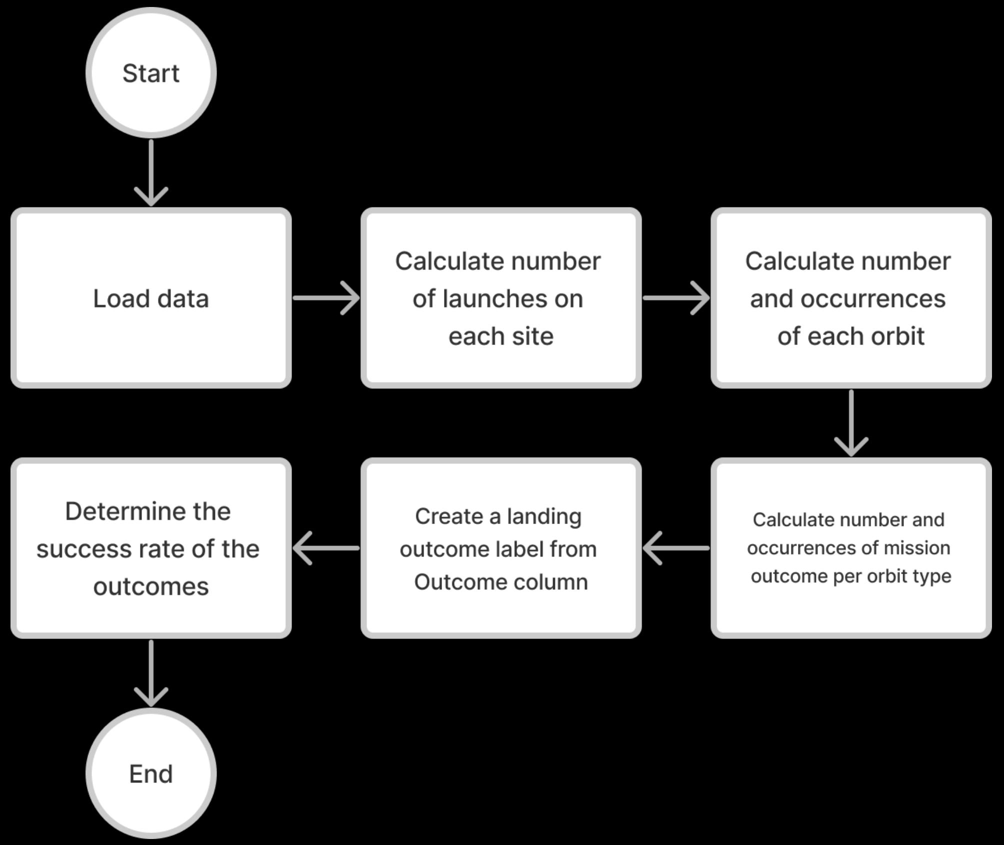
The dataset must be fixed first to gather meaningful insights

- In the LandingPad column of the dataset, some rows have NULL values. These NULL values will be represented as unused landing pads.
- Missing values in the PayLoadMass column is fixed by getting the mean of the whole column and replaced by the mean.

| | |
|---|---|
| Calculate the mean of the PayLoadMass column | `meanpayloadmass = data_falcon9['PayloadMass'].mean()` |
| Replace the nan values with the mean | `data_falcon9['PayloadMass'].replace(np.nan, meanpayloadmass, inplace = True)` |

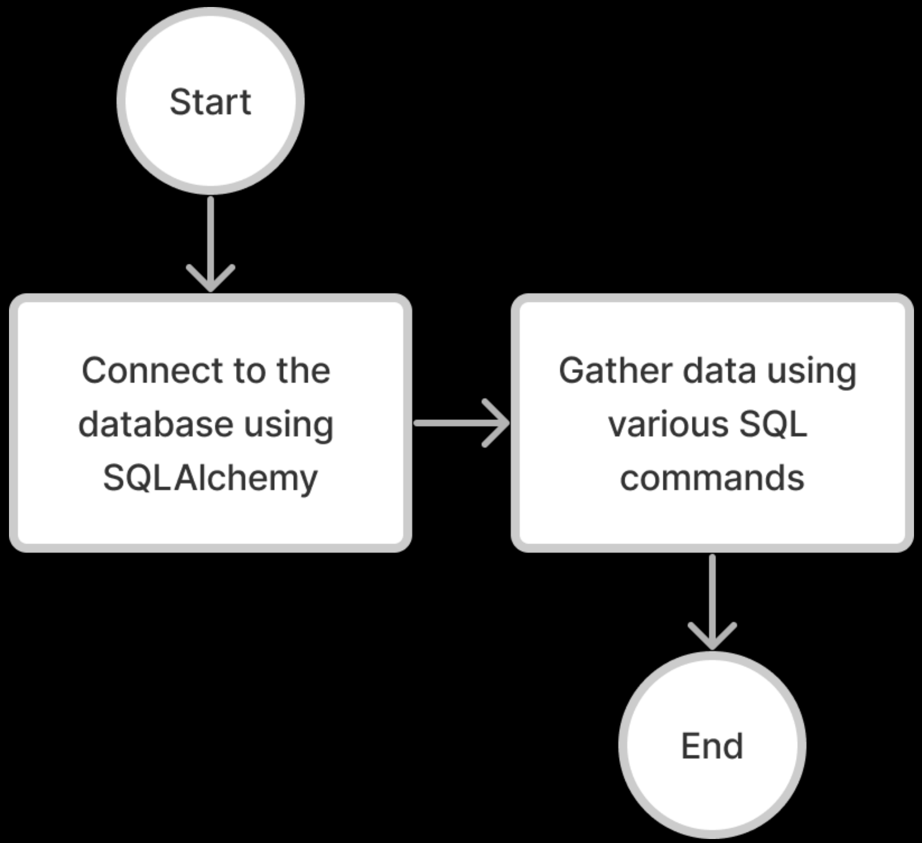
# Data Wrangling

- Calculate number of launches
- Determine the outcome of rocket launches
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/3%20-%20Data%20Wrangling.ipynb



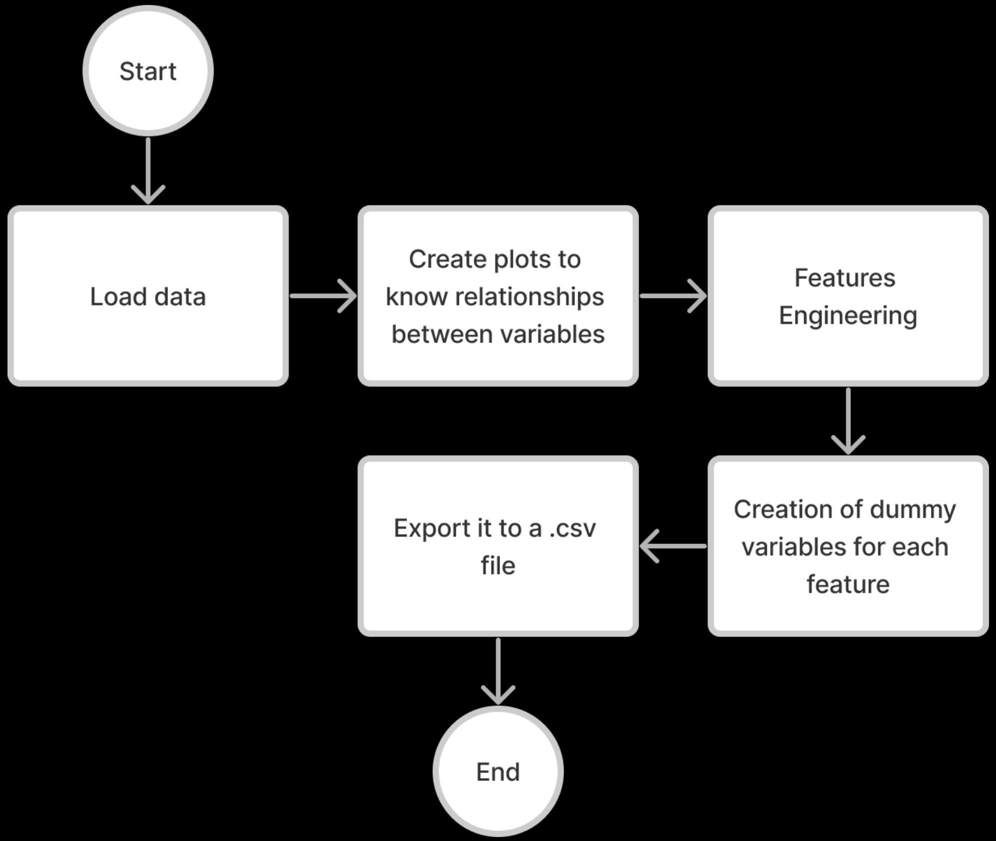
# EDA with SQL

- The dataset is accessed using SQLAlchemy
- Various SQL commands are used to gather data
  - Collect the names of launch sites
  - Payload mass by boosters launched
  - Successful and Failed launch outcomes
  - Etc.
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/4%20-%20EDA%20with%20SQL.ipynb



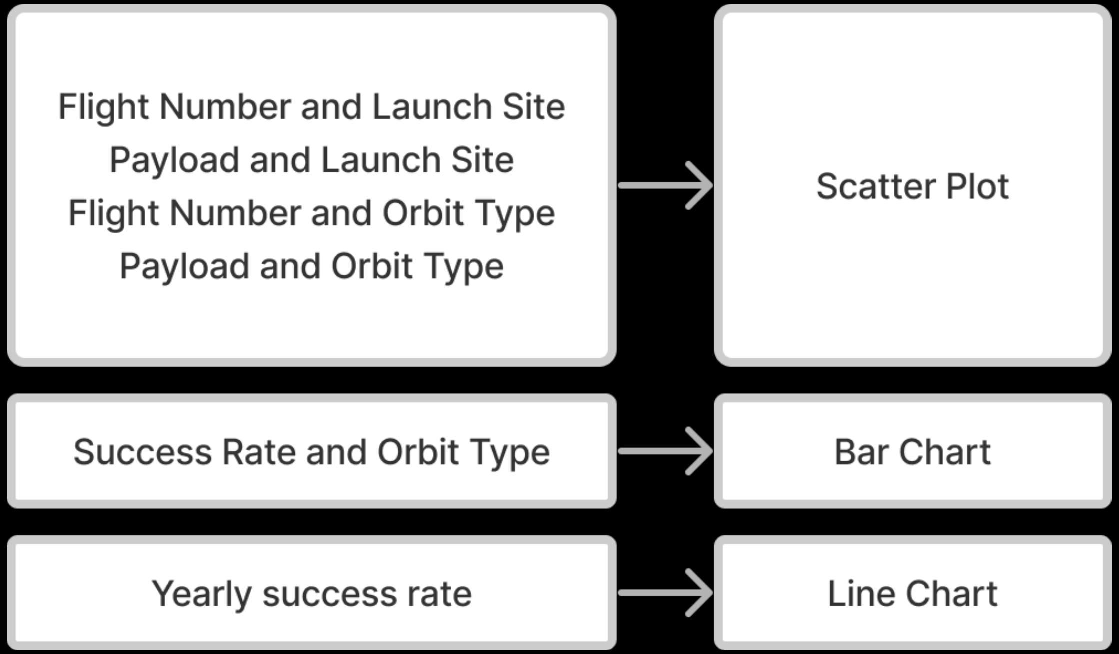
# EDA with Visualization

- Create plots to know relationships between variables
  - Flight Number
  - Launch Site
  - Orbit Type
  - Payload
- Features Engineering
  - One hot encoding
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/5%20-%20EDA%20with%20Visualization.ipynb



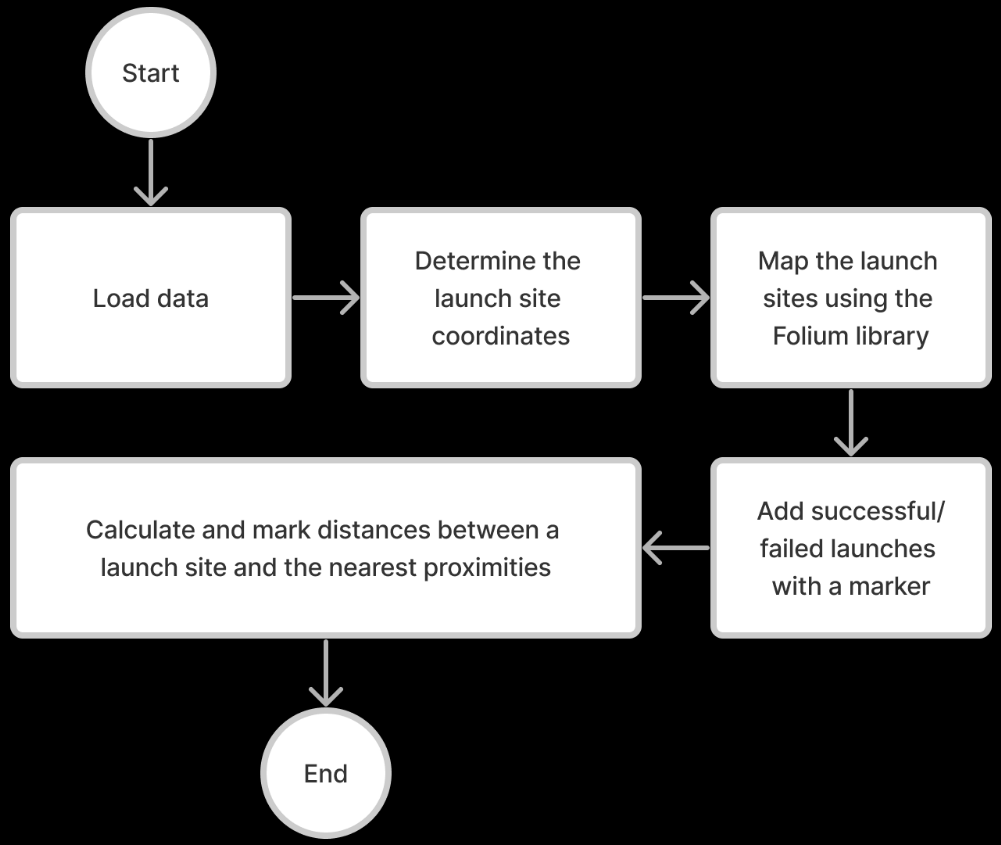
# EDA with Visualization (cont.)

- Scatter plot is appropriate to use with the first four sets of variables to see which launch site and orbit type has the most flight numbers and payloads based on class
- Bar chart is used to know the success rate of the orbit type
- Line chart is used to know yearly success rate of launches



# Interactive Map with Folium

- Load geographical data
- Mark points of interests on the map
- The interactive map must show the markers on the map based on the geographical and launch data provided.
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/6%20-%20Interactive%20Visual%20Analytics%20using%20Folium.ipynb

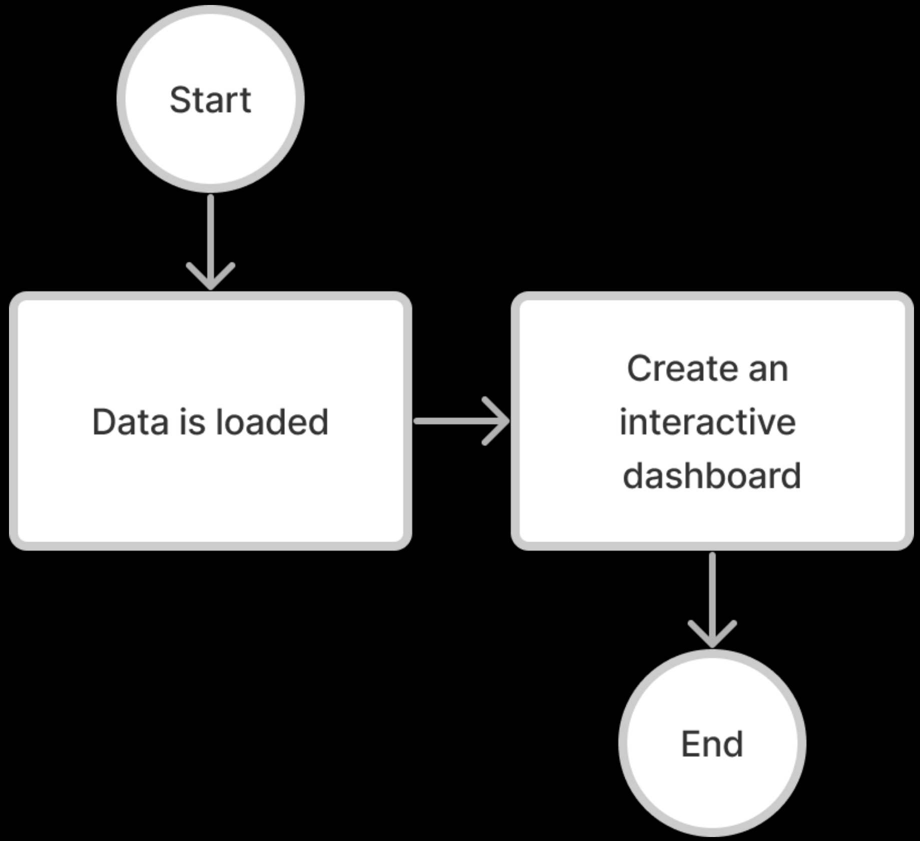


# Interactive Map with Folium (cont.)

- Launch sites are marked with an orange circle with a name.
- Launch successes outcomes are assigned as 1 and 0 respectively.
- Marker clusters determine the amount of successful and failed launches.
- Markers, circles and lines are added to those points of interest.
- Answer the questions:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

# Dashboard with Plotly Dash

- Create an interactive dashboard with Plotly
- A pie chart and a scatter plot with a slider are added.
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/7%20-%20Interactive%20Dashboard%20using%20Plotly%20Dash.py

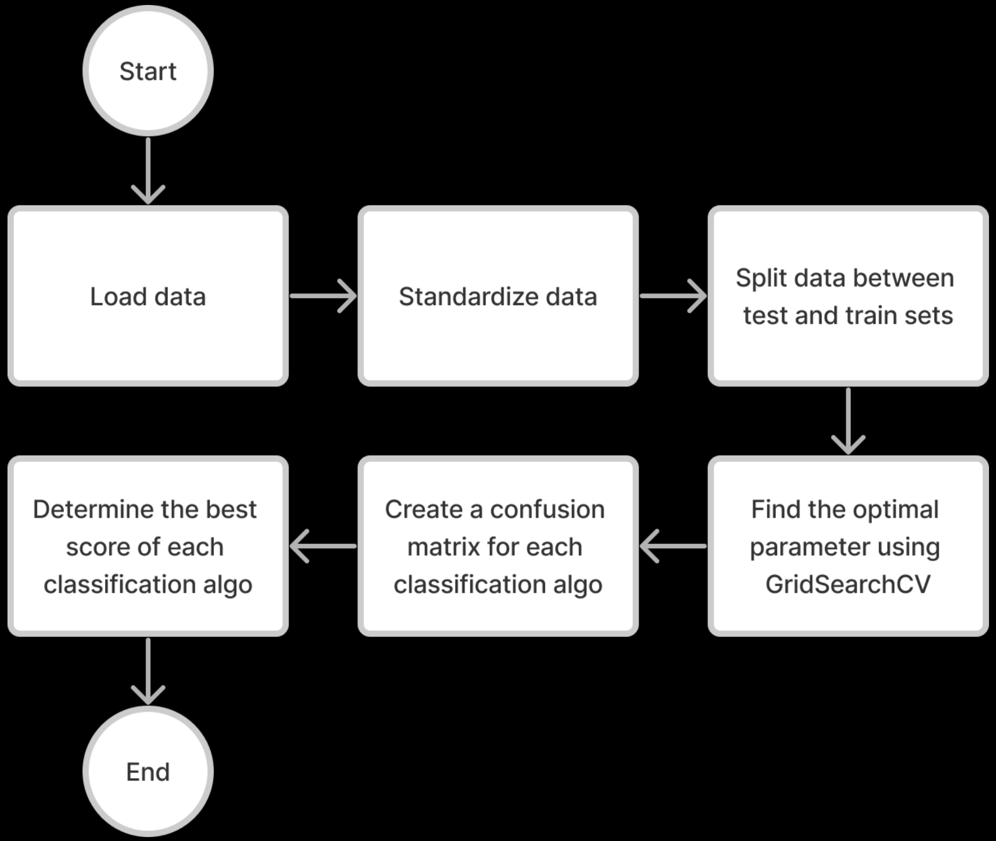


# Dashboard with Plotly Dash (cont.)

- A pie chart is used to determine the successful or failed launches of each launch site.
- A scatter plot with a slider is used to know the successful and failed Outcomes of each launch site with the Payload Mass in kg

# Predictive Analysis

- Load Data
- Standardize the data using StandardScaler
- Split data into train and test sets
- Find the optimal parameter using GridSearchCV
- Create a confusion matrix
- Determine the scores of each classification algorithm
- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/blob/main/8%20-%20Predictive%20Analysis.ipynb



# Predictive Analysis (cont.)

- Predictive analysis uses a machine learning model to predict the next successful launch based on the variables of a future launch.
- Each classification algorithm is tested and are compared to each other with scoring.
- The best classification algorithm will be used for future launches.

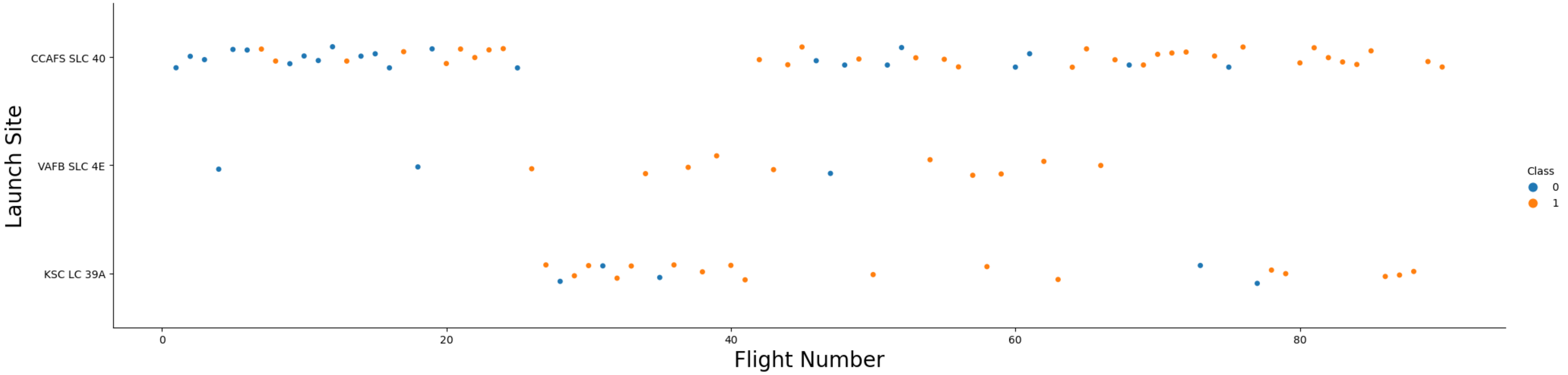
# Results

- Exploratory Data Analysis results
- Interactive Analytics demo in screenshots
- Predictive Analysis results

# Insights Drawn from EDA

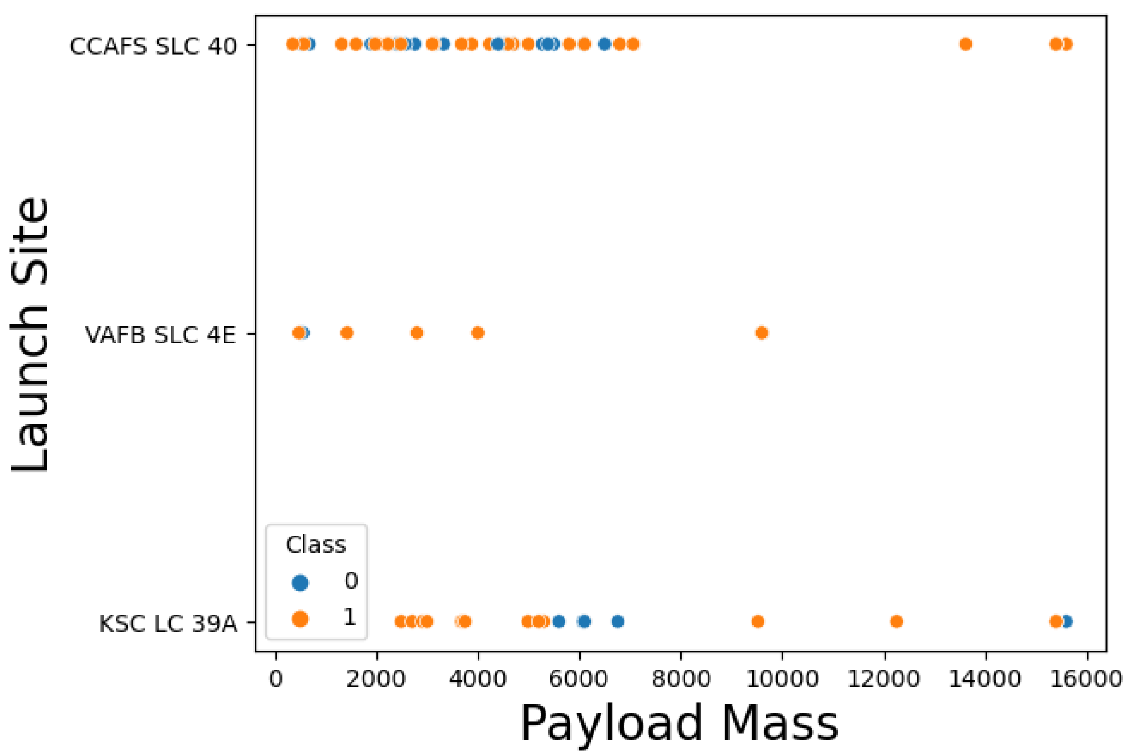
# Flight Number vs. Launch Site

It can be found that CCAFS SLC 40 Launch Site has the greatest number of flights.



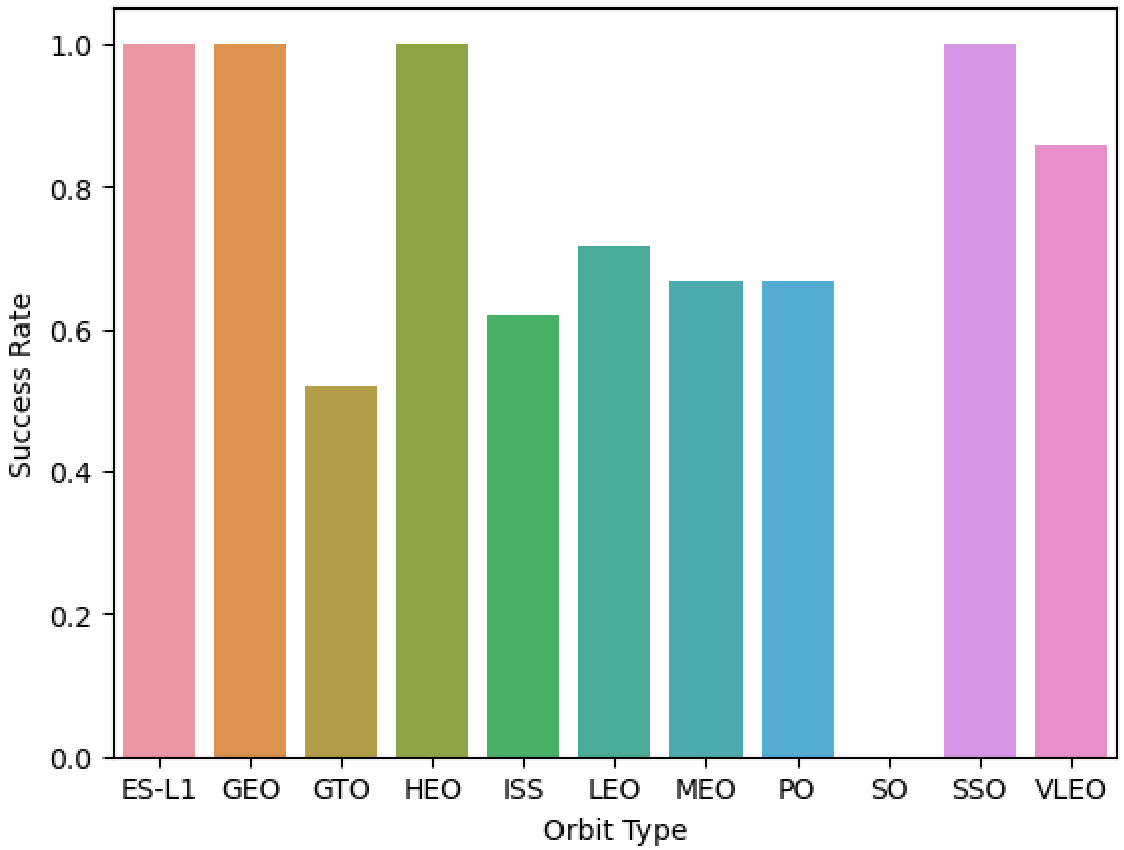
# Payload vs Launch Site

Most launches have a small payload mass (below 6000 KG) which is clear with the CCAFS SLC 40 and KSC LC 30A Launch Sites. However, both launch sites also have launches with payload masses above 14000 KG.



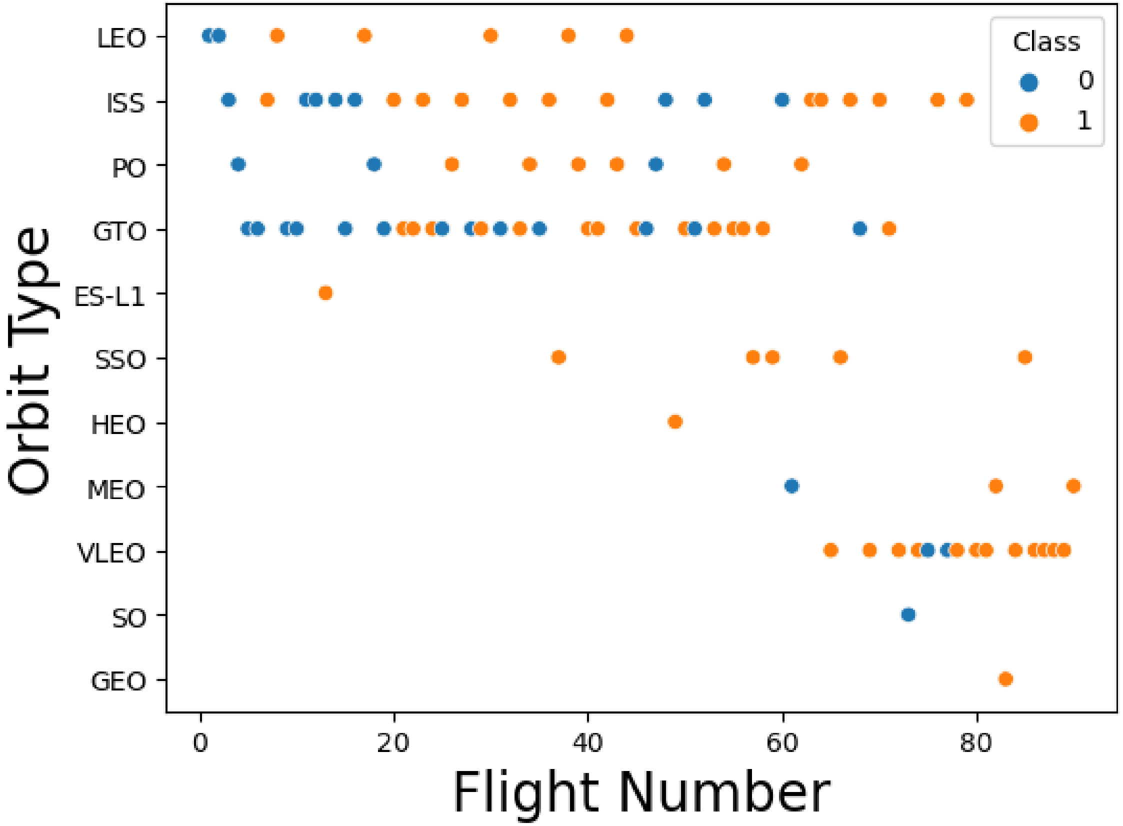
# Success Rate vs. Orbit Type

ES-L1, GEO, HEO and SSO have a very high success rate.



# Flight Number vs. Orbit Type

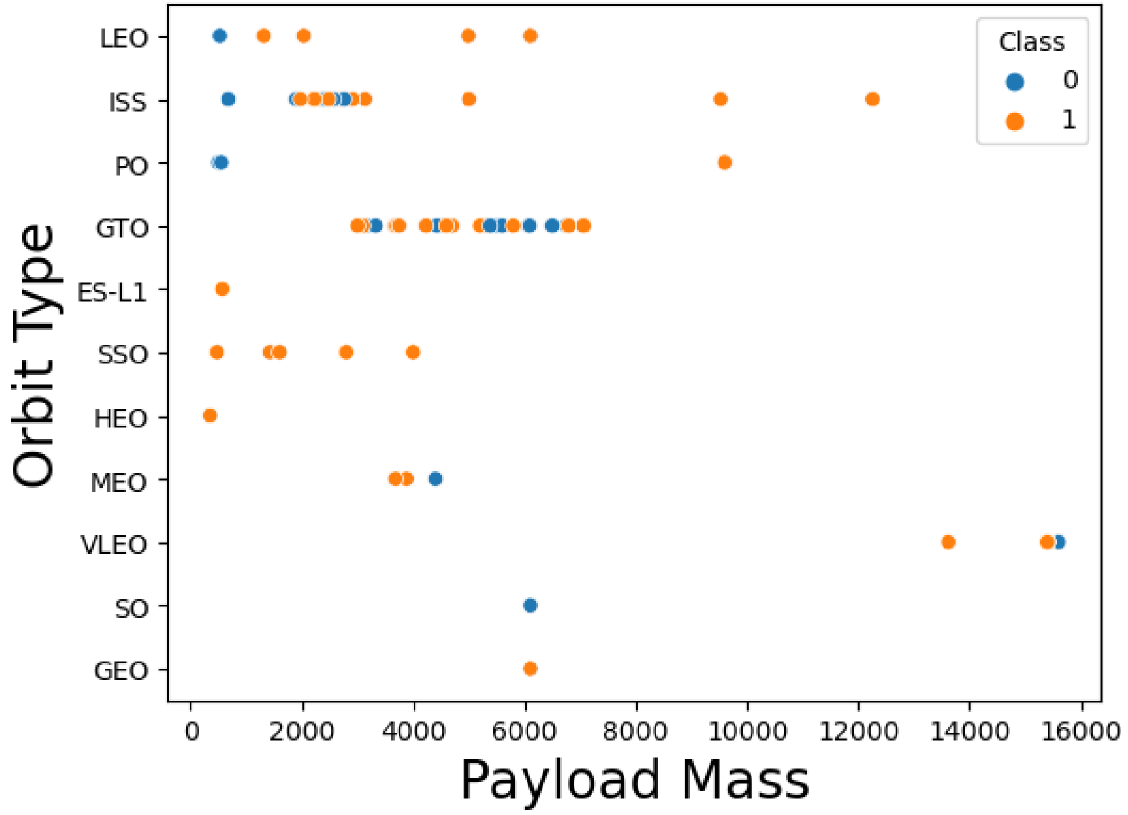
This plot shows that LEO, ISS, PO GTO and VLEO have the most flight numbers.



# Payload vs. Orbit Type

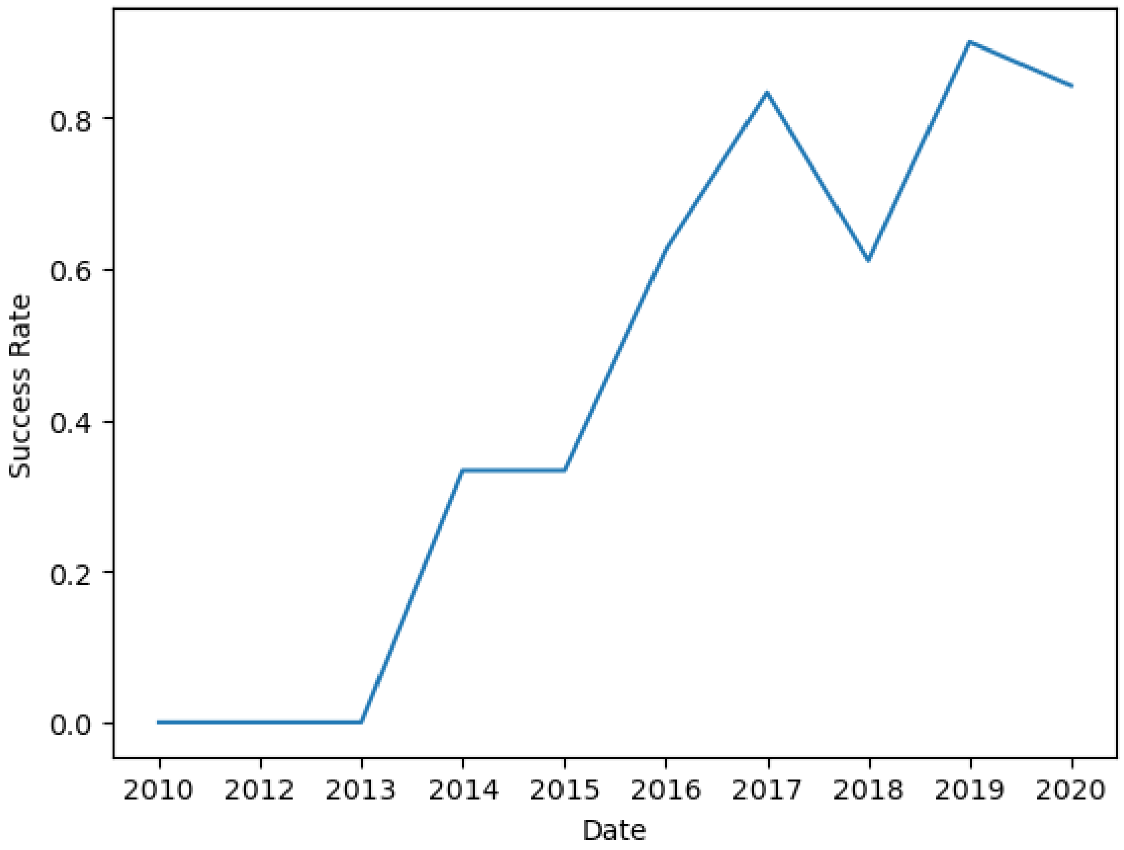
It shows that GTO has the greatest number of launches.

VLEO has the greatest number of launches that have a payload mass of >12000 KG



# Launch Success Yearly Trend

In this line chart, it shows that there is a gradual increase in launch successes.



# All Launch Site Names

- This query returns a list of all unique launch sites from the SpaceX database.
- **DISTINCT** keyword tells the database to only return unique rows.

[5]:
```
%sql select distinct launch_site from spacex
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[5]:

| launch_site |
|---|
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |
| None |

# Launch Site Names that Begin with `CCA`

- Selects all rows from the SpaceX table where the launch_site column contains the string `CCA` followed by any characters and returns only 5 results.
- **LIKE 'CCA%'** tells the database to return rows with the string CCA followed by any characters
- **LIMIT 5** returns the first five results of the database

Display 5 records where launch sites begin with the string 'CCA'

```
[6]: sql select * from spacex where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[6]:

| DATE | time__utc_ | booster_version | launch_site | payload | payload_m |
|---|---|---|---|---|---|
| 2010-04-06 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | |
| 2010-08-12 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | |
| 2012-08-10 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | |
| 2013-01-03 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | |

# Total Payload Mass

- Calculates the total payload mass carried by the boosters launched by NASA.
- Total payload mass is **45596 KG**
- **SELECT SUM(payload__mass_kg_)** returns the sum of the payload mass
- **WHERE customer = ‘NASA (CRS)’** tells the database to return rows where the customer name is NASA (CRS)

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[7]: %sql select sum(payload_mass__kg_) from spacex where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tg
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[7]:

| 1 |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

- This returns the calculated average payload mass carried by the booster F9 v1.1 which is **2928 KG**
- **SELECT AVG(payload__mass_kg_)** returns the average of the payload mass
- **WHERE booster_version = ‘F9 v1.1’** tells the database to select rows where the booster version is F9 v1.1.

```
[8]: sql select avg(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1'

 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0l
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[8]:

| 1 |
|---|
| 2928 |

# First Successful Ground Landing Date

- Returns the earliest date on which a SpaceX booster successfully lands on the ground which is **2015-12-22**.
- **SELECT MIN(date)** tells the database to return the minimum (or earliest) date
- **WHERE landing_outcome LIKE '%(ground pad)`** tells the database to return the rows where landing outcome has the string '(ground pad)'

```
[10]: %sql select min(date) from spacex where landing_outcome like '%(ground pad)'

 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0l
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[10]:

| 1 |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing w/ Payload between 4000 and 6000

- Returns the booster versions of all boosters that have a payload mass between 4000 and 6000 kg and have landed successfully.
- **WHERE (payload_mass__kg_ BETWEEN 4000 and 6000)** tells the database to return rows where payload mass is between 4000 and 6000.

•[13]:
```
%sql select booster_version from spacex where
(payload_mass__kg_ between 4000 and 6000) and
landing_outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0l
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[13]:

| booster_version |
|---|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

**AND landing_outcome = 'Success (drone ship)'** tells the database to return rows where the landing outcome has the string equal to 'Success (drone ship)'

# Total Number of Successful and Failed Mission Outcomes

- Selects the mission_outcome column and the count of the mission_outcome aliased as missionoutcomes and grouped by the mission_outcomes column.
- Returns a table with two columns, mission_outcome and missionoutcomes which contains the count of the mission outcomes.

```
%sql select mission_outcome, count(mission_outcome)
as missionoutcomes from spacex group by mission_outcome
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177t
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

| mission_outcome | missionoutcomes |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |
| None | 0 |

# Boosters Carried Maximum Payload

- Selects the booster version where the payload mass is equal to the maximum value of the payload mass column.
- Returns one column which contains booster versions that have carried the maximum payload mass.
- **MAX(payload_mass__kg_)** takes the maximum payload mass

•[15]:
```
%sql select booster_version from spacex where
payload_mass__kg_ in (select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[15]:

| 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

- Returns the launch records in 2015.
- Selects the booster version, landing outcome and launch sites where the date is in the date 2015 and landing outcome is a failure.
- Returns a table with 3 columns, booster_version, landing_outcome and launch_site

[17]:
```
%sql select booster_version, landing_outcome, launch_site
from spacex where year(date) = 2015 and landing_outcome = 'Failure (drone ship)'
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0l
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[17]:

| booster_version | landing_outcome | launch_site |
|---|---|---|
| F9 v1.1 B1012 | Failure (drone ship) | CCAFS LC-40 |
| F9 v1.1 B1015 | Failure (drone ship) | CCAFS LC-40 |

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

- Returns a table with two columns that has the landing outcome and its count.

[19]:
```
%sql select landing_outcome, count(landing_outcome)
as count from spacex where (date between '2010-06-04' and '2017-03-20')
group by landing_outcome order by count desc
```

```
 * ibm_db_sa://trv42132:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0l
qde00.databases.appdomain.cloud:30426/bludb
Done.
```

[19]:

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

# Launch Sites Proximities Analysis

# Launch Site Locations

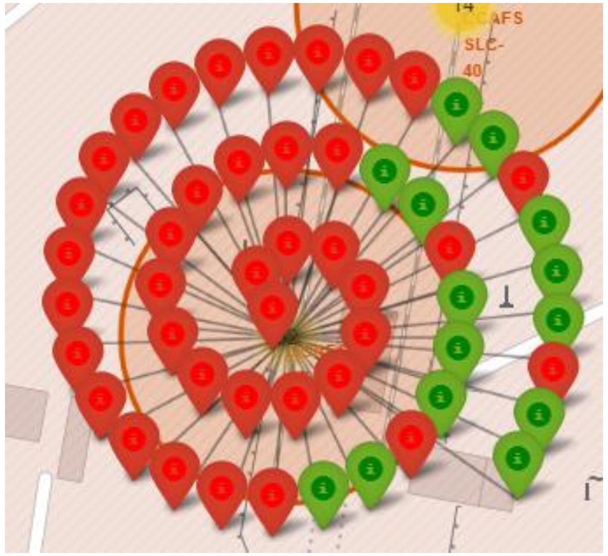
- The map shows the launch sites with a circular marker.
- They are properly labeled.

# Launch Sites Outcome Map

- These show the number of launches for each launch site
- Red marker shows a failed launch
- Green marker shows a successful launch
- CCAFS LC-40 has the most failed launches
- KSC LC-39A has the most successful launches

CCAFS SLC-40

KSC LC-39A

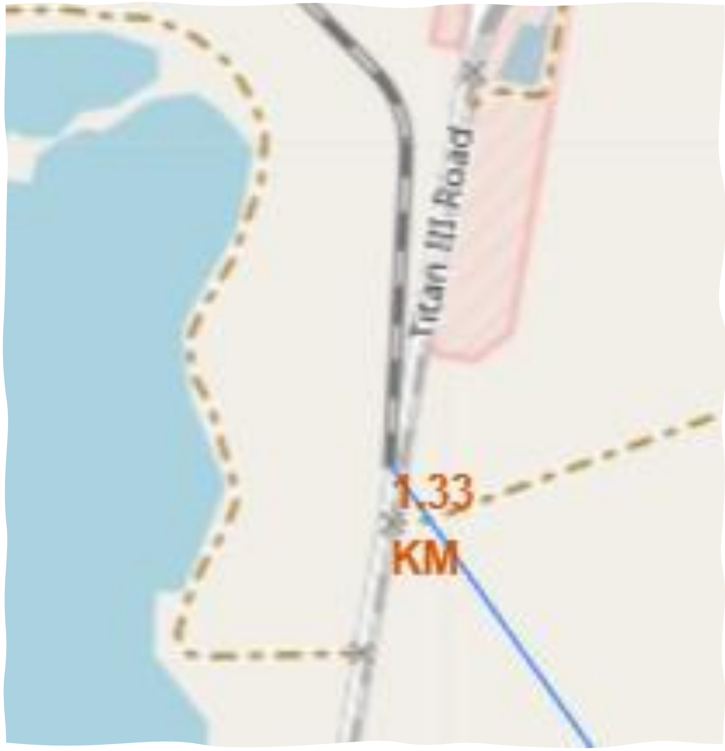


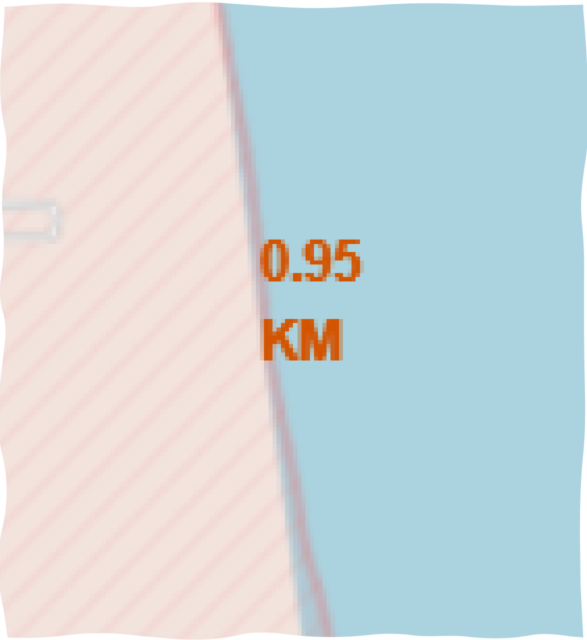
CCAFS LC-40

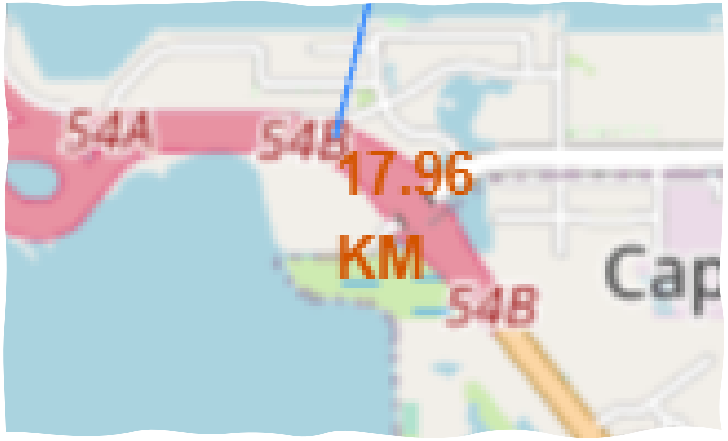

VAFB SLC-4E

# Proximities from Points of Interests

- These points of interests are the areas nearest to the launch site CCAFS LC-40.
- Cape Canaveral is 18.10 KM away
- Nearest Railway is 1.33 KM away
- Nearest Coastline is 0.95 KM away
- Nearest highway is 17.96 KM away

# Questions

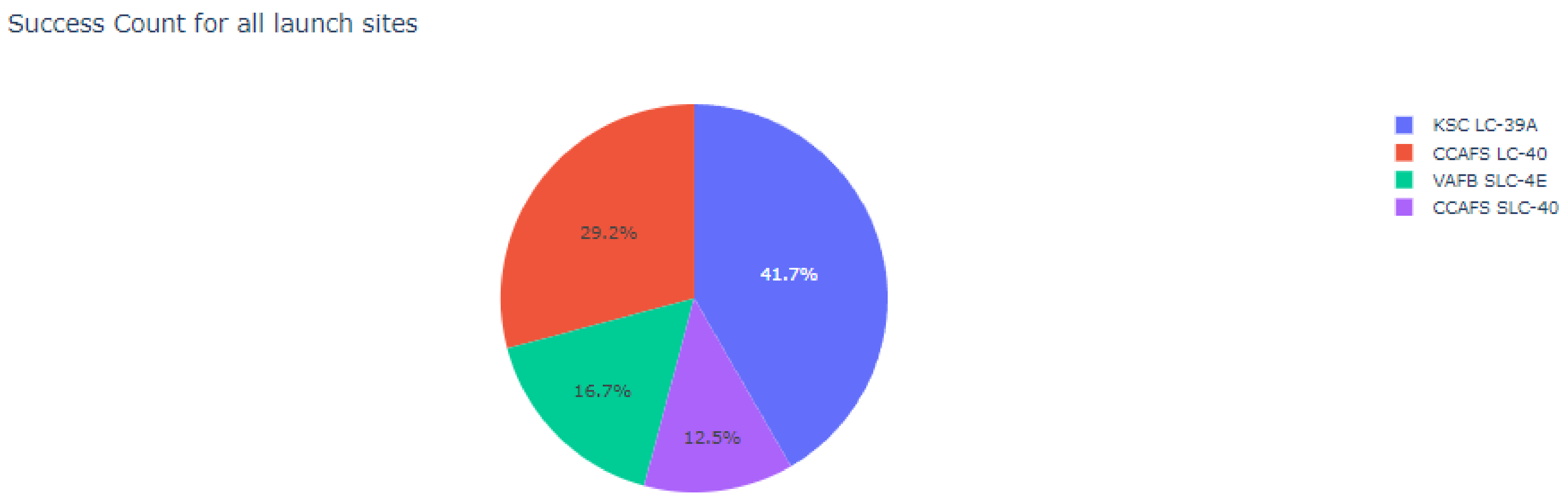
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

These launch sites must be far away from most of the human population to avoid losses in case there are errors in launch. They are also accessible to railways since a train can handle heavy payloads that will be delivered to these launch sites.

# Dashboard with Plotly Dash

# Launch Site Success

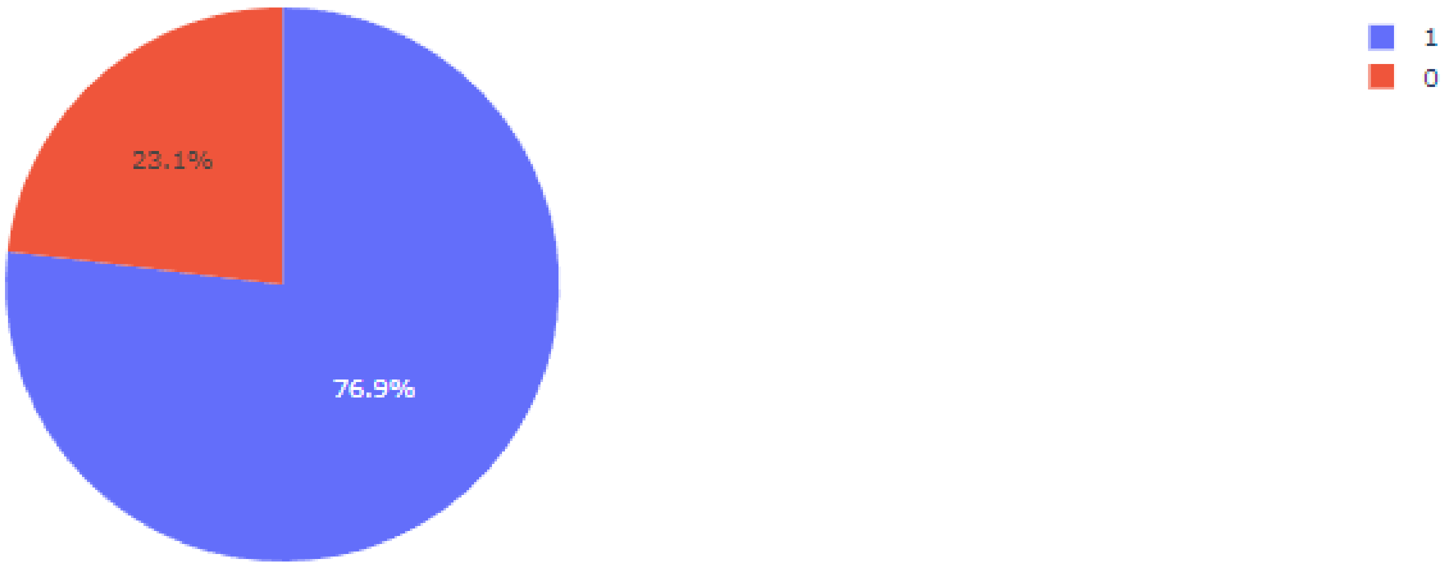
- This pie chart shows that the CCAFS SLC-40 has the highest success rate among the four launch sites.



# Site with the Highest Success Rate

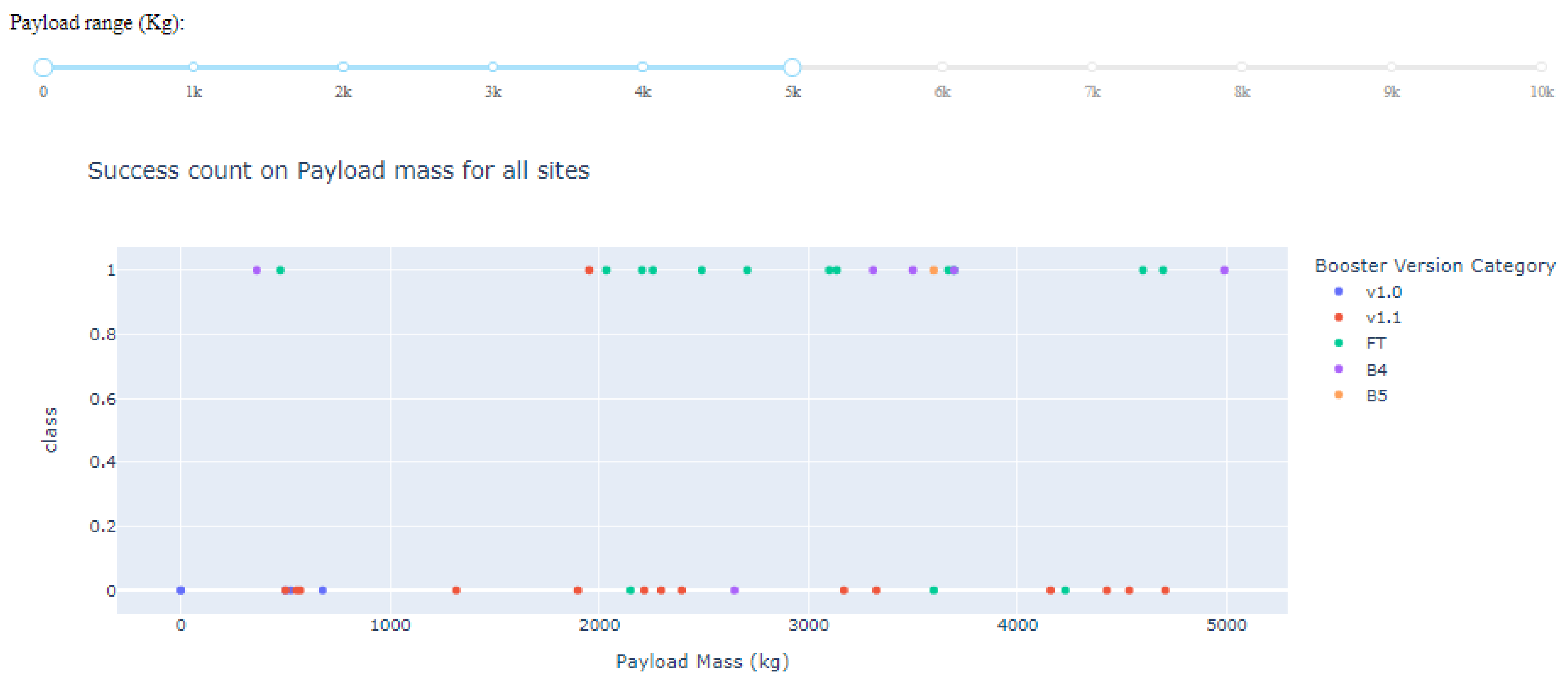
- KSC LC-39A has the success to failed launch ratio.

Total Success Launches for site KSC LC-39A



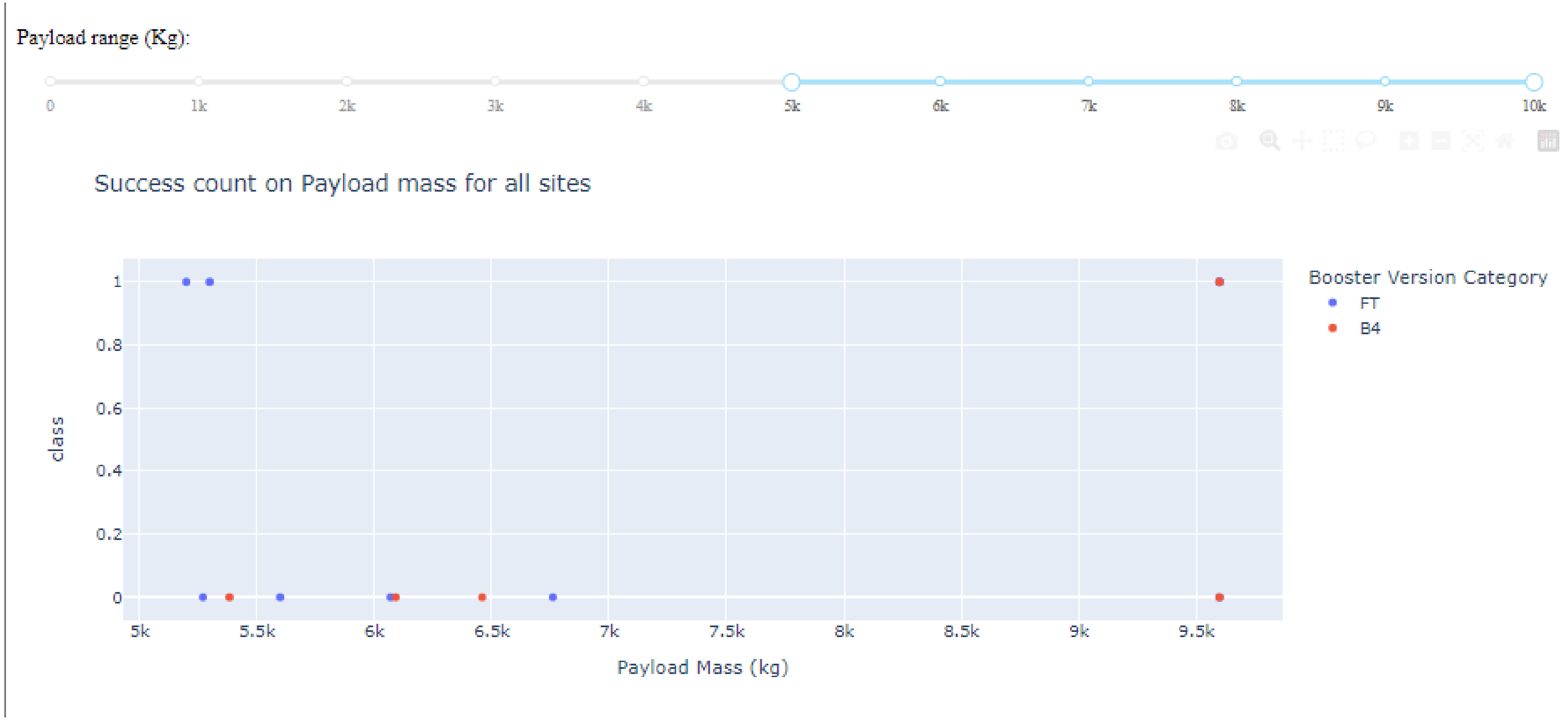
# Payload vs Launch Outcome for all Sites

- There is a higher probability of a successful launch with smaller payloads



# Payload vs Launch Outcome for all Sites (cont.)

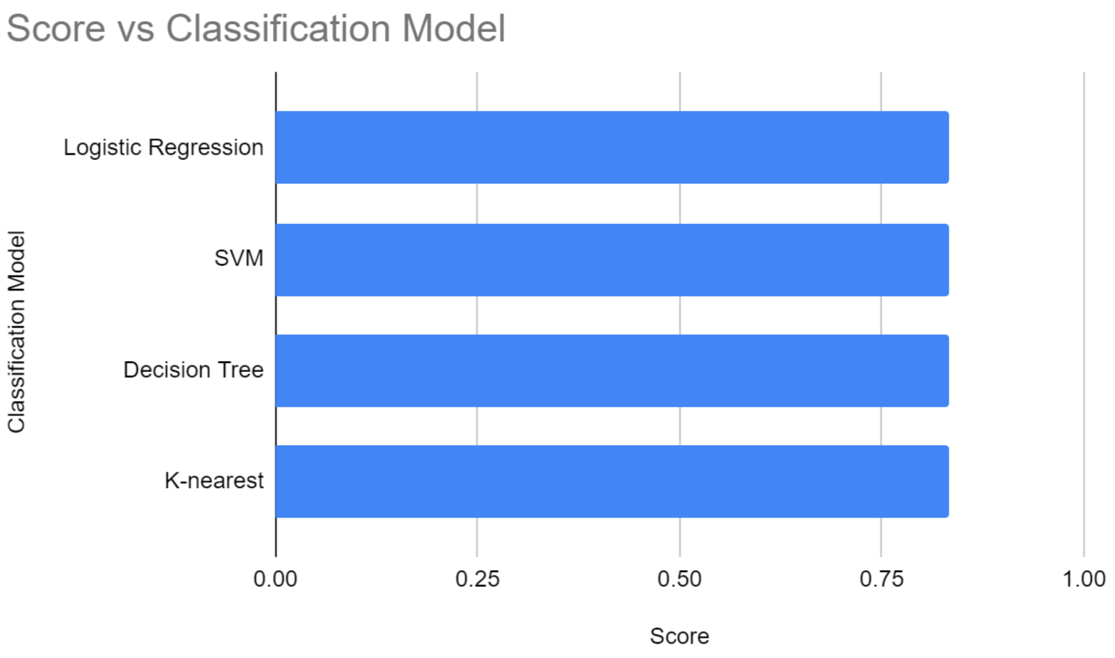
- There are only 3 successful launches with a >5000 kg payload



# Predictive Analysis (Classification)

# Classification Accuracy

- All classification models showed the same accuracy.



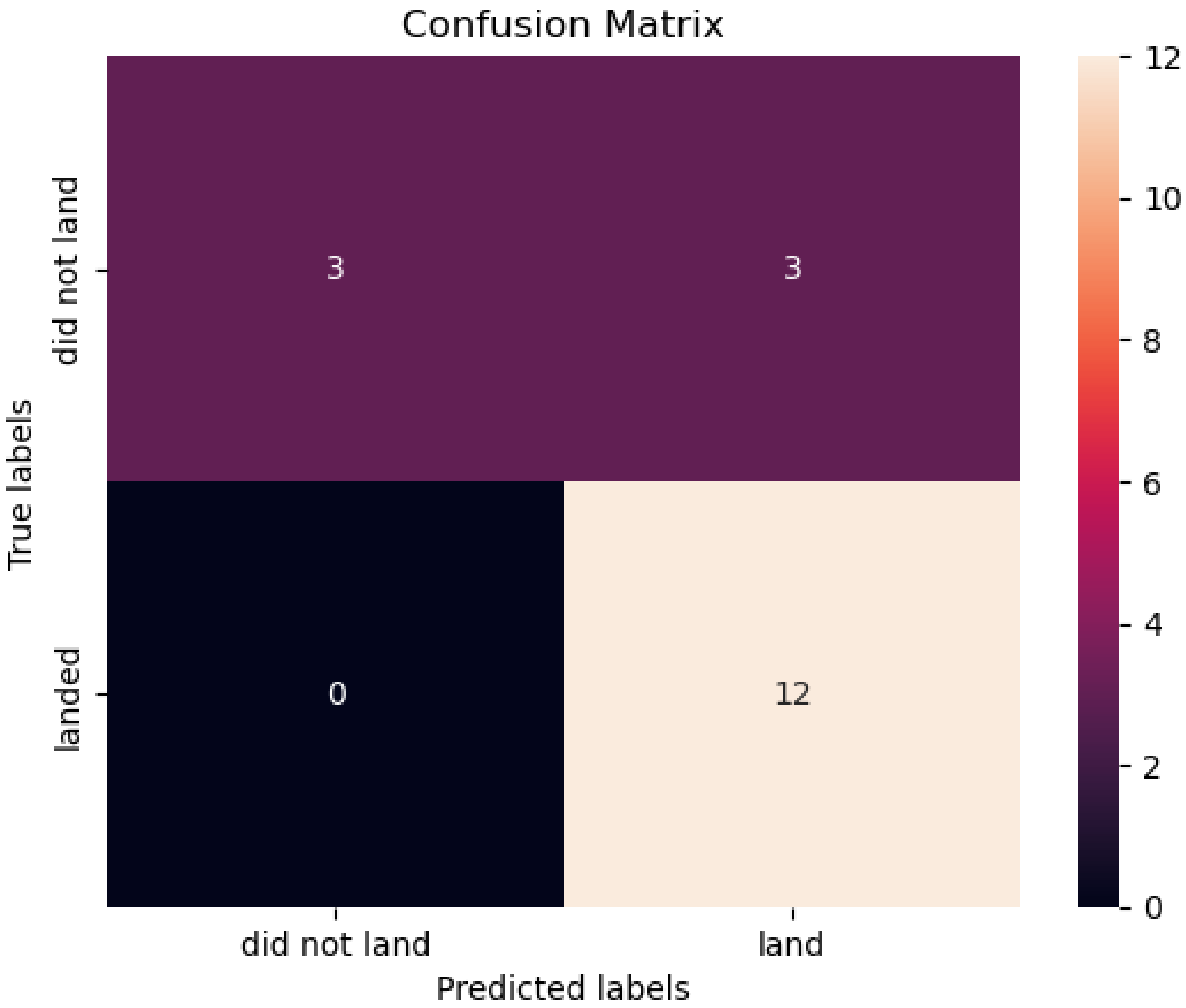

Find the method performs best:

```
[44]: print('Logistic Regression', logreg_cv.score(X_test, Y_test)
      print('SVM', svm_cv.score(X_test, Y_test))
      print('Decision Tree', tree_cv.score(X_test, Y_test))
      print('K-nearest', knn_cv.score(X_test, Y_test))
```

```
Logistic Regression 0.8333333333333334
SVM 0.8333333333333334
Decision Tree 0.8333333333333334
K-nearest 0.8333333333333334
```

# Confusion Matrix

- The model shows that the true positive and false positives have the same value of 3.



# Insights and Conclusions

# Insights and Conclusions

- The main factor behind the success of a flight is primarily the payload mass and orbit type.
- Launches with orbit types specifically used for Earth observation/communication have higher launch successes.
- In addition to point #2, these successful launches have short orbital distances from the surface of the Earth.
- More data is needed for better machine learning prediction.

# Insights and Conclusions

- KSC LC-39A has the highest flight success because most of the payloads are under 5000 kg
- It is observed that with CCAFS SLC-40 and VAFB SLC-4E, a smaller payload has a higher probability of a successful launch
- The launch success increases every year
- Rockets with lighter payloads can be launched in areas further inland due to a higher probability of a launch success

# Recommendation

# Recommendation

- More launch data is needed for better analysis

# Appendix

# Github Link

- https://github.com/Olrak29/IBM-Data-Analytics-Capstone/tree/main

# Thank You!