IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Nada Nemr
2024/6/1

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection using 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

# Introduction

- Project background and context
  - With the recent successes in private space travel, space industry is becoming more and more mainstream and accessible to general population. Cost of launch continues to remain a key barrier for new competitors to enter the space race
  - SpaceX with its first stage reuse capabilities offers a key advantage against its competitors. Each SpaceX launch costs around 62 million dollar and SpaceX can reuse stage 1 for future launches. This provides SpaceX a unique advantage where other competitors are spending around 165 mission plus for each launch
- Problems you want to find answers
  - Determine if the first stage of SpaceX Falcon 9 will land successfully
  - Impact of different parameters/variables on the landing outcomes (e.g., launch site, payload mass, booster version, etc.)
  - Correlations between launch sites and success rates

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
    - SpaceX API
    - Get requests to the Space X API and web scraping from Wikipedia
    - Perform data wrangling
    - Determined labels for training the supervised models by converting mission outcomes in to training labels (0-unsuccessful, 1-successful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Created a column for ‘class’; standardized and transformed data; train/test split data; find best classification algorithm (Logistic regression, SVM, decision tree, & KNN) using test data

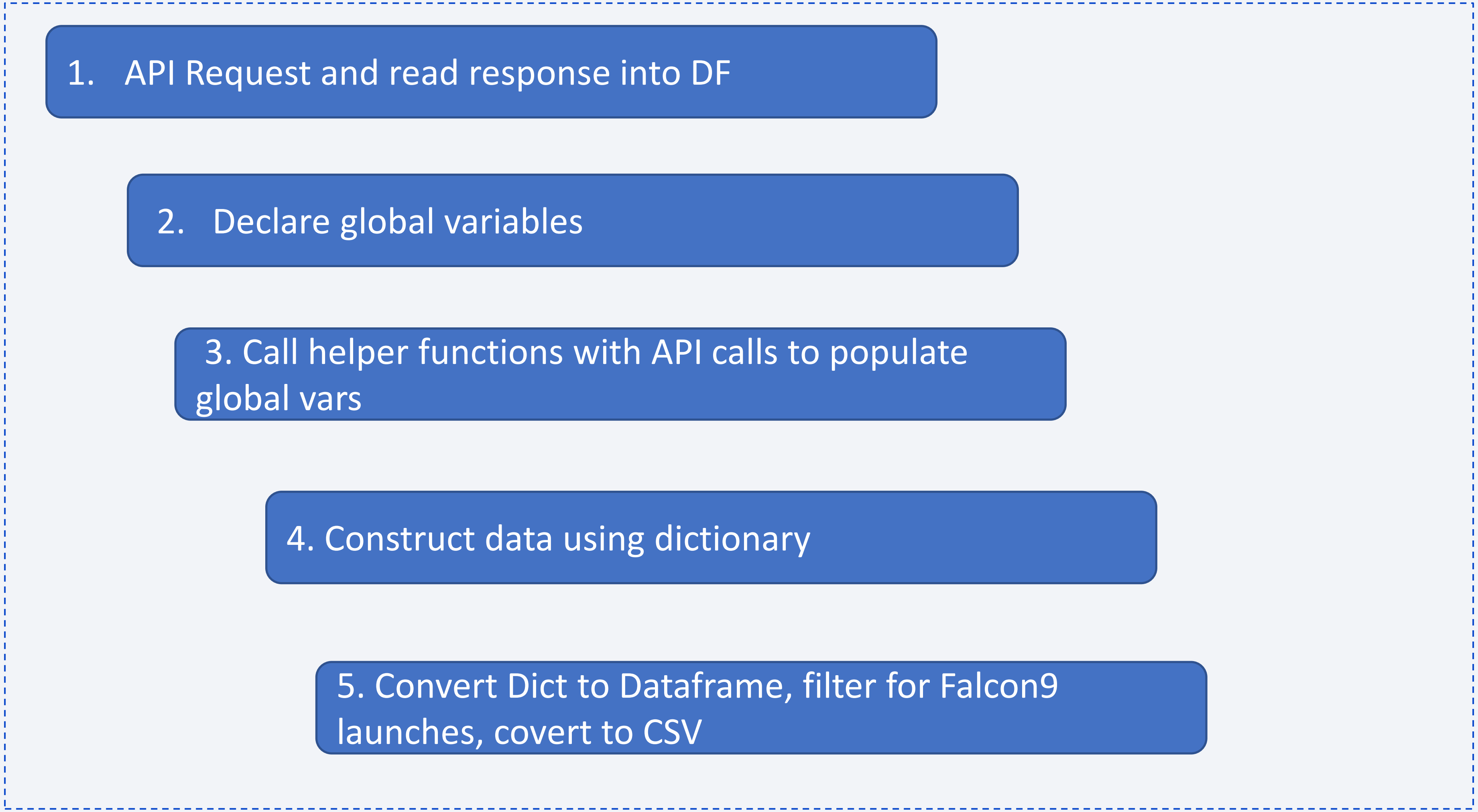
# Data Collection

- Data collection is the process of gathering data from available sources. This data can be structured, unstructured, or semi-structured. For this project, data was collected via SpaceX API and Web scrapping Wiki pages for relevant launch data.

  1. Enter the URL of the page you want to analyze for this project
  2. Request and parse the SpaceX launch data using the GET request
  3. decode the response content as a JSON and turn it into a Pandas data frame
  4. now use the API again to get information about the launches using the IDs given for each launch
  5. Filter the data frame to only include Falcon 9 launches and replace null values and get required output

# Data Collection – SpaceX API

- GitHub



1. API Request and read response into DF

2. Declare global variables

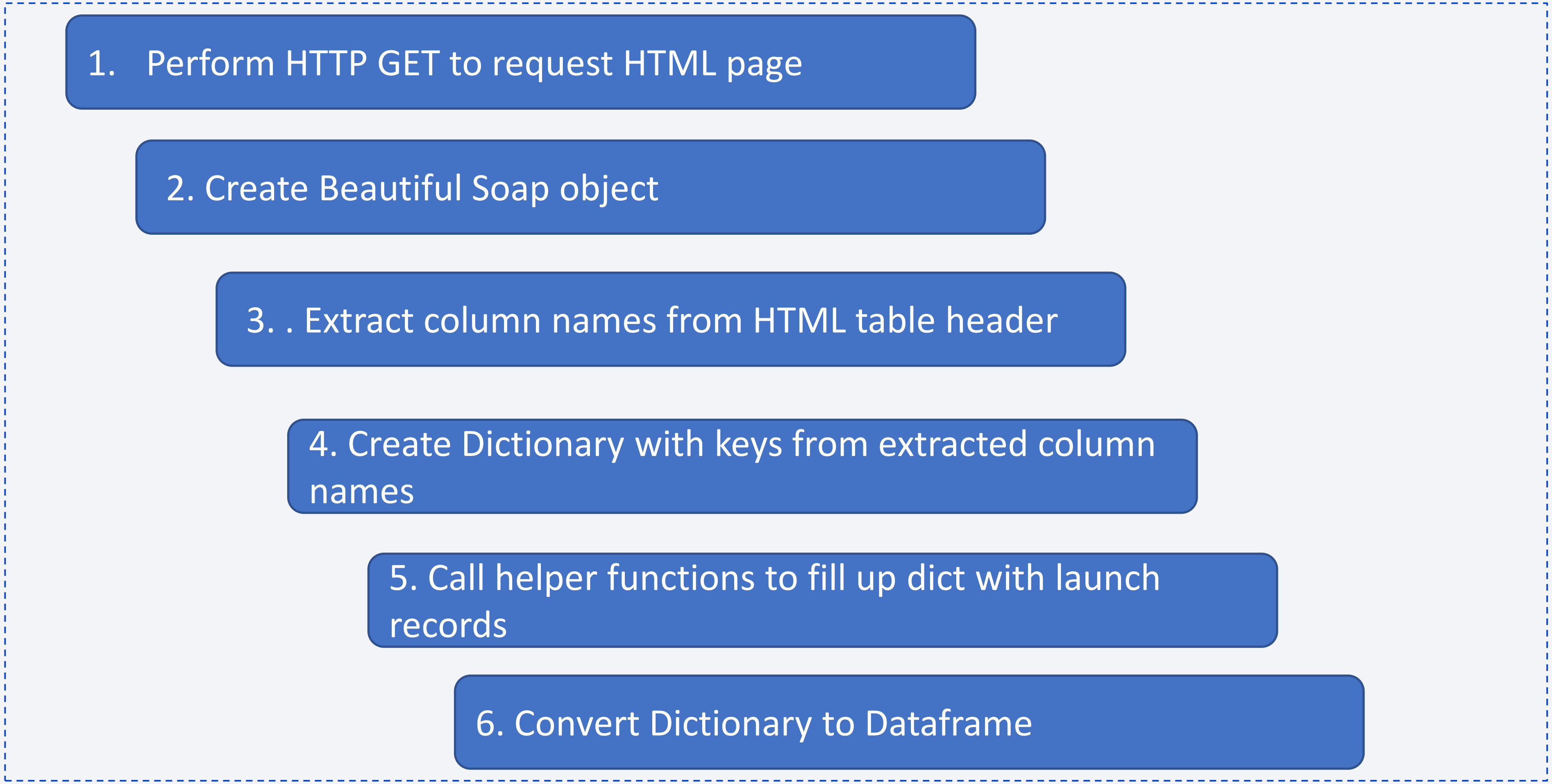
3. Call helper functions with API calls to populate global vars

4. Construct data using dictionary

5. Convert Dict to Dataframe, filter for Falcon9 launches, covert to CSV

# Data Collection - Scraping

- GitHub



1. Perform HTTP GET to request HTML page

2. Create Beautiful Soap object

3. . Extract column names from HTML table header

4. Create Dictionary with keys from extracted column names

5. Call helper functions to fill up dict with launch records

6. Convert Dictionary to Dataframe

# Data Wrangling

- GitHub

1. Load dataset in to Dataframe

2. Calculate the number of launches on each site

3. Calculate the number and occurrences of each orbit

4. Calculate the number and occurence of mission outcome per orbit type

5. Create a landing outcome label from Outcome column

# EDA with Data Visualization

- **As part of the Exploratory Data Analysis (EDA), following charts were plotted to gain further insights into the dataset**:

  1. **scatter plot :** Flight Number vs Payload Mass , Flight Number vs Launch Sites , Payload and Launch Sites , Flight Number and Orbit Type , Payload and Orbit Type
  2. **Bar chart :** Success rate of each orbit
  3. **Line plot :** success rate and Date

- GitHub

# EDA with SQL

## Summary of SQL queries that were used:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string ‘CCA
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 - List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Github

# Build an Interactive Map with Folium

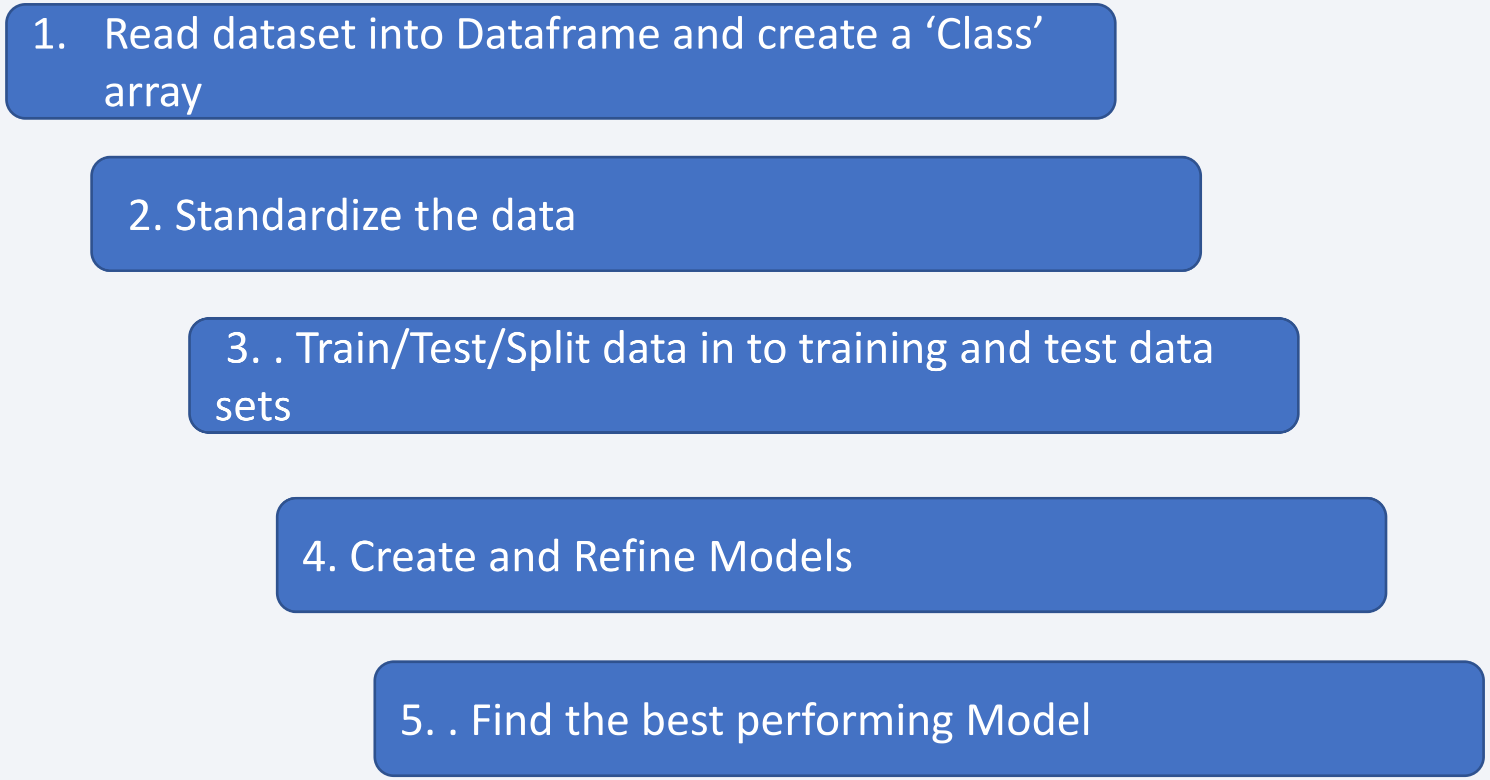
- Folium interactive map helps analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact launch success rate.
- Mark all launch sites on the map. This allowed to visually see the launch sites on the map.
  - Added ‘folium.circle’ and ‘folium.marker’ to highlight circle area with a text label over each launch site
- Added a ‘MarkerCluster()’ to show launch success (green) and failure (red) markers for each launch site.
- GitHub

# Build a Dashboard with Plotly Dash

- Pie charts and scatter charts were used to visualize the launch records of Space X.
- These charts displayed the rocket launch success rate per launch site. We are were able to get an understanding of the factors that may have been influencing the success rate at each site. Such as the payload mass and booster versions.
- Successful launches were represented by 1 while failures were represented by 0.
- Github

# Predictive Analysis (Classification)

- GitHub



1. Read dataset into Dataframe and create a ‘Class’ array

2. Standardize the data

3. . Train/Test/Split data in to training and test data sets

4. Create and Refine Models

5. . Find the best performing Model

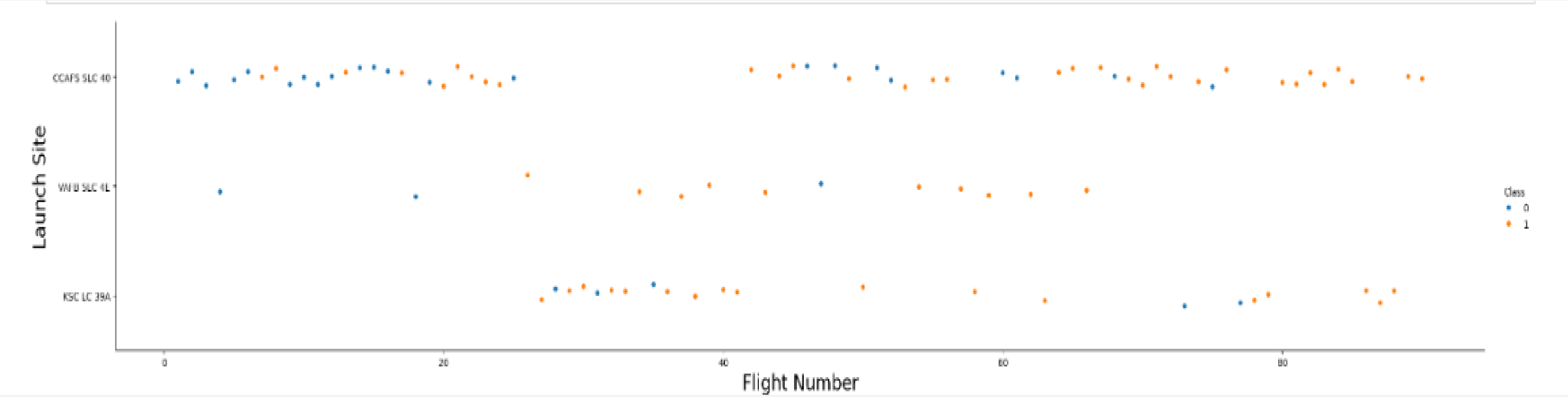
# Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning were able to predict the landing success of rockets with an accuracy score of 83.33%

Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site



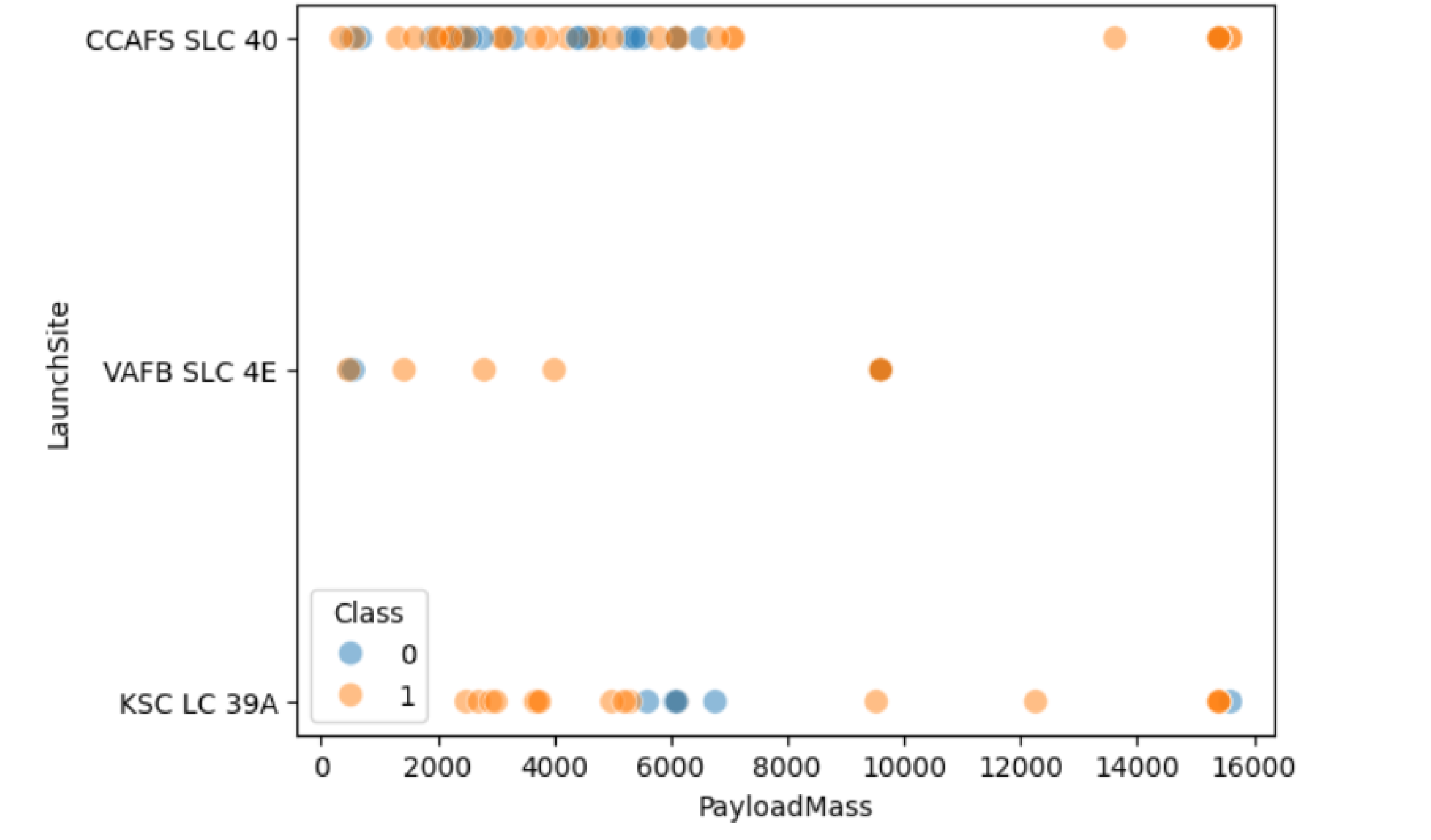
- Success rates (Class=1) increases as the number of flights increase
- For launch site ‘KSC LC 39A’, it takes at least around 25 launches before a first successful launch

# Payload vs. Launch Site

- For launch site ‘VAFB SLC 4E’, there are no rockets launched for payload greater than 10,000 kg

- Percentage of successful launch (Class=1) increases for launch site ‘VAFB SLC 4E’ as the payload mass increases

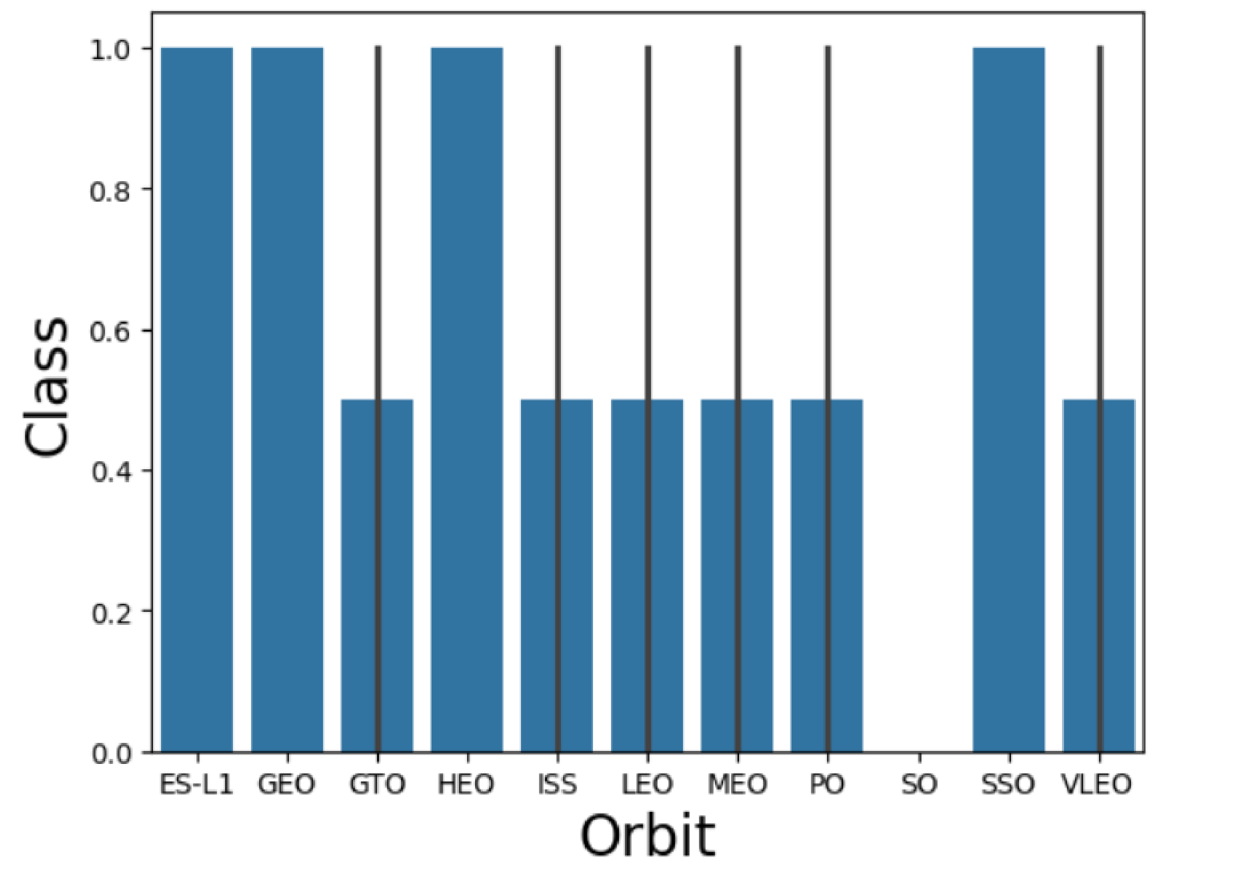
- There is no clear correlation or pattern between launch site and payload mass



# Success Rate vs. Orbit Type

- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates

- GTO orbit has the lowest success rate

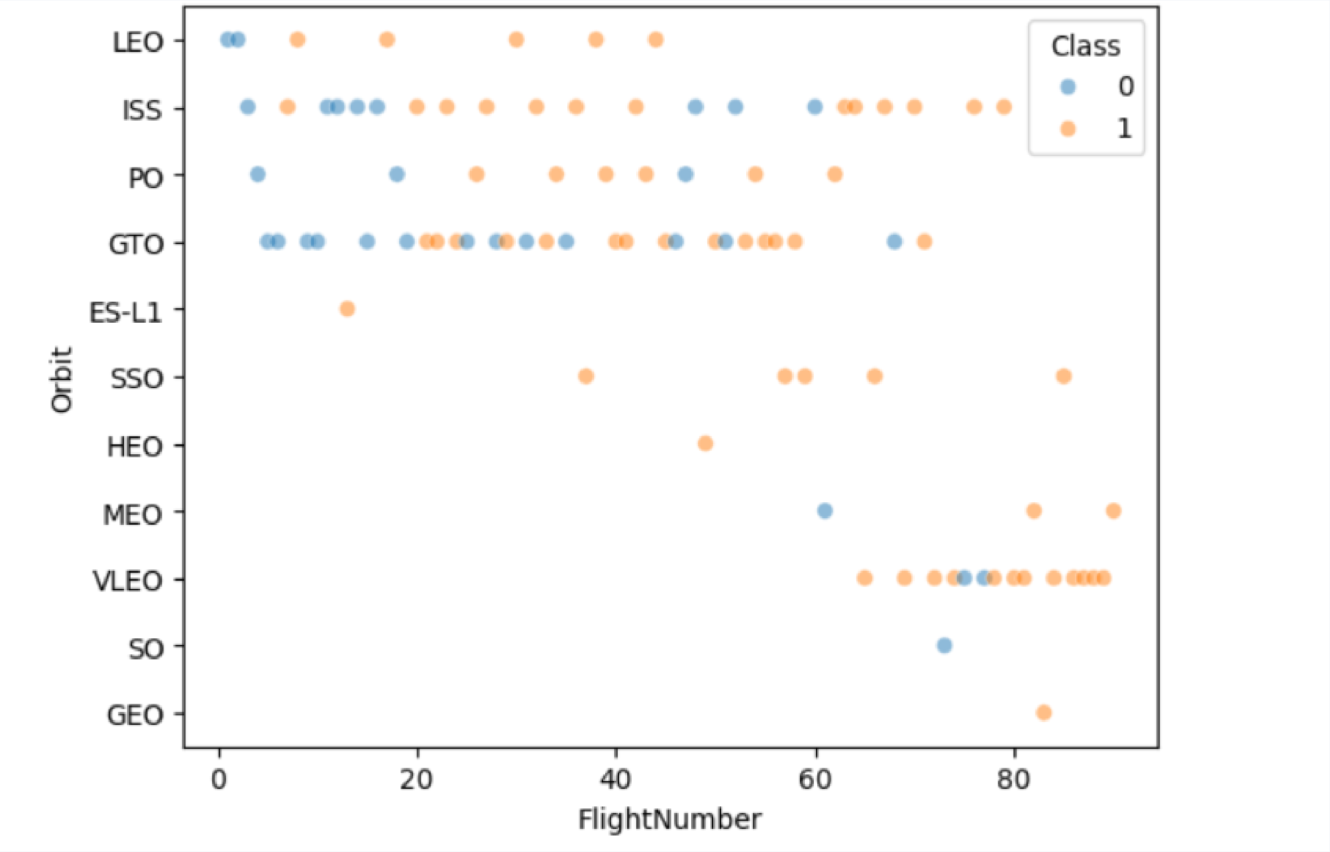


# Flight Number vs. Orbit Type

- For orbit VLEO, first successful landing (class=1) doesn’t occur until 60+ number of flights

- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers

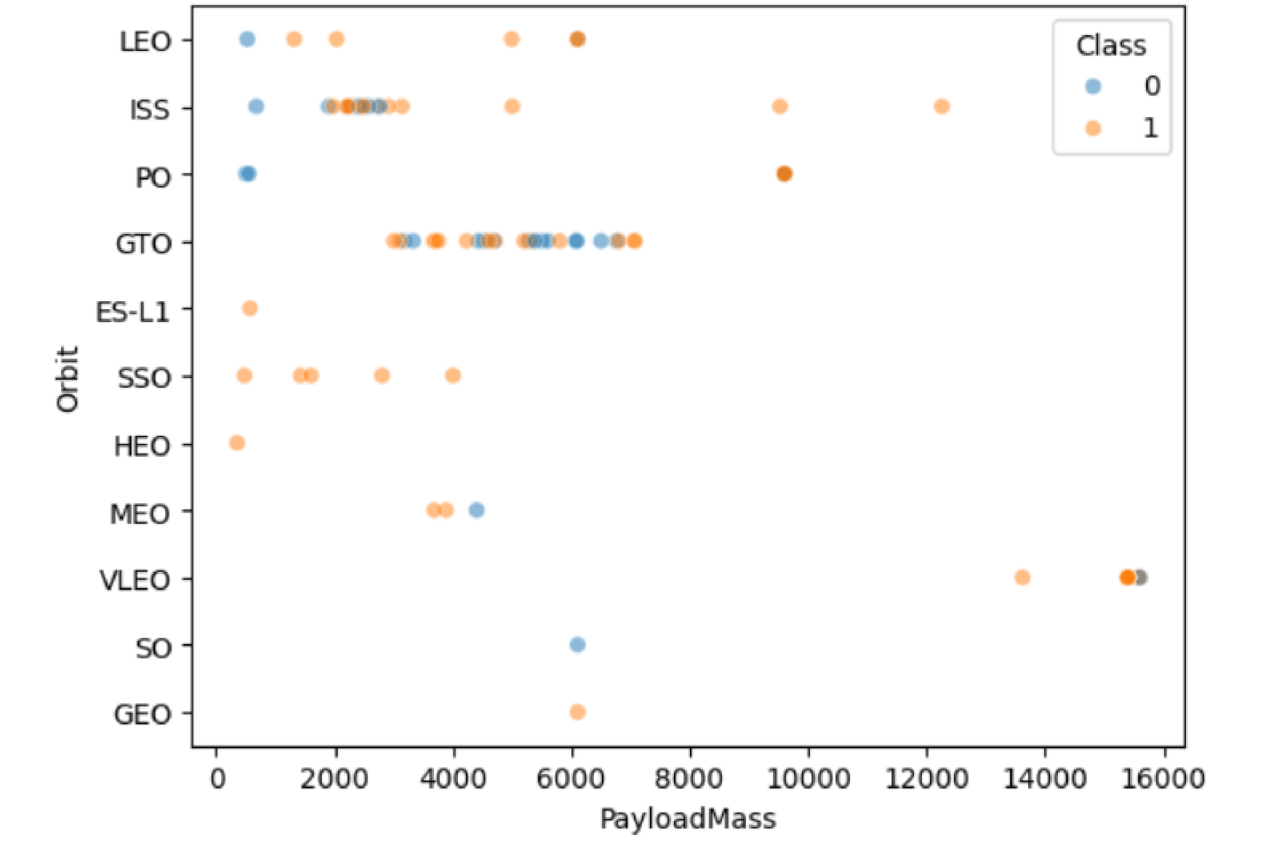
- There is no relationship between flight number and orbit for GTO



# Payload vs. Orbit Type

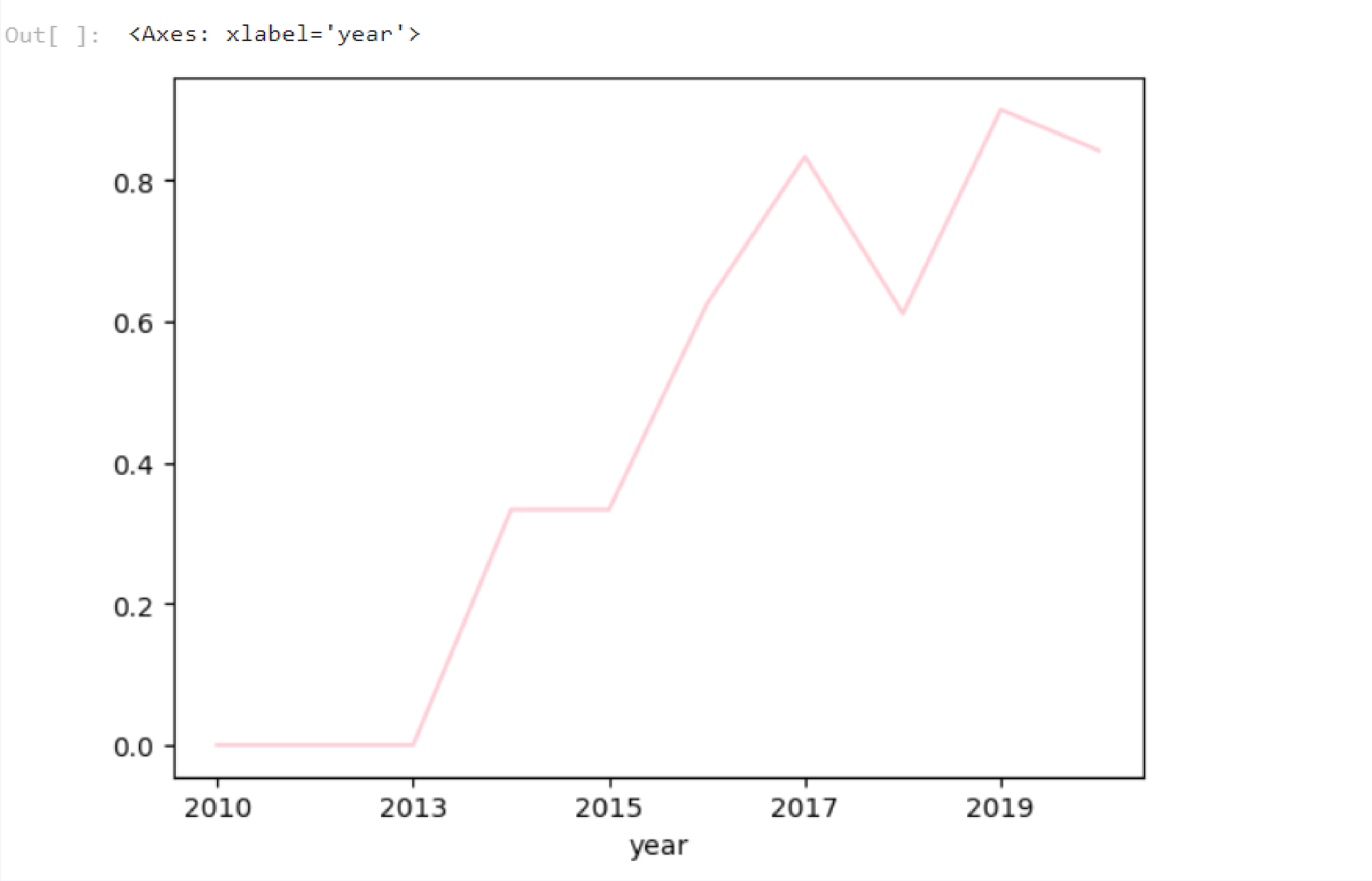
- Successful landing rates (Class=1) appear to increase with pay load for orbits LEO, ISS, PO, and SSO

- For GEO orbit, there is not clear pattern between payload and orbit for successful or unsuccessful landing



# Launch Success Yearly Trend

- Success rate (Class=1) increased by about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015
- Success rates decreased between 2017 and 2018 and between 2019 and 2020

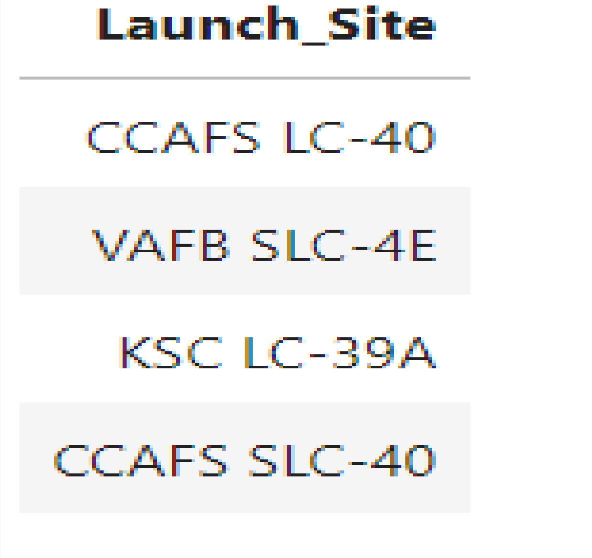


```
Out[ ]:  <Axes: xlabel='year'>
```

# All Launch Site Names

**Query:**

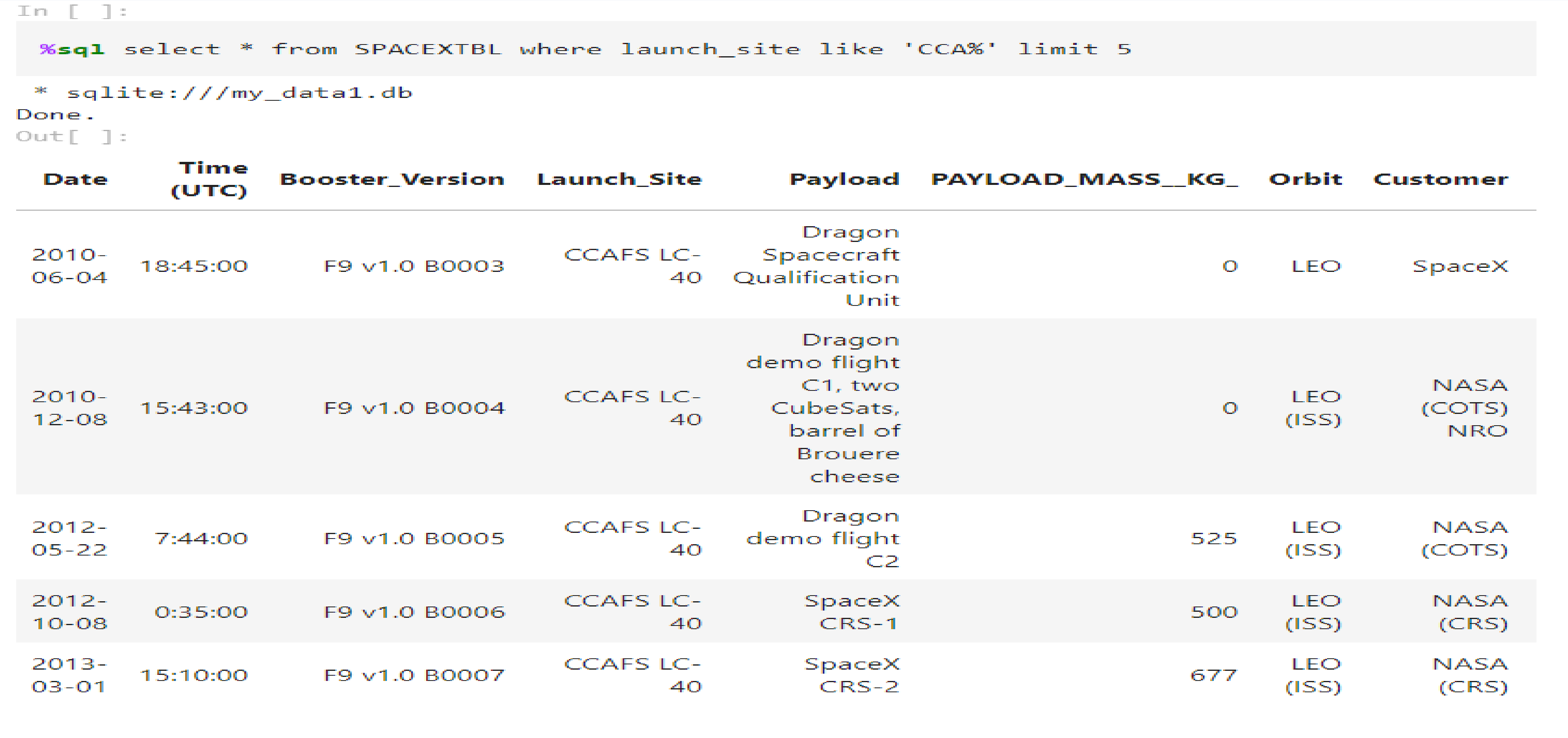
**%sql** select distinct launch_site from SPACEXTBL



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

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

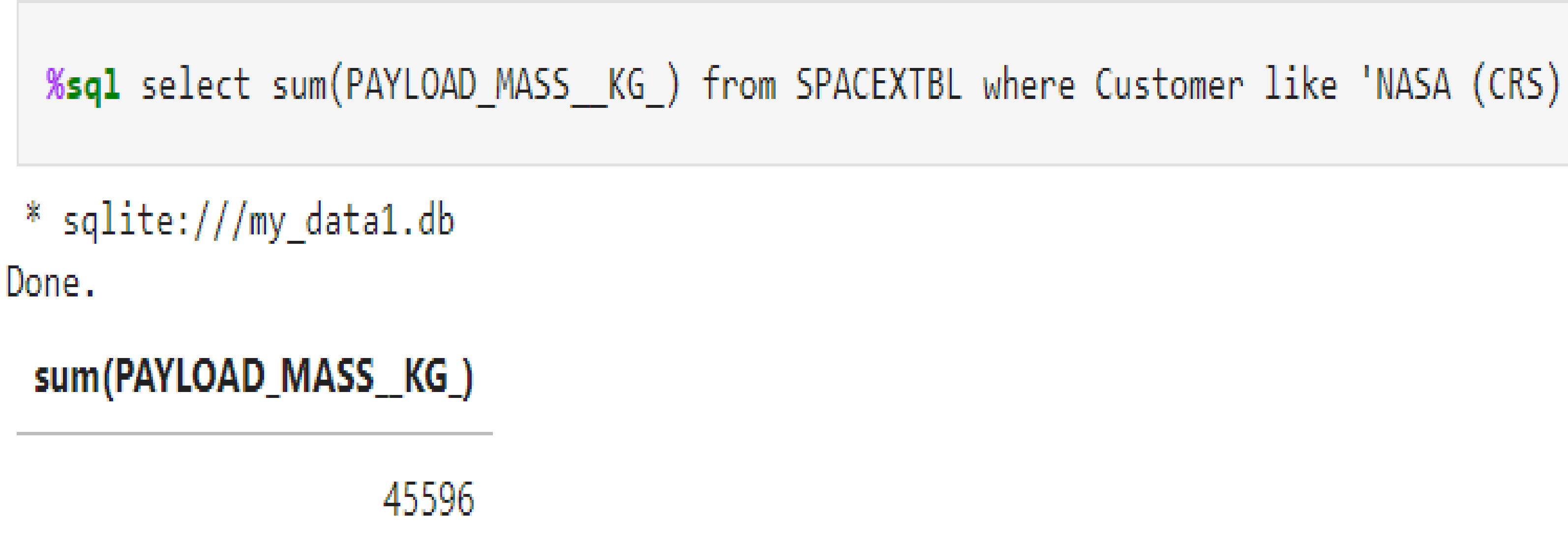


```
In [ ]:
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
Out[ ]:
```

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer |
|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX |
| 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 |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) |

# Total Payload Mass

- Calculate the total payload carried by boosters from NASA



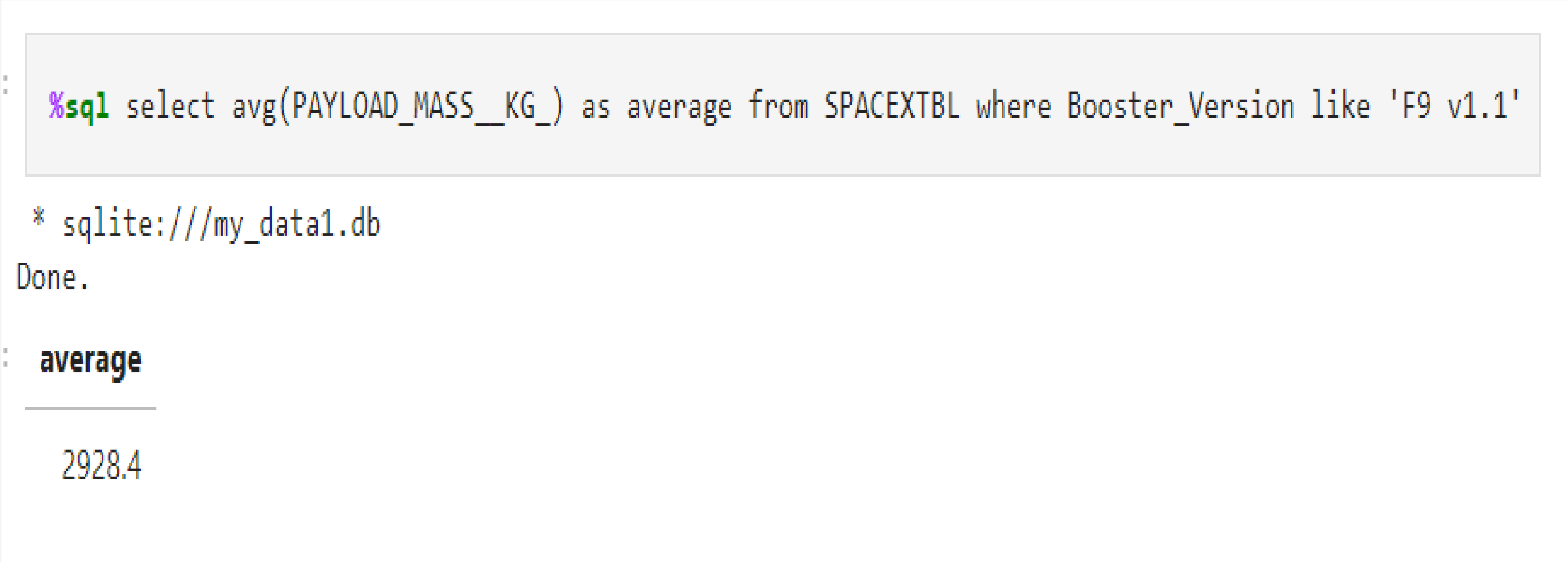
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA (CRS)
```

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

| sum(PAYLOAD_MASS__KG_) |
|---:|
| 45596 |

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1



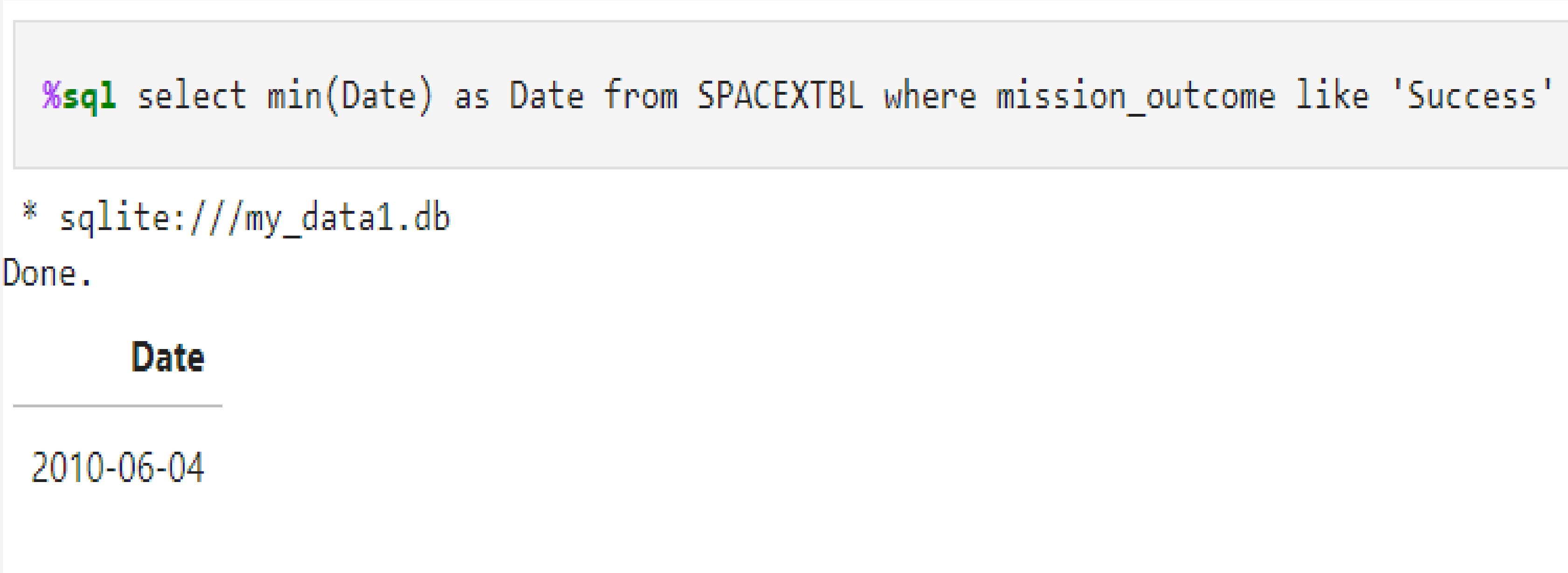
```
%sql select avg(PAYLOAD_MASS__KG_) as average from SPACEXTBL where Booster_Version like 'F9 v1.1'
```

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

| average |
|---|
| 2928.4 |

# First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad



```
%sql select min(Date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

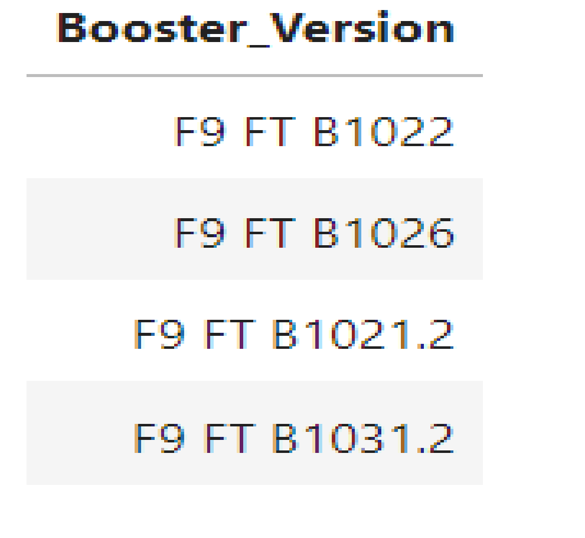
```
 * sqlite:///my_data1.db
Done.
```

| Date |
|---|
| 2010-06-04 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query:

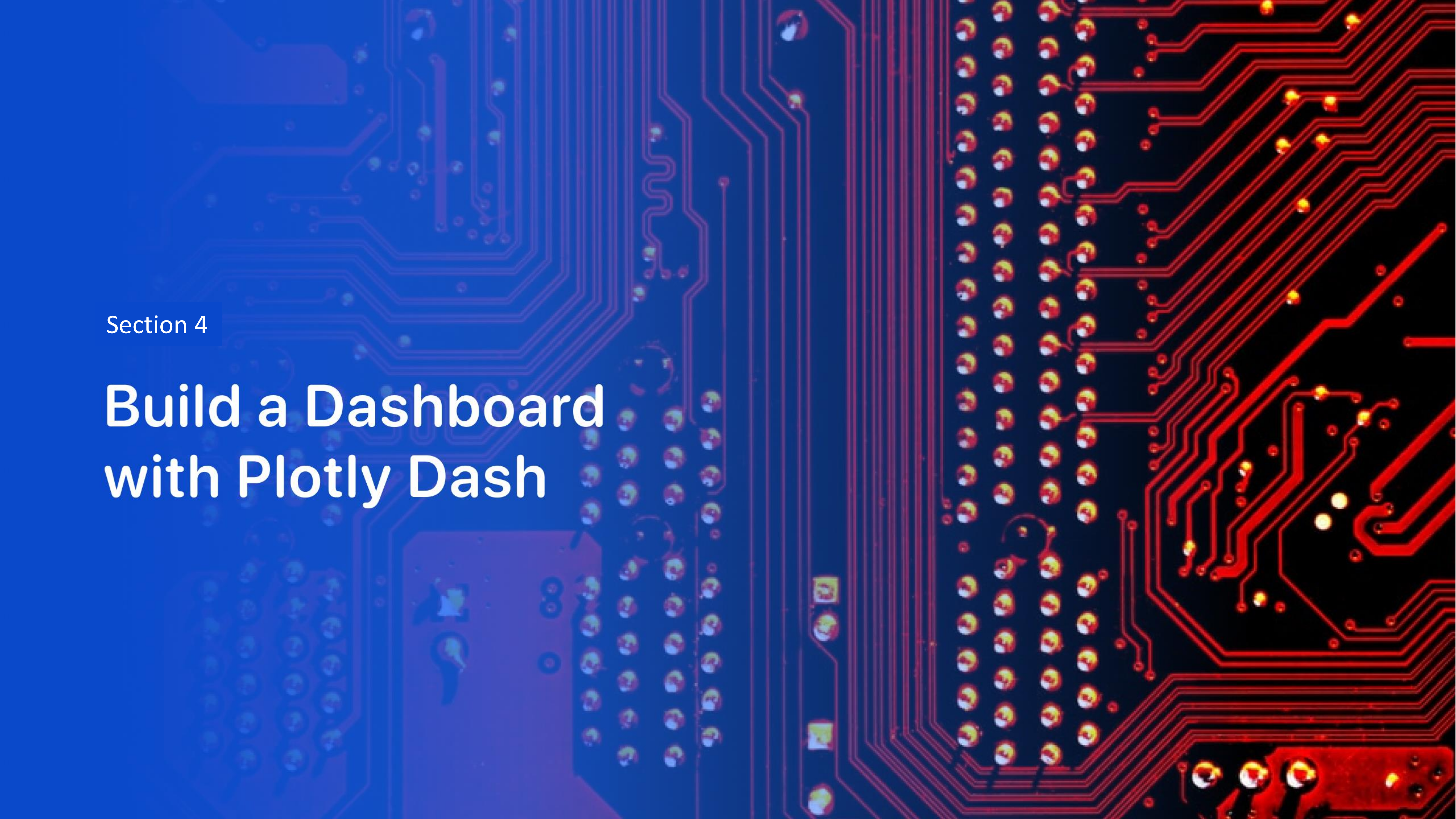
```
%sql select Booster_Version from SPACEXTBL where (mission_outcome like 'Success') and (PAYLOAD_MASS__KG_ between 4000 and 6000) and
(Landing_Outcome like 'Success (drone ship)')
```



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

Section 4

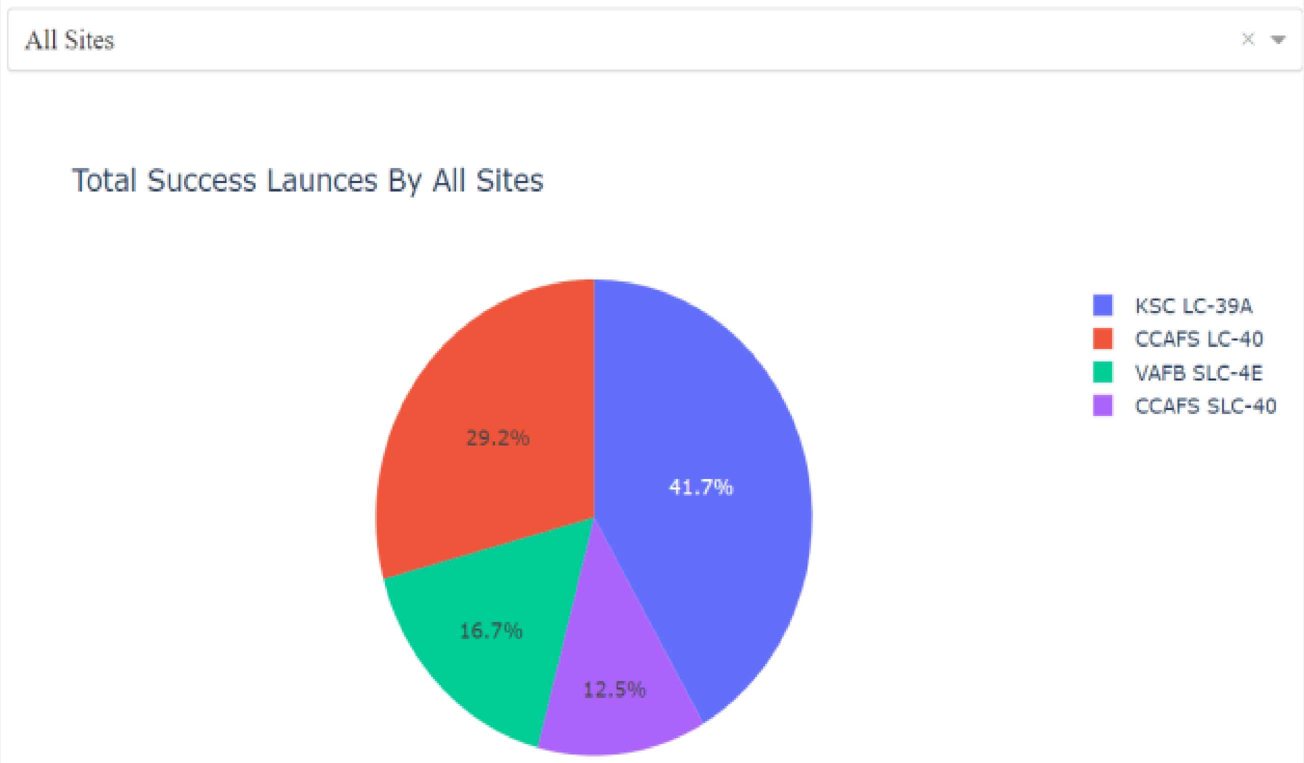
# Build a Dashboard with Plotly Dash



# Successful Launches by Site

- Launch Site ‘KSC LC-39A’ has the highest launch success rate

- Launch Site ‘CCAFS SLC40’ has the lowest launch success rate



All Sites

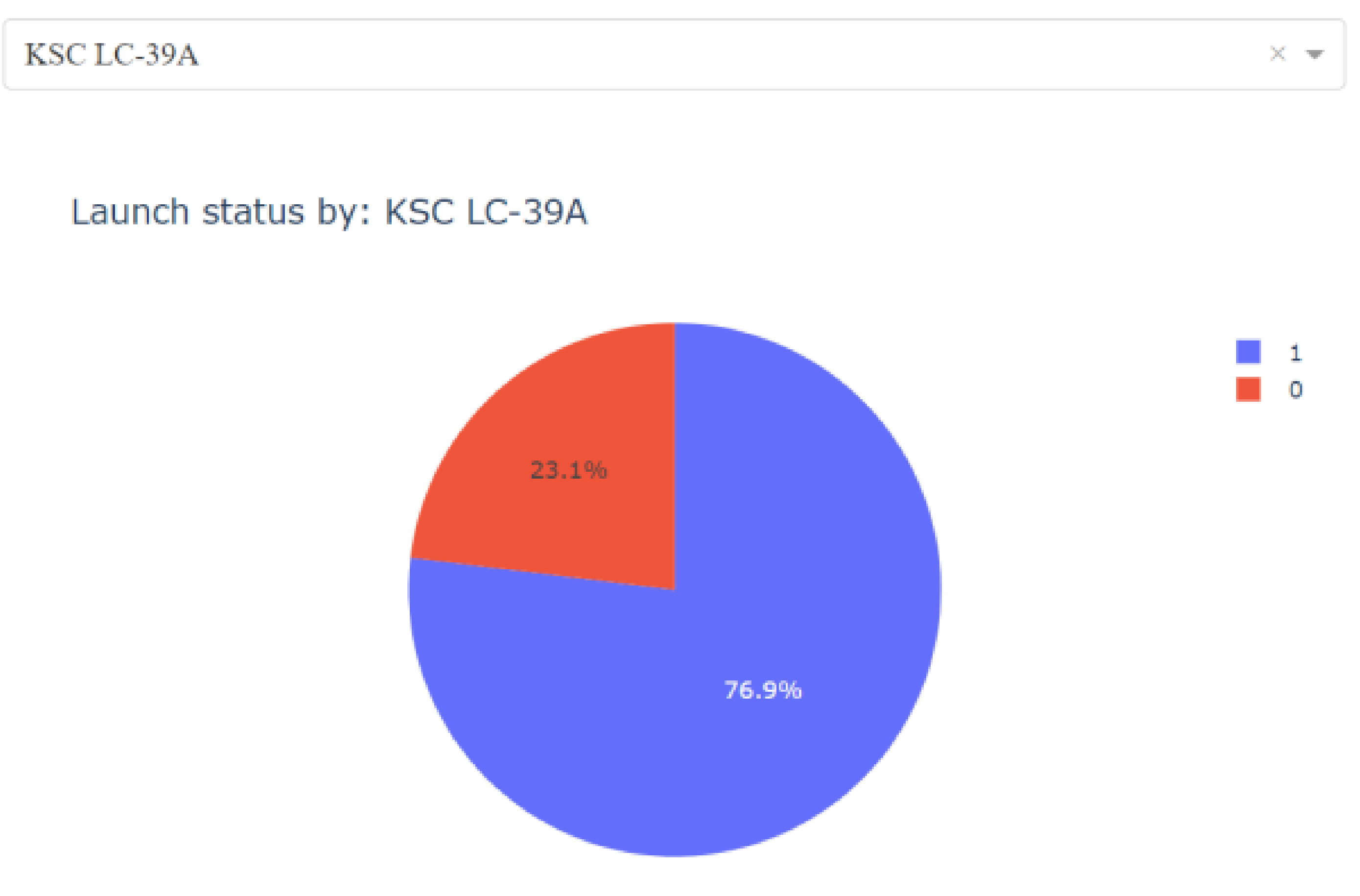
# Total Successful Launches for Site KSC LC-39A

- KSC LC-39A Launch Site has the highest launch success rate and count

- Launch success rate is 76.9%

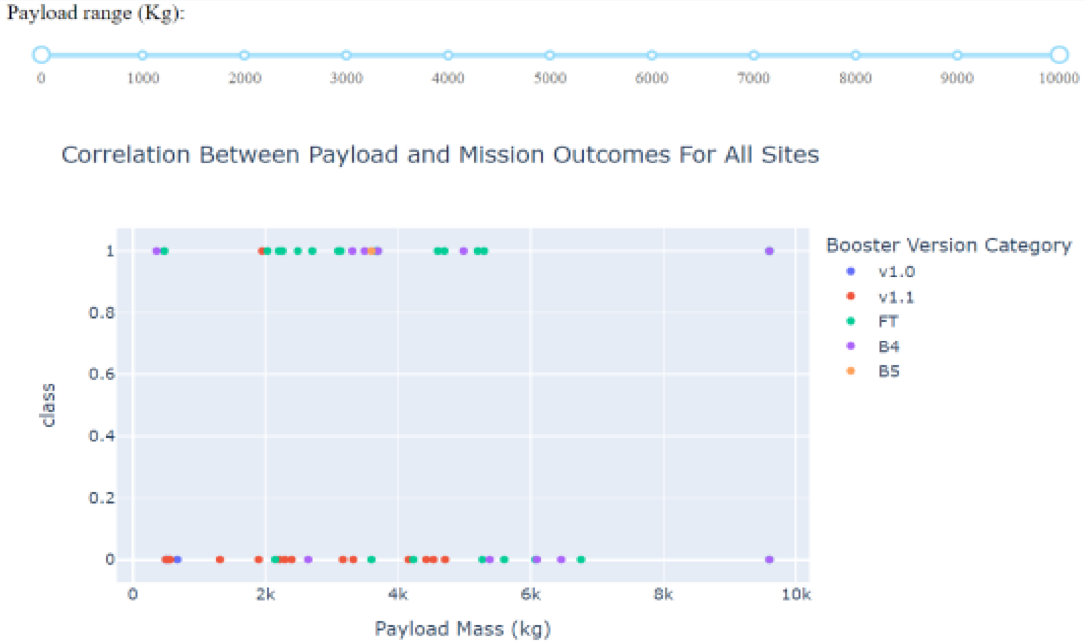
-

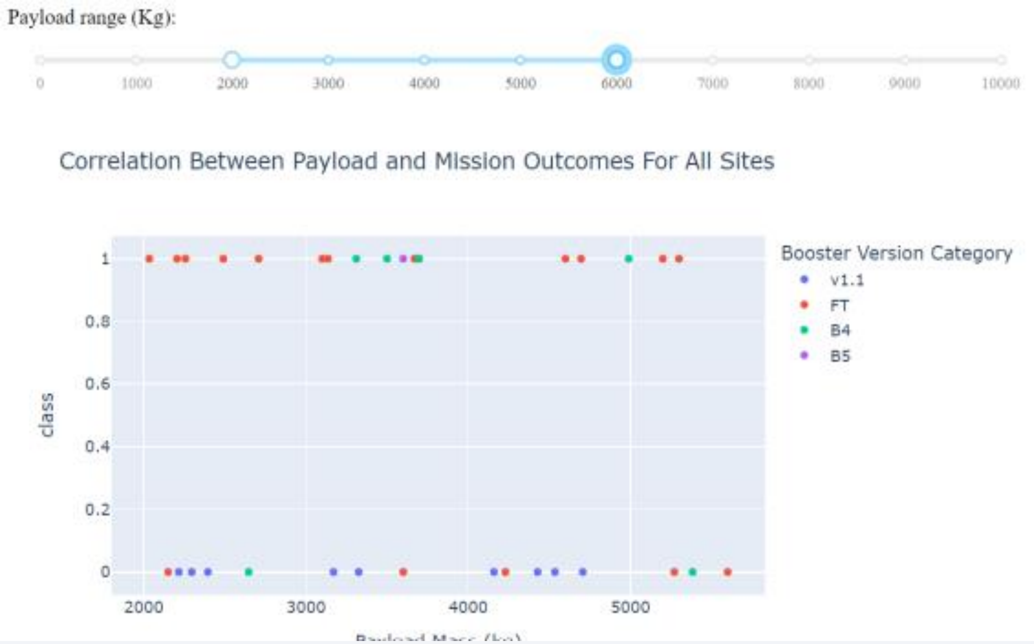
- Launch success failure rate is 23.1%



KSC LC-39A

# Payload vs. Launch Outcome Scatter Plot for All Sites





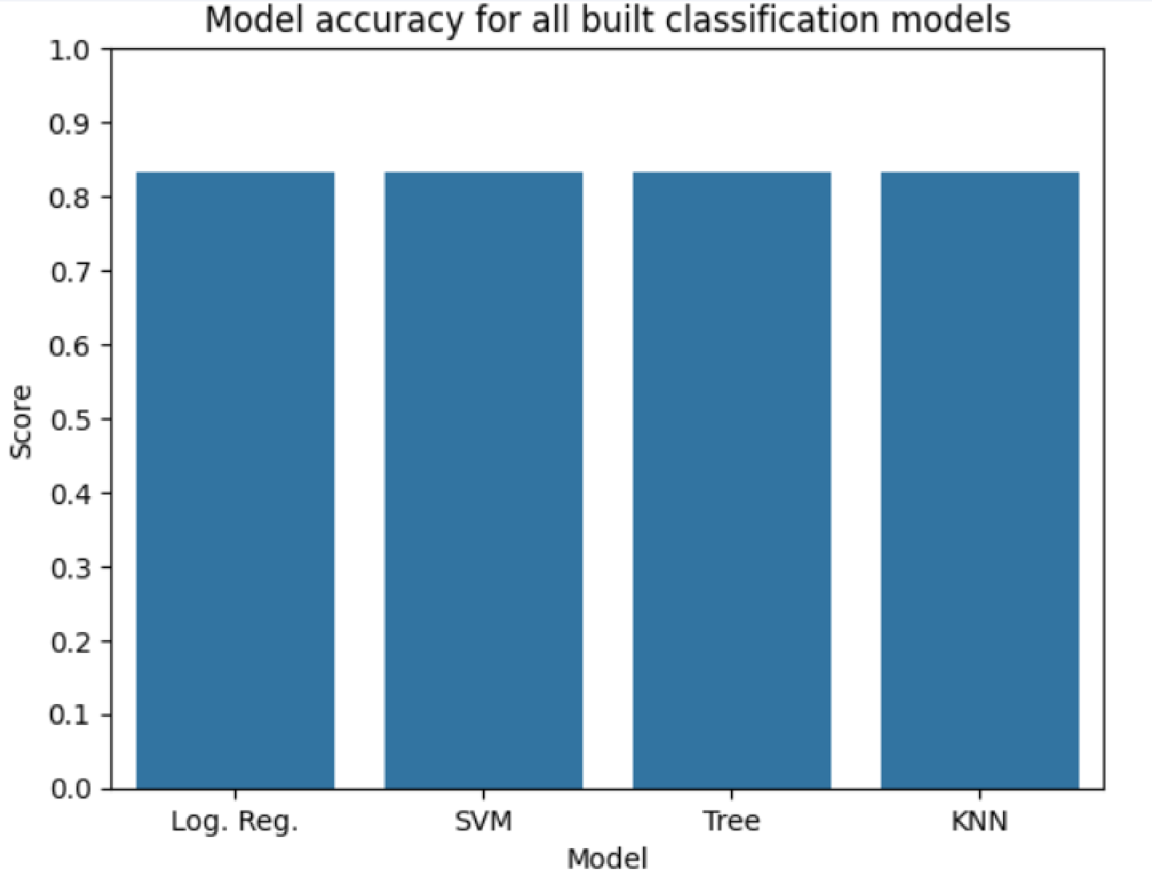
- Most successful launches are in the payload range from 2000 to about 5500
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

Section 5

# Predictive Analysis (Classification)

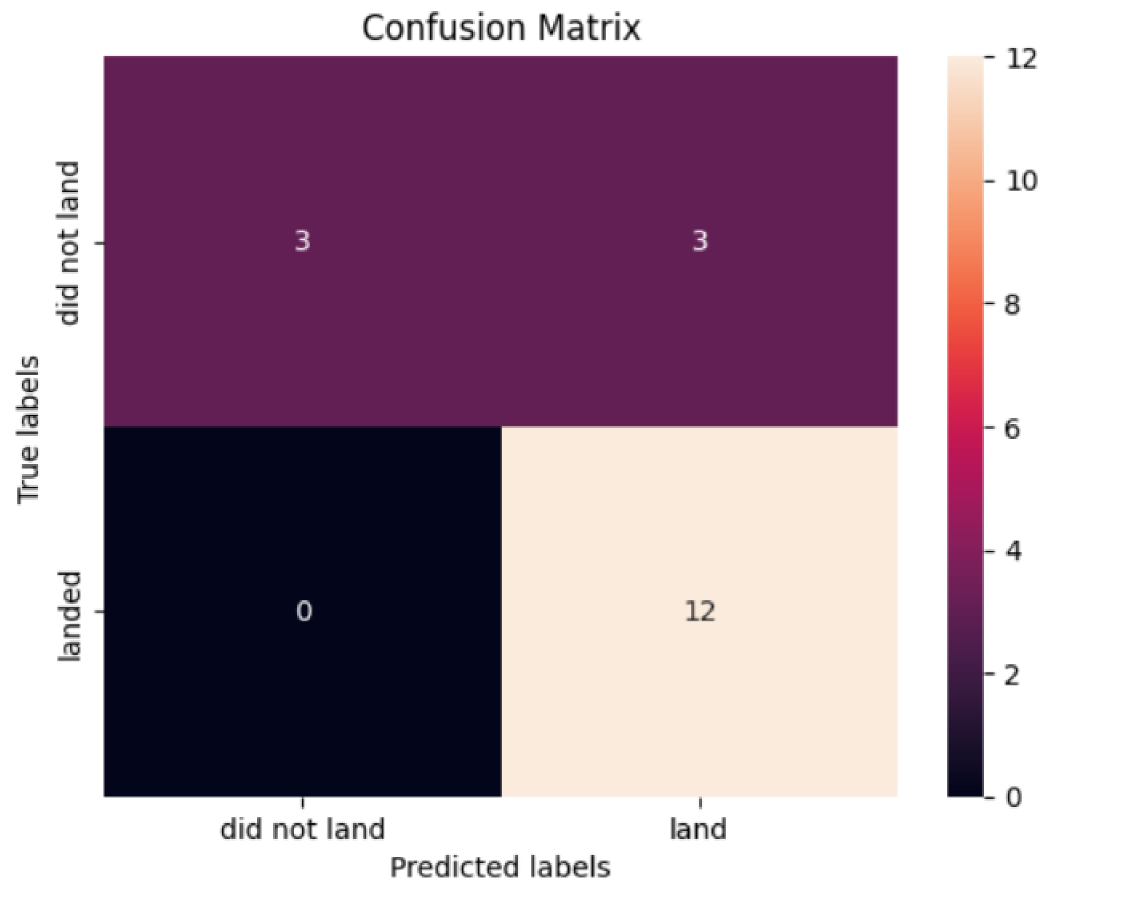
# Classification Accuracy

- Based on the Accuracy scores and as also evident from the bar chart, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333
- Given that the Accuracy scores for Classfication algorithms are very close and the test scores are the same, we may need a broader data set to further tune the models



# Confusion Matrix

- The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%



# Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully
- Success rates appear go up as Payload increases but there is no clear correlation between Payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- The best performing Machine Learning Classfication Model is the Decision Tree with an accuracy of about 87.5%. When the models were scored on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

# Thank you!