



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Shivankar Pandey  
30/09/2023



# Outline

---

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

# Executive Summary

- **Executive Summary: Determining the Cost of SpaceX Falcon 9 Rocket Launches and First Stage Reusability**
- **1. Introduction:**
- The commercial space industry has experienced a significant transformation, with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX making space travel more accessible.
- **2. SpaceX's Success and Cost Advantage:**
- SpaceX stands out as one of the most successful players in the space industry, with achievements such as manned space missions, satellite internet with Starlink, and affordable rocket launches.
- SpaceX's cost advantage lies in the reusability of its Falcon 9 rocket's first stage, which significantly lowers launch costs.
- **3. The Importance of First Stage Reusability:**
- Understanding the first stage's fate —whether it successfully lands or not— is crucial in determining the cost of each SpaceX launch.
- The first stage, being the largest and most expensive component, impacts launch economics.
- **4. Role of Data Science in Competitive Rocket Industry:**
- The capstone project revolves around a new rocket company, Space Y, aiming to compete with SpaceX.
- As a data scientist in this venture, your primary tasks are to determine launch prices and predict the first stage's reusability.
- **5. Data Gathering and Analysis:**
- Information gathering involves comprehensive research on SpaceX's past launches, outcomes, and costs.
- Publicly available data will serve as the foundation for your analysis.
- As the commercial space sector continues to evolve, the insights and techniques developed during this project will position Space Y for long-term success and growth in the industry.

# Executive Summary

- **6. Dashboards for Informed Decision-Making:**
  - Your role includes creating data-driven dashboards to present critical insights to the Space Y team.
  - These dashboards will facilitate strategic decision-making by providing real-time information on SpaceX's launch costs and first stage reusability.
- **7. Machine Learning for Predicting First Stage Reusability:**
  - In contrast to traditional rocket science methods, you will employ machine learning techniques to predict whether SpaceX's first stage will be reusable.
  - Publicly accessible data will be utilized to train and validate the machine learning model.
- **8. Competitive Edge for Space Y:**
  - By accurately estimating SpaceX's launch costs and predicting first stage reusability, Space Y can make informed decisions to compete effectively in the commercial space travel industry.
  - This data-driven approach will enable Space Y to offer competitive pricing and improve overall cost-efficiency.
- **9. Conclusion:**
  - The project's success hinges on your ability to leverage data science and machine learning to gather insights into SpaceX's launch operations.
  - Equipped with this knowledge, Space Y can make informed decisions and establish itself as a competitive player in the rapidly evolving commercial space industry.
- **10. Future Prospects:**As the commercial space sector continues to evolve, the insights and techniques developed during this project will position Space Y for long-term success and growth in the industry.



# Introduction

---

- Project background and context:
  - Space Y, a new entrant in the commercial space industry, aims to compete with SpaceX. SpaceX's cost advantage, driven by Falcon 9's first stage reusability, is a key factor in its success. This project seeks to determine SpaceX's launch costs and predict first stage reusability to inform Space Y's competitive strategy.
- Problems you want to find answers:
  - Estimating the true cost of SpaceX Falcon 9 rocket launches, considering variables like first stage reusability.
  - Predicting whether SpaceX's first stage will be successfully reused or not, using machine learning and publicly available data.
  - Providing Space Y with accurate insights to make informed decisions and compete effectively in the commercial space industry.



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

- **Source:** Data was primarily collected from the SpaceX website dataset, which contains information on rocket launches, costs, and outcomes.
- **Web Scraping:** Automated web scraping tools were utilized to systematically extract data from SpaceX's website.
- **Dataset Creation:** Extracted data was organized into structured datasets, each containing specific launch-related information.
- **Data Fields:** Datasets included fields such as launch date, mission details, payload information, launch success status, and launch costs.
- **Validation:** Data validation was performed by cross-referencing the collected data with external sources and verifying launch outcomes.
- **Cleaning:** Data cleaning procedures were applied to address missing values, inconsistencies, and outliers.
- **Documentation:** A comprehensive record was maintained to document the data collection methodology, sources, and any data modifications made during the process.





# Data Collection – SpaceX API

---

GitHub Link:

<https://github.com/Shivankar28/Final-Project-Report->

Start

|

|--- Access SpaceX Website

|

| |--- Web Scraping

|

| |--- Organize Data into Datasets

|

| |--- Validate Data

|

| |--- Clean Data

|

| |--- Documentation

|

|--- End

# Data Collection - Scraping

- GitHub Link:  
<https://github.com/Shivankar28/Final-Project-Report->

```
Start
|
|--- Access SpaceX Website
| |
| |--- Use Web Scraping Tool
| | |
| | |--- Extract Relevant Data
| |
| |--- Organize Extracted Data
|
|--- End
```

# Data Wrangling



**Data Collection:** Obtain raw data from diverse sources, such as databases, files, or APIs.



**Data Cleaning:** Rectify issues like missing values, duplicates, outliers, and data format inconsistencies.



**Data Transformation:** Enhance data usability by creating new features, standardizing types, and scaling variables.



**Data Integration:** Merge data from different sources, ensuring uniformity and coherence.



**Data Storage:** Safeguard cleaned data for analysis, typically in databases or organized files, while maintaining documentation for traceability.



GitHub Link:

<https://github.com/Shivankar28/Final-Project-Report->



# EDA with Data Visualization



**Purpose:** Explore and understand the dataset's characteristics and patterns.



**Steps:**

- Summarize data statistics.
- Visualize distributions (histograms, box plots).
- Identify correlations (scatter plots, heatmaps).
- Detect outliers and anomalies.
- Explore categorical variables (bar charts, pie charts).



**Benefits:**

- Reveal insights and trends.
- Identify data issues.
- Guide feature selection.
- Support decision-making.
- Communicate findings effectively.

GitHub Link:

<https://github.com/Shivankar28/Final-Project-Report->

# EDA with SQL

**1.Purpose:** Explore and analyze dataset using SQL queries.

**2.Steps:**

- 1.Select columns for inspection.
- 2.Calculate summary statistics (COUNT, AVG, SUM).
- 3.Group data (GROUP BY) for aggregations.
- 4.Filter and sort data (WHERE, ORDER BY).
- 5.Join tables for context (INNER JOIN, LEFT JOIN).


**3.Benefits:**

- 1.Efficient data exploration.
- 2.Complex queries for deeper insights.
- 3.Data segmentation and aggregation.
- 4.Seamless integration with databases.
- 5.Scalable for large datasets.

# Build an Interactive Map with Folium

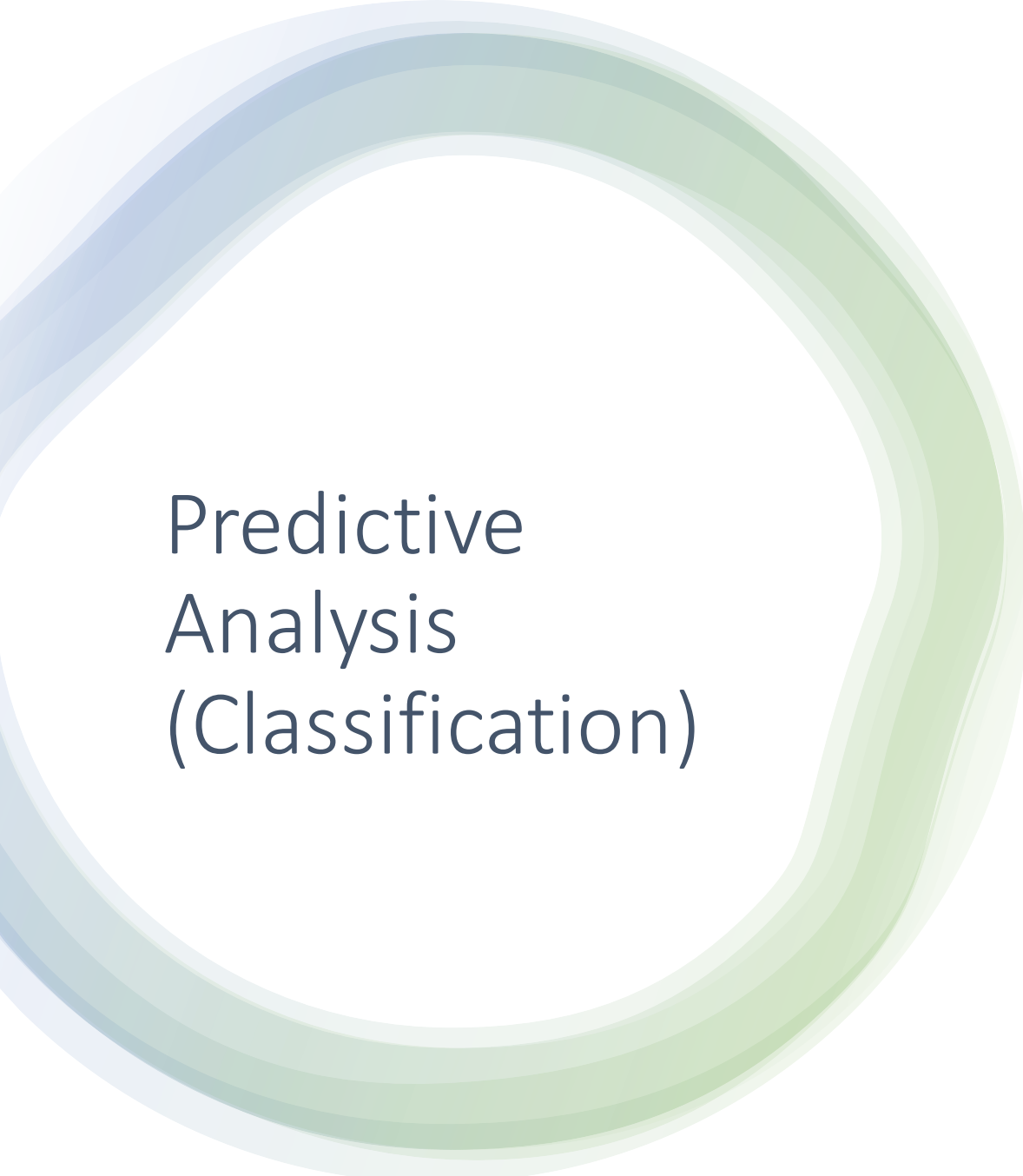
- **Objects added to Folium Map:**
- **Markers:** Used for pinpointing specific locations of interest, providing visual cues.
- **Circles:** Represent areas of influence or coverage, showcasing reach.
- **Lines:** Depict routes, boundaries, or areas of interest for visual clarity.
- **Pop-up Information:** Added to objects for extra details, enhancing interactivity.
- **Reasons for Adding Objects:**
- **Enhanced Visualization:** Objects offer a richer visual representation of data.
- **Geospatial Context:** Provides geographical context to data.
- **Interactivity:** Enables users to access additional details.
- **Communication:** Effectively conveys spatial information and data.
  - GitHub Link:  
<https://github.com/Shivankar28/Final-Project-Report->





# Build a Dashboard with Plotly Dash

- **Dashboard with Plotly Dash:**
- **Plots and Interactions:**
- **Line Charts:** Display time trends.
- **Bar Charts:** Compare categories.
- **Pie Charts:** Show part-to-whole relationships.
- **Scatter Plots:** Visualize correlations.
- **Heatmaps:** Identify patterns.
- **Why Added:**
- **Data Insight:** Visualize complex data.
- **User Control:** Customize views with interactions.
- **Actionability:** Enable specific actions.
- **Context:** Provide information on hover.

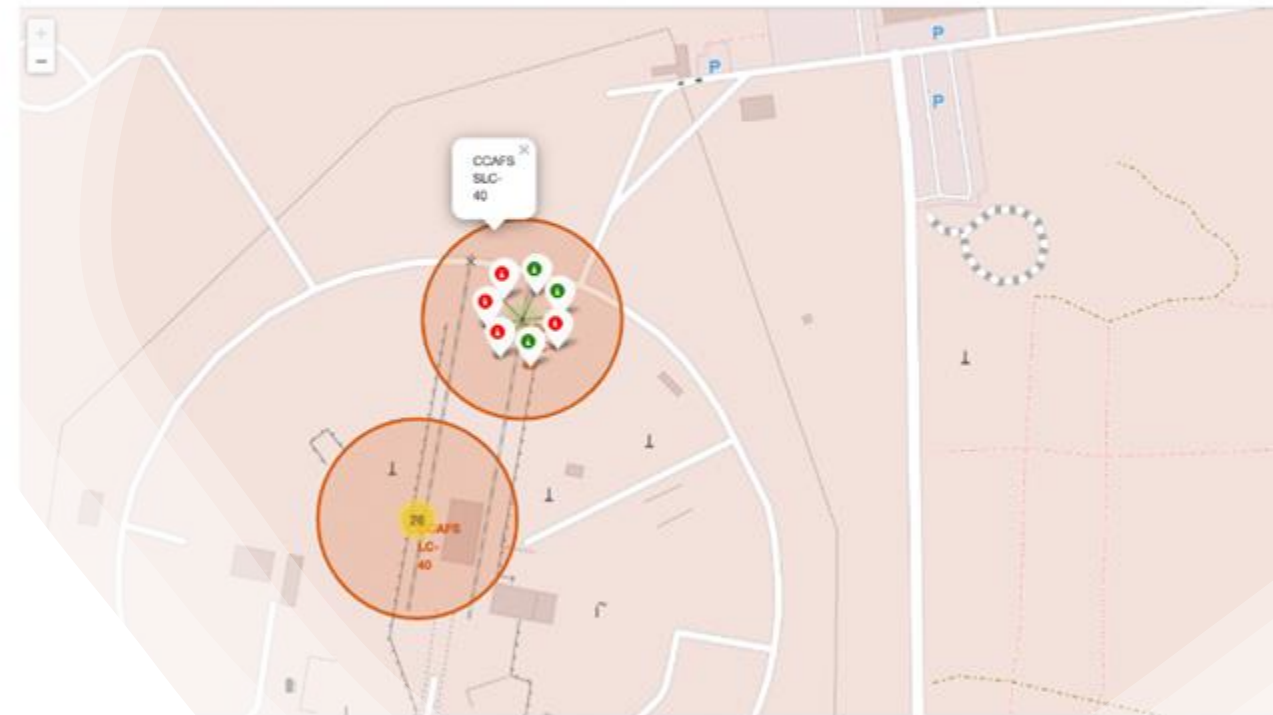
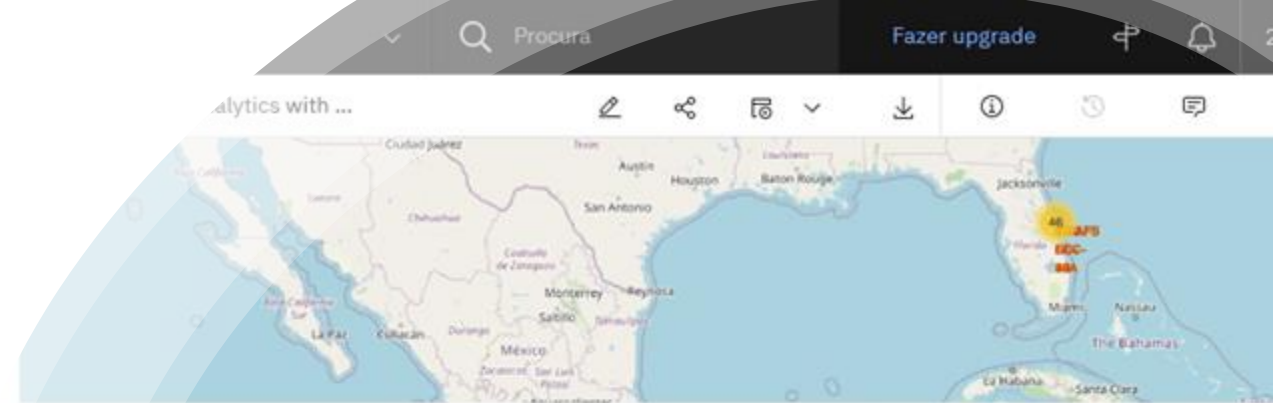


# Predictive Analysis (Classification)

- **Model Development Process:**
  - **Data Prep:** Collect, clean, split data.
  - **Feature Eng.:** Select/reengineer features.
  - **Model Selection:** Choose algorithms.
  - **Model Evaluation & Improvement:**
    - **Initial Eval:** Train, test, evaluate.
    - **Comparison:** Compare model performance.
    - **Tuning & Iteration:** Optimize parameters/features.
  - **Best Model Selection:**
  - **Final Model:** Choose best performer.
  - **Validation:** Validate on separate data.

# Results

- **Exploratory Data Analysis (EDA) Results:**
- Summary stats, data distributions, correlations.
- Visualizations, trends, and insights.
- **Predictive Analysis Results:**
- Model performance metrics, confusion matrix.
- Feature importance, model visualizations.
- Validation and model interpretability.



markers in marker clusters, you should be able to easily identify which launch sites have relatively high

the distances between a launch site



The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

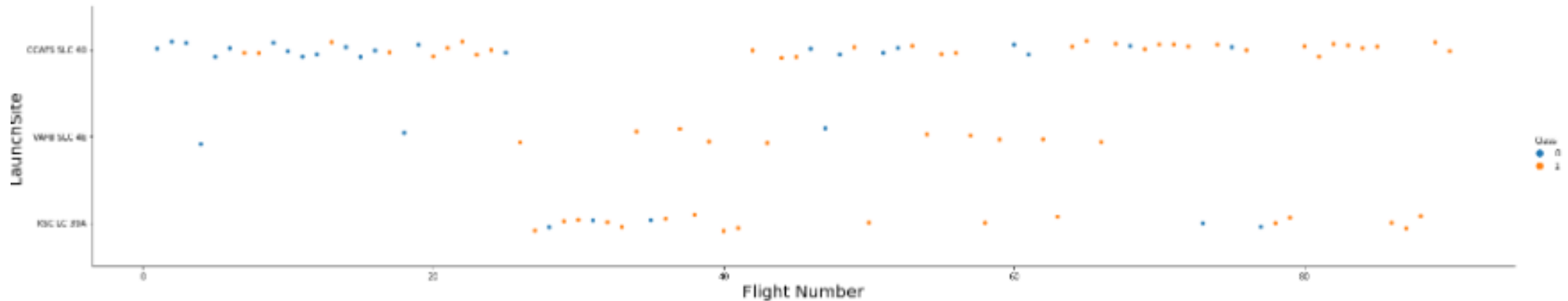
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

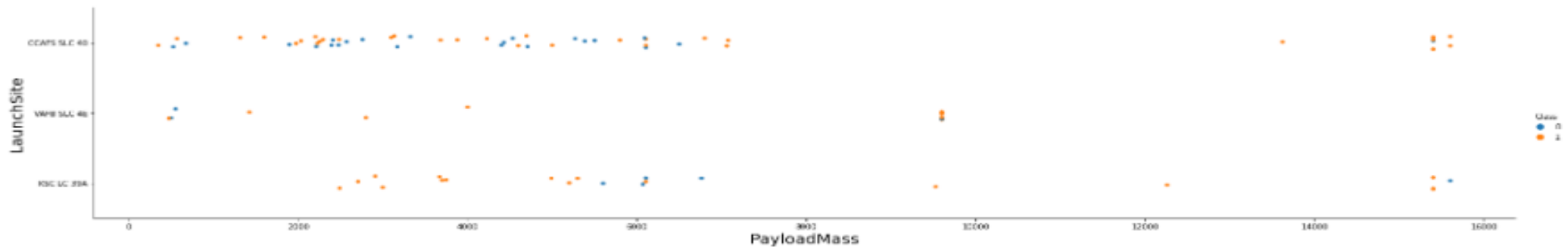
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



# Payload vs. Launch Site

---

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the c
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



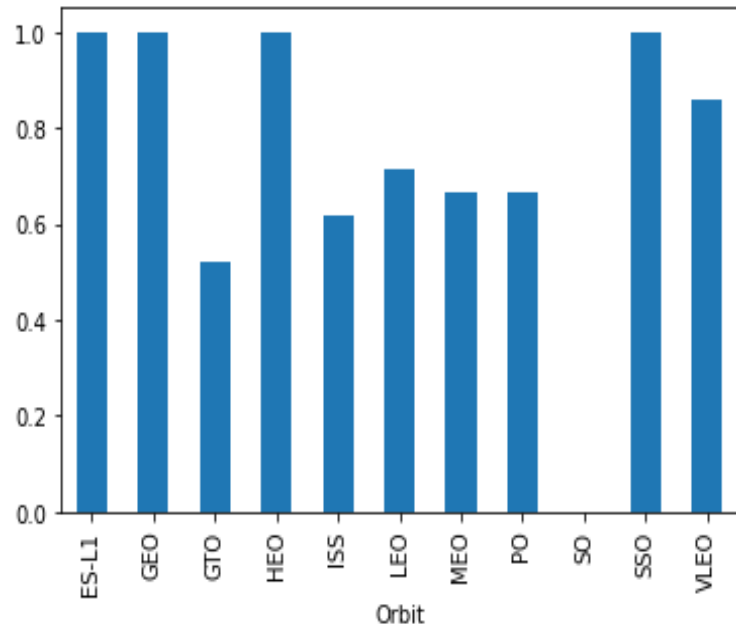


# Success Rate vs. Orbit Type

---

```
# HINT use groupby method on Orbit column and get the mean of Class column  
df.groupby('Orbit')['Class'].mean().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2169a52ad0>
```



# Flight Number vs. Orbit Type

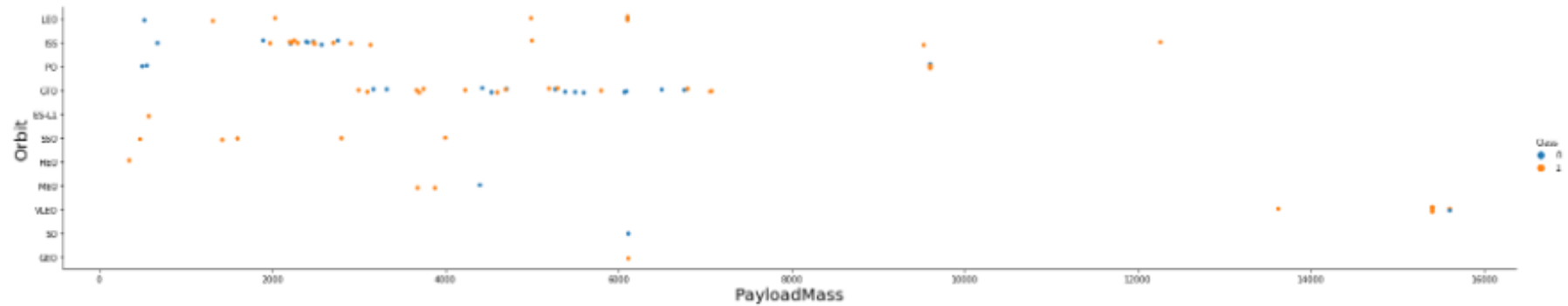
[10]:

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

# Payload vs. Orbit Type

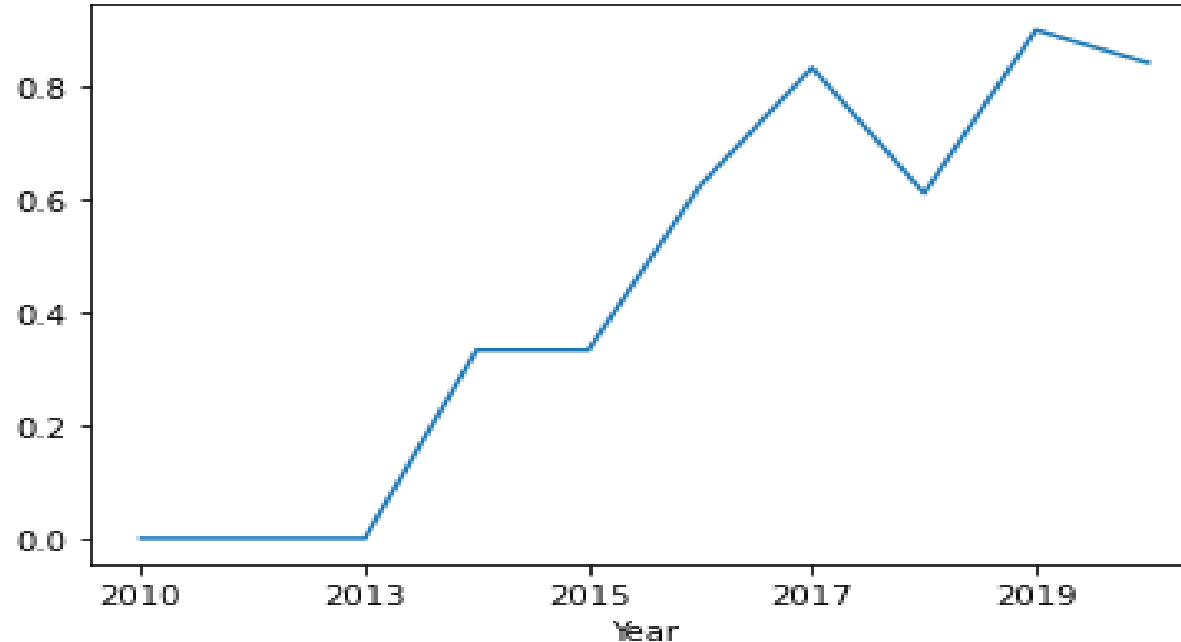
In [11]:

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



# Launch Success Yearly Trend

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7f2168e1c410>
```



you can observe that the success rate since 2013 kept increasing till 2020



# All Launch Site Names

---

Display the names of the unique launch sites in the space mission

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

```
* ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.clogj3sd0t  
ludb  
Done.
```

**launch\_site**

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

# Launch Site Names Begin with 'CCA'

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

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

```
* ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/b1udb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-02-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

# Total Payload Mass

---

## Task 5

List the date when the first successful landing outcome in ground pad was achieved.

*Hint: Use min function*

```
] sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';  
  
* ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud  
ludb  
Done.  
  
]: first_success_gp  
2015-12-22
```

# First Successful Ground Landing Date

---

## Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
1]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOM
* ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/b
ludb
Done.
```

```
1]: booster_version
```

```
F9 FT B1021.2
```

```
F9 FT B1031.2
```

```
F9 FT B1022
```

```
F9 FT B1026
```

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

---

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

```
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP
```

```
* ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/  
ludb  
Done.
```

landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

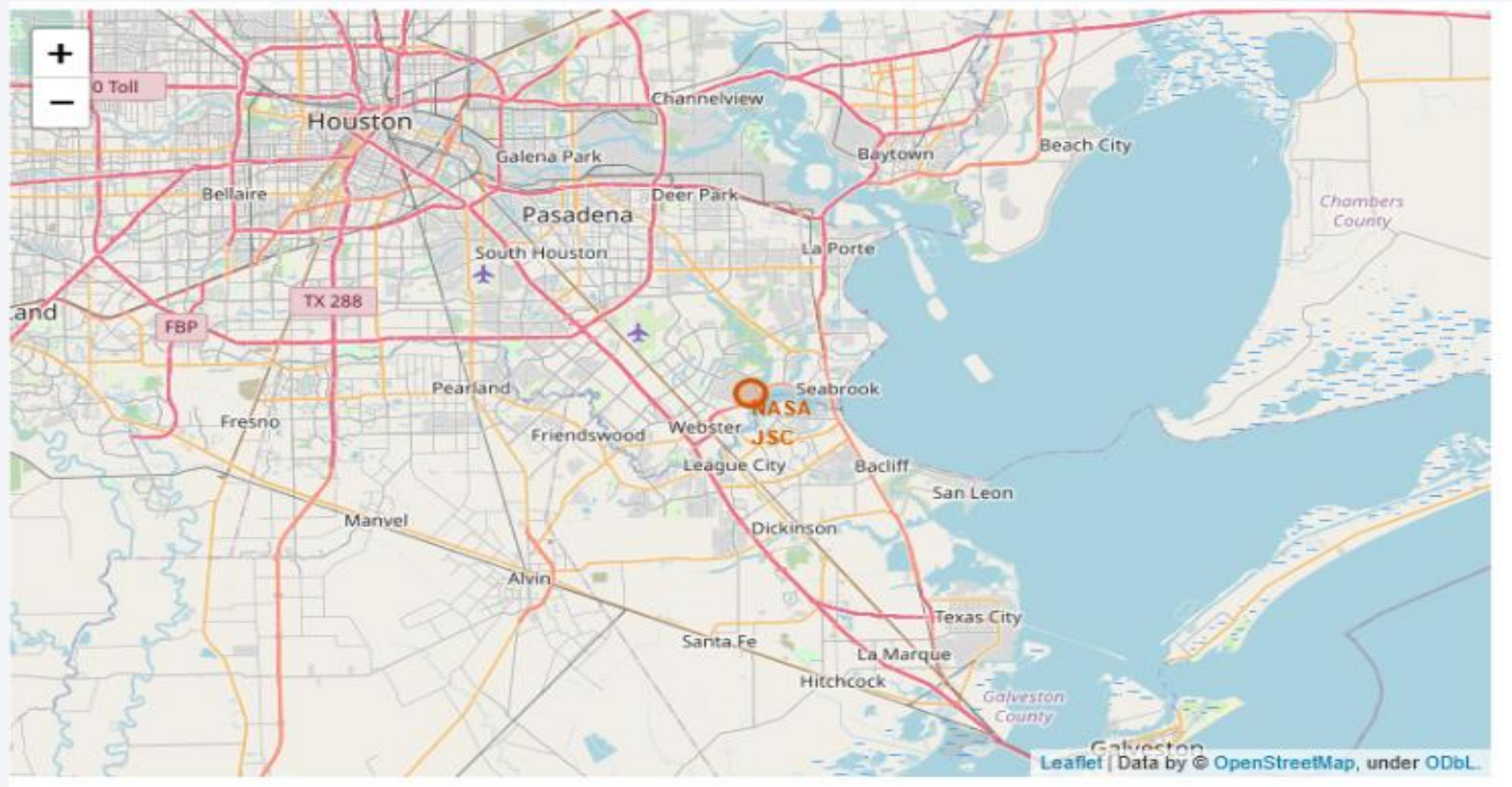


A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark blue, with numerous bright yellow and orange lights representing cities and urban areas. The horizon line of the Earth is visible, separating the dark surface from the blackness of space.

Section 3

# Launch Sites Proximities Analysis

# Folium Map 1



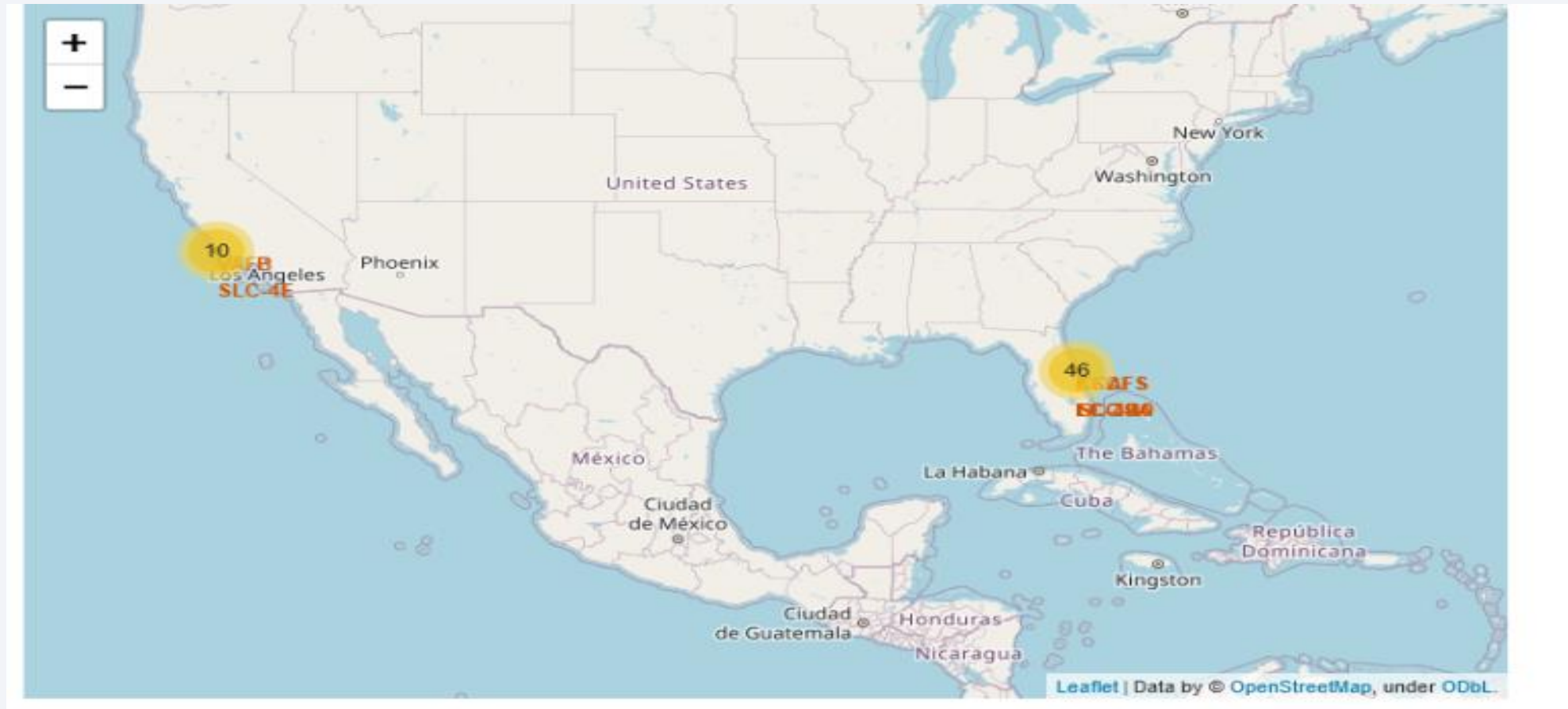
# Folium Map 2

---





# Folium Map 3





Section 4

# Build a Dashboard with Plotly Dash



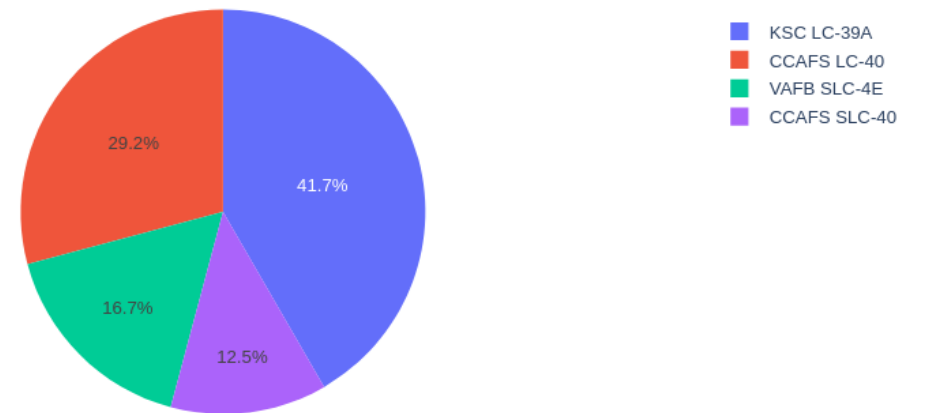
# Dashboard 1

## SpaceX Launch Records Dashboard

All Sites

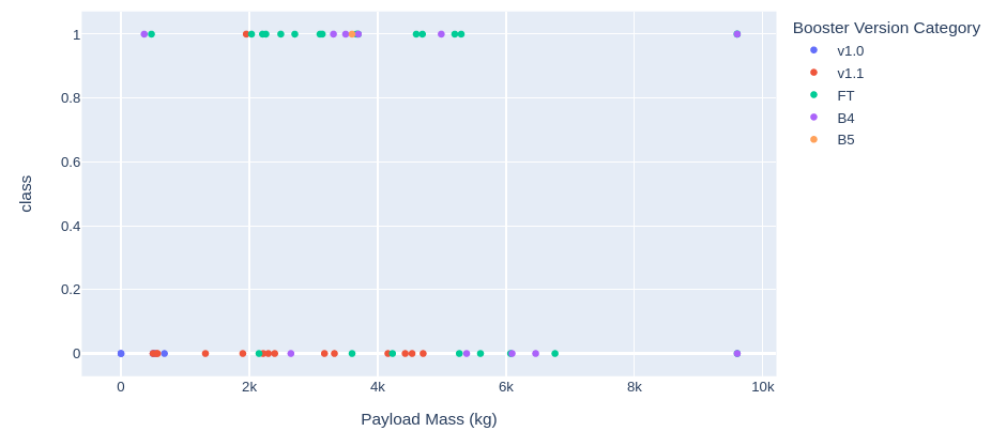


Total Success Launches By Site



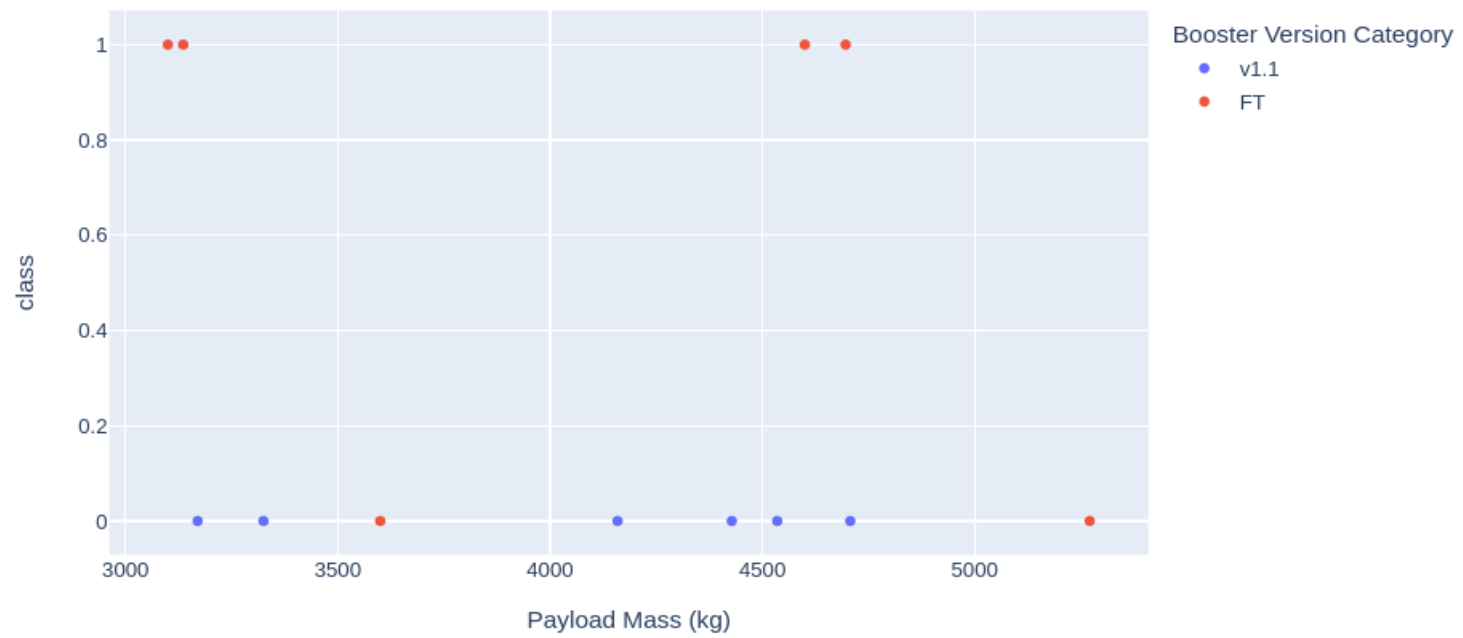
# Dashboard 2

Payload range (Kg):



# Dashboard 3

Payload range (Kg):





Section 5

# Predictive Analysis (Classification)

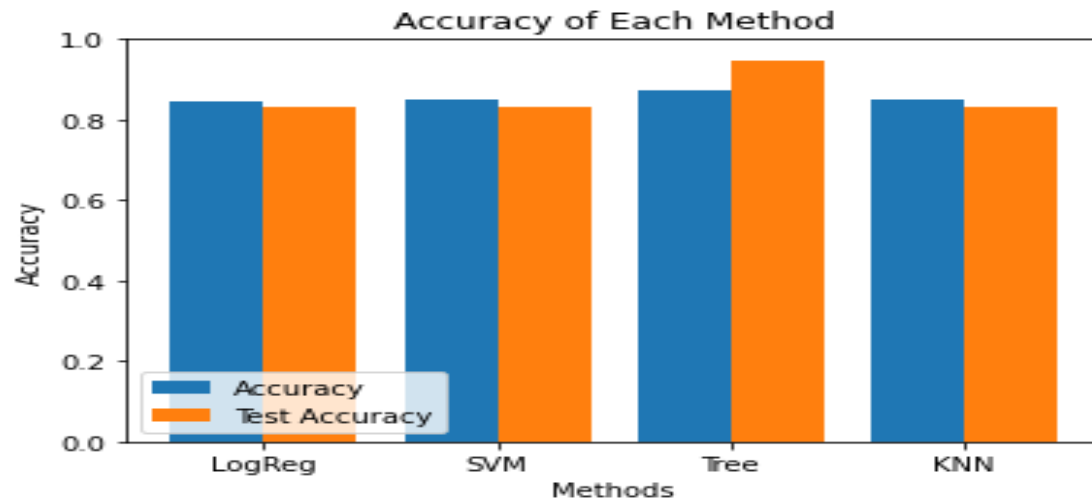


# Classification Accuracy

---

```
plt.ylim([0,1])
plt.xticks(x_axis, x)

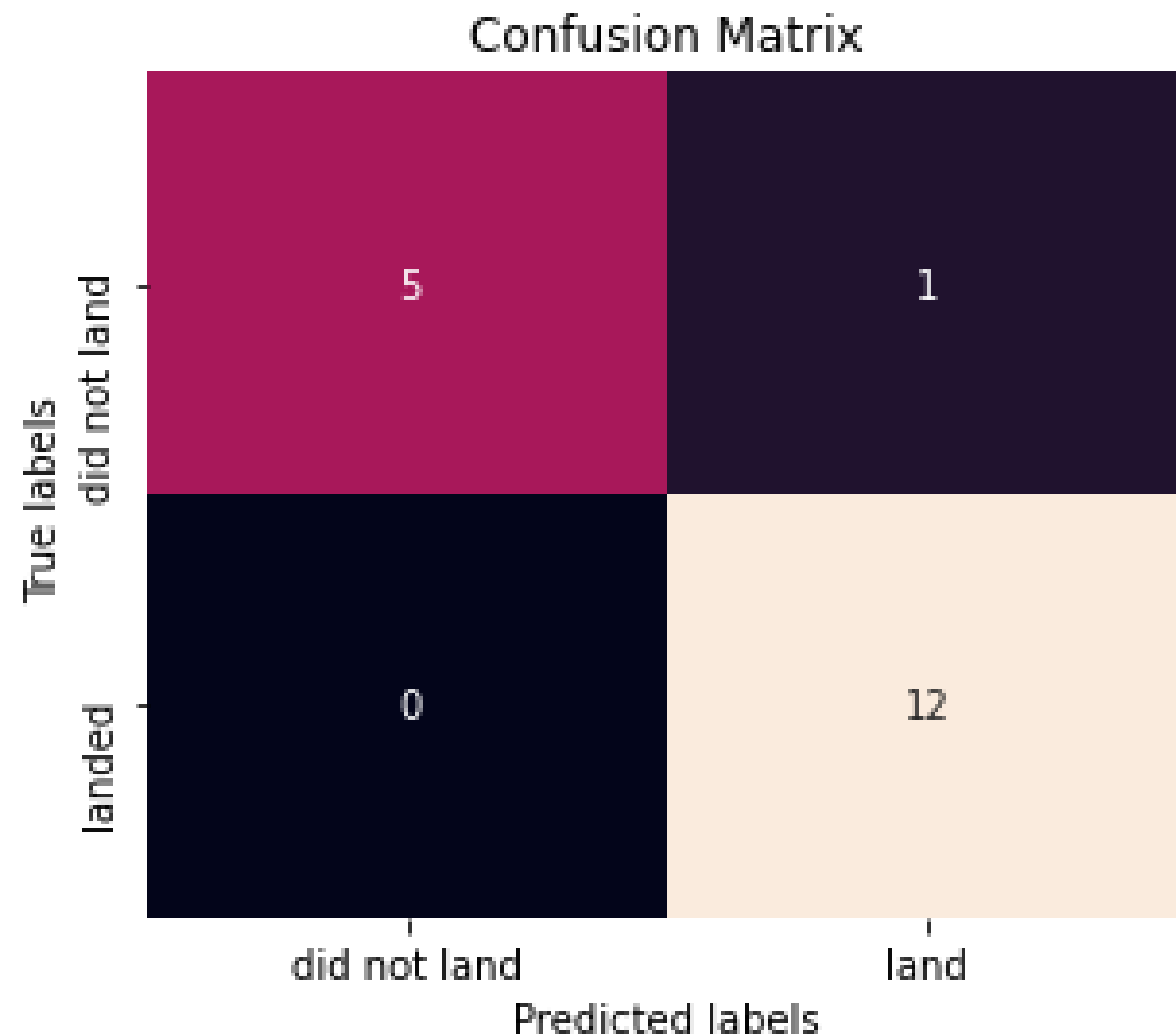
plt.xlabel("Methods")
plt.ylabel("Accuracy")
plt.title("Accuracy of Each Method")
plt.legend(loc='lower left')
plt.show()
```



# Confusion Matrix

---

```
yhat = tree_cv.predict(X_test)  
plot_confusion_matrix(Y_test,yhat)
```



# Conclusions

- **Promising Space Age:** The project underscores the era of commercial space travel, with companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX revolutionizing accessibility to space exploration.
- **SpaceX's Success:** Among these, SpaceX stands out as a trailblazer, achieving remarkable milestones such as International Space Station missions, the Starlink satellite constellation, and manned spaceflights.
- **Cost Efficiency:** SpaceX's key advantage lies in cost-efficiency, prominently exemplified by the reuse of Falcon 9 rocket first stages, significantly reducing launch expenses.
- **Data-Driven Approach:** The project showcases a data-driven strategy, employing machine learning to predict Falcon 9 first stage reusability. This innovative approach leverages public data to determine cost-effectiveness.
- **Competitive Aspirations:** Space Y's aspiration to compete with SpaceX, spearheaded by industrialist Elon Musk, exemplifies the enduring entrepreneurial spirit in the commercial space industry, underpinned by data science and analysis.
- In summary, the project illuminates the exciting prospects of affordable space travel, SpaceX's achievements, cost-efficient practices, data-driven decision-making, and the continued drive for competitiveness in the evolving space industry.

# Appendix

---

- <https://github.com/Shivankar28/Final-Project-Report->

Thank you!

