**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

Space Missions

**Submitted by**

**NAME:** K.Sai Kalyan Reddy

**REGISTRATION NO:** 12304905

**PROGRAMME AND SECTION:** K23GN57

**COURSE CODE:** INT375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **K. Sai Kalyan Reddy** bearing Registration no. **12304905** has completed **INT375** project titled, **“Mrs.Aashima”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, K. Sai Kalyan Reddy student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12304905 Name of the student: K. Sai Kalyan Reddy

Introduction

Exploratory Data Analysis (EDA) plays a critical role in the data science process. It helps transform raw data into meaningful insights by identifying patterns, trends, relationships, and potential anomalies. Through EDA, we gain a deeper understanding of the dataset, which is essential before applying any advanced modeling or predictive techniques.

In this project, we perform EDA on a dataset related to space missions, using Python and its robust data analysis and visualization libraries, such as Pandas, Matplotlib, Seaborn, and Plotly. The dataset includes information about launch dates, locations, mission outcomes, organizations, vehicle types, and associated costs.

Our primary objective is to explore and visualize the data to uncover insights into the global space industry. Specifically, we aim to analyze:

Trends in the number and frequency of space launches over time

The geographic distribution of launch sites and countries involved

The financial aspects of space missions, including cost trends

The performance and success rates of different space organizations and launch vehicles

By the end of this analysis, we expect to have a clearer picture of how the space industry has evolved over time and which players are leading the charge in space exploration.

**Source of Dataset**

The dataset used in this project is titled **"space\_missions.csv"**. It was sourced from a publicly available dataset repository that aggregates historical information on space missions conducted by various organizations around the world. The dataset provides comprehensive details, including the launch date, company name, location, mission details, rocket type, and mission status.

This dataset is particularly valuable for exploratory analysis as it captures both temporal and categorical variables, allowing for multifaceted insights into the evolution of space exploration.

* **File Name:** space\_missions.csv
* **Encoding:** Latin-1

Before analysis, the dataset was inspected and preprocessed to handle encoding issues and ensure clean, structured data for visualization and interpretation.

**EDA Process**

The Exploratory Data Analysis (EDA) process in this project involved several key steps to better understand the dataset and extract meaningful insights. Each step helped transform raw data into visual and statistical narratives that shed light on trends and patterns in space missions.

The steps followed include:

* **Importing the dataset:** Loading the space\_missions.csv file into a DataFrame using appropriate encoding (Latin-1).
* **Data cleaning and preprocessing:** Handling missing values, correcting inconsistent entries, and formatting date fields for uniformity.
* **Feature extraction and engineering:** Creating new columns such as Year, Month, and Country to enable time-based and geographical analyses.
* **Data visualization:** Plotting distributions, trends, and comparisons using various types of graphs to uncover insights.
* **Statistical analysis:** Identifying outliers, correlations, and summary statistics to support deeper data understanding.

The following Python libraries were used throughout the analysis:

* **Pandas** and **NumPy** for data manipulation and numerical operations
* **Matplotlib** and **Seaborn** for creating static visualizations and uncovering visual patterns in the data

These tools enabled efficient data exploration and helped bring clarity to the key insights derived from the dataset.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a crucial step in the data science process, used to explore, understand, and summarize the main characteristics of a dataset before applying formal modeling techniques. It helps uncover hidden patterns, detect anomalies, and test underlying assumptions, all of which are vital for making informed, data-driven decisions.

**Why EDA Is Important:**

EDA enhances the overall quality of analysis by:

* Providing a clear understanding of the dataset’s structure and content
* Highlighting patterns, trends, and potential anomalies
* Revealing relationships between variables
* Ensuring data is clean, complete, and ready for modeling

By identifying issues early, EDA also improves the accuracy and performance of predictive models.

**How EDA Was Performed in This Project:**

In this space missions analysis, we applied the following EDA techniques:

* **Data Cleaning:** Handled missing values, removed duplicates, and standardized date and string formats
* **Feature Engineering:** Created new features such as Year (from launch dates) and Country (from launch locations)
* **Visualization Tools:** Used line plots, bar charts, and heatmaps to visualize trends, distributions, and correlations
* **Statistical Techniques:** Applied correlation matrices and Z-scores to detect relationships and identify outliers

Together, these methods allowed us to gain meaningful insights into global space missions—highlighting trends over time, company performance, geographical distribution, and more.

Analysis on Dataset

Analysis 1: Launch Trends Over Time

i. Introduction:

This analysis explores how the number of space launches has changed over the years.

ii. General Description:

We count the number of launches per year to observe long-term trends.

iii. Functions and Formulas:

df['Year'].value\_counts().sort\_index()

matplotlib.pyplot.plot()

iv. Analysis Results: Certain years saw peaks in launch activity, reflecting global interest or advancements in space technology.

Analysis 2: Company-wise Mission Success and Failures

i. Introduction:

Evaluates which companies have had the most success or failure in launching missions.

ii. General Description:

The 'Company' and 'MissionStatus' columns are used to create grouped bar charts.

iii. Functions and Formulas:

sns.countplot(x='Company', hue='MissionStatus')

iv. Analysis Results:

Top-performing companies consistently have more successful missions. Some have notable failure rates.

**Analysis 3: Mission Type Distribution**

**i. Introduction:**  
This analysis focuses on categorizing space missions based on their purpose, specifically distinguishing between **communication-related missions** and **others**. Understanding the types of missions helps shed light on the primary objectives driving space exploration efforts.

**ii. General Description:**  
To perform this analysis, we created a new column called Mission\_Type by analyzing the mission descriptions provided in the dataset. Missions that included keywords like "communications", "telecom", or similar were labeled as *Communication*, while the rest were grouped under *Other*.

**iii. Functions and Formulas Used:**  
The classification and visualization were performed using the following approach:

df['Mission\_Type'] = df['Mission'].apply(...)

sns.countplot(x='Mission\_Type')

**iv. Analysis Results:**  
The analysis revealed that a significant portion of space missions are geared toward **communication**, reflecting the global demand for satellite-based telecommunications. The remaining missions fall under other categories such as earth observation, scientific research, navigation, and defense-related activities.

**Analysis 4: Cost Analysis Over Years**

**i. Introduction:**  
This analysis examines how the **cost of space missions** has varied over the years. Tracking cost trends can offer insight into technological advancements, economic factors, and evolving mission complexities in the space industry.

**ii. General Description:**  
We calculate the **average mission cost per year** using the Price column from the dataset. By grouping the data by year and plotting the yearly averages, we can visualize long-term financial patterns in space exploration.

**iii. Functions and Formulas Used:**  
The following code was used to compute and plot average annual costs:

df.groupby('Year')['Price'].mean()

plt.plot()

**iv. Analysis Results:**  
The analysis shows that mission costs have **fluctuated over time**, reflecting changes in **mission objectives, inflation, and technology**. Some years show spikes—likely due to high-profile or complex missions—while others indicate cost optimization through innovation or commercial efficiencies.

**Analysis 5: Launch Distribution by Country**

**i. Introduction:**  
This analysis investigates the **geographic distribution of space launches** by examining the number of missions launched from various countries. Identifying leading countries in space exploration highlights global players and helps understand the global distribution of space infrastructure.

**ii. General Description:**  
To analyze this, we derived a new column, Country, from the Location field. Using this column, we calculated the number of launches by country and visualized the top countries with the highest number of space missions.

**iii. Functions and Formulas Used:**  
The following steps were used for classification and visualization:

**iv. Analysis Results:**  
The results show that the top countries in space exploration include **the USA, Russia, and China**, which dominate global launch activities. These countries have well-established space programs and infrastructure, leading to a high frequency of space missions compared to other nations.

**Analysis 6: Relationships and Correlations**

**i. Introduction:**  
This analysis explores potential correlations between numerical variables in the dataset to identify any notable relationships or patterns.

**ii. General Description:**  
Pairwise relationships between variables were visualized using scatterplots and correlation matrices. The goal was to assess the strength and direction of linear associations using correlation coefficients.

**iii. Functions and Formulas Used:**

* sns.pairplot(df) – Generates scatterplots and histograms to visualize distributions and relationships between variables.
* sns.heatmap(df.corr()) – Displays a heatmap of Pearson correlation coefficients, highlighting the degree of linear correlation between each pair of numeric variables.

**iv. Analysis Results:**

* A **minor correlation** was detected between **cost** and **year**, indicating a slight upward or downward trend over time.
* No other **strong linear relationships** were found among the variables, suggesting relative independence between most numeric features.
* The lack of high correlation values implies that multicollinearity is not a significant concern for further modeling or statistical analysis.

**Analysis 7: Price Outlier Detection**

**i. Introduction:**  
This analysis focuses on identifying outlier values in the **Price** variable to detect missions that are significantly more expensive than typical cases.

**ii. General Description:**  
Outliers can reveal unique, high-cost missions—potentially due to advanced technology, mission complexity, or special objectives. Identifying these helps in understanding cost distribution and flagging anomalies.

**iii. Analysis Results:**

* A few missions stood out as **significantly more expensive** than the rest.
* These high-cost missions may reflect **specialized operations or high-budget launches**.
* The presence of outliers suggests that while most missions fall within a normal cost range, some deviations exist and may warrant further investigation.

**Extra: Top Launching Companies**

**i. Introduction:**  
This section ranks space launch companies based on the total number of missions they conducted, offering insight into industry leaders.

**ii. General Description:**  
The top 10 companies were identified by counting the number of launches attributed to each. The results were visualized using a bar chart to clearly show the comparative frequency of launches.

**iii. Analysis Results:**

* The **space launch industry is dominated by a few major companies**, with the top players responsible for a significant portion of all missions.
* These leaders include both government and private entities, reflecting their capabilities, infrastructure, and consistent presence in the market.

**5. Conclusion**

This analysis provided a comprehensive overview of global space missions, examining various aspects such as launch costs, temporal trends, geographic distribution, and company-level activity. Through data visualizations and statistical techniques, several key insights emerged:

* **Costs and outliers** revealed the presence of high-budget missions, likely linked to complex or specialized operations.
* **Temporal trends** indicated an upward trajectory in space activity, reflecting technological progress and increased global interest.
* **Geographical analysis** showed that while traditional spacefaring nations continue to lead, new players are entering the domain.
* **Company-level rankings** highlighted a handful of dominant organizations, though there's a clear trend toward **greater participation from private companies** and emerging space agencies.

Overall, the findings underscore the **dynamic and evolving nature of space exploration**, marked by growing international involvement, technological advancements, and increased accessibility. These trends suggest a promising future for the industry, with further innovation and collaboration on the horizon.

**6. Future Scope**

The current analysis lays a strong foundation for understanding global space missions, but there are several opportunities to expand and deepen the insights:

* **Integration of Additional Variables:**  
  Incorporating more features such as **payload type**, **mission objectives**, and **launch outcomes** could provide a richer context for analysis and uncover deeper patterns.
* **Predictive Modeling:**  
  Developing **machine learning models** to estimate mission costs based on inputs like year, company, payload, and location can support budgeting and feasibility assessments.
* **Interactive Dashboards:**  
  Creating **interactive dashboards** using tools like **Plotly Dash, Tableau, or Power BI** would enable users to dynamically explore data, filter by criteria, and visualize trends in real time.
* **Geospatial Analysis:**  
  Utilizing **launch site coordinates** to perform spatial analysis can reveal geographic patterns, clustering of launch activity, and proximity to key space infrastructure.

These enhancements would not only improve analytical depth but also support decision-making for researchers, industry stakeholders, and policymakers in the space sector.

**7. References**

* **Kaggle Public Datasets:**  
  Source of the primary dataset used for analyzing global space missions.  
  *(e.g., “Space Missions” dataset on Kaggle)*
* **Python Libraries:**
  + **Pandas** – for data manipulation and analysis
  + **NumPy** – for numerical operations
  + **Matplotlib** and **Seaborn** – for data visualization
* **Scientific Literature and Official Sources:**
  + Academic articles and reports on global space activity
  + Official websites and databases of space agencies (e.g., **NASA**, **ESA**, **ISRO**, **SpaceX**) used to validate or supplement findings

These resources were essential for conducting thorough and credible analysis throughout the project.

