AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH
(AIUB)

FACULTY OF SCIENCE & TECHNOLOGY

DEPARTMENT OF ENGINEERING

INTRODUCTION TO DATA SCIENCE

Spring 2024-2025

Section: G

Report name:

Project Report on Data Analysis

Supervised By:

DR. ASHRAF UDDIN

Submitted By:

NAME	ID
1. ARSHAD ABEDIN ABIR	22-47188-1
2. KAMRUZZAMAN SONY	22-46797-1
3. MRIDUL KANTI KUNDU	22-47182-1
4. MAYSHA MARIUM	22-47197-1

Date of Submission: April 27, 2025

1. Introduction

This project aims to analyze and preprocess a dataset containing information related to movie ticket purchases. The goal is to gain insights from the data and apply necessary data conversion, transformations, handling missing values, and outliers to prepare the data for further analysis tasks.

2. Dataset Creation

The dataset used in this project was downloaded from Kaggle and initially contained no missing values or outliers. After downloading the dataset, we introduced missing values and outliers for the purpose of demonstrating data preprocessing techniques.
The dataset contains the following columns:

- **Ticket_ID:** A unique identifier for each ticket purchased.
- **Age:** The age of the individual purchasing the ticket.
- **Ticket_Price:** The price of the movie ticket.
- **Movie_Genre:** The genre of the movie (e.g., Comedy, Drama, etc.).
- **Seat_Type:** The type of seat chosen (e.g., Standard, VIP, Premium).
- **Number_of_Person:** The number of people in the group for the purchase.
- **Purchase_Again:** Whether the person plans to purchase tickets again (Yes/No).

After downloading the dataset, we manually introduced the following issues for analysis:

- **Missing Values:**

We added missing values in the Movie_Genre and Seat_Type columns. These missing values were later handled using appropriate preprocessing techniques, such as filling missing genres with the most frequent value and removing rows with missing Seat_Type. Additionally, some missing values in the Number_of_Person column were labeled as 'Alone.'

- **Outliers:**

We introduced outliers in the Age and Ticket Price columns to demonstrate how outliers can be managed. For example, an extreme age value (142) was replaced with the most frequent age, and extreme ticket prices (63.78, 40.09) and the lowest price (1.81) were replaced with the median ticket price.

The dataset now consists of 1,439 records, capturing moviegoer preferences and behaviors, with added missing values and outliers for preprocessing demonstrations.

3. Data Preprocessing Steps

❖ *Handling Missing Values*

We first checked for empty values within the dataset by replacing any empty strings ("") with NA. This step was necessary because in R, NA is recognized as a missing value, while empty strings may not be treated the same way in analyses. To do this, we ran the command df[df == ""] <- NA. This effectively marked all empty string cells as missing values.

Next, we used the function colSums(is.na(df)) to check for missing data in the columns. This gave us a clear picture of which columns had missing values. We noticed that columns such as Seat_Type and Movie_Genre had a few missing entries, which were expected since we had intentionally introduced missing values.

For categorical variables like Movie_Genre, we decided to fill the missing values with the most frequent genre in the dataset. This decision was based on the idea that replacing missing values with the most frequent category preserves the overall distribution of genres in the dataset. To accomplish this, we calculated the frequency of each genre, sorted them in descending order, and used the most frequent genre to fill the missing values.

For numerical variables, such as Number_of_Person, we replaced any missing values(Alone) with 0. This was done because a missing entry for the number of people attending could reasonably be interpreted as “no one attended” or “no data.”

Finally, after applying these changes, we ran the check colSums(is.na(df)) one more time to ensure that there were no remaining missing values in the dataset.

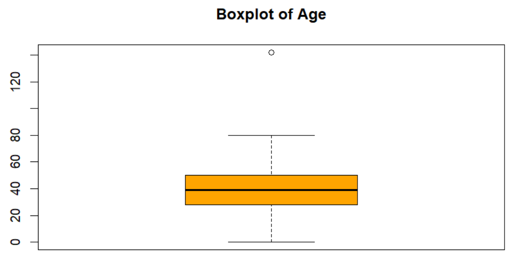
This method of handling missing values was chosen to ensure the dataset was complete and ready for analysis while maintaining the integrity of the data’s overall structure. The steps we followed were based on filling the missing data in a way that kept the dataset balanced and suitable for the analyses we planned to conduct.

❖ *Outlier Handling*

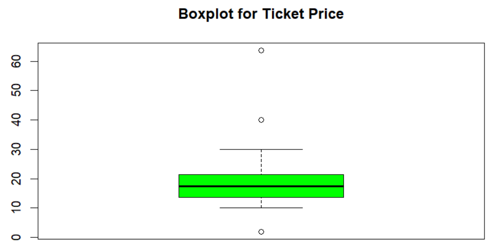
Once we handled missing values, the next step was to detect and manage outliers. Outliers are values that significantly differ from the rest of the data. These can appear due to data entry mistakes, measurement errors, or sometimes due to valid, but rare, occurrences. However, outliers can have a strong impact on analysis results, which is why it’s important to detect and handle them properly.

Outliers can skew the results of any analysis. For example, they can pull the mean (average) toward extreme values, leading to inaccurate conclusions. By identifying and handling outliers, we ensure the analysis reflects the true distribution of the data, allowing for more reliable insights.

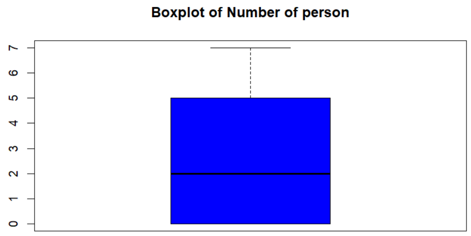
To identify outliers, we used a boxplot. This is a visual tool that helps us see the distribution of the data. The boxplot shows the median (middle value), the spread of the data (quartiles), and any data points that fall outside the typical range (which are identified as outliers).



• **For the Age column,** we used a boxplot to visualize the data. The boxplot clearly showed a value of 142, which was far beyond the typical age range, making it outlier.



• **For the Ticket Price column,** we used the same method and found that there were extreme values such as 40.09, 1.81, and 63.78 that were outliers, standing outside the whiskers of the boxplot.



• **For the Number of Person column,** no outliers were detected. The data for the number of people seemed to fall within a reasonable and consistent range, so we didn't need to take any action here.

After identifying the outliers, we decided to handle them by replacing them with more reasonable values. This ensures that the extreme values do not distort the data and impact the results.

• For the Age column, the outlier value of 142 was replaced with the most frequent age in the dataset. This helps maintain the consistency of the data and avoids the outlier from affecting the analysis.

• For the Ticket Price column, the extreme values were replaced with the median ticket price. The median is a good choice because it reflects the middle value in the data, making it a more realistic replacement for extreme outliers.

By handling missing values and outliers, we ensure that the dataset is clean, accurate, and ready for further analysis. Outliers, if left unhandled, can distort statistical measures and predictions, so detecting and managing them is essential. Replacing outliers with reasonable values helps maintain the integrity of the data and ensures that subsequent analysis is based on valid, realistic information.

❖ *Data Conversion*

For this project, we ensured the data was in the correct format for efficient analysis. Here’s what we did:

1. Removed Unneeded Column:

We removed the Ticket_ID column from the dataset. This column contained only ticket numbers, which were not relevant for our analysis. Removing unnecessary columns helps keep the dataset clean, focused, and easier to work with.

2. Changed “Yes” and “No” to Numbers:

The Purchase_Again column contained the values “Yes” and “No,” indicating whether a person plans to purchase tickets again. For better analysis and ease of use, we converted these categorical values into numerical values: we replaced “Yes” with 1 and “No” with 0. This transformation allows us to use the data in mathematical and statistical models, as most models require numeric input.

Numerical data is required for most analytical models, and these conversions made the dataset compatible with a wide range of algorithms that can perform predictive analysis or classification.

By doing these conversions, we ensured that the data was in the right format for analysis, making it more efficient and ready for modeling.

❖ *Data Transformation*

we applied some key transformations to make our data ready for analysis. Here’s what we did:

1. Created Dummy Variables:

• For columns that contained categorical data (like “Movie_Genre” and “Seat_Type”), we used dummy variables. This process converts text data into numerical values so that it can be easily used in statistical analysis or machine learning models.

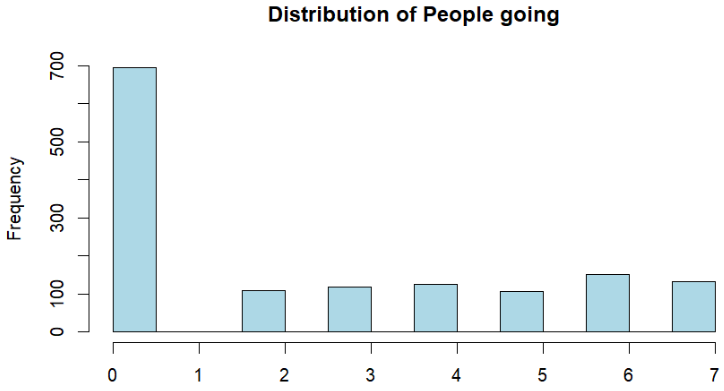
• For example, the “Movie_Genre” column, which contains categories like “Comedy” and “Horror”, was converted into separate columns with 0s and 1s for each category. This allows models to understand the data better.

2. Min-Max Scaling:

• We applied Min-Max scaling to the numerical columns, including “Age”, “Ticket_Price”, and “Number_of_Person”. This scaling process normalizes the values in each column, making sure that they fall between a minimum value (0) and a maximum value (1).

• By doing this, we avoid any individual column from dominating the analysis due to its range of values. This ensures fairness and makes the model perform better.

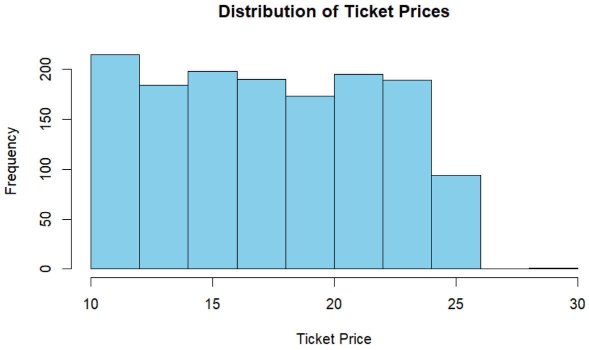
4. *Key Findings from the Analysis*



1. **Distribution of People Going**:

• A large number of people prefer attending alone (0). This indicates that most ticket purchases are for single-person attendance.

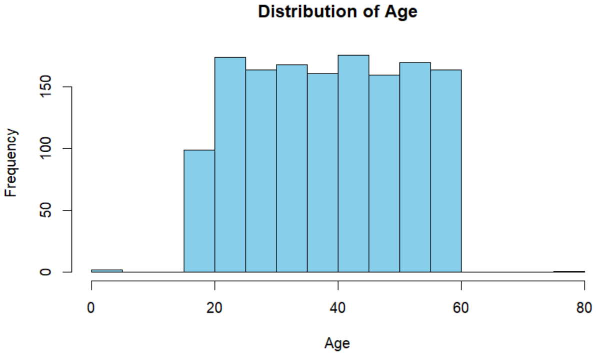
• The frequency of people attending in groups (1-7) is much lower.



2. **Distribution of Ticket Prices**:

• Ticket prices are fairly evenly distributed, with most tickets falling between 10 and 25.

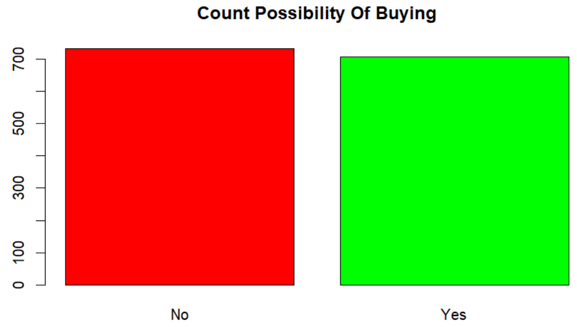
• There is no significant skew, and the prices are relatively spread out.



3. **Distribution of Age**:

• The age distribution is quite spread out, with most attendees being in their 20s to 50s, showing a diverse audience for the event.

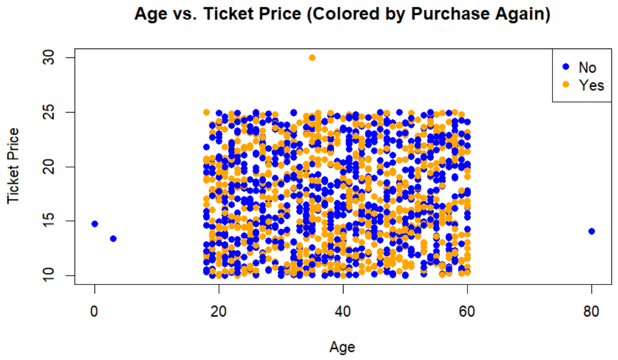
• There are fewer younger and older participants.



4. **Count Possibility of Buying**:

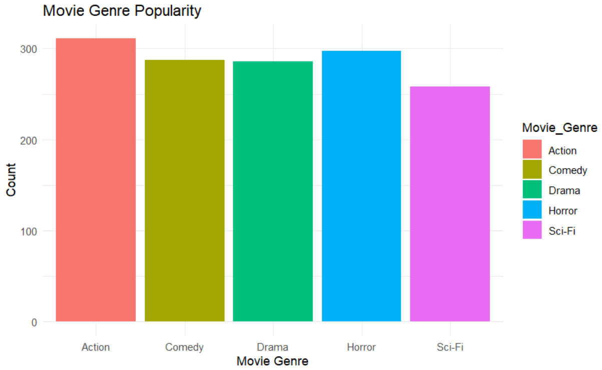
• There is a clear contrast between those who would purchase tickets again (“Yes”) and those who wouldn’t (“No”).

• The number of “No” responses is much higher, indicating a lower likelihood of repeat purchases.



5. **Age vs. Ticket Price (Colored by Purchase Again)**:

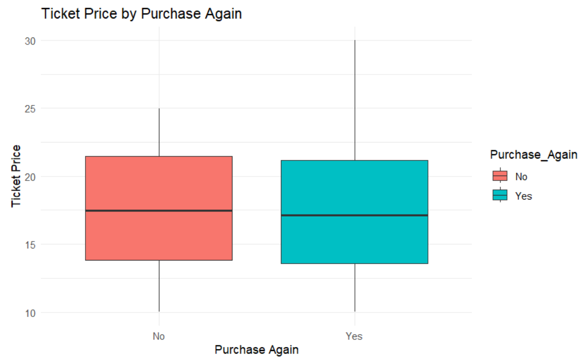
• People of all ages and price ranges are equally likely to either buy again (“Yes”) or not buy again (“No”).



6. **Movie Genre Popularity**:

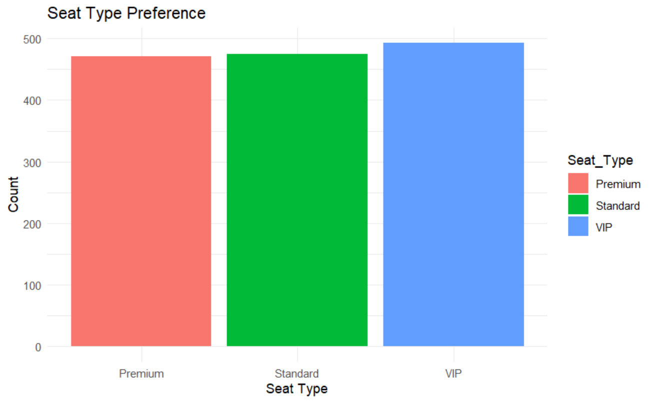
• Action, Comedy, Drama, and Horror movies are the most popular genres, with action leading in terms of ticket preference.

• Sci-Fi has the least popularity but still holds a solid portion of the audience.



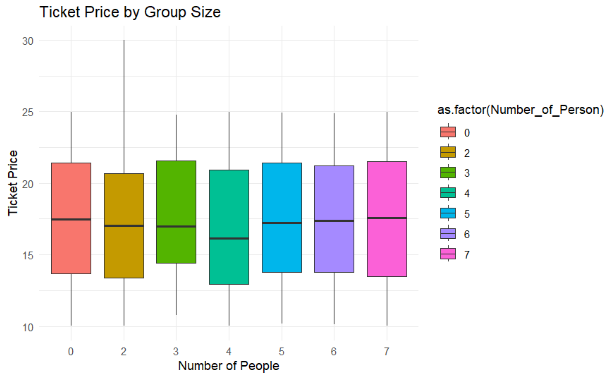
7. **Ticket Price by Purchase Again**:

- People who bought tickets again (“Yes”) tend to pay slightly higher prices on average compared to those who didn’t (“No”).
- Both groups show a similar spread of prices, though the median is higher for repeat buyers.



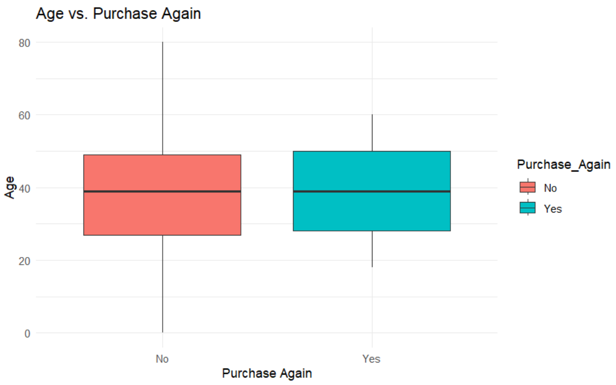
8. **Seat Type Preference**:

- VIP seats are the most popular, with Standard and Premium seats closely following.
- This shows a balanced preference across all seat types, but VIP seats dominate.



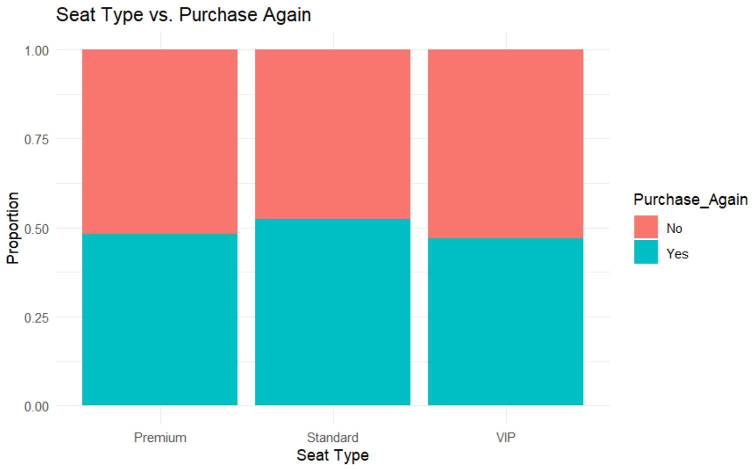
9. **Ticket Price by Group Size**:

- People attending in groups of 3 and 7 have the highest ticket prices on average, while those attending in groups of 2 and 4 pay slightly lower.
- Smaller and larger groups appear to prefer more expensive tickets.



10. **Age vs. Purchase Again**:

• There is a similar likelihood of purchasing again ("Yes") or not purchasing again ("No") across both younger and older age groups.



11. **Seat Type vs. Purchase Again (Proportion Chart)**:
• The proportion of customers who bought again is almost equal across all seat types (Premium, Standard, VIP).
• The chart shows that people who purchased again (represented by the blue color) have a similar distribution for each seat type, indicating no strong preference for re-purchasing based on seat type.

5. *Justification for Transformations and Outlier Handling*

5.1. Outlier Handling

Outliers are data points that are significantly different from the rest of the data and can have a disproportionate impact on statistical analysis, leading to biased or incorrect results. Proper handling of outliers is essential for the following reasons:

1. **Accuracy of Analysis:** Outliers can distort key statistical metrics like the mean, making them unrepresentative of the data as a whole. For example, a single extreme ticket price (e.g., 63.78) can skew the average, leading to conclusions that do not reflect typical behavior. By identifying and managing outliers, we ensure that these extreme values do not distort the overall analysis, leading to more reliable insights.
2. **Improving Model Performance:** Outliers can negatively affect machine learning models by influencing predictions and overfitting. Models trained on datasets with outliers can exhibit poor performance because the model may give undue weight to these outliers. Handling outliers ensures that the model is not distracted by these extreme values and focuses on the underlying trends in the data.
3. **Consistency and Integrity:** Replacing extreme values with more reasonable ones (such as the median or most frequent value) maintains the consistency of the data without discarding potentially useful information. It ensures the dataset remains realistic and reliable for further analysis, preventing anomalies from distorting the results.

5.2. Data Transformation:

Data transformation is a crucial step in preparing the dataset for analysis or machine learning. The transformations applied, such as creating dummy variables and Min-Max scaling, are done for the following reasons:

1. **Converting Categorical Data (Dummy Variables):** Many statistical techniques require numerical input. Categorical columns (like “Movie_Genre” and “Seat_Type”) need to be converted into numerical values to make them usable in models. Creating dummy variables allows us to transform these categorical variables into a format.
2. **Normalizing Numerical Data (Min-Max Scaling):** Numerical columns like “Age”, “Ticket_Price”, and “Number_of_Person” can have varying ranges of values. For instance, ages may range from 0 to 80, while ticket prices might range from 1 to 100. If these features are not scaled, variables with larger ranges can dominate the model's predictions. Min-Max scaling transforms all values into a consistent range (between 0 and 1), ensuring that each feature contributes equally to the model. This makes the analysis fairer and allows the model to better interpret relationships between features.
3. **Improving Model Performance:** Models perform better when the input features are on a similar scale and are appropriately formatted. By normalizing the data and handling categorical variables properly, we improve the model’s efficiency and the accuracy of predictions. Transformations make the data more suitable for algorithms that assume or work better with data in a specific range or format.

6. Conclusion

The data preprocessing steps outlined above were essential in preparing the dataset for further analysis and modeling. By handling missing values, transforming categorical variables, and addressing outliers, the dataset was cleaned and normalized. The key findings provide useful insights into the behavior of moviegoers, and the transformations ensure that the dataset is ready for predictive modeling or further statistical analysis.