BS(SE)V1.0

|  |  |  |  |
| --- | --- | --- | --- |
| **Assignment # 2** | | |  |
| **Mapped CLO** | **Mapped GA** | **Mapped Learning Level** | **SDG** |
| CLO2 | GA 5 (Modern Tool Usage) | C3 (Apply) | 4 & 8 |

**Data Science Assignment# 02**

**Group Members:**

Syed Imran Murtaza (65196)

Muhammad Asjad (63915)

**Question:**

You are provided with a consolidated dataset of online gamers behavior collected as a result of Assignment 1. The data set is raw and have many issues. Your task is to **clean**, **preprocess**, and apply **exploratory data analysis (EDA)** on this fragmented dataset to derive meaningful insights. Task includes

1. **Data Cleaning:**
   * Handle missing values, incorrect data types, and duplicates.
   * Identify and treat outliers if necessary. o Data Cleaning resultant and justification should be available (as statistical reports or visualizations)

1. **Exploratory Data Analysis (EDA):**
   * Use tool to analyze and visualize trends, correlations, and patterns.
   * Create summary statistics for key columns.
   * Plot histograms, scatter plots, and box plots to identify distributions and relationships.
2. **Insight & Interpretation** o Write a **brief report (150–200 words)** interpreting your findings o What patterns or behaviors are most prominent?
   * Any surprising relationships or anomalies?

**1. Data Cleaning Process:**

### **1. Handling Missing Values**

**Detection Method:**  
To detect missing values in each column, we used the =COUNTBLANK(A:A) function in Excel. This helped us identify how many data points were missing in each part of the dataset.

**Missing Values:**

* **age**: 1 missing value
* **gender**: 0 missing values
* **city**: 20 missing values
* **country**: 36 missing values
* **education\_level**: 20 missing values
* **occupation**: 40 missing values
* **monthly\_income\_bracket**: 82 missing values
* **earning\_bracket**: 95 missing values
* **hours\_played\_per\_day**: 38 missing values
* **how\_often\_do\_you\_play\_online\_games**: 68 missing values
* **what\_motivates\_you\_to\_play\_online\_games**: 17 missing values
* **PlayerLevel**: 0 missing values
* **GameDifficulty**: 0 missing values

## **How We Handle Missing Values:**

### We apply different strategies for each:

### **Age Column**

In the age column, we noticed that some values were written as ranges such as "18–24", "25–34", or "Under 18" instead of exact numbers. To make the data consistent and usable, we assigned a single representative number to each range:

* "18–24" was replaced with 21
* "25–34" was replaced with 30
* "Under 18" was replaced with 16

After converting all entries to numeric values and removing any invalid data, we calculated the median age, which is 21.0. We used this value to fill in any missing age entries.

### **Hours Played Per Day**

The hours\_played\_per\_day column had many unrealistic values, such as 108 or 144, which exceed the number of hours in a day. We assumed these were recorded in minutes instead of hours, so we converted them by dividing the values by 60.

We applied the following Excel formula to clean the data:

=IF(ISNUMBER(I2), ROUND(IF(I2>24, I2/60, I2), 2), "")

This formula keeps valid hour entries as they are, converts larger values from minutes to hours, and leaves non-numeric cells blank.

After applying the formula, we found that some cells were still empty. We filled those with the median of the cleaned data, which was 1.93 hours. This ensured consistency and prevented gaps in the analysis.

### **Categorical Text Columns**

For columns that contain text-based responses—such as city, education\_level, how\_often\_do\_you\_play\_online\_games, and what\_motivates\_you\_to\_play\_online\_games—we used the most frequently occurring answer (mode) to fill in the missing values. Based on our dataset:

* Missing cities were filled with "Karachi"
* Missing education levels were filled with "Bachelor's"
* Missing play frequency was filled with "Daily"
* Missing motivation was filled with "Competition"

Using the most common answers helped us keep the data natural and reliable without inserting guesses.

### **Other Fields**

For columns with moderate missing entries, we used general but appropriate values:

* The occupation column was filled with "Not Specified"
* The monthly\_income\_bracket column was filled with "Not Disclosed"
* The country column was filled with "Pakistan"

This helped us keep those entries in the dataset while clearly showing that the actual values were not provided.

### **Dropped Column**

The earning\_bracket column had more than 75 percent missing values. Because this made the data unreliable and difficult to analyze, we decided to remove this column entirely to maintain the overall quality of the dataset.

2. Incorrect data type:

Some columns in our dataset were not in the right format. For example, numbers were stored as text, and some values were written in words or ranges. This makes it hard to calculate or analyze properly.

We fixed these problems while cleaning the missing values. This saved time and made the data ready for use and fill the missing blanks.

* In the **age column**, we changed ranges like "18–24" into numbers like 21. Then we made the whole column numeric and filled the blanks using the median, which is 21.
* In the **hours\_played\_per\_day column**, we found big numbers like 108 or 144. We treated them as minutes and divided them by 60 to get hours. Then we filled empty cells with the median, which is 1.93.

3. Duplicates:

We also checked the dataset for **duplicate rows** these are rows where all values are exactly the same across every column. Duplicate entries can affect analysis by repeating the same data and giving inaccurate results (like double counting users or changing averages).

To detect duplicates, we used Excel’s **Remove Duplicates** feature:

* We selected the entire dataset.
* Then clicked **Data > Remove Duplicates**.
* All columns were selected to make sure full rows were compared.

After checking, we found that:

**There were no exact duplicate rows** in our data.

## **Data Formatting and Normalization**

After fixing missing values and wrong data types, we focused on giving our dataset a clean and consistent format. Many columns had spelling mistakes, extra spaces, or different styles of writing. Here’s how we fixed that, step by step:

### **City Column**

City names were written in all sorts of ways some in small letters like "lahore", some in all caps like "ISLAMABAD", and some had typos like "Faisalbad". To clean this:

* We used **title case** to make them look like "Lahore", "Islamabad", etc.
* Removed extra spaces
* Fixed spellings like "Faisalbad" = "Faisalabad"

This helped us avoid duplicate values that looked different but meant the same city.

### **Country Column**

In the country column, people wrote "pakistan", "Pakstan", or even "paki". To fix this:

* We changed all of them to "Pakistan" using proper spelling and formatting
* Filled missing values with "Pakistan" too, since most users were from there

This made the data clean and consistent for location-based analysis.

### **Education Level**

Education data was messy — it had typos, mixed formats, and unclear phrases like:

* "Matric", "Metric", "9th" grouped as **"Matric"**
* "Inter", "Enter", "Intermidi" = **"Intermediate"**
* "BBA", "Graduated" = **"Bachelor's Degree"**
* "Postgraduate" = **"Master's Degree"**

We shortened long phrases and used standard names to make the data easier to compare.

### **Occupation Column**

This was the messiest column. It included:

* Typos like "buisnesse" instead of "Business"
* Repeated roles like "Freelancing" and "Freelancer"
* Long or unclear job descriptions like "Student hybrid Job remote internships"
* Some entries were not jobs at all, like "Karachi"

To fix this:

* We cleaned spelling, fixed formatting, and removed wrong entries
* Grouped jobs into clear categories like:

|  |  |
| --- | --- |
| Example Value | Final Category |
| "student", "Undergraduate" | Student |
| "Freelancing", "Freelancer" | Freelancer |
| "Graphic Designer" | Designer |
| "Doctor", "Pharmacist" | Medical |
| "CEO" | Executive |
| "Nil", "Nothing" | Not Specified |

### **Monthly Income Bracket**

This column had different formats like "Below PKR 20,000", "PKR 20k–50k", and some in dollars like "$500 - $1000". Many entries were just "Not Disclosed". So we grouped them into five clean ranges:

* **Low Income**: Below PKR 20,000
* **Lower-Middle Income**: PKR 20k–50k
* **Upper-Middle Income**: PKR 50k–100k
* **High Income**: Above PKR 100k or dollar-based ranges
* **Not Disclosed**: For skipped or unclear answers

This made income levels easier to work with while keeping privacy in mind.

### **How Often Do You Play Online Games?**

People answered this in different ways: "Daily", "3–5 times a week", "1–2 times a week", etc. We grouped them into three simple categories:

* **Daily** – plays every day
* **Frequently** – 3–5 times a week
* **Occasionally** – 1–2 times a week

This helped us understand and compare gaming habits better.

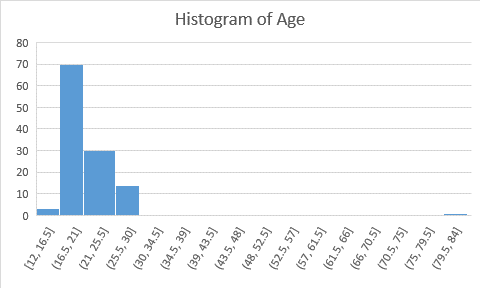
**Exploratory Data Analysis (EDA):**

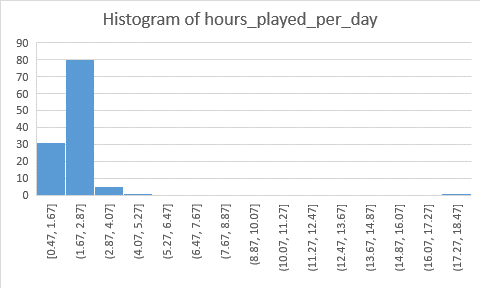
To understand the behavior and trends in our dataset, we performed Exploratory Data Analysis (EDA) using Microsoft Excel. This helped us identify patterns, relationships, and outliers in the data.

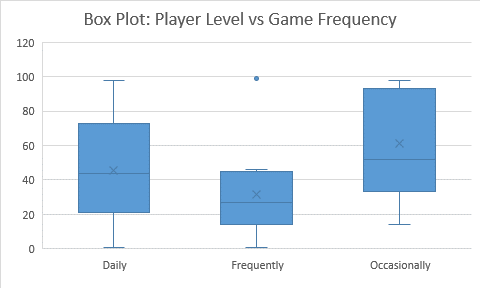
* **Summary statistics for key columns:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Mean** | **Median** | **Minimum** | **Maximum** | **Standard Deviation** |
| age | 22.03 | 21.0 | 12 | 81 | 118.0 |
| hours\_played\_per\_days | 1.99 | 1.93 | 0.47 | 18 | 1.63 |
| player\_level | 44.92 | 41.0 | 1.0 | 99.0 | 28.73 |

* **Visual Analysis**







* **Insight & Interpretation (Summary)**

### **Histogram of Age::**

The histogram shows that most online gamers fall within the **16 to 25 age range**, with the largest group being between **16.5 and 21 years old**. This confirms that online gaming is especially popular among teenagers and young adults. A smaller number of users are in the **25 to 30** range, and very few are older than 35. One rare outlier appears in the **75–84 age bracket**, which may be a data entry mistake or an exception. Overall, the chart clearly indicates that gaming is mostly preferred by a young audience.

### **Histogram of: Hours Played Per Day Interpretation:**

This histogram shows how many hours users spend on online games each day. Most users fall between **1.67 to 2.87 hours**, with another noticeable group around **0.47 to 1.67 hours**. This means the majority of players spend **1 to 3 hours per day** gaming. There are very few users who play beyond **4 hours**, and almost none who go over **10 hours daily**. One rare outlier falls between **17 to 18.5 hours**, which is likely an error or an extreme case. Overall, this chart shows that most gamers have a balanced and limited daily play time.

### **Scatter Plot: Hours Played vs Player Level:**

This scatter plot explores the relationship between how many hours a user plays per day and their game level. Most data points are grouped between **0 to 4 hours** of gameplay, and levels ranging from **0 to 100**. Although there is no strong linear pattern, we can see a **slight upward trend** users who play more tend to reach higher levels. However, there are exceptions. For example, one user playing around **18 hours** has a **moderate level**, and another playing only about **5 hours** is at a **very high level**. These exceptions show that while play time contributes to progress, other factors (like skill or type of game played) may also matter.

### **Box Plot: Player Level vs Game Frequency:**

This box plot compares player levels across three groups based on how often they play: **Daily**, **Frequently**, and **Occasionally**.

* **Daily players** have the widest range in levels, from very low to near maximum. Their average level is moderate to high.
* **Frequent players** (3–5 times a week) mostly fall in a smaller range, with a lower average level than daily gamers.
* Surprisingly, **occasional players** (1–2 times a week) have a higher median and many high-level outliers, suggesting that some skilled players progress quickly even with limited playtime.

Overall, this chart shows that while playing daily often helps players level up, skill and strategy can also play a role in performance especially for occasional users.

**Member’s Work Distribution:**

|  |  |  |
| --- | --- | --- |
| ID: | Name: | Work Done: |
| 65196 | Syed Imran Murtaza | * Handle missing values and incorrect data types, Data formatting and Normalization, Visual Analysis, Summary |
| 63915 | Muhammad Asjad | Handle Duplicate values, Data formatting and Normalization, Summary statistics for key columns, Summary |

### **Submission File Included:**

* **Cleaned Excel Dataset**
* **Final Report (Word Document)**
* **Reflection Notes from Each Group Member**

**"Without data, you're just another person with an opinion."**  
— W. Edwards Deming