PS6 - ECON 5253

Hannah Bermudez

March 2025

1 Introduction

This document details the steps taken to clean and transform a dataset downloaded from Kaggle, analyze its structure, and visualize key relationships.

2 Loading Required Packages

To begin, I loaded the necessary R packages for data manipulation, visualization, and modeling:

```
library(reticulate)
library(tidyverse)
library(ggplot2)
library(modelsummary)
library(broom)
```

3 Installing and Setting Up Kaggle API

The Kaggle API was installed using the following command in the command prompt:

```
python –m pip install kaggle
```

In R, I ensured the API was available and set the API key:

```
reticulate::py_require("kaggle")
Sys.setenv(KAGGLE_CONFIG_DIR = "C:/Users/berm0006/Downloads/")
```

4 Downloading the Dataset

The dataset was downloaded from Kaggle and extracted:

```
import kaggle
kaggle.api.dataset_download_files('deadier/play-games-and-success-in-
students', path='C:/Users/berm0006/Downloads/', unzip=True)
```

5 Loading and Exploring the Data

The dataset was then loaded into R:

```
game <- read.csv("C:/Users/berm0006/Downloads/gameandgrade.csv")
head(game)
```

6 Data Cleaning

I removed any rows that contained missing values:

```
game <- na.omit(game)
```

I also eliminated a data misclassification with the Playing.Games variable:

```
game <- game[game$Playing.Games %in% c(0, 1), ]
```

Finally, the Grade variable was converted to numeric format in order to properly use it:

```
game$Grade <- as.numeric(game$Grade)
```

7 Data Visualizations

7.1 Visualization - 1

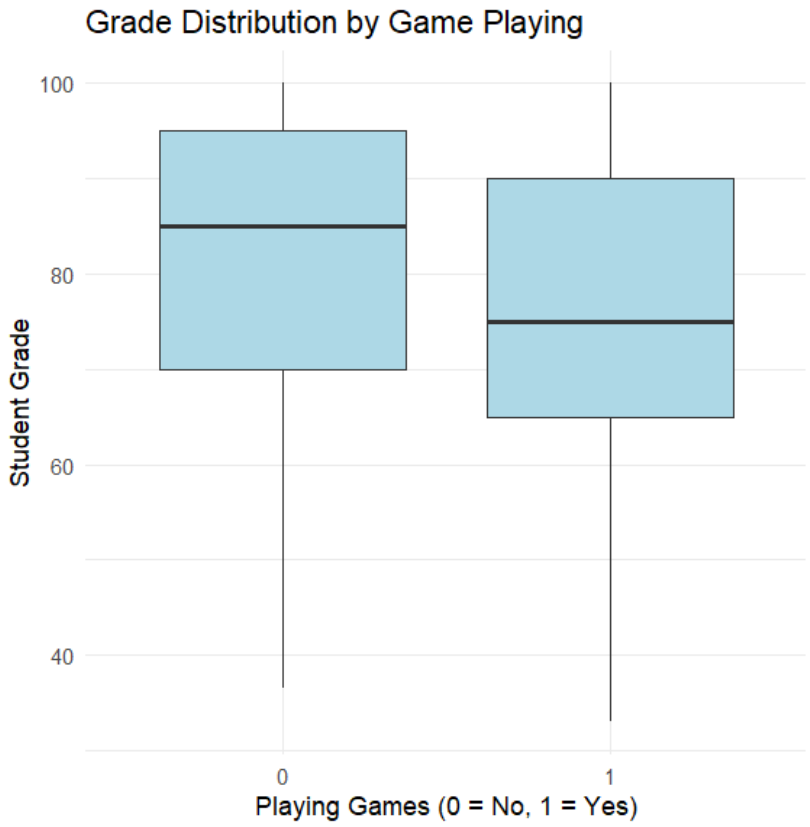

Figure 1: Boxplot of Grade Distribution by Game Playing

- The median grade is higher for non-gamers than for gamers.
- The whiskers suggest that both groups have a wide spread of grades, but gamers have a lower median.
- There are no extreme outliers present.

Interpretation: Students who do not play games tend to have slightly higher grades on average compared to those who do play games.

7.2 Visualization - 2

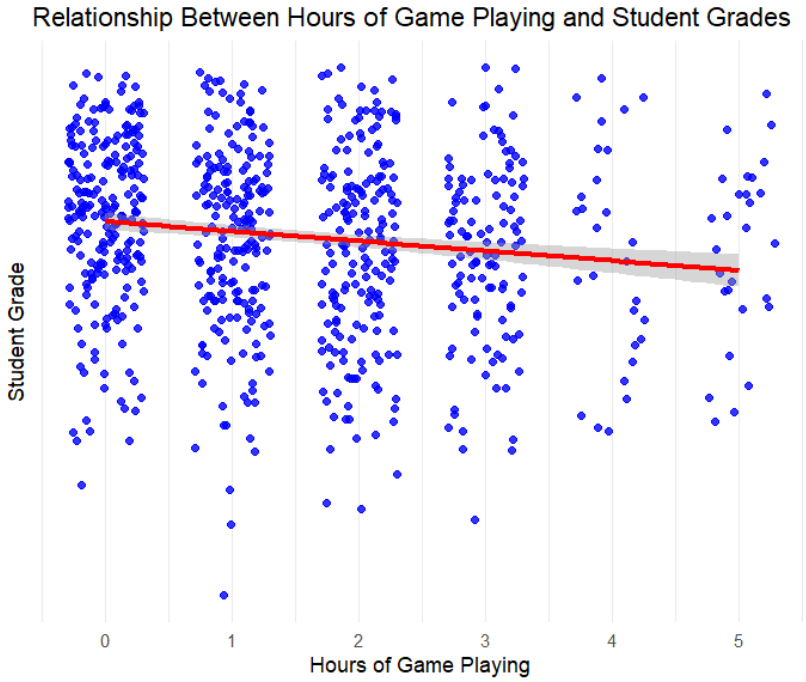

Figure 2: Scatterplot with Regression Line

- The red regression line with a shaded confidence interval shows a slight negative trend, suggesting that as hours of game playing increase, student grades tend to decrease.
- The data points are widely spread, indicating high variability in student performance regardless of gaming hours.
- The negative slope of the regression line suggests that increased gaming hours may be associated with lower grades, but the effect appears relatively small.

Interpretation: There is a negative relationship between the number of hours spent playing games and student grades, meaning that an increase in hours played will lead to a decrease in student grades.

7.3 Visualization - 3

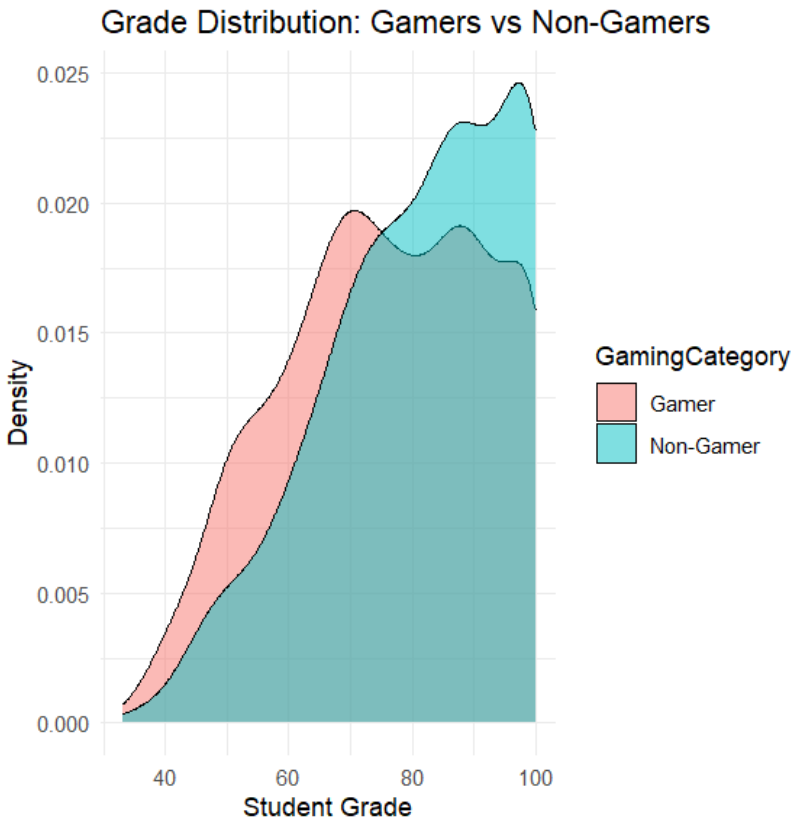

Figure 3: Density Plot of Grade Distribution

- Non-gamers have a higher concentration of grades near the upper end (closer to 100), indicating that they tend to achieve higher grades overall.
- Gamers show a broader distribution, with a peak around the mid-range but fewer high-performing students compared to non-gamers.
- The overlapping areas suggest that while some gamers achieve high grades, non-gamers are more likely to perform better academically on average.

Interpretation: Non-gamers generally tend to have higher grades, whereas gamers exhibit more variability in academic performance, with a greater presence in the lower and mid-range grade categories.