BSAN7204 A1: EXPLORATORY DATA ANALYSIS

Coert, Miss Madeline

Word count: 991 Student no. 49736577

Table of Contents

INTRODUCTION **2**

ANALYSIS **3**

UNIVARIATE ANALYSIS 3
  QQPlot for Hours: *3*
  Hours Histogram: *4*
  Purchases Density Plot: *4*
  Experience Level Bar Chart: *5*
  Insights *5*
BIVARIATE ANALYSIS 6
  Hours vs Type: *6*
  Time vs Type: *6*
  Purchases vs Type: *7*

A/B HYPOTHESIS TESTING **8**

  Hours by Type: *8*
  Purchases by Type: *9*
  Insights *9*

CONCLUSION: **10**

APPENDIX **11**

AI ACKNOWLEDGEMENT 11
RSCRIPT 12

REFERENCES: **15**

INTRODUCTION

GameLab is an emerging video game company hoping to release their next main title game. In a rapidly evolving market, the company focus is clear: increase in-game purchases to drive long-term success. However, as AI features have transformed the gaming landscape, innovation is no longer optional, with high performing games at the baseline of consumer expectations. Yet, there is a critical consideration of how an over reliance on these features could undermine the core experience that players love, immersive storytelling and the sense of escapism (Calleja, 2010). Therefore, an experiment was designed to compare two game types with the same core features except one has AI enhancements. The following analysis aims to address whether consumers prefer AI gameplay over traditional gameplay and if this preference influences their engagement and spending behaviour.

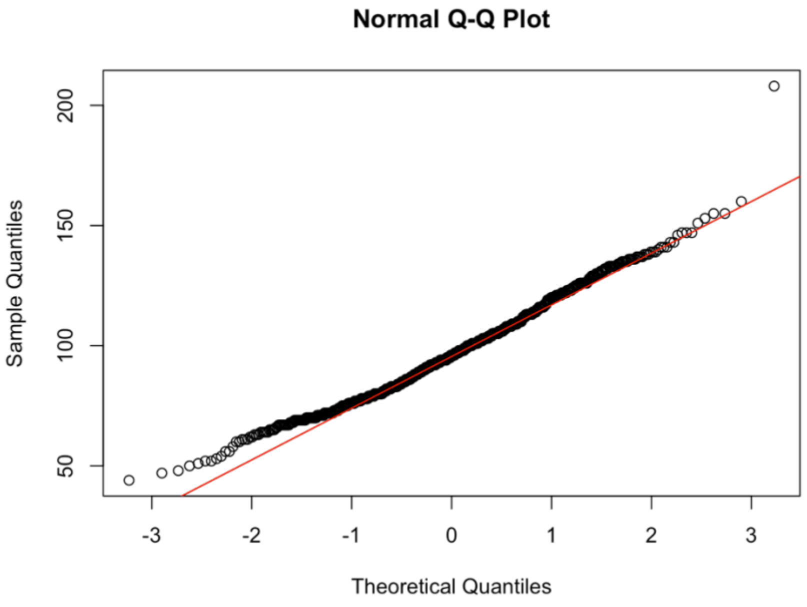
Variable name:	Description:
Game Type (2 levels)	1. A = AI-driven gameplay enhancements 2. B = Traditional gameplay, no AI features
Hours	• Total hours spent playing the designated game over the course of the experiment period.
Experience Level (5 levels)	1. F = Novice 2. G = Advanced beginner 3. H = Competent 4. K = Proficient 5. L = Expert
Time	• Time spent per session, measured in minutes.
Pay	• Hourly pay rate ($).
Purchases	• In-game purchases made during the experiment, with participants given unlimited access to make purchases without real financial costs ($).

ANALYSIS

Univariate Analysis

Engagement is crucial for the business, as more time spent playing leads to increased exposure to in-game content. To address the business problem, the key variables to analyse are hours and purchases as they offer the strongest insights into player engagement.

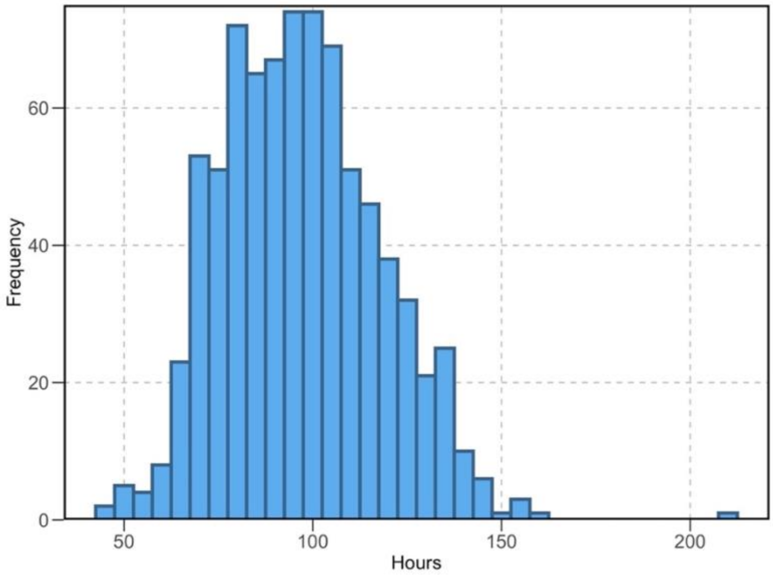
QQPlot for Hours:



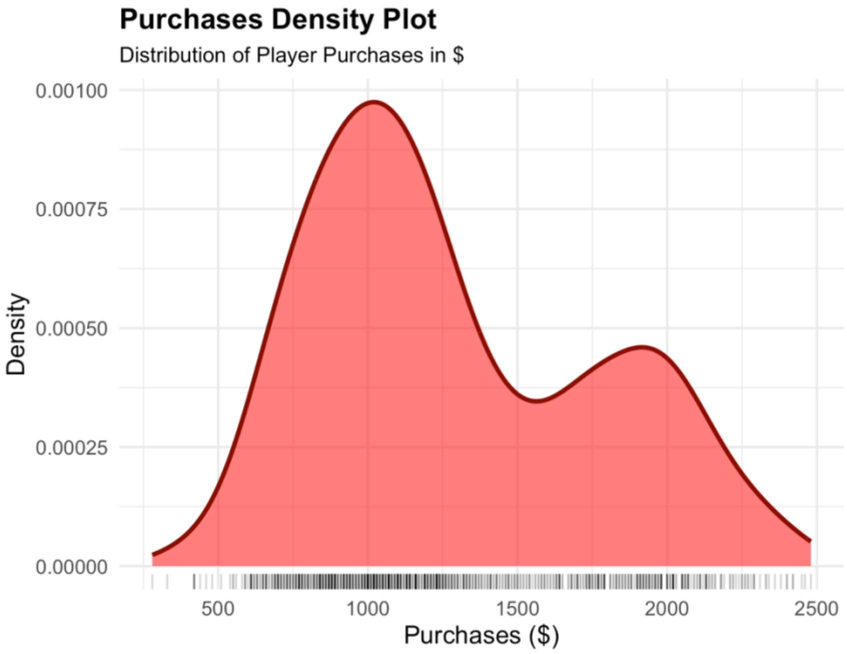
The data appears to be normally distributed as most points are close to the reference line. However, there are slight deviations at the tails indicating the presence of some outliers or mild skewness. Thus, using the mean of 97 hours is appropriate for summarising central tendency.

Hours Histogram:

The histogram displays a positively skewed distribution to the right. Hence, outlining that most players spend a moderate amount of time playing and a smaller number of players engage in longer playtime.

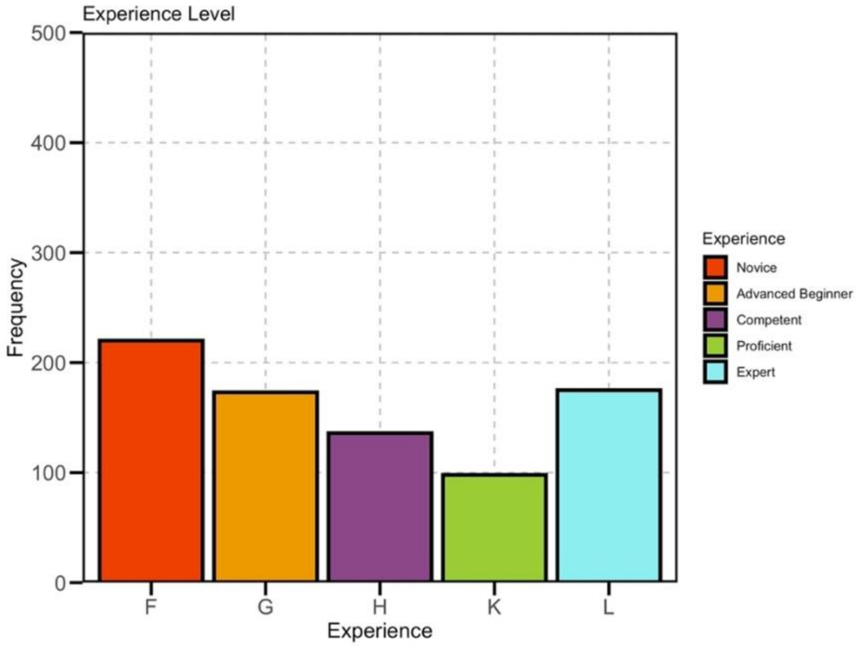


Purchases Density Plot:



This suggests that most players spend a moderate amount on in-game purchases with a few high spenders significantly increasing the average. The distribution appears right skewed**,** indicating that a small group of players contribute disproportionately to revenue.

Experience Level Bar Chart:



The bar chart reveals an imbalance in experience levels, with some more represented than others. This uneven distribution could bias engagement comparisons between game types.

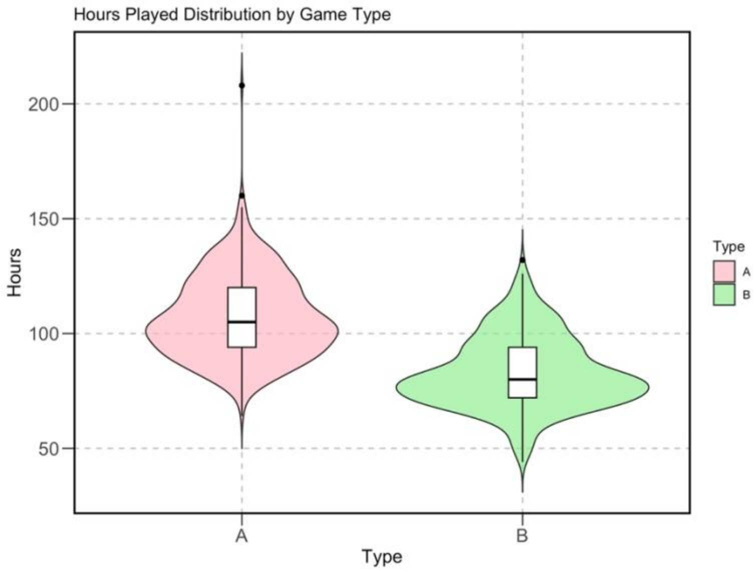
Insights

The univariate analysis outlined a need to discover whether GameLab should focus on increasing median playtime or targeting highly engaged users. Analysing the purchasing behaviour indicated a wide spread suggesting high variability in spending behaviour. Therefore, if purchases are heavily concentrated among a few players, the company should focus on segmenting high spenders rather than increasing playtime. Understanding the differences between experience levels aimed to provide a guide into marketing strategies to attract beginners or help developers adjust mechanics to retain less experienced players. However, the sample created difficulty for identifying experience against engagement factors as the data was not equal across groups hence, introducing bias.

Bivariate Analysis

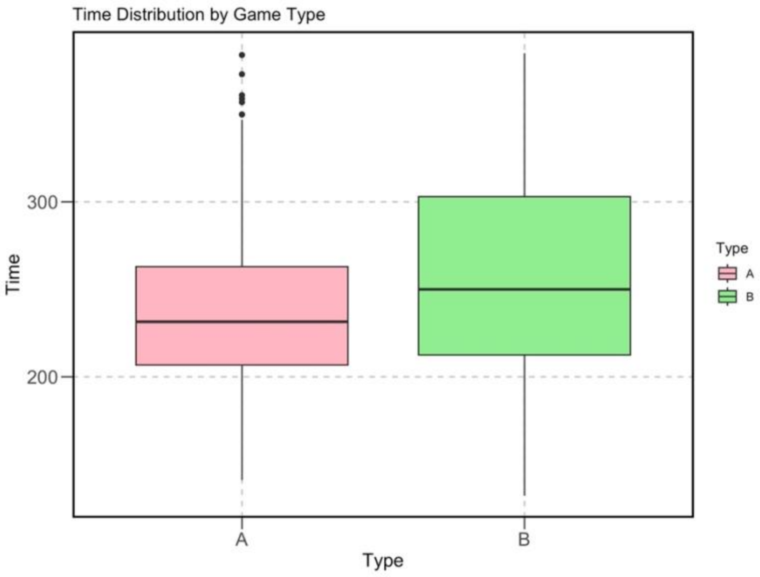
Hours by Type:

The violin plot displays Type A with a higher median playtime and greater variability, suggesting that some players are highly engaged whilst, others are disengaged. In contrast, Type B has a lower median playtime and a more consistent distribution, indicating steadier but less immersive engagement.

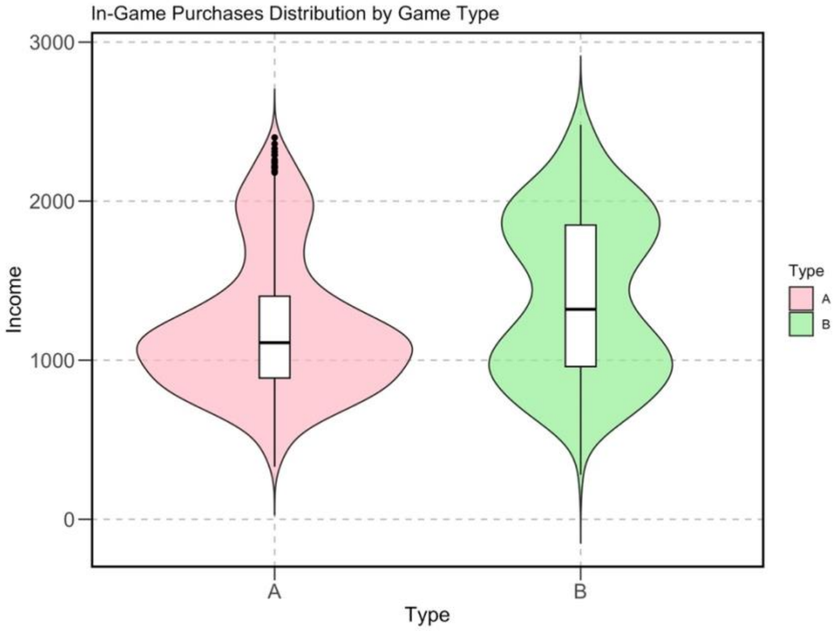


Time by Type:

This visualisation demonstrates Type B with a higher median session duration and a wider variability, implying more immersive gameplay. Type A has shorter and more consistent sessions but includes outliers with significantly longer playtimes, indicating a subset of highly engaged players.



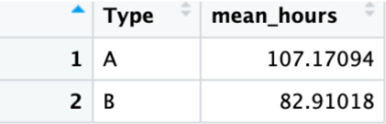
Purchases by Type:



Both games have similar median spending**,** Game A has more extreme outliers, indicating a small group of high spenders. However, Game B has a more evenly distributed spending pattern across players. This suggests that Game A relies on a few big spenders, while Game B fosters more consistent spending**.**

A/B Hypothesis Testing

Hours by Type:



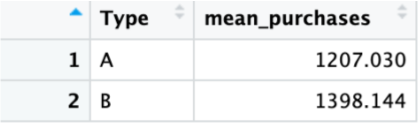
	Type	mean_hours
1	A	107.17094
2	B	82.91018

Two explanations for difference in mean hours spent:

Null Hypothesis (H_0)	- No true difference in mean hours spent on Game A and Game B.
Alternative Hypothesis (H_1)	- There is a true difference in the mean hours spent on Game A and Game B.

When interpreting the results, it is important to acknowledge the potential for statistical errors. A Type I error could lead to incorrect inclusions for an observed difference when no true effect exists. To mitigate this, appropriate significance thresholds have been applied in the analysis. Based on the results from the Welch Two Sample t-test, the p-value was less than 2.2e-16, which is well below the significance level of 0.05. This strongly supports rejecting the null hypothesis, providing compelling evidence that the mean hours spent on both games are significantly different.

Purchases by Type:



	Type	mean_purchases
1	A	1207.030
2	B	1398.144

Two explanations for difference in mean purchases:

Null Hypothesis (H_0)	• No true difference in mean purchases spent on Game A and Game B.
Alternative Hypothesis (H_1)	• There is a true difference in the mean purchases spent on Game A and Game B.

The p-value of 4.75e-08 is statistically significant and the likelihood that the observed difference in purchases is due to random chance is extremely low. As this value is much lower than the significance threshold of 0.05, we can confidently reject the null hypothesis.

Insights

The bivariate analysis identified Game A as more effective in creating engagement based on total hours played however, users spent more time per session on Game B. This suggests that Game B’s longer sessions may enhance retention, while Game A’s outliers suggest potential for deeper engagement. Session length did correlate with spending as Game B’s in-game purchases were consistent. Therefore, to optimise revenue it is suggested that GameLab should optimise monetisation strategies based on consistent spending rather than relying on few high spenders. These conclusions were confirmed with the A/B testing analysis, discovering that players are spending more time and money on one game over the other and that the difference is not due to random chance.

Conclusion:

As GameLab aims to increase in-game spending, Game B emerges as the ideal option, as its consistent spending patterns suggest a clear demand for further upgrades. While incorporating some AI features have the potential to enhance player satisfaction, the spending patterns in Game B suggest that players are already engaged without AI-driven upgrades. These findings challenge the hypothesis that AI features alone drive innovation for GameLab. It is also important to note, however, that because this study used virtual currency rather than participants own money, the results may not fully reflect how players would behave with their own financial investment. Therefore, it is recommended that further experimental research be conducted, focusing on segmenting players more effectively based on their individual preferences to gain deeper insights into their spending behaviour.

Appendix

AI Acknowledgement

AI Generative technologies were utilised in this analysis from the platform ChatGPT.

<table>
<tr><th>Prompt</th><th>Result</th></tr>
<tr><td>How do I code a legend into a bar chart in R?</td><td><ul><li>In ggplot2, the legend is automatically created when you map an aesthetic (like fill or color) to a variable in your data.</li><li>In base R, you manually specify the legend() function to add a legend, passing in the legend labels and the colors corresponding to those categories.</li></ul></td></tr>
<tr><td>How would you rate a 5-point scale measuring gamer experience levels?</td><td><ul><li>Novice, advanced beginner, competence, proficient, expert.</li></ul></td></tr>
<tr><td>How do I make my introduction more persuasive?</td><td>1. Start with a Hook: Begin with a statement or question that grabs attention. It could be a startling fact, a compelling statistic, or a thought-provoking question that relates to the topic you're addressing.
2. Clarify the Issue or Problem:
Clearly define the problem or issue you're addressing. Make it relevant to your audience, showing them why it's important and worth paying attention to.
3. Present the Benefit or Solution: Outline the solution or the main point of your argument. Make sure it's clear how your approach will address the issue effectively.
4. Establish Credibility: Establish your authority or experience in the subject. Mention any relevant research, expertise, or background to show that your opinion is grounded in facts or experience.
5. Create a Sense of Urgency or Importance: Make your audience feel that the topic is urgent and important, or that it's something that requires immediate action or attention.
6. Conclude with a Strong Statement or Preview: End the introduction with a compelling statement that sets up your argument and intrigues your audience, making them eager to read more.</td></tr>
<tr><td>What features can I add in my code to make it more engaging? (see appendix for purchases density plot code)</td><td>Add a Rug Plot (Show Individual Data Points)
A <b>rug plot</b> along the x-axis highlights individual data points:
r
CopyEdit<pre>Purchases_density <- Gaming %>%
  ggplot(aes(x = Purchases)) +
  geom_density(fill = "#B22222", alpha = 0.6) +
  geom_rug(alpha = 0.2, colour = "black") +
  labs(title = "Purchases Density Plot", x = "Purchases ($)", y = "Density") +
  theme_minimal()</pre>💡 <b>Why?</b> It gives insight into the <b>density of individual purchases.</b></td></tr>
<tr><td>Can you review my report and advise on where I can cut words down without removing key info?</td><td>Refinement Suggestions
<b>Introduction</b> – Reduce redundancy and make the problem statement clearer. The discussion on AI as both an opportunity and a challenge is strong but could be more direct.
<b>Analysis</b> – Some explanations are detailed but could be made more concise, particularly when describing visualisations.
<b>Repetitive phrasing</b> – Some points (e.g., engagement differences between game types) are discussed in multiple places. Consolidating insights will help save words.</td></tr>
<tr><td>How can I improve my problem statement to make it answerable with data?</td><td>To ensure your problem statement is <b>answerable with data</b>, it should be <b>clear, specific, and measurable</b>. Right now, your statement asks whether consumers <i>prefer</i> AI-enhanced gameplay, but "preference" can be subjective and hard to quantify unless it's linked to measurable behaviours (e.g., playtime, purchases, engagement).Improved Problem Statement:<i>"This analysis examines whether AI-enhanced gameplay influences player engagement, in-game spending, and overall satisfaction compared to traditional gameplay."</i></td></tr>
</table>

RScript

```
#Read Data
Gaming <- read.csv("EDA.csv")
#Installing packages - ggplot2 and dplyr
install.packages("ggplot2")
install.packages("dplyr")
#Loading packages
library(ggplot2)
library(dplyr)
#Formatting Variables - converting chr to factors
Gaming[c("Type", "Experience")] <- lapply(Gaming [c("Type", "Experience")], as.factor)
#Format Variables - convert int to num
Gaming[c("Hours", "Time", "Purchases", "Income")] <- lapply(Gaming [c("Hours", "Time", "Purchases", "Income")], as.numeric)
#Check for missing values
colSums(is.na(Gaming))

#UNIVARIATE ANALYSIS
#Summary Statistics
summary(Gaming$Hours)
#Distribution mean or median?
qqnorm(Gaming$Hours)
qqline(Gaming$Hours, col="red")
#Histogram - Levels
table(Gaming$Hours)
#Histogram for Hours
Hours_histo <- Gaming %>% ggplot(aes(x=Hours)) + geom_histogram(binwidth = 5, fill="steelblue2", colour="steelblue4", linewidth = 1.2) + labs(Title = "Hours Histogram", x="Hours", y="Frequency") + scale_y_continuous(limits = c(0, 75), expand = c(0, 0)) + theme(panel.background = element_blank(), plot.background = element_blank(), panel.border = element_rect(colour = "black", fill = NA, linewidth = 1.5), axis.text = element_text(size = 14), axis.title = element_text(size = 14), axis.ticks.length = unit(0.3, "cm"), panel.grid.minor=element_blank(), panel.grid.major=element_line(colour = "grey", linetype = 2))
#View Histogram
Hours_histo
#Save
ggsave(filename = "Hours_Histogram.jpg", path = "/Users/maddycoert/", width = 8, height = 6, dpi = 300)
#Purchases Density Plot
Purchases_density <- Gaming %>%
  ggplot(aes(x = Purchases)) + geom_density(fill = "red", colour = "red4", alpha = 0.6, linewidth = 1.2) + geom_rug(alpha = 0.2, colour = "black") + labs(title = "Purchases Density Plot", subtitle = "Distribution of Player Purchases in $", x = "Purchases ($)", y = "Density") + theme_minimal(base_size = 14) + theme(plot.title = element_text(face = "bold", size = 16), plot.subtitle = element_text(size = 12), plot.caption = element_text(size = 10, colour = "grey50"))
```

```
#View Density
Purchases_density
#Bar Chart for Experience
Exp_bar <- Gaming %>%
  ggplot(aes(x = Experience, fill = Experience)) +
  geom_bar(colour = "black", linewidth = 1.2) +
  labs(title = "Experience Level", x = "Experience", y = "Frequency") +
  scale_y_continuous(limits = c(0, 500), expand = c(0, 0)) +
  scale_fill_manual(values = c("F" = "orangered2", "G" = "orange2", "H" = "orchid4", "K" = "olivedrab3", "L" = "darkslategray2"), labels = c("Novice", "Advanced Beginner", "Competent", "Proficient", "Expert")) +
  theme(panel.background = element_blank(),plot.background = element_blank(), panel.border = element_rect(colour = "black", fill = NA, linewidth = 1.5), axis.text = element_text(size = 14), axis.title = element_text(size = 14), axis.ticks.length = unit(0.3, "cm"), axis.ticks = element_line(size = 1, color = "black"), panel.grid.minor = element_blank(), panel.grid.major = element_line(colour = "grey", linetype = 2))
#View Bar
Exp_bar

#BIVARIATE ANALYSIS
#Hours by Game Type Boxplot
Hours_Type_boxplot <- Gaming %>% ggplot(aes(x = Type, y = Hours, fill = Type)) + geom_boxplot() + labs(title = "Hours Played Distribution by Game Type", x = "Type", y = "Hours") + scale_fill_manual(values = c("A" = "lightpink", "B" = "palegreen2")) + theme(panel.background = element_blank(), plot.background = element_blank(), panel.border = element_rect(colour = "black", fill = NA, linewidth = 1.5), axis.text = element_text(size = 14), axis.title = element_text(size = 14), axis.ticks.length = unit(0.3, "cm"), panel.grid.minor = element_blank(), panel.grid.major = element_line(colour = "grey", linetype = 2))
#View
Hours_Type_boxplot
#Hours by Game Type Violin
Hours_Type_violin <- Gaming %>% ggplot(aes(x = Type, y = Hours, fill = Type)) + geom_violin(trim = FALSE, alpha = 0.7) + geom_boxplot(width = 0.1, fill = "white", colour = "black") + labs(title = "Hours Played Distribution by Game Type", x = "Type", y = "Hours") + scale_fill_manual(values = c("A" = "lightpink", "B" = "palegreen2")) + theme(panel.background = element_blank(), plot.background = element_blank(), panel.border = element_rect(colour = "black", fill = NA, linewidth = 1.5), axis.text = element_text(size = 14), axis.title = element_text(size = 14), axis.ticks.length = unit(0.3, "cm"), panel.grid.minor = element_blank(), panel.grid.major = element_line(colour = "grey", linetype = 2))
#View Violin
Hours_Type_violin
```

```
#Time vs type
Time_Type_boxplot <- Gaming %>% ggplot(aes(x = Type, y = Time, fill = Type)) +
geom_boxplot() + labs(title = "Time Distribution by Game Type", x = "Type", y = "Time") +
scale_fill_manual(values = c("A" = "lightpink", "B" = "palegreen2")) +
theme(panel.background = element_blank(), plot.background = element_blank(),
panel.border = element_rect(colour = "black", fill = NA, linewidth = 1.5), axis.text =
element_text(size = 14), axis.title = element_text(size = 14), axis.ticks.length = unit(0.3,
"cm"), panel.grid.minor = element_blank(), panel.grid.major = element_line(colour = "grey",
linetype = 2))
#View
Time_Type_boxplot
#Purchases vs Type
#T-test: check for equal variance
t.test(Purchases ~ Type, Gaming, var.equal=TRUE)
#Correlation
ggpairs(Gaming)
#Comparing means for Hours and Type
mean_hours <- Gaming %>%
  group_by(Type) %>%
  summarise(mean_hours = mean(Hours, na.rm = TRUE))
#T-test for Hours by Type:
t_test_Hours_Type <- t.test(Hours ~ Type, data = Gaming, var.equal = FALSE)
#Comparing means for Purchases and Type
mean_purchases <- Gaming %>%
  group_by(Type) %>%
  summarise(mean_purchases = mean(Purchases, na.rm = TRUE))
#T-test for Purchases by Type
t_test_Purchases_Type <- t.test(Purchases ~ Type, data = Gaming, var.equal = FALSE)
#3 way scatter plot
ggplot(Gaming, aes(x = Purchases, y = Experience, color = Type)) + geom_jitter(alpha = 0.6,
size = 3, width = 30, height = 0.2) + scale_color_manual(values = c("A" = "lightpink", "B" =
"green3")) + theme_minimal() + labs(title = "Scatter Plot of Purchases by Experience Level
and Game Type", x = "Number of Purchases", y = "Experience Level", color = "Game
Type")
```

References:

Calleja, G. (2010). Digital games and escapism. *Games and Culture*, *5*(4), 335-353. Retrieved from: https://journals.sagepub.com/doi/abs/10.1177/1555412009360412