

Gaming Patterns in Higher Education

Author 1 and Author 2

2024-10-23

Header

Author Contributions

Author 1: Did the data cleaning and preliminary analysis, as well as contributed to questions #1 and #2, along with doing the advanced data analysis.

Author 2:

Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

Introduction

The data provided is from a survey done at UC Berkeley for students enrolled in a statistics course in 1994. Out of an eligible 314 students, 91 students submitted eligible responses which will be the data used in this EDA. The data focuses on student's play time of video games, likes and dislikes of video games, and more. One important thing to note is that responses that were left unanswered or improperly answered were coded as 99, and students who had never played video games or expressed a strong dislike for them were instructed to skip many questions. Additionally, the survey was distributed into two parts, where the first part asks about variable data, while the second part focuses on attitude questions.

Main Research Questions

1. Provide an estimate for the fraction of students who played a video game in the week prior to the survey.
2. Check to see how the amount of time spent playing video games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison?
3. Provide a point estimate and an interval estimate for the average amount of time spent playing video games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.

4. Consider the attitude questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many nonrespondents to the questions as to whether they think video games are educational, where they play video games, etc.
5. Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.
6. Further investigate the grade that students expect in the course. How does it match the target distribution used in grade assignment of 20% A's, 30% B's, 40% C's and 10% D's or lower? If the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture?

Outline

The remainder of the report will go through a basic analysis of the data, including our cleaning methods, basic analysis on various variables in the study, and more. Additionally, we will analysis the questions posed above, along with the conclusions that we came up with in our data. We will also pose an advanced analysis question based on the relationship between the gestation period of a pregnancy and see how the relationship between a child's birthweight and their mother's smoking-status are intertwined.

Basic Analysis

Data Processing and Summaries

Methods

In order to analyze our data, we wanted to make sure that our data was clean, and was able to have analysis performed on it. In order to do this, we looked through the data to identify if there were any inconsistencies that we wanted to catch and either fix or get rid of. One such inconsistency was included when we went to check the data types for the second part of the survey.

Through this, we noticed that while the first part of the survey all contained numeric types, which was to be expected, the second part of the survey contained two non-numeric columns, 'other' and 'other2', both of which were not supposed to be included in the second part of the survey. In order to clean the data up, we removed these two columns.

```
survey_p2 <- survey_p2[, !(names(survey_p2) %in% c("other", "other2"))]
all(sapply(survey_p2, is.numeric))
```

```
## [1] TRUE
```

Furthermore, in order to clean up the data and keep it consistent, we wanted to combine our data into one data frame. By doing this, we are using the assumption that each row in 'survey_p1' is the same person that was in the same row in 'survey_p2', meaning that they are the same person. Additionally, we also got rid of rows that contained NA values, just to make sure we got rid of inconsistent data, because that could

mess up our survey. From this, we noticed that we have 87 rows of data left, which means that 4 of our rows had NA values.

```
combined_table <- cbind(survey_p1, survey_p2)
cleaned_df <- na.omit(combined_table)
dim(cleaned_df)
```

```
## [1] 87 34
```

We also noted down the types of variables that were present in our data, since many were binary, ordinal, nominal, and more. We also graphed out histograms of the variables in the two surveys to see if we could identify any outliers, but since responses that were left unanswered or were improperly answered were coded as 99, there were no such outliers. Therefore, we went with the assumption that there were no outliers in the dataset, since they were all coded as 99 and could be filtered out later.

Question 1: Estimate of Students who played a Video Game

Methods

In order to provide a solid estimation of the amount of people who played a video game in the week prior to the survey, let's filter the amount of people who had over 0 hours last week as well as filter out those who didn't respond correctly or did not answer the question (99)

```
## [1] "Point Estimate: 0.3678"
```

Furthermore, in order to get a more concrete range for the true measurement of this estimate, let's take a 95% confidence interval and get an interval estimate for the true population proportion who played a video game during an exam week.

```
## [1] "95% Confidence Interval: [0.2665, 0.4691]"
```

Analysis

Our examination of video game engagement among UC Berkeley Statistics students during the Fall 1994 semester revealed that 36.78% of surveyed students reported playing video games in the week following their second examination. The 95% confidence interval for this proportion ranges from 26.65% to 46.91%, indicating the true population parameter lies within this range with 95% confidence.

Key analytical findings include:

- The point estimate of 36.78% represents a substantial portion of the student population engaging in gaming activities in an academically intense period.
- The confidence interval width of ~20% suggests a high variability amongst students in terms of gaming behavior.
- The timing of the survey, during an exam week, provides context for interpreting these gaming patterns.

Conclusions

Based on our analysis, we can draw several significant conclusions:

- Video gaming was a common activity amongst Statistics students, even in a academically intense time. Our confidence interval shows us that at minimum, 25% of the student population was engaged in gaming activities during the survey period.
- Video games were played by a significant amount of people in the Statistics course, yet our wide confidence interval suggests that there is a lot of variability amongst the people in the course, possibly suggesting that different populations of students would have a different estimate about the percentage of students who played video games regularly.
- The percentage found from the data indicates that gaming was a prevalent hobby even in the mid-1990s, and can perhaps provide insight into student gaming behavior into the future, as perhaps it has become even more significant as time has passed and gaming has become easier to access for everyone.

Question 2

Methods

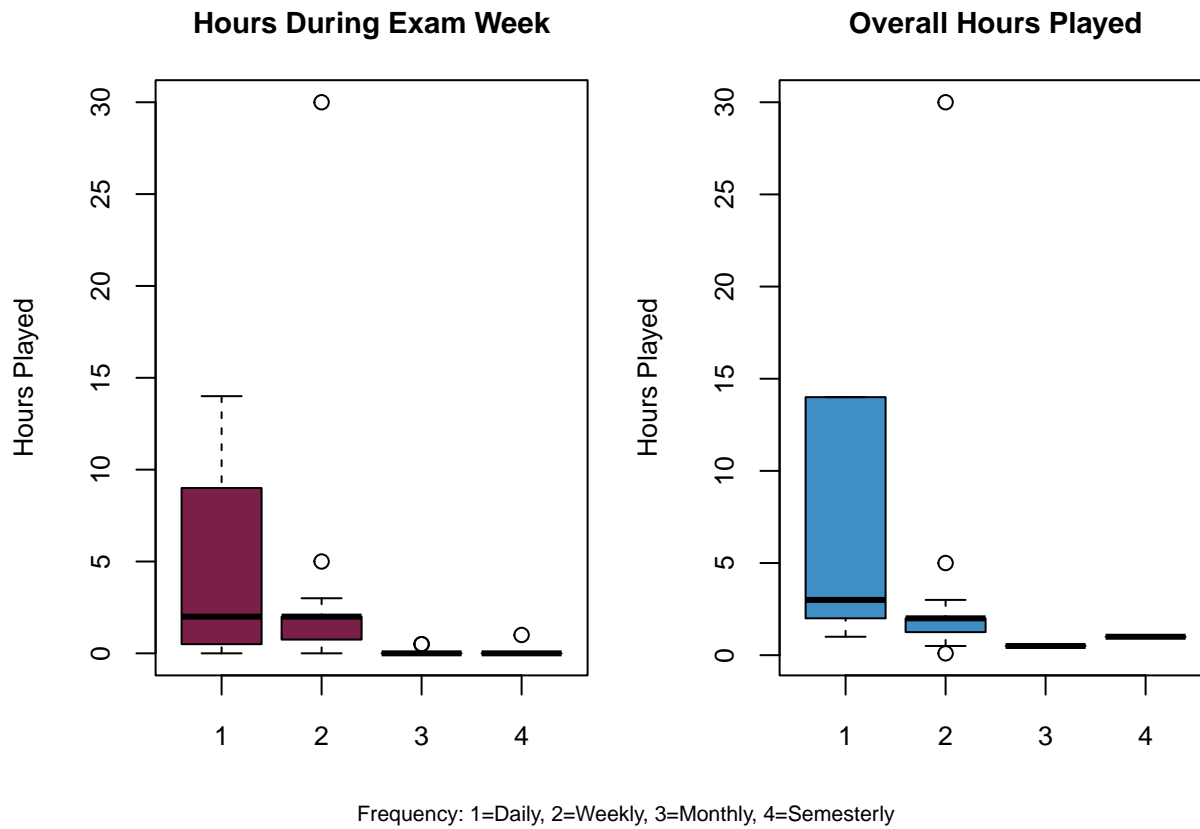
In order to compare the amount of time spent playing video games during the exam week versus a normal week, let's analyze what we need to do. First, we should find the amount of time the students spent playing video games, which we can do by creating a summary of the 'time' column in our cleaned dataframe.

```
## frequency avg_time median_time count
## 1          1 4.62500000          2      8
## 2          2 2.61481481          2     27
## 3          3 0.06250000          0     16
## 4          4 0.04347826          0     23
```

In order to filter for the fact that the week was an exam week, let's make the assumption that many students that were studying for the exam were not playing games that week. In order to account for this, we filtered out students who played 0 hours of video games in the exam week. The reason we made this assumption is that it would mostly affect those that normally play daily or weekly, as compared to monthly or semesterly. If our assumption is that those who play monthly and semesterly would probably not even play, or play very little during an exam week, their distributions should remain relatively similar.

```
## frequency avg_time median_time count
## 1          1 6.166667          3.0      6
## 2          2 3.069565          2.0     23
## 3          3 0.500000          0.5      2
## 4          4 1.000000          1.0      1
```

Now, let's use a boxplot to compare the two distributions, between students playing video games during exam week and those who didn't.



Analysis

Looking through these two boxplots and their data summaries, there is an obvious difference, especially for students with a daily or weekly frequency of play. For instance, our adjusted comparison shows that those who played ('freq' = 1) daily during an exam week typically had a lower amount of hours played, with a 1.5 hour difference between our exam week data and our filtered data. Similarly, our boxplot also shows a difference in the interquartile range between the two datasets for those who played daily, as during a normal week, students were more likely to play more and had a wider variety of playtimes, which can be inferred from the two box plots. A similar comparison can be made from students who played weekly ('freq' = 2). There's a ~0.4 hour increase in normal weeks versus the exam week. Another interesting observation is that the adjusted data seems to have a smaller interquartile range compared to its original exam week data, somewhat implying that those who play weekly usually play a more set amount of hours that usually doesn't diverge. Frequencies 3 and 4 don't have a large difference, just due to the nature of their descriptions, since most students in these categories would probably play very little to none already. If someone only plays games monthly or semesterly, then they are less likely to have major differences during an exam week versus a normal week.

Conclusions

Overall, the fact that there was an exam in the week prior to the survey tells us this:

- On average, students played less hours than they normally would on a week without an exam.
- Those who reported that they play daily or weekly typically have the largest difference between their average hours played during an exam versus a normal week, with around a ~1.5 hour difference for those who play daily and a ~0.45 hour difference with those who play weekly.

- Those who reported that they play monthly or semesterly have small or negligible differences between their average hours played due to the nature and scarcity of their data, along with the assumption that we made when filtering.

Question 3

Methods

```
# Your R code for methods related to Question 3
```

Analysis

```
# Your R code for analysis related to Question 3
```

Conclusions

Your conclusions for Question 3.

Question 4

Methods

```
# Your R code for methods related to Question 4
```

Analysis

```
# Your R code for analysis related to Question 4
```

Conclusions

Your conclusions for Question 4.

Advanced Analysis

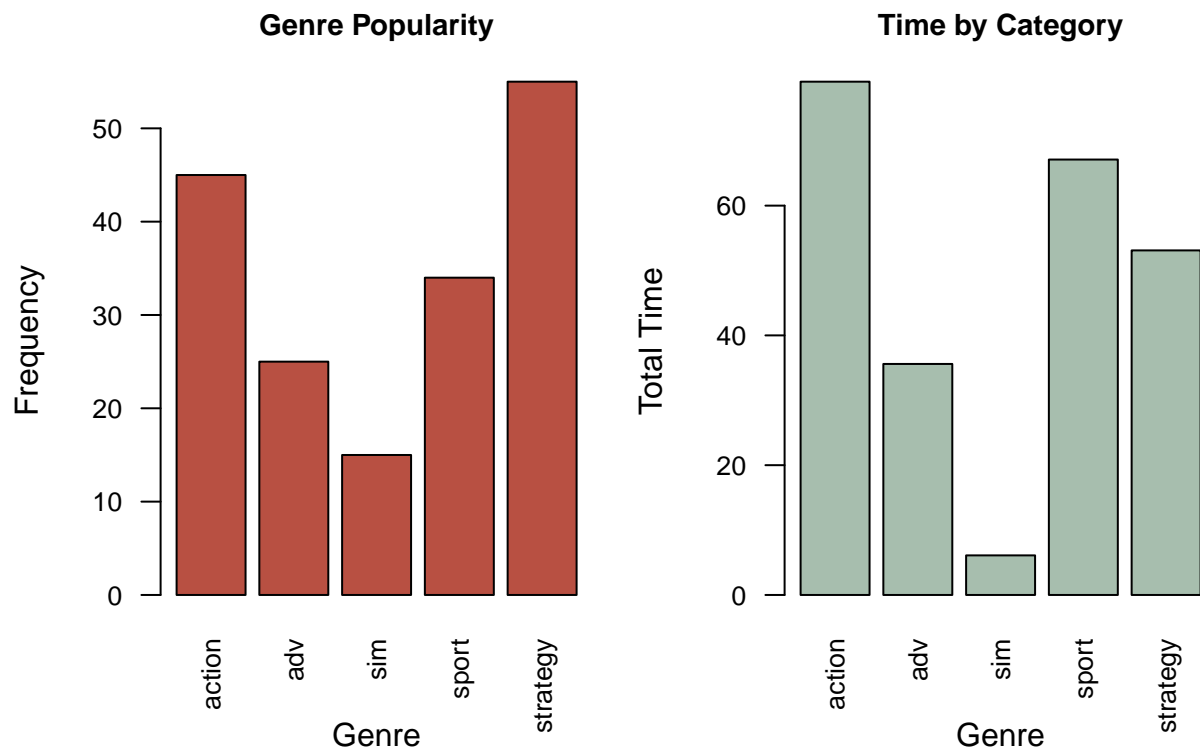
Our advanced analysis focuses on the intersectionality between people that like video games and which types of video games they prefer. By doing this, we can see the most popular types of video games that these students prefer, along with the average amount of time they play video games a week, potentially showing us which games people spent the most time on.

Methods

In order to do this, we filtered the data. It's important to note that we ignored people who skipped this question, because those aren't the people we want to analyze. We went with the assumption that those who put all 0s in each of the columns that corresponded to the game type, because presumably if someone didn't like games or haven't played it, they wouldn't know the genres and skip.

```
## [1] 75 34
```

From this, we can see that we have 75 respondents that still enjoy games, which we can further filter out. WE can do this by capturing the most popular video game genres from this data, while also comparing the amount of time people spent playing those games.



Analysis

Our two graphs, especially when put next to each other, show us another interesting detail. Although strategy was the most popular game genre chosen by people, with the action and sport genres falling slightly behind, when we look through the total time spent playing each game, action was the highest, with sport and then strategy falling behind. This means that there's a decent chance that the people who play strategy games don't typically play the game for a longer time, while those who played action and sports games play for longer durations. On another note, sim was the least popular genre of game in the survey, and that pattern still shows in the total time spent. People who mentioned they like simulation games spent very little time actually playing them, leading us to believe that those who play these simulation games either don't play them very often or play them for short periods of time compared to their peers who played action or sport games.

Conclusions

Our conclusions for this advanced analysis show us that although a video game genre like ‘strategy’ could perhaps be the most accessible, it’s not necessarily the most played time-wise. While ‘action’ and ‘sport’ genres were not as popular as ‘strategy’ games in terms of total picks, the students who played these games far exceeded the play-time of students who played ‘strategy’ games.

It’s also important to note that this isn’t conclusive. Firstly, there’s inherent bias because this survey was taken during an exam week, likely leading to students playing less than they normally would. Additionally, the way we filtered our data may be potentially flawed, as students were allowed to choose up to 3 of their favorite genres, not just one. This means that our graph of total time spent on each category is layered and can be duplicated, because if a student likes to play multiple genres of games, they are stacked on top of their respective genres, potentially leading to more visible and exaggerated graphs than what would normally be. However, even with that discrepancy, we still believe that these two graphs show us concrete differences in many of the genres and how students viewed the genres differently, both in terms of popularity as well as personal enjoyment.

Conclusions and Discussion

Summary of Findings

Reprise the questions and goals of the analysis stated in the introduction. Summarize the findings and compare them to the original goals.

Discussion

Additional observations or details gleaned from the analysis section. Discuss relevance to the science and other studies, if applicable. Address data limitations. Raise new questions and suggest future work.

Appendix

Include any additional technical details, tables, or figures that support your analysis but would disrupt the flow of the main text if included earlier.

Additional R code or output can be included here