NUS
National University
of Singapore

GEA1000 Group Project Report

TD39 Group 1

Name	**Matriculation No.**
Ang Chun Juay	A0253008N
Chan Wei Ning	A0259497E
Jeong Harin	A0224097A
Ng Yixuan	A0259104J
Wang Helin	A0258383U

Part A

Section 1

1. The aim of the study is to find out the effects of gaming on youths' health by examining the prevalence and clinical correlates of gaming, reported problems associated with video games, and the prevalence and correlates of problematic gaming via a survey data.

2. The main finding of the study is "The prevalence of problematic gaming is low but not insignificant, and problematic gaming may be contained within a larger spectrum of externalising behaviours."

3. The subjects studied were 4028 High School Students from various high schools in Connecticut aged 14-18.

4. The target population of the study is Adolescents.

Section 2ii (Observational Study)

5a. The *main exposure variable* is Time Spent playing video or computer games in a typical week. This was assessed through self-reported time spent on games. Respondents who reported "none" were classified as non-game players. The remaining respondents who were classified as game players were categorised based on the frequency of the play: less than 7, 7 to 14, 15 to 20, and 21 or more hours per week.

The *main response variable* is Health correlated problematic gaming behaviours. It includes numerous measuring factors such as smoking, use of marijuana, drug usage, alcohol consumption frequency, depression, and aggression.

The exact date when these variables were assessed was not noted in the study. Yet, as this paper was accepted for publication on August 12, 2010, it can be inferred that the study was conducted before 2010 or in early 2010.

5b. Potential confounders that were controlled include 'Gender' and 'Family Structure'.

Gender: Gender is a potential confounder because it is associated with time spent gaming as among the 4028 adolescents surveyed, 76.3% of the boys and 29.2% of the girls reported gaming. Thus, male adolescents are positively associated with gaming while female adolescents are negatively associated with gaming. Therefore, gender is associated with gaming. Females may be less likely to smoke as compared to males because researchers have shown that smoking can lead to more painful and irregular periods, thus females are negatively associated with smoking. Hence, this shows that gender is associated with health. Therefore, gender is a potential confounder.

Family Structure: For the total sample, respondents with 2 parents have a higher rate for 'played video games ever,' as 52.32% than 1 parent of 48.02%. Hence, family structure is associated with time spent gaming. For health, children with 2 parents may have a lower risk of depression or drug usage as they may receive more sufficient love and concern from their parents as compared to children with only

1 parent, thus children with 2 parents may be positively associated with health. Therefore, family structure is a potential confounder.

Section 3

6. The non-probability sampling method was employed. This includes:

1) Convenience sampling. The researchers have only recruited schools in the state of Connecticut, where the researchers are based. Therefore, they have chosen subjects to form a sample among those that are most easily available to participate in the study. Furthermore, implying that other demographics of the population such as high school students from other states would be left out.

2) Volunteer sampling. Students that completed the survey have volunteered themselves into the sample as they were allowed to refuse to complete the survey.

We are unable to calculate the response rate because the researchers did not report the number of students who did not respond to their invitations. To be able to calculate the response rate, we would need to calculate the number of responses divided by the total number of invitations sent.

7. The null hypothesis is that there is no association between problematic gaming and the development of health problems. At a 5% level of significance, for 'Sad or hopeless ≥ 2wk', the p-value is less than 0.0001. Since the p-value is less than the level of significance, we have sufficient evidence to reject the null hypothesis.

8. In order to address potential confounders, data of other suspected variables have to be collected so that we are able to investigate the association between the suspected variable to both the independent and dependent variables through the slicing of data. In this case, the independent variable would be the hours spent playing games and the dependent variable would be health-correlated problematic gaming behaviours. We acknowledge that the researchers collected extensive amounts of data from the respondents but it is still possible that there is another potential confounder which is not controlled for in this study where data is not collected as well.

One such potential confounding variable would be the genre of games played by adolescents. This could be a potential confounder as playing different genres of games would lead to different health outcomes, for example, youths who played more physical games such as Wii Sports could be healthier compared to those who only played online games which do not involve significant physical movements such as first person shooter (FPS) games. It can also affect hours played as typically, adolescents who play FPS will play for long hours.

9. There are sources of potential bias, namely selection bias and non-response bias which arise from the non-probability sampling method used in this study. Addressing selection bias, researchers were biased on the selection of units in this study, this can be seen where the researchers recruited various high schools in the state of Connecticut but the population interested were adolescents. This leads to an imperfect sampling frame as adolescents from other states in the US are left out.

In the initial attempt of sampling, the researchers acknowledged that there are insufficient responses and to reduce non-response bias, it led to targeted contacts of high schools but this, unfortunately, resulted in increased selection bias.

Another non-response bias is where the researchers allowed students to voluntarily refuse to complete the survey and this could lead to a scenario where those who opt out of the survey have significantly different health conditions to those who opted in for the survey, especially since this study focuses on health of the subjects where they might feel uncomfortable sharing information.

10. The findings cannot be generalised to the target population who are adolescents (age 10 to 19 as defined by WHO). To confidently generalise the findings to the target population, it is important that a probability sampling method is used, there is a good sampling frame and high response rate. Firstly, the sampling frame was not equal or larger than the population of interest as only specific high schools in Connecticut were chosen which were further cut down to targeted contacts to get a higher response rate. Secondly, the convenience sampling method is also a form of non-probability sampling method which led to a biased sample as no randomisation was done.

Due to the nature of non-probability sampling, we are unable to generalise the finding to a subpopulation which are the high school students in Connecticut. This is because there are other possible confounding factors that can influence the study between the variables of interest. At the same time, high school students from other schools in Connecticut are left out due to selection bias. Thus, we are unable to generalise to the subpopulation.

Part B

1. Without any data cleaning, there are 1500 points in the data set. Two categorical variables we chose are BORN and FINRELA, and two numerical variables we chose are CONRINC and EDUC.

Summary Statistics for BORN and FINRELA

In finding Summary Statistics for these categorical data, we used Excel. Using Home > Filter, we selected relevant variables and calculated proportions based on the number of data set points of 1500.

[*For BORN (Born in USA or Not)]*

Rate(Born in USA): 1086/1500 = 0.724

Rate(Not Born in USA): 164/1500 = 9.109

[*For FINRELA (Compared with American families in general, Respondent's Family Income level)]*

Rate(far above average): 31/1500 = 0.0207

Rate(above average): 341/ 1500 = 0.227

Rate(average): 539/1500 = 0.359

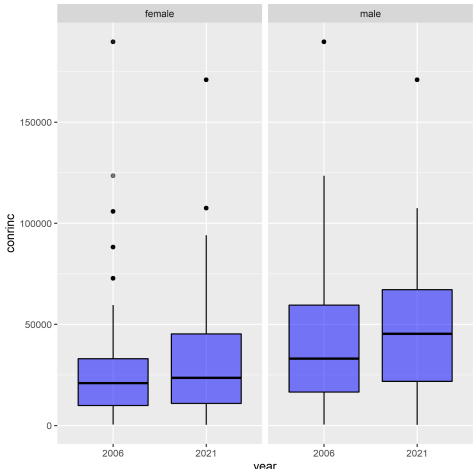
Rate(below average): 278/1500 = 0.185

Rate(far below average): 62/1500 = 0.0413

Summary Statistics for CONRINC and EDUC

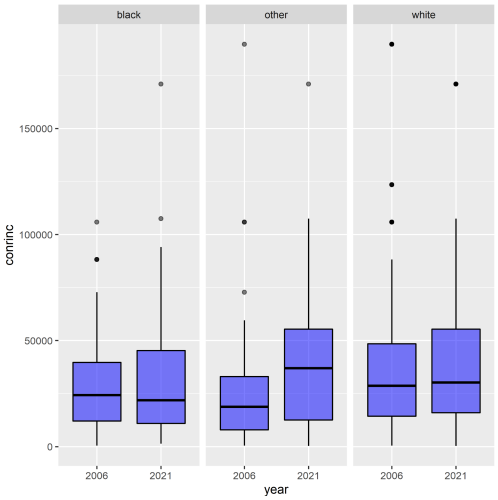
variable	mean	median	sd	p25	p75	IQR	min	max
All	All	All	All	All	All	All	All	All
conrinc	38,537.911	28,668	37,106.988	14,280	48,516	34,236	336	189,740
educ	14.450	14	2.926	12	16	4	0	20

2. The respondents' change in income between the years 2006 and 2021, across gender, race, educational level, and whether one was born in the US is illustrated below using bar plots. We used different filtered data sets for each variable. For each variable, we filtered the missing data entry, eg. for gender > Transform > Select variable: 'gender', Transformation type: remove missing values > press 'enter' > Store changes in: 'Project_B_filtered_gender'. We created the box plot by Visualise > plot type: box plot, Y-var: 'conirc{numerical}', X-var: 'year{factor}', Facet column: 'sex{factor}' > Create plot.The same is done for the rest of the variables.



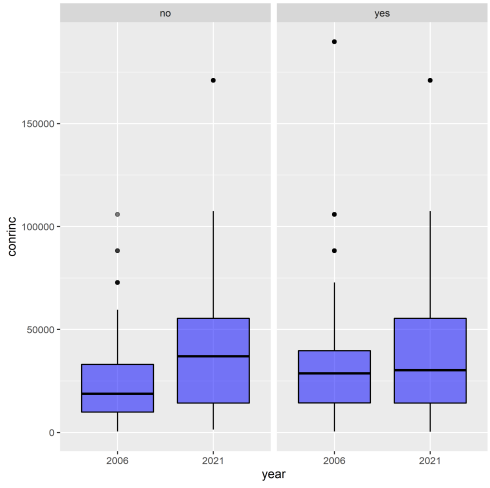
Gender

Between 2006 and 2021, the Interquartile Range(IQR) and median income have increased for both males and females. Both genders also displayed a decrease in the number of outliers. For females, the distribution shifted from relatively normal distribution to positively skewed. For Males, the distribution shifted from a positively skewed to a relatively negatively skewed distribution as the median increased.



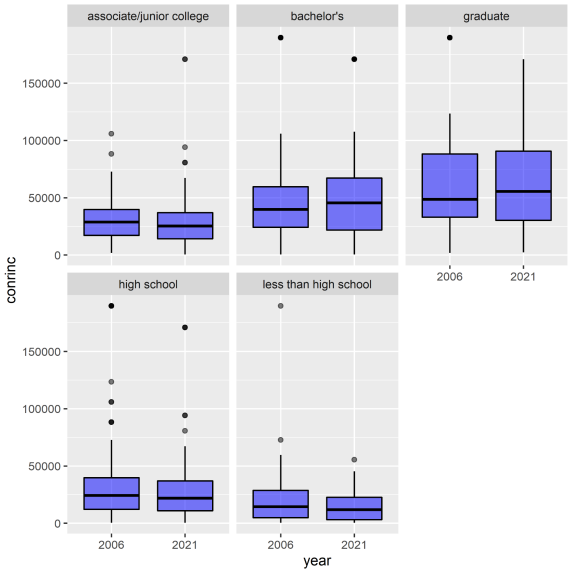
Race

Between 2006 and 2021, the median income and Q1 increased for whites and others but decreased for blacks. However, the IQR for all of them increased, suggesting an increase in the dispersion of individuals' income from 2006 to 2021. Furthermore, the number of outliers decreased for both other and white, while it remained the same for black. The distribution remained relatively positively skewed for both 2006 and 2021 for black and white. Meanwhile, the distribution shifted to negatively skewed in 2021 for others.



Born in the US

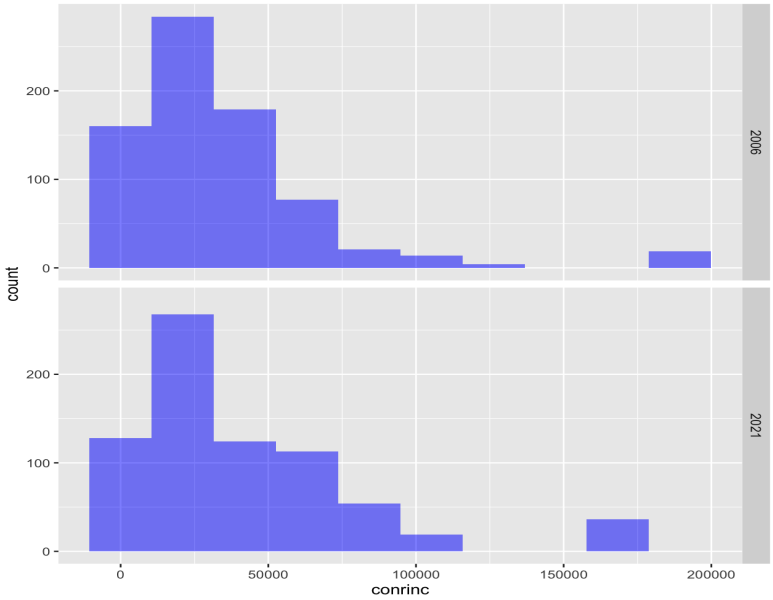
Between 2006 and 2021, the median income and IQR for both respondents born in the US and not born in the US increased. The number of outliers decreased for those born in the US and those not born in the US. The distribution shifted from relatively positively skewed to negatively skewed for respondents not born in the US as data became more dispersed.For the respondents born in the US, the distribution shifted from relatively negatively skewed to positively skewed.



Education Level

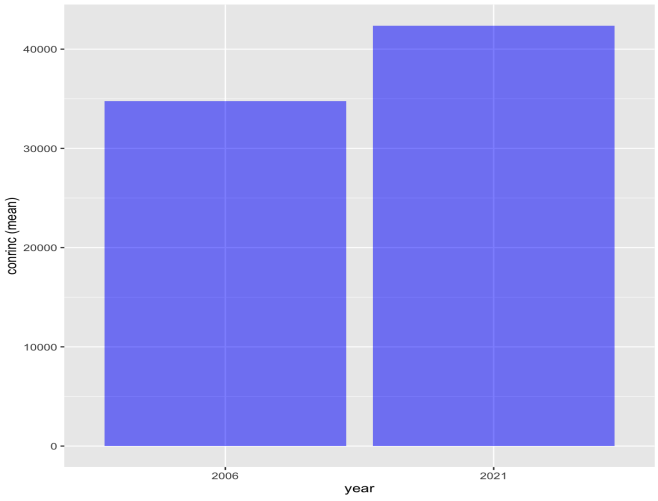
Between 2006 and 2021, the median income for associate/junior college, high school, and less than high school decreased, but the median income increased for bachelor's and graduates. The IQR decreased for all education levels except high school and less than high school, which displayed decreased IQR and dispersion. Additionally, the distribution remained relatively the same across education levels, except for that of bachelor's which shifted from positively skewed to negatively skewed distribution.

3. We used Bar Plots and Histograms as 2 other visualisation tools to make two comparisons between the years 2006 and 2021.
Visualise > Dataset: 'GSS_group1_1_' > Distribution/Bar > For Distribution: X-var: 'conrinc{numerical}' and Facet row: year{factor}, For Bar: Y-var: 'conirc{numerical}', X-var: 'year{factor}' > Create plot to obtain the histogram and bar plot respectively.



Histogram

Between 2006 and 2021, the maximum income is different as seen from the histogram above whereby the maximum income decreased.
However, the histograms in the 2 years are similar in the sense that both are right-skewed.



Bar Plot

Between 2006 and 2021, we can see that the mean incomes are different as seen from the bar chart above whereby the mean income increased from around 35,000 to around 42,000.

Both the mean income for 2006 and 2021 are above the overall median income of 28,668, suggesting that the data has positively skewed distribution.

(Median income can be found on Radiant: Manage > Dataset: 'GSS_group1_1_' > Display: 'summary')

4.

	Happy		
Income	**unhappy**	**happy**	**Total**
All	All	All	All
high	89	533	**622**
low	133	494	**627**
Total	**222**	**1,027**	**1,249**

To create this table in Radiant, go to Transform > Select the cleaned data set for the variable happy (used the steps mentioned in Q2 to filter missing values in variable "Happy") > Transformation Type: 'Create' > Type: Happy = happy == "pretty happy" | happy == "very happy" under 'Create:' > Store changes as: Q4_newdataset > repeat these steps for income but type: Income = 'conrinc > 28,668' instead. For the table, go to Pivot > select the new data set created > Categorical variables: 'Happy{factor}' and 'Income{factor}' > Click 'Create pivot table' to generate the table as shown above.

Rate(Happy|High income) = 533/622 = 0.857

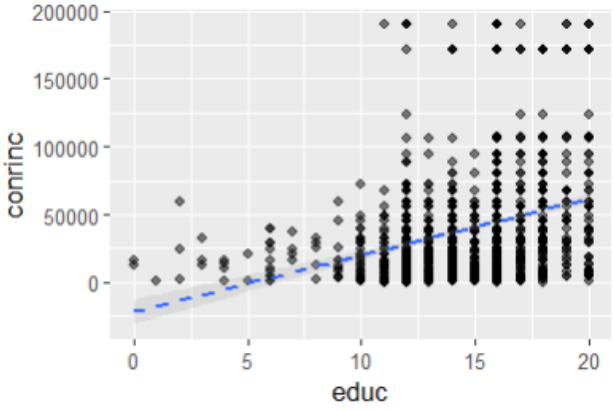
Rate(Happy|Low income) = 494/627 = 0.788

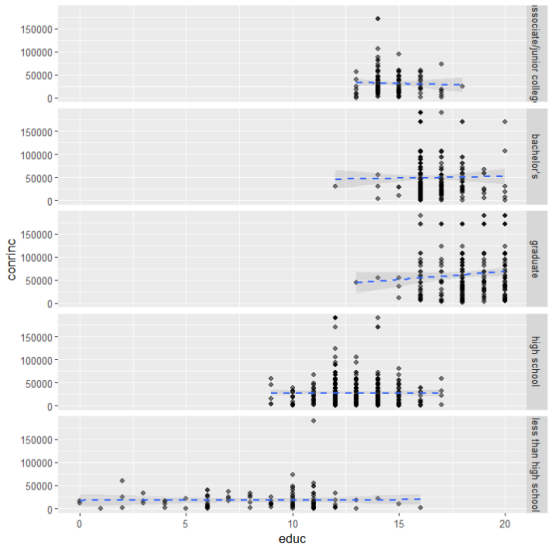
Since rate(Happy|High income) is larger than rate(Happy|Low income), there is a positive association between being happy and earning a high income.

5. On Radiant, using the steps mentioned for filtering in Q2, we filtered the missing values for "educ{numerical}" and stored it as a new dataset "gss_cleanededuc". Using this dataset, at the "Model" tab, select "Linear regression (OLS)" > Response variable: 'conrinc{numerical}', Explanatory variable:

'educ{numerical}' > Estimate Model to get the equation and to retrieve the scatter plot, select "Plot", "Scatter" and "All" at the dropdown menu, click "Line" to show the blue line. Results are shown below:

conrinc = -21724.042 + 4173.905 (education) - Overall





We can see that the respondent's income is positively associated with the number of years of education they had. This is because there is a positive gradient (ie. upward slope) between the 2 variables.

To split the data into subgroups, on Radiant, with dataset "gss_cleanededuc", at "Data" Tab, select the variables "conrinc{numerical}", "educ{numerical}" and "degree{factor}" with Ctrl + Click. At "View", under "Degree", select the degree subgroup you are interested in such as "associate junior college" then store it as a new filtered dataset "associate junior college". Repeat the steps above for the equation and graph. Repeat for the rest of the subgroups. We then get:

1. conrinc = 46681.239 - 1058.047(education) - Associate junior college
2. conrinc = 36363.448 + 788.542 (education) - Bachelor
3. conrinc = -1422.708 + 3499.432 (education) - Graduate
4. conrinc = 29064.229 - 91.028 (education) - High school
5. conrinc = 17540.916 + 113.058 (education) - Less than high school

We can see that the majority of the degree subgroups, namely, bachelor, graduate and less than high school, have positive associations between respondent's income and the number of years of education which is in line with the overall positive association above. Only 2 degree subgroups have negative association and they are associate junior college and high school. Thus, there is no Simpson's Paradox observed here.

6. Chi-square Test was conducted at a 5% level of significance to determine if there is a significant association between being born in the US and having a self-perceived "above average" family income level. Using steps mentioned in Q2, we filtered out the 249 missing data entries from 'finrela{factor}' and 250 missing data entries from 'born{factor}'. We performed the chi-squared test and calculated the p-value by > Basics > Cross-tabs > Categorical variable: 'finrela{factor}' and 'born{factor}', tick the boxes for 'Observed', 'Expected' and 'Chi-squared' > p-value is found in results under chi-squared: … p-value 0.871.

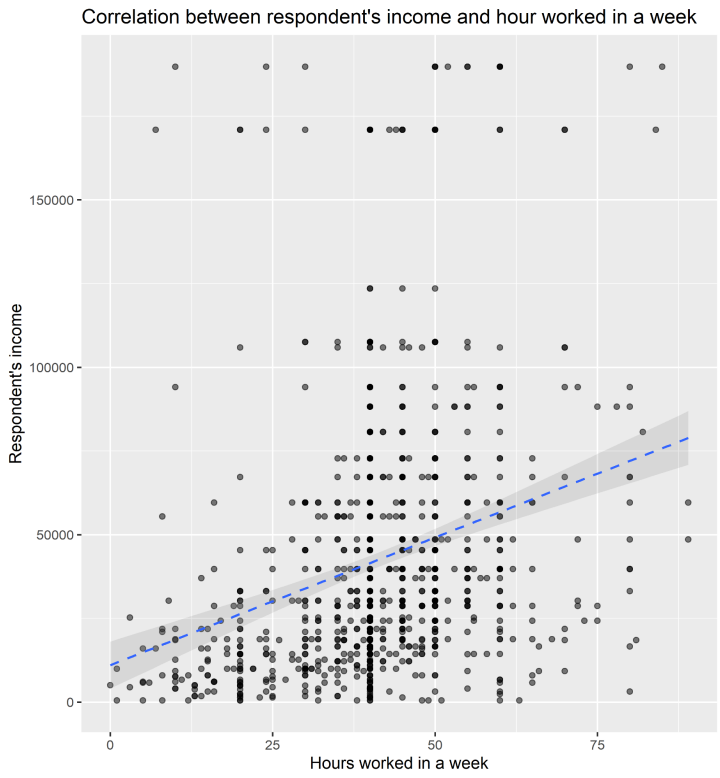
Assumptions: We made several assumptions in conducting the Chi-square Test. We assumed that the sample taken from the population is random, all observations are independent, and the different categories of the variables are mutually exclusive.

Null hypothesis: H0: There is no association between being born in the US and having a self-perceived "above-average" family income level.
Alternative hypothesis: H1: There is a significant association between being born in the US and having a self-perceived "above average" family income level.
Significance level: 0.05 [taken from 5% level of significance]
P-value: 0.871

```
Cross-tabs
Data     : GSS_group1_cleanedbornandfinrela
Variables: born, finrela
Null hyp.: there is no association between born and finrela
Alt. hyp.: there is an association between born and finrela

Contribution to chi-squared: (o - e)^2 / e
       finrela
born    perceived above average perceived average and below Total
  no    0.02                    0.01                         0.02
  yes   0.00                    0.00                         0.00
  Total 0.02                    0.01                         0.03

Chi-squared: 0.026 df(1), p.value 0.871

0.0 % of cells have expected values below 5
```

Based on the observation in the sample (P-value 0.871 > 0.05 Level of Significance), there is insufficient evidence to support the alternative hypothesis, at the significant level that we have set our test.

7. Additional analysis that we have conducted aimed to find the association between respondents' individual income and hours worked in a week. Using steps mentioned in Q2, we filtered out the missing 339 data entries from the 'hrs1' and calculated correlation coefficient by > Basics > Correlation > Select variable: 'conrinc{numerical}' and 'hrs1{numerical}' > Calculate correlation > Correlation coefficient will be under Correlation matrix. We created the scatter plot by > Visualise > Plot type: Scatter Plot > Y-var: 'conrinc{numerical}', X-var: 'hrs1{numerical}', Number of data points plotted: all > Create plot.
Correlation coefficient between respondent's income and hours worked in a week: 0.27
There is a weak positive linear association between respondent's income and hours worked in a week.




```
Correlation
Data        : Project_B_Q7_remove_age
Method      : Pearson
Variables   : conrinc, hrs1
Null hyp.   : variables x and y are not correlated
Alt. hyp.   : variables x and y are correlated

Correlation matrix:
     conrinc
hrs1 0.27

p.values:
     conrinc
hrs1 0.00
```