Analysis of General Social Survey 2021 with a focus on resident's time spent online*

Age, Education level and Quality of life bring huge influence on Internet access time

Xueru Ma

01 May 2022

Abstract

With the extensive application of the Internet in various fields, the time spent on the Internet by the residents has shown a constant increase. We analyzed the relationship between time spent online and age, income, education level, gender, email hours, mental health, physical health, and quality of life. The dataset we used was obtained from the General Social Servey 2021, selecting some of the variables of interest, and we identified that respondents with lower age, weaker education level and poorer quality of life evaluations exhibited a greater tendency to spend more time online. Our analysis provides a guideline for future regulations of the Internet access time of residents.

keywords: time spent online, income,sex,email hours, age, education level, mental health, physical health, quality of life

Contents

1 Introduction **3**

2 Data **4**

2.1 Data Source 4

2.2 Methodlogy 4

2.3 Data Description 5

2.4 Strength 6

2.5 Weaknesses 6

2.6 Key Features 7

3 Model **16**

3.1 Linear Regression Model 16

4 Data Simulation **19**

*Code and data are available at: https://github.com/maxueru001023/sta304_Final_paper.

5 Result **21**

5.1 Time spent online vs. Age . . . 21

5.2 Time spent online vs. Sex . . . 22

5.3 Time spent online vs. Age Group . . . 23

5.4 Time spent online vs. Quality of life . . . 24

5.5 Time spent online vs. Physical Health . . . 25

5.6 Time spent online vs. Mental Health . . . 26

5.7 Time spent online vs. Education level . . . 27

5.8 Time spent online vs. Quality of life . . . 28

6 Discussion **29**

6.1 Age vs. Web hours . . . 29

6.2 Mental health vs. Web hours . . . 29

6.3 Life quality vs. Web hours . . . 29

6.4 Limitation . . . 30

6.5 Future improvements . . . 30

7 Appendix **31**

7.1 Respondent's age and education years distribution . . . 31

7.2 wwwhr vs. sex . . . 32

7.3 wwwhr vs. quallife . . . 33

7.4 wwwhr vs.hlthphys . . . 34

7.5 wwwhr vs. hlthmntl . . . 35

7.6 wwwhr vs. educ_level . . . 36

7.7 residual plots . . . 37

Bibliography **39**

1 Introduction

The Internet has invariably revolutionized cities, countries, and the entire world. As the Internet is adopted in more and more industries, people can find its traces almost everywhere in their life, work and studies. The Internet of Things integrates a large number of technologies and envisions a variety of things or objects around us that, through unique addressing schemes and standard communication protocols, are able to interact with each others and cooperate with their neighbors to reach common goals (Iera 2012). In life, people tend to choose the online social networking mode and communicate with their relatives, friends and colleagues through the social media; online shopping provides great convenience for people's consumption, residents only need to place orders on electronic devices and their expected products could be delivered to their doorsteps; online games provide relief from the stress in people's life and have become an indispensable entertainment in morden society. At work, people use the Internet to edit and transmit documents efficiently; they hold online meetings with colleagues to discuss the considerations in their work. In studies, professors send announcements and study materials to students through the Internet, and students learn through electronic textbooks rather than the real textbooks. As the Internet has boomed in different aspects, the amount of time people spend online has soared.The latest Digital 2019 report shows that people're spending on average of 6 hours and 42 minutes online each day (Hughes 2019). It is identical to say that more than 100 days are spent online every year for each Internet user. That even accounts for more than 27 percent of time of a year. In 2021, the number of people using the Internet worldwide reached 4.66 billion, which is an increase of 316 million people from the same period last year. The global Internet penetration rate is 59.5%. Currently, there are 4.2 billion social media users across the world. These staggering statistics show the enormous amount of Internet users and the significant average time spent online by Internet users.

The data was collected from December 2020 to May 2021 by National Organization Research Office (NORC) at the University of Chicago.The General Social Survey 2021 had a sample size of 27,591 adults who were 18 or older in the United States as well as living in noninstitutional housing at the time of interviewing.

The residents in the United States pay a lot of attention to the Internet. Almost everyone spends a certain amount of time online each week. According to the survey released by the market research firm Forrester Research, the amount of time Americans spend online has increased by 121 percent over the past five years. Adults under 30 years old spend about 12 hours per week online; and the people over 66 years old spend about eight hours per week online. Forty percent of U.S. Internet users' time is spent on three activities, namely social networking sites, gaming and email. This phenomenon has also led to a steady decline in the percentage of time spent online by other Internet activities. In terms of Internet assess time, significant differences were detected among respondents in terms of different genders. Residents with higher education level tends to use the Internet in a more reasonable time interval. The younger respondents seem to be more addicted to the Internet, and they will spend more time using the Internet than the older individuals.

This paper explores the distribution of time spent online by respondents and attempts to relate it to eight factors, including gender, age, income, education, length of email, quality of life, physical health, and mental health, to analyze the time spent time spent online of Americans who were 18 or older and lived in noninstitutional housing in 2021. Factors were chosen in consideration of literature and also what studies have repeatedly highlighted to be affecting the amount of time spent online in the real life: low family income, poor self-control ability due to younger age, fairly low education level, worse evaluations in terms of meantal health and physical health, are the most backward issues influencing the accessed time online. In addition, based on the existing data of Internet access time, we also predicted the average distribution of Internet access time for the whole country, utilizing the above factors to analyze the tendency of Internet use by macro-national residents.

The remainder of the whole paper is divided into six parts. Section 2 illustrates the data source and the methodology conducted in the analysis, as well as the main characteristics of our selected data. Section 3 demonstrates the linear regression model between time spent online and predictors. The fourth part is about data simulation and the exploration on confidence interval. Section five shows the relationship between time spent online and other factors that we obtained through methodology. Section 6 discusses the results we found as well as the weaknesses and limitations of this paper. The last part is the appendix, in which we show the supplementary figures and tables for our paper.

2 Data

2.1 Data Source

This paper applies the time spent online of the residents obtained from the General Social Survey 2021(GSS 2021). The GSS has been administered by NORC at the University of Chicago (NORC) and funded by the National Science Foundation (NSF) since its inception. Currently, the GSS is designed by a set of Primary Investigators (PIs), with input from the GSS Board, comprised of notable researchers within the scientific community.

R studio (R Core Team 2020)is used to process the code and complete the analysis. The package 'tidyverse' (Wickham et al. 2019) is then used to gather the data. The package 'knitr' (Y. Xie 2021)enables integration of R code into LaTeX, allowing reproducible research in R through the means of literate programming. The package 'car' (Fox and Weisberg 2019) compliments applied regression techniques by providing numerous functions that perform tests, creates visualizations, and transform data. The package 'dplyr' (Wickham et al. 2022) provides a consistent set of verbs that help us solve the most common data manipulation challenges. The package 'haven' (Wickham and Miller 2021) enables R to read and write various data formats used by other statistical packages. The package 'tidyr' (Wickham and Girlich 2022) makes it easy to "tidy" our data.

The package 'ggplot' (Wickham 2016) is used for making the figures and graphs for the analysis. The package 'patchwork' (Pedersen 2020) is applied to make plot composition in R extremely simple and powerful. The package 'kableExtra' (Zhu 2021) is used to extend the basic functionality of tables produced using knitr::kable(). The package 'hrbrthemes' (Rudis 2020) offers extra 'ggplot2' themes, scales and utilities, including a spell check function for plot label fields and an overall emphasis on typography. The package 'viridis' (Garnier et al. 2021) improves graph readability for readers with common forms of color blindness and/or color vision deficiency. The package 'forcats' (Wickham 2021) reorders factor levels and modifies factor levels.

2.2 Methodlogy

2.2.1 General Social Survey 2021 Methodlogy

The General Social Survey 2021 is a national representative survey in the United States. The survey was conducted by 27,591 adults 18 or older in the United States who live in noninstitutional housing at the time of interviewing. This report summarizes the findings of the General Social Servey(GSS) 2021. The GSS 2021 collects data on American society to monitor and explain trends in opinions, attitudes, and behaviors. The GSS 2021 has adapted questions from earlier surveys, combing with some of the issues occurring in real time. The GSS has been administered by NORC at the University of Chicago (NORC) and funded by the National Science Foundation (NSF) since its inception. Currently, the GSS is designed by a set of Primary Investigators (PIs), with input from the GSS Board, comprised of notable researchers within the scientific community. The GSS 2021 contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. The GSS staff conducted an independent fresh cross-sectional address-based sampling push-to-web study (referred to in this document as the 2021 GSS Cross- section but also known as the 2020 cross-sectional survey in previous documents). This survey provides details of the second study—namely, the 2021 GSS Cross-section, where newly selected respondents answered a GSS questionnaire from December 2020 to May 2021. We refer to the second study as the 2021 GSS Cross-section because the majority of the data was collected in 2021.

GSS variables appear at three different frequencies. Items in the Replicating Core, Household Composition, and Contact/Validation typically appear every year and are rarely altered (although, in 2020, they feature modification for the web mode). Items in some topical modules, such as ISSP modules, can appear in multiple years, but not every year. Finally, items on the rest of the topical modules typically only appear in a single year. In the 2021 GSS, the NSF Science Knowledge and Attitudes (reduced set of questions),

Healthy People, ISSP Social Inequality, and ISSP Environment modules are repeats from previous years, while the Social Inequality (sponsored), Religion (sponsored), and Board-Initiated modules are new.

Due to the global COVID19 pandemic, the 2021 GSS implemented significant methodological adaptions for the safety of respondents and interviewers. Since its inception, the GSS has traditionally used in-person data collection as its primary mode of data collection. However, the 2021 GSS Cross-section used an address-based sampling with push to web and a web self-administered questionnaire, with phone interviews as a secondary mode. While the data will contribute to our understanding of society, any changes in public opinion seen in the 2021 GSS data could be due to either changes in actual opinion and/or changes the GSS made in the methodology to adapt to COVID-19. When evaluating the GSS for trend changes over time, the GSS staff caution their users to carefully consider changes in the GSS methodology from a total survey error perspective, examining how a change they are observing in a trend may have been impacted by the methodological differences employed in 2021. Total survey error is a way of comprehending the impact on estimates due to measurement, non-response, coverage, and sampling error.

2.2.2 Data Cleaning Methodlogy

I select the 9 variables from the GSS 2021 raw dataset, which are income, age, educ, emailhr, quallife, hlthphys, hlthmntl, age_group, educ_level, sex. By applying the 'mutate' function, I replace the numbers under variable educ by corresponding education levels, substitute the numbers under variable age by corresponding age group, making it easier to observe the selection of the respondents. With the 'filter' function , I deleted all of the missing responses existing under the variables of our interests. Finally, by the 'select' function, I choose all variables necessary in this analysis to conduct my cleaned dataset.

2.3 Data Description

The glimpse of the final dataset that we cleaned is shown in Table 1 below.

Table 1: Clean Dataset

wwwhr	income	age	educ	emailhr	quallife	hlthphys	hlthmntl	age_group	educ_level	sex
35	refuesed	60	16	30	EXCELLENT	VERY GOOD	VERY GOOD	60-69	college or above	male
40	$25,000 or more	20	12	1	VERY GOOD	GOOD	FAIR	20-29	high school	male
3	$25,000 or more	76	13	2	GOOD	GOOD	GOOD	70-79	college or above	male
80	$25,000 or more	37	11	40	GOOD	FAIR	FAIR	30-39	high school	male
2	refuesed	75	18	4	EXCELLENT	GOOD	EXCELLENT	70-79	college or above	male
30	$25,000 or more	22	15	2	VERY GOOD	GOOD	FAIR	20-29	college or above	female
50	$25,000 or more	33	17	4	FAIR	GOOD	GOOD	30-39	college or above	female
95	$25,000 or more	31	16	10	VERY GOOD	FAIR	FAIR	30-39	college or above	male
20	$25,000 or more	37	14	1	GOOD	VERY GOOD	VERY GOOD	30-39	college or above	male
6	$25,000 or more	53	12	5	GOOD	GOOD	FAIR	50-59	high school	male

As Table 1 shown, there are 11 variables that are:

- wwwhr: The amount of hours respondents spend online per week. This is a numeric variable.
- income: The income level of the respondents, which is divided into 12 intervals except for refused. This is a categorical variable.
- age : The exact age of the respondents.This is a numeric variable.
- educ: The number of years that the respondent has accepted for education. This is a numeric variable.
- emailhr: The number of hours respondent spent on sending emails. This is a numeric variable.
- quallife: The evaluation of respondent on his/her quality of life, which is divided into 5 categories. This is a categorical variable.

- hlthphys: The evaluation of respondent on his/her physical health, which is divided into 5 categories. This is a categorical variable.
- hlthmntl: The evaluation of respondent on his/her mental health, which is divided into 5 categories. This is a categorical variable.
- age_group: The age group of the respondents, which is divided into 8 intervals. This is a categorical variable.
- educ_level: The education level of the respondents, which is divided into 4 categories. This is a categorical variable.
- sex: The gender of the respondents, which is divided into 2 categories. This is a categorical variable.

2.4 Strength

This questionnaire is very informative covering a wide range of areas, it contains 565 different questions including a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. We can use this questionnaire not only to study the relationship between the amount of time spent online and some personal characteristics, but also to investigate many issues related to civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. The content of this questionnaire is less limited and there is a great deal of research that can be done based on this questionare.

2.5 Weaknesses

We are all aware that the United States is a country of immigrants gathering people from all over the world. According to statistics, there are between 350 and 430 languages spoken in the United States of America (Yamazaki 2021). However, this questionnaire is only available in English and Spanish. Although we know that English and Spanish are the most widely spoken languages in the United States, there is still a segment of the population that is not fluent in English and Spanish. It was difficult for this group of people to participate in the survey.

2.6 Key Features

2.6.1 Time spent online

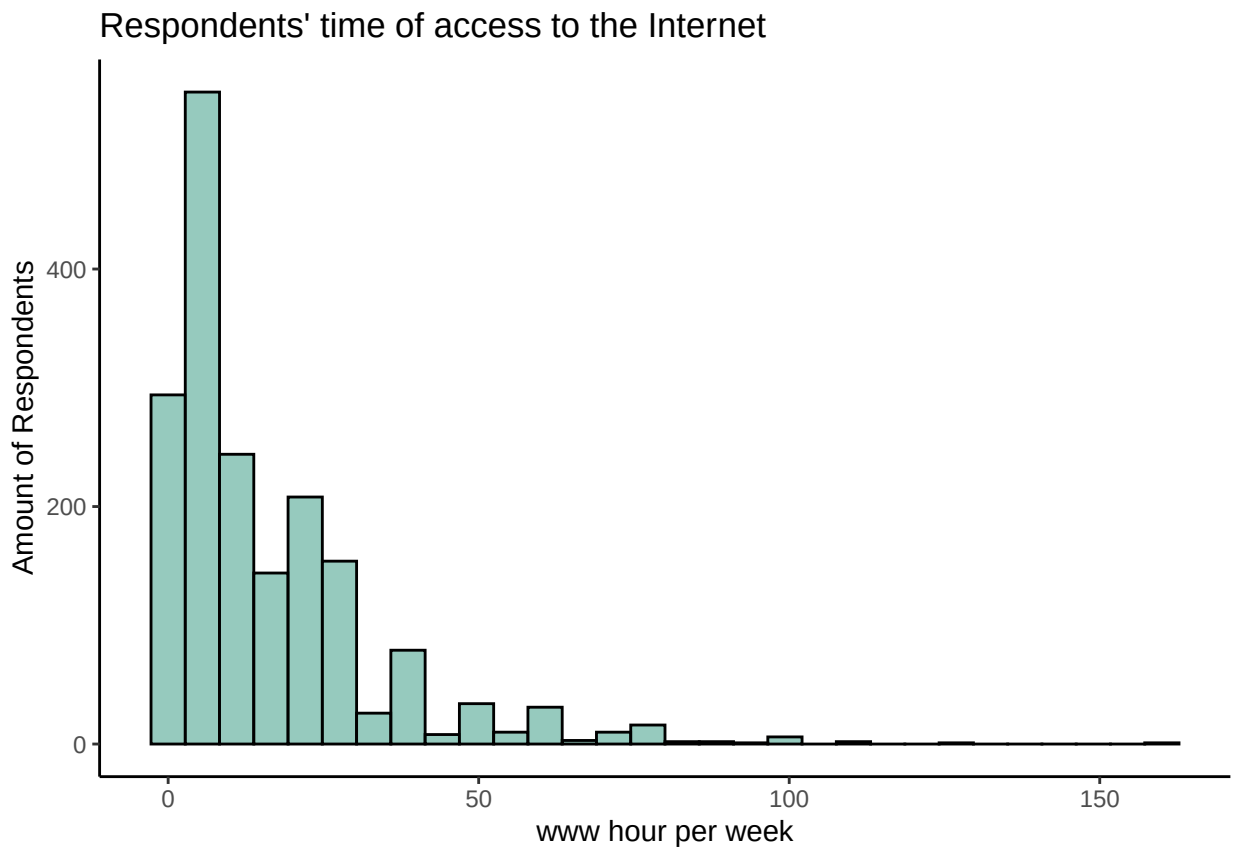


Figure 1: The distribution of time spent online and the email hours

Figure 1 depicts the distribution of weekly Internet browsing time of the respondents. As we can see, the number of hours spent online is mainly concentrated in the range of 0-50 hours. It is observed that the largest amount of respondents, even more than 600, spend 8-14 hours per week on the Internet. When the time spent online exceeds 14 hours, we find that with an increasing number of hours spent online, the number of corresponding respondents is continuously reducing. Since we notice that this histogram is organized in right skewed distribution, it is clear that the majority of respondents have less time on the Internet and only a small amount of respondents spend very long periods of time on the Internet.

2.6.2 Income level

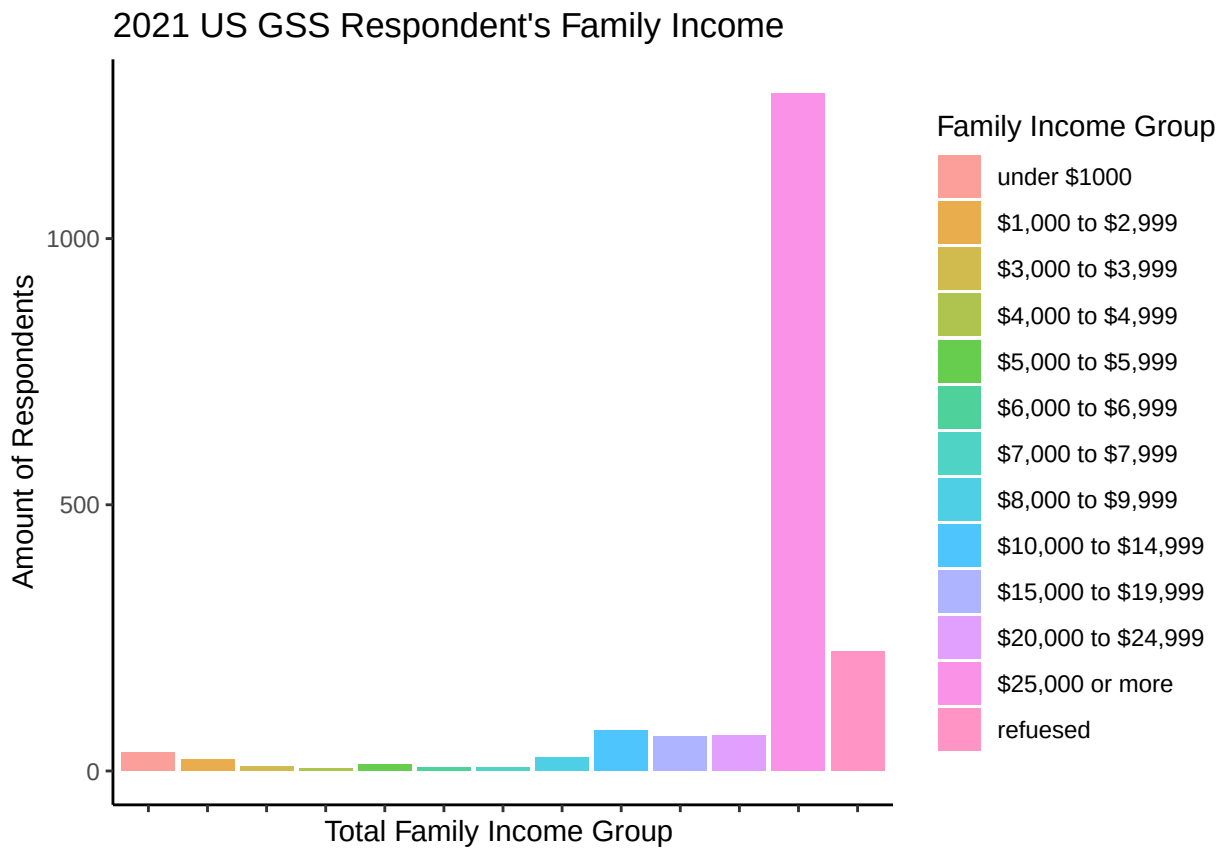


Figure 2: The majority of respondents have medium to high household incomes

Figure 2 presents the intervals of household income of the respondents. We find a left-skewed trend in the figure distribution, with very few respondents having household incomes below 10,000 dollars, therefore we can infer that only a small percentage of respondents have low family incomes within the total sample. Interestingly, we find that people's responses are mainly concentrated in the $25,000+ range. It reflects the relatively high level of income among the respondents. However, due to the fact that the questionnaire categorizes respondents with household incomes greater than or equal to 25,000 dollars into a whole bracket, we are unable to observe whether the income gap between the affluent and mid-class respondents is very large. Generally speaking, there are fewer low-income groups among the respondents, the majority of the respondents have household incomes higher than average.

2.6.3 Time spent on sending emails

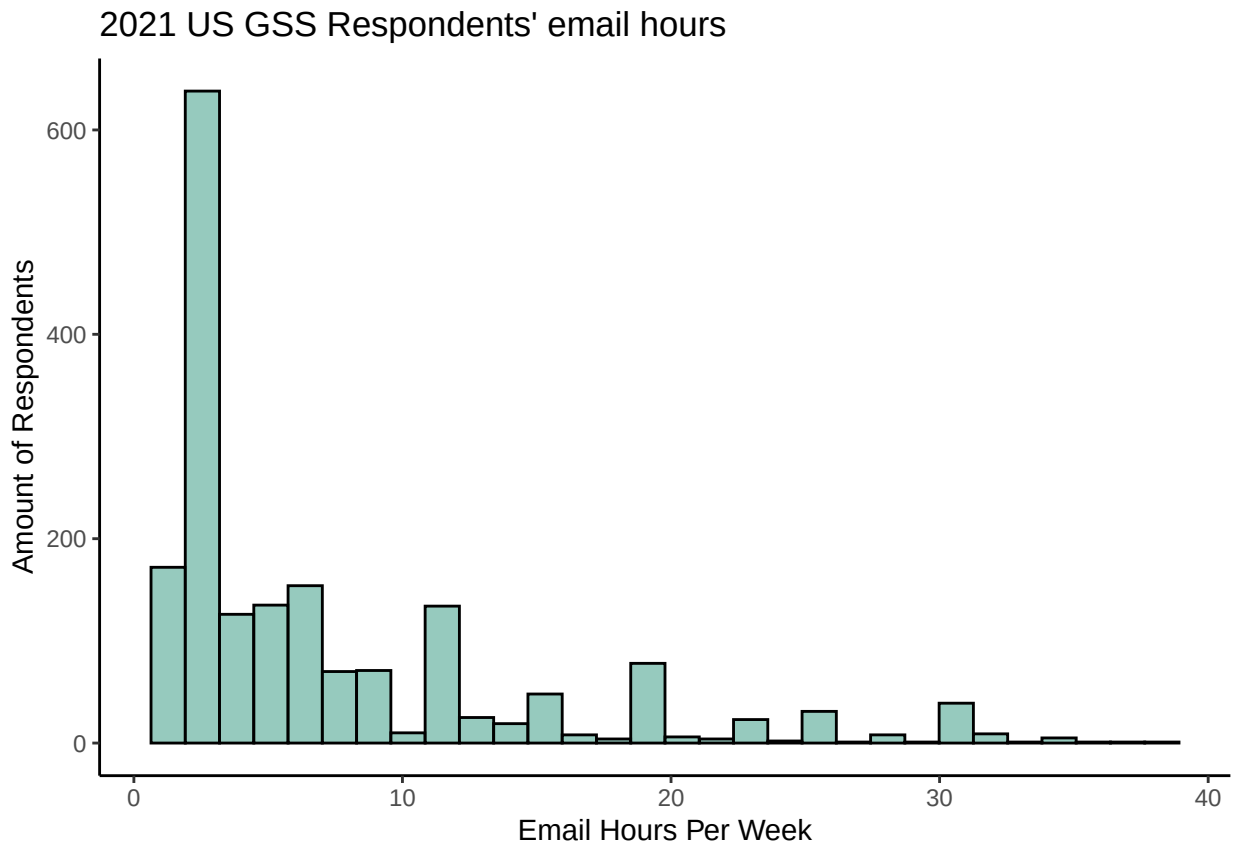


Figure 3: Respondents spend a wide range of time on emailing each week

Figure 3 illustrates the distribution of respondents' weekly emailing hours. We observe that respondents spend a wide range of time on emailing each week, but the majority of them spend 0-13 hours on sending emails, which is relatively short time each week. Among them, the largest number of respondents indicated that they spent 1.5-3 hours per week on posting emails. The number of respondents who spent 0-1.5 hours, 3-4.5 hours, 4.5-6 hours and 6-7.5 hours on emailing was very similar, with about 170 respondents. Since the entire distribution shows right-skewed trends, we realize that the very minority of respondents spend long periods of time on sending emails, but there exists a fraction of respondents who devote extremely large amounts of time to emails. Therefore, we infer that most people spend some time on sending emails for work and social reasons, but usually it does not take up too much of their time, However, there exists a small group of respondents who dedicate a lot of time on emails probably due to the requirements of their work.

2.6.4 Quality of life

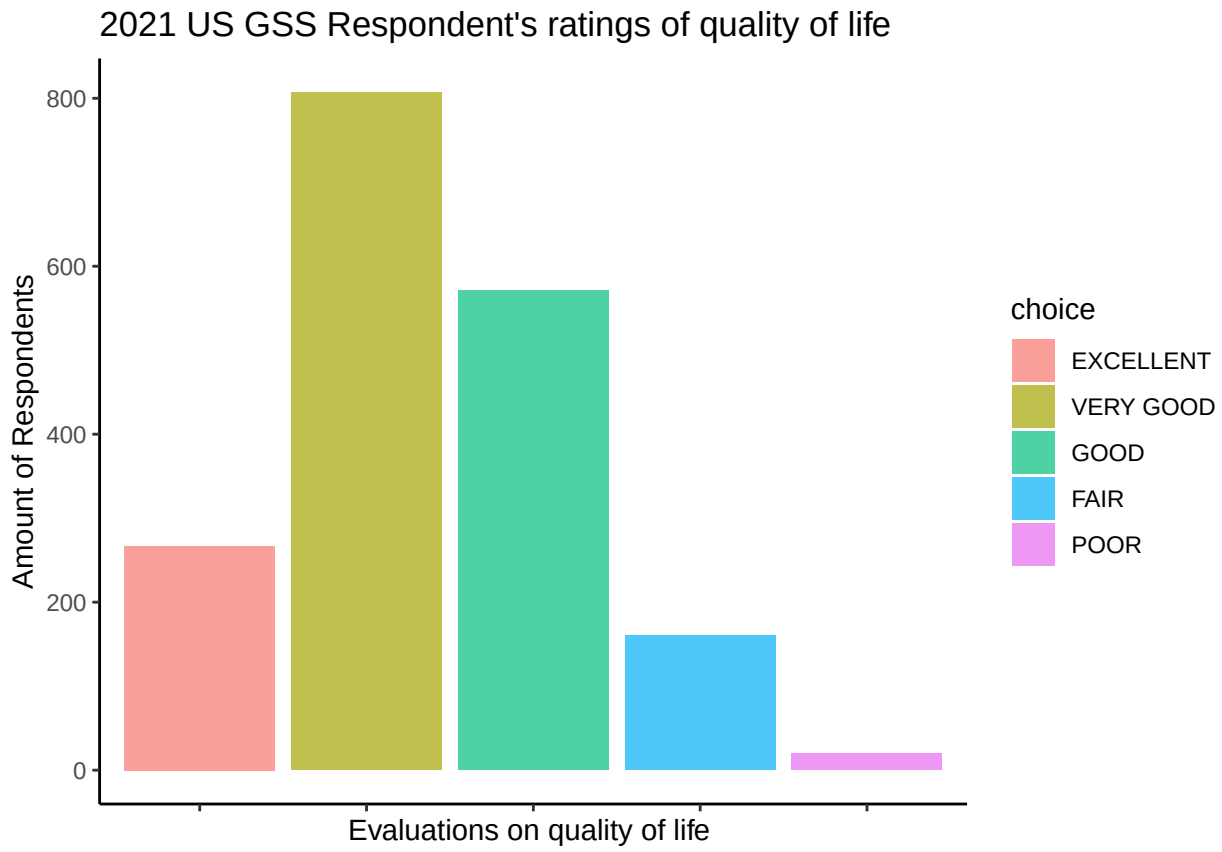


Figure 4: Most respondents have relatively positive evaluations of their current quality of life

Figure 4 presents the respondents' evaluation of their quality of life. The ratings are divided into five main levels, namely excellent, very good, good, fair and poor. We found that the number of respondents who assessed their quality of life as excellent was about 250, which is less than the number of respondents who chose very good and good. It reflects that most people still think there could be some improvements in their quality of life and they are not completely satisfied with their life at this stage. We observe that the number of respondents who chose very good is almost 800, making it the most popular choice among the respondents. Similarly, the number of respondents who chose good is also relatively high. Among the respondents, less than two hundred people chose fair, whereas we also noticed that very few people chose poor. By combining these points, we find that most people are not 100% satisfied with their current life, but they still have a relatively positive evaluation of their current quality of life.

2.6.5 Physical health

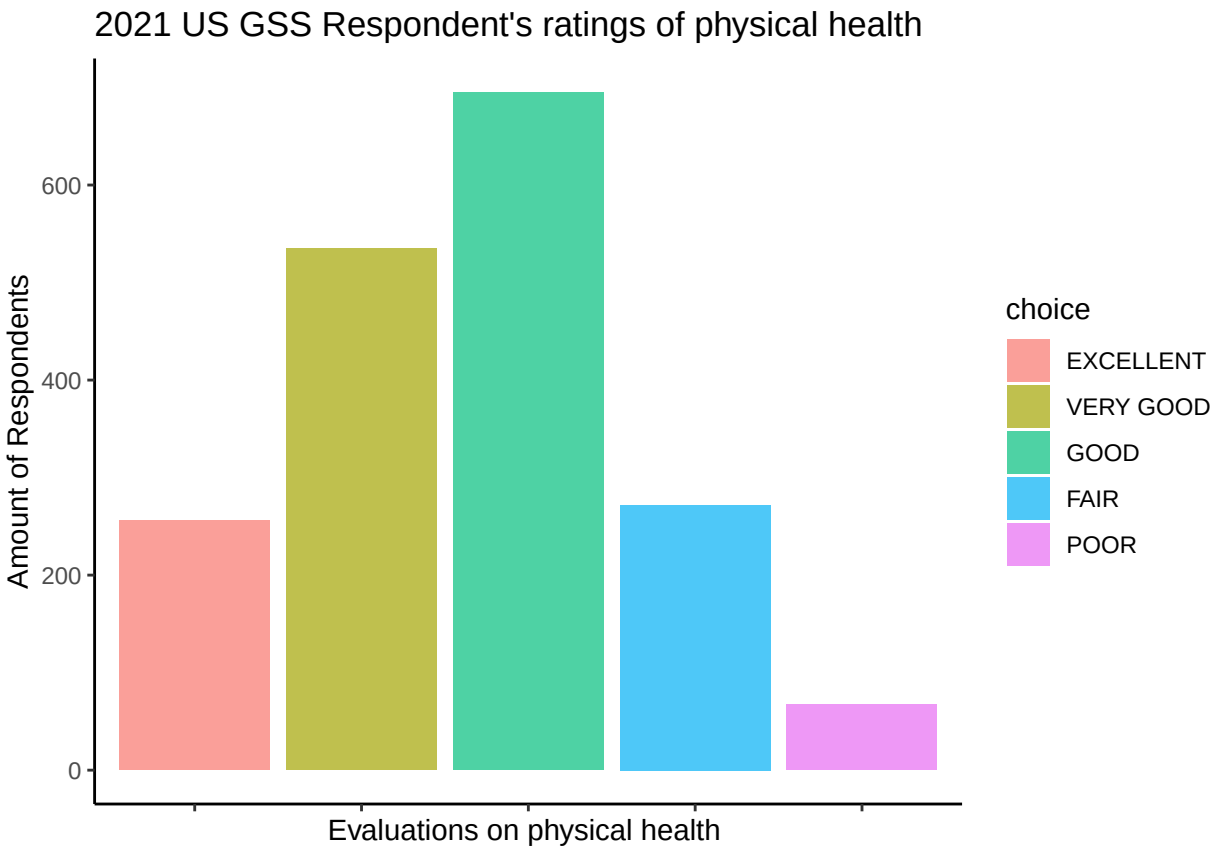


Figure 5: A significant proportion of respondents are still concerned about their physical health

Figure 5 provides an overview of the distribution of the respondents' ratings of their physical health. Similarly, the assessments are categorized into five levels: excellent, very good, good, fair and poor. Among the relatively positive ratings, the number of respondents decreases as people's satisfaction with their health improves. It reflects the fact that only a small proportion of respondents are completely satisfied with their current physical status, while most people are not completely confident about their physical health. We found that the largest number of respondents chose the option of good, as over 600 respondents consider their current physical health to be in a relatively average state, not particularly good, but not bad either. Compared with the previous evaluations of quality of life, we find that people's evaluation of physical health is not that positive, and it seems that the number of people choosing fair and poor has increased a lot. A significant proportion of respondents are still concerned about their physical health.

2.6.6 Mental health

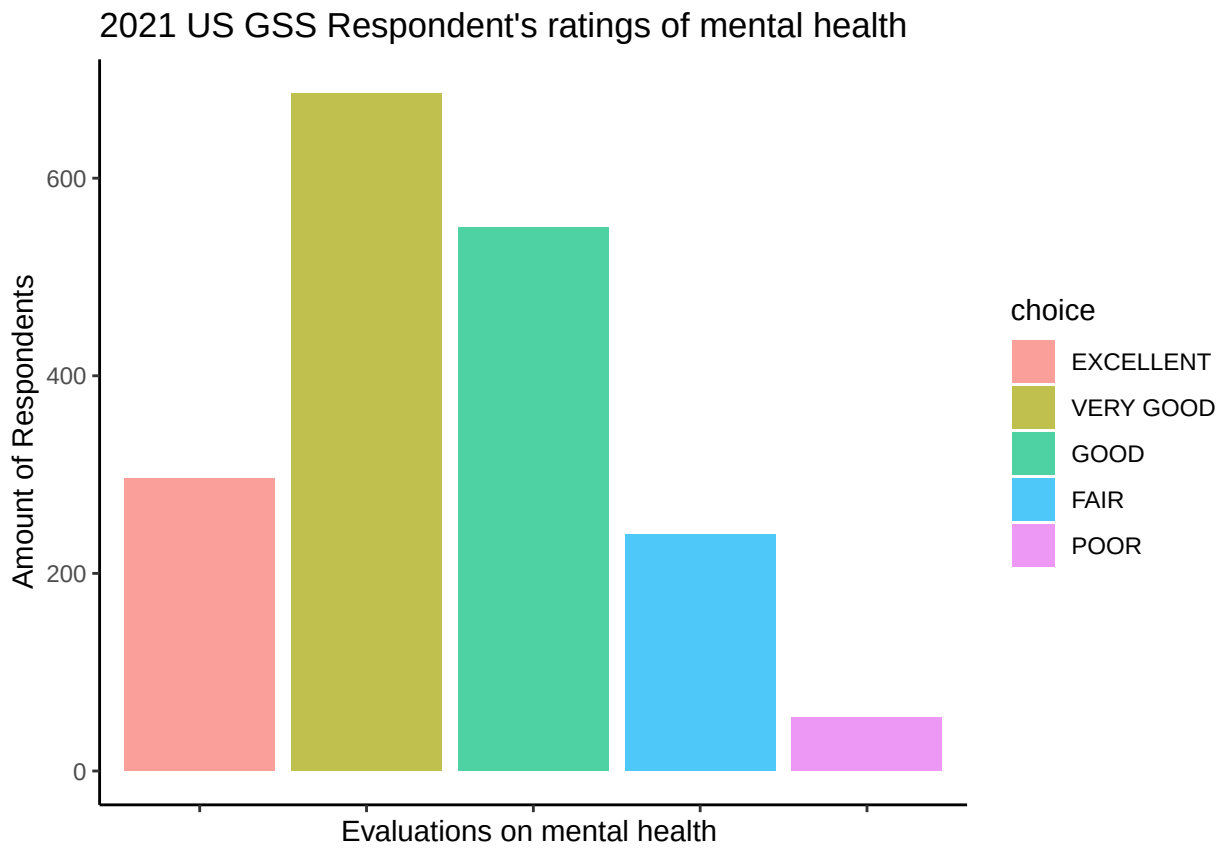


Figure 6: the spike in the number of people choosing very good option reveals an increased level of satisfaction with people' s mental health status

Figure 6 demonstrates the distribution of respondents' responses to their mental health evaluations. Like the previous evaluations, the assessment of mental health was classified into 5 levels: excellent, very good, good, fair and poor. Nearly 300 respondents rated their mental health status as excellent, a significantly higher number than those who chose this option for the quality of life and physical health ratings. Respondents seemed to be more positive about their mental health status. In addition, the spike in the number of people choosing very good option also reveals an increased level of satisfaction with people' s mental health status. More than 600 people believe they have very good mental health currently. The number of people who selected for this option even surpassed the most popular option in the previous assessment, which is good. It is notable that about 550 people feel that their mental health status is on the good scale. Most people have very aggressive ratings for mental health. Only a small percentage of respondents felt that their current mental state was fair or poor.

2.6.7 Age group

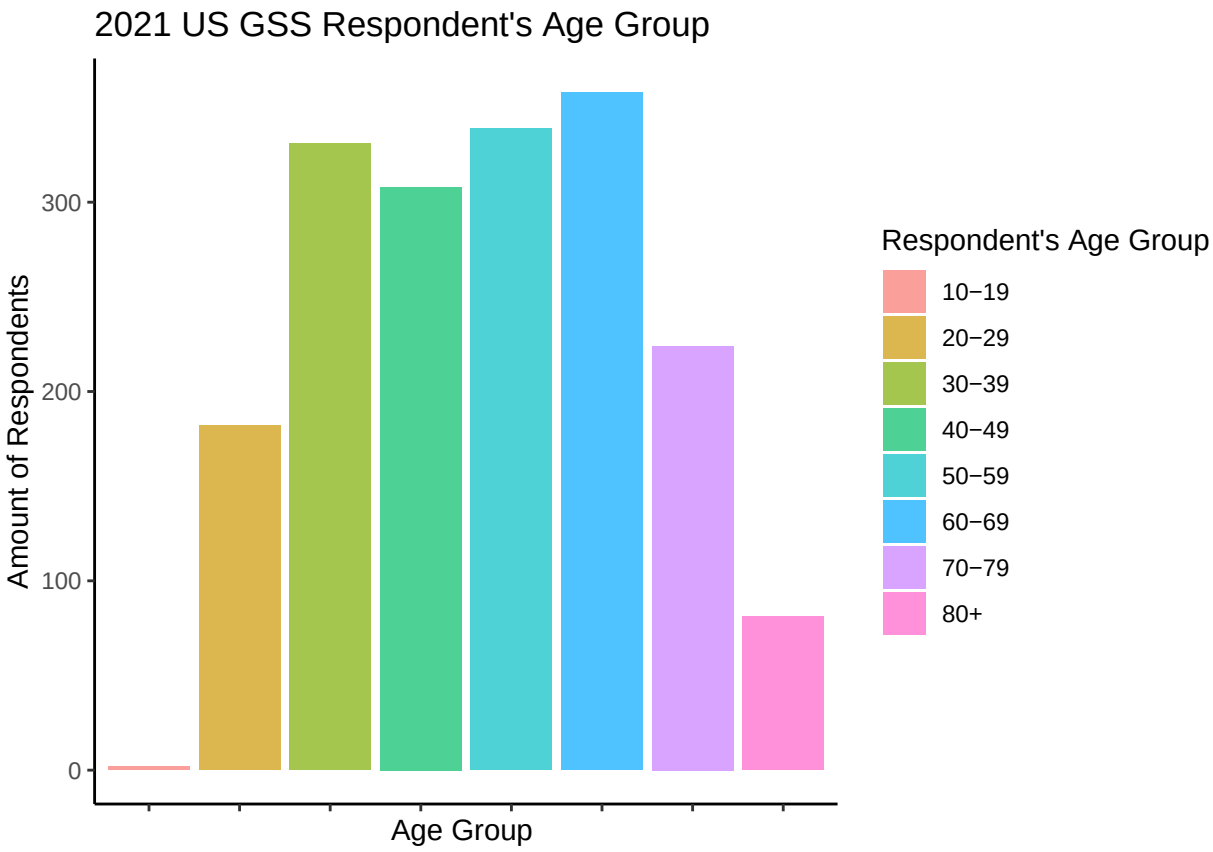


Figure 7: In most cases, respondents were in the 20 to 69 age range

Figure 7 displays the age distribution of the respondents. It can be observed that almost no respondents are in the age range of 10-19 years old. In addition, the number of respondents whose age is between 20 and 29 years old is less than 200. It seems that in our survey, there are only limited number of respondents in the lower age range. The amounts of respondents in the 30-39, 40-49, 50-59 and 60-69 age ranges are very close to each other, with the number of respondents in these age groups being around 350. Moreover, we could identify that there are only a small number of respondents who are over 70 years old, accounting for a small percentage of the total number of respondents. In general, the age distribution of respondents was mainly in the 30-69 age range, with the majority of respondents being middle-aged.

2.6.8 Education level

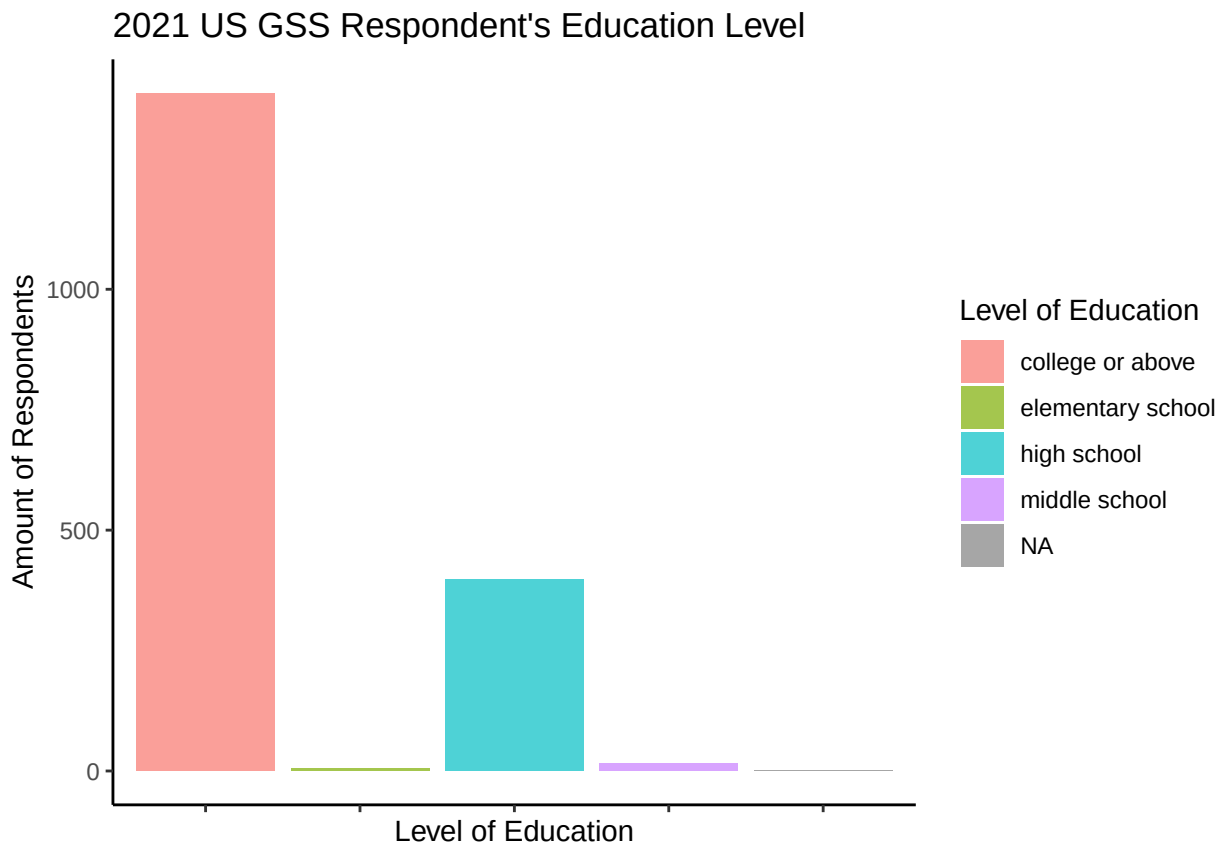


Figure 8: Most of the respondents have received education beyond high school or college level

Figure 8 reveals the distribution of education levels of the respondents. The graph shows that over 1,300 respondents indicated that they have completed college or higher education. About 400 respondents received high school education. Only a few respondents indicated that they had only attended elementary or junior high school. This phenomenon reflects the fact that the average education level of people in the United States is very high, and most people have access to at least high school education.

2.6.9 Gender

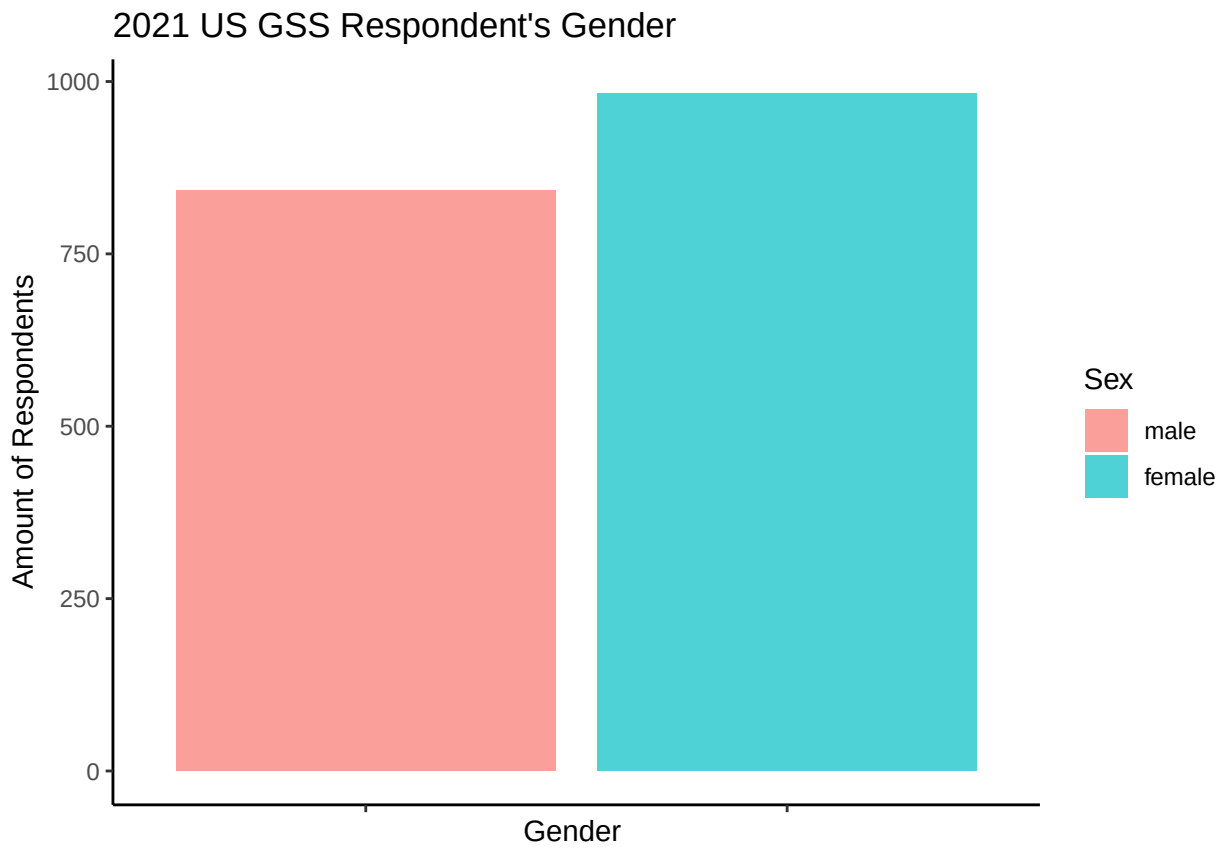


Figure 9: There were obviously more female respondents than male respondents

Figure 9 exhibits the gender distribution of the respondents. We find that the number of male respondents is close to 1000, while the total amount of female respondents is only approximately 800. Based on our common sense, we realize that the ratio of male respondents to female respondents should be close to 1:1, whereas here the number of male respondents is much higher than the number of female respondents, which reflects the gender imbalance in our respondent group.

3 Model

3.1 Linear Regression Model

Firstly,we choose a starting model with the response variable wwwhr and 8 potential predictors, which are income, age, educ, emailhr, quallife, hlthphtys, hlthmntl and sex. Then we need to ensure that there's no multicollinearity presented in our model since the multicollinearity can make our important variables look unimportant by expanding the variance of all predictors. Under the impact of multicollinearity, we will get inaccurate conclusions when choosing our model based on the p-value of each predictor. We will only select the predictors with low variance inflation factors which are less or equal to 5. After observation, the variance inflation factors of all chosen predictors are smaller than 5 so no predictor should be removed.

Table 2 : *The condition table*

Condition	Request
Condition 1	Conditional mean response is a single function of a linear combination of the predictor
Condition 2	Conditional mean of each predictor is a linear function with another predictor

After that, we need to check the 2 conditions to make sure that we can use redsidual plots to disgnose the model. In condition checking part, for condition 1, the conditional mean response is a single function of a linear combination of the predictors. We observed that there is a pattern between the real value y_i(wwwhr) and the fitted value $\hat{y}$, and the shape of the graph is not in a fanning pattern, so it satisfies the condition 1. For condition 2, the conditional mean of each predictor is a linear function with another predictor. We observed that there are no relations between the predictors, so it satisfies the condition 2. Thus we can use residual plot to explore what exactly is being violated.

Then we need to use residual plots to identify potential violations against model assumptions, there are four assumptions that need to be satisfied in order to apply the multiple linear regression model.
1. linearity
2. normality
3. constant variance
4. uncorrelatedness

In the residual vs. fitted plot, there is no obvious pattern shown here. However, there is a potential linear relationship in the residual vs. fitted plot and a small positive linear trend in the residual vs. age plot. These two figures indicates that there might exist the violation of the independence of linear model. In the residual vs age plot, we indicate that the plots have no fanning pattern, so we can conclude that there's no violation in the model's constant variance assumption. In the residual Q-Q plot, the points are mostly distributed around the line with angel of 45 degrees. Also, there is no obvious deviation, so there is no violation in he normality assumption of the model. By the assumption check, I found that there are violations in the model's independence assumption for the residual vs. fitted plot and residual vs. age plot. In order to avoid the violations, we did model transformation by applying box-cox transformation. We get the estimated power of each predictor, forming a transformed model here. The powerTransform is used to determine numerical variable transformation. The following table 3 shows the transformation.

Table 3 : *The transformation table*

Variable Name	Est. Power	Round Power
wwwhr	0.0830	0.08
age	0.7066	0.71

By the estimated power, I mutated a new variable wwwhr_trans and referring the value of that as wwwhrˆ0.08, similarly, I muatated age_trans as ageˆ0.71. So the numerical variables are transformed to normality, such that it can satisfy the assumption of independence. A new model, model_trans is generated by using the transformed variables, with the response variable: wwwhr_trans and 8 predictors, which are income, age_trans, educ, emailhr, quallife, hlthphys, hlthmntl and sex. Then, it is necessary to re-assess condition 1 and condition 2 of the transformed model.By repeating the steps as above, I used yi vs. y_hat model and pairs plots to test condition 1 and 2. It is obvious that the condition 1 is satisfied because there is a pattern between the real yi and fitted value y hat, and the shape of the graph is not fanning patter.The condition 2 is also satisfied because there is no linear relationship between predictors.

The residual plots is used on the transformed model to see whether it is an important model over the original model. There is no obvious pattern in the residual vs. transformed age and transformed y indicating that there is no violation in the model's independence assumptions now. Also the plots are not in a fanning pattern, so there is no violations in the model's constant variance assumptions. The QQ plot fits more line than the last QQ plot, which indicates that there is no violation in the normality assumption of the model. Thus, all assumptions are satisfied in the transformed model.

Since all of the assumptions of linear regression model are satisfied now, the linear model should be conducted by using the transformed dataset. By comparing the p-value of each predictors, the predictors with lower p-value is keeped, which indicates that it is significant predictor.
The only variable that was deleted was hlthphys, indicating that there is no significant linear relationship between the physical health and the time people spend on internet.
M2 is the reduced model after deleting the variable hlthphys. After using partial F-test, the p-value for the reduced model is 0.4458, which is larger than the significance level 0.05. Thus, we can conclude that the reduced model is better than the original model after transformation.
The next step is to check two conditions to make sure that we can use residual plots to diagnose the reduced model.

The y vs. y hat graph and pairs plots are similar as the graphs of transformed model. Condition 1 and condition 2 are both satisfied. The residual plots also have no pattern, indicating no violation in the model's independence and constant variance assumptions.

Table 4 : *The summary table*

Variable	Estimate	Std. Error	t value	Pr(>abolute(t))
income$1,000 to $2,999	2.838e-02	2.406e-02	1.180	0.23822
income$3,000 to $3,999	7.719e-02	3.274e-02	2.358	0.01850*
income$4,000 to $4,999	2.360e-02	4.971e-02	0.475	0.63501
income$5,000 to $5,999	9.740e-04	2.809e-02	0.035	0.97234
income$6,000 to $6,999	1.530e-02	3.683e-02	0.415	0.67796
income$7,000 to $7,999	-3.171e-02	3.675e-02	-0.863	0.38834
income$8,000 to $9,999	5.153e-02	2.290e-02	2.250	0.02458*
income$10,000 to $14,999	3.505e-02	1.778e-02	1.972	0.04882*
income$15,000 to $19,999	-1.416e-05	1.820e-02	-0.001	0.99938
income$20,000 to $24,999	2.876e-02	1.823e-02	1.578	0.11485

Variable	Estimate	Std. Error	t value	Pr(>abolute(t))
income$25,000 or more	2.963e-02	1.530e-02	1.937	0.05291.
incomerefuesed	2.681e-02	1.633e-02	1.642	0.10072
age_trans	-5.436e-03	5.647e-04	-9.626	< 2e-16 ***
educ	1.979e-03	8.049e-04	2.459	0.01402 *
as.numeric(emailhr)	3.183e-03	2.753e-04	11.560	< 2e-16 ***
quallifeVERY GOOD	-6.470e-03	6.247e-03	-1.036	0.30053
quallifeGOOD	-1.580e-02	7.151e-03	-2.209	0.02728*
quallifeFAIR	-2.670e-02	9.847e-03	-2.711	0.00677**
quallifePOOR	4.393e-02	2.297e-02	1.912	0.05600.
hlthmntlVERY GOOD	1.675e-03	6.065e-03	0.276	0.78247
hlthmntlGOOD	1.257e-02	6.643e-03	1.892	0.05862.
hlthmntlFAIR	2.640e-02	8.209e-03	3.216	0.00132**
hlthmntlPOOR	3.087e-02	1.445e-02	2.137	0.03275*
sexfemale	-1.231e-02	3.977e-03	-3.095	0.00200**

Finally, we want to get the form of the final model and interpret model coefficients. By observing the p value of each predictors of our transformed model, we keep only the predictors with smaller p-value and form the reduced transformed model.The final model has predictors of income, transformed age, education level, email hours, life quality, mental health and sex. The equation for the final model is shown as:

$Y_{log\_wwwhr} = \beta_0 + \beta_1 X_{income} + \beta_2 X_{transformed\_age} + \beta_3 X_{email\_hours} + \beta_4 X_{quality\_of\_life} + \beta_5 X_{mental\_health} + \beta_6 X_{female}$

Income Income levels are separated into 11 different groups and not every group is significant correlated with the response variable wwwwhr. There are only three levels are significant relate with wwwhr. The income level of $3,000 to $3,999 has positive impact of web hours with p-value of 0.0185. For this level, when income level increase by 1 unit, then wwwhr increase by 7.719%. The income level of $8,000 to $9,999 has positive impact of web hours with p-value of 0.02458. For this level, when income level increase by 1 unit, then wwwhr increase by 5.153%. The income level of $10,000 to $14,999 has positive impact of web hours with p-value of 0.04882. For this level, when income level increase by 1 unit, then wwwhr increase by 3.505%.

Age The transformed age is age^0.71 as shown in SECTION XXXX and it is significantly related with wwwhr with p-value of < 2e-16. It has estimate of -5.436e-03 that indicates it is negatively correlated with wwwhr. When transformed age increase by 1, then wwwhr decrease by 0.5436%.

Email hours It is positively related with wwwhr with p-value of < 2e-16. When email hour increase by 1 hour, then wwwhr increase by 0.3183%.

Life quality There are five levels of life quality. With dividing line of Fair, people who have better than fair life quality, their life quality is negatively related with wwwhr. And people who have worse life quality than fair, such as poor life quality would have a positive relationship with wwwhr.

Mental health The mental health is also divided into five different levels from VERY GOOD to POOR and they are all positively related with the response variable. However, only fair and poor mental health are significantly related with wwwhr. For people who have fair mental health, when their mental health increase by 1, then wwwhr increase by 2.64%. For people who have poor mental health, when their mental health increase by 1, then wwwhr increase by 3.087%.

Sex Only females are significantly related with wwwhr with p-value of 0.002 and it is negatively correlated with the response variable. When the proportion of female increase by 1, then wwwhr decrease by 1.231%.

4 Data Simulation

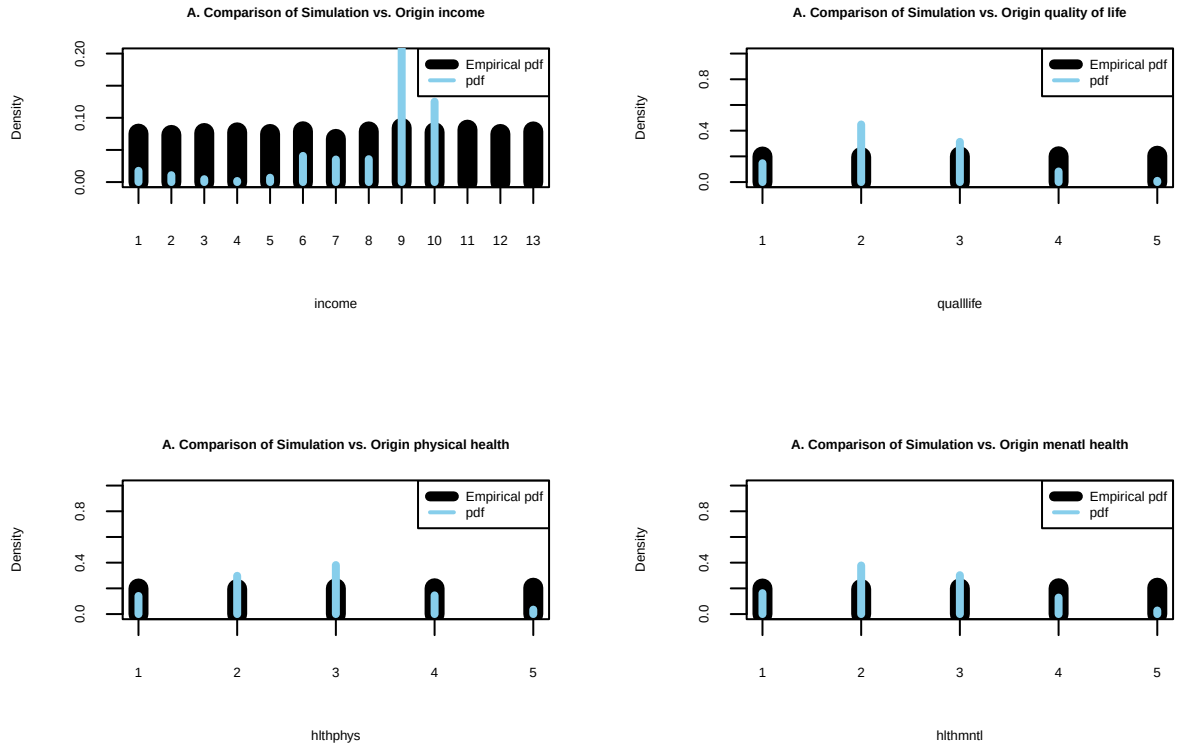


Figure 10: Simulation Result

Figure 10 and Figure 11 show the results of our simulation. We observed that the simulated results differed significantly from the actual values. In the actual income distribution of the respondents, there are few low-salary respondents and most of them have a higher income level. In contrast, the simulated low-salary respondents are much more than the actual, while the simulated high-salary respondents are significantly lower than the reality. When we consider respondents' actual self-assessment (in terms of quality of life, physical health and mental health) of their own conditions, we find that the proportion of respondents who consider their conditions to be very good or good is significantly higher than the proportion under simulation. People were more likely to show their satisfaction with themselves. By comparing the distribution of the actual age range of the respondents with the distribution of the simulated results, we found that the age distribution of the respondents was relatively uneven, with most of the respondents being middle-aged. And the actual distribution of education level differs very much from the simulated distribution results, we found that more people actually received higher education and only a small percentage of them did not receive education beyond high school. For education, the factual results seem to be more positive than the simulated results. For the gender of the respondents, we find that the simulated results are closer to the original results, but there are in fact slightly fewer females than males, which indicates that in general the gender ratio is balanced, whereas the actual quantity of males is slightly more than that of females.

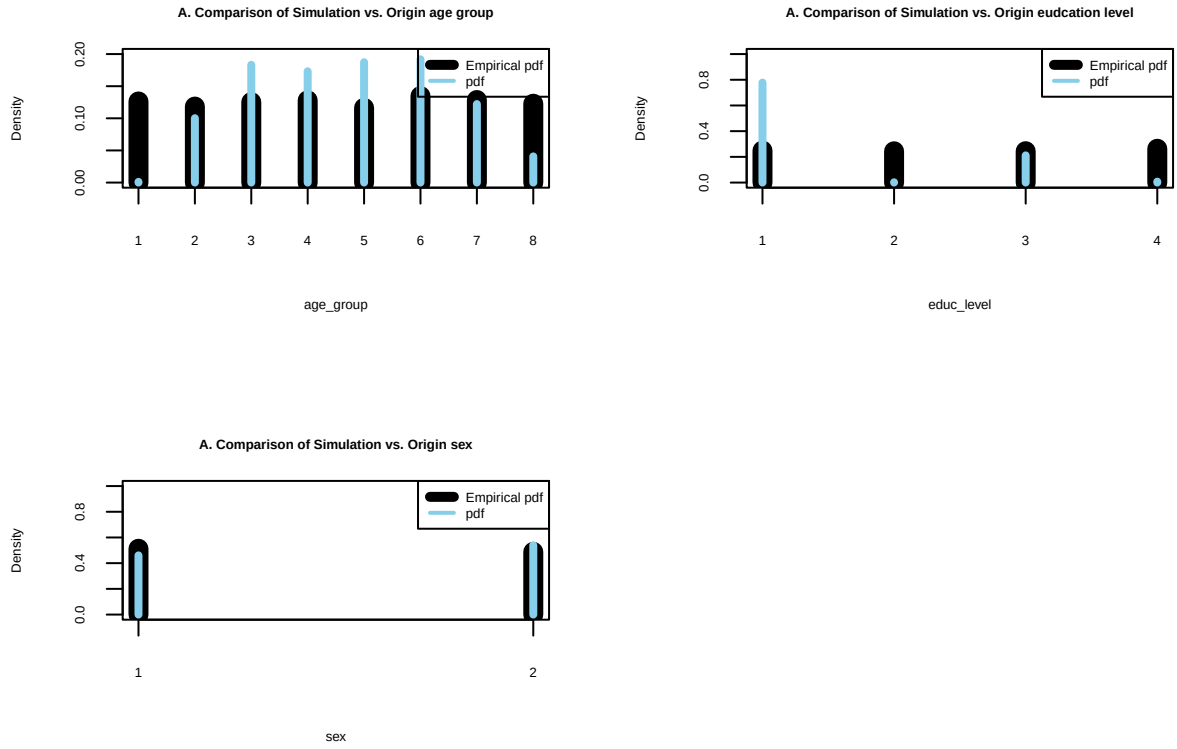


Figure 11: Simulation Result

5 Result

5.1 Time spent online vs. Age

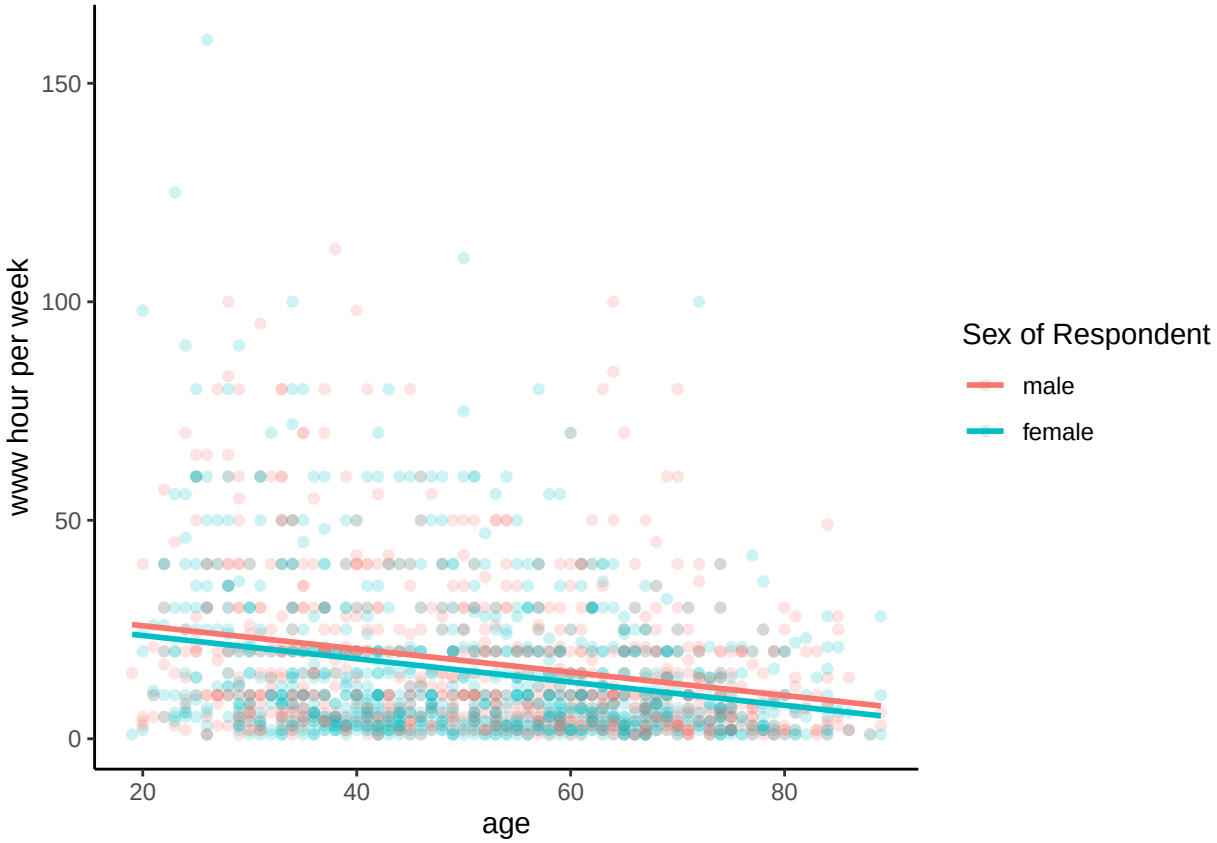


Figure 12: There is a strong linear relationship between age and time spent online

Figure 12 demonstrates the relationship between age and the amount of time spent online per week of respondents in the various genders. We discovered that there is a strong linear relationship between age and time spent online. Age has a negative association with time spent online. As the age increases, the time spent online tends to decrease for both genders. The red and green lines have almost the same coefficient, which reflects that for respondents in different genders, the same amount of increase in age leads to the same amount of decrease in time spent online. It appears that the same age, male respondents would spend more time online than female respondents, as the red line is slightly higher than the green line. However, we also detected some extreme cases in the sample, so the linear relationship presented in the figure can only describe the general trend of the association between time spent online and the age of the respondents.

5.2 Time spent online vs. Sex

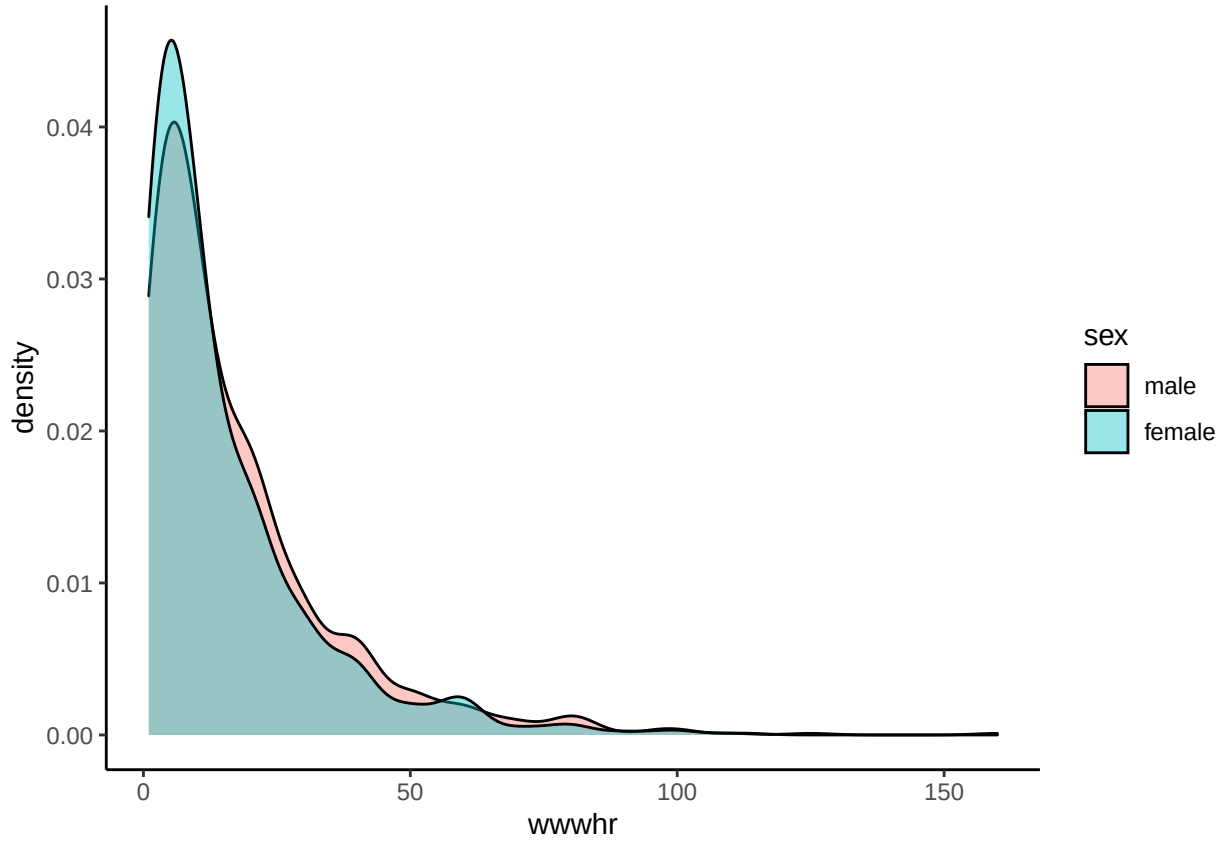


Figure 13: The majority of male and female spend 0-25 hours on the Internet each week

Figure 13 exhibits the distribution of time spent online by respondents in different genders. Since we know the number of male respondents is significantly higher than that of female respondents, at almost all of the time spent online, the corresponding number of males is higher than that of females. The distribution of time spent online is right-skewed for both genders. The majority of male and female spend 0-25 hours on the Internet each week. The largest number of male and female respondents spent 5-10 hours per week online. As the internet accessing time exceeds 60 hours, for a certain fixed time spent online, the proportion of male and female respondents among all was almost the same. Generally speaking, a larger proportion of male respondents tend to spend 0-17 hours per week on the Internet compared with females. In contrast, a greater proportion of females than males spend 17-60 hours per week online. The proportion of female respondents who spend more than 60 hours online is very similar to that of male respondents.

5.3 Time spent online vs. Age Group

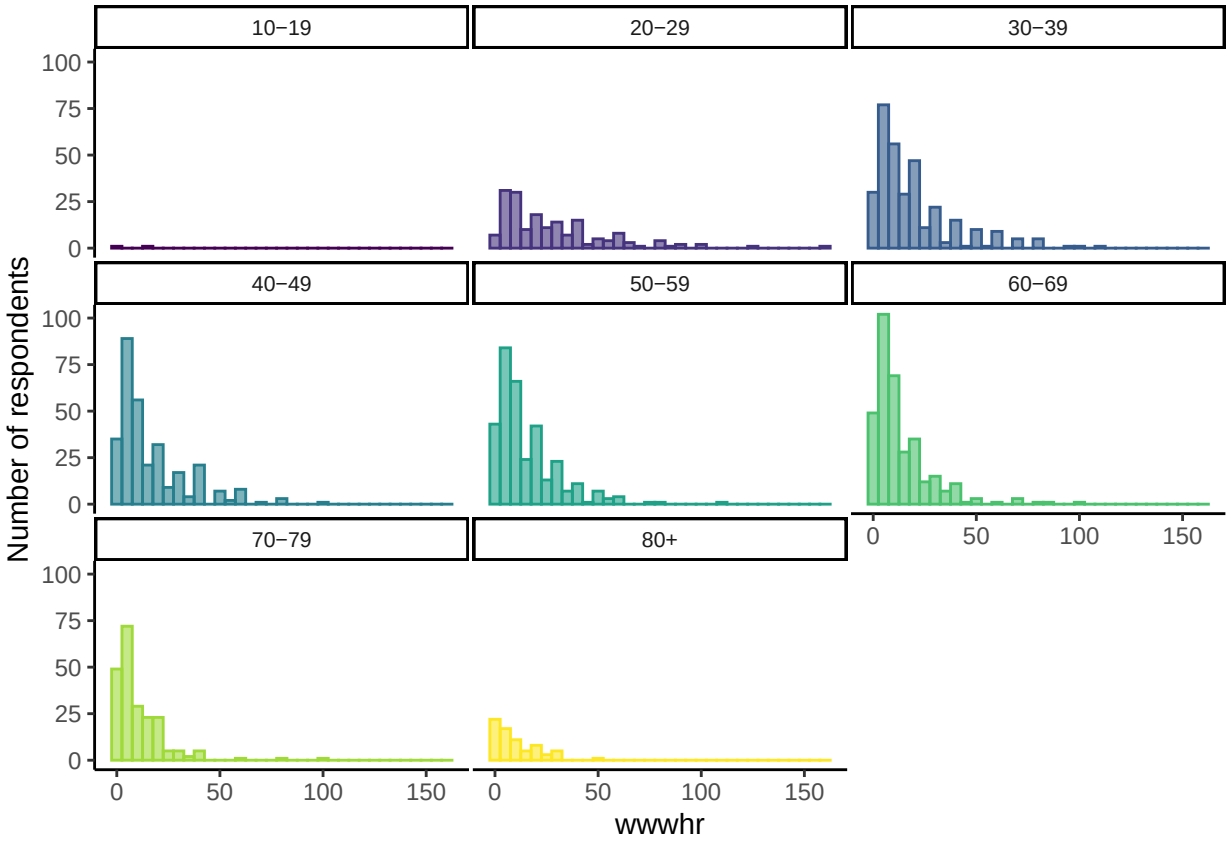


Figure 14: The distribution of time spent online is right-skewed for almost all age groups

Figure 14 displays the distribution of time spent online by respondents of different age groups. We notice that very few respondents are in the age range of 10-19 years old, and only a small fraction of the total number of respondents are over 80 years old. The distribution of time spent online is right-skewed for almost all age groups. In addition, it is found that the time spent online by respondents of different ages is mainly concentrated in the interval of 0-30 hours. Moreover, we also observe that a much larger proportion of respondents aged 30-39, 40-49, and 50-59 have spent more than 40 hours online than other age groups. We speculate that this phenomenon may be due to the fact that middle-aged respondents are facing greater work pressure and have to spend a lot of time on the Internet each week because of their jobs.

5.4 Time spent online vs. Quality of life

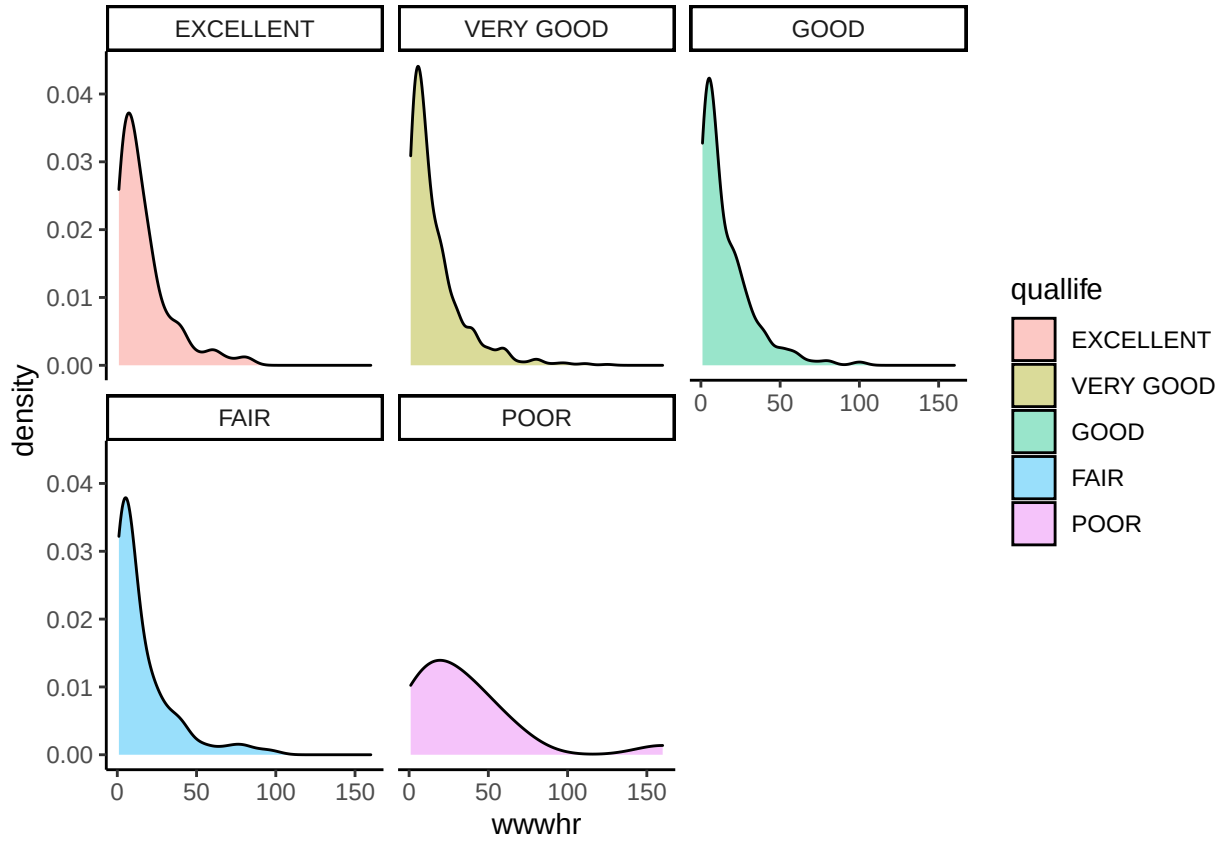


Figure 15: Respondents with poor quality of life are more likely to stay online for linger time

Figure 15 depicts the proportion of time spent online per week by respondents with different levels of quality of life evaluations. Since we already know the distribution of respondents with different quality of life, we are aware that the amount of respondents who chose excellent is smaller than the number of respondents who chose good and very good. However, here it is found that the distribution of time spent online per week is very similar for all three levels of respondents, with most of them spending a short amount of time online and only a small percentage of them choosing to spend a lot of hours on the Internet. For respondents with weaker quality of life ratings, we find that most of them still prefer to spend 0-30 hours online. However, we found that in these two groups, the proportion of respondents who spent more than 30 hours online was much higher than that of the previous groups. In particular, for respondents who rated their quality of life as poor, a large proportion of them were identified as having been online for more than 150 hours. In general, the respondents' Internet surfing time was basically concentrated in the range of 0-30 hours, but the proportion of respondents with poor quality of life who spent more than 150 hours online was significantly higher than that of the other groups.

5.5 Time spent online vs. Physical Health

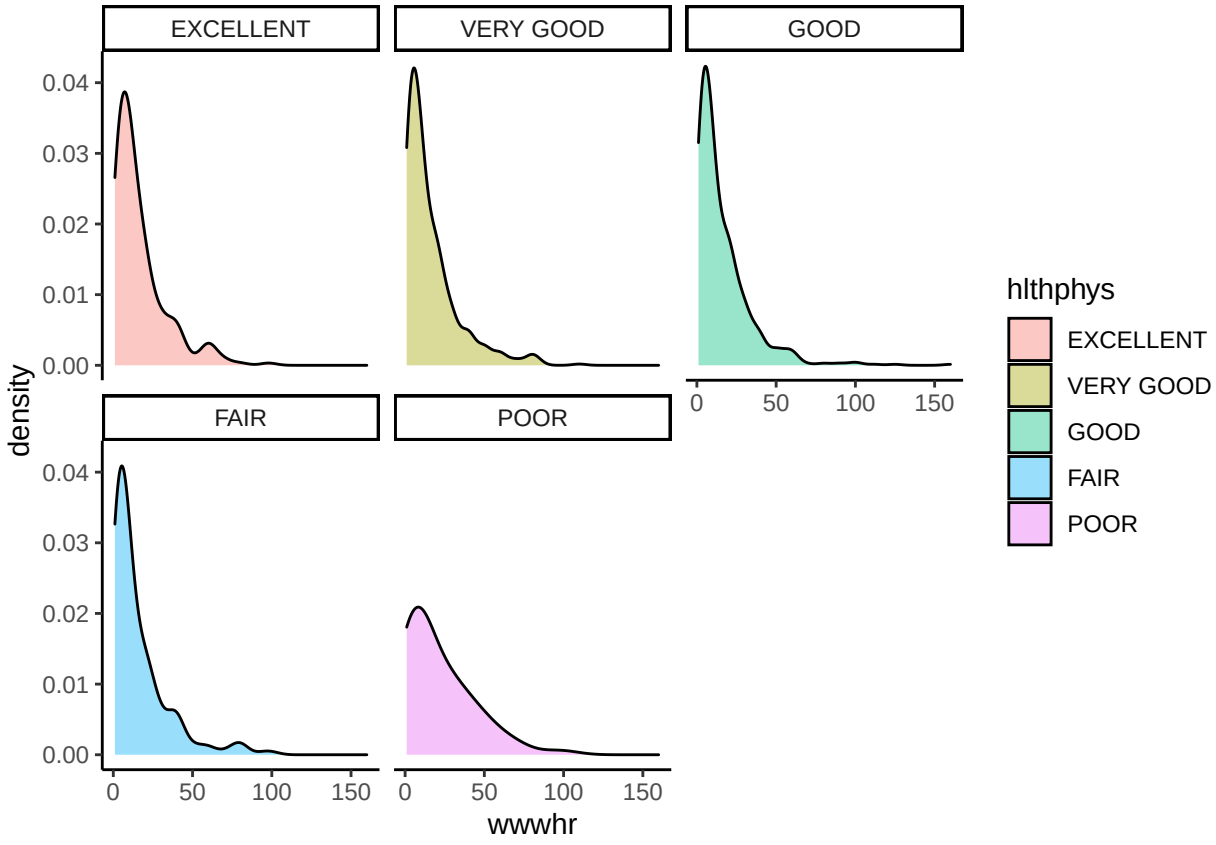


Figure 16: The proportion of respondents with short Internet access hours for respondents with poor physical health was much lower than that of the other levels

Figure 16 portrays the proportional distribution of time spent online per week among respondents who gave different ratings of physical health. For the majority of respondents who selected excellent, very good, good, and fair, they tended to use the Internet for 0-50 hours per week, with the largest proportion of respondents spending 15 hours per week on the Internet. However, we found that the proportion of respondents who indicated excellent physical health and access to the Internet for 55 hours per week is slightly greater than those who answered very good and good in terms of physical health. For those who classified themselves in the poor physical health, we noticed that the proportion of respondents with short Internet access hours was much lower than the previous levels, and the proportion of respondents with 50-100 hours of Internet access per week was significantly higher than the rest of the groups.

5.6 Time spent online vs. Mental Health

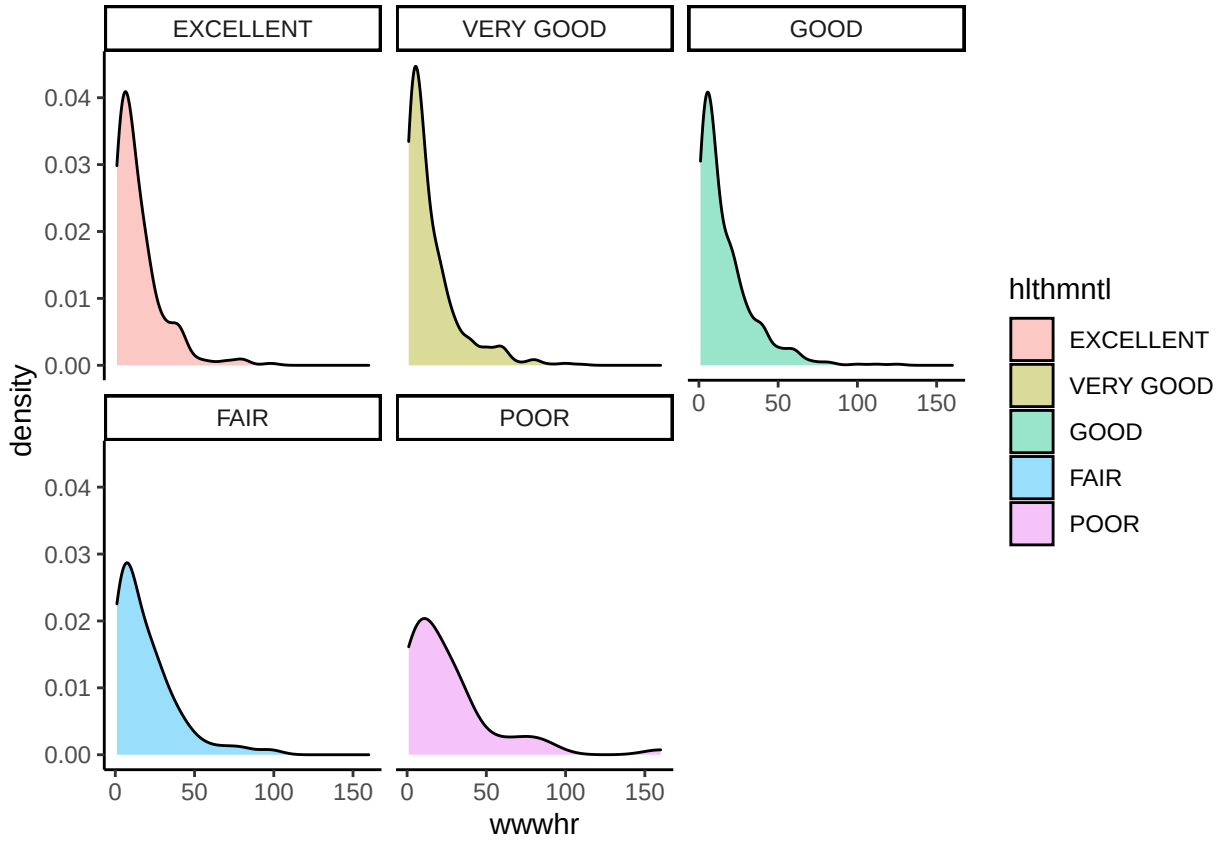


Figure 17: The largest proportion of respondents would like to spend 15 hours per week on the Internet

Figure 17 reflects the proportional distribution of weekly Internet access time among respondents with different levels of mental health. Similar to the prior, respondents with above or equal average mental health were more likely to use the Internet for 0-30 hours per week, with the largest proportion of respondents spending 15 hours online. However, for respondents who were accompanied by fair or poor mental health, the proportion of those who used the Internet for 0-30 hours was not as high as the previous groups, as the proportion of those who accessed the Internet for more than 30 hours was significantly higher than the other categories. The ratio of time spent online was significantly less for respondents who chose poor than for the other groups. Respondents with poor psychological status were considerably more likely to spend over 70 hours online compared to the other groups.

5.7 Time spent online vs. Education level

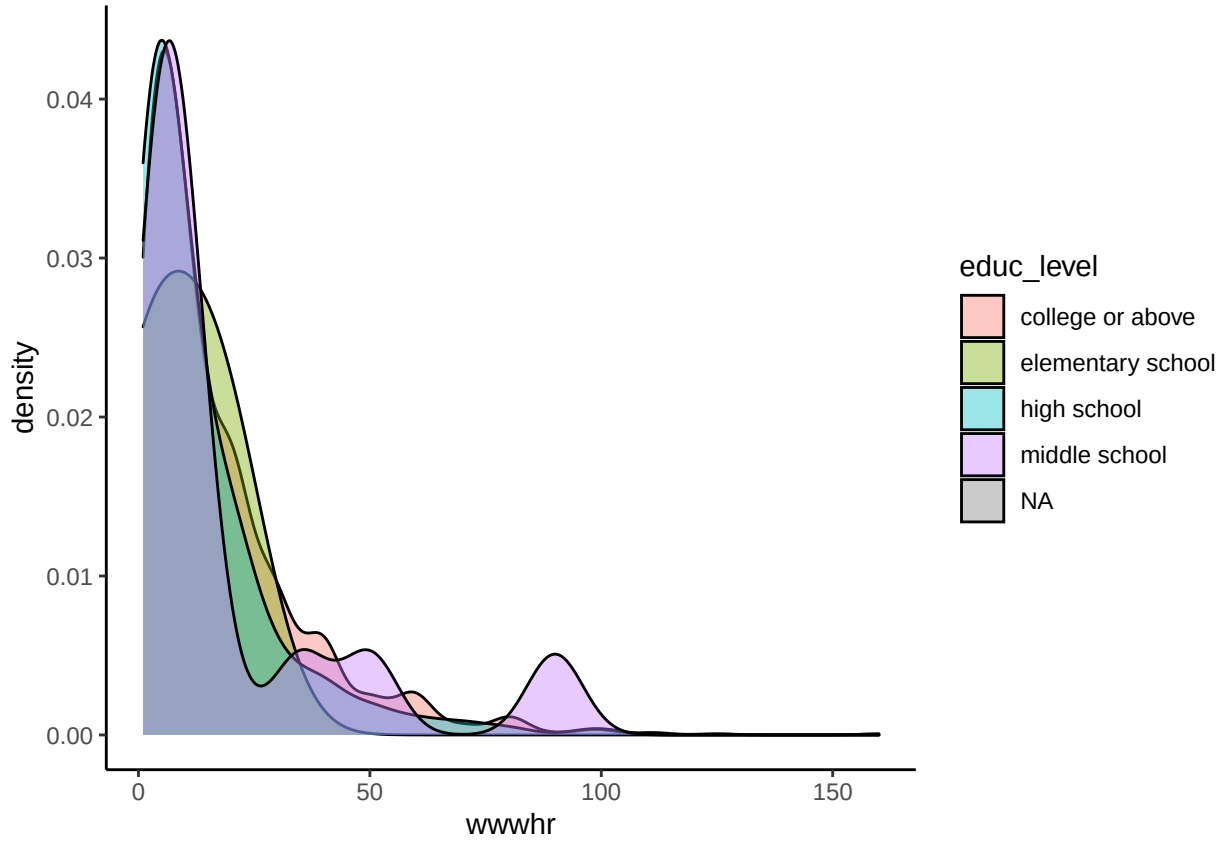


Figure 18: Respondents who have received high school education or college and above education have basically the same proportion distribution of time spent online per week

Figure 18 provides the percentage distribution of weekly time spent online by respondents with different education levels. It is noted that respondents who have received high school education or college and above education have basically the same proportion distribution of time spent online per week. For respondents with only middle school experience, most of them also tend to use the Internet for 0-30 hours per week. However, among the group that had only attended middle school, the proportions of those who used the Internet for 40-60 hours and 90-100 hours were significantly higher than that of respondents who had received high school education or higher. A narrower range of time spent online was observed for respondents who had only attended elementary school, with the largest number of respondents remaining on the Internet for 15 hours, but corresponding to a much lower proportion compared to respondents with other levels of education.

5.8 Time spent online vs. Quality of life

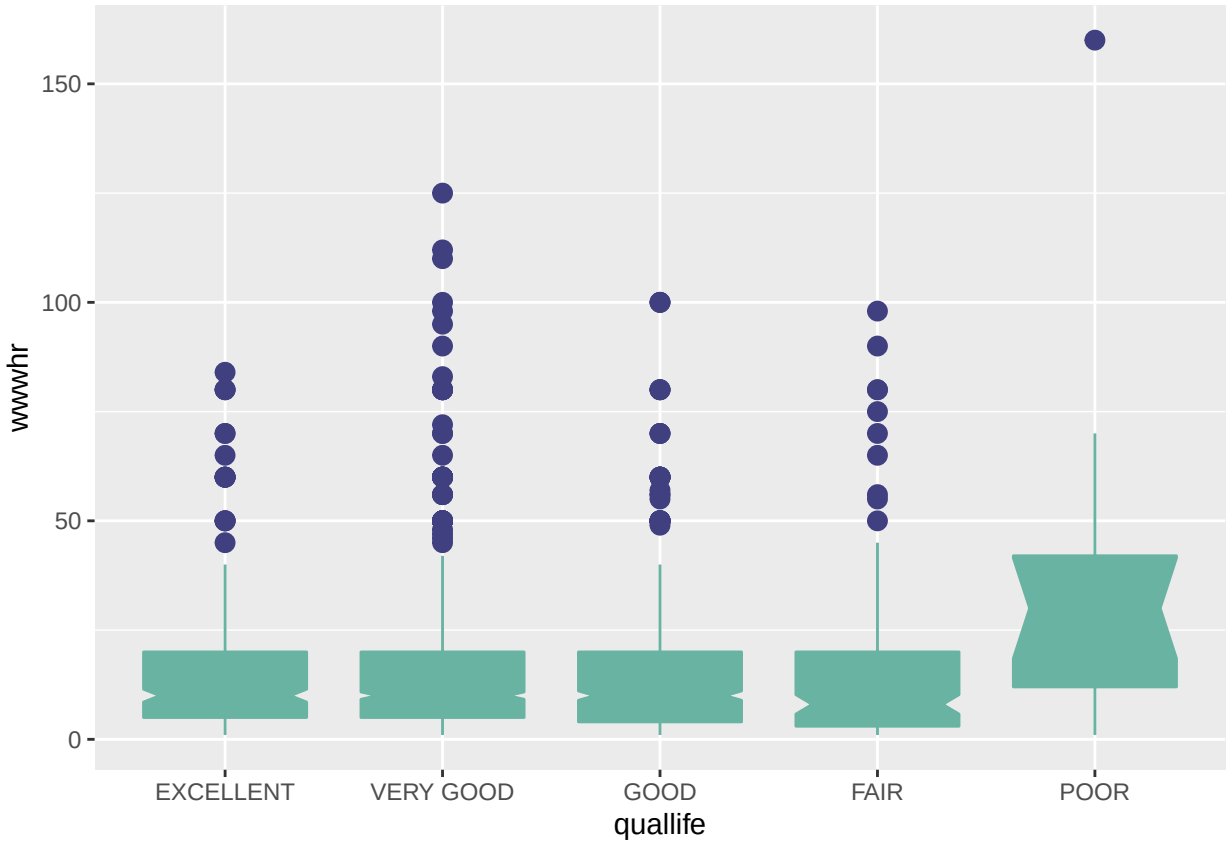


Figure 19: The distribution of time spent online is right-skewed for almost all age groups

Figure 19 displays the distribution of time spent online by respondents with different ratings of quality of life. The distribution of time spent online is similar for respondents who rated their quality of life as excellent, very good, good, and fair. The majority of these respondents spent between 5 and 20 hours online each week, and the median time spent on the Internet was close to 12 hours. There are many outliers in each of these groups, suggesting that most people access the Internet for a relatively short period of time, but there exist individuals who spend significant amounts of time online. However, for respondents who rated their quality of life as poor, the overall distribution of their online time was higher than the other groups. The median number of hours spent online per week for the respondents with poor quality of life even reached a striking 30 hours. Respondents with poor quality of life seemed to be more inclined to spend more time online.

6 Discussion

6.1 Age vs. Web hours

The analysis contains sample from age of 19 to 89, and from the linear regression model, age and the web hours has a negatively correlation, which indicates that the younger people would have longer time spending online(Canada 2017), people aged from 16 to 24 years has 89.9% of population who are internet users, which is the largest group of internet users. Previous research found that this could be caused by both parents and children believe they ca do better in school with a personal computer(Wang 2005). Thus, teenagers need to spending more time on websites for academic use, gathering information and entertainment.

6.2 Mental health vs. Web hours

The results from the above linear model has found that worse mental health could have longer time spending on websites. People who have VERY GOOD mental health have estimate of 1.675e-03, GOOD mental health have estimate of 1.257e-03, FAIR mental health have estimate of 2.640e-02 and people who have POOR mental health have estimate of 3.087e-02.

The data from the Health and Retirement Survey (HRS) in the United States from 2002 to 2008, Cotten et al. found that older adults were 33% less likely to be depressed when using the Internet(L. Xie 2021). The previous paper from 2021 has also found that the use of internet can improve people's mental health by enriching their society activity, expanding the social interactions and increasing the contacts with families, friends and social networks(L. Xie 2021).

6.3 Life quality vs. Web hours

There are many factors to define quality of life, such as personality, work environment, leisure, and social capital (Qian 2022). Nowadays, the use of internet closely relate and affect our lives. With dividing line of FAIR life quality, the previous linear model has found that people who have FAIR, GOOD and VERY GOOD life quality spend less time online, and people have POOR life quality could rather spend more time online. An article from 2022 has found that the internet for leisure purposes has a negative effect on risk perception and positive effect on internet addition. In other words, the more time people spending on the websites, then they would have more chance to have internet addition, thus they would have a lower quality of life. However, people who have POOR life quality has a positive impact on wwwhr, this could also explained by (Qian 2022), that more time people spending internet on leisure purpose, then they could be harder to live without the internet, and this could result them to have a worse life quality. Notably, life quality is usually determined by physical health and mental health. My model suggests that physical health does not have significant relationship with the time people spending on websites and mental health status have positive impact on the response variable, which lead to the hypothesis that life quality should also have positive impact on the response variable, which contradicts with my finding. Scientists found that older adults reported lower levels of depression than younger adults and used internet far less frequently as well. Additionally, within the older adult population, those who used the internet frequently were more depressed than other older adults who did not use the internet frequently(Boman 2021). The more time people spending online might lead to more chance of depression, which caused a lower life quality, and this could be caused by the use of internet decrease face to face interactions and activities. Especially for older population, the in person interaction is important for their life quality. People could have much less emotional interactions from networking online, which could seriously impact mental health for older population, and lead to higher chance of depression.

6.4 Limitation

First of all, we know that this questionnaire is only available in English and Spanish. However, as a country with many immigrants, the United States has populations from all over the world. Over 350 languages are spoken by people of different ethnicities. Although English and Spanish are the official languages, there are still some people who are not fluent in English or Spanish. In the dataset, we also observed that almost all respondents have completed high school education or higher, but only residents who are proficient in English and Spanish were able to read and fill out the survey. Those with lower education levels, who may have immigrated from another country, were not able to understand the contents of the questionnaire because of the language barrier. Therefore, the language limitation causes the questionnaire to be addressed to not all of the U.S. residents.

Secondly, we noticed that this questionnaire was too long that a lot of options were skipped by the respondents when they completed it. As a consequence, our data contained many missing values. in order to make our study proceed smoothly, we removed all the missing values in the data cleaning section, thus more than half of the data were removed. Although we derived a relationship between time spent online and several factors based on the available data, this is not representative because we are actually using only a small fraction of the data and the information reflected by them is not generalizable to the whole group.

6.5 Future improvements

Since we found that our exploratory variables contain many categorical variables, there is not necessarily a linear relationship between them and our response variable (time spent online). In the future, we may learn new models that can be more accurate in detecting the relationship between numerical response variable and categorical exploratory variables.

In addition to that, we cannot ensure the honesty of the respondents when they fill out the survey. Therefore, we may consider proactively obtaining the recorded respondent information from some relatively authoritative organizations in the next data collection process

7 Appendix

7.1 Respondent's age and education years distribution

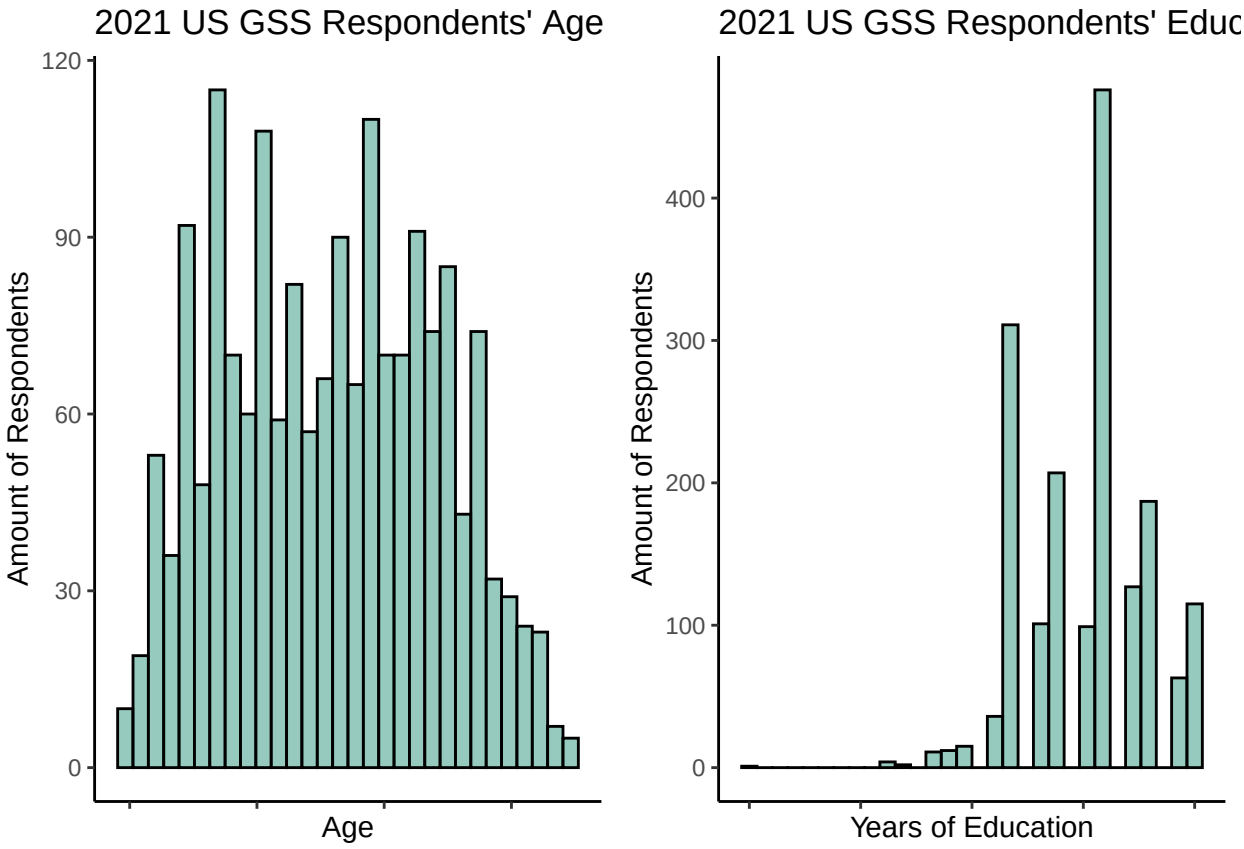


Figure 20: Respondent's age and education years distribution

7.2 wwwhr vs. sex

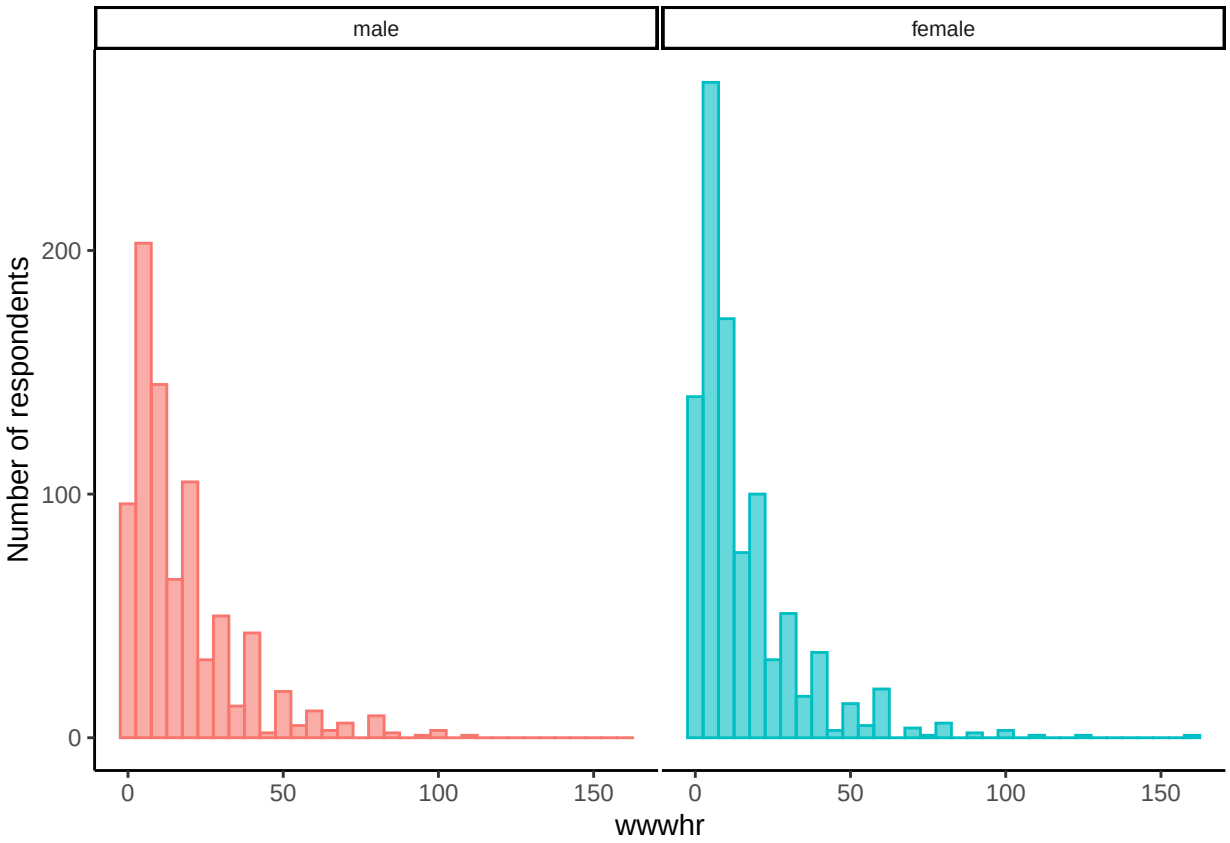


Figure 21: wwwhr vs. sex

7.3 wwwhr vs. quallife

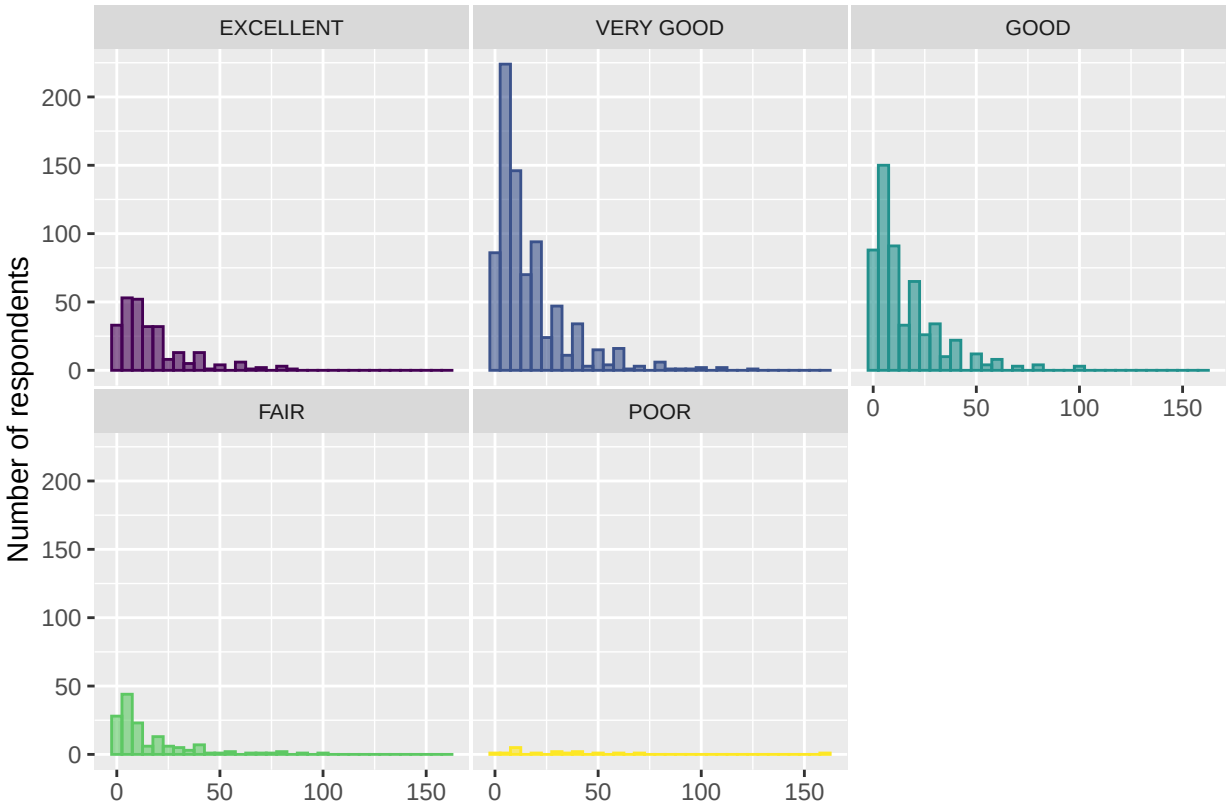


Figure 22: wwwhr vs. quallife

7.4 wwwhr vs.hlthphys

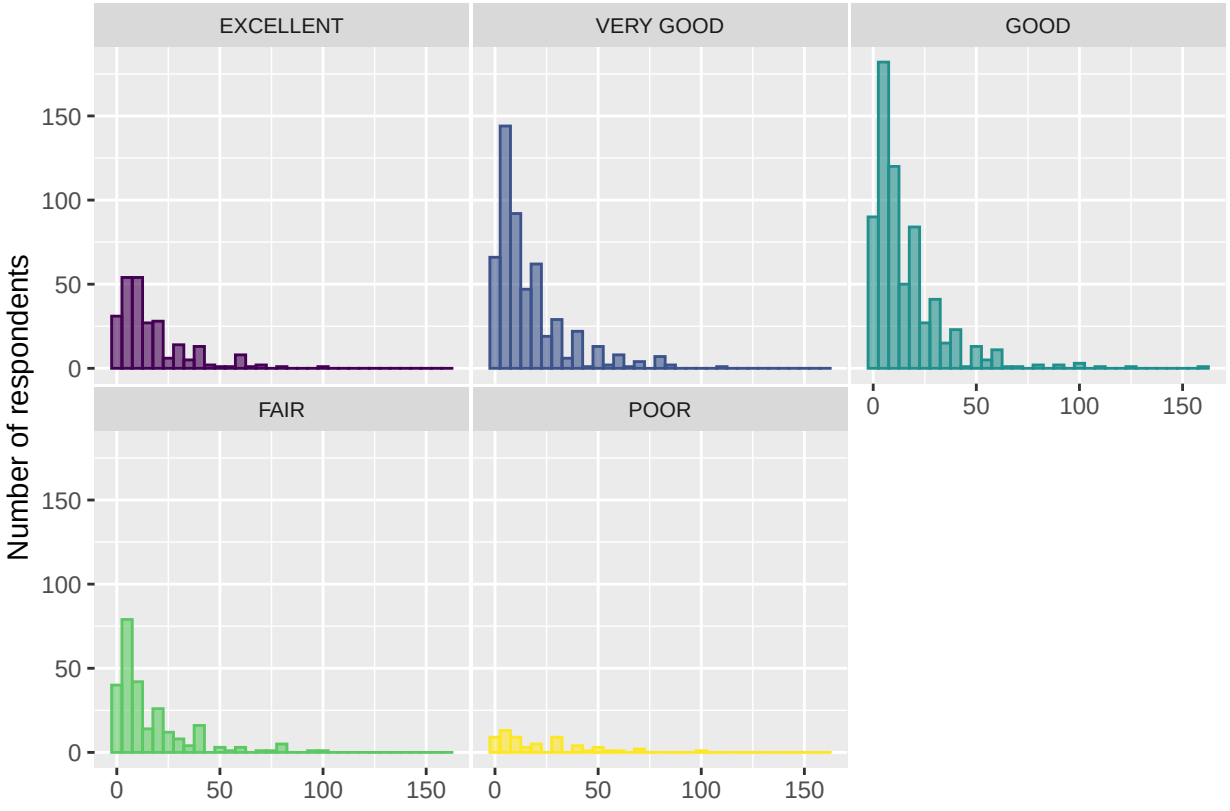


Figure 23: wwwhr vs.hlthphys

7.5 wwwhr vs. hlthmntl

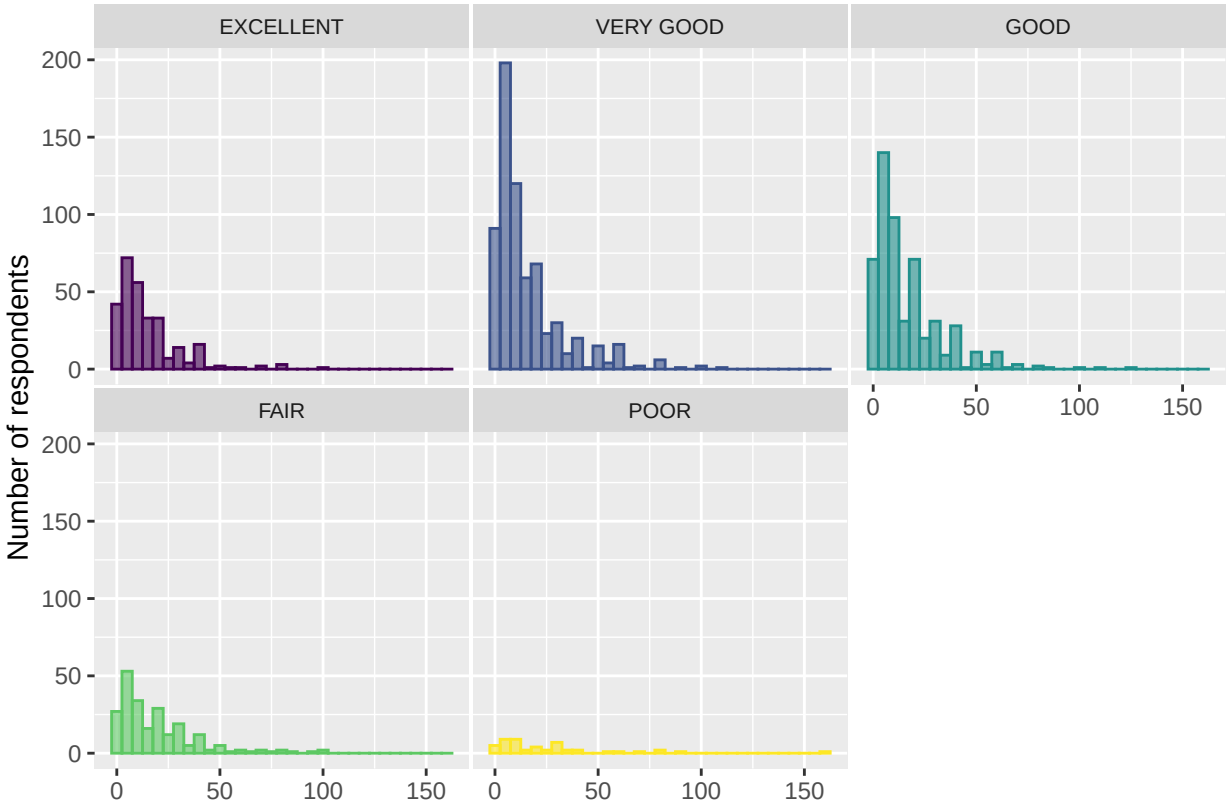


Figure 24: wwwhr vs. hlthmntl

7.6 wwwhr vs. educ_level

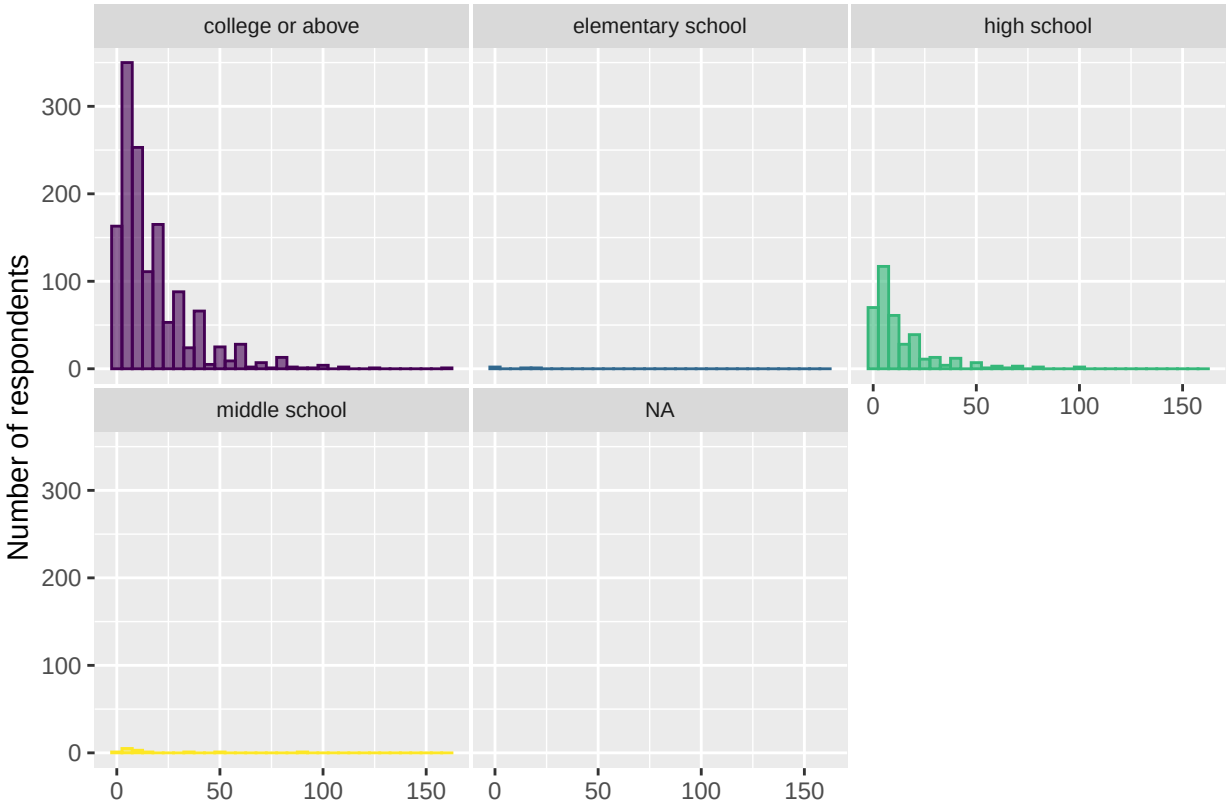
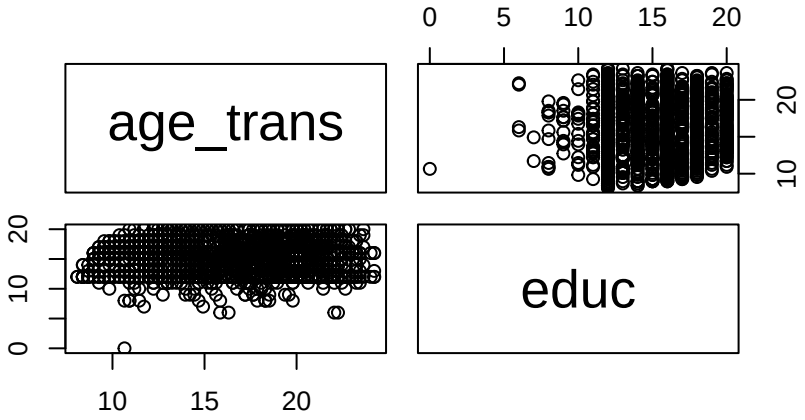
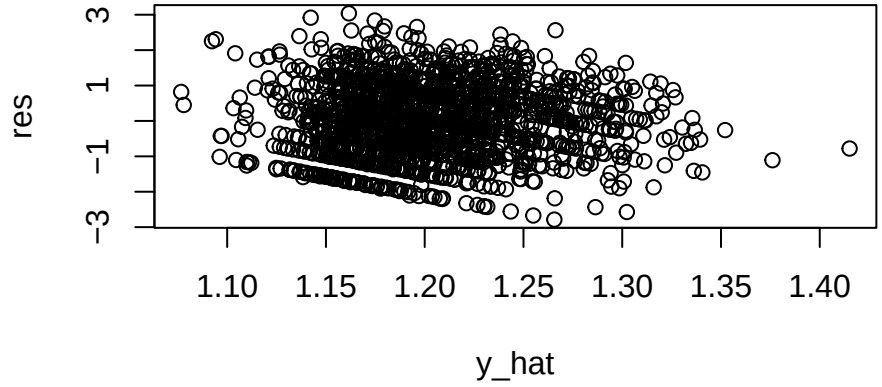
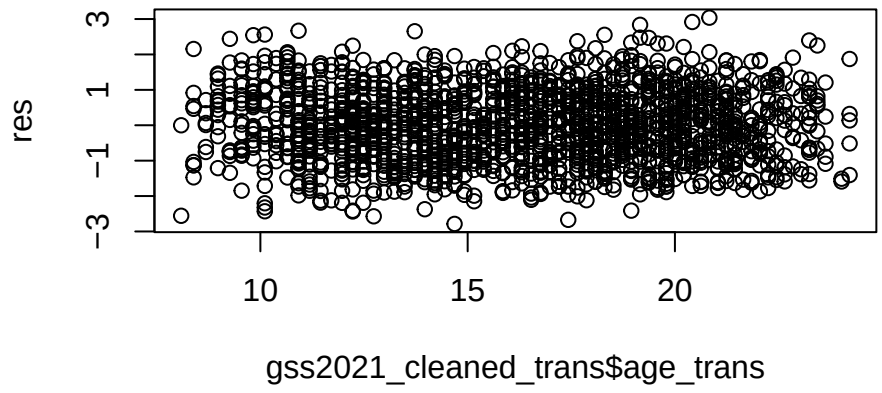


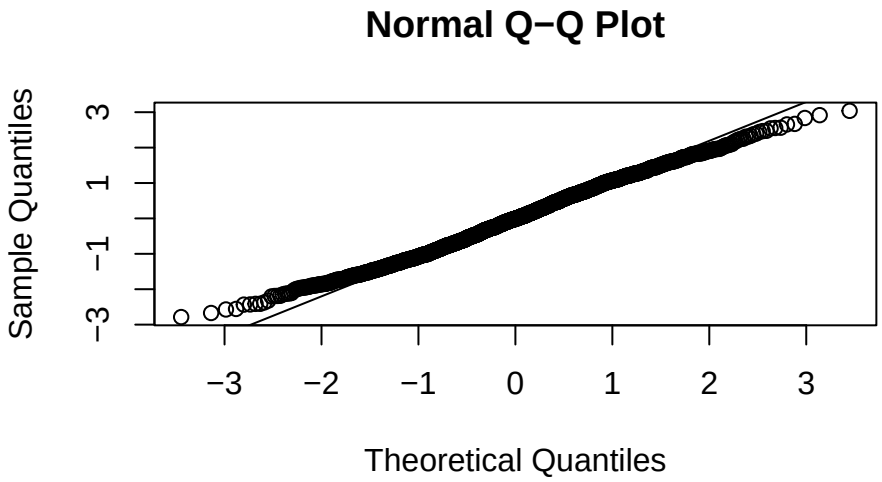
Figure 25: wwwhr vs. educ_level

7.7 residual plots






res
3
1
−1
−3
10
15
20
gss2021_cleaned_trans$age_trans


Normal Q−Q Plot
Sample Quantiles
3
1
−1
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Bibliography

Boman, Hamilton. 2021. "The Impact of Age and Internet Use on Depression." https://www.proquest.com/docview/2520840100?accountid=14771&parentSessionId=8El70OcPN0x9l%2FyJgz24J5XjBm%2FPvOWMhwmmlkUC06U%3D&pq-origsite=primo.

Canada, Statistic. 2017. "Internet Use by Frequency of Use, Age Group and Sex." https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2710001801.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2021. *viridis - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679424.

Hughes, Matthew. 2019. "Study Shows We're Spending an Insane Amount of Time Online." https://thenextweb.com/news/study-shows-were-spending-an-insane-amount-of-time-online.

Iera, Antonio. 2012. "The Social Internet of Things (SIoT) – When Social Networks Meet the Internet of Things: Concept, Architecture and Network Characterization." https://doi.org/10.1016/j.comnet.2012.07.010.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Qian, Bo. 2022. "Internet Use and Quality of Life: The Multiple Mediating Effects of Risk Perception and Internet Addiction." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835165/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rudis, Bob. 2020. *Hrbrthemes: Additional Themes, Theme Components and Utilities for 'Ggplot2'*. https://CRAN.R-project.org/package=hrbrthemes.

Wang, Rong. 2005. "Teenagers' Internet Use and Family Rules: A Research Note." https://www-jstor-org.myaccess.library.utoronto.ca/stable/3600310?sid=primo&seq=1.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Maximilian Girlich. 2022. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, and Evan Miller. 2021. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Xie, Lin. 2021. "Does the Internet Use Improve the Mental Health of Chinese Older Adults?" https://www.frontiersin.org/articles/10.3389/fpubh.2021.673368/full.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Yamazaki, Kazuki. 2021. "What Are the Top Languages Used in the u.s.?" https://telelanguage.com/blog/top-languages-usa/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.