Potential factors that affect alcohol consumption*

A analysis of data from the 2016 General Social Survey on Canadians at Home and Work

Qingya Li

20 March 2022

Abstract

The paper inputs the data from the 2016 General Social Survey (GSS) on Canadians at Home and Work. By conducting graphs and models, we found alcoholic beverages are positively associated with the individual's sex and highest education level. The findings may help the researchers study the individual motivation for drinking and enhance the persuasiveness of other alcohol studies.

Contents

*Code and data are available at: https://github.com/QingyaLi/Potential-factors-that-affect-alcohol-consumption.git.

1 Introduction

The use of alcohol, cigarettes, cannabis, and other drugs has increased in recent years raised people's concerns. The paper intends to study the factors that influenced individual alcohol consumption in Canada. We imported the data from the 2016 General Social Survey (GSS). Specifically, we focused on the relationship between alcohol consumption and gender, age, education level. We predicted males, older, and low-educated individuals consume more alcoholic drinks.

The paper includes three main sections:2, 3 and 4. In 2, we cleaned the dataset and only included the variables we needed. We perform exploratory data analysis (EDA) for each variable along with linear regression models. In 3, We found respondents drank 2-3 times alcohol in a month. Males consumed 0.468995 more alcohol than females on average. Finally, we discussed the weakness of the survey design. To improve, we included additional survey questions with Google form.

It is important to hold such a study since it can help us know more about alcohol consumption behaviors in society. The findings can also help with studying the personal motives for drinking and increase the persuasiveness of the relevant alcohol research. Last but not least, provide a pathway for prevention and treatment efforts.

2 Data

2.1 Data Source

The GSS program designed the telephone survey among the ten provinces of Canada to collect cross-sectional data for trend analysis (S. Canada 2018b). There are two principal objectives of the program 1) provide background information, estimate the social trends from the data, and predict Canadian's living conditions and well-being over time 2) include the information about the social policy problems (S. Canada 2018b). To meet the goals, the GSS dataset contains 1) core contents, which are the social change regarding living conditions and well-being, could be used to estimate the effectiveness of specific policy 2) classification variables such as gender, age, and income which can identify the population groups (S. Canada 2018b).

2.2 Methodology and Data Collection

The 2016 GSS on Canadians at Home and Work was collected between 2016.8.2-2016.12.23 (S. Canada 2018b). The population of the data frame consisted of all the non-full-time institutional people who were over or equal to 15 years old in Canada, excluding the Yukon, Northwest Territories, and Nunavut residents (S. Canada 2018b). Aiming to find the sample, the researchers divided ten provinces into different geographic areas. Numerous of the Census Metropolitan Areas (CMAs) such as Toronto, Montreal, Quebec City, Ottawa, Vancouver was considered as separated strata (S. Canada 2018b). The rest of the CMAs were grouped into three strata (S. Canada 2018b). The non-CMAs were grouped into 27 strata (S. Canada 2018b). The sample was randomly selected by household telephone number. Their telephone information was collected by two components, including 1) a list of the phone number (landline and cellular) from a new frame that created by statistics Canada in 2013, and 2) the address register (AR) which refers to a list of addresses among the ten provinces (S. Canada 2018b). The AR combines all the telephone numbers with the significant address (S. Canada 2018b). When multiple phone numbers were associated with one address, the primary phone number would be considered the best response phone number (S. Canada 2018b). The selected respondents need to have a telephone and at least one person who was 15 years of age or older in the household to be considered eligible (S. Canada 2018b). To determine the eligibility, the respondents need to answer several questions before the survey (S. Canada 2018b). Then respondents can decide to complete the survey by either electronic questionnaires (EQ) or phone interviews (CATI) (S. Canada 2018b). The interviews were conducted from 9:00 a.m. - 9:30 p.m. from Mondays to Fridays, 10:00 a.m. to 5:00 p.m. on Saturdays, and 1:00 p.m. to 9:00 p.m on Sundays, in the five Statistics Canada offices: Halifax, Sherbrooke, Sturgeon Falls,

Winnipeg, and Edmonton (S. Canada 2018b). The interviewers would introduce the interview and randomly ask one of the members in the household to answer the questions (S. Canada 2018b). If the respondents refuse to participate in the survey, the interviewers would explain the interview again to encourage them to participate (S. Canada 2018b).

2.3 Data Characteristics

The paper was written by the **R** statistical language (R Core Team 2020), the packages of **tidyverse** (Wickham et al. 2019) and **dplyr** (Wickham et al. 2021). **RMarkdom** (Allaire et al. 2020) and **bookdown** (Xie 2021a) were used to output the results. The bar plots were created by **ggplot2** (Wickham 2016) and tables were created by (Xie 2021b) and **tidy** function from (Silge and Robinson 2016). We used the ANOVA test to choose the most appropriate model to estimate the relationship between drink and the predictor variables.

Overall, the dataset included 19,609 observations and various variables. The dataset is accessible at the GSS dataset through the UofT library. The dataset was in CSV format, we only downloaded demographic and household composition derived variables; drinking; education highest degree block v.3. We imported the dataset with **read_csv** function and saved later with **write_csv** function. In the cleaning process, we first created a new data frame called *df*. Since we intended to find how gender, age, and education levels influenced respondents' alcohol consumption. We used the **select** function to select the variables (CASEID, sex, agegr10, ehg3_01, drr_110) that we are interested in analyzing. Sex refers to the gender of respondents, agegr10 refers to the age group of respondents, ehg3_01 refers to the highest education levels of respondents, and drr_110 refers to alcohol consumption of respondents in the past month. We used use **rename** function to rename "CASEID" as "id", "agegr10" as "age", "ehg3_01" as "edu", "drr_110" as "drink". Then we created another new data frame called *df1* that we mutated the variables from double numbers into categorical as follows regarding the codebook (S. Canada 2018a). It is more convenient for us to plot the data:

sex = male ~ 1, = female ~ 2, = valid skip ~ 6, = don't know ~ 7, = refusal ~ 8, = not stated ~ 9

age = 15-24 ~ 1, = 25-34 ~ 2, = 35-44 ~ 3, = 45-54 ~ 4, = 55-64 ~ 5, = 65-74 ~ 6, = valid skip ~ 96, = don't know ~ 97, = refusal ~ 98, = not stated ~ 99

edu = less than high school diploma ~ 1, = high school diploma ~ 2, = trade certificate ~ 3, = college diploma ~ 4, = university diploma below bachelor's degree ~ 5, = bachelor's degree ~ 6, = university diploma above bachelor's degree ~ 7, = valid skip ~ 96, = don't know ~ 97, = refusal ~ 98, = not stated ~ 99

drink = everyday ~ 1, = 4-6 times a week ~ 2, = 2-3 times a week ~ 3, = once a week ~ 4, = once or twice in the past month ~ 5, = not in the past month ~ 6, = never drink ~ 7, = valid skip ~ 96, = don't know ~ 97, = refusal ~ 98, = not stated ~ 99

2.4 Figures

Figure 1-4 demonstrates the distributions of the selected variables. Figure 1 displays the distribution of survey respondents' sex. The female and male respondents were around 10,000 and 8,000, respectively. Our results of respondents' age, grouped by 10 in figure 2 followed a normal distribution. The plot is left-skewed where the mode falls into the 55-64 years of the age column. 4,040 respondents were at the age of 55-64 years. Figure 3 indicates the distributions of respondents' highest education level. Most respondents had a high school diploma (4,739), followed by a college certificate and bachelor's degree diploma. Figure 4 answered the question of "how often did you drink alcoholic beverages in the past months?" The options included from every day, 4-6 times a week, 2-3 times a week, once a week, once or twice in the past month, not in the past month, never drink, etc. There were 4,712; 3,145; 3,117 respondents who claimed they drink once or twice in the past month, not in the past month, and 2-3 times a week, respectively.

The results of Figures 5-7 explore the factors that affect the respondents' alcohol consumption. For example, does male/older/higher education respondents consume more alcohol than female/younger/low education respondents? Figure 5 illustrates respondents' alcohol consumption grouped by gender. We observed the female respondents' alcohol consumption is slightly higher than males except for the everyday column. Gender may be the factor that affects drinking, but further analysis is needed. Since the reason might be the number of females is higher than the male. The results of figure 6 show the distribution of alcoholic drinks consumption grouped by age. We could see that alcohol consumption among the different age groups is most likely the same. The group of 15-24 years of age consumed less than other groups. We believe the reason might be the minimum legal drinking age in Alberta, Manitoba, and Quebec is 18 years old; and 19 years old in the rest of the provinces and territories in Canada. Figure 7 displays the consumption of alcoholic beverages grouped by highest education level. We noticed the frequency of high school diploma respondents consuming drinks is generally higher than other education level respondents. And the frequency of university diplomas below bachelor's degree, in general, is lower than other education level respondents. The respondents who have less than a high school level prefer never to drink.

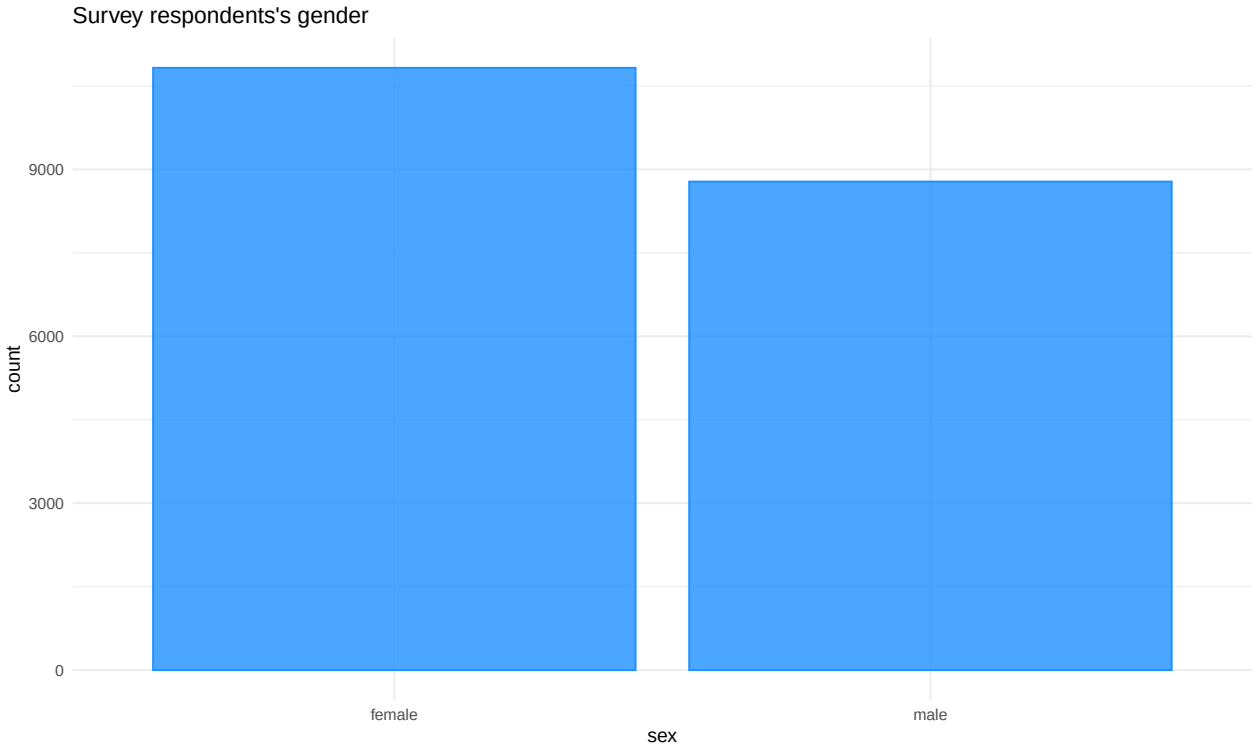


Figure 1: Distribution of survey respondents gender

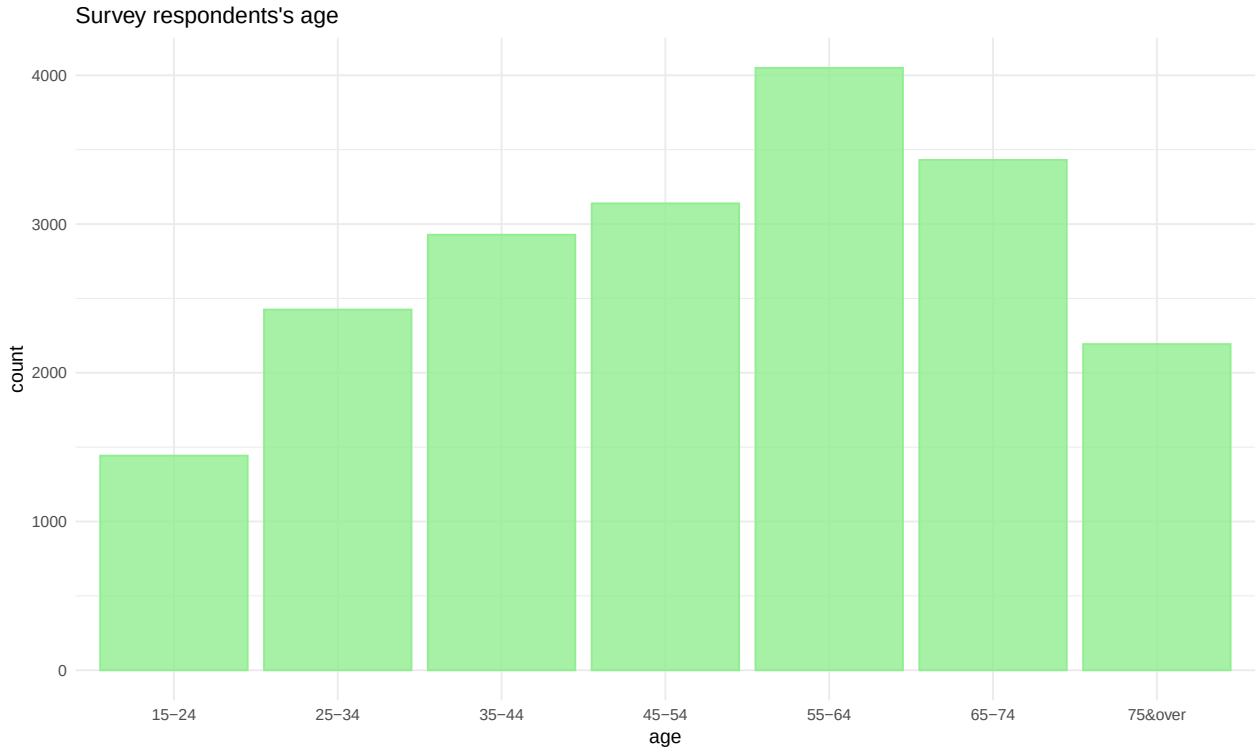


Figure 2: Distribution of survey respondents age, group by 10

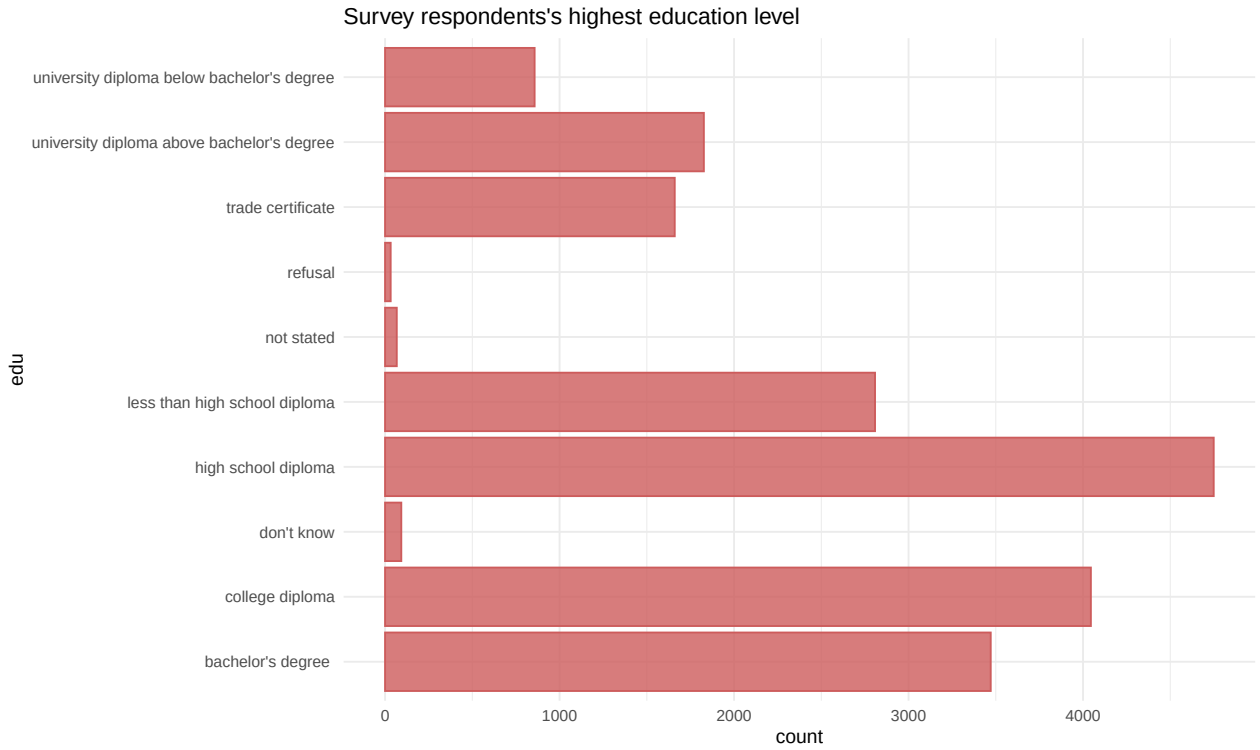


Figure 3: Distribution of survey respondents highest education level

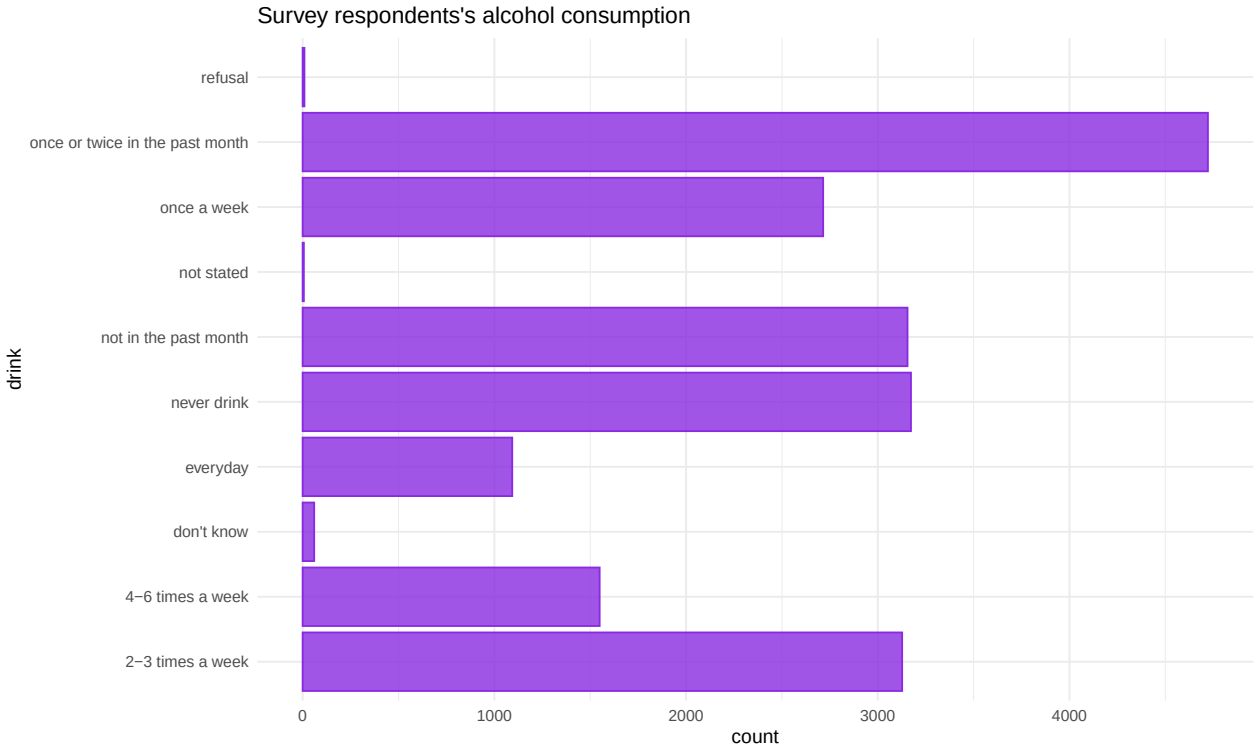


Figure 4: Distribution of survey respondents alcohol consumption

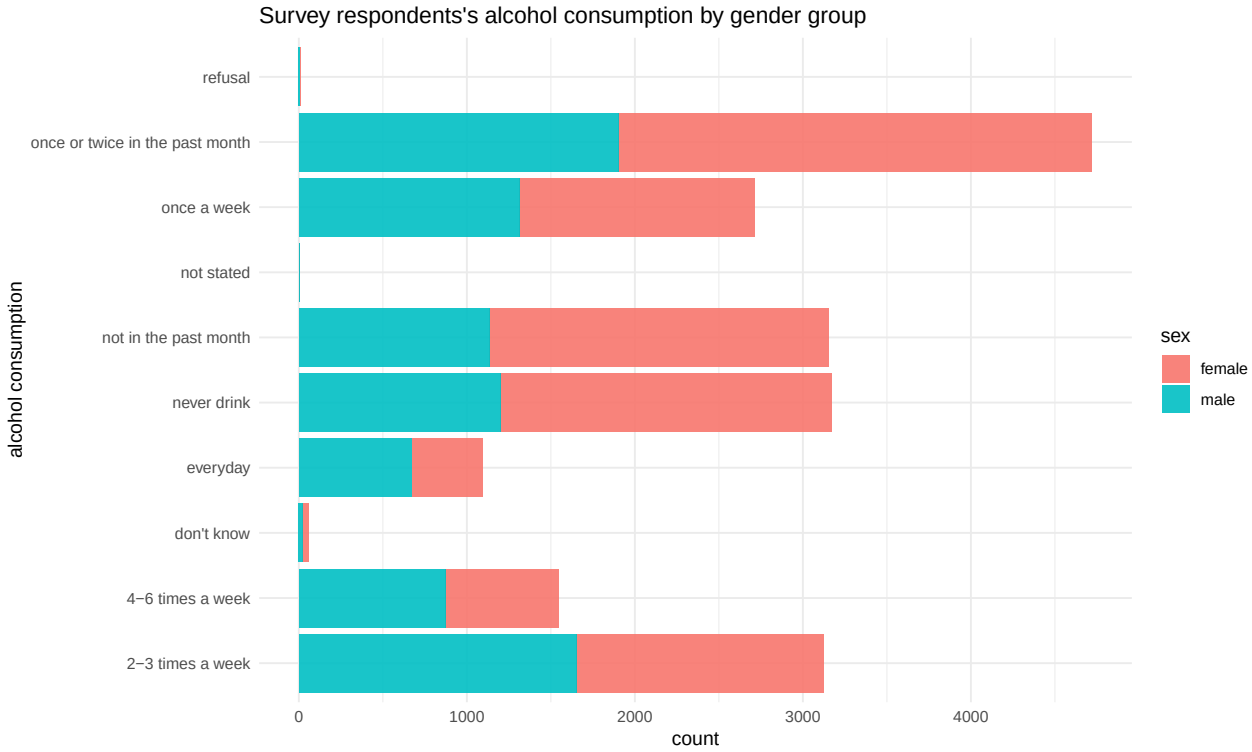


Figure 5: Survey respondents alcoholic beverages consumption in the past month by gender group

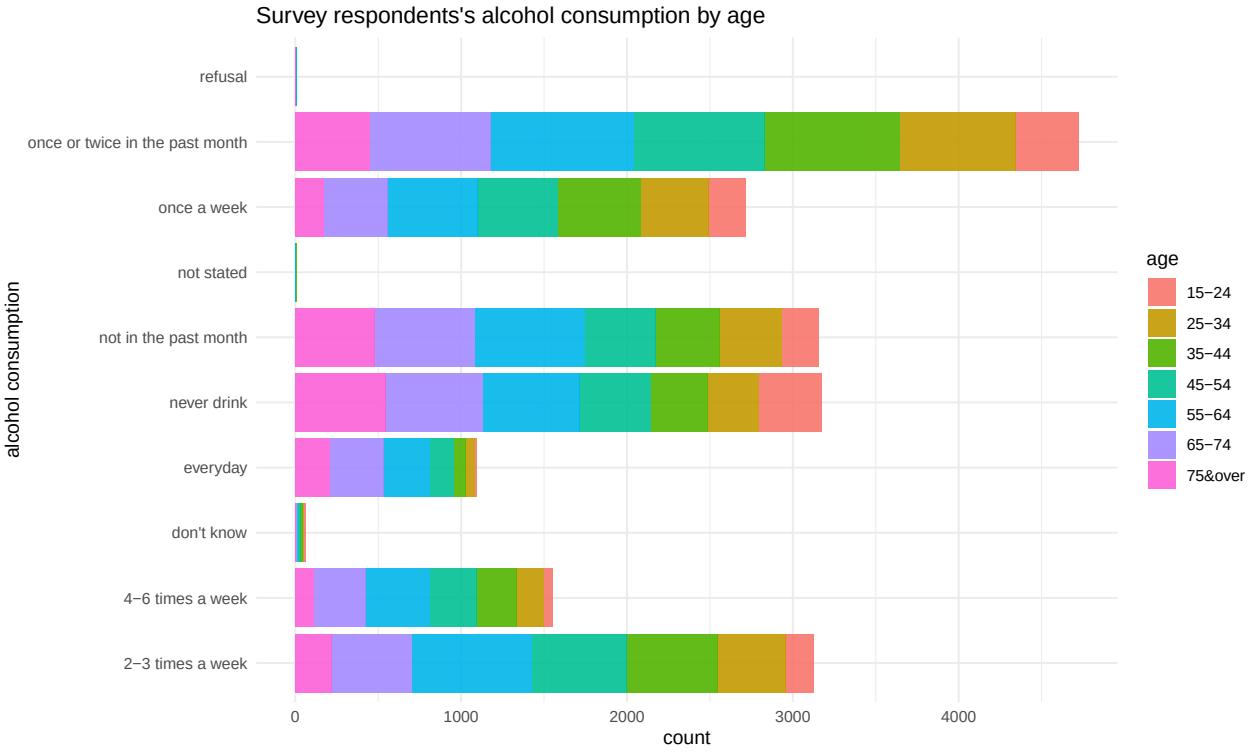


Figure 6: Survey respondents alcoholic beverages consumption in the past month by age group

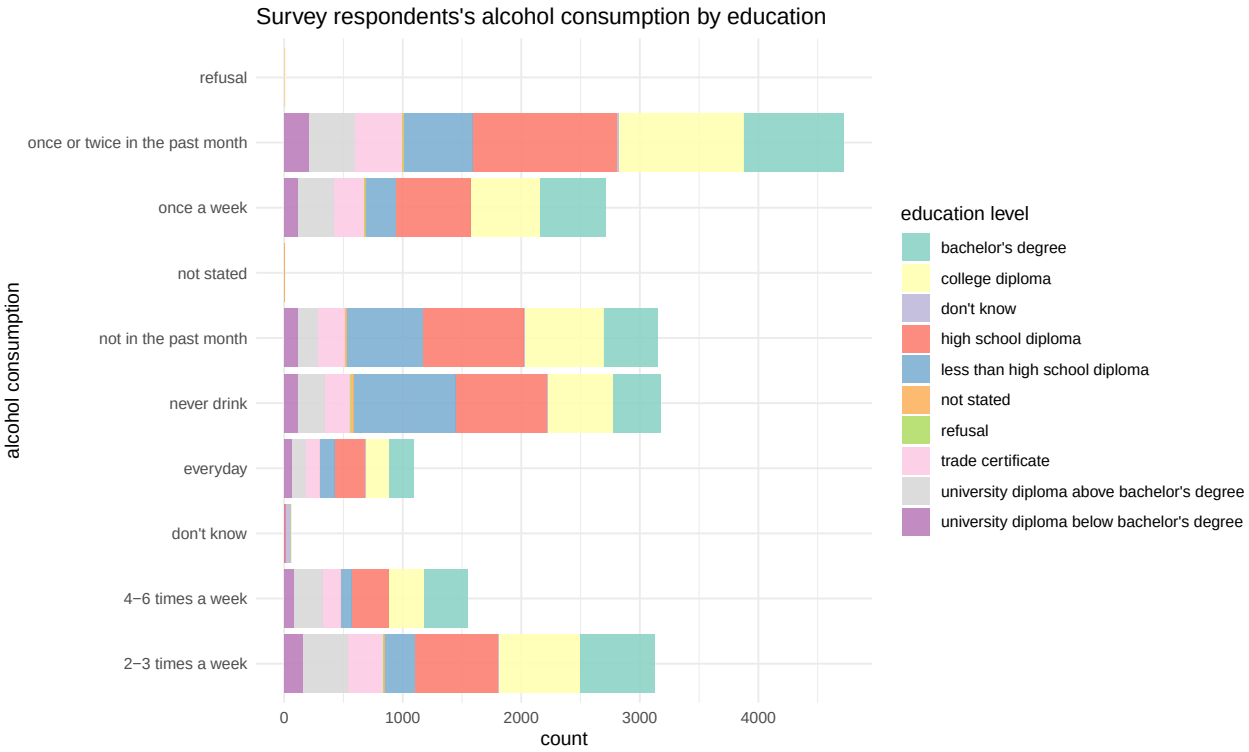


Figure 7: Survey respondents alcoholic beverages consumption in the past month by education level

Table 1: Summary table for mod1

term	estimate	std.error	statistic	p.value
(Intercept)	3.1673616	0.1565190	20.236279	0.0000000
sex	0.4759260	0.0787760	6.041508	0.0000000
age	-0.0334824	0.0221068	-1.514575	0.1298963
edu	0.2516906	0.0041286	60.962859	0.0000000

Table 2: Summary table for mod2

term	estimate	std.error	statistic	p.value
(Intercept)	3.0339147	0.1293684	23.451741	0
sex	0.4689946	0.0786456	5.963395	0
edu	0.2519261	0.0041258	61.061206	0

2.5 Models

The model is shown below which consist with three all explanatory variables (sex, age, highest education level) and the response variable is the alcoholic drink consumption.

$$y_{idrink} = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 edu + \epsilon_i$$

- y_{idrink} is survey respondents' alcohol consumption (in the past month)
- β_0 is the grand mean of alcohol consumption in the past month
- $\beta_s ex$ are fixed effects for respondents' sex
- $\beta_a ge$ are fixed effects for respondents' age
- $\beta_e du$ are fixed effects for respondents' highest education level
- ϵ_i is the error term that $\epsilon_i \sim N(0, \sigma^2)$

3 Results

The models were conducted with *df* dataset and **lm** function. mod1 is the full model, mod2 contains with two explanatory variables (sex, higest education level), and mod3 only included variable of sex. We applied the **ANOVA** test between mod2 and mod3; mod2 and mod1, to see which model is the most appropriate.

After applying ANOVA test, we chose mod2 is the most appropriate model. The p value between mod2 and mod3 is $<$ 2e-16 which is less than 0.05; the p value between mod2 and mod1 is 0.1298963. The small p value means that we have strong evidence to against the null hypothesis that there is no difference between simpler model and more complex model. Thus, we think the education term is significant but age is not. The final model is mod2 as shown below:

$$y_i drink = 3.033915 + 0.468995 sex + 0.251926 edu + 5.476 \quad (1)$$

Table 3: Summary table for mod3

term	estimate	std.error	statistic	p.value
(Intercept)	4.1055930	0.1398260	29.362168	0
sex	0.5247873	0.0857903	6.117092	0

Table 4: The ANOVA test between mod2 and mod3

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
19606	587929.1	NA	NA	NA	NA
19607	699735.5	-1	-111806.4	3728.471	0

Table 5: The ANOVA test between mod2 and mod1

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
19606	587929.1	NA	NA	NA	NA
19605	587860.3	1	68.7842	2.293937	0.1298963

Equation (1) shows the average of respondents' alcoholic beverages consumption in the past month is 3.033915 (2-3 times a week). An 0.468995 increase in alcohol consumption when the sex is male. An 0.251926 increase in alcohol consumption when education level is higher. The standard error of residual is 5.476.

4 Discussion

4.1 First discussion point

In the paper, we conducted EDA to find the distribution of sex, age, education level, alcoholic drinking consumption as well as grouped by these three variables. Moreover, we conduct three linear regression models using alcoholic consumption as response variables and the rest were predicted variables.

4.2 Second discussion point

The 2016 GSS on Canadians at Home and Work illustrates the proportion of female respondents was slightly higher than males. The distribution of respondents' age indicates the aging population issue in Canadian society. Another statistic report from Statista depicted 15.9%, 67.65%, and 16.46% of the overall population in Canada were 0-14 years of age, 15-64 years of age, and over 65 years of age, respectively in 2016 (Statista 2022). The proportion over 65 years of age increasing in recent years and reached 18.1% in 2020 (Statista 2022). The distribution of highest education level depicts the proportion of high school diploma is the highest. Other researches pointed out that over 54% of Canadians aged 25-64 years had either a university or college diploma in 2016 (Canada 2017). The alcohol consumption distribution indicates most respondents reported they drink once or twice in the past month, followed by 2-3 times a week, not in the past month, and never drink. Figures 5-7 predict the relationship between alcoholic beverages consumption and sex, age, highest education level. The model we chose proved individual sex and highest education level positively correlated to alcohol consumption.

4.3 Third discussion point

There is some relevant research that studied alcohol consumption. The researchers created an informal survey, including the question to ask how many alcoholic drinks did you consume last weekend (Paul Roback 2021). Similarly, they fitted generalized (Poisson) regression models with predict variables off campus (the students live on campus or off campus) and sex (Paul Roback 2021). They found off-campus students' alcohol consumption is 0.8976 higher than on campus (Paul Roback 2021). Also, males' alcohol consumption is 1.1154 higher than females (Paul Roback 2021).

Besides, we should consider including other factors that affect alcoholic consumption such as income, ethnicity, and mental health disorders. For example, publications of the government of Canada reported that

increased alcohol consumption with income and specifically significant performance in males. 3.5 times higher among white Canadians than among South Asian, East/Southeast Asian, and Arab/West Asian Canadians (P. H. A. of Canada 2018).

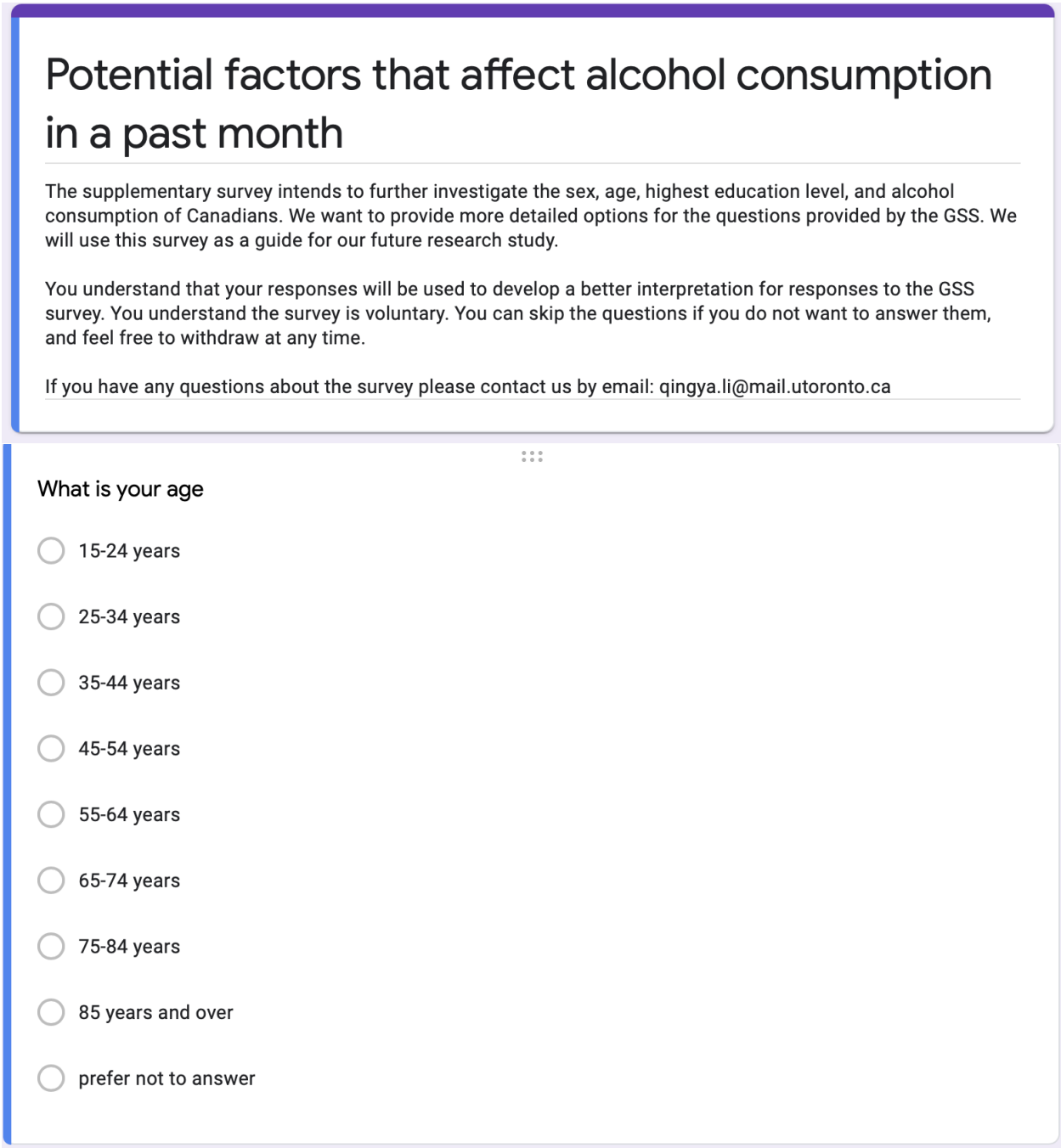
4.4 Weaknesses and next steps

There were some limitations in the survey. First, the survey only required one person from each selected household to complete the questions. Bias may exist because the respondents cannot represent the opinions of other family members. Second, the options for the survey questions could be further refined. For example, the options of sex questions can also include transgender, gender-neutral, non-binary, agender, and so on. Therefore, we include an additional survey link in the 4.4.

Appendix

.1 Supplementary survey

Our supplementary survey is available here: https://forms.gle/f7f4mBsjq5QVU4as8



what is your sex

- male
- female
- transgender
- gender neutral
- non-binary
- agender
- pangender
- genderqueer
- two-spirit
- third gender
- all, none or a combination of these
- prefer not to answer

what is your highest education level

- ○ no certificate
- ○ less than high school diploma
- ○ high school diploma
- ○ trade certificate
- ○ college diploma
- ○ bachelor's degree diploma
- ○ master's degree diploma
- ○ phd's degree diploma
- ○ prefer not to answer
- ○ 其他...

How often do you drink in the past month

- 7 times or more a week
- 4-6 times a week
- 2-3 times a week
- once a week
- 2-3 times in the past month
- once a month
- did not drink in the past month
- never drink
- prefer not to answer
- 其他…

How much money do you spend on alcohol in the past month

- ○ $100 and over
- ○ $80-99
- ○ $60-79
- ○ $40-59
- ○ $20-39
- ○ $1-19
- ○ none
- ○ prefer not to answer
- ○ 其他...

What type of alcoholic drinks did you consume in the past month

- [ ] Beer
- [ ] Cider/cooler
- [ ] Wine
- [ ] Spirits
- [ ] prefer not to answer
- [ ] 其他…

Thank you for taking your time to complete the survey, your question will be recorded. Do you have any questions about the survey?

简短回答文本

Thank you for your participation!

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

Canada, Statistics. 2017. "The Daily-Education in Canada: Key Results from the 2016 Census." https://www150.statcan.gc.ca/n1/daily-quotidien/171129/dq171129a-eng.htm.

———. 2018a. *General Social Survey (Canadians at Work and Home) 2016 Main Code Book (Excludes Sports and Culture Sections) Public Use Microdata File*. Statistics Canada. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss30/gss30/more_doc/GSSC30EN_cbk_person.pdf.

———. 2018b. *General Social Survey Cycle 30: Canadians at Work and Home*. Statistics Canada. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss30/gss30/more_doc/GSSC30ENgid.pdf.

Canada, Public Health Agency of. 2018. "Inequalities in High Alcohol Consumption in Canada." https://publications.gc.ca/collections/collection_2018/aspc-phac/HP35-106-4-2018-eng.pdf.

Paul Roback, Julie Legler. 2021. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. CRC Press. https://doi.org/https://doi.org/10.1201/9780429066665.

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

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Statista. 2022. "Age Distribution in Canada 2010-2020." https://www.statista.com/statistics/266540/age-distribution-in-canada/.

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

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. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021a. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://github.com/rstudio/bookdown.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.