Investigation on Potential Demographic Variables Affect Alcohol Consumption in Canada, 1979[*]

Yiying Chen

27 April 2022

Abstract

As alcohol abuse becomes a raising problem in the society, it has been closely related to mental and physical health. Exploring the reasons behind alcohol intake was critical, which helps to improve the efficiency of health promotions. The National Study regard to alcohol consumption published by ODESI gives us an exhaustive database from various perspective. Based on this study, we extend further from demographic perspective to investigate how personal background affect people' intake of alcohol. Linear regression was conducted to construct correlation of each variable statistically. At the end of analysis, we also indicate drawbacks of this study and give advice on further improvements. Key words: Alcohol Consumption, Demography, Health Promotion, Data Visualization, Linear Regression Model, Model Optimization and Selection

Contents

[*] Code and data avaliable at https://github.com/YiyingChen0523/Final__Paper-Alcohol-Consumption-.git

Introduction

Alcohol intake was always considered as a key standard of assessing health status. Suitable amount of alcohol intake could help to release pressure; however, the excessive amount of alcohol will cause serious alcohol addiction and other diseases. This paper will be based on Alcohol Consumption in Canada, a National Study, 1979, which was published on ODESI (Archive 2022). The study that focuses on factors behind alcohol consumption was meaningful, it stimulates public awareness of drinking behavior and help people to understand reasons behind alcohol addiction.

In this paper, we mainly analyze demographic variables, including their living regions, sex, level of education, tenure and other variables related to their life. Those are background information of individuals, largely affects people' attitude toward life. According to their personal background and their drinking habit, we would be able to reflect on the current situation and know a clear direction for future health promotions.

The data was collected through face-to-face interviews, which includes family, cultural background, economics, traits of the respondents. The result was influential, more than 2000 people between age 15-90 was interviewed across various provinces (Limited 1979). Due to the methodology of the data collection, the data collected was effective and reliable. Face-to-face enables investigators to obtain more information and relevant details.

This paper will construct statistically by linear regression model to analyze the correlation between alcohol intake and individual's personal background, distinguish the different levels of importance. Since most of the variables are numerical, we select most correlated variables from the dataset, then constructing a linear model to compare their importance. The final model from linear regression filtered the most important variables that contribute to the alcohol consumption in Canada: sex and age of respondent, number of children per household, income of individual, tenure and whether knowing someone with drinking problem. Other statistical graphs including scatterplot and histogram will also be included to visualize them more easily. At the end of the paper, we will develop extensions about future health promotion based on our result.

Data

This paper was based on the dataset of alcohol consumption in Canada, 1979, which contains various aspects of factors. Based on the variable description, most demographic variables were selected and make a deep analysis to explore how does individual's context could influence their drinking habits. The data will be analyzed by R(R Core Team 2020), tidyverse(Wickham et al. 2019) and dplyr(Wickham et al. 2020) packages. All graphs will be created by ggplot(Wickham 2016) and the entire file will be knitted to pdf format by knitr(Franbois, Henry, and Miller 2021). Other packages, janitor (Firke 2021), pdftools(Ooms 2022), purrr(Henry and Wickham 2020) are also used.

Data Source

As a service provided by Ontario Council of University Libraries, ODESI provides data from many sectors, including agriculture, business, health, education, employment… The dataset we used in this paper was viewed and downloaded from ODESI, which was an internet tool that provides extraction, exploration and analysis for the dataset (COU 2022).

"Alcohol Consumption in Canada, a National Study, 1979" was established based on the national drinking survey, which was a dialogue campaign of drinking lasted from 1977 to 1979. This dataset collected data in the last year, aiming to concern alcohol use and associated problem. It provides as an exhaustive study on drinking habits of Canadians. ODESI provides us a complete dataset and the related codebook (Limited 1979). The codebook documents the overview of the national study and the specific meaning of each variable in the dataset, which was used as a reference in our paper.

Data Collection

Geographically, the alcohol consumption in Canada was collected by provinces. By utilizing the method of random sampling, a four-stage random sample selection was designed based on locality, enumeration area, block and household (Limited 1979). However, it excludes residents in Northwest Territories and Yukon because of sparsely distributed population and least accessibility. The dataset collects information of individuals between ages of 15-90 years, which was persuasive to represent the alcohol intake of residents in Canada.

As the primary investigator, Canadian Facts Limited was responsible for data collection process. The sampling frame was created by randomly selected qualified residents from various areas, if they at home at the time of interview, the face-to-face interview in English or French will hold. The interview question was recorded in a questionnaire, convenient for arranging a result dataset.

The population was residents in Canada, to sample the population, residents from 10 provinces was included in the sample. There are 2068 respondents were selected between age 15-90 across various areas. Each province was separated into separate strata in order to make the sample become representative. Due to the high efficiency of face-to-face interview, 2056 responds were recorded, the response rate over 98%.

The respondents will be asked questions from the questionnaire that made by Canadian Facts Limited, then the result will be summarized and transferred to a numerical dataset. Numbers directly represents categorical data, which makes the dataset was easier for further exploration. For instance, the 10 provinces in Canada were represents by numbers from 0 to 9. The corresponding code book was also posted ODESI as reference for scholars to interpret the data.

Data Overview and Cleaning

In the data cleaning process, we aim to extract important variables; furthermore, we also convert the form of some variables to visualize them. For instance, the sex of respondents was represented as number; to visualize them in a bar graph, we need to convert it back to categorical variable, clearly reflect the distribution of male and female.

We also made calculation between variables to get a new useful variable: in this dataset, it provided number of people in a household and number of people that aged over 18 years old in a household. We subtract those two variables to get number of people that was younger that 18 years old, which is called people_under_18. This was also an important variable that possibly affect the alcohol intake for people.

Exploratory Data Analysis

In this section, we will run diagrams for variables that will be used in the linear regression. It was a necessary step before constructing a model, we made an initial investigation on data, hence, to determine their distribution and the correlation with outcome variable y.

Figure 1: Days having one or more drinks

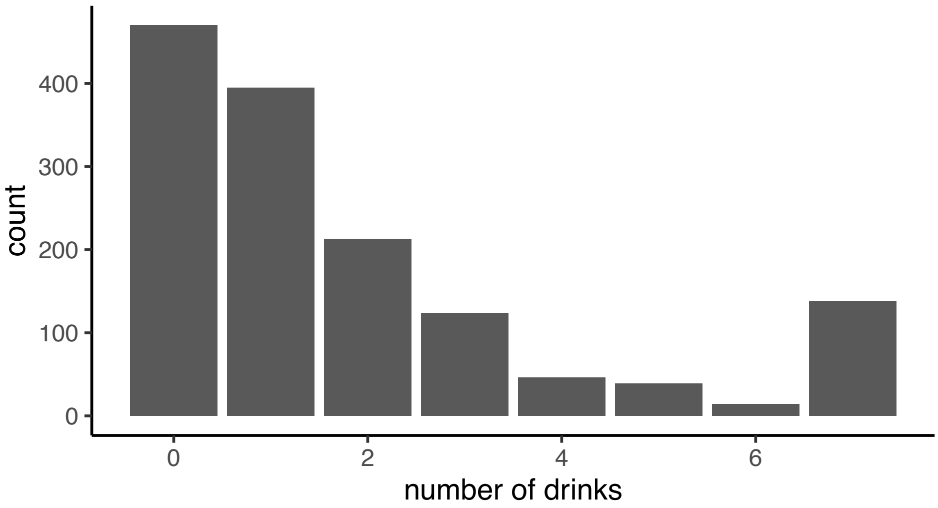
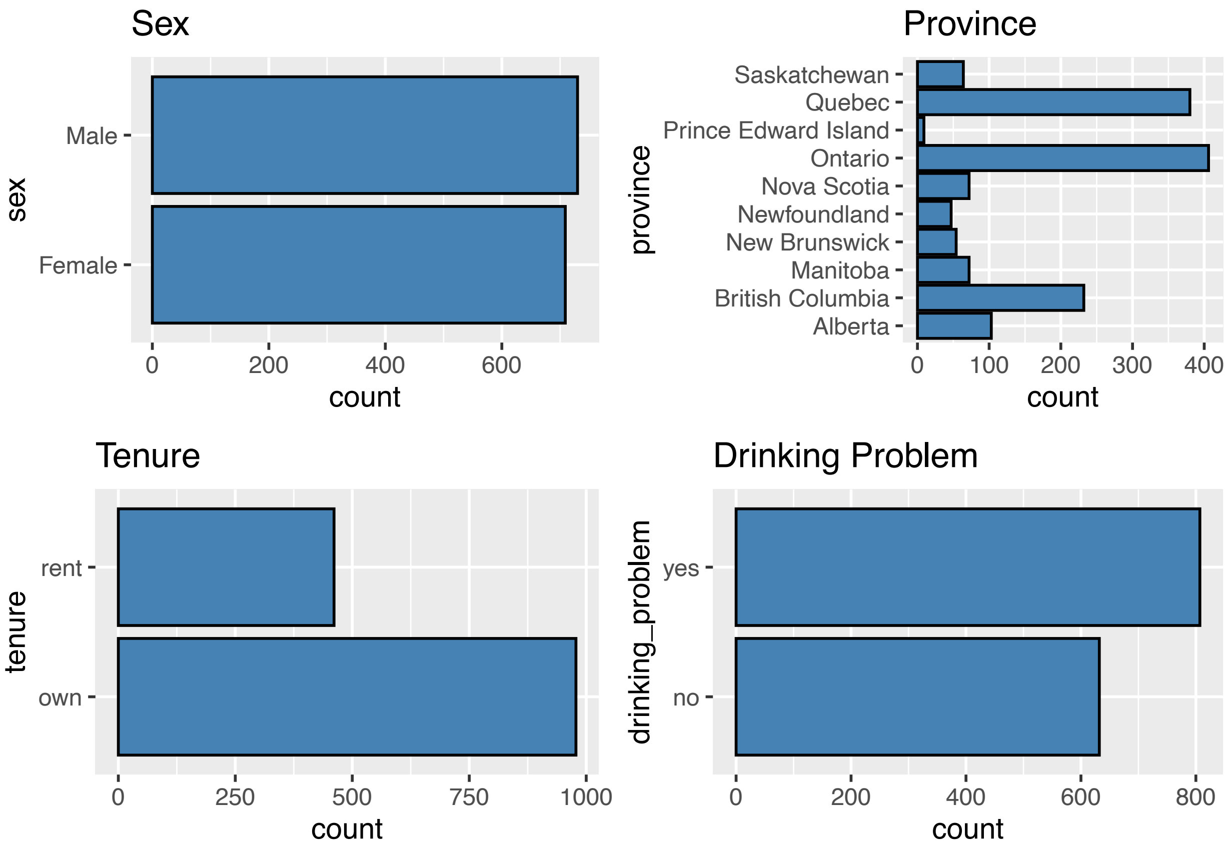
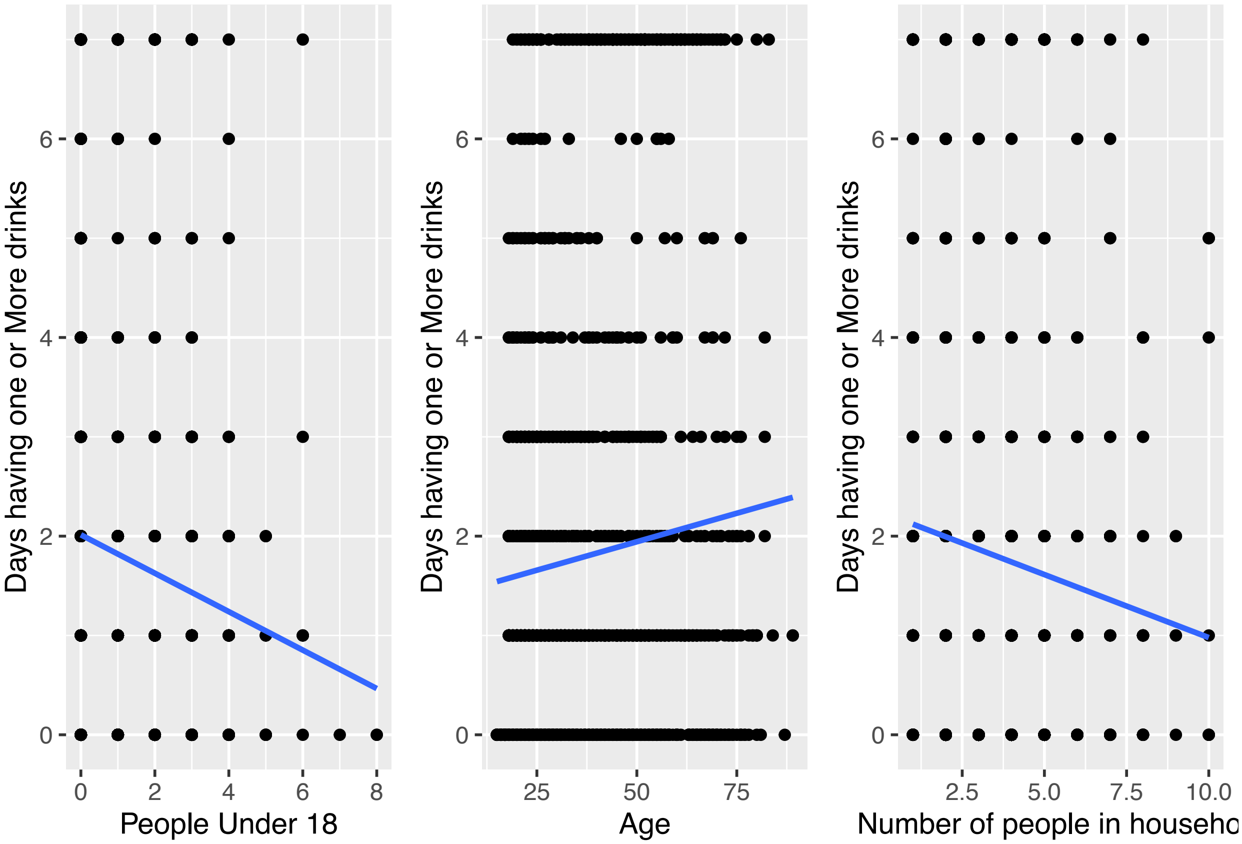


Figure 1 shows the distribution of y, which was the frequency of alcohol use every week. It shows a right skewed tendency, majority of the respondent using alcohol 0~2 times per week. There is also a small peak on the right end, which means some alcohol-addicted people use alcohol every day. This histogram reflects the drinking habit of Canada's residents.



The figure above shows the distribution of four important input variables by boxplot: sex, province, tenure and drinking problem. Sex represents the gender of respondence, and province indicates their place of residence. Tenure corresponds to whether their home was rented or owned, respondents will answer this question with rented or owned. Drinking problem means whether respondents know one more person with drinking problem, this was a yes or no question.

The y-axis of boxplot divided data within a variable into subgroups, and the x-axis was responsible for counting the amount of data in each subgroup, this reflects whether the data could represent the population. For sex, the number male respondent and female respondent does not have a big difference. Then the province could also represent the population distribution in Canada: Quebec, Ontario and British Columbia have most residences in the country. Tenure section also could reflect the real-life situation: majority of Canadians owned their home instead of renting. From the result of statistics Canada, approximately 19.1% of Canadians aged 12 and older are reported as heavy drinker (Canada 2022), so it was reasonable that more than half of the respondents know people with drinking problem.



The three scatterplots show the correlations between numerical variables and the frequency of drinking per week. The left side diagram shows the number of children (younger than 18 years old) and the drinking frequency; it shows that if a household has more than 4 kinds, there will be relatively lower frequency of drinking. The scatterplot located in the middle represent how age influences the frequency of drinking: it tends to be randomly distributed. The right scatterplot reflects total number of people in a household could affect the frequency of drinking, more people in a household lead to less intake of alcohol.

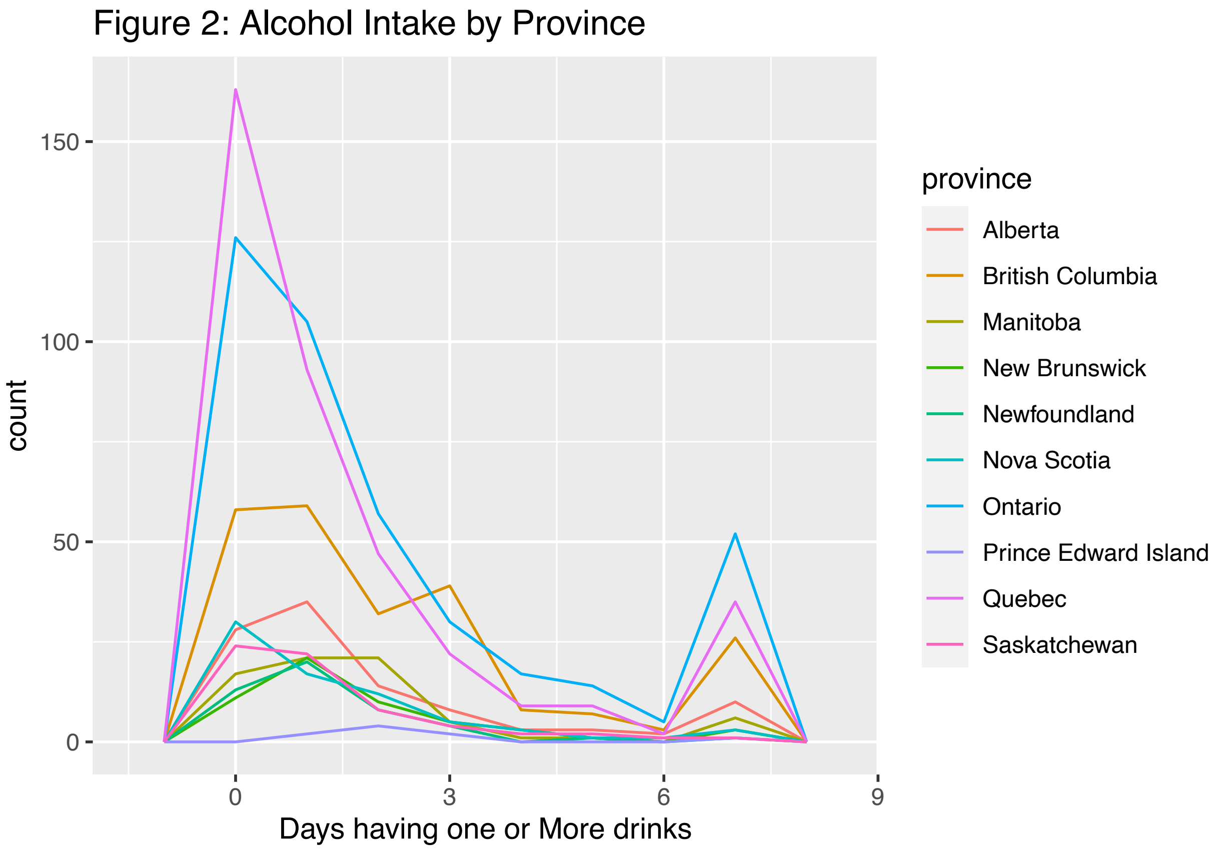


Figure 2 analyze the drinking frequency across province through a frequency polygon, each color represents a province. The lines were ranged one by one from top to the bottom, which represent different number of respondents for each frequency; they showed a similar patten of fluctuation on zero and seven.

Result

From the previous data section, we obtained a primary investigation on important variables that will be selected in the linear selection. In this section, we will be constructing a full linear model, then develop a best model by different methods. Both automated selection and manual selection of variables will be used. Hence, we can draw conclusion of what variables will influence the frequency of drinking and their correlations.

Linear Model: Candidate model 1 (Full Model)

1. Model Construction

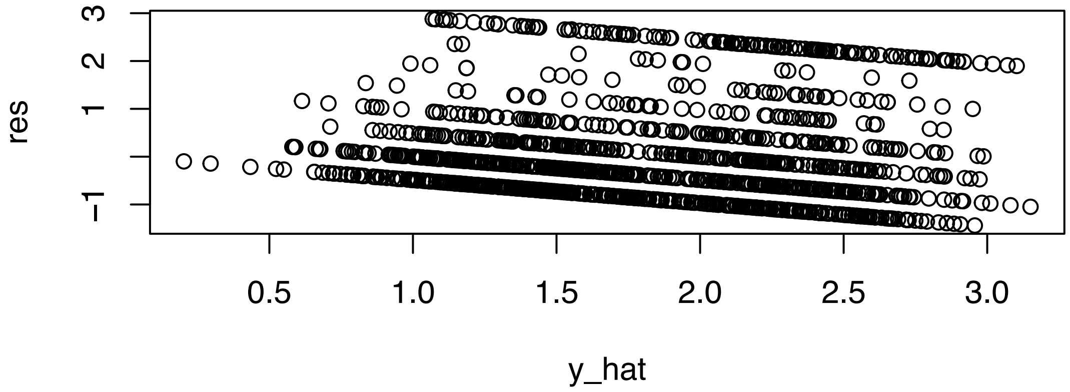
The first linear regression model we constructed was called full model. It contains all possible variables that could affect the frequency of drinking: province, sex, number of people younger than 18, age of respondent, income, tenure and drinking problem. Table 1 presents the importance of each variable by p-value. Then the correspondent residual plot and normal q-q plot was also generated below.

At this point, our linear regression expression could be written as: Y = β_0 + β_1(V07_PROVINCE) + β_2(V08_SEX) + β_3(people_under_18) + β_4(V30_RESPAGE) + β_5(V43_SOCCLASS) β_6(V36_TENURE) + β_7(V70_DRINKPROB)

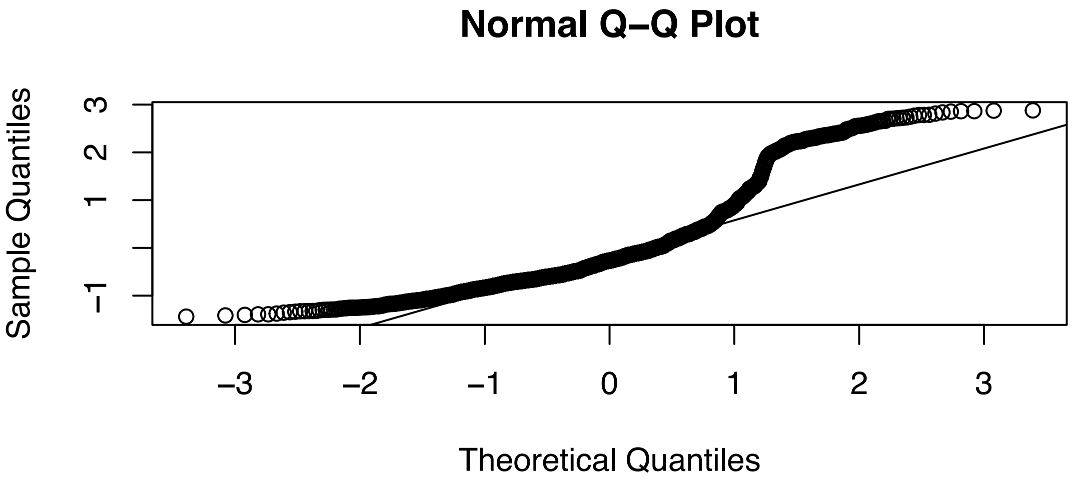
Table 1	Estimate	Std. Error	t value	Pr(>
(Intercept)	2.482505	0.411112	6.039	1.98e-09
V07_PROVINCE	-0.017817	0.020915	-0.852	0.394436

Table 1	Estimate	Std. Error	t value	Pr(>
V08_SEX	-0.926761	0.109301	-8.479	< 2e-16
people_under_18	-0.123151	0.046145	-2.669	0.007699
V30_RESPAGE	0.012373	0.003519	3.516	0.000452
V43_SOCCLASS	0.089418	0.033312	2.684	0.007353
V36_TENURE	0.230404	0.127662	1.805	0.071317
V70_DRINKPROB	-0.215464	0.111588	-1.931	0.053694

2. Residual plot



3. Normal Q-Q



Linear Model: Candidate model 2 (Manual Selection)

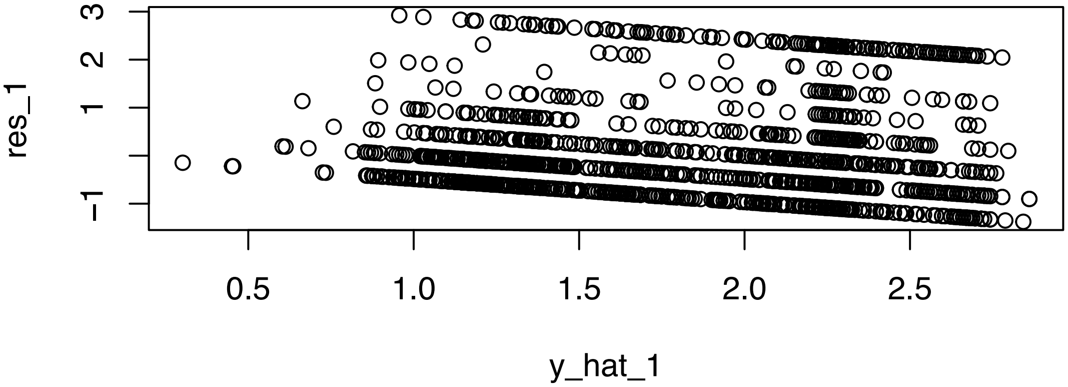
1. Model Construction From the result of last linear regression, we can use p-value to determine the significance of each variable. Then we manually select significant variables in order to improve the accuracy of the model and checked by residual plot and normal q-q plot. The selected variables were presented in Table 2.

The expression of manual selected linear regression was written as:
Y = β_0 + β_1(V08_SEX) + β_2(people_under_18) + β_3(V30_RESPAGE)

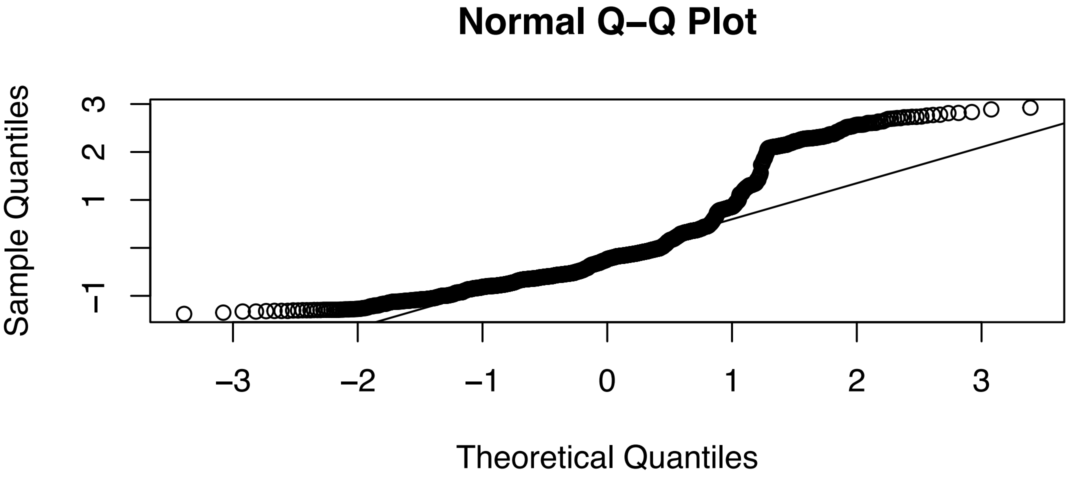
Table 2	Estimate	Std. Error	t value	Pr(>
(Intercept)	2.962966	0.222837	13.297	< 2e-16
V08_SEX	-0.911434	0.109443	-8.328	< 2e-16

Table 2	Estimate	Std. Error	t value	Pr(>
people_under_18	-0.140011	0.044887	-3.119	< 0.00185
V30_RESPAGE	0.009092	0.003321	2.738	0.00626

2. Residual plot



3. Normal Q-Q



As we could see, the residual plot became more randomly distributed, it proves that our manual selection makes the model become more accurate. However, the plot of normal q-q plot still obtains deviations. We could not determine this manual selected model as the final model.

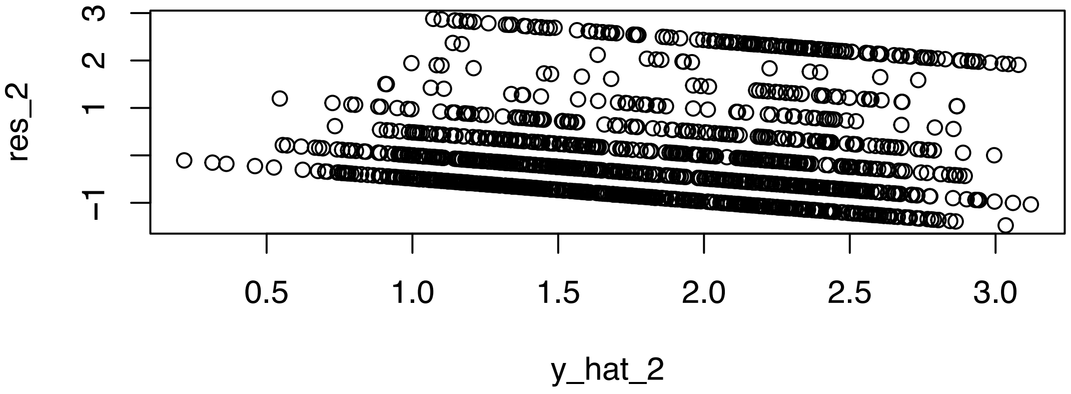
Linear Model: Candidate model 3 (Automated selection)

1. Model Construction The auto selection process was using R to select most related variables and construct corresponding linear regression model. The auto selection was based on the output of full model, which contains all related variables from the dataset, then it automatically generates a new model . After comparison, we found that the auto-selected model deleted the variable province, which represents the respondent' place of residence. Table 3 shows the variables in the auto-selected model and their significance. The p-value for province was 0.394436, it was larger than $\alpha = 0.05$, shows insignificance of the variable.

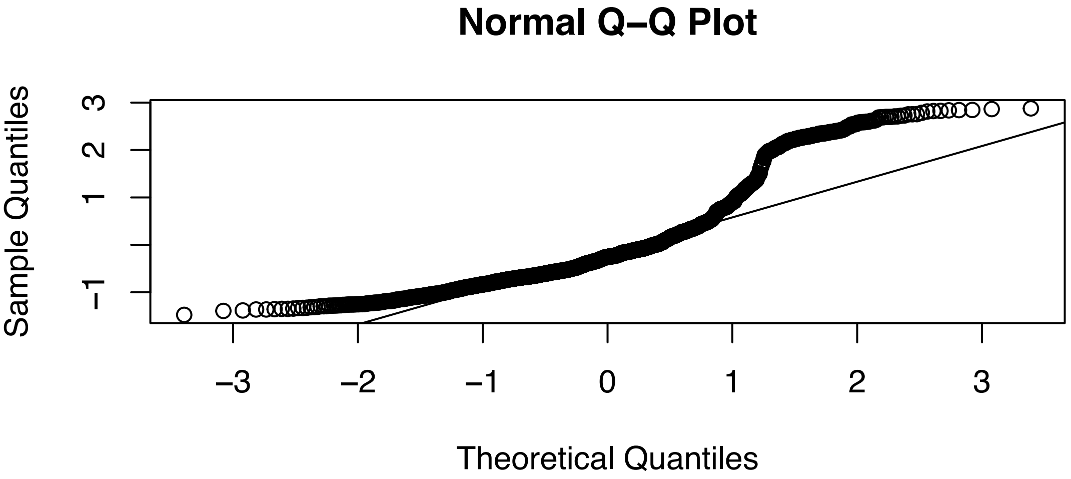
Hence, the expression for the auto-selected model was: Y = β_0 + β_1(V08_SEX) + β_2(people_under_18) + β_3(V30_RESPAGE) + β_4(V43_SOCCLASS) β_5(V36_TENURE) + β_6(V70_DRINKPROB)

Table 3	Estimate	Std. Error	t value	Pr(>
(Intercept)	2.407024	0.401411	5.996	2.55e-09
V08_SEX	-0.929928	0.109227	-8.514	< 2e-16
people_under_18	-0.123263	0.046140	-2.671	0.007638
V30_RESPAGE	0.012389	0.003519	3.521	0.000444
V43_SOCCLASS	0.090072	0.033300	2.705	0.006914
V36_TENURE	0.227247	0.127596	1.781	0.075127
V70_DRINKPROB	-0.217755	0.111545	-1.952	0.051112

2. Residual plot



3. Normal Q-Q



The distribution of the residual plot tends to be more separated and without specific patterns, which was better than the full model and the manual-selected model. For the normal q-q plot, the auto-selected model was more consistent and closer to the normality line. Hence, it was chosen as the final model.

Model Interpretation

After a careful comparison between full model, manual selected model and auto-selected model, we determined the auto-selected model as our final model because it satisfied linear regression assumptions. The model does not experience any mathematical complex transformation; hence it clearly reflects the correlation between each variable and the frequency of drinking.

Table 3 has listed six variables that selected in the final model, presenting their correlation with y-value numerically. In this final model, sex, number of people under 18 years old and drinking problem has negative correlation with the frequency of drinking. It reflects that female, less number of people under 18 years old, known people around had drinking problem will lead to less heavy drinkers.

On the other side, age, income and tenure had positive impact on the frequency of drinking; higher age will lead to more frequent drinking but not correlation was not very strong due to the small value of coefficient. Then, higher income causes more intake of alcohol, they are relatively more affordable and had more choices of alcohol. At the same time, people who renting their home also shows the tendency of more alcohol intake.

Discussion

Conclusion

The investigation on alcohol consumption in Canada utilizes statistics model to explore the significance of various variables. As we downloaded data from ODESI (Archive 2022), in fact, we already complete the preliminary variable selection. The focus in this paper was demographic variables, which was an analysis of respondent' background. From this result, we can analyze the reasons behind high alcohol assumption and adjust advertisement to make it more efficient, this study contributes to the health promotion activities.

Through the usage of linear regression, it mathematically presents how different demographic variables might contribute to alcohol consumption. Linear regression model was easy and efficient to interpret the correlation between each variable and y-value. The result of the linear regression was also could be used as extrapolation beyond the specific dataset, which also correspond to the main idea of this paper: exploring alcohol consumption for population in Canada.

Three different linear regression models were constructed and compared, contains full model, manual selected model and auto-selected model. The full model includes variables in our dataset, both categorical and numerical. The last two models did selection based on the result from full model, which contains all variables in the dataset. The manual selection was based on the output of full model, only selected model with high significance; and the auto-selection was using R system, selecting important variable. By the analysis through residual plot and q-q plot, we check which model could satisfied the linear assumption mostly.

The coefficient of linear regression model for each variable indicates their relation; whether negatively impacted or positively impacted, hence, to draw a conclusion. Based on the result of final model, we targeted groups of people that specifically needed health promotion of alcohol: male, families that have many children, people who do not know anyone with drinking problem, elder people, group of high-income and people who rent houses. In the entire population of Canada, we should more emphasis on those groups of people. The advertisement would become more effective when we determine the target groups of people.

Improvement

The original dataset on alcohol assumption of Canada was conducted in 1979, at that time, the technology of collecting data was not advanced. All questions were recorded as a questionnaire, then Canadian Facts Limited used face-to-face interview to collect data, which ensures the respondent rate and accuracy of the study, however, the number of respondents was not enough. To make our study more effective, we can use other methods at the meantime, send the questionnaire directly to more respondents. Hence, we can have a larger sample size for future analysis.

The sample was used to present the entire Canadian population. The study covered ten provinces in Canada, British Columbia, Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan and Alberta. Those are regions that had majority of the population. However, Northwest Territories and Yukon, places that are sparsely populated and least accessible are not included. They constitute 7% population in Canada, according to consumption statistics and other known correlates of alcohol abuse, they represent the most likely sources of heavy drinkers (Limited 1979). The data would be more representable if we included those regions next time. Due to least accessible, we can send the questionnaire online or mail to their home directly.

We utilized linear regression model to analysis the impact of each variable in this paper, however, it obtains some drawbacks that cannot be neglected. The linear regression was very sensitive to outliers, it requires a strict linear relation between variables, any outliers could largely affect the value of coefficients. This model only nicely fits dataset that could fit into a single linear correlation, it is prone to multicollinearity. Linear regression only presents linear relation between variables, which is not exhaustive enough. Many numerical values do not have linear regression, they distributed in other patterns. The final model also shows that some variables do not have enough significance, they slightly influence the y-value, which is alcohol consumption. Next time we can also try other models out to test its validity, not only linear model.

Appendix

Datasheet

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The national study contains a code book, includes interpretation for each variable. Tables, histograms are also included, showing the distribution of data for each variable separately.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 9 tables, 70 histograms.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The paper contains all tables and histograms. The questionnaire was conducted to the represent general population of Canada.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - All data from the questionnaires are recorded in the dataset and converted into numerical value. The tables and histogram we see in the paper are all processed and properly laid out by sections.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Yes. The paper was divided based on general topics and sorted under different topics. Including administration, demographic, household and alcohol use. Each instance is displayed under the corresponding sections with titles.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - No
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Yes, the organization of the paper made the connection explicit.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - No
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - The results are affected by non-sampling error and sampling error. The sampling method used was not simple random. It was due to the data collection method, the investigators use face-to-face method to collect data, which means only people who willing to take interview and they stay at home while data collection will respond to the questionnaire. It does not randomly select people across Canada. On the other hand, some least accessible and sparsely populated area are not included in this dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - It is self-contained. The paper was based on the dataset they collected.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected*

by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

- No

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The data was collected randomly from Canadians between age 15-‘90. The study was focused on alcohol consumption, so younger kids are not included in the study.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - No
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data was collected based on face-to-face interview, the raw data consisted of 2056 observations. The results were organized into the paper. The data was not inferred or derived from other data.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - The staff from Canadian Fact Limited conducted the data collection procedure. A four-stage random sample selection was employed according to locality, enumeration area, block and household. A sample of 2068 respondents between the ages of 15 and 90 years was randomly selected from the qualifying household members at home at the time of interview.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - The survey was a national study, target people who aged between 15-90. By the face-to-face interview, the response rate was over 98%.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - The staff from Canadian Fact Limited was involved, includes investigators that could speak English and French. The paper did not mention the compensation.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data collection conducted between November 1978 to February 1979.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- The data was obtained directly from the respondents, then converted to a numerical dataset.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - Yes, they are aware of the collection. Before the interview, the investigator would make acknowledgement and respondents were allowed to not answer questions that they considered as sensitive.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - By responding, they consent to give their information.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - No
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - The data processing process was not mentioned in the paper.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - N/A
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - N/A
4. *Any other comments?*
    - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No
3. *What (other) tasks could the dataset be used for?*
    - The result of this will be share with Health Promotion and Health and Welfare Canada, it helps to improve the efficiency of health promotion on alcohol consumption.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - No
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The paper did not mention how the dataset will be distributed. The paper is accessible to the general public in many databases.
3. *When will the dataset be distributed?*
   - The paper did not mention when the dataset will be distributed.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - N/A
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The survey was conducted by Canadian Facts Limited. The paper did not mention if they are going to conduct future surveys.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Alcohol Consumption in Canada: A National Study, 1979 could be found on ODESI, the detailed code book http://odesi1.scholarsportal.info/documentation/alcohol-consumption/1979-02/1979-alcohol-consumption-cdbk-DRG.pdf
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - N/A
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - No
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - No
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- They would contact the Canadian Facts Limited, the primary investigator of this study.

References

Archive, Canadian Opinion Research. 2022. *Ontario Data Documentation, Extraction Service and Infrastructure Initiative*. http://odesi2.scholarsportal.info/webview/.

Canada, Statistics. 2022. *Heavy Drinking, 2018*. https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00007-eng.htm.

COU. 2022. *Ontario Council of Universities*. https://ocul.on.ca/node/2116.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Franbois, Yihui Xiein, Lionel Henry, and Kirill Miller. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*.

Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

Limited, Canadian Facts. 1979. *Alcohol Consumption in Canada: A National Study, 1979*. http://odesi1.scholarsportal.info/documentation/alcohol-consumption/1979-02/1979-alcohol-consumption-cdbk-DRG.pdf.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

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

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 Franbois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.%0A%20%20%20%2021105/joss.01686.

Wickham, Hadley, Romain Franbois, Lionel Henry, and Kirill Miller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.