Smoking and Alcohol Drinking Will be Effected by Some Factors*

Education, Gender and Living area are the main factors which will affect smoking and alcohol Drinking

Yijun Shen

13/04/2022

Abstract

This report conducts a secondary analysis of the Canadian Tobacco, Alcohol and Drugs Survey 2017, and selectively analyzes data on Tobacco and Alcohol. The results showed that area of life determined the frequency of smoking and drinking, the timing of first drinking or smoking was determined by education level, and household composition was a key determinant of secondhand smoke exposure. And the statistical programming language R was used for logistic regression analysis. The results can help us understand the causes of smoking and drinking, and reduce our exposure to tobacco and alcohol so that we can lead healthier and more active lives.

Keywords: Cigarettes, Health, Nicotine, Nonsmokers, Smoke, Teenage smoking, Youth, Alcohol, Drinking, Mentalhealth, Family, Education

Contents

*Code and data are available at: https://github.com/Nckshen/final-paper.

1 Introduction

In today's society, tobacco and alcohol have been fully integrated into people's lives. Smoking and drinking have become the current status of some people. One of the reasons why people are so vulnerable to cigarettes and alcohol is because of excessive stress in their lives.(Azagba 2011)(Azagba and Sharaf 2011) In recent years, with the impact of the global epidemic, the world economy has suffered unprecedented damage. The characteristics of migrant workers are heavy workload and low salary, which leads to the imbalance between labor output and income, which makes people's life pressure more serious. When people's stress builds up to a certain level, they will erupt, and the results may be uncontrollable. One way people avoid stress outbreaks is by overeating and using more alcohol, tobacco and drugs.(Marks 2021)(Marks 2021) Excessive use of alcohol, tobacco and drugs can lead to addiction. As described by the National Cancer Institute, cigarettes are tubular tobacco products made from finely cut cured tobacco leaves wrapped in thin paper. Nicotine is the most addictive in tobacco leaves, and tar is a major cause of cancer.(NCI 2021)(Institute 2022) Alcohol, according to CAMH, is a sedative and reduces stress by slowing down the parts of the brain that control thinking, behavior, breathing and heart rate.(CAMH 2021)(CAMH 2022) Moderate drinking can relieve stress, but excessive drinking can lead to dependence on alcohol and even death after excessive drinking.(Statistics Canada 2021)(Canada 2021) According to Statistics Canada, 21 million people drank alcohol in 2022, while about 4.7 million smoked.(Statistics Canda 2019)(Canada 2020)

In this article, we will first clean up the raw data so that all relevant data can be used directly for drawing and modeling. Secondly, we introduced the detailed information of each variable of the data used, and listed all the problems existing in the data. Then we made several pictures based on the data in order to better explain the advantages and disadvantages of the data. Next, we matched a model according to the characteristics of the data, and explained the role and use of the model and some drawbacks in detail. Then the data are analyzed in detail and the corresponding conclusions are drawn. At the end of this paper, a summary of this paper, Ethics and Limitation and Future work are also included. The Appendix section includes a datasheet and a short survey. The purpose of this article is to find out the causes that influence people's use of tobacco and alcohol, in terms of education, region, and gender. And it turns out that all of these factors have an impact on tobacco and alcohol use. From the results obtained, we can effectively reduce the harm caused by tobacco and alcohol to people, thus making the society more harmonious.

2 Datas

2.1 Dataset Infromation

To conduct a correlation analysis on Alcohol and Tobacco use, I used Odesi to find a dataset called Canadian Tobacco, Alcohol and Drugs Survey 2017 from Statistics Canada, It's a livelihood survey done by the Canadian government. The data recorded the results of a 2017 survey on alcohol, tobacco and drug use. The initial data included the results of a survey of 16,349 people and 407 related questions, including some on the frequency of drinking and smoking, family background, education and health. Questions about whether people use drugs are also included. The data set also record the specific time of the survey. The data will be cleaned and retained some relative variables for the further analysis. The data will also be analysed by using R(R Core Team 2021), tidyverse(Wickham et al. 2019), ggplot(Wickham 2016), knitr(**knitr?**),gridExtra(Auguie 2017) and tibble(Müller and Wickham 2021).

2.2 Variables Information

In order to get the data most relevant to the research topic, I first deleted all variables unrelated to drinking and smoking in the initial data. In the initial data, all variables were named by combining letters and numbers, and some variables could not be understood without reading the Code book, so I used rename to rename all variables. So that the data can be better understood. In the remaining data, all variable results are displayed by numeric codes. For example, in the case of gender, the number 1 represents male and the number 2 represents female. Therefore, I will combine mutate and case when to change all the numbers in variable into specific answers.After modifying variables, I will use select and -c to delete all unmodified variables to make the data set look clearer. In some numerical variables, there are some numbers such as 97,98,99, which mean that there are no specific answers in the survey. Therefore, I use filter to delete all the results containing these answers, so as to reduce the influence of these answers on the overall analysis.

After I cleaned the data, there are only 42 variables and 15933 obs left, the variables left are mainly about household information, personal information, alcohol frequency, smoke frequency. Among these variables, "-14," "15-24," "25-44," "45-,""household_type," "Household_size," are the variables about the ages of all household members; "num_peo_smok," "Day_peo_esmok," "Year_smoke," "first_smoke," "first_smoke_from," "num_cigarette_daily," "frequency_ofsmoke," "monthly_sec_smok," these variables are all relate to smoke, including smoke frequency and smoke number; Variables include "age_of_drink," "drink_SUN," "drink_MON," "drink_TUES," "drink_WED," "drink_THUR," "drink_FRI," "drink_SAT," "drink_underage," "drinked_before," "Yearly_drinking_beverage," "Drink_status" are about alcohol drinking, which include drinking history, drinking beverage and status.

2.3 Gaps of Datas

There are incomplete data in government surveys on alcohol and tobacco. A large proportion of respondents chose not to answer questions about the frequency of alcohol and tobacco use, reducing the availability of data on the frequency of alcohol and tobacco use. As a result, the accuracy of subsequent analysis is greatly reduced, and there may be some bias. For example, if all first-time smokers who were younger than the legal age to smoke chose not to answer the question, there would be no early smokers in the survey, which would skewing the analysis of the data to a wrong result. Therefore, some important questions can be marked in the survey to guide participants to fill in these important questions completely for subsequent data analysis.

Secondly, the data did not explain the job and income of the participants. The overall survey was basically conducted around family and personal history, such as how many people there were in the family, the ages of family members, and the age when they smoked and drank for the first time. As a result, the analysis could not be based on participants' jobs and income. The resulting results may be incomplete because of the absence of data on participants' jobs and incomes.

Table 1: Table of Respndents' Ages

Min	Q1	Median	Q3	Max	IQR	Mean	SD	Mode
15	18	22	45	85	27	31.2	17.71	numeric

2.4 Tables and Plots

The table above shows the information of the age of respondents, which can seen that the youngest respondents are only 15 years old, and the oldest respondents are 85 years old, a median age of 22 and with a mean of 31. The standard deviation of the age are about 17.7, which means that the average age will add or subtract up to 17 years from the average age of 31. Through the median age of the respondents, the median age of 22 means that most of the respondents are around 20 years old. A large deviations in data range may lead to unrepresentative final analysis results, since the most of the respondents are under 30 years old. But due to the lack of internet using skills, those people who over 50 years old may have problem to do the survey online, which means the unbalance of the age of respondents has a proper reason, because of the social trends favor the young people. Somehow, this survey may be created for young people, possibly to make them more vigilant about alcohol and tobacco use.

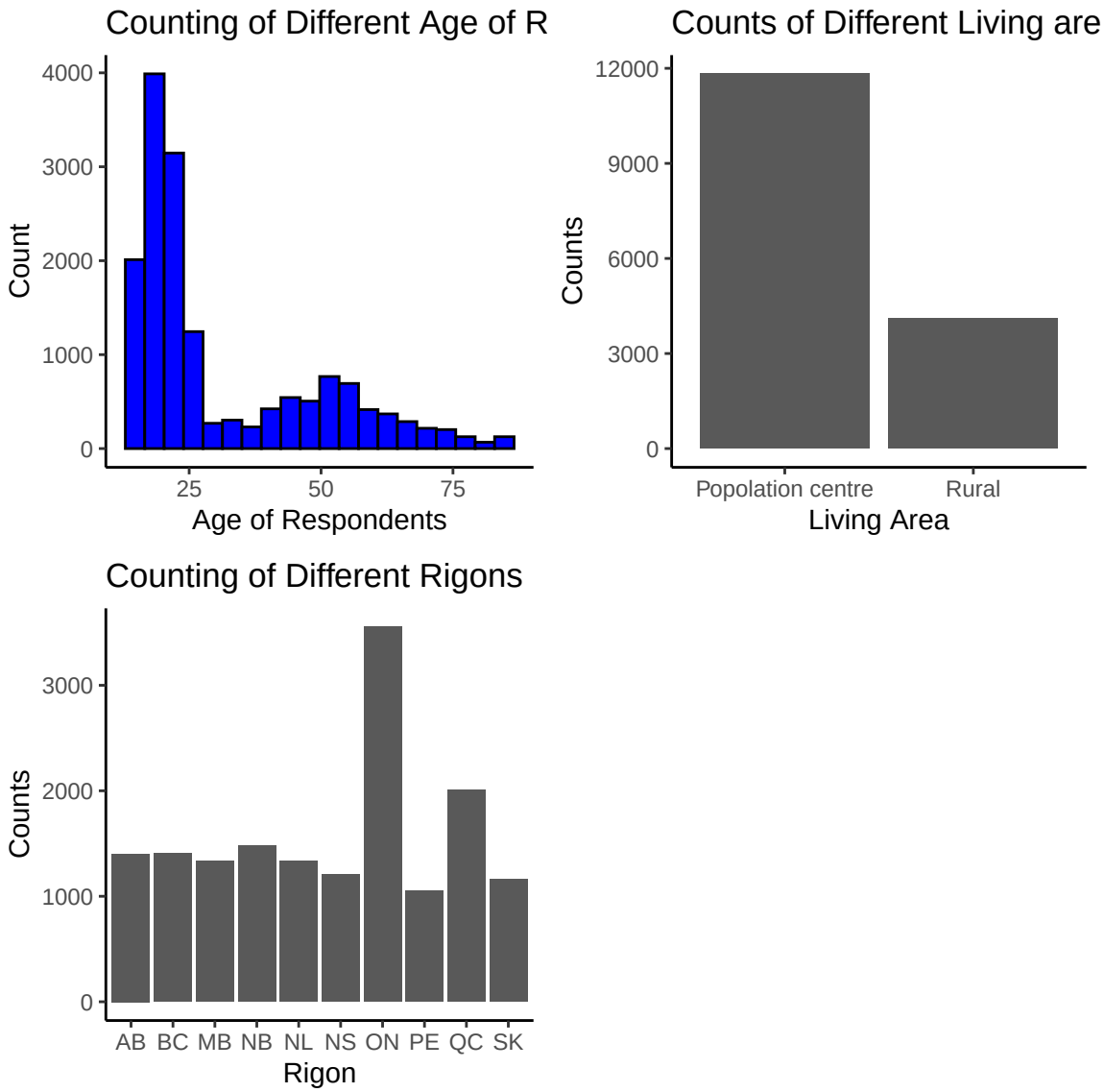


Figure 1: Figures include age, gender and rigons

In this histogram of age of respondents, the number of respondents with different ages drops off in a cliff from young to old, it resulting the plot become not normal. Although this positive skewness will have an unsteady average (mean), but the median and the mass of the respondents' ages could not be replaced, which is steady.

According to the bar chart of regions in the respondents' survey, Ontario and Quebec had the highest number of respondents by province, while the other provinces had a similar number of respondents, each about one third of the number of respondents in Ontario. This may be due to the fact that Ontario and Quebec, as Canada's two most economically advanced provinces, also account for a large part of the population of Canada. Thus, Ontario and Quebec had the most respondents on this survey.

Based on the observation of the bar chart showing the respondent's living environment, most of them live in a densely populated area, while only a small number live in the countryside. This means that the results of this analysis may best describe densely populated areas, not all areas. Therefore, this is also the shortcoming of the data, and will affect the integrity of the data analysis, thus leading to the deviation of the final results.

3 Methodologys

The aim of the study was to find out what might influence alcohol and tobacco use, not only in families but also in individuals. A generalized additive model can be used to find out the relationship between different variables and alcohol and tobacco use. The model will be shown that:

$$Y_i \sim G(\mu_i, \theta)$$
$$g(\mu_i) = X_i\beta + Z_iU + f(W_i)$$

The Generalized Additive Model(GAM) is a fancy Generalized Linear Model(GLM), which GAM can fit both linear and non-linear model. Although the linear model is much more intuitive and easy to understand. However, in real life, the function of variables is usually not linear, and the linear assumption of variables often can not meet the actual demand, and may lead to the reduction of the accuracy of analysis. Thus, the Generalized Additive Model(GAM) is the best model to use to find out the relationship between different variables and the alcohol and tobacco use. Since the GAM is based on GLM, so, $Y_i \sim G(\mu_i, \theta)$ is the initial model of GLM, inside of this model, Y_i are the responses, and G is the responses distribution, μ_i and θ are the parameters of the responses. In the equation $g(\mu_i) = X_i\beta + Z_iU + f(W_i)$, $g(\mu_i)$ can be any form, which both linear and non-linear variables are available here, so, in this study, the frequency of alcohol and tobacco use; drinking under age; health status and mental health status can be represented as $g(\mu_i)$, which is the y of the equation. X_i, Z_i and W_i are covariates, which are the variables might have effects on y. β is the fixed effects, and U is the random effects, β will stay with the variable X_i and the random effect U will stay with the variable Z_i, and that means the variable X_i and Z_i are having fixed and random effects. In this study, the fixed effect can be age, gender, education and family, and the random effect can be drinking number in a week, yearly drinking beverage and current drinking status.

There are defects in many nonlinear data in the data set. For example, some people answer "Don't know," so in order to avoid large deviations in the established model, it is necessary to remove all answers that are "don't know." Therefore, taking the variable drink_SUN as an example, there are 16,349 responses in the original data, 416 were removed after screening, leaving 15,933 valid responses. By taking weekly drinking frequency as a random effect and education level as a fixed effect, we can judge whether the two will affect health through GAM model. The variable health as the $g(\mu_i)$ in the formula, it is non-linearity, this means that each answer in variable health may dependent on the frequency of smoke and the education level.

In addition, multiple fixed and random effects can be added to this model as predictors of health or mental health. Region and gender can also be plugged into the model as fixed effects. However, depending on the focus of this study, region and gender may not have an impact on health. Place of residence, a seemingly unrelated statistic for drinking and smoking, plays an important role in this model. The aim of the model is to find out why smoking and drinking affect health, and where people live can affect the frequency of smoking and drinking. In the data, there are two options for places of residence: rural and population centre. Representing countries and cities respectively, an article in the National Library of Medicine states that there are more people in population centre than in rural areas when it comes to smoking due to economic pressures. And because people in population centers have less demand for alcohol than people in rural areas, drinking is more frequent in rural areas than in population centers. (Dixon 2016)(Mark A. Dixon and Karen G. Chartier 2016) (Doogan 2017)(N. J.Doogan and T.Higginse 2017)

All factors affecting health through drinking and smoking should be verified in combination with the actual situation, so that all variables in the model will not affect the accuracy and representativeness of the final results. In addition, the data used in the model should be transparent and open to the public, so as to obtain different opinions and judgments from the public on the research.

4 Results

In this paper, all the reference group are Canadians all around Canada, between age 15-85, they have different education level, and live in population center and rural area. Since the model above can include both numerical and dummy variables, and all the variables inside of the model will show whether the variable has the effects on the yearly drinking or smoking beverage. For the family situation, education level, gender and living area, all of them have the effects, which the beverage may get changed based on the level of education and the area they live.

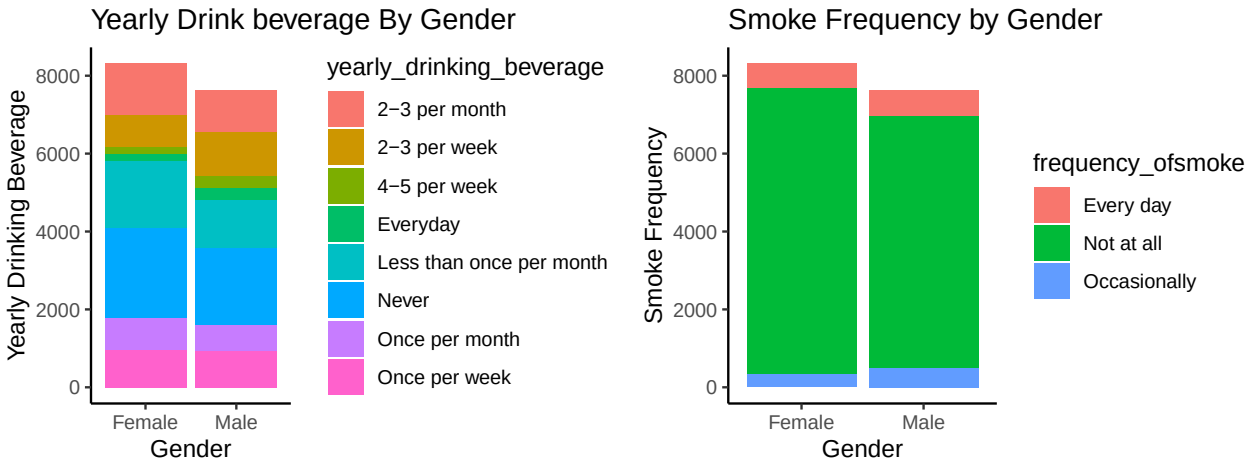


Figure 2: Bar plots of Gender and alcohol & smoke

Figure 2 shows two bar plots related to gender and frequency of alcohol and tobacco use. As shown in the first plot, it can be seen that there are more women than men participating in the survey, however this does not affect the final result. According to the proportions of the different areas in the first plot, it can be seen that females drink more frequently than males, while males drink more frequently and even daily. According to this, it can be concluded that the frequency of drinking is influenced by gender, and to some extent it indicates that the frequency of drinking is much higher in men than in women. The second plot includes the frequency of smoking among men and women, with those who have never smoked accounting for nearly 90%. However, according to the data provided by the smoking participants, the proportion of women who smoked every day was similar to that of men who smoked every day, but the proportion of men who did not smoke regularly was higher than the proportion of women. So women smoke more often than men. Through the observation and analysis of the two plots, it is concluded that different genders will affect the frequency of smoking and drinking, and will increase or decrease the frequency in actual situations

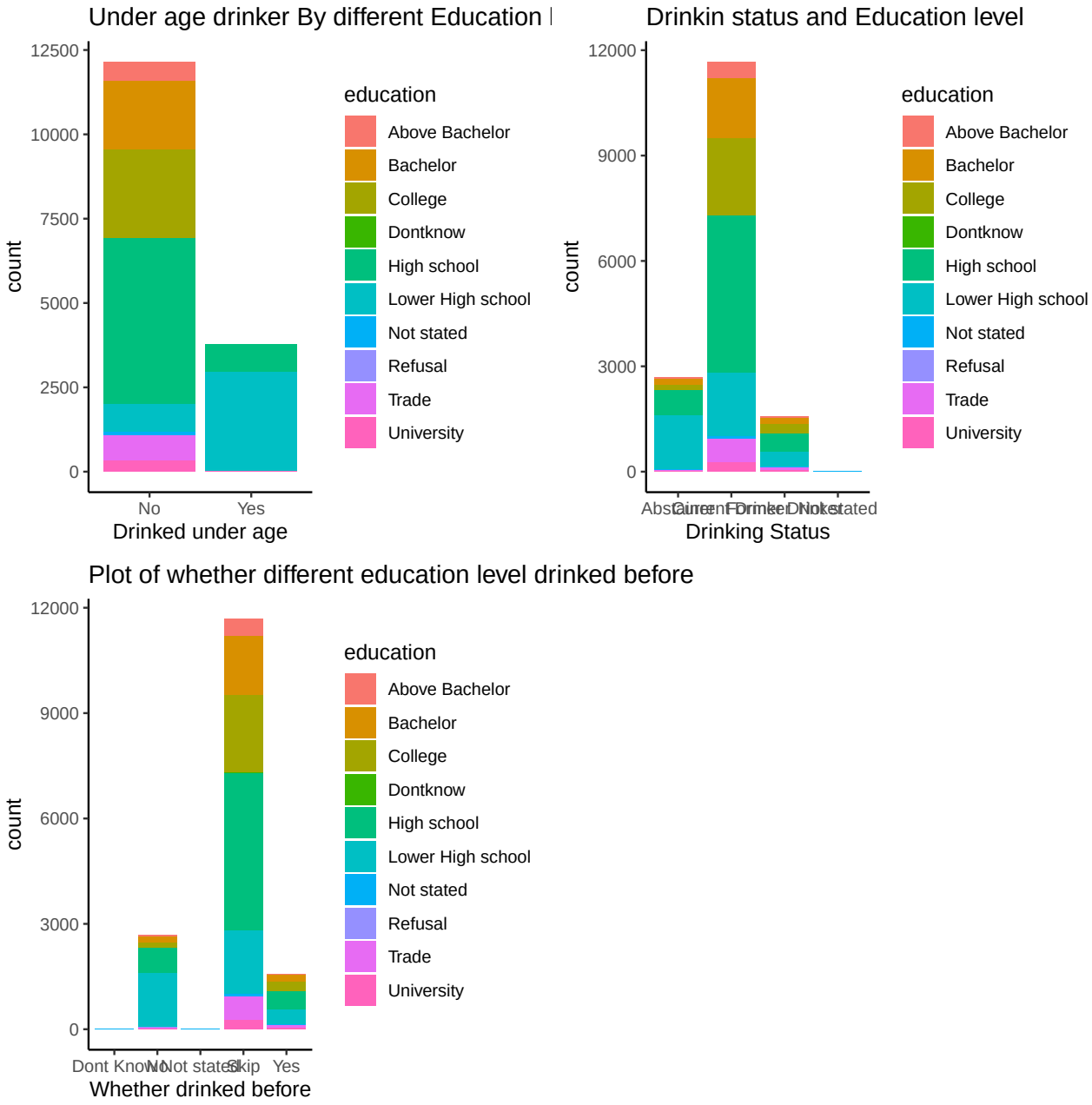
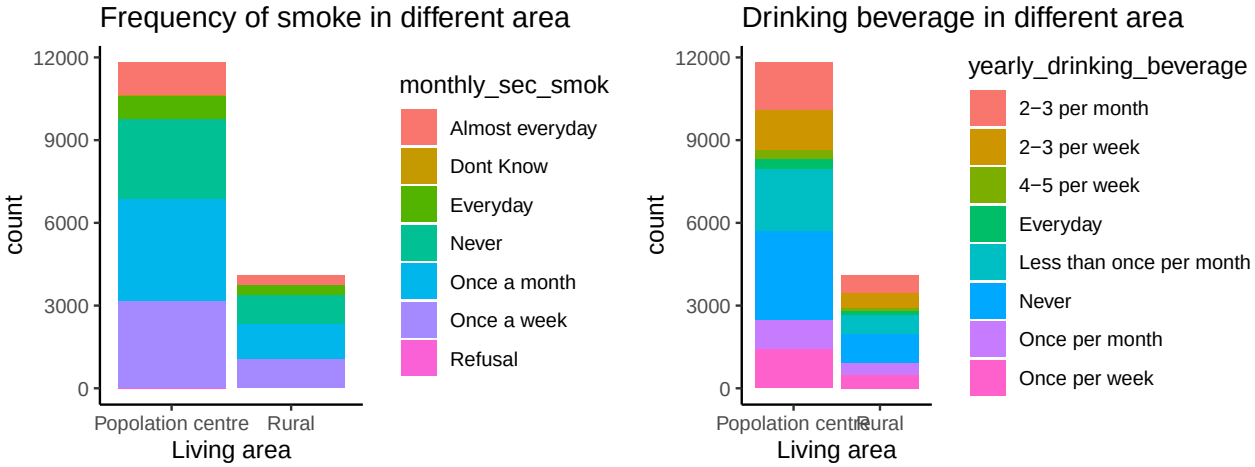


Figure 3: Bar plots about education level

Figure 3 shows three bar plots related to education and drinking. The first plot contained answers for different levels of education and whether the participants had drunk alcohol before the age of 19. As can be seen from the graph, those who drank alcohol before the age of 19 were basically those with a high school education or below, while those with a high school education or above did not drink alcohol before the age of 19, which also confirms that illegal drinking behavior will decrease with the increase of participants' education. In the second plot, according to each participant's academic background, their current drinking states were recorded, namely, abstinence, current drinker, former drinker and no stated. Of those states, current Drinker has the largest number of people with high school degrees, followed by College and bachelor degrees. Those with less than a high school education do not make up the largest proportion of current Drinker, while those with less than a high school education make up the largest proportion of current drinker. Combined with the information in Figure 1, it can be concluded that people with less education start drinking early and

may have some physical reasons for abstinence. Those with a high school education or more may be the biggest current Drinker because of life stress. It can be concluded that education determines the current state of people's drinking. The third plot shows whether participants of different educational levels have ever consumed alcohol, which does not include foods containing alcohol. Most of them skipped, and only a few answered. From these minority answers, it can be seen that the majority of those who have never drunk alcohol are those with less than a high school education. And those with a high school diploma or above are almost all alcoholics. Thus, to verify the conclusion in plot 2, education determines whether people have drunk alcohol or not, i.e. the current state of drinking.



In Figure 4, there are two bar plots related to residence and frequency of smoking and drinking. On the X-axis of the two graphs are the areas where the respondents live, and the frequency of smoking and drinking is segmented according to different areas. As can be seen from the graph, there are far more people living in the population center than in the rural area. By comparing the differences of different frequencies in different regions according to the proportion of people in different regions, we find that the proportion of people who smoke every day in the population center is much higher than that of people who smoke every day in the rural area. From the finding, it can be concluded that people living in population centers are more dependent on tobacco than those in rural areas. In terms of alcohol, we used the same method to compare the ratio of different frequency between the two regions, and found that the proportion of people who drink alcohol every day in the rural area is much higher than that in the population center. Based on these two findings, we can conclude that the demand of people living in different areas for tobacco will decrease with the decrease of population density, while the demand for alcohol will increase with the decrease of population density.

5 Discussion

5.1 What is done?

This paper analyzed the data by selecting alcohol and tobacco from the statistics Canada 2017 data set on drug, drug, alcohol and tobacco use. To understand how different factors affect people's use of alcohol and tobacco. I first cleaned up the data selected from the original data set to make it easy to analyze. Secondly, I made an introduction to the cleaned data, and explained the meaning and connection of the data used in detail. The data were then analyzed by comparing and analyzing the different bar plots that were created, and it was concluded that gender, education background and place of residence all affected people's alcohol and tobacco use. During the analysis, I also presented a statistical model called the generalized Additive Model (GAM) that might be suitable for this analysis. And explain the model and match the data in the data set, so as to prove that the statistical model can be used to analyze the data. At the end of the paper, there will be a complete datasheet and a simple survey, which aims to help readers better understand the paper and collect current data on alcohol and tobacco use for subsequent research. Finally, the paper completes the data analysis of alcohol and tobacco use and provides a detailed explanation for the reader.

5.2 Ethics

Through the research of this paper, we found that the tobacco and alcohol have reached everywhere in people's life, and there is a lot of the people under the legal age for time, by means of alcohol and tobacco to try to satisfy their inner vanity, resulting in the future life is dependent on the use of alcohol and tobacco. These days, e-cigarettes are popular among smokers, and they don't have the same restrictions as regular tobacco, such as addiction, but the tar they contain makes them more dangerous. According to the pictures shown in the paper, it is found that people in cities have a great demand for tobacco. According to a CDC report on smokers, smoking is responsible for one-fifth of all deaths in the United States each year. Most smokers are between the ages of 25 and 64, and the most common cause of smoking is financial stress. Driven by stress, people turn to smoking and drinking to relieve the pain of stress. According to these findings, it turns out that the underlying cause of smoking and drinking is the growing economy and the financial burden on people, especially those living in cities.

5.3 Limitation and Futrue Work

In this paper, the data set used contains too much useless data, such as Dont Know in the data set, these data cannot be analyzed, which leads to a great decrease in the actual number of some data, and results of analysis have certain deviation and are not representative enough. In the data analysis part of the paper, the influence of different factors on smoking and drinking was not completely compared, but some representative data were selected for analysis and comparison. The results of a purposeful analysis of selected data can be emotional and cannot be completely fair. However, the establishment of a single model also has certain shortcomings. The establishment of multiple statistical models can ensure the matching degree of the model to the data. Finally, a result with minimum deviation can be obtained through the comparison of multiple models. In general, this paper does not do a good job in establishing models and creating charts, which leads to the shortcomings of this paper. In the future, we need to further study the model and cartographic analysis, and in the future work, the analysis of such data should be more detailed, and the range of data involved should be broader and more representative. In the aspect of model analysis, more factors should be considered and more comparisons should be made. Only through comparison can the most suitable model be obtained. In terms of drawing and analysis, the selection of and data needs to be more random. Only by drawing and comparing the details of relevant data can the conclusion be more accurate. Only by doing this can the analysis be more representative and recognized by readers in many future analyses.

6 Appendixs

6.1 Data Sheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The data set was created to find out what influences people to smoke and drink. People smoke and drink, but why and what affects them? Find out why people smoke and drink and some of the factors that influence them.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The Canadian government created this dataset to represent the country.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The creation of this dataset was funded by the Government of Canada, through the Program Statistics Canada.
4. *Any other comments?*
   - No other comments.

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.*
   - Examples of the dataset are representative of respondents across Canada. And there are multiple types, with gender, place of residence, province, age and education as the main types.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 38 cases in total, and the main types are sex, residence, province, age and education background.
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).*
   - This dataset was selected from a larger dataset containing data on drugs, drugs, alcohol, tobacco and other related matters across Canada. This sample cannot represent a larger set, because the data only selects two related data sets, and each sample represents different information, so it cannot represent a larger data set.
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.*
   - Each instance contains frequency, quantity, and status. This is true for both alcohol and tobacco.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - There is no fixed tag or target association.
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.*
   - There is a lack of information, which is caused by the lack of answers to questions in the process of investigation and collection of original data, resulting in the loss of follow-up dat
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.*
   - There is no very clear relationship between the delay instances, only through data analysis can find the relationship.
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.*

- There is no recommended data split.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - Only missing data exists in the data set. Please refer to question 6 for details.
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.*
    - The data was extracted from ODESI and the source website is Statistics Canada. First of all, this data will always exist, because it is publicly available data provided to the public by the Canadian government, and it only records data from 20171, so it will not change. Secondly, the data details and corresponding Codebook are available on the statistics Canada website, and each instance is well described. Finally, the data is available completely free of charge at no cost, and all information about the data can be found on the Statistics Canada website.
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.*
    - There is no secret in this data, it is all transparent and open.
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.*
    - This data does not contain any data which may be offensive, insulting, threatening, etc.
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.*
    - This dataset does not identify any subpopulations.
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.*
    - The data does not contain any data that may directly or indirectly identify individuals.
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.*
    - The dataset does not contain any data that might be considered sensitive.
16. *Any other comments?*
    - No other comments.

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.*
    - Data for each instance were obtained from the citizen awakening questionnaire and released after statistics Canada validated the 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 data was collected using Internet software and telephone consultations.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - If the sample is drawn from a larger set, the sample should be drawn at random. The same number can be drawn at random for different ages or genders to be used as a new data set.

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)?*
    - No one was involved in the data collection process because it was all done by software programs and telephone calls.
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 were collected in 2017 from February to December, making the data more representative due to the wide time range.
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 ethical review process.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - Data provided by a third party website – Odesi.
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.*
    - The parties concerned were not informed of the collection of information as the data were transparent and public.
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.*
    - The person concerned consents to the collection and use of the information as it is open and transparent.
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 need to provide a mechanism.
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.*
    - There is no analysis of the potential impact of data sets and their use on data topics.
12. *Any other comments?*
    - There is no other commonts.

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 has been processed and cleaned. All the data that can be transformed into dummy variable are transformed into words from numerical, and some data irrelevant to the studied problem are appropriately deleted.
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.*
    - https://search1.odesi.ca/#/details?uri=%2Fodesi%2FCTADS-82M0020-E-2017-person.xml
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R Studio
4. *Any other comments?*
    - No other comments.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - The dataset is not currently being used for any task.
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.*
   - None
3. *What (other) tasks could the dataset be used for?*
   - Data sets record drug, drug, alcohol and tobacco use and trends across Canada.
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?*
   - The collection or composition of a data set does not affect future use because the data is fixed.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No bad data.
6. *Any other comments?*
   - No other comments.

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.*
   - The data set will not be distributed to the public for reference, as the data is for personal analysis only.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The data will be stored in Github and in doi.
3. *When will the dataset be distributed?*
   - Dec. 2017
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.*
   - None
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.*
   - None
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.*
   - None
7. *Any other comments?*
   - Not other comments

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The Government of Canada will support, host and maintain the dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Contact Statistics Canada.
3. *Is there an erratum? If so, please provide a link or other access point.*
   - None
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)?
   - The data will not be updated.
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.*
   - There is no human-related data in the data set, and no privacy exists.
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.*
   - The data records relevant data for 2017. Consumers can extract information from past data sets and compare it with current data to find trends.
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.*
   - There are no mechanics, unless they have a time machine to travel back to 2017.
8. *Any other comments?*
   - No other comments.

6.2 Survey

Please follow up the link below to fill up the survey. Thank you!

https://docs.google.com/forms/d/e/1FAIpQLScueZ0SvEsU6vkhTG-BC3FHiQAQoQqjNob1IeZbDXLPZ_2wyw/viewform?usp=sf_link

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Azagba, Sunday, and Mesbah F Sharaf. 2011. "The Effect of Job Stress on Smoking and Alcohol Consumption." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3403311/#B1.

CAMH. 2022. "Alcohol." https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/alcohol.

Canada, Statistic. 2020. "Smoking, 2019." https://www150.statcan.gc.ca/n1/pub/82-625-x/2020001/article/00003-eng.htm.

———. 2021. "Alcohol and Cannabis Use During the Pandemic: Canadian Perspectives Survey Series 6." https://www150.statcan.gc.ca/n1/daily-quotidien/210304/dq210304a-eng.htm.

Institute, National Cancer. 2022. "Cigarette." https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cigarette.

Mark A. Dixon, L. C. S. W., and M. S. W. Karen G. Chartier Ph.D. 2016. "Alcohol Use Patterns Among Urban and Rural Residents." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4872615/.

Marks, Hedy. 2021. "Stress Symptoms." https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body.

Müller, Kirill, and Hadley Wickham. 2021. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

N. J.Doogan, M. E.Wewers, M. E.Roberts, and S. T.Higginse. 2017. "A Growing Geographic Disparity: Rural and Urban Cigarette Smoking Trends in the United States." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600673/.

R Core Team. 2021. *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 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.