A Brief Overview of Tabacco, Vaping and Cannabis Use in Canada*

Shenghui Qiao

27 April 2022

Abstract

Use of tobacco, cannabis and other addictive products has always been an important but controversial social topic. This report analyzes public use of cigarettes, vaping and cannabis in Canada based on the data collected in the national Canadian Tobacco and Nicotine Survey (CTNS). Results showed that respondent's personal conditions including gender and age, as well as information of respondent's first usage of these addictive products - age of respondent and type of product at the first usage - are important in determining whether this respondent is likely to be a smoker or vaper now. These findings may have great significance for future policy makers in regulating and controlling addictive products.

Keywords: tobacco, nicotine, cannabis use, Canada, Canadian Tobacco and Nicotine Survey (CTNS), health, society and community

1 Introduction

Appropriate control of addictive products has always been an important issue in society. As early as in 1997, government of Canada enacted the Tobacco Act in response to the national public health problem posed by tobacco use. In May 2018, the new Tobacco and Vaping Products Act (TVPA) retained key components from the former act, and further regulated the manufacture, sale, labelling and promotion of tobacco and vaping products sold in Canada (Canada (2018)). The TVPA created a new legal framework that protects the health of Canadians, especially young persons, from tobacco related disease. However, later in the same year, access to cannabis was legalized by the Cannabis Act, which provided more freedom of smokers and other addicts. Statistics showed that by 2019, 16.8% of Canadians aged 15 or older reported use of cannabis, which is 1.9% higher than the percentile in 2018 before cannabis legalization (Rotermann (2020)).

While more and more people are using tobacco, nicotine or cannabis plant as a relief from stress, concern and opposition to these addictive products exist all the time. Research showed that tobacco has causal relationship to some serious lung-related diseases, such as lung, colon or rectal cancers. In the United States, over 400,000 deaths could be attributed to tobacco smoke each year (Robert (2005)). Similarly, according to Health Canada, use of cannabis may cause symptoms including confusion, fatigue, ability impairment, anxiety, and lung infections (H. Canada (2022)). Besides, cannabis abuse is regarded as the second primary factor of causing bipolar disorder (BD), which is one of the most serious psychiatric disorders in terms of morbidity and mortality (Catherine M Cahill (2006)).

Analysis of data reveals that most people started smoking or vaping at an early age of 15-19. Findings about the frequency of usage show that a large proportion of people have a daily consumption of these addictive products - 76.77% consumes cigarettes daily, about one half of respondent reports vaping daily, and 38.76% smokes cannabis every day. Besides, while cigarette is preferred by elder age groups, young people tend to use more vaping or cannabis. The interrelation among cigarette, vaping and cannabis is also investigated. An important reason that people started vaping is to either cut down smoking cigarette or to try quit smoking. This could be because a large proportion of people perceive the harm of vaping the same as, or less than the

*Code and data are available at: https://github.com/qiaoshe1/STA304-Final-Paper.git

harm of cigarette. People's preference of pure cannabis and tobacco-cannabis mixture is also investigated, however, tobacco-cannabis does not turn out to be more attractive than pure cannabis.

A generalized linear model (GLM) is chosen to investigate what factors potentially affect the respondent's choice of becoming a user of cigarette, vaping product or cannabis. Results showed that gender, age, first product that respondent tried, and the age when the respondent used cigarette, vaping product or cannabist for the first time are statistically significant in determining the respondent's status of smoking (vaping) or not.

This report aims to provide a brief overview of the use of tobacco, vaping products and cannabis in Canada. It also provides critical commentary on the data collection and survey methodology. All the analysis will be based on the 2020 Canadian Tobacco and Nicotine Survey (CTNS) conducted by the Statistics Canada. Section 2 provides an overview of the data set and comments on the data collection. Section 4 builds a model based on response variable and certain predictors. Section 5 presents results and section 6 presents discussions of these findings, as well as limitations and future directions of the study. Finally, section 6.1 includes important enhancement to the study.

2 Data

The data set is from the Public Use Microdata File (PUMF) of the Canada Tobacco and Nicotine Survey (CTNS) conducted by Statistics Canada in 2020 (Alcohol and (CADS) (2021)). Statistics Canada conducted CNTS on the behalf of Health Canada in order to fill the important information gap of vaping, cannabis, and tobacco usage in Canada. The CTNS data includes information of several topics - the prevalence of cigarette smoking, vaping, cannabis use, alcohol use, as well as some basic information of the respondents. This report will analyze frequencies of smoking, vaping and cannabis use and potential factors which influence these frequencies.

The data set contains answers from 8112 respondents who are aged 15 years and older across 10 provinces. Answers are collected through an electronic questionnaire or telephone interview. In total, the data set has 99 variables, each corresponds to respondent's answer to one question. Variables that are informative on a larger scale are kept, such as age of respondent when he/she started using tobacco products or cannabis, frequency of smoking or vaping in the last 30 days, number of times tried quit smoking, etc. Variables related to respondents' background information such as gender, age and provice are also included. Because this report aims to provide an overview of each addictive product's usage in Canada, some variables that are too specific are dropped, such as number of cigarettes smoked or vaped in each day. These would likely be redundant for the purpose of this report. In the original data set, answers are coded by numbers. In the cleaning process, I changed the variable names and uncoded answers with the *CTNS Data Dictionary book* in order to make the new data set more informative and easy for people to understand.

Figure 1 shows the distribution of some background information of the 8112 respondents who took part in the survey. The number of female and male respondents are very close, meaning that this survey avoids potential bias on selecting gender of respondents.However, table 1 further shows that although the overall number of female respondent is slightly larger than that of male respondents, the number of male who reported usage of cigarettes, vaping products and cannabis is actually higher than number of female in all three categories. That is, a higher proportion of male would report usage than female does. In all 3876 male respondents, 48.74% smoked cigarettes, 23.25% vaped, and 42.11% smoked cannabis; while these proportions of female are 41.36%, 16.76% and 35.6% respectively.

Figure 1 also shows that most respondents of this survey are aged 65 and older, while least people are in the 25-34 age group. Number of respondents in all other age groups are about 1000. The distribution of age group does not show any pattern, but because one age group has much more respondents than other groups, this inclination may have an effect on the analysis of frequency of usage. Simlarly, table 2 is created to show the relation between age groups and number of people who smoked or vaped. The age group with most people smoked cigarettes is 65 and older - in 1912 respondents in this age group, 1206 of them smoked cigarettes (63.08%). We can also observe a clear tendency of youth usage of vaping products and cannabis,

as the age group 20-24 years old reported most people vaping and smoking cannabis. Furthermore, table 2 shows that elder people have a preference of smoking cigarettes, while young people tend to use vaping products. However, cannabis seems to be equally attractive to all ages.

The bottom two plots in figure 1 show province that respondent is from and the first product that the respondent tried. A large proportion of respondents are from either Ontario or Quebec, while there are fewer people from other provinces. This may introduce sampling bias to the further analysis of usage across Canada because the data set has too many observations from the same group that is province. Lastly, the most popular product that people try first is cigerattes. Among all 2709 people who have tried cigarettes, vaping or cannabis, 1870 of them started from smoking cigarettes. That is, cigarettes is the first tried product for 69.03% of smokers based on this survey.

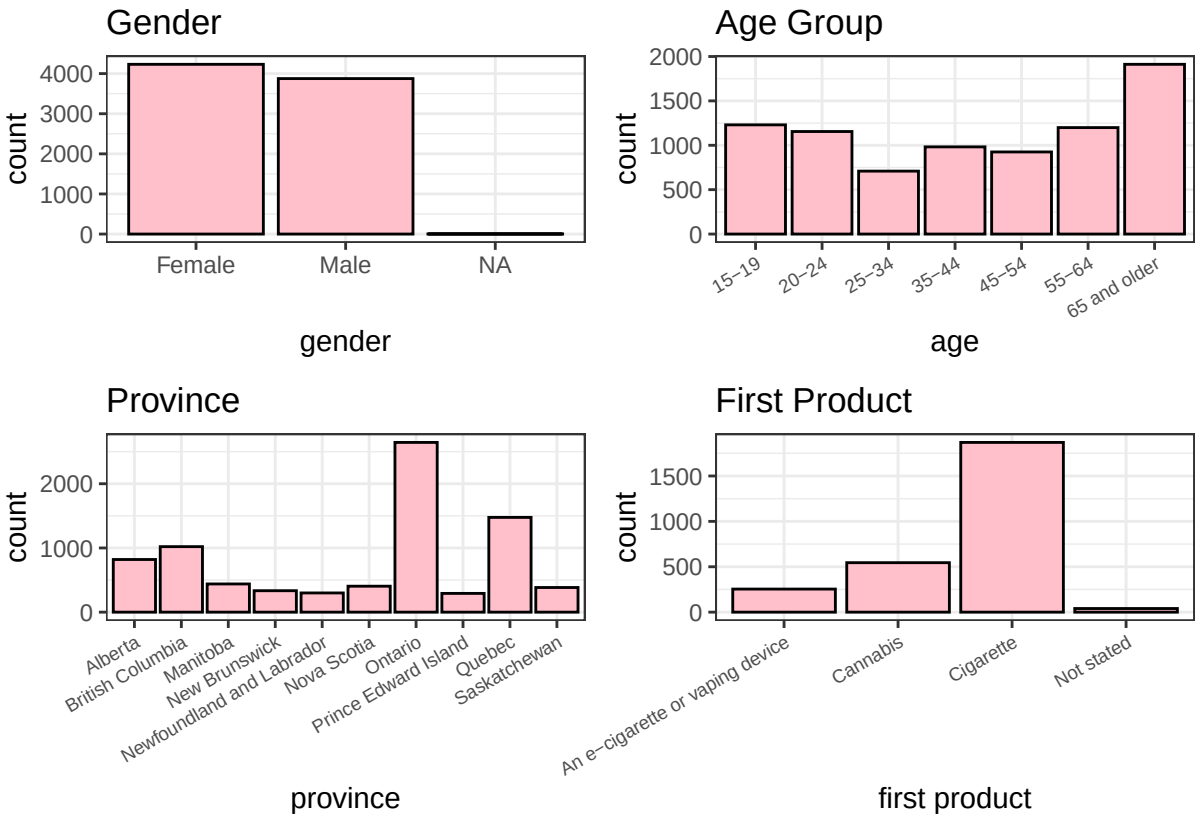

Figure 1: Background Information of Respondents

Table 1: Number of People Smoked Ciarettes, Vaped and Smoked Cannabis by Gender

gender	Female	Male
number of people smoked cigarettes	1752	1889
number of people vaped	710	901
number of people smoked cannabis	1508	1632

Table 2: Number of People Smoked Ciarettes, Vaped and Smoked Cannabis by Age

age	15-19	20-24	25-34	35-44	45-54	55-64	65 and older
number of people smoked cigarettes	125	328	294	467	485	739	1206
number of people vaped	433	511	198	167	98	137	71
number of people smoked cannabis	275	595	412	492	367	541	462

Throughout the report, We used R to conduct our analysis (R Core Team (2020)). R packages `tidyverse` (Hadley Wickham [aut 2021), `knitr` (Yihui Xie ORCID iD [aut 2021), `reshape2` (Wickham 2020), `janitor`(Sam Firke [aut 2021), `kableExtra` (Hao Zhu ORCID iD [aut 2021), `patchwork` (Pedersen 2020), `lme4` (Douglas Bates ORCID iD [aut] 2022), and `lmtest` (Torsten Hothorn ORCID iD [aut] 2022) are used for data cleaning, analysis and discussion.

3 Survey Methodology

The target population of CTNS survey is non-institutionalized persons aged 15 years or older, living in Canada, who are not members of collectives or living on reserves. In total, 8112 respondents across all 10 provinces took part in the survey. They represent a weighted total of 31.3 million Canadian residents who are over 15 years old. The process of data collection lasted six weeks. It commenced on December 8th, 2020, and ended January 16th, 2021.

The survey was conducted through electronic questionnaires, which is designed by Statistics Canada in consultation with Health Canada, or telephone follow-up interviews. Data are collected directly from survey respondents and on a voluntary basis. The questionnaires are divided into several different sessions, each containing related questions. These sessions are age-order selection (AOS), demographics (DEM and DEM2), gender (GDR), tobacco (TBC), other tobacco product status (OTP), vaping (VAP), cannabis (CAN), initial use (IU), alcohol (ALC), and feedback (FDB). Among them, age-order selection, demographics and gender include questions of respondent's background information; tobacco, other tobacco product status, vaping and cannabis concern the frequency, reason and variability of respondent's usage; initial use, alcohol and feedback provide extra information to the survey and gather comments from the participants.

To detect and avoid error in the survey, the "raw" electronic survey file which contains the completed respondent survey records were created. Before further processing, verification was performed to identify and eliminate potential duplicate records and to drop non-response and out-of-scope records. At the cleaning and verification stage, further editing was performed to identify errors, remove empty records and invalid data, and format remaining data.

Each region who participated in the survey adopted a uniform, cross-sectional sample design which included a systematic, stratified sample. The CTNS sample has a one-stage design for persons who are 15 to 24 years old, and a two-stage design for persons who are 25 years old and older - the first stage being the dwelling and the second stage being the person. Also, the stratification method is different based on the age groups. For the 15-24 years old, the frame was stratified by age group and province, and sample was selected independently within each age group and province. And for the 25 years old and older, the frame was stratified by province, and a simple random sample of dwellings was selected independently within each province.

All content in this session are from CTNS description and summary by Statistics Canada (Alcohol and (CADS) (2021), G. of Canada (2022a)).

4 Model

This report wants to investigate what factors have an effect on people's decision of using cigarette, vaping product, and cannabis. Respondents are divided into category of "smokers" and "non-smokers" based on their frequency of smoking or vaping in past 30 days. We thus have a binary outcome. And the model relies on binomial logistic regression.

The model assumes that:

$$P(Y = 1|X_k) = \frac{exp(\beta_0 + \beta_1 X_1 + ...\beta_k X_k + ... + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + ...\beta_k X_k + ... + \beta_n X_n)} \tag{1}$$

where:

- Y is the binary response that indicates whether the respondent used cigarette, vaping, or cannabis in the past 30 days;
- X_k is predictors or independent variables;
- β_0 is y-intercept;
- Other β_k is the coefficient for each corresponding predictor;
- n is number of predictors.

The model estimates the probability P of the outcome variable Y to be successful given the predictor variable(s) X_k. Cigarette, vaping, and cannabis are treated as three independent cases, and have different response variables. The binary response variables for every model are created based on variable "frequency of smoking cigarettes (vaping/smoking cannabis)". If the answer to frequency of smoking is "Not at all" or a "Valid skip", this respondent is identified as a non-smoker; in contrast, if the respondent reported smoking or vaping at least once in the past month, he or she is identified as a smoker or vaper.

The independent variables for these three cases are the same, which are age, gender, province, age when respondent started smoking or vaping, first product that respondent tried, and frequency of drinking alcohol in the past 30 days. These independent variables capture background information of the respondent, early effect of first product and age started smoking/vaping, as well as the effect of alcohol. Because these predictors are categorical, dummy variables are also used to represent each option of answer, so that the modelling will be simpler. For example, the age group "15-19" is expressed as 1 and "20-24" is expressed as 2. Note that for any product, its model will include predictors of age that the respondent tried the other products only. For instance, model of cigarette includes variable `age_first_time_vaped` and `age_first_time_cannabis`, as we want to see whether the usage of vaping and cannabis affect usage of cigarette. In contrast, variable `age_first_time_cigerette` will not be included in this model, because this variable explicitly indicates that the respondent is a smoker.

Observations in our data set are independent from each other because they are from different individuals. Also, our preferred response type is binomial. Therefore, a generalized linear model (GLM) would fit our data. This type of model provides some flexibility for the non-linear response type, but is a good fit because the observations do not have any grouping. For each case of cigarette, vaping and cannabis, three models are built. These 3 models are nested; they have the same response but different predictors.

For instance, in case of cigarette:

Model 1_1: predictors are gender, age, and province

Model 1_2: predictors are gender, age, province, age when first tried smoking, and first product

Model 1_3: predictors are gender, age, province, age when first tried smoking, first product, and frequency of alcohol usage in past 30 days

With ANOVA test and likelihood ratio test, I compared these three models using cigarette as response. The results for ANOVA are shown in table 3 and likelihood ratio test are shown in table 4

Table 3: ANOVA Results

	Df	Deviance	Resid. Df	Resid. Deviance
Model 1_1				
Intercept	NA	NA	8107	5206.21
Gender	1	23.93	8106	5182.28
Age	1	52.80	8105	5129.48
Province	1	0.07	8104	5129.41
Model 1_2				
Intercept	NA	NA	8107	5206.21
Gender	1	23.93	8106	5182.28
Age	1	52.80	8105	5129.48
Province	1	0.07	8104	5129.41
Age first time vape	1	373.58	8103	4755.83
Age firt time smoke cannabis	1	136.96	8102	4618.87
First product	1	201.95	8101	4416.92
Model 1_3				
Intercept	NA	NA	8107	5206.21
Gender	1	23.93	8106	5182.28
Age	1	52.80	8105	5129.48
Province	1	0.07	8104	5129.41
Age first time vape	1	373.58	8103	4755.83
Age firt time smoke cannabis	1	136.96	8102	4618.87
First product	1	201.95	8101	4416.92
Frequency of alcohol usage	1	4.05	8100	4412.87

Table 4: Likelihood Ratio Test Results

	X.Df	Loglikelihood	Df	Chisq	Pr(>F)
Model 1_1	4	-2564.705	NA	NA	NA
Model 1_2	7	-2208.460	3	712.489	0.000
Model 1_3	8	-2206.437	1	4.047	0.044

From the likelihood ratio test results, the second model suits best for our data. That is, the model with all predictors except for frequency of drinking alcohol. Applying these three models to the case of vaping and cannabis generated similar results. Therefore, model 2 will be our choice of final model.

The composition equation of Model 2, which is our chosen model, looks like:

$$log(Y) = \beta_0 + \beta_1 * gender + \beta_2 * age + \beta_3 * province + \beta_4 * age_1 + \beta_5 * age_2 + \beta_6 * firstproduct$$

where age_1 and age_2 are the age when respondent first try the other two products. For instance, when Y is whether respondent smokes cigarettes, age_1 and age_2 are ages when respondent first try vaping and smoking cannabis.

5 Results

5.1 Cigarettes

Among all 8112 respondents, 3644 of them had smoked cigarettes in the last 30 days of the survey, that is, 44.92% of the population. As shown in figure 2, 53.07% of these people started smoking at an early age of 15-19 years old, and 28.05% of them started smoking even younger, at an age of 10-14 years old. Youth use of tobacco therefore seems to be a crucial problem in Canada. Also, this results further confirms discovery of an early research in the United States which suggests that "nearly 9 out of 10 adults who smoke cigarettes daily first try smoking by age 18" (Health and Services (2012)). Compare to proportion of teenager smokers, people who start smoking at an elder age is significantly smaller.

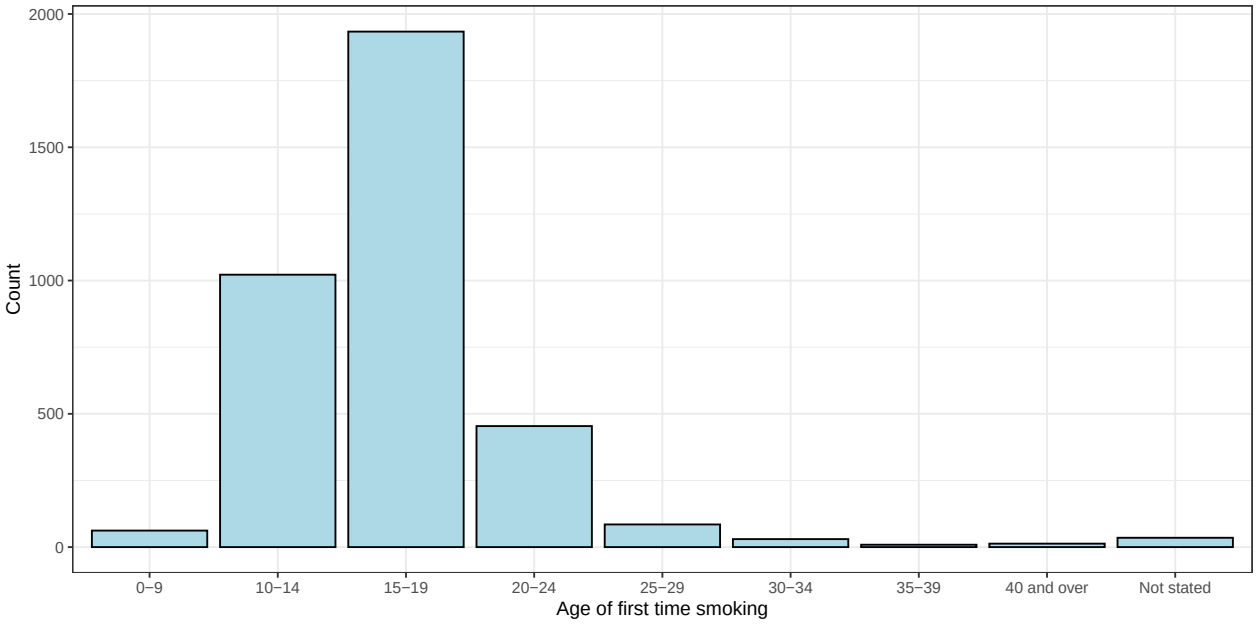

Figure 2: Histogram of Respondents' Age of First Smoking

Among the 8112 respondents, a large proportion of them reported smoking at least one tobacco product in the past 30 days. As figure 3 shows, the undoubtly most popular tobacco product to smokers is cigarette. Specifically, 75.43% of smokers smoked cigarettes recently. Cigars and little cigars are the second and third popular options, which make up 8.86% and 5.9% respectively. It is worth noticing that a majority (i.e. 76.77%) of cigarette smokers would smoke every day, and 56.45% of little cigar smokers also do this daily. However, none of the other tobacco products are consumed daily. Instead, we can see from figure 3 that some people report usage of chewing tobacco, cigars, tobacco pipe or tobacco waterpipe at least once a week, but a majority of them would use these products less than once a week, but at least once in the past month. Compare to cigarettes and little cigars, other tobacco products are used by less people and consumed with a much lower frequency.

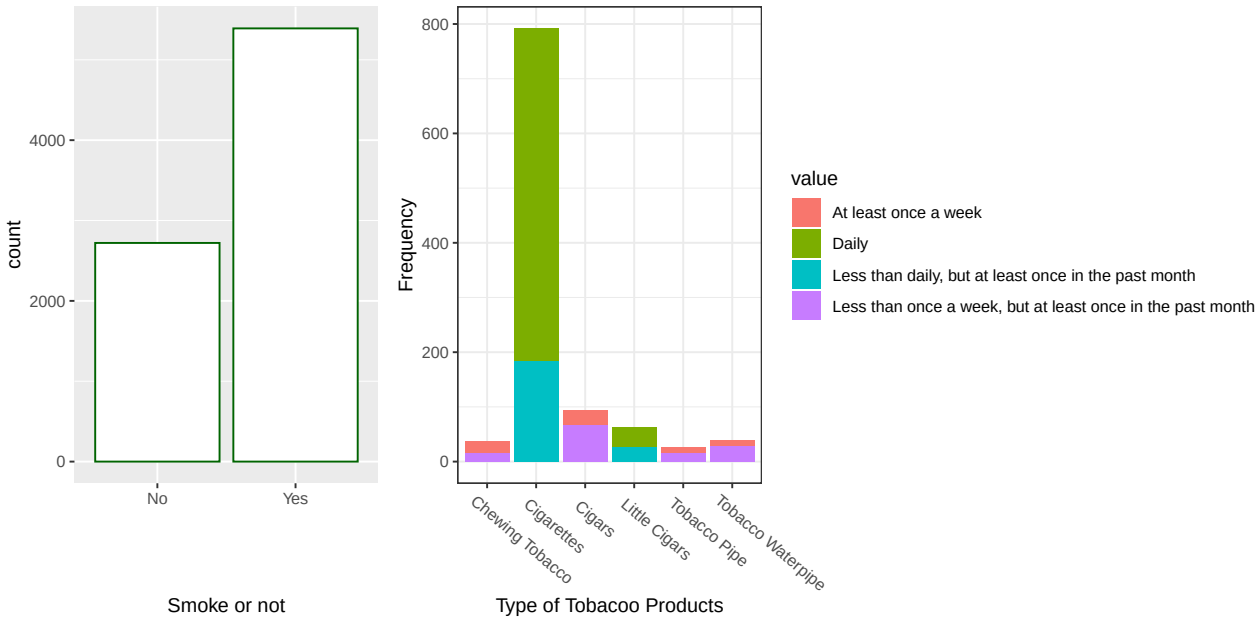

Figure 3: Proportion of smoker and non-smoker, and frequency of smoking across different tobacco products

Lastly, figure 4 shows the distribution of number of times one respondent tried to quit smoking by any method. Despite that there are other options for the question about when the respondent quit smoking - 1 to 2 years ago, 3 to 5 years ago, and more than 5 years ago - all respondents who answered this question stopped smoking only less than one year ago, based on our data. Among these people, 27.38% of them quit smoking successfully in the first time. 22.62% of people tried once, 28.57% tried twice or three times, and 21.43% tried 4 times or more. The addictiveness of nicotine in tobacco products make quitting smoking difficult. Thus, most people tried and failed at least once in quitting smoking before they succeed.

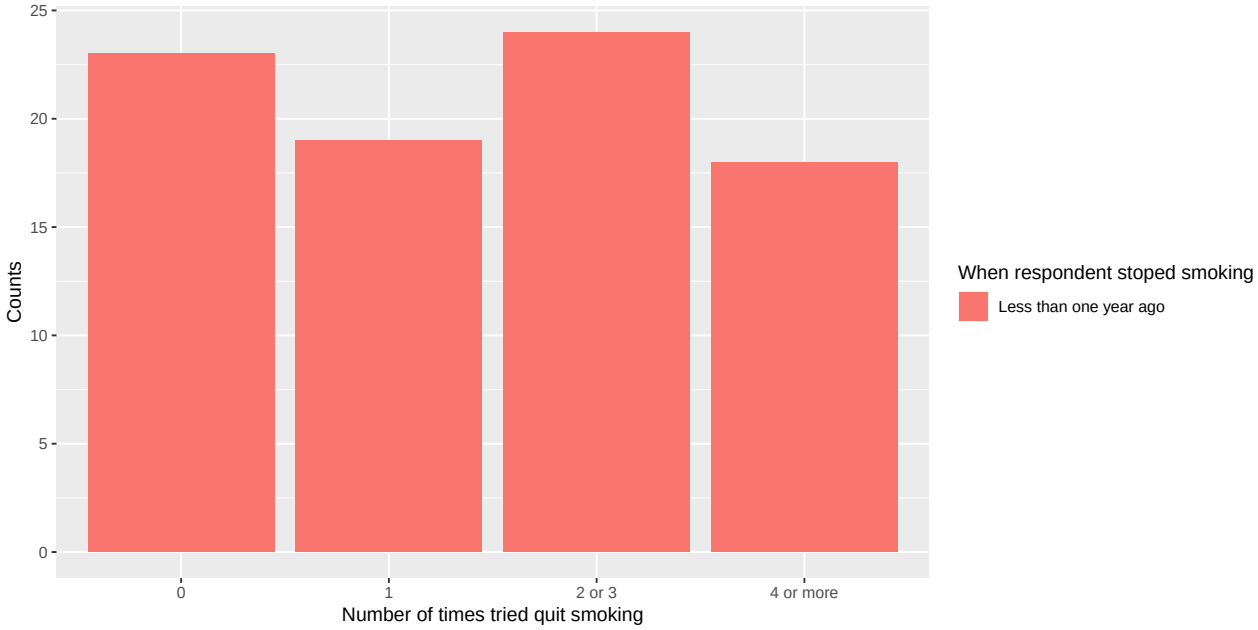

Figure 4: Histogram of Number of Times Respondent Tried Quit Smoking

5.2 Vaping

Among all respondents, 19.91% of them has vaped in the past 30 days. Figure 5 shows the distribution of respondents' age when they first tried vaping. Similar to the case of cigarettes, 649 respondents (40.19% of vapers) were exposed to vaping at an early age of 15-19 years old. Besides, 13.99% and 8.54% of people who vaped started vaping at 20-24 and 10-14 years old, respectively. This reveals a vaping prevalence among teenagers and young people in Canada. As it is illegal to sell or provide vaping products to anyone under 18 (in some provinces, this age is increased to 19 or 21), the majority of youth obtained vaping products from social sources such as acquaintances, friends and family, rather than from retail (G. of Canada (2022b)). Although Canada has a regulatory framework for vaping products, especially with a focus on preventing uptake by youth, the results of the CNTS data actually suggest a incompetence of this regulation, and more serious punishment should be promoted. It is also worth noticing that a small proportion of people started vaping at an elder age of 50-59 years old, mainly to cease smoking or to avoid returning to smoking. The main reason of vaping would further analyzed at the end of this section.

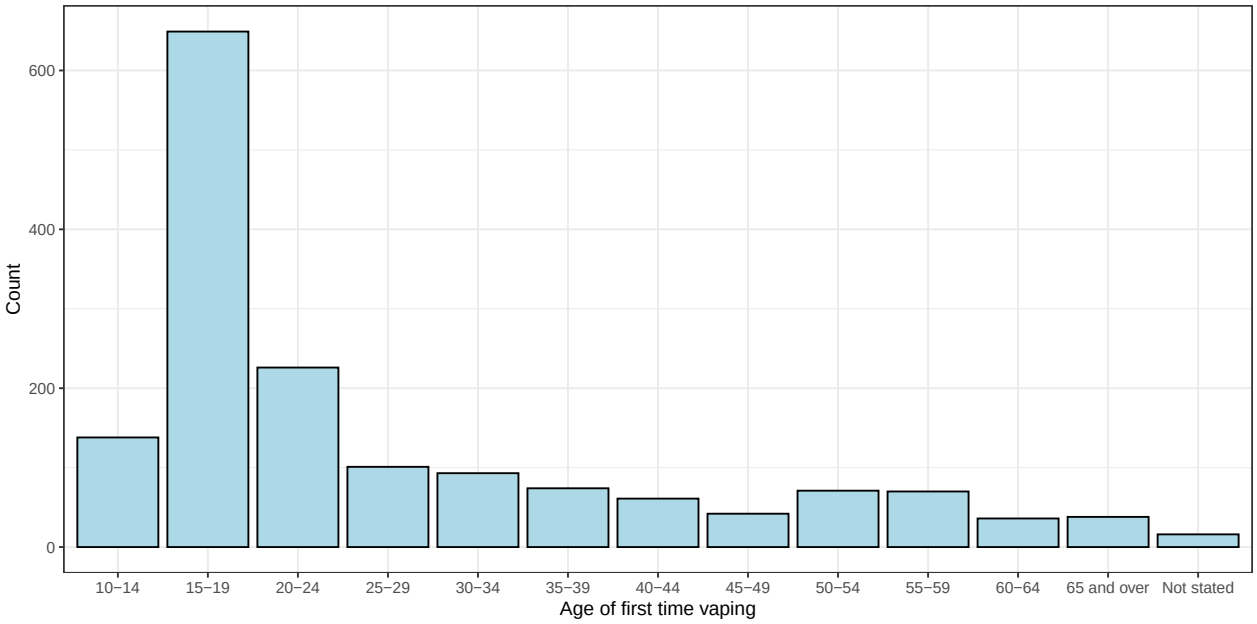

Figure 5: Histogram of Respondents' Age of First Vaping

As shown in figure 6, 70.32% of people did not vape in the past 30 days at all, which is a majority of respondents. Among those who vaped, about half of the sample - 46.76% - vaped everyday, and the rest half vaped less frequently, either less than daily or less than once a week but at least once in the past month. Compare to the case of cigarettes, the number of vapers is smaller and the frequency of vaping products usage is also lower. Vaping thus seems less addictive or attractive to its users.

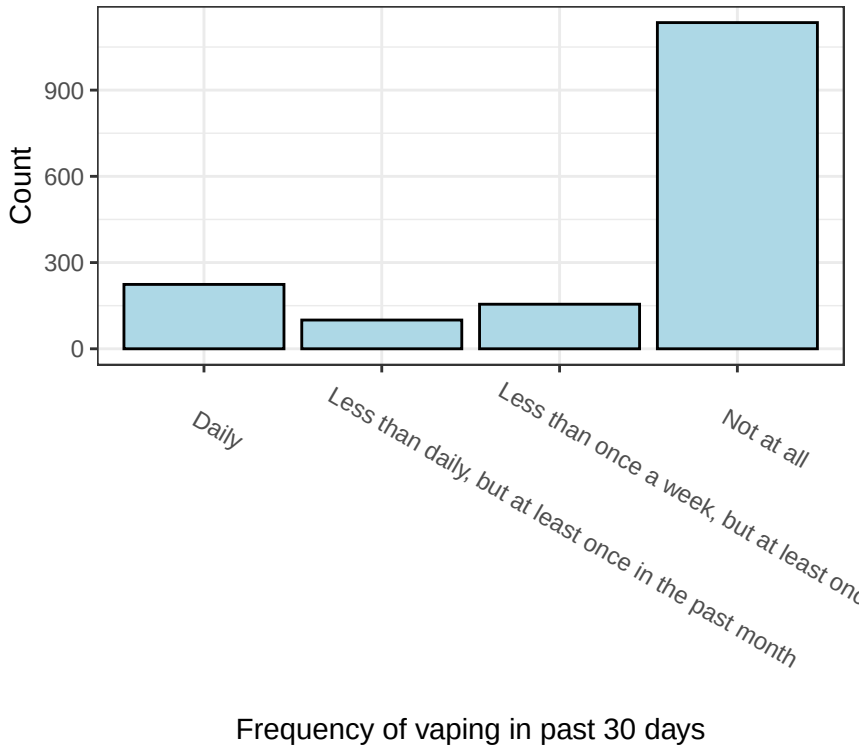

Figure 6: Frequency of Vaping in Past 30 Days

Figure 7 shows number of times that respondents tried to quit vaping and their perception of harms of vaping. These two variables have a casual relationship so they are plotted in the same figure. The distribution of number of times tried quit vaping in figure 7 seems a bit "extreme", because most vapers either did not try quitting at all or tried four times or more in that month. While a large proportion (i.e. 57.47%) of vapers tried zero times, there are 20% of people have stopped vaping for one day or longer for 4 times or more because they want to quit vaping. Number of people who tried 1, 2 or 3 times are smaller.

Figure 7 also states current vapers' perception of harm of vaping. About one quarter of respondents think the harm of vaping is about the same as cigarettes. 43.02% of people perceive vaping as either somewhat or much less harming than cigarette, which explains the reason why some people use vaping products as a replacement of cigarettes. 14.28% of respondents think vaping is either somewhat or much more harmful than cigarettes, but they did not stop vaping or replace vaping by other products. It is worth noticing that 13.72% of vapers actually don't know the harm of vaping or don't have a perception of how harmful it would be to their health. The results indicate that although to most people who are currently vaping, vaping is either same harmful as or less harmful than cigarettes, there is no consensus among the public. In fact, the harm of vaping on human body is still uncertain and is a controversial topic today. Besides, some vapers did not realize the negative effect of vaping at all. More research and publicity about vaping products and their effects are needed to clear the mystery.

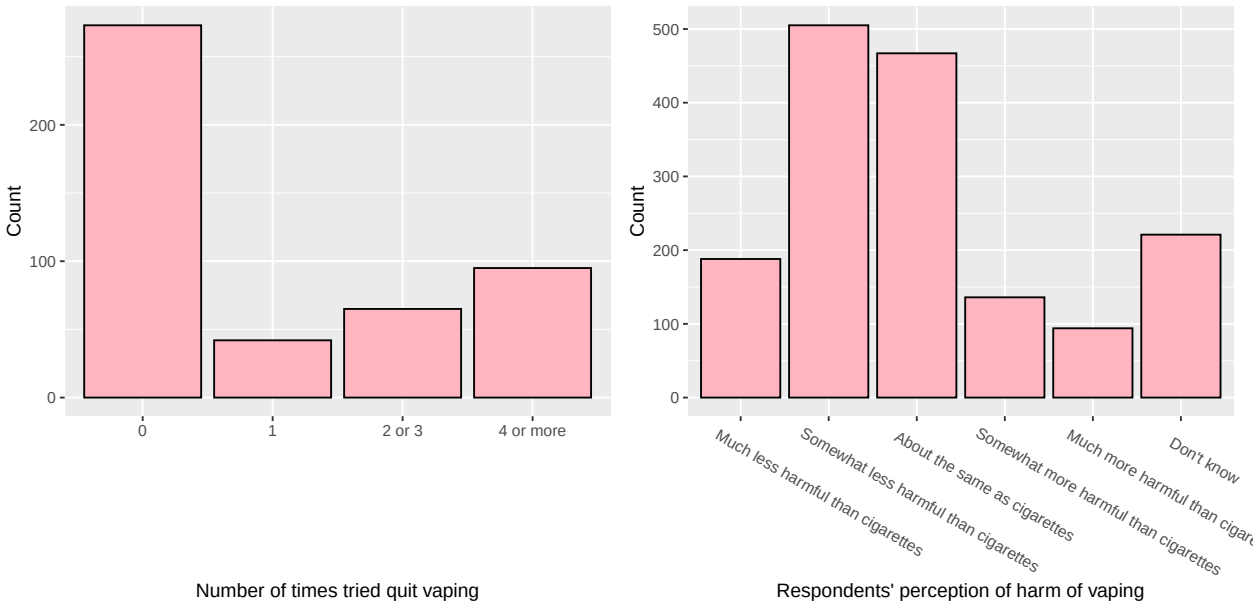

Figure 7: Number of Times Tried Quit Vaping and Respondent's Perception of Harm of Vaping

Lastly, figure 8 shows the main reason of vaping usage. The most popoular reason of vaping is related to the smoking of cigarettes, and the results show that many people use vaping as an effective alternative of cigarette. In total, 31.45% of people started vaping because they want to quit smoking cigarettes, to cut down on smoking cigarettes, to avoid returning to smoking cigarettes, or to use when cannot/not allowed to smoke cigarettes. To these people, usage of vaping is in fact an attempt to get rid of cigarettes, which may cause more harm to them. This further suggests that more information about the harm of cigarettes and vaping is needed. Other popular reasons of vaping are to reduce stress, to enjoy it, or to satisfy curiosity. And the proportion of respondents who choose these three options are 20.13%, 19.92% and 17.82% respectively. Besides, 10.69% of people vape because of other reasons.

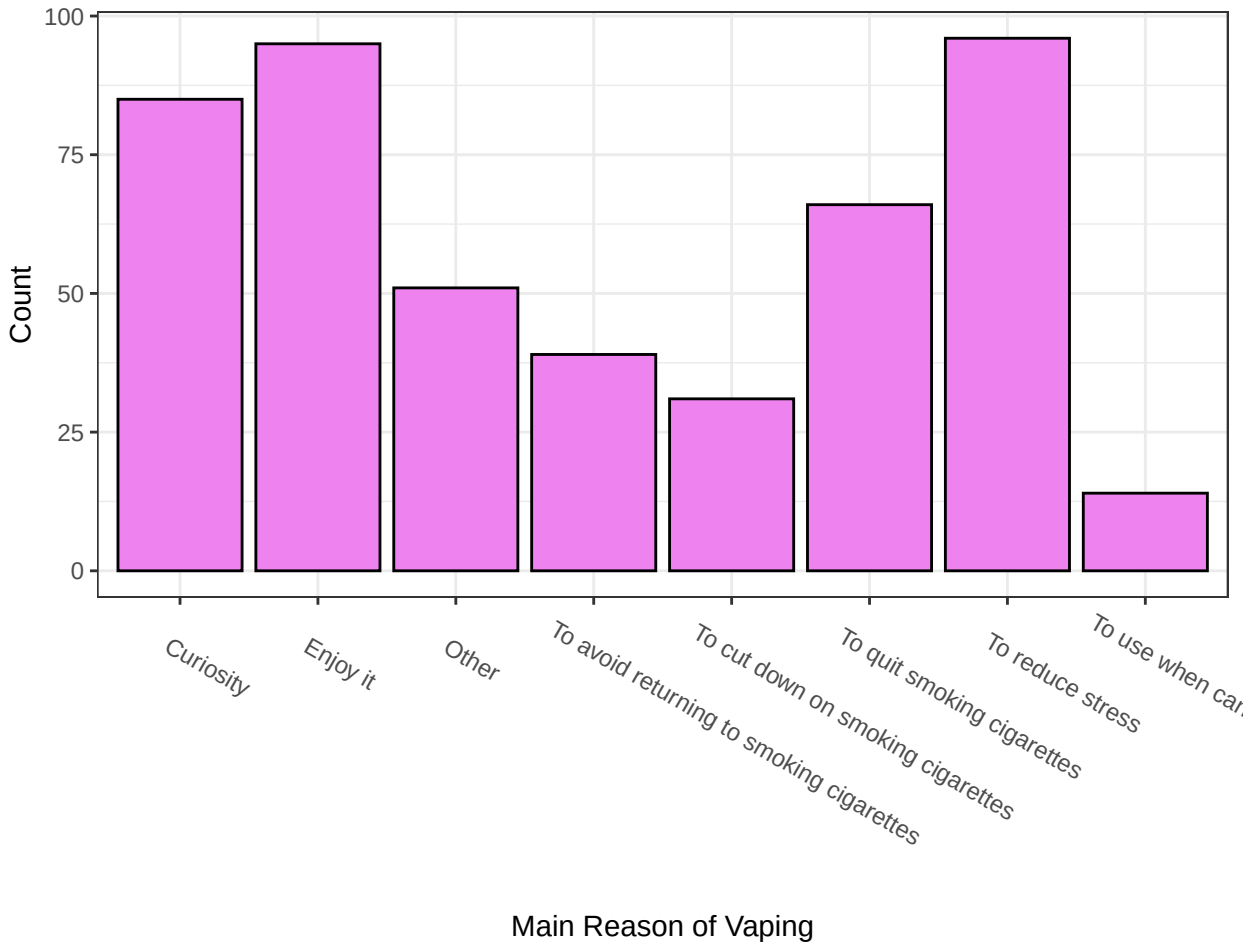

Figure 8: Main Reasons of Vaping

5.3 Cannabis

In total, 38.76% (i.e. 3144 people) of all respondents reported cannabis smoking in the last 30 days. More than half of the respondents (59.38%) first smoked cannabis at 15-19 years old. And similar to the case of cigarettes and vaping, the two age groups very likely to be exposed to cannabis were those aged 20-24 (16.48%) and 10-14 (12.69%). This finding further suggest the importance of regulating youth usage of tobacco and cannabis. The histogram is rightly skewed just like the histograms for cigeratte and vaping. As age increases, the number of people who started smoking cannabis decreases, and this is intuitive.

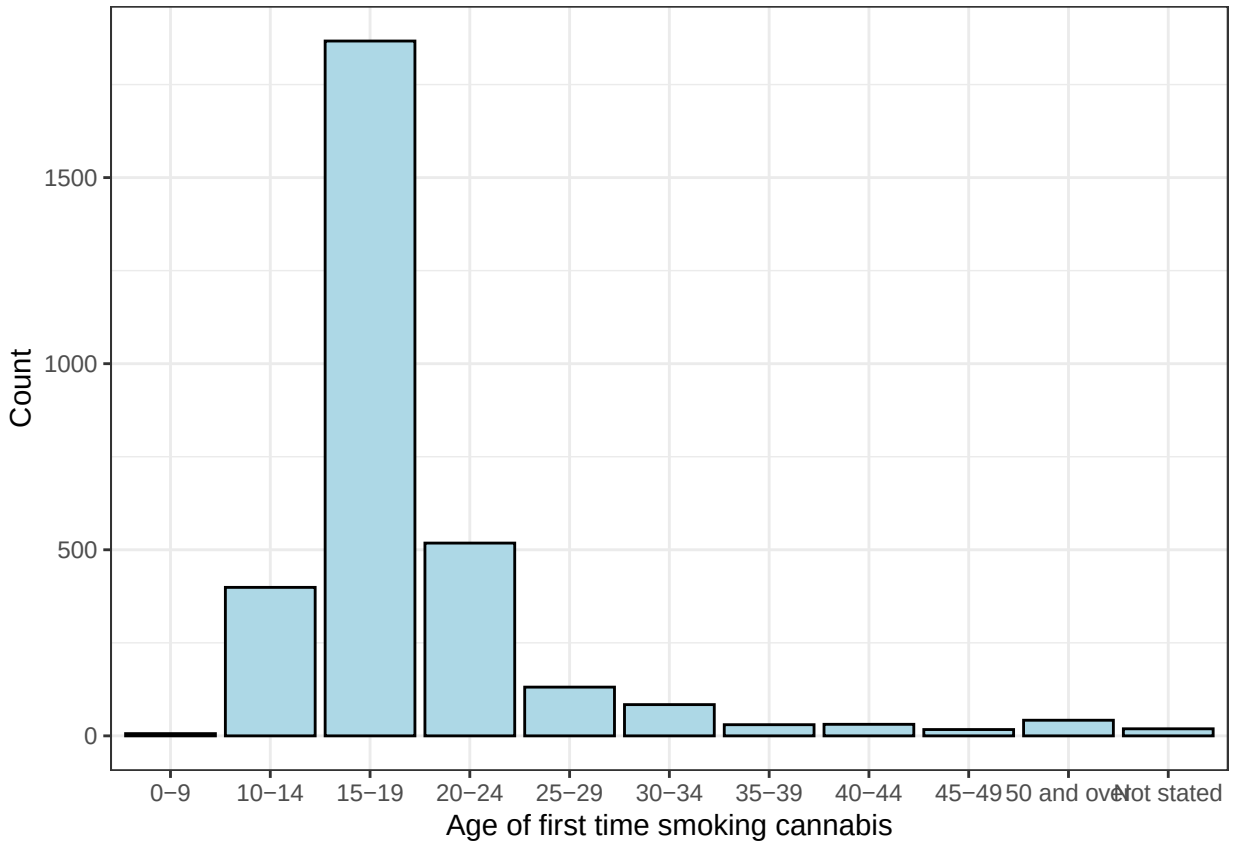

Figure 9: Histogram of Respondents' Age of First Smoking Cannabis

According to figure 10, 80.66% of cannabis smokers smoke pure cannabis only, and 19.34% smoke both pure cannabis and a tobacco-cannabis mixture. The pure cannabis product is still very popular, and the mixture product hasn't gained much attention from consumers. None of the respondents smoke only the tobacco-cannabis mixture.

The plot on the right of figure 10 shows the frequency of usage for cannabis and tobacco-cannabis mixture. It is worth noticing that all users of the tobacco-cannabis mixture are users of pure cannabis products, based on our finding above. The distribution of frequency of usage seems very balanced for both types of product, with each frequency makes up about one-third of the count. Specifically, 30.24% of people smoke cannabis or mixed product every day, 28.8% smoke less than daily, but at least once a week, and the rest 40.96% smokes less frequent, but at least once in the past month. From the figure, there is no obvious relationship between the frequency of smoking and the type of product.

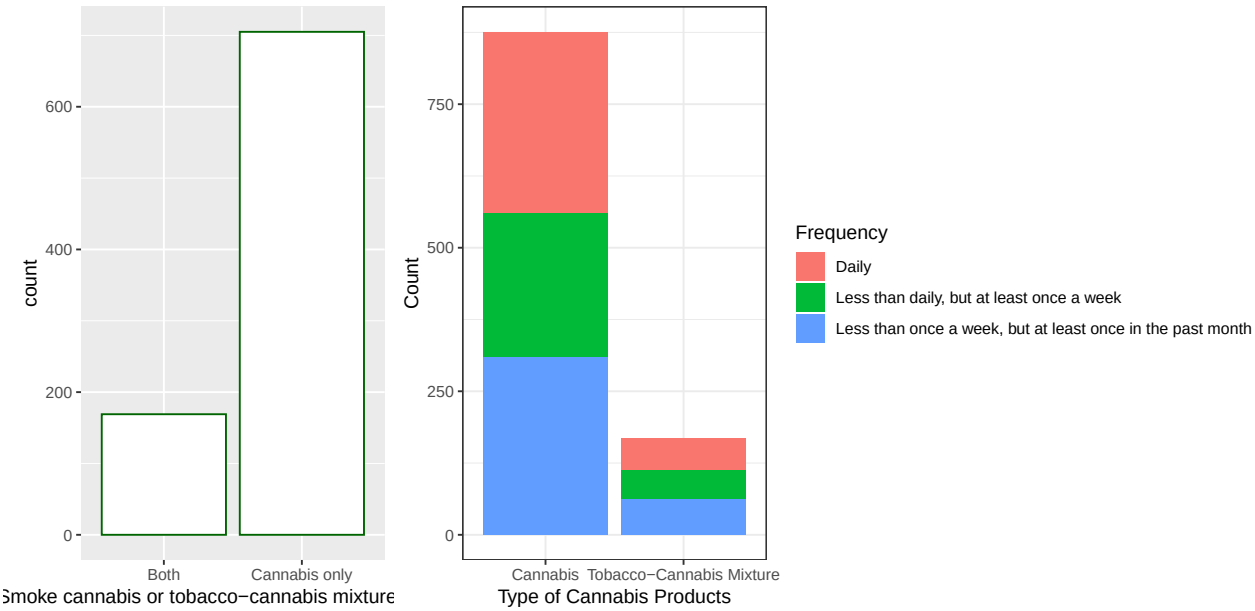

Figure 10: Proportion of smoker and non-smoker, and frequency of smoking for pure and mixed cannabis

6 Discussion

Based on our choice of model, we want to investigate what factors are most influencial on the respondent's usage of each product. The model for analysis looks like:

$$log(Y) = \beta_0 + \beta_1 * gender + \beta_2 * age + \beta_3 * province + \beta_4 * age_1 + \beta_5 * age_2 + \beta_6 * firstproduct$$

This model has six predictors in total, which are gender, age, province, ages when respondent first try other two products, and the first product respondent tried. Some factors could be more significant than others. To begin with, I created table 5, which shows the correlation among variables. The rightest three columns, also most bottom three rows, `smoke_cigarette`, `vape` and `smoke_cannabis`, are our binary response variables; and all other variables are predictors. From table 5, we may see that first product has a fairly strong relation to every response variables. Three variables of agea of first try of each product is less significant, but still have some significance on the respnose. Respondent's age also has some significance, though it does not have strong relation to any of the responses. Gender and province do not seem to have any significance or relation with the response, so we can hypothesize that these two are not important in the model as well.

Table 5: Correlation among variables

	gender	age	province	age_first_time_cigarettes	age_first_time_vape	age_first_time_cannabis	first_product	smoke_cigarette	vape	smoke_cannabis
gender	1.000	0.053	0.008	0.073	0.082	0.067	0.070	-0.054	-0.058	-0.096
age	0.053	1.000	-0.021	-0.352	0.435	0.091	-0.069	0.077	-0.214	-0.161
province	0.008	-0.021	1.000	0.039	0.016	0.043	0.036	0.001	-0.002	-0.013
age_first_time_cigarettes	0.073	-0.352	0.039	1.000	0.103	0.392	0.624	-0.353	-0.136	-0.203
age_first_time_vape	0.082	0.435	0.016	0.103	1.000	0.377	0.319	-0.148	-0.503	-0.386
age_first_time_cannabis	0.067	0.091	0.043	0.392	0.377	1.000	0.698	-0.185	-0.231	-0.445
first_product	0.070	-0.069	0.036	0.624	0.319	0.698	1.000	-0.292	-0.204	-0.311
smoke_cigarette	-0.054	0.077	0.001	-0.353	-0.148	-0.185	-0.292	1.000	0.160	0.219
vape	-0.058	-0.214	-0.002	-0.136	-0.503	-0.231	-0.204	0.160	1.000	0.298
smoke_cannabis	-0.096	-0.161	-0.013	-0.203	-0.386	-0.445	-0.311	0.219	0.298	1.000

The summary of final model with response variable being `smoke_cigarette` is displayed in table 6 below. As shown in the table, taking 0.05 as the threshold, `gender`, `age`, `age_first_time_vape` and `first_product`

are statistically significant. They have very small p-values, indicating that we have strong evidence to reject the null hypothesis that this variable has no effect on observed outcome. According to this model, for each 1 unit increase in age (which means changing to an older age group as this is a categorical variable), the probability that respondent is a smoker will increase by 0.332. Similarly, each 1 unit decrease in age that respondent first tried vaping - that is, if respondent started vaping earlier - will lead to an increase of 0.163 in probability of being a smoker. Gender and first product have a negative relationship with the response because of the dummy variable assigned to each of them. Because male is assigned 0 and female is assigned 1, the coefficient -0.228 of gender actually indicates that males are more likely to be a smoker than females do. Also, variable first product is arranged in an order to "Cigarette", "An e-cigarette or vaping device", "Cannabis" ~ 3, "Valid skip", and "Not stated", where "Cigarette" is assigned to 1 and "Not stated" is assigned to 5. The coefficient -0.562 indicates a negative relationship between first product and response. Therefore, if the first product that respondent used is cigarette, then he or she is most likely to become a cigarette smoker. If this person started with vaping or cannabis, he or she is less likely to be a cigarette smoker. And if the answer is a valid skip or not stated, meaning that we are uncertain of the first product that respondent used, or the respondent has not used any of these product, then certainly this person is very unlikely to be a smoker of cigarette now.

Table 6: Summary of Model for Cigarette

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.459	0.149	-3.090	0.002
gender	-0.228	0.080	-2.857	0.004
age	0.332	0.031	10.624	0.000
province	0.019	0.014	1.305	0.192
age_first_time_vape	-0.163	0.014	-11.455	0.000
age_first_time_cannabis	0.013	0.015	0.864	0.388
first_product	-0.562	0.040	-13.980	0.000

Similarly, table 7 and 8 display the summary of model when the response is vaping and cannabis, respectively. In the case of vaping, predictors `age`, `age_first_time_cigarettes`, `age_first_time_cannabis`, and `first_product` are statistically significant; while in the case of cannabis, predictors `gender`, `age`, `age_first_time_cigarettes`, `age_first_time_vape`, and `first_product` are statistically significant, according to their p-values.

Compare the results of model for all three products, gender has some statistical importance, but is not always significant. The coefficient of gender is negative in all three models, which means that males are more likely to use all three tobacco and cannabis product than females do. Age of respondent is very significant according to all three models. In case of cigarette, if the respondent is older, he or she is more likely to be a smoker. But in case of vaping and cannabis, younger people are more likely to use these products. This result is corresponding to Statistic Canada's findings that the prevalence of current cigarette smoking among young adults aged 15 to 19 was 3% (63,000) and aged 20 to 24 was 8% (201,000), but the prevalence of vaping among youth aged 15 to 19 and 20 to 24 were 35% (711,000) and 43% (1.0 million) respectively, and the prevalence of cannabis among the same age groups were 23% (471,000) and 52% (1.2 million) respectively (G. of Canada (2022a)). Cigarette is more attractive to people in elder age groups while vaping and cannabis are preferred by young people.

Province or location does not seem to have any statistical importance in any of these models. The age when respondent started using any of cigarette, vaping products or cannabis, has significant effect on the response. Even the age when respondent tried vaping, for example, seems to have no direct casual relation to whether this person smokes cigarette currently, these two variables are closely connected. This is corresponding to Berry KM's previous study that "the use of e-cigarette and other non-cigarette tobacco products may increase the odds of cigarettes initiation and use, particularly among low-risk youths" (Berry KM (2019)). Besides, a study of Allison N. Kristman-Valente discovered a reciprocal relationship between conventional

cigarette smoking and marijuana use during adolescence, that increased cigarette smoking predicted increased marijuana use and vice versa (Kristman-Valente (2017)).

Lastly, the first product that the respondent tried is very significant to his or her current status. Intuitively, there is a relation between type of first product and the respondent's choice of whether to become a smoker. For example, if the respondent has smoked cigarette before, he or she is more likely to continue smoking than those who did not smoke or tried other products.

Table 7: Summary of Model for Vaping

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.749	0.225	12.209	0.000
gender	-0.245	0.106	-2.302	0.021
age	-0.784	0.039	-20.041	0.000
province	0.010	0.019	0.527	0.598
age_first_time_cigarettes	-0.240	0.026	-9.323	0.000
age_first_time_cannabis	-0.130	0.019	-7.040	0.000
first_product	-0.283	0.063	-4.492	0.000

Table 8: Summary of Model for Cannabis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.482	0.168	14.739	0.000
gender	-0.419	0.082	-5.112	0.000
age	-0.206	0.030	-6.954	0.000
province	-0.003	0.014	-0.184	0.854
age_first_time_cigarettes	-0.143	0.019	-7.382	0.000
age_first_time_vape	-0.148	0.011	-13.493	0.000
first_product	-0.450	0.041	-10.987	0.000

6.1 Weaknesses and next steps

Survey errors come from a variety of different sources. One dimension of survey error is sampling error. Sampling error is defined as the error that arises because an estimate is based on a sample rather than the entire population (Alcohol and (CADS) (2021)). Because it is nearly impossible to collect data of the entire population of Canada, the CTNS survey selects a sample of 8112 individuals from 10 provinces of Canada. However, as we mentioned in the data session, figure 1 showed that a majority of respondents are from either Ontario or Quebec, which means the sampling would have some bias. A survey as large as the CTNS is bounded to have sampling error. Sampling error could be expressed in a form of confidence interval (Alcohol and (CADS) (2021)). According to data published by Statistic Canada, the approximate sampling error estimate of 95% confidence intervals for a proportion of 50% (Canada level) is 4.0%, and the estimate of 95% confidence intervals for a proportion of 10% (Canada level) is 2.5% (Alcohol and (CADS) (2021)).

Non-response bias and self-report bias may arise as well. The response rate for the Canadian Tobacco and Nicotine Survey was 41% (Alcohol and (CADS) (2021)). Although the survey estimates are adjusted to account for non-response through the survey weights, the non-responding households and persons may still lead to a bias in results. Self-report bias is a methodological problem that arises when researchers rely on asking people to describe their behaviours rather than measuring directly. People may not have given fully correct answers, perhaps because talking about their usage of addictive products may be uncomfortable for

some people, given the society's perception of harm of these products. Since the questionnaire provides fixed responses to participants, some respondent may select options which are not their truly condition.

Furthermore, the data collection may encounter coverage errors. Coverage errors arise when there are differences between the target population and the observed population (Alcohol and (CADS) (2021)). The targe population is any person in Canada who is aged 15 or above, while the observed population is persons living in dwellings with mailable addresses on the frame. Approximately, 95% of the dwellings in Canada had mailable addresses. To the extent that the excluded population differs from the rest of the target population, the results may be biased (Alcohol and (CADS) (2021)).

Appendix

Extract of the questions from Gebru et al. (2021)

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 Canada Tobacco and Nicotine Survey (CTNS) conducted by Statistics Canada in 2020. Statistics Canada conducted CNTS on the behalf of Health Canada in order to fill the important information gap of vaping, cannabis, and tobacco usage in Canada.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Statistic Canada with consultation of Health Canada.
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.*
    - We took the dataset from the CTNS website where the final report is published. No funding was needed on our part.
4. *Any other comments?*
    - No, 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.*
    - The instances that comprise the dataset represent the prevalence of usage of cigarettes, vaping products and cannabis among Canadian aged 15 and older. All of the instances have to do with people and their personal information.
2. *How many instances are there in total (of each type, if appropriate)?*
    - 29 instances.
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 dataset has information on the prevalence of tobacco and nicotine usage among Canaidans aged 15 and older, and all of that information was important to our report, therefore the dataset contains all possible instances.
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 consists of the usages of cigarette, vaping product or cannabis, as well as some basic information of the respondent.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - There is no target associated with each instance.

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 no information missing from individual instances.
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.*
    - Relationships between individual instances is made explicit by graphing. The graphs highlight trends and differences between individual instances.
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, there is no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - There are no errors, sources of noise, or redundancies in the 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.*
    - This dataset is archived by the Statistic Canada, so since it is a government institution, there are guarantees that it will exist over time.
    - There are official archival versions of the complete dataset.
    - There are no licenses or fees that would restrict someone's access to any of the external resources.
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.*
    - Since the data is anonymous, it is not considered confidential.
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.*
    - The data is not offensive, insulting, threatening, and does not cause anxiety.
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.*
    - No.
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.*
    - It is impossible to identify individuals directly or indirectly from the dataset because all of the information that was collected was anonymous.
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 data that might be considered offensive in any way.

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.*
    - The data set can be directly observed or gathered from the Statistic Canada official website.
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?*
    - Downloading the csv file from website.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - This data set is not a sample from a larger population.
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)?*
    - Until 2017, Statistics Canada conducted the Canadian Tobacco, Alcohol and Drugs Survey (CTADS), which collected data on tobacco as well as alcohol and drug use in Canada. In 2019, the Canadian Alcohol and Drugs Survey (CADS) was conducted to collect data on alcohol and drug use independently from the Canadian Tobacco and Nicotine Survey (CTNS) which was conducted to primarily collect data on tobacco and nicotine.
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 is collected from 2020-12-08 to 2021-01-16.
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)?*
    - Data is collected from individuals through electronic questionnaires and telephone interviews.
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.
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.*
    - No.

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).*
    - Consent was not obtained.
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.
12. *Any other comments?*
    - 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 is cleaned, and some labels are changed.
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://www150.statcan.gc.ca/n1/pub/13-25-0001/132500012021001-eng.htm
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 was usd.
4. *Any other comments?*
    - No.

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?*
    - Further research on tobacco and nicotine usages.
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 data set is only used for STA304 final paper, which is this task.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
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.*
    - 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 dataset will not be distributed.
3. *When will the dataset be distributed?*
    - The dataset will not 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.*
    - No.
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.
7. *Any other comments?*
    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will not be maintained.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - N/A
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)?*

- The dataset 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.*

   - N/A

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.*

   - N/A

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.*

   - N/A

8. *Any other comments?*

   - No.

References

Alcohol, Canadian, and Drugs Survey (CADS). 2021. *Canadian Tobacco and Nicotine Survey (Ctns)*. https://doi.org/https://doi.org/10.25318/132500012021001-eng.

Berry KM, Benjamin EJ, Fetterman JL. 2019. "Association of Electronic Cigarette Use with Subsequent Initiation of Tobacco Cigarettes in Us Youths." *JAMA Netw Open* 2 (2): 7794. https://doi.org/10.1001/jamanetworkopen.2018.7794.

Canada, Health. 2022. *Health Effects of Cannabis*. https://www.canada.ca/en/health-canada/services/drugs-medication/cannabis/health-effects/effects.html.

Canada, Government of. 2018. *Tobacco and Vaping Products Act*. https://www.canada.ca/en/health-canada/services/health-concerns/tobacco/legislation/federal-laws/tobacco-act.html.

———. 2022a. *Canadian Tobacco and Nicotine Survey (Ctns): Summary of Results for 2020*. https://www.canada.ca/en/health-canada/services/canadian-tobacco-nicotine-survey/2020-summary.html.

———. 2022b. *Preventing Kids and Teens from Vaping*. https://www.canada.ca/en/health-canada/services/smoking-tobacco/preventing/vaping.html.

Catherine M Cahill, Belinda Ivanovski, Gin S Malhi. 2006. "Cognitive Compromise in Bipolar Disorder with Chronic Cannabis Use: Cause or Consequence?" *Expert Review of Neurotherapeutics* 6 (4): 591–98. https://doi.org/10.1586/14737175.6.4.591.

Douglas Bates ORCID iD [aut], Ben Bolker ORCID iD [aut, Martin Maechler ORCID iD [aut]. 2022. *Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Hadley Wickham [aut, RStudio [cph, cre]. 2021. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Hao Zhu ORCID iD [aut, Thomas Travison [ctb], cre]. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Health, U. S. Department of, and Human Services. 2012. *Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General*. Atlanta: U.S. Department of Health; Human Services, Centers for Disease Control; Prevention, National Center for Chronic Disease Prevention; Health Promotion, Office on Smoking; Health. https://www.cdc.gov/tobacco/data_statistics/sgr/2012/index.htm.

Kristman-Valente, Hill, A. N. 2017. "The Relationship Between Marijuana and Conventional Cigarette Smoking Behavior from Early Adolescence to Adulthood." *Prevention Science : The Official Journal of the Society for Prevention Research* 18 (4): 428–38. https://doi.org/https://doi.org/10.1007/s11121-017-0774-4.

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

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

Robert, Melamede. 2005. "Cannabis and Tobacco Smoke Are Not Equally Carcinogenic." *Harm Reduct Journal 2*, no. 21. https://doi.org/https://doi.org/10.1186/1477-7517-2-21.

Rotermann, Michelle. 2020. "What Has Changed Since Cannabis Was Legalized?" *Statistics Canada, Catalogue No. 82-003-X - Health Report* 31 (2): 11–20. https://doi.org/https://www.doi.org/10.25318/82-003-x202000200002-eng.

Sam Firke [aut, Bill Denney [ctb], cre]. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://cran.r-project.org/package=janitor.

Torsten Hothorn ORCID iD [aut], cre], Achim Zeileis ORCID iD [aut. 2022. *Lmtest: Testing Linear Regression Models*. https://CRAN.R-project.org/package=lmtest.

Wickham, Hadley. 2020. *Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. https://cran.r-project.org/package=reshape2.

Yihui Xie ORCID iD [aut, Abhraneel Sarma [ctb], cre]. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.