Analyzing the Gateway to Vaping

Andy Vu - 1005244932

December 19, 2021

Abstract

Vaping is a relatively new product that has become increasingly popular throughout the last few years. Even though vaping is marketed towards smokers, there has been an influx in teen vaping exposing youth to nicotine, a harmful and addictive drug. This analysis aims to investigate if smoking is the biggest factor for one to try vaping. The Canadian Tobacco and Nicotine Survey has been able to capture data related to vaping, smoking, and other government-regulated drugs. Variables such as gender, age, province, and smoking, vaping, and cannabis usage, as well as which was used first have been taken from the survey data for analysis. Propensity Score Matching was used to designate a treatment and control group based on whether or not the respondent has tried smoking. A logistics model was then created from the resulting subset of the data with trying vaping for the first time as the response variable and tried smoking as one of the explanatory variables. The model showed that the biggest factor that affected one trying vaping for the first time was age. Predictably, those who are younger had a much higher probability of trying vaping for the first time than those who are older. Youth should not be trying vaping especially since it can contain nicotine. Ultimately, it is a problem that a product marketed to smokers is being tried more commonly amongst youths.

Keywords: Canadian Tobacco and Nicotine Survey, Teen Vaping, Smoking and Cannabis, Causal Inference, Logistic Regression, Propensity Score Matching, Odds and Probability.

Introduction

Smoking tobacco is probably one of the most heavily researched topics of the 21st century. There have been extensive studies on smoking, connecting it to various adverse health effects (Canada, 2016). Moreover, not only is smoking bad for one's health, but it is also highly addictive (Canada, 2013). The main component that makes smoking tobacco addictive is the drug known as nicotine (Canada, 2013). Vaping claims to give the same feeling like smoking tobacco, but without the negative effects (Canada, 2020). Thus, it seems like many smokers have turned to vaping as an alternative (Canada, 2020). However, the underlying problem is that not all those who vape were previously smokers (Canada, 2021). Therefore, vaping itself has spawned its own wave of users (Canada, 2021).

Generally, many factors influence a person to try a drug such as nicotine. With vaping being the new "cool" thing and its accessibility, many teenagers are trying it and getting hooked (Canada, 2021). A major concern with vaping is that nicotine may become a gateway drug to other more extreme drugs (Canada, 2021). That being said, the goal of this analysis is to **determine the effect that smoking tobacco has on the probability of a person trying vaping for the first time.** It is also important to understand which factors affect the probability of one vaping for the first time if it is not because of smoking. With the given facts, the hypothesis is that smoking will have some effect on the probability of a person trying vaping for the first time, but it will not be the biggest factor. Knowing that teens can easily be socially influenced, it would not be a surprise if they have a high probability of trying vaping for the first time without having previously smoked tobacco (Canada, 2021).

The data that will be used in this analysis is from the Canadian Tobacco and Nicotine Survey (CTNS) (Government of Canada, 2020). This survey collects data relating to vaping, cannabis, and tobacco usage throughout Canada (Government of Canada, 2020).

Data

Data Collection Process

The data from the Canadian Tobacco and Nicotine Survey (CTNS) was collected between December 2020 to January 2021 (Government of Canada, 2020). The survey targeted people 15 years of age or older residing in any of the ten provinces in Canada (Government of Canada, 2020). A stratified sampling and cross-sectional design were used to survey the population electronically (Government of Canada, 2020). All the responses from the survey are voluntary and collected directly from the respondent (Government of Canada, 2020). The overall response rate was about 41% of the approximately 20,000 surveys that were sent out (Government of Canada, 2020).

Data Summary

The dataset from the Canadian Tobacco and Nicotine Survey includes all the necessary variables such as whether or not the respondent had tried vaping, smoking, or using any other regulated drugs such as cannabis or alcohol in their lifetime (Government of Canada, 2020). Besides this, the dataset also provides an age group of the respondent, as well as their gender and province of residence (Government of Canada, 2020). A key aspect of this survey is that there is a specific question that asks the respondent which of the 3 possible products they have tried first. This variable is important to the model that will be used later in the analysis. Since the variable of interest is the probability that a person would try vaping for the first time, the model's estimate should be nearly 100%, if not 100% if the respondent had responded with vaping as their first product tried. Overall, the goal of this survey is to provide a snapshot of cigarette smoking, vaping, and cannabis use throughout Canada (Government of Canada, 2020). The sample size of the dataset is 8112 with 99 variables.

Data Cleaning

Of the 99 possible variables, there will only be 7 specific variables that will be taken from the dataset to be used in this analysis. The variables are the respondents' gender, age group, province of residence, whether or not they have tried smoking, whether or not they have tried vaping, whether or not they have tried cannabis and which of the three they had tried first.

Of the 99 possible variables, there will only be 7 specific variables that will be taken from the dataset to be used in this analysis. The variables are the respondents' gender, age group, province of residence, whether or not they have tried smoking, whether or not they have tried vaping, whether or not they have tried cannabis and which of the three they had tried first.

The first responses to map are the ones for the Gender variable. In the case of the survey, there were only two possible responses, either Male or Female. The following responses to map are the ones corresponding to the age groups. There are a total of 7 age groups: 15-19 years old, 20-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, and 65 years old and older. The next responses are the ones for the province variable. In this case, there are the standard 10 provinces of Canada: Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Colombia. The next three responses to map are the ones corresponding to whether or not the respondent has tried smoking, vaping, or cannabis. These responses are mapped to 1 if they have tried the corresponding product and 0 if they have not tried the corresponding product. The last set of responses to map is the one for the first product that the respondent has tried. The possible responses are Smoking, Vaping, Cannabis, or None. After mapping each response, any observations with invalid responses will be omitted from the analysis to ensure a complete dataset.

The most important step now is to create a new variable for whether or not the first product that the respondent tried is Vaping. This variable will be created based on the response from the variable that asks which of the products the respondent has tried first. This variable will be either 1 if Vaping was the first product tried and 0 otherwise. It is important to create this variable instead of using the variable for whether the respondent has tried vaping in their lifetime because the model should be able to determine the probability of trying vaping for the first time of those who have never tried any of the products before. For example, one respondent may have tried vaping before so it would not the first product they had tried, whereas for another respondent vaping would be their first product if they have has never tried any of the three products before.

Finally, 8 observations responded with having tried Cannabis as their first product but previously responded to never having tried Cannabis before. These are invalid observations and are hence omitted.

Description of Varibles

Table 1: Description of Important Variables

Variable	Description
Gender	The gender of the respondent. (Male or Female)
Age Group	The age group of the respondent ranging from 15 to 65 years of age and older.
Province	The province in which the respondent resides.
Smoked	Whether or not the respondent had tried smoking in their lifetime. (1 if they have, 0 if not)
Vaped	Whether or not the respondent had tried vaping in their lifetime. (1 if they have, 0 if not)
Cannabis	Whether or not the respondent had tried cannabis in their lifetime. (1 if they have, 0 if not)
Tried First	The product that the respondent has tried first. (Smoking, Vaping, Cannabis, None)
Vaping First	Whether or not the respondent had tried Vaping first. (1 if they have, 0 if not)

Numerical Summaries and Plots

As seen in Table 2, the majority of the respondents have tried either smoking, vaping, or cannabis, with 3020 responding to having tried smoking as their first product. This is followed by 1264 respondents who have tried cannabis as their first product and only 484 who have tried vaping as their first product. On the other hand, 3237 respondents have never tried any of these products before.

Following up with Table 2, Table 3 shows the distribution of the variable of interest. Table 3 shows that 484 respondents tried vaping as their first product and 7521 respondents have tried other products first or no products at all.

Table 2: Summary of the First Product Tried

Product	Frequency
Smoking	3020
Vaping	484
Cannabis	1264
None	3237

Table 3: Summary of Vaping being the First Product

Vaping First	Frequency
False	7521
True	484

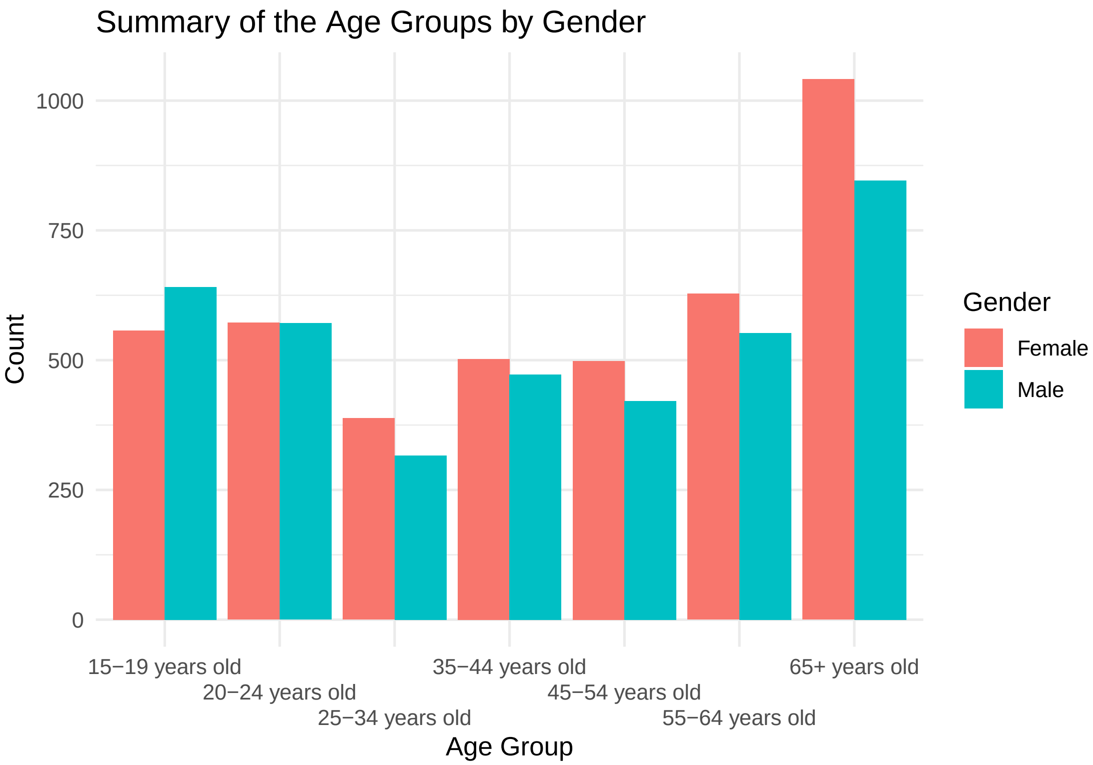

Figure 1: Summary of the Age Groups by Gender

As seen in Figure 1, the distribution of Gender seems to be about the same across all age groups. This means that there are as many respondents who identify as the Male gender as there are those who identify as the Female gender. The number of respondents in each age group also seems to be equally distributed. Although there is a significantly greater portion of the sample who is part of the 65 years of age and older group, this can easily be explained by the age range that is covered. 65 years of age and older covers a larger range of ages in comparison to the 5-year or 10-year ranges that the other age groups cover.

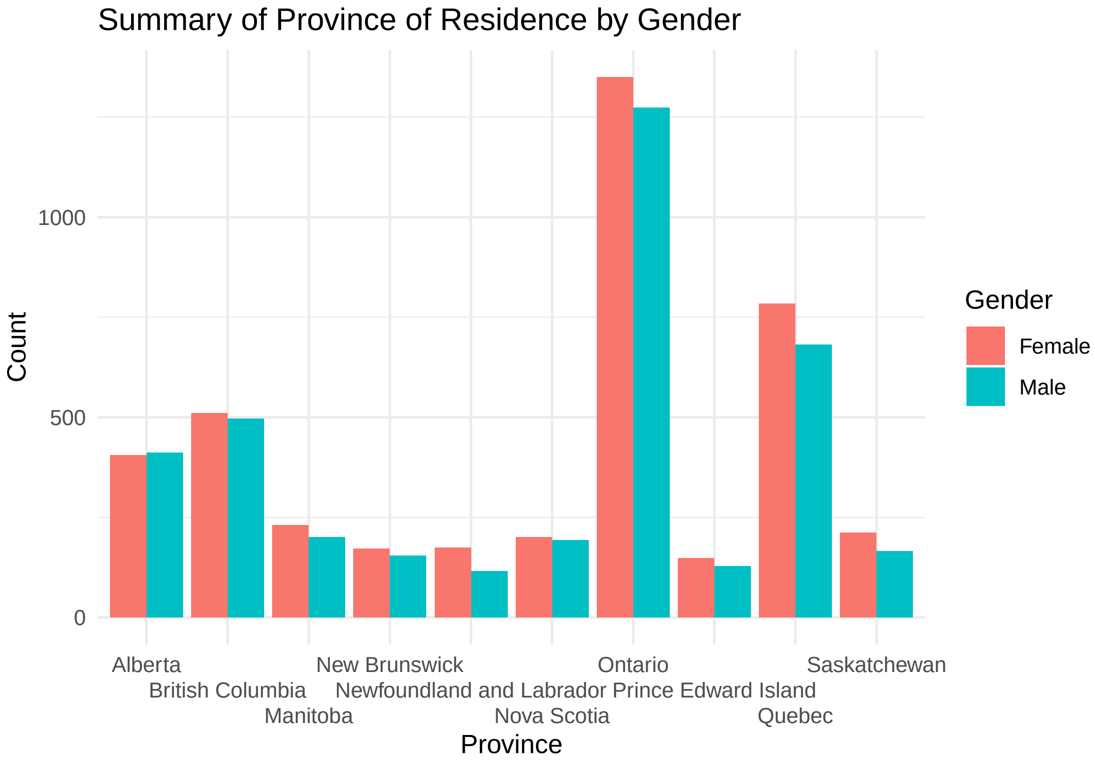

Figure 2: Summary of Province of Residence by Gender

In Figure 2, the Gender distribution is as expected with each gender having about the same distribution in each province. The only significant observation is that most of the respondents in the sample reside in Ontario, followed by Quebec. This may imply that the survey is heavily weighted by the responses of those living in Ontario and Quebec. However, it is important to note that the majority of the population of Canada resides in Ontario, followed by Quebec (Government of Canada, 2018). Thus, it follows that the sample used for this survey does in fact give a good representation for the population of Canada as expected.

Figure 3 depicts the products tried by the respondent by the first product they have tried. Some key observations are that there is a larger proportion of the respondents who also tried cannabis and responded to having tried smoking as their first product. Similarly, there is a large proportion of respondents who tried cannabis and responded to having tried vaping as their first product. On the other hand, for those who responded to having tried cannabis as their first product, there is about an equal proportion who also tried smoking and/or vaping afterward. As for the respondent who tried none of the products, there is nothing notable here.

All analysis for this report was programmed using `R version 4.0.5`.

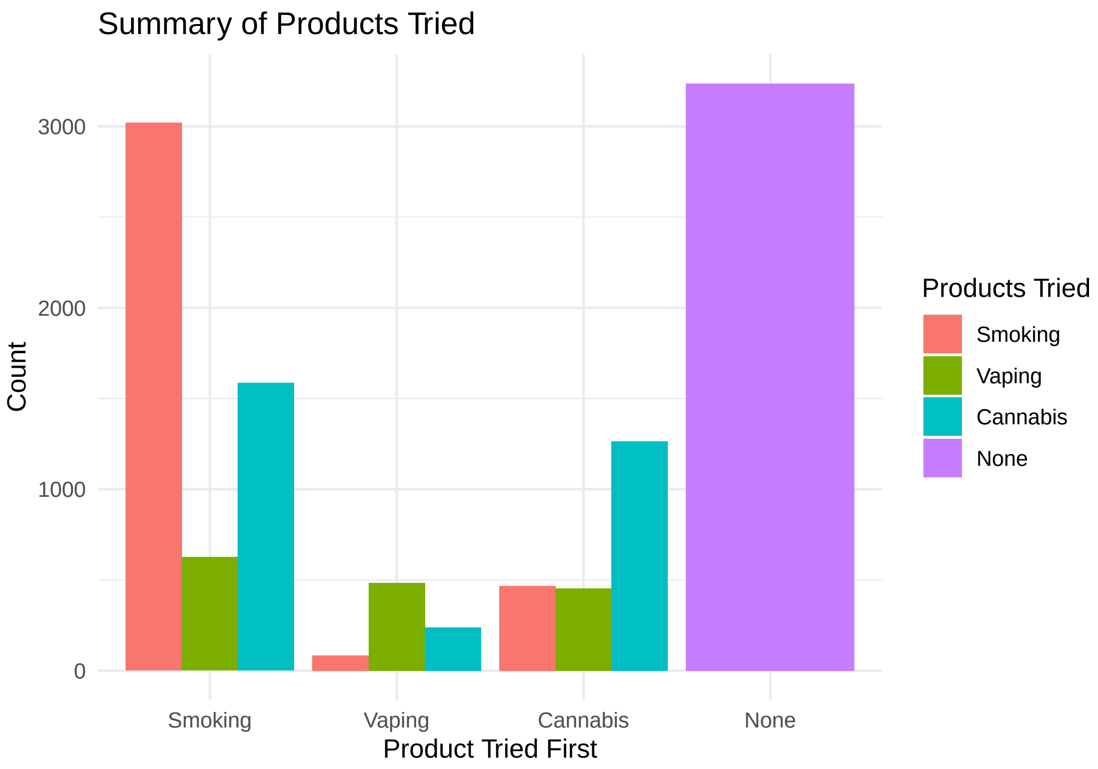

Figure 3: Summary of Products Tried

Methods

Since the data that is used in this analysis is based on a survey, it is classified as observational data. The goal of this analysis is to possibly find a causal relationship between trying vaping for the first time and having previously tried smoking. However, the problem is that there is no clear indication of treatment or control groups. For this reason, the main method that will be executed in this analysis is Propensity Score Matching.

The first step in Propensity Score Matching is to construct a logistic regression model to assign a probability/propensity score to each observation. In the case of this analysis, the response for the model will be whether or not the respondent has tried smoking in their lifetime. Similarly, since the overall variable of interest is whether or not the respondent will try vaping as their first product, the corresponding variable will be excluded from this model that is being used for assigning the propensity scores. Thus, the model that will be used to assign the propensity scores will have all of the remaining variables as explanatory variables. In addition, using this model assumes all the observations are independent and identically distributed.

After constructing the logistic regression model, a vector of probability estimates from the model will be appended to the data. This is so that each observation has a propensity score for whether or not the respondent has tried smoking for the first time, regardless of if they did or not. The next step is to now compare the observations and match observations of similar propensity scores, but different outcomes. That is two respondents with the same propensity score, but with one who has indeed tried smoking for the first time, and one who has not. A nearest neighbour matching approach will be used, which can easily be implemented using the matching function from the `arm` package in `R` (Gelman & Hill, 2006).

After matching, the resulting dataset will be reduced. This is due to the possibility that there are observations that had no matches. Regardless, the new dataset will then be used to construct a new logistic regression model with the response variable being whether or not the respondent has tried vaping for the first time.

Model selection will then be implemented starting with a full model of all the explanatory variables, where the variable for whether or not the respondent has tried smoking for the first time is included here. A simple method of removing variables with non-significant p-values, i.e., p-values greater or equal to 0.05, will be used to select the explanatory variables and determine the final model.

Model

In both cases, the logistic regression model will be of the form,

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k,$$

where $\hat{y}$ is the estimated log-odds, each β_i represents the average change in log-odds corresponding to the ith factor level and x_i is only either 1, if the corresponding ith factor level is present, and 0 if not, since all of the explanatory variables are categorical.

Results

After executing the Propensity Score Matching method, the resulting matched dataset shrunk to 7140 observations from the 8005 observations before matching. As seen in Table 7 in the Appendix, there is the same number of observations for those who had tried smoking in their lifetime and those who had not. This dataset is then used to create a logistic regression model where the variable for whether or not the respondent tried Vaping as their first product is the response variable and the remaining variables are the explanatory variables.

The first model as seen in Table 8 in the Appendix is the full model. It seems like the variable for the first product that the respondent tried with Vaping as one of its factor levels carried all of the weight. This left the coefficients for all of the other factor levels to essentially be 0 and the algorithm to not converge. This should have been expected since the response variable was created from this variable. In either case, it is the first variable to be removed from the final model.

With the new model, it is evident from Table 8 in the Appendix, that the province factor levels have insignificant p-values. Hence, the variable for the province is removed from the final model. Since the model is to estimate the probability of a person trying Vaping for the first time, it also makes sense to remove the explanatory variable that corresponds to whether or not the respondent has vaped for the first time. Thus, the resulting model is the one that has the explanatory variables for gender, age group, whether or not the respondent has tried smoking, and whether or not the respondent has tried cannabis. It is also important to note here that the matched variable, that is whether or not the respondent has tried smoking in their lifetime, is significant from its low p-value.

A summary of the final model can be seen in Table 4, where all of the estimates are in terms of log-odds. Then, to better interpret the results, the log-odds estimates have been converted to probabilities in Table 5 and odd ratios in Table 6, by the formula below. That is, since the model is of the form,

$$\begin{aligned}
\hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \\
\log\widehat{(\frac{p}{1-p})} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \\
\widehat{\frac{p}{1-p}} &= \exp{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} \\
\hat{p} &= \frac{\exp{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}}{1 + \exp{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}}
\end{aligned}$$

Table 4: Summary of Final Model

	(1)		
(Intercept)	0.660	***	(0.142)
genderMale	-0.566	***	(0.128)
age20-24 years old	-1.560	***	(0.125)
age25-34 years old	-3.781	***	(0.258)
age35-44 years old	-6.292	***	(0.715)
age45-54 years old	-6.922	***	(1.006)
age55-64 years old	-7.091	***	(1.006)
age65+ years old	-20.589		(402.425)
factor(smoked)1	-0.710	***	(0.153)
factor(cannabis)1	-0.130		(0.134)
N	7140		
logLik	-910.281		
AIC	1840.563		

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Therefore, since the coefficients in the model are multiplicative, it is easier to interpret the reference group with the estimated probability and odds ratio for each explanatory variable. Then, the probability of trying Vaping for the first time in the default group with the gender being Female, the age group being 15-19 years old, never tried smoking and cannabis for the first time is estimated to be approximately 66%.

If the respondent identifies as Male and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.57 times the odds of those who identify with the Female gender. This means that by holding all other values constant, the odds of the respondent trying vaping for the first time of those who identify with the Female gender is $\frac{1}{0.57} \approx 1.75$ times that of those who identify as Male.

If the respondent is part of the 20-24 years old age group and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.21 times the odds of those who are part of the 15-19 years old age group. This means that by holding all other values constant, the odds of the respondent trying vaping for the first time of those who are part of the 15-19 years old age group is $\frac{1}{0.21} \approx 4.76$ times that of those who are part of the 20-24 years old age group.

Similarly, if the respondent is part of the 25-34 years old age group and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.02 times the odds of those who are part of the 15-19 years old age group. This means that by holding all other values constant, the odds of the respondent trying vaping for the first time of those who are part of the 15-19 years old age group is $\frac{1}{0.02} \approx 50$ times that of those who are part of the 25-34 years old age group.

Table 5: Summary of Final Model Probability Estimates With Respect to the Default Group

	Estimated Probability
Default Group *	0.6592
Male	0.3621
20-24 years old	0.17364
25-34 years old	0.0223
35-44 years old	0.00185
45-54 years old	0.00098
55-64 years old	0.00083
65 years old and older	0
Tired Smoking	0.32951
Tried Cannabis	0.46749

* Default Group: Female, 15-19 years old, Never Tried Smoking, and Never Tried Cannabis

Table 6: Summary of Final Model Odds Ratio Estimates With Respect to the Default Group

	Estimated Odds Ratio
Default Group *	1.93425
Male	0.56765
20-24 years old	0.21013
25-34 years old	0.0228
35-44 years old	0.00185
45-54 years old	0.00099
55-64 years old	0.00083
65 years old and older	0
Tired Smoking	0.49146
Tried Cannabis	0.87792

* Default Group: Female, 15-19 years old, Never Tried Smoking, and Never Tried Cannabis

Then, if the respondent is part of the 35-44 years old age group, 45-54 years old age group, 55-64 years old age group, or 65 years old and older age group and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.002, 0.001, 0.0008 and 0 times the odds of those who are part of the 15-19 years old age group, respectively.

If the respondent has tried smoking in their lifetime and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.33 times the odds of those who have never tried smoking in their lifetime. This means that by holding all other values constant, the odds of the respondent trying vaping for the first time of those who have never tried smoking in their lifetime is $\frac{1}{0.33} \approx 3.03$ times that of those who have tried smoking in their lifetime.

Similarly, if the respondent has tried cannabis in their lifetime and by holding all other values constant, the model estimates that the odds of the respondent trying vaping for the first time is on average 0.47 times the odds of those who have never tried cannabis in their lifetime. This means that by holding all other values constant, the odds of the respondent trying vaping for the first time of those who have never tried cannabis in their lifetime is $\frac{1}{0.47} \approx 2.12$ times that of those who have tried cannabis in their lifetime.

These results seem reasonable since teenagers are expected to have a higher probability to try vaping compared to older age groups (Canada, 2021). Surprisingly, the Female gender has a higher probability of trying vaping in comparison to the Male gender. As for the variables on whether or not the respondent has tried smoking or cannabis before, this result follows from what was expected. That is, the odds of the respondent trying vaping for the first time is higher if they had tried smoking before, compared to those who had tried cannabis before. This is probably because vaping is more commonly marketed to smokers than cannabis users (Canada, 2020).

Conclusions

Predicting the probability of trying vaping for the first time based on a survey is not as straightforward. Many aspects need to be taken into account given that it is observational data. Overall, the results were somewhat as predicted in the hypothesis. That is, having tried smoking before will affect the probability of a person trying vaping for the first time, but it is not the biggest factor. In short, the main method used in this analysis was Propensity Score Matching. This method was used to emphasize distinct treatment and control groups from observational data. In the case of this study, the observational data came from the Canadian Tobacco and Nicotine Survey, and the intended treatment and control groups were those who had tried smoking before and those who had not, respectively. In the end, this created a reduced dataset with an equal size of observations that are in the treatment group as there are in the control group. This new dataset was then used to create a Logistic Regression model to determine the probability of trying vaping for the first time where whether or not the respondent had previously tried smoking is one of the explanatory variables.

All in all, the model estimated that having previously tried smoking increased the estimated odd of trying vaping for the first time more than those who had only previously tried cannabis. Furthermore, it was also hypothesized that those who are of the teenage age group have a higher probability to try vaping for the first time compared to other age groups. The model confirmed this part of the hypothesis by comparing the 15-19 years old age group to the other age groups. For instance, the odds of 15-19 years old trying vaping for the first time is nearly 5 times that of 20-24 years old. These odds increased exponentially as the 15-19 years old age group were compared to the even older age groups. For this reason, it had become clear that age is the biggest factor in determining whether or not one would try vaping for the first time.

This key result was no surprise because teenagers are known to be susceptible to social influences. It was found that those who had friends who vaped were more likely to try vaping themselves (Canada, 2021). However, the most unexpected result from this analysis was that the odds of trying vaping for the first time of those who identify as the Female gender are nearly double that of those who identify as the Male gender. That is, those who identify as the Female gender are twice as likely to try vaping compared to those of the

Male gender. The only conclusion that can be drawn from this result is that those of the Female gender may be more susceptible to social influences.

Weaknesses and Limitations

There are many limitations to all analyses, but the major limitation of this analysis is the fact that the results may appear causal but are not. This is because the dataset used for the model is based on a survey, which is observational data. Given that the treatment in question is smoking, there is no way to get experimental data due to ethical concerns. Consequently, it is the main reason for using Propensity Score Matching as it is a popular way to form mock treatment and control groups from observational data.

A limitation of the smoking and cannabis usage explanatory variables in the model is that there is no difference between a respondent who tried the product once and one who uses it often. For example, a respondent who had one puff of smoke and then never smoked again and would be considered the same observation as a respondent who smokes daily.

Besides weaknesses of the overall analysis, there are also some weaknesses from using the Propensity Score Matching method as well. A notable weakness is that when the data is being matched, it is only being matched to the observations that are already in the dataset (Alexander, 2021). This means that it is only matching based on the variables and samples of the original dataset and nothing else. This brings up the main component that they are being matched by which is the propensity score determined by a model. It is no doubt that different models will produce different results (Alexander, 2021). Depending on the model used, there can be significantly more matches or no matches at all.

Next Steps

There are many ways that this analysis can be taken further. One for instance is that there are many other explanatory variables directly related to the response variable that can be used such as accessibility to vaping products, number of friends who vape, personal health concerns about vaping, etc. It is also possible to further classify respondents to the order in which they tried each product, i.e., smoking, cannabis, and/or vaping if they had tried more than one. A deeper analysis to consider is looking at the factors that contribute to a user who vapes regularly.

It would also be beneficial to take a deeper look into the effect caused by gender. A limitation of the dataset is that it assumed that gender is binary, prompting some to not respond to the question in the survey, hence rendering their observation obsolete. Thus, it is possible to redo the survey with a non-binary gender response to possibly get a different result than the one in this analysis.

Discussion

In essence. youth are found to be most prone to trying vaping, which is a call for concern. As a product that is relatively new to the market, there is a lack of research on its health effects compared to smoking. What is known is that vaping contains nicotine which is a harmful and addictive drug (Canada, 2021). This analysis has found that although vaping is marketed to smokers, the odds of a smoker trying vaping for the first time are lower than that of a youth. In other words, age is the biggest factor in affecting the change in the probability of one trying vaping for the first time. Given these points, this result is solely based on this analysis from observational data. To better understand any topic more research is needed to be done and more time is needed for longitudinal studies.

Bibliography

1. Alexander, R. (2021, August 24). *Telling Stories With Data. Telling Stories With Data.* Retrieved December 15, 2021, from https://www.tellingstorieswithdata.com/.

2. Andrew Gelman and Jennifer Hill. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

3. Canada, H. (2013, March 7). *Government of Canada.* Canada.ca. Retrieved December 10, 2021, from https://www.canada.ca/en/health-canada/services/smoking-tobacco/effects-smoking/smoking-your-body/nicotine-addiction.html.

4. Canada, H. (2016, May 17). *Government of Canada.* Canada.ca. Retrieved December 10, 2021, from https://www.canada.ca/en/health-canada/services/smoking-tobacco/effects-smoking/smoking-your-body/risks-smoking.html.

5. Canada, H. (2020, June 12). *Government of Canada.* Canada.ca. Retrieved December 10, 2021, from https://www.canada.ca/en/health-canada/services/smoking-tobacco/vaping/smokers.html.

6. Canada, H. (2021, February 5). *Government of Canada.* Canada.ca. Retrieved December 10, 2021, from https://www.canada.ca/en/health-canada/services/smoking-tobacco/preventing/vaping.html.

7. Government of Canada, S. C. (2018, February 7). *Population and dwelling count highlight tables, 2016 census.* – Canada, provinces and territories. Retrieved December 15, 2021, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&SR=1&S=3&O=D#tPopDwell

8. Government of Canada, S. C. (2020, December 4). *Canadian tobacco and Nicotine Survey (CTNS).* Surveys and statistical programs. Retrieved November 28, 2021, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1291039.

Appendix

A1: Ethics Statement

As an individual doing research and analysis in statistics, it is of utmost priority to exercise proper ethical practices. The dataset used in this analysis has been directly sourced from open data. This data is made publicly available and free to be used by anyone. The survey data has been ensured to be free of any private information that may disclose the identities of the respondents. Following the use of open data, all sourced information, including the data, has been properly referenced and cited in the bibliography. This guarantees that the research and work done by other individuals are given credit where credit is due. Lastly, the code for analysis has been made to be open source to provide transparency and allow for reproducibility of the results. This is important to avoid any bias or p-hacking that can be done to manipulate the findings in the analysis. It is recognized that upholding these aspects will encourage others to do the same to create a better statistics community.

A2: Materials

Here is a glimpse of the Canadian Tobacco and Nicotine Survey dataset:

```
## Rows: 8,112
## Columns: 99
## $ PUMFID   <dbl> 50000, 50001, 50002, 50003, 50004, 50005, 50006, 50007, 50008~
## $ HHLDSIZE <dbl> 3, 1, 1, 1, 3, 5, 2, 5, 5, 3, 3, 1, 4, 1, 3, 2, 4, 3, 4, 5, 1~
## $ GENDER   <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 2, 2~
## $ TBC_05AR <dbl> 1, 2, 2, 1, 2, 1, 1, 2, 2, 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1~
## $ TBC_05BR <dbl> 4, 96, 96, 3, 96, 3, 2, 96, 96, 96, 96, 3, 96, 3, 3, 96, 96, ~
## $ TBC_10AR <dbl> 4, 6, 6, 1, 6, 4, 1, 6, 6, 6, 6, 4, 6, 4, 2, 6, 6, 4, 6, 4, 4~
## $ TBC_10BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 1, 6, 6, 6, 6, 6, 6~
## $ TBC_15R  <dbl> 2, 6, 6, 1, 6, 2, 1, 6, 6, 6, 6, 1, 6, 1, 2, 6, 6, 2, 6, 2, 1~
## $ TBC_20R  <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 4, 6, 4, 6, 6, 6, 6, 6, 6, 4~
## $ TBC_25QR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 9~
## $ TBC_30AR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_30BR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_30CR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_30DR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_30ER <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_30FR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 1, 96, ~
## $ TBC_30GR <dbl> 96, 96, 96, 5, 96, 96, 5, 96, 96, 96, 96, 96, 96, 96, 0, 96, ~
## $ TBC_35R  <dbl> 6, 6, 6, 1, 6, 6, 1, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_40R  <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_41AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_41BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_41CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45ER <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45FR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_45GR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50DR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50ER <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ TBC_50GR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 9, 6, 6, 6, 6, 6, 6~
## $ OTP_05A  <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ OTP_05BR <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ OTP_05CR <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4~
## $ OTP_05DR <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ OTP_05ER <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4~
## $ VAP_05AR <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1~
## $ VAP_05BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 2, 96, 96, 96, 96, 3, 96,~
## $ VAP_10R  <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 3, 6, 6, 6, 6, 3, 6, 6, 4, 6, 4, 4~
## $ VAP_15AR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 1, 96, 96, 96, 96, 0, 96,~
## $ VAP_15BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 0, 96, 96, 96, 96, 0, 96,~
## $ VAP_15CR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 0, 96, 96, 96, 96, 0, 96,~
## $ VAP_20R  <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 3, 96, 96, 96, 96, 3, 96,~
## $ VAP_21R  <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 2, 96, 96, 96, 96, 3, 96,~
## $ VAP_30R  <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 2, 96, 96, 96, 96, 9, 96,~
## $ VAP_35R  <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 8, 96, 96, 96, 96, 3, 96,~
## $ VAP_40AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40DR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40ER <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 1, 6, 6, 6, 6, 6, 6~
## $ VAP_40FR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40GR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_40HR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 1, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41DR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41ER <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 1, 6, 6, 6, 6, 6, 6~
## $ VAP_41FR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41GR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_41HR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 1, 6, 6, 6, 6, 2, 6, 6, 6, 6, 6, 6~
## $ VAP_45R  <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 1, 6, 6, 6, 6, 1, 6, 6, 6, 6, 6, 6~
## $ VAP_60   <dbl> 2, 3, 2, 7, 3, 7, 3, 3, 3, 7, 5, 2, 4, 4, 2, 7, 4, 3, 7, 3, 7~
## $ CAN_05AR <dbl> 1, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 1, 1, 2, 1, 2, 1, 1~
## $ CAN_05BR <dbl> 4, 96, 96, 96, 96, 4, 4, 96, 96, 2, 96, 3, 96, 96, 2, 3, 96, ~
## $ CAN_10AR <dbl> 4, 6, 6, 6, 6, 4, 4, 6, 6, 2, 6, 1, 6, 6, 2, 1, 6, 4, 6, 4, 1~
## $ CAN_10BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 3, 96, 96, 96, 96, 3, 96,~
## $ CAN_15AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 4, 6, 6, 3, 4, 6, 6, 6, 6, 4~
## $ CAN_15BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 4, 96, 96, 96, 96, 1, 96,~
## $ CAN_20AR <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 1~
## $ CAN_20BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 2, 96, 96, 96, 96, 3, 96,~
## $ CAN_25AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 4, 6, 6, 6, 6, 4, 6, 6, 4, 6, 6, 4~
## $ CAN_25BR <dbl> 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 96, 9~
## $ CAN_30AR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30BR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30CR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30DR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 1, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30ER <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30FR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30GR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 1, 6, 6, 1, 6, 6, 2~
## $ CAN_30HR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 2~
## $ CAN_30IR <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 2, 6, 6, 6, 6, 2, 6, 6, 2, 6, 6, 1~
## $ IU_05R   <dbl> 3, 6, 6, 6, 6, 1, 1, 6, 6, 3, 6, 1, 6, 6, 3, 6, 6, 1, 6, 1, 1~
## $ ALC_05   <dbl> 2, 4, 3, 4, 4, 4, 1, 4, 3, 4, 4, 4, 4, 1, 4, 2, 2, 3, 4, 3, 3~
## $ ALC_10   <dbl> 4, 8, 7, 8, 8, 8, 5, 8, 7, 7, 8, 7, 8, 5, 8, 8, 5, 7, 8, 8, 7~
## $ AGEGROUP <dbl> 2, 5, 4, 6, 3, 3, 5, 2, 2, 1, 4, 6, 1, 7, 1, 4, 4, 3, 1, 1, 7~
## $ PROV_C   <dbl> 12, 59, 59, 13, 11, 12, 24, 59, 35, 35, 35, 24, 48, 35, 35, 4~
## $ DV_SSR   <dbl> 3, 3, 3, 1, 3, 3, 1, 3, 3, 3, 3, 2, 3, 2, 1, 3, 3, 3, 3, 3, 2~
## $ DV_CN30R <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 2, 2, 1~
## $ DV_VP30R <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2~
## $ DV_VC30R <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ DV_ALC30 <dbl> 1, 2, 1, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2, 1, 2, 1, 1, 1, 2, 1, 1~
## $ FIRSTTRR <dbl> 3, 6, 6, 1, 6, 1, 1, 6, 6, 3, 6, 3, 6, 1, 3, 3, 6, 1, 6, 1, 3~
## $ VERDATE  <chr> "29JUN2021", "29JUN2021", "29JUN2021", "29JUN2021", "29JUN202~
## $ WTPP     <dbl> 971.3926, 1821.0276, 4717.9118, 1805.0580, 2323.2049, 2833.87~
```

Here is a glimpse of the cleaned dataset:

```
## Rows: 8,005
## Columns: 8
## $ gender      <chr> "Male", "Female", "Male", "Female", "Male", "Male", "Male"~
## $ age         <chr> "20-24 years old", "45-54 years old", "35-44 years old", "~
## $ province    <chr> "Nova Scotia", "British Columbia", "British Columbia", "Ne~
## $ smoked      <dbl> 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1~
## $ vaped       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1~
## $ cannabis    <dbl> 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1~
## $ first       <fct> Cannabis, None, None, Smoking, None, Smoking, Smoking, Non~
## $ vaped_first <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Here is a glimpse of the matched dataset:

```
## Rows: 7,140
## Columns: 10
## $ gender      <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "M~
## $ age         <chr> "15-19 years old", "15-19 years old", "15-19 years old", "~
## $ province    <chr> "Quebec", "Quebec", "Quebec", "Quebec", "Quebec", "Quebec"~
## $ smoked      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ vaped       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ cannabis    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ first       <fct> None, None, None, None, None, None, None, None, None, None~
## $ vaped_first <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ .fitted     <dbl> 2.21166e-10, 2.21166e-10, 2.21166e-10, 2.21166e-10, 2.2116~
## $ cnts        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Supplementary Tables and Plots

Table 7: Summary of Matched Dataset by whether or not they Smoked

Smoked Before	Frequency
False	3570
True	3570

Table 8: Summary of Models for Model Selction

	Full Model		Without First Product Tried First		Without Province and Whether or Not They Vape	
(Intercept)	-26.566	(31429.569)	-18.847	(509.832)	0.660 ***	(0.142)
genderMale	0.000	(8583.468)	0.445 *	(0.180)	-0.566 ***	(0.128)
age20-24 years old	0.000	(20527.094)	-1.479 ***	(0.191)	-1.560 ***	(0.125)
age25-34 years old	0.000	(23036.876)	-3.072 ***	(0.339)	-3.781 ***	(0.258)
age35-44 years old	0.000	(22388.483)	-4.760 ***	(0.776)	-6.292 ***	(0.715)
age45-54 years old	0.000	(22894.043)	-4.831 ***	(1.039)	-6.922 ***	(1.006)
age55-64 years old	0.000	(22330.404)	-5.193 ***	(1.035)	-7.091 ***	(1.006)
age65+ years old	0.000	(21873.373)	-19.456	(803.431)	-20.589	(402.425)
provinceBritish Columbia	0.000	(17825.523)	-0.508	(0.342)		
provinceManitoba	0.000	(22079.745)	-0.447	(0.477)		
provinceNew Brunswick	0.000	(24319.208)	0.027	(0.482)		
provinceNewfoundland and Labrador	-0.000	(26656.393)	-1.222	(0.662)		
provinceNova Scotia	-0.000	(23153.141)	-0.498	(0.484)		
provinceOntario	0.000	(15301.735)	-0.512	(0.297)		
provincePrince Edward Island	0.000	(26210.154)	-0.643	(0.553)		
provinceQuebec	0.000	(16557.733)	-0.521	(0.335)		
provinceSaskatchewan	0.000	(23837.296)	-0.076	(0.435)		
factor(smoked)1	0.000	(19177.098)	-2.232 ***	(0.185)	-0.710 ***	(0.153)
factor(vaped)1	-0.000	(14066.482)	22.842	(509.832)		
factor(cannabis)1	0.000	(12565.284)	-2.564 ***	(0.238)	-0.130	(0.134)
firstVaping	53.132	(28781.153)				
firstCannabis	0.000	(18642.762)				
firstNone	0.000	(22159.794)				

*** p < 0.001; ** p < 0.01; * p < 0.05.