Understanding Personal Factors Affecting Alcohol Use in Canada

Dhanraj Patel -1003965168

2021-12-17

Introduction

In Canada, the Canadian Community Health Survey (CCHS) is a national survey taken in a joint venture between Statistics Canada and the Canadian Institute for Health Information that consists of a volunteer based response survey that is open to Canadians. The goal of the survey is to better understand the health of Canadians, the survey asks health related questions in many categories including nutrition, drug use, alcohol use and other health related topics. In this study, by using linear regression in conjunction with the 2017-2018 CCHS [1] we will investigate the central research question of which personal factors of the survey participants most strongly linearly correlate to their alcohol consumption trends. The personal factors include the personal attributes of the survey participants which include their sex, age, family history and more. Through a linear regression model we will be able to answer the goal of this study of better understanding the alcohol consumption trends of Canadians by analyzing which personal factors most heavily influence heavy consumption of alcohol.

This study is important as by understanding which personal factors most strongly linearly coincide with heavy drinking we will be able to take the first steps to help those who suffer from alcoholic addiction. Many studies have shown drinking excessive amounts of alcohol can lead to a decline in one's health such as a study released by the World Health Organization (WHO) where they found a link between excessive drinking and an increased risk of cancer [2]. These negative effects impact the lives of everyday Canadians as we can see from a Global News report which reported that there is a "silent epidemic of alcoholism" in Canada and that seven Canadians pass away everyday from alcohol related symptoms [3]. From this we can see that alcoholism is an important issue in Canada.

Methods

Variable Selection

When selecting which variables to use in your final model it is important to consider many different factors in tandem. The first step to finalizing the model was to use contextual information to organize a subset of predictors that could potentially be incorporated into the final model. Since the goal of the study is to better understand which personal factors have the greatest linear correlation with alcohol consumption and the CCHS is a study that considers that studies the health of Canadian in various different health fields the first step taken was to organize a subset of variables that are pertinent to alcohol consumption and any factors that can influence alcoholic habits of the survey participants. Also, continuing to use contextual reasons for the final model the total number of alcoholic drinks the survey participants drank in the previous week (of the time when the survey was taken) was selected to be the response variable as the purpose of the study is to measure their personal factors against how much alcohol they

consume and in addition the seven variables containing the number of alcoholic drinks they drank in each day of the week was removed due to redundancy and issues of multicollinearity.

For selecting the predictor, backwards selection was used where we start off with every possible variable and slowly decrease the number of predictors based on the value of the model's BIC and adjusted R^2 values. Using this method we were able to arrive at a model that balances the highest possible adjusted R^2 value and lowest BIC value. The BIC value was chosen as it most heavily penalizes a model with a large number of predictors which leads to a model that is easy to interpret. Backwards selection was used as it allows us to use a mechanical way to maximize the favorability of the BIC and adjusted R^2 value, however it is important to note that this method introduces bias so as to compensate for the model diagnostics. F-Test and the T-test values were included to add contextual reasoning to the model that will offset the bias. Using all these methods together is how the final model was decided upon.

Model Violations and Diagnostics

When using backwards selection to finalize our model it is also important to perform model violations and diagnostics to ensure that the variables satisfy the assumption of linear regression and to understand and alleviate the limitations of the variables. When a potential is modeled using this method an F test is necessary to ensure at least one predictor is linearly related to the response. In addition, when selecting the predictors for the model it is important to know if the individual variables are significantly linearly related to the response variable. This was done by performing a T-test and those variables that were not significantly related were not included in the final model. This is important as our study goal seeks to understand the most relevant personal factors pertinent to the survey participants' alcohol consumption and it is important that all predictors be linearly related to the response variable.

The next model diagnostics performed was to see if the model followed the assumptions of linear regression. Before checking the assumptions it is necessary to check to see if the two conditions have been met. The first condition states that the conditional mean response is a single function of a linear combination of the predictors and the second condition states that the Conditional mean of each predictor is a linear function with another predictor. We can analyze these conditions by using pairwise plots of the predictors and a y vs. y_hat graph of the prospective model. Moving on, we can then use a standardized residual plot and qq plot of the variables to assess if the prospective model has any issues in linearity, non constant variance, normality and uncorrelated errors. When analyzing these assumptions we use box plot transformations to improve the normality and linearity of the model and we can use variance stabilizing transformations to improve issues with non constant variance. We would ideally want all the assumptions to be satisfied in our final model and the predictors in the final model should be able to get as close to this as possible. Lastly, it is important to check the assumptions in our model to verify the legitimacy of using linear regression on the model and identify biases.

Moving on we will then use the VIF score of the model to identify if the predictors have an issue with multicollinearity. Ideally we should want a VIF score as low as possible, but it is important that the final model's VIF score is less than 5 and ideally lower than 1. To fix high levels of multicollinearity we would remove/replace predictors to lower the score while also balancing the

adjusted R^2 value of the model. Lastly, it is important to identify outliers, leverage points and influential points in the data. These points may deviate the results and lead us to acquiring less accurate results. Ideally we should have as few as possible, but if the predictor with them can not be removed then they will be noted in limitations. These steps are important to ensure the results are accurate and reproducible.

Model Validation

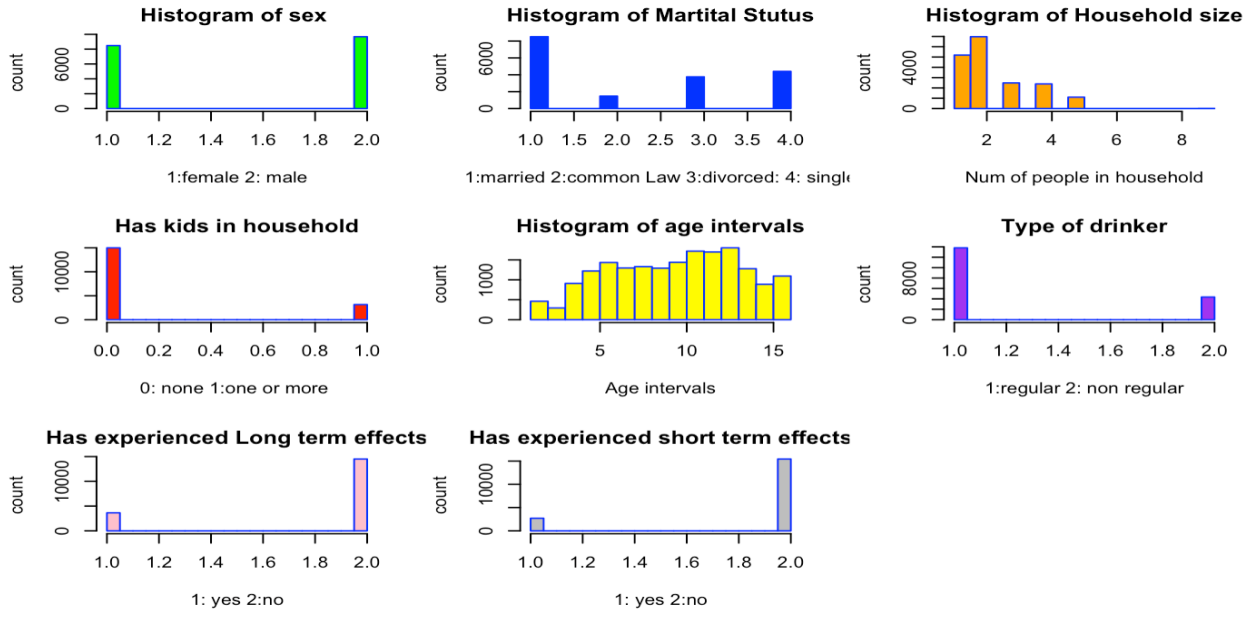
It is important to validate models to better understand how reproducible the model is and how significant of an impact the limitations had on the accuracy of the model. To validate our model at the very beginning the data was randomly separated into two groups. The two groups were split 50/50 and by comparing we can identify any differences in characteristics we can identify the reproducibility and limitations of the model.

Results

Description of Data

The first step of obtaining a model was using context to narrow down the number of variables to the ones that are pertinent to answering the central question of the study. The response variable was selected to be the number of alcoholic drinks the survey participants drank in the previous week of when the survey was taken. As for the predictors after narrowing down potential personal factors that may have contributed to the alcohol consumption trends, nine possible variables were selected. The graphical summaries are present following:

(graph_1)



caption: Histograms of all potential predictors considered for final model

In Graph_1 we see that the predictor variables are all discrete variables and there seems to be varying levels of skew when comparing normality of all the variables. As for the sex of the participants we see that there is roughly an equal number of men and women with slight more men, sex should not skew the results. As with marital status we see that most people are married, roughly equal number of people are divorced and single with the lowest number of people in common law. Since it is discrete it should not affect normality of data. As for household size we see that most people live alone or with one other person with a lot less people in the survey living with at least two people, this can be further seen as not many more people reported having no kids in the survey compared to those who have kids. This may lead to a skew in data and data that underrepresented adults with children. As for the age intervals we see a slight left skew, but the data seems generally normal. As for the type of drinker, long and short term and house ownership we see that all four of these are binary options where one option has a much higher count than the other. Like with children this may increase the bias in the results and lead to under representation in the results.

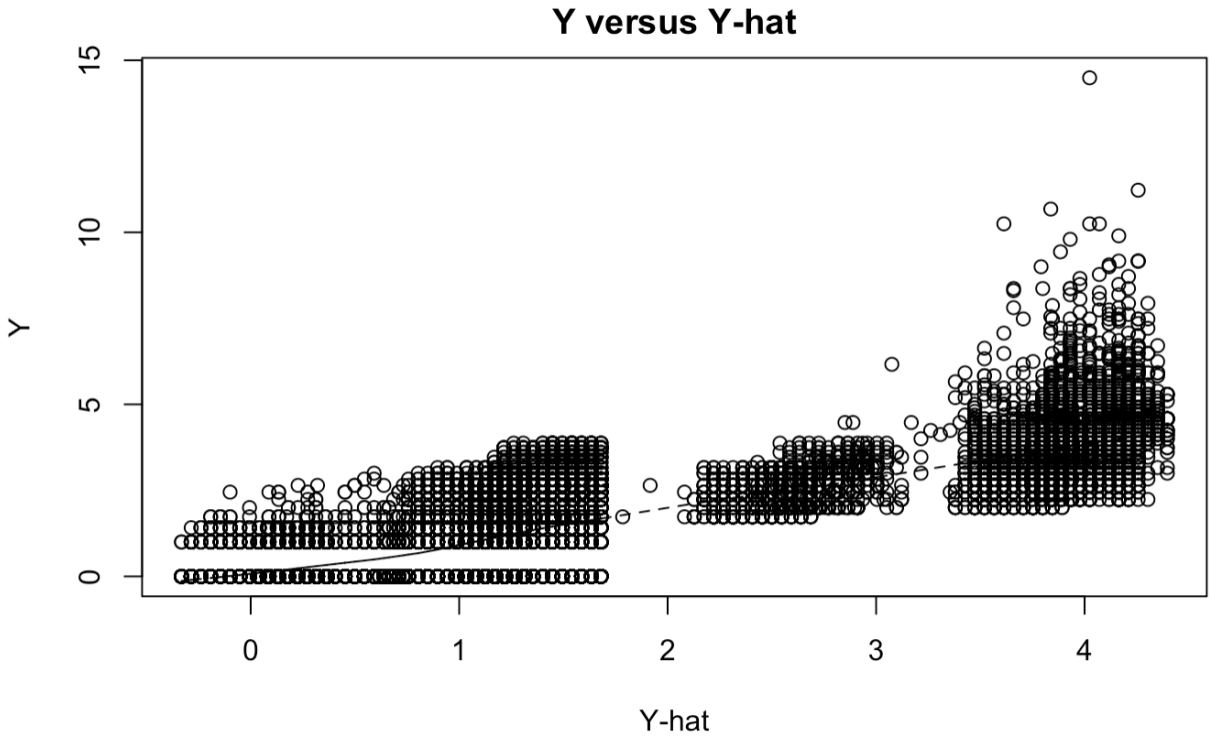
In addition to the predictors using numerical summaries we can examine the properties and skew of the response variable. In the response we found a large right skew which impacts the assumptions. Please see Table_1 for more detail.

Process of Obtaining Final Model

The first process of obtaining the final model after the EDA was performing a backwards selection process and comparing it to a stepwise selection for further context. The model compared the BIC and adjusted R^2 of the models and suggested that the best model included seven predictors. When comparing it against an stepwise selection model that the variables marital status and household size add very little adjusted R^2 values and removing them decreased the BIC value by a significant value so I decided to remove those variables leaving me with the variables age, sex, short term risk, long term risk and type of drinker. The model with the five predictors resulted in an adjusted R^2 value of 0.5127, F statistic of 3820 and using T tests all five were found to be significantly related to the predictor.

Moving on it was then important to check the assumptions of the model. From a look at the two conditions and the standardized residuals it was evident that the model would be required to improve the violations present. The variables in the model were transformed using the predictors using the following box cox transformations. To see exact transformations done please see Table_2 in appendix.

(Graph_2)

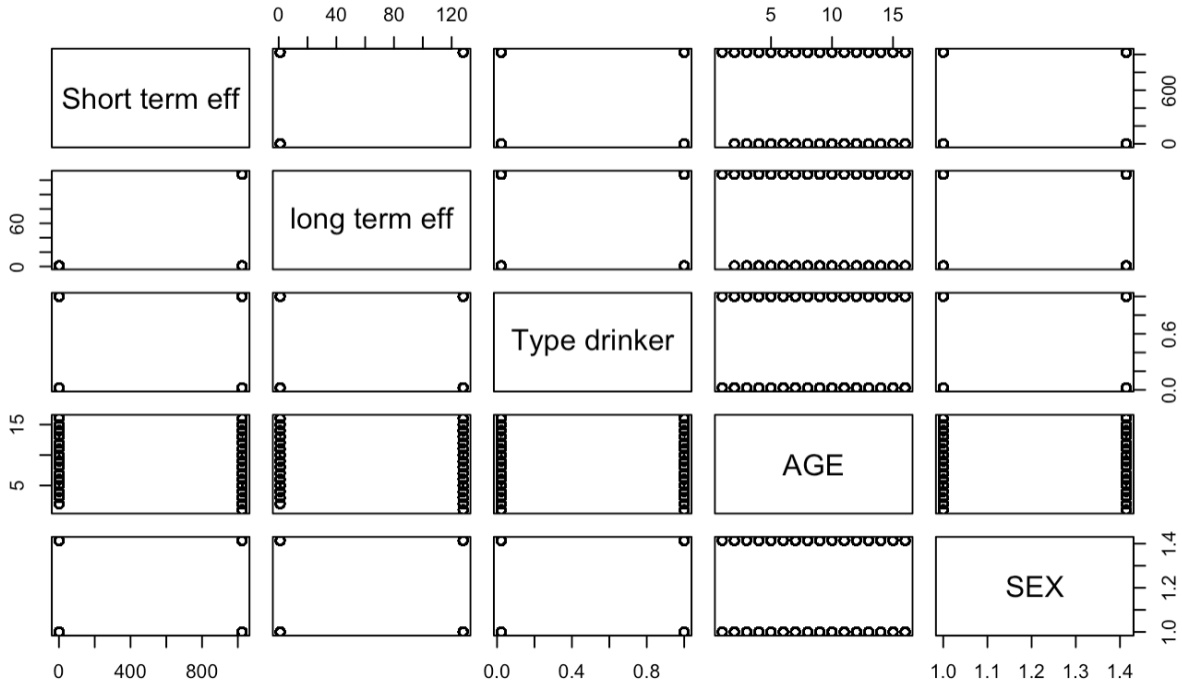


Caption: condition 1 y vs y_hat plot

From the y vs y_hat graph we see that there are issues in the first condition. We see that there is a non random scatter around the identity function, three scattered sections and an upwards curve. The Y vs Y_hat is after the transformation and the best found results after attempting many transformations. The issue will impact the veracity of the residual plots.

The first condition of the model can be seen with pairwise graphs which is shown below:

(Graph_3)

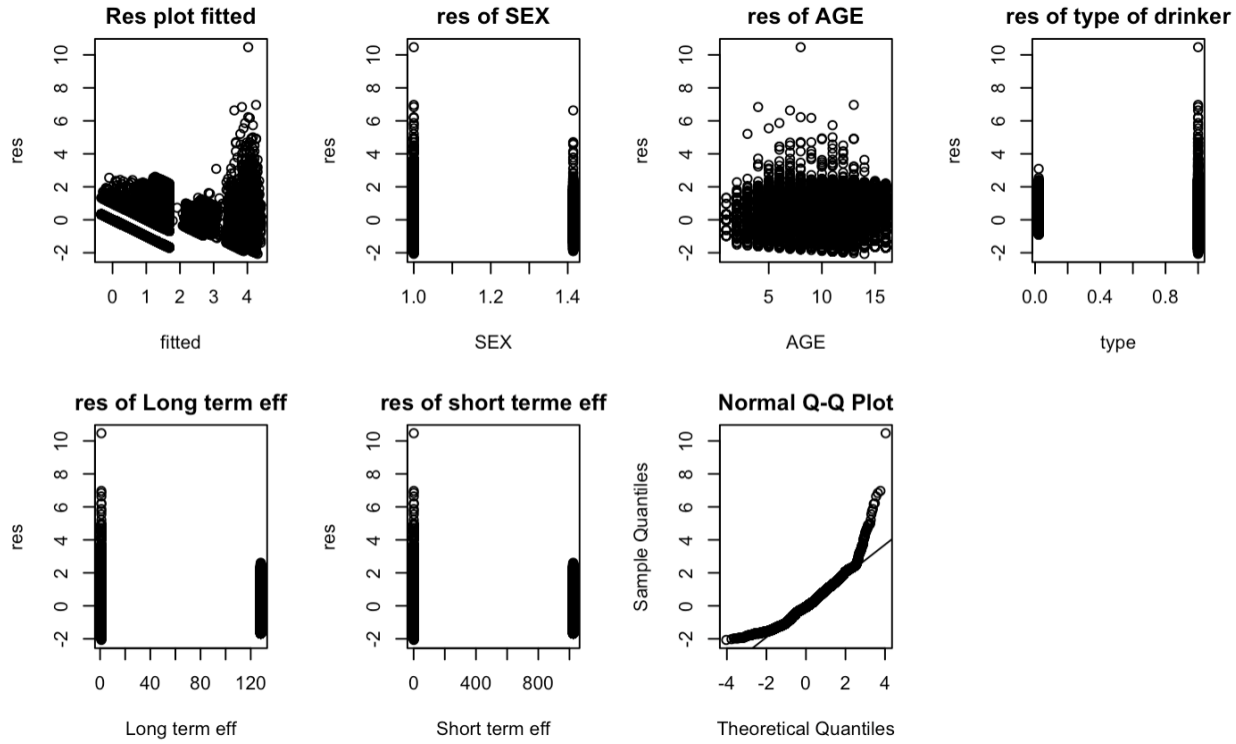


Caption: pairwise functions of predictors to check condition 2

From graph_3 it is important to know that all of the predictors were discrete options and many of the variables were also binary. After performing the transformations we see a random scatter with no curves or other issues and leaves us to believe that the condition holds.

After the conditions are checked we can move onto the residuals and qq plot to assess the assumptions.

(Graph_4)



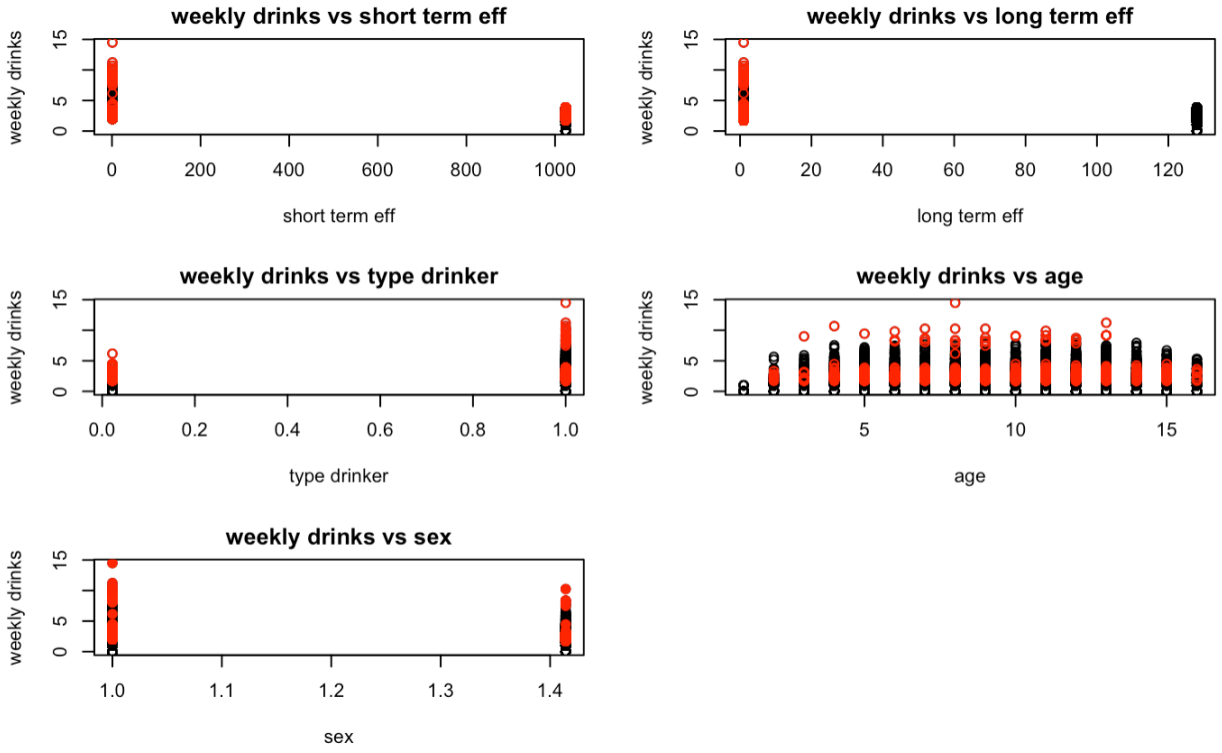
Caption: residual plots of data

With graph 4 we can see the residuals of the fitted value, the predictors and the qq plot. The fitted values seem to have issues with uncorrelated errors as we see two large clusters present with outliers seen at the top. The predictor residuals have no notable issues and with the QQ plot we see that the values curve above the line at the ends meaning we have issues with normality.

In addition after the transformations it is important to check the VIF score for multicollinearity issues. All predictors had scores between 1 and 1.5 except the short and long term effects variables which had scores between 3 and 3.2. These variables were not removed as removing them affects the adjusted R^2 heavily and the VIF scores are below 5. No variables were removed in this step.

Finally it is important to check for the presence of outliers, leverage points and other influential points that might affect data. They can be seen from the following graph:

(Graph_5)



Caption: special points as seen in the data

From the red dots we see the influential points on our graphs that could be (not guaranteed) to be either outliers or influential points. Since these predictors are discrete the issue of a large number of these special points (in red) come from the response variable. Since in the survey while many people drink normal amounts there are those we drink heavily and effect results by being outliers. The presence of these special points hinders the ability of the results of the model from portraying the general trend that we want by being influenced by these special points. No outliers were removed.

Validation

In the beginning after cleaning the data the data was split into two halves, one for training and one for validation. With our model finalized we perform all the steps that are done on the training model and compare. From the original dataset the training and testing datasets were split evenly. When calculated the mean of the variables were found to be similar in both datasets, please see Table_3 in appendix for exact values.

After checking their means it is important to check all other aspects that were done to ensure similar results. When checking the model both datasets had similar adjusted R^2, T-tests results,

F statistic values, coefficient values, special points, residual plots (and conditions) and VIF scores before and after transformation. We can conclude and say that validation was a success, but it is important to know that issues in the training also persisted in the validation set.

DISCUSSION

The goal of the study is to identify the personal factors that are most linearly related to alcohol consumption. After performing linear analysis from all the possible predictors it has been determined that the most significant personal factors relating to the consumption of alcohol are sex, age, type of drinker and increased long and short term risk due to drinking identification.

However, from this conclusion we need to consider the limitations of the analysis in which these results were attained. One major limitation was all the missing data, in the CCHS one has the option to skip any question and in our data cleaning we removed any observation that skipped even one question. This leads to an increase in non response bias in our results. In addition, we see that the survey had a large number of influential points in the response of heavy drinkers which also led to biased results. It led to issues with normality and non constant variance, this can be seen in the qq plot and y vs y_hat graph. Lastly all the predictors were discrete which led to imprecise data, this can be seen with the pairwise plots.

Conclusion

In Canada, drinking alcohol is a huge part of the culture; however, reports by the WHO and global news show that drinking alcohol is linked to many health concerns such as cancer. Using the 2017-2018 CCHS we chose the total number of alcoholic drinks consumed in the previous week as our response variable and using context indicators we narrowed down the possible set of predictors of personal factors that could influence the response variable. From this point we performed a statistical analysis to determine which factors were the most significantly related to the response variable. From our model we came to the conclusion that sex, age, type of drinker, and short and long term effects that are related to alcohol were the most pertinant variables. We can use this knowledge to help reduce the overall consumption by targeting these particular factors in alcohol awareness campaigns to reduce the number of alcohol related fatalities in Canada.

Appendix

(1)
(Table_1)

Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean	Var	SD	Range
18152	0	0	2	6	210	6	4.45	2.8	60	7.76	210

caption: Numerical summaries of response variable (total number of alcoholic drinks consumed in past week)

From Table_1 we see that the response variable has a large right skew. In the data the presence of heavy outliers caused by heavy alcoholic drinkers causes a great skew in the data. We see from the min value up to q3 the average number of drinks was between 0 and 6, but from q3 to the max there is a range from 6 to 126 This seriously dampans the normality in the response variable and could cause a lot of issues with the model assumptions (especially normality).

(2)

(Table_2)

variable	Transformation
Number of drinks in a week	var**(0.5)
Short term effects present	var**(10)
long term effects present	var**(7)
Sex	var**(0.5)
Type of drinker	var**(-5.5)
Age	Not transformed

Caption: These box cox transformations were performed to improve assumptions in the model.

(3)

Table_3

Variable name	Training dataset means	Validating dataset means
Sex	1.527931	1.530895
Age interval	9.742893	9.737228
Short term effects	1.847014	1.851263
Long term effects	1.794348	1.799294
Type of drinker	1.232536	1.243187
Weekly num of drinks	4.446838	4.250414

Caption: From table_3 we see that both the training and validation datasets have very close means for their variables.

Bibliography

1. ODESI2. Scholars Portal. http://odesi2.scholarsportal.info/webview/ (Last Accessed: Dec 17, 2021)

2. (2021, July 13) *New WHO study links moderate alcohol use with higher cancer risk*.Centre for Addiction and Mental Health https://www.camh.ca/en/camh-news-and-stories/new-who-study-links-moderate-alcohol-use--with-higher-cancer-risk (Last Accessed: Dec 17, 2021)

3. Bains C. (2019, June 13). *Alcohol-related deaths remain a 'silent epidemic' in Canada: expert.* Global News The Canadian Press. https://globalnews.ca/news/5386829/alcohol-deaths-hospital-study/ (Last Accessed: Dec 17, 2021)