What factors influence personal health satisfaction level in Canada?

Jinwen Tao, Yuanxi Gong, Zhifei Cao

Abstract

The national statistical office — Statistics Canada had conducted a General Social Survey that collected the key features of Canadian's health condition in 1991 to ensure citizens' fitness and social services' performances. Based on the dataset from the survey, we are expected to plot regression models and graphs in order to discover some factors such as stress levels, sleeping time, taking flu shots and years of smoking that could possibly influence individuals' health satisfaction levels. By analysing the results, smoking generates an adverse effect on personal health and citizens with longer sleeping time or relatively lower anxiety levels are usually satisfied with their current health conditions. These social trends monitor the changes in people's health and living conditions, and allow the government and society to take effective actions to improve health services in the future.

Introduction

The Canadian General Social Survey (GSS) , which focuses on gathering information on economy, society and health services in Canada, provided both the government and citizens with a clear and unbiased perspective on how the living conditions of Canadians change over time and what are Canadians' views on some current political issues. In the General Social Survey in 1991, it investigated the health conditions of approximately 10,000 Canadian citizens who are older than 15 or more in ten provinces. Since the dataset of the General Social Survey contains hundreds of variables, we specifically focus on how variables like stress levels, sleeping time, taking flu shots and years of smoking influence individuals' health satisfaction levels.

By plotting a poisson regression model, some correlations between these variables are revealed. We figure out that taking the flu shot has a marginal impact on their choice of health satisfaction levels. After drawing two separate pie charts for taking flu shots, people who took the flu shot in the recent one-year period surprisingly have lower health satisfaction levels. Then we plot the bar charts and tables for degree of relaxation, sleeping time and years of smoking respectively to summarise the data based on their health satisfaction levels. Citizens with longer years of smoking have collectively lower health satisfaction levels, while those who sleep relatively longer and stay relaxed have better health conditions.

Individuals' health satisfaction levels demonstrate citizens' current health states and their living conditions. By investing in factors that could affect persons' health satisfaction levels, it allows the society and government to draw more attention to health services and gather information on Canadians' health conditions. However, the dataset which was collected in 1991 was too old and thus it might fail to reflect the health conditions of Canadians in 2020. Since the survey was taken by telephone, there are many unresponded questions and this largely affects the accuracy of our dataset. As a consequence, our outcome will be less representative for responders who did not answer the phone. Aside from that, we cannot ignore the fact that respondents will not always provide accurate and honest answers during the survey. While, our data strongly relies on respondents' answers , so we can only suppose that they provide honest answers. However, we do find a few unreasonable answers that will affect our result in the study.

Data

The dataset —- “General social survey, cycle 6 - health, 1991” is downloaded from the series of Canadian General Social Survey(GSS) on the CHASS website. This dataset was created and distributed in July.9th, 1991 by the Producers Statistics Canada and Housing, Family and Social Statistics Division in Ottawa. In this cycle 6 General Social Survey, Distributors state that main objectives are gathering data on social issues and provide information on social stipulation to monitor Canadians’ satisfaction and social interests on different aspects of health problems including social satisfaction, use of alcohol and smoke, time of sleep, services on medical care and so on. The dataset has the strength that contains lots of variables which could reflect Canadians’ social well-beings in different aspects. However, a serious weakness has been revealed in analyzing works as most of the data was categorical variables.

This questionnaire contains 22 sections and every section has multiple survey questions. Since it covers a wide spectrum of social and personal questions, this questionnaire can be considered as a comprehensive survey that covers all-sided problems. It ranges in depth from the simple question like the personal information to the deep feelings such as social satisfaction. Moreover, this questionnaire has two types of official languages, English and French, which makes respondents more convenient to choose their familiar language to answer the questionnaire.

Nonetheless, the weak part of this survey is lack of the numerical data. Going through the whole questionnaire, we could find out that 90 percent of questions are multiple choice with provided answers, and only few of them are open questions. Thus, this directly causes more categorical variables and less numerical variables for the dataset, which creates difficulties on data analyzation. In addition, the questionnaire can be done by Proxy if the participant is ill or have language barriers. The replacement of taking surveys by Proxy may influence the choice of the real participant and cause uncertain responses.

According to the survey documentation from the CHASS website, non-institutionalized persons with the age greater than 15 years old who live in the ten provinces in Canada are the target population for this survey. The survey takes a sample of approximately 10000 people from the target population.

The frame of the Canadian General Social Survey contains a list of telephone numbers of the survey samples — 15-year-old or above individuals that live in the geographic domain of provinces in Canada in the whole target population.

In order to find participants to involve, the survey was taken via telephone interview by using two-stage stratified random digit dialing (RDD), Waksberg method and Elimination of Non-Working Banks method. However, there are plenty of not stated or unanswered questions from the survey participants. Since the survey is conducted by telephone and it took hours to reply to all the survey questions , some people might hang up the telephones or refuse to answer all the questions. This causes many non-response questions which influence the accuracy when we analyse the survey.

As this health survey is a form of telephone interview, two-stage stratified random digit dialing (RDD) was used. For two-stage stratified random digit dialing, the first sample group was randomly selected according to the specific strata like located provinces of the telephone numbers. Then, the distributor would randomly select the second sample group from the first sample group. For ten provinces, the Waksberg method was used. The Waksberg method is a random sampling method that increases the probability of reaching households by collecting the telephone area code and prefix number combinations into the survey area in 10 provinces. For the exclusion of three territories, we used the Elimination of Non-Working Banks method. Since it was difficult to reach a household by telephone companies in territories, the Elimination of Non-Working Banks method was used to exclude the three territories in order to orient households that can be located by the telephone numbers.

We lack information on the cost of the survey in 1991 and it was not provided on the documentation of the CHASS website. However, one of the sampling approaches — Waksberg method will reduce the costs of the survey dramatically compared with the complete randomization of dialling the telephone numbers.

```
## Rows: 6,088
## Columns: 6
## $ `Health Satisfication Level` <chr> "Level 2 - Somewhat dissatisfied", "Leve…
## $ `Hours of Sleep per night`   <int> 10, 7, 6, 7, 8, 8, 7, 7, 6, 7, 8, 7, 4, …
## $ `Take Flu shot?`             <chr> "NO", "YES", "NO", "NO", "NO", "YES", "N…
## $ `Years of Smoking`           <int> 13, 12, 50, 30, 12, 38, 15, 3, 20, 2, 14…
## $ `Degree of Relaxation`       <int> 4, 4, 4, 2, 2, 3, 4, 3, 3, 3, 2, 3, 2, 4…
## $ Gender                       <chr> "Male", "Male", "Male", "Male", "Female"…
```

```
table_overall
```

Health Satisfication Level	Hours of Sleep per night	Take Flu shot?	Years of Smoking	Degree of Relaxation	Gender
Level 2 - Somewhat dissatisfied	10	**NO**	13	4	**Male**
Level 4 - Very satisfied	7	**YES**	12	4	**Male**
Level 1 - Very dissatisfied	6	**NO**	50	4	**Male**
Level 2 - Somewhat dissatisfied	7	**NO**	30	2	**Male**
Level 3 - Somewhat satisfied	8	**NO**	12	2	**Female**
Level 4 - Very satisfied	8	**YES**	38	3	**Female**
Level 4 - Very satisfied	7	**NO**	15	4	**Female**
Level 3 - Somewhat satisfied	7	**NO**	3	3	**Male**
Level 1 - Very dissatisfied	6	**NO**	20	3	**Female**
Level 2 - Somewhat dissatisfied	7	**NO**	2	3	**Male**
Level 3 - Somewhat satisfied	8	**NO**	14	2	**Female**
Level 3 - Somewhat satisfied	7	**NO**	20	3	**Male**
Level 1 - Very dissatisfied	4	**NO**	41	2	**Female**
Level 3 - Somewhat satisfied	7	**NO**	51	4	**Male**
Level 1 - Very dissatisfied	6	**NO**	45	3	**Male**
Level 4 - Very satisfied	9	**YES**	30	2	**Female**
Level 3 - Somewhat satisfied	8	**NO**	54	4	**Male**
Level 4 - Very satisfied	9	**NO**	56	1	**Male**
Level 3 - Somewhat satisfied	8	**NO**	57	3	**Male**

Health Satisfication Level	Hours of Sleep per night	Take Flu shot?	Years of Smoking	Degree of Relaxation	Gender
Level 4 - Very satisfied	4	**YES**	11	4	**Female**

Here is our cleaned-up data (raw data) and the table illustrates a preview of 20 lines from our dataset.

Six variables, which are chosen from raw data and renamed by using rename() function in Rstudio, are “Health Satisfaction Level”(dvn2a), “Hours of Sleep per night”(h2hrs), “Take Flu shot?”(d3), “Years of Smoking”(dvyrsmod), “Degree of Relaxation”(n3) and “Gender”(dvsex).

Since these variables contain many useless choices for the analysis, we need to clean up our data. After filtering out the choice of “Dissatisfied-degree not stated”, “Satisfied - degree not stated”, “No opinion”, “Not stated” and changing the order of the options “Somewhat satisfied” and “Very satisfied” in the “Health Satisfaction Level” variable, we have four health satisfaction levels in total from level 1 to level 4. The higher level you choose, the more satisfied you are likely to be with your health conditions. As moving out the options “Do not know” and “Not stated” for the variable “Hours of Sleep per night”, we obtain a qualitative variable for providing the number of hours of sleep per night for each respondent.

For the “Take Flu shot?” variable, we filter out the choices of “Do not know” and “Not stated”. Then, we get the cleaning variable “Take Flu shot?” that shows whether a responder takes a flu shot or not. We clean up the variable “Years of Smoking” by filtering out the missing data and “Not stated” option, then we could get numerical data that indicates the number of years of participants’ smoking. While cleaning the “No opinion” and “Not stated” choices, the “Degree of Relaxation” variable ranges in levels from “Very stressful” to “Not at all stressful”. Responders who are more stressful will choose a lower degree of stress according to those 4 levels. For the “Gender” variable, the raw data only includes the selections of male and female, so we can directly select this variable to analyze.

Among all the variables of satisfaction, we want to investigate the health satisfaction since most people take care of themselves’ health well-being nowadays. In order to see what factors affect the health satisfaction, we choose four variables(“Hours of Sleep per night”, “Take Flu shot?”, “Years of Smoking” and “Degree of Relaxation”) that we think are related to the choice of people’s health satisfaction levels. Compared with those variables in the dataset, these four variables are more representative to fit into the regression model in the model section, and are more easier to analyze each relationship between themselves and health satisfaction levels in the result parts. In addition, we select the Gender variable to indicate the number of females and males participants in different health satisfaction levels and discover the gender differences in health satisfaction levels.

Model

```
##
## Call:
## lm(formula = Health_Satisfaction_level ~ Hours_of_Sleep_per_night +
##     Smoking_Years + Flu_shot_yes + Degree_of_Relaxation, data = Effects_on_health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8281 -0.4164  0.2781  0.6496  1.3887
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.5743831  0.0642047  40.096  < 2e-16 ***
## Hours_of_Sleep_per_night   0.0366170  0.0080502   4.549 5.51e-06 ***
## Smoking_Years             -0.0041021  0.0006374  -6.436 1.32e-10 ***
## Flu_shot_yes1             -0.1324671  0.0297344  -4.455 8.54e-06 ***
## Degree_of_Relaxation       0.2280344  0.0120757  18.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8206 on 6083 degrees of freedom
## Multiple R-squared:  0.06561,    Adjusted R-squared:  0.06499
## F-statistic: 106.8 on 4 and 6083 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## glm(formula = Health_Satisfaction_level2 ~ Hours_of_Sleep_per_night +
##     Smoking_Years + Flu_shot_yes + Degree_of_Relaxation, family = "poisson",
##     data = Effects_on_health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7388  -0.2271   0.1329   0.3487   0.7552
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.9728877  0.0433221  22.457  < 2e-16 ***
## Hours_of_Sleep_per_night  0.0111035  0.0053972   2.057  0.03966 *
## Smoking_Years            -0.0012500  0.0004295  -2.910  0.00361 **
## Flu_shot_yes1            -0.0402114  0.0200893  -2.002  0.04532 *
## Degree_of_Relaxation      0.0687624  0.0080833   8.507  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1539.0  on 6087  degrees of freedom
## Residual deviance: 1452.2  on 6083  degrees of freedom
## AIC: 20014
##
## Number of Fisher Scoring iterations: 4
```

We choose the poisson regression model to analyze how sleeping time, years of smoking, taking the flu shot and degree of relaxation affect individuals' health satisfaction level. Our response variable is the health satisfaction levels of our respondents and it is a discrete variable. We have 4 explanatory variables which are sleeping time, years of smoking, taking the flu shot and degree of relaxation. More specifically, taking the flu shot is a categorical variable with the outcomes of "yes" or "no". The sleeping time per night, years of smoking and degree of relaxation are numeric variables.

Poisson regression is used to model response variables that are counts. Since health satisfaction level is classified as a count variable, poisson regression model could be the most appropriate model to demonstrate the relationships between health satisfaction level and other independent variables. Although a negative binomial regression model could also deal with response variables that are discrete, we could not use it since we include both categorical explanatory variables and numeric explanatory variables. Therefore, a poisson regression model is still the best choice in this situation.In addition, we choose the years of smoking instead of classifying the years into groups in order to have a numeric variable. If we classify the years of smoking into different groups, this would be a categorical variable which could not reflect the health satisfaction levels accurately in a poisson regression model.

As assume the response variable Y (health satisfaction level) following a poisson distribution and the log of E(Y) (expected health satisfaction level) could be modelled by a linear combination of parameters, we can apply poisson regression model and obtain a formula for our data:

$log (E(Y))= B0 + B1x1 + B2x2 + B3x3 + B4x4+ ln(t) $

$$log(E(healthsatisfactionlevel)) = B0 + B1 * (hoursofsleepingpernight) + B2 * (lengthofsmokingyears) + B3 * (flushotornot) + B4 * (degreeofrelaxa$$

, where ln(t) is an offset which represents the denominator of the rate. Although ln(t) appears on the right side of the equation, it is forced to have a value of 1. B0 is the intercept parameter and reflects the log of expected health satisfaction level when the other explanatory variables are equal to zero. B1, B2, B3, B4 are coefficients that reflect how the changes in x could affect our response variable.

After using R software to run the model we can summarise the values for B1, B2, B3, B4. Mathematically, we have

$log (E(Y))= 0.97 + 0.011x1 - 0.001x2 - 0.040x3+ 0.068x4 + ln(t) $

, where B0 with the value of 0.9728, B1 with the value of 0.011, B2 with the value of - 0.0012, B3 with the value of -0.04 and B4 with the value of 0.068.

Since our x3 is a categorical variable, it demonstrates the difference in health satisfaction levels of taking the flu shot or not. If a person took the flu shot, then we have x3 equals to 1, otherwise, x3 remains zero. According to our data, taking flu shot has an unexpected outcome that people with flu shot injecting actually have 0.040 unit lower in log of expected health satisfaction levels while those who did not take flu shot have 0.040 unit higher in log of expected health level.

Therefore, when people take the flu shot, the formula will be:

$log (E(Y))= 0.97 + 0.011x1 - 0.001x2 - 0.040x3+ 0.068x4 + ln(t) $

When people did not take flu shot, the formula will be:

$log (E(Y))= 0.97 + 0.011x1 - 0.001x2+ 0.068x4 + ln(t) $

In both cases, X1 is the length of sleeping per night and it has a positive influence on individual health satisfaction level since B1 is a positive coefficient. This illustrates that 1 unit increased length of sleeping time could boost up 0.011 unit in log of expected health satisfaction level. We could interpret this by taking the exponentiated value, e^0.011 = 1.011. This illustrates that there is a 11% increase in the health satisfaction level for 1 more hour of sleep.

Similarly, X4 shows the degree of relaxation and people with higher degrees of relaxation tend to have higher health satisfaction levels. This means 1 unit increase in degrees of relaxation could lead to 0.068 increase in log of expected health satisfaction levels.This represents an increase in 1 level of relaxation degree will have 7% (e^0.068 = 1.07) increase in health satisfaction level.

X2 represents the length of smoking years. The longer years as a smoker has, not surprisingly, a negative effect on health satisfaction level. Specifically, 1 unit increase in years as a smoker, 0.001 unit decrease in log of expected health satisfaction level. By taking the exponentiated value e^0.001 = 1.001, this shows that 1 more year of smoking could decrease the health satisfaction by 0.1%.

The null hypothesis assumes that the response variable and the explanatory variable are not related, that B0,B 1, B2, B3 ,B4 are zero . By summarizing the model, the p-value for the slope parameter is close to zero which indicates we can reject our null hypothesis that B0 is zero. Therefore, we have the p-value of 0.97 for our B0. The p-value for explanatory variables X1,X2,X3,X4 are 0.0396, 0.0036, 0.04 and 0 respectively. Since all of them have a p-value less than 0.05, they are statistically significant and reject the null hypothesis of equaling to zero. Although p values for years of smoking and flu shot have relatively large, we still have weak evidence to reject null hypothesis. This means all of X1,X2,X3,X4 have correlations with our response variables. Therefore, stress levels, sleeping time, taking flu shots and years of smoking could influence individual health satisfaction levels to different degrees.

The alternative model we use is the multiple linear regression model for analyzing the same variables that we used in the poisson regression

model. We use lm() function in the R Software to create the linear regression lines, $ y = b_0 + b_1x_1+ b_2x_2 + b_3x_3 + b_4x_4$, where y is the health satisfaction level, b_0 is the intercept parameter with value of 2.574, x_1 is the first slope parameter — "Hours of Sleep per night", x_2 is the second slope parameter — "Years of Smoking", x_3 is the third slope parameter — "Take Flu shot?", x_4 is the fourth slope parameter — "Degree of Relaxation", b_1 is the coefficient of slope parameter x_1 and has the value 0.037, b_2 is the coefficient of slope parameter x_2 and has the value -0.004, , b_3 is the coefficient of slope parameter x_3 and has the value -0.132, and b_4 is the coefficient of slope parameter x_4 and has the value 0.228.

Mathematically, for people who did not take the flu shot(baseline): $ Health Satisfaction Levels = 2.574+ 0.037*(Hours of Sleep per night) - 0.004*(Years of Smoking) -0.132*(Take Flu shot?) + 0.0688*(Degree of Relaxation)$ $= 2.574 + 0.037*(Hours of Sleep per night) - 0.004*(Years of Smoking) - 0.132*0+ 0.228*(Degree of Relaxation)*

$$= 2.574 + 0.037 * (Hours of Sleep per night) - 0.004 * (Years of Smoking) + 0.228 * (Degree of Relaxation).$$

While, for people who took the flu shot in mathematical formula: $ Health Satisfaction Levels = 2.574 + 0.037*(Hours of Sleep per night) - 0.004*(Years of Smoking) - 0.132*(Take Flu shot?)+ 0.228*(Degree of Relaxation) $ = 2.574 + 0.037*(Hours of Sleep per night) - 0.004*(Years of Smoking) - 0.132*1+ 0.228*(Degree of Relaxation) $ $ = 0.932 + 0.037*(Hours of Sleep per night) - 0.004*(Years of Smoking) + 0.228*(Degree of Relaxation)$.

For these two regression lines that illustrate the health satisfaction levels of people who did not take and took the flu shots, as we increase 1 unit of slope parameter "Hours of Sleep per night", the y variable "Health Satisfaction Levels" increases 0.037 unit on average. Similarly, as we increase 1 unit of slope parameter "Years of Smoking", the y variable "Health Satisfaction Levels" decreases 0.004 unit on average. For an increase in 1 unit of slope parameter "Degree of Relaxation", the y variable "Health Satisfaction Levels" increases 0.228 unit on average. When those four slope parameters are unchangeable, we have the intercept parameter "Health Satisfaction Levels" to maintain in the 2.574 unit.

As we assume the health satisfaction levels of people who did not take flu shots is the baseline of the model, the slope parameter x_3 — "Take Flu shot?" can be considered as 0 to represent that people did not take the flu shots. So, if people did not take flu shots, the y variable "Health Satisfaction levels" will not change at all. Surprisingly, if people have taken flu shots before, the slope parameter x_3 — "Take Flu shot?" can be considered as 1 and the y variable "Health Satisfaction levels" will decrease 0.132 unit. Since we doubt this result, we decide to do further discovery in the result section to investigate the distribution of the number of people who took the flu shots and analyze why the model shows a negative relationship between "Health Satisfaction levels" and "Take Flu shot?".

Since all the p-values for each slope parameter are less than 0.05, we can conclude that our model has evidence against the null hypothesis. Compared with the poisson regression model, the multiple linear regression model is more simple to fit into the regression lines and is more easier to understand the interpretation of the model. However, there exists a big weakness that we can only fit the data into this linear regression model as an alternative model. A linear regression model is better to use the continuous variable as the response variable, thus our alternative model will be inaccurate as the response variable "Health Satisfaction Levels" is a discrete variable that contains only 4 levels.

Results

```
Table_of_Flu_shot_yes
```

Health Satisfication Level	Number of People	Proportion	People with Flu shot (%)
Level 1 - Very dissatisfied	72	0.07280081	7.3

Health Satisfication Level	Number of People	Proportion	People with Flu shot (%)
Level 2 - Somewhat dissatisfied	109	0.11021234	11.0
Level 3 - Somewhat satisfied	329	0.33265925	33.3
Level 4 - Very satisfied	479	0.48432760	48.4

Table 1

Table 1 shows the number and the proportion of people with flu shot for each health satisfaction level. There are four health satisfaction levels in total. Level 1 means very dissatisfied. Level 2 means somewhat dissatisfied. Level 3 means somewhat satisfied. Level 4 means very satisfied. That's, the bigger number you choose, the more satisfied you are likely to be with your health. In all the answers, about half of (about 48.4%) responders choose "very satisfied" while only a few of responders(about 7.3%) choose "very dissatisfied".

```
Table_of_Flu_shot_no
```

Health Satisfication Level	Number of People	Proportion	People without Flu shot (%)
Level 1 - Very dissatisfied	237	0.04647970	5
Level 2 - Somewhat dissatisfied	491	0.09629339	10
Level 3 - Somewhat satisfied	1670	0.32751520	33
Level 4 - Very satisfied	2701	0.52971171	53

Table 2 Table2 shows the number and the proportion of people without flu shot for each health satisfaction level. There are four health satisfaction levels in total. Level 1 means very dissatisfied. Level 2 means somewhat dissatisfied. Level 3 means somewhat satisfied. Level 4 means very satisfied. That's, the bigger number you choose, the more satisfied you are likely to be with your health. In all the answers, More than half of (about 53%) responders choose "very satisfied" while only a few of responders(about 5%) choose "very dissatisfied".

Figure A **Figure B**

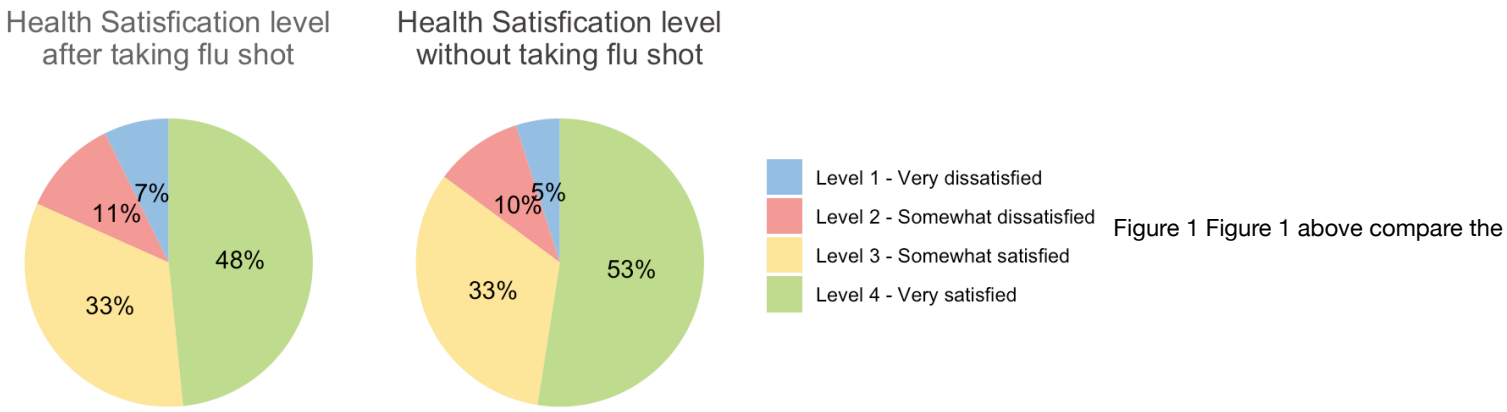

Figure 1 Figure 1 above compare the people who take flu shots with people who did not. Out of all citizens surveyed, “Very satisfied”(level 4) is the most popular answer, with 48% and 53% respectively. Meanwhile, about an equal number of responders in two groups answered “Somewhat satisfied”(level3) for this question. “very dissatisfied” (level1) is considered as the least favorite choice for both groups, with 7% and 5%respectively. While, responders who take flu shots tend to choose “very dissatisfied” (level1) and "somewhat dissatisfied” (level2) than those who don’t. Although both groups show similar results, it is clear that responders without flu shot are more likely to be satisfied with their health conditions than responders who take flu shot.

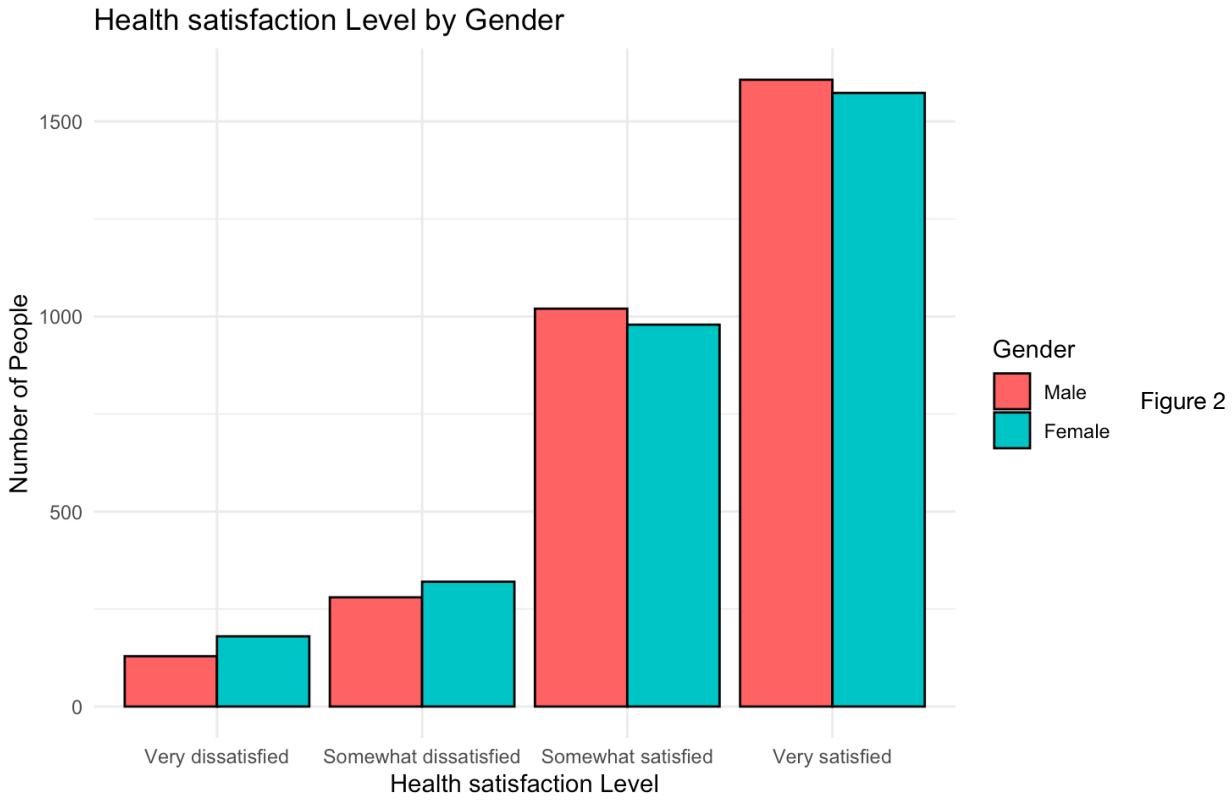

Figure 2

In Figure 2,the histogram clearly shows people's satisfaction towards their health. It is obvious that more than two thirds of citizens choose "satisfied" rather than "dissatisfied". Out of all citizens surveyed, "Very satisfied" is the most popular answer, with about 1650 male responders and 1600 female responders. While, about equal number of male responders and female respondents answer "Somewhat satisfied" for this question. In the graph, we can find that both male and female groups give similar results for this question. It is interesting that male responders are more likely to be satisfied with their health conditions than female responders.

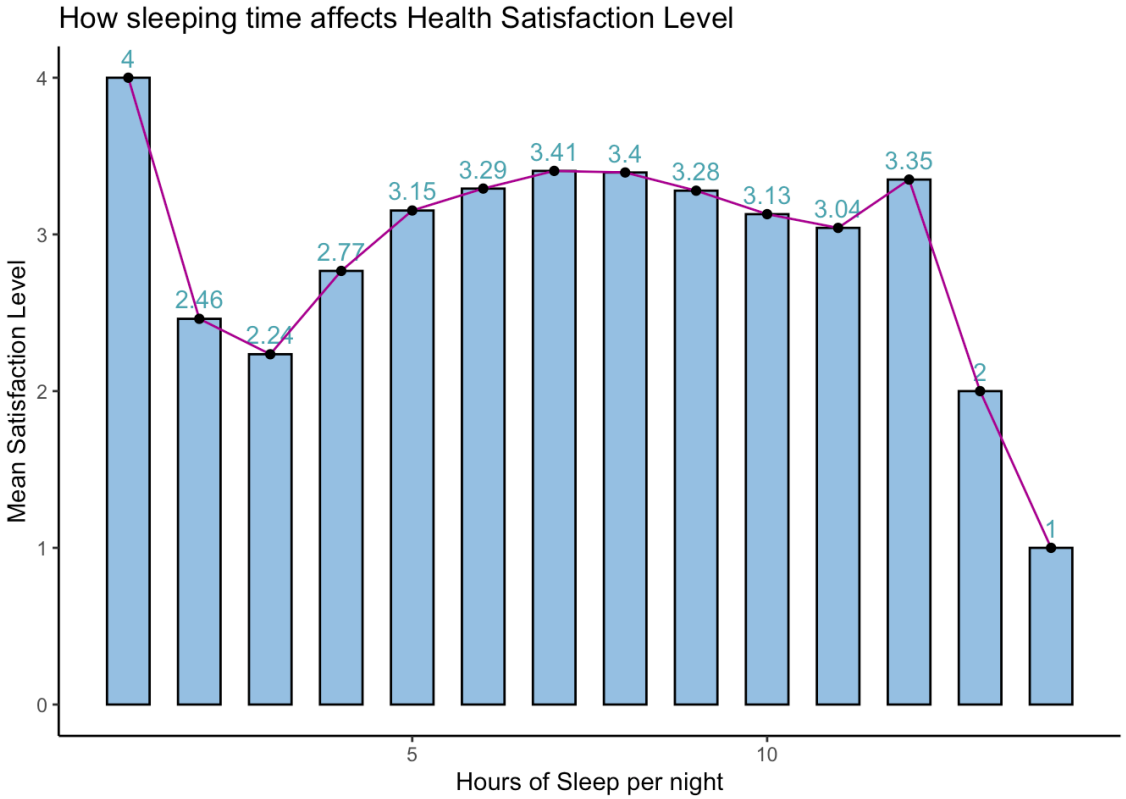

Figure 3

In Figure 3, a bar chart shows mean health satisfaction levels of responders who have different hours of sleeping. We can find an outlier on the graph. It indicates that responders who sleep 1 hour per night show the highest mean value of health satisfaction level (level4). Out of the rest of citizens surveyed, the responders who sleep 12 hours per night display the highest mean value of health satisfaction level (level 3.35). It is obvious that an increasing trend of mean health satisfaction levels is shown when sleeping hours grow from 3 to 7 hours. It peaks at 3.41 level when the sleeping hours is 7. After that, the mean value declines slowly when sleeping hours grow from 7 to 11 hours. However, when sleeping hours increase over 12 hours, the mean of health satisfaction levels decreases dramatically to 1 level.

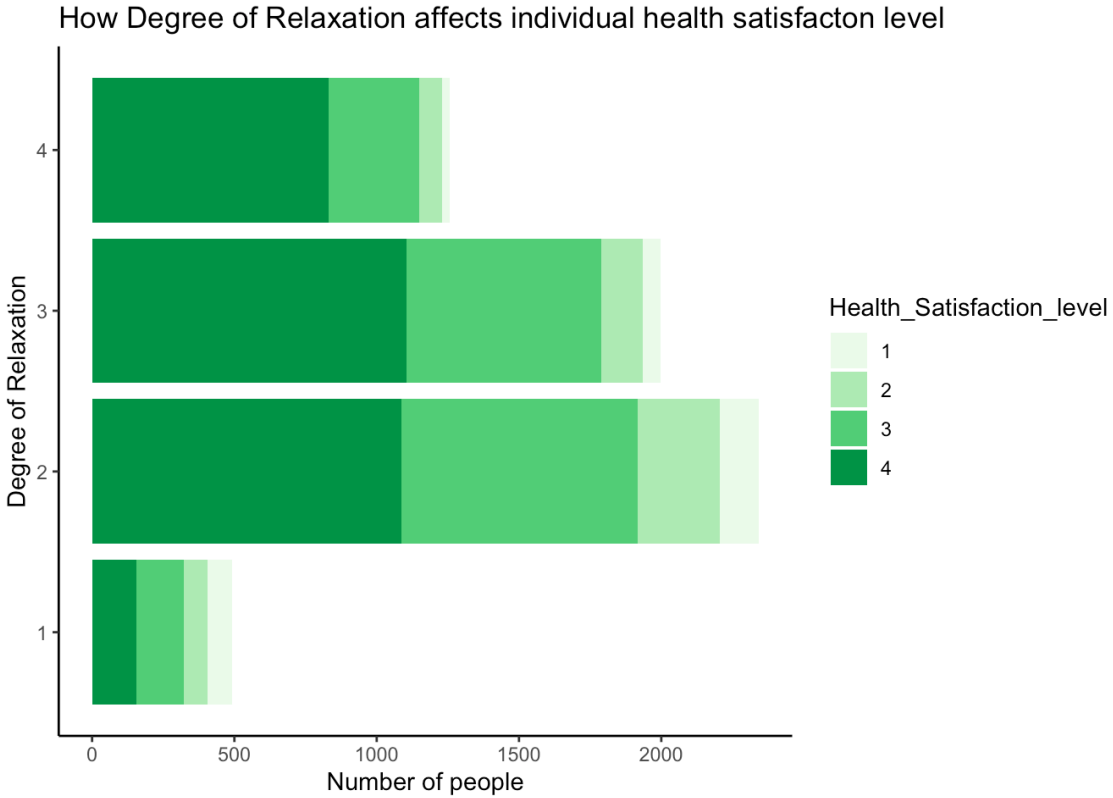

Figure 4

In Figure 4, the bar chart above shows responders' health satisfaction levels at different degrees of relaxation. The chart displays that the most people choose the degree of relaxation at 2 and the least responders choose the relaxation degree at 1. It is clear that about an equal number of responders from relaxation degree1 group and relaxation degree2 group are very satisfied with their health, though group2 have more responders. Meanwhile, responders who are very dissatisfied with their health tend to have relaxation degrees of 2. It is interesting that the responders with highest relaxation degrees are less likely to show "very dissatisfied" towards their health condition.

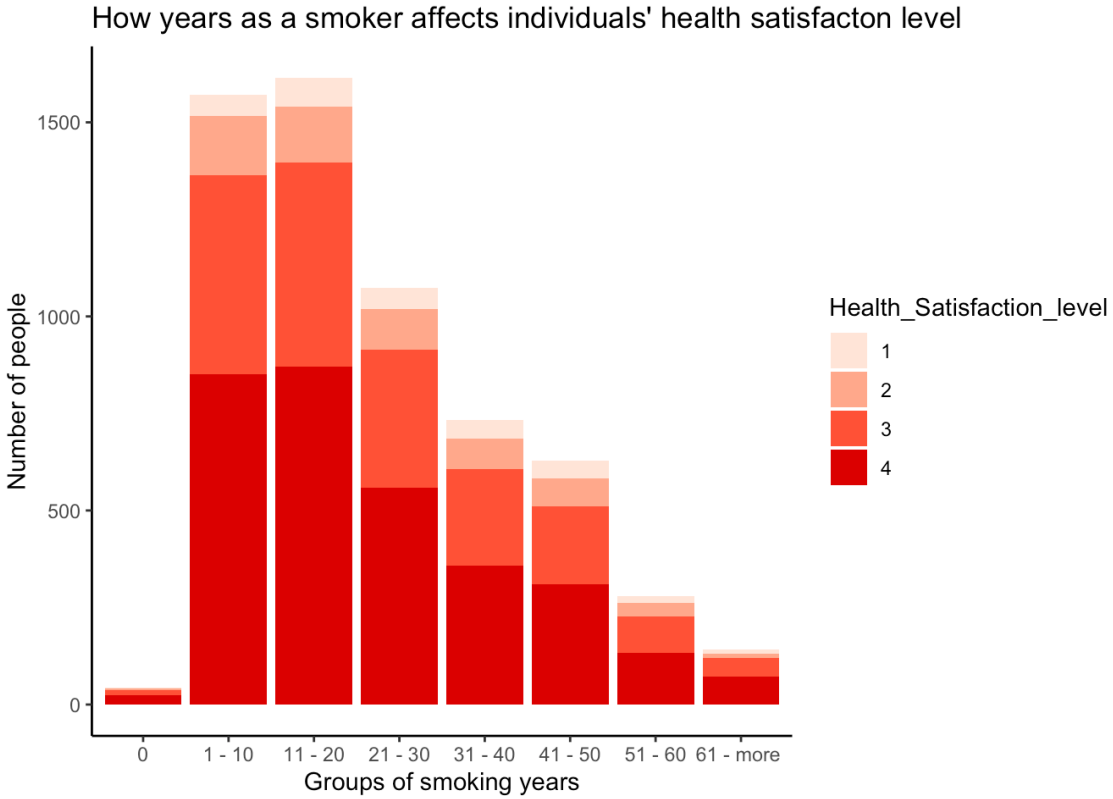

Figure 5 In Figure 5, a bar chart indicates the health satisfaction level of people who have different smoking years. Responders are divided into eight groups based on their smoking years. Group 1 represents citizens who don't smoke. Group2 represents responders who smoke within 10 years. In this order, group 8 represents participants who smoke longer than 60 years. Among all the responders surveyed , it is clearly to display that most people smoke between 10 to 20 years while only a few responders don't smoke. From the graph, about an equal number of participants from group 2 and 3 show "very satisfied" towards their health condition. While most responders who are very unsatisfied with their health are from group 3.

Discussion

It is interesting that male responders are more likely to be satisfied with their health conditions than female responders.It is obvious that an increasing trend of mean health satisfaction levels is shown when sleeping hours grow from 3 to 7 hours. However, when sleeping hours increase over 12 hours, the mean of health satisfaction levels decreases dramatically to level 1. Among all the responders surveyed , it is clearly to display that most people smoke between 10 to 20 years while only a few responders don't smoke. Meanwhile, responders who are very dissatisfied with their health tend to have relaxation degrees of 2. It is interesting that the responders with highest relaxation degrees are less likely to show "very dissatisfied" towards their health condition.

By analysing the results, smoking generates an adverse effect on personal health and citizens with longer sleeping time or relatively lower anxiety levels are usually satisfied with their current health conditions. Since all of them have a p-value less than 0.05 in regression model, they are statistically significant and reject the null hypothesis of equaling to zero. Although p values for years of smoking and flu shot have relatively large,

we still have weak evidence to reject null hypothesis. This means stress levels, sleeping time, taking flu shots and years of smoking could influence individual health satisfaction levels to different degrees.

Individuals' health satisfaction levels demonstrate citizens' current health states and their living conditions. By investing in factors that could affect persons' health satisfaction levels, it allows the society and government to draw more attention to health services and gather information on Canadians' health conditions. However, the dataset which was collected in 1991 was too old and thus it might fail to reflect the health conditions of Canadians in 2020. Since the survey was taken by telephone, there are many unresponded questions and this largely affects the accuracy of our dataset. As a consequence, our outcome will be less representative for responders who did not answer the phone. Aside from that, we cannot ignore the fact that respondents will not always provide accurate and honest answers during the survey. While, our data strongly relies on respondents' answers , so we can only suppose that they provide honest answers. However, we do find a few unreasonable answers that will affect our result in the study.

Nonetheless, the weak part of this survey is lack of the numerical data. Going through the whole questionnaire, we could find out that 90 percent of questions are multiple choice with provided answers, and only few of them are open questions. Thus, this directly causes more categorical variables and less numerical variables for the dataset, which creates difficulties on data analyzation. In addition, the questionnaire can be done by Proxy if the participant is ill or have language barriers. The replacement of taking surveys by Proxy may influence the choice of the real participant and cause uncertain responses.

Reference

• Canada 1991 to 2031." Journal of Rheumatology, 25 (1), 1998: 138-144.

• Basavaraj, S. "Smoking and Loss of Longevity in Canada." Canadian Journal of Public Health, Vol. 84, No. 5, 1993: 341-345.

• Boyle, M. H., W. Furlong, D. Feeny, G.W. Torrance and J. Hatcher. "Reliability of the Health Utilities Index - Mark III Used in the 1991 Cycle 6 Canadian General Social Survey Health Questionnaire." Quality of Life Research, Vol. 4, No. 3, 1995: 249-257.

• CBC/Radio-Canada (2019).How do the main parties compare on these issues? https://newsinteractives (https://newsinteractives). cbc.ca/elections/federal/2019/party-platforms/ • Education in Canada (2019).Council of Ministers of Education. https://www.canada.ca /en/immigration- (https://www.canada.ca/en/immigration-) refugees-citizenship/services/new-immigrants/new-life-canada/enrol-school.html • Erin Duffin (2020).Education in Canada - Statistics & Facts. https://www.statista.com/topics/2863/ (https://www.statista.com/topics/2863/) education-in-canada/#dossierSummary__chapter1 • Google services (2020). Google Account.https://www.google.com/drive/ • Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales (https://CRAN.R-project.org/package=scales) • Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. https://devtools.r-lib.org/,https://github.com/r-lib/devtools (https://devtools.r-lib.org/,https://github.com/r-lib/devtools). • H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.https://ggplot2.tidyverse.org • JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com (https://rmarkdown.rstudio.com). • Kerri Breen (2019). The 'genderation' gap: political divisions exist between men, women, different age groups,

polls show.https://globalnews.ca/news/5988160/genderation-gap-political-divisions-men- women/ • R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (https://www.R-project.org/). • Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian Election Study - Online Survey', https://doi.org/10.7910/DVN/DUS88V (https://doi.org/10.7910/DVN/DUS88V), Harvard Dataverse, V1 • Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686) • Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26. https://yihui.name/knitr/ (https://yihui.name/knitr/) • Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963. https://yihui.name/knitr/ (https://yihui.name/knitr/) • Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595. https://yihui.name/knitr/ (https://yihui.name/knitr/) • Yihui Xie and Christophe Dervieux and Emily Riederer (2020). R Markdown Cookbook. Chapman and Hall/CRC. ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook (https://bookdown.org/yihui/rmarkdown-cookbook). • Faculty of Arts & Science, University of Toronto. http://www.chass.utoronto.ca (http://www.chass.utoronto.ca) • Canadian general social surveys (GSS).https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm •Strike, Carol Overview of the 1991 General social survey on health (GSS-6). [Ottawa, ON]: Statistics Canada, July 8, 1991 (General social survey; working paper no. 4) (RA 407.5 C2 S87 1991 c.1 DATA) • Canada. Statistics Canada. General Social Survey. General social survey: bibliography. [Ottawa: Statistics Canada, n.d.] • Papers using the Canadian General social survey data – GSS. By Jong Won Min Calgary, AB: University of Calgary. Faculty of Social Work, n.d.