Logistic Regression and Propensity Score Matching Analysis of Canadians and Flu Shots from the 2017-2018 Canadian Commmunity Health Survey

Linhan Zhang 1004734353

14/12/2020

Abstract

The goal of this study was to evaluate the factors which affected Canadian influenza vaccination rates and determine whether education is a factor which causes increased/decreased vaccination. The methods include a logistic regression model evaluating multiple factors and their association with whether an individual received a flu shot within the past year and a propensity score matching analysis of the observational data. First, the data for both analysis were retrieved as subsets of the CCHS; cleaned by removing NAs, renaming variables, and redefining age and vaccination categories; and finally the logistic regression model data was split into training (75%) and testing (25%). Subsequently, a logistic model was created using forward selection. The final model included variables sex, marital status, age, education, and access to health care. Diagnostics through internal validation revealed a mean square error of 0.003 and an AUC of 0.6855 as well as a prediction error of 1.207982 through the testing data. The PSM revealed that education is a significant predictor of whether someone has had their flu shot within the last year. This suggests first and foremost that improvements can be made in public health education. Through this analysis, health experts will be able to target those less likely to get their flu shot either through education, outreach, or additional incentives; subsequently, these measures would be crucial in maintaining public health and ensuring high quality of life in Canada.

(Github link:https://github.com/Linlinzhang28/STA304_Final_Project.git)

Keywords: logistic regression, propensity score matching, flu shots, Canada, Canadian Community Health Survey, Education, population health, health, influenza, vaccination

Introduction

The influenza is a common, epidemic, respiration illness which is caused by a virus (Dickin et al., 2017). Each year, there are estimated more than 12, 200 hospitalization cases and 3500 deaths caused by the influenza in Canadian communities which puts elderly people and those with chronic conditions at great risk (Dickin et al., 2017). The most effective way of preventing the spread of the influenza is through getting the vaccination before each flu season. Indeed, the government shows that vaccination can prevent up to 60% of the population from getting the flu (Ministry of Health, 2020). However, data from the Canadian Community Health Survey 2017-2018 (CCHS) illustrates that only 56.6% of the people within the last year have gotten their shot and 32.5% of the people have not gotten any flu shots for more than 2 years (Health Statistics Division, 2017). Studies conducted in the United States have found that factors such as race, gender, side effect concerns, lack of knowledge of the ingredients in a flu vaccine all exert significant effects on influenza vaccination rate (Zimmerman et al., 2003).

This study aims to find out what factors affect the Canadian influenza vaccination rate by using a logistic regression model. Accordingly, variables are selected through the CCHS from 2017-2018 (Health Statistics Division, 2017). This model will specifically examine the factors which are associated with the likelihood of not having any flu shots within the past year. Further, a propensity score matching analysis will also be conducted to explore any possible causal relationships between variables of interest and flu shot data (Caetano, 2020a). This tool, popularized since the 1980s, allows causal inferences to be made from observational data under specific circumstances. That is, by matching observations based on a variety of covariates, differences between treatment and outcome groups are essentially adjusted for, thereby creating a pseudo-randomized sample from observational data for analysis. In this analysis, the outcome variable of interest will also be whether the individual had any flu shots in the past year, however, the main treatment variable of interest is education. That is, it is hypothesized that non-belief in the benefits or effectiveness of vaccines is what causes most to forego the shot. Overall, this study will point to factors of focus that need to be addressed to improve Canadian vaccination rates. Also, by identifying factors associated with decreased vaccination, Health Canada will also be able to implement improvements to their current policies and vaccination outreach and practices.

Two data sets will be employed throughout the analysis, though both are subsets of the 2017-2018 CCHS (Health Statistics Division, 2017). The first data set will be used in the main logistic regression analysis of factors associated with lack of flu shots within the past year, while the second data set will be used in the propensity score matching analysis. This report will overview the methodologies by highlighting the important aspects of the CCHS and its data and the process of creating the model. Then the results of the study will be presented and a discussion on the meaning of the results will follow. Finally, the limitations and improvements as well as possible future steps will be explored and reasoned. An appendix is included at the end and will provide supplemental details, plots, graphs, and calculations relevant to the analysis.

Methodology

Data

The two data sets for the logistic regression analysis and PSM of this study were retrieved as subsets from the 2017-2018 CCHS and contained sex, age, education, marital status, last time had flu shot, access to health care, health insurance coverage, and income (Health Statistics Division, 2017). The Canadian Community Health Survey is a cross sectional survey which attempts to collect a sample representative of the Canadian population (Health Statistics Division, 2017). To achieve this, they weigh populations by their respective size in Canada and sample accordingly. Generally, the CCHS aims to collect 130,000 responses of which 10,000 from teens aged between 12 and 17 years. While the target population is all Canadians, the sampling frame excludes full-time members of the Canadian Armed Forces, institutionalized people, children in foster homes, and people living in special Quebec health regions; though altogether, these populations represent less than 3% of Canada's population (Health Statistics Division, 2017). Overall, the major strengths include the large sample size, which are well weighted to represent the Canadian population, and the large number of variables measured; however, major weaknesses lie in their low response rate which can skew their population weights. Indeed, no methodology regarding how non-response was treated was provided; this may affect how well the final data sample represents Canada.

The variables selected for the data were hypothesized to be associated with variability in likelihood of people getting their flu shots. The data set used for the main analysis and construction of the logistic regression initially comprised of 113,290 observations of 12 variables. These main variables from both raw data subsets were plotted and visualized in Appendix 1. Most trends in the raw data do not appear unusual; perhaps the most notable is the fact that a large majority of the respondents reported having no difficulty accessing health care. Nonetheless, all these variables will undergo cleaning and will be part of the analysis. However, during the data cleaning process, observations were removed for NA responses and 3 variables were removed because they were not analyzable in the logistic regression. In fact, those variables were answers to secondary questions regarding flu shots and were therefore not relevant. After cleaning there were 63,980 observations of 8 variables. Finally, within the predictor variable, ‘last time had flu shot’, the response category ‘within last 1 year’ was codified to the value ‘1’ while all other levels of response (‘more than 1 year but less than 2 years’ and ‘more than 2 years ago’) were codified to the value ‘0’ for the logistic regression. This was done to simplify the data and make the response variable more precise since the point of interest is what currently affects flu shot rates. Table 1 below presents the baseline characteristics of the subjects from the first data set. Moreover, this first data set was also separated into training and testing sections by 0.75 and 0.25 proportions respectively.

Table 1: Baseline Characteristics Table LRM

	1	0	p	test
n	38601	25379		
Sex = Female (%)	22735 (58.9)	13647 (53.8)	<0.001	
Marital.status (%)			<0.001	
Married	18158 (47.0)	9838 (38.8)		
Common-law	2723 (7.1)	2915 (11.5)		
Widowed/Divorced/Separated	10166 (26.3)	4508 (17.8)		
Single	7554 (19.6)	8118 (32.0)		
Age (%)			<0.001	
Adolescence	2261 (5.9)	2535 (10.0)		
Young Adult	5674 (14.7)	7992 (31.5)		
Middle-aged Adult	8562 (22.2)	8393 (33.1)		
Old Adult	22104 (57.3)	6459 (25.5)		
Education (%)			<0.001	
Less than secondary school graduation	8134 (21.1)	4571 (18.0)		
Secondary school graduation, no post-secondary education	7983 (20.7)	5727 (22.6)		
Post-secondary certificate diploma or univ degree	22484 (58.2)	15081 (59.4)		
Last.time.had.flu.shot = 0 (%)	0 (0.0)	25379 (100.0)	<0.001	
Access.to.health.care = No (%)	36206 (93.8)	24057 (94.8)	<0.001	
Health.insurance.coverage = No (%)	36902 (95.6)	24203 (95.4)	0.171	
Income (%)			<0.001	
No income or less than $20,000	3149 (8.2)	2135 (8.4)		
$20,000 to $39,999	7189 (18.6)	3678 (14.5)		
$40,000 to $59,999	6302 (16.3)	3706 (14.6)		
$60,000 to $79,999	5131 (13.3)	3216 (12.7)		
$80,000 or more	16830 (43.6)	12644 (49.8)		

Subsequently, the data used for propensity score matching originally contained 113,290 observations of 9 variables. After cleaning out the NA responses, there were 61,780 observations remaining. This time however, the response variable of interest (outcome) was ‘does not believe benefits’. The ‘does not believe benefits’ is a variable derived from those who responded ‘no’ to having taken the current flu shots. In this category not believing in the benefits was codified as 0 while the opposite as 1. The baseline characteristics of the subjects in this data set are presented in Table 2 below.

Table 2: Baseline Characteristics Table PSM

	0	1	p	test
n	23908	37872		
Province (%)			<0.001	
NEWFOUNDLAND AND LABRADOR	859 (3.6)	1006 (2.7)		
PRINCE EDWARD ISLAND	388 (1.6)	538 (1.4)		
NOVA SCOTIA	805 (3.4)	1254 (3.3)		
NEW BRUNSWICK	803 (3.4)	1036 (2.7)		

	0	1	p	test
QUEBEC	5490 (23.0)	9972 (26.3)		
ONTARIO	6394 (26.7)	10815 (28.6)		
MANITOBA	1456 (6.1)	1653 (4.4)		
SASKATCHEWAN	1192 (5.0)	1352 (3.6)		
ALBERTA	2883 (12.1)	4325 (11.4)		
BRITISH COLUMBIA	2957 (12.4)	5139 (13.6)		
YUKON	180 (0.8)	278 (0.7)		
NORTHWEST TERRITORIES	224 (0.9)	331 (0.9)		
NUNAVUT	277 (1.2)	173 (0.5)		
Sex = Female (%)	12019 (50.3)	19979 (52.8)	<0.001	
Marital.status (%)			<0.001	
Married	8520 (35.6)	16539 (43.7)		
Common-law	2808 (11.7)	5375 (14.2)		
Widowed/Divorced/Separated	5229 (21.9)	6519 (17.2)		
Single	7351 (30.7)	9439 (24.9)		
Family.arrangement (%)			<0.001	
Unattached individual living alone.	7576 (31.7)	10942 (28.9)		
Unattached individual living with others.	1051 (4.4)	1292 (3.4)		
Individual living with spouse/partner.	6560 (27.4)	10173 (26.9)		
Parent living with spouse/partner and child(ren).	3776 (15.8)	9980 (26.4)		
Single parent living with children.	1441 (6.0)	2483 (6.6)		
Child living with a single parent with or without siblings.	818 (3.4)	513 (1.4)		
Child living with two parents with or without siblings	1577 (6.6)	1166 (3.1)		
Other	1109 (4.6)	1323 (3.5)		
Age (%)			<0.001	
Adolescence	1220 (5.1)	112 (0.3)		
Young Adult	6751 (28.2)	13924 (36.8)		
Middle-aged Adult	7596 (31.8)	14945 (39.5)		
Old Adult	8341 (34.9)	8891 (23.5)		
Type.of.drinker (%)			<0.001	
Regular drinker	13831 (57.9)	26512 (70.0)		
Occasional drinker	4592 (19.2)	5752 (15.2)		
Did not drink in the last 12 months	5485 (22.9)	5608 (14.8)		
Does.not.believe.benefits = No (%)	20091 (84.0)	30883 (81.5)	<0.001	
Income (%)			<0.001	
No income or income loss	550 (2.3)	471 (1.2)		
Less than $20,000	9070 (37.9)	7234 (19.1)		
$20,000 to $39,999	7759 (32.5)	11060 (29.2)		
$40,000 to $59,999	3475 (14.5)	7395 (19.5)		
$60,000 to $79,999	1510 (6.3)	4860 (12.8)		
$80,000 or more	1544 (6.5)	6852 (18.1)		

Finally, the age category in both data sets were widened to 4 levels: 12-19 years (Adolescence), 20-39 years (Young Adult), 40-59 years (Middle-aged Adult), and 60 and older (Old Adult).

Models

Logistic Regression Model

The logistic regression model was built and run in R. A binary logistic regression model was chosen because the variables were categorical, meaning that other models like linear regression would be meaningless if applied to the data (Caetano, 2020b). Since the response variable was simplified to ‘0’ or ‘1’ levels, the model is binary representing two levels of the response variable. Specifically, this model was built using forward selection. The first step included using Chi-square tests to evaluate the first and foremost significant variable correlated with flu shots within the past year (Wiley & Wiley, 2019). As detailed in Appendix 2, the variable sex was significant and was added to the model. Then other variables of interest were added step-wise, first starting with marital status; then using an ANOVA test the significance of marital status in the model was determined (Appendix 3). In this case, a conventional significance level of 0.05 was used. This process was repeated for the other variables (age, education, access to health care, health insurance coverage, and income). The results of each step’s ANOVA test are detail in Appendix 3 (Wiley & Wiley, 2019). The final model included variables age, sex, education, marital status, and access to health care. The formula for the final model is:

$$log(\frac{Pr(Last.\widehat{time.have}.flu.shot = 1)}{Pr(Last.time.have.flu.shot = 0})=$$

$$-0.316 - 0.204 * Marital.status_{Common-law} + 0.197 * Marital.status_{Widowed/Divorced/Separated}$$

$$+0.260 * Marital.status_{Single} + 0.284 * Age_{Young.adult} - 0.008 * Age_{Middle-aged.adult}$$

$$-1.208 * Age_{Old.adult} + 0.198 * Education_{Secondary.school.graduation.no.post-secondary.education}$$

$$0.052 * Education_{Post-secondary.certificate.diploma.or.univ.degree} + 0.239 * Access.to.health.care_{no.difficulties}$$

The formula illustrates each predictor’s coefficient (Appendix 4). A coefficient details the magnitude of the change in the likelihood of the response variable when all other factors remain equal. For example, OldAdult has a coefficient of -1.206729 compared to Young Adult with a coefficient of 0.284098. What these mean is that, when all other factors are equal, adults in the old category have higher log odds of not believing in the benefits of the flu shot as negative coefficients bring the overall value closer to ‘0’.

The model was first diagnosed using cross-validation a technique used to test the model by using different samples of data (Stefan, 2020). The training set contained 75% of the observations while the testing set contained the remaining 25% of the observations. The cross-validation checks if the model predicts well for training data set. Essentially, this tool validates the accuracy of the model’s predictions. The calibration plot (Appendix 5) shows that the bias-corrected line closely follows the ideal line which suggests that the model performs strongly to predict responses internally. Then, an area under the curve (AUC) also aids in assessing the performance of the model. This specific tool is used because it is created for instances where predictors are ordinal, a type of categorical data (Agresti, 2014, Wiley & Wiley, 2019). The AUC of the model, in Appendix 6, is 0.6855. Finally, a prediction error was calculated using the testing data to be 1.207982 (Appendix 7).

Propensity Score Regression (PSR)

PSR unlike the logistic regression analysis looks to evaluate causality between treatment and outcome (Caetano, 2020a). While the earlier model evaluated the associations between several factors and decreased vaccination rates, this propensity score regression aims to establish whether our treatment, in this case education, is a significant predictor of refusal to vaccinate due to the belief that there are no actual benefits to the flu shot. The code from this analysis is retrieved partially from Alexander. Essentially, PSR adjusts for differences between two groups and aims to highlight the true ‘treatment effect’. These two analyses together provide a foundation of evidence upon which government and health experts can develop strategies to increase vaccination rate and trust in vaccinations. This part of the analysis relied on the second data set in which education was the treatment and belief that vaccinations are not effective is the outcome. In the propensity score matching there were 47,816 matches generated. The final propensity model is:

$$\begin{gathered}
\widehat{y} = 2.018 - 0.37 * Province_{Prince.Edward.Island} - 0.063 * Province_{Nova.Scotia} - 0.0.086 * Province_{Brunswick} \\
-0.106 * Province_{Quebec} - 0.066 * Province_{Ontario} - 0.0.034 * Province_{Manitoba} \\
-0.030 * Province_{Saskachewan} - 0.069 * Province_{Alberta} - 0.047 * Province_{British.Columbia} \\
-0.068 * Province_{Yukon} - 0.054 * Province_{Northwest.Territories} \\
-0.003 * Province_{Nunavut} - 0.015 * Sex_{Female} 0.01 * Marital.status_{Common-law} \\
-0.013 * Marital.status_{Widowed/Divorced/Separated} - 0.01 * Marital.status_{Single} \\
+0.007 * Family.arrangement_{Unattached.individual.living.with.others} \\
-0.018 * Family.arrangement_{individual.living.with.spouse/partner.and.children} \\
-0.012 * Family.arrangement_{Single.parent.living.with.children} \\
+0.03 * Family.arrangement_{Children.living.with.a.single.parent.with.or.without.siblings} \\
-0.012 * Family.arrangement_{Other} - 0.025 * Age_{Young.adult} - 0.08 * Age_{Middle-aged.adult} \\
-0.098 * Age_{Old.adult} - 0.004 * Type.of.drinker_{Occasional.drinker} \\
+0.012 * Type.of.drinker_{did.not.drink.in.the.last.12.month} - 0.026 * Income_{less.than.20,000} \\
-0.028 * Income_{20,000.to.39,999} - 0.027 * Income_{40,000-59,000} \\
-0.031 * Income_{60,000-79,000} - 0.029 * Income_{80,000.and.more} \\
-0.019 * Education_{1}
\end{gathered}$$

The output from this model summarized will provide information regarding the significance of the predictors. Should education be a significant factor then, it would be suggested that lack of education can cause people to not believe in the benefits of vaccinations.

Results

Through the first part of the analysis, the table summary of the logistic regression model (Appendix 4) highlights interesting correlations to vaccination within the past year. The table shows the coefficients of each level of each categorical variable. Larger positive values indicate greater likelihood of having gotten a vaccine in the past year and larger negative numbers indicate greater chance of not having gotten any flu shot in the past year. First, most noticeably, the population of older aged adults (60 years and older) are significantly more likely to have not gotten their vaccination within the past year with a coefficient of -1.206729. Furthermore, all marital statuses are positively associated with having gotten the flu shot with the coefficients varying between 0.196808 and 0.376134; women are less likely to have gotten a shot with a coefficient of -0.203871; those with no difficulty accessing health care are more likely to have gotten a shot with a coefficient of 0.238615. Finally, between education levels, surprisingly those with secondary education only or lower have almost 4 times the likelihood of having gotten a vaccination within the past year compared to those with post-secondary education (coefficients are 0.198286 and 0.051684 respectively). Overall, most variables lean in one direction over another however the magnitudes are not very large. The most interesting associations are those regarding older age adults and education levels.

Finally, through the propensity score regression analysis, as illustrated in Appendix 8, education1 is a significant predictor (-0.19) of whether individuals believe in the benefits of vaccination. However, education1 refers to those with post-secondary education and the effect direction is negative meaning that the results suggest that higher education tends to cause individuals to not believe in the benefits of flu shots.

Discussion

Summary

This study was conducted on subsets of data from the 2017-2018 Canadian Community Health Survey (Health Statistics Division, 2017). For the first part, variables of interest for the logistic regression model predicting whether an individual had a flu shot within the past year included age, sex, access to health care, health insurance coverage, income, education, and marital status. The data cleaning process removed many NA responses, re-organized the time of last flu shot variable, and re-defined the age variable into four more general levels. For the final model though, only age, sex, marital status, education, and access to health care were included. Then a propensity score matching analysis and regression was conducted on the second subset of data. In this analysis, education was the treatment variable and the outcome variable was whether individuals believed in the benefits of flu shots. Finally, it was determined that education is a significant predictor of whether an individual believes in the benefits of flu shots; however, the 'treatment effect' was in the opposite hypothesized direction and the 'effect size' was not very large.

Conclusions

In the first part of the study, a logistic model predicting the likelihood of an individual having gotten a flu shot within the past year was created. The most notable variables and levels included older aged adults, education, and access to health care. That is while marital status and sex had significant results, the variation between those of different marital statuses did not differ drastically. Moreover, women had just slight higher likelihood of not having gotten a flu shot in the past year. Nonetheless, it was evident that older aged adults had drastically higher chances of not having been vaccinated within the last year. This is especially concerning given that the older population suffers more from chronic diseases and is generally more compromised when it comes to community flu outbreaks. This variability could be explained by the fact that elderly people may have diminished connection to their greater communities and therefore are not aware of vaccination campaigns; another aspect which may influence their vaccination rate is the fact that many may not have the physical ability or resources to travel and have their flu shots each year. Similarly, no difficulty with access to health care was associated with greater likelihood of receiving a flu shot. These results, in tandem, may suggest that improvements must be made in health care accessibility and community outreach for the elderly.

Finally, the education variable was surprising since lower education levels were associated with greater likelihood of having received a vaccination within the past year. While many often assume the higher educated must be more aware and have greater critical thinking when it comes to the science of vaccinations, this assumption appears to be false and misleading. A reason for this is that education is not correlated with higher vaccination rates since not everyone attends higher education institutions in immunology and sciences, therefore there is no reason for higher education to be associated with greater scientific understanding which would propel individuals to get vaccinated. Overall, education at all levels are associated with slightly greater vaccination rates; however, for public health officials, it is important to note that to increase the effectiveness of schooling, it may be important to implement more pro-vaccination courses or modules into all levels of education since the magnitude of the positive coefficients of education were still small. Nonetheless, based on the first part of this study, it becomes apparent that improvements can be made in Canada to improve vaccination rates; indeed, starting with increasing community outreach and health care accessibility is crucial in taking care of Canada's aging population. Then, it would be important to focus on addressing the limitations of the current education system regarding public health practices like getting the flu shot.

Subsequently, in the propensity score regression, it was determined that education is a significant predictor of whether an individual believes in the benefits of the flu shot; however, similarly to the first part of the study, it was determined that higher education causes individuals to not believe in the benefits of the vaccine. These results, though surprising, again highlight a new aspect to consider. Indeed, those with higher education should not, in society, be granted 'a pass' and assumed to be conforming to proper public health practices just because they have a degree in higher education. Nonetheless, it is not important to dwell on the minor differences in the vaccination rates of those with different aquired education. What is more relevant is an evaluation of the effectiveness of education in informing Canadians on the importance of getting their flu shot. Indeed, in both the analyses the magnitude of the effects of education on vaccination are not very large. This should be an area of focus for Health Canada. While Canadians undergo their general education and highschool, they should be specifically informed on proper practices like getting the flu shot. The ideal observation in the Canadian population is that all those who have gotten their education in Canada have high vaccination rates. Of course there would still be other factors like those previously mentioned concerning accessibility and outreach; however, education is a low cost and effective area to start.

Limitations and Next Steps

First, though the CCHs is a very comprehensive survey, asking a wide variety of questions and presenting over 1000 variables, many of these variables do not have enough responses. Indeed, some variables though very interesting and relevant to this study like the frequency of doctor visits. A variable such as this one could be strongly associated with Canadian vaccination rates and could have provided important insight alongside access to health care variable; also, information could have given more context regarding where the government needs to improve in order to increase vaccination rates. Without many of these more precise measurements, this study is only able to draw general conclusions and inferences. Another limitation of this study is the somewhat arbitrary definition of the levels of certain categories like the age brackets. While the survey had responses from people as young as 12 and as old as 80+, this analysis used 4 age brackets to cover everyone. This reduces the meaning of many associations and conclusions because of how broad the categories are. However, this was necessary in order to evaluate categorical variables. Finally, in the propensity score matching analysis, while there were over 40,000 matches, the number of covariates that were adjusted for, again, was relatively small. This, though due to the reduced number of variables with sufficient observations is also due to a fundamental weakness of propensity score matching. Indeed, using too many variables will reduce the effectiveness of the matching. However, too few also weakens the strength of the results.

Overall, this study incorporated many methods into its analysis, though for the future many improvements could be made. First, given the size of the sample, more variables could be considered for the model to make associations. This would also rely on improved responses from the survey. It is suggested that questions be made clearer or the survey be conducted in modules in order to increase the response rate to each question (e.g. it may be easier for a person to respond to all the health care questions if they were asked together and not rushed). Of course, in the future, a more meaningful propensity score matching analysis is due and this should be accomplished by including more covariates into the matching process, other important variables to consider could include more specific education levels, cultural background, religious beliefs, health conditions (which may limit vaccine accessibility like allergies). Ultimately, Health Canada should understand the results of this current study and begin forming and implementing changes in order to encourage Canadians to get their flu shot.

References

Agresti, A. (2014). Categorical Data Analysis. Hoboken: Wiley.

Caetano, S. J. (2020a). STA304 Propensity Score Matching. [Handout]. Retrieved from: https://q.utoronto.ca/courses/184060/files/9632737?module_item_id=1891982

Caetano, S. J. (2020b). STA304 Logistic Regression Intro. [Handout]. Retrieved from: https://q.utoronto.ca/courses/184060/files/9632737?module_item_id=1891982

Dickin, J.„ & Bailey, P.„ & James-Abra, E., Influenza (Flu) in Canada (2017). In The Canadian Encyclopedia. Retrieved from: https://www.thecanadianencyclopedia.ca/en/article/influenza

Health Statistics Division. (2017). Canadian community health survey, 2017-2018: Annual component Health Statistics Division, Statistics Canada. Retrieved from http://odesi.ca/#/details?uri=/odesi/cchs-82M0013-E-2017-2018-Annual- component.xml

Ministry of Health. (2020). Flu shot safety and effectiveness. Retrieved from: https://www.ontario.ca/page/flu-vaccine-safety-effectiveness

Alexander, R. (11 November 2020). Running Through a Propensity Score Matching Example. Retrieved from: https://q.utoronto.ca/courses/184060/files/10649324?module_item_id=1998838

Stefan, G. (2020). STA303/STA1002: Methods of Data Analysis II Lecture 6 [PowerPoint slides]. Retrieved from: https://q.utoronto.ca/courses/159686/modules

Zimmerman, R. K., Santibanez, T. A., Janosky, J. E., Fine, M. J., Raymund, M., Wilson, S.A.,… Nowalk, M. P. (2003). What affects influenza vaccination rates among older patients? an analysis from inner-city, suburban, rural, and veterans affairs practices. The American Journal of Medicine, 114(1), 31-38. doi:10.1016/S0002-9343(02)01421-3

Appendix

Appendix 1

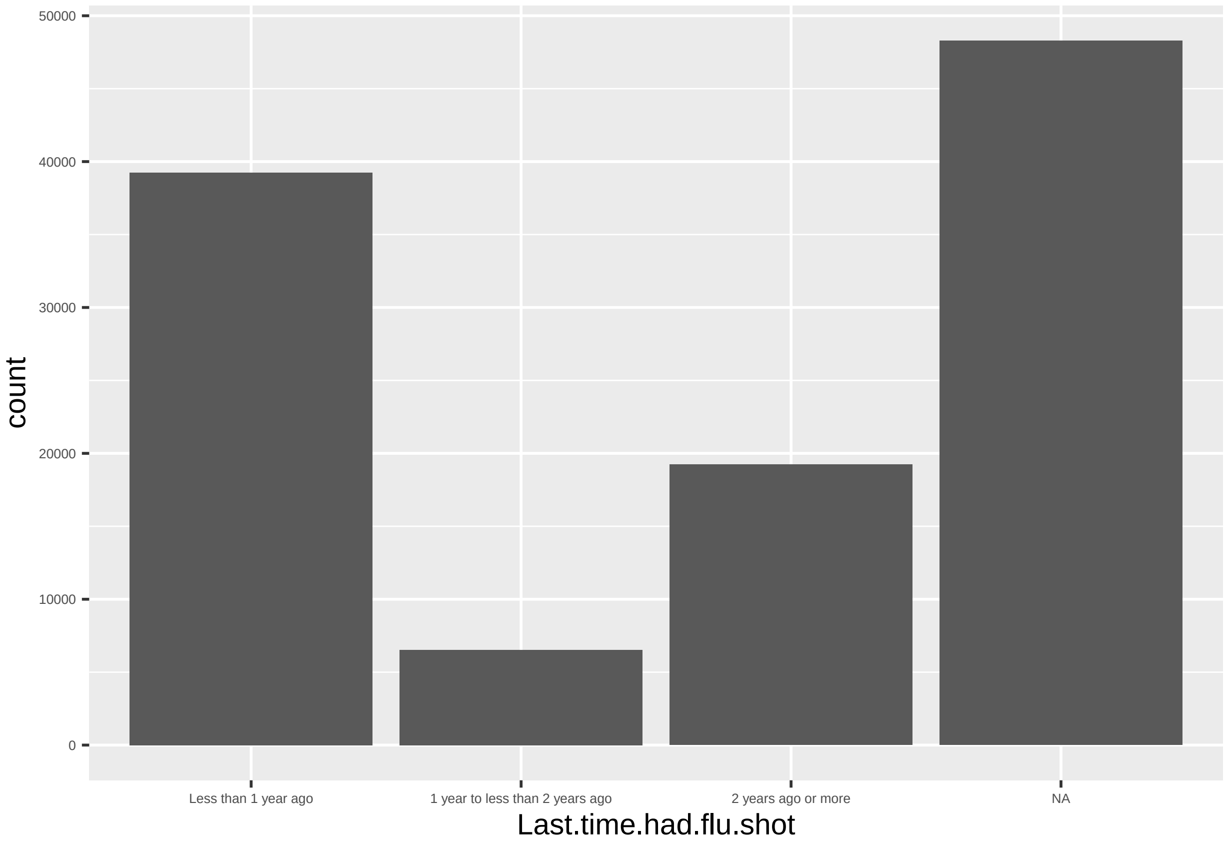


Figure 1: Distribution for Last.time.had.flu.shot

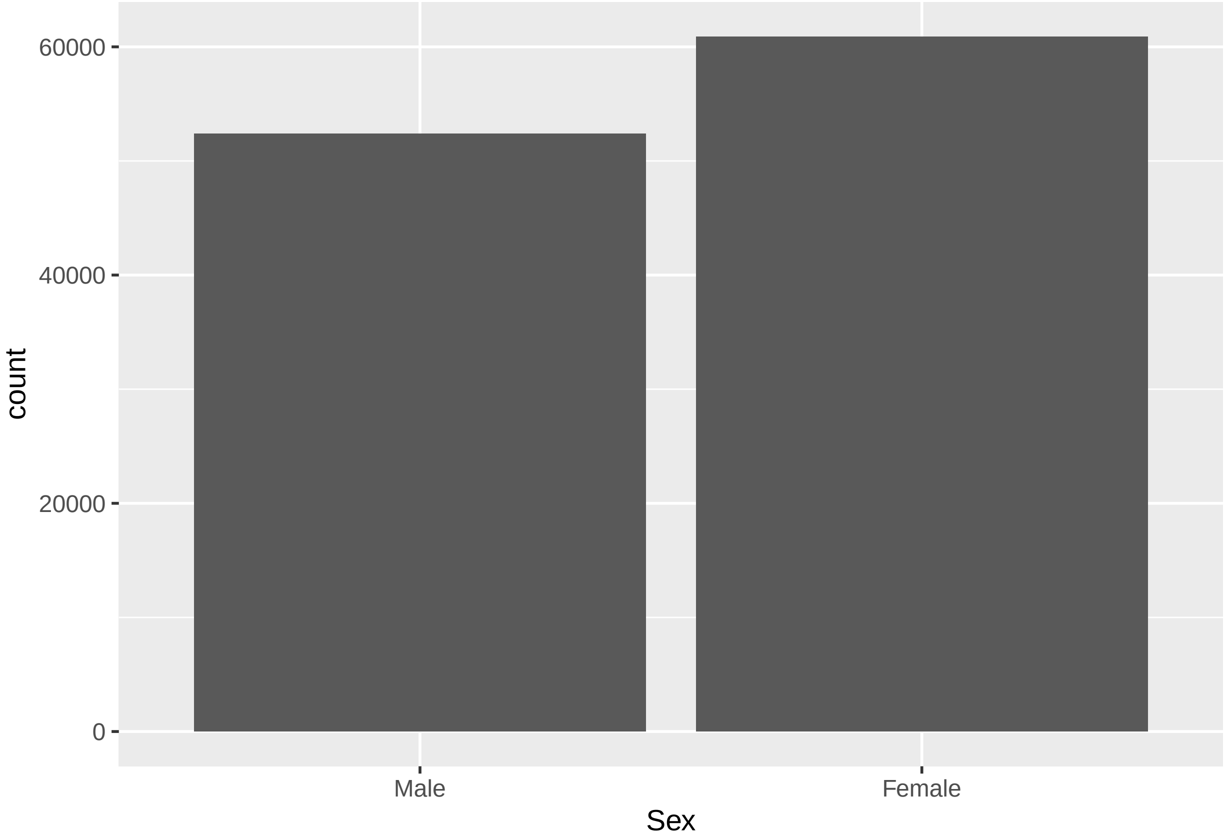


Figure 2: Distribution for Sex

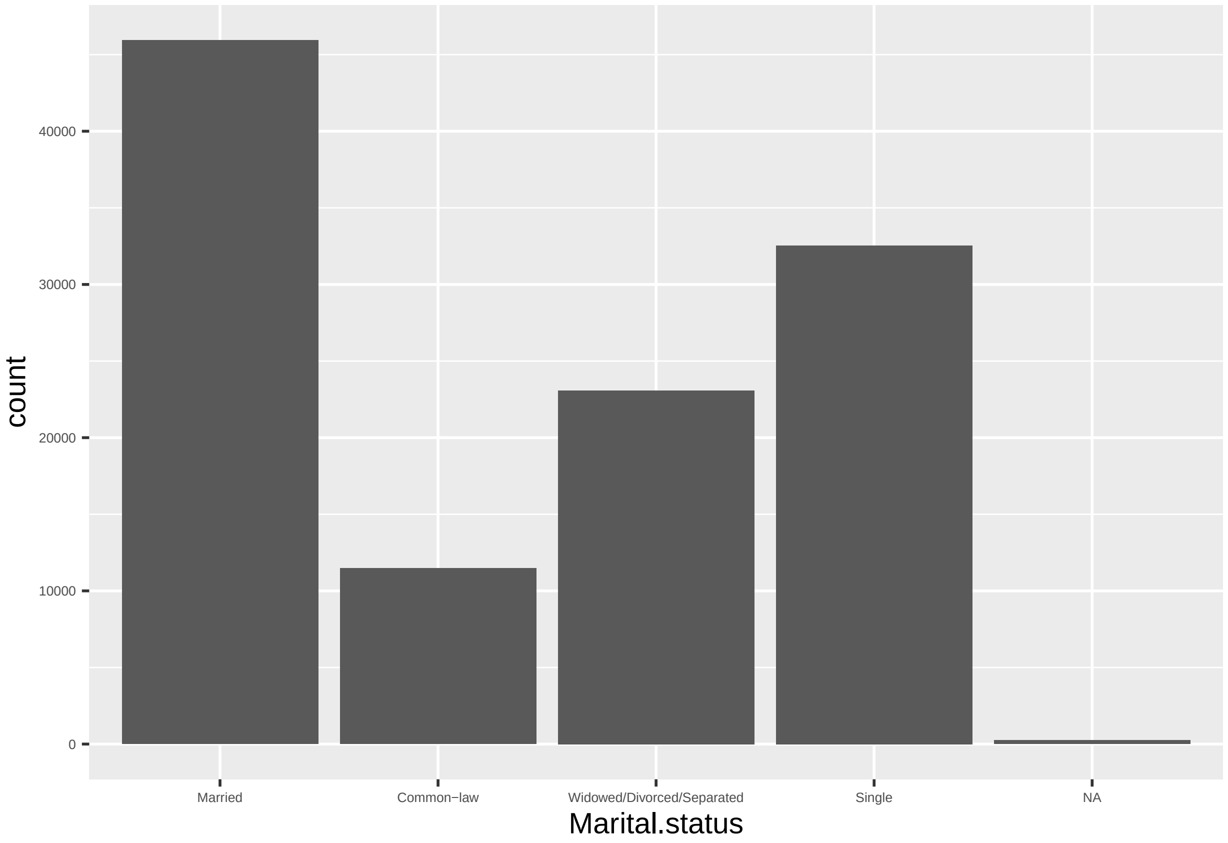


Figure 3: Distribution for Marital.Status

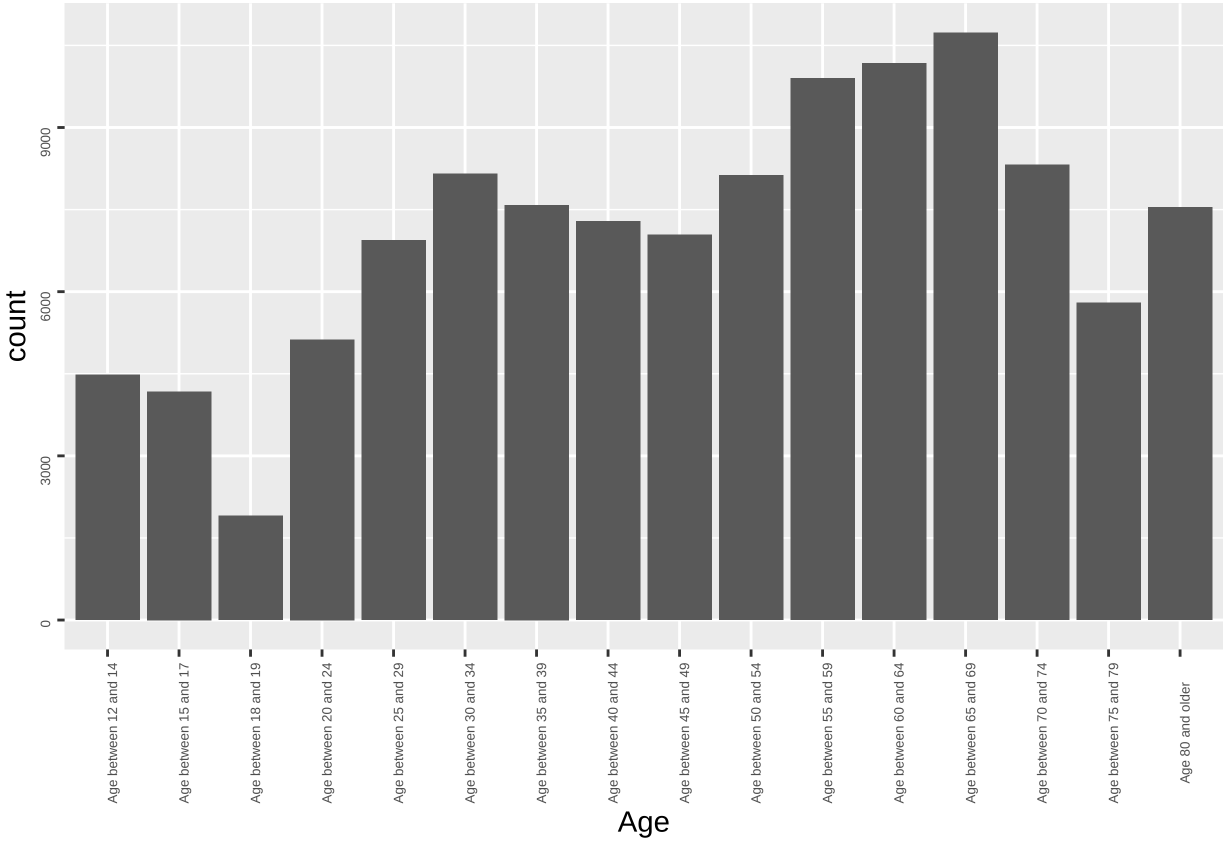


Figure 4: Distribution for Age

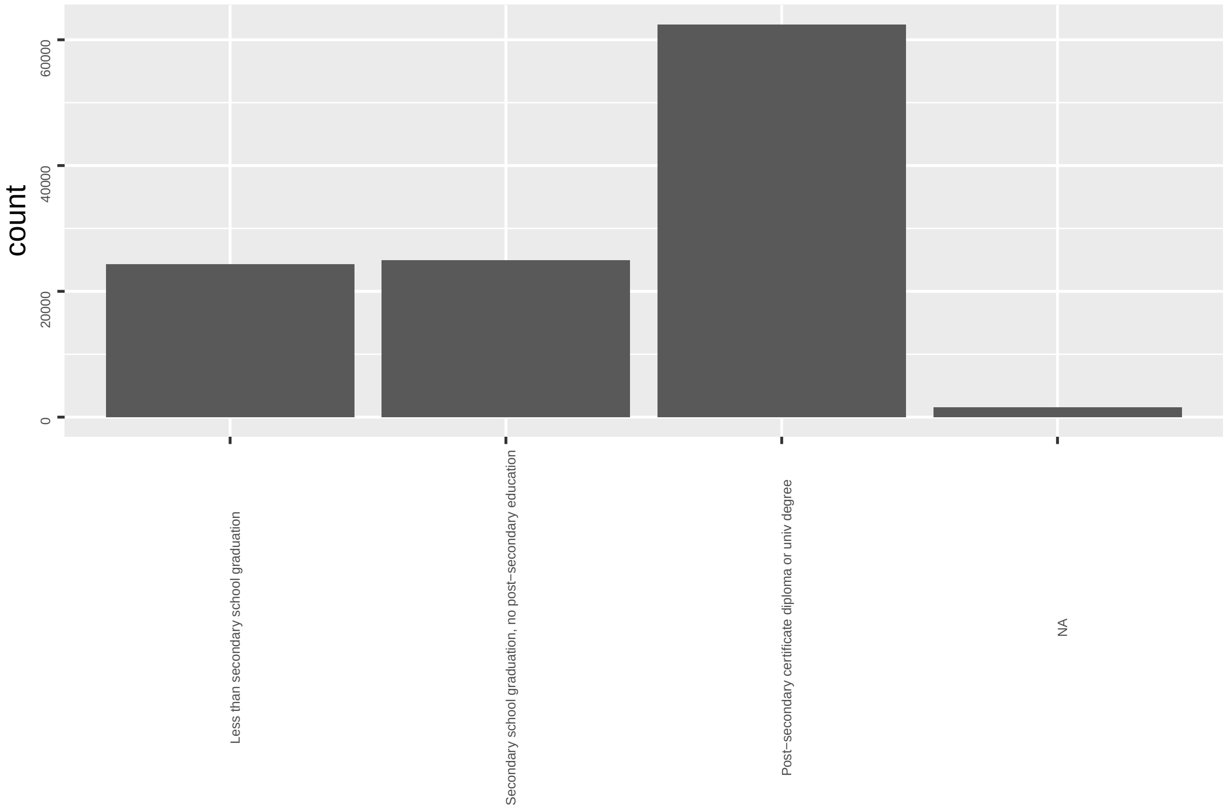


Figure 5: Distribution for Education

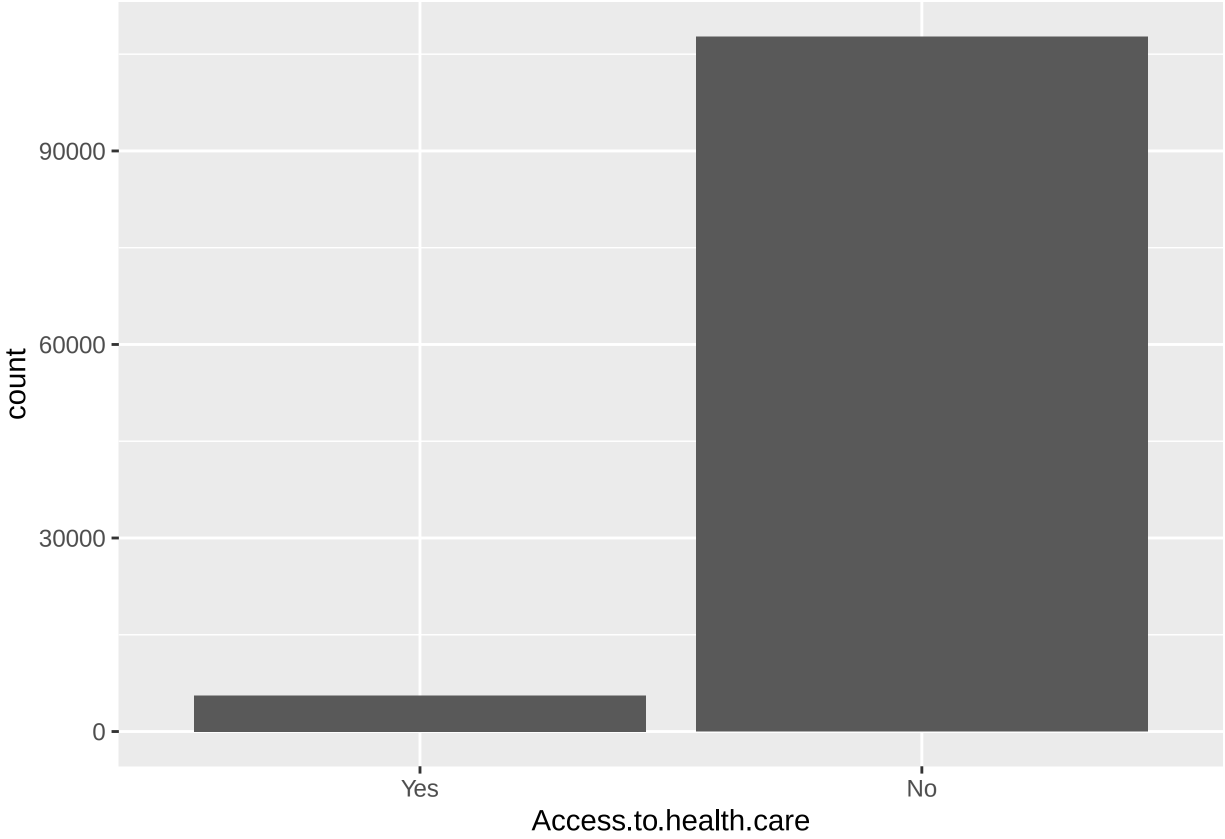


Figure 6: Distribution for Access to Health Care

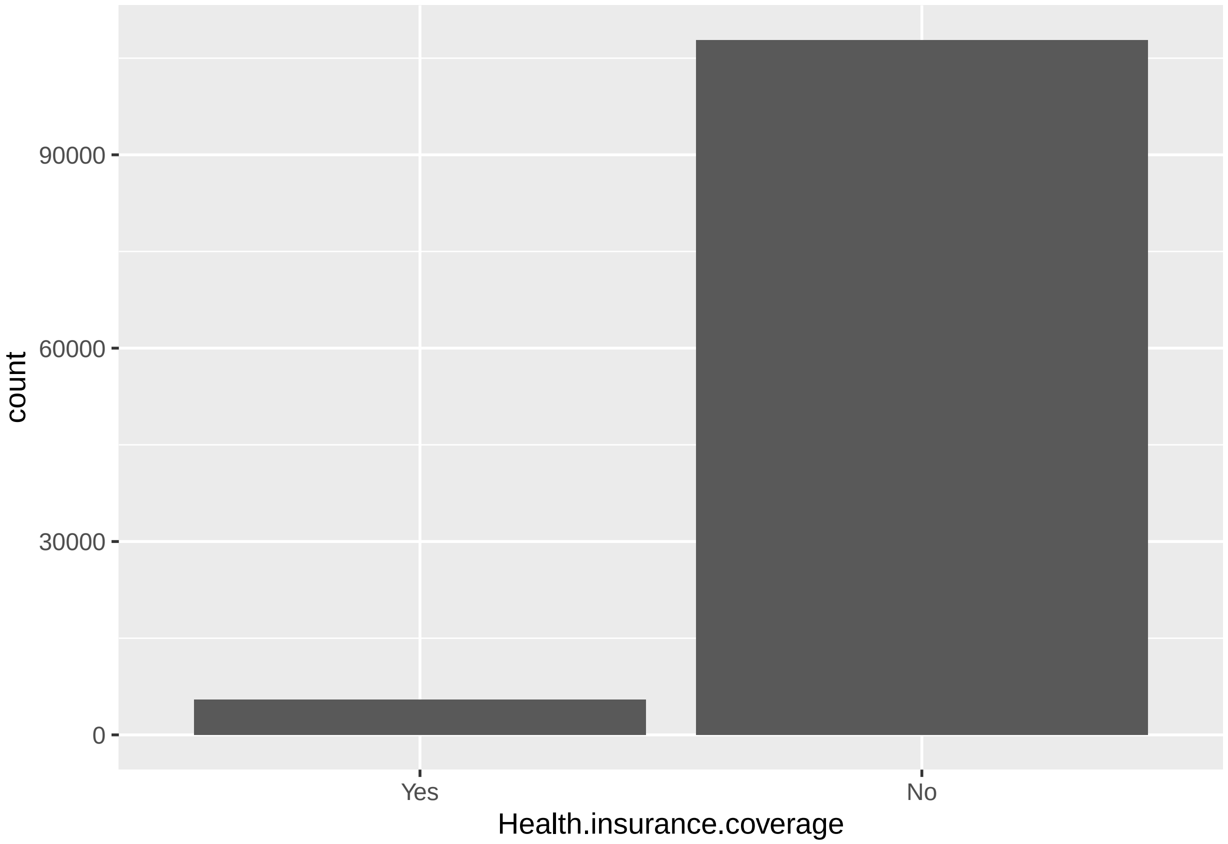
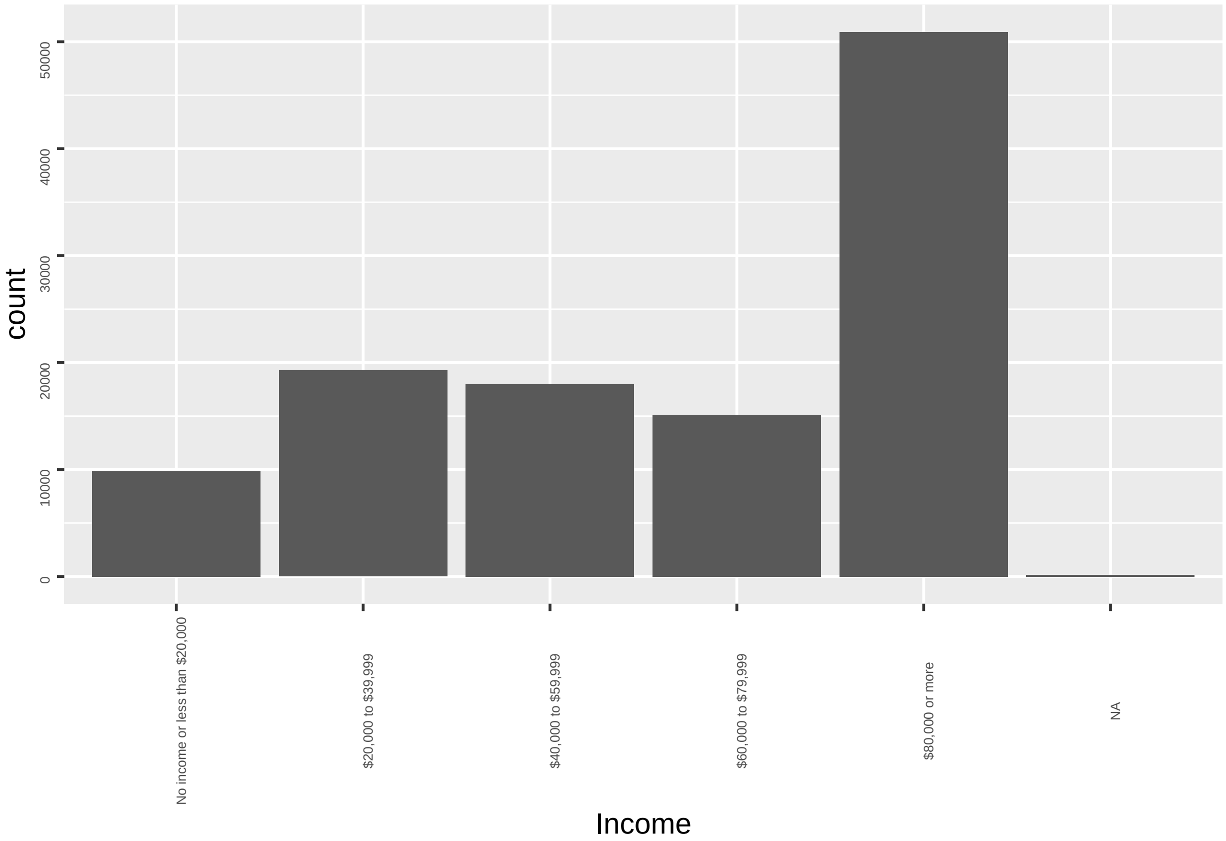


Figure 7: Distribution for Health Insurance Coverage


count
0
10000
20000
30000
40000
50000
No income or less than $20,000
$20,000 to $39,999
$40,000 to $59,999
$60,000 to $79,999
$80,000 or more
NA
Income

Appendix 2

Table 3: Associations between Last.time.had.flu.shot and Sex

	Last.time.had.flu.shot
Sex	2.2e-16

Appendix 3

Table 4: P Values from the Chi-Square Tests

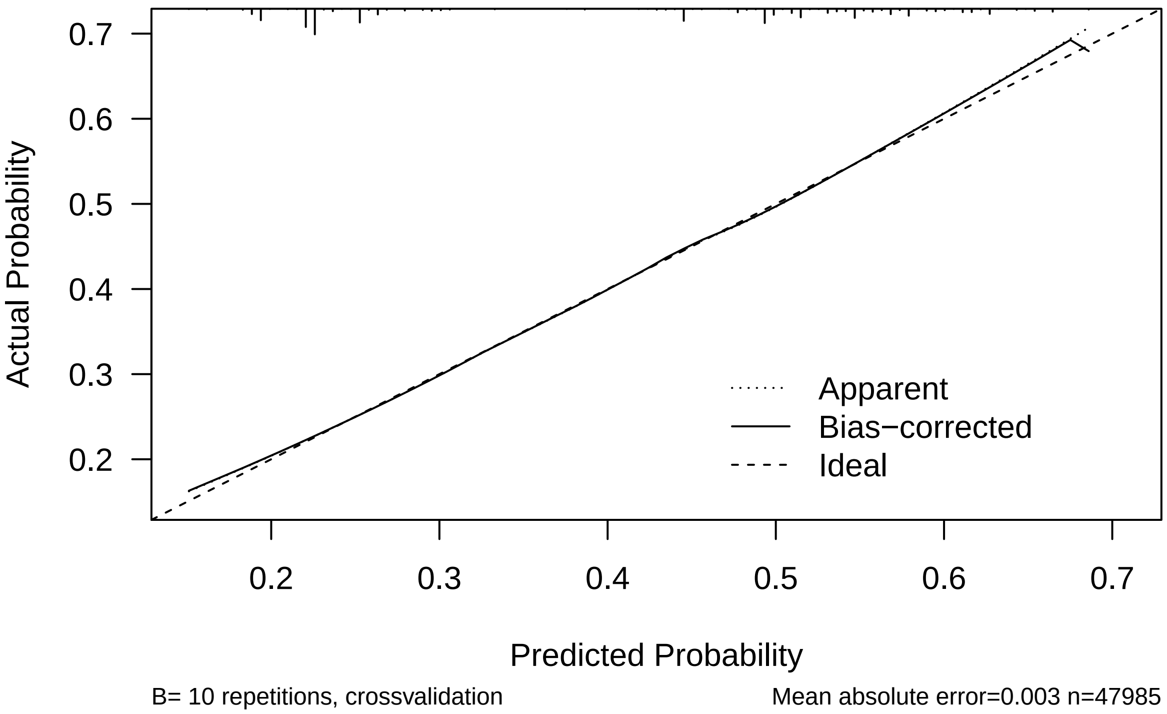
	P.Value
Add Marital.Status	2.2e-16
Add Age	2.2e-16
Add Education	5.423e-11
Add Access to Health Care	2.413e-08
Add Health Insurance Coverage	0.2115
Income	0.2533

Appendix 4

Table 5: Summary Table for Final Model

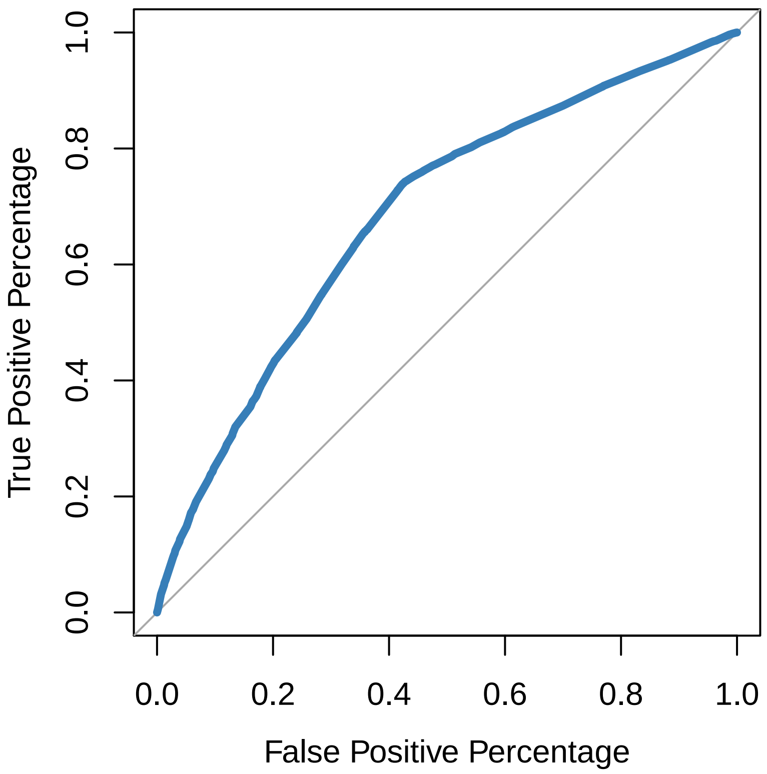
term	estimate	std.error	statistic	p.value
(Intercept)	-0.3158194	0.0615204	-5.133568	0.0000003
SexFemale	-0.2038707	0.0201537	-10.115789	0.0000000
Marital.statusCommon-law	0.3763141	0.0359613	10.464412	0.0000000
Marital.statusWidowed/Divorced/Separated	0.1968083	0.0274686	7.164852	0.0000000
Marital.statusSingle	0.2600571	0.0288360	9.018491	0.0000000
AgeYoung Adult	0.2840982	0.0480272	5.915364	0.0000000
AgeMiddle-aged Adult	-0.0078479	0.0490522	-0.159991	0.8728882
AgeOld Adult	-1.2067292	0.0483877	-24.938758	0.0000000
EducationSecondary school graduation, no post-secondary education	0.1982864	0.0336824	5.886940	0.0000000
EducationPost-secondary certificate diploma or univ degree	0.0516840	0.0308867	1.673340	0.0942604
Access.to.health.careNo	0.2386154	0.0430807	5.538803	0.0000000

Appendix 5



```
##
## n=47985   Mean absolute error=0.003   Mean squared error=2e-05
## 0.9 Quantile of absolute error=0.006
```

Appendix 6



```
## Area under the curve: 0.6855
```

Appendix 7

```
##Validating on the test data
test$logodds <-
  final_mod%>%
  predict(newdata = test)
test$estimate <- exp(test$logodds)/(1+exp(test$logodds))
mean((test$estimate - as.numeric(test$Last.time.had.flu.shot))^2)
```

```
## [1] 1.207982
```

Appendix 8

	names	model1
		(1)
1	(Intercept)	2.01775412614763 ***
2		(0.0211884059237421)
3	ProvincePRINCE EDWARD ISLAND	-0.0369002721035462 *
4		(0.0151512038993555)
5	ProvinceNOVA SCOTIA	-0.0627093011496101 ***
6		(0.0120527678664672)
7	ProvinceNEW BRUNSWICK	-0.0863926362368119 ***
8		(0.0123869879041716)
9	ProvinceQUEBEC	-0.106328657325233 ***
10		(0.00929912052511313)
11	ProvinceONTARIO	-0.0657183908293454 ***
12		(0.00920978723516253)
13	ProvinceMANITOBA	-0.0342889646181077 **
14		(0.0110501443858525)
15	ProvinceSASKATCHEWAN	-0.0297713249351331 **
16		(0.0115006708374254)
17	ProvinceALBERTA	-0.069334567792165 ***
18		(0.00982453672346468)
19	ProvinceBRITISH COLUMBIA	-0.0470703454239474 ***
20		(0.00969854886652943)
21	ProvinceYUKON	-0.0684560525586596 ***
22		(0.0196774789533523)
23	ProvinceNORTHWEST TERRITORIES	-0.0542861748786722 **
24		(0.0182663817577894)
25	ProvinceNUNAVUT	-0.0030739207346048
26		(0.0198996914844473)
27	SexFemale	-0.0148379945532533 ***
28		(0.0031903397988681)
29	Marital.statusCommon-law	-0.00989292842036699
30		(0.0050785653296609)
31	Marital.statusWidowed/Divorced/Separated	-0.0131091646014184
32		(0.00960454675826318)
33	Marital.statusSingle	-0.00967361029377286
34		(0.00947543428685721)
35	Family.arrangementUnattached individual living with others.	0.00702058128580241
36		(0.00854231686259143)
37	Family.arrangementIndividual living with spouse/partner.	-0.0179343118647671
38		(0.00956165269587343)
39	Family.arrangementParent living with spouse/partner and child(ren).	-0.0221522962934762 *
40		(0.00972210669225507)
41	Family.arrangementSingle parent living with children.	-0.0123079266951091
42		(0.00683130746712371)
43	Family.arrangementChild living with a single parent with or without siblings.	0.0302186338802559 **
44		(0.0113606107574223)
45	Family.arrangementChild living with two parents with or without siblings	0.0466786940017284 ***
46		(0.00892528088498665)
47	Family.arrangementOther	-0.0122832297162998
48		(0.00982939622831936)
49	AgeYoung Adult	-0.0250266839661732 *

	names	model1
50		(0.0117905535133912)
51	AgeMiddle-aged Adult	-0.0795801946416326 ***
52		(0.0122665950525457)
53	AgeOld Adult	-0.0978152427110827 ***
54		(0.0125531327479718)
55	Type.of.drinkerOccasional drinker	-0.00376075518639901
56		(0.00424255915652161)
57	Type.of.drinkerDid not drink in the last 12 months	0.0115876626456821 **
58		(0.00420895837080038)
59	IncomeLess than $20,000	-0.0257512812149929 *
60		(0.0122065524420548)
61	Income$20,000 to $39,999	-0.0281131259659129 *
62		(0.0122177144869012)
63	Income$40,000 to $59,999	-0.027193801967795 *
64		(0.0124787525662489)
65	Income$60,000 to $79,999	-0.0309485283630318 *
66		(0.0128835975197879)
67	Income$80,000 or more	-0.0290434687935204 *
68		(0.0127334023549276)
69	Education1	-0.0189198139201703 ***
70		(0.00332782541716072)
1.1	N	61780
2.1	R2	0.0172993966499535
3.1	logLik	-27327.8932704564
4.1	AIC	54727.7865409128
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	