STA304 Group 107: Topics on discrimination

Yifei Zhang, Ziyi Qu, Zihua Wang, Peilin Chen

2020/10/19

To Determine The Relationship Between Five Factors and Discrimination

Yifei Zhang, Ziyi Qu, Zihua Wang, Peilin Chen

2020/10/19

Code and data supporting this analysis is available at: https://github.com/EmiChen/To-Determine-The-Relationship-Between-Five-Factors-and-Discrimination.git

Abstract

The main topic of our discussion is Discrimination, which is based on data selected from Canadian General Social Survey (GSS) 2014, Cycle 28: Canadians' Safety and Security (Victimization). The methodology of our group work was running logistic regression, and the methodology of the survey itself from which we obtained data was stratified sampling method. Specifically, by running Logistic Regression, our group aimed to find the relationship between occurrence of discrimination, the response variable, and five explanatory variables: self-rated mental health, drinking habits, drug uses, school attendance, and number of relative friends. The final result of our study was: self-rated mental health and school attendance had statistically-significant negative relationship with being discriminated, yet frequency of drinking and drug usage had statistically-significant positive relationship with being discriminated, with number of relative friends having no significant relationship with the occurrence of discrimination. The final finding was reasonable and fitted perfectly into our goal of study since we found factors that were highly correlated with discrimination, from which we were able to conclude meaningful information to reduce the potential occurrence of discrimination, benefit people, and increase social welfare.

Introduction

The main topic of our discussion is Discrimination, which is understandably passionate as it has far-reaching implications for all citizens. Acted as an issue of public health and safety, discrimination can be as severe as a national problem affecting the overall development of the country, or as narrow as an affliction harming individual mental health.

The goal of our study was to find some factors that were highly correlated with discrimination so that certain actions could be taken to protect people's mental health and increase social welfare, from a long-term perspective.

To be more specific, the purpose of our study was to find the relationship between five factors we chose (independent variables) and the occurrence of discrimination (dependent variable) by conducting logistic regression. In our study, we chose victims of discrimination—5 years (discrim) to be dependent variable

and selected 5 independent variables: self-rated mental health(srh_115), drinking habits(drr_110), drug uses(dur_110), school attendance(esc1_01), and number of relative friends(isl_100). By observing the P-value, we hoped to determine the statistical significance of coefficients of these explanatory factors.

The hypothesis that we were trying to test was: self-rated mental health, school attendance, and number of relative friends would have statistically-significant negative relationship with being discriminated, while drinking frequency and drug usage would have statistically-significant positive relationship with being discriminated.

The structure of our report is:

1.Abstract: to show what was done, what was found, and why this matters

2.Introduction: to introduce topic/background/goal of our study

3.Data: to introduce everything about our data set

4.Model: to introduce our model, the methodology used, and the purpose of conducting each model

5.Result: to display the result from modeling

6.Discussion: to interpret the result

7.Weakness & Next Step: to discuss the weaknesses of our study and potential improvements

(Note: Please refer to each section about detailed analysis. The structure written above may not be as accurate or as comprehensive as information in each section.)

Data

We collected the data from Canadian General Social Survey (GSS) 2014, Cycle 28: Canadians' Safety and Security (Victimization), conducted statistical study based on modeling, and furthered our interest through making necessary recommendations. The survey itself used stratified sampling method, where the target population was partitioned into sub-populations based on geographic areas (different 10 provinces in Canada), each of which were further divided into strata. Strata among provinces represented different features while information within each strata was alike.

The original data wad comprised of 33090 rows and 790 columns, that was, 33089 observations and 789 variables (including both categorical and numerical). Among the 789 variables, our group chose five variables (four categorical variables and one numerical variable) that we first assumed are most relevant to the issues of discrimination.

The data was collected mainly from telephone, and 81.5% of the telephone numbers dialed reached eligible households. Few methods were used to deal with non-response. For example, if the timing of the first call was inconvenient, rearrangement of appointments to call back was scheduled. Also, the survey used re-contacting up to two times, or elaborated the importance of the survey to solve non-response issue.

The target population was the set of people the survey want to examine. In this case, the target consisted of all people of 15 years or above in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions.

The frame was where our survey sample can be drawn from, and the frame population in our survey included people whose telephone numbers in use were listed on various platforms (ie. Telephone companies' records and census of population) which were accessible to Statistics Canada, combined with people whose dwellings were listed on the Address Register within the ten provinces in Canada (Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan).

The sample population was the group of people the survey finally reached. In this case, the sample population consisted of 33127 eligible respondents the survey actually contacted with. For each province, minimum

sample sizes were determined to ensure that the survey results would generate acceptable sampling variability in certain estimates at the stratum level.

The response variable that we chose was if_discrim (discrim), which yielded binary response on whether this respondent had experienced any discrimination in the past five years.

The explanatory variables that we chose were:

x1:if_mental_healthy(srh_115) –>ranging from 1(excellent) to 5(poor) on mental health from self-rating.

x2:if_drink(drr_110) –> ranging from 1(everyday) to 7(never drinks) on respondent's drinking habits.

x3:if_drugs(dur_110) –> yielding binary response on whether the respondent used drugs.

x4:school_attendance(esc1_01) –> yielding binary response on whether the respondent had ever received education.

x5:number_relative_friends(isl_100) –>yielding number of relative friends the respondent had.

First of all, we chose if_mental_healthy(srh_115) to be one of the explanatory variables because we believed it had strong relationship with discrimination. If the respondent thought he had a good state of mind and that he was mentally healthy, he would in fact act more easygoing and in good-mood, which would decrease the possibility of being discriminated by others. Also, the reasons why we selected if_drink(drr_110) and if_drugs(dur_110) were similar, as we predicted people who endowed less drugs or drunk less would show more normal behaviour and thus would be less likely to be discriminated. Whether a person received education was also strongly correlated with discrimination. We believed that schooling and formal education could reduce the occurrence of discrimination. Last, we speculated that people who were more socially active (having more friends) tended to be mentally healthy so as to have less chance of encountering discrimination.

Our data had the advantage that the variables that we selected to construct the model were various. Independent variables included not only binary variables but also numerical variables. Therefore, we could rely on both categorical response and numerical answers to analyze the discrimination issue. Furthermore, the GSS data set used stratified sampling method, which had several advantages over simply sampling method. Because of the greater precision from stratified sampling, the smaller sample size that stratified method allowed was cost-saving. Also, a stratified sample was less likely to be unrepresentative since it ensured sufficient sample points from each strata based on subgroups' weights. That being said, when sample points were appropriately allocated to subgroups, disproportionate stratification would be more precise than proportionate stratification if variances existed across strata.

However, our data had some drawbacks. The first drawback was that the the result might be biased because every respondent could have different interpretation and understanding of words based on personal beliefs and judgments. For example, respondent would have diversified understanding of "excellent" when they rated their mental health, which made the data have personal bias. The second drawback of out data was that the model we tried to build would not be the best fit. To be more specific, factors other than the five we chose might also influence the chance of discrimination. There were also drawbacks from our dataset, such as low response rate and biased survey result, which would be demonstrated in details in Weakness Section.

Model

We believed that there was such a relationship in the population(true model):

$$Y_{if\_discrim} = \beta_0 + \beta_1 X_{if\_mental\_healthy\_2} + \beta_2 X_{if\_mental\_healthy\_3} + \beta_3 X_{if\_mental\_healthy\_4}$$
$$+\beta_4 X_{if\_mental\_healthy\_5} + \beta_5 X_{if\_drink\_2} + \beta_6 X_{if\_drink\_3} + \beta_7 X_{if\_drink\_4} + \beta_8 X_{if\_drink\_5}$$
$$+\beta_9 X_{if\_drink\_6} + \beta_{10} X_{if\_drink\_7} + \beta_{11} X_{if\_school\_attandence} + \beta_{12} X_{if\_drugs} + \beta_{13} X_{number\_relative\_friends}$$

β_0 was the intercept parameter, with $\beta_1 \sim \beta_{13}$ being the slope parameters.

In this study we built up a model to estimate the true model:

$$\hat{Y_{if\_discrim}} = \hat{\beta}_0 + \hat{\beta}_1 X_{if\_mental\_healthy\_2} + \hat{\beta}_2 X_{if\_mental\_healthy\_3} + \hat{\beta}_3 X_{if\_mental\_healthy\_4}$$

$+\hat{\beta}_4 X_{if\_mental\_healthy\_5} + \hat{\beta}_5 X_{if\_drink\_2} + \hat{\beta}_6 X_{if\_drink\_3} + \hat{\beta}_7 X_{if\_drink\_4} + \hat{\beta}_8 X_{if\_drink\_5} + \hat{\beta}_9 X_{if\_drink\_6}$

$+\hat{\beta_{10}} X_{if\_drink\_7} + \hat{\beta_{11}} X_{if\_school\_attandence} + \hat{\beta_{12}} X_{if\_drugs} + \hat{\beta_{13}} X_{number\_relative\_friends}$

$\hat{\beta}_0$ is the estimate of β_0, with $\hat{\beta}_1 \sim \hat{\beta}_{13}$ being estimates of $\beta_1 \sim \beta_{13}$.

We ran a logistic regression, using R, with y variable being if_discrim and x variables being if_mental_healthy, if_drink, if_drugs, school_attendance, and number_relative_friends.

The code that we used to run our model in R was:

myy <- glm(if_discrim~ as.factor(if_mental_healthy) + as.factor(if_drink) + if_school_attandence + if_drugs + number_relative_friends, data = data, family = "binomial")

There were reasons why we chose theses variables instead of others. We select discrim(victims of discrimination–past 5 years) rather than dis_10(discrimination–sex) and dis_40(discrimination–Age) because we wanted to analyze the issue of discrimination as a whole instead of demonstrating the relationship between five factors and a specific kind of discrimination. Also, we chose dur_110(Use of marijuana, hashish, hash oil or other cannabis - Respondent) to become one of our independent variables instead of dur_105 (How often (marijuana, etc.) - Respondent) because we wanted to analyze respondent's 'yes or no' choice of using drugs rather than their frequency of using. The reason why we selected self-rated mental health instead of self-rated general health to be one of our independent variables was that we believed mental health had a stronger relationship with discrimination than general health, which could help us build a stronger model.

The reason why we chose to run logistic regression was that our response variable yielded binary response: whether this respondent had ever experienced discrimination in the past five years. For the explanatory variables, if_drugs and school_attendance were binary variables which indicated whether a respondent used drugs or whether a respondent had attended formal educational institutions. Because if_drugs and school_attendance would yield binary responses, there was only one case other than the base-group case and we did not need to use as.factor() function. For if_drugs, a variable describing multiple degree of drug usage, and if_mental_healthy, a variable describing multiple mental health state, we needed to use as.factor() function to separate each degree and to use different coefficient to measure each degree.

Logit regression did not have model convergence, so discussion of this part could be skipped. For model checks and diagnostic issues, multicollinearity and cross validation could be involved for investigation. Multicollinearity was a common issue when there was a large number of covariates in a dataset. The severe problems it caused varied from unstable estimates to inaccurate variances. The multicollinearity diagnostic was highly necessary to ensure the accuracy and reliability of a logit regression model. There were several ways to implement such diagnostics. Correlation matrix was a helpful but not sufficient tool. Linear regression with the option tolerance, Vif, condition indices and variance proportions were better ways to examine the multicollinearity problem.

Model validation could assess logit regression's prediction capability on the relevant dataset. Specifically, cross-validation would be conducted on the training data. The algorithm process was, for example, to randomly divide the original training dataset into 10 subsets using R, then randomly used 9 of them as new "training" dataset while the remaining 1 became the new "test" dataset. Each time, by fitting a new model based on the new "training" data and making prediction on the new "test" data, we could check the MPE to assess the prediction accuracy of our model. Usage of calibration plot for checking the accuracy of regression estimates by visually observing whether the prediction line align with the observed line or not could also be helpful.

We also wanted to build up a bar chart to visualize the relationship between one of the explanatory variables, if_mental_healthy, and discrimination.

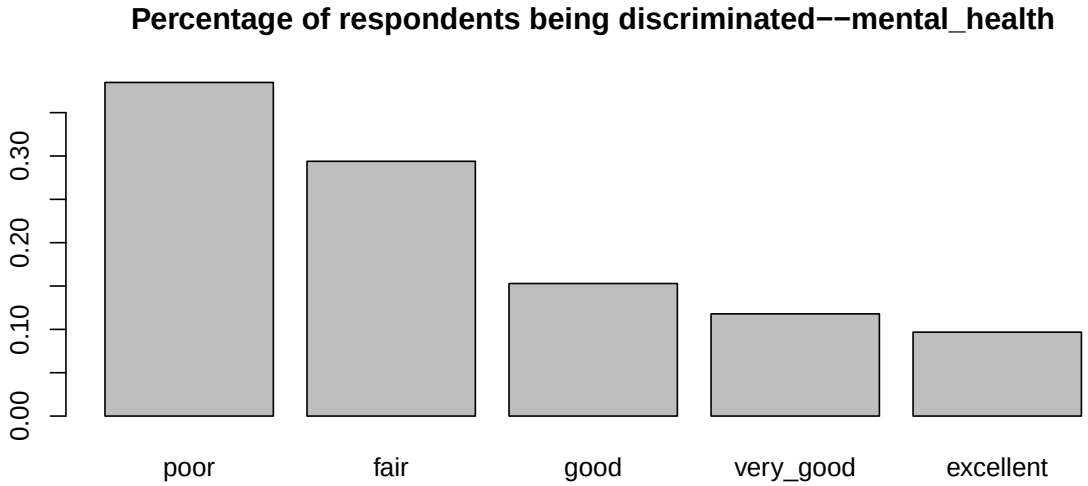
Results

By conducting logistic regression, we built up such a model:

$\hat{Y_{if\_discrim}}$ = -2.599+0.222$X_{if_mental_healthy_2}$ +0.533$X_{if_mental_healthy_3}$
+1.325$X_{if_mental_healthy_4}$+1.783$X_{if_mental_healthy_5}$ + 0.24$X_{if_drink_2}$+0.188$X_{if_drink_3}$
+0.232$X_{if_drink_4}$+0.367$X_{if_drink_5}$+0.471$X_{if_drink_6}$+0.267$X_{if_drink_7}$+ 0.615$X_{if_school_attandence}$
+X_{if_drugs}-0.0001$X_{number_relative_friends}$

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic   p.value
##    <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                   -2.60      0.0911     -28.5   4.03e-179
##  2 as.factor(if_mental_healthy)2  0.222     0.0431       5.16  2.53e-  7
##  3 as.factor(if_mental_healthy)3  0.533     0.0450      11.8   2.60e- 32
##  4 as.factor(if_mental_healthy)4  1.33      0.0660      20.1   9.83e- 90
##  5 as.factor(if_mental_healthy)5  1.78      0.118       15.1   9.02e- 52
##  6 as.factor(if_drink)2           0.240     0.110        2.18  2.95e-  2
##  7 as.factor(if_drink)3           0.188     0.0971       1.93  5.30e-  2
##  8 as.factor(if_drink)4           0.232     0.0991       2.34  1.92e-  2
##  9 as.factor(if_drink)5           0.367     0.0928       3.95  7.81e-  5
## 10 as.factor(if_drink)6           0.471     0.103        4.57  4.88e-  6
## 11 as.factor(if_drink)7           0.267     0.0927       2.88  4.02e-  3
## 12 if_school_attandence           0.615     0.0461      13.4   1.16e- 40
## 13 if_drugs                       1.00      0.181        5.54  3.08e-  8
## 14 number_relative_friends       -0.000126  0.000918    -0.137 8.91e-  1
```

We also built up a bar chart to visualize the negative relationship between mental health and discrimination.



Discussion

$\hat{\beta}_0$= -2.5986974 means that when a person has "excellent" mental health, drinks every day, has not attended school, has used to drug usage and has no close friend, the log odds of being discriminated for that person is -2.5986974.

$\hat{\beta}_1$= 0.2220942 means that, compared with people who have "excellent" mental health, people who have "very good" mental health will have 0.2220942 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_2$= 0.5329586 means that, compared with people who have "excellent" mental health, people who have "good" mental health will have 0.5329586 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_3$= 1.3252361 means that, compared with people who have "excellent" mental health, people who have "fair" mental health will have 1.3252361 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_4$= 1.7829967 means that, compared with people who have "excellent" mental health, people who have "poor" mental health will have 1.7829967 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_5$= 0.2403246 means that, compared with people who drink every day, people who drink 4-6 times a week are expected to have 0.2403246 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_6$= 0.1878515 means that, compared with people who drink every day, people who drink 2-3 times a week are expected to have 0.1878515 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_7$= 0.2319451 means that, compared with people who drink every day, people who drink once a week are expected to have 0.2319451 higher log odds of being discriminated, keep holding everything else the same.

$\hat{\beta}_8$= 0.3667383 means that, compared with people who drink every day, people who drank once or twice in the past month are expected to have 0.3667383 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta}_9$= 0.4710193 means that, compared with people who drink every day, people who did not drink in the past month are expected to have 0.4710193 higher log odds of being discriminated,holding everything else the same.

$\hat{\beta_{10}}$= 0.2666597 means that, compared with people who drink every day, people who never drink are expected to have 0.2666597 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta_{11}}$= 0.6153804 means that, compared to people who have not attended schools, people who have attended schools are expected to have 0.6153804 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta_{12}}$= 0.9999243 means that, compared to people who do not use drugs, people who use drugs are expected to have 0.9999243 higher log odds of being discriminated, holding everything else the same.

$\hat{\beta_{13}}$= -0.0001259 means that people who have one more relative friend are expected to have 0.0001259 lower log odds of being discriminated, holding everything else the same.

By analyzing P-values, we could find the significance of each estimate in predicting the odds of being discriminated for a person with certain characteristics. For example, given a 5% significance level, we would reject null hypothesis and conclude that a β was statistically different from zero if the p-value is smaller than 5%. A close scrutiny on the summary of our logistic regression revealed that the explanatory variables of four categorical factors: if_mental_healthy, if_drink, if_school_attendance and if_drugs, were all significant in predicting the log odds of being discriminated for a person, while the numerical variable, number_relative_friends, showed less significant in predicting a person's log odds of being discriminated, with a high P-value of 0.89092.

To visualize the relationship between x variables and discrimination, we draw a bar chart with x representing different rating of mental health and y representing percentage of number of respondent in different mental state received discrimination. The negative relationship between mental health and discrimination could be easily seen from the graph. From the bar chart we could find out that as a respondent became more mentally healthy, he or she were less likely to face discrimination.

Based on the results we obtained in Result Section, we draw the conclusion that self-rated mental health and school attendance had statistically-significant negative relationship with being discriminated, yet drinking frequency and drug usage had statistically-significant positive relationship with being discriminated, with number of relative friends having no significant relationship with the occurrence of discrimination. Once again, the original goal of our study was to find factors that had high correlation with occurrence of discrimination so that actions and policies could implemented to reduce discrimination and to increase social welfare.

The conclusion drawn from the result perfectly fitted our goal, as some useful information could be inferred to benefit the society. For example, by knowing education had negative relationship with being discriminated, the government could promote education to mitigate the problem of discrimination. Also, community could hold consulting services to help individual with mental illness so as to diminish the occurrence of discrimination.

Weaknesses

The first weakness was the low response rate. According to the user guidebook provided by Statistics Canada, the overall response rate was only 52.9%, even though a few techniques had been used to prevent non-response such as rearranging appointments to call back when the timing of the first call was inconvenient, and re-contacting up to two times or elaborating the importance of the survey. The relatively low response rate was a major source of non-sampling errors in survey, which could potentially devalue the results. Moreover, the extent of non-response varied from partial non-response to total non-response, for which the non-response would be discussed in details below. A survey's response rate was long being viewed as a significant indicator of survey quality, and low response rate could give rise to sampling bias. Furthermore, largely because of the non-response, the actual sample size of 33,127 records was approximately 5,000 less than target size of 39,674, adding on another layer of risk of non-representative bias caused by the smaller than expected sample size.

Besides total non-response, partial non-response should be regarded as another weakness that needed to be discussed with attention. Such non-response occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information. All the above scenarios would lead to missing value, yet with distinct degrees of bias and different groups of population being under- or over-represented. Some of the possible approaches to mitigate the disadvantages brought by missing value would be discussed in the following section.

The next weakness came from the sampling methodology used by General Social Survey (GSS). GSS used a stratified design, with significant differences in sampling fractions between strata. As a result, some areas were over-represented in the sample while some were relatively under-represented. Together with the non-response issue we argued above, such sampling method made the sample even less representative.

The limited number of explanatory variables to be chosen from GSS was the last weakness to discuss in this section. Besides self-rated mental health, drinking habits, drug usage, school attendance and number of close friends, there could have been other reasonable and significant factors, which allowed us to gain a deeper insight into the topic of discrimination. Thus, our current model could have drawn more thorough and valuable conclusions if we had access to other stronger variables that were excluded by the GSS 2014, for example, language and caste.

Next Steps

Acknowledging the fact that total non-response appeared due to interviewer's failure to contact the respondent, inability of the household to provide the information or the refusal of participation, some techniques can be considered to assuage bias to the largest possible extent. Groves, Cialdini, and Couper (1992), who

modeled factors of survey participation, combining socio-demographic, survey design, and psychological considerations, have suggested that one of the most efficient strategies to increase response rates was incentives and mode of contact. Dillman (2007) drawn the same conclusion. Specifically, by offering financial incentives (e.g. payment of a perceived gift) and non-financial incentives (e.g. underlying charities or donations), respondents will more willing to participate and respond to a survey. In addition to motivations, Dillman supposed that personal contact enhanced response rates. To clarify, specific elements like the timing for the first contact, the personalization of the contact and the words used all contribute to level of response rate of a survey.

Concerning the next step for solving missing value (partial non-response) issue, three general strategies for analyzing incomplete data are summarized by Little and Rubin (Little and Rubin 1987, 1989; Rubin 1987; Little 1988) and by others more recently: direct analysis of the incomplete data, weighting, and imputation. Imputation method will be focused on for now. Basically, imputation involves replacing missing values with suitable estimates by using an appropriate model that incorporates random variation. Skipping the underlying machine learning concepts covered by imputation method, there are various tools available for performing imputation, including R, STATA, SAS and so on.

As mentioned in the last second section, the topic of discrimination is large in scope. Our present study is still shallow. More factors should be involved to better examine such topic. Having access to data set of other potentially more relevant variables, for instance language and caste, could contribute to the enhancement of effectiveness of our analysis results. Caste discrimination might not be suitable to address in the Canadian context, but language is suitable. People are sometimes discriminated because their preferred language is associated with a particular ethical group. The Anti-Quebec sentiment in Canada which targeted people who speak French would be an example. A subsequent study based on adding language as an explanatory variable to predict probability of being discriminated can be considered as a next step.

References

Vera Toepoel & Matthias Schonlau (2017) Dealing with nonresponse: Strategies to increase participation and methods for postsurvey adjustments, Mathematical Population Studies, 24:2, 79-83, DOI: 10.1080/08898480.2017.1299988 https://www.tandfonline.com/doi/full/10.1080/08898480.2017.1299988

Maria Pampaka, Graeme Hutcheson & Julian Williams (2016) Handling missing data: analysis of a challenging data set using multiple imputation, International Journal of Research & Method in Education, 39:1, 19-37, DOI: 10.1080/1743727X.2014.979146 https://www.tandfonline.com/doi/full/10.1080/1743727X.2014.979146

https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss28/gss28/more_doc/GSSc28gid.pdf

Habshah Midi, S.K. Sarkar & Sohel Rana (2010) Collinearity diagnostics of binary logistic regression model, Journal of Interdisciplinary Mathematics, 13:3, 253-267, DOI: 10.1080/09720502.2010.10700699 https://www.tandfonline.com/doi/abs/10.1080/09720502.2010.10700699?journalCode=tjim20

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag NewYork, 2016.