Types of housing needs

Henry Lik-Tin Cheng, Zefeng Weng, Weizhi Guo, Ziping Liu

12/10/2020

Abstract

Housing is an indispensable part of people's lives, we want to discover people's attitudes towards whether to own a house in a diffetnent situations through this survey. We found that when people have enough money, they will tend to own houses instead to rent houses.This is because people want to own a house to ensure their quality of life.

Introduction

Several decades ago, people worked for life and would like to spent all their savings to buy a house, since in their mind, having their own house is the reason for working hard. However, people's mind might have changed now. In this fast-paced modern society, people are tending to rent the house instead of buying a house to avoid carrying a burden of debt. In order to investigate the relationship among various factors and making the decision of whether to rent a house, a large amount of data was collected. By removing some unnecessary factors that are not of interest, a logistic model was built for predicting the probability of renting a house for people in different ages, genders, marital status and the children conditions. After building up the model, some critical findings related to the model will be briefly discussed and some essential diagnostics are also checked to validate the model.

Data

The dataset was from General Social Survey (GSS) on the family. In this dataset, more than 20000 subjects were surveyed about the information related to the family situations, including ages, marital status, living conditions and etc. The targeted population for this survey is all the Canadian citizens that are over 15 years old living in the 10 province in Canada.

The survey method is through telephone survey and may involves selection bias and voluntary response bias. To conduct the survey, a combination of telephone numbers and Statistics Canada's Address Register is used. The sampling frame is from a list of telephone numbers created in 2013. And more than 300 variables are collected.

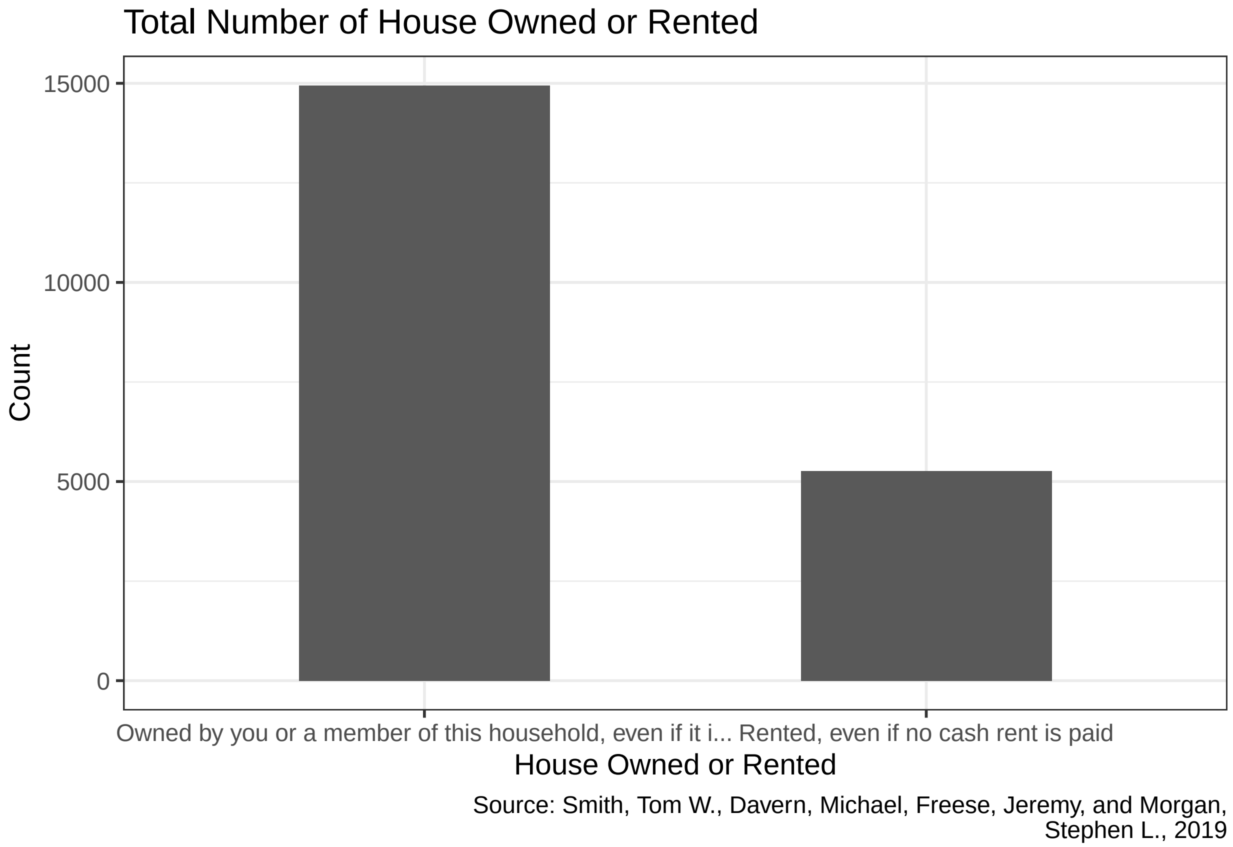
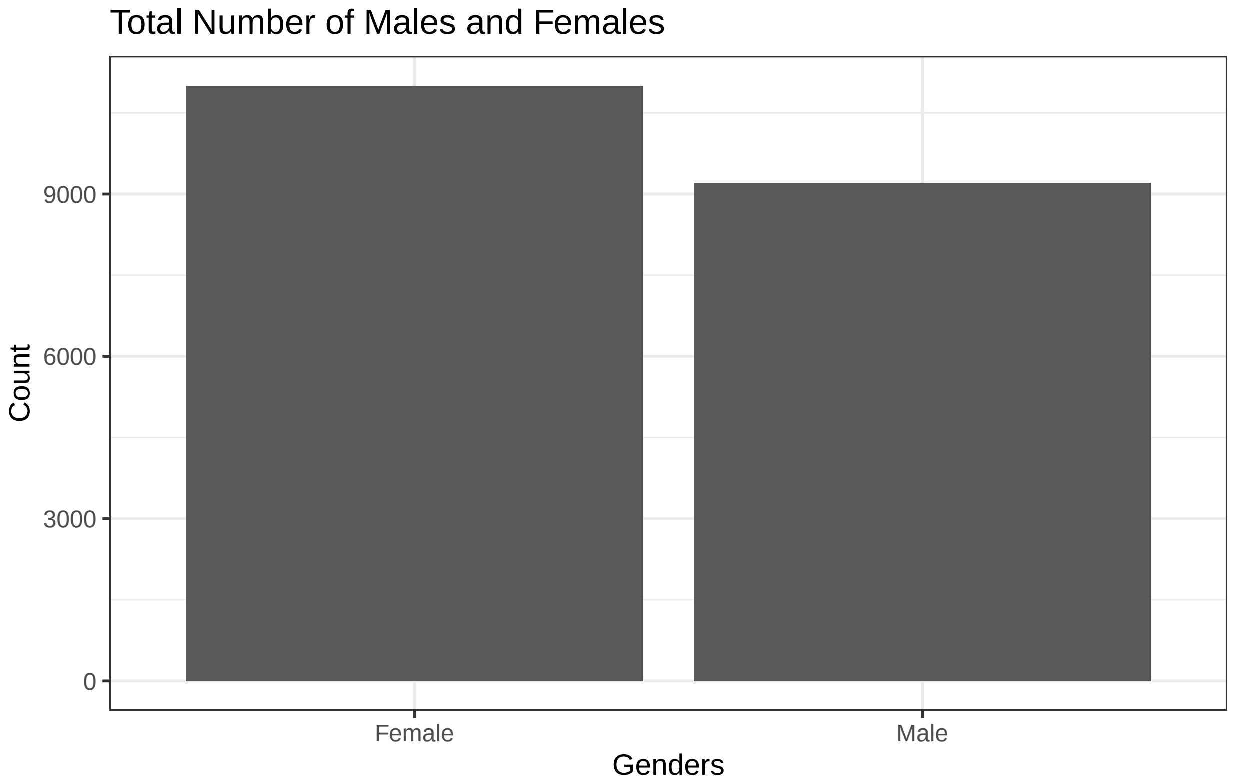


Figure 1: Total Number of House Owned or Rented

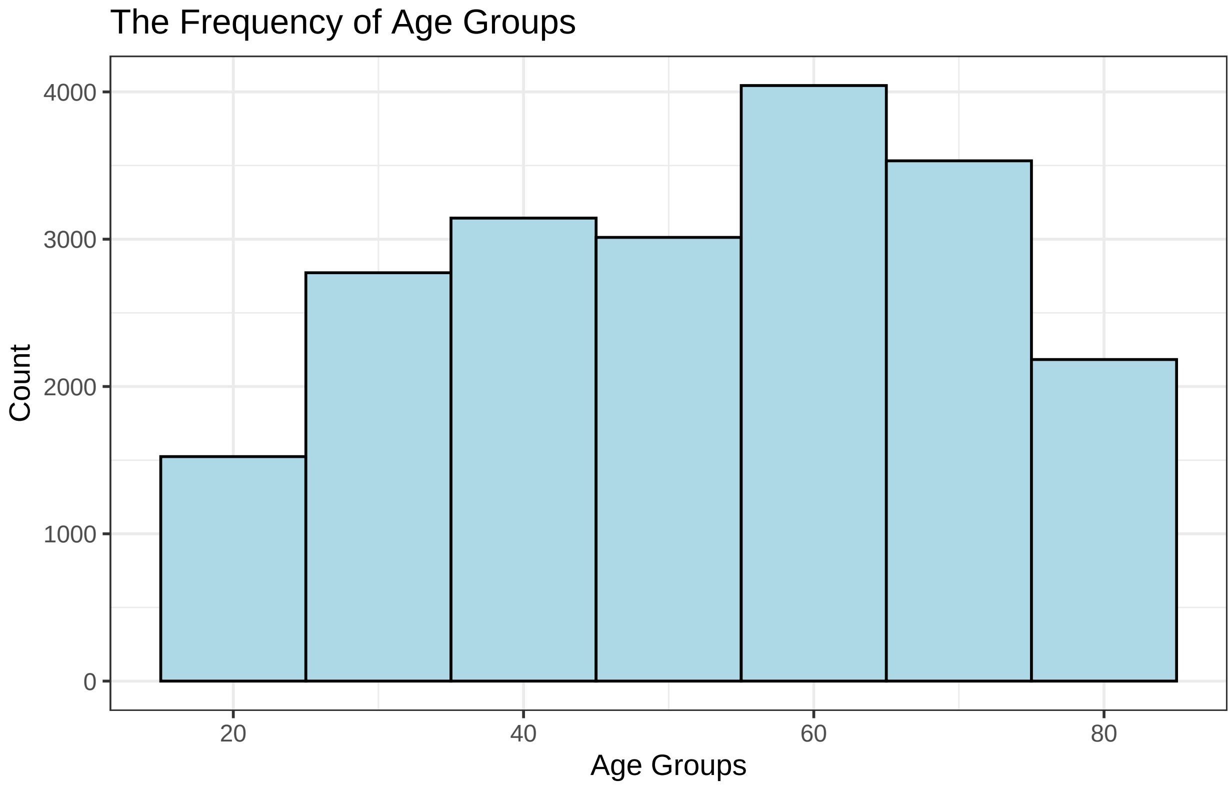
In this survey, for the response we got, owned by you or a member of this household has a relatively large number, which is almost 3 times that of rented houses.



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 2: Total Number of Males and Females

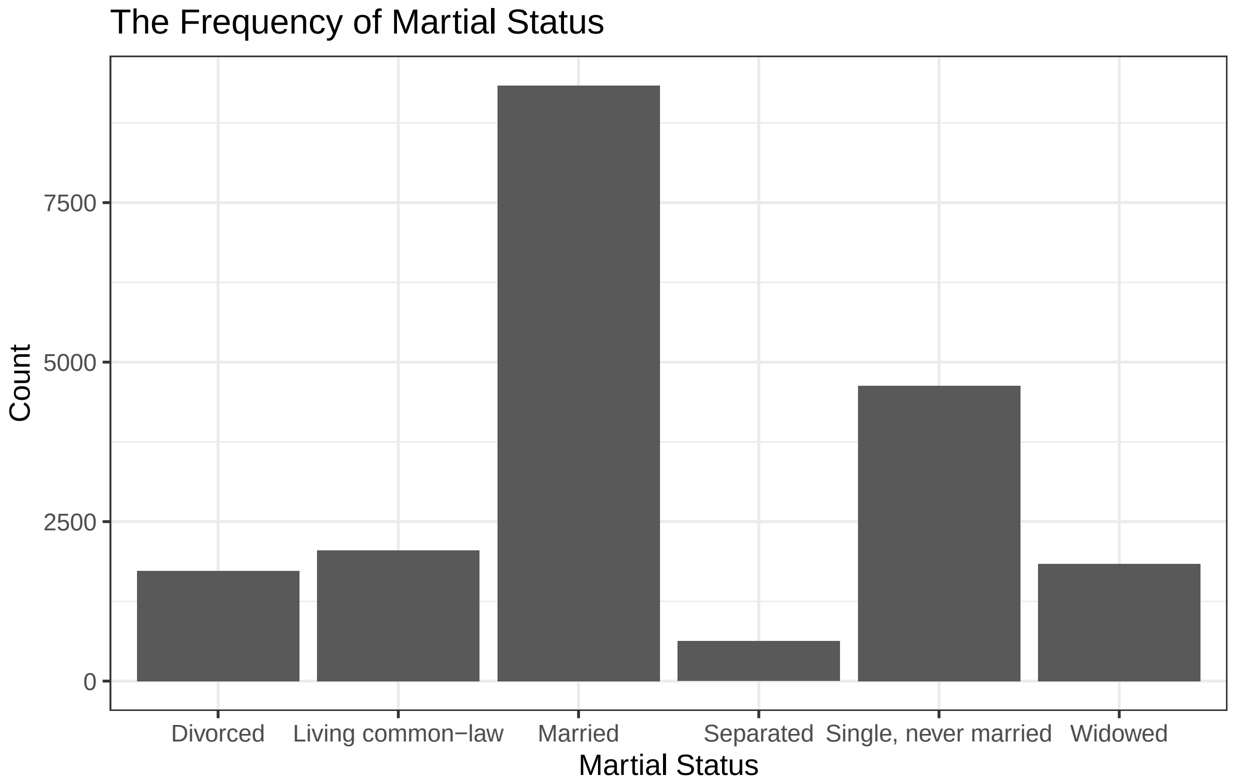
In our survey, among the responses we got, women accounted for more than men.



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 3: The Frequency of Age Groups

The population of our survey is mainly concentrated in the age group of 40 to 70 years old. 20 and 80 years old are relatively less


The Frequency of Martial Status

Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 4: The Frequency of Martial Status

In the response we got, most people's marital status is married, and the proportion of separated is the least.

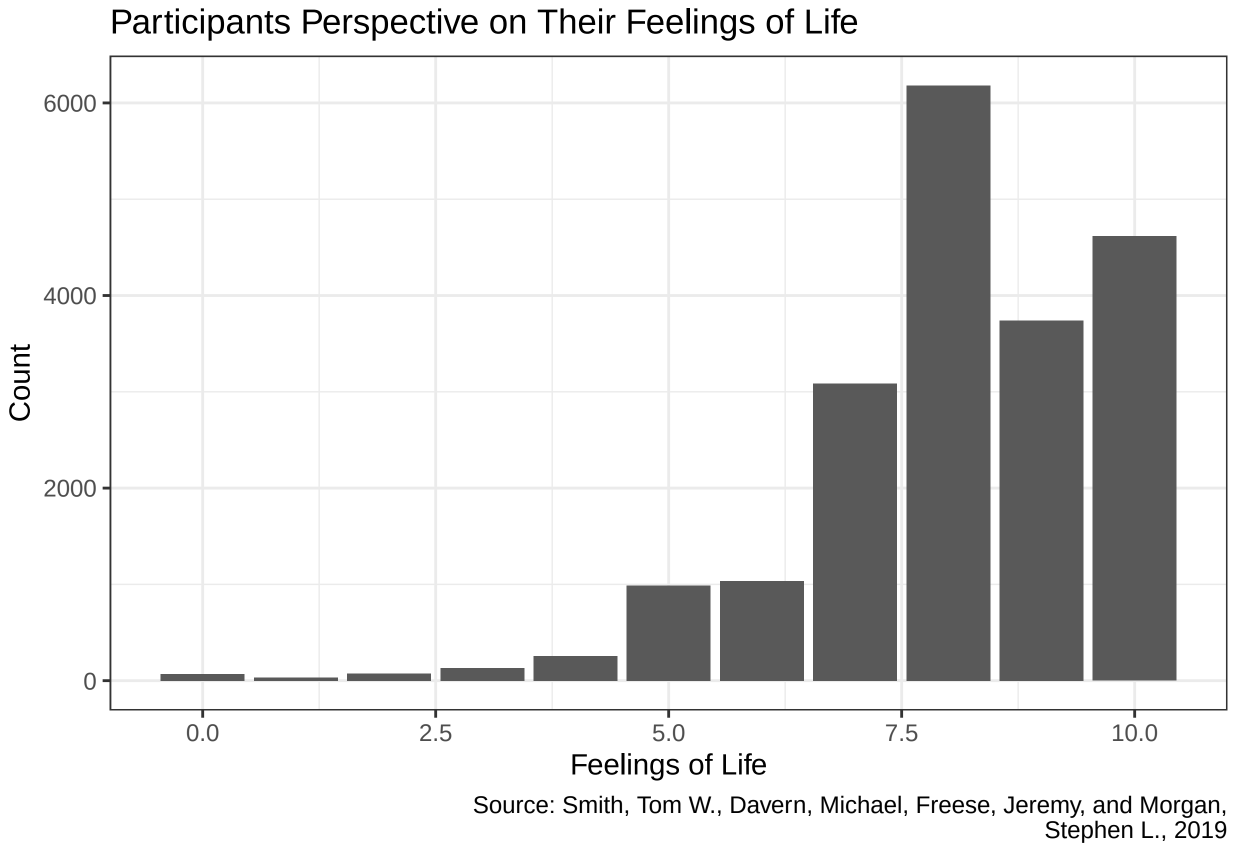
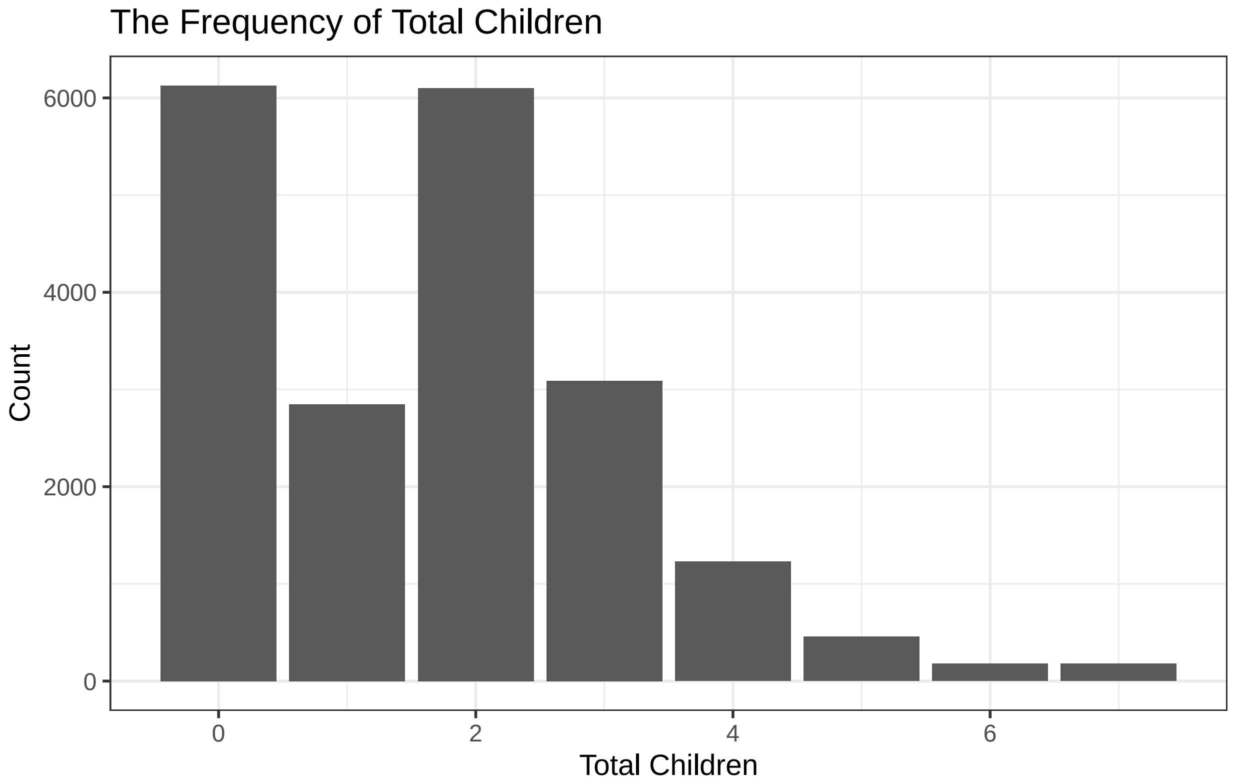


Figure 5: Participants Perspective on Their Feelings of Life

In the response, most people would rate Feeling of Life above 7 . Most of people give a score of 7.7 and the person with a score of 1 is the least.


The Frequency of Total Children

Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 6: The Frequency of Total Children

Among the participants in the survey, families with 0 children and 2 children are the largest number, while the proportion of families with 6 or more children is relatively small

Model

The code below estimates the logistic regression model by using function called 'glm', which convert to a factor to indicate the level should be treated as categorical variable.

```
##
## Call:
## glm(formula = y ~ age + sex + marital_status + feelings_life +
##     total_children, family = "binomial", data = gss1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2412  -1.0969   0.5176   0.8173   1.5095
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.136597   0.116807  -9.731  < 2e-16 ***
## age                                  0.008368   0.001207   6.933 4.11e-12 ***
## sexMale                              0.071984   0.034900   2.063   0.0392 *
## marital_statusLiving common-law      0.782061   0.074194  10.541  < 2e-16 ***
## marital_statusMarried                1.530467   0.059709  25.632  < 2e-16 ***
## marital_statusSeparated              0.119909   0.096610   1.241   0.2145
## marital_statusSingle, never married  0.029949   0.065821   0.455   0.6491
## marital_statusWidowed                0.284847   0.072493   3.929 8.52e-05 ***
## feelings_life                        0.127268   0.010068  12.641  < 2e-16 ***
## total_children                       0.003779   0.013899   0.272   0.7857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23182  on 20208  degrees of freedom
## Residual deviance: 20937  on 20199  degrees of freedom
## AIC: 20957
##
## Number of Fisher Scoring iterations: 4
```

$$Pr(y_i = 1) = \text{logit}^{-1}(5.29845 - 0.01946\alpha^{Age}_{a[i]} + 0.15543\alpha^{sexMale}_{b[i]} - 0.782061\alpha^{statusLiving}_{c[i]} -$$

$$1.530467\alpha^{statusMarried}_{c[i]} - 0.119909\alpha^{statusSeperated}_{c[i]} - 0.029949\alpha^{statusSingle}_{c[i]} - 0.284847\alpha^{statusWidowed}_{c[i]} -$$

$$0.127268\alpha^{life}_{d[i]} - 0.003779\alpha^{children}_{e[i]})$$

where the α are age groups, genders, martial statuses that include living together; married; separated; single and widowed, feelings of life and total children effects, respectively. The notation $a[i]$ refers to age a which individual i belongs. These are modeled as:

$$\alpha^{age}_a \sim N(0, sigma_{age}) \text{ for } a = 1, 2, \ldots, A$$

where A is the total number of age groups.

We named the summary above "logistics" which stands for logistics analysis model we just mentioned. In the output above, we called 'glm' formula, and set if a person lives in a rented or owned residence. On the other hand, we combine situations like their age, gender, marital status, feelings of life and total children. After that, we could see deviance residuals, which show the distribution for individual situation we want to observe in the logistic model. We will elaborate the data below from the summary. The second column which has title named “Estimate” shows the coefficient; ‘age’, ‘sexMale’,‘feelings of life’ and ‘total children’ are statistically significant. As marital status, the logistic regression coefficients express the change in the log odds of the outcome for any one unit increase in the predictor variable. For every one unit change in ‘age’, the log odds of ‘own_rent’ decreased by 0.008368. For every one unit change in ‘sexMale’, the log odds of ‘own_rent’ decreased by 0.071984. For every one unit change in ‘feelings_of_life’, the log odds of ‘own_rent’ decreased by 0.127268. For every one unit change in ‘total_children’, the log odds of ‘own_rent’ decreased by 0.003779. As marital status, it is apparently an indicator variable (‘divorced’,‘living common-law’, ‘married’, ‘separated’, ‘single, never married’, ‘widowed’ totally 6 levels), which has different interpretation as above: we might say marital status of living in common-law comparing with divorces, the change of log odds of ‘rent_own’ will decrease 0.782061.

Confidence Interval

Now, we would like to figure out the confident interval of the coefficient estimates by using ‘confint’ function. For, logistics models, confidence intervals are based on the log-likelihood function. If we want to show the confident interval by adding standard errors, we may use the default method as second line shows. Both confident intervals use default alpha level with 5%(0.05). As we could conclude from results, confident intervals of separated, Single, never married and total children include ‘0’, which mean there have close relationships between confidence intervals and significance tests. All values included inside the confidence interval are plausible values for the parameter, whereas values outside the confidence interval should be rejected as plausible values for the parameter.

```
##                                           2.5 %      97.5 %
## (Intercept)                        -1.365785647 -0.90787614
## age                                 0.006004567  0.01073574
## sexMale                             0.003613649  0.14042720
## marital_statusLiving common-law     0.636919130  0.92778965
## marital_statusMarried               1.413359307  1.64744310
## marital_statusSeparated            -0.068937953  0.30988242
## marital_statusSingle, never married -0.099173006  0.15886658
## marital_statusWidowed               0.142883847  0.42708373
## feelings_life                       0.107547711  0.14701749
## total_children                     -0.023411330  0.03107882
```

```
##                                           2.5 %      97.5 %
## (Intercept)                        -1.365534143 -0.90766041
## age                                 0.006002174  0.01073306
## sexMale                             0.003579937  0.14038734
## marital_statusLiving common-law     0.636643914  0.92747845
## marital_statusMarried               1.413438939  1.64749447
## marital_statusSeparated            -0.069443471  0.30926097
## marital_statusSingle, never married -0.099058530  0.15895658
## marital_statusWidowed               0.142763105  0.42693006
## feelings_life                       0.107535473  0.14700126
## total_children                     -0.023463115  0.03102185
```

From results shown above, only separated, single but never married marital status and total children confident intervals contain ‘0’, which mean there is a close relationship between confidence intervals and significance tests. Specifically, if a statistic is significantly different from 0 at the 0.05 level, then the 95% confidence interval will not contain 0.

Chi-squared Test

We want to test the overall effect of marital status, so we use ‘wald.test’ function of ‘aod’ library, which means we are going to compute a Wald Chi-squared test for 1 or more coefficients, given their variance-co-variance matrix. Here we should make sure the order in which the coefficients are given in the table of coefficients have same order in the model because ‘wald.test’ function refers to the coefficients by their order in the model. We use the ‘wald.test’ function where ‘b’ stands for the coefficients; ‘Sigma’ stands for the variance, co-variance matrices of the error terms. Ultimately, ‘Terms’ stands for which terms are going to be tested in the model. (i.e Terms 3:7 are all the marital status listed in the logistics model summary)

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 1252.0, df = 5, P(> X2) = 0.0
```

The result above saying the Chi-squared test statistic of 1252.0, with degree of freedom of 5, associated with a P-value of infinitely approaching to 0, which implies the overall effect of marital status is statistically significant.

We now would like to test the additional hypotheses about the differences in the coefficients for the different levels of marital status. First of all, we pull out marital status are married and single, but never married, and compare one another. Basically, the first line of code we create a vector called ‘l1’ that defines the test we intend to perform. In this case, we just simply test the difference between married and single status (we may test other different term if we want), we multiply one of them by 1, and the other by -1, respectively. Due to other terms are irrelevant, therefore we denote them as ‘0’. The second line we let R know we wish to base the test on the vector ‘l1’.

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 872.6, df = 1, P(> X2) = 0.0
```

The result above shows that the Chi-squared test statistic of 872.6, degree of freedom of 1, with associated P-value infinitely approaching to 0; thus, indicating that the difference between married and single, but never married status are statistically significant.

Odd-ratios

Now we want to figure out the odds-ratios, which R studio would do this for us. To get the eponentiated coefficients, we just simply use ‘exp’ functions, and the terms we want to exponentiate is called the coefficients. To put it all in one table, we use ‘cbind’ function to bind those coefficients and confidence intervals together, where is shown in the second line of code below.

```
##                         (Intercept)                                 age
##                           0.3209091                           1.0084027
##                             sexMale     marital_statusLiving common-law
##                           1.0746378                           2.1859733
##               marital_statusMarried             marital_statusSeparated
##                           4.6203326                           1.1273940
## marital_statusSingle, never married               marital_statusWidowed
##                            1.0304020                           1.3295580
##                        feelings_life                      total_children
##                            1.1357218                           1.0037865
```

```
##                                            OR     2.5 %   97.5 %
## (Intercept)                         0.3209091 0.2551801 0.403380
## age                                 1.0084027 1.0060226 1.010794
## sexMale                             1.0746378 1.0036202 1.150765
## marital_statusLiving common-law     2.1859733 1.8906471 2.528913
## marital_statusMarried               4.6203326 4.1097381 5.193683
## marital_statusSeparated             1.1273940 0.9333846 1.363265
## marital_statusSingle, never married 1.0304020 0.9055860 1.172182
## marital_statusWidowed               1.3295580 1.1535958 1.532781
## feelings_life                       1.1357218 1.1135440 1.158374
## total_children                      1.0037865 0.9768606 1.031567
```

We could straightforwardly find out for each unit increase in ‘age’, the odds of being “rent_own” increases by a factor of 0.9916673 etc.

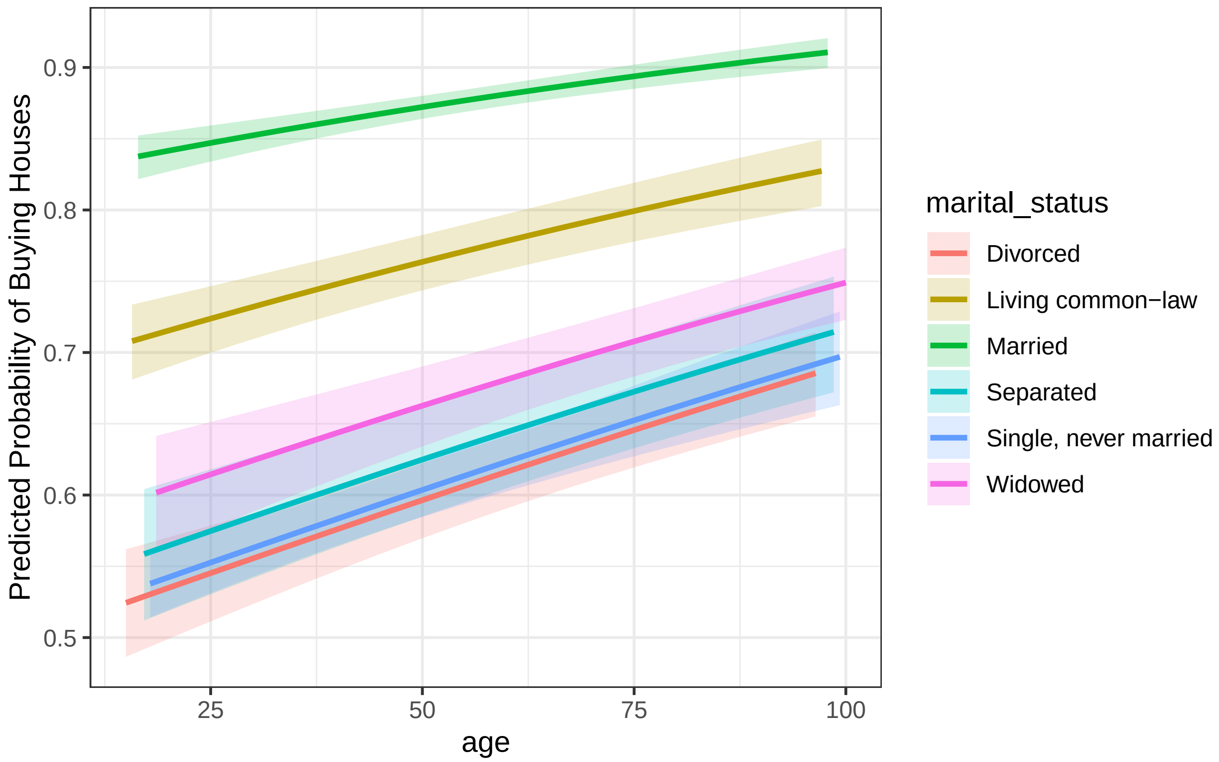
Graphs

We want to use predicted probabilities to help understanding the model, which could be computed for both categorical and continuous variables. So we use function called ‘with’ to build up a new data frame with the values we want the independent variables to create our predictions.

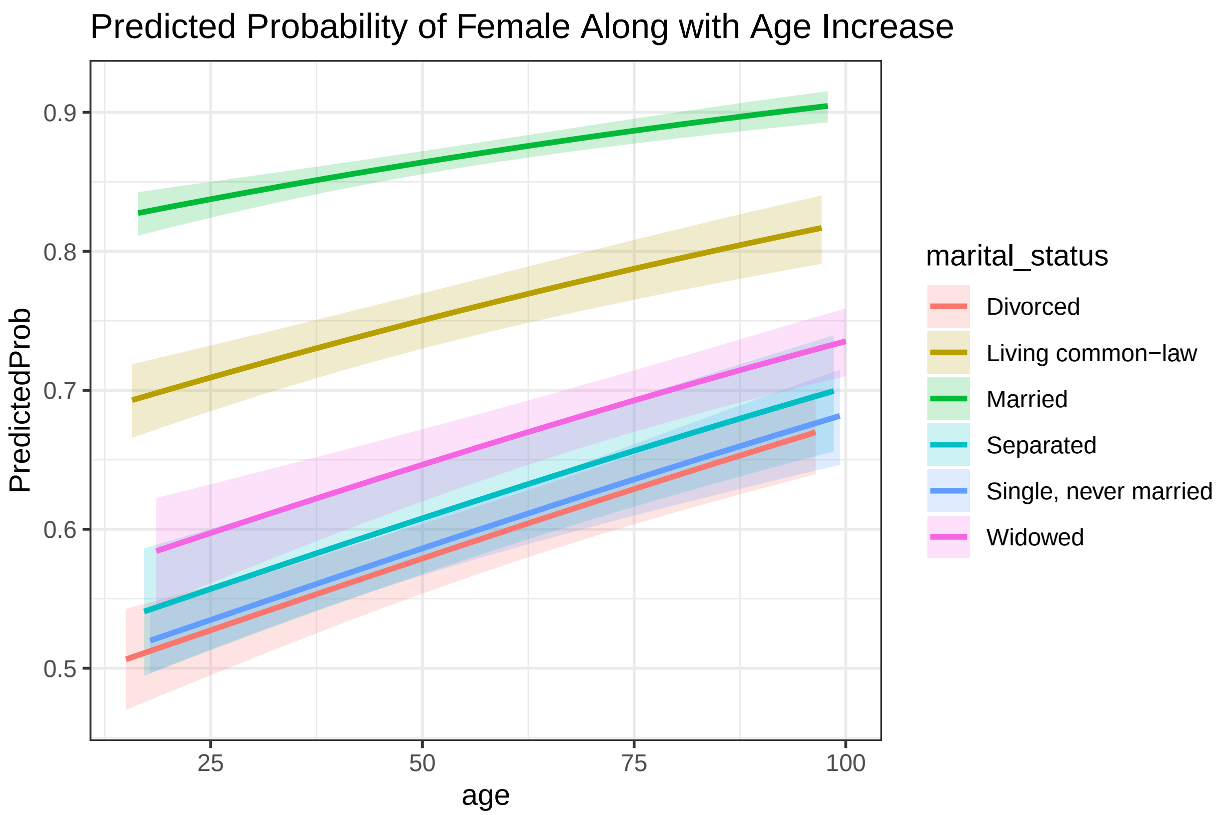
```
##   own_rent      age    sex        marital_status feelings_life total_children
## 1        0 52.12087 Female              Divorced      8.096046       1.675046
## 2        1 52.12087   Male     Living common-law      8.096046       1.675046
## 3        0 52.12087 Female               Married      8.096046       1.675046
## 4        1 52.12087   Male             Separated      8.096046       1.675046
## 5        0 52.12087 Female Single, never married      8.096046       1.675046
## 6        1 52.12087   Male               Widowed      8.096046       1.675046
```

Now that we have the data frame we want to use for calculating the predicted probabilities, as we let R Studio to create the prediction.

Predicted Probability of Male Along with Age Increase

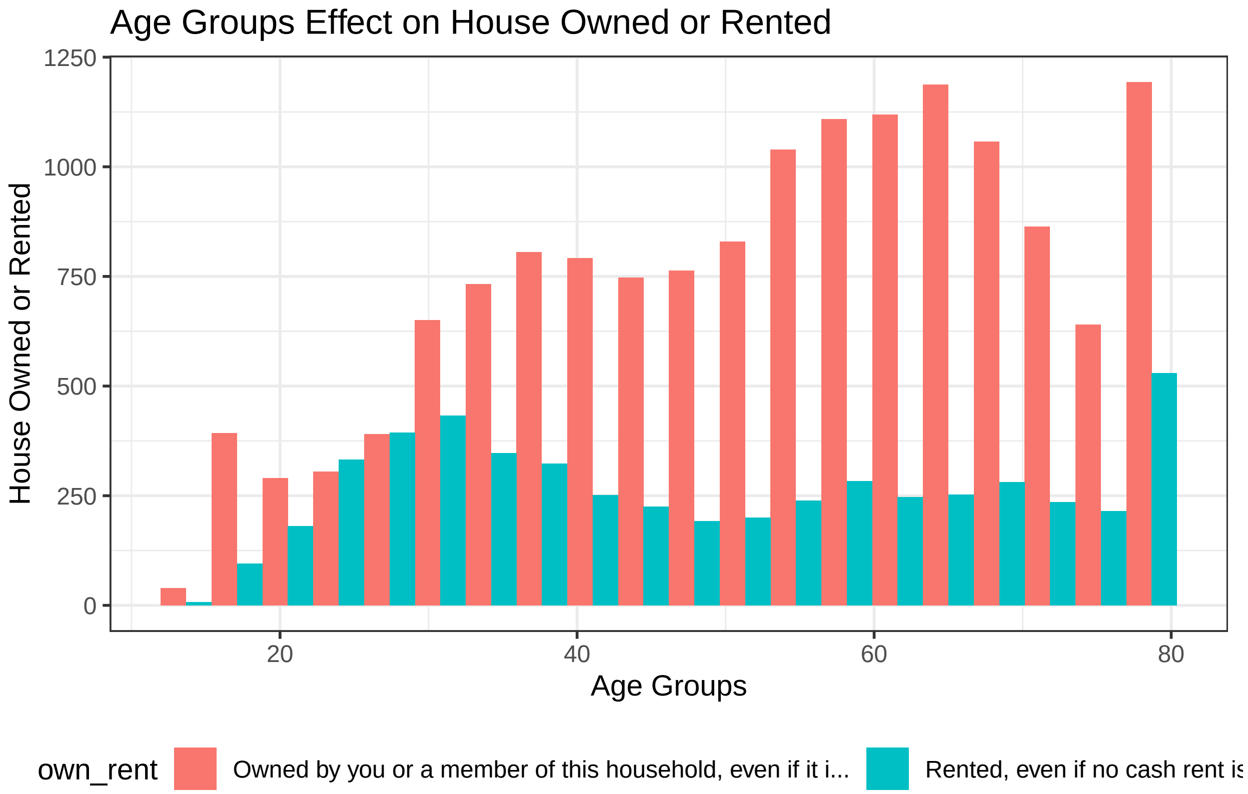


ce: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019



ce: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

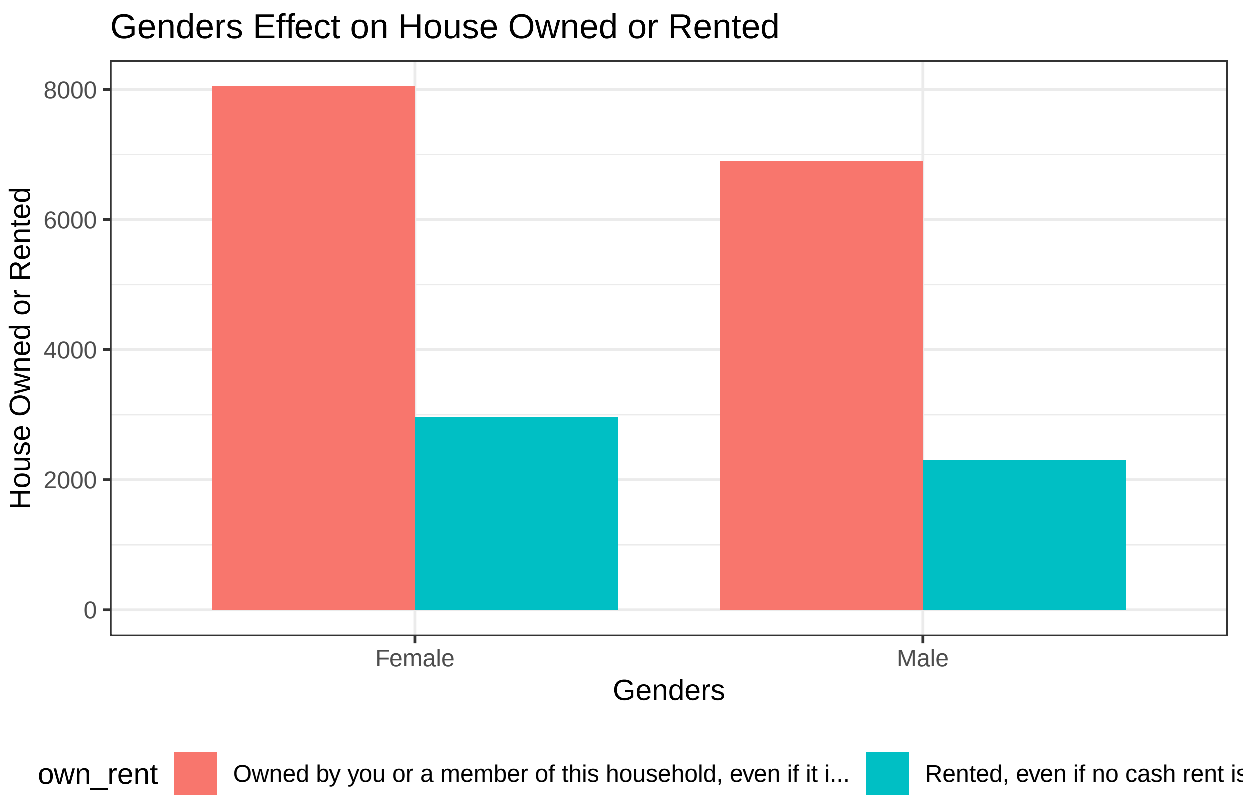
Result



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 7: Age Groups Effect on Owned or Rent a House

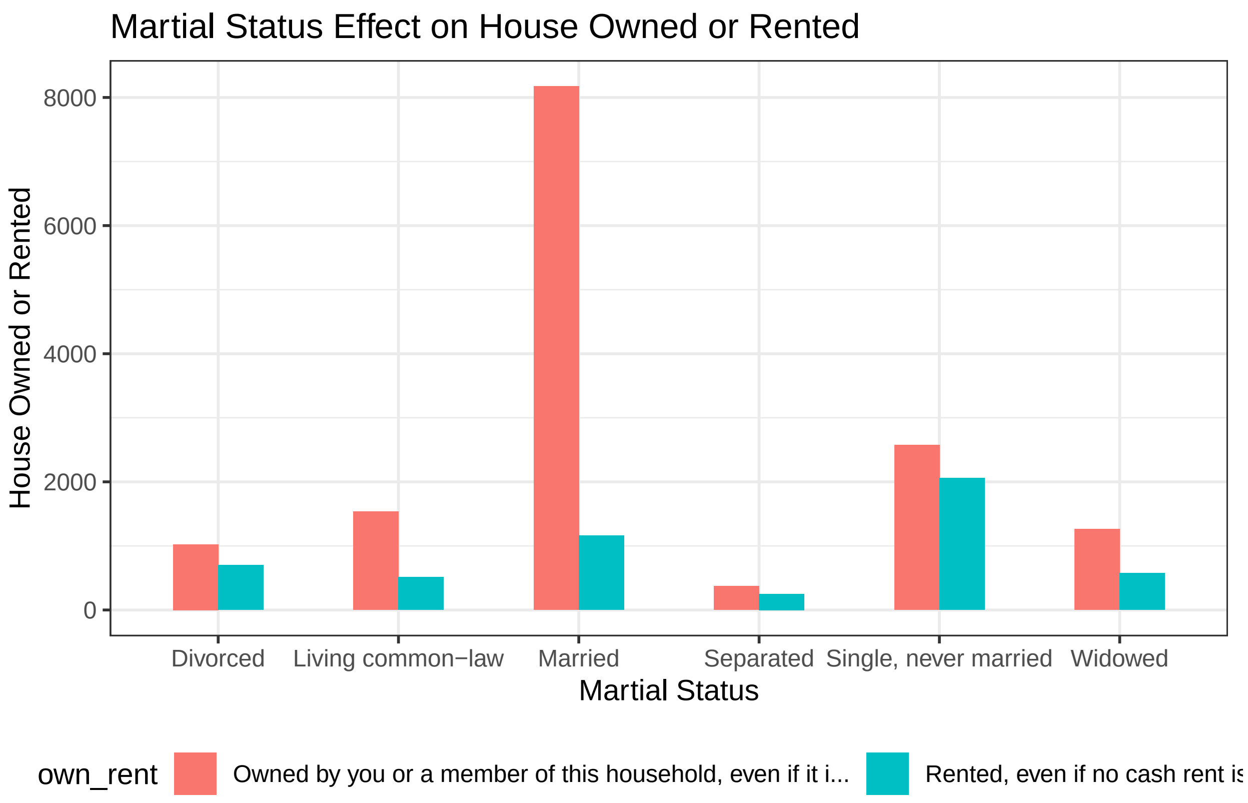
We can see from this picture that there are more people of all ages who own houses. The number of people who own houses becomes larger as age increases. Among 25-28 of age, the proportion of people who rented houses is larger than the other working age groups. The trend declines after the age of 28, but the proporion of seniors who rent for houses is the highest among all age groups.



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 8: Genders Effect on Owned or Rent a House

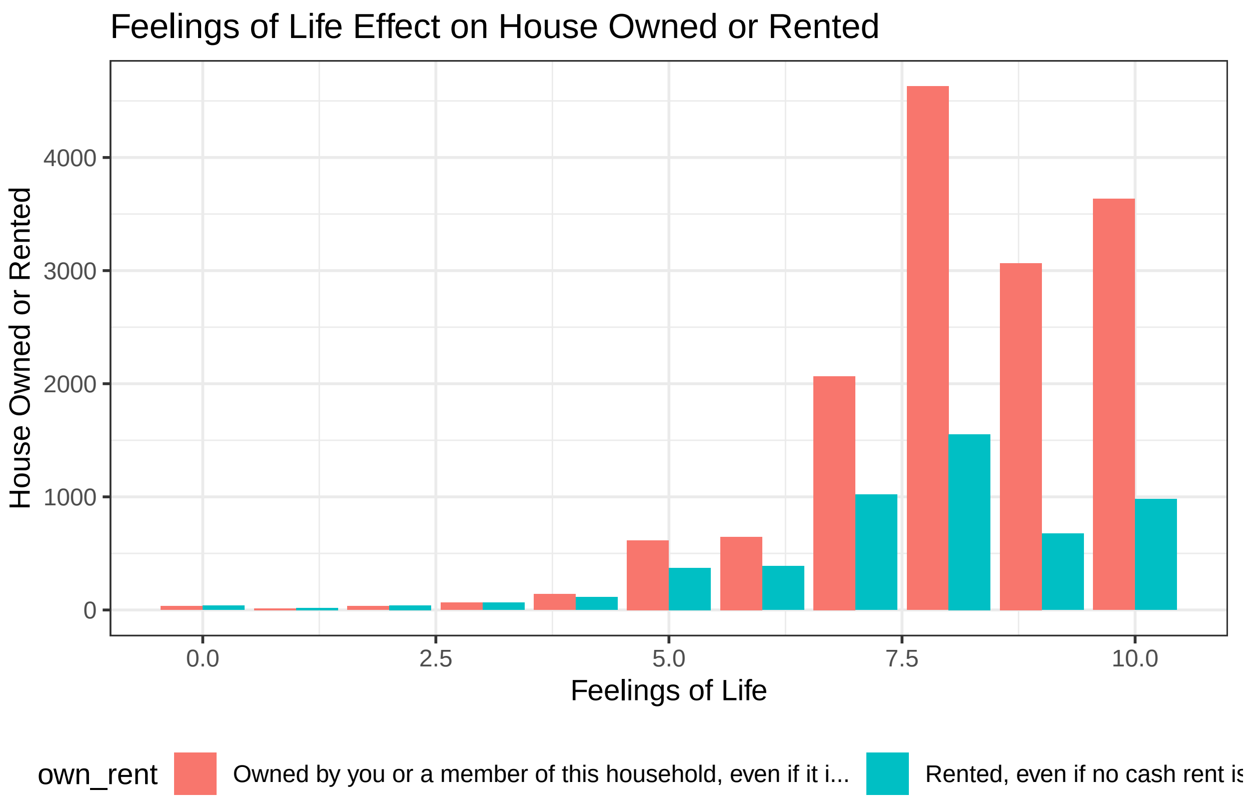
In this picture, we find that no matter what gender it is, the proportion of owning a house is greater than that of renting a house. Compared with men, females own and rent more houses than males



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 9: Martial Status Effect on Owned or Rent a House

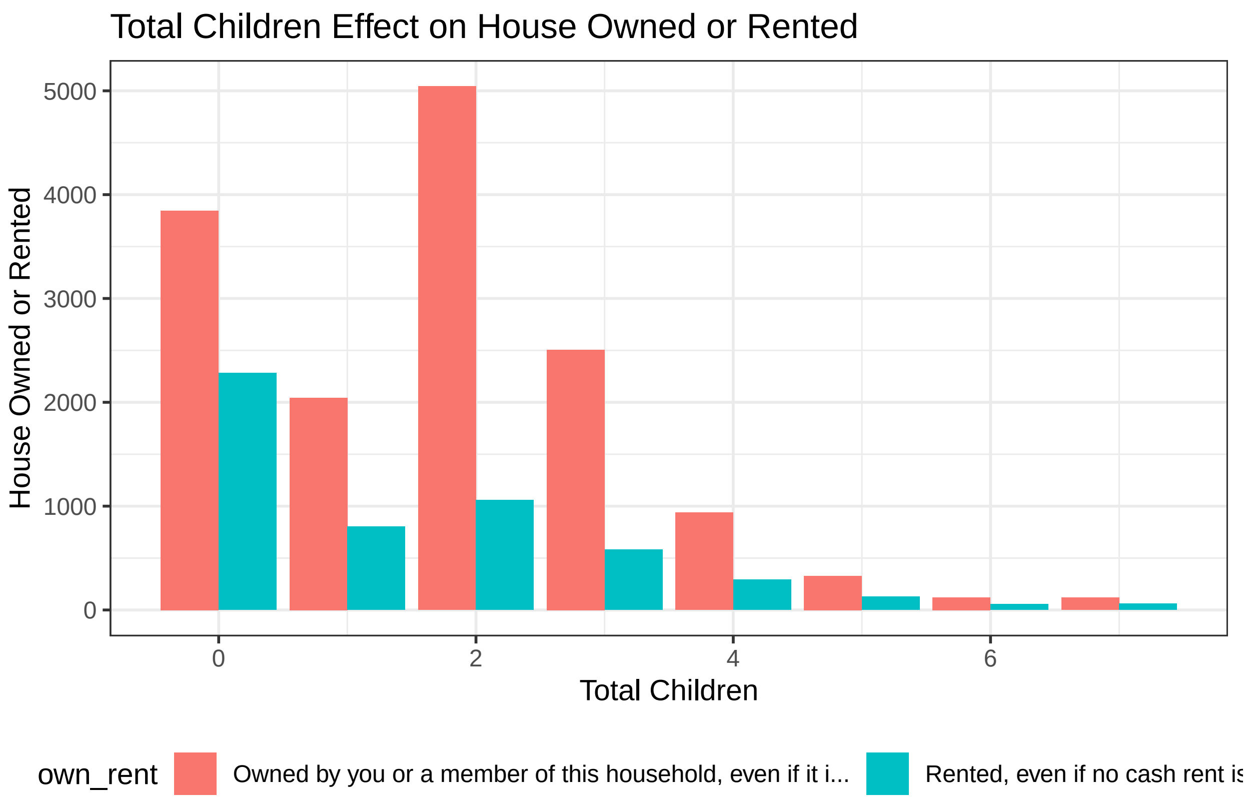
This picture mainly describes the relationship between marital status and housing ownership. We can see that no matter what the marital status, the number of houses owned is greater than that of rented houses. The marital status is married, and the proportion of owning a house is very high, which greatly exceeds other situations, and separated has the smallest proportion.



Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

Figure 10: Feelings of Life Effect on Owned or Rent a House

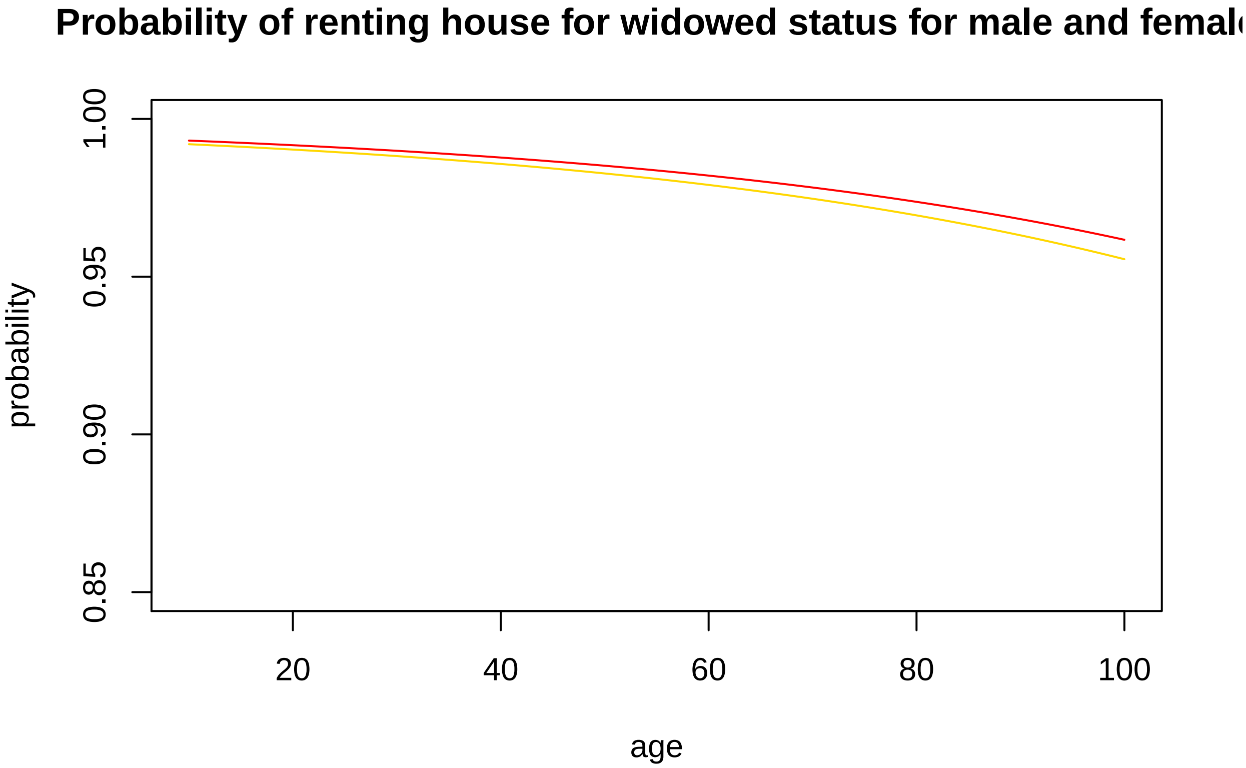
We can see that in the range of feeling of life from 0 to 3, the number of houses owned is similar to the number of houses rented. Starting from 4 points, there are more houses than rented houses. At 7.7 points, the number of houses owned and the number of rented houses reached the maximum, then fell back, and rose again at 9 points.

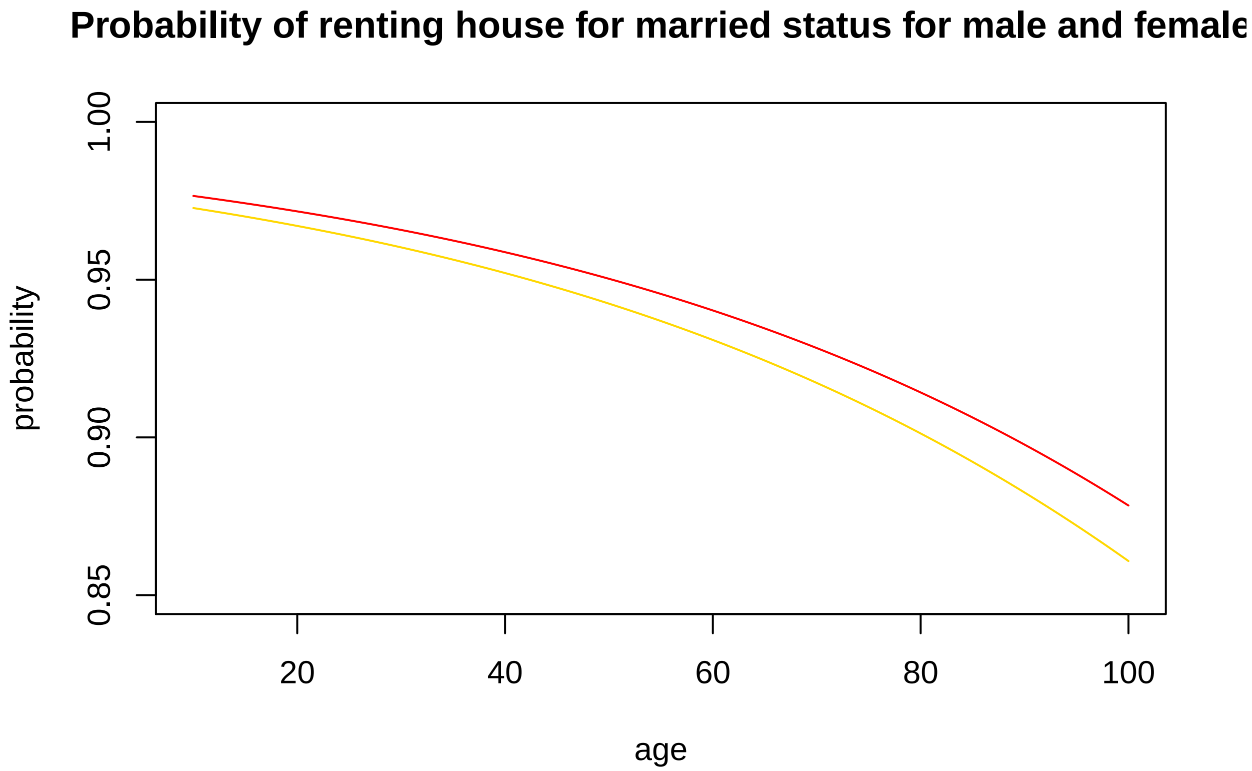

Total Children Effect on House Owned or Rented

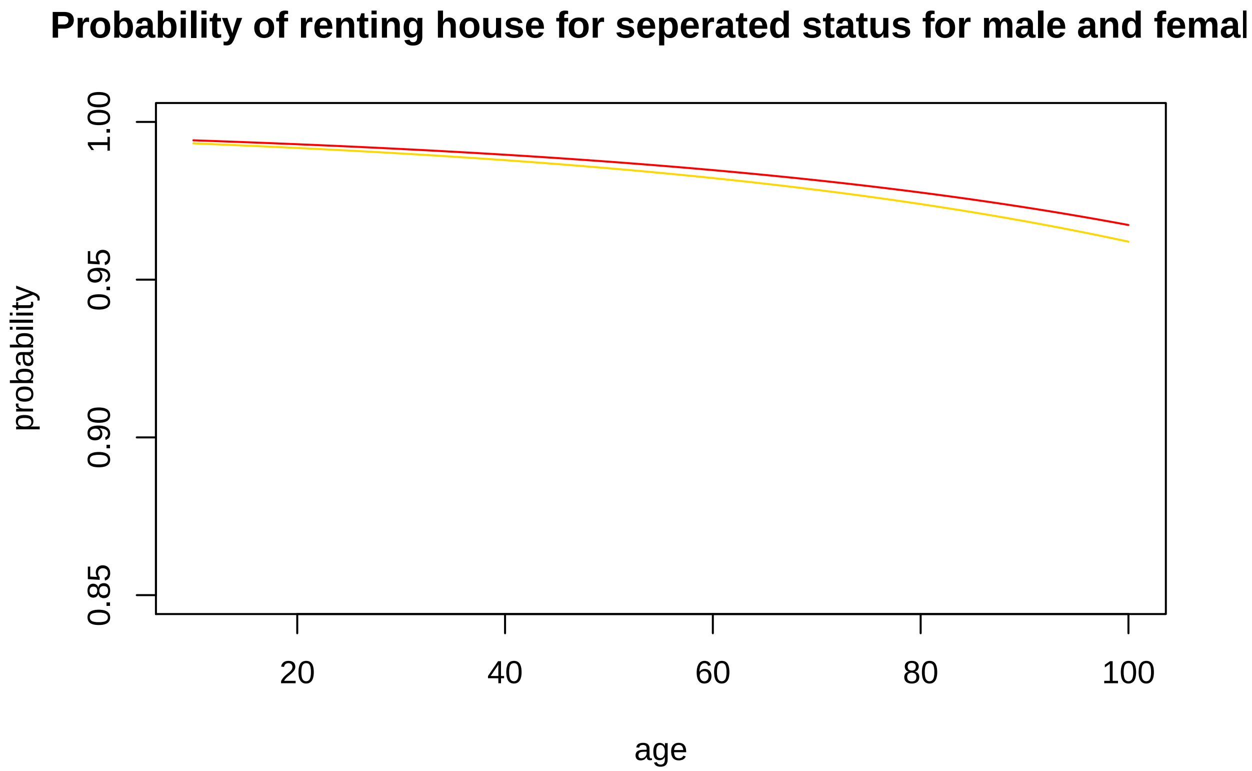
Source: Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., 2019

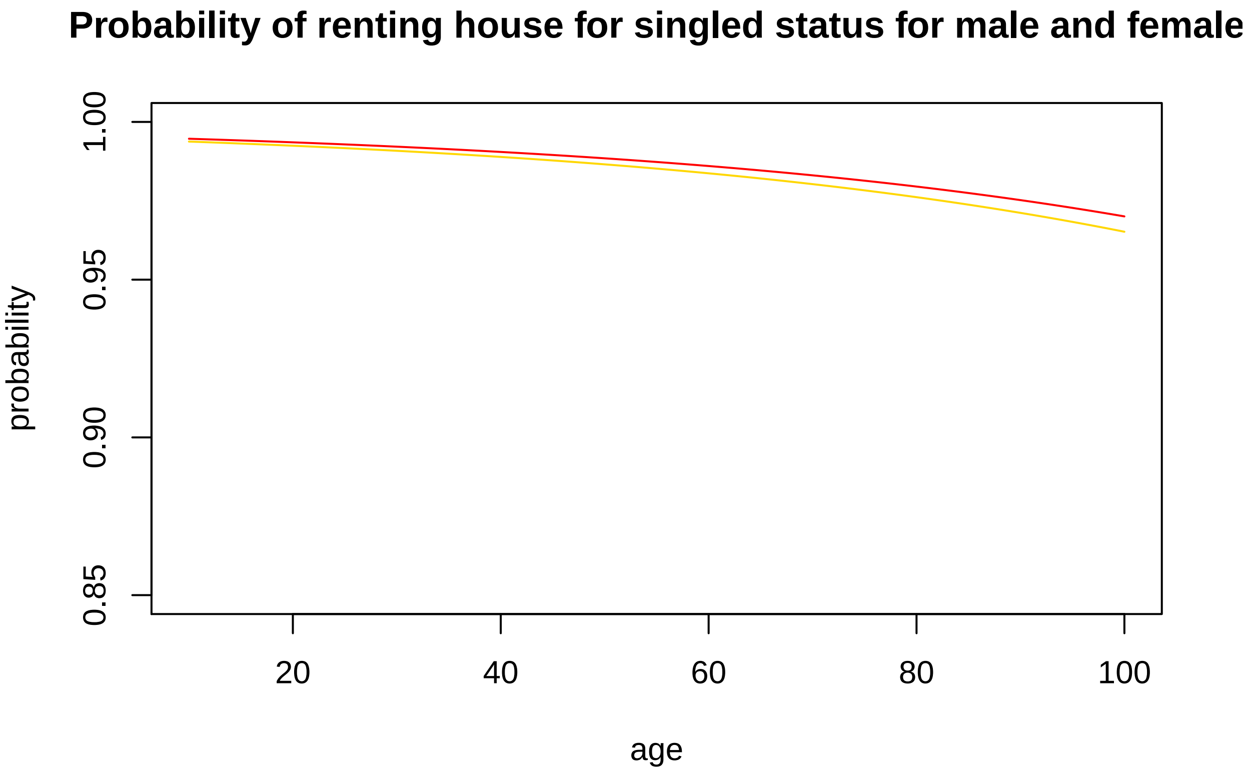
Figure 11: Total Children Effect on Owned or Rent a House

On the whole, no matter how many children you have, there are more houses owned than rented houses. The number of rented houses peaked when there were 0 children and then began to decline. In families with two children, the number of rented houses rebounded and then gradually declined. In a family with two children, the number of home ownership peaked and then slowly declined.


Probability of renting house for widowed status for male and femal
probability
1.00
0.95
0.90
0.85
20
40
60
80
100
age


Probability of renting house for married status for male and femal
probability
1.00
0.95
0.90
0.85
20
40
60
80
100
age


Probability of renting house for seperated status for male and femal
probability
1.00
0.95
0.90
0.85
20
40
60
80
100
age



From the model that was built, a comparison between male and female is conducted. To make this model simpler, an assumption of no children condition and no feeling of life is made. So basically, these comparisons are about the probability of renting house for male and female as age increases controlling for the same marital status.

From the above four graph, both male and female follows the same decreasing trend as age increases. This makes sense, since as age becomes older, people tend to have enough money to afford their own house. And also, older people will be less likely to enjoy renting the house than the young people.

Besides, by comparing the decreasing trend for all four graph, a significant drop in probability is found in the married people. This is also similar to what people would expect. Since the married people will have more likelihood to consider about having a house for the family or the baby. And two people together will be more likely to afford the house. Moreover, the rank for probability of renting house is single, seperated, widowed and married from highest to lowest.

Now, considering about adding the other two factors, the feeling of life and having children. Without drawing the graph, an even deeper drop in probability would be expected. Since the higher life satisfaction feeling tend to be more desired for buying a house. On the contrast, the lower life feeling will lead to not enough ability to afford a house. For the families with children, they tend to need their own house then the other families as well. These conclusions can be proved by looking at the logistic regression. All the coefficients are negative, which suggest a negative relationship between the probability and these factors.

Discussion

Dataset and Result

The dataset is from GSS. We used the data cleaning code from authors: Rohan Alexander and Sam Caetano. And further more, we filter out the "NA" from the dataset. We build up the models of house owned or rented against every single other variables, and we select those that have stronger correlation with house owned or rented. Which are age, gender, martial status, feelings of life and total children. From the result, we can conclude that age has a positive impact on owning a house; male has a little bit advantage than female for owning a house; martial status as married have significant positive impact on owning a house, the others are less impacted or has a negative impact on owning a house; better feelings of life have greater positive impact on owning a house; and last but no least, increasing number of total children has slightly negative impact on owning a house. The results we found match with the logistic model.

Weaknesses of The Survey and Sampling Method

The questionnaire is long and it is only expose to telephone users, which can be associated with non-response error. And portion of the respondents might hesitate to reveal their privacy such as their income or housing information. There is a chance for respondents to reply false information, which could be result in response error. Also there is possibly a selection bias in the sampling method. Contact with telephone users would increase the likelihood of someone that owning or renting a house. This will greatly affect the accuracy of the model because the dataset is skewed to people that already owing or renting a house.

Weaknesses of The Model

Although the logistic model shows a correlation between dependent variable[house owned and rented] vs. independent variables[age groups, genders, martial status, feelings of life and total children]. However we can not conclude the cause and effect relationship such that those independent variables might not be the causes for the dependent variable. From Figure.10, we can see a positive relationship between feelings of life vs. house owned and rented. Which implies people have better feelings of life are more likely to own or rent a house. Feelings of life is the cause for owning or renting a house. However in reality, people owned a house are tend to have a better feelings of life, because they have less financial pressure. This proves that there might be reverse causation in our model. In another word, owning a house is the cause for better feelings of life instead of better feelings of life is the cause for owning a house. Also house owned or rented vs. total children might also be a reverse causation. People owning a house are possibly wealthier. They are less worry about having more children due to their financial situation.

Another weakness of the model might be we missed some important variables. Some variables might have weak correlation to housed owned or rented, but they can be the determine factors in reality. Also we exclude income in the model because it makes the model much more complicated and we have to filter out a lot of data due to the 'NA' from respondents. Yet, income can be more significant than any other variables in our model. Income itself is sufficient enough to form a model with owning or renting a house.

Appendix

Github link : https://github.com/ChephonN/hihihi

References

- Data cleaning and dataset was used from : Authors: Rohan Alexander and Sam Caetano Contact: rohan.alexander@utoronto.ca Date: 7 October 2020 License: MIT
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Lesnoff, M., Lancelot, R. (2012). aod: Analysis of Overdispersed Data. R package version 1.3.1, URL http://cran.r-project.org/package=aod
- Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
- Significance Testing and Confidence Intervals Author(s):David M. Lane
- Smith, Tom W., Davern, Michael, Freese, Jeremy, and Morgan, Stephen L., General Social Surveys, 1972-2018 [machine-readable data file] /Principal Investigator, Smith, Tom W.; Co-Principal Investigators, Michael Davern, Jeremy Freese and Stephen L. Morgan; Sponsored by National Science Foundation. –NORC ed.– Chicago: NORC, 2019.