Young people with low income or attending school are prefered living with parents

Xiye Zhong, Jiesen Cui, Ruihan Zhang

2020/10/19

#Abstract:

Living with parents was always a complex and sensitive topic, sometimes living with parents was seen as the symbol of immature and may cause many family issues, sometimes living with parents can reduce the pressure of living and economics, however, there were 9% of Canadian adults(aged 25 to 64) live with parents in 2019 and has been since 1995. In this report, we are studying what factors aspect the families living arrangement and what kinds of adults or families would rather to live with parents, such that, we selected some variables we thought may be related to living arrangements based on data from the General Social Survey (GSS) on Family, then analysed the data by using a binomial logistic regression model and expressed the probability of living with parents in graph. We found that age, income level, education and language(culture) all significant to living arrangement, furthermore, the young adults below 30 years old are more likely to live with parents and the adults with low income or had lower education may prefer to live with parents. About 4% (1 million) Canadians said that they considered moving back with parents, this research may help the people who are hesitating whether to live with parents to make their decision and may help the government predict the pension pressure.

This file and data cleaning code can also be found at https://github.com/Xiye-Zhong/Analysis_GSS

#Introduction:

More and more Canadian adults considered living with parents under the pressure of house price. The rent for an apartment was not a small cost for young people, especially during COVID-19. About 2.8 million (one out of ten Canadian) changed their living arrangement under the pressure of COVID-19 (Frank O'Brien,Business in Vancouver,JULY 24, 2020 08:39 AM). For the married people, parents could help to company children such that the spouse had more time to focus on work. "It is safe to assume that many of the parents who moved in with their adult children are also grandparents who are helping to provide childcare for exhausted working parents of young children, who have limited or no childcare options until school begins," said by Scott Birke, publisher at Finder.com. Living with parents was also a good way to care about aged parents. If the elderly parents were uncomfortable, their children could help their parents get treatment in time. Although living with parents had many advantages, it was a complex issue for most people. Not all Canadian fit for living with parents, it might lead to arguments for some families. Culture, education, income level or household size may all be related with living arrangement. In this essay, we aimed to find out what factors would affect the living arrangement and what kind of families were more likely to live with parents(The Daily, 2019-02-15).

We obtained the data from the General Social Survey (GSS) and selected nine factors such as age, family income, marital status etc. and built a logistic regression with living arrangement. The target people of this data covered ten provinces in Canada who were older than 15 years old. The actual number of respondents was 20,602 while the planned sample size was 20,000. This survey frame combined telephone number and registered address. This survey used a simple random sample without replacement as the sampling method and collected data by telephone interviews. After using binomial logistic regression to analyse the selected data, we found that age, income level, education, sex, household size and language were all significant related to living arrangement.

To test the goodness of fit for the model we used, we tested the AIC value and built a AUC curve. The result showed the model we selected was fit for our analysis. After that, we studied the probability distribution of living arrangement for various significant variables. However, this research still had some shortcomings. For example, most people over 69 years old don't need to consider living with parents, we could remove this group of data next time to reduce the cost of survey. In addition, the language at home cannot completely stand for the culture difference. We should collect more variables about culture if we designed the survey by ourselves.

We hope this essay could help hesitate people to make a better decision on living arrangements. and for the families that do not fit for living with parents, we hope our study could help them solve their problems earlier. For the next step, we want to explore the advantages and disadvantages on living arrangement issues for different types of families.

#Data:

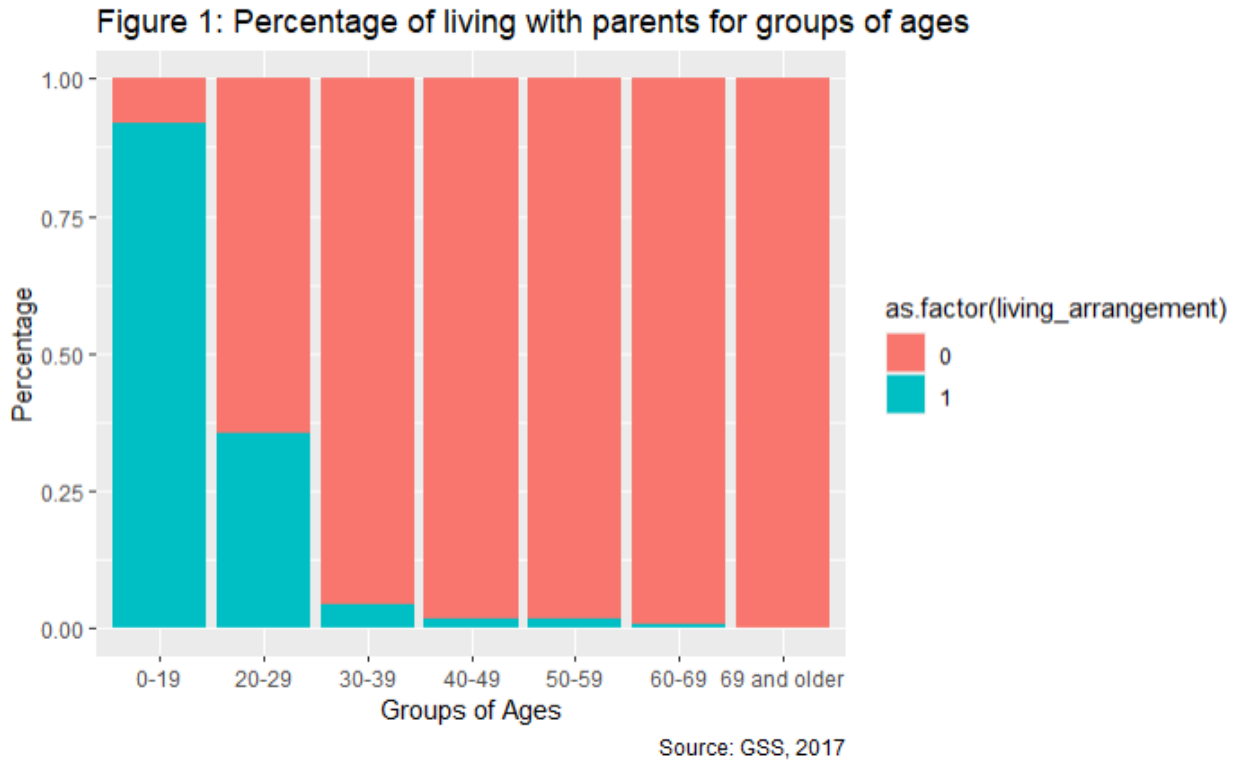
This dataset is produced by the General Social Survey(GSS), Statistics Canada in 2017. It is designed as a cross-sectional sample survey, whose target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada(Yukon, Northwest Territories, and Nunavut are excluded). The frame of this survey combines telephone number and registered address, the sample size is planned to be 20,000 while the actual number of respondents is 20,602. The staff of this program construct several linked sources such as bills rather than just phone numbers, which makes the telephone numbers eligible to reach a valid address.(91.8% of the selected telephone numbers reached eligible households) Moreover, those respondents who are not eligible would be dropped after an initial set of questions. Those policies increase the coverage of this survey and make it more accurate.

The sampling strategy this survey uses is a simple random sample without replacement and the overall response rate is 52.4%. The collection process is an interview via telephones, they would re-contact those who initially reject to participate in this survey and reschedule the interview for those who do not have a convenient timing. Meanwhile, for the cases where there is no one at home, there would be numerous call backs.

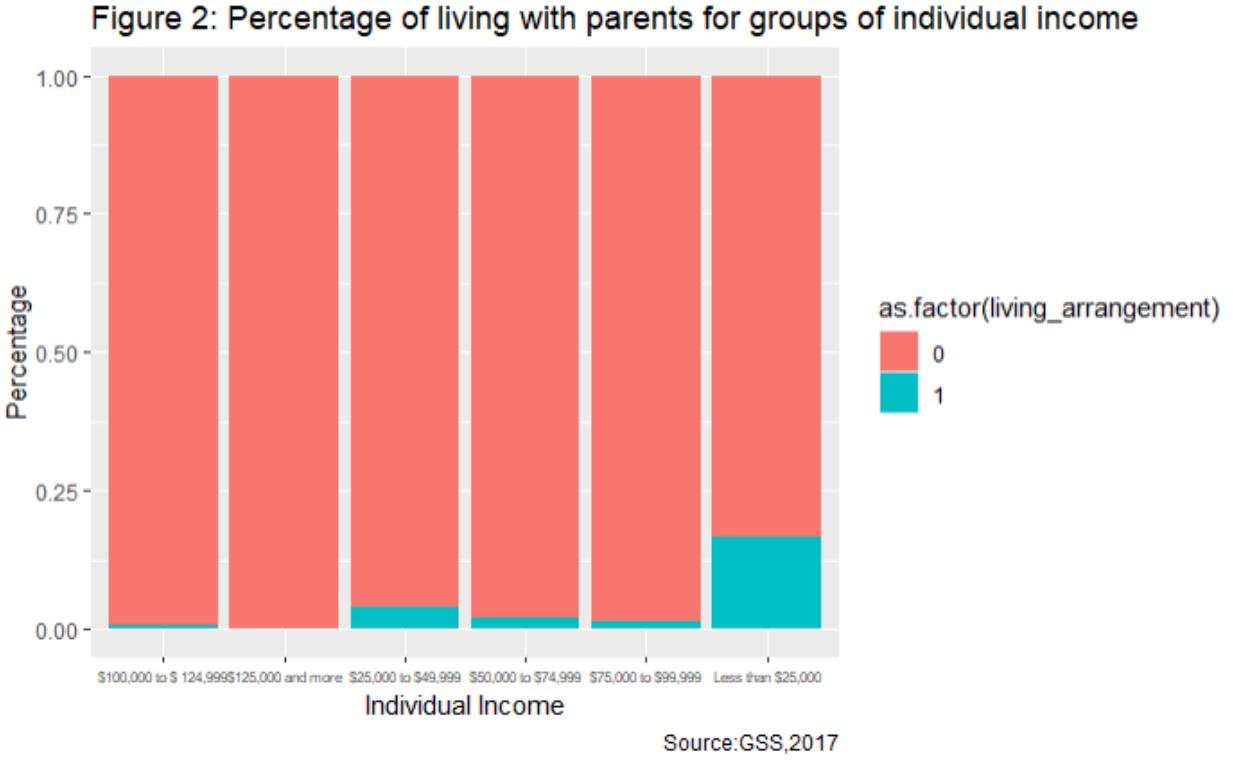
There are 14 key concepts in this survey which contains the basic information, health, marital status, financial and education information, social and family information for the respondents and their families. This survey is detailed and the questions are clear, it makes the response rate high and the data reliable. However, there are 559 cases where the respondents are under 19, which means they are unlikely to build their own family. Those cases may affect the result of our model.

We are interested in whether people live with their parents. As mentioned before, only 9% of Canadian adults choose to live with their parents. We consider several variables that may affect. The first one is the ages, for those who are under 19 are more likely to live with their parents while those who are over 70 are very unlikely to live with their parents. The second one is sex, we are interested in whether the difference in gender would affect the model. The third variable we select is education, we think different educational backgrounds may result differently. The situation of whether owning real estate is another variable, people who do not have their own house may prefer to live with their parents. The house type and house size are also chosen since people may live with their parents if they have a big house. Culture may also be a factor. Hence, we also consider the languages used at home as a cultural background. Since we are interested in the relationship between the individual choice and individual information, we do not choose the other variables like family income or partner's information.

Among those variables we choose, we think age, education, individual income and culture are the most significant factors. Hence we plot bar charts to show the percentage of people living with their parents in groups.

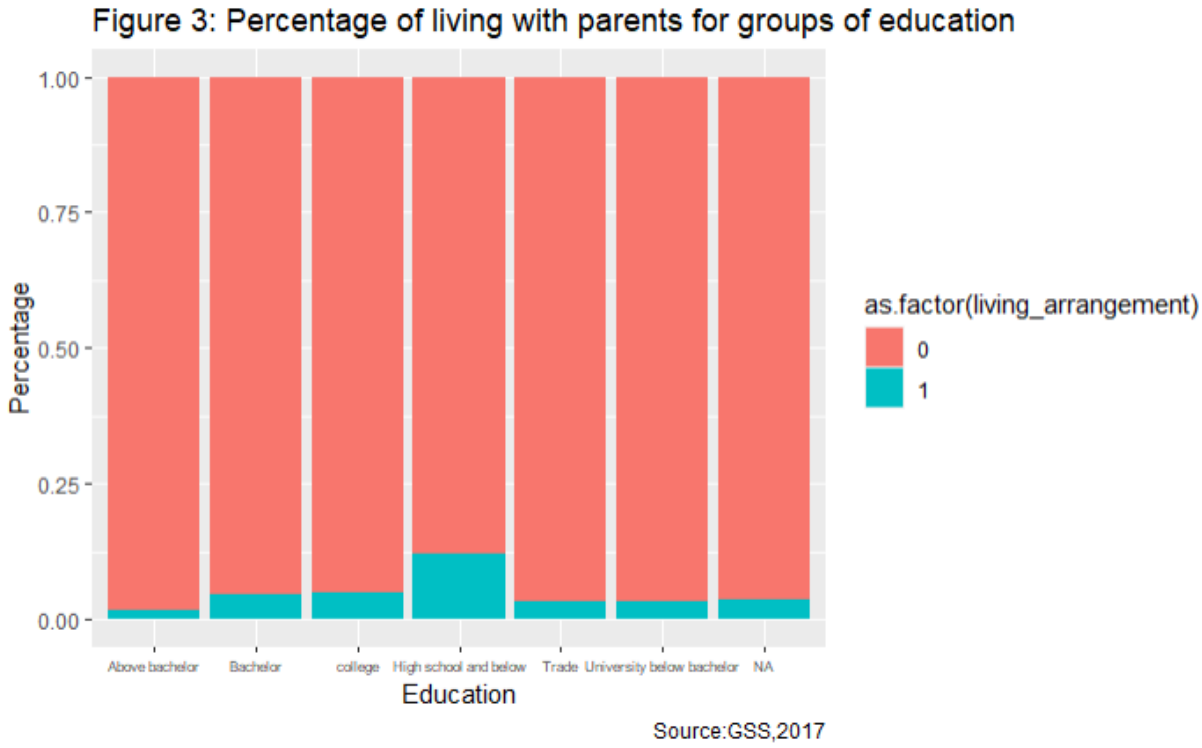


As figure 1 shows, most of the cases who are under 19 are living with their parents and there are also plenty of people who are 20-29 choosing to live with parents. For those who are over 40, they almost do not live with parents.

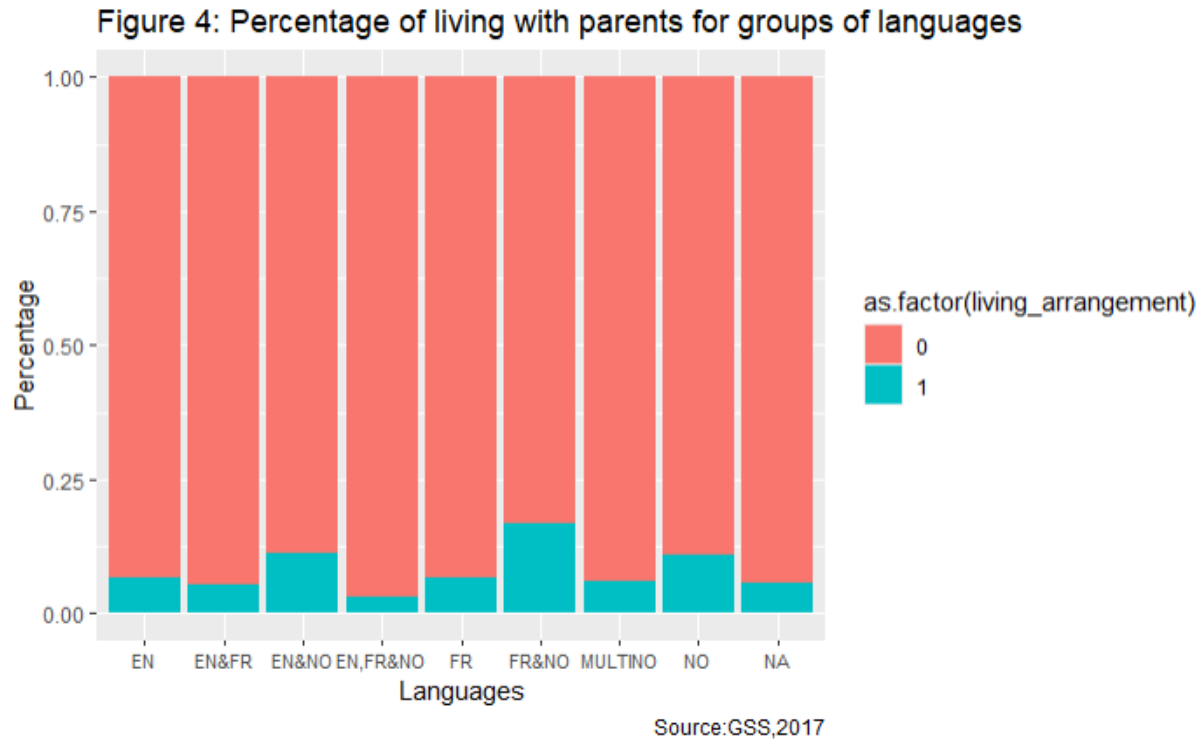


In figure 2, we can see that people whose individual income is lower than $50,000 are more likely to live

with parents.



In figure 3, the percentage of living with parents is highest for those who have a high school and below degree. On the contrary, people who receive a degree above bachelor are those most unlikely to live with parents.



As figure 4 shows, the different cultural backgrounds result in differently.

#Model:

The model we used to fit our data and target is the logistic regression model in general linear model. The logistic regression model can be used when the data is categorical. Our target is to find the factors which influence whether living with parents. For our target variable(living arrangement), it's a binary data, "yes" stands for living with parents, and "no" stands for not living with parents. For other variables we selected, we treat them as independent variables, and then we used the binomial logistic regression, to predict which independent variable may influence the living arrangement. In a general linear model:

$$y_i = \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i$$

where Yi is the response variable,beta(i)..... beta(k) is unknown parameters, x is modelled by a linear function of explanatory variables xi , i = 1, . . . , k , and epsilon(i) is the random error term.

In a logistic model, Consider a simple k variable regression model where k = n +1, whose general form for the general regression model is given by:

$$Y = \mathrm{f}\left(\beta_0 + \sum_{i=1}^{n}(\beta_i x_i)\right)$$

$$Ln(P/1-P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where beta(0) has a constant value, and the beta(i). for all i = 1,2, . . . , n are the estimated weights of Xi., the transformed raw data. For the second formula,it is a link function. We replaced Y for ln(p/1-p), and p stands for the conditional mean of Y. In our model, Y represents the variable living arrangement, and beta represents the independent variables we selected such as age, individual income..... Our purpose is to find the factors that make people live with their parents. For Y, because we are fitting our data to a binomial logistic model, which requires the dependent variable to be a binary data. For beta, this model does not have any specific data category limits for beta, so when we process independent variables beta, we did not change the category of them. Therefore, we selected the binomial logistic model to fit our data. To further test the goodness of fit for our model, we used AIC value to compare with the alternative model and ROC curve to test it. Our alternative model is poisson regression, because our variable is numeric and continuous, poisson regression can fit the situation when one variable determines whether or not an event happens at all and another that determines how many times the event happens when it does. In our situation, the other independent variables, for example : age, family income, .,... can determine the event of living with parents.

With our data, we run two different models in R studio and want to find whether our first choice binomial logistic regression is the most suitable model for our analysis.

The AIC value for binomial logistic regression is 4909.4, and the AUC value is 0.9369; and for the poisson regression model, the AIC value is 6282.7 and AUC value is 0.9377. When we are comparing the goodness of fit for two models, AIC value is a distinguishing result we can use, and if the AIC value is smaller than another one, it means the model is better fitted for our data. For the AUC curve, the range for the AUC curve is 0.5 to 1. It means how accurately our model can predict the data. In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

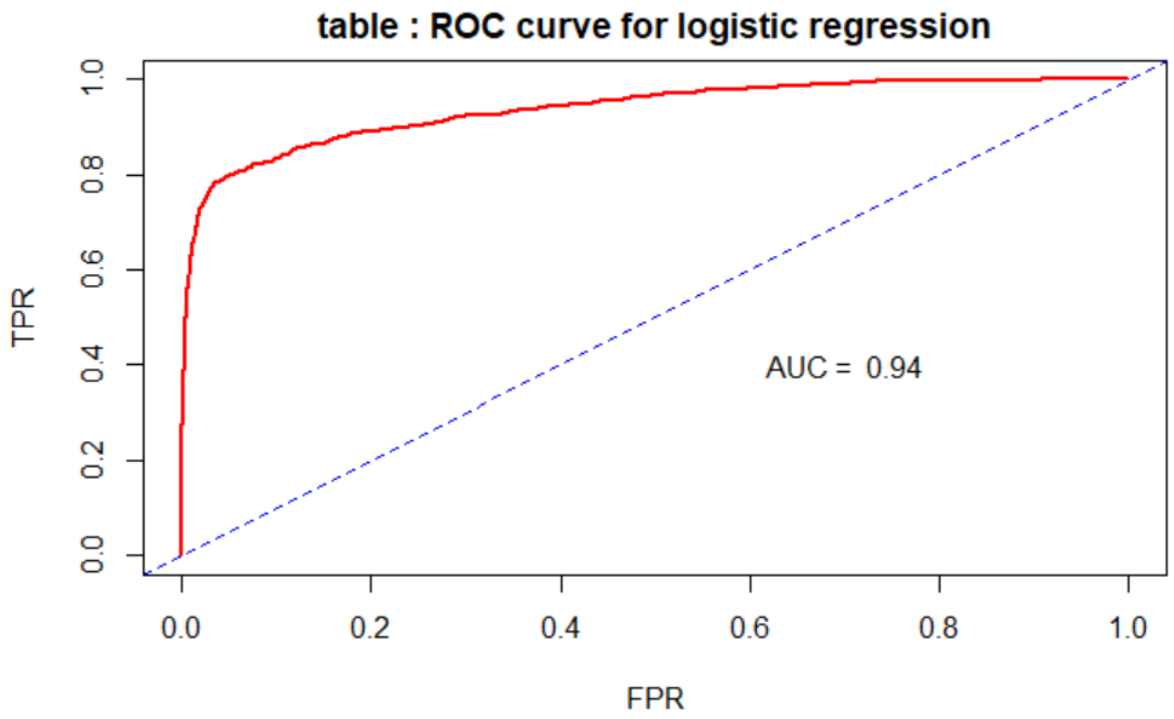


Figure 5:the ROC curve for binomial logistic regression

From Figure 5, we can see the ROC curve is close to the left corner, and AUC value is close to 1, which means our binomial logistic regression model fits the data well.

In the models we selected, the AIC value for binomial logistic regression model is smaller than poisson regression model, which means the binomial logistic regression model is better fitted to our data; the AUC value for two model are both outstanding, it is hard to say which one is better. Therefore, the final model we selected is the binomial logistic regression model.

For our model, there are some weaknesses. The binomial logistic regression is also a generalized linear model, so when this model processes the variables that are not completely independent, it tends to have a poor performance. Overfitting is also a problem of our model. In our further work, we will consider cancelling some noisy data, or try to select a smaller part of our data to fit the model, in order to cope with the weaknesses.

#Result:

Variables	P-value	Estimates
age20-29	< 2e-16	-2.179232
age30-39	< 2e-16	-4.981439
age40-49	< 2e-16	-6.163973
age50-59	< 2e-16	-6.06725
age60-69	< 2e-16	-6.684792
age69 and older	< 2e-16	-9.186724
sexMale	5.75E-09	0.467536
educationHigh school diploma or its equivalent	4.52E-08	0.639959
educationLess than high school diploma or its equivalent	8.36E-06	0.757221
educationUniversity certificate, diploma or Bechalor above	0.006345	-0.589088
hh_size	< 2e-16	0.366386
language_homeEnglish and non-official language	0.008985	0.391104
language_homeFrench and non-official language	0.040868	0.962021
language_homeNon-official languages	0.591805	0.626124
income_family$125,000 and more	0.000886	0.446869

Table1 : P value and estimates for significant variables

From table 1, we can see the P-value of the significant variables. After we fitted the processed data into our binomial logistic regression model, we found that there are 5 major variables that influence whether people live with their parents, for example the age, gender, education level, household size, mother language, and family income. Then we selected all significant variables and made the table.

The first row are the specific types of variables that affect living arrangement, and P-value represents whether the variable determines the living arrangement. And the estimates means the slope of the data.

Our result shows whatever age group you are, it is a significant factor to decide whether you live with your parents. From age 20-29 to 69 and older, different age groups show a different slope as age grows. For example, for people from the 20-29 age group, they have a slope of around -2.18, which means as age changes from 20-29, the possibility of them living with parents will decrease to -2.18. As age grows up, for 69 and older, the slope comes to -9.18, which is apparently different from the 20-29 age group. So age is a significant variable to determine people's living arrangement.

For education level, as we can see from the table, people with highschool, or less than highschool diploma will be more likely to live with parents. We analysed the possible reason, for people with less than highschool diploma, they are probably still going to school. They are not economically independent, they do not have a source of income. They have no choice but to rely on their parents and live with them.

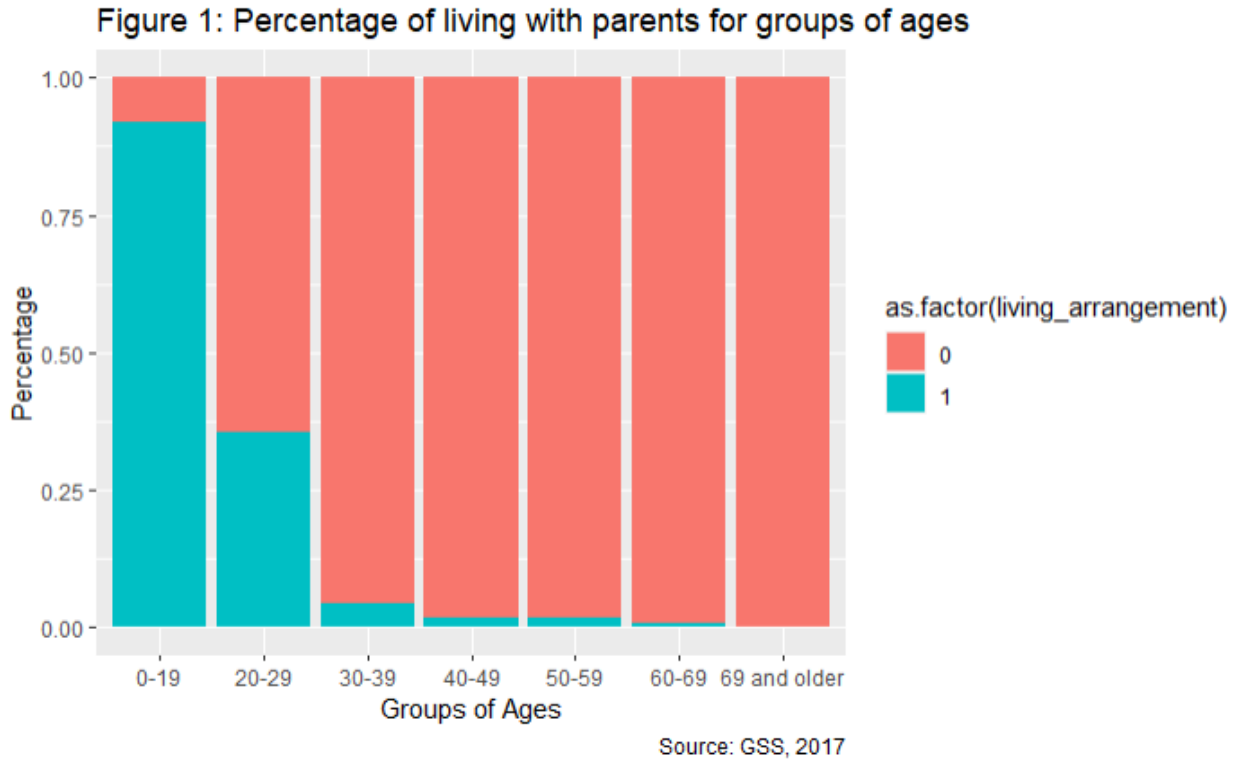
And for household size, it is a factor that influences the living arrangement as well. For people who have a small house or condo, they may not have extra money to live outside or rent a room for living, they have to live with their parents. And if people can afford a larger size of house or condo, they can choose to live with their parents or live alone.

For the language they use, it is about the culture. From our analysis, people with diverse cultural backgrounds tend to live with their parents. In some cultural traditions, it is an adult's responsibility to raise their parents, so they will live with their parents.

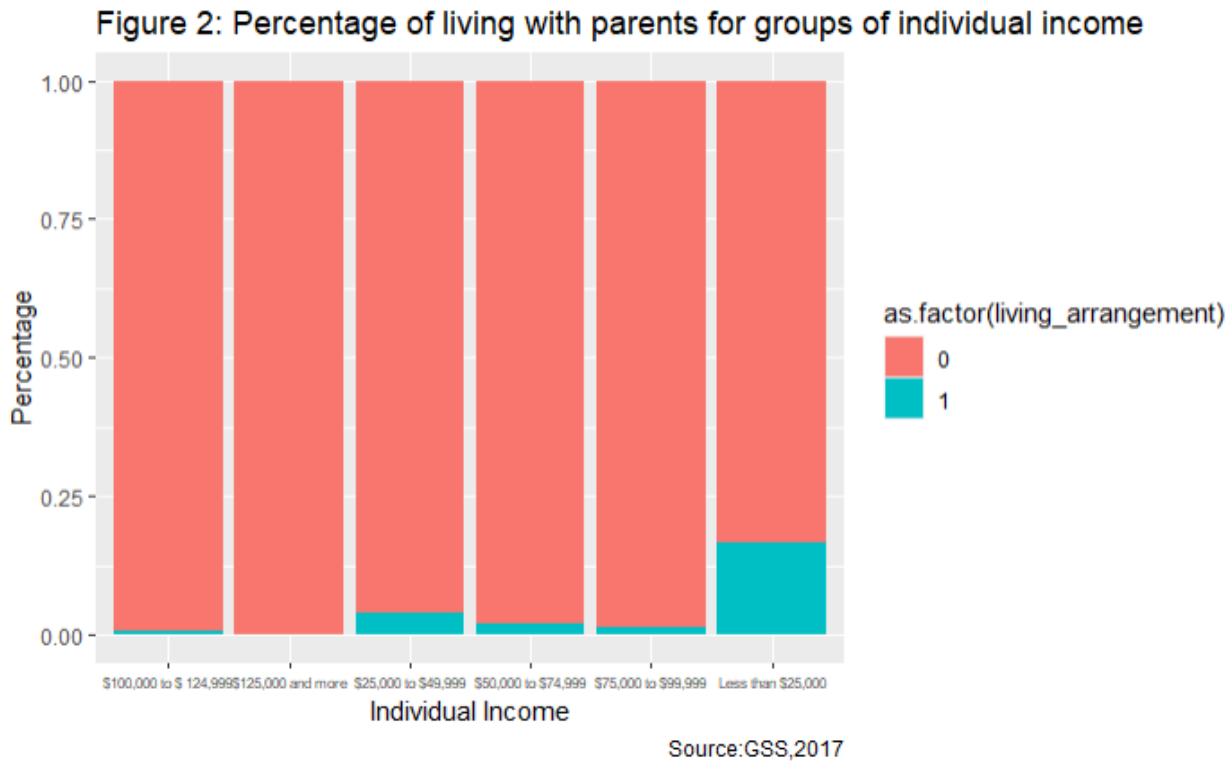
For the family income, we found a family's income above 125000 is a significant factor to decide their living arrangement. Based on our analysis, this group people are not likely to live with their parents. They are economically independent and rich enough to afford their own house and even to buy a house for their

parents.

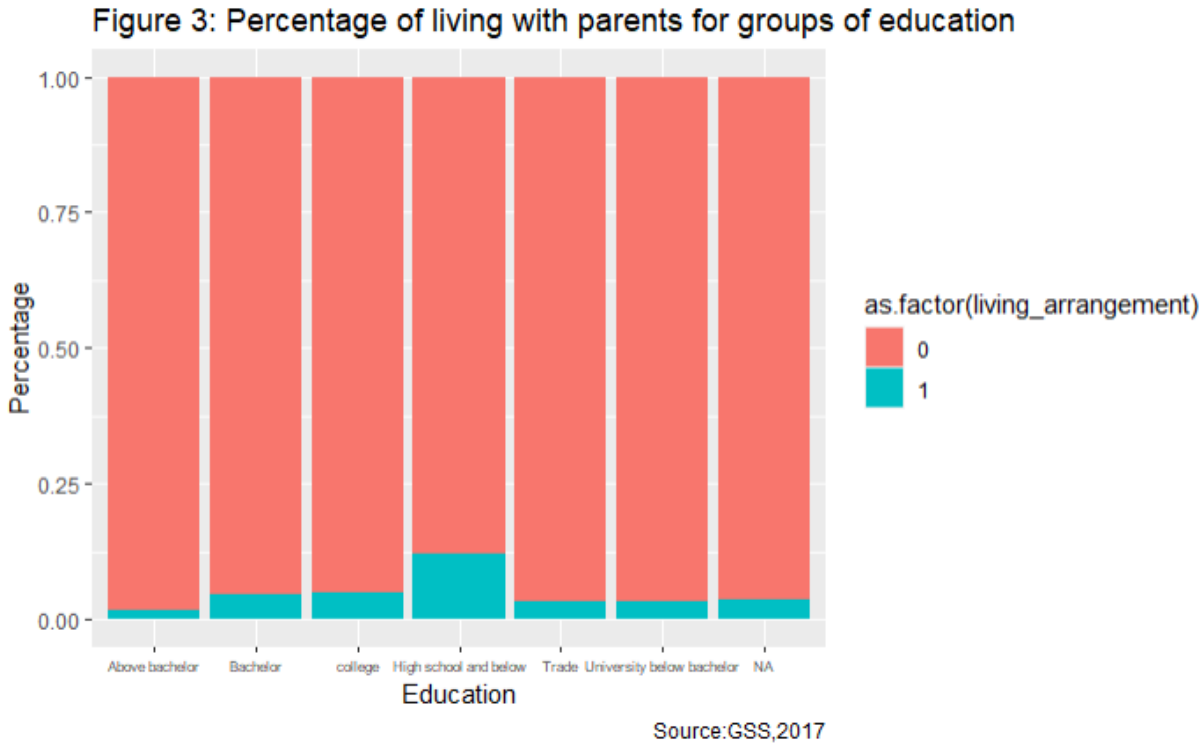
Among those variables we choose, we think age, education, individual income and culture are the most significant factors. Hence we plot bar charts to show the percentage of people living with their parents in groups.



As figure 1 shows, most of the cases who are under 19 are living with their parents and there are also plenty of people who are 20-29 choosing to live with parents. For those who are over 40, they almost do not live with parents.

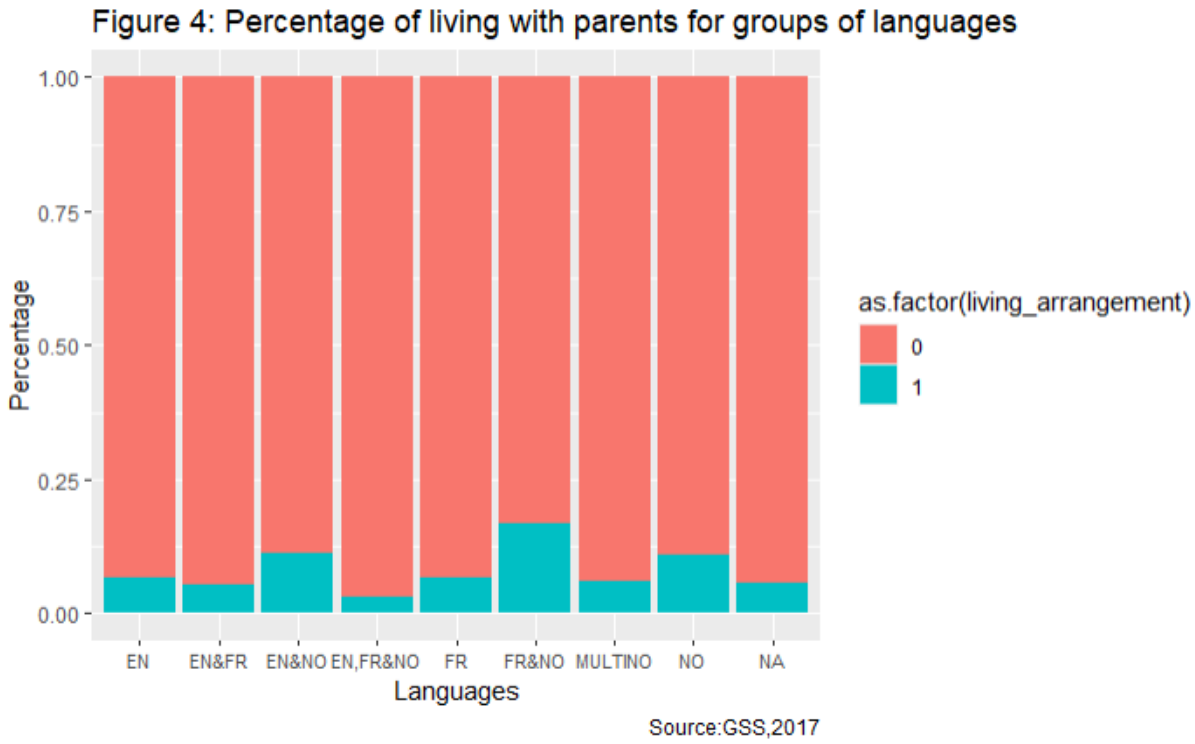


In figure 2, we can see that people whose individual income is lower than $50,000 are more likely to live with parents.



In figure 3, the percentage of living with parents is highest for those who have a high school and below degree. On the contrary, people who receive a degree above bachelor are those most unlikely to live with

parents.



As figure 4 shows, the different cultural backgrounds result in differently.

#Weakness and shortcomings:

For the data set of this research, there were some weaknesses because the data was not collected by ourselves. The language variable could not describe the culture difference well in our subjects. we could collect some culture related data if we designed the survey by ourselves. For example, 21% of South Asians and 19% of Chinese adults aged 25 to 64 lived with parents, while 9% of Canadians aged 25 to 64 lived at home(Family Matters: Adults living with their parents, The Daily, 2019-02-15). Second, the non-sampling error overcoverage may occur in our research. The people above 69 years old and younger than 19 could be removed because most aged people do not need to consider living with parents and most teens younger than 19 need to live with their parents. In this way, we could reduce the cost of survey.

As we discussed in the model selection part, the model we selected has an issue of overfitted. The further step to improve the model is we can cancel more noisy data for example some data mentioned above(age younger than 19, older than 69), with less data, we hope the model can be well-fitted.

#Referrence:

- Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

- Rohan Alexander and Sam Caetano (2020). gss-cleaning: To clean-up the 2017 GSS data obtained from the U of T library.
- A SURVEY ON GENERALIZED LINEAR MODELS (GLMS) AND THEIR DIAGNOSTIC TOOLS,FAHAD ABDULAZIZ ALSIDRANI and DR. MUNNI BEGUM - ADVISOR, may 2017.
- Jonathan Vespa,The Changing Economics and Demographics of Young Adulthood: 1975–2016, April 2017
- In Focus 2019: Adults Living with Parents, vanierinstitute.ca, published on February 15, 2019
- Statistical Modeling and Analysis Results for the Topsoil Lead Contamination Study, Scott M. Lesch,Daniel R. Jeske,Javier Saurez, Jan,2017
- Link Functions and Errors in Logistic Regression, Karen M, 2020
- Goodness of Fit in Logistic Regression, Mcgill Medicine School
- Family Matters: Adults living with their parents, The Daily, 2019-02
- Frank O'Brien,Business in Vancouver,JULY 24, 2020
- Statistics Canada,General Social Survey - Family (GSS), 2019