Data Analysis on the Well-being of Canadians from 2017 GSS Data Set

Rutvik Gupta (1004939837), Elyssa Plaza (1004356760), Yubing Xia (1005063244), Hongbo Zhou(1004832862)

Oct 19, 2020

Abstract

In this report, our team collected findings that would contribute to a better understanding of what factors promote one’s well-being in life. This study performed descriptive and inferential data analysis, model fitting and diagnostics at the *2017 General social survey on Family (GSS)* data set. We aim to investigate what variables can potentially affect an individual’s feeling of life. We studied the relationship of four variables of interest and feeling of life. After model fitting, we excluded three insignificant variables: Average working hours, income, and age. The finalized model consist of two independent vairables (household size and mental health state) and a response variable.

Introduction

This assignment will be looking at the *2017 General social survey on Family (GSS)*. *2017 GSS* is a sample survey of cross sectional design. It was conducted from February 2nd to November 30th, 2017. This survey followed the below two objectives:

“a) To gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time; and b) To provide information on specific social policy issues of current or emerging interest.” (CHASS)

Our team were interested to analyze what factors affects an individual’s feelings of life. From the GSS data set, we extracted the following six variables: feelings of life, hours of work, income, mental health, age, and household size. Due to a large amount of NA values within the dataset, we decide to first clean the data before drawing graphs and plots for data analysis. We then performed descripive data analysis by graphing bar charts and scatter plots. Clear positive or negative relationships between the feelings of life and all five independent variables were observed. For instance, those with higher incomes generally have better rankings of both feelings in life and mental health state than those with lower incomes.

In order to further determine the trend, we performed logistic regression model fitting and diagnostics. From the fitted model, we are able to conclude that mental health and house sizes have a statistically significant impact at an individual’s feelings of life, whilst average working hours, income, and age does not.

This survey also encountered several weaknesses. First, the overall completion rate of the survey is low. Next, there are too many missing values in the vairable column in the dataset. Presence of these missing responses or answers are possibly due to the fact that some participants feel that the information they were sharing was too sensitive. This poses a loss of collection in this data population and could potentially lead to biased response pool. Another weakness would be the self rating design for answers to topics including mental health state, feeling of life and household size. These self-rated questions can be answered subjectively by participants. They might rate their state by the feelings at the moment, which can be highly variable and biased. For future research, our team can improve our survey practicability by specifically asking participants objective questions for hope of collecting honest responses. This move could effectively avoid unnecessary skewed data and ensure a statistically significant data collection. Additionally, in future interviews, the following questions could be added on the topic of individual feelsing of life: whether or not the participants suffer from a psychological disorder, their frequency of mental breakdown and counseller meetings.

Code and data supporting this analysis is available at: https://github.com/RutvikGupta/STA304-PS2 (https://github.com/RutvikGupta/STA304-PS2)

Data

Data summary

Name	gss_data
Number of rows	20602
Number of columns	81
Column type frequency:	
character	60
logical	1
numeric	20

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
sex	0	1.00	4	6	0	2	0
place_birth_canada	97	1.00	10	19	0	3	0
place_birth_father	203	0.99	10	19	0	3	0
place_birth_mother	47	1.00	10	19	0	3	0
place_birth_macro_region	16457	0.20	4	18	0	6	0
place_birth_province	4289	0.79	6	39	0	11	0
year_arrived_canada	16550	0.20	19	27	0	14	0
province	0	1.00	6	25	0	10	0
region	0	1.00	6	16	0	5	0
pop_center	0	1.00	20	53	0	3	0
marital_status	7	1.00	7	21	0	6	0
aboriginal	3855	0.81	2	10	0	3	0
vis_minority	140	0.99	10	22	0	3	0
age_immigration	17225	0.16	12	14	0	16	0
landed_immigrant	16450	0.20	2	10	0	3	0
citizenship_status	1143	0.94	8	17	0	3	0
education	341	0.98	28	60	0	7	0
own_rent	120	0.99	10	59	0	3	0
living_arrangement	0	1.00	5	51	0	12	0
hh_type	76	1.00	5	40	0	5	0
partner_birth_country	7697	0.63	6	22	0	3	0
partner_birth_province	7883	0.62	6	39	0	12	0
partner_vis_minority	7719	0.63	10	22	0	3	0
partner_sex	20407	0.01	4	10	0	3	0
partner_education	8259	0.60	28	60	0	7	0
average_hours_worked	7166	0.65	6	19	0	6	0
worked_last_week	23	1.00	2	10	0	3	0
partner_main_activity	7907	0.62	5	51	0	10	0
self_rated_health	99	1.00	4	10	0	6	0
self_rated_mental_health	106	0.99	4	10	0	6	0
religion_has_affiliation	282	0.99	10	25	0	3	0
regilion_importance	253	0.99	10	20	0	5	0
language_home	448	0.98	6	41	0	8	0
language_knowledge	105	0.99	10	26	0	5	0
income_family	0	1.00	17	21	0	6	0
income_respondent	0	1.00	17	21	0	6	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
occupation	7297	0.65	9	59	0	11	0
childcare_regular	18756	0.09	10	35	0	3	0
childcare_type	19365	0.06	14	38	0	6	0
childcare_monthly_cost	19962	0.03	3	18	0	7	0
ever_fathered_child	13604	0.34	2	10	0	3	0
ever_given_birth	12769	0.38	2	10	0	3	0
number_of_current_union	18600	0.10	11	21	0	4	0
lives_with_partner	0	1.00	2	3	0	2	0
children_in_household	0	1.00	8	22	0	4	0
has_grandchildren	4	1.00	2	3	0	2	0
grandparents_still_living	9	1.00	2	10	0	3	0
ever_married	5	1.00	2	10	0	3	0
current_marriage_is_first	10416	0.49	2	10	0	3	0
religion_participation	199	0.99	10	23	0	6	0
partner_location_residence	18978	0.08	10	36	0	4	0
full_part_time_work	18852	0.08	9	25	0	3	0
time_off_work_birth	18855	0.08	2	10	0	3	0
reason_no_time_off_birth	20283	0.02	5	48	0	10	0
returned_same_job	19451	0.06	2	3	0	2	0
satisfied_time_children	19691	0.04	9	17	0	5	0
provide_or_receive_fin_supp	19578	0.05	10	35	0	5	0
fin_supp_agreement	19937	0.03	5	60	0	5	0
future_children_intention	13438	0.35	6	18	0	6	0
age_diff	10430	0.49	10	42	0	16	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
main_activity	20602	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
caseid	0	1.00	10301.50	5947.43	1.0	5151.25	10301.5	15451.75	20602	▇▇▇▇▇
age	0	1.00	52.19	17.75	15.0	37.30	54.2	66.77	80	▃▆▆▇▇
age_first_child	6835	0.67	30.57	17.10	0.0	15.00	32.0	44.00	60	▇▆▇▇▆
age_youngest_child_under_6	18488	0.10	2.41	1.60	0.0	1.00	2.0	4.00	5	▇▅▃▃▃
total_children	19	1.00	1.68	1.49	0.0	0.00	2.0	3.00	7	▇▆▃▁▁
age_start_relationship	18566	0.10	33.63	11.20	18.0	25.00	30.5	40.62	60	▇▇▅▃▂
age_at_first_marriage	15248	0.26	24.10	5.41	15.0	20.50	22.8	26.40	50	▇▇▂▁▁
age_at_first_birth	7865	0.62	26.86	5.42	18.0	22.80	26.4	30.30	45	▆▇▅▂▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
distance_between_houses	19476	0.05	17.13	18.18	0.0	4.00	10.0	24.75	90	▇▂▁▁▁
age_youngest_child_returned_work	19466	0.06	6.59	6.17	0.2	0.50	6.0	12.00	48	▇▆▁▁▁
feelings_life	271	0.99	8.09	1.65	0.0	7.00	8.0	9.00	10	▁▁▂▇▇
hh_size	0	1.00	2.35	1.26	1.0	1.00	2.0	3.00	6	▇▂▂▁▁
number_total_children_intention	12202	0.41	0.90	1.18	0.0	0.00	0.0	2.00	4	▇▁▃▁▁
number_marriages	0	1.00	0.80	0.62	0.0	0.00	1.0	1.00	4	▅▇▁▁▁
fin_supp_child_supp	20057	0.03	0.77	0.42	0.0	1.00	1.0	1.00	1	▂▁▁▁▇
fin_supp_child_exp	20057	0.03	0.34	0.47	0.0	0.00	0.0	1.00	1	▇▁▁▁▅
fin_supp_lump	20057	0.03	0.06	0.23	0.0	0.00	0.0	0.00	1	▇▁▁▁▁
fin_supp_other	20057	0.03	0.06	0.23	0.0	0.00	0.0	0.00	1	▇▁▁▁▁
is_male	0	1.00	0.46	0.50	0.0	0.00	0.0	1.00	1	▇▁▁▁▇
number_total_children_known	0	1.00	0.41	0.49	0.0	0.00	0.0	1.00	1	▇▁▁▁▆

The source data we perform analysis on is the the *2017 GSS dataset*. The methods used to collect and process data are stratified sampling and simple random sampling. First, 27 stratas were created and within each strata simple random sample without replacement was performed to randomly select households and then one respondent in each chosen household.

Due to respondent language barriers or other personal reasons, non-response scenarios are present in the dataset. The non-response data were collected and categorized for further analysis. For respondents who answered the survey phone call but were unwilling to answer certain questions during the survey, the institution was able to impute the value based on data from other sources. If unable to find information from other sources, these will be left as NA/Not willing to participate in the survey dataset.

As it can be seen from the preview of the data above and skimmed summary of all the labels avaiable, the dataset consists of 81 labels and there are 20602 samples per label. One of the primary advantage of using this dataset is that it consist of sufficient information for us to develop and fit a statistically significant regression model on the dataset and analyse the relationship between some of the labels. The sample size is sufficiently large to produce estimates and we can take multiple explanatory variables for our response variables so that our model will be unbiased and precise and show a better depiction of the target population.

For our model, the target population is a sample size of approximately average hours worked, life feelings score, self rate mental health and age of 20,000 respondents. The frame population are all the citizens of Canada who are eligible to respond to the survey using Statistics Canada’s common telephone frame. Since the target population is so large, the sampled population are 1000 randomly selected respondent from entire dataset with given labels.

One disadvantage of using this dataset is that the there too many missing values for most of the labels and variables, and the complete_rate is also quite low overall. This implies there are many variables in the dataset that involves sharing sensitive information and people wont feel confident sharing them while filling out the survey. Therefore, our anaylsis for some labels can be a bit skewed or not statistically significant. Also many people can lie when responding to some general life related labels like self_rated_mental_health or feeling_life, and that also can lead to poor or biased analyis of the dataset.

We selected our sample population by only taking randomly selected 1000 respondents and their feelings_life, average_hours_worked, income_respondent, self_rated_mental_health, age and hh_size labels from 20000 respondents. Since many of them can consist of incomplete samples for some of the labels, we will remove them from the dataset by filtering them out and then randomly sample out 1000 from the filtered dataset to get samples with complete labels.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, age [824]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1             9 0.1 to 29.9 hou~ Less than $25,0~ Very good         15
##  2             6 0.1 to 29.9 hou~ Less than $25,0~ Good              15.6
##  3             9 30.0 to 40.0 ho~ Less than $25,0~ Excellent         15.7
##  4             9 0.1 to 29.9 hou~ Less than $25,0~ Excellent         15.8
##  5             7 0.1 to 29.9 hou~ Less than $25,0~ Good              16
##  6             8 40.1 to 50.0 ho~ Less than $25,0~ Very good         16.3
##  7            10 0.1 to 29.9 hou~ Less than $25,0~ Excellent         16.8
##  8            10 0.1 to 29.9 hou~ Less than $25,0~ Excellent         16.8
##  9             8 0.1 to 29.9 hou~ Less than $25,0~ Very good         16.9
## 10             7 30.0 to 40.0 ho~ Less than $25,0~ Very good         17.3
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

When we arrange our sample population with respect to age, we see that younger people (15 to 25) have a frequent higher feelings_life score ranging from 8 to 10 compared to the older people (50 to 80) who have lower feelings_life score (6 to 8) showing a negative relationship between the two variables.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, average_hours_worked [35]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1            10 0.1 to 29.9 hou~ Less than $25,0~ Very good         47
##  2            10 0.1 to 29.9 hou~ Less than $25,0~ Good              49.4
##  3            10 0.1 to 29.9 hou~ Less than $25,0~ Excellent         47.1
##  4             9 0.1 to 29.9 hou~ Less than $25,0~ Excellent         38.9
##  5             7 0.1 to 29.9 hou~ $25,000 to $49,~ Excellent         69.6
##  6             8 0.1 to 29.9 hou~ Less than $25,0~ Excellent         23.8
##  7             9 0.1 to 29.9 hou~ Less than $25,0~ Very good         17.5
##  8             7 0.1 to 29.9 hou~ $25,000 to $49,~ Fair              61.2
##  9            10 0.1 to 29.9 hou~ $50,000 to $74,~ Good              32.3
## 10             8 0.1 to 29.9 hou~ Less than $25,0~ Excellent         17.7
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

When we arrange our sample population with respect to average hours worked, we see that people who work less hours on average (1 to 30) have a frequent higher feeling_life score ranging from 8 to 10 compared to the people who work more hours on average (40 and more) who have a lower feelings_life score (5 to 8) showing a positive relationship between the two variables.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, income_respondent [45]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1             9 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         56
##  2             8 50.1 hours and ~ $100,000 to $ 1~ Good              48.6
##  3             8 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         59
##  4             8 40.1 to 50.0 ho~ $100,000 to $ 1~ Very good         57.1
##  5             8 50.1 hours and ~ $100,000 to $ 1~ Good              64
##  6             9 40.1 to 50.0 ho~ $100,000 to $ 1~ Very good         36.5
##  7             8 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         43.2
##  8             7 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         48.9
##  9             9 50.1 hours and ~ $100,000 to $ 1~ Very good         58.3
## 10            10 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         59.3
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

When we arrange our sample population with respect to income_respondent, we see that people who are in a higher income bracket ($100k, $129k) have a frequent higher feelings_life score ranging from 8 to 10 compared to the people who are in lower income bracket (less than $25k) who have a lower feelings_life score (5 to 8) showing a positive relationship between the two variables.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, self_rated_mental_health [38]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1             8 30.0 to 40.0 ho~ $25,000 to $49,~ Excellent         65  
##  2             8 50.1 hours and ~ $50,000 to $74,~ Excellent         55.5
##  3            10 30.0 to 40.0 ho~ $75,000 to $99,~ Excellent         40.4
##  4            10 30.0 to 40.0 ho~ Less than $25,0~ Excellent         58.3
##  5             9 40.1 to 50.0 ho~ $125,000 and mo~ Excellent         43  
##  6            10 30.0 to 40.0 ho~ $75,000 to $99,~ Excellent         31.6
##  7            10 30.0 to 40.0 ho~ $125,000 and mo~ Excellent         39.6
##  8            10 0.1 to 29.9 hou~ Less than $25,0~ Excellent         47.1
##  9             9 0.1 to 29.9 hou~ Less than $25,0~ Excellent         38.9
## 10             7 0.1 to 29.9 hou~ $25,000 to $49,~ Excellent         69.6
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

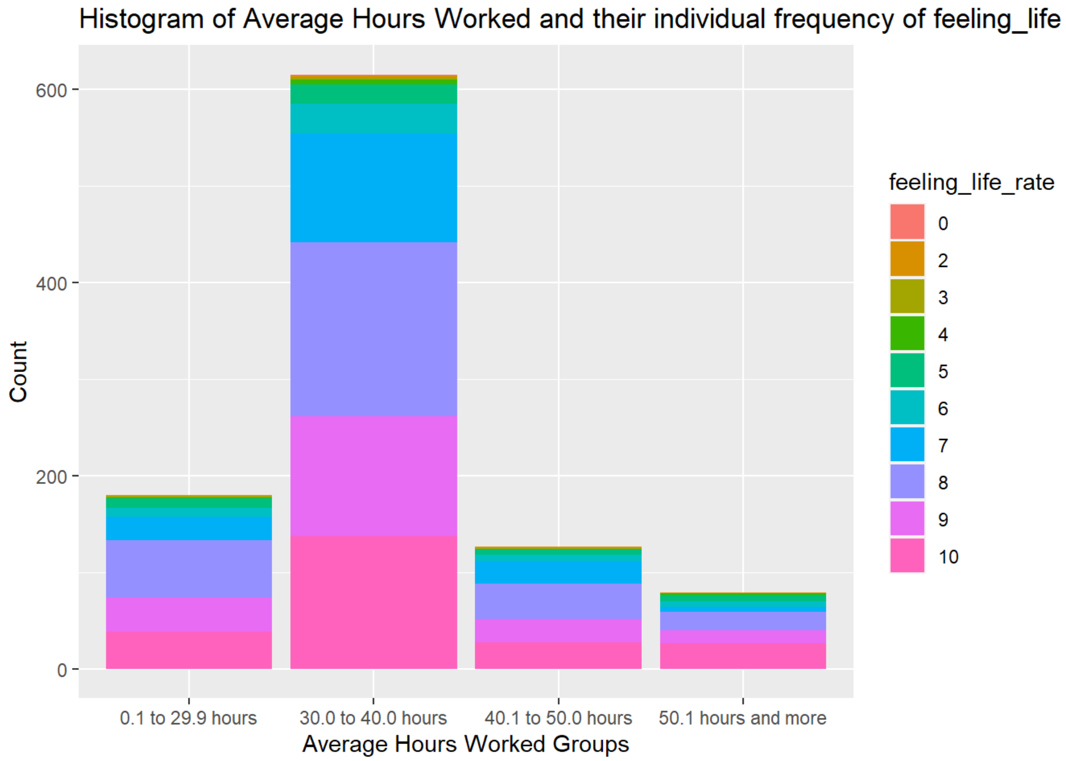
When we arrange our sample population with respect to income_respondent, we see that people who are in higher income bracket ($100k, $129k) have a frequent higher feeling_life score ranging from 8 to 10 compared to the people who are in lower income bracket (less than $25k) who have a lower feelings_life score (5 to 8) showing a positive relationship between the two variables.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, hh_size [45]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1             9 30.0 to 40.0 ho~ $100,000 to $ 1~ Very good         56  
##  2            10 30.0 to 40.0 ho~ Less than $25,0~ Excellent         58.3
##  3             7 30.0 to 40.0 ho~ $25,000 to $49,~ Fair              23.4
##  4             8 40.1 to 50.0 ho~ $75,000 to $99,~ Very good         47  
##  5             2 40.1 to 50.0 ho~ $50,000 to $74,~ Fair              58.3
##  6             7 0.1 to 29.9 hou~ $25,000 to $49,~ Excellent         69.6
##  7             8 30.0 to 40.0 ho~ $75,000 to $99,~ Excellent         52.5
##  8            10 40.1 to 50.0 ho~ $75,000 to $99,~ Excellent         33.7
##  9             8 30.0 to 40.0 ho~ $50,000 to $74,~ Excellent         68.3
## 10             7 0.1 to 29.9 hou~ $25,000 to $49,~ Fair              61.2
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

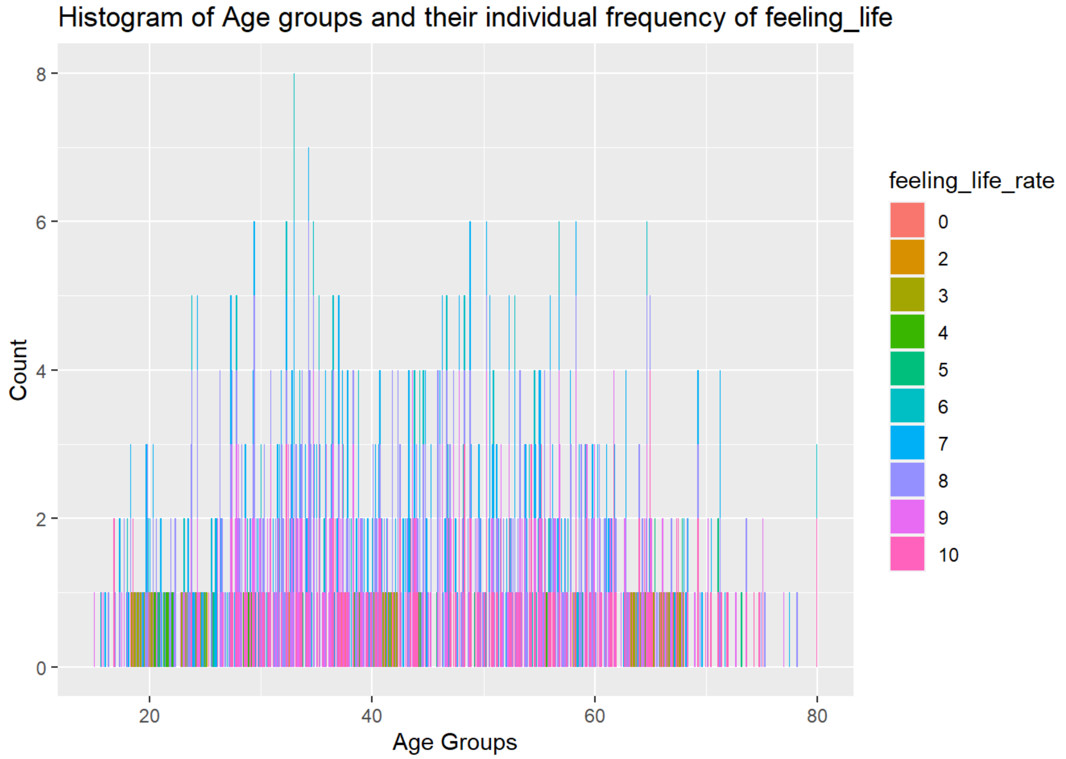
When we arrange our sample population with respect to hh_size, we see that people with bigger house size (4 to 6) have a frequent higher feeling_life score ranging from 8 to 10 compared to the people with smaller house size (1 to 3) who have a lower feelings_life score (5 to 8) showing a positive relationship between the two variables.

```
## # A tibble: 1,000 x 6
## # Groups:   feelings_life, self_rated_mental_health [38]
##    feelings_life average_hours_w~ income_responde~ self_rated_ment~   age
##            <int> <chr>            <chr>            <chr>            <dbl>
##  1             0 30.0 to 40.0 ho~ $125,000 and mo~ Excellent         37.4
##  2             2 40.1 to 50.0 ho~ $50,000 to $74,~ Fair              58.3
##  3             2 50.1 hours and ~ $50,000 to $74,~ Poor              58.1
##  4             2 0.1 to 29.9 hou~ $50,000 to $74,~ Fair              32  
##  5             2 30.0 to 40.0 ho~ $25,000 to $49,~ Poor              32.6
##  6             2 30.0 to 40.0 ho~ $50,000 to $74,~ Good              44  
##  7             3 30.0 to 40.0 ho~ Less than $25,0~ Fair              64.3
##  8             3 40.1 to 50.0 ho~ $25,000 to $49,~ Good              66.9
##  9             3 0.1 to 29.9 hou~ Less than $25,0~ Good              19.4
## 10             3 0.1 to 29.9 hou~ Less than $25,0~ Poor              24  
## # ... with 990 more rows, and 1 more variable: hh_size <int>
```

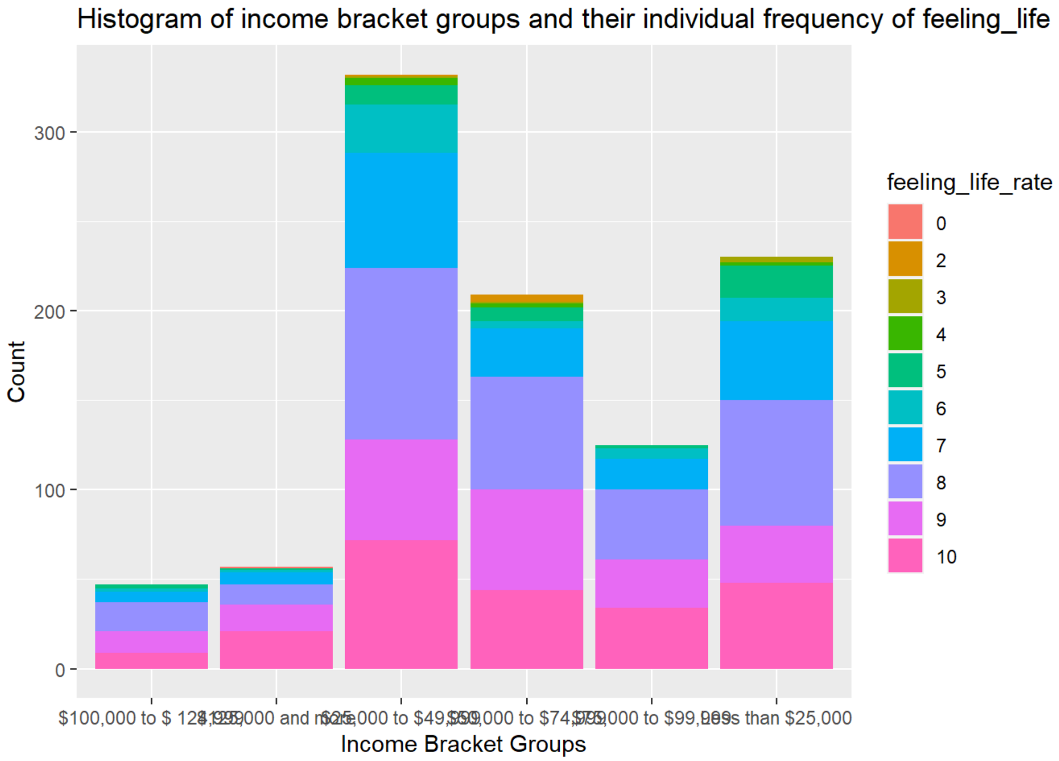
When we arrange our sample population with respect to descending order of feelings_life, we see that people with “very good” or “excellent” self rated mental health have a frequent higher feeling_life score ranging from 8 to 10 compared to the people with “poor or”fair" self rated mental health who have a lower feelings_life score (5 to 8) showing a positive relationship between the two variables.



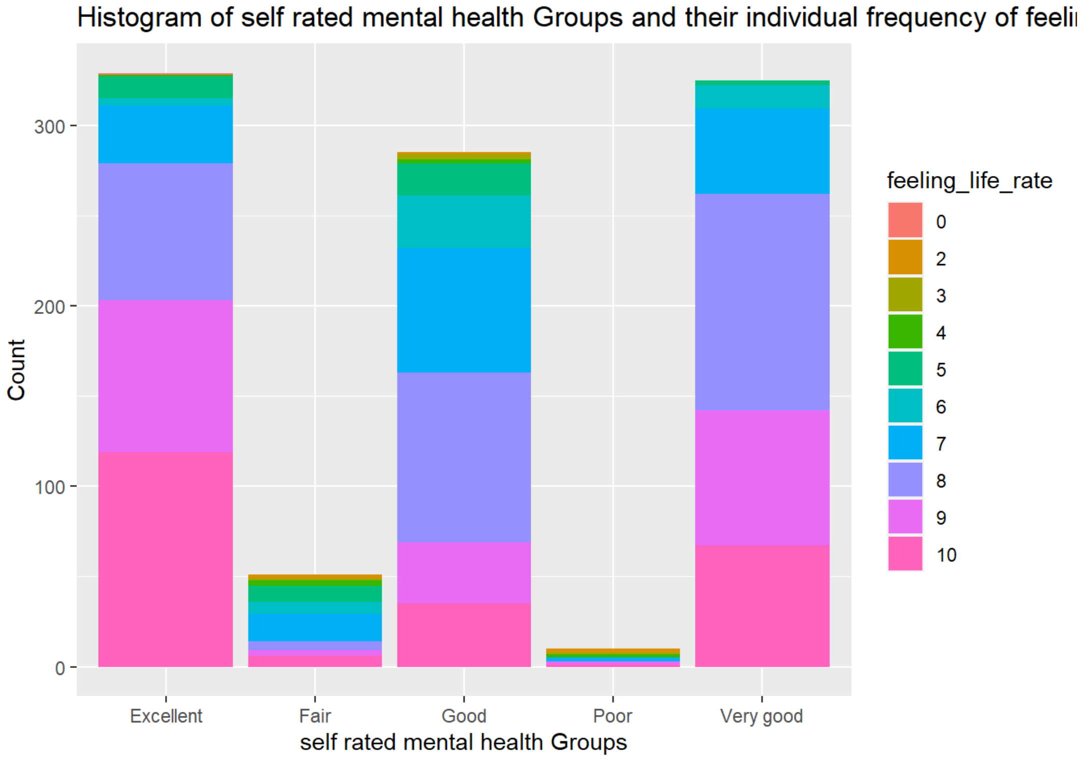
The “Histogram of Average Hours Worked and their individual frequency of feeling_life” supports our findings in the table analysis showing that people working 30 to 40 hours on average a week have more frequent rating of high feeling_life variable (8 to 10) as compared to other. This histogram also shows that our data can be normally distributed because of its bell shaped curve.



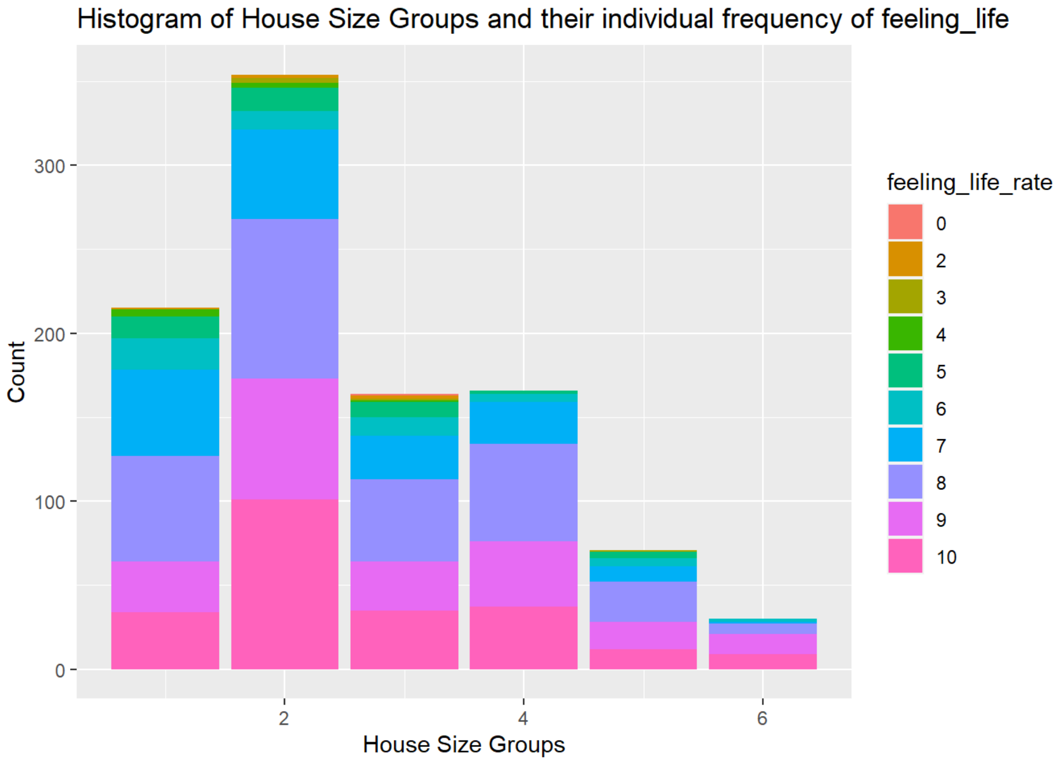
The “Histogram of Age groups and their individual frequency of feeling_life” supports our findings in the table analysis showing that people in 20 to 40 age groups have more frequent rating of high feeling_life variable (8 to 10) as compared to other. This histogram also shows that our data can be normally distributed because of its bell shaped curve.



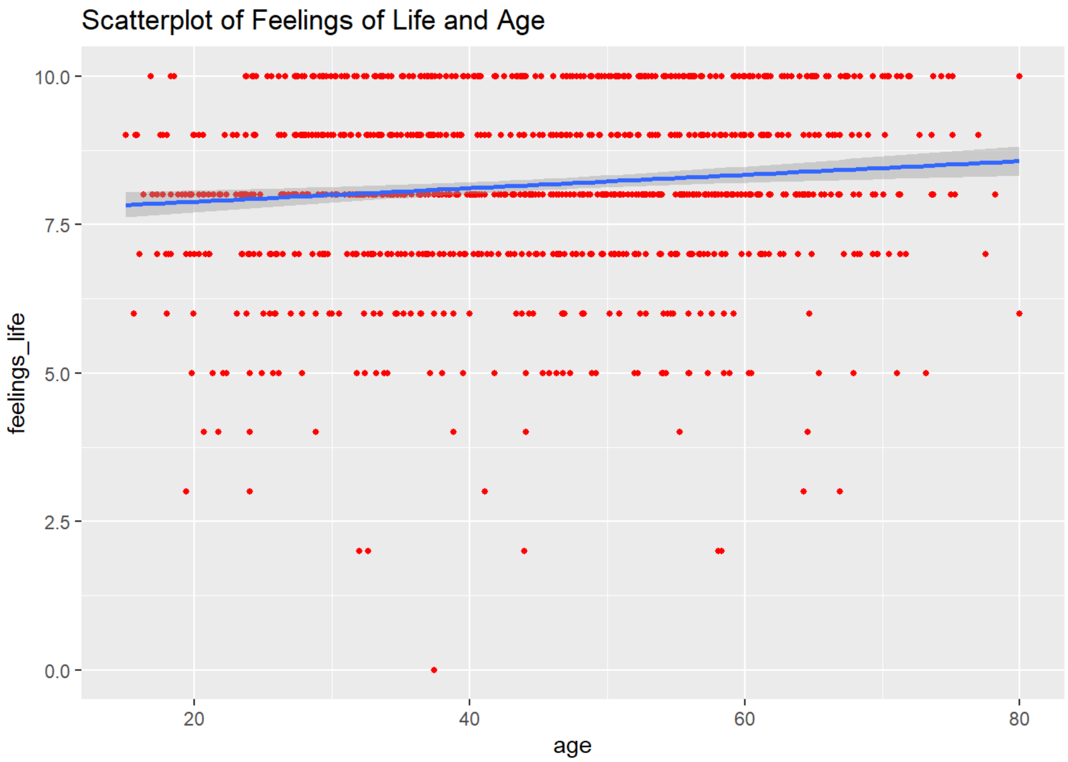
The “Histogram of income bracket groups and their individual frequency of feeling_life” supports our findings in the table analysis showing that people in higher income bracket ($50k and above) have more frequent rating of high feeling_life variable (8 to 10) as compared to other. This histogram also shows that our data can be normally distributed because of its bell shaped curve.



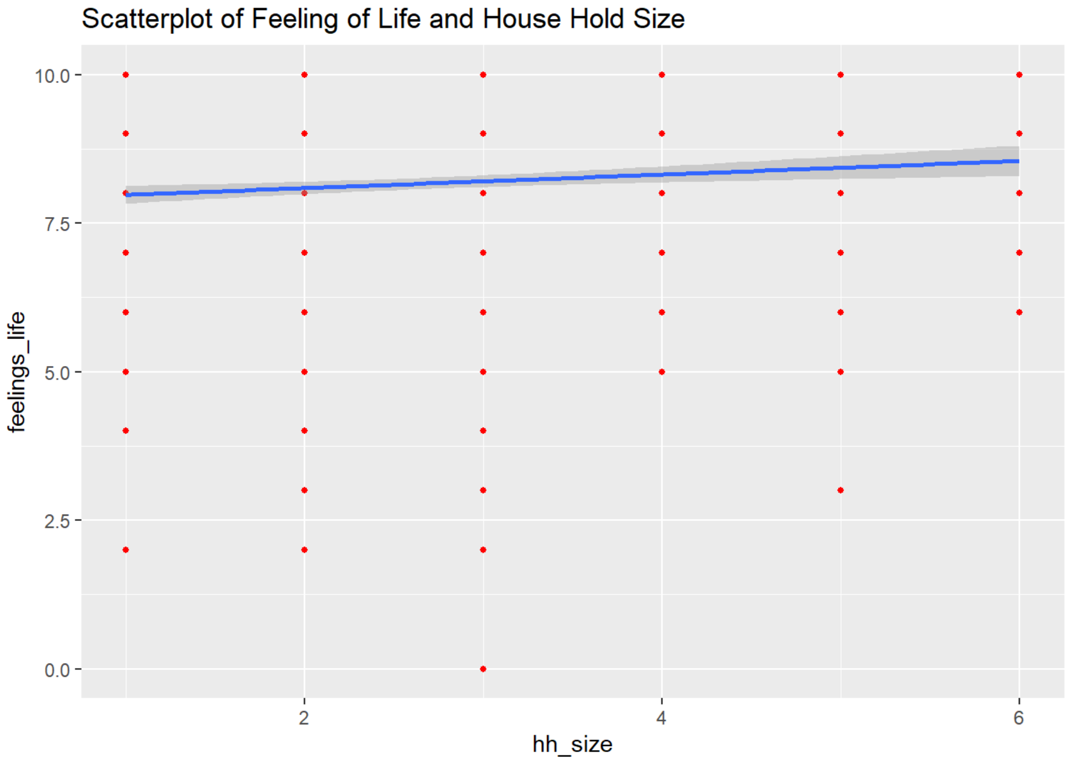
The “Histogram of self rated mental health Groups and their individual frequency of feeling_life” supports our findings in the table analysis showing that people with “excellent” and “very good” self rated mental health have more frequent rating of high feeling_life variable (8 to 10) as compared to other.


Histogram of House Size Groups and their individual frequency of feeling_life

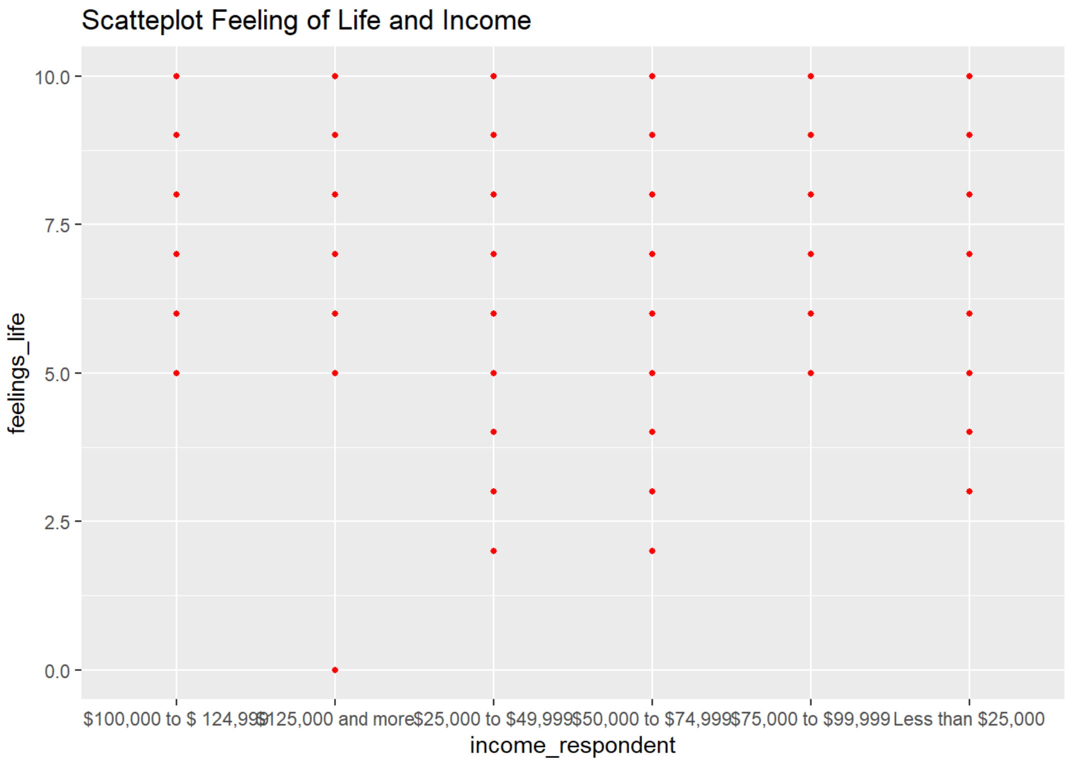
The “Histogram of House Size Groups and their individual frequency of feeling_life” shows an opposite result as compared to our findings in the table analysis showing that people with with smaller house size (1 to 3) have more frequent rating of high feeling_life variable (8 to 10) as compared to other. This implies that we need to rely on our anaysis from fitting the model to justify our findings.


Scatterplot of Feelings of Life and Age

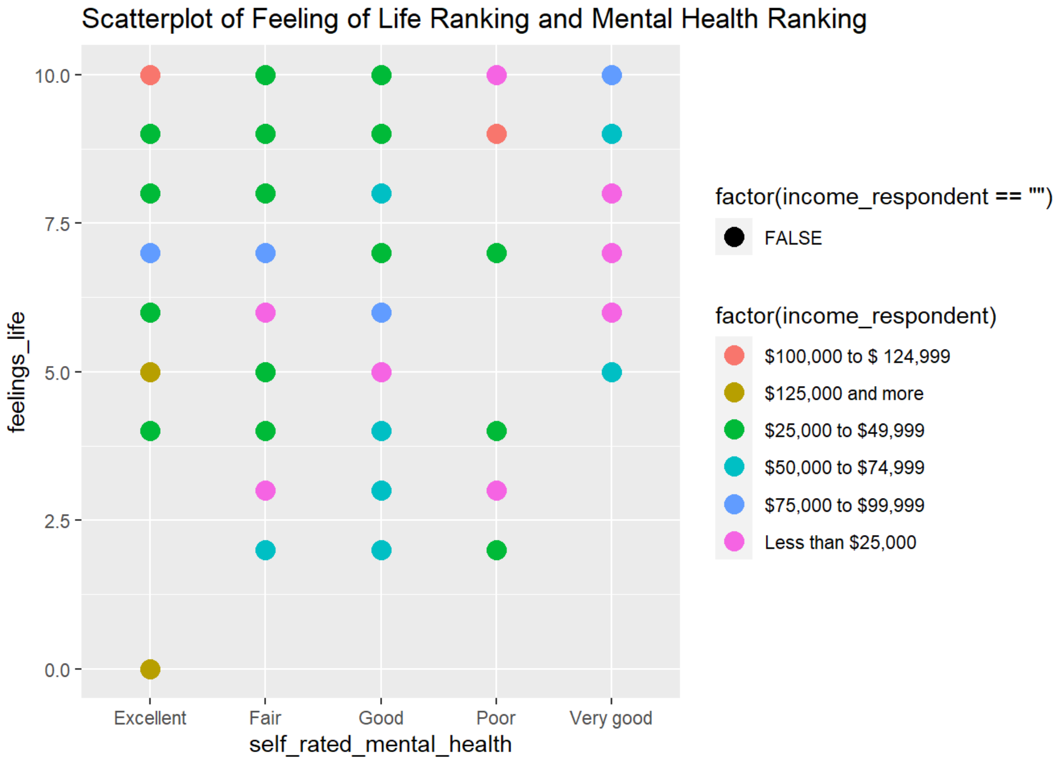
The “Scatterplot of Feelings of Life and Age” plots the variables of age (x-axis) and feelings_life (y-axis), ranking of life from 0 to 10 (bad feeling of life to good feeling of life).The plot shows that many of the points are concentrated when the feeling of life is at 7.5 - 10, implying many have high feelings of life. In addition to that, the positive line of best fit shows that the feelings_life variable tend to be higher as age increases.



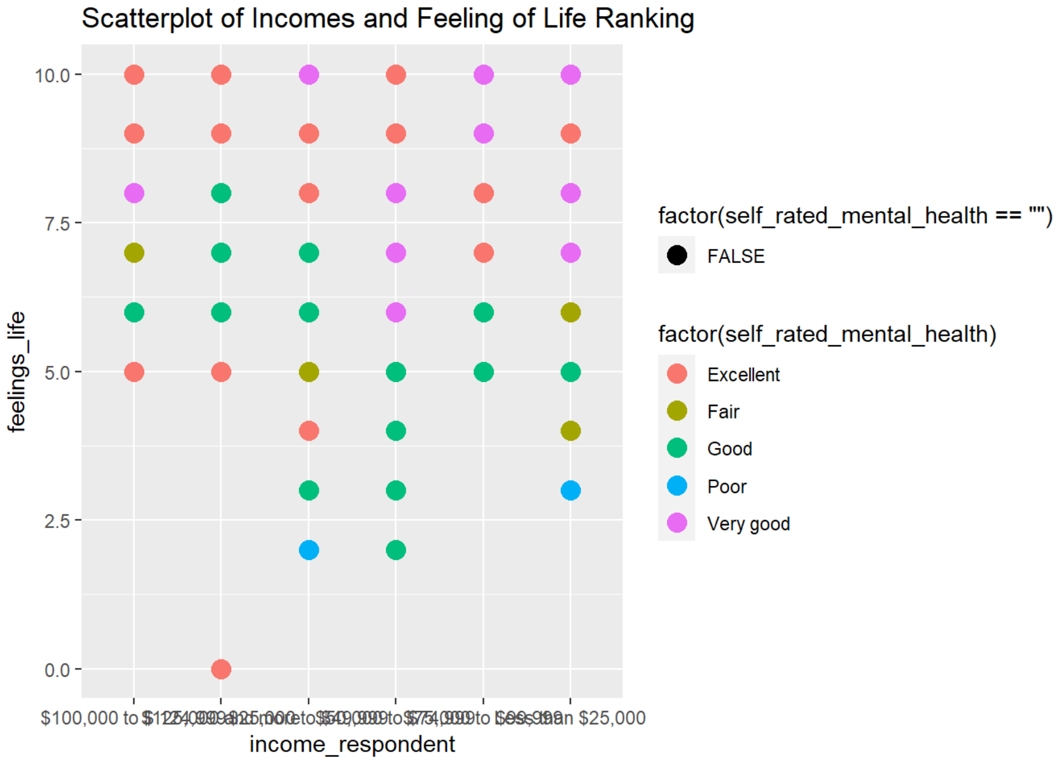
The “Scatterplot of Feeling of Life and House Hold Size” plots the variables hh_size (x-axis) and feelings_life (y-axis). The plot shows that many of the points are concentrated when the feeling of life is at 7.5 - 10, implying many have high feelings of life. In addition to that, the positive line of best fit on the plot shows that the feelings life increase when house hold size is larger.



The “Scatteplot Feeling of Life and Income” plots the variables income_respondent (x-axis) and feelings_life (y-axis). The plot shows that lower feelings of life (0-5) are more evident from those making an income of less than $75k. Whereas, those who have an income greater than $75k tend to have higher feelings of life (6-10).



The “Scatterplot of Feeling of Life Ranking and Mental Health Ranking” plots the variables self_rated_mental_health (x-axis) and feelings_life (y-axis), with income_respondents as the factor. The red point show those who have an income from $100k to $124,999. This point is evident when self_rated_mental_health is “Excellent” and feelings_life ranking is ranked high (10). This red point is also evident when self_rated_mental_health is poor, however feelings_life ranking is high (9).



The “Scatterplot of Incomes and Feeling of Life Ranking” plots the variables income_respondent (x-axis) and feelings_life (y-axis), with self_rated_mental_health as the factor. The red point represents an “Excellent” self_rated_mental_health, these points are highly concentrated at higher incomes greater than $25k. These red points are also concentrated when feelings_life are high (7.5 -10).

All the scatterplots above strongly supports our findings and observations in the tables and historgram analysis. The line of best fit in some variables shows positive slope hence showing a strong relationship between the explanatory variables and our response variables feelings life.

Model

The main response variable we are interested in is feelings of life, which is a categorical variable ranging from 0 to 10. In order for us to perform a logistic regression model analysis on this categorical variable, we first dichotomize it into 1 – a good feeling of life, and 0 – a not good feeling of life. Values from 0 to 5 of the original ranking would now be 0, and 6 to 10 would be 1.

Then we examine other variables of interest: average working hours, a numerical variable; income, another numerical variable; self-rated mental health, a categorical variable; age, also a numerical variable; and household size, a categorical variable. We filter some NA values and sample 1000 data entries from the entire 20602 observations of the survey.

Putting all these into consideration, we decide to fit both a standard logistic regression model and a survey estimation for logistic regression using R and RStudio. The model would be

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{workhours30-40} + \beta_2 x_{workhours40-50} + \beta_3 x_{workhours>50} + \beta_4 x_{income25-49} + \beta_5 x_{income50-74} + \beta_6 x_{income75-99} + \\ \beta_7 x_{income<25} + \beta_8 x_{fair} + \beta_9 x_{goodmental} + \beta_{10} x_{poormental} + \beta_{11} x_{verygoodmental} + \beta_{12} x_{age} + \beta_{13} x_{hhsize2} + \\ \beta_{14} x_{hhsize3} + \beta_{15} x_{hhsize4} + \beta_{16} x_{hhsize5} + \beta_{17} x_{hhsize6}$$

, which is just the model modelling feelings of life with average working hours, income, mental health, age and household size.

After we finish the modelling using both techniques, we would first compare the coefficient estimates between the standard model and the survey model. Then we would check for the individual p-values of the coefficients to see if a smaller model can be fitted. If so, that is, some variables do not seem to be statistically significant, we would try to fit a smaller model with the variables with statistically significant p-values again using the two techniques.

After fitting the models, we would check for model assumptions and perform diagnostics for the models to see if constant variance property is violated and the presence of possible outliers. Also, we would be interpreting all the results and discussing possible weaknesses and next steps in later sections.

Results

Model Fitting

Based on the cleaned dataset of 20602 respondents, we performed standard logistic regression model fitting. Below is the summary for the initial standard model buildup.

```
##
## Call:
## glm(formula = feelings_life ~ average_hours_worked + income_respondent +
##     self_rated_mental_health + age + as.factor(hh_size), family = "binomial",
##     data = gss_select)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1722   0.1015   0.1890   0.3402   1.3342
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              2.67360    1.15050   2.324  0.02013
## average_hours_worked30.0 to 40.0 hours    0.32826    0.38275   0.858  0.39109
## average_hours_worked40.1 to 50.0 hours    0.15185    0.52926   0.287  0.77418
## average_hours_worked50.1 hours and more  -1.10032    0.54890  -2.005  0.04501
## income_respondent$125,000 and more       -0.41362    1.12649  -0.367  0.71349
## income_respondent$25,000 to $49,999      -0.10256    0.88254  -0.116  0.90749
## income_respondent$50,000 to $74,999      -0.89490    0.88952  -1.006  0.31439
## income_respondent$75,000 to $99,999       0.96289    1.11639   0.863  0.38841
## income_respondentLess than $25,000       -1.08575    0.89176  -1.218  0.22340
## self_rated_mental_healthFair             -2.50442    0.44806  -5.589 2.28e-08
## self_rated_mental_healthGood             -0.65085    0.36307  -1.793  0.07303
## self_rated_mental_healthPoor             -2.96874    0.75907  -3.911 9.19e-05
## self_rated_mental_healthVery good         1.72851    0.65072   2.656  0.00790
## age                                       0.01079    0.01061   1.017  0.30910
## as.factor(hh_size)2                       0.40508    0.37009   1.095  0.27372
## as.factor(hh_size)3                      -0.11348    0.41476  -0.274  0.78440
## as.factor(hh_size)4                       2.39037    0.80577   2.967  0.00301
## as.factor(hh_size)5                       0.19081    0.59082   0.323  0.74672
## as.factor(hh_size)6                      15.38999  667.50238   0.023  0.98161
##
## (Intercept)                             *
## average_hours_worked30.0 to 40.0 hours
## average_hours_worked40.1 to 50.0 hours
## average_hours_worked50.1 hours and more *
## income_respondent$125,000 and more
## income_respondent$25,000 to $49,999
## income_respondent$50,000 to $74,999
## income_respondent$75,000 to $99,999
## income_respondentLess than $25,000
## self_rated_mental_healthFair            ***
## self_rated_mental_healthGood            .
## self_rated_mental_healthPoor            ***
## self_rated_mental_healthVery good       **
## age
## as.factor(hh_size)2
## as.factor(hh_size)3
## as.factor(hh_size)4                     **
## as.factor(hh_size)5
## as.factor(hh_size)6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 459.42  on 999  degrees of freedom
## Residual deviance: 348.60  on 981  degrees of freedom
## AIC: 386.6
##
## Number of Fisher Scoring iterations: 16
```

As labelled with *'s, in this model p-values of variables mental health ratings and household size are lower than 0.05, the significance level. Variables average hours of work, income and age have p-values above 0.05. We also performed a survey estimation of logistic regression from the sample data of 1000 participants. Here is the summary for the fitted survey model.

```
## 
## Call:
## svyglm(formula = feelings_life ~ average_hours_worked + income_respondent + 
##     self_rated_mental_health + age + as.factor(hh_size), design = gss_design, 
##     family = "binomial")
## 
## Survey design:
## svydesign(id = ~1, data = gss_select, fpc = fpc.srs)
## 
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              2.673599   1.192187   2.243 0.025145
## average_hours_worked30.0 to 40.0 hours   0.328263   0.366149   0.897 0.370191
## average_hours_worked40.1 to 50.0 hours   0.151849   0.512221   0.296 0.766948
## average_hours_worked50.1 hours and more -1.100318   0.517013  -2.128 0.033567
## income_respondent$125,000 and more      -0.413619   1.302437  -0.318 0.750876
## income_respondent$25,000 to $49,999     -0.102559   1.039560  -0.099 0.921432
## income_respondent$50,000 to $74,999     -0.894904   1.043891  -0.857 0.391501
## income_respondent$75,000 to $99,999      0.962892   1.278772   0.753 0.451641
## income_respondentLess than $25,000      -1.085746   1.038353  -1.046 0.295984
## self_rated_mental_healthFair            -2.504416   0.428100  -5.850 6.69e-09
## self_rated_mental_healthGood            -0.650851   0.345476  -1.884 0.059871
## self_rated_mental_healthPoor            -2.968739   0.698476  -4.250 2.34e-05
## self_rated_mental_healthVery good        1.728509   0.620791   2.784 0.005466
## age                                      0.010791   0.009515   1.134 0.257021
## as.factor(hh_size)2                      0.405078   0.353315   1.147 0.251865
## as.factor(hh_size)3                     -0.113475   0.392315  -0.289 0.772455
## as.factor(hh_size)4                      2.390370   0.679131   3.520 0.000452
## as.factor(hh_size)5                      0.190813   0.509429   0.375 0.708066
## as.factor(hh_size)6                     15.389987   0.412146  37.341  < 2e-16
##                                                                              
## (Intercept)                             *  
## average_hours_worked30.0 to 40.0 hours     
## average_hours_worked40.1 to 50.0 hours     
## average_hours_worked50.1 hours and more *  
## income_respondent$125,000 and more         
## income_respondent$25,000 to $49,999        
## income_respondent$50,000 to $74,999        
## income_respondent$75,000 to $99,999        
## income_respondentLess than $25,000         
## self_rated_mental_healthFair            ***
## self_rated_mental_healthGood            .  
## self_rated_mental_healthPoor            ***
## self_rated_mental_healthVery good       ** 
## age                                        
## as.factor(hh_size)2                        
## as.factor(hh_size)3                        
## as.factor(hh_size)4                     ***
## as.factor(hh_size)5                        
## as.factor(hh_size)6                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 0.8765813)
## 
## Number of Fisher Scoring iterations: 16
```

In this survey model, variable estimates are the same and their p-value significances follow the similar pattern as in the standard model.

In both models, we find p-values of several independent variables (average hours of work, income and age) to be greater than the significance level at 0.05. Those variables don’t impact the response variable, feelings of life, in a statistically significant sense. We perform further modifications to the models by deleting the insignificant variables.

We fit smaller models with the remaining two independent variables: mental health ratings and household sizes.

Below is the summary for the fitted standard model. The p-value of both variables are under 0.05.

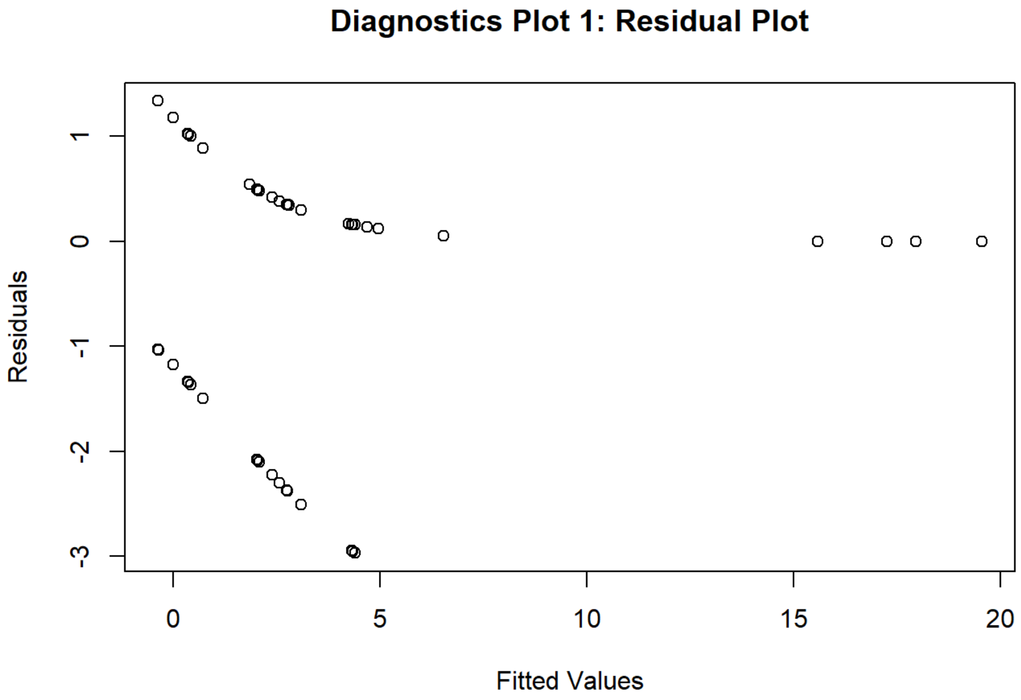
```
##
## Call:
## glm(formula = feelings_life ~ self_rated_mental_health + as.factor(hh_size),
##     family = "binomial", data = gss_select)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9697   0.1348   0.1684   0.3528   1.3366
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         2.74574    0.35538   7.726 1.11e-14 ***
## self_rated_mental_healthFair       -2.38158    0.42631  -5.586 2.32e-08 ***
## self_rated_mental_healthGood       -0.71218    0.34926  -2.039   0.0414 *
## self_rated_mental_healthPoor       -3.11244    0.71254  -4.368 1.25e-05 ***
## self_rated_mental_healthVery good   1.58836    0.64261   2.472   0.0134 *
## as.factor(hh_size)2                 0.36206    0.35391   1.023   0.3063
## as.factor(hh_size)3                 0.01833    0.40200   0.046   0.9636
## as.factor(hh_size)4                 2.21565    0.77389   2.863   0.0042 **
## as.factor(hh_size)5                 0.06313    0.55597   0.114   0.9096
## as.factor(hh_size)6                15.22505  674.36611   0.023   0.9820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 459.42  on 999  degrees of freedom
## Residual deviance: 370.85  on 990  degrees of freedom
## AIC: 390.85
##
## Number of Fisher Scoring iterations: 16
```

Below is the summary for the finalized survey model. Two independent variables have significant p-values. Therefore, we conclude the model fitting process complete.

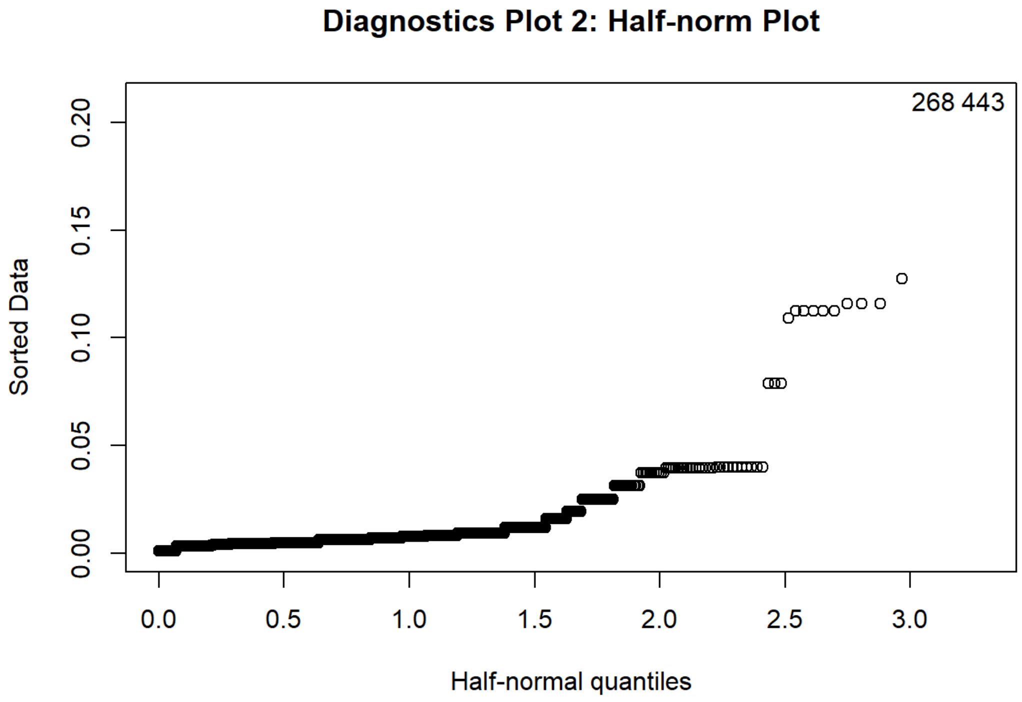
```
##
## Call:
## svyglm(formula = feelings_life ~ self_rated_mental_health + as.factor(hh_size),
##     design = gss_design, family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = gss_select, fpc = fpc.srs)
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2.74574    0.33428   8.214 6.65e-16 ***
## self_rated_mental_healthFair      -2.38158    0.42309  -5.629 2.36e-08 ***
## self_rated_mental_healthGood      -0.71218    0.33860  -2.103  0.03569 *
## self_rated_mental_healthPoor      -3.11244    0.68733  -4.528 6.67e-06 ***
## self_rated_mental_healthVery good  1.58836    0.62118   2.557  0.01071 *
## as.factor(hh_size)2                0.36206    0.34525   1.049  0.29458
## as.factor(hh_size)3                0.01833    0.40884   0.045  0.96426
## as.factor(hh_size)4                2.21565    0.71543   3.097  0.00201 **
## as.factor(hh_size)5                0.06313    0.51985   0.121  0.90337
## as.factor(hh_size)6               15.22505    0.42062  36.197  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.8159634)
##
## Number of Fisher Scoring iterations: 16
```

Model Diagnostics

We continue to diagnose the fitted models and check for potential assumption violations. Below are the graphs generated for model diagnostics.



```
## Warning: package 'faraway' was built under R version 4.0.3
```



And we also check the discrimination ability of the model.

```
## Warning: package 'pROC' was built under R version 4.0.3
```

```
## [1] 0.82
```

Discussion

2017 GSS Survey

In this report, we performed descriptive and inferential analysis on the *2017 GSS dataset* (General Social Survey on Family). This survey was conducted from February to November in 2017 via a cross-sectional study design. The survey team reached out to the sampling population and collected participants' responses on their overall life well-being by interviewing the participants a list of questions.

The survey sampling design followed a stratified sampling methodology. First, all ten provinces in Canada are divided into smaller geographical areas and each defined as one strata, with many CMAs (Census Metropolitan Areas) considered as separate stratas. In total, this dataset created 27 stratas. Following the strata splitting, within each strata SRS (Simple Random Sample) without replacement was performed to randomly select households and then subsequently one respondent in each chosen household.

This stratified sampling methodology isn't a perfect tool, however, it poses several tradeoffs. First and foremost, to collect and align telephone numbers and household addresses within each strata requires an exhausting amount of research. The linking of information is achieved through checking with numerous sources, including phone companies, Census of population, and many more. To add on, the majority of houses have more than one resident, resulting in numerous phone lines for each household. Some residents might have moved and haven't changed their address list yet. Or a resident might have registered their information in several addresses of their family members'. In these scenarios, researchers might not be capable of confidently classifying, therefore, would require more time to confirm and sort the multiple phone numbers by source and type (landline first and cellular last). These issues lead to high costs for the institutions to hire workers as well as a long period of time to finalize data arrangement.

Target, Frame, and Sample population

In the 2017 GSS handout, the researchers explained their clear definitions of target, frame, and sampling population for this survey. This dataset targets the population of Canadians who are over 15 years of age, except the Canadians residing in Yukon, Northwest Territories, and Nunavut or Canadians who work full-time at the institutions.

The frame population was defined with a delicate design. The researchers combined lists of telephone numbers and address information from two sources: Statistics Canada and AR (Address Register) and aligned these data to each household. The completion rate of the information alignment was around 86%. The frame population for the survey was this 86% of all households with valid telephone as well as address data entries.

20,602 respondents participated and provided useful data for the survey in the 2017 GSS. This sampling population size was over the estimated 20,000 participants. This sample size was large enough to ensure acceptable sampling variability estimates within stratum.

Non-response

The survey plans to sample from a population of 39,323 households. In this frame population, 26,602 usable survey responses were collected. Among the 12721 non-responsive calls, there are 1,525 cases that refused to participate, 17,196 cases that did not pick up the survey phone calls on either household or personal level. Among the 17,196 households that did not pick up, 14.6% were due to respondent language barriers or other personal reasons. Although this survey failed to collect information from these non-responsive data, the non-response data were collected and categorized into different subtypes for further analysis. These personal/household non-responses will be put into consideration in the design of future surveys.

For respondents who answered the survey phone call but were unwilling to answer certain questions during the survey (most commonly for sex or age related questions), the institution was able to impute the value based on data from other sources. If unable to find information from other sources, these will be left as NA/Not willing to participate in the survey dataset.

Dataset

The dataset analyzed in this report was filtered and cleaned from the dataset of 2017 GGS Study, which has 20602 observations on 81 variables. This report aims to analyze Canadians' feelings of life in relation to the following five variables: average hours of working, income of the respondent, self-ranked mental health state, age of the respondent and household size. We cleaned the data by filtering out the observations with incomplete data (NA input) and selecting on the six variables of interest. We then randomly sampled 1000 respondents from the sample population for the survey population.

In this survey study, the design of the self-rating questions on current mental health state, feeling of life, and household sizes pose a potential bias in survey analysis. Participant self evaluations are subjective and variable by their condition at the moment. This leads to a possible biased or skewed dataset. However, these biases are most possibly minor and it is beyond our reach to correct for them, so we will use the dataset as it is.

Possibilities of bias in the survey were evaluated by the institution. A considerable amount of time and effort were spent to reduce non-sampling errors. At every step during the data collection, quality assurance and monitor measures were performed. This includes the training and overwatch on interviewers.

This survey design improves efficiency compared to the previous years of GSS practices of the random dialing method. The useof several linked sources including the Administrative and billing data files also promoted an increase of coverage on the stratum level. However, this survey also possesses several weaknesses. Collecting responses by telephone restricts participation from the sub-population of households without telephones. Also, during the survey phone calls, institution workers have to spend extra time to go through a list of eligibility criteria to make sure the participants meet the age, residing area and job requirements. 9.2% of the selected participants did not meet the criteria and the call was terminated. To add on, another disadvantage of this survey is the high occurrence of missing values in many variables. This implies that participants are uncomfortable for sharing certain sensitive information, which could affect our analysis on certain variables.

Fitted Model

The final survey model and standard model share the same parameter estimates. We set the response variable as the dichotomized feelings of life variable from the original dataset. Below is the finalized model equation:

$$log(\frac{\hat{p}}{1-\hat{p}}) = 2.74574 - 2.38158x_{mentalhealthFair} - 0.71218x_{mentalhealthGood} - \\ 3.11244x_{mentalhealthPoor} + 1.58836x_{mentalhealthVeryGood} + 0.36206x_{hhsize2} + \\ 0.01833x_{hhsize3} + 2.21565x_{hhsize4} + 0.06313x_{hhsize5} + 15.22505x_{hhsize6}$$

We set the response value equal to 1 as having good feelings of life (6-10 rating from the original variable), and to 0 for having not so nice feelings of life (0-5 ratings from the original variable). The independent variables were reduced to respondent self-rated mental health state and household size. From the p-values generated from the model summary, we find p-values for both variables significant (< 0.05). Therefore, we can conclude that mental health state and household sizes have a significant impact on the participant general well being in life.

During the model fitting process, we find average working hours, income and age do not seem to have a statistically significant impact on life feelings of people. This contradicts our descriptive analysis on the cleaned dataset, which suggests a strong connection between these three variables and the respondent feeling of life. This is possible since during descriptive analysis, we only look at numerical and graphical representations of the dataset without performing further statistical analyses to confirm our thoughts.

Model Diagnostics

The “Diagnostics Plot 1: Residual Plot” and “Diagnostics Plot 2: Half-norm Plot” are graphed for model diagnostics. The residual plot exhibits a pattern scattering around 0. This suggests that the constant variance property is not severely violated, although there seems to be outliers trailing off. For the half-norm plot, the distribution follows a general trend with a slight tailing off at the end. This doesn’t seem to violate the model assumptions. AUC (Area Under The Curve) value is also computed for diagnostics purposes. 0.82 is a high number indicating the model’s fine performance at distinguishing between categories. Therefore, we can conclude that no model violations were detected during the diagnostics process and the model has a high performance score for future use.

Model Interpretation

With careful data screening and cleaning, we produced our data of interest and performed model fitting. The final model consists of two independent variables, household size and mental health. Based on the model, the larger the house is, the better the participants feel about life in general (the log odds ratio increases as can be seen from the coefficients). This is easy to comprehend since household size suggests the owner’s financial and purchasing power. Income was one of the variables of interest, but was deleted from the model because it shows no significant relationship with the general well-being in life of a person. An individual might be earning a lot from their job, but they might choose to spend little or prefer to invest or store in a bank. Household sizes inherently suggest one’s income as well as purchasing will, and therefore is more significantly linked to an individual’s well-being in life.

The other independent variable in the model, mental health state, also has a positive relationship with the general life well being (the log odds ratio increases as can be seen from the coefficients). This makes sense since a healthy mental state is characterized by an individual’s capability of performing daily activities and social interactions. An individual’s average hours working does not have a significant connection with their feelings of life. This is understandable: when a person enjoys their work, long hours of working would not make them feel exhausted and unhappy; but when a person dislikes their job, even short hours of working might be irritable for them. Therefore, hours of work cannot accurately depict their feelings about their job. Self-rated mental health, however, covers one’s feelings. Hence, being able to function well in social and work life provides enjoyment in life, and directs toward a better feeling of life.

In “Statistical Rethinking”, Statistician McElreath introduced the concept of model performance in the small and big world. The small world refers to the 1000 respondents randomly selected for this study, whilst the big world consists of the entire target population in the 2017 GSS study - all Canadians who have registered household address and meet the age, job prerequisites.

Our finalized model is capable of accurately representing the big world. This is supported by our choice of data collection and sampling techniques, which is explicitly explained in the previous Model Diagnostics subsection. We performed stratified sampling and simple random sampling without replacement. These techniques allow us to sample in a non-biased random way. However, there are some aspects in our model fitting process that might challenge the representation performance of the model in the large world. Since the small world is limited by sampling size, some scenarios might fail to be included. For instance, people who didn’t provide answers to somes questions during the phone interview are filtered from the small world sample for a purpose of better analysis quality. Excluding data from this population of participants might have an influence on how accurate our small world sample model is able to depict the target large world. But in all, the finalized model will be a good representative of both the small world and the large world.

Therefore, one's wealth or working condition is not that significant when considering their general well-being in life. Instead of focusing on how to earn more money and find a better job with less working hours or other requirements, people should focus on how to purchase wisely and always check themselves for a healthy mental state. What is more important is always your decision on how to treat yourself, rather than reaching for the subjective things in life.

Weakness and Next Steps

Self-rated variables in the model includes evaluations of self mental health state, feeling of life, as well as household size. This self-rating design induces biasedness in inferential analysis. This is because participants can be subjective and rate the variables by their condition at the moment, which can be highly variable. This poses a weakness in this modelling design, since two of the three variables in the model are rated by participants themselves.

In the future, a new survey should be drawn to collect responses on well being. New variables should be added to directly investigate one's purchasing power. This would include self-rated purchasing power (could be biased), weekly/monthly purchasing power measured by dollars spent, preference of circulating funds via investment or other or storage in banks. Similarly, the mental health state variable size could be expanded to the occurrence of psychological disorder, the frequency of breakdown, the frequency of visits to mental health services etc. These added variables would help us gain a more thorough understanding on how an individual's purchasing will and mental health state have to do with their well-being in life.

Conclusion

In conclusion, our final model brings insightful thoughts on how to be nicely-placed in life. It is a decent tool for future analysis on one's general well being, giving that they are Canadians who fit in the target population. By inputting data on their household size and self rated mental health state, the model is able to generate a reliable result on their current feeling of life, and vice versa.

References

1. Ebner, Joshua, et al. "How to Make a Scatter Plot in R." Sharp Sight, 26 Nov. 2019, www.sharpsightlabs.com/blog/scatter-plot-in-r/.
2. "General Social Survey, Cycle 31 : Families." Statistics Canada, Minister Responsible for Statistics Canada, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf.
3. McElreath, Richard, 2020, Statistical Rethinking, 2nd Edition, CRC Press.
4. "What Is Good Mental Health?" Mental Health Foundation, 5 Aug. 2016, www.mentalhealth.org.uk/your-mental-health/about-mental-health/what-good-mental-health.