Analysis on key factors affecting life satisfaction

Woolim Kim, Yena Joo, Guemin Kim

Oct 19, 2020

Analysis on key factors affecting life satisfaction

Code and data supporting this analysis is available at: https://github.com/Guemin/Problem_Set_2.git

Abstract

We use the 2017 General Social Survey(GSS) data obatained from the CHASS website to study and analyze some potential factors affecting Canadians' life satisfactions, as well as to observe the most significant factor among them. Linear model and adjusted R^2 values[1] are used to determine the significant factors, and plots are drawn to show the linear trend or to exaplain some weaknesses of the analysis. Through the analysis, We find positive linear relationships between a dependent variable, life_satisfaction rate and two independent variables self_rated_mental_health and self_rated_health, which means people who rate themselves as "mentally and/or physically healthy" are relatively more satisfied with their lives; on the other hand, there is no overall effects of hours worked on life satisfaction level. Together, these results suggest that no matter how much the income or working hours are, the most important factor that decides people's life satisfaction level is their mental health.

Introduction

There are many factors that determine one's well being and satisfaction of life. It could be health condition, economic status, relationships, religion, or any other element one would value or prioritize.
Particularly, in these uncertain and unprecedented times caused by the COVID-19 pandemic, many people across the world are feeling more stressed due to the changes that the pandemic has brought into their lives and those concerns negatively influence their own well being. Therefore, as the ongoing COVID-19 crisis reminds us of the importance of well being, our group decided to investigate and determine some factors that influence one's well being and satisfaction of life.

Throughout this report, we are going to determine and analyze some potential factors that affect one's life satisfaction, as well as to identify the most significant factor among them.
To be more specific, in the following sections, we will use statistical methods to build a regression model of life satisfaction score by potential factors, and interpret the regression output to find relationships between the life satisfaction score and potential factors.
This process includes cleaning the given data into a simpler, but efficient version, linear regression modelling, graphical visualizations and interpretations of the outputs.

Data

The data set we chose for this assignment contains responses of the General Social Survey conducted in 2017. The contents of the survey include some characteristics of diverse families in Canada, their socio-economic status, as well as other subjective information such as the respondent's life satisfaction and health conditions.

[1]*adjusted R^2 is explained in "Model" section

The target population in the GSS data includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The frame population is everyone who is registered combining both landline and cellular with Statistics Canada's address registers, and the sampled population is whoever is reached via telephone. The target population was divided into 27 strata by geographic areas, and simple random sampling without replacement of records was performed in each stratum (which means, from each stratum/group, everyone has an equal probability of being chosen).

Since we want to identify some key factors affecting one's life satisfaction as mentioned previously, the focus of our analysis will be "Health and subjective well-being".

The reason for choosing the 2017 GSS is because it is the most recent[2] survey that includes the "Health and subjective well being" topic.

One of the characteristics of our data set is that the majority of the variables are categorical. However, some drawbacks of using such data is that there is a limit to the kinds of statistical analysis that we can use with our data, as well as numerical operations or quantitative analysis cannot be performed on such data.
Not only this, but there are numbers of columns in our data set with too many NAs in them, and this indicates that many of the observations in certain variables are not available or missing. If we want to use such variables in our analysis, we would first need to exclude those 'NA' observations from our data; however, any results drawn from the data could possibly be biased or misleading due to the small number of observations available.

Since the original data set contains too many variables that are not necessary, we are going to clean the data set prior to analysis by removing them. Also, we are going to look for non-responses in our variables of interest and simply remove them from our data; removing the non-responses would not influence the overall performance of the regression model since we still have 13007 observations left in the new data.

These are the first six observations in our newly created data:

ID	satisfaction_score	self_rated_health	self_rated_mental_health	family_income	work_hours
1	8	5	5	$25,000 to $49,999	30.0 to 40.0 hours
2	10	3	3	$75,000 to $99,999	50.1 hours and more
5	8	3	3	$50,000 to $74,999	30.0 to 40.0 hours
9	8	4	4	Less than $25,000	30.0 to 40.0 hours
11	10	5	5	Less than $25,000	0.1 to 29.9 hours
12	6	4	3	$25,000 to $49,999	50.1 hours and more

Our data contains 6 variables: ID, satisfaction_score, self_rated_health, self_rated_mental_health, family_income, and work_hours.
Detailed descriptions on variables are provided in the footnote[3].

Since we want to observe how the life satisfaction score is related to potential factors such as health or financial conditions, the response variable of our analysis will be satisfaction_score and the predictors will be the potential factors: self_rated_health, self_rated_mental_health, family_income, and work_hours.

(Note: the scatter plot of the raw data is eliminated since the explanatory variables are categorical, which do not show a good visualization of the data)

[2]As it is stated in the documentation of the GSS, one of the primary objectives of the General Social Survey is to monitor the well being of Canadians over time. As a result, every survey conducted so far contain the responses related to the questions asking for the respondents' well being, and the 2017 GSS is the most recent survey with such responses.

[3]* satisfaction_score indicates the life satisfaction score on a scale of 0(very dissatisfied) to 10(very satisfied).
* self_rated_health and self_rated_mental health are the physical and mental health ratings, respectively, on a scale of 1(poor) to 5(Excellent) given by the respondent.
* work_hours indicates the average number of hours worked per week.

Model

Now, we are going to fit a multiple linear regression model in order to find linear associations between satisfaction_score and other predictor variables: self_rated_health, self_rated_mental_health, family_income, and work_hours, using R software.

Note, the equation for our regression line looks like this:

satisfaction_score $= \hat{B}_0 + \hat{B}_1 * x_{health2} + \hat{B}_2 * x_{health3} + \hat{B}_3 * x_{health4} + \hat{B}_4 * x_{health5} + \hat{B}_5 * x_{mental2} + \hat{B}_6 * x_{mental3} + \hat{B}_7 * x_{mental4} + \hat{B}_8 * x_{mental5} + \hat{B}_9 * x_{income2} + \hat{B_{10}} * x_{income3} + \hat{B_{11}} * x_{income4} + \hat{B_{12}} * x_{income5} + \hat{B_{13}} * x_{income6} + \hat{B_{14}} * x_{work2} + \hat{B_{15}} * x_{work3} + \hat{B_{16}} * x_{work4} + \hat{B_{17}} * x_{work5}$

(The detailed descriptions on the x-variables are found in the footnote[4].)

As it is shown above, we are going to have a very long equation for our regression line; however, this is inevitable since each of our predictor variables has several levels in them.

One thing we should notice in our data is that both the income and work_hours variables are categorical. To be more specific, we are given a range of values instead of an exact amount as income or average work hours in each variable. However, since our response variable is numerical, and we want to determine if each predictor has a linear relationship with it, we are going to treat our observations in both income and work_hours as numbers by replacing each category in income and work_hours with the midpoint.

This process will allow us to investigate the linear relationships between the life satisfaction and the two predictors numerically; however, there is a chance where the true values for income or work_hours could be very different from the midpoint. Therefore, we need to take into account when interpreting the regression model, that the two predictor variables could be biased and so does the result.

After finding some relationships between our response and predictor variables using the regression model, we are going to identify which predictor is the most significant factor in explaining the variability in the response variable. In other words, we are going to find out which of the potential factors (among physical health condition, mental health condition, income and average work hours) can explain the variability in life satisfaction the most.

[4]* x_{health_i} is a physical health rating indicator for i from 2 to 5
(i.e. $x_{health_5} = 1$ if the respondent's self_rated_health = 5, and $x_{health_5} = 0$ otherwise).
* x_{mental_i} is a mental health rating indicator for i from 2 to 5.
* x_{income_i} is an average income range indicator for i from 2 to 6
(i.e. $x_{income_2} = 1$ if the family income is in the second category "$25,000 to $49,999", and $x_{income_2} = 0$ otherwise).
* x_{work_i} is a working hours range indicator for i from 2 to 5
(i.e. $x_{work_2} = 1$ if the average hours of work is in the second category "0.1 to 29.9 hours", and $x_{work_2} = 0$ otherwise).

Results

Here is the summary of the multiple linear regression model:

Summary 1:

coefficients	estimates	p_values
(Intercept)	3.4455178	0.0000000
as.factor(self_rated_health)2	0.5147058	0.0000007
as.factor(self_rated_health)3	0.8347525	0.0000000
as.factor(self_rated_health)4	0.9858410	0.0000000
as.factor(self_rated_health)5	1.1752236	0.0000000
as.factor(self_rated_mental_health)2	1.6327434	0.0000000
as.factor(self_rated_mental_health)3	2.6995119	0.0000000
as.factor(self_rated_mental_health)4	3.2449885	0.0000000
as.factor(self_rated_mental_health)5	3.7214505	0.0000000
as.factor(family_income)37499.5	0.1135077	0.0226291
as.factor(family_income)62499.5	0.3340758	0.0000000
as.factor(family_income)87499.5	0.3590310	0.0000000
as.factor(family_income)112499.5	0.4795057	0.0000000
as.factor(family_income)125000	0.5281599	0.0000000
as.factor(work_hours)15	0.2960824	0.2774738
as.factor(work_hours)35	0.2325230	0.3922313
as.factor(work_hours)45.05	0.3205580	0.2409347
as.factor(work_hours)50.1	0.3947597	0.1503998

With the estimates from the regression output, we know that our regression line has a following equation:

$$\text{Satisfaction Score} = 3.44552 + 0.51471 * x_{health2} + 0.83475 * x_{health3} + 0.98584 * x_{health4} + 1.17522 * x_{health5} + 1.63274 * x_{mental2} + 2.69951 * x_{mental3} + 3.24499 * x_{mental4} + 3.72145 * x_{mental5} + 0.11351 * x_{income2} + 0.33408 * x_{income3} + 0.35903 * x_{income4} + 0.47951 * x_{income5} + 0.52816 * x_{income6} + 0.29608 * x_{work1} + 0.23252 * x_{work2} + 0.32056 * x_{work3} + 0.39476 * x_{work4}$$

(*Note, you can find the descriptions for x-variables from the footnotes in the previous page.)

As we can observe from the output, estimated satisfaction_score increases as each of self_rated_health, self_rated_mental_health and family_income increases; however, the slope estimates for work_hours are found to be quite inconsistent, because there is a decrease in slope estimates from 0.29608 to 0.23252 in the first two categories of work_hours, but they increase again in the third and fourth categories.
Furthermore, unlike self_rated_health, self_rated_mental_health, and family_income where the p-values are much less than the significance level of 0.05, p-values of work_hours are greater than 0.05; this finding provides us with more evidence that there is no linear relationship between the average working hours and the life satisfaction score. Hence, the regression model suggests that only the physical health, mental health and financial conditions have positive linear relationships with the life satisfaction score.

Now that we've found that the life satisfaction score has linear relationships with physical health, mental health, and financial conditions, we want to ask ourselves: are those factors equally important in terms of explaining the variability in satisfaction_score?

The answer is 'No'.
Although they all have positive linear relationships with the response variable, not all variables may contribute significantly in explaining the variability of the response variable.
Hence, we would now like to identify the most important predictor variable in this regression model.

There are two ways to do this[5]. In this analysis, specifically, we are going to use the method where we

[5]One way is to compare the standardized regression coefficients, and the other way is to compare the increases in adjusted R^2 when each predictor is added to the regression model. For this analysis, specifically, we cannot use the first method, because our model contains categorical predictor variable which cannot be standardized.

compute and compare the changes in adjusted R^2 for the last variable added to the model[6]. This is a valid method for identifying which predictor explains the most variability in the response variable, because by the definition, R^2 gives the percentage of variation in the response variable explained by the regression line, and also, if the newly added predictor variable is the only difference between the two models, the associated change in 'adjusted R^2' will represent the 'goodness-of-fit'.

We are going to begin with fitting a linear model with only one predictor, and then add another predictor to the model at a time to see how much $R^2_{adjusted}$ changes when each variable is added.

Key point in this method is to identify the predictor variable with the largest increase in $R^2_{adjusted}$ when it is the last variable added to the model.

This is the table with $R^2_{adjusted}$ values in each model, and the change in $R^2_{adjusted}$ as more variables are added:

Table 1:

Predictors	R_squared_adjusted	Change
self_rated_health	0.1172437	0.1172437
self_rated_health + self_rated_mental_health	0.2395756	0.1223319
self_rated_health + self_rated_mental_health + family_income	0.2533018	0.0137262

As it is shown in the table, there is a greatest increase in $R^2_{adjusted}$ when the second predictor, self_rated_mental_health is added to the model. This increase is quite close the $R^2_{adjusted}$ of our initial simple linear regression model with a predictor, self_rated_health.

On the other hand, the change in $R^2_{adjusted}$ is relatively small when family_income is added to the regression model compared to the previous changes, and therefore, we know that the family income does not contribute significantly in explaining the variation in the satisfaction score.

[6] $R^2_{adjusted} = 1 - (1 - R^2) * (\frac{n-1}{n-p-1})$ where n is the total sample size and p is the number of additional predictor variables. Hence, unlike R^2 that will be inflated as a new predictor is added to the model, regardless of its significance, if the newly added predictor does not explain variation in the response variable well, $R^2_{adjusted}$ will go down.

Discussion

The goal of this process is to find some factors that affect one's satisfaction on life. Data cleaning of the 2017 General Social Survey(GSS), obtained from the CHASS website, is done in the "Data" part, but there may have some bias since many responses are not available. Therefore, We focused more on the variables with less NA responses to do the anlaysis as accurate as possible. Further discussions on the biases that our data may contain can be found in the "Weaknesses" section.

Here is some graphical visualization of positive correlation of the explanatory variables used in the regression models and life satisfaction score:

Figure 1. Boxplot of Satisfaction score vs. Self rated Mental health

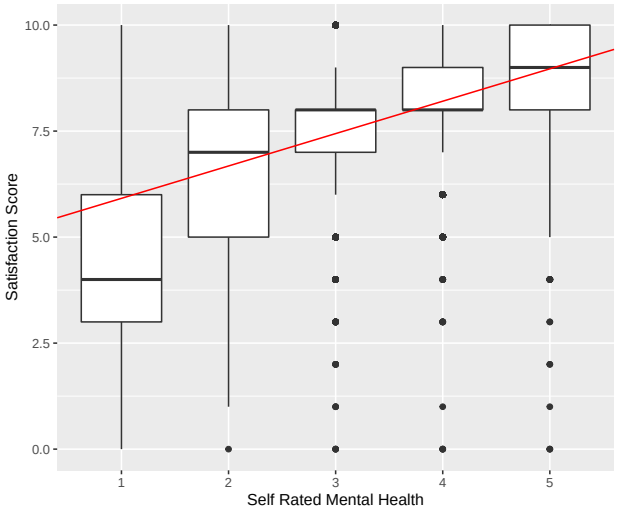


Figure 2. Boxplot of Satisfaction score vs. Work hours

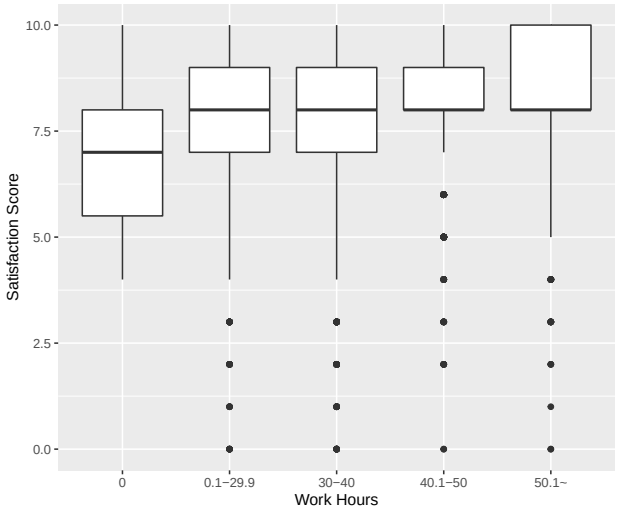


Figure 3. Boxplot of Satisfaction score vs. Self rated Health

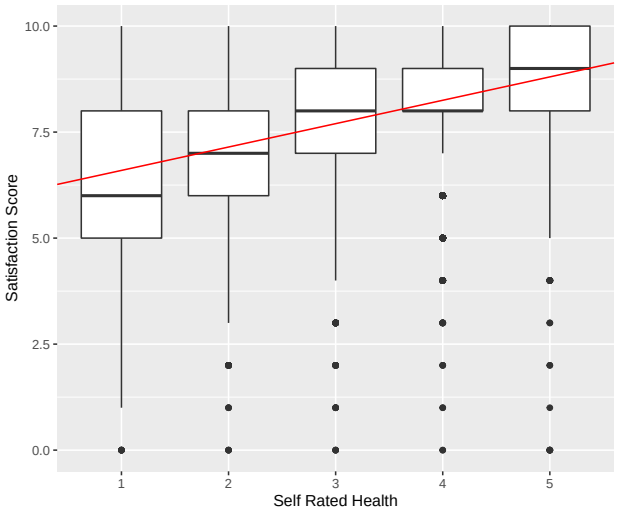
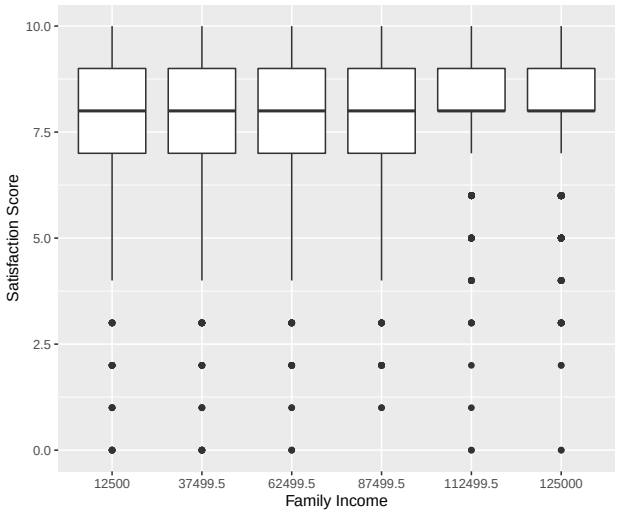


Figure 4. Boxplot of Satisfaction score vs. Family income



In Figure 1, The boxplot has a positive linear relationship between the two variables, mental health and life satisfaction. It signifies that average of life satisfaction score increases as the self rated mental health increases.

Figure 2 shown above is the boxplot of satisfaction_score and work_hours. Work_hours variable includes 0 hour to 50 hours and more, which means that both unemployed and employed are included in this category. As it was mentioned previously, employed hours show no positive or negative relation with the satisfaction score signifying that there is no significant relationship with satisfaction score.

Figure 3 is a boxplot of satisfaction_score and self_rated_health, and it shows a positive linear relationship

between the two variables. The slope of the red line shows weaker change in 1 unit of health rate than what shows in figure 1, which is mentioned previously that self_rated_health is not as well related as self_rated_mental_health.

Figure 4 shows a consistent median of satisfaction score over all income ranges, which shows family_income and satisfaction_score does not have a significant linear relationship. The boxplot suggests that family income does not contribute much in explaining the variation in satisfaction score.

From Table 1 and Summary 1 in the result section, we find that physical health also has a significant positive linear relationship (using estimates and p-value, and r-squared). Which explains, that physical and mental health are the two major factors that affect life satisfaction. However there is no strong relationship shown between income, hours worked and life satisfaction(from summary 1 using p-values and estimates). Hence, the result suggests that both physical and mental health are significant predictors. Between the two variables, mental health is a more important factor than physical health in terms of explaining the variability of the life satisfaction score.

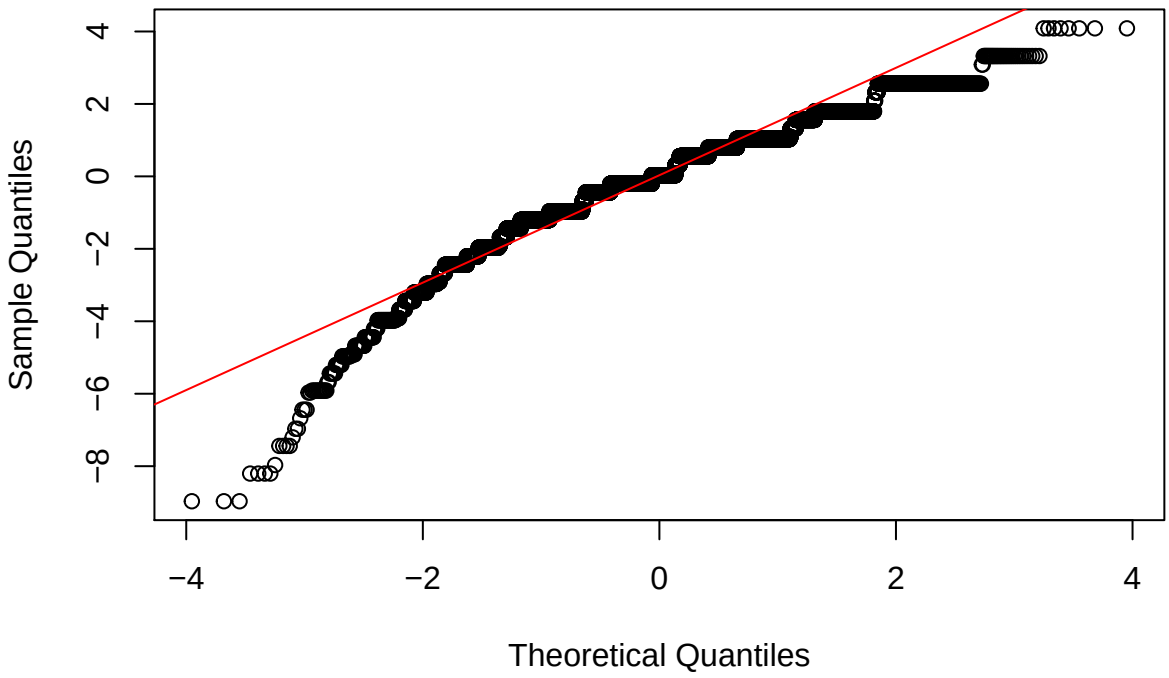
Weaknesses

Every statistic data study and analysis includes some biases and weaknesses. One of our weaknesses is that the predictor variables are categorical variable, which made it challenging to visualize the linear regressions model we found and show the significance between the life satisfaction score and other factors. In order to overcome these problems, we replaced the ranges given in the categorical predictor variables with their midpoints.
This allowed us to perform quantitative analysis on our data; however, we still need to take potential biases into account when interpreting the result, because the true values for the categorical predictor variables could be far away from their newly assigned midpoints.

Another possible weakness is that we do not know if there is any omitted variable bias. Omitted variable should be correlated with the dependent variable, and correlated with the explanatory variables included in the model. There might be an important variable that would affect the model, but it is hard to figure out since the variable might be missing in our data set, or might be impossible to measure.

Also, the units of self_rated_health and self_rated_mental_health are based on different standards of the participants and might not be a reliable measurement to analyze. There might be some errors in the measurement of the variables since it is not a measurable unit.

Figure 5. Normal Q−Q Plot



The normal QQ plot shown above graphically analyze the residuals in our regression model.
The x-axis represents the quantiles for the standard normal distribution and the y-axis are the data points for the residuals.
As we can see in the plot, the data points do not trend the theoretical line, and the points at each tail of the data seem to fall off the line, revealing that the distribution of residuals may have long tails.
Normality is one of the assumptions in the linear regression model. However, the normal QQ plot above suggests that our model does not satisfy the normality assumption on the error terms. Therefore, we need to take into account that the result drawn from the regression model could be misleading or biased.

Next Steps

For the next steps, we could do a follow up survey on the related topic (life satisfaction vs mental and physical health), and compare the data of the prior and post COVID-19. We could see how people's mental health changed due to COVID-19, the life satisfaction rate is expected to decrease in this case. Also, we could seek for interaction effects between few independent variables and do ANOVA tests to figure out if there is any independant variable that is dependent to another independant variable. (ie) how age and mental health rate could be interactive).

Furthermore, since we have found in the previous section that the normality assumption was violated in our model, we could fit other models that satisfy all of the required assumptions on linear regression model, and find the best model among them.

References

1. 2017 GSS Data: General Social Survey, Cycle 31: 2017: Family. (n.d.). Retrieved October 16, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

2. 2017 GSS Data Documentation: General Social Survey Cycle 31 : Families - Public Use Microdata File Documentation and User's Guide. (2017). Retrieved October 17, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

3. Data Cleaning Code: Alexander, Rohan, and Sam Caetano.(2019, Sept 16). "gss_cleaning.R". Retrieved Oct 10. 2020, from https://www.tellingstorieswithdata.com/01-03-r_essentials.html

4. Identifying the most significant predictor: Frost, J., Adusei, C., Peeyush, Siyabonga, Sreeja, Narayanan, J., . . . Sachin. (2019, June 13). Identifying the Most Important Independent Variables in Regression Models. Retrieved October 16, 2020, from https://statisticsbyjim.com/regression/identifying-important-independent-variables/

5. QQ normal plot code: Sheather, S. J. (2010). A Modern approach to regression with R. New York: Springer.

6. Plotting multiple plots on one page: R Draw Multiple ggplot2 Plots Side-by-Side (Example): Plot on One Page. (2020, September 30). Retrieved October 17, 2020, from https://statisticsglobe.com/draw-multiple-ggplot-plots-side-by-side

7. Knitr package: Xie, Y. (n.d.). Knitr v1.30. Retrieved October 19, 2020, from https://www.rdocumentation.org/packages/knitr/versions/1.30

8. Wrapping column name in pdf table using package "kableExtra": Feder, A. (2018, April 10). Wrap column name in pdf table, from knitr::kable. Retrieved October 19, 2020, from https://community.rstudio.com/t/wrap-column-name-in-pdf-table-from-knitr-kable/3278/4 (apa6)