Exploratory Data Analysis

S. Arora, J. Harmse, V. Mulholland

April 22, 2018

```
library(tidyverse); theme_set(theme_bw())
library(cowplot)
library(ggjoy)
library(reshape2)
library(pwr)
```

Overview

The purpose of this study to measure whether a person is driven by money or not. We found it reasonable to assume that a person who is driven by money would expect to earn more than the average person who has the same skillset and experience.

Our survey has captured the salary of what a participant thinks an average person with their skills and experience should earn, as well as the salary that the participant expects to receive in 1 year's time. The average inflation rate world-wide was of approximately 3.15 per the Statista Statistics Portal. Taking inflation and other micro-factors into account, a participant's expected salary in a year's time shouldn't be much higher than the average person with the same skills and experience.

The survey captured the participant's salary in their unique currency. The survey was answered by people from various countries with different currencies. This means that we cannot compare the captured salary values between participants. An easy way of standardising these values is to handle the salary values as a ratio of expected salary over average salary. The ratio should be consistent across different currencies.

For the purpose of this study, social standards will be defined as a person's inclination for a high relative consumption on leisure activities and non-essential expenditure. Our hypothesis relies on the theory that prevailing social conditions will influences one's relationship with money which would translate in whether increase in income is the priority. The study is inspired by Hirsh's hypothesis which states that prevailing social conditions will influences one's spending on status goods and activities (status-seeking) which would translate in those with a higher disposable income being more inclined to spend on non-essential expenditures.

Data Pre-processing

Anonymity

In order to maintain user privacy a few manipulations were handled before the raw data was uploaded to the analysis repository. Any confidential information such as IP addresses were ommited, as well as any respondents that did not accept the confidentiallity agreement.

Pre-processing Workflow

These were the first steps applied to `surveydata_clean.rds` when the data was downloaded raw from *Survey Monkey*.

```
# removing confidential data
#survey_results <- read_csv(file = '../../survey_data/Demographic Survey.csv', skip = 1)
survey_results <- survey_results[, 10:ncol(survey_results)]

#import data
# survey_results <- read_csv(file = '../../survey_data/Demographic Survey.csv') # local path - remove i

# redefine column names
colnames(survey_results) <- c('consent', 'country', 'salary_base', 'salary_expect', 'no_increase_accepta
                            'living_expenses', 'savings', 'vacation', 'daily_leisure', 'consumption_good
                            'sports_hobbies', 'other')

# spending categories
spending_cats <- c('living_expenses', 'savings', 'vacation', 'daily_leisure', 'consumption_goods',
                            'sports_hobbies', 'other')

# remove no consent
survey_results <- survey_results %>% filter(consent %in% c('Yes'))

# add observation id
survey_results$id <- 1:nrow(survey_results)

# save raw clean data
saveRDS(survey_results, file = '../data/processed/surveydata_clean.rds')

# remove all traces
rm(survey_results)
```

Once the data is pre-processed, it is reimported and the columns and categories are defined.

```
# import clean data
survey_results <-  readRDS(file = '../data/processed/surveydata_clean.rds')
# redefine column names
colnames(survey_results) <- c('consent', 'country', 'salary_base', 'salary_expect',
                               'no_increase_acceptance',
                               'living_expenses', 'savings', 'vacation', 'daily_leisure',
                               'consumption_goods',
                               'sports_hobbies', 'other', 'id')

# spending categories
spending_cats <- c('living_expenses',
                   'savings',
                   'vacation',
                   'daily_leisure',
                   'consumption_goods',
                   'sports_hobbies',
                   'other')
```

The way that we are measuring a person's drive for financial weatlh is by comparing a person's expected salary to the base salary the person would expect to earn. We are using the equation $(Salary_{expect} - Salary_{base})/Salary_{base}$. In most cases a person would not expect to receive a large salary increase in a year's time. If a person's salary ratio is higher than that of the average person, it could likely be due to personal drive above and beyond that of a similar person with no expectional drive for earning a higher salary.

```
# ensure any NA values are set to 0
survey_results[ , spending_cats][is.na(survey_results[,spending_cats])] <- 0

# converting char to numeric
survey_results$salary_base <- as.numeric(as.character(survey_results$salary_base))
survey_results$salary_expect <- as.numeric(as.character(survey_results$salary_expect))


# add ratio
survey_results <-
  survey_results %>%
  mutate(ratio = (salary_expect - salary_base)/salary_base)
```

Outlier Handling

Having chosen to remove outliers on the basis that with a small number of observations applying the statistical method of removing outliers greater than two standard deviations could be erroneous since it cannot be deduced with certainty which distribution is being represented. That being said, a combination of visual assessments and box-plot/quantile analysis allowed a reasonable upper and lower limit to be chosen.

```
## saving a copy with all outliers.
survey_results_all_outliers <- survey_results

outliers <- boxplot.stats(survey_results$ratio)$out
dh<- data.frame(outliers)
knitr::kable(dh)
```

outliers
0.4500000
0.6000000
0.8000000
11.0000000
9.0000000
1.5000000
-0.5254237
-0.7500000
0.8181818
-0.9100000
9.0000000
2.3333333

It was decided to remove the values beyond ~95% confidence level. The box-plot method performs a more sophisticated outlier selection than the alternative, the quantile approach, that is more rigid in the 95% threshold. Since we have less observations than ideal, it seemed more appropriate. The visualization below shows the contrast when the most extreme outliers are removed.

Remove Outliers

```
# remove outliers
survey_results <- survey_results %>%
  filter(!ratio %in% boxplot.stats(survey_results$ratio)$out)
```

Divide into GROUP A and GROUP B

As mentioned, our hypothesis is comparing people who are driven by financial wealth with people who are not driven by financial wealth. Up until now we have quantified financial wealth as a measurement of salary. Even though the salary ratio is a fairly good measurement, it has its flaws. It could be the case that a person expects a higher salary, not because of financial drive, but rather because of something like good job performance, and thus representing satisfaction rather than financial imperatives. In this case, a person with a high salary ratio would not necessarily be a person who is driven by financial wealth.

To make the criteria more strict for how we define someone who is driven by financial wealth or not, we take into account whether the person would maintain a job if they do not receive a salary increase in 2 years, even though the job satisfaction is high. A person who is driven by financial wealth, would most likely not stay in a job that doesn't provide them with a realistic salary increase. Whereas a person who is more motivated by job satisfaction would stay in the job position, regardless of not receiving an increase.

The respondants are divided into two groups - those who expect a salary increase and have a high salary ratio, and those who are content with high job satisfaction and have a low salary ratio. The respondants who do not fall in one of these two groups cannot be classified as either a person who is driven by money or someone who is not driven by money (it may be a more complex classification).

```
# grouping by wealth or not wealth groups
survey_results <- survey_results %>% mutate(group = ifelse(no_increase_acceptance == "No" & ratio > medi
                                              group_bin = ifelse(group == "wealth", 1, ifelse(group == "no

survey_results_grouped <- survey_results %>% filter(!group %in% c("other"))
```

The questions were designed to minimize the potential for entry mistakes when participants entered their responses. A rule was included to ensure that the expenditure percentages summed up to 100 points, but this was not possible with the user salary through the *Survey Monkey* interface. This process of removing outliers will filter out major mistakes in currency where the user entered that they expected a very disproportionate salary increase.

Below each variable is summarized. Since it is difficult to highlight important information from a summary table containing so many variables, a jitter-violin plot was also generated.

```
sum.tb <- summary(survey_results)
sum.tb
```

```
##    consent             country            salary_base
##   Length:71          Length:71          Min.   :    3000
##   Class :character   Class :character   1st Qu.:   70000
##   Mode  :character   Mode  :character   Median :   80000
##                                         Mean   : 1166042
##                                         3rd Qu.:  117500
##                                         Max.   :65000000
##
##   salary_expect      no_increase_acceptance living_expenses    savings
##   Min.   :    3000   Length:71              Min.   : 0.00   Min.   : 0.00
##   1st Qu.:   75000   Class :character       1st Qu.:25.00   1st Qu.: 7.00
##   Median :   90000   Mode  :character       Median :40.00   Median :10.00
##   Mean   : 1265775                          Mean   :38.92   Mean   :14.27
##   3rd Qu.:  120000                          3rd Qu.:50.00   3rd Qu.:20.00
##   Max.   :70000000                          Max.   :90.00   Max.   :50.00
##
##     vacation      daily_leisure   consumption_goods sports_hobbies
##   Min.   : 0.000   Min.   : 1.00   Min.   : 0.000    Min.   : 0.000
##   1st Qu.: 5.000   1st Qu.: 5.00   1st Qu.: 5.000    1st Qu.: 4.000
```

```
##  Median :10.000   Median :10.00   Median :10.000    Median : 5.000
##  Mean   : 9.394   Mean   :12.45   Mean   : 8.366    Mean   : 6.268
##  3rd Qu.:10.000   3rd Qu.:17.50   3rd Qu.:10.000    3rd Qu.:10.000
##  Max.   :30.000   Max.   :60.00   Max.   :30.000    Max.   :25.000
##
##      other             id            ratio              group
##  Min.   : 0.00   Min.   : 2.00   Min.   :-0.25000   Length:71
##  1st Qu.: 5.00   1st Qu.:24.00   1st Qu.: 0.00000   Class :character
##  Median :10.00   Median :45.00   Median : 0.06250   Mode  :character
##  Mean   :10.34   Mean   :44.41   Mean   : 0.05949
##  3rd Qu.:10.00   3rd Qu.:65.50   3rd Qu.: 0.12500
##  Max.   :66.00   Max.   :83.00   Max.   : 0.38462
##
##    group_bin
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :0.0
##  Mean   :0.3
##  3rd Qu.:1.0
##  Max.   :1.0
##  NA's   :31
```

Analysis

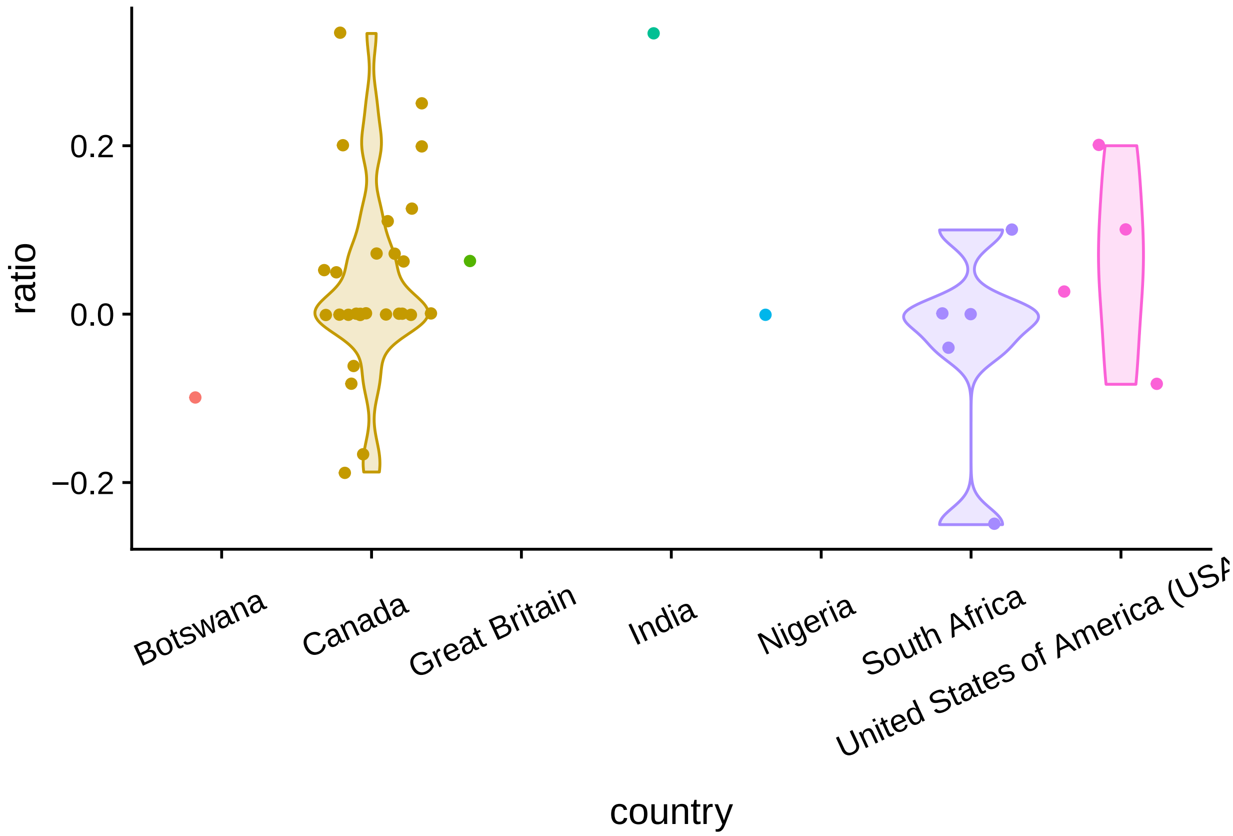
We want to determine whether people who are driven by wealth have different spending habits than people who are not driven by wealth. The nature of the hypothesis and the data makes a logistic regression model the obvious choice for determining whether any expense category is significantly different between the two groups.

Before we start building the model we should consider whether we are dealing with any confounding variables.

Each person who completed the survey had to report their country of residence. Data was collected from a number of different countries. The study's response variable is standardised salary ratio, which does not require taking the person's country into account. However, the country a person live has the potential to play a role in a person's spending habits regardless of their salary ratio. For example, a person from Africa would not necessarily spend a lot on vacations in comparison to a person from North America which may be a result of something like cultural differences.

We need to consider whether country has any significant effect on either out explanatory expense variables or our response variable.

```r
ggplot(data = survey_results_grouped, aes(x = country, y = ratio, colour = country, fill = country)) +
  geom_jitter() +
   geom_violin(alpha = 0.2) +
    theme(axis.text.x = element_text(angle = 25, hjust = 0.7, vjust = 0.8), legend.position = "none")
```



After removing the observations that do not match our criteria for either a person who is driven by wealth or not driven by wealth, we are left with a relatively small number of observations. The visualization above shows a number of single observations for different countries, but quite a few observations from Canada, South Africa and the US.

The simplest way of determining whether country has an influence on our model, we should influde a person's country as an explanatory variable and determine whether this variable has any statistical significance. Country is a categorical variable which makes the single observations difficult to work with when trying to determine its effect. A arguably logical solution to the handling of the single observations would be to group the countries by similarities. Seeing that Canada and the United States of America are neighbouring countries it we can group these two countries as `North America`.

South Africa, Nigeria and Botswana have a lot in common in terms of cultural lifestyle, which means that we could group these observations as `Africa`. The other single observations should be omitted for this comparison, because it would require arbitrary assumptions where these data points would fit in.

```
survey_results_grouped <- survey_results_grouped %>% mutate(continent =
                                                     ifelse(country %in% c("United States of Am
                                                                           "Canada"),
                                                            'North America',
                                                            ifelse(country %in% c("Botswana",
                                                                                  "Nigeria",
                                                                                  "South Africa
                                                                   'Africa',
                                                                   NA)))

survey_results_continent <- survey_results_grouped %>% filter(!is.na(continent))
```

```r
# testing the effect of continent
lm_confound_not <- glm(group_bin ~ living_expenses +
                  savings +
                  vacation +
                  daily_leisure +
                  consumption_goods +
                  sports_hobbies +
                  other,
                data = survey_results_continent,
                family = binomial(link = "logit"))

# summary(lm_confound_not)

# testing the effect of continent
lm_confound <- glm(group_bin ~ living_expenses +
                  savings +
                  vacation +
                  daily_leisure +
                  consumption_goods +
                  sports_hobbies +
                  other + continent,
                data = survey_results_continent,
                family = binomial(link = "logit"))

# summary(lm_confound)

anova(lm_confound_not, lm_confound, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: group_bin ~ living_expenses + savings + vacation + daily_leisure +
##     consumption_goods + sports_hobbies + other
## Model 2: group_bin ~ living_expenses + savings + vacation + daily_leisure +
##     consumption_goods + sports_hobbies + other + continent
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        31     38.156
## 2        30     34.329  1   3.8262  0.05046 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our analysis of variance test shows that the continent of a pariticipant does not significantly influence the model. The probability value isn't significant, but it is extremely close to our 0.05 significance threshold. To be safe, it would be better to influde the confounding variable in our model.

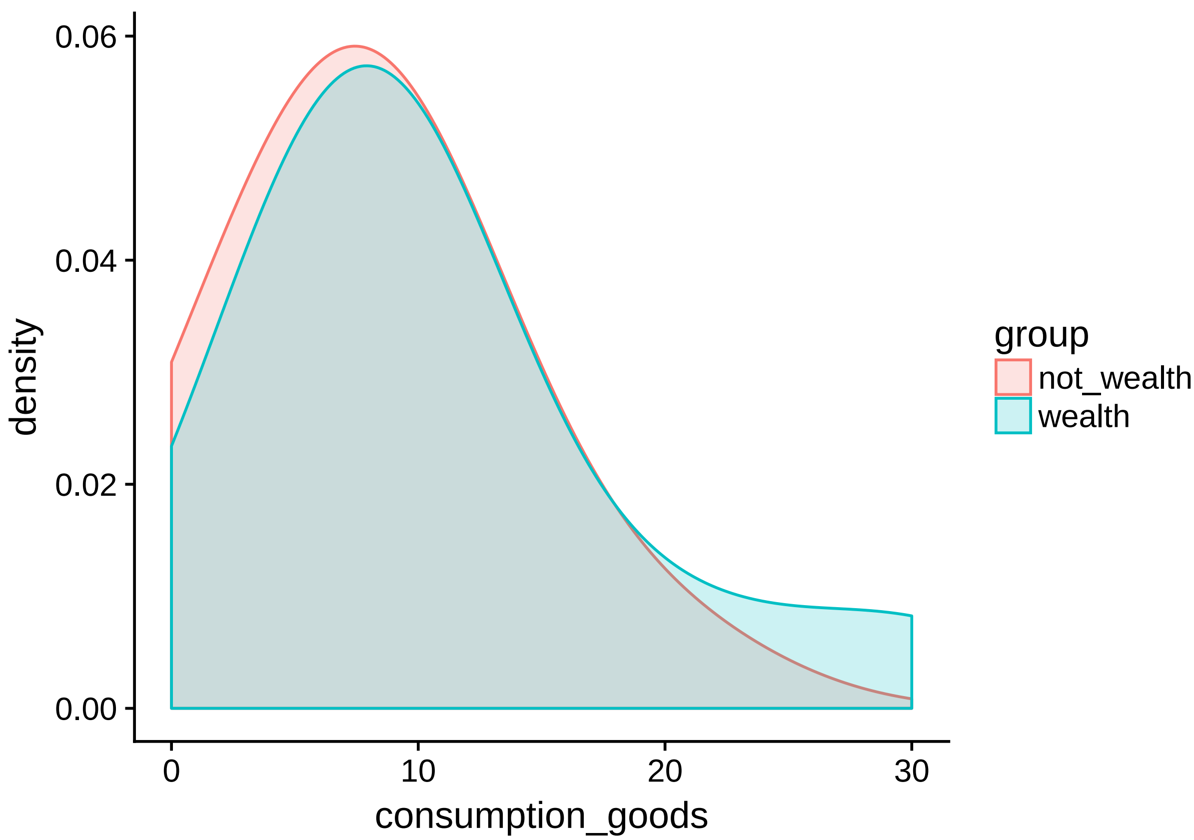
```r
summary(lm_confound)
```

```
##
## Call:
## glm(formula = group_bin ~ living_expenses + savings + vacation +
##     daily_leisure + consumption_goods + sports_hobbies + other +
##     continent, family = binomial(link = "logit"), data = survey_results_continent)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6035  -0.6246  -0.4260   0.8331   2.4827
```

```
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -15.27124    8.41726  -1.814   0.0696 .
## living_expenses         0.12446    0.08338   1.493   0.1355
## savings                 0.04958    0.08715   0.569   0.5694
## vacation                0.10286    0.10092   1.019   0.3081
## daily_leisure           0.04807    0.10355   0.464   0.6425
## consumption_goods       0.26942    0.13456   2.002   0.0453 *
## sports_hobbies          0.33138    0.17093   1.939   0.0525 .
## other                        NA         NA      NA       NA
## continentNorth America   2.75956    1.70609   1.617   0.1058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.728  on 37  degrees of freedom
## Residual deviance: 34.329  on 30  degrees of freedom
## AIC: 50.329
##
## Number of Fisher Scoring iterations: 6
```

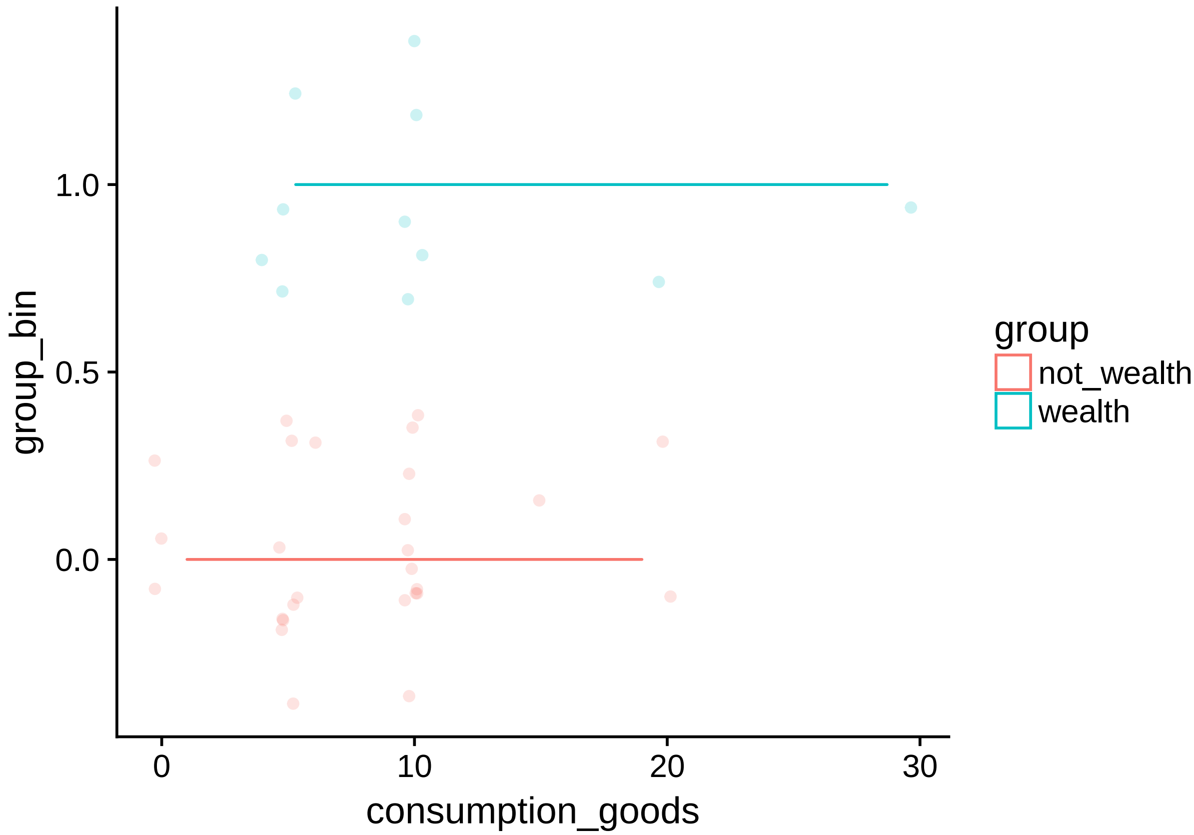
Consumption goods shows some significance. This indicates that people who are classified as driven by wealth tends to spend more on consumption goods than people who are not driven by wealth according to our data.

```
ggplot(data = survey_results_continent, aes(x = consumption_goods, group = group, colour = group)) +
  geom_density(alpha = 0.2, aes(fill = group), bw = 5)
```



```
ggplot(data = survey_results_continent, aes(x = consumption_goods, y = group_bin, group = group, colour
  geom_jitter(alpha = 0.2, aes(fill = group)) +
  geom_violin(scale = "width")
```

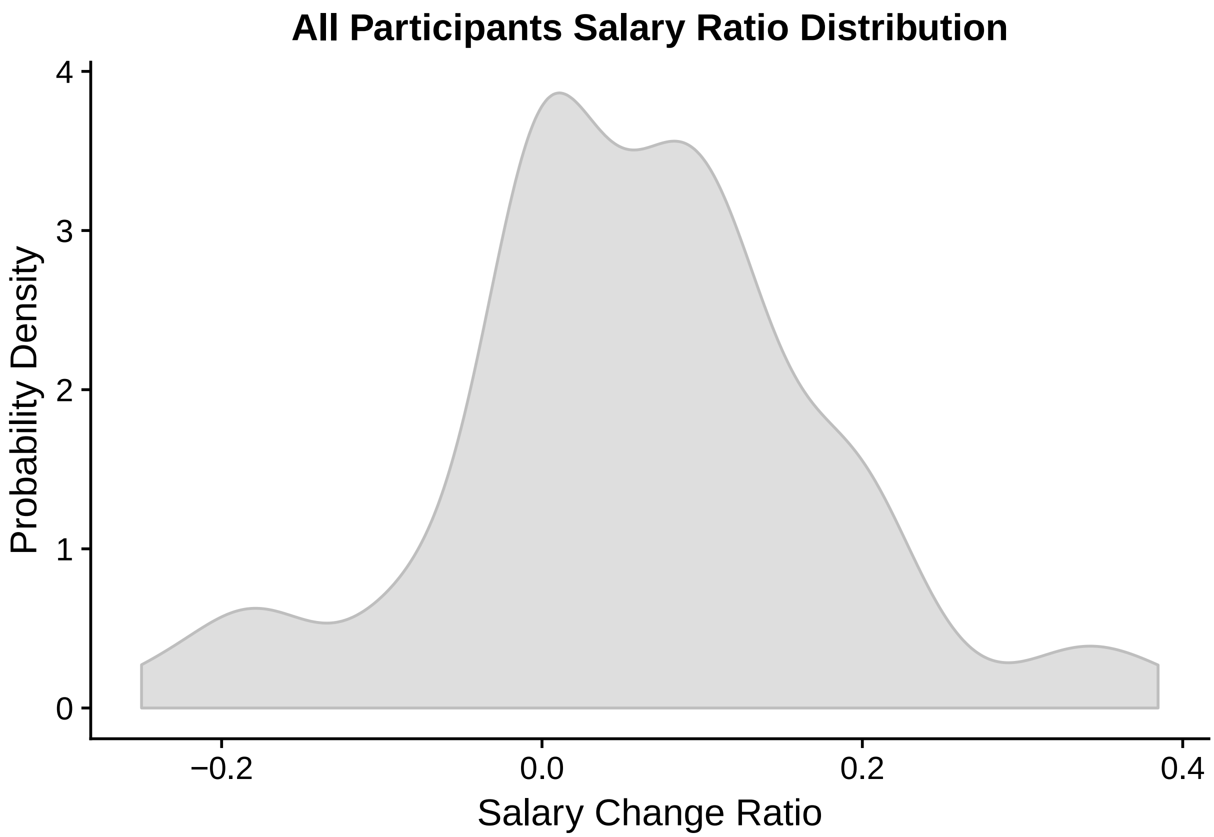
```
## Warning: position_dodge requires non-overlapping x intervals
```



The density plots show a slight difference between the two groups. The amount of data that we have available is fairly limited and makes it difficult to draw accurate conclusions.

Our assumption seems to be accurate with regards to countries not varying too greatly in their responses. There is no country that has a significantly higher or lower ratio distribution. As a sanity check, it is a good idea to combine survey answers from all participants to verify that the variance around our mean is somewhat normally distributed (the plot above makes it seem intuitive that this would be the case, but cannot make the assumption). This would verify that we are dealing with a t-distribution.

```
ggplot(data = survey_results, aes(x = ratio)) +
  geom_density(colour = 'grey', fill = 'grey', alpha = 0.5) +
  labs(x = 'Salary Change Ratio', y ='Probability Density', title = 'All Participants Salary Ratio Dist
```



Evaluating the Response

The study is interested in the ratio distribution above. Is there any correlation between the above ratio and social standards? The premise of the study was to develop a metric that would indicate the inclination of individuals to see financial gain as the main driver for success and determine if there is a relationship with the way their income is spent. Three variables were collected that pertain to our model's dependent variable which include:

Dependent Features	Description
`salary_base`	An indicator meant to be a subjective baseline of what salary a person of their expertise would earn.
`salary_expect`	The expected salary combined with the base salary provides a relative indicator to the respondents pursuit of monetary gains.
`no_increase_acceptance`	A binary metric serves as a safety check against false positives, that is respondents that may have over-exagerated their expected salary skewing the impression of interest in monetary gain while in reality being content with their current situation.
`ratio`	This is a calculated metric that simplifies handling respondent's country selection.

The survey also captured the percentages of the main expenses of each participant. Each participant had to assign percentages that adds up to 100%. The different expense categories were strategically chosen which are believed to relate to a person's social standards. For example, it is believed that a person who spends a large percentage on vacations and daily leisure most likely has higher social standards than a person who contributes most of their salary to savings. The hypothesis is that a person with higher social standards will have a higher salary ratio as described above.

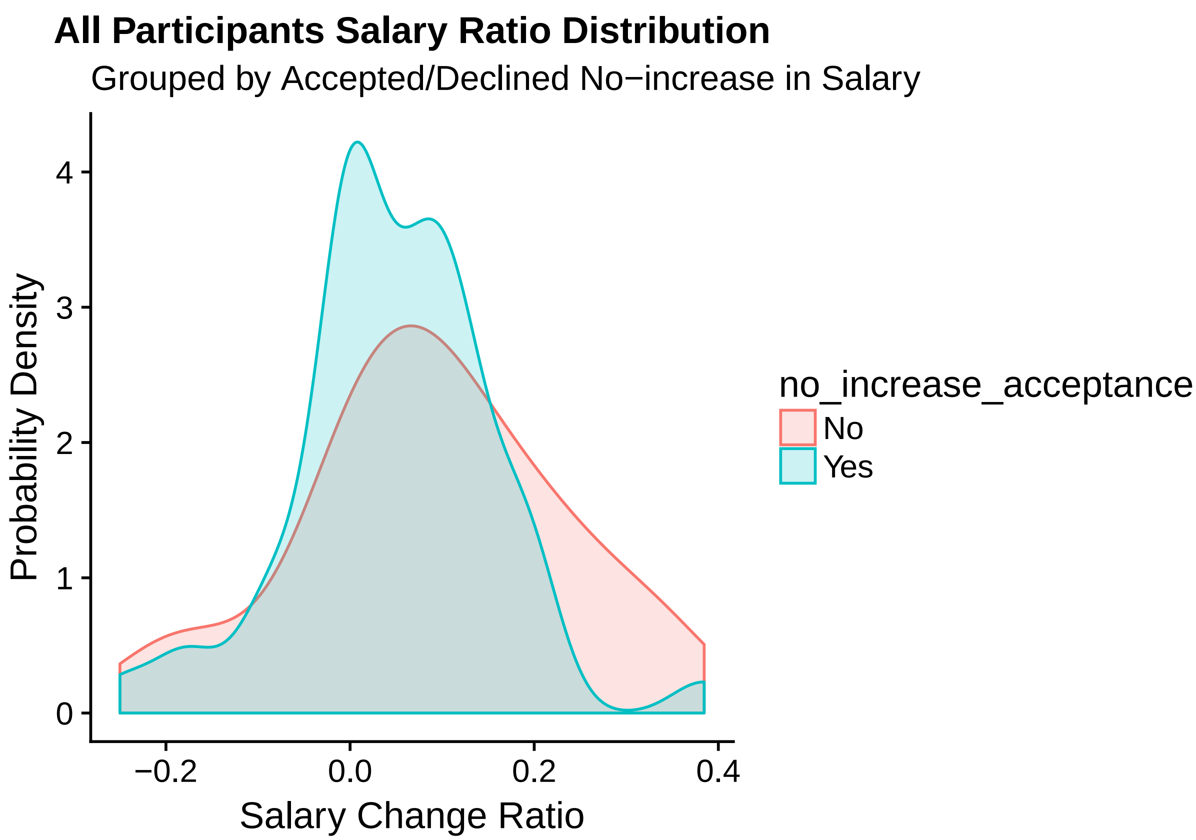
In theory this makes sense to simply compare these expense percentages to the salary ratios and look for any significant correlation. But in the real world there are many confounders that have to be accounted for. For example, a person who is close to retirement will most likely not expect an increase in the coming year, but may spend a large portion of their salary on vacations and daily leisure.

It isn't always as clear-cut as to say that the closer you are to retirement, the more you will spend on vacation. Or on the other side of the spectrum, it cannot be assumed that a young person won't spend a large percentage of their income on traveling.

The first confounder that we believe is of importance, is whether a person prefers job satisfaction over an increase in salary. The survey raised the question whether a person would keep their job if they don't receive a salary increase in two years, given high job satisfaction.

A person who spends a lot on vacation and leisure (which can be either the younger or older generation) may strive for a higher salary, but the possibility exists that they don't - possibly depending whether they value job satisfaction over a salary increase.

```
ggplot(survey_results, aes(x = ratio, group = no_increase_acceptance, colour = no_increase_acceptance))
  geom_density(aes(fill = no_increase_acceptance), alpha = 0.2)  +
  labs(x = 'Salary Change Ratio', y ='Probability Density', title = 'All Participants Salary Ratio Dist
```



The plot above shows similar salary ratio distributions for participants who prefer high job satisfaction as those who prefer a salary increase. It does seem as if a person who has a higher salary ratio has a higher probability of prefering an increase over job satisfaction, even though this probability is not significant. However, it will be of more importance if the distributions looked different for people with different types of expenses.

It is difficult to visualize the interaction between expenses, salary ratio and job satisfaction versus salary increase preference. It seems more logical and of statistical importance to fit comparitive models and observe whether the confounder variable adds any value to the model.

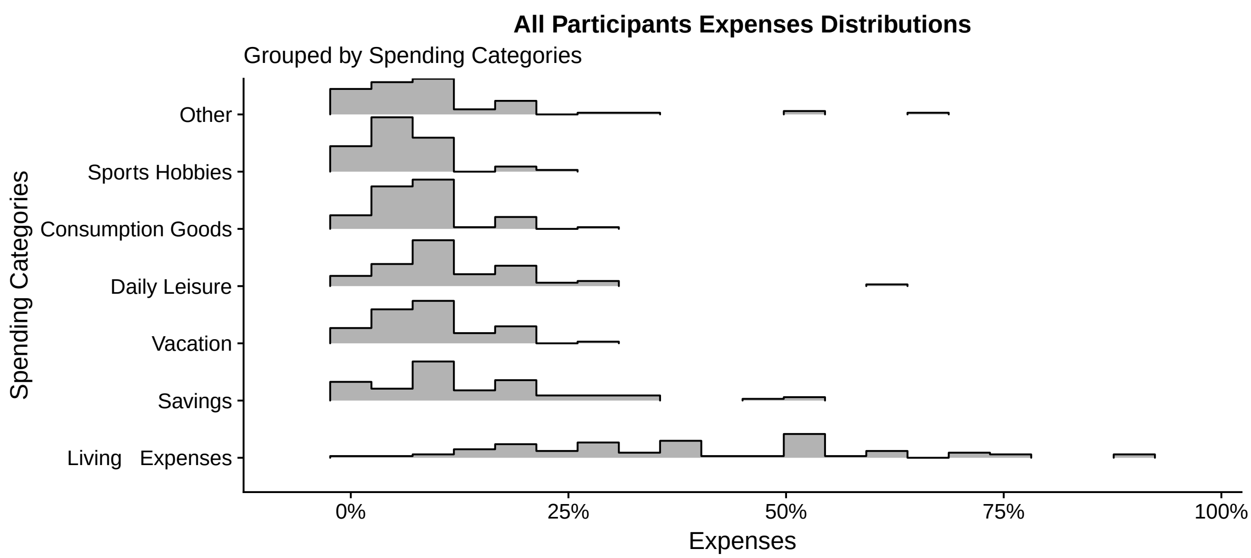
The salary ratio is a continuous variable and from our ratio probability distribution earlier, we saw that the standard deviation is fairly normally distributed around the mean after removing outliers. For this reason a linear regression model seems like a sensible model to fit to our data.

We want to determine whether the preference for job satisfaction interacts with with our explanatory variables. The explanatory variables in our case are the expense categories. We need to compare an additive linear model with a model that considers job satisfaction as a variable that interacts with our expense categories. The following joy plot displays the distribution of participant spendings.

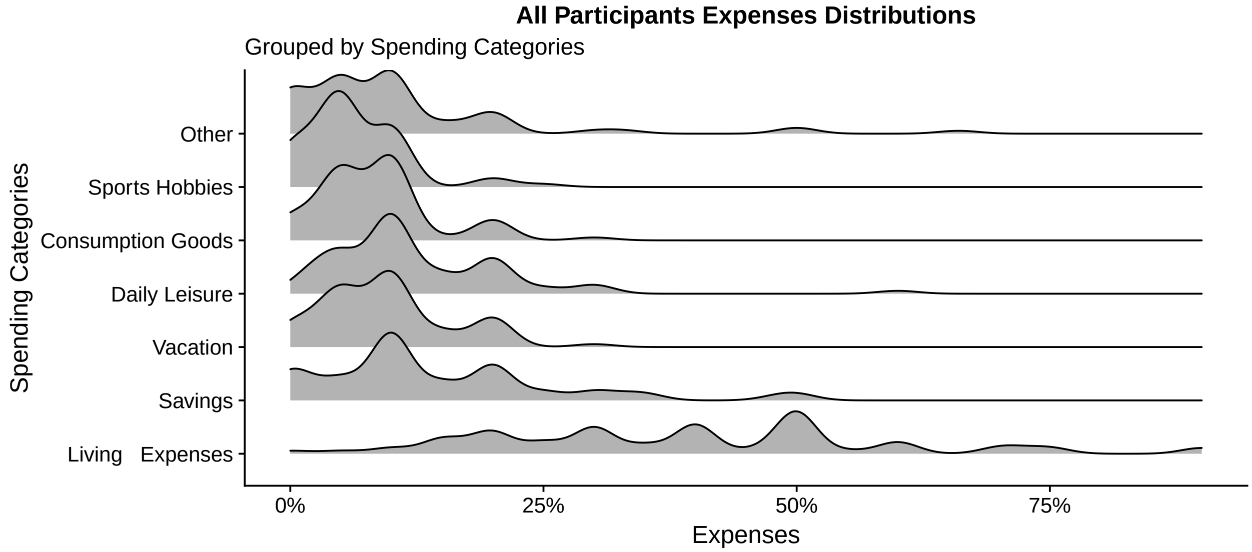
```
# additional wrangling for plotting purposes
survey_results_spendings <- survey_results %>% select(spending_cats)
survey_results_spendings <- map_df(survey_results_spendings, as.numeric)
survey_results_spendings<- melt(survey_results_spendings)
```

```
## No id variables; using all as measure variables
```

```
# joy plot per participant
ggplot(survey_results_spendings, aes(x = value, y = variable))+
  geom_joy(stat = "binline",bins = 20, scale = 0.95, draw_baseline = FALSE)+
    scale_y_discrete(breaks = c("living_expenses","savings","vacation","daily_leisure","consumption_goo
  theme(legend.position = "None") +
    scale_x_continuous(labels = function(x) paste0(x, "%"))+ # Add percent sign
      labs(x = 'Expenses', y ='Spending Categories', title = 'All Participants Expenses Distributions',
```




```
# joy plot per participant
ggplot(survey_results_spendings, aes(x = value, y = variable, height = ..density.. ))+
  geom_joy(stat = "density", bw=2)+
    scale_y_discrete(breaks = c("living_expenses","savings","vacation","daily_leisure","consumption_goo
  theme(legend.position = "None") +
    scale_x_continuous(labels = function(x) paste0(x, "%"))+ # Add percent sign
      labs(x = 'Expenses', y ='Spending Categories', title = 'All Participants Expenses Distributions',
```



Additional Modelling and Exploratory Analysis

```r
# model without interaction
lm_survey <- lm(ratio ~ living_expenses +
                  savings +
                  vacation +
                  daily_leisure +
                  consumption_goods +
                  sports_hobbies +
                  other, data = survey_results)

summary(lm_survey)
```

```
##
## Call:
## lm(formula = ratio ~ living_expenses + savings + vacation + daily_leisure +
##     consumption_goods + sports_hobbies + other, data = survey_results)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32290 -0.05802 -0.00767  0.06133  0.30502
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0961144  0.1122856   0.856    0.395
## living_expenses   -0.0007987  0.0012893  -0.619    0.538
## savings           -0.0008328  0.0016622  -0.501    0.618
## vacation          -0.0017813  0.0024973  -0.713    0.478
## daily_leisure     -0.0003918  0.0020366  -0.192    0.848
## consumption_goods  0.0021812  0.0031719   0.688    0.494
## sports_hobbies     0.0015482  0.0033636   0.460    0.647
## other                     NA         NA      NA       NA
##
## Residual standard error: 0.1209 on 64 degrees of freedom
## Multiple R-squared:  0.05402,    Adjusted R-squared:  -0.03467
## F-statistic: 0.6091 on 6 and 64 DF,  p-value: 0.7221
```

Without any interaction, none of the expenses carry any statistical significance. Below we allow the job satisfaction versus salary increase preference to interact with the expense explanatory variables.

```r
# model with interaction
lm_survey <- lm(ratio ~ no_increase_acceptance:(living_expenses +
                  savings +
                  vacation +
                  daily_leisure +
                  consumption_goods +
                  sports_hobbies +
                  other), data = survey_results)

df_survey_values<- summary(lm_survey) %>% broom::tidy()
knitr::kable(df_survey_values)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1663530	0.1116164	1.4903987	0.1416328
no_increase_acceptanceNo:living_expenses	-0.0008974	0.0014477	-0.6199059	0.5377902
no_increase_acceptanceYes:living_expenses	-0.0015262	0.0013120	-1.1632031	0.2495936
no_increase_acceptanceNo:savings	-0.0047901	0.0040656	-1.1782051	0.2436104
no_increase_acceptanceYes:savings	-0.0004291	0.0016620	-0.2581976	0.7971844
no_increase_acceptanceNo:vacation	-0.0013712	0.0058800	-0.2332058	0.8164370
no_increase_acceptanceYes:vacation	-0.0056610	0.0028051	-2.0181460	0.0482912
no_increase_acceptanceNo:daily_leisure	-0.0061126	0.0028297	-2.1601688	0.0349783
no_increase_acceptanceYes:daily_leisure	0.0015329	0.0026293	0.5830020	0.5621915
no_increase_acceptanceNo:consumption_goods	0.0079150	0.0048841	1.6205807	0.1106273
no_increase_acceptanceYes:consumption_goods	-0.0027200	0.0040829	-0.6661885	0.5079780
no_increase_acceptanceNo:sports_hobbies	0.0147686	0.0105318	1.4022854	0.1662529
no_increase_acceptanceYes:sports_hobbies	-0.0000498	0.0034745	-0.0143312	0.9886158
no_increase_acceptanceNo:other	-0.0082628	0.0046396	-1.7809337	0.0802520

```r
adjusted_p_values_fdr<- p.adjust(df_survey_values$p.value, method = "fdr")
adjusted_p_values_bonf <- p.adjust(df_survey_values$p.value,method = "bonf")

df_survey_values<- cbind(df_survey_values,adjusted_p_values_bonf)
df_survey_values <- cbind(df_survey_values,adjusted_p_values_fdr)

df_survey_values <- df_survey_values %>% select(term,p.value,adjusted_p_values_fdr,adjusted_p_values_bon

knitr::kable(df_survey_values)
```

term	p.value	adjusted_p_values_fdr	adjusted_p_values_bonf
(Intercept)	0.1416328	0.3879234	1.0000000
no_increase_acceptanceNo:living_expenses	0.5377902	0.7155165	1.0000000
no_increase_acceptanceYes:living_expenses	0.2495936	0.4367889	1.0000000
no_increase_acceptanceNo:savings	0.2436104	0.4367889	1.0000000
no_increase_acceptanceYes:savings	0.7971844	0.8792399	1.0000000
no_increase_acceptanceNo:vacation	0.8164370	0.8792399	1.0000000
no_increase_acceptanceYes:vacation	0.0482912	0.3380386	0.6760772
no_increase_acceptanceNo:daily_leisure	0.0349783	0.3380386	0.4896957
no_increase_acceptanceYes:daily_leisure	0.5621915	0.7155165	1.0000000
no_increase_acceptanceNo:consumption_goods	0.1106273	0.3871956	1.0000000
no_increase_acceptanceYes:consumption_goods	0.5079780	0.7155165	1.0000000

term	p.value	adjusted_p_values_fdr	adjusted_p_values_bonf
no_increase_acceptanceNo:sports_hobbies	0.1662529	0.3879234	1.0000000
no_increase_acceptanceYes:sports_hobbies	0.9886158	0.9886158	1.0000000
no_increase_acceptanceNo:other	0.0802520	0.3745096	1.0000000

Power Analysis

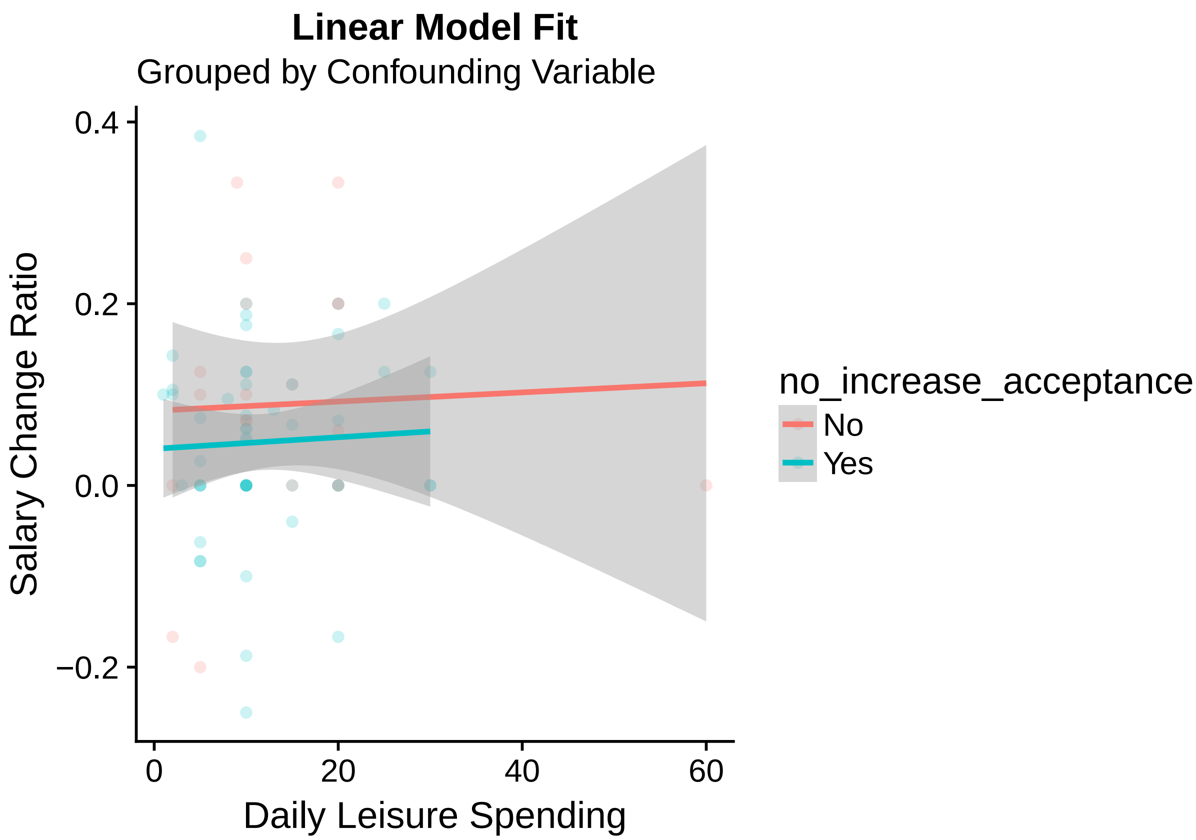
```
# Source: https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

## u is the number of coefficients in our model (minus the intercept).
no_of_coefficients <- length(df_survey_values$term)-1

adj_r_squared <-  summary(lm_survey)$adj.r.squared

## The effect size, f2, is adj_R2/(1???adj_R2), where R2 is the coefficient of determination
effect_size <-  adj_r_squared /(1 - adj_r_squared)

power_analysis<- unlist(pwr.f2.test(u = no_of_coefficients, f2 = effect_size, sig.level = 0.05, power =
```

To estimate the sample size (n) for our general linear model, we use the `pwr.f2.test` from the `pwr` package. The null hypothesis is that "the people who are driven by financial wealth and those that are driven by job satisfaction share the same social standards in terms of spending patterns". This would mean their regression coefficients are statistically indistinguishable from 0. The alternative is that at least one of the coefficients is not 0. The power test has numerator and denominator degrees of freedom. The numerator degrees of freedom, `u` is the number of coefficients (minus the intercept) in our model. The denominator degrees of freedom,`v`, is the number of error degrees of freedom: $v = n - u - 1$; where n is the number of samples. We use adjusted R^2 in $AdjR^2/(1 - adjR^2)$ to calculate the effect size,`f2`. Adjusted R^2 indicates how well our predictors fit a curve or line, but also adjusts for the number of terms in the model. The estimated value of v from the test is ≈ 288. Using the formula from above we get n as $288 + 13 + 1 = 302$. This means that we need a sample size of 302 to ensure that the test of hypothesis will have 80% power to detect a significant impact of the covariates on the outcome.

Above we see that that the job satisfaction confounder variable does contribute towards the correlation between daily leisure, vacation and salary ratio.

Below we visualize daily leisure while accounting for our confounder variable.

```
ggplot(survey_results, aes(y = ratio, x = daily_leisure, group = no_increase_acceptance, colour = no_in
  geom_point(aes(fill = no_increase_acceptance), alpha = 0.2) +
  geom_smooth(method = "lm") +
  labs(x = 'Daily Leisure Spending', y ='Salary Change Ratio', title = 'Linear Model Fit', subtitle= "G
```

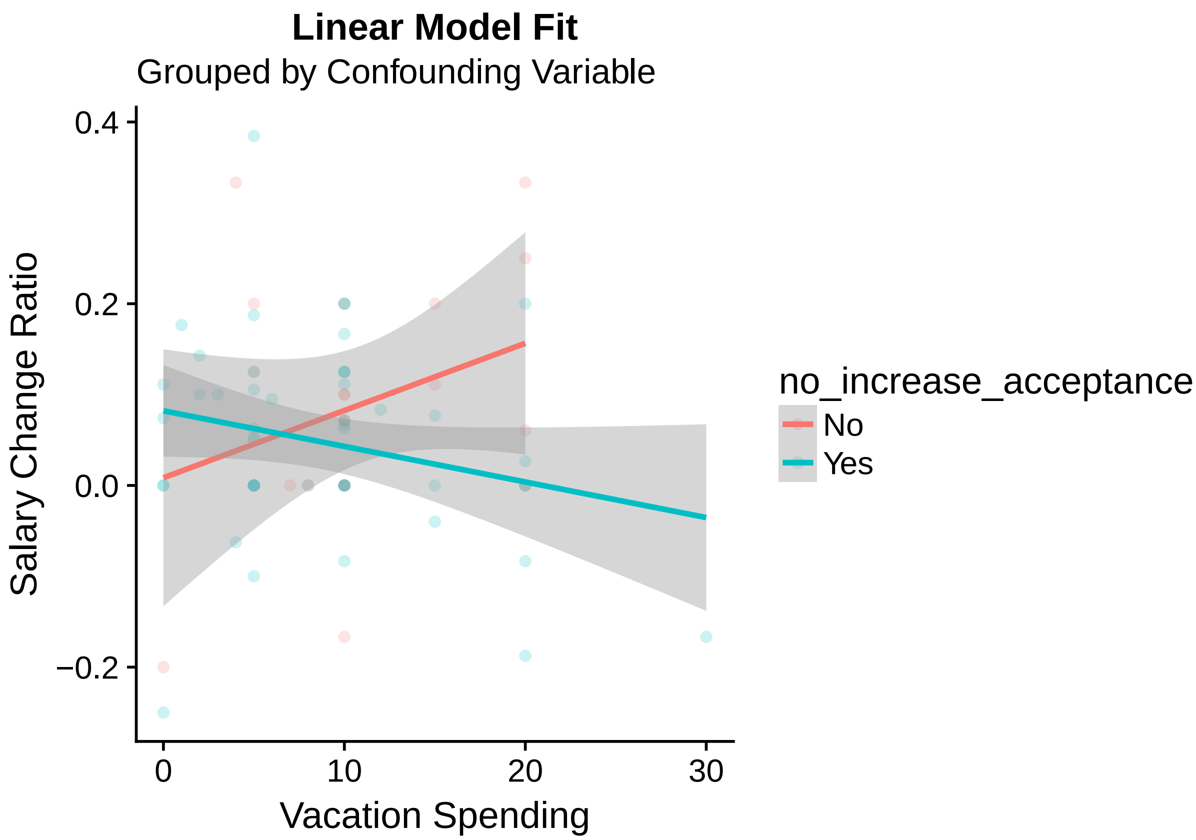


Even though the model found some significance, our visualization seems to disagree to an extent. It might be the daily leisure outlier value that is contributing towards the difference in slopes. The difference in slopes is also quite marginal.

We aren't directly interested in a person's preference between job satisfaction and salary increase, but we do need to take into account how this variable is influencing our study. There are various ways of dealing with confounding variables, but given our our dataset size, our options are limited. For now, including this interaction in our model should be sufficient to maintain awareness of its effect. We should also strongly consider removing higher leverage outliers for the different expense categories which may eliminate the effect of the confounding variable, especially in the case above as linear regression model are highly suceptible to outliers.

Below we visualize vacation while taking our confounding variable into account.

```
ggplot(survey_results, aes(y = ratio, x = vacation, group = no_increase_acceptance, colour = no_increase
  geom_point(aes(fill = no_increase_acceptance), alpha = 0.2) +
  geom_smooth(method = "lm") +
  labs(x = 'Vacation Spending', y ='Salary Change Ratio', title = 'Linear Model Fit', subtitle= "Groupe
```



The difference in slopes is more radical in this case. It would appear that people who spend a larger percentage on vacation have a larger salary ratio **only** if they prefer a salary increase. The confidence intervals are fairly wide, but there might be some truth in the finding. It could contribute towards our hypothesis - people who spend a large percentage on vacation may be the people who are driven by money. In this case, it seems as if our confounding variable interaction could support our hypothesis - people who prefer a salary increase above job satisfaction are those with (possibly) higher social standards (we should be careful to assume that vacation is a direct indication of social standards) and are the same people who expect a higher salary ratio. However, the lack of statistical significance (we aren't yet considering adjusted p-values) and small number of observations mean that we cannot draw any conclusions. However, it is important to differentiate between the people who prefer job satisfaction and those who prefer an increase.