# EDA

*Johannes Harmse*

*April 9, 2018*

```r
# removing confidential data
# temp <- read_csv(file = '../../survey_data/Demographic Survey.csv', skip = 1, header = T)
# temp <- temp[, 10:ncol(temp)]
```

```r
library(tidyverse)
```

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

```
## Warning: Duplicated column names deduplicated: 'Response' =>
## 'Response_1' [2], 'Response' => 'Response_2' [5]
```

```
## Parsed with column specification:
## cols(
##   Response = col_character(),
##   Response_1 = col_character(),
##   `Annual Salary (before deductions)` = col_integer(),
##   `Annual salary (before deductions)` = col_integer(),
##   Response_2 = col_character(),
##   `Living Expenses (utilities, rent, mortgage, transportation, property taxes if owner, etc.)` = col_
##   `Savings (retirement, investments, emergency funds, etc.)` = col_integer(),
##   `Vacation (lodging, transportation, day trips, etc.)` = col_integer(),
##   `Daily Leisure (eating out, books, movies, self-care, etc.)` = col_integer(),
##   `Consumption Goods (clothing, electronics, other luxury items, etc.)` = col_integer(),
##   `Personal Sports and Hobbies (sporting goods and services, gym, arts and crafts, etc.)` = col_inte
##   `Other (health care, taxes, dependent expenses, etc.)` = col_integer()
## )
```

```r
# redefine column names
colnames(survey_results) <- c('consent', 'country', 'salary_base', 'salary_expect', 'no_increase_accepta
                              'living_expenses', 'savings', 'vacation', 'daily_leisure', 'consumption_goods
                              '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/raw/raw_clean.rds')

survey_results %>% head()
```

```
## # A tibble: 6 x 13
##   consent country             salary_base salary_expect no_increase_accep~
##   <chr>   <chr>                     <int>         <int> <chr>
## 1 Yes     United States of A~          NA            NA <NA>
## 2 Yes     Canada                    70000         70000 Yes
## 3 Yes     Canada                    90000        100000 No
## 4 Yes     Canada                    65000         90000 Yes
## 5 Yes     Canada                    80000         75000 Yes
## 6 Yes     Canada                    95000        105000 Yes
## # ... with 8 more variables: living_expenses <int>, savings <int>,
## #   vacation <int>, daily_leisure <int>, consumption_goods <int>,
## #   sports_hobbies <int>, other <int>, id <int>
```

```
# readRDS(file = '../data/raw/raw_clean.rds')
```

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

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

summary(lm_survey)
```

```
##
## Call:
## lm(formula = ratio ~ no_increase_acceptance + living_expenses +
##     savings + vacation + daily_leisure + consumption_goods +
##     sports_hobbies + other, data = survey_results)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2530 -0.4988 -0.2299  0.1048  7.6121
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.368683   2.649100  -0.139    0.890
## no_increase_acceptanceYes  0.250355   0.537186   0.466    0.644
## living_expenses            0.010001   0.029113   0.344    0.733
## savings                    0.003538   0.035576   0.099    0.921
## vacation                   0.073248   0.050322   1.456    0.155
## daily_leisure              0.007683   0.042829   0.179    0.859
## consumption_goods          0.001092   0.054564   0.020    0.984
## sports_hobbies             0.047661   0.059265   0.804    0.427
## other                            NA         NA      NA       NA
##
## Residual standard error: 1.523 on 32 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.1142, Adjusted R-squared:  -0.07954
## F-statistic: 0.5895 on 7 and 32 DF,  p-value: 0.7594
```

```
# remove outliers
survey_results <- survey_results %>%
  filter(ratio < 10 &
           ratio > 0.1)

# replace NA spendings with 0

survey_results[ , spending_cats][is.na(survey_results[ , spending_cats])] <- 0
```

```
# generic first model (outliers removed and data cleaned)
lm_survey <- lm(ratio ~ no_increase_acceptance +
                  living_expenses +
                  savings +
                  vacation +
                  daily_leisure +
                  consumption_goods +
                  sports_hobbies +
                  other, data = survey_results)

summary(lm_survey)
```

```
## 
## Call:
## lm(formula = ratio ~ no_increase_acceptance + living_expenses + 
##     savings + vacation + daily_leisure + consumption_goods + 
##     sports_hobbies + other, data = survey_results)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.44724 -0.15351 -0.01762  0.08516  1.71976 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                1.0019531  0.3906479   2.565   0.0152 *
## no_increase_acceptanceYes -0.1894531  0.1328235  -1.426   0.1635  
## living_expenses           -0.0001961  0.0041568  -0.047   0.9627  
## savings                    0.0102204  0.0057979   1.763   0.0875 .
## vacation                  -0.0060749  0.0102128  -0.595   0.5561  
## daily_leisure              0.0038172  0.0087573   0.436   0.6658  
## consumption_goods         -0.0010248  0.0110443  -0.093   0.9267  
## sports_hobbies             0.0157987  0.0119227   1.325   0.1945  
## other                      0.0053635  0.0073922   0.726   0.4734  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3674 on 32 degrees of freedom
## Multiple R-squared:  0.2449, Adjusted R-squared:  0.05617 
## F-statistic: 1.298 on 8 and 32 DF,  p-value: 0.2798
```

```
survey_tidy <- NULL

non_spendings <- colnames(survey_results)[!(colnames(survey_results) %in% spending_cats)]
```

```
for (spending in spending_cats){
  temp <- survey_results[ , non_spendings]
  temp$spending_cat <- spending
  temp$spending_val <- survey_results[[spending]]
  survey_tidy <- rbind(survey_tidy, temp)
}
```

```
for (i in unique(survey_tidy$id)){
  temp <- survey_tidy %>% filter(id == i)
  user_living <- as.numeric(temp %>% filter(temp$spending_cat == 'living_expenses') %>% select(spending_
  survey_tidy[survey_tidy$id == i, 'spending_ratio'] <- temp$spending_val/user_living
}
```

```
p_vals <- data.frame('category' = character(length(spending_cats)), 'slope' = numeric(length(spending_ca

count <-  0

for (i in spending_cats){
  count <- count + 1
  temp <- survey_tidy %>% filter(spending_cat == i)
  temp_lm <- lm(ratio ~ spending_ratio, data = temp)
  lm_summary <- summary(temp_lm)
  p_vals[count, 'category'] <- as.character(i)
  p_vals[count, 'slope'] <- temp_lm$coefficients[2]
  p_vals[count, 'p_value'] <- ifelse(nrow(lm_summary$coefficients) > 1, lm_summary$coefficients[2 , 4],
}
```

```
p_vals
```

```
##            category     slope      p_value
## 1   living_expenses        NA           NA
## 2           savings 0.3289472 1.311452e-06
## 3          vacation 0.2041850 1.816509e-01
## 4     daily_leisure 0.1359407 2.794008e-01
## 5 consumption_goods 0.1669850 1.044669e-01
## 6    sports_hobbies 0.8587598 8.193711e-05
## 7             other 0.3754592 1.979317e-03
```