EDA

Johannes Harmse

April 9, 2018

```
library(tidyverse)
```

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

```
## Warning: Missing column names filled in: 'X1' [1], 'X2' [2], 'X3' [3],
## 'X4' [4], 'X5' [5], 'X6' [6], 'X7' [7], 'X8' [8], 'X9' [9]
```

```
## Warning: Duplicated column names deduplicated: 'Response' =>
## 'Response_1' [11], 'Response' => 'Response_2' [14]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   X2 = col_integer(),
##   `Annual Salary (before deductions)` = col_integer(),
##   `Annual salary (before deductions)` = col_integer(),
##   `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()
## )
```

```
## See spec(...) for full column specifications.
```

```
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_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)
```

```r
# 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~      100000        110000 Yes
## 2 Yes     South Africa             160000        120000 Yes
## 3 Yes     South Africa                NA            NA <NA>
## 4 Yes     Canada                   30000         30000 Yes
## 5 Yes     South Africa            3000000       3300000 Yes
## 6 Yes     United States of A~       75000         77000 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>
```

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

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

```r
# 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
## -2.5572 -0.7988 -0.1067  0.3220  8.0703
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.875809   1.761593  -0.497    0.621
## no_increase_acceptanceYes   0.284242   0.490828   0.579    0.565
## living_expenses             0.007298   0.019068   0.383    0.703
## savings                     0.028358   0.024205   1.172    0.246
## vacation                    0.147477   0.031670   4.657 1.83e-05 ***
## daily_leisure              -0.002480   0.030399  -0.082    0.935
## consumption_goods          -0.005303   0.047016  -0.113    0.911
```

```
## sports_hobbies             0.023158   0.049494   0.468    0.642
## other                           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 60 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.3138, Adjusted R-squared:  0.2337
## F-statistic: 3.919 on 7 and 60 DF,  p-value: 0.001399
```

```
# 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
## -0.68667 -0.14044 -0.02061  0.06512  1.86415
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.1309956  0.1426541   7.928 2.27e-11 ***
## no_increase_acceptanceYes -0.1560024  0.0910668  -1.713   0.0911 .
## living_expenses           -0.0010387  0.0018352  -0.566   0.5732
## savings                    0.0071305  0.0034563   2.063   0.0428 *
## vacation                  -0.0024537  0.0068022  -0.361   0.7194
## daily_leisure             -0.0005872  0.0052386  -0.112   0.9111
## consumption_goods         -0.0012078  0.0083464  -0.145   0.8853
## sports_hobbies             0.0072588  0.0086850   0.836   0.4061
## other                      0.0019409  0.0035539   0.546   0.5867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3542 on 71 degrees of freedom
## Multiple R-squared:  0.1117, Adjusted R-squared:  0.0116
## F-statistic: 1.116 on 8 and 71 DF,  p-value: 0.3632
```

```
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 <- temp %>% filter(!is.na(spending_ratio) & abs(spending_ratio) != Inf)
  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 0.23516271 8.284805e-05
## 3          vacation 0.14432323 2.600583e-01
## 4     daily_leisure 0.05701884 4.969810e-01
## 5 consumption_goods 0.11015343 2.013140e-01
## 6    sports_hobbies 0.25028137 5.877919e-02
## 7             other 0.01648614 6.326853e-01
```