EDA

Johannes Harmse

April 9, 2018

```
library(tidyverse)
```

```
#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<U+00A0>(utilities, rent, mortgage, transportation, property taxes if owner, etc.)
##   `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()
## )
```

```
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_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 id
survey_results$id <- 1:nrow(survey_results)

# 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_acceptance
##   <chr>   <chr>         <int>         <int> <chr>
## 1 Yes     Canada        70000         80000 Yes
## 2 Yes     Canada        90000        100000 Yes
## 3 Yes     Canada        80000         80000 Yes
## 4 Yes     Canada       100000        120000 No
## 5 Yes     Canada      1000000         90000 No
## 6 Yes     Canada        70000         70000 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')
```

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

```
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
## -1.6783 -0.6026 -0.2512  0.2574  7.2035
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -3.55114    4.80731  -0.739    0.467
## no_increase_acceptanceYes  0.02486    0.66484   0.037    0.970
## living_expenses            0.04615    0.05322   0.867    0.394
## savings                    0.03878    0.05830   0.665    0.512
## vacation                   0.11253    0.06905   1.630    0.115
## daily_leisure              0.04483    0.06600   0.679    0.503
## consumption_goods          0.03298    0.07677   0.430    0.671
## sports_hobbies             0.09997    0.08412   1.188    0.245
## other                           NA         NA      NA       NA
##
## Residual standard error: 1.647 on 26 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1533, Adjusted R-squared:  -0.07461
## F-statistic: 0.6727 on 7 and 26 DF,  p-value: 0.6932
```

```
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
}
```

```
for (spending in spending_cats){
  temp <- survey_tidy %>% filter(spending_cat == spending)
  temp_lm <- lm(ratio ~ spending_ratio, data = temp)
  print(spending)
  print(summary(temp_lm))
}
```

```
## [1] "living_expenses"
## 
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.2591 -0.3491 -0.2866 -0.2241  8.6509 
## 
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.349      0.265   5.091 1.31e-05 ***
## spending_ratio       NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.568 on 34 degrees of freedom
## 
## [1] "savings"
## 
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.1712 -0.3890 -0.2470 -0.1735  8.7388 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      1.2612     0.3239   3.894 0.000454 ***
## spending_ratio   0.1608     0.3328   0.483 0.632177    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.586 on 33 degrees of freedom
## Multiple R-squared:  0.007024,   Adjusted R-squared:  -0.02307 
## F-statistic: 0.2334 on 1 and 33 DF,  p-value: 0.6322
## 
```

```
## [1] "vacation"
##
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0927 -0.3952 -0.2411 -0.1436  8.5922
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.1827     0.3566   3.316  0.00223 **
## spending_ratio   0.4501     0.6396   0.704  0.48655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.579 on 33 degrees of freedom
## Multiple R-squared:  0.01478,    Adjusted R-squared:  -0.01507
## F-statistic: 0.4952 on 1 and 33 DF,  p-value: 0.4865
##
## [1] "daily_leisure"
##
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1469 -0.3972 -0.2693 -0.2054  8.6466
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.2311     0.3775   3.261  0.00258 **
## spending_ratio   0.2446     0.5507   0.444  0.65985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.586 on 33 degrees of freedom
## Multiple R-squared:  0.005942,   Adjusted R-squared:  -0.02418
## F-statistic: 0.1973 on 1 and 33 DF,  p-value: 0.6598
##
## [1] "consumption_goods"
##
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2523 -0.3535 -0.2893 -0.2331  8.6480
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.33365    0.33314   4.003 0.000347 ***
## spending_ratio  0.07344    0.44529   0.165 0.870042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.613 on 32 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0008493,  Adjusted R-squared:  -0.03037
## F-statistic: 0.0272 on 1 and 32 DF,  p-value: 0.87
##
## [1] "sports_hobbies"
##
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0394 -0.3728 -0.2630 -0.1282  8.6170
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.1294     0.3592   3.145  0.00358 **
## spending_ratio   1.0142     1.0107   1.004  0.32313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.588 on 32 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.03051,    Adjusted R-squared:  0.0002152
## F-statistic: 1.007 on 1 and 32 DF,  p-value: 0.3231
##
## [1] "other"
##
## Call:
## lm(formula = ratio ~ spending_ratio, data = temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2581 -0.3597 -0.3040 -0.2264  8.6540
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.34604    0.39115   3.441  0.00163 **
## spending_ratio  0.05758    0.87261   0.066  0.94780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.613 on 32 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0001361,  Adjusted R-squared:  -0.03111
## F-statistic: 0.004355 on 1 and 32 DF,  p-value: 0.9478
```