Exploratory Data Analysis

S. Arora, J. Harmse, V. Mulholland

April 9, 2018

```
library(tidyverse)
```

Anonymity

In order to maintain user privacy

```
# 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 ic

# 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/processed/surveydata_clean.rds')

# remove all traces
rm(survey_results)
```

```
# import clean data
survey_results <-  readRDS(file = '../data/processed/surveydata_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        145000 Yes
## 2 Yes     Canada                   140000        150000 No
## 3 Yes     Canada                    60000         65000 Yes
## 4 Yes     South Africa             250000        400000 No
```

```
## 5 Yes     South Africa             550000        550000 Yes
## 6 Yes     Canada                    50000         90000 No
## # ... with 8 more variables: living_expenses <int>, savings <int>,
## #   vacation <int>, daily_leisure <int>, consumption_goods <int>,
## #   sports_hobbies <int>, other <int>, id <int>
```

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

```
# data summary table
summary(survey_results)
```

```
##    consent            country           salary_base
##  Length:83          Length:83          Min.   :    3000
##  Class :character   Class :character   1st Qu.:   70000
##  Mode  :character   Mode  :character   Median :   80000
##                                        Mean   : 1033071
##                                        3rd Qu.:  120000
##                                        Max.   :65000000
##
##  salary_expect      no_increase_acceptance living_expenses    savings
##  Min.   :    3000   Length:83              Min.   : 0.00   Min.   : 0.00
##  1st Qu.:   75000   Class :character       1st Qu.:25.00   1st Qu.: 5.00
##  Median :   90000   Mode  :character       Median :40.00   Median :10.00
##  Mean   : 1149657                          Mean   :39.49   Mean   :14.43
##  3rd Qu.:  120000                          3rd Qu.:50.00   3rd Qu.:20.00
##  Max.   :70000000                          Max.   :90.00   Max.   :50.00
##                                                            NA's   :1
##     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.: 5.000
##  Median :10.000   Median :10.00   Median :10.000    Median : 5.000
##  Mean   : 9.928   Mean   :12.46   Mean   : 8.568    Mean   : 6.372
##  3rd Qu.:11.000   3rd Qu.:17.50   3rd Qu.:10.000    3rd Qu.:10.000
##  Max.   :50.000   Max.   :60.00   Max.   :30.000    Max.   :25.000
##                                   NA's   :2         NA's   :5
##      other             id
##  Min.   : 0.00   Min.   : 1.0
##  1st Qu.: 5.00   1st Qu.:21.5
##  Median :10.00   Median :42.0
##  Mean   :10.39   Mean   :42.0
##  3rd Qu.:10.75   3rd Qu.:62.5
##  Max.   :66.00   Max.   :83.0
##  NA's   :7
```

———basic plot showing outliers

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

———plot ratio

The questions were designed to make use of a monetary ratio to avoid additional manipulation of currency data.

vvvvvvvvvvvvvvvvv ——not required

```
# 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
## -2.4069 -0.7512 -0.1195  0.3144  8.1632
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.8670565  1.6641282  -0.521    0.604
## no_increase_acceptanceYes   0.2663450  0.4425869   0.602    0.549
## living_expenses             0.0075329  0.0181016   0.416    0.679
## savings                     0.0284803  0.0228386   1.247    0.217
## vacation                    0.1413033  0.0295466   4.782 9.88e-06 ***
## daily_leisure               0.0045063  0.0279493   0.161    0.872
## consumption_goods          -0.0001305  0.0433870  -0.003    0.998
## sports_hobbies              0.0143647  0.0454546   0.316    0.753
## other                             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.695 on 67 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.3017, Adjusted R-squared:  0.2287
## F-statistic: 4.134 on 7 and 67 DF,  p-value: 0.0007849
```

Outliers

Considerations were made to account for outliers. Justification for removing outliers beyond two standard deviations from the mean - typos -

————————- boxplot function to remove outliers ————————- boxplot plot

```
# ---------------------------------fix for 2SD
# 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
```

Model

A first take at modelling the data .....

```
# 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
## -0.71269 -0.15281 -0.04237  0.06155  1.88255
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.3474319  0.3548949   3.797 0.000306 ***
## no_increase_acceptanceYes -0.1578734  0.0937777  -1.683 0.096673 .
## living_expenses           -0.0030799  0.0038333  -0.803 0.424392
## savings                    0.0043620  0.0048473   0.900 0.371220
## vacation                  -0.0066359  0.0073956  -0.897 0.372609
## daily_leisure             -0.0002022  0.0059387  -0.034 0.972931
## consumption_goods         -0.0013920  0.0093466  -0.149 0.882029
## sports_hobbies             0.0028296  0.0098079   0.289 0.773801
## other                             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3657 on 71 degrees of freedom
## Multiple R-squared:  0.09716,    Adjusted R-squared:  0.008143
## F-statistic: 1.091 on 7 and 71 DF,  p-value: 0.378
```

```
# gathered data
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)
}
```

Standardizing the spendings according to their living expenses.

```
# spending as a ratio of living expenses
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
}
```

```
# store variable p-values
p_vals <- data.frame('category' = character(length(spending_cats)), 'slope' = numeric(length(spending_ca

# run linear models for each spending category individually
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 0.22085290 0.0002101268
## 3          vacation 0.12517657 0.3191775703
## 4     daily_leisure 0.07049266 0.3917040658
## 5 consumption_goods 0.11580023 0.1726792238
## 6    sports_hobbies 0.23850548 0.0676555677
## 7             other 0.01411300 0.6800794939
```