Exploratory Data Analysis

S. Arora, J. Harmse, V. Mulholland

April 9, 2018

```
library(tidyverse); theme_set(theme_bw())
```

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_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)
```

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')  # local path - remove iden

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

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

A new variable was created as a measurement of relative expected increase in salary. This variable corresponds to the response variable of our linear regression model. The benefits of using a ratio meant that there would be less extra manipulations and potential confounding variables behind adjustments for foreign currencies. This ratio of the expected salary over the base salary of a respondent is indicative of a person's desire for financial gains. This type of metric introduces a set of confounding aspects when we do not account for specific demographic groups that would naturally behave a certain way with regards to financial inclinations. For instance, people of a certain age will typically not expect major salary increases when nearing retirement while students working while pursing a higher education degree may be recipients of a very small income in comparison to a year from now when they will enter the job market. Had the survey been organized to reach a larger demographic, more questions could have been added to delineate the confounding variables into groups including demographics such as age, occupation, etc. Another solution to handle this would have been to adjust the wording of the questions in a way that was more explicit, say instead of "" we could have specified that if they are part-time employed or a student, to imagine they are on the full-time job market when considering their qualifications and salary estimates. **IS THIS WHERE PAUL TALKS ABOUT BLOCKS**

```
# 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 = survey_results$salary_expect/survey_results$salary_
```

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 quantile analysis allowed a reasonable upper and lower limit to be chosen.

EXPLAIN REASONING HERE

———basic plot showing outliers

```
# ---------------------------------------remove 2 standard deviations
# remove outliers
survey_results <- survey_results %>%
  filter(ratio < 10 &
           ratio > 0.1)
```

Each variable is summarized. Since it is difficult to highlight important information from a summary table containing so many variables, a _____plot was generated.

```
# data summary table
sum.tb <- summary(survey_results)
sum.tb
```

```
##    consent            country           salary_base
##  Length:79          Length:79          Min.   :    3000
##  Class :character   Class :character   1st Qu.:   70000
##  Mode  :character   Mode  :character   Median :   80000
##                                        Mean   : 1072411
##                                        3rd Qu.:  120000
##                                        Max.   :65000000
##  salary_expect      no_increase_acceptance living_expenses    savings
##  Min.   :    3000   Length:79              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   : 1203532                          Mean   :39.53   Mean   :14.34
##  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.: 3.000
##  Median :10.000   Median :10.00   Median :10.000    Median : 5.000
##  Mean   : 9.291   Mean   :12.49   Mean   : 8.342    Mean   : 6.038
##  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
##  Min.   : 0.000   Min.   : 1.00   Min.   :0.250
##  1st Qu.: 5.000   1st Qu.:22.50   1st Qu.:1.000
##  Median : 9.000   Median :42.00   Median :1.071
##  Mean   : 9.962   Mean   :42.27   Mean   :1.120
##  3rd Qu.:10.000   3rd Qu.:63.50   3rd Qu.:1.134
##  Max.   :66.000   Max.   :83.00   Max.   :3.333
```

```
# labels(survey_results) <-
# library(knitr)
# library(papeR)
# xtable(summarize(survey_results, type = "numeric"))
# xtable(summarize(Orthodont, type = "factor", variables = "Sex"))
# xtable(summarize(Orthodont, type = "numeric", group = "Sex"))
#
# library("knitr")
# summarize(survey_results, type = "factor")
# kable(summarize(Orthodont, type = "numeric"))
# kable(summarize(Orthodont, type = "factor", variables = "Sex", cumulative = TRUE))
# kable(summarize(Orthodont, type = "numeric", group = "Sex", test = FALSE))
```

Response Variable

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:

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

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.

————plot ratio

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
## -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
```

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
```