Exploratory Data Analysis

S. Arora, J. Harmse, V. Mulholland

April 9, 2018

```
library(tidyverse)
```

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 id

# 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

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>
```

```
# 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')
```