Reddit Data Analysis

Pramod Duvvuri

4/2/2019

```
# Reading Data
reddit_data <- read.csv('./reddit.csv')
head(reddit_data)
```

```
##   id gender age.range                           marital.status
## 1  1      0     25-34                                     <NA>
## 2  2      0     25-34                                     <NA>
## 3  3      1     18-24                                     <NA>
## 4  4      0     25-34                                     <NA>
## 5  5      1     25-34                                     <NA>
## 6  6      0     25-34 Married/civil union/domestic partnership
##    employment.status military.service children         education
## 1 Employed full time             <NA>       No Bachelor's degree
## 2 Employed full time             <NA>       No Bachelor's degree
## 3          Freelance             <NA>       No      Some college
## 4          Freelance             <NA>       No Bachelor's degree
## 5 Employed full time             <NA>       No Bachelor's degree
## 6 Employed full time               No       No Bachelor's degree
##         country      state      income.range    fav.reddit      dog.cat
## 1 United States   New York  $150,000 or more  getmotivated         <NA>
## 2 United States   New York  $150,000 or more        gaming         <NA>
## 3 United States   Virginia     Under $20,000 snackexchange         <NA>
## 4 United States   New York  $150,000 or more    spacedicks         <NA>
## 5 United States California $70,000 - $99,999           aww         <NA>
## 6 United States   New York  $150,000 or more        gaming I like dogs.
##    cheese
## 1    <NA>
## 2    <NA>
## 3    <NA>
## 4    <NA>
## 5    <NA>
## 6 Cheddar
```

```
table(reddit_data$employment.status)
```

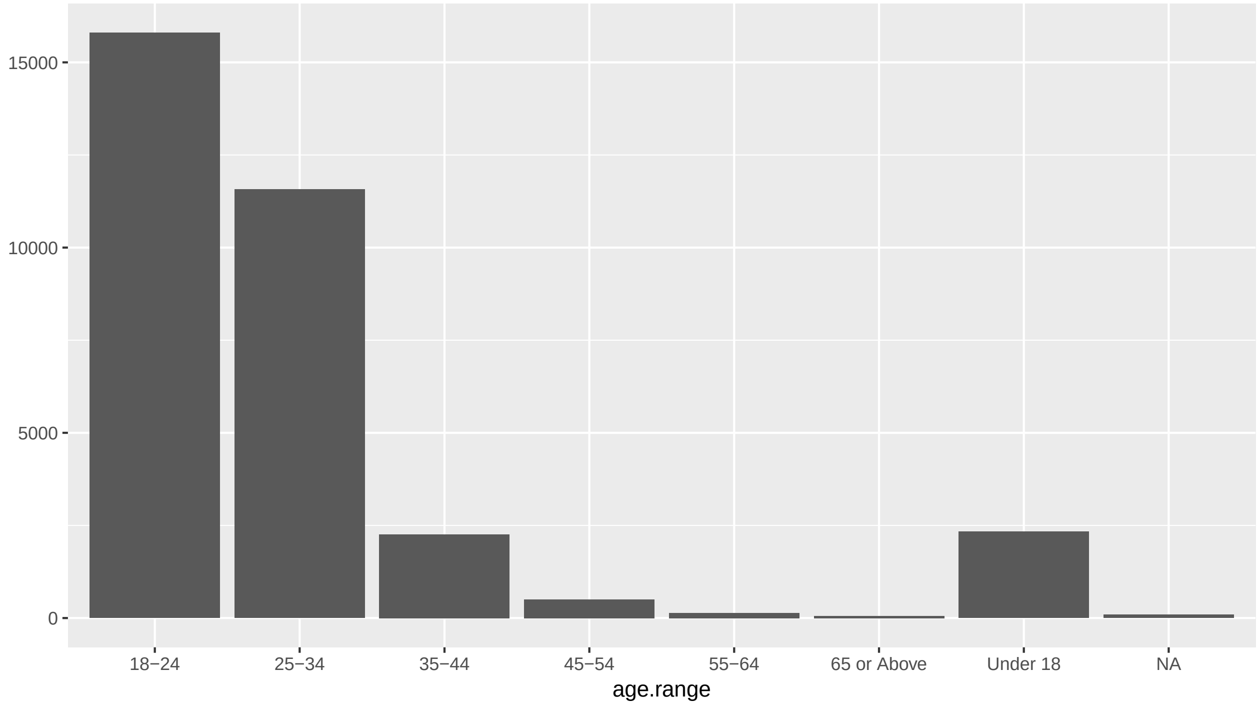
```
## 
##                     Employed full time 
##                                  14814 
##                              Freelance 
##                                   1948 
## Not employed and not looking for work 
##                                    682 
##    Not employed, but looking for work 
##                                   2087 
##                                Retired 
##                                     85 
##                                Student 
##                                  12987 
```

```
#summary(reddit_data)
```

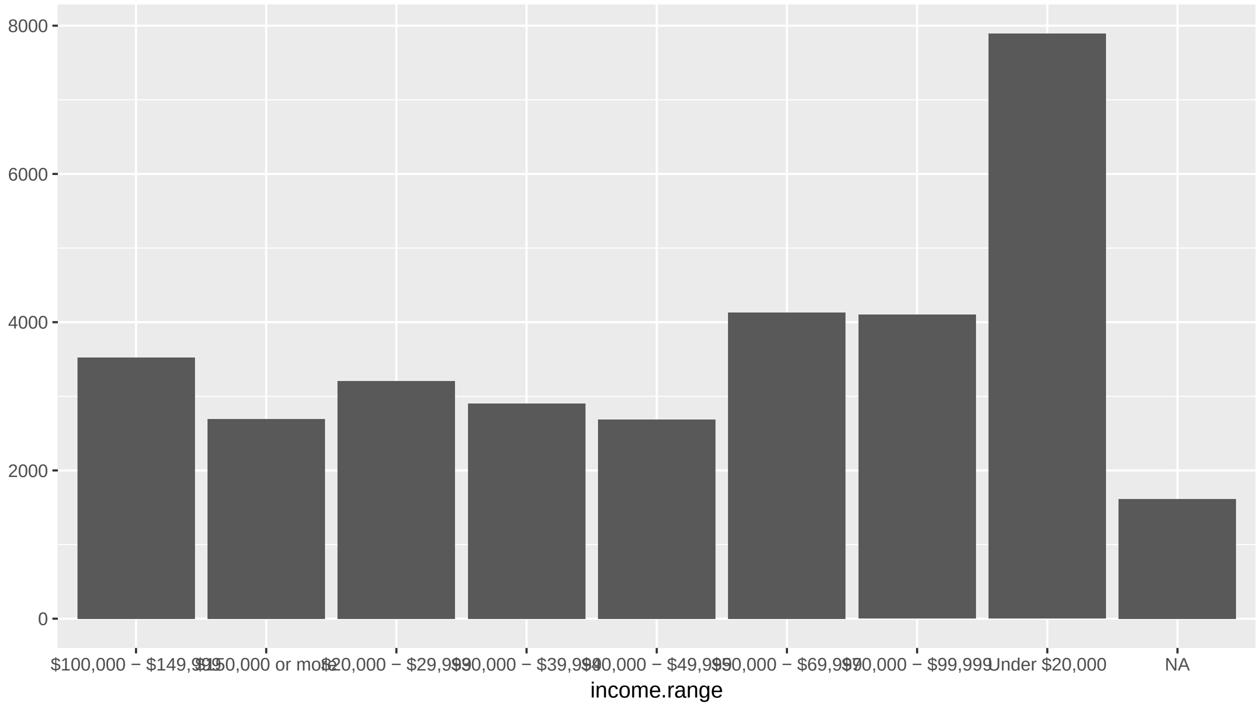
```
levels(reddit_data$age.range)
```

```
## [1] "18-24"       "25-34"       "35-44"       "45-54"       "55-64"
## [6] "65 or Above" "Under 18"
```

```
# Basic Plot of unordered factors
library(ggplot2)
qplot(data = reddit_data, x = age.range)
```

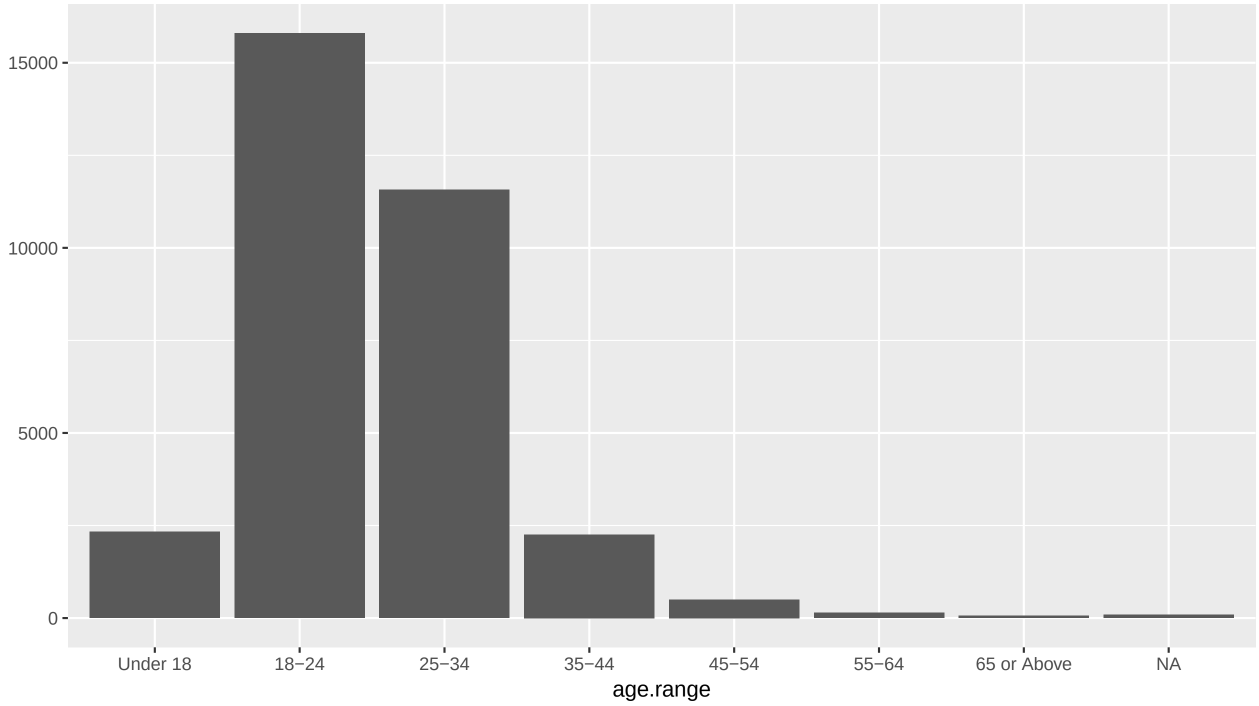


```
qplot(data = reddit_data, x = income.range)
```



```r
# Creating ordered variables for better plots
reddit_data$age.range <- ordered(reddit_data$age.range,
                                 levels = c("Under 18","18-24" ,
                                            "25-34","35-44","45-54",
                                            "55-64","65 or Above" ))
```

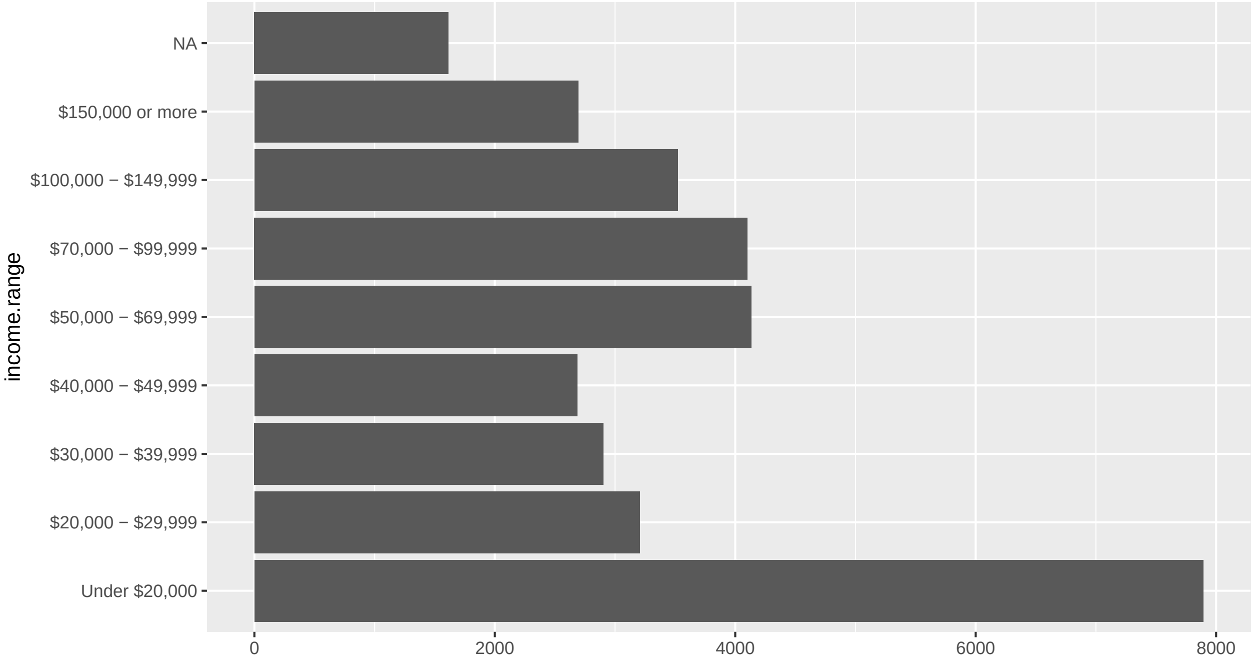
```r
qplot(data = reddit_data, x = age.range)
```



```
levels(reddit_data$income.range)
```

```
## [1] "$100,000 - $149,999" "$150,000 or more"    "$20,000 - $29,999"
## [4] "$30,000 - $39,999"   "$40,000 - $49,999"   "$50,000 - $69,999"
## [7] "$70,000 - $99,999"   "Under $20,000"
```

```
# Creating ordered variables for better plots
reddit_data$income.range <- ordered(reddit_data$income.range,
                             levels = c("Under $20,000","$20,000 - $29,999",
                                        "$30,000 - $39,999","$40,000 - $49,999",
                                        "$50,000 - $69,999","$70,000 - $99,999",
                                        "$100,000 - $149,999","$150,000 or more"))
qplot(data = reddit_data, x = income.range) + coord_flip()
```


NA
$150,000 or more
$100,000 − $149,999
$70,000 − $99,999
$50,000 − $69,999
$40,000 − $49,999
$30,000 − $39,999
$20,000 − $29,999
Under $20,000
income.range