FE8828 Assignment for Exploratory Data Analysis

Toh Zhao Zhi *<Email: tohz0016pg@e.ntu.edu.sg (mailto:tohz0016pg@e.ntu.edu.sg)>*

Nov 19, 2017

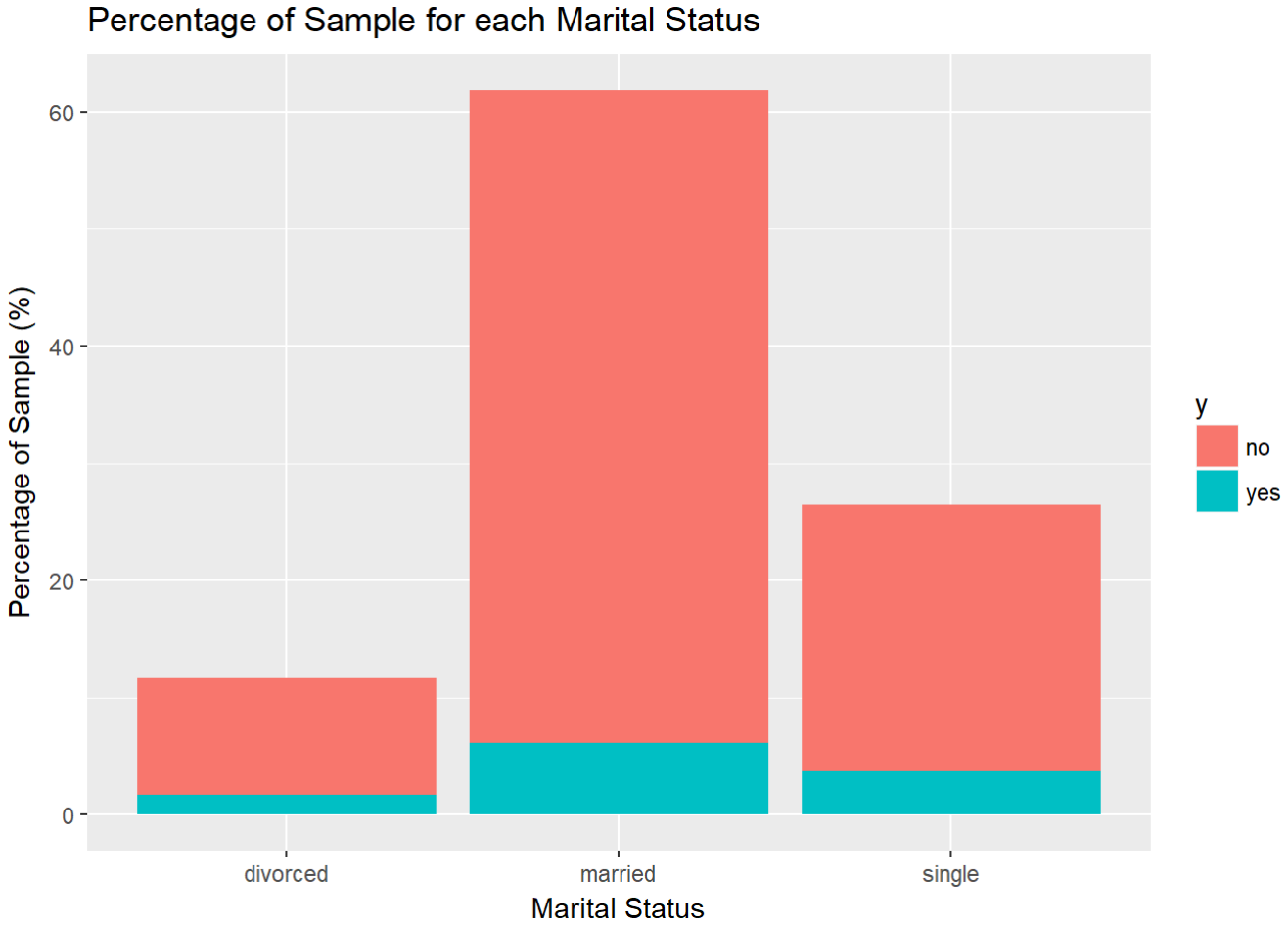
Finding #1

This data contains 4521 number of data points.

Finding #2 Marital Status Demographics of Sample

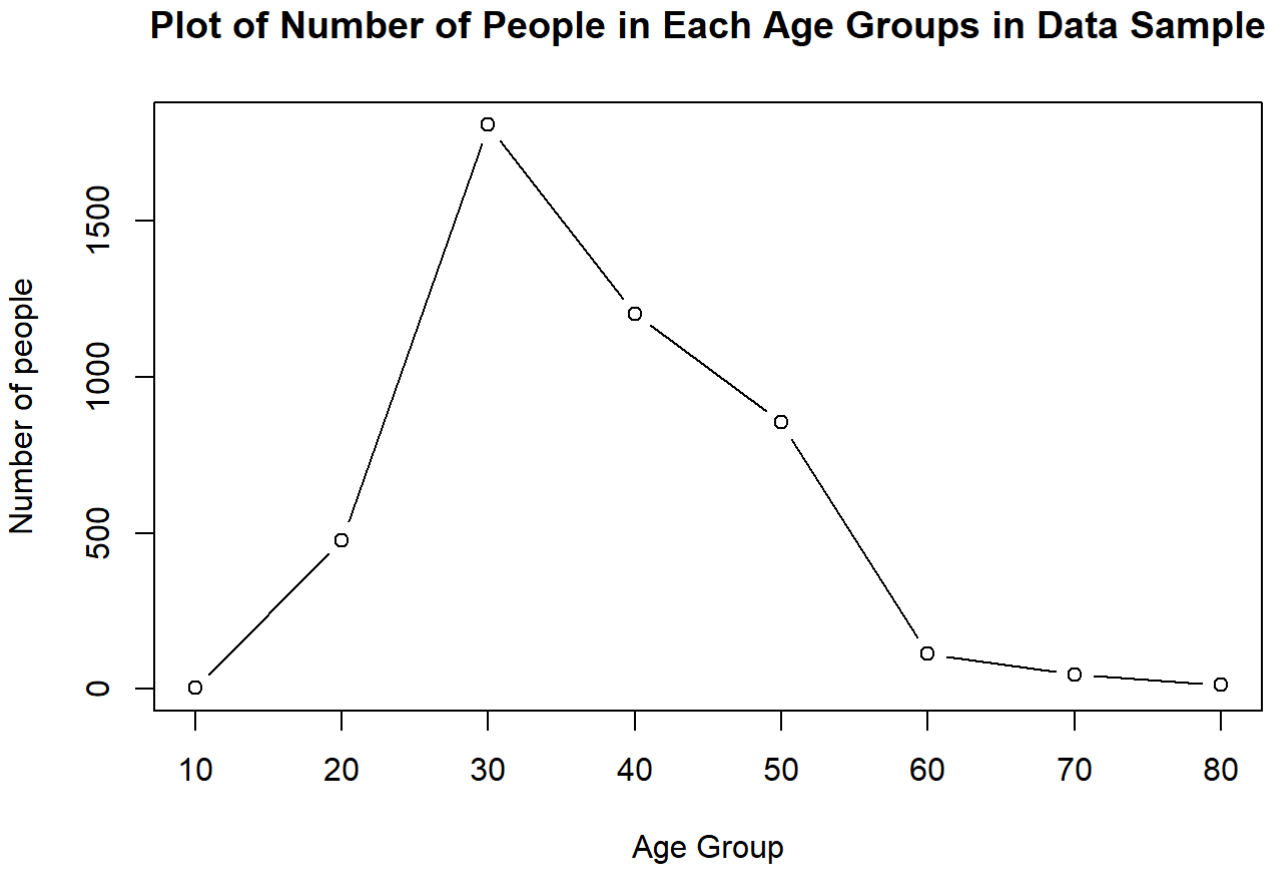
Proportion of data sample for each marital status:

```
## # A tibble: 3 x 2
##    marital     n
##     <fctr> <int>
## 1 divorced   528
## 2  married  2797
## 3   single  1196
```



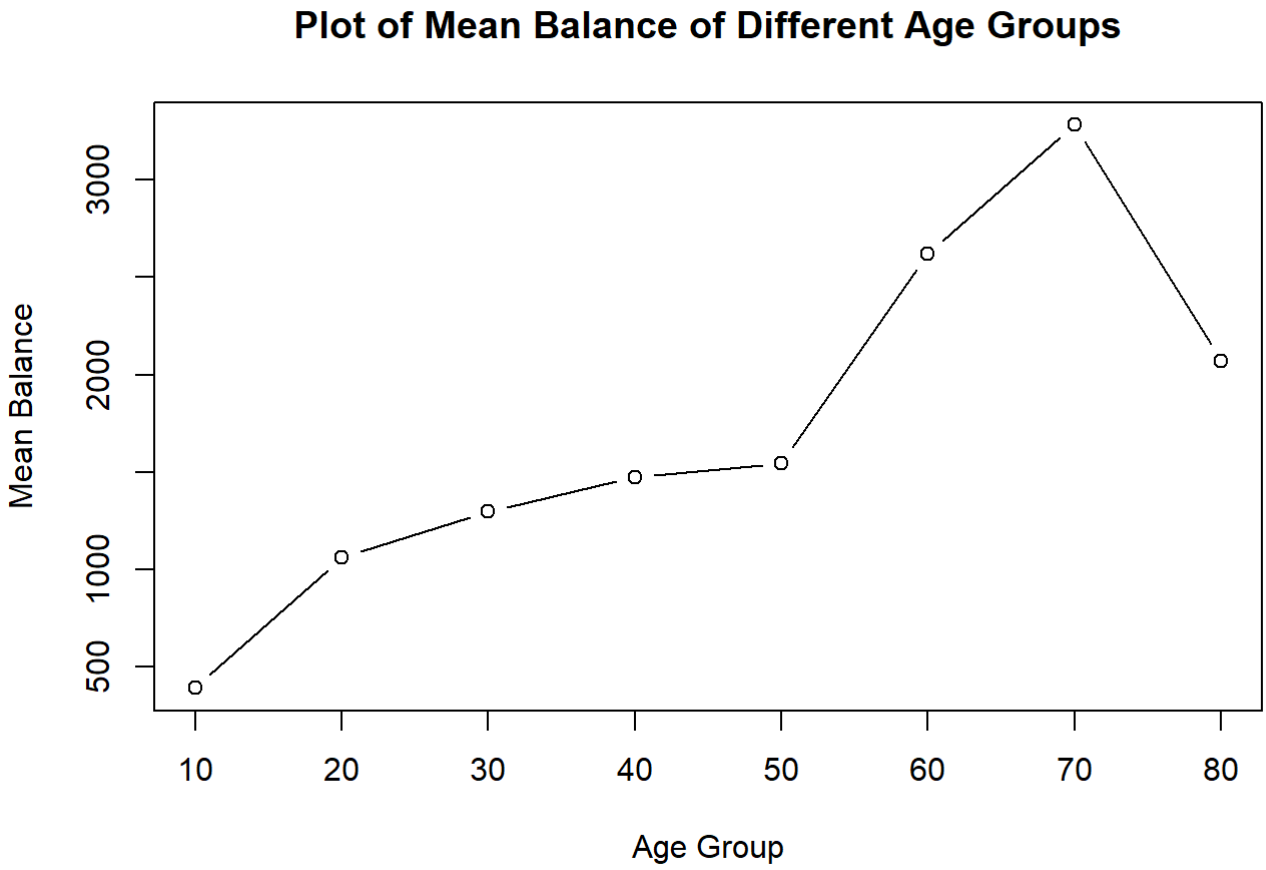
528 of the people in the sample are divorced, 2797 are married and 1196 are single of which majority of the people in each category have not subscribed to the term deposit.

Finding #3 Age Demographics of Sample



In the above analysis, people in the sample were categorized into age groups of 10+, 20+, 30+,…, 80+. It can be seen that the majority of the data sample is made up of people in their 30s and the second largest group would be the people in their 40s.

Finding #4 Mean Balance across Age Groups

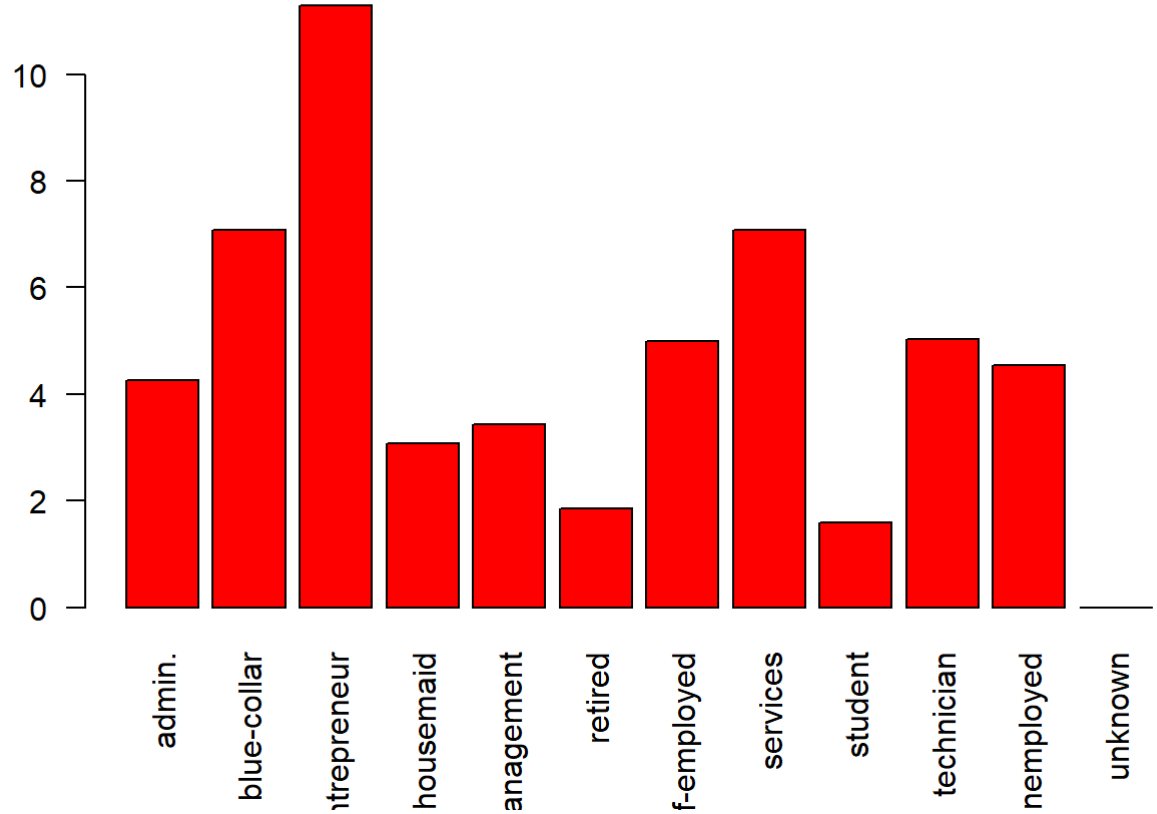



```
## # A tibble: 8 x 3
##   age_group count balance_mean
##       <dbl> <int>        <dbl>
## 1        10     4      393.500
## 2        20   478     1063.657
## 3        30  1808     1298.147
## 4        40  1203     1474.692
## 5        50   854     1547.420
## 6        60   113     2619.779
## 7        70    47     3280.872
## 8        80    14     2071.143
```

The above plot shows the mean balance across different age groups and it can be observed that on average, people above 60 years old have higher balances as compared to people below 60 years old. People in their 70s have highest mean balance, while people aged 10+ have the lowest mean balance.

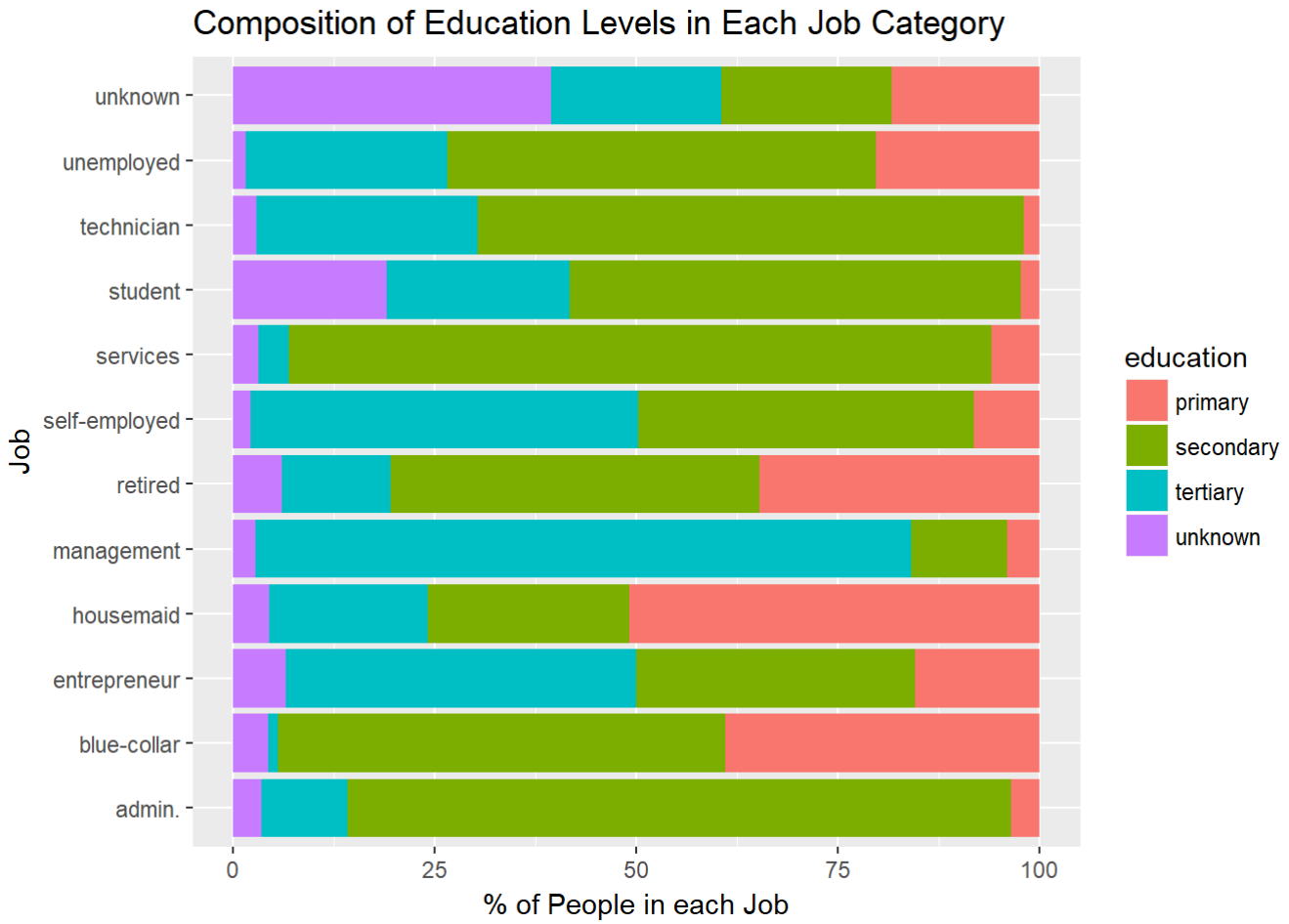
Finding #5 Likelihood of default in each job category

```
## # A tibble: 12 x 4
##              job default_count total likelihood.default
##           <fctr>         <dbl> <dbl>              <dbl>
##  1        admin.             6   141           4.255319
##  2   blue-collar            14   198           7.070707
##  3  entrepreneur             7    62          11.290323
##  4     housemaid             2    65           3.076923
##  5    management            14   407           3.439803
##  6       retired             3   161           1.863354
##  7 self-employed             4    80           5.000000
##  8      services             7    99           7.070707
##  9       student             1    63           1.587302
## 10    technician            15   298           5.033557
## 11    unemployed             3    66           4.545455
## 12       unknown             0    37           0.000000
```



The above bar chart shows the default likelihood across different job categories. It can be seen that entrepreneurs have the highest possibility of defaulting on their loans, while unknown, students and retired have the lowest possibility of defaulting on their loans.

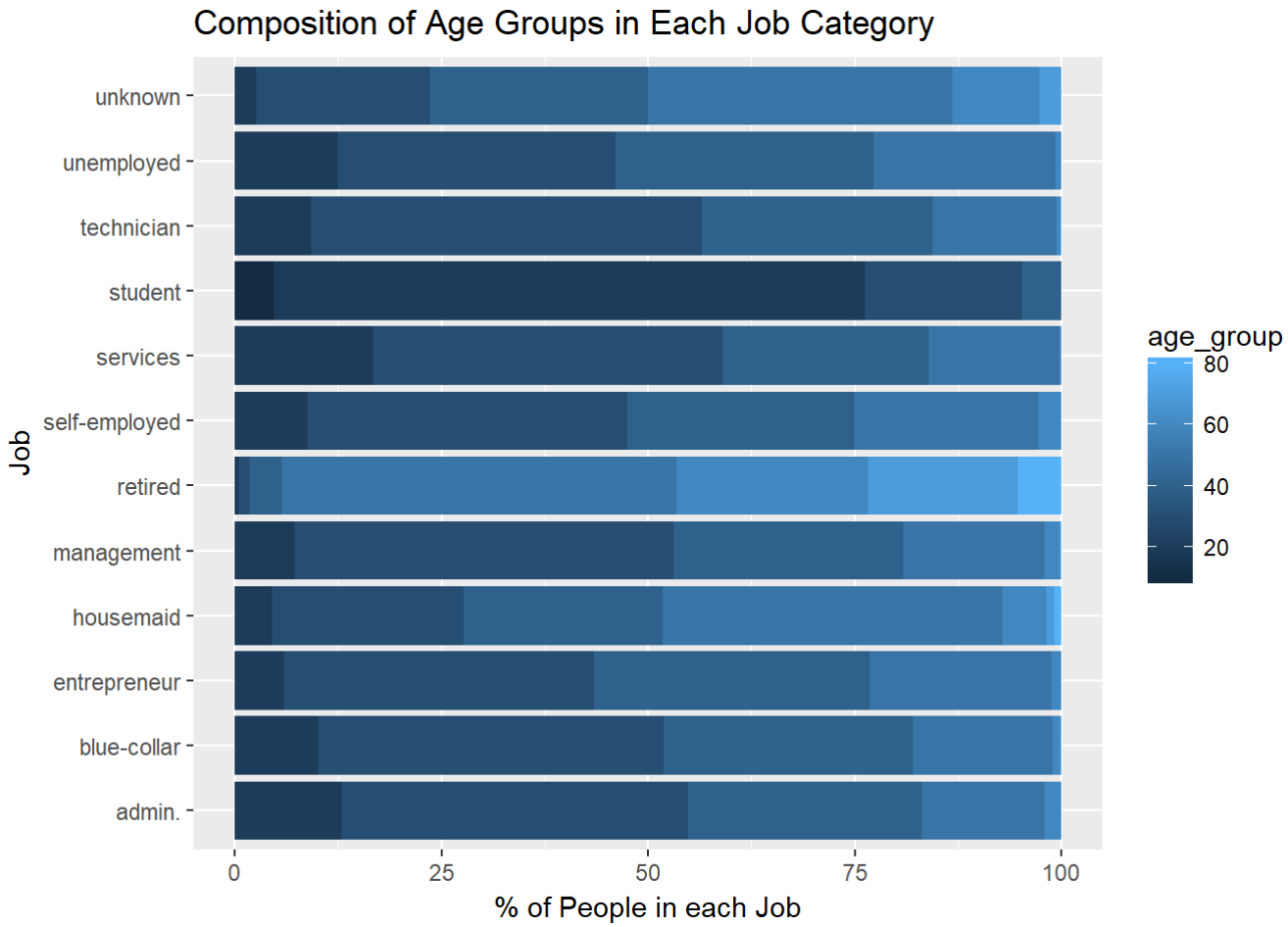
Finding #6 Education-Job Relationship




```
## # A tibble: 48 x 4
## # Groups:   job [12]
##             job education     n `Percentage in Job Category`
##          <fctr>    <fctr> <int>                        <dbl>
##  1       admin.   primary    17                     3.556485
##  2       admin. secondary   393                    82.217573
##  3       admin.  tertiary    51                    10.669456
##  4       admin.   unknown    17                     3.556485
##  5  blue-collar   primary   369                    39.006342
##  6  blue-collar secondary   524                    55.391121
##  7  blue-collar  tertiary    12                     1.268499
##  8  blue-collar   unknown    41                     4.334038
##  9 entrepreneur   primary    26                    15.476190
## 10 entrepreneur secondary    58                    34.523810
## # ... with 38 more rows
```

The above chart shows the composition of Education Levels in Each Job Category. It can be seen that management level jobs are mostly taken up by people with tertiary education level, while services, admin, services, technicians are mostly taken up by people with secondary education. Housemaids is mainly composed of people with primary education.

Finding #7 Job-Age Relationship

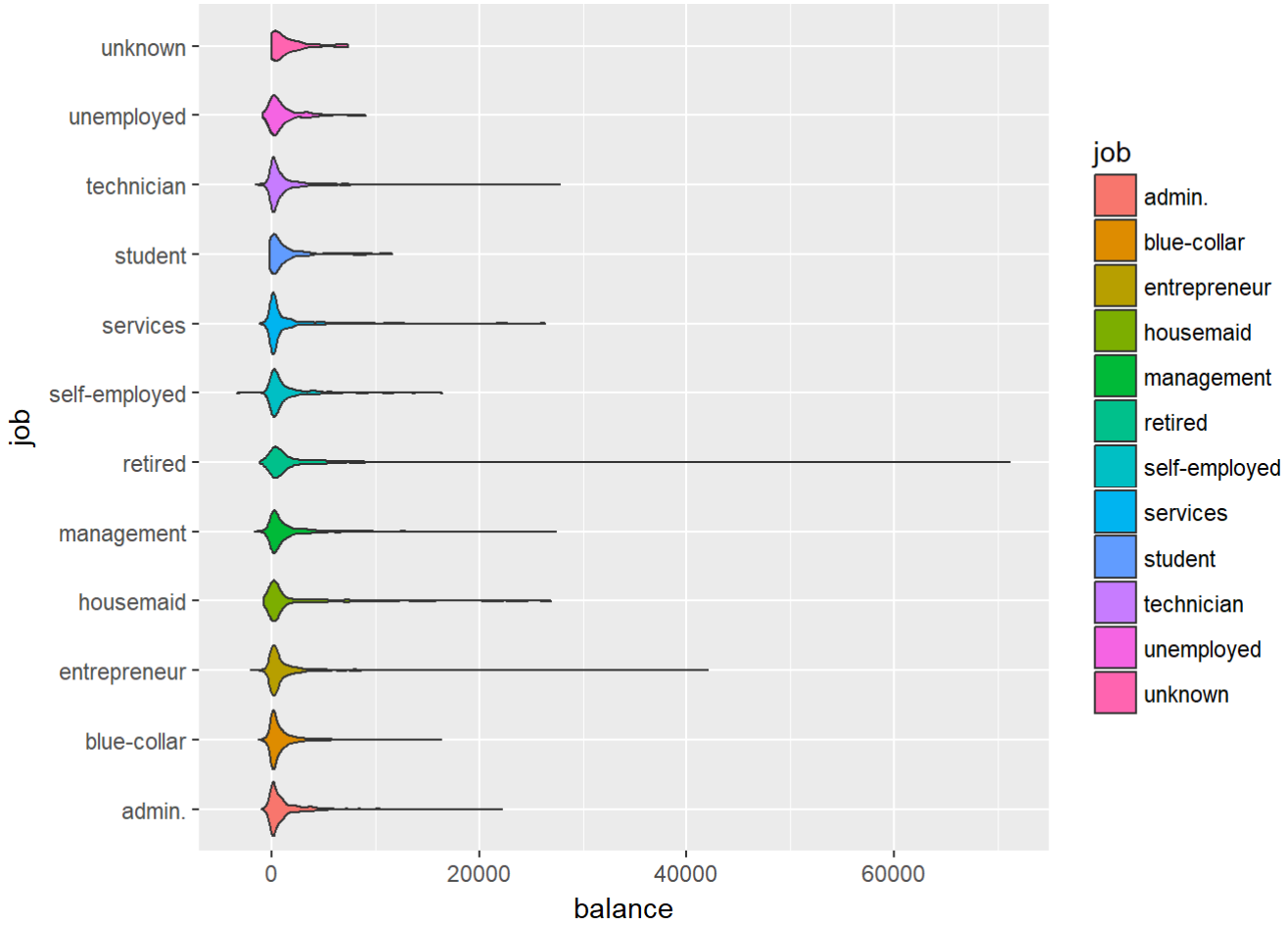



```
## # A tibble: 67 x 4
## # Groups:   job [12]
##            job age_group     n `Percentage in Job Category`
##         <fctr>     <dbl> <int>                        <dbl>
##  1      admin.        20    62                     12.970711
##  2      admin.        30   200                     41.841004
##  3      admin.        40   135                     28.242678
##  4      admin.        50    71                     14.853556
##  5      admin.        60    10                      2.092050
##  6 blue-collar        20    96                     10.147992
##  7 blue-collar        30   395                     41.754757
##  8 blue-collar        40   285                     30.126850
##  9 blue-collar        50   160                     16.913319
## 10 blue-collar        60     8                      0.845666
## # ... with 57 more rows
```

The above chart shows the age composition in the different job categories. It can be observed that students are mostly composed of people between 20-40 years old, with the bulk being in their 20s. Retired are mainly people older than 60 years old. Managerial roles are generally taken up by people in their 30s and 40s.

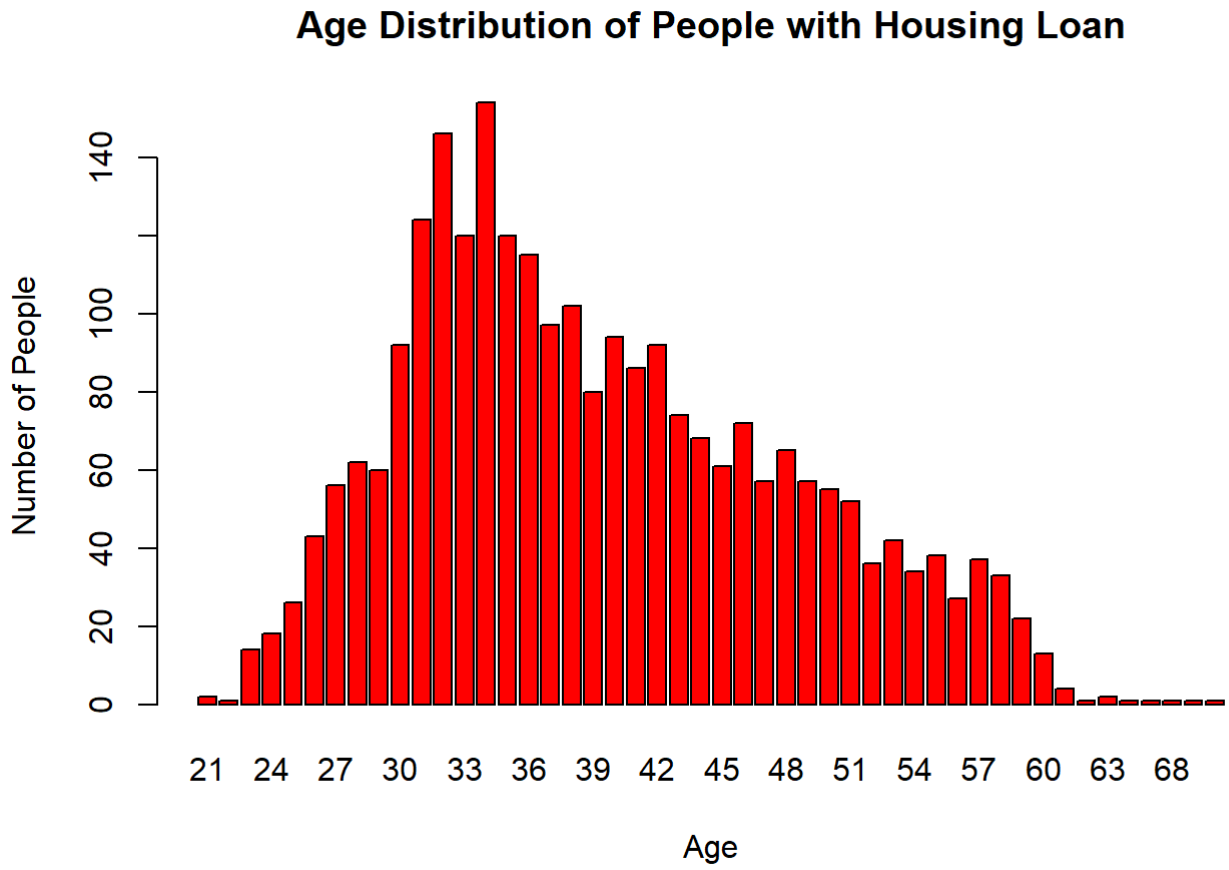
Finding #8 Min and Max Balance in each job category

```
## # A tibble: 12 x 4
##              job min_balance max_balance mean_balance
##           <fctr>       <dbl>       <dbl>        <dbl>
##  1        admin.        -967       22171     1226.736
##  2   blue-collar       -1400       16353     1085.162
##  3  entrepreneur       -2082       42045     1645.125
##  4     housemaid        -759       26965     2083.804
##  5    management       -1746       27359     1766.929
##  6       retired       -1206       71188     2319.191
##  7 self-employed       -3313       16430     1392.410
##  8      services       -1202       26394     1103.957
##  9       student        -230       11555     1543.821
## 10    technician       -1680       27733     1330.996
## 11    unemployed        -872        9019     1089.422
## 12       unknown           0        7337     1501.711
```



The above plot shows the distribution of balance in the different job categories. The maximum balance of the sample comes from the retired category (71188), while the minimum balance of the sample comes from the self-employed category (-3313).

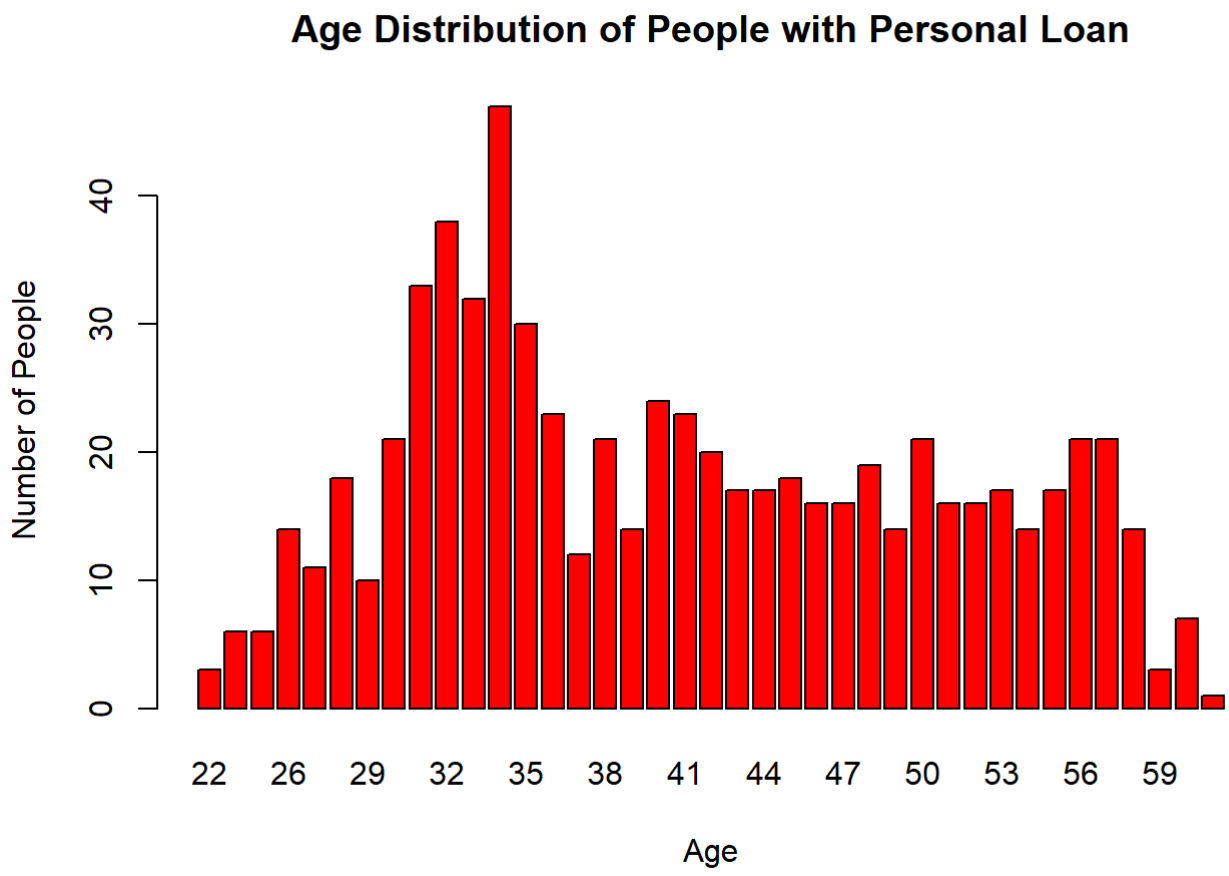
Finding #9 Age range of people with housing loan




```
## # A tibble: 48 x 3
## # Groups:   age [48]
##      age housing     n
##    <int>  <fctr> <int>
##  1    21     yes     2
##  2    22     yes     1
##  3    23     yes    14
##  4    24     yes    18
##  5    25     yes    26
##  6    26     yes    43
##  7    27     yes    56
##  8    28     yes    62
##  9    29     yes    60
## 10    30     yes    92
## # ... with 38 more rows
```

The above chart shows the age distribution of people with housing loan. The age range of people with housing loan is 21 to 75. It can be seen that the bulk of people with housing loans are aged 30-42, with the age 34 having the most people with a housing loan. In general, as age increases/decreases from the 30s, the number of people having housing loans decreases.

Finding #10 Age range of people with personal loan




```
## # A tibble: 39 x 3
## # Groups:   age [39]
##      age   loan     n
##    <int> <fctr> <int>
##  1    22    yes     3
##  2    24    yes     6
##  3    25    yes     6
##  4    26    yes    14
##  5    27    yes    11
##  6    28    yes    18
##  7    29    yes    10
##  8    30    yes    21
##  9    31    yes    33
## 10    32    yes    38
## # ... with 29 more rows
```

The above chart shows the age distribution of the people with personal loans. The age range of people with personal loans is 22 to 61. The age with the most people having personal loans is similar to that having housing loans at age 34. However, the number of people with personal loans decreases less gradually as age increases. In fact, the number of people with personal loans are relatively equal for ages 40 to 58.