PROJECT 2~ CODING CLUB

PROJECT 2: CODING CLUB

Author:Weru Joan Nyokabi

1. Import the data to R in csv format(this is SPSS data) and name it financial_data.csv

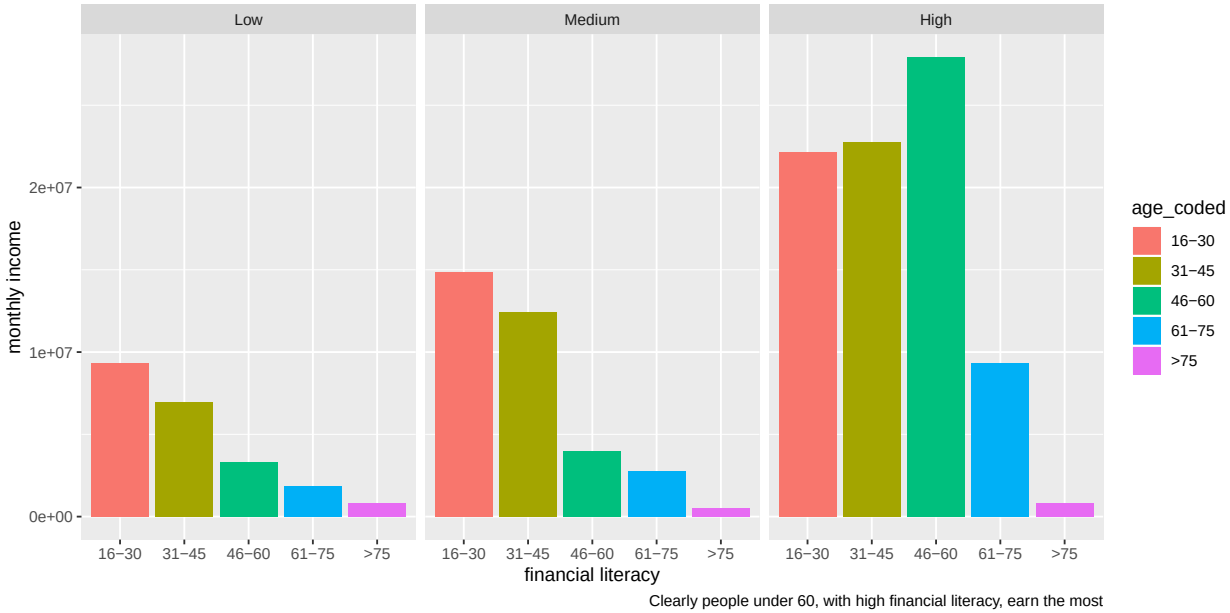
2.clean the data before undertaking any analysis

```
library(tinytex)
library(haven)
library(foreign)
library(tidyverse)
setwd("~/R- CLASS DATASETS")
finance<-read.spss("FinAccess_SPSS.sav",to.data.frame=TRUE,use.value.labels=TRUE)
View(finance)
```

3.Conduct exploratory analysis of the data and write a few bullet points on any descriptive statisics(summary statistics, tables and graphs) you find interesting and why you find them interesting

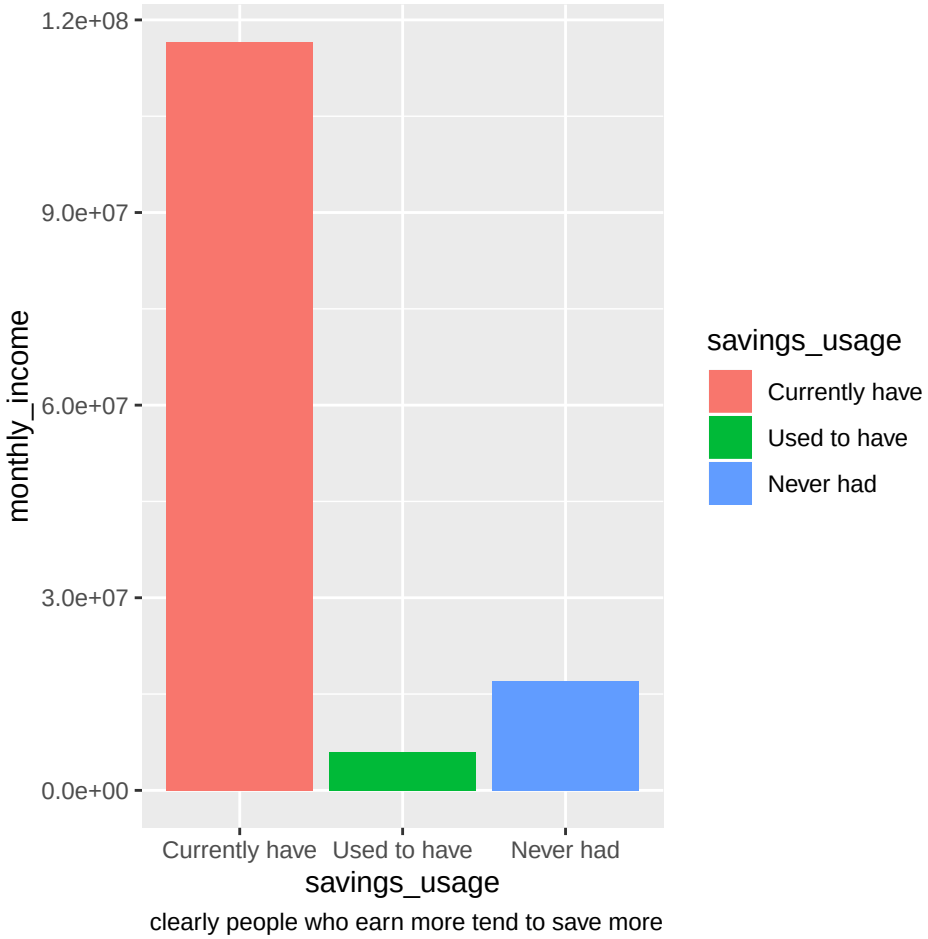
- We now want to see how monthly income differs among the different ages and different scales of financial literacy

```
library(ggplot2)
library(tigerstats)
finance%>%ggplot(aes(x=age_coded,y=monthly_income,fill=age_coded))+geom_col()+facet_wrap(~fin_literacy)+
```



- We now check to see if people's earnings differ in terms of their saving usage

```
ggplot(finance,aes(x=savings_usage,y=monthly_income,fill=savings_usage))+geom_col()+labs(caption = "clea
```



clearly people who earn more tend to save more

- We now seek to find out which age group has the highest financial literacy

```
library(knitr)
library(tigerstats)
rowPerc(xtabs(~age_coded+fin_literacy,data=finance))%>%knitr::kable()
```

	Low	Medium	High	Total
16-30	34.69	36.55	28.76	100
31-45	30.73	36.70	32.57	100
46-60	36.78	34.26	28.97	100
61-75	54.44	26.78	18.79	100
>75	73.67	17.79	8.54	100

people in the 31-45 age group have the highest financial literacy

4.Create a cross table of age_coded and cluster type

```
library(gmodels)
ct<-CrossTable(finance$age_coded,finance$cluster_type)
```

```
##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  8665
##
##
##                    | finance$cluster_type
## finance$age_coded |     Rural |     Urban | Row Total |
## ------------------|-----------|-----------|-----------|
##             16-30 |      1890 |      1976 |      3866 |
##                   |    34.879 |    44.383 |           |
##                   |     0.489 |     0.511 |     0.446 |
##                   |     0.390 |     0.518 |           |
##                   |     0.218 |     0.228 |           |
## ------------------|-----------|-----------|-----------|
##             31-45 |      1483 |      1130 |      2613 |
##                   |     0.269 |     0.342 |           |
##                   |     0.568 |     0.432 |     0.302 |
##                   |     0.306 |     0.296 |           |
##                   |     0.171 |     0.130 |           |
## ------------------|-----------|-----------|-----------|
##             46-60 |       803 |       426 |      1229 |
##                   |    19.156 |    24.376 |           |
##                   |     0.653 |     0.347 |     0.142 |
##                   |     0.165 |     0.112 |           |
##                   |     0.093 |     0.049 |           |
## ------------------|-----------|-----------|-----------|
##             61-75 |       467 |       209 |       676 |
##                   |    20.678 |    26.312 |           |
##                   |     0.691 |     0.309 |     0.078 |
##                   |     0.096 |     0.055 |           |
##                   |     0.054 |     0.024 |           |
## ------------------|-----------|-----------|-----------|
##               >75 |       209 |        72 |       281 |
##                   |    16.956 |    21.577 |           |
##                   |     0.744 |     0.256 |     0.032 |
##                   |     0.043 |     0.019 |           |
##                   |     0.024 |     0.008 |           |
## ------------------|-----------|-----------|-----------|
##      Column Total |      4852 |      3813 |      8665 |
##                   |     0.560 |     0.440 |           |
## ------------------|-----------|-----------|-----------|
##
##
```

5.Select the variables:age and monthly_income. Test for normality

We use the **Kolmogorov_Smirnov test** which checks whether the two sample data comes from the same distribution.

- D- value of the test statistic
- H_o(null hypothesis)- The two samples were drawn from the same population
- H_1(alternative hypotheses)- The two samples were not drawn from the same distribution
- p-value- a measure of strength of the evidence against the null hypothesis

```
ks.test(finance$age,finance$monthly_income)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  finance$age and finance$monthly_income
## D = 0.98107, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

our p-value is less than 0.05(level of significance) so we reject the null hypothesis and accept the alternative hypothesis; the two samples were not drawn from the same distribution

if possible, does monthly income depend on the age of the respondents

```
y<-finance[,"monthly_income"]
x<-finance[,"age"]
model1<-lm(y~x)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##   -27672   -12715    -9212    -1134 15131178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9233.2     4796.2   1.925   0.0543 .
## x              184.4      117.8   1.566   0.1175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 181700 on 8663 degrees of freedom
## Multiple R-squared:  0.0002828,  Adjusted R-squared:  0.0001674
## F-statistic: 2.451 on 1 and 8663 DF,  p-value: 0.1175
```

p-value is more than 0.05(level of significance),we accept the null hypothesis; there is no significant relationship between age and monthly income

6.Select the categorical variables on age and monthly income. Perform a Chi_square test. State what you are testing and your findings.

Chi square test is a statistical method which is used to determine if two categorical variables have a significant correlation between them. In this case our two categorical variables are monthly income and finance

- H_0- The two variables are independent

- H_1-The two variabes relate to each other

```
cq<-chisq.test(finance$age,finance$monthly_income)
cq
```

```
##
##  Pearson's Chi-squared test
##
## data:  finance$age and finance$monthly_income
## X-squared = 39095, df = 39093, p-value = 0.4956
```

Our p-value is more than 0.05(level of significance) so we accept the null hypothesis and conclude that our two variables are independent

7.Perform a one way ANOVA between monthly income and highest education of the respondent. Does monthly income depend on the level of education?

- H_o-The means between groups are identical
- H_1 least the mean of one group is different
- F-statistic= Variation among sample means/variation within groups; there is a relationship between the groups

```
anova1<-aov(finance$monthly_income~finance$education)
summary(anova1)
```

```
##                     Df    Sum Sq   Mean Sq F value   Pr(>F)
## finance$education    3 2.241e+12 7.470e+11    22.8 1.07e-14 ***
## Residuals         8661 2.838e+14 3.276e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our p-value is less than our level of significance(0.05) we therefore reject the null hypothesis and conclude that monthly income does indeed depend on education

8.Test for statistical significance using one sample t test to find out if the average income is less than Kshs.15,000?

- H_0-true mean is greater or equal to 15000
- H_1true mean is less than 15,000

```
t.test(finance$monthly_income,mu=15000,alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  finance$monthly_income
## t = 0.55951, df = 8664, p-value = 0.7121
## alternative hypothesis: true mean is less than 15000
## 95 percent confidence interval:
##      -Inf 19302.81
## sample estimates:
## mean of x
##  16092.05
```

Our p value is more than 0.05(level of significance) we accept the null and conclude average income is greater than or equal to 15,000

9.Using the generalized linear model, perform the following

- a)Does mobile usage depend on gender, religion, education and credit usage
- b)Does bank usage depend on financial literacy, numeracy, savings usage and credit usage

```
attach(finance)
```