R Project - Household

*EDA, Summarization, Visualization*

*By: Peng Wang*

*Instructor: Mr. Hamid Rajaee*

*Metro College Toronto*

*Due: Friday, March 5, 2021*

RStudio version 1.4.1103.

Project Scope

**Data Source**

- <http://stat511.cwick.co.nz/homeworks/acs_or.csv>

**Industry Orient**

* + Real Estate / Agent
  + Banking / Mortgage / Loan
  + City Hall / CRA / Property Tax
  + Household Utilities Providers

**Analysis Tasks**

* + Identify the distribution of Income Group, i.e., Low Income, Middle Class, and High-Income families;
  + Analysis of the relationship between Communication Mode and Income Group;
  + Identify and handle outlier, if any;
  + Analysis of the relationship between the number of bedrooms and internet accessibility;
  + Identify the distribution of household ownership;
  + Identify the distribution based on the built decade of houses;
  + Analysis of the relationship between house owner's age and income;
  + Analysis of the relationship between house owner's age and internet accessibility;
  + Analysis of the relationship between ownership and income group.

**R Learning Points and Skills**

**Data Set Cleaning**

* + - Changing Working Directory
    - Importing and Reading Data
    - Understanding Data
    - Cleaning Data
    - Processing and Amending Data
    - Outliers Handling

**Data Set Summarization**

* + - Distribution Analysis
    - Segmentation
    - Contingency Table (Two-way Table)

**Data Set Visualization**

* + - Pie Chart
    - Simple Bar Chart
    - Histogram Plots
    - Stacked Bar Chart
    - Grouped Bar Chart
    - Mosaic Plots
    - Association Plots

**Data Set Relationship Analysis**

* + - Bivariate Analysis
    - Chi-square Test
    - T-test

Question 1 – Data Cleaning

Add a new column of 'income\_total' which is the sum of 'income\_husband' and 'income\_wife'. Then segment to 'income\_group' by 'income\_total'.

*# 1.1: Add the 1st column - income\_total*

*hh$income\_total <- hh$income\_husband + hh$income\_wife*

*summary(hh$income\_total)*

## Min. 1st Qu. Median Mean 3rd Qu. Max.

-9000 42300 69700 88816 106000 979000

*# 1.2: Add the 2nd column - income\_group*

*hh$income\_group = ifelse(hh$income\_total < 50000, "Low Income",*

*ifelse(hh$income\_total < 300000, "Middle Class",*

*ifelse(hh$income\_total >= 300000, "High Income", "")))*

*str(hh)*

## 'data.frame': 7811 obs. of 15 variables:

$ household : int 48 218 279 612 947 1373 1733 1858 1947 1962 ...

$ age\_husband : int 64 63 56 71 37 86 67 70 33 41 ...

$ age\_wife : int 62 64 51 68 33 91 67 74 31 47 ...

$ income\_husband : int 11000 100000 31000 51700 16600 77500 8400 73670...

$ income\_wife : int 29200 3100 0 8800 26000 30000 4800 11000 600 ...

$ bedrooms : num 1 4 2 3 3 4 4 0 1 3 ...

$ electricity : int 90 230 200 170 260 20 70 180 20 80 ...

$ gas : int 3 30 40 3 3 30 150 80 30 200 ...

$ number\_children: int 0 0 0 0 2 0 0 0 0 2 ...

$ internet : chr "Yes" "Yes" "No" "Yes" ...

$ mode : chr "followup" "mail" "followup" "internet" ...

$ own : chr "Owned with mortgage or loan" "Owned with mortgage or loan" "Rented" "Owned free and clear" ...

$ decade\_built : int 1940 1990 1950 1950 1990 1980 1980 2000 1930 ...

$ income\_total : int 40200 103100 31000 60500 42600 107500 13200 ...

$ income\_group : chr "Low Income" "Middle Class" "Low Income"...

Question 2 – Distribution & Charts

What is the distribution of the variable 'income\_group'?

*# 2.1: list the distribution*

*tbl <- aggregate(hh$income\_group,list(hh$income\_group),length)*

*tbl*

## Group.1 x

1 High Income 250

2 Low Income 2489

3 Middle Class 5072

*# 2.2: 3D Pie Chart*

*install.packages('plotrix')*

*library(plotrix)*

*count <- table(hh$income\_group)*

*pct <- round(count/sum(count)\*100)*

*lbls <- c("High Income", "Low Income", "Middle Class")*

*lbls <- paste(lbls, pct) # add pct to label*

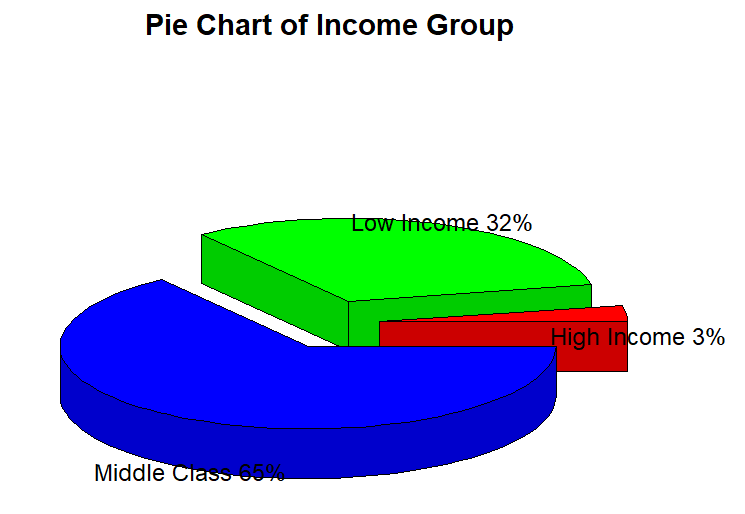
*lbls <- paste(lbls, "%", sep = "") # add % to pct*

*pie <- pie3D(count,*

*explode=0.2,*

*main = "Pie Chart of Income Group")*

*pie3D.labels(pie, labels = lbls)*



*# 2.3: Simple Bar Plot*

*counts <- table(hh$income\_group)*

*counts*

*barplot(counts,*

*main = "Simple Bar Plot: Income Group",*

*xlab = "income\_group",*

*ylab = "Frequency",*

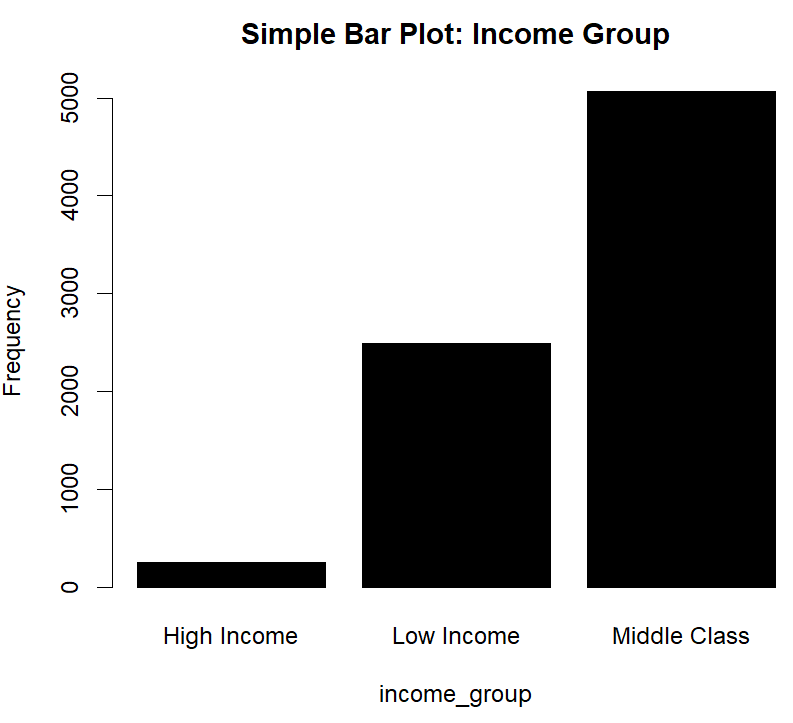
*col = 'black',*

*horiz = FALSE)*

##

High Income Low Income Middle Class

250 2489 5072

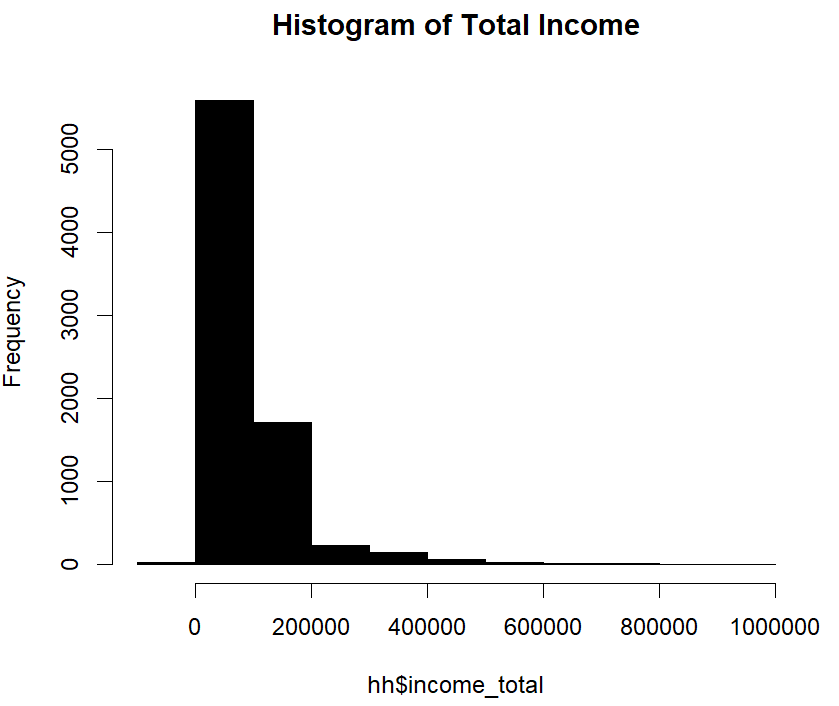


*# 2.4: Histogram of income\_total*

*hist(hh$income\_total,*

*main = "Histogram of Total Income",*

*col = "black")*



Question 3 - Bivariate Analysis

Is there any relation between communication mode and target(income\_group)?

**Bivariate analysis for categorical vs. categorical**

For visualization:

* Stacked bar chart or grouped bar chart

For summarization:

* Contingency table(two-way table)

For the test of independence:

* chi-square test

*# 3.1 - Visualization: Stacked Bar Plot*

*tbl <- table(hh$mode,hh$income\_group)*

*tbl*

*counts <- tbl[1:3,1:3]*

*counts*

*barplot(counts,*

*main = "Communication Mode Vs. Income Group",*

*xlab = "Income Group",*

*col = c("black","red", "yellow"),*

*legend = rownames(counts),*

*args.legend = list(x ='top', bty='n', inset=c(0,0)))*

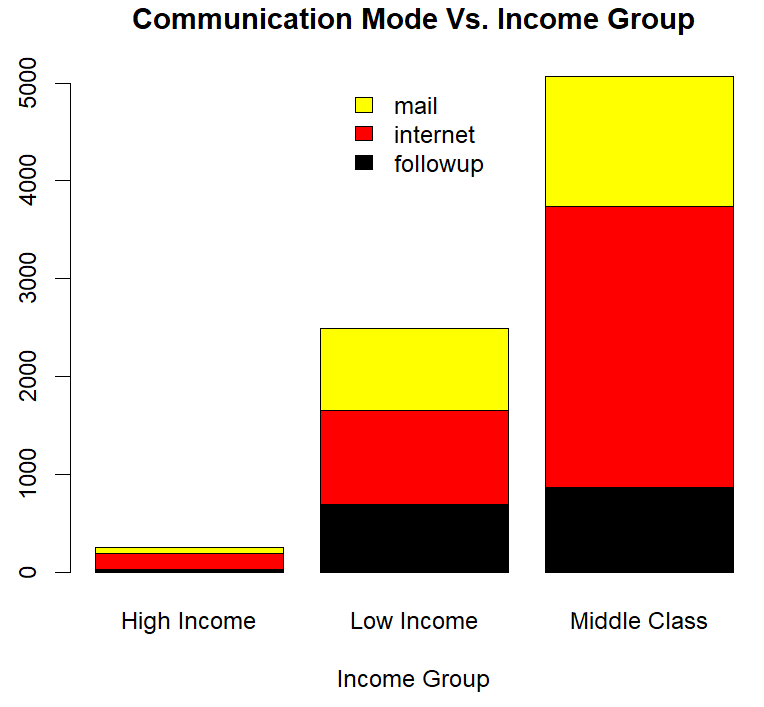
*# Legend position:* [*https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r*](https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r)

## High Income Low Income Middle Class

followup 27 694 865

internet 166 960 2878

mail 57 835 1329



*# 3.2 - Summarization: Contingency Table*

*add <- addmargins(xtabs(~ mode + income\_group,data=hh))*

*add[1:4,1:4]*

*proportions(xtabs(~ mode + income\_group,data=hh))[1:3,1:3]*

income\_group

mode High Income Low Income Middle Class

followup 0.003456664 0.088849059 0.110741262

internet 0.021252080 0.122903597 0.368454743

mail 0.007297401 0.106900525 0.170144668

*# 3.3 - Indipendency: Chi-square Test*

*# 3.3.1 Problem:*

*# Test whether the communication mode is independent of the income group at a 0.05 significance level.*

*# Null hypothesis: Communication Mode is independent of Income Group*

*# 3.3.2 Solution:*

*# p-value*

*library(MASS)*

*tbl <- table(hh$mode,hh$income\_group)*

*tbl*

*chisq.test(tbl) # the p-value < 2.2e-16*

##

Pearson's Chi-squared test

data: tbl

X-squared = 261.59, df = 4, p-value < 0.00000000000000022

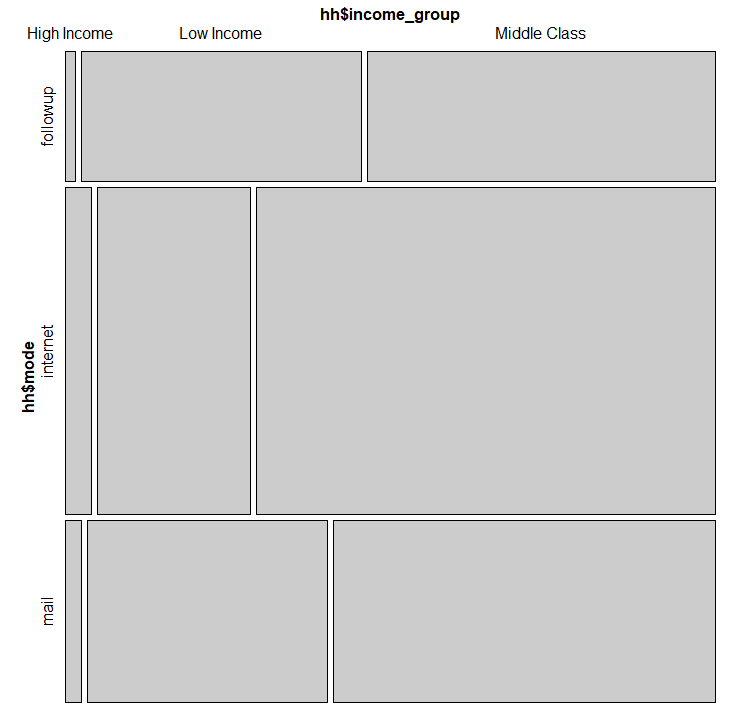
*# Mosaic Plots*

*library(vcd)*

*library(grid)*

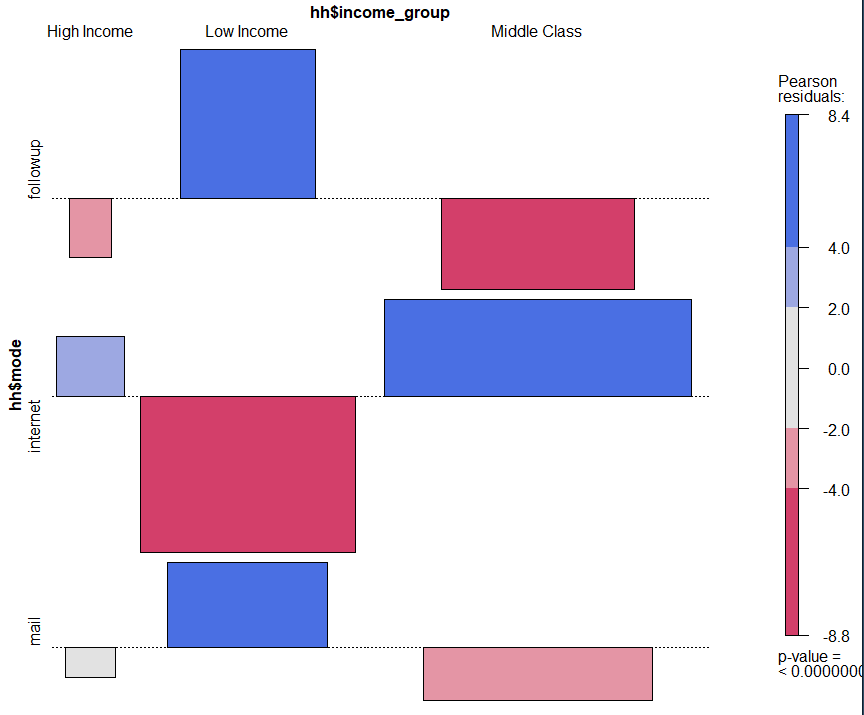
*mosaic(structable(hh$income\_group ~ hh$mode))*

*# structable:* [*https://stackoverflow.com/questions/14547162/missing-value-where-true-false-needed-error-vcdmosaic*](https://stackoverflow.com/questions/14547162/missing-value-where-true-false-needed-error-vcdmosaic)



*# Association Plots*

*assoc(hh$income\_group ~ hh$mode, shade=TRUE)*



**# 3.3.3 Conclusion:**

As the p-value 2.2e-16 is less than the 0.05 significance level, we **reject** the null hypothesis that Communication Mode is independent of the Income\_Group and conclude that in our data, the 'mode' and the 'income\_group' are statistically significantly associated (p-value = 0).

Question 4

What is Bedrooms distribution, how to handle the outlier, if any?

*# 4.1 summary*

*summary(hh$bedrooms) # 10 rooms seems too much*

##

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 3.000 3.000 3.117 4.000 10.000

*# 4.2 histogram*

*hist(hh$bedrooms,*

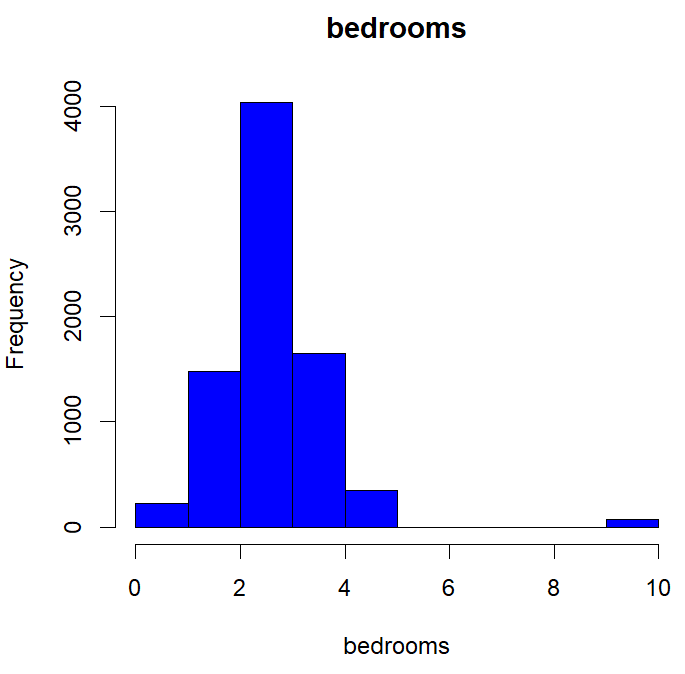
*breaks = 8,*

*main = "bedrooms",*

*col = "blue",*

*xlab = "bedrooms",*

*ylab = "Frequency")*



*# 4.3 Boxplot of Bedrooms by internet*

*boxplot(bedrooms ~ internet,*

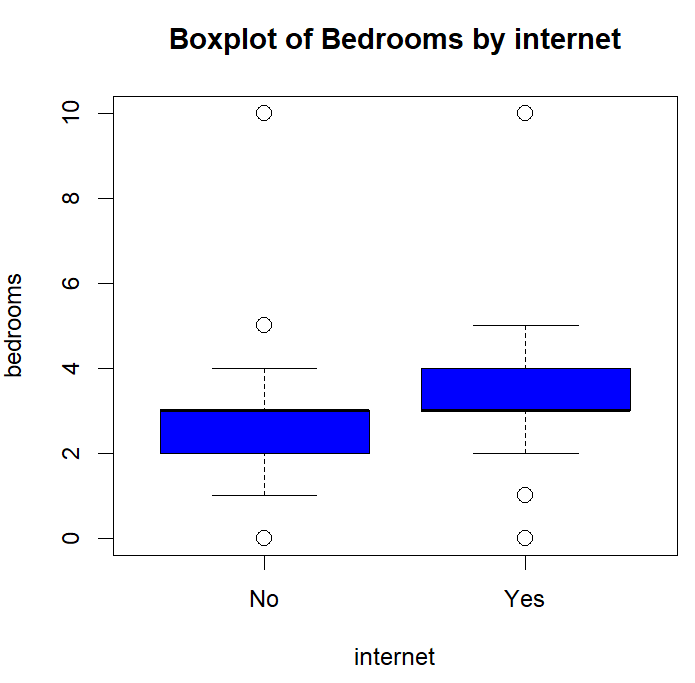
*data = hh,*

*main = "Boxplot of Bedrooms by internet",*

*xlab = "internet",*

*ylab = "bedrooms",*

*col = "blue")*



*# 4.4 pattern of outlier*

*bed\_out <- hh[which(hh["bedrooms"]==10),]*

*bed\_out$bedrooms # total 72 obs cross all types of ownership, income\_group, built years...*

*summary(bed\_out["bedrooms"])*

*nrow(bed\_out)*

bedrooms

Min. :10

1st Qu.:10

Median :10

Mean :10

3rd Qu.:10

Max. :10

> nrow(bed\_out)

[1] 72

*# 4.5 prove the bedroom numbers = 10 are just scaled up by 10.*

*count <- 0*

*for (val in bed\_out$bedrooms){*

*if (val%%10 !=0) {count = count+1}*

*}*

*count # count = 0 means all the bedrooms equal to 10 are scaled up by 10*

*# 4.6 amend outlier by deviding by 10*

*hh$bedrooms <- ifelse(hh$bedrooms == 10, hh$bedrooms/10, hh$bedrooms)*

*summary(hh["bedrooms"]) # Max reduced to 5.*

##

bedrooms

Min. :0.000

1st Qu.:3.000

Median :3.000

Mean :3.034

3rd Qu.:4.000

Max. :5.000

Question 5 – T-test

Is there any relationship between Bedrooms and Internet(Yes/No)?

**Continuous Vs. Categorical**

For summarization:

group by categorical column an aggregate for numerical column

For visualization:

Grouped box plot

For the test of independence :

1) if the categorical column has only two levels: t-test

2) if the categorical column has more than two levels: ANOVA

*# 5.1: Summary grouped by Internet(Yes/No)*

*agg1 <- aggregate(bedrooms ~ internet, hh , mean)*

*agg1*

##

internet bedrooms

1 No 2.732087

2 Yes 3.060957

*# 5.2: Visualization by qplot*

*library(ggplot2)*

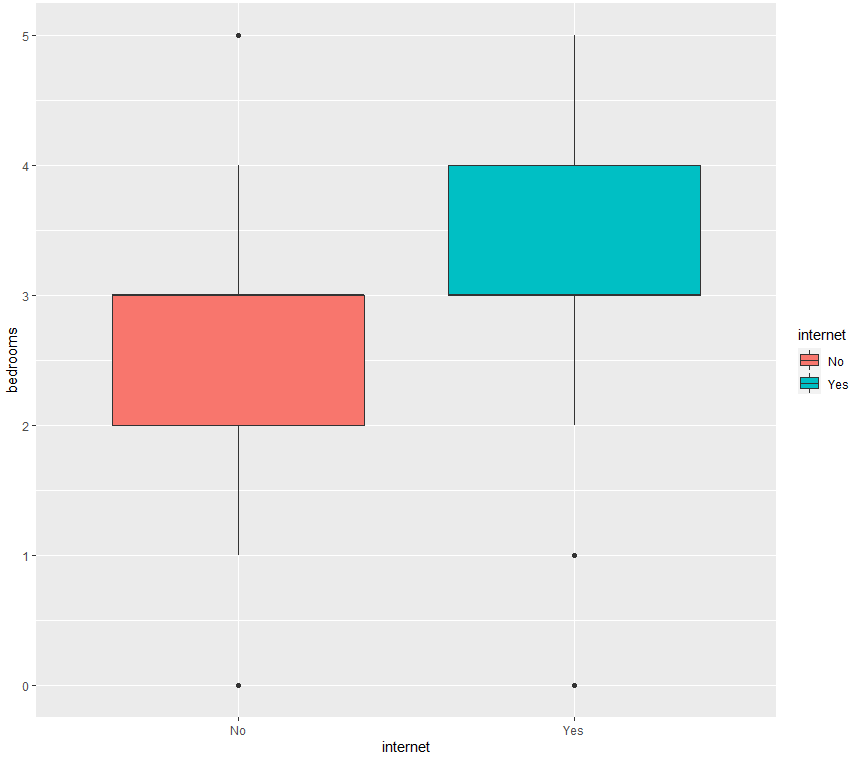
*qplot(internet,*

*bedrooms,*

*data = hh,*

*geom="boxplot",*

*fill = internet)*



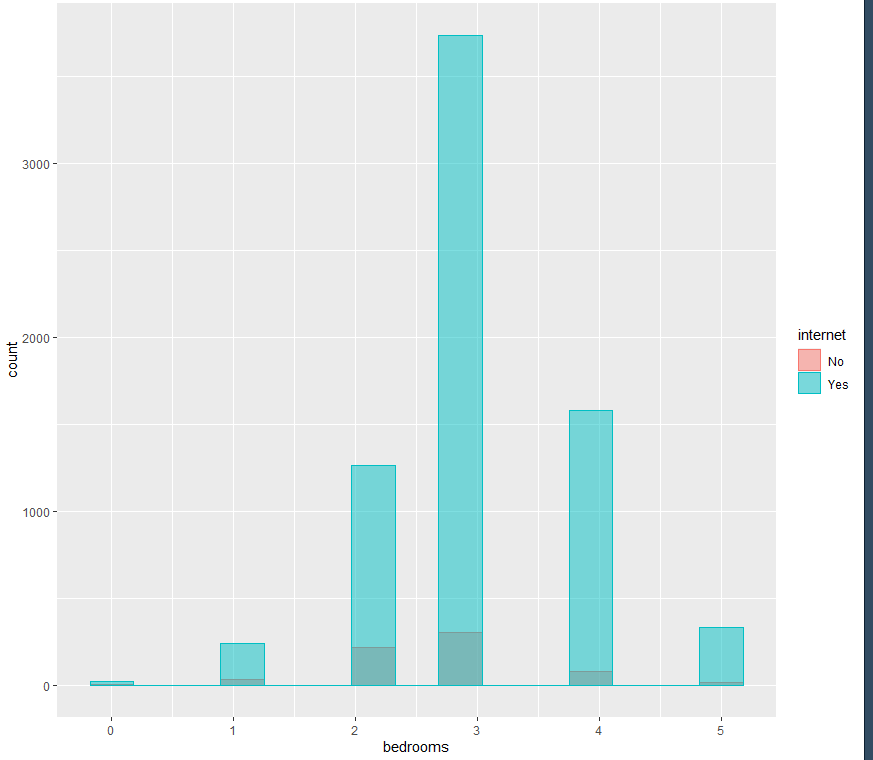
*# 5.3: Changing histogram plot fill colors by internet and usinging semi-transparent fill*

*p <- ggplot(hh,aes(x=bedrooms, fill=internet, color=internet)) +*

*geom\_histogram(position="identity", bins=15, alpha=0.5)*

*# bins: https://stackoverflow.com/questions/34774120/set-number-of-bins-for-histogram-directly-in-ggplot*

*p*



*# 5.4: Add mean lines*

*library(plyr)*

*mu <- ddply(hh, "internet", summarise, grp.mean=mean(bedrooms,na.rm=T))*

*head(mu)*

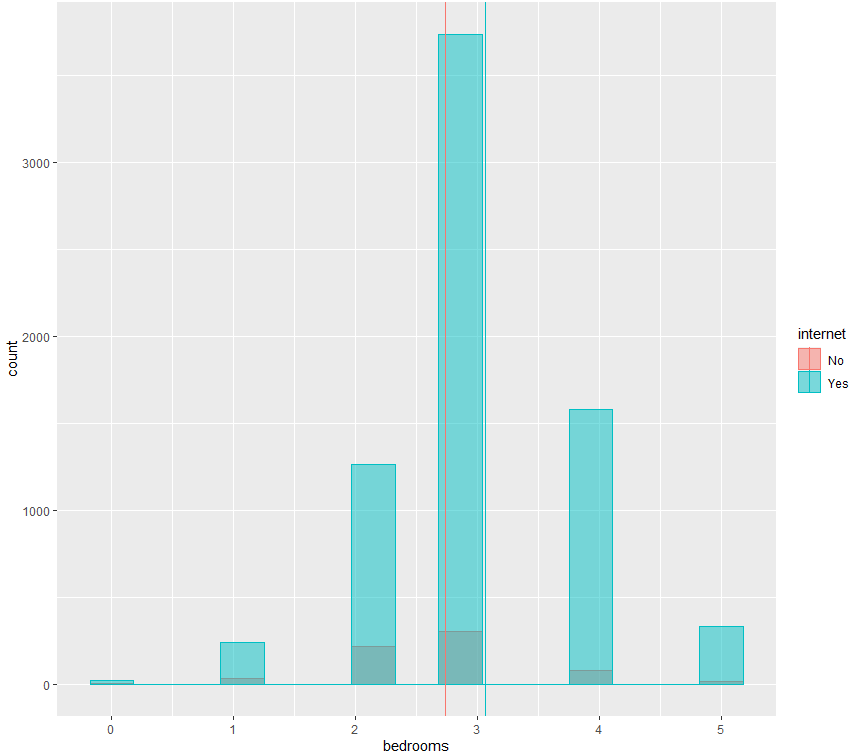
*p <- p + geom\_vline(data=mu, aes(xintercept=grp.mean, color=internet),linetype="solid")*

*p*

internet grp.mean

1 No 2.732087

2 Yes 3.060957



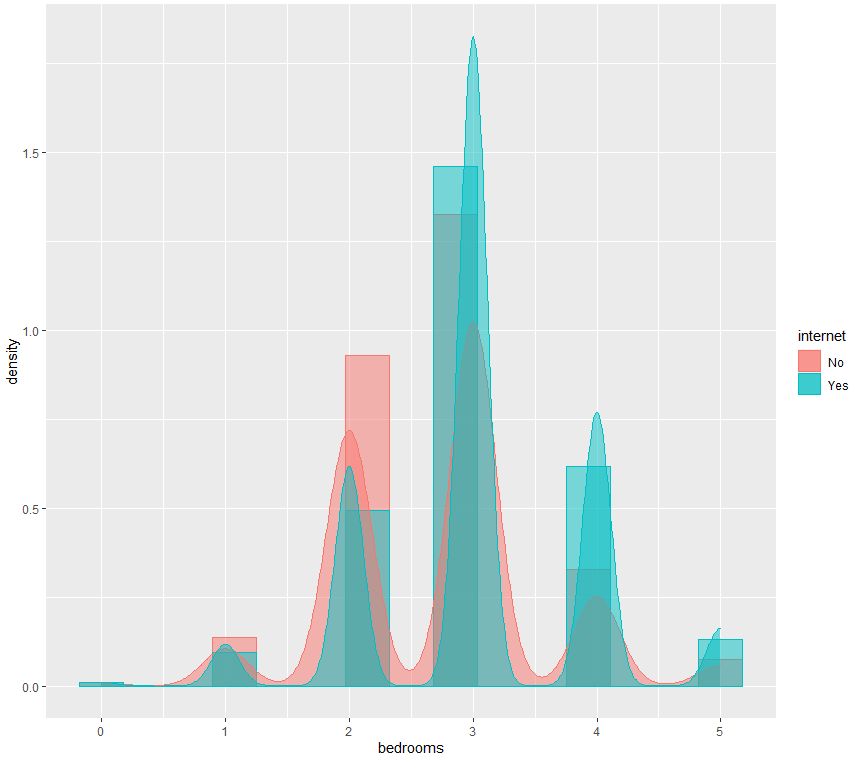
*# 5.5: Add density*

*p <- ggplot(hh, aes(x=bedrooms, fill=internet, color=internet)) +*

*geom\_histogram(aes(y=..density..),bins=15, position="identity", alpha=0.5)+*

*geom\_density(alpha=0.5)*

*p*

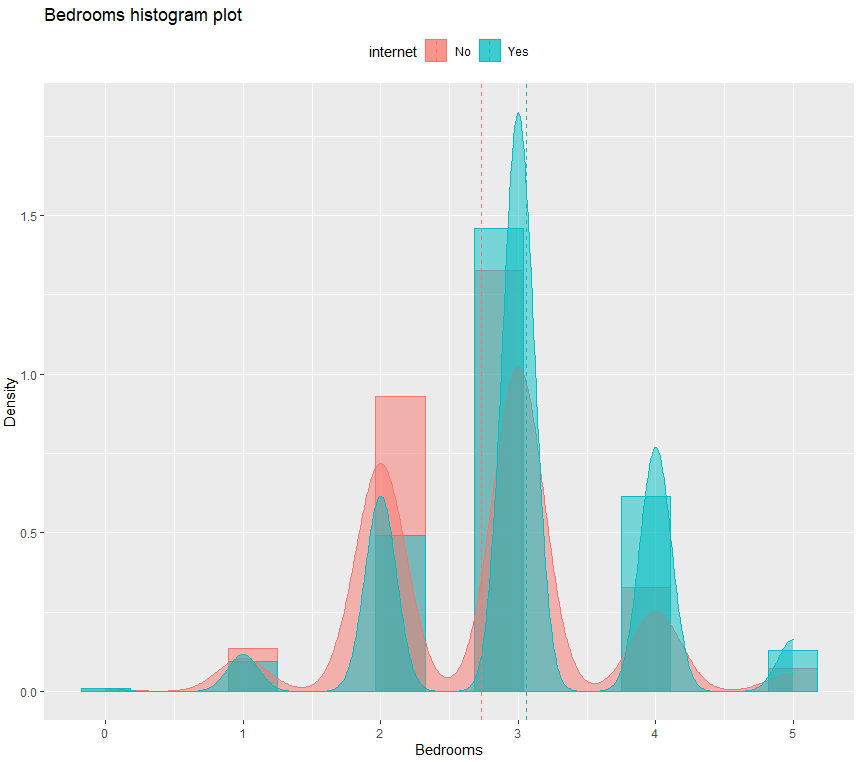


*# 5.6: Add mean lines and Change the legend position*

*p + geom\_vline(data=mu, aes(xintercept=grp.mean, color=internet),linetype="dashed")+*

*theme(legend.position="top")+*

*labs(title="Bedrooms histogram plot", x="Bedrooms", y = "Density")*



*# 5.7 t-test*

*# Yes: House has internet, No: House has no internet*

*# Null Hypothesis: µYes = µNo (the means of both populations are equal)*

*# Alternate Hypothesis: µYes <> µNo (the means of both populations are not equal)*

*t.test(bedrooms ~ internet, data=hh )*

Welch Two Sample t-test

data: bedrooms by internet

t = -9.4886, df = 766.41, p-value < 0.00000000000000022

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3969082 -0.2608311

sample estimates:

mean in group No mean in group Yes

2.732087 3.060957

*# Conclusion: p-value is less than 0.05, so the mean values between uYes and uNo are not equal.*

Question 6

What is the distribution of ownership?

*# 6.2: Pie Chart*

*count <- table(hh$own)*

*count*

*freq1 <- c(count[1], count[2], count[3], count[4])*

*lbls <- c("Occupy", "Own", "Mortgage", "Rent")*

*pct <- round(freq1/sum(freq1)\*100)*

*lbls <- paste(lbls, pct) # add percents to labels*

*lbls <- paste(lbls,"%",sep="") # ad % to labels*

*pie(freq1,*

*labels = lbls,*

*col = rainbow(length(lbls)),*

*main = "Pie Chart of Ownership")*

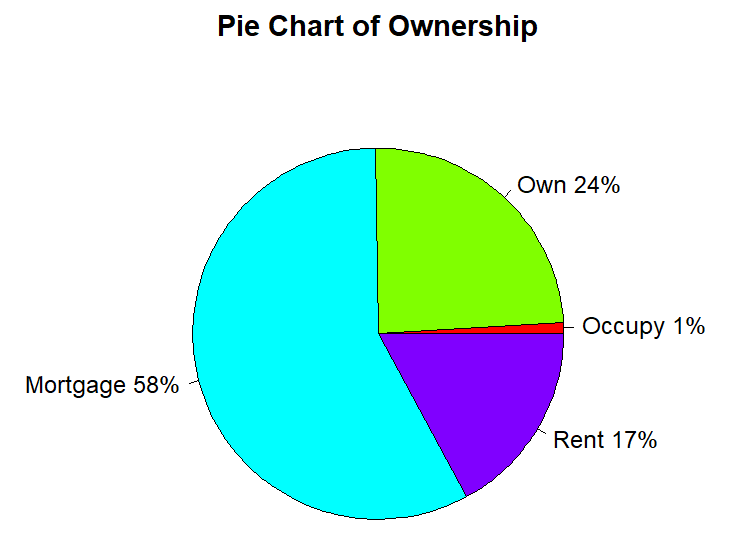
Group.1 x

1 Occupied without payment of rent 76

2 Owned free and clear 1896

3 Owned with mortgage or loan 4505

4 Rented 1334



*# 6.3: Simple Bar Plot*

*counts <- table(hh$own)*

*counts*

*barplot(counts,*

*main = "Simple Bar Plot: Ownership",*

*xlab = "Ownership",*

*ylab = "Frequency",*

*col = 'black',*

*horiz = FALSE)*

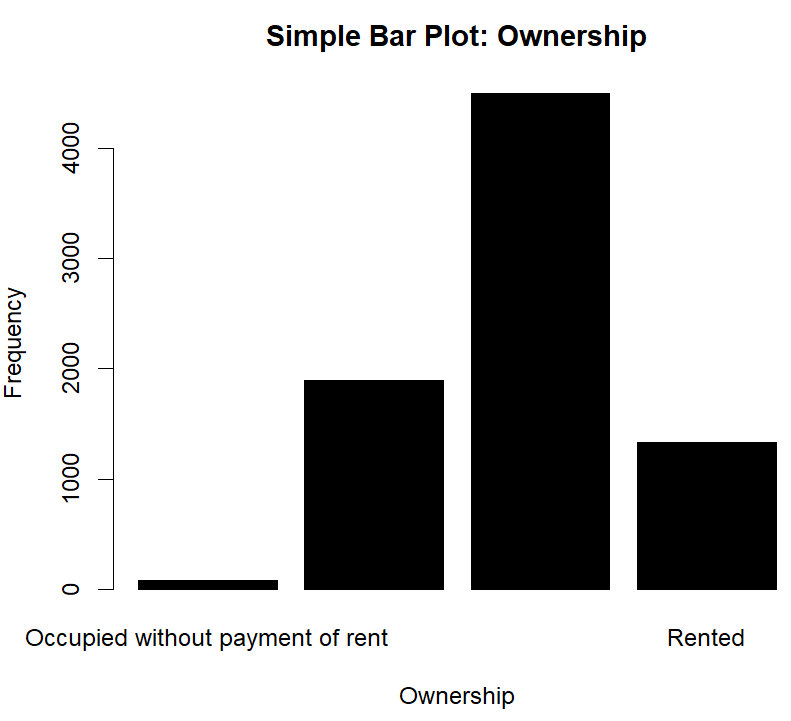
Group.1 x

1 Occupied without payment of rent 76

2 Owned free and clear 1896

3 Owned with mortgage or loan 4505

4 Rented 1334



Question 7

What is the distribution of built year?

tbl <- aggregate(hh$decade\_built,list(hh$decade\_built),length)

tbl

Group.1 x

1 1930 1021

2 1940 435

3 1950 671

4 1960 684

5 1970 1415

6 1980 803

7 1990 1444

8 2000 1234

9 2010 104

# 7.2: Pie Chart

count <- table(hh$decade\_built)

pct <- round(count/sum(count)\*100)

lbls <- hh$decade\_built

lbls <- paste(lbls, pct) # add pct to label

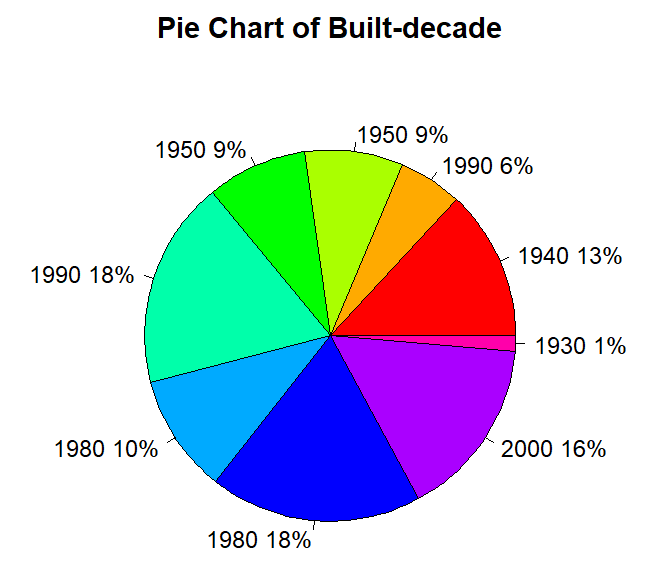
lbls <- paste(lbls, "%", sep = "") # add % to pct

pie(count,

labels = lbls,

col = rainbow(length(pct)),

main = "Pie Chart of Built-decade")



# 7.3: Simple Bar Plot

counts <- table(hh$decade\_built)

counts

barplot(counts,

main = "Simple Bar Plot: Built Decade",

xlab = "Ownership",

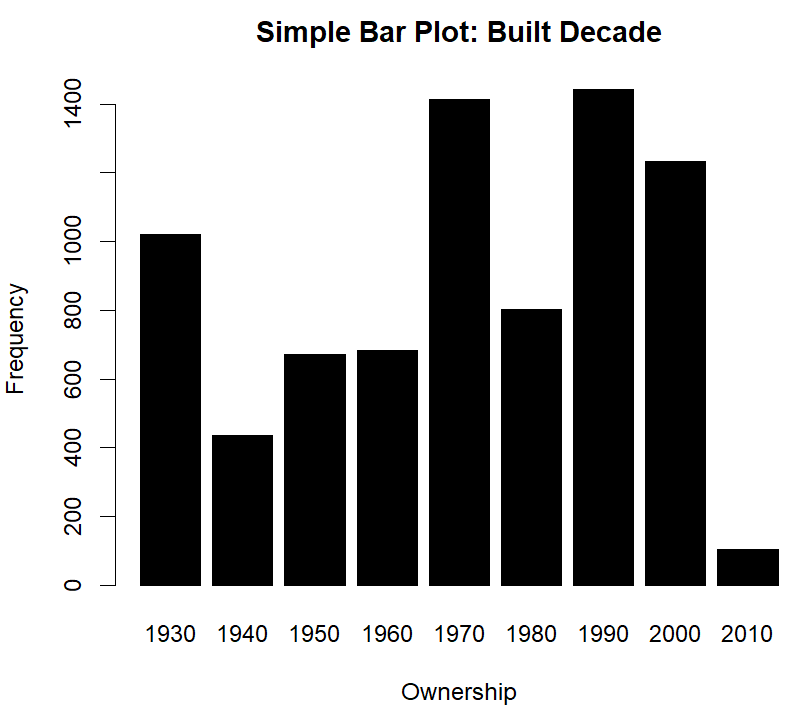
ylab = "Frequency",

col = 'black',

horiz = FALSE)

1930 1940 1950 1960 1970 1980 1990 2000 2010

1021 435 671 684 1415 803 1444 1234 104

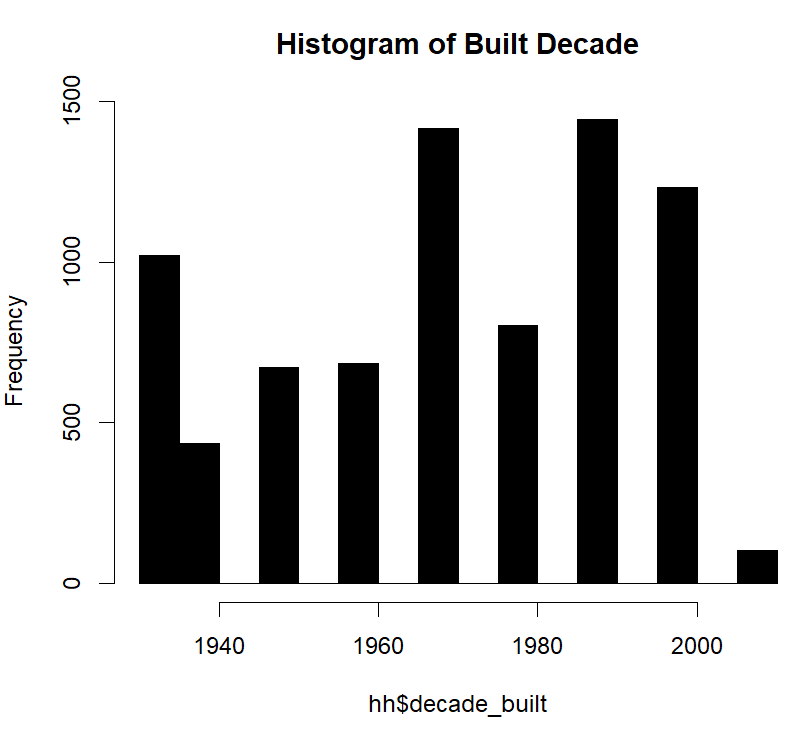


# 7.4: Histogram

hist(hh$decade\_built,

main = "Histogram of Built Decade",

col = "black")

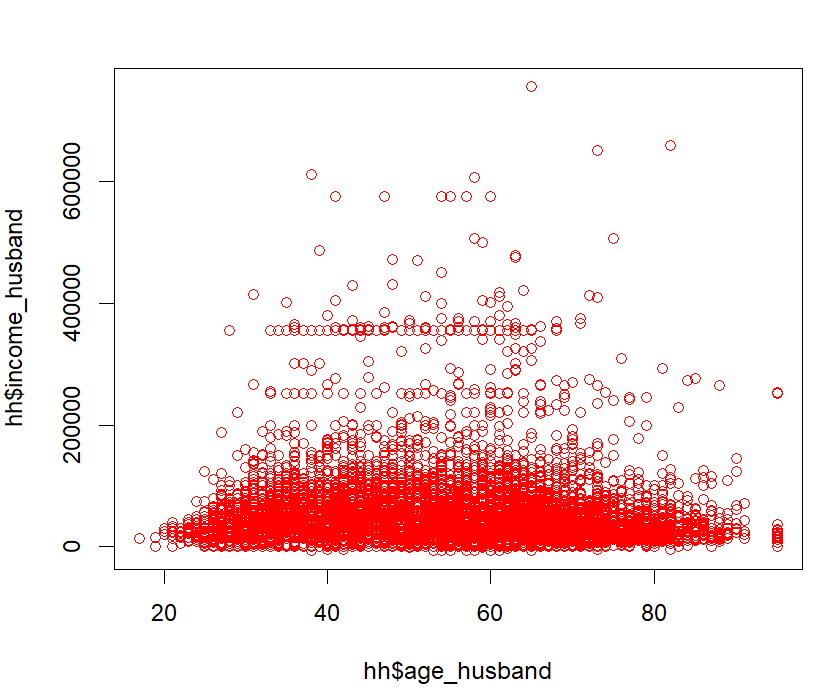


Question 8

Are there any relationships between the husband's age and his income?

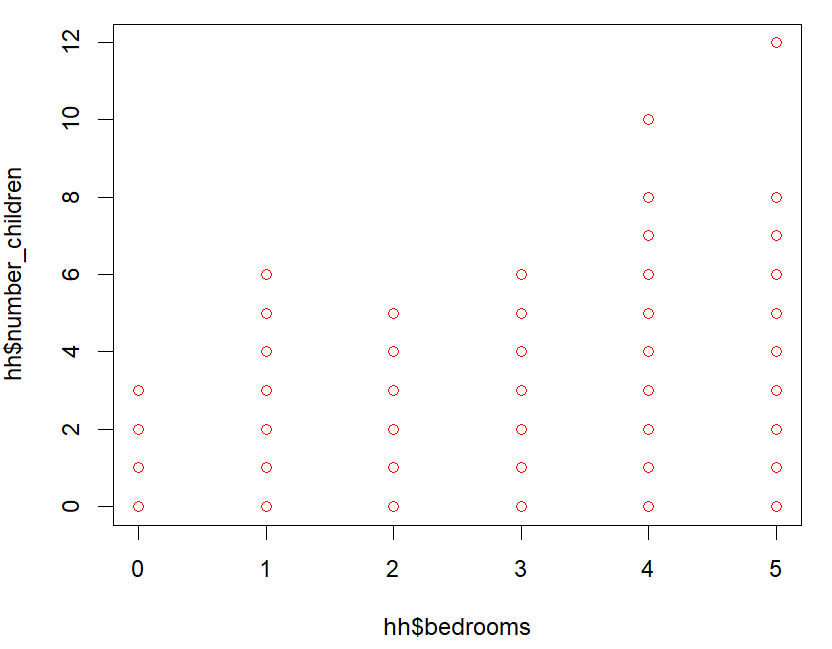
# create a scatter plot of a data set

plot(x = hh$age\_husband , y = hh$income\_husband, type = 'p', col="red")



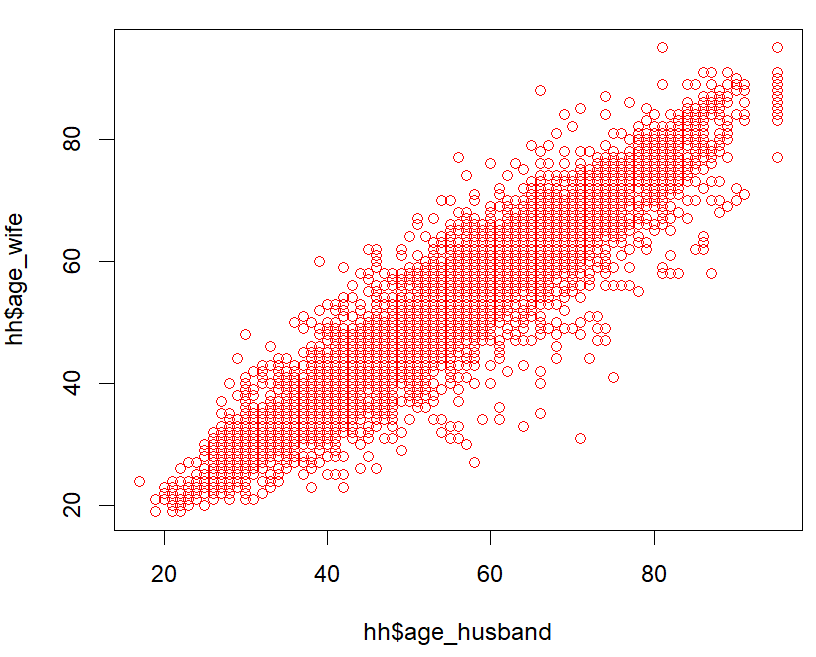
# Answer: In this data set, there's **NO** evidence to prove there's a relation between the husband's age and his income.

plot(x = hh$bedrooms , y = hh$number\_children, type = 'p', col="red")



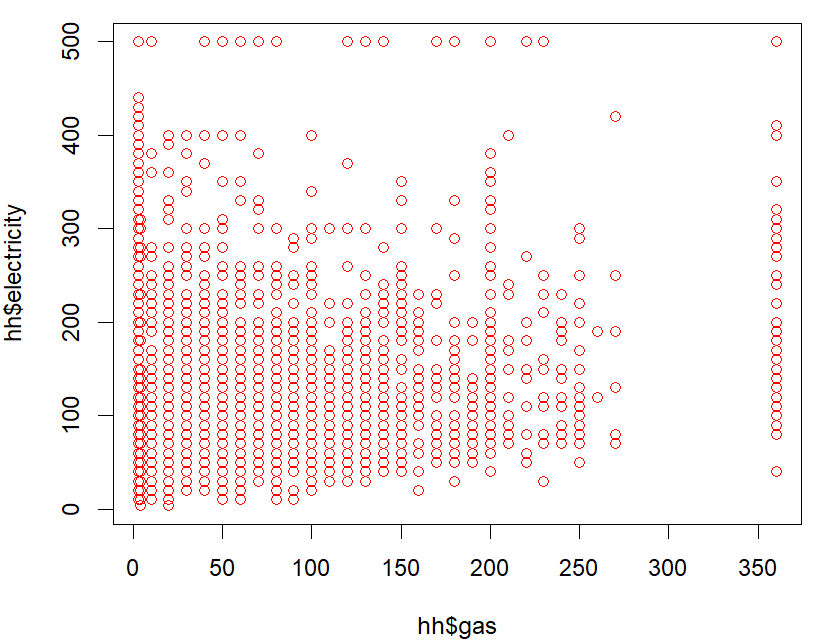
# Answer: In this data set, there's evidence to prove there's a relation between the bedrooms and number of children.

plot(x = hh$age\_husband , y = hh$age\_wife, type = 'p', col="red")



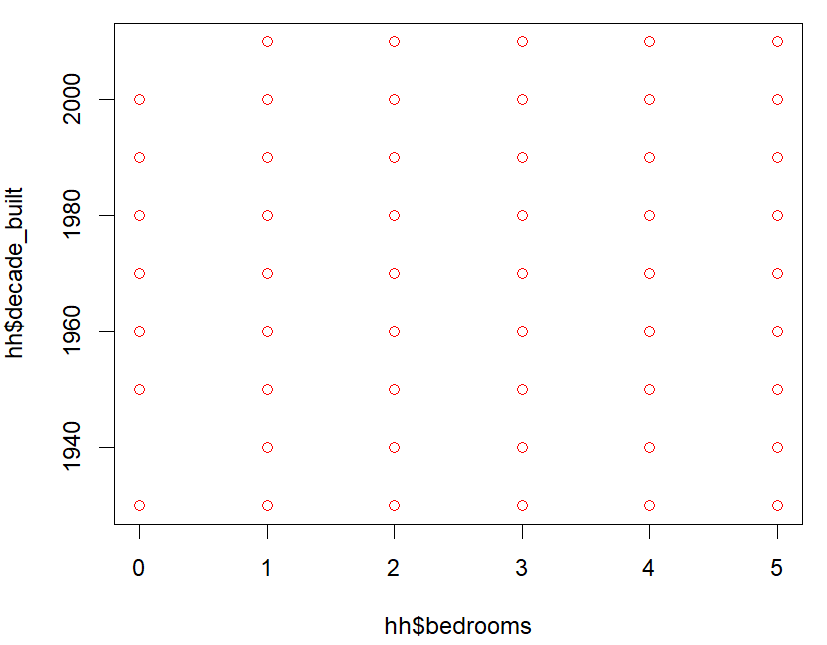
# Answer: In this data set, there's evidence to prove there's a relation between the husband’s age and wife’s age.

plot(x = hh$gas , y = hh$electricity, type = 'p', col="red")



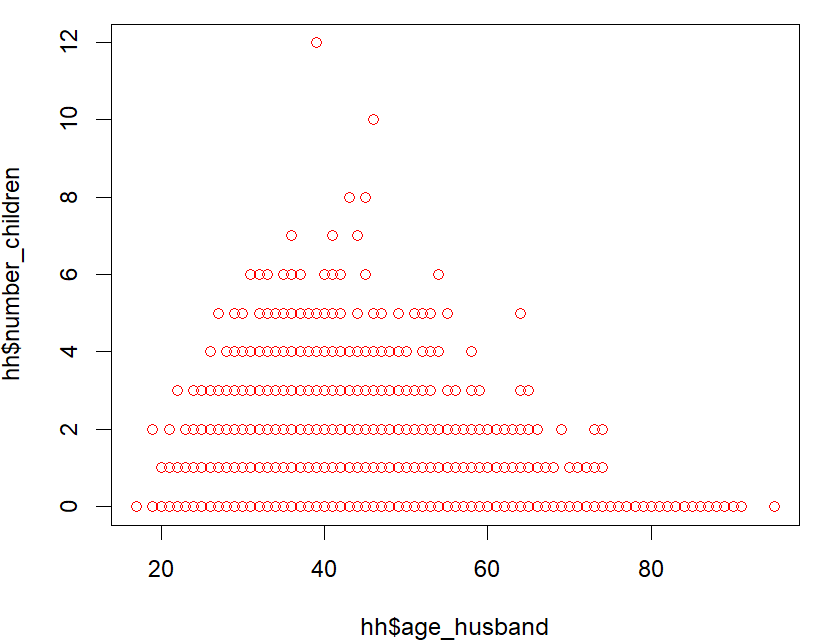
# Answer: In this data set, there's NO evidence to prove there's a relation between the gas expense and electricity expense.

plot(x = hh$bedrooms , y = hh$decade\_built, type = 'p', col="red")



# Answer: In this data set, there's evidence to prove there's a relation between the bedrooms and decade of built year.

plot(x = hh$age\_husband , y = hh$number\_children, type = 'p', col="red")



# Answer: In this data set, there's evidence to prove there's a relation between the husband's age and the number of children.

Question 9

Is there any relationship between wife's age and Internet availability(Yes/No)?

*# 9.1: Summary grouped by Internet(Yes/No)*

*agg1 <- aggregate(age\_wife ~ internet, hh , mean)*

*agg1*

##

internet age\_wife

1 No 60.27103

2 Yes 51.34565

*# 9.2: Visualization by qplot*

*library(ggplot2)*

*qplot(internet,*

*age\_wife,*

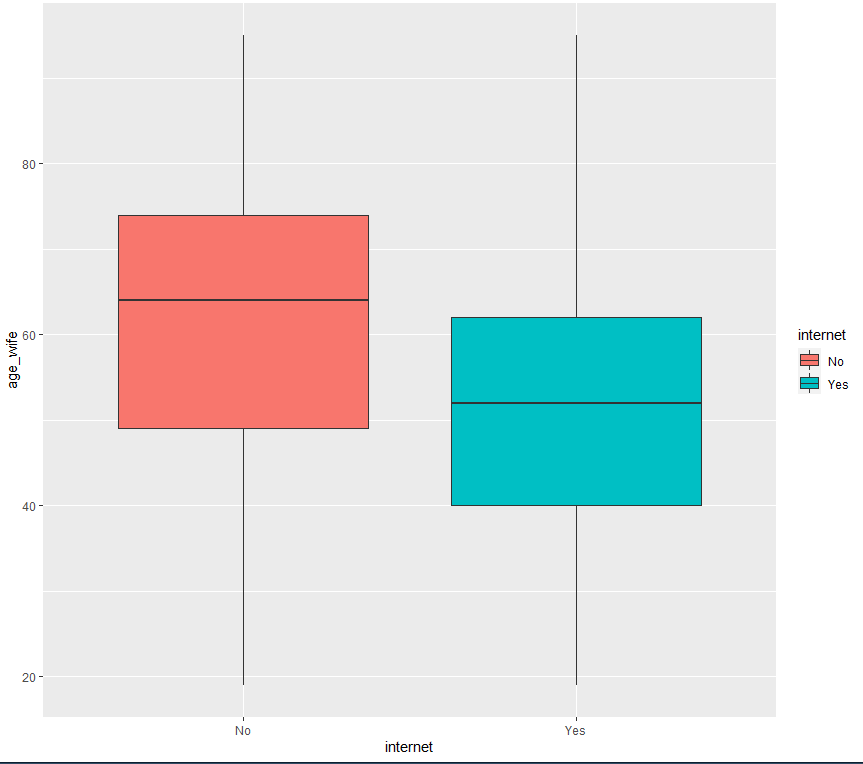
*data = hh,*

*geom="boxplot",*

*fill = internet)*

*library(ggplot2)*

*# Conclusion: the younger the age, the more internet access*



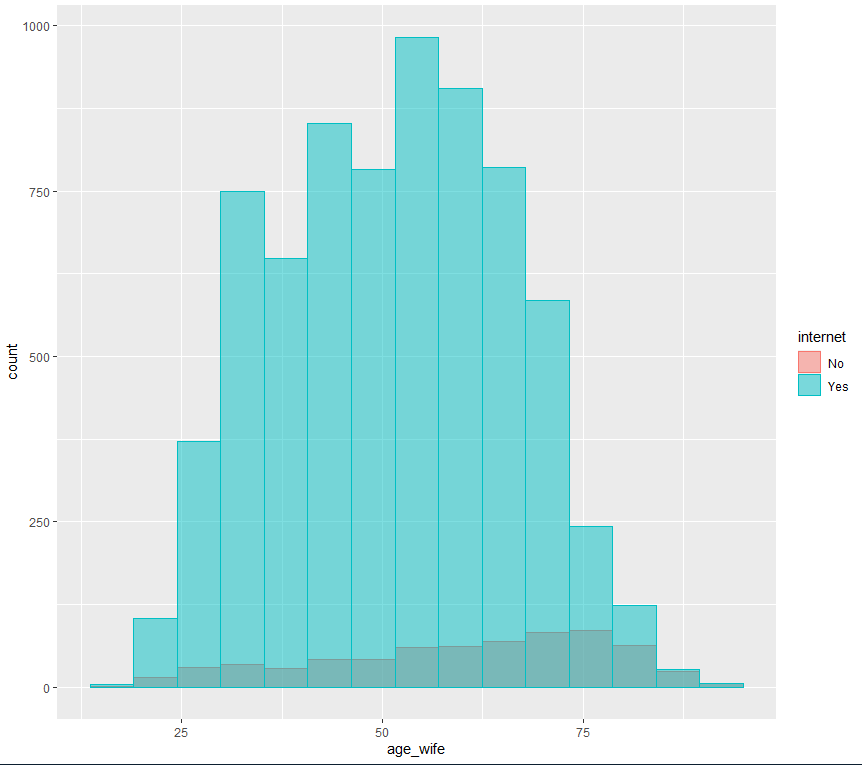
*# 9.3: Changing histogram plot fill colors by internet and usinging semi-transparent fill*

*p <- ggplot(hh,aes(x=age\_wife, fill=internet, color=internet)) +*

*geom\_histogram(position="identity", bins=15, alpha=0.5)*

*# bins: https://stackoverflow.com/questions/34774120/set-number-of-bins-for-histogram-directly-in-ggplot*

*p*



*# 9.4: Add mean lines*

*library(plyr)*

*mu <- ddply(hh, "internet", summarise, grp.mean=mean(age\_wife,na.rm=T))*

*head(mu)*

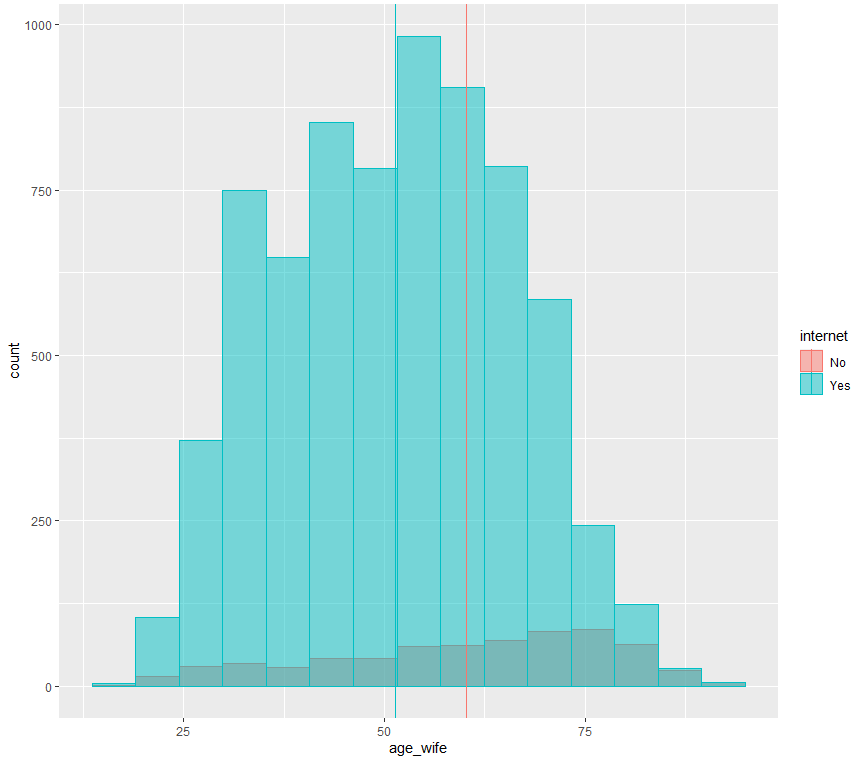
*p <- p + geom\_vline(data=mu, aes(xintercept=grp.mean, color=internet),linetype="solid")*

*p*

internet grp.mean

1 No 60.27103

2 Yes 51.34565



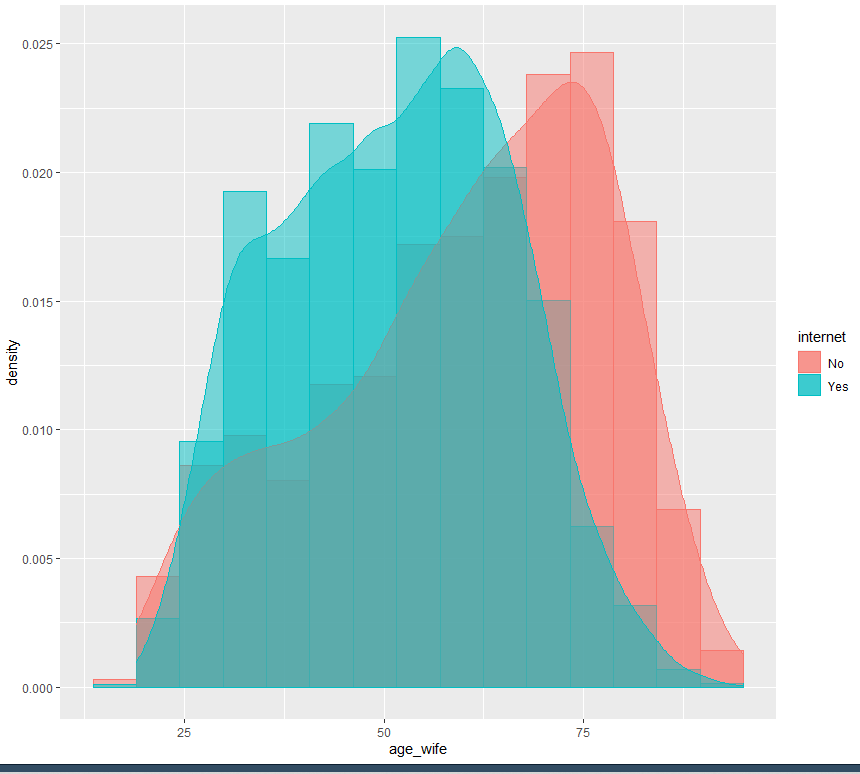
*# 9.5: Add density*

*p <- ggplot(hh, aes(x=age\_wife, fill=internet, color=internet)) +*

*geom\_histogram(aes(y=..density..),bins=15, position="identity", alpha=0.5)+*

*geom\_density(alpha=0.5)*

*p*

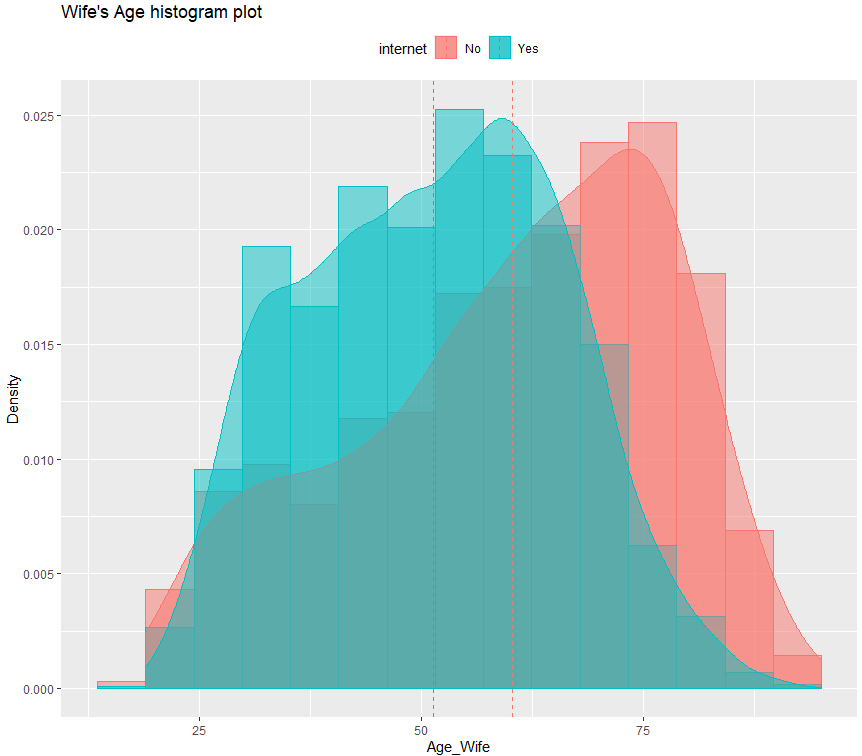


*# 9.6: Add mean lines and Change the legend position*

*p + geom\_vline(data=mu, aes(xintercept=grp.mean, color=internet),linetype="dashed")+*

*theme(legend.position="top")+*

*labs(title="Wife's Age histogram plot", x="Age\_Wife", y = "Density")*



*# 9.7 t-test*

*# Yes: House has internet, No: House has no internet*

*# Null Hypothesis: µYes = µNo (the means of both populations are equal)*

*# Alternate Hypothesis: µYes <> µNo (the means of both populations are not equal)*

*t.test(age\_wife ~ internet, data=hh )*

*# Conclusion: p-value is less than 0.05, so there is association between wife's age and internet at 5% significant level*

Welch Two Sample t-test

data: age\_wife by internet

t = 12.603, df = 721.29, p-value < 0.00000000000000022

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

7.534967 10.315779

sample estimates:

mean in group No mean in group Yes

60.27103 51.34565

*# Conclusion: p-value is less than 0.05, so the mean values between uYes and uNo are not equal.*

Question 10

Is there any relation between ownership and target(income\_group)?

*# 10.1 - Visualization: Stacked Bar Plot*

*tbl <- table(hh$own,hh$income\_group)*

*tbl*

*counts <- tbl[1:4,1:3]*

*counts*

*barplot(counts,*

*main = "Ownership Vs. Income Group",*

*xlab = "Income Group",*

*col = c("black","red", "yellow", "green"),*

*legend = rownames(counts),*

*args.legend = list(x ='topleft', bty='n', inset=c(0,-0.1)))*

*# Legend position:* [*https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r*](https://stackoverflow.com/questions/27688754/bar-chart-legend-position-avoiding-operlap-in-r)

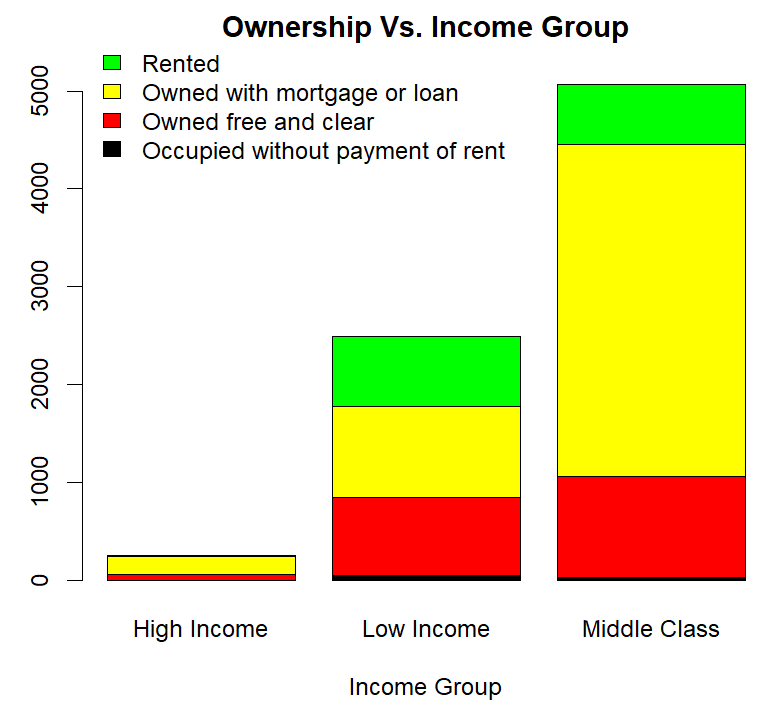
High Income Low Income Middle Class

Occupied without payment of rent 0 49 27

Owned free and clear 61 799 1036

Owned with mortgage or loan 180 932 3393

Rented 9 709 616



*# 10.2 - Summarization: Contingency Table*

*add <- addmargins(xtabs(~ own + income\_group,data=hh))*

*add*

*add[1:5,1:4]*

*proportions(xtabs(~ own + income\_group,data=hh))[1:4,1:3]*

income\_group

own High Income Low Income Middle Class

Occupied without payment of rent 0.000000000 0.006273204 0.003456664

Owned free and clear 0.007809499 0.102291640 0.132633466

Owned with mortgage or loan 0.023044425 0.119318909 0.434387402

Rented 0.001152221 0.090769428 0.078863142

*# 10.3 - Indipendency: Chi-square Test*

*# 10.3.1 Problem:*

*# Test the hypothesis whether the ownership is independent of the income group at .05 significance level.*

*# Null hypothesis: Ownership is independent of Income Group*

*# 10.3.2 Solution:*

*# p-value*

*library(MASS)*

*tbl <- table(hh$own,hh$income\_group)*

*tbl*

*chisq.test(tbl) # the p-value < 2.2e-16*

data: tbl

X-squared = 680.48, df = 6, p-value < 0.00000000000000022

Warning message:

In chisq.test(tbl) : Chi-squared approximation may be incorrect

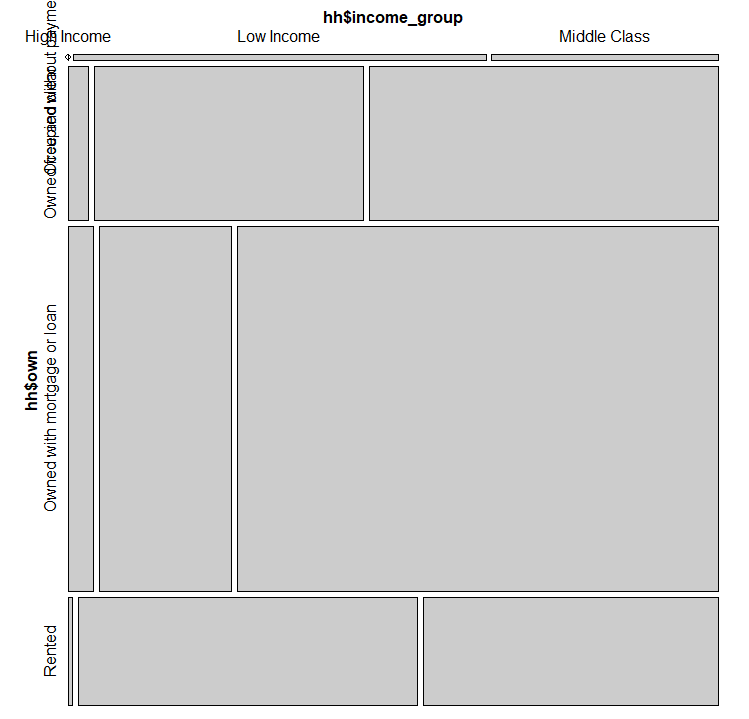
*# Mosaic Plots*

*library(vcd)*

*library(grid)*

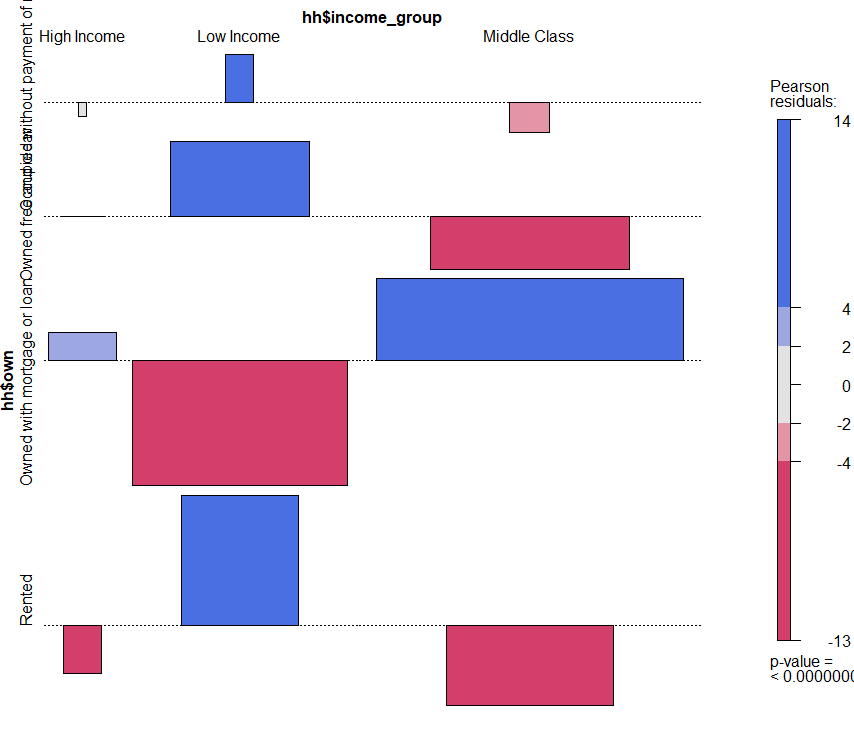
*mosaic(structable(hh$income\_group ~ hh$own))*

*# structable:* [*https://stackoverflow.com/questions/14547162/missing-value-where-true-false-needed-error-vcdmosaic*](https://stackoverflow.com/questions/14547162/missing-value-where-true-false-needed-error-vcdmosaic)



*# Association Plots*

*assoc(hh$income\_group ~ hh$own, shade=TRUE)*



**# 10.3.3 Conclusion:**

As the p-value 2.2e-16 is less than the 0.05 significance level, we **reject** the null hypothesis that Communication Mode is independent of the Income\_Group and conclude that in our data, the 'own' and the 'income\_group' are statistically significantly associated (p-value = 0).