Project in R : Household

By: Peng Wang
Instructor: Mr. Hamid Rajaee
Metro College Toronto
March 5, 2021

Project Scope

Source Data:

http://stat511.cwick.co.nz/homeworks/acs_or.csv

- 7811 household records contains 15 variables
- husband and wife's income,
- number of bedrooms and children,
- age of the house, household expenses, and
- the availability of internet access etc.

Project Objective:

Review and visualize the relationship between each factor.

Industry Oriented:

- Real Estate / Estate Agent
- Mortgage Provider
- Property Tax / City Hall / CRA

Data Structure

```
> 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 55050 42000 ...
 $ income_wife    : int  29200 3100 0 8800 26000 30000 4800 11000 600 36000 ...
 $ bedrooms       : int  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 2000 ...
 $ income_total   : int  40200 103100 31000 60500 42600 107500 13200 84670 55650 78000 ...
 $ income_group   : chr  "Low Income" "Middle Class" "Low Income" "Middle Class" ...
```

Raw data was clean, but I still gone through and coded the each data clearing process for practice purpose.

shape/null/duplicate/add/drop/copy etc.

Q1

What is the distribution of variable 'income_group'?

What is the distribution of variable 'income_group'?

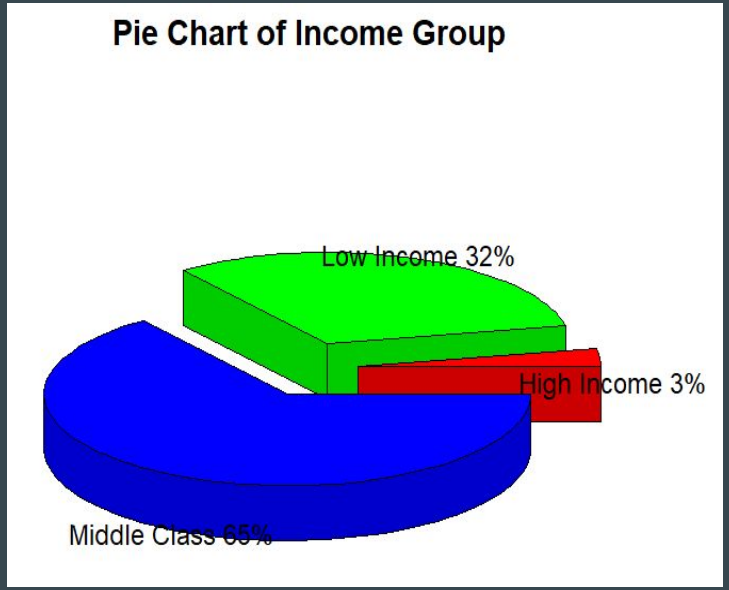
Add Columns ‘Income Group’, to facilitate analysis

```r
# 1.1: Add the 1st column - income_total

hh$income_total <- hh$income_husband + hh$income_wife
View(hh)
summary(hh$income_total)

# 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", "")))
View(hh)
str(hh)
```

What is the distribution of variable 'income_group'?

3D Pie Chart

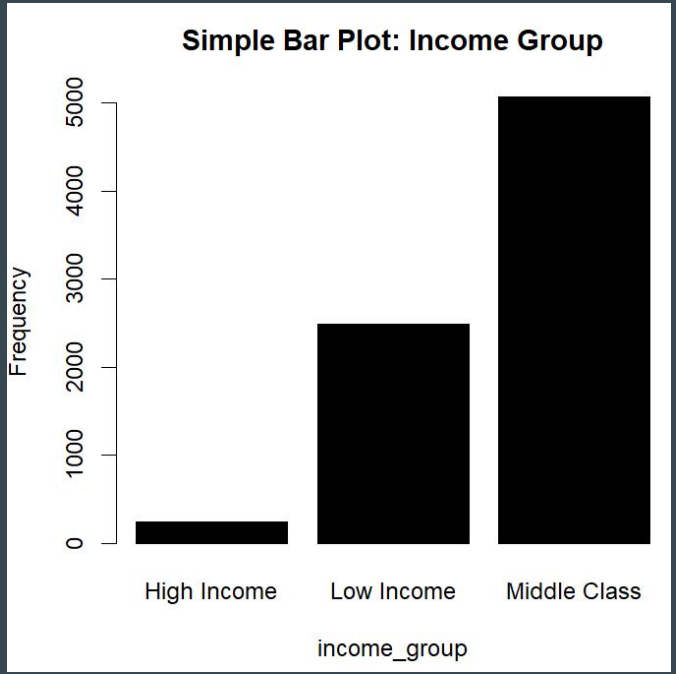
```
install.packages('plotrix')
library(plotrix)
count <- table(hh$income_group)
pct <- round(count/sum(count)*100)
lbls <- c("High Income", "Low Income", "Middle Class")
lbls <- paste(lbls, pct) # add pct to label
lbls <- paste(lbls, "%", sep = "") # add % to pct
pie <- pie3D(count,
             explode=0.2,
             main = "Pie Chart of Income Group")
pie3D.labels(pie, labels = lbls)
```



What is the distribution of variable 'income_group'?

Simple Bar Plot

```r
counts <- table(hh$income_group)
counts
barplot(counts,
        main = "Simple Bar Plot: Income Group",
        xlab = "income_group",
        ylab = "Frequency",
        col = 'black',
        horiz = FALSE)
```



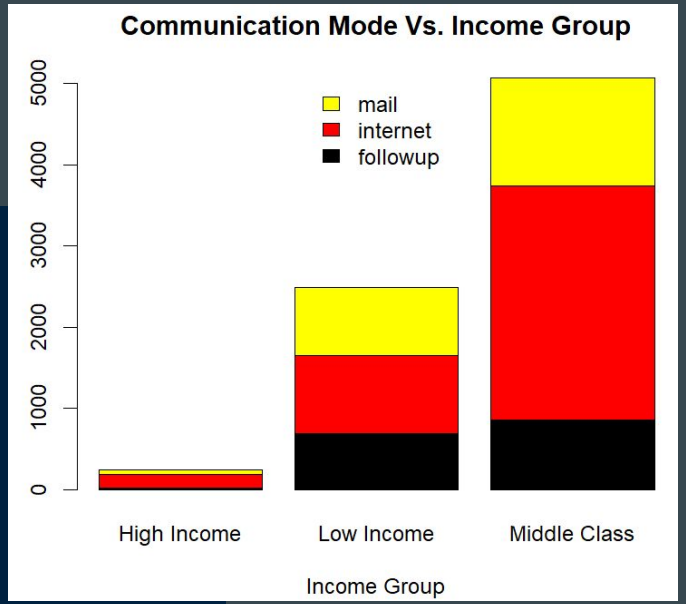
Q2

Is there any relation between communication mode and target(income_group)?

Relation between communication mode vs. income_group

Bar plot

```r
# 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
```



Relation between communication mode vs. income_group

Chi-Sq Test & Conclusion

```
# p-value
library(MASS)
tbl <- table(hh$mode,hh$income_group)
tbl
chisq.test(tbl) # the p-value < 2.2e-16
```

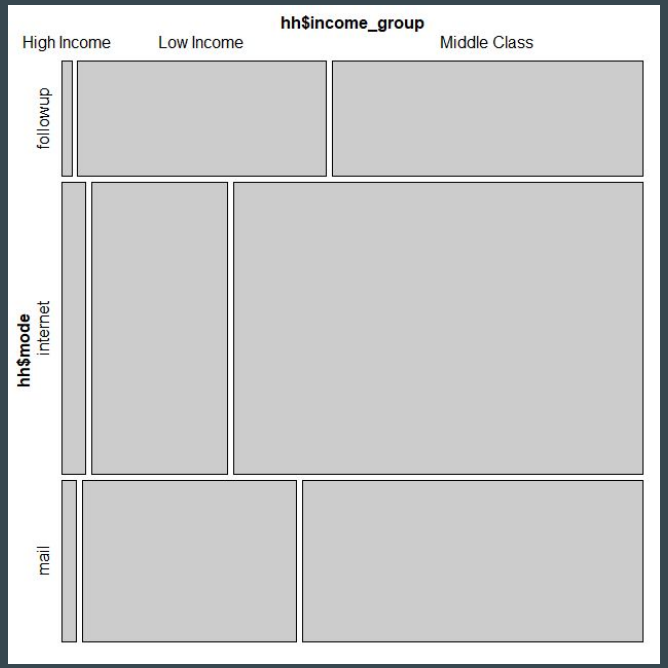
```
> chisq.test(tbl) # the p-value < 2.2e-16

        Pearson's Chi-squared test

data:  tbl
X-squared = 261.59, df = 4, p-value < 0.00000000000000022
```

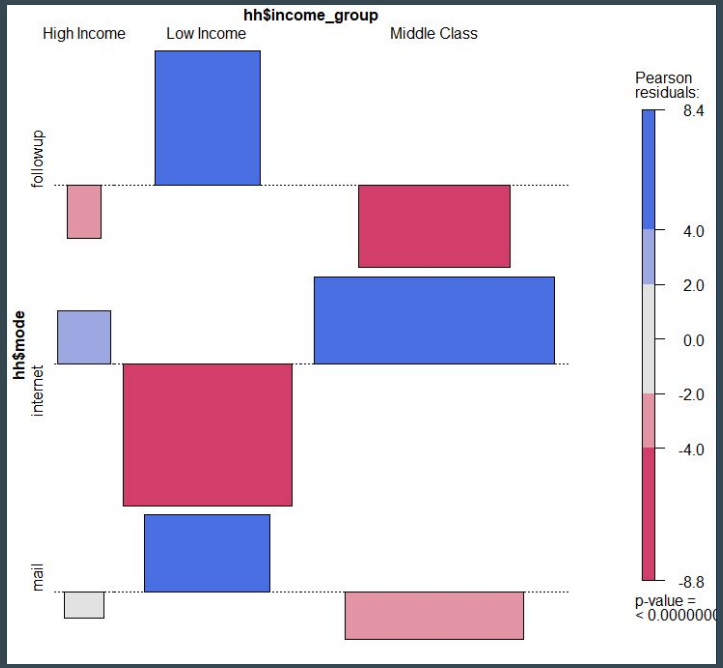
As the p-value 2.2e-16 is less than the 0.05 significance level, reject the null hypothesis that Communication Mode is independent of the Income_Group.

The 'mode' and the 'income_group' are statistically significantly associated.

Relation between communication mode vs. income_group

Mosaic & Association Plots





Q3

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

Spot outliers

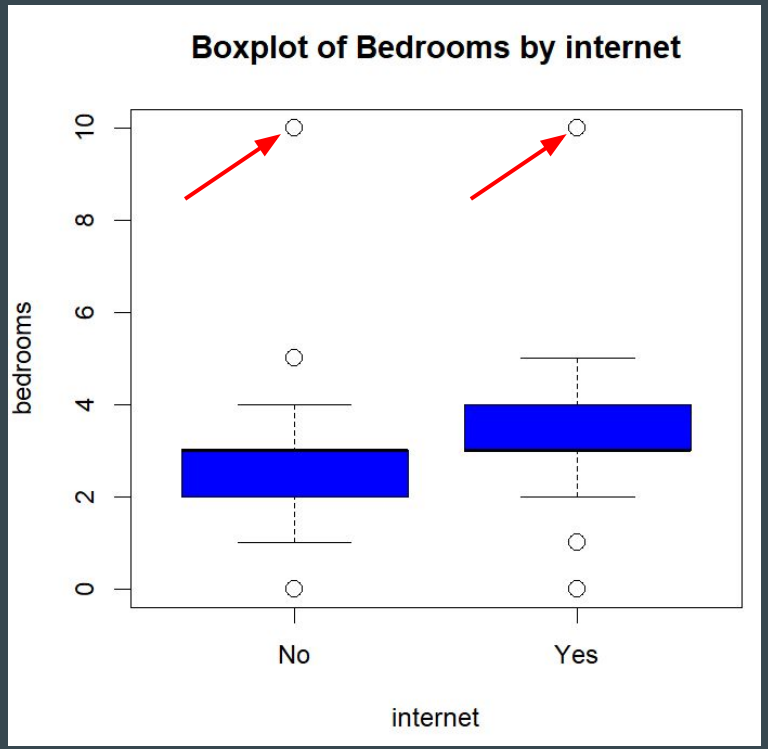
Histogram

```r
# 4.2 histogram
hist(hh$bedrooms,
     breaks = 8,
     main = "bedrooms",
     col = "blue",
     xlab = "bedrooms",
     ylab = "Frequency")
```

Spot the outlier

Boxplot of Bedrooms by internet

```r
# 4.3 Boxplot of Bedrooms by internet
boxplot(bedrooms ~ internet,
        data = hh,
        main = "Boxplot of Bedrooms by internet",
        xlab = "internet",
        ylab = "bedrooms",
        col = "blue")
```



Handle the outlier.

Find outlier's pattern, and PROVE it

```
# 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 y
summary(bed_out["bedrooms"])
nrow(bed_out)

# 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
```

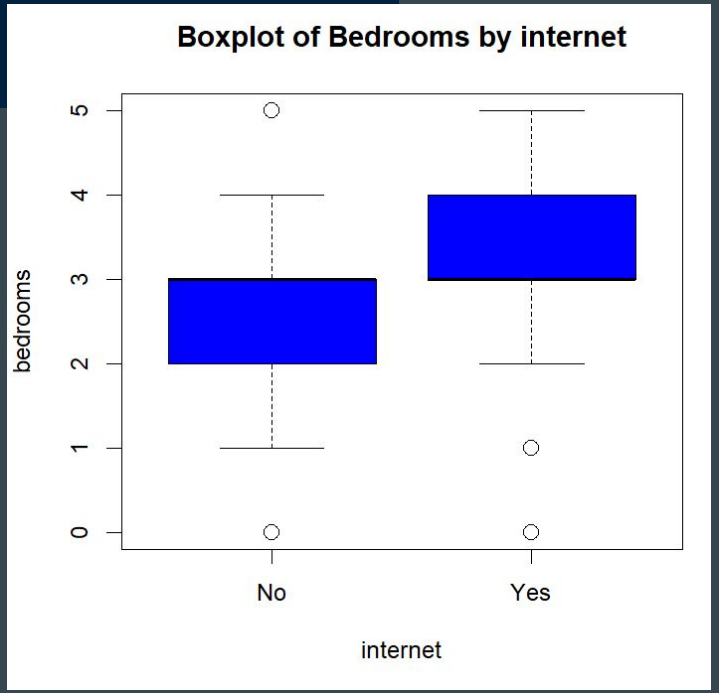
Handle the outlier - BEFORE

```
> # 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 year
s...
 [1] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
[28] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
[55] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
> summary(bed_out["bedrooms"])
    bedrooms
 Min.   :10
 1st Qu.:10
 Median :10
 Mean   :10
 3rd Qu.:10
 Max.   :10
> nrow(bed_out)
[1] 72
>
```

Handle the outlier - AFTER

```r
# 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.
# copy file
hh4 <- hh
View(hh4)
```

```
> 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
```



Q4

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

Relationship between Bedrooms and Internet(Yes/No)

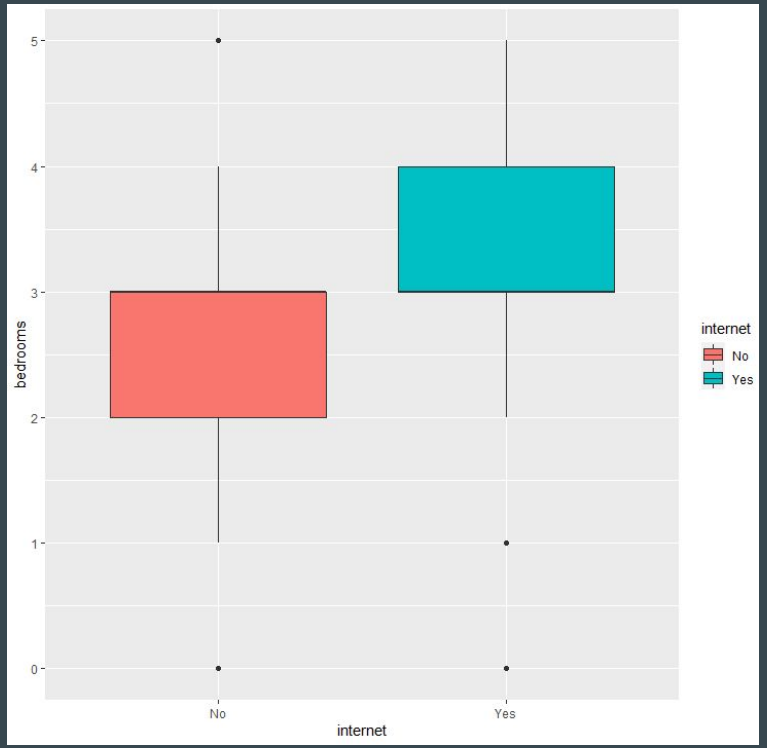
```
# 5.1: Summary grouped by Internet(Yes/No)
agg1 <- aggregate(bedrooms ~ internet, hh , mean)
agg1
```

```
> # 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
```

Relationship between Bedrooms and Internet(Yes/No)

qplot

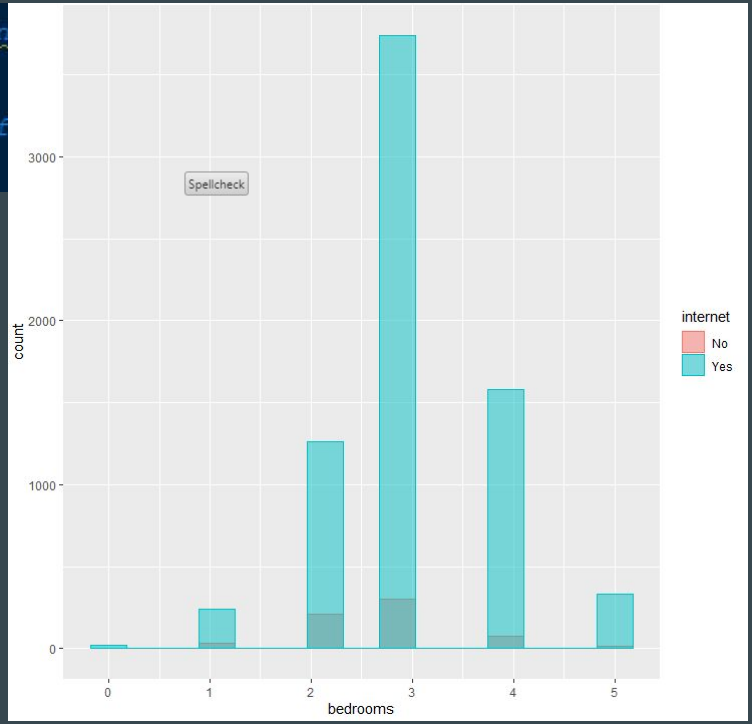
```
# 5.2: Visualization by qplot
library(ggplot2)
qplot(internet,
      bedrooms,
      data = hh,
      geom="boxplot",
      fill = internet)
```



Relationship between Bedrooms and Internet(Yes/No)

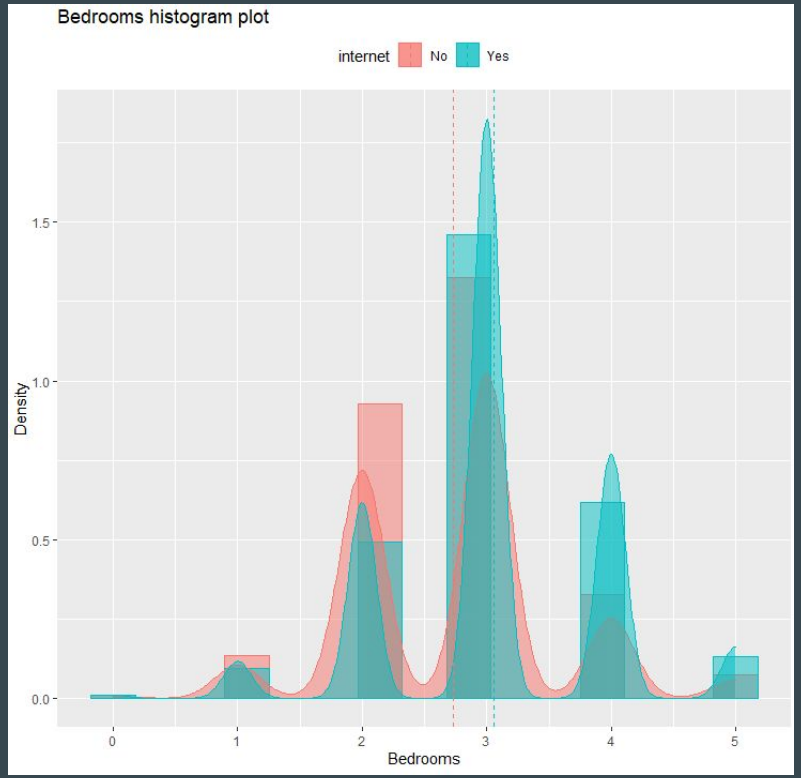
Histogram

```
# 5.3: Changing histogram plot fill colors by internet and usingin
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
p
```



Relationship between Bedrooms and Internet(Yes/No)

Histogram - after adding mean line and density line.



Relationship between Bedrooms vs. Internet(Yes/No)

T-Test

```
> 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
```

Relationship between Bedrooms and Internet(Yes/No)

Hypothesis -

- 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 conclusion -

p-value is less than 0.05, the mean values between uYes and uNo are not equal. Null hypothesis rejected.

Thank you!