Statistics 501 final Project

Debabrata Halder, Piusha Gullapalli, Snehil Verma

2022-12-25

Contents

1 Introduction and data Background

This data was extracted by Barry Becker from the 1994 Census database.

The data was extracted to be used for a prediction task to determine whether a person makes over 50K a year.

Conversion of original data as follows:

1. Discretized agrossincome into two ranges with threshold 50,000.
2. Convert U.S. to US to avoid periods.

3. Convert Unknown to "?"
4. Run MLC++ GenCVFiles to generate data,test.

Description of fnlwgt (final weight):

The weights on the CPS files are controlled to independent estimates of the civilian non institutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau.

1.1 Attribute Information

Parameters -

age: the age of an individual

workclass: a general term to represent the employment status of an individual

fnlwgt: final weight. This is the number of people the census believes the entry represents.

education: the highest level of education achieved by an individual.

education_num: the highest level of education achieved in numerical form.

marital_status: marital status of an individual.

occupation: the general type of occupation of an individual

relationship: represents what this individual is relative to others.

race: Descriptions of an individual's race

sex: the sex of the individual

capital_gain: capital gains for an individual

capital_loss: capital loss for an individual

hours_per_week: the hours an individual has reported to work per week

native_country: country of origin for an individual

NOTE: Some values in the dataset is marked as "?". It means the value is unknown.

1.2 Loading the data

```
adult <- read.table("adult.data", sep = ",")
colnames(adult) <- c("age", "workclass", "fnlwgt", "education", "education_num", "marital_status", "occu
```

```
summary(adult)
```

```
##       age         workclass             fnlwgt          education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education_num   marital_status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
```

```
##  1st Qu.: 9.00   Class :character   Class :character   Class :character  
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :10.08                                                           
##  3rd Qu.:12.00                                                           
##  Max.   :16.00                                                           
##      race               sex             capital_gain    capital_loss  
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0  
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0  
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0  
##                                        Mean   : 1078   Mean   :  87.3  
##                                        3rd Qu.:    0   3rd Qu.:   0.0  
##                                        Max.   :99999   Max.   :4356.0  
##  hours_per_week  native_country       fifty_k         
##  Min.   : 1.00   Length:32561       Length:32561      
##  1st Qu.:40.00   Class :character   Class :character  
##  Median :40.00   Mode  :character   Mode  :character  
##  Mean   :40.44                                        
##  3rd Qu.:45.00                                        
##  Max.   :99.00
```

2 Testing realtionship between capital gain and sex

Motivation: we want to find out if the capital gain differs based on sex.

2.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.
4. And the final assumptions is homogeneity of variance.

2.2 Hypothesis

H0: capital gain is equal for both gender

Ha: capital gain is not equal.

```r
# adult %>%
#   group_by(sex) %>%
#   summarise(record_count = n())

female <- filter(adult, str_detect(sex, 'Female'))
male <- filter(adult, str_detect(sex, 'Male'))

t.test(capital_gain ~ sex, data=adult) # Unpooled
```

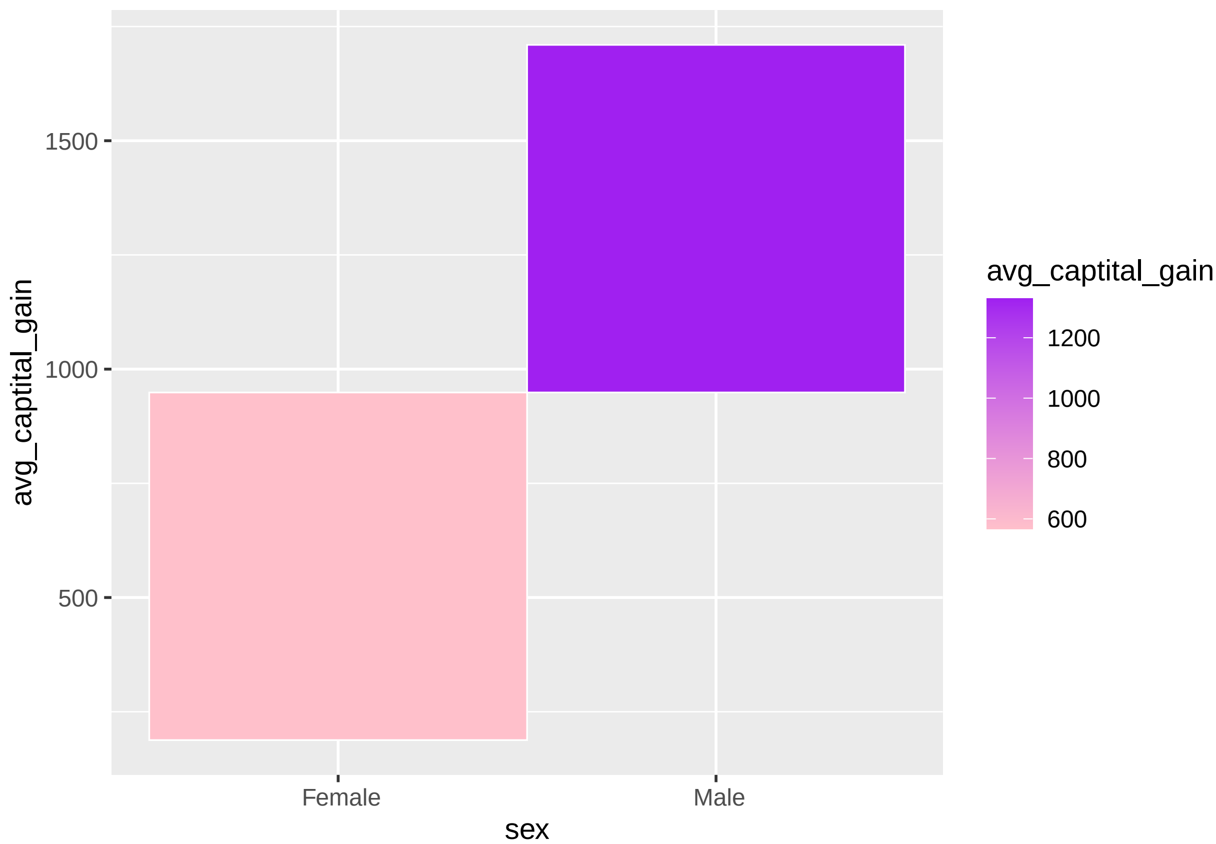
```
## 
##  Welch Two Sample t-test
## 
```

```
## data:  capital_gain by sex
## t = -10.324, df = 31563, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  Female and group  Male is not equal t
## 95 percent confidence interval:
##  -905.4303 -616.4888
## sample estimates:
## mean in group  Female   mean in group  Male
##              568.4105             1329.3701
```

```
t.test(capital_gain ~ sex, var.equal=TRUE, data=adult)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  capital_gain by sex
## t = -8.758, df = 32559, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  Female and group  Male is not equal t
## 95 percent confidence interval:
##  -931.2616 -590.6575
## sample estimates:
## mean in group  Female   mean in group  Male
##              568.4105             1329.3701
```

```
gain_sex<-adult %>%
  group_by(sex) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_sex %>%
  ggplot(aes(x=sex, y=avg_captital_gain,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="pink",high="purple")
```



2.3 Conclusion

Looking at the p value which is close to 0, we can reject the null hypothesis.

We have evidence that suggests that the true difference in means between group Female and group Male is not equal to 0.

We have evidence to say that there is a difference in the average capital gain of Male and Female

```
t.test(capital_loss ~ sex, data=adult) # Unpooled
```

```
##
##  Welch Two Sample t-test
##
## data:  capital_loss by sex
## t = -8.8911, df = 26312, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  Female and group  Male is not equal t
## 95 percent confidence interval:
##  -47.62897 -30.42238
## sample estimates:
## mean in group  Female   mean in group  Male
##              61.18763            100.21331
```

```
t.test(capital_loss ~ sex, var.equal=TRUE, data=adult)  # Pooled
```

```
##
##  Two Sample t-test
##
## data:  capital_loss by sex
## t = -8.2308, df = 32559, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  Female and group  Male is not equal 
## 95 percent confidence interval:
##  -48.31906 -29.73229
## sample estimates:
## mean in group  Female   mean in group  Male
##              61.18763             100.21331
```

3 Testing relationship between capital gain and race

Motivation: we want to find out if the capital gain differs based on race.

3.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.
4. And the final assumptions is homogeneity of variance.

3.2 Hypothesis

H0: capital gain is equal for all race

Ha: there exist a pair of race for which capital gain is not equal.

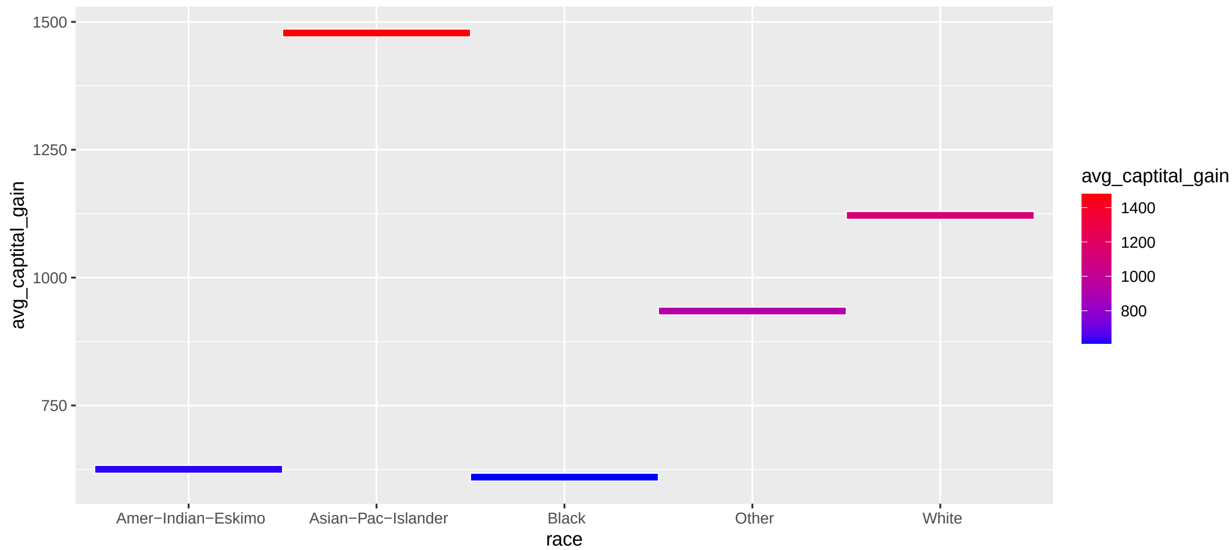
```
# adult %>%
#   group_by(race) %>%
#   summarise(record_count = n())

anov_race <- aov(capital_gain ~ race, data = adult)
summary(anov_race)
```

```
##                Df    Sum Sq   Mean Sq F value  Pr(>F)
## race            4 9.733e+08 243318824   4.463 0.00132 **
## Residuals   32556 1.775e+12  54519345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

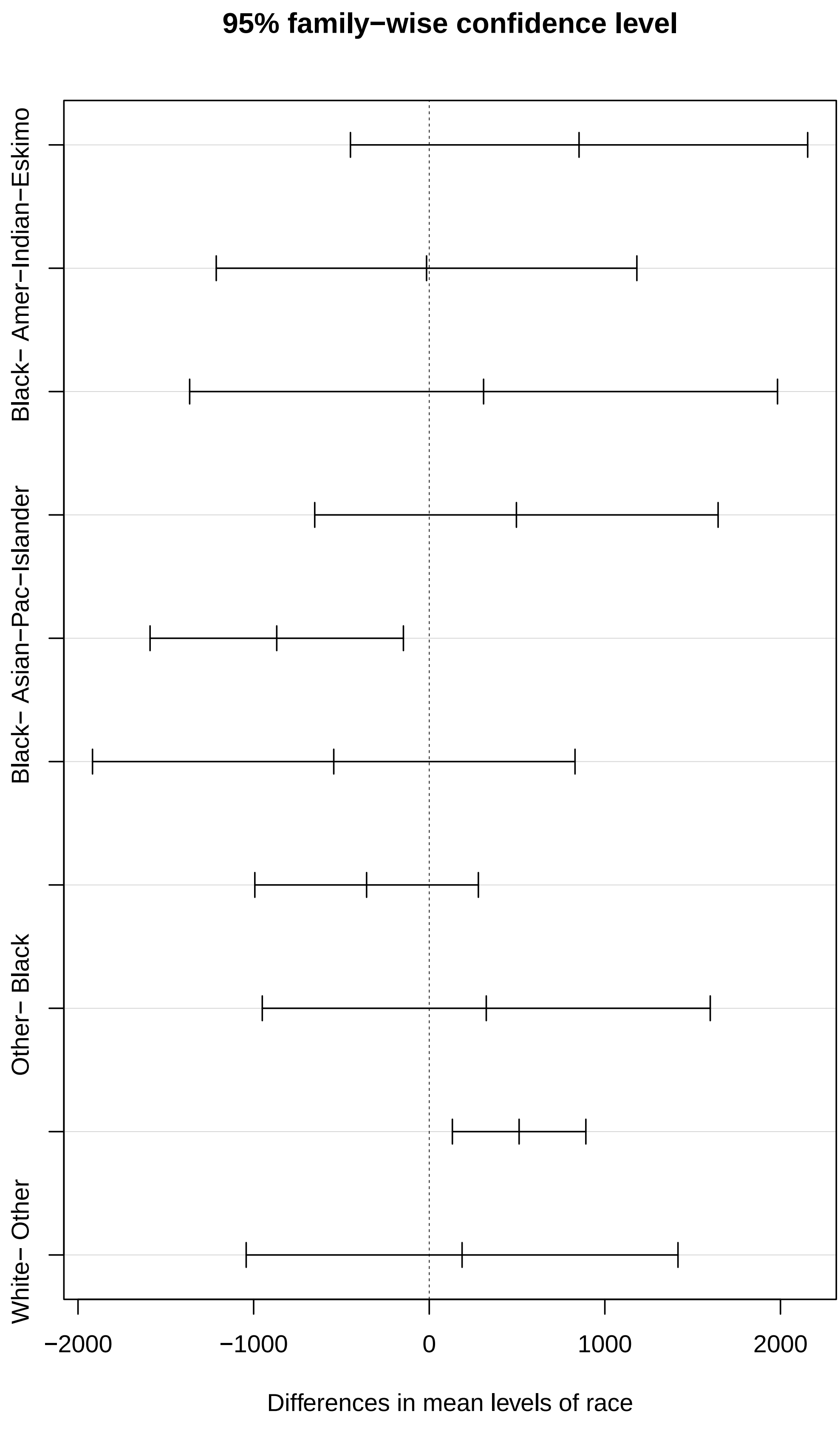
```
#TukeyHSD(anov_race)
```

```
gain_race<-adult %>%
  group_by(race) %>%
  summarize(avg_captital_gain=mean(capital_gain))
```

```r
gain_race %>%
  ggplot(aes(x=race, y=avg_captital_gain,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="blue",high="red")
```




```r
plot(TukeyHSD(aov(capital_gain ~ race, data = adult)))
```



95% family−wise confidence level
Black− Amer−Indian−Eskimo
Black− Asian−Pac−Islander
Other− Black
White− Other
−2000
−1000
0
1000
2000
Differences in mean levels of race

3.3 Conclusion

Since the p-value in our ANOVA table (0.00132) is less than .05, we have sufficient evidence to reject the null hypothesis.

This means we have sufficient evidence to say that the mean capital gain is not equal across different races.

From the Tukey Test, we can see that there is a significant difference between the means for Black- Asian-Pac-Islander and White- Black, and the p values are below the significance level.

From the plots, we can see that the maximum average capital gain is in the race Asian-Pac-Islander.

4 Testing relationship between capital gain and occupation

Motivation: we want to find out if the capital gain differs based on occupation.

4.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.
4. And the final assumptions is homogeneity of variance.

4.2 Hypothesis

H0: capital gain is equal for all occupation

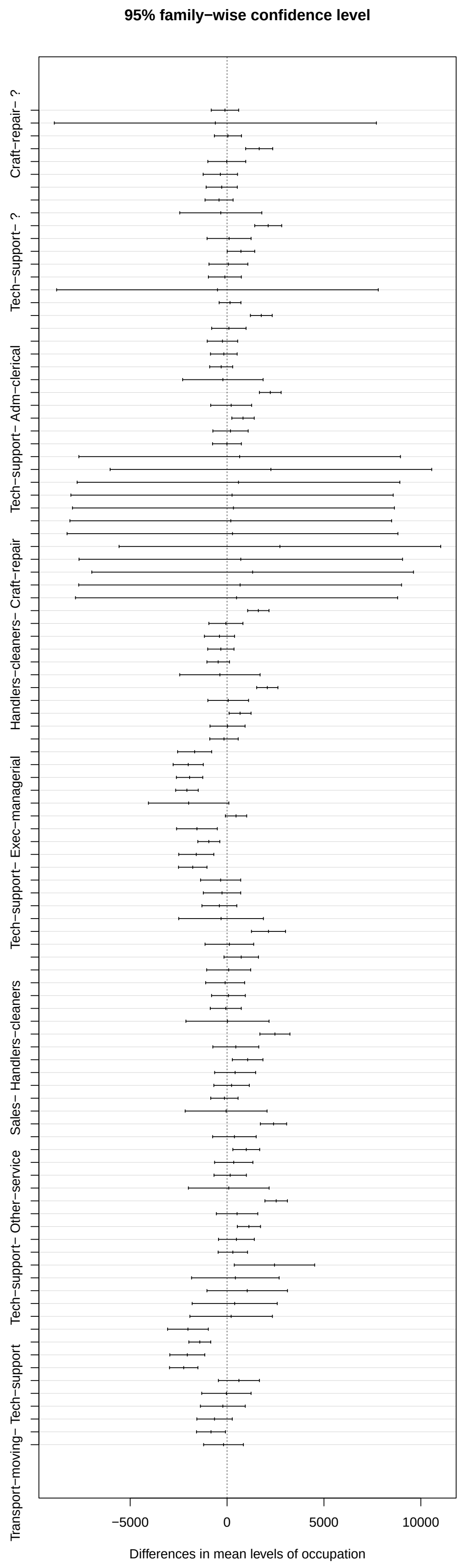
Ha: there exist a pair of occupation for which capital gain is not equal.

```
anov_occ <- aov(capital_gain ~ occupation, data = adult)
summary(anov_occ)
```

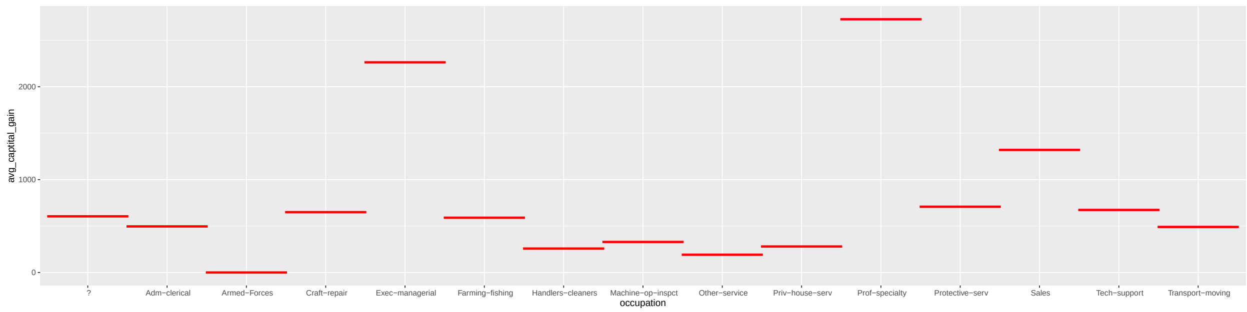
```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## occupation     14 2.539e+10 1.813e+09   33.72 <2e-16 ***
## Residuals   32546 1.751e+12 5.379e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#TukeyHSD(anov_occ)
```

```
plot(TukeyHSD(aov(capital_gain ~ occupation, data = adult)))
```


95% family−wise confidence level
Craft−repair− ?
Tech−support− ?
Tech−support− Adm−clerical
Handlers−cleaners− Craft−repair
Tech−support− Exec−managerial
Sales− Handlers−cleaners
Tech−support− Other−service
Transport−moving− Tech−support
−5000
0
5000
10000
Differences in mean levels of occupation

```
gain_occupation<-adult %>%
  group_by(occupation) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_occupation %>%
  ggplot(aes(x=occupation, y=avg_captital_gain))+
  geom_tile(color="red",size=1)
```



4.3 Conclusion

Since the p-value in our ANOVA table (10^-16) is less than .05, we have sufficient evidence to reject the null hypothesis.

This means we have sufficient evidence to say that the mean capital gain is not equal across different occupation.

From the Tukey test, we can see the p-values for different occupation pairs, and the difference in average capital gain.

From the plots, we can see that the maximum average capital gain is in the occupation of Exec-managerial.

5 Testing relationship between capital gain and workclass

Motivation: we want to find out if the capital gain differs based on workclass.

5.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.
4. And the final assumptions is homogeneity of variance.

5.2 Hypothesis

H0: capital gain is equal for all workclass

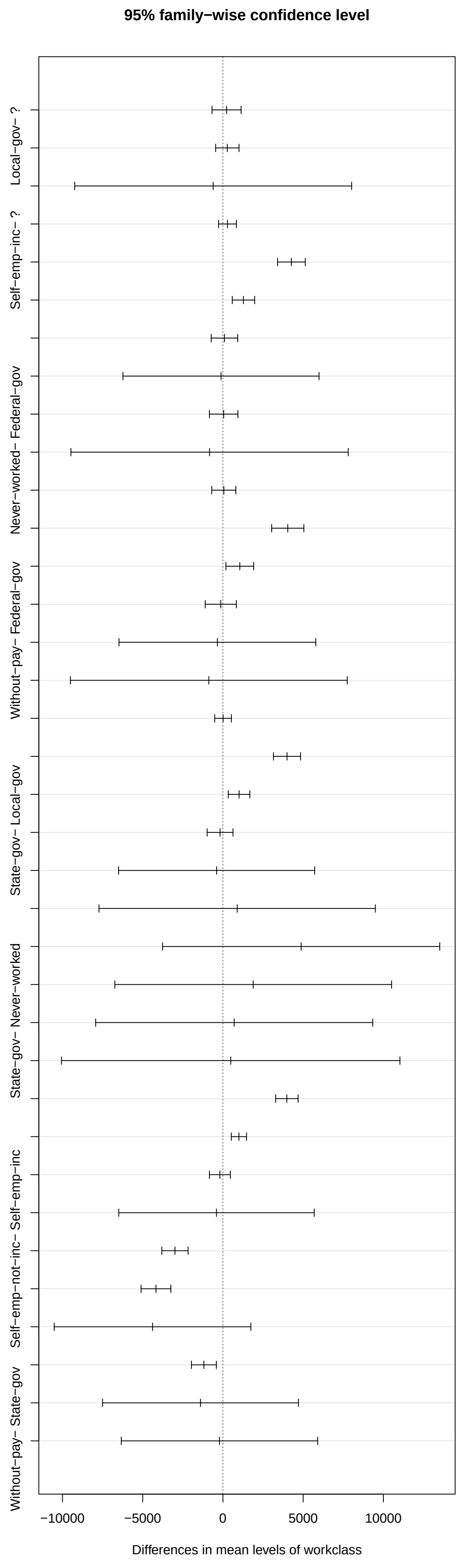
Ha: there exist a pair of workclass for which capital gain is not equal.

```r
anov_wc <- aov(capital_gain ~ workclass, data = adult)
summary(anov_wc)
```

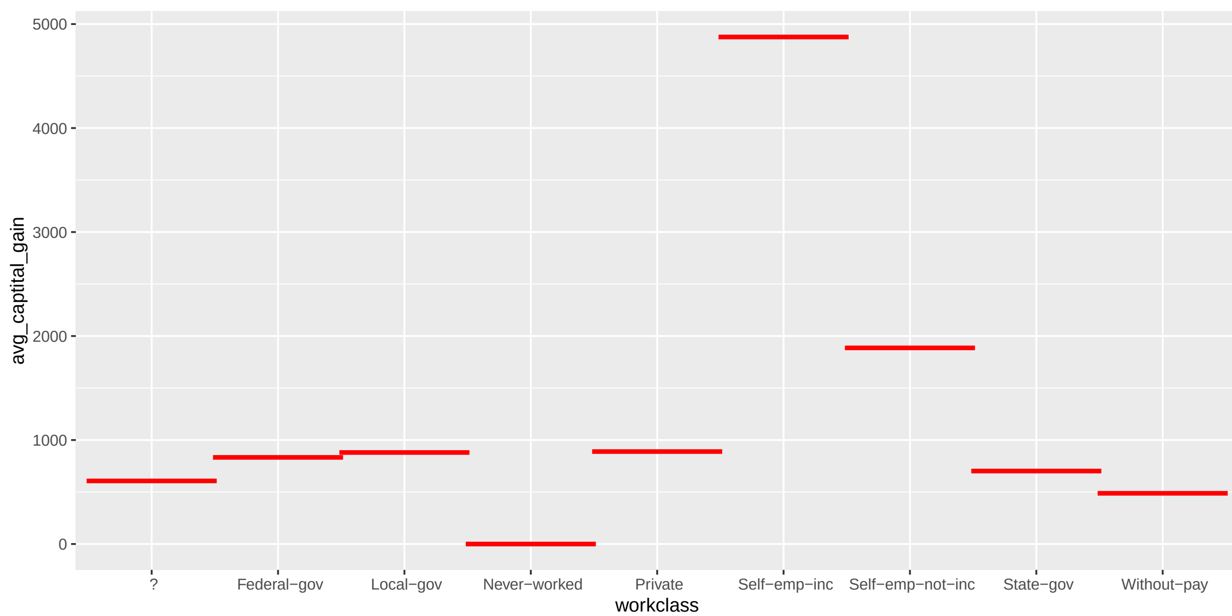
```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## workclass         8 1.931e+10 2.413e+09   44.72 <2e-16 ***
## Residuals     32552 1.757e+12 5.396e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#TukeyHSD(anov_wc)
```

```r
plot(TukeyHSD(aov(capital_gain ~ workclass, data = adult)))
```


95% family−wise confidence level
Local−gov− ?
Self−emp−inc− ?
Never−worked− Federal−gov
Without−pay− Federal−gov
State−gov− Local−gov
State−gov− Never−worked
Self−emp−not−inc− Self−emp−inc
Without−pay− State−gov
−10000
−5000
0
5000
10000
Differences in mean levels of workclass

```r
gain_wc<-adult %>%
  group_by(workclass) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_wc %>%
  ggplot(aes(x=workclass, y=avg_captital_gain))+
  geom_tile(color="red",size=1)
```



5.3 Conclusion

Since the p-value in our ANOVA table (10^-16) is less than .05, we have sufficient evidence to reject the null hypothesis.

This means we have sufficient evidence to say that the mean capital gain is not equal across different workclass.

From the Tukey test, we can see the p-values for different occupation pairs, and the difference in average capital gain.

From the plots, we can see that the maximum average capital gain is in the occupation of Self-emp-inc.

6 Testing relationship between capital gain and education level

Motivation: we want to find out if the capital gain differs based on education level.

6.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.

4. And the final assumptions is homogeneity of variance.

6.2 Hypothesis

H0: capital gain is equal for education level

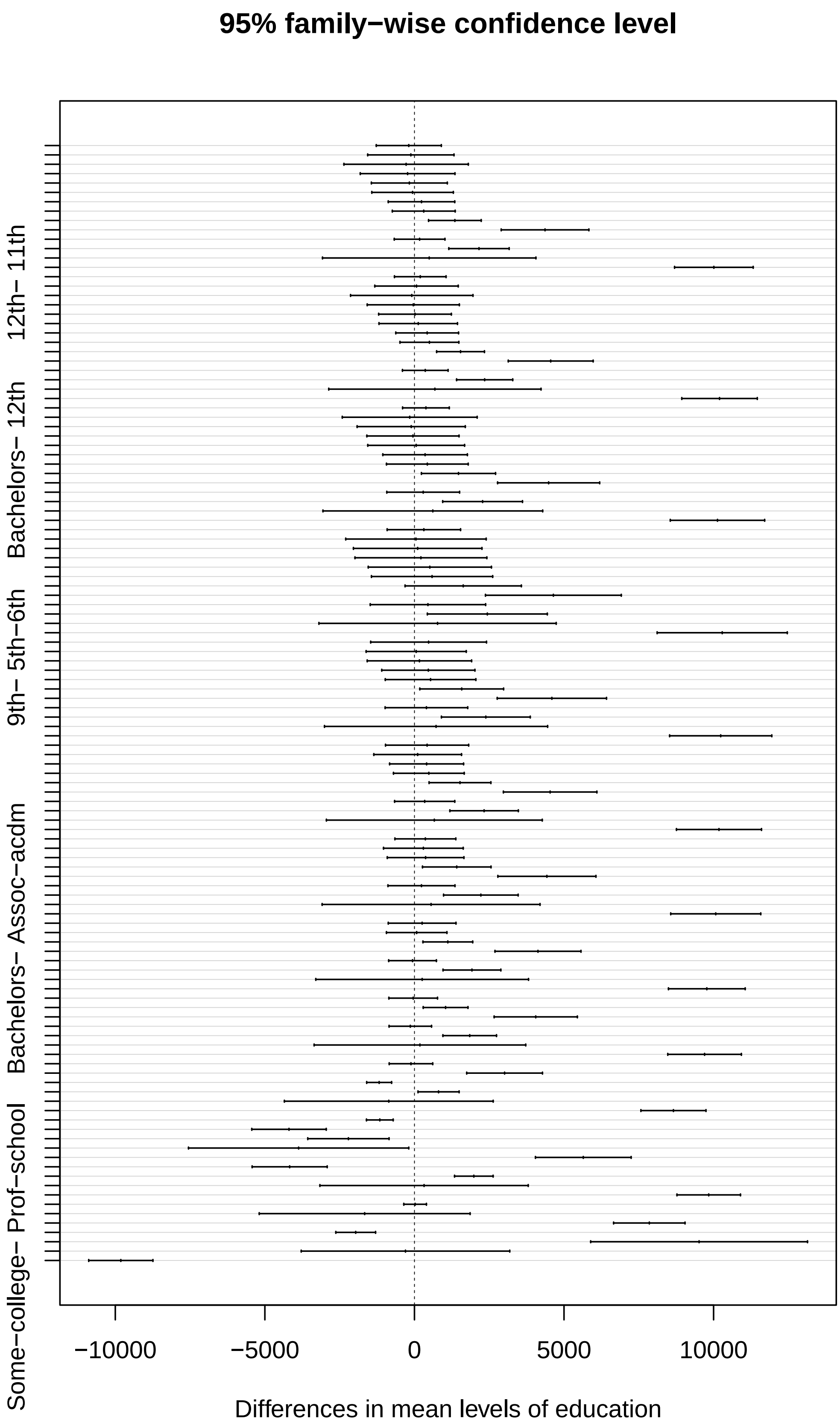
Ha: there exist a pair of education level for which capital gain is not equal.

```
# adult %>%
#   group_by(education) %>%
#   summarise(record_count = n())

anov_edu <- aov(capital_gain ~ education, data = adult)
summary(anov_edu)
```

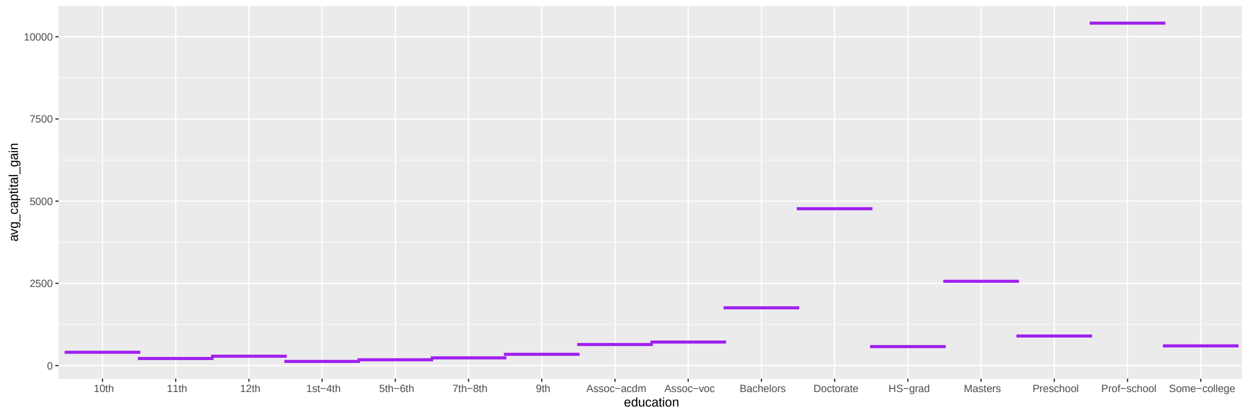
```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## education        15 6.953e+10 4.636e+09   88.41 <2e-16 ***
## Residuals     32545 1.706e+12 5.243e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#TukeyHSD(anov_edu)
```

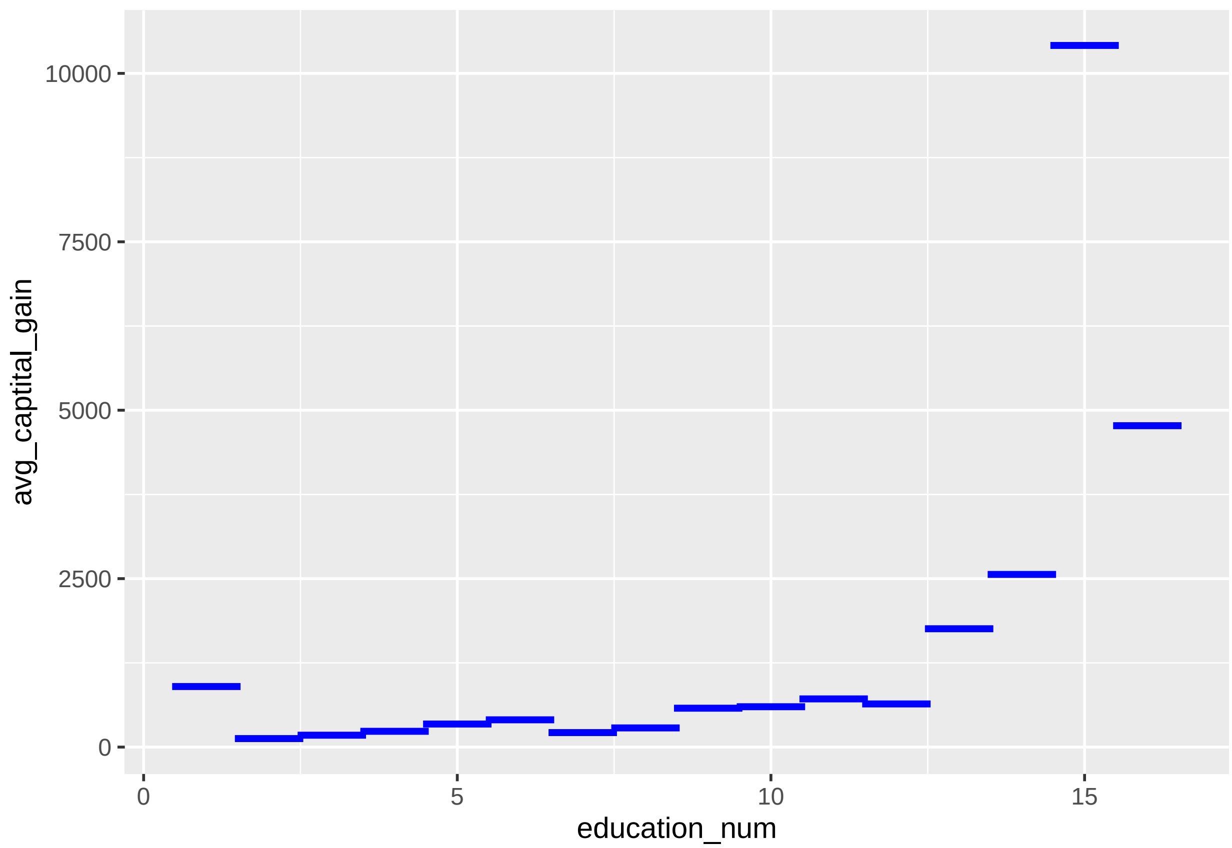
```
plot(TukeyHSD(aov(capital_gain ~ education, data = adult)))
```


95% family−wise confidence level
12th- 11th
Bachelors- 12th
9th- 5th-6th
Bachelors- Assoc-acdm
Some-college- Prof-school
−10000
−5000
0
5000
10000
Differences in mean levels of education

```
gain_edu<-adult %>%
  group_by(education) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_edu %>%
  ggplot(aes(x=education, y=avg_captital_gain))+
  geom_tile(color="purple",size=1)
```




```
#Checking for education number

anov_edu_num <- aov(capital_gain ~ education_num, data = adult)
# summary(anov_edu_num)
# anov_edu_num

gain_edu_num<-adult %>%
  group_by(education_num) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_edu_num %>%
  ggplot(aes(x=education_num, y=avg_captital_gain))+
  geom_tile(color="blue",size=1)
```



6.3 Conclusion

Since the p-value in our ANOVA table is less than .05, we have sufficient evidence to reject the null hypothesis.

This means we have sufficient evidence to say that the mean capital gain is not equal across different education levels.

From the Tukey test, we can see the p-values for different education pairs, and the difference in average capital gain.

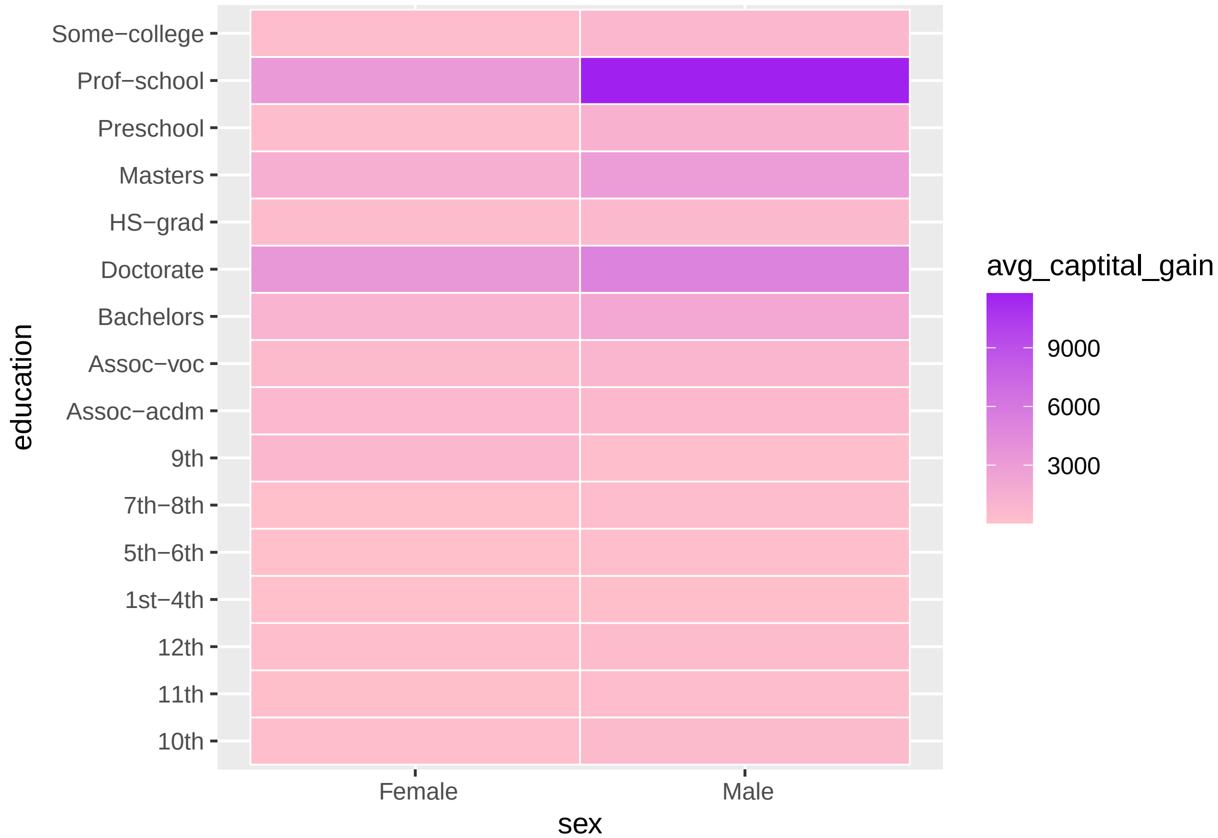
From the plots, we can see that the maximum average capital gain is with the education prof school.

7 Plotting capital gain on education and sex

```
education_sex<-adult %>%
  group_by(sex, education) %>%
  summarize(avg_captital_gain=mean(capital_gain))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```r
education_sex %>%
  ggplot(aes(x=sex,y=education,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="pink",high="purple")
```

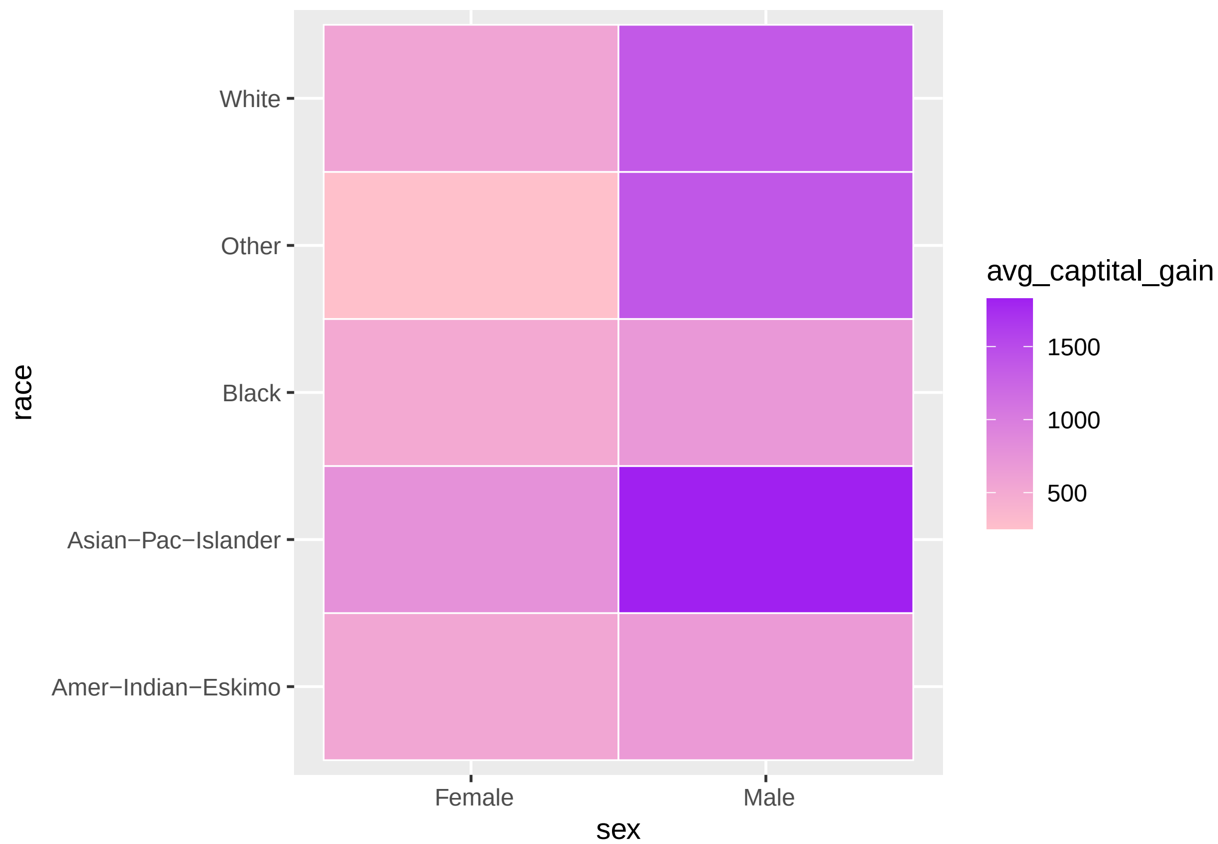


8 Plotting capital gain on race and sex

```r
race_sex<-adult %>%
  group_by(sex, race) %>%
  summarize(avg_captital_gain=mean(capital_gain))
```

```
## 'summarise()' has grouped output by 'sex'. You can override using the '.groups'
## argument.
```

```r
race_sex %>%
  ggplot(aes(x=sex,y=race,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="pink",high="purple")
```



9 Average capital gain vs earning greater than or less than or equal to 50k

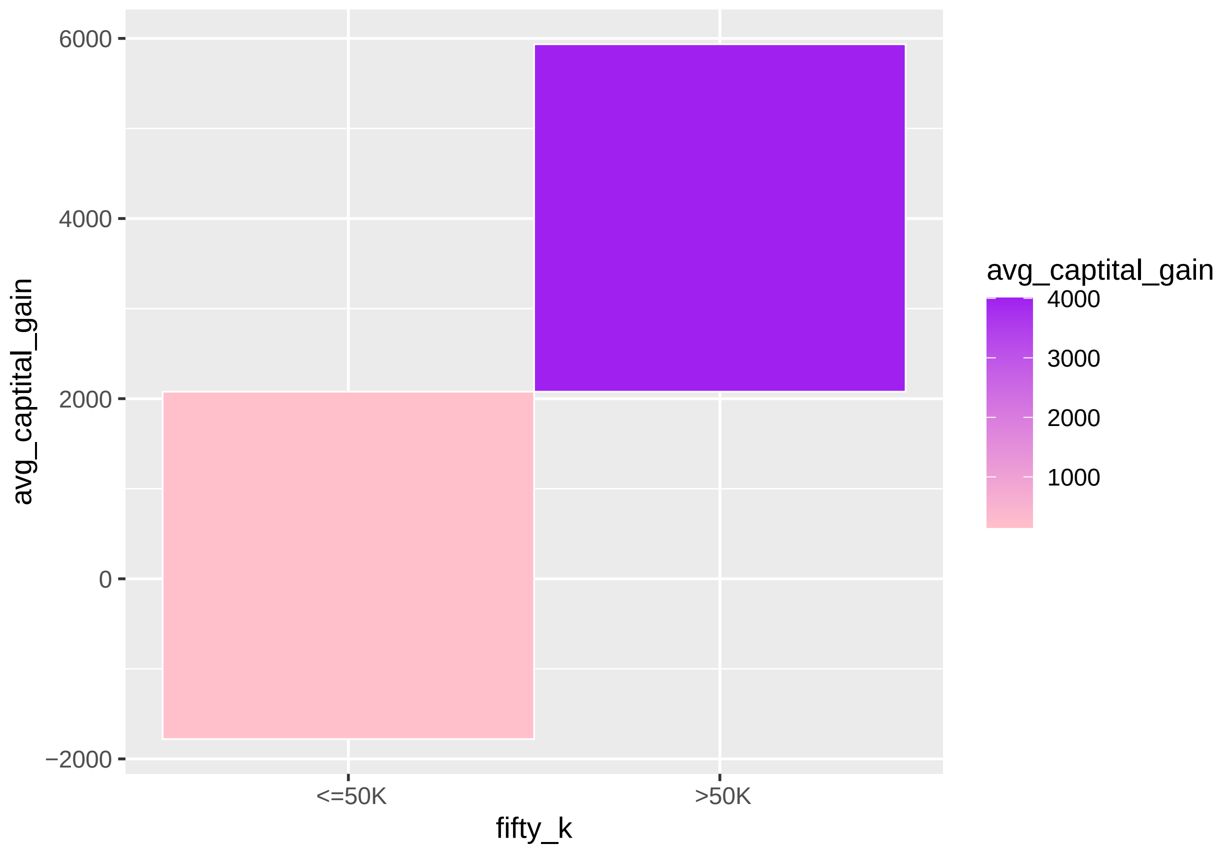
```
# adult %>%
#   group_by(fifty_k) %>%
#   summarise(record_count = n())


t.test(capital_gain ~ fifty_k, data=adult) # Unpooled
```

```
##
##  Welch Two Sample t-test
##
## data:  capital_gain by fifty_k
## t = -23.427, df = 7861.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  <=50K and group  >50K is not equal to
## 95 percent confidence interval:
##  -4180.166 -3534.614
## sample estimates:
## mean in group  <=50K  mean in group  >50K
##             148.7525            4006.1425
```

```r
t.test(capital_gain ~ fifty_k, var.equal=TRUE, data=adult)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  capital_gain by fifty_k
## t = -41.342, df = 32559, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group  <=50K and group  >50K is not equal to
## 95 percent confidence interval:
##  -4040.271 -3674.509
## sample estimates:
## mean in group  <=50K  mean in group  >50K
##             148.7525            4006.1425
```

```r
gain_fifty<-adult %>%
  group_by(fifty_k) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_fifty %>%
  ggplot(aes(x=fifty_k, y=avg_captital_gain,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="pink",high="purple")
```



9.1 Conclusion

Looking at the p value which is close to 0, we can reject the null hypothesis.

We have evidence that suggests that the true difference in means between group that earns less than or equal to 50k and more than 50 is not equal to 0.

We have evidence to say that there is a significant difference in the average capital gain.

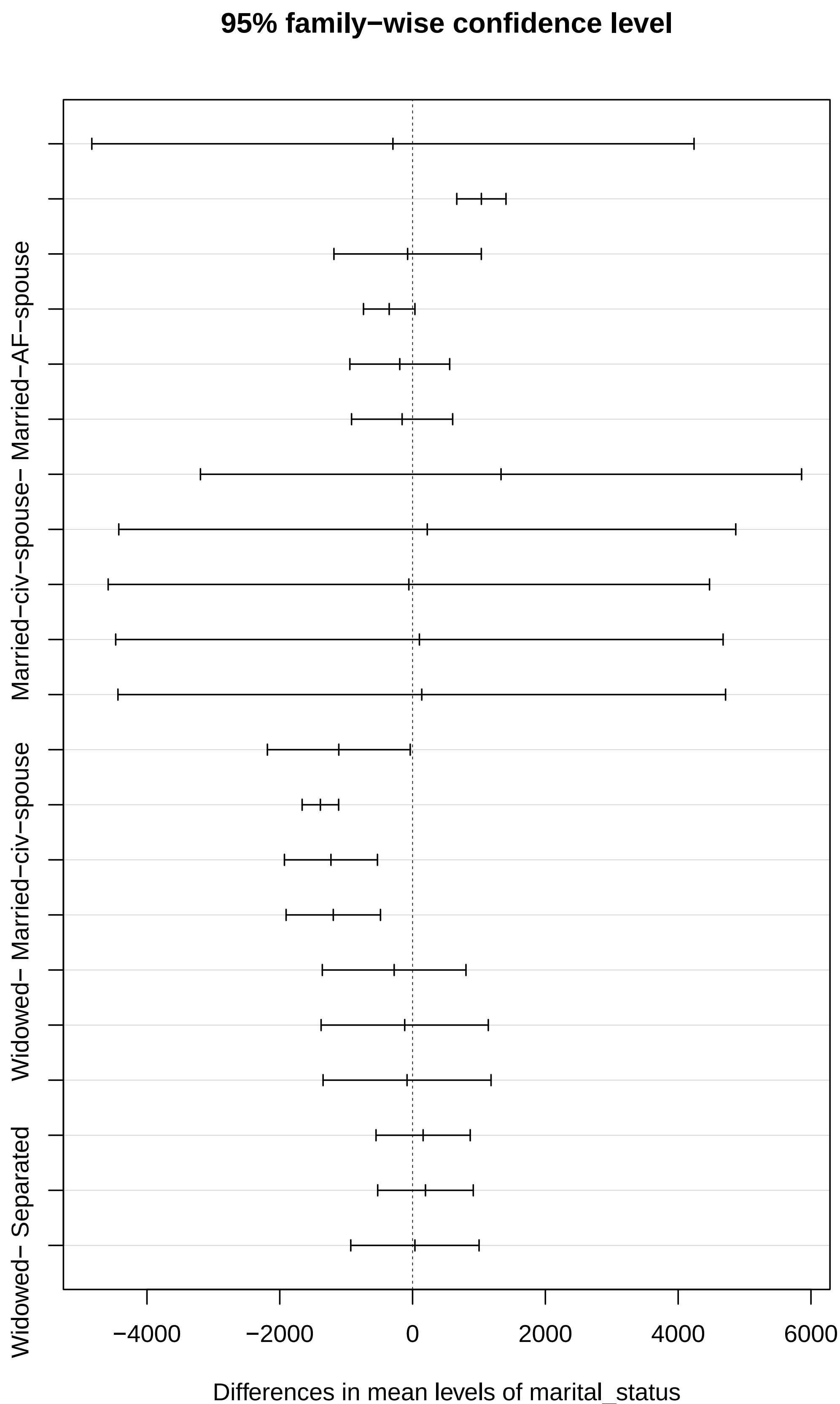
10 Testing relationship between capital gain and marital status

```
# adult %>%
#   group_by(race) %>%
#   summarise(record_count = n())

anov_race <- aov(capital_gain ~ marital_status, data = adult)
summary(anov_race)
```

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## marital_status     6 1.351e+10 2.251e+09   41.58 <2e-16 ***
## Residuals      32554 1.762e+12 5.414e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

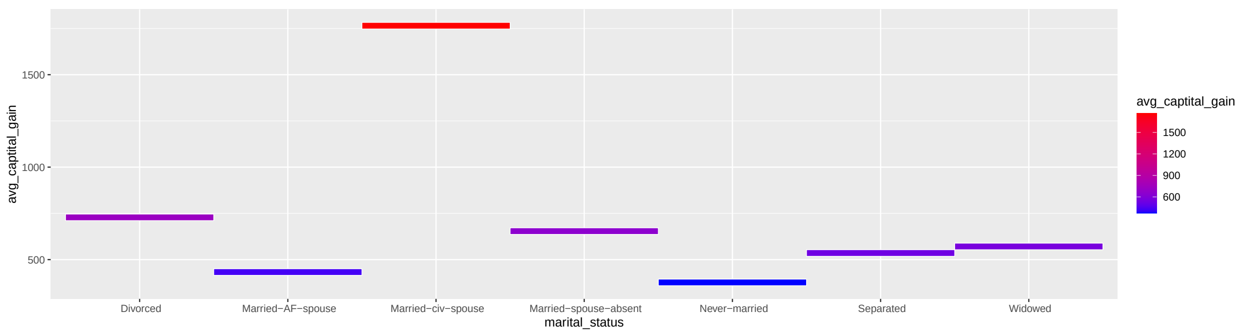
```
#TukeyHSD(anov_race)
```

```
plot(TukeyHSD(aov(capital_gain ~ marital_status, data = adult)))
```



95% family−wise confidence level
Widowed− Separated
Widowed− Married−civ−spouse
Married−civ−spouse− Married−AF−spouse
−4000
−2000
0
2000
4000
6000
Differences in mean levels of marital_status

```r
gain_marital<-adult %>%
  group_by(marital_status) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_marital %>%
  ggplot(aes(x=marital_status, y=avg_captital_gain,fill=avg_captital_gain))+
  geom_tile(color="white",size=0.3)+
  scale_fill_gradient(low="blue",high="red")
```



10.1 Conclusion

Since the p-value in our ANOVA table is less than .05, we have sufficient evidence to reject the null hypothesis.

This means we have sufficient evidence to say that the mean capital gain is not equal across different marital-status.

From the Tukey test, we can see the p-values for different marital status pairs, and the difference in average capital gain.

From the plots, we can see that the maximum average capital gain is with married-civ-spouse.

11 Testing relationship between capital gain and native country

Motivation: we want to find out if the capital gain differs based on native country.

11.1 Assumptions

1. The dataset is a random sample of original population.
2. The data comes from a normal distribution.
3. The sample size is large enough to conduct any test.
4. And the final assumptions is homogeneity of variance.

11.2 Hypothesis

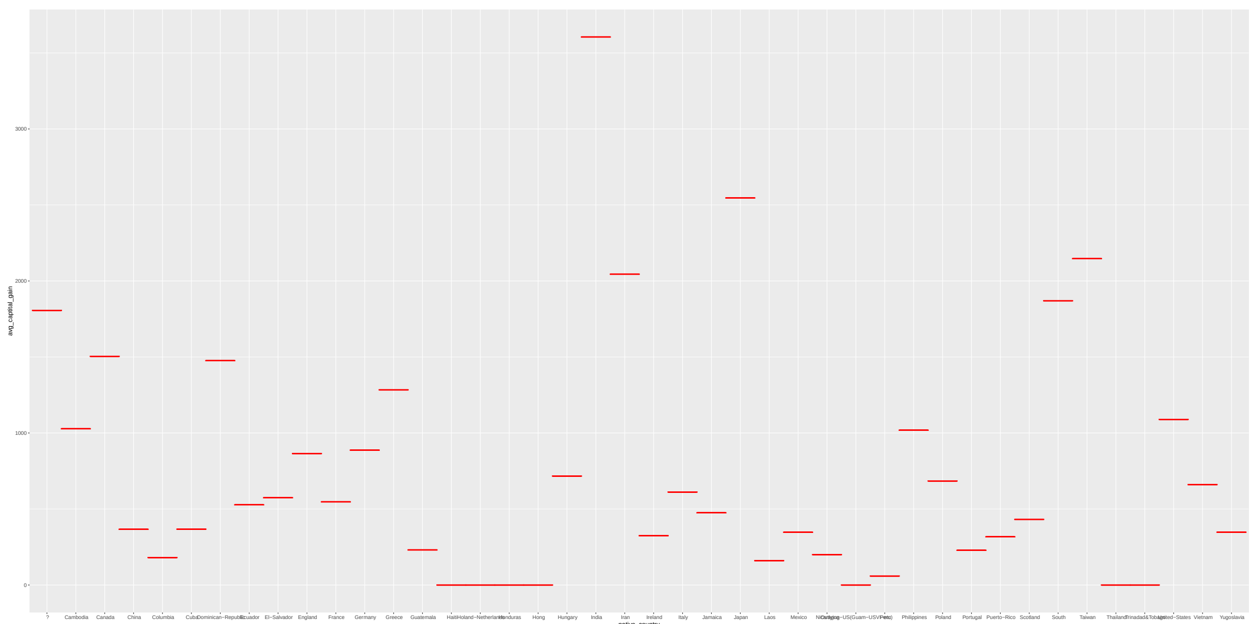
H0: capital gain is equal for different native countries

Ha: there exist a pair of native countries for which capital gain is not equal.

```r
# adult %>%
#   group_by(native_country) %>%
#   summarise(record_count = n())

anov_country <- aov(capital_gain ~ native_country, data = adult)
summary(anov_country)
```

```
##                    Df    Sum Sq  Mean Sq F value Pr(>F)
## native_country     41 2.256e+09 55022066   1.009  0.455
## Residuals       32519 1.774e+12 54541935
```

```r
gain_country<-adult %>%
  group_by(native_country) %>%
  summarize(avg_captital_gain=mean(capital_gain))

gain_country %>%
  ggplot(aes(x=native_country, y=avg_captital_gain))+
  geom_tile(color="red",size=1)
```



11.3 Conclusion

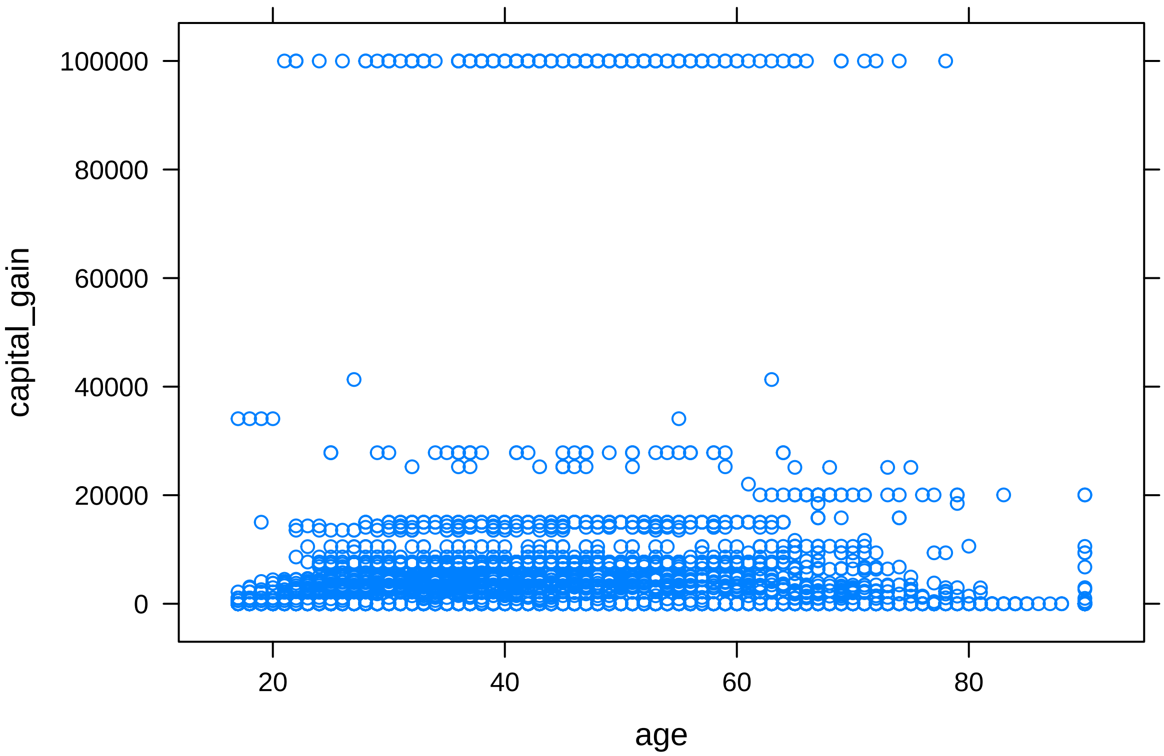
Since the p-value in our ANOVA table is greater than .05, we do not have sufficient evidence to reject the null hypothesis.

This means we do not have sufficient evidence to say that the mean capital gain is not equal across different native countries.

From the plots, we can see that the maximum average capital gain is for native country India.

12 Linear Regression on Census Data

```
xyplot(capital_gain ~ age, data=adult)
```



As from initial plot, we can see that capital gain and age not in linear relationship, there is no point to fit a linear model with the data.

We checked if linear regression can be used for any of the numerical attributes vs capital gain, but we noticed no significant information that points to a linear relationship, hence we did not use any linear models.

13 Summary:

After analyzing the dataset, we concluded following points.

1. Capital gain is not equal and varies based on multiple factors.
2. On average a male has higher capital gain than a woman.
3. The capital gain differs based on race.
4. From hypothesis tests, we also observed that work class, occupation and education levels also play major role deciding the capital gain. Capital gain is not equal across all these attributes.

14 Real Estate data set

14.1 Introduction

We performed analysis on an additional real estate data set since we found it interesting. This dataset has various attributes which could potentially affect house prices such as number of convenience stores, distance from metro stations and we aim to analyze and find if a linear relationship exists between the house price (Y) and any of the other parameters. This dataset has 414 observations.

14.1.1 Attributes

1) X1.transaction.date - Date of transaction
2) X2.house.age - Age of the property
3) X3.distance.to.the.nearest.MRT.station - Distance to nearest metro station
4) X4.number.of.convenience.stores - Number of convenience stores
5) X5.latitude - Latitude
6) X6.longitude - Longitude
7) Y.house.price.of.unit.area - Price of house per area (Response variable)

14.1.2 Loading the data

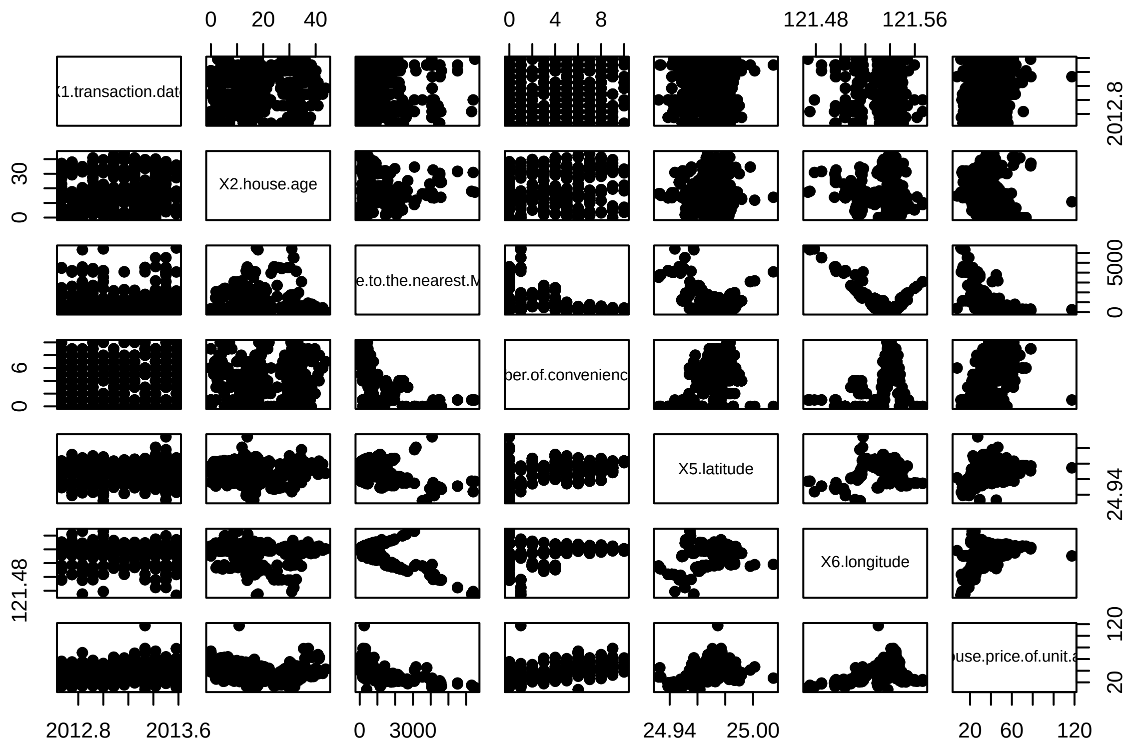
```
real_estate <- read.csv("Real_Estate.csv")
summary(real_estate)
```

```
##       No        X1.transaction.date  X2.house.age
##  Min.   :  1.0   Min.   :2013       Min.   : 0.000
##  1st Qu.:104.2   1st Qu.:2013       1st Qu.: 9.025
##  Median :207.5   Median :2013       Median :16.100
##  Mean   :207.5   Mean   :2013       Mean   :17.713
##  3rd Qu.:310.8   3rd Qu.:2013       3rd Qu.:28.150
##  Max.   :414.0   Max.   :2014       Max.   :43.800
##  X3.distance.to.the.nearest.MRT.station X4.number.of.convenience.stores
##  Min.   :  23.38                        Min.   : 0.000
##  1st Qu.: 289.32                        1st Qu.: 1.000
##  Median : 492.23                        Median : 4.000
##  Mean   :1083.89                        Mean   : 4.094
##  3rd Qu.:1454.28                        3rd Qu.: 6.000
##  Max.   :6488.02                        Max.   :10.000
##   X5.latitude     X6.longitude    Y.house.price.of.unit.area
##  Min.   :24.93   Min.   :121.5   Min.   :  7.60
##  1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70
##  Median :24.97   Median :121.5   Median : 38.45
##  Mean   :24.97   Mean   :121.5   Mean   : 37.98
##  3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60
##  Max.   :25.01   Max.   :121.6   Max.   :117.50
```

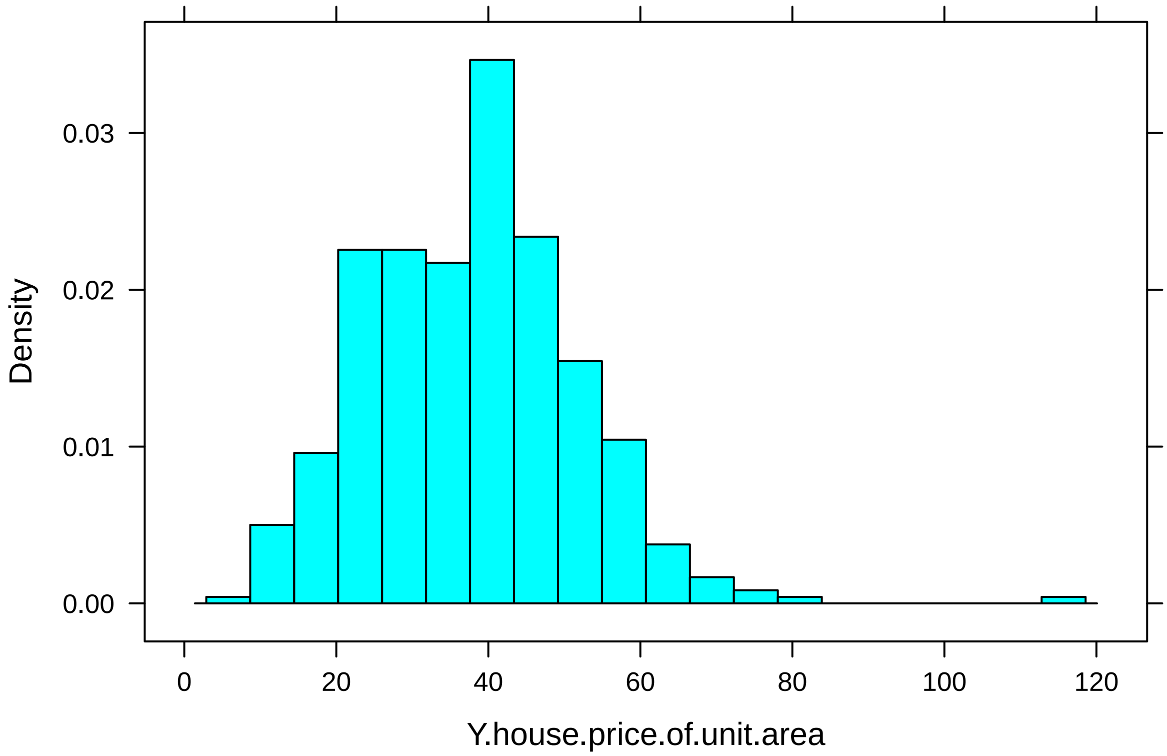
```
ls(real_estate)
```

```
## [1] "No"
## [2] "X1.transaction.date"
## [3] "X2.house.age"
## [4] "X3.distance.to.the.nearest.MRT.station"
## [5] "X4.number.of.convenience.stores"
## [6] "X5.latitude"
## [7] "X6.longitude"
## [8] "Y.house.price.of.unit.area"
```

```
pairs(real_estate[,2:8], pch=19)
```




```
#xyplot(Y.house.price.of.unit.area ~ X4.number.of.convenience.stores,data=real_estate)  # positive tren
#xyplot(Y.house.price.of.unit.area ~ X3.distance.to.the.nearest.MRT.station,data=real_estate)  # negati

# study with distance to metro station.

#check value distribution.
histogram(~Y.house.price.of.unit.area, data=real_estate, nint=20)
```



14.2 Testing if house price varies with distance to metro station

```
#check correlation between house price and distance to metro station.
cor(Y.house.price.of.unit.area ~ X3.distance.to.the.nearest.MRT.station, data=real_estate) # -0.673
```

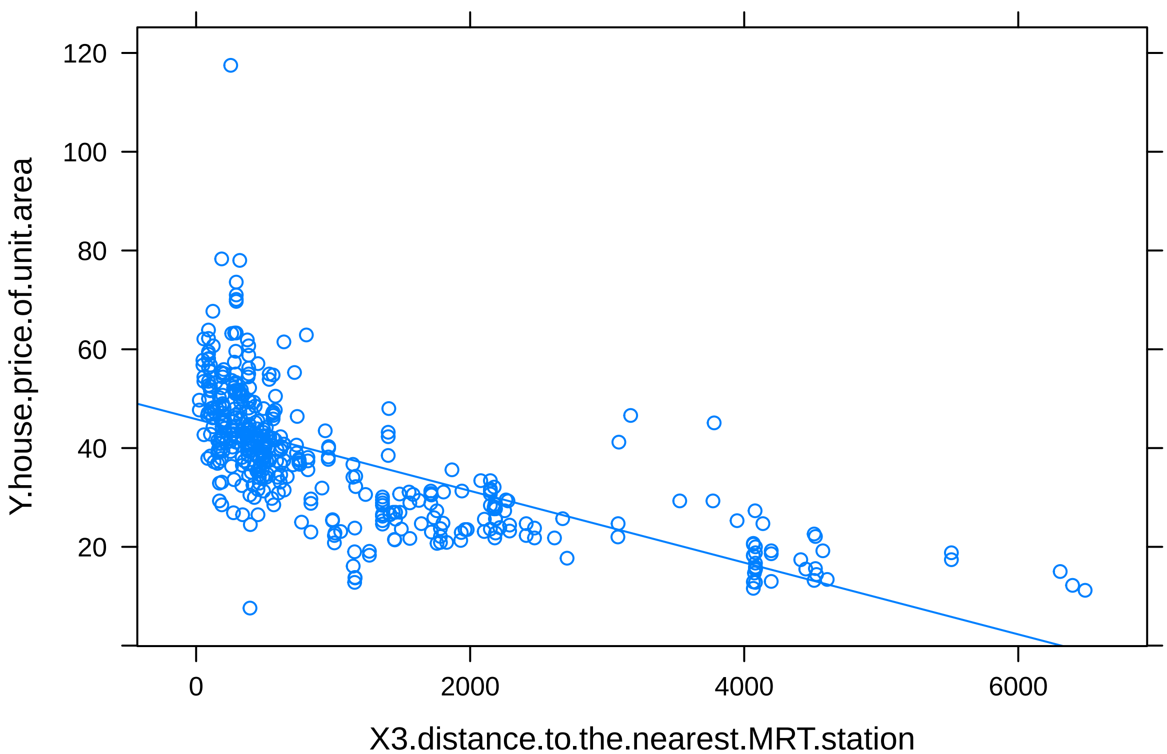
```
## [1] -0.6736129
```

```
#the least squares line  regression line.
m1 <- lm(Y.house.price.of.unit.area ~ X3.distance.to.the.nearest.MRT.station, data=real_estate)
summary(m1)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X3.distance.to.the.nearest.MRT.station,
##     data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.396  -6.007  -1.195   4.831  73.483
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           45.8514271  0.6526105   70.26   <2e-16
```

```
## X3.distance.to.the.nearest.MRT.station -0.0072621  0.0003925  -18.50   <2e-16
##
## (Intercept)                            ***
## X3.distance.to.the.nearest.MRT.station ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 412 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4524
## F-statistic: 342.2 on 1 and 412 DF,  p-value: < 2.2e-16
```

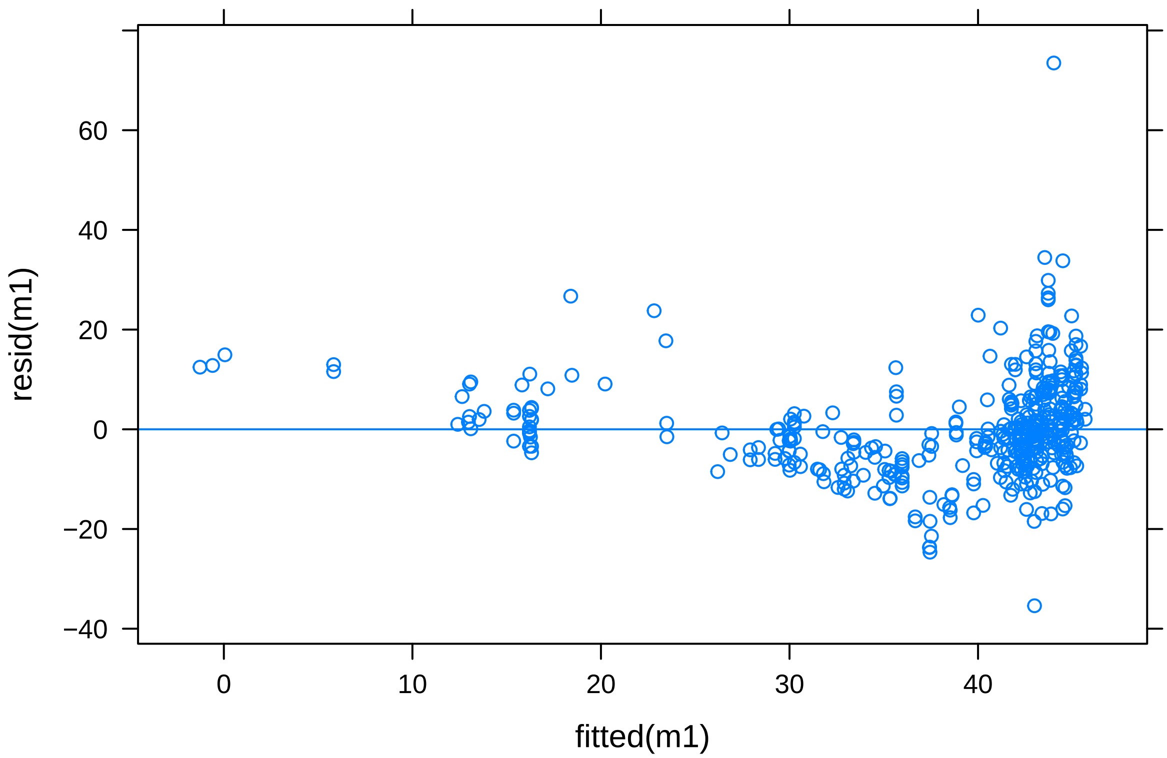
```
#xy plot
xyplot(Y.house.price.of.unit.area ~ X3.distance.to.the.nearest.MRT.station,data=real_estate,type=c("p",
```



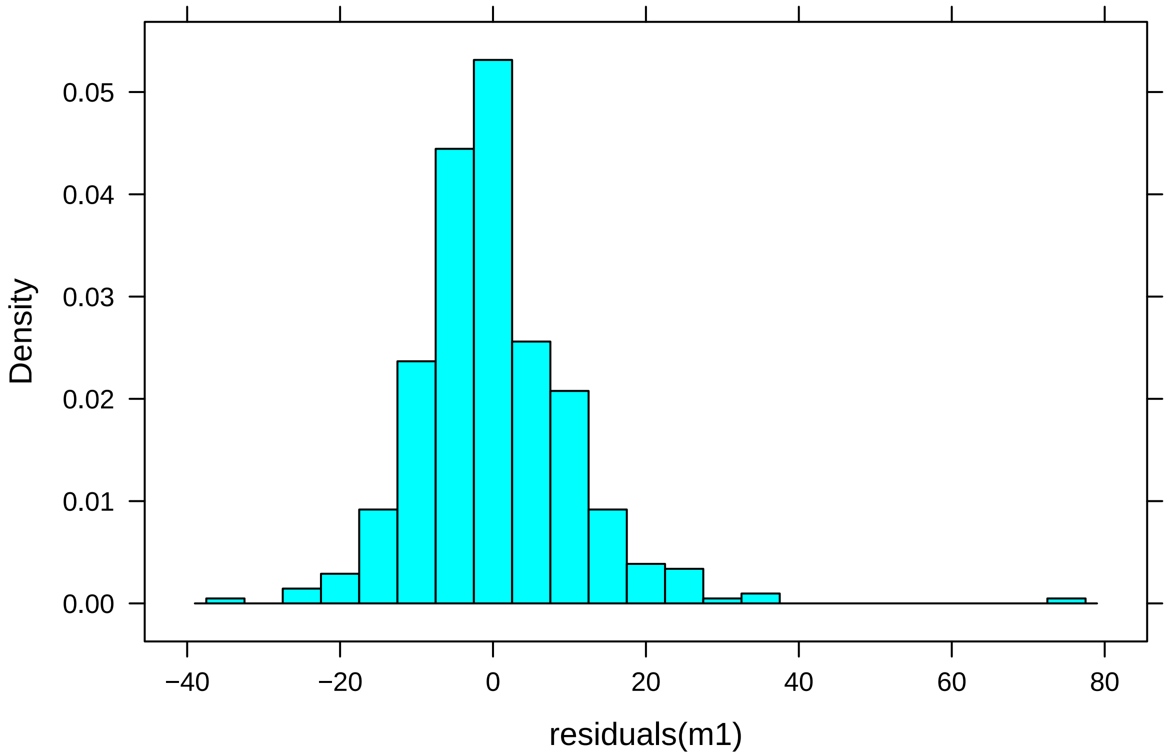
14.2.1 Assumptions

1. Residual are uniformly distributed around y=0 horizontal line.
2. Residual follows normal distribution.
3. The relationship between two variables should be linear.
4. The observation should be independent of each other.

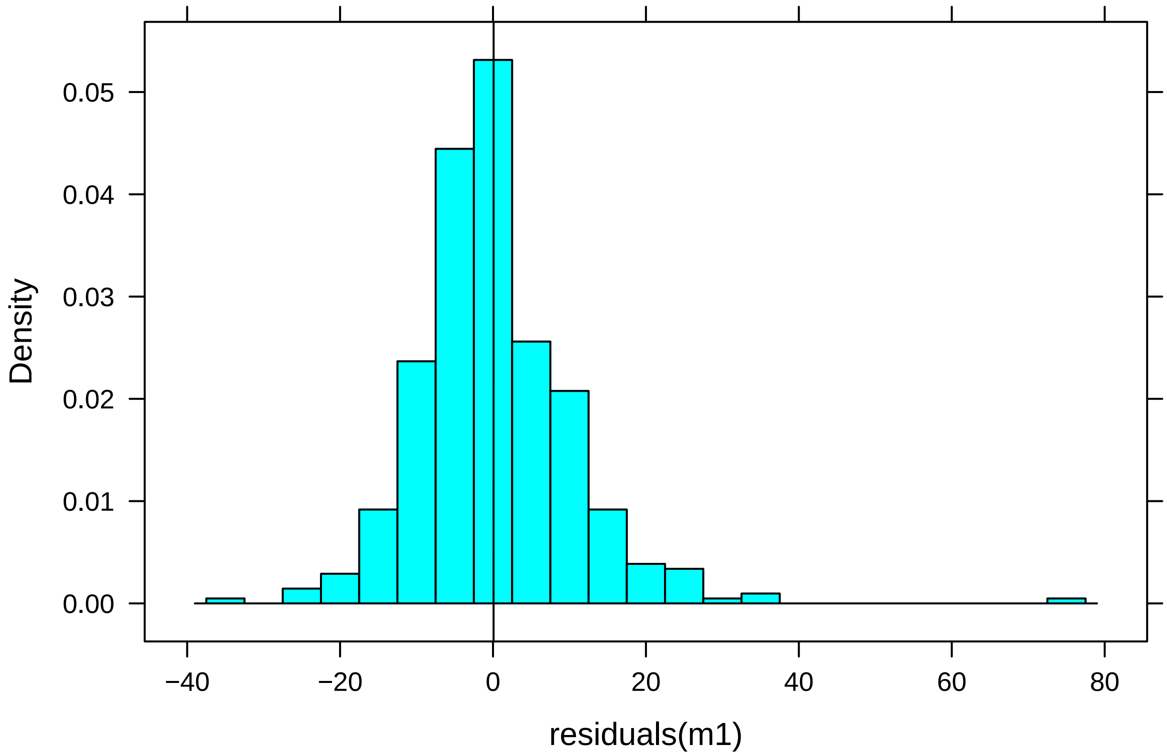
```r
#normalty check of errors/residual and assumptions check *
xyplot(resid(m1)~fitted(m1), data=real_estate, type=c("p","r"))
```



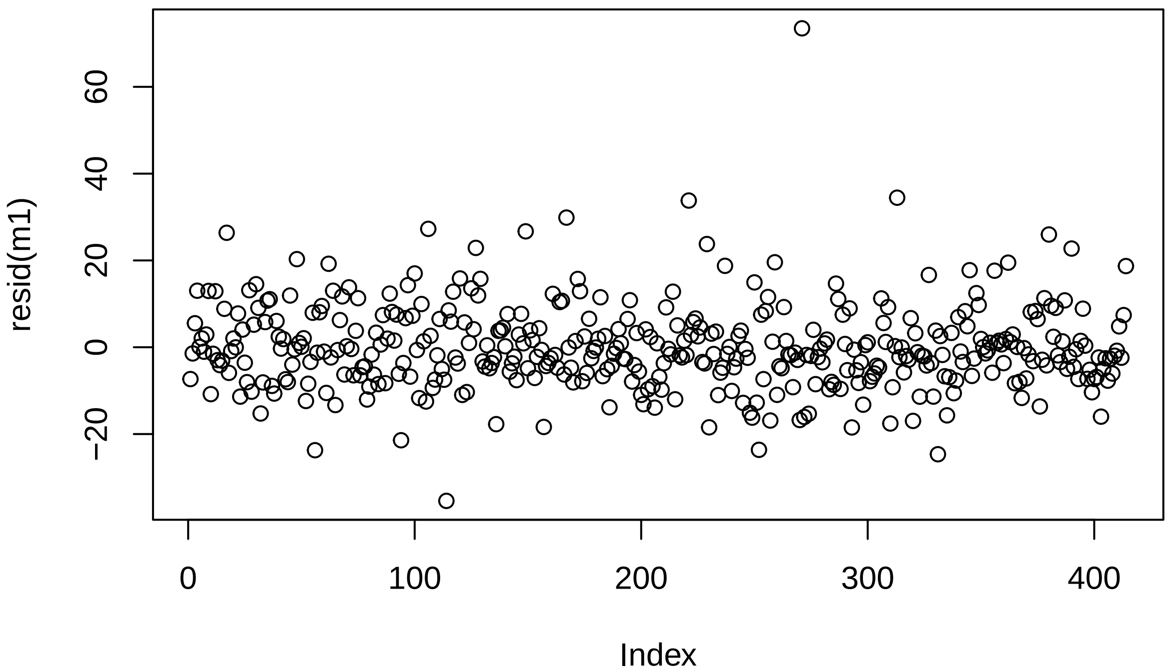

```r
histogram(residuals(m1),width=5)
```



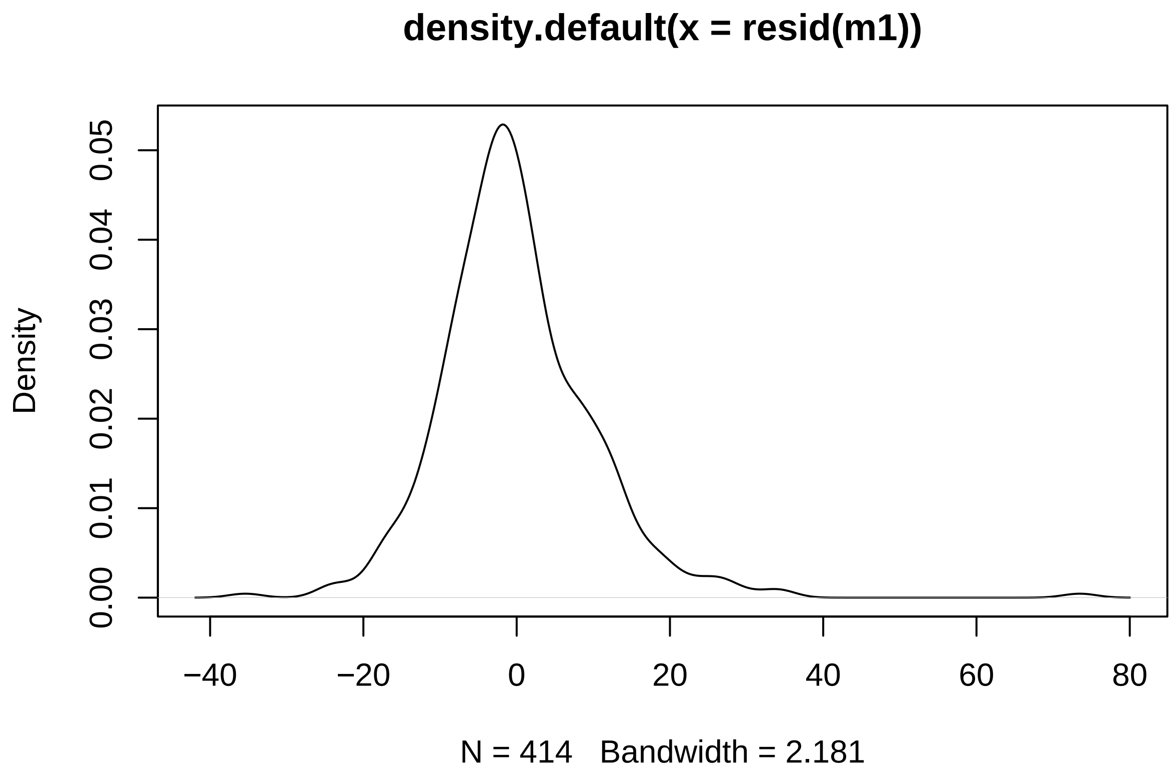

```
ladd(panel.qqmathline(resid(m1)))
```



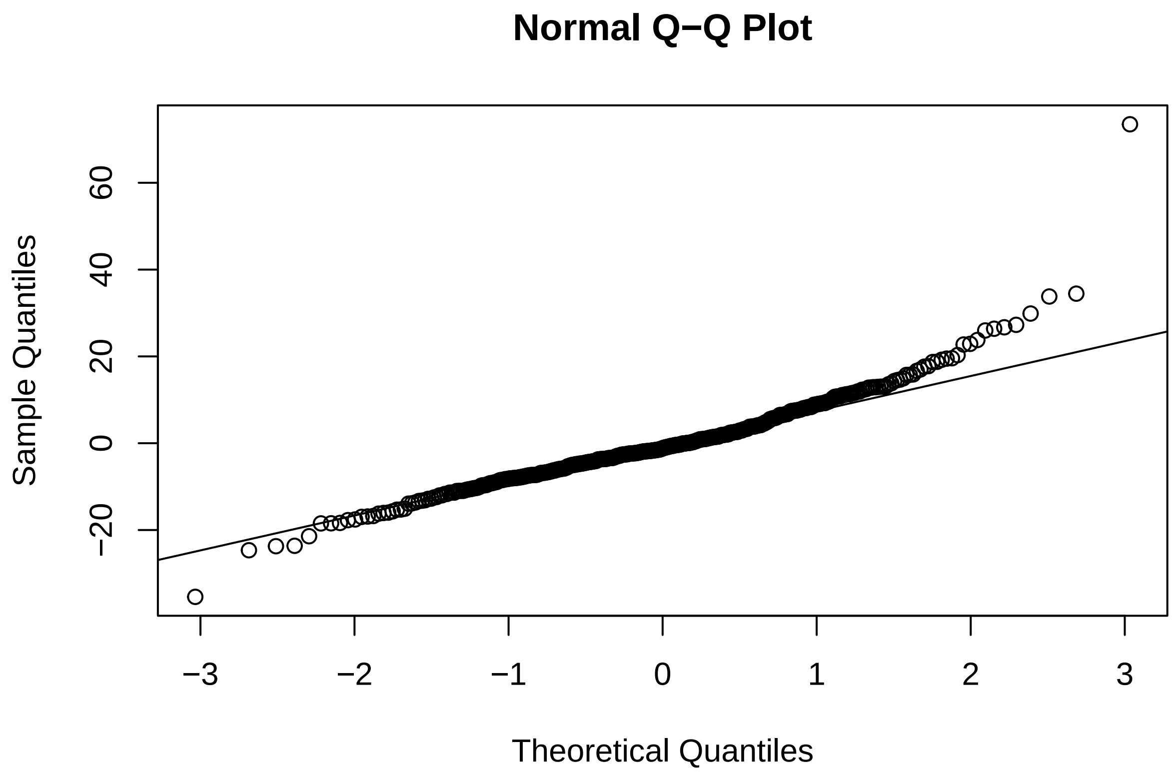

```
plot(resid(m1))
```




```
plot(density(resid(m1)))
```




```
qqnorm(resid(m1))
qqline(resid(m1))
```



14.2.2 Conclusion

All assumptions holds here. From different graphs we can see that the conditions for linear model fitting holds.

14.3 Testing if house price varies with number of convenience stores

```
m2 <- lm(Y.house.price.of.unit.area ~  X4.number.of.convenience.stores,  data=real_estate)
summary(m2)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X4.number.of.convenience.stores,
##     data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.407  -7.341  -1.788   5.984  87.681
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      27.1811     0.9419   28.86   <2e-16 ***
## X4.number.of.convenience.stores   2.6377     0.1868   14.12   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 412 degrees of freedom
## Multiple R-squared:  0.326,  Adjusted R-squared:  0.3244
## F-statistic: 199.3 on 1 and 412 DF,  p-value: < 2.2e-16
```

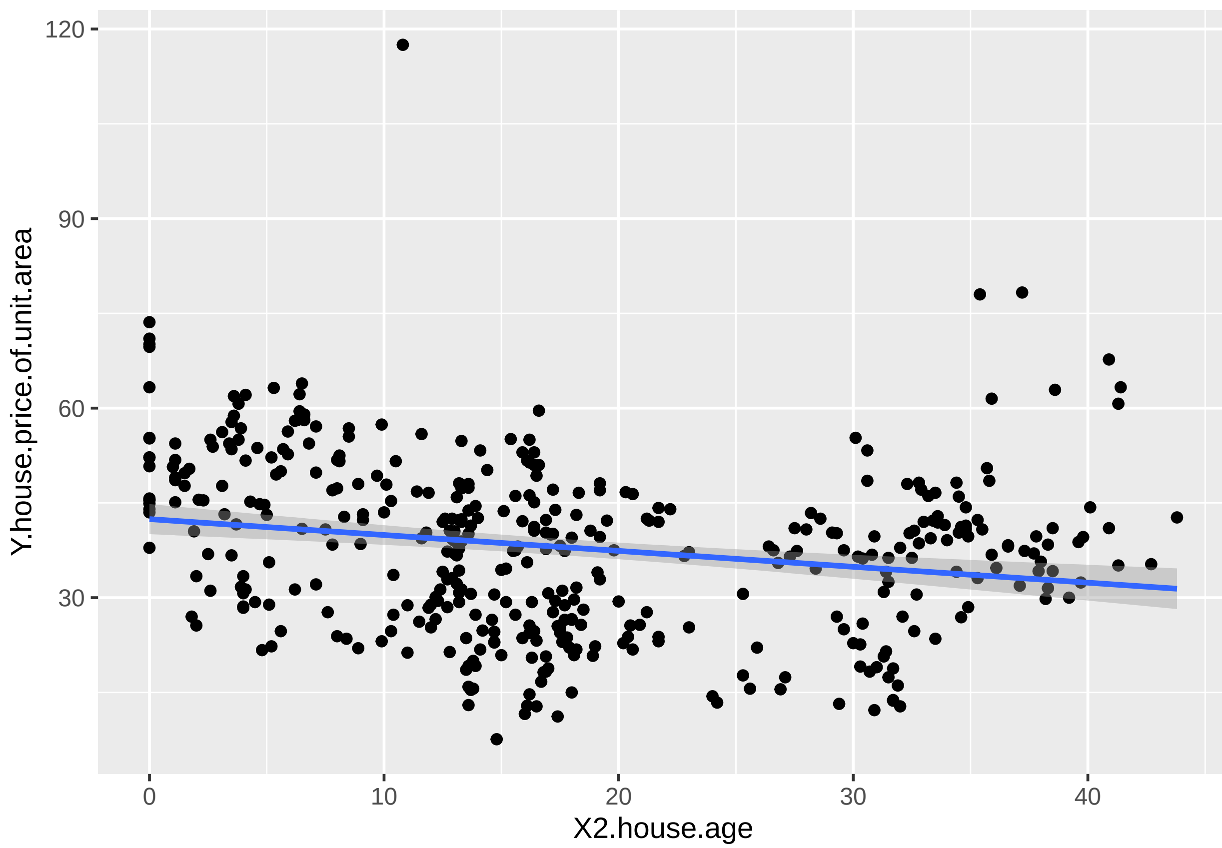
14.4 Testing if house price varies with house age.

```
m3 <- lm(Y.house.price.of.unit.area ~ X2.house.age,  data=real_estate)
summary(m3)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X2.house.age, data = real_estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.113 -10.738   1.626   8.199  77.781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.43470    1.21098  35.042  < 2e-16 ***
## X2.house.age -0.25149    0.05752  -4.372 1.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.32 on 412 degrees of freedom
## Multiple R-squared:  0.04434,    Adjusted R-squared:  0.04202
## F-statistic: 19.11 on 1 and 412 DF,  p-value: 1.56e-05
```

```
ggplot(real_estate, aes( X2.house.age, Y.house.price.of.unit.area)) + geom_point() + stat_smooth(method
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Assumptions

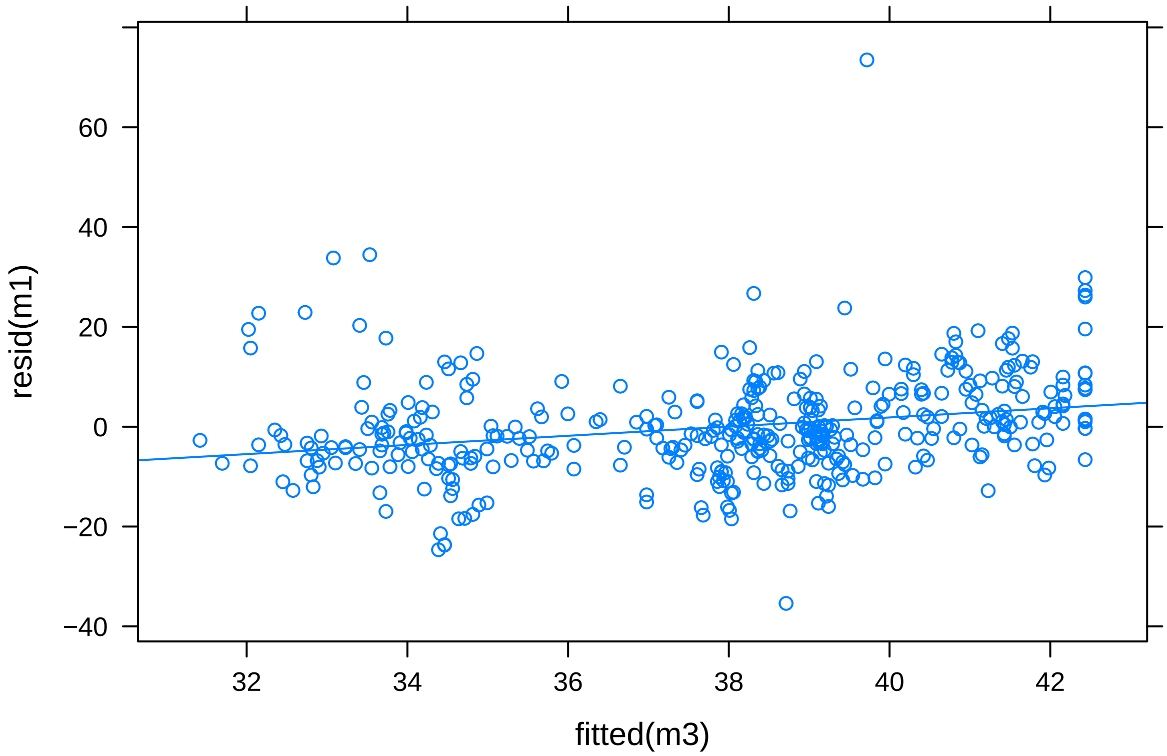
1. Residual are uniformly distributed around y=0 horizontal line.

2. Residual follows normal distribution.

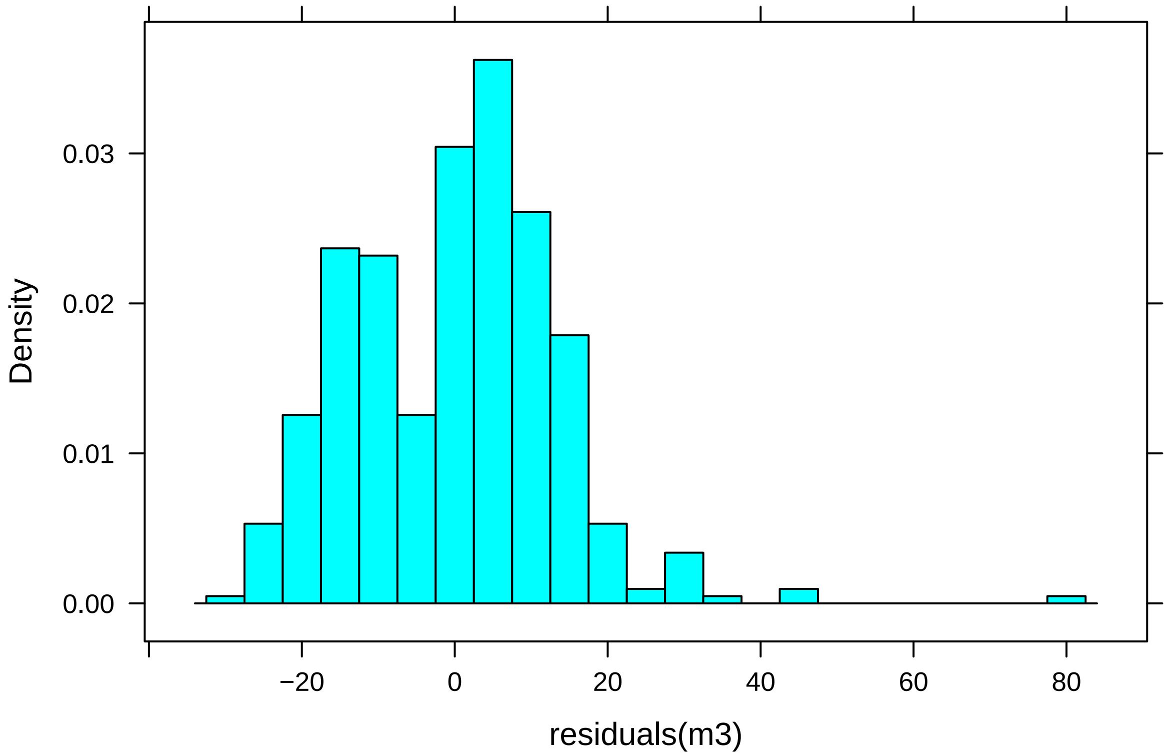
3. The relationship between two variables should be linear.

4. The observation should be independent of each other.

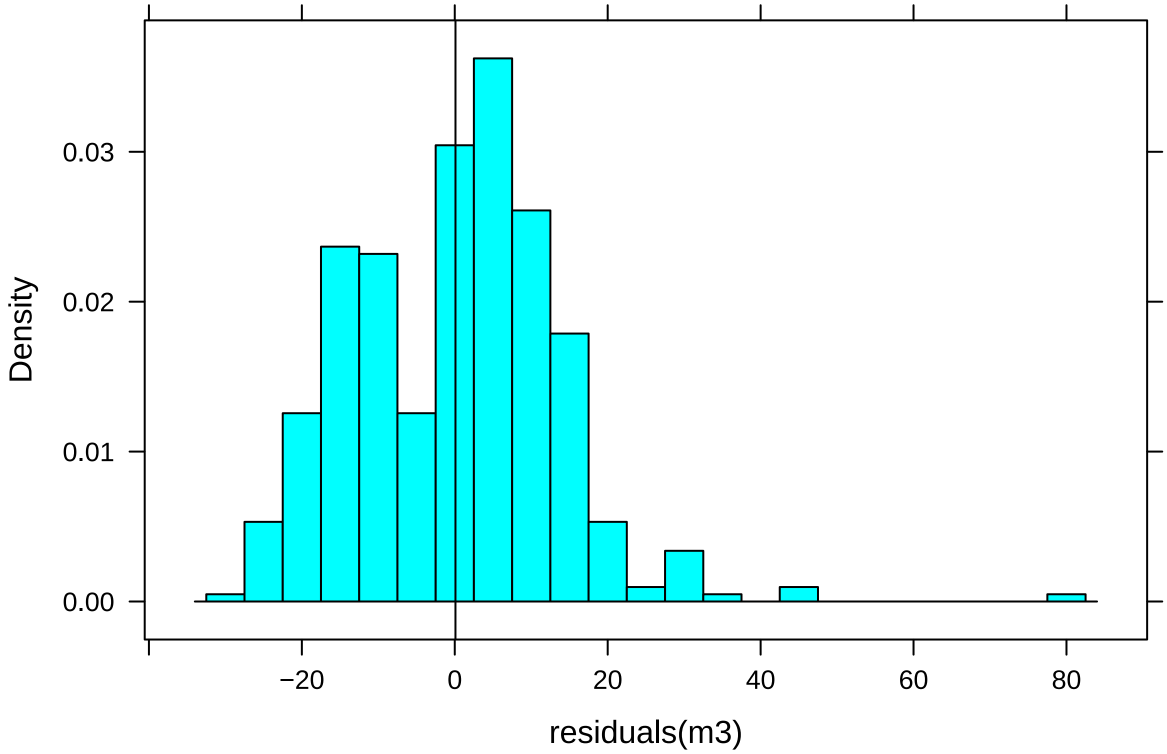
```
xyplot(resid(m1)~fitted(m3), data=real_estate, type=c("p","r"))
```



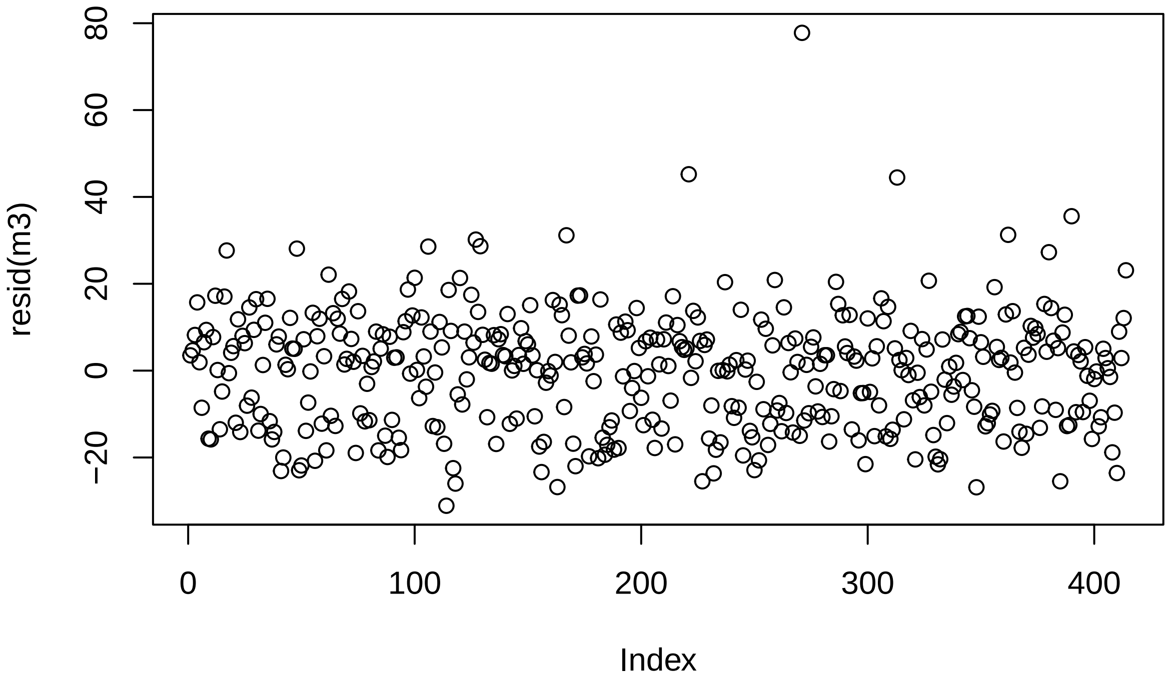

```
histogram(residuals(m3),width=5)
```



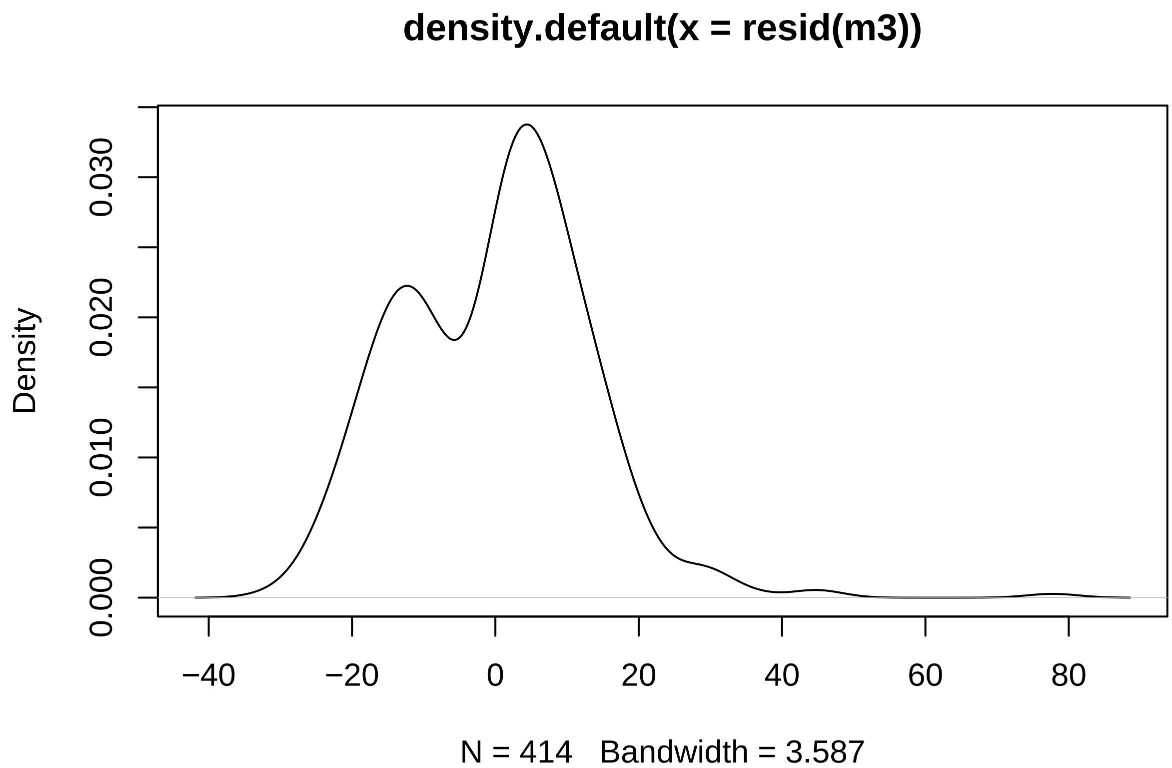

```
ladd(panel.qqmathline(resid(m3)))
```



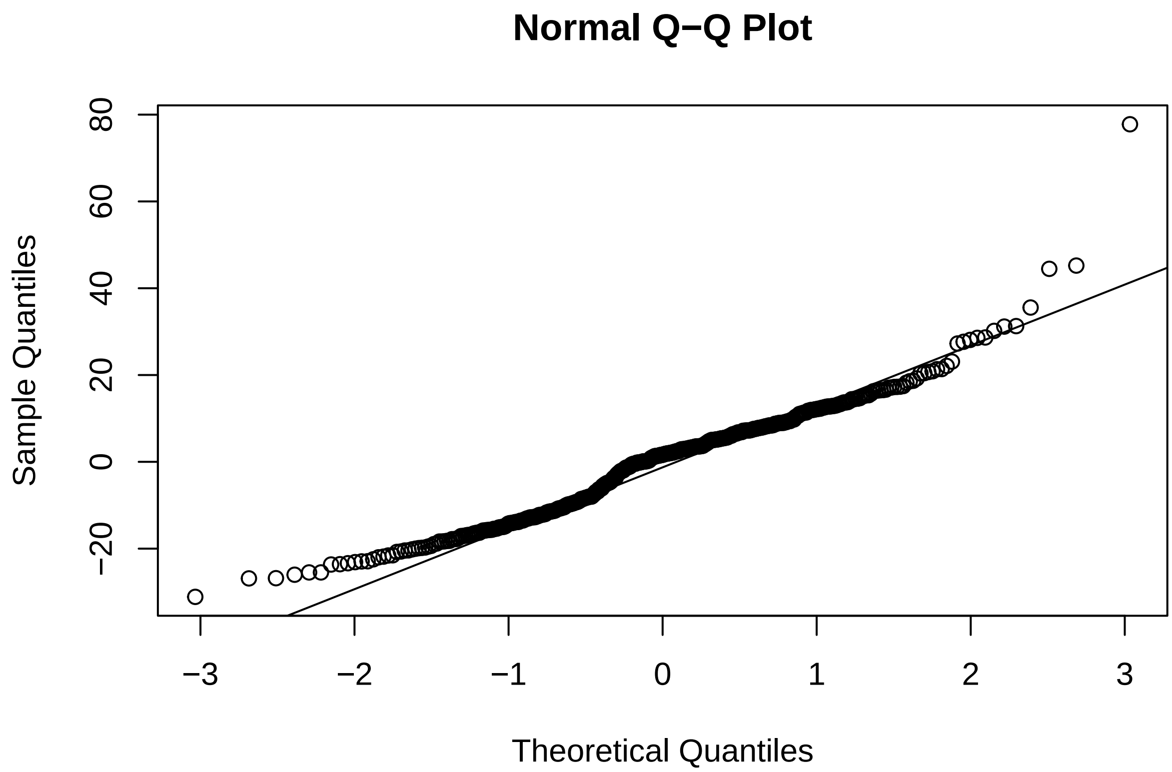

```
plot(resid(m3))
```




```
plot(density(resid(m3)))
```




```
qqnorm(resid(m3))
qqline(resid(m3))
```



14.4.1 Conclusion

From the above graphs, the relationship between house price and house age is not linear, and from Q_Q plot also, we can see that the residuals are not on a straight lines and uniform distribution of error around y=0 horizontal lines doesn't hold also, so we should not use linear model to predict the house price based on house age. And if we build the model, we can see that the R-squared value is around 4%, which also indicates linear model is not suitable to predict the house price based on house age.

15 Summary:

After analyzing the the dataset we found following points:

1. The price of a house depends on house age, distance to the nearest metro stations, number of convenience stores nearby etc, but it need not to be they are linearly dependent.
2. For house price and distance to nearest metro station, there exist a liner relationship and for that reason, we are able to fit a linear model.
3. Whereas the number of convenience stores and house age is not linearly dependent, so we could not fit any linear model there. We checked the assumptions and reported where it failed to hold the conditions.