FINAL EXAM - MA4142

Dependent Variable - Housing

- Printing the head of the data

```
> # head of the data
> head(data)
  Income Commute Literacy JobGrowth Physicians RapeRate Restaurants
1 26,000    49.2     5.15      10.8       1987     51.3        5582
2 29,300    45.3     5.97       9.5        517     50.8        9988
3 24,800    39.8     9.41       8.2        592     77.7       20511
4 27,900    46.8     4.61       7.6       3310     51.2        8946
5 37,500    39.9     5.64      12.2        975     40.1        4000
6 31,900    49.5     4.80       7.7       2238     38.0        8970
  Housing MedianAge HouseholdIncome
1 109,400      35.3          68,000
2  97,000      43.2          70,400
3 114,700      29.5          60,500
4  99,100      40.5          65,900
5 122,200      47.1          84,700
6 145,300      39.3          75,800
```

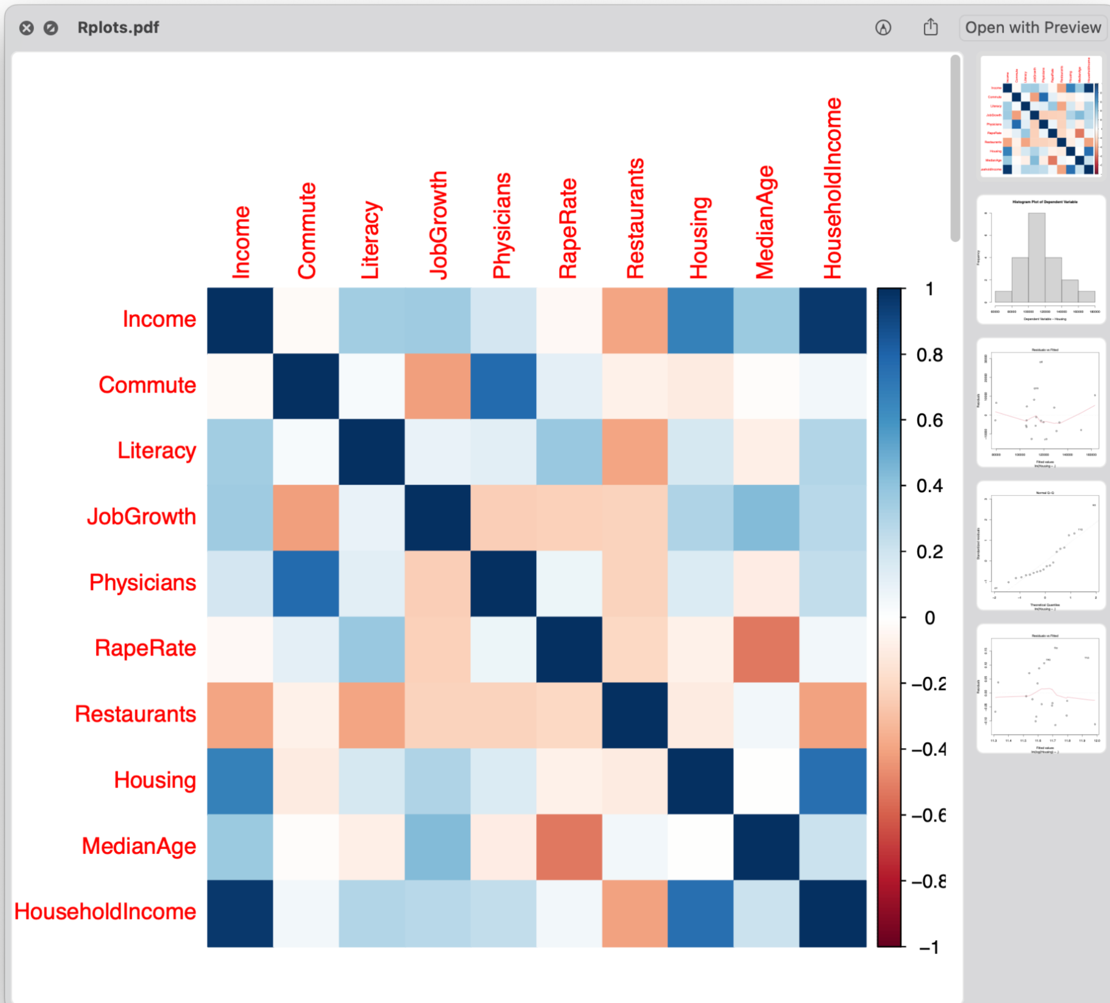
The above summary tells that the data has a total of 10 variables, among which *Housing* is the dependent variable, and rest 9 are the independent variables

- Getting the summary of the data

```
> # summary of the data
> summary(data)
    Income             Commute         Literacy        JobGrowth     
 Length:20          Min.   :37.80   Min.   :1.660   Min.   : 4.700  
 Class :character   1st Qu.:40.83   1st Qu.:3.405   1st Qu.: 7.325  
 Mode  :character   Median :44.60   Median :4.915   Median : 8.150  
                    Mean   :44.12   Mean   :4.671   Mean   : 8.360  
                    3rd Qu.:45.67   3rd Qu.:5.610   3rd Qu.: 9.125  
                    Max.   :53.50   Max.   :9.410   Max.   :13.900  
   Physicians        RapeRate      Restaurants       Housing         
 Min.   : 166.0   Min.   :17.80   Min.   : 2655   Length:20         
 1st Qu.: 359.2   1st Qu.:42.05   1st Qu.: 8964   Class :character  
 Median : 530.0   Median :51.05   Median :11978   Mode  :character  
 Mean   :1059.2   Mean   :51.80   Mean   :15512                     
 3rd Qu.:1617.8   3rd Qu.:59.85   3rd Qu.:16179                     
 Max.   :4143.0   Max.   :83.60   Max.   :65804                     
   MedianAge     HouseholdIncome   
 Min.   :29.50   Length:20         
 1st Qu.:35.10   Class :character  
 Median :38.95   Mode  :character  
 Mean   :38.84                     
 3rd Qu.:41.62                     
 Max.   :52.70                     
```

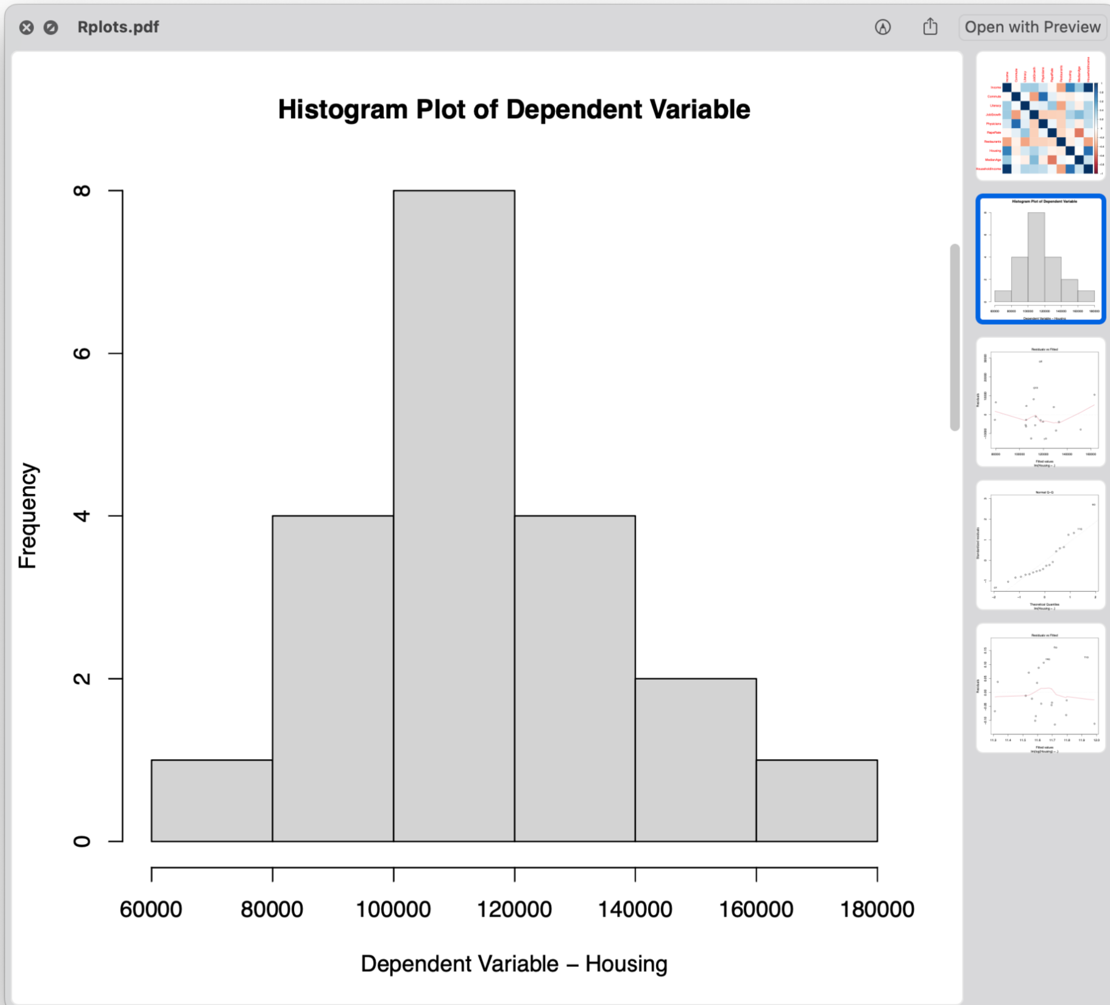
This summary of the data tells us about the statistics of each variable. As we can see, each attribute of the data gave us the summary of the least (Min.) value, highest (Max.) value, the mean, median, and 1st and 3rd quartiles.

- Plotting the Correlation Matrix of the data with their variables



The above image is the correlation matrix plot consisting of all 10 variables. As we can clearly see, income and household income are highly correlated with a factor of 1. And Housing (dependent variable) is highly correlated with income more than any other independent variable.

- Plotting the histogram of the dependent variable - Housing



The above graph is the histogram plot of the dependent variable (Housing). As we can clearly see, the housing value between 100000 and 120000 takes up the highest frequency of 8 (out of 20).

Fitting an appropriate model to the data

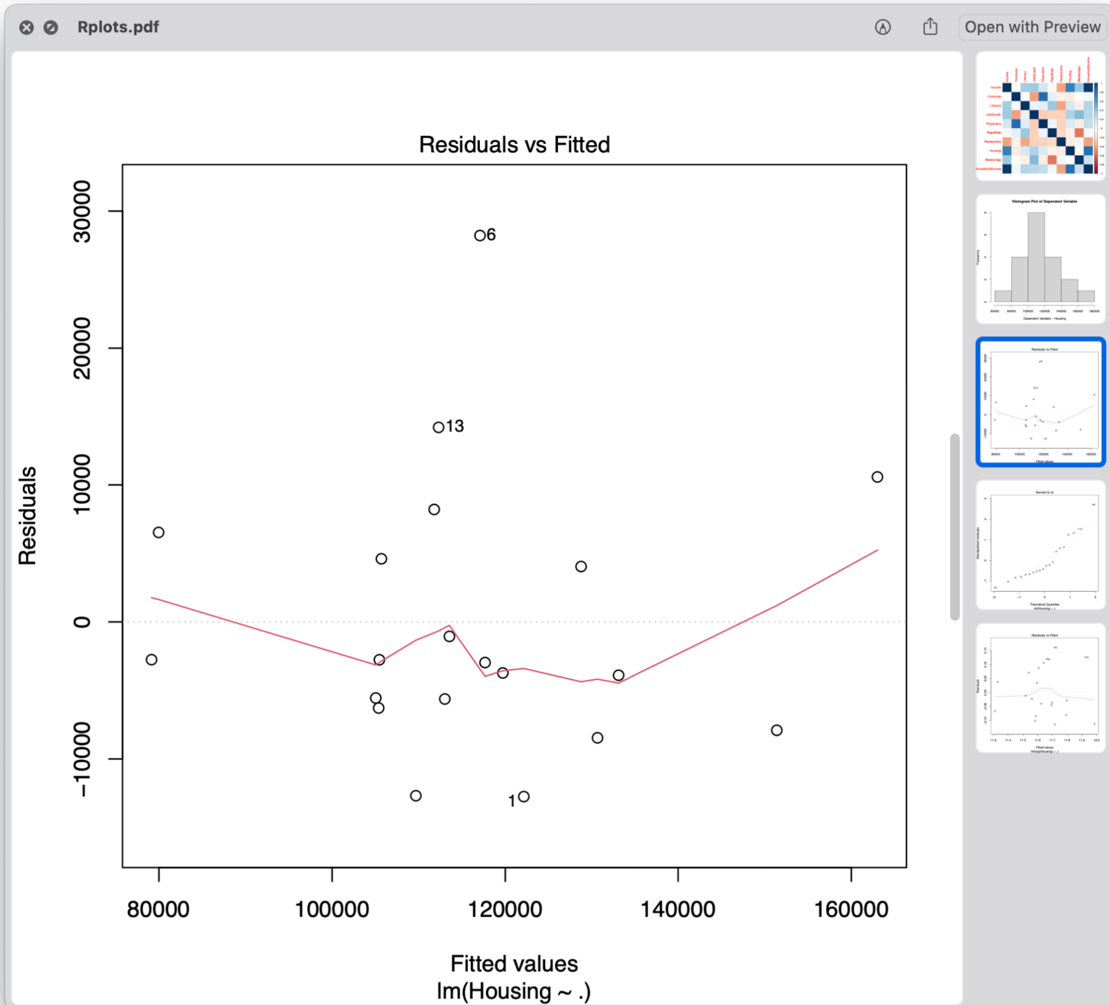
- We first fit a linear regression model to the data. The summary of the model is as follows

```
> # fit linear regression model on the dataset
> model <- lm(Housing ~ ., data = data)
> # summary of the model
> summary(model)

Call:
lm(formula = Housing ~ ., data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-12741  -5786  -2862   5086  28217 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)  
(Intercept)     89192.0526 65702.0891   1.358   0.2045  
Income             -4.6119     3.3923  -1.360   0.2038  
Commute         -1418.9681  1564.3143  -0.907   0.3857  
Literacy         2155.2877  2262.7216   0.953   0.3633  
JobGrowth        2057.0238  1847.6049   1.113   0.2916  
Physicians          4.1323     5.1789   0.798   0.4435  
RapeRate         -336.7469   242.5468  -1.388   0.1952  
Restaurants         0.5078     0.2756   1.843   0.0952 .
MedianAge        -767.2750  1053.0830  -0.729   0.4830  
HouseholdIncome     3.2746     1.3832   2.367   0.0394 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13670 on 10 degrees of freedom
Multiple R-squared:  0.7979,	Adjusted R-squared:  0.6159 
F-statistic: 4.385 on 9 and 10 DF,  p-value: 0.0152
```

Here, we have first taken a linear regression model and tried to fit it into our dataset. From the summary of the model, we can see that the Multiple R-Squared value is 0.7979, stating that this regression model fits almost 79% of the data accurately. And the Adjusted R-Squared value is 0.6159. We can initially assume that this linear regression model satisfies all the assumptions of a linear model. Now, we try to verify them.

- Checking whether all the assumptions of a linear regression model are satisfied or not

- Linearity Assumptio



Here, we can see that the plot between residuals and fitted values yields almost linearly. The red line in the graph is almost linear, henceforth, we can say that this assumption is correct.

- Autocorrelation Assumption

```
> # 2. Independence of errors (Auto correlation)
> # Performing Durbin-Watson test for Auto correlation
> # Durbin-Watson statistic close to 2 implies no auto correlation
> dwtest(model)

        Durbin-Watson test

data:  model
DW = 2.2172, p-value = 0.5357
alternative hypothesis: true autocorrelation is greater than 0
```

For the autocorrelation, we perform the Durbin-Watson test. And the DW Statistic is given as 2.2172, with p-value = 0.5357. Ideally, the DW Statistic should be around 1.5 to 2.5 in order to say that the autocorrelation is absent. Here, we get that value as 2.21, which says that we failed to reject our null hypothesis. (It is to be also noted that our p-value > 0.05)

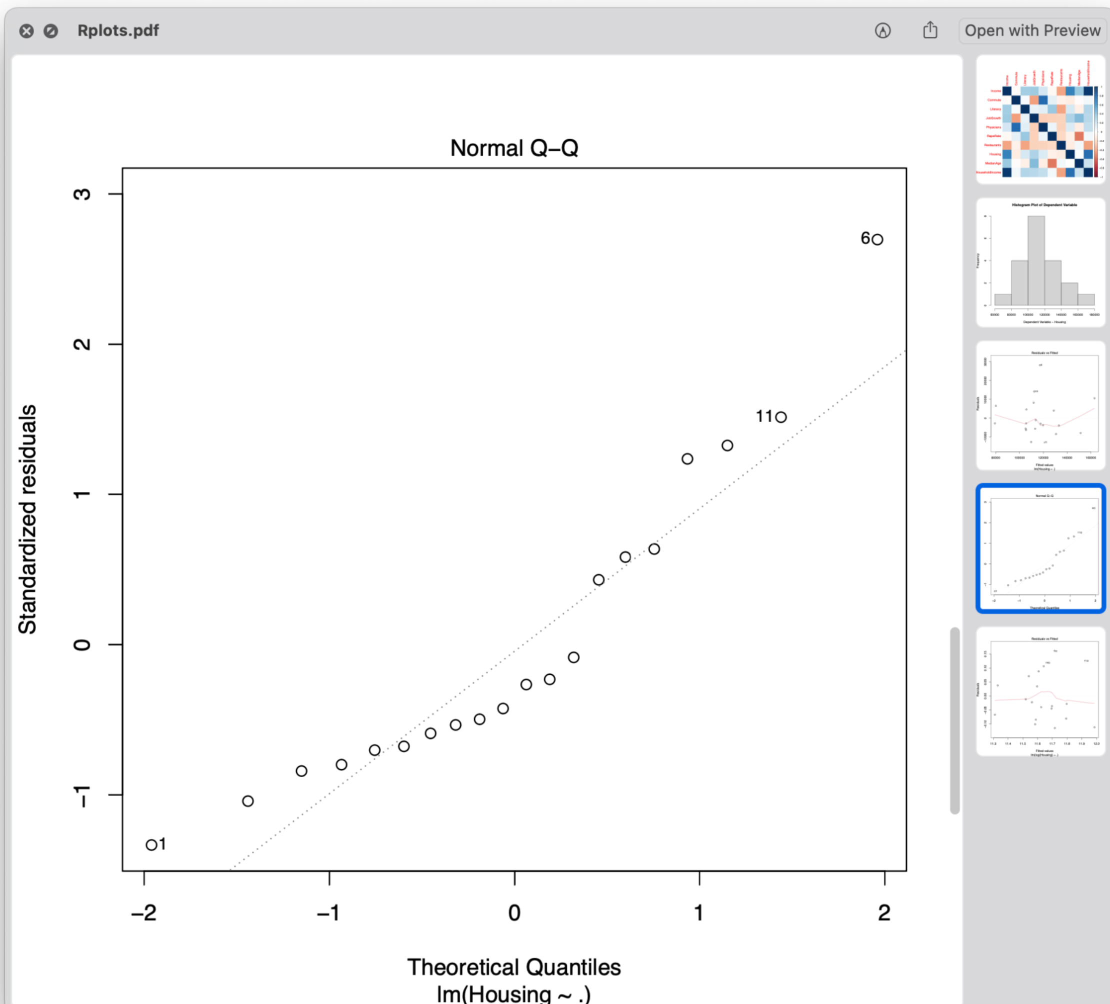
- Homoscedasticity Assumption

```
> # 3. Homoscedasticity
> # Performing Breusch-Pagan test for Homoscedasticity
> # Null hypothesis: Homoscedasticity
> # Alternate hypothesis: Heteroscedasticity
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 8.589, df = 9, p-value = 0.476
```

For the homoscedasticity, we perform the Breush-Pagan test to check if we get heteroscedasticity in the data. The test statistic says the p-value = 0.476 > 0.05 Therefore, we can say that this data model is not heteroscedastic and follows homoscedasticity.

- Normality of errors Assumption

```
> # 4. Normality of errors
> # Performing Shapiro-Wilk test for Normality of errors
> # Null hypothesis: Normality of errors
> # Alternate hypothesis: Non-normality of errors
> # Q-Q plot
> plot(model, which=2)
> shapiro.test(resid(model))

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.89898, p-value = 0.03947
```



We can see that the normality of errors assumption fails. Therefore, we try to correct this, and update our current model with a new one. We try to perform log transformations.

- Correction for Normality of Errors - Log Transformation
- Re-check for normality of errors Assumption

```
> shapiro.test(model_log$residuals)

        Shapiro-Wilk normality test

data:  model_log$residuals
W = 0.92177, p-value = 0.1072
```

After correction, we can see that the errors follow a normal distribution.

- Multicollinearity Assumption

```
> # 5. Multicollinearity
> # Performing Variance Inflation Factor (VIF) test for Multicollinearity
> # VIF > 10 implies Multicollinearity
> vif(model_log)
         Income         Commute        Literacy       JobGrowth      Physicians
      42.152047        4.096666        1.799669        1.841323        3.392345
       RapeRate     Restaurants       MedianAge HouseholdIncome
       1.815191        1.479778        3.411101       35.332742
```

- Correction for multicollinearity
- Removing the highest VIF Value - Income variable
- Checking the multicollinearity again and finding the summary of the new updated model

```
> # Income has high VIF value
> # Removing Income from the model
> new_data <- data[, -1]
> model_log <- lm(log(Housing) ~ ., data = new_data)
> #check VIF
> vif(model_log)
        Commute        Literacy       JobGrowth      Physicians        RapeRate
       3.490046        1.403594        1.814860        3.254180        1.792495
    Restaurants       MedianAge HouseholdIncome
       1.473944        2.004123        1.484061
```

- Getting the summary of the final updated model

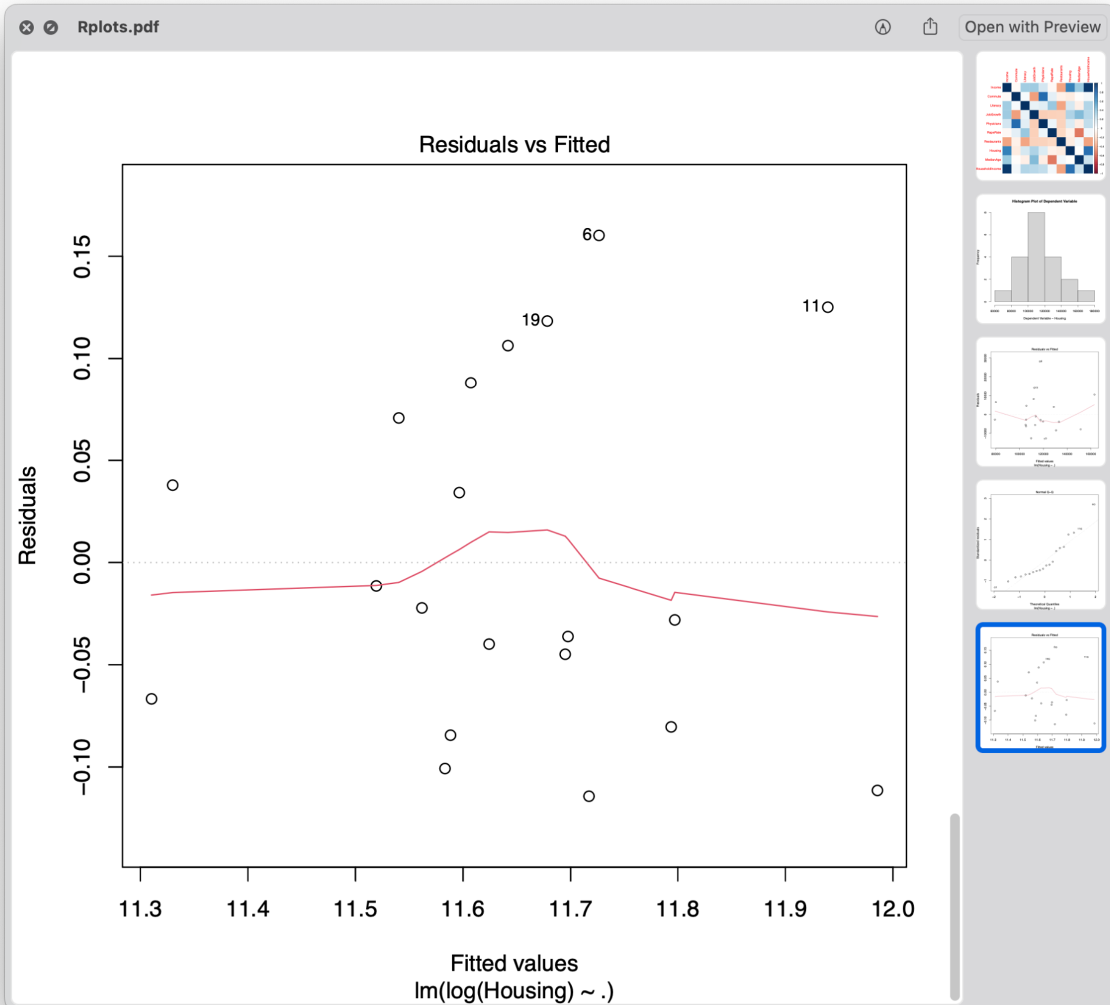
```
> # updated model summary
> summary(model_log)

Call:
lm(formula = log(Housing) ~ ., data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11430 -0.07011 -0.02513  0.07503  0.16020 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.153e+01  5.457e-01  21.123 2.97e-10 ***
Commute         -4.932e-03  1.199e-02  -0.411  0.68885    
Literacy         8.703e-03  1.660e-02   0.524  0.61052    
JobGrowth        2.507e-02  1.524e-02   1.645  0.12823    
Physicians       3.053e-05  4.214e-05   0.725  0.48387    
RapeRate        -3.253e-03  2.002e-03  -1.625  0.13247    
Restaurants      5.150e-06  2.285e-06   2.254  0.04555 *  
MedianAge       -1.712e-02  6.706e-03  -2.553  0.02683 *  
HouseholdIncome  1.175e-05  2.355e-06   4.988  0.00041 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1136 on 11 degrees of freedom
Multiple R-squared:  0.7862,	Adjusted R-squared:  0.6308 
F-statistic: 5.057 on 8 and 11 DF,  p-value: 0.00786
```

Thus, our new model's summary says that this model fits 78% of the data.

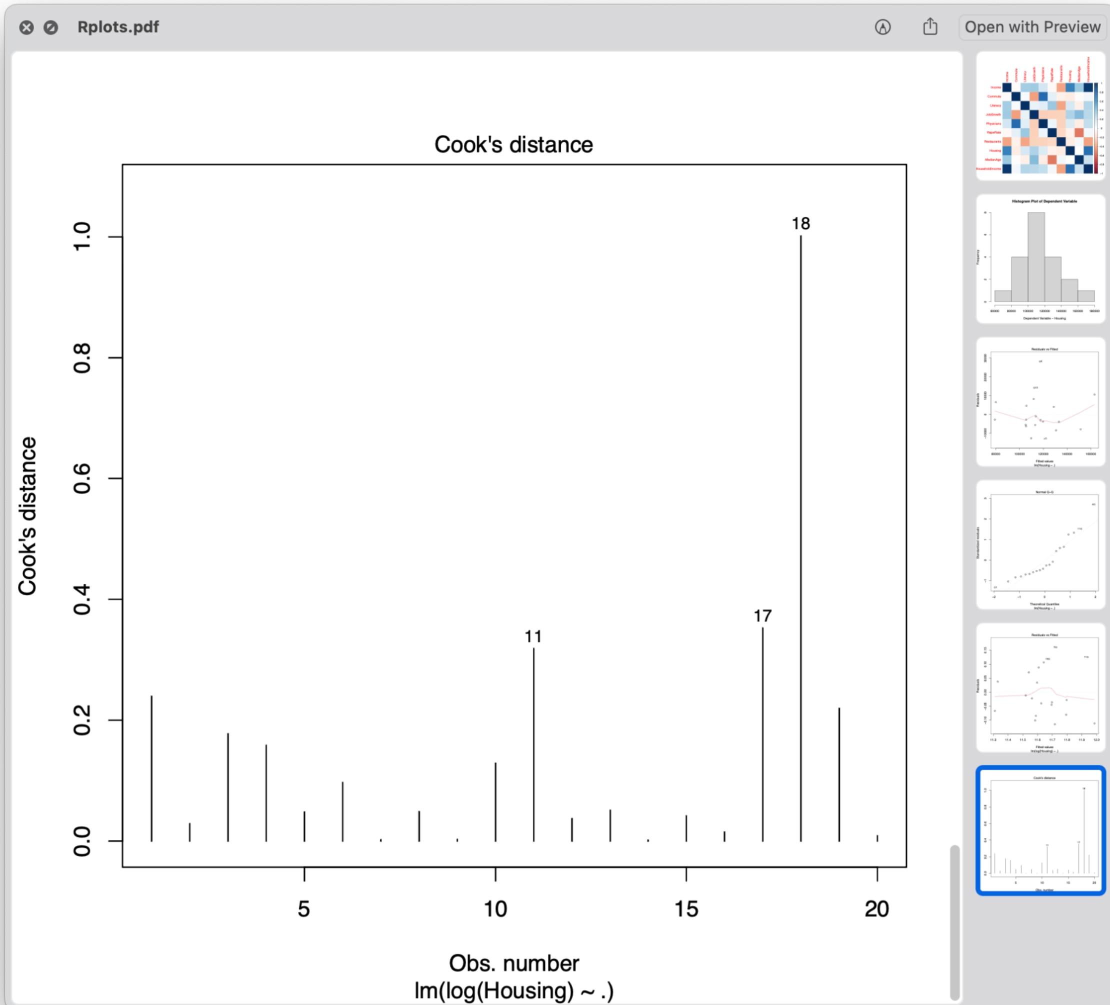
- Recheck all the assumptions for this final model

```
> # checking all the assumptions once again on the updated final model
> # 1. Linearity
> plot(model_log, which=1)
> # 2. Independence of errors (Auto correlation)
> dwtest(model_log)

        Durbin-Watson test

data:  model_log
DW = 2.003, p-value = 0.3453
alternative hypothesis: true autocorrelation is greater than 0

> # 3. Homoscedasticity
> bptest(model_log)

        studentized Breusch-Pagan test

data:  model_log
BP = 9.7785, df = 8, p-value = 0.2809

> plot(model_log, which=2)
> # 4. Normality of errors
> shapiro.test(resid(model_log))

        Shapiro-Wilk normality test

data:  resid(model_log)
W = 0.92997, p-value = 0.1542

> # 5. Multicollinearity
> vif(model_log)
      Commute        Literacy       JobGrowth      Physicians        RapeRate
     3.490046        1.403594        1.814860        3.254180        1.792495
  Restaurants       MedianAge HouseholdIncome
     1.473944        2.004123        1.484061
```

And the corresponding linearity plot



- Check for outliers



We can see the outliers of our dataset in the above Cook's distance graph

Henceforth, our new model is a better fit than our earlier linear regression model.

R - Code

```r
# load necessary packages for regression
library(car)
library(lmtest)
library(corrplot)
library(readr)
library(MASS)


# Load data
data <- read_delim("/Users/knvardhan/Desktop/final.csv", delim = ",")

# remove 1st column
data <- data[, -1]

# head of the data
head(data)

# change column names
names(data) <- c("Income", "Commute", "Literacy","JobGrowth", "Physicians",
"RapeRate", "Restaurants", "Housing", "MedianAge", "HouseholdIncome")

# summary of the data
summary(data)

# correlation matrix and plot
corr_matrix <- cor(data)
corrplot(corr_matrix, method = "color")

# histogram plot of dependent variable - Housing
hist(data$Housing, main="Histogram Plot of Dependent Variable",
    xlab = "Dependent Variable - Housing", ylab = "Frequency")


# fit linear regression model on the dataset
model <- lm(Housing ~ ., data = data)

# summary of the model
summary(model)

# Checking the correctness of the assumptions
# 1. Linearity
# 2. Independence of errors
```

```r
# 3. Homoscedasticity
# 4. Normality of Errors
# 5. Multicollinearity

# 1. Linearity
plot(model, which=1)

# 2. Independence of errors (Auto correlation)
# Performing Durbin-Watson test for Auto correlation
# Durbin-Watson statistic close to 2 implies no auto correlation
dwtest(model)


# 3. Homoscedasticity
# Performing Breusch-Pagan test for Homoscedasticity
# Null hypothesis: Homoscedasticity
# Alternate hypothesis: Heteroscedasticity
bptest(model)

# 4. Normality of errors
# Performing Shapiro-Wilk test for Normality of errors
# Null hypothesis: Normality of errors
# Alternate hypothesis: Non-normality of errors
# Q-Q plot
plot(model, which=2)
shapiro.test(resid(model))


# correction for normality
# Log transformation
model_log <- lm(log(Housing) ~ ., data = data)

# updated model summary
summary(model_log)

# 5. Multicollinearity
# Performing Variance Inflation Factor (VIF) test for Multicollinearity
# VIF > 10 implies Multicollinearity
vif(model_log)

# Income has high VIF value
# Removing Income from the model
```

```r
new_data <- data[, -1]
model_log <- lm(log(Housing) ~ ., data = new_data)

#check VIF
vif(model_log)

# updated model summary
summary(model_log)

# checking all the assumptions once again on the updated final model
# 1. Linearity
plot(model_log, which=1)

# 2. Independence of errors (Auto correlation)
dwtest(model_log)

# 3. Homoscedasticity
bptest(model_log)

# 4. Normality of errors
shapiro.test(resid(model_log))

# 5. Multicollinearity
vif(model_log)

# All the assumptions are satisfied

# check for outliers
# Cook's distance
plot(model_log, which=4)
```