Final Examination ~ MA4142

~Varunaditya Singhal
~MA20BTECH11021

From the data given in the R first task is loading all the necessary packages. Below are the same :

```
# Load necessary libraries

library("dplyr")

library("car")

library("readr")

library("corrplot")

library("lmtest")
```

After changing the working directory the next task was to load the data , since the First attribute(City) is a character datatype, we remove it cause in later stages index no. can be mapped to the city name (A – T). Below is the code for the same :

```
data <- read_delim("D:/IITH MnC/Sem 6/Regression/cities.csv", show_col_types = FALSE)

data <- data[,-1]

names(data) <- c("Income",      "Commute","Literacy","Job_Growth",   "Physicians",
          "Rape_Rate","Restaurants","Housing","Median_Age", "Household_Income")
```

The updated data looks like :

	Income	Commute	Literacy	Job_Growth	Physicians	Rape_Rate	Restaurants	Housing	Median_Age	Household_Income
1	26000	49.2	5.15	10.8	1987	51.3	5582	109400	35.3	68000
2	29300	45.3	5.97	9.5	517	50.8	9988	97000	43.2	70400
3	24800	39.8	9.41	8.2	592	77.7	20511	114700	29.5	60500
4	27900	46.8	4.61	7.6	3310	51.2	8946	99100	40.5	65900
5	37500	39.9	5.64	12.2	975	40.1	4000	122200	47.1	84700
6	31900	49.5	4.80	7.7	2238	38.0	8970	145300	39.3	75800
7	25300	44.4	6.84	5.4	611	38.8	9570	99500	38.6	62600
8	22000	44.8	2.79	6.2	272	65.7	19101	76400	41.6	54800
9	29400	44.9	4.48	7.8	381	48.7	12099	112500	41.8	72900
10	42400	44.7	5.16	8.0	1812	45.4	10953	143500	41.2	100000
11	40500	40.0	6.41	10.9	294	69.6	2655	173600	41.7	102000
12	24700	38.7	1.66	9.0	196	19.0	15796	129200	33.4	65300
13	24400	41.1	5.60	8.7	404	77.2	16001	126500	30.6	62200
14	22400	42.8	2.16	8.3	534	57.9	16712	102700	34.5	59200
15	22200	37.8	2.72	8.4	166	50.9	11856	110300	35.4	57100
16	27500	48.4	4.03	8.1	1553	83.6	12348	107400	34.3	72000
17	23100	44.5	2.07	4.7	502	42.7	65804	116000	38.5	59400
18	25000	41.4	3.61	13.9	172	17.8	36151	120000	52.7	57300
19	25800	53.5	5.03	5.3	4143	57.4	14310	132800	36.2	71900

I did summary statistics for the initial data :

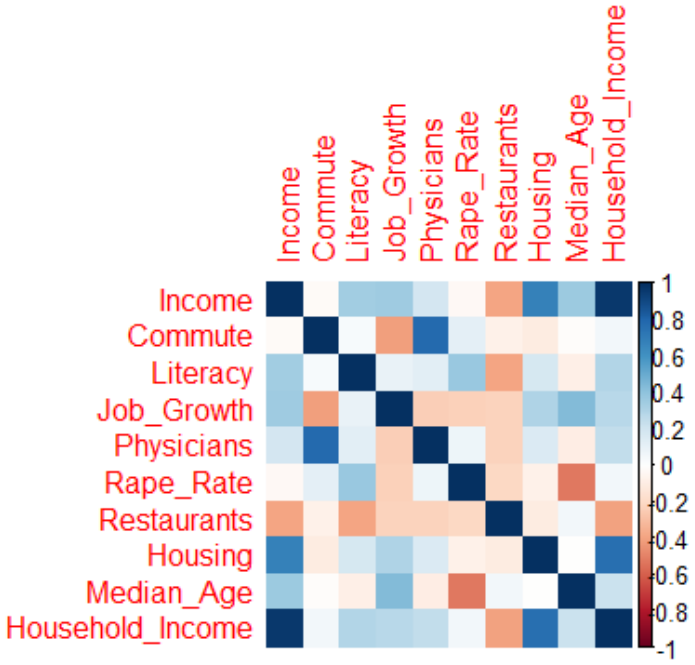
```
> # Summary statistics
> summary(data)
     Income         Commute         Literacy       Job_Growth      Physicians      Rape_Rate
 Min.   :22000   Min.   :37.80   Min.   :1.660   Min.   : 4.700   Min.   : 166.0   Min.   :17.80
 1st Qu.:24075   1st Qu.:40.83   1st Qu.:3.405   1st Qu.: 7.325   1st Qu.: 359.2   1st Qu.:42.05
 Median :25550   Median :44.60   Median :4.915   Median : 8.150   Median : 530.0   Median :51.05
 Mean   :27735   Mean   :44.12   Mean   :4.671   Mean   : 8.360   Mean   :1059.2   Mean   :51.80
 3rd Qu.:29325   3rd Qu.:45.67   3rd Qu.:5.610   3rd Qu.: 9.125   3rd Qu.:1617.8   3rd Qu.:59.85
 Max.   :42400   Max.   :53.50   Max.   :9.410   Max.   :13.900   Max.   :4143.0   Max.   :83.60
  Restaurants       Housing        Median_Age    Household_Income
 Min.   : 2655   Min.   : 76400   Min.   :29.50   Min.   : 54000
 1st Qu.: 8964   1st Qu.:101900   1st Qu.:35.10   1st Qu.: 59350
 Median :11978   Median :113600   Median :38.95   Median : 65600
 Mean   :15512   Mean   :116230   Mean   :38.84   Mean   : 68800
 3rd Qu.:16179   3rd Qu.:127175   3rd Qu.:41.62   3rd Qu.: 72225
 Max.   :65804   Max.   :173600   Max.   :52.70   Max.   :102000
>
```

From here we can make many conclusions

1) The minimum salary of all the cities is 22,000 which is fairly decent for someone who is looking for house.
2) None of the cities are free from Rape a minimum of 17.80 rate exists and may shoot up to an incredibly high no of 83.60.
3) None of the city seems to be in outpost of the districts , since they have good no of Housing and Restaurants located.
4) None of the cities are of new age , which is good for our analysis since new cities have high rate of Housing.

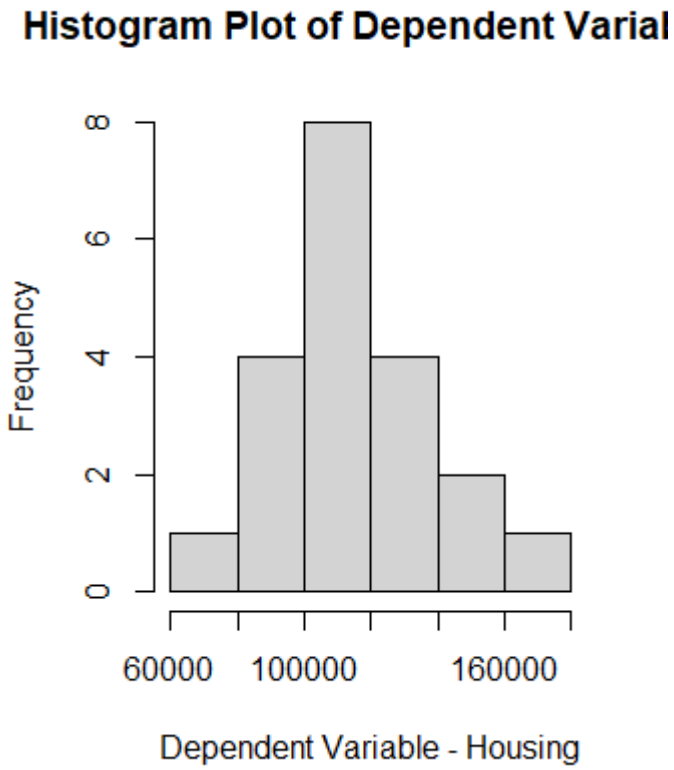
Next, I plotted the Correlation matrix and made some conclusions from it :

```
# Correlation matrix and plot

corr_matrix <- cor(data)

corrplot(corr_matrix, method = "color")
```



1) We can infer from this that for Housing income is highly correlated and commute being the least correlated.
2) Literacy is highly correlated with Income(which make sense) , but it is with the Rape Rates as well, which is a very fascinating observation.
3) If Literacy rate is high then the restaurants are less correlated.

Histogram of our dependent variable

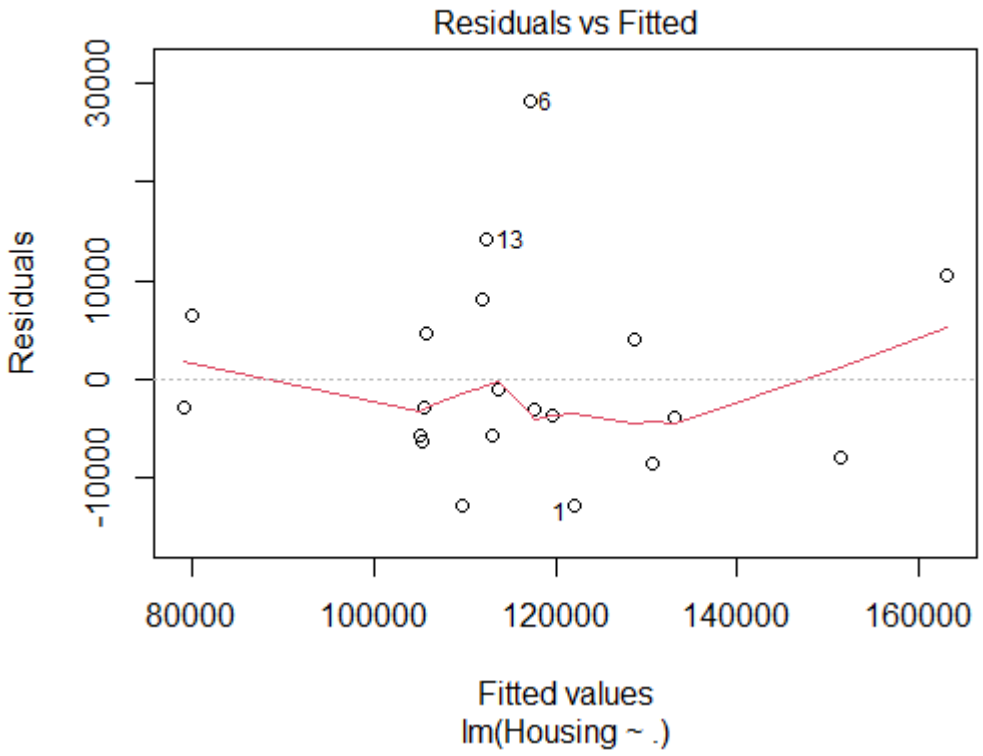


We made our initial model with Housing as dependent variable and observe the summary for it :

```
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
Job_Growth        2057.0238  1847.6049   1.113   0.2916
Physicians           4.1323     5.1789   0.798   0.4435
Rape_Rate         -336.7469   242.5468  -1.388   0.1952
Restaurants          0.5078     0.2756   1.843   0.0952 .
Median_Age        -767.2750  1053.0830  -0.729   0.4830
Household_Income     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
```

We check the assumptions for our model :

Linearity :



We can assume our model to be linear.

Autocollenearity :

```
        Durbin-Watson test

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

The result is near 2 and we conclude the model has no aurocollenearity.

Homosedastic :

```
        studentized Breusch-Pagan test

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

We conclude out model is Homodedastic.

Updated Summary :

```
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
Job_Growth        2057.0238  1847.6049   1.113   0.2916
Physicians           4.1323     5.1789   0.798   0.4435
Rape_Rate         -336.7469   242.5468  -1.388   0.1952
Restaurants          0.5078     0.2756   1.843   0.0952 .
Median_Age        -767.2750  1053.0830  -0.729   0.4830
Household_Income     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
```

Normality :

```
        Shapiro-Wilk normality test

data:  data$Housing
W = 0.96649, p-value = 0.6797
```

The test claims to have normal distribution on the dependent variable.

Multicolleniarity :

```
> vif(model)
          Income          Commute          Literacy       Job_Growth
       42.152047         4.096666         1.799669         1.841323
      Physicians        Rape_Rate      Restaurants       Median_Age
        3.392345         1.815191         1.479778         3.411101
Household_Income
       35.332742
>
```

VIF values >5 shows that they are collinear to other independent variables.

After deleting Icome , updated VIF values

```
> vif(final_model)
       Commute        Literacy      Job_Growth       Physicians
      3.490046        1.403594        1.814860        3.254180
     Rape_Rate     Restaurants      Median_Age Household_Income
      1.792495        1.473944        2.004123        1.484061
```

Thus our updated model doesn't show any multicollinearity.

Final Model Sumath

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

Residuals:
   Min     1Q Median     3Q    Max
-14192  -8060  -2760   7704  20633

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      90969.5728 68174.9232   1.334 0.209052
Commute           -600.5931  1498.4966  -0.401 0.696245
Literacy           712.1475  2073.8964   0.343 0.737781
Job_Growth        2358.1536  1903.6938   1.239 0.241230
Physicians           2.7113     5.2643   0.515 0.616712
Rape_Rate         -299.8756   250.1468  -1.199 0.255794
Restaurants          0.5313     0.2854   1.861 0.089595 .
Median_Age       -1686.7585   837.7395  -2.013 0.069188 .
Household_Income     1.4341     0.2942   4.874 0.000491 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14190 on 11 degrees of freedom
Multiple R-squared:  0.7605,    Adjusted R-squared:  0.5863
F-statistic: 4.366 on 8 and 11 DF,  p-value: 0.01357
```