Ecotrix Project

2024-03-25

Aakansha Rawat

Basics

Q1: Write down all code with the interpretation of the result. (Vector, Matrix, Summary of Vector and Matrix, make data frame (with 20 columns), Summary of data frame, Interpretation of Summary)

a) Vector

eg 1:

```
vector1<-c(1:8)   ;   vector1
```

```
## [1] 1 2 3 4 5 6 7 8
```

```
summary(vector1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.75    4.50    4.50    6.25    8.00
```

Here we have created a vector named “vector1” which is from 1 to 8; And some results of this is as follows :

Min of vector1 is **1**

Max of vector1 is **8**

Median (Q2 i.e 50th percentile) of vector1 is **4.5**

Mean of vector1 is **4.5**

First quartile (Q1) i.e 25th percentile of vector1 is **2.75**

Third quartile (Q3) i.e 75th percentile of vector1 is **6.25**

eg 2:

```
vector2<-seq(1,20,2)
vector2
```

```
##  [1]  1  3  5  7  9 11 13 15 17 19
```

```
summary(vector2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     5.5    10.0    10.0    14.5    19.0
```

Here we have created a vector named "vector2" which is from 1 to 20 with gap of 2 each; And

Min of vector2 is **1**

Max of vector2 is **19**

Median of vector2 is **10**

Mean of vector2 is **10**

First quartile (Q1) i.e 25th percentile of vector2 is **5.5**

Third quartile (Q3) i.e 75th percentile of vector2 is **14.5**

eg 3:

```
vector3<-rep(c(1:5),2)
vector3
```

```
##  [1] 1 2 3 4 5 1 2 3 4 5
```

```
summary(vector3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2       3       3       4       5
```

Here we have created a vector named “vector3” which is from 1 to 5 with repetition of each term twice; And

Min of vector3 is **1**

Max of vector3 is **5**

Mean of vector3 is **3**

Median of vector3 is **3**

First quartile (Q1) i.e 25th percentile of vector3 is **2**

Third quartile (Q3) i.e 75th percentile of vector3 is **4**

b) Matrix

eg 1:

```
matrix1<-matrix(1:20,5,4)
matrix1
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    6   11   16
## [2,]    2    7   12   17
## [3,]    3    8   13   18
## [4,]    4    9   14   19
## [5,]    5   10   15   20
```

```
summary(matrix1)
```

```
##        V1          V2          V3          V4
##  Min.   :1   Min.   : 6   Min.   :11   Min.   :16
##  1st Qu.:2   1st Qu.: 7   1st Qu.:12   1st Qu.:17
##  Median :3   Median : 8   Median :13   Median :18
##  Mean   :3   Mean   : 8   Mean   :13   Mean   :18
##  3rd Qu.:4   3rd Qu.: 9   3rd Qu.:14   3rd Qu.:19
##  Max.   :5   Max.   :10   Max.   :15   Max.   :20
```

Here we have created a matrix named "matrix1" with elements from 1 to 20. By default the matrix is formed with column dominance; And No of columns of matrix1 is **4**

No of rows of matrix1 is **5**

Other than that here all other statistic is calculated for each column Notice that name of each column is V1, V2, … this is default.

Taking an example like for V3:-

Min of V3 is **`min(matrix1[, 3])`**

Max of V3 is **15**

Mean of V3 is **13**

Median of V3 is **13**

First quartile (Q1) i.e 25th percentile of V3 is **11, 12, 13, 14, 15**

Third quartile (Q3) i.e 75th percentile of V3 is **11, 12, 13, 14, 15**

eg 2:

```
matrix2<-matrix(1:20,5,4,byrow= TRUE)
matrix2
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
## [4,]   13   14   15   16
## [5,]   17   18   19   20
```

```
summary(matrix2)
```

```
##        V1           V2           V3           V4
##  Min.   : 1   Min.   : 2   Min.   : 3   Min.   : 4
##  1st Qu.: 5   1st Qu.: 6   1st Qu.: 7   1st Qu.: 8
##  Median : 9   Median :10   Median :11   Median :12
##  Mean   : 9   Mean   :10   Mean   :11   Mean   :12
##  3rd Qu.:13   3rd Qu.:14   3rd Qu.:15   3rd Qu.:16
##  Max.   :17   Max.   :18   Max.   :19   Max.   :20
```

```
colnames(matrix2) <- c("alpha", "beta", "gamma", "sigma")
matrix2
```

```
##      alpha beta gamma sigma
## [1,]     1    2     3     4
## [2,]     5    6     7     8
## [3,]     9   10    11    12
## [4,]    13   14    15    16
## [5,]    17   18    19    20
```

Here we have created a matrix named “matrix2” with elements from 1 to 20. Now to change the default setting we have used an argument “byrow” ; And

No of columns of matrix2 is **4**

No of rows of matrix2 is **5**

Also i have named the columns instead of default V1, V2…

c) Data Frame

```
head(dtfrm1)
```

```
##   S_no            Order_ID     Date Age Gender    Status Category Qty currency
## 1    1 402-5239558-4281957 04-05-22  44  Women   Shipped      Set   8      INR
```

```
## 2    2 403-4106553-1691525 04-23-22  29  Women Cancelled      Top   6      INR
## 3    3 407-3483041-9133135 04-16-22  67  Women Cancelled   Bottom   3      INR
## 4    4 406-7195963-2374765 05-02-22  20  Women   Shipped      Set   5      USD
## 5    5 404-4712917-7300332 06-06-22  62  Women   Shipped    Saree   2      USD
## 6    6 408-9885439-6698733 04-13-22  49    Men   Pending      Set   3      INR
##   Amount Size      shipstate ship_service_level ship_postal_code
## 1   5584   XS ANDHRA PRADESH          Expedited           522007
## 2   2295  3XL         PUNJAB           Standard           144022
## 3   4235    M      RAJASTHAN           Standard           302012
## 4   3036    M  UTTAR PRADESH          Expedited           226002
## 5   2894  3XL    MAHARASHTRA          Expedited           400081
## 6   2864    S      KARNATAKA          Expedited           560032
```

```
summary(dtfrm1)
```

```
##       S_no        Order_ID             Date                Age
##  Min.   : 1.0   Length:19          Length:19          Min.   :18.00
##  1st Qu.: 5.5   Class :character   Class :character   1st Qu.:26.50
##  Median :10.0   Mode  :character   Mode  :character   Median :44.00
##  Mean   :10.0                                         Mean   :44.05
##  3rd Qu.:14.5                                         3rd Qu.:57.00
##  Max.   :19.0                                         Max.   :76.00
##     Gender             Status            Category              Qty
##  Length:19          Length:19          Length:19          Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :3.000
##                                                           Mean   :3.368
##                                                           3rd Qu.:4.500
##                                                           Max.   :9.000
##    currency             Amount         Size            shipstate
##  Length:19          Min.   :2295   Length:19          Length:19
##  Class :character   1st Qu.:2598   Class :character   Class :character
##  Mode  :character   Median :2664   Mode  :character   Mode  :character
##                     Mean   :2910
##                     3rd Qu.:2862
##                     Max.   :5584
```

```
##  ship_service_level ship_postal_code
##  Length:19          Min.   :123029
##  Class :character   1st Qu.:316708
##  Mode  :character   Median :410206
##                     Mean   :401288
##                     3rd Qu.:500030
##                     Max.   :700016
```

Now here i have created a data frame of random amazon sales data with S.no, order_ID, Date, Age, Gender, Status, Category, Qty, currency, Amount, Size, shipstate, ship_service_level and ship_postal_code as its header.

Here apart from numerical data I also have categorical data for eg:
gender and currency is a type of a [1]**categorical data;** whereas qty and and age is a [2]**numerical type** of data.
And in summary it has given us different values of statistics similar to a matrix and a vector.

Q2 : Explain process of importing data from Windows (CSV file, Excel file) and packages. Interpret imported data with a summary. Plot histogram, box plot, frequency density.

```
## [1] "C:/Users/Hamada Salim G Trd/Desktop"
```

```
## [1] "C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming"
```

The process of importing data from csv/excel file is as follows:

1. Download any dataset online or made up in csv/excel format and save it in your directory. In this case i have a csv/excel file named wage1/Amazon_sale_report stored in my [3]**directory** i.e a folder in desktop named "coding_or_programming"

2. Then simply using the inbuilt command i.e read.csv()/read_excel() (now here's the catch we cant use read.excel in markdown for this purpose we have read_excel now in order to use it, you need to download the package named '**readx1**') with name of the file you want to import in "" (within brackets).

3. If for example your file is not saved in the directory but somewhere else then simply enter the full path of the document rather then only its name.

a) csv file

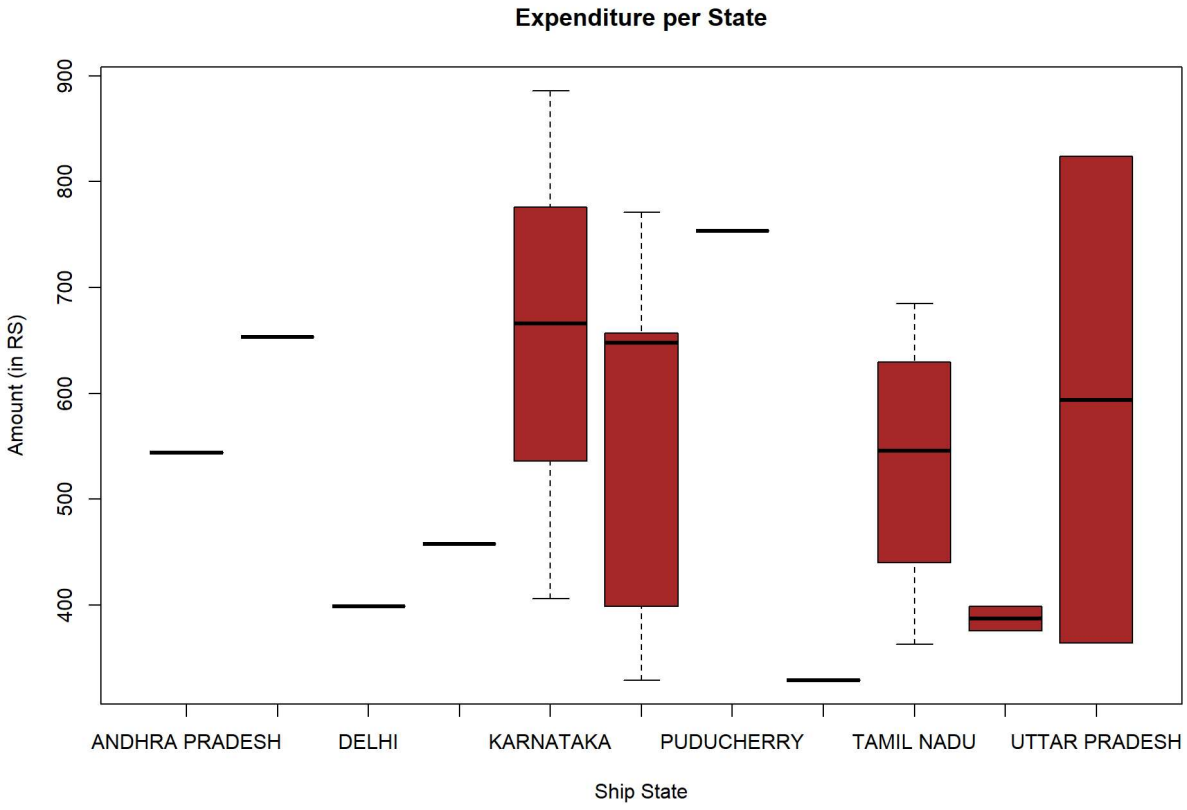
```
ds1<-read.csv("C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming/Amazon2.csv
head(ds1)
```

```
##              Order.ID     Date                        Status Fulfilment
## 1 405-8078784-5731545 04-30-22                     Cancelled   Merchant
## 2 171-9198151-1101146 04-30-22 Shipped - Delivered to Buyer   Merchant
## 3 404-0687676-7273146 04-30-22                       Shipped     Amazon
## 4 403-9615377-8133951 04-30-22                     Cancelled   Merchant
## 5 407-1069790-7240320 04-30-22                       Shipped     Amazon
## 6 404-1490984-4578765 04-30-22                       Shipped     Amazon
##   ship.service.level      Category Size Courier.Status Qty currency Amount
## 1           Standard           Set    S                  0      INR 647.62
## 2           Standard         kurta  3XL        Shipped   1      INR 406.00
## 3          Expedited         kurta   XL        Shipped   1      INR 329.00
## 4           Standard Western Dress    L                  0      INR 753.33
## 5          Expedited           Top  3XL        Shipped   1      INR 574.00
## 6          Expedited           Set   XL        Shipped   1      INR 824.00
##     ship.city    ship.state ship.postal.code
## 1      MUMBAI   MAHARASHTRA           400081
## 2   BENGALURU     KARNATAKA           560085
## 3 NAVI MUMBAI   MAHARASHTRA           410210
## 4  PUDUCHERRY    PUDUCHERRY           605008
## 5     CHENNAI    TAMIL NADU           600073
## 6   GHAZIABAD UTTAR PRADESH           201102
```

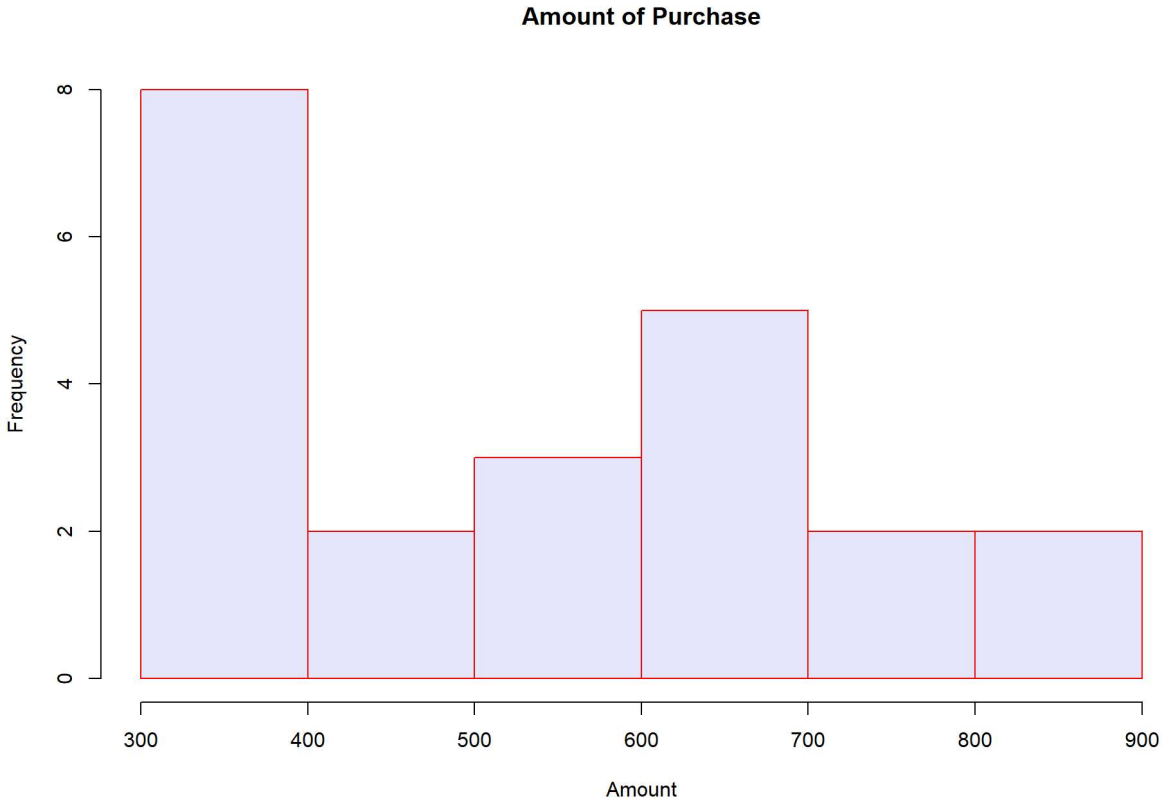
```
summary(ds1)
```

```
##    Order.ID             Date              Status           Fulfilment
##  Length:23          Length:23          Length:23          Length:23
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  ship.service.level   Category             Size           Courier.Status
##  Length:23          Length:23          Length:23          Length:23
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       Qty           currency             Amount       ship.city
##  Min.   :0.0000   Length:23          Min.   :329.0   Length:23
##  1st Qu.:1.0000   Class :character   1st Qu.:399.0   Class :character
##  Median :1.0000   Mode  :character   Median :530.5   Mode  :character
##  Mean   :0.8696                      Mean   :545.5
##  3rd Qu.:1.0000                      3rd Qu.:663.8
##  Max.   :1.0000                      Max.   :886.0
##                                      NA's   :1
##   ship.state        ship.postal.code
##  Length:23          Min.   :110074
##  Class :character   1st Qu.:351037
##  Mode  :character   Median :500008
##                     Mean   :429253
##                     3rd Qu.:560063
##                     Max.   :620018
##
```

```r
data(ds1, package = "MASS")
boxplot(Amount~ship.state, data = ds1, main= "Expenditure per State",
        xlab = "Ship State", ylab = "Amount (in RS)", col="brown")
```




```r
hist(ds1$Amount, main = "Amount of Purchase", xlab = "Amount", col = "lavender", bo
```



b) excel file

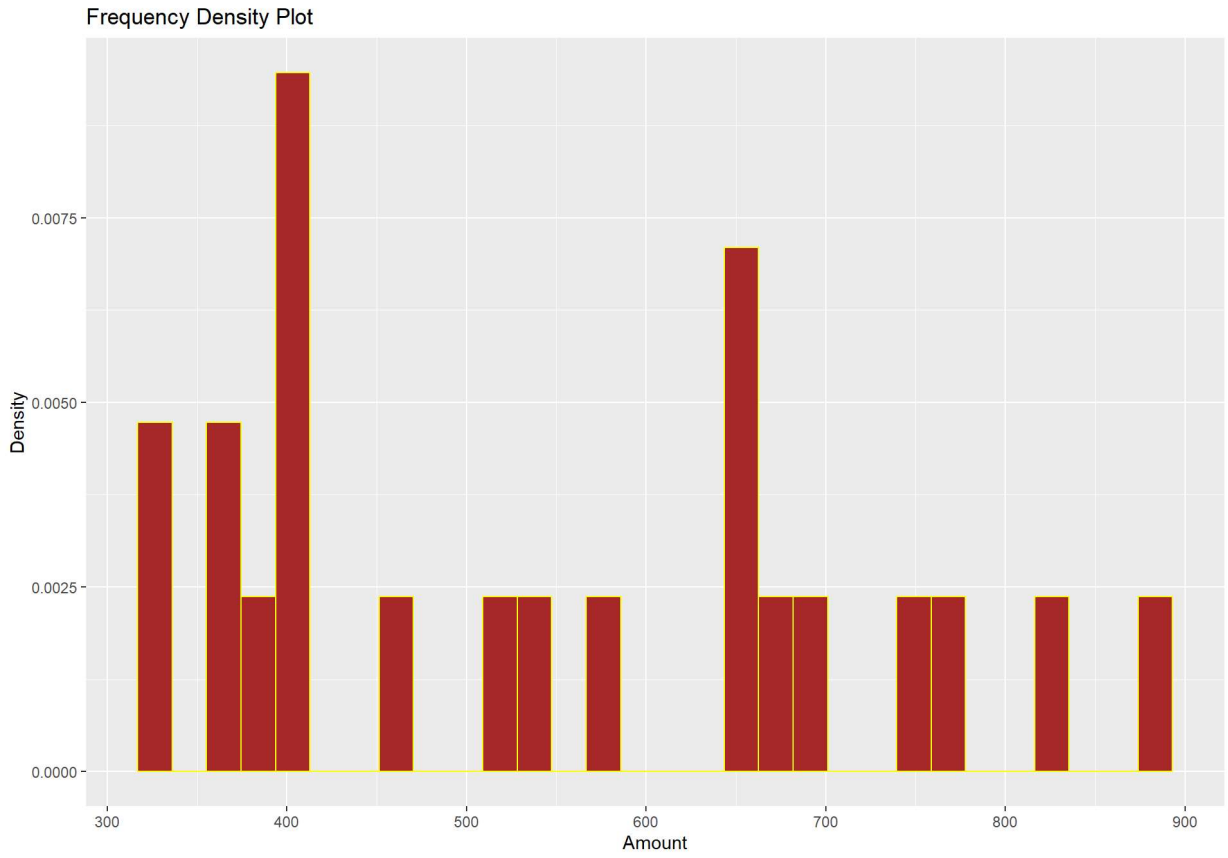
```
ds2<-read_excel("C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming/Amazon1.x
head(ds2)
```

```
## # A tibble: 6 × 11
##   `Order ID`          Date   Status Category Size  `Courier Status`   Qty Amount
##   <chr>               <chr>  <chr>  <chr>    <chr> <chr>            <dbl>  <dbl>
## 1 405-8078784-5731545 04-30… Cance… Set      S     <NA>                 0   648.
## 2 171-9198151-1101146 04-30… Shipp… kurta    3XL   Shipped              1   406
## 3 404-0687676-7273146 04-30… Shipp… kurta    XL    Shipped              1   329
## 4 403-9615377-8133951 04-30… Cance… Western… L     <NA>                 0   753.
## 5 407-1069790-7240320 04-30… Shipp… Top      3XL   Shipped              1   574
## 6 404-1490984-4578765 04-30… Shipp… Set      XL    Shipped              1   824
## # ℹ 3 more variables: `ship-city` <chr>, `ship-state` <chr>,
## #   `ship-postal-code` <dbl>
```

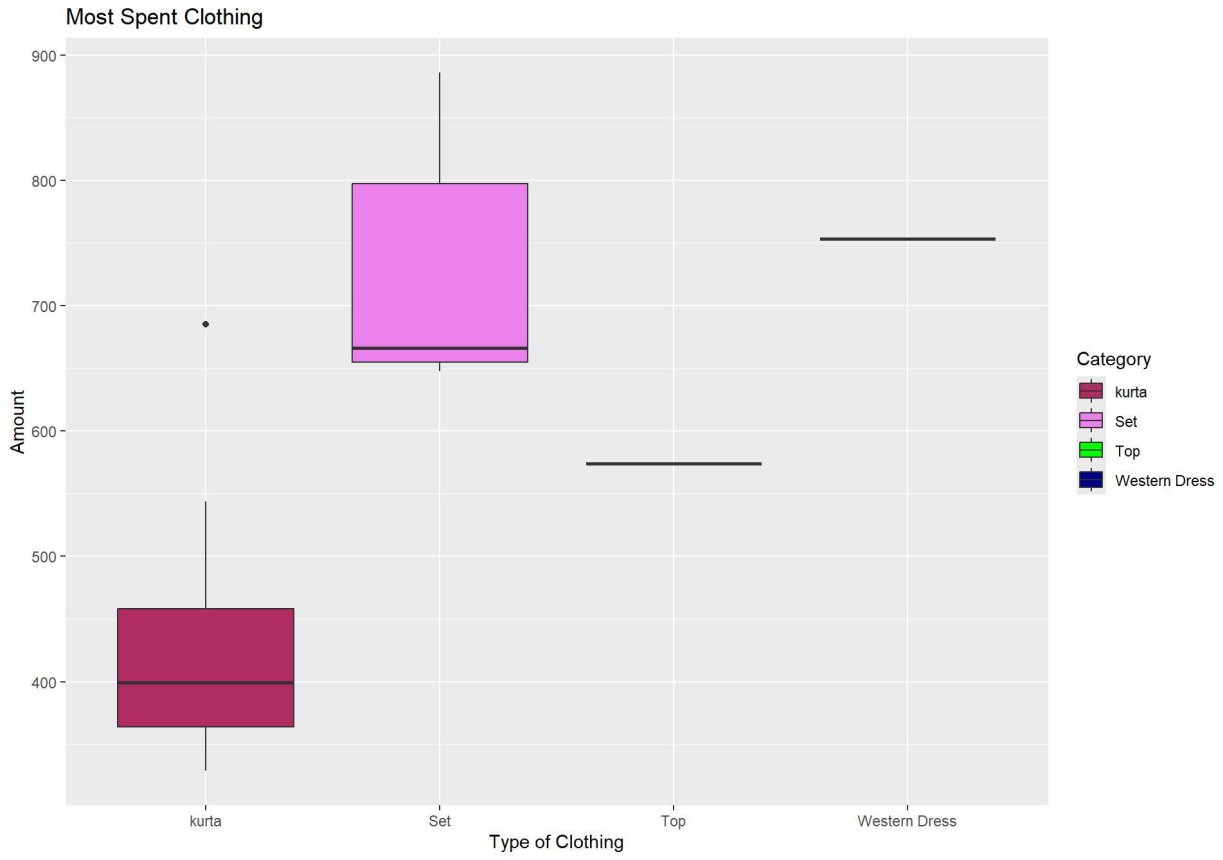
```r
summary(ds2)
```

```
##    Order ID             Date              Status            Category
##  Length:23          Length:23          Length:23          Length:23
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      Size           Courier Status          Qty             Amount
##  Length:23          Length:23          Min.   :0.0000   Min.   :329.0
##  Class :character   Class :character   1st Qu.:1.0000   1st Qu.:399.0
##  Mode  :character   Mode  :character   Median :1.0000   Median :530.5
##                                        Mean   :0.8696   Mean   :545.5
##                                        3rd Qu.:1.0000   3rd Qu.:663.8
##                                        Max.   :1.0000   Max.   :886.0
##                                                         NA's   :1
##   ship-city          ship-state        ship-postal-code
##  Length:23          Length:23          Min.   :110074
##  Class :character   Class :character   1st Qu.:351037
##  Mode  :character   Mode  :character   Median :500008
##                                        Mean   :429253
##                                        3rd Qu.:560063
##                                        Max.   :620018
##
```

```r
ggplot(data = ds2, aes(x = Amount)) +
  geom_histogram(aes(y = ..density..), bins = 30, color = "yellow", fill = "brown",
  labs(title = "Frequency Density Plot",
       x = "Amount",
       y = "Density")
```




```
my_colors <- c("maroon", "violet", "green", "darkblue", "black")
ggplot(data = ds2, aes(y = Amount, x = Category, fill = Category)) +
  geom_boxplot() +
  labs(title = "Most Spent Clothing", y = "Amount", x = "Type of Clothing") +
  scale_fill_manual(values = my_colors)
```



Linear Regression

Q3 :Take other imported data and do correlation with interpretation, simple linear regression with interpretation (Check and interpret: Intercept and Slope, variance of parameter and error term, t-test, p-value, Adj R square, F test). Plot regression model on the graph.

```
ds3<-read.csv("C:/Users/Hamada Salim G Trd/Desktop/coding_Or_progrmming/HTV.csv")
head(ds3)
```

```
##        wage     abil educ ne nc west south exper motheduc fatheduc brkhme14
## 1 12.019231 5.027738   15  0  0    1     0     9       12       12        0
## 2  8.912656 2.037170   13  1  0    0     0     8       12       10        1
## 3 15.514334 2.475895   15  1  0    0     0    11       12       16        0
## 4 13.333333 3.609240   15  1  0    0     0     6       12       12        0
## 5 11.070110 2.636546   13  1  0    0     0    15       12       15        1
## 6 17.482517 3.474334   18  1  0    0     0     8       12       12        0
##   sibs urban ne18 nc18 south18 west18 urban18   tuit17    tuit18    lwage
## 1    1     1    1    0       0      0       1 7.582914  7.260242 2.486508
```

```
## 2    4     1    1    0       0      0       1 8.595144  9.499537 2.187472
## 3    2     1    1    0       0      0       1 7.311346  7.311346 2.741764
## 4    1     1    1    0       0      0       1 9.499537 10.162070 2.590267
## 5    2     1    1    0       0      0       1 7.311346  7.311346 2.404249
## 6    2     1    1    0       0      0       1 7.311346  7.311346 2.861201
##   expersq      ctuit
## 1      81 -0.3226714
## 2      64  0.9043922
## 3     121  0.0000000
## 4      36  0.6625338
## 5     225  0.0000000
## 6      64  0.0000000
```

So here first i have imported a file named "HTV.csv" from my directory. This data consists of **1230** observations with **23** variables.

```
sm<-lm(wage ~ educ, data = ds3)
sm
```

```
##
## Call:
## lm(formula = wage ~ educ, data = ds3)
##
## Coefficients:
## (Intercept)         educ
##      -4.857        1.392
```

```
summary(sm)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = ds3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -17.370  -4.558  -1.267   2.579  69.722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.8574     1.3601  -3.571 0.000369 ***
## educ          1.3918     0.1027  13.557  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.474 on 1228 degrees of freedom
## Multiple R-squared:  0.1302, Adjusted R-squared:  0.1295
## F-statistic: 183.8 on 1 and 1228 DF,  p-value: < 2.2e-16
```

```
pv1<-vcov(sm)
pv1
```

```
##             (Intercept)        educ
## (Intercept)   1.8499325 -0.13741670
## educ         -0.1374167  0.01054019
```

```
variances1 <- diag(pv1)
variances1
```

```
## (Intercept)        educ
##  1.84993251  0.01054019
```

```
anova(sm)
```

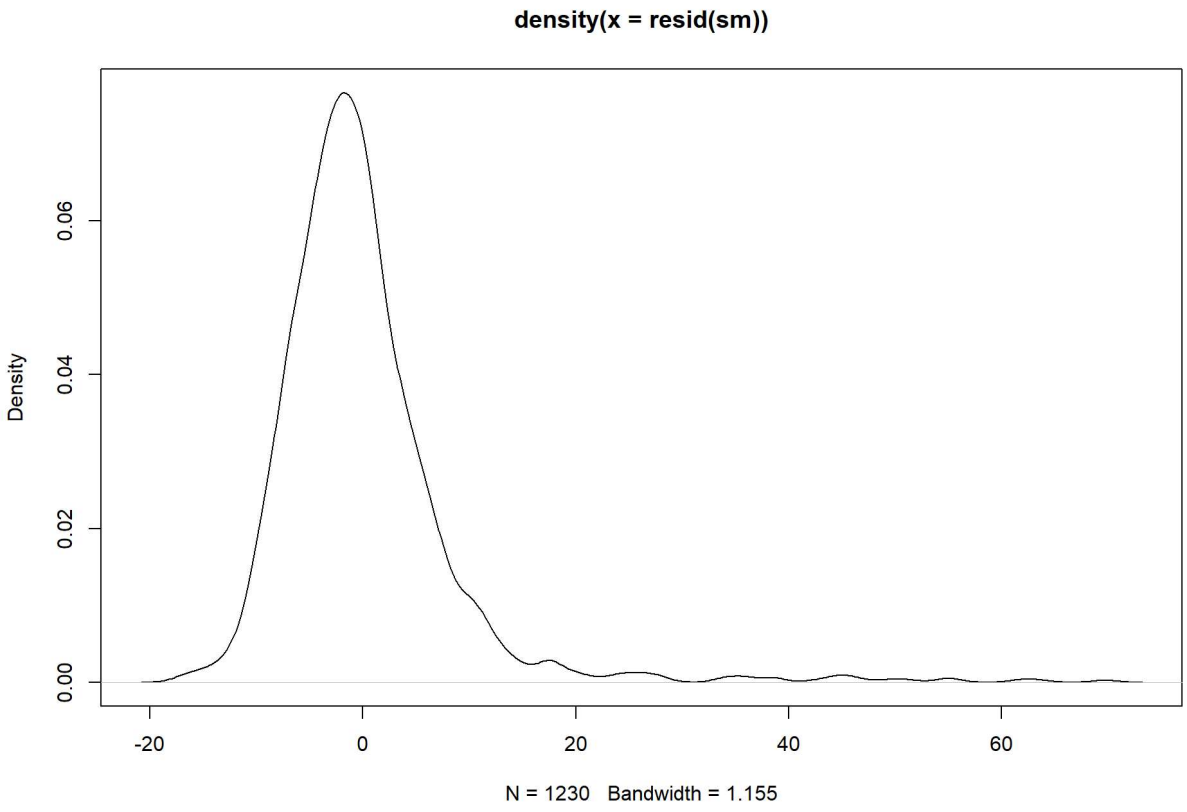
```
## Analysis of Variance Table
##
## Response: wage
##             Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## educ         1  13196 13196.3  183.79 < 2.2e-16 ***
## Residuals 1228  88174    71.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
resid1<-resid(sm)
var(resid1)
```

```
## [1] 71.74432
```

```
plot(density(resid(sm)))
```



Process:

- Now i have first formed a linear regression model named "sm" with Y = B_1 + B_2X_2 + u *as the* ***PRF*** *and* $Y\ hat = b_1 + b_2X_2$ + e is the **SRF** where Y is the wage and X is the education.

- So the summary shows us that $\mathbf{b_1 = -4.857}$ and $\mathbf{b_2 = 1.392}$ the 3 stars (***) in the intercept and educ shows that both are significant we also have residual standard error i.e sigma hat which is **8.4736501** also called as *standard error of regression* ([4]**SER**). also notice that degree of freedom is given 1228 i.e (n-2) in this case as only two unknown parameters are there which is consistent with what we studied in theory.

- r^2 value is **0.1302** i.e only 13.02% which means only this percentage of variation in Y was explained by the model also known as [5]**goodness of fit.**

- vcov() function is used to extract the variance-covariance matrix in which the diagonal elements of this matrix represent the variances of the parameter estimates. Similary, we find the variance of residuals.

- The density plot of residuals shows us that they are normally distributed.

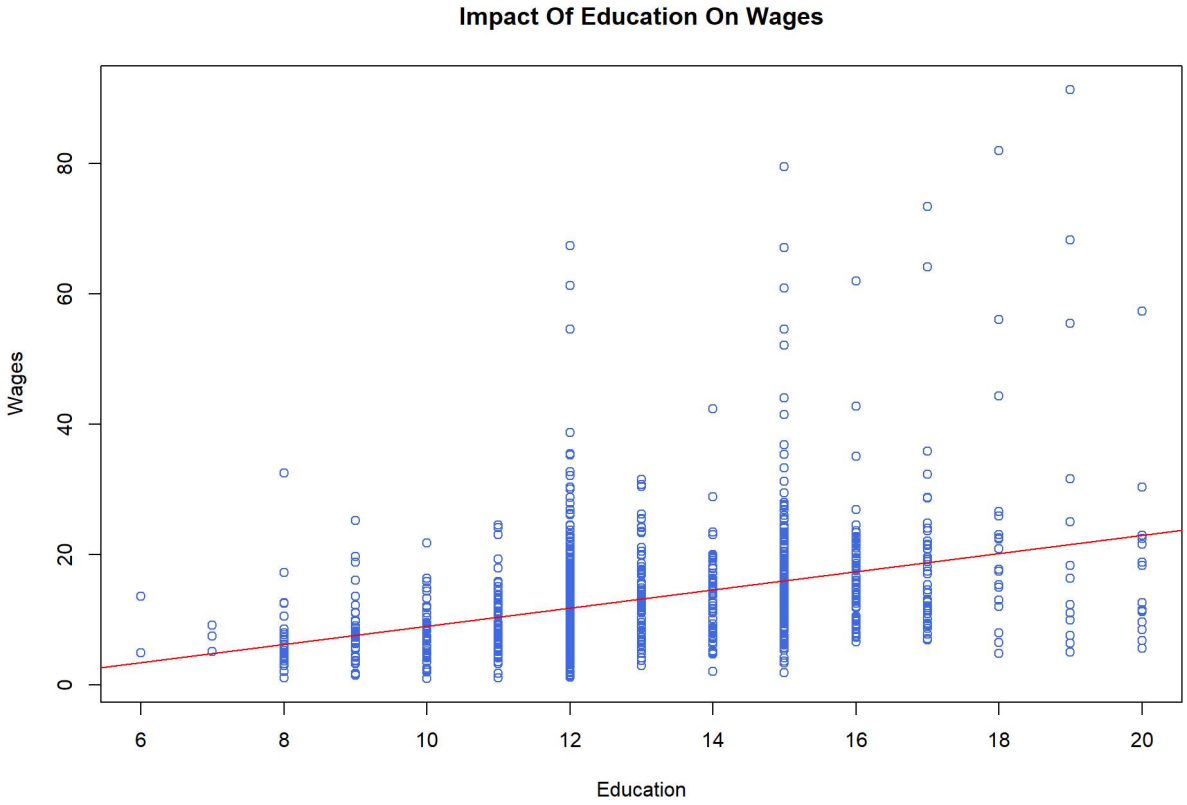
- There is the [6]**adjusted $\mathbf{R}^2$** *value* which is **0.1294712** it can be -ve or +ve in this case it is +ve, it doesn't have much significance in linear model but more in multiple model so we'll talk about it then.

- Along with this we have F stat value so now we can check the property $t^2_{\ k} = F_{1,\ k}$ where k is the no of parameters (i.e 2); Taking level of significance as (5%) alpha = 0.05

  So, $t^2_{\ 0.025,\ 2} = F_{\ 1,\ 2,\ 0.05}$ = **approx 18.513**

Now representing the model graphically :

```
plot(ds3$educ,ds3$wage,ylab = "Wages",xlab = "Education",
     main = "Impact Of Education On Wages", col = "royalblue")
abline(sm, col = "red")
```

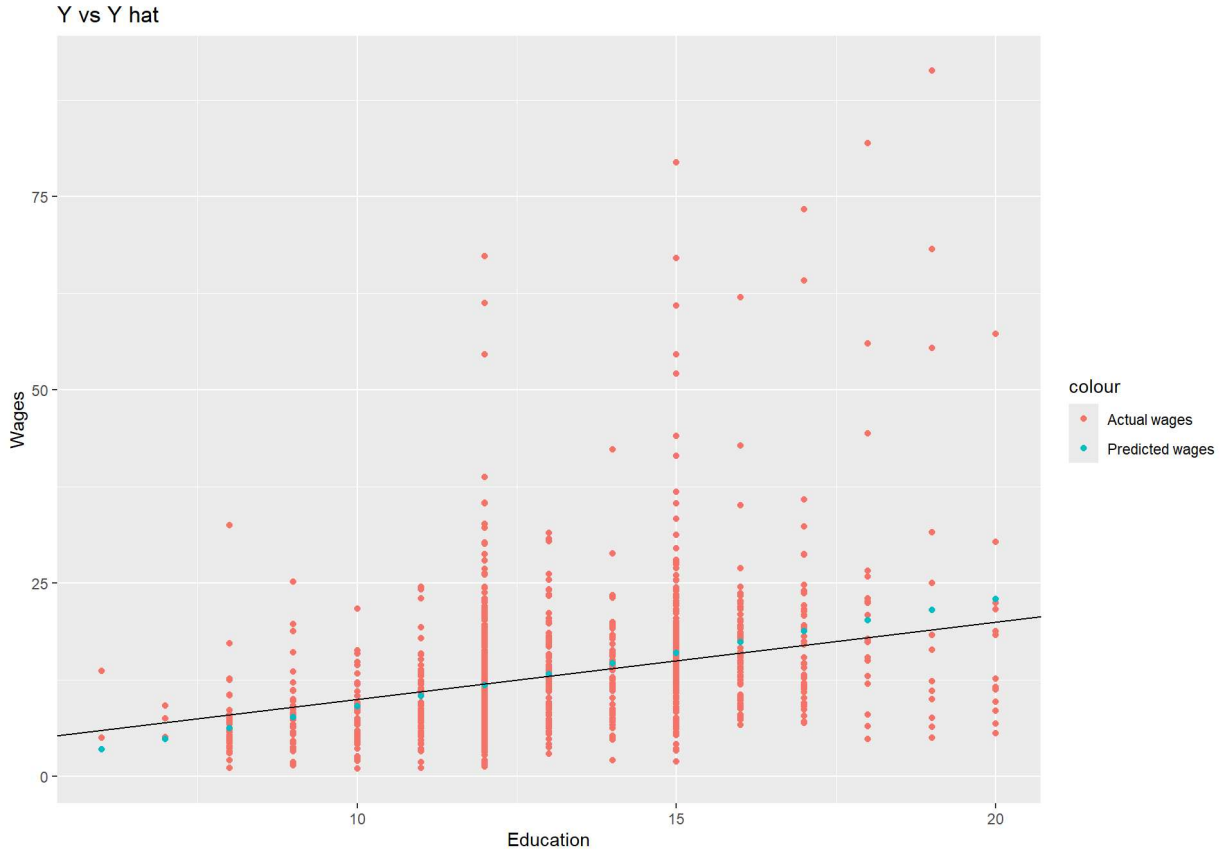



```
cor(ds3$wage,ds3$educ)
```

```
## [1] 0.360804
```

```
ds3 %<>% mutate(wage_hat_1=fitted(sm))

ggplot(data = ds3,mapping = aes(x=educ))+
  geom_point(mapping = aes(y=wage, col='Actual wages'))+
  geom_point(mapping = aes(y=wage_hat_1, col='Predicted wages'))+
  labs(title = "Y vs Y hat",
       x = "Education",
       y = "Wages")+
  geom_abline(intercept = 0,slope = 1)
```



- First I have created a simple scatter plot where the *regression line slopes* upward, it suggests a positive relationship between education and wages, indicating that individuals with higher levels of education tend to earn higher wages.
- Also if you look at the predicted values how they are so close to the regression line which shows us that the good fit of the model.

The *correlation* of wage and education also came out to be positive i.e **0.360804** which means that as one variable (education) increases, the other variable (wages) tends to increase as well, but the relationship is not that strong.

Multiple Regression

Q4 :Do multiple regression model and compare with a simple Linear regression model (basis of comparison: Intercept and Slope, variance of parameter and error term, t-test, R Square, Adj R square, F test).Plot regression model on the graph.

```
mm<-lm(wage ~ educ + exper, data = ds3)
mm
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = ds3)
##
## Coefficients:
## (Intercept)         educ        exper
##    -18.7048       1.9478       0.6143
```

```
summary(mm)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = ds3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.329  -4.586  -1.210   2.636  69.320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.7048     2.7237  -6.867 1.04e-11 ***
## educ          1.9478     0.1390  14.013  < 2e-16 ***
## exper         0.6143     0.1051   5.842 6.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.362 on 1227 degrees of freedom
## Multiple R-squared:  0.1537, Adjusted R-squared:  0.1523
## F-statistic: 111.4 on 2 and 1227 DF,  p-value: < 2.2e-16
```

```
pv2<-vcov(mm)
pv2
```

```
##             (Intercept)        educ       exper
## (Intercept)   7.4187987 -0.35936669 -0.24919785
## educ         -0.3593667  0.01932028  0.01000614
## exper        -0.2491979  0.01000614  0.01105472
```

```
variances2 <- diag(pv2)
variances2
```

```
## (Intercept)        educ       exper
##  7.41879872  0.01932028  0.01105472
```

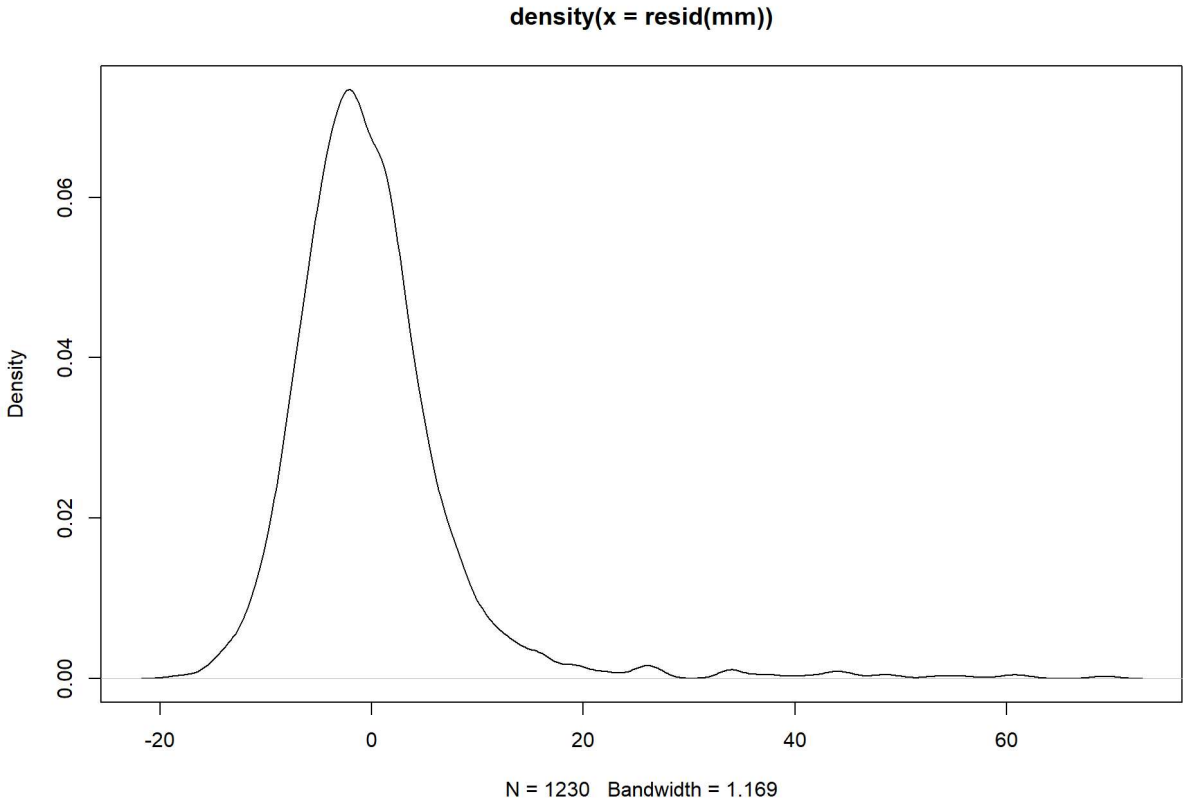
```
anova(mm)
```

```
## Analysis of Variance Table
##
## Response: wage
##             Df Sum Sq Mean Sq F value    Pr(>F)
## educ         1  13196 13196.3 188.745 < 2.2e-16 ***
## exper        1   2387  2386.6  34.135 6.581e-09 ***
## Residuals 1227  85787    69.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
resid2<-resid(mm)
var(resid2)
```

```
## [1] 69.80245
```

```
plot(density(resid(mm)))
```



Process :

- Now here i have formed a multiple linear regression model named "mm" with Y = B_1 + B_2X_2 + B_3X_3 + u as the **PRF** and Y hat = b_1 + b_2X_2 + b_3X_3 + e is the **SRF** where Y is the wage, X1 is the education and X2 is experience.

- The summary shows us that $\mathbf{b_1 = -18.7048}$, $\mathbf{b_2 = 1.9478}$ and $\mathbf{b_3 = 0.6143}$ the 3 stars (***) in the intercept, educ and exper shows that all are significant we also have residual standard error i.e sigma hat which is **8.3615923** also called as standard error of regression (SER). also notice that degree of freedom is given 1227 i.e (n-3) in this case as now we have three unknown parameters which is again consistent with what we studied in theory.

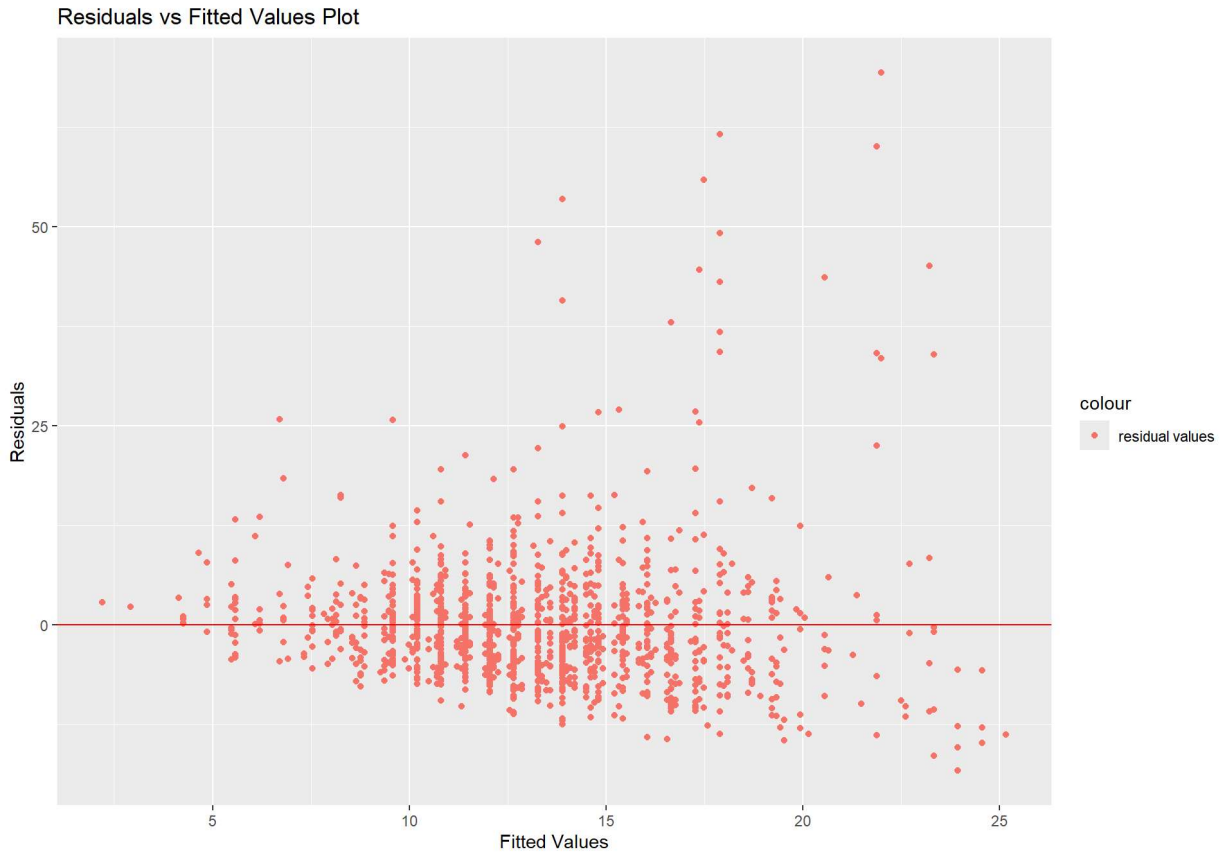
- R^2 value is **0.1537** i.e only 15.37% which means only this percentage of variation in Y was explained by the model also known as goodness of fit.

- Here the adjusted R^2 value which is **0.1523432** it can be -ve or +ve in this case it is +ve.

Now representing the model graphically :

```
ds3 %<>% mutate(wage_hat_2=fitted(mm))
ds3 %<>% mutate(resid2=residuals(mm))

ggplot(data = ds3, mapping = aes(x=wage_hat_2))+
  geom_point(mapping = aes(y=resid2, col = 'residual values' ))+
  labs(x = "Fitted Values", y = "Residuals",
       title = "Residuals vs Fitted Values Plot")+
  geom_hline(yintercept = 0, color = "red")
```



- We can see here that there are some points that deviate substantially from the horizontal line at y = 0 which represents *outliers* or influential observations.

- Also *Cluster of points* may suggest that there are subgroups or patterns in the data that are not captured by the model, these clusters could also indicate that there are additional variables or nonlinear relationships that should be included in the model to better explain the variability in the response variable. Alternatively, the clusters may arise due to inherent variability in the data that cannot be explained by the predictors included in the model.

```r
cor(ds3$wage,ds3$educ)
```

```
## [1] 0.360804
```

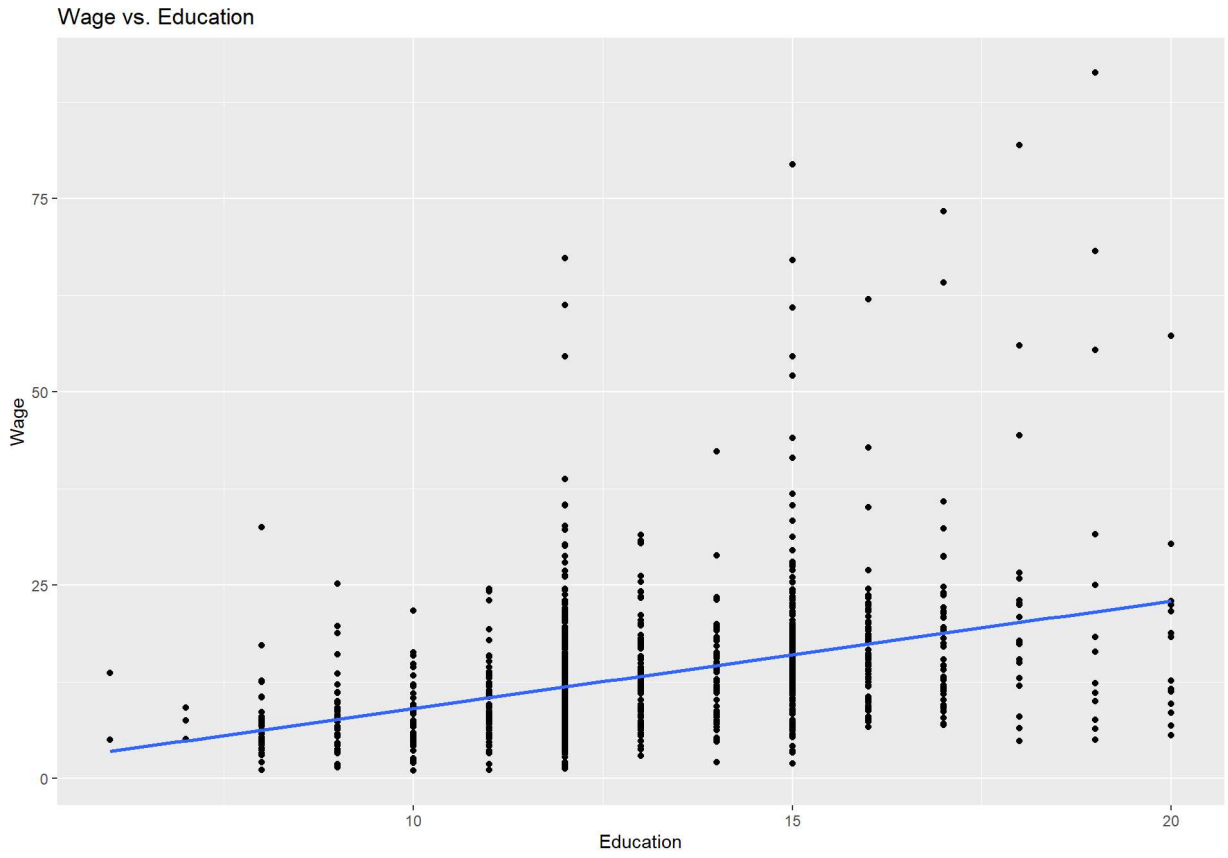
```r
cor(ds3$wage,ds3$exper)
```

```
## [1] -0.1352018
```

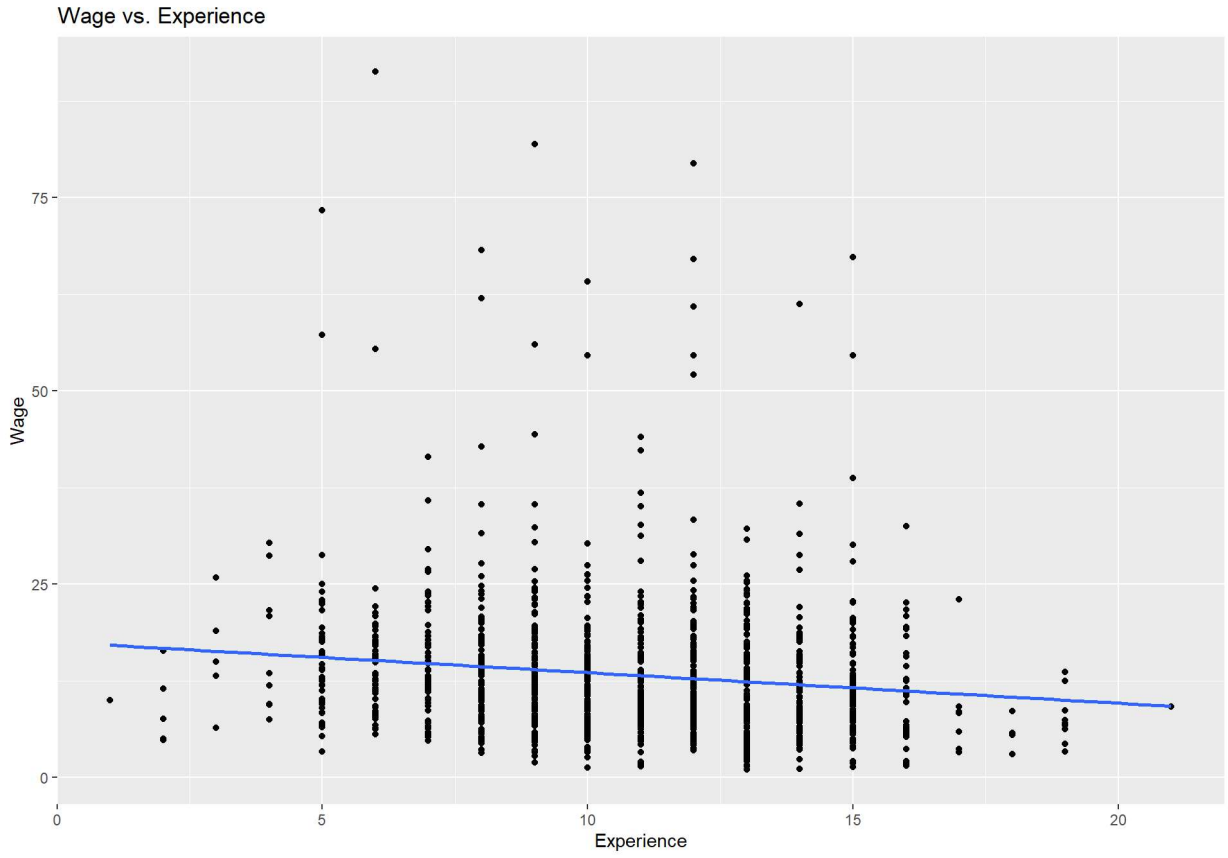
```r
plot1 <- ggplot(ds3, aes(x = educ, y = wage)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Wage vs. Education",
       x = "Education",
       y = "Wage")

plot2 <- ggplot(ds3, aes(x = exper, y = wage)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Wage vs. Experience",
       x = "Experience",
       y = "Wage")

plot1
```


Wage vs. Education
Wage
Education
0
25
50
75
10
15
20

```
plot2
```


Wage vs. Experience
Wage
Experience
0
25
50
75
0
5
10
15
20

Now here I have calculated the correlation of each explanatory variables with the dependent variable which is **0.360804** and **-0.1352018** for education and experience respectively. We see there is a -ve relationship between wage and exper whereas a +ve relationship between wage and educ, though in both the cases there is a *moderate correlation.* Also very evident from the corresponding graphs.

Now if we compare it with linear regression:

	R^2	Adjusted R^2	SER (sigma hat)	Var of residuals (sigma^2 hat)	Var of Intercept	Var of Slopes	Correlation	F Sta Val
Linear Model	0.1302	0.1295	8.474	71.74432	1.8499	educ = 0.01054	educ = 0.3608	183.
Multiple Linear Model	0.1537	0.1523	8.362	69.8024	7.4187	educ = 0.01932 exper = 0.011054	educ = 0.3608 exper = -0.1352	111.

We can see few interesting things:

1. In multiple regression both R^2 and adjusted R^2 is greater than in the former model which means that the latter model explains or fits the data better.
2. Also the fact that we have increased the explanatory variables gives us the necessary penalty.
3. And then we have variance whose literal meaning is the spread of data points so here more the variance, more will be the spread.

---

1. Categorical data refers to data that can be divided into groups or categories, typically representing qualitative characteristics or attributes. Examples include gender, color, and type of car.↩
2. Numerical data, on the other hand, consists of numbers and can be measured or quantified. It represents quantities or values that can be counted or measured, such as height, weight, temperature, and age.↩
3. In R, a directory is essentially a folder in your computer's file system. These are fundamental for organizing and accessing files and data within R.↩

4. *SER represents the standard deviation of the errors/residuals in a regression model. It measures the accuracy of the regression predictions.*↩

5. r^2 is a statistical measure used in regression analysis to assess the goodness of fit of a regression model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables.↩

6. is a modification of the standard r^2 value that adjusts for the number of predictors in a regression model. While r^2 tends to increase as more predictors are added to the model, even if those predictors are not relevant, Adjusted R^2 penalizes the addition of unnecessary variables.↩