honor statement: "I have completed this work independently. The solutions given are entirely my own work."

1. Short Essay (20 pts.) For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.
a. (10 pts.) When building a model, you make four assumptions about the residuals. Explain what they are and how you can verify that your assumptions are correct.

1. The mean of the residuals is 0 - The least square regression model always produces a sum of error at 0
2. The residuals are homoscedastic - the variance of errors is constant throughout the independent variables
3. The residuals are normal - about half of the error will be above the regression line and about half below,most will be close to the regression line and some further away.
4. The residuals are independent - one error should not depend on the other error.

b. (10 pts) Define 'interaction term'. From your own experience, identify an instance in which you believe an interaction term would be appropriate.
Define 'interaction term':
An interaction Model relating E(y) prediction to two quantitative independent variables interact to have a n effect that is different from the sum of their parts.
EX:
$(E(y) = \beta 0+\beta 1x1+\beta 2x2+\beta 3x1*x2)$
In the interaction model, when we change x1, it is going to impact β1, but it is also going to impact β3. For every unit increase x1,holding x2 fixed, y is going to change by β1 plus β3x2 ; for every unit increase x2,holding x1 fixed, y is going to change by β2 plus β3x1 .
The interaction term is putting a twist in our predictive plane.

EX: Drug interaction- a certain pill was used to depress a human's central nervous system , the alcohol has the same impact on the human's central nervous system, so if someone takes a certain pill and alcohol at the same time, pill and alcohol would work together and strengthen the effect to even cause to fall into a coma.
We could use the interaction term model to see how these two variables work together to make the introduction to people how to take the pill and alcohol adequately.

2. BANKING (30 pts.) Use the Banking dataset for this question, found under content on the D2L. This dataset consists of data acquired from banking and census records for different zip codes in the bank's current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The fields in the dataset: Median age of the population (Age) Median years of education (Education) Median income (Income) in

$ Median home value (HomeVal) in $ Median household wealth (Wealth) in $ Average bank balance (Balance) in $

a. (5 pts.) In R, you can create a scatter plot by using the plot command, i.e. plot(x, y). Create scatterplots to visualize the associations between bank balance and the other five variables. Paste them (5 in total) into your submission. Describe the relationships.

Balance & Age
Form: linear
Strength :Weak
Direction: Positive

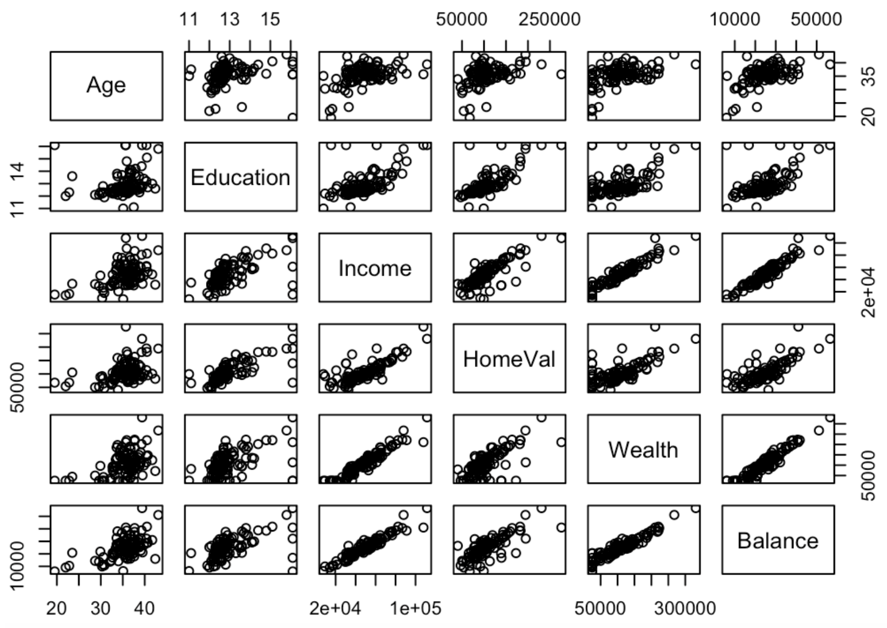
Balance & Education
Form: linear
Strength :Weak
Direction: Positive

Balance & Income
Form: linear
Strength :strong
Direction: Positive

Balance & HealthVal
Form: linear
Strength :mediocre
Direction: Positive

Balance & Wealth
Form: linear
Strength :strong
Direction: Positive

plot(banking)



b. (5 pts.) In R, you can compute correlations between two variables by using the cor command, i.e. cor(x,y) where x and y are the names of your variables, or you can compute pair-wise correlations by using cor(D), where D is the name of your dataframe. Compute correlations for the bank data. Paste them into your submission. Describe which variables appear to be strongly associated? Interpret any correlation values you deem important.

Trying to predict banking Balance
Correlation value between Balance and Age is 56% which is not strong enough to keep this variable in the data.
Correlation value between Balance and Education is 55% which is not strong enough to keep this variable in the data.
Correlation value between Balance and Income is 95% which is a strong positive correlation that could be used for further analysis, so we could keep this variable in the data.
Correlation value between Balance and HomeVal is 76% which is a mediocre that we could keep this variable to see how it would affect the model
Correlation value between Balance and Wealth is 94% which is a strong positive correlation that could be used for further analysis, so we could keep this variable in the data.
The strongest correlation is income which could be the best predictor for balance.

> cor(banking)

	Age	Education	Income	HomeVal	Wealth	Balance
Age	1.0000000	0.1734611	0.4771474	0.3864931	0.4680918	0.5654668
Education	0.1734611	1.0000000	0.5731467	0.7489426	0.4681199	0.5521889
Income	0.4771474	0.5731467	1.0000000	0.7953552	0.9466654	0.9516845
HomeVal	0.3864931	0.7489426	0.7953552	1.0000000	0.6984778	0.7663871
Wealth	0.4680918	0.4681199	0.9466654	0.6984778	1.0000000	0.9487117

```
Balance   0.5654668   0.5521889   0.9516845   0.7663871   0.9487117   1.0000000
```

c. (5 pts.) Fit a single regression model of balance vs the other five variables. Present the estimated regression model and evaluate it. Recall that you can build a linear regression model by using the lm command and display the model by using the summary command.

```
model<-lm(banking$Balance~banking$Age+banking$Education+banking$Income+banking$Ho
meVal+banking$Wealth)
> summary(model)

Call:
lm(formula = banking$Balance ~ banking$Age + banking$Education +
    banking$Income + banking$HomeVal + banking$Wealth)

Residuals:
    Min      1Q  Median      3Q     Max
-5365.5 -1102.6   -85.9   868.9  7746.5

Coefficients:
                    Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)       -1.033e+04   4.219e+03  -2.449    0.016160 *
banking$Age        3.175e+02   6.104e+01   5.201    1.12e-06 ***
banking$Education  5.903e+02   3.151e+02   1.873    0.064085 .
banking$Income     1.468e-01   4.083e-02   3.596    0.000512 ***
banking$HomeVal    9.864e-03   1.099e-02   0.898    0.371591
banking$Wealth     7.414e-02   1.120e-02   6.620    2.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2059 on 96 degrees of freedom
Multiple R-squared:  0.9468,  Adjusted R-squared:  0.944
F-statistic: 341.4 on 5 and 96 DF,  p-value: < 2.2e-16
```

By checking the p-value in F-test, we could reject the null hypothesis and accept the alternative one, which is great, it means we could apply this model to predict the balance appropriately.The R-square is 94% which means 94% variance of dependent variable is explained by the model.

```
> m1 <- lm(banking$Balance~banking$Age)
> summary(m1)

Call:
lm(formula = banking$Balance ~ banking$Age)

Residuals:
   Min     1Q Median     3Q    Max
-18236  -3890  -1152   3404  26685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19976.5     6582.6  -3.035  0.00307 **
banking$Age   1265.5      184.6   6.856 5.93e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7209 on 100 degrees of freedom
Multiple R-squared:  0.3198,  Adjusted R-squared:  0.313
F-statistic: 47.01 on 1 and 100 DF,  p-value: 5.931e-10
```

Balance$Age
By checking the p-value in this model, we failed to reject the null hypothesis, so we could not include this variable in our data. The R-square is 31% which is not good enough to let us include this variable.

```
> m2 <- lm(banking$Balance~banking$Education)
> summary(m2)

Call:
lm(formula = banking$Balance ~ banking$Education)

Residuals:
   Min     1Q Median     3Q    Max
-33793  -3266    115   4871  16820

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -36850.8     9349.5  -3.941  0.00015 ***
banking$Education   4757.7      718.3   6.623 1.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 7288 on 100 degrees of freedom
Multiple R-squared: 0.3049, Adjusted R-squared: 0.298
F-statistic: 43.87 on 1 and 100 DF, p-value: 1.784e-09

Balance$Education
By checking the p-value in this model, we failed to reject the null hypothesis, so we could not include this variable in our data. The R-square is 29% which is not good enough to let us include this variable.

```
> m3 <- lm(banking$Balance~banking$Income)
> summary(m3)

Call:
lm(formula = banking$Balance ~ banking$Income)

Residuals:
    Min      1Q  Median      3Q     Max 
-9132.5 -1656.2  -179.4  1329.1  9447.5 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.020e+03  7.239e+02   5.554 2.31e-07 ***
banking$Income 4.275e-01  1.379e-02  30.992  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2684 on 100 degrees of freedom
Multiple R-squared:  0.9057,  Adjusted R-squared:  0.9048
F-statistic: 960.5 on 1 and 100 DF,  p-value: < 2.2e-16
```

Balance$Education
By checking the p-value in this model, we could reject the null hypothesis, so we could include this variable in our data. The R-square is 90% which is good enough to let us include this variable.

```
> m4 <- lm(banking$Balance~banking$HomeVal)
> summary(m4)

Call:
lm(formula = banking$Balance ~ banking$HomeVal)

Residuals:
     Min       1Q   Median       3Q      Max 
-17397.4  -2252.9    607.8   2999.3  12948.7 
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6529.6074  1636.1372   3.991 0.000126 ***
banking$HomeVal    0.1718     0.0144  11.930  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5615 on 100 degrees of freedom
Multiple R-squared:  0.5873,  Adjusted R-squared:  0.5832
F-statistic: 142.3 on 1 and 100 DF,  p-value: < 2.2e-16
```

Balance$HomeVal
By checking the p-value in this model, we could reject the null hypothesis, so we could include this variable in our data. The R-square is 58% which is a mediocre number, we could include this variable to see how it would affect the model.

```
> m5 <- lm(banking$Balance~banking$Wealth)
> summary(m5)

Call:
lm(formula = banking$Balance ~ banking$Wealth)

Residuals:
    Min      1Q  Median      3Q     Max
-7344.2 -1650.8  -162.2  1248.1  7271.8

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.853e+03  5.709e+02   17.26   <2e-16 ***
banking$Wealth 1.379e-01  4.595e-03   30.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2763 on 100 degrees of freedom
Multiple R-squared:  0.9001,  Adjusted R-squared:  0.8991
F-statistic: 900.5 on 1 and 100 DF,  p-value: < 2.2e-16
```

Balance$Wealth
By checking the p-value in this model, we could reject the null hypothesis, so we could include this variable in our data. The R-square is 89% which is good enough to let us include this variable.

.

d. (5 pts.) Which of the five predictors have a significant (a=.05) effect on balance? Explain.

By checking the P-value of each variable, the variable of Income and Wealth is so small that we can reject the null hypothesis and accept the alternative one which is Beta is not equal to zero so they have a significant (a=.05) effect on balance.

e. (5 pts.) A good model should only contain significant independent variables, so remove the variable with the largest p-value (>0.05) and refit the regression model of balance versus the remaining four predictors. Analyze if all four predictors have a significant association with balance? (a=.05) If not, continue to remove one insignificant variable at a time until all the remaining predictors are significant. Present the final regression model.

The R-square did not improve by removing the the worst p-value in variable -HomeVal which means that we should include all variables in our model

```
> model1<-lm(banking$Balance~banking$Age+banking$Education+banking$Income+banking$Wealth)
> summary(model1)

Call:
lm(formula = banking$Balance ~ banking$Age + banking$Education +
    banking$Income + banking$Wealth)

Residuals:
    Min      1Q  Median      3Q     Max
-5403.9 -1234.1   -75.0   998.6  7430.7

Coefficients:
                   Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)       -1.214e+04  3.704e+03  -3.278    0.00145 **
banking$Age        3.242e+02  6.051e+01   5.358    5.68e-07 ***
banking$Education  7.498e+02  2.600e+02   2.884    0.00484 **
banking$Income     1.615e-01  3.738e-02   4.321    3.75e-05 ***
banking$Wealth     7.265e-02  1.106e-02   6.566    2.57e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2057 on 97 degrees of freedom
Multiple R-squared:  0.9463,  Adjusted R-squared:  0.9441
F-statistic: 427.4 on 4 and 97 DF,  p-value: < 2.2e-16
```

```
> model2<-lm(banking$Balance~banking$Age+banking$Income+banking$Wealth)
> summary(model2)

Call:
lm(formula = banking$Balance ~ banking$Age + banking$Income +
    banking$Wealth)

Residuals:
    Min      1Q  Median      3Q     Max
-4991.0 -1201.0  -166.8  1059.5  7281.3

Coefficients:
                 Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     -3.115e+03  2.054e+03  -1.517     0.133
banking$Age      3.019e+02  6.222e+01   4.852     4.61e-06 ***
banking$Income   2.119e-01  3.425e-02   6.188     1.42e-08 ***
banking$Wealth   6.381e-02  1.102e-02   5.789     8.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2132 on 98 degrees of freedom
Multiple R-squared:  0.9417,  Adjusted R-squared:  0.9399
F-statistic: 527.7 on 3 and 98 DF,  p-value: < 2.2e-16
```

f. (5 pts.) Interpret each of the regression coefficients for the final model. Discuss the adjR 2 for the final model. Is this a good model? Explain.

The model did no be improve by removing these variable with the highest p -value, so we should include all variables in the model, the adjR 2 for the final model is 94.4% which means 94% variance of dependent variable is explained by the model.It could be a good model