Big Data
Lab 5
(Logistic Regression)

Name	ID	SEC	BN
Norhan Reda Abdelwahed Ahmed	9203639	2	31
Hoda Gamal Hamouda Ismail	9203673	2	33

Supervisor: Eng. Omar Samir

(Q1) Write the variable pairs that are not correlated at all to each other.

Ans

```
       Price     Income        Age
Price      1 0.00000000 0.00000000
Income     0 1.00000000 0.09612083
Age        0 0.09612083 1.00000000
```

It seems from the correlation matrix that the pairs that are not correlated are : (price,age),(price,income)

(Q2)Are there any highly correlated variables in this dataset?

No

(Q3)How many categories are there for the Price variable?

There are 3 categories for the Price variable which

```
   10, 20 ,30
```

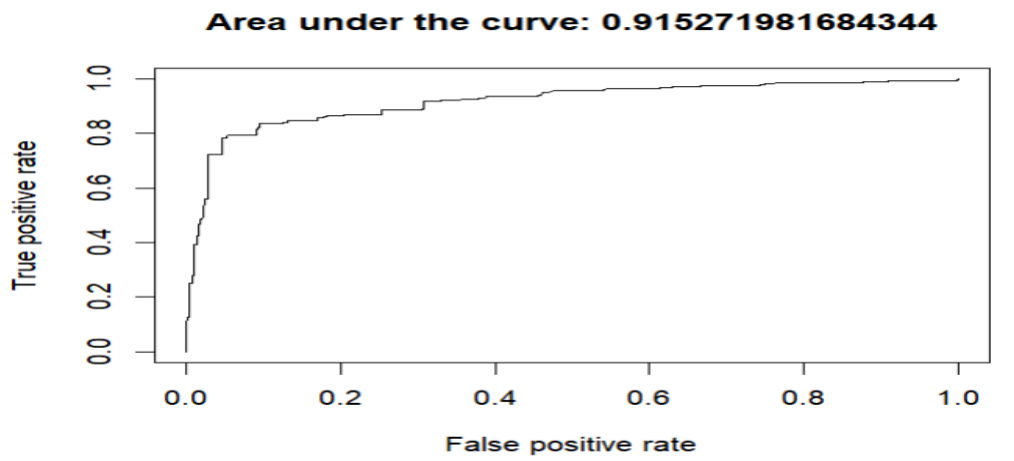
```
    MYDEPV
Price   0   1
   10 115 135
   20 137 113
   30 174  76
```

(Q4) Why is it divided into two entries only in the model?

This is because when using a categorical variable with k levels, the model includes k-1 dummy variables to represent the different levels. In this case, the reference level is "Price" 10, so it is not explicitly included as a separate coefficient in the model. The coefficients "as.factor(Price)20" and "as.factor(Price)30" represent the differences in the log-odds of the outcome variable between the reference level (10) and the other levels (20 and 30)

(Q5.1) Write the value of this expression (just the number)

AUC= 0.915272



(Q5.2) What is the maximum value of AUC (ideal case)?

the maximum value of AUC (Area Under the Curve) in the ideal case is 1. AUC ranges from 0 to 1, where a value of 1 represents a perfect classifier that can perfectly separate the classes.

(Q6)What does each point in the ROC graph represent?
#In other words, what is the value that changes and drives TPR and FPR to change too
#from one point to another in the graph?

In a ROC (Receiver Operating Characteristic) graph, each point represents the performance of a classification model at a specific threshold value. The ROC graph plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold values.

The value that changes and drives TPR and FPR to change from one point to another in the ROC graph is the threshold value used to classify the predictions. The threshold determines the point at which the predicted probabilities or scores are converted into class labels. By adjusting the threshold, we can control the trade-off between the TPR and FPR.

(Q7)How is the predicted probability affected by changing only the price holding all other variables constant?

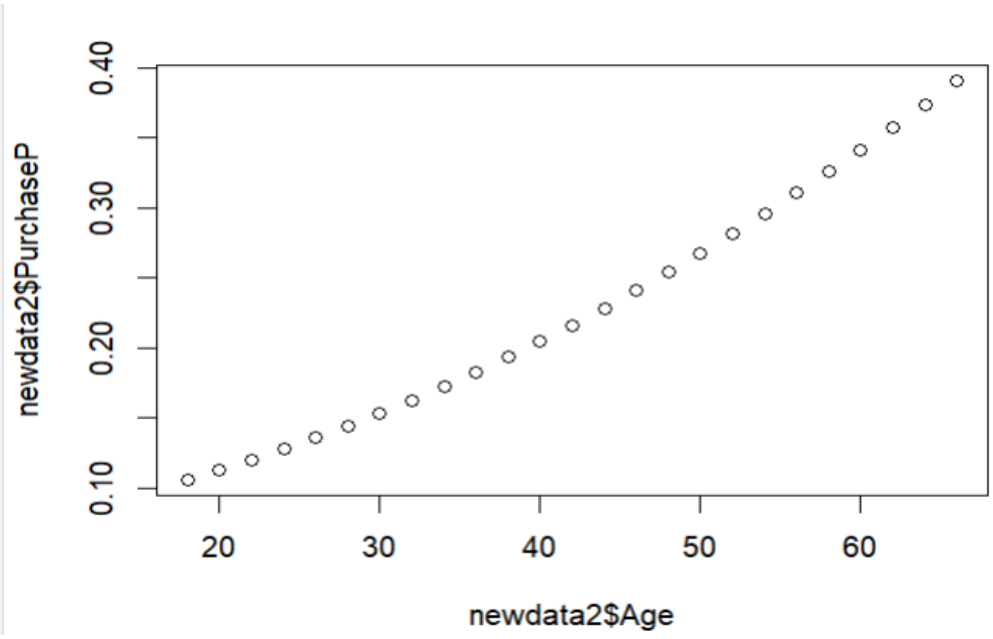
	Income	Age	Price	PurchaseP
1	42.492	35.976	10	0.6707408
2	42.492	35.976	20	0.4918407
3	42.492	35.976	30	0.1826131

By examining the values of "PurchaseP" for these observations, we can see that as the "Price" increases from 10 to 30 while keeping "Income" and "Age" constant, the predicted probability of "PurchaseP" decreases.
Therefore , holding the "Income" and "Age" variables constant, increasing the "Price" is associated with a decrease in the predicted probability of "PurchaseP." so there is a negative relationship between them

(Q8)How is the predicted probability affected by changing only age holding all other variables constant?

	Age	Income	Price	PurchaseP
1	18	42.492	30	0.1063052
2	20	42.492	30	0.1131540
3	22	42.492	30	0.1203845
4	24	42.492	30	0.1280103
5	26	42.492	30	0.1360445
6	28	42.492	30	0.1444993
7	30	42.492	30	0.1533864
8	32	42.492	30	0.1627160
9	34	42.492	30	0.1724975
10	36	42.492	30	0.1827387
11	38	42.492	30	0.1934457
12	40	42.492	30	0.2046231
13	42	42.492	30	0.2162731
14	44	42.492	30	0.2283958
15	46	42.492	30	0.2409892
16	48	42.492	30	0.2540483
17	50	42.492	30	0.2675657
18	52	42.492	30	0.2815308
19	54	42.492	30	0.2959303
20	56	42.492	30	0.3107477
21	58	42.492	30	0.3259636
22	60	42.492	30	0.3415553
23	62	42.492	30	0.3574973

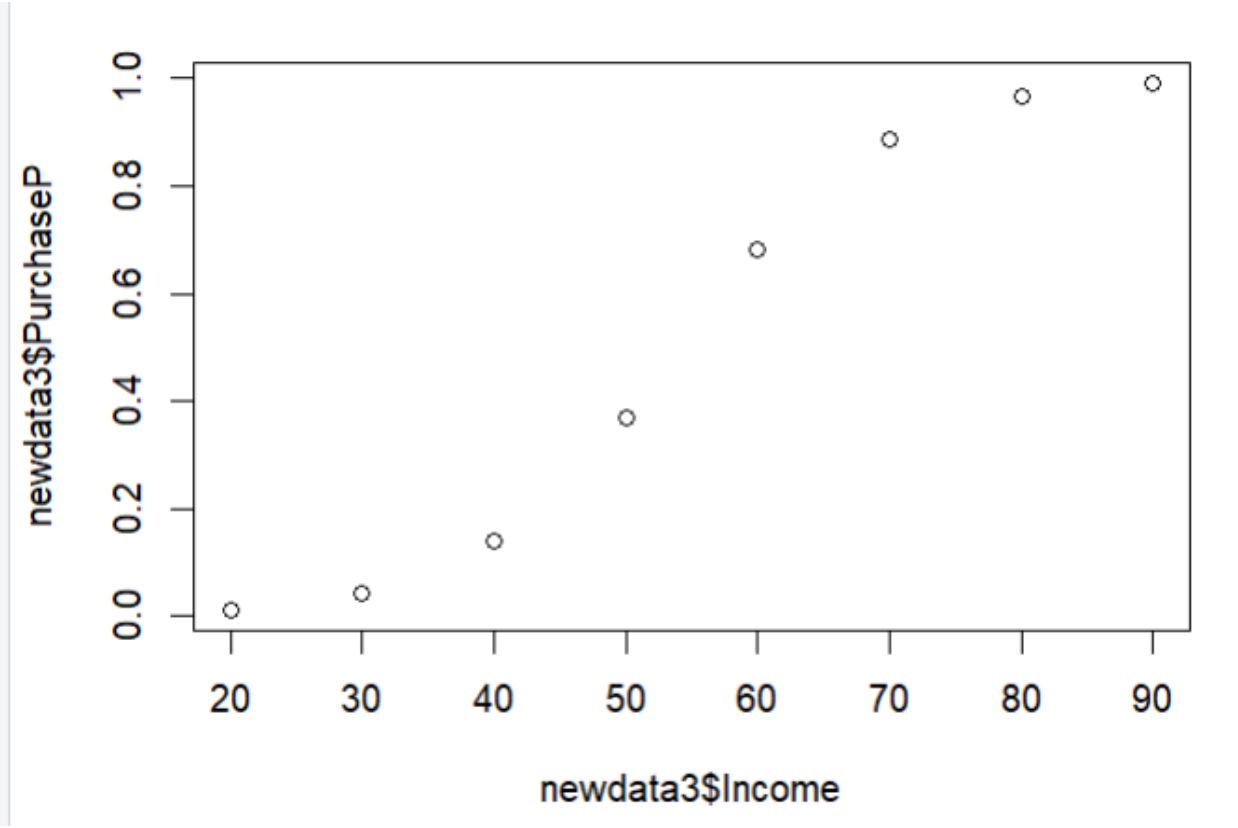
```
24   64    42.492     30   0.3737609
25   66    42.492     30   0.3903150
```



As the "Age" variable increases from 18 to 66 while holding "Income," "Price," and other variables constant, the predicted probability of "PurchaseP" also increases. This indicates that there is a positive relationship between age and the likelihood of making a purchase, according to the model.

(Q9)How is the predicted probability affected by changing only income holding all other variables constant?

```
    Age Income Price   Prob
    Income    Age   Price  PurchaseP
1     20    35.976    30   0.01219091
2     30    35.976    30   0.04281102
3     40    35.976    30   0.13948050
4     50    35.976    30   0.37004640
5     60    35.976    30   0.68039246
6     70    35.976    30   0.88525564
7     80    35.976    30   0.96546923
8     90    35.976    30   0.99022745
```



From the given dataset, it appears that as the "Income" variable increases while keeping "Age," "Price," and other variables constant, the predicted probability of "Prob" generally increases This suggests a positive relationship between income and the likelihood of making a purchase according to the model.