Homework – 1
Sai Kiran Putta

Question – 1:

Following are the results using different methods (cosine and jaccard) for different vectors.

Part-A using Cosine similarity.

```
> distance(A, method="cosine")
          v1        v2        v3
v1 1.0000000 0.6666667 0.6666667
v2 0.6666667 1.0000000 0.6666667
v3 0.6666667 0.6666667 1.0000000
```

Part-A using Jaccard similarity.

```
> distance(A, method="jaccard")
    v1  v2  v3
v1 0.0 0.5 0.5
v2 0.5 0.0 0.5
v3 0.5 0.5 0.0
```

Part-B using Cosine similarity.

```
> distance(A, method="cosine")
          v1        v2        v3
v1 1.0000000 0.5773503 0.5000000
v2 0.5773503 1.0000000 0.2886751
v3 0.5000000 0.2886751 1.0000000
```

Part-B using Jaccard similarity.

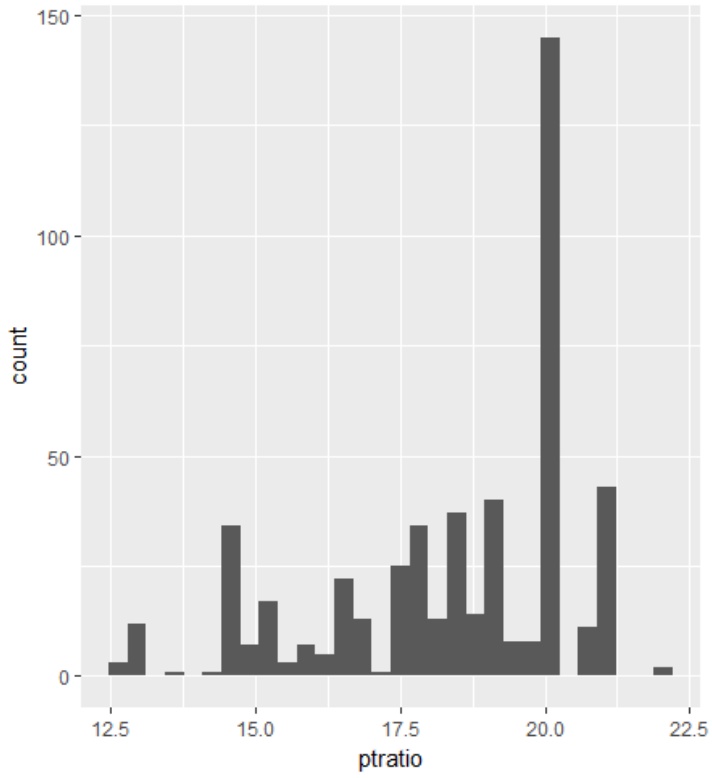
```
> distance(A, method="jaccard")
          v1        v2        v3
v1 0.0000000 0.6000000 0.6666667
v2 0.6000000 0.0000000 0.8333333
v3 0.6666667 0.8333333 0.0000000
```

Part-C using Cosine similarity.

```
> distance(A, method="cosine")
          v1         v2         v3
v1 1.0000000  0.5465040  0.1634083
v2 0.5465040  1.0000000 -0.3125615
v3 0.1634083 -0.3125615  1.0000000
```

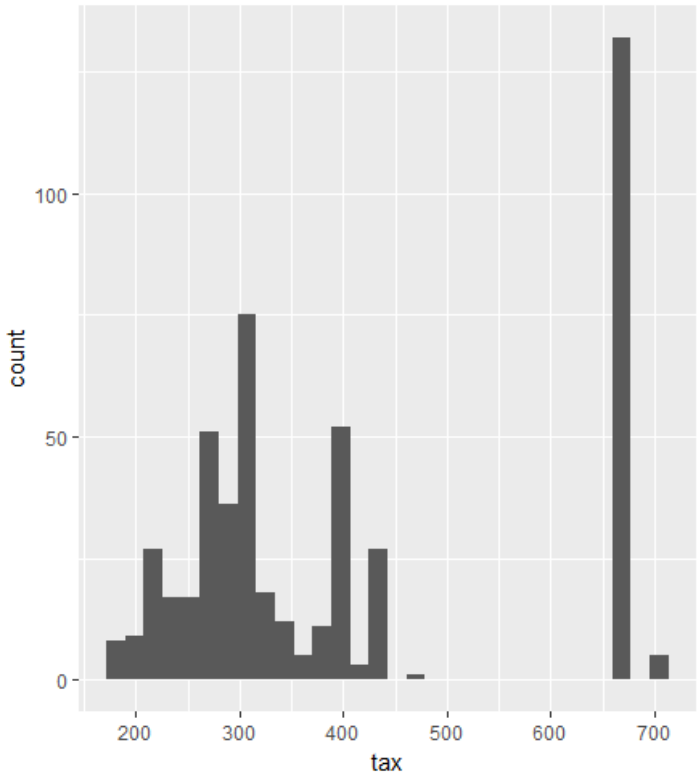
Question – 2

To better understand the data, let us plot the histograms of the data. And let us binarize the data using the plots and the summary of the feature.
For example,



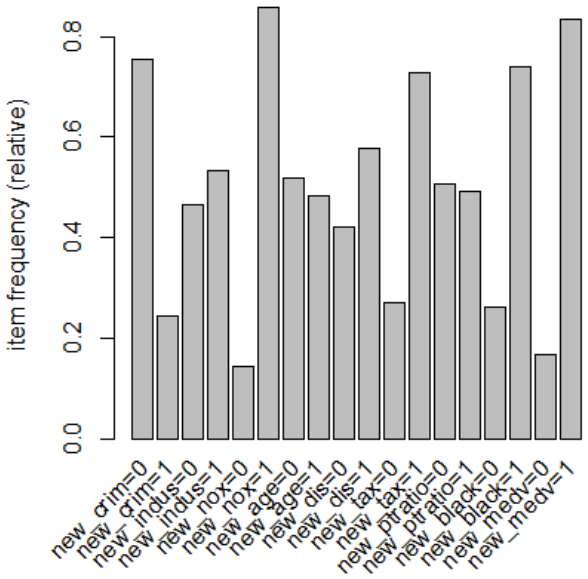
A binarised split of >19 is taken for ptratio after looking at the graphs.

As an other example,



A binarised split of >500 is used for the feature tax.

Following is the frequency plot obtained.



From the graph above we can see that new_nox = 1 (low) and new_medv = 1 (low) have more frequency.
Taking a support of 0.05, building association rules for this data gives us around 6900 rules.

c)
Irrespective of the limit that we set to the support, we are not able to get any rules that match both low crime area and place that is as close to city as possible.
Hence dividing the rules and checking, we get,
Low crime rules:

```
> inspect(head(sort(rules_lowcrime, by = "lift")))
    lhs                                                  rhs          support    confidence lift     count
[1] {new_dis=0,new_black=1,new_medv=0}                => {new_crim=1} 0.05533597 0.8        3.264516 28
[2] {new_age=1,new_dis=0,new_black=1,new_medv=0}      => {new_crim=1} 0.05533597 0.8        3.264516 28
[3] {new_dis=0,new_ptratio=1,new_black=1,new_medv=0} => {new_crim=1} 0.05533597 0.8        3.264516 28
[4] {new_indus=1,new_dis=0,new_black=1,new_medv=0}    => {new_crim=1} 0.05533597 0.8        3.264516 28
[5] {new_dis=0,new_tax=1,new_black=1,new_medv=0}      => {new_crim=1} 0.05533597 0.8        3.264516 28
[6] {new_nox=1,new_dis=0,new_black=1,new_medv=0}      => {new_crim=1} 0.05533597 0.8        3.264516 28
```

Inference: One interesting observation from the rules above is that, for the crime to be low, it is better to opt for a place that is far away from the employment center! And also live in high valued occupied homes.

Low distance rules:

```
> inspect(head(sort(rules_lowdistance, by = "lift")))
    lhs                          rhs          support   confidence lift     count
[1] {new_nox=0}               => {new_dis=1} 0.1422925 1          1.726962 72
[2] {new_nox=0,new_tax=0}     => {new_dis=1} 0.1106719 1          1.726962 56
[3] {new_indus=0,new_nox=0}   => {new_dis=1} 0.1422925 1          1.726962 72
[4] {new_nox=0,new_ptratio=0} => {new_dis=1} 0.1106719 1          1.726962 56
[5] {new_nox=0,new_age=0}     => {new_dis=1} 0.1422925 1          1.726962 72
[6] {new_nox=0,new_black=1}   => {new_dis=1} 0.1146245 1          1.726962 58
```

Inference: One interesting observation from the rules above is the following that, the NOX values are high. That is since it is nearer to the employment centers the Nitric Oxide concentrations are more.

d)Low ptratio rules:

```
> inspect(head(sort(rules_lowpupilratio, by = "lift")))
    lhs                                    rhs              support    confidence lift     count
[1] {new_crim=1,new_black=1,new_medv=0}  => {new_ptratio=1} 0.05928854 1          2.032129 30
[2] {new_dis=0,new_black=1,new_medv=0}   => {new_ptratio=1} 0.06916996 1          2.032129 35
[3] {new_age=1,new_black=1,new_medv=0}   => {new_ptratio=1} 0.08300395 1          2.032129 42
[4] {new_indus=1,new_dis=1,new_medv=0}   => {new_ptratio=1} 0.05928854 1          2.032129 30
[5] {new_indus=1,new_black=1,new_medv=0} => {new_ptratio=1} 0.11462451 1          2.032129 58
[6] {new_dis=1,new_tax=1,new_medv=0}     => {new_ptratio=1} 0.06916996 1          2.032129 35
>
```

Inference: From the rules above it is better to opt for a place where,
the crime is less
the black proportion is less
the distance to employment center is less
property tax to be less and

median value of homes is more (possibly as a consequence)

The summary of the regression model is as follows.

```
Call:
lm(formula = new_ptratio ~ ., data = subset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97694 -0.07909  0.07719  0.32456  0.63415

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.30758    0.09197   3.344 0.000887 ***
new_crim     0.12589    0.04585   2.746 0.006255 **
new_indus   -0.12126    0.05040  -2.406 0.016505 *
new_nox     -0.15629    0.05520  -2.831 0.004823 **
new_age      0.12954    0.04666   2.776 0.005706 **
new_dis      0.07302    0.04677   1.561 0.119119
new_tax      0.69147    0.05135  13.466  < 2e-16 ***
new_black    0.05416    0.03792   1.428 0.153800
new_medv    -0.35566    0.04769  -7.457 3.96e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3696 on 497 degrees of freedom
Multiple R-squared:  0.4631,    Adjusted R-squared:  0.4544
F-statistic: 53.58 on 8 and 497 DF,  p-value: < 2.2e-16
```

The regression model too follows a similar trend where crime is low, black proportion is low, distance to employment centers is low and the median value of houses is more. But the property tax seems to be high.
In the current case where the features are binary it is easier to interpret it using association rules. Generally regression can be preferred when the features are mostly non-categorical (numeric).

Question – 3

Firstly importing the data – marketing from library ElemStatLearn. Adding column target with class as 1.
Randomly permuting the features in the dataset and naming the target as 0. Combining both the dataframes.

After building a decision tree model we can observe that it only has one single root indicating that the features do not have predictive power to do a classification.

To verify it, we can predict the model on the training set itself and we observe that the probability for every row is 0.5 for both the classes, indicating that there is no predictive power.