Question 1: Taxi
Author: Alden Steve Arduo

The sample was randomly taken from different taxi companies over a period of three months. Distance (in kilometres) is how far the taxi has travelled and fare (in $) is the fee for the taxi ride.

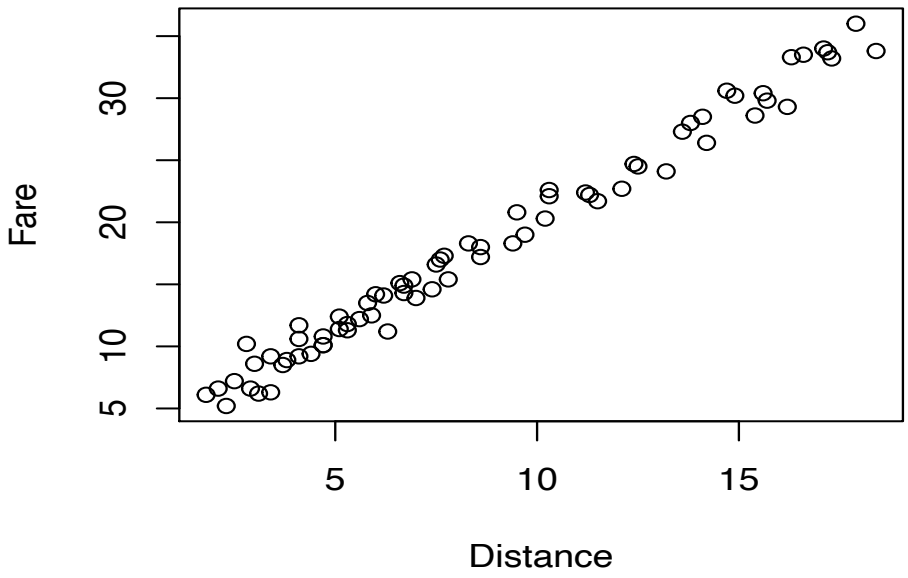
In this assignment, I will investigate if I can use the distance travelled by a taxi to predict its fare.

1.

```
> setwd("~/RStudio")
> taxi.df = read.csv("taxi.csv")
> head(taxi.df, n=6)
  Distance Fare
1      2.3  5.2
2      3.1  6.2
3      3.4  6.3
4      3.7  8.5
5      4.4  9.4
6      4.7 10.1
> tail(taxi.df, n=6)
   Distance Fare
67     13.6 27.3
68     14.1 28.5
69     14.7 30.6
70     15.6 30.4
71     16.6 33.5
72     17.1 34.0
> taxi.df[3:10,]
   Distance Fare
3       3.4  6.3
4       3.7  8.5
5       4.4  9.4
6       4.7 10.1
7       5.3 11.8
8       5.3 11.3
9       6.2 14.1
10      6.3 11.2
```

2.

```
> plot(Fare ~ Distance, data = taxi.df)
```



The plot shows a positive relationship between distance the taxi travelled and the fare of the taxi ride. As the distance increases, the fare increases as well.

```
> taxi.lm = lm(Fare ~ Distance, data = taxi.df)
> summary(taxi.lm)

Call:
lm(formula = Fare ~ Distance, data = taxi.df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47565 -0.79446 -0.04503  0.86890  2.82383

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.33659    0.27550   8.481 2.37e-12 ***
Distance     1.79985    0.02756  65.310  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.117 on 70 degrees of freedom
Multiple R-squared:  0.9839,    Adjusted R-squared:  0.9836
F-statistic:  4265 on 1 and 70 DF,  p-value: < 2.2e-16
```

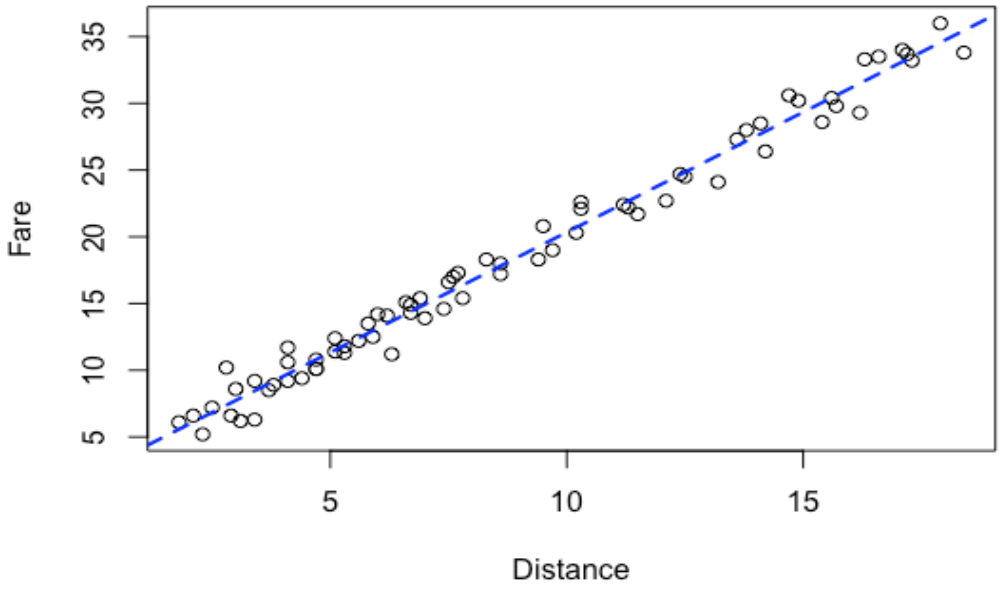
The summary supports my comment about the positive relationship of the two variables. In fact, for every increase in the distance, the fare increases by $ 1.8 (2sf). Also, when there is no distance travelled, 0 km, the fare is already at $ 2.3 which is not applicable because the rider was not even in the taxi at this point.

3.

```
> summary(taxi.lm)$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.336595 0.27549518  8.481436 2.371651e-12
Distance    1.799850 0.02755855 65.310035 1.835202e-64
```

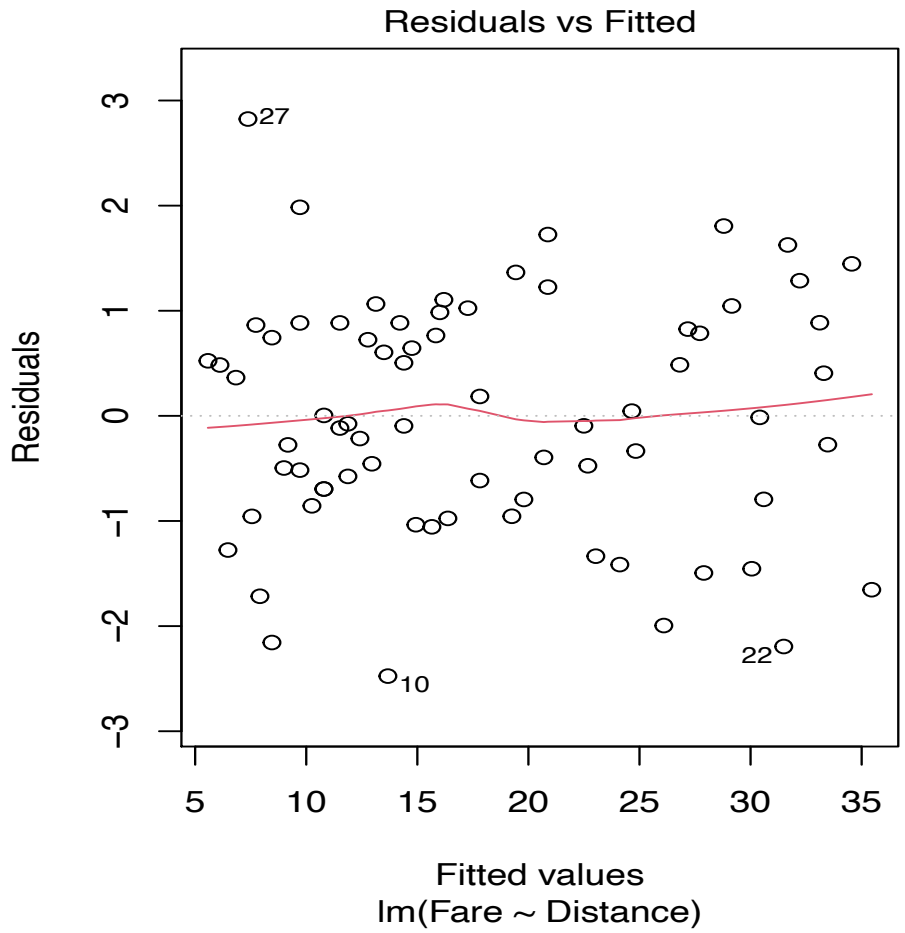
4.

```
> plot(Fare ~ Distance, data = taxi.df)
> abline(2.34, 1.80, lty = 2, lwd = 2, col = "blue")
```



5.

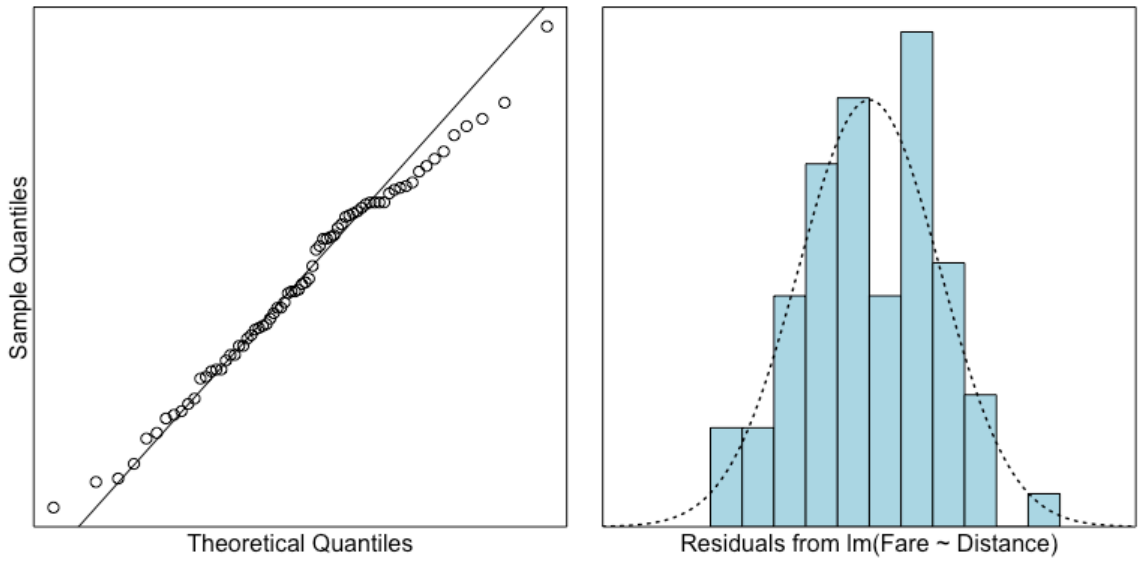
```
> plot(taxi.lm, which = 1)
```



The residual plot has a constant scatter around 0 showing no trend. Most of the points fall into the +- 1 residual band from 0.

6.

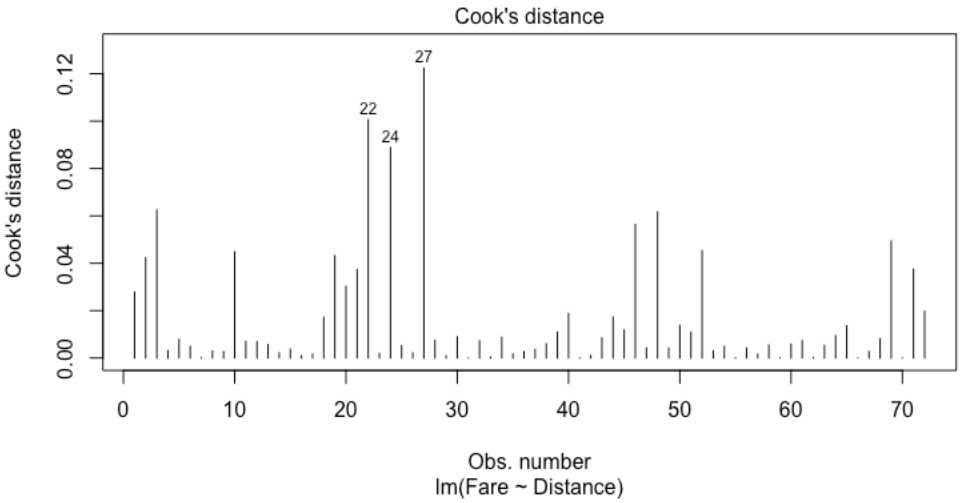
```
> library(s20x)
> normcheck(taxi.lm)
```



To further check for the normal distribution of the sample then we would need normal quantile-quantile plot and histogram. The quantile-quantile plot follows the normal distribution as they are close to the regression line. In addition, the histogram also shows a normal distribution. The sample is also random therefore independent.

7.

```
> plot(taxi.lm, which = 4)
```



The plot does not show any influential points as each point fall under 0.4. Thus, no point should be removed. So far, the sample meet the assumptions of a linear model and requirements for Cook's distance so I will carry on with my analysis.

8.
The sample gives a regression line equation of E[Fare] = 2.34 + 1.80 * Distance.
This equation is useful for predicting the fare of a taxi ride using the distance its travelled.

9.
For example, if a ride travelled 12 km then its fare would be:
E[Fare] = 2.34 + 1.80 * 12
E[Fare] = $ 23.94 (2dp)

10.
We can use the same equation to calculate residuals. For example, the 10^{th} observation in the original dataset has a distance of 6.3 km. To get its residual, then do number 9 like usual:
E[Fare] = 2.34 + 1.80 * 6.3
E[Fare] = $ 13.68 (2dp)
Then substitute E[Fare] to the residual equation = 10^{th} observation fare - E[Fare]
11.2 – 13.68 = $ -2.48 (2dp)

Therefore, the residual of the 10^{th} observation is $-2.48 which means that the actual fare below the regression line.

Conclusion:

I conclude that the sample is useful for predicting the fare of the taxi ride using the distance travelled by the taxi ride. This would be helpful for frequent riders of taxi or Uber.

Question 2: Diamonds

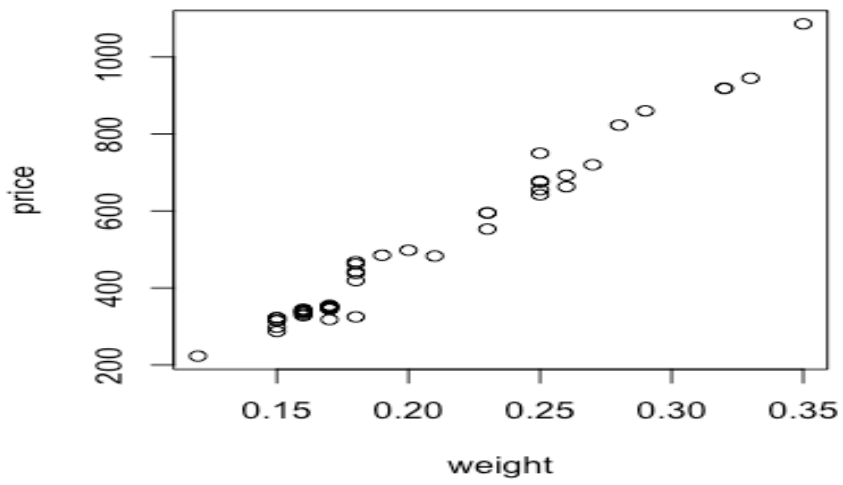
The sample was randomly taken from a single retailer of diamond rings. Weight (in carats) is the number of carats of diamond rings and price (in $) is the cost of the diamond rings.

In this assignment, I will investigate if I can use the weight of the diamond rings to predict the price of the ring.

1.

```
> diamonds.df = read.csv("diamonds.csv")
> View(diamonds.df)
> head(diamonds.df, n=6)
  weight price
1   0.17   355
2   0.16   328
3   0.17   350
4   0.18   325
5   0.25   642
6   0.16   342
```

```
> plot(price ~ weight, data = diamonds.df)
```



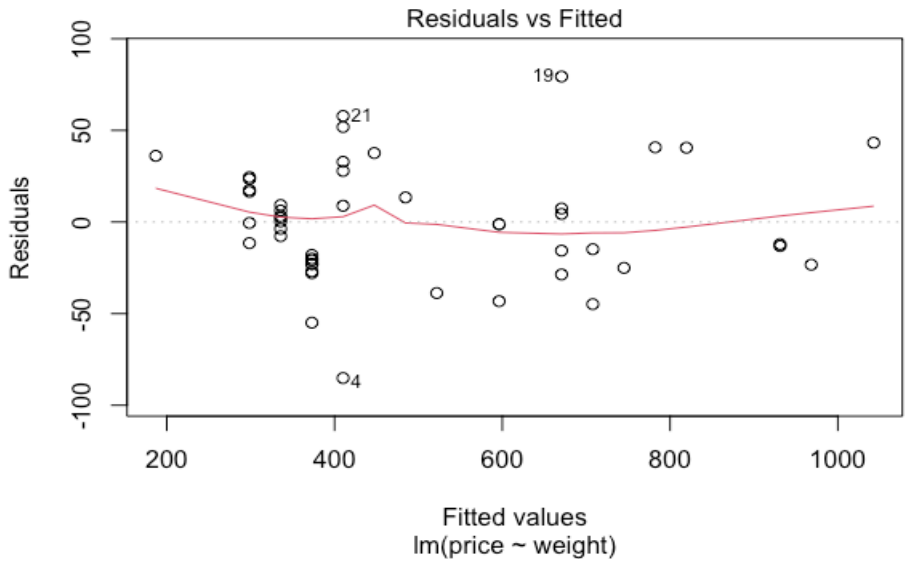
The plot shows a positive relationship between the weight and the price of diamond rings. As the weight increases, the price increases as well.

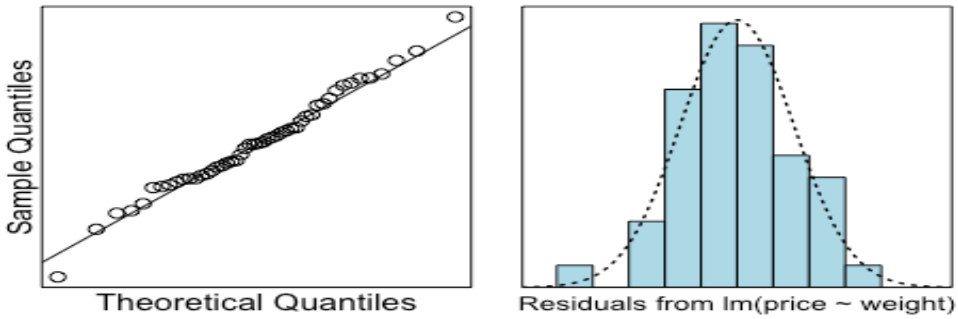
2.

```
> diamonds.lm = lm(price ~ weight, data = diamonds.df)
> summary(diamonds.lm)

Call:
lm(formula = price ~ weight, data = diamonds.df)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

3.

```
plot(diamonds.lm, which = 1)
library(s20x)
normcheck(diamonds.lm)
```

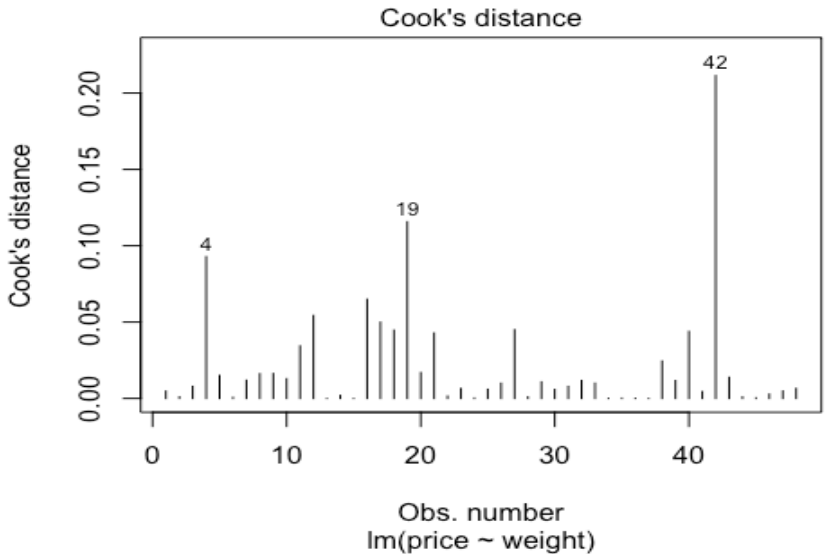




The residual plot shows a reasonable scatter around 0 with no trend. Also, most of the points fall into the +- 50 residual band. The normal quantile-quantile plot shows a normal distribution as the points are closed to the regression line. This is supported by the histogram as it is also showing a normal distribution. The sample of the diamond rings is randomly selected thus independent. Overall, the sample satisfies the assumptions of a linear model.

4.

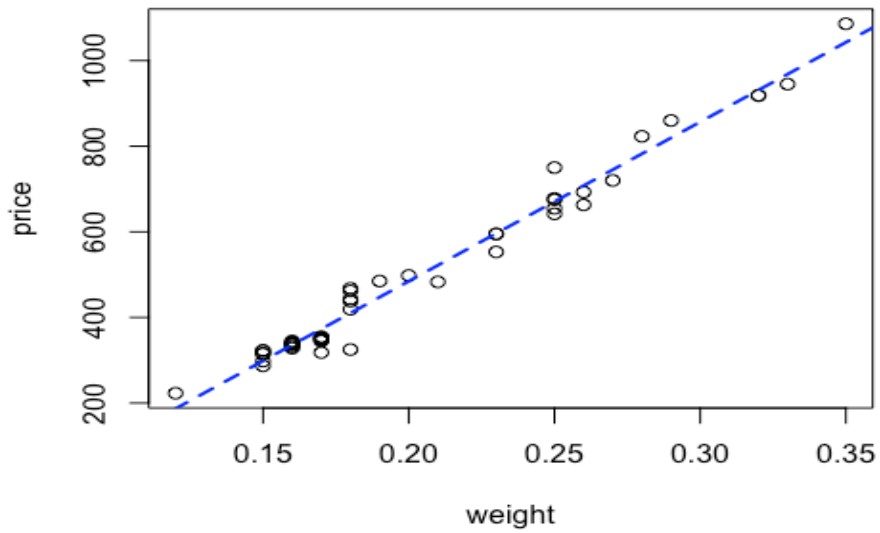
```
plot(diamonds.lm, which = 4)
```



The plot shows no influential points as all the points fall under 0.4 Cook Distance. Therefore, I will carry on with my analysis because the sample meet the assumptions of a linear model and the requirement of Cook's distance.

5.

```
plot(price ~ weight, data = diamonds.df)
abline(-259.6, 3721.02, lty = 2, lwd = 2, col = "blue")
```



6.
The regression line equation of the sample is:
E[price] = -259.63 + 3721.02 * weight.
This equation is useful for predicting the price of a diamond using its weight (in carats).

7.
For example, if a ring has a weight of 0.2 carats then its price would be:
E[price] = -259.63 + 3721.02 * 0.2
E[price] = $ 484.57 (2dp)

8.
We can use the same equation above to calculate residuals. For example, the 5th observation in the original dataset, has a 0.25 weight. To get its residual, then 7 like usual
E[price] = -259.63 + 3721.02 * 0.25
E[price] = $ 670.625
Then substitute E[price] to the residual equation = 5th observation price - E[price]
$ 642 – $ 670.625 = $ -28.63 (2dp)

Therefore, the residual in the 5th observation is $-28.63 which means that the actual price is below the regression line.

Conclusion:

I conclude that the sample is useful for predicting the price of diamond rings using its weight (in carats). This would be helpful for jewellery shop for putting the right price for their diamond rings.