TASK 1

- Refer to the dataset of incinerator
- Find five problem statements from the entire dataset
- Model the relationships
- Find the solutions of the problem statements
- Justify the solutions with concrete detail

Five problem statements from the entire dataset

1. Is there any significant correlation between the selling price (`price`) and distance from the incinerator (`dist`) which can be further used for prediction?
2. How does each variable altogether impact the selling price?
3. Is there any significant correlation between the selling price (`price`) and logarithm of distance from the incinerator (`ldist`) which can be further used for prediction?
4. Which relation model will be the best to predict selling price according to backend elimination?
5. Which polynomial regression model will be the best to predict selling price with age?

Models of the relationships

I. Linear regression model of selling price and distance from the incinerator:

To extract data from the file called 'Incinerator':-

```
library(readxl)
Incinerator <- read_excel("C:/Users/Aysha Emelda/Downloads/Incinerator.xlsx")
View(Incinerator)
```

To preprocess the data:

```
library(caTools)
set.seed(123)
split <- sample.split(dataset$price, SplitRatio = 2/3)
training_set <- subset(dataset, split == TRUE)
test_set <- subset(dataset, split == FALSE)
View(test_set)
```

Variables price and dist are made fit into a linear model where price is considered the dependent variable and dist as the independent variable.

```
regressor <- lm(formula = price ~ dist,
                data = training_set)
regressor
```

```
summary(regressor)
```

The summary of `regressor` is shown in the image below:

```
Call:
lm(formula = price ~ dist, data = training_set)

Coefficients:
(Intercept)         dist
  72715.662        1.014


Call:
lm(formula = price ~ dist, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-66379 -28841 -11990  22594 211574

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.272e+04  7.104e+03  10.236  < 2e-16 ***
dist        1.014e+00  3.184e-01   3.184  0.00167 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40090 on 212 degrees of freedom
Multiple R-squared:  0.04563,   Adjusted R-squared:  0.04113
F-statistic: 10.14 on 1 and 212 DF,  p-value: 0.001673
```

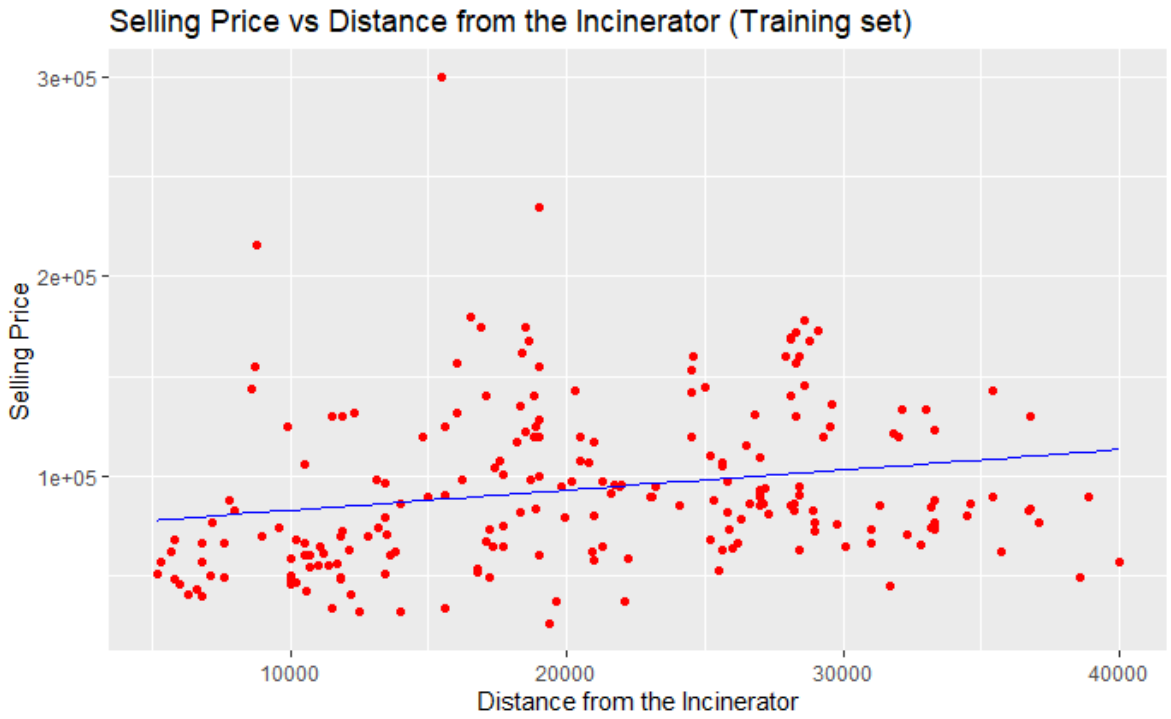
Prediction is done with test set:

```
y_pred <- predict(regressor, newdata =
test_set)
```

A scatter diagram of price versus `dist` from the `training_set` is made for visualisation of data:

```
library(ggplot2)

plot1<-ggplot() +

  geom_point(aes(x = training_set$dist, y =
training_set$price),

             colour = 'red') +
```

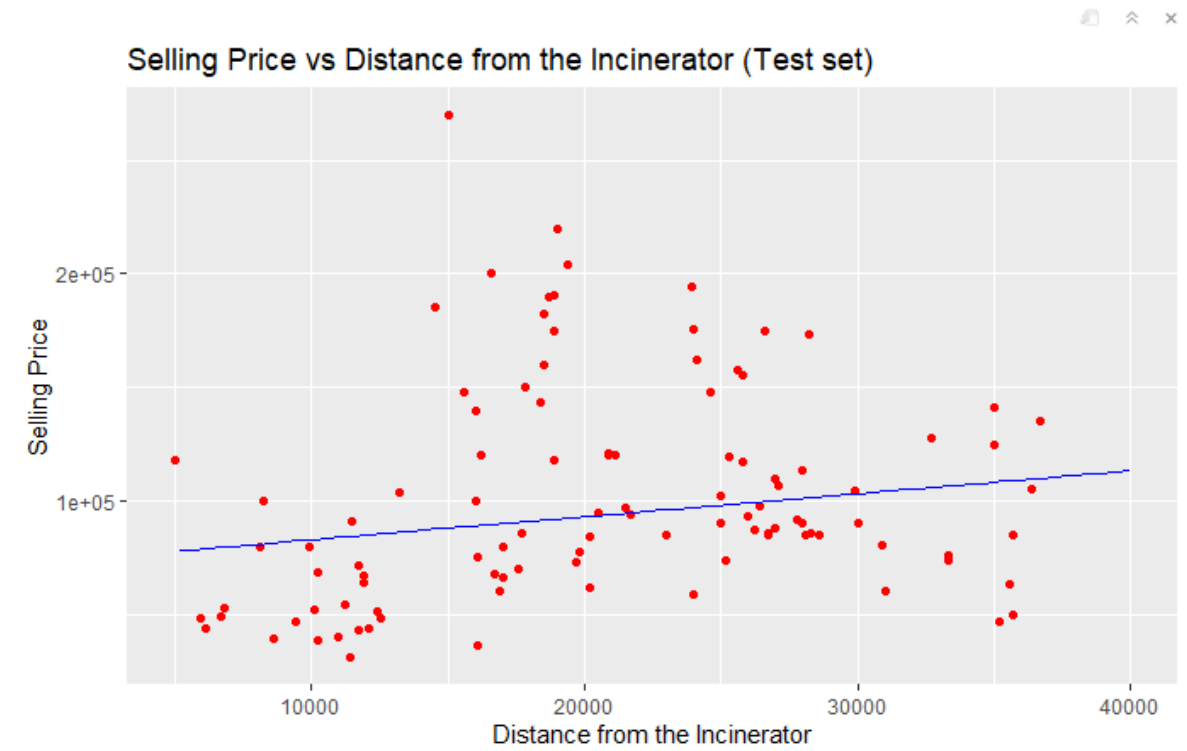
```
  geom_line(aes(x = training_set$dist, y =
predict(regressor, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Selling Price vs Distance from the
Incinerator (Training set)') +
  xlab('Distance from the Incinerator') +
ylab('Selling Price')
print(plot1)
```



A scatter plot of prediction with `test_data` is produced for visualisation of data:

```
plot2 <-ggplot() +
  geom_point(aes(x = test_set$dist, y =
test_set$price),
             colour = 'red') +
```

```
  geom_line(aes(x = training_set$dist, y =
predict(regressor, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Selling Price vs Distance from the
Incinerator (Test set)') +
  xlab('Distance from the Incinerator') +
  ylab('Selling Price')
print(plot2)
```


Selling Price vs Distance from the Incinerator (Test set)
Selling Price
2e+05
1e+05
10000
20000
30000
40000
Distance from the Incinerator

II. Multiple linear regression model of selling price and all other variables:

```
mregressor <- lm(formula = price ~ .,
                 data = training_set)
summary(mregressor)
```

y_pred1 <- predict(mregressor, newdata1 = test_set)

```
Residuals:
     Min       1Q   Median       3Q      Max
-17830.6  -2264.6   -211.5   2376.7  18170.7

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.375e+10  2.508e+10  -2.940  0.00369 **
year         3.728e+07  1.268e+07   2.940  0.00369 **
age         -5.745e+01  4.113e+01  -1.397  0.16411
agesq        3.433e-01  2.569e-01   1.336  0.18307
nbh         -2.617e+02  1.808e+02  -1.448  0.14937
cbd          5.560e-01  4.285e-01   1.297  0.19606
intst        3.227e-02  6.646e-01   0.049  0.96132
lintst       1.936e+04  2.947e+04   0.657  0.51209
rooms       -3.991e+02  4.537e+02  -0.880  0.38020
area         4.821e-01  2.329e+00   0.207  0.83623
land        -3.194e-03  1.001e-02  -0.319  0.75013
baths        1.529e+02  6.988e+02   0.219  0.82708
dist        -5.150e-02  4.352e-01  -0.118  0.90594
ldist       -5.026e+03  5.044e+03  -0.996  0.32031
wind         1.583e+02  3.096e+02   0.511  0.60961
lprice      -4.238e+08  1.442e+08  -2.940  0.00369 **
y81                 NA         NA      NA       NA
larea        1.281e+03  4.638e+03   0.276  0.78262
lland       -1.023e+03  8.417e+02  -1.216  0.22553
y81ldist     3.097e+03  2.467e+03   1.255  0.21091
lintstsq    -1.309e+03  1.920e+03  -0.682  0.49620
nearinc      5.847e+02  1.921e+03   0.304  0.76119
y81nrinc    -6.035e+03  2.497e+03  -2.417  0.01659 *
rprice       1.051e+00  3.484e-02  30.154  < 2e-16 ***
lrprice      4.239e+08  1.442e+08   2.940  0.00369 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4171 on 190 degrees of freedom
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9896
F-statistic: 883.9 on 23 and 190 DF,  p-value: < 2.2e-16
```

III. Linear regression model of selling price and logarithm of distance from the incinerator:

A linear regression model of `price` and `ldist` is made:

```
regressor <- lm(formula = price ~ ldist,
                data = training_set)
regressor
summary(regressor)
```

```
Call:
lm(formula = price ~ ldist, data = training_set)

Coefficients:
(Intercept)        ldist
    -127235        22469


Call:
lm(formula = price ~ ldist, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-68606 -27128 -11210  21000 210437

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -127236      54747  -2.324   0.0211 *
ldist          22470       5564   4.038 7.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39540 on 212 degrees of freedom
Multiple R-squared:  0.07143,   Adjusted R-squared:  0.06705
F-statistic: 16.31 on 1 and 212 DF,  p-value: 7.524e-05
```

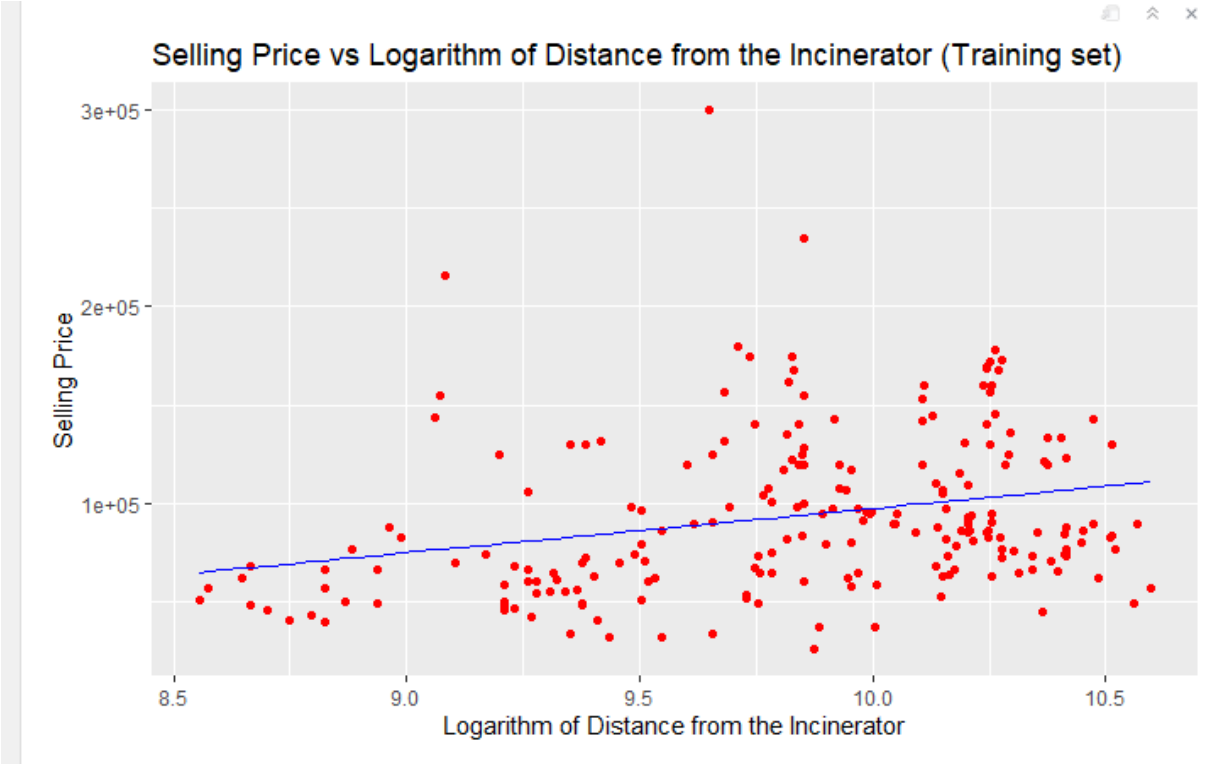
The summary of the linear model is above.

Prediction is done with test set:

```
y_pred <- predict(regressor, newdata = test_set)
```

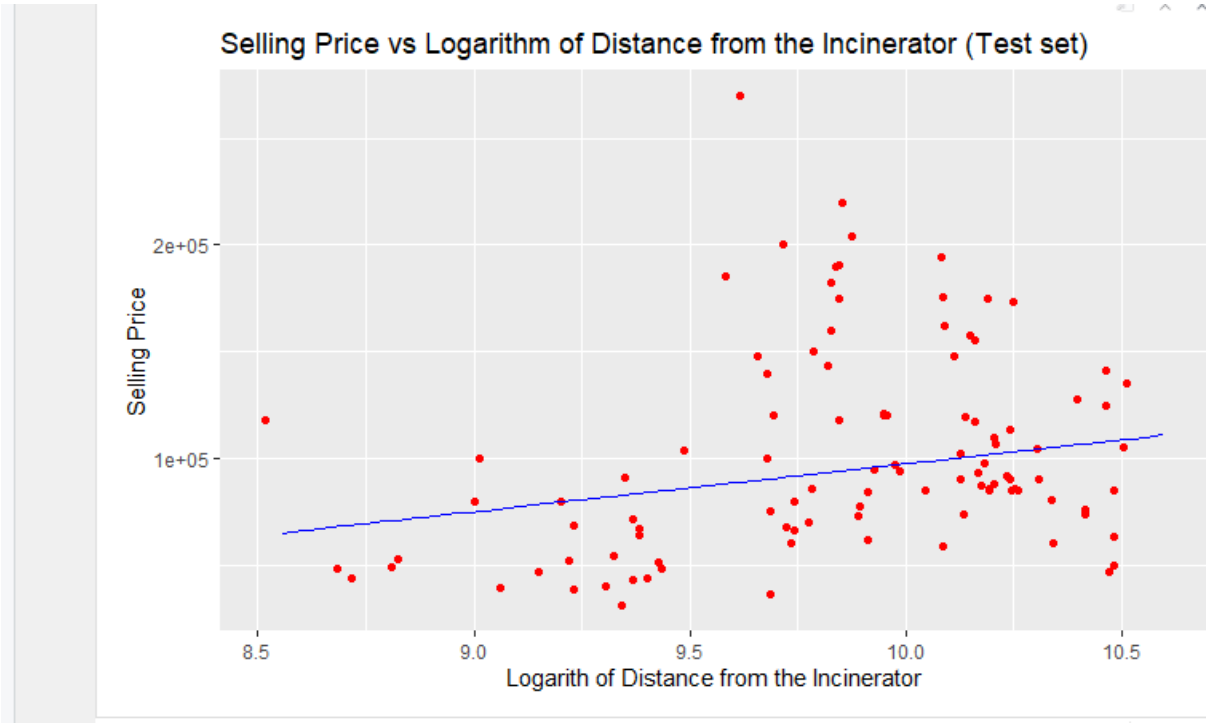
A scatter diagram of price versus `ldist` from the `training_set` is made for visualisation of data:

```
plot1<-ggplot() +
  geom_point(aes(x = training_set$ldist, y =
training_set$price),
             colour = 'red') +
  geom_line(aes(x = training_set$ldist, y =
predict(regressor, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Selling Price vs Distance from the
Incinerator (Training set)') +
  xlab('Logarithm of Distance from the
Incinerator') +
ylab('Selling Price')
print(plot1)
```



A scatter plot of prediction with `test_data` is produced for visualisation of data:

```
plot2 <-ggplot() +
  geom_point(aes(x = test_set$ldist, y =
test_set$price),
             colour = 'red') +
  geom_line(aes(x = training_set$ldist, y =
predict(regressor, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Selling Price vs Logarithm of Distance
from the Incinerator (Test set)') +
  xlab('Logarith of Distance from the
Incinerator') +
  ylab('Selling Price')
print(plot2)
```



IV. Backend elimination from multiple linear regression

From the multiple linear regression we did already, we know the variables with the least significance. Having eliminated them, a model is made out of remaining variables:

```
mregressor <- lm(formula = price ~
age+agesq+nbh+cbd+rooms,
                 data = training_set)
summary(mregressor)
```

```
Call:
lm(formula = price ~ age + agesq + nbh + cbd + rooms, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-90668 -21109  -7193  17816 180326

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 40907.7559 22087.3888   1.852 0.065430 .
age         -1245.9366   244.8807  -5.088 8.07e-07 ***
agesq           6.1275     1.6631   3.684 0.000292 ***
nbh         -2459.0156  1105.7063  -2.224 0.027228 *
cbd            -0.2694     0.3173  -0.849 0.396757
rooms       11675.6821  3085.7701   3.784 0.000202 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33310 on 208 degrees of freedom
Multiple R-squared:  0.3535,    Adjusted R-squared:  0.338
F-statistic: 22.75 on 5 and 208 DF,  p-value: < 2.2e-16
```

`cbd` has no significance, so eliminate `cbd` and make the model.

```
Call:
lm(formula = price ~ age + agesq + nbh + rooms, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-89600 -20917  -7943  17135 182135

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36820.483  21542.188   1.709 0.088893 .
age         -1143.829    213.184  -5.365 2.14e-07 ***
agesq           5.570      1.527   3.648 0.000334 ***
nbh         -2742.968   1053.229  -2.604 0.009866 **
rooms       11577.807   3081.559   3.757 0.000223 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33290 on 209 degrees of freedom
Multiple R-squared:  0.3513,    Adjusted R-squared:  0.3389
F-statistic: 28.29 on 4 and 209 DF,  p-value: < 2.2e-16
```

In the new model, the significance of `nbh` increased. The adjusted R-squared increased by a small fraction. These indicate that it is a better model than the previous one.

Since the significance of `nbh` is comparatively low, in the next model we eliminate `nbh`.

```
mregressor <- lm(formula = price ~
age+agesq+rooms,
                 data = training_set)
summary(mregressor)
```

```
Call:
lm(formula = price ~ age + agesq + rooms, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-84280 -22292  -7419  19575 187395

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28855.263  21615.568   1.335 0.183345
age         -1173.606    215.788  -5.439 1.49e-07 ***
agesq           5.658      1.547   3.656 0.000323 ***
rooms       11964.253   3120.074   3.835 0.000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33740 on 210 degrees of freedom
Multiple R-squared:  0.3302,    Adjusted R-squared:  0.3207
F-statistic: 34.51 on 3 and 210 DF,  p-value: < 2.2e-16
```

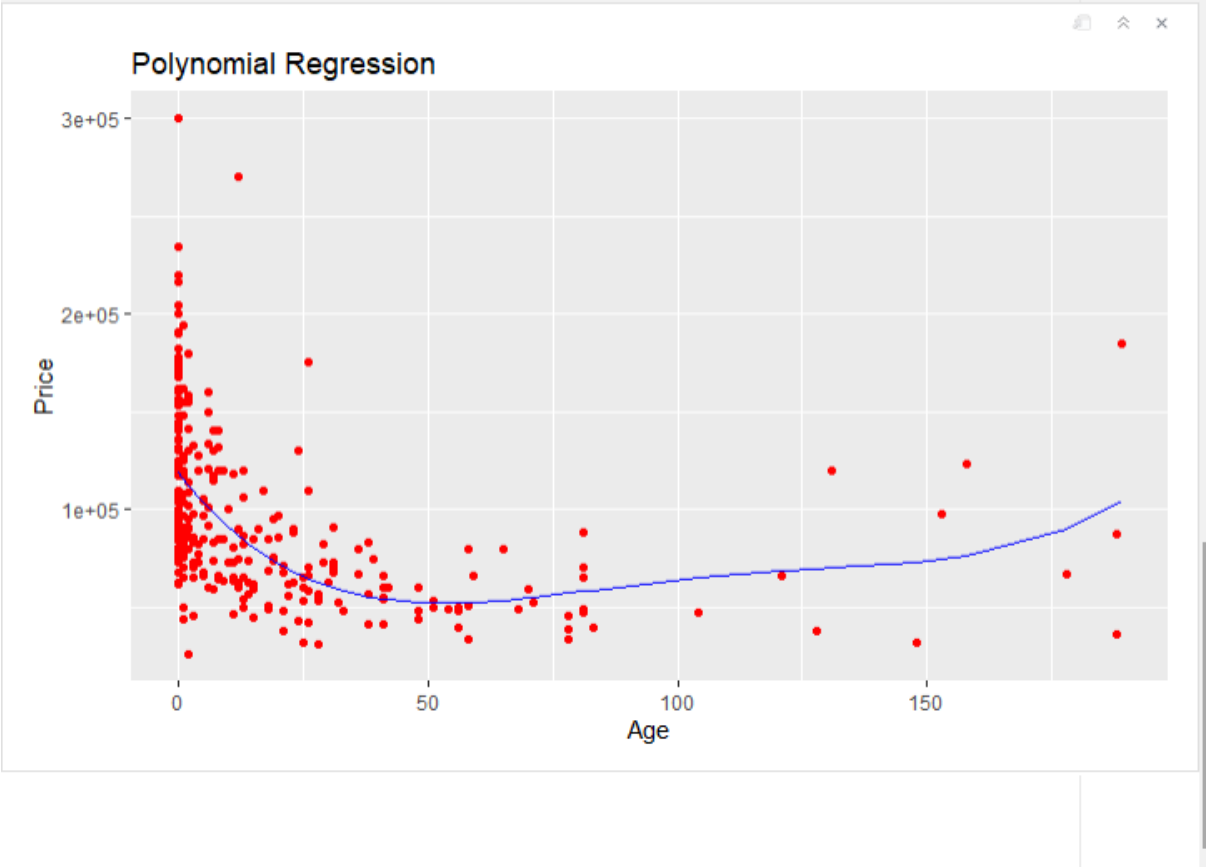
The adjusted R-squared is lower than that of the previous model. It implies that the previous model is a better one.

V. Polynomial regression model of selling price and age:

First we consider age until the power of 4:

```
Incinerator$agecube<- Incinerator$age^3

Incinerator$agepower4<- Incinerator$age^4

poly_reg <- lm(formula= price~ age+agesq+agecube+agepower4, data=
Incinerator)

summary(poly_reg)

library(ggplot2)

ggplot()+

  geom_point(aes(x=Incinerator$age,
y=Incinerator$price),colour='red')+

  geom_line(aes(x=Incinerator$age, y=predict(poly_reg,
newdata=Incinerator)), colour='blue')+

  ggtitle('Polynomial Regression')+

  xlab('Age')+

  ylab('Price')
```

```
Call:
lm(formula = price ~ age + agesq + agecube + agepower4, data = Incinerator)

Residuals:
   Min     1Q Median     3Q    Max
-86358 -24691  -5748  20408 183328

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.187e+05  2.853e+03  41.602  < 2e-16 ***
age         -3.275e+03  4.829e+02  -6.781 5.88e-11 ***
agesq        5.470e+01  1.561e+01   3.505 0.000523 ***
agecube     -3.580e-01  1.557e-01  -2.299 0.022138 *
agepower4    8.368e-04  4.679e-04   1.789 0.074645 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36010 on 316 degrees of freedom
Multiple R-squared:  0.3145,    Adjusted R-squared:  0.3058
F-statistic: 36.24 on 4 and 316 DF,  p-value: < 2.2e-16
```



Now we create another model excluding agepower4 because its significance is only in the border. The summary of the model is below:

```
Call:
lm(formula = price ~ age + agesq + agecube, data = Incinerator)

Residuals:
   Min     1Q Median     3Q    Max
-86092 -26137  -5852  18396 182828

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.172e+05  2.733e+03  42.871  < 2e-16 ***
age         -2.597e+03  3.004e+02  -8.643 2.75e-16 ***
agesq        2.864e+01  5.609e+00   5.105 5.71e-07 ***
agecube     -8.269e-02  2.352e-02  -3.517 0.000501 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36140 on 317 degrees of freedom
Multiple R-squared:  0.3076,    Adjusted R-squared:  0.301
F-statistic: 46.93 on 3 and 317 DF,  p-value: < 2.2e-16
```

Find the solutions of the problem statement

Solution to problem statement 1:

A simple linear regression model of `price` and `dist` was made. The analysis of the summary of `regressor` gave insights on the signifance of the variable. Correspoding to that, the scatter plot gave us insights, visually, on the correlation between `price` and `dist`.

The summary of the linear model provided p-value of `dist` which is below 0.05 (threshold value), between 0.001 and 0.01. Thus, it is indicated that `dist` is significant as a dependent variable to `price` which is the

independent variable here. However, it is not of highest significance. From the positive sign of coefficient, it is derived that correlation is positive. In the test dataset, the prediction did not perform as good as it performed in the training dataset.

Solution to Problem statement 2:

The impact of all the variables altogether on the price is assessed with the summary of `mregressor.` With the analysis of p-values, in this multiple linear regression, it is concluded that all variables about which we are keen to see the impact are insignificant. That does not mean that all the variables are insignificant individually. We can see that in the solution of problem statement 1.

Solution to Problem Statement 3:

According to the summary of the simple linear model, logarithm of distance from the incinerator is of very high significance. In the scatter plot also we can visually experience the significance. As the sign of the coefficient is positive, the correlation is positive.

Solution to Problem Statement 4:

Through backward elimination, the model which as variables `age`, `agesq`, `rooms` and `nbh` is the best one to predict `price`.

Solution to problem statement 5:

A polynomial regression model between `price` and `age` are possible. Through backward elimination from powers till 4, the model with the first 4 powers of age is better than that with lesser powers. This is derived by considering the values of adjusted R-squared from the summary.

Justify the solutions with concrete details

We made the models, went through various steps and made assumptions based on p-value and adjusted R-squared. Models were selected through backend elimination.

The p-value has to lesser than 5 percent for the variable to be significant. The lower it is, the higher the significance is. This particular detail is used to accept or reject independent variable and form new models.

The models are compared using the adjusted R-squared. Adjusted R-squared is used instead of R-squared, because the R-squared is highly dependent on the number of variables. The higher the value of adjusted R-squared, the better the model is.

We did backend elimination to select the better model. Variable with the least significance determined using p-value was removed and thus the next models were made until the best model was found by looking into the value of the adjusted R-squared.

In the model for first problem statement, we see that the p-value is between 0.01 and 0.001. This means that the correlation is significant, however not of highest significance. Since the coefficient is signed positive, it is a positive correlation.

In the model for second problem statement, we see that the significance is low in most of the variables. This is because we considered a model with all the variables impacting each other and the price. That does not mean that all the variables are insignificant individually.

Consider the model of third problem statement. According to the summary of the simple linear model, logarithm of distance from the incinerator is of very high significance. In the scatter plot also we can visually experience the significance. As the sign of the coefficient is positive, the correlation is positive.

Consider the models of fourth problem statement. Independent variables with least significance were eliminated. Thus a group of models is made. Among these one model is selected according to the adjusted R-squared. Its value is supposed to get higher, closer towards 1, as the model betters through backend elimination. At a particular model, the value went lower than that of previous model. We selected the previous model as the best one.Through backward elimination, the model which has independent variables `age`, `agesq`, `rooms` and `nbh` is the best one to predict the dependent variable `price`.

In the models of the fifth problem statement, we tried polynomial regression with age. As the power increased to 4, the better the model was. This was examined with the help of adjusted R-squared as mentioned in the solution.

****