Statistical packages - report 1

Urszula Grochocińska, Marcin Mazurkiewicz

December 3, 2018

Contents

1 Introduction

For this report, we have decided to analyze "Diamonds" dataset about the features and prices of stones which we have found on Kaggle [https://www.kaggle.com/shivam2503/diamonds]. It contains 54940 records and each one has 11 columns with different variables.

First one is irrelevant because it's just index number. Then we have numerical variables - x, y, z that describe a size of a diamond in each of three dimensions. There is also weight given in carats. Today, a carat is equal to exactly 0.2 grams. Carat weight is unrelated to the similar sounding karat, which refers to gold's purity. Next, we have two variables describing percentage relation between appropriate measures within the diamond 1. Depth is the height of a diamond, measured from the culet to the table, divided by its average girdle diameter. The table is the width of the diamond's table expressed as a percentage of its average diameter. The last numerical variable is a price. Coming to discrete variables we have cut describing the quality of a diamond's cut: Fair, Good, Very Good, Premium, Ideal. The cut describes the symmetry proportioning and polish of the diamond. Then there is a color of diamond ranged descending from D to J. The later letter in the alphabet the more yellow is the diamond. And the last is clarity of particular diamond with possible values: FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 (ordering from best to worse). It refers to the absence of inclusions and blemishes.

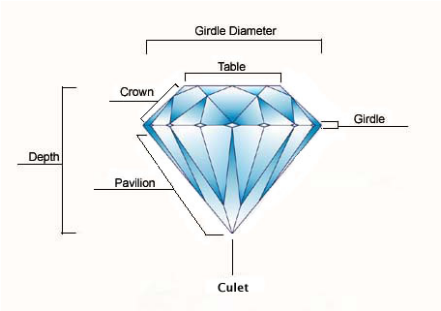

Figure 1: Anatomy of diamond. *source: https://www.everything-wedding-rings.com*

2 Aim of analysis

Jewelry is a kind of art made of expensive raw material. This art is based on precision and integrity but because of the fact that it is an art sometimes, the defects increase the price. These defects make diamond unique and uniqueness is the main part of the art. In our analysis we would like to check which features influence the price and at which level. Besides we are interested in how important are all features that are influenced by a jeweler(shape, cut) and that what is natural (color, clarity). The figure 2 is a visualization of the distribution of prices. The chart shows that the most common are relatively cheap diamonds. Then if the price increase the diamond becomes rare.

3 Data preprocessing

As a first step of working with data we start from removing unnecessary variables. We remove the index column because it is irrelevant information. Then we look for missing data. Dataset doesn't contain any NaNs, but some values are impossible to occur in reality. To handle such situations we change every zero from x, y or z column to NA, because this makes our upcoming work easier. We also check if other numerical columns contain zeros, but they don't. Next, we focus on ensuring that categorical variables don't contain typos. In order to do that, we just look at all values they take, but there is no need for cleaning. We don't deal with outliers, because this is part of further steps in our analysis.

```
> diamonds_data <- diamonds_data[2:11] #data cleaning - removing unnecessary variables
> #changing 0 to NAs
> diamonds_data$z[diamonds_data$z==0] = NA
> diamonds_data$x[diamonds_data$x==0] = NA
> diamonds_data$y[diamonds_data$y==0] = NA
> any(is.na(diamonds_data)) #check if there are any nulls in our dataset

[1] TRUE

> diamonds_data_omited <- na.omit(diamonds_data)
> #checking for typos
> summary(diamonds_data$cut)

     Fair      Good     Ideal   Premium Very Good
     1610      4906     21551     13791     12082

> summary(diamonds_data$color)

    D     E     F     G     H     I     J
 6775  9797  9542 11292  8304  5422  2808
```

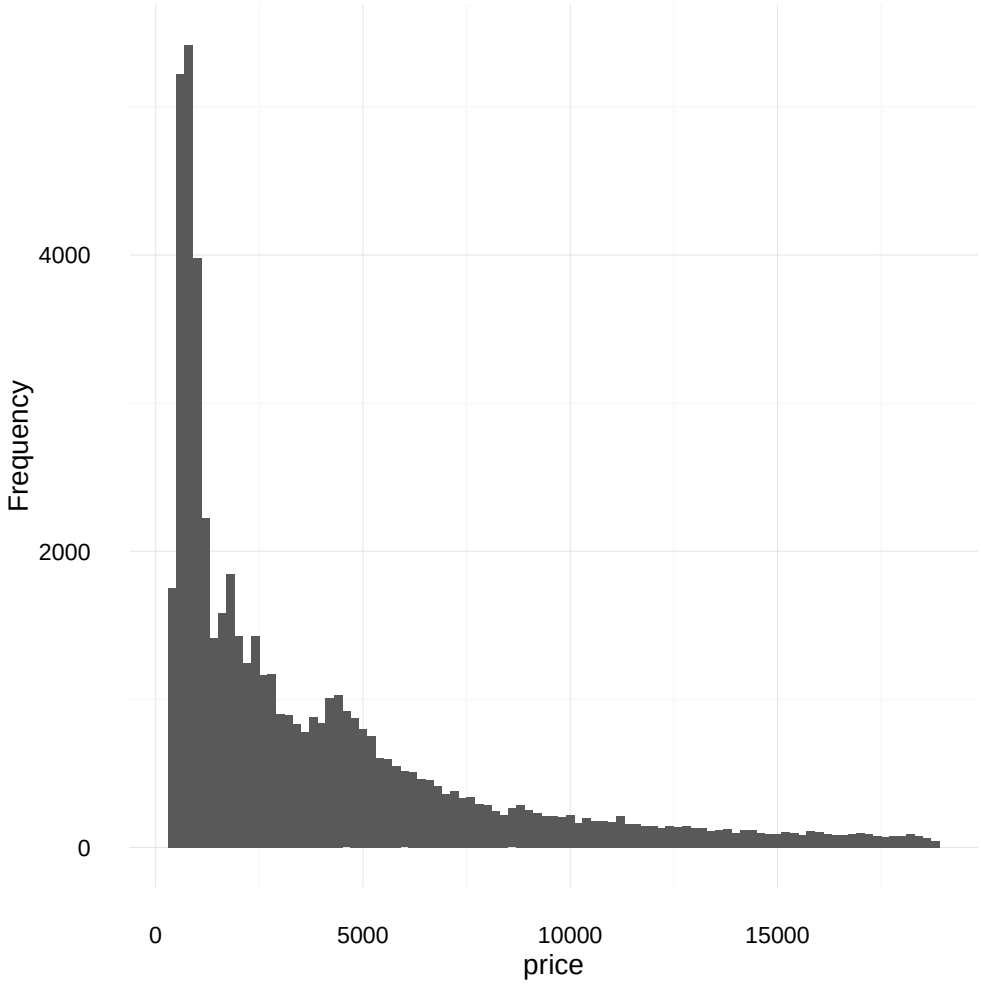

Figure 2: Distribution of values of price in our dataset

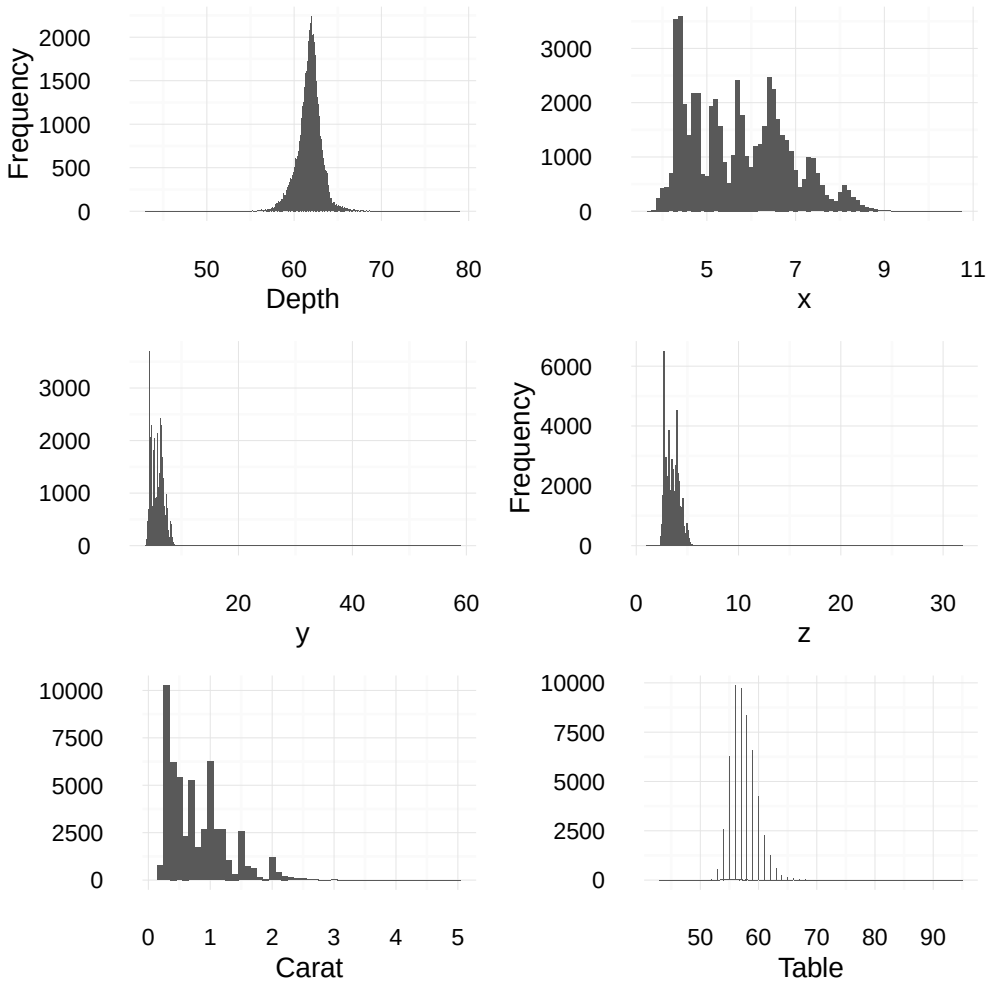

Figure 3: Distributions of each of continuous variables

```
> summary(diamonds_data$clarity)

   I1    IF   SI1   SI2   VS1   VS2  VVS1  VVS2
  741  1790 13065  9194  8171 12258  3655  5066
```

Before we start fitting the model we make quick look on the data. In figures 3 and 4 we can see histograms and box plots. The plot of depth is symmetric similar to the bell curve. The next one (x) is much more dispersed. The distributions of y, z and especially table show that this data have small differences in each group.

4 Model fitting

4.1 Continuous variables

To fit model to data we use function glmulti from package glmulti. To reduce complexity of model at this step we decide to choose maximum number of used variables to 3. We check models with different distribution

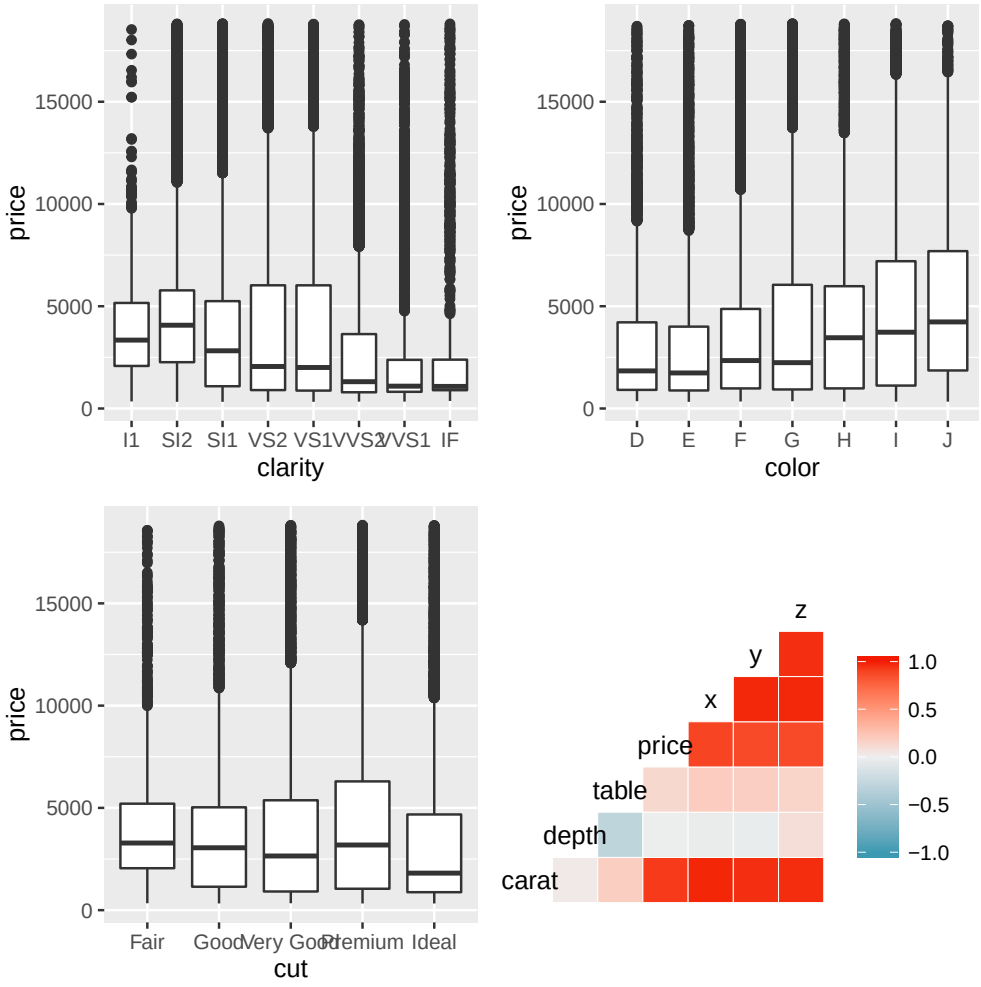

Figure 4: Boxplots of price depending on values of categorical variables and the correlation chart between each two variables

color(C_i)	E	F	G	H	I	J
coefficient	-214.2	-282.6	-486.8	-993.2	-1477.5	-2397.4

Table 1: Table with coefficients of color depending on type of diamond's color.

clarity(D_i)	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
coefficient	5707.1	3922.3	2952.6	4878.2	4559.3	5339.5	5266.3

Table 2: Table with coefficients of clarity depending on type of diamond's clarity.

families. These are Gaussian, Poisson and Gamma. It turns out that Gamma model provides the best fit under the BIC criteria. In first model the function which finds the best model under AIC critera choose two size variables: x, z and the carat. In the figure 5 is shown the output of the model. Because of the fact that our dataset has about 10k records we have ploted only a part of them. The visible results of the model cover the real values.

At the plot we can clearly see that there are outliers, but we handle them in next sections.

4.2 Additional categorical variables

In case of additional categorical variables, we use a very similar approach as in continuous variables. But this time we increase the maximum number of variables to 4 and consider Gaussian family. The best model we find consists of two continous variables: x, carat and two categorical: color and clarity. It has different coefficients. For the x variable it is equal to -1016.9, for carat it is on the level 11192.5. The coeficetnts for categorical variables are shown in the table 1 (for colors) and table 2 (for clarity). Similar as in the first case we plot 6 the choosen points of real prices and the calculated points from model. Also in this case the fited linas cover the real values.

4.3 Model with additional interactions

To find model with additional interactions we use very similar approach as in previous cases. Maximum number of variables is 4 and the family is Gaussian. The best model we find consists of two continous variables: x, carat and two categorical: color and clarity. It has different coefficients. For the x with interaction with carat the coeficient is equal to 1527.5, for single carat it is on the level - 6912.8. The coeficetnts for categorical variables are shown in the table 3 (for colors with carat interaction) and table 4 (for clarity with carat interaction). Similar as in the first case we plot 7 the choosen points of real prices and the calculated points from model. As in previous cases the difference beetween model is not visible on the plot.

5 Other models

We use glmulti function to find all of considered models. But it turned out, that in second and third case the best model doesn't include z variable. That arouses our curiosity if model excluding this variable is significantly better than with it. Under BIC criteria best model with categorical variables with additional z is only 0.02 % better than without it. Although it gives better BIC it contains more variables so it may be considered as worse model.

carat $\cdot$ color(C_i)	E	F	G	H	I	J
coefficient	-219.1	-342.1	-737.5	-1373.5	-1920.8	-2786.8

Table 3: Table with coefficients of carat $\cdot$ color depending on type of diamond's color.

```
> model_simple <- glm(formula = price~z+x+carat, data = diamonds_data_omited,
+                     family = Gamma(link = 'log'))
> plot(seq(1, to=54000, by = 50), model_simple$fitted.values[seq(1, to=54000, by = 50)],
+      cex=.8, xlab = 'Number of observation', ylab = 'Diamond price')
> points(seq(1, to=54000, by = 50), diamonds_data_omited$price[seq(1, to=54000, by = 50)],
+        col = 'red', pch = '.', cex=2)
> legend(35000, 21000, legend=c("Fitted values", "Real values"), col=c("black", "red"), lty=3, cex=.8)
```

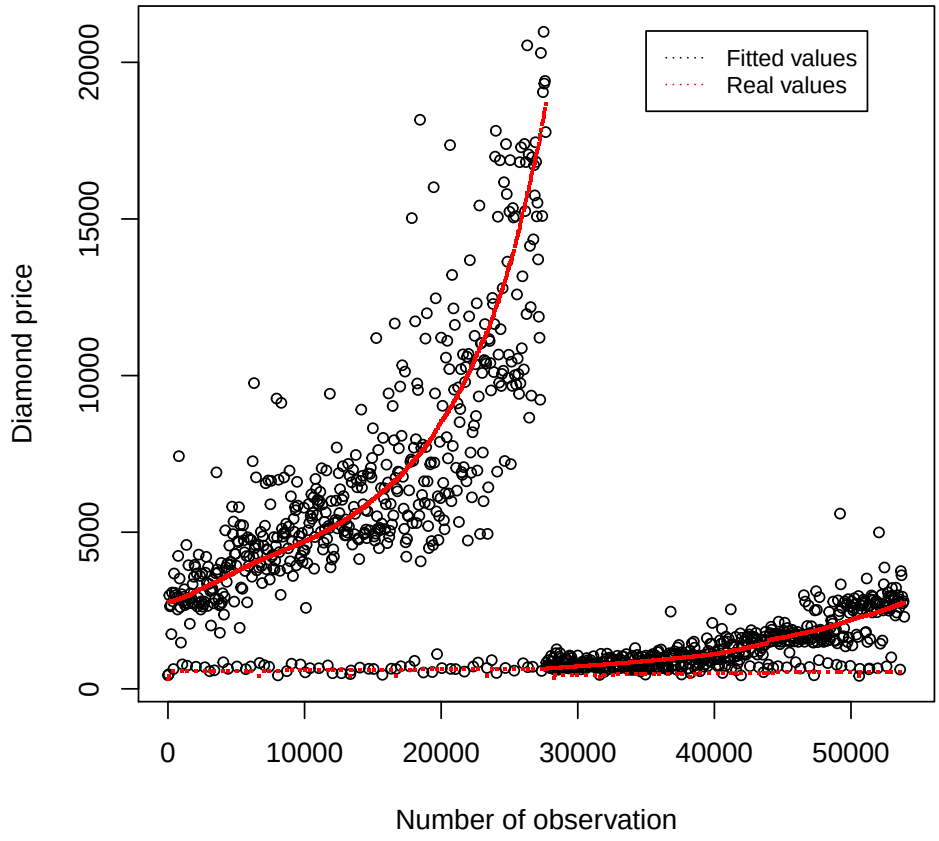

Figure 5: Result of GLM fit and real data in a model with continious and categorical variables

carat · clarity(D_i)	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
coefficient	6891.3	3657.7	2664.9	4883.4	4448.8	6154.9	5788.6

Table 4: Table with coefficients of carat · clarity depending on type of diamond's clarity.

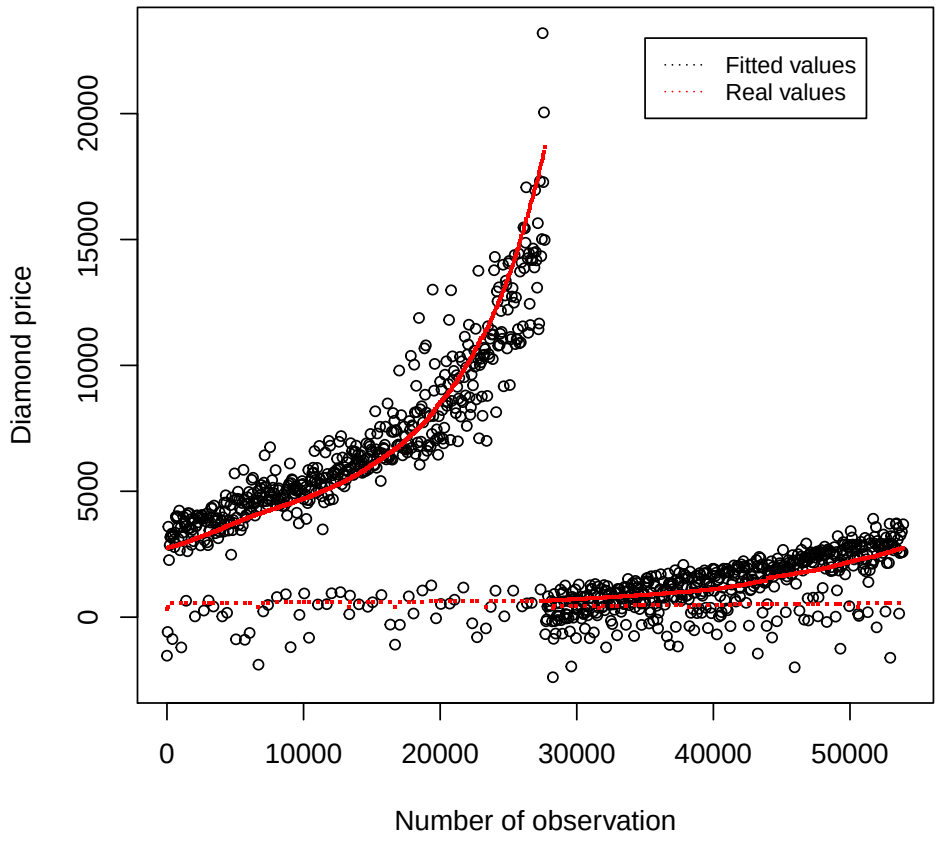

Figure 6: Result of GLM fit and real data in a model with continious and categorical variables

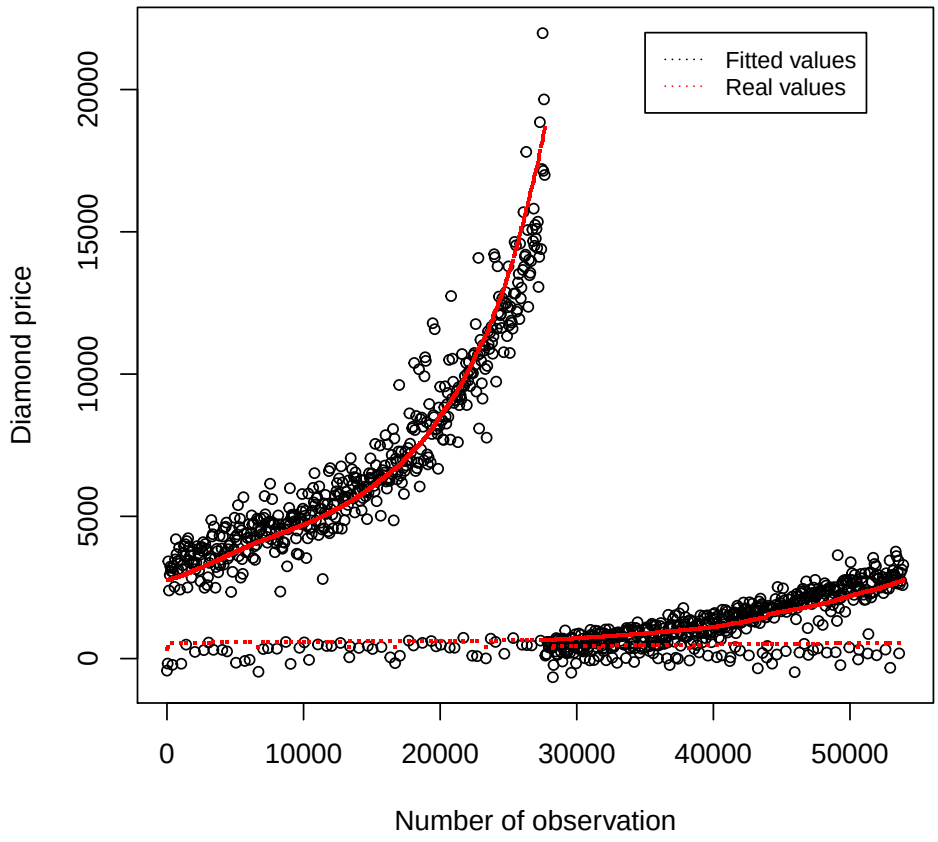

Figure 7: Result of GLM fit and real data in a model with continious and categorical variables

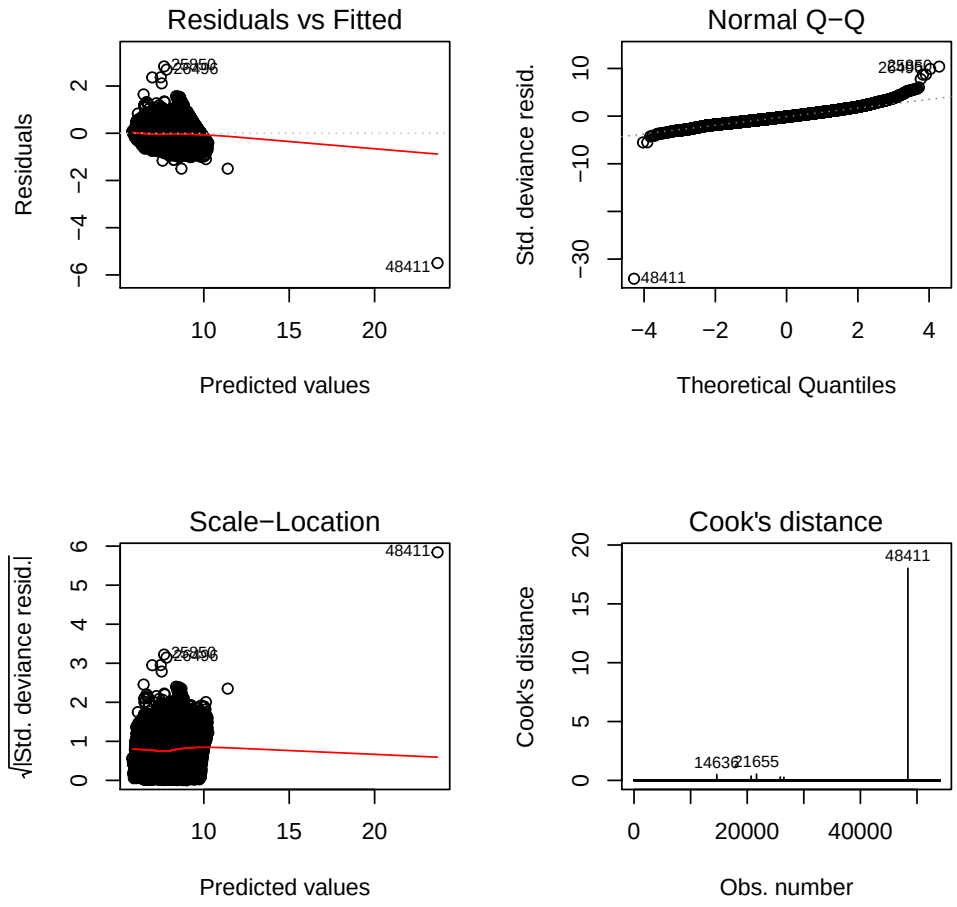

Figure 8: Charts presenting residual analysis for model only with continous variables

6 Check model assumptions

We check if fitted models fulfill appropriate assumptions. It turns out that each of them is far from having normal distributed residulas or being homogenous. On the plots below 8, 9 and 10 we can see that non of them has normally distributed residuals. Chart 'Residuals vs Fitted' present the distribution of residuals which should be placed around 0. In each of our model the assumption about homogeneity is not met.The normal Q-Q shows if your residuals are normally distributed. Residuals should go around the diagonal line but in our cases they are not. Cook's distance for residuals helps to detect outliers. They are visible in our cases and we try to handle them in next step of our analysis.

7 Outliers handling

To find outliers in our fits we use outlierTest function. It provides us a list of most significant outliers. We decide to delete outliers that are too far from rest of the data. This decision is made because of the fact that this observations are much different then others in only chosen characteristics and it could be the mistake.

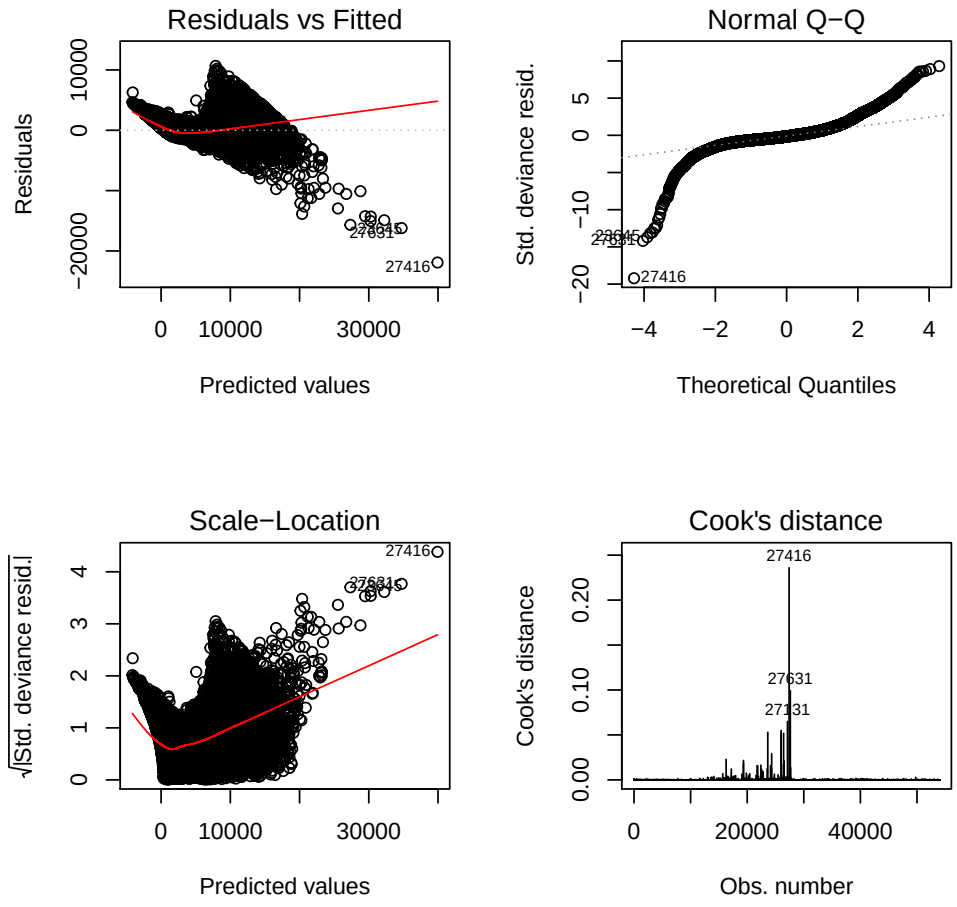

Figure 9: Charts presenting residual analysis for model with continous and categorical variables

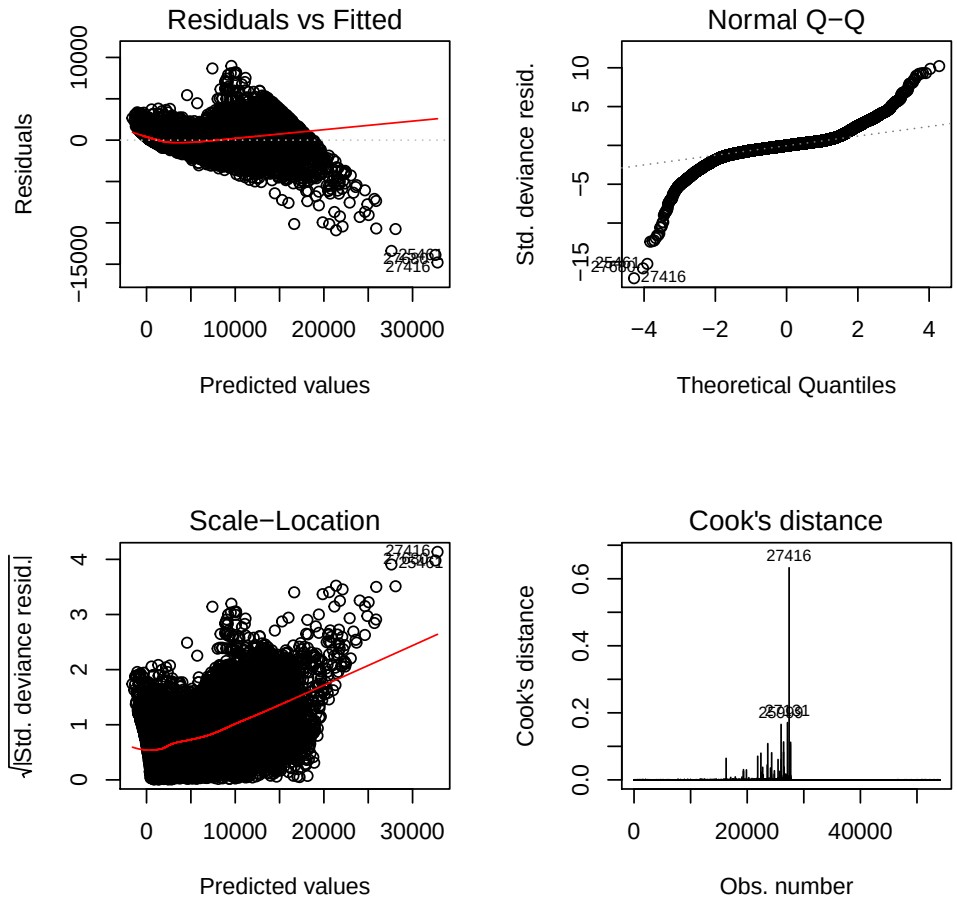

Figure 10: Charts presenting residual analysis for model with interactions

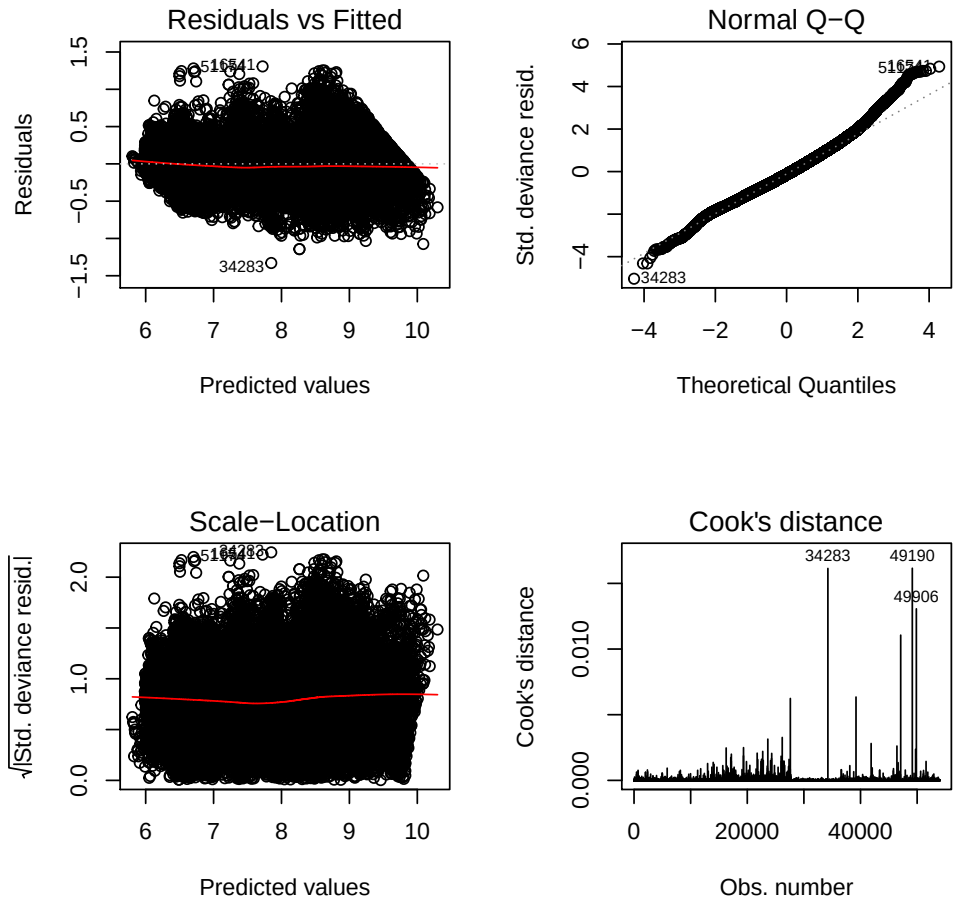

Figure 11: Charts presenting residual analysis for model with interactions

After that we tried to fit the best model again. On the charts 11 we could see that removing outliers make the residuals better.

8 Order models

It turns out that the best fit we manage to make is with the simplest model with just continuous variables. It becomes even more accurate when we delete outliers. The worst fit of those three models is obtained for the model with continuous and categorical variables, but without interactions. The most complex model with additional interactions gives fit somewhere between two previous ones.

9 Check significance of parameters

To check if all parameters in our models are significant we use function summary called at the best model that we find. Then we look at the significance codes and we can find out what parameters are significant. All summaries are shown below:

	number of variables	AIC	difference between the model and the best model
Model with only continuous variables without outliers	3	848300	0.0%
Model with only continuous variables with outliers	3	848800	-0.06%
Model with only continuous and categorical variables	4	912700	-7.6%
Model with interactions	4	884100	-4.2%

Table 5: Comparison of 4 models from analysis

```
> summary(model_simple_out)

Call:
glm(formula = price ~ z + x + carat, family = Gamma(link = "log"),
    data = diamonds_data_omited_simplout)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.32910  -0.19555  -0.03556   0.13746   1.30666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.000568   0.019825   50.47   <2e-16 ***
z            0.733609   0.013255   55.34   <2e-16 ***
x            0.894331   0.008363  106.94   <2e-16 ***
carat       -1.133497   0.011945  -94.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.07007919)

    Null deviance: 52787.9  on 53889  degrees of freedom
Residual deviance:  3576.4  on 53886  degrees of freedom
AIC: 846485

Number of Fisher Scoring iterations: 4

> summary(model_middle)

Call:
glm(formula = price ~ color + clarity + x + carat, family = gaussian(),
    data = diamonds_data_omited)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-21911.2    -592.4    -177.9     383.6   10661.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2826.45      93.97  -30.08   <2e-16 ***
```

```
colorE       -214.17      18.14  -11.80   <2e-16 ***
colorF       -282.59      18.35  -15.40   <2e-16 ***
colorG       -486.83      17.96  -27.11   <2e-16 ***
colorH       -993.16      19.09  -52.02   <2e-16 ***
colorI      -1477.46      21.46  -68.83   <2e-16 ***
colorJ      -2397.39      26.50  -90.46   <2e-16 ***
clarityIF    5707.09      51.01  111.89   <2e-16 ***
claritySI1   3922.27      43.75   89.66   <2e-16 ***
claritySI2   2952.57      44.01   67.09   <2e-16 ***
clarityVS1   4878.18      44.59  109.41   <2e-16 ***
clarityVS2   4559.32      43.90  103.86   <2e-16 ***
clarityVVS1  5339.55      47.12  113.31   <2e-16 ***
clarityVVS2  5266.32      45.85  114.87   <2e-16 ***
x           -1016.88      21.46  -47.38   <2e-16 ***
carat       11192.46      50.74  220.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1313862)

    Null deviance: 8.5723e+11  on 53919  degrees of freedom
Residual deviance: 7.0822e+10  on 53904  degrees of freedom
AIC: 912687

Number of Fisher Scoring iterations: 2
```

```
> summary(model_complex)

Call:
glm(formula = price ~ carat + carat:x + carat:color + carat:clarity,
    family = gaussian(), data = diamonds_data_omited)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14794.9    -401.1     -26.9     332.4    8974.6

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -783.901     17.108  -45.82   <2e-16 ***
carat             -6912.758     86.359  -80.05   <2e-16 ***
carat:x            1527.515      9.482  161.10   <2e-16 ***
carat:colorE       -219.073     18.503  -11.84   <2e-16 ***
carat:colorF       -342.142     17.934  -19.08   <2e-16 ***
carat:colorG       -737.484     17.290  -42.65   <2e-16 ***
carat:colorH      -1373.460     17.299  -79.40   <2e-16 ***
carat:colorI      -1920.826     18.141 -105.89   <2e-16 ***
carat:colorJ      -2786.799     19.940 -139.76   <2e-16 ***
carat:clarityIF    6891.256     42.282  162.98   <2e-16 ***
carat:claritySI1   3657.671     24.449  149.61   <2e-16 ***
carat:claritySI2   2664.851     24.041  110.84   <2e-16 ***
carat:clarityVS1   4883.444     25.999  187.83   <2e-16 ***
carat:clarityVS2   4448.760     24.833  179.15   <2e-16 ***
carat:clarityVVS1  6154.864     34.457  178.62   <2e-16 ***
```

```
carat:clarityVVS2  5788.636     29.482  196.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 773362.4)

    Null deviance: 8.5723e+11  on 53919  degrees of freedom
Residual deviance: 4.1687e+10  on 53904  degrees of freedom
AIC: 884111

Number of Fisher Scoring iterations: 2
```

It turns out that in all models that were fitted by us all parameters are significant. It means that we shouldn't get rid of them.

10 Conclusions

At the beginning of this analysis we set the question about how estimate the prices basing on given parameters. The second question was about the variables which has big influence on the price. It turns out that the most important feature is the catarat, because it appears in all our models. Also the color and clarity have relativly big influence on the price. Taking into consideration that facts we are thinking that obviously the carat (so the weight) is important but on the other hand the shape (table, depth) are not so important. Besides when we look at the features in our dataset it turns out that most of them have very litle interval. The fact that we did not found the model which meet the asumption of GLM provides to question if it would be a good idea to use it in this case. This analysis should led to other analysis in which other approach would be considered.