Diamond Price Prediction

Marco Cazzola

Matr: 964573

Abstract

In this paper, we are going to apply parametric and non-parametric learning algorithm to the `diamonds` data set (attached to the `ggplot2` package) to predict the diamonds' price according to their characteristics.

Data definition

The data set consists of 53940 rows and 10 columns, representing observations and variables respectively. More precisely, the available variables are:

- `price`: price of the diamond in US dollars. This will be the target variable of the analysis.
- `carat`: weight of the diamonds (1 carat = 0.2 grams)
- `cut`: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- `color`: diamond color, from D (best) to J (worst)
- `clarity`: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- `x`: length of the diamond, in mm
- `y`: width of the diamond, in mm
- `z`: depth of the diamond, in mm
- `depth`: depth of the diamond expressed relatively to the other two dimensions (`depth = z / mean(x, y)`)
- `table`: width of top of diamond relative to widest point.

Data summary

First of all, let's check if there is any NULL or NA value in the data set.

```
## [1] FALSE
```

Luckily, this is not the case. Let's then have a look at the data summary:

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
```

```
##  Max.   :5.0100                    I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

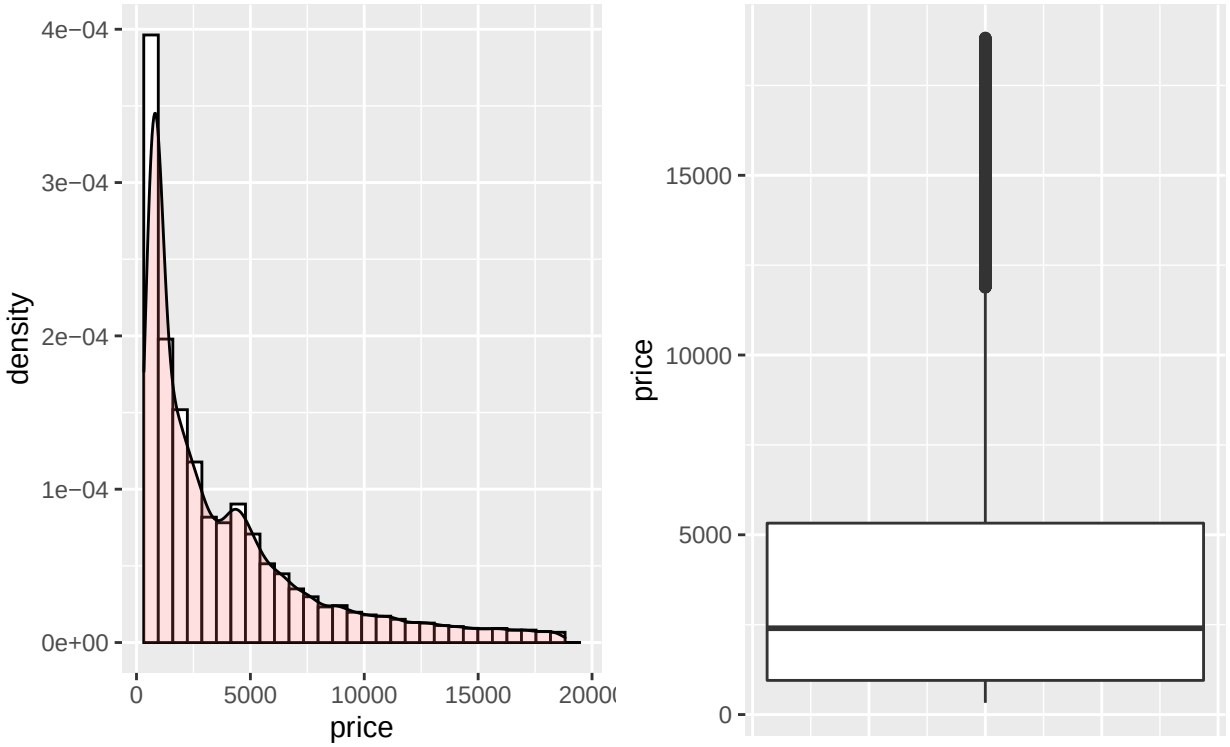
The price ranges between 326 USD and 18.823 USD; however, most of the observations are below the level of 5.000 USD. Also `carat` shows a similar pattern, with most of observations being below the level of 1 carat, while the maximum is 5.01. `depth` and `table` seem to be more fairly distributed, even though `table` shows a rather extreme maximum value: let's keep that in mind for the moment. Proceeding in the analysis, we notice some inconsistency in the dimension measurements: some diamonds have zero length, width and/or depth. Since the carat is not null for those diamonds, these zeros probably denote missing values. We therefore replace them with NA. Another thing we notice is that, while the maximum value for `x` is around 11, `y` and `z` show quite large maximum values. To investigate the possibility these are outliers, let's see how many observations have a value larger than 11 either for `y` or for `z`, and how many diamonds have a `table` value larger than 80.

```
## # A tibble: 4 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  2    Premium   H     SI2      58.9  57   12210  8.09 58.9   8.06
## 2  2.01 Fair      F     SI1      58.6  95   13387  8.32  8.31  4.87
## 3  0.51 Very Good E     VS1      61.8  54.7  1970  5.12  5.15 31.8
## 4  0.51 Ideal     E     VS1      61.8  55    2075  5.15 31.8   5.12
```

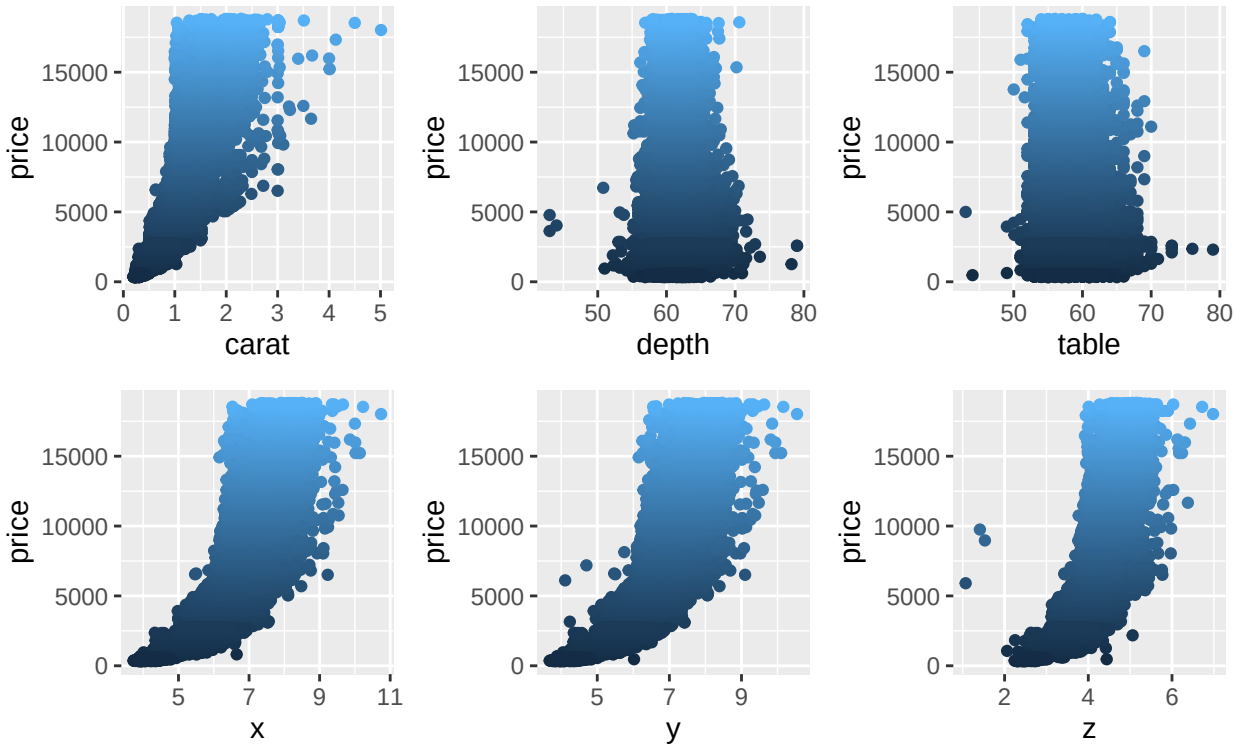
Over almost 54.000 observations, just 4 of them show such large levels of `y`, `z` or `table`; in other words, these are probably typos. Since those four examples represent a very negligible part of our data set, we drop them. Moving along with the categorical variables, most of the diamonds in the data set show an ideal cut, and also high-quality color diamonds are more frequent than low-quality color ones. However, diamonds with of best clarity class are very rare.

Descriptive statistics

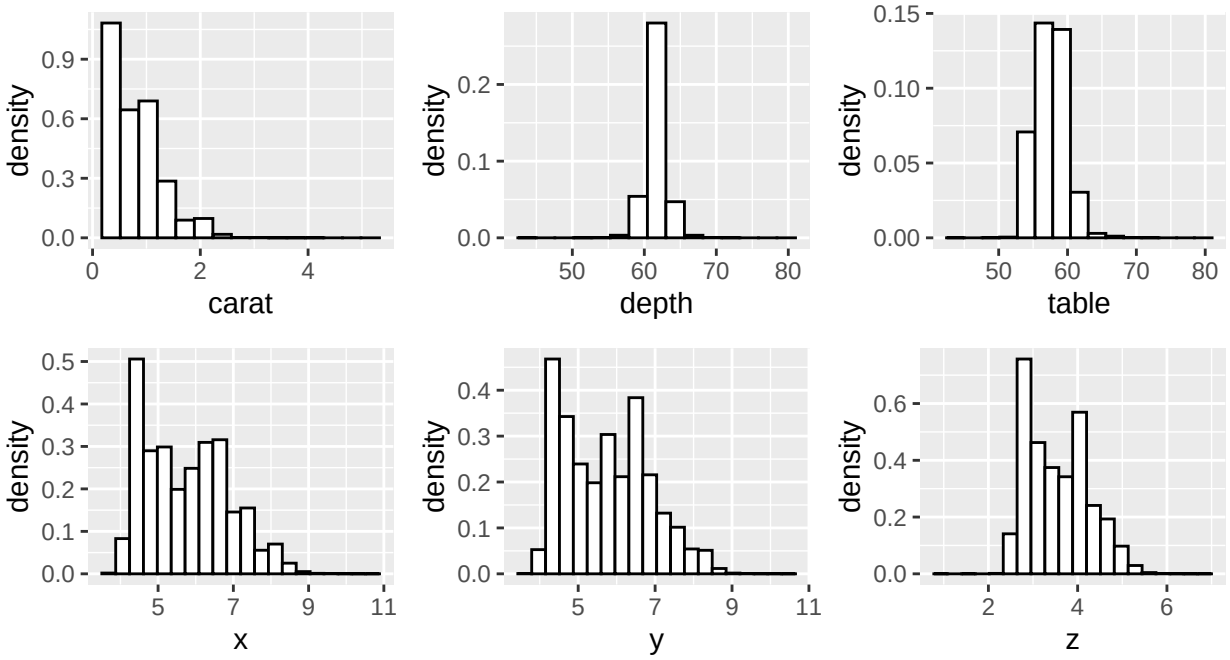
Target variable. Our target variable is definitely not normally distributed, and the boxplot shows a very large number of outliers. We decide to keep all the observations for the moment, since those high levels of prices could be justified by some premium characteristics of the diamonds.



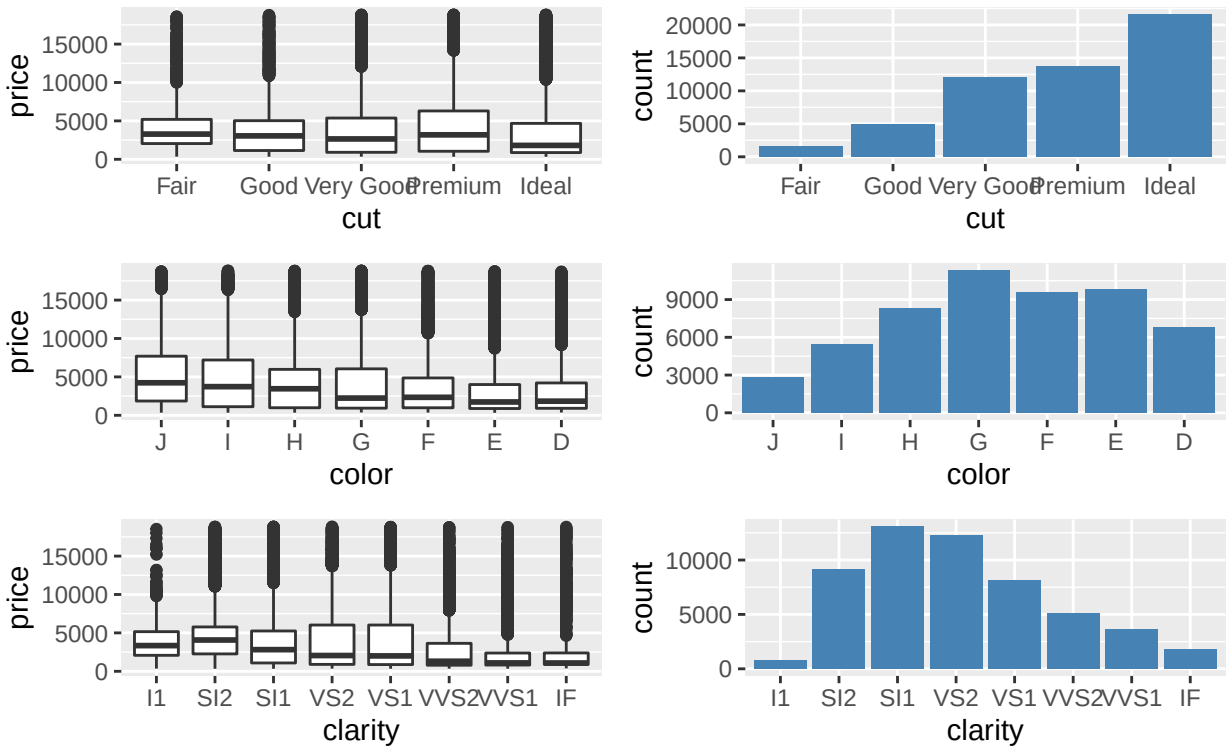
Quantitative regressors. The dimensionality variables (`x`, `y` and `z`) show a very similar pattern to `carat`, which measures weight: not surprisingly, weight and dimensions are probably correlated and they have both a strong and positive correlation with price. `depth` and `table` instead seem to have no relation with price.



For what concerns the distributions, notice how the distribution of `carat` is similar to the one of `price`. `depth` and `table` are quasi-normal (especially `depth`), while `x`, `y` and `z` are more uniformly distributed.

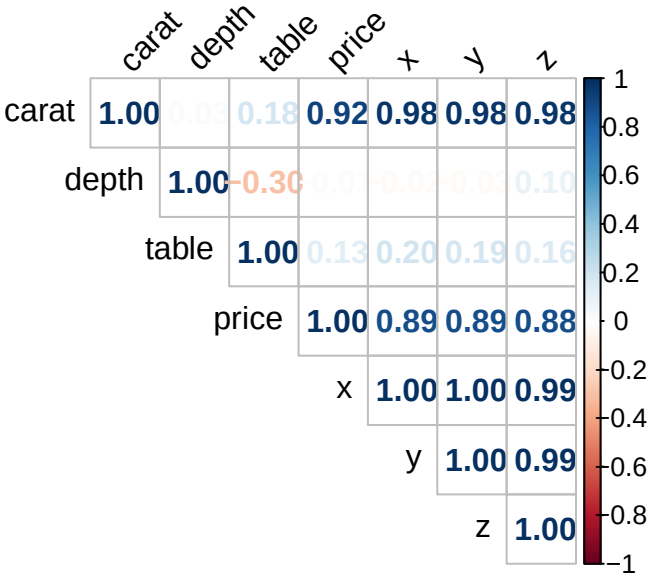


Qualitative regressors. None of the categorical predictor alone seems to be able to tell us something about the price variability: even outliers are fairly distributed among classes. For what concerns the distributions, most of the diamonds in the data set present an ideal cut; `clarity` is almost normally distributed, similarly to `color`.



Collinearity checks

If two regressors are correlated, including both of them adds no information but variability to the model, resulting in an overall worst performance. That is why it is important to check correlation among the features, displayed below through the correlation matrix.



As guessed in the previous analysis, `x`, `y`, `z` and `carat` are extremely correlated: they basically carry the same information. However, `carat` is the most strongly correlated with `price` among the four, and that is why we keep it and drop the others. Moreover, recall that there were some `NA` values in `x`, `y` and `z`, while `carat`'s column is complete. The other regressors do not show any correlation, but `depth` and `table` are very poorly correlated with our target, as shown in the graphical analysis.

In the following, we will use Goodman and Kruskal tau index to measure the strength of association between our categorical variables, which cannot be computed through 'standard' correlation. Reassuringly, the plot tells us there is no association between the categorical variables, which means each of them carries independent information from the others and it is therefore worth including in the model.