# CONTENTS

# 1 DATA

# 2 CORRELATION COEFFICIENT

To understand which pair of features is the best for constructing a linear regression on, we have to take into account their distributions. As shown on fig. 1, there are several features of interest: ”x”, ”y”, ”z”, ”carat” and ”price”. Features ”x”, ”y” and ”z”, with the exception of some outliers, are distributed in a linear pattern and their distributions wrt ”carat” look like polynomial functions. This is an expected behavior, since ”carat” is roughly the product of ”x”, ”y”, ”z”, which are linearly distributed with each other, and diamond density.

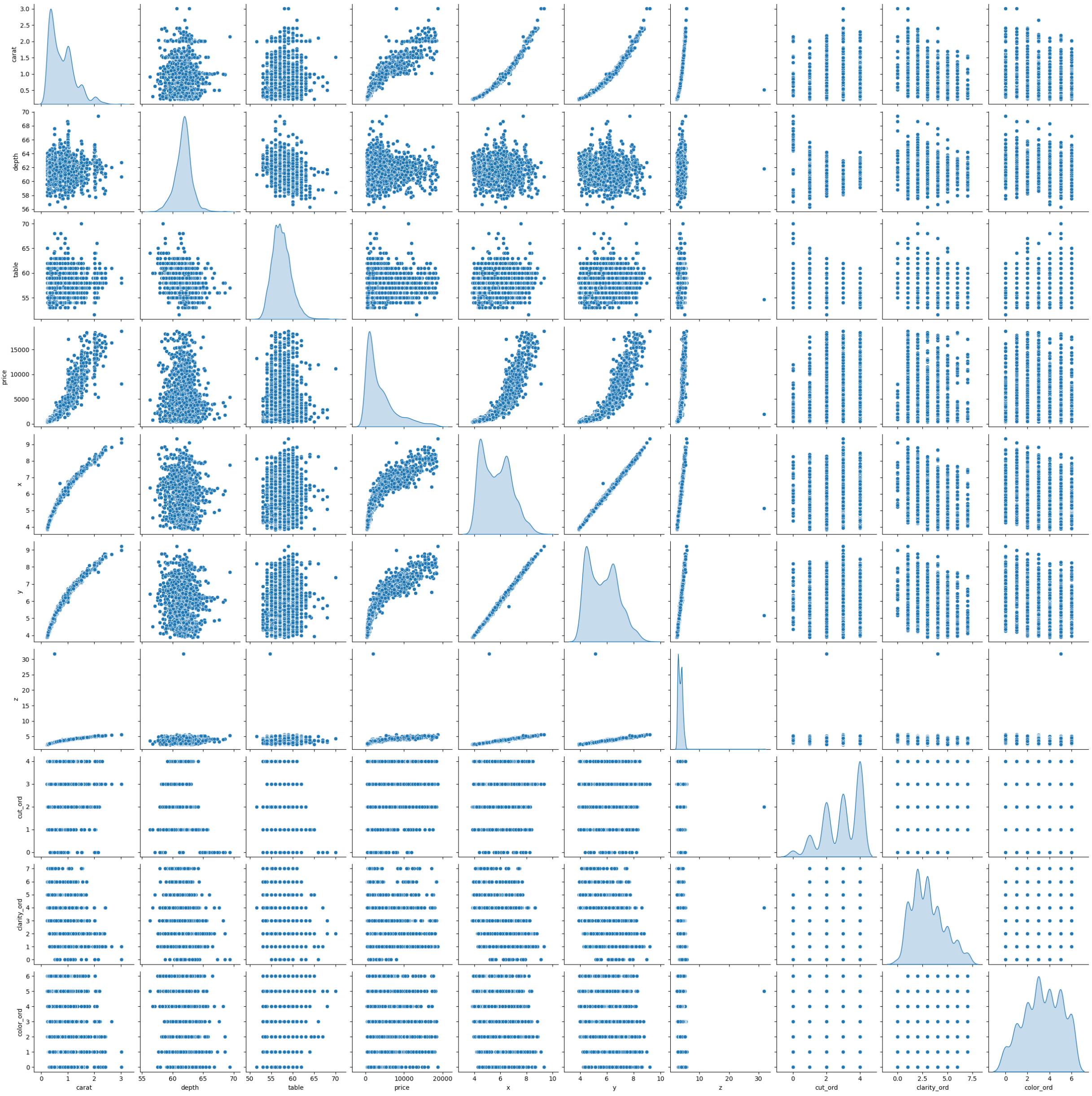

**Figure 1.** Pairwise scatter plots of all the features in data

In this task we will consider the pair "carat" and "price", because their distribution is not as trivial as those of diamonds' physical dimensions wrt each other or their mass (fig. 2).

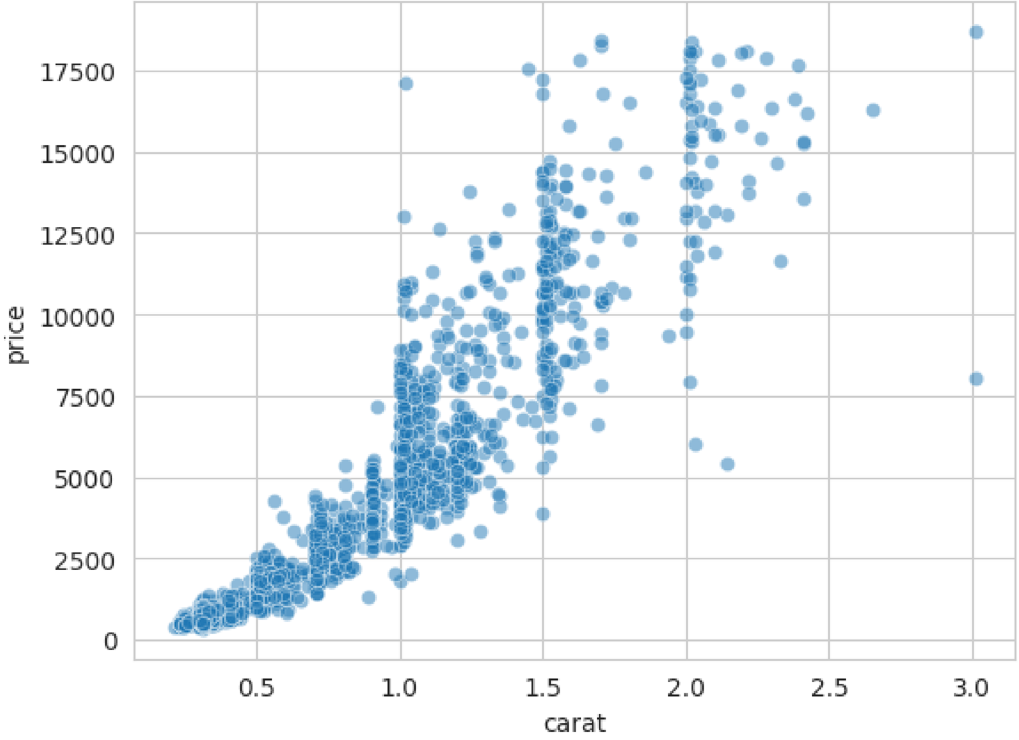

**Figure 2.** Distribution "carat" and "price" wrt each other

First, a regular linear regression of "price" over "carat" was constructed. As shown on fig. 3, this line has a positive slope, hence the diamond's price tends to positively correlate with its mass, which is to be expected. Next, the correlation and $R^2$ coefficient, also known as *the coefficient of determination*, were computed:

$$\rho = 0.9240,\ R^2 = 0.8537.$$

Since $R^2$ coefficient is the proportion of variance of the target feature taken into account by linear regression, in our case 85.37% of "price" variance is taken into account.

After that, a set

$$x = \begin{pmatrix} 0.3000 & 863 \\ 0.3100 & 788 \\ 0.4000 & 662 \\ 2.0100 & 17078 \end{pmatrix}$$

of four random pairs of "carat" and "price" values was chosen to compare the predicted and true "price" values. To do this, the percentile deviations were found:

$$\Delta y_{\text{true}} = (89.9245\%, 79.1276\%, -30.2385\%, 21.8683\%)^T ,$$
$$\Delta y_{\text{pred}} = (892.5046\%, 379.1018\%, -23.2178\%, 27.9890\%)^T ,$$

where $\Delta y_{\text{true}}$ is relative deviation of predicted from true values and $\Delta y_{\text{pred}}$ is the other way round. From this we can conclude that the regression line tends to underestimate the price of diamonds with lower mass.

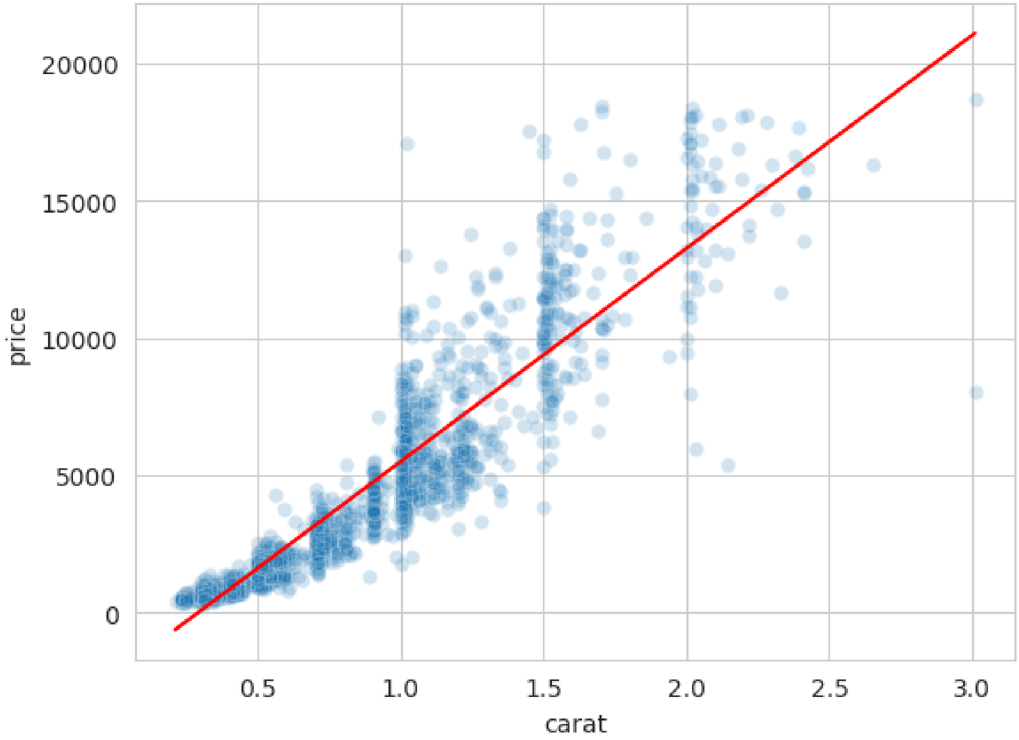

**Figure 3.** Linear regression of "price" over "carat"

Finally, mean absolute percentile error(MAPE) relative to both true and predicted over the whole data was considered in order to measure the mean deviation of the data from the predictor line and vice versa:

$$\text{MAPE}_{\text{true}} = 39.2063\%, \ \text{MAPE}_{\text{pred}} = 121.4897\%.$$

These values indicate that, similarly to the previous case, regression on this pair features tends to underestimate the "price" feature. There is also another interpretation: $\text{MAPE}_{\text{true}}$ is data analysis(DA) view on MAPE, while $\text{MAPE}_{\text{pred}}$ is machine learning(ML) view on MAPE; hence, while in terms of DA the line we constructed is good, while the opposite is true in terms of ML.

We also tried polynomial regression with a third degree polynomial, as seen on fig. 4. This resulting in following MAPE values:

$$\text{MAPE}_{\text{true}} = 39.2063\%, \ \text{MAPE}_{\text{pred}} = 37.0984\%.$$

This indicates that the polynomial curve uses more data and does not underestimate the price.

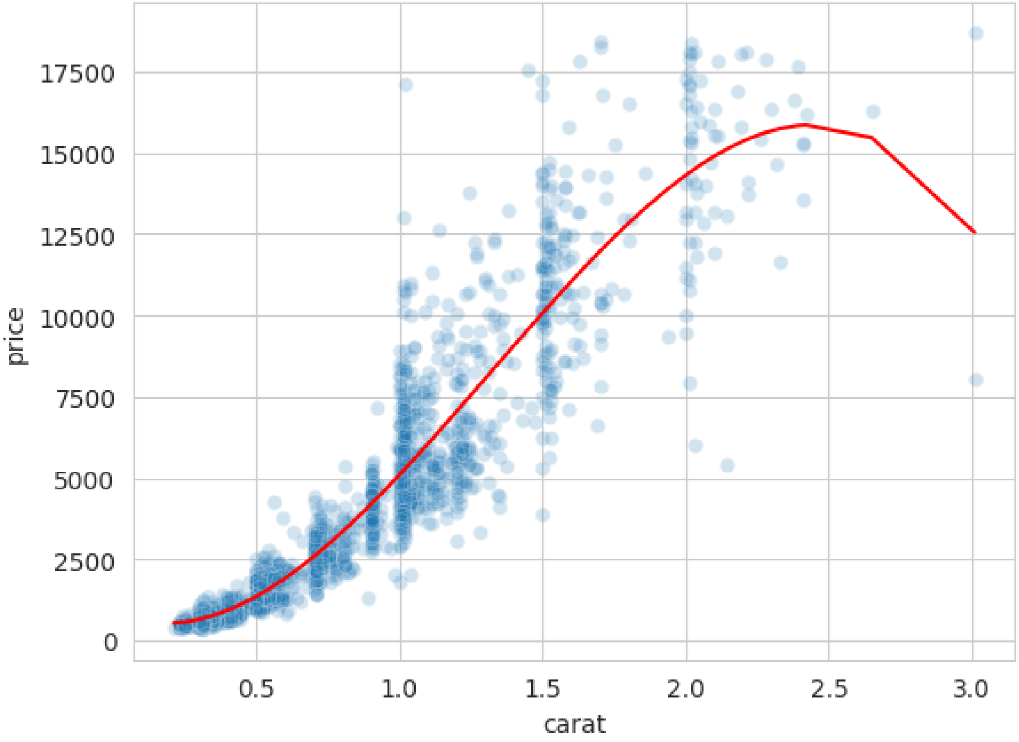

**Figure 4.** Polynomial regression of "price" over "carat"

# 3 CLUSTER ANALYSIS AND CLUSTER INTERPRETATION

To understand the general pricing patterns a following subset of features was explored via K-Means algorithm:

- "depth";
- "table";
- "carat";
- "price".

These features denote the physical parameters of a diamond and were scaled to have normal distribution before the clustering algorithm was applied to them.

Then, using the build-in K-Means algorithm with 4 and 7 clusters were applied to the data. It was initialized to start with twelve sets of random centers for each number of clusters used. Algorithm's performance was measured with inertia metric:

$$L = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} (y_{iv} - c_{kv})^2 ,$$

where $K$ is the number of clusters;

$S_k$ is the number of objects in $k$-th cluster;

$V$ is the set of features;

$y_{iv}$ is value of feature $v$ of the $i$-th object from cluster $k$;

$c_{kv}$ is value of feature $v$ of the center of cluster $k$.

After the evaluation, for each number of clusters the best-performing cluster set was chosen for further analysis. On those the absolute and relative deviations from grand mean for each feature were calculated:

$$d_{abs} = (c_{kv} - c_v),\ d_{rel} = \frac{c_{kv} - c_v}{c_v} \cdot 100\%.$$

For each cluster in each table deviations in a feature of more than 30% were marked with green(positive deviations) and red(negative deviation) backgrounds.

**Table 1.** Best set for K-Means with 4 clusters

| | **depth** | **table** | **carat** | **price** |
|---|---|---|---|---|
| grand mean | 61.7409 | 57.3868 | 0.7865 | 3858.1980 |
| **cluster 1 (234 instances)** | | | | |
| center | 61.4547 | 58.0838 | 1.6865 | 12463.5812 |
| grand mean devation | -0.2861 | 0.6970 | 0.9000 | 8605.3832 |
| rel. grand mean deviation | -0.46% | 1.21% | 114.44% | 223.04% |
| **cluster 2 (854 instances)** | | | | |
| center | 62.0034 | 56.1681 | 0.4438 | 1276.4660 |
| grand mean devation | 0.2625 | -1.2187 | -0.3427 | -2581.7320 |
| rel. grand mean deviation | 0.43% | -2.12% | -43.57% | -66.92% |
| **cluster 3 (512 instances)** | | | | |
| center | 62.5969 | 57.3387 | 1.0456 | 5258.8340 |
| grand mean devation | 0.8560 | -0.0481 | 0.2592 | 1400.6360 |
| rel. grand mean deviation | 1.39% | -0.08% | 32.95% | 36.30% |
| **cluster 4 (400 instances)** | | | | |
| center | 60.2520 | 59.6425 | 0.6599 | 2543.2325 |
| grand mean devation | -1.4889 | 2.2557 | -0.1266 | -1314.9655 |
| rel. grand mean deviation | -2.41% | 3.93% | -16.10% | -34.08% |

According to the table 1, those four clusters correspond to extremely valuable, extremely cheap, moderately valuable and moderately cheap diamonds respectively. Interestingly, ”depth” and ”table” do not deviate that much from their respective grand means and do not correlate with neither ”price” nor ”carat”.

Similar conclusion can be made for the case with 7 clusters. They correspond to various ”price” and ”mass” ranges with clusters in similar range being divided based on their ”depth” and ”table” deviations.

**Table 2.** Best set for K-Means with 7 clusters

| | **depth** | **table** | **carat** | **price** |
|---|---|---|---|---|
| grand mean | 61.7409 | 57.3868 | 0.7865 | 3858.1980 |
| **cluster 1 (193 instances)** | | | | |
| center | 61.5523 | 58.2466 | 1.7496 | 13071.0829 |
| grand mean devation | -0.1885 | 0.8598 | 0.9631 | 9212.8849 |
| rel. grand mean deviation | -0.31% | 1.50% | 122.46% | 238.79% |
| **cluster 2 (290 instances)** | | | | |
| center | 60.5655 | 57.3117 | 0.4779 | 1450.8103 |
| grand mean devation | -1.1753 | -0.0751 | -0.3085 | -2407.3877 |
| rel. grand mean deviation | -1.90% | -0.13% | -39.23% | -62.40% |
| **cluster 3 (350 instances)** | | | | |
| center | 62.3123 | 58.0829 | 0.4388 | 1231.0829 |
| grand mean devation | 0.5714 | 0.6961 | -0.3476 | -2627.1151 |
| rel. grand mean deviation | 0.93% | 1.21% | -44.20% | -68.09% |
| **cluster 4 (192 instances)** | | | | |
| center | 59.7719 | 60.8776 | 0.8353 | 3704.7188 |
| grand mean devation | -1.9690 | 3.4908 | 0.0488 | -153.4792 |
| rel. grand mean deviation | -3.19% | 6.08% | 6.21% | -3.98% |
| **cluster 5 (184 instances)** | | | | |
| center | 63.7891 | 58.3761 | 1.0024 | 4422.3478 |
| grand mean devation | 2.0483 | 0.9893 | 0.2159 | 564.1498 |
| rel. grand mean deviation | 3.32% | 1.72% | 27.45% | 14.62% |
| **cluster 6 (431 instances)** | | | | |
| center | 62.1624 | 55.0935 | 0.4575 | 1338.5963 |
| grand mean devation | 0.4216 | -2.2933 | -0.3290 | -2519.6017 |
| rel. grand mean deviation | 0.68% | -4.00% | -41.83% | -65.31% |

**Table 2.** Best set for K-Means with 7 clusters

| | **depth** | **table** | **carat** | **price** |
|---|---|---|---|---|
| **cluster 7 (360 instances)** | | | | |
| center | 61.7317 | 56.6878 | 1.1141 | 6222.5278 |
| grand mean devation | -0.0092 | -0.6990 | 0.3277 | 2364.3298 |
| rel. grand mean deviation | -0.01% | -1.22% | 41.66% | 61.28% |

Overall, "depth" and "table" seem to have no correlation to both the diamond's price and its mass. Increasing the number of classes seems to subdivide the price ranges into smaller pieces determined by their "table" and "depth".