ML: Assignment 1 Report

Team

- Kevin Pirkelbauer
- Michael Breiteneder
- Stefan Schwaighofer
- Lorenz Graf

Tasks

Part 1: lecture script

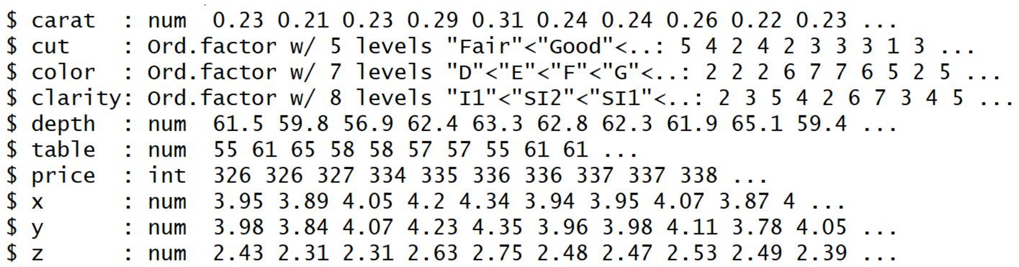
This task was performed by the team members individually and there is no hand in required.

Part 2: Diamond Prices

1. Give an overview of the dataset structure:

• How many samples and features are in the dataset?
The dimensions of the diamonds dataset correspond to the amount of samples and features. By using the dim() command we get the information that there are 53940 samples and 10 features.

• What are the feature data types
We can get information about the data types using the str() or structure command. While most of the features are numeric or integers the cut, color and clarity are labels or factors. The resulting types can be seen here:



```
$ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
$ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
$ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
$ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
$ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
$ table  : num  55 61 65 58 58 57 57 55 61 61 ...
$ price  : int  326 326 327 334 335 336 336 337 337 338 ...
$ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
$ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
$ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

• Are diamonds balanced across color, cut and clarity?
Color: Fairly balanced, even though I and J are a bit fewer.
Cut: Not balanced. Most “ideal” with very few “fair”.
Clarity: Not balanced. Very few I1 and IF and for example more SI1 and VS2.

2. Visualize diamond prices using a histogram, boxplot and densityplot:

• Is there a visible trend? If yes, which is it and from which plots can you derive it?
There are more cheap diamonds than expensive ones (from hist and density plot). The lowest priced diamond is at 326, 25% of diamonds have a price below 950 (Q1), 50% below 2401 (Median), 75% below 5324 (Q3). While most diamonds cost below 12000 some outliers cost up to 18823 (from boxplot in combination with summary for exact numbers).

3. Calculate and state the mean, median, standard deviation, median absolute deviation, 1st and 3rd quantile, and inner quantile range of the diamond price:

Mean: 3933
Median: 2401
Standard Deviation: 3989.44
Median Absolute Deviation: 2475.942
1st Quantlie: 950
3rd Quantile: 5321.25
Inner Quantile Range: 4374.25

4. Plot the diamond price against the carat values as an xy-plot:

• Is there a trend visible in the plot? If yes, which is it?
The price of a diamond increases with its carat value.

5. Analyze the correlation between diamond price and diamond x, y, and z dimensions:

• Is there a trend visible between x, y, and z? If yes, which is it? Is there a trend visible between the dimensions and the price? If yes, which is it?
The plot shows logarithmic corellation between price and features x, y, z (bigger dimensions equal a higher price). The plot alsow shows linear corellation between the features x and y, x and z, as well as y and z.

6. Analyze diamond prices per diamond color:

• Create boxplots showing diamond price boxes for each diamond color. Use a single plot command to obtain this plot (hint: use the formula interface of boxplot with a “price ~ color” formula) Create densityplots showing diamond prices for each diamond color. Is there a trend visible? If yes, which is it?
Both plots shows that most of the diamonds in each color category have a price range between 0 and <= 5000.

7. Use vectorized commands to answer these questions:

• How many diamonds have a price above 9500?
5734

• How many diamonds have a price above 9500 and have color “D”?
461

- **For all color “D” diamonds with cut “Fair” show the summary for price and carat**

```
     carat
 Min.   :0.2500
 1st Qu.:0.7000
 Median :0.9000
 Mean   :0.9201
 3rd Qu.:1.0100
 Max.   :3.4000

     price
 Min.   :  536
 1st Qu.: 2204
 Median : 3730
 Mean   : 4291
 3rd Qu.: 4797
 Max.   :16386
```

- **For all color “J” diamonds with cut “Ideal” show the summary for price and carat**

```
     carat
 Min.   :0.230
 1st Qu.:0.540
 Median :1.030
 Mean   :1.064
 3rd Qu.:1.410
 Max.   :3.010

     price
 Min.   :  340
 1st Qu.: 1132
 Median : 4096
 Mean   : 4918
 3rd Qu.: 6732
 Max.   :18508
```

8. Utilize functions from the plyr library or the standard apply function on the diamonds data.frame to compute the following:

1. Calculate the mean (?mean) for all numerics features using a single command.

```
sapply(diamonds, function(x) if (!is.numeric(x)) NA else mean(x))
```

2. Do the same with median (?median).

```
sapply(diamonds, function(x) if (!is.numeric(x)) NA else median(x))
```

3. Do the same with standard deviation (?sd).

```
sapply(diamonds, function(x) if (!is.numeric(x)) NA else sd(x))
```

4. Do the same with median absolute deviation (?mad) and set the constant parameter of the mad function to 1.

```
sapply(diamonds, function(x) if (!is.numeric(x)) NA else mad(x, constant = 1))
```

Part 3: Cell Body Segmentation Data

1. Which classes exist? Are they (roughly) balanced?

PS and WS.
They are not balanced (1300:719).

2. Which noteworthy trends of features and relations between features as well as features and Class do you see?

AwgIntenCh1 is linearly related to DiffIntenDensityCh1. Between all other attributes there were no clear relations visible.

3. If you would need to distinguish between classes, which features do you think would be most helpful? Why?

There is no real good attribute to distinguish the classes. One of the more fitting attributes is AvgIntenCh2. Further attributes that may help distinguish classes are AvgIntenCh1 and ConvexHullPerimRatioCh1. These attributes show fewer overlaps in the density-, box- and hist-plot.

However these are still not an optimal indicator. We would recommend collecting other features.