# MCM ML 520 Assignment 1

## Assignment 1

### R Setup

- Install R and RStudio
- Install the packages plyr, lattice, ggplot2 and caret. In the R terminal:

```
install.packages('plyr')
install.packages('lattice')
install.packages('ggplot2')
install.packages('caret')
```

### Hand in

- Hand in the R scripts used to compute the answers/figures asked in for each tasks and the report (details in first slide set) in which you document your approach, results, and findings.

### Part 1: lecture script

Go through the `01_intro-to.r` script: execute it line by line and understand what each line achieves. You need this understanding to solve the following tasks. No hand in is required for this part.

### Part 2: Diamond Prices

#### Dataset

Use the diamonds dataset delivered with the ggplot2 library:

```
library(ggplot2)
data(diamonds)
```

**Task**

Make yourself familiar with the dataset using `?diamonds`. Prepare an R script that fulfills the following requirements and answer the included questions (put results into the report that support your answers, like command outputs, figures, tables, etc):

1. Give an overview of the dataset structure:
   - How many samples and features are in the dataset?
   - What are the feature data types?
   - Are diamonds balanced across color, cut and clarity? (Hint: roughly 1:1 means balanced, e.g. 1:2 is a "1:2 imbalance")
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?
3. Calculate and state the mean, median, standard deviation, median absolute deviation (MAD), 1st and 3rd quantile, and inner quantile range of the diamond price:
   - If you are not familiar with those functions, try Google, Wikipedia, ...
   - Hint: `summary()` will be useful.
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?
   - Hint: plotting many samples will be slow. Changing the plot symbol to `'.'` will cause a speedup.
5. Analyze the correlation between diamond price and diamond x, y, and z dimensions:
   - Create pairwise plots for these features.
   - 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?
   - Hint: remember what a linear correlation between 2 variables (=features) looks like:
     - Linear correlation: feature A low –> feature B low, and feature A high –> feature B high.
     - (Inverse) linear correlation: feature A low –> feature B high, and feature A high –> feature B low: inverse linear correlation. Usually also just called linear correlation.
     - When plotting feature A against feature B, both will cause a "straight line" (+ some noise = scatter). This means: if you see straight lines, there is a linear relation between features.
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 (hint: `featurePlot`).
   - Is there a trend visible? If yes, which is it?
7. Use vectorized commands to answer these questions:
   - How many diamonds have a price above 9500?
   - How many diamonds have a price above 9500 and have color "D"?
   - For all color "D" diamonds with cut "Fair" show the summary for price and carat
   - For all color "J" diamonds with cut "Ideal" show the summary for price and carat
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.
2. Do the same with median (`?median`).
3. Do the same with standard deviation (`?sd`).
4. Do the same with median absolute deviation (`?mad`) and set the `constant` parameter of the `mad` function to `1`.
5. Hint: usually, you would compute the first 2 using the functions `colMeans` and `summary`, but for the purpose of applying arbitrary function on data.frames, don't use these functions for this task.

**Hints**

- The Udacity Data Analysis with R course also focuses on this dataset for introduction. You can watch their short series of videos on how they analyze this dataset to dive into more details.

## Part 3: Cell Body Segmentation Data

**Dataset**

Use the segmentationData dataset provided as csv-file. The class we differentiate is stated by the "`Class`" feature.

**Task**

Analyze the data using the same techniques as for the last task. Decide for yourself which and how to use the specific commands. Answer the following questions in the report and include figures supporting your answers:

1. Which classes exist? Are they (roughly) balanced?
2. Which noteworthy trends of features and relations between features as well as features and Class do you see?
3. If you would need to distinguish between classes, which features do you think would be most helpful? Why?

**Hints**

- Ensure the "`Class`" feature is a factor after loading data/before passing it as parameter to functions that require factors.
- Separate classes in plots/use colors to distinguish between them.
- Pairwise feature plots and `caret::featurePlot` might come in handy again.