# Exercise Sheet 3 - Data Visualization

Timo Meiendresch

WS 20/21

- For reference, you may check the `ggplot2` cheatsheet by clicking on "Help" > "Cheat sheets" > "Data Visualization with ggplot2" in the RStudio IDE.

Before starting: Recall the basic structure of the first three layers, namely

- `data` - Dataset you want to plot
- `aes`thetic mapping - How to do the mapping of variables
- `geom_` - Indicates the geometric object used for the mapping.

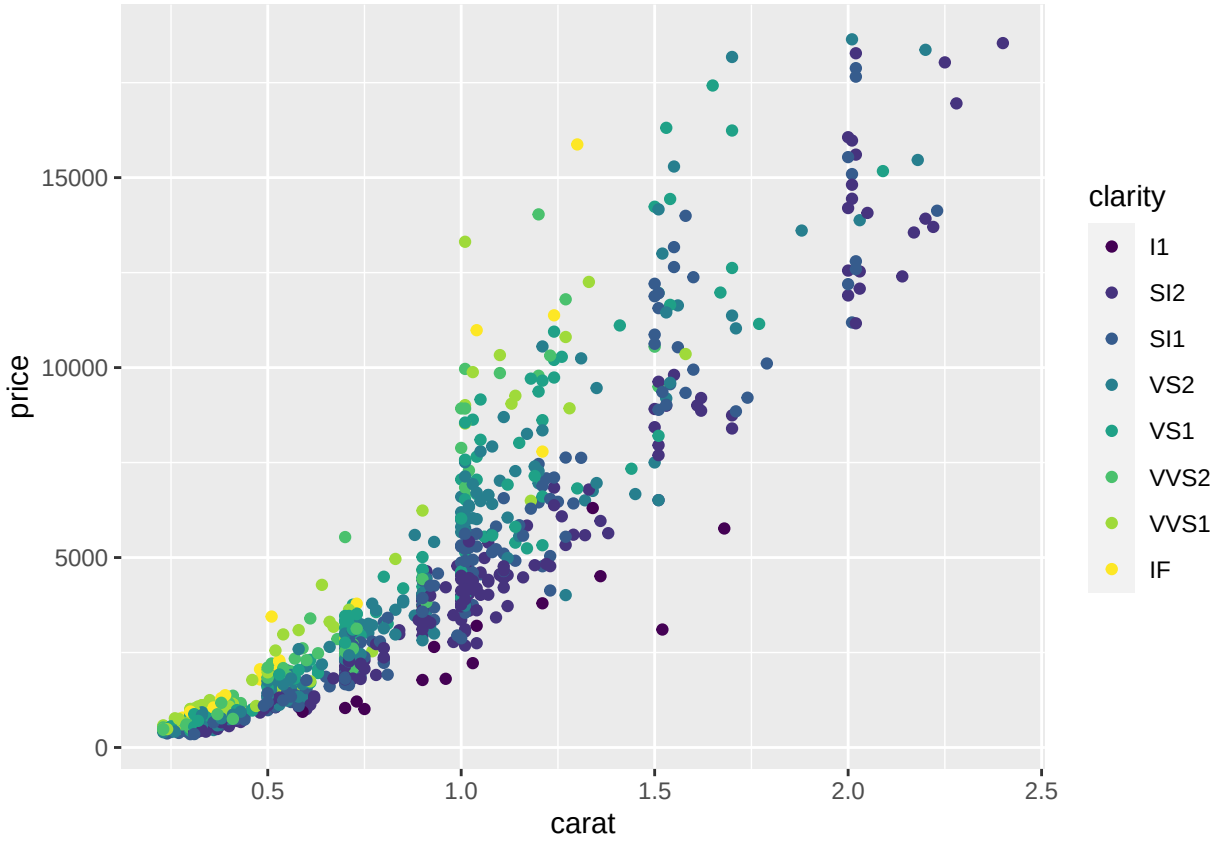
The basic template for thousands of different data visualizations in `ggplot2` looks like this:

```
# template
ggplot(data = <MYDATA>, mapping = aes(<MAPPINGS>)) +
  <GEOM_FUNCTION>()
```

## 1 Exercise: Data Visualization

### 1.1 `diamonds` - data

Look at the following plot that uses the `diamonds` dataset, which is part of the `ggplot2` library:




- Briefly describe its elements:
  - dataset
  - mapping
  - geometric object
- What is the relation between `carat` and `price` of diamonds? Functional form?

## 1.2 `mpg`-data

- Check out the dataset `mpg` which is part of the `ggplot2` library
  - What's in the dataset: Number of observations? Variables/features?
- What relationship do you expect between the variables?
  - Focus on the relationship between the variables `displ` and `hwy`.
  - `displ` stands for engine displacement, in litres and is a measure of an engine's size and an indicator of how powerful a car is.
  - `hwy`, i.e. how many miles (1.6 kilometers) a car can drive per gallon (3.8 litres)
- Check your intuition by calculating the correlation (`cor()`) between these two variables.

### 1.2.1 Scatterplot/Point plot (`geom_point()`)

- What are scatterplots used for ...?
  - Hint: You may check the "Description" (`?geom_point`) to answer this question.

Next, a simple scatterplot:

- use `data = mpg`
- `aes`thetic mapping, where you map the variable `displ`on the x-axis and the variable `hwy` on the y-axis.

  - Note that the mapping can be placed either within the `ggplot()` call or in the geometric object `geom_point()`.

- Use `geom_point()` as geometric object.
- Is the relationship as expected?
  - The correlation coefficient is around `-0.77`. Briefly explain what this means?

**Additional mappings**

- Check your visualization for possible outliers, i.e. points that seem to escape the overall relationship.
  - Hint: Check points around the area ($x = 6, y = 25$)
- To get a clearer picture, we'll add another mapping. Can you add the mapping of the variable `class` to color?
  - Hint: `mapping = aes(x = displ, y = hwy, color = ...)`. You may have to check the number of open/closed parentheses (at least that's one of my favorite mistakes)

Note that the legend is added automatically

**Experiment with geometric object**

Within the `geom_point()`-call you can also change the characteristics of the geometric object, i.e. how the points of the plot are displayed. Experiment with `alpha`, `size`, `color`, `shape`. + For example: Change the color to blue, change the size of the dots, etc. + Hint: `geom_point(alpha=0.42, ...)`.

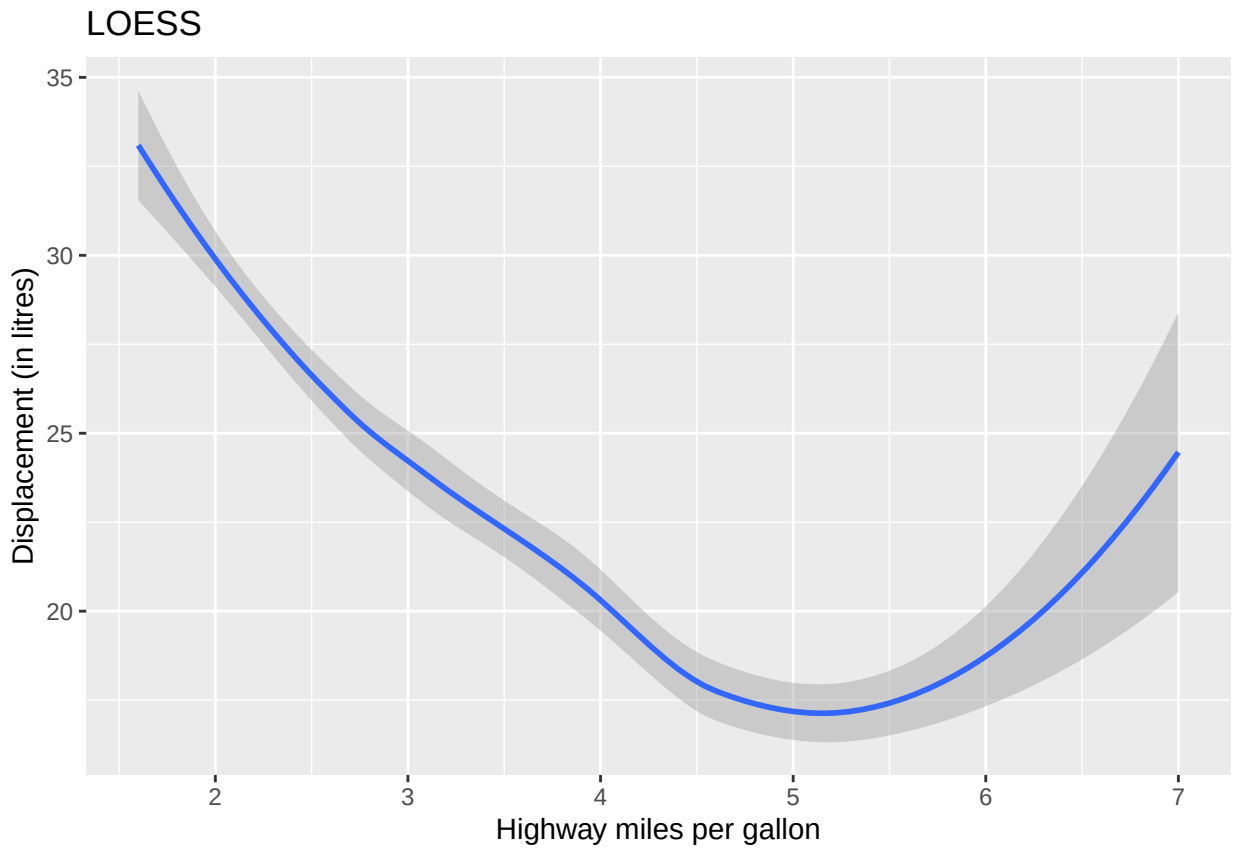
**Adding a `geom_smooth()` layer**

- Quantify the relationship between fuel consumption (`hwy`) and engine power (`displ`) using a simple bivariate linear model
  - Hint: Remember the linear model function `lm()`?
- Can you visualize this relationship in addition to a visualization of the model coefficients?
  - Start from the basic ggplot2 template
  - data: `mpg`
  - mapping: `x=displ`, `y=hwy`
  - geometric object 1: Scatterplot
  - geometric object 2: `geom_smooth()` - Check the function description, specify `method="lm"`
- Briefly explain the relationship between coefficients of your linear model call to the visual representation of the same model.

The following exercise may sound advanced but is actually pretty simple once you understand the task

- LOESS, which stands for locally estimated scatterplot smoothing, is a simple form of approximating two-dimensional relationships.
- It is also the default smoothing method in `geom_smooth()`
- Recreate the following visualization which uses LOESS to approximate the underlying functional form

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




### 1.2.2 Histograms (?) and density (`geom_density()`)

- What does a histogram show?
- Check the cheat sheet for the respective `geom_`
- Create a histogram of the variable `hwy`.
  - Change the binwidth to "2".
  - Hint: A histogram does not require a mapping for the y-axis.

We can also plot the densities using `geom_density()`

- Plot the density of `hwy` colored by class.
- What does this visualization show?

### 1.2.3 Boxplot (`geom_boxplot()`)

- The so-called "five-number summary" is often used to provide information about a variable of interest. It consists of five statistical measures, namely:
  - the **minimum** value of the sample
  - the **first quartile** (`p=0.25`), indicating that 25% are equal or below this value.
  - the **median** or **second quartile** (`p=0.5`), i.e. 50% of values in the sample are equal or below this value
  - the **third quartile** (`p=0.75`)
  - the **maximum** value of the sample

The boxplot is a visual representation of the five-number summary and a way to represent the distribution of variables. They are particularly useful when used side by side separated conditional on a categorical variable. These side by side boxplots can be used to compare distributions by factor.

- Visually compare the distribution of "highway miles per gallon" (`hwy`) by "type" of car (`class`).
  - Check the two variables. Which one is categorical?
  - Start with the standard template from the beginning
  - Name of the dataset is `mpg`
  - map the variable `class` to the x-axis
  - map the variable `hwy` to the y-axis
  - use `geom_boxplot()` as geometric object.
- Note that the `class` labels are barely readable. Add another layer that flips the coordinates (`coord_flip()`)
- Add a layer that fixes the labels (title, x-axis, y-axis) using the `labs(title="", ...)` layer which overwrites the automatic labels.

#### 1.2.4 Bar plots

- Create a bar plot of the `class` variable. Use the cheatsheet if you get stuck
  - flip the coordinates.
  - What does the bar chart display on the y-axis by default?

A bar plot just needs one variable for a mapping. By default it simply counts occurences of the specified variables and displays this count on the y-axis. A column plot (`geom_col()`) is a generalized bar plot and let's you specify what to plot on the other axis.

# 2 `economics`-data

## 2.1 Line plot (`geom_line()`)

For sequential data, i.e. time series we have to plot data in sequential order.

- Check the `economics` dataset: Observations? Variables?
- What is the average unemployment rate in the dataset?
- Plot the unemployment rate over time. What are the respective variables and on which axis do you want to map them?
  - Hint: `unemployment_rate=unemploy/pop`