Exercise Series 3

We need the dataset `d.jura`, which you get by typing

```
read.table('http://stat.ethz.ch/ stahel/courses/R/jura.txt',header=T)
```

The `d.jura` dataset is about a survey of heavy metals in soils in of region in the Swiss Jura mountains. It reports measurements for 259 locations and lists for each location

- the coordinates `x`, `y` (unit: km),
- the content of `cd`, `co`, `cr`, `cu`, `ni`, `pb`, `zn` in the topsoil (unit: mg/kg),
- and two factors characterizing the `rock` type and the `landuse`.

Type `str(d.jura)` and `summary(d.jura)` to gain some overview of the data.

1. Boxplots. First, we study the frequency distribution of the metal measurements.

a) Create a boxplot of the lead (`pb`) content. Add a title and label the axis with the variable name and the unit of the measurements.

b) As **a)**, but use a logarithmic scale for the axis and display the data as a horizontal boxplot. Add lines to mark the *guide* (50 mg/kg) and the *trigger value* (200 mg/kg) of the Swiss Soil Protection Ordinance. How many observations exceed the guide and the trigger value, respectively?

Hint: You can use `table()` in combination with `cut()` to answer the last question (cf. Problem 4, Exercise Series 2).

c) Create boxplots of the lead content for all categories of `rock`, using again a logarithmic scale on the axis. Add notches to display intervals for approximate 95 % two-sample tests of the group medians. Use shorter names to label the categories.
Does the lead content differ between the rock types? Why is the notch for category *portlandian* much wider than the notches of the other categories?

Hint: Use the argument `names` to pass the names of the categories to `boxplot()`.

d) Create the boxplots of all 7 metals in a single plot. Use again a logarithmic scale for the axis.

Scatterplots. Next, we create plots of the sampling locations and explore the spatial distributions of `rock`, `landuse` and `pb`.

2. **a)** Create a plot of the positions of the sampling locations such that both axes display a given distance by the same length. Add a title to the plot and label the axis with some names and the distance unit (km).

b) Create a plot that displays the spatial distribution of the rock type by different symbols. Add a legend to annotate the various symbols by the rock categories. Use just the first 3 letters for labelling the categories in the legend and suppress the bounding box around the legend.

Hints:

- Use `col=as.numeric(d.jura[,"rock"])` or `pch=as.numeric(d.jura[,"rock"])` to display the rock type. The function `as.numeric()` converts the *levels* (=categories [character strings]) of a factor to integers.
- `levels(d.jura[,"rock"])` extracts the names of the categories of the factor `rock` and `nlevels(d.jura[,"rock"])` returns the number of categories. Use theses functions along with `substr()` to create the labels of the legend.
- The argument `bty` controls the plotting of the bounding box (type `?legend` for details).

c) As **b)**, but add another set of larger, circular symbols to show the landuse by the different colors. Add for `rock` and `landuse` separate legends with the labels `Rock types` and `Landuse` as first rows.

Hint: Use `NA` to code missing items when defining a vector.

d) Now create a "bubble plot" for the lead content which shows the position of the sampling locations and the lead content by the *area* of circular symbols. Color the circles by rock type. Add one legend with the symbols corresponding to a lead content of 20, 50, 100, 200 mg/kg and one for `rock`. Can you detect any patterns in the spatial distribution of `pb`?

Hint: Use the argument `cex=sqrt(...)` to set the size of the symbols. Divide by a suitable constant to get a reasonable size of the bubbles.

e) Read the lead content for a couple of points from the plot generated in **d)** by using `identify()`. Create a similar plot but just for the data in the square defined by $1 < x \leq 2$ and $1 < y \leq 2$ and `identify` the lead measurements within clusters of adjacent points. What do you notice about the spatial variation of `pb` at different spatial scales?

Hint:

- Right-click to stop `identify`ing points.
- You can choose the displayed ranges of x and y by the arguments `xlim` and `ylim` in the call of `plot`.

f) Re-create a plot with the sampling locations and use `locator()` and `polygon()` to draw a bounding polygon around the survey area.

g) Find out what happens if you type `plot(y~x, data=d.jura, type="l")`. Read the help page `?plot` to find out what other values the argument `type` accepts. Explore them. What rule is used to connect the points?

h) Generate again a plot of the sampling locations and annotate them by the row numbers of the observations in the dataframe.