MATHS 7107
Data Taming
Practical 5

1 Preliminaries

- Setup an `RStudio` project for this prac.
- Download the following datasets and put them in a `/data` subdirectory:
  - `population.csv`
  - `wordrecall.txt`
- Open a new script.
- At the top of your script write the code to load these packages
  - `tidyverse`
  - `inspectdf`
  - `caret`
  - `moments`
- Save the script.

2 Population dataset

Questions:

1. Load the population dataset.
2. Check the `population.csv` variable for `NA`s.
3. Standardise the population variable.
   - You will probably need the `na.rm = TRUE` option.
   - Make sure you confirm the scaling worked.
4. Apply min-max scaling to the population variable.
   - Make sure you confirm the scaling worked.

3 Word recall dataset

Questions:

5. Load the `wordrecall.txt` dataset.
   - This is a tab-delimited file. Use `read_tsv()`.

6. Draw a scatter plot for `time` and `prop`.

7. Log transform each or both of the variables to find a linear relationship.

8. Try to find the the equation relating `time` and `prop`. Then make a new plot with `time` (unscaled) on the horizontal axis.

4 `meuse` dataset

The *meuse* dataset gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Variable *zinc* in *meuse* contains the topsoil zinc concentration.

We will use this dataset to practice a Box-Cox transformation on some univariate data.

Questions:

9. Load the `meuse` data in package `sp`.

   - remember to use `install.packages()`
   - then `library()`

10. Plot a histogram of the `zinc` data. Calculate the `skewness`.

> **IMPORTANT!**
>
> Note that the `skewness()` function in the `moments` package is the one we use for this course. There are other `R` functions to calculate the moments, but their algorithms are often slightly different. So make sure you use the command `moments::skewness()`.

11. Now log transform the zinc data and produce a histogram and find the skewness.

12. Use Box-Cox to find the optimal λ scaling.

13. Apply the scaling and produce a histogram and find the skewness.

14. Which scaling gives the best output?