This compulsory assignment covers some of the prerequisites of the DAT320 course, as well as some basic R programming skills. The goal is to revisit statistical fundamentals required throughout the course.

Exercise 1 (R syntax & data structures)

The goal of this exercise is to become familiar with the syntax and data structures in R. In particular, the R packages `dplyr` and `ggplot2` can be useful when working with R `data.frame`s. The dataset used in this exercise contains a time series of life expectancies in different countries.

Material provided:

- `gapminder.csv`

Tasks:

(a) *Load the dataset and convert categorical variables to factors and all other variables to an appropriate type. Print a summary of the variables in the* `data.frame`*.*

(b) *Plot the time series of each country (*`Year` *versus* `lifeExp`*), coloured by continent.*

(c) *Give a summary of the lifeExp for each continent and for each year. Compute minimum, median, mean, maximum, and standard deviations. Display this as a* `data.frame`*.*

Hint: Use the group_by(continent, year) and summarise(across())

(d) *For each continent, plot the average lifeExp across all countries per year (*`Year` *versus* `lifeExp`*) as a line plot. Add error bars or ribbons indicating ± 1 standard deviation. Hint: facet_grid(.∼continent) creates a graph for each continent.*

Exercise 2 (Elementary data analysis and model training)

In this exercise, we will perform an analysis of a dataset containing the weather history for a given location. The data set contains 96 453 samples and 12 variables, where one target variable is the Apparent Temperature (C).

Material provided:

- `weatherHistory.csv`

Tasks:

(a) *Import the file* `weatherHistory.csv`*. Look at the first few rows of the data. Discuss what types of variables you have. Which variables are categorical and which are numerical? Visualize the data (use bar plots and histograms with kernel density estimates). Discuss the data.*

(b) *Perform preprocessing using dummy/onehot encoding and standard scaling. Divide the data into* $75\% - 25\%$ *train-test split randomly.*

(c) *Train a linear regression model with all the predictors you find logical to include based on your analysis in A) and use Apparent Temperature (C) as the target. Calculate the metrics below and interpret the hypotheses tests for the model parameters.*

- *RMSE (root mean squared error)*
- *MAE (mean absolute error)*
- *coefficient of determination (R^2 score)*

Exercise 3 (Linear Regression and Diagnostic Plots)

In this exercise you will explain some fundamental concepts of linear regression and distributions.

Material provided:

- `diagnosticplot.Rmd`

Tasks:

(a) *Give a brief explanation of the four assumptions of linear regression and how to diagnose violations of these assumptions:*

a) ***Linearity***

b) ***Homoscedasticity***

c) ***Independence***

d) ***Normality***

Formula for Linear regression with multiple regressors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

(b) *Give a brief explanation of the four plots generated when you run the plot command on a model created with lm() in R:*

- *Residuals vs Fitted*
- *Normal Q-Q*
- *Scale-Location (or Spread-Location)*
- *Residuals vs Leverage*

(c) *Use the script in "diagnosticplot.Rmd" and generate data for linear regression that:*

- *Holds to all the assumption*
- *Breaks the assumption of Homoscedasticity*
- *Breaks the assumption of Linearity*

- *Breaks the assumption of Normality.*

Generate a plot for the data and diagnostic plots

`Hint:` If the relationship between Y and X is nonlinear, any non-linear component will remain in the residuals of the models.

(d) *Give an explanation of how the relationships between the Y and X in c) violates the assumptions for linear regression.*

Exercise 4 (correlation and partial correlation)

The goal of this exercise is to become familiar with the concept of partial correlation in relation to regular correlation.

Material provided:

- `weatherHistory.csv`

Tasks:

(a) *Give a brief explanation of the concept of correlation. How does scaling of the different variables X and Y affect the correlation (use formulas below for guidance)?*

$$\begin{aligned}\rho_{W_1W_2} &= \frac{cov(\alpha_1 X, \alpha_2 Y)}{\sqrt{var(\alpha_1 X)var(\alpha_2 Y)}} \\ &= \frac{\alpha_1\alpha_2 cov(X,Y)}{\sqrt{\alpha_1^2 var(X)\alpha_2^2 var(Y)}} \\ &= \frac{\alpha_1\alpha_2 cov(X,Y)}{|\alpha_1|\cdot|\alpha_2|\sqrt{var(X)var(Y)}} \\ &= sgn(\alpha_1\alpha_2)\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \\ &= sgn(\alpha_1\alpha_2)\rho_{XY}\end{aligned}$$

(b) *Do some research on what partial correlation is (e.g., Wikipedia) and explain the formula below. Give 3 examples of scenarios to investigate partial correlations and explain the relationship between X, Y and Z.*

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}$$

(c) *Does the property from part a) of this exercise hold for partial correlation? Justify the answer with the formulas given above.*

(d) *Import the dataset `weatherHistory.csv` and do the following:*

- *Select the features: `Temperature (C)`, `Apparent Temperatur (C)`, `Humidity`*

- *Compute the pairwise correlation of each pair of features.*
- *Compute the partial correlation once for each of the three possible conditional variables Z.*

Investigate the values from the partial correlation and pairwise correlation. Comment on the relationship between the variables and if they have any confounding effect on each other.