Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 0: Data IO

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Rilind Sahitaj, Artem Tsikiridis

Exercise H0.1 *Loading and Describing Data with Pandas*

In this exercise, you will perform basic pandas operations to analyze a dataset, provided in *LaborSupply1988.csv*. We recommend the *"10 minutes to pandas" guide*[1] for a quick overview of basic pandas functionality.

a) Read *LaborSupply1988.csv* into a pandas dataframe.

b) How many features (columns) and datapoints (rows) does the dataset have?

c) Which attributes does the dataset have?

d) List the first 10 datapoints of the dataset.

e) Determine the value range of the attribute "age".

f) Calculate the average of log annual hours (lnhr) worked by the labourers with $0, 1, \ldots, 6$ kids each.
Hint: `.groupby()`

g) Compute the average number of kids of the 40 year olds.

Exercise H0.2 *Plotting Data with Pandas and Matplotlib.Pyplot*

In this exercise, you will visualize the data provided in *LaborSupply1988.csv*. For common plot types and settings, pandas provides functions that can be accessed directly from the dataframe. More involved plots can be created via matplotlib.pyplot, or via other libraries such as seaborn.

a) Read the file *LaborSupply1988.csv* into a pandas dataframe.

b) Plot a histogram of the attribute "age". Which is the most frequent age?

c) Plot the average number of "kids" against "age" and interpret the resulting graph.
Compute the correlation between "kids" and "age" to check your interpretation.

d) Plot "log of hourly wage (lnwg)" against "age".

e) Plot the mean of "log of hourly wage (lnwg)" against "age".
Compute and discuss the type of correlation between "lnwg" and "age".

f) Plot "lnhr" against "age" with different colors for "disab=0" and "disab=1".

g) Create a boxplot of "lnhr" against "kids".
What can be observed regarding median and variance?
Is the observation meaningful for large values of kids?

[1] `https://pandas.pydata.org/docs/user_guide/10min.html`