Chair of Decision Sciences and Systems
Department of Informatics
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 0: Data-IO

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao
October 19, 2023

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". If you didn't already, have a look at https://pandas.pydata.org/docs/user_guide/10min.html for a quick overview of basic pandas functionality.

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

b) Determine the amount of attributes and datapoints inside the dataframe.

c) Which attributes does the dataset have?

d) Have a look at the first 10 datapoints.

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

f) Compute 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 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 "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 of "kids" and "age" to check your interpretation.

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

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