Business Analytics & Machine Learning
Homework sheet 6: Data Preparation

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise H6.1 *Data aggregation*

In an ideal world, you would obtain a tidy data set for your modelling purposes. However, it is usually necessary to aggregate your data set from multiple sources in many real-world scenarios before you can use it. In this exercise you will learn how to approach this problem. You can find a template for this exercise on Moodle.

1. Read the customer table (*customers.csv*), and order table (*orders.csv*) files into two separate tibbles.
2. Create a table that displays for each customer the associated information in the order table. Use a join to combine the information from the customer and orders table. Keep all entries of the customer table.
3. You want to use the information from the two table to generate another table that contains all information about the customers and their orders. We want to see only those customers that have placed an order yet.
4. In the next step, you are interested in merging the two columns and keeping all entries from both tables.
5. Finally, you want to derive a table of customer that have not placed an order yet, i.e., that do not have an entry in the orders table.

Exercise H6.2 *General data preparation*

The data set (*raw_data.csv*) contains data from an online shop. Table 1 describes the attributes and values.

1. Load the *raw_data.csv* and rename all attributes to match the *description* column in Table 1.
   Hint: *read_csv(), rename()*
2. Correct the data types for all nominal attributes and assign the corresponding labels from the *comment* column in Table 1.
   Hint: *replace(), as_type()*
3. Correct the data type for the ordinal attribute size and assign the corresponding labels from the comment column in Table 1.
   Hint: *upper(), Categorical()*
4. Correct the data type for all date attributes. Create separate attributes for weekday, year, month, day, and quarter of order date.
   Hint: *to_datetime()*

Attribute	Description	Comment
ID	ID	
od	order_date	
dd	delivery_date	
size	size	ordinal: S < M < L < XL < < XXL < XXXXL
tax	tax	
a6	salutation	nominal: 2 = Company, 3 = Mr., 4 = Mrs.
a7	date_of_birth	
a8	state	nominal: 1 = BW, 2 = BY, 3 = BE, 4 = BB, 5 = HB, 6 = HH, 7 = HE, 8 = MV, 9 = NI, 10 = NW, 11 = RP, 12 = SL, 13 = SN, 14 = ST, 15 = SH, 16 = TH
a9	return_shipment	nominal: 0 = No, 1 = Yes

Table 1 Attributes of the data set

5. Find missing values (only NA), fill missing prices/ tex with averages or remove the instances.
   Hint: *isna(), fillna(), dropna()*

6. Calculate a new attribute *delivery time* as the difference of *order* and *delivery date* in days. Inspect the values for error and set the value to NA for corresponding instances.
   Hint: Negative delivery time is impossible.

7. Plot a histogram for the new *delivery time* column. Then discretize, i.e., bin, it to levels "NA", "$\leq$ 5d", and "> 5d" in a new attribute *delivery_time_discrete* and plot a bar chart for it.
   Hint: *hist(), bar()*

8. Compute the correlation matrix for the numerical attributes only. Plot the matrix of the scatterplots. Plot the heat map of the correlation matrix.
   Hint: *corr(), scatter_matrix()*