Tutorial Business Analytics

Tutorial 7 - Exercise

Exercise 7.1

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

Table 1: Attributes

Attribute	Description	Comment
ID	ID	
od	order_date	
dd	delivery_date	
size	size	ordinal: "S"<"M"<"L"<"XL"<"XXL"<"XXXL"
price	price	
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	0: “no” 1: “yes”

a) Load “raw_data.csv” and rename all attributes to match the “description” column in Table 1.
Hint: *read_delim(), rename()*

b) Correct the data types for all nominal attributes and assign the corresponding labels from the *comment* column in Table 1.
Hint: *mutate(), factor()*

c) Correct the data type for the ordinal attribute size and assign the corresponding labels from the *comment* column in Table 1.
Hint: *toupper(), table()*

d) Correct the data types for all date attributes. Create separate attributes for weekday, year, month, day, and quarter of *order date*.
Hint: *mutate(), across(), as_date()* from package “lubridate”

e) Find missing values (only NA), fill missing prices/tax with averages or remove the instances.
Hint: *mutate(), across(), if_else(), na.omit()*

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

g) Plot a histogram for the new *delivery time* column. Then discretize (“bin”) it to levels "NA", "<= 5d", and "> 5d" in a new attribute *delivery_time_discrete* and plot a bar chart for it.
Hint: *hist(), barplot()*

h) Compute the correlation matrix for the numerical attributes only. Plot the matrix of the scatterplots. Plot the heat map of the correlation matrix.
Hint: *cor(), pairs(), ggcorr()* from package “GGally”

i) Standardize all numerical values and again compute their correlation matrix.
Hint: *scale()*

Exercise 7.2

In this exercise, you will implement a workflow for developing a model that predicts the power production of a power plant given historical data. The data set "power_plants.csv" contains this historical data.

a) Before we create our recipe, we want to incorporate external information on the country where the power plants operate. Read and join the "country.csv" on: "country_long" = "country" and "primary_fuel" = "fuel"
Hint: Use the *left_join()* function

Including external data allows us to create two additional variables that might be insightful for our modeling purpose.

b) Create a new recipe *rec* using the *recipe()* function of the tidymodels package. We will now add preprocessing steps to this recipe:

1. Each recipe requires a formula describing the set of independent and dependent variables like the model definition of regression models and a data set. Add these basic arguments to the recipe, where our dependent variable is *generation_gwh_2017,* and all remaining variables build the independent ones. We will use the train set as the data set.

2. Since our data set contains an *ID* column, which we do not want to include in our estimation, we update the role of this column to "ID".
Hint: *update_role()*

3. Turn dates into decimal values using the *decimal_date()* function in the lubridate package.
Hint: *step_mutate_at()*

4. Replace all NA values of the columns *cap_share_of_country_gen_by_fuel* and *cap_share_of_country_gen_total* with zero.
Hint: *step_mutate_at(), ~replace_na(.,0)*

5. Impute all other missing values of the remaining **numeric** columns by calculating their average.
Hint: *step_impute_mean()*

6. Impute all missing nominal values with the value "none".
Hint: *step_unknown()*

7. Convert all strings to factors, which do not have the role "ID" and are not "outcome" variables, i.e., dependent variables.
Hint: *step_string2factor()*

8. Finally, remove all constant columns of the predictors.
Hint: *step_zv()*

c) Add the recipe to a workflow.