Bank Loans Data Exploration Exercise

General information for candidates

This assignment tests your ability to explore data.

Business Problem

On the real exam it can be bewildering to look at the statistics and graphs and then interpret them in a way that is easy to explain. The data set is from the UCU machine learning repository, which is a massive online archive of data sets.

Specific Tasks

1. (5 points) Examine the target variable

    - Comment on the values of the target and what they represent.
    - Choose two categorical variables and look for trends in the percentage of subscribers.
    - If there are any categorical variables which have levels with fewer than 50 observations, put these into an "other" category.

2. (10 points) Decide on which variables (if any) to remove

    Read the data dictionary and if appropriate, discard any variables that you believe should be removed and explain your reasoning.

3. (5 points) Examine the numeric variables

    Comment on the statistical summaries and histograms.

4. (5 points) Examine the factor variables

    Inspect the bar graphs for three factor variables that you did not already inspect in Task 1 and comment on any trends.

Data Dictionary

Variable	Description
age (numeric)	Age of applicant
job	type of job (categorical
marital	marital status (‘divorced’,‘married’,‘single’,‘unknown’; note - ‘divorced’ means divorced or widowed)
education (categorical	‘basic.4y’,‘basic.6y’,‘basic.9y’,‘high.school’,‘illiterate’,‘professional.course’,‘university.degree’,‘unknown’)
default	has credit in default? (‘no’,‘yes’,‘unknown’)
housing	has housing loan? (‘no’,‘yes’,‘unknown’)
loan	has personal loan? (categorical
contact	contact communication type (‘cellular’,‘telephone’)
month	last contact month of year. Values are ‘jan’, ‘feb’, ‘mar’, …, ‘nov’, ‘dec’.
day_of_week	last contact day of the week. Values are ‘mon’,‘tue’,‘wed’,‘thu’,‘fri’
duration	last contact duration, in seconds (numeric). Important note - this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
campaign	number of contacts performed during this campaign and for this client including the most recent contact. Values are integers from 1 – 40.
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client (numeric)
poutcome	outcome of the previous marketing campaign (‘failure’,‘nonexistent’,‘success’)
emp.var.rate	employment variation rate
cons.price.idx	consumer price index
cons.conf.idx	consumer confidence index
euribor3m	euribor 3 month rate
nr.employed	number of employees
y	has the client subscribed a term deposit?