APP 2: Univariate statistics (task)

Input – The updated flat data file in csv format where the columns are the variables and the rows are the observations. The first row is a list of variable names. Representation of missing data as an exhaustive list of strings (all possible representations are included; e.g. ‘null’, ‘missing’, ‘Null’). The variable names would identify the type of feature, for instance, categorical variable name would start with “c\_”, and numeric variable name would start with “n\_”, other variable name such as “patient\_id” would start with “o\_”. The descriptive table only work for “c\_” and “n\_” variables.

Test:

In the resultant file generated:

1. For every cell, identify the variable name and the statistic name; verify if that statistic is correct

Univariate Statistics, Output 1 (overview):

Function: compute number of observations, number of numerical variables, number of categorical variables, list of all variables names and their type (numerical or categorical).This should be presented as in the sheet called OVERVIEW in the file ‘example\_of\_descriptive\_statistics.xlsx.

Univariate Statistics, Output 2 (categorical variables):

One row per variable.

Columns:

. Variable name

. Number of non-missing obs

. Percentage non-missing

. Number of missing

. Percentage Missing

. Number of levels

. First level

. Number of obs

. % of obs (calculated with respect to the number of non-missing observations), i.e. sum up the % of each level it would be 100%.

. continuing for all levels.

Please see example in sheet ‘univar – CAT’ in file example\_of\_descriptive\_statistics.xlsx

Univariate Statistics, Output 3 (NUMERICAL variables):

Output for all NUMERICAL variables can be put on a single table.

Columns:

. Variable name

. Number of non-missing obs

. Percentage non-missing

. Number of missing obs

. Percentage of missing

. Mean

. Standard deviation

. Min

. P1

. P5

. P10

. P25

. P50

. P75

. P90

. P95

. P99

. Max

Example of output is shown on sheet univar- NUM in file example\_of\_descriptive\_statistics.xlsx.