DANA4800 – Team Project – Phase 1 - EDA

Understand the data

A general strategy: In exploring a new dataset, the following basic sequence is often useful:

1. Assess the general characteristics of the dataset, e.g.:

   a) How many records do we have? How many variables?
   b) What are the variable names? Are they meaningful?
   c) What type is each variable—e.g., numeric, categorical, discrete, or logical? (Table 1)
   d) How many unique values does each variable have?
   e) What value occurs most frequently, and how often does it occur?
   f) Are there missing observations (vertically and horrizontally)? If so, how frequently does this occur?

2. Examine descriptive statistics for each variable
   For categorical variables, answer the main questions like:
   a) How many distinct values or “levels” does the variable exhibit
   b) How often does each of these levels occur in the dataset?
   c) How does the behavior of another variable X vary over the levels of C?

   For numerical variable, answer the main questions like:
   a) What is the mean, median, standard deviation?
   b) Does the data follow the normal distribution?

3. Where possible—certainly for any variable of particular interest—examine exploratory visualizations and identify anomalies
4. Look at the relations between key variables using the ideas of visualization and statistical tests

Table 1: An example of total number of features and their measures for mechanically ventilated patient dataset

Target variable	1	Binary
Demographic variables	2	Binary Discrete
Medical history variables	12	Binary
Disease severity variables	3	Discrete
Diagnosis variables	14	Binary
Vital signs variables	15	Continuous
Lab results variables	21	Continuous

Total	67 predictors 1 target variable	36 continuous 27 binary 4 discrete

Your task

Based on the ablove strategy, you will conduct an EDA, including:

- Missing values (horizontally and vertically) – identify patterns of missing values and then discuss the imputation methods with the instructors
- Outliers
- Univariate data distributions, including normality check (visualization + Statistical test)
- Pair-wise data distributions (scatter plots + correlation + chi-square test)
- Standarize data

Statistical tests for the descriptive analyses
Conduct t-test, or chi-square tests to identify the difference **between** survival and mortality groups. Conduct variance analysis to identify differences **within** the survival group. Same analysis for mortality group.

Note that:
- If an independent variable is nominal, chi-square is used to identify an association.
- If an independent variable is discrete, Wilcoxon signed rank test is used to identify an association.
- If an interdependent variable is numerical, t-test is used to identify the mean difference.

The report should include the following:

- Interpretation and findings of patient characteristics [overall and each group of patient]
- Interpretation and findings of medical history [overall and each group of patient]
- Interpretation and findings of disease severity [overall and each group of patient]
- Interpretation and findings of diagnosis [overall and each group of patient]
- Interpretation and findings of vital signs [overall and each group of patient]
- Interpretation and findings of lab results [overall and each group of patient]

The python notebook includes codes for analysis and the report includes your interpretation and findings. Note that you need to use both graphs and statistical tests to support your conclusions.

References:

Chi-square: https://ethanweed.github.io/pythonbook/05.01-chisquare.html

Compare two means: https://ethanweed.github.io/pythonbook/05.02-ttest.html

Mathematical explanation of between and within group variance: https://www.statology.org/within-between-group-variation-anova/

The analysis of variance: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3382318/