**Villanueva - Comini**

**Prova d’esame**

**Analisi descrittiva**

Per avere una prima visione delle variabili raccolte ho caricato il dataset “heart.csv” nella nuova variabile dataset con la funzione ‘read.csv’ e ho di conseguenza analizzato la struttura dei dati raccolti.

Leggendo il dataset mi sono accorto della presenza di valori NA dovuti ad esempio al data-entry manuale. Quindi ho creato un ciclo che controllasse la presenza dei valori. Tutti i valori NA sono stati cancellati definitivamente dal dataset con la funzione ‘dataset <- na.omit(dataset)’.

Ho continuato controllando e rimuovendo le colonne non ritenute necessarie, quindi ho visualizzato il dataset attraverso la funzione ‘view(dataset)’ e ho rimosso le colonne non necessarie attraverso la funzione ‘subset()’. Ho deciso di eliminare la colonna X perché in questo caso ritenuta inutile.

Terminato il passaggio dell’eliminazione delle colonne superflue, ho continuato rinominando le colonne in maniera appropriata e descrivendo il tipo di ogni attributo di ciascuna. Infatti ho stampato il mio dataset con la funzione ‘str(dataset)’, con la funzione ‘names(dataset)[names(dataset) == "vecchio"] <- "nuovo"’ ho rinominato le colonne in modo appropriato, come ultima azione ho assegnato ad ogni attributo il suo tipo.

names(dataset)[names(dataset) == "cp"] <- "chest\_pain"

names(dataset)[names(dataset) == "trestbps"] <- "rest\_bp"

names(dataset)[names(dataset) == "chol"] <- "cholesterol"

names(dataset)[names(dataset) == "thalach"] <- "max\_hr"

names(dataset)[names(dataset) == "exang"] <- "exercise\_angina"

names(dataset)[names(dataset) == "thal"] <- "thalassemia"

names(dataset)[names(dataset) == "target"] <- "heart\_disease"

names(dataset)[names(dataset) == "ca"] <- "n\_vessels"

names(dataset)[names(dataset) == "restecg"] <- "rest\_ecg"

age int ORDINALE

sex chr NOMINALE

chest\_pain int NOMINALE

rest\_bp int DI RAPPORTO

cholesterol chr DI INTERVALLO

fbs int DI RAPPORTO

rest\_ecg int NOMINALE

max\_hr int DI INTERVALLO

exercise\_angina int NOMINALE

oldpeak num ORDINALE

slope int NOMINALE

n\_vessels int ORDINALE

thalassemia int NOMINALE

heart\_disease int NOMINALE

Ho eseguito poi un controllo per correggere la consistenza del tipo di dato per ogni variabile, infatti, con la funzione ‘[colonna del dataset] <- as.[tipo nel quale voglio cambiare i dati](colonna del dataset)’, ho trasformato i dati in modo che siano tecnicamente corretti.

* Ho trasformato, nella colonna sex, i semplici valori "0" e "1" in "F" per femmina e in "M" per maschio e poi ho cambiato il tipo di dato da int a factor (quindi diviso in più livelli) per i miei attributi “F” e “M”.
* Ho cambiato il tipo di dato per la colonna chest pain da int a factor quindi diviso in più livelli (0 - 1 - 2 - 3).
* Per la colonna cholesterol ho trasformato in primo luogo tutti i valori "undefined" nella mediana dei valori di tutta la mia colonna e in secondo luogo ho traformato il tipo di dato da char a integer.
* Ho cambiato il tipo di dato per la colonna fbs da int a factor quindi diviso in più livelli (1 - 0).
* Ho cambiato il tipo di dato per la colonna rest\_ecg da int a factor quindi diviso in più livelli (0 - 1 - 2).
* Ho cambiato il tipo di dato per la colonna exercise\_angina da int a factor quindi diviso in più livelli (1 - 0).
* Ho cambiato il tipo di dato per la colonna slope da num a factor quindi diviso in più livelli (0 - 1 - 2).
* Ho cambiato il tipo di dato per la colonna thalassemia da int a factor quindi diviso in più livelli (0 - 1 - 2 - 3).
* Ho cambiato il tipo di dato per la colonna heart\_disease da int a factor quindi diviso in più livelli (1 - 0).

Per vedere se le modifiche fossero avvenute con successo ho stampato nuovamente il dataset.

Per far emergere gli outlier e le anomalie, come prima cosa ho, con la funzione hist(dataset$minimum\_nights), estratto un grafico a barre dei valori dei dati forniti in ingresso dal dataset. Di conseguenza ho deciso che il numero minore di notti non sia superiore a 365 e che il numero maggiore di notti sia il valore medio della variabile.

Ho rappresentato con un altro grafico più espilcativo, il boxplot, gli outliers relativi alla variabile prezzo boxplot(datset$price). Ho calcolato il 3° e il 1° quantile e li ho sostituiti nella formula IQR<-(Q3-Q1) per trovare lo scarto interquartile. Ho rilevato, di conseguenza, il range della differenza interquartile per creare il nuovo boxplot.

Ristampando a console il mio dataset, aggiornato e modificato, ho notato che è più consistente e che dunque non necessita di altre trasformazioni.