2024年1月29日

#### ハンズオン課題（データ前処理）

作成者：西村FE

パワーポイント資料「第2講 データの前処理」の内容を確認しつつ，以下の問いに答えなさい．

≪問題≫

1. 同一階層にある「01\_data.csv」ファイルを読み込み，各カラムの欠損割合を計算しなさい．
2. 1で読み込んんだデータフレームにて欠損値が確認できた場合，欠損値が存在する行を削除したデータフレームを作成せよ．また，欠損値が存在する列を削除したデータフレームを作成し，2つを作成して欠損値を削除する場合のデメリットを考えよ．
3. 1で読み込んんだデータフレームにて欠損値が存在する場合，ゼロで欠損値を補完したデータフレームを作成せよ．
4. 1で読み込んんだデータフレームに，新たにすべてのデータがNA値である列を追加し，pandas.dropna() メソッドの引数（how = “”）により，1つでもNA値が含まれる場合に列を削除したデータフレームとすべての列がNA値である場合に列を削除したデータフレームの2つを作成せよ．ただし，新たに追加する列のカラム名は特段指定しない．
5. 1で読み込んんだデータフレームにて，ディクショナリ（辞書）を利用して，Age列の欠損値を「1」，Salary列を列の平均値で補完したデータフレームを作成せよ．
6. 同一階層にある「02\_data.csv」ファイルを読み込み，すべての列が一致している行が10行存在する．このとき，この10行を削除したデータフレームを作成せよ．
7. 6で読み込んだデータフレームの各カラムに格納されたデータには重複するものが存在する．このとき，重複の検出対象を「Column1，Column5」とした場合における重複データを削除したデータフレームを作成せよ．
8. 同一階層にある「03\_data.csv」ファイルを読み込み，外れ値の有無を確認せよ．外れ値が存在した場合は，各カラムの平均値で外れ値を置き換えよ．  
   ヒント：NA値で保管した後に，辞書を用いて補完すればよい（他にも方法はある）
9. 8にて作成したデータフレームにて，データのランダムサンプリングを実施せよ．
10. 同一階層にある「04\_data.csv」ファイルを読み込み，カテゴリ変数「性別・地域・商品カテゴリ」のダミー変数変換を実施したデータフレームを作成せよ．また，もとのデータフレームと比較して，どの程度データが大きくなったか確認してみよ．
11. 同一階層にある「movies\_data.csv」ファイルを読み込み，「genres」列のダミー変数を行ったデータフレームを作成せよ．