Errors

1. In column “Age”, there is mix data which contains both number and characters.
2. In column “Weight” and “Height”, there is serval empty data which is unneeded.
3. The data in the column “Systolic_BP&Diastolic_BP” is wrong. It has two variables. It should be seperated into two columns: “Systolic_BP” and “Diastolic_BP”.
4. And the two colums: “Systolic_BP” and “Diastolic_BP” should be the type of numeric.

Steps to perform the data wrangling

Part I

1. For column “Age”:
   a) Use split syntax: data[“Age”].str.split(“ ”, expand = True) to split the number and character into two colums: “Age”, “Years”.
   b) Use data.drop('Years', axis = 1, inplace = True) to delete unneeded column “Years”.
   c) Use data['Age'] = pd.to_numeric(data['Age']) to change the type of “Age”.
2. For empty data in column “Weight” and “Height”:
   a) Use data.dropna(axis = 0, how='any', inplace = True) to delete the rows which contain empty data.
3. For column “Systolic_BP&Diastolic_BP”:
   a) Use split syntax: data["Systolic_BP&Diastolic_BP"].str.split("/", expand = True) to split it into two colums “Systolic_BP” and “Diastolic_BP”.
   b) Use data.drop("Systolic_BP&Diastolic_BP", axis = 1, inplace = True) to drop the original column "Systolic_BP&Diastolic_BP".
4. Change the data type in colum “Systolic_BP&Diastolic_BP”:
   a) Use data["Systolic_BP"] = pd.to_numeric(data["Systolic_BP"]) and data["Diastolic_BP"] = pd.to_numeric(data["Diastolic_BP"]) to change the type of data from object to numeric in this two columns.

Part II

1. delete columns that all rows are empty:
   a) data2.dropna(axis = 1, how='all', inplace = True).
2. delete inneeded (duplicate information) column “Value”
   a) data2.drop("Value", axis = 1, inplace = True)
3. delete column "Dim1 type" and change the column name of "Dim1" to "sex
   a) data2.drop("Dim1 type", axis = 1, inplace = True)
   b) data2 = data2.rename(columns={"Dim1":"Sex"})
4. use Boxplots to identify outliers of column "FactValueNumeric".
   a) plt.boxplot(data2["FactValueNumeric"])
   b) plt.show()
5. show missing values
   a) data2.isnull().sum()
6. split date into “Date” and “Time”, and delete “Time”
   a) data2[["Date", "Time"]] = data2["DateModified"].str.split(" ", expand = True)
   b) data2.drop("DateModified", axis = 1, inplace = True)
   c) data2.drop("Time", axis = 1, inplace = True)
7. change “Date” to date format and split them into three columns "month", "day, "year"
   a) data2["Date"] = pd.to_datetime(data2["Date"])
   b) data2["Month"] = data2["Date"].dt.month
   c) data2["Day"] = data2["Date"].dt.day
   d) data2["Year"] = data2["Date"].dt.year