# ADS Assignment 7.

You have been provided with 2 csv data files, 'heart' and 'insurance'. They shall both be used interdorm for analysis.

1. Import the 2 data sets, 'insurance' and 'heart'.
2. For the 'heart' data frame, rename the column 'target' to' heart disease'.
3. In the 'insurance' data frame, map encode the gender categories in the following procedure: a. Female – 0 b. Male - 1
4. Explore each data frame using at least 2 data exploratory tools of your choosing in pandas and interpret your observation in a markdown cell.
5. Assume the 2 data frames where taken from the same hospital. As a result, a few of the individuals who went through a heart check-up had insurance coverage. Utilize the 2 common columns to combine the 2 data frames to a singular data frame called df_all.
6. Visualize the age distribution for the column 'age' in both the df_all and the heart data frame. (Ensure your visualization is of an appropriate size for effective analysis)
7. What effects did the combination of the 2 data frames have on the age distribution? (Interpret your observation in a markdown cell.)

(Exclusively work with the data frame df_all from this point)

8. Isolate all the numerical column names into a list named 'numerical_continuous'.
9. Create a list containing all the numerical discrete column names called 'numerical_discrete'.
10. Visually identify if there is presence of any outliers in the columns and resolve them using a zscore test and a pvalue threshold of your choosing.
11. Validate that your analysis above was successful by visualizing the value distribution in the resulting columns using an appropriate visualization method.
12. Assuming the column 'charges' is your target for your regression analysis, feature select the best 'numerical_continuous' columns using the backward elimination method.
13. Isolate all the categorical column names into a list named 'categorical'.
14. Assuming the column 'heart_disease' is the target for your classification analysis, run a chi contingency test to identify the best categorical and numerical_discrete features to proceed with the analysis.
15. Using ColumnTransformer, OneHotEncode the categorical columns.