**Introduction to Machine Learning**

**Assignment for Week 1**

Given, UCI Riverside Heart Disease dataset, which has 14 attributes and 1 target variable.

**Description of attributes:**

1. age: Age of the patient in years.
2. sex: Sex of the patient (1 = male, 0 = female).
3. cp: Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic).
4. trestbps: Resting blood pressure (in mm Hg).
5. chol: Serum cholesterol level (in mg/dl).
6. cigs: Number of cigarettes smoked per day.
7. years: Number of years of smoking.
8. fbs: Fasting blood sugar (1 = fasting blood sugar > 120 mg/dl, 0 = otherwise).
9. dm: Presence of diabetes mellitus (1 = yes, 0 = no).
10. famhist: Family history of heart disease (1 = yes, 0 = no).
11. restecg: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable/definite left ventricular hypertrophy).
12. thalach: Maximum heart rate achieved.
13. exang: Exercise-induced angina (1 = yes, 0 = no).
14. thal: Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect).

**Target Variable:**

1. num: Target variable (0 = no heart disease, 1,2,3,4 are levels of heart disease ).

**Objectives:**

Impute missing data (don’t just throw it away)

'-9' means missing

Look at the data with EDA

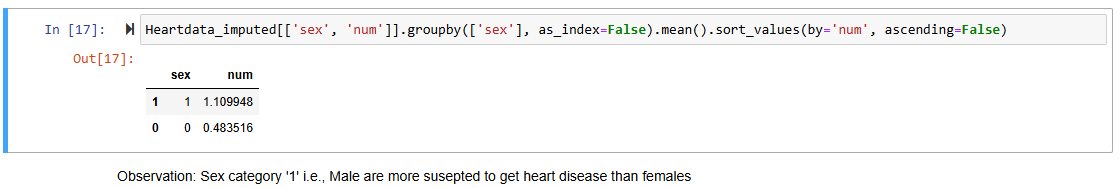
Drop unnecessary columns, if needed, and defend your decision to do so.

Deal with outliers (maybe make a boxplot)

Provide a brief summary of how you prepared your data for future analysis.

**Solution:**

1. **Aquiring data:**
   1. Read the data file and converted to data frame
   2. Looked at the top few columns and bottom few columns to see how the data looks like
   3. Seen the column names, and understood the description of each attributes
2. **Analyze the data:**
   1. Preview of the data
   2. Check for null values – not available as they are denoted as ‘-9’
   3. Replace the ‘-9’ with na
   4. Count the no of null values
   5. Check for which features are numerical and categorical
3. **Imputation of missing data**
   1. Check for the no of missing values
   2. Observation: the attributes that have missing values are cigs, years, dm, and thal
   3. Check the statistics of the features to see the count as well
   4. Clearly ‘dm’ attribute has very less data i.e., 23 datapoints and all these datapoints are equal to 1. And if we impute also all the datapoints will be 1 and hence it is constant throughout and hence not useful for modeling
   5. Removed the ‘dm’ attribute from the data set
   6. Used mean imputation for the numerical attributes i.e., 'age', 'trestbps', 'chol', 'thalach', 'cigs', 'years'
   7. Used mode imputation for the categorical attributes i.e., 'sex', 'cp', 'fbs', 'restecg', 'exang', 'thal', 'num'
4. **Univariate EDA**
   1. Analyze by pivoting features (analyze our feature correlations by pivoting features against each.

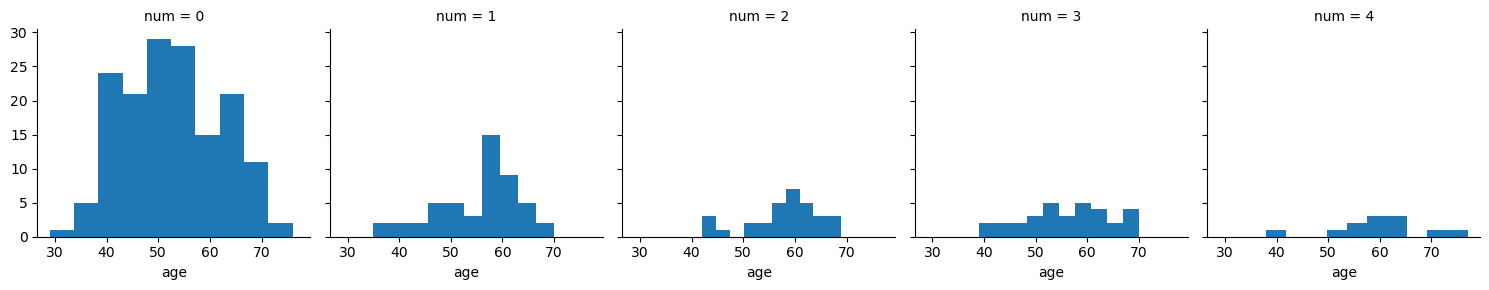


Based on this analysis of all the categorical variables, the key observations are:

1. Male are more suspected to get heart disease than females
2. Asymptomatic chestpain (cp category ‘4’) is having major effect on heart disease
3. The people who have fasting blood sugar > 120 mg/dl ('1') is suspected to have heart disease
4. It is well known that, the people who have family hiostory of heart diseases is suspected to have to heart disesase and is well alligned as per the results, but there is no major difference between with family history and without family history
5. Restcg category '2' ie., showing probable/definite left ventricular hypertrophy is suspected to have heart disease
6. exang category '1' i.e, people who has Exercise-induced angina are having high chances of getting heart disease
7. Thalassemia with category '7' i.e., reversible defect is suspected to have heart disease.

**Data visualization**

**Sample Univariate analysis:** Analyzing how each attribute is distributed wrt to

****

Key observations:

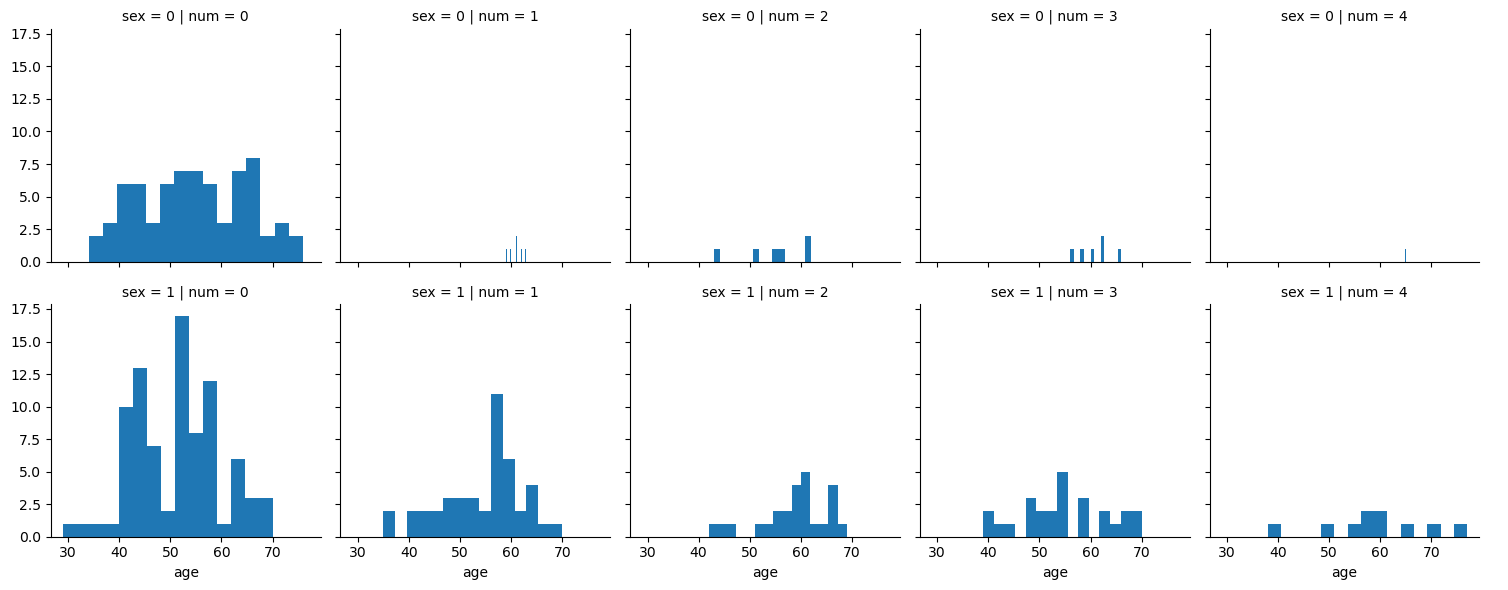
1. Age > 35 is tend to have heart disease

2. Age group 55 - 60 is more suspected to have heart disease of category '1'

3. Category 4 is very less frequent and is more found in the age group of 50 - 65 years of age

4. more people is suspected to have heart disease at 50-60 years of age

**Sample bivariate analysis:** Analyzing the bivariate dependency on heart disease. Here we analyse the heart desase wrt to two variables at a time.



Here, we have analyzed the integrated effect of sex and age on heart disease. Based on this analysis it is observed that :

1. Sex '0' i.e., females are less susceptible to heart disease, and the high severity i.e., cat 4 is only for female of age ~65 yrs
2. Females are very less susceptible to heart diseases irrespective of their age, and most of the people in females were identified at stage 2.
3. No of people having stage '4' heart disease is very less in both the sex categories

As we have 13 variables, doing univariate analysis and bivariate analysis manually is tediuous. For loop is used to do EDA.

**Univariate analysis :**

In case of numerical values, histograms and box plots are used to understand the distribution of data and identification of any outliers. the features like chol, cigs and years are not normally distributed and they have skewness. features such as trestbps, chol and cigs have skewness and outliers

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

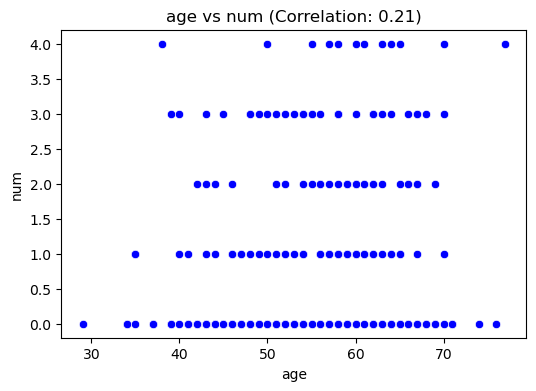
In case of categorical values, histograms are used to understand the distribution of data. From this analysis it is found that we have categorical varables is that are not equally distributed particularly fbps, restcg, thal.

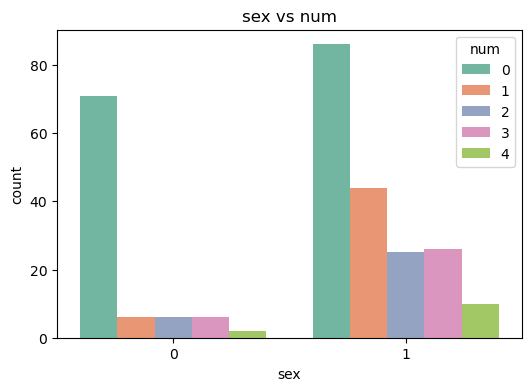
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | |

Bivariate analysis:

Scattered plot is used to analyze the data for bivariate analysis, initially we have seen how the input features are correlated to target variables.

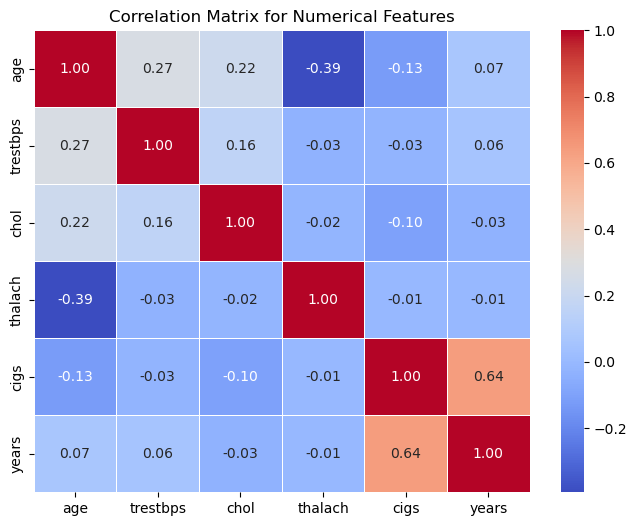
Based on corelation values for numerical attributes age, trestbps, thalach, are highly correlated to the heart disease.





For categorical variables, based on chi-square test, sex, cp, restecg, exang, and thal is significantly related to num.

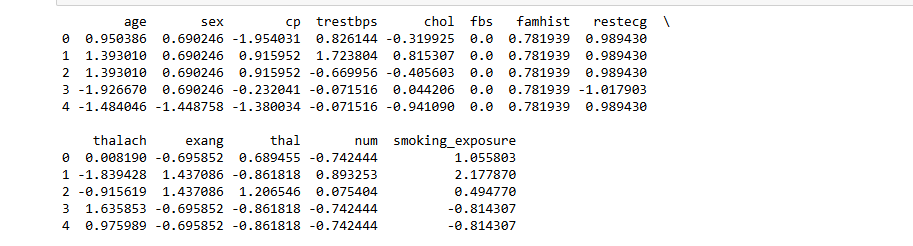
Correlation matrix:



Based on correlation matrix, No features are majorly correlated and hence these can not be removed.

**Feature engineering:**

As the Cig and years are releated to the exposure of a person to smoking they can be combined. A new attribute called Smoking\_exposure is made by combining cigs\*years\*365.



Outliers are also removed and then the data is normal.

**summary**

We have done various activities to clean, and make the data useful for modelling  
1. Viewed the data clearly, understood meaning of each features  
2. identified both the categorical and numerical variable  
3. Identified the quantity of data missing  
4. Removed insignificant features  
5. Imputation of datapoints  
6. univariate analysis and bivariate analysis  
7. outlier detection and treatment  
8. Combined the features, to form cigs and years and developed a new variable called smoking exposure.  
9. normalizing the data.