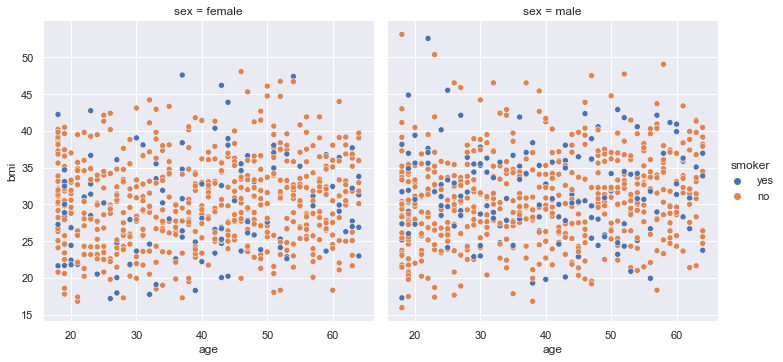
ML-Assignment 02

Sarvesh S

21BDA05

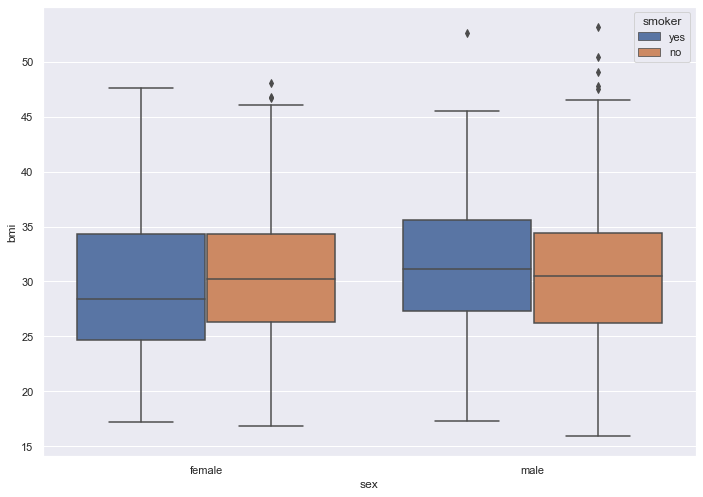
**1.Explaining EDA**

**1.relplot:**

****

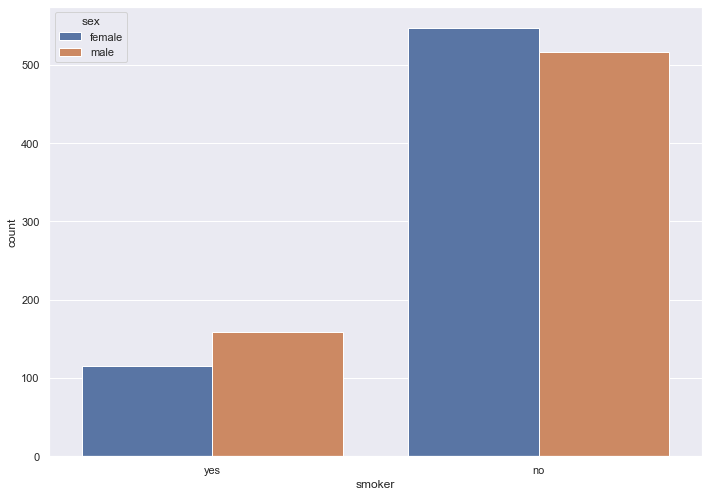
with the help of above relplot we can say that in both the gender the smoker has more bmi on an average than other

**2.boxplot:**



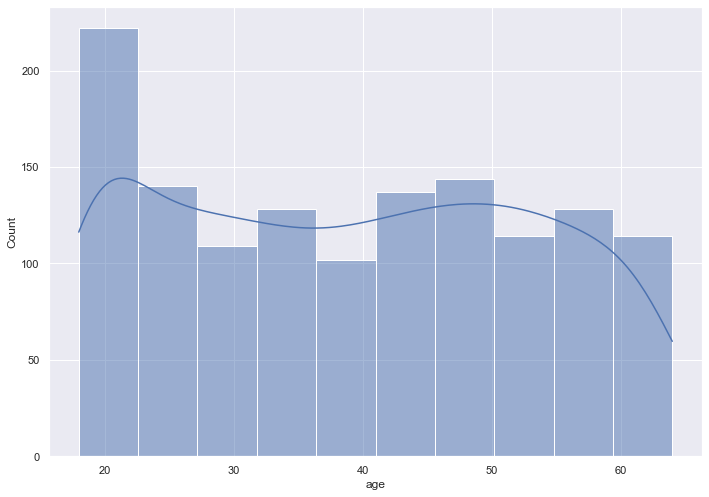
The boxplot is basically a visualisation of summarisation of our data here we ploted sex against bmi as we can see that the avg non smoker has bmi of 27 and the avg smokers bmi is 31

**3.countplot:**

****

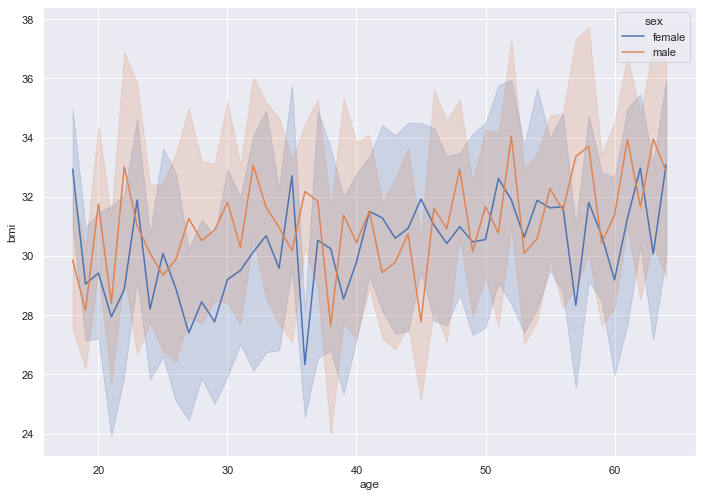
As the name suggest this plot is used to count how many values does the particular categories of value has and it represent that in the form of barchart.

**4.histogram:**

****

**Histogram** is a bar chart without space between that and it is useful to say that particular value is fallows normal distribution or not(i.e symmetric or not).

**5.lineplot:**

****

The above plot is used to see at what particular age(x-value) we can get corresponding bmi(y-value) with the help of this we can find trend in our data. Is there any irregular pattern in our data.

**2.Answer the following questions:**

**i. What are the assumptions of linear regression?**

1.Linearity: There must exists a linear relationship between independent variable ‘y’ and dependent variable ‘x’

2.Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3. Homoscedasticity: The residuals have constant variance at every level of x(predictor)

4.Normality: The residuals of the model are normally distributed.

5.multicolinearity: there should no correlation between the independent variables

**ii.How can we evaluate a Regression model? Define each metric and its interpretation.**

we can calculate regression model using 3 metric:

* Mean Squared Error (MSE).
* Root Mean Squared Error (RMSE).
* Mean Absolute Error (MAE)

**Mean square error:**

Here “least squares” refers to minimizing the mean squared error between predictions and expected values.The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

**Root mean square error:**

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

**Mean absolute error:**

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

**iii.Can R squared be negative?**

Yes, r square can be negative (i.e) r square doesn’t always be the square of anything sometimes if the fit is actually worse than just fitting horizontal line leads to negative r square (R square can have a negative value when the model selected does not follow the trend of the data, therefore leading to a worse fit than the horizontal line)

**iv. What is dummy variable trap?**

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables.so if there are m categorical variable we should have to take m-1

**v.Is One Hot Encoding different from Dummy Variables?**

The disadvantage of one hot coding leads to dummy variable trap and this can be overcome by droping a column of dummy variable

**vi.How is polynomial regression different from linear regression?**

Polynomial Regression is a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x).Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable. But through linear regression we can able to find the linear relationship b/w the variable

**vii.Interpret the screenshot below from the notebook we discussed in class today**

at first we split our data into two testing and training in training there is 80% of data and in testing the rest thing why we are splitting like this means we can know the model is working for new values or not

score denotes the r square value (i.e how much amount variation explained by the independent variable) for eg. Model score b/w x\_test and y\_test is 0.9085 that represent 90.85% amount of variation in y\_test is explained by x\_test