**ML Lab Assignment 2**

**Jugvi Mankad**

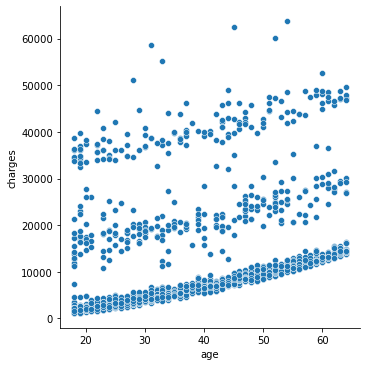
**21BDA49**

**Question 1:**

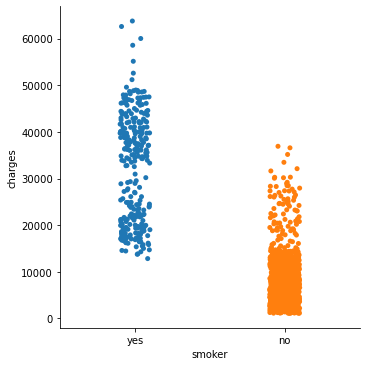
**Document 5-6 key insights from EDA and support each point with a visualization.**

**Answer:**

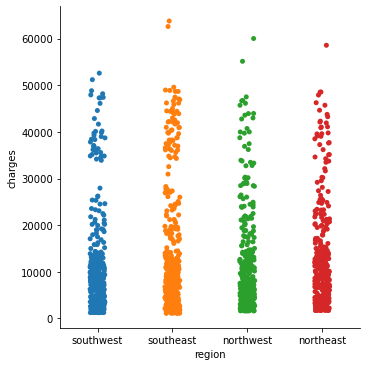
1)As age increases, there is a slight increase in the insurance charges.



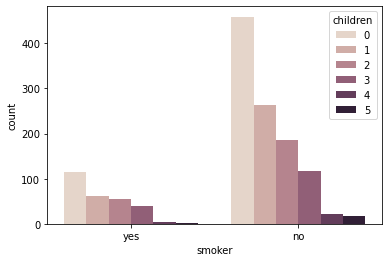
2)Base insurance charges for smokers are more than base charges for non-smokers.



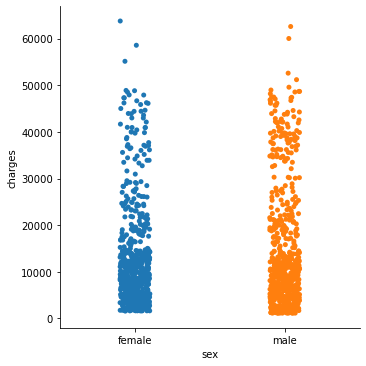
3)Insurance charges do not depend upon the residential region of a person. (All regions have almost equally distributed insurance charges)



4)Smokers usually have fewer children than non-smokers.



5)Insurance charges are independent of sex of a person.



**Question 2:**

**Answer the following questions:**

**Answer:**

**1)What are the assumptions of linear regression?**

🡪There are 4 assumptions of linear regression:

**(i) Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

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

**(iii) Homoscedasticity:**The residuals have constant variance at every level of x.

**(iv) Normality:**The residuals of the model are normally distributed.

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

🡪There are 4 main metrics for evaluation of a regression model.

(i)Mean Absolute Error (MAE):

It is measured by taking the average of the absolute difference between actual values and the predictions.

The less the value of MAE the better the performance of your model.

(ii)Root Mean Square Error (RMSE):

The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It represents the sample standard deviation of the differences between predicted values and observed values

(iii)R-Squared:

It measures how well the actual outcomes are replicated by the regression line. It helps you to understand how well the independent variable adjusted with the variance in your model. That means how good is your model for a dataset. Usually, the value of R^2 lies between 0 to 1. The closer its value to one, the better your model is.

(iv)Adjusted R-Squared:

The value of R^2 keeps on increasing with the addition of more independent variables even though they may not have a significant impact on the prediction. To solve this pitfall, an Adjusted R^2 value is used instead of R^2 value. Adjusted R^2 will penalize the model whenever you add a new variable to it. Hence, using an Adjusted R^2 value you can better understand the effect of the additional variables to your model.

**3)Can R-Squared be negative?**

🡪 Yes, it can be negative if the regression line somehow has a worse fit than the average.

**4) What is dummy variable trap?**

🡪Regression is usually used to quantify the relationship between one or more predictor variables and a response variable.

To use categorical variables as predictor variables, we use dummy variable trap. We use variables called dummy variables. These are variables that we create specifically for regression analysis that take on one of two values: zero or one.

We use the categories as columns/fields and put 0 or 1 as yes or no/ true or false accordingly.

**5) Is One Hot Encoding different from Dummy Variables?**

🡪Both are conceptually same, but the writing approach is different.

In dummy variables, we create different columns for categorical variables and put 0 or 1 in the rows, 0 being false/no/absent and 1 being true/yes/present.

In one hot encoding, we make a 1\*k array of combination of 0 and 1 (similar to dummy variables) instead of putting them in different columns. This is a more compact representation.

**6) How is polynomial regression different from linear regression?**

🡪 Linear regression is a basic and commonly used type of predictive analysis which usually works on continuous data. We fit a straight line to check for errors and minimize them.

But what if we saw two variables in our data set and knew that they were correlated, but the relationship doesn’t look linear? So hence depending on what the data looks like, we can do a polynomial regression on the data to fit a polynomial equation to it.

Polynomial regression line is not a straight line like a linear regression line.

**7) Interpret the screenshot below from the notebook we discussed in class today:**

🡪 model.score(x\_test, y\_test) : This gives the accuracy of the testing set.

model.score(x\_train, y\_train) : This gives the accuracy of the training set.

**8)Automated EDA options:**

🡪(i)Pandas Profiling

(ii)Autoviz