# Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

**Answer:** After a thorough analysis of the given dataset, it can be stated that variables like *temp* and *yr* are highly correlated to the dependent variable. It is clearly evident in the *EDA* as in the final model, *weathersit* was highly negatively correlated and two dummy variables from it are present in the model as mist and rain. Seasonal variables like *spring*, *summer*, & *winter* depict high multicollinearity evident from the high *VIF* values. *holiday* feature showed low correlation with *cnt* and is in the model with the lowest *VIF* count depicting least collinearity with other variables and hence becomes an important predictor. Other than *yr* and *holiday* all the variables in the model are in some way related as all of them are some sort of seasonal or climatic variables. This means that the business / dependent variable's fate resides highly on the weather climate and seasonal factors/predictions.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Answer: For categorical variables like *Attendance* which can only have two possible values *Present* and *Absent*, if we form a dummy variable with both the variables as new features when one will be high other will logically be low this will render the features to possess highest form of negative collinearity (-1) with each other. Similarly, if variables with more than two categories are converted to a feature each, they will introduce high multicollinearity between each other and might seem redundant to the algorithm, to include in a model, even if they might be significant. These things can further cost us the accuracy of the model and hence it is rather highly important that we use **drop_first=True** while creating new dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Among the numerical variables *temp* has the highest correlation with the target variable, which looks logical now that we know that our model is highly dependent on climatic, seasonal and weather-related features which possess collinearity with temperature (*temp*).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: The assumptions of the linear regression can be stated as follows:

A. Linearity

Linearity can be explained from the EDA analysis of the pair plot itself. Significant variables like temp are colinear to the dependent variable.

B. Homoscedasticity

Homoscedasticity can be observed in the distribution of errors as the data does not follow any patterns of heteroscedasticity like linear, quadratic or funnel shaped distribution, hence the data is Homoscedastic.

C. Mean of Residuals

The mean of the residuals here is close to zero, so we can say that this is a good regression model

D. Independence

I can be clearly observed that the variables of the model are not correlated. The corelation values of the features is close to zero or very low, hence multicollinearity is very low in the given model.

Also, during testing the autocollinearity the residual terms were found to be in the zone of the confidence interval region.

E. Normality

The distribution of the Residuals of the model follows a normally distributed curve with the peak right over zero and the bell-shaped curve tapering towards the ends.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The features like yr, temp, and rain are the top contributing features as the absolute values of the coefficients of these features is the highest. Due to this high value of their coefficients these features can drastically change the value of the dependent variable, even with small changes in the feature values. As these features contribute the most to the predicted value of the dependent variable, they hold the highest significance to the model.

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is a method to analyse the linear relationship between the predictor variables and a dependent variable. It takes into consideration that the data follows a linear relation among the variables, this means that if the value of one related variable changes the value of the other variable associated to it will change accordingly.

Mathematically given by

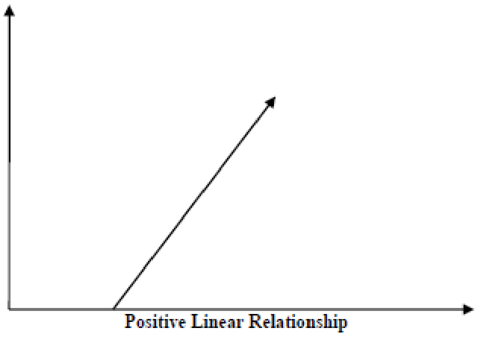
$$y = \beta_0 + \beta_1 X$$

which is also the equation of a straight line where y is the dependent variable, and X is the predictor or independent variable. $\beta_0$ is the y intercept and $\beta_1$ is the slope of the line or the coefficient of the predictor variable, its value decides the extent of the relationship between X and y.

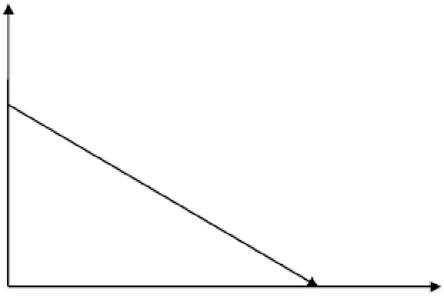
Linear Relation can be of two types positive and negative

If the value of the dependent variable (y) increases when the value of the predictor variable increases then this relation is call as positive linear relation.

Whereas if the dependent variable (y) decreases when the value of the predictor variable increases then this relation is call as negative linear relation.



Positive Linear Relationship



Negative Linear Relationship

Linear regression can be of two types Simple linear regression and multiple linear regression. Simple linear regression has only one predictor variable influencing the value of the dependent variable.

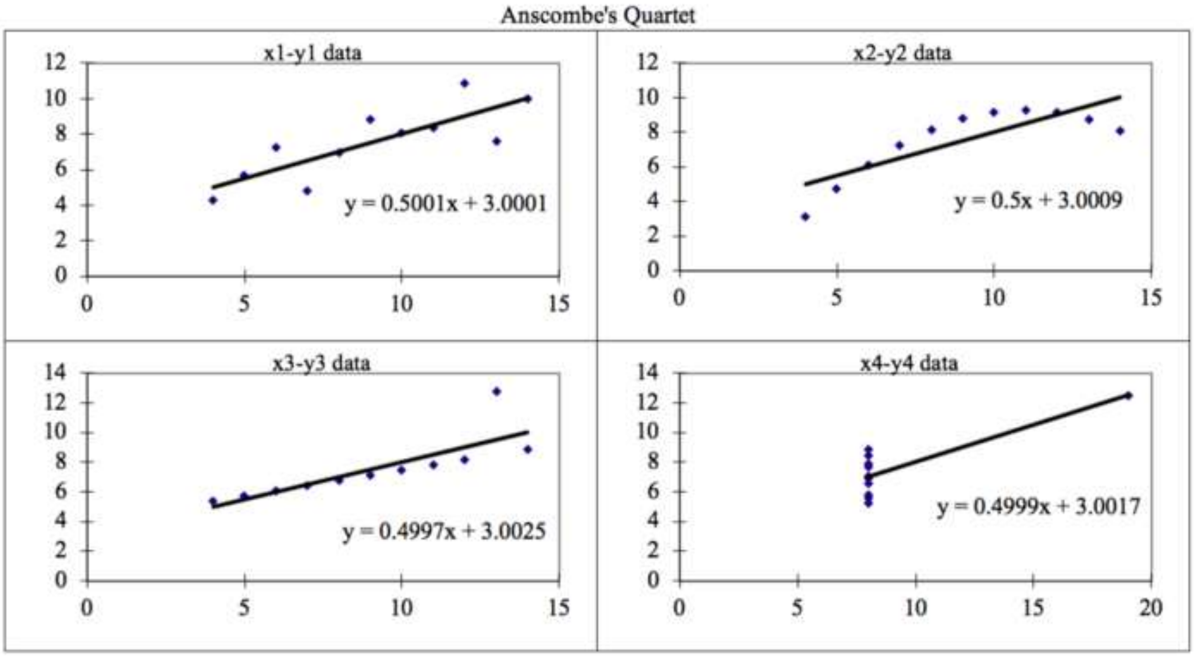
In multiple linear regression there are multiple predictor variables involved in the equation which is used to predict the value of the dependent variable

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots\ldots \beta_n X_n$$

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet is a pair of four data sets which have almost similar descriptive statistics but have a completely different distribution. These four data sets can be immediately be differentiated from one another when represented graphically. I was discovered by an English statistician Francis Anscombe. It was developed in order to explain the importance of plotting data before analysis, as at that time calculations were considered more exact and graphs were considered rough.

- In the figure the first graph appears to be a linear relationship.
- The second is not normally distributed and is neither linear.
- The third graph should have had a different regression line but is a little offset due to the outlier.
- The fourth has all same x values and one outlier that proves to be the high leverage point to give high correlation coefficient.



3. What is Pearson’s R? (3 marks)

Answer: Pearson’s r is a measure of covariance of two variables divided by the product of their standard deviations. Covariance can only reflect the linear relation between variables and neglects all other relations. The value of person’s correlation is normalised that is it is always between -1 and 1, this makes it a great measure foe comparison of multiple corelation variables with drastically different values.

It is mostly used in hypothesis testing that is if the corelation coefficient is 0 or not. It is also used to derive the confidence intervals.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: In Machine learning scaling is done to bring all the value of all the numerical features in the same range.

Consider in a model $y = \beta_0 + \beta_1X_1 + \beta_2X_2$, say if X1 is temp in degree Celsius and X2 is the price of gold per 10 grams. The value of temp will probably be in two digits and the value of gold per 10 grams will be in thousands, if we do not scale such values even if the coefficient of the temp feature is high, due to its significance, it will not be able to make much contribution to the model as it’s value will always be way smaller than that of gold, and hence scaling is performed to bring those values in the same range.

Normalised scaling brings the data to the values between 0 and 1. Standardised scaling will replace the value of the feature with their calculated z values. Normalisation will tend to lose outlier data in all cases where as standardising them will retain the information.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: VIF or Variance inflation factor is an indicator of multicollinearity. If in a dataset the features are correlated then the VIF values will be high, The higher the collinearity between the features higher the VIF values be if there is perfect correlation between the features the VIF values will be infinite.

VIF = $1/1-R^2$
$R^2$ = square of correlation
Hence if correlation is 1 then $R^2 = 1$
Which means

VIF = $1/1-(1)^2$ = $1/1-1$ = $1/0$ = $\infty$

Hence mathematically proven when correlation is high (1) then the VIF becomes infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q plot, short for quantile-quantile plot is a graphical plot to display the data from two sources or distributions. It is generally use to identify if the two given datasets follow the same distribution. In that case if the data of the two distributions fall in the same line of about 45-degree from x and y axis than it can be said that the data from the two sources follow the same distribution or it can also be said that they belong to the same population. And if the plot has datapoints that are scattered away from the 45-degree line, then the two data sets follow different distributions and are not from the same population.