# Subjective Answers

Assignment based subjective Questions

Ques 1 From your analysis what can you infer about the categorical variables from the dataset, what could you infer about their effect on the dependent variable??

Answer. The categorical variables in analysis like windspeed, weather sit has negative impact on demand since they have negative coefficients while year, holiday and month have positive impact on demand of the Bikes.

Ques 2 why it is important to use drop first =True during dummy variables?

Answer it is important to use drop first =True during performing dummy variables to ensure to drop redundant variables to reduce unnecessary columns that increases the efficiency of the linear model. It forms n-1 where n is the no. of the levels in the column.

Ques 3 Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer - Looking at the pair-plot among the numerical variables, “Temp” and “Atemp” has the highest correlation with the target variable(cnt)

Ques 4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer to validate the assumptions of the linear regression we need to check for thr residual analysis, error should be normally distributed and finally have to check for r2 for the train and test model and difference should be within the limits.

Ques 5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are ‘humidity’, ’year’ and ‘season’. with their large coefficient’s values compared to others.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer Linear regression model is one of the type of SUPERVISED LEARNING which is basically used on the “Continuous Variables”. It is basically used to predict the target variable using the given dataset of previous records by putting best fit regression line to predict parameter using “least squared method” Examples: calculating the marks of the students, predicting the sales of the company etc.

Ques 2. Explain the Anscombe’s quartet in detail.

Answer **Anscombe's quartet** comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

It is discovered by statistician Francis **Anscombe's** to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Ques3 What is Pearson’s R?

Answer In statistics, the Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is**a measure of linear correlation between two sets of data.** It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1.

Ques 4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer Scaling in model formation refers to the reducing all the variables within a certain limit that it will not affect the result. Scaling does not affect the data spread and variance. Scaling need to be done because all values have different scale leads to wrong interpretation about the data like their coefficients.

Normalised scaling refers to scaling the values within 0 and 1 which is done by Minmaxscaler and all values lies within it whereas standardised scaling refers to the just standardising the values using the mean of he data and dividing with it

Difference between both the two is that there is less variance in normalisation as compared to the standardisation

Ques 5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer An **infinite** **value** of **VIF** for a given independent variable indicates that it can be perfectly predicted by other variables in the model. Looking at the equation above, this happens when R 2 approaches 1. It happens when the corresponding variable may be expressed exactly by a linear combination of other variables. VIF value below 5 are acceptable whereas below 2 is good and above 5 shows relationship between variables so they needed to be dropped.

Ques 6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer -- A **q-q plot** is a **plot** of the quantiles of two distributions against each other, or a **plot** based on estimates of the quantiles. The pattern of points in the **plot** is used to compare the two distributions. The main step in constructing a **Q**–**Q** **plot** is calculating or estimating the quantiles to be **plotted**. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don’t, your residuals aren’t Gaussian and thus your errors aren’t either.

Quantile-Quantile (Q-Q) plot, is a graphical tool for determining if two data sets come from populations with a common distribution such as a Normal, Exponential, or Uniform distribution. This **helps** in a scenario of linear regression when we have the training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.