**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?**

Ans. In the bike sharing dataset, lets consider the effect of the categorical variable ‘weathersit’ on the target variable ‘cnt’. While performing EDA, I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately. Similarly, certain inferences could be made ‘season’ and ‘yr’ as well.Also, during model building on inclusion of categorical features such as yr,season etc we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

**2. Why is it important to use drop\_first=True during dummy variable creation?**

Ans. During dummy value creation (dummy encoding) it is advisable to use drop\_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

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

Ans. The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt’.

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

Ans. The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no and little linearity is present. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

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

Ans. Based on final model top three features contributing significantly towards explaining the demand are:

1.Temperature (0.5174).

2.weathersit\_2 (-0.282).

3.year (0.232).

**General Subjective Questions**

**1. Explain the linear regression algorithm in detail.**

Ans. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Regression analysis is used for three types of applications:

1.Finding out the effect of Input variables on Target variable.

2.Finding out the change in Target variable with respect to one or more input variable.

3.To find out upcoming trends.

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

Ans. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

**3. What is Pearson’s R?**

Ans. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. "Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

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

Ans. WHAT?

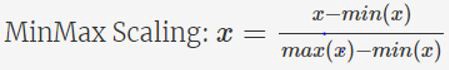
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

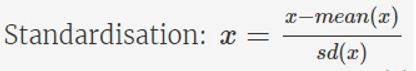
Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler**helps to implement normalization in python.



Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).



**sklearn.preprocessing.scale** helps to implement standardization in python.

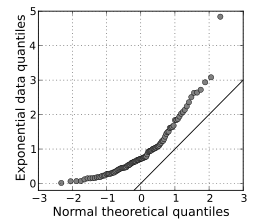
One disadvantage of normalization over standardization is that it **loses** some information in the data,especially about **outliers**.

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

Ans. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. A Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.