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)**

   Ans: The temp and atemp has the highest correlation with the target variable cnt temp and atemp are highly co-related with each other.

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

   Ans: It helps to reduce the extra column created during dummy variable creation. So that it reduces the correlations created among dummy variables.

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

   Ans: The highest correlation was between Cnt (target) and temp (numerical variable).

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

   Ans: VIF seems to be almost accepted (<5). p-value for all the features is almost 0.0 and R2 is 0.821. Difference between R-squared and Adjusted R-squared values for this model is very less, which also means that there are no additional parameters that can be removed from this model.

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

   Ans: The top 3 features contributing demand of the shared bikes are
   - temp (temperature)
   - weathersit_bad (weathersit bad)
   - yr (year)

General Subjective Questions

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

   Linear Regression is a machine learning algorithm. It is supervised learning. Linear regression models a target variable using a predictor. The predictor is an independent variable and target is dependent variable. It is used to find relationship between dependent and independent variable. Linear regression predicts a dependent variable value (y) based on the independent variable value(x). It shows linear relationship between x (input) and y (output). The equation for linear regression is as below:

   $$y = \theta 1 + \theta 2 * x$$

   Here,
   y= target variable,

$\theta 1$= intercept,
$\theta 2$= slope/coefficient of x,
x= predictor

Linear regression algorithm is to find the best values for $\theta 1$ and $\theta 2$, so that we have the best fit line. The best-fit regression line, will give minimum difference between predicted y value to actual y value. So, it is very important to update the $\theta 1$ and $\theta 2$ values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). Cost function, also known as Mean Squared Error (MSE), is a function that measures the performance of a Machine Learning model for given data. Cost Function quantifies the error between predicted values and expected values and presents it in the form of a single real number. The equation for MSE is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2$$

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

Anscombe's quartet comprises four data sets that have nearly identical simple statistics but when they are plotted, very different distributions appear and very different when graphed. Each dataset consists of eleven (x, y) points.
Anscombe's quartet shows why data visualization is important even before analyzing the data. • Here, the first plot shows simple linear relationship between two variables.
• The second one is not normally distributed but shows some relationship between variables but it is not a linear relationship.
• The third graph has a linear relationship but has different regression line. The outlier will influence the correlation in this case.
• Finally, the fourth graph shows an example when one high leveraged point is enough to produce a high correlation coefficient while other points are not showing any relationship.

3. What is Pearson's R? (3 marks)

Pearson's R is a linear correlation between two variables X & Y. It has a value between +1 and −1. Here 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The formula is as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

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

Scaling is a method used to standardize the range of features of data. Scaling is necessary because te data value may vary widely. Scaling is important in data pre-processing while using machine learning algorithms. During data pre-processing standardization and normalization are used.

Standardization:
Standardization is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with mean as zero and standard deviation as 1.

μ=0 and σ=1 where μ is the mean and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

Whereas Normalization is Min-Max scaling shrinks the data to a range of -1 to 1 or 0 to 1. It works well if the distribution is not Gaussian or standard deviation is very small. If the data is having outliers then normalization is not good to use and in such cases standardization can be used.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

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

(3 marks)

The **variance inflation factor** (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It helps to diagnose multicollinearity in a model.

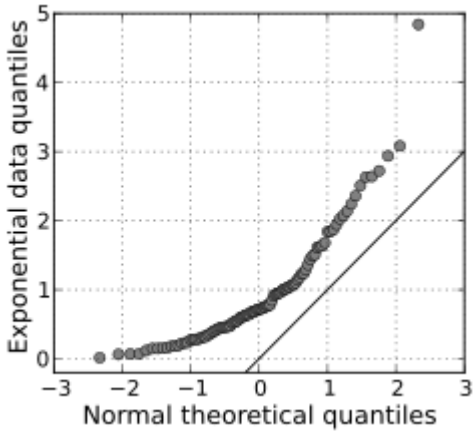
$$VIF = \frac{1}{1 - R^2}$$

VIF is computed for all the predictor in a model. If the value of VIF is 1 then predictor is not correlated to other predictors. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

(3 marks)

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.