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 our data set we have following categorical variables:

Season: We found that in spring our target variable cnt has least value and in fall it is quite high

yr: Demands were high in 2019 as compared to 2018

mnth: highest rentals were made in September and least in January which is in line with weather condition.

holiday: Rentals were reduced during holiday.

weathersit: High demands were on good weather days like Clear, Few clouds, Partly cloudy.

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

Ans: For handling categorical variable we create dummy variable and use it for better interpretation and scaling afterwards. Since one important aspect of regression problem is that predictor variable should not be correlated but in case, we don't drop first column of dummy variable it will lead to multicollinearity issue and it will become difficult for our model to tell how strongly a particular variable affects the target. In such case the coefficient of regression model will not convey the correct information.

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

Ans: temp and atemp are highly correlated with target variable (cnt).

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

Ans: We found in residual analysis that error terms are normally distributed around the mean (0).

Graphically it was clear after plotting distribution plot.

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

Ans: a) **yr** having coefficient **1991.30598653**

b) **temp** having coefficient **3072.6222**
c) **weathersit** having coefficient **687.1111**

General Subjective Questions:

1) Explain the linear regression algorithm in detail?

Ans: Linear regression is a form of supervised machine learning algorithm which is used for prediction of numerical values. A simple linear regression model attempts to explain the relationship between a dependent and independent variable using a straight line (y=mx+c).

Here independent variables are also called as predictor variables and dependent variable is called as target variable. Regression is broadly divided into two categories.

a) **Simple Linear Regression**: When dependent variable is predicted using only one independent variable.
b) **Multiple Linear Regression**: When dependent variable is predicted using multiple independent variables.

Equation for MLR is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Also, for linear regression we make few assumptions noted below:

- Linear relationship between X and Y
- Error terms are normally distributed
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

2) Explain the Anscombe's quartet in detail?

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. The core motive behind this was to lay emphasis that visualising data can at times gives a holistic view which can be easily overlooked in plain descriptive statistics.

3) What is Pearson's R?

Ans: In statistics, the Pearson correlation coefficient (PCC), 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.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

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

Ans: 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.

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 must 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 or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as: X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes, but the age is close to uniform.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected. Standardization does not get affected by outliers because there is no predefined range of transformed features.

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. 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.