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?

   There are two different types of categorical variables, and we know that they need to be handled and made numeric before using to predict the dependent variable. Both the categories should be treated differently before modelling.

   Ordinal Variables - Here the variables are categorical but, have some sort order based on their magnitude (Intensity). We normally try and handle them by Numeric Encoding, which is first sorting on magnitude (ascending or descending anything will work), and then assigning them numeric values corresponding to their intensity. I am not sure whether I am right or not, but if I am not very sure of intensity difference between any two pairs of ordinal variables, I will use one hot encoding in those cases as I do not want the model to get biased by getting assigned single coefficient which can be avoided by using one hot encoding.

   Nominal Variables – Here the variable is categorical and do not have any sort of order between the values. The best way to handle these variables is to one hot encode. If there are lot many different categories, we can also just keep top N occurring values and assigning one dummy value representing the other categories.

   For One Hot Encoded values, the response variable gets affected by the coefficient assigned to each different category if the encoding is 1. No effect if it is 0.

   For Numeric Encoded values, the response variable gets affected by the product of the common coefficient assigned times the values of the encoding assigned.

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

   It is important to use drop_first= True, while creating dummy variables because the first category can be represented by the other variables getting assigned 0 in their one hot encodings. If we do not drop these then it will increase the multicollinearity between the variables, which can make the coefficients of the model unstable, and we won't get proper stimuli on the response variable by increasing or decreasing other independent variables. It should be noted that it will not affect the prediction of the response variable, it will only affect the explainablity of the model.

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

   Even tough I did not make pair plot of the numeric variable with the target variable. I did plot the correlation matrix with correlation coefficients. 'temp' and 'atemp' were the two variable that had the highest correlation with the response variable 'cnt'. Correlation coefficient of both the variables were 0.63.

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

There are 5 major assumptions made with Multiple Linear Regression:

Linearity and Additivity: Linear Regression is a parametric model; we first assume of linearity, (change in independent variable by one unit keeping other independent variables would change the response variable by a value of the coefficient amount). Additivity meaning effect of every variable change should get added up and effect the response variable. Not sure how to check this assumption, but when we use the liner equation, we are assuming this fact.

No Multicollinearity: This assumption while very important, is not practically possible. A change one independent variable should ideally not affect any other independent variable. In practical situations we try to minimize this. I performed checks by validating that the correlation coefficients of the dependent variable should not be very high and dropped variables in situations where it happened. I also used Variation Inflation Factor to eliminate other variables that were collinear finally having all VIFs less than 4.

Normal Distribution of Residuals: I checked that the residuals should be normally distributed.

No pattern followed by the residuals: I plotted the residuals and made sure that the scatter plot of the residuals did not have any patterns. They should not be dependent on value taken by residuals over time and should be random, 0 cantered and 0 mean.

Homoscedasticity: I did check for homoscedasticity by plotting the scatter plot of residuals vs predicted y values. This check is made to make sure that the variance of residuals across the predicted value should not change at different locations of the scatter plot.

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

Based on the values of the coefficients shown below. Highest contributing features were, ‘atemp’, ‘weathersit_light_snow_rain’ and ‘year’ contributing significantly towards the demand.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2299.5501	212.469	10.823	0.000	1882.106	2716.994
holiday	-733.9257	233.136	-3.148	0.002	-1191.974	-275.877
atemp	3889.0827	272.207	14.287	0.000	3354.269	4423.896
windspeed	-1117.2627	224.244	-4.982	0.000	-1557.841	-676.685
year	2054.0806	73.592	27.912	0.000	1909.492	2198.670
season_spring	-1047.4453	133.812	-7.828	0.000	-1310.350	-784.541
season_winter	322.8149	108.302	2.981	0.003	110.031	535.598
weathersit_Li_sn_rn	-2322.3385	222.091	-10.457	0.000	-2758.688	-1885.989
weathersit_Mist	-699.3971	78.456	-8.914	0.000	-853.542	-545.252
month_July	-394.5762	146.763	-2.689	0.007	-682.925	-106.227
month_September	482.7971	139.522	3.460	0.001	208.675	756.920

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is parametric model where we try to model the response variable as the sum of the independent variable times a coefficient associated with that variable + some random error (constant coefficient). The same linear equation can be extended to multiple independent variables. As the equation formed is linear equation, hence the name.

An increase of variable x by one unit keeping the other variables constant will result into the target variable to increase by the coefficient units.

After we have formed the equation (the same behaves as cost function) we try to minimize the value of cost function and try to determine the values of the coefficients for which we achieve the minimum. This minimization is mostly done using gradient descent which minimizes the cost function iteratively converging towards the local minimum.

We need to make certain assumptions and check these assumptions before we conclude if the model is good fit or if we should even use linear regression. I have already explained the assumptions and how to validate the assumptions in 4th Question of Assignment Based subjective questions.

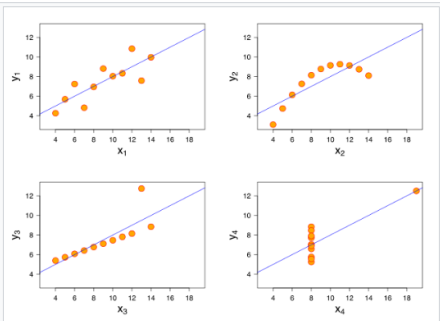
Finally, when we have learnt the coefficients (after training) we use the same equation we had learnt with new data points and calculate the response variable for the new sets of inputs (prediction)

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to visualize the importance of proper Exploratory Data Analysis, why we should not use the summary statistics only to make decisions on the data.

The visualization shows four different scatter plots who have same mean, standard deviation, variance, correlation, linear regression slopes and intercepts even tough the scatter plots are totally different.

The first dataset show that the response variable is linearly dependent on x.
Second shows that we should use a higher order curve to get the best fit.
Third shows how the presence of even one outlier can affect the slope and intercept of exactly liner dataset.
The fourth dataset with one x producing different y with the help of only one high leverage point can totally change to slope, intercept of regression and correlation coefficient.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson's R?

Pearson's R is the most used correlation coefficient to calculate linear correlation. It can take values between -1 and 1. -1 and 1 values indicate very high correlation in negative and positive directions, respectively. 0 value indicates no correlation. The magnitude determines the strength and sign determines the direction of relationship between the two variables. It can be calculated using below formula. Also expressed as covariance of the variables by the product of their standard deviation. As it uses mean for calculation, its value is highly impacted by outliers. In such cases one can use spearman's rank correlation coefficient.

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

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?

Scaling is used to get all the variables in similar scale. It may or may not impact the distribution of the variables based on the type of scaling being used. It might impact the model's performance (accuracy or any evaluation score) based on the scaler being used. It makes gradient descent faster, achieving convergence faster.

Difference between normalized scaling and standardized scaling:

a. Normalized Scaling does not change the distribution of independent variable while Standardizing does.
b. Normalized scaling rescales the range between 0 and 1 while Standardizing centres the data around mean with 1 standard deviation.
c. Normalized scaling is more sensitive to outliers present in the data while Standardizing is less sensitive to outliers.
d. Scaling cannot impact model evaluation while normalizing can have an impact.

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

We know that Variance Inflation Factor is calculated for a variable by calculating 1/(1-R^2) where R^2 is the coefficient of determination of the variable we are computing VIF for using other independent variables. We know that R_squared can take values till 1. When it takes the value of 1 it means that we can exactly fit the data, in our case we exactly determine the value of variable that we are calculating VIF for. And when this happens the denominator approaches 0 in the above formula, making VIF value infinite. So, when we have a variable that can be exactly determined by using other independent variable the determinable variable will have infinite value as its VIF.

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

A Q-Q plot is used to compare two types of distribution. One can use a theoretical distribution and make a scatter plot with another unknown distribution and decide on the type of distribution. If nearly all the points fall on the 45-degree line then, both the distributions are having same type of distribution.

In linear regression one needs to check the type of distribution followed by the residuals and validate one of assumptions. One can confirm if normal distribution is being followed by the residuals by making a Q-Q plot (by making scatter plot of a theoretical normal distribution and the residuals).