Assignment-based Subjective Questions

Ques1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
-- The cnt of bike sharing for first season is least
--The number of bike shares increased in 2019
--The cnt values increases in summer months
--The cnt values are less during holidays
--Weekday and working day do not show major variation in their results
--The cnt has zero values for weather situation - 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Ques2. Why is it important to use **drop_first=True** during dummy variable creation?
--Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Ques3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
--temp and cnt has highest correlation as temp increases count for bike sharing also increases.

Ques4. How did you validate the assumptions of Linear Regression after building the model on the training set?
--**Assumption 1**: The Dependent variable and Independent variable must have a linear relationship.A simple pairplot of the dataframe can help us see if the Independent variables exhibit linear relationship with the Dependent Variable.
Assumption 2: No Perfect Multicollinearity. The Variables with high Multicollinearity can be removed altogether, or if you can find out which 2 or more variables have high correlation with each other, you could simply merge these variables into one. Make sure that VIF < 5.
Assumption 3: Residuals must be normally distributed. If the Residuals are not normally distributed, non–linear transformation of the dependent or independent variables can be tried.

Ques5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
-- Temperature, Seasons and month are significantly contributes the demand of the shared bikes.

General Subjective Questions

Ques1. Explain the linear regression algorithm in detail.

-- Linear regression is defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change, the value of dependent variable will also change accordingly. Mathematically the relationship can be represented with the help of following equation –

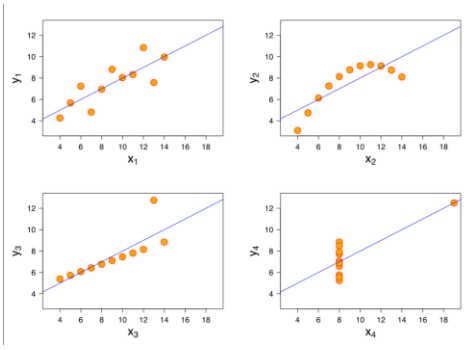
$$Y=mX+b$$

Here, Y is the dependent variable. X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y. b is a constant, known as the Y-intercept.

Ques2. Explain the Anscombe's quartet in detail.

-- Anscombe's quartet comprises four data sets 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. Anscombe's Quartet warns of the dangers of outliers in data sets. if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict. For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably *should* be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analyzing it.



Ques3. What is Pearson's R?

-- The Pearson correlation coefficient, also called Pearson's R, is a statistical calculation of the strength of two variables' relationships. In other words, it's a measurement of how dependent two variables are on one another. The Pearson product-moment correlation coefficient depicts the extent that a change in one variable affects another variable. This relationship is measured by calculating the slope of the variables' linear regression. The value of Person r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between

the variables. If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in first variable will result in a decrease in the second variable. If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.

Ques4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
-- Scaling means that you transform your data to fit into a specific scale, like 0-100 or 0-1. To scale the data when we use methods based on measurements of the distance between data points, such as supporting vector machines and the k nearest neighbors. With these algorithms, a change of "1" in any numeric characteristic has the same importance.
Scaling helps utilize available resources to a maximum, and makes a trade-off between marginal cost and accuracy. Minimizing human involvement.
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation**.** In this case, the values are not restricted to a particular range.Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

Ques5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
--This is happened when estimated regression coefficient increases and predictors are correlated in this situation model isn't going to be as reliable so in this situation precise estimates is needed.

Ques6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
-- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.