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)

***Answer***:

The following are the inferences after analyzing the effects of categorical variables on dependent variable 'cnt':

a. The categorical variable 'yr' (year) suggests that the demand for shared bikes increased from previous year to the current year. This gives us a potential trend of the demand going up in the upcoming year.
b. The highest demand was during the 'fall' season.
c. The demand steadily increases from the beginning of the year in the month of January till the month of September. From September the demand gradually drops.

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

***Answer***:

a. The *pd.get_dummies()* is used for 'one-hot encoding' of categorical variables. This means that for every unique value in the column, a new column of dummy variable is created.
b. The attribute *drop_first=true* of *pd.get_dummies()* is used to drop the first such created dummy column since the same amount of information can be represented with ***n-1*** columns, where ***n*** is the number of unique values (or layers). This is done to reduce the number of redundant columns in the dataset.

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

***Answer***:

a. The numerical variable 'temp' 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? (3 marks)

***Answer:***

a. Validated the 1st Linear regression assumption of 'Linear relationship between dependent and independent variables' using pairplots.
b. Validated the 2nd Linear regression assumption of 'Error terms must be normally distributed (Normality)' by plotting a *sns.distplot()* of error terms.
c. Validated the 3rd Linear regression assumption of 'Error terms distribution must have mean at zero' with the above plot.
d. Validated the 4th Linear regression assumption of 'Homoscedasticity - The residuals have constant variance at every level of x' by plotting a line graph showing the variance of error (y_train – y_train_pred) against the error_range (length of X_train).

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

***Answer:***

a. 'temp' (temperate) has a positive impact on demand of the shared bikes.
b. 'windspeed' has a negative impact on the demand.
c. There will be less demand during moderate weather (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and bad weather conditions (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), so the business can focus on maintenance activities during this time.

General Subjective Questions:

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

***Answer:***

a. Linear regression is one of the very basic ***supervised*** models of machine learning, where we train a model to predict the behavior of a given dataset based on some variables (or features). In the case of linear regression, there exists a linear relationship (y=mx +c) between the target variable (dependent variable) on the y-axis and the independent variable on the y-axis.

   Mathematically, we can write a linear regression equation as: $y = \beta 0 + \beta 1 x$, where, y is the dependent variable(target), β1 is the slope and β0 is the intercept and x is the independent variable.

b. The goal of linear regression algorithm is to get the best values for β0 and β1 to find the best fit, with least possible error. In order to minimize the error(variance), *Ordinary Least Squares (Cost function)* method is used to make the vertical distance from the data points to the regression line as small as possible.
c. Gradient Descent is a method for updating β0 and β1 to minimize the above cost function.
d. We validate the fitment of a LR model using R-squared and adjusted R-squared metrics. Moreover, residual analysis can also help validate the fitment of a good ML model.
e. After performing residual analysis, we need to validate the following assumptions of Linear regression algorithm:
    a. Linearity i.e., linear relationship between dependent and independent variables.
    b. Homoscedasticity - The residuals have constant variance at every level of independent variable.
    c. Normality – The error terms are normally distributed with mean at zero.

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

***Answer***:

a. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
b. 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 analyzing it and the effect of outliers on statistical properties.
c. The 4 graphs plotted as 4 quartets (with identical summary points) where – linear, non-linear, single-point with one outlier, constant value a x-axis with one outlier.
d. The descriptive summary statistics that were identical are mean, standard deviation, and correlation between x and y.
e. This goes on to prove that we shouldn't reply simply on crunched statistical numbers, but instead should visualize the data points to understand the story data is trying to say.

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

***Answer***:

a. In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's R, 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.
b. Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
c. The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 8 means there is a strong association

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

***Answer***:

a. Scaling is a data pre-processing step before training the ML model, 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.
b. Normalization typically means rescaling the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
c. In normalization, the scaling values are bounded. Whereas in standardization there is no bounded ranges.
d. Sklearn python package provides MinMaxScaler for normalization scaling, whereas it provides StandardScaler for standardization scaling.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

***Answer***:

a. VIF stands for Variance Inflation Factor. This indicates whether a variable can be well explained by any other variable in the feature set.
b. If there is perfect correlation, then VIF equals infinity. This shows a perfect correlation between two independent variables.
c. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this issue, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
d. 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. (3 marks)

***Answer***:

a. 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.
b. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
c. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. If they from a common distribution, the points will fall on that reference line making a 45-degree angle on the Q-Q plot.
d. 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.