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 categorical variables do show a correlation to the target "count".

The ***Year*** (difference of year from start year) affected the count the most, probably accounting for the popularity increase in the 2nd year.

The *weather-related variables* had similar effects on the count, wherein Summer/summer months/higher temperatures increased the count with a similar proportion.

Working day/holiday related based categories had least effects, where they increased the count only marginally.

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

ANS:

Categorical variables have finite set of values. In these cases, we can have *(n-1) dummy variables* where n is the number of values in a categorical column. Simply because a 0 in all n-1 dummy columns would imply a value of 1 for the nth column.

This would ***reduce in the number of features*** that we have to deal with. Drop_first =True, drops the first value and uses the values in all other columns to derive column 0.

Smaller number of features also ***increase the Adjusted R-Squared*** value of a Linear Model

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

ANS:

A pair plot or a heat map would show that the temp or atemp have highest correlation with target.

Also temp/atemp variables themselves are highly correlated, implying only one of them would be needed for analysis.

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

ANS:

1) We validated the assumptions of Linear Regression with ***the VIF values that showed reduced collinearity***.
2) Scatter Plots for linearity (between actuals/predictions) show that there is a linear relation
3) Plotted error terms that shows the mean is approximately zero and error distribution is normalized.
4) Homoscedasticity that shows equal variance of the predictions. However, the model is still not perfect in terms of homoscedasticity.

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:

1) The top feature contributing to the demand is ***temperature***,
2) This is followed by ***humidity***.
3) The ***year*** flag which is a categorical value is the 3rd most contributing feature although ***windspeed*** is very close to it.

General Subjective Questions

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

ANS:

Linear Regression is a machine learning algorithm that performs a regression task to predict a dependent variable value (y) from one or more independent variable (X_i). The algorithm assumes a linear relationship between y and X_i and the output is a linear coefficient for each X_i.

The output on a plot is a regression line which is a best fit for the model. The mathematical function explaining the Linear model can be written as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + - - - - - - - - - - + \beta_i X_i$$

The goal of the Linear Regression algorithm is to identify the independent variables (X_i) and to determine the β_i coefficient of each (X_i), such that the linear equation line is a best fit for the data.

A best fit line determined by the linear regression model, also means that the error terms (i.e. the difference od each actual y from the best fit line) for any given X follows a normal distribution with the mean on the best fit line.

The best fit for the linear regression model can also be determined by the R-squared value of the model, which is a measure of the proportion of the variance of the dependent variable explained by the independent variable(s) in the model.

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

ANS:

- Anscombe's quartet is a demonstration, involving 4 completely different sets of statistics, to show how completely different distributions can still have the same descriptive statistics.
- It goes on to prove effectively, that descriptive statistics by itself may not describe the distributions accurately.
- Effect of outliers and influential observations are well described by graphical representations rather than numerical properties such as mean, sample variance, correlation or even Linear Regression line.

Hence it is imperative that graphical analysis is part of and sometimes the initial step to analyze relationship of data within a dataset.

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

ANS:

Pearson's R, also known as Pearson's Correlation Coefficient (PCC), quantifies the linear correlation between two variables. The value of PCC varies between -1 and 1, where -1 represents a total negative linear correlation and 1 represents total positive correlation.

The formula for PCC for two random variables is given by:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$ where *cov* is covariance, σ_i is the standard deviation of i

Or,

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$ where E is expected value, μ_i is the mean of i.

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

ANS:

- Scaling is the process of standardizing the variable values of a distribution, to bring them into a common scale.
- This is required or preferred in order to reduce the variability of the data the lies because of the unit value. For example, when plotting a weight measurement, a 1KG increase in KG would have a different slope than the same increase when depicted in LB.

Normalized	Standardized
Normalized scaling rescales the values to fit in range between [0,1]. So, the minimum value in the range maps to 0 and the maximum value maps to 1, all values in between are a fraction obtained by dividing the value with the difference between max and min.	Standardize Scaling rescales so that all the values are depicted by the units of 'standard deviation' σ that the values deviate from the mean. In statistical terms, the values are converted to the Z values.
$V_n = \frac{V}{X_{max} - X_{min}}$	$V_s = \frac{V - \mu}{\sigma}$
Normalized values are always positive, between 0 and 1.	Standardized values are centered around 0 and can be positive or negative.
Outlier values may need to be excluded to normalize the values.	Standardization does not need to exclude the outliers but would be depicted with a higher Z-score.

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

VIF is a measure of correlation between coefficients. The value of VIF is determined by:

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

Where R^2 is the coefficient of determination. A perfect value of R^2 i.e. 1 means that the variable can be perfectly determined by the other variables.
Also, in that case the denominator of the above equation equals to 0, hence VIF is infinite.
So, a VIF of infinite would mean that the coefficient is highly corelated with other coefficients and hence can be dropped as a feature.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANS: Q-Q plot is the abbreviation for quantile-quantile plot, which is the plot used to compare 2 distributions to see if they come from or belong to the same distribution. Quantiles are intervals of equal probabilities for a distribution. A Q-Q plot between 2 distributions would therefor depict the how the distribution of the probabilities relates to each other,

In Linear regression, a Q-Q plot can therefore be used to plot the theoretical and standardized residuals. If the plot follows a linear slope, then the residuals are normally distributed.

If they deviate from the straight line, then the residual distribution is not normalized, and the linear regression model needs to be looked at.