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 :

- `season` seems to be having quite impact on count of Bike Riders, the Bike Riders are increasing mostly in Autumn/Fall season, followed by summer season and being the lowest in spring season.
- The Bike Rides can be seen having huge rise in year 2019 due to company's increasing awareness in the market.
- `mnth` seems to be showing increase in count of bike rides from the beginning of the year and being at the peak in mid year months (highest in Sept) and then later decrease depicting the season cycle again.
- `weathersit` depicts decrease in bike rides with increase in mist, rainfall or snowfall.

2. **Why is it important to use <u>drop_first=True</u> during dummy variable creation? (2 marks)**

   Answer: While creating dummy variables we use drop_first because a column having N levels, can be depicted by N-1 levels combined. Like in the assignment season variable/feature had 4 levels and while dummy variable creation it has been turned to 4-1 ie 3 variable columns.

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

   Answer : Variable 'atemp' has the highest correlation with target variable at 63.0685%

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

   Answer : To validate the assumption like normality I plotted histogram of error terms depicting a normal distribution with mean centered to 0.

   To check Homoscedasticity, plotted scatter plot with dependent variable & residuals to check any pattern.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Answer: Below mentioned are highest impacting features –

   a. atemp – Higher temperature attracts more bike rentals
   b. yr – Higher rentals made due to growth made over time
   c. Light Rain/Snow Weather – It had a negative impact, thus, increase in Rain/Snow decreases the rentals.

General Subjective Questions

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

   Answer: Linear regression is a widely used statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit straight line that represents the linear relationship between the variables. The best fit line is given by –

   $y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n + e$

   where, **y** is dependent variable, **β** is coefficient of **X** independent variable, **e** is error term

   Linear regression is one of the easiest and most popular machine learning algorithms. It is a statistical method used for predictive analysis.

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

   Answer : Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

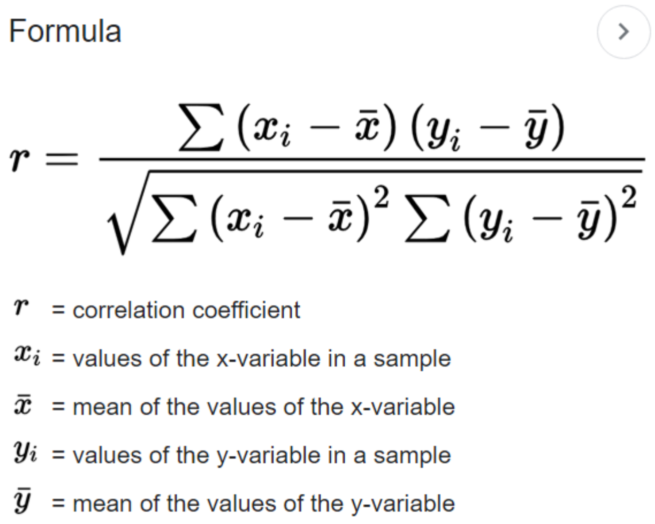
   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

   Answer: Pearson's R is a measure of linear correlation between 2 numeric variables. It takes the values between +1 and -1, depicting the strength and direction of the linear relationship where +1 is perfect positive correlation and -1 is for perfect negative correlation.

   Generally a high positive correlation means when variable1 increases, variable2 also usually increases and vice versa. A high negative correlation means when variable1 increases, variable2 usually decreases and vice versa. Correlation of low magnitude means when variable1 increasing or decreasing has little to no bearing on the direction variable2 moves and vice versa.

   Like in assignment, **atemp** & **temp** variables are highly correlated as there correlation is 0.99 which is almost 1



Formula

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

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? (3 marks)**

   Answer : Scaling is a step of data pre-processing which is applied to independent variables in a way that brings it within a range of values. It is performed to ensure that all variables are on a similar scale and have comparable importance during model training. Scaling is necessary to avoid variable bias, comparing coefficients.

   Normalised or Min Max Scaling brings all the data in the range of 0 and 1.

   Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

   In the assignment, I have used Normalised / MinMax Scaling

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

   Answer : VIF is a measure of collinearity between predictor variables. The formula is 1/(1-Rsquared) where Rsquared refers to the model of the other predictor against the variable in question. If the variable in question is almost completely explained by the other predictors, it will have an Rsquared close to 1, which means the denominator will be close to zero, leading to an infinite VIF score.

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

   Answer: A Q-Q Plot or a Quantile-Quantile plot is a probability plot, which measures the probability of two probability distributions graphically by plotting their quantities against each other. A quantile mean the fraction (or percent) of points below the given value, ie, the 0.4 (or 40%) quantile is the point at which 40% percent of the data fall below and 60% fall above that value.

   Q-Q Plot helps in Linear Regression in the scenario of training and testing dataset, as Q-Q plot helps verify the that both testing and training datasets are from populations from same distributions.

-- By Ruchir Dhingra