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?**
    a. Categorical Variables and their effects:
        i. **Season**: Total number of bike users are maximum in Fall and minimum in Spring.
        ii. **Year:** Total bike users increased significantly by more than 30% in 2019
        iii. **Month:** Bike users are significantly less in Janauary, whereas it increases more than 50% in March, after that the total bike users are almost same till October, after that there is a steep decline.
        iv. **Holiday:** People use marginally less bike on holidays.
        v. **Weather:** The number of bike users decreases by more than 40% when it's snow fall in comparison to when the weather is clear.

2. **Why is it important to use drop_first=True during dummy variable creation?**
    a. It's important because doing so will drop the first category while creating the dummy variable and thereby reducing the number of columns by one.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    a. Pair plot reveals that "Feel Temperature" has the highest correlation with the target variable.

4. **How did you validate the assumption of Linear Regression after building the model on the training set?**
    a. Linear relationship by looking at the Scatter plot.
    b. By visualizing the plots between predicted values and test set.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

a) It's a Supervised learning model, used when there is a linear relationship between independent and target variables.

b) Assumptions of a linear regression:

i) Linear relationship.
ii) Multivariate normality.
iii) No or little multicollinearity.
iv) No auto-correlation.
v) Homoscedasticity.

c) Equation: Y = $\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$

(1) Y = Target Variable
(2) X1, X2,…, Xp = Independent Variables or Features
(3) $\beta 0$ = Intercept or Constant
(4) βp = Coefficients
(5) $\epsilon =$ Error Term

2) Explain the Anscombe's quartet in detail.

a) It has four data sets, all having identical descriptive statistics but different distributions. Each data sets contains 11 data points.

b) It demonstrates the importance of charting the data for analysis because seemingly similar data in a table can be actually quite different in terms of distributions.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Source: https://builtin.com/data-science/anscombes-quartet

3) **What is Pearson's R?**
   a) It is known by many different names, such as:
      i) Pearson correlation coefficient (PCC)
      ii) Pearson product-moment correlation coefficient (PPMCC)
      iii) Bivariate Correlation.
   b) It measures the linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.
   c) The Pearson's correlation coefficient varies between -1 and +1 where:
      i) r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
      ii) r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
      iii) r = 0 means there is no linear association
      iv) r > 0 < 5 means there is a weak association
      v) r > 5 < 8 means there is a moderate association
      vi) 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?**
   a) Scaling: It is a data pre-processing step used to normalise data within a specific range by applying it to independent variables.
   b) It helps in accelerating calculations in algorithms. It easy for a model to learn and understand the problem when the data fed to the model is scaled.
   c) Normalized Scaling: It brings all of the data in the range of 0 and 1.
   d) Standardized Scaling: Standardization 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).

5) **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   a) A Variance Inflation Factor (VIF) index measure of how much collinearity increases the variance of an estimated regression coefficient.
   b) When VIF is infinite, it means that there is a perfect correlation between two independent variables.

6) **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   a) Using the Quantile-Quantile (Q-Q) plot, we can visually determine whether a set of data is likely to have originated from a theoretical distribution like the Normal, Exponential, or Uniform distribution.

b) This is useful when performing a linear regression because it allows us to verify using a Q-Q plot that the training and test data sets are from populations with similar distributions.
c) Steps to make a Q-Q plot:
    i) Order the items from smallest to largest.
    ii) Draw a normal distribution curve.
    iii) Divide the curve into n+1 segments. Here, n= number of items.
    iv) Find the z-value (cut-off point) for each segment.
    v) Plot your data set values against your normal distribution cut-off points.