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)

- Bike demand in 2019 is higher than 2018.
- Bike demand is high on fall season
- Bike demand high between the month from April to October
- Almost constant demand on weekdays
- Demand is slightly high on working day
- Demand is high on Clear, Few clouds, Partly cloudy, Partly cloudy weather condition

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

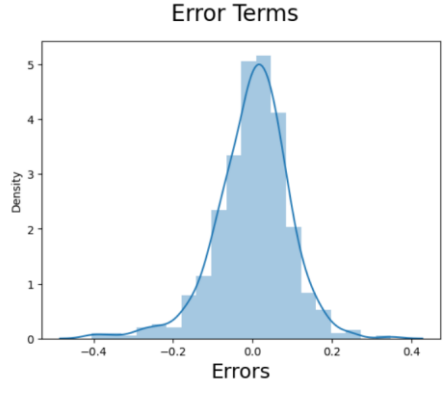
While creating dummy variables, number of columns or variables should be considered as n-1. "**drop_first=True"** used to drop first dummy variable column.

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

Temperature (temp) or feeling temperature(atemp) shows highest correlation with respect to count of total rental bikes including both casual and registered(cnt).

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

It can be validated by residual analysis of train data. It can be obtained by calculating the error term and viewed through distplot graph. Error terms are normally distributed and mean in the distribution error is zero. Refer below image.



As well verifying the dependent variable spread between test and predicted values of test helps to validate the assumptions.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Below are the three variables identified from coefficient values as we have performed scaling.
1. Temperature (temp) or feeling temperature (atemp)
2. Weather condition 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (Negatively influencing)
3. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression algorithm is part of supervised learning method in machine learning. This will find linear equation using the historical data set that best describes the linear relationship between independent variables(predictor) and a dependent variable. This is achieved by fitting a line to the data using least squares and the output variable to be predicted is a continuous variable.

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by minimizing the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:

1. Differentiation
2. Gradient descent method

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

RSS: Residual Sum of Squares

TSS: Total Sum of Squares

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: When the number of independent variables is 1
- Multiple linear regression: When the number of independent variables is more than 1

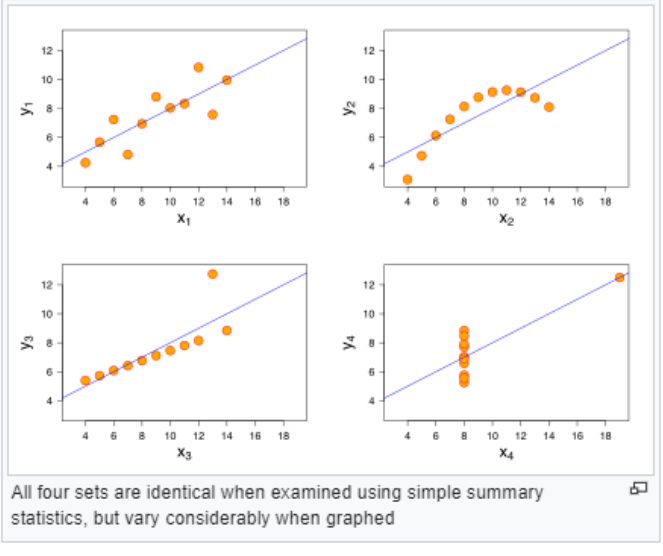
Assumptions of simple linear regression model.

1. The target variable and the input variables are linearly dependent.
2. The error terms are normally distributed.
3. The mean in the distribution of the error terms is zero.
4. The error terms have constant variance. It should be Homoscedasticity.
5. There is no high correlation between the independent variables. That means no multicollinearity.

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

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



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

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

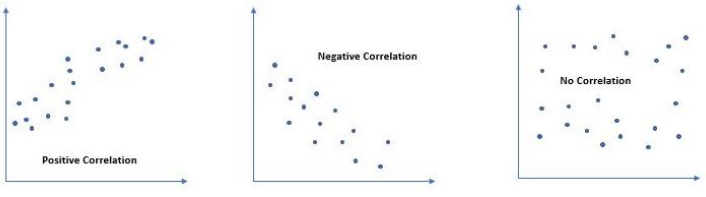
Pearson's Correlation Coefficient, often denoted as *r*, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.



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

scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values. Gradient descent as an optimization technique requires the scaled data which ensure that the gradient descent moves smoothly towards minima for all the features.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are two major methods to scale the variables, i.e. standardization and MinMax scaling.

- Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.

$$x = \frac{x - mean(x)}{sd(x)}$$

- MinMax (normalized) scaling, on the other hand, brings all of the data in the range of 0 and 1.

$$\frac{x - min(x)}{max(x) - min(x)}$$

Normalization	**Standardization**
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points

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

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. VIF measures the number of inflated variances caused by multicollinearity.
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.

The Quantile-Quantile plot is used for the following purpose in linear regression:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.