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?

   Below are the inferences based on categorical variables and the target variable

   1. Bike rentals are more during Summer and Fall as the median value of both seasons are higher compared to spring and Winter.
   2. Bike rentals are higher in 2019.
   3. The number of bike rentals are higher from the month of Mar to Oct.
   4. Significant use of bike rentals weekdays and working days.
   5. Bike rentals are at its higher when the weather situation is Clear/Partly cloudy.

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

   It is essential to use **drop_first=True** because this helps us to create k-1 dummies out of k categorical levels by removing the first level during dummy variable creation and helps to fit a better model.

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

   Highest correlation is between temp and atemp - 0.99

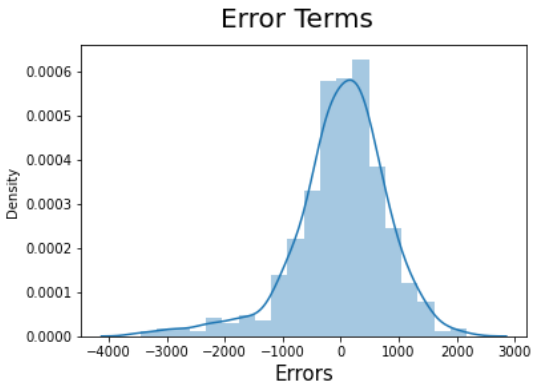
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   The major assumptions of Linear Regression after building the model are zero mean, independent and normally distributed error terms with constant variance.

   This is validated by finding **Residual given by predicted dependent variable from actual dependent variable**

res = ytrain - ytrain_pred

where the res is visualized as below with zero mean and normally distributed



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

   The below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

   1. yr_2019
   2. atemp
   3. season_winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression algorithm is the process of estimating relationship between (dependent and one or multiple independent Variables/Predictors). Linear Regression is represented by Equation of a linear line:

   - $Y = \beta_0 + \beta_1 X$
     - Y – Dependent variable
       X– Independent variable
     - where the β_1 is the slope or gradient and β_0 is the intercept

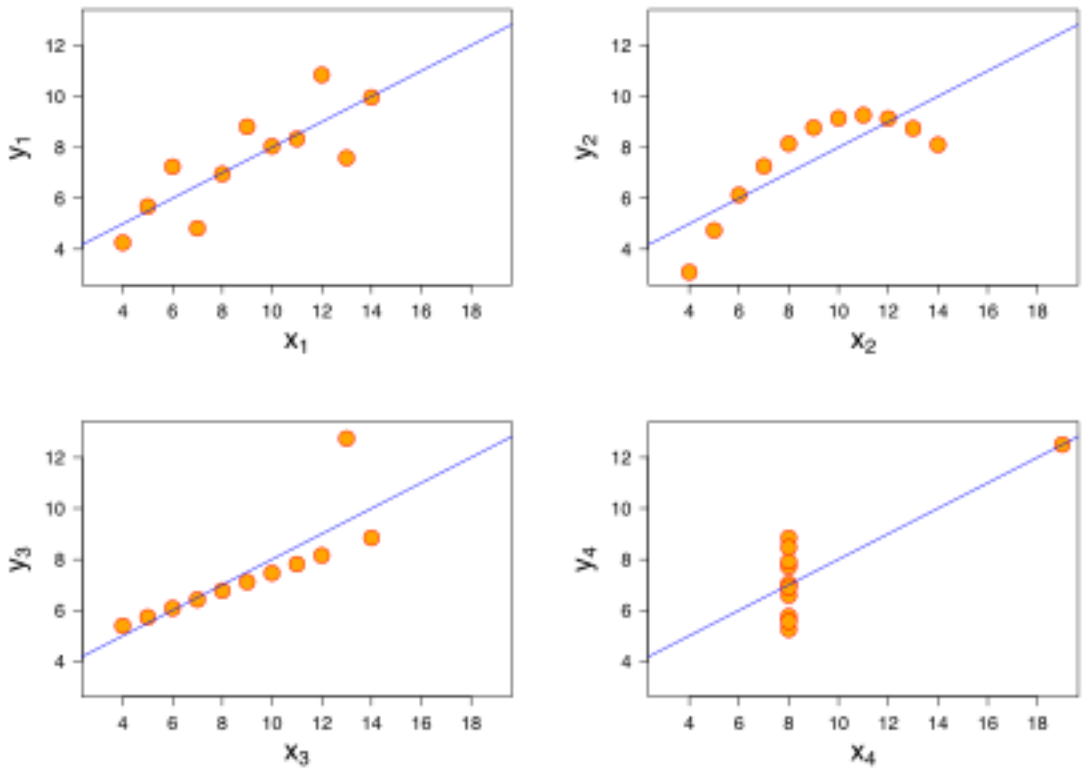
Linear regression models are classified into two types depending upon the number of independent variables:

- **Simple linear regression,** when the number of independent variables is 1.
- **Multiple linear regression,** when the number of independent variables is more than 1.

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by minimising the cost function – RSS (Residual sum of squares) using the ordinary least squares method.
Where the residual is calculated by subtracting predicted value of dependent variable from actual value of dependent variable.

The strength of a linear regression model is mainly explained by R^2 and RSE.

2. Explain the Anscombe's quartet in detail.
   Anscombe's quartet constructed by statistician Francis Anscombe and consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when visualized.
   Anscombe's quartet explains the importance of visualising data before analysing and also understanding effect of outliers and other influential observations on statistical properties.



1. The first scatter plot (top left) appears to be a simple linear relationship
2. The second graph (top right) is not distributed normally and not a non-linear relationship.
3. In the third graph (bottom left), the distribution is linear except one outlier.
4. Finally, the fourth graph (bottom right) shows an example when one high-leverage point indicating to produce high correlation coefficient.

3. What is Pearson's R?

   Pearson's R is also known as Pearson's correlation coefficient is used to find linear relationship between two continuous variables.

   It tells us about the magnitude of correlation, or association and the direction of the relationship.

When R = -1, the data is perfectly linear with a negative slope and R=1, the data is perfectly linear with a positive slope. R = 0means there is a no correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Feature Scaling is one of the most important steps during data preparation for ML. Scaling means normalizing the features in the same standing so that one feature with a significant number does not affect the model.

   Scaling directly affects the coefficients and performing scaling can help us build a strong model.

   Normalization and Standardized Scaling are common method of Feature Scaling.

   During Normalization, the data points are bound between two numbers, say from 0 to 1, whereas during standardization scaling, data are converted to have zero mean with Variance=1 thus making the data unitless.

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

   An infinite VIF indicate that there is perfect correlation between two independent variables.

   VIF is given by 1/(1-R2)

   For a perfect correlation R2 is 1. Thus, this makes VIF value infinite.

   This indicate multicollinearity and one of the independent variables should be dropped.

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

   Quantile-Quantile(Q-Q) plot is a graphical tool used to determine if two data sets belong to some theoretical distributions such as normal or uniform distribution.

   Q-Q plot is obtained by plotting the quantile of the first dataset vs the quantiles of the second dataset.

   In the case of Linear regression, Q-Q plots helps us to identify if the test and train datasets come from population with common distribution.