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)**

   The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables. spring, winter falss under season category and have been dummy encoded. Light Snow and (Mist + Cloudy)falls under weathersit category and have been dummy encoded. Similarly, december,november ,september variables fall under mnth category and have been dummy encoded. We can infer from above image that these variables are statistically significant and explain the variance in model very well.

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

   drop_first=True is important to use, because it helps in reducing the extra column created during dummy variable creation.

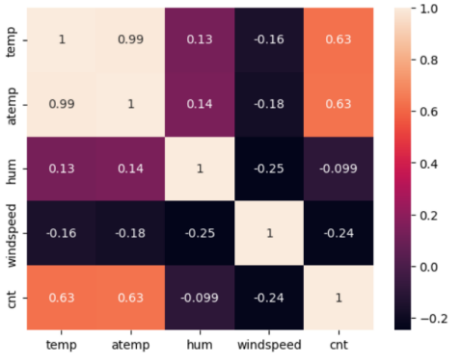
Example:

1. Let's say we have 3 types of values in Categorical column(Gender: Male, Female, Others) and we want to create dummy variable for this columns. If one variable is not Male and Others, then It is obvious Female. So we do not need 3rd variable to identify the Female.
2. if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.
3. Advantages:

- Number of columns can be reduced
- Reduces the multicolinearity created among dummy variables.

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

   Atemp and temp are having highest correlation with target variable cnt. As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.63. And correlation coefficient between temp and cnt is 0.63.

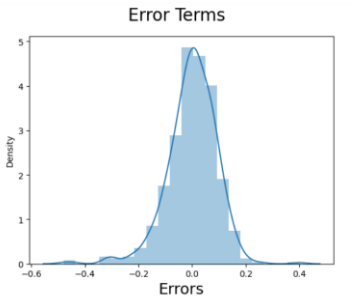


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- Linear relationship between predictor variables and target variable: This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0. 828 and adjusted R-Squared value on training set is 0. 824. This means that variance in data is being explained by all these predictor variables.

- Residual Analysis: Error terms are normally distributed. histogram of the error terms are as below



residuals are normally distribution with a mean 0.

- Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

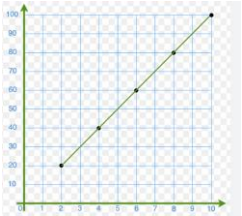
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Top 3 features significantly contributing towards demand of shared bikes are:

- `temp (coefficient: 0. 0.3870)`
- `Light Snow (coefficient:-0.2679`
- `year (coef: 0.2297)`

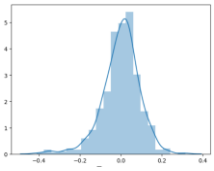
General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - Linear regression is a supervised learning technique conducted on continuous variable.
   - It is a method of finding the best straight-line fitting to the given data between independent and dependent variables.
   - The assumptions of linear regression are:
     a. It is assumed that there is a linear relationship between the dependent and independent variables.



b. Assumptions about the residuals:

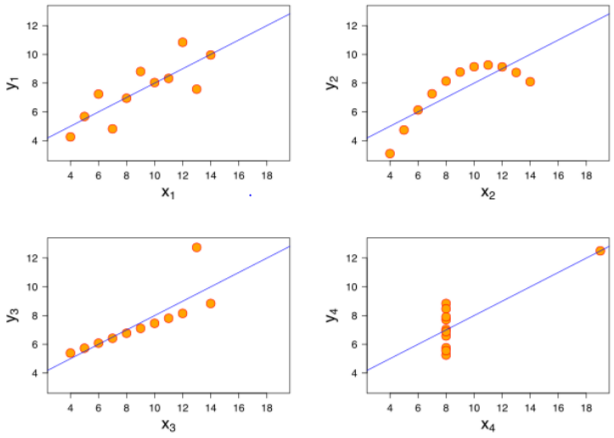
1. It is assumed that the error terms, ε(i), are normally distributed.



2. the error terms are normally distributed around zero.
3. the residual terms have the constant variance(homoscedasticity)
4. It is assumed that the residual terms are independent of each other(no Multicolenearity)

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

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



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

  1) The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

  2) The second graph is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

  3) In the third graph ,the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

  4) the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)
    - Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data.
    - It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
    - As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.
      1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
      2) A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in perfect correlation with speed.
      3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related. The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

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

- Scaling is a method used to normalize the range of independent variables or features of data.
- For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature.
- Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
- Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.
  Normalization: Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

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

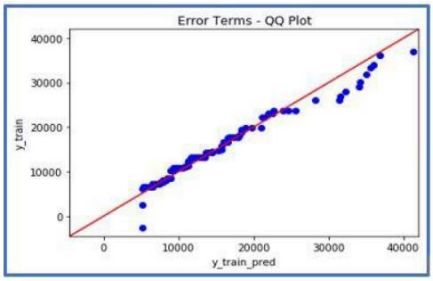
  Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.
  Standardization: standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

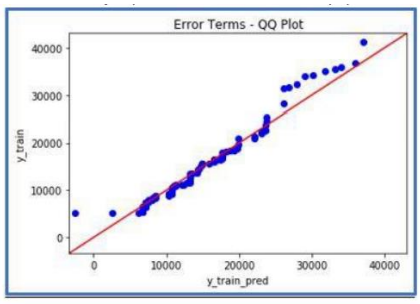
  Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

4. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
    - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
    - To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
    - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

5. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
1) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis
statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.