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?

Ans. Following categorical variables present in the dataset and I have kept for analysis are:

- Season: During Spring boom bike has the lowest business where as during fall season it has got strong trend wrt to target variable.
- Year:2019 is showing strong wrt to the target variable
- Weathersit: During clear weather boom bike has highest business
- Weekday: All weekday (from saturday till friday) has almost same trend wrt target variable 'cnt'
- Holiday: Boom bike has more business during working day.

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

Ans. I have used one-hot encoding for creation of dummy variables. They are created to cover all the values of a categorical variable. Dummy variable acquire value 1 which denotes presence and 0 for absence of the particular value of the categorical variable. This means if the category variable has 3 categories, there will be 3 dummy variables.
The drop_first = True is used while creating dummy variables to drop the reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in all the columns of a single row for all the other dummy variables of a particular category.

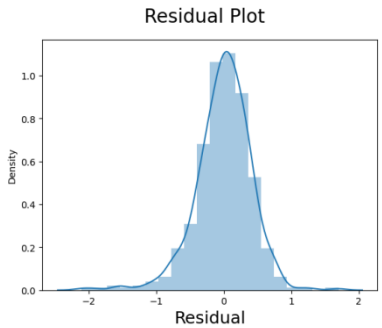
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Highest correlation coefficient value(0.63) is for variable ‘temp’ and ‘atemp’ with the target variable.

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

Ans. Assumptions of Linear regression are:

- Linear relationship between X and Y – This can be established using scatter plot between y predicted and y actual
- Error terms are normally distributed (not X, Y) Frequency plot of errors- is a normal distribution, and it is centered around 0.



- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

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

Ans. The top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- **Temp**: Coefficient is 0.43
- **year** (2019): Coefficient is 1.04, hence with every unit increase in year demand increases.
- **Winter** : Coefficient is 0.34

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised machine learning algorithm used to predict a continuous numerical outcome based on one or more input features. It assumes that there is a straight-line relationship between the input variables and the target variable. The main objective of linear regression is to find the best line that fits the data points and minimizes the difference between the predicted and actual values of the target variable. Linear Regression is only possible when there is a linear relation found between different variables (also referred as features) , if there is no linear relation is visible on a plotted scatter chart of different variables , the model will not work as expected and the accuracy of model will get down drastically. Linear regression assumes linearity, independence of errors, constant variance (homoscedasticity), normality of errors, and absence of multicollinearity between independent variables. These assumptions should be verified and validated.

Linear regression equation is referred as Y = β0 + β1X1 + β2 x X2 + β3 x X3 ... βn x βn where n is the number of independent variable. This equation is basically the best fit line equation that has the least distance between actual and predicted values of the target variable. While establishing , this equation and during the process of Feature selection , it is very important to make sure that the selected features for the model creation should not have multicollinearity issue between them because if they have multicollinearity , the weightage of similar feature will be taken into account making our model accuracy not as expected. In simple linear regression, we have only one input feature (X), while in multiple linear regression, we can have multiple input features (X1, X2, X3, ...). The linear regression model assumes that the relationship between the input features and the target variable can be represented by an equation.

For example, in simple linear regression, the equation is y = b0 + b1 * X, where b0 is the starting point (y-intercept), b1 is the slope, and X represents the input feature. For multiple linear regression, the equation becomes y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn, where each X represents a different input feature. During model training, we try to find the best values for the coefficients (b0, b1, b2, ..., bn) in the equation. These coefficients determine the shape and position of the line. We use an optimization algorithm, such as Ordinary Least Squares (OLS), to minimize the difference between the predicted and actual values of the target variable.
After training the model, we evaluate its performance. We use evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-

squared to measure how well the model predicts the target variable. These metrics help us understand the accuracy of the predictions and how well the model fits the data.

2. Explain the Anscombe's quartet in detail.

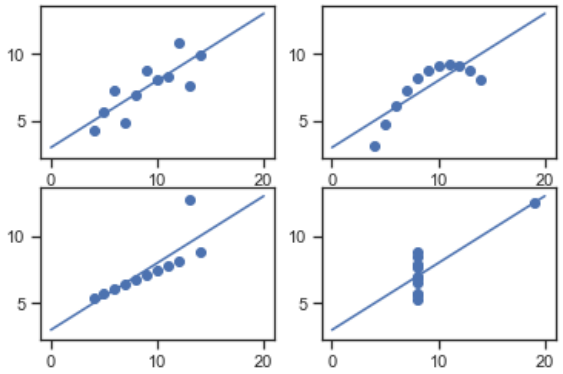
Ans. Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics but exhibit starkly different properties when plotted and analyzed. These datasets were originally created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and to challenge the reliance on summary statistics alone. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set,

- Average Value of x = 9
- Average Value of y = 7.50
- Variance of x = 11
- Variance of y =4.12
- Correlation Coefficient = 0.816
- Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



Graphical Representation of Anscombe's Quartet:
Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3. What is Pearson's R?

Ans. Pearson's correlation coefficient, often denoted as Pearson's R or simply R, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by the British statistician Karl Pearson and is widely used in various fields, including statistics, social sciences, and data analysis.
Pearson's R ranges from -1 to 1, where:
• A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
• A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
• A value of 0 indicates no linear relationship between the variables.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

The calculation of Pearson's R involves the following steps:
1. Standardizing the variables: Each variable is transformed by subtracting its mean and dividing by its standard deviation. This step ensures that the variables are on a comparable scale.
2. Computing the covariance: The product of the standardized values of the two variables is calculated for each data point, and the average of these products is computed. This measure is known as the covariance.
3. Computing the standard deviations: The standard deviations of the two variables are calculated separately.
4. Calculating Pearson's R: The covariance is divided by the product of the standard deviations of the two variables to obtain the correlation coefficient, which is Pearson's R.

Pearson's correlation coefficient is widely used because it is easy to interpret and provides a measure of the linear association between variables. However, it is important to note that Pearson's R only measures the strength and direction of linear relationships and may not capture other types of relationships, such as non-linear or curvilinear associations. Additionally, Pearson's R is sensitive to outliers, and its validity can be affected by violations of assumptions, such as nonnormality or heteroscedasticity.

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

Ans. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. 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.

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

Ans. In the context of linear regression analysis, the Variance Inflation Factor (VIF) is a measure that quantifies multicollinearity, which is the correlation or high interdependency among predictor variables. VIF is used to assess the extent to which the variance of the estimated regression coefficients is inflated due to multicollinearity.

The formula for calculating the VIF of a predictor variable is as follows: VIF = 1 / (1 - R^2) where R^2 represents the coefficient of determination obtained by regressing the predictor variable against all other predictor variables. The VIF value provides insight into how much the variance of a particular predictor's coefficient is inflated due to multicollinearity.

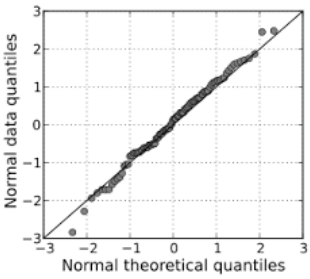
In some cases, the VIF value can be infinite. This occurs when the coefficient of determination (R^2) for a particular predictor variable is equal to 1. A perfect correlation between a predictor variable and other predictor variables can lead to an R^2 of 1, resulting in an infinite VIF. There are a few scenarios that can cause a predictor variable to have an R^2 of 1 and an infinite VIF:

1. Perfect Linear Relationship: The predictor variable is perfectly linearly related to one or more other predictor variables in the model. In this case, the VIF becomes infinite because the variance of the coefficient estimate cannot be determined separately from the other correlated variables.
2. Redundant Predictor: The predictor variable is a linear combination or a duplicate of another predictor variable(s) in the model. When two or more predictor variables provide the same information, it results in perfect multicollinearity and leads to an infinite VIF.

Having an infinite VIF suggests severe multicollinearity, indicating that the predictor variable is perfectly predictable from the other variables in the model. This can pose challenges in interpreting the model and estimating the effect of individual predictors. In such cases, it is necessary to address multicollinearity by identifying and resolving the high interdependency among the predictor variables. Techniques such as removing redundant variables, transforming variables, or using dimensionality reduction methods can help mitigate multicollinearity issues.

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

Ans. Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below:



- Interpretations
    - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
    - Y values < X values: If y-values quantiles are lower than x-values quantiles.
    - X values < Y values: If x-values quantiles are lower than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.