## 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?

Answer:

The analysis has been done on categorical variables with boxplot

So, from the analysis of the categorical variables from the dataset, and their effect on dependent varible the insights are:

1.Season 'fall' seems to have highest demand for rental bikes.

2.And the demand is increased from 2018 to 2019 for each season.

3.Demand is growing 'jan' to 'jun' and highest in 'september'.  4.'clear' weather seems to have higher demand.

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

Answer:

Dropping variable is one of the methods of Multicolinearity.

So, drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. It reduces the correlation created among the dummy variables.

so, generally N-1 dummy variables can be used to describe a categorical variable with N levels.

If we have n categorical variables we need n-1 dummy variables.

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

Answer:

Looking at the pair-plot among the numerical variables, 'temp' variable which is temperature has highest correlation with the target variable.

It is positively correlated with target variable.

And target variable is linearly increasing with temp indicating linear relationship

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

Answer:

I have validated the assumptions of Linear Regression after building the model on the training set based on some assumptions:

1. Error terms should be normally distributed.
2. Multi colinearity check
3. Linear relationship validation.
4. There should be no visible pattern in residual values.
5. No auto correlation.

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

The top 3 features are:

1.temp

2.winter

3.sep

## General Subjective Questions

### 1. Explain the linear regression algorithm in detail.

### Answer:

Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the input features and the target variable. The algorithm aims to find the best-fitting line, known as the regression line or the best-fit line, that minimizes the distance between the predicted values and the actual values.

Here's a detailed explanation of the linear regression algorithm:

Problem Setup: Suppose we have a dataset consisting of m instances or examples, where each instance is characterized by n input features and a corresponding target variable. We represent the input features as X and the target variable as y. The goal is to find a linear function that maps the input features to the target variable.

Model Representation: In linear regression, we assume a linear relationship between the input features and the target variable. The linear function is represented as:

y = β0 + β1*x1 + β2*x2 + ... + βn*xn

Here, y is the target variable, β0 is the intercept term (y-axis intercept), β1, β2, ..., βn are the coefficients or weights associated with each input feature x1, x2, ..., xn.

Cost Function: To estimate the best-fit line, we need to define a cost function that quantifies the difference between the predicted values and the actual values. The most commonly used cost function for linear regression is the Mean Squared Error (MSE). The MSE is calculated as:

MSE = (1/m) * Σ(yi - ŷi)^2

Where yi represents the actual target value for the i-th instance, and ŷi represents the predicted target value for the i-th instance. The goal is to minimize this cost function.

Parameter Estimation: To estimate the coefficients β0, β1, β2, ..., βn, we use a technique called Ordinary Least Squares (OLS). The OLS method finds the values of the coefficients that minimize the MSE. The formulas for estimating the coefficients are as follows:

β̂= (X^T * X)^(-1) * X^T * y

Where X is the design matrix, which consists of the input features and an additional column of 1s for the intercept term, X^T denotes the transpose of X, and y is the target variable.

Model Evaluation: Once the coefficients are estimated, the linear regression model is ready for predictions. To evaluate the performance of the model, various metrics can be used, such as the coefficient of determination (R-squared), root mean squared error (RMSE), mean absolute error (MAE), etc.

Making Predictions: Given a new set of input features, we can predict the corresponding target variable by substituting the values into the linear function:

ŷ = β0 + β1*x1 + β2*x2 + ... + βn̂*xn

Here, ŷ represents the predicted value of the target variable.

Linear regression is a simple yet powerful algorithm used in various fields, such as economics, finance, social sciences, and machine learning. It provides interpretable results and serves as a building block for more complex regression techniques.

## 2. Explain the Anscombe's quartet in detail.

### Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between

x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

## 3. What is Pearson's R?

### Answer

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

## Answer

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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

### Answer

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.