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:

   Categorical columns have a major effect on the value of dependent variable. From our analysis we found that 9 out of the 10 significant features in the model are categorical features.
   We can infer the following about the effect of categorical variables:

   - **Year:** There's a significant growth in shared bike's demand in the year 2019.
     This is due to the fact that BoomBikes must have increased their number and availability of shared bikes in 2019 in comparison to the year 2018.
   - **Holiday:** Presence of holiday is negatively impacting the demand of shared bikes. As during holidays, people tend to stay at home mostly and hence the demand is negatively affected.
   - **Season:** Winters are positively impacting, whereas spring is negatively affecting the demand.
     As we know, USA has around 50 states, so the winters might happen during different months for different states.
     Spring has negative impact as the weather is warmer, so people would prefer air-conditioned vehicles.
   - **Month:** During the months of July, November and December demand is negatively affected.
     This is because November and December are the holiday months, so they are negatively affecting the demands. During July, it's rainy in USA hence people prefer car over bikes.
   - **Weather:** Snowy, Rainy, Cloudy, and Misty weather is having a negative effect on the demands.
     As during these weather conditions people would prefer to travel in car rather than bikes, as they won't have a roof to protect them in case of bikes or two-wheelers. That's why these weather conditions are negatively affecting the demand of shared bikes.

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

   ANSWER:

   During dummy variable creation it is important to use **drop_first=True** because:

   1) It helps in reducing the extra column created during dummy variable creation.
   2) It reduces the correlations created among dummy variables.
   3) If we won't drop the extra column that will add unnecessary complexity into our model.

   Though, it is not mandatory to use drop_first argument. We can manually drop the column any one of the created columns. Eg. One can drop the column which is irrelevant to business understanding.

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

   ANSWER:
   Among the numerical variables, **temp (temperature)** has the highest correlation with the target variable (i.e., **cnt**). The correlation coefficient between temperature and the target variables is **0.63**.

   We can ignore the **casual** and **registered,** as the sum of these two columns equates to the value of **cnt**. So, they together are able to match with the target and hence should not be considered, as they don't add any value to the analysis.

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

   ANSWER:
   I validated the assumptions of Linear Regression after building the model using the following:
   1) **For validating normal distribution of errors,** I plotted a distribution plot of the residuals and checked whether the distribution was normal.
   2) **For validating mean = 0,** I checked in the distribution plot whether the distribution was normal with mean = 0.
   3) **For validating homoscedasticity,** I plotted a scatterplot between the residuals and target values (i.e., total count of bike shares) to check whether the residuals are uniformly distributed around y=0. This tells us that error terms have constant variance i.e., they are homoscedastic in nature.
   4) **For validating error independence,** I checked in the scatterplot whether they the errors increase or decrease with the increase of the target variable.

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

   ANSWER:
   Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

   I. **Temperature** – Temperature is positively affecting the demands of the shared bikes (having coefficient as *0.4736*).

   II. **Light Snow/ Rain** – Presence of Light Snow or Rain (or Thunderstorm) type of weather has negative impact of the demands of the shared bikes (having coefficient as *-0.3178*).

   III. **Year 2019** – The demand of shared bikes has increased during 2019 in comparison to the year 2018 (having coefficient as *0.2556*).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

ANSWER:

Linear Regression is a type of supervised learning algorithm which is used to define a linear model, i.e., a model that assumes a linear relationship between the input variables (X) and the single output variable (y). More specifically, that the dependent variable (y) can be calculated from a linear combination of the input variables (X). In Linear Regression it is assumed that the dependent variable (y) can be calculated as a linear combination of the input variables (X). It also assumes that error terms are normally distributed with their mean at 0. Moreover, residuals are assumed to be independent of each other and have a constant variance i.e., homoscedastic.

Linear Regression is of two types:

i. **Simple Linear Regression:** Simple Linear Regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable.
E.g., We want to predict the salary of an employee by knowing his/her years of experience.
For this, we have the *salary* as the dependent variable and *years of experience* as the predictor. Hence, we will use Simple Linear Regression here.

Y = b0 + b1*X

ii. **Multiple Linear Regression:** Multiple Linear Regression is useful for finding relationship between a continuous variable and multiple other variables. One is the response or dependent variable and others are the predictor variables or independent variables.
E.g. We want to predict blood pressure (the dependent variable) from independent variables such as sex, height, weight, age, and hours of exercise per week. Hence, we will use Multiple Linear Regression for this problem.

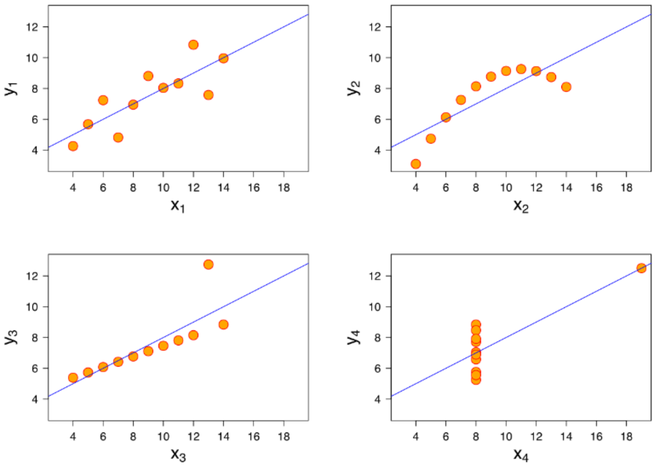
Y = b0 + b1*X1 + b2X2 + b3X3 + ………bnXn

2. Explain the Anscombe's quartet in detail.

ANSWER:

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 Linear Regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. It also talks about the sensitivity of Linear Regression model with the outliers in the data.

Anscombe’s Quartet has the following four scatterplots:



We can observe the following from these four plots:

- **Plot 1:** this fits the linear regression model pretty well.
- **Plot 2:** this could not fit linear regression model on the data quite well as the data is non-linear.
- **Plot 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model
- **Plot 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

This tells us the importance of data visualisation and how regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Moreover, if the outliers not been present, we could have gotten a great line fitted through the data points. So, we should never ever run a regression without having a good look at our data.

3. What is Pearson’s R?

ANSWER:

Pearson’s R correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. Its formula is given by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

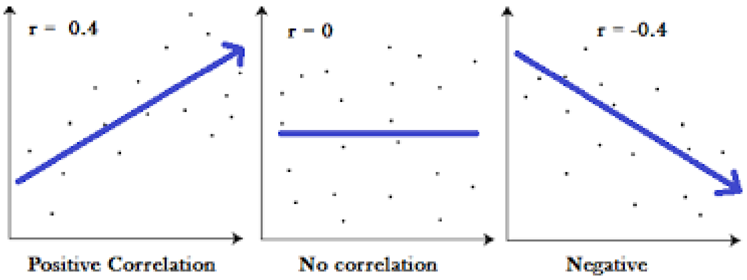
where,

r = Pearson’s R between x and y

x_i = value of x (for i^{th} observation)

y_i = value of y (for i^{th} observation)

Pearson's R correlation coefficient is designed to determine the correlation between linearly dependent continuous variables only and it might not be a good measure for if the relationship between the variables is non-linear. Pearson correlation coefficient (r) varies with the strength and the direction of the relationship. This can be seen in the below plots:



E.g., we can use the Pearson correlation to evaluate whether an increase in age leads to an increase in blood pressure.

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

ANSWER:

Scaling is the process of normalizing the features to a certain range such that the numerical features become comparable to each other. In other words, feature scaling is a method used to normalize the range of independent variables or features of data. Scaling is performed to:

1) It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
2) 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.
3) To make the Gradient Descent or the optimization algorithm to converge faster.

There are mainly two types of scaling, Normalized Scaling and Standardized Scaling. The difference between them is:

1) Normalized Scaling (Min-Max)

This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

2) Standardized Scaling (Mean=0 & Variance=1)

It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

ANSWER:

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).
Variance Inflation Factor (VIF) is given by:

$$\text{VIF} = \frac{1}{1 - R^2}$$

Infinite VIF shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

E.g., Let's suppose we are having three columns Height, Weight and BMI.
We know,

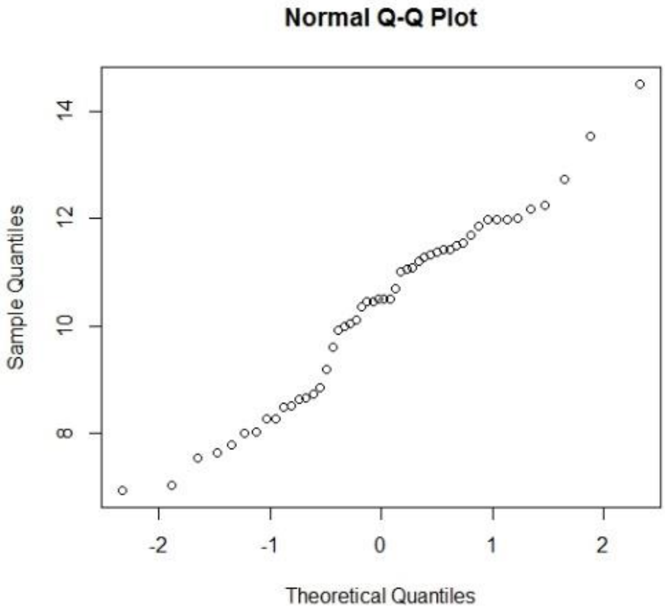
$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m}^2\text{)}}$$

BMI can be calculated using Height and Weight.
So, the VIF for BMI would be infinite (∞).

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

ANSWER:

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Use and Importance in Linear Regression:

1. This helps in scenarios 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.
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check if two data sets:

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

If we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It allows us to see if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.