Assignment-based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. There were 6 categorical variables in the dataset.
I used Box plot to study their effect on the dependent variable ('cnt').
The inference that we could derive are:

season: Most of the bike booking were happening in season 3(Fall) with a median of over 5000 booking (period of 2 years). This was followed by season 2(Summer) & season 4(Winter) respectively total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Most of the bike booking were happening in the months June, August, September and October with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Maximum of the bike booking were happening during weathersit 1(Clear) when weather is clear with a median of close to 5000 booking (period of 2 years). This was followed by weathersit 2(Mist) of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

yr: Maximum of the bike booking were happening during year 2019 with median 6000 approx. in compare to year 2018 with median less than 4000. This indicates, yearly biases the demand and booking of bikes keep increasing. So, yr is a good predictor for the dependent variable.

weekday: weekday variable shows total booking on all days of the week having their medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: workingday variable shows total booking happening with a median of close to 5000 booking (period of 2 years) no matter if it is working day or not. This indicates, workingday variable have no influence towards the predictor.

Q2. Why is it important to use ***drop_first=True*** during dummy variable creation?
Ans. ***drop_first=True*** is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans. There is linear relationship between *temp* and *atemp*. Both of the parameters cannot be used in the model due to multicollinearity. We will have to decide which parameters to keep based on understanding dataset and variables w.r.t other variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans. There are several assumptions that need to be validated after building a linear regression model on the training set

***Linearity*:** The first assumption of linear regression is that there is a linear relationship between the dependent variable and the independent variables. We can use scatterplots to visualize the relationship between the dependent variable and each independent variable.

Residual Analysis: The second assumption is that the residuals (the difference between the predicted and actual values) are normally distributed. You can use a histogram plot to check for normality.

Homoscedasticity: The third assumption is that the variance of the residuals is constant across all levels of the independent variables. You can use a scatterplot of the residuals against the predicted values to check for homoscedasticity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans. As per final model, the top 3 predictor variables that influences the bike booking are:

1. *Temperature (temp)*
2. *Year(yr)*
3. *Light_rain_snow*

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

Ans. *Linear regression* assumes a linear or straight-line relationship between the input variables (X) and the single output variable (y). More specifically, that output (y) can be calculated from a linear combination of the input variables (X). Linear regression is a statistical method that aims to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting straight line through a set of data points, which can then be used to make predictions about new data.

Here's a step-by-step overview of the linear regression algorithm:

Data Collection: Collect data on the dependent variable (also called the target variable or response variable) and the independent variables (also called predictors or features) from various sources.

Data Preparation: Prepare the data by cleaning it, eliminating missing values, and dealing with outliers. The data should also be normalized or standardized to ensure that the variables are on the same scale.

Split Data: Split the data into training and testing sets to evaluate the accuracy of the model. The training set is used to fit the model, and the testing set is used to evaluate the performance of the model.

Define the Model: In linear regression, the relationship between the dependent variable and the independent variables is defined by a linear equation of the form Y = b0 + b1X1 + b2X2 + ... + bn*Xn, where Y is the dependent variable, X1, X2, ... Xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that need to be estimated.

Estimate Coefficients: The coefficients in the linear equation are estimated by minimizing the sum of squared errors between the

predicted and actual values of the dependent variable. This is typically done using an optimization algorithm such as gradient descent.

Model Evaluation: The accuracy of the model is evaluated by comparing the predicted values of the dependent variable to the actual values in the testing set. This can be done using various metrics such as the mean squared error, mean absolute error, or R-squared.

Model Deployment: Once the model has been evaluated and found to be accurate, it can be deployed to make predictions on new data.

Q2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a set of four datasets that were introduced by the statistician Francis Anscombe in 1973. The datasets are often used to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

Each of the four datasets has the same mean, variance, correlation coefficient, and linear regression line, but they are very different in terms of their underlying structure and relationships between variables. This is illustrated by plotting the data and visualizing the relationships between variables.

Here's a brief overview of each of the four datasets in Anscombe's quartet:

Dataset I: This dataset is a simple linear relationship between X and Y, with a slope of 0.5 and an intercept of 3. It has no outliers or unusual points.

Dataset II: This dataset also has a linear relationship between X and Y, but with an outlier point that significantly affects the regression line. If

this point were removed, the linear relationship would be much stronger.

Dataset III: This dataset has a non-linear relationship between X and Y, with a quadratic curve that fits the data well. The regression line is not a good fit for this data and does not capture the true relationship between the variables.

Dataset IV: This dataset has an outlier point that completely dominates the regression line, resulting in a misleading summary of the relationship between X and Y. Without this point, the relationship between X and Y would be non-linear.

The *key takeaway* from Anscombe's quartet is that relying solely on summary statistics can be misleading and may not capture the true relationship between variables. It is important to visualize the data and understand the underlying structure and patterns before drawing conclusions or making predictions.

Q3. What is Pearson's R?

Ans. Pearson's R, also known as the *Pearson correlation coefficient* or simply the correlation coefficient, is a measure of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to 1, where -1 indicates a perfectly negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfectly positive linear relationship.

Pearson's R is calculated as the covariance between two variables divided by the product of their standard deviations. The formula for Pearson's R is:

$$r = (\Sigma ((x - \bar{x}) (y - \bar{y}))) / (\sqrt{} (\Sigma (x - \bar{x}) ^2) * \sqrt{} (\Sigma (y - \bar{y}) ^2))$$

where x and y are the two variables, $\bar{x}$ and $\bar{y}$ are their respective means, and Σ represents the sum of the values.

The Pearson correlation coefficient is widely used in statistics and machine learning to understand the relationship between variables, to identify patterns in data, and to make predictions. It is particularly useful when the relationship between variables is assumed to be linear. However, it should be noted that Pearson's R only measures the strength of the linear relationship between variables and may not capture non-linear relationships. Additionally, it can be influenced by outliers in the data and may not be appropriate for variables with non-normal distributions.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling 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 have to 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.

1. ***Normalization/Min-Max Scaling:*** It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

2. ***Standardization Scaling***: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. The *variance inflation factor (VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model.

It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

$$VIF = 1/1-R^2$$

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. This would mean that that standard error of this coefficient is inflated by a factor of 2. The standard error of the coefficient determines the confidence

interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

Q5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. *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.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small. for example, if we have N observations, then each step traverses 1/(N-1) of the data. So, we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).