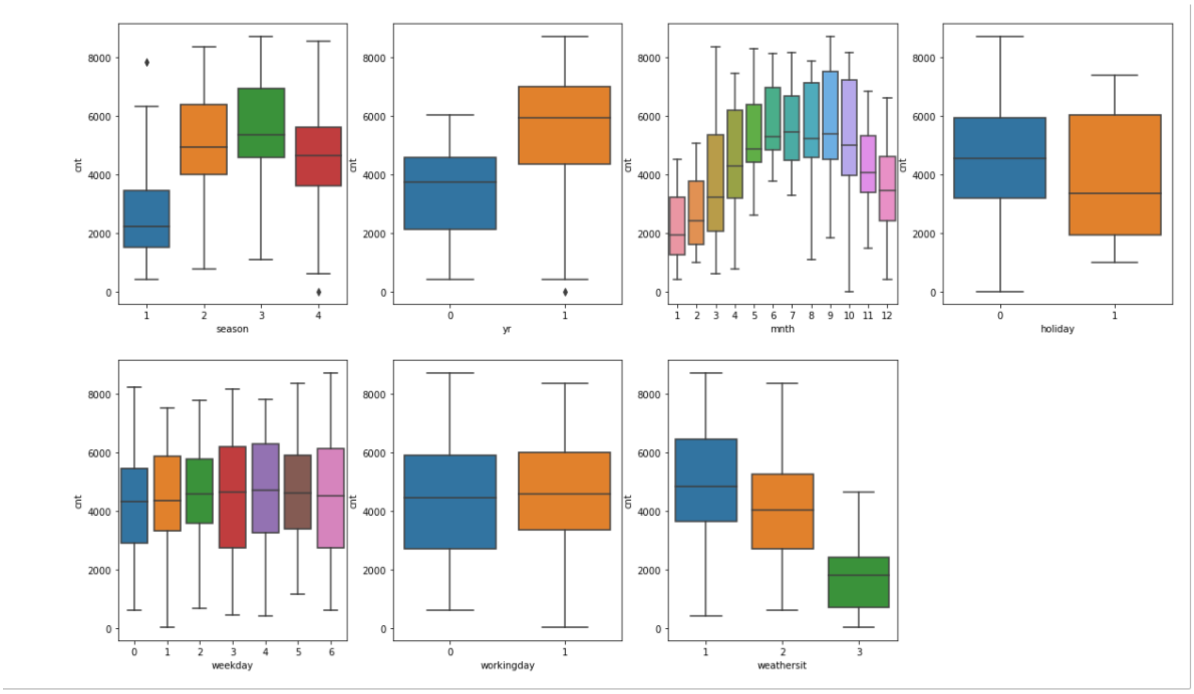
Assignment-based Subjective Questions

Ques1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



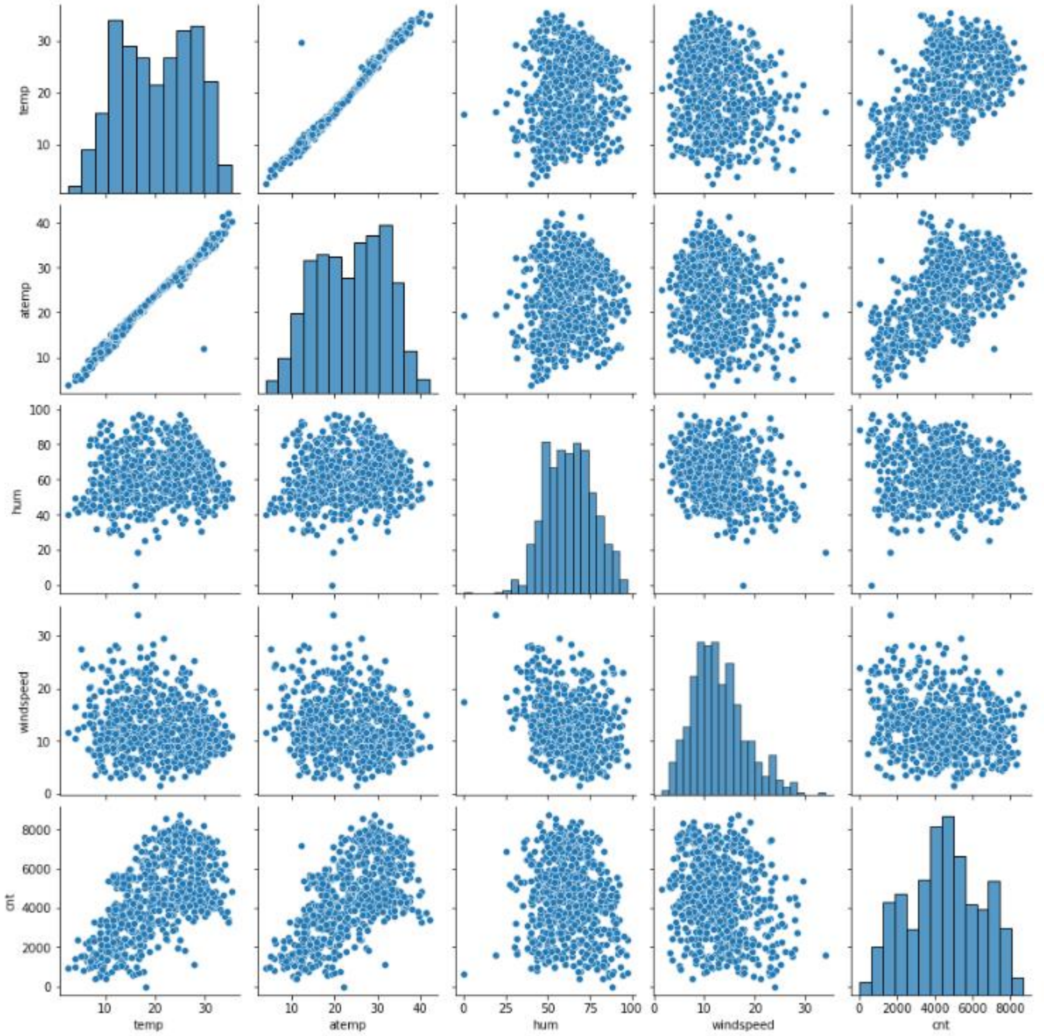
- For the variable season, we can clearly see that the category 3 : Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring
- The year 2019 had a higher count of users as compared to the year 2018
- The count of total users is in between 4000 to 6000 (~5500) during clear weather
- The count is highest in the month of September
- The count of users is less during the holidays

Ques2: Why is it important to use drop_first=True during dummy variable creation ?

It is important to use drop_first = True during dummy variable creation because if we do not drop the first column data will become redundant. It helps in reducing the extra column created during dummy variable creation. It reduces the correlation created among dummy variables.

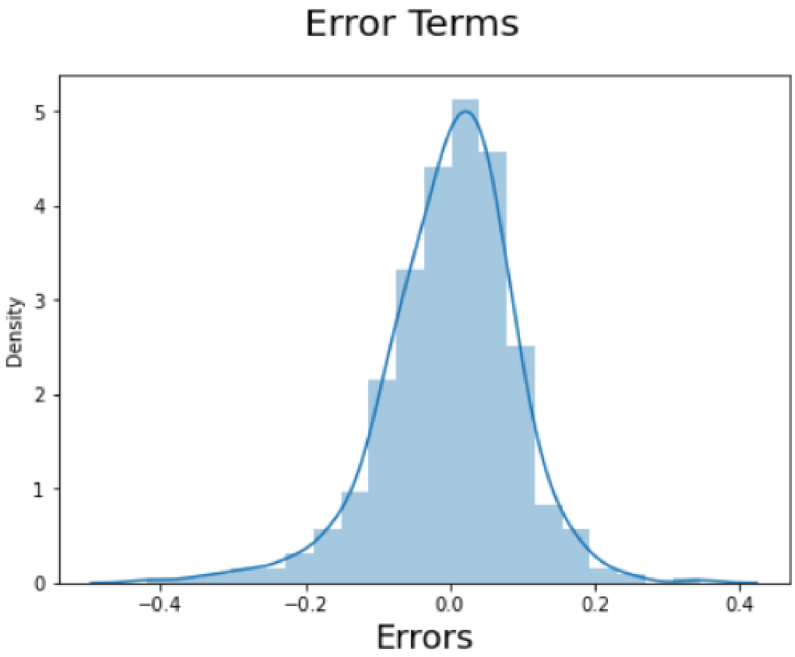
Ques3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable ?

Ans



Using the above pairplot it can be seen that , “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

Ques4: How did you validate the assumptions of Linear Regression after building the model on the training set?



1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

2. Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.

3. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

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

weathersit_Light Snow & Rain, Temp & year are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

General Subjective Questions

Ques1: Explain the linear regression algorithm in detail.

Linear regression is a ML algorithm used for predictive modeling. It is a supervised learning method that tries to establish a relationship between a dependent variable (target variable) and one or more independent variables (feature variables). The goal of linear regression is to create a linear equation that best predicts the value of the target variable based on the values of the feature variables.

Following steps involved in the linear regression algorithm:

Step 1: Data Preparation The first step in linear regression is to collect and preprocess the data. This includes cleaning the data, handling missing values, encoding categorical variables, scaling/normalizing features, and splitting the data into training and testing sets.

Step 2: Feature Selection Next, we need to select the relevant feature variables that will be used in the model. We can use techniques like correlation analysis, VIF and recursive feature elimination to identify the most important features.

Step 3: Model Formulation In this step, we formulate the linear regression model using the selected feature variables. Let's assume we have n feature variables x1, x2, ..., xn and a target variable y. The linear regression model can be represented as:

y = w0 + w1x1 + w2x2 + … + wn*xn + ε

where w0, w1, w2, ..., wn are coefficients that need to be learned, and ε is the error term representing the random variation in the target variable not explained by the feature variables.

Step 4: Parameters Estimation Now, we need to estimate the parameters (weights) w0, w1, w2, ..., wn that minimize the difference between the predicted values and actual values of the target variable. We use a cost function called Mean Squared Error (MSE) to measure this difference. MSE is calculated as:

MSE = (1/m) * Σ(y_actual - y_predicted)^2

where m is the number of samples in the training dataset, and y_actual and y_predicted are the actual and predicted values of the target variable, respectively. To minimize the MSE, we use an optimization algorithm such as gradient descent.

Step 5: Model Evaluation After training the model, we evaluate its performance using various metrics such as mean squared error (MSE), root mean squared error (RMSE), R-squared value, mean absolute error (MAE), etc. We can also use cross-validation techniques to ensure that our model generalizes well to new, unseen data.

Step 6: Model Deployment Finally, once we are satisfied with the model's performance, we can deploy it to make predictions on new data. We can use the trained model to forecast the target variable values for given input feature values.

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

Anscombe's Quartet is a set of four datasets that were created by Francis Anscombe in 1973 to demonstrate the importance of considering the causal relationships between variables when analyzing data. The datasets are identical in terms of their statistical properties (mean, variance, correlation, and regression line), yet they have very different visual appearances.

The four datasets are:

Y vs X: This dataset contains two variables, Y and X, where Y is the response variable and X is the predictor variable. The relationship between Y and X is a straight line, with a positive slope. The dataset has a mean of 10, a variance of 1, and a correlation coefficient of 0.816.

Y vs X^2: In this dataset, the relationship between Y and X is no longer a straight line, but rather a parabola. The dataset still has a mean of 10 and a variance of 1, but the correlation coefficient has increased to 0.943.

Y vs sin(X): In this dataset, the relationship between Y and X is now a sine wave. The mean and variance of the dataset remain the same as the previous two datasets (10 and 1), but the correlation coefficient has decreased slightly to 0.785.

Y vs tan(X): In this final dataset, the relationship between Y and X is now a tangent function. Again, the mean and variance of the dataset remain the same as the previous three datasets (10 and 1), but the correlation coefficient has dropped significantly to 0.357.

Despite having the same statistical properties, the four datasets display vastly different visual patterns, illustrating the importance of considering the underlying structure of the data when interpreting statistical results. The first dataset displays a strong positive linear relationship between Y and X, while the second dataset shows a quadratic relationship. The third dataset exhibits a cyclical pattern, and the fourth dataset appears to have no discernible pattern.

Anscombe's Quartet highlights several important lessons for statisticians and data analysts:

Correlation does not imply causality: Two variables can have a strong correlation without being causally related. In fact, in some cases, the relationship between the variables may be due to a third, unobserved variable.

Visual inspection is essential: Statistical measures alone are not sufficient to understand the nature of the relationship between variables. Visual inspection of plots and charts can reveal important details about the structure of the data that statistical measures cannot capture.

Different transformations can change the appearance of the data: Applying different mathematical functions to the data can entirely change the visual appearance of the relationship between variables, even though the underlying statistical properties remain the same.

Be cautious when interpreting statistical results: Statistical analysis should always be accompanied by domain knowledge and a deep understanding of the underlying data. Otherwise, misinterpretations and incorrect conclusions can arise from seemingly robust statistical analyses.

Overall, Anscombe's Quartet serves as a reminder that statistics and machine learning models are only tools, and that careful consideration must be given to the context and meaning of the data when drawing conclusions or making decisions.

Ques3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that evaluates the strength and direction of the linear relationship between two continuous variables. It is a widely used metric in statistics and data analysis to quantify the association between two variables.

Pearson's R ranges from -1 to 1, where:

A value of 1 indicates a perfect positive linear relationship between the two variables.

A value of 0 indicates no linear relationship between the two variables.

A value of -1 indicates a perfect negative linear relationship between the two variables.

The calculation of Pearson's R involves determining the covariance between the two variables and dividing it by the product of their standard deviations. The formula is:

Pearson's R = Cov(X,Y) / (σ_X * σ_Y)

where X and Y are the two variables, Cov(X,Y) is the covariance between them, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Interpreting Pearson's R can be tricky, as the magnitude of the coefficient does not necessarily indicate the importance of the relationship. For example, a small correlation coefficient can still represent a strong relationship if the variables are highly correlated in a non-linear manner. Therefore, it's essential to consider other factors, such as the context of the data and the goals of the analysis, when interpreting Pearson's R.

Some common applications of Pearson's R include:

Assessing the strength of a linear relationship between two variables.

Identifying patterns in data and exploring potential associations between variables.

Testing hypotheses about the relationships between variables.

Building predictive models, such as linear regression, which relies on the assumption of a linear relationship between the independent and dependent variables.

However, it's important to note that Pearson's R assumes a few conditions are met:

The data must be normally distributed.

The observations must be independent and identically distributed.

There must be no multicollinearity between the variables.

Violating these assumptions can lead to biased or inconclusive results, so it's crucial to check the validity of the assumptions before interpreting Pearson's R. Additionally, there are alternative correlation measures, such as Spearman's rank correlation, that may be more appropriate for non-normally distributed data or ordinal variables.

Ques4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling refers to the process of transforming a dataset or a signal to have a specific range or distribution. The purpose of scaling is to improve the quality of the data or signal for further processing or analysis.

There are several reasons why scaling is performed:

To improve the interpretation of data: By scaling the data to a familiar range, it becomes easier to interpret and analyze. For example, temperature measurements in Celsius or Fahrenheit are more intuitive than measurements in Kelvin.

To reduce the effect of outliers: Outliers can dominate the analysis of data, leading to inaccurate results. Scaling can help to reduce the impact of outliers by compressing or expanding the data range.

To improve model performance: Some machine learning algorithms are sensitive to the scale of the data. By scaling the data, it is possible to improve the performance of these algorithms.

To facilitate comparison: Scaling can help to compare data from different sources or experiments. By scaling the data to a common range, it is possible to compare the relative magnitudes of the data.

Normalized scaling and standardized scaling are two types of scaling methods. The main difference between them lies in how they transform the data.

Normalized scaling transforms the data to a specific range, usually between 0 and 1, by dividing the data by a constant value, such as the maximum or minimum value of the data. This type of scaling is useful when the data has a fixed upper or lower bound, and the relationships between the data points are relative rather than absolute. Examples of normalized scaling methods include min-max scaling, vector scaling, and histogram equalization.

Standardized scaling, on the other hand, transforms the data to have a mean of zero and a standard deviation of unity. This type of scaling is useful when the data has a bell-shaped distribution and the relationships between the data points are absolute rather than relative. Standardized scaling methods include z-score scaling and Fisher transformation.

In summary, scaling is a data preparation technique used to improve the quality of data or signals for further processing or analysis. Normalized scaling and standardized scaling are two types of scaling methods, each suitable for different types of data distributions and relationships. While normalized scaling is useful for data with fixed bounds, standardized scaling is useful for data with bell-shaped distributions.

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

Value of VIF is infinite due to below reasons:

Multicollinearity: When there are redundant predictors in the model, the variance of the regression coefficients can become very large, leading to an infinite VIF.

Non-identifiable models: If the model is not identifiable, meaning that the parameters cannot be uniquely determined from the data, the VIF can become infinite.

Model misspecification: If the model is incorrectly specified, such as including irrelevant predictors or omitted variables, the VIF can become infinite.

Outliers or anomalous data points: Presence of outliers or anomalous data points in the dataset can cause the VIF to become infinite.

Large sample size: As the sample size increases, the degrees of freedom in the model increase, which can cause the variance of the regression coefficients to decrease. However, if the model is misspecified or there are too many irrelevant predictors, the variance of the regression coefficients may not decrease sufficiently, leading to an infinite VIF.

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

A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used to compare the distribution of residuals from a linear regression model to a theoretical distribution, usually a normal distribution. The plot shows the quantiles of the residuals on the y-axis against the quantiles of the standard normal distribution on the x-axis.

The use and importance of a Q-Q plot in linear regression are as follows:

Checking normality: A Q-Q plot helps determine whether the residuals are normally distributed, which is a crucial assumption in linear regression. If the residuals are not normally distributed, it may indicate a problem with the model or the data.

Identifying outliers: The plot highlights outliers or influential points that may affect the model's fit. Residuals that fall outside the 95% confidence interval for the mean of the error term are considered outliers.

Detecting heteroscedasticity: A Q-Q plot can detect heteroscedasticity, where the variance of the errors changes over time or with the level of the predictor variable(s). If the residuals have a non-constant variance, the plot will deviate from a straight line.

Evaluating model assumptions: A Q-Q plot can assess the validity of linear regression assumptions, such as homoscedasticity, normality, and independence of errors.

Diagnosing ommited variable bias: A Q-Q plot can help identify omitted variable bias, which occurs when a variable that should be included in the model has been left out.

By examining the Q-Q plot, analysts can quickly spot potential issues with their linear regression model and take appropriate measures to address them, leading to better model performance and increased accuracy in predictions.