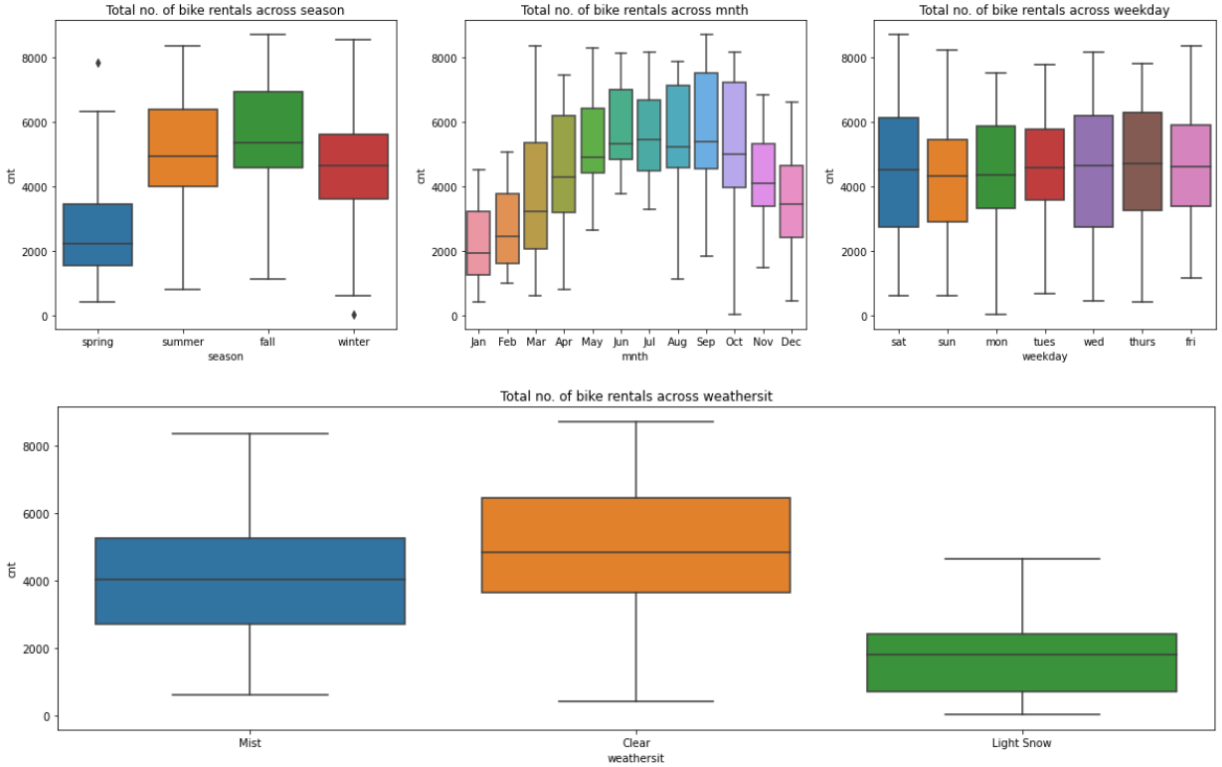
Assignment-Based Subjective Questions

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

   For the categorical variables, season, mnth, weekday and weathersit we have observed the following relationship with the target variable cnt.
   Following points can be inferred about the categorical variables effect on the dependent variable.

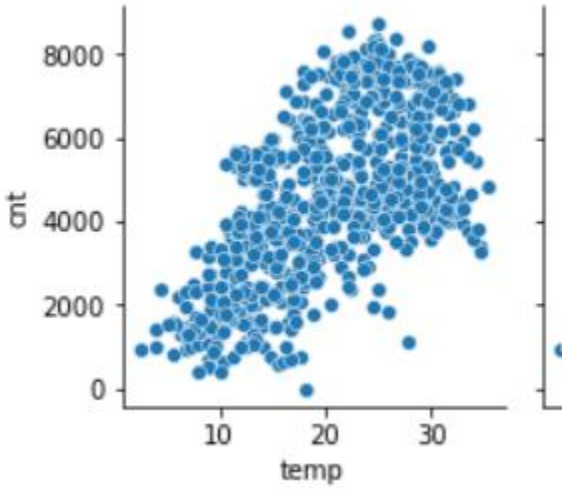
   - The total number of bike rentals is least during the spring season and maximum during the fall season. Also, during fall, total number of bike rentals is generally higher than other seasons.
   - Number of bike rentals is increasing from Jan to Sep (exception in Jul month) and then starts decreasing from Oct to Dec. Maximum number of bike rentals are in the month of Sep.
   - During sat, wed and thurs number of bike rentals is max.
   - During the clear weather number of bike rentals is maximum and minimum during the light snow.



2. **Why is it important to use drop_first=True during dummy variable creation?**
   For a categorical variable with n levels, n-1 dummy variables are required, indicating the levels.

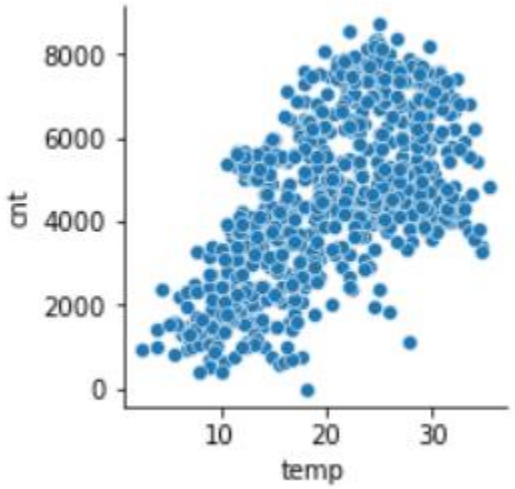
However, during the dummy variable creation, n variables are created, each variable representing one level. So, we have the redundant information in one column which can be deduced by using other n-1 dummy variables.
Therefore, drop_first=True is used to drop the first column of dummy variables and make the dummy variable count to n-1.
In short, to reduce data redundancy it's important to use drop_first=True during dummy variable creation.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Since atemp and temp predictor variables are highly correlated, therefore atemp variable is dropped from the data.
   By looking at the pair-plot we can see **temp** has highest correlation with the target variable.

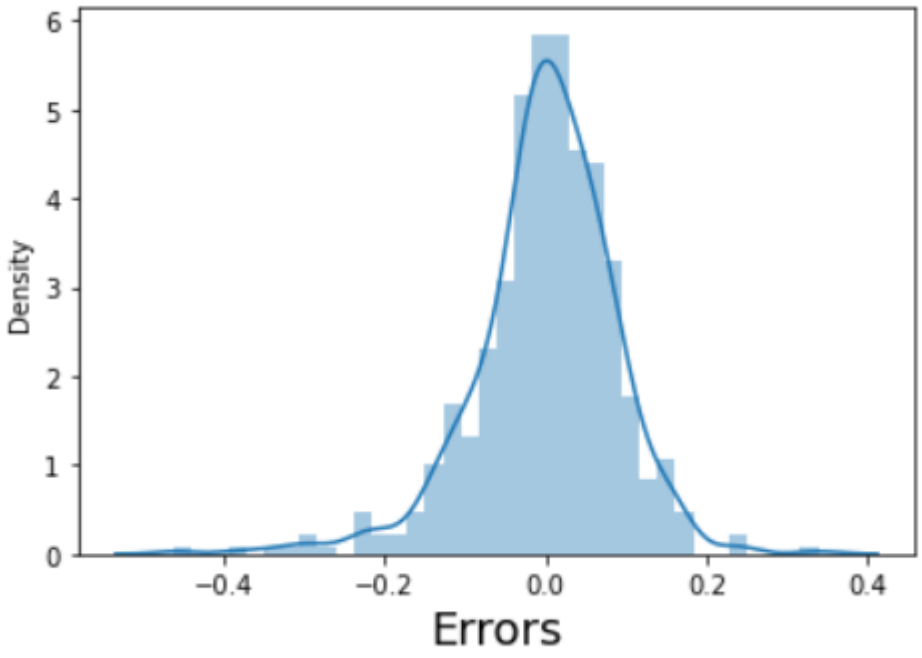


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

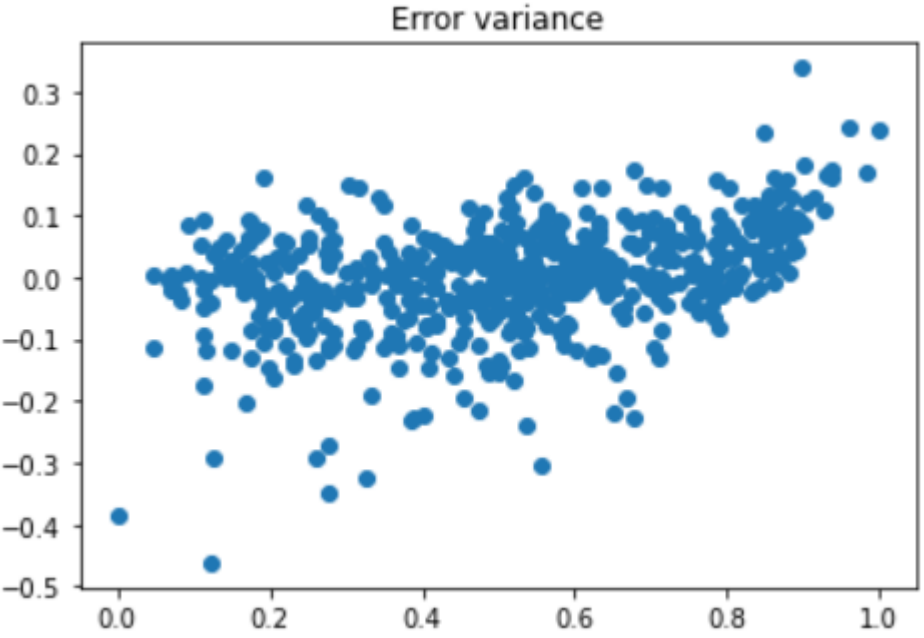
- Atleast one of the predictor variable must follow a linear relationship with target variable. Here we can see 'temp' follows a pretty linear relationship with 'cnt' variable.



- Error terms are *normally distributed* with mean zero. This is required to make further interpretations.



- *Error terms are* independent *of each other*. **i.e.** *The variance should not increase/decrease as error values changes*. i.e. Variance shouldn't follow any pattern as the error terms change.



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

General Subjective Questions

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

   Linear Regression is a machine learning algorithm based upon the supervised learning.
   Using this algorithm, we can model a target value depending upon one or more independent predictor variables. Hence, this method is mostly used for forecasting and finding out the cause and effect relationship between variables.

   The whole dataset is split into training and test data. Model is built on training data and is tested on test data.

   For Multiple Linear Regression (where target variable is dependent on more than 1 independent variable), After model building, Multicollinearity (phenomenon of having related predictor variables in the input dataset, giving large VIF values) and overfitting (While adding more variables, the model may become far too complex and hence end up memorizing all the training data and fail to generalize) issues might occur which need to be catered.

   Feature selection is performed to get the significant predictor variables and the non significant ones. The selection is done on various metrics like R squared, adjusted R-squared, p value, VIF etc. After removing the non-significant variables, model is rebuilt and again the same exercise is carried until we have satisfactory metrics available. Model building can be done using RFE (for larger number of predictor variables, it is highly recommended), Forward Selection, Backward selection and stepwise selection techniques.

   When the model is trained, it fits the best line/hyperplane to predict the values of target(y) variable for given values of predictor(x) variables.

   Here, Coefficients of each predictor variable are obtained by minimizing the sum of squared errors, the least squares criteria. For making inferences on the entire data set based upon the results on sample data, following assumptions are required to be followed.

   1. Atleast one of the predictor variable must follow a linear relationship with target variable.
   2. Error terms are normally distributed with mean zero.
   3. Error terms are independent of each other.
   4. Error terms have constant variance.

So, after fitting the best line/hyperplane, Residual analysis is carried on the training data to verify the above assumptions. After that the model is predicted and verified on the test data.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet consists of 4 data sets that have nearly identical simple descriptive statistics, yet their data distributions are very different when graphed. Each of the 4 datasets have 11(x,y) data points. This demonstrates the importance of graphing the data before analyzing it and the effect of outliers and other influential observations on statistical properties. This means, numerical observations can be exact but graphs can be rough for 4 data sets.
   The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. **What is Pearson's R?**
   Pearson's R coeffcient checks for the linear correlation between two sets of data. This is highly used in Linear Regression. It describes the covariance of two variables divided by product of their standard deviations. It's value lies from -1 to 1 , so this is normalized measurement of covariance. A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other, A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other and Zero means that for every increase in one variable, there isn't a positive or negative increase in other.
   Given a pair of random variables X and Y, Pearson's coefficient can be written as covariance of X and Y divided by product of their standard deviations.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling is pre processing step used in multiple linear Regression model. In this, we scale all the numerical variables in the dataset to fit in a particular range.
   It is important to have everything on the comparable scale for the model to be easily interpretable and to derive the inferences.
   If we don't have comparable scale, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. Hence, scaling is performed so that the units of coefficients obtained are all on the same scale.

   We have two types of scaling technique available.
   1. Min-Max: All the numeric data is compressed in the range of min and max value of the variable.
   2. Standardized Scaling: It follows a normal distribution with mean centered at 0 and standard deviation of 1.

The difference between Min max and Standardized scaling is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is extreme data point (outlier).

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   VIF suggests the correlation of a variable with a set of other independent variables. If the variable can be explained/derived by the set of other variables , then the correlation value would be high leading to increase in R squared value (as the sum of squares would be minimum).
   VIF = 1/(1-Rsquared)
   Hence higher the R squared value, higher the VIF.
   When R squared value=1, VIF becomes infinite.

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

   Q-Q is a quantile quantile plot. So, this is the plot of two quantiles against each other.
   Quantile is a fraction where certain values fall below that quantile.
   The main purpose of Q-Q plot is to find if the two sets of data come from the theoretical distribution like Uniform, normal, exponential or if it comes from populations with common distribution.
   A 45 degree angle is plotted on the Q-Q plot. If the two data sets come from same distribution, then points will fall on that reference line. On y axis we have exponential data quantiles and on x axis we have normal data quantiles. If all the data points on 45 degree, then the both data sets have same distribution.

   Use and Importance in Linear Regression: This helps in a scenario 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, have common location and scale, have similar tail behavior, have similar distributional shapes etc or not.