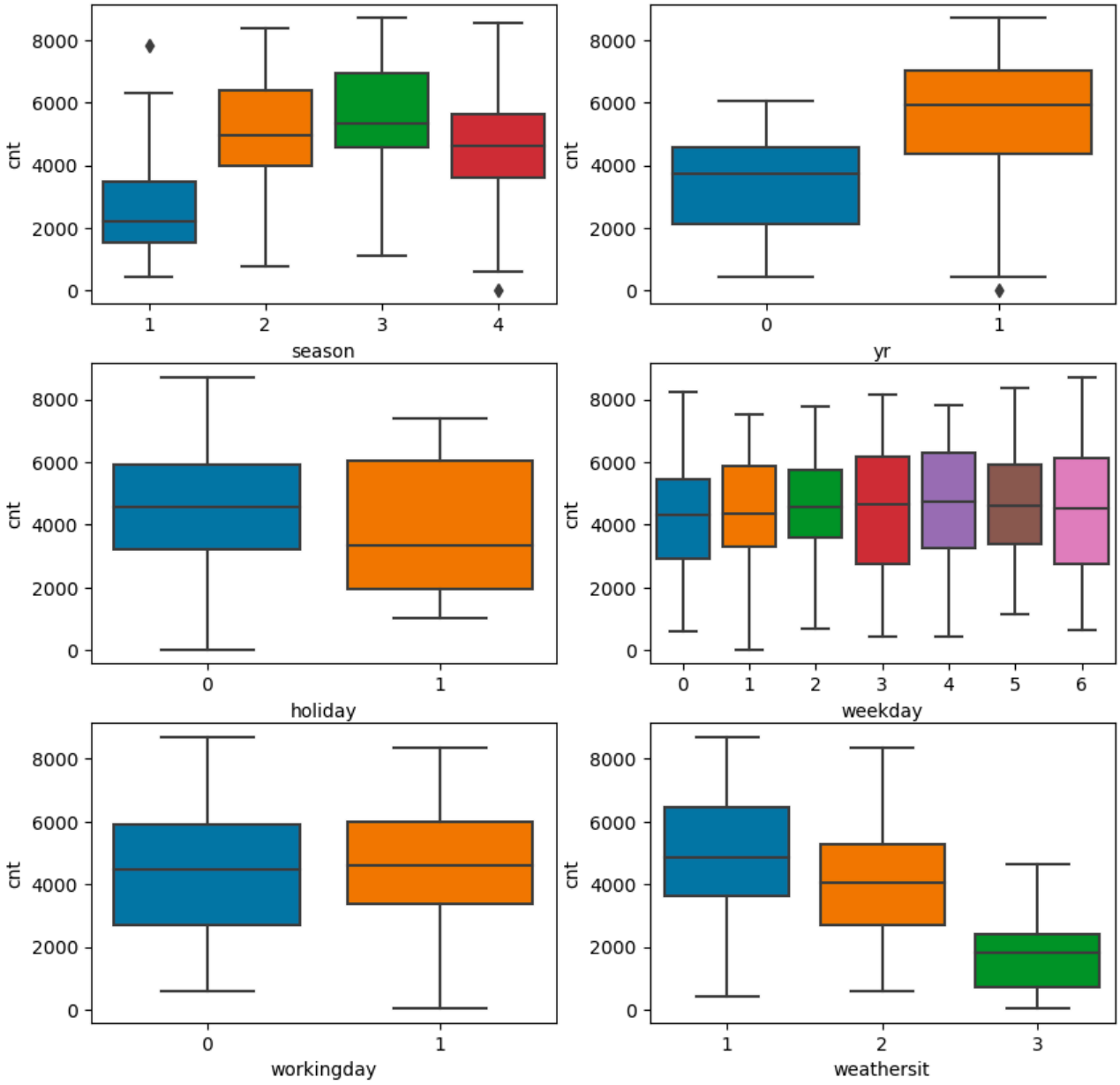
Assignment-based Subjective Questions and Answers

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

ANS 1:



- During Fall there is considerably high demand
- 2019 year was having highest demand compared to 2018
- On average 4500 counts of demand on weather situation like Clear, Few clouds, Partly cloudy, Partly cloudy compared to other weather situations
- There is no considerable pattern found in between weekdays. Still 6th and 3rd day of week have slightly high demand
- During holidays there is slightly less demand compared to non-holidays
- During working days there is more demand

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

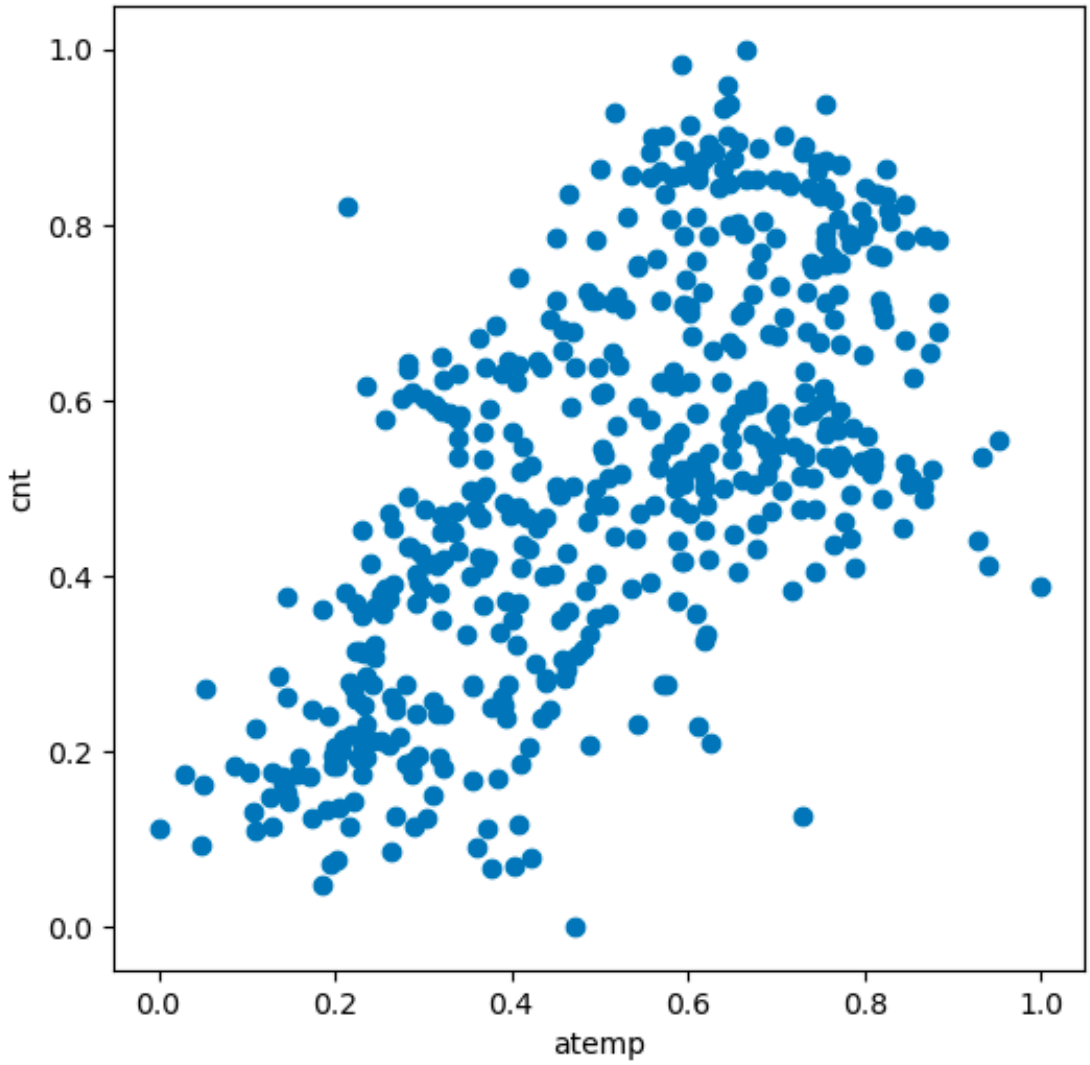
ANS 2: The main importance of adding drop_first=True is to avoid the perfect collinearity between the dummy variables.

For example, consider a categorical variable having 3 levels of values, and when you add dummy variables for this categorical variable, it creates 3 new dummy_variables. It will be evident that value first of those 3 new dummy variables is easily predicted when we know the value of remaining 2. So this first variable can add perfect collinearity to the model which can add extra complexity to the model predictors. So, by dropping it using drop_first

- Avoids Multicollinearity
- Improved Model Performance
- Simpler Models

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

ANS 3:



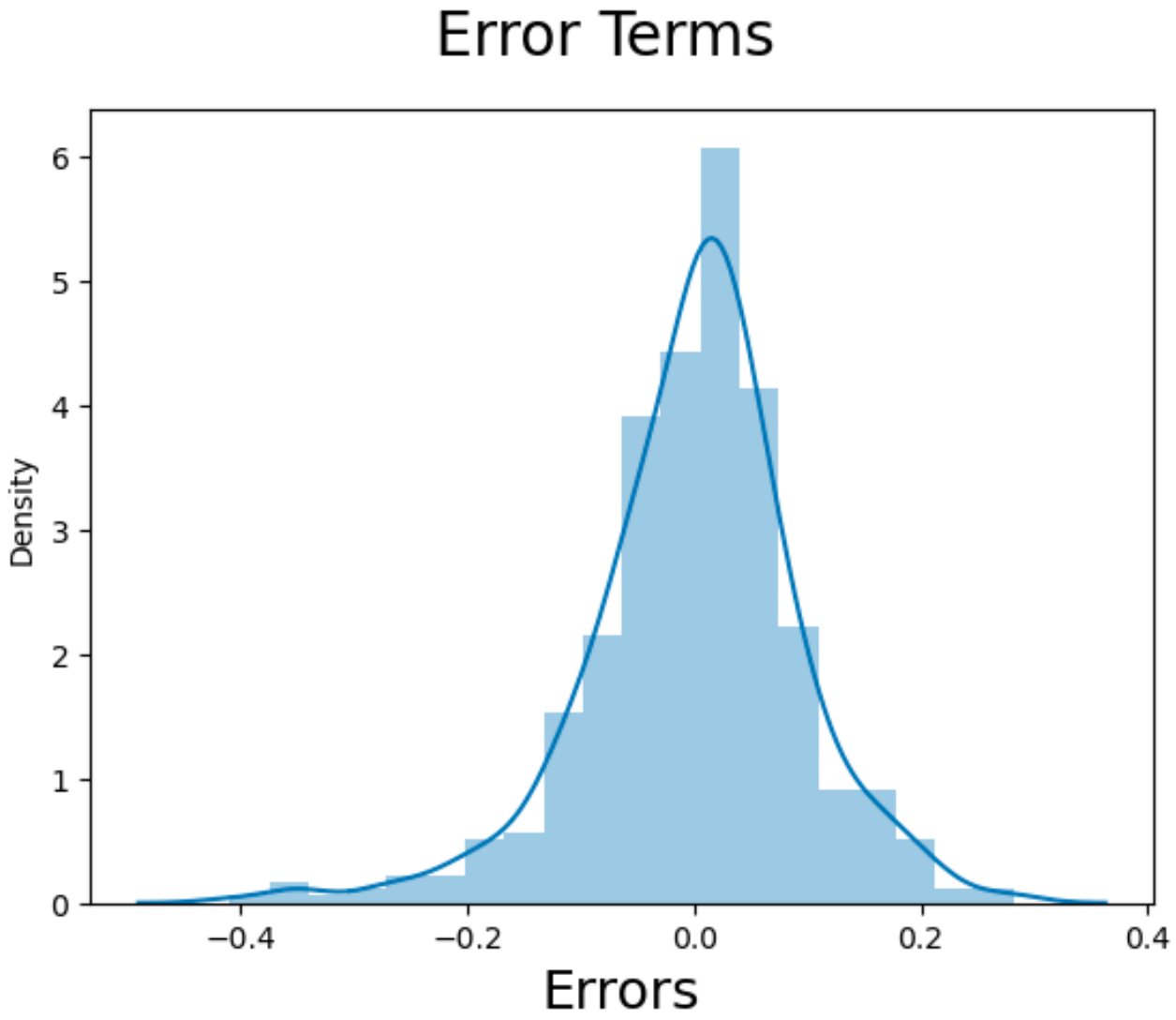
Looking at the pair-plot, the temp is having highest correlation with cat (target variable)

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

ANS 4: The linear regression assumptions are validated using Residual Analysis. This is to check if the errors are normally distributed.

I followed following steps to validate this

- Calculated predicted values of target variable of training set (y_train_pred) fro the model
- Plotted the distribution plot for actual values of target variable of training set - predicted target variable values (y_train - y_train_pred)
- And observed that the normal distribution of residuals in the dist_plot and ensured that residuals are normally distributed around 0.



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

ANS 5: Top 3 features contributing significantly towards explaining the demand of shared bikes

1. **Temp -** highest positive coefficient = 0.5735
2. **weathersit_3 -** Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - Highest negative coefficient = -0.2814
3. **Yr** - year - Second highest positive coefficient - 0.2320

General Subjective Questions and Answers.

1. Explain the linear regression algorithm in detail.

ANS 1:

- Linear regression is the is a statistical method used to model the relationship between a dependant variable and one or more independent variables. The main goal is to find the linear equation (y = beta_0 + beta_1 *X1 + beta_2 * X2 +… beta_n * Xn) that best predicts the dependant variable from the independent variables (X1, X2, ..Xn)
- The coefficients beta_1, beta_2, beta_n are estimated using the methods of Least Squares that minimised the sum of squared differences between observed value and values predicted by the model.

- Linear regression assumptions are linearity, independence, homoscedasticity, normality of error terms (residuals) and no multicollinearity and no overfitting of the models.
- The models's fit is evaluated using metrics such as R-Squared value that measures the proportion of the variance of the dependent variables explained by the independent variables.
- Due to its simplicity, the linear regression is treated as the fundamental tool in predictive modelling and data analysis.

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

ANS 2:

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties (mean, variance, correlation, and linear regression line) but appear very different when graphed.
- the quartet illustrates the importance of graphing data before analyzing it.
- Each dataset has the same mean and variance for both x and y, the same correlation coefficient between x and y, and nearly identical linear regression equations.
- However, their scatter plots reveal different patterns: one shows a linear relationship, another a non-linear relationship, the third includes an outlier affecting the regression line, and the fourth contains a vertical outlier.
- The quartet demonstrates that relying solely on summary statistics without visualizing the data can be misleading, highlighting the need for exploratory data analysis to understand the true nature of the data.

3. What is Pearson's R?

ANS 3: Pearson's correlation coefficient (Pearson's R) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association, ranging from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is sensitive to outliers and assumes that the relationship between variables is linear.

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

ANS 4:

- Scaling is the process of transforming data to fit within a specific range or distribution. It's performed to ensure that all features have a similar scale, making them comparable and preventing features with larger scales from dominating the model's training process.
- Normalized scaling, also known as min-max scaling, rescales the features to a specified range, typically between 0 and 1. It preserves the original distribution of the data and is useful when the data does not follow a normal distribution.
- Standardized scaling, also known as z-score scaling, transforms the features to have a mean of 0 and a standard deviation of 1. It centres the data around 0 and scales it to unit variance.
- Standardized scaling is beneficial when the data has outliers or follows a normal distribution, as it maintains the shape of the distribution and helps to compare variables measured in different units.

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

ANS 5:

- VIF infinite means a predictor variable is perfectly predictable from other predictor variables in the model.
- It happens when predictors are strongly related to each other. When this happens, the model struggles to estimate the individual effect of each predictor accurately that results in inflated uncertainties in the model predictions
- During this, the R-Squared value of a predictor variable with respect to another predictor variable become 1, that leads to VIF value as infinite as VIF is calculated as 1/(1-R-squared) where denominator of equation become 0 that leads to infinite.

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

ANS 6:

- A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given set of data follows a particular probability distribution, typically the normal distribution.

- It compares the quantiles of the observed data to the quantiles of a theoretical distribution, such as the normal distribution.
- Q-Q plots provide a simple yet powerful visual tool for assessing the validity of assumptions and diagnosing potential issues in linear regression analysis.
- They help ensure the robustness and reliability of the regression model by guiding appropriate adjustments and interpretations.
- Its importance includes
  - Assumption Checking
  - Testing Outliers
  - Model Diagnostics
  - Comparison across models