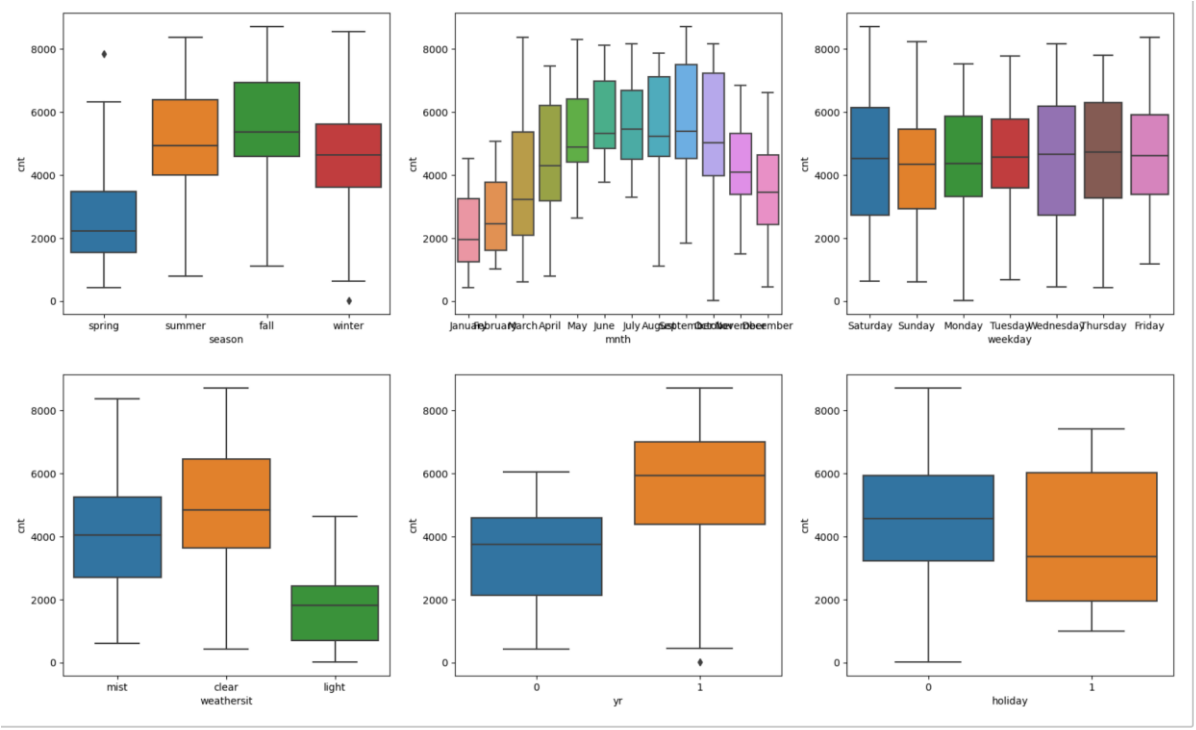
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   Following are the categorical variables found in the dataset.



   Following are the observation

   a. Season – fall season has more bike booking, may be fall is considered as off-season for outdoor activities
   b. Mnth – Bike demand is high in the months from May to October
   c. Weekday – The demand of bike is almost same throughout the weekdays
   d. Weathersit – Bike demand is high in clear weather
   e. Yr – Bike demand is high in 2019 compared to 2018
   f. Holiday – No impact of holiday in bike booking

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

   'drop_first=True' is used to achieve k-1 values with categorical data. It will drop the first extra column.

   Ex: In the below if Summer and Fall is 0 then it implies Fall. So no need of extra column. We can drop this.

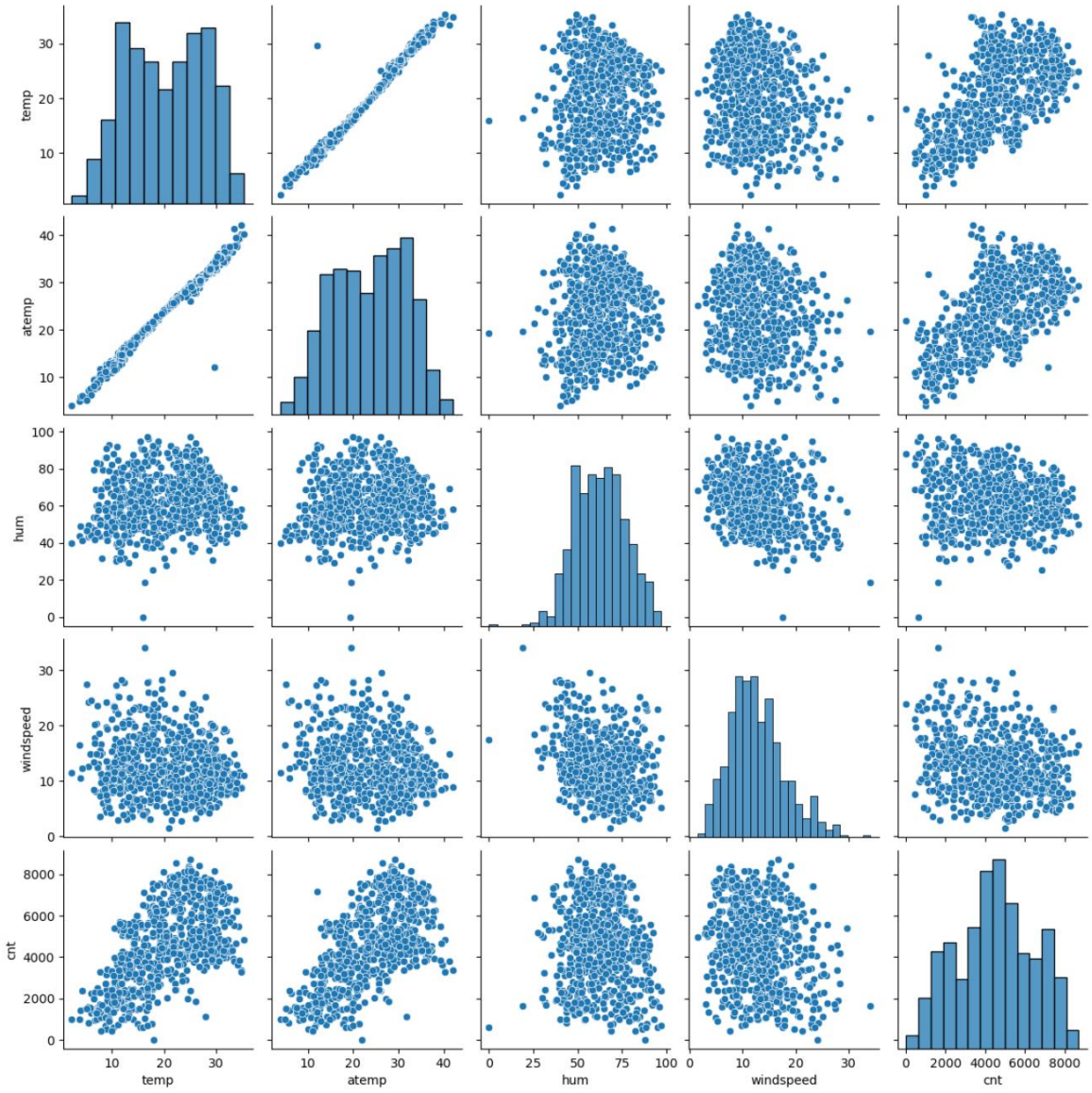
Spring	Summer	Fall
1	0	0
0	1	0
0	0	1

   Dropping un-necessary variables makes the model simple and reduces the redundancy.

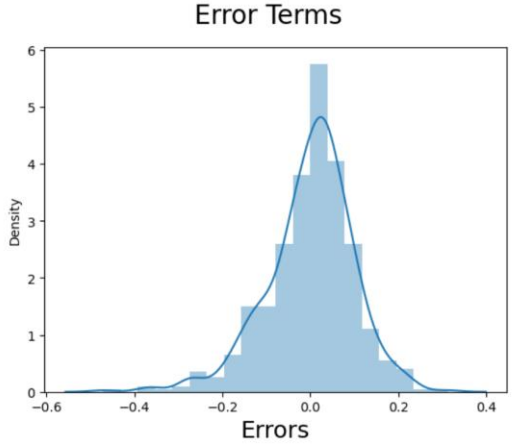
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   From the below pair-plot atemp and temp looks like highest correlation with target (cnt)

   From the heat map – it's atemp(0.65)



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   a. Residual Analysis: After building linear regression model, I performed residual analysis. Plot the error (ytrain and ypred) and observe normal distribution with mean 0 as follows



   b. linearity: plot between dependent and independent variable almost linear

c. Multicollinearity: Calculating VIF helps to identify multicollinearity
d. homoscedasticity: Observe patterns in residuals

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   correlation , P-value and VIF

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is like finding the best-fit line to predict something based on known factors. It's about adjusting the line to minimize how wrong your predictions are.

The coefficients in the equation tell you how much each factor influences the predicted outcome.

Given a problem, Identify dependent and independent variable. The dependent variable is also called target variable that we need to predict based on independent variables.

$$Y=\beta_0+\beta_1 \cdot X$$

Y – target variable

X – independent variable

β_0 – intercept

β_1 – Slope of the line.

goal is to find the best β0 and β1 values that make the line come as close as possible to actual one.

Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

- Following are the assumptions used for Linear regression Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.

2. Explain the Anscombe’s quartet in detail. (3 marks)
   Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
   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

regression model if build. They have very different distributions and appear differently when plotted on scatter plots. The four data set plots which have nearly same statistical observations, Which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data etc, Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. **What is Pearson's R? (3 marks)**
   The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Scaling is a step used in data preparation 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. And helpful in visualization.
   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.

   Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   Standardization Scaling: It brings all of the data into a standard normal distribution which has mean(μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.
   Removing variables with high VIF values can help reduce multicollinearity and improve the accuracy and stability of the regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.
   Use of Q-Q plot:
   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45- degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.