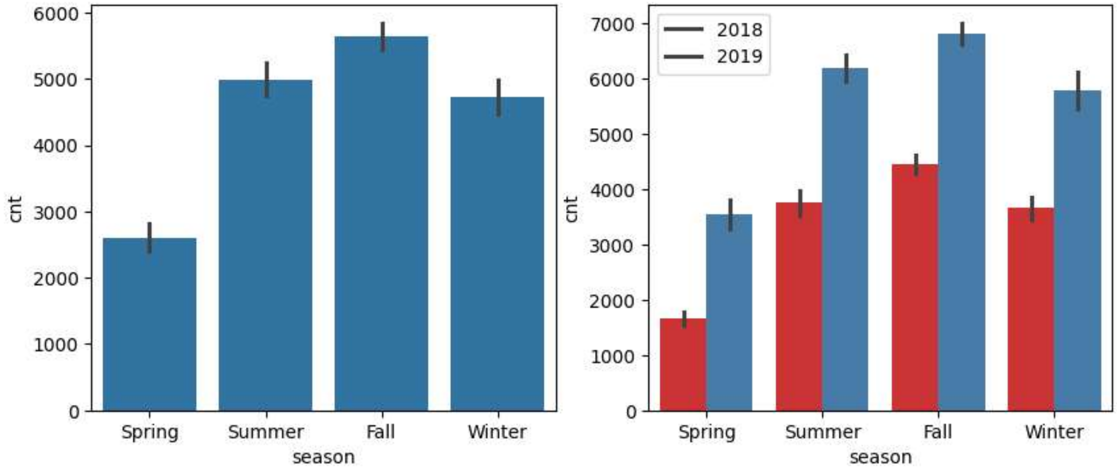
Assignment-based Subjective Questions

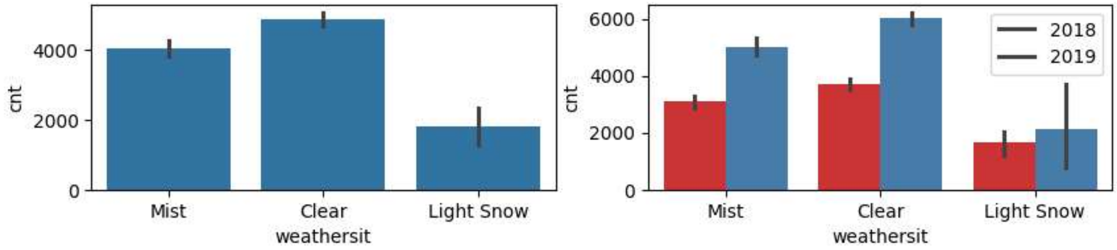
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

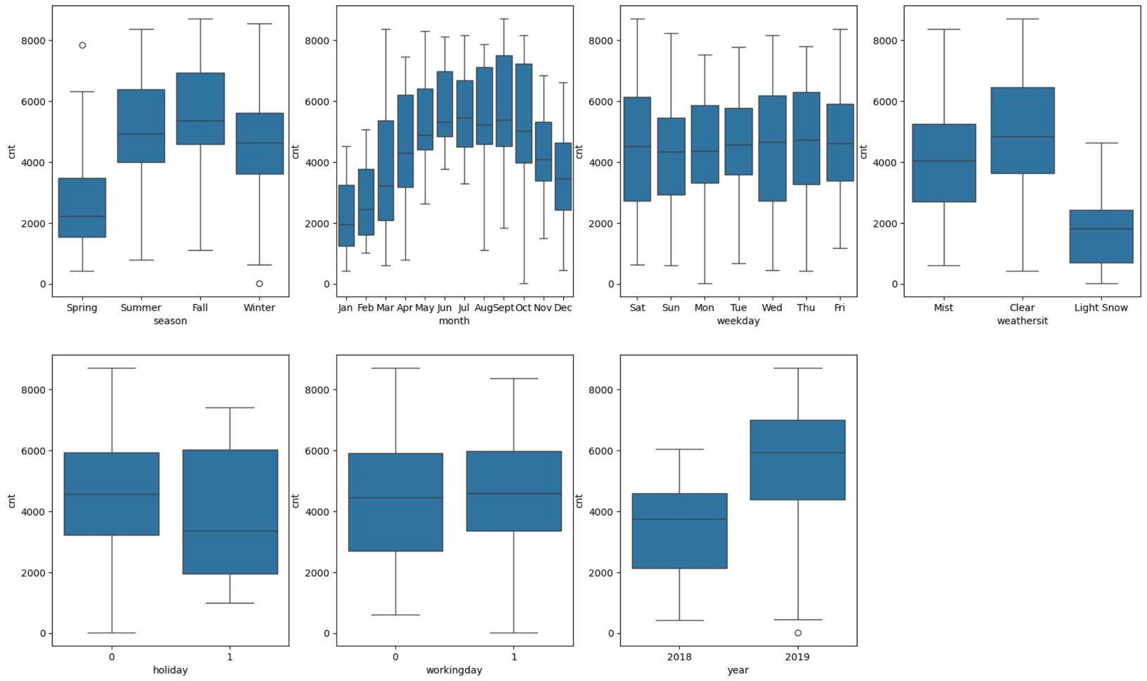
   I have done analysis on categorical columns using the boxplot and bar plot.

   Below are the few points we can infer from the visualization –

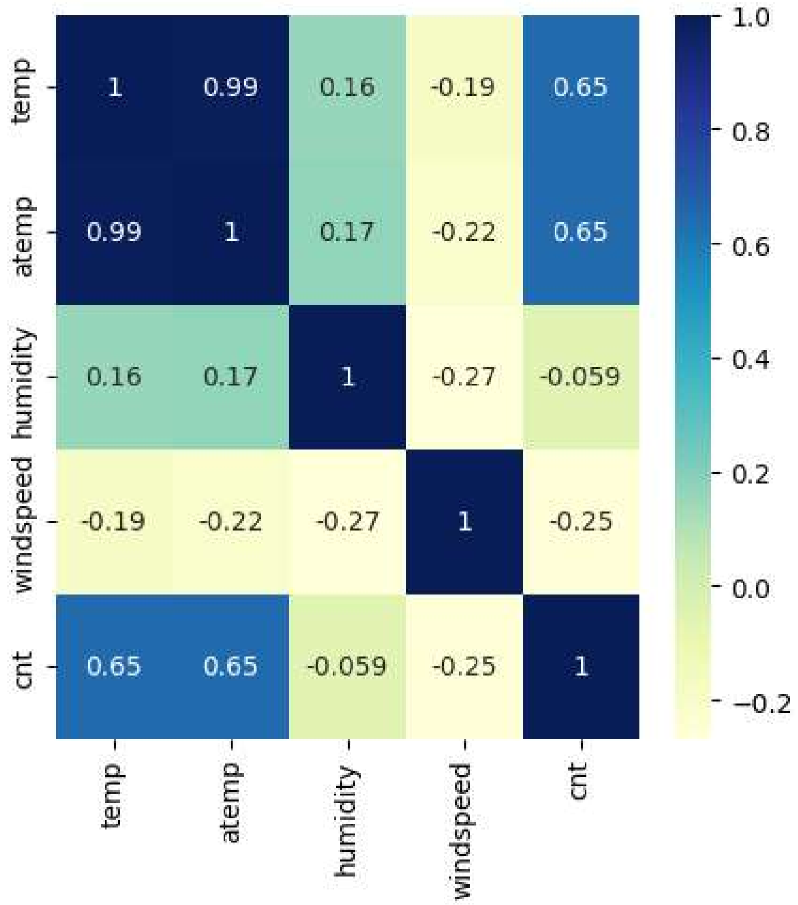
   - Fall Season seems to have attracted more booking. i.e Fall Season has more number of Count. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Count of users always increased in all the months of 2018 when compared to the same months of 2019.
   - The mean count of 2018 is much lower than mean count of 2019.
   - The increase in the users trend increased from Jan till Sep just other than July. And after Sep the count decreased.
   - On holidays the count was less when compared to Work day.







2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Drop_first=True is used to drop the base column on which the dummy variables are created. This helps in reducing one extra column for analysis. This also helps in reducing the correlations created among the dummy variables.
   When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   "temp" is the numerical variable which has the highest correlation with the target variable.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   I had validated the below assumptions of Linear Regression –
   - ➢ Error Terms are normally distributed.
   - ➢ Multicollinearity
   - ➢ Homoscedasticity
   - ➢ LR validation
   - ➢ Residuals are independent
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Based on the final model, below are the 3 top features contributing significantly towards the demand of shared bikes –
   - ✓ Year
   - ✓ Spring
   - ✓ Light Snow

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. It plays a pivotal role in predicting continuous outcomes. Linear regression predicts the relationship between two variables by assuming a linear connection between the

independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.

Applied in various domains like economics and finance, this method analyzes and forecasts data trends. When there is only one Independent variable upon which the prediction is made then it is Simple Linear Regression, where as if there is more than one independent variable, it is Multiple Linear Regression. Multiple linear regression can involve several independent variables.

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

Y = β0 + β1X

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression

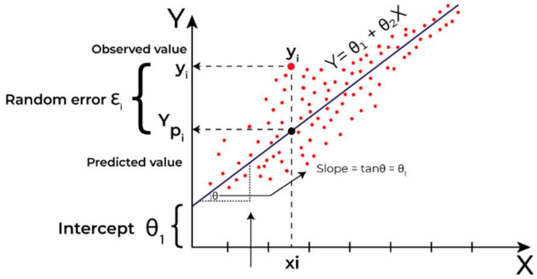
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

Y = β0 + β1X1 + β2X2 + ……………………… βnXn

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

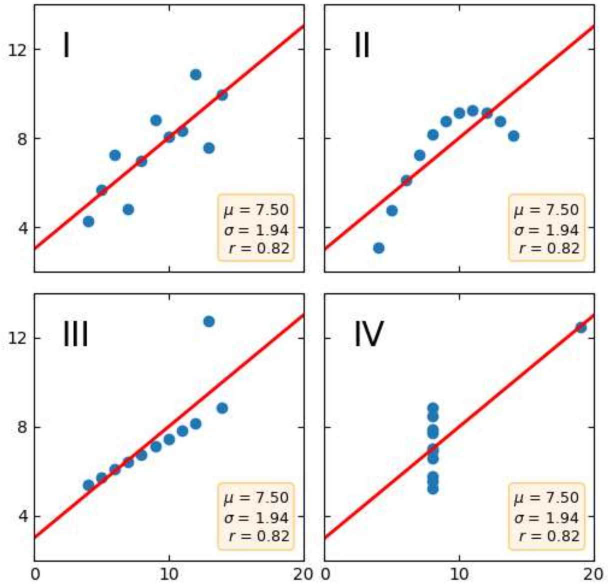
The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.


Y
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
Slope = tanθ = θ_1
θ
Intercept θ_1
xi
X

2. **Explain the Anscombe's quartet in detail. (3 marks)**

   Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

   It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



3. **What is Pearson's R? (3 marks)**

   The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

   - Pearson's r
   - Bivariate correlation
   - Pearson product-moment correlation coefficient (PPMCC)
   - The correlation coefficient

   The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

   Feature scaling is a method used to normalize the range of independent variables or features of data. Since the dataset contains different columns & the values of the columns will vary in a wide range, if we build a model on top of the column data as such, ML algorithm will tend to weigh values based on the higher values. Hence to avoid them, we bring all the values in common specific range, so that weightage of each feature is calculated approximately proportionately. Some of the methods of scaling are

   - Rescaling (Min-max Normalization)
   - Standardization

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

   VIF – Is the Variable Inflation Factor. VIF calculates how well one independent variable is explained by all the other independent variables combined.

   VIF is calculated as below –

$$VIF_i = \frac{1}{1-{R_i}^2}$$

So if we analyze why VIF becomes infinite, it becomes infinite when denominator is 0, i. e, when R2 =1. R2=1 for that variable with it has perfect correlation with the other independent variables. Hence for this if we drop one of the independent variable, the VIF drops.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.