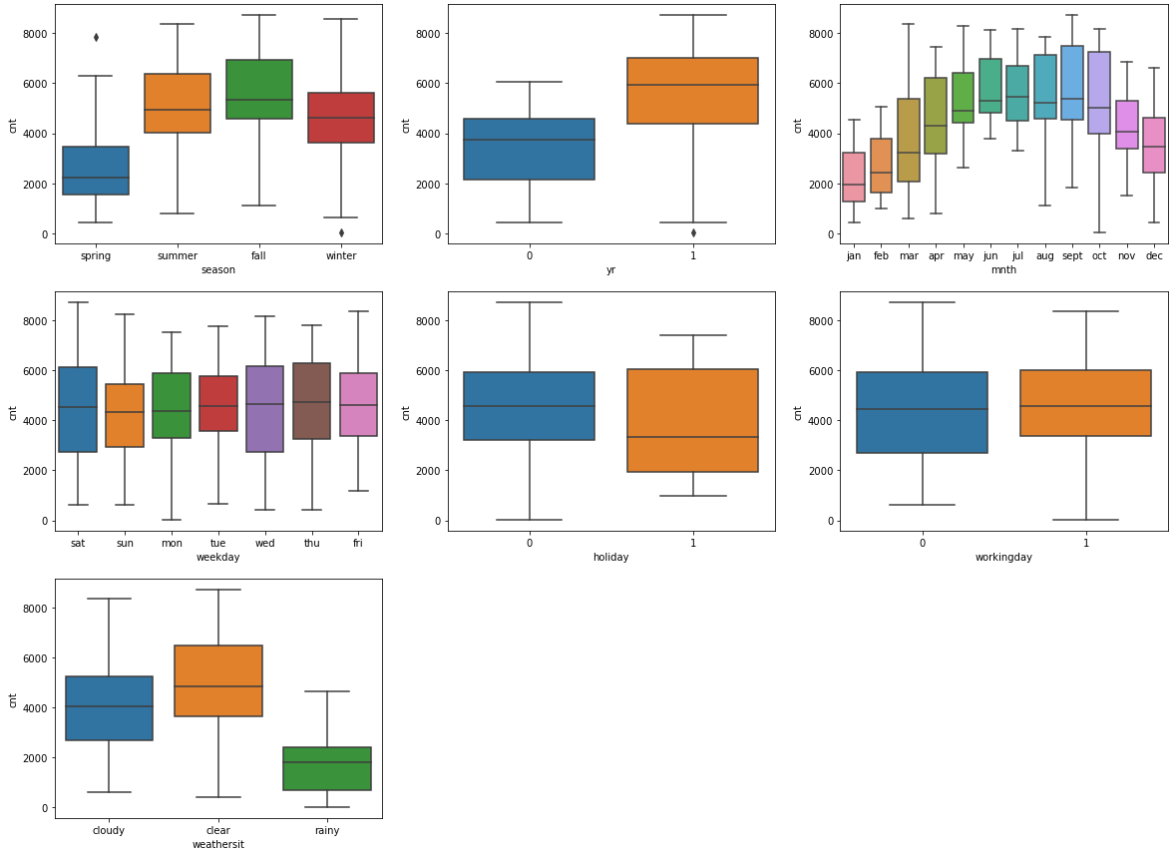
Assignment-based Subjective Questions

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

   Answer: Most of the Categorical variables are correlated to the dependent variable under analysis. The correlation is clarified in the below box plots. The relation is high with many of the variables like Year, season, and months, and weathersit, and is less with other categorical variables like weekday and workingday.

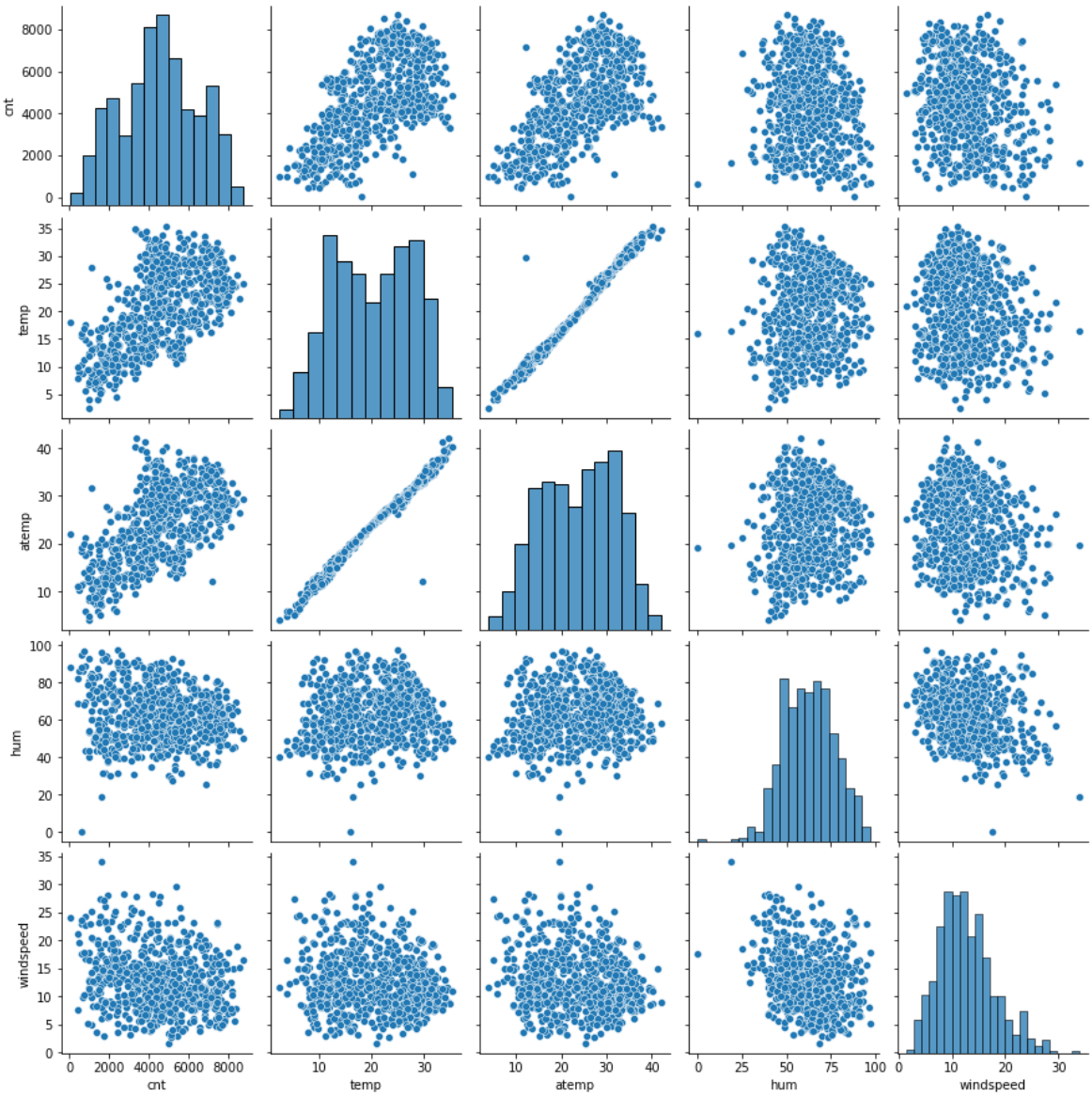


2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Answer: Adding more variables in the model incur a penalty. A categorical variable with n categories can be defined with (n-1) dummy variables. For example weathersit (Clear, cloudy, rainy) the 3 values can be defined by 2 variables Cloudy(1/0) and Rainy(1/0) a 00 value means not cloudy and not rainy which means it is clear. (10: Cloudy, 01: Rainy, 00: Clear). Thus for any categorical variables with n possible values we need n-1 dummy variables to represent. The drop_first= true option in create dummies tells the library to drop the first possible value and only create dummies for n-1 values. In the example Clear, Cloudy, Rainy instead of using 100, 010, 001 to represent, we use 10 and 01 while 00 replace 100 to represent the first possible value is true.

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

From the pairplots, temp and atemp should the highest correlation with the target variable cnt.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: Through residual analysis, I confirmed that the errors are normally distributed with mean near to 0. Confirmed independency between dependent variables and dropped one of the variables that has high strict correlation with another (dropped atemp) as it is highy correlated with temp.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top 3 features contributing significantly toward demand of shared bikes are:

Yr, Temp, and season.

Holiday has also high contribution to the target variable (negative correlation), but I dropped in the final model to ensure zero p-values for all variables in the model.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	286.0
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	1.55e-181
Time:	16:09:53	Log-Likelihood:	-4154.4
No. Observations:	511	AIC:	8327.
Df Residuals:	502	BIC:	8365.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2316.3078	152.930	15.146	0.000	2015.847	2616.769
yr	1991.3060	73.684	27.025	0.000	1846.539	2136.073
temp	3072.6222	215.967	14.227	0.000	2648.311	3496.934
season_spring	-1539.1340	118.741	-12.962	0.000	-1772.424	-1305.844
mnth_mar	562.6487	157.010	3.584	0.000	254.170	871.127
mnth_oct	810.7397	133.176	6.088	0.000	549.089	1072.390
mnth_sept	706.5239	132.487	5.333	0.000	446.228	966.820
weathersit_cloudy	-687.1111	78.386	-8.766	0.000	-841.117	-533.105
weathersit_rainy	-2417.2048	221.832	-10.897	0.000	-2853.039	-1981.371

Omnibus:	94.106	Durbin-Watson:	2.107
Prob(Omnibus):	0.000	Jarque-Bera (JB):	195.618

General Subjective Questions

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

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

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

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions, so they look completely different from one another when you visualize the data on scatter plots.

3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. The target is to fit all variables on same scale like between 0 and 1. Virables

visualization can't be understood if different variables use different scales, so scaling normalize all variables into a standard scale.

Normalization also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

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

Answer: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Answer: Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line.

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.