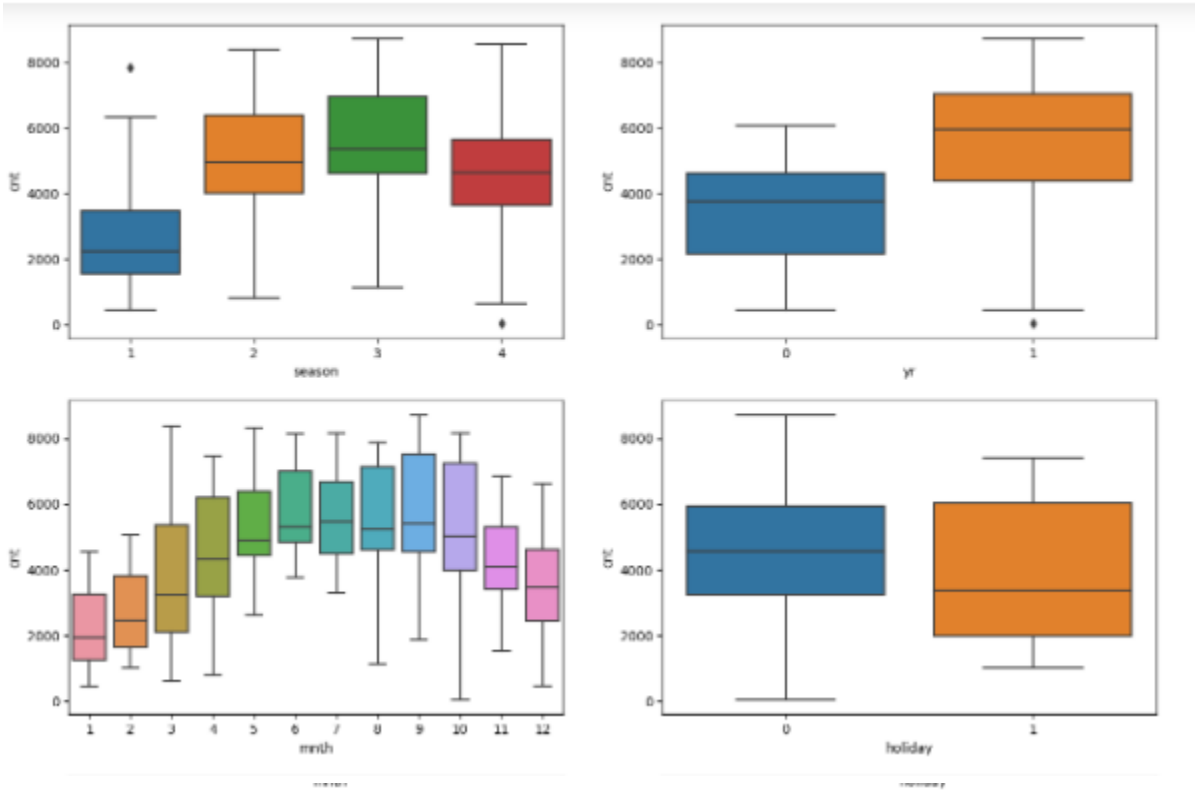
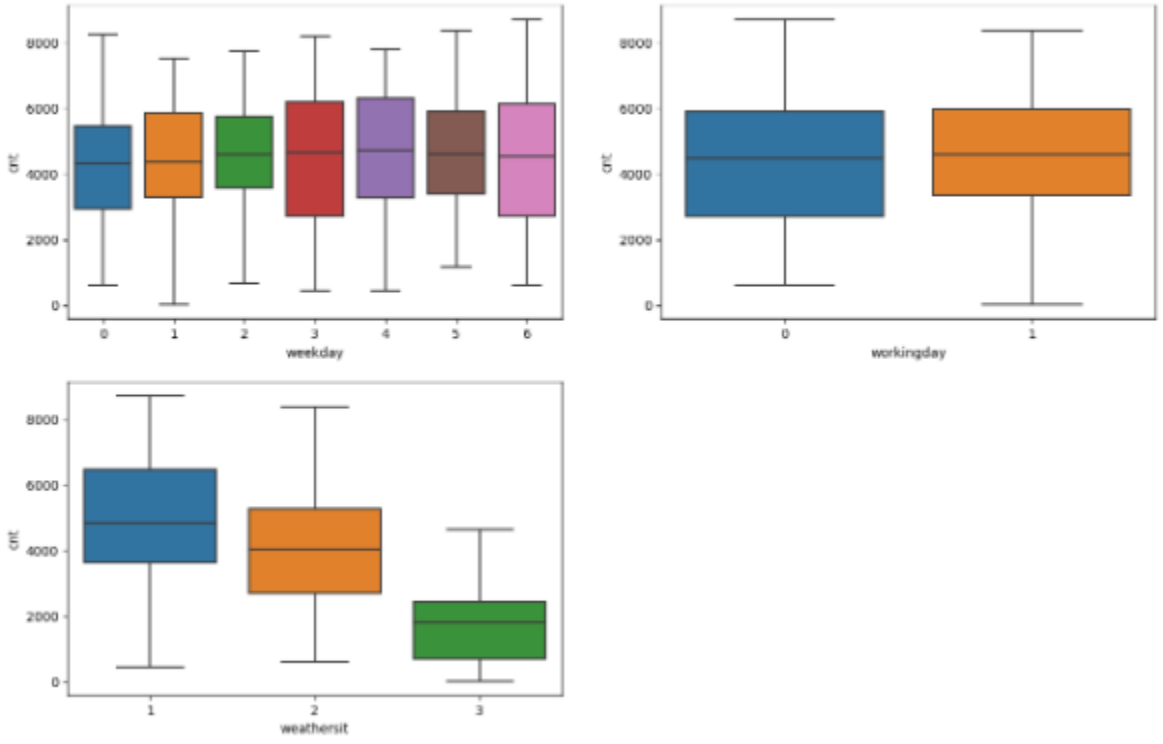
Assignment-based Subjective Questions

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

   a. As it can be seen below from boxplots, there are some variables which have considerable impact on the dependent variables. Specially Seasons, yr, month and weather has most impart on the number of bikes rented.

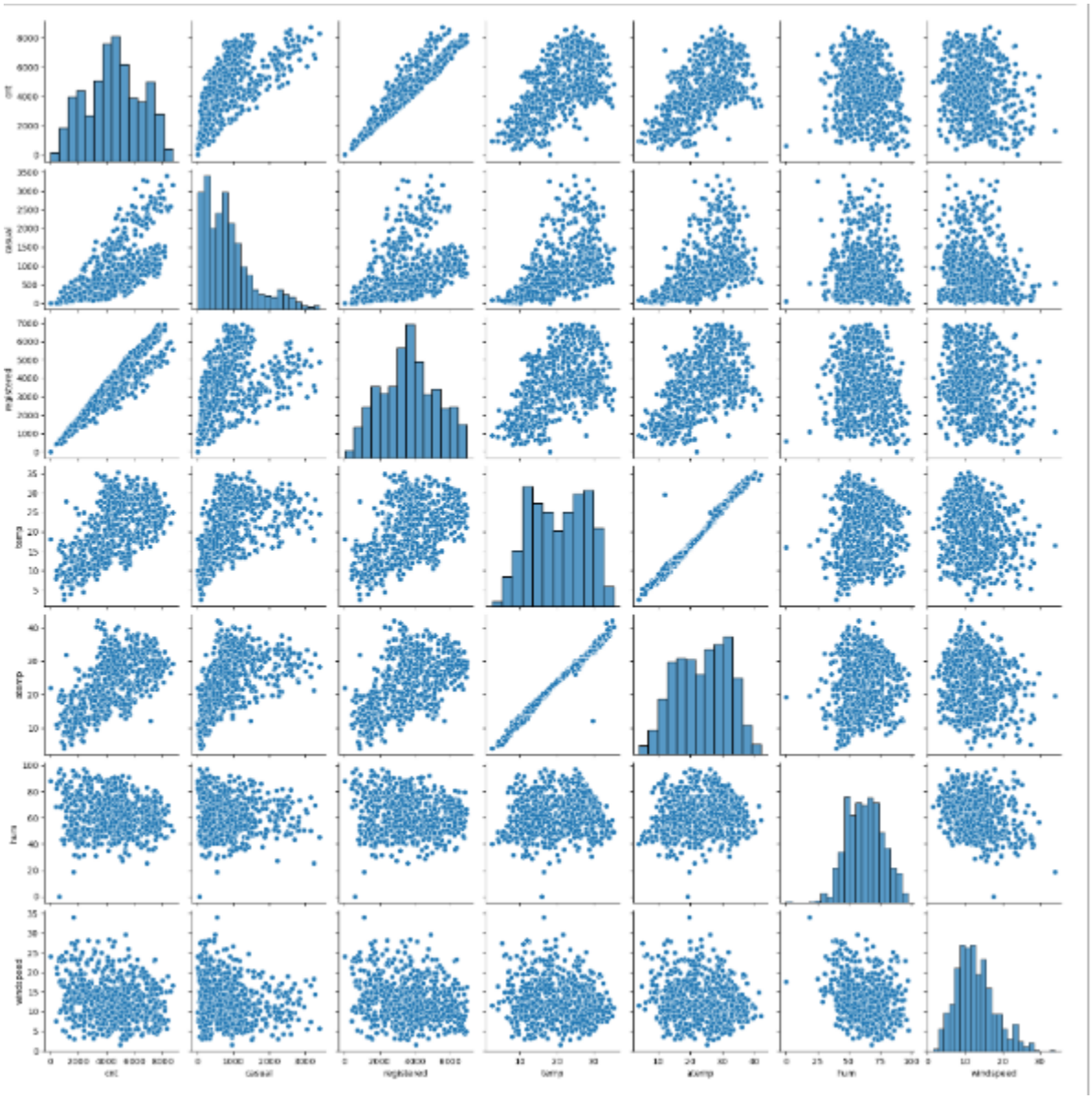


   b.

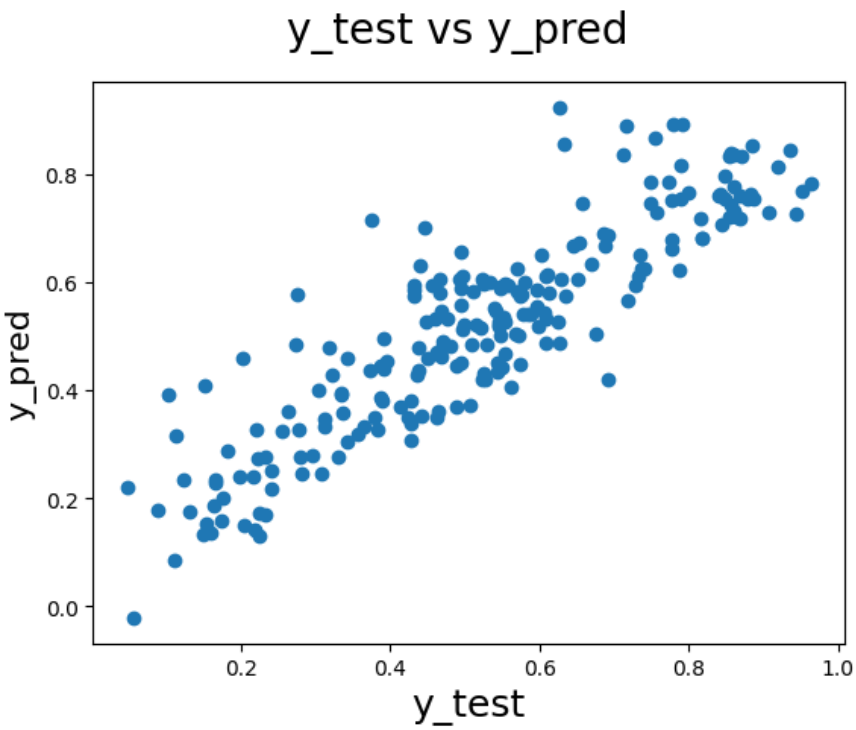


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    a. While creating dummy variables there could be columns which can be inferred from the other available columns, and hence to optimize the size of the data and improve model building efficiency this extra column is dropped from the dataset.

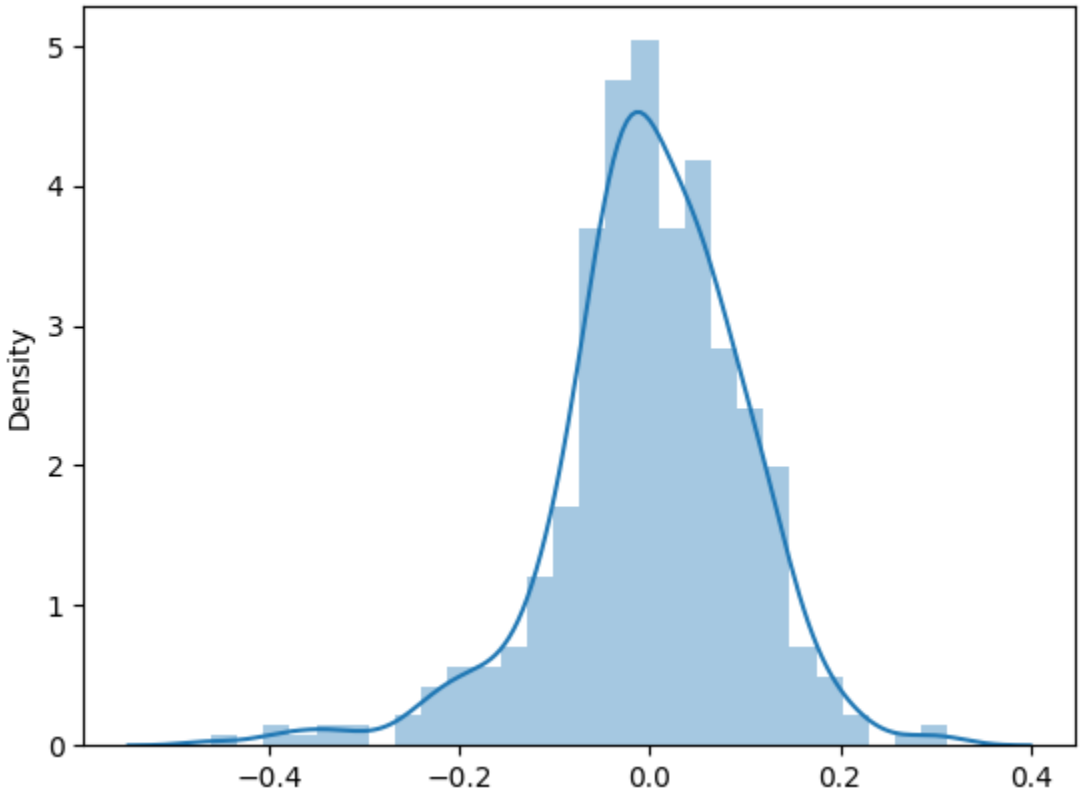
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. Looking at the pair-plot, we can see that temp, atemp registered have highest correlation.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Following points were confirmed for Liner regression.
        i. Its Linear – there is a linear relationship between independent and dependent variables.
        ii. There is no multicollinearity in the data used for building the model.
        iii. Residuals are in equal variance.



iv. Predictors are distributed normally.



v.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   a. As can be seen below, atemp, yr and season are the major contributors.

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.803
Model:	OLS	**Adj. R-squared:**	0.799
Method:	Least Squares	**F-statistic:**	226.3
Date:	Thu, 10 Aug 2023	**Prob (F-statistic):**	4.80e-170
Time:	09:53:18	**Log-Likelihood:**	453.03
No. Observations:	510	**AIC:**	-886.1
Df Residuals:	500	**BIC:**	-843.7
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1778	0.032	5.565	0.000	0.115	0.241
season	0.1295	0.013	9.869	0.000	0.104	0.155
yr	0.2353	0.009	26.088	0.000	0.218	0.253
holiday	-0.0678	0.029	-2.315	0.021	-0.125	-0.010
weekday	0.0481	0.013	3.585	0.000	0.022	0.075
workingday	0.0198	0.010	2.014	0.045	0.000	0.039
weathersit	-0.1561	0.021	-7.461	0.000	-0.197	-0.115
atemp	0.5104	0.023	21.845	0.000	0.465	0.556
hum	-0.1028	0.041	-2.492	0.013	-0.184	-0.022
windspeed	-0.1507	0.028	-5.306	0.000	-0.207	-0.095

Omnibus:	63.967	**Durbin-Watson:**	2.031
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	153.102
Skew:	-0.658	**Prob(JB):**	5.68e-34
Kurtosis:	5.339	**Cond. No.**	21.0

b.

General Subjective Questions

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

   a. 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. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

b. Assumption for Linear Regression Model - Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

   i. Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
   ii. Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
   iii. Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
   iv. Normality: The errors in the model are normally distributed.
   v. No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. Explain the Anscombe's quartet in detail. (3 marks)
   a. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
   b. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R? (3 marks)
   a. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
      i. If r = -1, then there is a perfect negative linear relationship between x and y.
      ii. If r = 1, then there is a perfect positive linear relationship between x and y.
      iii. If r = 0, then there is no linear relationship between x and y.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   a. It is a step of data Pre-Processing 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.
   b. 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.
   c. Normalization/Min-Max Scaling, brings all of the data in the range of 0 and 1.
   d. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

a. An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.
b. Identify those variables and try dropping them.

6. What is a Q–Q plot? Explain the use and importance of a Q–Q plot in linear regression. (3 marks)
    a. The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.