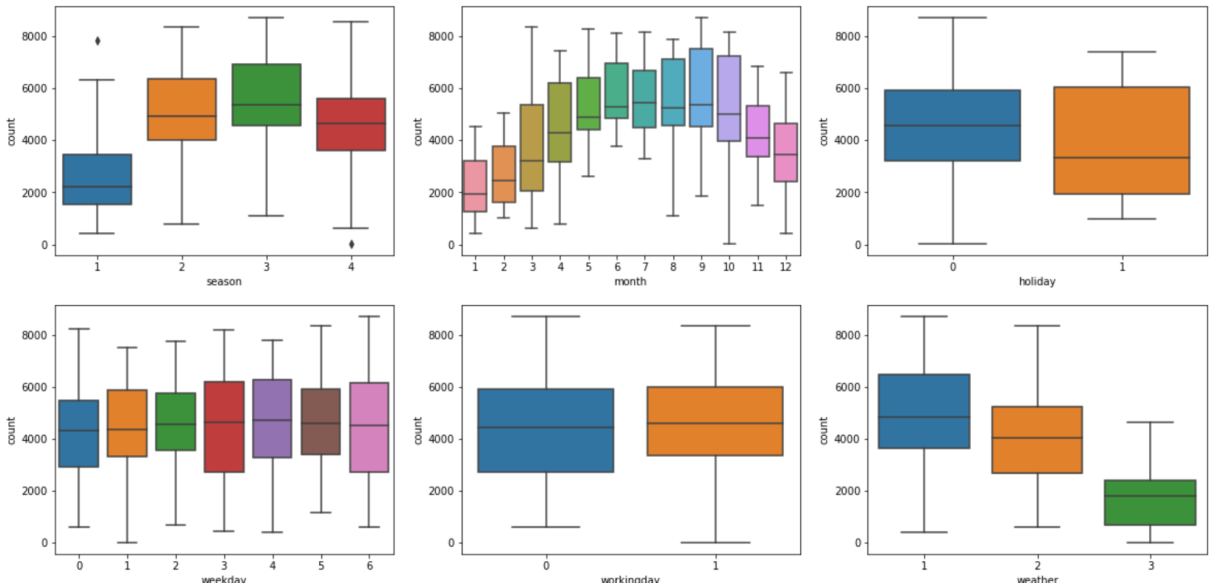
Assignment-based Subjective Questions

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

   Categorical variables impact the target variables either in positive or negative ways.

   In this analysis we proved that temperature, season, and month variables affect the bike rentals positively but weather and Holiday affects the bike rental negatively.

   Have a look at the categorical variable graph and notice that each variable has a dynamic impact on target value count.



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

   When preparing the dummy variable one of the cases will be with none of the above all of the above cases , so we can assume that case as the first column. So as a standard dummy variable preparation is always N-1 columns.

   Take an example:

	season_1	season_2	season_3	season_4
0	1	0	0	0
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0

Take a look at the above table and see that if all the other columns are zero means the first column will be 1 , that's sure and that's the way dummy variables work . So it's safe to remove the first column.

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

   Initially the registered variable got the highest correlation value as per the pair plot. But at a later point of time temperature got the highest correlation value after applying the Multiple Linear Regression methods.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - Error terms should be normally distributed
   - Error terms are centered at zero
   - Error terms have constant variance
   - Model fits a hyperplane instead of line
   - Coefficient obtained by minimizing the sum of Squared error applies to the least squares criteria.
   - Make sure no overfitting in the model and eliminate the overfitting variables.
   - Look at the F-static and determine the overall significance of the model fit.
   - Correlation coefficient specifies how strong is the relationship between variables , high values indicate the strong relationship.
   -
   -

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

- Temperature
- Season
- Weather

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression one of the type in Regression Machine learning Algorithm supervised learning method which is continuous Label based.
   This approach is most commonly used for predicting and determining cause-and-effect correlations among variables. When the dependent variable is of continuous data type, regression can be used, irrespective of the data type of the predictors or independent variables. The regression approach aims to identify the best fit line that accurately depicts the connection between the dependent and independent variables. The most fundamental type of regression analysis is linear regression, which is a supervised machine learning algorithm. The Linear Regression assumes that there is a linear relationship between the dependent variable and independent variables, the accuracy of train data and test data are almost equal so that there is not over/under fitting, and the error terms are normally distributed to ensure there are no non-linear relationships in the data. When all these conditions are satisfied, LR can be applied to develop a model for accurate predictions of the target variable. It strives to find the best fit line in regression to explain the connection between the predictors and the predictive variable. In linear regression, the output variable is formulated as a function of the independent variables, their coefficients, and an error term of regression as $y = \beta 0 + \beta i Xi$. where Xi is i^th independent variable in the training input data, β0 is the y-intercept, βi is the coefficient of i^th independent variable in the input X data, and y is the predicted target variable. The model seeks to predict y value in such a way that the error difference between predicted and real value is as little as possible by reaching the best-fit regression line. When there is only one predictor variable a Simple Linear Regression model is created, whereas when multiple predictor variables are present Multi Linear Regression is used. In MLR, there are a few more assumptions that need to be satisfied in addition to the three specified in SLR. They are the absence of Multicollinearity, Homoscedasticity of residuals, Sample size and Categorical Variable Conversion. Multicollinearity condition requires no two independent variables to be highly correlated. Homoscedasticity of residuals means the variance of errors must be constant across all independent variables.

.

2. Explain the Anscombe’s quartet in detail .

    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.

    Each dataset consists of eleven a (X,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

3. What is Pearson’s R?

    It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

    Scaling is a normalizing technique used to rewrite the data in a certain range instead of having different big values for each variable.

    Normalization Scaling is done to bring all data into a range of 0 and 1 so that all variables under consideration are in the same range.

    MinMax Scaling x = x - min(x) / max(x) - min(x)

    In Standardized scaling, the data is converted into a standard normal distribution which has zero mean value and 1standard deviation, that is all data values are replaced with their Z scores.

    Standardization x = x - mean(x)/ sd(x)

    NOTE: Standard scaling will affect the values of dummy values but MinMax scaling won't. MinMax handles dummy variables effectively .

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

   When VIF is infinite, it means that there is perfect correlation between the two independent variables under consideration. That's the R2 will be one with these two variables in the model. The equation of VIF is: VIF = 1/(1-R2) So, when R2 = 1; the equation becomes VIF = 1/(1-1) = 1/0 = Inf In a data when two independent variables are highly correlated, it's said to have multicollinearity . Thus, to overcome multicollinearity, one of the variables under consideration that produces infinite VIF must be removed, then the other variable will have a finite VIF.

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 tool for detecting if two data sets are from the same population. A Q-Q plot is a comparison of two dataset's quantiles. Quantiles are cut points that divide the range of a probability distribution into equal-probability continuous intervals or the observations in a sample in the same way. Q-quantiles are values that divide a finite collection of values into almost equal-sized Q subgroups. There exists one Q-quantile for each integer k that is satisfied by 0 < k < Q, resulting in Q-1 total quantiles. Continuous distributions also can use quantile to apply rank statistics to the continuous data.