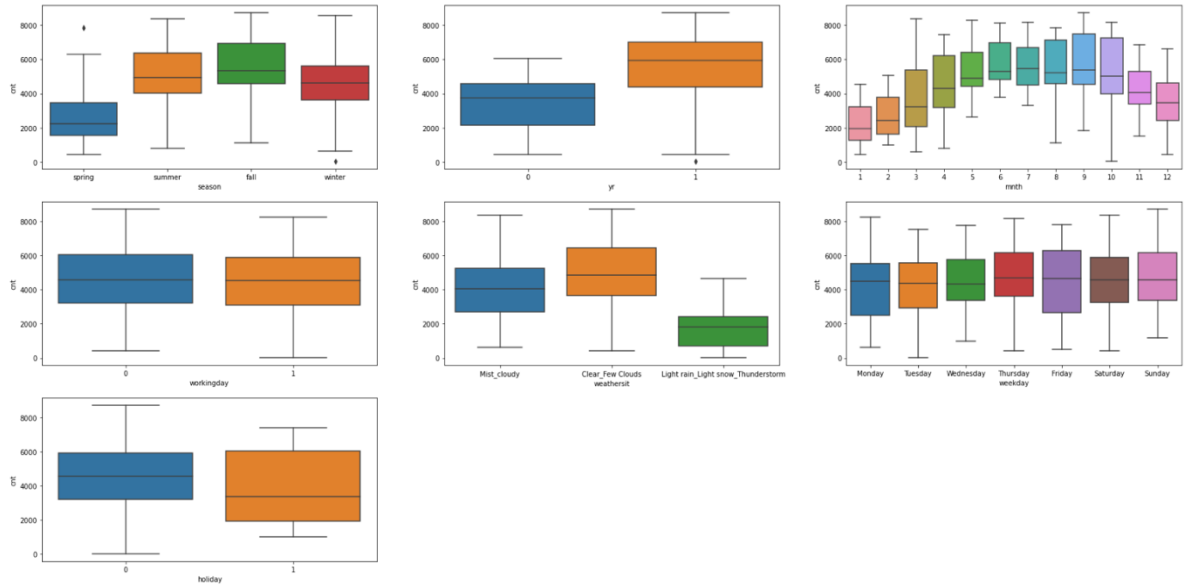
Assignment Questions

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)

From the dataset categorical variables were found to be ‘season’, ‘workingday’, ‘mnth’, ‘weekday’, ‘weathersit’, ‘holiday’ and ‘yr’. We found that the dependent variable is most dependent on ‘season’, ‘weathersit’ and ‘month’ regarding the highest user numbers. We find few outliners in their graphs.

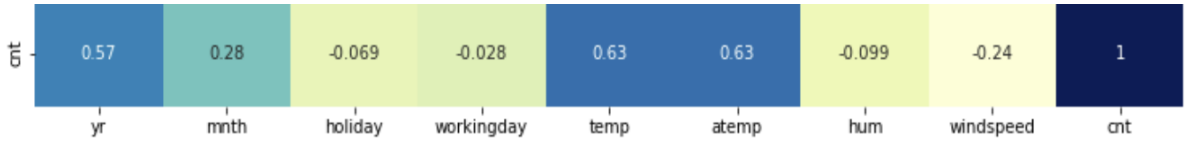


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Based on the pair-plot numerical variables, the highest correlation with the target variable is ‘atemp’ & ‘temp’ with 0.63 and the next is ‘yr’ with a correlation value of 0.57.

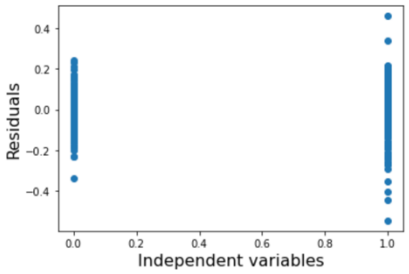


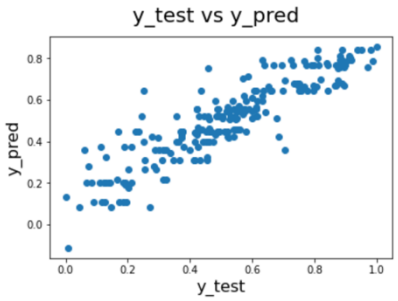
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The assumptions of Linear Regression after building the model on training set was validated as follows:

- Linear relationship between the predicted data (y_pred) and test data (y_test) through scatter plot. We also check the R-squared value of predicted data (0.794) and test data (0.798) which is very close.
- We plot the residuals to check for autocorrelation and homoscedasticity. When we have time series data (e.g. yearly data), then the regression is likely to suffer from autocorrelation because demand next year will certainly be dependent on demand this year. Hence, error terms in different observations will surely be correlated with each other.

Assignment Questions





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 chosen, the three main features contributing significantly towards explaining the demand of the shared bikes are- weathersit, mnth, weekday and holiday.
Demand increases in the 'mnth': 3,5,6,7,8,9,10 and 'Sunday'.
Demand decreases if it is 'holiday', 'Spring', 'Light rain_Light snow_Thunderstorm', 'Mist_cloudy' or 'Monday'.

General Subjective Questions

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

Linear Regression Algorithm (LRA) is a machine learning algorithm based on supervised learning. It performs the task to predict a dependent/target variable value (Y) based on independent variable (x). Mathematically, we can write a linear regression equation as:
$Y = a + bx$,where a and b are given by the formulas:
b (slope) $= [n\Sigma xY – (\Sigma x)(\Sigma Y)] / [n\Sigma x^2 – (\Sigma x)^2]$
a (intercept) $= [n\Sigma Y - b(\Sigma x)]/n$

Some assumptions of LRA are:
Linear relationship between x and Y. It can be checked through a scatter plot of the model.
The data is normally distributed (homoscedasticity) and can be checked through a histogram plot.
To have minimised collinearity, checked through the VIF value of the model.
Error terms/residuals – to have no autocorrelation, constant variance and must be normally distributed. Can be checked through QQ plot.

R-square – The variance in 'Y' is being explained by the 'x' variables. If a new variable is added and the variance remains same, there is no need to consider that variable as it has no significant impact. Disadvantage is it never decreases, it remains same or increases on addition of new variable 'x'

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

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. They were constructed by statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it and the effect of outliers and other influential observations on statistical properties. It also highlights the importance of plotting data to confirm

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

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

Assignment Questions

variables. A positive correlation signifies that if variable A goes up, then B will also go up. Whereas if the value of correlation is negative, then if A increases, B decreases.

$R = (n (\sum xy) - (\sum x)(\sum y)) / (\sqrt{([n\sum x^2 - (\sum x)^2] [n\sum y^2 - (\sum y)^2]})$

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

Scaling refers to the means with which we deal with the categorical variables in LR model building. Scaling affects only the coefficients and none of the other parameters such as T-statistics, F-statistics, R-squared value and so on.

Normalised Scaling	Standardised scaling
We map the feature value to a predefined level of 0 to maximum 1. Hence, the feature values are mapped into the [0,1] range. $z = (x - \min(x)) / (\max(x) - \min(x))$	In standardisation, we don't enforce data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1. $z = (x - \mu) / \sigma$ It also helps centralise the data.

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. This shows a perfect relationship between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ = infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

A Q-Q plot are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on the line but not necessarily on the line y=x. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.