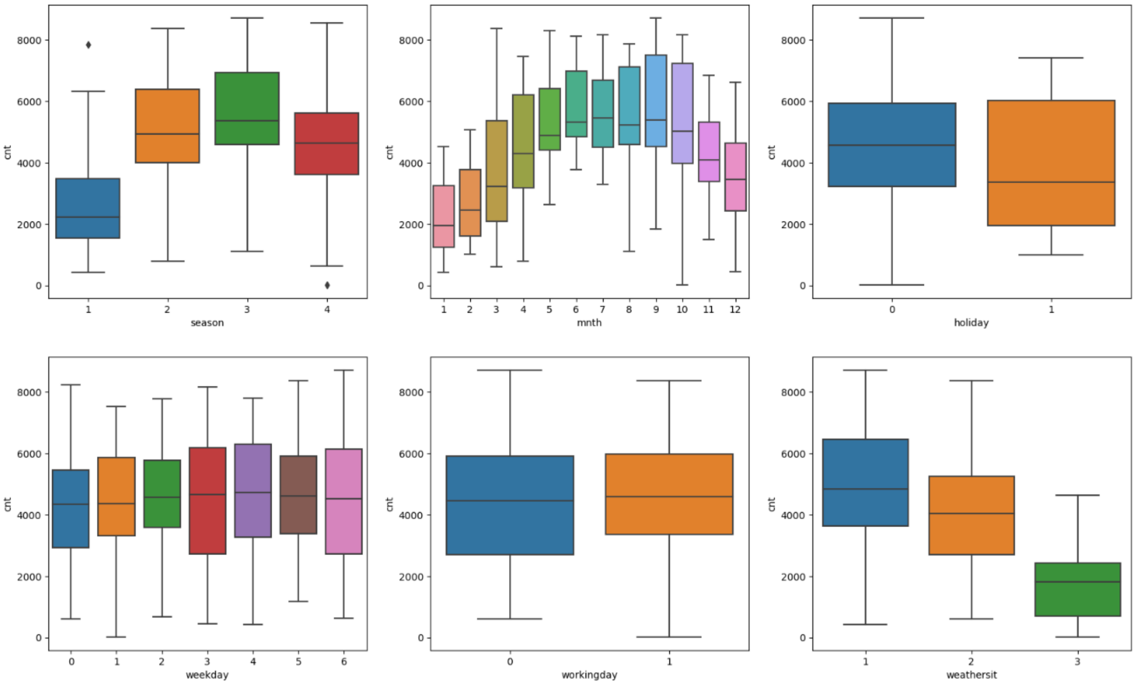
Assignment-based Subjective Questions

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



(image ref: assignment notebook)

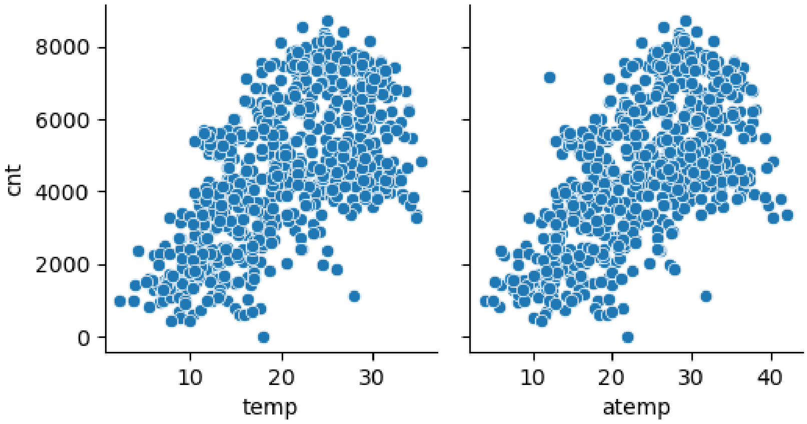
- The weather situation value is never 4, It can indicate either missing data or the weather was never type 4 (Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog) in the recorded two years dataset.
- Fall seems to have the highest rentals across the season. Which also correlates to the corresponding months having higher rentals
- There are more rentals recorded when the weather is clear.
- Holidays have slightly lower avg rentals than non holidays. Which shows a similar pattern in working days having slightly higher rentals than that of non working days.
- The avg rentals are more or less similar across the days of the week.

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

If we don't use drop_first we get the number of columns the same as the possible values for the categorical column. Which can be reduced to n-1 by dropping the first column, When all the other columns are `0` it can be inferred as the dropped column. Model would still be able to learn the effect of that column this way.

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

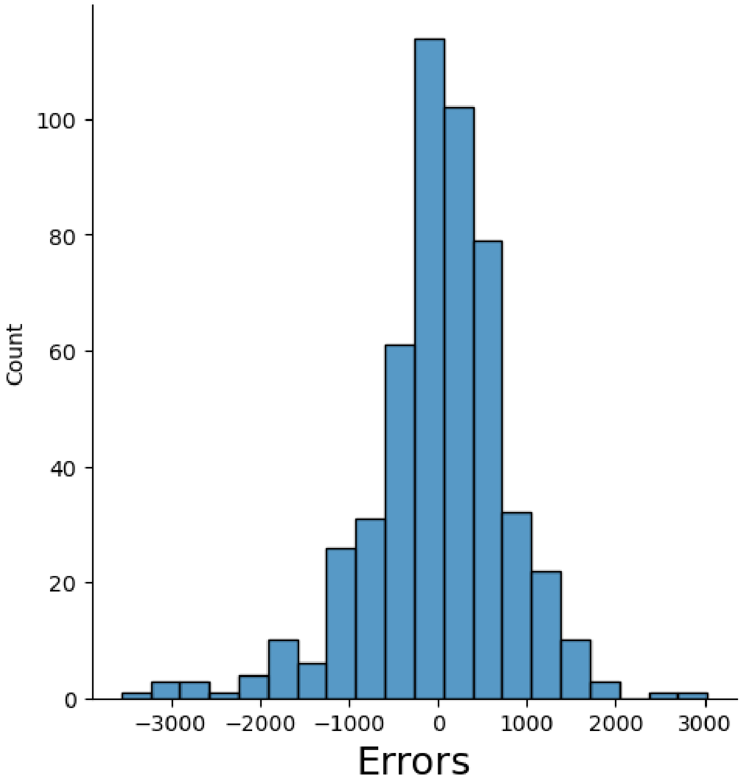
Temperature variables seem to have a good correlation with the target variable.



(image ref: assignment notebook)

On checking the correlation values within the training data, `***atemp***` seems to have slightly higher as 0.65.

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



(image ref: assignment notebook)

As part of residual analysis, I plotted the errors resulting in a normal distribution as expected for the linear regression model.

A scatter plot of residuals can't be made because there are multiple variables in the X_train.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
`temp` is the most significant feature contributing towards the demand of the shared bikes, followed by `yr` (with a positive effect in the 2019 year) and `weather_snowy` (with a negative effect on the shared bikes). These are based on the coefficients from the trained model. Higher the absolute value of the coefficient, higher the significance of it. Since linear regressions is essentially a weighted avg of the independent variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a way to find a linear equation that can explain the dependent variable. Regression algorithms try to reduce the error on residuals on the target variable and find a best fit line as close to all the target variables as possible. This works best for when there is a linear relation between target variable and independent variables.

$y = \beta 0 + \beta 1 * x$, y = target variable, x = independent variable, Betas are the coefficient and intercept defining the line equation that will be calculated as part of training.

In a multivariate linear regression, linear regression tries to find a hyperplane that represents a weighted avg of the independent variables finding the best fit that has the least sum of squares of residual errors.

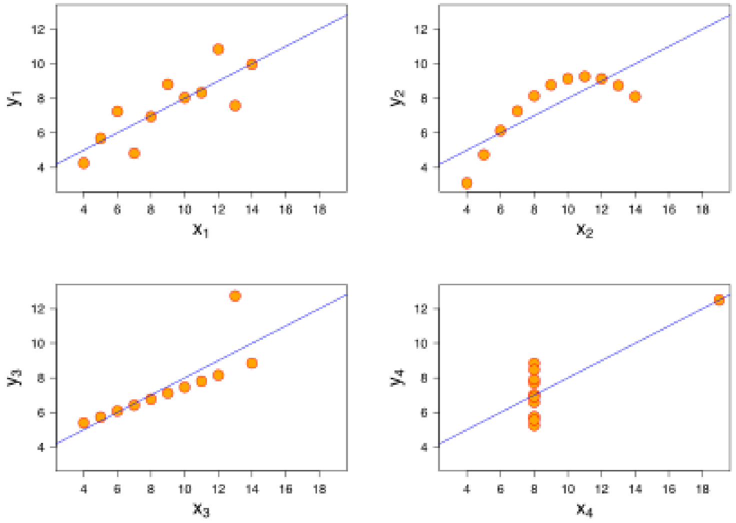
$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + .. + \beta n * xn,$ y = target variable, xi = independent variables, Betas are the coefficients defining the line equation that will be calculated as part of training.

Gradient descent is one way to find the best-fit for linear regression. In this algo, at every step we find a gradient and adjust coefficients to slide towards a minimal point so as to approach a local/global minima for the linear equation.

2. Explain the Anscombe’s quartet in detail.

It is a set of four datasets created by Francis Anscombe that looks similar from a statistical metrics perspective like mean, variance, correlation and the linear regression model. But in reality not all the models fit a linear pattern and are either skewed by outliers and/or nonlinear patterns that linear regression might not do the full justice to.

As shown in the image below, the unfitness of the Linear regression model can be clearly pointed in the datasets 2, 3 and 4 when plotted as a scatter plot. So, It is always better to plot the data where possible to determine the quality of the predicted models and



(image ref: Wikipedia)

3. What is Pearson’s R?

Pearson’s R (aka Pearson’s correlation coefficient) is one way to determine correlation between two normally distributed datasets. The values for the correlation coefficient lies in between -1 and 1.

R close to -1 implies that the variables have a negative effect on each other (one increases as other decreases),
R close to 1 implies that the variables have a positive effect on each other (one increases as other increases).
R closer to 0 implies that the variables don’t have much relation with each other.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is to transform values for variables in the training data to a similar range of values. It helps transform the independent variables to similar scale of values while maintaining their data characteristics. A regression model can be trained quickly on a scaled datasets and can help achieve a minima point in the hyperplane faster.

Normalised scaling is where we use maximum and minimum values of the dataset to scale the values in the range of [0, 1]. A standardised scaling is where we use mean and standard deviation to adjust the values in the datasets according to them. Standardised scaling assumes the dataset is in a normal distribution.

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

A very high VIF indicates that there is a high multicollinearity among variables and infinity indicates that a certain variable is having 1 correlation with some other variables. In other terms, other variable(s) might be perfectly explaining the variable with infinite VIF. This could possibly happen on a derived metric variable which is completely derived from other independent variable(s) in the dataset.

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

A Q-Q plot is a quantile-quantile plot used to compare two datasets visually to answer questions like, Comparing distributions of datasets in the same plot, Comparing scale and location for the datasets etc. To plot this, we determine the values at quantiles(percentiles) of both the datasets and plot them one on each axis. Since we are dealing with quantiles, we can still plot the datasets with an uneven number of data points.

In Linear Regressions, When training and test datasets are received separately, this Q-Q plot can be used to compare the distributions of both the datasets and determine if test data is a good comparison to use on the model trained on given training data.