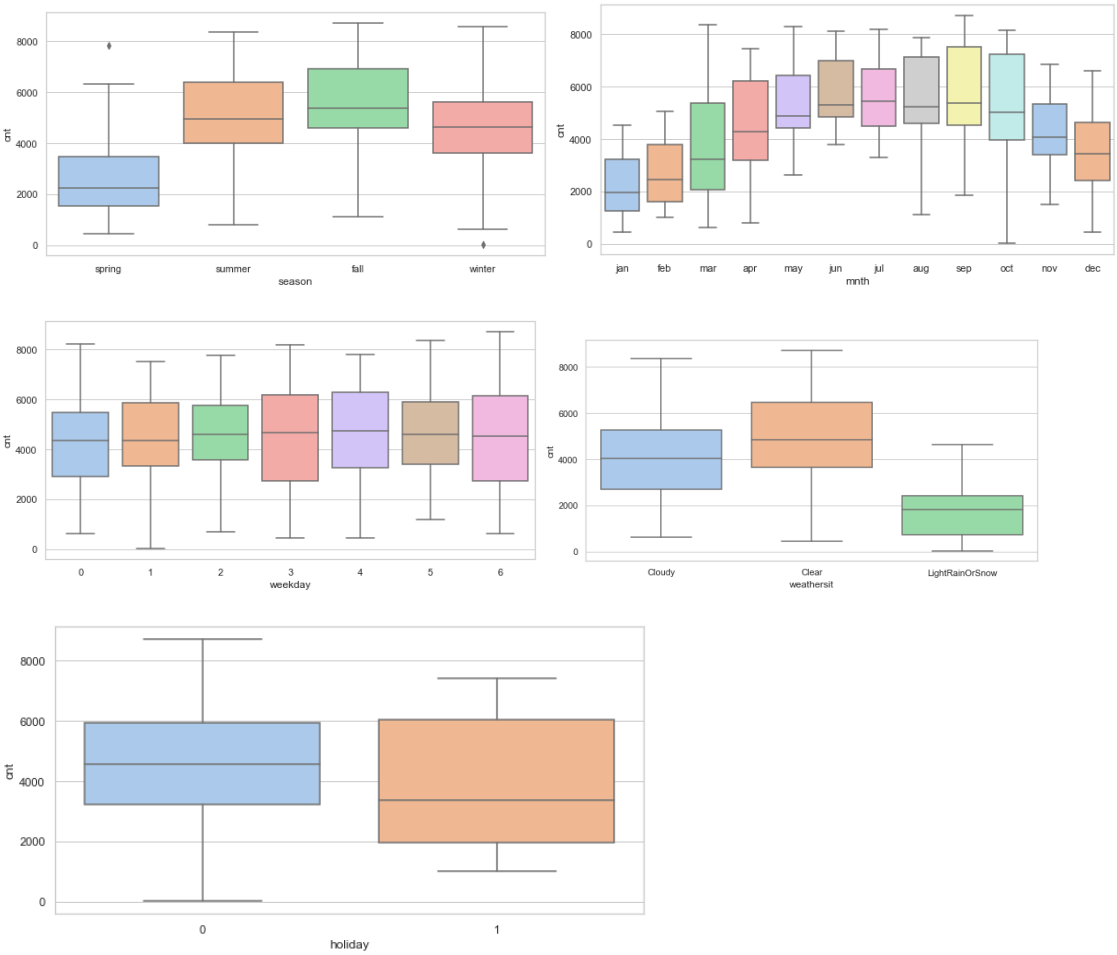
Assignment-based Subjective Questions

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



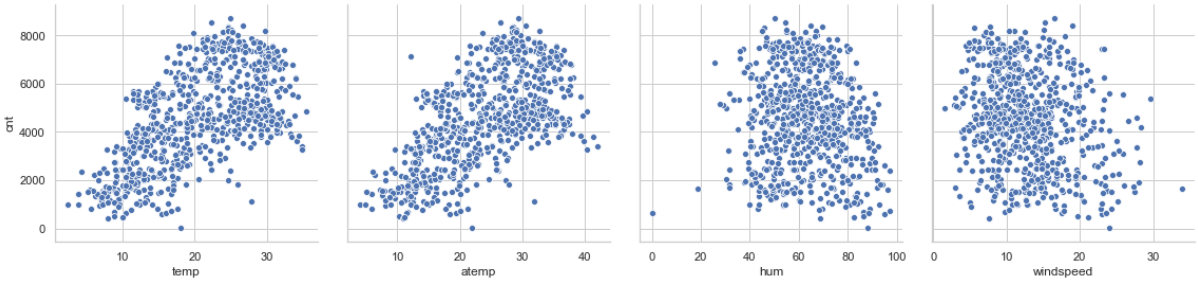
The boxplots show the effect of categorical variable over bike rental.

- Bike rental is higher in June to September and i.e. fall.
- As expected, bike rental is much lower if it is raining or snowfall. Clear weather has highest median rental count.
- In holidays, rental count is lower.

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

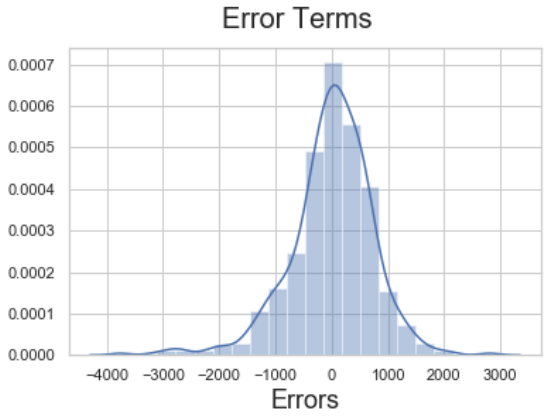
Dummy variables are made from a categorical variable where each unique value can take 0 or 1 (like '0' = no, '1' = yes). But one unique value can be calculated if rest of the values are 0 (No) as all the dummy variables are mutually exclusive. If a categorical variable has n number of unique values, it can be expressed with (n-1) number of dummy variables where first unique value can be identified if all other values are 0.

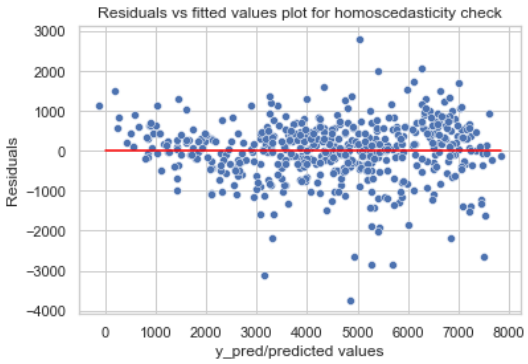
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



Temperature and feel temperature has the highest correlation as depicted in this figure (Correlation coefficient +0.63)

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





A. Linearity – dependent and independent variables had linear relationship – evident from scatterplot shown above

B. The residuals had a normal distribution with mean of 0 as shown in the figure.

C. Residual vs fitted value shows no pattern in variance i.e. heteroscedasticity

D. **Goldfeld Quandt Test** - to check for heteroscedasticity showed a non-significant p-value

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 final model the top 3 features were –

- Temperature,
- Weather – Light Rain or Snow
- Year - 2019

General Subjective Questions

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

Linear regression is an ML algorithm where a linear relationship is established between dependable and independent variables using best straight-line fitting method. Here, dependent variable is a continuous variable and we try to predict its value using one predictor variable (Simple Linear Regression) or multiple predictors (Multiple regression). The formula is like –

$y = aX + c$ (Simple Linear regression)

$y = aX_1 + bX_2 + ... + zX_n + c$ (Multiple regression)

To find the best fit line, we calculate the distance from data points to the regression line, square it, and sum all of the squared errors together. This is called Residual Sum of Squares (RSS). It is done to avoid positive and negative distances negate each other. This RSS value is then divided by number of data points to get Mean Square Error (MSE). Minimizing the MSE is the method of optimization adopted during modelling.

Before fitting and interpreting a linear regression model, we should look for possible violations of assumptions like –

- linear relationship graphically between variables
- presence of multicollinearity
- heteroscedasticity
- residuals must be normally distributed with mean 0

It is not advisable to include all available parameters too fit a model. There is a method called adjusted R^2 which helps us determining optimum number of variables.

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

Anscombe's quartet is a set of four datasets in the form of x-y pairs. The number of data pairs is four, so it was named as quartet. The datasets have eleven points with similar summary statistics like mean, variance. When a regression line is drawn with these datasets, the lines have almost same gradient and constant value. But if they are plotted as scatterplot, the differences between their distribution becomes apparent.

This demonstration shows why it is always advisable to check the data graphically before concluding with statistics alone. Since its publication in 1973, several similar datasets have been developed.

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

The Pearson correlation coefficient is a statistical method to find the strength of association between two variables and is denoted by r. The range of r is from -1 to +1. Here the value indicates the strentgth and +/- sign denotes the direction. There are a few assumptions to be met before calculating Pearson's r -

- Both variable should be continuous scale
- always they must be in 'pair'
- both the variable should be distributed normal (gaussian). If not, we may need to use Spearman's rho/ Kendall's tau

- relation should not linear
- there should not be Heteroscedasticity

finally, we have to remember, Pearson's R tells only about correlation. We can not assume cause-effect relation with this value.

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

- Scaling is a statistical method to convert continuous variables into almost the same scale, causing features equally essential and making it easier to process by most ML algorithms.
- Often dataset contains columns with very narrow range (say 0 to 5) and some other columns with huge range (say 0 to 5 million). This variability of range causes a few problems like –
    - Feature with higher range gets more importance in model building
    - ML algorithms which use gradient descent as an optimization technique particularly suffers while selecting step size for each feature
- Normalization vs Standardization:
  Normalization is a scaling technique where values are converted between 0 to 1 where 0 indicates minimum and 1 indicates maximum value. The formula is-
  $X' = (X- X_{min})/(X_{max} – X_{min})$

  Standardization is another scaling technique where values are centred around the mean with a unit standard deviation.
  $X' = (X – mean)/SD$
  This method has no bounding limit like 0 and 1. So, outlier values retains their property.

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

As per the formula of VIF –

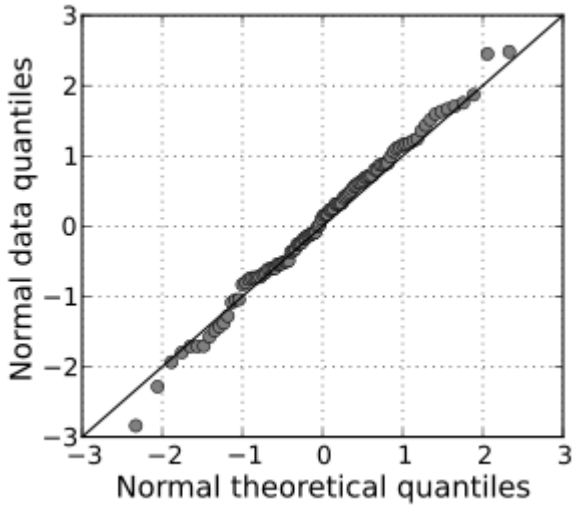
$VIF = 1/(1-R^2)$

where R2 is the coefficient of determination of the regression equation where one predictor variable is on the left-hand side, and all other predictor variables (all the other X variables) on the right-hand side.

Now, if one variable can be 100% accurately calculated with other predictors, then R^2 becomes very close to 1 and $(1- R^2)$ becomes almost 0. In this scenario, VIF is infinite and it indicates presence of severe multicollinearity.

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

QQ plot is a visual way to compare a given dataset with a desired distribution. It evaluates the quantile values between two distributions. Usually, we use this plot to check whether our dataset follows gaussian, exponential or other distribution. the Q–Q plot follows the 45° line (i.e. y = x), if the two distributions being compared are identical as shown below –



(Source of picture – Wikipedia)

The same principle can be applied to check whether two datasets are from same population or not. In linear regression if we can compare one feature from train and test data using QQ plot to test whether they have common distribution.