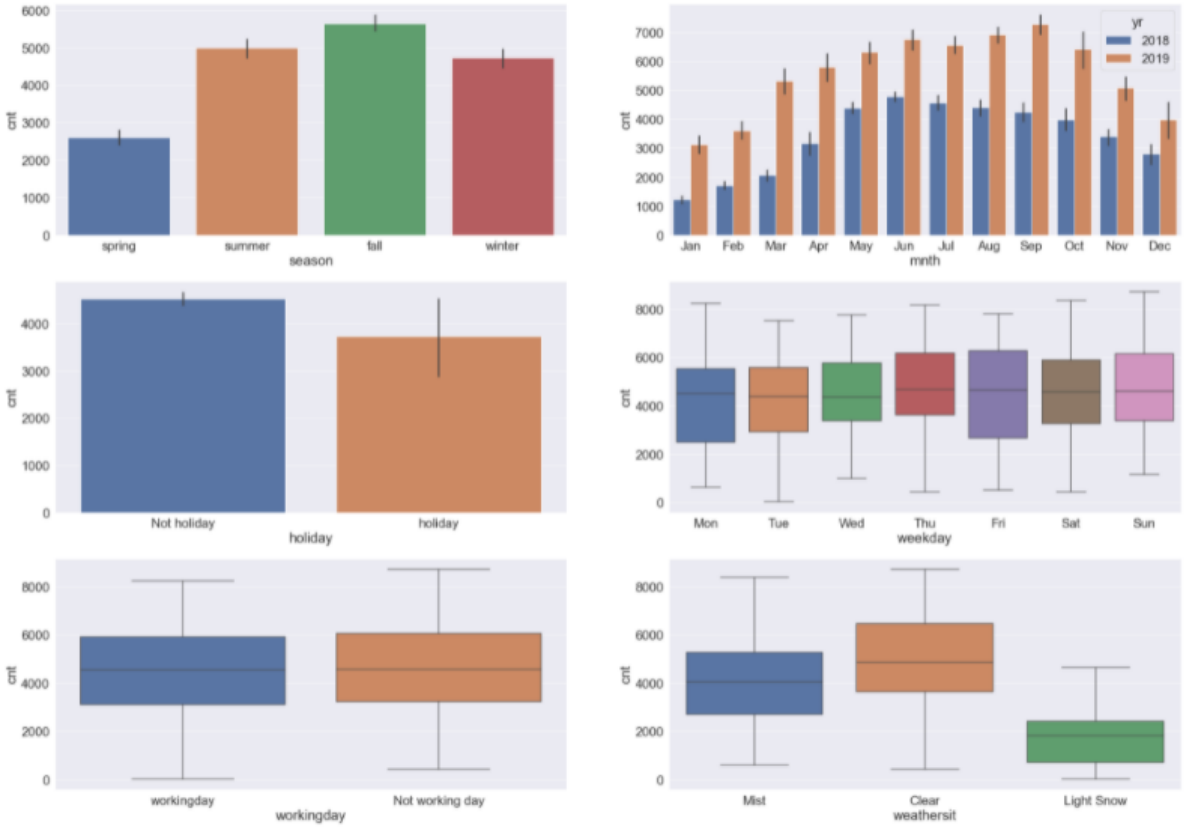
Assignment-based Subjective Questions

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 below graphs have been created using the dataset. The inferences are mentioned below:



- ***Season***: Highest demand in "Fall" and lowest in "Spring". This shows, the demand increases after "Spring", keeps increasing in "Summer", reaches peak in "Fall" and then starts decreasing "Winter" onwards
- ***Month and Year*****:** Highest demand can be observed in the month of September. Also, 2019 has higher demand.
- ***Holiday*****:** On holiday, the demand decreases
- ***Weekday*****:** Demand is almost same every day.
- ***Working day*****:** Demand is same, be it working or non-working day
- ***Weather*****:** Higher demand on Clear weather days and minimum when there is Light Snow.

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.

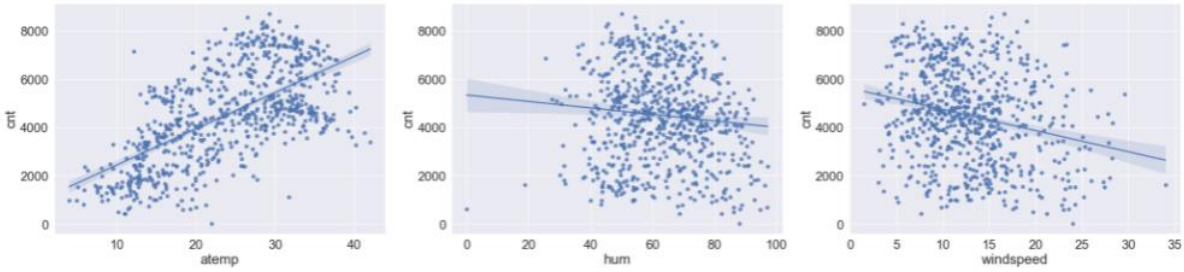
   For example: In the given dataset we have some variables like season, month, year etc, where we need to create the dummy variable columns.

   For n-levels, we need to create n-1 columns. Like for month, the levels are 12 (since we have 12 months), the dummy columns created will be 11.

   Here we drop "Jan", and create column for all other months, with values as 0 or 1 (0 as FALSE and 1 as TRUE). In case, for a specific row, values are 0 in all months, then that would be considered as Jan.

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

   Among the numeric variables, "atemp" as the highest and positive correlation with the Target variable "cnt"



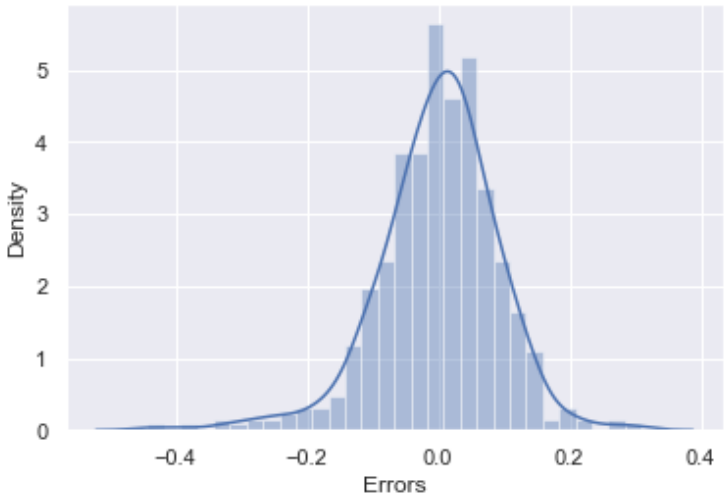
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   While building the model, we kept checking p-values and VIF. We kept dropping the variables with p-values > 0.05 (which was none), and VIF > 5.

   Finally we got a model with all variables having p-values less than 0.05 and VIF less than 5.

   We also used RFE, and did Residual Analysis which gave a Normal distribution curve for errors (as shown in the graph)

   Finally, the R-squared values for train and test data set were 0.835 (83.5%) and 0.807 (80.7%) respectively. This shows that model is accurate.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Final model equation is as follows:

   cnt = 0.209 + atemp(0.444) + windspeed(-0.149) +season_summer(0.073) + season_winter(0.122) + yr_2019(0.235) + mnth_Aug(0.057) + mnth_Dec(-0.056) + mnth_Feb(-0.055) + mnth_Jan(-0.086) + mnth_Nov(-0.041) + mnth_Sep(0.104) + holiday_holiday(-0.085) + weathersit_Light Snow(-0.286) + weathersit_Mist(-0.081)

   From this equation, the tope 3 features are:
   1) atemp – which has high positive correlation with cnt. On increase of 1 unit of atemp, the cnt will increase by 0.444
   2) yr_2019 – High positive correlation with target variable
   3) weathersit_Light Snow – High negative correlation. Increase in this variable will result in decrease of target variable.

General Subjective Questions

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

Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one response variable (denoted by y) and one or more explanatory variables (denoted by X1, X2, X3…. Xn).
The response variable will be dependent on how the explanatory variables changes and not the other way round. Response variable is also known as target or dependent variable while the explanatory variable is known as independent or predictor variables.

There are 2 types of Linear Regressions:
1) Simple Linear Regression
2) Multiple Linear Regression

Simple Linear Regression*: It is a type of linear regression model where there is only independent or explanatory variable.*

Multiple Linear Regression*: It is similar to simple linear regression but here we have more than one independent or explanatory variable.*
Linear Regression can be written mathematically as follows:

$$Y = \beta 0 + \beta 1.X1 + \beta 2.\ X2 + \beta 3.\ X3 + \beta 4.\ X4 + \beta 5.\ X5 + \beta 5.\ X6 + \epsilon$$

β0 = Y-intercept (always a constant)
β1, β2, β3, β4, β5 = regression coefficients
ϵ = Error terms (Residuals)

In Linear Regression, we always lookout for best fit line which can be decided by gradient descent.

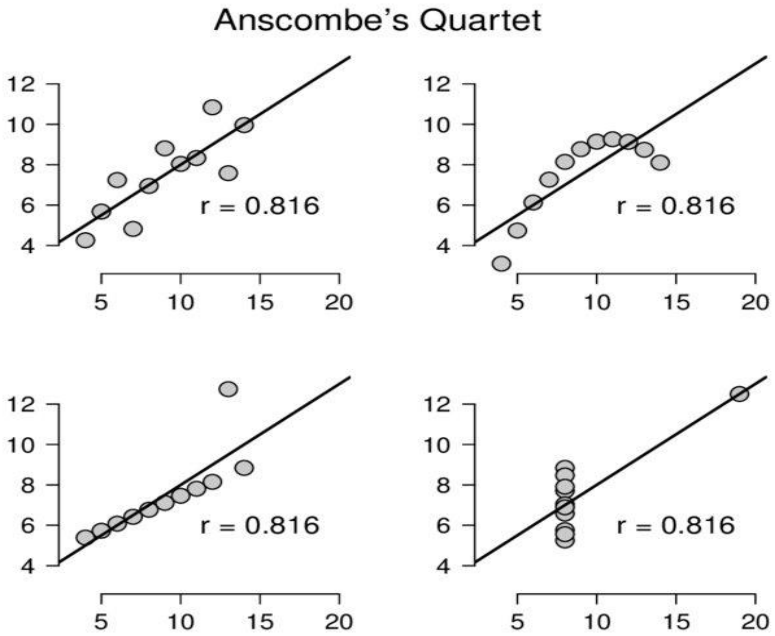
2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.

It 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.

For example : Look at the graphs shown:
-- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
-- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
-- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
-- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.



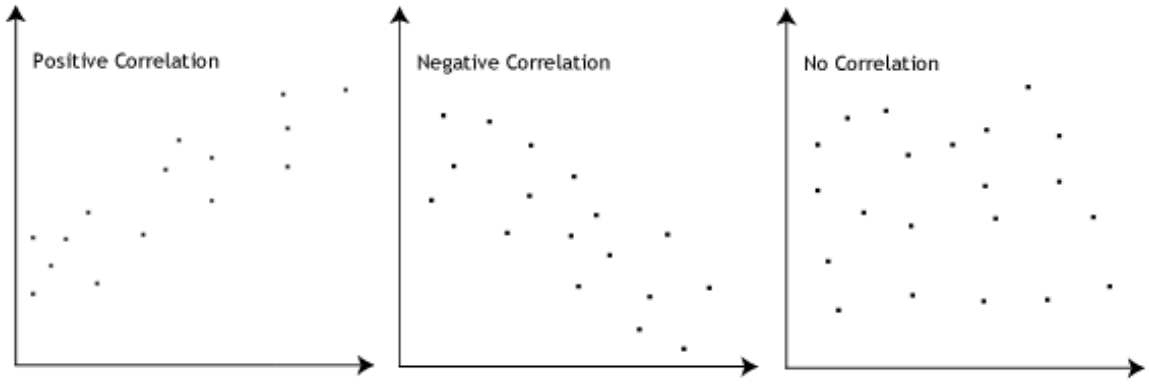
3. **What is Pearson's R? (3 marks)**

Pearson's R is a coefficient correlation mainly used in Linear Regression. It measures how strong two variables are. The value of coefficient correlations lie between -1 and 1.

If r = 1, it shows perfectly positive correlation, i.e., if one variable increases, other increases too.
If r = -1, it shows negative correlation, i.e., if one variable decreases, other increases or vice versa.
If r = 0, shows no correlation.
If 0 < r < 5, there is weak association
If 5 < r < 8, there is moderate association
If r > 8, there is strong association

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r=correlation coefficient
xi=values of the x-variable in a sample
x=mean of the values of the x-variable
yi=values of the y-variable in a sample
y=mean of the values of the y-variable



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

   Scaling, also called feature scaling is a method in which we convert the numerical features in fixed range, i.e., it normalizes the data within a particular range. It also helps speeding up the calculations in an algorithm.

   Scaling is performed because datasets often contains features that are varying in degrees of magnitude, range, and units. Hence, for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

   Difference between Normalized and Standardised scaling :
   -- *Normalized:*
   Also called MinMax() method which rescales values from 0 to 1.
   MINMAX: X = x-min(x) / max(x)-min(x)

   -- *Standardized:*
   It rescale in such a way that the mean is 0 and variance is 1. It replaces the values by their Z scores.
   Standardization scaling: x = x-mean(x) / sd(x)

   One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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 correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

In statistics, a Q-Q (quantile-quantile) plot is a probability plot, which is graphically compared the two quantiles against each other.

A q-q plot is used to compare the shapes of distributions, providing a graphical views of how properties such as location, scale and skewness are similar or different in the two distributions.

Q-Q plot importance are:
1. It can be used with sample sizes.
2.Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

In this dataset, the Q-Q plot is as follows: