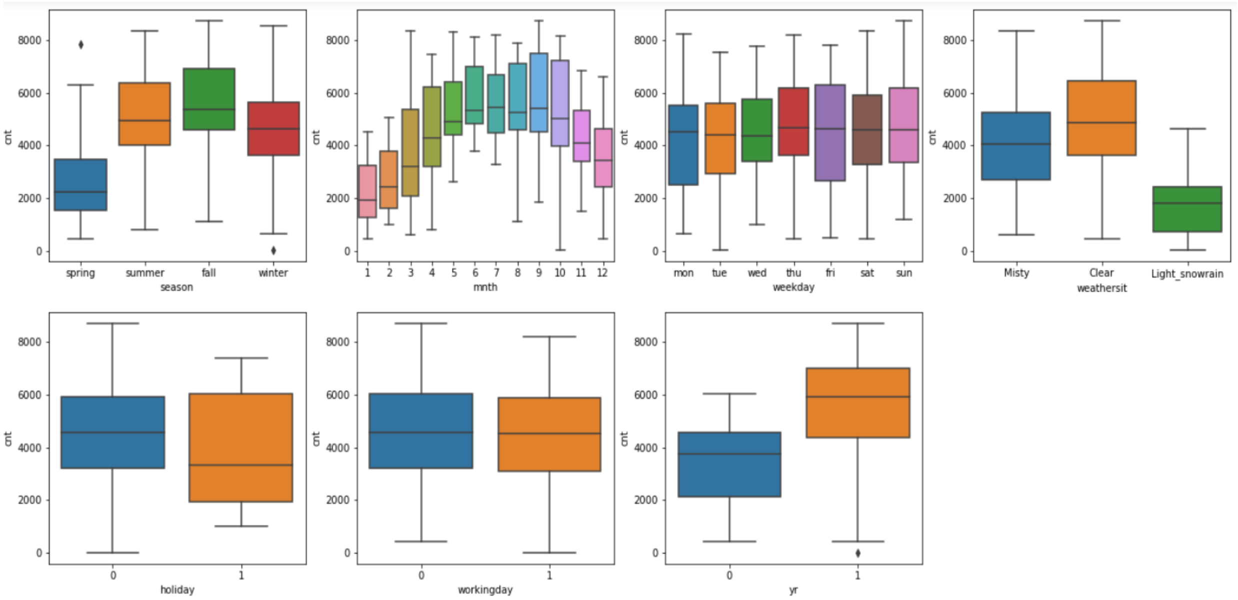
Assignment-based Subjective Questions

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

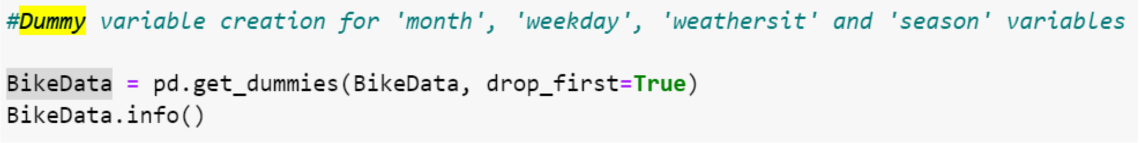
Ans. *The following were the categorical variables within the dataset - 'season','yr','mnth','holiday','weekday','workingday','weathersit'. These categorical variables were interpreted by creating boxplots for them against the dependent variable 'cnt'. Following are the inferences about their effect on 'cnt' -*

- ***Season*** - *Some variance can be observed across the different levels, with fall having the highest median that's around 5000 whereas it is lowest for spring at about 2000. The IQR for fall was higher than others as well, followed by summer and winter and finally for spring. Hence, season seems to be a good predictor variable.*
- ***Mnth*** - *A lot of variance in cnt can be observed for different levels of mnth as well, with the period of may to september having higher median values, that is again hinting at this being agood predictor variable. This could be because these are pleasant months without much snow as the value of cnt is especially less for months december to february.*
- ***Weekday*** - *The median value for all levels of weekday are similar at about 4500, hinting that it is not a decent predictor variable.*
- ***Weathersit*** - *Weathersit has variance across its levels especially having a high median when it is clear but quite low when there is snow/rain. This seems to be in tandem with the observation from mnth.*
- ***Holiday*** - *On days that are holidays, the median cnt can be observed to be less, hinting that this might be a good predictor for the dependent variable.*
- ***Workingday*** - *Workingday has a similar median across both its levels for cnt with less variance and might not be a good predictor for dependent variable.*
- ***Yr*** - *The total rented bikes i.e. cnt also significantly increased in 2019 as compared to 2018, hinting at yr to be relevant and a good predictor variable for cnt.*



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

Ans. It is important to use drop_first = True during dummy variable creation as using this condition, the first column for dummy variables is dropped hence avoiding multicollinearity between dummy variables and being a redundant factor. For n categorical levels in the data, n-1 dummy variables should be created as standard procedure. The effect of this phenomenon is especially greater when cardinality is small. There might be issues in converging and distortion of variable importances in the case of iterative models.The following code was used in the assignment as well -

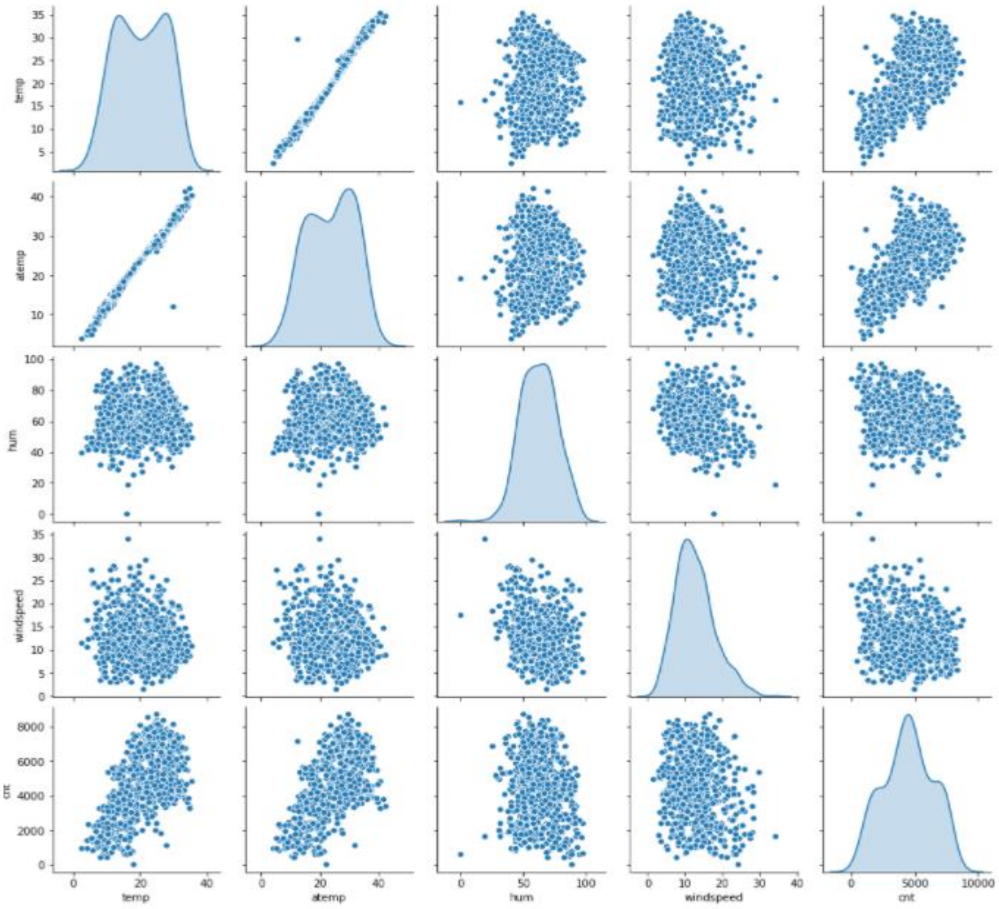


```
#Dummy variable creation for 'month', 'weekday', 'weathersit' and 'season' variables

BikeData = pd.get_dummies(BikeData, drop_first=True)
BikeData.info()
```

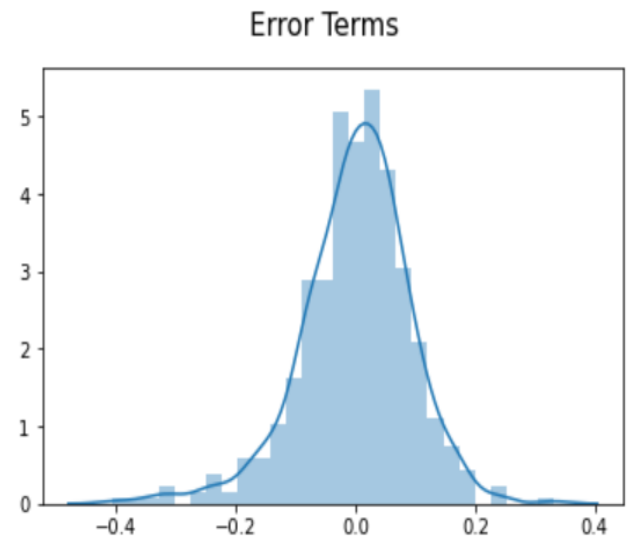
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. After observing the pair plot amongst numerical variables, it can be seen that temp and atemp have the highest correlation with the target variable 'cnt' and this is a positive correlation. These both- temp and atemp have miniscule differences and are similar in theory, hence, both have been suggested for this answer, however only one was used in the model building(temp).



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

*Ans. Through residual analysis the assumptions for linear regression were validated after building the model on the training set. The errors i.e y_train - y_train_predicted (difference between actual and predicted values of cnt by the model on the training data) were visualized using a distplot. The assumption for Linear Regression state that the residuals should be distributed normally and the mean should be centered at 0. The graph shown below is from the assignment and validated these assumptions about the errors being normally distributed and their mean equalling 0***.**



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

Ans. Basis the final model, the top 3 features that contribute significantly in explaining the demand of shared bikes are as follows -

- ***temp*** *- It has the highest coefficient of 0.469496 suggesting that temperature has a significant impact upon the count of total rental bikes.*
- ***weathersit_Light Snow & Light Rain*** *- It has a coefficient of -0.299283 that impacts count of total bikes negatively*
- ***yr*** *- It has a coefficient of 0.233180 and is impacting the count of total rental bikes positively*

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. The linear regression algorithm is a supervised machine learning algorithm wherein there is a predefined notion of labels and the output variable that is predicted is a continuous variable. This is most commonly used as a predictive analysis model and basically models a target prediction value that is commonly called as the target variable based upon single/multiple independent variables.

*The hypothesis function for Linear Regression can be given as - Y = ϑ1: + ϑ2 * x*

Wherein,

x: input training data (univariate – one input variable(parameter)) y: labels to data (supervised learning)

ϑ1: intercept (constant)

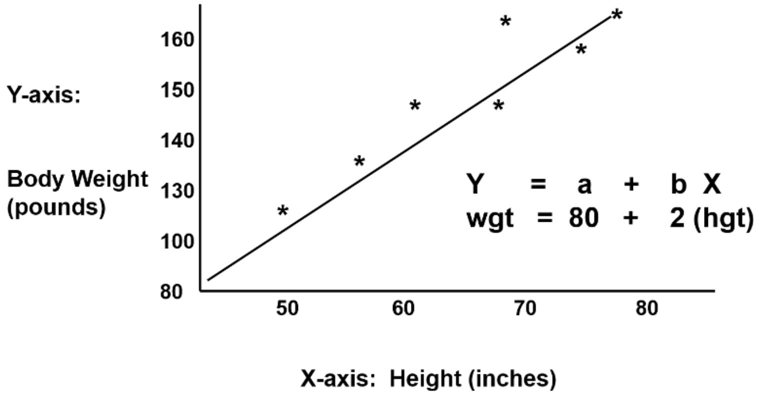
ϑ2: coefficient of x (Slope)

While training the model the best fit line is formed so as to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best ϑ1 and ϑ2 values.Once we find the best ϑ1 and ϑ2 values, we get the best fit line.

There are two types of Linear Regression -

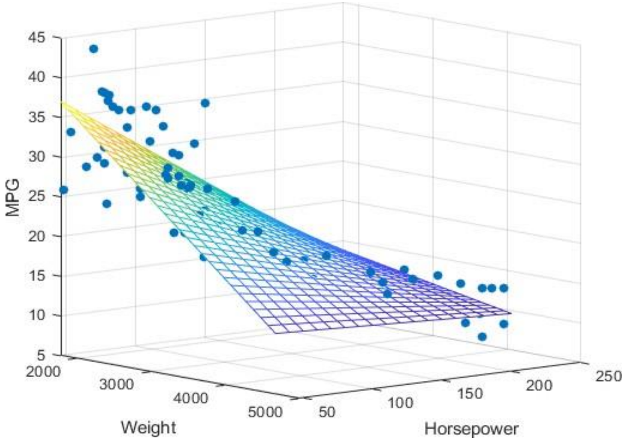
1. *Simple Linear Regression - This is the most elementary form of linear regression wherein the dependent variable is predicted using only one independent variable using a straight line.Standard equation for the regression line for simple linear regression is given by the following expression:* $Y = \beta_0 + \beta_1 X$

   Example Image-



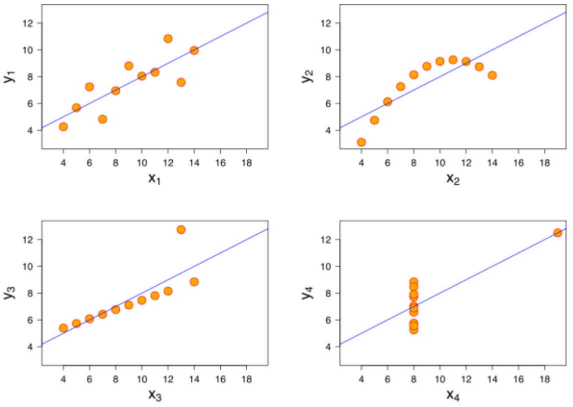
2. *Multiple Linear Regression - This is a statistical technique wherein the dependent variable is predicted and the relation of multiple independent variables is analyzed upon the target variables value. However, adding more variables isnt always best as it may cause overfitting and multicollinearity. The equation for multiple linear regression is given as -*

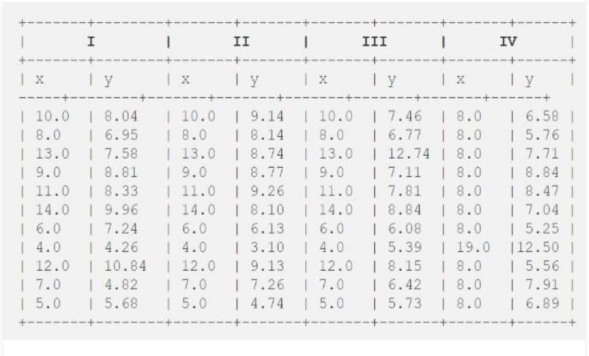
$$\widehat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$



3. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is comprised of four different data sets that have an approximately identical simple descriptive statistics(eg. mean), but appear very different once visualized in the form of scatter plots. Each dataset comprises 11 data points and these were created by an English statistician Francis Anscombe in 1973. The significance of the Anscombe's quartet is basically about the importance of visualizing data through graphs before analysis and checking for the effect of outliers on statistical properties. The following is a pictorial representation of the Anscombe's quartet and the original data points -





I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

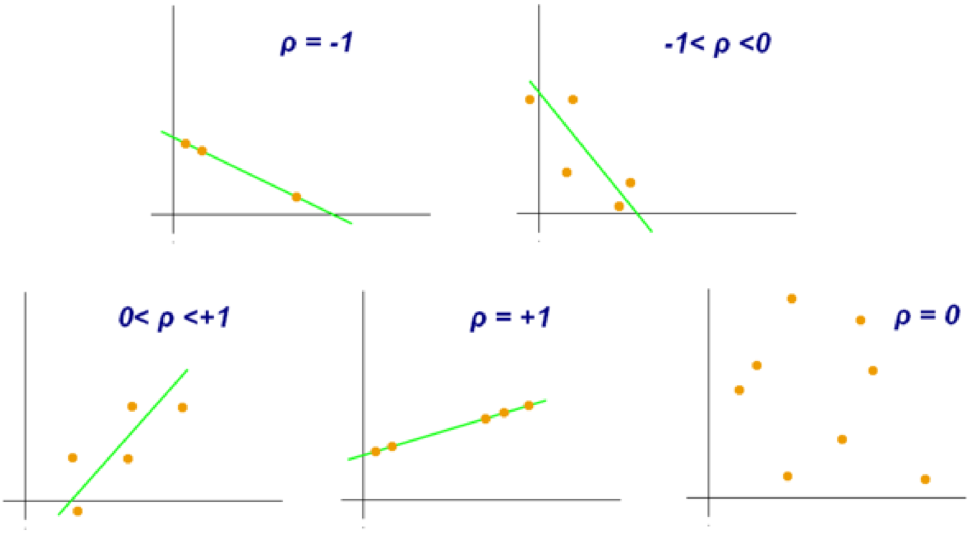
The following inferences can be observed from the 4 different graphs -

- *For the first scatter plot(top left) there seems to be a linear relationship between x and y.*
- *For the second scatter plot(top right) there is a non-linear relationship between x and y and the data points are not distributed normally.*

4. What is Pearson’s R?

Ans. Pearson’s R is another term for the Pearson correlation coefficient that is used in statistics to measure the linear correlation between two sets of data. It is the ratio between covariance of two variables and product of their standard deviations. It is a normalized measurement of covariance with values ranging between -1 and 1. A value of 0 is indicative that there’s no association between two variables, whereas a value greater than 0 indicates a positive association and finally a value below 0 is indicative of a negative association. The stronger the association of two variables, the closer the Pearson’s R would be to either +1 or -1 indicating a positive or negative relationship.

Examples of scatter plots with different correlation coefficients-



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

Ans. Scaling of features is a crucial step while building a machine learning model and deals with pre-processing of data prior to creation of the model. Through this method, the range of independent variables or features of data with significant differences in numerical ranges can be normalized or standardized.

Scaling is performed primarily because machine learning algorithms like linear regression use gradient descent as an optimization technique and require data to be scaled. If there are differences in ranges of features, the step size of gradient descent would be different for each feature and hence having features on a similar scale can help the gradient descent converge more quickly towards the minima. It is needed to bring every feature in the same footing without any upfront importance. If scaling is not done then the ML algorithm would weigh features based upon their numerical value with no consideration about their units or what the feature represents, that may lead to coefficients of some variables being significantly higher in comparison to others.

- *Normalized Scaling - This is a scaling technique wherein values are shifted and rescaled so that they may range between 0 and 1, and is also more commonly known as min max scaling. Normalized scaling was used during this assignment for scaling of features as shown in code below -*

```
In [33]: scaler = MinMaxScaler()
```

The formula for Normalized Scaling is given as -

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Such scaling responds well when standard deviation is small and distribution is not gaussian, however, it is sensitive to outliers.It is useful for algorithms like Neural Networks where any distribution of data is not assumed

- *Standardized Scaling - This is a scaling technique where values center around the mean with unit standard deviation, meaning that the mean of new attribute is 0 and the new distribution has a unit standard deviation. It assumes data is normally distributed within each feature.*
  The formula for standardization scaling is given as -

$$X' = \frac{X - \mu}{\sigma}$$

If data is not normally distributed it is not good to use this scaling method. In cases where there is Gaussian distribution it is helpful however and isn't affected by outliers.

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

Ans. The variance inflation factor or VIF is an indicator that provides the measure of how much variance of an estimated regression coefficient increases due to collinearity. A higher VIF suggests severe multicollinearity with any value greater than 10 being severe, while 5 is taken as moderate and 1 shows no multicollinearity.
Hence, if the value of VIF is infinite it means that there is a perfect correlation between the variables.This suggests that the particular independent variable for which VIF is infinite is explained completely by the other independent variables

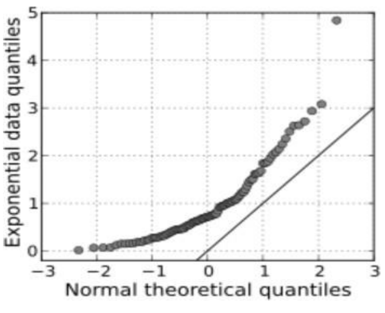
The formula for VIF is given by - 〖VIF〗 _1=1/(1-R_1^2)
Hence, if VIF is infinite, it would mean that the r squared equals 1, which would make VIF = 1/(1-1) = infinity.In such a case corrective measures need to be taken before using multiple regression like elimination of variables that are redundant and don't explain much variation in the model. Another approach is to increase the sample size which will make the confidence intervals of model coefficients much more narrow and will overcome multicollinearity.

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

Ans. Q-Q plot is also known as quantile-quantile plot, which is a probability plot that helps compare two probability distributions by plotting their quantiles against each other. The first quantile plotted is of the variable that is being tested for hypothesis and the second is for the actual distribution that we are testing against, The main step in constructing it is calculating or estimating the quantiles to be plotted. The purpose of the Q-Q plot is to understand if two data sets come from the same distribution. In case two distributions that are compared are similar, then the plot will lie on the line y = x, whereas if the distributions are linearly related then points in the Q-Q plot would lie approximately on a line but not y=x.
The Q_Q plot can be also used for the following -

- *Estimating parameters in a location - scale family of distributions*
- *Comparing shape of distributions*
- *Similarity in location, scale and skewness of two distributions*

The Q-Q plot is important for interpretation of data sets as it helps us determine if a dataset follows any particular type of probability distribution like normal, uniform,exponential.Examples of Q-Q plot -


Exponential data quantiles
Normal theoretical quantiles