# 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?**

A) **Season, Weather situation, holiday, month, working day and weekday were the categorical variables in the dataset. A boxplot was used to visualise these. These variables influenced our dependent variable in the following :**

- **Season: The Boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt. Summer and winter had cnt values that wee in the middle.**

- **Weather Situation: When there is a heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was 'Clear, partly cloudy'.**

- **Holiday: Rentals were found to be lower during the holidays.**

- **Month: September had the most rentals, while December had the fewest. The weather in the December is typically cold and snowy.**

- **Weekday: The overall median for the weekdays and working days are same.**

- **Working Day: It had little effect on the dependent variable.**

- **Year: The Median of the Bike Rentals has increased in the year 2019 compared to 2018.**

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

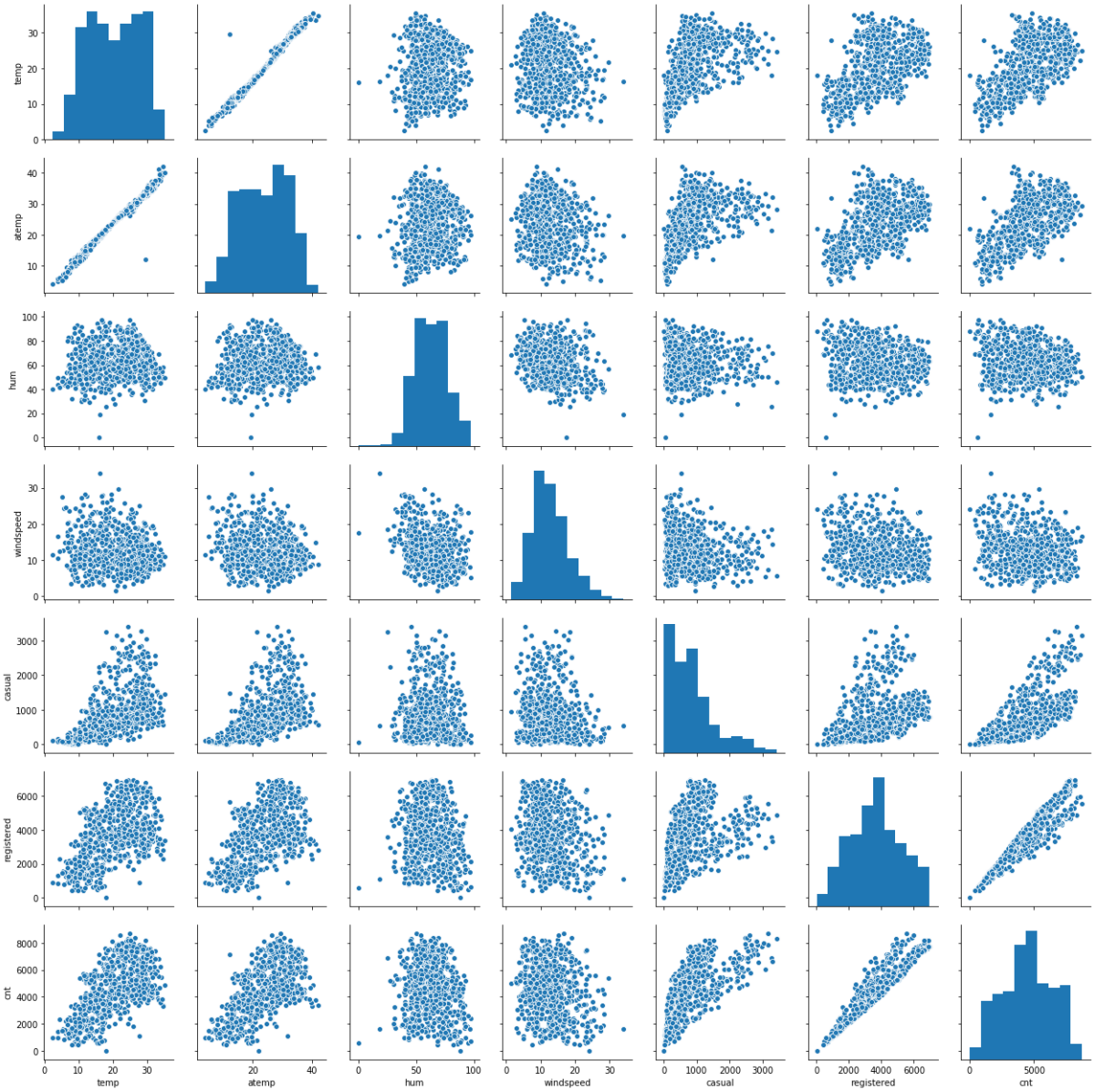
**It is important to use drop_first = True as it helps in reducing the extra column created during dummy variable creation. It helps to reduce the correlations created among dummy variable.**

**Dummy variable will be correlated if you don't remove the first column. This may have negative impact on some models, and the effect is expanded when the cardinality is low.**

**For example, we have different convergent and lists of variable importance may be distorted. Another argument is that having all dummy variables result in multicollinearity between them. We lose one column to keep everything under control.**

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

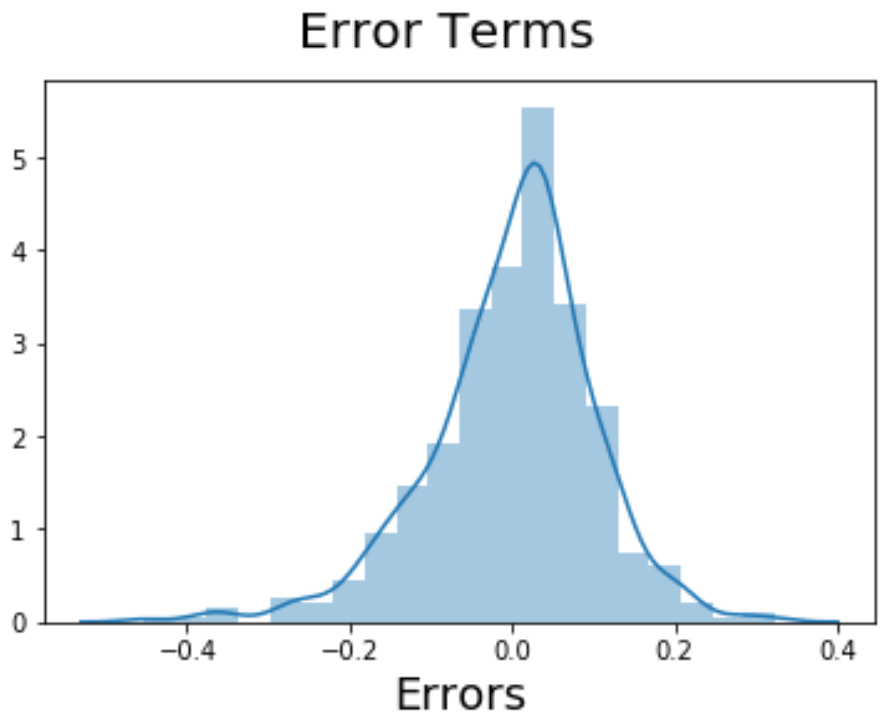
**A) "Temp" and "atemp" are the two numerical variables. They are highly correlated with the target variable cnt.**




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

**A) By plotting a distplot of the residuals we validate the assumptions of the linear regression and to see if it is a normal distribution or not.**

**The distribution of residuals should be normal and the mean is 0. We test the residual assumption by plotting a distplot of residuals to see whether it is a normal distribution or not. The residuals are scattered around the mean = 0 as seen in the below diagram.**




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

**A) The following are the 3 predictor variables that influence the bike booking to our final model.**

**Temperature: With a coefficient of 0.4279, a unit increase in the temp variable increases the number of bike rentals by 0.4279 units.**

**Weather Situation 3: With a coefficient of -0.2413, a unit increase in the weathersit3 variable reduces the number of bike hires by 0.2413 units.**

**Year: With a coefficient of 0.2360, a unit increase in the year variable increases the number of bike rentals by 0.2360 units.**

# General Subjective Questions

**1) Explain the linear regression algorithm in detail?**

**A) Linear Regression is a type of supervised Machine Learning Algorithm that is used for the prediction of numeric value. Linear Regression is the most**

**basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear Regression is based on the equation.**

$$Y = mx + c$$

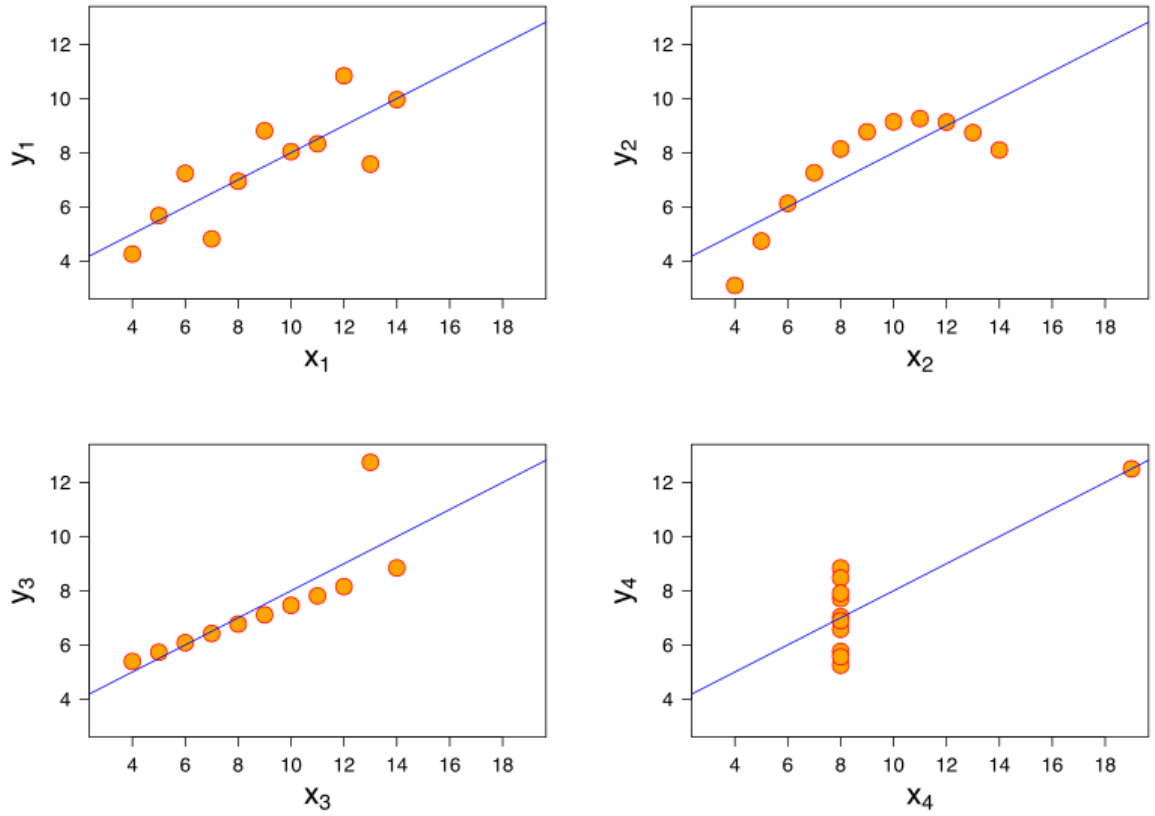
**It assumes that there is a linear relationship between the dependent variable(y) and the predictor/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and predictors or independent variables could be of any data type like continuous, nominal/categorical etc. This method tries to find the best fit line which shows the relationship between the dependent and predictor with least error. In, regression, dependent variable is the function of an independent variable and the coefficient and the error term.**

**Regression is divided into 2 types:**

1. **Simple Linear Regression: simple linear regression is used when the dependent variable is predicted using only one independent variable.**
2. **Multiple Linear Regressions: Multiple linear regressions is used when the dependent variable is predicted using only multiple independent variables.**
   **Equation of Multiple Linear Regressions**
   **y = mx1 + mx2+ mx3+ b**

**2) Explain the Anscombe's quartet in detail?**

**A) Anscombe's Quartet was developed by statistician Francis Anscombe. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.**




**Statistical Properties:**

- **The first scatter plot appears to be a simple linear relationship.**
- **The second graph is not distributed normally, while there is a relationship between them. But it's not linear.**
- **The third graph distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816**
- **The fourth graph shows an example when one high – leverage point is enough to produce a high correlation coefficient.**

**3) What is Pearson's R?**

**A) The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1. It depicts the linear relationship of two sets of data. R= -1 meaning a total negative linear correlation**

**R= 0 being no correlation**

**R= 1 meaning a total positive correlation.**

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

**A) Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If features scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.**

- **Normalization is generally used if you know the distribution of your data which does not follow a Gaussian distribution. This algorithm can be useful when don't assume any distribution of the data like K-Nearest Neighbour and Neural Network.**
- **Standardization can be helpful in cases where the data follows a Gaussian distribution. So, if you have outliers in your data, it doesn't affect the standardization. Standardization and normalization doesn't have a bounding range.**

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

**A) If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. The VIF indicates how much collinearity has increased the variance of the coefficient estimate. If an independent variable can be completely described by other independent variables, it has perfect correlation and R-squared value of 1.**

**VIF = 1/ (1-1)**

**VIF = 1/0 which is infinity.**

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

**A) Q-Q plots are also known as Quantile-Quantile plots. As the name says, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. The quantile of the first data set are plotted against the quantile of the second data set in a q-q plot. This is one of the tools to compare the shapes of different distributions. A scatterplot plotting two sets of quantiles against each other is known as a Q-Q plot.**

**Both sets of quantiles came from the same distribution; those points form a line that line is called straight line.**

**The Q-Q plot is used to answer these types of questions.**

1. **Two data sets have similar distributional shapes?**
2. **Both sets have common location and scale?**
3. **Two data sets have same tail behaviour?**
4. **Both the sets come from populations with a common distribution?**