# 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?**

these are the below categorical data variables and analyzed the effects on the dependent variable.

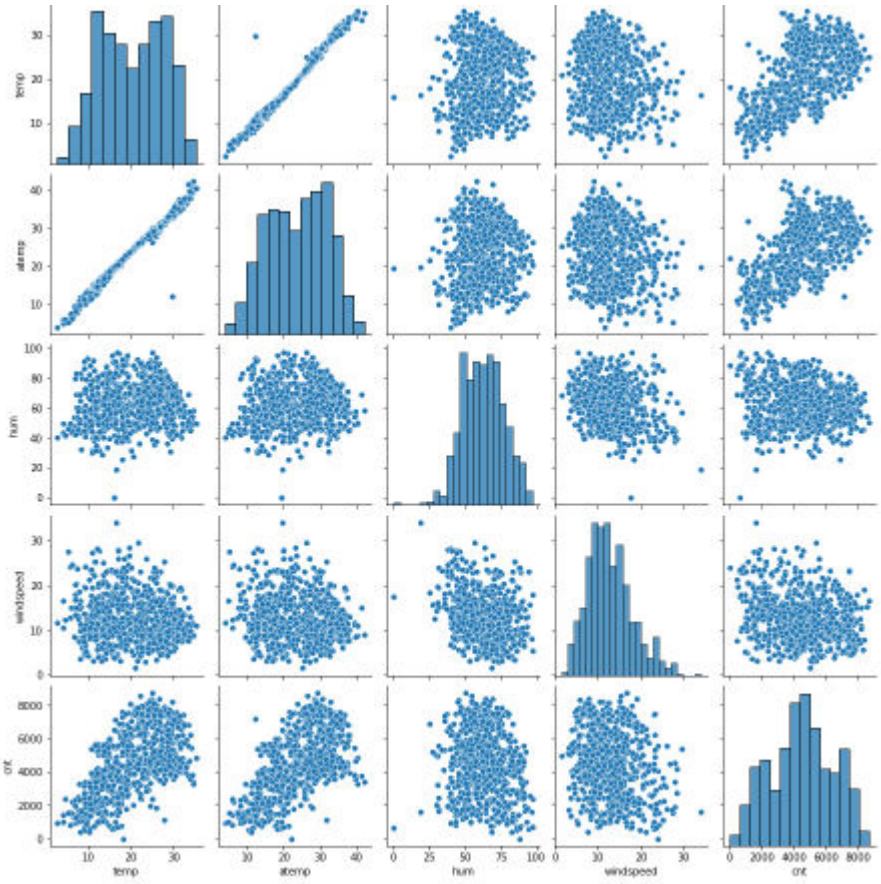
1. season: as we look in generated box plot for season vs cnt, found that spring season had very least count however fall season had the highest count.
2. mnth: Bike rental count was high in September month, and this is also part of waethersit.
3. holiday: Bike rental counts reduced during the holiday.
4. weathersit: When weather is Clear or partly or few cloudy then bike rental had the highest counts in comparison to others.
5. Yr: Bike rental count in 2019 was greater than 2018.

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

we use 'drop_first=True' for removing redundant columns, if we don't use it, it will create redundant columns and may affect analysis during creating of model and training data set as multicollinearity.

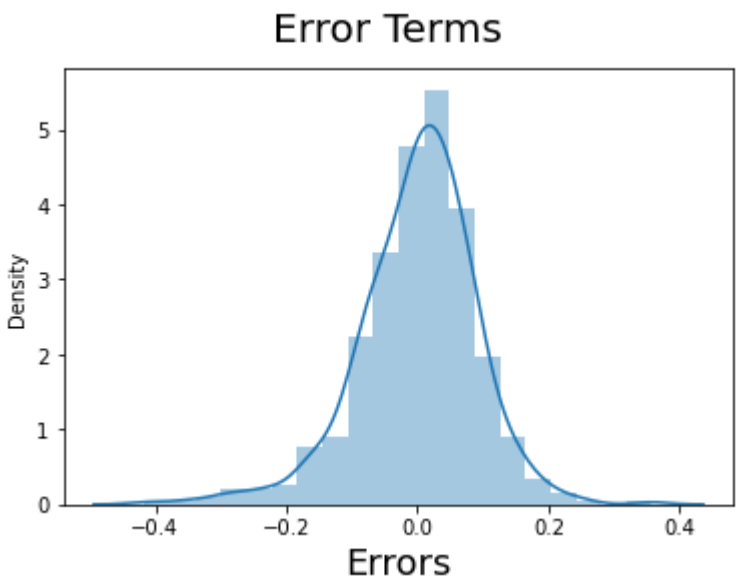
**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

there is two variables 'temp' and 'atemp' which are the highest correlation with the target variable(cnt).




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

We usually follow the Residual analysis for validation, we create a displot on the training dataset and see if it follow a normal distribution or not and make a decision, the below diagram it's the normal distribution, and the mean is approximate 0.




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

| var_name | coffiecient |
|---|---|
| • temp: | 0.491508 |
| • yr | 0.233482 |
| • weathersit_Light Snow & Rain | -0.285155 |

## General Subjective Questions

**1. Explain the linear regression algorithm in detail.**

Linear Regression comes under a supervised Machine Learning algorithm and it is used for the prediction of numeric values and it follows the basic equation "y = c+mx", It means that there is a linear relationship between a dependent(y) and independent variable(x). We evaluate the best fit between a dependent variable and independent variable. It works when the dependent variable is continuous data however independent variable could be any data set type as nominal, continuous and categorical. The regression method is used to find the best fit on the regression line with the least error.

It is divided into SLR and MLR.

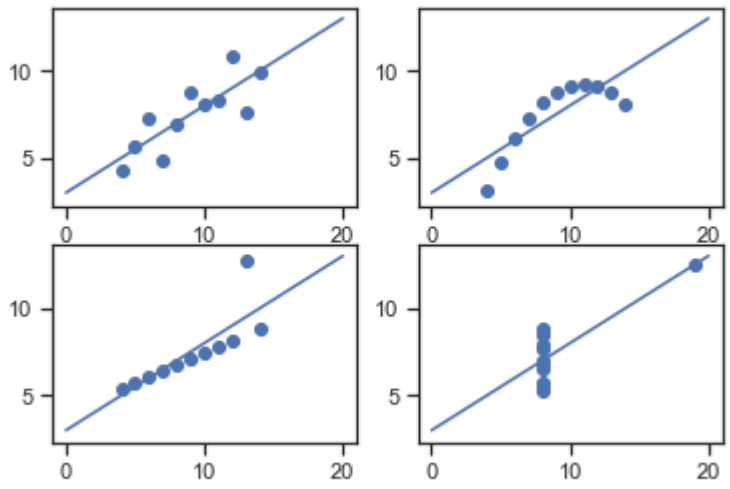
a) SLR(Simple Linear Regression): SLR is used if we are making a prediction using only one independent variable.

b MLR(Multiple Linear Regression): MLR is used if we are making a prediction using more than one independent variable.

**2. Explain Anscombe's quartet in detail.**

Anscombe's Quartet is modal and it was constructed in 1973 by Francis Anscombe to demonstrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

Suppose there are 4 datasets and have the same descriptive statistics(mean, variance, standard deviation, etc) but different graphical representations. Each graph plot shows the different behavior irrespective of statistical analysis.




The basic thing to analyze about these data sets is that they all share the

Dataset 1— it is consists of a set of (x,y) points that represent a linear relationship with some variance.

Dataset 2 — it shows a curve shape but doesn't show a linear relationship.

Dataset 3 — It looks like a tight linear relationship between x and y, except for one large outlier.

Dataset 4 — it looks like the value of x remains constant, except for one outlier as well.

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

Person'R is used to measure how strong relationship between two variables, it is also called Pearson correlation and is commonly used for linear regression and value range between -1 to 1.

- -1: It indicates a strong negative relationship.
- 0: it indicates No relationship.
- 1: It indicates a strong positive relationship.

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

Feature scaling is a method used to **normalize** or Standardlizationthe range of independent variables or features of data and it is generally performed during the data preprocessing step.

if we ignore the feature scaling, we might end up with higher or smaller value irrespective of the units of value.

- Normalization: it is used when the distribution of data do not follow Gaussian Distribution and value are evaluated between 0 and 1 also called min-max scaled
- Standardization: it is used when the distribution of data follows Gaussian Distribution, it rescales data to have a mean and standard deviation of unit variance.

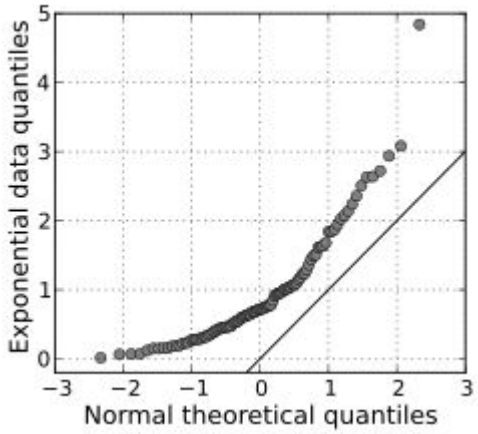
**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

VIF is Variance Inflation Factor, and evaluated using this formula- 1/(1-R_1^2 ), if R square is equal to one then VIF would be infinite, which means that there is a perfect correlation.

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

Q-Q Plots (Quantile-Quantile plots) are plots of the quantiles of data set against each other. It is used for comparing the distribution of shapes. it is created using scatterplot by plotting the quantiles of data set against each other. The purpose of the Q-Q plot is to find out if two data sets come for the same distribution.

**Example:**


Exponential data quantiles
Normal theoretical quantiles
5
4
3
2
1
0
−3
−2
−1
0
1
2
3