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 Season, month, weekdays, weather situation, holiday these are the categorical variables**.**
   - **Season:** Spring season have very low cnt values, Fall season has Very High count of cnt values. Summer & winter season have good count of cnt values**.**
   - **Weather Situation:** Clear Weather season has higher cnt values count & In Heavy rain/snow there is no cnt count indicating that weather is unfavourable for this business**.**
   - **Weekday:** All days of week have similar count of cnt values**.**
   - **Holiday:** On working day bike demand has been increase as compare to non-working day.
   - **Month:** Sept have highest rentals, Jan Dec month have lower count because of Cold & snowy season**.**

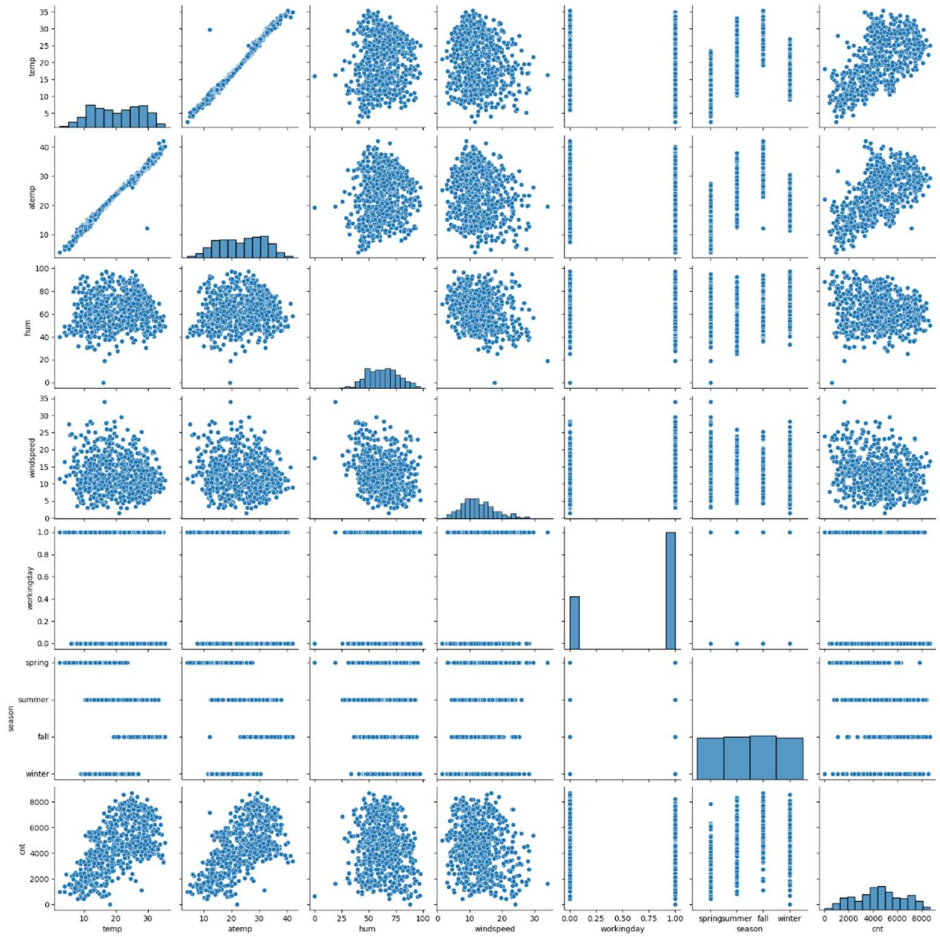
2. **Why is it important to use drop_first=True during dummy variable creation?**

   Ans:

   - It's is important to use drop_first=True, helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables
   - Let's say we have 3 types of categorical variables, we want to create dummy variables. If one variable is not furnished and semi furnished then it will un furnished. We not need to create 3rd variable.

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

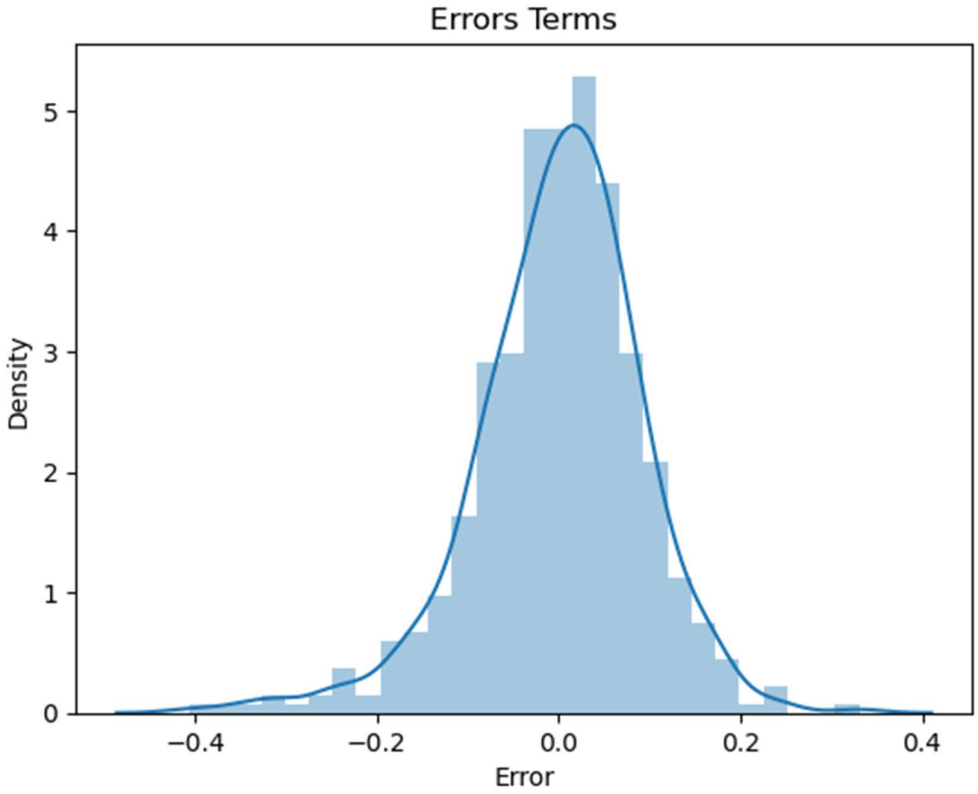
   Ans:



- Looking at temp & atemp variable which are highely corelated with each other.

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

Ans:



We Can validate assumption of linear aggression by using distplot of residual analysis. The Diagram say that its normally distributed mean = 0

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans: Top 3 as follows

- **Temp:** Coefficient values of 0.548 indicates that temp variable increases the number of rental bikes by 0.548
- **Year:** coefficient value of 0.23 indicates that year variable increase the number of rental bikes by 0.23
- **Weather situation 3 Light_Snow_rain:** Coefficient values of -0.2 indicates that a unit increase in weather situation 3 variable, decrease the bike hire number by 0.20.

General Subjective Questions

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

Linear regression is a supervised machine learning method that is used by the train using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

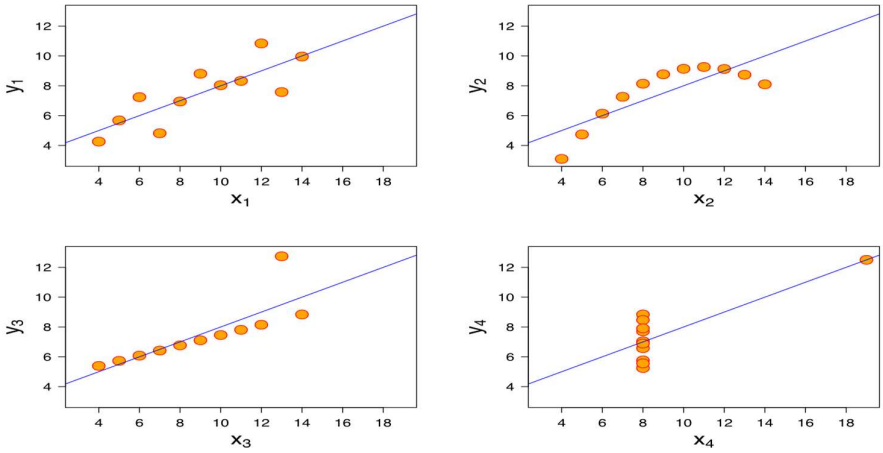
y = B0 + B1*X1 + B2*X2 + ….. + Bn*Xn.
B0 = Constant/Intercept
B1 = Coefficient of X1 Variable
B2 = Coefficient of X2 Variable
Linear Regression is divided into 2 pars

- **Simple Liner Regression:** It is used when dependent variable is predicted using only one independent variables.
- **Multiple Linear Regression:** It is used when dependent variable is predicted using multiple independent variables.

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

   Ans:

   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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

   Ans:

   The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

   - The values between 0 to 1 shows Positive corelation that means when one variable changes, the other variable changes in same directions.
   - 0 values indicated No corelations that means there is no relationship between variables.
   - The values between 0 to -1 shows Negative corelation that means when one variable changes, the other variable changes in opposite directions.

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

   Ans:

   - Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm

   - Mostly collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude

- Normalized Scaling: *It brings all of the data in the range of 0 and 1.* sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardized Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). sklearn.preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans: The formula for VIF is
   1 / 1-R^2
   Basically, if R^2 is 1 then VIF becomes infinite. It means there is perfect corelation between variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**
   Ans: A Q-Q plot is scatter plot of tow set of quantiles against each other. Its purpose is to check of the two sets of data came from the same distributions. It is visual check of data. If the data is from same source than the plot will appear as a line