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: - We can infer the following points about the effect of independent variable on dependent variable: -

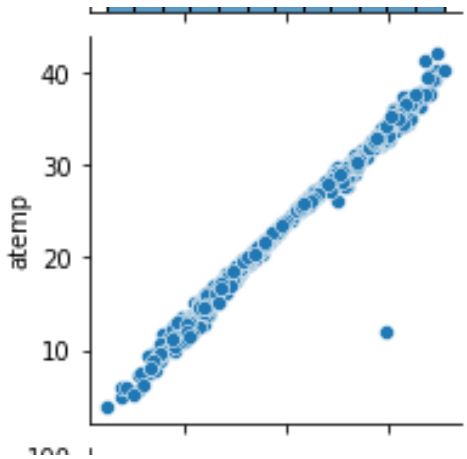
   - The demand is high in Fall and Summer season as compared to other seasons.
   - The demand is increasing significantly from 2018 to 2019.
   - The demand is high from the month of March to October as compared to other months.
   - The demand is slightly higher on holidays and weekends.
   - The demand is highest when the weather is clear, few clouds, partly cloudy, partly cloudy and lowest when there is heavy rain, ice pallets, thunderstorm, mist, snow, fog.
   - The demand is slightly high on Saturday and Sunday as compared to other days.
   - The demand is very high on holidays.

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

   Ans: - During the dummy variable creation it helps to reduce extra categorical column which is created while conversion. For example: - If we have 7 types of values in categorical column and we want to create dummy variable from that column, then if one variable is not any of the six variables then it is obviously 7th variable. We need **n-1** columns of categorical values if we have **n** values of categorical column.

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

   Ans: - From the pair-plot it is very clear temp and atemp are very closely related to each other as they form a close straight line in a scatter plot.

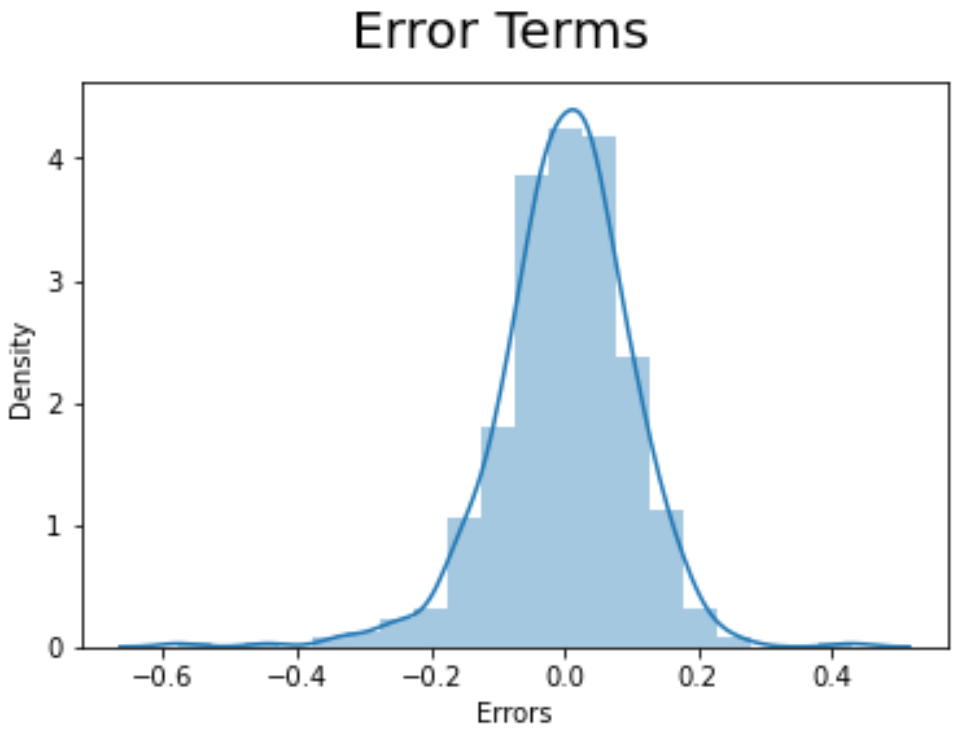


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: - Assumptions of Linear Regression after building the model are: -

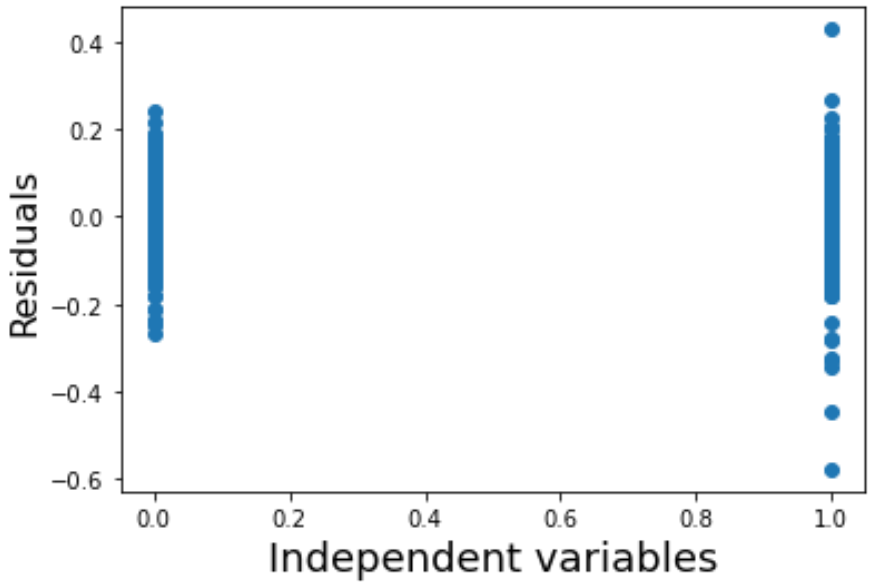
- **Linear Relationship between X and Y: -**
  In a model linear relationship is defined by
  cnt = 0.6 - ($windspeed$ *0.1685) - ($seasonspring$ * 0.2682) - ($seasonsummer$ * 0.0458) + ($yr2019$ * 0.2470) - ($mnthDec$ * 0.1369) - ($mnthJan$ * 0.1141) - ($mnthNov$ * 0.1446) - ($weekdaySunday$ * 0.0436) - ($weathersitLightSnow,LightRain+Thunderstorm+Scatteredclouds,LightRain+Scatteredclouds$ * 0.2895) - ($weathersitMist+Cloudy,Mist+Brokenclouds,Mist+Fewclouds,Mist$ * 0.0847)

- **Error terms are normally distributed: -**
  The Error terms are normally distributed as shown in figure for train set



(FIG 1)

- **Error terms are independent of each other: -**
  This shows that the Error terms are independent of each other.



(FIG 2)

- **Error terms have constant variance: -**
  As shown in Fig 2 there is constant variance in the model. This phenomenon is called homoskedasticity.
- **The independent variables should not be correlated: -**
  For a model building, we can remove the variable that are most correlated to avoid multicollinearity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans: - The Top 3 features that defined the demand of the bikes are: -
   - **yr_2019** with coefficient of 0.2479
   - **weekday_Sunday** with coefficient of -0.0436
   - **season_summer** with coefficient of - 0.0458

General Subjective Questions

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

   Ans: - The following steps are followed in Linear regression algorithm: -

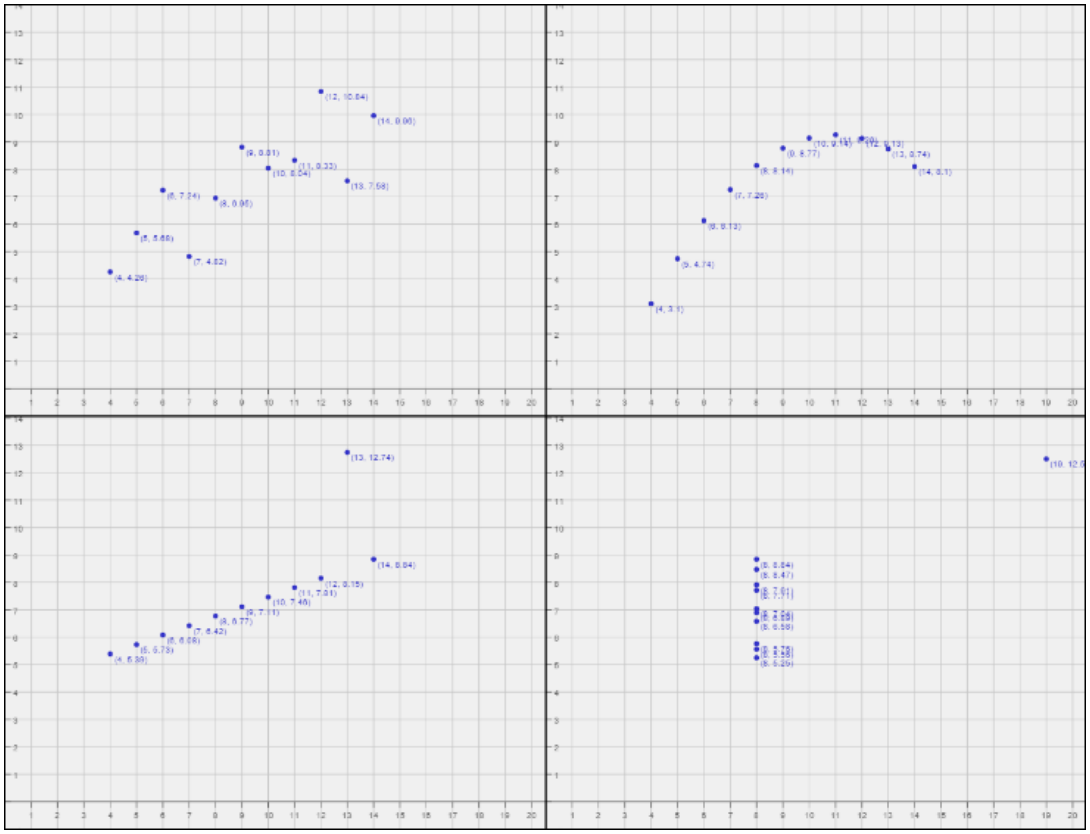
   - **Step 1: Reading and Understanding Data**
     In this step we need to read the data from the given file and then understand it in terms of numeric data and categorical data.
   - **Step 2: Visualising the Data**
     Visualising the different variables with target variable and try to find most correlated variables from that and choose it for further step of model building.
   - **Step 3: Performing Simple Linear Regression**
     Now divide the current data set into training and testing datasets. Then on training dataset try to estimate the model coefficients. Then we check the significance of these coefficients using R-squared and F-statistics method.
   - **Step 4: Residual Analysis**
     Now we look for the error terms in the training dataset and looking for the patterns in the residual to ensure the training model is correct.
   - **Step 5: Prediction on the Test Set**
     After all of the above steps, we ensure our model's working with the testing dataset and check it's behaviour and visualise the residual.

   After all of the above steps, we get the linear equation which is used to predict the values for actual data.

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

   Ans: - Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
   Let's assume we have four dataset and all four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression as shown in (FIG 3).
   Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still different from each other when plotted.

   How to solve the problem: -
   When we plot the data, we need to analyse the graph whether it is close linear regression or not. So, we go with that data set with close linear relationship visually.
   Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analysing data, and statistics about a data set do not fully depict the data set in its entirety.

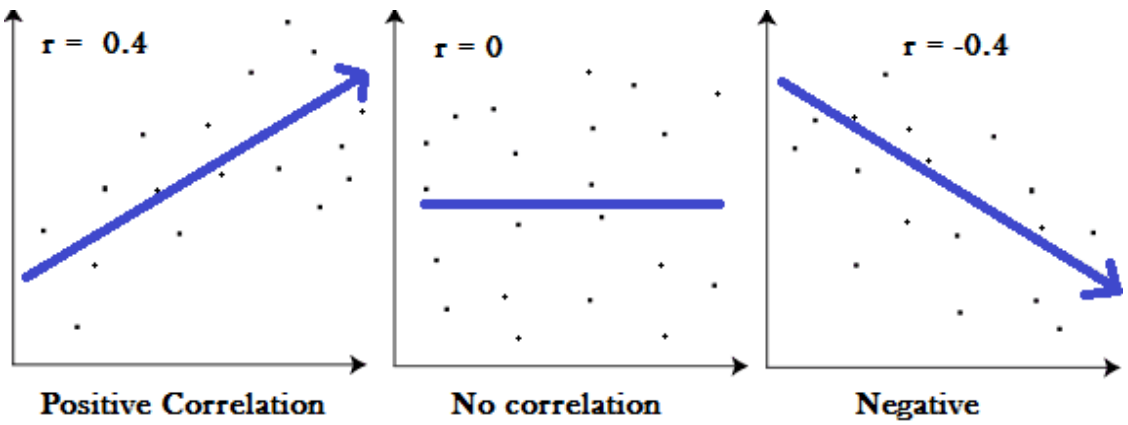


(FIG 3)

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

Ans: - The Pearson's R is a correlation coefficient i.e. commonly used to show the linear relationship between two variables. It is used to find how strong is the relationship between variables. The value of the coefficient is between -1 to 1 where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



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

   Ans: - Scaling is defined as pre-processing of the data to standardising or normalising it. The two most common techniques of feature scaling are Normalisation and Standardisation. Normalisation is generally bound to it's value between either [0,1] or [-1,1] while standardisation transforms the data to have 0 as mean and 1 as a variance.

   Need of Scaling: -

   - Ease in interpretation
   - Faster convergence to gradient descent methods
   - It improves the execution time of algorithm
   - Features with high scale dominates over low scale

   Normalisation: -

   $$y = (x - min) / (max - min)$$

   Standardization: -

   $$y = (x - mean) / standard\_deviation$$

   Difference between Normalisation and Standardization: -

   - Standardization can be used in case when data represent Gaussian Curve
   - Normalization is great when Non-Gaussian Curve Representation
   - Standardization is less impacted by outlier values.
   - Impact of Outliers is very high in Normalization

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

   Ans: - If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. for e.g. if the VIF is 5, it means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity.
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear formula with another variable.
   Mathematically, the VIF is infinite when the R-squared value is equal to 1.

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

   Ans: - In Q-Q(Quantile-Quantile) plot we plot quantiles of two variables against each other, then we obtain this plot. The plot provides inference that distribution of variables are similar or not.

   There are following interpretations from the q-q plots: -

   - If the all points of quantile lie on the straight line at 45-degree angle. It indicates the sample have similar distribution.

- If the y-quantile are lower than x-quantiles hence the distribution is not same and it is tendency towards x-value lower than y-value of the plot.
- If the x-quantile are lower than y-quantiles hence the distribution is not same and it is tendency towards y-value lower than x-value of the plot.
- There are some points where the x-quantile is lower than y-quantile and vice-versa. Then there is some mixed distribution where the x-value is dominating over y – value and vice-versa.