Assignment-based Subjective Questions and Answers

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

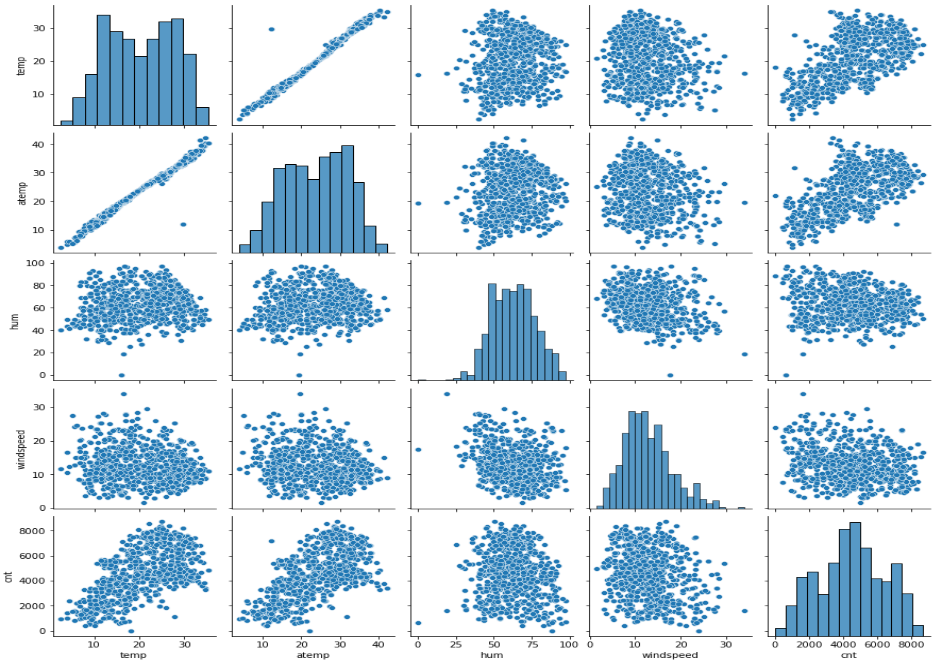
   Answer: Inferences for Analysis of categorical variables from the dataset on the dependent variable.

   a. Year 2019 has a higher median than 2018, it might be due to the fact bike rentals are getting popular and people are becoming more aware about environment.
   b. Overall spread in the month plot is reflection of season plot
   c. People rent more on non-holidays compared to holidays.
   d. Overall median across all days is same but spread for Saturday and Wednesday is bigger. (Saturday is non-working day)
   e. Working and non-working days have almost the same median although the spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that
   f. Clear weather is most optimal for bike renting, as temperature is optimal, humidity is less, and temperature is less.

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

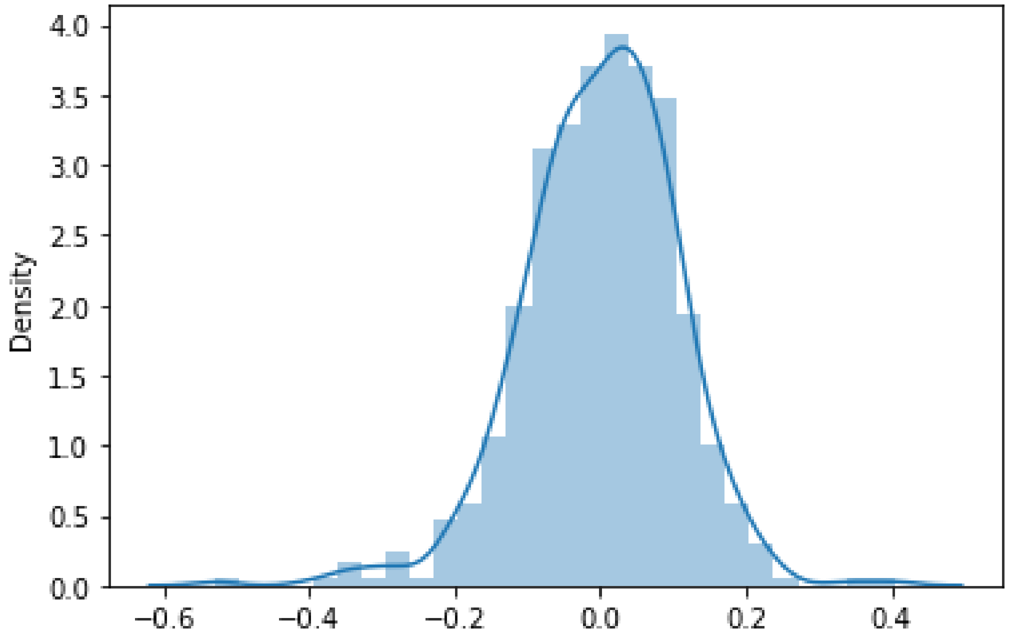
   Answer: A variable with n levels can be represented by n-1 dummy variables. So, if we remove the first column then also, we can represent the data. If the value of variable from 2 to n is 0, it means that the value of 1st variable is 1.

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



"temp" and "atemp" are the two numerical variables which 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?



Distribution of Error Terms

Answer: Residuals distribution should follow normal distribution and centred on 0 (mean 0). We Validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.

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

The top 3 features are:

1. Temperature (temp) - A coefficient value of '0.4910' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4910 units.
2. Weather Situation(weathersit_Light Snow & Rain) - A coefficient value of '-0.2842' indicated that, a unit increase in weathersit_Light Snow & Rain variable decreases the bike hire numbers by 0.2842 units.
3. Year (yr) - A coefficient value of '0.2336' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2336 units.

General Subjective Questions

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

   Answer: Linear regression is one of the most widely used models (both in academic as well as industry). It is type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Regression is the most commonly used predictive analysis model.

   Linear regression equation is “y = mx + c”

   It assumes that there is a linear relationship between the dependent variable(y) and predictors(s)/independent variable(x).
   In Regression, we can calculate the best fit line which describes the relationship between the dependent and independent variable.
   Regression is performed when the dependent variable is of continuous data type or independent variables could be of any data type like continuous, categorical etc.

   Regression used to find the best fit line which shows the relationship between the dependent variable and predictors with least error.
   Linear regression models can be classified into two types depending upon the number of independent variables:

   Simple linear regression: When the number of independent variables is 1
   Multiple linear regression: When the number of independent variables is more than 1

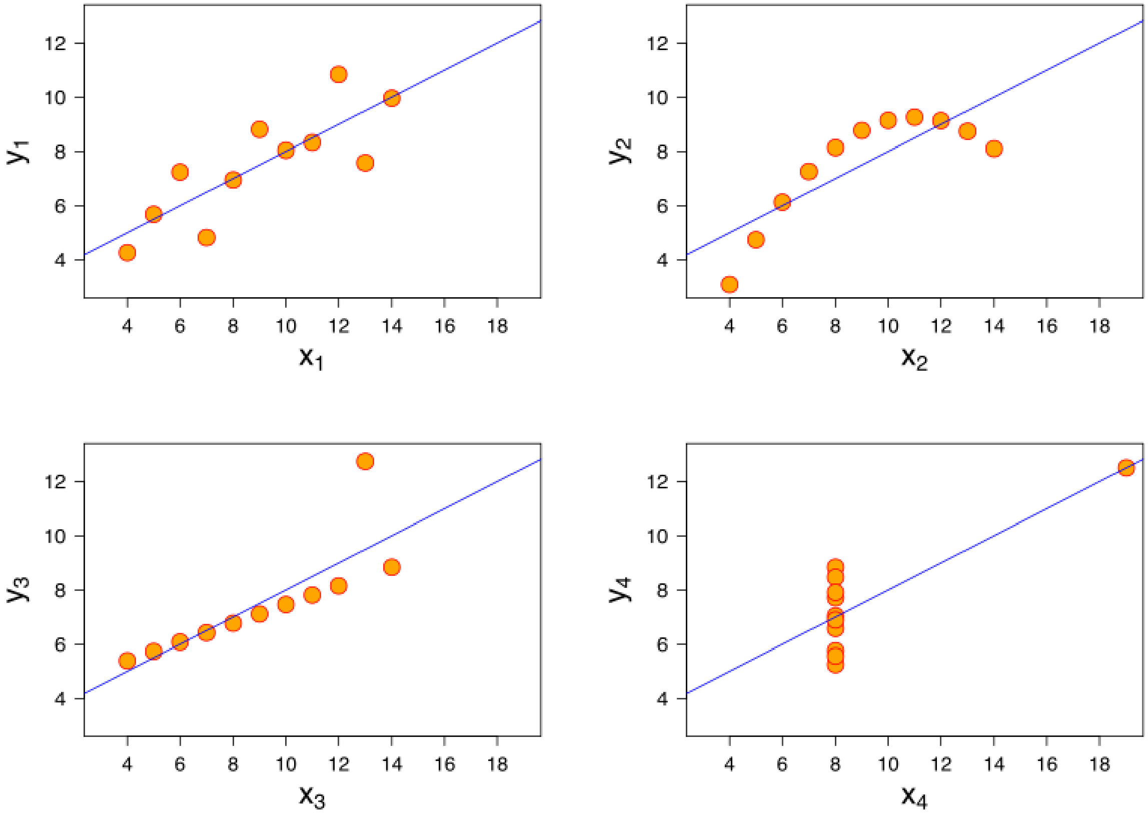
   The equation for multiple linear regression:

   $$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\epsilon$$

   β_1 - Coefficient for X_1 variable
   β_2 - Coefficient for X_2 variable and so on..
   β_0 is the intercept.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet was developed by statistician Francis Anscombe. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.



i. The first scatter plot appears to be simple linear relationship.

ii. The second plot is not distributed normally, it's not linear, and there is a relationship between them.

iii. The third plot, the 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.

iv. The fourth plot 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?

Answer: Pearson's R is a numerical summary of the strength of the linear association between the variables. It values ranges between -1 to +1. It shows the linear relationship between two sets of data.

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

r=1 means the data is perfectly linear with a positive slope
r=-1 means the data is perfectly linear with a negative slops
r=0 means there is no linear association

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

   Answer: Scaling of variables is an important step because, it is important to have everything on the same scale for the model to be easily interpretable.

   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 feature scaling is not done, then a machine learning algorithm tends to weigh higher values and consider smaller values as the lower values, irrespective of the units of the values.

   Normalization: MinMax scaling: Converts or compress the data between 0 and 1.
   It is generally used when you know that the distribution of your data does not follow a Gaussian distribution.
   Standardization: (mean-0, sigma-1): In cases where the data follows a Gaussian distribution. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

   Answer: VIF (Variance Inflation Factor): The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

   VIF = 1/1-R^2

   If there is a perfect correlation, then VIF = infinity. R-1 is the R-Square value of the independent variable. How well the independent variable is explained well by other independent variables.

   If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1.
   So, VIF = 1 / (1-1) which give VIF=1/0 which results in infinity

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

Answer: In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Plot of the quantiles of the first data set against the quantiles of second data set.

It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

If both the sets of quantiles from the same distribution, we should see the points forming a line that's roughly straight.

Used to find:

i. Two data sets come from populations with a common distribution or not
ii. To find two data sets have common location and scale.
iii. Two data sets have similar distributional shapes or not and have similar tail behaviour or not.