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? (3 marks)

The categorical variable in the dataset were season,weathersit,holiday,mnth,yr and weekday.These were visualized using a boxplot .These variables had the following effect on our dependant variable:-

1. **Season -** The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt.Summer and winter had intermediate value of cnt.
2. **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable.Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
3. **Holiday -** rentals reduced during holiday.
4. **Mnth -** September saw highest no of rentals while December saw least.This observation is on par with the observation made in weathersit.The weather situation in december is usually heavy snow.
5. **Yr** - The number of rentals in 2019 was more than 2018

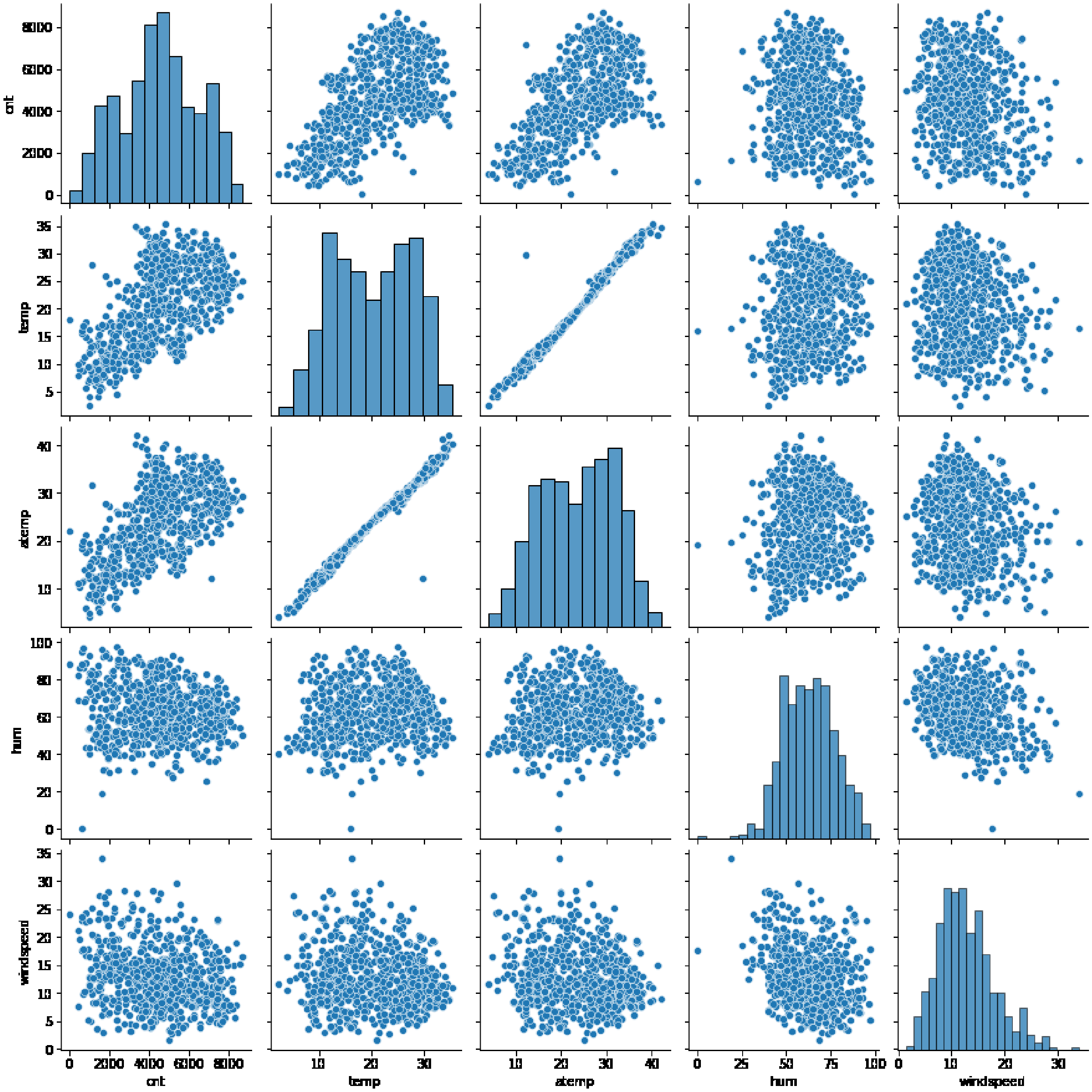
6. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Example:

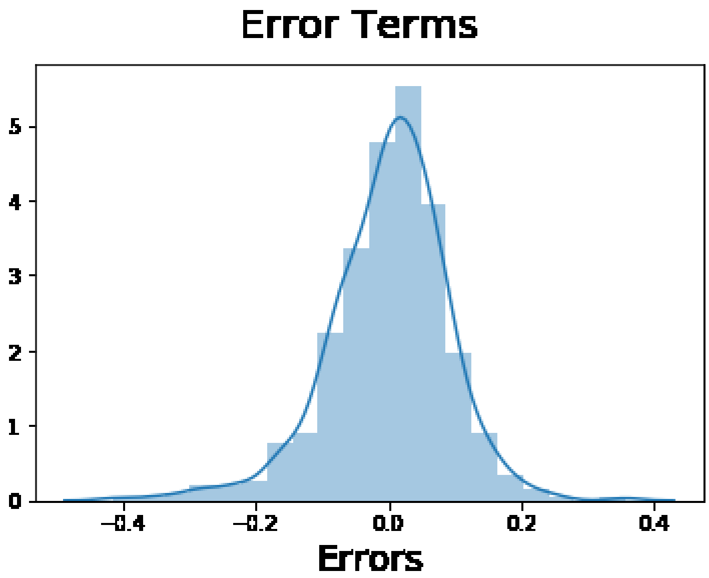
   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished

2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



“temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt)

3. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)



Residuals distribution should follow normal distribution and centred around 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.

4. 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-
Temp- coefficient value –'.5636'
Weather situation (weathersit_3)-'-.3070'
Year (yr)- coefficient value –'.2308'

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is based on the popular equation **"y = ax + b".**

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

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. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression : SLR** is used when the dependent variable is predicted using only **one** independent variable.
2. **Multiple Linear Regression :MLR i**s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots ,$$

β1 = coefficient for X1 variable

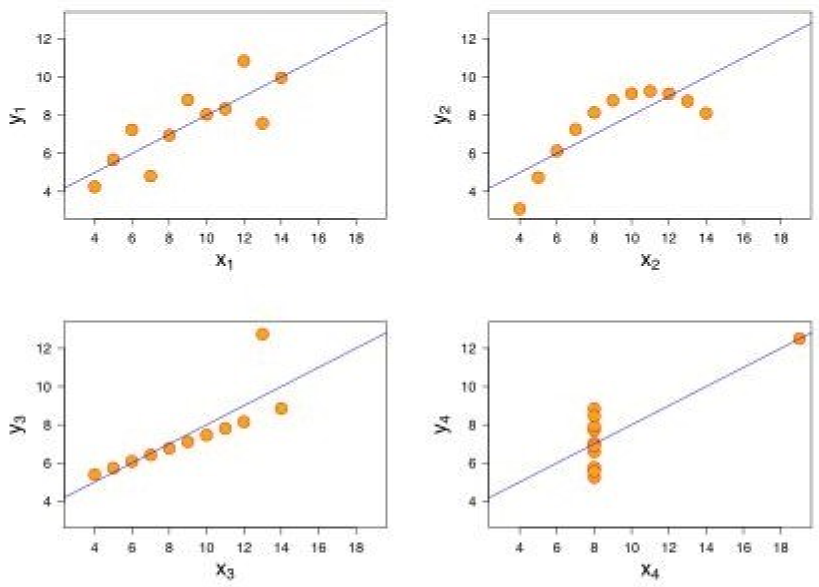
β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



The four datasets can be described as:

1. Dataset 1(top left): this fits the linear regression model pretty well.
2. Dataset 2:(top right) this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3(Bottom- left): shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4(Bottom-right): shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1.

It shows the linear relationship between two sets of data. In simple terms, it tells us can we *draw a line graph to represent the data?*

r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
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? (3 marks)

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance). Normalization is often called as Scaling Normalization while standardization is often called as Z-score Normalization

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. 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 ? (3 marks)

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this 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 it’s R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives 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. (3 marks)

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?