Assignment-based Subjective Questions

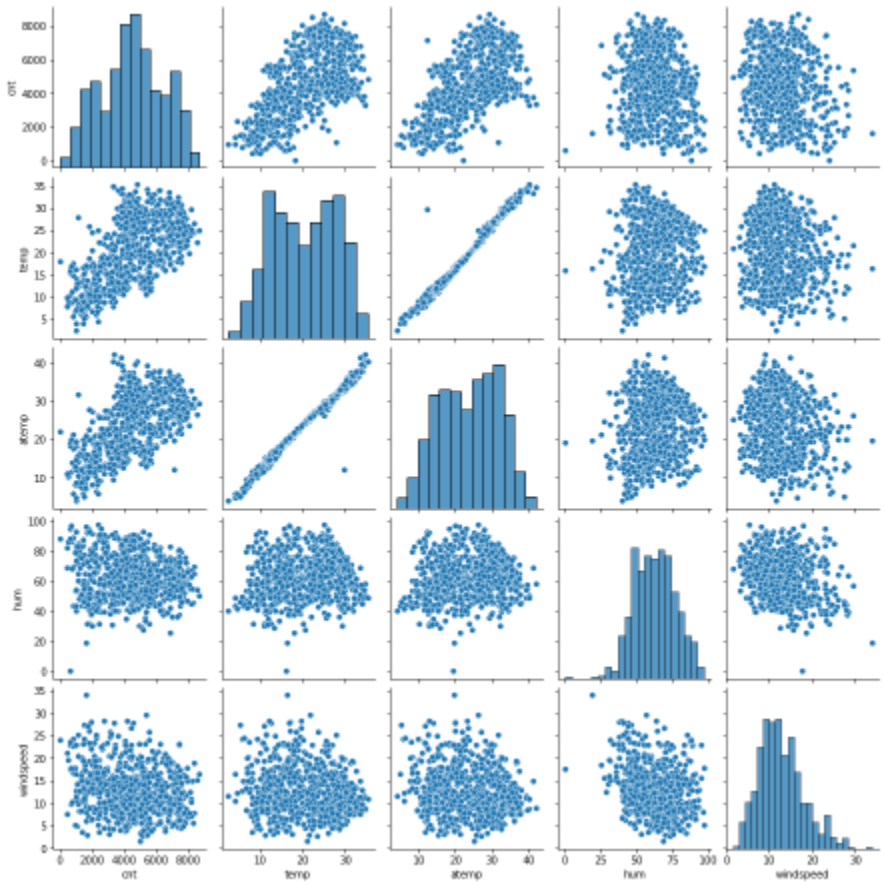
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- **Temp** - Temperature plays a crucial role, as it can be seen from the model that as the temperature increases the number of people renting the bike also increases.
- **Workday** - Although workweek has some influence on the number of bikes that are hired by people, specific days of a workweek does not pose any dependency on the 'cnt' target variable
- **Season** - The seasons winter and spring have some impact on the target variable. Spring has a negative effect which winter has a positive effect
- **Month** - In the month of July the people rent less bikes. Apart from this month none of the other months in an year bears any significance
- **Windspeed** - If the wind speed increase the number of bike users decrease. It is the most detrimental factor out of the ones that are observed
- **Year**- The Year variable has a positive co-efficient of 0.234 which implies that the count of bike hires has increased over the years

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

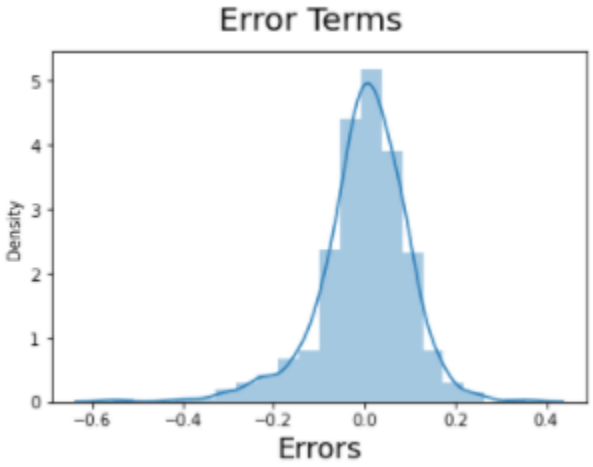
The argument drop_first = 1 drops the first dummy variable created for a categorical variable of n levels. The first dummy variable serves as a redundant variable since its value can be effectively inferred from the values in the other columns created from the same categorical variable.
Ex – The gender columns has three values M(male), F(Female), U(Unknown). After creating the dummy variables the will be three columns having binary values of 1 and 0. Now it can be known for sure that if any one of the columns will be 1 the rest have to be 0. Therefore if both F = 0 and U = 0, then M will be 1.
Using this analogy it will be safe to drop the M column for the analysis

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



The above plots suggest that the ‘temp’ and ‘atemp’ variables above have 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? (3 marks)



The above visualization was plotted for the errors that were obtained after predicting the values using the multiple linear regression model. As it seen form the graph the residuals follows a normal distribution with the mean equal to zero thereby confirming the assumption of the linear regression.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three variables that contribute to the demand of shared bikes are

Variable	**Co-efficient**
temp	0.483505
yr	0.234174
windspeed	-0.17334

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning model that is used for prediction of values leveraging the historic data. This model can be broadly classified as

- Simple Linear Regression - one dependent variable one independent variable
- Multiple Linear Regression – one dependent variable multiple independent variables

The objective is to find the best fit line that passes through the most of the points provided covering as much variance and use the characteristics of this line to predict the information of the target or the dependent variable. The effectiveness with which the line is fit can quantified using concepts in the Residual analysis such as **Residual Sum of Squares**

The independent variable have that are fed into the model have to be continues variables and the target variable can be continues or categorical etc.

The equation that underpins the model is

$$y= mx + c$$

- **m -** The inclination of the line with respective to the positive x-axis
- **x -** Value of the independent variable
- **c –** Intercept of the line with y-axis
- **y –** The target or the dependent variable

The equation has only one depended and independent variable. However for multiple linear regression we have to incorporate multiple independent variables for the prediction of the target variable and therefore the above formula has to be modified as shown below

$$Yi = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + ... + \beta pxip + \epsilon$$

where,

for i=n observations:

yi=dependent variable

xi=explanatory variables

β0=y-intercept (constant term)

βp=slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

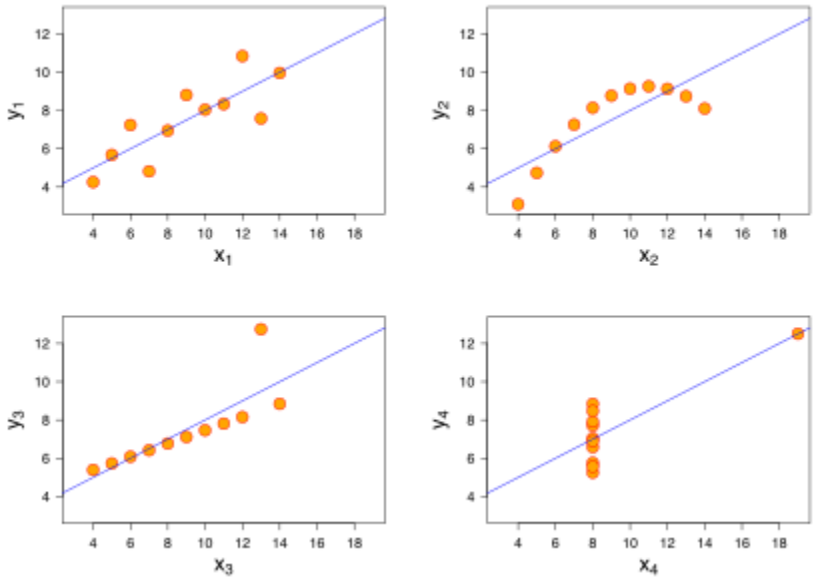
The Assumptions of Linear regression are as follows

1. The dependent and the independent variable show be having a linear relationship between them, which implies that when x changes there should be an equivalent change in y.
2. The error values should be independent of each other.
3. The error terms must be normally distributed.
4. Multi collinearity among the variables have to be eliminated in order to remove redundant variables
5. The error terms must have constant variance. (homoscedasticity)

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was developed by an English statistical analyst name Francis Anscomb. The ideology of this method is to prove the importance of visualizing data before analyzing it.

This theory is supported using four datasets that exhibit similar statistical features but when represented as plots demonstrate varying patterns.



- The first scatter plot (top left) appears to be a simple linear relationship
- The second graph (top right) is not distributed normally; while a relationship between the two variables it is not linear

- In the third graph (bottom left), 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 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?

The Pearson's R is a numerical measure of the relationship or association between two variables. The values ranges from – 1 to 1. The basic utility of this feature is to understand if it is possible to draw a line using the two variables under consideration. Pearson's R also helps us to detect multi collinearity while building model models using linear regression.

The values of Pearson's R can be interpreted as follows

- -1 : With increase in the value of one variable the other value of the other variable decreases .If the co-efficient is nearer to -1 it shows perfectly linear negative slope
- 1 : With increase in the value of one variable the other value of the other variable increase as .If the co-efficient is nearer to -1 it shows perfectly linear positive slope
- 0 : Linear relationship between the two variables do not exist

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

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.

Most of the times, 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.

Normalisation:

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$X_{new} = (X - X_{min}) / (X_{max} - X_{min})$$

- **X:** It is a set of the observed values present in X.

- **X_{min}**: It is the minimum values in X
- **X_{max}:** It is the maximum values in X

Standarsidation:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

(X – mean)/(standard deviation)

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

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?

,