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?

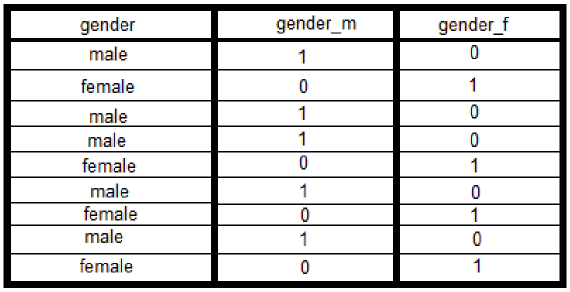
Ans. The categorical variables & their effect on dependent variable in the dataset are as follows:

- **Season** - The Spring saw the least rentals while Fall saw the most. The rentals started increasing from Summer & started dipping in Winters.
- **Yr** - The rentals count showed a rise in 2019 compared to 2018
- **Mnth** – The most number of rentals are in September and least is in January. The count is generally up in between April – October
- **Holiday-** Rentals are down in holidays and rises in non holidays
- **Weathersit** – It is observed there is no rentals during heavy rain, thunderstorm or snow as it should be while the most is during clear, partly cloudy
- **Weekday**- the rentals are more or less equal throughout the week
- **Workingday-** the rentals are high in working days

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

Ans. The idea behind dummy variables is that for a categorical variable with 'n' levels requires 'n-1' levels new columns to define whether that level is existing (1) or not (0) using 0 & 1. If we do not drop the first column it will lead to Dummy Trap leading to the problem of multicollinearity which is a violation of one of the assumptions of Linear regression. It also may lead to incorrect interpretations and wrong insights as change in 1 feature may change the correlated feature too.

For eg, gender contains Male & Female it can be encoded as

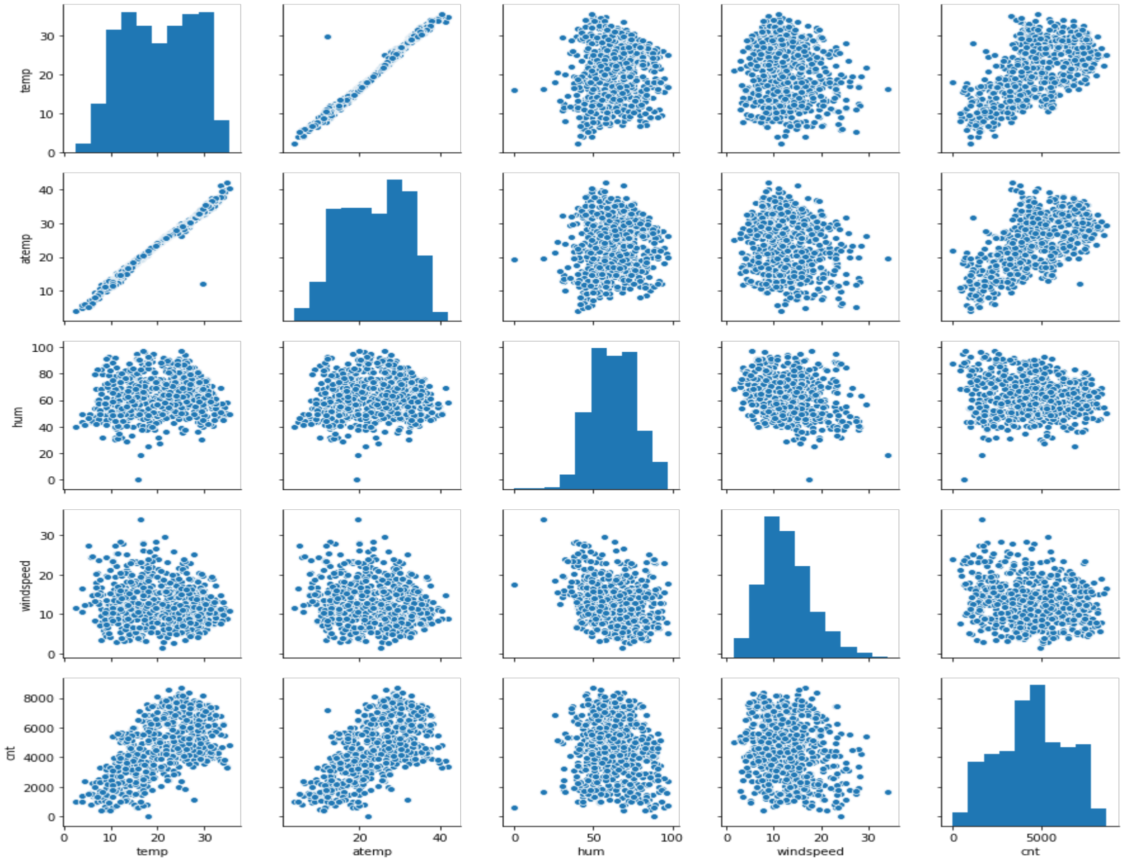


gender	gender_m	gender_f
male	1	0
female	0	1
male	1	0
male	1	0
female	0	1
male	1	0
female	0	1
male	1	0
female	0	1

But, we can drop gender_m yet have the same meaning as if gender_f is 1 then female else it is male. Thus, we remove it.

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

Ans. The features temp and atemp has the highest correlation with cnt

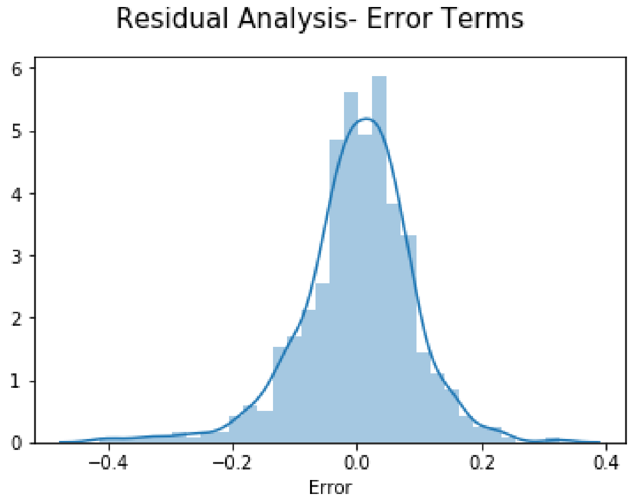


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

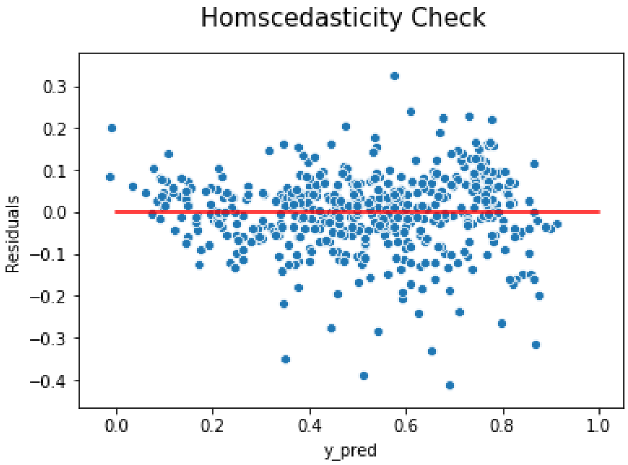
Ans. The assumptions are:

a. Linear Relationship between X and Y
   - The above answer contains the image which depicts the data distribution among the features, it is not exactly linear but temp and atemp is mostly

b. Error terms are normally distributed
   - The error terms are normally distributed with mean = 0



c. Error terms have constant variance (homoscedasticity)
   - There is no definite pattern so we can conclude homoscedasticity is present



d. Error terms are independent of each other
   - The error terms should be independent of each other.

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

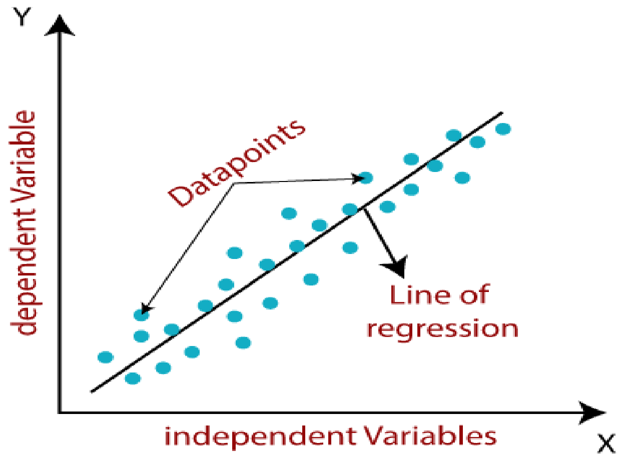
Ans. The features are mentioned based on the coefficient value irrespective of the sign :

- temp
- weathersit_LightSnow_LightRain_Thunderstorm_ScatteredClouds
- yr_2019

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. Linear regression is a machine learning algorithm that finds best linear relationship between independent (X) and dependent (Y) variables on any given data. It is a supervised model mainly used for predictive analysis. It follows the basic equation of straight line **Y = mX + C** and assumes a linear relationship between X and Y. Following the formula it tries to fit a best line with least error using the generated coefficient (m) and intercept(C) .



There are mainly 2 types of linear regression models:

1. Simple Linear Regression (SLR) - it is used when only 1 independent variable predicts the dependent variable
2. Multiple Linear Regression (MLR) - it is used when we have >1 independent variables to predict the dependent variable

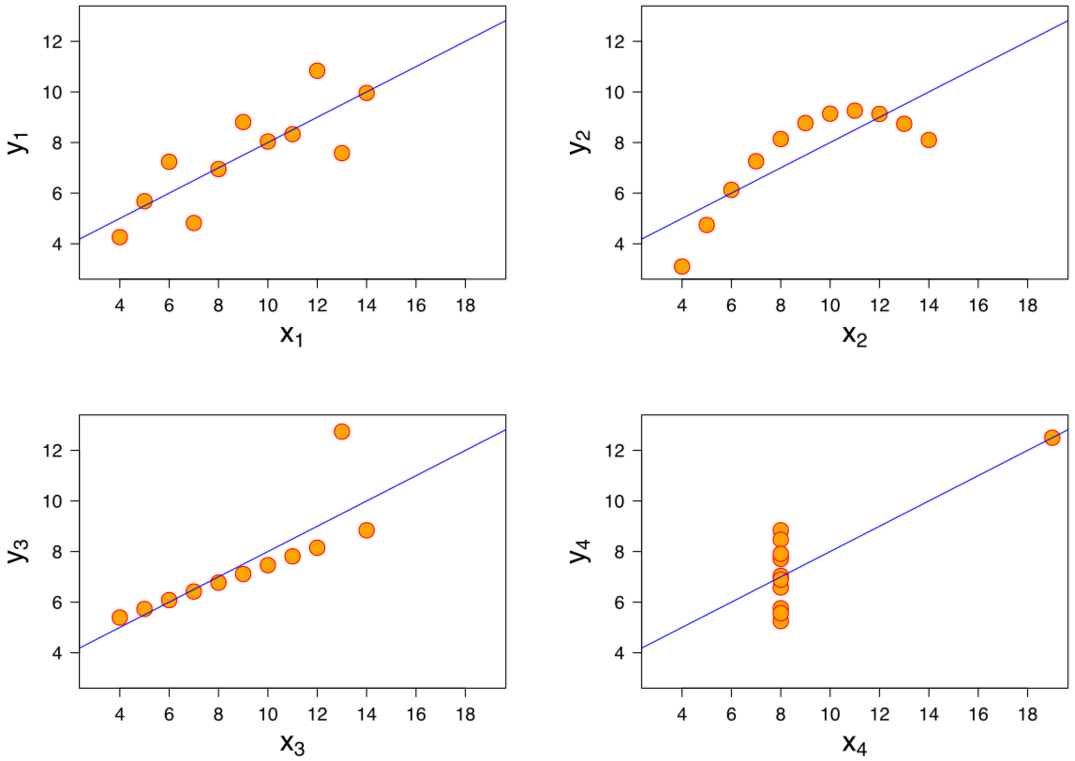
The MLR has the formula modified as :

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots + \beta_i X_i}$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet explains how data with almost identical properties can be different when visualized. It is depicted with the following four plots :



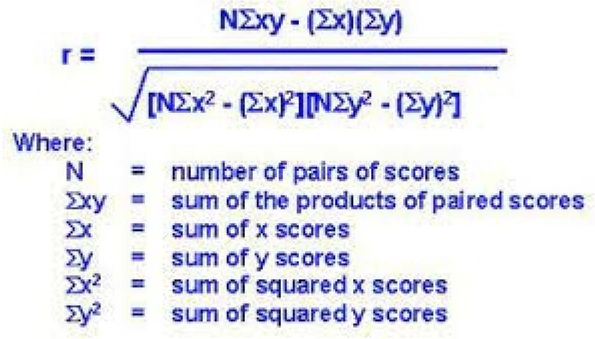
- **Top left** – it shows a normal linear relationship with best fit line
- **Top right** – it is not showing any linear relationship but the model cannot handle it too
- **Bottom Left** – it shows how an outlier can change the regression line, the line should have been different if the outlier was not present
- **Bottom right** – the data points do not suggest a linear relationship but due to one high point the line is drawn differently

The main idea is to always visualize and have a good look at data before modeling it so that our model is not fooled.

3. What is Pearson’s R?

Ans. It is a measure of strength of linear relationship between two variables. It is always between –1 and +1. It basically helps in deciding whether a line can be used to explain the data.

It is calculated using



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

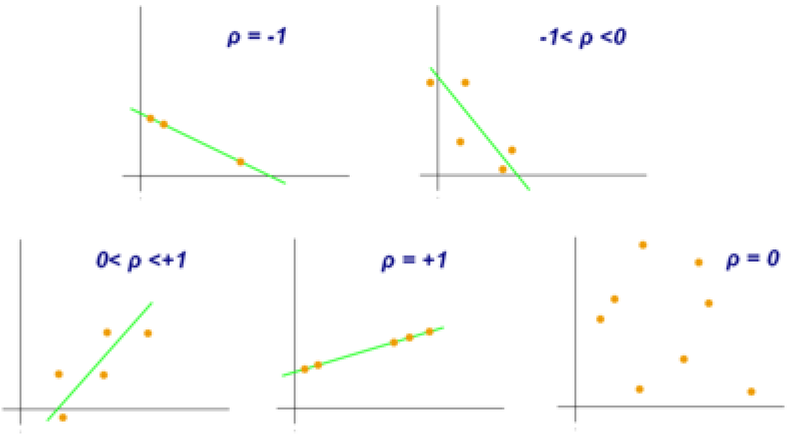
Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

r = -1 : data points have perfect linear relation with negative slope

r= +1 : data points have perfect linear relation with positive slope

r= 0 : no linear relationship



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

Ans. Scaling is the process of normalizing or standardizing the data points within a range. It is mainly executed during data preprocessing stage to deal with differing values.

It is performed to deal with varying values of data points, if not done then the model tends to treat values with high weightage with more priority and low weightage one with less priority irrespective of the units which can lead wrong information to the model.

Normalized scaling is the process of scaling in which all the values are shifted and scaled to lie in the range of 0 to 1. It is also known as Min-Max scaling. It is good to use when we know data does not follow Gaussian distribution

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized scaling is the process of scaling in which the data points are scaled in such a manner that their mean is 0 and standard deviation is 1. It is good to use when we know data follows Gaussian distribution.

$$X' = \frac{X - \mu}{\sigma}$$

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

Ans. Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset. It states how much the variance of the coefficient is inflated by collinearity

$$VIF_i = \frac{1}{1 - R_i^2}$$

VIF being infinite means it is perfectly collinear. Here Ri squared is the R squared value of the independent variable which we want to check. If other independent variables are explaining it perfectly then R squared is 1. Thus, as per the formula if R squared is 1 then

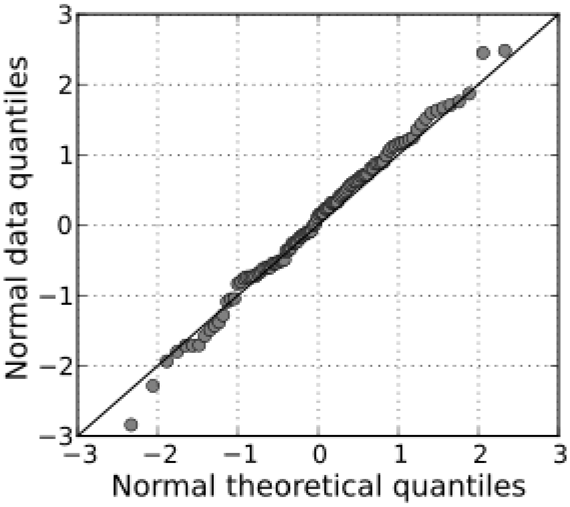
VIF = 1 / (1-1)

=1/0

= infinite

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 the plot of quantiles of first dataset against the quantiles of the second dataset. If the two datasets come from a population having same distribution then it should make a rough straight line along the reference line of 45 degree.



The main use of Q-Q plot is for following scenarios:

- If the 2 datasets are from the population having same distribution
- It can be used to prove the linear regression assumption
- If the 2 datasets have similar tail behaviors
- If they have similar distribution shape