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?

   Answer: This are my observations after the analysis of categorical variables and its effect on the dependent variable:-

   a) As the temperature increases, the count of total rental bikes increases by 43%
   b) If the weathersit is 'Light Snow', the count of rental bikes decreases by 31% and if the weathersit is 'mist', the count decreases by 8% and the demand in both type of weather situation is less than the weather situation when it is 'clear'
   c) For the month of Jul,Nov,Dec,Jan, the count of rental bikes decreases by 6%, 5%,4% and 4% respectively and for the month of Sep, the count increases by 6%. The demand is less than the month of 'Apr' with comparison to Jul, Nov,Dec, Jan and similarly when compared with 'Sep' , the demand is more than in the month of 'Apr'
   d) In winter, the count of rental bikes increases by 7% and in 'Spring', the count decreases by 10% that means in 'winter' , the demand is more as when compared with 'fall' season and in Spring, the demand is less when compared with 'fall' season
   e) If there is a Holiday - There will be 9% less demand for shared bikes

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

   Answer : In dummy variable creation, the total variable will be always be total variable minus 1 i.e if the total variables are 10, then we have to create 9 Dummy variables and we need to use drop_first=True to drop 1 dummy variables so that dummy variables does not become redundant which can adversely effect the model

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

   Answer: 'temp' and 'atemp' are highly correlated

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

   Answer:

   There are various assumption that can be drawn :-

   a) If I can see the residual analysis for train data, I can find that the Error terms are Normally Distributed
   b) There is a Linear relationship between dependent variable('cnt') with any of the independent variable
   c) There is no Multi Collinearity between the Independent variables as the maximum VIF score is of value 2.46

d) If I can see Durbin-Watson value, a value of 2.017 shows that there is less auto-correlation between various variables

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

   Answer: a) Temp- 43% positively correlated

   b) light_snow- 31% negatively correlated

   c) Spring season- 11% negatively correlated

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables.

   There are two types of Linear Regression

   a) Simple Linear Regression (SLR)- Model with 1 independent variable
   It explains the relationship between a dependent variable and an independent variable using a straight line. The independent variable is known as Predictor Variable and the Dependent Variable is known as Target Variable or Output Variable

   The equation is: -
   Y=mX+b
   Where Y is the dependent variable, we are trying to predict
   X is the independent variable we are using to make predictions
   m is the slop of the regression line which represents the effect X has on Y
   b is a constant, known as the Y-intercept

   b) Multiple Linear Regression (MLR)- Model with more than 1 independent variable
   Consider a dataset having **n** observations, **p** features i.e. independent variables and **y** as one response i.e. dependent variable the regression line for p features can be calculated as follows

   Y= bo+b1x1+b2x2+….+bpxp
   Here Y is the predicted value and b0,b1,b2,…,bp are the regression coefficients

   As MLR always include errors in the data known as residual error which changes the calculations as follows

   Y= bo+b1x1+b2x2+….+bpxp+ ei

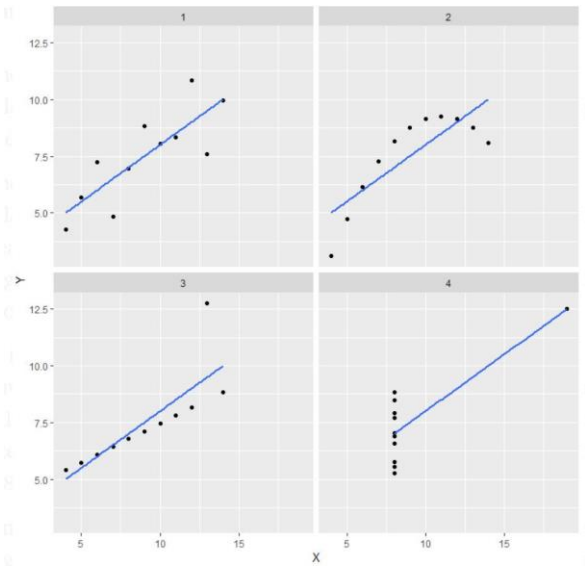
   Where ei=summation of (Yactual – Ypredict)^2

In regression, there is a concept of Best Fit Line which is a line for which the error between the predicted values and actual values is minimum.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet was developed by statistician Francis Anscombe. It is a **set of four datasets** that have the same mean, variance and correlation but look very different.

Each dataset consists of eleven (x,y) points. It is used to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Answer: In statistics, the **Pearson correlation coefficient** (**PCC**), also referred to as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**), or the **bivariate correlation**, is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

Pearson's correlation coefficient, when applied to a **population**, is commonly represented by the Greek letter **ρ (rho)** and may be referred to as the **population correlation coefficient or the population Pearson correlation coefficient**.

Pearson's correlation coefficient, when applied to a **sample**, is commonly represented by **r** and may

be referred to as the **sample correlation coefficient or the sample Pearson correlation coefficient.**

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

   Answer: 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 pre-processing step. When the range of values are very distinct in each column, we need to scale them to the common level. The values are brought to common level and then we can apply further machine learning algorithm to the input data

   Normalized Scaling brings all data in the range of 0-1. It is given as

   X= {x-min(x)} / {max(x)-min(x)}

   Standardized Scaling brings all the data into Standard Normal Distribution with mean as 0 and Standard Deviation as 1. It is given as

   X= {x- mean(x)} / Standard Deviation(x)

   - Normalization is good to use 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 Neighbours 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?

   Answer: As the squared multiple correlation of any predictor variable with the other predictors approaches unity i.e if there is a perfect correlation between all the independent variables, the VIF will be infinite

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

   A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or **theoretical distributions.** The use of Q–Q plots to compare two samples of data can be viewed as a **non-parametric** approach to comparing their underlying distributions

   In linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions

   It has few advantages also such as :-

a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot

As Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set, we can have following interpretation for two data sets

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis