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?**

As per EDA process, we observed that Season,Year,Weathersit,Holiday,Month,Working day and Weekday are categorical variables in given dataset.

These variables influenced our dependendent variable

1. Season: Boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt.Summer and winter had cnt values that were in the middle.

Positive trend in number of customers in 2-Summer,3-Fall and 4-Winter seasons

2.Year: Overall business shows a increasing trend in their user base year on year.

3.Weathersit: In case of weathersituation,whenever there is heavy rain /snow there are no users indicating that the weather is extremely unfavourable.

The highest count was observed when the weather forecast was 'Clear, Partly cloudy'.

4.Holiday: On holidays, the users show a wider spread in the counts. On normal days, the users are more than holidays.

5.Month: Similar to the season trend, there is a postive trend in the months of summer, fall and winter.

September has the most rentals,while December had the fewest.

6.Working day: Minimal effect on dependent variable.

7.Weekday: Weekdays or weekends do show any specific trend.

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

   Dummy variables will result in redundancy and will result in multicollinearity between them. It may also result in negative impact on some models with list of variable importance's may be distorted. We can lose one column to keep everything under control.

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 target variable (cnt).

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

   The distribution of residuals should be normal and centered around 0

   We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not meaning residuals are scattered around mean = 0.

   The linear relationship between target and feature variables should exist.

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 predictor variables that influence bike booking,

   In our case study lr3.params has shown the below mentioned values which are related to Model 3 , according to our final model are:

   - Temperature(temp): With a coefficient of 0.418961, a unit increase in the temp variable increases the number of bike rentals by 0.418961 units.

   - Weather situation 3 [Cloudy]: With coefficient of -0.274585,a unit increase in the Weathersit3 variable reduces the number of bike hires by 0.274585 units as compared to Weathersit1.

   - Year: With a coefficient of 0.236410 , a unit increase in the year variable increases the number of biker rentals by 0.236410

General Subjective Questions

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

   Linear Regression is a type of supervised Machine learning algorithm which is used for prediction of numerical values. Linear Regression is the most basic form of regression analysis. Regression is most commonly used predictive analysis model which is based on mathematical equation "y=mx+c".

   It assumes that there is a linear relationship between dependent variable(y) and the predictor(s) / independent variable(x).In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

   Regression is performed when the dependent variable is of continuous data type and predictors or independent variables could be of any data type like continous,nominal/categorical etc. Regression method tries to find the best fit line which shows relationship between dependent variable and predictors with least error.

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

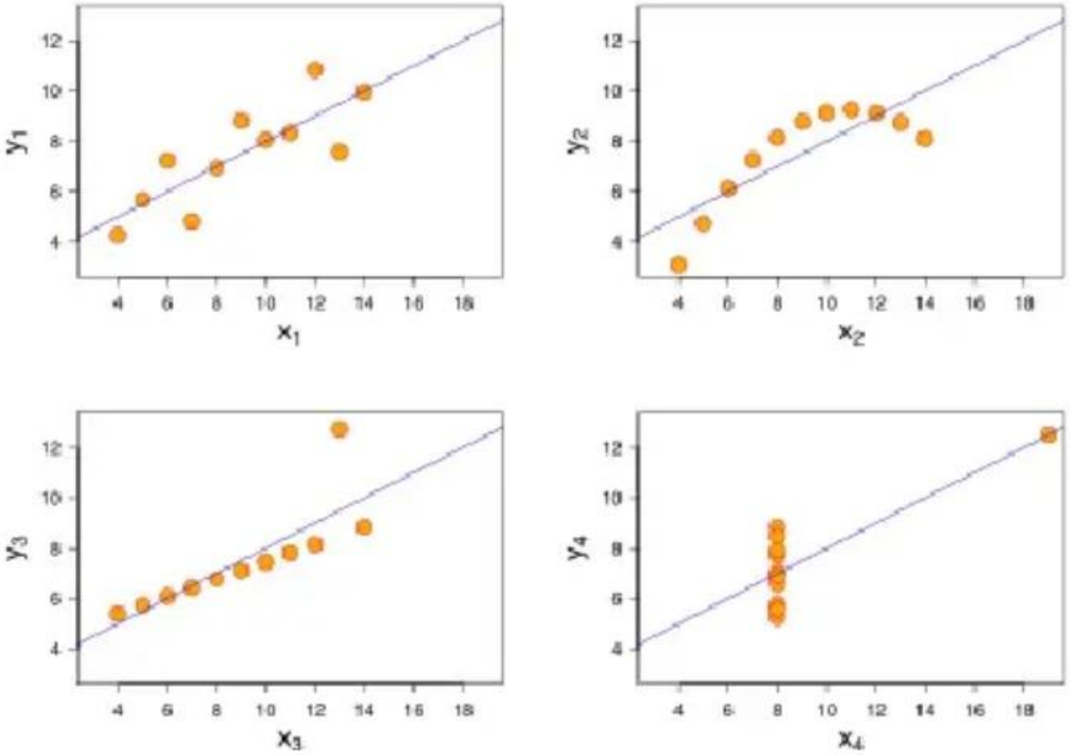
   Regression is broadly divided into simple linear regression and multiple linear regression.

   1.Simple Linear Regression: Used when dependent variable is predicted using only one independent variable.

   2.Multiple Linear Regression: Used when dependent variable is predicted using multiple independent variables.

2. **Explain the Anscombe's quartet in detail.**

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



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). 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?

Pearson's R is a numerical representation of the strength of the linear relationship between the variables. It's value ranges from -1 to +1 which depicts the linear relationship of two sets of data by mentioning if line graph can be drawn to represent the data.

r = 1 means data is perfectly linear with positive slope

r = -1 means data is perfectly linear with 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**

   Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of units of values.

   - Normalization is generally used when we know that distribution of data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of data like K-nearest neighbours and Neural networks.
   - Standardization can be helpful in cases where 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 we have outliers in our 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?**

   VIF - Variance inflation factor: The VIF indicates how much collinearity has increased the variance of the coefficient estimate. VIF is equal to 1/(1-Ri^2).VIF=infinity if there is a perfect correlation where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1.As a result, VIF=1/(1-1) provides VIF=1/0 which is infinity.

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

   The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It’s a tool for comparing the shapes of different distributions such as a Normal, exponential or uniform distribution. A scatterplot generated by plotting two set of quantiles against each other is known as a Q-Q plot.

   Because both sets of quantiles came from the same distribution, the points should form a line that's a fairly straight 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 behaviour?