# Linear Regression Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - In the bike sharing dataset, let's consider the effect of the categorical variable 'weathersit' on the target variable 'cnt'.
    - While performing EDA, I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy) a high number of bike rentals were made, with the median being 50,000 approximately.
    - Similarly, certain inferences could be made 'season' as well. Also, during model building on inclusion of categorical features such as season etc. we saw a significant growth in the value of R-squared and adjusted R-squared.
    - This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset

2. **Why is it important to use drop_first=True during dummy variable creation?**
    - Drop_first=True is important to use because it helps in reducing the extra column created while creation of dummy variable. Hence it reduces the correlations created among dummy variables.
    - For ex, we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is semi-furnished and unfurnished, then its obvious that the 3$^{rd}$ variable is furnished. So, we don't need that 3$^{rd}$ variable to be identify the furnished.
    - Hence, if we have categorical variable with n values, then we need to use n-1 columns to represent dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Looking at the pair-plot and heat map, the numerical variable 'temp' and 'atemp' both have 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?**
    - Assumption 1: The dependent variable and independent variable must have a linear relationship.
      A pair plot can help us to validate if the independent variable exhibits relationship with dependent variable.
    - Assumption 2: No correlation in residuals.
      DW statistics must lie between 0 and 4. If DW=2, implies no autocorrelation. If between 0 and 2 implies positive correlation. If between 2 and 4implies negative correlation.

- Assumption 3: No heteroskedasticity.
  Residual vs fitted values plot can tell if Heteroskedasticity is present or not. If the plot shows tunnel shape pattern, then Heteroskedasticity is present.
- Assumption 4: No prefect Multicollinearity.
  In case of less variables we can use heatmap. Another way to check is to calculate VIF. If VIF <= 2 less/no multicollinearity, VIF >=10 very high multicollinearity.
- Assumption 5: Residuals must be normally distributed.
  Using distribution plot on the residuals and see if it is normally distributed.

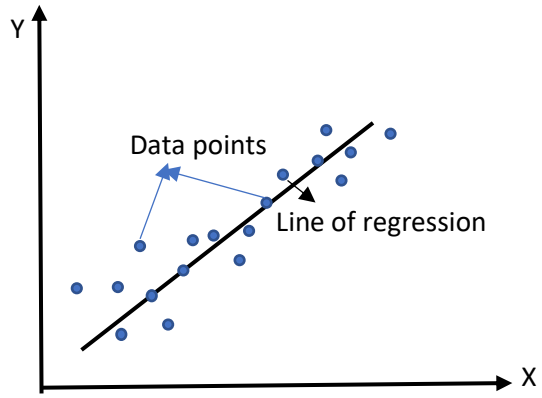
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

   The top three features contributing significantly towards explaining the demand of shared bikes are:

   - Temp
   - Year
   - Winter season

# General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   - Regression is a method of modelling a target value based on independent predictors.
   - Linear Regression is the most popular Machine Learning algorithms. Linear Regression is a statistical method that is used for predictive analysis. It makes the predictions for continuous/real or numeric variables.
   - Linear Regression algorithm shows a linear relationship between a dependent variable (y) and one or more independent variables. It finds how the value of the dependent variable is changing according to the values of the independent variable.
   - The linear regression model provides a sloped straight line representing the relationship between the variables.




Mathematically we can represent linear regression as:

**$Y=\beta 0+\beta 1x$**

Y=dependent variable (target/output variable)
x= independent variable (predictor variable)
β0=intercept of the line
β1= slope

- Linear Regression is further divided into:
  1. Simple Linear Regression: If a single independent variable is used to predict the values of a numerical dependent variable, then such a linear regression algorithm is called as Simple Linear Regression.
  2. Multiple Linear Regression: If more than one variable is used to predict the values of a numerical dependent variable, then such a liner regression algorithm is called as Multiple Linear Regression.

- When working with Linear Regression, main goal is to find the best fit line that means the error between the predicted values and the actual values should be minimized. The best fit line should have the least error.
- Cost Function helps us to figure out the best possible values for β0, β1 which would provide the best fit line for the data points

Assumptions of Linear Regression:

- Linear regression assumes the linear relationship between the dependent and independent variable.
- Multicollinearity means high-correlation between the independent variables. Due to this, it may be difficult to find the true relationship between the predicted and target variables. So, the model assumes either little or no multicollinearity between the variables.
- Error terms is the same for all the values of the independent variables.
- Linear regression assume that the error terms should follow the normal distribution pattern. If the error terms are not normally distributed, the confidence interval will become either too large or small, which may cause the difficulty in finding coefficients.
- The linear regression model assumes no autocorrelation in error terms.

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

- **Anscombe's Quartet** can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
- There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
- When the statistical information of these four datasets are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities.
- The four datasets can be described as:
  1. **Dataset 1:** this **fits** the linear regression model pretty well.
  2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
  3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
  4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

**3. What is Pearson's R?**

- Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
- For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed.
- There should be no significant outliers. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient.
- Each variable should be continuous.
- The two variables have a linear relationship. Scatter plots will help to tell whether the variables have a linear relationship. If the data points have a straight line, then the data satisfies the linearity assumption.
- The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.
- Homoscedasticity describes a situation in which the error term is the same across all values of the **independent** variables. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.
- Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

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

- It is a step of data pre-processing which is applied to independent variables to normalize the data within a range. It also helps in speeding up the calculations in an algorithm.
- If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- 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 must do scaling to bring all the variables to the same level of magnitude.
- Scaling only affects the coefficients but not the other parameters like t-statistics, F-statistics, p-values, R-squared.

  **Normalized/ Min-Max Scaling:**

  1.This technique re-scales a feature or observation value with distribution value between 0 and 1.

  2.It brings all the data in the range of 0 and 1.

  3. Sklearn.preprocessing.MinMaxScalar helps to implement normalization in python.

**MinMaxScaling:** $x = \frac{x - \min(x)}{\max(x) - \min(x)}$

**Standardized Scaling:**

1.It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.
2. Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
3.sklearn.preprocessing.scale helps to implement standardization in python.

**Standardization:** $x=\frac{\mathbf{x-mean(x)}}{\mathbf{sd(x)}}$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

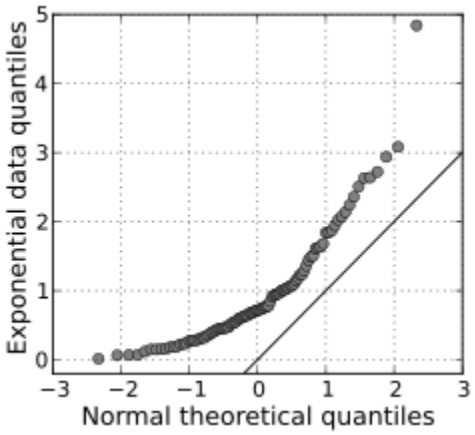
- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

$$\mathbf{VIF=}\frac{1}{1-R^2}$$

- The higher the value the greater the correlation of the variables. Values more that 5 is regarded as high VIF.
- If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
- In case of perfect correlation, we get R^2=1, which lead to 1/(1-R^2) infinity.
- To drop this, 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 variables 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.**

- 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.




- If two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y=x.
- If the distribution is linearly related, the points in the Q-Q plot will approximately lie on the line, but not necessarily on the line y=x.
- Q-Q plots can be used as graphical means of estimating parameters in a location-scale family of distribution.