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?

   The dependent variable is highly dependent on categorical variable such as **season, weather, workingday** etc.
   Season: company should focus on summer & fall season
   Weather: Users prefer when the weather is pleasant.

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

   drop_first=True is very important, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   temp & **atemp** have the highest correlation with target variable (cnt). Also, temp & atemp are highly correlated (0.99) so we can drop one feature.

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

   By checking the assumption of normality using histogram & checking assumption of homoscedasticity and autocorrelation. Checking the Residual Analysis using histogram. The errors should be normally distributed.

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

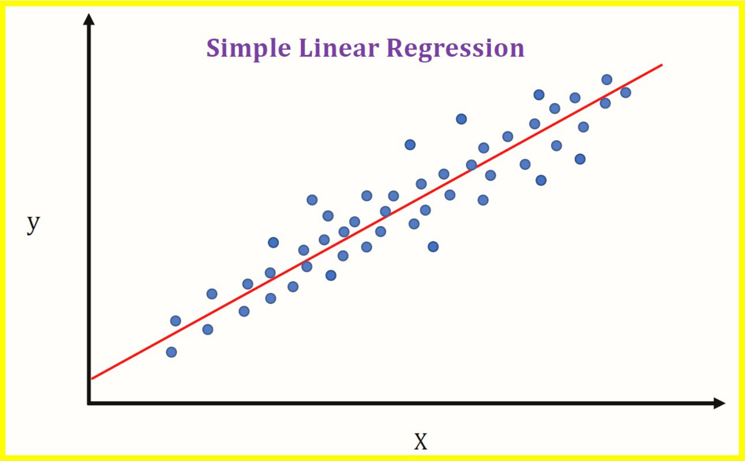
   The below features contribute significantly:

   - year: company should wait for pandemic to get over and things become normal
   - season: company should focus on Summer & Fall
   - weather: Users prefer the bikes when the weather is pleasant
   - temp: Users prefer bikes when the temperature is moderate

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Regression algorithm is a statistical method to model the relationship between a dependent and independent variables with one or more independent variables Regression helps us to understand how the value of dependent variable changes with corresponding to independent variable. It predicts continuous values like temperature, rent, salary etc.



   The red line in the above graph is referred as the best fit straight line. Based on the given data points we try to plot a line that models the points the best.

   The mathematical equation for Linear regression is Y= a + bX

   Where: Y: dependent variable (target variable)

   X: independent variable (predictor variable)

   b: the slope of the line

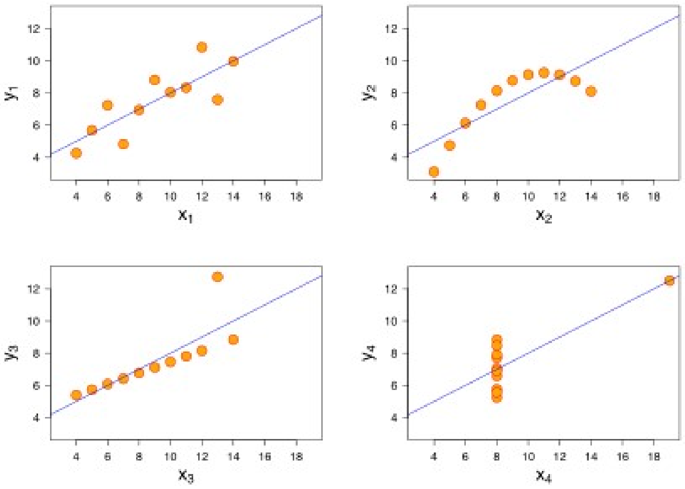
   a: intercept (the value of y when x=0)

   We also need to understand the concept of Gradient descent. This is a method to update **a** & **b** to reduce the cost function (MSE). The idea is to try iteratively the points and find the best fit line and with least error function.

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.
   It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. 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.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.



3. What is Pearson's R?

In statistics, the Pearson correlation coefficient, 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[a].
It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844.The naming of the coefficient is thus an example of Stigler's Law.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
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.

Given a pair of random variables (X,Y), the formula for ρ is
ρX, Y = cov(X, Y)/sigma(X) * sigma(Y)
where:
- cov is the covariance
sigma(X) is the standard deviation of X
sigma(Y) is the standard deviation of Y

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

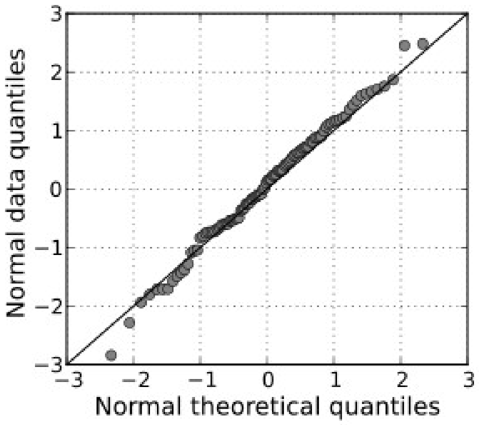
   Scaling is a method used to normalize the range of independent variables or features of data. It is also known as data normalization. This is usually performed during the data pre-processing step.
   Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant numbers starts playing a more decisive role while training the model.
   The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represent completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same. So these more significant numbers starts playing a more decisive role while training the model. Thus feature scaling is needed to bring every feature in the same footing without any upfront importance.
   Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it.
   Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
   Standardization is another scaling technique where the values are centred 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.

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

   This shows a perfect correlation between two independent variables. These VIFs tell you there is perfect collinearity. 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.

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

   In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.



The purpose of Q-Q plots is to find out if two sets of data come from the same distribution or not. 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.

- The slope shows us the steps in the dataset are too big or too small. For example, if we have N observations, then each step traverses 1/(N-1) of the data.
- A sharply sloping section of the Q-Q plot shows that the observations are more spread out than if they were normally distributed.
- A flat Q-Q plot means that the data is more grouped together than we would expect from a normal distribution. The Q-Q plot of a uniformly distributed variable will have a very shallow slope.