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?**

    1. Fall has the highest bike demand
    2. The demand is increasing monthly till sep reaching max and then it is going down
    3. Not much difference in demand for bikes daywise
    4. Most demand on Clear weathersit condition
    5. The demand is decreased on holidays
    6. Demand has increased a lot compared to previous year

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

    drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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 the target variable (cnt).

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    I have done following tests to validate assumptions of Linear Regression:

    a. There should be linear relationship between independent and dependent variables. In my pairplot the linear relationship is observed between cnt and temp,atemp
    b. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). When the plotted histogram of residuals, it was satisfying above condition

    c. linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. I calculated the VIF (Variance Inflation Factor) and it is not greater than 5

    d. Plotted a scatterplot between residuals and ytrain and there was no visible trend confirming Homoscedasity

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 significant features are:
        1. temp - coefficient : 0.4907
        2. year - coefficient : 0.2335
        3. weathersit_Light Snow & Rain - coefficient : -0.29 approx

General Subjective Questions

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

   Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.

   The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:

   $y = \theta 0 + \theta 1x$

   where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

   To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

   $MSE = (1/n) * \sum(y - y')^2$

   where n is the number of data points, y is the actual value, and y' is the predicted value.

   The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

   Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

   $y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

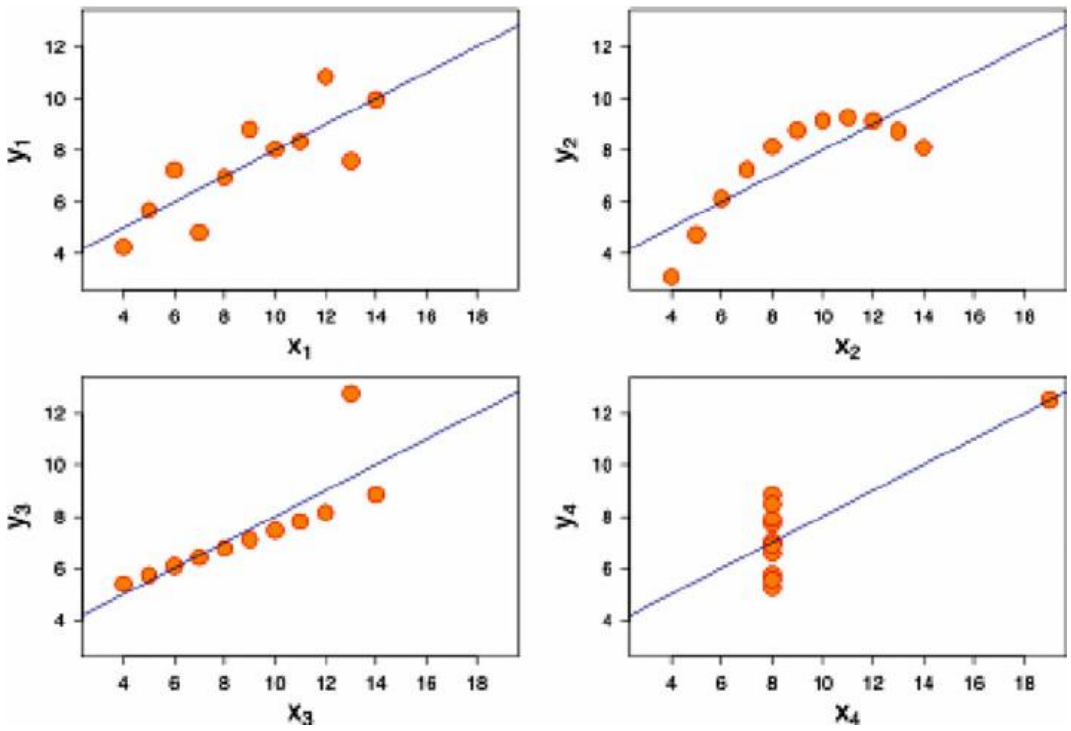
   Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

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 statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

   - The first scatter plot (top left) appears to be a simple linear relationship.
   - The second graph (top right) is not distributed normally; while there is a relation between them is not linear.

● In the third graph (bottom left), the distribution is linear, but should have a different regression line 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 summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data?"

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

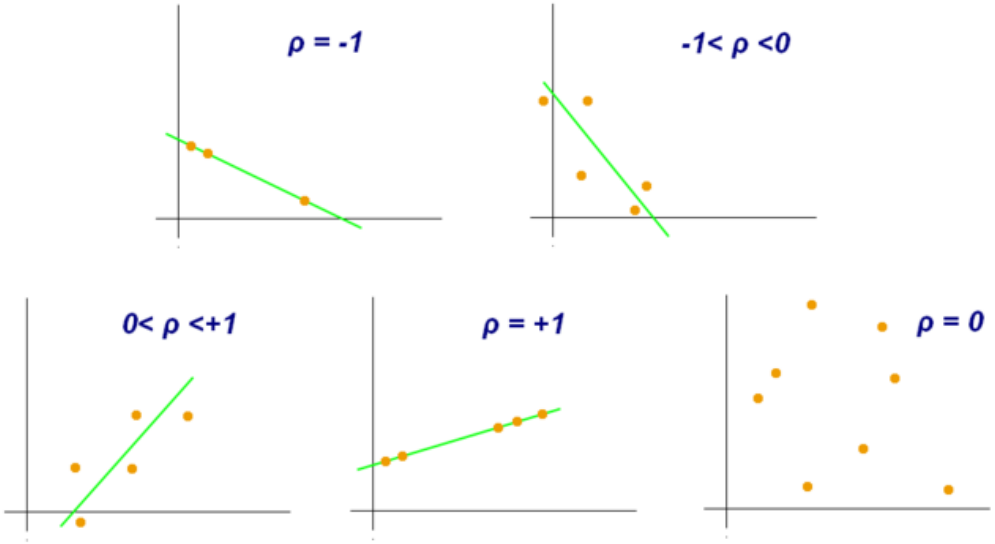
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a 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 technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue. S.NO. Normalized scaling Standardized scaling 1. Minimum and maximum value of features are

used for scaling Mean and standard deviation is used for scaling. 2. It is used when features are of different scales. It is used when we want to ensure zero mean and unit standard deviation. 3. Scales values between [0, 1] or [-1, 1]. It is not bounded to a certain range. 4. It is really affected by outliers. It is much less affected by outliers. 5. Scikit-Learn provides a transformer called MinMaxScaler for Normalization. Scikit-Learn provides a transformer called StandardScaler for standardization.

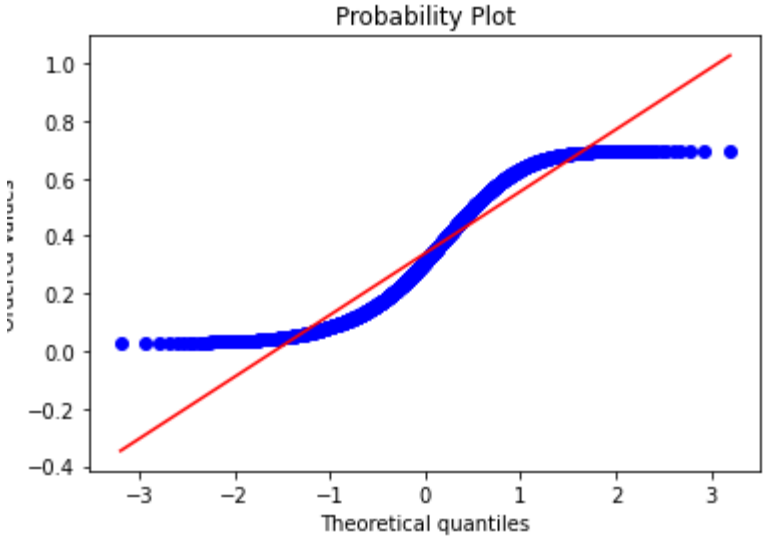
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this 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.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests


Probability Plot
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−3
−2
−1
0
1
2
3
Theoretical quantiles