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 categorical variables have an impact on the final count / usage prediction. Below are the observations,

   ➔ Seasons have an impact on the count / usage.

   Spring has a coefficient of -0.13 which shows a significant negative impact on usage.
   Winter shows a slight increase in the usage with a coefficient on 0.06

   ➔ Weather conditions have a significant impact on usage.

   Light rain, snow and thunderstorm has significant negative impact (second highest coefficient after temperature) on usage / count.
   Misty / Cloudy conditions also have a negative impact on the final count / usage.

   ➔ Months have an impact on count / usage.

   Some months like March(0.03) and September(0.06) shows slight increase in usage or counts whereas months like Nov(-0.04), Dec(-0.04), Jul(-0.06) shows slight decrease.

2. Why is it important to use drop_first=True during dummy variable creation?
   When pandas create dummy variables for a categorical variable, it creates as many dummies as there are categories. For examples in this case for seasons, it will create 4 dummy variables as there are four (n) seasons((1:spring, 2:summer, 3:fall, 4:winter). However to represent these variables we only need 3(n-1) dummy variables. This means we drop one variable by specifying 'drop_first=True' while creating dummy variables using pandas 'get_dummies' call. This also prevents multi collinearity as if we keep four variables the sum of the four variables will always be 1 whereas if we keep only 3 variables this will not be the case allowing the model to interpret the impact of the variables clearly.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Temperature (atemp) has the highest correlation with a positive correlation of 0.65.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Please find below assumptions and how they are validated.

   ➔ Linearity

   There is linearity between variables which is clear from the pair plot. However, after building the model, plotting residuals vs fitted values did not show any discernible pattern which is a check that confirms linearity.

   ➔ The residuals are independent of each other

   Calculated the Durbin-Watson statistic which came out to be 2.0. As per the Durbin-Watson test, if the statistic comes out to be 2.0, there is no correlation between the residuals. In other words the residuals are independent.

   ➔ The residuals have constant variance at every level of the independent variables.

   Plotted residuals vs fitted values and this follows no pattern and are scattered around the horizontal axis. This proves that residuals have constant variance.

- ➔ The independent variables does not have collinearity (Multicollinearity)

  ViF or variance inflation factor is a measure that is arrived at by trying to predict an independent variable using all other independent variables trying to find how a variable is related to the others (Multicollinearity). After creating the final model, ViF for each of the feature in the final feature set was calculated and they were all below 5 which confirms there is no significant collinearity between the independent variables.
- ➔ The residuals are normally distributed

  Plotted the distribution of residuals and as expected they are normally distributed.
- ➔ Independent variables are not correlated with error term.

  Plotted residuals vs fitted values and this follows no pattern and are scattered around the horizontal axis which also confirms no correlation between independent variables and error terms / residuals.

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

   Top 3 features with highest impact are below,

   1) Temperature(atemp) with a coefficient of 0.44
   2) Weather condition **Light Snow, Light Rain and Thunderstorm / Scattered clouds** with a coefficient of -0.29
   3) Year(yr) with a coefficient of 0.24

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression is a statistical method to predict a continuous variable (say 'y' – also can be called a dependant variable or target variable) dependant on one or more independent variables (say – x1, x2, x3…xn – also called predictors). This is assuming that there is a linear relationship between the target variable 'y' and the predictors.

   There are broadly two types of linear regressions based on the number of predictors as below,

   - ➔ Simple Linear Regression (SLR):

     There is only one predictor and one target variable in SLR. This is represented by y = B0 + B1x + e (y = beta zero + beta1 times x + error (to capture unexplained variance)

     In this case:

     y: dependant / target variable that is to be predicted

     B0: intercept (this where line crosses the y-axis)

     B1: Slope (rate of change of y with respect to x, in other words coefficient of x which determines how much 'y' will change with respect to a unit change in x)

     E: error term to capture unexplained variance
   - ➔ Multiple Linear Regression (MLR):

     MLR is more or less the same as SLR, in this case the difference is that we will have multiple independent / predictor variables to predict the target / dependant variable y.

     In this case the equation changes to y = B0 + B1x1 + B2x2……..BnXn + e where,

x1…..xn: independent / predictor variables
B0: Intercept
B1….Bn: coefficients of each predictor (how much y will change for a unit change in corresponding predictor variable provided the coefficients of other predictor variables are kept constant)
E: error term to capure unexplained variance

Linear Regression assumes the following for the model to work.

Linearity: There is a linear relationship between the predictor variables and the target variables.
Independence: Observations are independent of each other
Homoscedasticity: The variance of residuals (this is the difference between actual value and predicted value of target variable) is constant across all levels of predictors
Normality of residuals: Residuals (this is the difference between actual value and predicted value of target variable) are normally distributed and centred around zero.
No perfect Multicollinearity: The predictors are not correlated to each other, in other words predictor variables are independent / not related to each other

The predictions are made by fitting a line based on the data / observations we have to train the model. The idea is that the fitted line holds the predictions. A best fit line is arrived at by using a technique called RSS (Residual sum of squares). Residuals are the difference between actual value and predicted value and the idea is that the sum of the squares of residuals have to be the minimum for the model to have the most accurate prediction. The idea is that the lowest sum of residuals will have the lowest variance of the predicted value from actual value. In other words this will explain the most variance in the data. So a line is arrived at using this technique which will be – 'y = B0 + B1x1 + B2x2……..BnXn' – giving us the coefficients of each predictor variable there by quantifying their impact on the target variable.

The model is trained with the existing data to arrive at a best fit line using the above technique. Which can be then used for prediction.

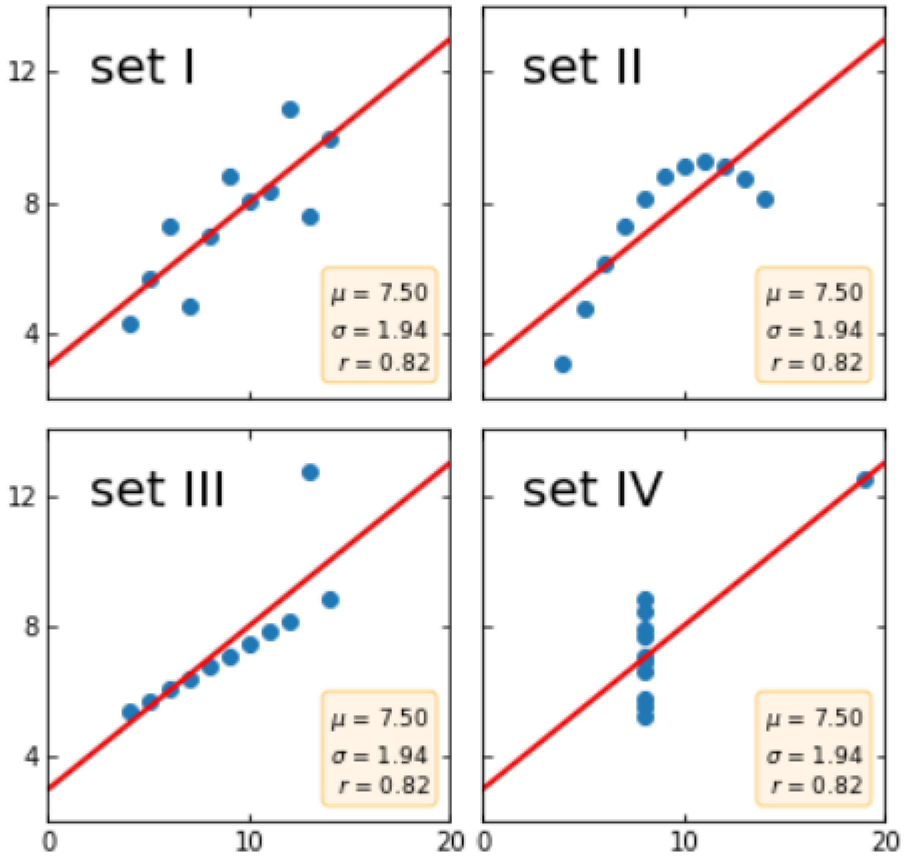
2. Explain the Anscombe's quartet in detail.
Anscombe's quartet is a set of four datasets created by statistician Francis Anscombe in 1973. The idea of the datasets is to demonstrate how different sets of data can have similar statistical properties like mean variance etc, yet look very different when visualized / graphed. Each dataset have eleven data points representing points on an 'x,y' coordinate.
Below are the 4 datasets,
Dataset 1:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
Dataset 2:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
Dataset 3:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
Dataset 4:

x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

The statistical properties of these datasets are same as below,

Mean of x in all 4 datasets is 9
Mean of y in all 4 datasets is 7.5
Variance of x in all 4 datasets is 11
Variance of y in all 4 datasets is 4.127
Correlation between x and y in all 4 datasets: 0.816
Linear regression line: y = 3 + 0.5x

Despite identical statistical properties the datasets when plotted looks very different. Please find below the screenshot (screenshot taken from my jupyter notebook)



This proves that summary statistics alone does not tell the whole story. Visualizing the data is very important and it can uncover patterns in the data that is otherwise not understood just using summary statistics.   In order to understand the data, a combination of graphical (visualizations) and numerical analysis is required.

This also tells us things like outliers can heavily influence statistical measures and knowing the domain / context of data is also extremely important. To quote the example shared in the EDA session in the course if we include Mr Bill Gates in the data for IT employees salary, this can heavily influence summary statistics like mean.

3. What is Pearson’s R?
   Pearson’s R (also known as Pearson’s correlation coefficient) is a measure of linear correlation between two variables. Pearson’s R quantifies the degree to which the variables are correlated and the value lies between -1 and 1 where 1 means a perfect positive correlation and -1 means a perfect negative correlation and zero means no linear relationship at all between the variables. Below is the formula to compute Pearson’s R (using an image acquired from internet)

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

In the above formula;
‘r’: Pearson’s R
x and y are individual sample points
x bar and y bar are the means of X and Y.

The numerator is basically the covariance of x and y / joint variability of x and y. As can be seen, this is basically the sum of products of deviations of x and y from their respective mean.

The denominator is the product of standard deviations of x and y

If the Pearson’s R is 1, it means that when x increases, y also increases perfectly in a linear manner

If the Pearson’s R is 0, it means that there is no linear relationship between x and y

If the Pearson’s R is -1, it means that when x increases, y decreases in a perfectly linear manner

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Scaling is a technique used in Machine Learning to adjust the range of values of features.

   This is done to make sure that the numerical values of all features lie in the same range, thereby ensuring no feature dominates the learning process due to its scale. For example, in Linear regression if the scales are different, the coefficients of individual features cannot be compared for their impact on the target variable. A feature with a higher scale may have a much higher coefficient than a feature with a lower scale there by making a comparison of the impact of features difficult. Changing the features to same scale makes comparing their impact easier.
   Scaling of variables also makes methods like Gradient descent work faster as the features are of the same scale.

   Normalized Scaling also known as MinMax scaling:
   This will convert the values of feature between into a specific range usually between 0 and 1. This is done by using minimum and maximum value of the feature and hence the name MinMax scaling.
   Formula for scaling;

X(normalized) = x – min(x) / max(x) – min(x)
Where:
X(normalized) is the normalized value of x
Min(x) is the minimum value of the feature x
Max(x) is the maximum value of the feature x

Standardised scaling also knows as Z-score Normalisation:
This scaling technique transforms the data in such a way that the resultant distribution of data after transformation will have a mean of 0 and standard deviation of 1. This is done by subtracting the mean from the value of the feature and dividing by standard deviation.
Formula for scaling:
X(normalized) = x-mu / sigma
Where:
X(normalised) is the normalized value of feature x
'mu' is the mean of feature x
'sigma' is the standard deviation of the feature x

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
ViF or variance inflation factor is an indicator of multicollinearity. The ViF of a particular feature / variable indicates if that feature can be explained by a combination of one or more other features making it redundant in the model (does not explain anything about the target variable that other variables does not yet provide). Generally, a ViF of 5 or more is considered to have high multicollinearity, in other words a feature with a ViF of 5 or more is explained by a combination of other features.
The formula for ViF is as below;
ViF(x) = 1 / 1-Rsquare
Where RSquare is the coefficient of determination of the regression of x on all other predictors.

So as per the above formula for ViF of a variable to become infinite, the Rsquare should be 1 which will make the denominator 0. This will happen when X can be perfectly explained by a linear combination of other variables or fearures / predictors.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Q-Q plot or Quantile-Quantile plot is a technique used to identify if dataset comes from certain type of distribution (say normal distribution). It plots the quantiles of the given data against quantiles of the theoretical distribution (say normal distribution) and if the points lie approximately along a straight line we can conclude that the given data follows the theoretical distribution.

In Linear regression Q-Q plot is used to confirm the assumptions pertaining to error terms or residuals. In linear regression the error terms are assumed to be normally distributed. Q-Q plot is used to plot the quantiles of residuals against quantiles of a normal distribution to confirm if the residuals are normally distributed.