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?

- Categorical variables such as *'season'*, *'yr'*, *'holiday'*, *'workingday'*, *'weathersit'* from the dataset are having effect on the dependent variable *'cnt'*. Among these categorical variables, variable *'yr' has most effect on the dependent variable 'cnt'*.

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

- Dummy variables are useful as they enable us to use a single regression equation to represent multiple groups present in the dataset. This helps us in reducing the extra column creation while we create dummy variable, which further 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?

- Variable *'atemp'* has the highest correlation with the target variable *'cnt'* in the given dataset.
  The second variable having highest correlation with the target variable *'cnt'* is *'temp'*.
  In the python code submitted by me, I have dropped the variable *'atemp'* and used *'temp'* variable, hence *'temp'* has the highest correlation with the target variable *'cnt'* in my python code.

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

- Validating the assumptions of Linear Regression after building the model on the training set is done by pair-wise scatterplots as linear regression assumes that there exists a linear relationship between the dependent variable and the independent variable and also it is easy to visualize a linear relationship on a plot.

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

- The *'temp'*, *'season'*, *'yr'* are top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a statistical regression method used for predictive analysis that shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). If there is a single input variable, such linear regression is called Simple Linear Regression and if there is more than one input variable, such linear regression is called Multiple Linear Regression**.** The linear regression model gives a sloped straight line describing the relationship within the variables. To calculate best-fit line, linear regression uses a traditional slope-intercept form as given below:

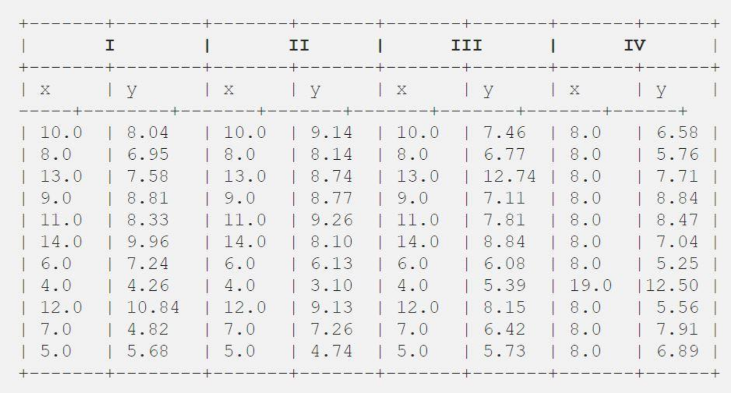
  $$y = mx+c \longrightarrow y=a0+a1x$$

  Where,
  y= Dependent Variable.
  x= Independent Variable.
  a0= intercept of the line.
  a1 = Linear regression coefficient.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, though appearing very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

  Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data points are given below.

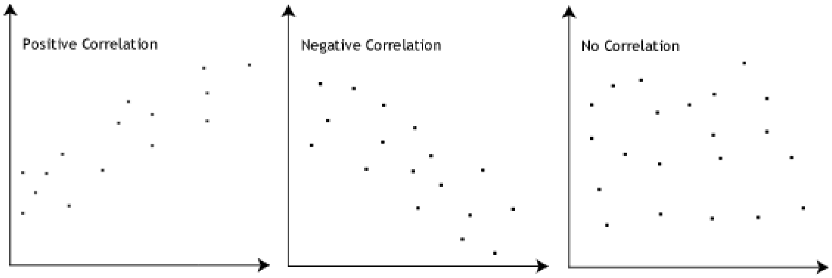


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

  After that, the council analyzed these sets & data points using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson's R?

- The Pearson's correlation coefficient varies between -1 and +1 where:
  r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
  r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
  r = 0 means there is no linear association
  r > 0 < 5 means there is a weak association
  r > 5 < 8 means there is a moderate association
  r > 8 means there is a strong association



Pearson R Formula

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

Where,
r = correlation coefficient
xi = values of the x-variable in a sample
$\overline{x}$ = mean of the values of the x-variable
yi = values of the y-variable in a sample
$\overline{y}$ = mean of the values of the y-variable

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

- Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range commonly 0 to 1. It also helps in speeding up the calculations in an algorithm. 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 have to do scaling to bring all the variables to the same level of magnitude.
  It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x\text{-}min(x)}{max(x)\text{-}min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

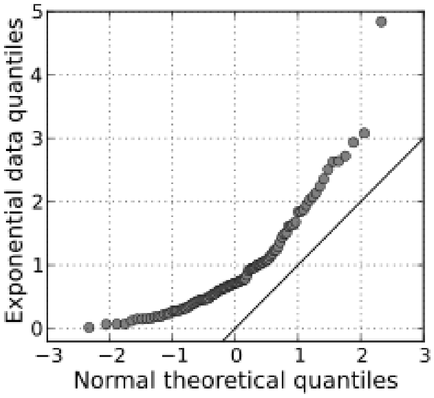
$$\text{Standardization: } x = \frac{x\text{-}mean(x)}{sd(x)}$$

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

- If there is perfect correlation between two independent variables, then VIF = infinity. 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. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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. 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.
A Q-Q plot showing the 45 degree reference line:



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

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.