Assignment based Subjective Question

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   A: Below are the observations:
   - More users tends to use on Holidays
   - There is an increasing trend in users year on year
   - People uses more bikes when its clear+partly cloud conditions. Volume drops during rains and snow.
   - Sat and Wed attracted more users
   - There is a positive trend from summer to fall to winter
   - For months, there is an increasing trend from Mar to Oct
   - Bike demand in working and non working day has almost no significant change

2. Why is it important to use drop_first=True during dummy variable creation?
   A. Setting drop_first=true helps in reducing the extra column created during the dummy variables creation step. If we don't set it to true and use the default value of false then it will create the dummy variables which are highly corelated leading to the Dummy variable trap. Dropping it reduces the correlations among the dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   A. temp and atemp has a high corr with cnt which is almost same for both (0.63)

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   A. Assumptions validated as below:
      a. Using displot - NO pattern visible in the residual indicating that the error terms are independent.
      b. Pairplot indicating no specific patter
      c. Durbin-Watsn values close to 2(1.91) indicating no first-order auto correlation
      d. VIF for all feature variable < 5 indicating there is no multicollinearity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   A. Below are the top three factors:
      a. atemp
      b. windspeed
      c. year

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of the data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = mx + b$$

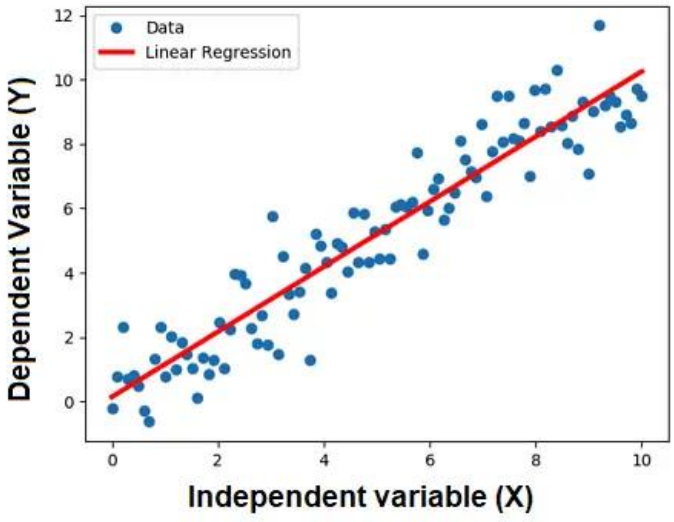
Here, x and y are two variables on the regression line.

b = Slope of the line
m = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset

However, machine learning experts, have a different notation to the above slope-line equation,

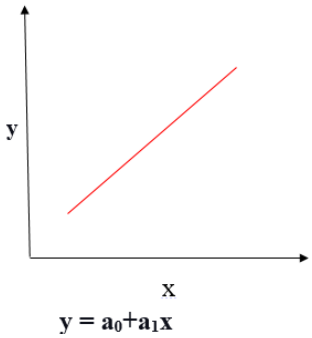
$$y(x) = p0 + p1 * x$$

where,
y = output variable. Variable y represents the continuous value that the model tries to predict.
x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.
p0 = y-axis intercept (or the bias term).
p1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.
pi = weights (in general).
Thus, regression modeling is all about finding the values for the unknown parameters of the equation, i.e., values for p0 and p1 (weights).



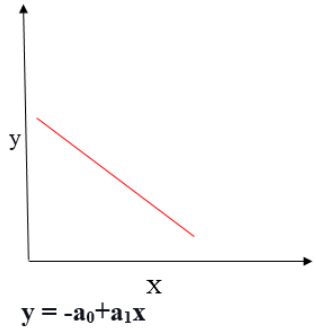
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

1. Positive Linear Regression: If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



2. Negative Linear Regression

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



2. Explain the Anscombe's quartet in detail

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven ***(x,y) points.***

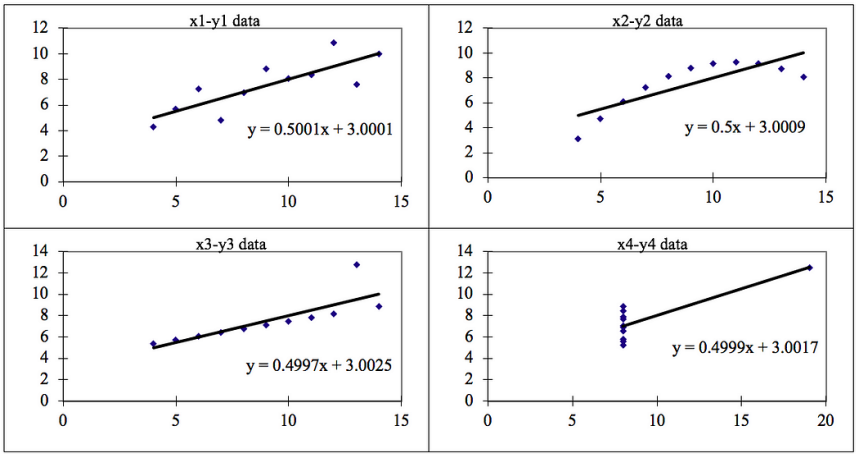
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.The datasets are as follows. The linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

The statistical information for these four data sets are approximately similar. We can compute

them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE'S QUARTET FOUR DATASETS:

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?
   Ans: In Statistics, the **Pearson's Correlation Coefficient** is also referred to as **Pearson's r**, **the Pearson product-moment correlation coefficient (PPMCC),** or **bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

For Population:

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.

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

For Sample:

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

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

   Ans: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. 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.

   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.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   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).

$$X' = \frac{X - \mu}{\sigma}$$

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

   Ans: The value of VIF is calculated by the below formula:

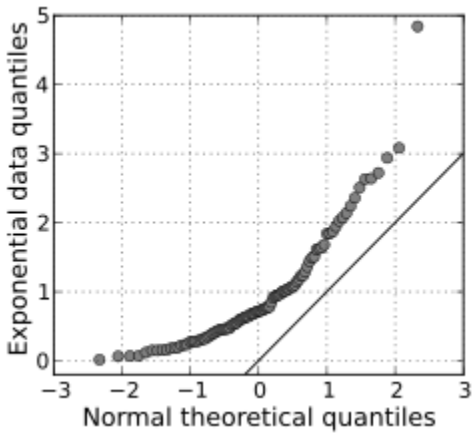
$$VIF_i = \frac{1}{1 - {R_i}^2}$$

   Where, 'i' refers to the ith variable.

   If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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