Ravinder Singh

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? (3 marks)

   Answer: With the help of boxplot, I have plotted categorical variables against dependent variables and following are point of analysis obtained:

   - Season: 3: fall has high demand for rental bikes
   - Observed that demand for next year has grown
   - Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
   - There is a decreased in demand for holiday.
   - Weekday is not giving clear picture about demand

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

   Answer: 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. In,Dummy variables drop_first=True drop the main column from where we create dummy column after we use entire data from main column. As we have all the data in new columns. So, there is no use of keeping main column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: The feature “temp” has highest correlation. It is very well linearly related with target “cnt”

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer: I have validated the following assumptions:

   - Error terms are normally distributed with mean 0.
   - Error Terms do not follow any pattern.
   - Multicollinearity check using VIF(s).
   - Linearity Check.
   - Ensured the overfitting by comparing the R2 value and Adjusted R2

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: “temp”, “year “and “season” are top 3 contributing features.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression Algorithm also known as simple linear regression is a machine learning algorithm based on supervised learning. it is a statistical technique for investigating and modeling the relationship between variables. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of 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. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique. Also because of its ease of interpretation used a lot.

As an example of a problem in which regression analysis may be helpful, suppose that an industrial engineer employed by a soft drink beverage bottler is analyzing the product delivery and service operations for vending machines. He suspects that the time required by a route deliveryman to load and service a machine is related to the number of cases of product delivered. The engineer visits 25 randomly chosen retail outlets having vending machines, and the in-outlet delivery time (in minutes) and the volume of product delivered (in cases) are observed for each. The 25 observations are plotted in Figure *a*. This graph is called a **scatter diagram.** This display clearly suggests a relationship between delivery time and delivery volume; in fact, the impression is that the data points generally, but not exactly, fall along a straight line. Figure *b* illustrates this straight-line relationship.

If we let *y* represent delivery time and *x* represent delivery volume, then the equation of a straight line relating these two variables is

$$y = \beta_0 + \beta_1 x \qquad (1.1)$$

where β_0 is the intercept and β_1 is the slope. Now the data points do not fall exactly on a straight line, so Eq. (1.1) should be modified to account for this. Let the difference between the observed value of *y* and the straight line ($\beta_0 + \beta_1 x$) be an **error** ε. It is convenient to think of ε as a statistical error; that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on delivery time, measurement errors, and so forth. Thus, a more plausible model for the delivery time data is

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1.2)$$

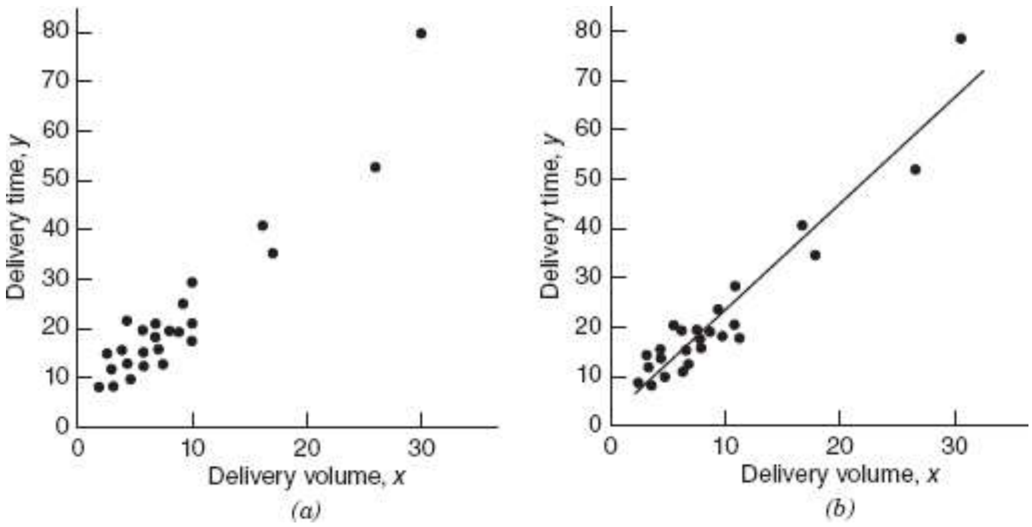


Figure (*a*) Scatter diagram for delivery volume. (*b*) Straight-line relationship between delivery time and delivery volume.

Here,

x and y are two variables on the regression line.

b1 = Slope of the line.

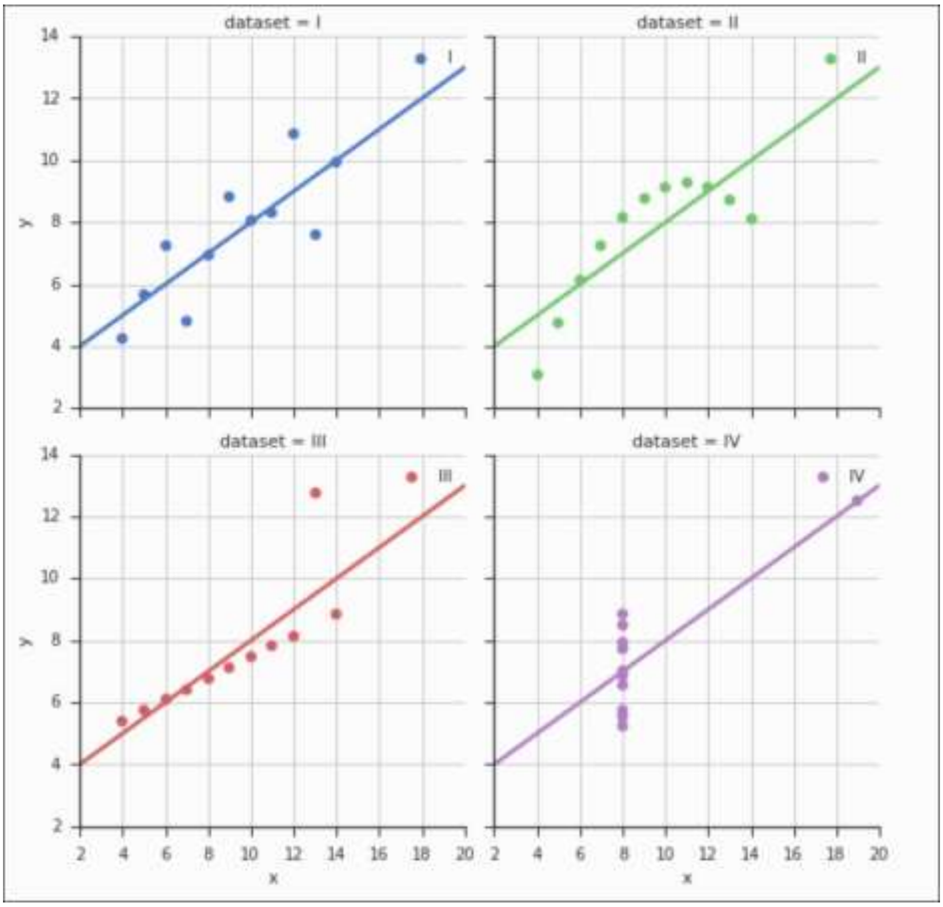
b0 = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer: **Anscombe's quartet** is a classic example that illustrates why visualizing data is important. The quartet consists of four datasets with similar statistical properties. Each dataset has a series of *x* values and dependent *y* values. 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.



The four datasets can be described as:

• Dataset 1: this fits the linear regression model pretty well.

• Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
• Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

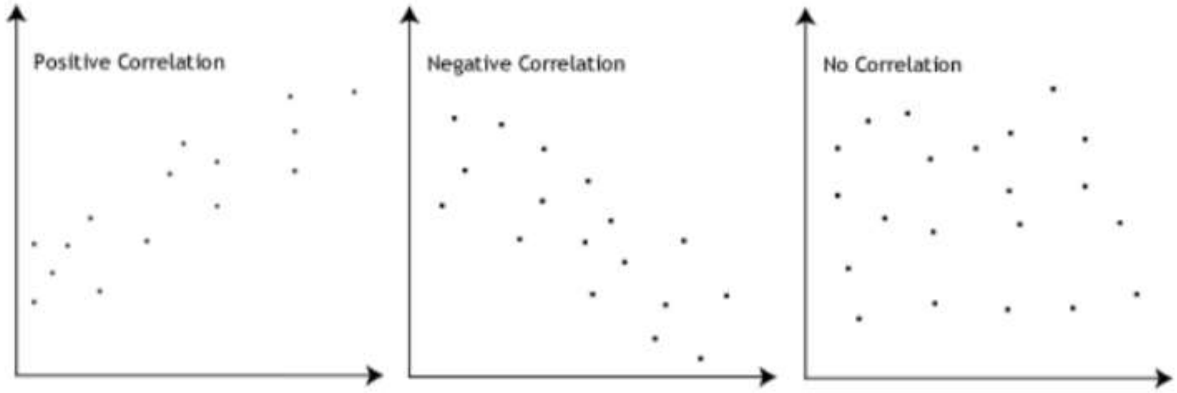
• Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)

Answer: Most common correlation method is Pearson's product moment correlation coefficient (PPMCC) , *r*. This method is appropriate for paired data that are quantitative and either interval or ratio data. If the data are only slightly skewed from normal, then Pearson's may still be close enough (although not perfect).

$$r_{xy} = \frac{\sum_i^N (x-\overline{x})\left(y-\overline{Y}\right)}{\sqrt{\sum_i^N (x-\overline{x})^2 \sum_i^N (y-\overline{y})^2}}$$

What this rather messy formula measures is the intensity or strength of the straight-line or linear relationship between the x data and the y data. In particular, the correlation coefficient *r* will be a number between -1 and 1. The stronger the correlation between the two variables, the closer to -1 or 1 the result will be. In this case, a strong correlation is one that looks almost like a straight line (also known as **linear**). Positive correlation tells us the two values increase/decrease in sync together. Negative correlation tells us the two values move in opposite directions.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider values as higher and smaller values as lower values. 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. Scale is used for ease of interpretation too.

eg. Price is 3$ and price is 5 INR. In this example machine learning algorithm will consider 5 INR as greater value which is not the case. And it will do wrong prediction. Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform. Scaling can be performed in two ways: Normalization: It scale a variable in range 0 and 1. It is given as:

MinMax Scaling:x = (x- min(x))/(max(x)-min(x))

Standardization: It transforms data to have a mean of 0 and standard deviation of 1.

Standardization:x = (x- mean(x))/sd(x)

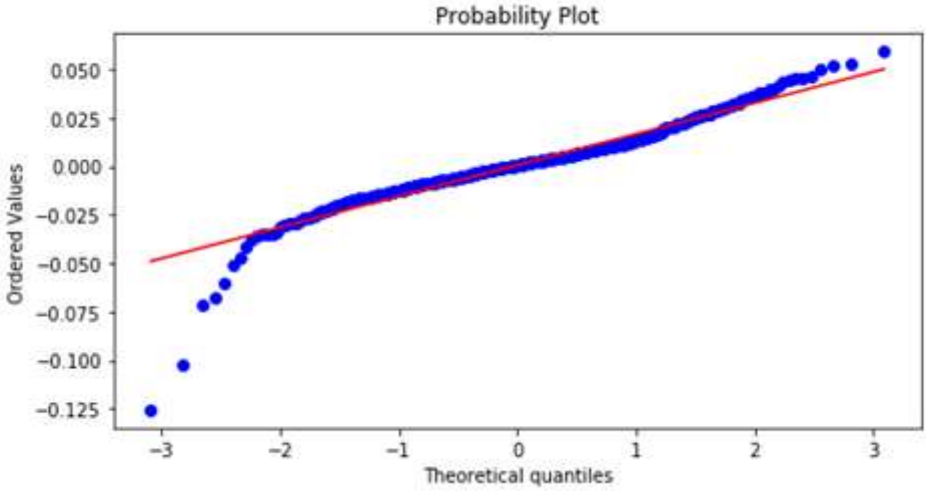
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: If there is perfect correlation, then **VIF** = **infinity**. A large **value** of **VIF** indicates that there is a correlation between the variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. 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. (3 marks)

Answer: A Q-Q (quantile-quantile) plot is a probability distribution plot, where the quantiles of two distributions are plotted against each other. If the distributions are linearly related, the points in the Q-Q plot will lie along a line. Compared to histograms, Q-Q plots help us to visualize points that lie outside the line for positive and negative skews, as well as excess kurtosis.

Let's take an example, we use the last prices of the ABN stock dataset and compute the daily percentage change for charting a Q-Q plot:



When all points fall exactly along the red line, the distribution of data implies perfect correspondences to a normal distribution. Most of our data is close to being perfectly correlated between quantiles -2 and +2. Outside this range, there begin to be differences in correlation of the distribution, with more negative skews at the tails.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.