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)**

Ans:

The relationship between category variables and a target variable are

• Bike Rentals higher in 2019 than 2018
• Bike rentals higher in partly clouded whether.
• Rentals are higher in the fall and then in the summer.
• Bike rentals are dip in spring • There are more bike rentals on Saturday, Wednesday, and Thursday.
• There are more bike rentals in months of June, July, Aug, Sep, Oct.

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

Ans: In order to avoid the dummy variable trap, which occurs when variables become highly correlated, **drop_first=True** must be used when creating dummy variables. It keeps one dummy variable out of the equation (often the reference category), prevents multicollinearity, and makes it easier to understand the coefficients in regression models.

Example:

Colour	Blue	Red	Green
Red	0	0	1
Green	0	1	0
Blue	1	0	0
Red	0	0	1
Green	0	1	0

In this example, the "Colour" variable has been converted into three dummy variables: "Blue," "Green," and "Red." However, notice that if you sum the values across all three dummy variables for each row, you get 1. This is a clear indication of multicollinearity, and it can cause issues in regression models.

To avoid the dummy variable trap, you would use **drop_first=True** when creating the dummy variables. Here's how you can do it:

Colour	Red	Green
Red	0	1
Green	1	0

Blue	0	0
Red	0	1
Green	1	0

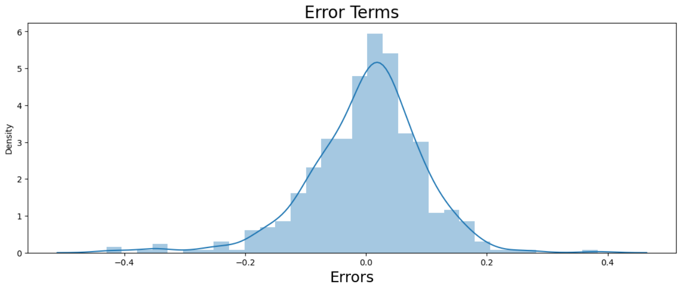
Now, by dropping the first dummy variable ("Blue"), we avoid the dummy variable trap, and the dummy variables "Green" and "Red" are sufficient to represent the original categorical variable without introducing multicollinearity.

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

Ans: ‘registered’ variable has highest correlation with the target variable with (0.93), ‘atemp’ and ‘temp’ are also having high correlation with target variable.

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

Ans: By creating a distplot of the residual and analyzing it to see whether or not it has a normal distribution and whether or not its mean is zero, we use linear regression to validate the assumptions. The diagram below demonstrates that the distribution is normal, with a mean of zero.



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

Ans: The following 3 features contributing significantly towards the demand of the shared bikes.

1. Temp : We can see that temperature variable is having the highest coefficient 0.5038, which means if the temperature increases by one unit the number of bike rentals increases by 0.5038 units.
2. winter : We can see that winter variable is having the highest coefficient 0.0829
3. Year: We can see that year variable is having coefficient 0.0232

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables only. Following steps are performed while doing linear regression:

- The dataset is divided into test and training data
- Train data is divided into features and target data sets
- A linear model is fitted using the training dataset. Internally the api"s from python uses gradient descent algorithm to find the coefficients of the best fit line. Te gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares (R2).
- In case of multiple features, the predicated variable is hyperplane instead of line.

  The predicated variable takes the following form:

  $Y = β0+β1x1+β2x2+ β3x3+……..+ βnxn$

- The predicated variable is than compared with test data and assumptions are checked.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet consists of four data sets with virtually similar simple descriptive statistics that, when shown graphically, have significantly distinct distributions. The mean, sample variance of x and y, correlation coefficient, linear regression line, and R-square value make up basic statistics. Anscombe's Quartet demonstrates how graphing several data sets with a great deal of statistical similarity can nonetheless result in very distinct outcomes.

The graphs are shown below:

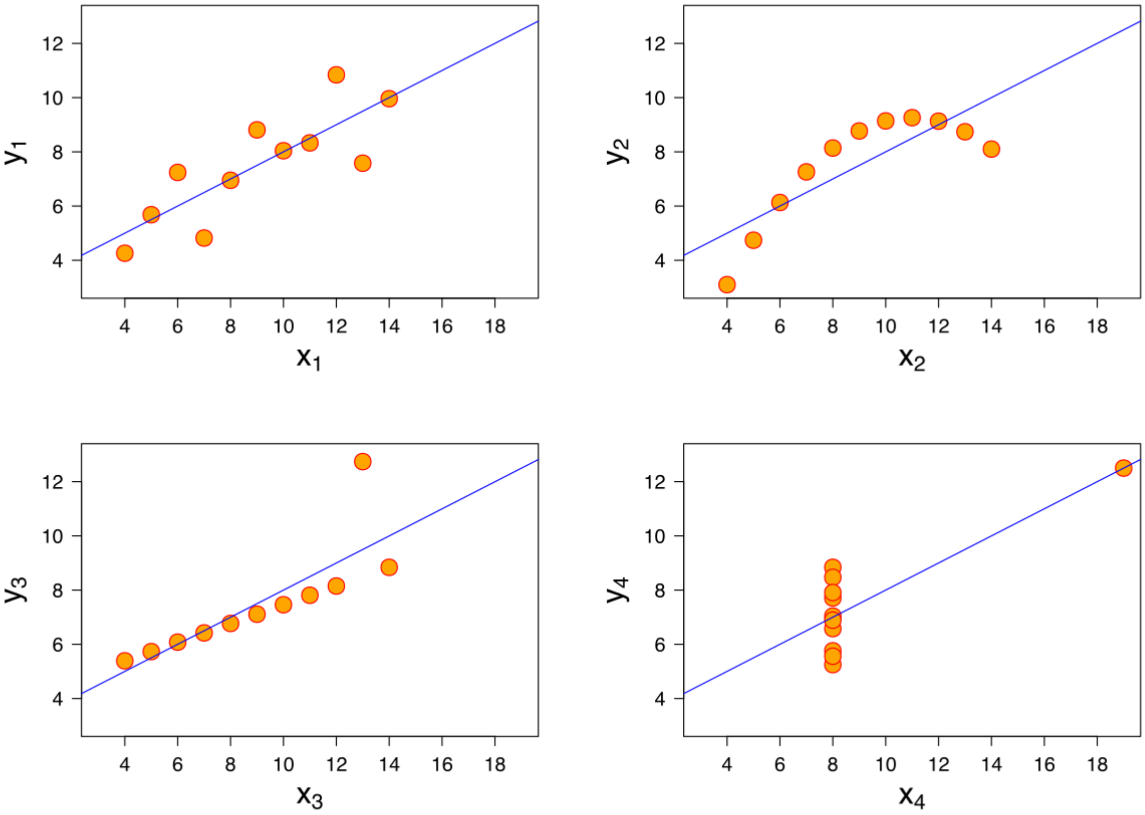


Image Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). 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?

Ans: Pearson's R measurer the strength of association of two variables. It is the covariance of two variables divided by the product of their standard deviation. It has a value from +1 to -1.

- A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase.
- A value of 0 means no correlation
- A value of -1 means a total negative correlation. It means that if one variable increase then other will decrease.

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

Ans: Scaling of a variable is performed to keep a variable in certain range. Scaling is a pre-processing step in linear regression analysis. The reason we scale a variable is to make the computation of gradient descent faster. The step size of gradient descent are generally low for accuracy, if the data has some small variables (values in the range 0-1) and some big variables (values in the range of 0-1000) than the time taken by the gradient descent algorithm will be huge.

Normalised Scaling:

- Called min max scaling, scales the variable such that the range is 0-1
- Good for non-gaussian distribution
- Value id bounded between 0-1
- Outliers are also scaled.

Standardized Scaling:

- Values are centred around mean with a unit standard deviation
- Good for gaussian distribution
- Value is not bounded
- Does not affect outliers

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

Ans: The formula for VIF is

$$VIFi = 1 / 1-Ri2$$

When there is perfect multicollinearity in regression analysis, the Variance Inflation Factor (VIF) may become infinite. When one independent variable is an exact linear combination of the others, perfect multicollinearity happens. In certain situations, the VIF formula's R square value increases to 1, resulting in a division by zero and an endless VIF. One possible solution to this problem could be to reduce the number of unnecessary variables, combine variables, increase the amount of data, or use regularization methods. Regression coefficients that are stable and dependable must take perfect multicollinearity into account.

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

(3 marks)

Ans: A graphical tool used in linear regression to determine if a model's residuals follow a normal distribution is called a Q-Q (Quantile-Quantile) plot.

It contrasts quantiles that are observed with those that are predicted by a theoretical distribution, like the normal distribution. In linear regression, Q-Q plots are useful for visually examining normality, spotting outliers, verifying model assumptions, and directing possible modifications. Plot deviations from a straight line indicate non-normality or further distributional problems. The normality assumption of residuals is verified by analysts using Q-Q plots in conjunction with statistical tests, which enhances the dependability of linear regression models.