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: Following are my observations :

- The demand of bike is less in the month of spring when compared with other seasons
- The demand for bikes is almost similar throughout the weekdays.
- The bike demand is low when the weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- We do not have any data for Heavy Rain + Ice Pellets + Thunderstorms + Mist, Snow + Fog , so we don’t have a final conclusion for these days.
- The demand for bikes increased in the year 2019 when compared with the year 2018.
- Month Jun to Sep is the period when bike demand is high.
- Bike demand is less in holidays/weekends in comparison to being in working days.

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

Answer: It is important to use drop_first = True, because If we do not use it, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: Temperature has highest positive correlation with target variable total users i.e. cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: Following are the assumptions :
* The error terms are normally distributed i.e Homoscedasticity Assumption.
* The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.
* The predicted values have a linear relationship with the actual values.
* The distribution plot of the error term shows the normal distribution with mean at Zero.

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: Based on final model top three features contributing significantly towards explaining the demand are:

- □ Temperature (0.4)
- □ weathersit : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (-0.365
- □ Spring season (-0.684)

General Subjective Questions

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

Answer: Linear regression is a technique that is commonly used to model the relationship between two variables, where one variable is considered the independent variable and the other is the dependent/target variable. This technique uses a linear equation algorithm to estimate the value of the dependent variable based on the independent variable(s).

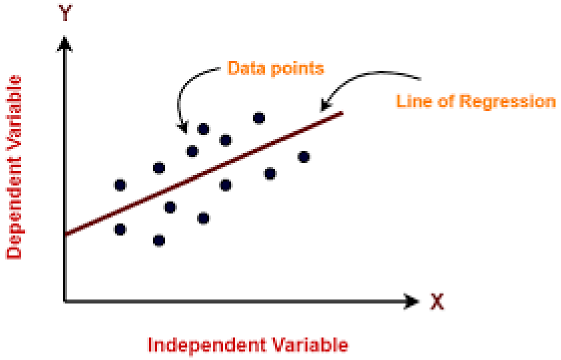
The linear regression algorithm tries to find a line of best fit that minimizes the distance between the predicted values and the actual values. The line of best fit is represented by the equation:

$y= \beta 0 + \beta 1 * x$

where y is the dependent variable, x is the independent variable, β0 is the y-intercept, and β1 is the slope of the line.

Ordinary Least Squares (OLS) regression process is the usual methodology we use to perform this regression where it first collects a set of data points for both the independent and dependent variables. These data points are plotted on a graph, with the independent variable on the x-axis and the dependent variable on the y-axis. The algorithm then calculates the values of β0 and β1 that minimize the sum of the squared differences between the predicted values and the actual values.

Once the values of β0 and β1 are calculated, the line of best fit can be plotted on the graph. This line can then be used to predict the value of the dependent variable for any given value of the independent variable.



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

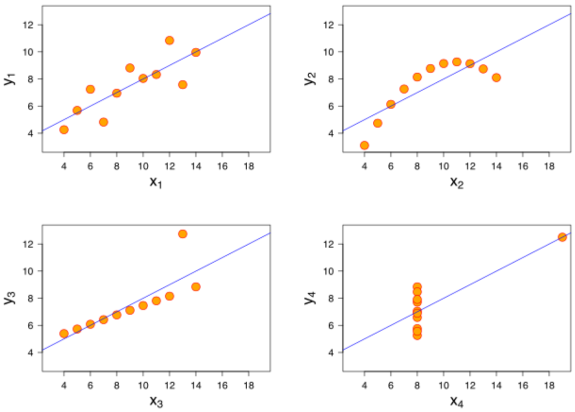
Answer: Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets have identical statistical properties, but when plotted, they have very different shapes and characteristics, illustrating the importance of data visualization in statistical analysis.

The four datasets each contain 11 (x,y) pairs of data. The summary statistics of each dataset are as follows:

1.Dataset I: This dataset has a linear relationship between x and y, with a slope of 0.5 and an intercept of 3. The correlation coefficient between x and y is 0.816, and the coefficient of determination (R-squared) is 0.67.

2.Dataset II: This dataset has a non-linear relationship between x and y, with a perfect quadratic curve. The correlation coefficient between x and y is also 0.816, and the R-squared value is 0.67.

3.Dataset III: This dataset has a linear relationship between x and y, but with one outlier that significantly affects the correlation coefficient and R-squared value. The outlier has an x value of 10 and a y value of 9.9.
4.Dataset IV: This dataset has a nearly perfect linear relationship between x and y, except for one outlier. The outlier has an x value of 10 and a y value of 9.



When the datasets are plotted, the differences become more apparent:
1.Dataset I: The data points form a linear trend that is easy to identify.
2.Dataset II: The data points form a clear quadratic curve, showing that the relationship between x and y is non-linear.
3.Dataset III: The data points appear to have a linear trend, but the outlier significantly affects the trend line and the R-squared value.
4.Dataset IV: The data points appear to have a linear trend, but the outlier affects the slope and intercept of the trend line.

3. What is Pearson’s R? (3 marks)

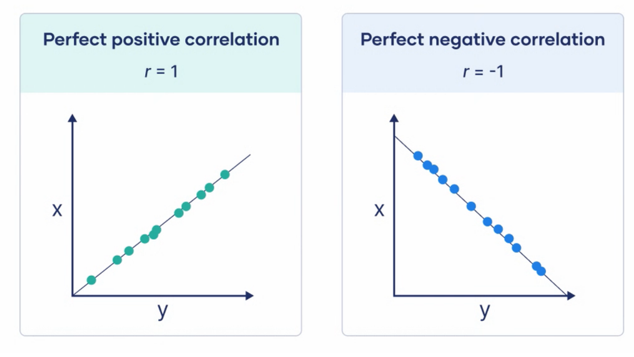
Answer: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson's r can take values between -1 and +1, where -1 indicates a perfect negative correlation (as one variable increases, the other decreases), +1 indicates a perfect positive correlation (as one variable increases, the other increases), and 0 indicates no linear correlation (there is no relationship between the variables).

The formula for calculating Pearson's r is:

$$\mathbf{r = (\Sigma(x - \overline{x})(y - \overline{y})) / (sqrt(\Sigma(x - \overline{x})\hat{}2) * sqrt(\Sigma(y - \overline{y})\hat{}2))}$$

where Σ is the sum of the indicated terms, $\overline{x}$ and $\overline{y}$ are the means of x and y, respectively, and sqrt denotes the square root.



The Pearson correlation coefficient is a good choice when all of the following are true:

- *Both variables are quantitative:* You will need to use a different method if either of the variables is qualitative.

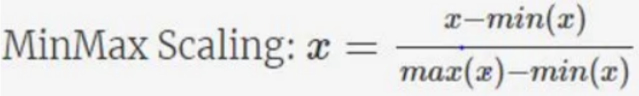
- *The variables are normally distributed:* You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- *The data have no outliers:* Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- *The relationship is linear:* "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatter plot to check whether the relationship between two variables is linear.

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

Answer: Scaling is a process of transforming a dataset to a common scale, typically to a range of values that is easier to work with or compare across different features. It is performed to bring different variables to a common scale so that they can be compared and analyzed together.

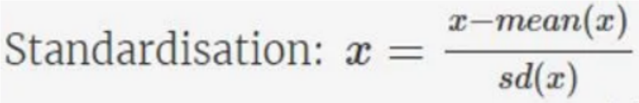
There are two types of scaling: normalized scaling and standardized scaling.

- *Normalized scaling* involves scaling the data to a range of 0 to 1, where the minimum value in the data is scaled to 0, the maximum value is scaled to 1, and all other values are scaled proportionally between these two values. Normalized scaling is useful when the absolute values of the data are not important, but only their relative values or proportions.



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

● *Standardized scaling*, also known as z-score scaling, involves scaling the data to have a mean of 0 and a standard deviation of 1. This scaling is useful when the absolute values of the data are important, and we want to compare different variables based on their deviation from the mean.



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

The main difference between normalized scaling and standardized scaling is that :

● Normalized scaling scales the data to a fixed range, while standardized scaling scales the data based on its mean and standard deviation.

● Normalized scaling preserves the shape of the data, while standardized scaling transforms the data to a standard normal distribution.

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

Answer: The Variance Inflation Factor (VIF) is a measure of the severity of multicollinearity in a regression model. A high VIF value indicates that a predictor variable is highly correlated with other predictor variables in the model, which can lead to unstable and unreliable regression coefficients.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = *Unadjusted coefficient of determination for regressing the ith independent variable on the*
remaining ones

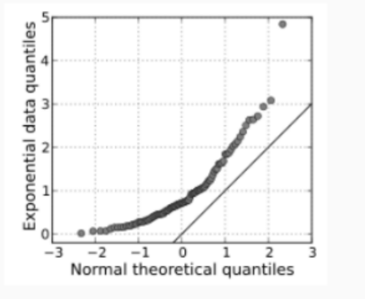
When the value of VIF is infinite, it indicates that there is a perfect linear relationship between the predictor variable of interest and at least one of the other predictor variables in the model. This means that the predictor variable is a linear combination of the other predictor variables, and can be expressed as a weighted sum of those variables.

An infinite VIF indicates that the predictor variable of interest is redundant and provides no additional information to the model beyond what is already provided by the other predictor variables. In such cases, it may be necessary to remove the redundant variable from the model to improve the stability and reliability of the regression coefficients.

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 plot (Quantile-Quantile plot) is a graphical method for comparing the distribution of a sample of data to a theoretical distribution, such as the normal distribution. The Q-Q plot compares the quantiles of the sample distribution with the quantiles of the theoretical distribution, and if the sample distribution is close to the theoretical distribution, the plot will show points close to a straight line.

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



Use of Q-Q plot:

1. In linear regression, a Q-Q plot is used to check the normality assumption of the errors, also known as residuals, which are the differences between the observed values and the predicted values.

2. The Q-Q plot can help to visually assess whether the residuals follow a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will be approximately along a straight line, while deviations from the straight line may indicate non-normality.

Importance of Q-Q plot:

The importance of a Q-Q plot in linear regression is that it helps to assess the validity of the normality assumption. A violation of the normality assumption can lead to biased and unreliable estimates of the regression coefficients and their standard errors, The Q-Q plot provides a simple and intuitive way to check the normality assumption and identify any potential violations, allowing for appropriate remedial actions to be taken, such as transformations or model modifications.