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?

   The season summer and winter having positive coefficients, these are influencing factor on the increase of bike demand (Dependent variable).

   Similarly, the weather categorical variables cloudy and Light rain and season spring having negative coefficients which are having impact on bike demand negatively. (reduces the value of dependent variable).

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

   Drop_first = true , is used to get the k-1 level of dummy values. This ensures that one category from each column is dropped to avoid multicollinearity of variables.

   Example,

   dummy_var = pd.get_dummies(bike_sharing['season'], drop_first=True) this creates only 3 dummy variables like below

Season_spring	Season_summer	Season_winter
1	0	0
0	1	0
0	0	1

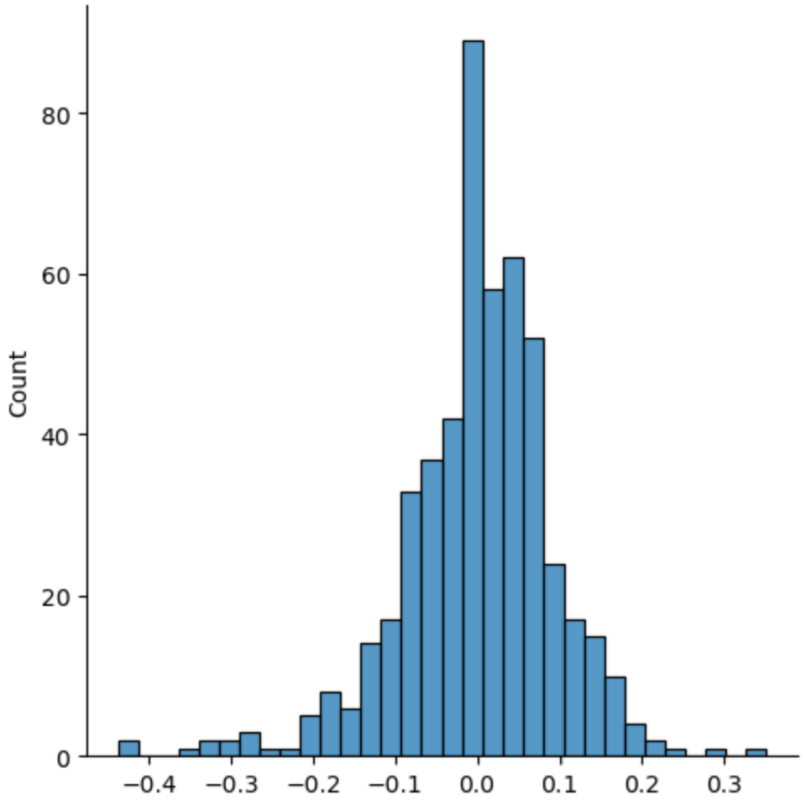
   The fourth season fall is represented as 0 0 0 values. Here drop_first = True removes the first column Season_fall to avoid the multicollinearity.

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

   a. **temp** is having the highest correlation with target variable **cnt.**

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

   Error terms are normally distributed with mean equal to zero. This is the assumption normally made in the linear regression.



Here Error term mean is zero. following normal distribution.

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

Top three contributors.

1. temp – having + coefficient of 0.4657
2. yr - having + coefficient of 0.2355
3. season_winter, summer – total having positive coefficient of 0.1156

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (also known as the response or target variable) and one or more independent variables(also known as predictors or features). This goal is to find the best-fitting straight line through the data points that minimizes the difference(error) between the observed values and the values predicted by the model.

Key concepts of Linear Regression

1. **Simple Linear Regression**
   a. Simple linear regression involves a single independent variable and is represented by the equation $y = \beta 0 + \beta 1x + \epsilon$

      y = Dependent variable
      x = Independent variable
      $\beta 0$ – intercept of regression line
      $\beta 1$ – slope of the regression line.
      ϵ – Error term (difference between the observed and predicted values.
2. **Multiple linear Regression**

   Multiple linear regression involves multiple independent variables and is represented by the equation

   $$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$$

      y = Dependent variable
      x1, x2, x3..xn are Independent variable
      $\beta 0$ – intercept of regression line
      $\beta 1$, $\beta 2$, $\beta 3$.. βn – Coefficients of the independent variables..
      ϵ – Error term (difference between the observed and predicted values.

Steps in Linear Regression

1. **Data collection**

   Gather the data that includes both dependent variable and independent variables.

2. **Data preprocessing**

   Prepare the data for analysis. This may include

   - Handling missing values
   - Encoding categorical variables
   - Normalizing or standardizing features
   - Splitting the data into training and testing sets.

3. **Model fitting**

   Estimate the coefficients β using the training data. The most common method for estimating the coefficients is OLS(Ordinary Least Square)

4. **Ordinary Least Squares (OLS) Method**

OLS aims to minimize the sum of the squared differences between the observed values and the predicted values. Mathematically, this involves finding β that minimizes:

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

where $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$.

The solution is found using:

$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

where $\mathbf{X}$ is the matrix of input features and $\mathbf{y}$ is the vector of observed values.

5. **Model Evaluation**

Evaluate the model using the testing set and various metrics such as:

- **R-squared (R^2)**: Proportion of variance in the dependent variable that is predictable from the independent variables.
- **Mean Absolute Error (MAE)**: Average absolute difference between observed and predicted values.
- **Mean Squared Error (MSE)**: Average squared difference between observed and predicted values.
- **Root Mean Squared Error (RMSE)**: Square root of the MSE.

Assumptions of Linear Regression

1. **Linearity**: The relationship between the dependent and independent variables should be linear.
2. **Independence**: Observations should be independent of each other.
3. **Homoscedasticity**: Constant variance of the errors (residuals) across all levels of the independent variables.
4. **Normality**: The residuals should be approximately normally distributed.
5. **No Multicollinearity**: Independent variables should not be highly correlated with each other.

2. Explain the Anscombe's quartet in detail.

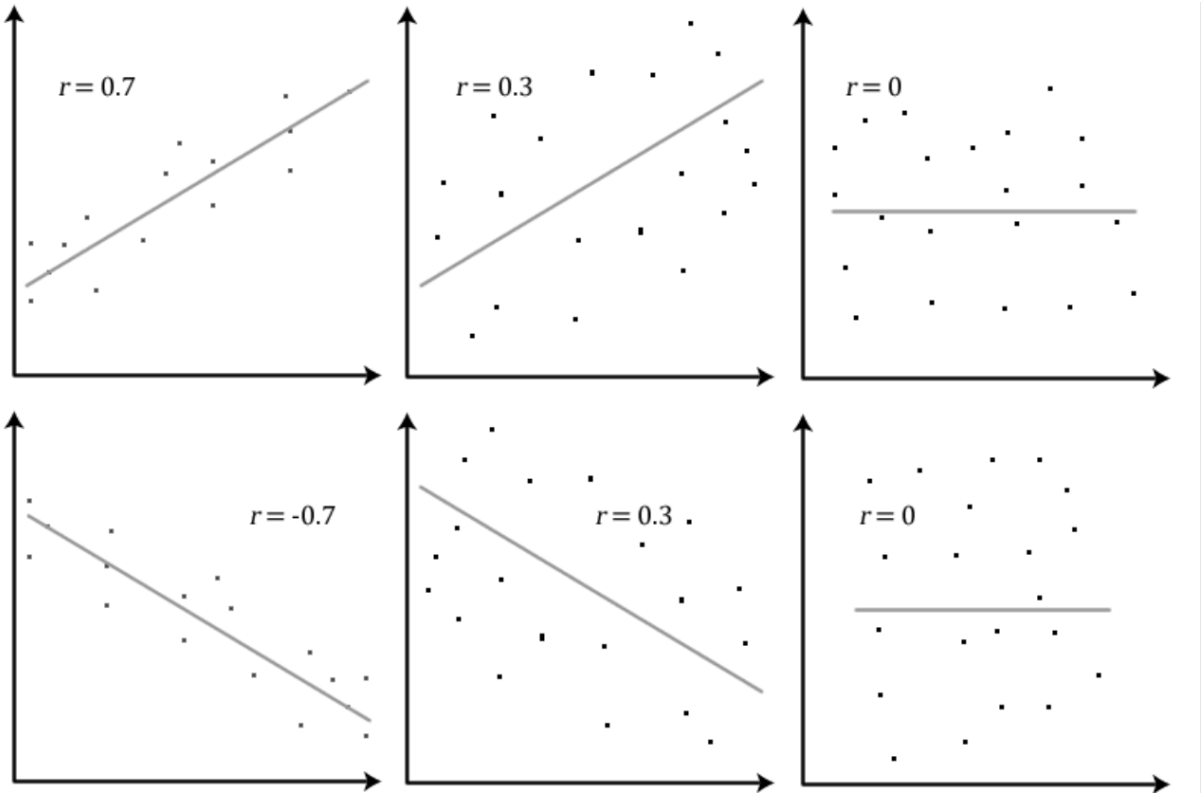
3. What is Pearson's R?

The value of the correlation coefficient, denoted as rr, will always be between -1 and 1, inclusive. This coefficient measures the strength and direction of the linear relationship between two variables.

- r=1 r=1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
- r=−1 r=−1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- r=0 r=0 indicates no linear relationship between the variables.

Values of rr between 0 and 1 indicate varying degrees of positive correlation, while values between 0 and -1 indicate varying degrees of negative correlation. The closer rr is to 1 or -1, the stronger the linear relationship between the variables.

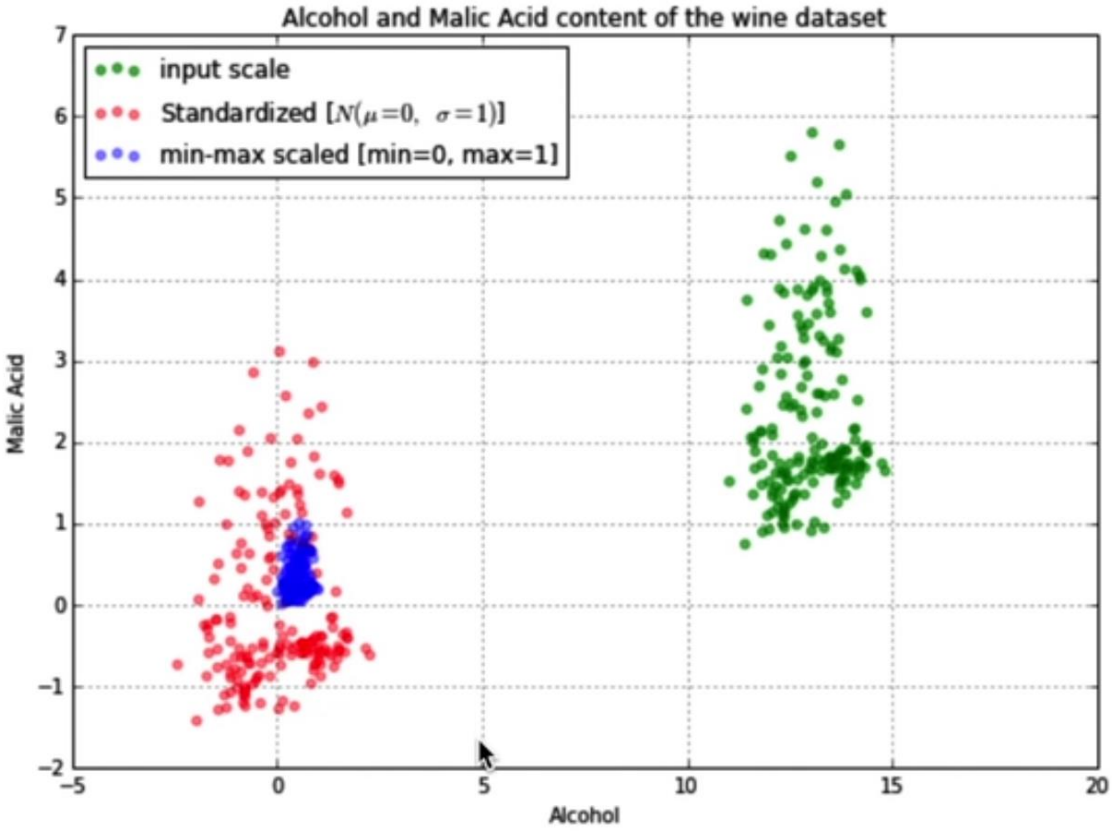
The below picture explains the correlation coefficient (R)



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

1. Ease of interpretation.
   If variables are having very different scales it is very difficult to keep in mind they are in different ranges, therefore the coefficients does not signifies importance of them.
   If we keep all the variables in in scale, it is easy to compare one variable to another and interpretation is very easy .
2. Faster convergence for gradient descent methods.

   If the features are in different ranges say one feature in -1 to 1 and another in 1000 to 2000 it takes lot of time for convergence.
   If you get them about to same ranges (not necessarily to exactly same ranges) say -1 to 1 or 0 to 1 or -3 to 3 i.e., in close ranges the convergence becomes much faster.



- The input data points in green varies between 0.8 to 6
- After normalisation, all the data points lies between 0 to 1. no data points lies outside 0 and 1. The standardisation points spread little more than the normalized version. But the mean is 0. and the standard deviation is 1.

Now the question Is which one to use? Standardization or Normalization?

One advice is to use Normalization. Because it takes care of outliers very well. Think of if any input point is above 8 or 9 in the pic, it would have mapped to 1. other data points would have mapped between 0 and 1.

In case of standardization, it would map the outlier somewhere around 4. So, the values are spread across.

Difference: it maps the any input values between 0 to 1. Whereas standardization does not map all input values between 0 to 1.

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

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression

model are highly correlated. High multicollinearity can make it difficult to estimate the relationship between each independent variable and the dependent variable accurately.

Understanding VIF

VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other independent variables.

VIF for an independent variable Xi is caluclated as

$$VIF_i = 1/1 - R_{i2}$$

where $Ri2$ is the coefficient of determination of the regression of Xi on all the other independent variables.

Why VIF Becomes Infinite

VIF becomes infinite when there is perfect multicollinearity. This happens when an independent variable is a perfect linear combination of one or more other independent variables in the model.

Mathematically, if $Ri2=1$, then:

$$VIFi = 1/1 - 1 = \infty$$

Cause of Perfect multicollinearity

1. Duplicate Variables.(if two columns that are exactly the same)
2. When using dummy variable if you don't drop one category you will have perfect multicollinearity because the dummy variables add up to the original categorical variables.

How to address infinite VIF,

1. Remove perfectly collinear variables.
2. Principal component analysis.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)