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.There are several categorical variables such as 'mnth','weekday','season','weathersit'

   a. months has a significant impact on target variable. January and February has negative correlation with target variable (cnt).
   b. weekday doesn't have much of the impact on cnt as corelation for most of them is zero except monday which has correlation of -0.1
   c. season show significant impact on target variable. There are 3 seasons namely Spring, Summer, and Winter is present in the data. Spring is strongly negatively related (-0.6) with the target variable.
   d. Weathersit has 4 categories, Out of which light snow(-0.2) and mist(-0.2) are both negatively and positively related with the target variable.

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

   Ans. It is crucial to employ drop_first=True when generating dummy variables using the get_dummies function in pandas. This is because the function creates dummy variables equivalent to the number of categories' level present, whereas we often prefer to represent the categorical levels as one less than the total number of levels. By setting drop_first=True, pandas automatically excludes one level, aligning with this convention.

   When Drop_first not given

   Below table represents weather category there are 4 levels without drop_first, get dummy will return

Spring	Summer	Autumn	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

   when drop_first=True, here we can represent spring via three zeros

   Three zeros represent spring

Summer	Autumn	Winter
0	0	0
1	0	0
0	1	0
0	0	1

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

   Ans. Looking at the pair plot of numerical variable. Feature "**Registered**" has the highest correlation with pair plot (**cnt**) target variable, but target variable(cnt) is direct sum of features registered and casual. This will result in high multicollinear issues as CNT is completely explained by casual and registered.

   <u>If we ignore registered and casual, highest correlation is with feature</u> **temp**

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

   Ans.

   To validate model after building the model with training set, below steps were followed

   a. There should be **linear relationship between dependent and independent variable**
   b. **Error terms needs to be normally distributed** plotted error terms (y_train_atual-y_train_pred). This can be verified via QQ plot or scatter plot
   c. Error terms should have constant variance (**homoscedasticity)**, can be verified by plotting residuals against predicted values.
   d. There should not be any visible patterns in the error terms, scatter plots can be used for the same.

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

   Ans. Based on the regression model below are the top 3 features which are significant contributors explaining demand for shared bikes.

   Features with low p-Value are highly significant contributors

   1. **Year** (yr):

      Coefficient: 0.2348, **p-value**: 0.000 (highly significant)

      This shows that over the years from 2018 to 2019 bike demand is increasing, since coefficient for year is positive

   2. **Light Snow** (light snow):

      Note Complete Value for Weathersit Light Snow is ""Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds". Light Snow is just abbreviated form

      Coefficient: -0.2877**, p-value**: 0.000 (highly significant)

      The weather situation is also a significant contributor. The negative coefficient indicates that, during light snow conditions, the demand for shared bikes decreases. This is reasonable, as inclement weather can discourage bike usage.

   3. **Temperature** (temp):

      Coefficient: 0.4335, **p-value**: 0.000 (highly significant)

      The positive coefficient implies that as the temperature increases, the demand for shared bikes also increases. Warmer weather is often associated with higher bike usage. This is in sync with weather situation as temp decreases with light snow bike demand decreases

General Subjective Questions

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

Ans. Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find the linear equation that best predicts the dependent variable based on the values of the independent variables. Here's a detailed explanation of the linear regression algorithm:

Simple Linear Regression:

In simple linear regression, there is only one independent variable. The linear relationship between the dependent variable (\(Y\)) and the independent variable (\(X\)) is represented by the equation:

$$Y = \beta_o + \beta_1 X + e$$

a. Y is the dependent variable,
b. X is the independent variable
c. β_o is the y-intercept (constant term)
d. β_1 is the slope of the line (regression coefficient)
e. **e** is the error term representing the unobserved factors affecting Y.

The objective is to find the values of β_o and β_1 that minimize the sum of squared differences between the predicted and actual values of Y i.e error terms

RSS: residue sum of Squares

i -> n

$$RSS = \Sigma (Y_{i\ (pred)} - Y_{i\ (actual)})^2$$

Multiple Linear Regression:

In multiple linear regression, there are multiple independent variables. The linear relationship is represented by the equation: Below equation assumes that per unit change in any independent variable will result in change in dependent variable when other independent variables are kept constant

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots\ldots\ldots + \beta_n X_n + e$$

a. Y is the dependent variable.
b. $X_1, X_2, X_3, X_4 \ldots X_n$ are the independent variables,
c. β_o is the y-intercept (constant term)
d. $\beta_1, \beta_2, \beta_3, \beta_4 \ldots \beta_n$ are the regression coefficients.
e. **e** is the error term representing the unobserved factors affecting Y

Again, the goal is to find the values of the coefficients that minimize the sum of squared differences as explained for simple linear regression

There are certain assumption we need to keep when working with linear regression

1. X independent and Y dependent variables are **linearly related**
2. **Error terms are normally distributed** and **mean of error term is centered around zero**
3. Error terms are independent of each other. Change in one error term does not affect other error terms
4. Error terms have zero variance i.e **homoscedasticity**, variance should not increase or decrease as error values Change.

Below are the steps involved in doing linear regression analysis

1. **Identify** dependent and independent variables
2. **Identifying linear relationship** between dependent and independent variables. This can be done by plotting scatter plot between dependent and independent variables.
3. **Calculate Pearson coefficient** and plot on heatmap to visualize strength of relationship between variables.
4. **Identify Categorical variable and create dummy variables** . Number of dummy variables will be n-1 where n is number of levels
5. **Scale numerical features** (column) excluding dummy variables. Scaling can be of two type normalized(Min max) or standard (Z-Score)
6. **Split the data into training and test data**
7. **Feature Selection:**. This can be done manually or using **RFE (Recursive Feature Elimination)**
8. **Build model** using features identified in step7 on training data
9. **Evaluate Model** significance of model by making sure that P-value of all identified feature is less than .05 and R-sq is as close to 1.

10. **Evaluate Multicollinearity** by calculating VIF, VIF need to be lower than 5. Drop columns with VIF greater than 5

    $VIF = 1 / (1 - R^{\wedge}_2)$
11. **Prediction** predict target variable using model created on training data.
12. **Error terms** Y(actual) - Y(pred) , Error terms need to be normally distributed and their mean needs to be centered around zero
13. **Test Data Evaluation:** Now Finally work on test data. Scale test data using scaler that was used to scale training data.
14. **Test Data Prediction:** predict using test data and calculated R-sqaured value of test and training data. Both of them should be similar.

    <u>If R2_score drops drastically from training to test data then it mean regression model has overfitting line.</u>

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

   Ans.

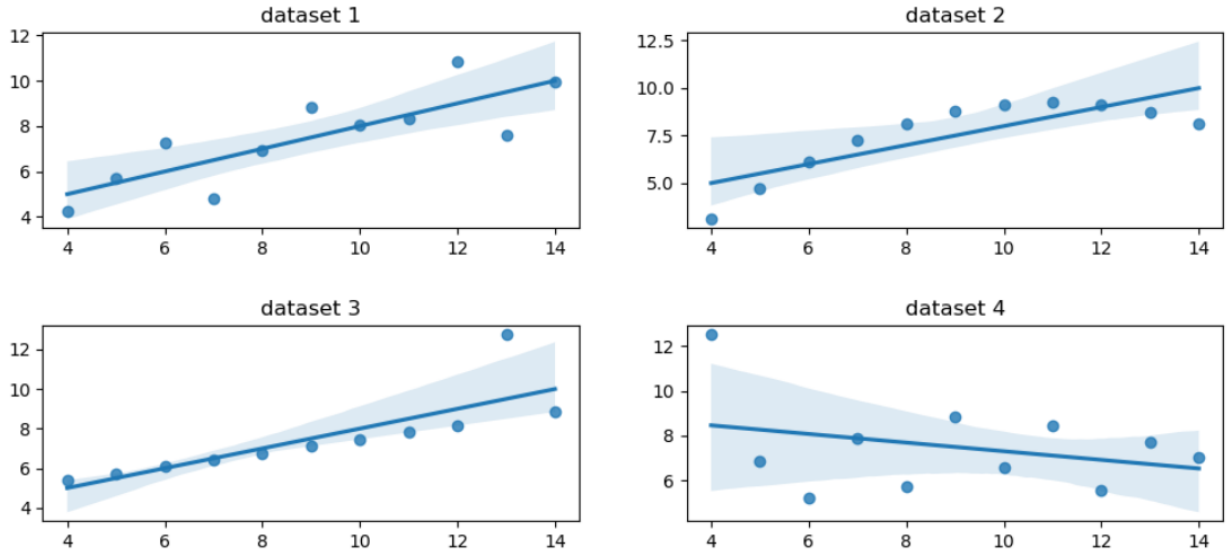
   The Anscombe quartet comprises four datasets with nearly identical statistical measures, including mean, variance, and correlation. However, their graphical representations exhibit substantial differences.

   Anscombe's insight underscores the necessity of not solely depending on statistical data. It serves as an illustration of the vital role of examining data visually, beyond relying solely on basic statistical properties. Relying exclusively on statistical descriptions may lead to misleading conclusions, as summary statistics might fail to unveil hidden patterns or relationships within the data.

   This concept resonates with the linear regression process, where plotting all independent variables against the dependent variable is crucial for a comprehensive understanding.

   <u>As it can be seen in below screenshot four dataset has same mean, variance and standard deviation, but graphical representation shows different pattern for each</u>

```
y 1: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
  Mean: 7.501  Variance: 3.752 SD: 1.937
y 2: [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]
  Mean: 7.501  Variance: 3.752 SD: 1.937
y 3: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
  Mean: 7.5  Variance: 3.748 SD: 1.936
y 4: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]
  Mean: 7.501  Variance: 3.748 SD: 1.936
```



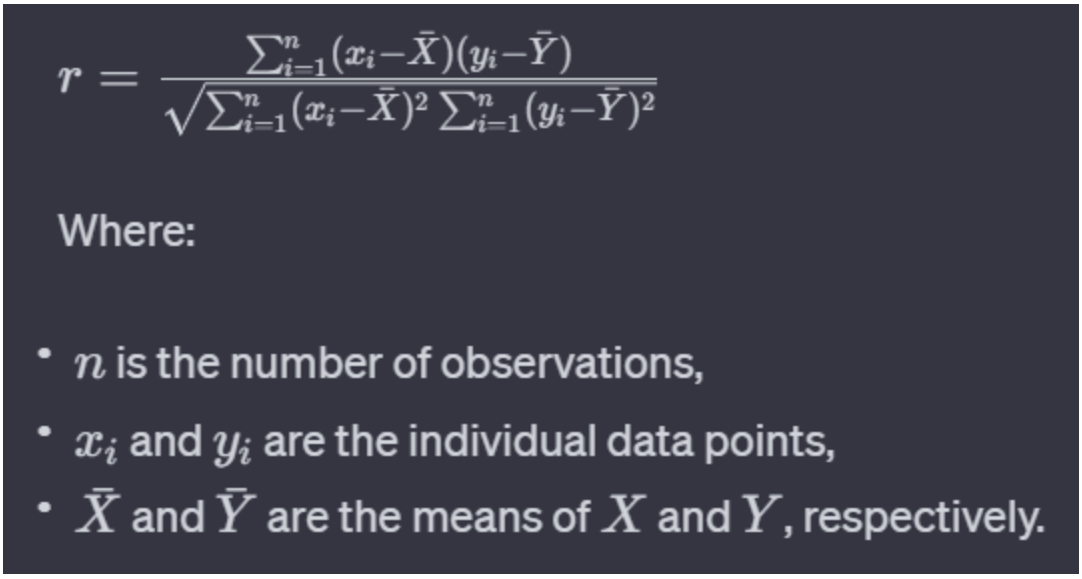
3. What is Pearson's R? (3 marks)

Ans.

Pearson's correlation coefficient, often represented by 'r' . It quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

1. (**r = 1**) indicates a perfect positive linear relationship,
2. (**r = -1**) indicates a perfect negative linear relationship
3. (**r = 0**) indicates no linear relationship between the variables.

The formula for Pearson's correlation coefficient between variables X and Y with observations x_i and y_i is given by:



Pearson's correlation coefficient is particularly useful for assessing the linear relationship between two variables, but it assumes that the relationship is linear and may be sensitive to outliers.

If the relationship between variables is not linear, or if there are influential outliers, other correlation measures or methods may be more appropriate.

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

   Ans.

   Scaling is a preprocessing technique used to adjust the range of values or features in a dataset. The primary goal is to bring all features to a similar scale.

   Scaling is crucial in machine learning, as algorithm doesn't have to deal with features that have widely varying magnitudes, making algorithm perform better.

   There are different methods of scaling, and two common approaches are:

   Normalized Scaling (Min-Max Scaling): This method adjusts the range of values between 0 and 1 using the formula:

Normalized Value = $(X - X_{min})/(X_{max} - X_{min})$

After normalization, the minimum value becomes 0, and the maximum value becomes 1.

Standardized Scaling (Z-score normalization): This method scales values to have a mean of zero and a standard deviation of 1 using the formula:

Standardized Value = $(X - X_{mean})/$Standard Deviation.

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

Ans. The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the independent variables in a regression model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, meaning that one can be predicted exactly using a linear combination of the others.

Formula for VIF

$$VIF = 1 / (1 - R^{\wedge}_2)$$

where (R^2) is the coefficient of determination from the regression of a particular independent variable against the other independent variables in the model.

If R^2 = 1, it implies that all the variation in that independent variable can be explained by the other independent variables, leading to a perfect linear relationship. When R^2 = 1 , the denominator in the VIF formula becomes zero, resulting in an undefined or infinite VIF.

In practical terms, an infinite VIF indicates that the variance of the estimated regression coefficients for that variable is inflated due to the perfect linear relationship with other variables. In such cases, it becomes impossible to obtain reliable estimates for the coefficients, and it can lead to issues in the interpretation and stability of the regression model.

To address this, it is crucial to identify and handle multicollinearity issues, perhaps by removing one of the perfectly correlated variables or using other techniques such as regularization.

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

Ans. A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess the similarity between the distribution of a dataset and a theoretical distribution, often assumed to be normal in the context of linear regression.

In linear regression, it is crucial to examine if the error terms conform to a normal distribution. To construct a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the theoretical distribution (usually a normal distribution) after sorting the data.

If the error terms truly follow a normal distribution, the points on the Q-Q plot will align along a straight line, indicating a close match between the empirical distribution and the assumed theoretical distribution. Departures from a straight line suggest deviations from normality.

Below screenshot shows when normally distributed error terms plotted in QQ plot they fall on a straight line. Right graph shows QQ plot of error terms from a linear regression model. Left graph shows same error terms are normally distributed.