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)**

Categorical variables in your dataset can have varying effects on the demand for shared bikes (cnt). Based on the final model, here's a potential inference:

- **Season (summer, winter)**: Positive coefficients suggest that bike demand increases in summer and winter compared to the reference season.
- **Weather Situation (bad, moderate)**: Negative coefficients indicate a decrease in bike demand under bad and moderate weather conditions.
- **Holiday**: A negative coefficient implies that bike demand is lower on holidays compared to non-holiday periods.
- **Working Day**: A positive coefficient indicates increased bike demand on working days compared to non-working days.
- **Month (September, October)**: Positive coefficients suggest higher demand in these months compared to the reference month.
- **Weekday (Saturday)**: A positive coefficient suggests increased bike demand on Saturdays.

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

When creating dummy variables for categorical features, using drop_first=True with k levels, it creates k-1 dummy variables with which we can predict the left out variable

- **Avoids Multicollinearity**: By dropping the first category, you prevent multicollinearity (the problem of having perfectly correlated features), which occurs if all categories are included.
- **Example**: Let’s say we have blood groups A+, O and B+ in the DB, if we create 2 dummies for A and B, the left out one will be the O combination and it doesn’t need an explicit mention.

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

Temperature ‘tmp’ and ‘atmp’ has the highest correlation with the target ‘cnt’ with almost greater than 0.6 correlation. (tmp and atmp are almost having perfect corrletaion with ~99%)

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

- After the model has been finalized on training dataset, the same will be tested against the test dataset to ensure if the model is properly fit or not. This can be decided looking at the below factors. If below values are rightly fit with a small change means the model is considered overall good.
  - R-squared
  - Adjusted R-squared

- **Linearity**: Plot residuals vs. fitted values. Residuals should be randomly scattered around zero, indicating that the model fits the data well and that the relationship between predictors and the response variable is linear.

- **Normality of Residuals**: Use a Q-Q plot to check if residuals are normally distributed. If the residuals follow a straight line on the Q-Q plot, the normality assumption is satisfied.

- **No Multicollinearity**: While not mentioned specifically, this assumption is important and can be checked using VIF (Variance Inflation Factor) scores, not R^2 or adjusted R^2.

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

Temperature and year are the top two **positively** correlated features and **weathersit_bad** is having great **negative** correlation which explains the demand for bikes, either by driving it up or down based on their coefficients.

- **temp (0.5610**): The highest positive coefficient indicates that temperature has a strong positive impact on bike demand. As temperature increases, bike demand rises, likely due to favorable weather conditions for biking.

- **yr (0.2282**): This coefficient shows that bike demand has increased over the years, possibly reflecting an overall growth in bike-sharing popularity or an increase in urban cycling infrastructure.

- **weathersit_bad (-0.2173):** The significant negative coefficient indicates that bad weather conditions drastically reduce bike demand, highlighting the sensitivity of bike-sharing usage to adverse weather.

General Subjective Questions

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

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. The core idea is to find the best-fitting linear equation (line) that predicts the dependent variable based on the independent variables. The linear relationship is modeled by the equation

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon, \text{ where:}$$

- Y is the dependent variable (**target**).
- β0 is the ***y-intercept*** (constant term).
- X1,X2,…, Xn are ***independent*** variables.
- β1,β2,…,βn are the ***coefficients*** (***slopes***) of the independent variables X1,X2,…, Xn.
- ϵ is the ***error term***, representing the difference between observed and predicted values, also known as ***residuals***.

Objective:

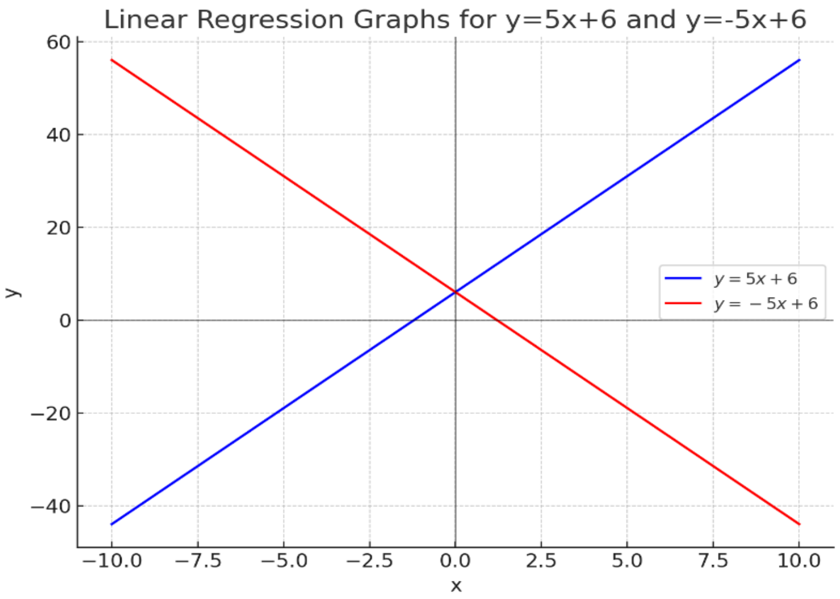
- The objective is to minimize the sum of squared residuals (differences between observed and predicted values). This method is known as ***Ordinary Least Squares (OLS)***.

Types of Linear Regression:

- ***Simple Linear Regression :*** Uses only one independent variable.
- ***Multiple Linear Regression:*** Uses two or more variables.

Types of Linear Relationships:

- ***Positive Linear Relationship:***
  - When one variable increases, the other increases proportionally.
  - The slope of the line will be upwards.
- ***Multiple Linear Relationship:***
  - When one variable increases, the other decreases proportionally.
  - The slope of the line will be downwards.



Steps involved in a Linear Regression :

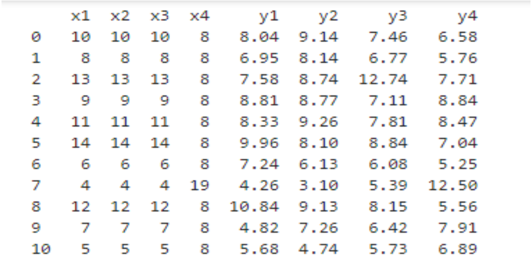
1. **Fit the Model**: Estimate the coefficients β0, β1, … , βn by minimizing the sum of squared residuals.
2. **Interpretation**: The coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables remain constant.
3. **Model Evaluation**: Common metrics include
    - *R-squared (R^2):* Measures the proportion of variance in the dependent variable explained by the model.
    - *Adjusted R^2*: Adjusted for the number of predictors, it penalizes adding unnecessary variables.
    - *p-values:* Indicate the significance of each coefficient in the model.
    - *F-statistic:* Tests the overall significance of the model.

Assumptions:

- **Linearity**: The relationship between the dependent and independent variables is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: Constant variance of residuals.
- **Normality**: The residuals are normally distributed.

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

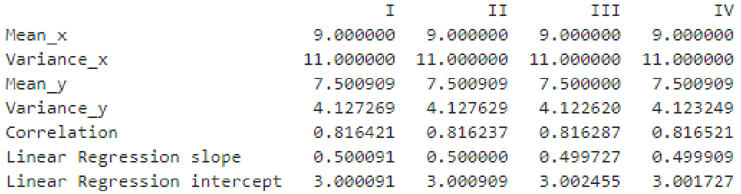
Anscombe's quartet is a collection of ***four datasets*** that have *nearly identical simple statistical properties*, such as mean, variance, and correlation, yet are vastly different when visualized.
This highlights the importance of data visualization alongside statistical analysis.



```
     x1  x2  x3  x4    y1    y2     y3     y4
0    10  10  10   8  8.04  9.14   7.46   6.58
1     8   8   8   8  6.95  8.14   6.77   5.76
2    13  13  13   8  7.58  8.74  12.74   7.71
3     9   9   9   8  8.81  8.77   7.11   8.84
4    11  11  11   8  8.33  9.26   7.81   8.47
5    14  14  14   8  9.96  8.10   8.84   7.04
6     6   6   6   8  7.24  6.13   6.08   5.25
7     4   4   4  19  4.26  3.10   5.39  12.50
8    12  12  12   8 10.84  9.13   8.15   5.56
9     7   7   7   8  4.82  7.26   6.42   7.91
10    5   5   5   8  5.68  4.74   5.73   6.89
```

Statistical Similarity: All four datasets have the same:

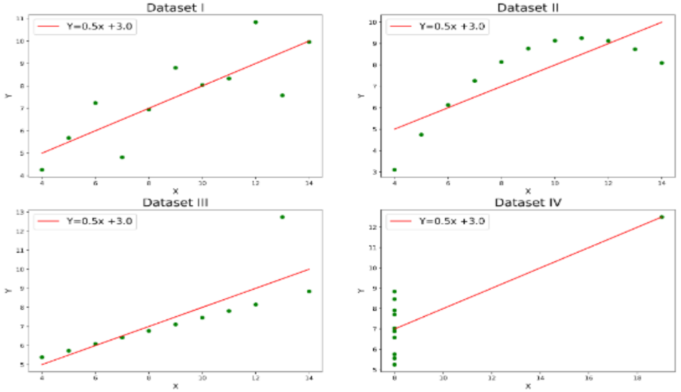
- Mean of X and Y.
- Variance of X and Y.
- Correlation coefficient between X and Y.
- Linear regression line equation.
- Coefficient of determination (R²).



```
                                    I          II         III         IV
Mean_x                       9.000000    9.000000    9.000000    9.000000
Variance_x                  11.000000   11.000000   11.000000   11.000000
Mean_y                       7.500909    7.500909    7.500000    7.500909
Variance_y                   4.127269    4.127629    4.122620    4.123249
Correlation                  0.816421    0.816237    0.816287    0.816521
Linear Regression slope      0.500091    0.500000    0.499727    0.499909
Linear Regression intercept  3.000091    3.000909    3.002455    3.001727
```

Graphical Differences:

- **Dataset 1**: A simple linear relationship with normally distributed data.
- **Dataset 2**: A quadratic relationship, showing a curve rather than a straight line.
- **Dataset 3**: A linear relationship with an outlier, which influences the regression line significantly.
- **Dataset 4**: A vertical line with one outlier that heavily influences the correlation.



Anscombe's quartet Plot

Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a ***linear relationship*** between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a ***non-linear relationship*** between x and y.
- In the third one(bottom left) you can say when there is a ***perfect linear relationship*** for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a ***high correlation coefficient***.

Importance:

- **Visualization**: Anscombe's quartet demonstrates that identical statistical properties can mask vastly different data distributions. Visualizing data can reveal patterns, relationships, or anomalies that simple statistics may miss.

- **Outliers and Leverage Points**: It underscores the effect of outliers and leverage points on statistical analysis and model fitting.

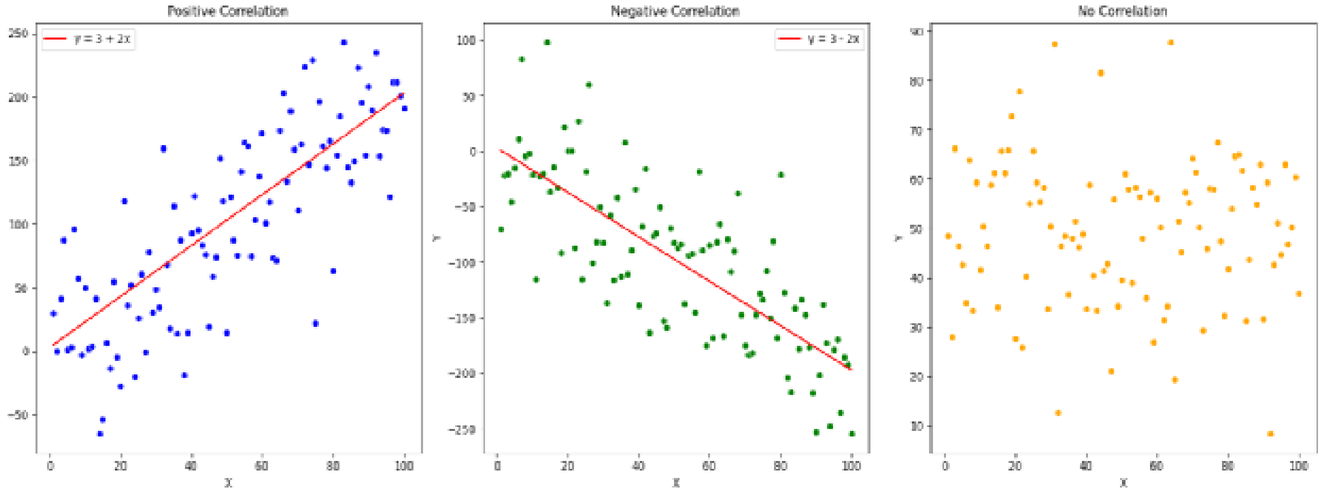
- **Caution in Analysis**: It serves as a cautionary tale against relying solely on summary statistics without examining the data visually

3. What is Pearson’s R? (3 marks)

Pearson’s correlation coefficient (Pearson’s R), is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between them and ranges between -1 and 1

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$ Where:

Xi and Yi are individual data points.
X¯ and Y¯ are the means of X and Y, respectively.



Interpretation (*Sign*: Indicates the direction of the relationship)**:**

- ***r = +1*** : Perfect positive linear relationship.
- ***r = −1*** : Perfect negative linear relationship.
- ***r = 0*** : No linear relationship between the variables.
- **|r| = 0 to 0.3**: Weak correlation.
- **|r| = 0.3 to 0.7**: Moderate correlation.
- **|r| = 0.7 to 1**: Strong correlation.

Assumptions:

- **Linearity**: Pearson’s R assumes a linear relationship between the variables.
- **Interval Data**: The variables should be measured on an interval or ratio scale.
- **Normality**: The variables should be approximately normally distributed, especially for small sample sizes.
- Pearson’s R is widely used in various fields to determine the strength and direction of the linear relationship between two continuous variables.

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

Scaling is the process of transforming the features of your data so they fit within a certain range or follow a particular distribution. This is especially important in machine learning, where different algorithms may require data to be scaled to perform optimally.

Why Scaling is Performed:

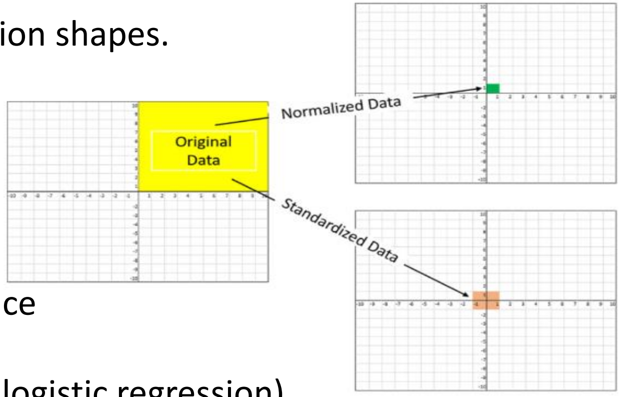
- ***Improved Model Performance*** : Most algorithms are sensitive to the scale of input data.
- ***Faster Responsiveness*** : Scaling helps gradient descent converge faster by ensuring that all features contribute equally to the learning process.
- ***Equal Contribution of Features*** : Ensures that each feature contributes proportionally to the model, preventing bias toward features with larger scales.

Types of Scaling:

1. *Normalized Scaling*: Rescales the data to a fixed range, typically [0, 1].

- Used when you want all features to have the same scale but retain their distribution shapes.
- Common in cases where data has varying units or ranges.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$



2. *Standardized Scaling*: Centers the data around the mean and scales it to have unit variance (mean = 0, standard deviation = 1). Also known as z-score normalization.

- Used when the model assumes normally distributed data (e.g., linear regression, logistic regression).
- Helps in stabilizing variance and making the feature distributions.

$$X_{std} = \frac{X - \mu}{\sigma}$$

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

Variance Inflation Factor (VIF) is a measure of multicollinearity among features in a regression model. When the VIF value is infinite, it indicates perfect multicollinearity among predictor variables. This typically happens due to:

1. ***Perfect Multicollinearity***: This occurs when one predictor variable is a perfect linear combination of others.
2. ***Redundant Predictors***: If there are redundant predictors in the model, such as including both a variable and its exact duplicate or linear transformation, it can lead to an infinite VIF.
3. ***Singular Matrix***: If R^2 = 1, which indicates perfect prediction, the denominator becomes zero, resulting in an infinite VIF.

$$VIF = \frac{1}{1 - R^2}$$

where, R2 is the coefficient of determination when regressing that predictor variable against all other predictor variables in the model.

When perfect multicollinearity is present, the determination coefficient (i.e. R) between the variables is equal to 1 (or -1), and this leads to the denominator in the VIF calculation becoming zero, which results in an infinite VIF value.

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

A ***Q-Q (Quantile-Quantile) plot*** is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

1. Assessing Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps visualize whether this assumption holds. If the residuals follow a normal distribution, the points in the Q-Q plot will approximately lie on a straight line.

2. Checking Model Fit: By examining the Q-Q plot, you can detect deviations from normality, such as skewness or kurtosis, which might indicate model mis-specification or issues with the data. This helps in diagnosing potential problems with the model and guiding corrective actions.

3. Validating Statistical Tests: Many statistical tests and confidence intervals rely on the assumption of normally distributed residuals. A Q-Q plot can help validate these assumptions, ensuring that the statistical inferences made from the model are reliable.