Assignment-based Subjective Questions

Q.1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical variables such as season, yr, mnth, holiday, weekday, workingday, and weathersit were analysed,

The following inference's can be made about their effect on the dependent variables

Season:

- Rentals are relatively consistent across the four seasons (spring, summer, fall, winter). There is no significant drop or spike, indicating that seasonality may not have a drastic impact on the overall number of rentals, but other seasonal factors could be influencing specific periods.

Weathersit (Weather Situation):

- Most rentals occur under 'Clear' weather conditions, followed by 'Mist'. 'Light Snow or Rain' conditions see a sharp drop in rentals. This suggests that adverse weather conditions discourage bike rentals, with clear weather being the most favourable.

Holiday:

- The majority of rentals happen on non-holidays. Very few rentals occur on holidays, suggesting that people are less likely to rent bikes on holidays, possibly due to alternative leisure activities or reduced commuting needs.

Workingday:

- More rentals occur on working days compared to non-working days. This indicates that a significant portion of bike rentals might be for commuting purposes during workdays.

Weekday:

- Rentals are fairly evenly distributed across all days of the week, with no significant peaks or troughs. This might suggest a steady demand for bike rentals throughout the week, without a strong preference for any particular day.

Month:

- The number of rentals is relatively consistent across different months, with a slight decrease in rentals during the summer months (July and August). This could be due to higher temperatures making biking less comfortable.

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

Using drop_first=True helps avoid the dummy variable trap, where the model might suffer from multicollinearity if all categories of a categorical variable are included. By dropping the first category, we prevent this issue and ensure that the model remains stable and interpretable. The category which we drop can be explained using other features.

Explanation:

- When you create dummy variables for a categorical feature with n categories, n dummy variables are generated by default.
- However, including all n dummy variables in a regression model introduces perfect multicollinearity, as the value of one dummy variable can be exactly determined by the others.
- To prevent this, drop_first=True is used to drop the first dummy variable. This leaves n−1 dummy variables, which are sufficient to capture the information while avoiding multicollinearity.

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

The pair-plot reveals that temp (temperature) and atemp (feeling temperature) has the highest positive correlation with the target variable count (total bike rentals).

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

The assumptions of Linear Regression were validated through several steps:

- **Linearity**: Checked by plotting residuals vs. fitted values to ensure no clear patterns.
- **Normality**: Residuals were plotted in a histogram or Q-Q plot to check for normal distribution.
- **Homoscedasticity**: Examined through residuals vs. fitted values to ensure constant variance.
- **Multicollinearity**: Variance Inflation Factor (VIF) was calculated to ensure no high correlation among predictors**.**

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

The top 3 features based on the final model are:

- **Temperature (temp)**: With the highest positive coefficient of 4160.0340, temperature strongly influences bike rentals. Higher temperatures lead to a significant increase in bike demand.
- **Year (year)**: The coefficient for the year variable (2007.7272) indicates a significant increase in bike rentals in 2019 compared to 2018. This suggests a strong positive trend in bike-sharing popularity over time.
- **Humidity (humidity)**: Humidity has a significant negative impact, with a coefficient of -1259.2241. Higher humidity levels tend to decrease bike rentals, making it an important factor in predicting demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as *Y*) and one or more independent variables (denoted as *X*). The algorithm tries to fit a line (in the case of simple linear regression) or a hyperplane (in the case of multiple linear regression) that best represents the data. The equation for a Multiple linear regression model is:

$$\mathrm{Y} = \mathrm{b}_0 + b_1\mathrm{X}_1 + \mathrm{b}_2\mathrm{X}_2 + \cdots + \mathrm{b}_\mathrm{n}\mathrm{X}_\mathrm{n} + \epsilon$$

Where:

- Y is the dependent variable.
- $X_1, X_2, \dots, X_n$ is the independent variable.
- b_0 is the intercept (the value of Y when X=0).
- $b_1, b_2, \dots b_n$ is the slope (the change in Y for a one-unit change in X).
- ϵ (epsilon) is the error term (the difference between the observed and predicted values).

The algorithm finds the best-fitting line by minimizing the sum of the squared differences between the observed values and the values predicted by the line (this method is known as **Ordinary Least Squares** or OLS).

Types of Linear Regression

- Simple Linear Regression: Involves one independent variable and one dependent variable, modelling the relationship as a straight line.
- Multiple Linear Regression: Involves two or more independent variables to predict the value of a dependent variable.

Assumptions of Linear Regression

Linear regression relies on several key assumptions:

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The residuals (errors) have constant variance across all levels of the independent variables.
- Normality: The residuals should be normally distributed, especially for inference (e.g., hypothesis testing).

Cost Function (Ordinary Least Squares)

The algorithm seeks to find the best-fitting line by minimizing the sum of the squared differences between the actual and predicted values of y. This is known as the Ordinary Least Squares (OLS) method. The cost function (also called the loss function) for OLS is:

The goal is to find the values of β0, β1, ..., βn that minimize this cost function.

$$J(\beta 0, \beta 1) = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

- y_i is the actual value of the dependent variable.
- $\widehat{y_i}$ is the predicted value from the linear model.
- n is the number of observations.

Optimization

To minimize the cost function, calculus (partial derivatives) is used to find the point where the gradient (slope) is zero. This involves taking the derivative of the cost function with respect to each parameter (i.e.β0, β1, …, βn) and setting it to zero.

Interpretation of Coefficients

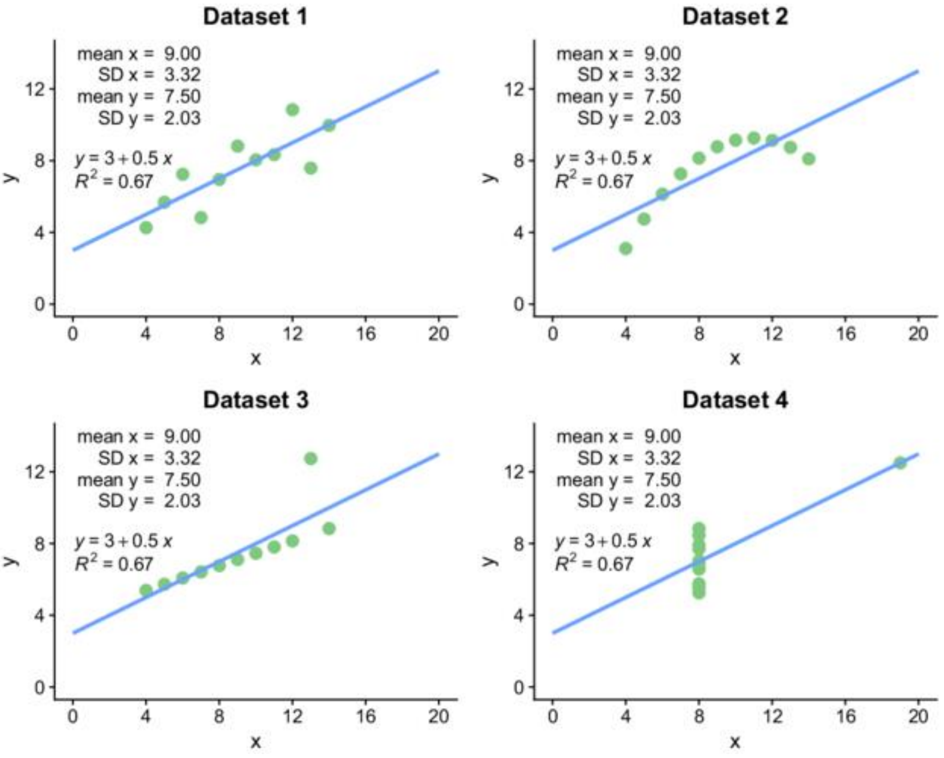
- **Intercept (β0)**: Represents the expected value of y when all independent variables are zero.
- **Slopes (β1, β2, …, βn)**: Represent the expected change in y for a one-unit increase in the corresponding independent variable, holding other variables constant.

Example: Suppose we have data on house prices (Y) and their sizes (X). A linear regression model could help you predict the price of a house based on its size

---

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four different datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but appear very different when graphed. These datasets were constructed by Francis Anscombe in 1973 to demonstrate the importance of data visualization before analysis.



Key Points:

- **Dataset 1**: Represents a simple linear relationship with a small amount of noise.
- **Dataset 2**: Has a clear non-linear relationship, yet the correlation is the same as Dataset 1.
- **Dataset 3**: Contains a single outlier, which heavily influences the correlation.
- **Dataset 4**: All the points lie on a vertical line except for one, drastically skewing the correlation.

---

3. What is Pearson's R?

Pearson's R (Pearson's correlation coefficient) is a measure of the linear correlation between two variables X and Y. Its value ranges between -1 and 1:

- **1**: Perfect positive linear correlation.
- **0**: No linear correlation.
- **-1**: Perfect negative linear correlation.

The formula for Pearson's R is:

$$r = \frac{\sum(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{[\sum(Xi - \bar{X})^2 * \sum(Yi - \bar{Y})^2]}}$$

Example: If the Pearson's R between study hours and exam scores is 0.85, it indicates a strong positive relationship, meaning as study hours increase, exam scores tend to increase.

---

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

Scaling is the process of adjusting the range of features in your dataset to ensure that all features contribute equally to the model. It's important because many machine learning algorithms (like gradient descent-based methods) are sensitive to the scale of data.

- **Normalization (Min-Max Scaling)**: Rescales the data to a fixed range, usually [0, 1]. The formula is:

$$Xscaled = \frac{X - Xmin}{(Xmax - Xmin)}$$

- **Standardization (Z-score Scaling)**: Rescales the data to have a mean of 0 and a standard deviation of 1. The formula is:

$$X_{scaled} = \frac{(X - \mu)}{\sigma}$$

---

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

Variance Inflation Factor (VIF) is used to detect multicollinearity in a regression model. A VIF value becomes infinite when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of one or more other predictor variables.

VIF for a predictor variable is calculated as:

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

Where, R^2 represents the coefficient of determination.

Reason: When multicollinearity is perfect, the denominator in the VIF formula (1 - R^2) becomes zero, leading to an infinite VIF value.

Example: If two variables, say X1 and X2, are perfectly correlated (e.g., X1 = 2 * X2), the VIF for these variables will be infinite.

---

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 plot** (Quantile-Quantile plot) is a graphical tool to compare the distribution of a dataset to a theoretical distribution (often a normal distribution). It plots the quantiles of the data against the quantiles of the theoretical distribution.

Use in Linear Regression: In linear regression, a Q-Q plot is used to check if the residuals (errors) follow a normal distribution. Normally distributed residuals are an assumption for many inferential statistics, like hypothesis testing.

Importance in Linear Regression: A Q-Q plot helps verify the normality assumption of residuals in linear regression. If the residuals deviate from the straight line in the plot, it suggests potential issues with the model, which could lead to unreliable predictions or conclusions**.**
