Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Following categorical variables have significant impact on dependent variable –

1) Weathersit
2) Season
3) Year
4) Month

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The numerical variable that have highest correlation with target variable is **temp.**

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I first calculated the residual error using –

res = y_train - y_train_pred

y_train – value of dependent variable (actual)
y_train_pred – value of dependent variable (predicted)

Then I plotted the residual error using seaborn distplot.

Finally to validate the assumption I observed that residual error is normally distributed with mean equal to 0.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Following 3 features contributes significantly –

1) Temp
2) Weather (Light Rain)
3) year

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is used to predict the value of a continuous target variable y based on the values of one or more input variables (features) x1,x2,…,xn.

There are 2 types of liner regression –

1) Simple Linear regression – It involves one independent variable and one dependent variable. It assumes a linear relationship between x and y.
   Equation for Simple linear regression – $y=\beta0+\beta1x+\epsilon$
   $\beta0$ – Intercept
   $\beta1$- Coefficient
   x – independent variable
   ϵ - error of estimation

2) Multiple Linear regression - Involves multiple independent variables. The model can be expressed as:
   $$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nx+\epsilon$$
   $\beta0$ – Intercept
   βn- Coefficient
   x – independent variable
   ϵ - error of estimation

Objective of Linear Regression –
The goal is to find the values of $\beta0$ and βi that minimise the difference between the predicted values and the actual values.
This difference is measured by the Residual Sum of Squares (RSS), which is the sum of squared errors (differences between observed and predicted values).

Evaluating the Model
Once the model is trained, its performance is evaluated using metrics such as:

- Mean Absolute Error (MAE): The average of absolute errors between predicted and actual values.
- Mean Squared Error (MSE): The average of squared errors.
- Root Mean Squared Error (RMSE): The square root of the average of squared errors.
- *R*2 Score: Indicates how well the independent variables explain the variance of the dependent variable (ranges from 0 to 1, with 1 indicating a perfect fit).

Assumptions of Linear Regression

- **Linearity**: The relationship between x and y is linear.
- **Independence of Errors**: Errors are independent of each other.
- **Homoscedasticity**: Constant variance of errors for all levels of x.
- **Normality of Errors:** Errors should be normally distributed.
- **No Multicollinearity (in multiple regression)**: Independent variables should not be highly correlated.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's correlation helps us understand the relationship between two quantitative variables when the relationship between them is assumed to take a linear pattern. The relationship between two quantitative variables (also known as continuous variables), can be visualized using a scatter plot, and a straight line can be drawn through them. The closeness with which the points lie along this line is measured by Pearson's correlation coefficient, also often denoted as Pearson's r.

Pearson's r can be thought of not just as a descriptive statistic but also an inferential statistic because, as with other statistical tests, a hypothesis test can be performed to make inferences and draw conclusions from the data

Pearson correlation coefficient formula –

The formula for Pearson's correlation coefficient, r, relates to how closely a line of best fit, or how well a linear regression, predicts the relationship between the two variables. It is presented as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling refers to the process of transforming data so that it fits within a particular range or has certain desired properties. It is a crucial step in data pre-processing, especially in machine learning, where the performance and convergence of many algorithms can be affected by the scale of the input features.

Scaling is performed because of following reasons -

1) **Improving Algorithm Performance**: Many machine learning algorithms (e.g., k-nearest neighbors, support vector machines, and neural networks) are sensitive to the scale of the data. Scaling can help these algorithms converge faster and improve accuracy.

2) **Ensuring Fair Comparisons**: When features have vastly different ranges (e.g., age in years vs. income in thousands), they may disproportionately influence the results. Scaling standardizes feature ranges so each variable contributes fairly.

3) **Handling Metrics in Distance-based Algorithms**: Algorithms that rely on distance metrics (like Euclidean distance in k-means or KNN) are sensitive to the scale of features. Scaling ensures that no single feature dominates the distance calculation due to a large range.

Types of Scaling Methods

Min-Max Scaling (Normalization):

- Scales data to a fixed range, typically [0, 1] or [-1, 1].
- Formula:

$$X' = \frac{X - X\ min}{X\ max - X\ min}$$

- Best for algorithms sensitive to feature ranges or data that doesn't follow a normal distribution.

Standardization (Z-score Scaling):

- Transforms data to have a mean of 0 and a standard deviation of 1.
- Formula:

$$X' = \frac{X - mean}{SD(X)}$$

SD(X) – standard deviation

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The **Variance Inflation Factor (VIF)** is a measure used to detect multicollinearity in a dataset, especially in multiple regression analysis. It indicates how much the variance of a regression coefficient is inflated due to the correlation with other predictor variables. Sometimes, the value of VIF can become **infinite**, which typically points to a severe multicollinearity issue.

VIF (X) = 1 / (1 – R2)

R2 – tells you how much variance in the data has been explained by the model.

If R2 is 1 then VIF can go to infinite.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, usually the normal distribution. It helps assess if the data follows a particular distribution by plotting the quantiles of the sample data against the quantiles of the theoretical distribution. If the data closely follows the theoretical distribution, the points in a Q-Q plot will roughly align along a straight line.

In linear regression, the Q-Q plot is a quick and effective way to visually assess whether residuals are normally distributed, ensuring the validity of regression assumptions. Properly diagnosing residual normality with a Q-Q plot allows for more reliable interpretations of statistical tests and improves the model's predictive quality.
