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)

When analyzing categorical variables in relation to the dependent variable (e.g., bike demand in a shared bike dataset), we look for patterns or relationships between categories and the outcome. For example:

- If the day of the week (categorical) is a variable, the demand for bikes could be higher on weekends and lower on weekdays, indicating that the type of day (weekend vs weekday) influences demand.
- If weather conditions (e.g., sunny, rainy, cloudy) are categorical, we might see a clear trend where sunny days lead to higher bike demand compared to rainy or cloudy days.

By analyzing the categorical variables, we infer how each category (or group) affects the dependent variable. Statistical techniques like ANOVA or Chi-square tests could also help in assessing whether there are significant differences in the means of the dependent variable across different categories.

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

Ans-When creating dummy variables for categorical features in regression models, drop_first=True is important because it prevents the model from having **multicollinearity**. Multicollinearity occurs when one of the dummy variables can be perfectly predicted from the others, which leads to instability in the model. By dropping the first category, we avoid creating redundant variables and ensure that the model does not suffer from perfect correlation between the categories. This makes the model more interpretable and statistically sound.

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

We need to look at the pair-plot of the numerical variables. The variable with the highest correlation (positive or negative) with the target variable would be the one showing the strongest linear relationship. For example, in a shared bike dataset, the number of **temperature** or **hour of the day** might have the highest correlation with bike demand, as demand could vary with temperature or peak during certain hours.

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

The assumptions of Linear Regression can be validated by:

- Linearity: Ensuring the relationship between independent variables and the dependent variable is linear. This can be checked using scatter plots.
- Independence: Checking if residuals are independent. Durbin-Watson test can help assess autocorrelation of residuals.
- Homoscedasticity: Ensuring the variance of residuals is constant across all levels of the independent variables. This can be checked by plotting residuals vs. fitted values, where random scatter indicates homoscedasticity.

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

After building the linear regression model, we look at the coefficients and their p-values to assess which features significantly contribute to explaining the demand. The features with the highest positive or negative coefficients and low p-values (indicating statistical significance) would be the top contributors. For example, temperature, hour of the day, and season might be the top features for bike demand, depending on the dataset.

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 and one or more independent variables by fitting a linear equation to observed data. In simple linear regression, the model assumes that the relationship between the dependent variable (Y) and independent variable (X) is of the form:

Y=β0+β1X+ϵY = \beta_0 + \beta_1X + \epsilonY=β0 +β1 X+ϵ

- β0\beta_0β0 is the intercept.
- β1\beta_1β1 is the slope (the coefficient of the independent variable X).

In multiple linear regression, the equation extends to multiple independent variables:

The algorithm works by minimizing the sum of squared errors (SSE), which is the difference between the observed values and the predicted values. Techniques like Ordinary Least Squares (OLS) are used to estimate the coefficients that minimize this error.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a set of four datasets created by Francis Anscombe in 1973 to demonstrate the importance of data visualization. All four datasets have nearly identical descriptive statistics, such as the same mean, variance, correlation, and regression line. However, when visualized, the datasets show completely different patterns:

1. Dataset 1: A linear relationship.
2. Dataset 2: A perfect quadratic relationship.
3. Dataset 3: A linear relationship with an outlier.
4. Dataset 4: A non-linear relationship with outliers.

The quartet emphasizes how summary statistics alone can be misleading and how data visualization is critical in understanding underlying patterns.

3. What is Pearson’s R?

Pearson’s R (also known as the Pearson correlation coefficient) measures the strength and direction of the linear relationship between two variables. The value of Pearson’s R ranges from -1 to +1:

- +1: Perfect positive linear relationship.
- -1: Perfect negative linear relationship.
- 0: No linear relationship.

It is calculated as:

r=∑(X−X¯)(Y−Y¯)∑(X−X¯)2∑(Y−Y¯)2r = \frac{\sum{(X - \bar{X})(Y - \bar{Y})}}{\sqrt{\sum{(X - \bar{X})^2} \sum{(Y - \bar{Y})^2}}}r=∑(X−X¯)2∑(Y−Y¯)2 ∑(X−X¯)(Y−Y¯)

Pearson’s R is sensitive to outliers and assumes a linear relationship between the variables.

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

Scaling refers to the process of transforming the features so that they have specific properties, such as a particular range or distribution. It is performed to make the data suitable for modeling, especially when the features have different units or scales.

-

  Normalized scaling (also known as min-max scaling) transforms the data to fit within a fixed range, usually [0, 1], by using the formula:

  - X′=X−XminXmax−XminX' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}X′=Xmax −Xmin X−Xmin
  -

  Standardized scaling (also known as z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1, using the formula:

  - X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ

Scaling ensures that no feature dominates others due to its scale and allows models that are sensitive to feature magnitudes, like linear regression or k-means clustering, to perform better.

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) measures how much the variance of the estimated regression coefficients is inflated due to multicollinearity. A VIF becomes infinite when there is perfect multicollinearity, meaning one of the independent variables is a perfect linear function of others. This happens when the independent variables are highly correlated, and one or more predictors provide redundant information. This results in unreliable estimates of the regression coefficients.

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 whether the residuals of a regression model follow a normal distribution. In the plot, the quantiles of the sample data are plotted against the quantiles of a standard normal distribution. If the residuals are normally distributed, the points will lie on or near a straight line.

The Q-Q plot is important in linear regression because one of the assumptions of linear regression is that the residuals should be normally distributed. If the points deviate significantly from the line, it indicates that the residuals are not normally distributed, suggesting potential problems with the model (such as non-linearity or outliers).