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?

Ans:- The specific inferences depend on the type of categorical variable and the context of the analysis. Here's a general approach to understanding their effects:

1. Use box plots, bar charts, or other visualizations to compare distributions across categories.
2. categorical variables are typically included as dummy variables (0/1). The coefficients of these dummy variables indicate the effect of each category on the dependent variable. You can analyse these coefficients to understand the impact.
3. The coefficient of a dummy variable representing a category shows the difference in the dependent variable compared to the baseline category. A positive coefficient indicates a higher value of the dependent variable relative to the baseline, while a negative coefficient indicates a lower value.
4. Statistical Significance: Look at the p-values associated with each coefficient to determine if the effect is statistically significant. A low p-value (typically < 0.05) suggests a significant effect.

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

Ans:- When there are n levels in a categorical variable it can be brought down to n-1 dummy variables, so one column can be removed, to remove one dummy variable we use this code.

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

Ans:- atemp

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

Ans:- By plotting a histogram on error terms.

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

Ans:- Month of the year, humidity and atemp are the 3 main features on which the demand can e forecasted.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans:- In the simplest form of linear regression (simple linear regression), the relationship between the dependent variable y and the independent variable x is modeled by the equation:

$y=\beta0+\beta1x+ \epsilon$

Where:

- y is the dependent variable.

- x is the independent variable.
- β0 is the y-intercept of the line.
- β1 is the slope of the line.
- ϵ is the error term (residual), which accounts for the variability in yyy that cannot be explained by x.

For multiple linear regression, the model generalizes to:

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets are used to demonstrate the importance of visualizing data and how statistical properties can be misleading if not accompanied by graphical analysis. Despite having nearly identical statistical properties, the datasets exhibit different patterns when visualized, highlighting how different datasets can have the same summary statistics but very different underlying distributions and relationships.

Visualization and Interpretation

The key lesson from Anscombe's quartet is that despite having nearly identical summary statistics, the datasets exhibit very different patterns when plotted:

1. **Dataset I**: Shows a linear relationship between x and y. This is the classic case of a linear relationship that fits a simple linear regression model well.
2. **Dataset II**: Shows a vertical line pattern. All x values are the same, and y values are distributed with some noise around a specific value. This indicates that x has no variability, making the correlation misleading.
3. **Dataset III**: Shows a clear non-linear pattern. The relationship between x and y is quadratic, indicating that a simple linear model would not capture the relationship well.
4. **Dataset IV**: Contains an outlier that dramatically affects the fit of the linear model. The outlier has a significant impact on the correlation and regression line, demonstrating how influential data points can skew results.

Importance of Anscombe's Quartet

1. **Emphasizes the Need for Graphical Analysis**: It illustrates that summary statistics like mean, variance, and correlation can be misleading without visualizing the data. Graphical representation is crucial for understanding the underlying patterns and relationships in the data.
2. **Demonstrates Data Integrity**: Shows how different types of data relationships can yield the same summary statistics but require different types of analysis.
3. **Educational Tool**: Used in teaching to highlight the importance of data visualization in exploratory data analysis and model fitting.

3. What is Pearson’s R?

Pearson's R (Pearson's correlation coefficient) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is one of the most commonly used methods to assess the correlation between variables.

Pearson's R is a measure of the linear correlation between two variables X and Y. It ranges from -1 to +1, where:

- **+1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

2. Formula

The Pearson correlation coefficient r is calculated as:

r=Cov(X,Y)/σXσY

Where:

- Cov(X,Y) is the covariance between the variables X and Y.
- σX is the standard deviation of X.
- σY is the standard deviation of Y.

• **Positive Correlation (+1 to 0)**:

- Values close to +1 indicate a strong positive linear relationship: as X increases, Y tends to increase.
- Values close to 0 suggest a weak positive linear relationship.

• **Negative Correlation (0 to -1)**:

- Values close to -1 indicate a strong negative linear relationship: as X increases, Y tends to decrease.
- Values close to 0 suggest a weak negative linear relationship.

• **Zero Correlation (0)**:

- An r value of 0 suggests no linear relationship between X and Y. However, this does not mean there is no relationship at all; there could be a non-linear relationship.

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

It’s a process of transforming data so that it fits within a specific scale, making comparison between different variables meaningful. It involves adjusting the range of values or standardizing the variables in a dataset.

Why Scaling is Performed?

Scaling is performed for several reasons:

1. **Equal Weightage**: Variables may have different units or scales (e.g., kilograms vs. meters). Scaling brings all variables to a comparable level, ensuring no single variable dominates due to its larger scale or range.
2. **Algorithm Requirements**: Many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and close to normally distributed. Algorithms like K-nearest neighbors (KNN), support vector machines (SVM), and neural networks are particularly sensitive to feature scaling.
3. **Interpretability**: Interpretation and visualization of data are often easier when variables are on a similar scale.

Types of Scaling

1. Normalized Scaling (Min-Max Scaling)

- **Method**: Normalized scaling rescales the data to a fixed range, usually between 0 and 1.
- **Formula**: For each feature

  x (scaled)= x−min(X)/ max(X)−min(X)

- Where min⁡(X)\min(X)min(X) and max⁡(X)\max(X)max(X) are the minimum and maximum values of X, respectively.
- **Benefits**:
  - Simple and intuitive.
  - Preserves the shape of the original distribution.
  - Useful when data needs to be on a specific scale, such as for algorithms that require input values to be in a bounded interval.
- **Drawbacks**:
  - Sensitive to outliers, as it compresses the range of the data.

2. Standardized Scaling (Z-score Normalization)

- **Method**: Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1.
- **Formula**: For each feature x:

  xscaled=x−μ/ σ

- Where μ\muμ is the mean of X, and σ\sigmaσ is the standard deviation of X.

- **Benefits**:

- Handles outliers better than normalized scaling, as it is less sensitive to the range of values.
- Useful when the distribution of data is not known or is not assumed to be Gaussian.

- **Drawbacks**:

  - Does not bound data to a specific range, which may be required for certain algorithms or interpretations.

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

The occurrence of infinite values in the Variance Inflation Factor (VIF) typically happens due to perfect multicollinearity in the data. Let's delve into why this happens:

Understanding VIF

The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in a regression analysis. It measures how much the variance of a regression coefficient is inflated due to multicollinearity among the predictor variables.

Infinite VIF values occur in regression analysis when perfect multicollinearity exists among predictor variables. This situation arises when one or more variables can be perfectly predicted by a linear combination of others, leading to indeterminate coefficient estimates. Identifying and addressing perfect multicollinearity is essential for accurate regression modeling and interpretation of results.

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

Ans:- A **Q-Q plot** (Quantile-Quantile plot) is a graphical technique used to assess if a dataset follows a specific theoretical probability distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the theoretical distribution.

Importance of Q-Q plot

a Q-Q plot is particularly useful for assessing the **normality of residuals**. The residuals (errors) are the differences between the observed values and the values predicted by the model. For valid inferences from a linear regression model, it is important that these residuals meet certain assumptions.

Use of Q-Q Plot

- **Fit the Linear Regression Model**: Run your regression analysis and obtain the residuals.
- **Create a Q-Q Plot**: Plot the quantiles of the residuals against the quantiles of a normal distribution.
- **An alyze**:

  - **Straight Line**: If the points lie along the 45-degree line, the residuals are approximately normally distributed.

- **Deviations**: If the points deviate significantly from the line, assess the pattern to determine if adjustments to the model or further investigation are needed.