Assignment based questions

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

   There are two categorical variables – seasons and weathersit. By visualizing the pairplot, we see the slope for each value in the variables are same.

   There are 4 seasons and therefore we see kind of 4 straight line in the histogram for seasons vs cnt. For summer, fall & winter, the count of bikes rental are almost similar whereas, in spring, it's lower as compared to the other seasons. Again, if we investigate the heatmap after creating dummy variable for seasons, there is a high negative correlation between spring and cnt.

   For the weathersit variable, as it's a categorical variable, again we see three separate straight line in the histogram created for weathersit vs cnt. From the same plot, we can see the number of rental bikes is higher when the weather is clear or somewhat misty. Whereas the number of bikes rental is lower when it's raining.

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

   While creating dummy variables, it's important to use drop_first=True for avoiding multicollinearity in regression analysis.
   When we include dummy variables for all the levels of categorical variables, there are high chances for multicollinearity as one level can be easily identified by the other levels. Therefore, it can produce unstable coefficient estimates and inflated standard errors.
   Also, if we include dummy variables for all the levels, that will lead to perfect multicollinearity because the sum of all dummy variables for each observation will equate to 1. Thus, the regression coefficients become uninterpretable and model becomes unstable.

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

   As per the pair-plot, temp variable has the highest correlation with the target variable.

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

   The assumption of Linear Regression is validated using the Residual Analysis.

   After creating the regression model with train dataset, we try to predict the target variable using the sort listed feature set. Thereafter, we calculate the error terms between the best fitted target variable (y) and predicted target variable(y bar).

   Visualizing the histogram plot of the error terms, we find out that the error terms are normally distributed, which satisfies the assumptions of linear distributions.

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

   We can analyze the top 3 features contributing significantly by analyzing their respective coefficients. From the final model , we can conclude that the most significant features to explaining the demand of the shared bikes are – temperature(temp), yr(year) and light-rain(weathersit).

   Temp & yr having high positive coefficient of 3646.34 and 2042.97 respectively. Which signifies that count will increase by 3646.34 and 2042.97 units for 1unit increment in temp and yr respectively.

   Whereas, light-rain having negative coefficient of -2444.17 signifies that cnt will decrease by 2444.17 unit for 1 unit increment in light-rain feature.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used for predicting the value of a dependent variable based on one or more independent variables. Here's a detailed explanation of the algorithm:

Problem Statement: Linear regression is typically used when there is a linear relationship between the independent variable(s) and the dependent variable. The goal is to find the best-fitting straight line that describes this relationship.

Model Representation: In simple linear regression, there is one independent variable (x) and one dependent variable (y). The relationship between (x) and (y) is modeled as:

y = beta_0 + beta_1 x + error

Where:

- (y) is the dependent variable
- (x) is the independent variable
- (beta_0) is the intercept (the value of (y) when (x) is 0)
- (beta_1) is the slope (the change in (y) for a one-unit change in (x))
- (error) is the error term, representing the difference between the observed and predicted values of (y)

Objective Function: The objective is to find the values of (beta_0) and (beta_1) that minimize the error between the predicted values and the actual values of (y). This is typically done by minimizing the sum of squared errors (SSE):

SSE = sum_{i=1}^{n} (y_i - (beta_0 + beta_1 x_i))^2

Where:

- (n) is the number of data points
- (y_i) is the actual value of the dependent variable for the (i^{th}) data point
- (x_i) is the value of the independent variable for the (i^{th}) data point

Optimization: This is usually done using optimization techniques like gradient descent or the normal equation. Gradient descent iteratively updates the values of (beta_0) and (beta_1) to minimize the SSE, while the normal equation directly computes the optimal values of (beta_0) and (beta_1).Model Evaluation: Once the model parameters are estimated, the model needs to be evaluated to assess its performance and generalization ability. This can be done using metrics such as mean squared error (MSE), R-squared, and others.

Assumptions: Linear regression relies on several assumptions, including:

Linearity: The relationship between the independent and dependent variables is linear.

Independence: The errors are independent of each other.

Homoscedasticity: The variance of the errors is constant across all levels of the independent variable.

Normality: The errors are normally distributed.

Extensions: Linear regression can be extended to handle multiple independent variables (multiple linear regression), polynomial relationships, interactions between variables, and more complex models.Overall, linear regression is a simple yet powerful algorithm for modeling the relationship between variables and making predictions. It's widely used in various fields including statistics, economics, finance, and machine learning.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear quite distinct when plotted. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of relying solely on summary statistics. Here's a detailed explanation of Anscombe's quartet:

Dataset Description: Anscombe's quartet consists of four datasets, each containing 11 (x, y) pairs. Despite having different distributions, all four datasets have nearly identical summary statistics, including means, variances, correlations, and regression lines.

Dataset Visualization: When plotted, each dataset in the quartet reveals a unique pattern:

Dataset I: Forms a simple linear relationship.

Dataset II: Forms a non-linear relationship, where a linear model would be inappropriate.

Dataset III: Appears to follow a linear relationship except for one outlier, which significantly influences the regression line.

Dataset IV: Has an outlier that heavily influences the correlation coefficient and regression line, despite the other data points forming a perfect relationship when the outlier is removed.

Statistical Summary: Despite their differences, all four datasets share the following summary statistics:

- Mean of (x) and (y)
- Variance of (x) and (y)
- Correlation coefficient between (x) and (y)
- Linear regression line equation (slope and intercept)

Implications: Anscombe's quartet illustrates several important points:

Visualizing data is crucial: Summary statistics alone may not reveal the true nature of the data. Visual inspection can uncover patterns, outliers, and relationships that summary statistics might overlook.

Caution with assumptions: Relying solely on summary statistics can lead to incorrect conclusions. For example, assuming linearity based on a high correlation coefficient can be misleading if the relationship is actually non-linear.

The importance of exploratory data analysis: Before applying statistical methods or building models, it's essential to explore and understand the data thoroughly.

Educational Tool: Anscombe's quartet is often used in statistics education to emphasize the importance of data visualization, exploratory analysis, and the limitations of summary statistics. It serves as a reminder to approach data analysis with skepticism and to verify assumptions through visualization and exploration.

In summary, Anscombe's quartet highlights the importance of visualizing data and the limitations of relying solely on summary statistics. It challenges analysts to explore and understand their data thoroughly before drawing conclusions or making decisions.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as (r), is a measure of the linear relationship between two variables. It quantifies the strength and direction of the relationship between two continuous variables. Pearson's (r) ranges from -1 to 1:

(r = 1): Indicates a perfect positive linear relationship. As one variable increases, the other also increases proportionally.

(r = -1): Indicates a perfect negative linear relationship. As one variable increases, the other decreases proportionally.

(r = 0): Indicates no linear relationship between the variables.

The formula for Pearson's correlation coefficient between variables (X) and (Y) is:

r = sum((X_i - bar{X})(Y_i - bar{Y})) / sqrt(sum((X_i - bar{X})^2) sum((Y_i - bar{Y})^2))

Where:

- (X_i) and (Y_i) are individual data points.
- (bar{X}) and (bar{Y}) are the means of variables (X) and (Y) respectively.
- The summations are over all data points.

Pearson's correlation coefficient measures only linear relationships and assumes that both variables are normally distributed. It is sensitive to outliers and can be influenced by extreme values.

Pearson's (r) is widely used in various fields such as statistics, economics, psychology, and biology to assess the strength and direction of relationships between variables.

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

Scaling is the process of transforming the values of variables to a specific range or distribution. It's often performed as a preprocessing step in data analysis and machine learning to ensure that all variables have similar scales or distributions. Scaling is important for several reasons:

Equal Weighting: Many machine learning algorithms use distance-based metrics (e.g., Euclidean distance) to measure similarity between data points. If the variables are not on the same scale, those with larger scales may dominate the distance calculation, leading to biased results. Scaling helps prevent this issue by giving each variable equal weight.

Convergence Speed: In optimization algorithms like gradient descent, variables with larger scales can lead to slower convergence. Scaling the variables can help speed up the convergence process.

Model Interpretability: Scaling makes it easier to interpret the coefficients or importance of variables in a model. Without scaling, it might be challenging to determine the true significance of a variable's contribution to the model.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized Scaling: In normalized scaling, also known as min-max scaling, the values of the variables are transformed to fall within a specific range, typically between 0 and 1. The formula for normalized scaling is:

X_scaled = X - X_min / X_max - X_min

Where:

- (X) is the original value of the variable.
- (X_{\text{min}}) is the minimum value of the variable.
- (X_{\text{max}}) is the maximum value of the variable.

Normalized scaling preserves the relative relationships between the values but does not preserve the original distribution or variance of the data.

Standardized Scaling: In standardized scaling, also known as z-score normalization, the values of the variables are transformed to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

X_scaled = X – mu / sigma

Where:

- (X) is the original value of the variable.
- (mu) is the mean of the variable.
- (sigma) is the standard deviation of the variable.

Standardized scaling preserves the mean and variance of the original data and transforms it into a standard normal distribution.

In summary, scaling is performed to ensure that variables have similar scales or distributions, which can improve the performance, convergence speed, and interpretability of machine learning models. Normalized scaling transforms variables to a specific range, while standardized scaling transforms them to have a mean of 0 and a standard deviation of 1.

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

When the value of VIF becomes infinite, it indicates severe multicollinearity among the predictors. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This high correlation leads to issues in estimating the regression coefficients accurately, as the predictors are providing redundant information about the dependent variable.

Here are some reasons why VIF might become infinite:

Perfect Multicollinearity: When two or more predictor variables are perfectly correlated, meaning one can be expressed as a perfect linear combination of the others, VIF becomes infinite. In this case, one predictor variable is entirely redundant, providing no additional information to the model.

Nearly Perfect Multicollinearity: Even if multicollinearity is not perfect but still very high, VIF can approach infinity. This occurs when there is a near-linear relationship among the predictor variables, making it difficult for the regression model to estimate the coefficients accurately.

Small Sample Size: In datasets with a small number of observations relative to the number of predictor variables, VIF estimates may become unstable, leading to inflated VIF values. Small sample sizes exacerbate the effects of multicollinearity.

Duplicated Data: In some cases, duplicated data or repeated measurements can lead to artificially high correlation among predictor variables, resulting in inflated VIF values.

Overfitting: When the model is overfitted, meaning it captures noise in the data rather than true relationships, VIF values can be inflated.

Dealing with infinite VIF values involves identifying and addressing the root cause of multicollinearity. This might include dropping redundant variables, combining correlated variables, collecting more data to increase sample size, or regularizing the model to reduce overfitting. It's essential to address multicollinearity to ensure the stability and accuracy of the regression model's parameter estimates.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to compare the distribution of a sample dataset to a theoretical distribution, typically a normal distribution. The main purpose of a Q-Q plot is to visually assess whether the data come from a specified distribution, such as a normal distribution. Here's how it works and its importance in linear regression:

Use and Importance in Linear Regression:

- Q-Q plots are commonly used in linear regression to assess the assumption of normality of residuals.
- In linear regression, residuals are the differences between the observed values and the values predicted by the regression model.
- One of the assumptions of linear regression is that the residuals are normally distributed with mean zero and constant variance (homoscedasticity).
- By examining the Q-Q plot of the residuals, we can visually inspect whether the residuals follow a normal distribution.
- If the Q-Q plot shows the points deviating significantly from the straight line, it indicates that the residuals may not be normally distributed.
- Departure from normality in the residuals could affect the validity of statistical inferences made from the regression model, such as hypothesis testing and confidence intervals.
- Addressing deviations from normality may involve transforming the response variable or using robust regression techniques.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression for assessing the normality assumption of residuals. By visually inspecting the plot, researchers can identify potential departures from normality and take appropriate measures to ensure the validity of the regression analysis.