# 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?**

Season:

Spring: It looks like lower bike rental numbers compared to others, as it follows winter, and temperatures might still be relatively low, which could deter people from cycling.

Summer: It is expected to have higher bike rentals due to warmer weather, making it more conducive for biking activities.

Fall: It is looking like, similar to summer, fall might also exhibit high bike rentals, especially in the early part, as the weather remains favorable for outdoor activities.

Winter: It is most likely to see a decrease in bike rentals, as colder temperatures and potential snowfall make biking less appealing.

Weather Situation (weathersit):

Clear: It looks like, days with clear weather or few clouds are expected to have the highest bike rental counts, as most people prefer biking in clear and sunny conditions.

Mist + Cloudy: It is looking that, the slightly lower rentals can be expected in misty or cloudy conditions compared to clear days, but the impact might not be as significant unless visibility or comfort is severely impacted.

Light Snow, Light Rain: Significant drops in bike rentals are expected during light snow or rain, as adverse weather conditions can deter casual and even some regular bikers.

Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog: Although not directly analyzed due to dataset limitations, it's reasonable to infer that severe weather conditions would lead to the lowest bike rental counts due to safety concerns and discomfort.

From the inferences, you can note that weather conditions and seasons significantly impact bike-sharing demand. Favorable weather (clear skies, moderate temperatures) and seasons (summer, fall) are conducive to higher bike rentals, while adverse conditions (rain, snow) and colder seasons (winter, and to some extent, spring) reduce demand.

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

It looks like using drop_first=True during dummy variable creation is an important practice to avoid the issue of multicollinearity in linear models, which include linear regression, logistic regression, and any other model that assumes little to no multicollinearity among the independent variables. Multicollinearity refers to the situation where one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. This can lead to various problems, including:

Reduced Interpretability: It looks like it becomes more challenging to assess the impact of independent variables on the dependent variable because changes in one predictor associated with changes in another predictor.

Increased Standard Errors: It is expected that the multicollinearity increases the standard errors of the coefficients. This increase means that coefficients for some variables may be deemed not statistically significant when they should be.

Unstable Coefficients: It looks like, small changes in the model or the data can lead to large changes in the coefficients, making the model sensitive and less reliable.

When you create dummy variables for a categorical variable with N categories, you introduce N new binary variables. However, these N variables are not independent of each other—they are linearly related since knowing the values of N-1 of these variables will perfectly predict the value of the Nth variable (if the first N-1 are all 0, the Nth must be 1, and vice versa). This perfect multicollinearity violates the assumptions of linear models.

By setting drop_first=True, you drop one of the dummy variables (usually the first category becomes the reference category), and thus only N-1 dummy variables are included in the model. This action effectively removes the perfect multicollinearity because now the presence of all N-1 dummies set to 0 does not automatically mean the dropped category is 1; it just reverts to being the baseline against which the other categories are compared. This approach simplifies the model without losing any information, as the effects of the categorical variable on the dependent variable can still be fully captured and interpreted relative to the reference category.

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

The pair plot generated shows the relationships between all selected numerical variables and the target variable cnt. While I cannot directly interpret the visual details from the plot in this text-based interface, typically in such analyses:

Temperature (temp and atemp): These variables often show a strong positive correlation with bike rental counts (cnt). This means as the temperature increases, the number of bike rentals tends to

increase as well, up to a certain point. Between temp and atemp, the one with a visually tighter and more linear relationship with cnt in the scatter plots would indicate the higher correlation.

Humidity (hum): Humidity might show a negative correlation or a less pronounced relationship with cnt. High humidity levels can be uncomfortable for physical activities like biking, potentially reducing demand.

Windspeed (windspeed): Windspeed might also negatively correlate with cnt, as higher winds can make cycling more challenging and less enjoyable, thereby reducing the number of rentals.

Based on typical outcomes from such analyses, temperature variables (temp and atemp) are likely to have the highest correlation with the target variable cnt, indicating that warmer conditions tend to encourage more bike rentals. To identify which specific variable among temp and atemp has the highest correlation with cnt, one would look for the scatter plot in the pair plot visualization that shows the closest to a linear upward trend and the tightest clustering of points along that trend line.

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

Validating the assumptions of linear regression is crucial to ensure the reliability and accuracy of the model's predictions. Linear regression assumptions include linearity, homoscedasticity, independence, and normality of residuals. Here's how these assumptions can be validated after building the model on the training set:

Linearity: The relationship between each independent variable and the dependent variable should be linear. This can be visually checked using scatter plots of observed vs. predicted values or individual feature vs. target plots before model training. Additionally, plotting residuals vs. predicted values can help identify non-linearity.

Homoscedasticity: The residuals should have constant variance across all levels of the independent variables. This means the size of the residual should be consistent across all values of the independent variables. You can check for homoscedasticity by looking at a plot of residuals vs. predicted values. If the plot displays a pattern (such as a funnel shape), it suggests heteroscedasticity.

Independence: Residuals should be independent of each other, which means the residuals from one prediction should not inform or predict the residuals from another. This assumption is intrinsic to the data collection process and is often validated through the study design. For time series data, autocorrelation in residuals can be tested with the Durbin-Watson test.

Normality of Residuals: The residuals of the model should be normally distributed. This can be checked using a histogram or a Q-Q (quantile-quantile) plot of the residuals. If the residuals are normally distributed, the data points in the Q-Q plot should fall approximately along a straight line.

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

Top 3 Features Contributing to Bike Demand:
one_hot__weathersit_light_snow_rain 2098.815375
remainder__yr 1970.702648
one_hot__season_spring 999.726420
dtype: float64

# General Subjective Questions

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

In statistical and machine learning, linear regression is a basic procedure that is used to predict a continuous dependent variable from one or more independent variables. It is expected that there is a linear relationship between the independent variable or variables and the dependent variable. Depending on how many independent variables are used, there are two variants of linear regression:

Simple Linear Regression: Involves only one independent variable to predict the dependent variable. It aims to find a linear relationship between the two variables. The model is represented by the equation:

$y=\beta0 + \beta1\ x + \epsilon$

y is the dependent variable (target).

x is the independent variable (predictor).

β0 is the y-intercept of the regression line.

β1 is the slope of the regression line, indicating the change in y for one unit change in x

ϵ is the error term, the difference between the observed values and the values predicted by the model.

Multiple Linear Regression: Extends simple linear regression by using two or more independent variables to predict the dependent variable. The model is represented by the equation:

$y=\beta0 + \beta1\ x1 + \beta2\ x2 + \beta3\ x3 + ... \beta n\ xn + \epsilon$

y is the dependent variable (target).

x1, x2, x3 are the independent variable (predictor).

β0 is the y-intercept of the regression line.

β1, β2, β3 are the coefficients of the independent variables

ϵ is the error term, the difference between the observed values and the values predicted by the model.

Best Fit Line/Plane: In linear regression, the "best fit" line (in simple linear regression) or plane (in multiple linear regression) is determined through the method of least squares, which minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Coefficients Estimation: The coefficients (β) are estimated during the training phase using the least squares method. For simple linear regression, explicit formulas can be used, while for multiple linear regression, numerical techniques (like gradient descent) or analytical methods (like the normal equation) might be employed.

Linear regression is widely used for prediction and forecasting, where its simplicity and interpretability make it a valuable tool for statistical analysis and machine learning.

**2. Explain the Anscombe's quartet in detail.**

Key Properties of Anscombe's Quartet:

Mean and Variance: For all four datasets, the mean of the x values is the same, the mean of the y values is also the same, the variance of the x values is the same, and the variance of the y values is the same.

Correlation: The correlation coefficient between x and y variables is the same for all four datasets, suggesting a similar linear relationship.

Linear Regression: When a linear regression is applied to each dataset, the slope and intercept of the regression line are the same (or very close) across the datasets.

Graphical Representation: Despite these similarities in summary statistics and linear regression results, the datasets look very different when graphed. Each set highlights different issues and patterns:

Dataset I shows a simple linear relationship between x and y variables, fitting the assumptions of linear regression closely.

Dataset II demonstrates a clear non-linear relationship (curvilinear pattern), for which linear regression is inappropriate despite the same statistical properties.

Dataset III includes an outlier that affects the slope of the regression line, showing a situation where a single outlier can have a significant impact on the statistical analysis.

Dataset IV consists of x values that are nearly the same for all points except an outlier, demonstrating how a single influential point can drastically affect the regression line.

It's Significance:

Visualization: Its underscores the critical importance of visualizing data before starting the analysis. Graphical representations can reveal data characteristics and anomalies that summary statistics cannot.

Analysis Assumptions: It illustrates that different distributions can yield similar statistical properties, highlighting the need to check the assumptions of statistical analyses (e.g., linearity, normality, and homoscedasticity in linear regression).

Influence of Outliers: It shows how outliers can significantly influence statistical summaries and analytical outcomes, emphasizing the need for robust outlier detection and handling.

Teaching Tool: Anscombe's quartet is widely used as a teaching tool to encourage students and researchers to graph their data and look beyond the numbers before drawing conclusions.

**3. What is Pearson's R?**

Pearson's R, also known as the Pearson product-moment correlation coefficient (Pearson's correlation coefficient), is a measure of the linear correlation between two variables X and Y. It gives an indication of the strength and direction of their linear relationship. Pearson's R values range from -1 to 1, where:

1 indicates a perfect positive linear relationship,

-1 indicates a perfect negative linear relationship, and

0 indicates no linear relationship.

Closer to 1 or -1: Indicates a stronger linear relationship between the two variables. A positive value suggests that as one variable increases, the other variable also increases (positive correlation), while a negative value suggests that as one variable increases, the other decreases (negative correlation).

Closer to 0: Suggests a weaker linear relationship, meaning the variables do not have a strong linear correlation.

Applications

Pearson's R is widely used in the fields of science, finance, social sciences, and data analysis to:

Examine the strength and direction of the linear relationships between variables.

Perform exploratory data analysis to guide further statistical analysis or model building.

Limitations

It only measures linear relationships; non-linear relationships might not be well represented by Pearson's R.

It can be influenced by outliers, which can artificially inflate or deflate the correlation coefficient.

The presence of a correlation does not imply causation; two variables may be correlated due to their relationship with a third variable or other underlying factors.

Use Cases

Pearson's R is particularly useful in preliminary data analysis when determining which variables might have linear relationships, either for inclusion in linear models or for identifying potential confounding factors in experimental designs.

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

In order to normalise the range of independent variables or data features, scaling is a technique used in data preprocessing. Scaling is important in machine learning and data analysis because data values can vary greatly in range. When characteristics have varying sizes, some may have a greater impact on the model than others because of their higher magnitude, which might result in biassed models that don't effectively reflect the underlying data. By scaling, you can make sure that every feature makes a roughly proportionate contribution to the final forecast.

Why scaling?:

Algorithm Performance: Many machine learning algorithms perform better or converge faster when features are on a similar scale. Gradient descent-based algorithms, in particular, benefit from scaling as it ensures an equal step size across features.

Distance Calculation: Algorithms that rely on distance calculations, such as k-Nearest Neighbors (k-NN), and clustering algorithms like k-means, are affected by the scale of features because the distance between points is influenced by the magnitude of the features.

Model Interpretability: Scaling can help make the model training process more interpretable. It becomes easier to understand the importance of each feature when they’re on the same scale.

Normalization vs. Standardization:

Normalization and standardization are two common scaling techniques, each useful in different scenarios:

Normalization (Min-Max Scaling): This technique scales and transforms features to a range between 0 and 1 or -1 and 1 if there are negative values. The formula for min-max scaling is:

X norm = (X - X min) / (X max - X min)

Normalization is useful when you know the distribution of the data does not follow a Gaussian distribution. This scaling compresses all features into a given range, which is particularly useful for algorithms that compute distances between data points.

Standardization (Z-score Normalization): This technique transforms the features so they have a mean of 0 and a standard deviation of 1, following a standard normal distribution. The formula for standardization is:

X std = (X - mu)/ (sigma)

where mu is mean and sigma is standard deviation

Differences

Range: Normalization maps the data to a specific range (0 to 1), while standardization shifts the data to have a mean of 0 and a standard deviation of 1 without bounding the data to a fixed range.

Distribution: Standardization does not assume any specific distribution of the data, whereas normalization can compress outliers in a fixed range, which might distort the distances between the points.

Use Case: Normalization is typically used when the algorithm assumes data to be bounded within a specific range (e.g., neural networks), whereas standardization is more common when the algorithm assumes data to follow a Gaussian distribution.

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

The Variance Inflation Factor (VIF) is a measure used to detect the presence and severity of multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, making it difficult to ascertain the individual effect of each variable on the dependent variable. The VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

A VIF value:

1 indicates no correlation between the independent variable and the other variables.

Between 1 and 5 generally suggests a moderate correlation, but not severe enough to require attention.

Greater than 5 or 10 indicates potentially problematic levels of multicollinearity, depending on the sources you consult.

Infinite VIF

An infinite VIF occurs under a specific condition where the independent variable is perfectly linearly correlated with other independent variables. This perfect multicollinearity means that the independent variable in question can be perfectly predicted from the others with no error.

In mathematical terms, during the calculation of VIF, the denominator becomes zero because it involves subtracting 1 from the R-squared value obtained by regressing the independent variable against all other

independent variables. When there's perfect multicollinearity, the R-squared value is 1 (indicating a perfect fit), VIF becomes infinite

Why Does Perfect Multicollinearity Happen?

Perfect multicollinearity can occur due to various reasons, including:

Data Collection Process: Duplicating an independent variable, either exactly or as a proportion of another.

Derived Variables: Including variables in the model that are calculated from other variables in the model (e.g., using both total income and individual incomes that sum to the total).

Lack of Data Variation: Sometimes, the dataset might not have enough variation in the independent variables, leading to situations where one variable can predict another with complete accuracy.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y=x. Q-Q plots are commonly used to assess the assumption of normality in a dataset, which is an important consideration in linear regression analysis.

Use of Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to evaluate the normality of residuals, which are the differences between the observed values and the values predicted by the regression model. The normality of residuals is one of the key assumptions of linear regression, ensuring the validity of statistical tests for the regression coefficients. Specifically, a Q-Q plot helps in the following ways:

Assessing Normality: By plotting the quantiles of residuals against the theoretical quantiles of a standard normal distribution, a Q-Q plot can visually assess whether the residuals follow a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will fall approximately along a straight line.

Identifying Deviations from Normality: A Q-Q plot can reveal departures from normality, such as skewness and kurtosis. For example, if the residuals are skewed, the Q-Q plot will deviate systematically from the y=x line, curving either upward or downward.

Detecting Outliers: Outliers can also be identified in a Q-Q plot. Points that fall far away from the main cluster of points may indicate outliers in the data, which could potentially influence the regression model disproportionately.

Importance of Q-Q Plot in Linear Regression

Model Diagnostics: Evaluating the normality of residuals is crucial for the reliability of hypothesis tests on the regression coefficients. Non-normal residuals can indicate that the linear regression model might not be appropriate for the data, potentially leading to inaccurate estimates and predictions.

Improving Model Fit: Identifying deviations from normality can guide analysts in transforming the dependent variable or applying different modeling techniques better suited to the data distribution, thereby improving the model fit.

Assumption Validation: The Q-Q plot is a simple yet powerful diagnostic tool for validating the assumptions underlying linear regression. Validating these assumptions is essential for making accurate inferences about the relationship between the independent and dependent variables.