Bike Sharing Assignment

Subjective Questions

Submitted by:
Ravikanth Nagaraj

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- **Effect of Categorical Variables on Bike Demand**

1. **Season:**
    1. **Fall:** The highest bike demand occurs in the Fall season, indicating favorable conditions for biking.
    2. **Summer:** Demand is also high in Summer, with counts being greater than Winter but not as high as Fall.
    3. **Winter:** Winter shows intermediate demand levels, suggesting that while biking is still popular, it is less so than during Fall and Summer.
    4. **Spring:** The lowest demand is observed in Spring, potentially due to less favorable weather conditions.
2. **Year:**
    1. **2019 vs. 2018:** There is a noticeable increase in bike demand from 2018 to 2019, suggesting that the bikesharing program gained popularity during this period.
3. **Month:**
    1. **High-Demand Months:** June, July, August, and September are the months with the highest bike rentals, with September having the peak demand. The warm summer weather likely contributes to this increased demand.
    2. **Low-Demand Months:** December shows the lowest rentals, while January and February also exhibit low demand. This can be attributed to colder winter weather, which discourages biking.
4. **Holiday:**
    1. **Holidays vs. Non-holidays:** Bike demand is higher on holidays compared to non-holidays. This increase may be due to people having more leisure time and choosing to bike for recreation or errands.
5. **Weekday:**
    1. **Even Distribution:** Demand is relatively evenly distributed across all weekdays, indicating consistent usage throughout the week.
    2. **Slight Increase on Fridays and Saturdays:** There is a slight uptick in bike usage on Fridays and Saturdays, though this difference is not very pronounced.
6. **Weather Situation (Weathersit):**
    1. **Clear Weather:** The highest bike demand occurs during clear weather conditions, which are conducive to biking.
    2. **Adverse Weather:** Demand decreases significantly during misty conditions, light snow/rain, and heavy snow/rain. There is no recorded bike usage during heavy snow/rain, while light snow/rain shows the least demand due to the hazards and discomfort associated with adverse weather.

- **Conclusion**
- These categorical variables illustrate significant trends and patterns that impact bike demand. Understanding these influences can help in strategic planning, such as adjusting bike availability, targeting marketing efforts, and enhancing user experience based on anticipated weather conditions and seasonal trends.

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

Importance of Using drop_first=True in Dummy Variable Creation

- Preventing Multicollinearity:

Including all dummy variables for a categorical variable can result in multicollinearity, where predictor variables are highly correlated. This makes it challenging to isolate the individual effect of each predictor on the target variable. Dropping the first dummy variable helps to mitigate this issue by removing redundancy.

- Reducing Redundancy:

By excluding the first category (the reference category), the total number of dummy variables is reduced. This simplification enhances model efficiency and performance.

- Enhancing Model Interpretability:

Dropping the first category ensures that the model remains interpretable. It avoids redundant variables, making it easier to understand and analyze the effects of the remaining predictors on the target variable.

- **Syntax and Example**
- drop_first: bool, default False
- This parameter specifies whether to generate n−1n−1 dummy variables from nn categorical levels by removing the first level.

2nd question contd..

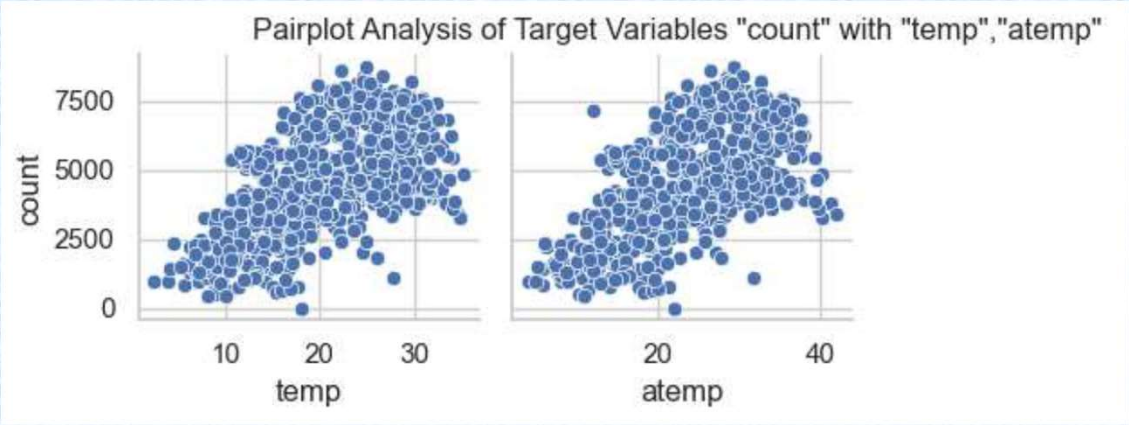
Example:

- Consider a categorical column with three values: A, B, and C. If we create dummy variables:
- A is the first category and can be inferred if both B and C are absent (00).
- By dropping the first column (A), we avoid redundancy.
- The resulting dummy variables will be:
- B: 10 (presence of B)
- C: 01 (presence of C)
- A is represented implicitly as 00 (absence of both B and C).

- This approach optimizes the model and enhances clarity in the interpretation of results.

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

- In the pair-plot analysis, **"temp" (temperature)** and **"atemp" (feeling temperature)** exhibit the highest correlation with the target variable **"count" (bike rental count)**. This strong positive correlation implies that as temperatures rise, there is a notable increase in bike rentals. Warmer conditions likely make biking more appealing, contributing to the higher demand observed.



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

- To validate the assumptions of the Linear Regression model, I used the following methods:

1. Normality of Error Terms:

1. **Histogram**: I plotted a histogram of the residuals, and the distribution resembled a bell curve, confirming that the residuals were normally distributed.
2. **Q-Q Plot**: I created a Q-Q plot of the residuals. The points followed the 45-degree line, further indicating normal distribution of residuals.

2. Multicollinearity Check:

1. **Variance Inflation Factor (VIF)**: I calculated the VIF for each predictor variable, and all values were less than 10, showing that multicollinearity was not an issue in the model.

3. Linear Relationship Validation:

1. **Residual Plot**: I plotted residuals against the predicted values. The residuals were randomly scattered around zero, confirming that a linear relationship existed between the predictors and the target variable.

4. Homoscedasticity:

1. **Residuals vs. Predicted Plot**: I plotted residuals against predicted values, and the absence of a clear pattern indicated constant variance (homoscedasticity), which supports this assumption.

5. Independence of Residuals:

1. **Durbin-Watson Test**: I calculated the Durbin-Watson statistic, which indicated no significant autocorrelation in the residuals, confirming their independence.

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

- The top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. **Temperature (temp)**: Higher temperatures lead to increased bike usage, making this the most influential factor in driving demand.
2. **Year (year_2019)**: The year 2019 shows a significant increase in bike usage, reflecting a growing trend in the popularity of bike-sharing programs.
3. **Light Snowy Rain Weather (weathersit_light_snow_rain)**: Adverse weather conditions like light snow or rain negatively impact bike demand, as fewer people opt to ride in such conditions.

General Subjective Questions

Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical model that analyzes the linear relationship between a dependent variable and one or more independent variables. It predicts how changes in the independent variables affect the dependent variable, assuming a linear relationship.

Key Elements of Linear Regression:

- **Dependent Variable (Y)**: The target variable we want to predict.
- **Independent Variables (X)**: Variables that influence the dependent variable.

Types of Linear Regression:

1. **Simple Linear Regression**: Models the relationship between a single independent variable (X) and the dependent variable (Y):
   - $Y=\beta0+\beta1X+\epsilon Y$
     - $\beta0$: Intercept (value of Y when X=0)
     - $\beta1$: Slope (change in Y for a one-unit change in X)
     - ϵ: Error term (difference between actual and predicted Y)
1. **Multiple Linear Regression**: Extends to multiple independent variables:
   - $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+$
   - X1,X2,..., Xn: Independent variables
   - $\beta1,\beta2,...\beta n$: Coefficients for each variable

linear regression algorithm contd

Assumptions of Linear Regression:

1. **Linearity**: The relationship between the dependent and independent variables must be linear.
2. **Independence**: The residuals (errors) should be independent.
3. **Homoscedasticity**: The residuals should have constant variance at every level of X.
4. **Normality**: The residuals should follow a normal distribution.

Steps in Linear Regression:

1. **Hypothesis**: Assume a linear relationship between the dependent and independent variables.
2. **Estimating Coefficients**: Use the **least squares method** to estimate the coefficients by minimizing the residual sum of squares (RSS This method minimizes the sum of the squareddifferences between the observed and predicted values of Y.
3. **Fitting the Model**: Fit the linear equation by adjusting the coefficients to minimize the RSS.

- The linear equation is fitted to the data by adjusting the coefficients to minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- where Yi_pred is the predicted value of Yi for the i-th observation.

Explain the Anscombe's quartet in detail. (3 marks)

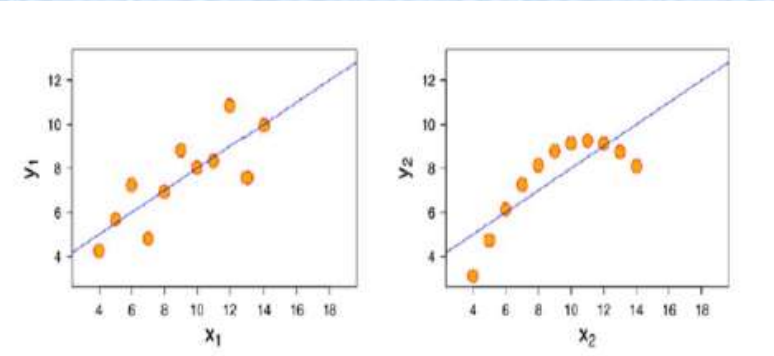
- Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets, each containing eleven pairs of (x, y) values. What makes these datasets notable is that, despite having nearly identical descriptive statistics, they exhibit significant differences when visualized through scatter plots. This demonstrates how statistical summaries can sometimes be misleading without proper visual analysis.

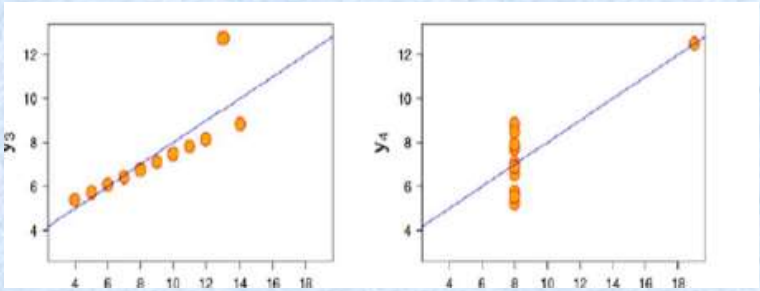
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- **Key Summary Statistics:**
- The mean of x is 9 and the mean of y is 7.50 for each dataset.
- The variance of x is 11, and the variance of y is 4.13 across all datasets.
- The correlation coefficient between x and y is 0.816 in each dataset.

Anscombe's quartet contd..

- **Graphical Interpretations:**
- **Graph 1 (Dataset 1):** Displays a well-fitted linear relationship, ideal for regression modeling.
- **Graph 2 (Dataset 2):** Shows a non-linear distribution, indicating that a linear regression model would not be suitable.
- **Graph 3 (Dataset 3):** Although mostly linear, the presence of an outlier skews the regression result.
- **Graph 4 (Dataset 4):** Features a high correlation coefficient, but this is entirely driven by a single outlier, emphasizing how a single data point can influence analysis.
- Despite having similar statistical summaries, the visualizations reveal distinct patterns and differences in distribution, underscoring the importance of using both visual and statistical tools in data analysis.

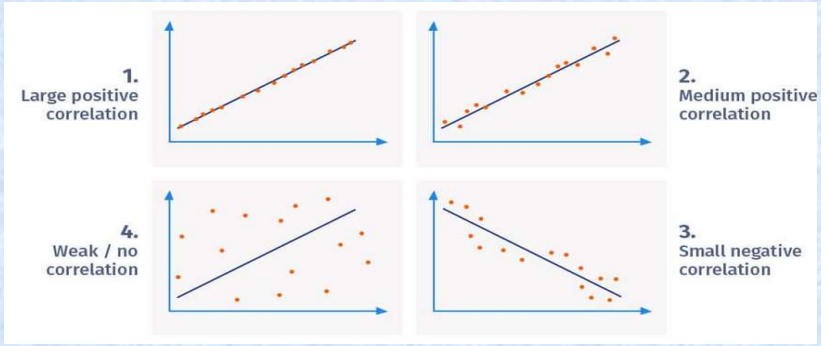




What is Pearson's R? (3 marks)

Pearson's R, also known as the **Pearson correlation coefficient (PCC)**, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where:

- **+1** indicates a perfect positive linear relationship (as one variable increases, the other also increases).
- **-1** indicates a perfect negative linear relationship (as one variable increases, the other decreases).
- **0** means there is no linear relationship between the variables.
- Pearson's R assumes the relationship is linear, meaning changes in one variable result in proportional changes in the other.



What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- **What is Scaling?**
- Scaling is the process of transforming the values of numerical features into a standard range or scale without distorting the differences in the ranges of values. It's a crucial step in data preprocessing, especially for machine learning algorithms, as it ensures that all variables contribute equally to the model.
- **Why is Scaling Performed?**
- Scaling is performed for several reasons:

1. **Preventing Bias in Models**: Many machine learning algorithms, especially those involving distance measures (e.g., K-Nearest Neighbors, SVMs, Gradient Descent), are sensitive to the magnitudes of features. Unscaled features with larger ranges can dominate those with smaller ranges, introducing bias in model performance.
2. **Improving Model Convergence**: In models like gradient descent-based algorithms (e.g., linear regression, logistic regression), scaling helps the algorithms converge faster by optimizing weight updates consistently.
3. **Ensuring Comparability of Features**: Features measured in different units (e.g., height in meters and weight in kilograms) need scaling to make them comparable, so the model learns appropriately from all features.

Difference Between Normalized Scaling and Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
   - **Definition**: Normalization transforms data to a fixed range, typically between 0 and 1. It rescales the data so that all values fall between the minimum and maximum values of the original data.
   - **Formula**:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Where

- Xmin is the minimum value of the feature
- Xmax is the maximum value of the feature

   - **Definition**: Standardization centers the data around a mean of 0 with a standard deviation of 1. It transforms the data by subtracting the mean and dividing by the standard deviation, essentially converting the data to a standard normal distribution.

   - **Use Case**: Normalization is suitable for cases where you know the distribution of data doesn't follow a normal (Gaussian) distribution or for algorithms like K-Nearest Neighbors (KNN) and Neural Networks, which depend on the distance between points.

1. **Standardized Scaling (Z-Score Normalization)**:
   - **Formula**:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Where:

- μ is the mean of the feature
- σ is the standard deviation of the feature

   - **Use Case**: Standardization is preferred when the data follows a normal distribution or for algorithms such as linear regression, logistic regression, and SVM, which assume normality of the feature distributions.

You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity between two or more independent variables in a regression model. This happens when one predictor variable is an exact linear combination of one or more other predictors.

Why Does This Happen?

$$VIF_i = \frac{1}{1 - R_i^2}$$

- The formula for VIF is:
- Where:
- Ri2R^2_iRi2 is the coefficient of determination for the regression of the ithi independent variable on all other predictors.
- When perfect multicollinearity exists, the R2R^2R2 value becomes 1, meaning that the independent variable can be completely explained by a linear combination of the other variables. In this scenario, the denominator of the VIF formula becomes 0, resulting in an infinite VIF.

Causes of Infinite VIF:

1. **Duplicate Variables**: If the same variable or a copy of a variable is used twice in the regression, perfect multicollinearity occurs.
2. **Linear Combinations**: When one variable is an exact combination of other variables (e.g., $X3=2\times X1+X2$)
3. **Dummy Variable Trap**: If all categories of a categorical variable are included as dummy variables without dropping one, this leads to perfect multicollinearity.

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 a dataset follows a particular theoretical distribution, most commonly the normal distribution. It does this by comparing the quantiles of the sample data against the quantiles of the specified distribution. Here's a detailed explanation of the use and importance of a Q-Q plot in the context of linear regression:

- **1. Definition and Construction**
- **Definition**: A Q-Q plot displays the quantiles of a dataset plotted against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points will approximately lie along a straight line (often the 45-degree line).
- **Construction**:
  - Calculate the quantiles of the observed data.
  - Calculate the corresponding quantiles of the theoretical distribution (e.g., normal distribution).
  - Plot the quantiles of the sample data on the y-axis and the quantiles of the theoretical distribution on the x-axis.
- **2. Use in Linear Regression**
- **Assumption Checking**: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of the residuals allows us to visually assess this assumption.
- **Detecting Deviations**: If the Q-Q plot shows that points deviate significantly from the reference line, this indicates that the residuals may not follow a normal distribution. Common deviations include:
  - **Heavy Tails**: Points that fall above the line at the ends suggest heavier tails than the normal distribution (e.g., presence of outliers).
  - **Light Tails**: Points falling below the line indicate lighter tails, suggesting that the variability in the data may not be adequately captured.
- **Improving Model Fit**: Identifying non-normality in the residuals may prompt further investigation, such as transforming the dependent variable or reconsidering the choice of predictor variables to improve model fit.

Q-Q plot continued

- **3. Importance of Q-Q Plots in Linear Regression**
- **Model Validation**: By confirming the assumption of normality in the residuals, a Q-Q plot aids in validating the reliability of the regression model. If this assumption is violated, it could lead to incorrect inferences about coefficients and significance tests.
- **Guiding Decision-Making**: Understanding the distribution of residuals helps in making informed decisions regarding model diagnostics, potential transformations, or the need for alternative modeling approaches (e.g., using generalized linear models for non-normal distributions).
- **Visual Insight**: Q-Q plots provide a quick and intuitive visual insight into how well the residuals fit a normal distribution, complementing quantitative tests for normality (e.g., Shapiro-Wilk test) and enhancing the overall analysis process.
- **Conclusion**
- In summary, a Q-Q plot is an essential diagnostic tool in linear regression for assessing the normality of residuals, which is critical for validating the assumptions underlying the regression analysis. Its use aids in ensuring the accuracy and reliability of the model's predictions and in guiding further model refinement if needed.