**Name:** Anshu

**Email ID:** [Pathakanshu576@gmail.com](mailto:Pathakanshu576@gmail.com)

**From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**  
  
Based on the analysis of the categorical variables in the dataset, it is clear that they influence the bike rental demand (cnt) in different ways. The **season** variable shows noticeable variation, with higher demand typically observed during summer and fall, likely due to more favorable weather conditions. The **year** feature also has a positive effect, as the second year (2019) records a general increase in bike usage, suggesting growing popularity or adoption of the service. Monthly patterns (mnth) reveal seasonal trends, while **working days** are associated with higher demand, indicating a strong link between weekday commuting and bike rentals. On the other hand, **holidays** tend to show a slight decrease in usage, possibly because fewer people commute. The **weather situation** significantly impacts demand, with clear weather encouraging more rentals, whereas bad weather like rain or snow leads to a drop in usage. Lastly, the **weekday** variable shows relatively consistent usage across the week, suggesting less variation based solely on the day. Overall, these categorical features provide useful insights into user behavior and contribute meaningfully to predicting bike demand.

**Why is it important to use drop\_first=True during dummy variable creation?**

Using drop\_first=True, when creating dummy variables is important to avoid the issue of **multicollinearity**, specifically the **dummy variable trap**. When all categories of a categorical variable are converted into dummy columns, one of them becomes redundant because it can be perfectly predicted from the others. This redundancy introduces multicollinearity into the model, which can distort the interpretation of coefficients and affect the stability of the model’s estimates. By dropping the first category, we eliminate this linear dependency without losing any meaningful information, allowing the model to remain interpretable and mathematically sound. The omitted category then serves as a baseline for comparison in the model.

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

By examining the pair-plot of the numerical variables, it is evident that the **‘registered’** variable shows the strongest positive relationship with the target variable **‘cnt’** (total bike count). The data points between these two variables form a tightly clustered upward trend, indicating a strong linear association. This suggests that the number of registered users is a major contributor to the overall bike usage, likely because these users are consistent commuters or frequent riders. While other variables like temperature (temp) and casual users (casual) also show positive correlations with cnt, none are as pronounced as the relationship observed with registered.

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

After building the linear regression model on the training set, the assumptions of linear regression were validated through both statistical metrics and visual evaluation. The **R-squared score** provided an initial indication of how well the model explained the variance in the target variable, suggesting a strong linear fit. To further validate the model, a **scatter plot of actual versus predicted values** was used. In this plot, most points were closely aligned with the diagonal line, indicating that the predictions were accurate and the errors were randomly distributed — a key assumption of linear regression. Additionally, the low values of **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** implied that the residuals (errors) were minimal, which supports the assumptions of homoscedasticity and linearity. While not explicitly shown in this analysis, a residual plot would further help in examining patterns or non-linearity in the residuals to confirm these assumptions more robustly.

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

Based on the final linear regression model, the top three features that contribute most significantly to explaining the demand for shared bikes are **‘registered’**, **‘temperature (temp)’**, and **‘feels-like temperature (atemp)’**. Among these, ‘registered’ has the strongest influence, indicating that a large portion of daily bike usage comes from users who are subscribed or frequently use the service. Both temperature-related features also play a critical role, as higher temperatures typically encourage more outdoor activity, making cycling more favorable. These features demonstrate a clear and consistent positive correlation with the target variable cnt, highlighting their importance in predicting overall bike demand.

--------------------------------------------------------General Subjective Questions-------------------------------------------

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

Linear regression is a foundational algorithm in statistical modeling and machine learning that aims to establish a relationship between one or more independent variables and a dependent variable. The goal is to model the dependent variable as a linear combination of the predictors.

The core idea behind linear regression is to find the best-fitting line through the data, which minimizes the difference between the observed values and the values predicted by the model. This difference is known as the residual. The line is determined by two parameters: the slope (or coefficient) and the intercept.

The equation for a simple linear regression model is:

y = β0​+β1x + E

Where:

* y is the dependent variable (target),
* β0​ is the intercept (the value of y when x=0),
* β1​ is the slope (coefficient) representing the change in y for a one-unit change in x,
* x is the independent variable (predictor),
* E is the error term, accounting for the difference between the observed and predicted values.

In the case of multiple linear regression, the model involves multiple predictors and is represented as:

y = β0​+β1x1+β2x2+………… + βnxn + E

Steps in Linear Regression:

1. Data Preparation: The data is preprocessed, which includes handling missing values, converting categorical variables into numerical representations, and scaling numerical features if needed.
2. Model Fitting: The model is trained using a dataset where the relationship between the predictors (independent variables) and the target (dependent variable) is learned by finding the coefficients that minimize the cost function. The most common cost function used is the mean squared error (MSE), which penalizes large errors.
3. Model Evaluation: The model's performance is evaluated using metrics like:
   * R-squared (R²): Indicates how much variance in the target variable is explained by the independent variables. A higher R² indicates a better model fit.
   * Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): Measure the average squared difference between the predicted and actual values. Lower values indicate better predictive accuracy.
   * Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values, which gives an intuitive measure of prediction accuracy.
4. Assumption Validation: Linear regression assumes that:
   * There is a linear relationship between predictors and the target.
   * Residuals (errors) are normally distributed and have constant variance (homoscedasticity).
   * There is no multicollinearity (predictors should not be highly correlated with each other).
   * Independence of observations.

**Explain the Anscombe’s quartet in detail.**

Anscombe's quartet is a famous collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but display very different data patterns when visualized. The purpose of Anscombe’s quartet is to demonstrate the importance of visualizing data before making inferences, especially when relying on summary statistics alone.

The quartet consists of four datasets, each with 11 data points. Despite having the same basic statistical properties, such as the same mean, variance, and correlation, they exhibit very different relationships between the variables when plotted. The four datasets are:

1. Dataset I: This dataset shows a strong linear relationship between the variables. A simple linear regression model would fit well and provide meaningful predictions.
2. Dataset II: Although the summary statistics are similar, this dataset contains one outlier that has a significant effect on the relationship between the variables. A regression model may still fit well for most data points, but it will be overly influenced by the outlier.
3. Dataset III: This dataset shows a perfect quadratic relationship, where the variables follow a curve rather than a straight line. A linear model would poorly capture the relationship, despite having similar summary statistics.
4. Dataset IV: This dataset consists of a vertical line, indicating no relationship between the variables, even though summary statistics might suggest otherwise. A regression line would be essentially meaningless here, and further analysis is needed to identify this pattern.

Anscombe’s quartet emphasizes the importance of visualizing data to understand underlying patterns and relationships that summary statistics alone cannot reveal. It serves as a cautionary tale about relying too heavily on metrics like mean and correlation without exploring the data graphically.

**What is Pearson’s R?**

Pearson’s R (also called the Pearson correlation coefficient) is a measure that quantifies 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 proportionally),
* -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally),
* 0 suggests no linear relationship between the variables.

The formula for Pearson’s R is:

A black and white math equation

Description automatically generated

Where:

* Xi and Yi​ are the individual data points for variables X and Y,
* Xˉ and Yˉ are the means of the variables X and Y.

Pearson’s R is widely used in statistics to assess the strength and direction of the relationship between two variables.

**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 of your data to a common scale without distorting differences in the ranges of values. This step is important because many machine learning algorithms, especially those that involve distance metrics like k-NN or gradient-based optimization techniques like linear regression, are sensitive to the magnitude of the features. If features vary widely, the model may place undue importance on features with larger values, which can affect the quality of predictions and training.

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

1. Normalized Scaling (Min-Max Scaling): This method rescales the data into a fixed range, typically [0, 1]. It is useful when the distribution of the data is not Gaussian. The formula for normalization is:

A black and white math equation

Description automatically generated

This ensures that all features are on the same scale, but it can be sensitive to outliers.

1. Standardized Scaling (Z-score Scaling): This method rescales the data so that the mean of each feature is 0, and the standard deviation is 1. It is useful when the data follows a Gaussian (normal) distribution. The formula for standardization is:

A math equation with black text

Description automatically generated with medium confidence

where μ is the mean and σ is the standard deviation of the feature. Standardization is less sensitive to outliers compared to normalization and is typically preferred for algorithms like linear regression.

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

In linear regression, the Variance Inflation Factor (VIF) is used to quantify how much the variance of the estimated regression coefficients increases due to multicollinearity. A VIF value of infinity typically occurs when there is perfect multicollinearity between two or more predictors. This means that one predictor variable can be perfectly predicted from another, causing the model to be unable to distinguish their independent effects. In such cases, the estimated coefficients become unstable, leading to extremely high VIF values, often approaching infinity. To avoid this issue, it’s essential to identify and remove highly correlated features before fitting a linear regression model.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular distribution, typically a normal distribution. In a Q-Q plot, the quantiles of the data are plotted against the quantiles of a chosen theoretical distribution. If the data follows the theoretical distribution, the points will lie approximately along a straight line.

In the context of linear regression, a Q-Q plot is useful for validating the normality assumption of the residuals (errors). One of the key assumptions in linear regression is that the residuals should be normally distributed. By plotting the residuals in a Q-Q plot, we can visually inspect if the residuals deviate from normality. If the points form a straight line, it suggests that the residuals are normally distributed, which supports the validity of the linear regression model. Conversely, significant deviations from the line could indicate problems like non-linearity, heteroscedasticity, or outliers that may require further analysis or model adjustments.