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:

There are noticeable differences in bike rental counts across seasons. It appears that some seasons (likely spring and summer) have higher median rental counts compared to others. This suggests seasonality has a significant impact on bike rentals.

Year (yr):

There seems to be a difference between the two years in terms of rental counts, with one year having a higher median count. This could indicate an overall increase in the usage of bike rentals from one year to the next.

Month (mnth):

The rental counts vary significantly across months, reflecting seasonal trends. Some months (likely warmer ones) show higher rental counts.

Holiday:

There's a difference in rental counts between holidays and non-holidays, with non-holidays having higher median rental counts. This could be due to more regular commuters or recreational users on non-holidays.

Weekday:

The impact of weekdays on rental counts seems relatively uniform, suggesting that the day of the week might not have a significant impact on overall rentals. However, some variation exists, possibly reflecting commuting patterns.

Workingday:

Working days and non-working days show some differences in rental counts. This could be related to commuting patterns, where working days might see more regular commuting usage.

Weather Situation (weathersit):

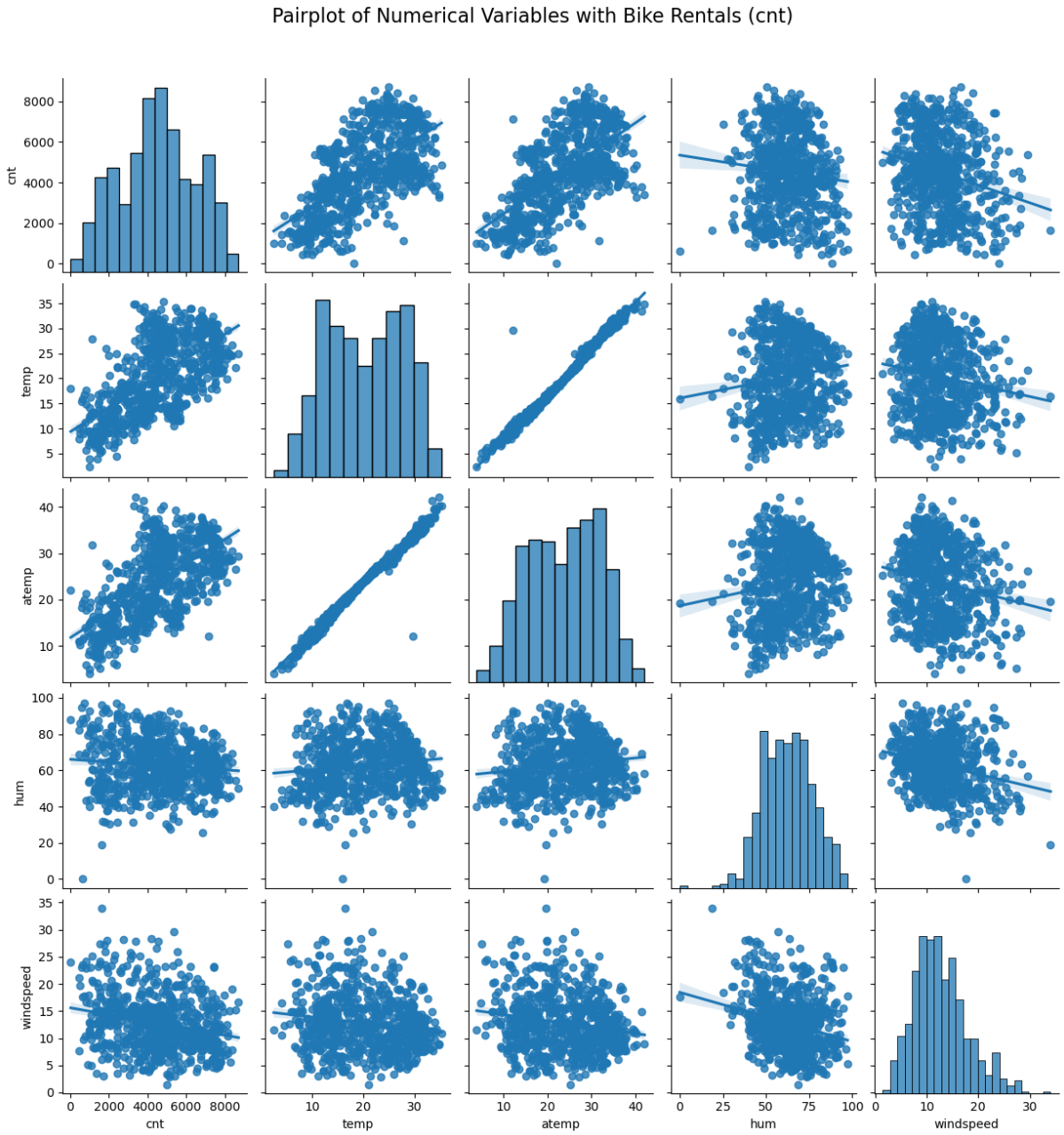
The weather situation has a clear impact on bike rentals. Better weather conditions (represented by lower category numbers) correspond to higher median rental counts. Adverse weather conditions significantly reduce bike rental usage.

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

In summary, using drop_first=True when creating dummy variables is a best practice that helps to avoid multicollinearity, simplifies the model, provides a clear reference category for interpretation, and can aid in preventing overfitting.

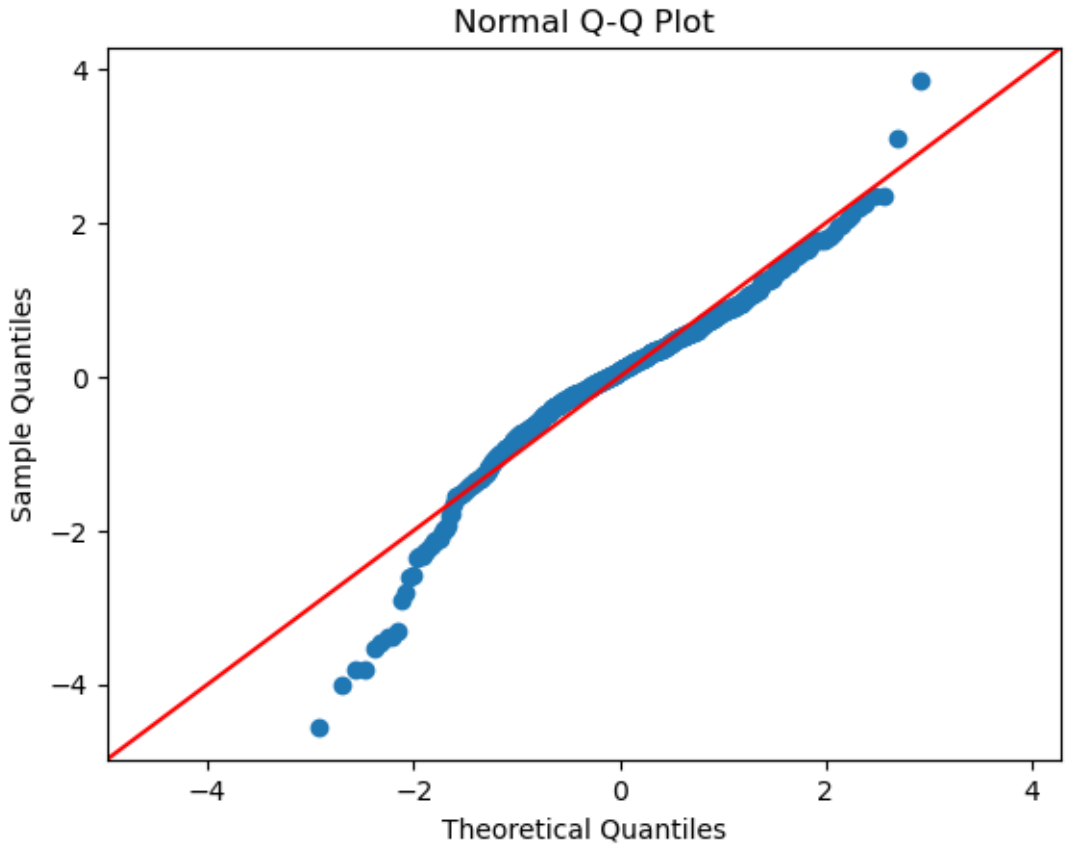
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The scatter plots for both temp vs cnt and atemp vs cnt show a clear upward trend, indicating that as temperature increases, the count of bike rentals also increases. This trend is linear, which is characteristic of variables that are correlated.



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

If the data were perfectly normally distributed, the points would lie exactly on the red line. However, the deviations from the line at both ends suggest that the data distribution has lighter tails than the normal distribution on the lower end and heavier tails on the upper end. This could indicate potential outliers or that the data are not perfectly normally distributed.



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

yr: Coefficient = 989.16333236

Represents the year. This large positive coefficient suggests a strong year-on-year increase in bike demand, possibly due to increased popularity or expansion of service.

temp: Coefficient = 1001.50540449

Represents the temperature. The positive coefficient indicates that higher temperatures lead to an increase in bike rentals, which is intuitive as more people are likely to ride bikes in warmer weather.

season_Winter: Coefficient = 376.13292269

Indicates the season. Despite being a seasonal coefficient, it's positive, suggesting that the winter season has a significant positive impact on bike rentals compared to the baseline season (probably omitted due to drop_first=True, likely spring). This may require context understanding; for instance, if winter refers to mild winters rather than harsh ones, or if there's increased bike usage during winter holidays.

General Subjective Question

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Here's a detailed explanation of how the linear regression algorithm works:

Formulation of the Model:

The simplest form of linear regression is a straight-line fit to data, which is known as simple linear regression. It assumes a linear relationship between the dependent variable y and a single independent variable x:

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

β0 is the intercept.

β1 is the slope of the line (the effect of x on y.

ϵ is the error term which accounts for the variability in y that cannot be explained by the linear relationship with x.

$$y=\beta0+\beta1x1+\beta2x2 +...+\beta nxn+\epsilon$$

β0 is still the intercept.

β1,β2,....βn are the coefficients for each independent variable

ϵ is the error term.

Assumptions:

Linear regression makes several key assumptions about the data:

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

Independence: The residuals (errors) are independent.

Homoscedasticity: The residuals have constant variance at every level of the independent variables.

Normality: The residuals of the model are normally distributed.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.

3.What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient or Pearson's product-moment correlation coefficient,

is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1

is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

It is widely used in the sciences as a measure of the degree of linear dependence between two variables.

Formula to calculate Pearson's R:

$r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / \sqrt{(\Sigma(xi - \bar{x})^2 \Sigma(yi - \bar{y})^2)}$

Where:

- xi and yi are the individual sample points indexed with i.

- $\bar{x}$ is the mean of the x values and $\bar{y}$ is the mean of the y values.

- Σ is the summation symbol, used to sum up all the individual components.

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

Scaling is the process of transforming data to fit within a specific scale. It is essential in data preprocessing for algorithms sensitive to the magnitude of data, such as SVMs and KNN.

Reasons to Perform Scaling:

1. Equal Contribution: Ensures each feature contributes proportionally to the final result in distance-based algorithms.

2. Gradient Descent: Helps in faster convergence in optimization algorithms.

3. Avoid Skewing: Prevents models from misinterpreting the scales of the data.

4. Learning Efficiency: Improves the performance of algorithms by normalizing the data distribution.

Normalization vs. Standardization:

Normalization, or Min-Max scaling, rescales the feature to a range of [0, 1] or [-1, 1] using the formula:

x' = (x - min(x)) / (max(x) - min(x))

Standardization, or Z-score normalization, rescales data to have a mean of 0 and standard deviation of 1, following the formula:

x' = (x - mean) / std_deviation

Normalization is useful for algorithms that require data within a bounded range. Standardization is crucial for features assumed to be normally distributed and is less sensitive to outliers.

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 of the amount of multicollinearity in a set of multiple regression variables. It indicates how much the variance of an estimated regression coefficient is increased because of collinearity.

The formula for VIF is:

$VIF_i = 1 / (1 - R^2_i)$

where R^2_i is the coefficient of determination of a regression of predictor i on all the other predictors.

A VIF becomes infinite when the independent variable i is perfectly collinear with at least one other independent variable. This means that the variable can be exactly linearly predicted from the others with no error.

Scenarios for Infinite VIF:

1. Duplicate Variables: Including the same variable twice in the regression.

2. Derived Variables: One variable is a combination of others (e.g., height in meters and height in centimeters).

3. Complete Separation: In logistic regression, an independent variable can perfectly separate the outcome variable.

Infinite VIF is a sign that the model needs reassessment. Solutions might include removing or combining features, or using regularization techniques like ridge or lasso regression.

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 the distribution of a dataset against a theoretical distribution, typically the normal distribution. It is particularly useful in linear regression for assessing the normality of residuals.

How a Q-Q Plot Works:

- It plots two probability distributions against each other by their quantiles.

- The x-axis represents the quantiles of the dataset, and the y-axis shows the quantiles from a theoretical distribution.

- If the dataset is similar to the theoretical distribution, points on the Q-Q plot will align closely with a straight line.

Importance in Linear Regression:

1. Checking Normality of Residuals:

  - Linear regression assumes that residuals are normally distributed. The Q-Q plot is a visual tool to check this.

  - Points lying on a straight line in the Q-Q plot indicate normal distribution of residuals.

2. Identifying Skewness and Outliers:

  - Deviations from the line in a Q-Q plot indicate skewness or outliers in the data, informing about potential issues in the regression model.

3. Improving Model Accuracy:

  - By identifying non-normality or outliers, the Q-Q plot guides adjustments to the model or data for better accuracy and reliability.

4. Comparative Analysis:

  - The Q-Q plot allows for an intuitive comparison of the sample data's distribution to a theoretical distribution.

A Q-Q plot is a vital diagnostic tool in linear regression. It helps in validating the assumptions of the model and guiding necessary transformations or adjustments.