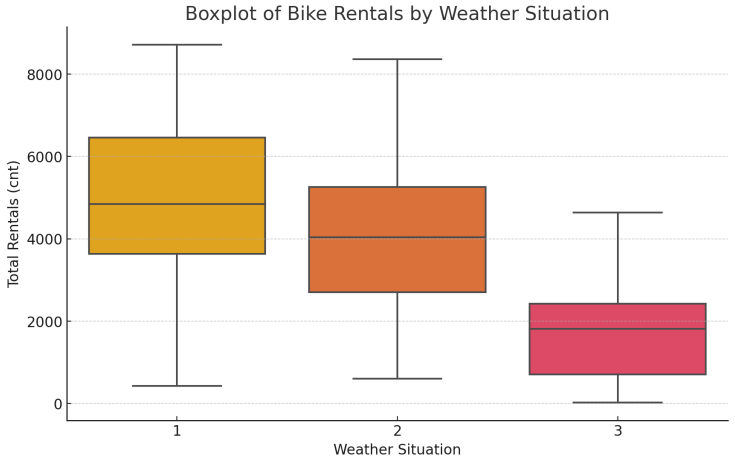
Linear Regression Assignment

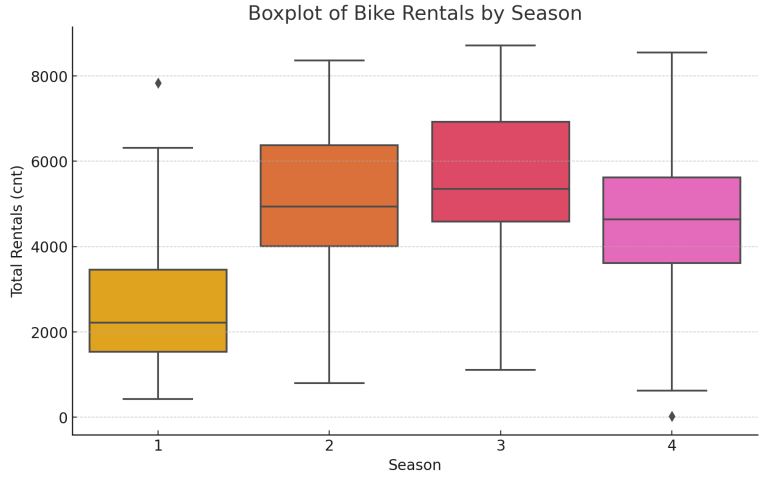
By Adila Parveen

Subjective Questions & Answers

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

1. To analyze the effect of categorical variables like weather and season from the given dataset on dependent variable (cnt), we can perform exploratory data analysis (EDA) using box plot
2. The boxplot of bike rentals by weather situation and season is as follows:





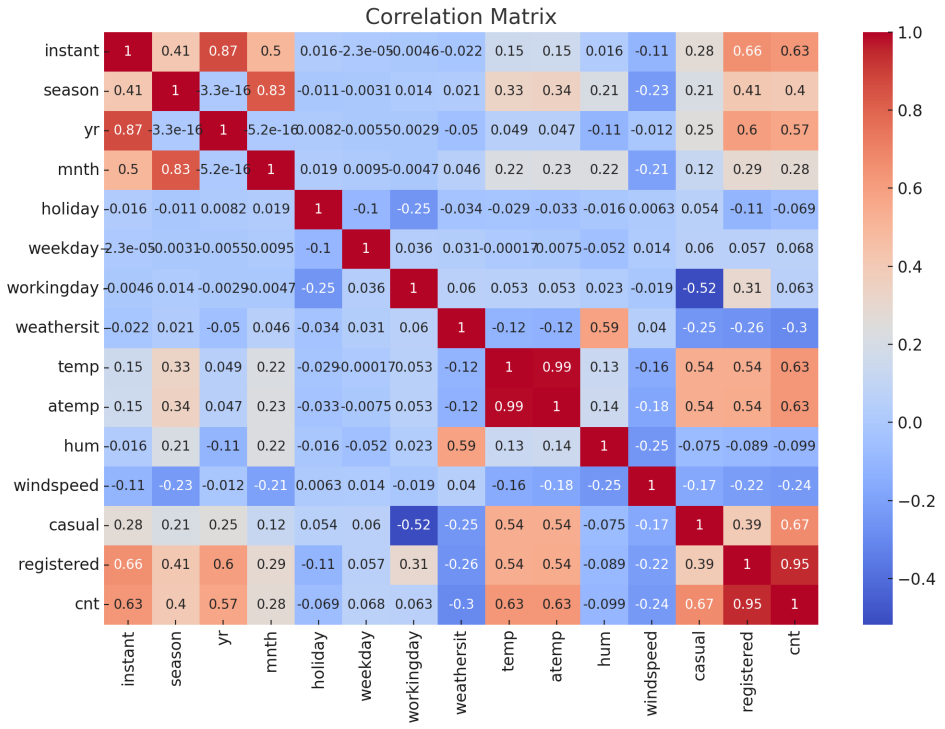
3. The boxplot shows, the bike rentals vary across seasons, with higher rentals in summer and fall and lower rentals in winter. It also shows the bike rentals were higher during clear weather conditions and comparatively lower to misty and rainy conditions.

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

- When creating dummy variables from categorical variables, the drop_first=True parameter is used to prevent multicollinearity issues in regression models. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. This can cause problems such as unstable coefficients estimates, inflated standard errors, and difficulties in interpreting the model.
- By setting drop_first=True, one level of each categorical variable is dropped during the creation of dummy variables. This means that if a categorical variable has kk levels, only k−1k−1 dummy variables are created. The dropped level becomes the reference category against which the other categories are compared.
- Dropping one level helps to avoid perfect multicollinearity because if all levels of a categorical variable were included as dummy variables, their sum would always be 1 for each observation, which is redundant information. Dropping one level removes this redundancy and prevents multicollinearity issues.
- Additionally, dropping one level also aids in the interpretation of coefficients in regression models. The coefficients of the dummy variables represent the difference in the outcome variable between each category and the reference category.
- Overall, using drop_first=True during dummy variable creation helps to ensure the stability and interpretability of regression models when dealing with categorical variables.

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

- **atemp** with the value 0.9916961786905638



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

- Using box plot
- Using correlation matrix
- Using heat maps

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

- weathersit: Weather situation (1: Clear, 2: Mist, 3: Light Snow/Rain).
- season: Season (1: winter, 2: spring, 3: summer, 4: fall).
- atemp: Normalized feeling temperature in Celsius.

General Questions & Answers

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable.

- Assumptions:
  - Linear relationship: Linear regression assumes that there is a linear relationship between the independent variables XX and the dependent variable YY. This means that the change in YY is proportional to the change in XX.
  - Independence: The observations should be independent of each other.
  - Homoscedasticity: The variance of the residuals (the differences between the observed and predicted values) should be constant across all levels of the independent variables.
  - Normality: The residuals should be normally distributed.
  - No multicollinearity: The independent variables should be linearly independent of each other (i.e., they should not be highly correlated).
- Model Representation:
  - Linear regression represents the relationship between the independent variables XX and the dependent variable YY as a linear equation: $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\varepsilon Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\varepsilon$
  - YY is the dependent variable.
  - $X1,X2,\ldots,XnX1,X2,\ldots,Xn$ are the independent variables.
  - $\beta0,\beta1,\ldots,\beta n\beta0,\beta1,\ldots,\beta n$ are the coefficients (also known as slopes) representing the effect of each independent variable on the dependent variable.
  - $\varepsilon\varepsilon$ is the error term, representing the difference between the observed and predicted values that cannot be explained by the independent variables.
- Continued…

- Objective:
  - The objective of linear regression is to estimate the coefficients β0,β1,…,βnβ0,β1,…,βn that minimize the sum of squared residuals (SSR), which is the sum of the squared differences between the observed and predicted values: SSR=∑i=1n(Yi−Y^i)2SSR=∑i=1n(Yi−Y^i)2
  - Y^iY^i represents the predicted value of YiYi based on the linear regression model.
- Estimation:
  - The coefficients β0,β1,…,βnβ0,β1,…,βn are estimated using the method of least squares, which minimizes the SSR.
  - The ordinary least squares (OLS) method finds the values of the coefficients that minimize the sum of squared differences between the observed and predicted values.
- Model Evaluation:
  - Linear regression models are evaluated based on various metrics such as R2R2 (coefficient of determination), mean squared error (MSE), root mean squared error (RMSE), and adjusted R2R2.
  - R2R2 measures the proportion of the variance in the dependent variable that is explained by the independent variables.
  - Lower values of MSE and RMSE indicate better model performance.
- Predictions:
  - Once the coefficients are estimated, the linear regression model can be used to make predictions on new data by plugging the values of the independent variables into the linear equation.

Linear regression is widely used for prediction and inference in various fields such as economics, finance, social sciences, and engineering, where there is a need to understand and quantify the relationship between variables.

2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet is a set of four datasets that have nearly identical statistical properties yet appear very different when graphed. These datasets were constructed by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical exploration and visualization in data analysis.
- Here's a detailed explanation of Anscombe's quartet:
- Construction:
  - Anscombe created four datasets, each consisting of 11 (x, y) pairs, to illustrate the limitations of relying solely on summary statistics (such as means, variances, and correlations) to understand data.
  - Despite having different distributions, relationships, and variances, the datasets share nearly identical statistical properties such as means, variances, correlations, and regression lines.
- Properties:
  - Each dataset consists of 11 pairs of xx and yy values.
  - All datasets have the same mean and variance for both xx and yy.
  - All datasets have the same correlation coefficient (rr) between xx and yy.
  - All datasets have the same linear regression line (y=ax+by=ax+b).
  - Despite these similarities in summary statistics, the datasets exhibit vastly different patterns when graphed.
- Graphical Exploration:
  - When plotted, the four datasets reveal different relationships between xx and yy.
  - Dataset I: Shows a linear relationship with some variability around the line.
  - Dataset II: Shows a non-linear relationship, where yy increases with xx but not in a strictly linear fashion.
  - Dataset III: Shows a linear relationship, but with one outlier that significantly influences the regression line.
  - Dataset IV: Shows no clear relationship between xx and yy, except for a single outlier that distorts the linear regression line.

Continued…

- Importance:
  - Anscombe's quartet highlights the importance of visualizing data to understand its underlying structure.
  - It demonstrates that summary statistics alone may not provide a complete picture of the data and can sometimes be misleading.
  - The quartet serves as a cautionary example against relying solely on numerical summaries and emphasizes the need for exploratory data analysis (EDA) and data visualization techniques.
- Educational Tool:
  - Anscombe's quartet is frequently used in statistics education to teach the importance of graphical exploration and to illustrate concepts such as the influence of outliers, the limitations of correlation coefficients, and the impact of nonlinear relationships on regression analysis.
- In summary, Anscombe's quartet is a set of four datasets that share identical statistical properties but exhibit vastly different patterns when graphed. It underscores the importance of visualizing data and conducting exploratory data analysis to gain insights into the underlying relationships within the data.

3. What is Pearson’s R?

- Pearson's correlation coefficient, often denoted by rr, is a measure of the linear relationship between two continuous variables. It quantifies the degree to which two variables are linearly related. Pearson's rr ranges from -1 to 1, where:
- r=1r=1: Indicates a perfect positive linear relationship. As one variable increases, the other variable also increases proportionally.
- r=−1r=−1: Indicates a perfect negative linear relationship. As one variable increases, the other variable decreases proportionally.
- r=0r=0: Indicates no linear relationship between the two variables.
- Mathematically, Pearson's correlation coefficient is calculated as:
- $r=\sum(Xi-\bar{X})(Yi-\bar{Y})\sum(Xi-\bar{X})2\cdot\sum(Yi-\bar{Y})2r=\sum(Xi-\bar{X})2\cdot\sum(Yi-\bar{Y})2\sum(Xi-\bar{X})(Yi-\bar{Y})$
- Where:
- XiXi and YiYi are the individual data points.
- $\bar{X}\bar{X}$ and $\bar{Y}\bar{Y}$ are the means of the XX and YY variables, respectively.
- Pearson's rr can be used to assess the strength and direction of the linear relationship between two variables. However, it assumes that the relationship between the variables is linear and that the data are approximately normally distributed. It may not capture non-linear relationships or relationships influenced by outliers.

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 such that it falls within a specific range. It is commonly performed in data preprocessing before applying certain machine learning algorithms.
- Here's why scaling is performed:

1. Normalization of Features: Scaling ensures that all features have a similar scale. This is important for algorithms that use distance-based metrics, such as k-nearest neighbors (KNN) and support vector machines (SVM). Without scaling, features with larger magnitudes may dominate the calculation of distances and influence the model's predictions disproportionately.
2. Faster Convergence: Scaling can help algorithms converge more quickly during optimization. Algorithms such as gradient descent converge faster when features are on a similar scale, reducing the number of iterations required to reach the optimal solution.
3. Improved Interpretability: Scaling can improve the interpretability of the model coefficients. For example, in linear regression, scaling ensures that the coefficients represent the change in the target variable corresponding to a one-unit change in the corresponding feature, regardless of the scale of the feature.
4. Stability of Algorithms: Some algorithms are sensitive to the scale of the features. Scaling can improve the stability and numerical robustness of these algorithms, preventing numerical overflow or underflow issues.

Continued…

- **Normalization and standardization are two common scaling techniques**:

1. Normalization (Min-Max Scaling):
    1. Normalization scales the features to a specific range, typically between 0 and 1.
    2. The formula for normalization is: $X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$
    3. This method preserves the original distribution of the data but may be sensitive to outliers.
2. Standardization (Z-score Scaling):
    1. Standardization scales the features to have a mean of 0 and a standard deviation of 1.
    2. The formula for standardization is: $X_{standardized} = \frac{X - \mu}{\sigma}$
    3. This method centers the data around 0 and scales it based on the standard deviation, making it less sensitive to outliers.
    4. Standardization assumes that the data are approximately normally distributed but is less affected by outliers compared to normalization.

- In summary, scaling is performed to ensure that features have a consistent scale, which can improve the performance, stability, and interpretability of machine learning algorithms. Normalization and standardization are two common scaling techniques, each with its own advantages and considerations.

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

- Yes, it's possible for the variance inflation factor (VIF) to become infinite in certain cases. VIF measures the extent of multicollinearity in a regression analysis, specifically how much the variance of the estimated regression coefficients is inflated due to multicollinearity among the predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated with each other.
- When the VIF is calculated, it involves computing the inverse matrix of the correlation matrix among the predictor variables. If the determinant of this correlation matrix is very close to zero, the inverse matrix cannot be computed, resulting in an infinite VIF value.
- This situation typically occurs when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity means that one or more of the predictor variables can be exactly predicted by a linear combination of the other variables. For example, if one predictor variable is an exact linear combination of other predictor variables, it will lead to a perfect correlation between them, causing the correlation matrix to be singular (i.e., having a determinant of zero), which makes it impossible to compute the inverse matrix and thus resulting in infinite VIF values.
- In summary, infinite VIF values occur when there is perfect multicollinearity among the predictor variables, making it impossible to compute the inverse matrix of the correlation matrix. This situation indicates a severe problem with the data that needs to be addressed, such as removing one of the correlated variables or re-evaluating the model specification.

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 assess whether a given sample of data follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the sample data against the quantiles of a theoretical distribution, typically a standard normal distribution (with mean 0 and standard deviation 1).

1. Construction of Q-Q Plot:
    1. The Q-Q plot is constructed by plotting the quantiles of the sample data on the horizontal axis against the quantiles of the theoretical distribution on the vertical axis.
    2. If the sample data follows the theoretical distribution closely, the points on the Q-Q plot will fall approximately along a straight line.
2. Interpretation of Q-Q Plot:
    1. If the points on the Q-Q plot fall along a straight line, it indicates that the sample data follows the theoretical distribution closely. This suggests that the assumptions underlying the statistical tests or models based on that distribution are met.
    2. Deviations from the straight line suggest departures from the assumed distribution. For example, if the points curve upward or downward, it suggests that the sample data has heavier or lighter tails than the theoretical distribution, respectively.
3. Importance in Linear Regression:
    1. In linear regression, Q-Q plots are commonly used to assess the normality of the residuals (the differences between the observed and predicted values).
    2. Linear regression assumes that the residuals are normally distributed with mean 0 and constant variance. Checking the normality of residuals is crucial because violations of this assumption can affect the validity of the regression coefficients and confidence intervals.
    3. By examining the Q-Q plot of the residuals, you can visually assess whether they follow a normal distribution. If the points on the Q-Q plot closely follow a straight line, it suggests that the normality assumption is reasonable. On the other hand, if there are significant deviations from the straight line, it indicates non-normality of the residuals, which may require further investigation or transformation of the data.

- In summary, Q-Q plots are important graphical tools used to assess the goodness-of-fit of a sample data to a theoretical distribution. In linear regression, Q-Q plots are particularly useful for checking the normality assumption of the residuals, which is essential for the validity of regression analysis.

Thank You