Assignment-based Subjective Questions

1. We used Box plot to study their effect on the dependent variable ('cnt') .The inference that We could derive were:

   a. season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

   b. mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

   c. weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

   d. holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

   e. weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

   f. workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

2. Using drop_first=True during dummy variable creation is important for a few reasons:
   a. Avoiding Multicollinearity
   b. Model Efficiency

3. Temp & atemp has the highest correlation with the target variable but due the exact same correlation we had removed atemp from our analysis.

4. To validate the assumptions of linear regression after building the model on the training set, you typically follow these steps:
   a. Linearity
   b. Homoscedasticity
   c. Normality of Residuals

d. No Multicollinearity
e. Model Specification

5. Temp, yr_2019, season_winter are the top 3 features contributing a +ve significance towards explaining the demand of the shared bikes whereas weathersit_Light Rain is the top feature contributing a -ve significance towards explaining the demand of the shared bikes

General Subjective Questions

1. Linear regression is a statistical method used to model the relationship between one or more independent variables (features) and a dependent variable (outcome). The relationship is expressed as:

Y=β0+β1X1+β2X2+…+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+…+βnXn+ϵ

Where YYY is the dependent variable, XXX represents the independent variables, β\betaβ are the coefficients, and ϵ\epsilonϵ is the error term.

Assumptions

Key assumptions include linearity, independence of observations, homoscedasticity (constant variance of residuals), normality of residuals, and no multicollinearity among independent variables.

Model Fitting

The Ordinary Least Squares (OLS) method estimates coefficients by minimizing the sum of squared residuals:

Minimize ∑(Yi−Y^i)2\text{Minimize } \sum (Y_i - \hat{Y}_i)^2Minimize ∑(Yi−Y^i)2

Prediction

Predictions are made using the fitted model:

Y^=β0+β1X1+…+βnXn\hat{Y} = \beta_0 + \beta_1X_1 + \ldots + \beta_nX_nY^=β0+β1X1+…+βnXn

Evaluation Metrics

Performance can be assessed using metrics like R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Limitations

Linear regression is sensitive to outliers and assumes linear relationships, which may not always hold. Extensions like multiple regression, regularization techniques, and polynomial regression can address some limitations.

Overall, linear regression is a foundational tool in data analysis, emphasizing the need to understand its assumptions and applicability.

2. Anscombe's quartet consists of four datasets with identical statistical properties (means, variances, correlations) but vastly different visual patterns. Created by Francis Anscombe in 1973, the datasets include:

- **Dataset I**: Linear relationship.
- **Dataset II**: Parabolic curve.
- **Dataset III**: Horizontal line with an outlier.
- **Dataset IV**: Vertical line with an outlier.

This quartet illustrates the importance of visualizing data before analysis, highlighting how different relationships can yield similar statistics and emphasizing the impact of outliers on statistical measures.

3. Pearson’s R, or Pearson correlation coefficient, quantifies the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 signifies no correlation. Calculated using the covariance of the variables divided by the product of their standard deviations, Pearson's R helps assess the strength and direction of the relationship. It’s widely used in statistics to determine how closely two variables are related, but it only captures linear relationships.

4. Scaling is the process of transforming features in a dataset to ensure they have a similar range or distribution, which is crucial for many machine learning algorithms. Scaling helps improve model convergence, accuracy, and interpretability, particularly for distance-based algorithms like k-nearest neighbors and gradient descent.

Normalized Scaling (Min-Max Scaling) rescales features to a fixed range, typically [0, 1], using the formula:

X′=X−min(X)max(X)−min(X)X' = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}X′=max(X)−min(X)X−min(X)

Standardized Scaling (Z-score Normalization) transforms features to have a mean of 0 and a standard deviation of 1:

X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ

In summary, normalized scaling adjusts the range, while standardized scaling adjusts the distribution, making both techniques suitable for different contexts in data preprocessing.

5. The Variance Inflation Factor (VIF) can be infinite when a predictor variable is perfectly collinear with one or more other predictors in the model. This means that one variable can be expressed as an exact linear combination of others, leading to undefined estimates of variance for that variable. In practical terms, this situation typically arises in cases of multicollinearity, where two or more variables provide redundant information, severely impacting the model's ability to estimate coefficients accurately. To resolve this, you may need to remove or combine correlated predictors.

6. A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, commonly the normal distribution. In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of the theoretical distribution.

In linear regression, a Q-Q plot is essential for assessing the normality of residuals. Since one of the assumptions of linear regression is that residuals should be normally distributed, a Q-Q plot helps visualize this. If the points fall approximately along a straight diagonal line, it indicates that the residuals are normally distributed. Deviations from this line suggest non-normality, which could impact the validity of statistical inferences and model predictions. Thus, the Q-Q plot is a crucial diagnostic tool for ensuring the robustness of linear regression results.