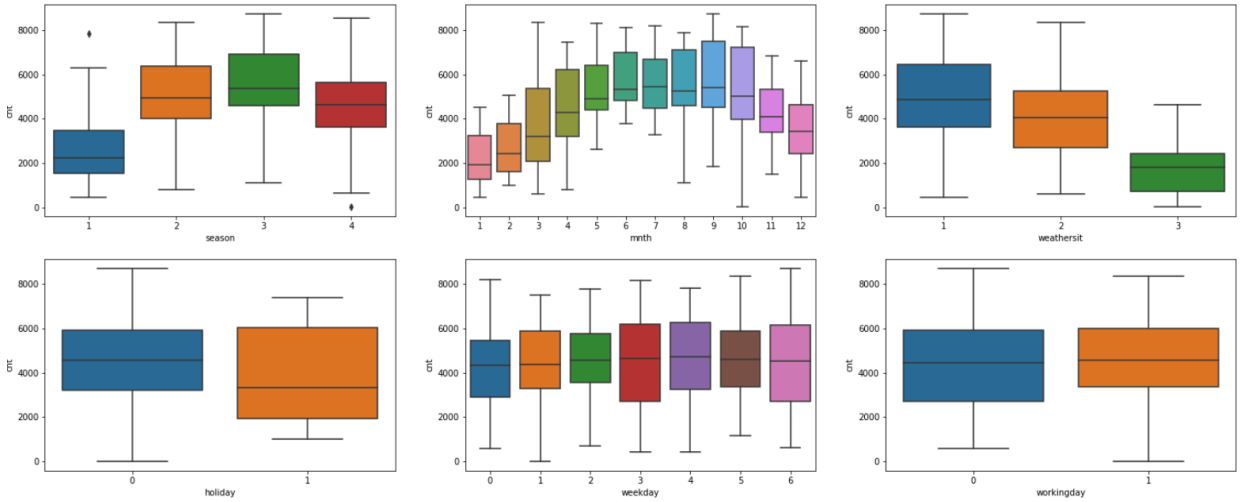
Assignment based Subjective Questions

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

   ANSWER: There are a few categorical variables namely season, mnth, weathersit, holiday, weekday and workingday. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same:



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

   ANSWER: drop_first=True will drop the first dummy variable for each set of dummies created. The intention behind the dummy variable is that for a categorical variable with 'n' levels, we create 'n-1' new columns each indicating whether that level exists or not using a zero or one.

   It is important to use **drop_first=True** as it helps in reducing the extra column created during dummy variable creation. It helps to reduce the correlations created among dummy variables.

   Using drop_first=True during dummy variable creation is important for a few reasons:

   - **Avoiding multicollinearity**: Including dummy variables for all categories of a categorical variable can lead to multicollinearity in statistical models. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This can cause instability in parameter estimates and inflate standard errors, making interpretation of the model less reliable. By

dropping the first level, you remove one redundant variable, which helps alleviate multicollinearity.

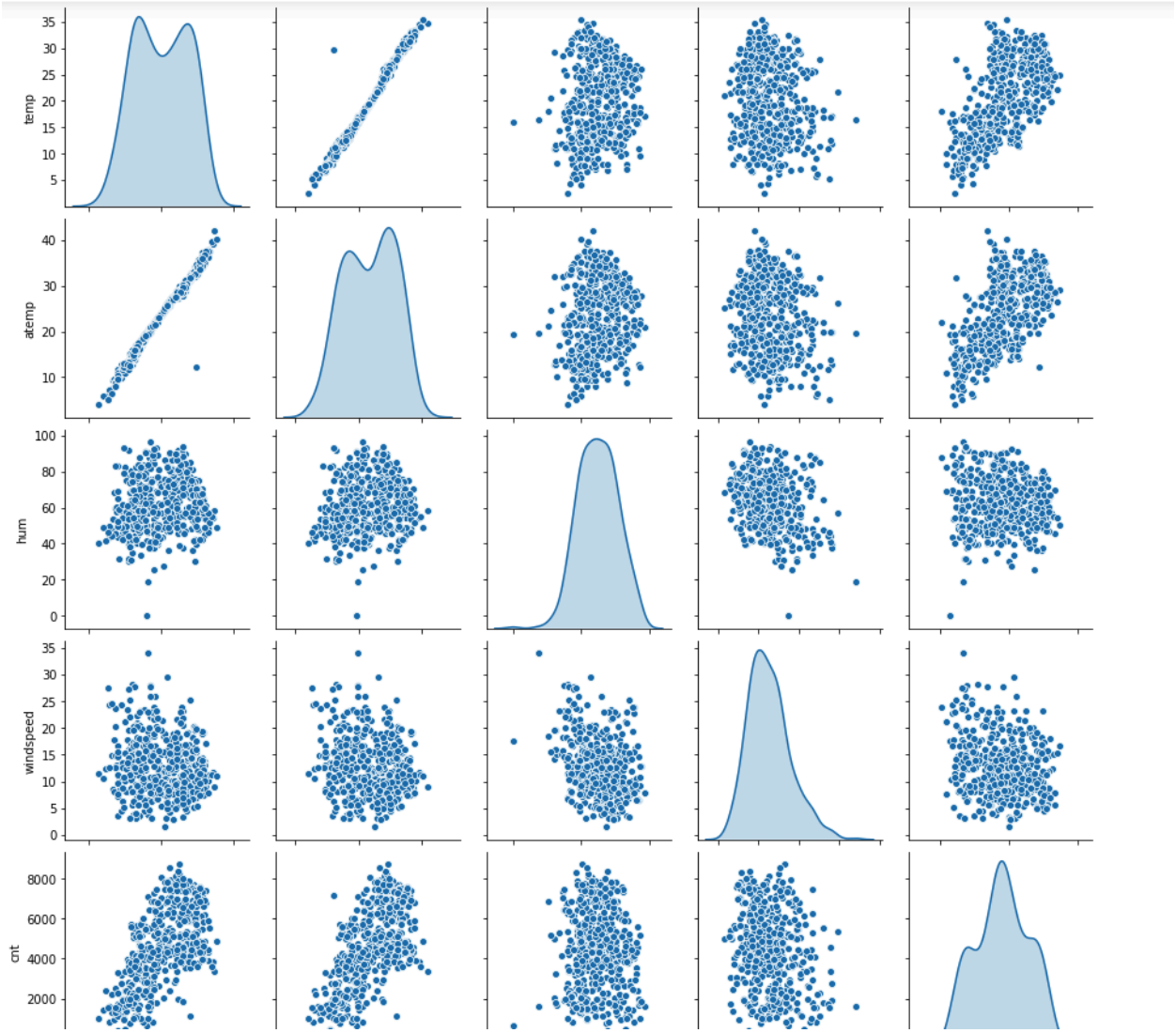
- **Interpretability**: When interpreting the coefficients of a regression model with dummy variables, it's often more straightforward if one category is left out as the reference category. The coefficients of the remaining dummy variables represent the difference between each category and the reference category. This makes the interpretation more intuitive and easier to understand.

- **Reducing dimensionality**: Including dummy variables for all categories increases the dimensionality of the dataset unnecessarily. Dropping the first level reduces the number of dummy variables needed to represent the categorical variable, which can help improve computational efficiency, especially in large datasets.

Overall, using drop_first=True helps improve the performance, interpretability, and efficiency of statistical models that involve categorical variables.

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

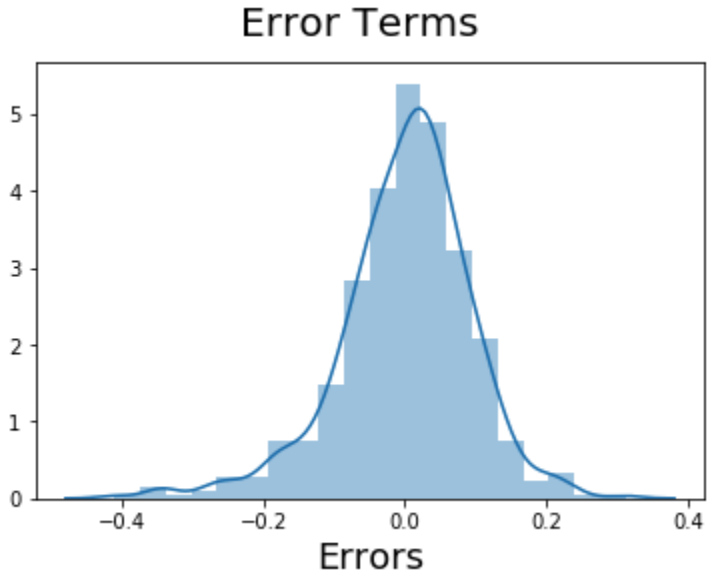
ANSWER: The ‘temp’ and ‘atemp’ variables have the highest correlation with the target variable ‘cnt’.



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

ANSWER: I have validated the assumption of Linear Regression by plotting a distplot of the residuals and analyzing it to see if it is a normal distribution or not and if it has a

mean=0. The diagram below shows that it is normally distributed with mean=0.



Assumptions of Linear model are validated based on:

- Normality of error terms
- Multicollinearity check
- Linear relationship validation
- Homoscedasticity
- Independence of residuals

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

   ANSWER: As per our final Model, the top 3 predictor variables that influences the bike booking are:

   - **Temperature (temp)** - A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
   - **weathersit (Light Snow)** - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
   - **Year (yr)** - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

   So, it's suggested to consider these variables of utmost importance while planning, to achieve maximum Booking.

General Subjective Questions

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

ANSWER: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It's one of the simplest and most commonly used techniques in predictive modeling and statistical analysis. Here's a detailed explanation of the linear regression algorithm:

Objective:

The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the independent variables X and the dependent variable y.

The linear equation takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where:

- y is the dependent variable (target).
- $x_1, x_2, ..., x_n$ are the independent variables (predictors).
- $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients (parameters) that represent the intercept and slopes of the linear equation.
- ϵ is the error term, representing the difference between the actual and predicted values.

Model Fitting:

- To find the optimal values of the coefficients, the algorithm minimizes the sum of squared differences between the actual and predicted values of the dependent variable. This optimization process is often achieved using the method of least squares.
- The coefficients are estimated using mathematical techniques such as ordinary least squares (OLS), gradient descent, or matrix factorization methods.

Assumptions:

- Linear regression assumes a linear relationship between the predictors and the target variable.
- It assumes that the errors (residuals) are normally distributed with constant variance (homoscedasticity).
- It assumes that the predictors are independent of each other (no multicollinearity).

- It assumes that the errors are independent and identically distributed (iid).

Model Evaluation:

- Once the model is trained, it is evaluated using various metrics such as:
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R-squared (R^2) - a measure of the proportion of variance in the dependent variable that is predictable from the independent variables.

Prediction:

- Once the model is trained and evaluated, it can be used to make predictions on new data by simply plugging in the values of the independent variables into the linear equation.

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

ANSWER: Anscombe's quartet is a famous example in statistics that consists of four datasets with nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but vastly different visual representations and underlying relationships. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical exploration in data analysis and to caution against relying solely on summary statistics.

The four datasets in Anscombe's quartet share the following characteristics:

- **Number of Observations:** Each dataset consists of 11 data points.
- **Number of Variables:** Each dataset has two variables: X (independent variable) and Y (dependent variable).
- **Descriptive Statistics:** Despite the datasets having different underlying distributions and relationships, their simple descriptive statistics (mean, variance, correlation, linear regression coefficients) are very similar.

However, despite these similarities, the datasets have dramatically different patterns when plotted graphically. Here's a brief description of each dataset:

- **Dataset I:** This dataset forms a linear relationship between X and Y.
- **Dataset II:** This dataset also forms a linear relationship between X and Y, but with an outlier that heavily influences the linear regression line.
- **Dataset III:** This dataset forms a non-linear relationship between X and Y, with a clear quadratic pattern. However, the correlation coefficient between X and Y is the same as in the other datasets.
- **Dataset IV:** This dataset is similar to Dataset I but with one extreme outlier. Removing this outlier would result in a perfect fit for a linear regression model.

3. What is Pearson’s R? (3 marks)

ANSWER: Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It's named after Karl Pearson, who developed the coefficient.

Pearson's r ranges from -1 to 1:

If r = 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

If r = −1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.

If r = 0 , it indicates no linear relationship between the variables.

Mathematically, Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where:

cov(X,Y) is the covariance of variables X and Y,

σX and σY are the standard deviations of variables X and Y, respectively.

Properties of Pearson's correlation coefficient:

- Symmetry: The correlation between variable X and variable Y is the same as the correlation between Y and X.
- Range: Pearson's r lies between -1 and 1.
- Scale Invariance: Multiplying or adding a constant to each value of either variable does not change the correlation coefficient.
- Outliers Sensitivity: Pearson's correlation coefficient can be sensitive to outliers.

Pearson's correlation coefficient is widely used in various fields such as statistics, data analysis, and machine learning to assess the relationship between variables and to make predictions based on the strength of that relationship.

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

ANSWER: Scaling is a preprocessing step in data analysis and machine learning that involves transforming the range of values of variables to a standard scale. The purpose of scaling is to ensure that all variables contribute equally to the analysis and prevent variables with larger scales from dominating those with smaller scales. Scaling is typically performed before applying certain algorithms or analyses, such as distance-based algorithms (e.g., K-means clustering, K-nearest neighbors) and gradient-based optimization algorithms (e.g., gradient descent).

Reasons for scaling:

- **Normalization:** Scaling helps to normalize the data, bringing all variables to a similar scale. This ensures that no single variable dominates the analysis solely because of its scale.

- **Convergence Speed**: Scaling can improve the convergence speed of optimization algorithms, such as gradient descent, by ensuring that the cost function is well-behaved and has a consistent scale.

- **Interpretability**: Scaling makes it easier to interpret the coefficients or weights associated with each variable in models like linear regression or neural networks.

- **Regularization**: Some regularization techniques, such as L1 and L2 regularization, assume that all variables are on a similar scale. Scaling ensures that regularization is applied uniformly across all variables.

- **Distance-based Algorithms**: Scaling is essential for distance-based algorithms like K-means clustering and K-nearest neighbors, as these algorithms are sensitive to the scale of the variables.

There are two main types of scaling: normalized scaling and standardized scaling.

Normalized Scaling:

- Normalized scaling (or min-max scaling) transforms the values of variables to a range between 0 and 1.
- The formula for normalized scaling is:

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

- Where X is the original value of the variable, X_{min} is the minimum value of the variable, and X_{max} is the maximum value of the variable.
- Normalized scaling preserves the relative relationships between the values of variables but does not handle outliers well.

Standardized Scaling:

- Standardized scaling (or z-score normalization) transforms the values of variables to have a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling is:

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

- Where X is the original value of the variable, μ is the mean of the variable, and σ is the standard deviation of the variable.
- Standardized scaling is robust to outliers and preserves the shape of the distribution, but it does not ensure that the transformed values fall within a specific range.

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

ANSWER: Sometimes the value of the Variance Inflation Factor (VIF) becomes infinite due to perfect multicollinearity among the predictor variables in the regression model. Perfect multicollinearity happens when one or more predictor variables can be perfectly predicted by a linear combination of other predictor variables in the model.

If there is perfect correlation, then VIF = infinity.

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get (R-Squared)=1, which leads to 1-(1-R^2) infinity.

To solve this, we need to drop one of the variables from the dataset which is causing the perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWER: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given sample of data follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the sample data to the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1).

Use in Linear Regression:

- Q-Q plots are often used in linear regression to assess the normality of residuals, which is an important assumption of linear regression models.
- After fitting a linear regression model, the residuals (the differences between observed and predicted values) are analyzed using a Q-Q plot.
- If the residuals follow a normal distribution, the points on the Q-Q plot will fall along a straight line. Deviations from this line indicate departures from normality.
- Departures from normality in the residuals may suggest that the linear regression model is not appropriate or may indicate the presence of outliers or influential points that should be investigated further.

Importance in Linear Regression:

- **Ensures Assumptions:** Normality of residuals is one of the key assumptions of linear regression. By using a Q-Q plot, we can assess whether this assumption holds or not.
- **Model Evaluation:** Q-Q plots provide a visual and quantitative way to evaluate the adequacy of a linear regression model. If the residuals are normally distributed, it indicates that the model adequately captures the variability in the data.
- **Diagnostic Tool:** Q-Q plots serve as a diagnostic tool to identify potential issues with the linear regression model, such as nonlinearity, heteroscedasticity, or outliers.

In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals, ensuring model assumptions are met, evaluating model adequacy, and diagnosing potential problems with the model.