Assignment-based Subjective Questions

1. From the analysis of the categorical variables from the dataset, not all categorical variables have strong significance on dependant variable **cnt**. Only few of them have strong relationship with the dependant variable. Those are,
   a. Season
   b. Year
   c. Month (Not all months)
   d. Holiday
2. By setting `drop_first=True`, one level of each categorical variable is excluded during the creation of dummy variables. This action removes the redundant information, and it avoids perfect multicollinearity among the predictors.
3. Temperature has highest correlation with the target variable in pair-plot.
4. Using Residual analysis. By plotting the error terms of training data and checking its distribution we can validate the assumptions of linear regression. If this plot is like normal or gaussian distribution, then we can say that the assumption of linear regression is valid.
5. The top 3 features explaining the demand are,
   a. Temperature
   b. Year
   c. Light Snow (-ve relation)

General Subjective Questions

1. Linear regression is a statistical method used for modelling the relationship between a dependent variable (often denoted as 'y') and one or more independent variables (often denoted as 'x'). Its goal is to establish a linear relationship between these variables, allowing for the prediction of the dependent variable based on the independent variable(s).
   Assumptions of Linear Regression:

   - **Linearity**: Assumes a linear relationship between the independent and dependent variables.
   - **Independence**: Assumes that the residuals (the differences between actual and predicted values) are independent of each other.
   - **Homoscedasticity**: Assumes that the variance of the residuals remains constant across all levels of the independent variables.
   - **Normality**: Assumes that the residuals are normally distributed.

   Types of Linear Regression

   a. **Simple Linear Regression**:
      In simple linear regression, there's a single independent variable predicting a dependent variable.
      The relationship between the independent variable 'x' and the dependent variable 'y' is represented by a straight line: y = mx + c where 'm' is the slope of the line and 'c' is the intercept.
   b. **Multiple Linear Regression**:
      In multiple linear regression, there are multiple independent variables predicting a dependent variable.

Fitting the Model:

- The coefficients (slope and intercept in simple linear regression, and coefficients for each independent variable in multiple linear regression) are estimated to best fit the data.
- This estimation is often done using optimization techniques like gradient descent or analytical methods that directly compute the coefficients that minimize the cost function.

Evaluation:

The model's performance is evaluated using metrics such as R-squared (coefficient of determination), Mean Squared Error (MSE), Root Mean Squared Error (RMSE)

Prediction:

Once the model is trained and evaluated, it can be used to make predictions on new or unseen data by applying the learned coefficients to the independent variables.

2. **Anscombe's quartet** is a set of four small datasets that have nearly identical simple descriptive statistics but display very different characteristics when graphed. It was created to emphasize the importance of visualizing data and not relying solely on summary statistics.

The four datasets in Anscombe's quartet consist of 11 data points each and contain pairs of x and y variables. Despite having identical means, variances, correlations, and linear regression parameters, they exhibit diverse patterns when graphed.

The characteristics of each dataset in Anscombe's quartet:

- **Dataset I**: This set forms a simple linear relationship with a scatterplot forming a clear linear pattern. It fits well with a linear regression line.
- **Dataset II**: It also represents a linear relationship but with an outlier that significantly affects the linear regression line. Removing the outlier might change the nature of the relationship.
- **Dataset III**: This set doesn't follow a linear pattern but exhibits a strong relationship when fitted with a quadratic curve. It emphasizes how a non-linear relationship might be overlooked if one only considers linear models.
- **Dataset IV**: Despite having an outlier like Dataset II, the presence of this outlier dramatically impacts the regression line. When the outlier is removed, the dataset forms a perfect relationship except for one point.

The main takeaway from Anscombe's quartet is that summary statistics alone, such as means, variances, and correlation coefficients, may not fully capture the nature and nuances of the data. It underscores the importance of visualizing data and checking for patterns using graphs and plots to better understand relationships, trends, and potential outliers. Relying solely on summary statistics can lead to misleading interpretations or missing crucial aspects of the data.

3. **Pearson's correlation coefficient** is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Pearson's r value ranges from -1 to +1:

- p=+1r=+1: Represents a perfect positive linear relationship, where one variable increases as the other variable increases in a linear fashion.
- p=−1r=−1: Indicates a perfect negative linear relationship, where one variable decreases as the other variable increases in a linear fashion.

- p=0r=0: Implies no linear relationship between the variables. However, note that it does not imply the absence of any relationship; there could still be a non-linear relationship between the variables.

Key characteristics of Pearson's correlation coefficient:

- **Strength of Relationship**: The absolute value indicates the strength of the linear relationship between the variables. The closer the absolute value is to 1, the stronger the linear relationship.
- **Direction of Relationship**: The sign (+ or -) indicates the direction of the linear relationship. A positive value indicates a positive correlation (both variables increase or decrease together), while a negative value indicates a negative correlation (one variable increases while the other decreases).
- **Assumption**: Pearson's *r* assumes a linear relationship between the variables and is sensitive to outliers. It measures only linear associations, so it might not capture non-linear relationships.

4. **Scaling** in the context of data preprocessing refers to the process of transforming or normalizing the features or variables of a dataset to a specific range or distribution. The primary goal of scaling is to bring all features to a similar scale or level, which can be crucial for certain machine learning algorithms to perform optimally.

Scaling is performed due to the following reasons.

a. **Algorithm Sensitivity**: Many machine learning algorithms are sensitive to the scale of features.
b. **Convergence**: Scaling can help algorithms converge faster during the training process, especially gradient descent-based algorithms, by making the optimization process smoother and more efficient.

Types of Scaling

a. **Normalized Scaling (Min-Max Scaling)**: Normalization scales the features to a range typically between 0 and 1. Normalization maintains the relative relationships between values but doesn't handle outliers well.
b. **Standardized Scaling (Z-score normalization)**: Standardization transforms features to have a mean of 0 and a standard deviation of 1. Standardization is effective in handling outliers as it centres the data around the mean, but it doesn't ensure a specific range for the transformed values.

5. **VIF** stands for Variance Inflation Factor, which measures the severity of multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other. When the VIF value is calculated to be infinite, it usually signifies an extreme case of multicollinearity.

6. A **Q-Q** plot, short for quantile-quantile plot, is a graphical tool used to assess whether a given dataset follows a certain probability distribution or to compare the distribution of a sample with a theoretical distribution like the normal distribution.

Working of Q-Q Plot

Theoretical Quantiles vs. Sample Quantiles: In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of a theoretical distribution.
If the data follows the theoretical distribution, the points in the Q-Q plot will approximately fall along a straight line.

If the points on the Q-Q plot form a relatively straight line, it indicates that the sample data closely follows the distribution specified by the theoretical line. Deviations from a straight line suggest departures from the assumed distribution.

Importance in Linear Regression

Q-Q plots are particularly useful in linear regression for several reasons:

Assumption Checking: A Q-Q plot of the residuals allows us to visually inspect whether the residuals approximate a normal distribution.

Identifying Outliers and Skewness: Q-Q plots can help identify outliers or skewness in the data. Outliers or heavy tails in the data distribution might deviate from the expected pattern on the Q-Q plot, appearing as points that fall far from the straight line.