Assignment-based Subjective Questions

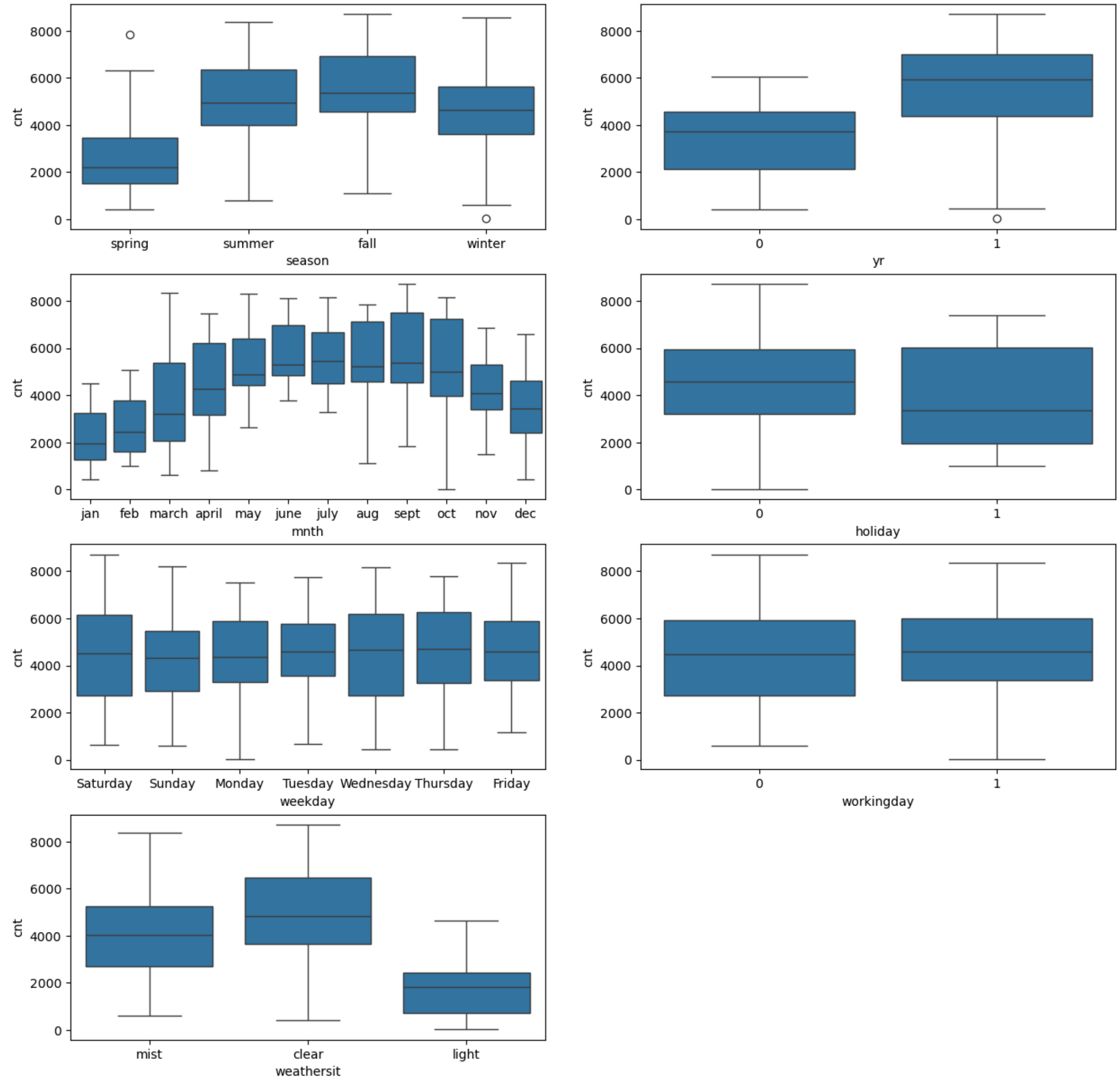
Question: 1

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

Answer:

As part of the analysis of the categorical variables from the dataset, the following set of box plots were plotted using the categorical variables on x-axis vs the dependent variable on the y-axis. The insights related to categorical variables vs dependent variable are -

- **season**: fall season has the highest demand.
- **year**: bike demand has increased from 2018 to 2019.
- **month**: bike demand increases from Jan to september and drops after that. dec has least bike demand.
- **holiday**: bike demand decreases during holiday time.
- **weekday**: there is no particular pattern.
- **workingday**: bike demand is little high on workingday.
- **weathersit**: bike demand is high, when weathersit is clear



.

Question: 2

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

Answer:

Using `drop_first=True` during dummy variable creation is important for avoiding multicollinearity in regression models and reducing the risk of the dummy variable trap.

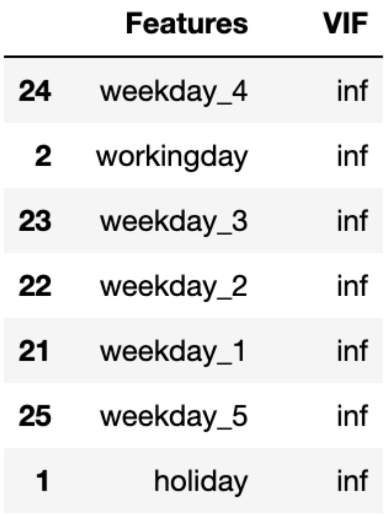
It is important because:

1. **Multicollinearity**: When you include dummy variables for all categories in a categorical variable, you introduce multicollinearity because the dummy variables are highly correlated with each other. This can lead to unstable estimates of the regression coefficients and inflated standard errors. By dropping the first dummy variable, you retain the necessary information about the categories while avoiding perfect multicollinearity.

2. **Dummy Variable Trap**: The dummy variable trap is a situation where two or more dummy variables are highly correlated (perfect multicollinearity). This happens when one or more dummy variables can be predicted exactly using the others. This leads to unreliable regression coefficients and invalid hypothesis testing. Dropping the first dummy variable prevents this trap.

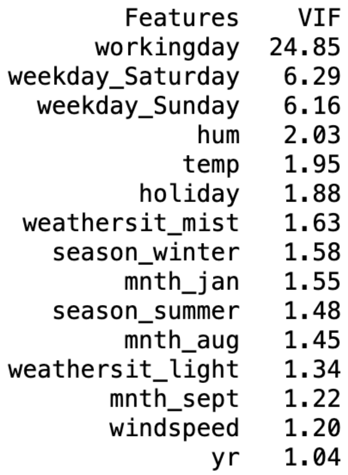
By using `drop_first=True`, you essentially use (n - 1) dummy variables for a categorical variable with (n) categories, which maintains the necessary information while mitigating multicollinearity issues.

For example: In the BoomBikes MLR assignment, if the ‘drop_first=True’ is not used then in the First model, one might see VIF= inf for many feature variables as follows -



	Features	VIF
24	weekday_4	inf
2	workingday	inf
23	weekday_3	inf
22	weekday_2	inf
21	weekday_1	inf
25	weekday_5	inf
1	holiday	inf

However, if ‘drop_first=True’ is used then the model behaves relatively better.



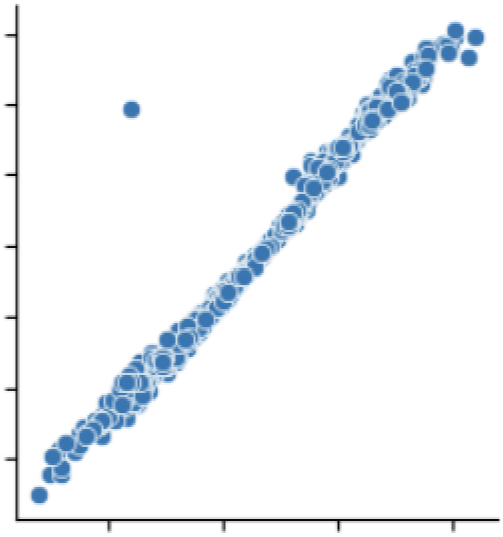
```
          Features    VIF
        workingday  24.85
   weekday_Saturday   6.29
     weekday_Sunday   6.16
               hum   2.03
              temp   1.95
           holiday   1.88
   weathersit_mist   1.63
     season_winter   1.58
          mnth_jan   1.55
     season_summer   1.48
          mnth_aug   1.45
  weathersit_light   1.34
         mnth_sept   1.22
         windspeed   1.20
                yr   1.04
```

Question: 3

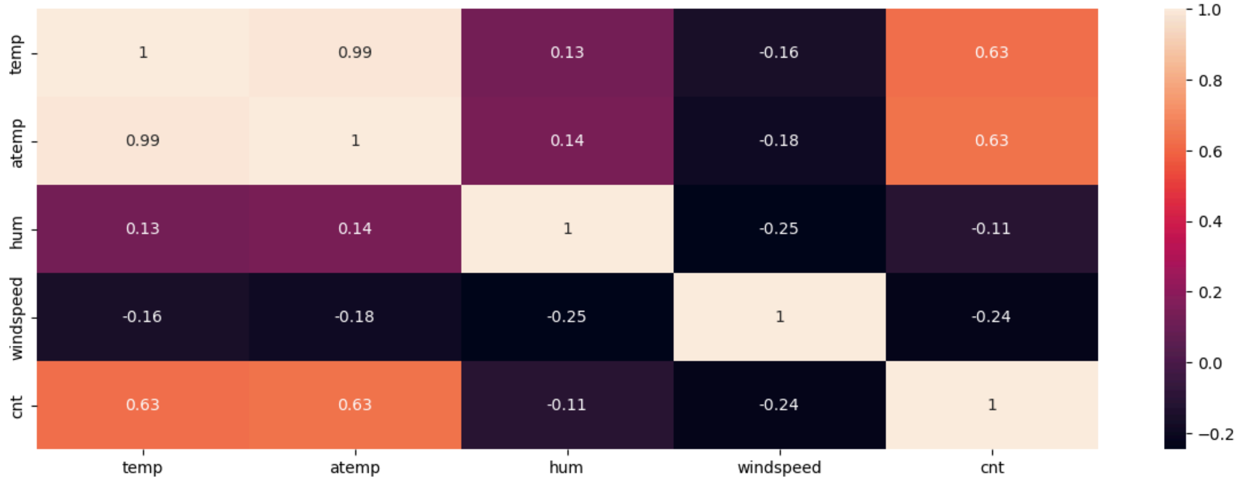
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

‘temp’ variable has the highest correlation with the target variable **‘cnt’**



This can also be validated using heatmap where the correlation coefficient is **0.63** between **‘temp’** and **‘cnt’**



‘atemp’ also has the same correlation coefficient but **‘temp’** has been selected because it is the primary cause which eventually affects **‘atemp’** also.

Question: 4

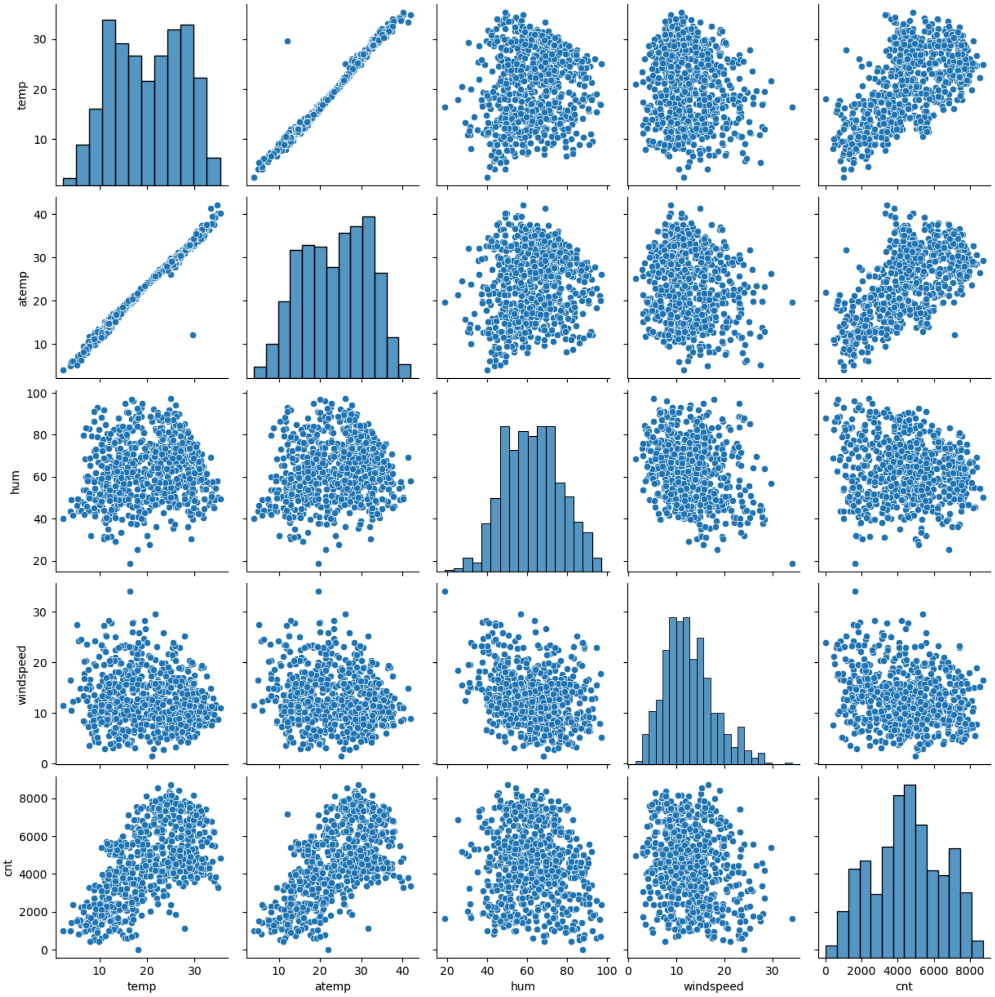
How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

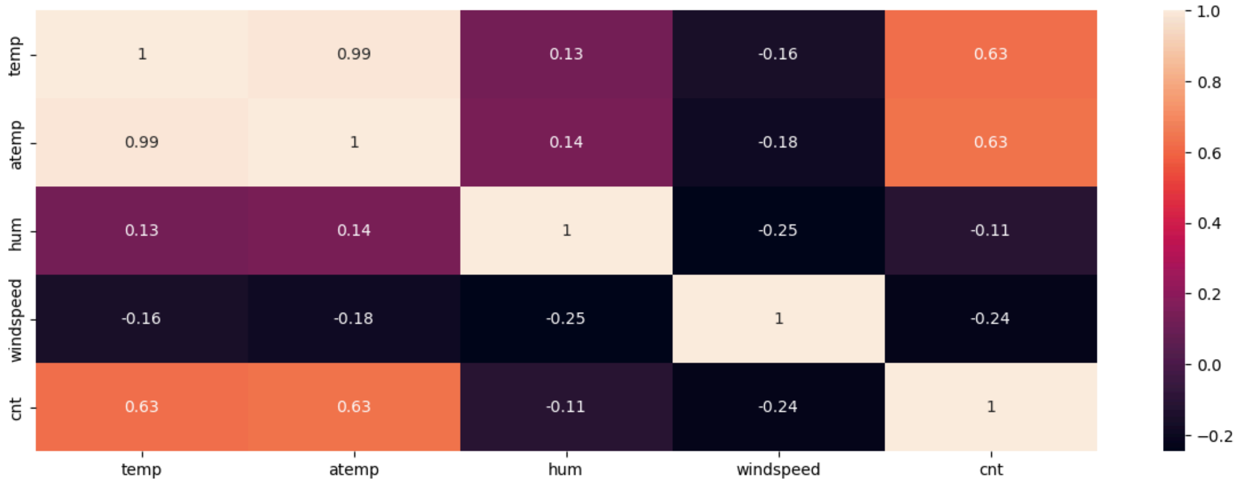
Answer:

I performed following evaluations on the Model to validate the assumptions of Linear Regression :

1. **Little to no Multicollinearity**

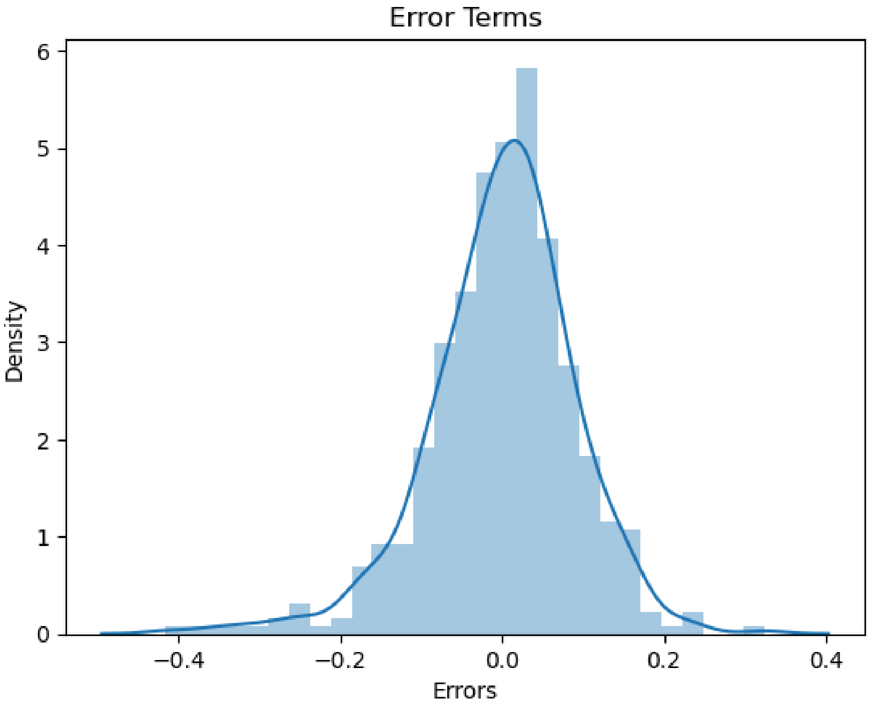
   Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model.Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features. Following pairplot and heatmap tells us the correlation among feature variables. We removed ‘atemp’ to create our model and hence reduced Multicollinearity.





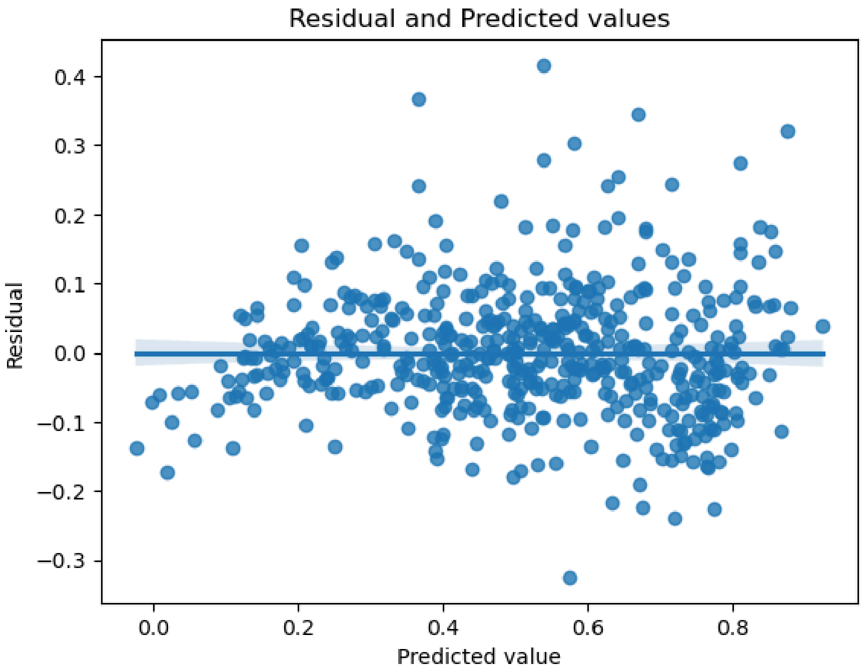
2. **Residual analysis**

The error(residuals) follow a normal distribution. If we plot a graph of Errors then the graph should follow a normal distribution. Below graph from our regression shows a Normal distribution graph with mean close to 0.



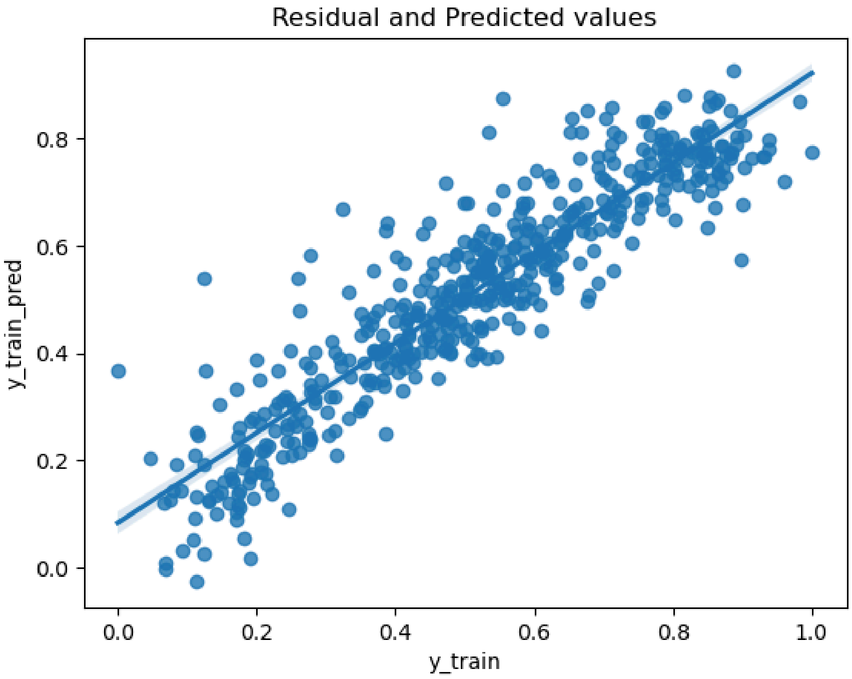
3. **Dependency among Error terms**

This happens when residual errors are dependent on each other.The presence of correlation in error terms drastically reduces model's accuracy.This usually occurs in time series models where the next instant is dependent on previous instant. This is called as Autocorrelation can be tested with the help of Durbin-Watson test. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation. This value for my model was - 2.103 which indicates the model passes this assumption.



4. Homoscedasticity of Error terms

Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables.A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity.There should be no clear pattern in the distribution and if there is a specific pattern. Following is the scatter plot from our model which clearly shows Homoscedastic error terms.

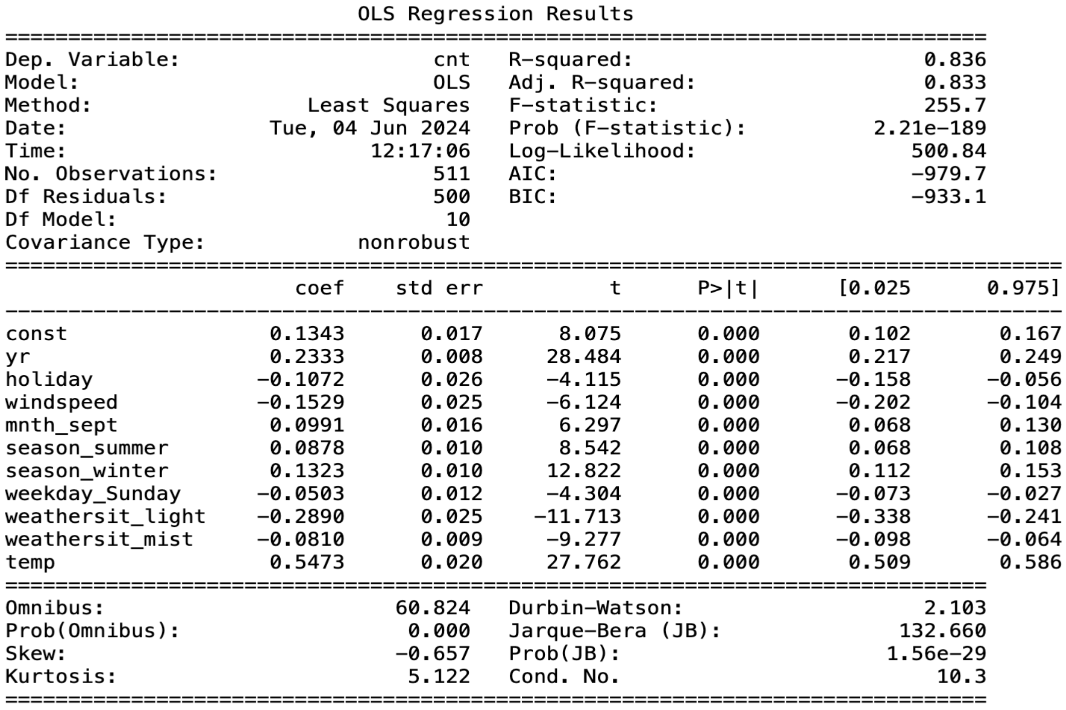


Question: 5

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

Answer:

I would like to present the following informational summary from the Final Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.836
Model:                            OLS   Adj. R-squared:                  0.833
Method:                 Least Squares   F-statistic:                     255.7
Date:                Tue, 04 Jun 2024   Prob (F-statistic):          2.21e-189
Time:                        12:17:06   Log-Likelihood:                 500.84
No. Observations:                 511   AIC:                            -979.7
Df Residuals:                     500   BIC:                            -933.1
Df Model:                          10
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                0.1343      0.017      8.075      0.000       0.102       0.167
yr                   0.2333      0.008     28.484      0.000       0.217       0.249
holiday             -0.1072      0.026     -4.115      0.000      -0.158      -0.056
windspeed           -0.1529      0.025     -6.124      0.000      -0.202      -0.104
mnth_sept            0.0991      0.016      6.297      0.000       0.068       0.130
season_summer        0.0878      0.010      8.542      0.000       0.068       0.108
season_winter        0.1323      0.010     12.822      0.000       0.112       0.153
weekday_Sunday      -0.0503      0.012     -4.304      0.000      -0.073      -0.027
weathersit_light    -0.2890      0.025    -11.713      0.000      -0.338      -0.241
weathersit_mist     -0.0810      0.009     -9.277      0.000      -0.098      -0.064
temp                 0.5473      0.020     27.762      0.000       0.509       0.586
==============================================================================
Omnibus:                       60.824   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              132.660
Skew:                          -0.657   Prob(JB):                     1.56e-29
Kurtosis:                       5.122   Cond. No.                         10.3
==============================================================================
```

Following is the Significant variables list:

- **'yr'**
- **'holiday'**
- **'windspeed',**
- **'mnth_sept',**
- **'season_summer',**
- **'season_winter',**
- **'weekday_Sunday',**
- **'weathersit_light',**
- **'weathersit_mist',**
- **'Temp'**

Also the equation of our best fitted line is:

$$cnt = 0.1343 + 0.2333 \times yr - 0.1072 \times holiday - 0.1529 \times windspeed + 0.0991 \times mnthsept + 0.0878 \times seasonsummer + 0.1323 \times seasonwinter - 0.0503 \times weekdaySunday - 0.2890 \times weathersitlight - 0.0810 \times weathersitmist + 0.5473 \times temp$$

Thus, top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **'Temp'**
2. **'yr'**
3. **'weathersit_light**

General Subjective Questions

Question: 1

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

Answer:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (often denoted as y) and one or more independent variables (often denoted as x). It's called "linear" because it assumes that the relationship between the dependent variable and the independent variable(s) can be approximated by a straight line.

Following is step by step guide to follow for creating a Linear Regression Model:

1. **Understanding the Problem**: Linear regression is typically used when you have a dataset containing pairs of observations ((x_i, y_i)), where (x_i) represents the independent variable(s) and (y_i) represents the dependent variable. The goal is to find a linear relationship between these variables.
2. **Assumptions**:
   - **Linearity**: The relationship between the independent and dependent variables is linear.
   - **Independence**: Observations are independent of each other.
   - **Homoscedasticity**: The variance of the residuals (the differences between the observed and predicted values) is constant across all levels of the independent variables.
   - **Normality**: The residuals are normally distributed.
3. **Simple Linear Regression**: In simple linear regression, there's only one independent variable. The relationship between (x) and (y) can be represented by the equation of a straight line: [$y = B_0 + B_1x + \boldsymbol{\epsilon}$]
   - B_0 is the intercept of the line (the value of y when x is 0).
   - B_1 is the slope of the line (the change in y for a one-unit change in x).
   - $\boldsymbol{\epsilon}$ represents the error term.
4. **Multiple Linear Regression**: When there are multiple independent variables, the equation becomes: [$y = B_0 + B_1x1 + ... + B_nx_n + \boldsymbol{\epsilon}$]
   - Here, (x_1, x_2, ..., x_n) are the independent variables, and (B_1, ..., Bn) are their respective coefficients.
5. **Fitting the Model**: The goal is to find the values of (B_0, B_1, ..., B_n) that minimize the sum of squared residuals (the vertical distance between the observed and predicted values). This is often done using the method of least squares.
6. **Ordinary Least Squares (OLS)**: In OLS, the algorithm minimizes the sum of squared differences between the observed and predicted values of (y) for all data points. This is achieved by finding the values of the coefficients that minimize the following cost function:

$$\text{Cost} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

   - y_i is the observed value.
   - $\hat{y}_i$ is the predicted value.
   - n is the number of data points.
7. **Model Evaluation**: After fitting the model, it's essential to evaluate its performance. Common metrics include (R^2) (coefficient of determination), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc.

8. **Interpretation**: Once the model is fitted and evaluated, the coefficients ((B_0, B_1, ..., B_n)) can be interpreted. The intercept B_0 represents the value of the dependent variable when all independent variables are 0, while the slopes ((B_1, ..., B_n)) represent the change in the dependent variable for a one-unit change in each independent variable, holding other variables constant.
9. **Predictions**: Finally, the model can be used to make predictions on new data

Question: 2

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

Answer:

Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. These datasets are used to demonstrate the importance of graphing data before analyzing it statistically and to show how different datasets can produce similar statistical properties while being very different in structure and appearance.

Each dataset in Anscombe's quartet consists of eleven (x, y) points. Despite having nearly identical simple descriptive statistics—such as mean, variance, correlation coefficient, and linear regression line—the datasets are significantly different from each other when plotted. Here’s a detailed look at each aspect of Anscombe's quartet:

For all four datasets, the following statistical properties are approximately the same:

- **Mean of x-values**: 9
- **Mean of y-values**: 7.5
- **Variance of x-values**: 11
- **Variance of y-values**: 4.125
- **Correlation between x and y**: 0.816
- **Linear regression line**: (y = 3 + 0.5x)
- **Coefficient of determination (R^2)**: 0.67

Key Take Aways:

1. **Importance of Visualizing Data**: Simply relying on summary statistics can be misleading. Graphical representations such as scatter plots can reveal patterns, trends, and outliers that are not apparent from summary statistics alone.
2. **Data Context and Structure**: Different datasets can exhibit the same statistical properties but have different distributions and relationships. This underscores the need to understand the context and structure of the data.
3. **Influence of Outliers**: Outliers can significantly impact statistical analyses. Identifying and understanding the role of outliers is crucial in interpreting data correctly.

Conclusion

Anscombe’s quartet illustrates that summary statistics alone cannot capture the nuances of data distributions and relationships. Visual analysis through plotting is essential to uncovering the true nature of the data. This insight is fundamental for effective statistical practice, ensuring that interpretations and conclusions are based on a comprehensive understanding of the data.

Question: 3
What is Pearson’s R? (3 marks)

Answer:

Pearson’s R, also known as the Pearson correlation coefficient, is a statistical measure that calculates the strength and direction of the linear relationship between two variables. It is denoted by the symbol (r) and ranges from -1 to 1.

Calculation of Pearson's R

Pearson’s R is calculated using the following formula:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

Where:

- (Xi) and (Yi) are the individual sample points.
- $\overline{X}$ and $\overline{Y}$ are the mean values of the X and Y variables, respectively.

Interpretation of Pearson’s R

- **r = 1**: Perfect positive linear relationship. As one variable increases, the other variable increases proportionally.
- **r = -1**: Perfect negative linear relationship. As one variable increases, the other variable decreases proportionally.
- **r = 0**: No linear relationship. The variables do not have any linear correlation.

Values between -1 and 1 indicate the degree of correlation:

- **0 < r < 1**: Positive correlation. Higher values of one variable are associated with higher values of the other variable.
- **-1 < r < 0**: Negative correlation. Higher values of one variable are associated with lower values of the other variable.

Strength of Correlation

The strength of the correlation can be qualitatively described as:

- **0.00 to ±0.10**: Negligible correlation.
- **±0.10 to ±0.39**: Weak correlation.
- **±0.40 to ±0.69**: Moderate correlation.
- **±0.70 to ±0.89**: Strong correlation.
- **±0.90 to ±1.00**: Very strong correlation.

Assumptions

For Pearson's R to be a valid measure, the following assumptions should be met:

1. **Linearity**: The relationship between the variables should be linear.
2. **Homogeneity of Variance (Homoscedasticity)**: The spread of the data points should be constant along the line of best fit.
3. **Normality**: The variables should be approximately normally distributed, especially for smaller sample sizes.
4. **Interval or Ratio Scale**: The data should be measured on an interval or ratio scale.

Applications

Pearson's R is widely used in various fields, including:

- **Psychology**: To measure the relationship between psychological traits.
- **Medicine**: To assess the correlation between variables such as age and blood pressure.
- **Economics**: To examine the relationship between economic indicators.

Limitations

- **Sensitivity to Outliers**: Pearson's R can be heavily influenced by outliers.
- **Linear Relationship**: It only measures linear relationships and does not capture nonlinear correlations.
- **Bivariate Measure**: Pearson's R only considers the relationship between two variables at a time.

In summary, Pearson's R is a versatile and commonly used statistic for measuring the strength and direction of linear relationships between two variables, but it is important to consider its assumptions and limitations when interpreting the results.

Question: 4

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

Answer:

Scaling is a data preprocessing technique used in machine learning and statistics to adjust the range and distribution of numerical features. The main goal of scaling is to ensure that all features contribute equally to the analysis, preventing any single feature from disproportionately influencing the model.

Scaling is performed for several key reasons:

1. Improving Model Performance: Many machine learning algorithms, such as gradient descent-based methods, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), are sensitive to the range of feature values. Features with larger ranges can dominate the learning process, leading to suboptimal model performance.
2. Faster Convergence: In optimization problems, particularly those using gradient descent, scaled features can lead to faster convergence because the algorithm does not need to compensate for differing feature magnitudes.
3. Distance-based Metrics: Algorithms that rely on distance calculations (e.g., KNN, clustering) can produce more accurate results when the features are on a similar scale, ensuring that each feature contributes proportionately to the distance metric.
4. Improved Interpretability: Standardized data can make the model coefficients easier to interpret, particularly in linear models where the coefficients represent the effect of each feature.

Difference Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling (Min-Max Scaling):
   - Definition: Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1].
   - Formula: $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
     - Where (x) is the original value, (x_{min}) is the minimum value in the dataset, and (x_{max}) is the maximum value in the dataset.
   - Purpose: Normalization is useful when the data does not follow a Gaussian distribution or when you want the data to be bounded within a specific range.
   - Example: If a feature ranges from 10 to 100, normalization would scale it to range from 0 to 1.
2. Standardized Scaling (Z-score Standardization):
   - Definition: Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
   - Formula: $x' = \frac{x - \mu}{\sigma}$
     - Where (x) is the original value, (μ) is the mean of the dataset, and (σ) is the standard deviation.
   - Purpose: Standardization is beneficial when the data follows a Gaussian distribution or when the algorithms assume normally distributed data (e.g., linear regression, logistic regression).
   - Example: If a feature has a mean of 50 and a standard deviation of 10, standardization would transform a value of 60 to $\frac{60-50}{10} = 1$

Summary

- Normalization scales data to a fixed range (e.g., [0, 1]), making it useful for non-Gaussian data and distance-based algorithms.
- Standardization scales data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms that assume normally distributed data and improving interpretability of model coefficients.

Both scaling methods aim to standardize the contribution of features to the model, but the choice between them depends on the nature of the data and the specific requirements of the machine learning algorithm being used.

Question: 5

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

Answer:

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when predictor variables (independent variables) are highly correlated with each other, which can cause problems in estimating the coefficients of the regression model.

VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. The formula for VIF for a predictor (X_j) is:

$$\mathrm{VIF}(X_j) = \frac{1}{1-R_j^2}$$

Where R_j^2 is the coefficient of determination obtained when X_j is regressed on all the other predictor variables in the model. The value of VIF becomes infinite when R_j^2 is equal to 1. This situation arises when there is perfect multicollinearity, meaning that one predictor variable is an exact linear combination of one or more other predictor variables.

Reasons for Infinite VIF

1. Perfect Multicollinearity:
   - Duplicate Variables: If you include the same variable more than once in the regression model, or if you have one variable that is a perfect multiple or sum of another, you will get perfect multicollinearity.
   - Linear Dependence: If one predictor can be perfectly predicted by a linear combination of other predictors.
2. Incorrect Model Specification:
   - Sometimes, perfect multicollinearity can occur due to mistakes in model specification, such as including dummy variables for all categories of a categorical variable without excluding one reference category (dummy variable trap).

Example

Consider a regression model with two predictors, X_1 and (X_2), where:

$X_2 = 2X_1$

Here, X_2 is a perfect linear function of X_1 When you try to calculate the VIF for X_1:

- Regress X_1 on X_2 : [$X_1 = \beta 0 + \beta 2\ X_2 + \epsilon$] Given ($X_2 = 2X_1$), the R_j^2 for this regression will be 1.

Thus,

$$\mathrm{VIF}(X_1) = \frac{1}{1-1} = \infty$$

Summary

An infinite VIF indicates perfect multicollinearity, where a predictor variable is an exact linear combination of other predictors. This causes severe issues in regression analysis, including numerical instability and difficulties in interpreting the model. Addressing this issue involves removing or combining collinear variables and ensuring the correct specification of the regression model.

Question: 6

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

Answer:

Q-Q Plot

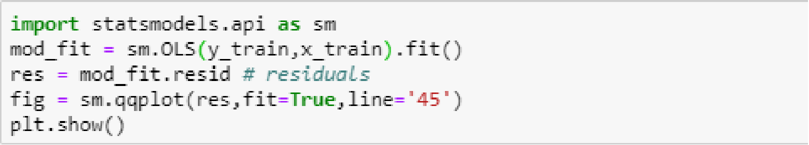
Normal distribution of the residuals can be validated by plotting a q-q plot.

Usage

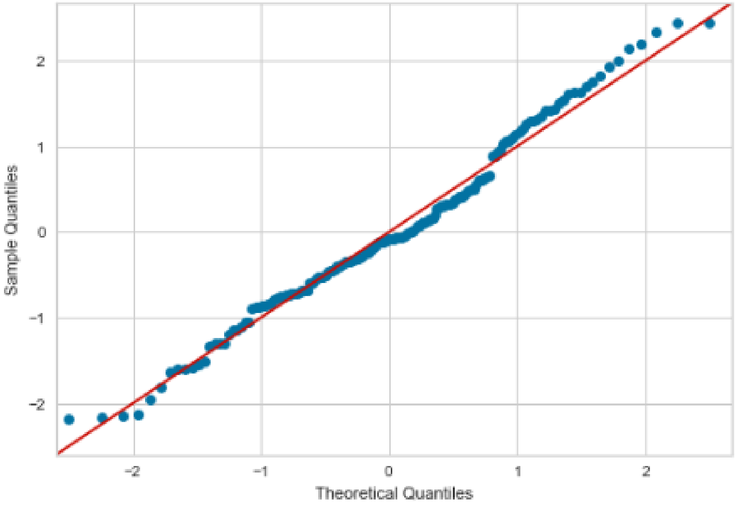
Using the q-q plot we can infer if the data comes from a normal distribution. If yes, the plot would show a fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

For example:

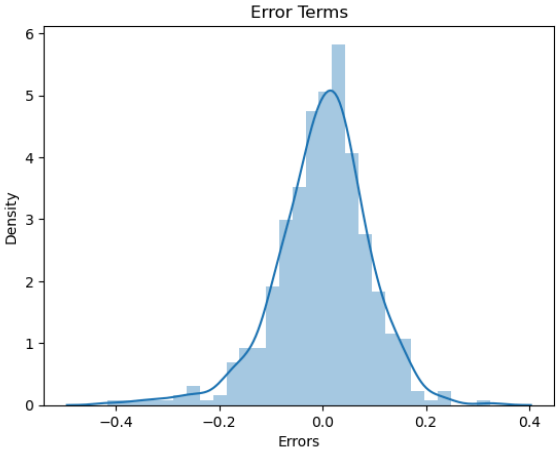
The following q-q plot of a sample advertising data set shows that the errors(residuals) are fairly normally distributed.



```
import statsmodels.api as sm
mod_fit = sm.OLS(y_train,x_train).fit()
res = mod_fit.resid # residuals
fig = sm.qqplot(res,fit=True,line='45')
plt.show()
```

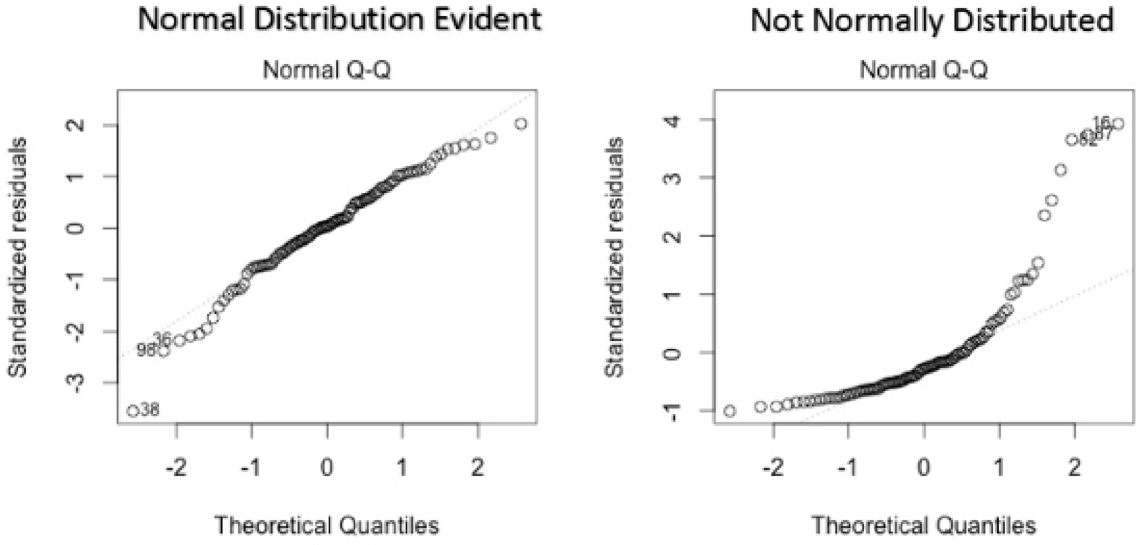


In this case, the histogram plot of “Error(residuals) vs Predicted values” will show that the errors are normally distributed with mean close to 0 as follows



Importance of a Q-Q plot in Linear Regression

1. One of the important aspects of the MLR Model evaluation is to evaluate the Assumptions of the Linear Regression.
2. One of these assumption is **<u>Normal distribution of error terms</u>** i.e the error(residuals) follow a normal distribution.
3. However, as sample sizes increase, the normality assumption for the residuals is not needed. To be precise, if we consider repeated sampling from our population, for large sample sizes, the distribution (across repeated samples) of the ordinary least squares estimates of the regression coefficients follow a normal distribution.
4. As a consequence, for moderate to large sample sizes, non-normality of residuals should not adversely affect the usual inferential procedures. This result is a consequence of an extremely important result in statistics, known as the **Central Limit theorem.**
5. Hence, using Q-Q plot to assess the Normal distribution of Error terms is vital for a successful Model.



Q-Q Plots