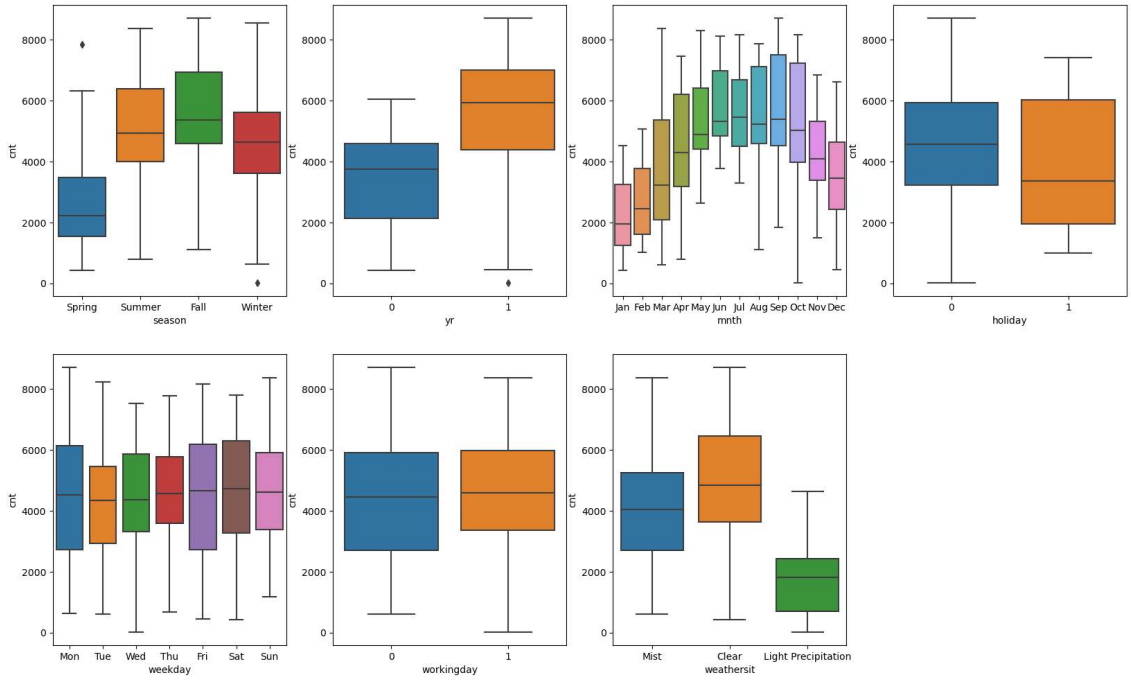
EPGML C55 July 2023
Syed Abdul Rahim

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?**
   Basing on the data the following categorical variables are identified, viz.,
   Season: Season
   Yr: Year
   Mnth: Month
   Holiday: Holiday
   Weekday: Weekday
   Workingday: Working Day
   Weathersit: Weather Situation

   Boxplot categorical_variable vs count:



   Observations:
   a. season: Highest bike demand is in Fall and lowest is in Spring
   b. yr: 2019 has seen more demand than 2018

c. mnth: Demand has varied much across months, with highest in September, while lowest in January. With an consistent demand from May to October
d. holiday: Demand on non-holiday vs holiday
e. weekday: No such noticeable difference
f. workingday: No such noticeable difference
g. weathersit: High demand in Clear weather, with the lowest on Light Precipitation and no demand at all on High Precipitation

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

It is important because it helps to avoid multicollinearity issues and prevents the "dummy variable trap."

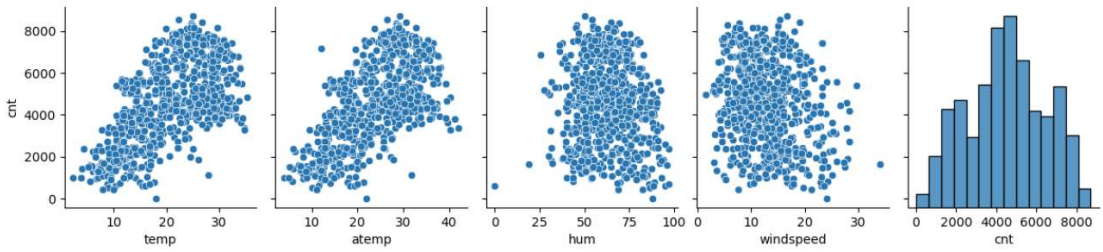
Multicollinearity: Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. If the dummy variables are created without dropping one-level, these dummy variables will be perfectly correlated because they can be derived from each other, which can cause problems in regression analysis, including unstable coefficient estimates.

The Dummy Variable Trap: The dummy variable trap is a specific case of multicollinearity. It occurs when dummy variables aren't dropped one-level. Because of which the model cannot distinguish the effect of the omitted category from the intercept term. As a result, the coefficients for the remaining dummy variables become collinear, and it becomes impossible to isolate the unique contribution of each category.

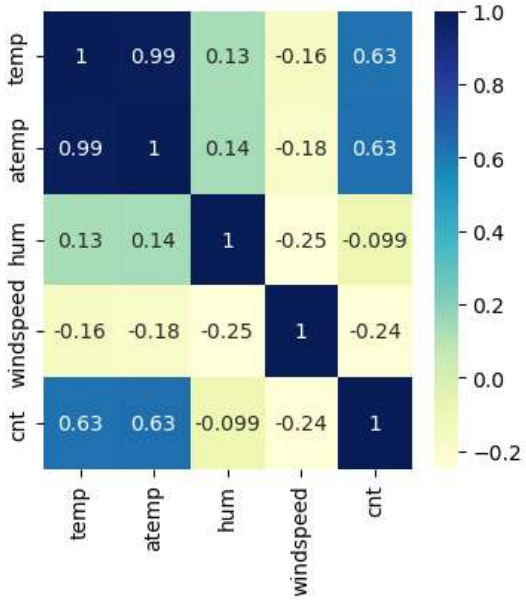
By setting drop_first=True, we are dropping dummy variables to one-level less, i.e., n-1 levels.

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

Looking at the pair-plot temp and atemp have higher correlation with cnt (count, which is the target variable).



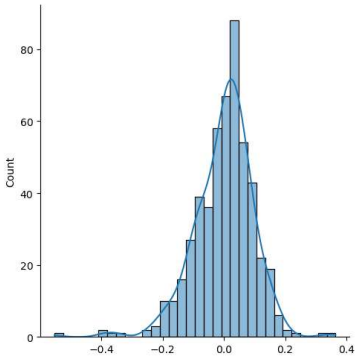
To be more precise we need to look at correlation heatmap, which tell us temp and atemp are the most correlated with a value of 0.63.



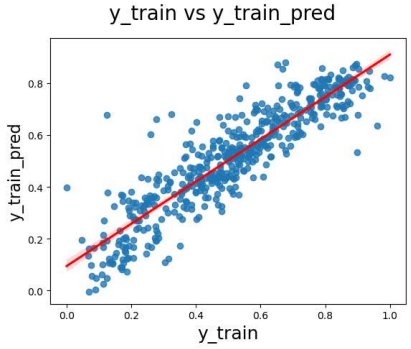
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validating the assumptions of linear regression after building the model can be done as follows,

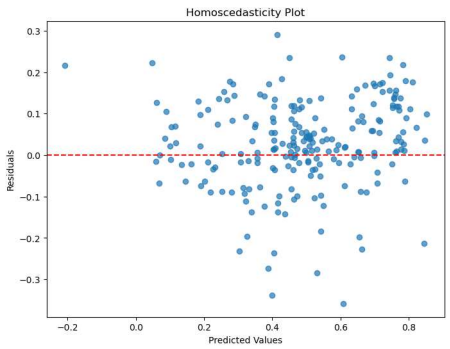
1. **Residual Analysis:**
   Computing the residuals, i.e., the differences between the actual and predicted values for the training data and then plotting a dist-plot with KDE curve to see a roughly symmetrical bell-shaped distribution exists.



2. **Linearity:**
   Creating a scatterplot of the predicted values against the actual values to visually check for linearity. The points should be roughly aligned around a 45-degree line.



3. **Homoscedasticity (Constant Variance):**
   Create a plot of residuals vs. fitted values. The spread of residuals should be roughly constant across all levels of the predicted values. If the spread widens or narrows systematically, it's an indication of heteroscedasticity.



4. **Independence of Errors:**
   Examine autocorrelation plots of residuals (e.g., ACF and PACF plots) to check for any patterns or correlations in the residuals over time, especially if the data has a time series component.

5. **Collinearity:**
   Calculating the Variance Inflation Factor (VIF) for each predictor variable to check for multicollinearity. High VIF values (typically greater than 5) may indicate collinearity issues.

	Features	VIF
2	windspeed	3.96
3	Spring	3.67
1	workingday	3.27
4	Summer	2.62
7	Jan	2.13
0	yr	1.87
6	Feb	1.84
9	May	1.67
14	Mist	1.56
12	Mon	1.55
8	Jun	1.28
10	Nov	1.23
5	Dec	1.21
11	Sep	1.21
13	Light Precipitation	1.08

6. **Cross-Validation:**
   Perform cross-validation on the training set to assess the model's predictive performance on unseen data. Cross-validation helps identify overfitting and generalization issues.

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

The below are the top 3 contributing factors explaining the demand of shared bikes,

#	Feature	Coefficient	Remarks
1	Light Precipitation	-0.3108	Negative Impact
2	yr	0.2474	Positive Impact
3	Spring	-0.1906	Negative Impact

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

   Linear Regression is a machine learning algorithm, which comes under Supervised Machine Learning, to model the relationship between dependent (target) and one or more independent variables (features) by fitting a linear equation.

   Model Assumptions: It assumes a linear relationship between the independent and dependent variable.

   a. **Simple Linear Regression (SLR):** If there is a single independent variable, then its called Simple Linear Regression.

Equation: Y = β0 + β1 * X

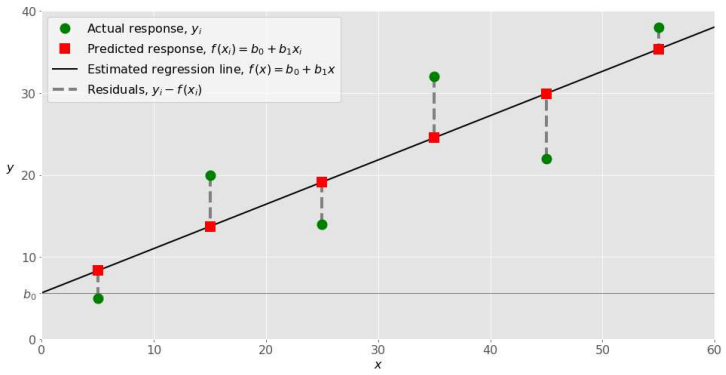
Where,

Y = Target

X = Single feature

β0 = Intercept

β1 = Slope



Metrics:

i. **R-squared:** It is the coefficient of determination that measure the proportion of the variance in the dependent variable. It can range from 0 – 1.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where,

RSS = Residual Sum of Squares

TSS = Total Sum of Squares

ii. **Residual Standard Error:**

It quantifies the average amount by which the ovserved values deviate from the predicted values.

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

Where,

n = number of data points

y_i = actual observed value of i-th data point

$\hat{y}_i$ = predicted value of i-th data point

b. **Multiple Linear Regression (MLR):** If there are more than one independent variables, then its called Multiple Linear Regression.
Equation: Y = β0 + β1 * X1 + β2 * X2 + … + βn * Xn + ε
Where,
Y = Target
X = Single feature
β0 = Intercept
β1 = Coefficient of 1st variable
βn = Coefficient of nth variable

Metrics:

i. **Adjusted R-squared:**
It's a modified version of R-squared, penalizing the inclusion of irrelevant variables.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

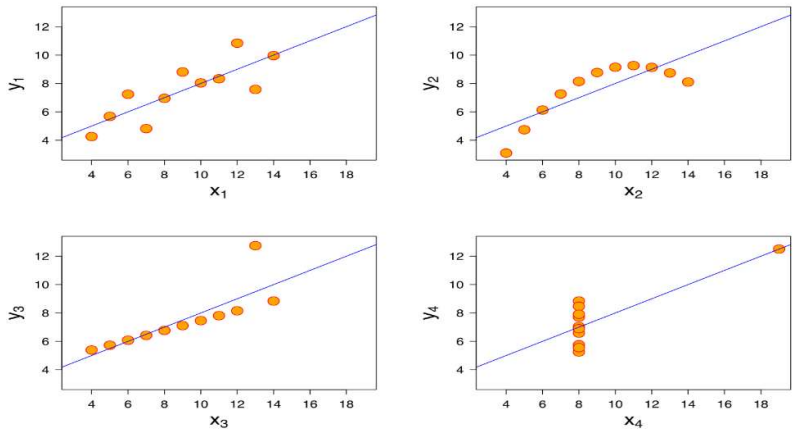
Where,
N = count of data points
P = predictors count

j. **Variance Inflation Factor (VIF):**
It measures the extent of multicollinearity among the independent variables. Typically >5 is considered bad and could be a potential candidate for dropping.

$$VIF = \frac{1}{1 - R^2}$$

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was created by Francis Anscombe to show the importance of plotting data rather than fully depending on summary statistics to take decision or direction.



Dataset I: Linear relationship between x & y **Model:** Linear Regression	**Dataset II:** Non-linear relationship between x & y **Model:** Quadratic function
Dataset III: Linear relationship between x &, y, with outlier that affects linear regression. **Highlight:** handling outliers	**Dataset IV:** Two distinct groups of data points. **Highlight:** Data distribution and subgrouping

Key takeaways:

a. It underscores the need to visualize data to understand its underlying structure and relationships effectively.
b. It demonstrates the limitations of relying solely on summary statistics, such as mean, variance, and correlation, to assess data.
c. It emphasizes the importance of considering the context and distribution of data in statistical analysis.
d. It serves as a reminder that different datasets can have the same summary statistics but require different models for accurate analysis.

3. What is Pearson's R?

Pearson's R or simply "r" is a statistical measure that quantifies the strength between two continuous variables. For non-linear or complex cases with more than two, this coefficient cannot be used.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Where,
n = number of data points
x_i & y_i = individual data values
$\bar{x}$ & $\bar{y}$ = means of x and y values

a. **Direction:** Sign of the coefficient indicates the direction of relationship, viz., Positive being increasing together or Negative being opposite to each other
b. **Strength:** The absolute value of coefficient shows the strength of relationship
c. **Range:** Between -1 to 1, with
   0: no linear relationship
   1: perfect positive linear relationship
   -1: perfect negative linear relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling: is the pre-processing step to transform data into standardized or normalized format. Which basically means attaching the whole set of values in dataset onto a common scale.

Need:

Avoiding Algorithm Biases: As some machine learning algorithms are sensitive to the scale of the input features, more valued features might get more weight. Which can lead to biased or suboptimal model performance.

Improving Convergence: It can help optimization algorithms converge more quickly and reliably when minimizing a cost or loss function.

Interpretability: Scaled data is easier to interpret since the coefficients in models represent the strength of the relationship between the variables.

Types:

Normalized Scaling (Min-Max Scaling):

In normalized scaling, data is transformed to a specific range, often [0, 1] or [-1, 1].

Formula: $\boldsymbol{X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}}$

Where,

X_{min} & X_{max} = minimum and maximum values in dataset

Normalized scaling ensures that the data falls within a specified range and maintains the relative relationships between the data points.

Standardized Scaling (Z-score Scaling):

In standardized scaling, data is transformed to have a mean of 0 and a standard deviation of 1.

Formula: $X_{std} = \frac{X - \mu}{\sigma}$

Where,

μ = mean (average) of the data

σ = standard deviation

Standardized scaling centers the data around the mean and scales it according to the spread of the data.

Differences:

Range:

Normalized scaling restricts the data to a specific range, such as [0, 1], while standardized scaling does not constrain the range. Standardized data can have negative values and values greater than 1 or -1.

Transformation:

Normalized scaling transforms the data to a common scale based on the minimum and maximum values.

Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

Importance:

Normalized scaling is suitable when preserving the original range is important, while standardized scaling is more appropriate when the mean and variance are critical.

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

VIF (Variance Inflation Factor) can approach or become infinite when there is perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity is a severe form of multicollinearity in which one or more independent variables can be exactly predicted from a combination of other independent variables. In other words, it indicates a linear dependency between two or more independent variables.

$$VIF = \frac{1}{1 - R^2}$$

If R^2 becomes 1, then the denominator becomes 0, which will lead to infinite VIF

Key Reasons:

Perfect Linear Relationship: Perfect multicollinearity occurs when one or more independent variables in the regression model can be expressed as an exact linear combination of others. For example, if you have two variables, X1 and X2, and X2 is exactly equal to 2*X1, this represents a perfect linear relationship, and VIF will be infinite.

Perfect Prediction: In the context of perfect multicollinearity, it is possible to perfectly predict one independent variable from the others. When you calculate the correlation matrix or the coefficient matrix in the regression, you will encounter problems like singular matrices or division by zero, which can lead to infinite VIF values.

Failed Regression Model: When there is perfect multicollinearity, the regression model cannot be estimated properly. The standard least squares method used for regression analysis cannot provide unique

coefficient estimates because the relationships among the variables are too predictable. Therefore, the software may not be able to compute VIF values properly.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Q-Q (Quantile-Quantile Plot): is a type of plot of quantiles of 2 distribution against each other. The pattern of points in the plot is used to compare the distributions. If the 2 distributions are similar or linearly related then the points will like on the line y=x, commonly called as 45-degree reference line. If the points are far from the reference line, the conclusion can be made that the datasets are from different distribution. Below is the q-q plot for Normal Distribution.

Use / Importance in Linear Regression: When training and test datasets are received from different source, we can use Q-Q plot to conclude that both are from populations with same distribution.

Advantages:

a. It can be used with varying sample sizes also.
b. We can use it to test many distributional aspects like shifts in location, shifts in scale, changes in symmetry and the presence of outliers.

[End of Document]