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?**

The dataset includes several categorical variables such as **season**, **year (yr)**, **holiday**, **weekday**, **workingday**, **weather situation (weathersit)**, and **month (mnth)**, which were visualized using boxplots. The impact of these variables on the dependent variable (bike demand) is as follows:

- **Season**: The Fall season (category 3) shows the highest median demand, indicating that bike rentals were highest during this period. In contrast, Spring (category 1) exhibited the lowest demand.
- **Year (Yr)**: Bike rentals were higher in 2019 compared to 2018, with a noticeable increase in the number of users in the later year.
- **Holiday**: Demand for bike rentals tends to decrease during holidays.
- **Weekday**: Bike demand remains relatively stable throughout the week, with little variation in rentals across weekdays.
- **Workingday**: The "Workingday" boxplot shows that most bookings occur between 4,000 and 6,000 rentals, suggesting a consistent median demand regardless of whether it's a working day or not.
- **Weather Situation (Weathersit)**: Bike rentals are non-existent during heavy rain or snow, highlighting the adverse impact of such weather. The highest rental counts were observed when the weather was clear or partly cloudy.
- **Month (Mnth)**: Bike demand peaked in September, while December saw a decline. This aligns with weather patterns, as heavy snowfall in December likely caused a drop in rentals.

This analysis suggests that certain seasons, weather conditions, and months have a significant impact on bike demand, with clear weather and fall months showing higher demand.

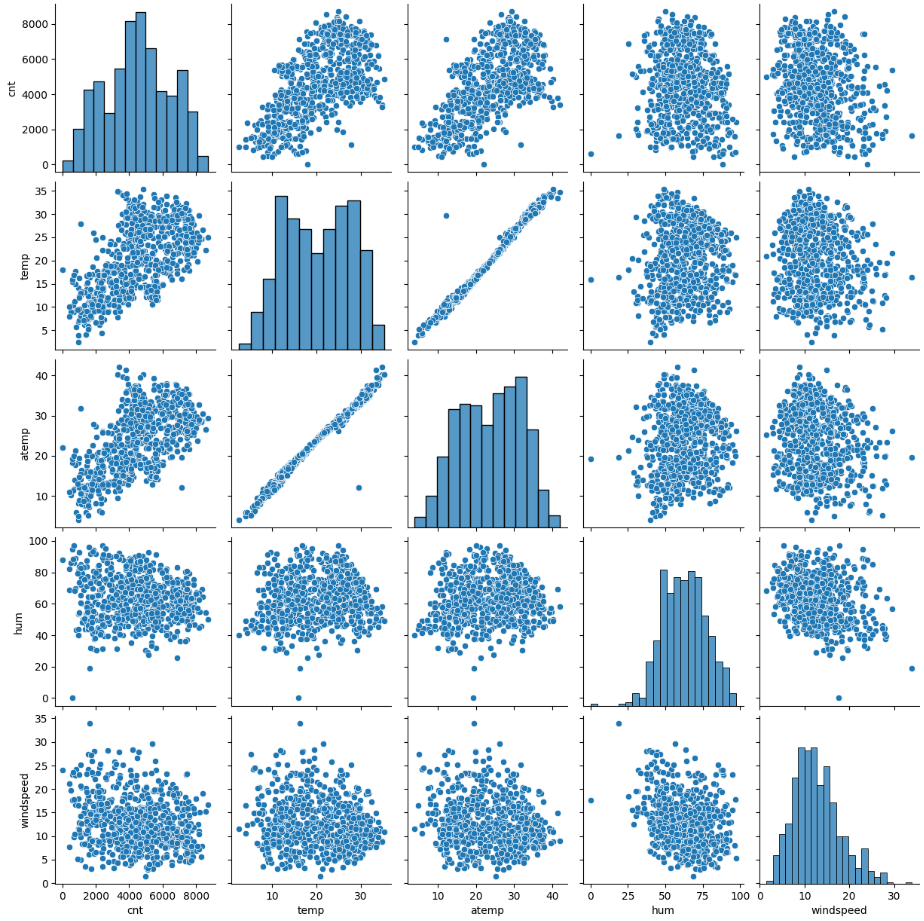
2. **Why is it important to use drop_first=True during dummy variable creation?**

Using drop_first = True is crucial because it prevents the creation of an extra column during the dummy variable transformation, which helps eliminate multicollinearity between the dummy variables.

When dealing with a categorical variable with **n** levels, only **n-1** dummy variables are required to represent it. This is because one category can be inferred from the others. For instance, if a categorical column has three possible values (e.g., "furnished", "semi_furnished", and "unfurnished"), by creating two dummy variables (e.g., "furnished" and "semi_furnished"), the third category ("unfurnished") can be automatically inferred when both dummy variables are 0. Therefore, there is no need for a third dummy variable to represent "unfurnished".

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

The two numerical variables, **"temp"** (temperature) and **"atemp"** (apparent temperature), exhibit a high correlation with the target variable, **"cnt"** (bike demand).

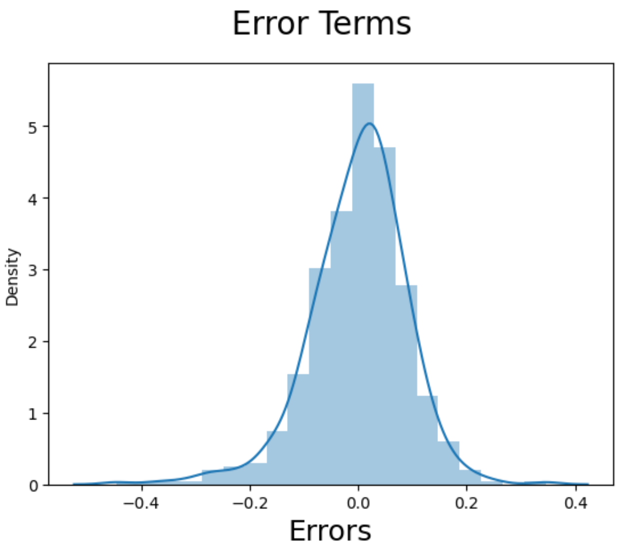


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

To validate the assumptions of Linear Regression, we conducted the following tests:

a. **Linear Relationship**: There should be a linear relationship between the independent and dependent variables. To assess this, we visualized the numerical variables using a pairplot to check if the variables exhibit linearity (as referenced in the pairplot above).

b. **Normal Distribution of Residuals**: The residuals should follow a normal distribution with a mean of 0. To validate this assumption, we plotted a distplot of the residuals to examine whether they follow a normal distribution.



c. **Multicollinearity**: Linear regression assumes minimal multicollinearity among the independent variables. Multicollinearity arises when the independent variables are highly correlated with each other. To assess this, we calculated the **Variance Inflation Factor (VIF)**, which quantifies the extent to which each feature variable is correlated with the others in the model.

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

The top 3 significant features are:

1. temp - coefficient : 0.438
2. yr - coefficient : 0.234
3. weathersit_Light Snow & Rain - coefficient : -0.291

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for regression tasks, where the goal is to predict a continuous output variable (y) based on one or more input variables (x). It is primarily employed to model the linear relationship between variables and for forecasting purposes.

The fundamental concept behind linear regression is to find a line that best represents the data points, minimizing the distance (or error) between the line and the actual data points. This line is described by the equation:

$y=\theta0+\theta1x$

where:

- $\theta0$ is the intercept (the value of y when x is 0),
- $\theta1$ is the slope (the change in y for a unit change in x).

These parameters, $\theta0$ and $\theta1$,, are known as the coefficients or parameters of the model.

To determine the optimal values for $\theta0$ and $\theta1$, a cost function is used to measure how well the line fits the data. A commonly used cost function is **Mean Squared Error (MSE)**, which calculates the average of the squared differences between the actual values (y) and the predicted values (y'):

$MSE= 1 /n \sum(y-y')2$

where:

- n is the number of data points,
- y is the actual value,
- y' is the predicted value.

The objective is to minimize the MSE by adjusting $\theta0$ and $\theta1$. This optimization can be performed using various methods, such as **gradient descent**, the **normal equation**, or through libraries like **scikit-learn**.

Linear regression can also be extended to multiple input variables, resulting in a multiple linear regression model. The equation in this case becomes:

$y=\theta0 + \theta1x1 + \theta2x2 + \cdots + \theta nxn$

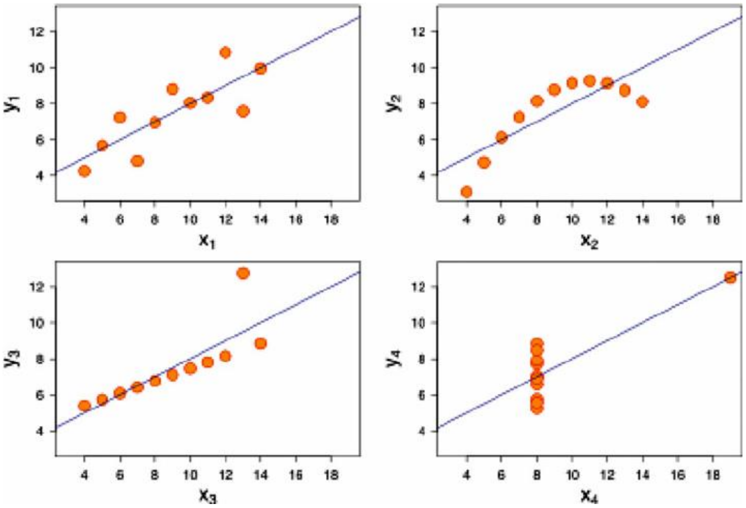
However, there are limitations to linear regression. It assumes a linear relationship between the input variables and the output variable, which may not always hold. Additionally, the model can be sensitive to **outliers** and **multicollinearity** (high correlation among the predictor variables), which can negatively impact its performance and accuracy.

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets that share nearly identical statistical properties, yet display very different distributions and behaviors when plotted. The purpose of the quartet is to highlight the importance of visualizing data before drawing conclusions and to demonstrate how outliers and influential data points can affect statistical results.

- The first scatter plot (top left) shows a clear linear relationship between the variables, suggesting a simple linear correlation.
- The second plot (top right) does not exhibit a normal distribution. While a relationship between the variables exists, it is not linear.
- The third plot (bottom left) shows a linear distribution but is influenced by a single outlier. This outlier has enough impact to distort the regression line, causing the correlation coefficient to drop from 1 to 0.816, despite the overall linearity of the data.
- The fourth plot (bottom right) demonstrates a case where a single high-leverage point is sufficient to create a high correlation coefficient, despite the other data points showing no discernible relationship between the variables. This highlights how a single influential data point can disproportionately affect the correlation and regression analysis.

Anscombe's Quartet serves as a reminder that statistical summaries alone can be misleading, and it's essential to examine the data visually before making inferences.



3. What is Pearson's R?

Pearson's r is a numerical measure that quantifies the strength and direction of the linear relationship between two variables. Its value ranges from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. In simpler terms, Pearson's r helps answer the question: "Can we represent the data with a straight line?" It shows how closely the data points follow a straight line when plotted.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

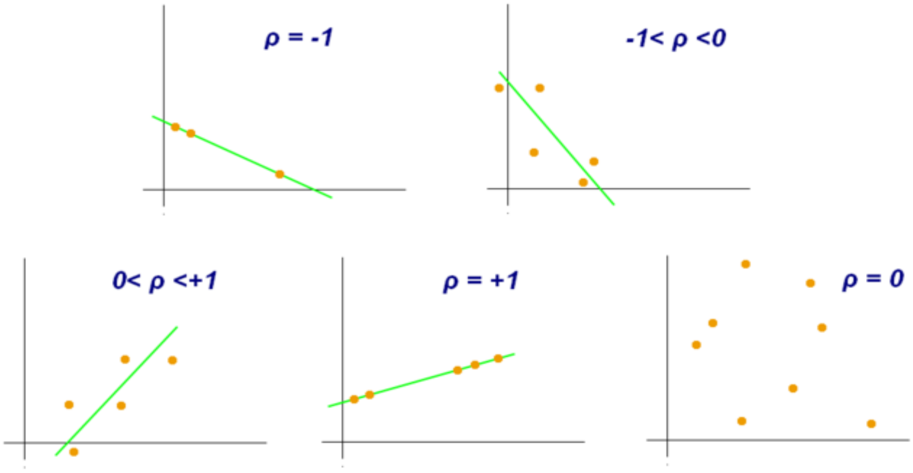
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As shown in the graph below, **r = 1** indicates a perfect positive linear relationship, meaning the data points align exactly along a straight line with a positive slope. **r = -1** represents a perfect negative linear relationship, where the data points form a straight line with a negative slope. **r = 0**, on the other hand, suggests that there is no linear relationship between the variables, meaning the data points do not follow any clear straight-line pattern.



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

Feature scaling is a technique used to normalize or standardize the range of independent variables (features) in a dataset. It is applied during the data preprocessing phase to handle varying scales in the data. Without feature scaling, machine learning algorithms might give more weight to features with larger values and less weight to those with smaller values, regardless of their units, which can lead to biased results.

- **Normalization** is typically used when the data does not follow a Gaussian (normal) distribution. It rescales the data to a specific range, often between 0 and 1. This method is especially useful for algorithms that do not assume any specific distribution, such as **K-Nearest Neighbors (K-NN)** and **Neural Networks**.
- **Standardization**, on the other hand, is beneficial when the data approximates a Gaussian distribution, although this is not a strict requirement. Standardization transforms the data to have a mean of 0 and a standard deviation of 1. Unlike normalization, it does not

bound the data to a specific range, meaning that outliers in the data will not be as heavily influenced by standardization.

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

The **Variance Inflation Factor (VIF)** measures the extent to which the variance of a regression coefficient is inflated due to multicollinearity among the predictor variables. A VIF value of infinity indicates perfect correlation between variables. Essentially, VIF provides a quantitative measure of how strongly each feature is correlated with the others in the model. It is a critical parameter to evaluate in order to assess the quality and stability of a linear regression model. High VIF values suggest that multicollinearity could be a problem, which may distort the model's estimates.

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

The term **R-1** refers to the R-squared value of an independent variable, which measures how well this variable is explained by the other independent variables in the model. If an independent variable can be perfectly explained by the other variables, its R-squared value will be 1, indicating perfect correlation. In this case, the VIF would be calculated as:

VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity". Thus, a VIF of infinity suggests perfect multicollinearity, where one variable is perfectly predicted by others.

The numerical value of the VIF indicates how much the variance (or standard error squared) of a coefficient is inflated due to multicollinearity. For example, a VIF of 1.9 means that the variance of a particular coefficient is 90% larger than it would be if there were no multicollinearity — i.e., no correlation with other predictors.

Guidelines for interpreting the VIF:

- **VIF = 1**: No correlation with other variables.
- **1 < VIF ≤ 5**: Moderate correlation.
- **VIF > 5**: High correlation, which may indicate problematic multicollinearity.

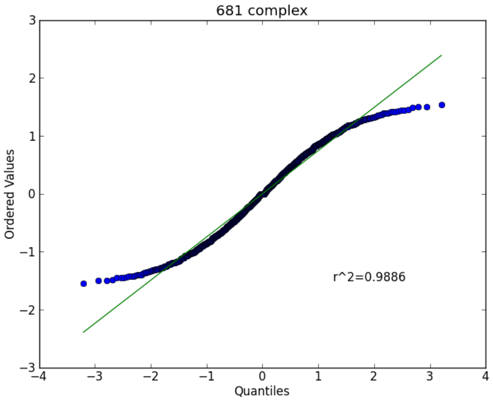
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the distributions of two datasets by plotting their quantiles against each other. It helps in assessing whether two datasets follow the same distribution. In a Q-Q plot, if the two datasets come from the same distribution, the points should form a straight line.

The Q-Q plot is useful for answering the following questions:

- Do the two datasets come from populations with the same distribution?
- Do the two datasets have similar location and scale? (i.e., similar central tendency and spread)
- Do the two datasets exhibit similar distribution shapes?
- Do the two datasets have similar tail behaviors? (i.e., the way extreme values or outliers are handled)

In essence, the Q-Q plot visually assesses whether two datasets share common statistical characteristics.


681 complex
Ordered Values
Quantiles
r^2=0.9886