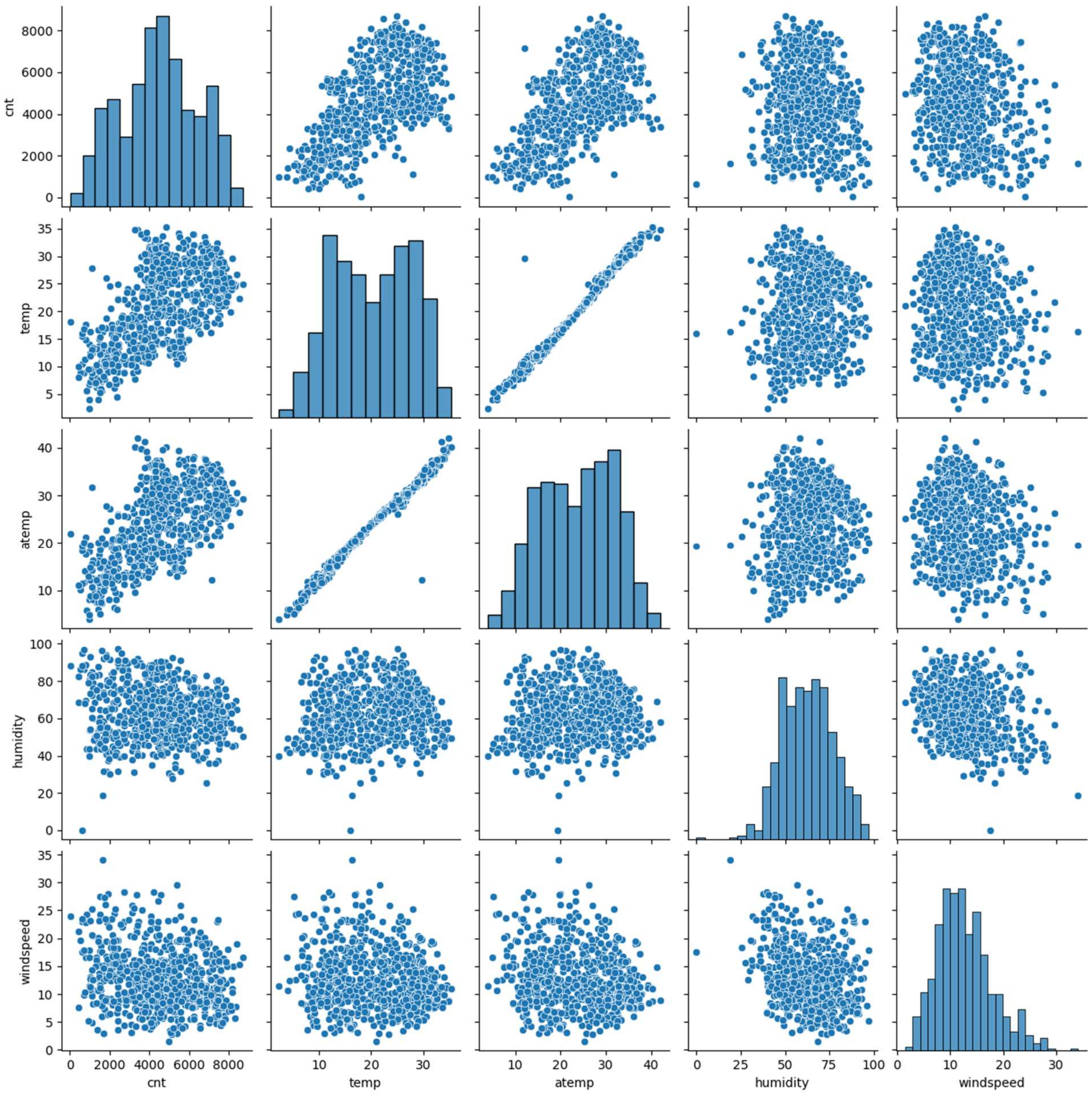
Assignment-based Subjective Questions – answered by Huang-Yin Tso

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - When the weather is clear, bike tends to be used which is reasonable.
    - Peoples use bike more in fall and not in spring which temperature ('atemp') factor can explain that temperature in springer is lower
    - Peoples use Boom bike more in year 2019 compared to 2018, peoples might be more aware of Boom bike in 2019 although the usage seems equal between working and non-working day.
2. **Why is it important to use drop_first=True during dummy variable creation?**

    Using drop_first=True when creating dummy variables is important to prevent multicollinearity in the model. Multicollinearity happens when dummy variables are too closely related, making it hard for the model to understand each variable's effect. By dropping the first dummy column, we avoid this problem and make sure the variables in the model are independent and do not overlap. This keeps the model stable and accurate.

(Q3. Continue in the next page)

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



it appears that temp has the highest correlation with cnt (the target variable)

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

My validation was carried out as follows:

- The error terms in the model are normally distributed.
- Residual analysis on the training data shows that the residuals follow a normal distribution.
- Confirming that the linear regression assumptions is valid.
- Variables were included or removed from the model based on VIF and p-values to avoid multicollinearity.

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

Temperature (temp):

- A coefficient value of 0.4624 indicates that a unit increase in the temperature variable results in an increase in bike hire numbers by 0.4624 units. This positive coefficient suggests that higher temperatures are associated with more bike bookings, likely due to more favorable weather conditions.

Weather Situation (weathersit_Light_snowrain)

- A coefficient value of -0.3060 indicates that a unit increase in the presence of weather conditions like light snow or rain decreases bike hire numbers by 0.3060 units. This significant negative coefficient reflects that adverse weather conditions lead to fewer bike hires.

Year

- A coefficient value of 0.2332 shows that a unit increase in the year variable results in an increase in bike hire numbers by 0.2332 units. This positive impact implies that over time, bike bookings have been increasing, potentially due to rising awareness, improved infrastructure, or greater popularity of bike-sharing services.

General Subjective Questions

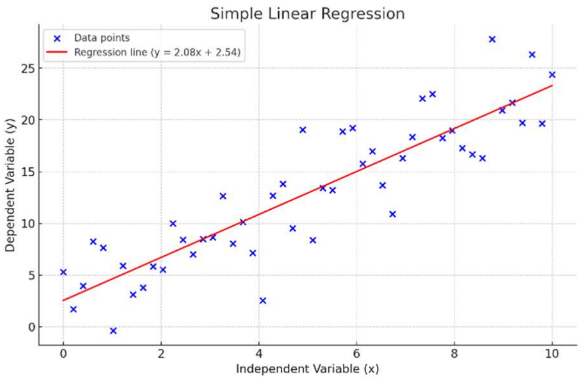
1. **Explain the linear regression algorithm in detail.**

- Linear regression is a simple machine learning algorithm used to model the relationship between a dependent variable (the one you want to predict) and one or more independent variables (the predictors). Here's how it works:
- Linear regression tries to fit a straight line through the data points in such a way that the line represents the best estimate of the dependent variable for given values of the independent variables.
- For one independent variable, this is called simple linear regression. When there are two or more independent variables, it's called multiple linear regression.
- The general equation of a linear regression line is:

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

And by training the model, it tries to find the best values to minimize the difference between the predicted and actual values. So it is very important that the model is valid.

Principles of linear regression is linearity, that relationship between the independent and dependent variables is linear. The error terms should be normally distributed. The independent variables should not be too closely related to each other. And observations should be independent of each other.



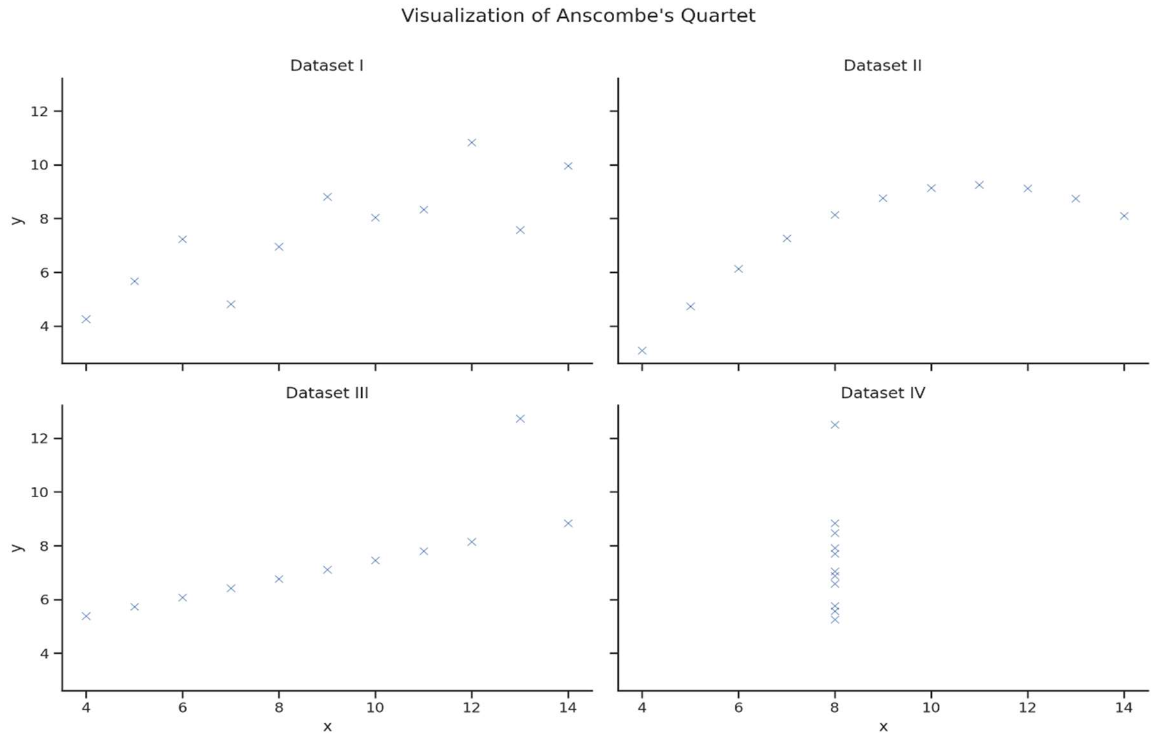
2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet was created to show that summary statistics (like the mean and correlation) can be the same for different sets of data, but the data can look very different when plotted. It teaches us the importance of visualizing data, not just relying on numbers.

When to Use Anscombe's Quartet:

- To explain why visualizing data is important: It helps show that statistics can sometimes hide the real pattern or problem in the data.
- To remind us to check data plots before analysis: It can be used to show that summary statistics alone are not enough to understand data.
- When teaching about data analysis: It's a good example for showing that different datasets can have the same statistics but behave differently when graphed.

In simple terms, use the lesson from Anscombe's quartet when you want to stress that "seeing" the data is as important as calculating statistics.



3. **What is Pearson’s R?**

Pearson’s R, also called the Pearson correlation coefficient, is a number that shows how strong the relationship is between two variables. It tells us if the relationship is positive, negative, or if there is no relationship at all.

Key Points:

Value Range: Pearson’s R can range from -1 to +1.

+1 means a perfect positive relationship (as one variable increases, the other also increases).

-1 means a perfect negative relationship (as one variable increases, the other decreases).

0 means no relationship between the variables.

Purpose: It helps us understand how well two variables move together. For example, we can use Pearson’s R to see if temperature and ice cream sales are related. This can help with sales prediction

Use: If the value is close to +1 or -1, it shows a strong relationship. If the value is near 0, it shows a weak or no relationship.

Pearson’s R is used in statistics and data analysis to check relationships between variables.

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

Scaling is the process of changing the range of data to make sure all features have a similar scale. And why we need this because some algorithms perform better when input features have similar ranges or units.

Normalized scaling means data changing range is between 0 and 1 (or -1 to 1). It is good to use when data is not normally distributed and needs to be in a fixed range.

Standardized scaling changes data so it has a mean of 0 and a standard deviation of 1. It is good to use when data follows a normal distribution and needs to be centered.

Scaling helps make sure all features have the same importance in a model.

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

The **Variance Inflation Factor (VIF)** measures how much one variable in a model is affected by other variables. A high VIF means there is a strong correlation between the variables, leading to **multicollinearity**.

To improve the model and avoid multicollinearity, I checked the VIF values of all variables. When I found variables with high VIF (indicating strong correlation with other variables), I removed them one by one. This process helps the model better understand the independent effect of each variable.

Explanation of Dropping High VIF Variables:

To improve the model and avoid multicollinearity, I checked the VIF values of all variables. When I found variables with high VIF (indicating strong correlation with other variables), I removed them one by one. This process helps the model better understand the independent effect of each variable.

By dropping variables with high VIF:

- Reduced Multicollinearity: This made sure the model's predictions were more reliable.
- Improved Model Stability: The coefficients became more accurate because the model was not confused by overlapping information.

After removing the high VIF variables, I built a new model with lower VIF values, which improved its performance and accuracy.

This approach ensures the model is simpler and more interpretable, as each variable contributes unique information.

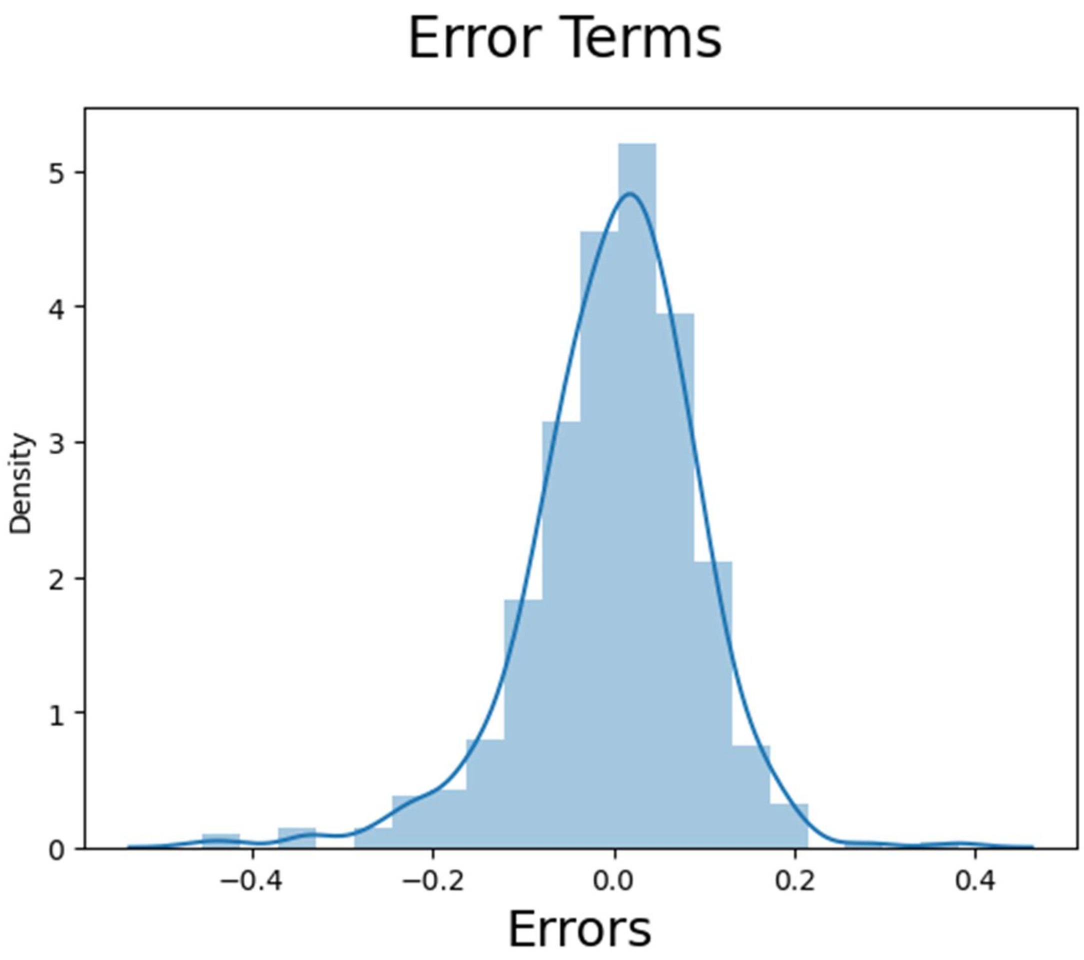
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A Q-Q plot (Quantile-Quantile plot) is a graph that shows if a dataset follows a specific distribution, usually a normal distribution.

- Checking Residuals: It helps check if the residuals (errors) from a linear regression model are normally distributed.

- How It Looks: The plot shows your data points compared to a line representing a perfect normal distribution.
- If the points follow the line closely, the data is normal.
- If the points move away from the line, the data is not normal and might be skewed or have outliers.
- A Q-Q plot helps confirm that the linear regression model meets the normality or valid assumption, making sure predictions are reliable.

In short, a Q-Q plot is important for checking if the errors in a linear regression model are normal, which helps make better predictions.


Error Terms
Density
Errors