# Assignment-based Subjective Questions

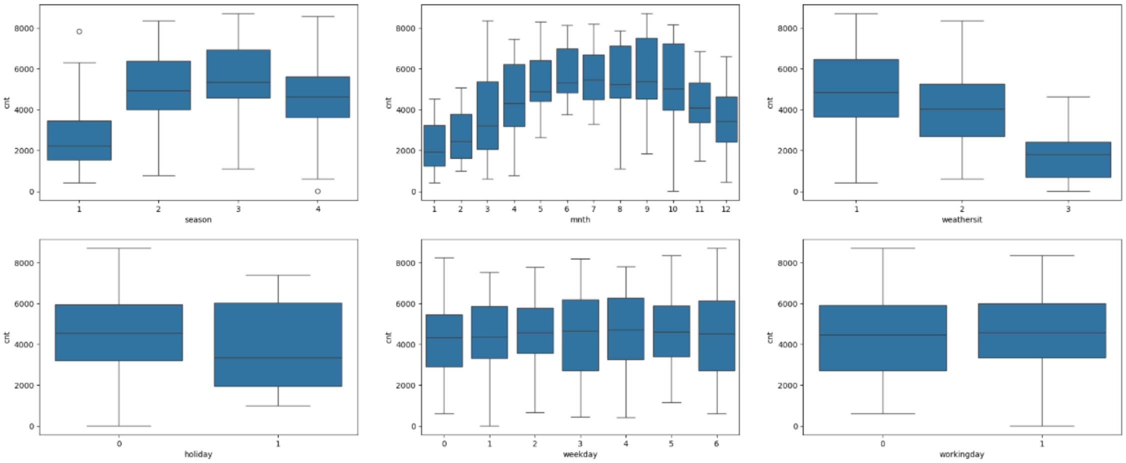
**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)

**Total Marks**: 3 marks (Do not edit)

**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

There are several categorical variables, including season, month, year, weekday, working day, and weathersit, which have a significant impact on the dependent variable ‘cnt’. The figure below illustrates the correlation among these variables.

Variables are visualized using Box plot:



---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation?  (Do not edit)

**Total Marks:**  2 marks (Do not edit)

**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

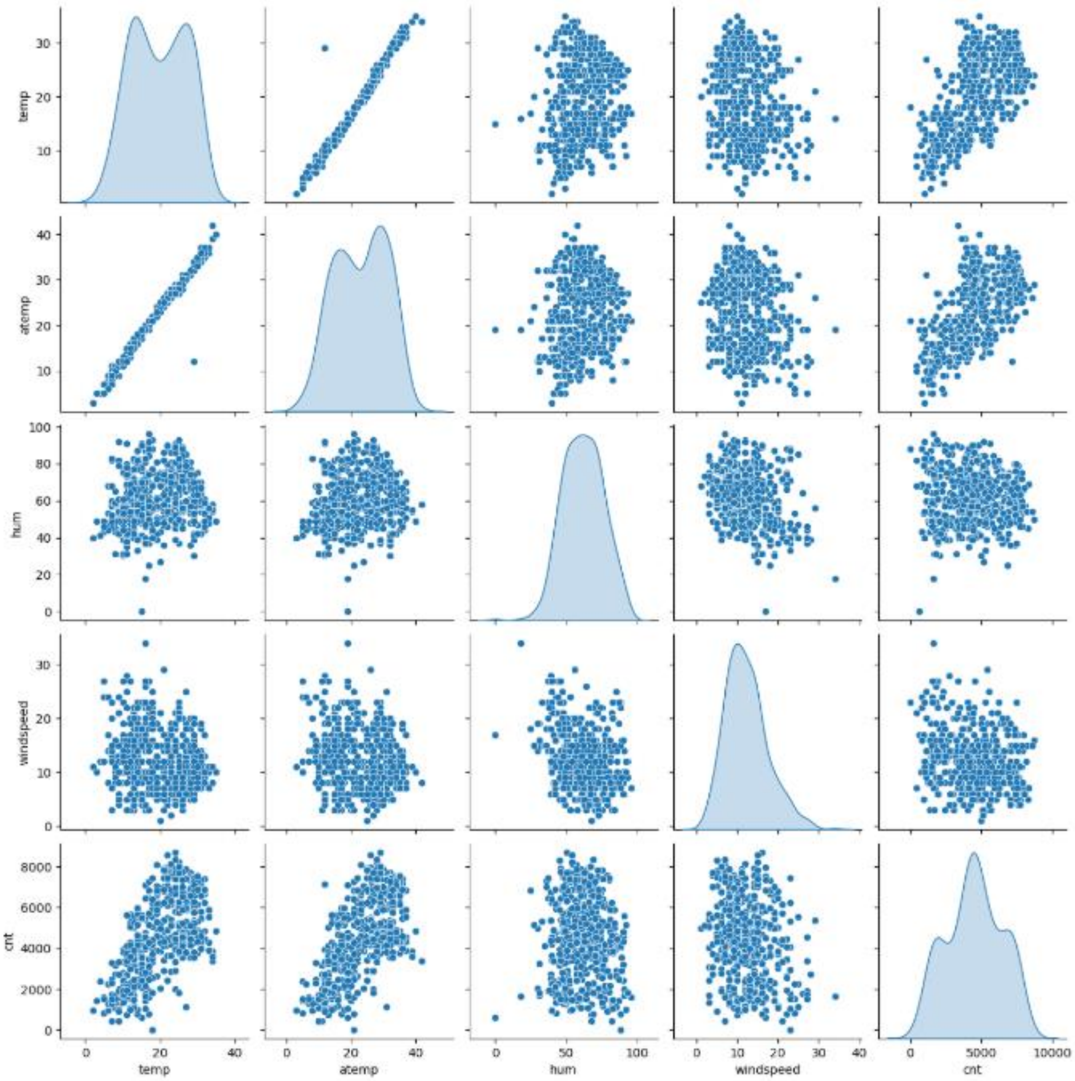
Using drop_first=True during dummy variable creation is important for the following reasons:

1. **Avoiding Multicollinearity**: When you create dummy variables for a categorical feature, you introduce multiple columns that represent the different categories. If you include all these dummy variables in your model, one of them becomes perfectly predictable from the others, leading to multicollinearity. This can distort the model’s coefficients and make it harder to interpret the results. By setting drop_first=True, you drop the first category (often called the reference category), which helps prevent multicollinearity.
2. **Ensuring Identifiability**: When one category is dropped, the remaining categories can be interpreted relative to the dropped category, which serves as a baseline. This allows the model to understand the effect of each category in comparison to this baseline, ensuring clear and meaningful coefficient estimates.
3. **Improving Model Stability**: Removing the first dummy variable makes the model more stable and efficient. Including all dummy variables can increase variance and make the estimation process less reliable, especially when the number of categories is large.

In summary, setting drop_first=True helps in creating a more interpretable and statistically sound model by avoiding redundancy and multicollinearity.

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
**Total Marks:** 1 mark (Do not edit)
**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.



**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

After building the linear regression model on the training set, validating the assumptions of linear regression is crucial to ensure the reliability and validity of the model. The key assumptions of linear regression are:

1. Linearity: The relationship between the independent variables and the dependent variable is linear.
2. Independence of Errors: The residuals (errors) should be independent of each other.
3. Homoscedasticity: The variance of residuals should be constant across all levels of the independent variables.
4. Normality of Errors: The residuals should be approximately normally distributed.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

The top three features that have a significant impact on explaining the demand for shared bikes are temperature, year, and season (weather).

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find a linear equation that best fits the data.

---

**1. Simple Linear Regression (Single Predictor)**
The model for simple linear regression is:
$y=\beta 0+\beta 1x+\epsilon$

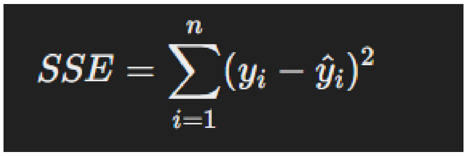
Where:
y = dependent variable
x = independent variable
$\beta 0$ = intercept
$\beta 1$ = slope (coefficient)
$\epsilon$ = error term (residuals)
The goal is to find $\beta 0$ and $\beta 1$ that minimize the error between actual and predicted values.

**2. Multiple Linear Regression (Multiple Predictors)**
For multiple predictors, the model extends to:
$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp+\epsilon$

Where x1,x2,…, xp are the independent variables.

**3. Estimating Coefficients (Ordinary Least Squares)**

Linear regression uses **Ordinary Least Squares (OLS)** to estimate coefficients by minimizing the sum of squared residuals:



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

**4. Making Predictions**

Once the model is trained, predictions are made by plugging input values into the equation:

y^=β0+β1x1+⋯+βpxp

**5. Model Evaluation**

**R-squared ($R^2$)**: Measures how well the model explains the variance in the target variable.

**Mean Absolute Error (MAE)**: Average of the absolute errors.

**Mean Squared Error (MSE)**: Average of the squared errors.

**Root Mean Squared Error (RMSE)**: Square root of MSE.

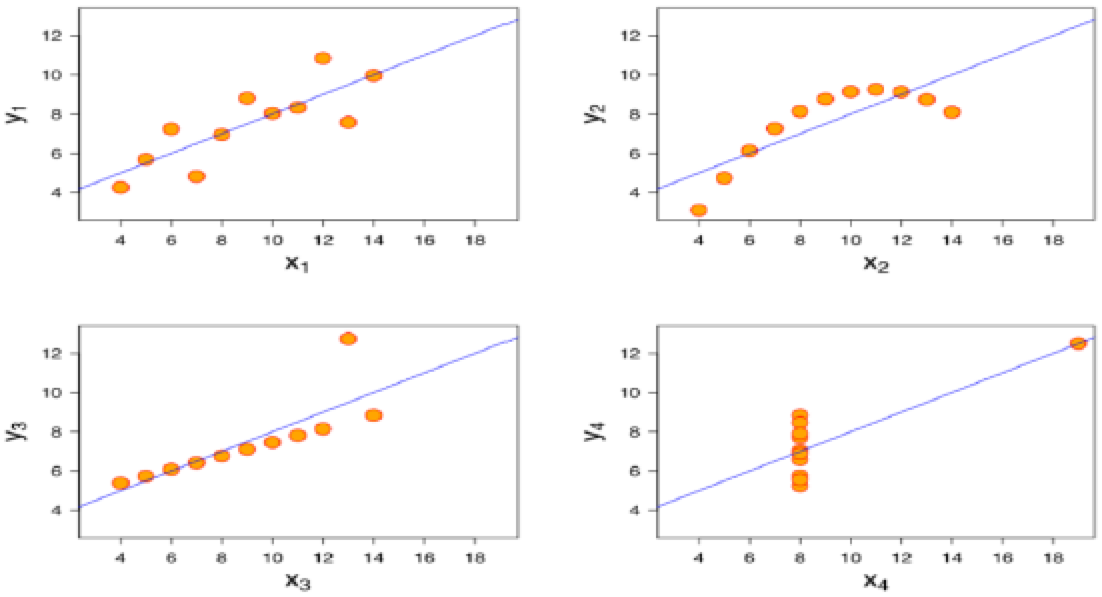
**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet refers to a collection of four datasets that are nearly identical in basic descriptive statistics but exhibit distinct characteristics that can mislead regression models if they are used without further analysis. Despite having the same mean and variance for both the x and y variables across all four datasets, their distributions are quite different and they appear noticeably distinct when plotted on scatter plots. The quartet was created to highlight the importance of visualizing data before drawing conclusions and building models, as well as to demonstrate how outliers or peculiar observations can affect statistical properties. Although the datasets have the same statistical summary (mean, variance), their underlying distributions and visual representations differ significantly.



● 1st data set fits linear regression model as it seems to be linear relationship between X and y

● 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.

● 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.

● 4th data set has a high leverage point means it produces a high correlation coefficient.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's r is a simple and widely used measure for determining the strength and direction of the linear relationship between two variables. However, it's important to remember that it only detects linear relationships and can be affected by outliers, so it should be used alongside visualizations (like scatter plots) and other analyses to confirm the findings.

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

Where:

- xi and yi are individual data points of variables x and y.

- x¯ and y¯ are the means of the x and y variables, respectively.
- The numerator is the covariance between x and y, while the denominator normalizes the covariance by the standard deviations of x and y.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling refers to the process of transforming data to fit within a specific range or scale. It is an important data preprocessing step that helps adjust the data, making it more suitable for algorithms and speeding up computations. Collected data often includes features with different magnitudes, units, and ranges. If scaling is not applied, algorithms might give more weight to features with larger values and overlook smaller ones, leading to inaccurate models.

**Difference Between Normalized Scaling and Standardized Scaling:**

1. **Normalization** uses the minimum and maximum values of the features to scale the data, while **standardization** uses the mean and standard deviation.
2. **Normalization** is typically used when the features are on different scales, whereas **standardization** ensures that the data has zero mean and a unit standard deviation.
3. **Normalization** scales the values between a specific range, like (0, 1) or (-1, 1), while **standardization** does not restrict the values to a fixed range.
4. **Normalization** is sensitive to outliers, whereas **standardization** is not significantly affected by them.
5. **Normalization** is often applied when the distribution of the data is unknown, while **standardization** is preferred when the data follows a normal distribution.
6. **Normalization** is also known as **scaling normalization**, whereas **standardization** is referred to as **Z-score normalization**.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) measures the relationship between a given independent variable and all other independent variables in the model. It helps identify how much the variance of a regression coefficient is inflated due to multicollinearity. The formula for calculating VIF is as follows:

A VIF value greater than 10 is considered very high, while a VIF above 5 should also be scrutinized and examined carefully.

A very high VIF indicates a strong correlation between two independent variables. In the case of perfect correlation, the $R^2$ value will be 1, which causes the VIF formula $1/(1- R^2)$to approach infinity. To resolve this issue, one of the variables causing this perfect multicollinearity should be removed from the dataset.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. It helps assess whether a set of data could have come from a specific theoretical distribution, such as Normal, Exponential, or Uniform.

Additionally, a Q-Q plot can be used to determine whether two distributions are similar. If they are, the plot will typically show a more linear pattern. The linearity assumption is often validated through scatter plots. For **linear regression**, it's essential that all variables follow a multivariate normal distribution. This assumption is best checked using either histograms or Q-Q plots.

**Importance of Q-Q Plot in Linear Regression:**

In the context of linear regression, when working with training and test datasets, a Q-Q plot can confirm whether both datasets come from populations with the same distribution.

**Advantages of Q-Q Plot:**

- It can be applied to datasets of any size.
- The plot can reveal various distributional characteristics such as shifts in location, scale, symmetry, and the presence of outliers.

**Uses of Q-Q Plot for Comparing Two Datasets:**

- To check if both datasets originate from populations with a common distribution.
- To verify if both datasets have the same location and scale.
- To compare the shape of the distributions in both datasets.
- To assess the tail behavior of the two datasets