# 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)

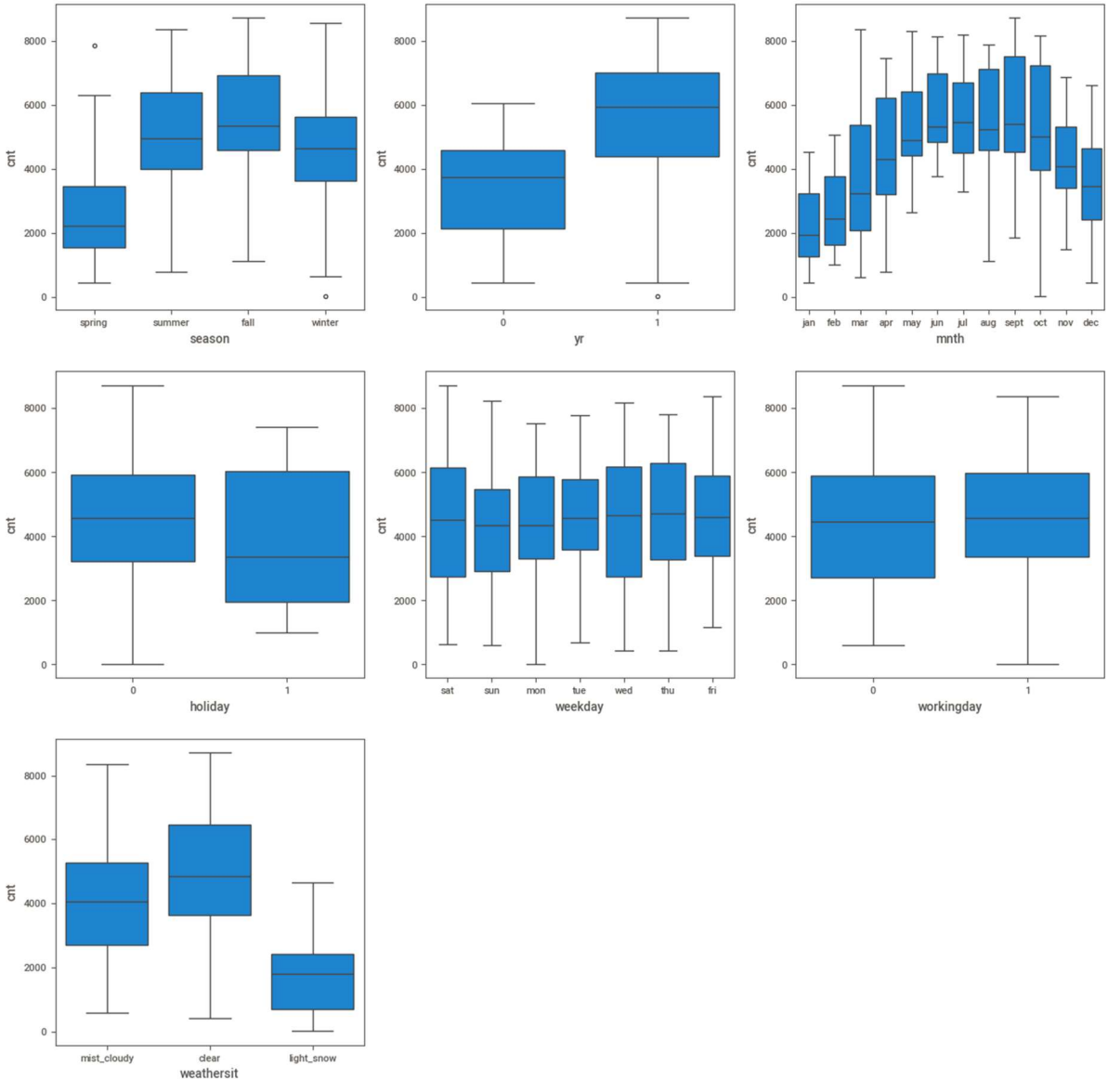
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

Yes, from analysis performed, it is found that categorical variables impact the target\dependent variable.

This is set of overall features which are predicting the target variable - ['yr', 'atemp', 'windspeed', 'spring', 'winter', 'dec', 'jul', 'mar', 'nov', 'sept', 'light_snow', 'mist_cloudy'].

---

- Season and month - The demand for shared bikes tends to vary by season and month, with higher demands observed during warmer months and lower demand during winter months.
- Weather situation – affects bike sharing, as clear weather leads to increased usage.
- Year wise usage - indicates increase in demand for 2019



---

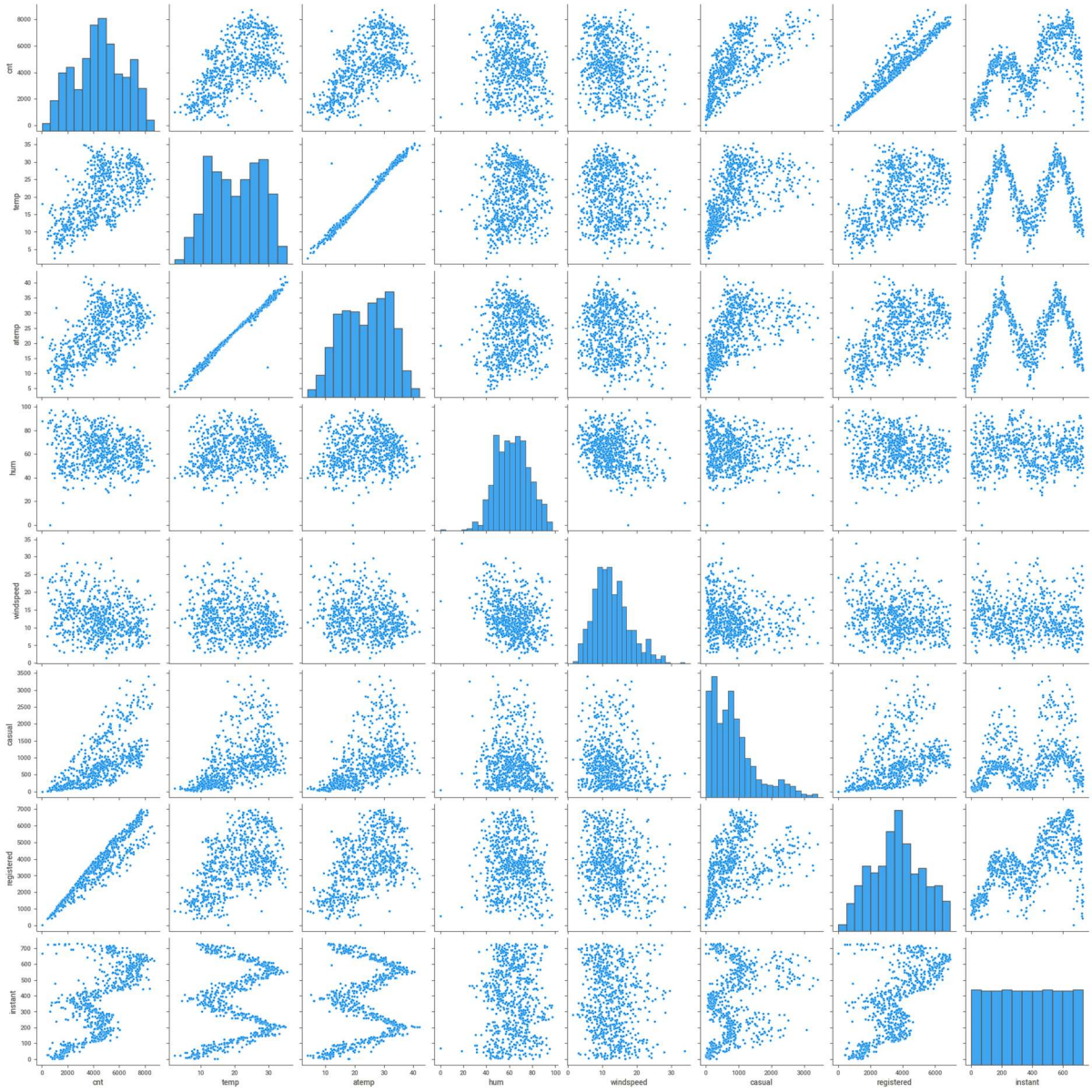
**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)
The drop_first=True parameter prevents multicollinearity in the model by avoiding the "dummy variable trap."

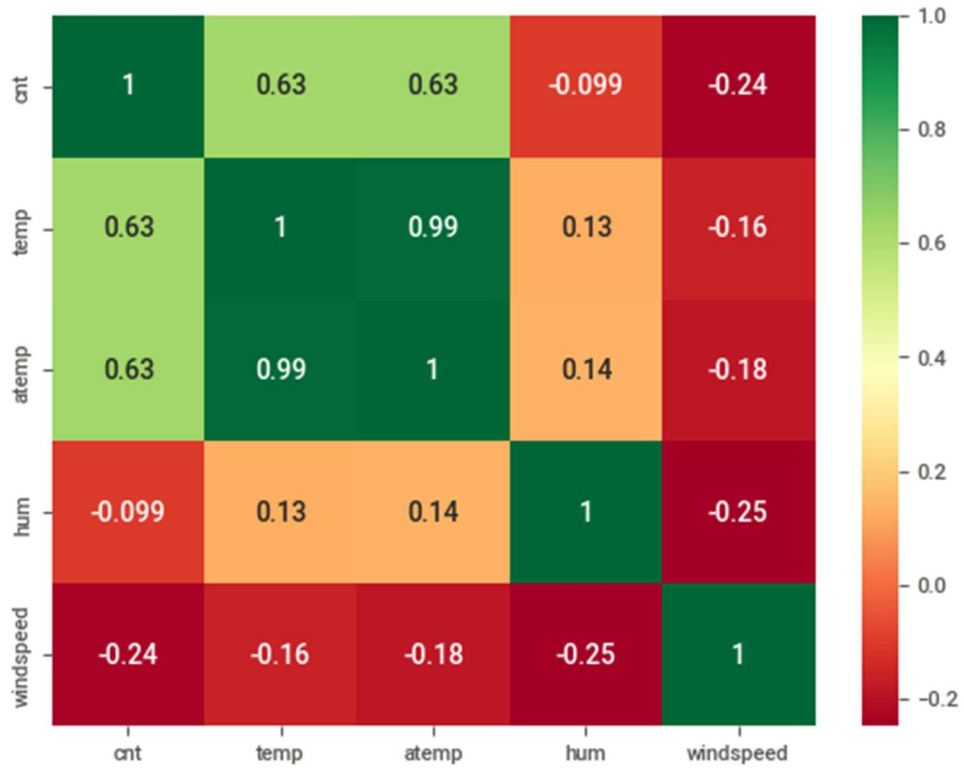
- The 'dummy variable trap' is a situation that arises when the model includes all categories for a variable, leading to perfect multicollinearity and makes it impossible for the model to effectively distinguish between the categories, thereby hindering the model's performance.
- By using drop_first=True, one category is dropped, making the dummy variables linearly independent, thereby improving model stability and interpretability.

---

**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)
Two variables have high correlation-

1. The variable "registered" shows a very high correlation with the count of bike rentals, indicating that most rentals are made by registered users rather than casual users.
2. Temperature (temp) and "feels like" temperature (atemp) demonstrate a strong correlation with the target variable. This suggests that warmer weather conditions tend to increase the demand for shared bikes.


cnt
temp
atemp
hum
windspeed
casual
registered
instant




---

**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 model on the training set you can validate the assumptions of Linear Regression using the following methods:

1. Residual Analysis: The residuals were plotted to assess their distribution. A normal distribution of residuals would indicate that the linear regression model is appropriate.
2. Homoscedasticity Check: The residual plot was checked to ensure that the errors were randomly scattered, which would indicate constant variance (homoscedasticity). A non-random pattern would indicate heteroscedasticity.
3. Normality of Residuals: The histogram of residuals was inspected to verify if they follow a normal distribution, which is important for valid inferences in linear regression.

---

**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)

Data belongs to US and as per the model summary following are the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. atemp(Coefficient: 3633.7761) positive contributors that increase the demand for shared bikes.
2. Yr(Coefficient: 2001.3503) positive contributors that increase the demand for shared

bikes.

3. light_snow(Coefficient: -2202.8805) is a significant negative contributor, as bad weather reduces the demand.

---

## 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 supervised learning algorithm used to predict a continuous dependent variable based on one or more independent variables. It aims to establish a linear relationship between the independent and dependent variables.

1. **Equation**: For linear regression model is:

   $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn + \epsilon$

   where:

   a. y is the predicted value of the dependent variable.

   b. β0 is the intercept(β called as beta).

   c. β1,β2,...,βn are the coefficients for the independent variables(β called as beta).

   d. x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn are the independent variables.

   e. ϵ\epsilonϵ is the error term, which represents the unexplained variance in y.

2. **Goal**: Find the values of β0,β1,...βn that minimize the **residual sum of squares (RSS)**, representing the difference between the actual and predicted values of y.

3. **Least Squares Method**: Linear regression typically uses the **Ordinary Least Squares (OLS)** method to minimize the residual sum of squares. OLS tries to find the line of best fit by minimizing the squared differences between the observed data points and the values predicted by the model.

4. **Assumptions**:

   a. **Linearity**: The relationship between the dependent and independent variables should be linear.

   b. **Homoscedasticity**: The residuals should have constant variance across different levels of the independent variables.

   c. **No Multicollinearity**: Independent variables should not be highly correlated.

   d. **Normality**: The residuals should be normally distributed.

5. **Metrics**: To evaluate the performance of the linear regression model, standard metrics include:

a. **$R^2$**: Represents the proportion of variance in the dependent variable explained by the independent variables.

b. **Adjusted $R^2$**: Adjusts for the number of predictors in the model, preventing overfitting.

c. **Mean Squared Error (MSE)**: This measure measures the average of the squares of the residuals, indicating the model's accuracy.

---

**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** is a set of four datasets with nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line, but are vastly different when graphed. Francis Anscombe created it to highlights the dangers of relying solely on summary statistics without visualising the data. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation, Correlation coefficient, Linear regression line fit) but different graphical representation

1. Despite having similar statistical properties, the four datasets have very different distributions when plotted. For instance:
    - **Dataset 1**: Shows a typical linear relationship.
    - **Dataset 2**: Has a clear non-linear pattern.
    - **Dataset 3**: Shows a linear relationship but has an outlier that significantly affects the regression line.
    - **Dataset 4**: All points are almost identical except for one extreme outlier.
2. Anscombe's Quartet is a powerful testament to the necessity of visualizing data. It unveils the structure and uncovers patterns, relationships, and anomalies that summary statistics alone cannot reveal. Data sets which are identical over several statistical properties, may produce dissimilar graphs so visualizing is important. This understanding is crucial in the field of statistics.

---

**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, or the Pearson correlation coefficient, measures the linear correlation between two variables and widely used in statistics to quantify the degree to which two variables have a linear relationship.

- Range: Pearson's R ranges from -1 to 1.
    - R = 1: Perfect positive linear relationship.
    - R = -1: Perfect negative linear relationship.
    - R = 0: No linear correlation.
- Interpretation:
    - Positive R: Indicates that the other increases as one variable increases.
    - Negative R: Indicates that the other tends to decrease as one variable increases.
    - Magnitude: The closer the value is to 1 or -1, the stronger the linear relationship.
- Formula: R=Cov(X,Y)/ sigmaX. sigmaY

  Cov(X,Y) is the covariance between variables X and Y,

  sigmaX. sigmaY are standard deviation of X,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 is a technique for rescaling features so that they are on the same scale. This is especially important for models such as linear regression, support vector machines (SVM), and neural networks, where differences in feature scales can impact model performance.

<u>1. Reasons for Scaling:</u>
- Feature Consistency: Different features may have varying scales. Without scaling, models may assign undue importance to features on larger scales.
- Model Convergence: In algorithms like gradient descent, unscaled features can lead to slow convergence or may prevent the algorithm from converging altogether.
- Distance Metrics: In distance-based algorithms such as K-nearest neighbours (KNN), differing scales can result in misleading distance calculations.

<u>2. Normalized Scaling:</u>
- Normalization rescales features to a range of [0, 1] or [-1, 1].
- The formula used for normalization is: X' = (X-Xmin)/(Xmax-Xmin) where Xmax, Xmin is max and min range
- Normalization is useful to maintain the relationships between features with non-Gaussian distributions.

<u>3. Standardized Scaling:</u>
- Standardization rescales features to have a mean of 0 and a standard deviation of 1.

- Formula: X'= (X- Xmean)/sigma where Xmean is mean and sigma is standard deviation.
- Standardization is appropriate when the features are normally distributed, ensuring the data conforms to a standard normal distribution.

---

**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>
A Variance Inflation Factor (VIF) value can be infinite when perfect multicollinearity exists among the independent variables. Perfect Multicollinearity occurs when one independent variable is an exact linear combination of other independent variables.

Division by Zero: VIF is calculated as: VIF=1/(1−Ri square) where Ri square is the coefficient of determination for regression of variable i on all other independent variables. When Ri square = 1, meaning perfect correlation, the denominator becomes zero, making the VIF infinite.

Solution: When encountering an infinite VIF, removing or combining the highly correlated features is important to avoid perfect multicollinearity.

---

**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 for comparing the distribution of a dataset against a theoretical distribution, such as the normal distribution.

The construction of the plot requires the quantiles of the observed data to be plotted against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points will lie along a straight line.

Use in Linear Regression: The Q-Q plot is used to validate the assumption of normality for residuals in a linear regression model. A Q-Q plot can help identify whether the residuals deviate significantly from normality, indicating that transformations or other modelling approaches might be needed.

In summary, the Q-Q plot is a crucial diagnostic tool for validating the normality assumption in linear regression. It ensures that the inferences drawn from your model are valid and reliable. The Q-Q plot provides a visual method to check for deviations from normality, such as heavy tails (kurtosis)

or skewness. If the residuals are not normally distributed, you might consider transforming the data or using a different regression technique, such as logistic regression.

---