# 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)

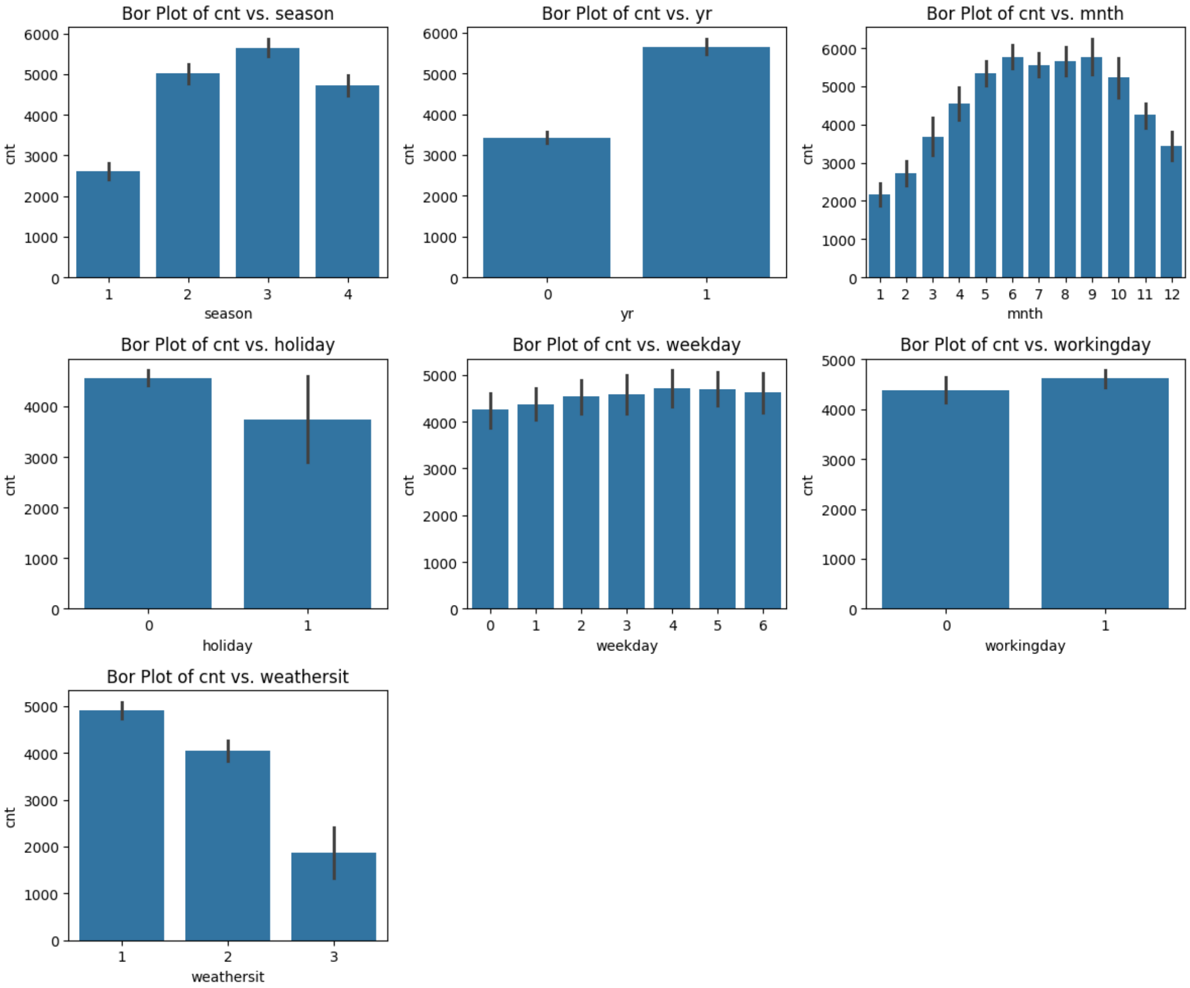
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

Bivariate analysis of categorial variables reveals that:

- **season**: The count of rentals varies across seasons. Season 3 (fall) seems to have the highest count, while season 1 (spring) has the lowest. This suggests that the season significantly affects the count.
- **yr**: The count of rentals is higher in year 1 (2019). This indicates that the count has increased over time.
- **mnth**: There is a clear seasonal trend, with the count peaking around months 6 to 9 (likely the summer months) and dropping in months 1 and 12 (likely winter months). This confirms a strong seasonal pattern in the data.
- **holiday**: The count is slightly lower on holidays (1) compared to non-holidays (0), suggesting that fewer people use the service on holidays.
- **weekday**: The count is consistent across all weekdays (0-6). This suggests that the day of the week does not significantly affect the count, possibly due to similar demand throughout the week.
- **workingday**: There is little difference in the count between working days (1) and non-working days (0). This might indicate that users rent similarly whether it's a working day or not, or the difference is negligible.
- **weathersit**: The count decreases as the weather situation worsens. Category 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) has the highest count, and category 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) has the lowest count. This suggests that weather significantly impacts the count, with fewer people engaging in the activity during poor weather conditions.

**In summary, the count variable is influenced significantly by seasonal factors, year, and weather; while holidays, weekdays, and working days have less impact.**

Plots created for Bivariate analysis of categorial variables is given below:



---

**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)

When creating dummy variables, using drop_first=True drops the variable's first category's dummy variable.

It is important to use this to avoid redundancy, multicollinearity and to help with model interpretation.

When dummies are created for a variable having n categories, n new dummy variables are created.

- This introduces **redundancy** since the values of n-1 dummy variables automatically determines the value of the nth dummy variable.
- This also introduces **multicollinearity**, since when n features are created it introduces a relationship between the features. This multicollinearity makes it difficult to estimate the coefficients, leading to unreliable coefficients being determined.

When dummies with drop_first=True, are created for a variable having n categories, n-1 new dummy variables are created.

- The dropped category becomes the **baseline category** and when performing linear regression, the coefficients for the remaining dummy variables are determined relative to this dropped category

Note that there are **exceptions to this** → when you want to manually determine the baseline category, then n dummies can be created, and the baseline category dummy variable can be manually dropped.

An example use of this will be if we are generating dummies for weather variable having three values- Rain, Clear, Sunny. In this case it is ideal to keep Clear as the baseline category since that is the default state, and rain and sunny are exceptions from / change from this default state. In this case we will create 2 dummy variables, and then manually drop the dummy variable for Clear.

This method also achieves the same outcome – avoids redundancy and multicollinearity. And helps with model interpretation.

---

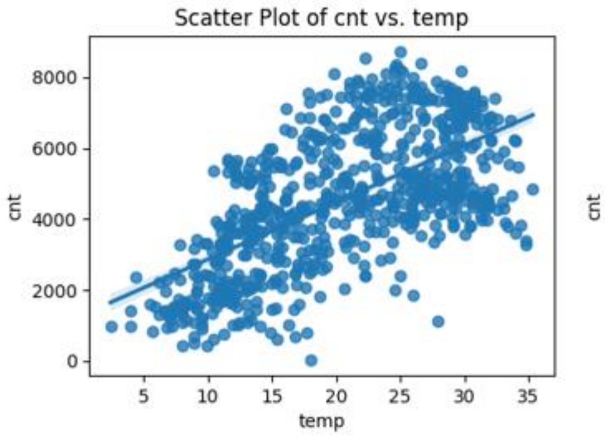
**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

Variable temp (Temperature) has the highest correlation with the target variable cnt. Refer plot below.

*Note that variable atemp also has a similar correlation with cnt, but we are dropping this variable since atemp itself has a high correlation with temp.*



---

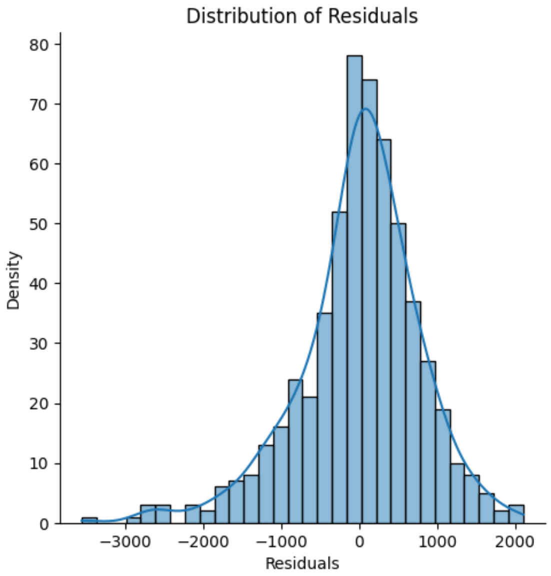
**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

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

**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of Linear Regression we validated as follows:

**Assumption 1: Errors are normally distributed** was validated using the distribution plot of residuals.



This plot assesses the normality assumption of the errors. A normal distribution is expected for the residuals.

**Observations:**

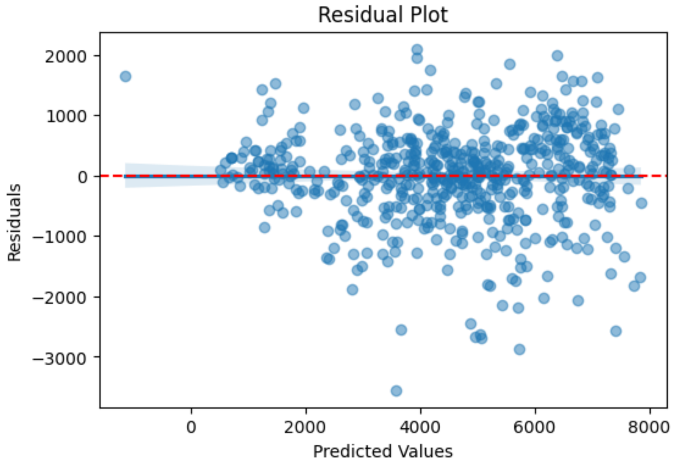
**Shape:** The histogram appears roughly bell-shaped, suggesting a normal distribution.

**Symmetry:** The distribution seems relatively symmetric around the mean (around 0).

**Outliers:** There are a few outliers, especially on the left tail. Outliers can indicate potential issues with the model or data points that might not fit the linear relationship well. But the amount of outliners here are within the acceptable range.

**Overall Assessment:** The distribution is generally consistent with normality.

**Assumption 2: Errors are independent of each other and Assumption 3: Homoscedasticity** were validated using the regplot of residuals.



This plot assesses the homoscedasticity assumption, which means the variance of the errors should be constant across different predicted values.

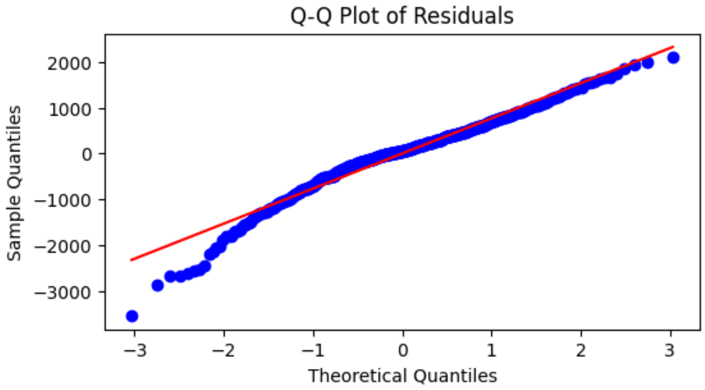
**Scatter:** The residuals seem to be scattered randomly around the horizontal line at 0. This is a good sign as it suggests there's no clear pattern in the spread of the residuals.

**Funnel Shape:** There doesn't appear to be a funnel shape, where the residuals spread out more as the predicted values increase or decrease. This is another positive indication.

**Overall Assessment:** The residual plot supports the homoscedasticity assumption and independent assumption. The residuals are evenly scattered, and there's no evidence of a changing variance based on the predicted values.

**Assumption 4: Linear Relationship** was validated using Bivariate analysis.

**Assessment was also done using Q-Q plot.**



This plot is another way to assess the normality assumption of the errors. It compares the quantiles of the observed residuals to the quantiles of a theoretical normal distribution.

**Linearity:** The points generally follow a straight line, suggesting that the residuals are normally distributed.
**Deviations:** There are some deviations from the line, particularly in the tails. These deviations indicate that the observed residuals might not perfectly align with the theoretical normal distribution.
**Overall Assessment:** While the Q-Q plot shows a general linear trend, the deviations from the line, especially in the tails, suggest that the normality assumption might not be perfectly met. This could be due to outliers or other factors that are skewing the distribution. But the skewness show is within acceptable range.

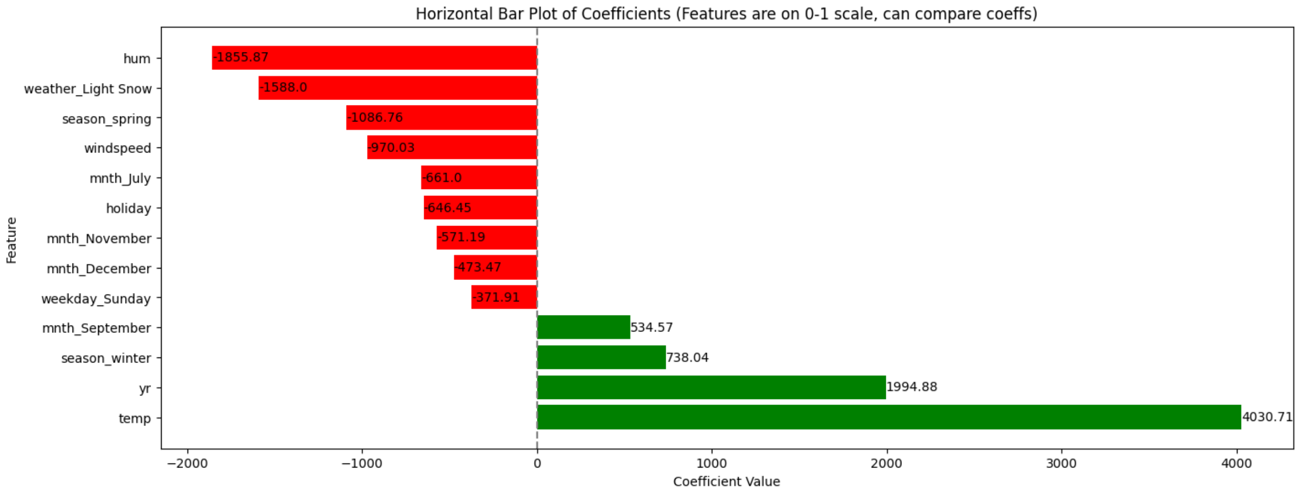
---

**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)

Top 3 features contributing significantly towards explaining the demand of the shared bikes are listed below in descending order of their impact:

1. temp (Temperature)
2. yr (year)
3. hum (humidity)

Below is a visual comparison of how each feature is contributing to explain the demand of shared bikes.



# 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)

Linear Regression is used to model relationship between one or more independent variables and a dependent variable. Independent variables are also called features, and dependent variables are also called the target variable.
It models this relationship by fitting a straight line (in the case of simple linear regression) or a hyperplane (in case of multiple linear regression) that best explains the data.

The equation for linear regression is:

$$y = \beta 0 + \beta 1*X1 + \beta 2*X2 + ... + \beta n*Xn + \epsilon$$

where:
y is the dependent variable
X is the independent variable
β0, β1, β2….. βn are the coefficients (β0 is also called the intercept)
ϵ is the error term.

For linear regression to perform well the below key assumptions need to be satisfied:

1. Linearity: relationship between independent and dependent variables should be linear
2. Independence: data points should be independent of each other
3. Homoscedasticity: The variance in error (also called the residuals) should be constant
4. Normality of residuals: The error (also called the residuals) should be normally distributed.

The primary objective of linear regression is to find values of the coefficients such the error in predicting the target variable is minimized. This error is usually used to quantify this error is MSE (Mean Squared Error). To find a best fit line for linear regression, values for coefficients are found such that MSE is minimized.

This minimization problem can be solved in two ways:

1. Closed form method, where in the normal equation is directly solved to compute the optimal values of the coefficients. This works well for simple linear regression. But for multiple linear regression with many more independent variables this this approach does not scale well.
2. The second approach is Gradient Descent, wherein we randomly default the coefficients and then vary them using the derivative of the MSE function as a guide such that the coefficients are adjusted in the direction that reduces error. This process is repeated until it converges to the minimum of the cost function.

Once the model is built, and we have the optimal coefficients, it is evaluated using the metrics such as Adjusted R-Squared, Mean Squared Error (MSE), Root Mean Square Error (RMSE) and mean absolute error (MAE).

The coefficients found can be used to evaluate relative and independent impact of independent variables on the dependent variable.

Linear Regression could be extended and enhanced by modifying the cost function. Popular enhancements are Ridge Regression, Lasso Regression and Elastic Net regression.

---

**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)

Anscombe's Quartet is a set of four separate datasets created by statistician Francis Anscombe in 1973. These datasets demonstrate the importance of visualizing data before analyzing it and caution against relying solely on summary statistics when drawing conclusions.

Each dataset in Anscombe's Quartet has nearly identical descriptive statistics such as Mean, Variance, Correlation, etc…

But when plotted using a scatter plot these can be clearly seen to be very different distributions.

Anscombe's Quartet servers as a caution for solely relying on descriptive statistics, and demonstrates that visualization and detailed analysis of data, its relationships and patterns is necessary before drawing conclusions.

---

**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)

Pearson's R is a statistic measure of the linear relationship between two variables. It quantifies the strength and direction of the relationship.

Formula:

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

$r$ = 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

It can vary from -1 to 1.
1 means the variables have a perfect positive linear relationship.
0 means the variables have no relationship and are independent of each other.
1 means the variables have a perfect negative linear relationship.

---

**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)

Scaling is the process of adjusting features to ensure that they are on a similar scale. Scaling ensures that all features contribute equally to the model.

Scaling is performed to improve convergence in gradient-descent based algorithms and also help with interpretability of the models.

Two common types of scaling are normalized scaling and standardized scaling.

Formulas:

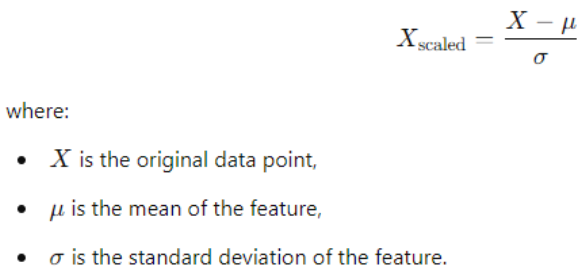
Normalized Scaling →

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

- $X$ is the original data point,
- $X_{\min}$ is the minimum value in the feature,
- $X_{\max}$ is the maximum value in the feature.

Standardized Scaling →



$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where:

- $X$ is the original data point,
- $\mu$ is the mean of the feature,
- $\sigma$ is the standard deviation of the feature.

Normalized scaling scales features to a fixed range which is usually 0 of 1.
Whereas Standard Scaling scales data such that after scaling the feature has a mean of 0 and standard deviation of 1.

Normalized scaling is more sensitive to outliers compared to standard scaling.

These methods are essential for preparing data for machine learning algorithms that are sensitive to feature scaling.

---

**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)

Formulae for VIF is VIF = 1 / (1 – R-squared)
Where R-squared is the coefficient of determination.

When there is perfect multicollinearity, i.e. when on feature is perfectly predicted by the other features, R-squared will be 1.

Substituting in the equation above we can see that when R-squared =1 VIF will be infinity.

Overall, VIF = Infinite can be used as a signal to perfect multicollinearity.

Methods to address this problem includes, dropping collinear variables and regularization techniques such as Ridge Regression and Lasso Regression.

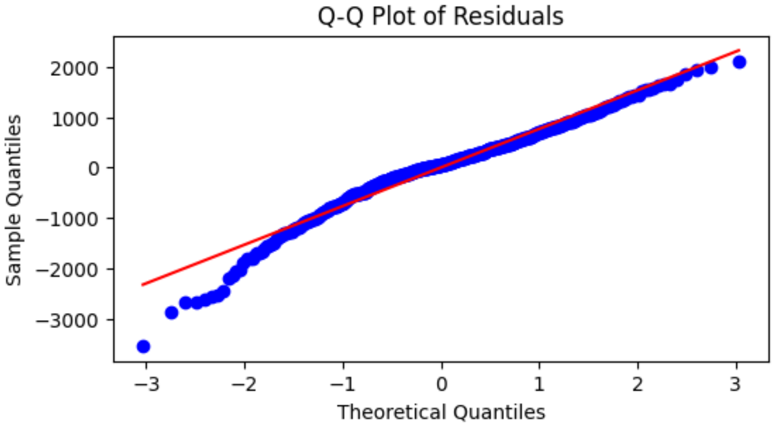
---

**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)

Q-Q plot (quantile-quantile plot) is plot used to compare the distribution of a dataset with another distribution. It plots and compares quantiles of a dataset against a theoretical distribution. If both distributions are, the plotted points will align on a straight line along the diagonal.

Q-Q plots can be used to check for normality, identify outliers and model interpretation. For example if the Q-Q plot of residuals aligns along the diagonal, this indicates that the assumption of normality is reasonable. If there is deviations along the bottom or the top end of the diagonal, it would indicate outliers.

Q-Q Plot from BoomBikes assignment →


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles