Assignment-based Subjective Questions

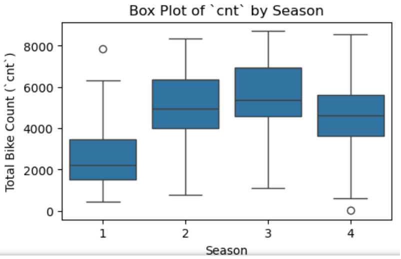
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

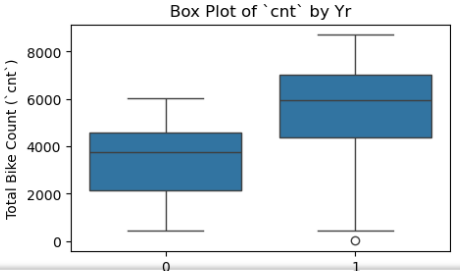
The interpretation of effect of categorical variables on dependent variable ‘cnt’ is given below:

a. **Season (Categories - 1: Spring, 2: Summer, 3: Fall, 4: Winter)**



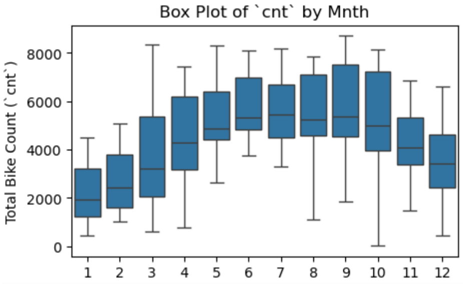
Season impacts the bike rentals quite well. Warmer seasons like Fall and Summer have favorable biking conditions so higher rental counts, then moderate demand in winter and lower in spring.

b. **Year Yr (Categories - 0: 2018, 1: 2019)**



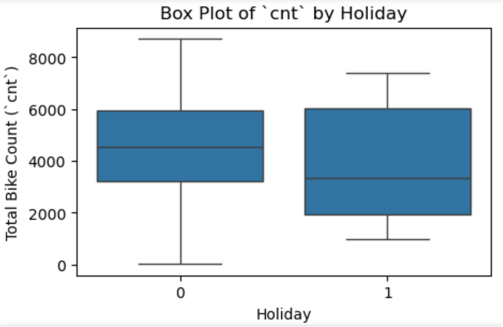
This categorical variable captures the annual trend in bike demand. The bike rental demand increased in year 2019 drastically as compared to 2018. There could be many reasons to it like more awareness, better service etc.

c. **Months mnth (Categories – 1 to 12)**



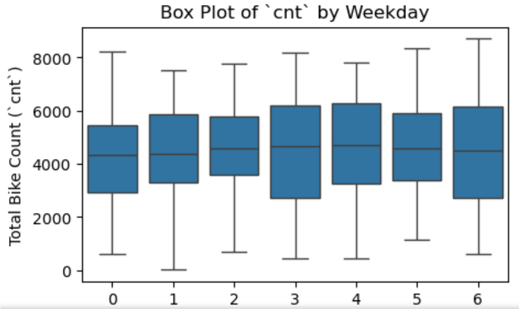
Month data confirms the demand observations in season category. The bike demand rises during Fall and Summer months like June to September and fall in colder months like January, February.

d. **Holiday (Categories – 0: Not a holiday, 1: Holiday)**



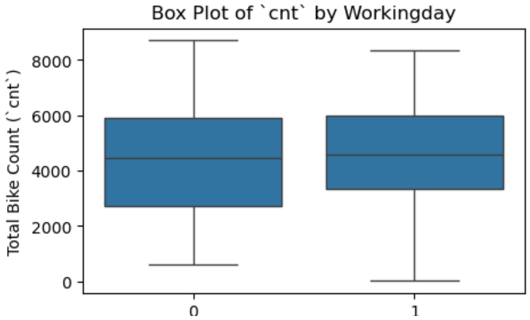
The bike rental demand on non-holiday is slightly higher than on holiday.

e. Weekday (Categories – 0 to 6 Sunday to Saturday)



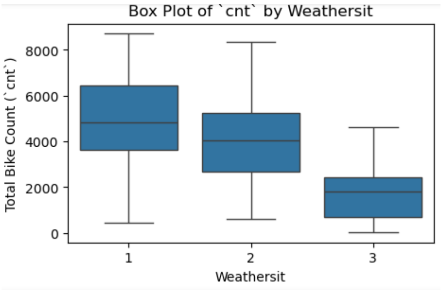
The bike demand is relatively consistent across all weekdays, suggesting that bike usage doesn't vary significantly based on the day of the week.

f. **Working Day (Categories - 0: Weekend or holiday, 1: Working day)**



The bike count is similar on both working days and non-working days, suggesting that bike usage doesn't vary significantly based on whether it's a working day or not. The demand on weekday could be slightly higher in case of weekday though when we compare median.

g. **Weather Situation (weathersit) (Categories –**
1 : Clear or partly cloudy, 2 : Mist or cloudy, 3 : Light snow or rain, 4 : Heavy rain, snow, or fog)



Weather significantly affects bike rentals, as people are more likely to rent bikes in clear weather and less likely in adverse conditions like rain or snow. Bike rental is highest for weather situation 1, followed by weather situation 2, and lowest for weather situation 3 and null for 4.

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 we create dummy variables for categorical variable, for n dummy variables, we generate n binary columns to represent each category. However, logically when we know value of n-1 variables, we can deduce the value of nth column. E.g When 1- Holiday, we can deduce that when value not 1 i.e. 0 is a non-holiday column. This means that these columns are linearly dependent. This introduces multicollinearity which can affect the model stability and interpretation.

So by setting up 'drop_first=True', we generate n-1 dummy variables for n categories.

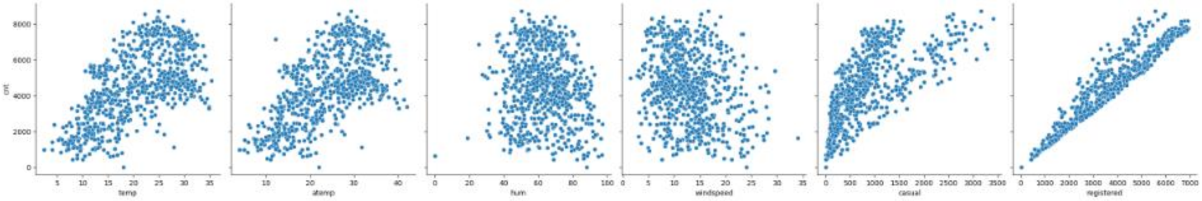
This step is important to avoid multicollinearity and make model more interpretable.

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 pairplot of various numerical variables is given below :



Looking at the pairplot, we can conclude that 'registered' variable has highest correlation with target variable, followed by 'temp' and 'atemp'.

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)

a. Validation on test data set

Predict the y values based on accepted model and look for 'R-squared' value. There shall not be much difference between R-squared values on train-data and test data.

b. Residual plot

The plot of errors or residuals should be randomly scattered around '0' without any specific pattern.

c. Histogram of residuals

The histogram of residuals shall follow normal distribution.

d. Variance inflation factor
The VIF of predictors from final accepted model shall be equal to or less than 5.

e. Interpret R-squared and adjusted R-squared
If the above conditions are met and R-squared value of model (both on train and test data) is high, it increases confidence in model's validity.

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)

Based on final model, the coefficient of the model comes out to be :

season	0.043348
yr	0.170890
workingday	0.219272
windspeed	-0.065219
casual	0.815867

Based on these coefficients the, 3 top most features explaining the demand of shared bikes are:

1. casual
2. workingday
3. yr

These features have strongest impact on 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)

<Your answer for Question 6 goes here>

The linear regression is supervised learning algorithm used to model the relationship between dependent or target variable and independent variables. Its goal is to find a best fitting line that minimizes the mean square error or ordinary least square. It assumes that the relationship between independent and dependent variables is linear.

There are two categories of linear regression :
1. Simple Linear Regression
2. Multiple Linear Regression

Simple linear regression –

In simple linear regression we assume that there is linear relationship between y (target variable) and x (single independent variable).

Representation:

y=β0+β1x+c

β0: the intercept, representing the predicted value of y when x=0.

β1: the slope, representing the change in y for a one-unit change in x.

c: the error term, capturing the difference between actual and predicted values.

Multiple linear regression –

In multiple linear regression we assume that there is linear relationship between y (target variable) and x1, x2.. xn (multiple independent variable).

y=β0+β1x1+β2x2+⋯+βnxn+c

Assumptions:

The relationship between independent and dependent variables is linear.

The variance of residuals (errors) is constant across all levels of the independent variables.

The residuals should follow a normal distribution.

The residuals should be independent.

Independent variables should not be highly correlated with each other.

The model performance is evaluated based on:

R-squared, Adjusted R-squared and RSME (Root mean square error).

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, designed by statistician Francis Anscombe in 1973, to demonstrate the importance of data visualization and the limitations of using simple statistical measures alone for data analysis.

Each dataset in Anscombe's Quartet has nearly identical basic statistical properties, such as mean, variance, correlation, and regression line. However, when plotted, they reveal very different relationships and patterns.

Composition of Anscombe's Quartet

Each of the four datasets in the quartet consists of 11 pairs of x and y values.

All four datasets have similar statistical summaries, including :

- The mean of x and y values (approx. 9 and 7.5 resp)
- The variance of x and y values (approx. 11 and 4.12 resp)
- The correlation coefficient between x and y (approx. 0.816)
- The slope and intercept of the linear regression line fitted to each dataset
- The coefficient of determination (R2) for the linear regression (approx. 0.67)

Despite these similarities, the datasets are structurally different, demonstrating that similar statistics can arise from very different data patterns.

Anscombe's Quartet serves as a classic example in data analysis to demonstrate that:
Understanding data requires both statistical measures and graphical exploration to make accurate, insightful 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)

<Your answer for Question 8 goes here>
Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies both the strength and direction of this linear association, giving a value between -1 and 1:
+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear correlation.

The closer the value of Pearson's R is to either +1 or -1, the stronger the linear relationship.
Assumption - Both variables are normally distributed and have a 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)

<Your answer for Question 9 goes here>
Scaling is a data preprocessing technique used to adjust the range and distribution of feature values in a dataset.
The goal of scaling is to bring all features to a similar scale or range, which can improve the performance and training stability of many machine learning algorithms.
Scaling is performed as :
Features with larger values can dominate models if not scaled, leading to biased results.

Normalized scaling :
Formula –
$(X-Xmin) / (Xmax-Xmin)$
Often used when we know the data distribution does not follow a normal distribution

Standardized Scaling :
Formula –
$(X-\mu)/\sigma$
Transforms data to have a mean of 0 and a standard deviation of 1.

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>

Infinite VIF maybe due to:

Perfect Multicollinearity: This occurs when one predictor variable is a perfect linear combination of one or more other predictor variables.

Zero Variance: If a variable has zero variance (i.e., it is constant across all observations), it means that the variable does not provide any information and can lead to infinite VIF.

Singular Matrix: When fitting a regression model, if the predictor matrix is singular (i.e., it cannot be inverted), this can result in infinite VIF.

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>

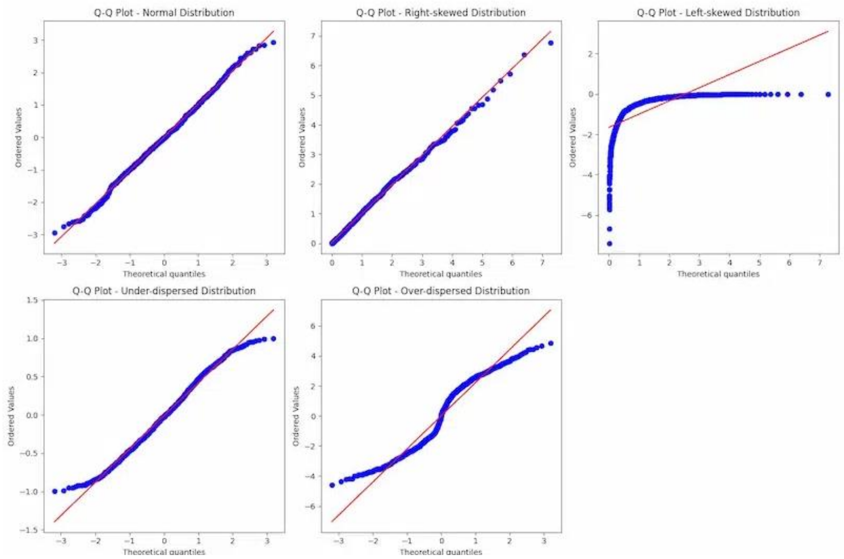
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution.

In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. The x-axis typically represents the quantiles of the theoretical distribution (e.g., normal), while the y-axis represents the quantiles of the sample data.

Q-Q are used in linear regression for :

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. A Q-Q plot helps to visually assess whether this assumption holds.

Example Q-Q plots :



Submission by - Harshada Kalsekar (ML65)