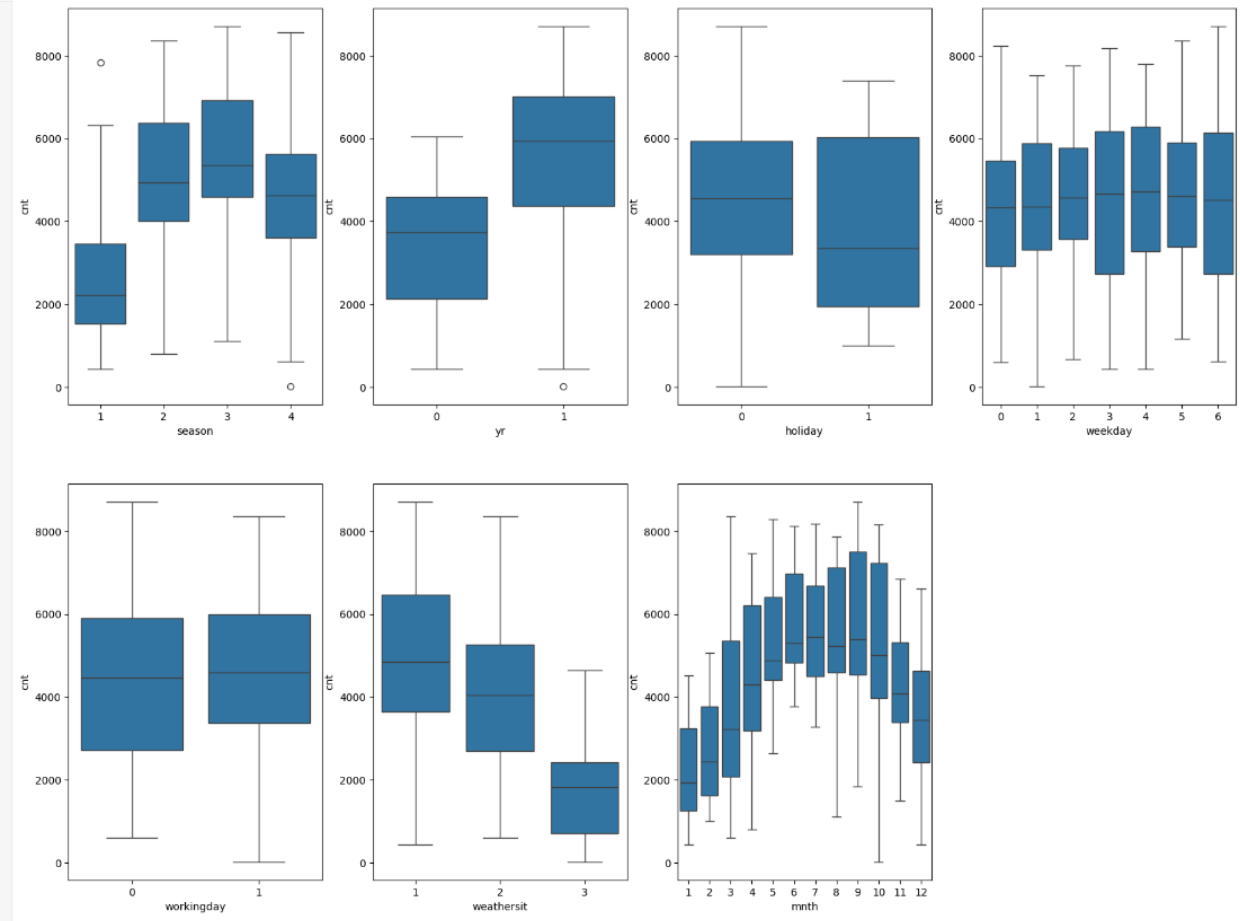
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Conslusions through above box plots:

1.Season: The category 3 (Fall), has the highest median, menaing that the demand was high during at this season. And it is lowest for 1: spring .

2.year: The year 2019 had a higher count of users as compared to the year 2018

3.weekday: The bike demand is almost constant throughout the week.

4.weathersit: The count of total users is in between 4000 to 6000 (~5500) during clear weather

5.mnth: August month has the highest user count

6.workday: During holidays the count is lowets

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Dropping the first column is important because not dropping will cause the dummy variables to be correlated (or redundant) which eventually may affect some models adversely and the effect is stronger when the cardinality is smaller.

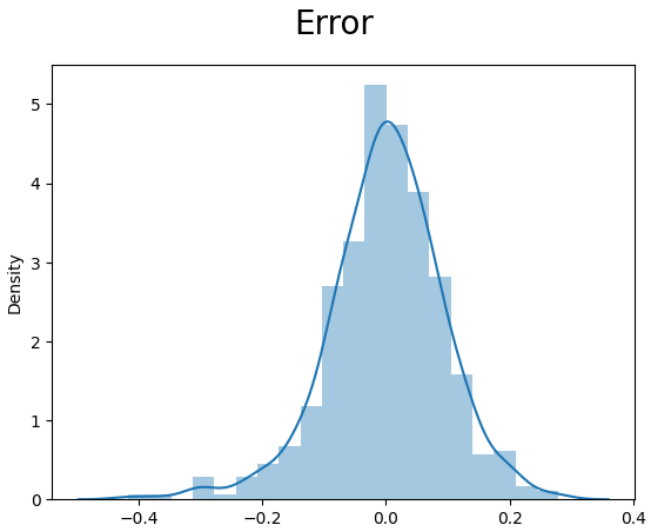
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Using the pairplot we can clearly observe that , “temp” and “atemp” are the 2 numerical variables which are highly correlated with the target variable (cnt).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Using the following tests:

1. Linear regression required the relationship between the independent and dependent variables to be linear. Which can be done through pairplot.
2. Residuals distribution should follow normal distribution and should be centered on 0 (mean = 0). This assumption about residuals is validated by plotting a distplot of residuals. This is to check that whether residuals are following normal distribution or not.
3. Linear regression assumes that there is no multicollinearity in the data. Which occurs when the independent variables are too highly correlated with each other. That can be calculated by calculating the VIF (Variance Inflation Factor).



Note: The error terms are centred around 0 and follows a normal distribution, this is in accordance with the stated assumptions of linear regression.

Verifying the above conclusion using a qq-plot

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features are:

1. temp with coefficient : 0.553883
2. yr with coefficient : 0.233054
3. weathersit_Light Snow & Rain with coefficient : -0.290127

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a statistical model that helps in understanding the relationship between a dependent variable with given independent variable through a straight line. This means that when there is a change in the value of one or more independent variables, whether increase or decrease, the value of the dependent variable also changes accordingly.

➢ Dependent variable, in any given task, is the variable we aim to predict. In linear regression, it is denoted by 'y'.

➢ Independent variables are the variables that affect the dependent variable and are denoted as X1, X2, ..., Xn.

➢ The algorithm estimates coefficients (b1, b2, ..., bn) for each independent variable and an intercept (b0).

Therefore, the linear regression equation formed by these terminologies is:

y = b0 + b1X1 + …… + bnXn

The primary objective in linear regression is to find the values of b0, b1, b2,……., bn that minimize the sum of squared differences between the predicted and actual values, and this is known as Ordinary Least Square(OLS) regression.

Following are the steps for the model:

1. The dataset is split into Training and Test sets.

2. It is then divided into X and Y sets for model building.

3. Recursive Feature Elimination is done that removes features, fits a model, and then evaluates the performance until the desired number of features is reached.

4. Next, the Variance Inflation Factor is measured which quantifies how much a variable is contributing to the variance of a model. It is used to detect multicollinearity among predictor variables in a regression analysis.

5. Then the Linear Model is built by using statsmodel.

6. After the model is finalized, Residual Analysis is done of the Train data.

7. At last, prediction is done using the final model, followed by Model Evaluation.

There are two types of Linear Regression:

a) Simple linear regression – it explains the relationship between a dependent variable and one independent variable using a straight line. Equation – $y = b_o + b_1.X$

b) Multiple linear regression – it is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
   Equation - y = b0 + b1X1 + …… + bnXn

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of plotting data before you analyze it and build your model.

- These four datasets have nearly the same statistical observations, which provide the same information for each x and y point in all four datasets.

- However, when you plot these datasets, they look very different from one another, that is, a set of four datasets that have nearly identical simple descriptive statistics but vary widely when graphed.
- It visualizes data before analyzing it and to highlight the limitations of relying solely on summary statistics.
- While sharing identical mean, variance, correlation, and linear regression line characteristics, the datasets within the quartet tell the significance of visually examining data and avoiding dependence solely on summary statistics.
- This quartet helps in understanding the importance of Visualization as by looking at the data structure graphically we can get a clear understanding of it.
- Therefore, we can say, it highlights that depending exclusively on numerical summaries can be misleading sometimes and we might fail to grasp the complexity within the data.

3. What is Pearson's R? (3 marks)

Pearson's R or Pearson's Correlation Coefficient is a statistical tool that measures the strength of linear relationship between the variables. It quantifies how well the relationship between variables can be described with the help of a straight line.

The Pearson's (r) or the value of (r) can range from -1 to 1:

- r = -1, indicates a perfect negative linear relationship
- r = 0, indicates no linear relationship
- r = 1, indicates perfect positive linear relation.
- The sign of (r) indicates the direction of the relationship.
- +ve sign means that when one variable increases, the other also increases, whereas
- -ve sign indicates that with the increase of one variable, the other will decrease.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

In machine learning and statistics, scaling is the process of converting numerical features of a data set into a standard range (or distribution). The main purpose of scaling is to make sure that all features contribute the same amount to the analysis and the model training process. Scaling is the process of reducing the values of variables in a data set to a specific scale so that they are comparable and that variables with larger values do not dominate the analysis.

Scaling is performed for various reasons:

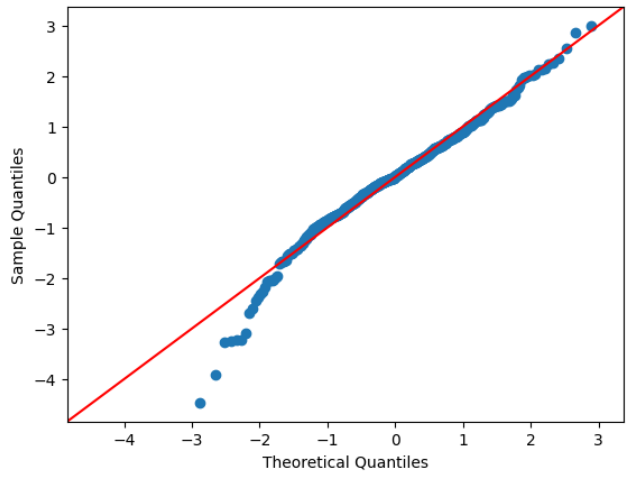
- Equal contribution: scaling all features ensures that they contribute the same amount to the model, otherwise, variables with larger scales will dominate and affect the model disproportionately.
- Gradient descent convergence: many machine learning algorithms (especially gradient descent optimization algorithms, such as linear regression and logistic regression, and neural networks) converge faster and work better when input features are similar in scale.
- Distance-based algorithms: algorithms that rely on distance between data points, such as K-nearest Neighbor and K-Means clustering, are sensitive to the size of features. Scaling prevents features with larger sizes from having a larger effect on distance calculations.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
The Variance Inflation Factor (VIF) serves as a tool for detecting multicollinearity in a regression model. Elevated VIF values suggest pronounced correlations among the independent variables, complicating the ability to discern the unique influence of each variable on the dependent variable. Although VIF is a valuable diagnostic tool, it can encounter issues and potentially reach infinity under specific circumstances.
Infinite VIF occurs when one or more independent variables are perfectly correlated, leading to a situation known as perfect multicollinearity. Perfect multicollinearity means that one or more variables can be exactly predicted using a linear combination of the other variables in the model. To solve this we need to drop one of the variables from the dataset which is causing this.
If a variable in the model is a linear combination of other variables, it can result in a perfect multicollinearity scenario and, consequently, an infinite VIF.
When there is a linear dependence among the independent variables, the matrix inversion required in the VIF calculation becomes problematic, resulting in an infinite VIF.
Infinite VIF values can lead to unreliable coefficient estimates in regression models.
The standard errors of the coefficients become extremely large, making it difficult to draw valid statistical inferences.

(3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess the normality of a dataset by comparing the quantiles of the observed data with the quantiles of a theoretical normal distribution. In a Q-Q plot, if the points closely follow a straight line, it indicates that the data is approximately normally distributed.

```
[239]: sm.qqplot((y_train - y_train_cnt), fit=True, line='45')
       plt.show()
```



So it can be firmly said that most of the data points lie on the straight line, which indicates that the error terms are distributed normally.

(3 marks)