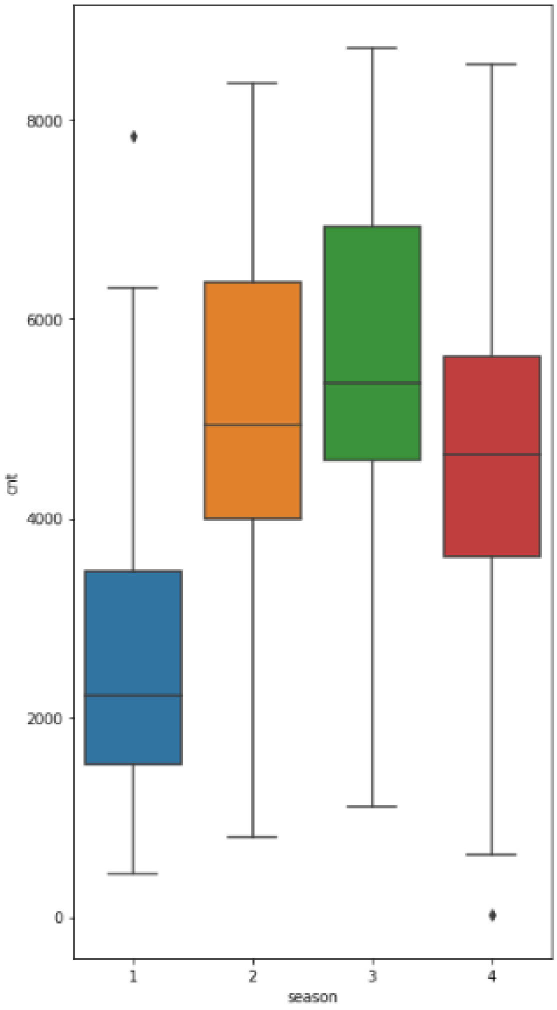
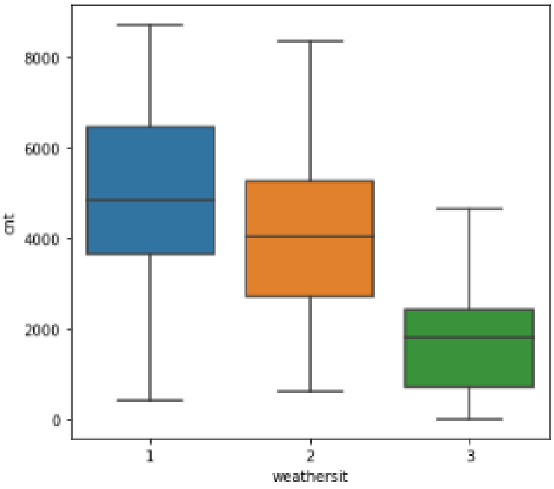
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   - The values of categorical variables like weathersit and season have high effect on our dependent variable cnt.





As you can see the cnt is very much different for every categorical value change.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - If there are supposedly 4 values in a categorical variable for eg. A,B,C,D we can assign them values like A = 1000, B = 0100, C = 0010, D = 0001.

   From the above example if we drop one of the variable like A then we can assume that B = 100, C = 010, D = 001 and with this logic A will be 000. This technique reduces the number of columns which makes it very easy for the analysis. That's why drop_first=True is important while creating dummy variable which drops the first column ('A' in our case).

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   -After building the model, I did the residual analysis on the training data which is very important. In this analysis we check if the error terms are normally distributed.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   -According to me the top 3 features are atemp, windspeed and light_snow are the features which explain the demand of the shared bikes significantly.

General Subjective Questions

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

- Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

Step 1: Suppose there is a dataset provided and we have to predict the dependent variable 'X' and some independent variables 'A','B','C', the first step is to convert these variables into numeric values and to check the corelation between the independent and the dependent variable

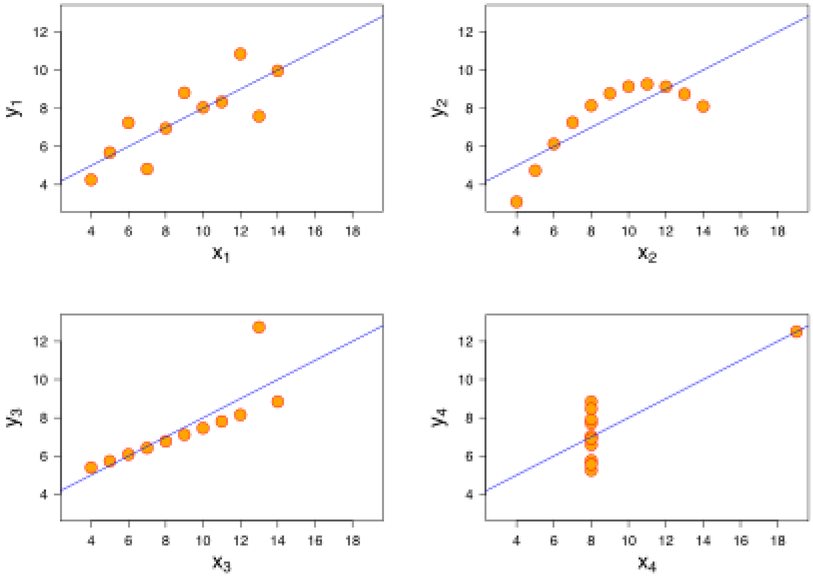
Step 2: After studying the corelation we will have to rescale the values of all the variables in the data using Min- Max Scaling or Standardisation

Step3: The next step would be creating the Linear regression model, in which we try to get the best fit straight line which passes through all or most of the points and its R squared values be as low as possible.

Step 4: This is the main step where we do the residual check in which we check if the error terms have a Normal distribution.

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

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

- In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/) — also known as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation,[1] or colloquially simply as the correlation — is a measure of linear correlation between two sets of data. It is the ratio between the [circular reference] of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

- This means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1.

  In regression, it is often recommended to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.

  Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

- If there is perfect correlation of the target variable with some other variable(s), then VIF = infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

  For regression, when checking if the data in this sample is normally distributed, we can use a Normal Q-Q plot to test that assumption.