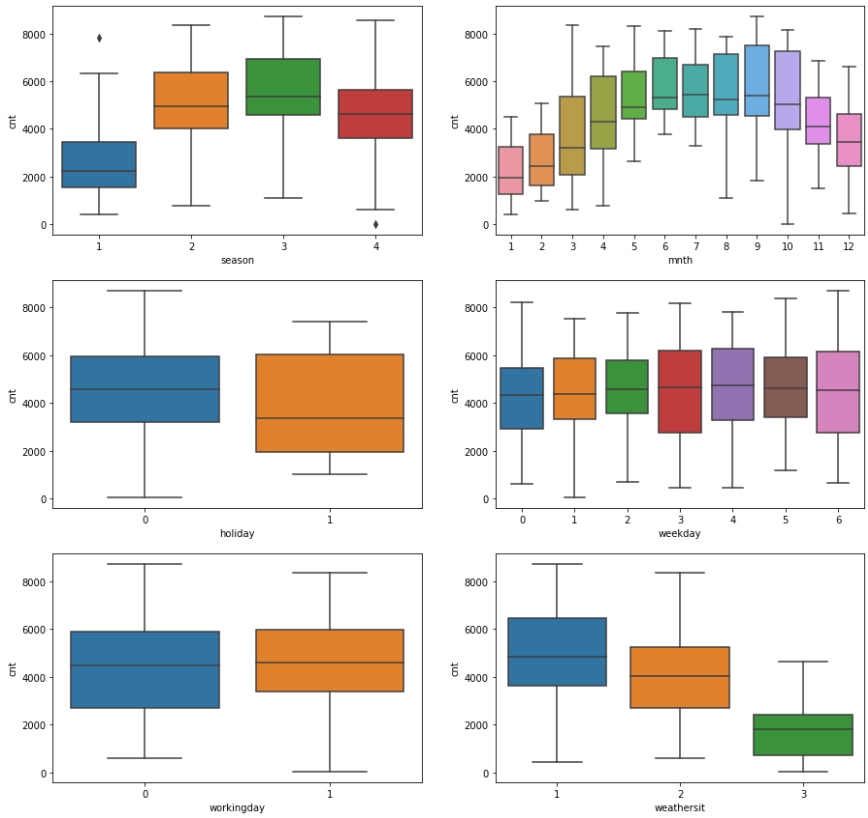
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans. I just plotted the box plot for categorical variables as follows -



General observations -

- Seasons 2 and 3 i.e. summer and fall have a higher count median as compared to seasons 1 and 4 i.e. spring and winter
- It's US demographic data, so in the US we have a lot of holidays in spring and winter that's why the share count dropped by some amount and same trends followed by month
- Holyday has some effect on sharing count, but weekdays and working days have no significate impact on sharing count.
- The weather condition has a huge impact on sharing count, as weather conditions drop sharing also drops.

Q2. Why is it important to use drop_first=True during dummy variable creation?
Ans. Drop first will drop on an extra categorical variable that is generated, it's always good to have less complex models which have optimal scores.
Let me explain with an example - We have categories for the weather as poor, normal, good

Category Values	poor	normal	good
poor	1	0	0
normal	0	1	0
good	0	0	1

Now, observed that 100 is poor, 010 is normal and 001 is good. This means 1 represent what kind of weather it is so for good 001 can be represented as it's not poor and not normal i.e. 00 so we can drop anyone which reduce model complexity and information remains the same.

Category Values	poor	normal
poor	1	0
normal	0	1
good	0	0

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans. The **temp** and **atemp** have the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans.

- I look in R-squared, p-value and VIF for model validation.
- R-square explains how well the model fit over the data and adjusted R-squares.
- P-value and VIF help to understand how well any feature contributes to the model target output.

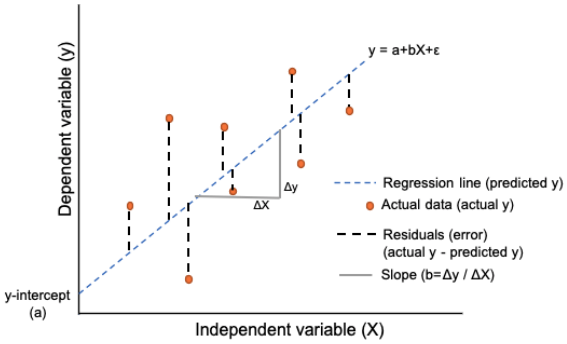
Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

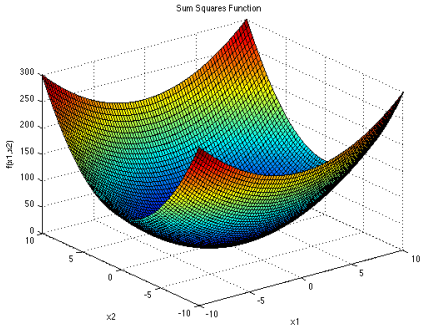
Ans. Season, weathersit and wind speed are the top 3 features contributing to share demand. Coffe. are giver bellow

Season - **(+0.25)** Weathers - **(-0.3)** Wind speed - **(-0.16)**

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.
Ans. Linear regression algorithm is used to find the best-fit line in iterations, in each iteration current line move toward the best fit line. This algorithm has three steps and these three steps are iteratively called to get the best line. There is two model hyperparameter with control, learning rate **(lr)** how fast the model moves toward the zero-error and number of titration **(epochs)** for repetitive steps.





Steps:

1. Initialise the trainable params i.e. **b0, b1, …** with a small random value, learning rate

$$b0, b1, b2,.., bn = 0.12, 0.34, 0.2,.., 0.3$$

2. Make predictions on train data using params from step 1, let's say **y_predicted**

$$y_{predicted} = b0 + b1 * x1 + b2 * x2 + ... + bn * xn$$

3. Compute the error, Sum Of Squared Error

$$E = \boldsymbol{\Sigma} (y_{actual} - y_{predicted})^2$$

4. Compute the gradient wrt to each variable (partial derivative)

$$\delta E/\delta b$$

5. Compute new trainable params, by moving a small step toward the gradient

$$b_{new} = b_{old} - lr * \delta E/\delta b$$

6. Repeat steps 2 to 5 as per the number of iterations defied **(epochs)**

Q2. Explain the Anscombe's quartet in detail.
Ans. As we start working on data and just look into the numbers for making decisions. So Anscombe's quartet explains how important visualisation is.

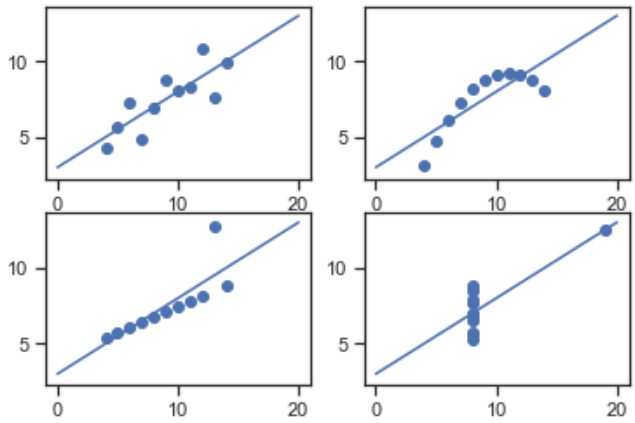
In this model four datasets were used for demonstrating the importance of visualisation in form of (x1,y1), (x2, y2), (x3, y3) and (x4, y4). These datasets share the same descriptive values like (mean, variance, SD, LR line etc..). But when plotted the graph they totally different.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

These are some descriptive data -

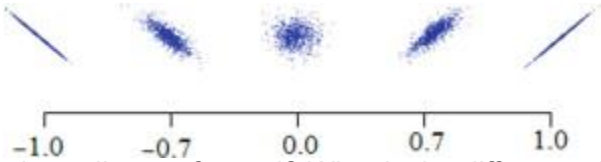
- **Average Value of x = 9**
- **Average Value of y = 7.50**
- **Variance of x = 11**
- **Variance of y =4.12**
- **Correlation Coefficient = 0.816**
- **Linear Regression Equation: y = 0.5 x + 3**

But when looking to plot they are totally different, hence plotting of data is very important before making any decision.



Q3. What is Pearson's R?
Ans. It is simple computation with the output number b/w [-1, 1], which define how two numerical data are correlated.

- If the coefficient is 0 then A and B are not correlated i.e. if A increases or decrease there is no effect on B and vice-versa.
- If the coefficient is 1 then A and B are highly +ve correlated i.e. if A increases or decrease there is with the same rate B increases or decreases and vice-versa.
- If the coefficient is -1 then A and B are highly -ve correlated i.e. if A increases or decrease there is with the same rate B decreases or increases and vice-versa.



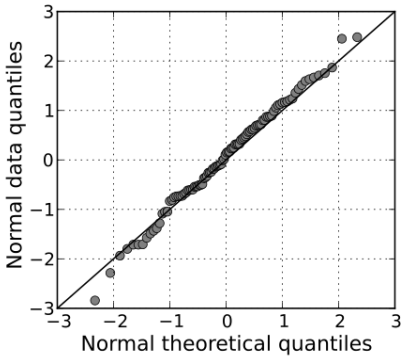
Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans. Scaling is a method to convert data variables into normalized ranges. Suppose the dataset have weights data of players and some of the data are in kgs and some are in grams so when we fit the model then it gets confused, so to handle this we have to rescale it at the same pace that's why data scaling is needed.
Difference b/w normalized and standardized scaling -

- Normalized is scaling transform data [0,1] range wherein standardized scaling data scaled where mean (μ) of 0 and standard deviation (σ) of 1 (unit variance)

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans. As we look into the formula, we found that VIF depends on the R2 value -

$$VIF = 1 / (1 - R2)$$

And R2 explain how two variables are correlated, so if R2 is 1 i.e. perfectly correlated then VIF will be 1/0 which is infinite.
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans. Q-Q plot is is a probability plot, which is plotted against two probability distributions.



Q-Q plot is used to find the type of distribution for example whether it be Uniform Distribution, Exponential Distribution or Gaussian Distribution, etc…