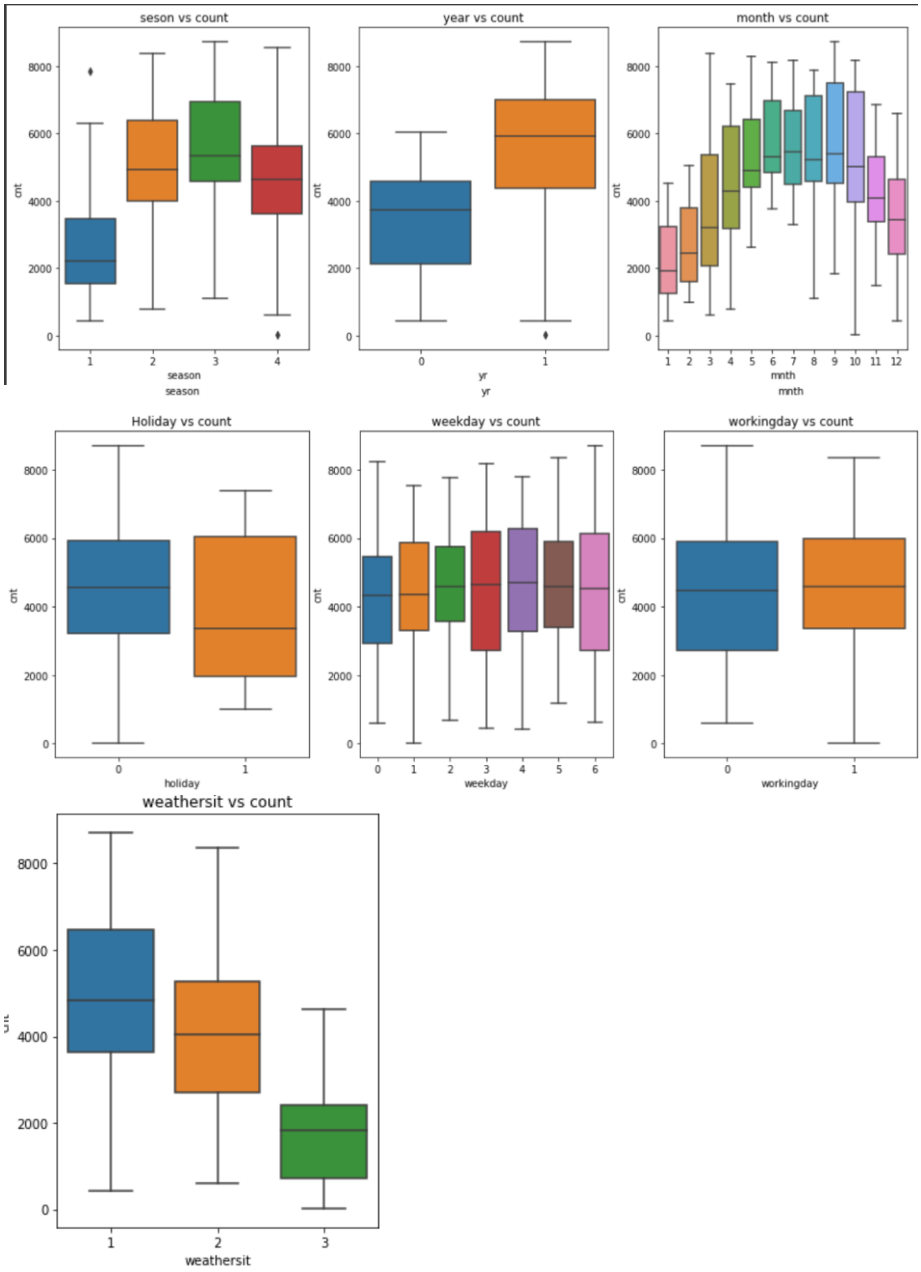
Assignment-based Subjective Questions

1. **From the categorical variables analysis from the dataset, what can be inferred about their effect on the dependent variable?**



Inferences:

- There is a low demand in Spring Season compared to summer, fall and winter. Generally, bike sharing demand remain high in summer and fall season.
- Popularity of Bike Sharing demand is increasing. Trend is visible. There is a demand increase in 2019 compared to past year 2018.
- From starting month to till 10th, In general demand increases then it starts to decline till Jan.
- Median of holiday and first Q1 is less than compared to non-holiday median and Q1.

- There is very low demand in snow season and no demand in heavy rainy season.

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

   Let's take example categorical value which having three values. Panda.get_dummies function generates three dummy variables if we give the same categorical variable.

value1	value2	value3	
1	0	0	if categorical value having value1
0	1	0	if categorical value having value2
0	0	1	if categorical value having value2

   If use parameter drop_first=True

value2	value3	
0	0	if categorical value having value1
1	0	if categorical value having value2
0	1	if categorical value having value2

   Still, we can represent three values using 2 variables and values1 seems redundant variable.

   In general rule, if there are n unique values in categorical column, all values can be represented by n-1 dummy variables.

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

   Temperature is highly corelated to target variable. Temperature is also corelated with atemp variable. These are almost identical.

4. **How can assumptions be validated of Linear Regression after building the model on the training set?**

   - **Non-Multicollinearity:** We used stats library to calculate VIF for each feature in train dataset. We found VIF of all chosen features are less than 3.
   - **Error Term followed perfect well curve cantered around zero mean**.
   - All chosen features have low P value. It means its correlation with target variable is significant.
   - After plotting Residual with Ytrue shows no heteroscedasticity. It means built model is good enough to predict.
   - We also plot scatter plot between ytrue and ypred for both test and train dataset. Plot forming linear curve. It means model behaving perfectly fine for both test and train dataset.

- Last, we also calculated R2-score for both train as well as test. Here, I have seen significant difference b/w test and train score. It leads overfit model.

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

   Top three features
   year – demand increases as year increases.

   light_snow - negative relation,

   winter – positive relation.

General Subjective Questions

1. **Explain the linear regression algorithm in detail?**

   Linear Regression is technique in which we try to explain linear relationship of features with target variable

   We can express relation with following expression

   yhat = B0 + B1X1 + B2X2 + B3X3+……

   Here yhat is predicted value and B0,B1,B2 and B3 is coefficient of respective features.

   If ytrue is true value of target

   Then error can be expressed as

   Error = |ytrue = yhat |

   Making square of this term and adding with all datapoints

   Squared error = Sum of (Yitrue – Yihat) (Yitrue – Yihat)

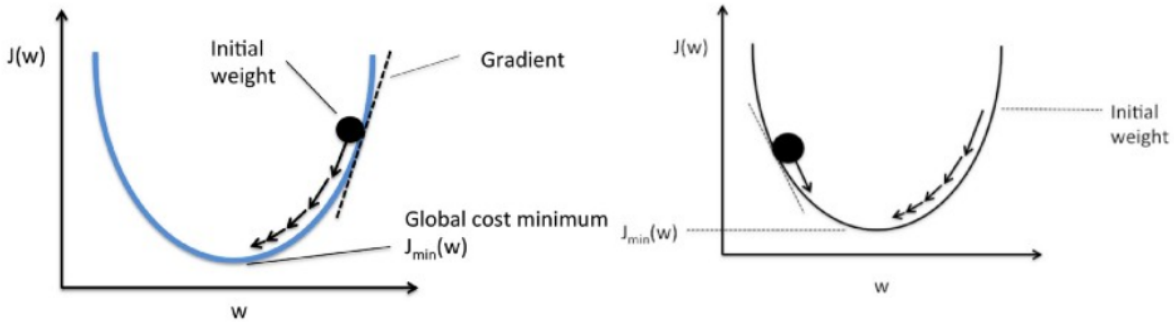
   We try to get best fit line that minimize this error.

   To get optimal best fit line, we follow Gradient decent Algorithm. In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.

   Gradient Descent:

   Gradient descent works like a small step which is called learning parameter towards down a graph. Learning parameter is hyper parameter if we choose very very small learning

parameter it will take time to reach at minimum position if we choose very high number of steps as learning parameter, it will jump from one side to another side. So we must choose optimal value of learning parameter.



Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression. It is given by

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Here, h is the linear hypothesis model, h=Θ0 + Θ1x, y is the true output, and m is the number of data points in the training set.

Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value.

The update is:

Repeat until convergence

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\, x_j^{(i)} \quad \text{for j = 1,2,...,n}$$

2. Explain the Anscombe's quartet in detail.

It is defined in Wikipedia as Anscombe's quartet comprises four datasets that have exact identical statistical properties, but they show different pattern or visual effect in graph.

These datasets were constructed in 1973 by statistician Francis Anscombe. He demonstrated importance of visualization before go with modelling or analysing algorithms.

Each dataset consists 11 datapoints (x,y). Mean, variance, standard deviation are identical for all four dataset, appear different in graph.

3. **What is Pearson's R?**

It measures correlation b/w two variables. It can be computed by using following formula:

R(xy) = sigma(xy) / sigma(x) sigma(y)

Where sigma(xy) is the covariance of population
sigma(x) => standard deviation x variable
sigma(y) => standard deviation of y variable

Its value lies between -1 to 1.

1 ➔ means x and y positively linear dependent
0 ➔ means no relationship b/w x and y
-1 ➔ means x and y negatively linear dependent

This correlation measurement can also be used to verify significance dependent one variable to another in Hypothesis testing.

The Pearson correlation of the sample is r. It is an estimate of rho (ρ), the Pearson correlation of the population. Knowing r and n (the sample size), we can infer whether ρ is significantly different from 0.

Null hypothesis (H0): ρ = 0
Alternative hypothesis (Ha): ρ ≠ 0

We can calculate the t value (a test statistic) using this formula:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Then finding critical value based on t value and infer significant label.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling definition:

In real life when we build supervise model or unsupervised model using dataset, it consists multiple features for example age, salary, weight. These features are in different units and measurements. So Scaling is a such technique which normalize the different features. In other words, it brings all features under the same scale.

Why scaling is important:

- Resulted model will generate features coefficients according to respective scale of feature. So, business interpretation will be almost impossible if all coefficients are in different scale. We can interpret easily that most top feature and less important feature.
- Gradient converges very quickly towards minima if all features are in same scale.

Difference b/w Normalized scaling and Standardized Scaling

Normalization:

Normalized scaling brings all features into 0 to 1 range. It is MinMaxScaler in skit learn library. It follows below formula

$$X\_norm = (X - X\_min)/(X\_max - X\_min)$$

It affects outlier because it is choosing max and min from dataset.

Standardization:

The basic concept behind the standardization function is to make data points centred about the mean of all the data points presented in a feature with a unit standard deviation. This means the mean of the data point will be zero and the standard deviation will be 1.

$$X\_Std = (X - X\_mean) / std$$

It does not affect outlier.

5. Sometimes the value of VIF is infinite. Why does this happen?

VIF measures multicollinearity of one variable among other variables in dataset. It follows below formula

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

Since R-Squared value lies between 0 to 1.

When R-Squared value is 0 ➔ it means one variable is perfectly orthogonal to all other variables in dataset. It means no relation also leads VIF value as 1.

When R-Squared value is 1 ➔ It means one variable able to explain all other variable perfectly. It means perfect co-linear also leads VIF value as infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

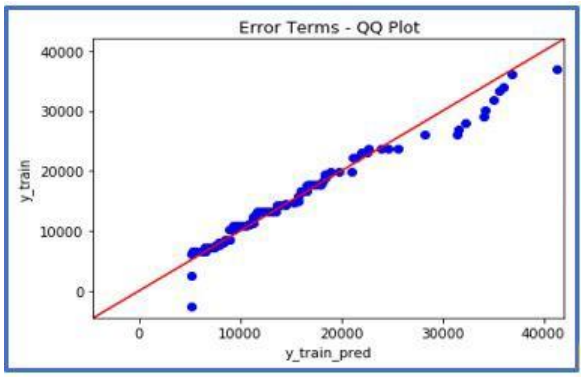
Definition :

A Q-Q plot is a plot of quantiles of first data set against the quantiles of second data set.

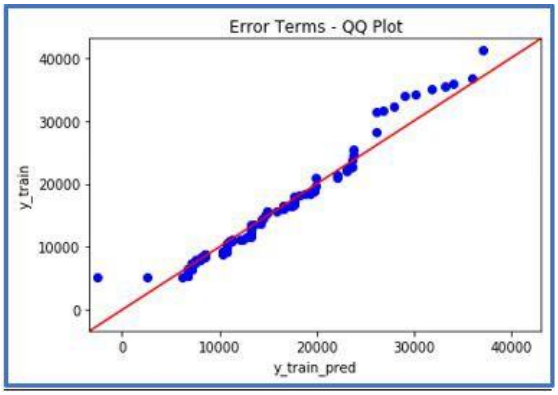
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Importance:

It is useful to determine following:

i. Two data come from populations with a common distributional shape, have similar tail behaviour.
ii. It ensures ml model is based on right distribution,
iii. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
iv. Skewness of distribution