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?**
   Ans. The categorical variables in the dataset were mapped to the integers, so first I mapped the integers to the labels like for months 1-> Jan, 2->Feb, etc. Then I plotted them against the target variable to get insights. Following are the insights from my analysis:
   1. Number of rental bikes increase when the season is summer or fall. Fall has the highest median.
   2. The number of rental are increased in the year 2019.
   3. Months follow the same pattern as seasons. The highest number of rentals are made in September followed by October.
   4. Weekdays do not say much about the number of rentals.
   5. People tend to rent the bikes on the working days than the holidays.
   6. If the weather is clear, people will rent the bikes than of other weather situations.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Ans. While creating dummy variables, we represent the dummy variables in n levels. For example, if we have four seasons, then dummy variables will be created for all the four seasons.

	Spring	Summer	Fall	Winter
Spring	1	0	0	0
Summer	0	1	0	0
Fall	0	0	1	0
Winter	0	0	0	1

Table 1.0

But we can use the n-1 levels of data to represent the n levels. Like, in seasons data, if spring has 0 value, summer has 0 value, fall has 0 value, they we know the season must be winter.

From Table 1.0, we can drop the winter column and we will still be able to represent all the four levels.

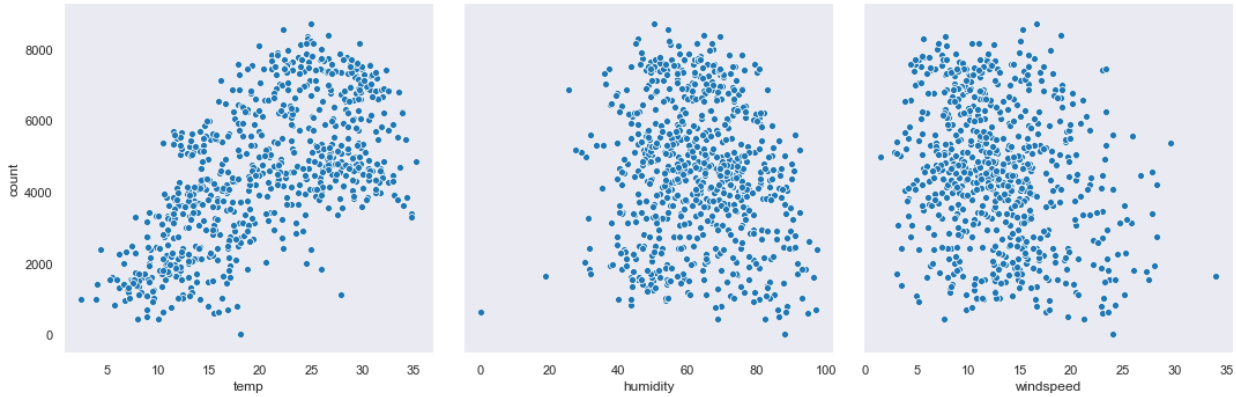
	Spring	Summer	Fall
Spring	1	0	0
Summer	0	1	0
Fall	0	0	1
Winter	0	0	0

Table 1.1

In Table 1.1, we just have 3 levels but we can interpret the data for all the four levels.

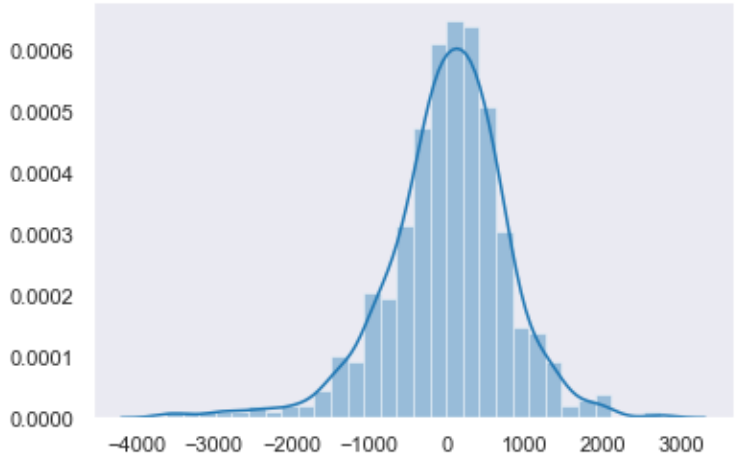
So, here we don't need to create a separate variable for winter season to represent it as winter, we can use the spring, summer and fall season to represent the winter.

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

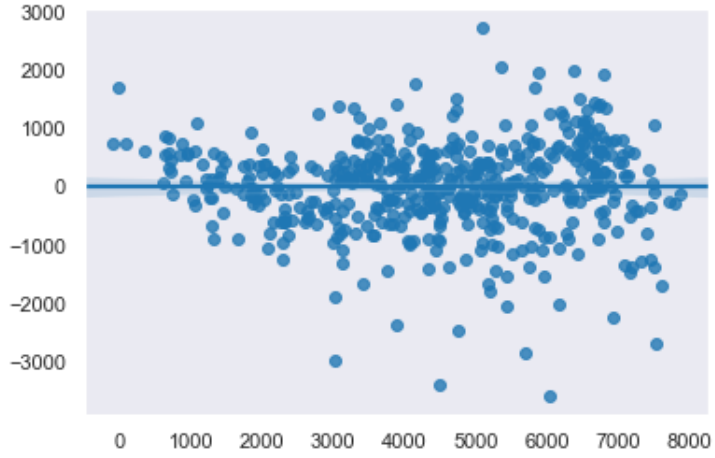


'temp' and 'atemp' has high correlation, so we dropped the 'atemp' variable. 'temp' has the highest correlation(0.627) with the target variable i.e. count.

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



By plotting the distance plot(histogram) of the error terms, we can see that error terms follow the normal distribution with mean 0.



By plotting the regplot() from seaborn library we can see there is no clear trend in the error terms.

And to check whether there is multicollinearity or not between the independent variables we used the VIF, i.e. Variance Inflation Factor.

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

   Ans. From the final model, three features significantly contributing to the demand of rental bikes are:

   a. temp(4963.951)
   b. Light Rain/Snow(-2075.683) – weather situation when there is little rain or snow
   c. 2019(1994.410) – when year is 2019
   d. windspeed(-1694.925)
   e. humidity(-1455.884)

General Subjective Questions

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

Ans. Linear regression is the linear model that represents the linear relationship between the input variable(s) x and the single output variable. In simple words, we can say that, the value of output variable y can be calculated using the linear combination of input variable(s). When there is a single input variable, then the method is referred as simple linear regression(SLR). When there are multiple input variables, we refer it as multiple linear regression(MLR).

The representation of linear regression combines the specific set of input variables (x), such that the predicted value is same as the output variable (y). Here, both the input and output variables are numeric.

The linear equation assigns a scaling factor to each input variable, called as coefficient represented by Beta. One additional coefficient is also added, giving line a degree of freedom, which is known as intercept.

For a simple linear regression, form of the model would be,

$$\mathbf{y = \beta 0 + \beta 1 x}$$

y = dependent variable

x = independent variable

β0 = intercept

β1 = coefficient of x

For a multiple linear regression, form of model becomes,

$$\mathbf{y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + \beta 3 x 3 \ldots. + \beta n x n}$$

To get a best-fit line for our model, we need to optimize the values of coefficients by minimizing the error of model on your training data set. And to minimize the errors, we use the Gradient descent method.

There are some assumptions we need to check while performing linear regression. They are as follows:

- There is a linear relationship between x and y.
- Error terms have constant variance.
- Error terms are normally distributed and are independent of each other.

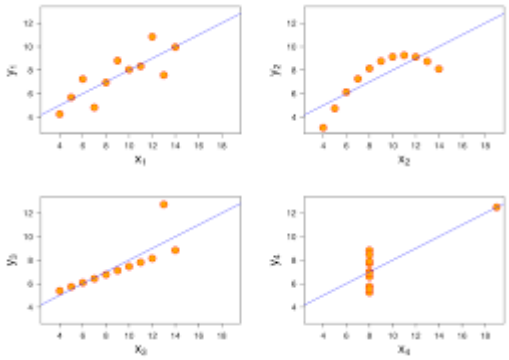
For multiple regression, with these assumptions, we need to check some other assumptions:

- The independent variables are not highly correlated with each other.
- The model is not overfitting on train dataset.

2. **Explain the Anscombe's quartet in detail.**

Ans. Anscombe's quartet compares the four different dataset with identical descriptive statistics and similar number of data points(i.e. 11 data points), but when plotted graphically the datasets follows different distributions.

These datasets were created by statistician Francis Anscombe in 1973 to demonstrate the importance of both the graphical plotting(visualization) and statistical summary while making a decision and how an outlier can influence the whole statistics of data.



1. First plot appeared as simple linear relationship.
2. Second plot does not follow a linear relationship and that's why the correlation coefficient for the same is irrelevant.
3. Third plot is linear relationship, but has a little different regression line due to outlier's presence in data.
4. Fourth plot does not follow a linear relationship, but a outlier is present, that made statistic look different and treat it as linear model.

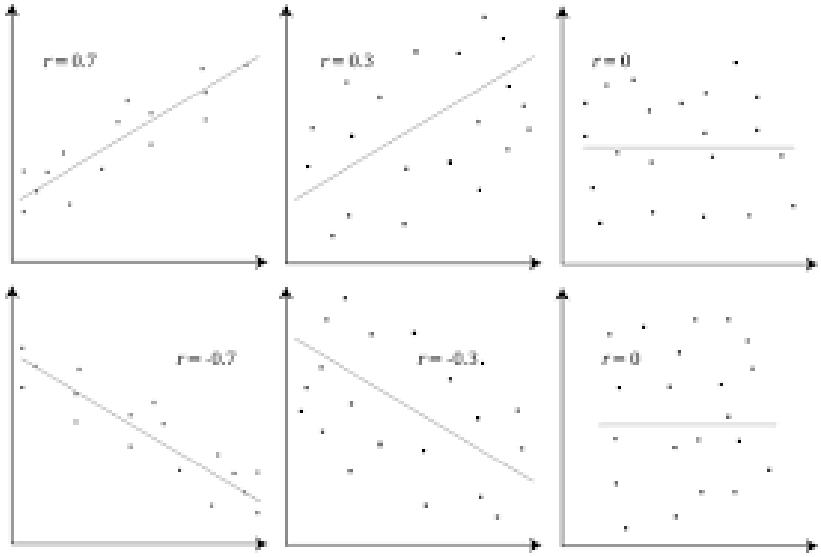
Overall, the Anscombe's quartet shows using just one of statistical summary or visualization is not enough to get the clear picture of data. Often, we need to use both the ways to see entire picture.

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

Ans. Pearson's R is defined as the measurement of strength of the relationship between two variables and their association with each other. In simple words, we can say that, it measures the effect of change in one variable when the other variable changes.

Pearson's R is calculated by taking the ratio between the covariance of two variables and product of their standard deviations.

Pearson's R is used to check the relationship between two variables. It's value ranges between -1 to +1.



• **Positive linear relationship** : A value nearing +1 is considered as positive linear relationship. It signifies that value of one variable increases when the value of other increases.

• **Negative linear relationship** : A value nearing -1 is considered as negative linear relationship. It signifies that value of one variable decreases when the value of other decreases.

• A value nearing 0 means there is no significant relationship between two variables.

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

Ans. Scaling is the technique applied to the independent variables to scale the data within a particular range. Also, this technique speeds up the calculations of the algorithm and saves the time.

The data is collected, stored at different times, different places having high or low magnitudes, different scales and different units. One variable can have range (0 – 50) and one variable can have range (1000-5000), we can see the difference between the magnitudes. If we create a model without changing the scale, it will influence the coefficients and low magnitude variables will get neglected as algorithm would not consider the units but the magnitude. To solve this issue, we can scale the variables and get all the variables on same scale i.e. same level of magnitude.

Difference between Normalized scaling and Standardized scaling :

	Normalized Scaling	Standardized Scaling
1.	Scales the data between 0 and 1.	Scales the data as their Z-scores
2.	It is used for non-gaussian distributions	It is used for gaussian distributions.
3.	Performing normalization can lose information.	Performing standardization does not lose information.
4.	Range is always between 0 and 1.	Range is not bounded between 0 and 1.
5.	It uses minimum and maximum value for scaling.	It uses mean and standard deviation for scaling.
6.	Outliers can influence the scaling.	Outliers cannot influence the scaling.
7.	Scikit-Learn provides MinMaxScaler() for performing transformation.	Scikit-Learn provides StandardScaler() for performing transformation.

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

Ans. VIF is used to check if multicollinearity is present between the independent variables or not. High VIF signifies the variance of variable can easily represented by all the other independent variables. Low VIF signifies the variance of the variable cannot be represented by other independent variables.

The formula to check the VIF is,

$$VIF_i = 1 / (1 - R_i^2)$$

The infinite VIF is only possible when R squared value is 1. And when R squared value is 1, we can definitely say that, there is a perfect correlation between the variables. In simple words, we can say that the respective variable can be easily represented by the linear combination of other independent variables.

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

Ans. Q-Q plot i.e. Quantile-Quantile plot is scatter plot used to assess if the two sets of data are from the same population having similar distribution. Also, we can see if the two sets of data follow any particular distribution like Normal distribution or uniform distribution.

Mainly, the Q-Q plot is used when we receive the data for training and testing separately. We can check their distributions using Q-Q plot.

We can detect the presence of outliers, the scale of variables, its change in symmetry from this plot. We can also use the plot to get a graphical view for scale and skewness are similar or different for two sets of data.

We can possibly infer the following with Q-Q plot:

1. **Similar distributions** : If all points of quantile lie around the straight line drawn at an angle of 45 degrees, we can say that, both data sets follow similar distributions.
2. **Different distributions** : If all points of quantile lie far away from the straight line, we can say that, both data sets follow different distributions.
3. **Y values < X values** : If y-quantiles lower than x-quantiles
4. **X values < Y values** : If x-quantiles lower than y-quantiles