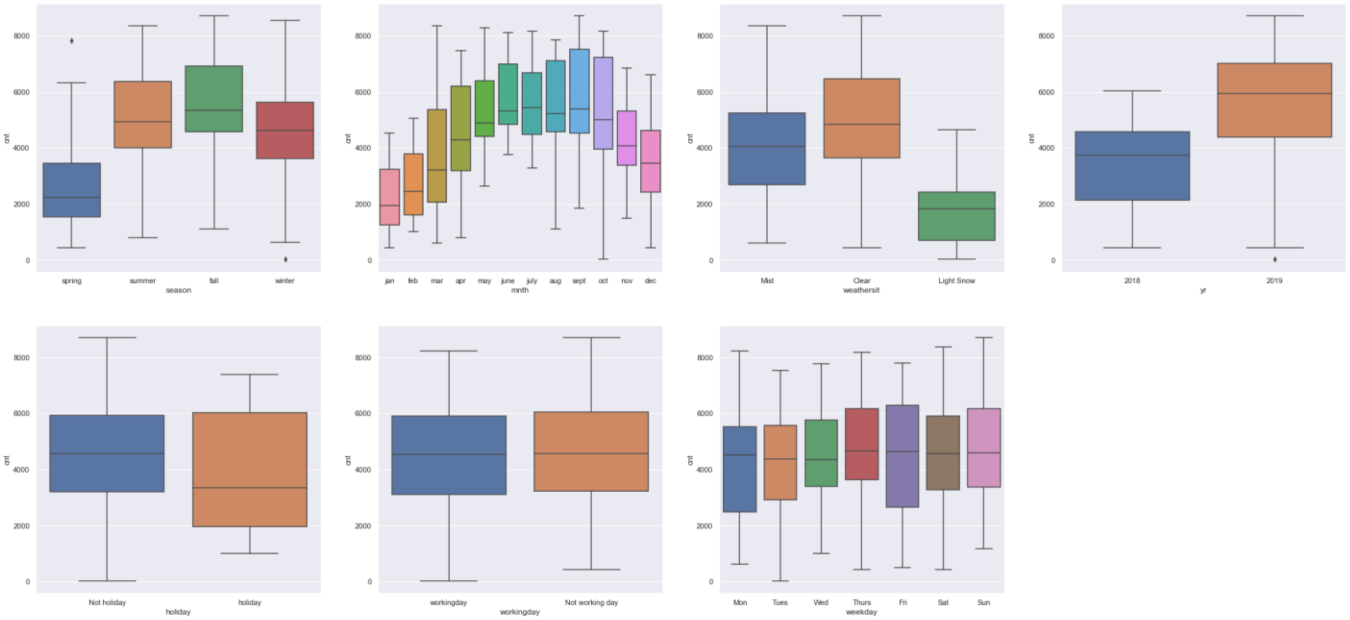
Name: Kumar Gaurav Sinha

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?**

A:



1. In the first boxplot that displays the relationship between seasons and cnt, we can find out that in the season fall, the demand is high. So, the pattern we observe is the demand starts rising from spring to summer and then to fall. Fall is the maturity stage, after which the demand declines i.e., means that the count of rental bike increases till fall and after that it decreases.
2. In the second boxplot, which tells us about the relationship between months and cnt, we can observe that the highest number of demands is in the month of June

3. In the third boxplot, which displays the relationship/pattern between weathersit and cnt, we can observe that the highest number of demands is when the weather is clear.
4. In the fourth boxplot, the relationship is between years and our target variables. We can observe that the count of rental bikes increases (demand of rental bikes) from next year onwards.
5. In the 5th boxplot, the relationship is between holiday and cnt, we can observe that when there's holiday the demand decreases
6. In the 6th boxplot, we can observe that the demand is almost same. It doesn’t matter if it’s a working day or not.
7. In the 7th boxplot, we can see that it doesn't matter which day it is. It's almost same. It doesn’t give any clear indication if it increases or decreases. But from the table, we can infer that the top day is Sunday.

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

A: Let’s say we have a column of gender, and we have 3 categories Male, Female and Transgender. So, we need to create a dummy variable. So, we will use get dummies function to get dummy variable. We will get something like this:

Male	Female	Transgender
0	0	1

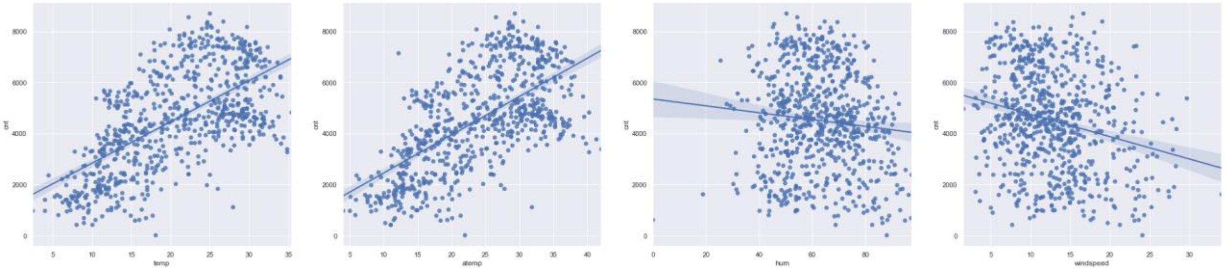
So, if we use drop_first = True, it will drop Male column. Result will be:

Female	Trangender
0	1

It reduces extra column that dummy variable created. It reduces correlation among each other.

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

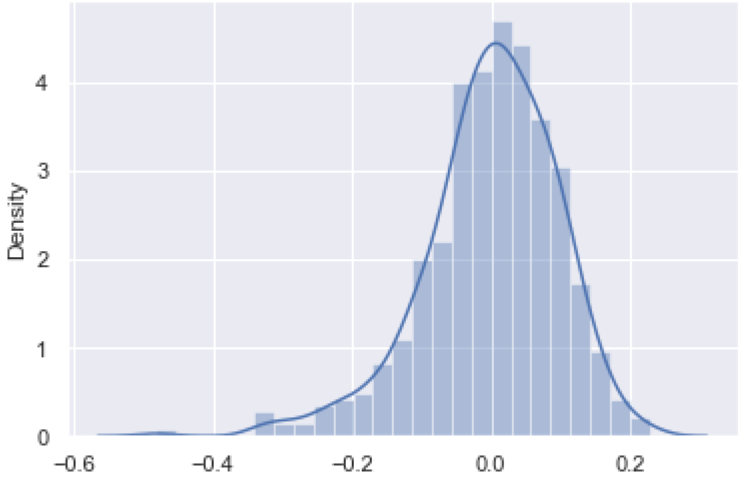
A:



We can observe that temp has the highest correlation with the cnt.

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

A: So, we created 3 models in training set. Last model is something which we accepted it. Because all variables have p-value less than 0.05 and their Variance inflation factor is also less than 5. So, that’s how I handled features. Additionally, I have used RFE to get some features automatically which are significant enough for our model. Also, we did the residual analysis.



We can clearly see from the bell-shaped curve that it is normally distributed. Also, R-square in my training set is 0.830 and R-square in test set is 0.826 which is almost same, and we can say that the model is almost accurate.

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

A:

cnt = 0.231 + temp * (0.427) + windspeed * (-0.098) + yr_2019 * (0.239) + season_spring * (-0.138) + season_winter * (0.078) + mnth_dec * (-0.046) + mnth_july * (-0.063) + mnth_nov * (-0.070) + mnth_sept * (0.051) + weathersit_Light Snow * (-0.280) + holiday_holiday * (-0.100)

From the final model, top 3 features we can say is:

a) **temp**: temp has a positive correlation with cnt, so if there's an increase of 1 unit in temp, the cnt also increases by 0.427.
b) **yr_2019**: A unit increase in yr_2019, it will increase cnt by 0.239.
c) **weathersit_Light Snow**: A unit increase in this variable, it will decrease cnt by 0.280.

General Subjective Questions

1. **Explain Linear Regression algorithm in detail?**

A: Linear Regression is a supervised machine learning algorithm. Supervised because we have a target variable that we need to predict. Target variable needs to be continuous variable. So, it basically defines cause and effect relationship between the dependent variable and the independent variable. Our dependent variable is our target variable that we need to predict, and the independent variables are our predictors. There are two types of linear regression, they are as follows:

1. **Simple Linear regression:** In this we have only one independent variable and 1 dependent variable.

2. **Multiple Linear regression:** In this we have more than 1 independent variables and 1 dependent variable. In this we basically find out how multiple independent variables can change the dependent variable.

General equation of any linear regression model is:

Y = c + (m1*x1) + (m2*x2) +…+(mn*xn)

Y = dependent variable

x1, x2, xn = they are independent variables

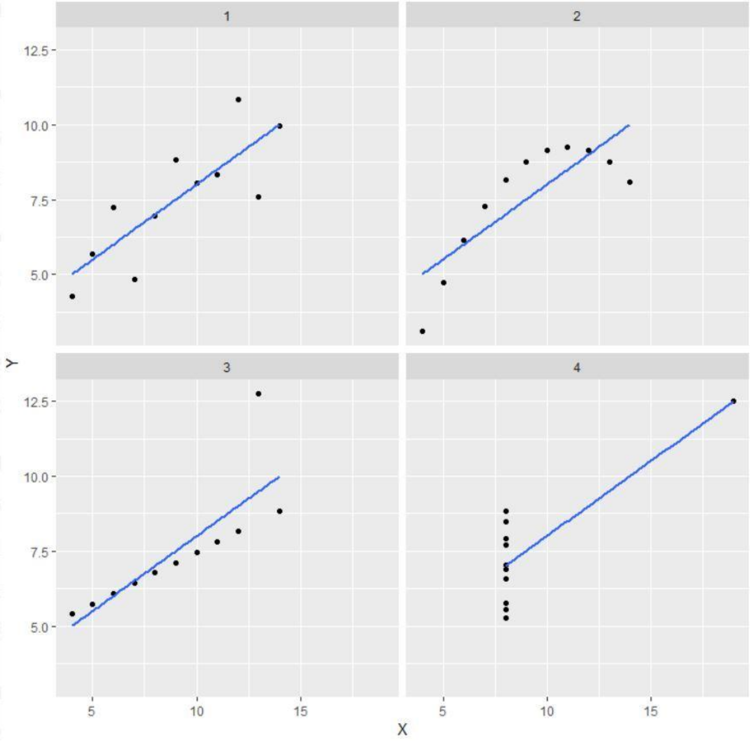
c = intercept

m1, m2, m3 = they are the coefficient of each independent variable. They are also known as slope.

In Linear regression, we always lookout for best fit line. Best fit line can be decided by using gradient descent.

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

A: Anscombe's quartet explains the importance of data visualization.



- In the first chart, the line is fitted and linear
- In the second chart, it is not normally distributed.
- It is linear in 3rd chart, but an outlier still exists
- In 4th chart, an outlier is enough to explain the correlation

So, quartet states that, although having 4 datasets with same statistics, their distribution is quite different. So, quartet gives more importance on data visualization during data analysis.

3. What is Pearson’s R?

A: Pearson’s R which is a coefficient correlation mainly used in Linear Regression. It measures how strong two variables are. The value of coefficient correlations lies between -1 and 1. If value is 1, it states that its perfectly positively correlated. It means that if one variable increases, other increases too. If it’s negative, then it means that it’s perfectly negatively correlated. It means that if one variable increases, other decreases or vice-versa. If the value is 0, then it means that there’s no correlation. Take a note that, it doesn’t talk about causation. It just describes the strength of relationship.

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

A: **Scaling** which is also known as **feature scaling** is a method in which we convert the numerical features in fixed range or in order words we normalize/standardize the features.

Scaling is performed because the datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

- Normalize: It is also known as MinMax() method. This method rescales the value from 0 to 1.
- Standardization: It rescale in such a way that the mean is 0 and variance is 1.

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

A: VIF calculates how well one independent variable is explained by all the other independent variables combined. If VIF value is greater than 5, then we need to remove it. If there’s perfectly positive correlation between the two independent variable, then VIF will be infinity. If R-square is equal to 1, then the VIF will be infinite.

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

A: A Q-Q plot also known as Quantile-Quantile plot. It is a scatterplot. It determines whether two data sets come from the sample or population with common distribution or not. A scatterplot is created by plotting two data sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that is straight.

Importance of Q-Q plot in Linear Regression is that the Q-Q plot is used to see if the points lie approximately on the line.