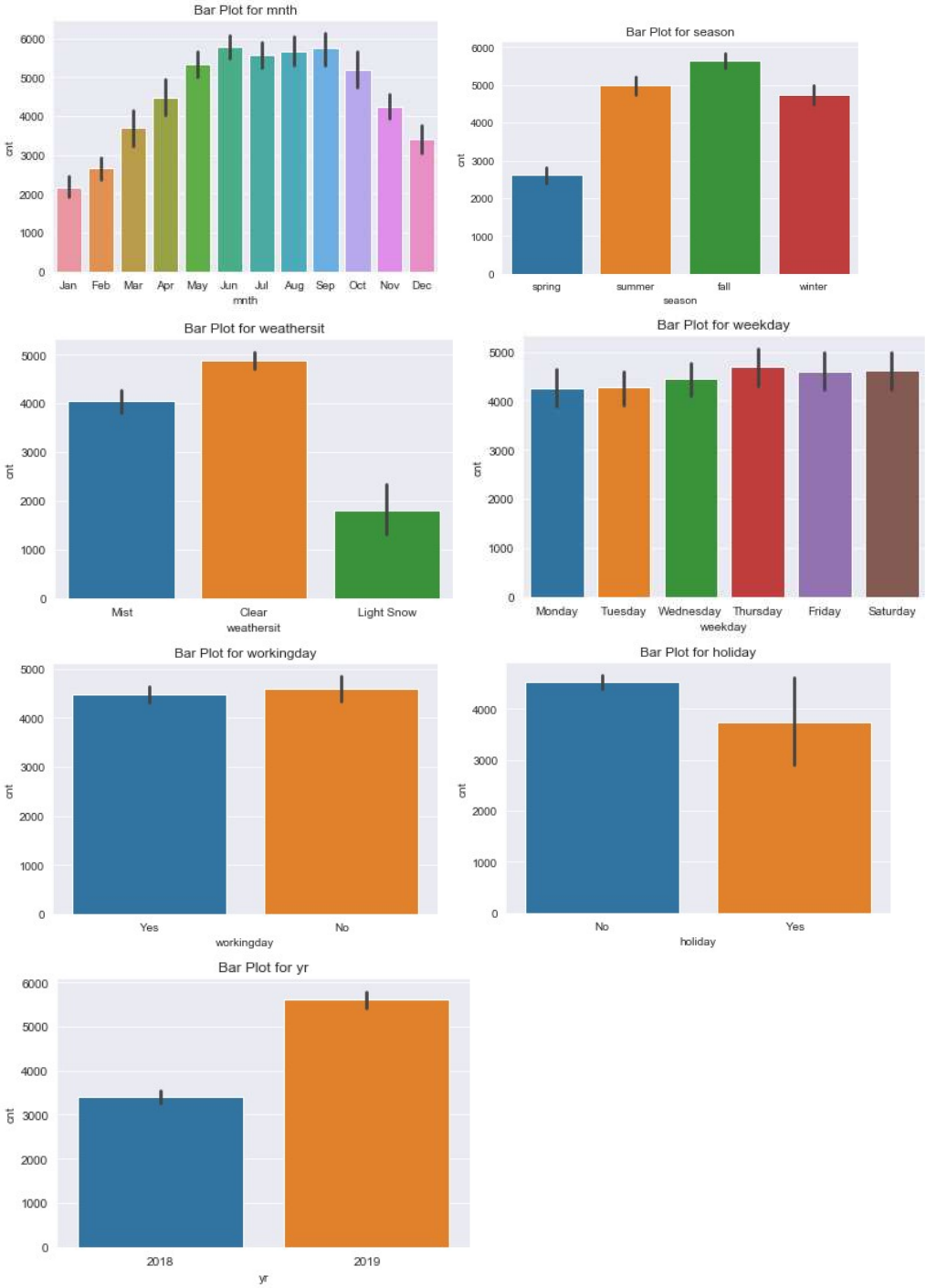
# Assignment Based Subjective Questions:

**Q1.** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?




- The most number of rentals in a season is *fall* followed by *summer*, *winter*. S*pring* seems to be the season with the least number of rentals.

- 2019 seems to have had more rentals than 2018.
- *September* seems to be the month with most rentals and it makes sense as we saw *fall* season with the most rentals. Also we can see that from *Jun* to *Sep* seems to be the period with the most rentals. *Jan* has the lowest rentals.
- Customers rent more when it's not a holiday. This means that possibly, people rent bikes to commute to their office/school/college.
- In a given week *Thursday* seems to be the day with the most rentals with *Friday* and *Saturday* just behind. *Monday* seems to show the lowest rentals during a week.
- Be it a working day or not, the rentals of the bike are somewhat the same.
- Customers rent the bikes most when it's a *Clear* day and of course no one rents bikes when there is a spell of heavy rain.

## Q2. Why is it important to use **drop_first=True** during dummy variable creation?

It is important to use **drop_first=True** during dummy variable creation because it helps us in reducing the extra column created during dummy variable creation. Hence, it reduces the correlations created among dummy variables. For e.g. in the column 'workingday' we had a yes and a no value being returned in it. While creating dummy variables it would have created 2 columns as workingday_Yes and workingday_No, but because we used **drop_first=True** it dropped one as we know if it's not yes it'll definitely be no.

## Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

'registered'

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

There are 4 assumptions associated with a Linear Regression Model:

1. Linearity: There is a linear relationship between the independent and dependent variables.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

Linearity check is the first check that is performed to see if there is a linear relationship between the dependent and independent variables, by plotting a pairplot we can visualize it.To check for normality I plotted a graph to check if there was a normal distribution or not between the error terms, it checked out. For Homoscedasticity and Independence check I plotted a scatter plot between the prediction of the testing dataset and the testing dataset itself, to see if there isn't a pattern between the plotted points and are independent from each other.

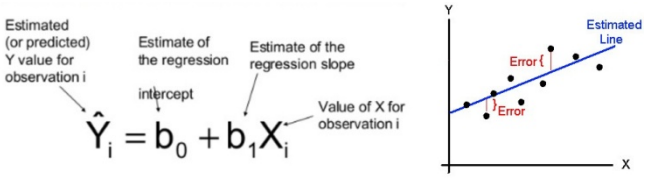
## **Q5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. season_winter
2. season_spring
3. yr_2019

# General Based Subjective Questions:

## Q1. Explain the linear regression algorithm in detail.




Above is the algorithm of linear regression.

Y = the value of the dependent variable or the target variable. This is basically the variable that is going to be predicted using the other independent variable (Here, X).

b0 = this is the intercept. It is the value of Y when X is 0.

b1 = this is the coefficient of X variable, also called the Model coefficient. This value gives us a lot of information about the nature of X. For e.g. if the sign of the coefficient is positive then it means that as the value of X increases the value of Y will increase. Alternatively, if the coefficient sign is negative then it means that as the value of X increases the value of Y will decrease. The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others.

X = this is the independent variable, also called feature variable. This is the feature variable that helps in creating the model to make predictions

In the figure we can see that there is a line in the middle of randomly placed dots. The line is called an estimated line or the regression line. The dots are value points of the actual data and the line is estimation. The difference between them is called error. The main goal is to reduce the mean of the errors or to get the dots to be closer to the line. The lesser the mean of error, the better the model.

## Q2. Explain the Anscombe's quartet in detail.

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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets. It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

## Q3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844. It is commonly represented by the Greek letter ρ (rho). Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient r.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

## Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling, as the name suggests, is a method of scaling the features in a dataset for the purpose of making them comparable against one another. Scaling only affects the coefficients and none of the other parameters like the t-statistic or the F-statistic etc. Scaling is performed for the sole reason of comparison. For example in a dataset having two features as distance travelled in kilometres and another feature as distance travelled in miles are not comparable. To make them comparable, we have to either convert the features having kilometres as unit to miles or miles as unit to kilometres. There are mainly two types of scaling methods:

- Standardisation – This method brings all the data into a standard normal distribution with mean 0 and standard deviation 1. The formula used for Standardisation is x = (x – x(mean))/(std. deviation(x))
- Normalisation – It is a method that scales all the data into the range of 0 to 1. The formula for Normalisation is x = (x – x(min))/(x(max) – x(min)). Normalisation is also referred to as MinMax Scaling. The major difference between scaling and normalisation is that Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

## Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Multicollinearity refers to the problem when the independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables. Two variables are perfectly collinear if there is an exact relationship between the two variables. If the independent variables are perfectly collinear, then our model becomes singular and it would not be possible to uniquely identify the model coefficients mathematically. Hence, collinearity is a problem that needs to be addressed when we are building a multiple regression model. There exist two ways to deal with multicollinearity : Looking at pairwise correlations of different pairs of independent variables. The drawback of this method is that if there are more than 50 such variables, then plotting them against each other would become a very tedious task and inefficient at the same time. There is also a possibility that instead of just one variable, the independent variable may depend upon a combination of other independent variables. This is where the Variance Inflation Factor comes in. Variance Inflation Factor (VIF) indicates how well one independent variable is explained by all the other independent variables. Generally, as a thumb rule we consider a VIF greater than 5 to be a high VIF and hence, such a feature is not a very desirable feature for the modelling process. The formula for calculating the

VIF is An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

## Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. The q-q plot is used to check if the error terms are normally distributed. The advantages of the q-q plot are: ➢The sample sizes do not need to be equal.

- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.