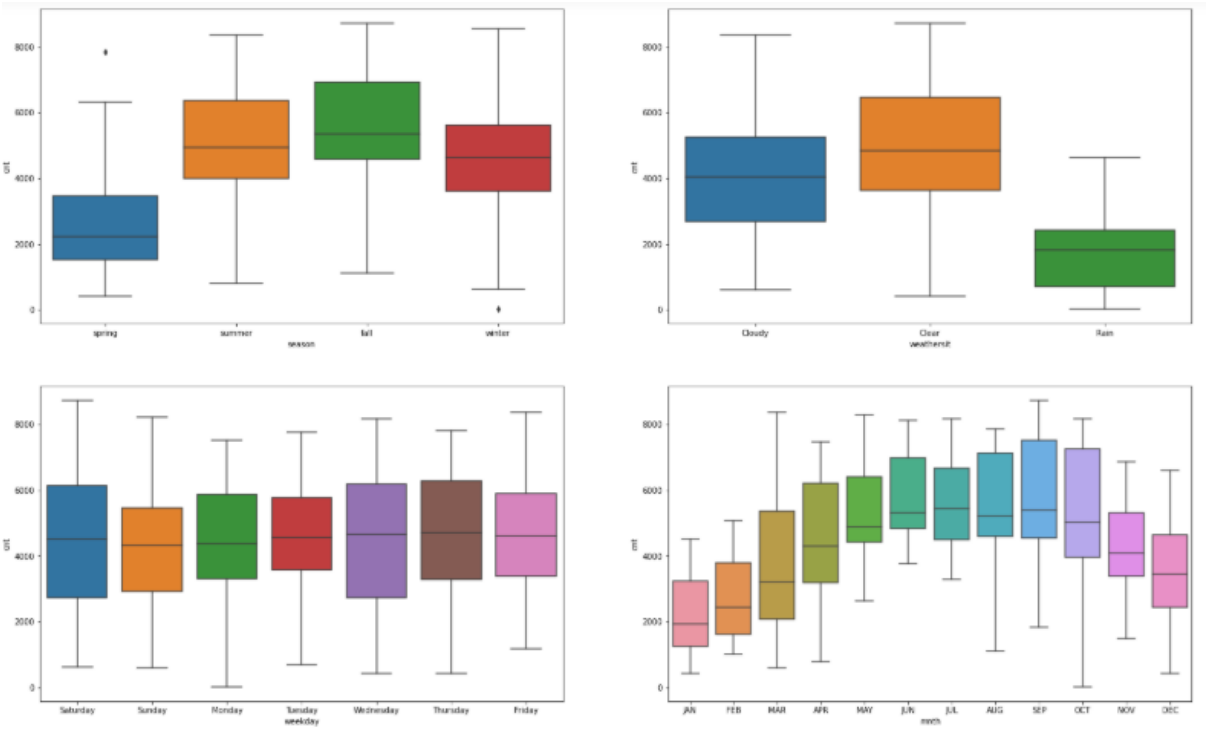
# LINEAR REGRESSION ASSIGNMENT

## 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)**




**From the above plots we can derive that cnt vs season:**

1. - In 1(spring) season the company has experienced a decline in demand

   - where as in 3(Fall) the company has experienced an exponential increase in demand.

2. - **cnt vs weathersit**: 1(clearsky) the demand is high whereas 3(rain/snow) the demand has declined.

3. - **cnt vs weekday** : 0(Sunday), 4(Thursday) during end of week demand seems high where as in weekdays the demand gradually decrease but it seems medians are almost same for weekend and weekdays.

4. - **cnt vs mnth** : From the plot we derive that in the month of 1(Jan) demand is low when compare to other month whereas 9(Sept) demand is high when compare to other months.

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

**A)** To drop a redundant dummy variable i.e the first column from the df we use

'drop_first = True' Which otherwise add unnecessary variables to data which can be even inferred from other dummy variables even if we drop it.

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

**A)** temp (temperature in Celsius ) is highly correlated with target variable (cnt) .

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

**A)** We validate the assumptions of training set after building the model on following parameters and methods/approaches:

- Residual Analysis of the train data : to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression), we plot the histogram of the error terms and infer from it , it should be normally distributed and centred around 0 .
- Making predictions using lm.predict i.e y_train_pred = lm.predict(model)
- We find residual (error terms ) by residual = y_train - y_train_pred & later plot it.
- Finally evaluate R-squared: for train i.e r2_score(y_train,y_train_pred)

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

**A)** From the final model which has a decent accuracy i.e above 80% the most 3 important features which helps in increasing demand are : ‘yr’ , ‘ temp’ , ‘winter ‘ (season) , as the units of these 3 variables increases , the demand in count of total no of rental bikes(cnt) increases.

## General Subjective Questions

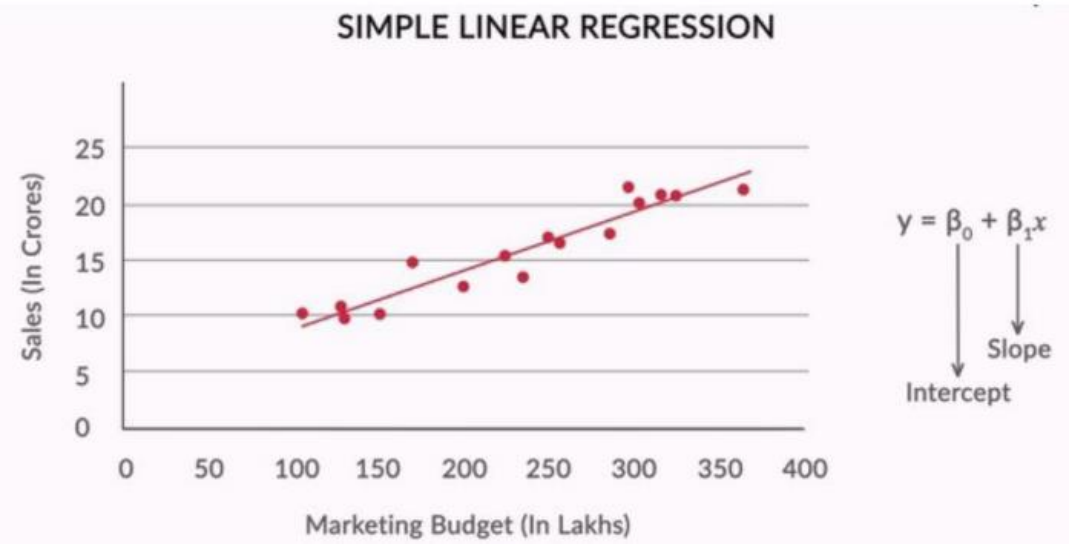
1. **Explain the linear regression algorithm in detail.**

**A)** There is 2 types of linear regression :

**Simple linear regression** is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

The standard equation of the regression line is given by the following expression:

$Y = \beta_0 + \beta_1 X$




**Multiple linear regression** is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

RSS(Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and

the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

RSquare is Mathematically, represented as: $R^2 = 1 - (RSS / TSS)$

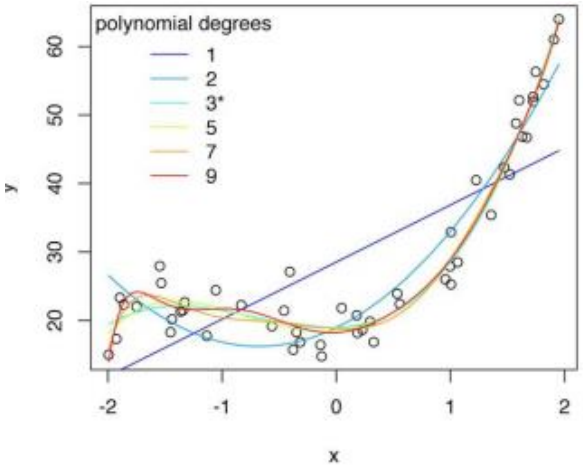



*Fig 12 - Overfitting*

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

**A)** Anscombe's quartet has 4 data sets that have nearly identical simple descriptive statistics, still has very different distributions and appear very different when plotted.

- Each dataset has 11 (x, y) points.
- Each graph tells different inferences of their similar summary statistics.
- The summary statistics show that the means and the standard deviation were identical for x and y across the groups and for the correlation coefficient of every group.

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

**A)** Pearson's R is a parameter which measures the strength of the linear relationship between two variables.

- The best way to check the Pearson's R is by plotting a scatter plot.
- If the variables tend to move up & down at the same time, the correlation coefficient will be positive and in contrast , If the variables tend to moves up and down in opposite/inversely to low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

A) Scaling : It is a technique used to standardize the independent features of the given data set, it is done during data pre-processing.

It is performed to scale/handle the higher range of values or units in comparison to other variables , if it is not done then the model , tends to weigh greater values, higher and consider smaller values as the lower values,

**Most commonly used 2 methods of Scaling are:**

- Normalization : (x-xmin)/(xmax-xmin) Getting min and max values within a fixed range i.e .0-1 . ((Most preferable)
- Standardisation : (x-mean)/S.D : It rescales so that the value has distribution with 0 mean and Variance/S.D equal to 1.

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

A)

- Generally VIF (Variance Inflation Factor) is used for calculating the correlation between the independent variables.
- We get an infinite VIF value because the corresponding variable may be expressed exactly same by a linear combination of other variables.

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

A) Quantiles means the cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

- When the quantiles of two variables are plotted against each other, then the plot obtained is known as Quantile-Quantile plot.
- This Q-Q plot helps to provide a summary of whether the distributions of two variables are similar or not with respect to the locations.

---------- END ----------