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?

Answer: *There were mainly four categorical variables season, month, weekday and weather Taken for analysis and its effect on target variable "cnt" was that:*

1. *For season average median value for cnt was high for summer, winter and fall (above 5000) compared to spring (around 2000)*
2. *For month we can see June to October months had high median value (around 5000) and lowest for months like Jan, Feb*
3. *For weekday we can see not much of variance in terms of median usage value for the bikes.(around 5000)*
4. *For weather clear days had more demand compared to misty and light rainfall and no usage on the days of thunderstorm*

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

Answer: *drop_first=True is important to use, as it helps in reducing the extra column created During dummy variable creation. Hence it reduces the correlations created among Dummy variables. Let's say we have 3 types of values in Categorical column and we Want to create dummy variable for that column. If one variable is not furnished and Semi furnished, and then it is obvious unfurnished. So we do not need 3rd variable to Identify the unfurnished. Hence if we have categorical variable with n-levels, then we Need to use n-1 columns to represent the dummy variables.*

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

*Answer: We can see highest correlation temp/atemp has the highest correlation of 0.63 with Target variable "**cnt**"*

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

Answer: The assumptions of linear regression are:
***Assumption about the form of the model**: We have checked this using various plots like pairplot, heatmaps etc. the dependent variable and different independent variables seem to have a linear relationship*
Assumptions about the residuals:
***Normality assumption**: It is assumed that the error terms, ε(i), are normally distributed. We have checked this using the distplot of residuals (difference between predicted y and trained y). The error Terms seem to have normal distribution and the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero. Also we have checked for **Constant variance assumption** and can see from the plots that the residual terms have the same (but unknown) variance, σ2. Also the residual terms are independent of each other, i.e., their pair-wise covariance is zero.*
*For checking the **multicollinearity** in the data we have constantly used VIF method and removed the variables showing high multicollinearity and the model developed has then been used for testing*

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

Answer: Based on the top 3 features are:

1.) *During the Light rain Light snow Thunderstorm scenario the demand of bike seems To decrease significantly (Coefficient value of -0.31 meaning impacts 31% of the varia nce)*
2.) *We are seeing year on increase of around (0.24) which means demand is increasing With every year*
3.) *We are seeing for the months September(9),August(8),MAY(5) seem to be the target months as they contribute to large part of the overall demand(combined coefficient of 0.41 meaning around 41% demand variance can be explained with this three months*

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer: Linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent Variables. This model is to find a linear relationship between the input variable(s) X and the single output Variable y.

***Simple linear regression**: When there is only single independent/feature variable X then it is called as simple linear regression.*
***Multiple linear regression**: When there are multiple independent/feature variables Xi then it is called as Multiple linear regression.*
*The **independent variable** is also known as the **predictor variable**.*
*The **dependent variables** are also known as the **output variables**.*

$Y = \beta_0 + \beta_1 X$ (SLR)

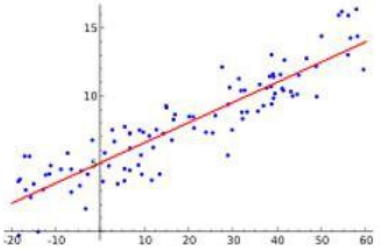
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$ (MLR)

Where:

Y = how far up ↑ and X = how far along →

$\beta_1, \beta_2 .. B_p$ = Slope or Gradient (how steep the line is)

β_0 = value of Y when X=0 (Y-intercept)

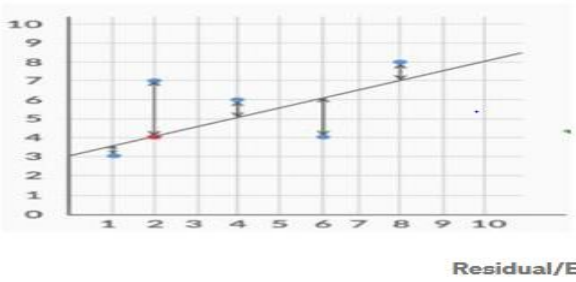


As part of linear regression, there can be multiple lines which can be drawn from the data points as part of scatter plot but regression model can help to identify model that is best fit line from the data points.

Cost Function:

The cost function helps to figure out the best possible values for ϐ0, ϐ1, ϐ2 etc… which would provide the best fit line for the data points. We need to convert this problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.
It means that given a regression line through the data, we calculate the distance from each actual data point to the regression line (predicated values), square it, and sum all of the squared errors together. This is called ***Residual Sum of Squares (RSS)***

Then we divide this RSS values by total number of data points which provides average squared error of all the data points and it is called ***Mean Square Error (MSE).*** *MSE is also known as cost function using which we need to identify optimal values of co-efficient and interceptor such that MSE values settles at minima.*

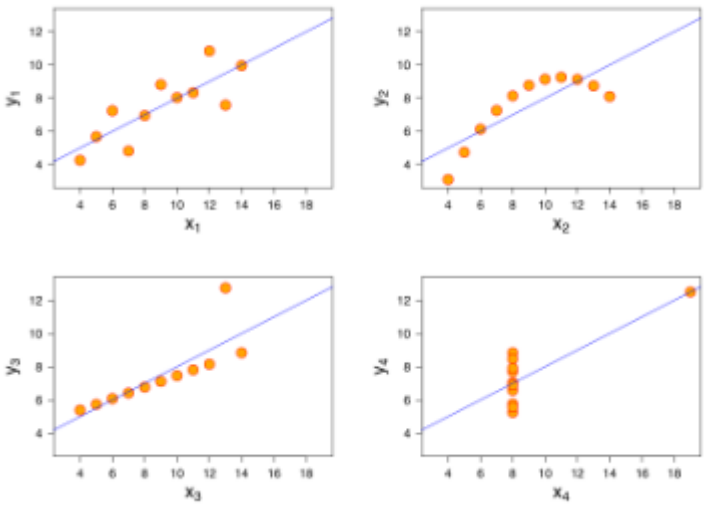


Residual/Error

$$\text{Residual/Error} = e_i = y_i - y_{pred}$$
$$RSS = e_1^2 + e_2^2 + ...+e_n^2$$
$$RSS = \sum_{i=0}^{n}(Yi - \beta 0 - \beta 1X)2$$
$$MSE = RSS/n$$

2. Explain the Anscombe's quartet in detail.

Answer: Linear Regression has shortcomings such as it is sensitive to outliers, models only Linear relationships and few assumptions have to be made to make inference about Model. This is where Anscombe's quartet helps :

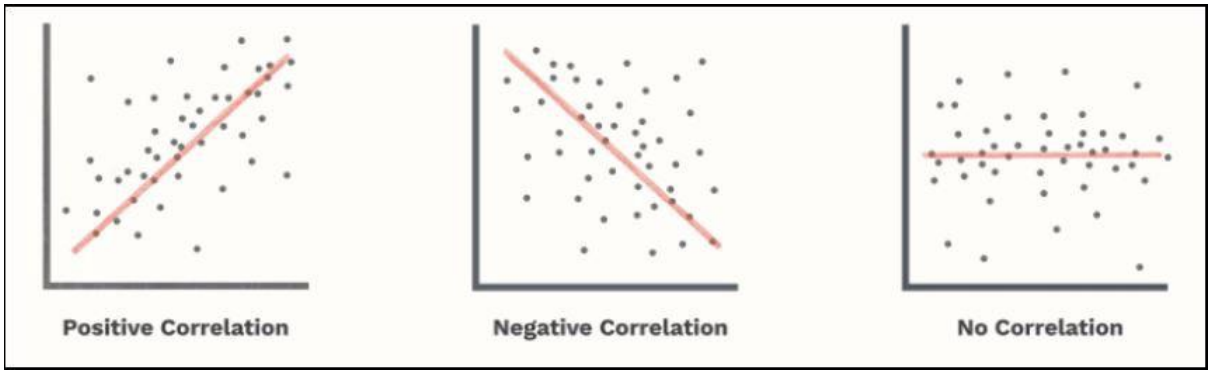


As we can see, all the four linear regression are exactly the same. But there are some Peculiarities in the data sets that have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth

images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have got a great line fitted through the data points. So, we should never run a regression without having a good look at our data.

3. What is Pearson's R?

Answer: Coefficient of Correlation which is also known as ***Pearson's R*** *measures the strength and the direction of a linear relationship between two variables (x and y) with possible values between -1 and 1.*
Positive Correlation*: It indicates that two variables are in perfect harmony. They rise and fall together. +1 is perfect +ve correlation*
Negative Correlation: *It indicates that two variables are perfect opposites. One goes up and other goes down. -1 is perfect -ve correlation*
No correlation: *If there is no linear correlation or a weak linear correlation, r is close to 0.*



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

Answer: We have seen that scaling doesn't impact our model. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer: Infinite VIF occurs for a variable when there is perfect correlation which means that a Variable having infinite VIF can be expressed exactly 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.

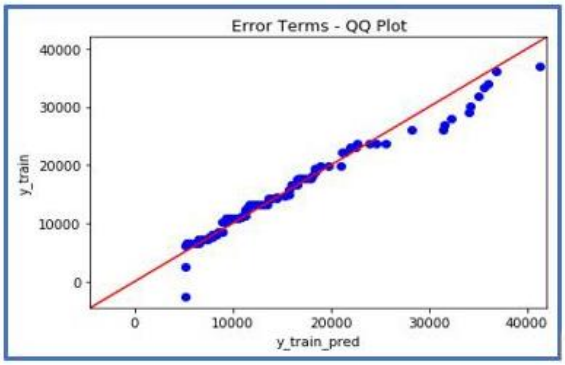
Answer: Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions

Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

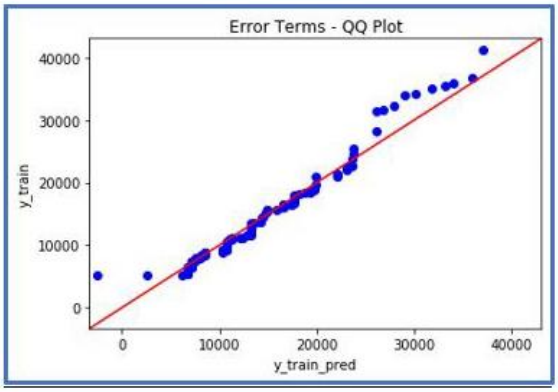
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

*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*

***Python:** statsmodels.api provides qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.*