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?**

   The categorical columns in the dataset were season, weekday, weathersit and mnth. These were visualized by using boxplot. These variables have following effect on the dataset

   - **Season:** Spring season has the least value of cnt, whereas fall season has the highest value of cnt.
   - **Weathersit**: There are no users when weathersit - 'Heavy Rain, Snow & Fog'. This is not a favourable weather condition. whereas the highest count was seen when the weathersit was Clear/ Partly cloudy
   - **Mnth:** Sep is the month with highest number of users, while Jan and Dec shows least number of users.
   - **Weekday:** On weekday the cnt is almost similar, while on weekends (specially on Sun), cnt is higher as compared to weekday.

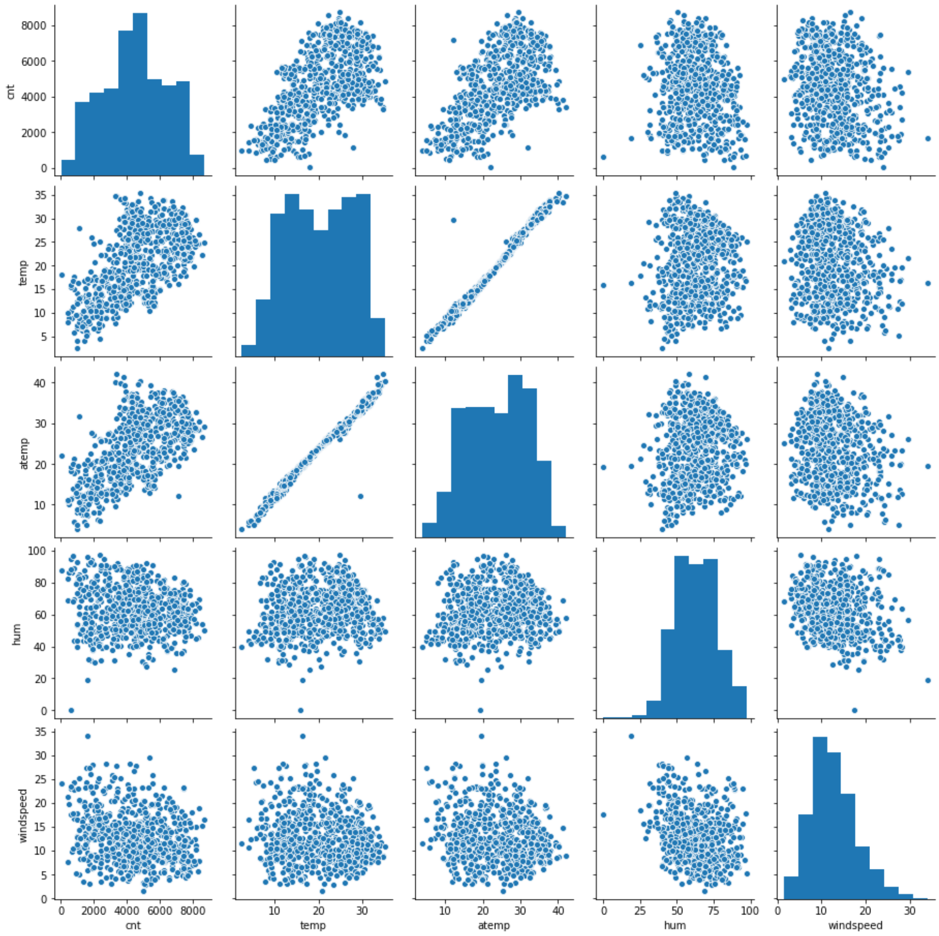
2. **Why is it important to use drop_first=True during dummy variable creation?**

   When we change a categorical variable into dummy variables, we will have one fewer dummy variable than we had categories. If there are p categories than p-1 dummy variable should use. The model should exclude one dummy variable. That’s because the last category is already indicated by having a 0 on all other dummy variables. Including the last category just adds redundant information, resulting in multicollinearity.

   During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

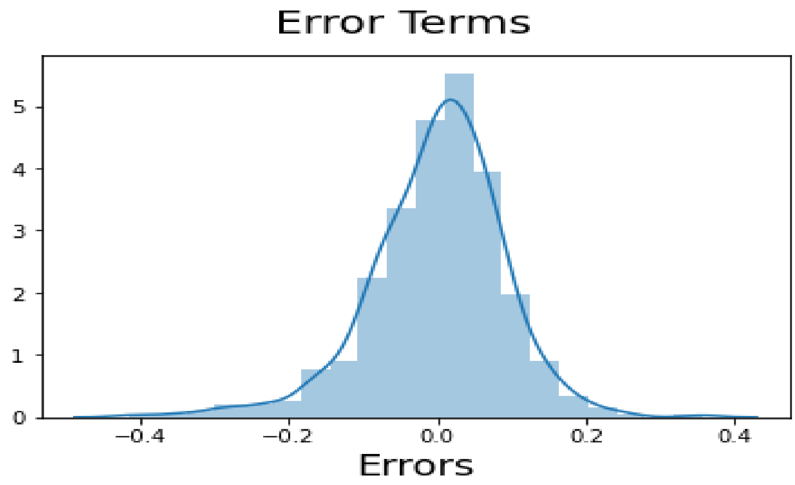
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   As per the graph below, numerical variables **temp and atemp** are the two variables which are having the highest correlation with the target variable **cnt**



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

   Residual distribution should follow normal distribution and centred around 0 i.e., mean =0. We validate this assumption by plotting a graph using distplot of residuals and check if residuals are following normal distribution or not. Below graph shows that residuals are normally distributed (mean=0).



5. **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 as below

   - temp -> Coefficient value: 0.491508
   - yr -> Coefficient value: 0.233482
   - weathersit_Light Snow & Rain-> Coefficient value: 0.285155

General Subjective Questions

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

Linear Regression:

Linear Regression is a statistical supervised learning technique to predict the quantitative variable by forming a linear relationship with one or more independent features.

It helps determine:

- If independent variable does a good job in predicting the dependent variable.
- Which independent variable plays a significant role in predicting the dependent variable.

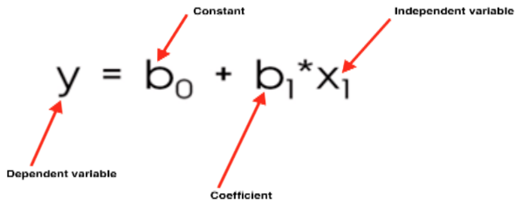
Assumption of Linear Regression

- The Independent variables should be linearly related to the dependent variables (we can use visualization techniques e.g. pairplot to examine the relationship between variables).
- There should be little or no multi-collinearity in the data. (we can perform VIF to check the multicollinearity between the variables).
- The mean of the residual is zero (residual is a difference between observed y-value and predicate y-value).
- Residuals obtained should be normally distributed.
- Variance of the residual throughout the data should be same. This is known as homoscedasticity.
- The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

Types of Linear Regression

- **Simple Linear Regression**

Simple Linear Regression helps to find the linear relationship between two continuous variables, One independent and one dependent feature. Formula can be represented as *y=mx+b* or



- **Multiple Linear Regression**

Multiple linear Regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
The independent variables can be continuous or categorical (dummy coded as appropriate)
Formula can be represented as *Y=mX1+mX2+mX3…+b* or

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

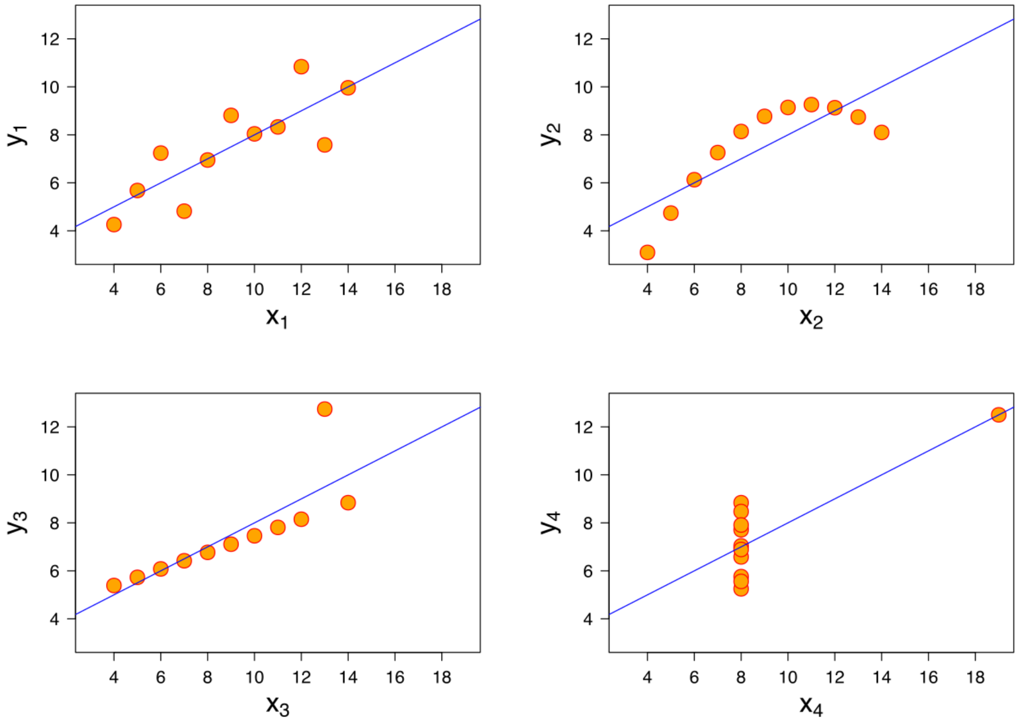
The whole idea of the linear Regression is to find the best fit line, which has very low error (cost function). This line is also called Least Square Regression Line (LSRL).

Properties of a Regression Line

- The line minimizes the sum of squared difference between the observed values (actual y-value) and the predicted value (ŷ value).
- The line passes through the mean of independent and dependent features.

2. 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



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

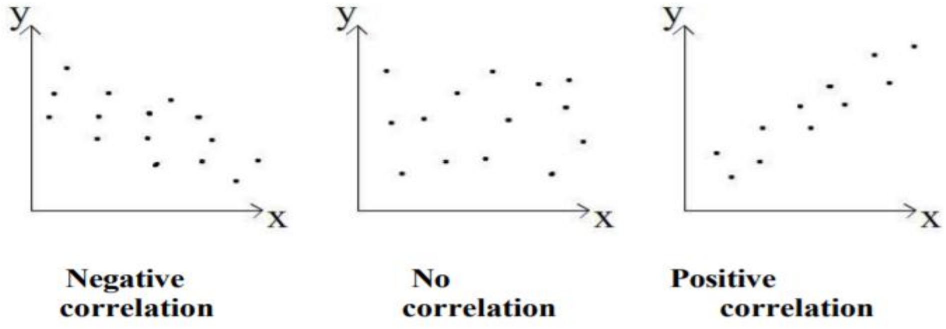
3. What is Pearson's R?

Pearson correlation coefficient also referred to as Pearson's R, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

We can categorise the type of correlation by considering as one variable increases what happens to the other variable:

- Positive correlation – the other variable has a tendency to also increase;
- Negative correlation – the other variable has a tendency to decrease;
- No correlation – the other variable does not tend to either increase or decrease.

The starting point of any such analysis should thus be the construction and subsequent examination of a scatterplot. Examples of negative, no and positive correlation are as follows.



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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units. This will result in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other statistical parameters like t-statistics, F-statistics, p-value, R-squared etc.

Normalized Scaling

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardized Scaling

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

sklearn.preprocessing.scale helps to implement standardization in python

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. VIF measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variable are not linearly related.
It is obtained by regressing each independent variable, for example X on remaining independent variable (Y and Z) and checking how much of it(X) is explained by these variables.

$$VIF = \frac{1}{(1-R^2)}$$

From the formula, if there is a perfect correlation, then VIF=infinite. If the independent variable can be explained by all the other independent variables, then it will have a perfect correlation and it's R-squared value will be equal to 1, So VIF=1/(1-1) which is VIF=1/0 will result in infinity.

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

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 —

- ✓ come from populations with a common distribution
- ✓ have common location and scale
- ✓ have similar distributional shapes
- ✓ have similar tail behaviour

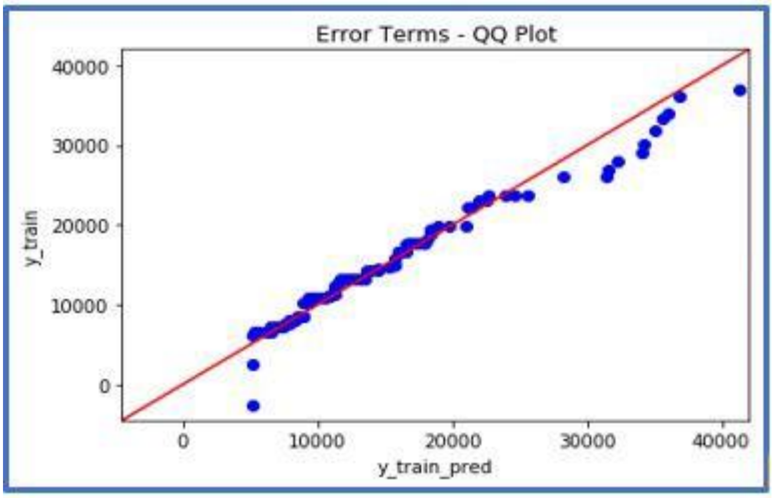
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

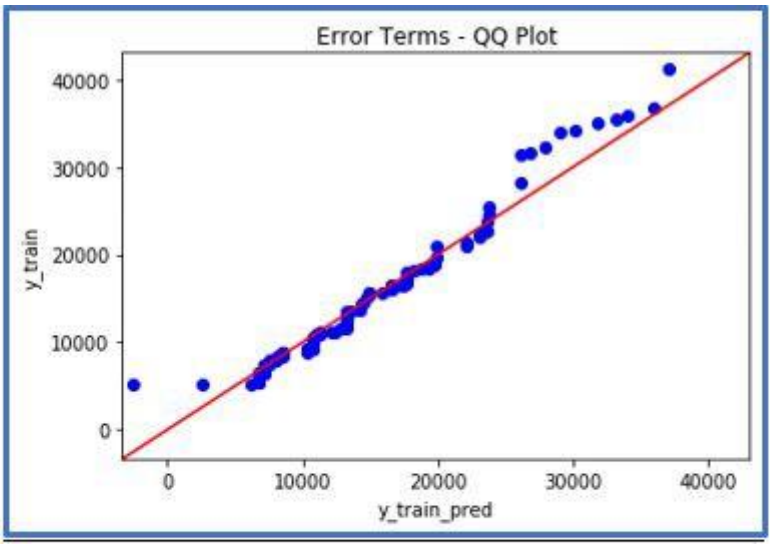
Below are the possible interpretations for two data sets.

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