Assignment-based Subjective Questions and Answers :

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

From the given dataset, the variables except 'dtetday' are numerical types. **Machine learning algorithms** sense these data as numerical ones which may lead to wrong interpretations. (For example, it may interpret 6 > 0, for weekdays). Therefore, we convert these variables to CATEGORICAL by mapping the numeric values with associated labels so that the dataset now contains **meaningful categories** of the variables.

One hot encoding is used here to convert **these categorical data variables** ready to be provided to machine learning algorithms which in turn helps to **improve predictions** as well as **classification accuracy** of the model.

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

It is very important to use **drop_first = True** during dummy variable creation because it **reduces correlation among dummy variables**, which as a result **resolves multicollinearity issue** when we give these dummy variables as inputs to the model building, i.e., the variables become dependent on each other if we use the full set of dummies.

Also, dropping one of the dummies **does not change the interpretation** of categorical variables and **memory** is also handled efficiently.

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

'temp' and **'atemp'** variables are highly correlated with the target variable **'cnt'**.

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

The assumptions made before model building are:

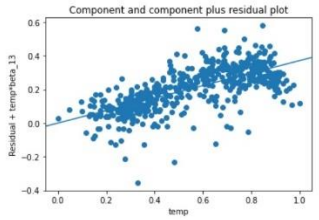
1. **Linear relationship** exists between X and Y
2. Error terms are **normally** distributed
3. Error terms are **independent** to each other
4. Error terms have **constant variance (homoscedasticity)**

These assumptions are validated through scatter plots, heatmap and histograms after model building.

Validation of Assumption 1 : Linearity check

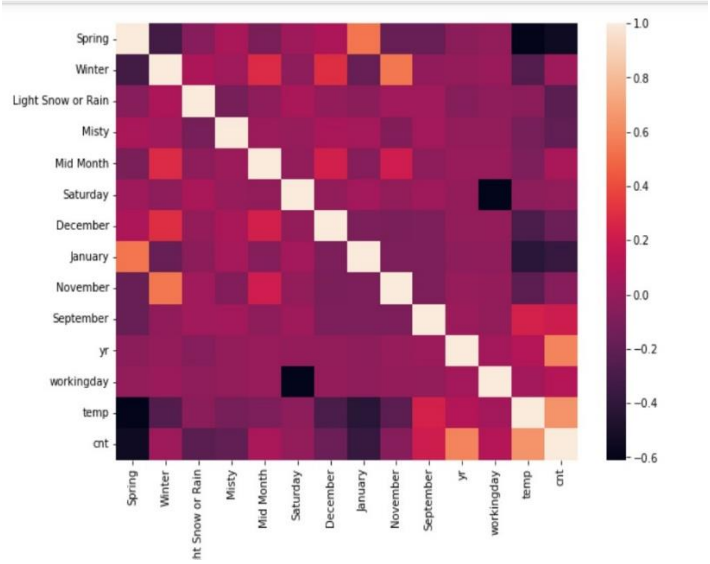
```
# Component and component plus residual plot

sm.graphics.plot_ccpr(lm_8, 'temp')
plt.show()
```



Inference :

The plot of temp with its corresponding predicted coefficient for demand shows that the predictor and the target variable are linearly related
[Linear relationship between X and y - ASSUMPTION No.1 VALIDATED]

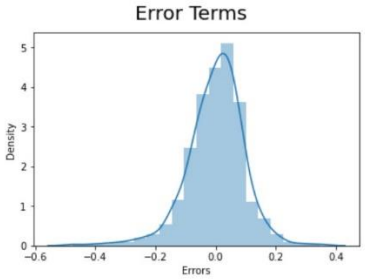


Validation of Assumption 2 : Normally distributed error terms

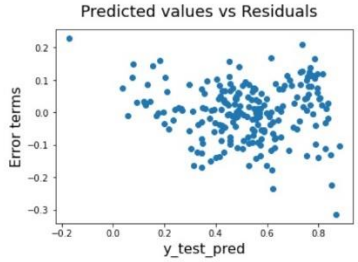
Validation of ASSUMPTION 2 - Error terms are normally distributed

```
# To check whether the error terms are normally distributed, plot the histogram of error terms

fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 10)
plt.show()
```



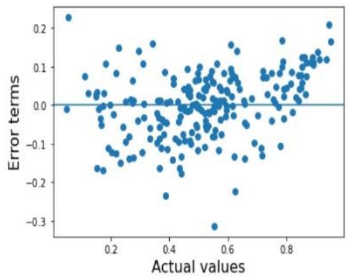
Validation of Assumption 3 : Error terms are independent to each other



Inference from the scatter plot of 'Predicted values vs Residuals' :

No specific patterns observed from the scatter plot of 'Predicted value' vs 'Residuals'. Thus, it shows that **the error terms are independent** to each other with respect to predicted values **[ASSUMPTION No.3 - VALIDATED]**

Validation of Assumption 4 : Homoscedasticity



Inference from the scatter plot of 'Actual values vs Residuals' :

The plot **doesnot show any specific cone or wedge shape**. Hence, it is clear that the variance of the error terms remain constant along the values of the dependent variable; i.e.; **homoscedasticity** exists **[ASSUMPTION No. 4 - VALIDATED]**

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

- Temp (Temperature in degree Celsius)
- Weathersit (Light Snow or Rain)
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression is one of the most popular Machine Learning algorithms based on supervised learning. It performs a regression task and it is a statistical method that is used for **predictive analysis**. Linear regression makes predictions for continuous/real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent **(y)** and one or more independent **(X)** variables, hence called as linear regression.

- **Analysing the correlation and directionality of the data**

  EDA is be used to analyse the data and check for directionality and correlation of data. Visualize the relationship between independent and dependent variables. Linear equation of a (multiple linear regression) model is given by :

  $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

  Where Y = dependent variable

  X1, X2, …,Xp = Independent variables

  β0, β1, …, βp = Coefficients of independent variables

- **Estimating the model, i.e., fitting the line**
  - In linear regression, we have to **find a line that best fits the data points** available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line. The **least squares method** is the most widely used procedure for developing estimates of the model parameters.
  - Least Squares method minimizes the sum of the squares of the residuals (**Residual Sum of Squares (RSS)** between the observed targets in the dataset, and the targets predicted by the linear approximation. Gradient descent, an optimization algorithm is used in the background to minimize the cost function.
  - Summary statistics used are : F-statistic, R-squared, coefficients and their p-values.
  - After fitting the line using least squares method, perform **residual analysis** which plots the histogram of the error terms to check normality. The error terms should be normally distributed with mean 0.
  - Predict the values of dependent variable based on the model built.

- **Evaluating the validity and usefulness of the model.**

  We can evaluate the model by using scatter plot of y_predicted vs y_actual values. Some of the evaluation metrics of how well a model can be fit are :

  Mean Squared Error (MSE)

  The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

  $$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

  Where n = number of samples, yi – predicted value, cap of yi = actual value

RMSE

Root Mean Squared Error is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals. It is an absolute measure.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

R-squared (R2)

It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by independent variables in a regression model. **R-squared** is a **relative measure** of fit. R-squared is given by :

R-squared = 1-(RSS/TSS)

TSS – Total sum of squares

RSS – Residual sum of squares

- If R^2 is high (say 1), then the model represents the variance of the dependent variable.
- If R^2 is very low, then the model does not represent the variance of the dependent variable.

Adjusted R – squared :

The main difference between adjusted R-squared and R-square is that R-squared describes the amount of variance of the dependent variable represented by every single independent variable, while adjusted R-squared measures variation explained by only the independent variables that actually affect the dependent variable.

$$R^2_{adjusted} = [\frac{(1-R^2)(n-1)}{n-k-1}]$$

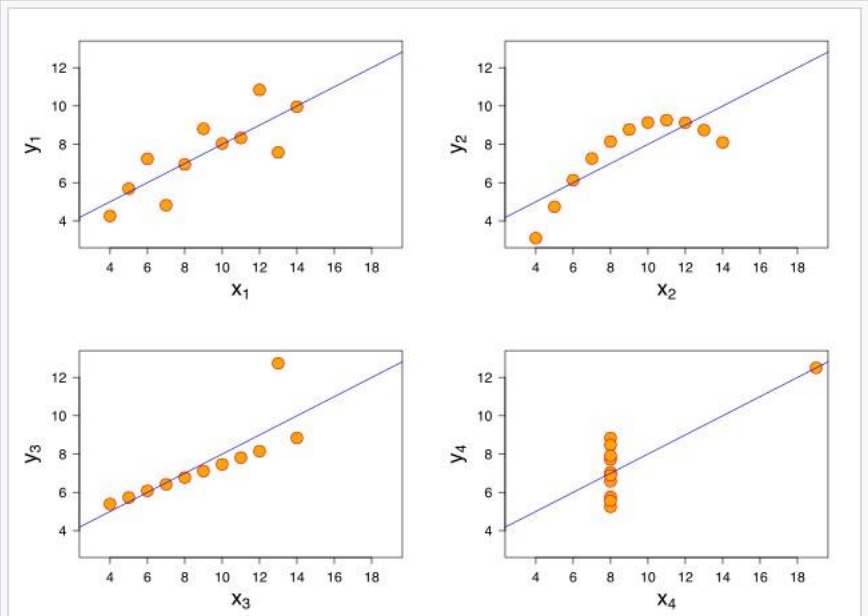
Where k = number of variables, n = number of datapoints

2. Explain the Anscombe’s quartet in detail

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

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 on statistical properties.

EXAMPLE :



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

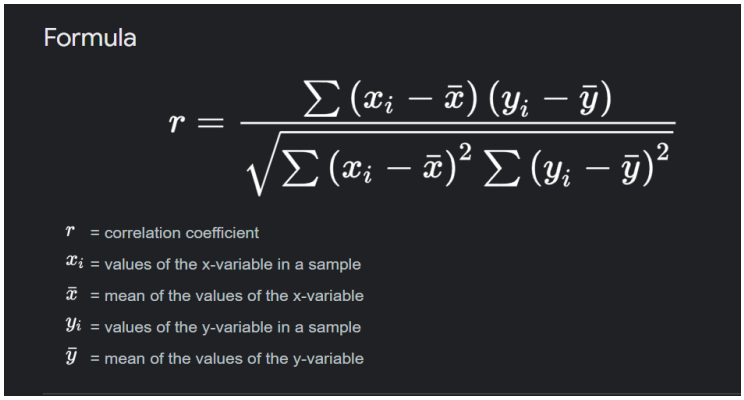
3. What is Pearson's R?

Pearson's R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value **between −1 and 1**.

Correlation shows the strength of the **relationship** between the two variables as well as the **direction** and is represented numerically by the **correlation coefficient (r).** The numerical values of the correlation coefficient lies between **-1.0 and +1.0.**

A **negative** value of the correlation coefficient means that when there is a change in one variable, the other changes in a proportion but in the **opposite direction**, and if the value of the correlation coefficient is **positive**, both the variables change in a proportion and the **same direction.**

Pearson's R formula is given by :



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 hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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-squared, etc.**

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Where x = datapoint, min(x) = minimum data point, max(x) = maximum data point

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Where x = datapoint, mean(x) = average of datapoints,

sd(x) = standard deviation of datapoints

sklearn.preprocessing.scale helps to implement standardization in python.

One **disadvantage of normalization** over standardization is that it loses some information in the data, especially about **outliers.**

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

If there is **perfect correlation**, then **VIF = infinity**. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared =1, which lead to 1/(1-R-squared) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

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,

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

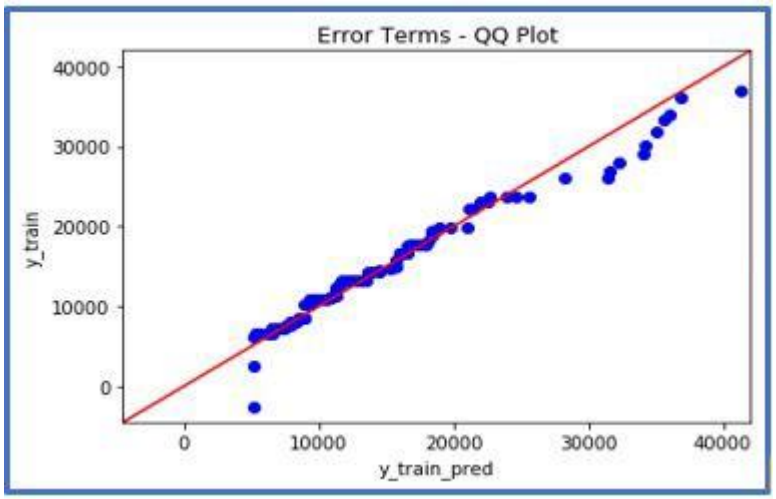
Interpretation:

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

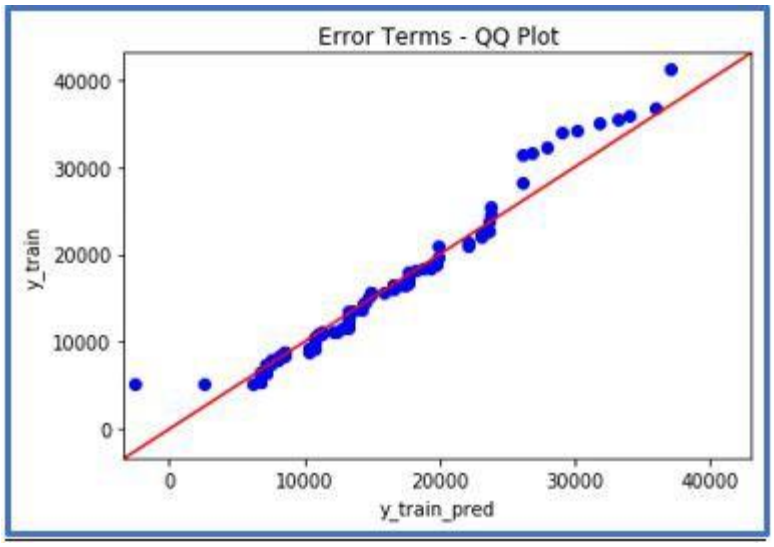
Below are the possible interpretations for two data sets.

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