Assignment-based Subjective Questions & Answers:

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

   - In the day.csv, we have four categorical variables(season, month, weekday, weathersit) out of which two have significant effect on the dependent variable. Those two variables are:
     i- Season
     ii- Month

     Winter from the season variable has a positive 0.0534 coefficient with count variable which states increment in winter season value will lead to demand for shared bikes.

     September from the month variable also has a positive 0.0654 coefficient with count variable which states that higher the value in September month value higher the demand of shared bikes.

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

   - **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.
     Example: Let us say we have 3 types of values in the Categorical column, and we want to create a dummy variable for that column. If one variable is not furnished and semi furnished, then It is obviously unfurnished. So, we do not need the 3rd variable to identify the unfurnished.

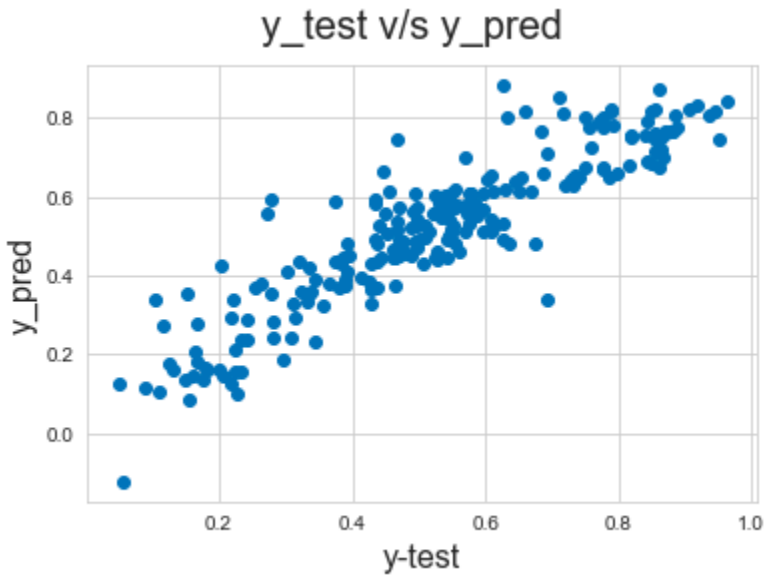
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   - Temperature has the maximum correlation with the target variable though Atemp also has the same correlation with count(dependent variable).

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

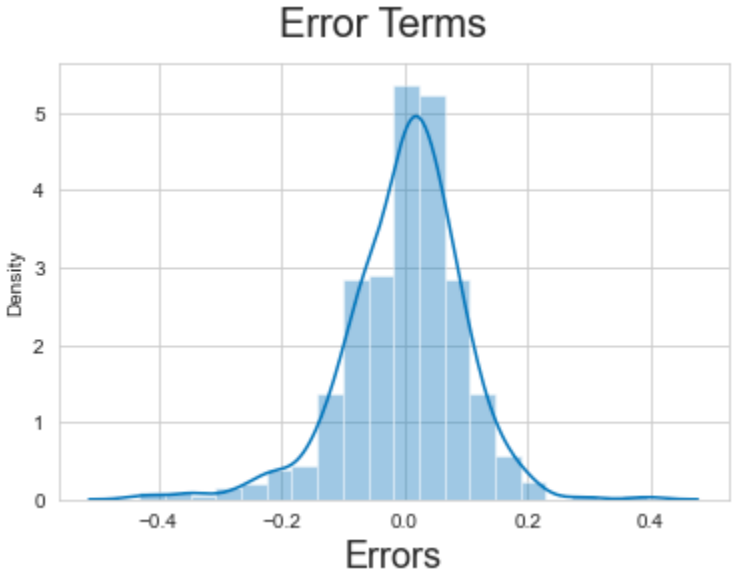
- Below are the assumptions with respect to the multiple linear regression:

  i- There is a linear relationship between x and y i.e., temp and count.



  Predicted values have a linear relationship with actual values

  ii- Error terms are normally distributed:



iii- Independent variables are not highly correlated with each other.
This assumption is tested using Variance Inflation Factor (VIF) values.

	Features	VIF
1	temp	2.99
4	year_2019	2.05
8	weather_mist and cloudy	1.51
3	season_winter	1.33
5	month_Jul	1.33
2	season_spring	1.25
6	month_Sep	1.19
7	weather_light snow and rain	1.06
0	holiday	1.04

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

- Top three most features contributing significantly towards the demand for the shared bikes are:

  i- Temperature

  ii- Winter Season

  iii- Year 2019

  Since these three variables have positive correlation 0.4695 , 0.0534, 0.2332 respectively with count which means increment in these variables will lead to demand for shared bikes.

General Subjective Questions & Answers:

1. Explain the linear regression algorithm in detail.

- Linear Regression is a machine learning algorithm based on supervised learning model. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and predictions. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.

  When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

  The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

  The linear equation assigns one scale factor to each input value or column, called a coefficient, and represented by Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

  In a simple regression problem (a single x and a single y), the form of the model would be:

  $$y = B0 + B1*x$$

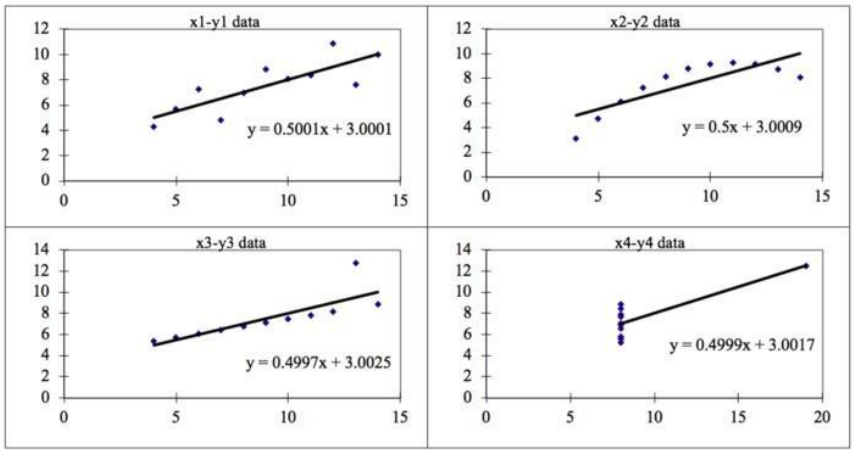
  In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients.

  It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

  When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

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.



The four datasets can be described as:

- Dataset 1: this fits the linear regression model well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

These four describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

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

- **Pearson's correlation coefficient** is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

  Assumptions:

  i- For the Pearson r correlation, both variables should be **normally distributed**. i.e., the normal distribution describes how the values of a variable are distributed.

  ii- There should be **no significant outliers**. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.

  iii- Each variable should be **continuous** i.e., interval or ratios for example weight, time, height, age etc.

  iv- The two variables have a **linear relationship**. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption.

  v- The observations are **paired observations.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

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

- Scaling It is a step of data Preprocessing 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 time, the collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then the algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

  Normalization/Min-Max Scaling:

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

MinMax Scaling: X = (X-min(X))/(max(X)-min(X))

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).

Standardization: X = (X - mean(X))/sd(X)

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?**

- Variance inflation factor (VIF) calculates how well one independent variable is explained by all the other independent variables combined, it is a measure of the amount of multicollinearity in a set of multiple regression variables. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

$$(VIF)I = 1/(1-Ri^2)$$

According to the formulae for VIF, VIF can be infinite only if R-squared value is 1. So, we can conclude that if there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

- A Q-Q plot or Quantile-Quantile plot is a plot of two quantiles between each other. It will help us assess if a set of data comes from some theoretical distribution i.e., Normal distribution or exponential distribution.

  While building a linear regression model, we assume that the residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption.

  Q-Q plot also 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.