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?**

   Following are the categorical variables:

   i. season
   ii. weatersit
   iii. holiday
   iv. mnth
   v. yr
   vi. weekday

   The following inferences can be drawn:

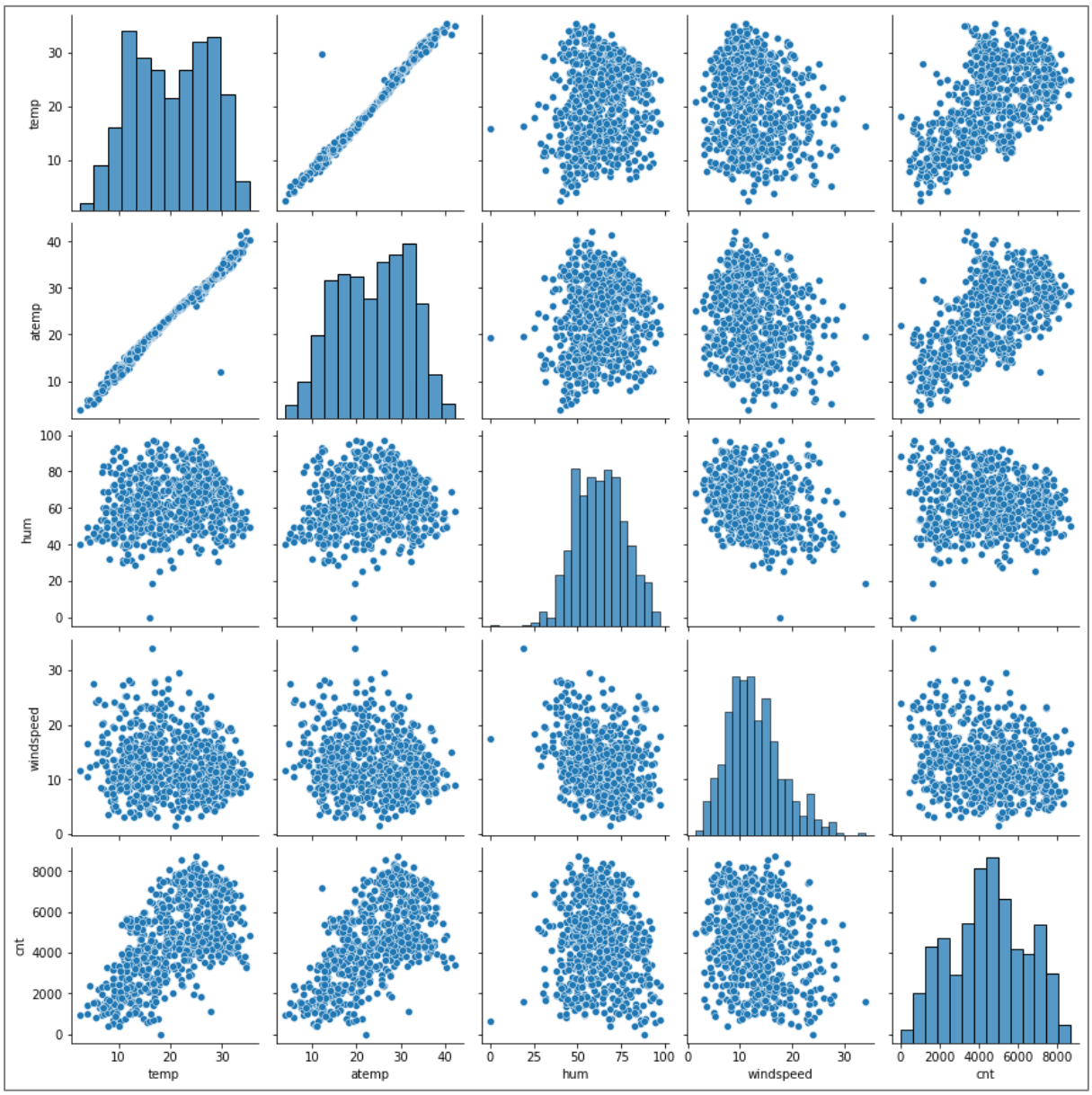
   - Bike sharing is least in the spring **season**.
   - The count variable is less during the **holidays**.
   - Fall season is having the highest demand for rental bike.
   - Demand for rental bike is increasing till the **month** of June, then there is a fallback of demand. For the month of September, demand is highest for the year and then demand for rental bike again decreases.
   - Demand for bike rent decreases in the **year** beginning and end. This may be due to bad weather condition. Also the number of rental is more in 2019 than 2018.
   - The demand does not give a clear picture whether it is a working day or **holiday**.
   - The demand increases for good **weathersit** drastically.

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

   drop_first = True reduces the extra variables created during creation of the dummy variables. This implies it also reduces the correlation created between the dummy variables.

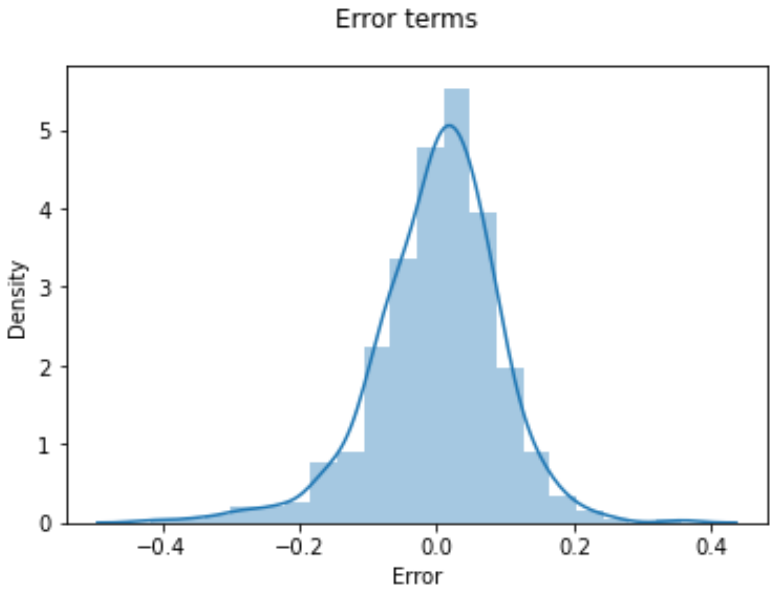
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

temp and atemp are the two variables who was having the highest correlation with the target variable. Please find below pair-plot for reference:

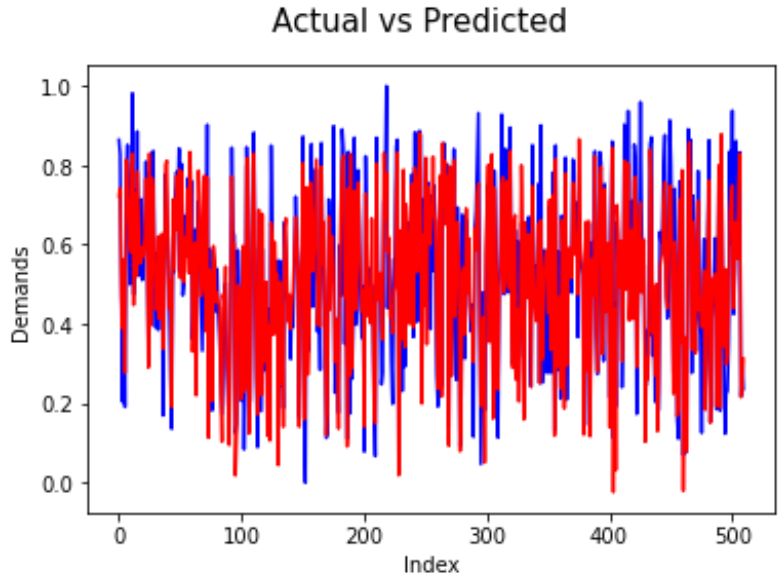


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

To validate the assumptions we should verify the residual to follow a normal distribution and mean = 0. Please find below chart which verifies the normal distribution along with mean = 0:



More over, the actual vs predicted should also be similar. Please find below chart which verifies that:



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

Following are the top 3 features contributing significantly to the demand of shared bikes to our final model:

i. temp is directly proportional with coefficient of 0.491508
ii. yr is directly proportional with coefficient of 0.233482
iii. weathersit_Bad is inversely proportional with coefficient of -0.203597

General Subjective Questions

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

Linear Regression in Machine Learning is a statistical regression method which can be used to predict the analysis and visualize the relationship between the continuous variables. It is based on the equation:

y **= mx + c**
where, m = gradient
c is the intercept of y-axis

Regression tries to find the best fit line between the dependent and the predicted variables with minimal error.

It shows the linear relationship between the dependent variable (y-axis) and the independent variable (x-axis). It can be broadly divided into two types:

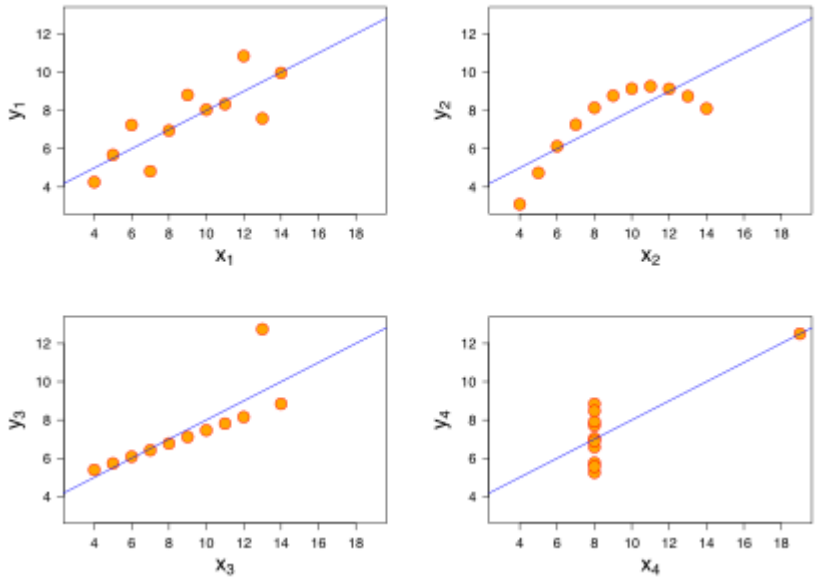
i. Simple Linear Regression – When the dependent variable is predicted using only one independent variable.
ii. Multiple Linear Regression – When the dependent variable is predicted using multiple independent variables.

Some use cases: We can use linear regression to predict the following:

- Predict the sales target for a company
- Predict scores of a student
- Predict change in stock price etc.

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

Anscombe's quartet is a group of four datasets which are identical statistical features but having different distribution and looks different when we try to plot in a scatter plot. It was developed by Francis Anscombe. It helps us to illustrate the importance of plotting the graph before we start analysing the model. The four types of charts are:



i. One chart appears to be having simple linear relationship.

ii. Second chart could not fit the linear regression model and shows as non-linear.

iii. Third chart shows the outliers involved

iv. Fourth chart also shows the outliers involved with one high leverage point which produce a high corelation coefficient.

3. What is Pearson's R?

It is a test statistic that measures the statistical relationship between two continuous variables. It ranges from -1 to +1 where;

r = 1 means the data is linear with positive slope

r = 0 means no linearity

r = -1 means the data is linear with negative slope

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

Scaling is a method which is used to normalize or standardize some independent variables of the data set. Scaling is performed at the pre-processing stage so that we can deal with the varying values in the entire dataset. Else if the units of the values are different and not standardized then it tends to give higher value for higher number and lower value for lower number.

Normalized scaling brings all the data in the range of 0 and 1. Minmaxscaler helps to implement normalized scaling where as in standardized scaling it replaces the values by z scores. One disadvantage of normalized scaling is it misses out the outliers as it ranges from 0 to 1.

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

   The Variance Inflation factor (VIF) is infinite when there is a perfect corelation. If the independent variable can be explained perfectly by another independent variable, then it results in perfect corelation. And its R squared value is 1.

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

   Quantile-Quantile plots otherwise called Q-Q plot are two quantiles against each other. It is used to compare the shapes of distributions through scatterplot. If the two data sets are from a common distribution, the points will fall on the reference line. If two distributions are compared and the Q-Q plot approximately lie on the line, then the distribution is linearly related. We can interpret the Q-Q plot using the following:

   i. Similar distribution: If all the quantiles lie on a straight line or close to a straight line at an angle 45 degree from the x axis.
   ii. X value < Y value: If x quantiles are lower than y quantiles.
   iii. Y value < X value: If y quantiles are lower than x quantile.
   iv. Different distributions: If all the quantiles are away from the straight line at an angle 45 degree from the x axis.