Linear Regression Subjective Questions :

by Aashka Vijapura

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? (3 marks)**

According to the analysis done there are these categorical variables present in the dataset, "season", "yr", "mnth", "holiday", "weekday", "workingday", "weathersit". And from the model chosen for prediction these are the categorical variables with their effects on the dependent variable:

a. mnth - This categorical variable represents 12 months from Jan to Dec and according to analysis September had the highest demand also in January, February and December the demand is least which is totally logical as the weather situation is severe as these months have heavy snow fall.

b. yr - This categorical variable represents two years 2018 and 2019 and according to analysis the demand was more in 2019 than in 2018 maybe because the company gained more popularity across the year 2018 so business increased in 2019.

c. holiday - This categorical variable represents whether it is a holiday or not, and the analysis shows that during holiday the demand increases, and that might be due to people going for outing etc.

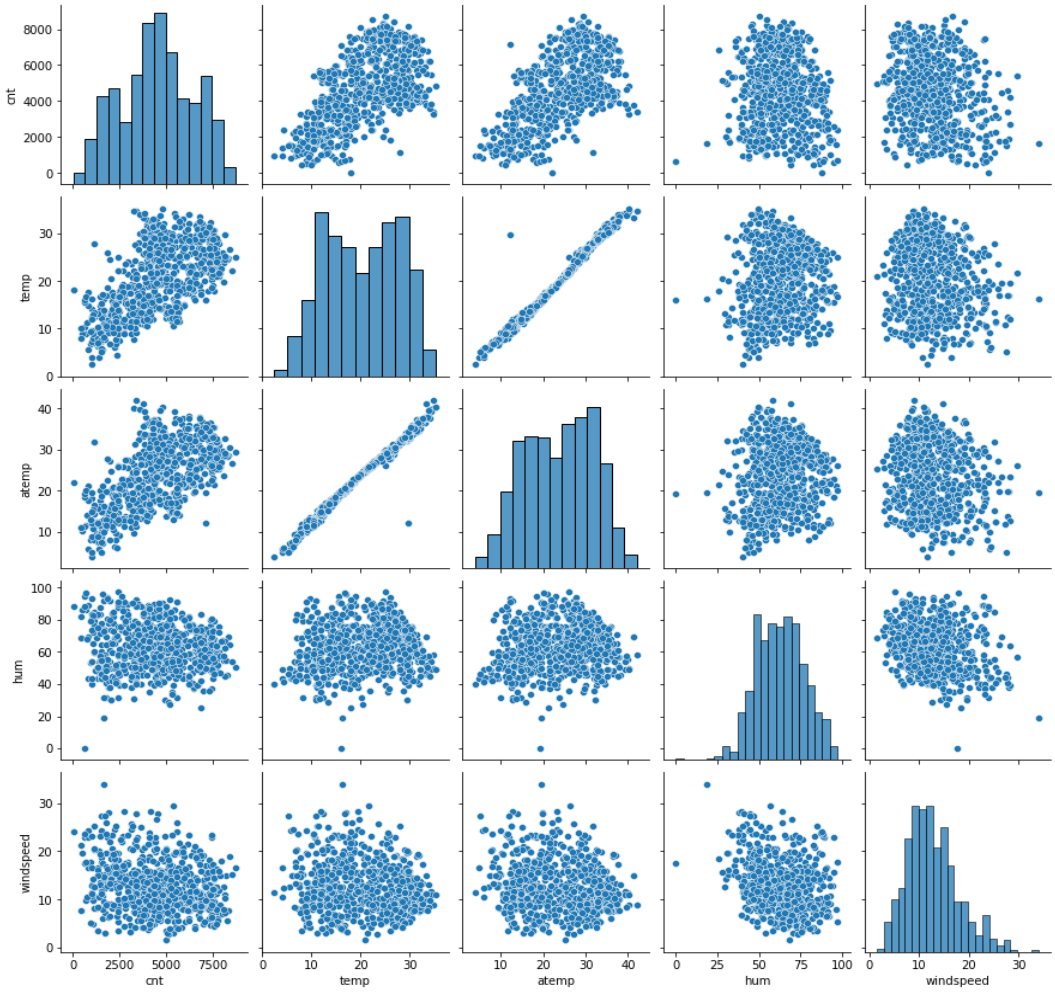
d. season - This categorical represents different seasons namely, spring, winter, summer, autumn, according to analysis the demand is the least in spring and the most in autumn and its almost medium in summer and winter, the company should focus on increasing their sales in spring.

e. weathersit – This variable represents the weather condition i.e., moderate, good, bad, severe, from the analysis it was visible that during good weather condition i.e., Clear, few clouds, partly cloudy, partly cloudy the demand is highest and during bad weather condition the demand is lowest.

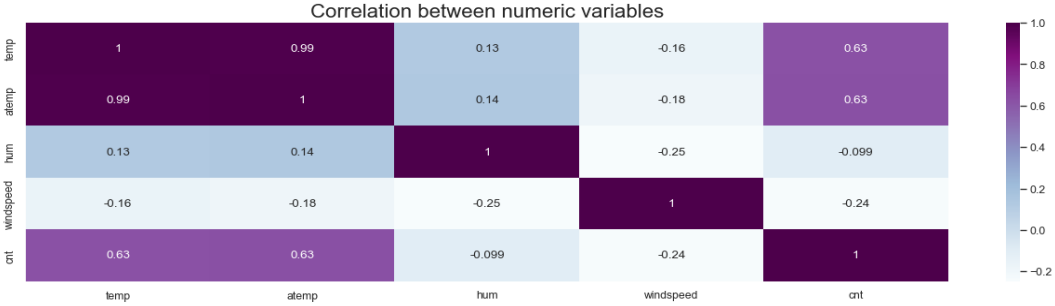
2. **Why is it important to use drop_first=True during dummy variable creation? (2mark)**

It is important to use drop_first=True because if we don't drop the first column then the dummy variables will be correlated which might affect the model majorly. Hence if there are n levels in a categorical variable we only need n-1 dummy variables to explain that categorical variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

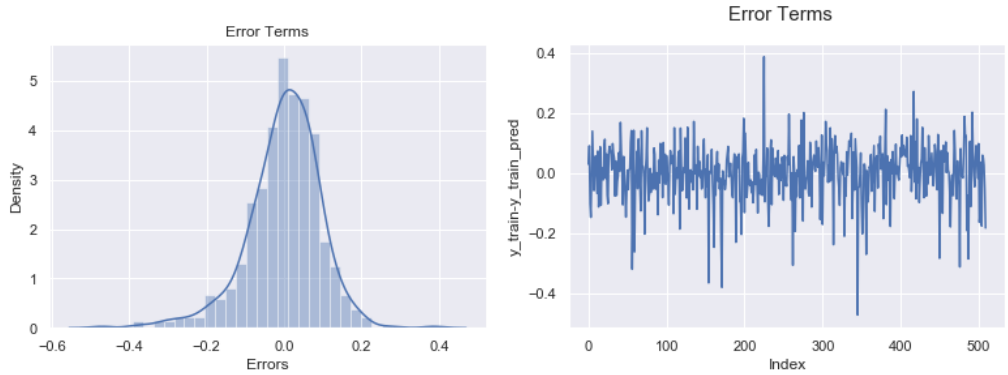


From the above given graph, we can clearly see that "temp" and "atemp" are the two variables which show highest correlation with the variable "cnt" which is the target variable and this can also be seen from the heat map as show below:

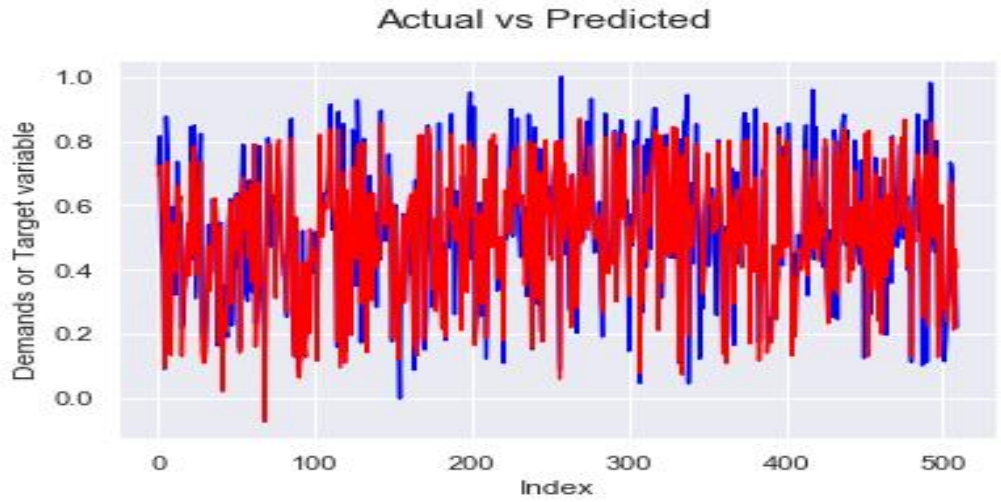


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

If a model is good, it should have its residuals normally distributed and should be centred at 0 and also should be independent of each other as seen from the below two graphs:



We can also prove that a model is valid by plotting the actual v/s predicted line graph and check how well they overlap as given below:



And its also safe and full proof to compare the r-squared value that we got for our model and the actual value.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

These are the top three features that contribute significantly towards explaining the demand of shared bikes:

Temperature (temp) - A coefficient value of '0.3746' indicated that a unit increase in temp variable increases the bike hire numbers by 0.3746 units.

weathersit_bad - A coefficient value of '-0.3160' indicates that with unit increase in weathersit_bad the demand of bikes will decrease by 0.3160 units.

Year (yr) - A coefficient value of '0.2357' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2357 units.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   Linear Regression is a linear approach and a machine learning algorithm based on supervised learning, used to predict the target variable based on the independent variable/variables. The target variable is also called the dependent variable as its value depends on the other variables called independent or explanatory variables. This approach will basically estimate the coefficients of all the independent variables involved in the linear equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_p X_{p+} \varepsilon$. Here β_0 is the intercept and β_1, β_2 etc are the coefficients. By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, the values of intercept and the coefficients are updated in such a way that the error reduces

   In regression with multiple independent variables, the coefficient tells how much the dependent variable is expected to increase when that independent increases by one, holding all the other independent variables constant.

   In simple linear regression there is only one independent variable and in multiple linear regression there are multiple independent variables.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

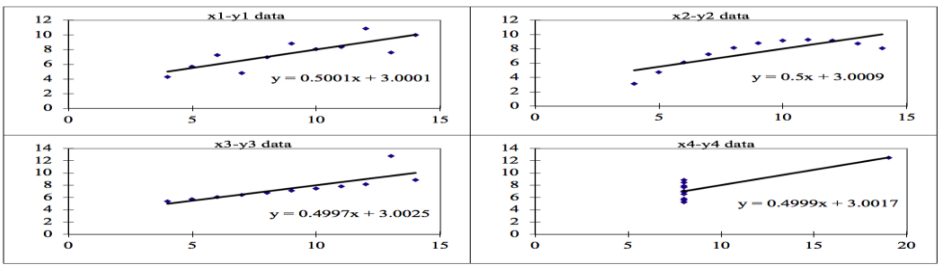
   Anscombe's quartet is basically a group of four data sets which has identical simple descriptive statistics, but when plotted they have very different distribution of values and this fools the regression model. This tells us the importance of data visualization before building the models which will explain different anomalies present in the dataset or any outliers which might affect the prediction majorly, linear regression is only fit for data showing some kind of linear relationship between its independent and dependent variables, same model wont work for the datasets having different type of distribution of values. Below given is an example of Anscombe's quartet

   Dataset 1: this fits the linear regression model pretty 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



3. **What is Pearson's R? (3 marks)**

Pearson's correlation coefficient measures the statistical relationship between two continuous variables, it is based on the method of covariance, it gives the magnitude as well as the direction of the relationship between the two variables.
It is not capable of capturing the non-linear relationships. It assumes numerical values lying between -1.0 and 1.0. It is also not capable of differentiating between independent and dependent variables. When r=1 means there is perfect linear relation with positive slope, if r= -1 means there is perfect linear relation with negative slope, r=0 which means there is no relation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling is the method used to standardize or normalize the values of the features or the variables. The scaling is performed in the data pre-processing stage to deal with varying values so the final results don't get affected, if scaling is not done the machine will give higher weightage to higher values lower weightage to lower values and hence it will prioritise the values and won't give the perfect result.
Normalization is used when you know that your dataset does not follow gaussian distribution and its only useful for algorithms that do not assume he distribution of the dataset. Itis 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.
Standardization is another scaling technique where the values are centred 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.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

The VIF (variance inflation factor) gives how much the variance of the coefficient estimate is being affected by collinearity. (VIF) $=1/(1-R_i^2)$. If there is perfect correlation, then VIF = infinity. Where R_i is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in infinite value.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of

quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution
- Do two data sets have common location and scale
- Do two data sets have similar distributional shapes
- Do two data sets have similar tail behaviour