Answers of Assignment-based Subjective Questions

(Md Merajul Islam)

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

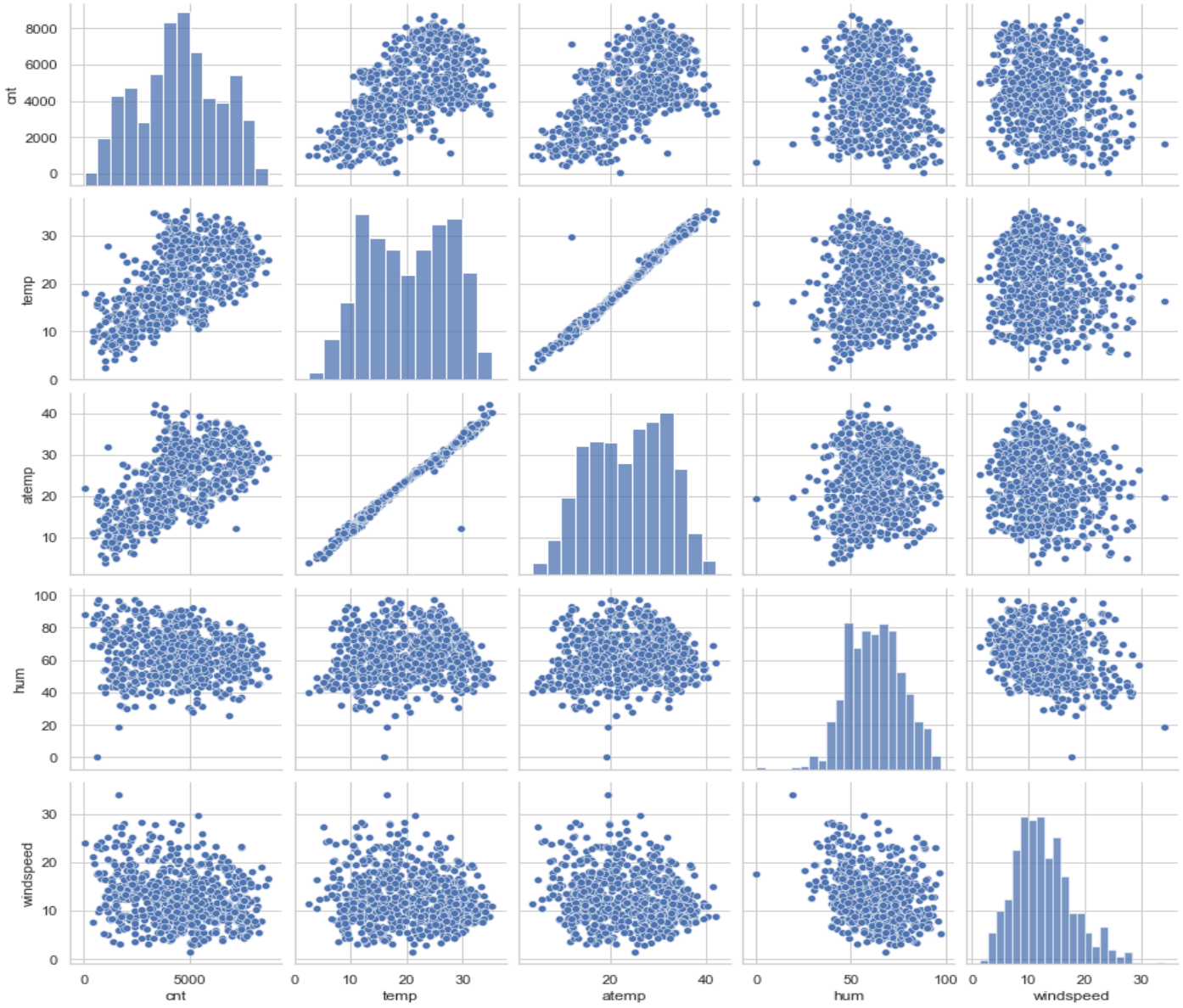
The categorical variable in the dataset were *season, weathersit, holiday, mnth, yr* and *weekday*. These were visualized using a boxplot. These variables had the following effect on our dependent variable:-

1. **Season** - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
2. **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. Highest count was seen when the weathersit was “Clear, Partly Cloudy”.
3. **Holiday** - Rentals reduced during holiday.
4. **Mnth** - September saw highest no of rentals while December saw least. This observation is on par with the observation made in weathersit. The weather situation in December is usually heavy snow.
5. **Yr** - The number of rentals in 2019 was more than 2018.

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

If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance’s may be distorted. Another reason is if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column.

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

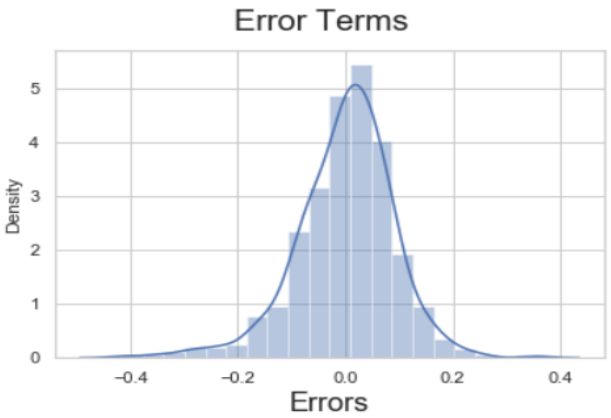


As per the above pair-plot “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

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

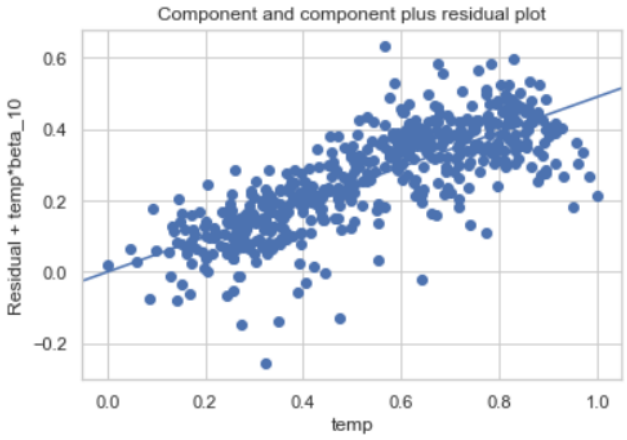
Below are the steps through which we have validated the assumptions of Linear Regression:

Normality of Errors



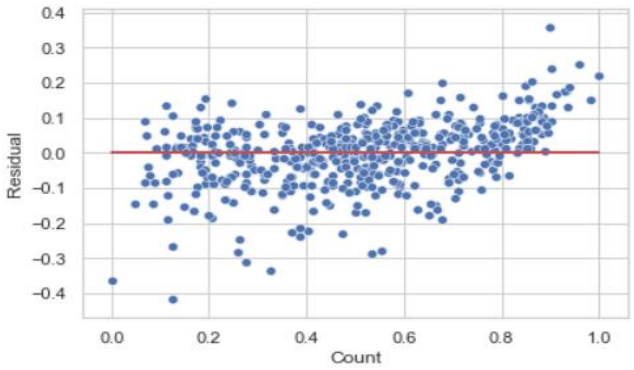
Residuals distribution should follow normal distribution and centered around 0(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The diagram shows that the residuals are distributed about mean = 0.

Linear Relationship



The partial residual plot represents the relationship between the predictor and the dependent variable while taking into account all the other variables. As we can see in the plot, the linearity is well respected.

Homoscedasticity



If all its random variables have the same finite variance then a sequence of random variables is homoscedastic.

As we can see in the above plot, Homoscedasticity is well respected since the variance of the residuals are almost constant.

Independence of residuals (absence of auto-correlation)

Autocorrelation refers to the fact that observations' errors are correlated. To verify that the observations are not auto-correlated, we have used the Durbin-Watson test. If the test output values between 0 and 4, the closer it is to 2, the less auto-correlation there is between the various variables (0–2: positive auto-correlation, 2–4: negative auto-correlation). While tested for our model we got a value of **2.0509**, which means There is almost nil auto-correlation.

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

The top 3 features are:

1. **temp** [coefficient : 0.490988]
2. **yr** [coefficient : 0.233570]
3. **weathersit_Light Snow & Rain** [coefficient -0.284199]

Answers of General Subjective Questions

(Md Merajul Islam)

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation "y = mx + c". It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression:** SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression:** MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

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

where, for i=n observations:

y_i = Dependent variable

x_i = Explanatory variables

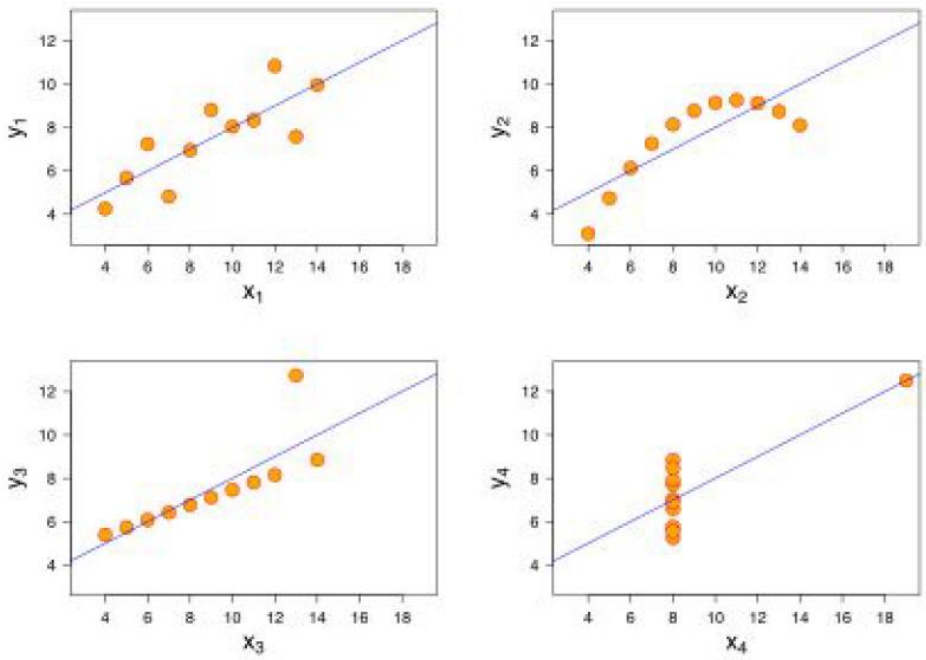
β_0 = y-intercept (constant term)

β_p = Slope coefficients for each explanatory variable

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

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing 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.
- The second graph (top right) is not distributed normally, while there is a relation between them, it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line 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? (3 marks)

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 normalized 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. In simple terms, it tells us *can we draw a line graph to represent the data?*

The Equation for calculating Pearson's r goes as below:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

r = 1 means the data is perfectly linear with a positive slope.
r = -1 means the data is perfectly linear with a negative slope.
r = 0 means there is no linear association.

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the 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? (3 marks)

The VIF [**Variance Inflation Factor**] gives how much the variance of the coefficient estimate is being inflated by collinearity. Equation for VIF is VIF =1/(1-R^2). If there is perfect correlation, then VIF = infinity. Where R-1 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 it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

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 behavior?