Assignment-based Subjective Questions

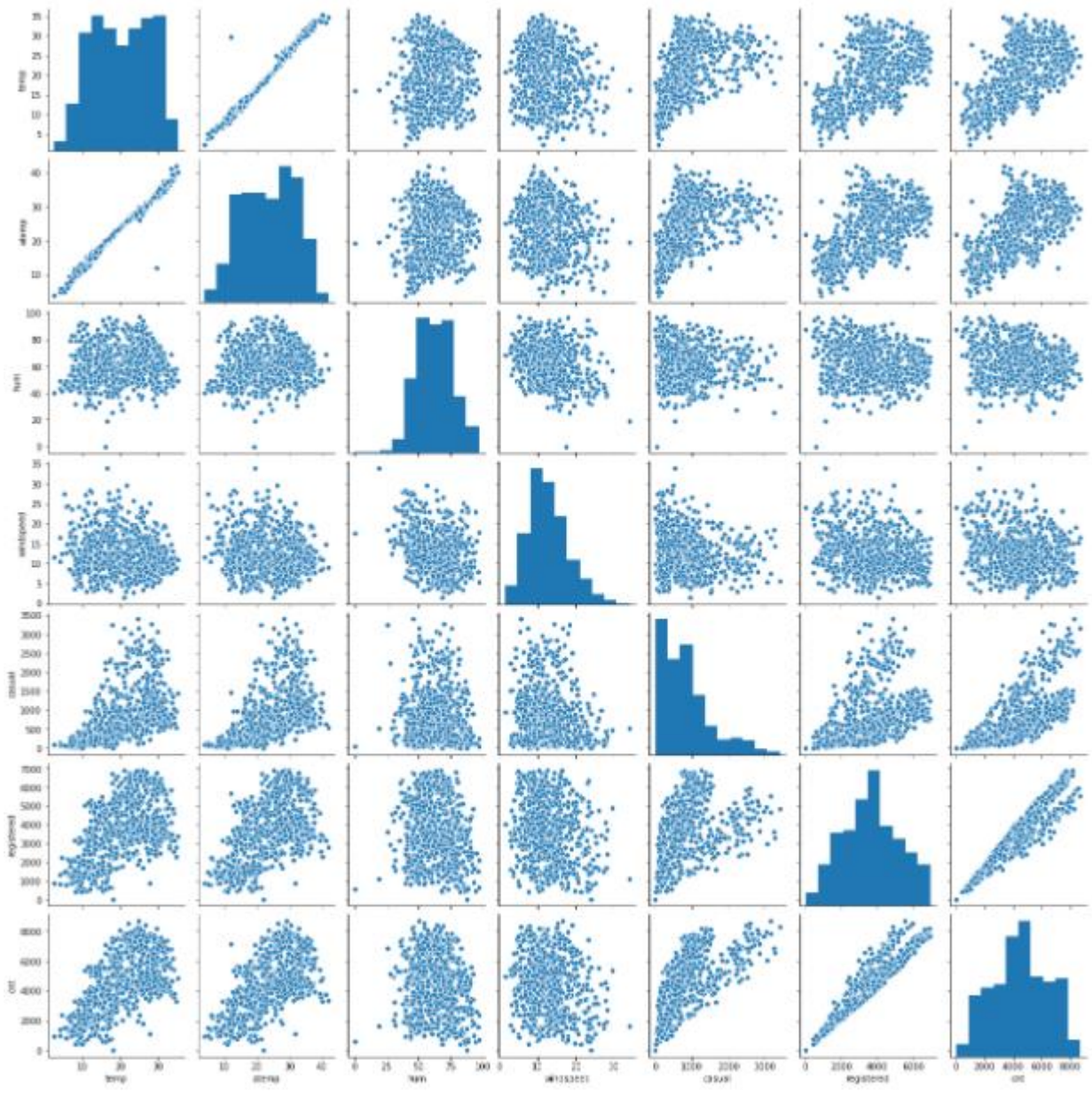
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

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:-
Season – We created boxplot and saw that that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt. This clearly tells us that season have a high effect on our business.
Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable.Highest count was seen when the weathersit was' Clear, Partly Cloudy'.Clearly weather have a huge impact on our business.
Holiday – There was a slight impact of Holiday on number of customers . On a holiday the number of customers decreases.
Mnth – June,September saw highest no of rentals while Jan saw the least number of customers.This observation is on par with the observation made in weathersit.The weather situation in Jan is usually heavy snow.
Yr - The number of customers have increased in 2019 compared to 2018.

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

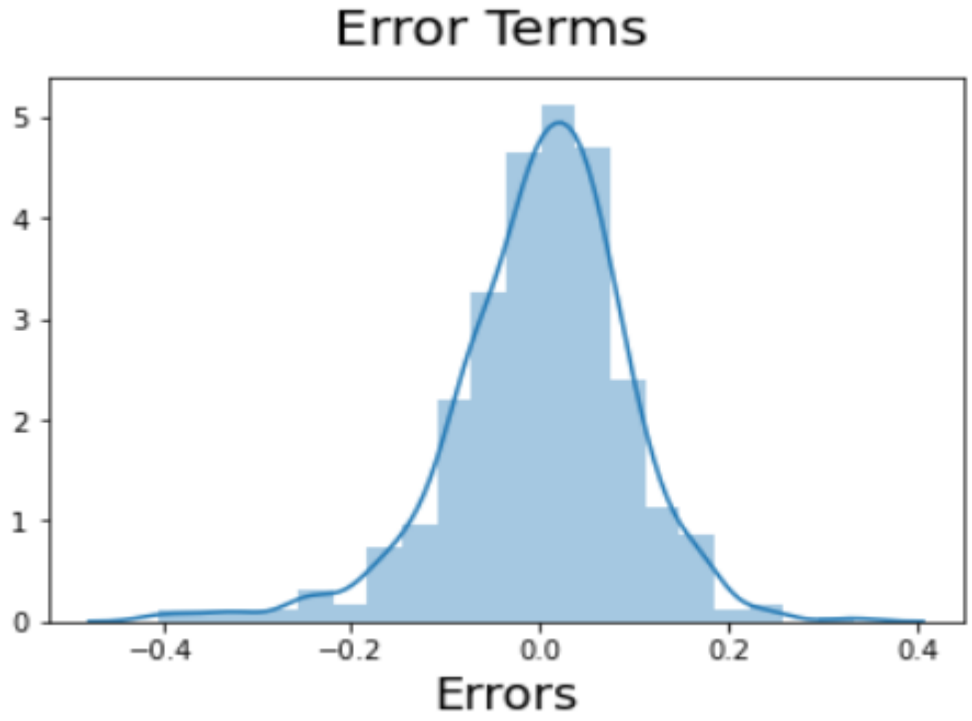
If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. Iterative models may have trouble converging and lists of variable importances may be distorted .Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables and the other variables as this can be represented using other 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)

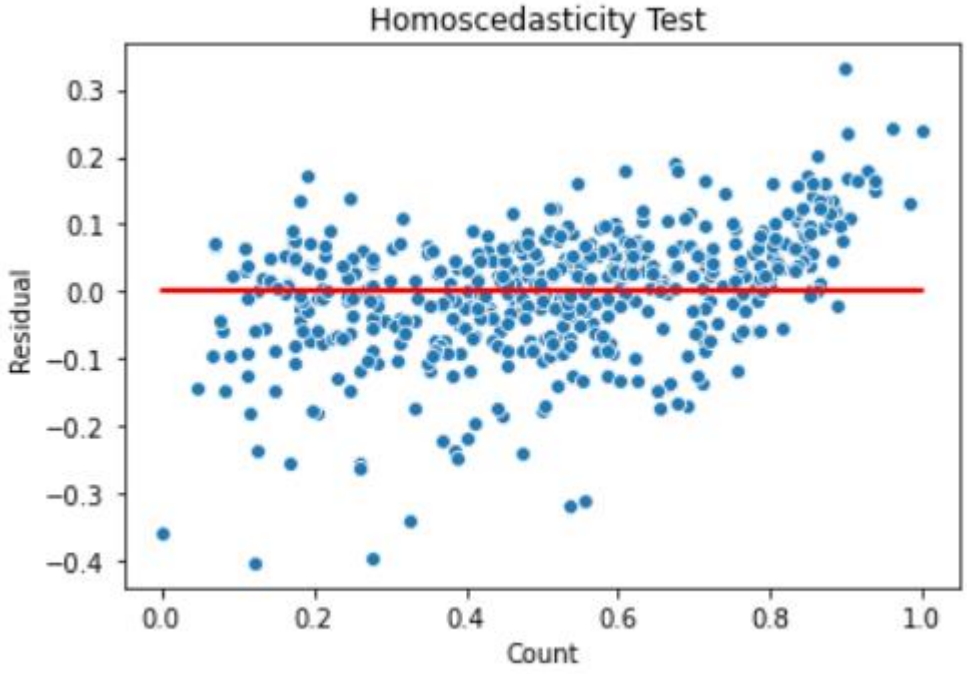


“temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt). “Registered” column also have high correlation with cnt but we won’t consider it as it also act as a target column as it indicates number of customers.

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



Residuals distribution should follow normal distribution and centred 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 above diagram shows that the residuals are distributed about mean = 0.



Residuals must have homoscedasticity . which means they should have similar variance

throughout the distribution and this can be validated by making a scatter plot of residuals and a horizontal line passing through 0. So we can observe that all the points are having almost similar variance throughout the distribution so we can say that Residuals have homoscedasticity.

Residuals must be independent. Even this assumption can be validated using the same plot and as we can see that there is no pattern in the residual we can say that only the random patterns are there between the residual so they are independent of each other. Also we calculates Durbin-Watson test of autocorrelation in which we got the value of 2.042 which is very near to 2 and signifies that there is no autocorrelation between the residuals.

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

The top 3 positive related features are :

1.temp - coeff: 0.553
2.yr - coeff: 0.233
3.season_Winter – coeff : 0.128

Top 3 negative realted features :

1.weathersit_Light rain – coeff : −0.279
2.windspeed- coeff: - 0.155
3. weathersit_Cloudy – coeff:-0.076

General Subjective Questions

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

The equation for MLR will be:

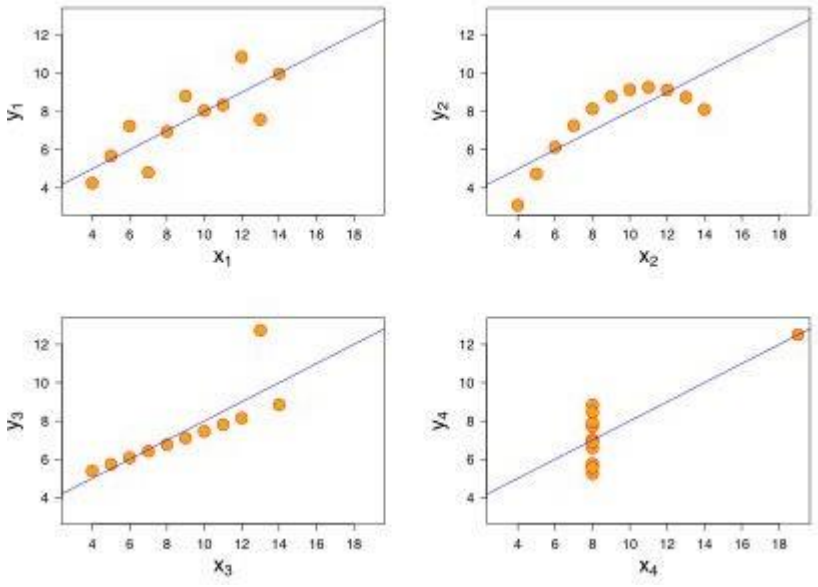
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots$$

β1 = coefficient for X1 variable β2 = coefficient for X2 variable
β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

2.Explain Anscombe's quartet in detail.

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 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 numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data?
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.

Difference between standardization and normalization:

Standardization : Mean and standard deviation is used for scaling in standardization. It is used when we want to ensure zero mean and unit standard deviation. It is not bounded to a certain range. It is much less affected by outliers, and given by the formula : X_new = (X - mean)/Std.

Normalization : Minimum and maximum value of features are used for scaling. It is used when features are of different scales. It is really affected by outliers, and is given by the formula : X_new = (X - X_min)/(X_max - X_min).

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

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). 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)

Q-Q plot stands for quantile quantile plot.It is used to tell if distribution of our data is a normal distribution or not. We will first arrange our data in ascending order and then divide it in quantiles then we will take a random dataset which we know is normally distributes and then divide this also in quntiles after that we will plot the theoretical data on x axis and smaple data on x axis and plot the values. We will have a line with slope 45 degrees if the plot follows the line we can conclude that our data is normally distributes else its not normally distributed.