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 variables in the dataset are season, weathersit, holiday, mnth, yr and weekday. They are visualized using a boxplot. The variables have the following effect on the dependent variable (cnt is the dependent variable) :-

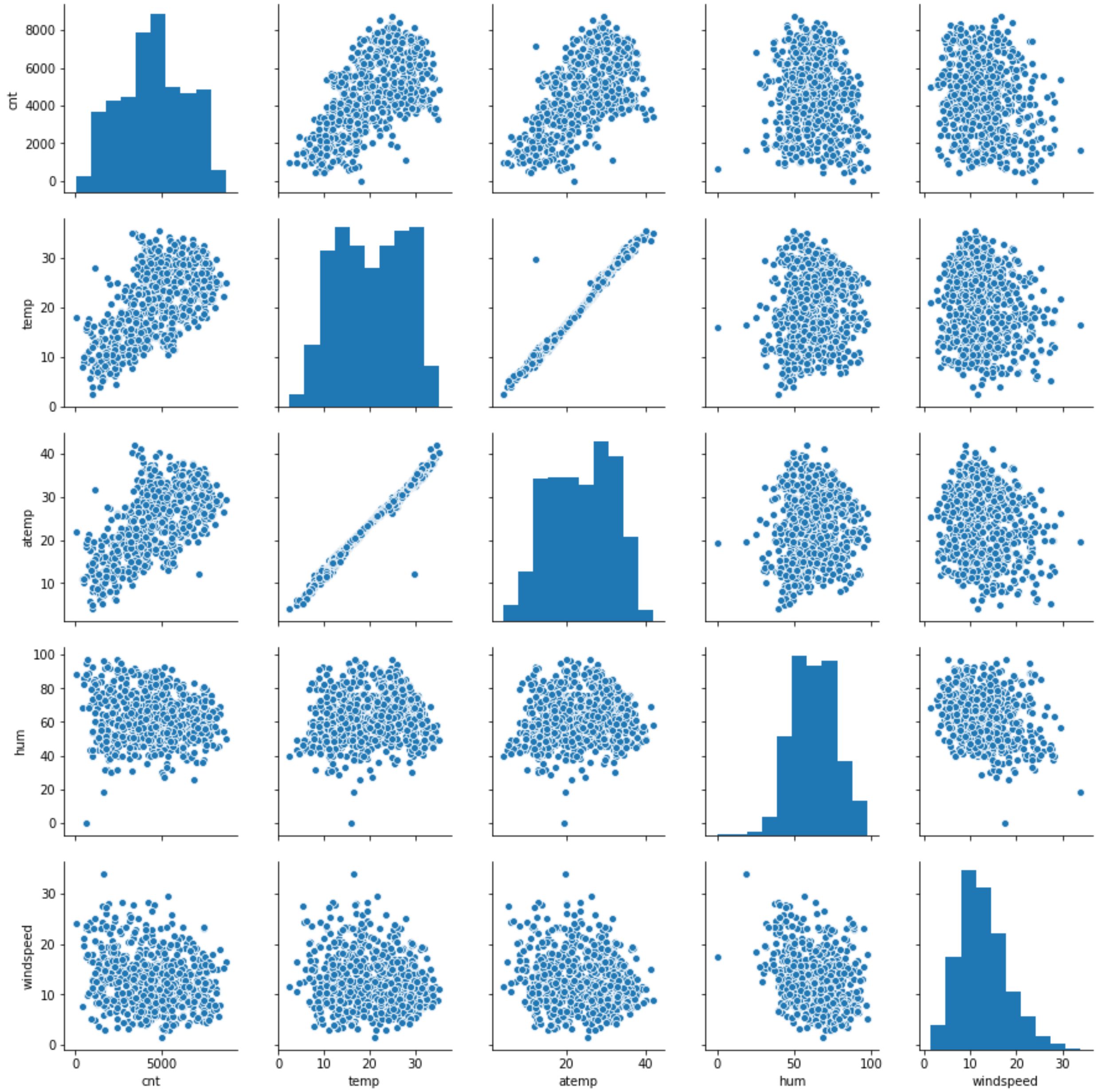
- **Season -** The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- **Weathersit(weather situation)** - There are no users when there is heavy rain and ice pallets indicating that the weather is extremely unfavorable. Highest cnt was seen when the (weather situation)weathersit was' Clear, Partly Cloudy'.
- **Holiday – (**cnt**)**rentals reduced during holiday.
- **Mnth -** September saw highest no of rentals(cnt) while December saw the least.
- **Yr -** The (cnt)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).

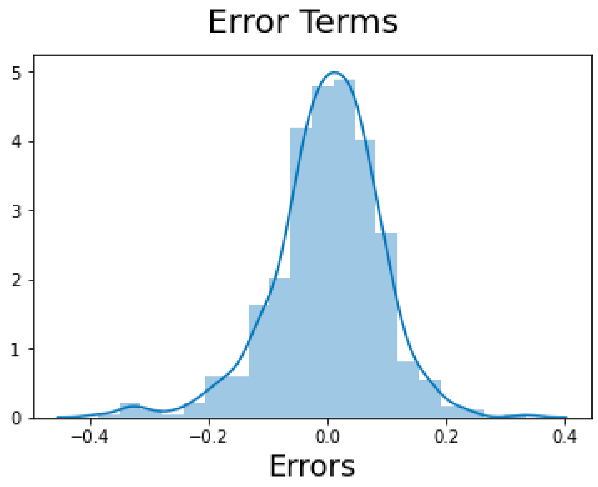
Setting drop_first= True, helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Multicollinearity concern will arise if we don't drop the first column.

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

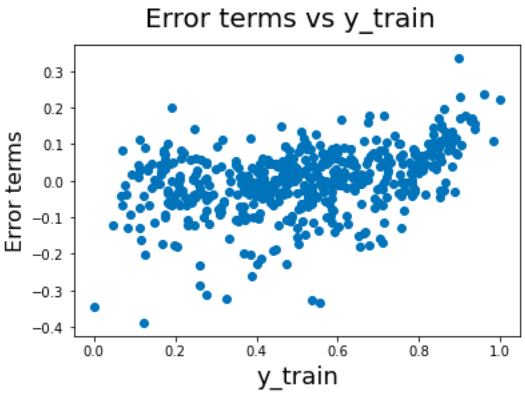


“temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt)

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 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 above diagram shows that the residuals are distributed about mean = 0.



The above scatter plot proves the independent error terms assumption. i.e the error terms are not dependent on each other.

The above scatter plot also shows that the error terms have constant variance(homoscedasticity).

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 features are

- temp - coefficient : 0.570606
- yr - coefficient : 0.2289
- weathersit_Light Snow & Rain - coefficient -0.236675

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 commonly used predictive analysis model.

It follows the equation of the straight line **“y = mx + c”.**
as ,

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we find the best fit line which describes the relationship between the independent and dependent variable with least error.

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. 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 i**s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

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

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on…

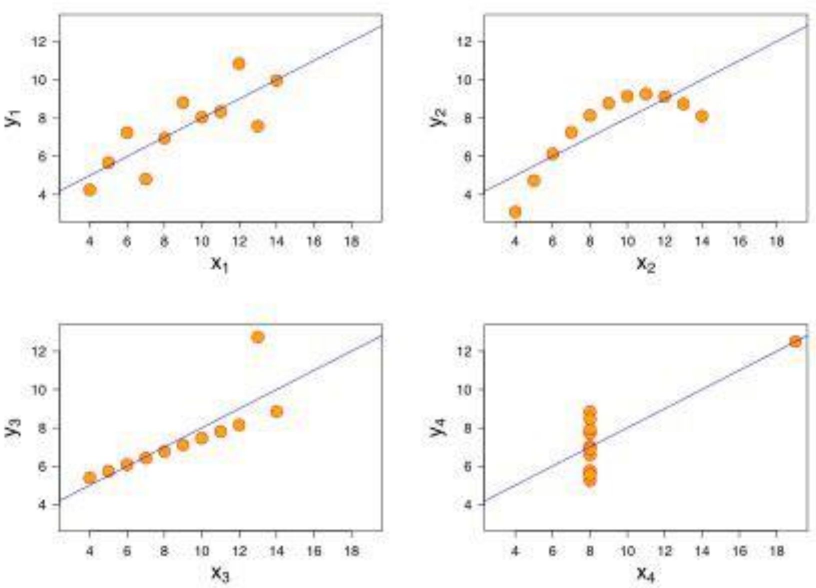
β0 is the intercept (constant term).

Assumptions of Linear Regression are as follows:

- Linear relationship between independent and dependent variable.
- Error terms are normally distributed
- Error terms have constant variance and are independent of each other.

2. Explain Anscombe's quartet in detail.

Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four datasets 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 is 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. Its value ranges between -1 to +1.

- r = 1 means the data is perfectly linear with a positive slope. It means both variables tend to change in the same direction
- r = -1 means the data is perfectly linear with a negative slope. This means both variables tend to change in different directions
- 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.

Feature scaling is essential for machine learning algorithms that calculate distances between data.. Hence, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Normalization means that the range of values are "normalized to be from 0.0 to 1.0. Normalization is 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.

Here's the formula for normalization: X' = (X-Xmin)/ (Xmax- Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

"Standardization" typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean .Standardization is where the values are centered 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.

Here's the formula for standardization: X' = (X- Mu) / Sigma

Feature scaling: Mu is the mean of the feature values and Feature scaling: Sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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^2). If there is perfect correlation, then VIF = infinity, where R^2 is the R-square value .We want to check how well this independent variable is explained well by the 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 Plots (Quantile-Quantile plots) are plots of two quantiles against each other. It is a plot of the quantiles of the first data set against the quantiles of the second data set .The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

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. Q-Q plot answers 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?

Importance of QQ plot in Linear Regression : Q-Q plots are used to visually check that your data meets the homoscedasticity and normality assumptions of linear regression. Q-Q plots let you check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed then they should form an approximately straight line