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?

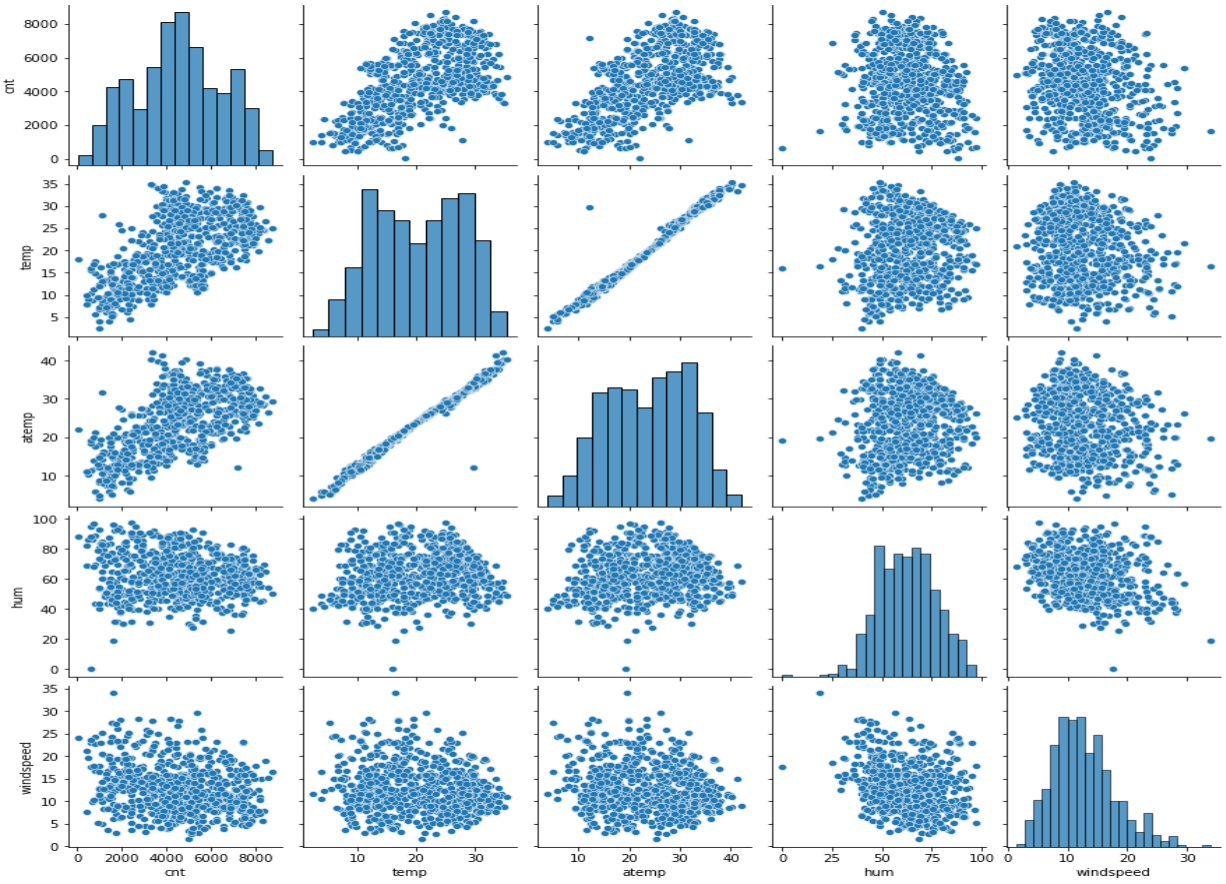
The categorical variable in the data set (year, season, weahersit, month, holiday). These were visualized using box plot

1. Year: 2018 & 2019
2. Mnth or month: September and August saw highest no of rentals while in January & December is very less in these months company can play for repair or services the bikes.
3. Holiday: More number of people are tent to stay in during the holiday and rents are reduced during this time & this time can be used for monthly maintenance of the bikes
4. Weathersit: As the company Boom Bikes is from US reign Due too cold temperature or snow or rainfall during that no users are not using the services.
5. Season: The spring and summer season are two season the more number of users used the services later followed by winter.

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

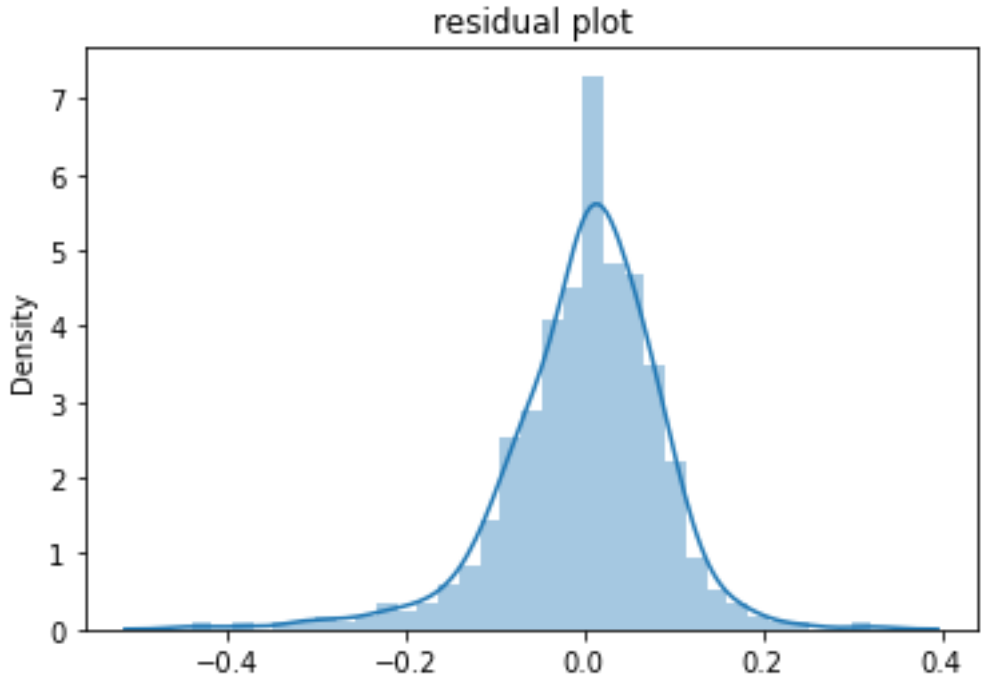
If we don't drop the 1st column then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance's may be distorted. Another 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?**



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?**



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

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

Top 3 are as follow:
Temp: 0.4524
Year: 0.2331
Spring: -0.0653

Satyanarayana
Subjective question – Assignment

General Subjective Questions

1. Explain the linear regression algorithm in detail.

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 "$y = mx + c$".

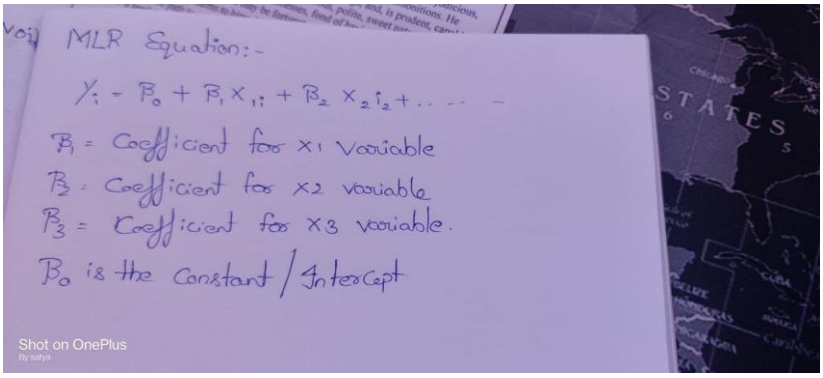
It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

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.

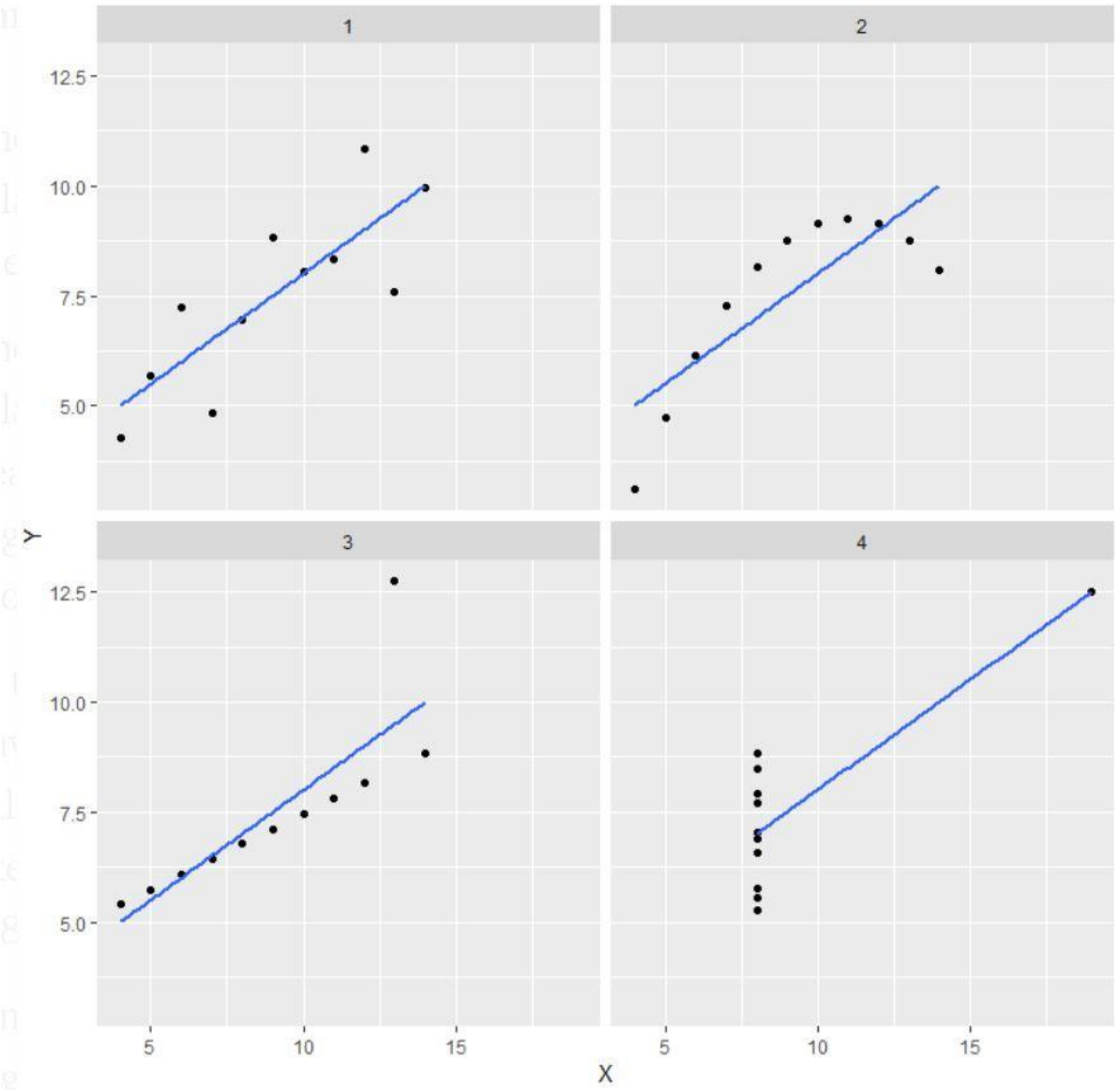
1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.



2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, but they have a very different distribution & look totally different when we plot it



Properties:

1. **Tope left:** simple linear relationship
2. **Top right:** it is a non-linear and not distributed one also
3. **Bottom left:** distribution is linear but have a different regression line and outlier are present
4. **Bottom right:** one high- level point is enough to produce a high correlation coefficient

3. What is Pearson's R?

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?

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 pre-processing 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?

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.

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