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?

Season, Weather Situation, holiday, month, working day, and weekday were the categorical variables in the dataset. A boxplot was used to visualize these. These variables influenced our dependent variable in the following ways:

1.**Season** : The boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt. Summer and winter had cnt values that were in the middle.

2.**Weather Situation** : When there is heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was 'Clear', or 'Cloudy'.

3.**Holiday** : Rentals were found to be lower during the holidays.

4.**Month :** September had the most rentals, while Jan had the fewest. This observation is comparable to the one made in weathersit. The weather in Jan is typically cold and snowy.

5.**Weekday :** Weekends saw a significant increase in book hiring compared to weekdays.

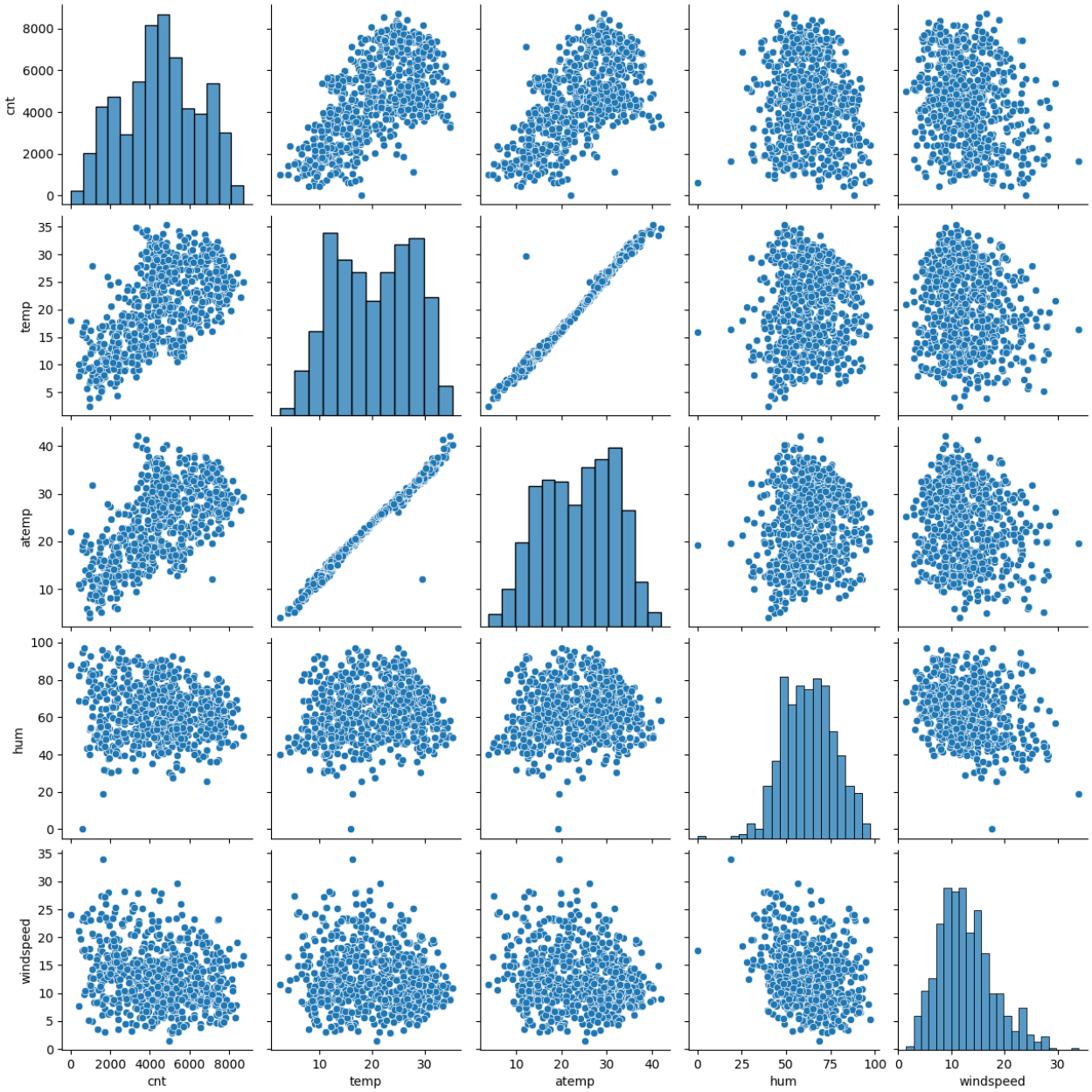
6.**Working day :** It had little effect on the dependent variable.

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

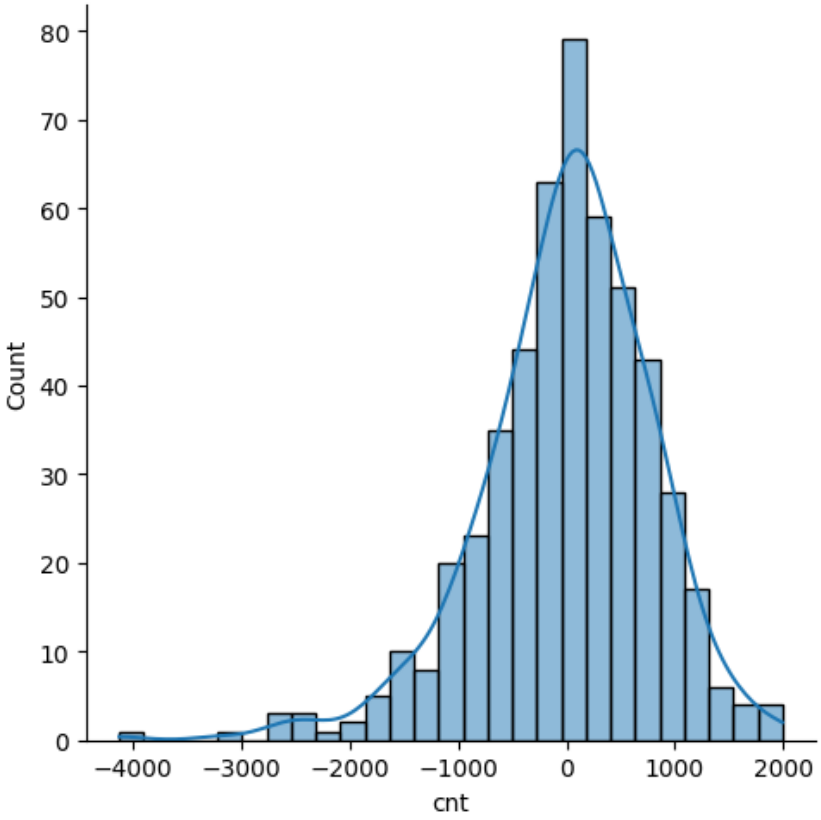
Your dummy variables will be correlated if you don't remove the first column (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted. Another argument is that having all dummy variables results in multicollinearity between them. We lose one column to keep everything under control.

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)

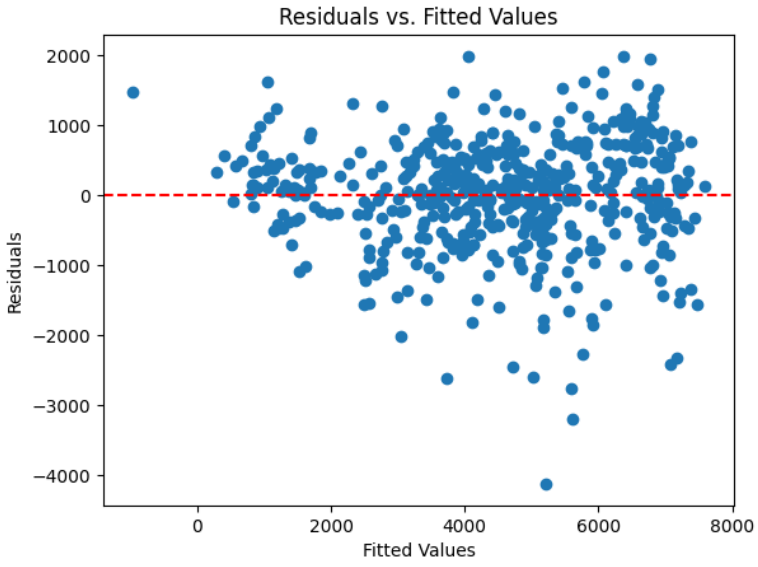


4. How did you validate the assumptions of Linear Regression after building the model on the training set?



i. The distribution of residuals should be normal and centred around 0. (The mean is 0).
The residuals are scattered around mean = 0 as seen in the diagram above.

ii. Homoscedasticity of Residuals or Equal Variances.



No clear pattern suggesting Heteroscedasticity.Thus equal variances is verified.

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

The top three predictor variables that influence bike booking, according to our final Model, are:

Temperature(temp) : With a coefficient of 3415, a unit increase in the temp variable increases the number of bike rentals by 3415 units.

Weather Situation "light snow & rain": With a coefficient of '-2242' a unit increase in this variable reduces the number of bike hires by 2242 units.

Year(yr) : With a coefficient of 2000, a unit increase in the yr variable increases the number of bike rentals by 2000. Which shows the demand might increase year over year.

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.

The equation for MLR will be:

$$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p$$

β1 = coefficient for X1 variable

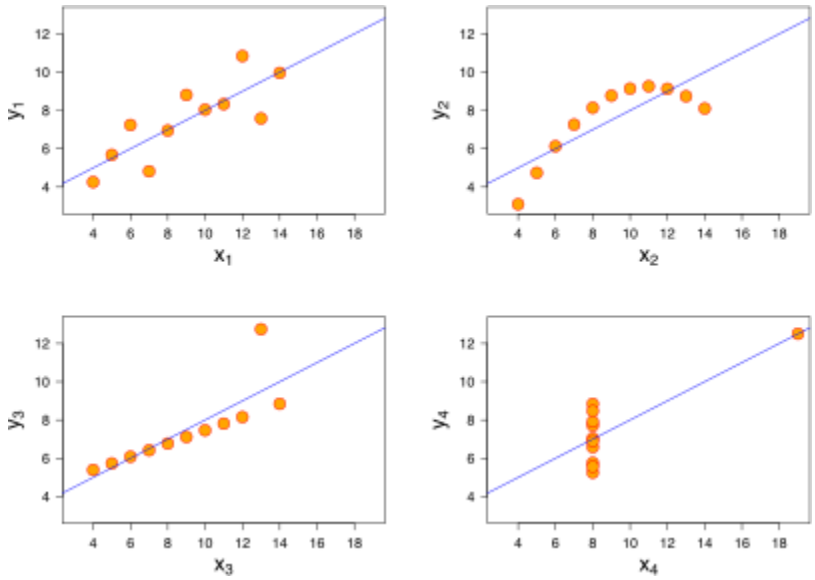
β2 = coefficient for X2 variable

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 of graphing data before analysing 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?

Pearson's r is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 to +1. It depicts the linear relationship of two sets of data. In layman's terms, it asks if we can 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 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 Neighbours 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 indicates how much collinearity has increased the variance of the coefficient estimate. (VIF) is equal to 1/(1-Ri^2). VIF = infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. - If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. As a result, VIF = 1/(1-1) provides VIF = 1/0, which is “infinity**.”**

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot.

Because both sets of quantiles came from the same distribution, the points should form a line.

That's a fairly straight line.

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?