Assignment-based Subjective Questions

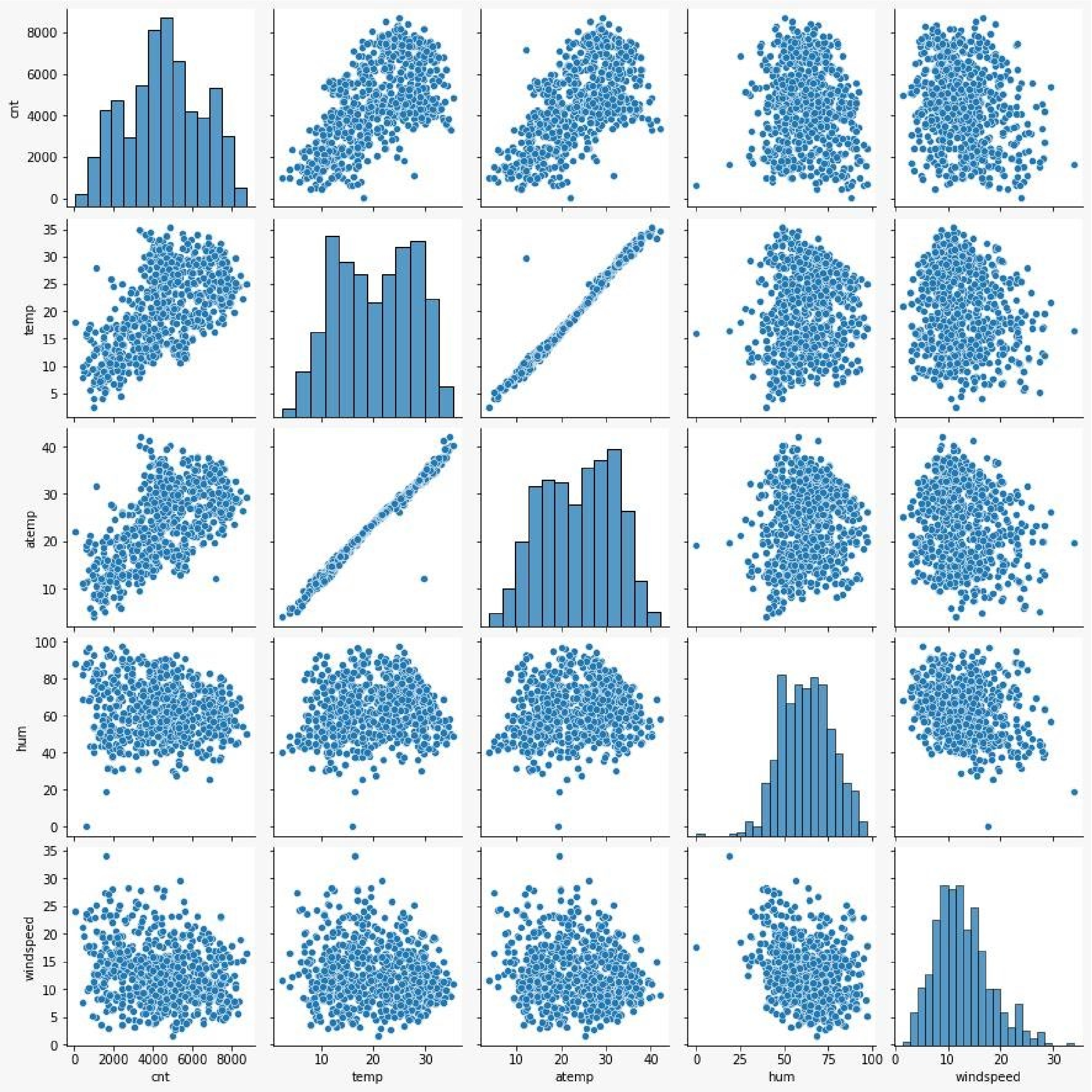
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The categorical variables in the dataset were season, weathersit, holiday, mnth, yr and weekday and were visualized using a boxplot. The effect of these on the dependent variable are:

1. **Season** – The spring season had the least value of cnt whereas fall had a maximum value of cnt as shown by the boxplot. Summer and winter had an intermediary value of cnt.
2. **Weathersit** – The heavy rain/snow has no users, thus indicating this weather to be extremely unfavourable. The highest count has been observed when the weathersit was' Clear, Partly Cloudy.
3. **Holiday -** rentals dropped during the holiday.
4. **Mnth –** Highest no of rentals were in September while the least were in December. This observation is at par with the observation made in weathersit. The weather situation in December is usually heavy snow.
5. **Yr** - The number of rentals in 2019 were higher than that in 2018

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

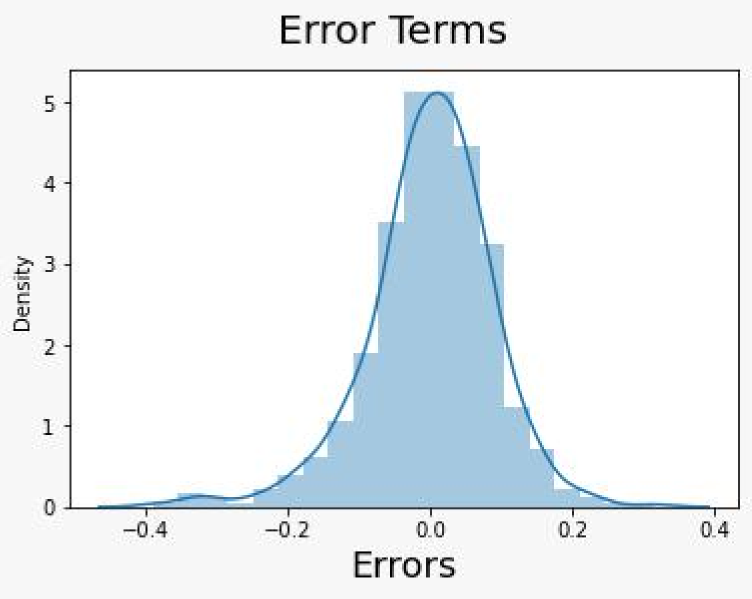
Not dropping the first column will lead to the dummy variables being correlated(redundant). Due to this, some models may be adversely affected and the effect would be stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance may be distorted. Moreover, having all dummy values will lead to Multicollinearity between the dummy variables. To keep this in check, 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)

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



Residual distribution must follow a normal distribution and should be centred around 0.(mean = 0).

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?

The top 3 features are

1.temp - coefficient : 0.594540

2.weathersit_Light Snow & Rain - coefficient -0.241359

3.yr - coefficient : 0.228354

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 and the most basic form of regression analysis. Regression is the most frequently 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, the best fit line is calculated, which describes the relationship between the independent and dependent variables.

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 regression method attempts to find the best fit line that demonstrates the relationship between the dependent variable and predictors with a least error.

The output/dependent variable in the regression is the function of an independent variable and the coefficient and the error term.

Regression is broadly classified 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

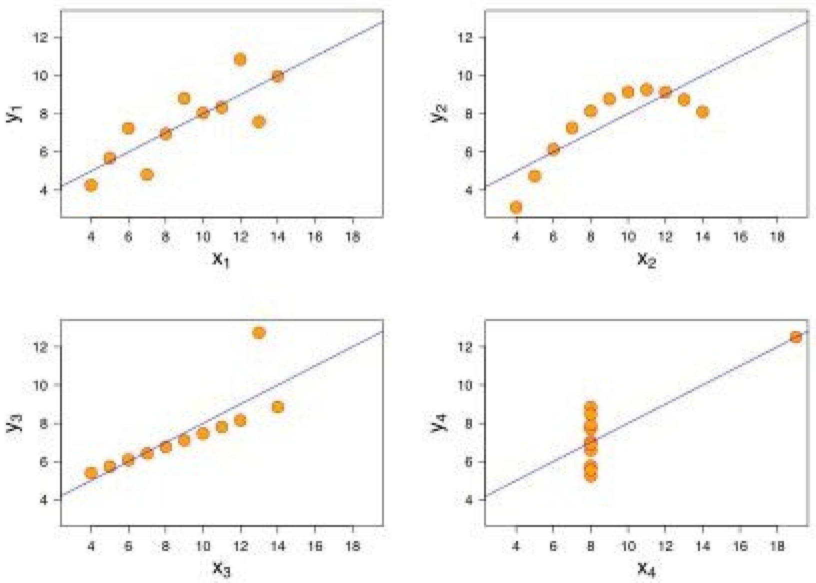
β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Statistician Francis Anscombe developed Anscombe’s Quartet. It comprises of four data sets that have almost identical statistical features, however, they have a very different distribution and when plotted on a graph, look totally different.. 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 on the top left, appears to be a simple linear relationship.
- The second graph on the top right is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph on the 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 on the bottom right shows an instance 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)

A numerical summary of the strength of the linear association between the variables is called Pearson'R. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. To put it in simple words, it tells us if we can draw a line graph to represent the data.
r = 1 indicates that the data is perfectly linear with a positive
slope r = -1 indicates that the data is perfectly linear with a
negative slope r = 0 indicates that 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)

Scaling is an approach used to normalize/standardize the range of independent variables or features of the data. It is executed during the data preprocessing stage to handle the varying values in the dataset. The machine learning algorithm tends to weigh greater values, higher and contemplate smaller values as the lower values, disregarding the units of the values, if the feature scaling is not done.

- Normalization is generally used it is known that the distribution of the data does not follow a Gaussian distribution. This can be insightful in the algorithms that does not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

- On the other hand, Standardization can be of great assistance in cases where the data follows a Gaussian distribution. However, it need not necessarily be true. Further, unlike normalization, standardization does not have a bounding range. So, even if there are outliers in the data, it shall not be affected by standardization.

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 the amount of variance of the coefficient estimate that is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is a perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we would like to check as to how well this independent variable is explained by other independent variables- It will be a perfect correlation, if that independent variable can be explained perfectly by other independent variables, 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 plot of the quantiles of the first data set against the quantiles of the second data set is called a q-q plot. They are 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 comes from the same distribution, the points forming a line that’s roughly straight shall be seen.

The q-q plot is used to answer the following questions:

- If the two data sets came from populations with a common distribution?
- Whether two data sets have a common location and scale?
- Does the two data sets have similar distributional shapes?
- Whether the data sets have similar tail behavior?