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)

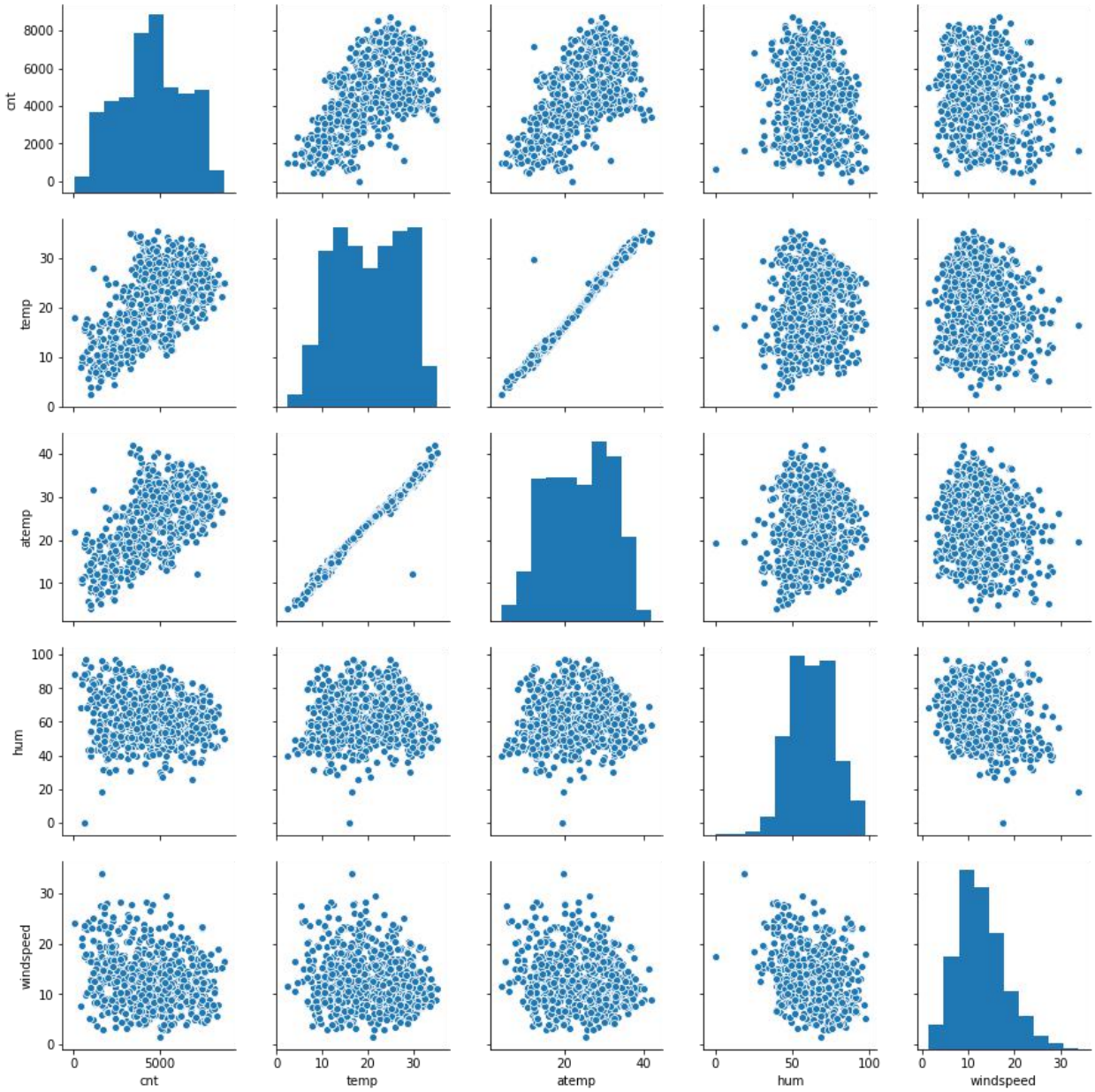
The categorical variable in the dataset are season,weathersit,holiday,mnth,yr and weekday. Boxplot is used to visualize this. These variables has the following effect on the dependant variable:-

1. **Weathersit** - When there is heavy rain/ snow, there are no users for bike rental indicating that this is extremely unfavourable. Highest count of uses are recorded when the weathersit was' Clear, Partly Cloudy'.
2. **Holiday -** Bike rentals reduced during holiday.
3. **Season -** The boxplot revealed that the spring season had the lowest cnt value, while the fall season had the highest. Summer and winter had cnt values that were in the middle.
4. **Mnth -** The biggest number of rentals were recorded in September, while the lowest were recorded in December. This observation corresponds to the one made in weathersit. The weather in December is typically cold and snowy.
5. **Yr** - The number of rentals increased in 2019 compared to 2018

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

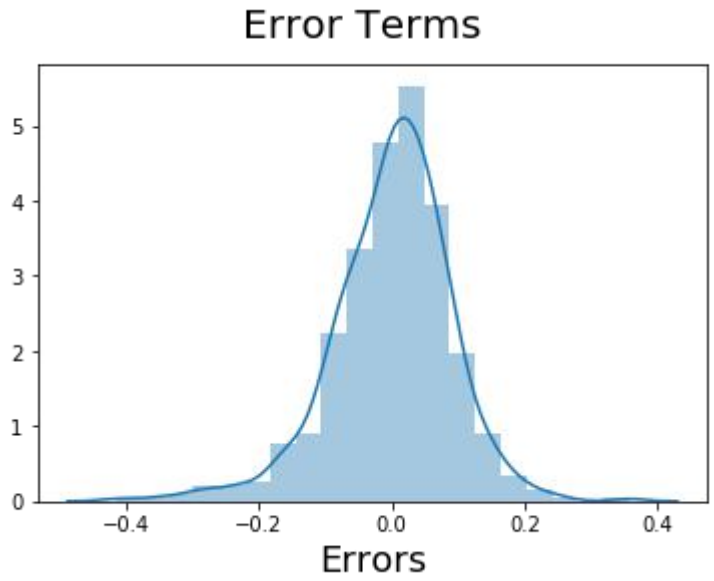
If the first column is not removed, your dummy variables will be correlated (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importances may be distorted. 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? (1 mark)



The numerical variables "temp" and "atemp" are closely associated with the target variable (cnt) as observed from the pairplot.

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



The distribution of residuals should be normal and centred around 0. (The mean is 0).
We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not.
The residuals are scattered around mean = 0 as seen in the diagram above.

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 contributing significantly towards explaining the demand of the shared bikes are:

1.temp - coefficient : 0.491508

2.yr - coefficient : 0.233482

3.weathersit_Light Snow & Rain - coefficient -0.285155

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a supervised Machine Learning approach for numeric value prediction. The most fundamental type of regression analysis is linear regression. The most widely used predictive analysis model is regression.

The popular equation "y = mx + c" underpins linear regression.

It presupposes that the dependent variable(y) and the predictor(s)/independent variable have a linear relationship (x). The best fit line, which defines the relationship between the independent and dependent variables, is calculated in regression.

When the dependent variable is a continuous data type, regression is used, and the predictors or independent variables can be of any data type, such as continuous, nominal, or categorical. The regression approach aims to identify the best fit line that accurately depicts the connection between the dependent variable and the predictors.

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

Simple linear regression and multivariate linear regression are two types of regression.

1.**Simple Linear Regression (SLR):** When only one independent variable is utilised to predict the dependent variable, SLR is used.

2.**Multiple Linear Regression (MLR):** MLR is employed when multiple independent variables are used to predict the dependent variable.

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

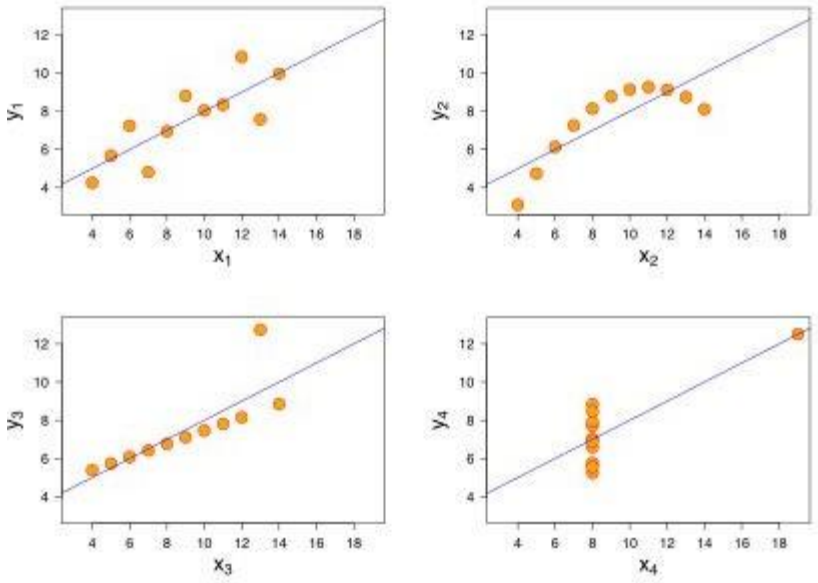
β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on.

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Francis Anscombe, a statistician, devised Anscombe's Quartet. It has four data sets with nearly equal statistical characteristics, but each has a drastically distinct distribution and appears on a graph in a completely different way. It was created to emphasise the significance of charting data before analysing it, as well as the impact of outliers and other significant observations on statistical features.



- The top left scatter plot looks to show a straightforward linear relationship.
- The second graph (top right) is not normally distributed; while a relationship exists between them, it is not linear.
- The distribution in the third graph (bottom left) is linear, but the regression line should be different. The estimated regression is thrown off by one outlier, which has a large enough impact to reduce the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) illustrates how one high-leverage point can yield a high correlation coefficient even when the other data points show no association between the variables..

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical representation of the strength of a linear relationship between two variables. Its value varies from -1 to +1. It depicts the relationship between two sets of data in a linear fashion. In simple terms, it asks if we can represent the data using a line graph.
The data is fully linear with a positive slope if r = 1. The data is fully linear with a negative slope if r = -1. There is no linear relationship if r = 0.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

The approach of feature scaling is used to normalise or standardise the range of independent variables or data features. It is used to deal with fluctuating values in the dataset during the data preprocessing stage. A machine learning algorithm will prone to fail if feature scaling is not done.

Regardless of the units of the data, consider larger values to be higher and smaller values to be lower.

- **Normalisation:** When you know that the distribution of your data does not match a Gaussian distribution, you should utilise normalisation. This is useful in algorithms like K-Nearest Neighbors and Neural Networks, which do not presume any data distribution.

- **Standardisation:** In circumstances where the data follows a Gaussian distribution, however, standardisation can be beneficial. This, however, does not have to be the case. Standardization, unlike normalisation, does not have a bounding range. As a result, even if your data contains outliers, normalisation will have no effect on them.

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

The variance inflation factor (VIF) indicates how much collinearity inflates the variance of the coefficient estimate. (VIF) is equal to 1/(1-R 12). VIF = infinity if there is perfect correlation. Where R-1 is the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. If that independent variable can be perfectly explained by other independent variables, then it has perfect correlation and its R-squared value is 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. (3 marks)

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

The following questions are answered using the q-q plot:

●Do the two data sets come from populations that have similar distributions?
●Do two data sets have common location and scale?
●Do two data sets have similar distributional shapes?
●Do two data sets have similar tail behavior?