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 following seven category variables are included in the data set:

1. weathersit (renamed to ‘weather)
2. season
3. mnth (renamed to ‘month’)
4. yr (renamed to ‘year’)
5. workingday
6. weekday
7. holiday

Based on boxplot visualisations, the following are the impacts on the target variable "cnt” as a result of these changes: -

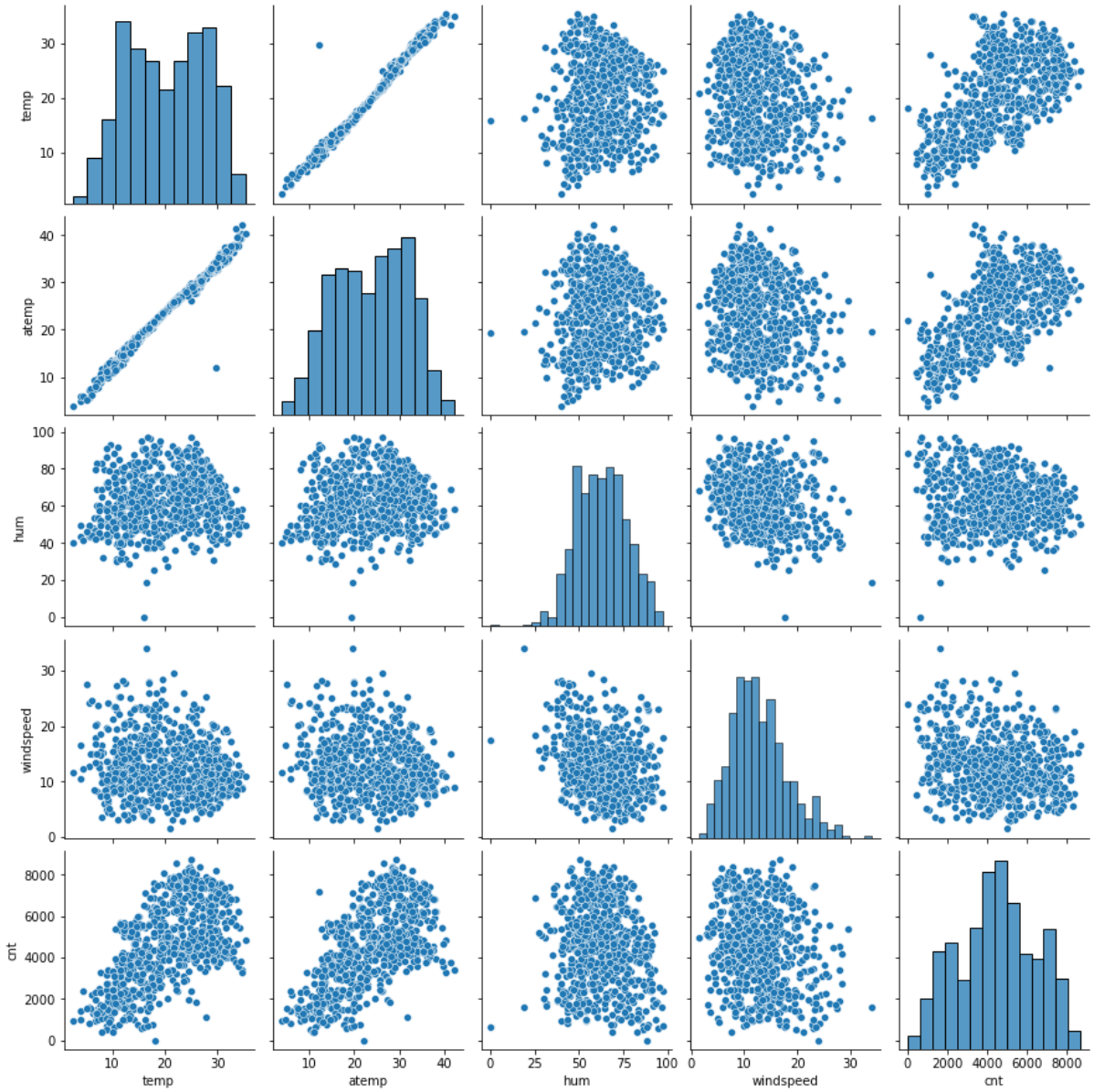
- In terms of **season**, the cnt has indeed been smallest in the spring and greatest in the fall, respectively. In the Summer and Winter seasons, cnt. values were in the middle.
- Heavy Rain & Thunderstorm was not in high demand, which indicates that the **weather** was exceedingly adverse. Clear weather has the greatest cnt.
- **Year** - In 2019, the cnt rose in comparison to '2018.'
- **Mnth** - September had the greatest cnt, while December had the lowest. Between the "Summer" and "Fall" periods, which correspond to the more positive as well as negative weather conditions in September and December, the number of months grows.
- In terms of **weekdays**, the median cnt doesn't really change substantially.
- When it comes to **'working day**,' the cnt is not influenced greatly.
- Bike rentals were lower during the **'holidays**,' as a result.

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

Dropping the first dummy column produced for a category variable is easy using the drop first=True option. The dummy parameters will be associated if we don't remove the first column (redundant). A lesser cardinality has a greater impact; however this might have an impact on certain models. For instance, iterative models might well have difficulties convergent, and items of variable importance may be skewed, etc. Having all dummy variables causes to multicollinearity, which is another reason to avoid them entirely. One column will be deleted to keep things in check.

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

The highest correlation with target variable ‘cnt’ is temp and atemp. The pair plot is given below.



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

The assumptions of simple LR are –

-The errors are spread in a regular manner.

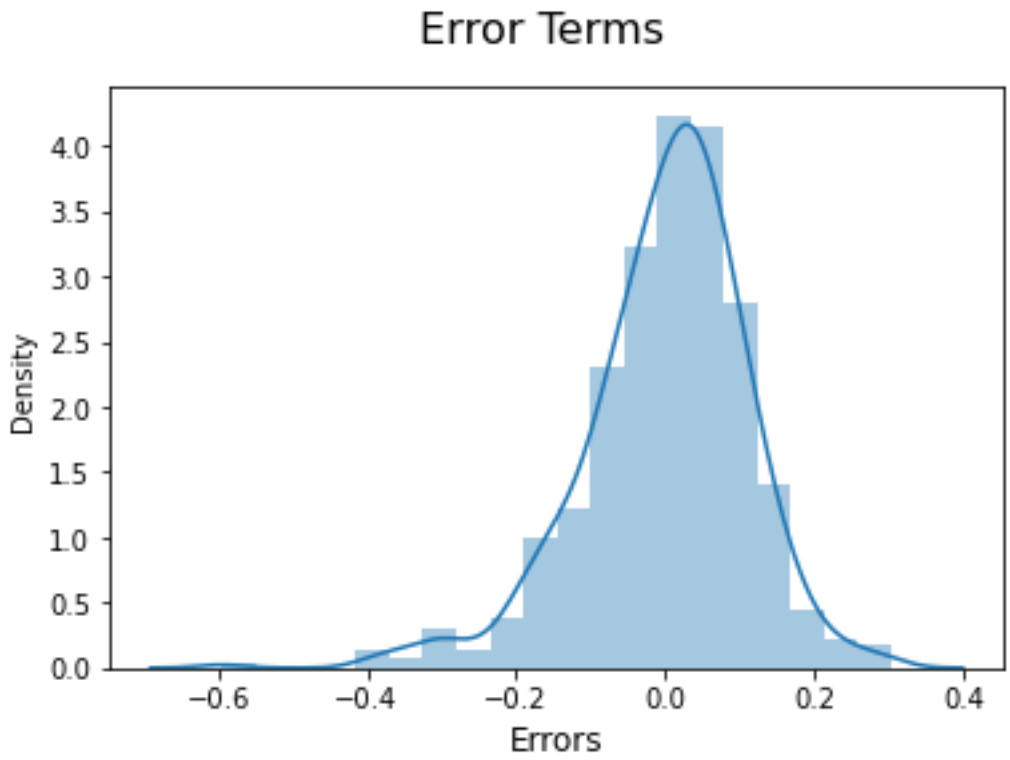
-The average of the residuals comes out to be zero.

- The error terms are independent.

Assumptions for multiple LR –

- First, the model is now fitted to a "hyperplane," rather than just one line.
- Coefficients still are generated by reducing the sum of squared errors. – (Least squares criterion).
- Assumptions made in Simple Linear Regression are still valid for inference.
- The model should never be overfitted and therefore should be devoid of multicollinearity.

The residual analysis plot is given below:



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

The variables that have positive impact on bike rental are

1. Yr – coef – 0.24
2. Temp – coef – 0.46
3. Winter – coef – 0.088

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A method called Linear Regression uses supervised learning. It's a regression job, in other words. Given a set of independent factors, regression calculates a predicted value for a target variable. It is primarily employed for establishing the link between different factors and the process of forecasting. The kind of link they examine between dependent and independent variables and the quantity of independent variables they utilise are two key differences among regression models.

An independent variable (x) is used to predict the value of a dependent variable (y) in a linear regression model (x). As a result, a linear connection exists between x (the input) and y (the output) (output). As a result, Linear Regression is the term given to this technique.

X (job experience) and Y (salary) are shown in the picture above, with X being the input and Y the outcome. Our model's regression line has the greatest fit.

Linear regression hypothesis function:

$$y = \theta_1 + \theta_2.x$$

During the process of training the model, we are provided with:

x: x is the input training data (parameter)

y: the assignment of data labels (supervised learning)

For each possible value of x, the model will try to find the line that best predicts the value of y. Using the best 1 and 2 values, the model obtains the best regression fit line.

θ1 and θ2 values.

θ1: intercept

θ2: coefficient of x

The best fit line is obtained after determining the optimal 1 and 2 values. Using our model to make predictions, we can find out the value of y based on the value of the input variable, x.

Simple linear regression and multiple linear regression are the two most common types of regression.

- When just one independent variable is used to predict the dependent variable, SLR is the most often employed method.
- Multiple Linear Regression (MLR) is utilised when the dependent parameter is forecasted utilizing multiple independent variables.

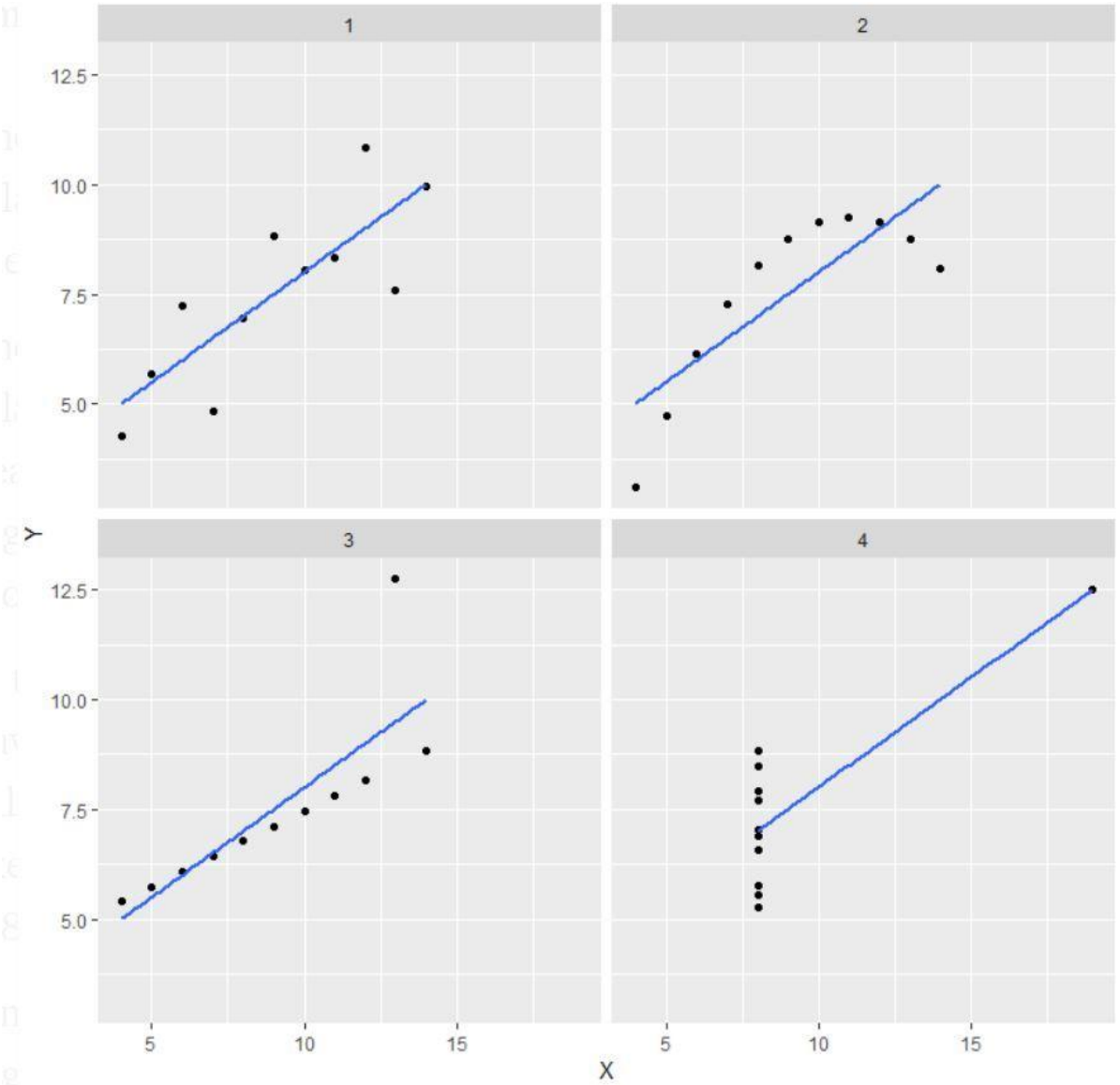
2. Explain the Anscombe’s quartet in detail. (3 marks)

If a regression model is developed using Anscombe's Quartet, it is likely to be fooled by the dataset's quirks, even if they are essentially similar in terms of descriptive statistics. Scatter plots show that they have wildly different distributions.

According to Francis Anscombe, a statistician, drawing graphs prior to analysing and developing models is critical to understanding the statistical features of a data set.

Plots of four data sets are essentially identical in terms of variance and mean for all x,y points in the datasets; hence, they convey the same statistical information.

There are a number of algorithms out there that can help you construct models out of the data, but if you don't have a good understanding of how the data is distributed and how it may help you discover anomalies such as outliers, you won't be able to develop a model that is accurate. Because it can only handle data with linear connections, the Linear Regression cannot handle any other form of datasets at all.



Anscombe's quartet consists of four datasets that have virtually equal statistical qualities but seem vastly different when plotted on a graph. This is indicated in the definition.

This output has been explained:

- There appears to be a linear connection between x and y in the first (top left) scatter plot.
- If you glance at the second graphic (top right), you may deduce that x and y are not linearly related.
- There is one outlier in the third chart (bottom left) that appears to be far off the line, indicating that there is a strong linear connection for all but one of the data points.
- One high-leverage point is all that is needed to achieve a correlation coefficient of a large magnitude in the fourth case (bottom right).

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

When two variables are linked, Pearson's r provides a numerical representation of how strong the linear relationship between them is. It may have a value ranging from -1 to +1. It depicts a straight line connecting two sets of numbers. In layman's words, it informs us if the data can be represented by a line graph or not.

- It signifies that the data is fully linear and has a positive slope if r = 1
- r = -1 denotes a negative slope for the data set.
- If r is equal to 0, then there is no linear relationship.

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

As a component of data preparation, it's used on independent variables in order to standardise their values. It also aids in the speed of algorithmic computations.

Almost all of the time, the acquired data comprises characteristics that are significantly variable in magnitude, units, and range. Incorrect modelling may occur if scaling is not performed, since the technique only considers magnitude and not units. Scaling is required to equalise the size of all variables in order to resolve this problem.

One must keep in mind that scaling has no effect on variables such as the t-statistic, F-statistic or other statistical measures such as p values.

- When the distribution of the data doesn't really match a Gaussian distribution, normalisation is often utilised. Algorithms like K-Nearest Neighbours and Neural Networks, which presume no distribution of the data, may benefit from this.

- Standardization - However, if the data has a Gaussian distribution, standardisation may be beneficial. This will not, nevertheless, have to be the case. It's also worth noting that, unlike normalisation, standardisation does not have an upper limit. Even if your data contains outliers, normalisation will not influence them.

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

If there is good correlation, then VIF = infinite. This exhibits a flawless correlation between two independent factors. In the event of perfect correlation, we have R2 =1, which result to 1/(1-R2) infinite. To overcome this issue we need to delete one of the elements from the set which is creating this perfect multicollinearity.

An infinite VIF value implies that the related variable may be stated perfectly by a linear combination of additional variables (which exhibit an infinite VIF as well) (which show an infinite VIF as well).

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

It is possible to use a graphical tool called the Quantile-Quantile (Q-Q) plot to determine whether or not a collection of data is consistent with a theoretical distribution. If two data sets originate from populations that are geographically dispersed, this method can tell you that.

Using a Q-Q plot, we can verify that both the training and test data come from populations with similar distributions in a linear regression model.

Here are a few perks of using Q-Q plot in LR:

a) It is also applicable to smaller sample sets.

b) It is possible to see variations in size, symmetry, and the existence of outliers in this distribution from this graphic.