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)

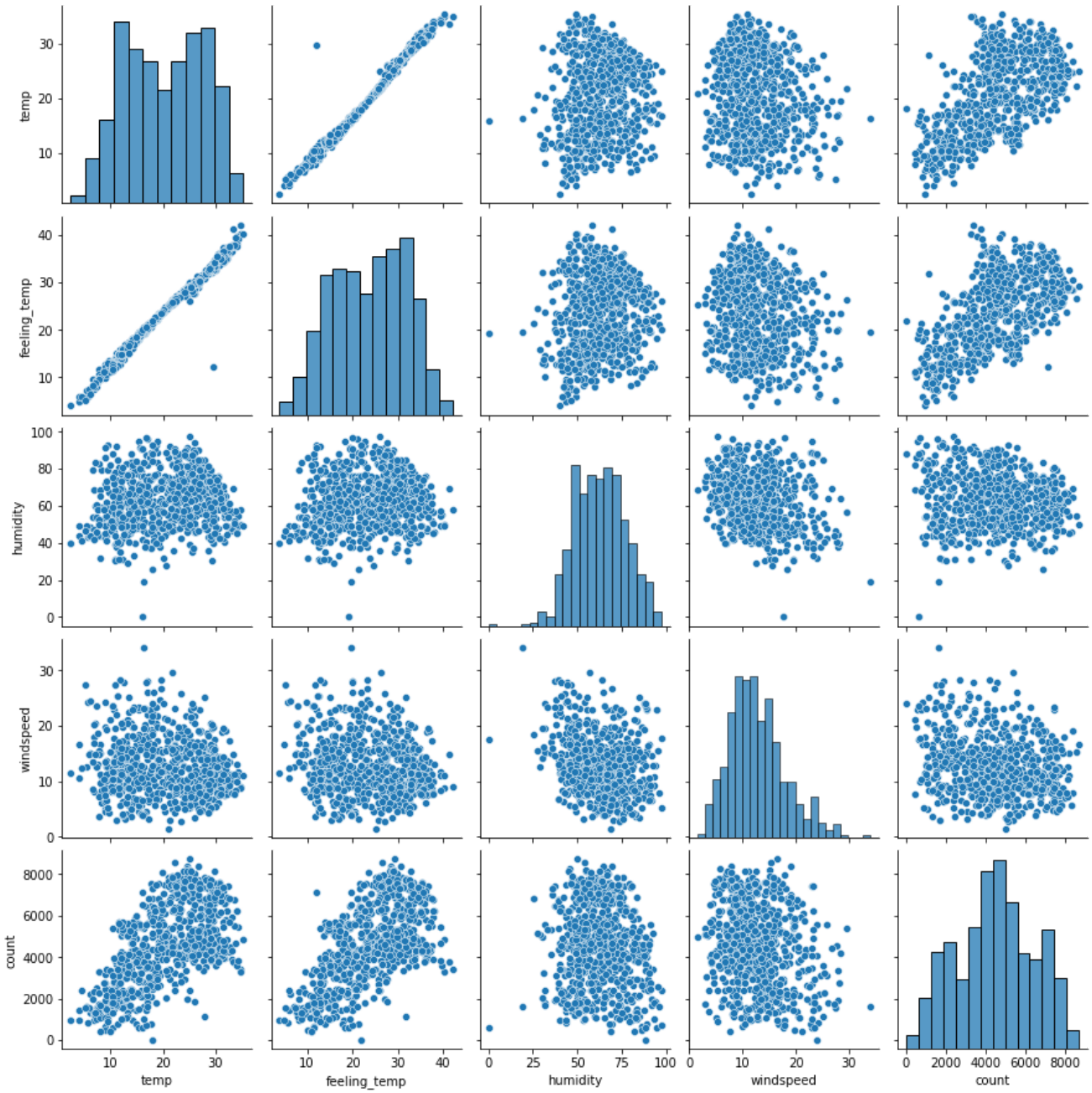
   The categorical variable in the dataset were season, weathersit, holiday, mnth,yr and weekday. These were visualized using a boxplot . These variables had the following effect on our dependant variable:-

   a. Season - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt.Summer and winter had intermediate value of cnt.
   b. Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable.Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
   c. Holiday - rentals reduced during holiday.
   d. Mnth - September saw highest no of rentals while December saw least.This observation is on par with the observation made in weathersit.The weather situation in december is usually heavy snow.
   e. Yr - The number of rentals in 2019 was more than 2018.

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

   If we don't drop the first column then your 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 may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between thedummy variables. To keep this under control, we lose one column.

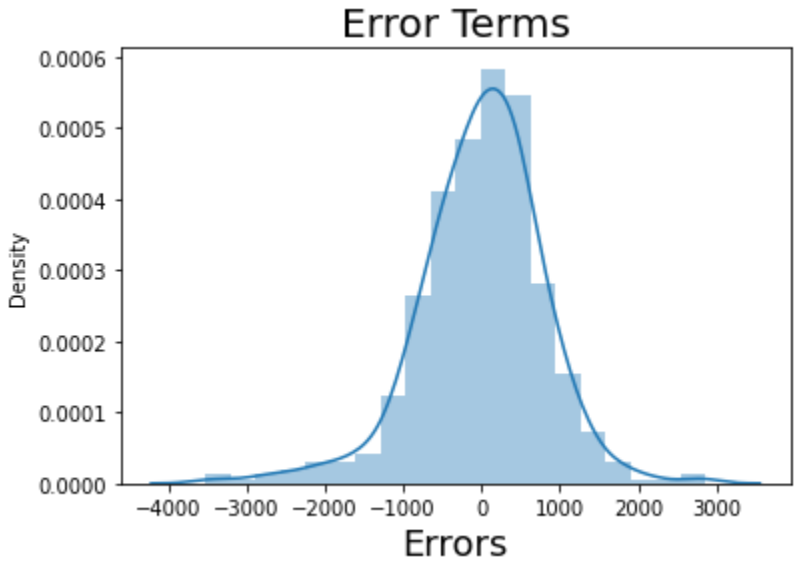
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



'Feeling_temp' which is 'atemp' as per given and 'temp' are more correlated with the target variable(cnt).

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

   We plotted the distribution of residual (Residual analysis).



   The plot was almost centered around zero, hence we can say that our linear regression model is giving the correct result.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

   1. feeling_temp 0.5069
   2. light_snow -0.2791
   3. year 0.2343

General Subjective Questions

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

   Ans: Linear regression is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable and the dependent variable,

   Different regression models differ based on – the kind of relationship between dependent and independent variables

   a. **Simple Linear Regression** : SLR is used when the dependent variable is predicted using only one independent variable.
   b. **Multiple Linear Regression** :MLR is used when the dependent variable is predicted using multiple independent variables.

   - Standard equation of the regression line is given by the equation:
     - $y = \beta0 + \beta1 X1 + \beta2 X2 + \ldots\ldots + \beta n Xn$

     $\beta1$ = coefficient for X1 variable

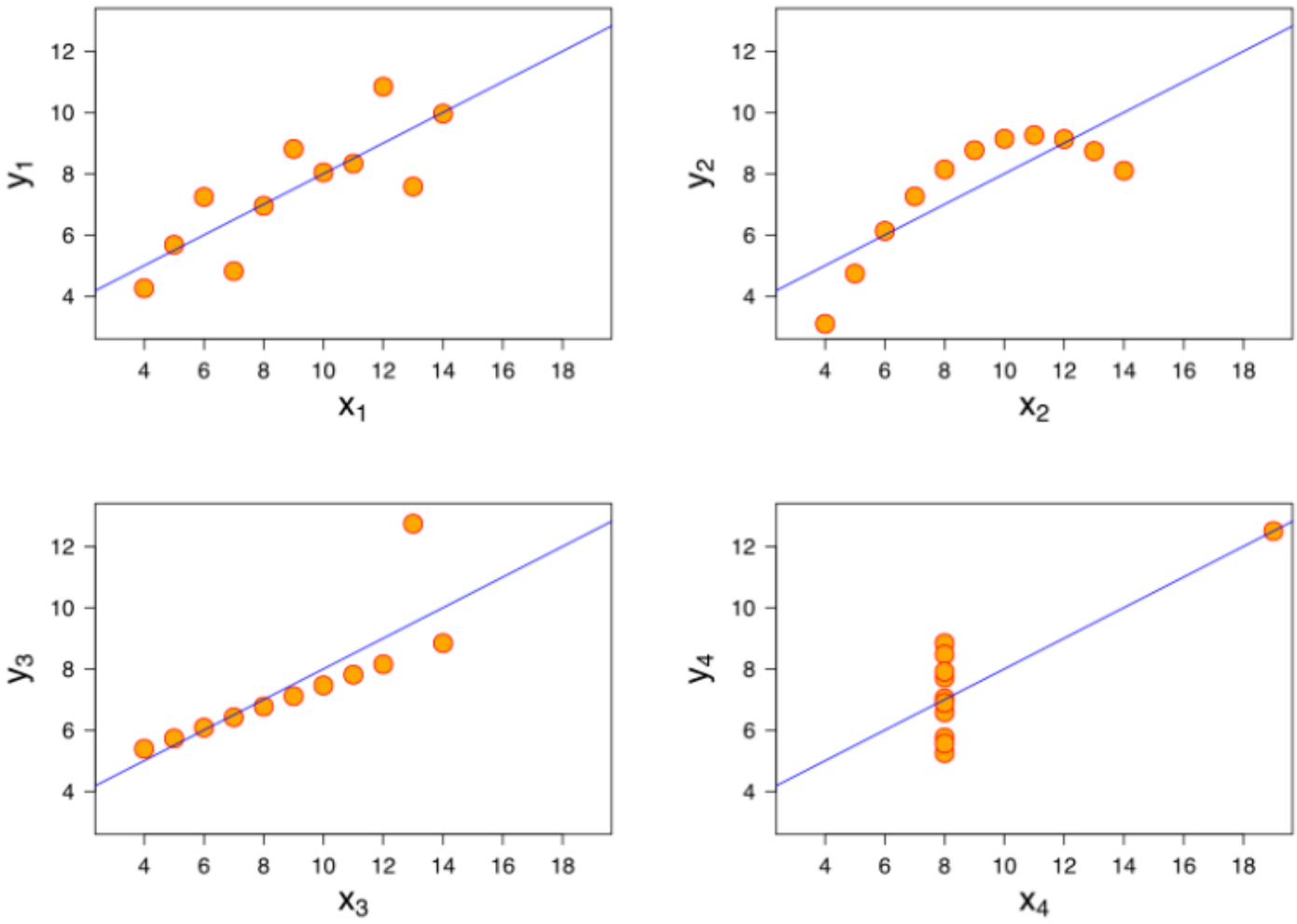
     $\beta2$ = coefficient for X2 variable

     $\beta3$ = coefficient for X3 variable and so on… $\beta0$ is the intercept (constant term).
   - Assumption:
     - X and Y has to have linear relation ship
     - Error terms are normally distributed
     - Error terms are independent of each other
     - Error terms have constant variance
   - The best-fit line is found by minimizing the Residual Sum of Squares(RSS) using Ordinary Least Squares method.
   - The strength of the linear regression model can be assessed by metrics like
     - R 2 = 1- (RSS/TSS) or Residual Standard Error (RSE).
   - Examples where linear regression can be used - predicting the housing prices, understanding relationship between advertising spending and revenue generated.

2. Explain the anscombe's quartet in detail. (3 marks)

   Ans: 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



- 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 Parson's capital R? (3 marks)

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? (3 marks)

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Scaling is done for:

- Ease of interpretation when the variables are at the same scale
- Faster convergence for Gradient descent methods.

Scaling just affects the coefficients and doesn't impact the parameters like p-value, R 2 , T-statistic or F-statistic.

Normalized scaling: Normalization technique brings all the data between the range 0 and 1. It is also known as Min-Max scaling. Given by the equation

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling

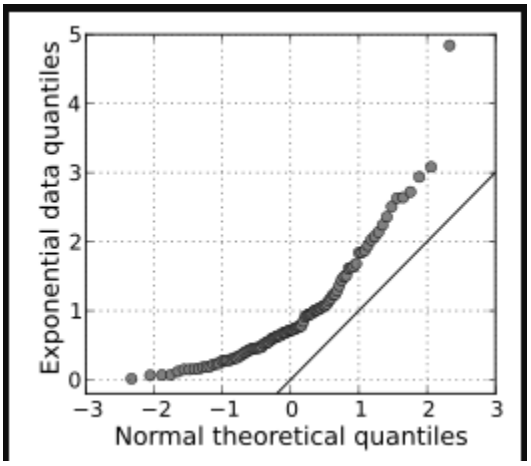
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans: If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, which leads to 1/(1-R2) becoming infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (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. (3 marks)
   Ans: Q-Q plot is a scatterplot created when two sets of quantiles are plotted against each other. If two datasets come from a population with the same distribution, the data points would fall approximately along the 45-degree reference line which is plotted.
   In general, it helps in answering the following:
   - If two datasets come from populations with a common distribution
   - If two datasets have a similar distribution shape
   - If two datasets have similar tail behavior
   - If two datasets have common location and scale

   Whenever there are two data sets, it is useful to know if the assumption of the common distribution is justified. The Q-Q plot can provide more insights into the nature of the difference between data sets.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.