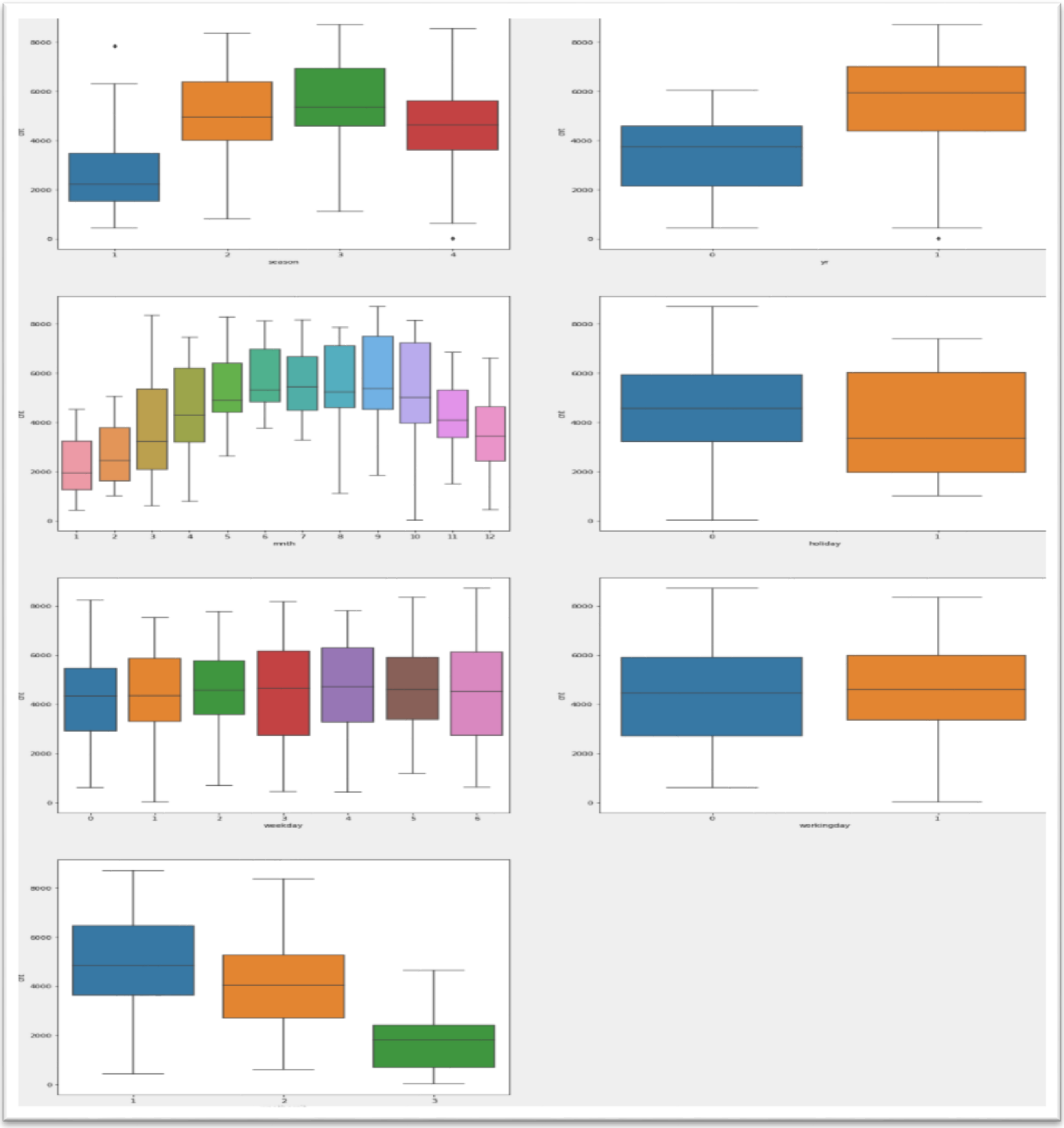
LINEAR_REGRESSION_ASSIGNMENT_SUBJECTIVE_QUESTIONS

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

My Observations after the analysis of categorical variables in the dataset provided are:

- Season 2,3 & 4 i.e. summer fall and winter shows higher number of riders as compared to season 1:spring
- 2019 surely have high numbers of riders which implies number of riders were increasing.
- There is sort of bell curve in the number of riders for a year starting from January and to December.
- Not such of difference between holiday and working day and weekdays.
- But surprisingly on holiday average is lower however compared to working and non-working day it is closely same

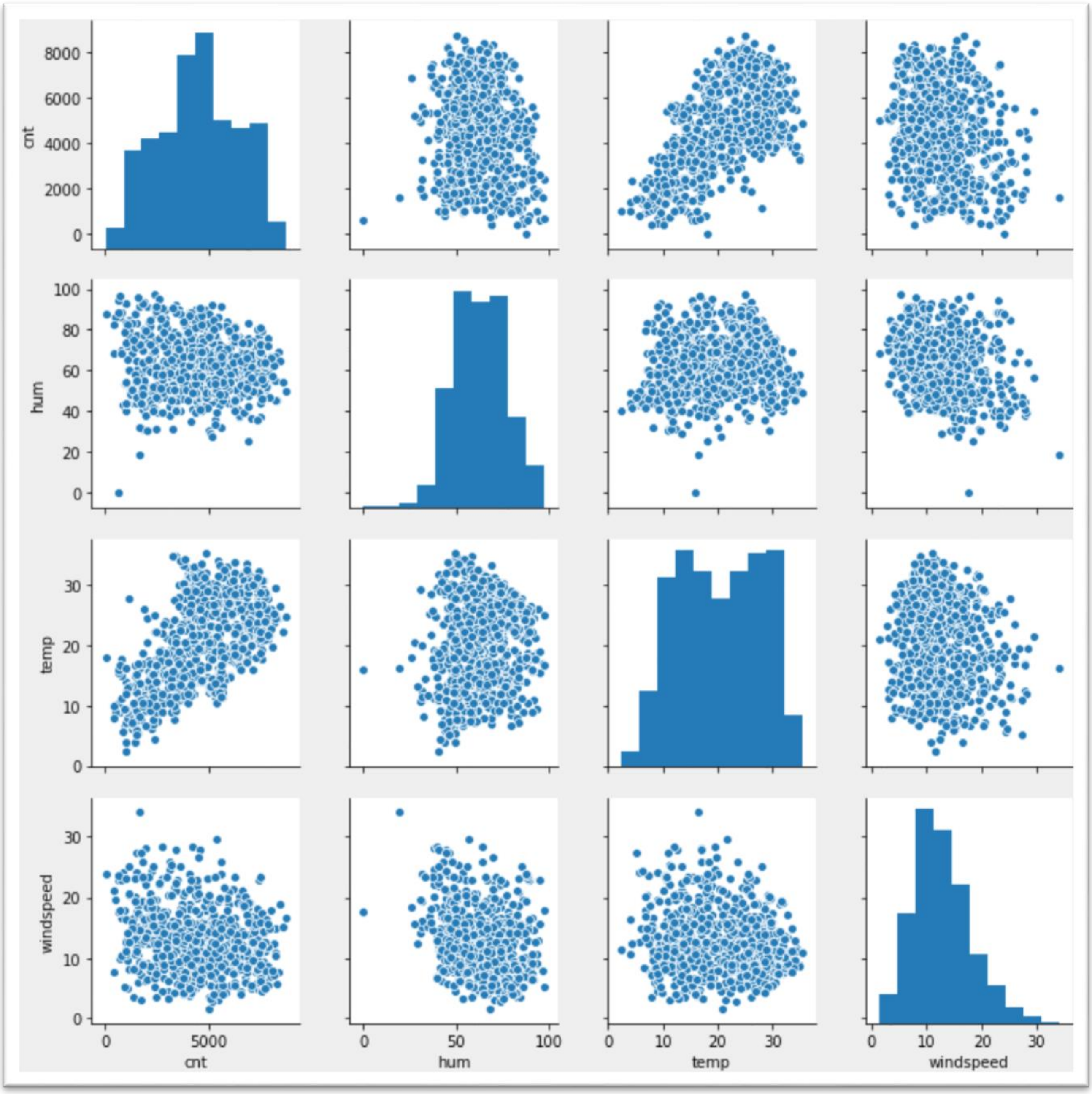


2. **Why is it important to use drop_first=True during dummy variable creation?**

   In case we do not drop_first=True during dummy variable creation, it will lead to high collinearity between dummy variables by creating an extra 1 variable which is not of such significance.
   I made this mistake in the assignment and it resulted to lead VIF for 3 of my dummy variables of Season to be inf, i.e. infinity.
   In addition to that due to high co-relations between the dummy variables it lead to the result where Residual, i.e. errors are not normally distributed and its mean is also not 0.

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

   It is clearly evident that temperature has the highest co-relation with our target variable.

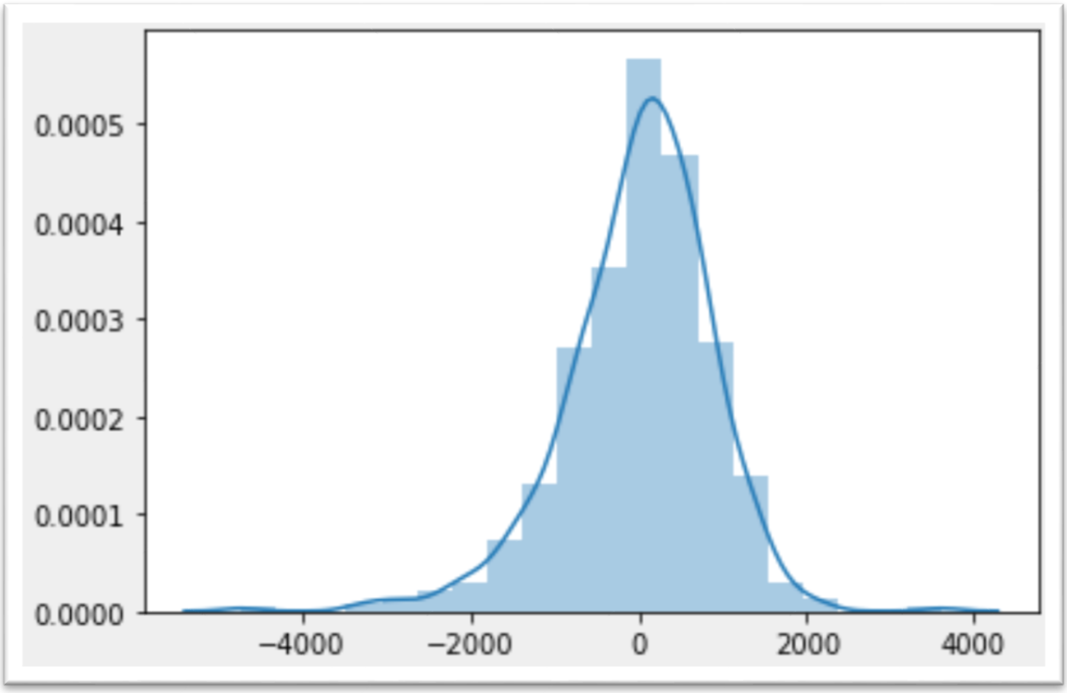


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

We validate the most important assumption of Linear Regression which is about residuals.

Normality assumption: Normality assumption describe the nature of distribution of the residuals/errors. It should be normally distributed.
Zero mean assumption: Residuals should have mean value about 0.

After building the Linear Regression model on training set in the assignment I performed above two validation check and below is the result for that.



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

Heavy Storm days with a co-efficient of 2770.76
Year with a co-efficient of 2103.10
Spring season with a co-efficient of 2049.78

6. **Explain the linear regression algorithm in detail.**

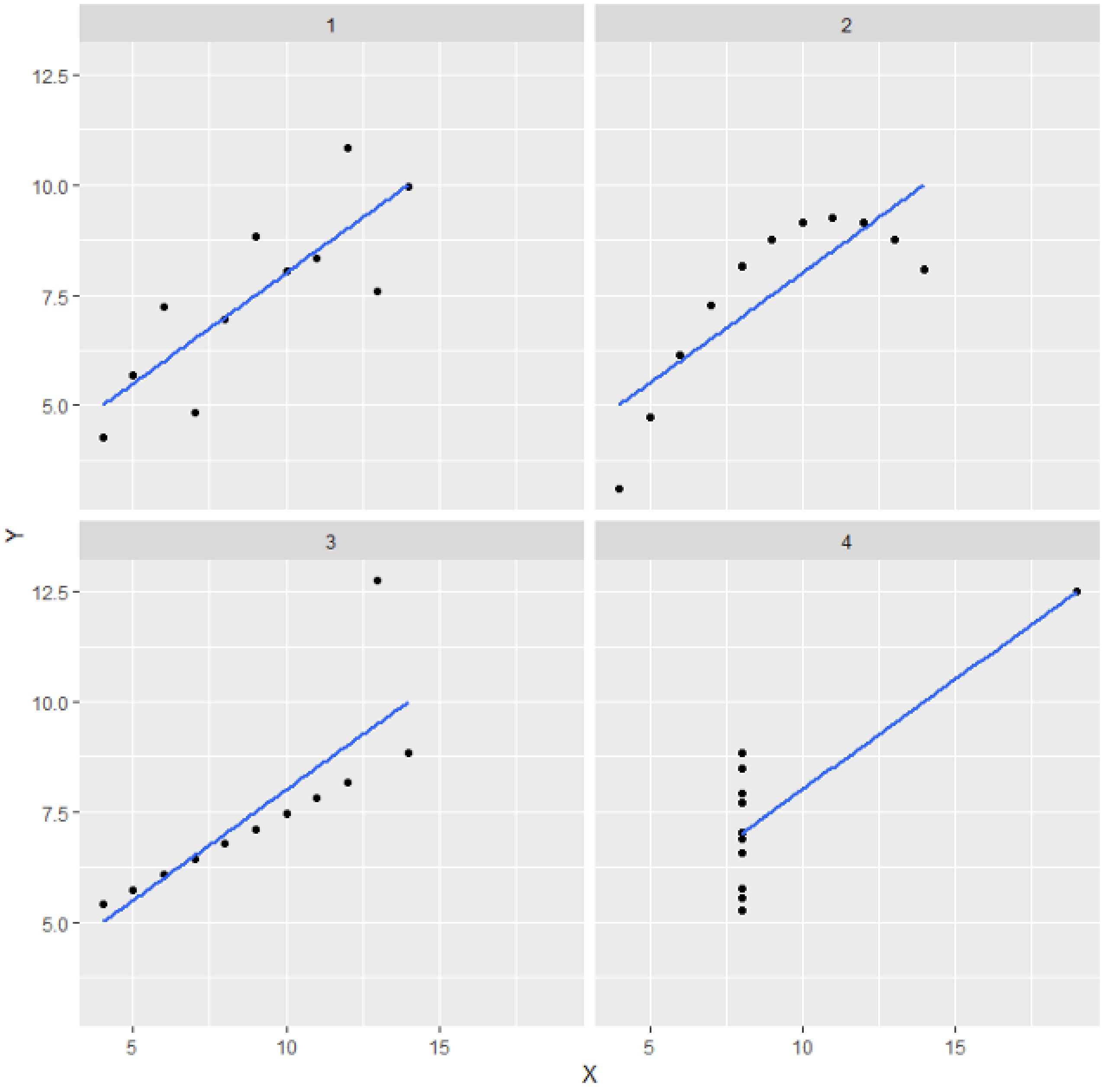
Linear regression is a machine learning algorithm which is of type supervised learning.
It performs a regression task in order to fit the best possible straight lie which defines the linear relationship between dependent and independent variables.
Mostly done by Residual Sum of Square method.
Assumptions of linear Regression

- Assumed to have linear relationship between dependent and independent variable
- Error terms are normally distributed, and have mean value of 0
- Residuals have same variance
- Residuals are independent of each other.
- No multicollinearity in data.

Hypothesis function for linear regression is y = mx + c, where c is intercept, x is independent variable and m is slope.

7. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four data-set which have exactly identical mean, standard variation and co-relation still their distribution is different from each other.



8. What is Pearson's R?

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.
It depicts the amount of change in one variable affects another variable.
It is measured by calculating the slope of the variables.

The value of Person r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.

9. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is a method to transform the continuous variables into a scale.
It is useful as when there are multiple independent variables in a model, many can be of different scales which could result out to be incorrect coefficient of the predictor/independent variables.

There are two type of scaling

- **Standardization** : The variables are scaled in such a way that their mean i s zero and standard deviation is one.
- **MinMax :** The variables are scaled in such a way that all the values lie between zero and one**.**

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

VIF (Variance Influence Factor) depicts the dependence of one variable on all other variables.
VIF = inf means infinity which mean your variable is highly cop-related with some other variable.
This sort of error can be encountered, in case while creating dummy variables first value is not dropped or you are calculating VIF with constant in RFE approach.

11. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
The quantile-quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line.