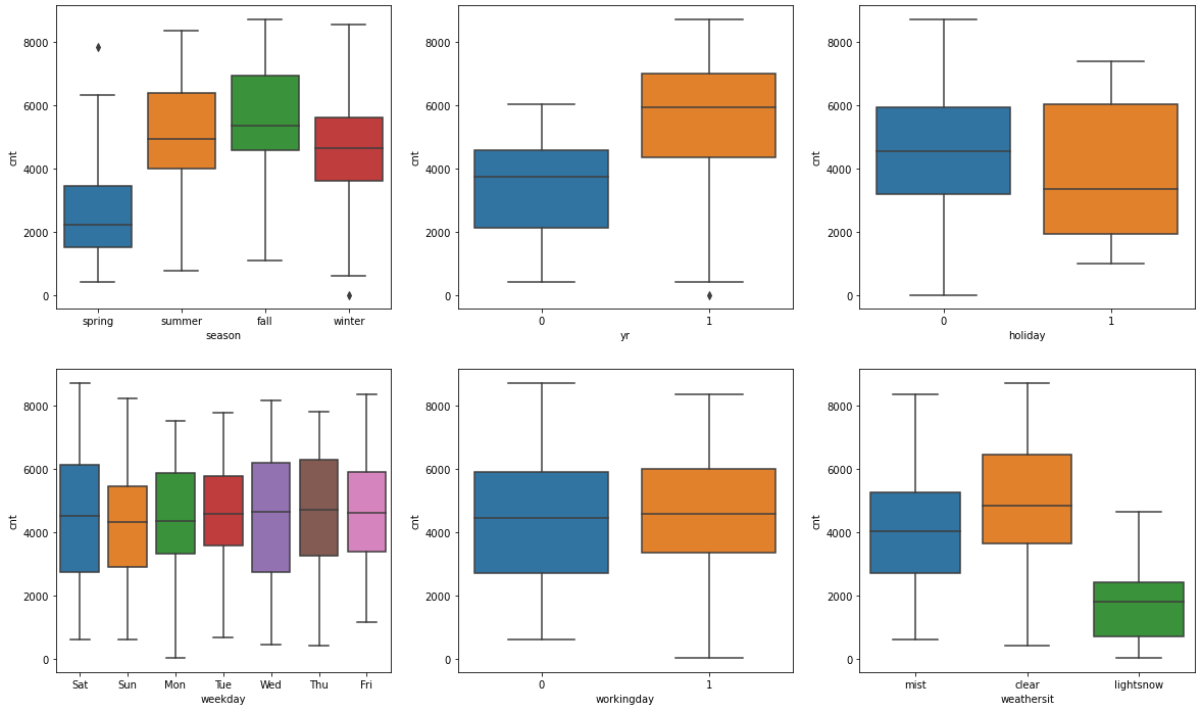
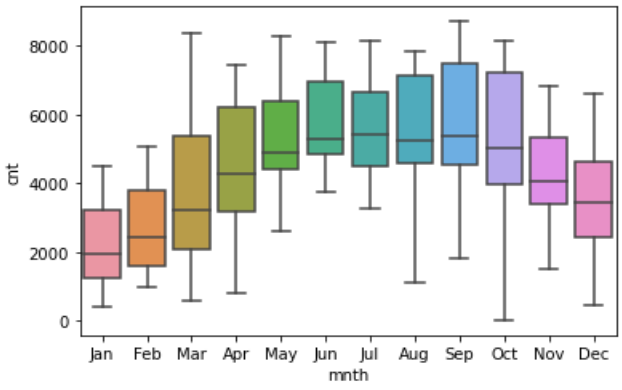
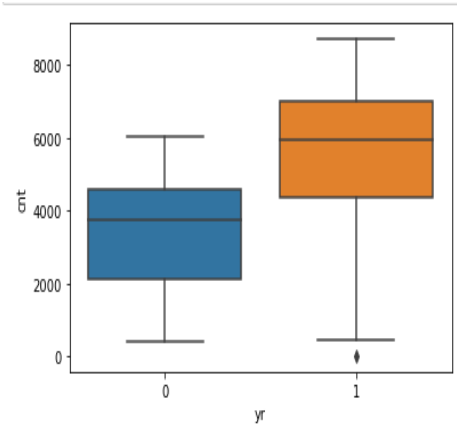
## Assignment-based Subjective Questions

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




- Year 2019 has been more demand for bikes than 2018
- May - October months has more demand than other months
- Weekday has more demands than weekend.
- Clear weather has more demand than mist and light snow.

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

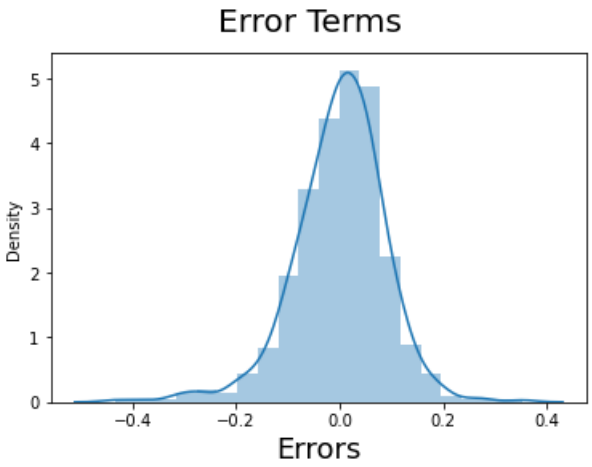
   It helps in reducing the extra column created during dummy variable creation. In other words, it reduces the correlations created among dummy variables.

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

with the target variable? (1 mark)

Attempt has highest correlation.

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




Error terms is distributed normally with mean 0

Other way to see that using VIF value

| | Features | VIF |
|---|---|---|
| 0 | const | 63.33 |
| 3 | atemp | 3.39 |
| 11 | season_spring | 2.69 |
| 12 | season_winter | 2.30 |
| 4 | hum | 1.91 |
| 9 | mnth_Nov | 1.70 |
| 7 | mnth_Jan | 1.65 |
| 14 | weathersit_mist | 1.57 |
| 6 | mnth_Dec | 1.44 |
| 8 | mnth_Jul | 1.30 |
| 13 | weathersit_lightsnow | 1.26 |
| 5 | windspeed | 1.21 |
| 10 | mnth_Sep | 1.12 |
| 1 | yr | 1.03 |
| 2 | holiday | 1.03 |

The VIF are below 5 so there is no multi collinearity exists between predictor variables

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   1) Atemp= Increase in one unit of temperature will increase the demand by 0.44 unit.
   2) Yr- Increase in year will increase the demand by 0.23 units
   3) Season_winter – has more demand than other season.

## General Subjective Questions

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

   Linear regression algorithm in ML is defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of following equation −

   Y = b+ m1*X1+m2*X2+m3*X3+…

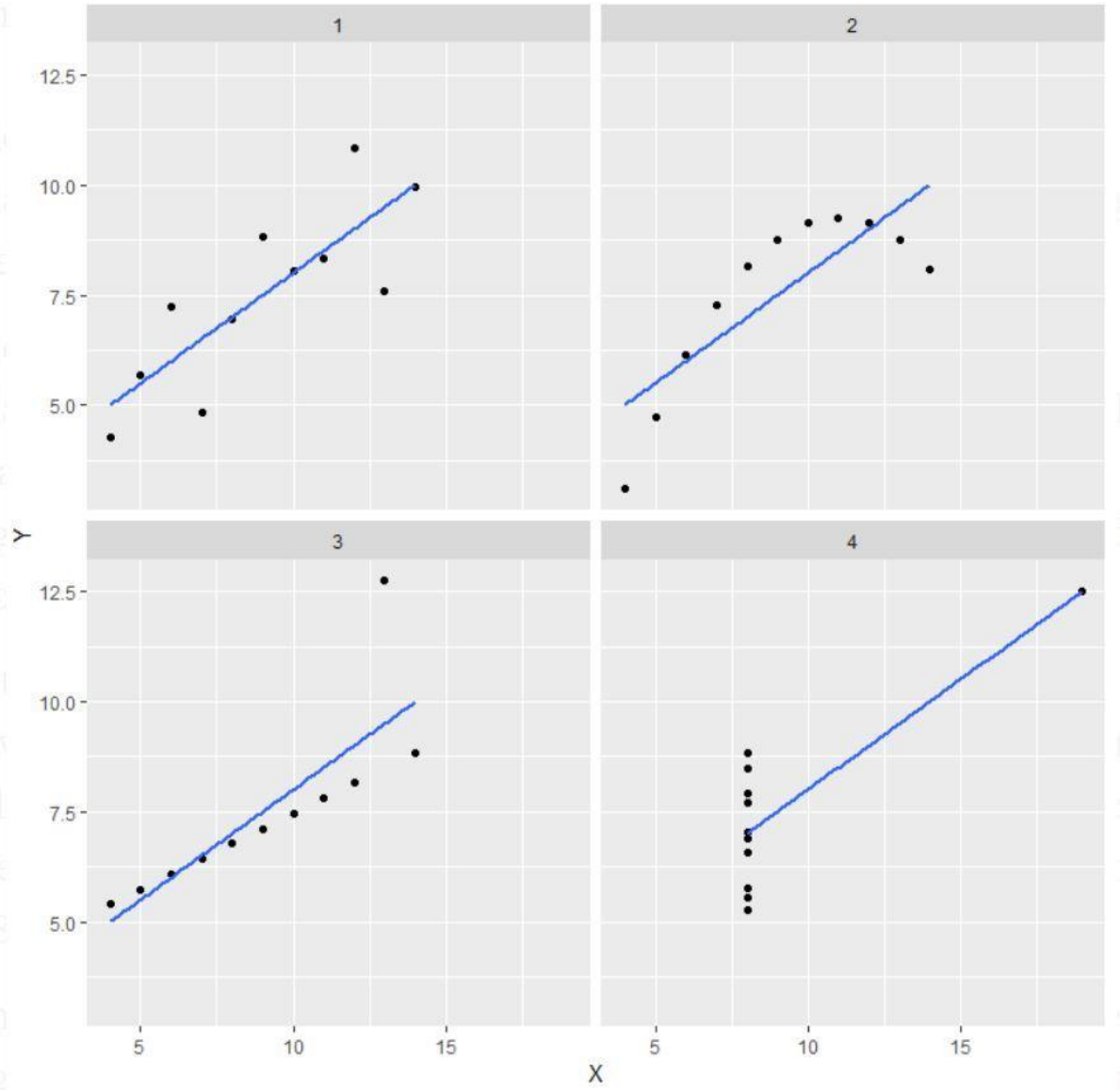
   Here, Y is the dependent variable we are trying to predict

   X1,X2,X3… is the dependent variable we are using to make predictions.

   M1,m2,m3 is the slop of the regression line which represents the effect X1,X2,X3 has on Y respectively considering X1, X2 and X3 has no relationship with each other.

   b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b.

7. Explain the Anscombe's quartet in detail. (3 marks)
   According to the definition given in Wikipedia, Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.




- In the first one(top left) in the scatter plot - there seems to be a linear relationship between x and y.
- In the second one(top right) - there is a non-linear relationship between x and y.
- In the third one(bottom left) -there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Here is the formula to calculate the Pearson's R.

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,

**N =** the number of pairs of scores

**Σxy =** the sum of the products of paired scores

**Σx =** the sum of x scores

**Σy =** the sum of y scores

**Σx2 =** the sum of squared x scores

**Σy2 =** the sum of squared y scores

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

   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

   This is performed to bring data consistencies among multiple independent variables when the scale/magnitude is different for their values. The variables might differ in magnitude and unit and ML models can take unit as a reference and might lead to incorrect modeling.

   Normalized Scaling- It brings all the data in the range of 0 and 1. This is also known as min max scaling.

   Min-max scaling x= x-min(x)/max(x)-min(x)

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

   Standardization x=x-mean(x)/sd(x)

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

    This shows a perfect correlation between two independent variables. In this case we might need to drop that variable which has perfect correlation.

11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

    Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

    In linear regression -This helps in a scenario of linear regression when we have the training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.