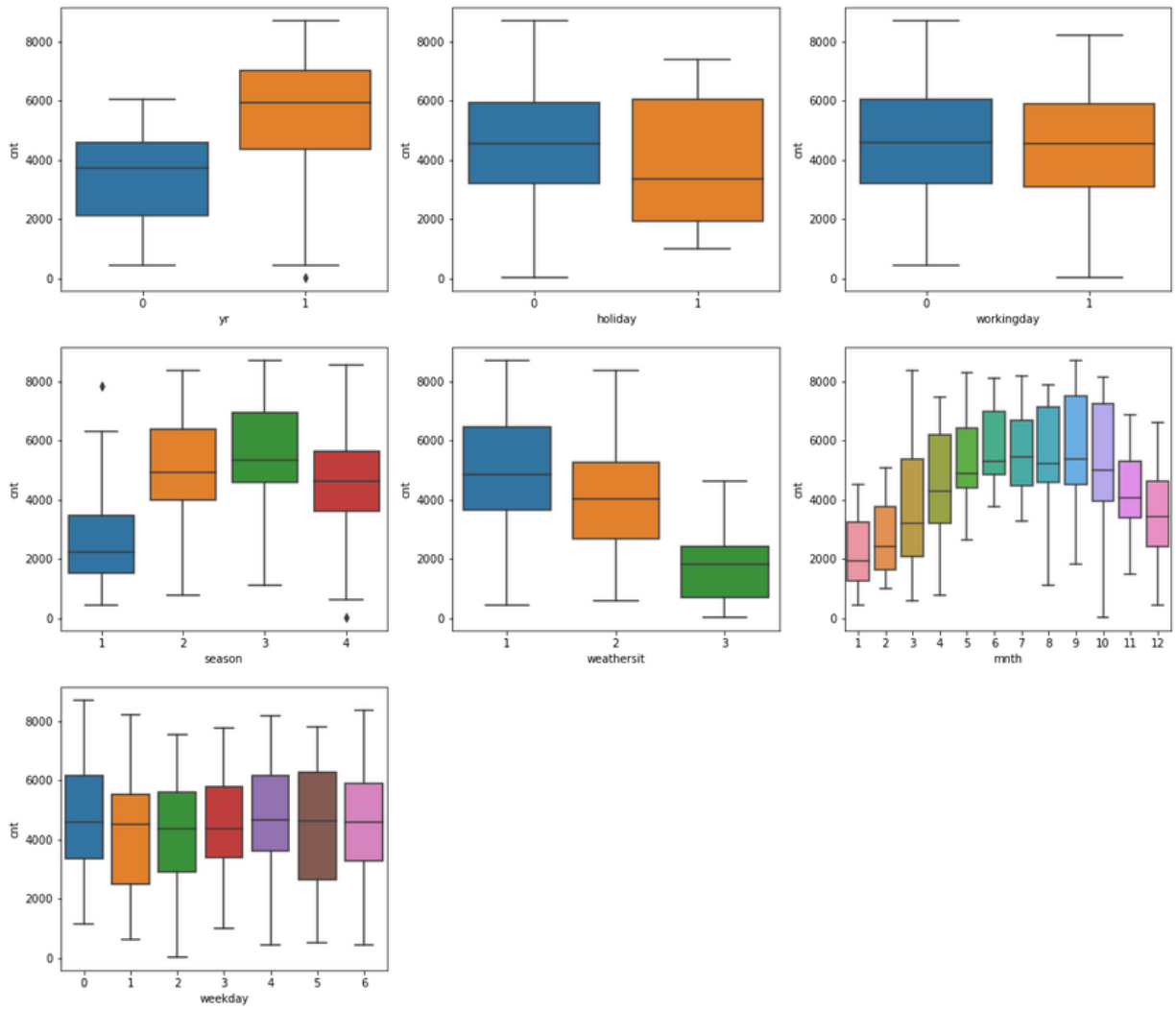
Assignment-based Subjective Questions

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

Ans.



- **Yr:** 2019 year has more bike booking than 2018
- **Holiday:** less bike booking when holiday & more bike booking in working day
- **Season:** 3->2->4->1 bike booking cnt order
- **Weathersit:** 1->2->3 bike booking cnt order
- **Mnth:** more bike booking in 5,6,7,8,9 mnth
- **Weekday:** not much influence of weekday

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

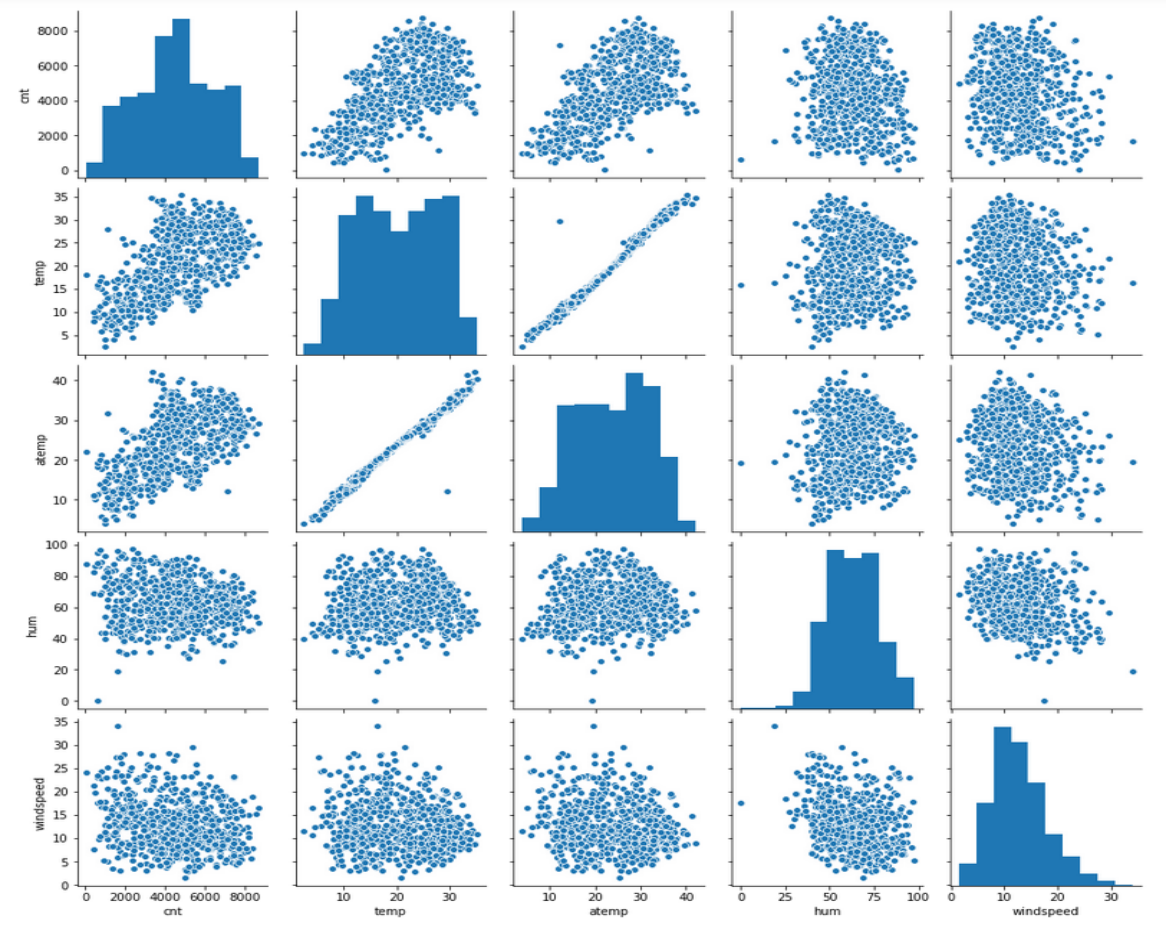
(2 mark)

Ans. If we don't drop the first column then our dummy variables will be redundant. If we keep all dummy variables it leads to multicollinearity between the dummy variable. And also, the first column is no of use it just polluting the data frame which may affect the model adversely.

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

(1 mark)

Ans.

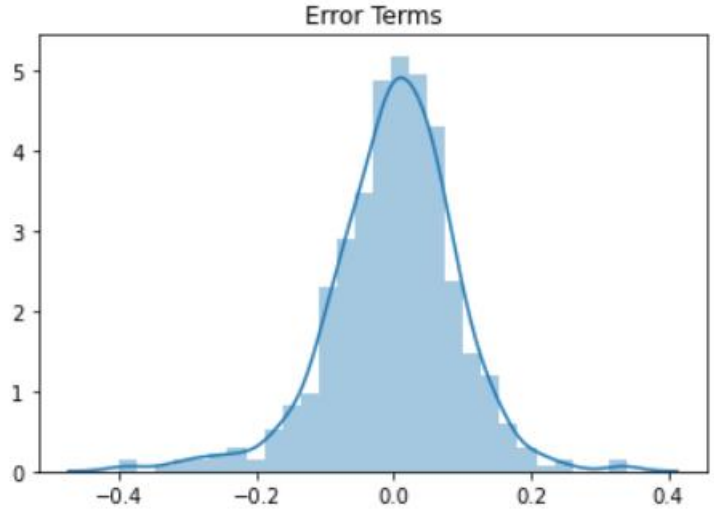


'temp' & 'atemp' are highly correlated with 'cnt'

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

(3 marks)

Ans.



Error terms are normally distributed i.e. mean is centered around hence assumption satisfied. We validate this assumption about residuals by plotting a distplot.

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

(2 marks)

Ans.

1. atemp: 0.536994
2. yr: 0.233436
3. season_winter: 0.131917

General Subjective Questions

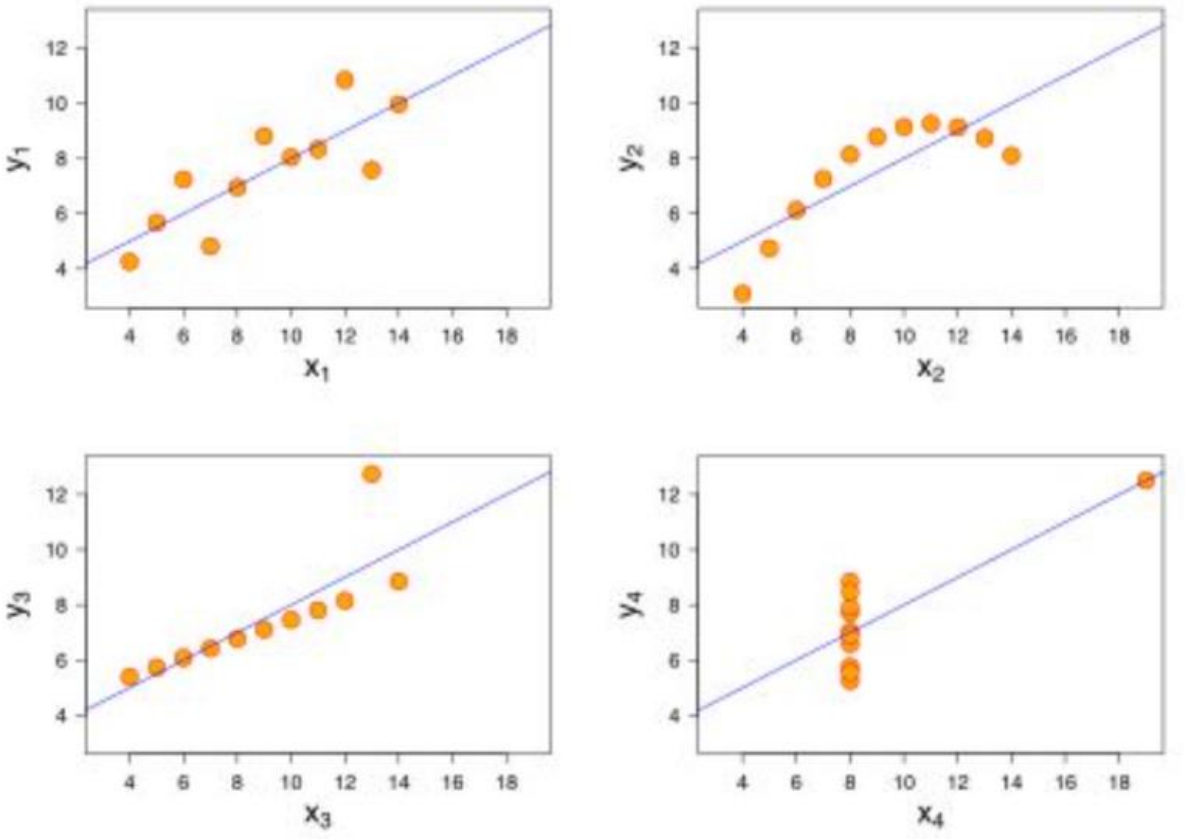
1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Linear Regression is a supervised Machine Learning Algorithm that is used for predicting numerical variables. It is a part of regression analysis. Linear regression is based on equation "**y = mx + c**" where m is the slope of line and c is the intercept. It assumes that there is linear relationship between dependent variable (y) and independent variable (x). In this first we find the best fit line which properly describe the relation between x(independent) & y(dependent). It only works when target variable is continuous in nature. It is used to finding out the effect on input variable on target variable, change in target variable with respect to one or more input variable and to find/predict upcoming trends. It is divided into two part

1. Simple Linear Regression – used when dependent variable is predicted using one independent/input variable. Equation => $y = mx + c$
2. Multiple Linear Regression - used when dependent variable is predicted using more than one independent/input variable. Equation => $y = c + m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_px_p$

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

Ans. It was constructed by Francis Anscombe in 1973 to illustrate the importance of plotting the graphs. Anscombe's Quartet includes fur 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 analyzing it and effect of outliers and other influential observations on statistical properties.



1. First plot is simple linear regression
2. Second plot is not distributed normally while there is a relation between then its nonlinear
3. Third plot the distribution is linear but should have a different regression line. Outlier involve in the data which cannot be handled by regression model
4. Fourth plot shows example of one high leverage point is not enough to produce a high correlation coefficient, even though the other data point should not indicate any relationship between the variables. Outlier involve in the data which cannot be handled by regression model

3. What is Pearson’s R? (3 marks)

Ans. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in

opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

(3 marks)

Ans. It 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. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

1. Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
2. Standardization Scaling:
   It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

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

(3 marks)

Ans. VIF is variance inflation factor. It gives how much the variance of coefficient estimate is being inflated by collinearity. VIF = 1/(1-r^2), if there is a perfect correlation, then VIF = infinity as r^2 is 1 for perfect correlation. VIF is infinity means 100% variance of that variable is explained by other variables. R^2 is the value of independent variable which we want to check how well this independent variable is explained well by another independent variable. If it is perfectly explained by other, then r^2 = 1 and VIF = infinity.

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

(3 marks)

Ans. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- 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?