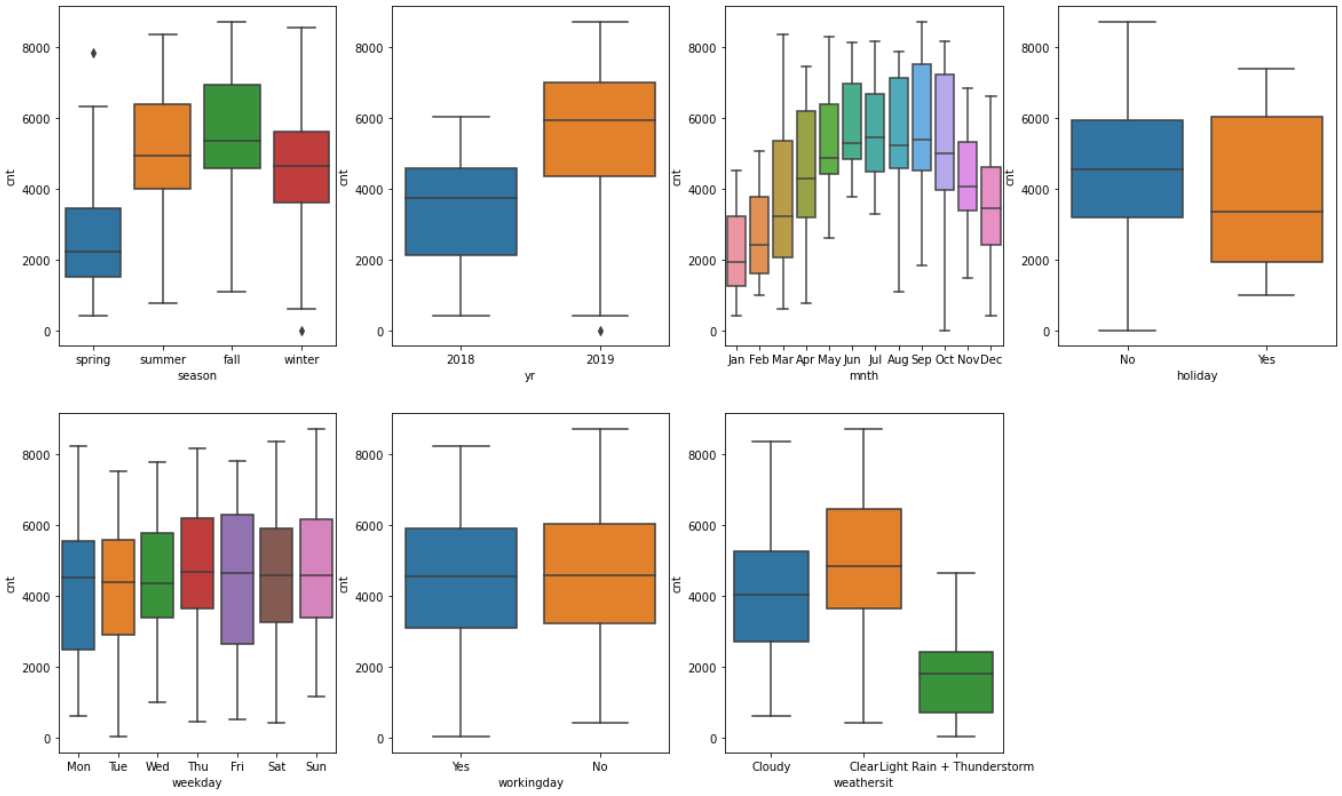
Assignment - based Subjective Questions

Question_1:

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

Answer:



- The categorical variable present in the bike_sharing dataset were, **`season, yr, mnth, holiday, weekday, workingday and weathersit`**
- By visualising using the box plot the following inferencens made,
  - **`Season`** - From the season box plot the `spring` season has the lowest demand and `fall` season has the highest demand for the bikes
  - **`yr`** - The Number of rentals was increased in `2019` as compared to `2018`
  - **`mnth`** - There is high demand in Sep whereas in Dec has the lowest demand, because in Dec usually, it will be heavy snow
  - **`holiday`** - Bike rentals reduced during holidays
  - **`weathersit`** - There is high demand in `Clear weather`, whereas low demand during `Light Rain + Thunderstorm days`, there are no users renting the bike during `Heavy snow`.

Question_2:

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

Answer:

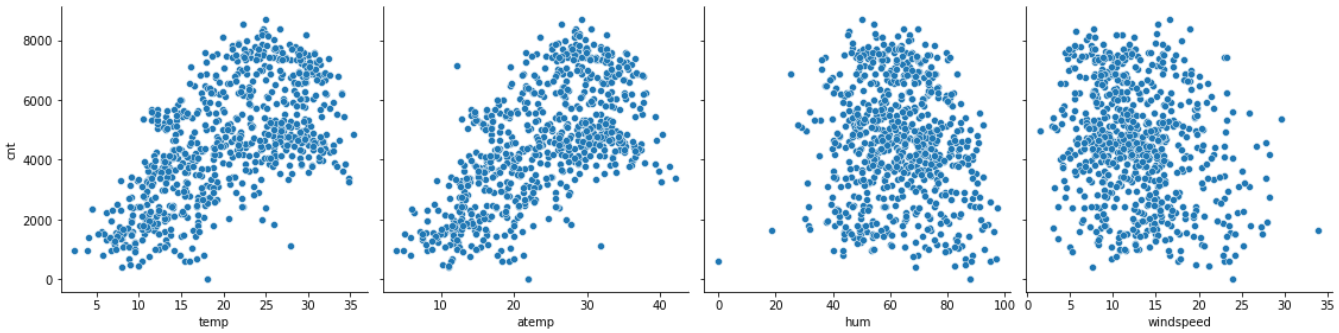
- It helps to reduce the correlations created among the dummy variables
- It helps in reducing the extra column created during dummy variable creation if we drop the first column of the created dummy variables still the information would remain the same

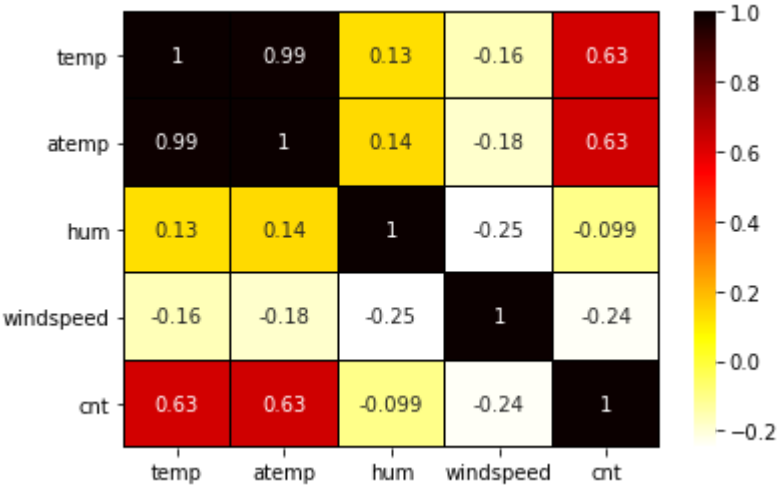
---

Question_3:

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

Answer:





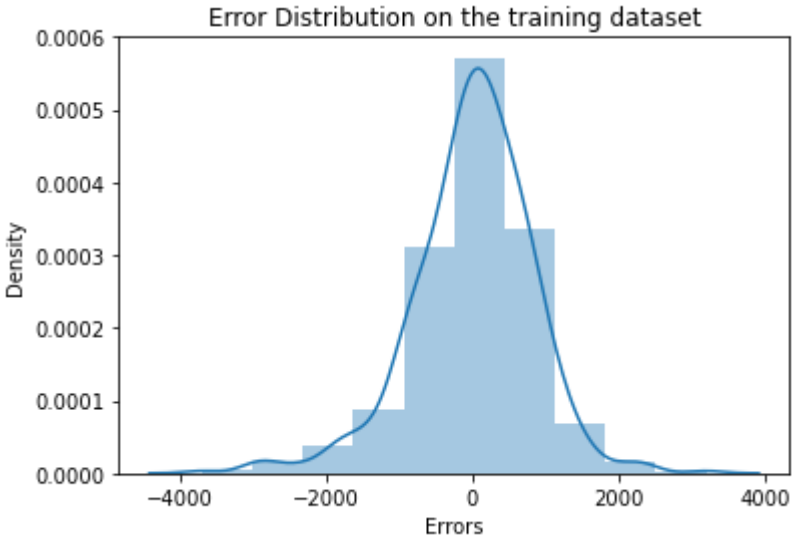
`temp` and **`atemp`** having the highest correlation with target variable **`cnt`** among numerical variables

---

Question_4:

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

Answer:



The error terms distribution should follow normal distribution and centred around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.The above diagram shows that the residuals are distributed about mean = 0.

Question_5:

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

Answer:

Top 3 features are,

1. `temp` has the coefficient of `3002.2414` with `cnt`
2. `yr` has the coefficient of `2064.0890` with `cnt`
3. `weathersit_Light Rain + Thunderstorm` has the coefficient of `-2404.2143` with `cnt`

General Subjective Questions

Question_1:

Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm based on the **supervised learning**. It performs **regression task** Regression models a target prediction value based on independent variables. It is mostly used to finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear Regression is based on the equation of $y = mx + c$

- It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).
- In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.
- Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.
- Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.
- In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

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

The equation for **MLR** will be,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

where,

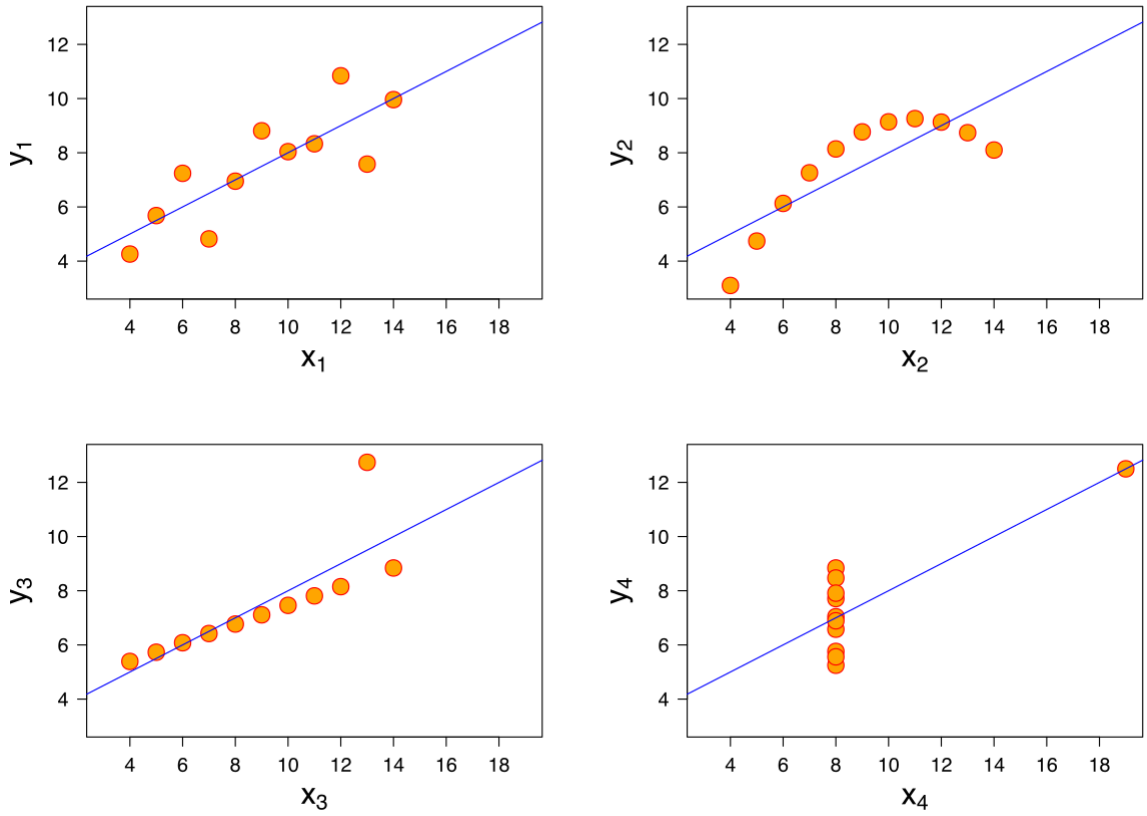
- β_0 - Intercept (constant term)
- β_1 - coefficient for the independent variable X_1
- β_2 - coefficient for the independent variable X_2
- β_p - coefficient for the independent variable X_p
- ϵ - is the error

Question_2:

Explain Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



statistical properties

- 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.

Question_3:

What is Pearson's R?

Answer:

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,

- 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

Question_4:

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

Answer:

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

Broadly there are two ways of feature scaling:

1. **Standardisation** - brings all the data into a standard normal distribution with mean 0 and standard deviation 1
2. **Normalisation** - brings all the data in the range of 0-1.

The formulae used in the background for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- Normalisation: $x = \frac{x - min(x)}{max(x) - min(x)}$

Note that scaling just affects the coefficients and none of the other parameters, such as t-statistic, F-statistic, p-values and R-squared.

Question_5:

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

Answer:

VIF - Variance Inflation Factor is the measure of amount of multicollinearity (How the independent variables correlating with each other) in a set of independent variable.

VIF is given by,

$$VIF = \frac{1}{1 - R^2}$$

- If there is a perfect correlation (R^2=1), then VIF will be infinity Where the R^2 value of that independent variable which we want to check how well this independent variable is explained well by other independent variables
- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1
- In that case the VIF will be

$$VIF = \frac{1}{1 - 1} = \infty$$

Question_6:

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Usage:

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

Advantages of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.

Type of Q-Q plots: