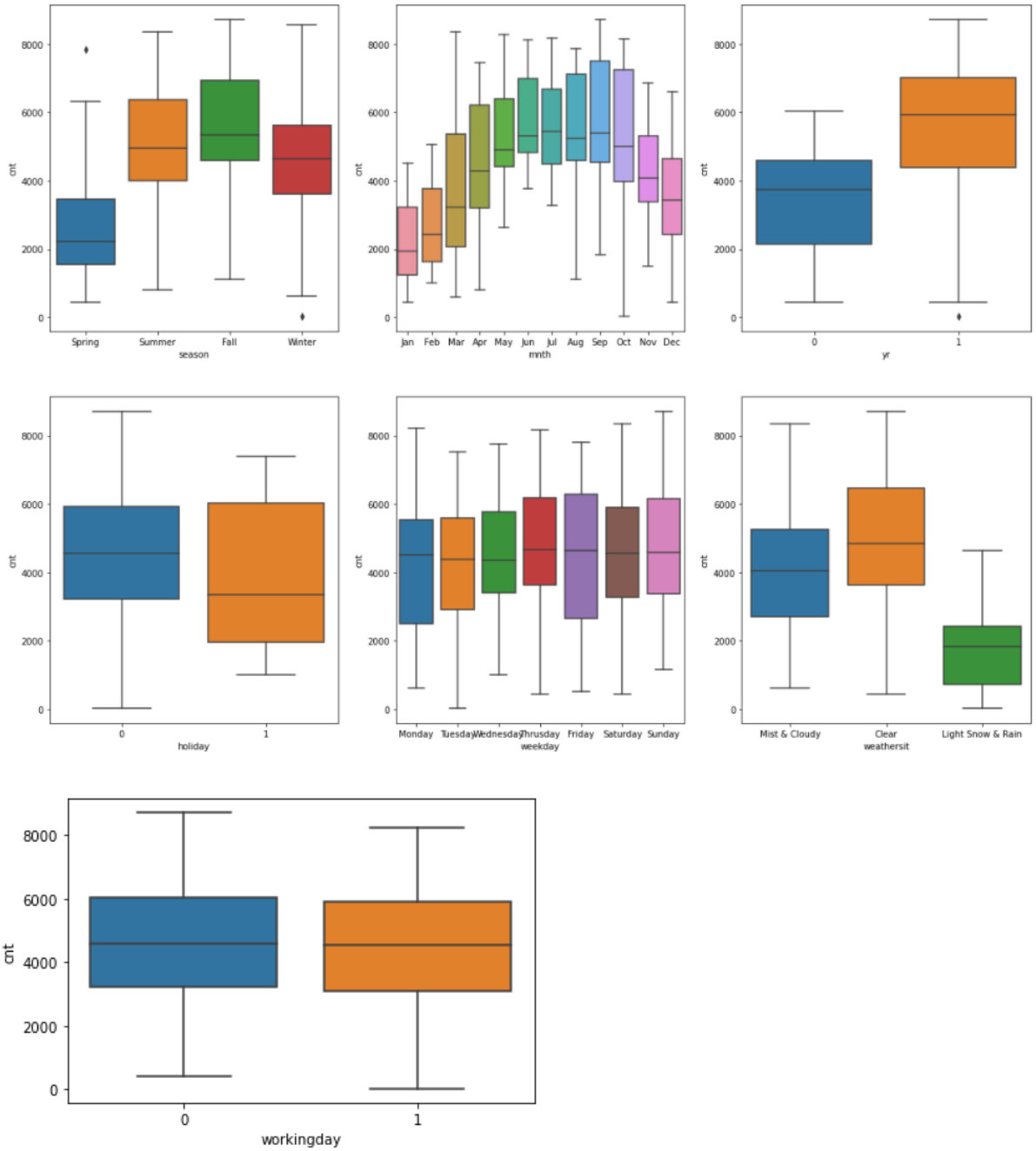
Assignment-based Subjective Questions

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

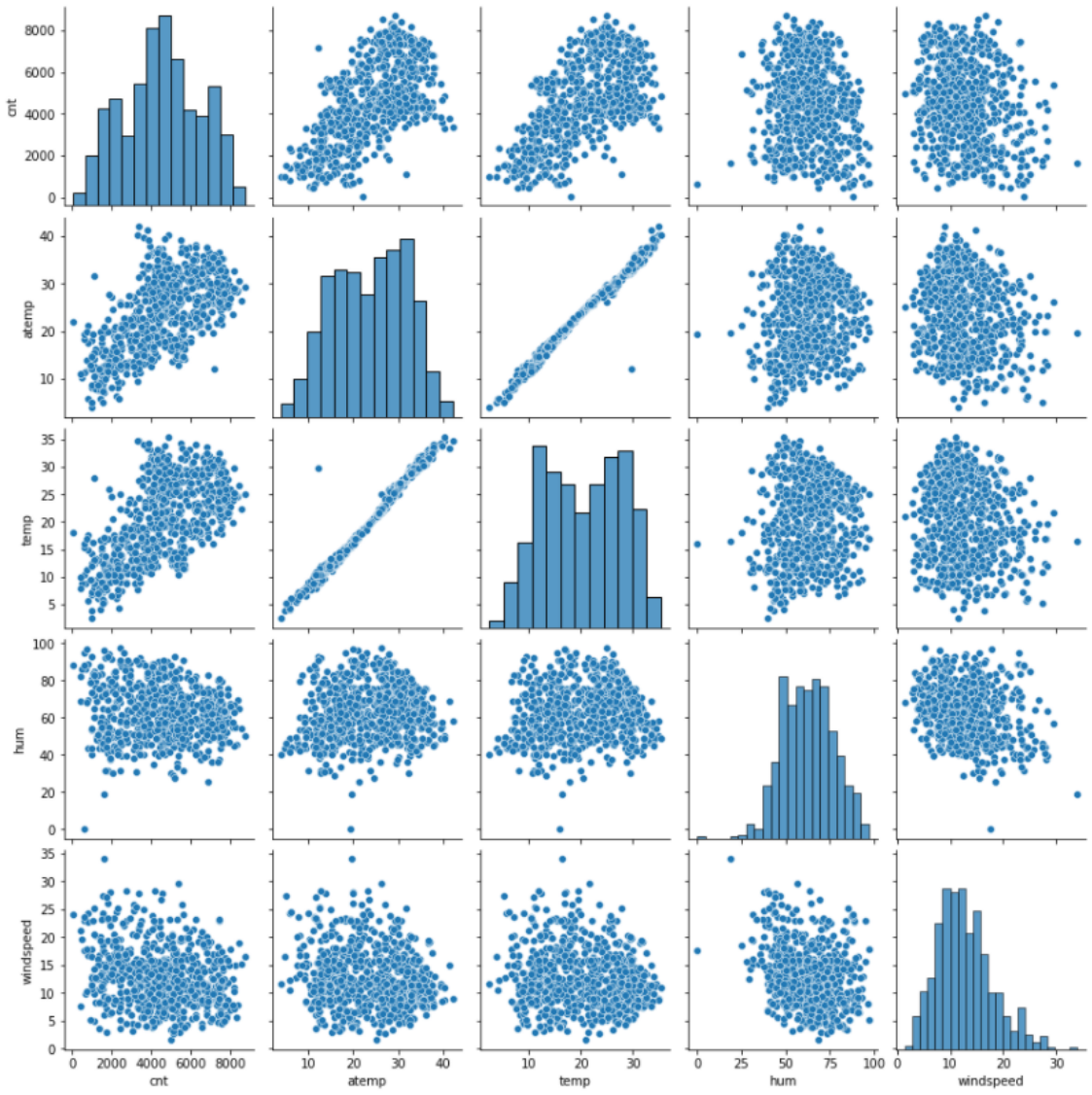


- Fall Season has higher count of rentals whereas spring has the least count of rentals
- Friday sees a higher count of rentals compared to other days of the week
- Clear weather situation sees the maximum rentals
- September has the highest number of rentals whereas December has the least count of rentals
- 2019 sees a higher count of rentals than 2018

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

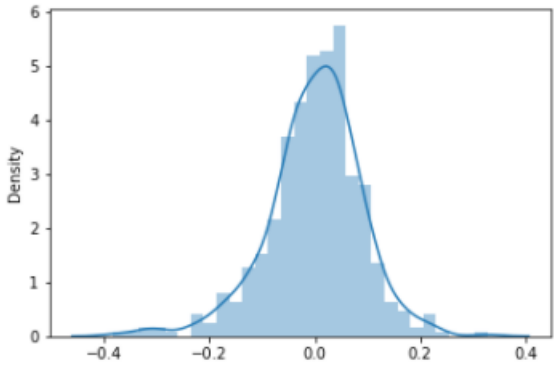
When dummy variation is created for all the categories in a categorical column, the n-1 columns can easily the explain the data for the nth column. So if we retain all the columns, it will lead to Multicollinearity. Hence, to avoid this situation, first column is dropped using drop_first = True

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



'temp' and **'atemp'** seem to have highest correlation with target variable.

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



For validation the density plot for residuals was plotted (y_actual – y_predicted).

It was seen the residuals are normally distributed about the mean 0, which validates the assumption for Linear Regression

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

```
const                          0.225733
yr                             0.227942
holiday                       -0.099419
temp                           0.597751
hum                           -0.182550
windspeed                     -0.189202
season_Summer                  0.080346
season_Winter                  0.135725
mnth_Jul                      -0.048116
mnth_Sep                       0.095936
weathersit_Light Snow & Rain  -0.233145
weathersit_Mist & Cloudy      -0.051698
dtype: float64
```

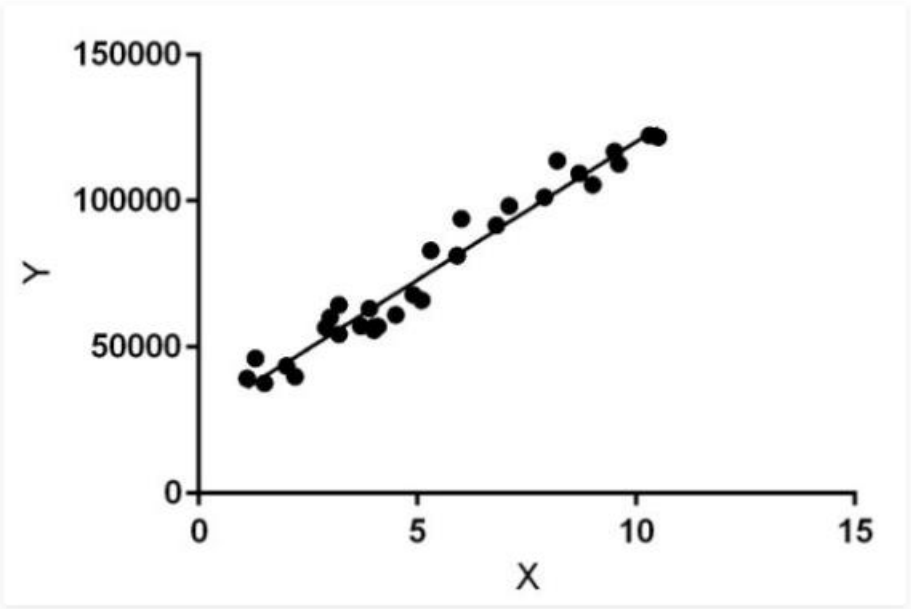
Below are the top 3 features

- Temp
- weathersit_Light Snow & Rain
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for 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 performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update θ_1 and θ_2 values to get the best fit line ?

Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

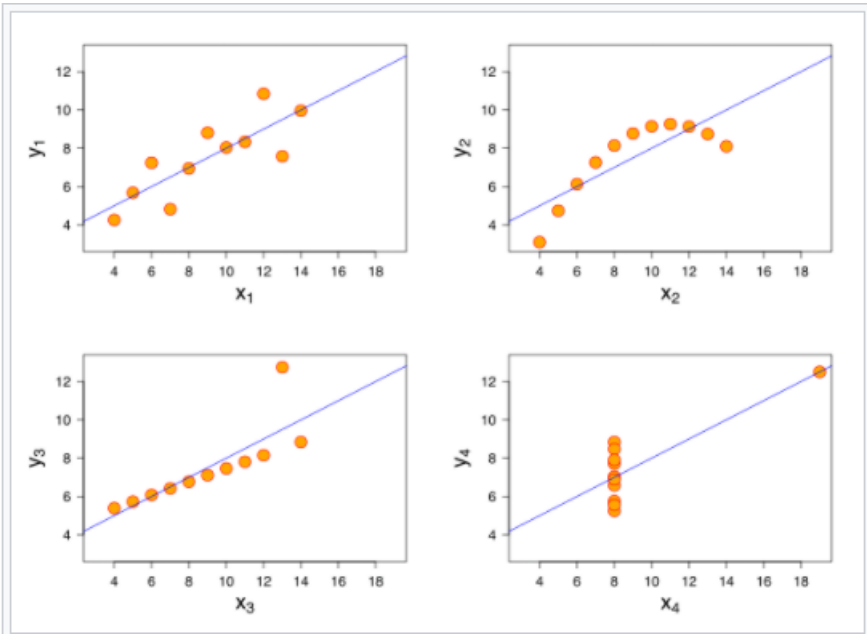
Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

Gradient Descent**:**

To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

2. Explain the Anscombe's quartet in detail.

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.



3. What is Pearson's R?

It's a coeff that denotes the strength of Linear relation between 2 variables

Its value ranges between +1 and -1

r = 1 : Perfectly linear with +ve slope

r = -1 : Perfectly linear with -ve slope

r = 0 : there is no linear association

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

Feature scaling is a method used to normalize the range of independent variables or features of data.

If not done, the ML model tends to weigh greater values, higher and consider small values as the lower values, irrespective of the units of the values

- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

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

   Formula for VIF is as follows

$$VIF_i = \frac{1}{1 - R_i^2}$$

   VIF will be infinite when R square = 1
   R2 will be 1 when that independent variable is perfectly explained by other independent variables
   In such a case VIF becomes Infinity

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

In statistics, a **Q–Q (quantile-quantile) plot** is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.