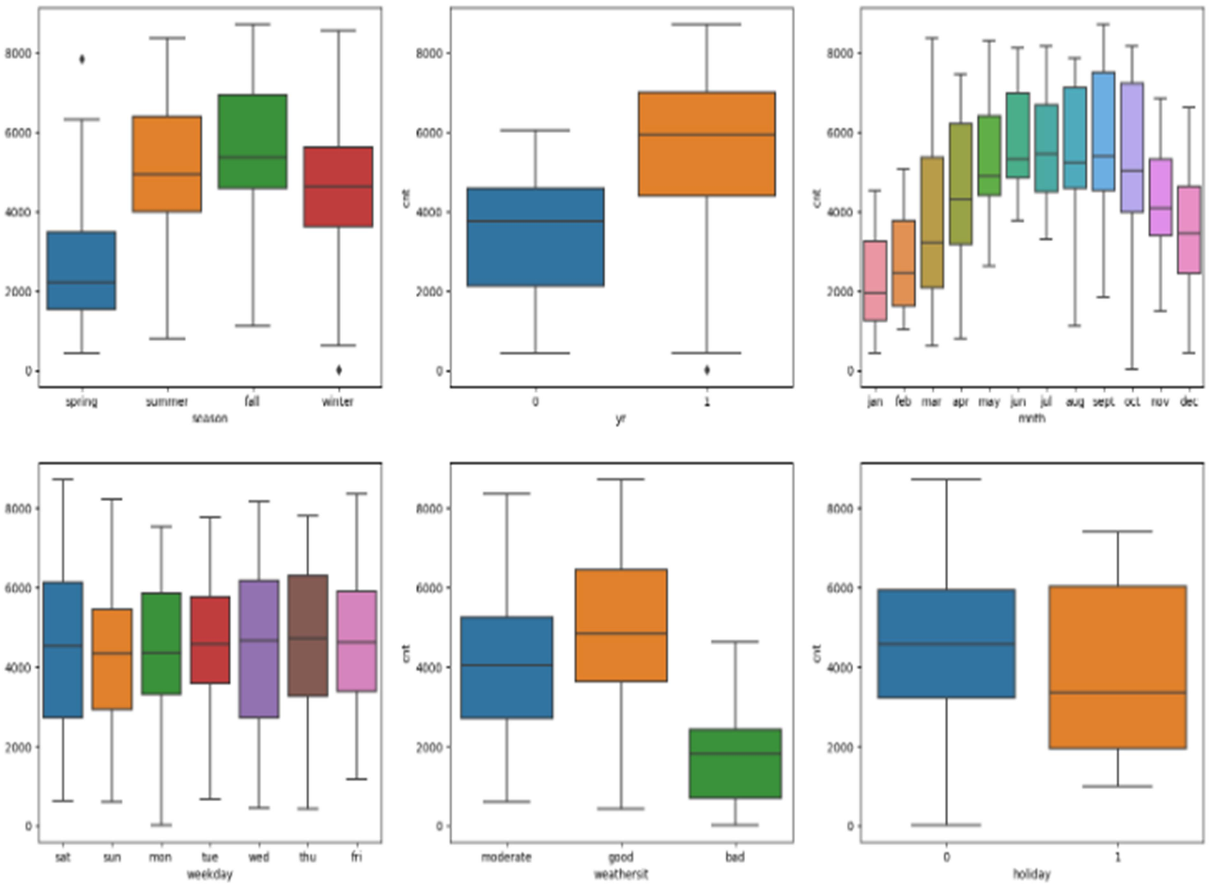
Assignment-based Subjective Questions

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- The categorical variables in the dataset are season , holiday, weathersit , mnth , weekday and yr .These can be visualized using a boxplot .
These variables had the following effect on our dependant variables

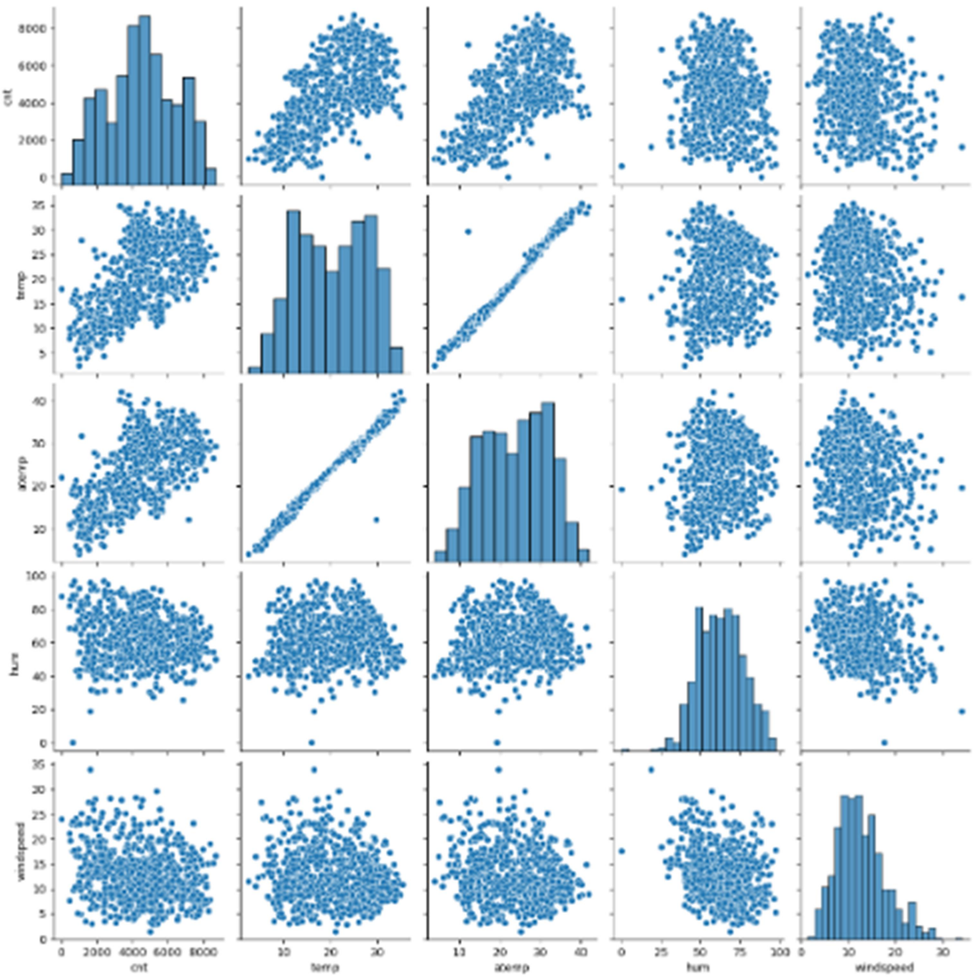


1. Season - The boxplot shows that spring season had least value of ‘cnt’(count of total rental bikes) whereas fall had maximum value of cnt means fall has highest demand for rental bikes . Summer and winter had intermediate value of cnt.
2. Weathersit – On good weathersit, demand is high .On un favourable weather conditions, there are no users.
   Holiday - rentals reduced during holidays.
   The clear(good) weathershit has highest demand.
3. Mnth - Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing
   September has the highest number of rentals, while December has least number.
   The weather condition in december is usually heavy snow.
4. Yr - The number of rentals in 2019 was more than 2018.

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

Ans-drop_first=True is used to delete first column in the dummy variable creation. It helps to reduce the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importances may be distorted.Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column.

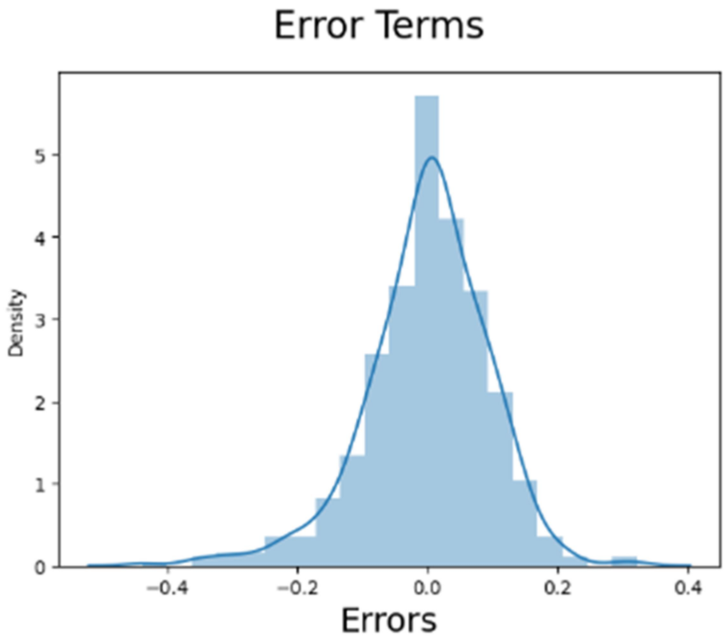
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



Ans - “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

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

Ans –



Residuals distribution is a normal distribution and centred on 0 (mean = 0). We can validate this assumption about residuals by plotting a distplot of residuals and can see if residuals are following normal distribution or not .The above diagram shows that the residuals are distributed and mean is zero.

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 -temp - coef : 0.4558

weathersit_Bad - coef : 0.2637

yr - coef : 0.2306

General Subjective Questions

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

   Ans –

   ○ Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.
   ○ Linear Regression is the most basic form of regression analysis .Regression is the most commonly used predictive analysis model.
   ○ Linear regression is based on the popular mathematical equation y = mx + c.
   ○ It assumes that there is a linear relationship between the dependent variable(y) and the 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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

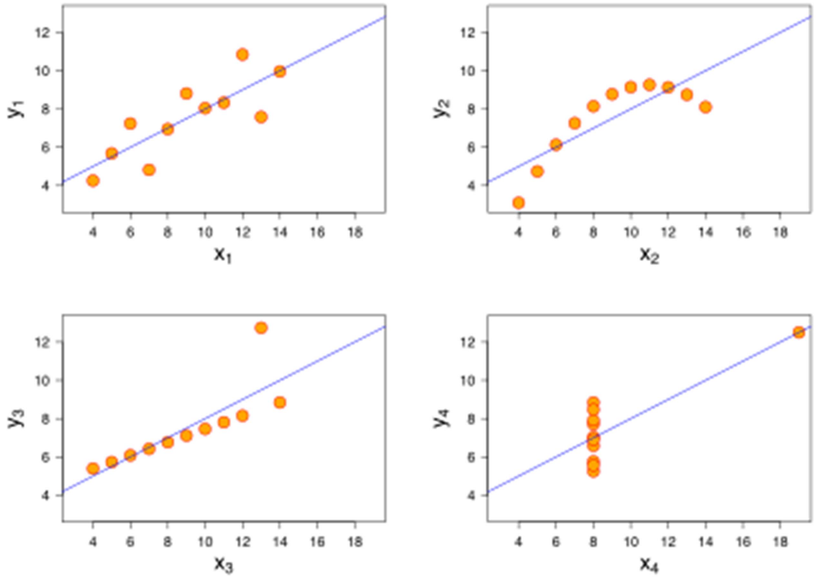
β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

β3 = coefficient for X3 variable and so on… β0 is the intercept (constant term).

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

Ans - Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but some peculiarities in data set that fools theregression model if built, 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 the effect of outliers and other influential observations .



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.

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.

   It's value ranges between -1 to +1.

   It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data?

   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 that means there is no o linear association.

4. 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. It also helps in speeding up the calculations in an algorithm.

   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

   Normalization

   It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

   Standardization

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

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

   Ans - VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

   (VIF) =1/(1-R_1^2). If there is perfect correlation, then

   VIF = infinity.Where R-1 is the R-square 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.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in infinity.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let's discuss some of the probability distributions.

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 compare the shapes of distributions, Providing a graphical view of how properties such as location, scale,and skewness are similar or different in the two distributions.