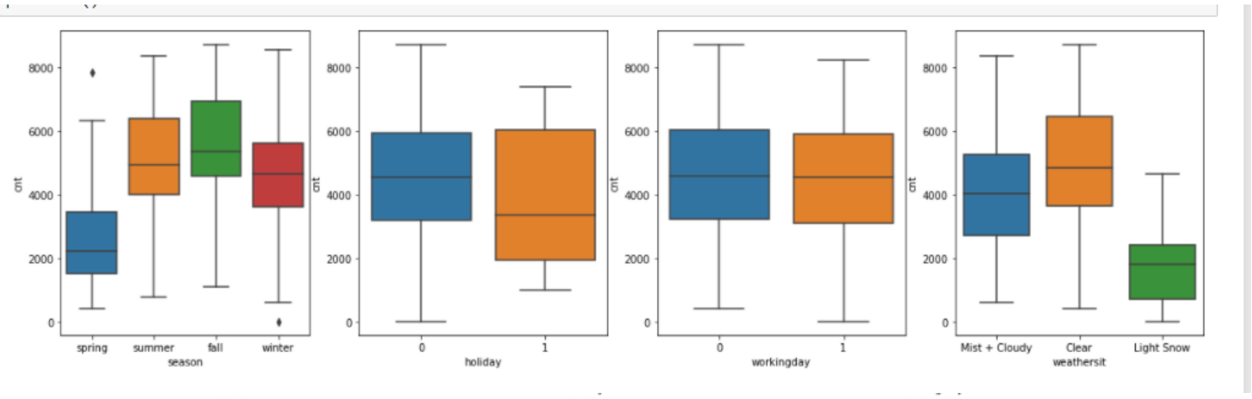
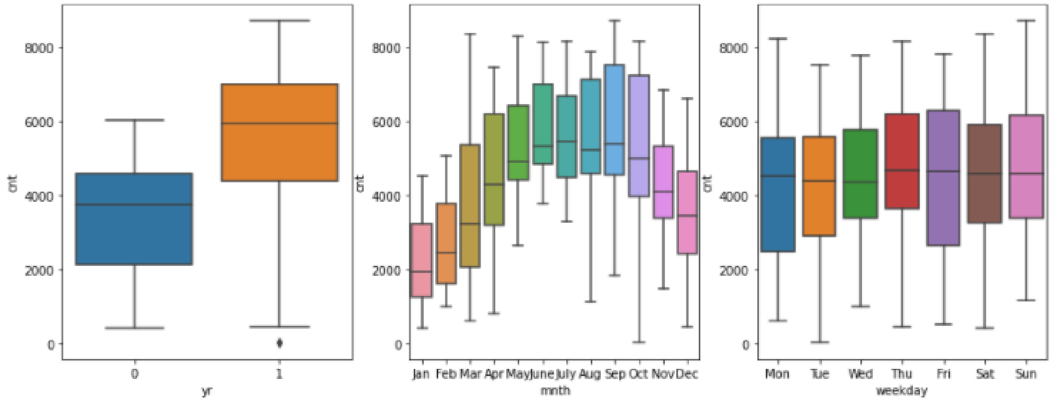
Assignment-based Subjective Questions

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

ANSWER





- season: -

   - season-spring bike rented count is less compared to 4 season and 75 count lies much below 25% of remaining season there is less usage of bike in this season.

   - season-summer in this season the bike is second most usage and there is slight difference between median.

   - season-fall is the most number of rent of bikes happened in this seson.

   - season-winter is 3 highest for renting the bike.

- yr.: -

   - in 2019 bike rented is more more compare to 2018.

- mnth.: -

   - September is the mnth for highest bike rent followed by oct then Aug.

- weekday: -

   - weekday variable shows very close trend.

   - Sunday is the day where bike rented was followed by Monday.

- Friday marginally higher than on Thursday but both median of Thursday and Friday are same.

- holiday: -

  - 0 and 1 which indicate working day and holiday we can see that on Woking day bike rented is more and we can see more number of bike were rented when it was not holiday.

- weathersit: -

  - bike rented is more in clear and low on light snow this show some relation with cnt.

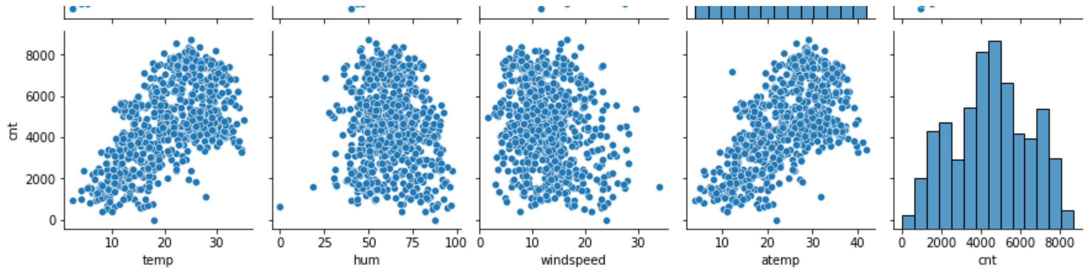
2. **Why is it important to use drop_first=True during dummy variable creation?**

ANSWER

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence 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?**

ANSWER



From the pair plot, temp and atemp has highest correlation with target variable.

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

Validate the assumption of linear Regression after building model.

1)Error term are normally distributed with mean 0.

2)Error terms do not follow any pattern.

3)Multicollinearity check using VIF.

4)Ensure the overfitting by looking at r2 and adjusted r2.

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

Light snow=-0.3354
It indicates that it has negative correlation ,1 unit in light snow would decrease the target variable to 0.3354

Spring =-0.2621
It indicates that it has negative correlation ,1 unit in increase would decrease the target variable to 0.2621

Yr= 0.2448
It indicates that it has positive correlation ,1 unit in increase would increase the target variable to 0.2448.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
ANSWER

- Linear regression is defined as the statistical model that analysis the linear relationship between the dependent and given set of independent variables.
- Linear relationship means when value of one or more variable independent variable increases the value of the dependent variable is affected means its is increased or decreased.
- Mathematically the relationship can be represented with the help of following equation

$$Y = mX + b$$

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

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

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

- Furthermore, the linear relationship can be positive or negative in nature as bsed on how it affects the variable explained below

  1)Positive linear relationship.

  If we increase indecent variable value which cause to increase in the dependent variable this is called as positive linear relationship.

  2)negative linear relationship.

  If we increase indecent variable value which cause to decrease in the dependent variable this is called as negative linear relationship.

- Further it can be divide into: -

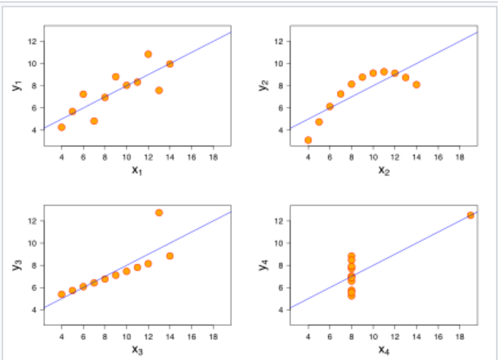
  Simple linear regression-which include modelling with 1 independent variable.

  Multiple linear regression-which include more then one variable.

2. Explain the Anscombe's quartet in detail.

ANSWER

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distribution and appear very different when graphed. Each dataset consists of eleven (x,y)points.



- The first scatter plot (top left) appears to be a simple linear relationship , corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). 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.

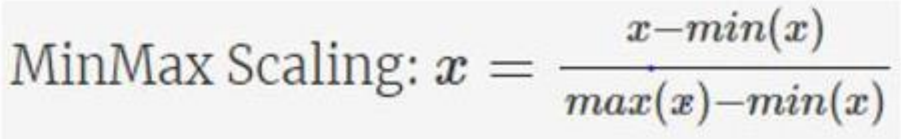
3. What is Pearson's R?

- Pearson r is a numerical summary of the strength of the linear association between the variable .if the variables tends to go up and down together , the correlation coefficient will be positive .
- Pearson r measure the strength of the linear two variables.
- Pearson r always between -1 and 1.
- If data lie on a perfect straight line with negative slope then r=-1

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

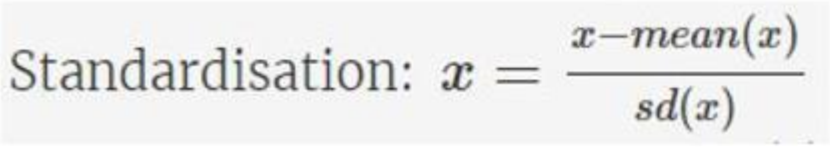
ANSWER

- Scaling is the method to normalize the range of independent variable.
- It is performed to bring all the independent variable on the same scale in regression. If scaling is not done the model will consider higher value is significant and lower value is non-significant.
- It is important to note that scaling only affects the coefficient and it will not affect p value r-square,f-statistics etc.
- Normalize scaling or min-max scaling :- brings the data into the range of 0 and 1



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardize scaling:- standardize scaling replace their values by z scores .it brings all the data into standard normal distribution which as mean 0 and standard deviation 1



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

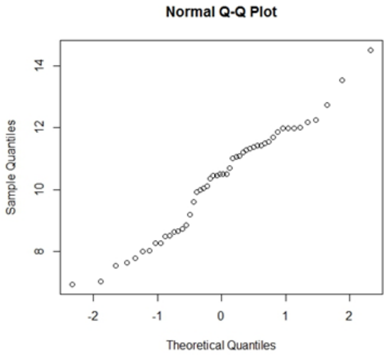
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.