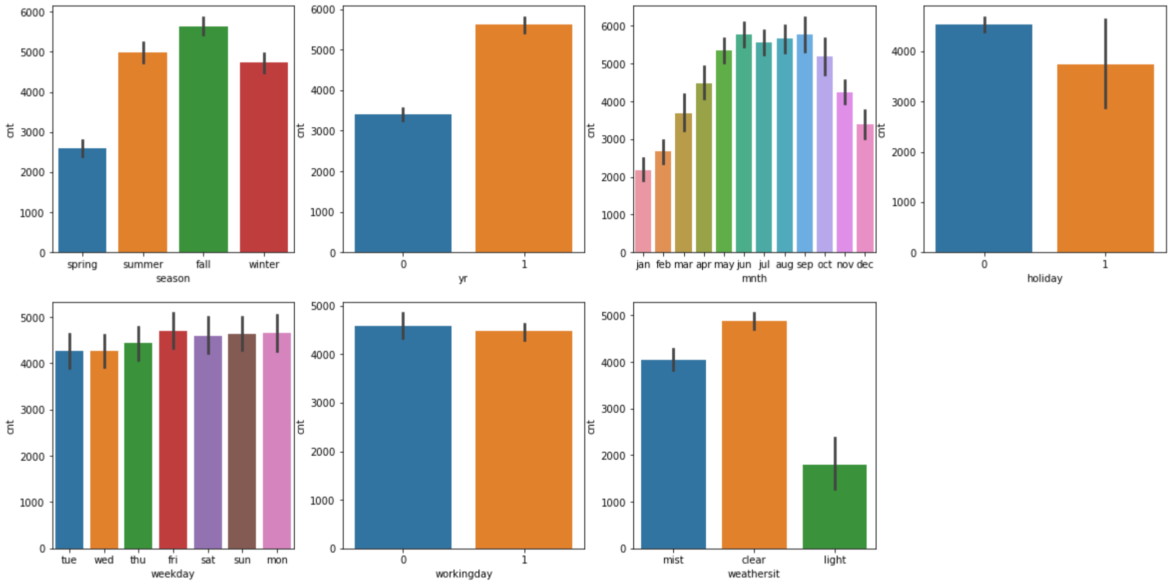
Assignment-based Subjective Questions

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

   Answer: (3 marks)



Observations:

- *Most bikes were rented in fall*
- *There are more bikes sold in 2019, then 2018*
- *From July to September rental rates were higher*
- *Bikes were rented more when it was not hliday*
- *Working days had higher rentals*
- *When the weather was Clear, Few clouds, Partly cloudy, Partly cloudy bikes were rented more*

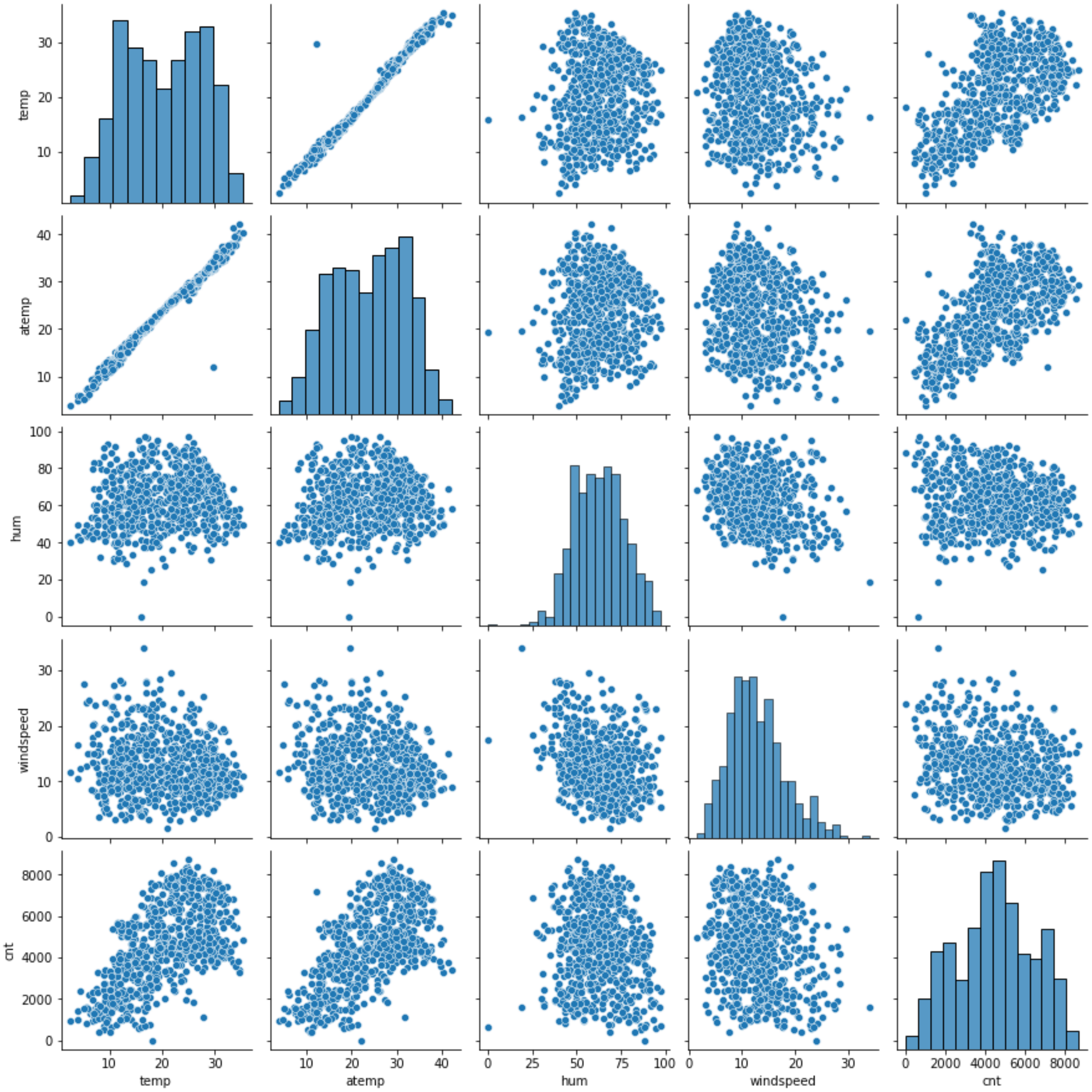
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Answer:

   It is important to use drop_first = True as it helps to reduce extra column during dummy variable creartion.

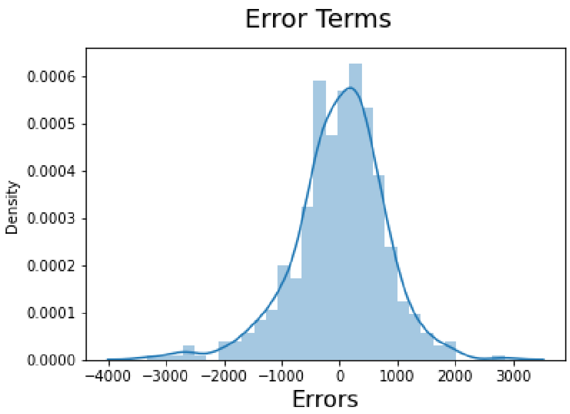
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:



“temp” and “atemp” had the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - Removed the column “atemp” since it was highly correlated to “temp”(0.99) to avoid multicollinearity.
    - Checked if error term was normally distributed



    - Checked data for homoscedasticity.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

- *Temp*
- *Yr*
- *Winter*

General Subjective Questions

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

Answer:

*Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on given data, between independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.*

There are two type of linear regression:

- ***Simple Linear Regression:** It explains the relationship between one dependent variable and only one independent variable using a straight line. A straight line is plotted on the scatter plot.*

  Formula for the Simple Linear Regression:

$$\boxed{Y=\beta_0+\beta_1X_1}$$

- ***Multiple Linear Regression:*** *It shows the relationship between one dependent variable and many independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.*

Formula for the Multiple Linear Regression:

$$\boxed{Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\epsilon}$$

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:

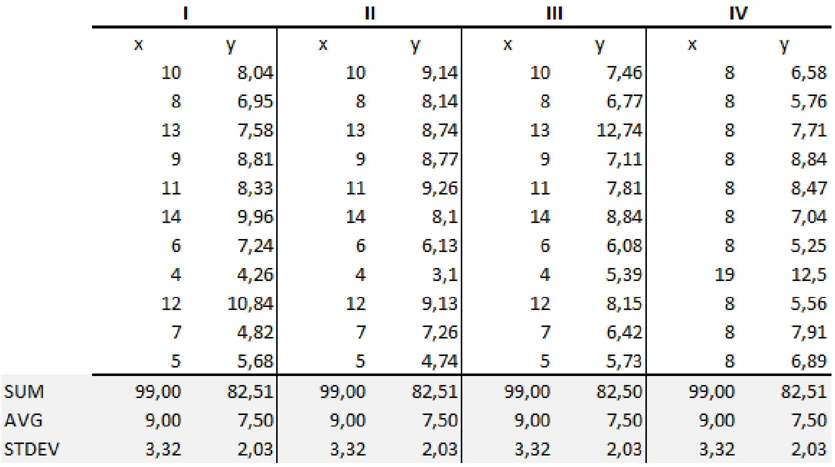
- *Differentiation*
- *Gradient descent*

Statsmodels or SKLearn libraries can be used in python for the linear regression.

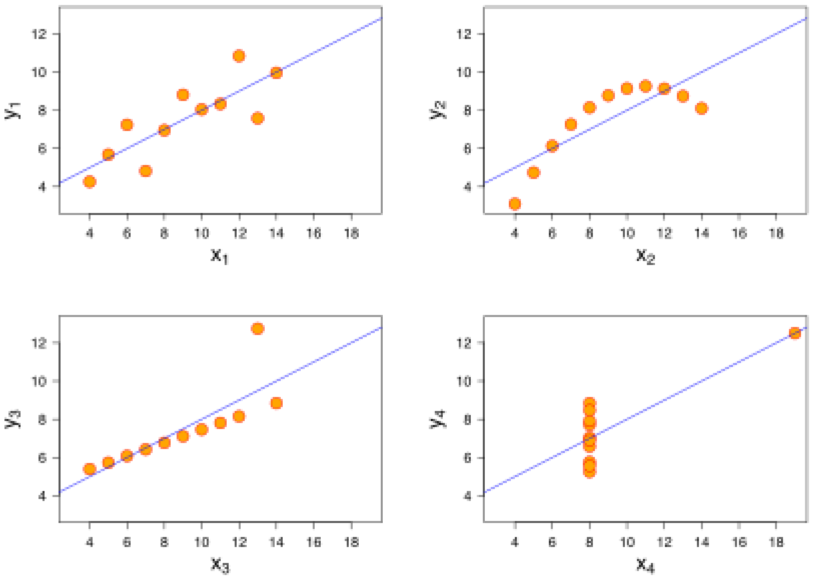
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

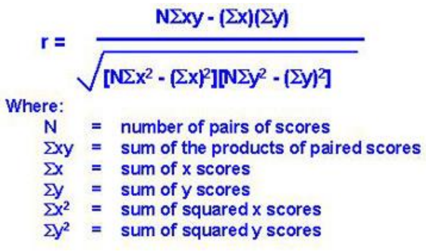


3. What is Pearson’s R? (3 marks)

Answer:

Pearson’s R was developed by Karl Pearson and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by ‘r’. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Formula



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

Example:

In our bike book case study, we saw that correlation between temp and atemp was 0.99 which means they are highly correlated.

4. 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling: *It is a preprocessing step, performed at independent variable to generalize daya within a particular range. It helps in speeding the calculations in an algorithm.*

Why Scaling: *Most of the time, data received contains features that vary highly in magnitude, units and range. If this data is not scaled, the model will calculate the coefficient incorrectly because higher data will have higher magnitude. Hence Scaling is done before building the model.*

Normailization: *Equation of normalization is derived by subtracting the variable with the minimum value of the column and dividing it by difference of min and max value of the column.*

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization: *Equation of Standardization is derived by subtracting variable with mean value od the column and dividing it by standard deviation of the column.*

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

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

Answer:

The value of VIF is calculated by the below formula:

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

If R-squared value = 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

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.*