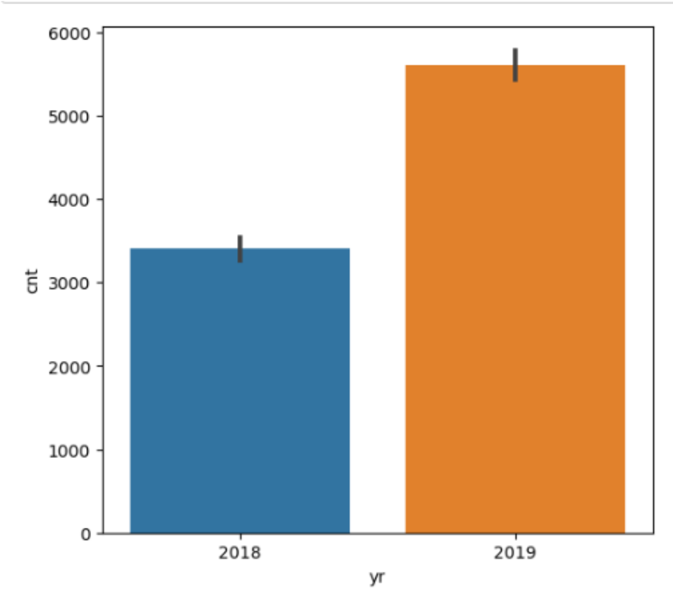
Assignment Questions-Linear regression

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

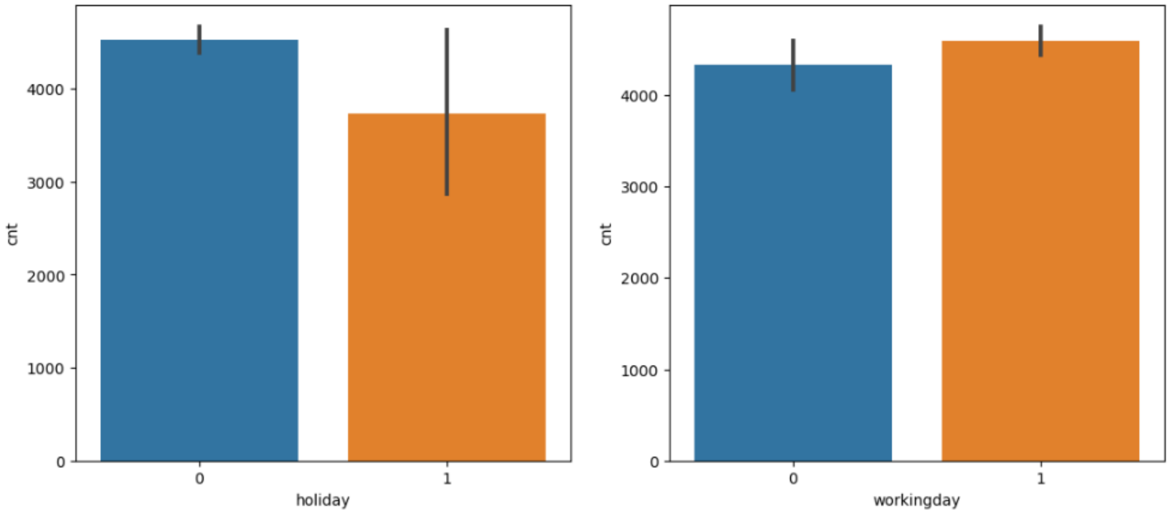
The final equation is
*0.1168+0.1115workingday+0.5865temp-0.1461hum-0.1625windspeed+0.2295*2019+0.0770sep+0.1123winter-0.0513Mist_few_clouds-0.2486snow_rain_thunderstrom+0.1190sat+0.0655sun*

<u>Impact of categorical variables on the target variable:</u>

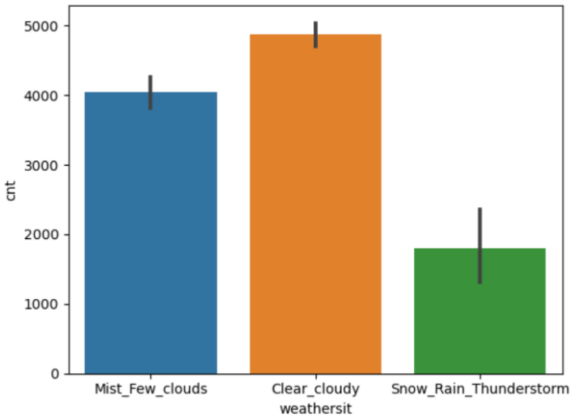
1. Year 2019 has seen surge in users



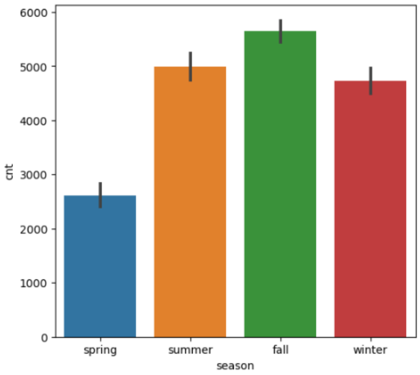
2. More number of users on Working day and non-holiday



3. When the weather is Clear, we can see more number of users hiring bikes



4. Maximum hiring can be seen in Fall

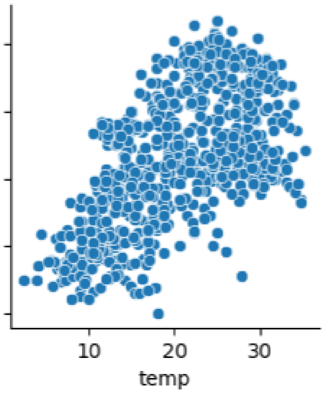


2. Why is it important to use drop_first=True during dummy variable creation?

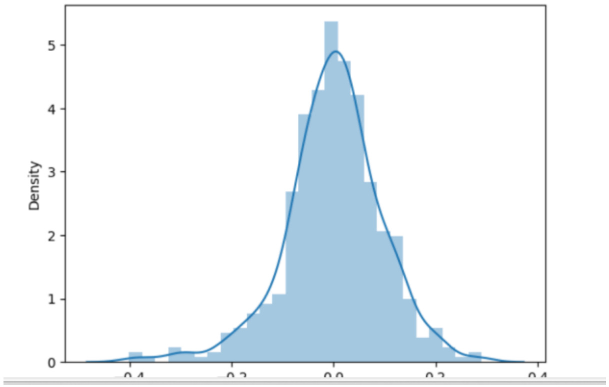
We need only n-1 dummies for n variables. This way we can reduce the number of variables, algorithm should train and test on.

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

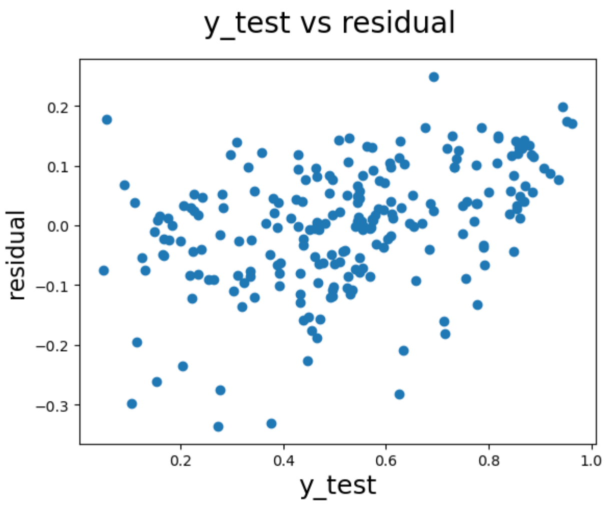
Temp and cnt has the highest correlation



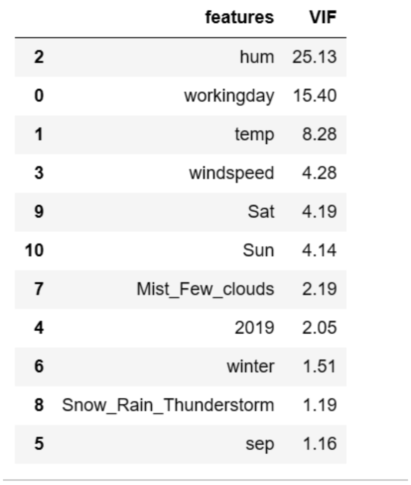
4. **How did you validate the assumptions of Linear Regression after building the model on the training set**?
    a. Linear relationship between temp and cnt as per pair plot
    b. Error terms are normally distributed



    c. Error terms have a constant variance(Homoscedasticity)



    d. Little multicorrelinearity(VIF<5 desired, but as per their P values, took them into consideration while building the model)



	features	VIF
2	hum	25.13
0	workingday	15.40
1	temp	8.28
3	windspeed	4.28
9	Sat	4.19
10	Sun	4.14
7	Mist_Few_clouds	2.19
4	2019	2.05
6	winter	1.51
8	Snow_Rain_Thunderstorm	1.19
5	sep	1.16

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
Clear, working day and fall contributes the demand of bikes.

General Subjective Questions

6. Explain the linear regression algorithm in detail.

Basic idea of linear regression algorithm is that independent(X) and target(Y) variable are linearly related where if we plot them, it will give us a straight line.

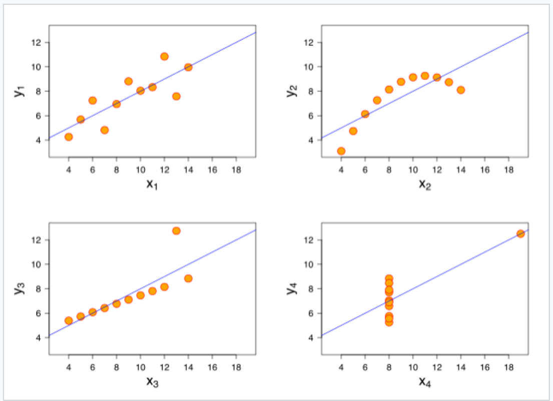
It can be written as Y=B0+B1X1+B2X2+B3X3+…

Steps:

1. Reading and understanding the data
2. Visualising the data
3. Prepare the data for modelling(split)
4. Training the model
5. Residual analysis
6. Predictions and evaluations on the test set

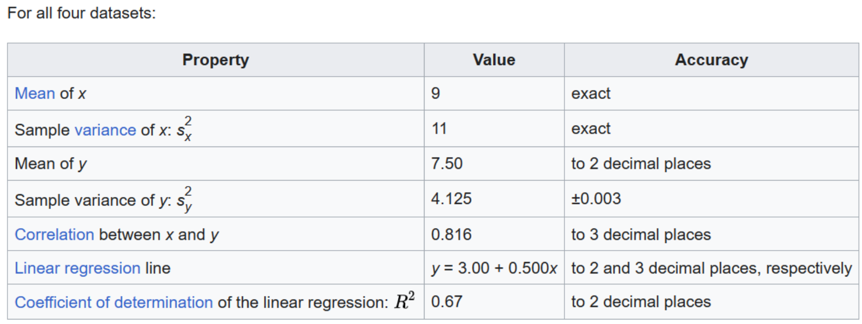
5. Explain the Anscombe's quartet in detail.

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.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Image ref: Wikipedia



For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

It is important to plot the data to pick the right model for building.

7. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. 0 to -1: Negative,0: No correlation,0 to 1: Positive correlation

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

Scaling is the process of bringing all variables in comparable measuring scale. This is needed specially to avoid coefficients going to extreme end, that leads to difficulty in interpretation of model. .

Normalized scaling or Min Max scaling tries to fit data in [0 and 1] scale by doing

(x-xmin)/(xmax-xmin)

Standardized scaling scales value in such a way that mean lies at 0. It is computed by

(x-mean(x))/standard deviation(x)

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

VIF=1/(1-R2)

Where R2=1, VIF becomes infinite. That means variables are highly correlated.

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

A 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 distribution1. It is also used to determine if two data sets come from populations with a common distribution1. Q-Q plots are useful in linear regression to determine if residuals follow a normal distribution, which is an assumption in regression2. Q-Q plots summarize any distribution visually