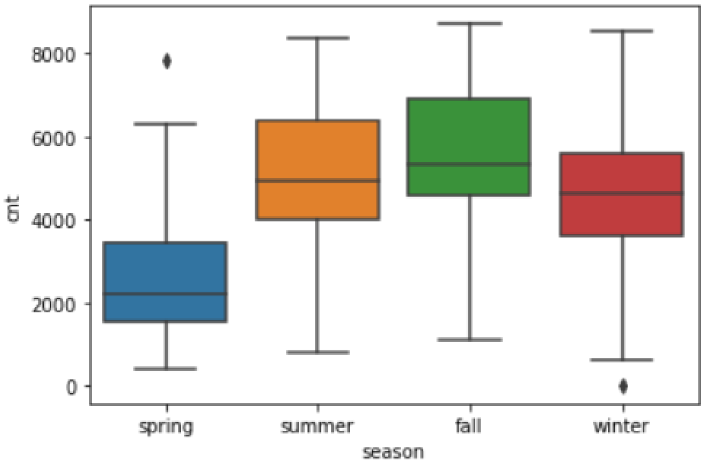
Assignment-based Subjective Questions

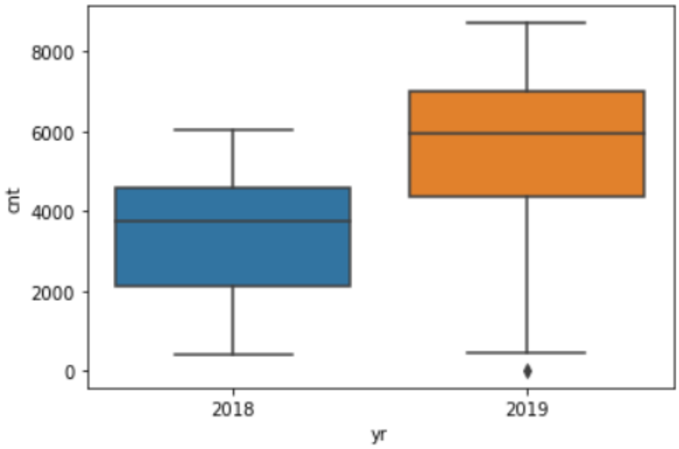
1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

For EDA of the categorical columns, the unique values of the column were found and then were replaced with the terms advised in the data dictionary file. The most important categorical features that are contributing to the target variable, i.e. 'cnt', are shown in the figures below. It can be observed that moving from one category to another category within the column shows changes in the rental bike count (cnt).

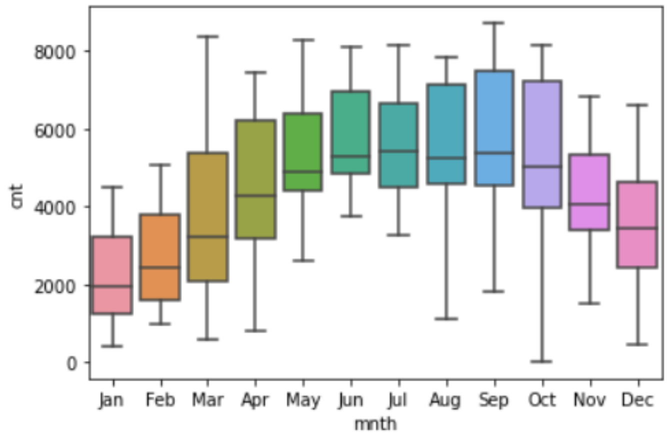
'cnt' vs 'season'



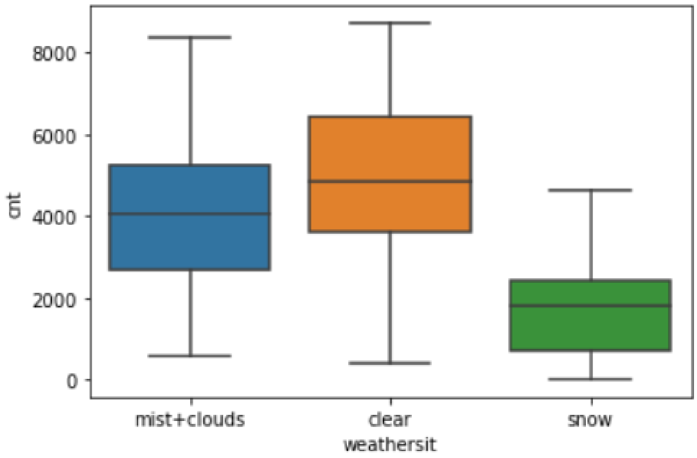
'cnt' vs 'yr'



'cnt' vs 'mnth'



'cnt' vs 'weathersit'



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

To avoid redundancy. E.g. Table 1 can also be represented with table 2, since bathroom = 0, and Store = 0, can represent Bedroom = 1. Therefore, in Table 1, the Bedroom column is redundant.

Table 1

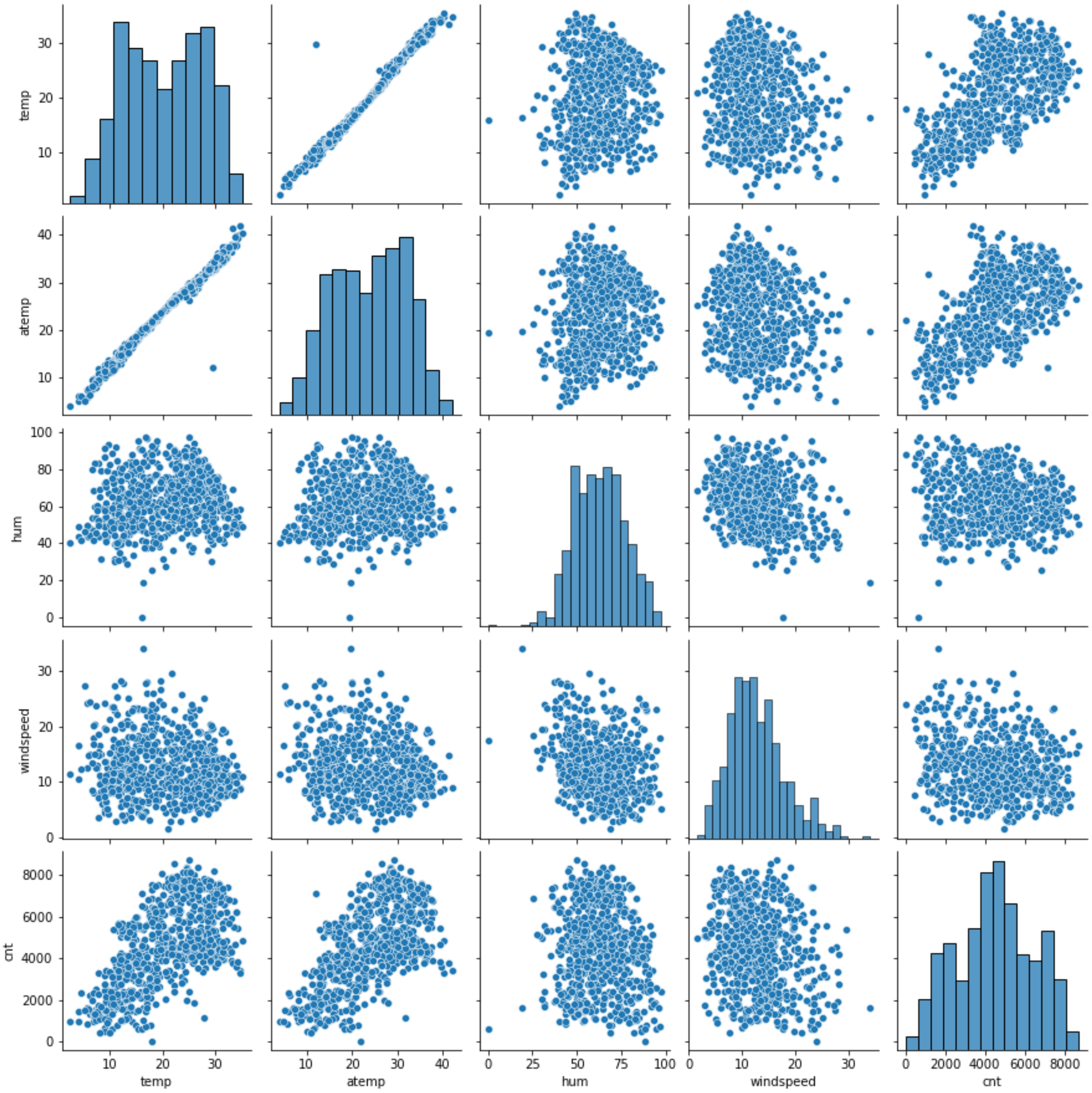
	Bedroom	Bathroom	Store
Bedroom	1	0	0
Bathroom	0	1	0
Store	0	0	1

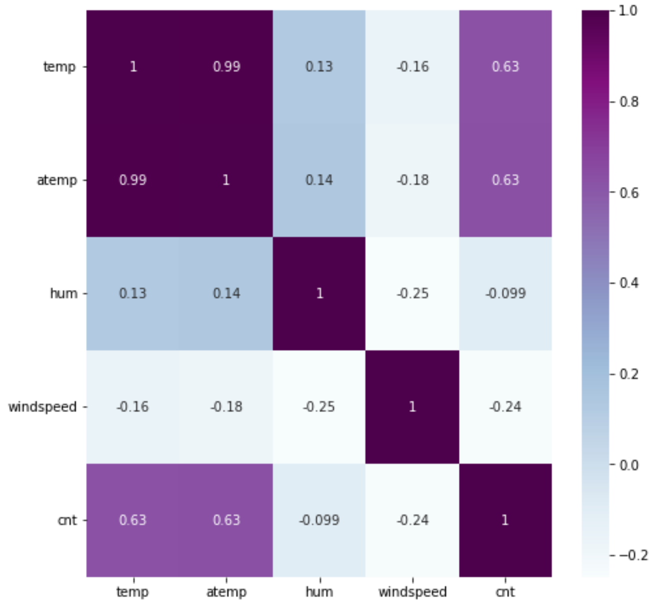
Table 2

	Bathroom	Store
Bedroom	0	0
Bathroom	1	0
Store	0	1

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

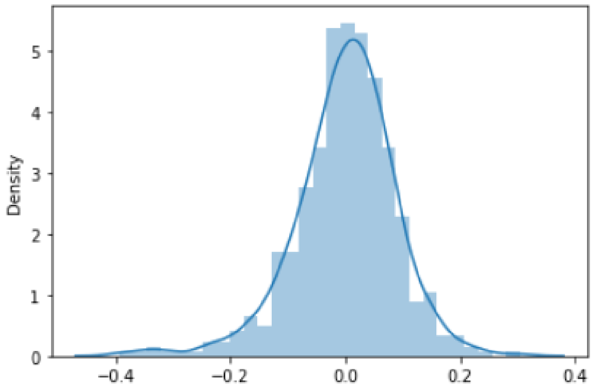
According to the pair-plots and the heatmap (of correlations) below, the 'temp' and 'atemp' have the highest correlation with the target variable, 'cnt', since cnt increases quite linearly with increase in temp and atemp.

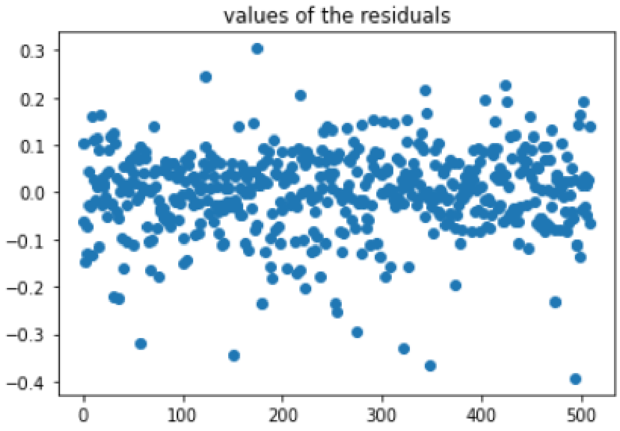




4- How did you validate the assumptions of Linear Regression after building the model on the training set?

By residual analysis below:



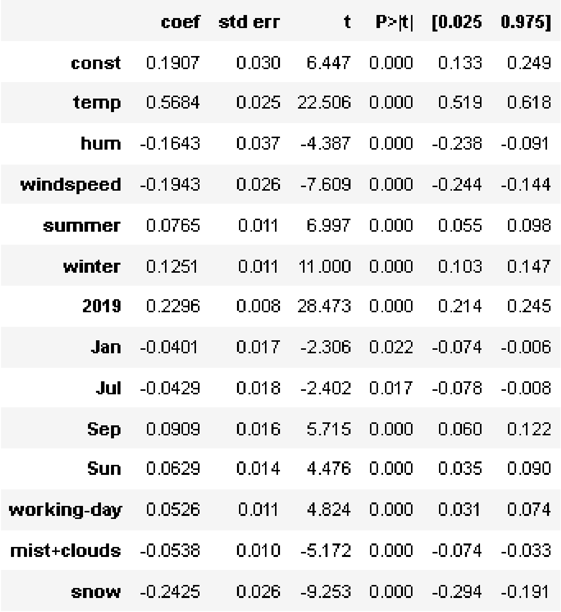


Analysis of the histogram and the scatter plot of the residuals reveals that:

- the residuals follow a normal distribution with a mean of zero, as expected as one of the conditions of the MLR assumptions.
- the residuals are scattered around 0, and are independent of each other.
- the residuals have constant variance (homoscedasticity).

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

Based on the coefficients estimated by the MLR model, the top 3 predictors for share bike rentals are: temperature (temp), snow conditions (snow) and the year 2019.



	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1907	0.030	6.447	0.000	0.133	0.249
temp	0.5684	0.025	22.506	0.000	0.519	0.618
hum	-0.1643	0.037	-4.387	0.000	-0.238	-0.091
windspeed	-0.1943	0.026	-7.609	0.000	-0.244	-0.144
summer	0.0765	0.011	6.997	0.000	0.055	0.098
winter	0.1251	0.011	11.000	0.000	0.103	0.147
2019	0.2296	0.008	28.473	0.000	0.214	0.245
Jan	-0.0401	0.017	-2.306	0.022	-0.074	-0.006
Jul	-0.0429	0.018	-2.402	0.017	-0.078	-0.008
Sep	0.0909	0.016	5.715	0.000	0.060	0.122
Sun	0.0629	0.014	4.476	0.000	0.035	0.090
working-day	0.0526	0.011	4.824	0.000	0.031	0.074
mist+clouds	-0.0538	0.010	-5.172	0.000	-0.074	-0.033
snow	-0.2425	0.026	-9.253	0.000	-0.294	-0.191

General Subjective Questions

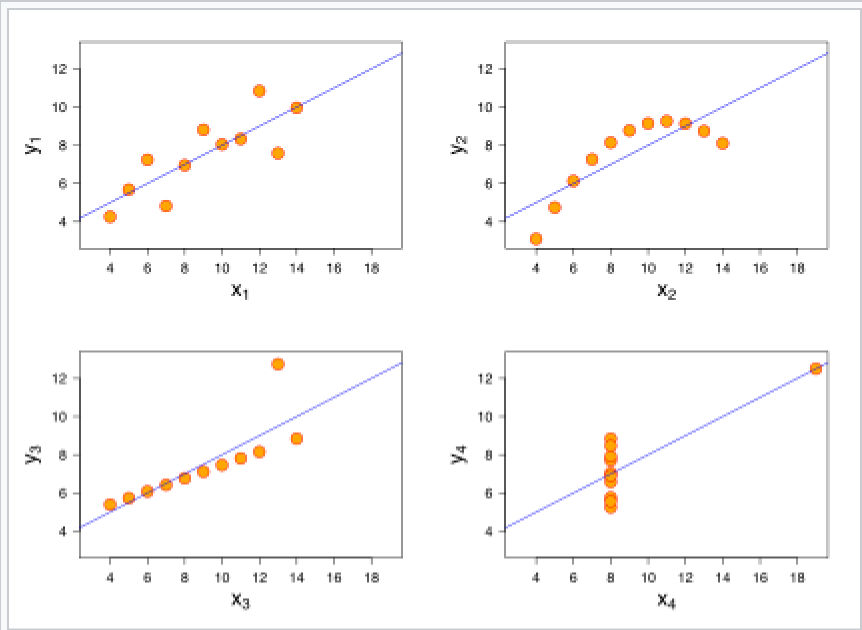
1- Explain the linear regression algorithm in detail.

The problem is about fitting a line (in SLR) or a hyper plane (in MLR) to best represent the data. The line/hyper plane are fitted based on the lowest value of the residual sum of squares. This is based on the assumptions below:

- there is some linear correlation between the target variable and few of the features.
- The residuals are normally distributed
- The residuals are independent of each other
- The residuals have constant variance.

2- Explain the Anscombe's quartet in detail.

The famous four data sets that have nearly identical statistics, i.e. mean of x, mean of y, variance of x, variance of y, linear regression line, and correlation between x and y. However, when they are plotted, they appear totally different in demonstration, as shown in Figure below[1].



3- What is Pearson's R?

Pearson's correlation coefficient (also known as Pearson's R) is a measure of linear correlation between two sets of data, and is calculated by dividing the covariance of the two variables by the product of their standard deviation

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

Scaling is about mapping a variable to a different range for some reason.

Why scaling:

- Better interpreting the results of the model
- Quicker convergence of the model

Normalised scaling: to map the feature to the interval between 0 and 1 (normally done for deep neural networks.

Standard scaling: to have the feature scaled so that it is centred around 0 with standard deviation of 1. For doing so, the values are subtracted by mean, and divide by the standard deviation.

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

[1] Image taken from: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

$$VIF = \frac{1}{1 - R^2}$$

Based on the above formula for VIF, when the R squared approaches one, then VIF becomes infinite.

Now, approaching of the R squared to one, means that the variable under study is perfectly linearly is estimated (explained) by the rest of the independent variables. Meaning that: that variable should definitely be dropped.

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

Quantile-Quantile (Q-Q) plot is a plot of the quantiles of two data sets against each other.

This is useful in determining if a distribution is related or similar to known distributions like normal, exponential, or uniform distributions.

For the case of linear regression, imagine we've been given one data set for training, and one for testing. Using q-q plot we can determine if these 2 data sets are coming from the same distribution or not.