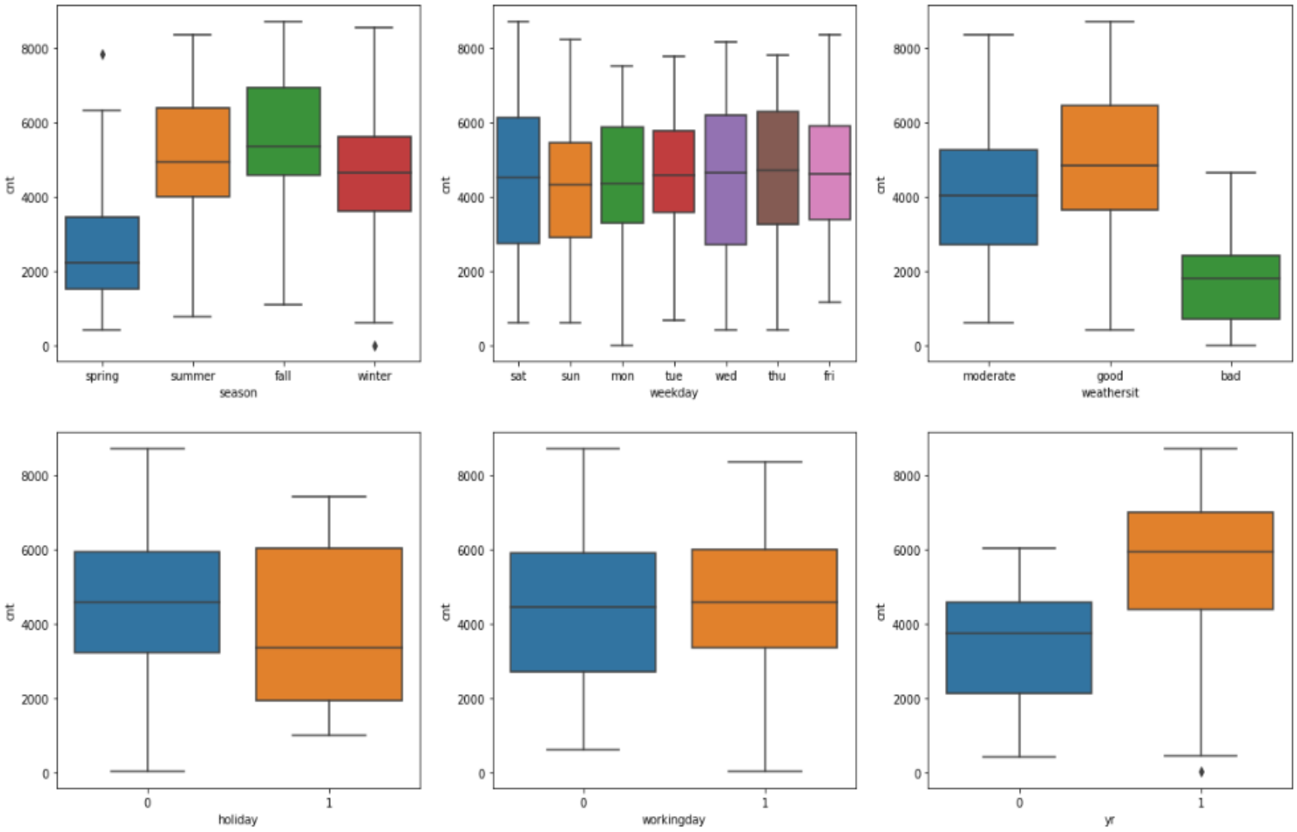
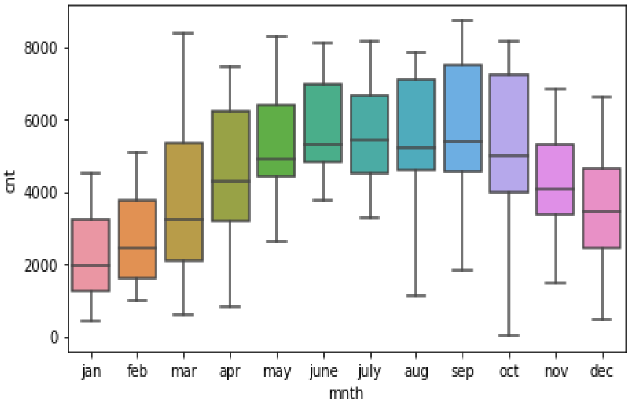
Assignment-based Subjective Questions

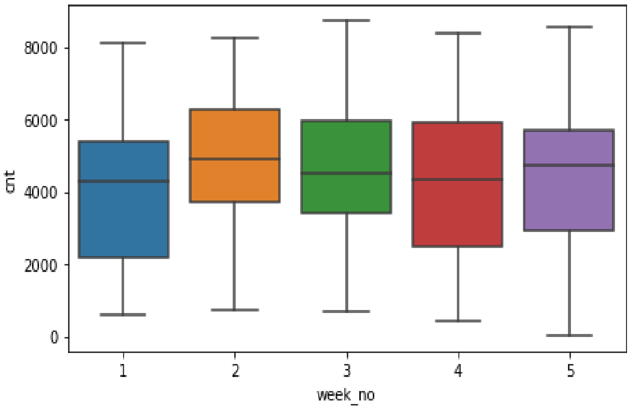
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

According to the analysis of the categorical variables from the dataset, following inferences were made about their effect on the dependent variable:-

- **season** seems to be a good factor, since it varies with target column
- **weekday** doesn't seem to be a good factor, since there is much less variance with target column
- **weathersit** seems to be a good factor, but no data is observed for weathersit_4 i.e. weathersit_severe label
- **holiday** seems to be a moderate factor
- **workingday** also seems to be a moderate factor
- **yr** seems to be a very good factor, where 0 indicates 2018 and 1 indicates 2019
- **mnth** seems to be a very good factor, especially, apr to oct months
- **week_no** doesn't seem to be a good factor







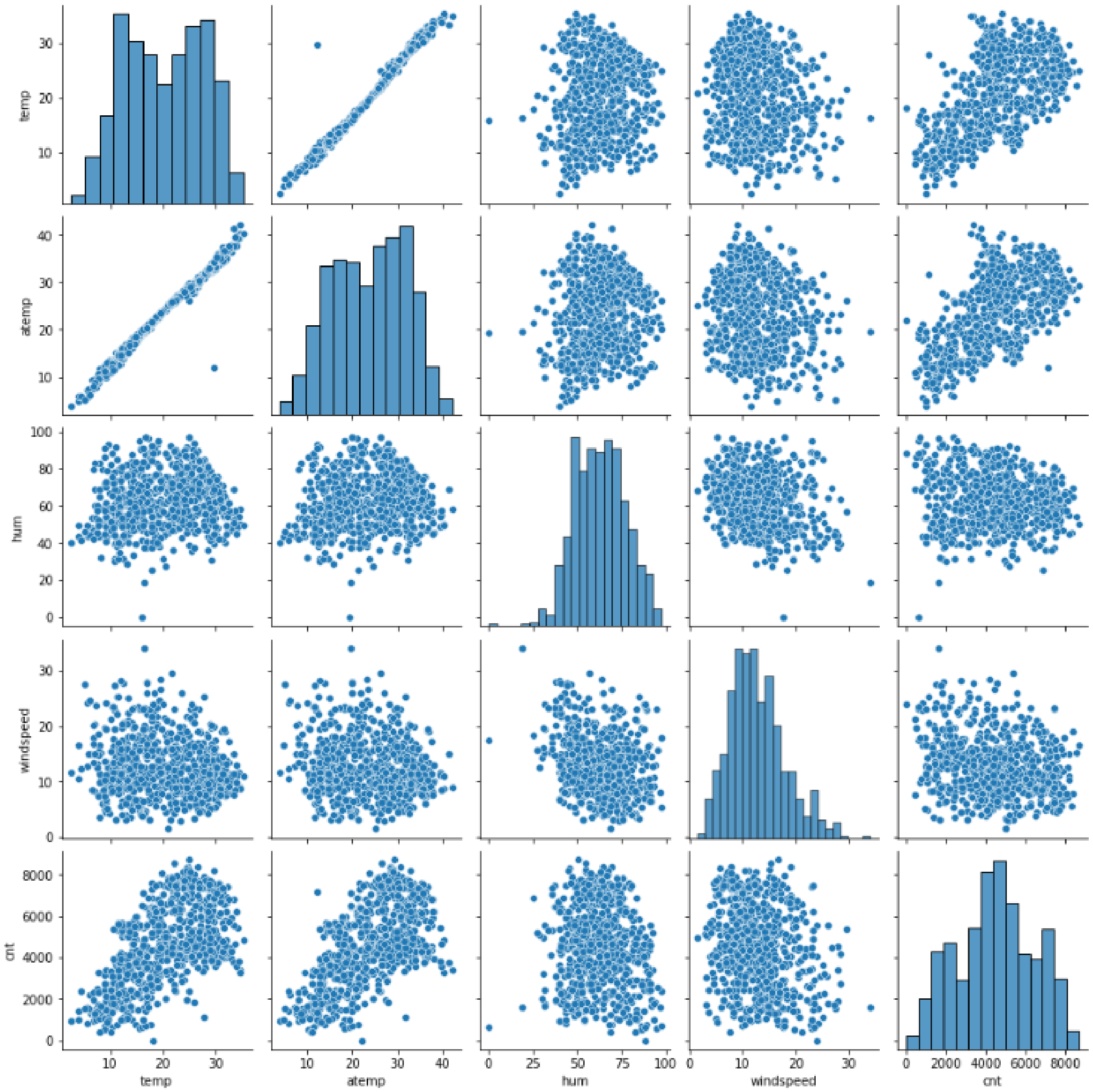
Q2. Why is it important to use drop_first=True during dummy variable creation?

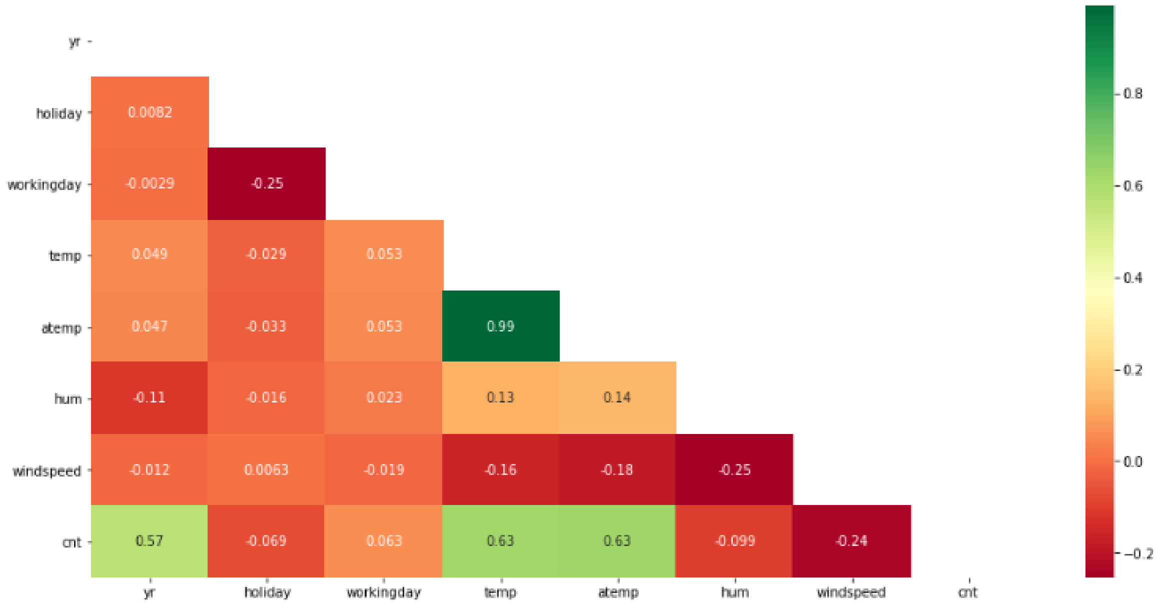
By dropping one of the one-hot encoded columns from each categorical feature, we ensure there are no reference columns—the remaining columns become linearly independent. Therefore, when using the normal equation to create an OLS model, you must drop one of the one-hot encoded columns from each categorical feature.

drop_first=True does the job of dropping one of the columns while creating one-hot encoded columns automatically for us, at the same time. It indicates whether to get k-1 dummies out of k categorical levels by removing the first level.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

“temp” and “atemp” have the highest correlation with the target variable, as per pair-plot among the numerical variables. But both are highly correlated to each other as well. Thus, we need to drop one of them.

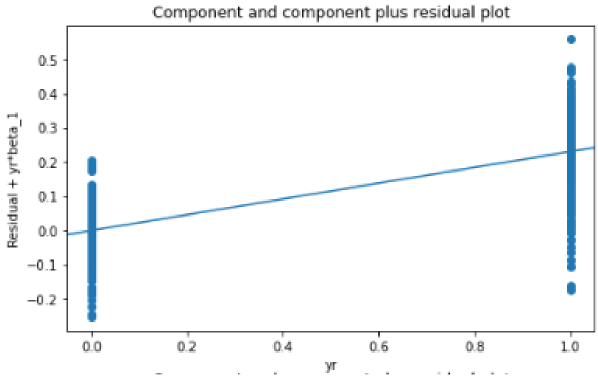

temp
atemp
hum
windspeed
cnt
10
20
30
40
0
25
50
75
100
2000
4000
6000
8000

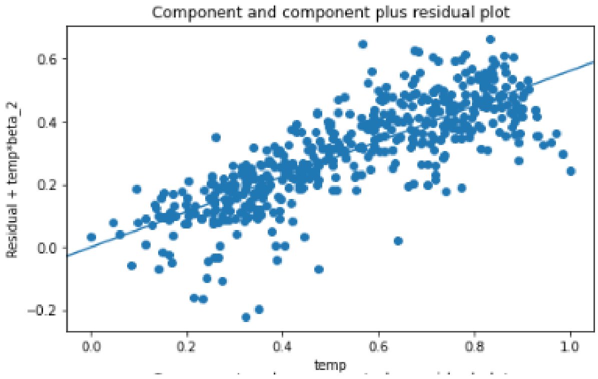

yr
holiday
workingday
temp
atemp
hum
windspeed
cnt
0.0082
-0.0029
-0.25
0.049
-0.029
0.053
0.047
-0.033
0.053
0.99
-0.11
-0.016
0.023
0.13
0.14
-0.012
0.0063
-0.019
-0.16
-0.18
-0.25
0.57
-0.069
0.063
0.63
0.63
-0.099
-0.24
0.8
0.6
0.4
0.2
0.0
−0.2

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

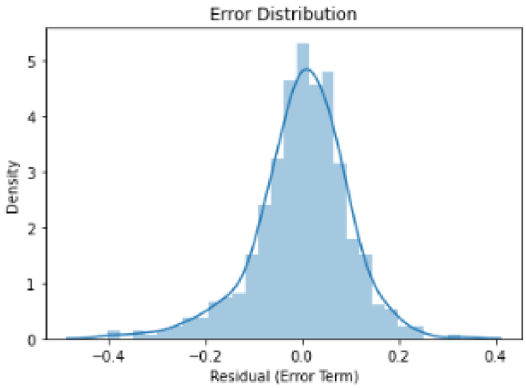
The assumptions of Linear Regression were validated after building the model on the training set as follows:-

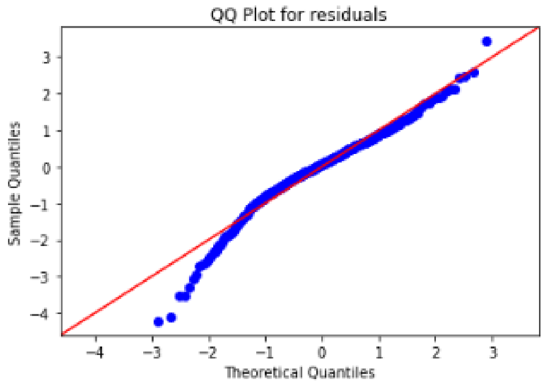
1) <u>Linear Relationship assumption</u>
    - By checking relationship of each feature with target variable using ccpr plot
    - For.eg, for yr (year) and temp (temperature)



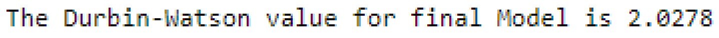


2) <u>Assumption of Normally Distributed Error Terms</u>
    - Tested using a histogram and a QQ Plot



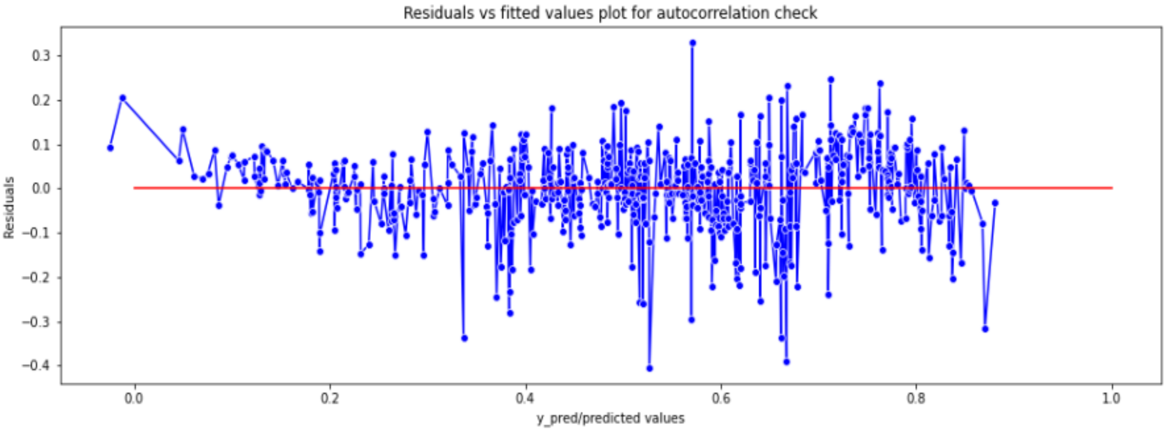


3) <u>Assumption of Error Terms Being Independent (Autocorrelation check)</u>
    - Independence of residuals means "absence of auto-correlation" Autocorrelation refers to the fact that observations' errors are correlated
    - a) <u>Durbin-Watson Test</u> - The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables (0–2: positive auto-correlation, 2–4: negative auto-correlation)



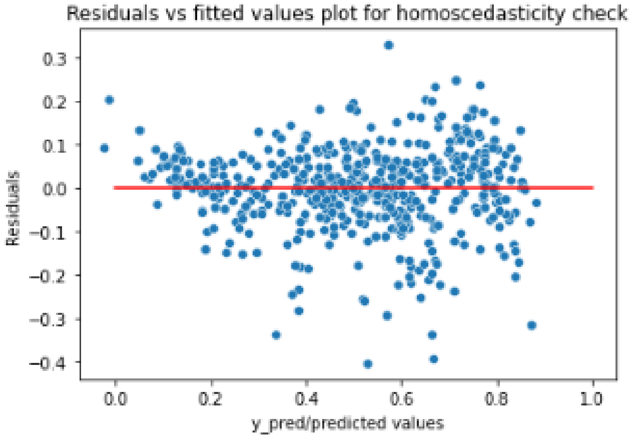
```
The Durbin-Watson value for final Model is 2.0278
```

- b) Residuals vs fitted values plot does not show any pattern



4) Homoscedasticity

- Homoscedasticity means that the residuals have equal or almost equal variance across the regression line. By plotting the error terms with predicted terms we can check that there should not be any pattern in the error terms.



- Goldfeld Quandt Test
    1. Null Hypothesis: Error terms are homoscedastic
    2. Alternative Hypothesis: Error terms are heteroscedastic.

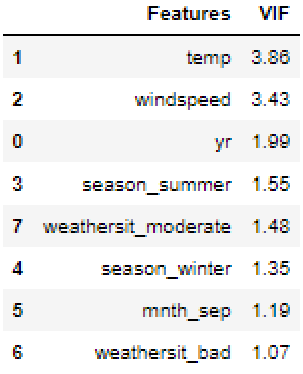
output-

```
[('F statistic', 1.2293168816942728), ('p-value', 0.06173211620657116)]
```

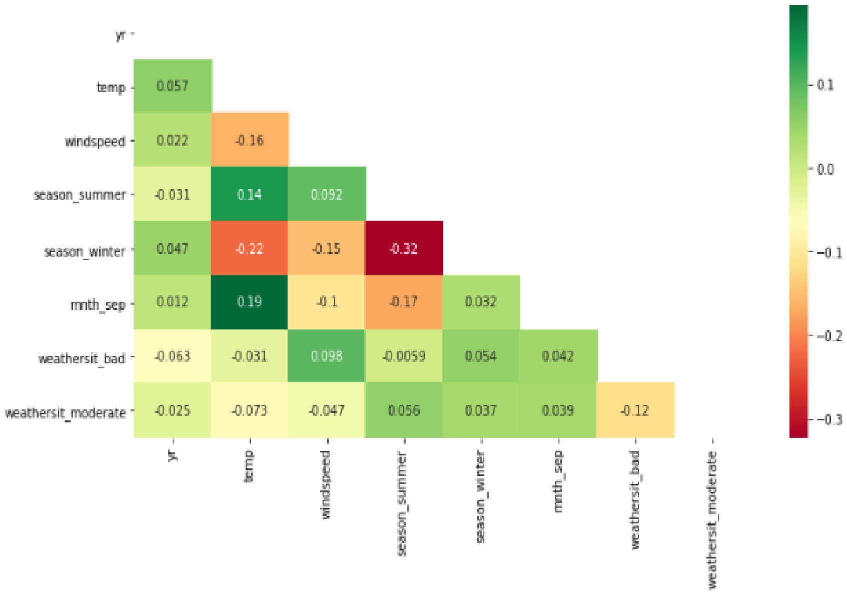
Since p value is more than 0.05, we can't reject it's null hypothesis that error terms are homoscedastic

5) Little or no Multicollinearity

- We can test for multicollinearity problems using the Variance Inflation Factor(VIF) or correlation matrix

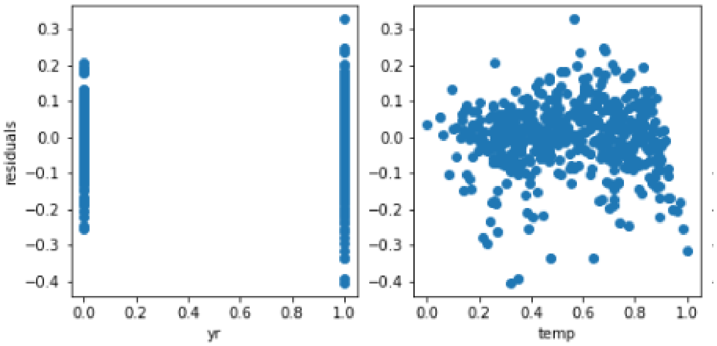


	Features	VIF
1	temp	3.86
2	windspeed	3.43
0	yr	1.99
3	season_summer	1.55
7	weathersit_moderate	1.48
4	season_winter	1.35
5	mnth_sep	1.19
6	weathersit_bad	1.07



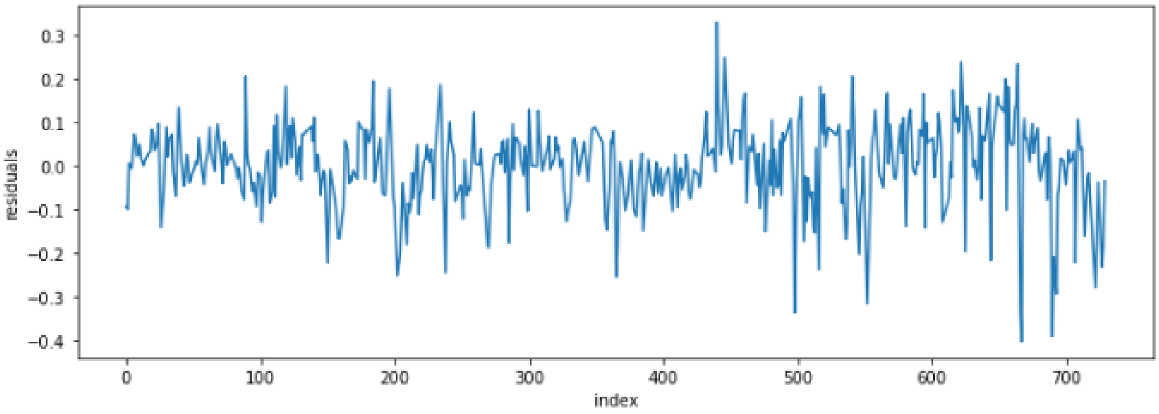
6) Independent variables are uncorrelated with the error term

- Tested using scatterplots. No patterns found.
- For.eg, for yr (year) and temp (temperature)



7) Observations of the error term are uncorrelated with each other

- This checks whether there is a correlation inside the observations of the error term. If this happens, then it means the assumption: the observations are drawn randomly, is violated. On plotting, residuals do not show any pattern.



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features contributing significantly towards explaining the demand of the shared bikes based on the final model are as follows:-

1) temp - Temperature (0.5616)
2) yr - Year (0.2307)
3) weathersit_bad - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (- 0.3023)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail

Linear Regression is a very basic form of supervised machine Learning algorithms, mostly used for regression tasks and forecasting.

Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It is used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.

Basically, Linear regression is used to find the best linear-fit relationship on any given data, between independent and dependent variables. This is usually done by minimising the cost function by using optimization techniques. Mostly, the cost function is RSS (Residual Sum of Squares) and optimization is done through Gradient Descent (or its variants).

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

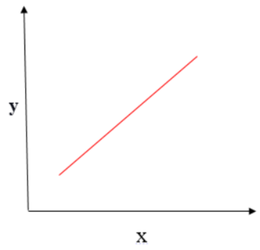
$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.
b = Slope of the line.
a = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset

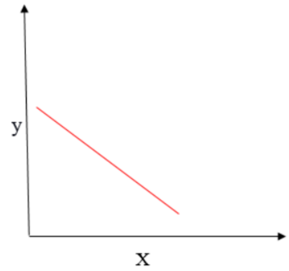
A linear regression line can have a Positive Linear Relationship or a Negative Linear Relationship.

1) If the dependent variable expands on the Y-axis and the independent variable progresses on the X-axis, then such a relationship is called a Positive linear relationship.



y = a + bx

2) If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called as a negative linear relationship



y = a - bx

Basically, there are two types of linear regression:-

1) Simple Linear Regression
    - It has one target value and only one predictor value
2) Multiple Linear Regression
    - It is an extension to simple linear regression
    - It can have more than one predictor values

There are some assumptions of Linear regression model which are as follows:-

1) Linear Assumption
    - There is a linear relationship between the dependent and independent variables
2) Assumptions about the residuals
    a) Normality assumption:
        - The error terms, ε(i), are normally distributed.
    b) Zero mean assumption:
        - The residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
    c) Constant variance assumption:
        - The residual terms have the same (but unknown) variance, σ2 . It is also known as the assumption of homogeneity or homoscedasticity.
    d) Independent error assumption:
        - The residual terms are independent of each other, i.e. their pairwise covariance is zero.

3) Assumptions about the estimators:

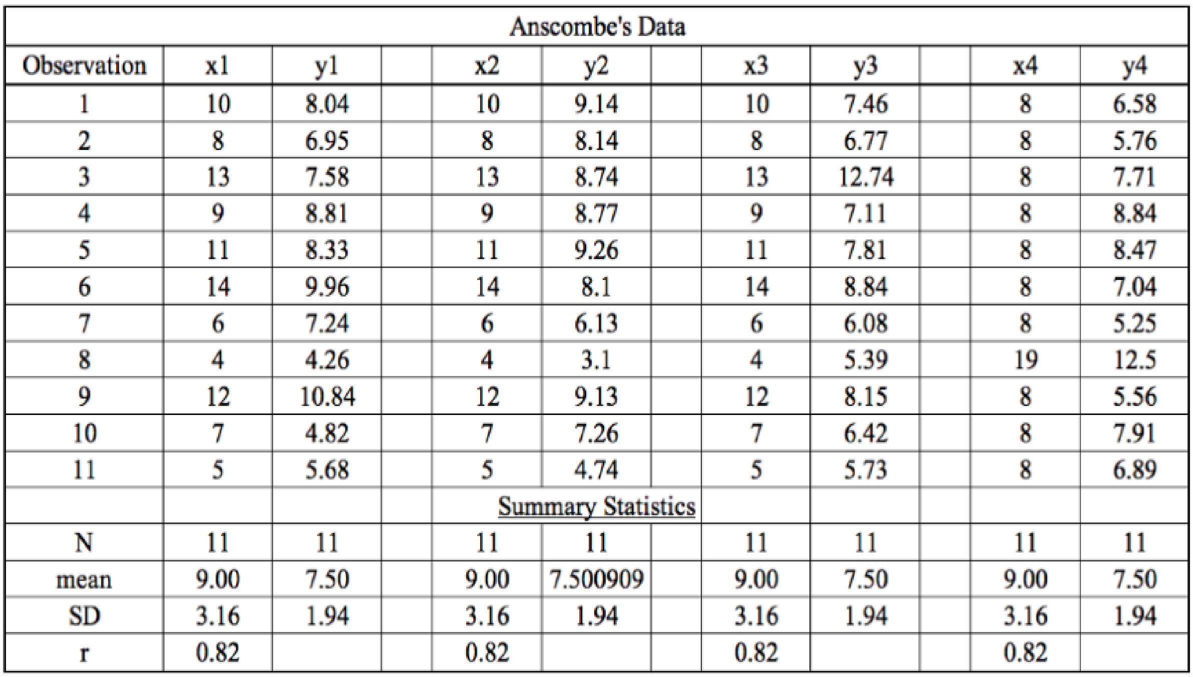
- The independent variables are measured without error.
- The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data. (Applies only in the case of Multiple Linear Regression)

Q2. Explain the Anscombe's quartet in detail.

Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimise against and use as a barometer for our business. But there's a danger in relying only on summary statistics and ignoring the overall distribution.

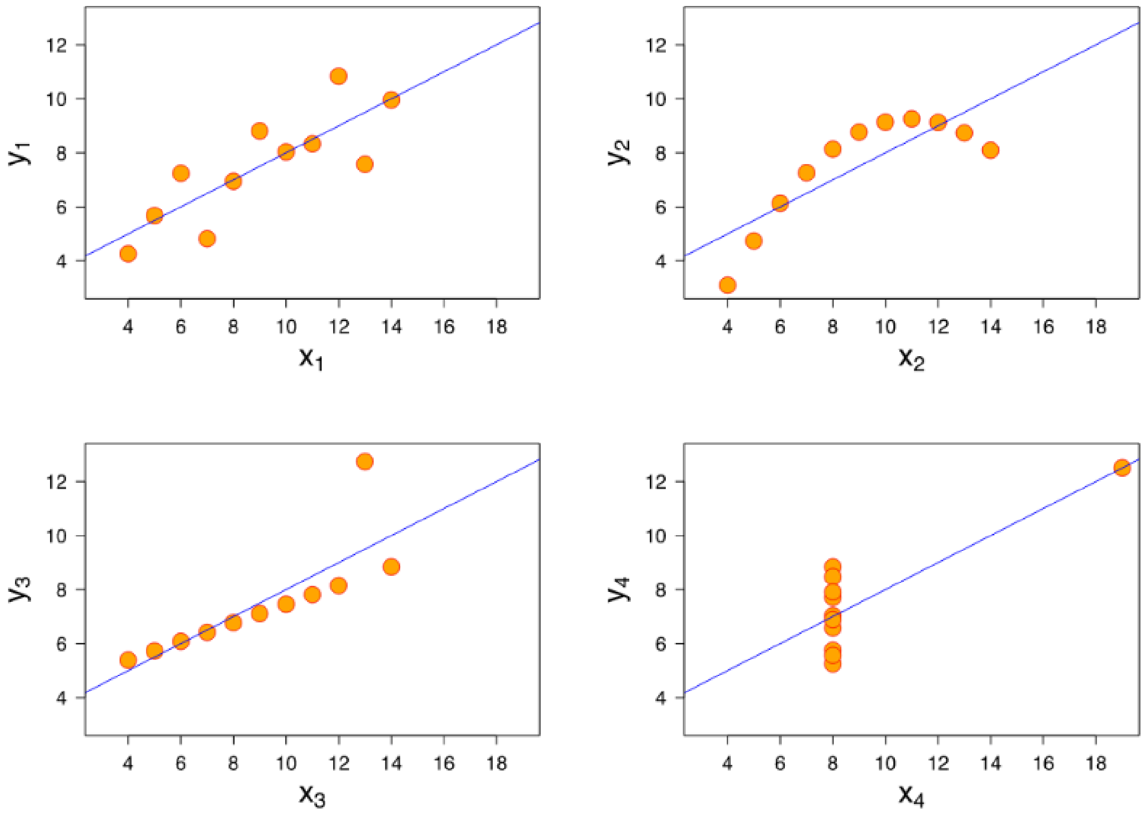
Anscombe's Quartet is the most elegant demonstration of the dangers of summary statistics. It can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Anscombe's Quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. Let's understand these from following data consisting of 4 datasets:-



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

So far these four datasets appear to be pretty similar. But when we plot these four datasets on an x/y coordinate plane, we get the following results:



The four datasets can be described as:

Dataset 1:

- This fits the linear regression model pretty well.

Dataset 2:

- This could not fit the linear regression model on the data quite well as the data is non-linear.

Dataset 3:

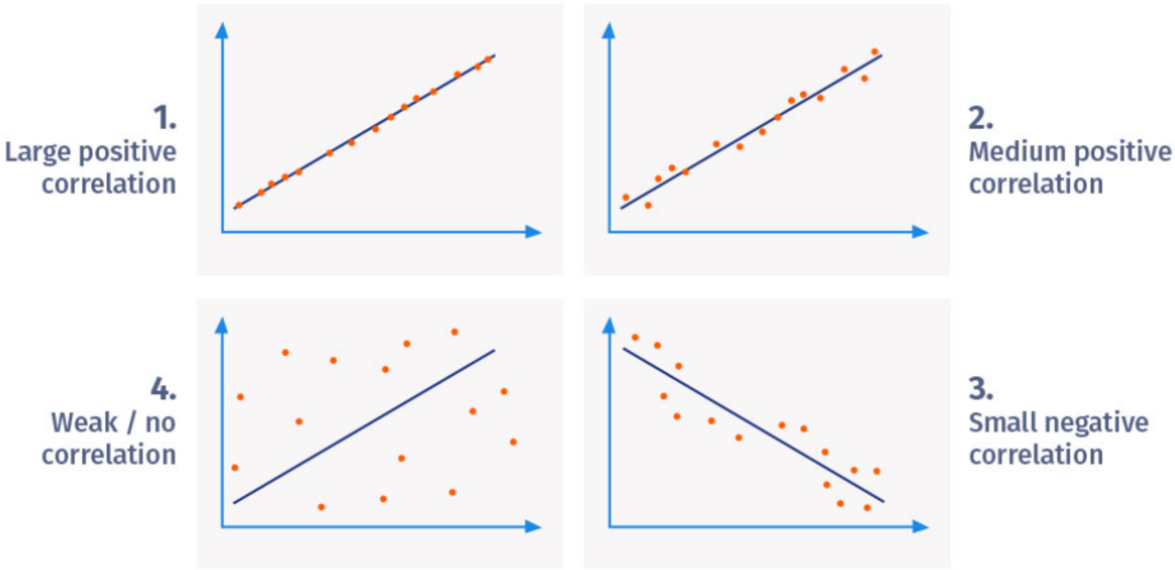
- This shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4:

- This shows the outliers involved in the dataset which cannot be handled by linear regression model

Q3. What is Pearson's R?

Pearson's correlation coefficient (or Pearson's R), developed by Karl Pearson, is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.



Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

There are certain requirements for Pearson's Correlation Coefficient though:

1) Scale of measurement should be interval or ratio
2) Variables should be approximately normally distributed
3) The association should be linear
4) There should be no outliers in the data

The formula for Pearson's R is as follows:-

$$\frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where :

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

For a sample, it is further evaluated as follows:-

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Rearranging, we get

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above, and s_x, s_y are defined below
- $\left(\frac{x_i - \bar{x}}{s_x}\right)$ is the standard score (and analogously for the standard score of y)

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

a) Scaling is a step of data pre-processing which is applied to independent variables to bring the data within a particular range. It also helps in speeding up the calculations in an algorithm.

b) Most of the time, the collected data set contains features highly varying in magnitudes, units and range. So, the statistics of models like t-statistics, f-statistic, p-value, R-squared, etc doesn't change, but the coefficients are affected.

So, basically, scaling is performed for the following two reasons:-

1) Ease of interpretation
2) Faster convergence of gradient descent methods, i.e. for faster optimization

c) The difference between normalized scaling and standardized scaling is as follows:-

Sr. No.	**Normalized Scaling**	**Standardized Scaling**
1	Minimum and maximum values of the features are used for scaling	Mean and standard deviation is used for scaling
2	It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales value in range 0-1	No limits on the range
4	$x = \frac{x - min(x)}{max(x) - min(x)}$	$\frac{x - mean(x)}{sd(x)}$
5	*sklearn.preprocessing.MinMaxScaler* helps to implement normalization in python	*sklearn.preprocessing.StandardScaler* helps to implement standardization in python
6	It is affected by outliers	It is much less affected by outliers
7	It loses some information in the data, especially about outliers	It does not lose such information i the data
8	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian
9	It is also known as min-max normalization or min-max scaling	It is also known as z-score Normalization

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model. A large variance inflation factor (VIF) on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

The value of VIF is calculated by the formula given below:

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

where:

i refers to the ith variable

Suppose, R-squared value is equal to 1, then the denominator equals 0. And mathematically, any value divided by 0 is infinite. Hence, when the R-squared value is equal to 1, the value of VIF becomes infinite.

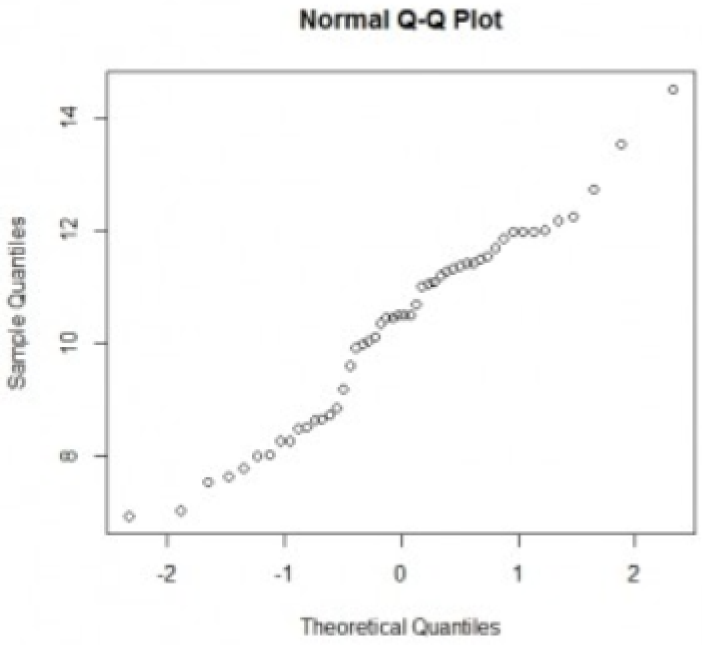
Now, what is meant by these is that there is a perfect correlation and the corresponding variable may be expressed exactly by a linear combination of other variables.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot, or quantile-quantile 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. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

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 and importance of a Q-Q plot in linear regression :-

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.

It is used to check if two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

Advantages —

a) It can also be used with sample sizes.
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

.

In Python, *statsmodels.api* provides qqplot and qqplot_2samples to plot Q-Q graphs for single and two different data sets respectively.