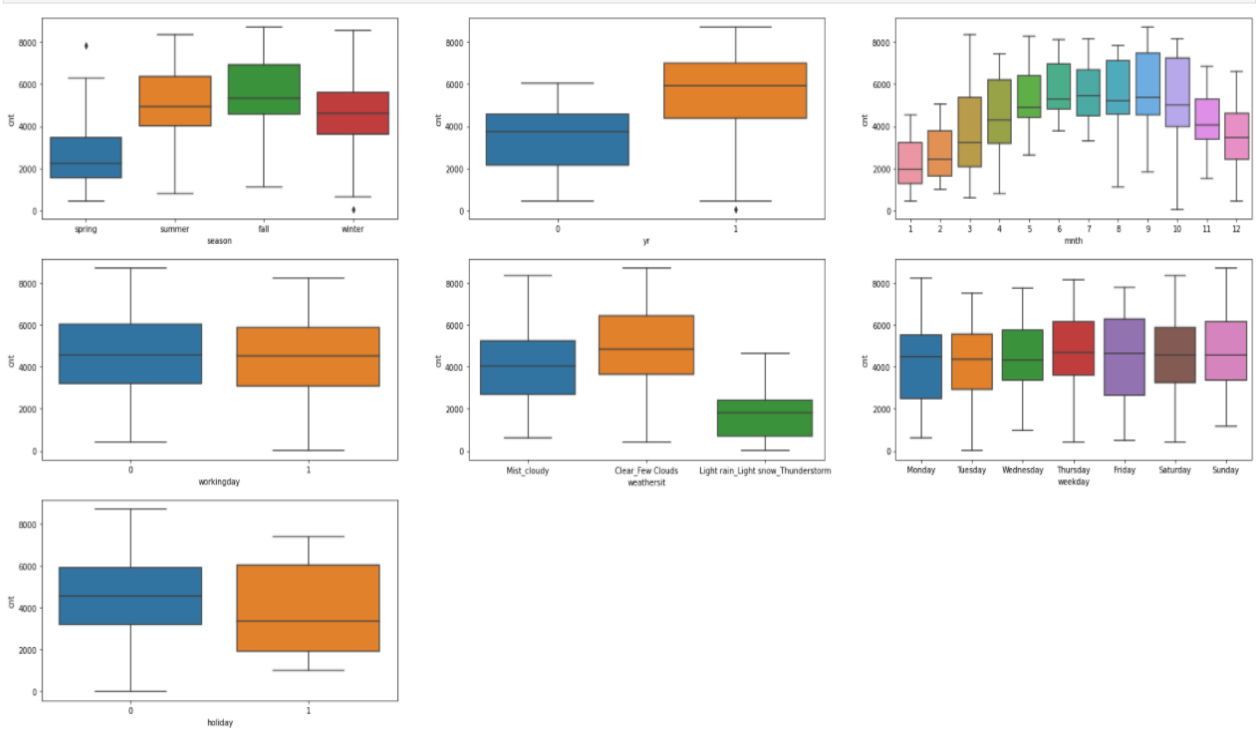
Assignment-based Subjective Questions:

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

For analysing the categorial variables from the dataset, used boxplot for all the categorial variables in 1 plot using subplots.



1. season:

Almost 32% of the bike booking were happening in season3 (fall) with a median of over 5000 booking (for the period of 2 years).

This is followed by season2 & season4 with 27% & 25% of total booking.

season can be a good predictor for the dependent variable cnt.

2. mnth:

Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month.

mnth has some trend for bookings and can be a good predictor for the dependent variable cnt.

3. weathersit:

Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking.

weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable cnt.

4. holiday:

Almost 97.6% of the bike booking were happening when it is not a holiday.

holiday cannot be a good predictor for the dependent variable cnt.

5. weekday:

weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings.

weekday can have some or no influence towards the predictor cnt.

6.workingday:

Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years).

workingday can be a good predictor for the dependent variable cnt.

7.Yr :

2019 has high average bookings compare to 2018 with median around 6000 (almost 80% of bookings if we consider 2 years)

yr can be a good predictor for dependent variable cnt.

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

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.

Below are some points on this:

1. The other variable is able to explain the values for this first variable as well, so we drop this
2. Dummy variable trap manifests itself directly from one-hot encoding applied on categorial variables.The size of one hot-vectors is equal to the unique values that a categorial column takes up and each such vector contains exactly one '1' in it, this ingests Multicollinearity in the data set.
3. Drop_first=True will make sklearn use the default assignment for this variable which is False.

Here, in our assignment,we are creating dummy_variables for below variables:

- season
- weathersit
- weekday
- mnth

Explaining it with **season**

```
Season_condition=pd.get_dummies(df['season'])
```

```
Season_condition.head()
```

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

Here, after converting the variable to category applied the dummies on dataset

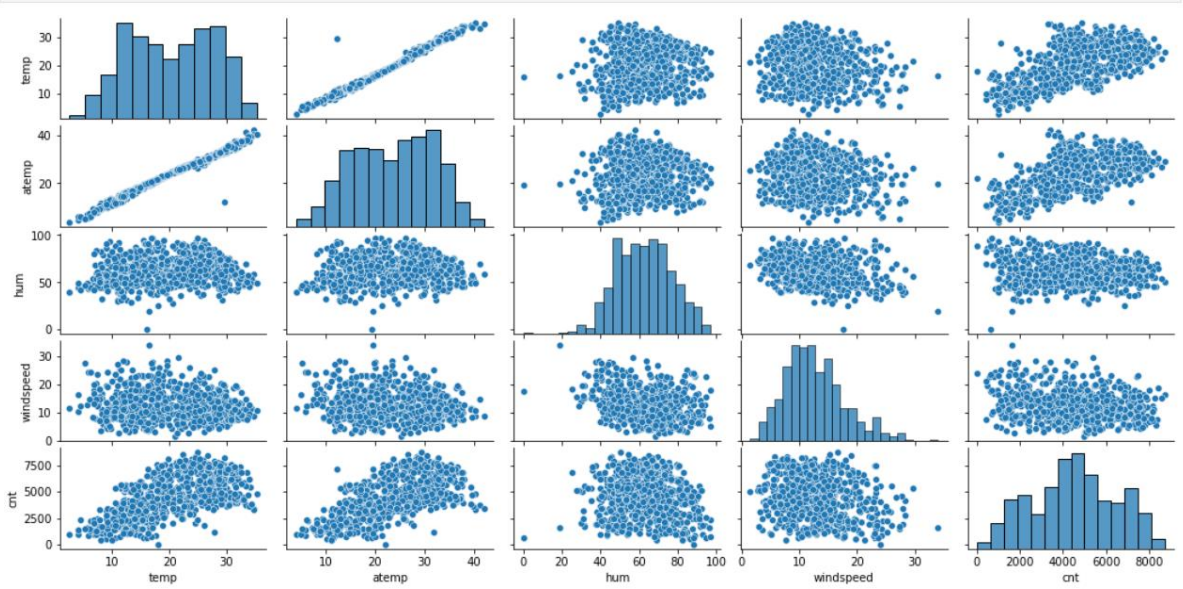
bike_new = pd.get_dummies(bike_new, drop_first=True)

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

temp and atemp has highest correlation with target variable ‘cnt’

atemp and temp are highly correlated among themselves and also with cnt

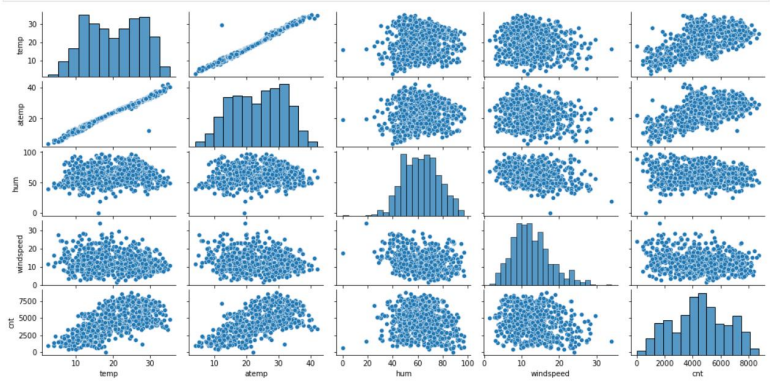
If we see the pairplot



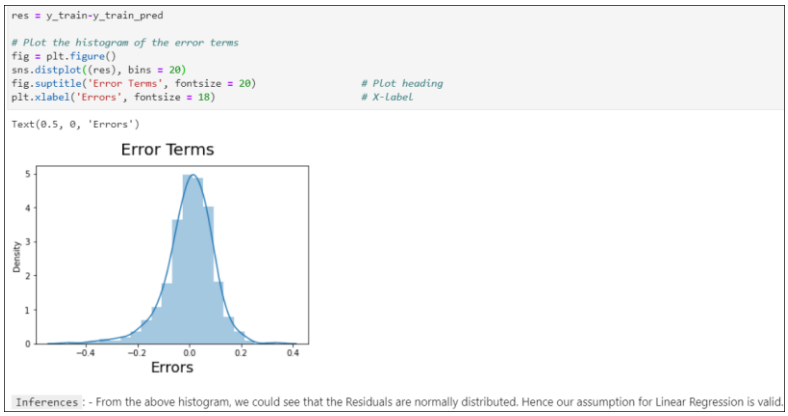
Q:4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Assumptions of Linear Regression:

1. There is a linear relationship between X and Y
   From the pairplot, we can see that there is linear relationship between temp and atemp variable with predictor cnt



2. Error terms are normally distributed with mean zero (not X,Y)



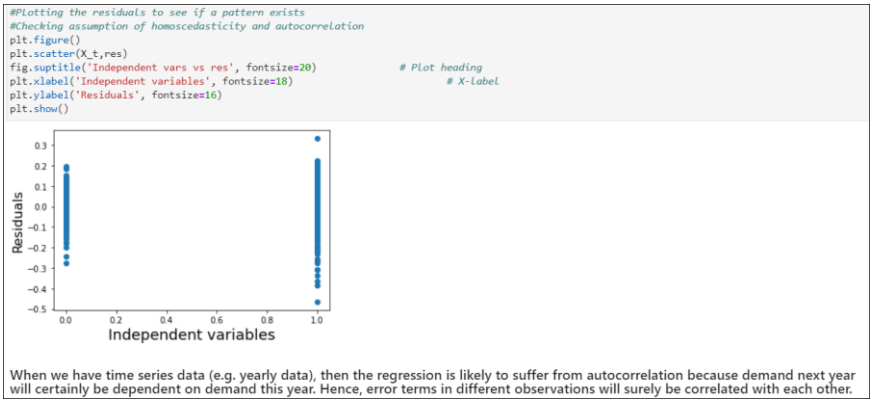
```
res = y_train-y_train_pred

# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((res), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

```
Text(0.5, 0, 'Errors')
```


Inferences : - From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

3. Error terms are independent of each other: Error terms should not be dependent on each other, like a time series data



```
#Plotting the residuals to see if a pattern exists
#Checking assumption of homoscedasticity and autocorrelation
plt.figure()
plt.scatter(X_t,res)
fig.suptitle('Independent vars vs res', fontsize=20)              # Plot heading
plt.xlabel('Independent variables', fontsize=18)                  # X-Label
plt.ylabel('Residuals', fontsize=16)
plt.show()
```

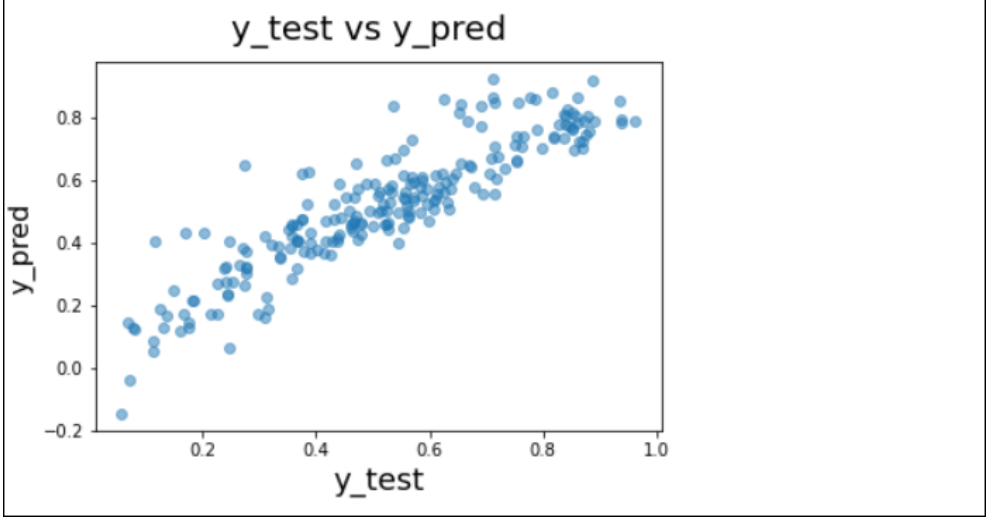

When we have time series data (e.g. yearly data), then the regression is likely to suffer from autocorrelation because demand next year will certainly be dependent on demand this year. Hence, error terms in different observations will surely be correlated with each other.

4. Error terms have constant variance is also seen.
5. Y_test and y_pred is coming approximately near and plotting scatter plot gives us overlapping data points which validates the model.

```
# Plotting y_test and y_pred to understand the spread
# import matplotlib.pyplot as plt
# import numpy as np


fig = plt.figure()
plt.scatter(y_test, y_pred, alpha=.5)
fig.suptitle('y_test vs y_pred', fontsize = 20)              # Plot heading
plt.xlabel('y_test', fontsize = 18)                          # X-Label
plt.ylabel('y_pred', fontsize = 16)
```

```
Text(0, 0.5, 'y_pred')
```



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

As per the final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of '**0.408372**' indicated that a unit increase in temp variable increases the bike hire numbers by 0.408372 units.
- **Weather Situation 3 (weathersit_Light Snow_Rain_Thunderstorm)** - A coefficient value of **'-0.319994'** indicates that a unit increase in weathersit_Light Snow_Rain_Thunderstorm variable decreases the bike hire numbers by 0.3070 units.
- Year (yr) - A coefficient value of '0.2336' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2336 units.

Recommendation for high demand is give these predictor variables more importance.
- The next best features that can also be considered are
- **season_winter** : - A coefficient value of **'0.032049'** indicates that, a unit increase in season_winter variable increases the bike hire numbers by 0.032049 units.

- **windspeed**: - A coefficient value of **'-0.140327'** indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.140327 units.

General Subjective Questions and Answers

Q1. Explain the linear regression algorithm in detail:

Ans: Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relations

Linear Regression, is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables.

Linear regression classification depending on number of variables:

1. **Simple Linear Regression** – Number of Independent variable is 1
2. **Multiple Linear Regression** -There are more than 1 independent variables

A **Simple linear regression** model attempts **to explain the relationship between a dependent variable and an independent one using a simple straight line** whose equation can be y=mx+c where m is the slope and c is the intercept.

The **independent variable** is also known as **predictor variable** and the **dependent variable** is also known as **target variable**.

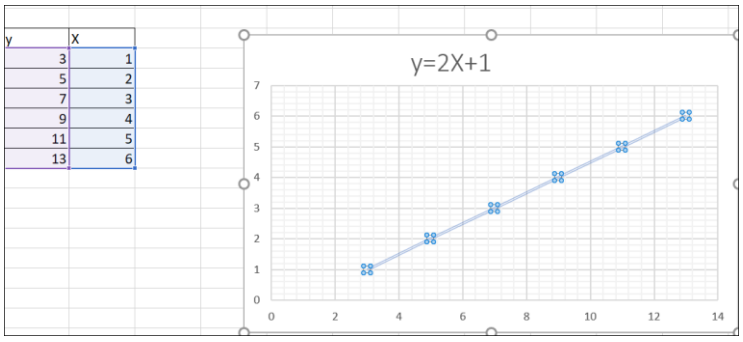
Example: Simple Linear Regression: We are trying to find Linear relationship between 2 variables using best fit line

We have x as independent variable and y as dependent variable. We plot the relationship between x and y using a scatter plot and below is the exact data points look like.

Looking at the data below, we can draw a basic linear equation between X and y

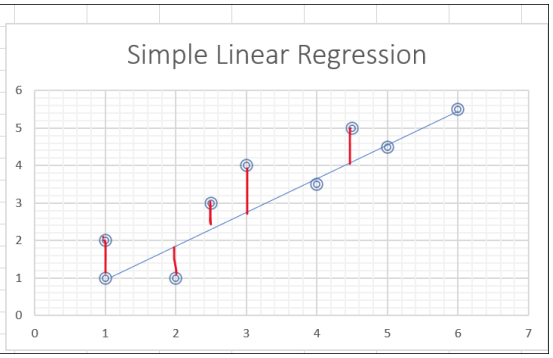
y= 2X+1 # This denotes Simple Linear equation which is also a best fit for all its data points.

It also depicts linear relationship between X and y



Now, suppose we have data (mostly the real time data) does not give the direct relationship so the data points and best fit line is not always as above. There can be multiple line which passes through all the data points but we consider the best fit line which has minimal error.

y	X
1	1
1	2
2	1
2.5	3
3	4
4	3.5
4.5	5
5	4.5
6	5.5



Equation of simple Linear Regression best fit line: $Y = \beta_0 + \beta_1 X$

Equation of simple Multiple Regression best fit line: $Y = \beta_0 + \beta_1 X + \beta 2X + \beta 3X + \ldots\ldots \beta nX$

Where β_1= slope and y-intercept = β_0

Slope: Slope or Gradient =change in y/change in x

Intercept: Value of y when X=0

Residuals (Measured Value – Predicted Value) difference between the yi and ypred values that is the actual data points and points which lies in the best fit predicted line.

These are also **errors** as it has some difference from actual data. These are also called as errors and

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

We square the error such that the summation is very less

y	X	Y-Fitted	error	error^2
1	1	1	0	0
1	2	2	1	1
2	1	1	-1	1
2.5	3	3	0.5	0.25
3	4	4	1	1
4	3.5	3.5	-0.5	0.25
4.5	5	5	0.5	0.25
5	4.5	4.5	-0.5	0.25
6	5.5	5.5	-0.5	0.25
				4.25

Residual Sum of Square(RSS) = Sum of all the Residual squares

For, every data point RSS is

$$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$$

Which can also be written as below:

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

As, we have Xi and Yi already for any data set for minimizing the RSS , we need to find the mimimal value of slope and coefficient

RSS is an absolute measure which changes if scale of a variable changes, so we need to define a relative measure which is also known as R-squared which is 1- RSS/TSS

$$R^2 = 1 - \frac{RSS}{TSS}$$

TSS (Total Sum of squares):

$$TSS = (Y_1 - \bar{Y})^2 + ...+ (Y_n - \bar{Y})^2$$

Or

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Physical significance of R2 can be seen as,when the data fits exactly fits the best fit R2=1 and when the data points does not fit the best bit line R2 becomes near to 0.

The strength of Linear regression model can be assessed using 2 metrics:

1. Coefficient of Determination or R2
2. Residual Standard Error (RSE)

RSE helps in measuring the lack of fit model on a given data. The closesness of the estimated regression coefficients to the true ones can be estimated using RSE. It is related to RSS by the formula

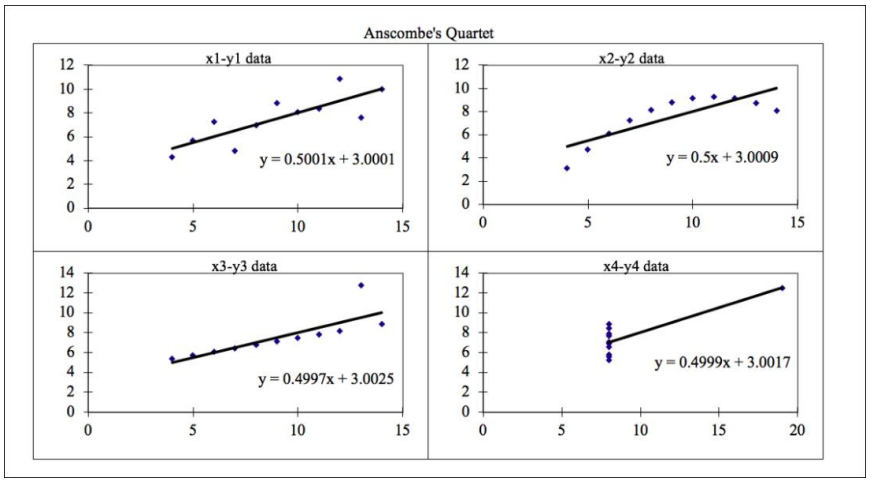
$$RSE = \sqrt{\frac{RSS}{df}}, \text{where } df = n - 2 \text{ and } n \text{ is the number of data points.}$$

Assumptions of Linear Regression:

6. There is a linear relationship between X and Y
7. Error terms are normally distributed with mean zero.
8. Error terms are independent of each other: Error terms should not be dependent on each other, like a time series data
9. Error terms have constant variance: Variance should not increase or decrease as the error value changes and variance should not follow any pattern as error terms change.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartlet comprises four data sets that have nearly identical simple descriptive statistics,yet have different distributions and appear very different when graphed.



It was constructed in 1973 by statistician, Francis Anscombe to illustrate the plotting the graphs before analyzing and model building and the effect of observations on statistical properties

Above figure shows same statistical observations (like variance, mean of all x,y points in all four datasets

The four plots shown above shares below data set and statistical observations:

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	

But the Visualization of data points using scatter plot tells us different story and cannot be interpretable by any of the linear algorithms

Q3. What is Pearson's R?

In Statistics, Pearson correlation coefficient (PCC) is also referred as Pearson's R the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation is a measure of linear correlation between two sets of data.

It is the covariance of two variables divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1

Covariance gives us direction of relation but Pearson's R give Strength as well as direction of the relation

This can be used in feature selection as well, say row(x,y,z) such that row(x,y)=1 that means both variables are highly correlated and one can be dropped.

As, with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Example: age and height of a sample of teenagers from a high school to have a Pearson coefficient significantly greater than 0 and less than 1

Formula for Pearson's R is given by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

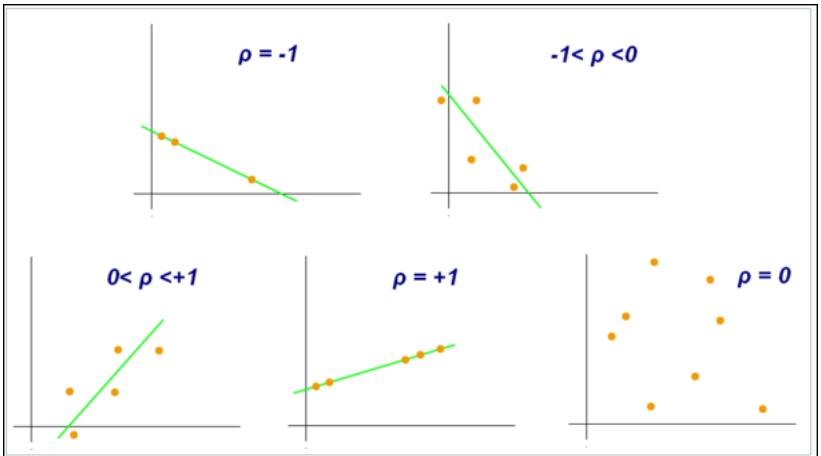
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable



Examples of scatter diagrams with different values of correlation coefficient (ρ)

In the diagram shown above, row (ρ)

=-1: x is decreasing when y is increasing

=1: x is increasing when y is increasing

=0: the data points are scattered and there is no relationship

=-1 <(ρ) <0: x is decreasing when y is increasing but not in a straight line

=0< (ρ) <1:x is increasing when y is increasing but not in a straight line

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

Scaling is to bring all the variables in one scale or same standing.

In many machine learning algorithms, to bring all the features in the same standing, we need to do scaling so that one significant number doesn't impact the model because of magnitude.

Feature scaling is one of the most critical steps during the pre-processing of data before creating a machine learning model.

Scaling is used for all of the below to bring them in one standing so the model build is more meaningful and accurate.

- Gradient descent based Algorithms
- Distance-based Algorithms
- Tree-based Algorithms

In Simple Linear Regression, scaling doesn't impact our data but where we have multiple independent variables to predict the model it is extremely important to scale the variables so that variables become comparable in one scale.

If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are two ways where scaling can be performed:

1. Normalization(Min-Max scaling)
2. Standardisation (mean-0, sigma-1)

Normalization – Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging 0 and 1. It is also known as Min-Max scaling.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where Xmax and Xmin is the maximum and minimum values of the feature respectively.

- In this case when value of x is mimimum value in the column the numerator will be 0 and hence X is 0
- When x is the maximum value in the column the numerator is equal to the denominator and hence x becomes 1
- The value always lies between 0 and 1

Standardization – Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

$$z = \frac{x - \mu}{\sigma}$$

This is also called Z-score normalization, the features will be rescaled so that they will have the properties of a standard normal distribution with mu=0 and sigma=1

The below plot describes the Normalization and Standardization very well.

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling.

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

VIF i.e **Variance Inflation factor** is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least square regression analysis. **It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.**

It calculates how well one independent variables is explained by all other independent variables combined.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

A **large value of VIF** indicates that there is a **correlation between the variables**.

So, if **VIF =infinity** that means there is a **perfect correlation** and the corresponding variable may be expressed exactly by a linear combination of other variables and hence, can be dropped for further model building process.

If R2=1 means perfect correlation and 1/(1-R2) = infinity. So, VIF becomes infinity when the R2=1

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

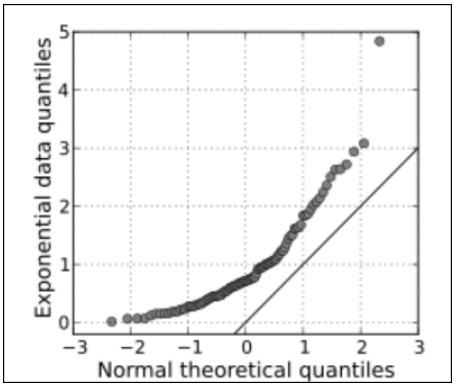
Q-Q Plots(Quantile-Quantile plots) are plots of two quantiles against each other. A Quantile is a fraction where certain values fall below that quantile. For example, median is a quantile where 50% data fall below the point and 50% data fall above the point

Purpose of Q-Q plot is to find if the two data sets come from a same distribution, a 45 degree angle is plotted on the Q-Q plot;if the two data sets come from a common distribution, the points will fall on that reference line.

Importance and Use:

Q-Q plots can be used to :

- It can be used in a single, simple way, **fit a linear regression model**, check if the points lie approximately on the line. This also helps in determining that your residuals are Gaussian
- We can **make use of CDF and standardization and explain non-normality**
- **Density plots** and **Residual KDE plots** can be drawn for testing Normality
- In case of Violations we can find the p-value and **reject the normality**
- **Z-scores can be plotted against data points and a linear relationship can be determined if exists**



The image shows quantiles from a theoretical normal distribution on the horizontal axis and its being compared to the set of data on y-axis and plot is called normal Quantile-Quantile plot.

The points are not clustered on the 45 degree line and in fact follow a curve, suggesting that the sample data is not normally distributed

Creation of Q-Q Plot:

1. Order the items from smallest to largest
2. Draw a normal Distribution curve
3. Find the z-value/cut off point for each segment and these segments are area and can be referred to z-table for their respective z-value.
4. Plot your data-set values(step1) against normal distribution cut-off points(step-3)
5. Infer the line (if straight line means data is approximately normally distributed)

Assumption of Normality:

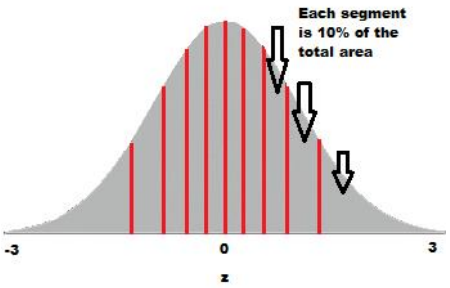
The assumption of Normality is an important assumption for many statistical tests; you assume you are sampling from a normally distributed population. This is just one way to assess normality.

Example:

Step1:

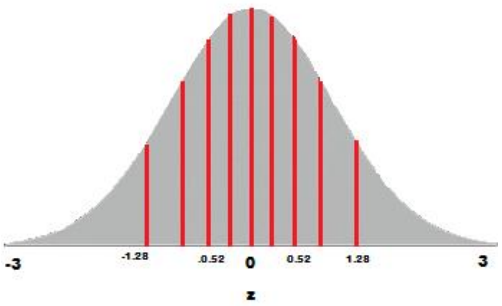
3.77
4.25
4.5
5.19
5.89
5.79
6.31
6.79
7.19

Step2: Divide curve into n+1 segments and draw a normal distribution curve:

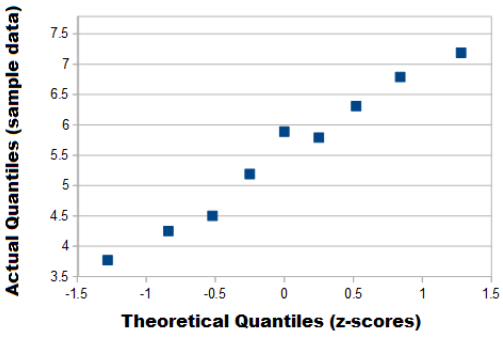


Step3 : Find the z-value cut-off point for each segment

10% -1.28
20% -0.84
30% -0.52
40% -0.25
50% 0
60% 0.25
70% 0.52
80% 0.84
90% 1.28
100% 3



Step 4: Plot step1 data Vs Step3 data



From the above method, we can create the linear regression best fit line if the residual is minimum. We can plot the error distribution and try to use the standardization concepts for a normal distribution of residuals.