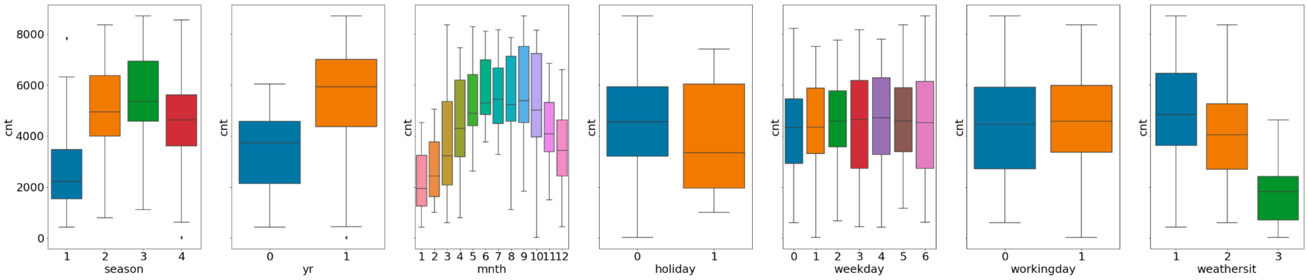
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**



Here Count refers to the count of total rental bikes.

A. Median Count increases from Spring, Summer, Fall and then drops in Winter.
Count has less range in Spring. Summer, Fall and Winter has similar range.
Count is right skewed in Summer whereas Summer, Fall and Winter are left skewed.

B. Median Count was higher in 2019 from 2018.
Count has more range in 2019 than in 2018.

C. Median count increases from month 1 to 6, then trend sideways in month 7, 8, 9 and then drops.
Count has less range in month 6 and 7 as compared to other months.

D. Median Count decreases on holiday.
Also, count has a lesser range on Holiday.
But, inter-quartile range is more on holiday.

E. Very slight variation in median count. **Drop in registered users on weekends is compensated by increase in casual users on weekend.**

F. Very slight variation in median count. **Drop in registered users on workingday is compensated by increase in casual users on weekend.**
Inter-quartile range is less on holiday.

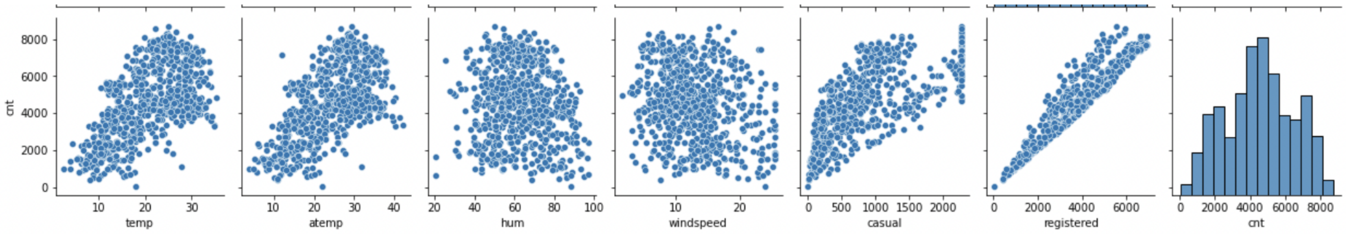
G. Median Count decreases sharply from Weather Situation 1 to 3.
Range of count also decreases in Weather Situation 3.

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

During dummy variable creation for a categorical variable having n levels, n columns are created for each level. Every column is binary in nature, i.e., it contains 0 and 1.
For example, we have 4 levels in season (categorical column) and we want to create dummy variables.
If one variable is not Spring, not Summer, not Fall then It is obviously Winter.
So we do not need the 4th variable to identify the Winter.
Here, the last row is Winter.

Spring	Summer	Fall
1	1	1
1	1	0
1	0	1
1	0	0
0	1	1
0	1	0
0	0	1
0	0	0

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



Looking at the scatter of the data, registered has the least scatter with count, i.e., for a value of x, y takes less values. Hence, it has the highest correlation with count.Alos, count increases with increases in registered users. So, it has a positive correlation.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Linear Regression Assumptions	**Validation**
Linear Relationship between x and y variable.	Scatterplots can show whether there is a linear or curvilinear relationship or no relationship between the variables.
No Linear Relationship between x variable and residuals or No Linear Relationship between y variable and residuals.	Scatterplot can be drawn considering Residuals as Y variable and independent variables as X variables or Residuals as Y variables and Fitted values as Y variables respectively.
Multicollinearity should not be there.	Pairwise correlation plot and Variance inflation factor (VIF).
Error terms should be normally distributed	Plot residual values on a histogram with a fitted normal curve or by reviewing a Q-Q-Plot. Normality can also be checked with a goodness of fit test e.g., the Kolmogorov-Smirnov test
Independence of error terms, i.e., Auto Correlation	Use the Durbin-Watson statistic. It is based on an assumption that errors are generated by a first-order autoregressive process. Compare displayed statistic with lower and upper bounds in a table. If D > upper bound, no correlation exists; if D < lower bound, positive correlation exists; if D is in between the two bounds, the test is inconclusive.
Constant variance of error terms, i.e, Homoscedasticity	Plot residuals against predicted values. There must be no pattern in the standard residuals scatter plot

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
In order to pick the top 3 variables contributing significantly, one needs to look at the significance, i.e., p-value and coefficient.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1602	0.024	6.760	**0.000**	0.114	0.207
yr	0.2289	0.008	28.648	**0.000**	0.213	0.245
holiday	-0.0555	0.027	-2.061	0.040	-0.108	-0.003
workingday	0.0443	0.012	3.842	**0.000**	0.022	0.067
season2	0.1044	0.011	9.702	**0.000**	0.083	0.126
season4	0.1353	0.012	11.342	**0.000**	0.112	0.159
mnth8	0.0563	0.016	3.491	0.001	0.025	0.088
mnth9	0.1256	0.016	7.779	**0.000**	0.094	0.157
mnth10	0.0404	0.017	2.358	0.019	0.007	0.074
weekday6	0.052	0.015	3.578	**0.000**	0.023	0.081
weathersit2	-0.0572	0.01	-5.462	**0.000**	-0.078	-0.037
weathersit3	-0.2443	0.026	-9.257	**0.000**	-0.296	-0.192
temp	0.5309	0.022	24.107	**0.000**	0.488	0.574
hum	-0.1363	0.03	-4.553	**0.000**	-0.195	-0.077
windspeed	-0.1372	0.019	-7.245	**0.000**	-0.174	-0.1

Temperature, Year, and Season 4, i.e., Winter has 0.5309, 0.2289 and 0.1353 coefficients respectively. These are the op 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

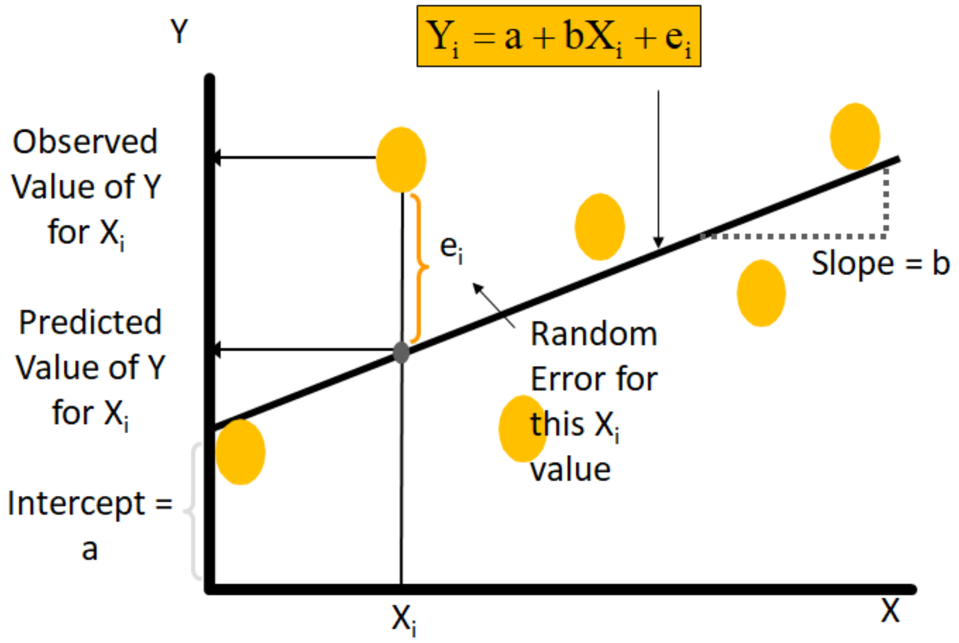
“A function is a set of ordered pairs of numbers (x,y) such that to each value of the first variable (x) there corresponds a unique value of the second variable (y)”. More intuitively, if there is a regular relationship between two variables, there is usually a function that describes the relationship. Functions are written in a number of forms. The most general is y = f(x), which simply says that the value of y depends on the value of x in some regular fashion, though the form of the relationship is not specified. The simplest functional form is the linear function where: $y = \alpha + \beta x$

α and β are parameters, remaining constant as x and y change. α is the intercept and β is the slope.

There can also be non-linear functions. Regression allows you to estimate directly the parameters in linear functions only, though there are tricks that allow many non-linear functional forms to be estimated indirectly. Regression also allows you to test to see if there is a functional relationship between the variables, by testing the hypothesis that each of the slopes has a value of zero.

The technique that specifies the dependence of the response variable on the explanatory variable is called regression.

When that dependence is linear, the technique is called linear regression.
The idea behind regression is that when there is significant linear correlation, you can use a line to estimate the value of the dependent variable for certain values of the independent variable.
Linear regression is therefore the technique of finding the line that best describes how the response variable linearly depends on the explanatory variable.

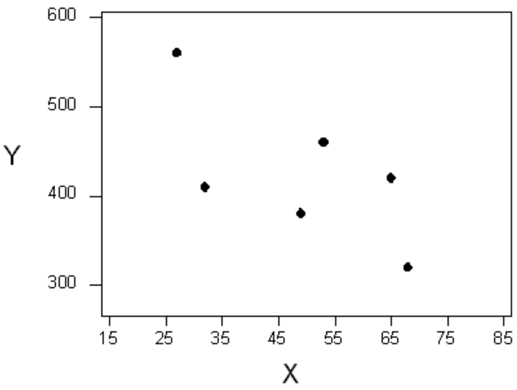


Y
$Y_i = a + bX_i + e_i$
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = a
e_i
Random Error for this X_i value
Slope = b
X_i
X

Best Fit Line

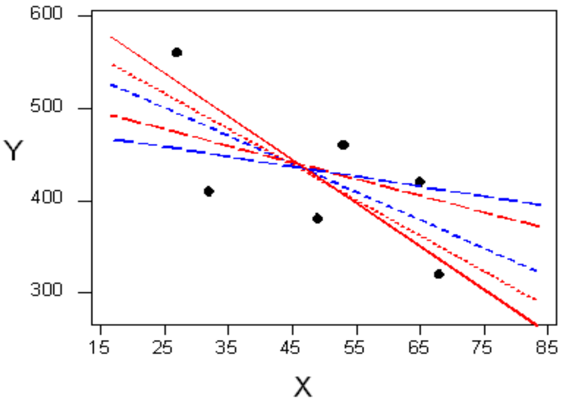
It can be of any shape depending on the number of independent variables (a point on the axis, a line in two dimensions, a plane in three dimensions, or a hyperplane in higher dimensions).
First, let us consider the simple case of a two-variable function. You believe that y, the dependent variable, is a linear function of x, the independent variable — y depends on x. Collect a sample of (x, y) pairs, and plot them on a set of x, y axes. The basic idea behind regression is to find the equation of the straight line that comes as close as possible to as many of the points as possible. The parameters of the line drawn through the sample are unbiased estimators of the parameters of the line that would come as close as possible to as many of the points as possible in the population, if the population had been gathered and plotted. In keeping with the convention of using Greek letters for population values and Roman letters for sample values, the line drawn through a population is $y = \alpha + \beta x$ while the line drawn through a sample is $y = a + bx$
In most cases, even if the whole population had been gathered, the regression line would not go through every point. Most of the phenomena that business researchers deal with are not perfectly deterministic, so no function will perfectly predict or explain every observation.

To understand how such a line is chosen, consider the following very simplified version of the age-distance example (we left just 6 of the drivers on the scatterplot):



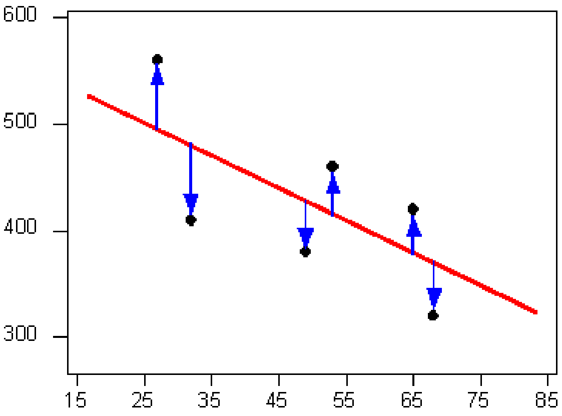
If we assume a straight line relationship is present between the variables, then draw a straight line that best captures the relationship between X and Y.

From scatter plot we can clearly say that there are many lines that look like they would be good candidates to be the line that best fits the data:

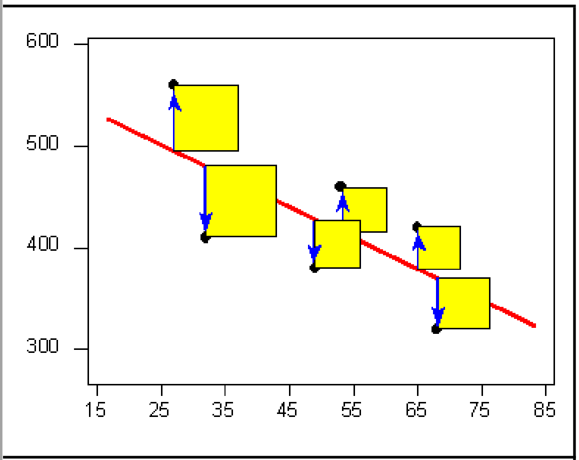


It is doubtful that everyone would select the same line in the plot above. We need to agree on what we mean by "best fits the data"; in other words, we need to agree on a criterion by which we would select this line. We want the line we choose to be close to the data points. In other words, whatever criterion we choose, it had better somehow take into account the vertical deviations of the data points from the line, which are marked with blue arrows in the plot below:

The red diagonal line is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction.



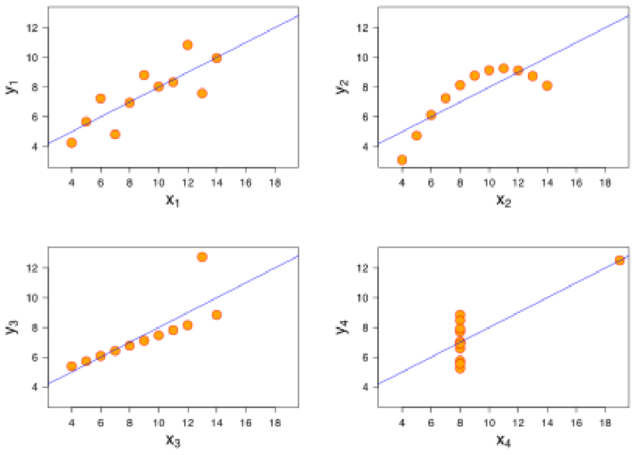
In estimating the unknown parameters of the population for the regression line, we need to apply a method by which the vertical distances between the yet-to-be estimated regression line and the observed values in our sample are minimized.
The most commonly used criterion is called the least squares criterion.
This criterion says: Among all the lines that look good on your data, choose the one that has the smallest sum of squared vertical deviations. Visually, each squared deviation is represented by the area of one of the squares in the plot below. Therefore, we are looking for the line that will have the smallest total yellow area.
This minimized distance is called sample error, though it is more commonly referred to as residual and denoted by e. In more mathematical form, the difference between the y and its predicted value is the residual in each pair of observations for x and y.
Obviously, some of these residuals will be positive (above the estimated line) and others will be negative (below the line). If we add all these residuals over the sample size and raise them to the power 2 in order to prevent the chance those positive and negative signs are cancelling each other out, we can write the following criterion for our minimization problem as shown on the right.

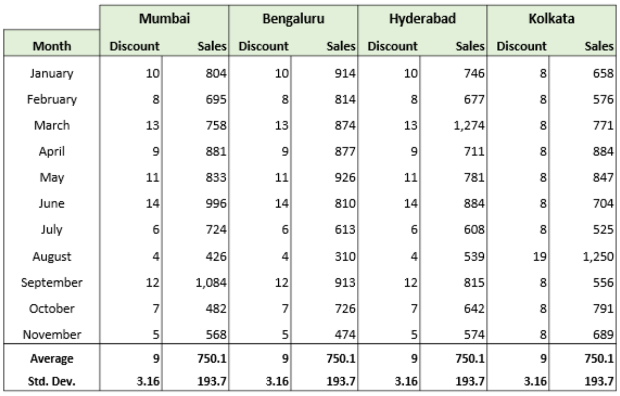


$$S = Min \sum_{i=0}^{n} (y - \hat{y})\hat{}2$$

$$Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

7. **Explain the Anscombe's quartet in detail. (3 marks)**
Anscombe's quartet comprises four data sets that **have nearly identical simple descriptive statistics, i.e., mean, standard deviation etc.,** 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."
Below is an example of discount and sales in 4 cities, i.e, Mumbai, Bengaluru, Hyderabad, Kolkata.
The distribution is very different but the mean and standard deviation is same.





	Mumbai		Bengaluru		Hyderabad		Kolkata	
Month	**Discount**	**Sales**	**Discount**	**Sales**	**Discount**	**Sales**	**Discount**	**Sales**
January	10	804	10	914	10	746	8	658
February	8	695	8	814	8	677	8	576
March	13	758	13	874	13	1,274	8	771
April	9	881	9	877	9	711	8	884
May	11	833	11	926	11	781	8	847
June	14	996	14	810	14	884	8	704
July	6	724	6	613	6	608	8	525
August	4	426	4	310	4	539	19	1,250
September	12	1,084	12	913	12	815	8	556
October	7	482	7	726	7	642	8	791
November	5	568	5	474	5	574	8	689
Average	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**
Std. Dev.	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**

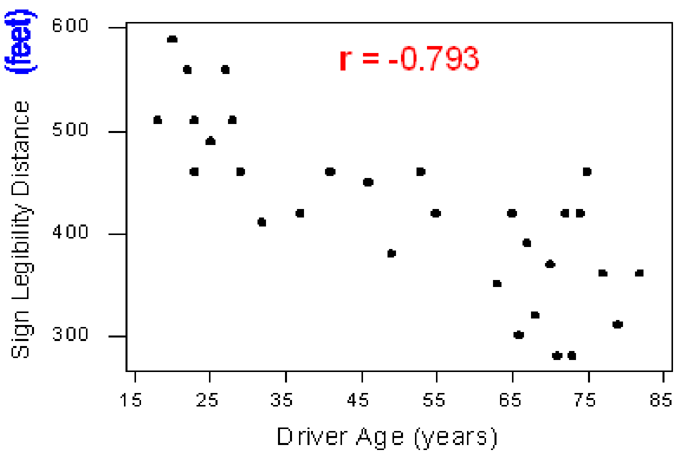
8. What is Pearson’s R? (3 marks)

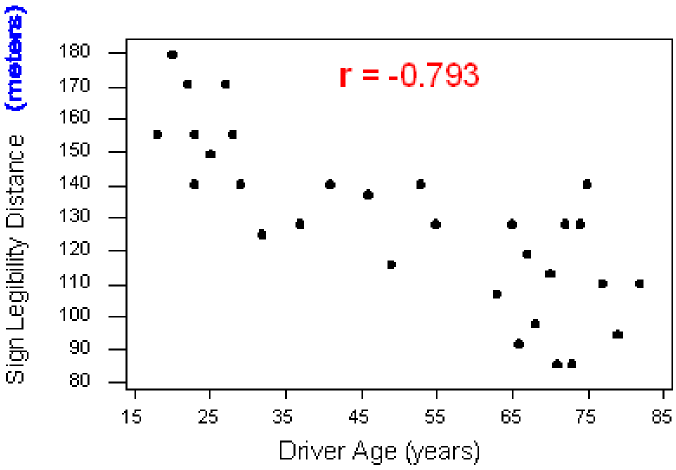
The correlation coefficient (r) is a numerical measure that measures the strength and direction of a linear relationship between two quantitative variables.

The population parameter is denoted by the greek letter ρ and the sample statistic is denoted by the roman letter r.

Here are some properties of r

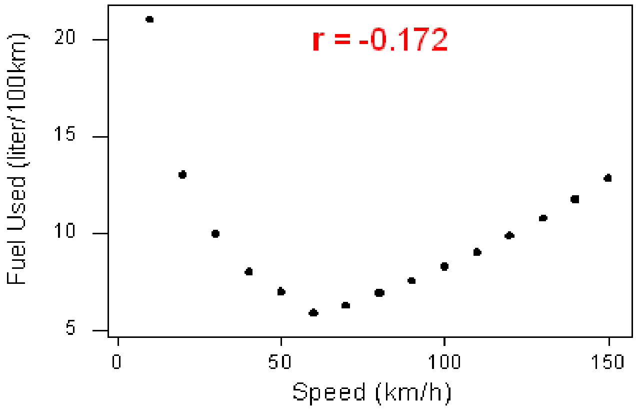
The correlation does not change when the units of measurement of either one of the variables change. In other words, if we change the units of measurement the explanatory variable and/or the response variable, the change has no effect on the correlation (r).
You may multiply, divide, add, or subtract a value to/from all the x-values or y-values without changing the value of r.
To illustrate, following are two versions of the scatterplot of the relationship between sign legibility distance and driver's age:
The top scatterplot displays the original data where the maximum distances is measured in feet. The bottom scatterplot displays the same relationship but with maximum distances changed to meters. Notice that the Y-values have changed, but the correlations are the same. This example illustrates how changing the units of measurement of the response variable has no effect on r, but as we indicated above, the same is true for changing the units of the explanatory variable, or of both variables.
This might be a good place to comment that the correlation (r) is unitless. It is just a number.



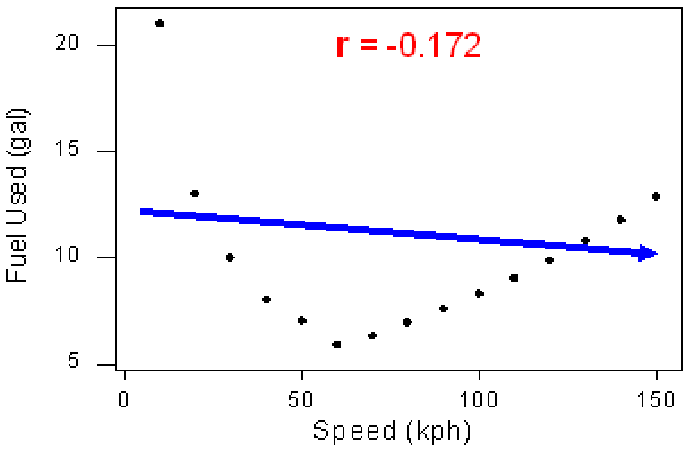


The correlation measures only the strength of a linear relationship between two variables. It ignores any other type of relationship, no matter how strong it is. For example, consider the relationship between the average fuel usage of driving a fixed distance in a car, and the speed at which the car drives:
Our data describe a fairly simple curvilinear relationship: the amount of fuel consumed decreases rapidly to a minimum for a car driving 60 kmph, and then increases gradually for speeds exceeding 60 kmph. The relationship is very strong, as the observations seem to perfectly fit the curve.

Although the relationship is strong, the correlation r = -0.172 indicates a weak linear relationship. This makes sense considering that the data fails to adhere closely to a linear form.

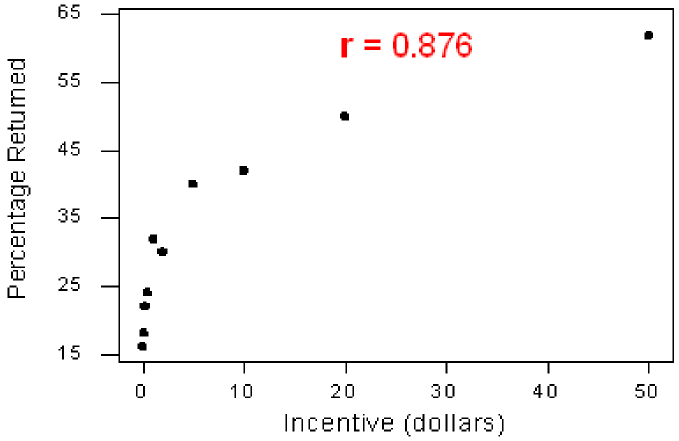


The correlation is useless for assessing the strength of any type of relationship that is not linear (including relationships that are curvilinear, such as the one in our example). Beware, then, of interpreting the fact that r is close to 0 as an indicator of a weak relationship rather than a weak linear relationship. This example also illustrates how important it is to always look at the data in the scatterplot because, as in our example, there might be a strong nonlinear relationship that r does not indicate.
Since the correlation was nearly zero when the form of the relationship was not linear, we might ask if the correlation can be used to determine whether or not a relationship is linear.



The correlation by itself is not sufficient to determine whether a relationship is linear. To see this, let's conside the study that examined the effect of monetary incentiv on the return rate of questionnaires. Below is the scatterplot relating the percentage of participants who completed a survey to the monetary incentive that researchers promised to participants, in which we find a strong curvilinear relationship:

The relationship is curvilinear, yet the correlation r = 0.8 is quite close to 1.
In the last two examples, we have seen two very strong curvilinear relationships, one with a correlation close to and one with a correlation close to 1. Therefore, the correlation alone does not indicate whether a relationsh is linear. The important principle here is:
Always look at the data!



The correlation is heavily influenced by outliers. As you will learn in the next two activities, the way in which the outlier influences the correlation depends upon whether or not the outlier is consistent with the pattern of the linear relationship.

r is always between -1 and 1 inclusive. -1 means perfect negative linear correlation and +1 means perfect positive linear correlation.

r has the same sign as the slope of the regression (best fit) line

r does not change if the independent (x) and dependent (y) variables are interchanged r has a Student's t distribution

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Here is the formula for r.

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

If you divide the numerator and denominator by n, then you get something which is starting to hopefully look familiar. Each of these values have been seen before in the Sum of Squares notation section. So, the linear correlation coefficient can be written in terms of sum of squares.

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} = \frac{(n-1)s_{xy}^2}{\sqrt{(n-1)s_x^2(n-1)s_y^2}} = \frac{s_{xy}^2}{s_x s_y} = \frac{COV(x,y)}{s_x s_y}$$

This is the formula that we would be using for calculating the linear correlation coefficient if we were doing it by hand.

(x,y)	Remarks
1	the variables have a perfect positive correlation. This means that if one variable moves a given amount, the second moves proportionally in the same direction. A positive correlation coefficient less than one indicates a less than perfect positive correlation, with the strength of the correlation growing as the number approaches one.
0	no relationship exists between the variables. If one variable moves, you can make no predictions about the movement of the other variable; they are uncorrelated.
-1	the variables are perfectly negatively correlated (or inversely correlated) and move in opposition to each other. If one variable increases, the other variable decreases proportionally. A negative correlation coefficient greater than –1 indicates a less than perfect negative correlation, with the strength of the correlation growing as the number approaches –1.

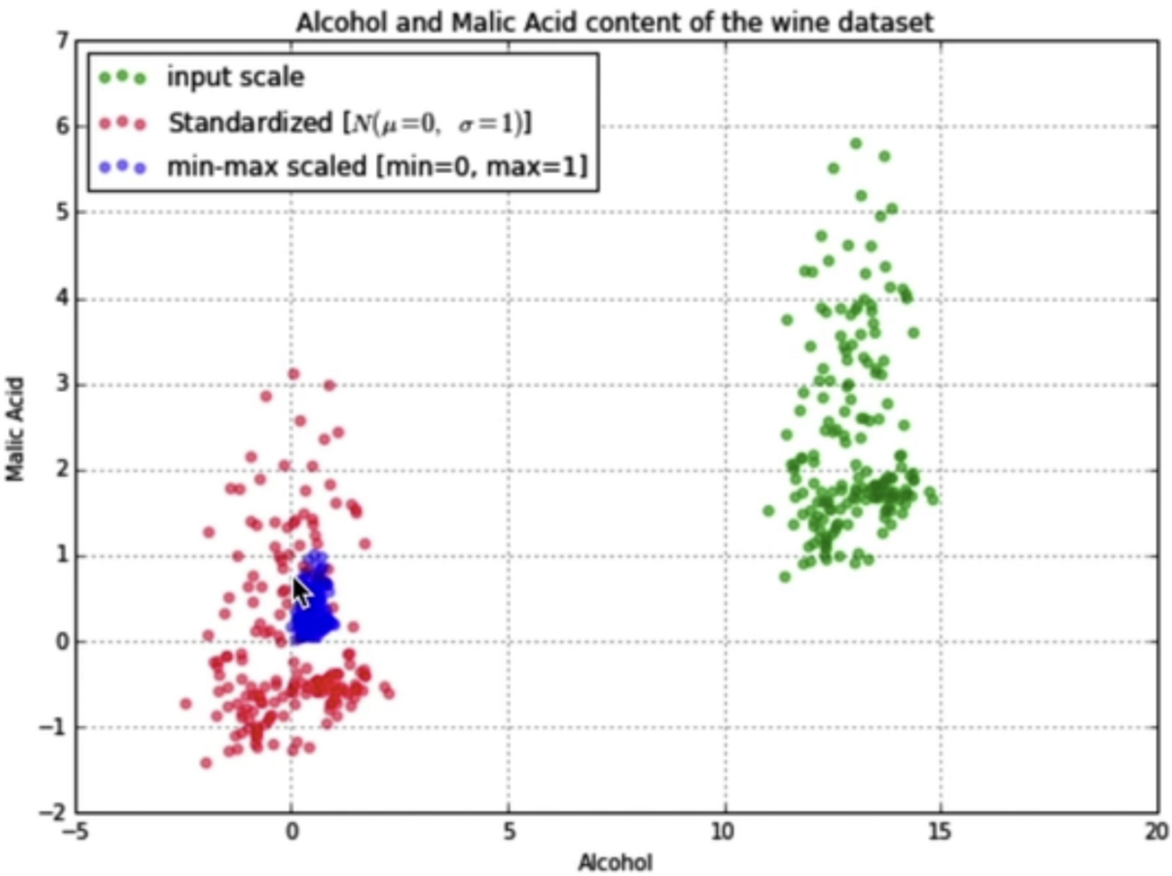
9. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Feature scaling helps you convert various variables having myriad units of measurement such as kilogram, Rupees, Years, etc into unitless measures.

We can speed up gradient descent by having each of our input values in roughly the same range. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

The way to prevent this is to modify the ranges of our input variables so that they are all roughly the same.

Normalization (Min Max Scaling)	Standard Scaling
Feature scaling involves dividing the input values by the range (i.e. the maximum value minus the minimum value) of the input variable, resulting in a new range of just 1. $x_i := \frac{x_i - x_{min}}{x_{max} - x_{min}}$	Mean normalization involves subtracting the average value from the values resulting in a new average value of zero. $x_i := \frac{x_i - \mu_i}{s_i}$
compress the data between a particular range of 0 to 1	useful, especially if there are extreme data point (outlier)



Feature normalization (or data standardization) of the explanatory (or predictor) variables is a technique used to center and normalise the data by subtracting the mean and dividing by the variance.

So if you're performing the test-train split earlier, you take the mean and variance of the whole dataset and introducing information regarding the data like the minimum and maximum values, mean, variance etc into the training explanatory variables.

Using any information coming from the test set before or during training is a potential bias in the evaluation of the performance.

Therefore, you should perform feature normalisation over the training data.

Test set plays the role of fresh unseen data, so it's not supposed to be accessible at the training stage.

Then perform normalisation on testing instances as well, but this time using the mean and variance of training explanatory variables.

Do not recalculate them on the test set, because they would be inconsistent with the model and this would produce wrong predictions.
In this way, we can test and evaluate whether our model can generalize well to new, unseen data points.

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

Variance Inflation Factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity.
It measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

How the VIF is computed
The standard error of an estimate in a linear regression is determined by four things:

1. The overall amount of noise (error). The more noise in the data, the higher the standard error.
2. The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a scale effect).
3. The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.
4. The extent to which a predictor is correlated with the other predictors in a model.

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

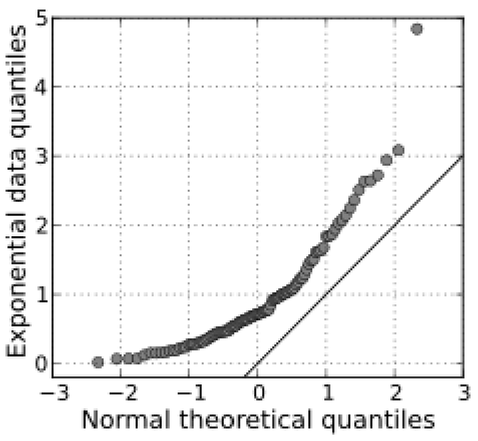
If VIF is infinite, it means 1 - R^2 is zero.

Or R^2 = 1.

It means 100% of the movement in the y-variable is explained by the x-variable.
It represents a perfect straight line with every y-variable falling on the line itself.

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

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, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the 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.
A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.