Anushka Bhagchandani B24 125

Assignment 2

Q1. Compare ggplot vs base plot with respect to their respective advantages and disadvantages. Give suitable examples.

[5 marks]
LO3, LO5

Pros and limitations of ggplot2

Pros:

- consistent, concise syntax
- Intuitive to most users
- visually appealing by default
- entirely customizable

Limitations:

- different syntax from the rest of R
- does not handle a few types of output well

Pros and limitations of base graphics

Pros:

- simple, straightforward for simple plots
- entirely customizable

Limitations:

- not visually appealing by default
- fiddly for adjusting positions, sizes, etc.
- syntax can get cumbersome for complex figures

A comparison between ggplot and base plot, based on the criteria of the various plots that can be plotted using both the packages.

Bar plot:

- In ggplot2, there is a call for each component, and you piece them together with the + operator.
- On the other hand, you use the barplot() function with base graphics and specify everything in the function arguments.
- The idea is that you can piece together various parts using the grammar for other visualization types. Whereas the single function call to barplot() is specialized to one thing.

Bins:
In ggplot2, you specify a binning by day through aes() and geom_bar(). However, in base graphics, you work with the data outside of the visualization functions. In this case, you can use table() to aggregate by day, and you pass that result to barplot().

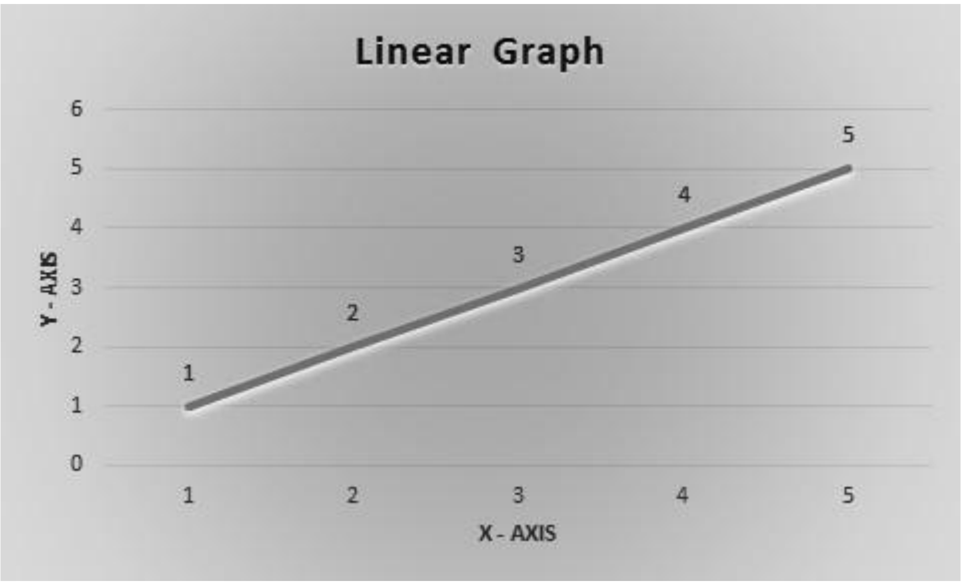
Line chart:
A component approach for ggplot2 with calls to geom_point() and geom_line(). In contrast, base plots make a call to plot() to make a line chart and then points() to add the circles at the end of the line.

Q2. Explain the mathematics behind linear regression with the help of a relevant dataset.

[5 marks]
LO4, LO6

Linear Regression is a predictive algorithm which provides a Linear relationship between Prediction ('Y') and Input ('X'). As we know from basic math, if we plot an 'X','Y' graph, a linear relationship will always come up with a straight line. For example, if we plot the graph of these values

(Input) X = 1,2,3,4,5 (Prediction) Y = 1,2,3,4 gives a perfectly straight line



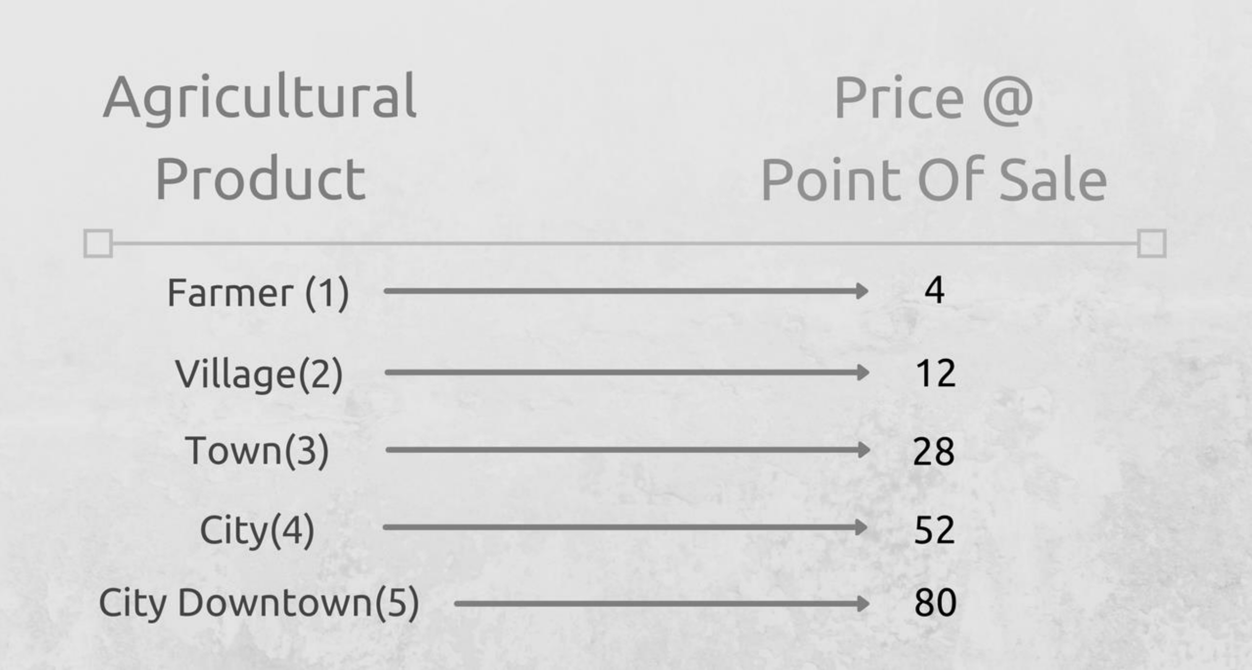
The equation of any straight line is written using the y = mx + b, where m is the slope (Gradient) and b is y-intercept (where the line crosses the Y axis).

Once we get the equation of a straight line from 2 points in space in y = mx + b format, we can use the same equation to predict the points at different values of x which result in a straight line.

Linear regression is a way to predict the 'Y' values for unknown values of Input 'X' like 1.5, 0.4, 3.6, 5.7 and even for -1, -5, 10 etc.

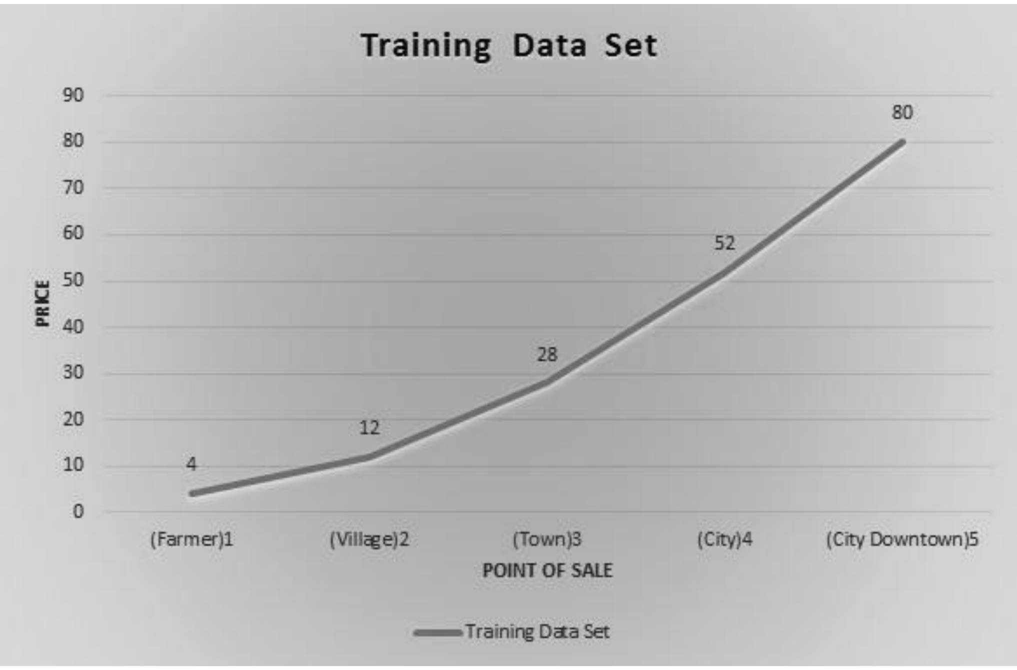
Let’s take a real world example to demonstrate the usage of linear regression and usage of Least Square Method to reduce the errors. Let’s take a real world example of the price of agricultural products and how it varies based on the location its sold. The price will be low when bought directly from farmers and high when brought from the downtown area.

Given this dataset, we can predict the price of the product in intermediate locations



Agricultural Product	Price @ Point Of Sale
Farmer (1)	4
Village(2)	12
Town(3)	28
City(4)	52
City Downtown(5)	80

In this example, if we consider Input 'X — Axis' as Sale Location and 'Y — Axis' as Price, we can plot the graph as



Problem Statement: Given this dataset, predict the price of agricultural product, if it’s sold in intermediate locations between farmers house and city downtown

Training DataSet: The dataset provided above can be considered as Training DataSet for the problem statement stated above, If we consider these inputs as Training Data for the model, we can use that model to predict the price at locations between
Farmers home — Village
Village — Town
Town — City
City — City Downtown

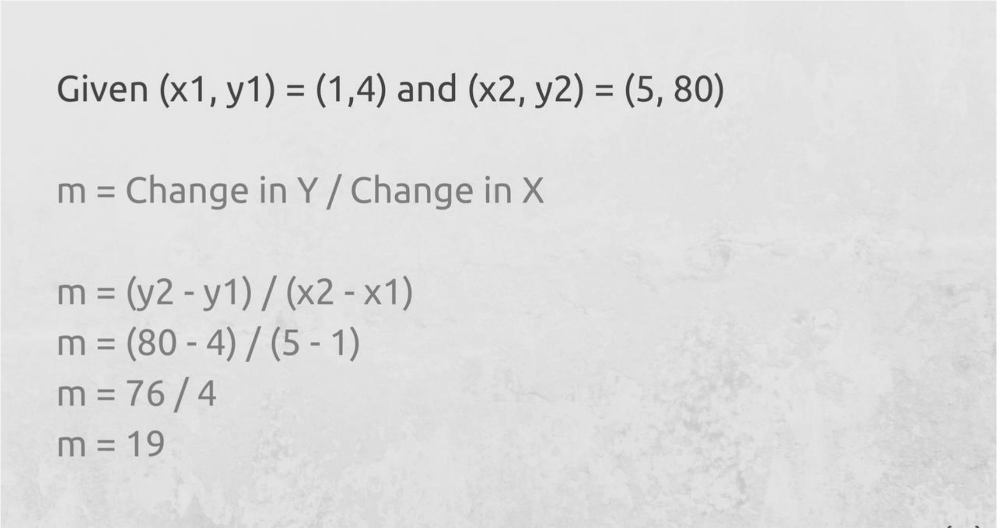
Our aim is to come up with a straight line which minimizes the error between training data and our prediction model when we draw the line using the equation of the straight line.

The math allows us to get a straight line between any two (x,y) points in a two dimensional graph. For this example, let’s consider farmers' homes and prices as the starting point and city downtown as the ending point.

The coordinates of the start and end points will be (x1,y1) = (1, 4), (x2,y2) = (5, 80) where x represents the location and y represents the price.

The first step is to come up with a formula in the form of y = mx + b where x is a known value and y is the predicted value.

To calculate the Prediction y for any Input value x we have two unknowns, the m = slope(Gradient) and b = y-intercept(also called bias)
Slope (m = Change in y/ Change in x)
The slope of the line is calculated as the change in y divided by change in x, so the calculation will look like

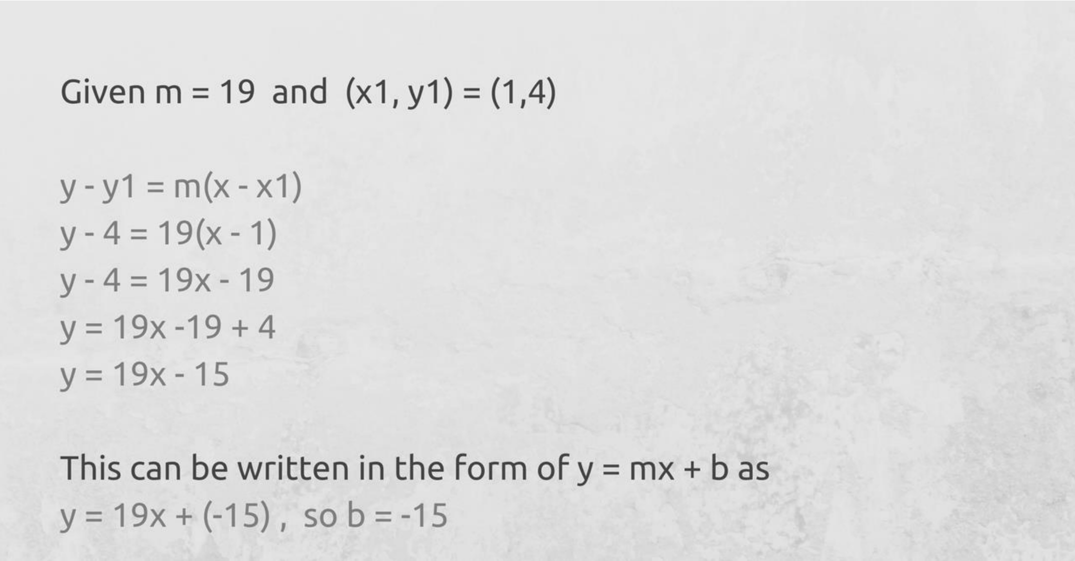


Given (x1, y1) = (1,4) and (x2, y2) = (5, 80)

m = Change in Y / Change in X

m = (y2 - y1) / (x2 - x1)
m = (80 - 4) / (5 - 1)
m = 76 / 4
m = 19

The y-intercept / bias shall be calculated using the formula $y-y1 = m(x-x1)$

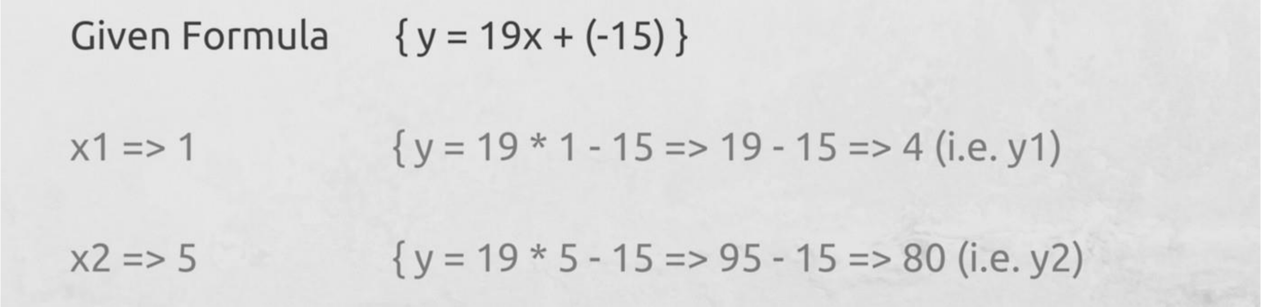


Given $m = 19$ and $(x1, y1) = (1,4)$

$y - y1 = m(x - x1)$
$y - 4 = 19(x - 1)$
$y - 4 = 19x - 19$
$y = 19x -19 + 4$
$y = 19x - 15$

This can be written in the form of $y = mx + b$ as
$y = 19x + (-15)$, so $b = -15$

Once we arrive at our formula, we can verify the same by substituting x for both starting and ending points which were used to calculate the formula as it should provide the same y value.

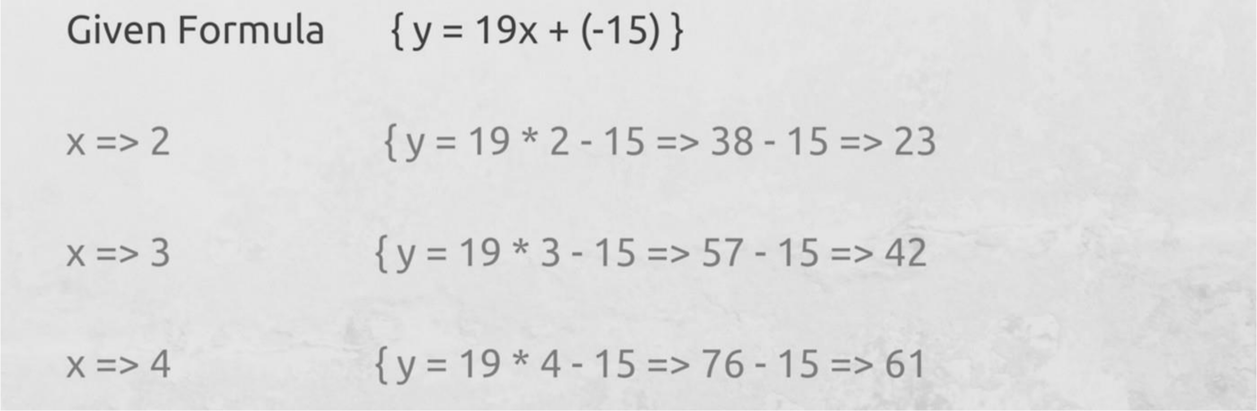


Given Formula { $y = 19x + (-15)$ }

x1 => 1 { $y = 19 * 1 - 15$ => $19 - 15$ => 4 (i.e. y1)

x2 => 5 { $y = 19 * 5 - 15$ => $95 - 15$ => 80 (i.e. y2)

Verifying $y = mx + b$



Given Formula { $y = 19x + (-15)$ }

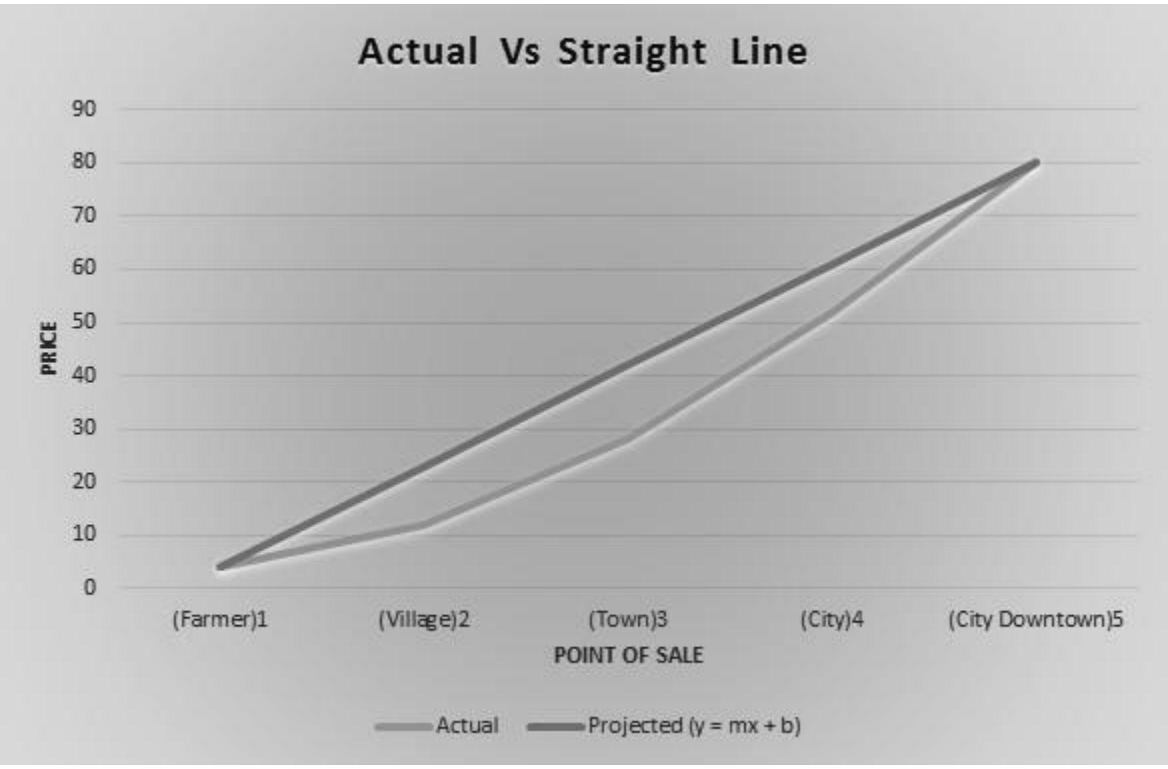
x => 2 { $y = 19 * 2 - 15$ => $38 - 15$ => 23

x => 3 { $y = 19 * 3 - 15$ => $57 - 15$ => 42

x => 4 { $y = 19 * 4 - 15$ => $76 - 15$ => 61

Predicting Y values for unknown X values

These values are different from what was actually there in the training set (understandably as original graph was not a straight line), and if we plot this(x,y) graph against the original graph, the straight line will be way off the original points in the graph of x=2,3, and 4.



Graph: Actual Line Vs Projected Straight Line