studentupsourcing.com

@supsourcing

Data Mining
Session #1

Student UPsourcing

A (Gentle) Intro To Data Mining And Machine Learning!

For Student UPSourcing 2015

studentupsourcing.com *@supsourcing*

Deolu Adeleye

Welcome!

I'm Deolu Adeleye, and I'm more than tickled pink to be the one introducing you into this exciting new world!

Let's dive right in, shall we?! :D

Something we should all be grateful for is that in the field of Computers and Electronics, nomenclature strongly follows real world phenomena.

In simple English, what I'm trying to say is that we (tend to) use words and terms, like 'Processing', 'Architecture', 'Server', 'Host' and so forth that have the same meaning or closely resemble their use in our everyday lives.

You will soon see too that in its most basic sense, 'machine learning' is about...well, machines that learn, or helping machines learn. And here, 'data mining' means just that: mining (raw) data into something more useful.

Data Mining

Thanks to the sharp and unabating rise in computing power over the years (your personal laptop is more powerful than the computer used to land the first man on the moon!), we find ourselves today with an increased capability to tackle some rather initially difficult problems. One of those problems is data *crunching* (or *mining*, but both mean the same thing)

Let's get a very basic idea of what that is by splitting the term and tackling its individual components:

- 'data'

and

- 'mining'

Mining

We tackle this first, because it is the most obvious: to *mine* something is to *excavate*, or it's rawest sense: *to dig*.

There, that's taken care of...

Data

So then: what is 'data'?

Well, we can simply define 'data' as a collection of facts and figures, be they recorded or measured.

The amount of bones in the human body...the sounds birds make...pictures posted on Instagram...these are all examples of 'data'.

Combining the two now: we can crudely define *Data Mining* as the process of excavating data from a source.

That's pretty much it.

Of course, as with practical mining, *you don't just dig anywhere* or *anyhow*. You have to

- *know where to* ***look***
- *know what you* ***want***
- *know what is* ***useful***
- *know what to do with* ***what you find***

You'll notice I made no mention of 'finding what ***you*** want'. As in real life, you will ***veeery rarely*** meet data that is exactly what you *need* and *already* in the *format* you need it: except someone does that for you, you'll usually need to do some pre-processing yourself to suit it to your purpose (like gold ores passing through a furnace...)

Continuing with our logic, we can break 'Data Mining' further into

- Data *Gathering* (our 'sourcing/digging' phase)
- Data *Analysis* (our 'figuring out what this is' phase)

(Note: it's actually *waaay* more than two, as there is still 'Cleaning Data' phase, Exploratory Data Analysis, Inference and Reporting...but this is a 'gentle' intro, remember? All that will come later :D)

Machine Learning

Thanks to Hollywood, I don't really have to work too hard to convince you: the concept of a machine displaying intelligent behaviour or 'doing human things' doesn't seem so far-fetched to us now. One such example is J.A.R.V.I.S., Tony Stark's personal virtual assistant (*Iron Man* movies).

Of course, as per Hollywood, things tend to get stretched *faar* beyond what the reality is...if we were to compare what the movies claim to where reality currently is, then J.A.R.V.I.S. can only be described as a hyper, *hyper*, **hyper**, ***hyper***, ***HYPER*** intelligent machine (and that's putting things mildly...)

The good news is, the J.A.R.V.I.S. reality may not be so far away in the future...

Once more, kindly again observe that I did *not* say 'making machines *intelligent*', but '*displaying* intelligent *behaviour*'. I point this out because it turns out not a lot of people have a unified agreement of what 'intelligence' is. Like,

- if you make a child memorize big words from the dictionary, does the child repeating said words at a later time mean he/she is intelligent, or just able to memorize?
- if your computer can predict what you'll eat tomorrow morning, is that intelligence, or just a 'trick' of statistics?

It's one of those areas philosphy bumps into electronics and computers. So, rather than argue or keep wondering about what 'intelligence' itself is (not) exactly, one thing we *do* all agree on is that there are clear attributes of intelligent *behaviour*, like *sensing*, *perception*, and *cognisance*, and these attributes can be recognised whenever displayed. As such, when something other than biological beings display said attributes, we refer to them as being ***artificially*** (as in *non-biological*) ***intelligent***.

Okay! I think I've fulfilled my philosophy quota. Moving on...

Practical Examples

In Data Mining

Let's look at some practical examples of data mining. Say someone came and dropped the following random numbers on your lap:

9.212224656 11.67968305 13.94964967 9.639728414 8.910806945 11.3190057 13.21611581 6.238036305 9.762514438 9.978639929 8.739640343 7.152172567 9.992850333 9.50009842 11.36732096 11.79591558 10.93099615 11.18536145 7.07887199 7.297773924 5.941906455 7.355677612 8.349606285 9.444760658 7.768624808 6.103134677 10.31485534 8.094337325 10.18542876 14.75314219 8.67245139 10.47949877 10.68095604 11.75593419 12.5729826 8.716186575 8.367377882 8.198182347 12.51647417 7.294255717 8.374946247 8.86546427 11.57918786 7.237818069 7.190257659 13.10920282 9.162478237 13.91980556 11.43241479 6.113264085 11.2511995 10.79678737 13.78289987 12.08424555 10.03607202 10.90793308 7.441513279 9.740836024 10.75969408 9.260381229 11.73454058 9.953715417 13.57099211 12.2819168 8.859859598 9.54402989 12.05035187 9.945217373 5.282841 11.55852947 11.89123733 11.87415334 11.74217006 7.094532407 9.339616756 10.38775672 9.24731488 7.293124319 7.772499374 11.609688 8.974806649 8.90764793 5.062099262 7.602209189 11.7539885 8.24401666 8.455199677 13.2339018 9.738811004 13.3083516 11.82376619 5.946682772 11.79436618 10.02177182 11.55182063 8.698614067 9.532949946 13.25678816 9.229040002 11.89212842

What could you say about them? Anything?

Not to worry! By the time we apply some data mining steps, these seemingly weird numbers will suddenly yield some surprising answers...

One of those steps is implemented here. ***NOW*** can you say anything about these numbers?

9.212224656	11.67968305	13.94964967	9.639728414	8.910806945
11.3190057	13.21611581	6.238036305	9.762514438	9.978639929
8.739640343	7.152172567	9.992850333	9.50009842	11.36732096
11.79591558	10.93099615	11.18536145	7.07887199	7.297773924
5.941906455	7.355677612	8.349606285	9.444760658	7.768624808
6.103134677	10.31485534	8.094337325	10.18542876	14.75314219
8.67245139	10.47949877	10.68095604	11.75593419	12.5729826
8.716186575	8.367377882	8.198182347	12.51647417	7.294255717
8.374946247	8.86546427	11.57918786	7.237818069	7.190257659
13.10920282	9.162478237	13.91980556	11.43241479	6.113264085
11.2511995	10.79678737	13.78289987	12.08424555	10.03607202
10.90793308	7.441513279	9.740836024	10.75969408	9.260381229
11.73454058	9.953715417	13.57099211	12.2819168	8.859859598
9.54402989	12.05035187	9.945217373	5.282841	11.55852947
11.89123733	11.87415334	11.74217006	7.094532407	9.339616756
10.38775672	9.24731488	7.293124319	7.772499374	11.609688
8.974806649	8.90764793	5.062099262	7.602209189	11.7539885
8.24401666	8.455199677	13.2339018	9.738811004	13.3083516
11.82376619	5.946682772	11.79436618	10.02177182	11.55182063
8.698614067	9.532949946	13.25678816	9.229040002	11.89212842

Just by re-arranging the numbers, we can tell a lot, such as the fact that there are ***100*** numbers in all (*20* rows by *5* columns)

Can we do more?

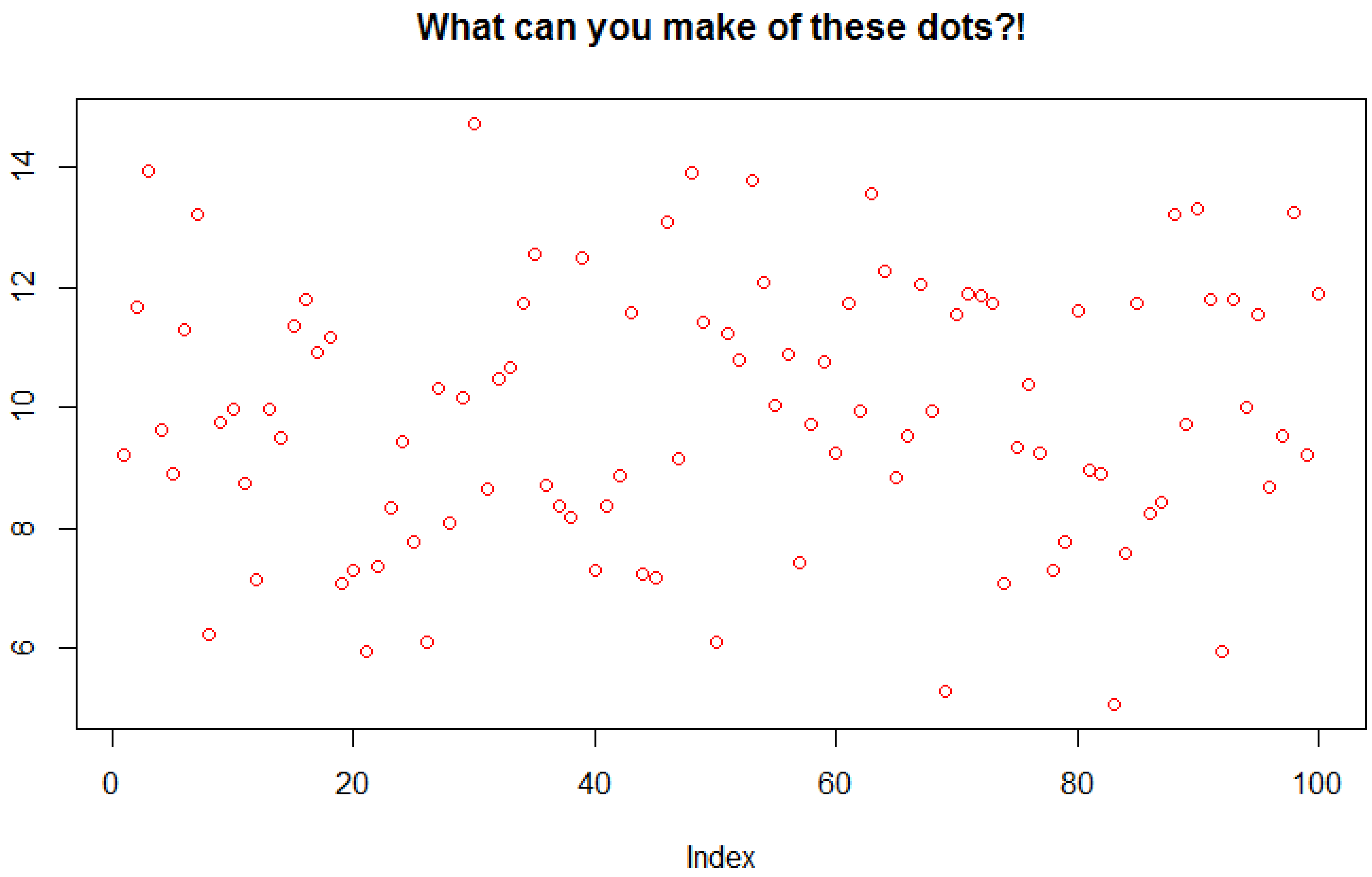
Yes!

```
5.062099    5.282841    5.941906    5.946683    6.103135
6.113264    6.238036    7.078872    7.094532    7.152173
7.190258    7.237818    7.293124    7.294256    7.297774
7.355678    7.441513    7.602209    7.768625    7.772499
8.094337    8.198182    8.244017    8.349606    8.367378
8.374946    8.455200    8.672451    8.698614    8.716187
8.739640    8.859860    8.865464    8.907648    8.910807
8.974807    9.162478    9.212225    9.229040    9.247315
9.260381    9.339617    9.444761    9.500098    9.532950
9.544030    9.639728    9.738811    9.740836    9.762514
9.945217    9.953715    9.978640    9.992850    10.021772
10.036072   10.185429   10.314855   10.387757   10.479499
10.680956   10.759694   10.796787   10.907933   10.930996
11.185361   11.251200   11.319006   11.367321   11.432415
11.551821   11.558529   11.579188   11.609688   11.679683
11.734541   11.742170   11.753988   11.755934   11.794366
11.795916   11.823766   11.874153   11.891237   11.892128
12.050352   12.084246   12.281917   12.516474   12.572983
13.109203   13.216116   13.233902   13.256788   13.308352
13.570992   13.782900   13.919806   13.949650   14.753142
```

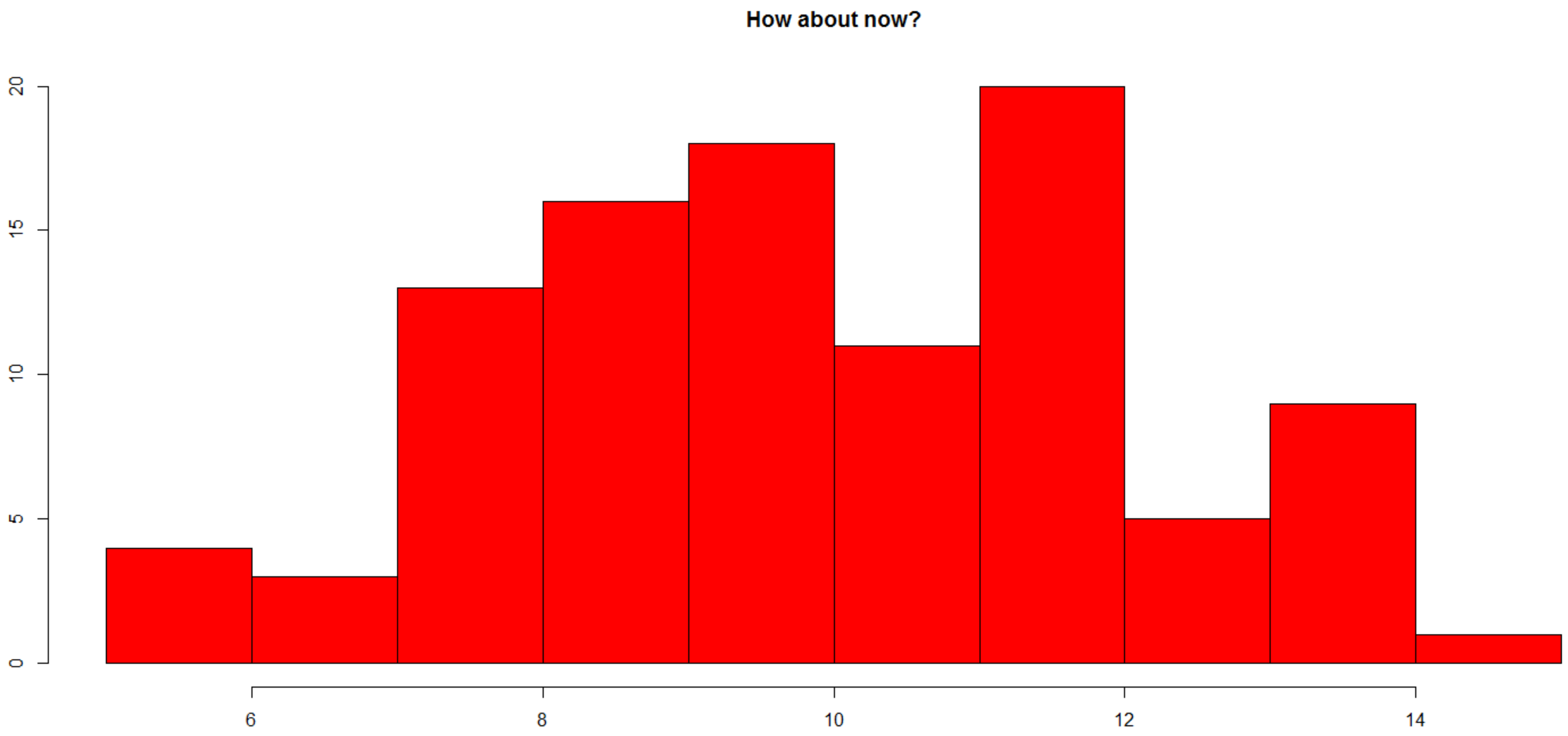
Now, by sorting, we can tell even more. We can now conclusively say that not only are there *100* numbers, but the lowest there is ***5.062099*** and the highest is ***14.753142***.

We're not through yet...

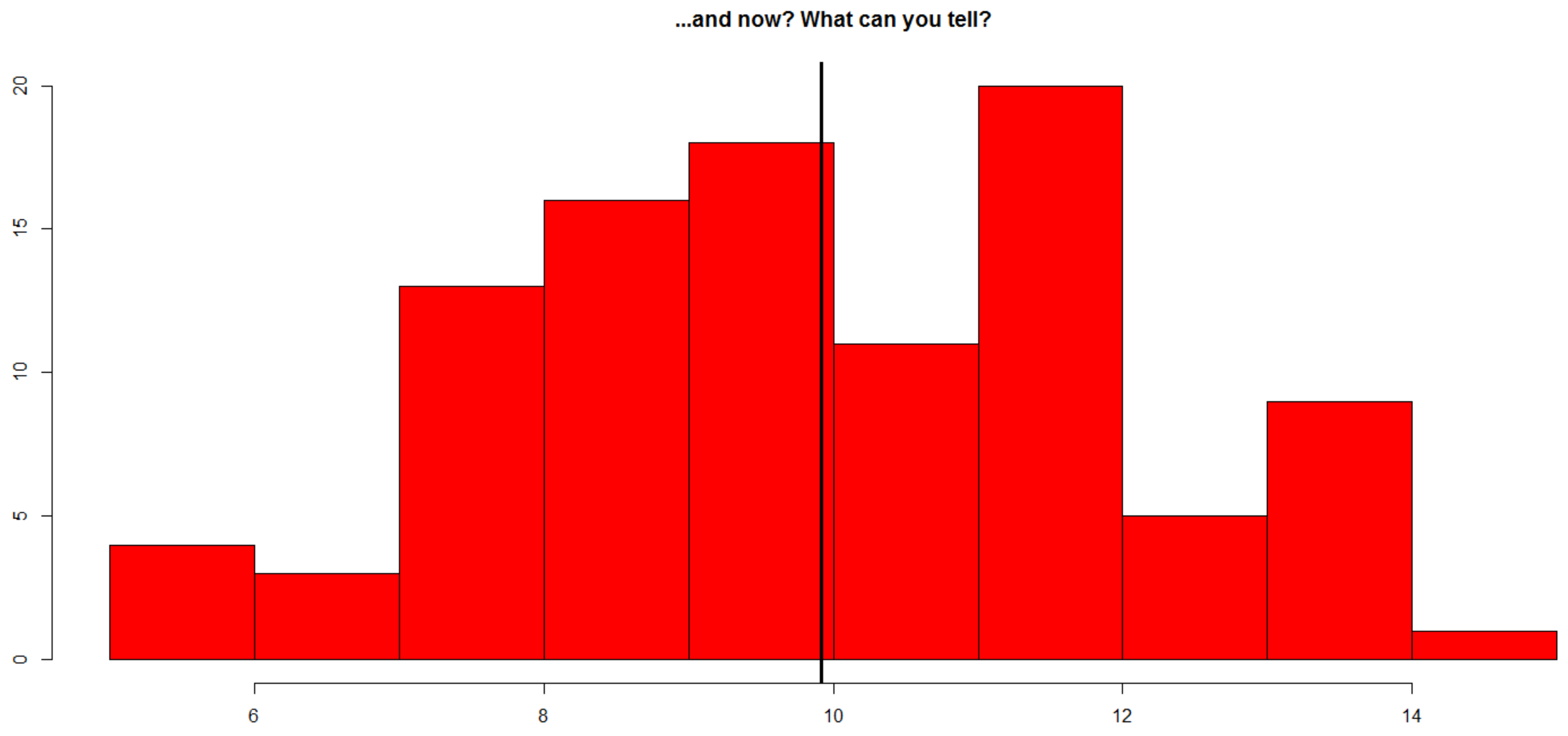
This is a plot of all those 100 numbers.



We'll keep going...making the 'random' data more communicable. Here’s a histogram:



Remember the mean, which is a measure of central tendency taught early in our mathematical curriculum? Here's a histogram with the mean shown:

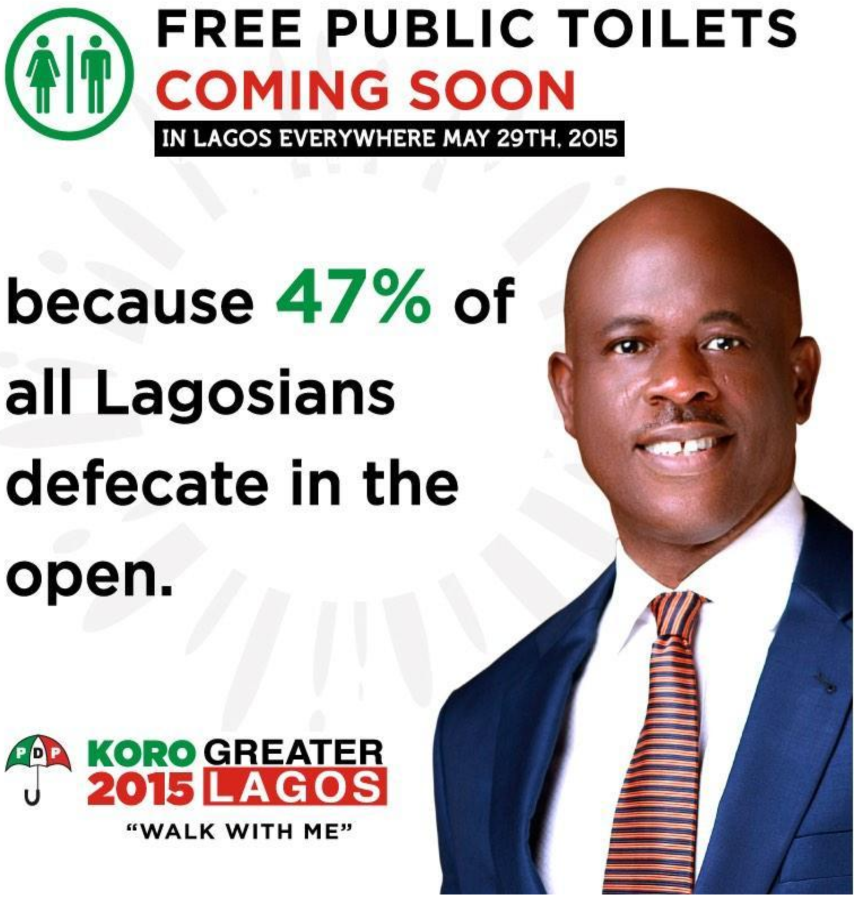


From this last plot, we're able to conclude that our data have a mean/central/'average' value of ***9.916226***.

Probably by now, you're excited at the prospects of what you can discover and do with data, which abounds all around us. Excellent!

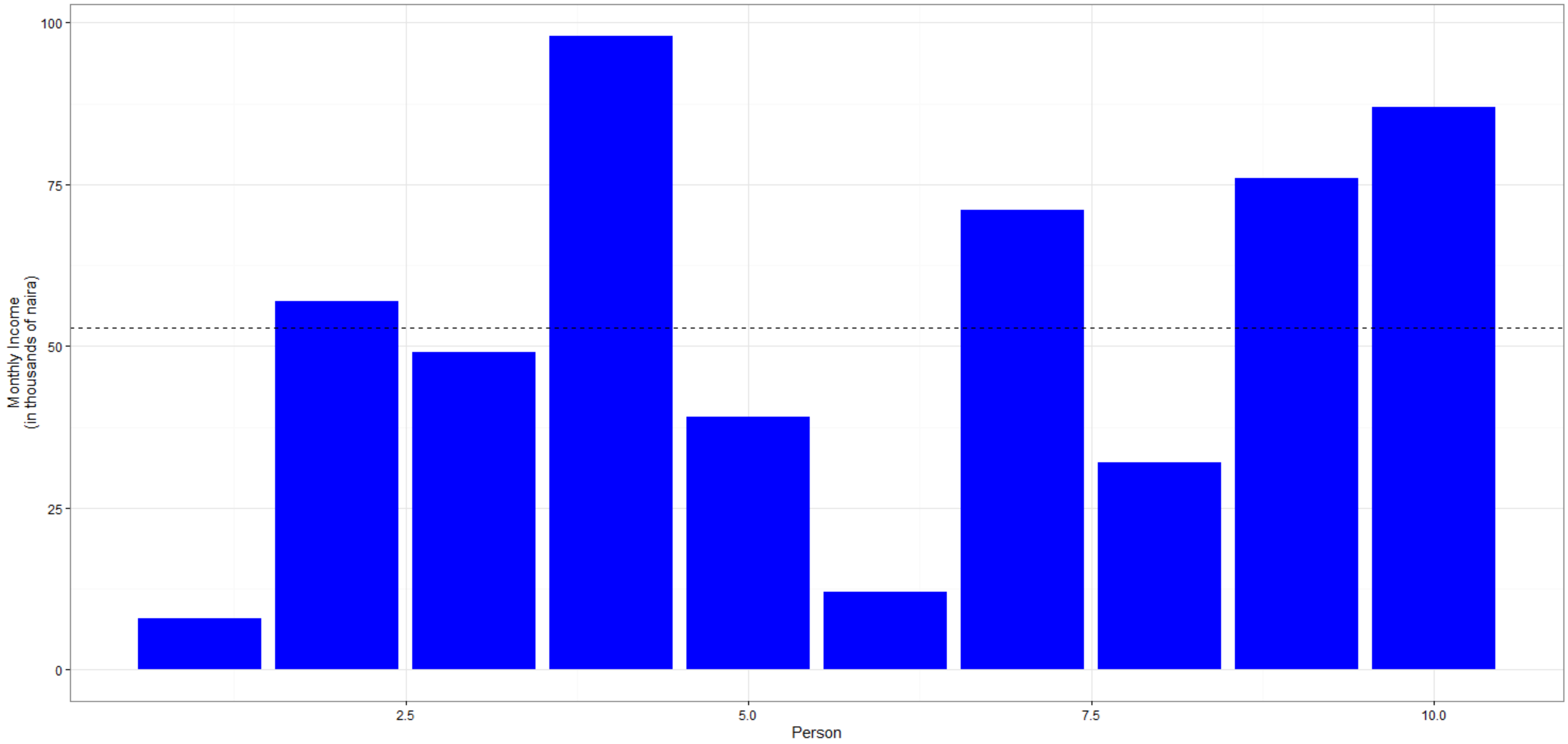
However, there are some pitfalls you'll need to watch out for...it's not all a smooth ride all the time.

For example, we have cases of *data confounding* (which you'll learn about later). In other cases, you just have what can only be described as abuse, or at most very poor statistical inference, such as in the picture below:


FREE PUBLIC TOILETS
COMING SOON
IN LAGOS EVERYWHERE MAY 29TH, 2015
because 47% of all Lagosians defecate in the open.
KORO 2015
GREATER LAGOS
"WALK WITH ME"

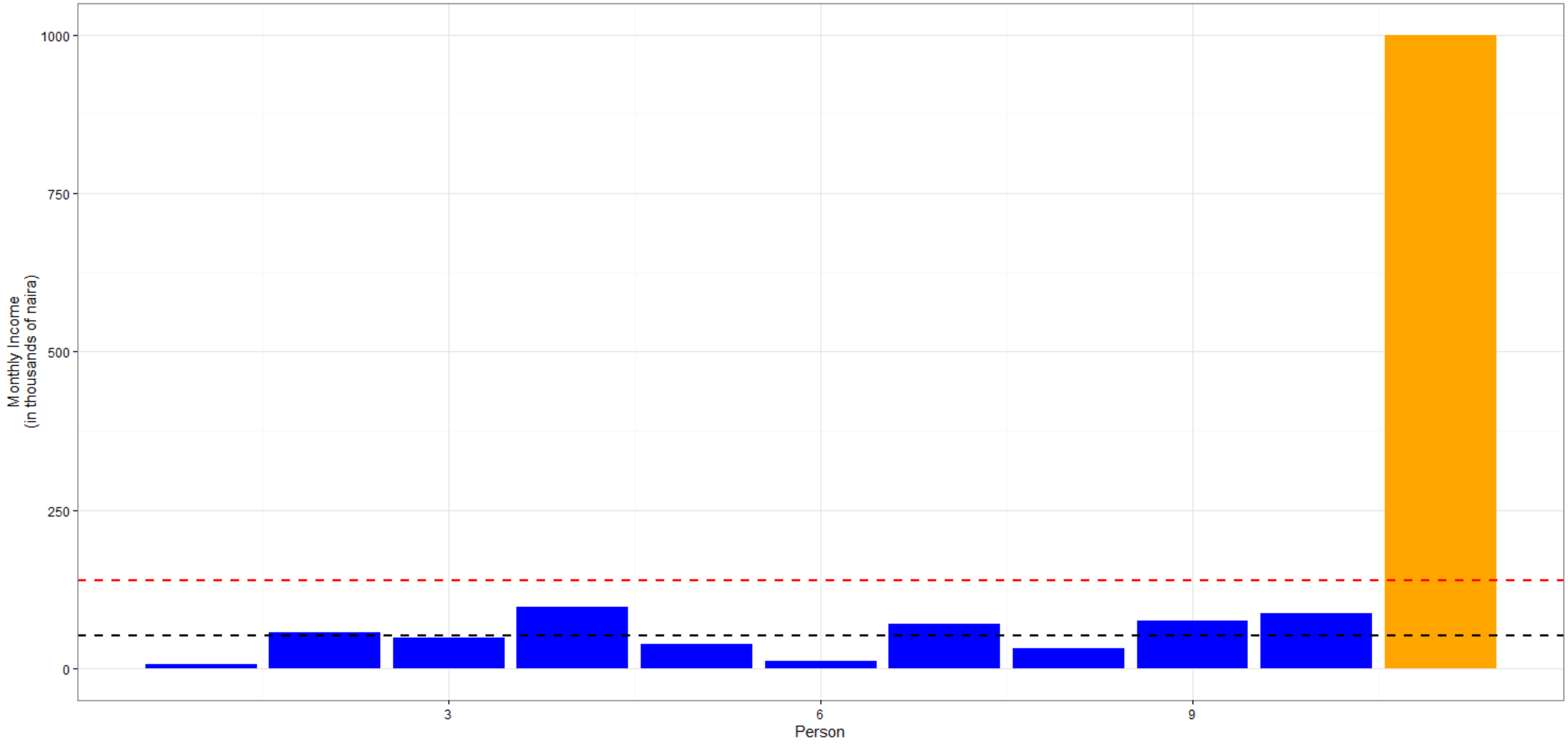
Who came up with this figure? How did they come up with it?!

Let's see another example: say a particular neighbour hood has 10 people living there, with the following incomes. Each person is plotted against their (in thousands of naira). To make it easier, the mean of all their incomes is highlighted by the dashed line:



Remember the mean being a measure used to detect the center of a group of data? We even used for our random data above, but we'll soon see it doesn't always give an accurate picture. In this case, we see the mean (or 'average') income in this community to be `52.9` thousand naira (₦52,900), right?

Well then, let's say just one person who earns much more than they all moves into that area, and let's say ***their*** own income is...let's modestly say one million naira (₦1,000,000). What happens when we add their own income?



Notice how the mean shifts considerably! The former 'average' (the black dashed line) was `52.9` thousand naira, but the addition of this one person has shifted it *waay* up to `139` thousand naira (₦139,000)! This person varies so much from the others, they're referred to in statistics as being an ***outlier***: something that differs considerably from the norm.

Now, here's where it gets even more interesting: say a shady/greedy realtor wants to sell you property in that area. To throw you off, rather than tell you "*Except for some outliers*, the average income here is ₦52900", they may just quote the average income as being **₦139,000**, making the neighbourhood appear to be what it's not! So, even though they **did** tell you the truth, ***they didn't give you the full picture***. Statistical figures, unfortunately, can be used to fool lots of people (examples abound today).

You'll probably be more wary of the word 'average' or 'mean' now I'm guessing... :)

Hopefully you won't fall into such potholes, and you'll be a great data scientist!

In this session, you've been given a very gentle introduction into this world of data mining and machine learning, and seen some potential (ab)uses of statistical analysis. Hopefully, your interest has been piqued to further investigate some of these topics more - there's an abundance of data around us all waiting to be explored and understood!

Happy hacking!