Bacon Bits

Delicious bits of Excel and Access Training from DataPig Technologies

Highlighting Outliers in your Data with the Tukey Method

datapig January 16, 2014 Business Statistics, Excel Formulas

42 Comments

I just recently completed a project that required a set of “Rank and Stack” reports for the purposes of identifying poor performers and recognizing good performers. In the process, it became clear that some of the data points had values that were so far removed from the norm that they were throwing off the ranking. So I had to go through an exercise of highlighting the outliers.

.

There are lots of statistical methods for identifying outliers. I used John Tukey’s method of leveraging the Interquartile Range. His method is applicable to most ranges since it isn’t dependent on distributional assumptions. It also ignores the mean and standard deviation, making it resistant to being influenced by the extreme values in the range.

.

Because I have no life, I thought it would be fun to explain Tukey’s method, and how to use it in Excel. So let me put my retainer in my mouth and let’s get started.

.

Understanding Interquartile Ranges

Try these Add-ins!

Search for a Topic

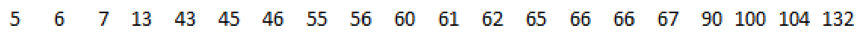
Search

Recent Posts

- Me vs Hungarian Notation
- Create SQL Server Tables from Excel Data – Free Tool
- Excel Against Humanity
- Best Practices for Organizing VBA Modules
- Pulling SSRS Data Directly into Excel with PowerPivot

When performing data analysis, we assume our values cluster around some central data point (a median). But sometimes, a few of the values fall too far from the central point – skewing the analysis. These values are called outliers (they lay outside the expected range).

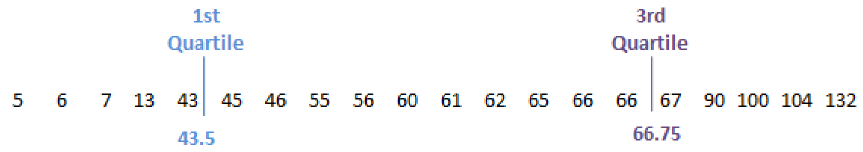
Let's use these numbers and try to find the outliers.



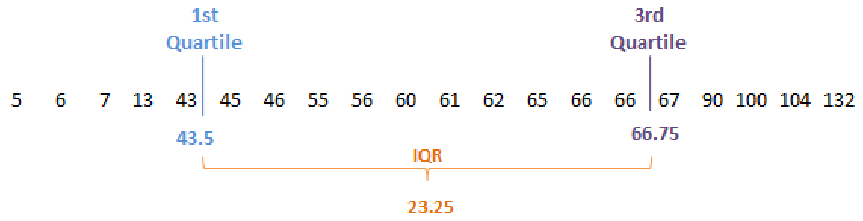
5 6 7 13 43 45 46 55 56 60 61 62 65 66 66 67 90 100 104 132

The first step in identifying outliers is to pinpoint the statistical center of the range. To do this, we start by finding the 1st and 3rd Quartiles. If you're not familiar with Quartiles, I explain them in detail in another post.

The 1st Quartile in this range is 43.5 – the 3rd Quartile is 66.75



Next, we subtract the 3rd Quartile from the 1st Quartile. This will give us an Interquartile Range (IQR). The IQR gives us a statistical way to identify where the bulk of the statistical data points (the middle 50%) sit in the range, and how spread out that middle 50% is. Note that statistical data points include not only the values you see, but all the unseen numbers between the values. In our example, the middle 50% spans 23.5 statistical data points.



Using IQR to Find Outliers

Now, the question is how far from the middle 50% can a value sit and still be considered a "reasonable" value? Leveraging the IQR, you can

Past Posts

Select Month ▼

Sites I Visit

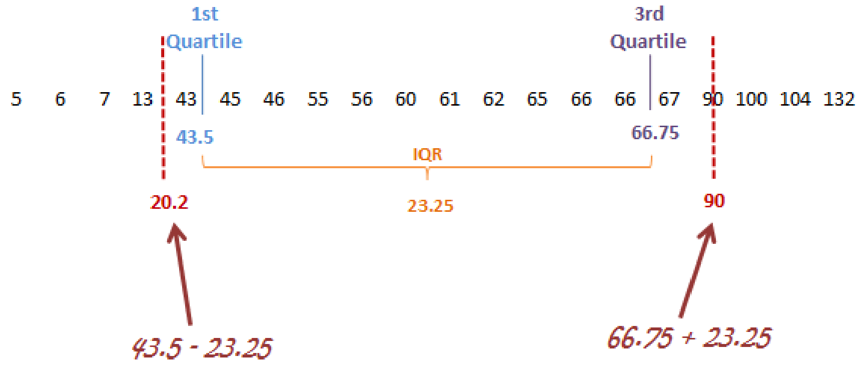
- BeExcellence
- Chandoo Excel Tips
- Chris Web BI Blog
- Clearly and Simply
- Contextures
- Daily Dose of Excel
- Excel Campus
- Excel TV
- ExcelGuru
- PowerPivot Pro
- PTS Blog
- SpreadsheetGuru
- Your Sum Buddy

establish a “fence” by extending the “reasonable” range using these calculations:

1st Quartile – IQR

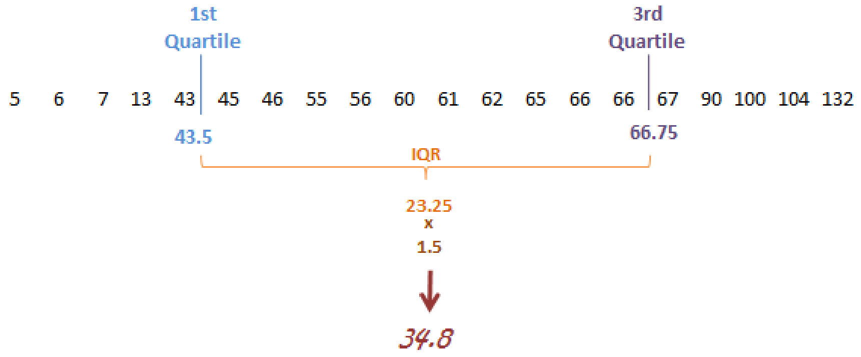
3rd Quartile + IQR

As you can see in our example, these calculations point to a new set of values which we call our Lower Fence (20.2) and our Upper Fence (90). Any values outside the fences are theoretically outliers.



.

That being said, applying fences based on the raw IQR would aggressively tag too many values as outliers. So Tukey established a 1.5 basis, stating that the IQR should be multiplied by 1.5.

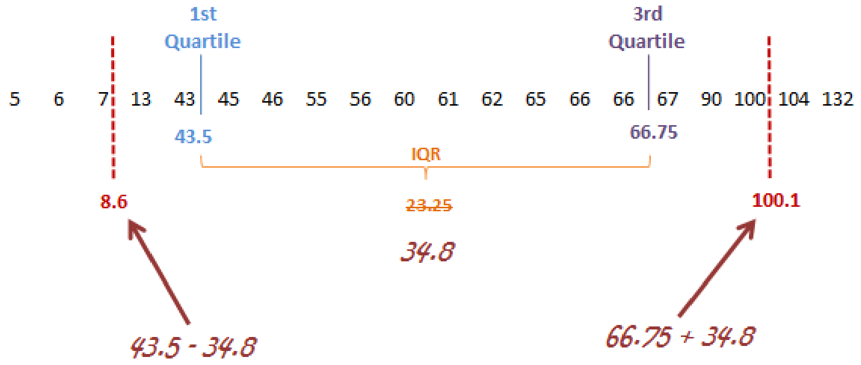


.

This expands the “reasonable” range and reduces the amount of outliers to a more appropriate set of values. With Tukey’s method, outliers are:

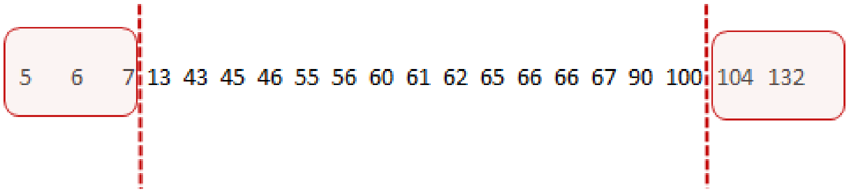
values below **(Quartile 1) – (1.5 × IQR)**

values above **(Quartile 3) + (1.5 × IQR)**



.

In our example, we see that 5, 6, 7, 104 and 132 are statistical outliers.



.

There doesn’t seem to be any statistically-driven reason Tukey uses 1.5 as a hard basis for his method. In fact, if you wanted to be more conservative, you could use 3 x IQR to identify the “extreme” outliers.

Extreme outliers are :

values below **(Quartile 1) – (3 × IQR)**

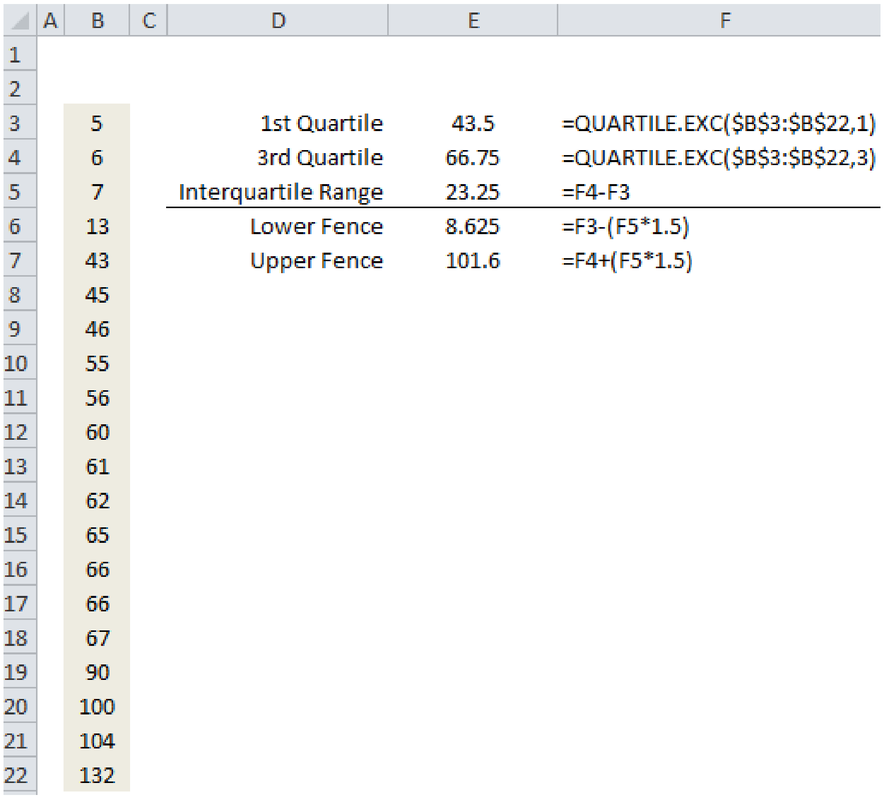
values above **(Quartile 3) + (3 × IQR)**

.

.

Using Tukey’s Method in Excel

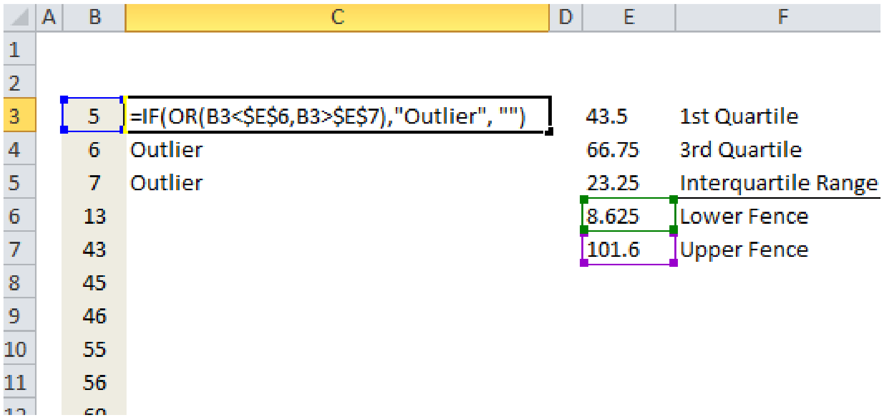
Tukey’s method is easy enough to apply in Excel. Calculate the needed values…



	A	B	C	D	E	F
1						
2						
3		5		1st Quartile	43.5	=QUARTILE.EXC($B$3:$B$22,1)
4		6		3rd Quartile	66.75	=QUARTILE.EXC($B$3:$B$22,3)
5		7		Interquartile Range	23.25	=F4-F3
6		13		Lower Fence	8.625	=F3-(F5*1.5)
7		43		Upper Fence	101.6	=F4+(F5*1.5)
8		45				
9		46				
10		55				
11		56				
12		60				
13		61				
14		62				
15		65				
16		66				
17		66				
18		67				
19		90				
20		100				
21		104				
22		132				

…then use those values in an IF statement to tag the outliers.

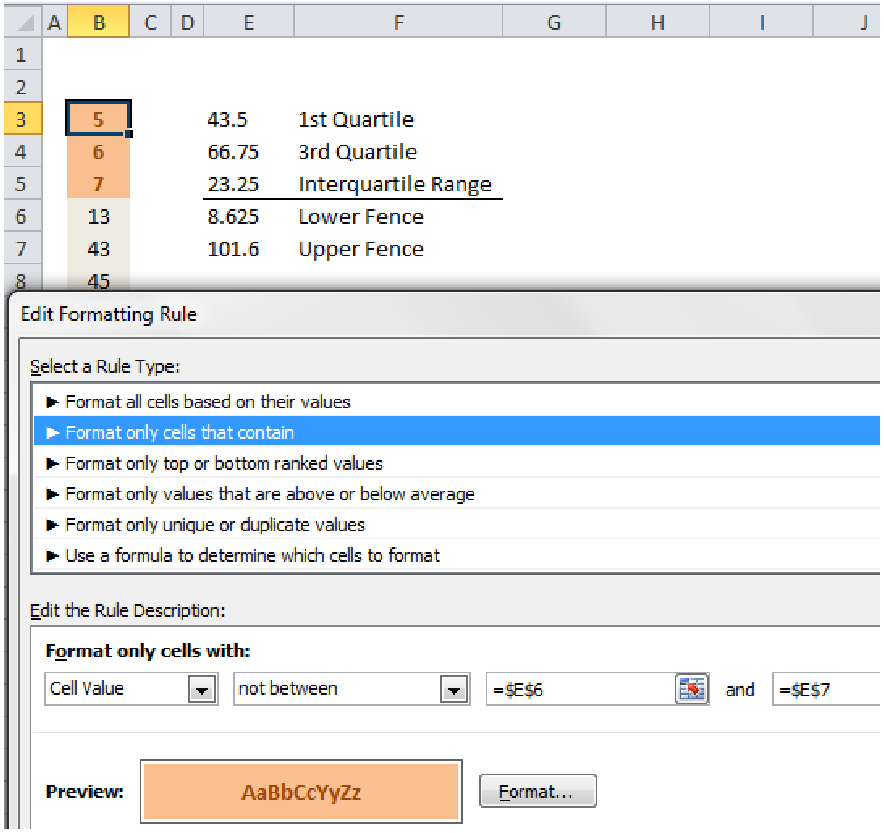
IF the value is less than the Lower Fence OR greater than the Upper Fence, then tag it as an outlier.



	A	B	C	D	E	F
1						
2						
3		5	=IF(OR(B3<$E$6,B3>$E$7),"Outlier", "")		43.5	1st Quartile
4		6	Outlier		66.75	3rd Quartile
5		7	Outlier		23.25	Interquartile Range
6		13			8.625	Lower Fence
7		43			101.6	Upper Fence
8		45				
9		46				
10		55				
11		56				

You can even get fancy and apply conditional formatting to highlight the outliers.

Simply create a rule that colors Cell Values not between the Lower Fence and Upper Fence.



	A	B	C	D	E	F
1						
2						
3		5			43.5	1st Quartile
4		6			66.75	3rd Quartile
5		7			23.25	Interquartile Range
6		13			8.625	Lower Fence
7		43			101.6	Upper Fence
8		45				

Edit Formatting Rule

Select a Rule Type:

- ▶ Format all cells based on their values
- ▶ Format only cells that contain
- ▶ Format only top or bottom ranked values
- ▶ Format only values that are above or below average
- ▶ Format only unique or duplicate values
- ▶ Use a formula to determine which cells to format

Edit the Rule Description:

Format only cells with:

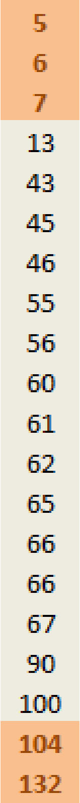
Cell Value | not between | =$E$6 and =$E$7

Preview: AaBbCcYyZz

Once you identify outliers, what do you do – delete them? mmm….No.

It's typically not good to simply delete any values in a legitimate data set.

Outliers could be a result of data entry error, changes in business rules, or some other random variation of factors. Instead of removing them, it's best practice is to investigate why they exist, where they came from, and what they mean to the analysis.



5
6
7
13
43
45
46
55
56
60
61
62
65
66
66
67
90
100
104
132

.

Wow….I just remembered that I used to be cool.

Business Statistics, Excel, Formulas

← Copy Just the Text from a Web Page

Using Power Query to Combine Data from Multiple Excel Files into One Table →

42 thoughts on "Highlighting Outliers in your Data with the Tukey Method"

1. Michael

   January 16, 2014 at 5:27 pm

   I enjoyed the post. Thanks for sharing your tips on practically identifying potential outliers!

2. damich

January 16, 2014 at 5:57 pm

Nice job explaining the statistics. I may have enjoyed statistics class if it was more like this. I hope you do more of them.

I also hope Mrs. Pig doesn’t read these posts, unless she’s into that whole geek thing. I’d say that I hope your kids don’t read them, but they already know how uncool you are.

3. AlexJ

January 16, 2014 at 7:00 pm

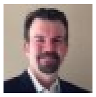
You ARE cool. And rather than using a retainer, you should be ON retainer. Thanks for a very good piece.

4.

Lukas

January 17, 2014 at 12:08 am

I’m saving this bad boy for Saturday night! I was never cool.

5. 

Rick Grantham

January 18, 2014 at 12:26 am

Interesting read. As I was going through it, I was wondering… Why the he** are they multiplying the IQR times 1.5? Whats the point?

Was glad to read your note… “There doesn’t seem to be any statistically-driven reason Tukey uses 1.5 as a hard basis for his method. In fact, if you wanted to be more conservative, you could use 3 x IQR to identify the “extreme” outliers.”

In my mind, Six Sigma and statistically driven analysis are the next steps in the evolution of BI (and hopefully Excel). I would typically need to use MiniTab or a similar program to perform this type of analysis. Thanks for providing some shortcuts here. Its appreciated.

Next… if you could dumb down multi-variate regression or ANOVA in Excel, I would appreciate it 🙂

Rick Grantham

6. Robert

January 19, 2014 at 2:24 am

Interesting article. I typically use the ZScore, or standard score, to identify outliers. I will see how this stacks up.

I use a UDF to simplify this process of identifying outliers based on the ZScore, and as such, have made a UDF based on your article. The UDF is as follows:

```
Option Explicit

Function Tukey(Rng As Range, Rng2 As Range) As String
Dim Quart1 As Double
Dim Quart3 As Double
Dim Lower As Double
Dim Upper As Double
Dim Interquartile As Double
Quart1 = Application.WorksheetFunction.Quartile_Exc(Rng2, 1)
Quart3 = Application.WorksheetFunction.Quartile_Exc(Rng2, 3)
Interquartile = Quart3 – Quart1
Lower = Quart1 – (Interquartile * 1.5)
Upper = Quart3 + (Interquartile * 1.5)
If Rng Upper Then Tukey = "Outlier"
End Function
```

Using the UDF above, the syntax to see if cell B4 is an outlier would be as follows:

```
=Tukey(B4,$B$3:$B$22)
```

If an outlier, it will say "Outlier" in the appropriate cell, otherwise, it will be blank.

7. Robert

January 19, 2014 at 2:28 am

in the above UDF, the one line should read

```
If Rng Upper Then Tukey = "Outlier"
```

8. John Walkenbach

January 21, 2014 at 3:49 am

Good article, Mike. But you missed a huge comedic opportunity by not posting this on Thanksgiving.

9.
Jeff Weir
January 22, 2014 at 4:27 am

Very funny, John.

10. Doris Choo
January 26, 2014 at 5:02 am

Thanks for interesting forecasting lesson 101. Most of all Mike's great sense of humor comes thru these very dry stuff, making it fun and enjoyable.

11.
Patricia
January 27, 2014 at 4:52 pm

Enjoyed your article and your sense of humor. The good news is that you wrote it on a Thursday and not a Saturday so you do have a life 🙂

12.
Bob
February 4, 2014 at 8:25 pm

Bacon Bits & Tukey. All that is missing is some toast & mayo… Great post.

13.
Rick
February 6, 2014 at 9:25 pm

Great article. I'm thinking about using it to determine outliers in student grades, but assuming the data is correct, is there even such a thing?
Also, second time this week I saw the use of the OR() function. Great UDF Robert.

14. Jeff
February 13, 2014 at 12:14 am

In my range the lower fense was a negative number. Makes me question the validity of this method. The 1.5 multipler seems to be the issue.

Now my range was 15 to 1500 and there were only 80 data points. Looking at the data you can see 15 stands out as an outlier but this negative lower fence suggests otherwise.

15. datapig Post author

February 13, 2014 at 5:50 pm

Jeff:
Your lower fence can absolutely be a negative number. If range is 15 to 1500, there are not only 80 data points. You see 80 numbers, but there are actually 1,485 statistical data points (every number from 15 to 1500) .

Remember Statistics involves the numbers between the ones you actually see as well.

Your lower and upper fences depends on the skew of your data. So the lower fence could absolutely be a negative.

16. Sandeep Singh

November 7, 2014 at 5:06 pm

One question : Do we need to arrange the data in ascending or descending order befoer applying the above method ?

17. Chris

March 25, 2015 at 7:11 pm

Just like Jeff above, using my data set on corn yields my lower fence was a negative number; yet when looking at the data it seemed clear that there were outliers skewing my data at the lower end. Also, a negative number isn’t possible in my real word. You can’t harvest a negative number of bags of corn from an acre; the least possible is ‘0’.

18. Melissa B

March 30, 2015 at 1:58 am

This was so incredibly helpful! I'm working on homework for a statistics class and could not wrap my head around how to find the outliers, but now I can 🙂

19. J Miller

July 30, 2015 at 4:16 pm

I am attempting to use this concept on some data but my lower fence is resulting in a negative number. I checked the formula and have everything correct. Any idea what direction I should take based on this result? I really need an efficient method of identify outliers in this massive data set. One thing I had to do was put my data into a pivot table and then referenced the cell range of the pivot table to find the quartiles. Not sure if that is causing any issues. I organized the data so that the data points did not aggregate together but are individual data points.

20. datapig Post author

July 30, 2015 at 5:06 pm

The lower fence can absolutely be a negative number. If your dataset contians many data points that are negative, this should be expected.

It just means your lower fence outliers are those that skew more negative than your dataset sample.

21. J Miller

July 31, 2015 at 6:25 pm

Well I actually have NO negative values in my data, so the fact that the lower fence is negative doesn't help to remove outliers below that number.

22. datapig Post author

July 31, 2015 at 9:24 pm

Jim: Then I would check to see if you didn't transpose the Interquartile range calculations.

IQR = (Quartile 3) – (Quartile 1)

Lower Fence = (Quartile 1) – (1.5 × IQR)
Upper Fence = (Quartile 3) + (1.5 × IQR)

23. J Miller

July 31, 2015 at 9:39 pm

I thought the same thing and checked the formula more than once. My Q1 is .21 my Q3 is .65 and my IQR is .44. So my Upper Fence = 1.3 and my Lower Fence = -.4

Any other thoughts on how to remove outliers?

24. wf

September 12, 2015 at 11:34 pm

Did Tukey also recommend the 3*IQR for extreme outliers, or was that introduced somewhere else later?

25. Joy

December 11, 2015 at 12:58 am

This was an excellent explanation please keep up the great post!

26. Alison

February 25, 2016 at 8:23 am

Even if i dont use the 1.5 multiplier my lower fence is lower than the lowest number in my data set.

J Miller did you figure out how to remove the outliers at the bottom end of your data set?

I'm looking to remove the bottom and top 10% of my numbers…..

27. David H

May 21, 2016 at 7:48 am

Mike, Great article. For those folks who are interested in different statistical methods of calculating outliers, take a look at this article:

A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets by Songwon Seo http://d-scholarship.pitt.edu/7948/1/Seo.pdf

The author discusses and compares several different methods of calcuating outliers, including Tukey, Z score, and modified Tukey.

28. Robert

August 8, 2016 at 12:17 am

“Next, we subtract the 3rd Quartile from the 1st Quartile.” Don’t you mean we subtract the 1st Quartile from the 3rd Quartile? I can’t quite see why we would want a negative value.

29. Pingback: Analizando los sueldos de la industria del software en Argentina (Parte 1) – sysarmy

30. Pingback: ???????? – Knowledge Discovery

31. Pingback: Project: Creating Customer Segments- Part 3 – One moment please..

32. RM

January 7, 2017 at 3:46 am

Hi,

Kinda new to all this stuff – I was wondering why the first quartile (quantile p=0.25) is 43.5 when it is between 43 and 45?

I calculated the quantiles in R and it says that the 1st quartile is just 44 – the mean of 43 and 45.

Is there some nuance to quartiles I’m missing here?

33. Giovani Ferreira

January 24, 2017 at 5:35 am

Amazingly written clear piece of information.
I’m really glad I found it.
Thank you for helping out other uncool retainer people.

34. John Park

February 5, 2017 at 4:28 pm

thanks~

your explanation make me understand about eliminating outliers 🙂

35. Pingback: How to Deal with Outliers in Your Data

36. Pingback: Customer Segmentation – A formula of maximum generality

37. Zhou Xiaorui

    March 24, 2017 at 12:13 pm

    Thank you! It’s really comprehensible.

38. Erinma Mary

    April 18, 2017 at 8:09 am

    Please, if i may ask. In what year was Tukey test proposed and by who (full name)

    Please help me out.

39. Muhammad Magdi

    April 21, 2017 at 10:42 pm

    Thanks a lot!

40. Pingback: Machine learning – Ideasinplain

41. Pingback: Customer Segmentation for Grocery Wholesaler - Schuman's Tech Blog

42. Rebecca Freddo

    April 13, 2018 at 8:20 am

    Well written and very informative. Thanks.

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

post comment

□ Confirm you are NOT a robot

Default footer text