Fundamentals of Analysis

Fundamentals of Analysis

Written by: Matt David
Reviewed by: Dave Fowler

Table of Contents

Choosing Metrics for Product Launch

You're about to launch a new feature or product. You need to be able to measure its impact so that you and your company can learn from it.

This article is intended to give you a framework for prioritizing which metrics matter

1. State your default position
2. Identify your Business Drivers and Levers
3. Determine Metrics that would change your default position

State Your Default Position

What would you do if you didn't have any metrics about the impact of your product or feature? We need to start from this position so that we know what it would take for us to change from this path

Think through what you would do about marketing, customer support, user flows, product plans without any more information than what you have right now. This is your default position, your current plan. It should include several departments and the status quo. It's helpful to write this down on paper.

For Example:

- Marketing: spend $10,000 on paid acquisition
- Customer Support: 2 employees manning ZenDesk between 9am and 5pm EST
- Product Plan: follow the current plan to release 2 new features every month
- Engineering: continue with current headcount, producing sprints at the current pace

This plan sounds great, so what sort of information would change your mind?

Identifying Your Business Drivers and Levers:

Business Drivers = the short list of measurable properties that leads to success or failure

When we launch products, make changes, hire people, it's all to drive the business forward. All businesses have a few key drivers of profitability.

- Cost of goods
- Revenue of goods
- Volume of goods sold
- Price of goods

- Number of paying customers

If one of these drivers change it will likely influence our default position. If marketing spend starts driving up the volume of goods sold you may invest more in it.

Business Levers = the changes you can make to influence your drivers

Some drivers are more easily changed than others. For example, we can more easily change our product's prices than we can control the costs of goods or control the number of people willing to pay these prices.

In a business, its capacity to influence its business drivers can be viewed as a series of levers. For example, the number of paying customers is affected by sales and marketing efforts. So one way a business can affect customer number is by investing more in sales and marketing. They can pull the lever from "the current sales effort is sufficient" to "we will invest more in sales efforts".

Department	Drivers This Department Affects	Levers This Department Owns
Marketing	Perceived value of productVolume of new leads	Marketing budget and channelsBranding
Sales	Number of paying customers	Sales team staffingSales strategy
Product	Price of goods	Product roadmap Sprint planning
Engineering	Cost of goods	Sprint planning

Ask Specific, Action-Oriented Questions

We need to map the right lever to the right driver. We must go beyond asking generic questions like "How do I get more customers?" to questions that deal in specific levers attached to a key drivers:

- "Which channels have the best conversion rate for customers?"
- "What is the ratio of earnings vs customer acquisition costs per channels so that I can adjust my strategy once this data is available?"

We want to find out what levers drive the business the most.

With generic questions like "how do I get more customers?", there are too many levers that come to mind so their effectiveness can become confused. Do more marketing and sales, is not helpful, we need to get more specific. Questions like "which channels should my marketing budget be spent on to get most paying customers for marketing dollar spent"? Focuses us on the lever of marketing channels.

Question	What business driver does this address?	If I knew the answer, what lever could I pull?	Is this a good question?
How do I get more customers?	Volume of paying customers	It depends on the answer… maybe I can act on the answer, maybe not.	Too vague…
Which marketing channels	Volume of	I could spend my marketing	Much

have the best conversion rate for customers?	paying customers	dollars differently	better
What is the ratio of customer revenue vs customer acquisition costs per channels?	Unit economics	I could invest in more profitable marketing channels or find ways to adjust my prices	Much better!

Looking at your current plan, what specific, action oriented questions do you have about it?

Indicate and enumerate them so that you have a list of specific, action oriented questions to look at. This will help us determine metrics in the next section.

Department	Default Position	Specific, Action-Oriented Question	If I knew, I would change my default position by pulling which levers?
Marketing	Spend $10,000 on paid acquisition	Which channels have the best conversion rate?	I could increase my budget on the best channels
Customer Support	2 employees manning ZenDesk from 9am - 5pm EST	How many tickets get missed because there wasn't a 3rd person available?	I could hire another employee
		How many tickets get missed because we don't have more hours of coverage?	I could hire a support service
Product	Follow the current plan to release 2 new features every month	How many features were released with bugs that forced them back into development?	I could adjust sprint planning to accommodate for more planning and iteration time.

Up to this point, we've not talked at all about metrics specifically, but you can see how this list of questions and "if I knew" statements can guide us to picking the right metrics.

Metrics That Would Change My Default Position

Let's focus in on the default position, the question, and the levers columns. Are there metrics that clearly answer the question and would change your position by telling you which direction to pull a lever?

Default Position	Specific, Action-Oriented Question	If I knew, I would change my default position by pulling which levers?	Metrics
Spend $10,000 on paid acquisition	Which channels have the best conversion rate?	I could increase my budget on the best channels	Conversion rate per channel
2 employees manning	How many tickets get missed because there	I could hire another employee	Missed tickets and ticket wait time

ZenDesk from 9am - 5pm EST	wasn't a 3rd person available?		between 9am and 5pm
	How many tickets get missed because we don't have more hours of coverage?	I could hire a support service	Number of tickets received outside of business hours
Follow the current plan to release 2 new features every month	How many features were released with bugs that forced them back into development?	I could adjust sprint planning to accommodate for more planning and iteration time.	Ratio of work units spent on feature planning vs development vs fixing issues over time

Once again, these metrics allow the business to make decisions about when and how to pull certain levers. And not all metrics are simple to construct. For example, there could be much more debate over what counts as "fixing issues" vs planning, or what conversions matter. But if you put in the work, you will find that your decision-making, company alignment, and outcomes will all improve over time.

Summary

In conclusion, companies and teams that put in the time to visualize potential outcomes and default actions, determine what matters to the business, and decide what can be changed under which circumstances will reap significant benefits.

There is no shortage of important decisions for product companies approaching a launch. And with every decision, it can feel like the stakes could not be any higher, but whenever possible remind yourself that your company's success has more to do with the overall quality of data-driven decisions made than any single analysis. Whether your next big data-driven decision comes in a quarter or in a week, take care to improve your process and appreciate the clarity, growth, and excitement that comes from a decision well made.

- State your default position - What would you do without data
- Identify your levers - What could you change if you did have data
- Determine metrics that would change your position - How much would that metric need to change in order to pull a lever

What is Data and Why Should I Care

When I was little, I went to a friends house for a play-date schedule by our mothers. As we were running around, playing with Nerf guns and foam swords, I stumbled upon a mysterious relic that I'd never seen before. Etched on the frame of the door were dozens of mysterious markings. Upon closer inspection I realized they were pencil marks - squiggly horizontal lines followed by a few letters and numbers. I asked my pal what I was looking at, to which he proudly replied: "It's the height of my family!"

I stared at the wall in wonder, realizing that the shortest marking on the wall was an indicator of the height of my beloved friend when he was only 5 years old. He was tiny! I then tilted my head back to see the highest marking, way above both our heads. "That's my cousin. She's a giant."

While it might just seem like a cute way to build family memories, rituals like measuring yourself at regular intervals is also a way to connect the narratives of your past to your present.

How Tall Were You As A Kid?

If I asked you right now whether you would consider yourself a tall kid when you were growing up, you'd probably consider a whole bunch of things before answering.

What do I mean by tall?

Do you mean a specific age, or over my entire childhood?

Tall compared to my classmates, or the entire population of kids?

While it might seem like a simple question on the surface, it's actually an impossibly difficult question to accurately answer without knowing more.

But, if your parents had a soft spot for tradition (and were willing to scuff up their walls), then you could tell me exactly how tall you were at a given time! You could also tell me the exact year that your growth spurt occurred, rocketing you into the "tall kid" category.

Those pencil scratches on the wall are observations that don't have any value, until a question is asked that they can help solve.

Information is only valuable if we're asking the right questions. Data is no different. It can be extremely useful by telling us how healthy our business is, what our customers are thinking, or where production bottlenecks are occurring. But the answers to those questions can only happen if we proactively do two things:

1. **We've taken the time to record observations of our surroundings.**
2. **We ask a specific question to get a specific answer.**

That's all there is to it!

If you are fine with general answers, like: "Well, I was the third tallest in first grade, just behind Tanya, but then she moved to Washington which made me second tallest after Mariah…" then don't worry about collecting any data.

But if you need your answers to be a little more solid than that, then you'll need to begin collecting data and asking the right questions.

Here are my top three tips if you're interested in taking the first step on your data journey, but are unsure of where to start:

1. Itch your own scratch

You know your needs better than anyone else. If you need a certain question answered on a daily basis, but find it difficult to get straight answers, begin with that. Ask yourself "What do you need to know in order to make a decision?".

Pro Tip: Keep it super simple to begin. You might have a thousand questions you need answering, but start with just one. This will help you to understand the anatomy of an answer and be able to replicate the process more effectively in the future.

2. Use human units

What I mean by this is to make the answer to your question as humanistic as possible. This means rounding answers to the nearest whole, use time frames that we can comprehend (like number of times per hour), and don't worry about being too specific. Pretend like you're answering your closest friend when they ask you a question. This will help you get comfortable with making judgement calls with data as a partner, rather than relying completely on data to give you the answer.

Pro Tip: When in doubt, think about your five senses (hearing, smelling, seeing, touching, tasting). What information from each of those is necessary to make a decisions and how should it be measured?

3. Apply Context

Our brains work better when the information we're presented realties to our world around us. Use techniques like comparison, over time, or "equivalent to" as ways to help yourself see your answer clearly. The statement "I grew 5 cm taller since I was last measured" has a very different meaning if you're 10 or 50 years old.

Pro Tip: You know you've applied enough context to a question if the answer can be said in one sentence. Ex. I am 166 cm tall as of September 5, 2019.

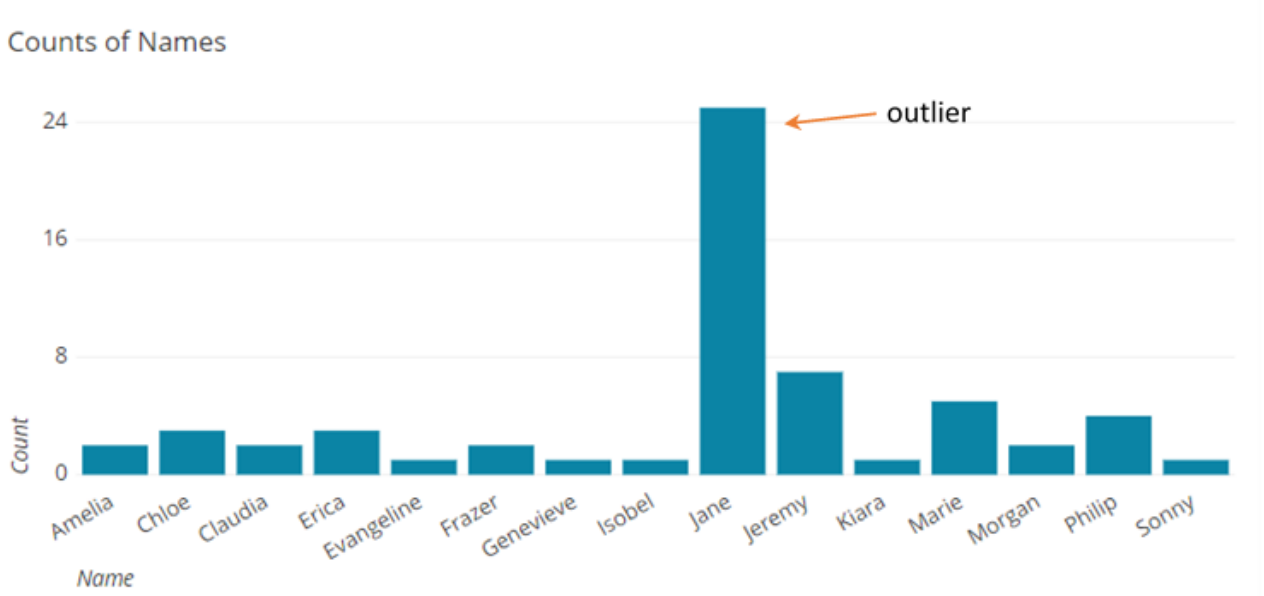
Keeping those tips in mind, you're now ready to get out and begin playing around with data. Don't worry if it gets confusing, practice will make perfect over time. Much like learning to read, it takes practice to begin working with data. Don't forget to celebrate the small wins and don't get discouraged if you get confused. I'm only a tweet away @stew_hillhouse if you've got any questions.

Stewart Hillhouse is a contributing writer to Data School. He's the co-founder of GoDo, a data experience studio. To get more, check out Connect With Data, his blog helping to demystify the complexities of data.

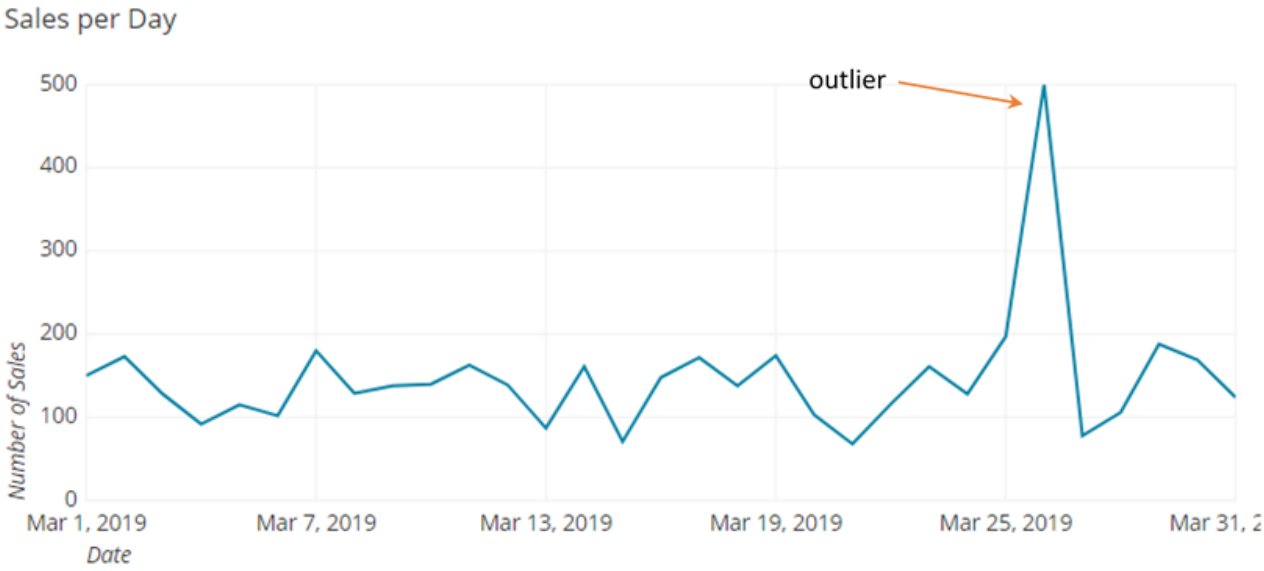
What is an Outlier?

An outlier is a value or point that [differs substantially from the rest of the data](#).

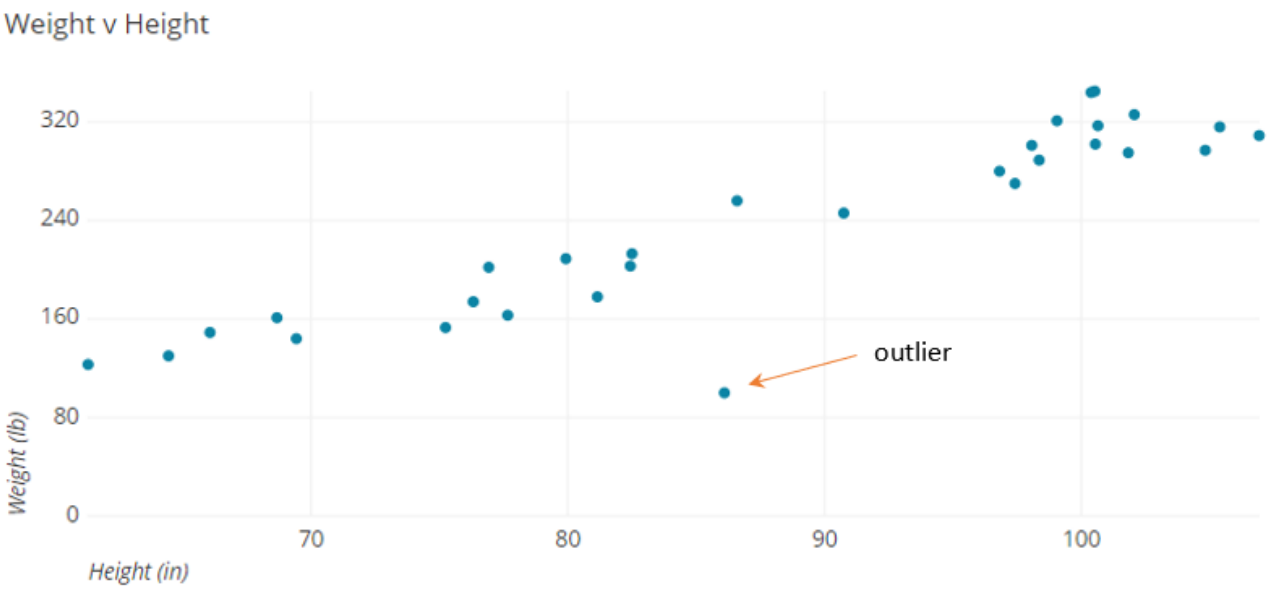
Outliers can look like this:



This:



Or this:



Sometimes outliers might be errors that we want to exclude or an anomaly that we don't want to include in our analysis. But at other times it can reveal insights into special cases in our data that we may not otherwise notice.

For example, in our names data above, perhaps the reason that Jane is found so many more times than all the other names is because it has been used to capture missing values(ie Jane Doe).

There is not a hard and fast rule about how much a data point needs to differ to be considered an outlier. As a result, there are a number of different methods that we can use to identify them.

Use of Domain Knowledge

resting_blood_pressure
145
130
130
120
120
140
140
120
172
150
140
130
130
110
150

Sometimes, the typical ranges of a value are known. For example, when measuring blood pressure, your doctor likely has a good idea of what is considered to be within the normal blood pressure range. If they were looking at the values above, they would identify that all of the values that are highlighted orange indicate high blood pressure. As a result, they may advise some course of action.

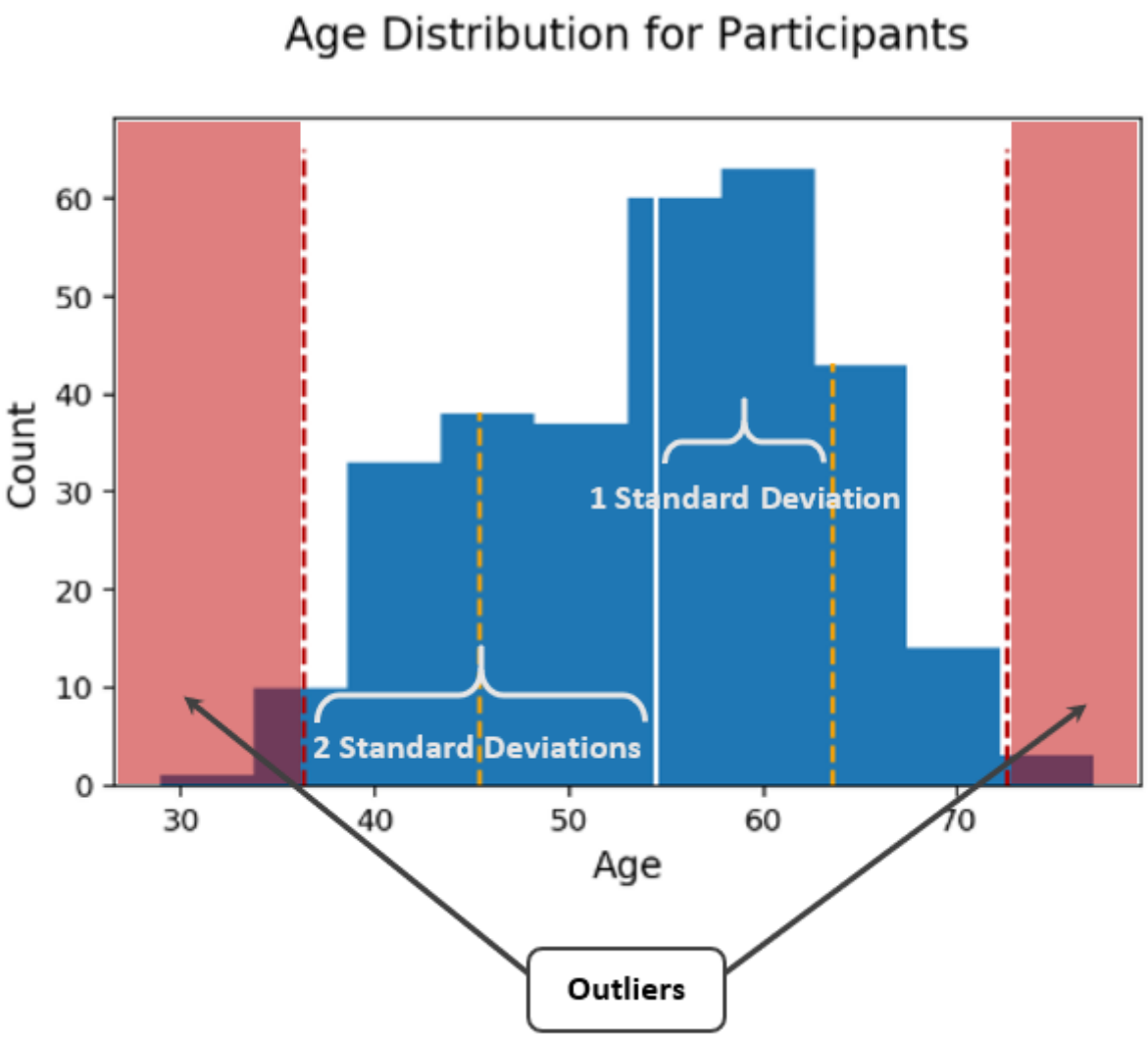
In this case, “outliers”, or important variations are defined by existing knowledge that establishes the normal range. It might be the case that you know the ranges that you are expecting from your data. If you identify points that fall outside this range, these may be worth additional investigation.

Statistical Indicators

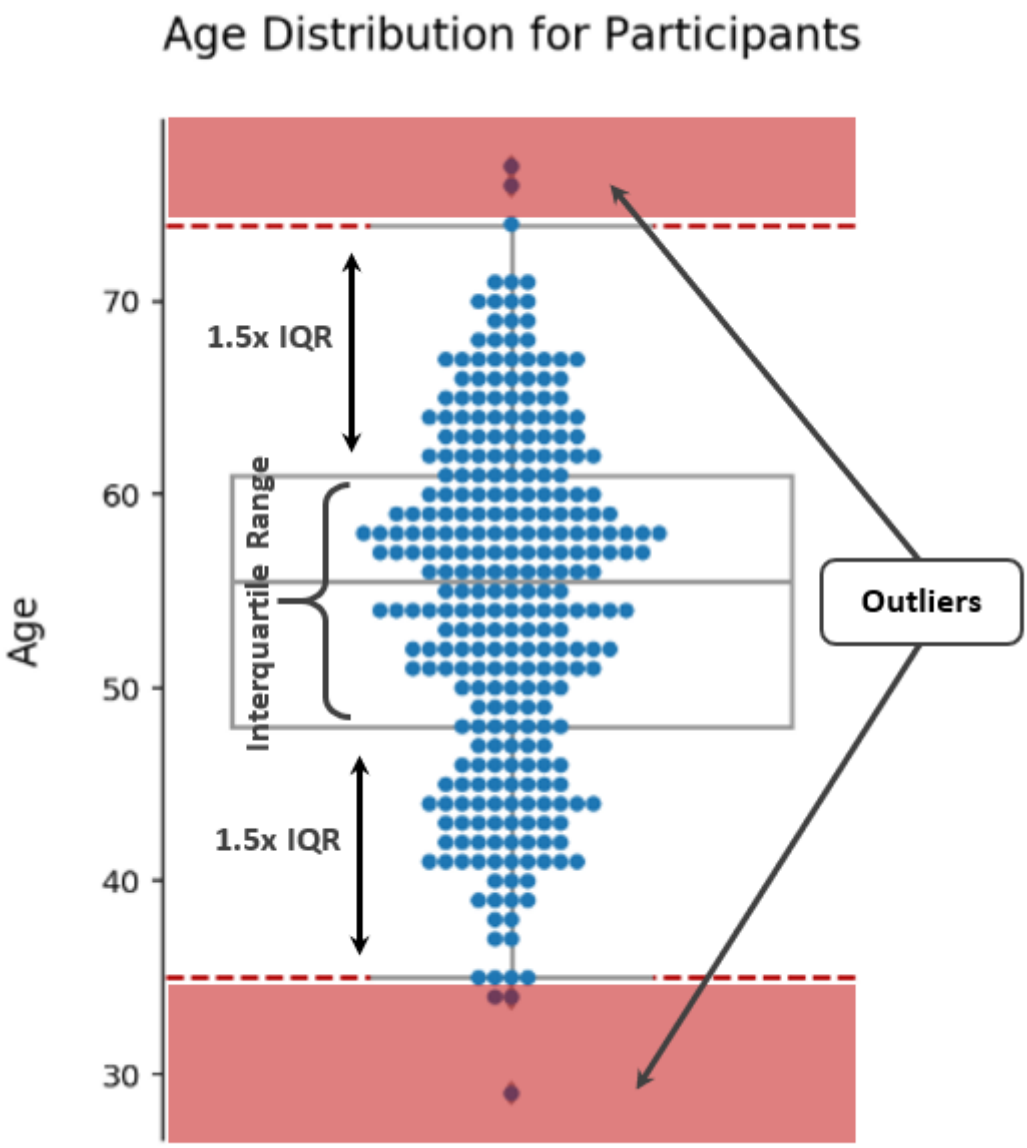
When using statistical indicators we typically define outliers in reference to the data we are using. We define a measurement for the “center” of the data and then determine how far away a point needs to be to be considered an outlier.

There are two common statistical indicators that can be used:

1. Distance from the mean in standard deviations



1. Distance from the interquartile range by a multiple of the interquartile range



For the purposes of our exploration, we're going to use the interquartile range, but for more information about using the mean and the standard deviation, you can check out this article.

Why is Finding Outliers Important?

Ensure Data Quality

One of the reasons we want to check for outliers is to confirm the quality of our data. One of the potential sources for outliers in our data are values that are not correct. There are different potential sources for these "incorrect values". Two potential sources are missing data and errors in data entry or recording.

Code for missing data

At times, when values are unknown, the person entering the data might use a value to indicate this. Some examples include:

- **Numeric values**: If there are values that are known to be outside of the expected range of values, these can be used to indicate missing values. Examples include:
    - 0
    - 9999
    - -9999
- **String values:** Often a repeating single character, punctuation, or specific words can be used for missing or unknown string. Examples include:
    - xxxx, aaa, yy
    - ., ,, ?, *
    - Unknown, Unspecified, Missing
- **Dates:** Dates are typically either the date of an event or a person's birth date. Dates that cannot be a true date can be used for missing values. For humans, this is usually dates that make the person's age impossible. For events, these can be dates before the event/activity began, or very far in the future. January 1 is also more common for a missing For example:
    - 1850-01-01, 1900-01-01
    - 2130-01-01, 3000-12-31
- **Names:** Missing names can be coded in similar methods as outlined for strings above, but there are some additional conventions that are often found for names. The names John and Jane Doe have long been used for those whose name is unknown, but other generic terms can often be used, depending on the area of business. Sometimes missing names are captured in a combination of first and last names, so if these are separate fields, it's good to combine them to double check. Missing name fields can include individual or combinations of:
    - Client, Customer, Person, Tenant
    - Man, woman, boy, girl, wife, husband, son, daughter
    - Other descriptive terms specific to the field

For all but the numeric values, often you won't be able to directly sort your data. However, if you complete a grouped count of these fields, it is often easy to identify "default" values. Because most of these are quite unique, if default values are used, they will often have much higher counts. You can quickly identify these counts using this type of query:

```
SELECT field_name, COUNT(*) AS value_counts
FROM my_table
GROUP BY 1
ORDER BY 2 DESC
```

Data entry/recording errors

- **Typos**: If someone is manually entering data, it can be easy to record something incorrectly. Most of the above examples can also be the result of typos, but some others include:
    - Digits in name fields
    - Misspelled words
    - Adding extras of the same character

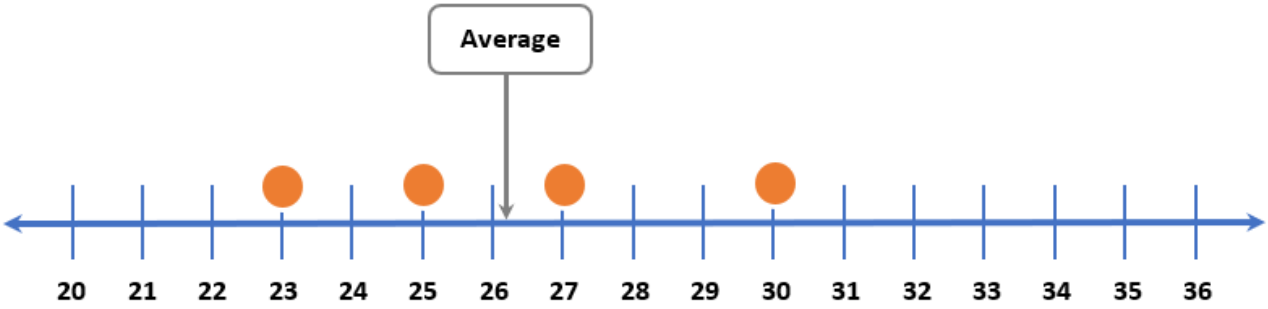
- **Incorrect units:** If different people are recording data, sometimes the information could have been recorded correctly, but a different unit of measure is used. Identifying these types of errors typically requires some knowledge of the expected range of values of values and can be trickier to identify. Some examples include:
    - Weights records in pounds and kilograms
    - Distances recorded in miles and kilometers
    - Temperature recorded in Fahrenheit and Celsius
    - Dates recorded in different orders, e.g. MM-DD-YY and DD-MM-YY
    - Times records in different units such as seconds, minutes, hours

If we find data that is in error or is missing, we may attempt to correct this data, or may need to exclude it from our analysis.

Provide Confidence in Analysis

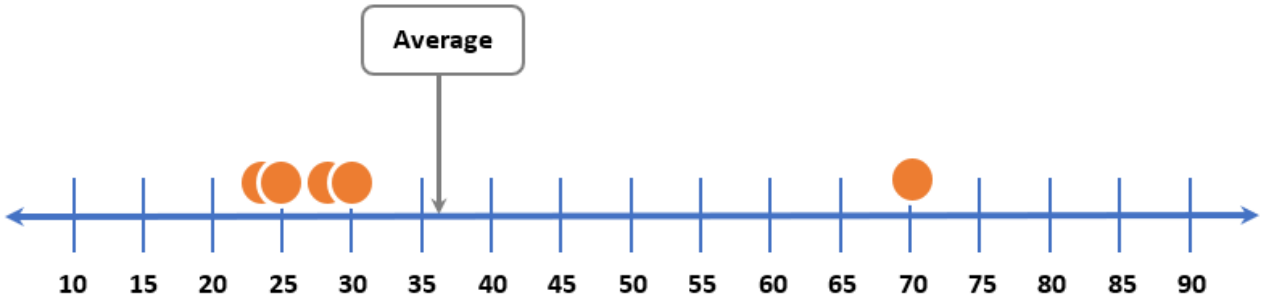
When outliers exist in our data, it can affect the typical measures that we use to describe it.

For example, if we had five friends with the ages of 23, 25, 27, and 30, the average age would be 26.25.



In this case we can have high confidence that the average of our data is a good representation of the age of a "typical" friend.

However, if we then change the value final value and we had friends with the ages of 23, 25, 27, and 70, the average age is now 36.25. This is quite a large increase, even though the majority of our friends are under 30 (mind the change in scale of the graphic).



In this case, we have much less confidence that the average is a good representation of a typical friend and we may need to do something about this.

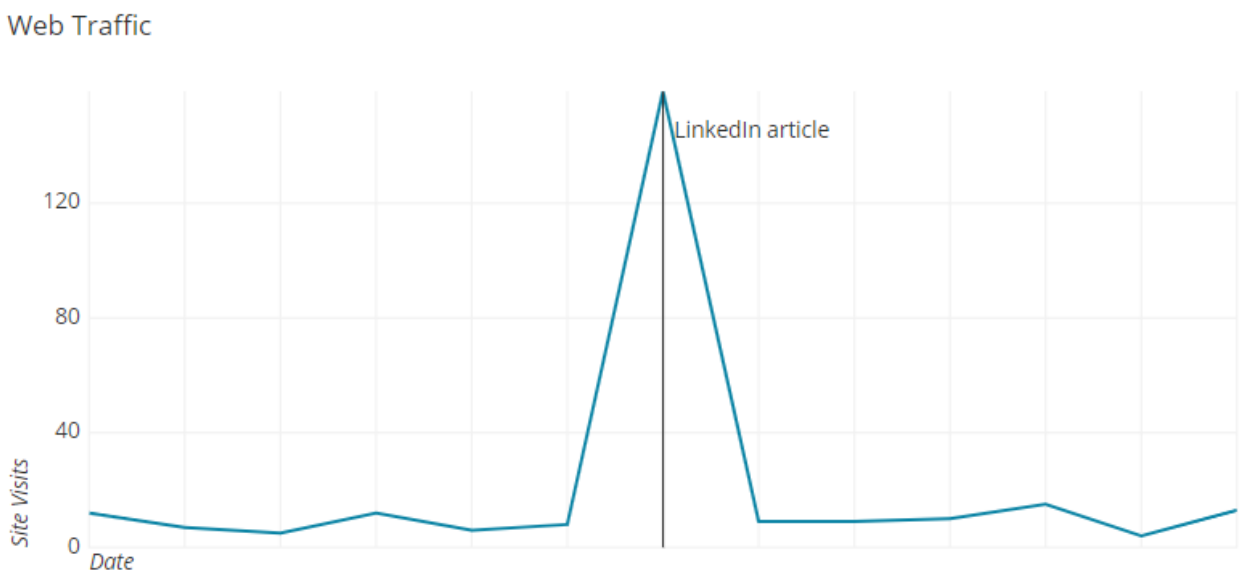
Being able to identify outliers can help to determine what is typical within the data and what are exceptions. If we don't have outliers, this can increase our confidence in the consistency of our findings.

Contextualize the Findings

Identify High Performers

Identifying outliers can also help to determine what we should focus on in our analysis. Sometimes what we wish to discuss is not what is common or typical, but what is unexpected. If results are extraordinarily good, it may be helpful to understand why a particular value is so much better than the rest - is there something that can be learned from this situation that can be applied elsewhere?

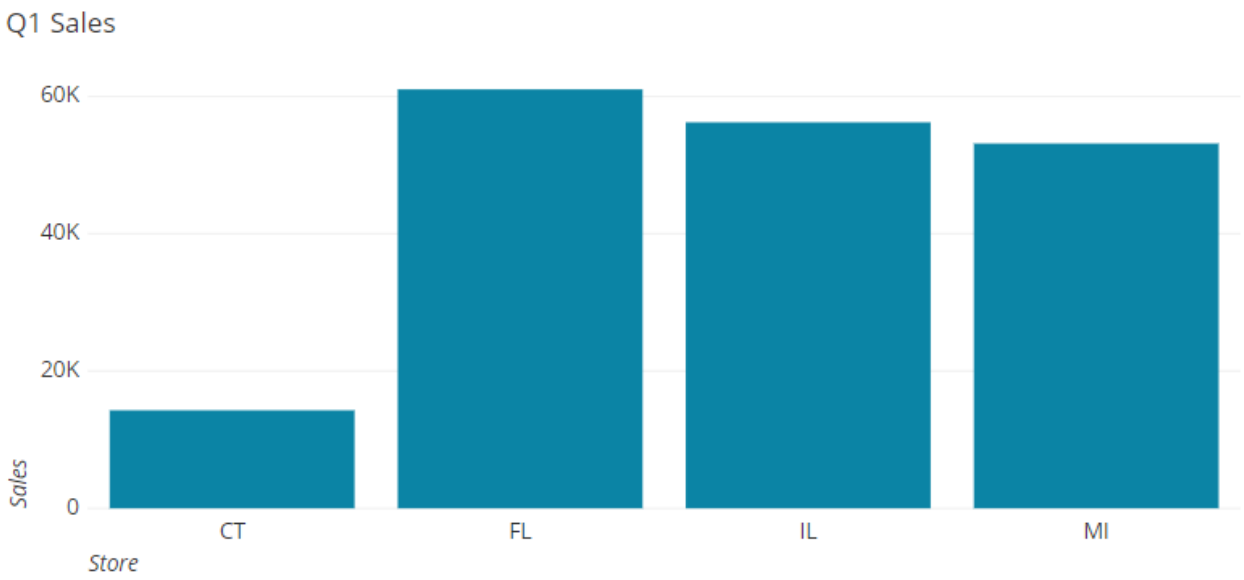
For example, let's say we're looking at our web traffic and we notice that we have some peaks that are much higher than others.



It can be helpful to try to understand the cause of these peaks. Did we start a new ad campaign on that day? Do these peaks always happen when we start an ad campaign? Are there some ad campaigns that have been associated with higher peaks than others? What can we learn from this? When presenting the information, we can add annotations that highlight the outliers and provide a brief explanation to help convey the key implications of the outliers.

Identify Low Performers

If something is particularly poor, it may alert us that there is an issue that needs to be addressed. For example, if you run four stores and in a quarter three are doing well in sales and one is not, this may be something to look into.

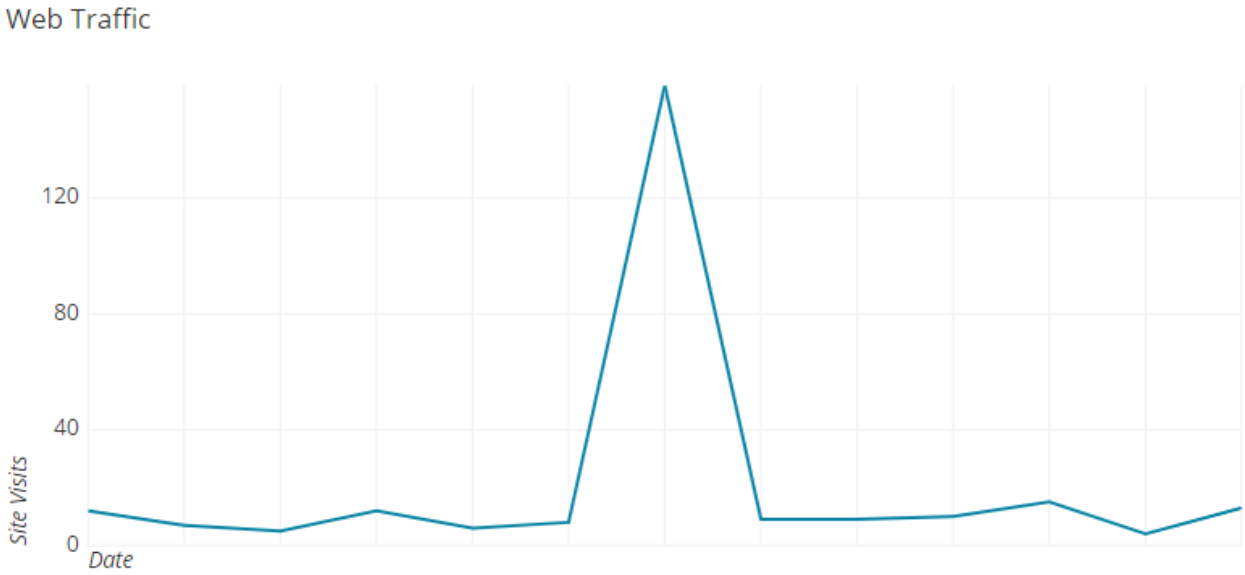


Is this consistent performance for the store? Was there something happening in the local neighborhood, such as construction on the street where it is located, that could have contributed to the lower sales? Are there practices that are implemented in the other stores that could be adopted here? Or, is it that this is a brand new store and it is still building up its customer base?

All outliers are not created equal! If we do identify them it's important to attempt to identify why they may have occurred. This will give us insights into how we manage them.

Visualization

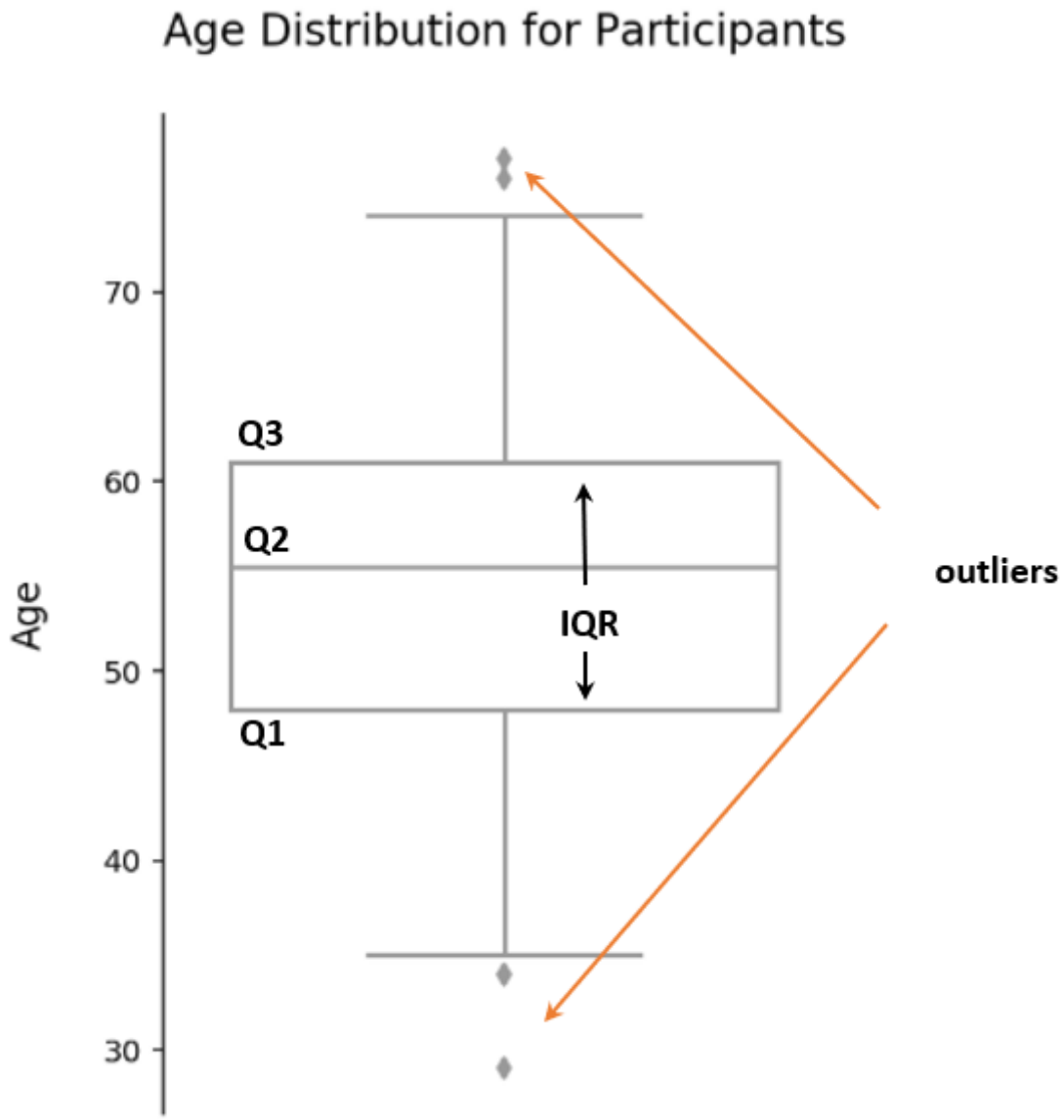
Visualizing data gives an overall sense of the spread of the data. Outliers in visualizations can dramatically skew the visualization making it hard to interpret the rest of the data.



In the above visualization, it is difficult to fully understand the fluctuation of the number of site visits because of one abnormal day.

There are visualizations that can handle outliers more gracefully. One such method of visualizing the range of our data with outliers, is the box and whisker plot, or just “box plot”.

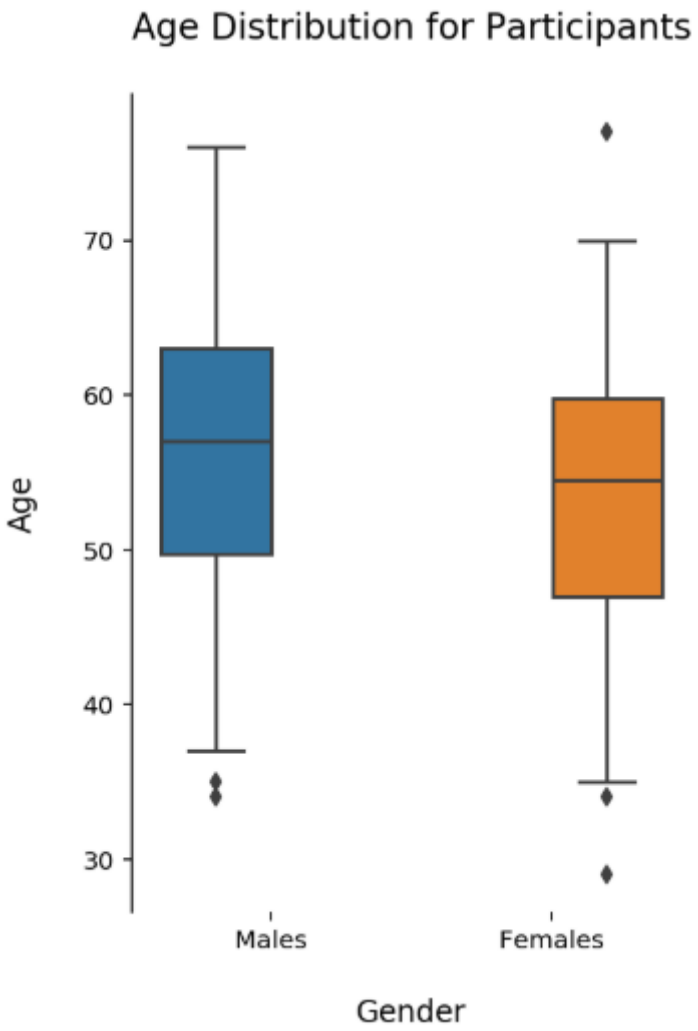
In a box plot we segment our data into four buckets or quartiles. The value that describes the threshold between the first and second quartile is called Q1 and the value that describes the threshold between the third and fourth quartiles is called Q3. The difference between the two is called the interquartile range, or IQR.



The boundaries of Q1 and Q3 create our box, and Q2 or the median is visualized as a line through the box.

From here, we add lines above and below the box, or “whiskers”. To easily visualize the outliers, it’s helpful to cap our lines at the IQR x 1.5 (or IQR x 3). Any points that fall beyond this are plotted individually and can be clearly identified as outliers.

If we want to look at different distributions of outliers we can plot different categories together:



For more detailed information on how outliers are found using the IQR, and how to use this method in SQL, check out these articles:

- What is IQR?
- How to Find Outliers with SQL

Conclusion

By now, it should be clear that finding outliers is an important step when analyzing our data! It helps us detect errors, allows us to separate anomalies from the overall trends, and can help us focus our attention on exceptions. While what we do with outliers is defined by the specifics of the situation, by identifying them we give ourselves the tools to more confidently make decisions with our data.

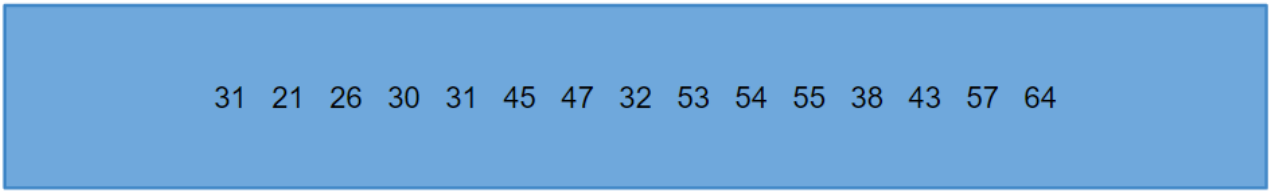
What is the Interquartile Range?

The interquartile range is a widely accepted method to find outliers in data. When using the interquartile range, or IQR, the full dataset is split into four equal segments, or quartiles. The distances between the quartiles is what is used to determine the IQR.

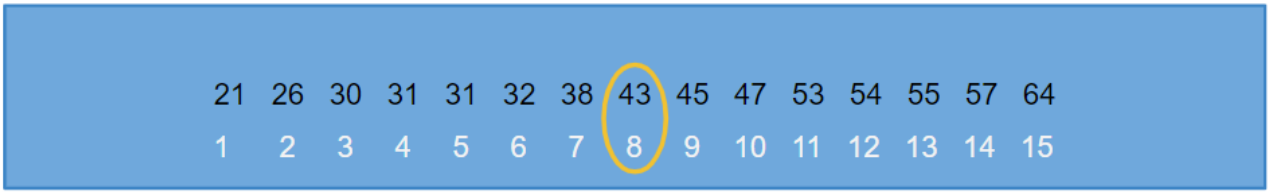
Here's how it works. Let's say that we had a pretty diverse group of 15 friends with the following ages: 31, 21, 26, 30, 31, 45, 47, 32, 53, 54, 55, 38, 43, 57, 64. If we wanted to find the IQR, we would do the following:

1. Order the ages from smallest to largest
2. Find the middle value and create a group above and below this
3. Find the middle value for each group that was created
4. Find the difference between the middle of the top and bottom groups

Let's work through that. First we start off will all of our ages unordered.



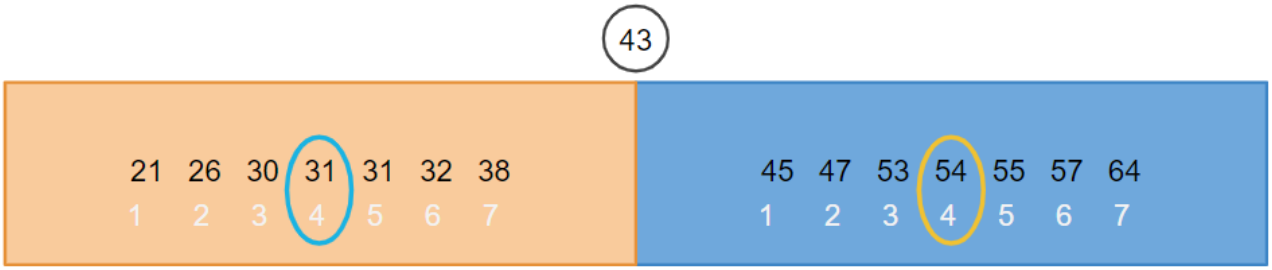
1. Once we've ordered them from smallest to largest, they'll look like this. You can also see their position below each number.



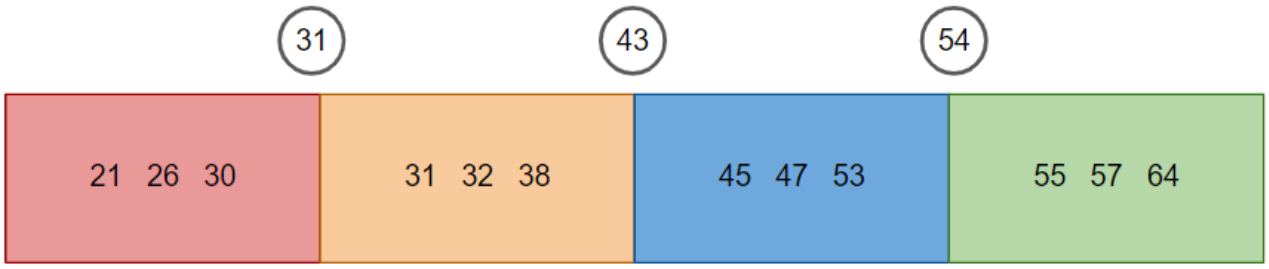
1. If we have 15 ages, the middle age will be at the 8th position. As we can see, the age 43 is in the 8th position.

We can use the median to split our two groups. All ages between 21 and 38 are in the bottom group, and all of our ages between 45 and 64 are in the top group.

1. Each group now has seven ages in it. So the middle value for each group will be in the 4th position.

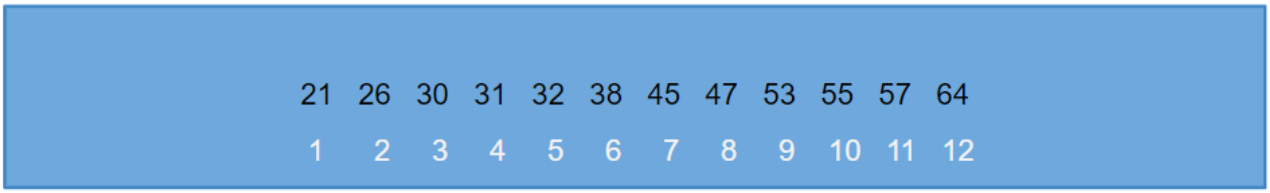


1. We calculate the interquartile range by first finding the value in the middle of the top group, which is 54 in this case. We then find the middle value in the bottom group, which is 31 in our example. The IQR is the difference between these two values. That is, 54 - 31, or, 23.



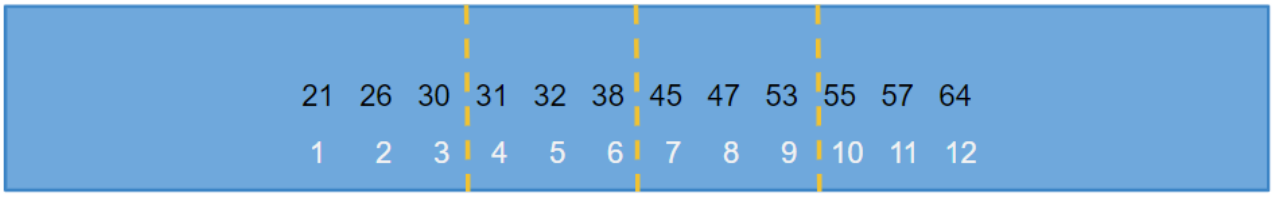
This method of breaking the groups in half, finding the middle number and repeating this for each half works perfectly with a collection of 15 ages. But we can still work out the interquartile range if we had an even number of ages and couldn't find middle values. Let's say we had these 12 ages, instead of our original 15.

With it ordered, it would look like this.



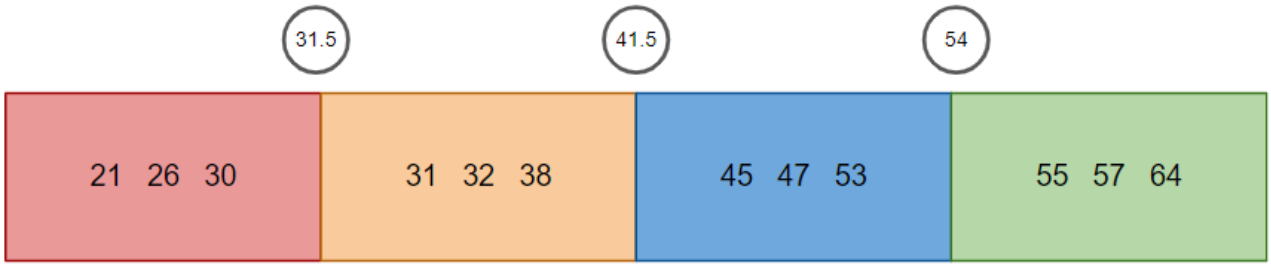
As we can see, because the total number is even, there isn't a number that falls in the middle of the groups. And, if we split the groups in half, there also isn't a number that falls in the middle of either half.

Instead of finding the middle number, we can break the ages in half, and then in half again.



In this case, the "middle" value, between each of the groups, is the average of the values on either side of the line:

(30 + 31)/2 = 31.5 (38 + 45)/2 = 41.5 (53 + 55)/2 = 54



This means that the interquartile range would be 54 - 31.5, or 22.5.

We can also refer to these values in the following way. For the above example:

The Q1 value is 31.5. This is also the 25th percentile.
That is, 25% of values are equal to or lower than 31.5.

The Q2 value is 41.5. This is also the 50th percentile or median.
That is, 50% of values are equal to or lower than 41.5.

The Q3 value is 54. This is also the 75th percentile.
That is, 75% of values are equal to or lower than 54.

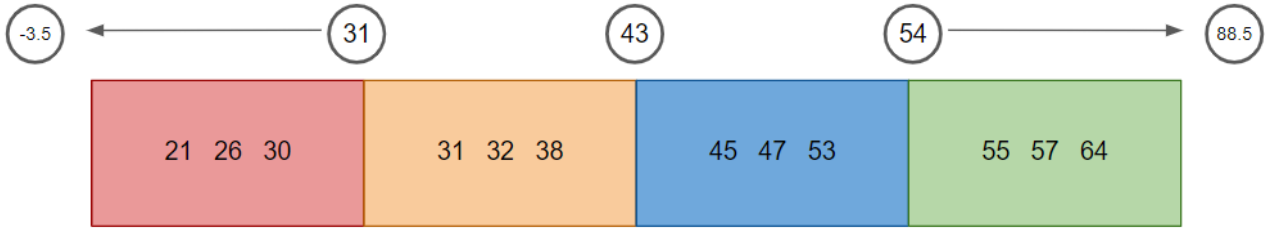
The interquartile range, or IQR, is 22.5

Finding Outliers with the IQR

Minor Outliers (IQR x 1.5)

Now that we know how to find the interquartile range, we can use it to define our outliers. The most common method of finding outliers with the IQR is to define outliers as values that fall outside of 1.5 x IQR below Q1 or 1.5 x IQR above Q3.

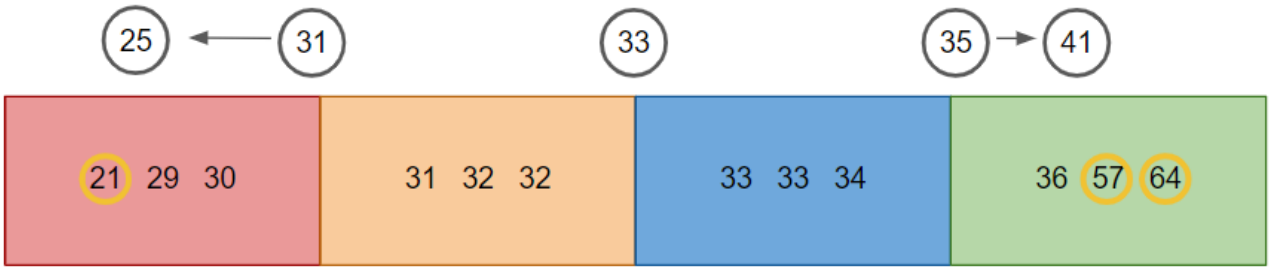
Let's break that down using our original example.



Our IQR was 23. If we multiply this by 1.5, we get 34.5. This means that we would consider any ages that are below -3.5 or above 88.5 to be outliers.

Notice that the thresholds for the outliers are simply defined by the data we use. Even though it's not possible to have a negative age, our outlier calculation only considers the numerical values. In our case, because we are using ages, this means that no matter how young our friend may be, we would not consider them an outlier. We'd also need a friend who was 89 years or older to consider them an outlier.

In the case above, we have a pretty broad range of ages for our friends. What would happen if the range of ages for our friends was much smaller?



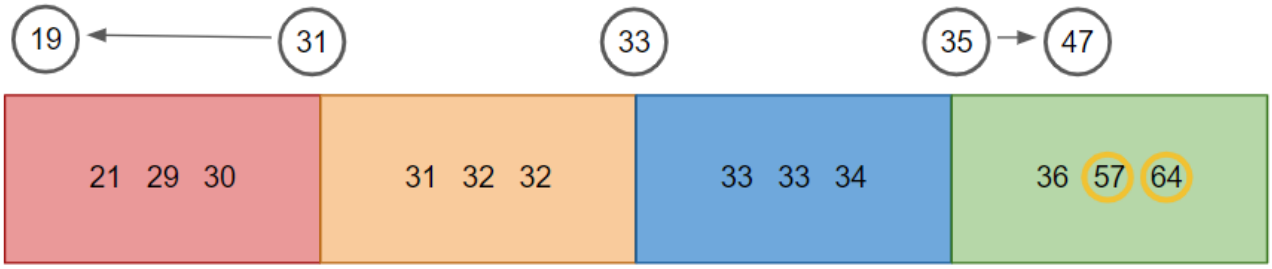
In this case, our Q1 value is 31 and our Q3 value is 35. This means that our IQR is only 4. Now, 1.5 times IQR is 6. Any values below 25, or higher than 41 will be considered outliers.

Now, our friends with the ages 21, 57, and 64 are considered outliers.

Major Outliers (IQR x 3)

This brings us to a second, less common threshold for assessing outliers. If we have a very small IQR, not all outliers are created equal. In the case above, while 21 and 64 are both outliers, 21 is only 10 years lower than our Q1 value of 31. But 64 is 30 years older than our Q3 value. How do we distinguish between "regular" outliers and "extreme" outliers.

A major outlier is defined as values that fall outside of 3 times IQR below Q1 or 3 times IQR above Q3.



If we go back to the previous example, 4 x 3 = 12. Major outliers will be those that are less than 19 and more 47.

This allows us to indicate some difference between 21 and the other two outliers. An age of 21 is not considered a major outlier, but 57 and 64 are major outliers.

Now that we know how to calculate our IQR and identify outliers, let's look at how we can use SQL to find outliers using the IQR.

Finding Outliers With SQL

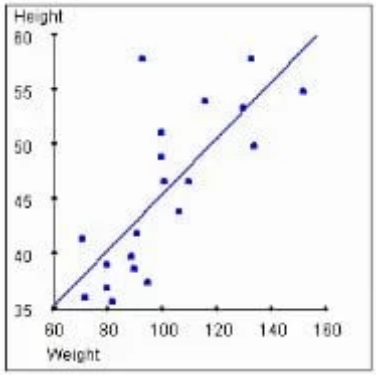
Correlation and P value

The two most commonly used statistical tests for establishing relationship between variables are correlation and p-value. Correlation is a way to test if two variables have any kind of relationship, whereas p-value tells us if the result of an experiment is statistically significant. In this tutorial, we will be taking a look at how they are calculated and how to interpret the numbers obtained.

What is correlation?

Correlation coefficient is used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficients (e.g. Pearson, Kendall, Spearman), but the most commonly used is the Pearson's correlation coefficient. This coefficient is calculated as a number between -1 and 1 with 1 being the strongest possible positive correlation and -1 being the strongest possible negative correlation.

A positive correlation means that as one number increases the second number will also increase. A negative correlation means that as one number increases the second number decreases. However, correlation does not always imply causation — correlation does not tell us whether change in one number is directly caused by the other number, only that they typically move together. Learn more about the Pearson correlation formula, and how to implement it in SQL here. To understand how correlation works, let's look at a chart of height vs weight.



We can observe that with increase in weight, the height also increases – which indicates they are positively correlated. Also, the correlation coefficient in this case is 0.88, which supports our finding. Learn more about correlation and how to implement it in Excel here.

What is a p-value?

P-value evaluates how well your data rejects the null hypothesis, which states that there is no relationship between two compared groups. Successfully rejecting this hypothesis tells you that your results may be statistically significant. In academic research, p-value is defined as **the probability of obtaining results 'as extreme' or 'more extreme', given that the null hypothesis is true** — essentially, how likely it is that you would receive the results (or more dramatic results) you did assuming that there is no correlation or relationship (e.g. the thing that you're testing) among the subjects. To understand what this means, let us look at an example.

We are going to conduct an experiment to check if a coin is biased or not. To do this, let's flip a coin 10 times. Intuitively, we can say that the probability of getting **5 heads and 5 tails** is highest, followed by **6 heads and 4 tails** or **6 tails and 4 heads,** and so on. So first, let's state the null and alternate hypothesis. Since the assumption is that the coin is fair, our null hypothesis is **"The coin is unbiased with equal probability of heads and tails".** We are conducting the experiment to prove/disprove the claim, so our alternative hypothesis is **"The coin is biased with unequal probability of heads and tails"**

Assuming the null hypothesis is true (the coin is fair), let's calculate the probabilities of the various possible outputs i.e 0 heads & 10 tails, 1 head & 9 tails, 2 heads & 8 tails, and so on.

The probabilities are calculated using the probability of a binomial distribution, which gives the probability of r successes in n trials using the formula :

nCr * (p)r * (1-p)n-r

Where,
n = no. of trials = 10
r = no. of successes (heads)
p = probability of a success = 1/2
1-p = probability of a failure = 1/2

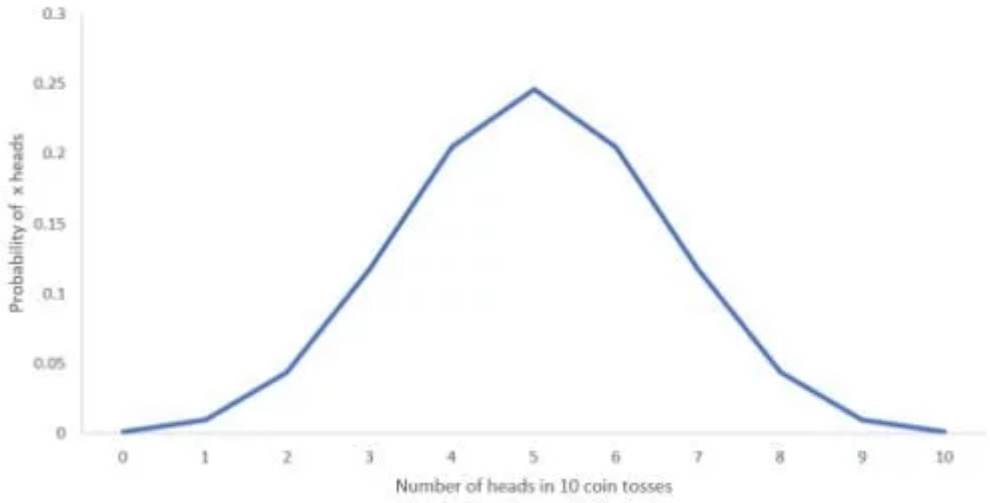
Let's consider a 'success' to be when heads appears in the coin toss. Also, it won't make a difference if 'success' is considered to be heads or tails. Let's first calculate the probability of obtaining 5 heads and 5 tails in 10 coin flips.

P(5 heads and 5 tails) = 10C5 * (½)5 * (½)5 = 0.24609375

Similarly, let's generate the probabilities of all other possible combinations of heads and tails:

No. of Heads	No. of Tails	Probability
0	10	0.000976563
1	9	0.009765625
2	8	0.043945313
3	7	0.1171875
4	6	0.205078125
5	5	0.24609375
6	4	0.205078125
7	3	0.1171875
8	2	0.043945313
9	1	0.009765625
10	0	0.000976563

Let's plot the probabilities to understand the intuition behind the above calculation:



We can observe from the chart that the probability of getting 5 heads is the highest, and the probability of getting 0 heads or 0 tails is the lowest. Now, let's assume we get the output of this experiment as **"9 heads and 1 tail"**.

Let us calculate the p-value of the experiment. To reiterate the definition – **"p value is the probability of obtaining results as extreme or more extreme, given the null hypothesis is true".**

Now, we add the probabilities of all the possible outputs of the experiment which are **as probable** as '9 heads and 1 tail' and **less probable** than '9 heads and 1 tail'.

P-value = P(9 heads and 1 tail) + P(10 heads and 0 tail) + P(9 tails and 1 head) + P(10 tails and 0 heads)

= 0.009765625 + 0.000976563 + 0.009765625 + 0.000976563 = 0.02148437 = 0.02 (approx.)

Now, we need to check whether the p-value is significant or not. This is done by specifying a significance cutoff, known as the alpha value. Alpha is usually set to 0.05, meaning the probability of achieving the same or more extreme results assuming the null hypothesis is 5%. If the p-value is less than the specified alpha value, then we reject the null hypothesis. Hence, we reject the hypothesis that "**"The coin is fair with equal probability of heads and tails"** and conclude that the coin is biased.

Conclusion

Though correlation and p-value provides us with the relationship between variables, care should be taken to interpret them correctly. Correlation tells us whether two variables have any sort of relationship and it does not imply causation. If two variables A and B are highly correlated, there are several possible explanations: (a) A influences B; (b) B influences A; (c) A and B are influenced by one or more additional variables; (d) the relationship observed between A and B was a chance error. Similarly, p-value should not be misused to produce a statistically significant result. If analysis is done by exhaustively searching various combinations of variables for correlation, then it is known as p-hacking.