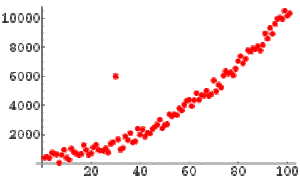
How should outliers in the data be handled from a visualization perspective? Should they just be removed?

**What are outliers?**

Outliers are data observations that significantly differ from others. These are beyond the range of what is anticipated. Researchers hope that their data should not have any outliers because it can challenge the validity of the data if not treated correctly. A simple visualization of an outlier is shown below. The below graph shows that there is strong co-relation between the attributes in x and y axis there is one data point which lies away from the other points, which is an outlier.

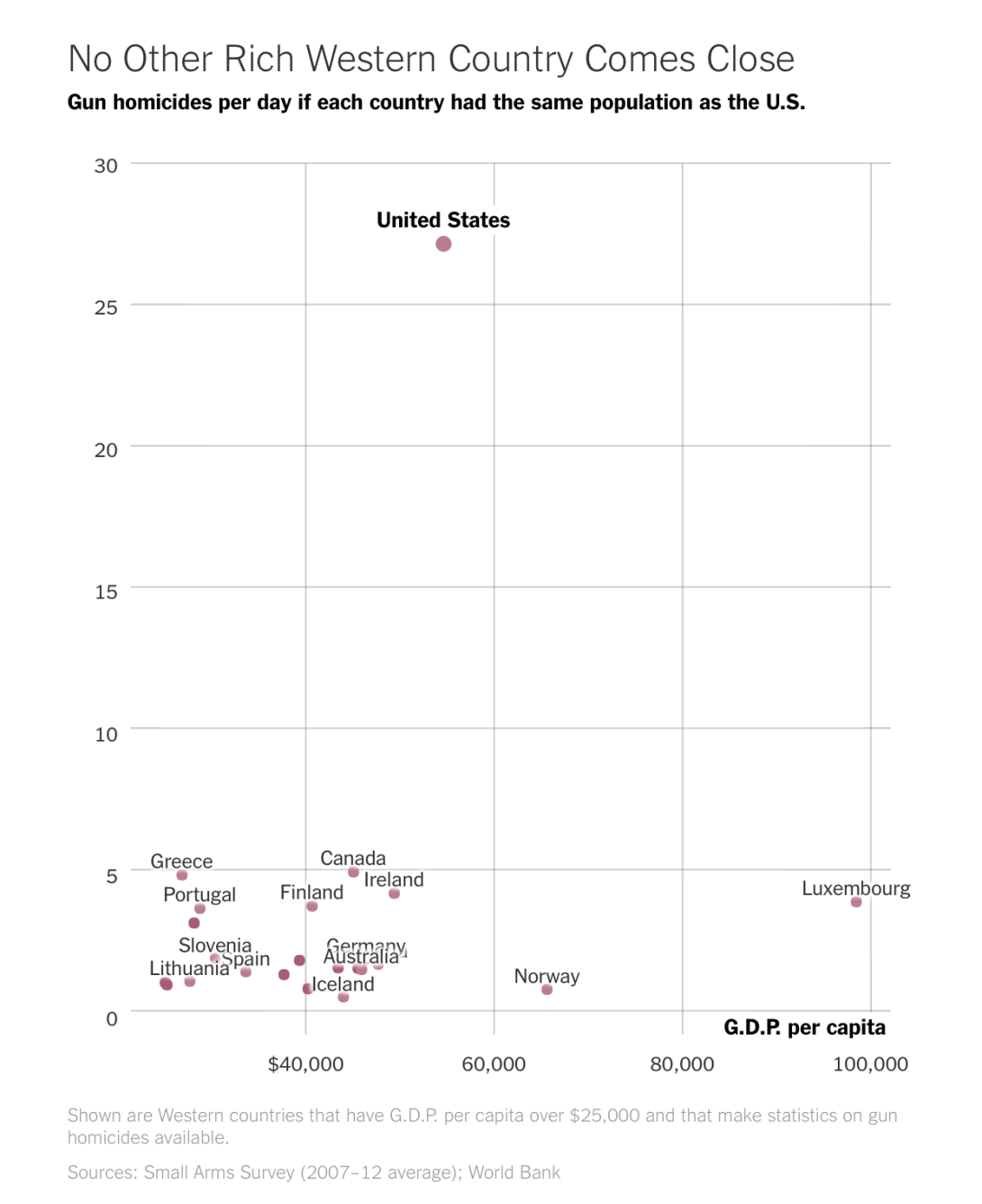


**What causes outliers?**

Outliers can be caused by manual errors while recording the data. For eg: when recording the heights of individuals in feet someone can enter data in centimeters and when plotted against the regular data set, this will be an outlier. Sometimes these outliers can be because of an exception or something extraordinary. We generally see this in sports. for eg: In test cricket, Sir Donald Bradman’s average is nearly 100(99.94 to be precise) whereas some of the other extraordinary players have a career average of 50. In the examples mentioned, one is just noise compared to other data whereas the other outlier deserves an applaud and a closer observation. When we are planning to plot, we need to take a closer look at our data to determine what is what first. This will help us in deciding whether to eliminate it or highlight it. There are few options that are mentioned in this week’s article. I have listed few of them below.

**Point of Focus:**

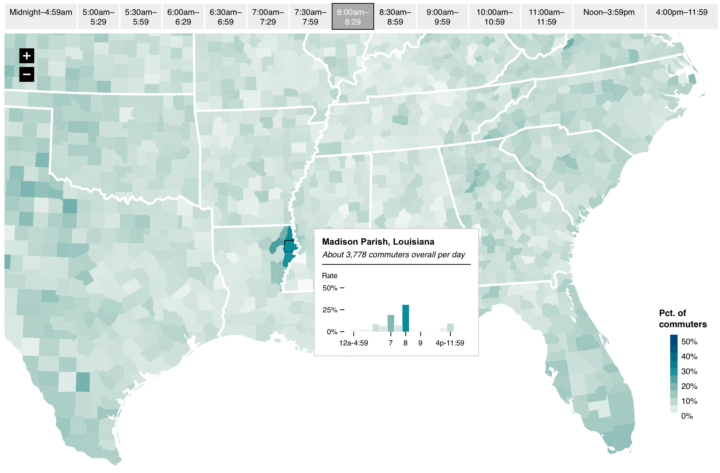
This focuses directly on the outlier and helps us in understanding how it varies from the other data points. For eg: Below is the plot highlighting the no. of gun homicides per day in the United States when compared with other Western Nations. These kinds of outliers tend to come up in the news a lot as they significantly differ from the average.



This kind of plots grabs the attention of the viewer’s easily from the rest as it highlights the single point. But highlighting these outliers can make the other data points look unimportant or that they are squeezed into one side. If we do not wish to highlight an outlier, use an alternative way of visualizing the data.

**Breakout:**

In this method, we’ll summarize the data normally and then explain the outliers by highlighting them. By this way, we can understand the total distribution of the data without focusing on the outliers.



In the example mentioned in the article[3], the author mapped the whole country first and summarized that most of the commuters leave home for work between 7:00 AM and 8:00 AM and then highlighted and explained the anomalies that stood from the rest.

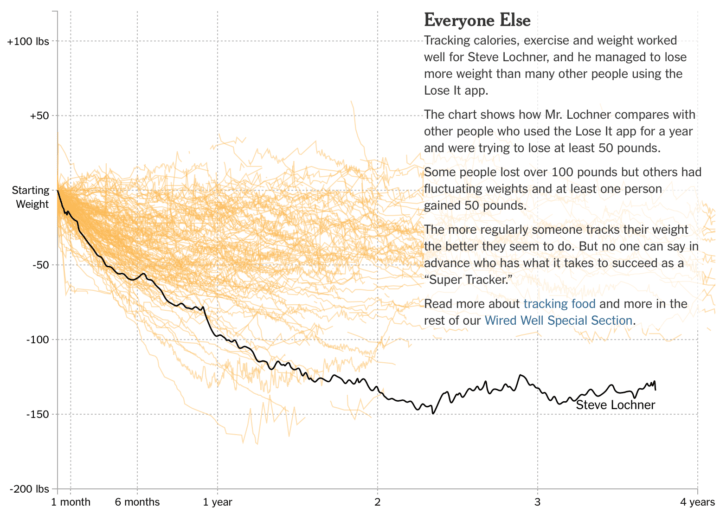
**Scale Adjustment:**

In some cases, these outliers can be best plotted on a different scale so that the averages and the extremes are displayed at the same time. For eg: when plotting data with large counts, logarithmic scale is often helpful. In the example mentioned by the author to show the emergency room visits [4], logarithmic scale was used for the vertical axis to cover the wide range of visits.

With this method, we can show the complete data set, without obscuring too much of data but there is a chance that the outlier may end up as an average data point.

**Reference point:**

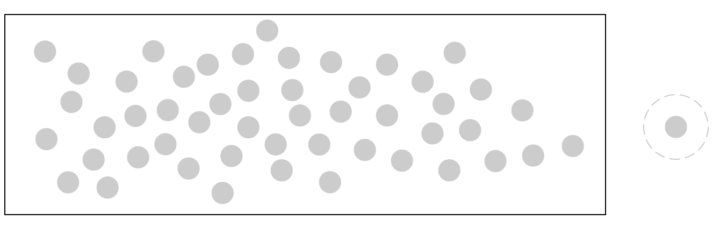
In this, we use the outliers to compare with the other data to make the data more relatable to the viewers.



In the above example, the story was told by the plot by focusing more on the weight loss of the individual compared to the others. This makes the data more familiar to the reader as we use this outlier as a reference point. But this will highlight the differences between the outlier and rest, and we should be cautious to not lose the general distribution.

**Providing Context:**

If we really do not want to highlight the outlier as it many not be important as the rest, just use it in the background. But this can end up as a side thought and can even be ignored later.



To conclude, we should primarily understand our data and outliers to see if it’s just a noise in the data or if it needs our attention. and decide what characteristics of the outliers we want to highlight and then proceed with the type of visualization.