**Outlier Detection**

Walkthrough Video:

<https://berkeley.zoom.us/rec/share/p4QARbqkhU-emsth3QaE817_w6ujGmUfnqL-P325hZc64bfhlp1MAktLNLKfiHw.9M430qmnpMZoh0Pr>

This week’s main focus is on Outlier Detection. We went through examples that showcased some of the ways in determining and highlighting outliers in our dataset. Some of the examples I used included BMI, although many parts of the data are relatively distributed smoothly based on how the sample was collected. However, there are still things that we can get some more insight from.

Here are some potential topics for you to explore.

* One of the parts of our data that is heavily skewed is in charges. Can you figure out how to use the examples shown to approach this, and draw more conclusions from our data on outliers with charges?
* You can also further explore BMI on your own if there may be other parts of the data not covered in the general walkthrough guide. Feel free to google more information on techniques to do this, since the example is not an exhaustive list of techniques for analyzing this!
* Anything else you can come up with on your own!

We gave an example of defining a function that helped us find outliers using the premise of boxplots, and getting data outside our IQR. Here are some potential methods for you to actually find these outliers.

* Standard deviation is a common measurement of how different our outliers might be. When our data is similar to a gaussian distribution (that is, a normal distribution), using standard deviation is a great way to measure outliers. What are common thresholds for doing this? Can you figure out how well this works for the data you choose to explore?
* To build off the above, Z-Scores are a great way to represent how far away our value might be from the mean with a “normalized” perspective. You can feel free to learn a bit more, and maybe try different thresholds of cutoffs for outliers to see what you can learn from the data.
* Again, if you are feeling adventurous, feel free to research more complex ways to determine unique outliers from our data.

Although some parts of the data might not be ideal for identifying outliers, in many real world datasets we will not have samples that are so evenly distributed and conform nicely to the situations we have set up. Identifying unique data points is crucial to understanding key characteristics from this, and this week’s topic will overall cover a small portion of statistics as well for you to familiarize yourself with.