**Week 4 - Disabilities Discrimination?**

**Process log**

Suhail Bhat

It is imperative to understand our dataset before exploiting it. Starting with the central measures, I gathered that the data is skewed, but not terribly so.

To get further clarity on that, I drew a normal distribution curve to gauge the dispersion of the data. To draw a normal distribution curve, we need mean and standard deviation calculated using built-in excel function. Run that NORM.DIST function for one cell and then drag it down to calculate for every cell. Select the distribution column and the ‘expenditure’ column, then insert a smooth line graph. This gives a curve that shows that we don’t have an exact normal distribution, but it is not that off from a normal distribution.

Now that we understand our data, we need to decide how we are going to target it to spot outliers and trends.

Find the first quartile(Q1) and the third quartile (Q3) using QUARTILE function and then calculate interquartile range(iqr) Q3-Q1.

Quartile measures are important for this dataset because it is not exactly normal distribution. Using iqr, calculate upper bound (=Q3+iqr \*1.5) and lower bound (=Q1-iqr\*1.5) values to spot outliers.

Now that we are done with the overall analysis, it is time to narrow it.

Let us check out if there is a gender bias in this dataset.

Count the number of females and males who received this funding by filtering the datasets. Also, find out the average expenditure (sum of expenditure divided by 1000) for each of these genders.

Plot this data to see if a trend pops out.

Let us check the dataset for racial bias.

There are eight categories in the ethnicity section, four of which are insignificant for analysis because of their small values.

The top four races in this dataset are Black, White not Hispanic, Asian, Hispanic.

Using filters, calculate the count and average expenditure for each of these races. Plot a graph.

Similarly, we can run analyses on age groups. Using filters, calculate the count and average expenditure for each of the age groups. I made five age groups based on my understanding of developmental disabilities. From the CDC’s website, I gather that development disabilities affect one in six children in the United States. So, I wanted my analysis to go in that direction. I divided it into 0-8 years, 9-17, 17-21, 22-50 and 51+. I made a bigger age group (22-50) because the developmental disability isn’t as big question for adults as it for children with unassailable innocence.

Now that we are done with analyzing bias using one variable, such as ethnicity, age and gender, at a time, we need to do run an analysis using two independent variables while keeping the third variable constant. I am taking a call at this point. My experience with this data tells me that the age group 0-17 and 55+ are interesting ones because together those two groups account for 60 percent of the data. Also, development disability is such an important issue for children- that is where my focus is going to be for the rest of the analysis.

Filter the data for the 55+ age group, now filter it for race and gender. Do a count of data points and average expenditure for each of these functions. Repeat this process for the four major races and plot the findings.

Repeat the process for the 0-17 age group.

Check your method again.