Step-by-Step Excel Data Manipulation Clustering Steps

Data Manipulation

1. Downloaded Baltimore City Employees' Salaries from the Open Baltimore publicly available data online
2. Filtered for FY2019 only to focus on employees that were employed and paid for the FY2019
3. Used the Text to Columns tool to separate the date and time for Hire Date, before changing the format to short date
4. Created a column, End_FY2019, to account for the end of FY2019 which is 06/30/19
5. Subtracted the difference between Hire Date and End_FY2019 and formatted the cells to Numbers to get the number of days, Days_Worked
6. Divided Days_Worked by 365 to get Years_Worked
7. There were a few negative outliers which were filtered out since these were people recorded as FY2019 but were hired during FY2020
8. Sorted Agency Name in Ascending Order to standardize jobs by the agency it fell under since there are multiple divisions within each agency
9. Each agency had several numerical denotations, i.e. City Council (001), City Council (002) etc. Filtered out using Text to Columns and parentheses as the delimiter to only include the main Agency Name

Creating Visuals

1. Calculated the Average Number of Years worked, Average Annual Salary, and Average Gross Pay for each department by creating a Pivot Table
    a. Decided to only include HLTH - Health Department as there were several locations with drastically varying years worked and average pay to give us a total of 56 agencies
2. Created the data visualizations by copying the Pivot Table average data into new tables on separate sheets
3. Sorted the data by Ascending for each of the three variables, Average Years Worked, Average Annual Salary, and Average Gross Pay
4. Created bar graphs for each of the three variables in ascending order

Clustering

1. Numbered the Agencies from 1 to 56
2. Calculated the Mean and Standard Deviation using AVERAGE and STDEV functions for the three variables
3. Calculated the z-eligible columns using the STANDARDIZE function for each of the three variables
4. Highlighted the whole data array to name it "Cluster" and numbered the z-scores columns in the table above
5. Used the VLOOKUP function to set up our anchors in finding the names for our random cluster numbers in the table above as well as respective z values for each of the variables
6. Used the SUMXMY2 function to calculate the distance between the variable z-scores for each agency and the z values of the anchors
7. Used the MIN function to find the smallest distance of the four distances in each agency row
8. Used the MATCH function to assign the smallest distance to an anchor
9. Calculated the sum of all the minimum distances by using the SUM function
10. Applied the Excel Solver operations for a cluster analysis on the dataset
    a. Set objective to our sum of the minimum distances cell
    b. Selected min
    c. By changing variable cells to the cluster numbers in the table
    d. Subject to the constraints with the cluster cells > = 1, < = 56, and = integer
    e. Selected Evolutionary as the solving method