Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Our main objective is to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?
   To make a decision as where to open the newest store, we need historical information of how the company's current stores are doing financially. We will be using a yearly model so all data needs to be in a yearly format and since we are analyzing at city level, if there are multiple stores in one city, the data needs to be appropriately aggregated to get a city level.

   To open a new store, we not only need to know financial health but also need to understand population demographics of the city like whether the population has decreased over the years or grown, age distribution, how many families are there in the city, how large and how dense is city.

Step 2: Building the Training Set

Round up to two decimal places, ex: 1.24

Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute?

Gillette, Cheyenne, RockSprings – 3 cities appeared as outliers using the IQR method for the six numerical variables we have used. I have decided to remove Cheyenne as it appears as an outlier in 3 of the 6 of the numerical variables we are using to perform our analysis. If we look at the Total sales, Population Density and Census for Cheyenne we see that those values are way above the Higher Fence in all three cases.

I got the idea of finding percentiles in order to filter outliers from the Alteryx community. Below is the link to the discussion –
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Calculating-Quartiles/m-p/28479