Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

My responsibilities are cleaning the data provided to perform the regression analysis afterward.
For now,
I need to clean, filter, join, and create an analytical dataset..

Key Decisions:

Answer these questions

Many decisions need to be made in order to set up the data properly.
1- clean data from unnecessary characters and create calculated columns for some data all csv.
2- prepare a training dataset to calculate the mean, sum, interquartile, IQR, and median.
3- find outliers to avoid misinterpreting the data.

Step 2: Building the Training Set

Review (cleaned data with IQR.xlsx)

Column	Sum	Average
Census Population	*213,862*	19,442
Total Pawdacity Sales	*3,773,304*	343,027.64
Households with Under 18	*34,064*	3,006.49
Land Area	*33,071*	3,096.73
Population Density	*63*	5.71
Total Families	*62,653*	5,695.71

Step 3: Dealing with Outliers

Answer these questions

Yes there are outliers. However, Cheyenne has 4 outliers and this actually does not consider outliers as the relationship with Cheyenne' 4 outliers are making sense because the population are large. Thus other 3 outliers in Cheyenne are also large which is not a mistake nor a typo. If

we delete Cheyenne, we will lose the 4 out of 6 data of Cheyenne. I concluded to keep outliers for Cheyenne. . I am considering imputing the Population Density of Cheyenne 20.34 and I will substitute it with the overall average 5.71. For Rock Springs, it has one outliers and I decided to keep too