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?
   The main decision is to find out the best city to establish the new Pawdacity Store (number 14), considering the sales forecast and business opportunity correlated with: population, the density of population, Households with Under18 years old members, a total of families, and other competitor's stores participation and their location and market share.

2. What data is needed to inform those decisions?
   The data need to report are:
   - The best city to establish the new Pawdacity store
   - The sales forecast calculated by month
   - Map ubication showing zone of influence and other comparison's stores
   - Potential amount of customers, based on population amount, the density of population, Households with Under 18 years old members, a total of families and competitor's participation and their market share.

Step 2: Building the Training Set

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

In addition, provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

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

Step 3: Dealing with Outliers

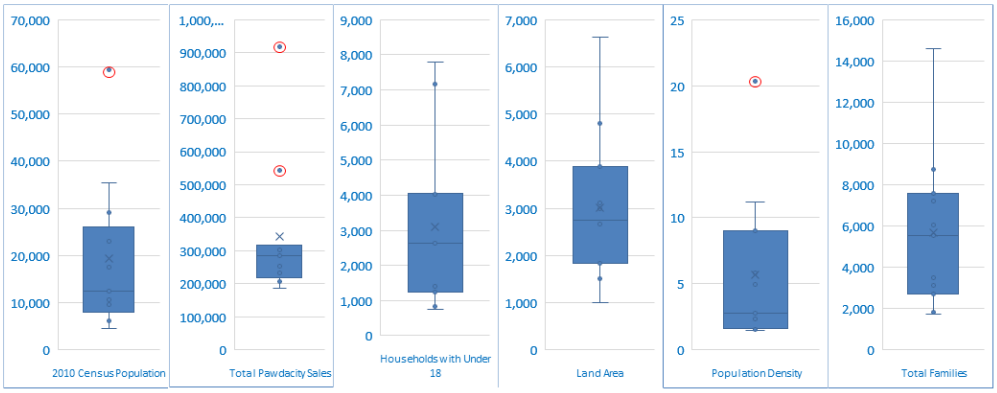
Answer these questions

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

According to the IQR method, there are two cities with outliers:

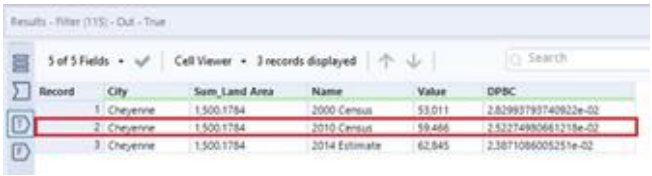
- Cheyenne city has 3 outliers: 2010 Census Population, Total Pawdacity Sales and Population Density
- Gillette city has one outlier: Total Pawdacity Sales

City	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4,585	185,328	746	3,116	2	1,820
Casper	35,316	317,736	7,788	3,894	11	8,756
Cheyenne	59,466	917,892	7,158	1,500	20	14,613
Cody	9,520	218,376	1,403	2,999	2	3,516
Douglas	6,120	208,008	832	1,829	1	1,744
Evanston	12,359	283,824	1,486	999	5	2,713
Gillette	29,087	543,132	4,052	2,749	6	7,189
Powell	6,314	233,928	1,251	2,674	2	3,134
Riverton	10,615	303,264	2,680	4,797	2	5,556
Rock Springs	23,036	253,584	4,022	6,620	3	7,572
Sheridan	17,444	308,232	2,646	1,894	9	6,040
Sum	213,862	3,773,304	34,064	33,071	63	62,653
Avg	19,442.00	343,027.64	3,096.73	3,006.49	5.71	5,695.71
QR1	6,314	218,376	1,251	1,829	2	2,713
QR2	12,359	283,824	2,646	2,749	3	5,556
QR3	29,087	317,736	4,052	3,894	9	7,572
IQR	22,773	99,360	2,801	2,065	7	4,860
Lower Fence	-27,846	69,336	-2,951	-1,268	-9	-4,577
Upper Fence	63,247	466,776	8,254	6,992	20	14,861
Outlieers	59,466	917,892|543,132			20	



The data to impute is Population Density from Cheyenne City

The population density calculus with the data from p2-partially-parsed-wy-web-scrape.csv (census) and p2-wy-demographic-data.csv (Land Area) shows a value of 2.25 for Cheyenne city.



Alteryx Workflow