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?

Pawdacity currently has 13 stores in Wyoming. The decision that needs to be made is; which city should the new store be located in.

2. What data is needed to inform those decisions?

The data needed to inform this decision is as followed at a city level:

- Total sales of current Pawdacity stores.
- Current population estimations for cities in Wyoming.
- Current demographic data for populations in Wyoming at a city level including; Total Families, Households with under 18, Population density.
- Land area of cities in Wyoming.
- Total sales of competitors.

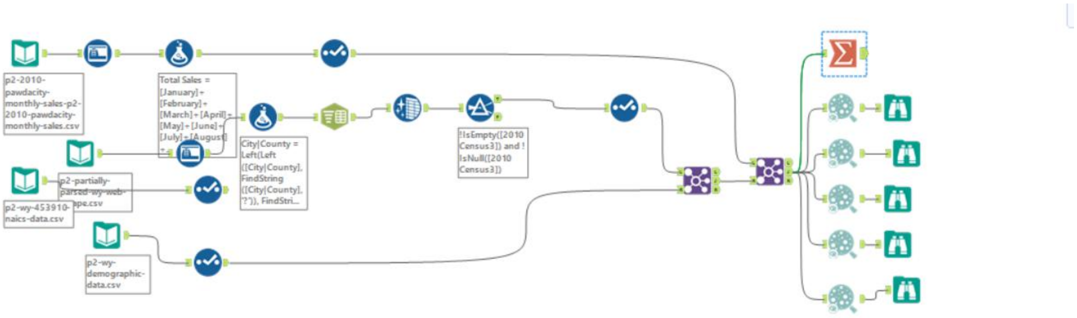
Step 2: Building the Training Set

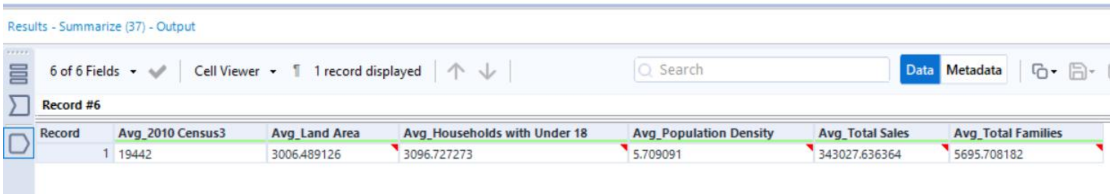
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

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*	*19442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
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*

Alteryx Workflow for cleaning the data below:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]+[August]+...
City|County = Left(Left([City|County], FindString([City|County], '?')), FindStri...
!IsEmpty([2010 Census3]) and !IsNull([2010 Census3])
p2-partially-parsed-wy-web-scrape.csv
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv



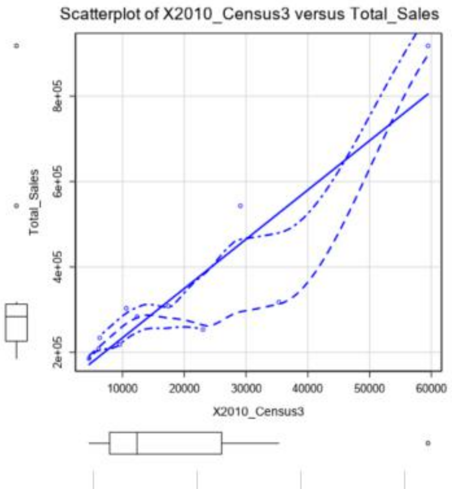
Record	Avg_2010 Census3	Avg_Land Area	Avg_Households with Under 18	Avg_Population Density	Avg_Total Sales	Avg_Total Families
1	19442	3006.489126	3096.727273	5.709091	343027.636364	5695.708182

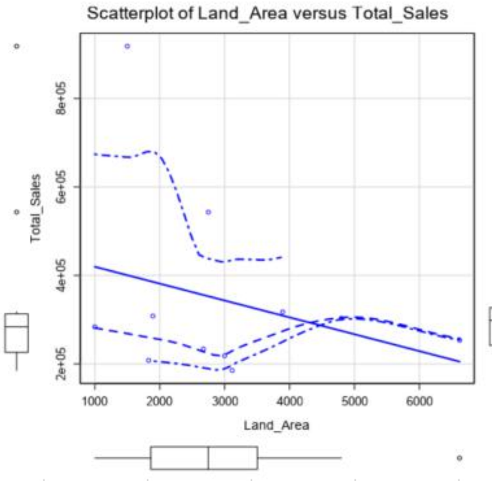
Step 3: Dealing with Outliers

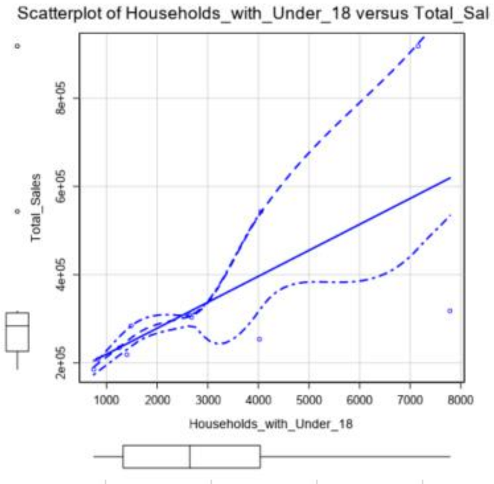
Answer these questions

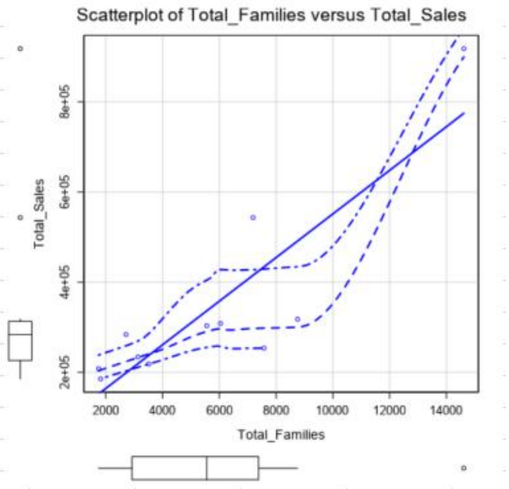
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

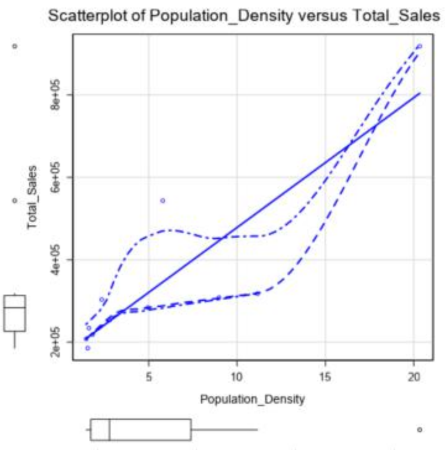
I will remove the city of **Cheyenne** as an outlier based on the following analysis.



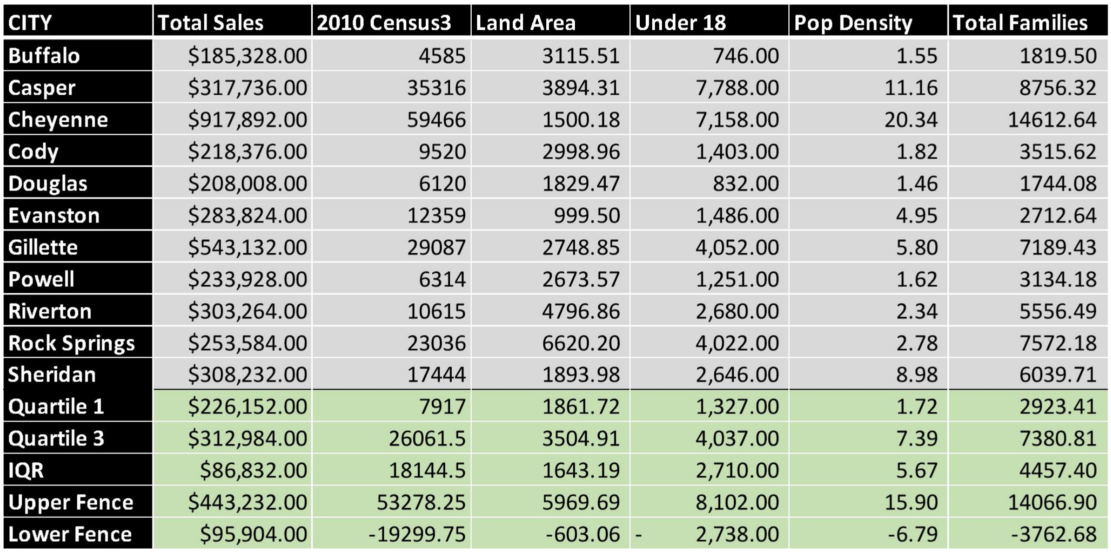






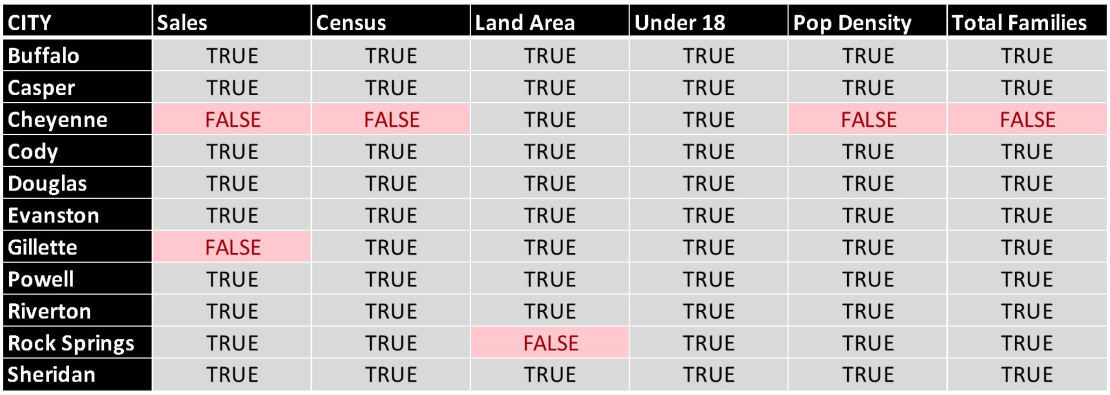


The following is an analysis of the outliers in each of the numeric variables:



CITY	Total Sales	2010 Census3	Land Area	Under 18	Pop Density	Total Families
Buffalo	$185,328.00	4585	3115.51	746.00	1.55	1819.50
Casper	$317,736.00	35316	3894.31	7,788.00	11.16	8756.32
Cheyenne	$917,892.00	59466	1500.18	7,158.00	20.34	14612.64
Cody	$218,376.00	9520	2998.96	1,403.00	1.82	3515.62
Douglas	$208,008.00	6120	1829.47	832.00	1.46	1744.08
Evanston	$283,824.00	12359	999.50	1,486.00	4.95	2712.64
Gillette	$543,132.00	29087	2748.85	4,052.00	5.80	7189.43
Powell	$233,928.00	6314	2673.57	1,251.00	1.62	3134.18
Riverton	$303,264.00	10615	4796.86	2,680.00	2.34	5556.49
Rock Springs	$253,584.00	23036	6620.20	4,022.00	2.78	7572.18
Sheridan	$308,232.00	17444	1893.98	2,646.00	8.98	6039.71
Quartile 1	$226,152.00	7917	1861.72	1,327.00	1.72	2923.41
Quartile 3	$312,984.00	26061.5	3504.91	4,037.00	7.39	7380.81
IQR	$86,832.00	18144.5	1643.19	2,710.00	5.67	4457.40
Upper Fence	$443,232.00	53278.25	5969.69	8,102.00	15.90	14066.90
Lower Fence	$95,904.00	-19299.75	-603.06	- 2,738.00	-6.79	-3762.68

IQR Outlier Table
(False means outside upper and lower fence base on IQR Method)



CITY	Sales	Census	Land Area	Under 18	Pop Density	Total Families
Buffalo	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Casper	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Cheyenne	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
Cody	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Douglas	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Evanston	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Gillette	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
Powell	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Riverton	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Rock Springs	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
Sheridan	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

As seen above, Cheyenne is an outlier in four out of the six numeric variables, where as only two other cities are an outlier in one numeric variable each. Although Cheyenne does seem to follow the trend if we extrapolate the linear trend-line in a couple of them, it is still the most appropriate to remove.