Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions need to be made?
- Choose a city location for Pawdacity's newest store based on predicted yearly sales.
2. What data is needed to inform those decisions?
- The monthly sales data for all of the Pawdacity stores for the year 2010;
- The WA's demographical data of each city (Households with individuals under 18, Land Area, Population Density, and Total Families).
- 2010 Census data of population for each city in WA.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales;

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*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

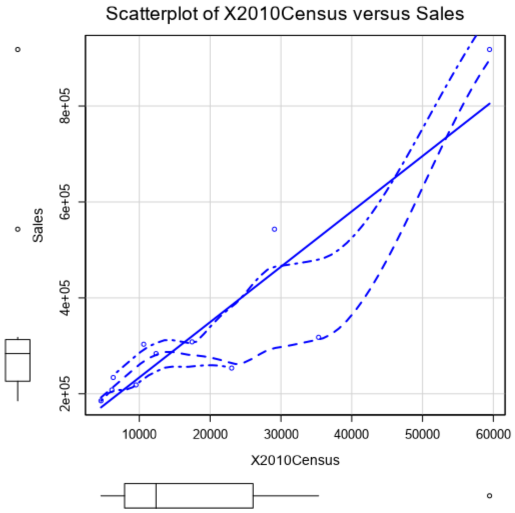
Step 3: Dealing with Outliers

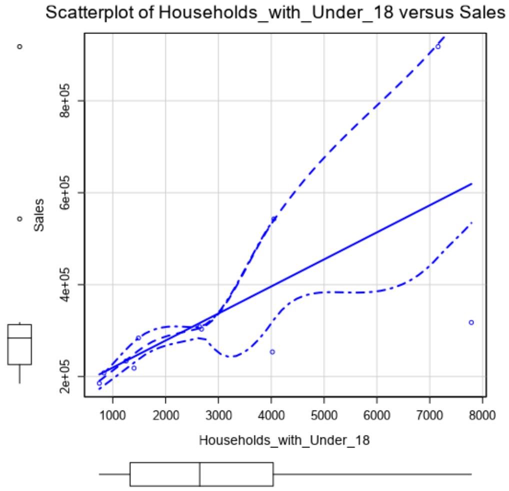
Answer these questions

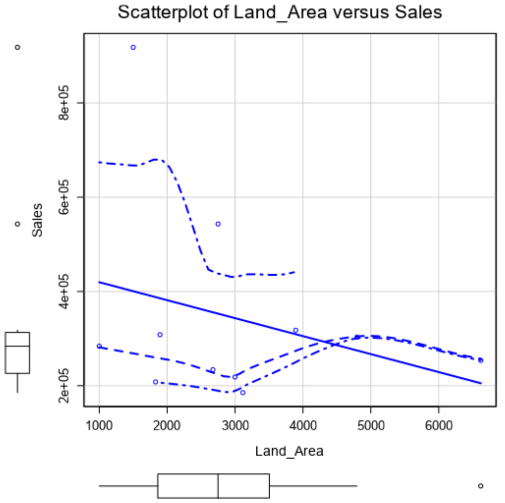
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

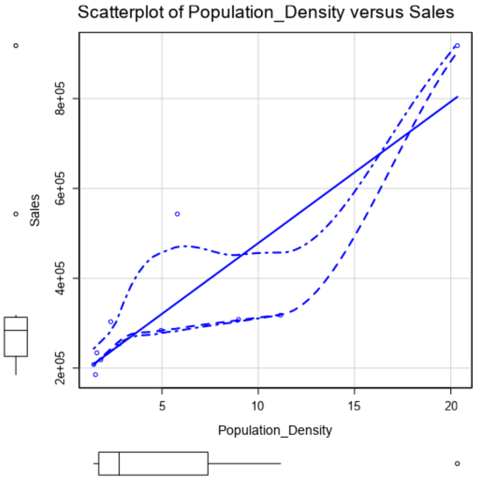
- Yes

- I started with the scatter plot to get an intuition of the relationship between each independent variable with the target variable. Here are the plots

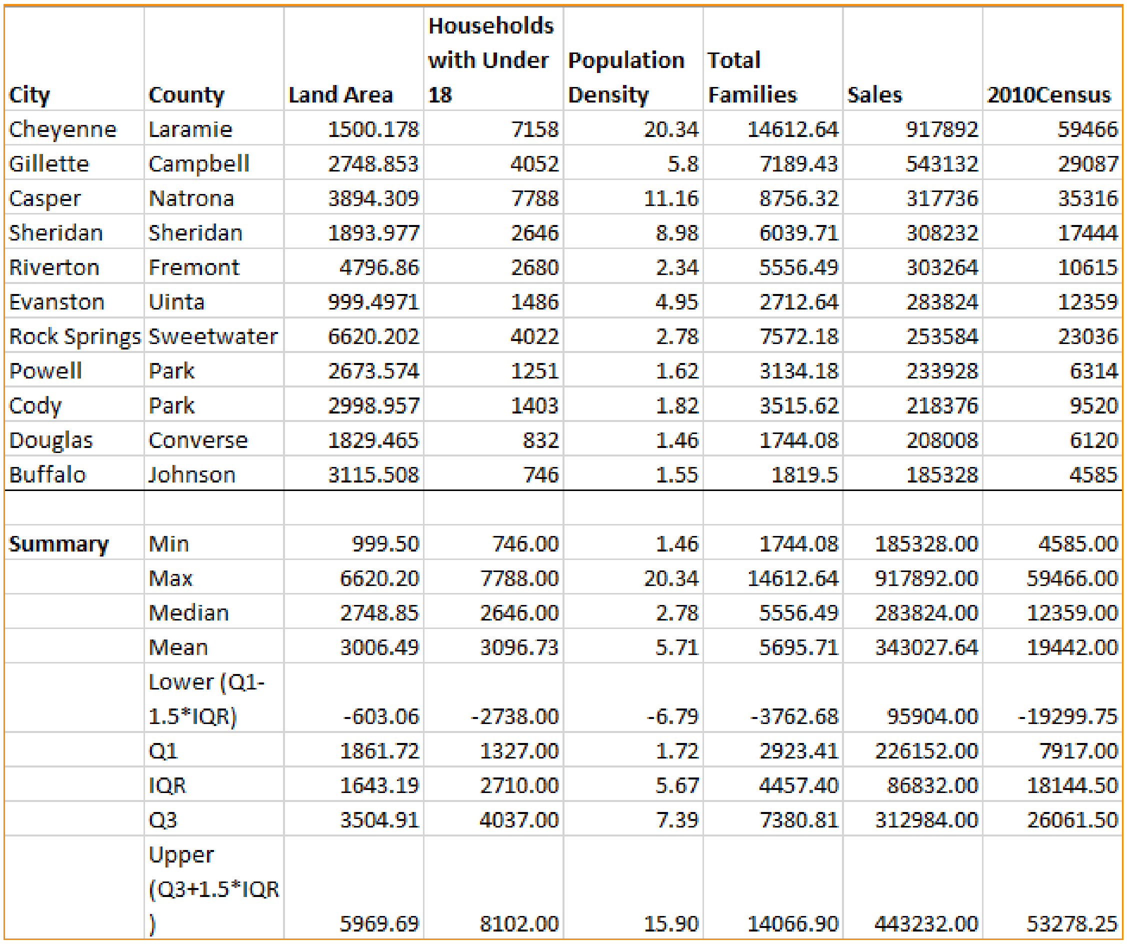








- Then i found there are outliers in plot 2 and 3 (the land_area and households_with_18). To better check the result, I used another technique which is the interquartile range summary to give the statistical overview to the outliers.
  Here is the summary statistics with min, max, median, mean, Lower Fence (Q1-1.5*IQR), Q1, IQR, Q3, and Upper Fence (Q3 + 1.5 *IQR).
- The condition of selecting outliers here is the value of the variable is greater than the upper fence or is less than the lower fence.

- 

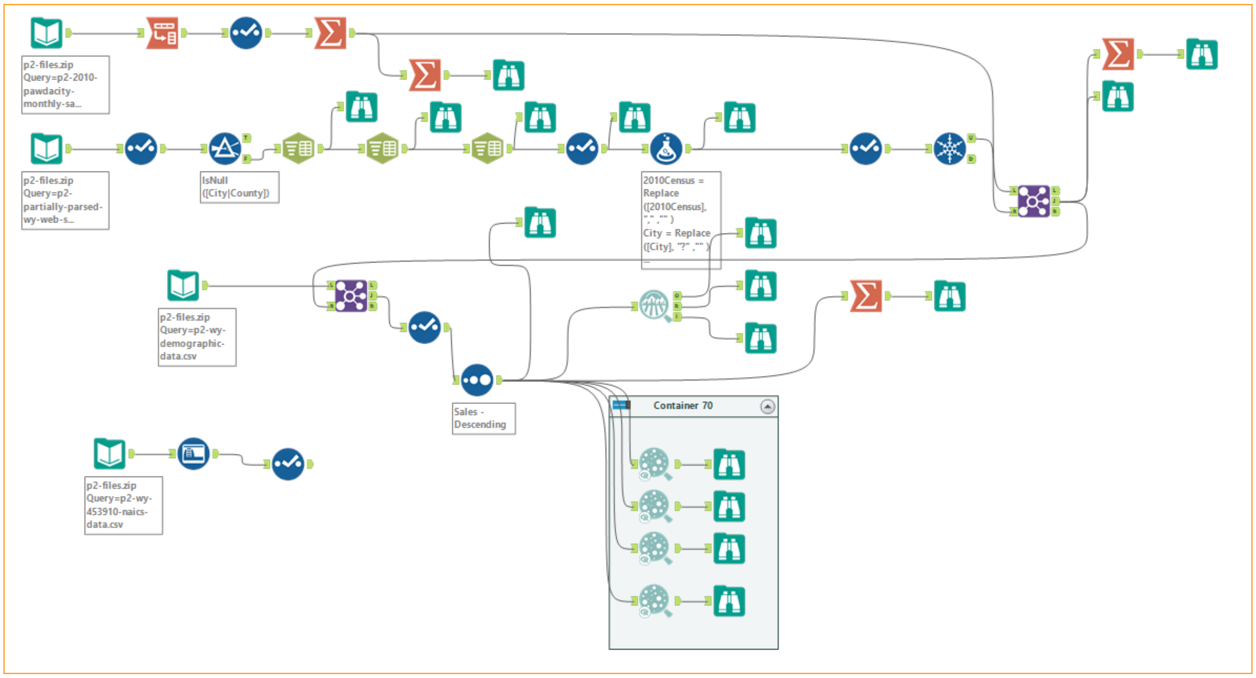
City	County	Land Area	Households with Under 18	Population Density	Total Families	Sales	2010Census
Cheyenne	Laramie	1500.178	7158	20.34	14612.64	917892	59466
Gillette	Campbell	2748.853	4052	5.8	7189.43	543132	29087
Casper	Natrona	3894.309	7788	11.16	8756.32	317736	35316
Sheridan	Sheridan	1893.977	2646	8.98	6039.71	308232	17444
Riverton	Fremont	4796.86	2680	2.34	5556.49	303264	10615
Evanston	Uinta	999.4971	1486	4.95	2712.64	283824	12359
Rock Springs	Sweetwater	6620.202	4022	2.78	7572.18	253584	23036
Powell	Park	2673.574	1251	1.62	3134.18	233928	6314
Cody	Park	2998.957	1403	1.82	3515.62	218376	9520
Douglas	Converse	1829.465	832	1.46	1744.08	208008	6120
Buffalo	Johnson	3115.508	746	1.55	1819.5	185328	4585
Summary	Min	999.50	746.00	1.46	1744.08	185328.00	4585.00
	Max	6620.20	7788.00	20.34	14612.64	917892.00	59466.00
	Median	2748.85	2646.00	2.78	5556.49	283824.00	12359.00
	Mean	3006.49	3096.73	5.71	5695.71	343027.64	19442.00
	Lower (Q1-1.5*IQR)	-603.06	-2738.00	-6.79	-3762.68	95904.00	-19299.75
	Q1	1861.72	1327.00	1.72	2923.41	226152.00	7917.00
	IQR	1643.19	2710.00	5.67	4457.40	86832.00	18144.50
	Q3	3504.91	4037.00	7.39	7380.81	312984.00	26061.50
	Upper (Q3+1.5*IQR)	5969.69	8102.00	15.90	14066.90	443232.00	53278.25

- If we examine each predictor variable, we can easily find that:
    - **Land Area:** Rock springs shows an outlier.
    - **Households with under 18:** no outlier was identified.
    - **Population Density:** Cheyenne stands out as an outlier.
    - **Total Families:** Cheyenne stands out as an outlier.
    - **Sales:** Cheyenne and Gillette stand out as outliers. And the sales of Cheyenne is way more beyond the upper fence level.
    - **2010 Census:** Cheyenne is the outlier.

Given this is a small aggregated dataset and the statistics of Cheyenne city stand out in several dimensions. It is reasonable to remove Cheyenne from the dataset.

Appendix:

Alteryx Workflow


p2-files.zip
Query=p2-2010-
pawdacity-
monthly-sa...
p2-files.zip
Query=p2-
partially-parsed-
wy-web-s...
IsNull
([City|County])
2010Census =
Replace
([2010Census],
",", "")
City = Replace
([City], "?" ,"")
...
p2-files.zip
Query=p2-wy-
demographic-
data.csv
Sales -
Descending
Container 70
p2-files.zip
Query=p2-wy-
453910-naics-
data.csv