Project 2.1: Data Cleanup

Business and Data Understanding

The task for this project is to determine the location of a new Pawdacity store. There are presently 13 stores in 13 cities, and we have been tasked with determining the best location for the new store in the state of Wyoming. This document will explore the steps to data preparation and outlier analysis in the effort to effectively predict potential factors within historic and demographic data that indicate a correlation to sales.

To build the predictive model using the training dataset, we need historic sales data to represent the target variable, as well as demographic and city level data to determine potential predictor variables. The predictor variables have been sourced from external sources requiring data wrangling techniques to clean and tidy up the data.

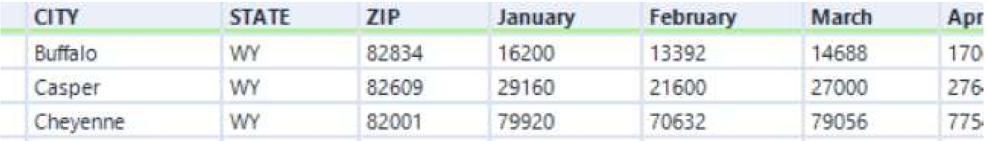
Building the Training Set

STEP 1: Gather and Clean

Gather the data from three different data sources using the Alteryx input tool. All three datasets require some extent of cleaning or formatting before further analysis. The three datasets can be categorized as follows:

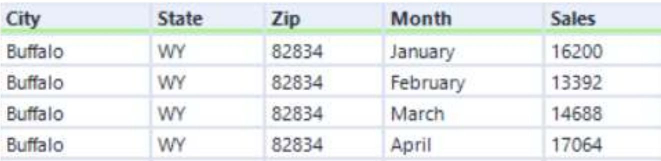
- Sales dataset
- Census dataset
- Demographic dataset

DATASET 1 – The original Pawdacity Sales dataset split sales by month, identified in a 'long' format with months as columns (as seen below).



CITY	STATE	ZIP	January	February	March	Apr
Buffalo	WY	82834	16200	13392	14688	170
Casper	WY	82609	29160	21600	27000	276
Cheyenne	WY	82001	79920	70632	79056	775

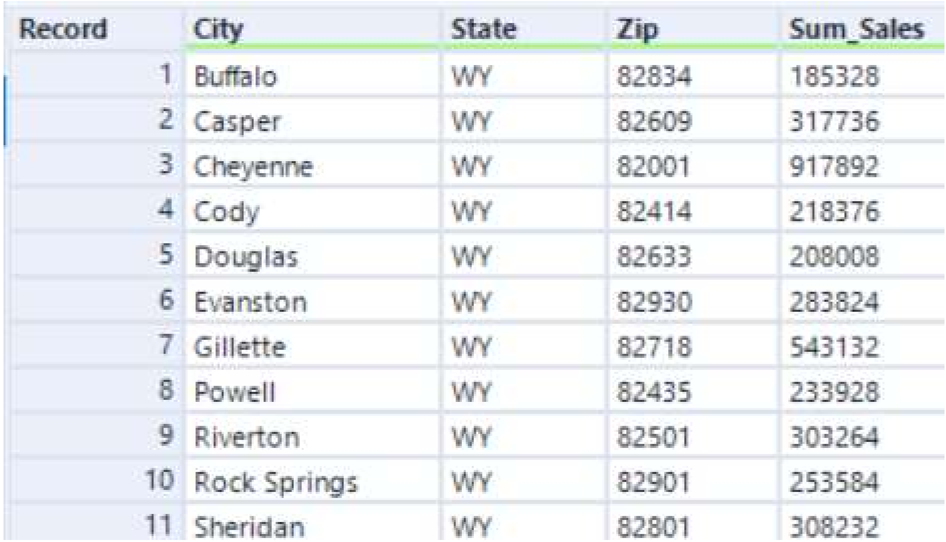
The dataset is reformatted into a 'Month' and 'Sales' column by using the Alteryx *Transpose* tool. The screenshot below shows the first few records in the dataset showing the new format:



City	State	Zip	Month	Sales
Buffalo	WY	82834	January	16200
Buffalo	WY	82834	February	13392
Buffalo	WY	82834	March	14688
Buffalo	WY	82834	April	17064

NOTE: The 'State' field is left in, even though it is the same across all the records. This will be useful later once all datasets are joined

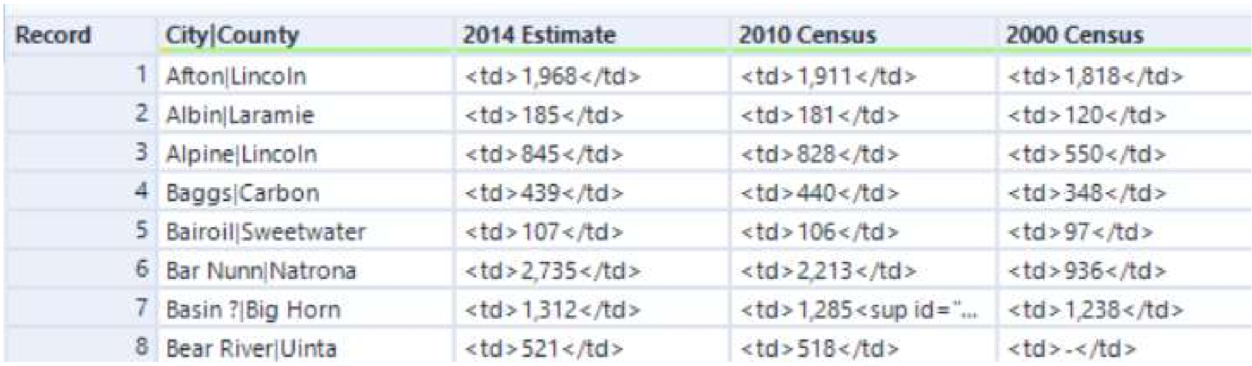
With the dataset reformatted, it becomes easy to aggregate the Sales by City by using the *Summarize* tool. Full results are shown below:



Record	City	State	Zip	Sum_Sales
1	Buffalo	WY	82834	185328
2	Casper	WY	82609	317736
3	Cheyenne	WY	82001	917892
4	Cody	WY	82414	218376
5	Douglas	WY	82633	208008
6	Evanston	WY	82930	283824
7	Gillette	WY	82718	543132
8	Powell	WY	82435	233928
9	Riverton	WY	82501	303264
10	Rock Springs	WY	82901	253584
11	Sheridan	WY	82801	308232

NOTE: While business identified that there are 13 Pawdacity locations, the original dataset is only provided sales data for 11 of those stores

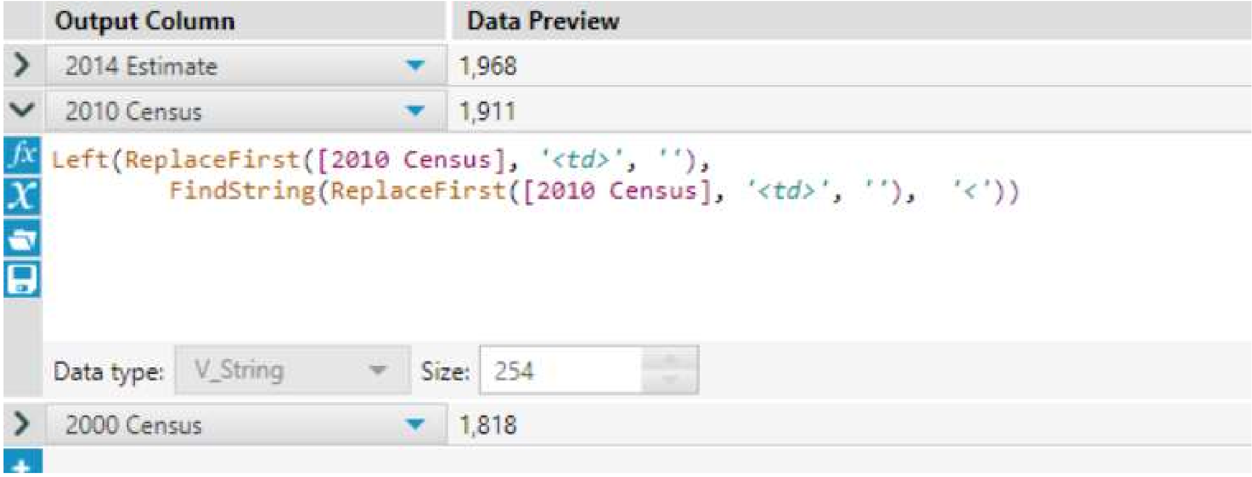
DATASET 2 - The Wyoming Census dataset provided includes Census information for different cities in Wyoming. The xml/html tags in the census data needs to be removed leaving just the integer values for the census. The 'commas' that remain will also be removed before converting to a numerical datatype



Record	City|County	2014 Estimate	2010 Census	2000 Census
1	Afton|Lincoln	<td>1,968</td>	<td>1,911</td>	<td>1,818</td>
2	Albin|Laramie	<td>185</td>	<td>181</td>	<td>120</td>
3	Alpine|Lincoln	<td>845</td>	<td>828</td>	<td>550</td>
4	Baggs|Carbon	<td>439</td>	<td>440</td>	<td>348</td>
5	Bairoil|Sweetwater	<td>107</td>	<td>106</td>	<td>97</td>
6	Bar Nunn|Natrona	<td>2,735</td>	<td>2,213</td>	<td>936</td>
7	Basin ?|Big Horn	<td>1,312</td>	<td>1,285<sup id="...	<td>1,238</td>
8	Bear River|Uinta	<td>521</td>	<td>518</td>	<td>-</td>

With the numbers handled, we need to separate 'City' from the 'County' values. This is important as the city names will be used as the key variable to combine the datasets later. Filter out any *null* values, we use the *Parse* tool, setting the delimiter to '|' and separate 'City' and 'County' into new auto incremented column names, to be renamed appropriately with a select tool later. A *Data Cleansing* tool is used to remove any extra character, like the question mark in record 7 above.

Next, we apply string formulas, using the *Formula* tool. The formula uses Find and Replace string functions to identify the variable position of the extra characters and then nest that inside a *Left* function to successfully extract the numerical value.



Output Column	Data Preview
2014 Estimate	1,968
2010 Census	1,911

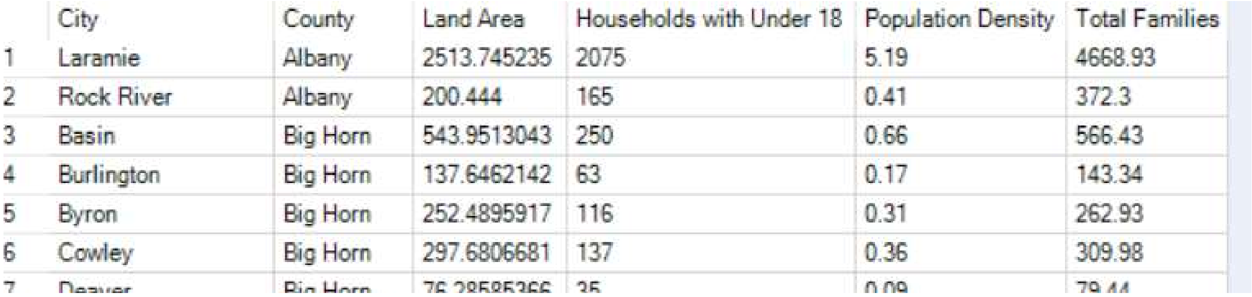
```
Left(ReplaceFirst([2010 Census], '<td>', ''),
        FindString(ReplaceFirst([2010 Census], '<td>', ''),  '<'))
```

Data type: V_String Size: 254

Output Column	Data Preview
2000 Census	1,818

NOTE: The formula was applied to all census fields including the 2010 data as they all followed the same patterns

DATASET 3 – The demographic dataset consists of demographic details for all cities in Wyoming. These are the main fields to be used as predictor variables when we build our model



	City	County	Land Area	Households with Under 18	Population Density	Total Families
1	Laramie	Albany	2513.745235	2075	5.19	4668.93
2	Rock River	Albany	200.444	165	0.41	372.3
3	Basin	Big Horn	543.9513043	250	0.66	566.43
4	Burlington	Big Horn	137.6462142	63	0.17	143.34
5	Byron	Big Horn	252.4895917	116	0.31	262.93
6	Cowley	Big Horn	297.6806681	137	0.36	309.98
7	Deaver	Big Horn	76.28585366	35	0.09	79.44

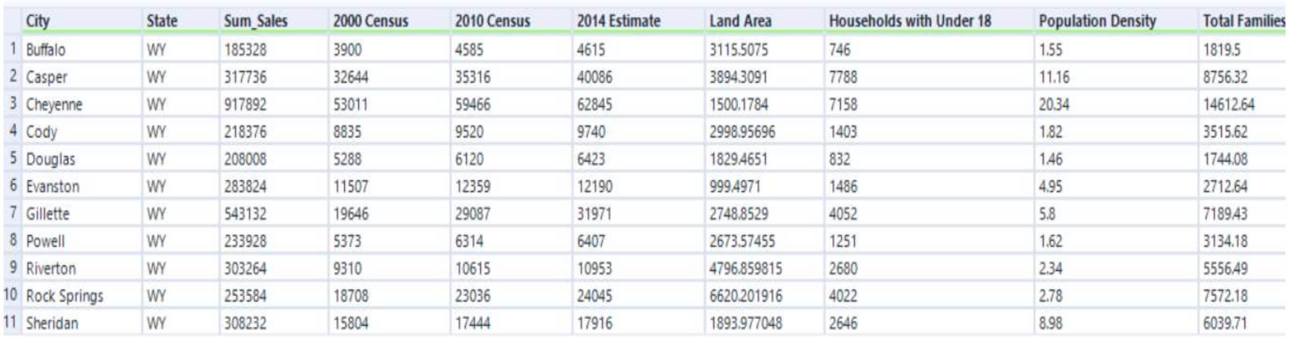
The demographic dataset is generally clean and well formatted and required basic select and datatype assignment steps. A addition to remove 'Trailing and leading whitespace' also resolved issues faced when doing the *joins* in the upcoming section. This is important to note as a great practice as it will always be missed by the naked eye.

STEP 2: Join

With all datasets prepared, in this step we combine them all to create a single dataset, using the unique *city* name as the key value for the joins.

The census and demographic datasets are joined first by 'city' name as the common field. This is then further joined with the sales data using 'city' again. Just as good practice, the common fields are left in to confirm the join, but then removed before proceeding to analysis. We can also check the 'outer' joins to see if any key joins were left out.

The **final dataset** includes the 11 store city locations with demographic and census data added in new columns. The complete dataset is shown below:



	City	State	Sum_Sales	2000 Census	2010 Census	2014 Estimate	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	WY	185328	3900	4585	4615	3115.5075	746	1.55	1819.5
2	Casper	WY	317736	32644	35316	40086	3894.3091	7788	11.16	8756.32
3	Cheyenne	WY	917892	53011	59466	62845	1500.1784	7158	20.34	14612.64
4	Cody	WY	218376	8835	9520	9740	2998.95696	1403	1.82	3515.62
5	Douglas	WY	208008	5288	6120	6423	1829.4651	832	1.46	1744.08
6	Evanston	WY	283824	11507	12359	12190	999.4971	1486	4.95	2712.64
7	Gillette	WY	543132	19646	29087	31971	2748.8529	4052	5.8	7189.43
8	Powell	WY	233928	5373	6314	6407	2673.57455	1251	1.62	3134.18
9	Riverton	WY	303264	9310	10615	10953	4796.859815	2680	2.34	5556.49
10	Rock Springs	WY	253584	18708	23036	24045	6620.201916	4022	2.78	7572.18
11	Sheridan	WY	308232	15804	17444	17916	1893.977048	2646	8.98	6039.71

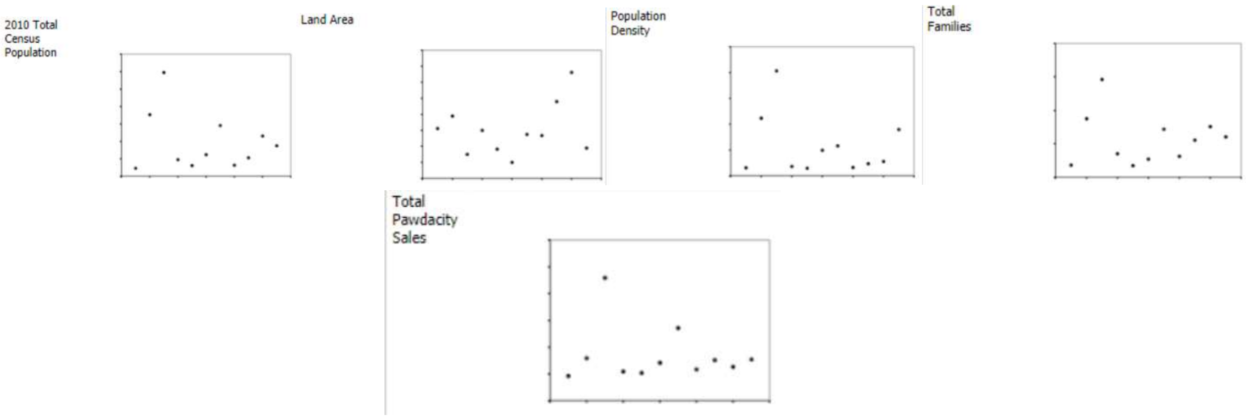
And below, are the aggregate for all the numerical:

Column	Sum	Average
Census Population	*213,862*	*19,442*
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*

With the data prepared for analysis, will continue outlier analysis in the next section.

Dealing with Outliers

We can get a general sense of the outliers by viewing the distributions of all the variables. For this we use the *Field Summary* tool in Alteryx

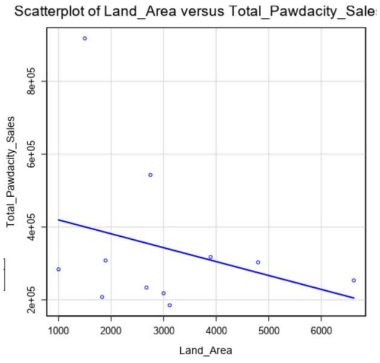


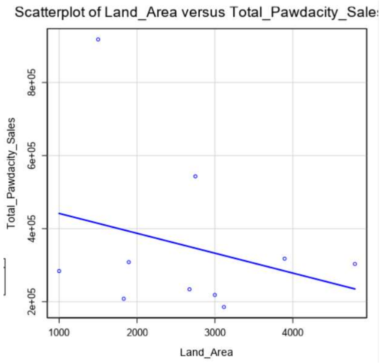
An interquartile analysis further confirms outliers outside the upper and lower fence. None of the values are below the lower fence, whereas we have a few values above the upper fence, which is calculated as being 1.5 times the IQR above the third quartile.

City	Total Pawdacity Sales	2010 Total Census Population	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	185328	4585	746	3115.5075	1.55	1819.5
Casper	317736	35316	7788	3894.3091	11.16	8756.32
Cheyenne	917892	59466	7158	1500.1784	20.34	14612.64
Cody	218376	9520	1403	2998.95696	1.82	3515.62
Douglas	208008	6120	832	1829.4651	1.46	1744.08
Evanston	283824	12359	1486	999.4971	4.95	2712.64
Gillette	543132	29087	4052	2748.8529	5.8	7189.43
Powell	233928	6314	1251	2673.57455	1.62	3134.18
Riverton	303264	10615	2680	4796.859815	2.34	5556.49
Rock Springs	253584	23036	4022	6620.201916	2.78	7572.18
Sheridan	308232	17444	2646	1893.977048	8.98	6039.71
Mean	140903.405	12570.36364	1933.289256	1163.803968	4.262479339	2853.043471
3Q	312984	26061.5	4037	3504.9083	7.39	7380.805
1Q	226152	7917	1327	1861.721074	1.72	2923.41
IQR	86832	18144.5	2710	1643.187226	5.67	4457.395
Upper Fence	443232	53278.25	8102	5969.689139	15.895	14066.8975
Lower Fence	95904	-19299.75	-2738	-603.059765	-6.785	-3762.6825

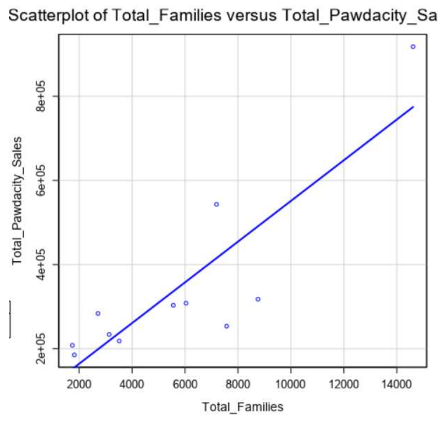
The results from the calculations provide the key cities and fields to observe when observing outliers. We will need to dive deeper into the outlier analysis of each city by observing the scatter plot of the predictor variables against the sales target variable.

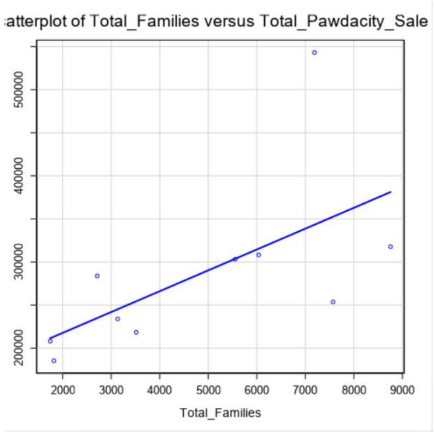
The *Land Area* for **Rock Springs** falls outside the fence, however when comparing the Sales-to-Land-Area relationship we can see that the trend line isn't different for the dataset with Rock Springs and the dataset without Rock Springs. We can also see that there isn't much of a trend but a negative relationship being created because of the sales above $50,000.

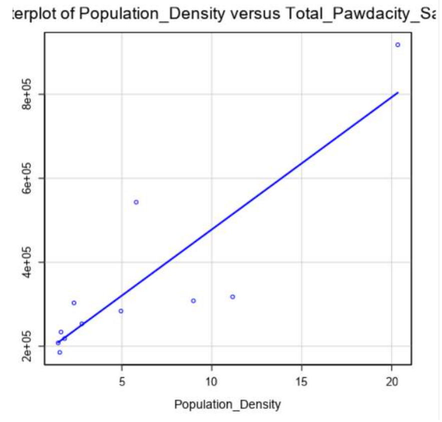


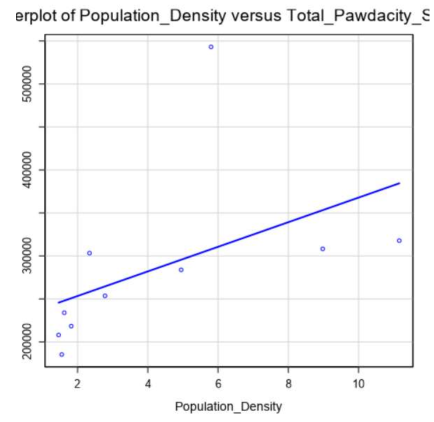


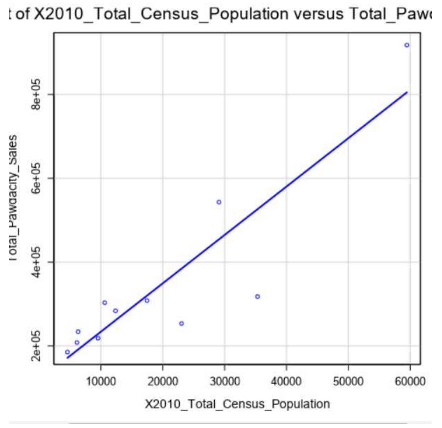
Cheyenne has the most outliers, however removing values for Cheyenne also maintains the same general trend, with the outliers veering very close to the trend line. The plots on the right with Cheyenne removed, identify the other sales value from **Gillette** to be more of an outlier, as without Cheyenne sales it has very little effect on the trend

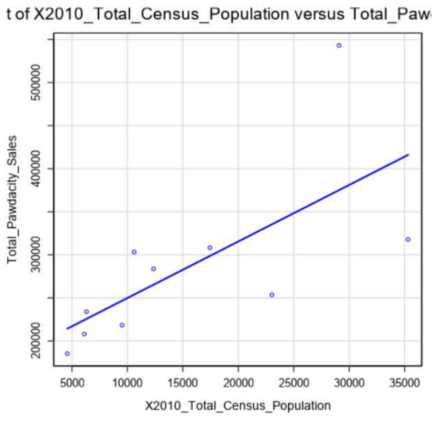












CONCLUSION: The recommendation moving forward is that the city of ‘Gillette’ be removed from the predictor models. This is also confirmed in spreadsheet analysis of points beyond the upper fence, as Gillette is the only city that has an outlier in the target variable, with no outlier identified amongst the potential predictor variables.