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 key decision to be made is determining the best location to open a new store. To determine the best location, the analyst would need to predict the yearly sales of various cities.

2. What data is needed to inform those decisions?
   The company would need to have historical data of the stores that includes the sales and location. External information that could be gathered such as population and average household income would be useful and impactful. Other data that may be difficult to obtain such as the location of competitors and their sales or even data on pet ownership could possibly positively influence the predicted yearly 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

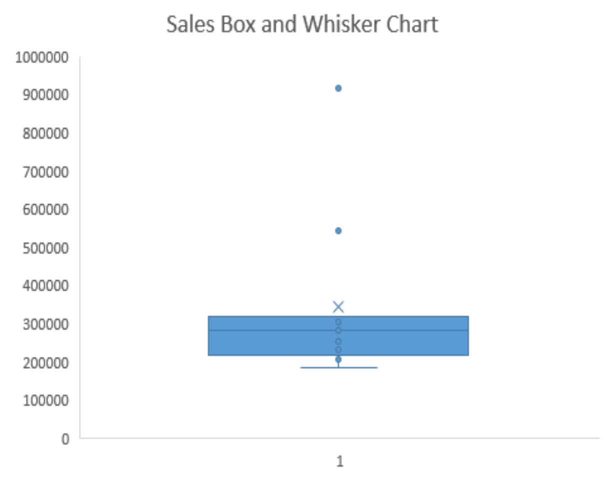
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

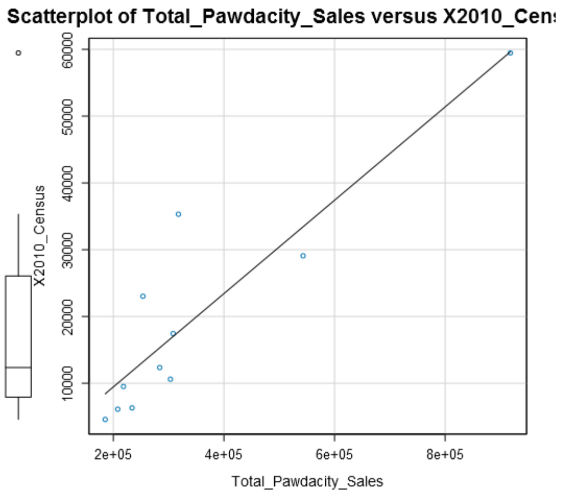
Step 3: Dealing with Outliers

Answer these questions

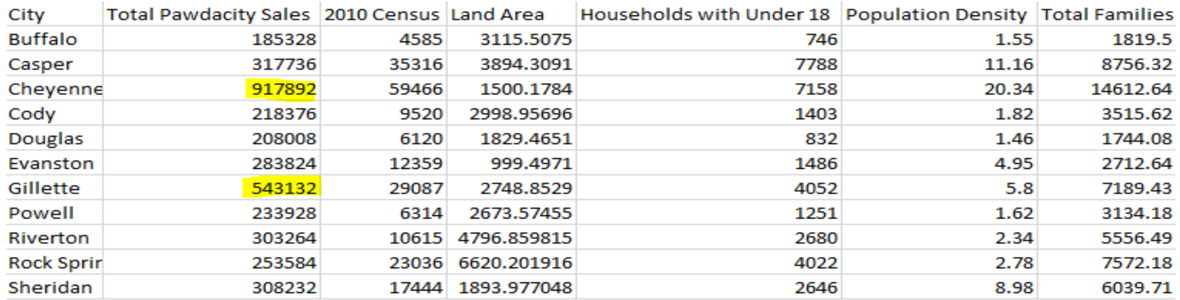
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

There is only 1 column that appears to have an outlier based from the field summary results, scatter plots and the Box and Whisker charts. The column being Total Pawdacity Sales.



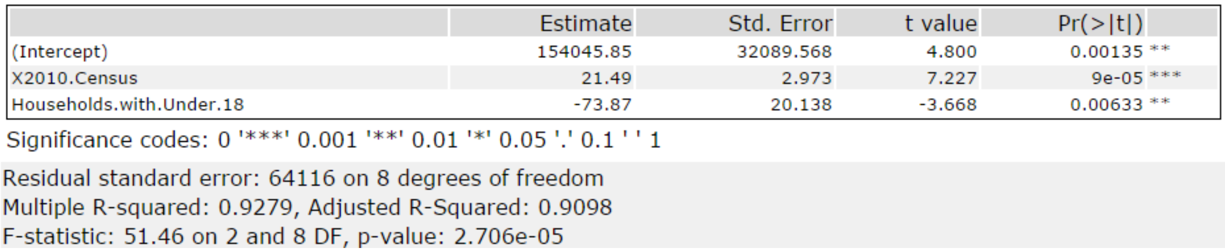


From the 11 cities, there are 2 cities that seem to have outliers and that is Cheyenne and Gillette.



City	Total Pawdacity Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.5075	746	1.55	1819.5
Casper	317736	35316	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Cody	218376	9520	2998.95696	1403	1.82	3515.62
Douglas	208008	6120	1829.4651	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.8529	4052	5.8	7189.43
Powell	233928	6314	2673.57455	1251	1.62	3134.18
Riverton	303264	10615	4796.859815	2680	2.34	5556.49
Rock Sprir	253584	23036	6620.201916	4022	2.78	7572.18
Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

I then made some linear models to see what values would make the best predictor variables which ended up being the census population and households under 18.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	154045.85	32089.568	4.800	0.00135 **
X2010.Census	21.49	2.973	7.227	9e-05 ***
Households.with.Under.18	-73.87	20.138	-3.668	0.00633 **

```
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64116 on 8 degrees of freedom
Multiple R-squared: 0.9279, Adjusted R-Squared: 0.9098
F-statistic: 51.46 on 2 and 8 DF, p-value: 2.706e-05
```

Now looking at Cheyenne, I can see that the population and the total number of families increase are consistent with the high total sales, so it is more likely for it not being that much of an outlier. However considering the best predictor variables from linear regression I did, Cheyenne's total number of households with under 18 is not the highest because Casper has the highest total and so it could leave it in question. So my decision for now is that I let Cayenne be part of the dataset but note it down. Gillette on the other hand has a population right below Casper but Gillette's total sales is not even close and that makes it odd. In fact with the data we have, Casper city has higher values in every field than Gillette has and so it would mean Casper should be having more sales. So considering everything, I think it would be best to consider Gillette as the outlier and remove it from the dataset.