Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

ANS. We need to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

: Awesome: Correct! This is indeed the main business decision to be made.

2. What data is needed to inform those decisions?

ANS. We have to predict the yearly sales for all the Pawdacity stores at the city level .To predict yearly sales we have to first format and blend together data from different datasets and deal with outliers.

The given datasets are:-

- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
- A partially parsed data file that can be used for population numbers.
- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

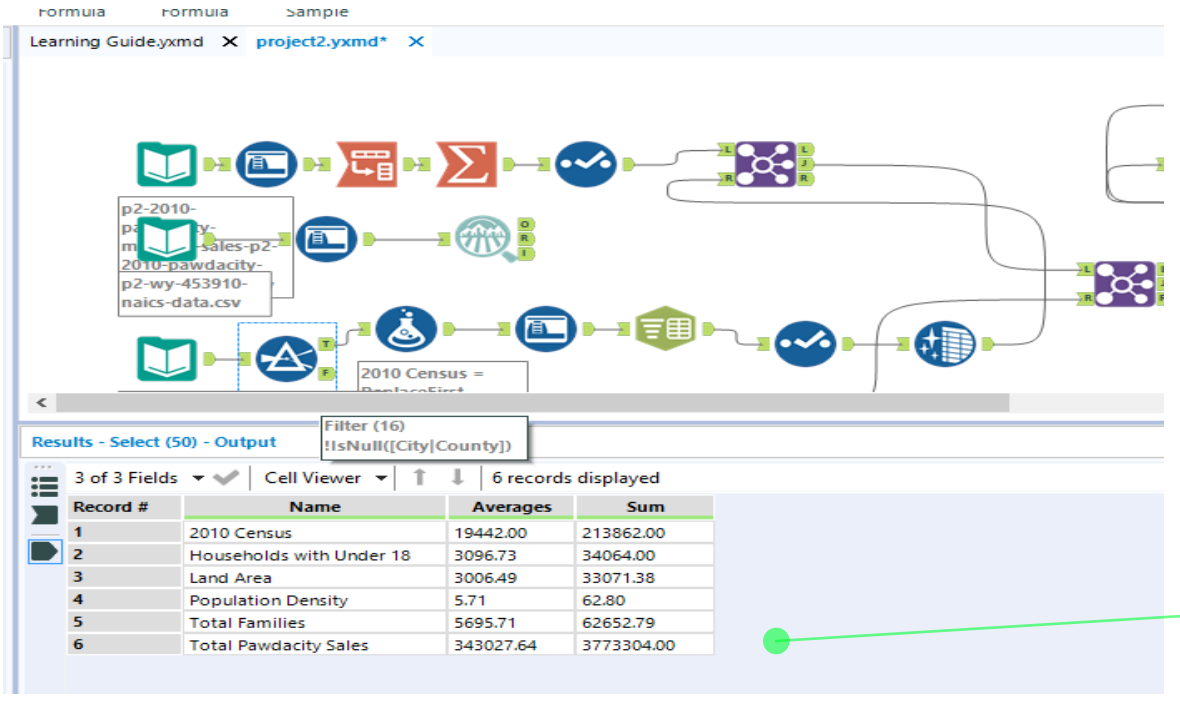
: Awesome: Good work identifying this. This data should be good enough for part 1 of the analysis.

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.

Column	Sum	Average
Census Population	*213,862*	
Total Pawdacity Sales	*3,773,304*	
Households with Under 18	*34,064*	
Land Area	*33,071*	
Population Density	*63*	
Total Families	*62,653*	

AFTER CLEANING THE DATASET THE REQUIRED SUM AND AVERAGES ARE (From Alteryx Workflow).

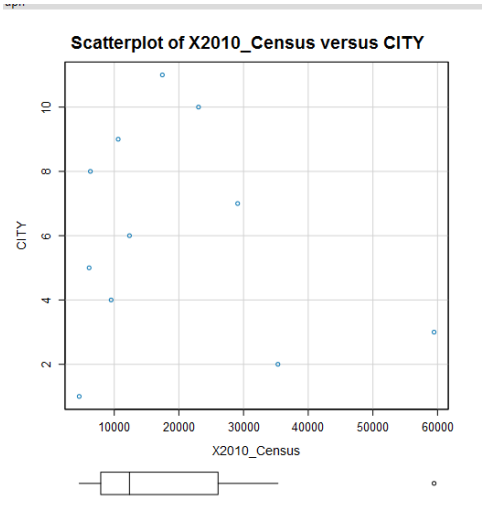


: Awesome: well done! All the sum & averages are perfectly correct!

Step 3: Dealing with Outliers

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.

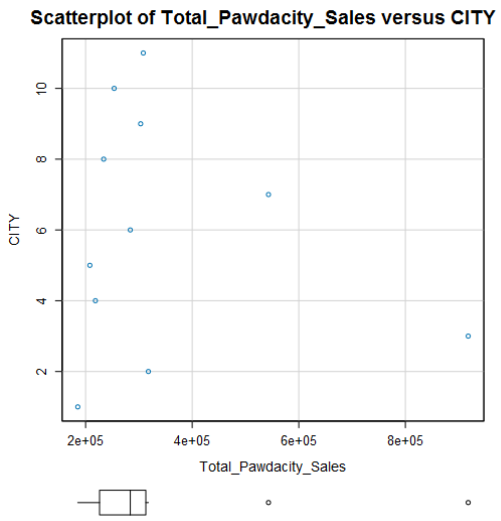
ANS. **Scatterplot for 2010_census versus City**:



As we can see from the scatter plot and box and whisker plot that City 3 i.e. Cheyenne has an outlier. 2010_census population for all the cites is around 35,000 but for the ‘Cheyenne’ it is about 60,000. So, the outlier exist at ‘Cheyenne’.

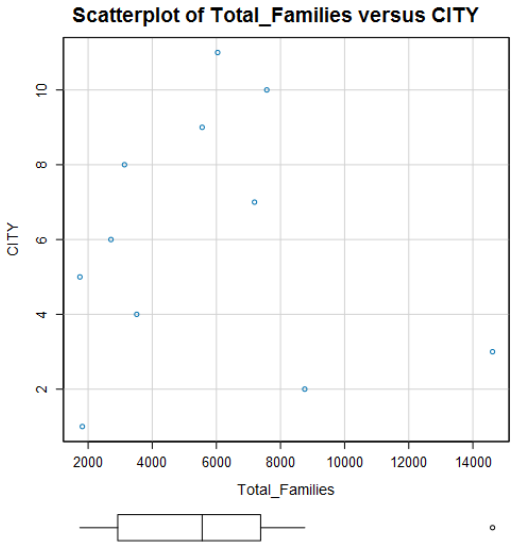
Scatterplot for Total Pawdacity Sales versus City:

As we can see from the scatter plot and box and whisker plot that there exists two outliers one for the City 7 i.e. ‘Gillete’ and the second for the City 3 i.e. Cheyenne. But the value of

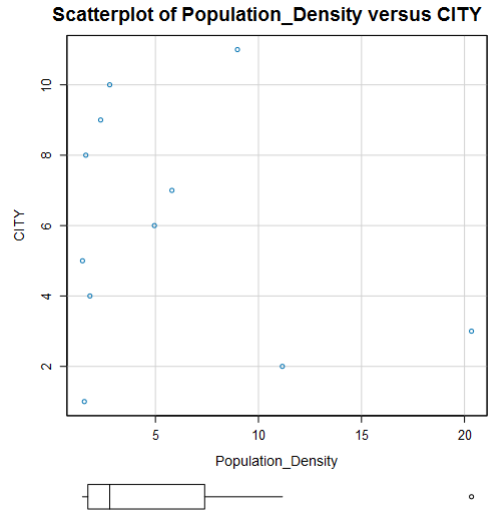


Total_Pawdacity_sales for the ‘Cheyenne’ is much higher than others which should be handled more carefully during analysis .So, the outlier exist at ‘Cheyenne’.

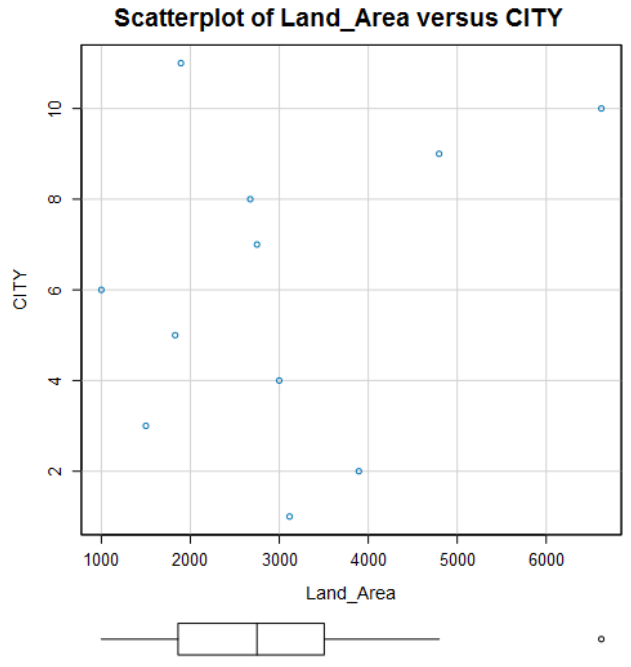
Scatterplot for Total FAMILY versus City:



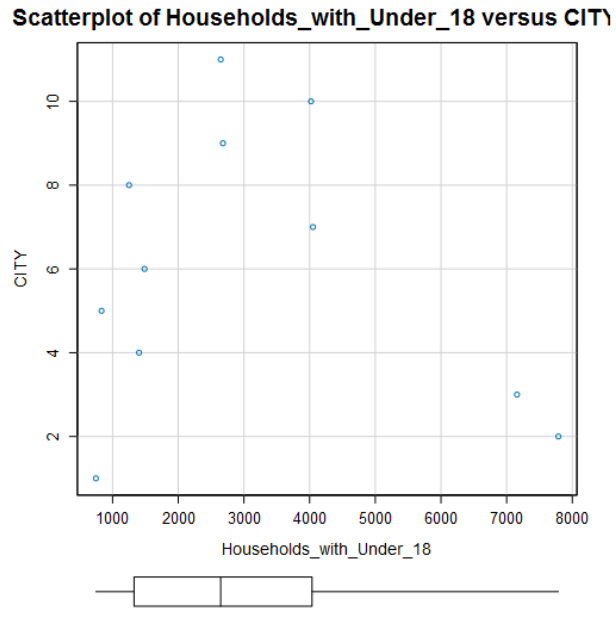
As we can see from the scatter plot and box and whisker plot that there exists a outlier for the City 3 i.e. Cheyenne. The total families existing in the city ’Cheyenne’ is much larger than other cities despite having a small area (as we can see from land area scatterplot) .So, the outlier exist at city ‘Cheyenne’.



Scatterplot for Population Density versus City:
As we can see from the scatter plot and box and whisker plot that there exists a outlier for the City 3 i.e. Cheyenne. Most of the cities have their population densities less than 6.0. But for the city ‘Cheyenne’ it is more than 20.0. So, the outlier exist at ‘Cheyenne’.



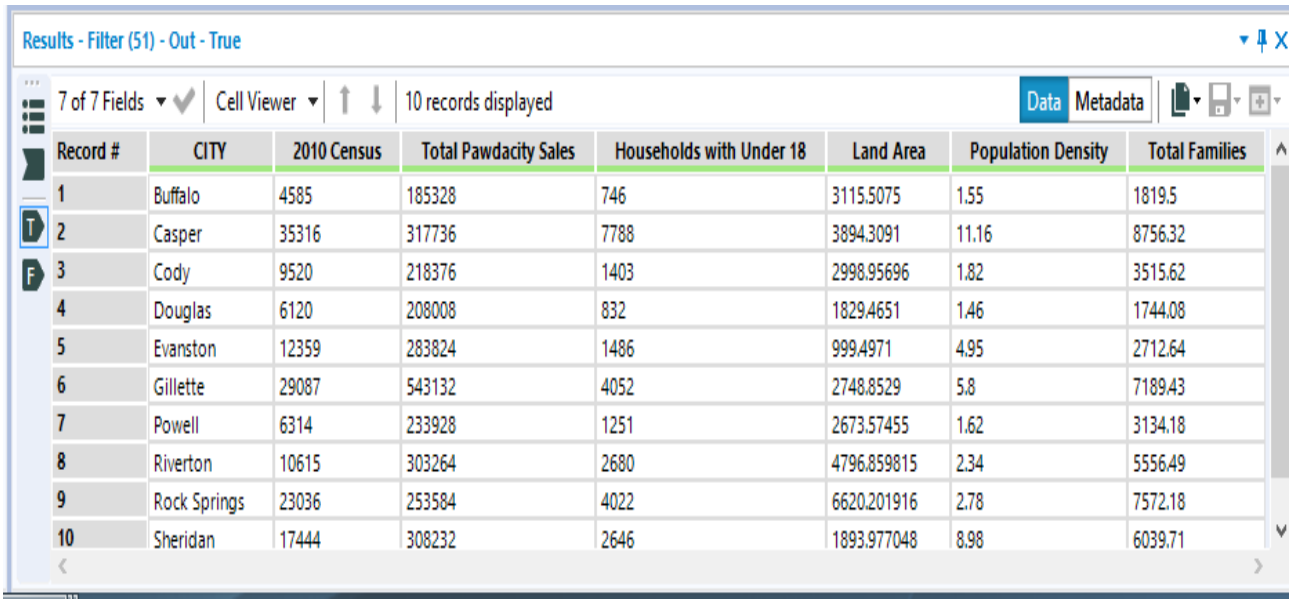
As we can see from the scatter plot and box and whisker plot that there exists a outlier for the City 10 i.e. Rock Springs.



There exists no outlier for this scatterplot

Since , most of the outliers are associated with the city 'Cheyenne ' so, It is better to remove this city from our cleaned dataset . Our dataset after all the necessary formatting ,blending and after removing outliers is (ready for analysis) :-

: Awesome: Cheyenne seems to be a big city , in midst of a dataset that contains small and medium sized cities. It has multiple outlier fields and even its other field values , are unlike the other cities in the dataset. Therefore, it can possibly skew our predictor model and thus, its removal or imputation from the dataset is justified.


Results - Filter (51) - Out - True

7 of 7 Fields | Cell Viewer | 10 records displayed | Data | Metadata

Record #	CITY	2010 Census	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
1	Buffalo	4585	185328	746	3115.5075	1.55	1819.5
2	Casper	35316	317736	7788	3894.3091	11.16	8756.32
3	Cody	9520	218376	1403	2998.95696	1.82	3515.62
4	Douglas	6120	208008	832	1829.4651	1.46	1744.08
5	Evanston	12359	283824	1486	999.4971	4.95	2712.64
6	Gillette	29087	543132	4052	2748.8529	5.8	7189.43
7	Powell	6314	233928	1251	2673.57455	1.62	3134.18
8	Riverton	10615	303264	2680	4796.859815	2.34	5556.49
9	Rock Springs	23036	253584	4022	6620.201916	2.78	7572.18
10	Sheridan	17444	308232	2646	1893.977048	8.98	6039.71