Project 2.1: Data Cleanup

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   Pawdacity wanted to open its 14th store, and the question here is to find which city would be the best choice. The rationale for the solution is that, first we conducted a linear regression model for the existing data, and made sales predictions for the new store. Finally, Pawdacity would choose the city that has good predicted sales based on the regression model.
   For this part, we conduct data wrangling and cleaning for predictions and business analysis.

2. What data is needed to inform those decisions?

   Based on the data provided, we can create a analytic dataset with information about sales in each city, along with city's demographic information such as population, number of households with people under 18, population density, land area, number of families.

   Data:
   2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.

   Partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.

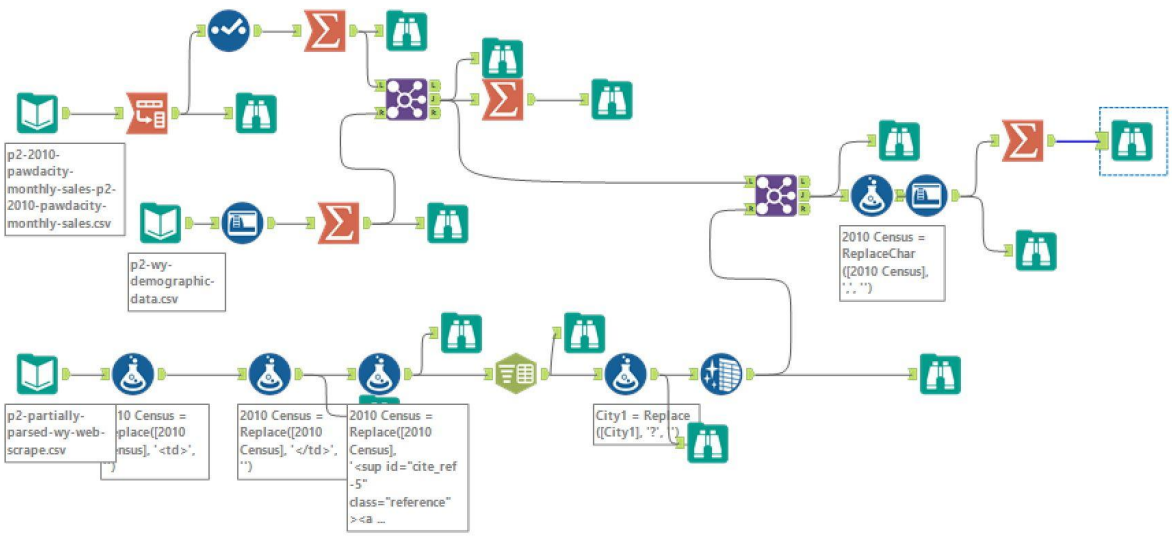
   wy-453910-naics-data.csv - NAICS data on the sales of all competitor stores where total sales is equal to 12 months of sales

   p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.

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.

Alteryx Workflow:



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

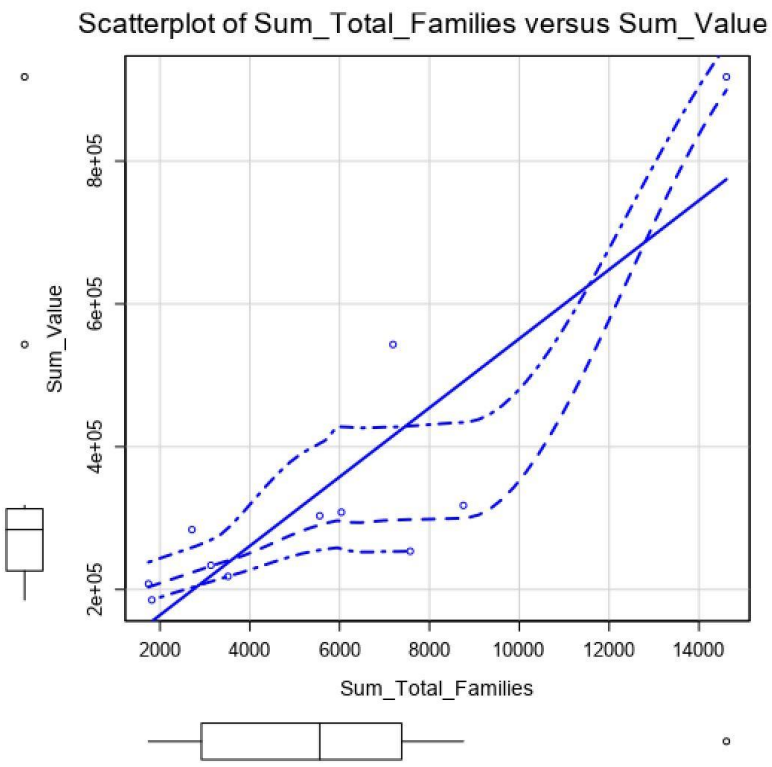
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

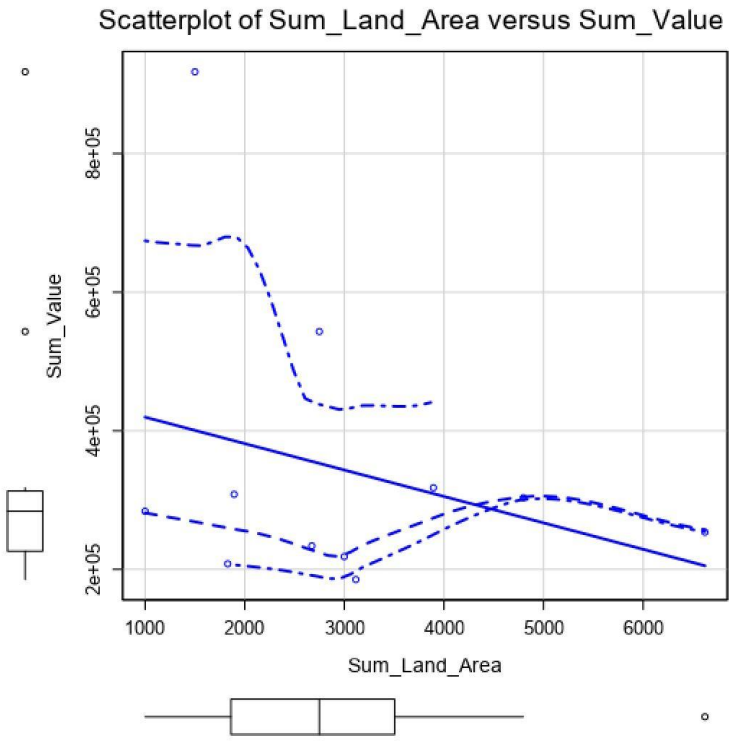
Below is our analytic dataset:

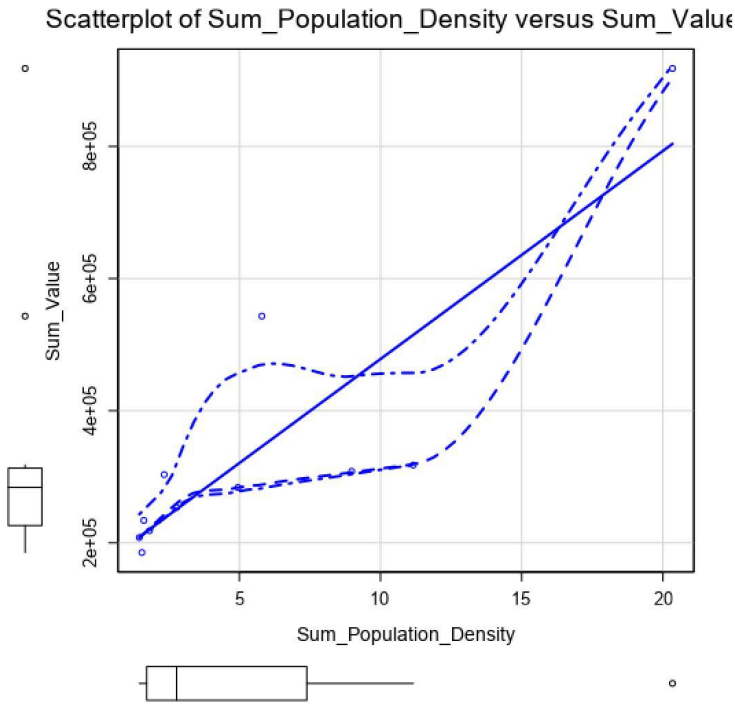
CITY	Total Pawdacity Sales	Land Area	Households with Under 18	Population Density	Total Families	Census Population
Buffalo	185,328.00	3,115.51	746.00	1.55	1,819.50	4,585.00
Casper	317,736.00	3,894.31	7,788.00	11.16	8,756.32	35,316.00
Cheyenne	917,892.00	1,500.18	7,158.00	20.34	14,612.64	59,466.00
Cody	218,376.00	2,998.96	1,403.00	1.82	3,515.62	9,520.00
Douglas	208,008.00	1,829.47	832.00	1.46	1,744.08	6,120.00

Evanston	283,824.00	999.50	1,486.00	4.95	2,712.64	12,359.00
Gillette	543,132.00	2,748.85	4,052.00	5.80	7,189.43	29,087.00
Powell	233,928.00	2,673.57	1,251.00	1.62	3,134.18	6,314.00
Riverton	303,264.00	4,796.86	2,680.00	2.34	5,556.49	10,615.00
Rock Springs	253,584.00	6,620.20	4,022.00	2.78	7,572.18	23,036.00
Sheridan	308,232.00	1,893.98	2,646.00	8.98	6,039.71	17,444.00
Quartile						
Q1	226,152.00	1,861.72	1,327.00	1.72	2,923.41	7,917.00
Q3	312,984.00	3,504.91	4,037.00	7.39	7,380.81	26,061.50
IQR	86,832.00	1,643.19	2,710.00	5.67	4,457.40	18,144.50
Upper Fence	443,232.00	5,969.69	8,102.00	15.90	14,066.90	53,278.25
Lower Fence	95,904.00	-603.06	-2,738.00	-6.79	-3,762.68	-19,299.75

EDA:


Scatterplot of Sum_Total_Families versus Sum_Value
Sum_Value
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Sum_Total_Families


Scatterplot of Sum_Land_Area versus Sum_Value
Sum_Value
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Sum_Land_Area



Outlier Analysis:
First, we can identify three outliers from the dataset.
Cheyenne: Total Sales, Population Density, Total Families and Censor Population
Gillette: Total Sales
Rock Springs: Land Area

Considering the small size of the dataset, I would only remove one outlier, Cheyenne's data. Because the result showed that for Total Sales, Population Density, Total Families and Censor Population, Cheyenne's records all exceeded their Upper Fence.
Also, from the scatter plots above, we noticed that Cheyenne's data for the relationship between Total Sales and Land Value is far from the trend line, so we may want to see if the Land Value have any impacts on total sales after excluding this extreme data.
Cheyenne is the capital city of Wyoming, considering its large population and other characteristics associated with capital cities, including these numbers in our model may be misleading.

I would keep Gillette and Rock Springs's data, because on the one hand, their values were not very far from Upper Fence as Cheyenne's data were. On the other hand, I would like to include these cities in my model so that we can have an understanding of the overall Wyoming environment, instead of putting too much weight on the capital city.