Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

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 Pawdacity pet shop, which is the leading pet shop in Wyoming with 13 stores, would like to open a 14th store this year. I was asked to perform an analysis to recommend the city for Pawdacity newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?
   - The monthly sales data for all 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.

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*	*19.44*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.5*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

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.

To check for outliers, I started with some scatterplots to visualize the relationship between each predictor variable and the target variable. Since there are five predictor variables, I dragged in five scatterplot tools and connected them to the cleaned dataset. Then I configured each one with a different predictor variable, attach browse tools, and run the workflow.
Additionally, I used another technique (Interquartile Range Method - IQR) to identify whether or not and how many cities has outliers

Total Pawdacity Sales
Looking for the Interquartile division we could see two outliers: Gillette city and Cheyenne city. For comparison purposes the Pawdacity Sales in Cheyenne City is 3x greater than the Q3.

2010 Census Population
For the census population, we have only one outlier: Cheyenne.

Households with under 18 years
Looking for the Households with under 18 years variable, there doesn't appear to be outliers.

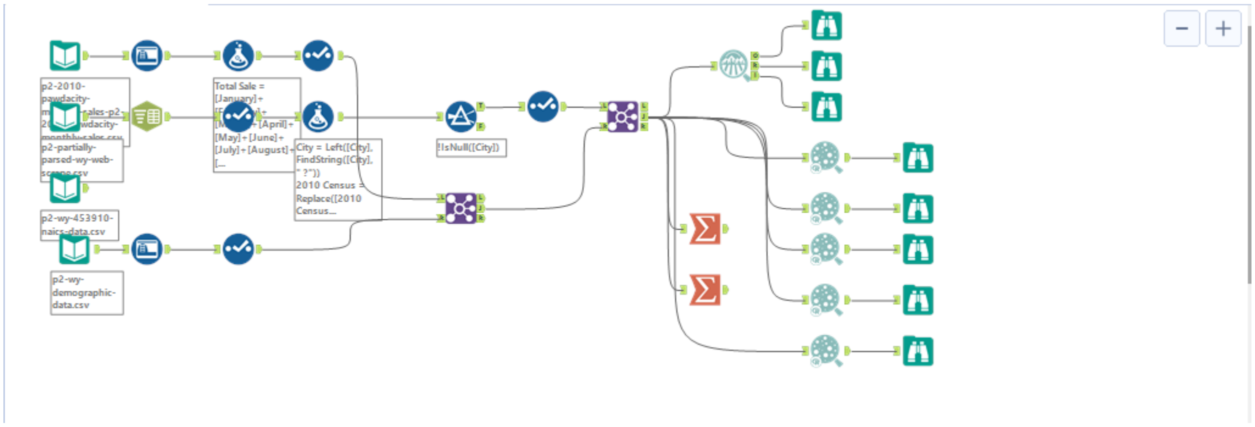
Land Area
For the Land Area variable, only the Rock Springs shows as an outlier.

Population Density
Looking at Population Density data, only one city that stands out from the all other as an outlier: Cheyenne

Total Families
Only in Cheyenne the Total Families data stands as an outlier.

Outlier Summary
After detailed observation of all the outliers of the data, I can conclude that doesn't seem like there is any typo error and all the data seem to like to be correct. On the other hand, Cheyenne city is clearly a very specific case in comparison with the other cities. Cheyenne has a high population density and very high revenue, even though it is one of the smallest cities of the state. The probability of having all these specificities in another city of the state is very low and because of it, this city will be removed from the dataset. So now we are ready for modeling.

Appendix
Alteryx Workflow


p2-2010-pawdacity-m…-sales-p2-20…wdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv
Total Sale = [January]+ [F…]+ [M…]+[April]+ [May]+[June]+ [July]+[August]+ [...
City = Left([City], FindString([City], " ?")) 2010 Census = Replace([2010 Census...
!IsNull([City])
T
F
L
J
R
−
+