Christopher W. Hong

Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. The company is considering opening a 14th store in the state this year.

2. What data is needed to inform those decisions?

   Historical data about the 13 stores, including total year sales, city population, population density, number of families, land area, etc.

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*	*19442*
Total Pawdacity Sales	*3,773,304*	343027.64
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*

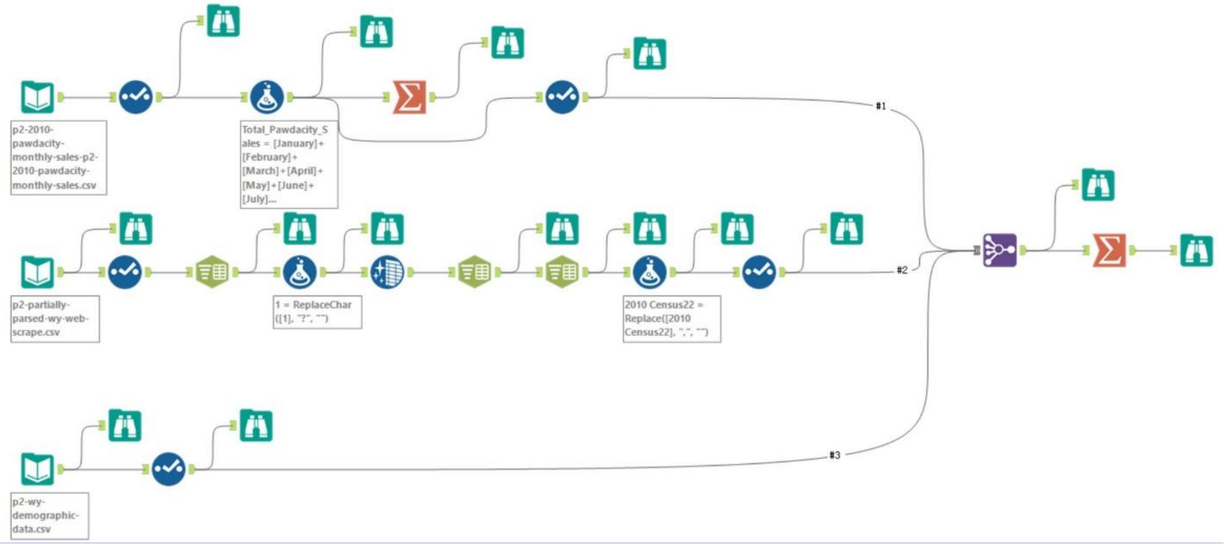
Step 3: Dealing with Outliers

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.

Based on the upper fence and lower fence values of each quantitative variables, Cheyenne is identified as outlier in variables Total Pawdacity Sales and Population Density; so is Gillette in variable Total Pawdacity Sales.

By inspecting the records of Cheyenne and Gillette, the are identified as outliers not because of input error. Because of the high value of Population Density of Cheyenne, it explains why the value of Total Pawdacity Sales is identified as outlier. Plus, the dataset is extremely small. Thus, it would be better to impute these two outliers instead of removing them. A linear regression model could be built in terms of predictors such as Total Families, Land Area, Households with Under, etc. to impute the Total Pawdacity Sales of Cheyenne and Gillette. If one outlier has to be removed, Cheyenne is suggested.

Below is the Alteryx workflow on cleaning the dataset:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total_Pawdacity_Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
#1
p2-partially-parsed-wy-web-scrape.csv
1 = ReplaceChar ([1], "?", "")
2010 Census22 = Replace([2010 Census22], ",", "")
#2
p2-wy-demographic-data.csv
#3