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:

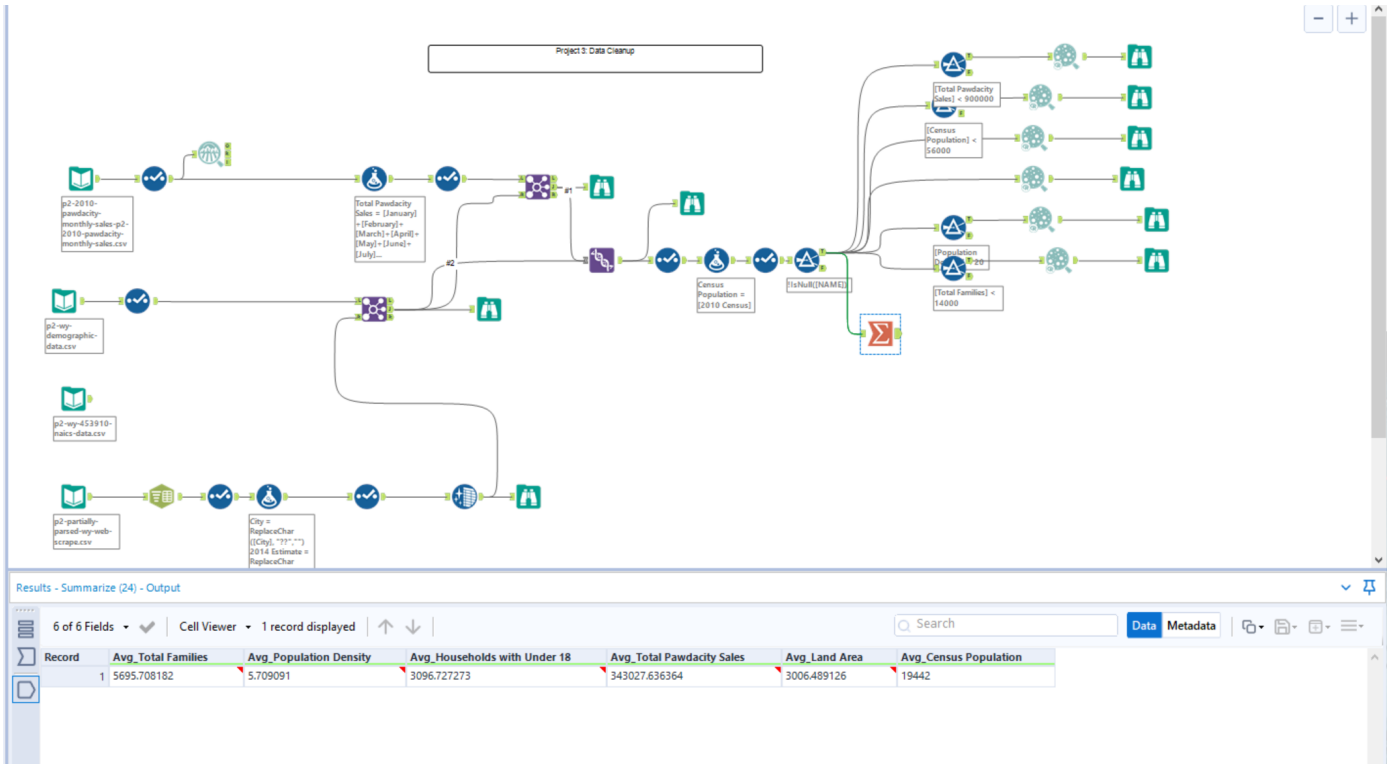
Answer these questions

1. What decisions needs to be made?
   The various streams of data have to be cleaned, proper columns made, missing data removed then joined together to produce the census population, Total Pawdacity sales, Household with under 18, Census population, Land area, Population density and Total families columns. These will be used a as a training set for a regression model.

2. What data is needed to inform those decisions?
   What method should be used to clean the data.
   What method should be used to handle missing data.
   What new columns should be made.
   What method should be used to deal with outliers.

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.
The outlier chosen is Cheyenne City because of the total families (14,612), population density (20.34) and Total Pawcity sales (917892) values. These values are significantly higher than values from other Cities so it will affect the slope of our plots.
I have chosen to remove Cheyenne because it is unlike other cities in the dataset.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.