Dhuha Baqarish
PAND

Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**
   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store .
   I will perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.
   I will that by build a training set which will include City, 2010 Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density and Total Families columns. Then I will calculate the upper fence and the lower fence to find the outlier cities and at last one outlier city will be Excluded**.**

2. **What data is needed to inform those decisions?**
    We are given four files of dataset (p2-2010-pawdacity-monthly-sales.csv, p2-partially-parsed-wy-web-scrape.csv, p2-wy-453910-naics-data.csv and p2-wy-demographic-data.csv), From this four files we need :
   - City
   - Census Population
   - Total Pawdacity Sales
   - Households with Under 18
   - Land Area
   - Population Density
   - Total Families

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.
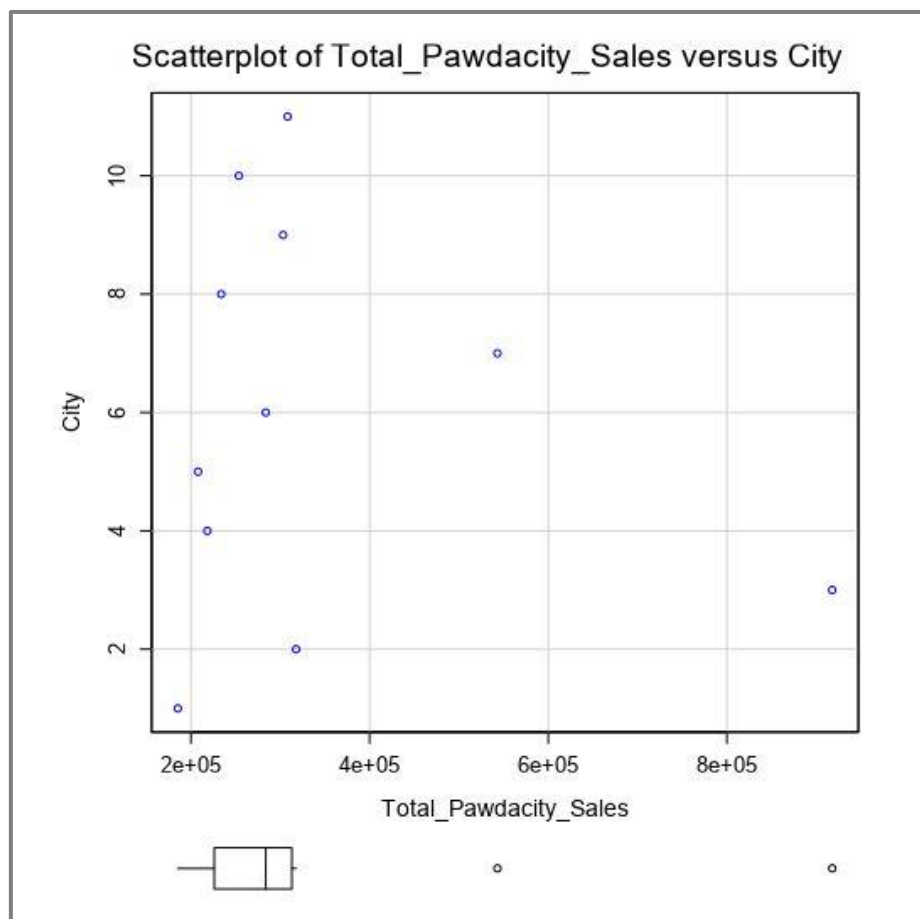
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24
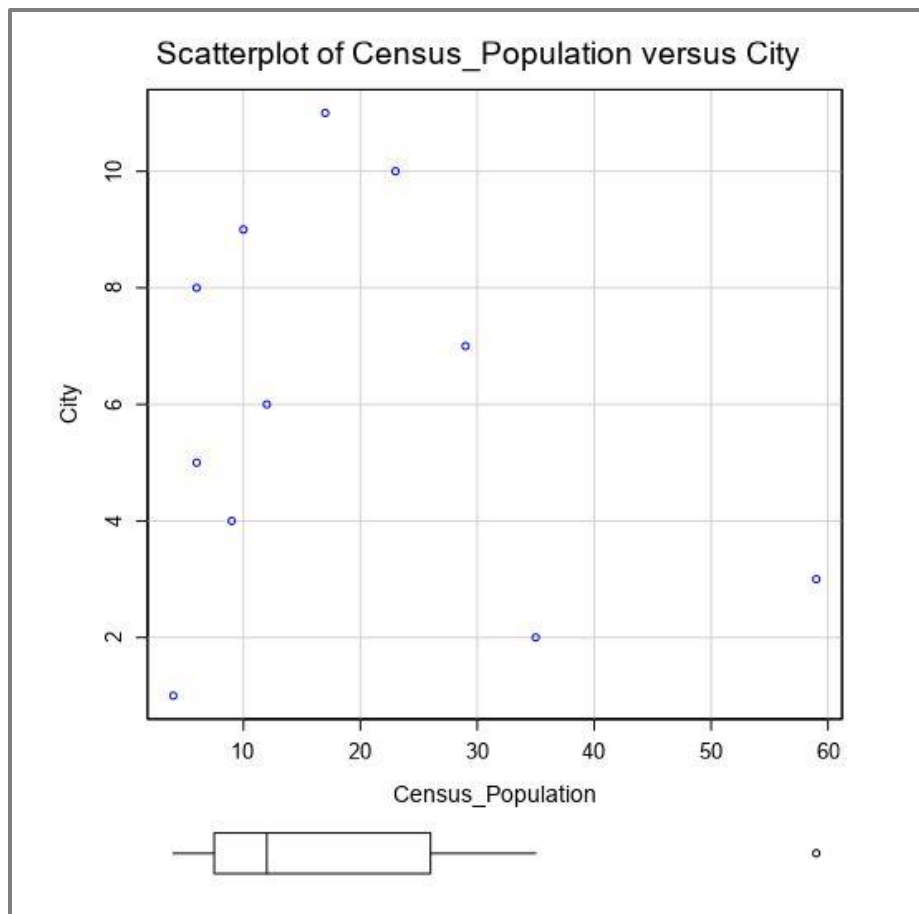
Column	Sum	Average
Census Population	*213,862*	19.09
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.49
Population Density	*63*	5.71
Total Families	*62,653*	5695.71

Step 3: Dealing with Outliers

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.

Scatterplot of Total_Pawdacity_Sales versus City

Scatterplot of Census_Population versus City

As we see in two different graph a city with number 3 " Cheyenne " was represent as outlier , so we will exclude this city from recommend and remove it from dataset .

Alteryx Workflow :