Project 2.1: Data Cleanup

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?

Where to open the new pet shop for Pawdacity? In which city?

2. What data is needed to inform those decisions?

Previous sales for all current branches, population information for each city (density, families, households with under 18).

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

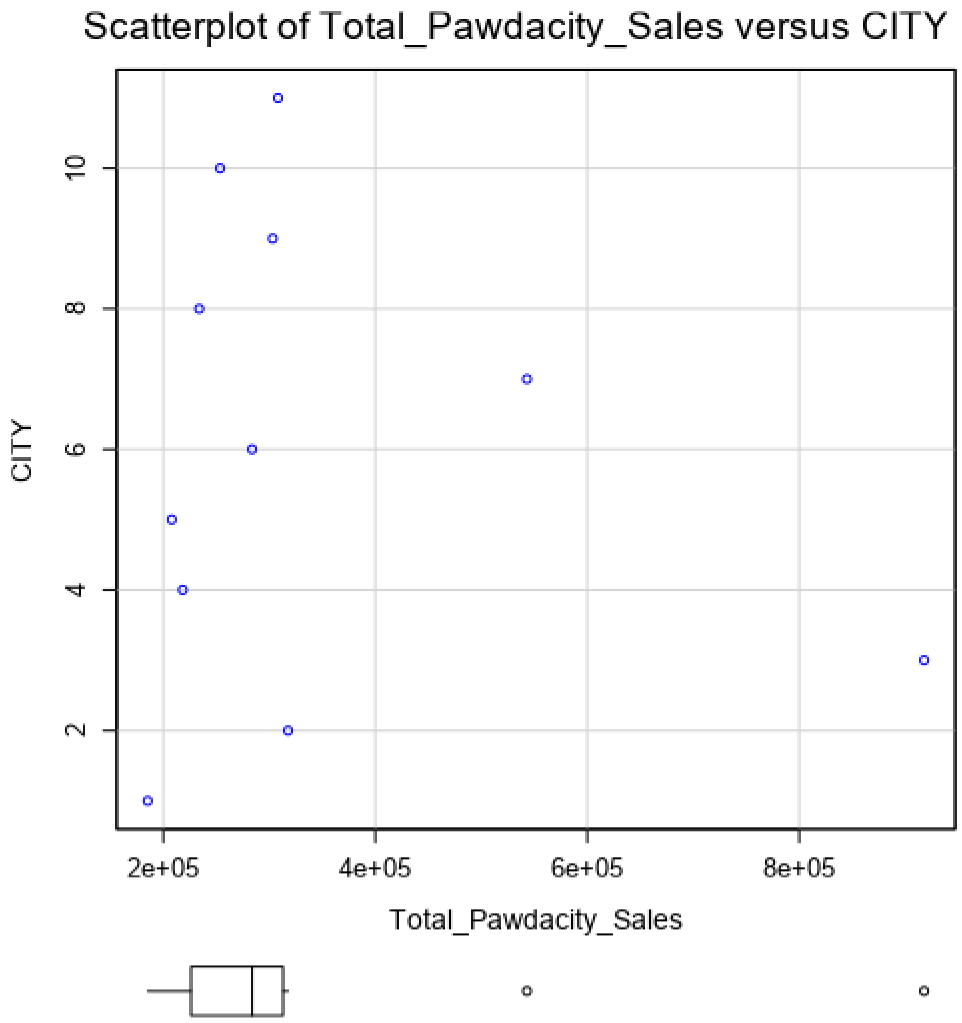
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.45*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

Step 3: Dealing with Outliers

Answer these questions

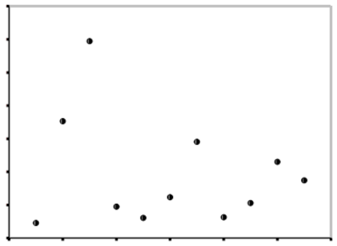
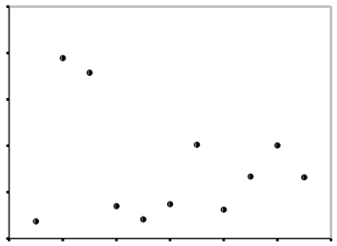
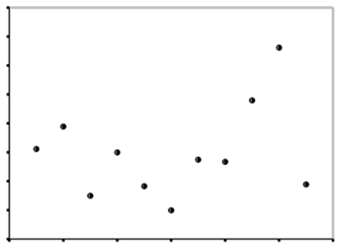
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

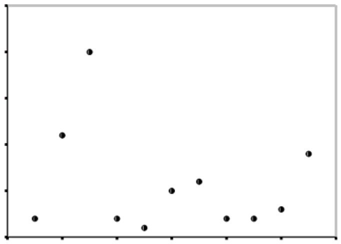
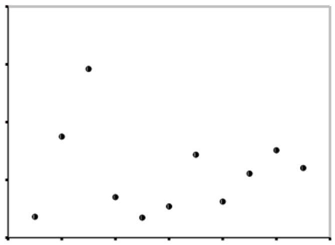
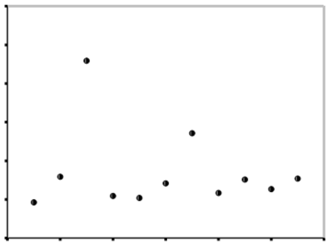
Since I care about total sales, so, I scattered plot it



We can see that, city 3 (Cheyenne) with sales amount = 917,892 $ and city 7 (Gillette) with sales amount = 543,132 $ are outliers, since they are above upper fence which is 443,232 $. After studying figures below:

Numeric Fields

Name	Plot
2010 Census	
Households with Under 18	
Land Area	

Population Density	
Total Families	
Total Pawdacity Sales	

We can see that, city 3 (Cheyenne) has a high population and families, so its high total sales is justified. While city 7 (Gillette) total sales is near the fence and the number can be considered as acceptable.

The decision to remove a city is complicated but for me, I will remove Cheyenne not because it wrong but because it will affect my model and as can be seen from above a few cities will have these characteristics of population.