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?

Decide the city for the Pawdacity's newest store.

2. What data is needed to inform those decisions?

    1. The monthly sales data for the Pawdacity 13 stores in 2010
    2. NAICS data on the total sales of all competitor stores
    3. The Wyoming's population numbers
    4. Demographic data for Wyoming's cities and countries, including Households with individuals under 18, Land Area, Population Density, and 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*	19,442
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.

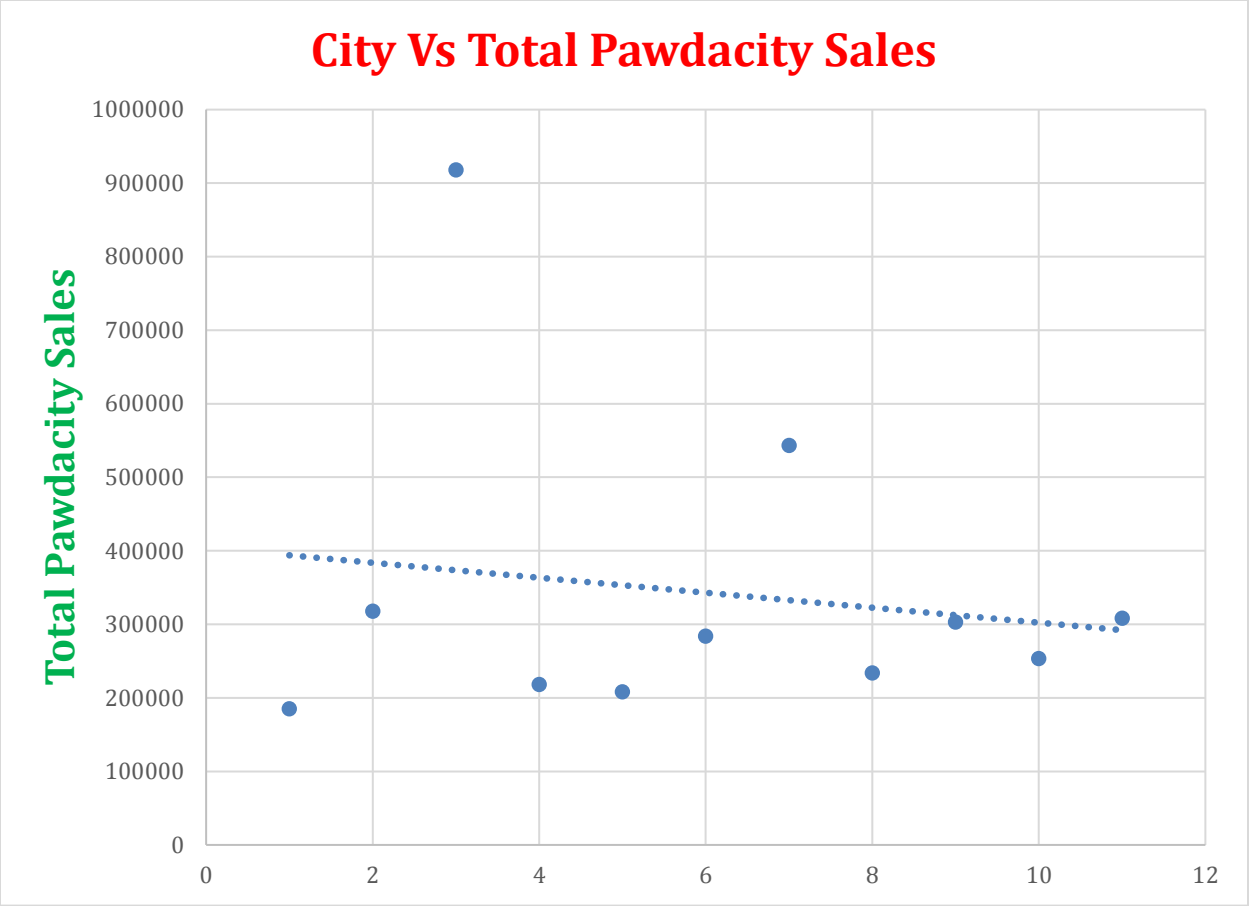
- After calculating the values of the upper fence and the lower fence for each field in the training set, **three outliers** were found, two of them were in the Total Pawdacity Sales field and the last one was in Population Density.

- There were **two** outliers cities in the training set:
  1. **Cheyenne city** that has **917,892** sales and **20.34** Population Density
  2. **Gillette city** that has **543132** sales

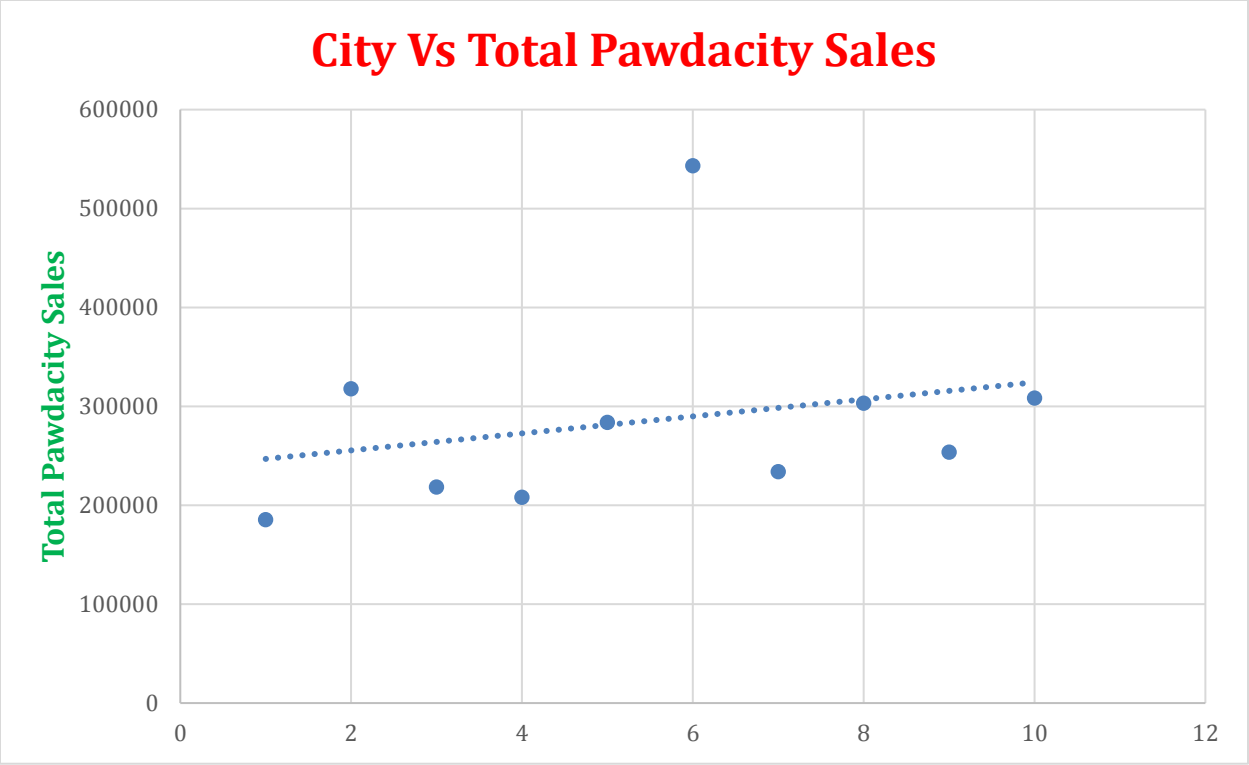
The **Cheyenne city** was chosen to be removed

- It is an abnormal data with two outlier fields.
- It has values closer to the upper fence of different fields.
- It has fields with larger values that the other cities in the training dataset.
- It has a significant impact on the linear model, when it's removed, the slope increases.

a. The scatter plot with Cheyenne city



b. The scatter plot without Cheyenne city



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.