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?

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. Based on predicted yearly sales, the manager of the company would like to understand which city is the best city to set up a Pawdacity’s newest store,

1. What data is needed to inform those decisions?

* The monthly sales data for all of the Pawdacity stores for the year 2010.
* A partially parsed data file that can be used for population numbers.
* Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

## 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* | *19422* |
| *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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CITY** | **Total Pawdacity Sales** | **Households with Under 18** | **Total Families** | **Land Area** | **Census Population** | **Population Density** |
| Buffalo | 185328 | 746 | 1819.5 | 3115.5075 | 4585 | 1.55 |
| Casper | 317736 | 7788 | 8756.32 | 3894.3091 | 35316 | 11.16 |
| Cheyenne | 917892 | 7158 | 14612.64 | 1500.1784 | 59466 | 20.34 |
| Cody | 218376 | 1403 | 3515.62 | 2998.95696 | 9520 | 1.82 |
| Douglas | 208008 | 832 | 1744.08 | 1829.4651 | 6120 | 1.46 |
| Evanston | 283824 | 1486 | 2712.64 | 999.4971 | 12359 | 4.95 |
| Gillette | 543132 | 4052 | 7189.43 | 2748.8529 | 29087 | 5.8 |
| Powell | 233928 | 1251 | 3134.18 | 2673.57455 | 6314 | 1.62 |
| Riverton | 303264 | 2680 | 5556.49 | 4796.859815 | 10615 | 2.34 |
| Rock Springs | 253584 | 4022 | 7572.18 | 6620.201916 | 23036 | 2.78 |
| Sheridan | 308232 | 2646 | 6039.71 | 1893.977048 | 17444 | 8.98 |
| **Q1** | 226152 | 1327 | 2923.41 | 1861.721074 | 7917 | 1.72 |
| **Q3** | 312984 | 4037 | 7380.805 | 3504.9083 | 26061.5 | 7.39 |
| **IQR** | 86832 | 2710 | 4457.395 | 1643.187226 | 18144.5 | 5.67 |
| **Upper Fence** | 443232 | 8102 | 14066.9 | 5969.689139 | 53278.25 | 15.895 |
| **Lower Fence** | 95904 | -2738 | -3762.683 | -603.059765 | -19299.75 | -6.785 |

Yes, based on the sales volume, Cheyenne and Gillette were identified as potential outliers using the Box and Whisker plot approach. Of both, Cheyenne had much more sales volume compared to the rest of the cities, including Gillette. Besides, Cheyenne also was found to be outliers with high population density, high total families, and high census population in 2010. The Cheyenne is probably the metropolitan city in Wyoming. For this reason, in order to match with other smaller cities, I would exclude Cheyenne in further analysis.

**Appendix – Alteryx workflow**

**Diagram

Description automatically generated**

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/382/view) here. Reviewers will use this rubric to grade your project.