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

### *Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. Company is planning to start another store in Wyoming.*

*Leadership needs an analysis which can recommend the location (City) where this new store can be setup based on the yearly sales.*

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

*City should be considered to setup a new Pawdacity’s 14th store.*

1. What data is needed to inform those decisions?

*We need to analyze the yearly sales for various cities in Wyoming. Below data would be useful:*

* *Sales data of Pawdacity stores for the year 2010*
* *Demographic data - Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming*
* *Overall Population numbers*
* *Competitors sales data for 12 months*

## 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* | *19442* |
| *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.

*As per the IQR method, there are 3 cities can be considered as outliers:*

***Cheyenne****: Outliers are present in “2010 Census Population”, “Total Pawdacity Sales”,*

*“Population Density”, “Total Families” columns.*

***Gillette:*** *Outlier is present in “Total Pawdacity Sales” column*

***Rock Springs:*** *Outlier is present in “Land Area”*

*We would choose to remove the Cheyenne city as outlier as there are multiple outliers present in for this city. We may need to impute the values for Gillette and Rock Springs cities.*

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.