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?
2. What data is needed to inform those decisions?

***Answer:***

**Question 1**

The aim of this project is to find a new city store location for business expansion. This decision will be based on the predicted yearly sales of the company.

**Question 2**

To make this decision, some data are needed to do this analysis. These are:

* Monthly sales data of Pawdacity stores
* The most current sales of all competitor stores
* Population numbers
* Demographic data of Wyoming

Then from those datasets we will find information about

* City
* 2010 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* | *19,442* |
| *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.

***Answer:***

Initially I used Excel to identify the outliers in our dataset. As a result, there are three city that identified as an outlier.

A screen shot of a building

Description automatically generated

Since only one city allowed to remove or impute, I will analyze the outlier one by one. First at Rock Springs is only slightly different in Land Area. Hence, I will keep this city. Next I will use scatter plot to see the trend of each variable compared to total sales before and after the outlier removal. Since this value will be our target variables.

|  |  |  |
| --- | --- | --- |
| Original | Remove Cheyenne | Remove Gillete |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

By Examining the scatter plot results and data on the table, even though Cheyenne is become an outlier in for variables. The total sales values still in the same proportion with other variable like total population. It makes sense that Cheyenne store have higher value due to the higher population. So, I decided to keep Cheyenne on the dataset.

On the other hand, Gillete total sales is not inline with other variable values. This city has higher total sales even though only have small population. Hence, Gillete will be excluded from the dataset in order to build an unbiased model

***Appendix***

![A close up of a map

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.