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?

Determine the Wyoming city where the new store for Pawdacity will be open based on the higher yearly sales.

1. What data is needed to inform those decisions?

Yearly sales by city, population, population density, land area, total families can be used to predict yearly sales.

## 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* | *19,442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *34,3027.64* |
| *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.

It has been removed the record for Gillette city, because it skews high in sales and potentially will affect the sales predictions, although not skew relative to the other data fields in the training set per the *Table 1 IQR analysis* shown below. And the record for Cheyenne city has been kept because per the *table 2 for Scatter plots analysis shown* below it is in line with the linear relationship and removing it will affect the sales predictions considerably.

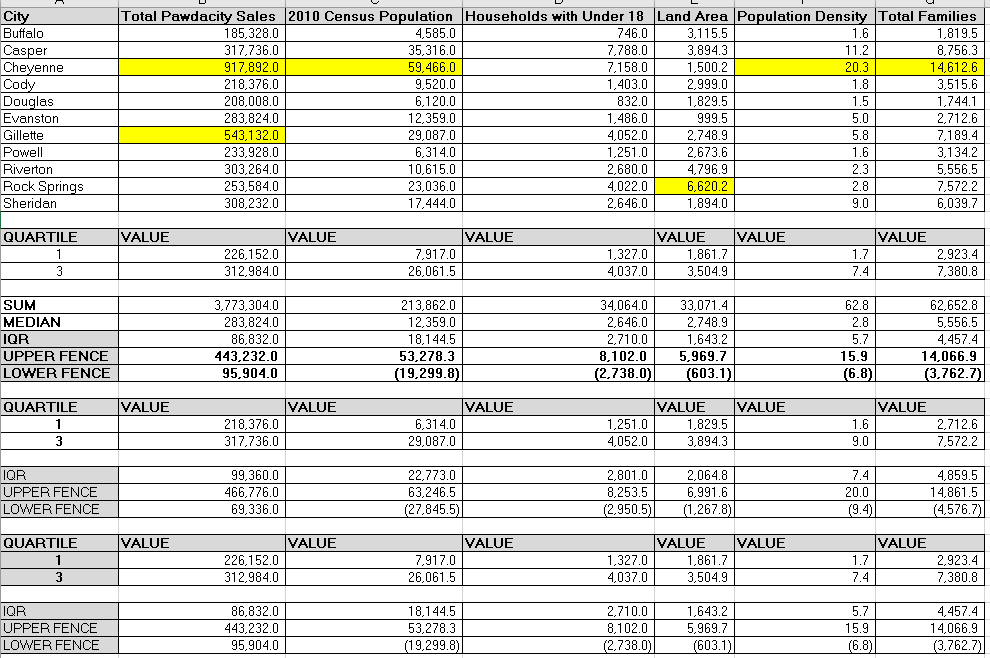


Table . IQR analysis

Also, it was analyzed the impact of removing the outlier record with scatter plot for every predictor variable and target variable, and the slope was not affected considerably.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **With outlier** | **Without Cheyenne outlier** | **Without Gillette outlier** |
| **Census population** |  |  |  |
| ***Households with Under 18*** |  |  |  |
| ***Land Area*** |  |  |  |
| ***Population Density*** |  |  |  |
| ***Total Families*** |  |  |  |

Table . Scatter plots analysis

Alteryx solution

