Project 2.1: Data Cleanup

Úrsula Pérez Ramírez

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?
   Based on predicted yearly sales, we have to recommend the city for Pawdacity's new store that will have more probability of success.

2. What data is needed to inform those decisions?
   First, to build a regression model and select predictor variables, we have to create a training data set containing the following information:
   - total sales data for all of the Pawdacity stores for 2010 (Y variable).
   - population numbers (2010 census)
   - demographic data (households with individuals under 18, land area, population density, and total families) for each city 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

Some steps to comment are the following:

- To replace the "?" in City|County, then parse to divide city and county.
- To obtain the 2010 Census population, the data needs to be parsed into 4 columns with delimiters <> and then with the Data Cleansing tool removing all whitespaces, letters and punctuation.
- The three csv file data are joined with two joins
- With the Summarize tool the sums and averages are calculated for each variable in the training data set

Table 1 contains the sum and average values of all the variables that form the training data set. The workflow is submitted with this pdf.

Sum_2010 Census Population	Sum_Households with Under 18	Sum_Land Area	Sum_Population Density	Sum_Total Families	Sum_Total Pawdacity Sales
213862	34064	33071	63	62653	3773304
Avg_2010 Census Population	**Avg_Households with Under 18**	**Avg_Land Area**	**Avg_Population Density**	**Avg_Total Families**	**Avg_Total Pawdacity Sales**
19442	3096.727	3006.455	5.727273	5695.727	343027.6

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*	*3,006.45*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

Table 1. Sum and average values of all the variables in the training data set.

The data set had no missing values.

Step 3: Dealing with Outliers

Answer these questions

The steps to calculate the outliers are the following:

1) Calculate the first (Q1) and third quartiles (Q3) with Excel QUARTILE.INC function
2) Calculate the interquartile range (IQR) as Q3 minus Q1
3) Calculate the upper threshold for outliers as Q3 + 1.5·IQR
4) Calculate the lower threshold for outliers as Q1 - 1.5·IQR

These values are shown in Table 2.

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.

Table 2. Statistics and thresholds for outliers.

	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales
Q1	7917	1861.721074	1327	1.72	2923.41	226152
Q3	26061.5	3504.9083	4037	7.39	7380.805	312984
IQR	18144.5	1643.187226	2710	5.67	4457.395	86832
Upper	53278.25	5969.689139	8102	15.895	14066.8975	443232
Lower	-19299.75	-603.059765	-2738	-6.785	-3762.6825	95904

In Table 3 it is highlighted in yellow the outliers in each of the variables.

Table 3. Outliers in each variable.

	City	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales
UPPER	Buffalo	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Casper	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Cheyenne	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
	Cody	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Douglas	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Evanston	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Gillette	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
	Powell	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Riverton	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Rock Springs	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
	Sheridan	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

	City	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales
LOWER	Buffalo	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Casper	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Cheyenne	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Cody	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Douglas	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Evanston	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Gillette	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Powell	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Riverton	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Rock Springs	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	Sheridan	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Cheyenne has an outlier in 2010 Census Population, Population Density, Total Families and Total Pawdacity Sales, which are feasible because if there is a high population, more Pawdacity sales make sense. Those outliers fit well the line on the scatterplot. For example, in Figure 1 it is shown the population density considering the 11 cities. If we remove Cheyenne (Figure 2) the slope is changed.

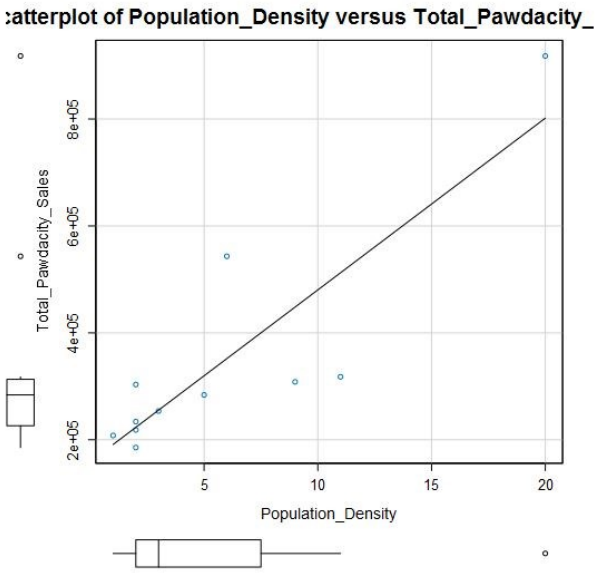

Figure 1. Population density vs total Pawdacity Sales.

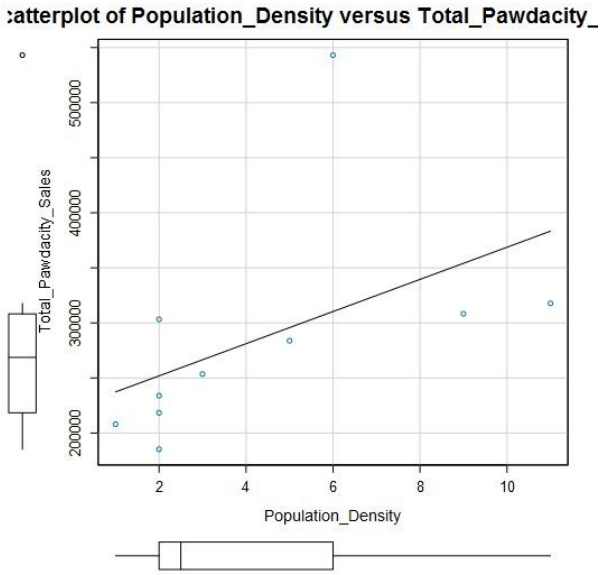

Figure 2. Population density vs total Pawdacity Sales, removing Cheyenne.

Rock Springs is a big city, which I checked the land area and is correctly inputted in the data set.

On the other hand, in Gillette all the variables except for Total Pawdacity Sales are at their average. Total Pawdacity Sales is an outlier which doesn't add up information. In Figure 3 Gillette is removed. The slope in Figure 3 has remained very similar to the slope of the original plot (Figure 1). Therefore the analysis with the outlier of Gillette would stay consistent with what we might have obtained without the outlier.

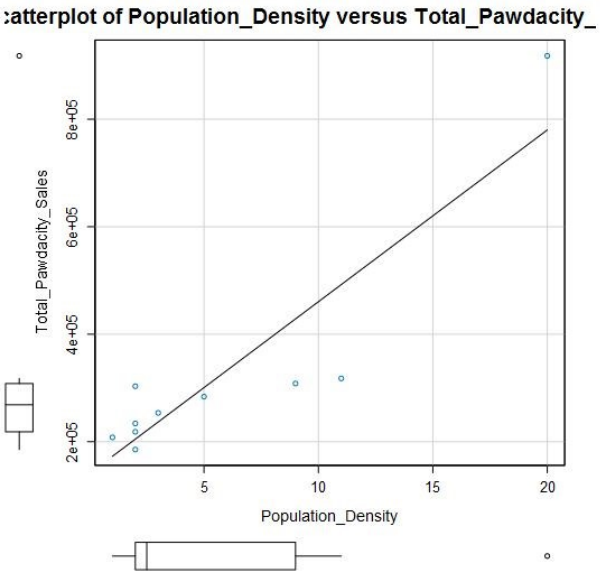

Figure 3. Population density vs total Pawdacity Sales, removing Gillette.

For the aforementioned reasons I consider that Gillette should be removed and the other outliers should be kept.