**Project 2.1: Data Cleanup**

**BY AYAME GOD’SWILL CLAUDE**

## 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?

**Answers:**

**There is the need to decide which city will be chosen as the location for Pawdacity’s newest (14th) store. The 14th store should be located in the city whose predicted yearly sales is found to be the highest.**

**To make this prediction, we need a historical (training) dataset for each of the 13 stores to build a linear regression model, which will then be used to determine which city has the highest sales and thus the potential of raking in the highest profit (predicted sales).**

**The following data, as given, are needed to inform those decisions:**

1. **The monthly sales data for all of the Pawdacity stores for the year 2010.**
2. **NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.**
3. **A partially parsed data file that can be used for population numbers.**
4. **Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming, taking into consideration that in the US city system, a state contains counties and counties contains one or more cities.**

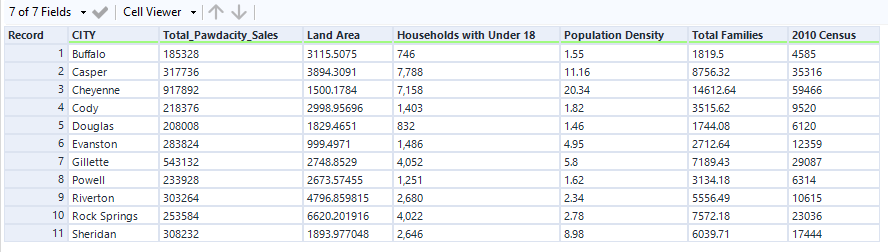
## Step 2: Building the Training Set

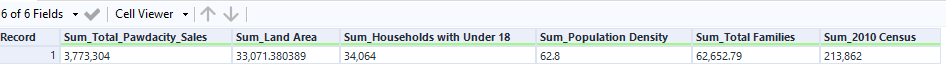
*Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.*

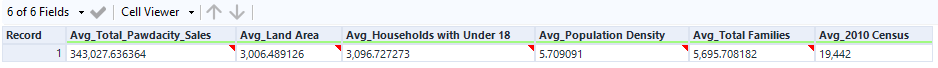
*In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24*

**Answers:**

**After due cleaning, the field summary report below shows that there are no duplicates or missing data.**

****

****

****

## 

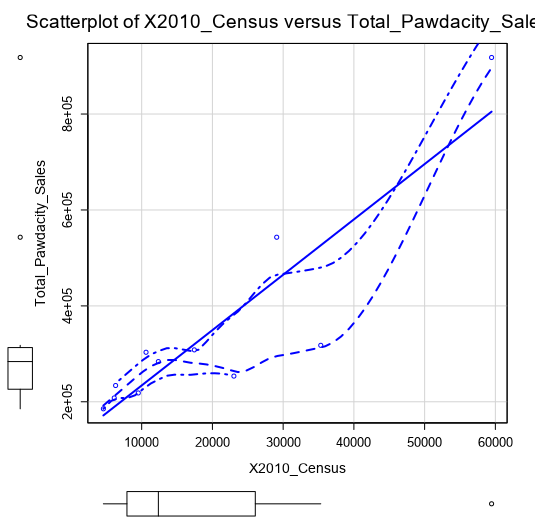
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19,442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.64* |
| *Households with Under 18* | *34,064* | *3,096.73* |
| *Land Area* | *33,071* | *3,006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653* | *5,695.71* |

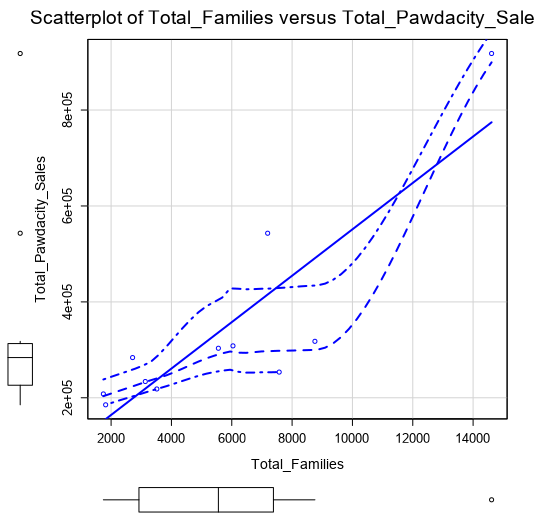
## Step 3: Dealing with Outliers

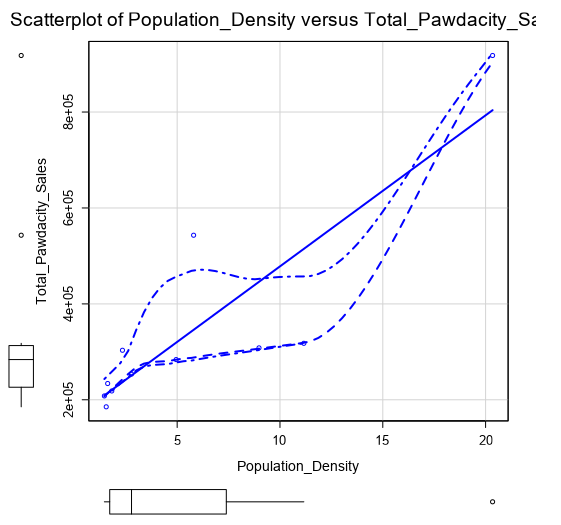
*Answer these questions*

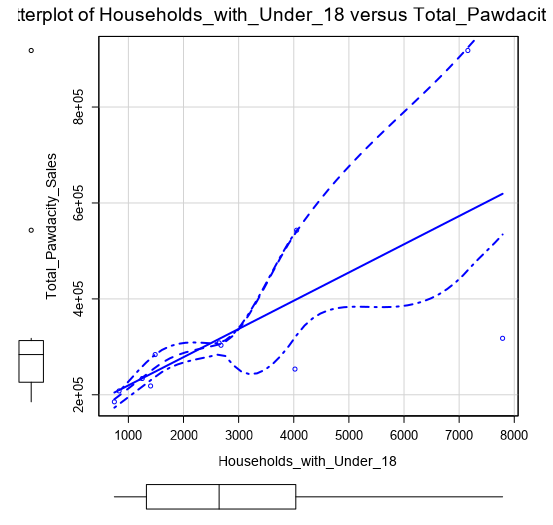
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

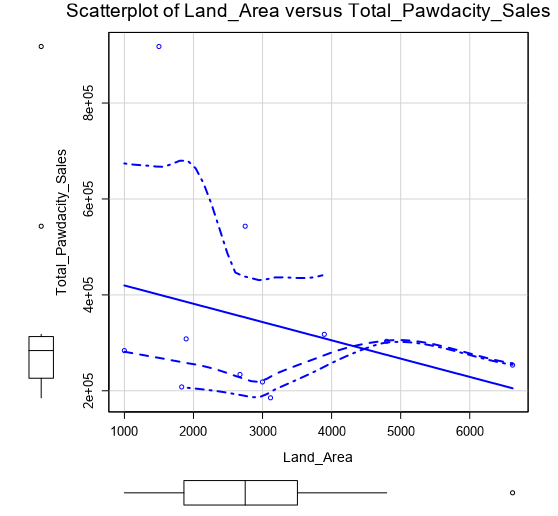
**Answers:**

****

****

****

****

****

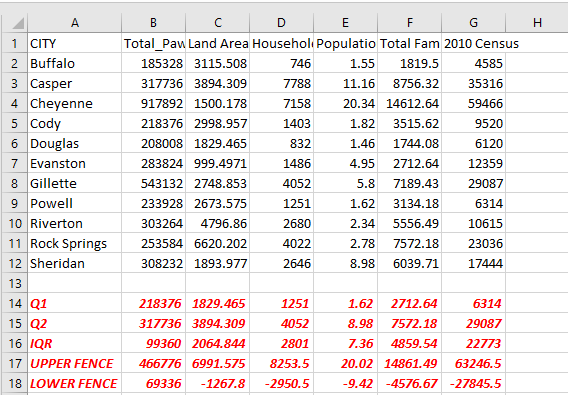
**The scatterplot data visualizations attached above reveal that there are three outliers.**

**A preliminary eye check of the dataset may suggest that Buffalo, Cheyenne and Gillette appear to be outliers in TOTAL PAWDACITY SALES*;* Evanston, Riverton and Rock Springs in LAND AREA; Buffalo and Douglas in HOUSEHOLDS WITH UNDER 18; Casper, Cheyenne and Sheridan in POPULATION DENSITY; Buffalo, Cheyenne and Douglas in TOTAL FAMILIES; and Casper, Cheyenne and Gillette in 2010 CENSUS.**

**However, to really confirm which cities are outliers, the IQR method (box-and-whisker plot) was used. The following calculations were done for each of the fields (columns):**

1. **Calculation of the 1st quartile Q1 and 3rd quartile Q3.**
2. **Calculation of the Interquartile Range: IQR = Q3 - Q1.**
3. **Next, the product of 1.5 & IQR was added to Q3 to get the upper fence**
4. **And finally, the product of 1.5 & IQR was subtracted from Q1 to get the lower fence.**

**The results of the above calculations are shown in the table below.**



**In each of the fields above, any value that is higher than the upper fence or lower than the lower fence is an outlier. In this vein, only Cheyenne records outliers in the fields TOTAL PAWDACITY SALES and POPULATION DENSITY. Also, Gillette has an outlier under the TOTAL PAWDACITY SALES field.**

### **Ordinarily, a city with outliers would be removed if its TOTAL PAWDACITY SALES value were particularly high with no corresponding high value in any other relevant field. A city may also be removed if it consistently records outliers in multiple fields, such that it is unlike other fields in the dataset for most fields, with the possibility of disrupting the model’s accurate predicting ability if added.**

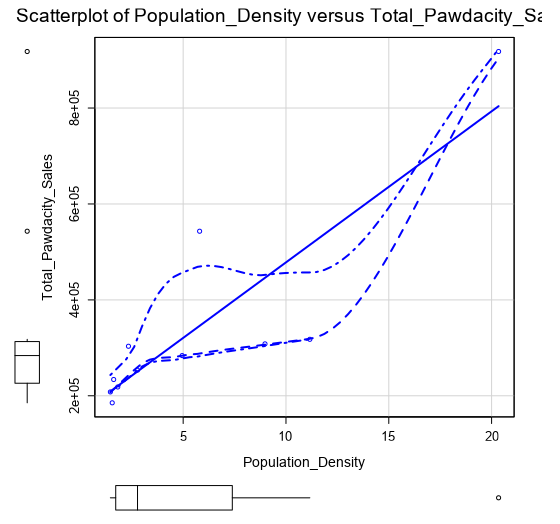
**On the other hand, an outlier can be retained in the dataset if it is in line with the linear relationship, or If the dataset is small and the city is an outlier in only one field.**

**In the case of Cheyenne, its outliers occur only in two fields, which are actually related, considering the fact that one would expect a high sales record from a city which is densely populated.**

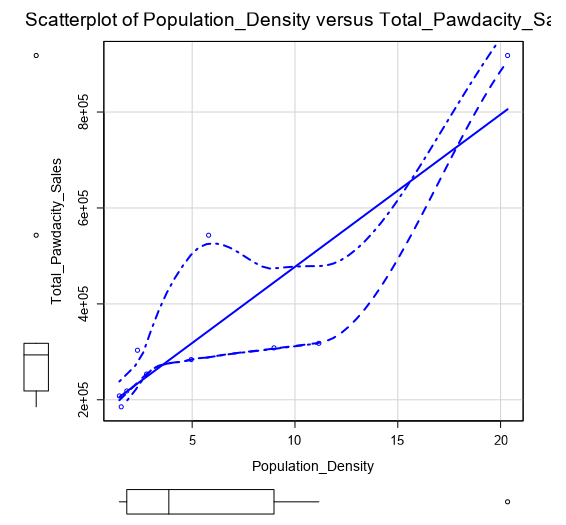
**But for Gillette, it has an outlier in the target variable (TOTAL PAWDACITY SALES) with no corresponding outlier in another relevant field like POPULATION DENSITY.**

**More so, scatterplots were used to check the relationship between the TOTAL PAWDACITY SALES and POPULATION DENSITY (for illustrative purposes only) with and without the outlier records (rows), and the followings were discovered:**

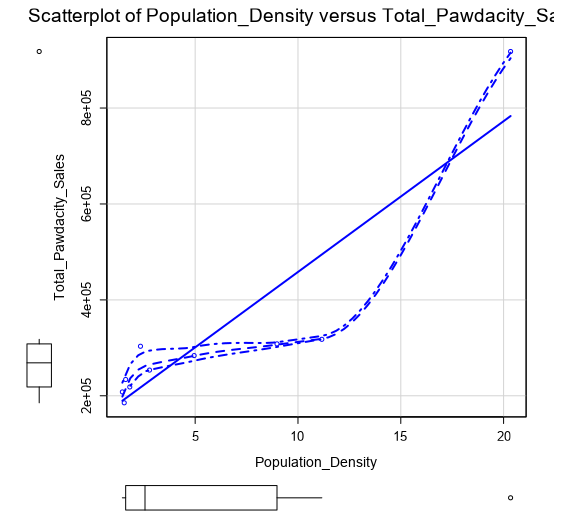
* + - 1. **With the outliers all included, the scatterplot shows a good linear relationship between both variables, as shown below**

****

* + - 1. **With the Cheyenne record removed, the scatterplot still maintains a good linear relationship between both variables with similar pattern and almost same slope as before.**

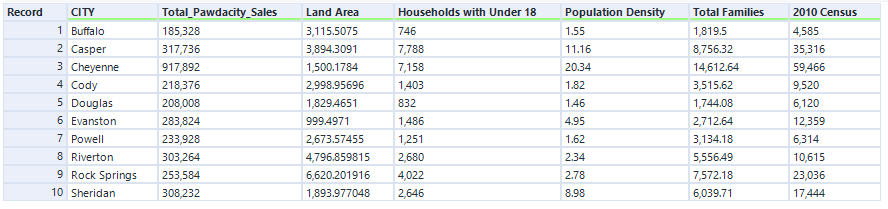
****

* + - 1. **With the Gillette record (row) removed, whilst the slope is quite similar, the scatterplot shows a pattern quite different from that of the original scatterplot containing all records.**

****

**Thus, from the foregoing the Gillette record will be removed, but the Cheyenne record will be retained with the rest of the dataset.**

**The final dataset and workflow solution are attached below.**

****

