Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

The primary decision to be made is to recommend a city for the newest store. The decision should be based upon the yearly sales. To make the recommendation we need to predict the yearly sales and based on the prediction (total sales) we can find which is the best location to open the new store.

2. What data is needed to inform those decisions?

The data that is need to make this recommendation is as follows. Firstly we need to find the total sales of the location. Next we need to analyze the census and the population density of the city. Even if we see that a particular city has very high sales that does not mean that the city is good location to open new store. Maybe the city is small and the single store in the city can handle all the customers and hence opening another store in that city will not help the company. Hence we also need the total area of the city to analyze the people per store ratio.

Step 2: Building the Training Set

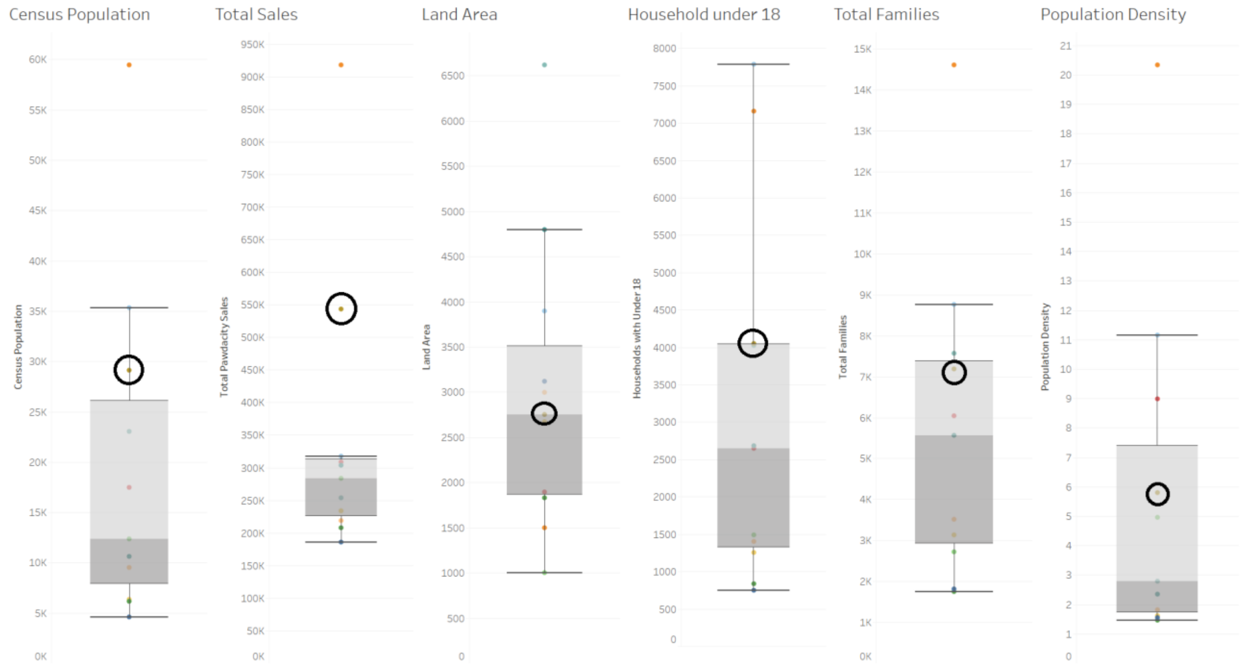
Column	**Sum**	**Average**
Census Population	*213,862*	19442
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.45
Population Density	*63*	5.73
Total Families	*62,653*	5695.73

Step 3: Dealing with Outliers

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.

At the first glance of the Box and Whisker Plot one might assume that Cheyenne is an outlier as it appears above the Upper Whisker in 4(Census, Total Sales, Total Families, Population Density) of the 6 variables. But on further analysis we see that Cheyenne has high Population

Density, Total Number of Families, Census Population and this may all lead to the High Sales in the City. Hence Cheyenne is not one of the outlier. On a deeper analysis of the plot we see that **Gillette** appears above the Upper Whisker in Total Sales, however all other values for Gillette are not high (which can be seen in the image below). Hence there is no justification for Gillette to have high sales and thus it can be considered as an outlier in the given dataset.
Imputation does not work in the current scenario as the total sales vary by a large margin and cannot be just imputed with average or any other number. Imputation may lead to unexpected results and thus should be avoided. Hence in the current scenario it is best to **remove** the city of Gillette.


Census Population
Total Sales
Land Area
Household under 18
Total Families
Population Density
Census Population
Total Pawdacity Sales
Land Area
Households with Under 18
Total Families
Population Density
60K
55K
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
950K
900K
850K
800K
750K
700K
650K
600K
550K
500K
450K
400K
350K
300K
250K
200K
150K
100K
50K
0K
15K
14K
13K
12K
11K
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K