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)

Our Client is the leading store in Wyoming with thirteen stores throughout the state, this year they want to expand and open store number fourteen. The objective is to do an analysis to give a recommendation of in which city the new store should open, according to the expected annual sales

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We are provided with three data sets (p2-2010-pawdacity-month-sales.csv, p2-partially-parsed-wy-web-scrape.csv and p2-wy-453910-naics-data.csv. We need to clean them, and know what data will be necessary to predict in which city the next branch would be

2. What data is needed to inform those decisions?
Using the datasets, we must combine them to create a dataset with the following fields

City
2010 Census Population
Total Pawdacity Sales
Households with under 18
Land Area
Population Density
Total Families

This data will be used later to create our prediction model for the location of the new store.

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*	19442
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.49
Population Density	*63*	5.71
Total Families	*62,653*	5695.71

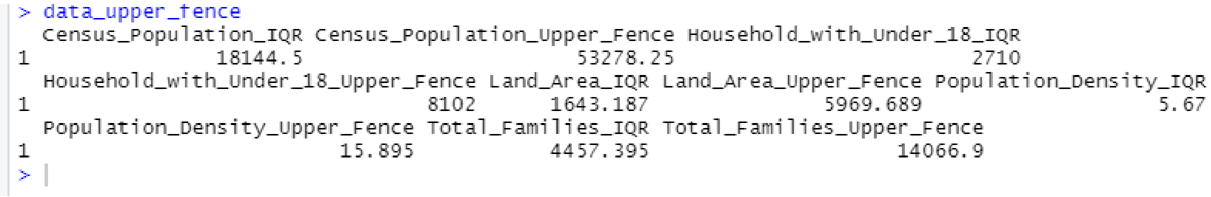
Step 3: Dealing with Outliers

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.

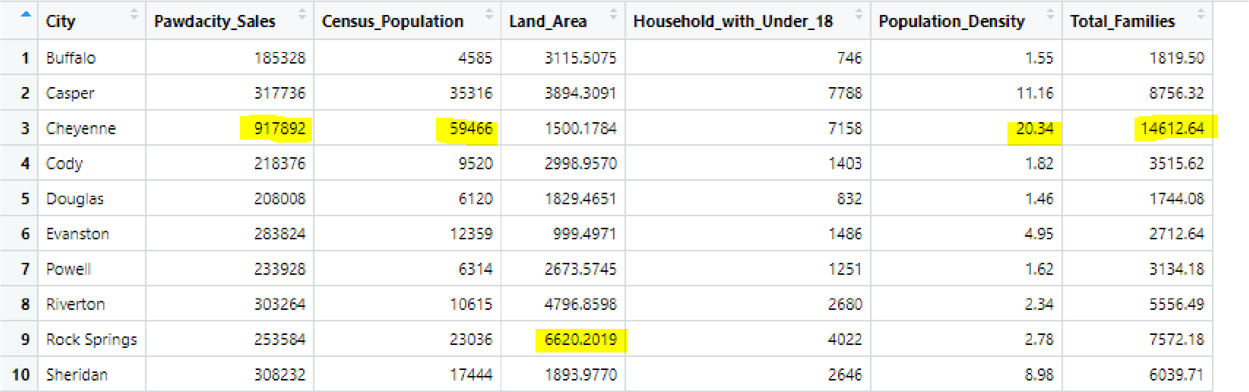
Let's observe the data summary, with an analysis of the interquantile ranges for the variables and their subsequent upper valley, the fence of this project will be [1.5 * interquantile range] + third quantile.

Name	Min	Max	Median	Mean	Std Dev.
Census_Population	4585.00	59466.00	12359.00	19442.00	16616.02
Household_with_Under_18	746.00	7788.00	2646.00	3096.73	2453.00
Land_Area	999.50	6620.20	2748.85	3006.49	1617.46
Pawdacity_Sales	185328.00	917892.00	283824.00	343027.64	213538.71
Population_Density	1.46	20.34	2.78	5.71	5.85
Total_Families	1744.08	14612.64	5556.49	5695.71	3816.05

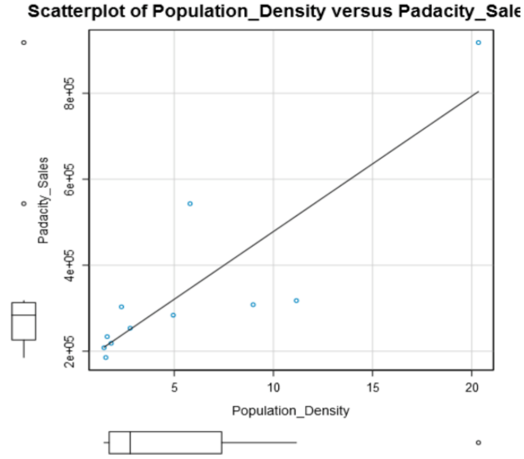


```
> data_upper_fence
  Census_Population_IQR Census_Population_Upper_Fence Household_with_Under_18_IQR
1               18144.5                      53278.25                        2710
  Household_with_Under_18_Upper_Fence Land_Area_IQR Land_Area_Upper_Fence Population_Density_IQR
1                                8102      1643.187              5969.689                   5.67
  Population_Density_Upper_Fence Total_Families_IQR Total_Families_Upper_Fence
1                         15.895           4457.395                    14066.9
> 
```

Here we realize that cheyenne, gillette and rock springs are outliers, and since we only have 11 cities, we eliminate Gillette



	City	Pawdacity_Sales	Census_Population	Land_Area	Household_with_Under_18	Population_Density	Total_Families
1	Buffalo	185328	4585	3115.5075	746	1.55	1819.50
2	Casper	317736	35316	3894.3091	7788	11.16	8756.32
3	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
4	Cody	218376	9520	2998.9570	1403	1.82	3515.62
5	Douglas	208008	6120	1829.4651	832	1.46	1744.08
6	Evanston	283824	12359	999.4971	1486	4.95	2712.64
7	Powell	233928	6314	2673.5745	1251	1.62	3134.18
8	Riverton	303264	10615	4796.8598	2680	2.34	5556.49
9	Rock Springs	253584	23036	6620.2019	4022	2.78	7572.18
10	Sheridan	308232	17444	1893.9770	2646	8.98	6039.71



However, you need to justify the decision to remove or maintain each one. Why did you decide to keep Cheyenne and Rock Springs and remove Gillette from the dataset?

I had a hard time deciding between Cheyenne and Gillette. Cheyenne on the one hand, has two branches and the data of both branches are joined, which could generate outliers, but since we are trying to place one more branch, we must look at the data of Cheyenne as a city, this makes Cheyenne the city that generates higher sales to give a justification to both stores. Gillette also has two branches, but looking from other perspectives, Gillette's data is seen in the range of outliers, making it difficult to explain and it would be better to remove this city from our dataset, however reluctantly removing since we only have 11 cities, being a small amount of data

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.