Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity a leading pet store chain in Wyoming with 13 stores throughout the state is set to expand and open their 14th store. They recommendation on which city their new store should be based. An analysis to recommend the city for Pawdacity's newest store will based on predicted yearly sales.

Key Decisions:

Pawdacity is set to make a decision on which city in Wyoming their 14th store will be located; this decision will be based on the recommendation gotten from yearly sales predictions.
Based on the predicted yearly sales to be gotten from the analysis, training, and modelling of the datasets, the city to be selected can be recommended.

To make the key decisions the dataset needed are the records of the monthly sales data for all of the Pawdacity stores for the year 2010.The NAICS data on the most current sales of all competitor stores will help in the yearly sales model training for yearly sales prediction. A partially parsed data file of population numbers of the cities in Wyoming and the demographic data for each city and county in the state of Wyoming.
These data sets will be cleansed, blended, trained and modelled to make the important decision of selecting the city where the 14th store of Pawdacity will be located.

Step 2: Building the Training Set

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	3,005.91
Population Density	63	5.73
Total Families	62,653	5,695.73

Step 3: Dealing with Outliers

The output from the data cleaning was a table with 7 columns and 11 rows. The columns consisted of City, Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, and Total Families. The 11 rows contained some cities from Wyoming State. See Table 1 below:

Record	CITY	Census_Population	Total_Pawdacity_Sales	Households with Under 18	Land Area	Population Density	Total Families
1	Buffalo	4,585	185,328	746	3,115	2	1,820
2	Casper	35,316	317,736	7,788	3,894	11	8,756
3	Cheyenne	59,466	917,892	7,158	1,500	20	14,613
4	Cody	9,520	218,376	1,403	2,998	2	3,516
5	Douglas	6,120	208,008	832	1,829	1	1,744
6	Evanston	12,359	283,824	1,486	999	5	2,713
7	Gillette	29,087	543,132	4,052	2,748	6	7,189
8	Powell	6,314	233,928	1,251	2,673	2	3,134
9	Riverton	10,615	303,264	2,680	4,796	2	5,556
10	Rock Springs	23,036	253,584	4,022	6,620	3	7,572
11	Sheridan	17,444	308,232	2,646	1,893	9	6,040

From Table 1 above, three cities were found to be outliers after calculating the Upper fence and Lower fence of the table. The cities are **Cheyenne, Rock Springs, and Gillette**.

Upper fence results: *'Census Population'* **53,278.25**, '*Total Pawdacity Sales'* ***443,232,*** *'Land Area'* **5,969.75,** *'Population Density'* ***15.75,*** and *'Total Families'* **14,066**.

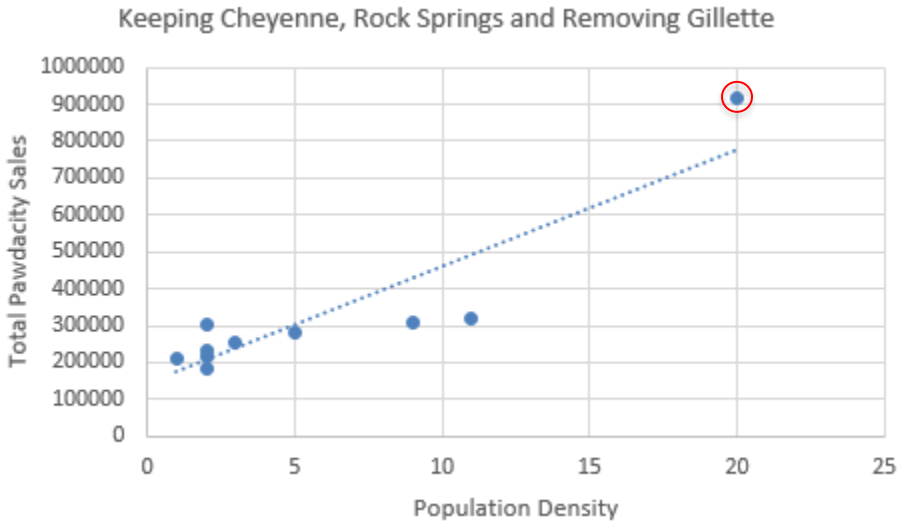
Cheyenne had a result greater than the Upper fence of the whole table for **Census Population (59,466**), **Total Pawdacity Sales (917,892)**, **Population Density (20)** and the **Total Families (14,613)**.

Rock Springs **Land Area (6620)** was greater than the Upper fence Land area of the whole table making it stand out as an outlier while Gillette's **Total Pawdacity Sales (543,132)** was greater than the table's Upper fence. Table 2 below shows these:

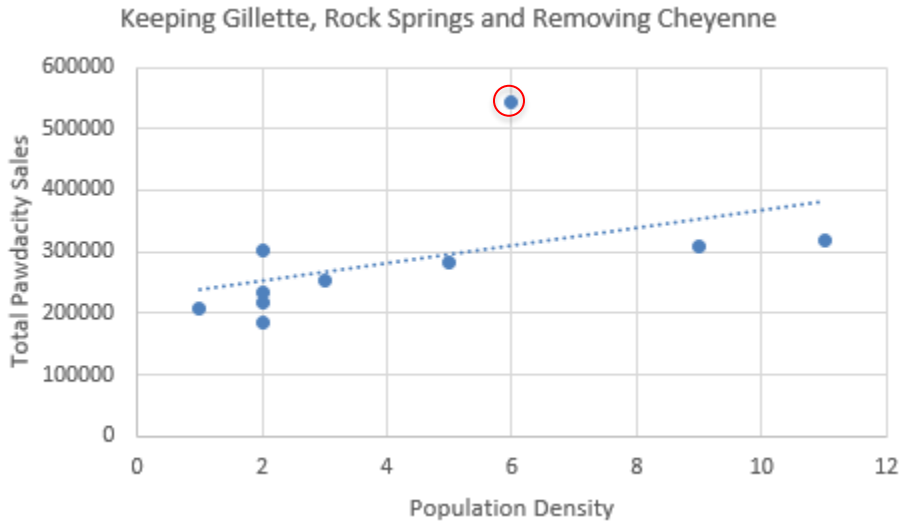
CITY	Census_Population	Total_Pawdacity_Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4585	185328	746	3115	2	1820
Casper	35316	317736	7788	3894	11	8756
Cheyenne	59466	917892	7158	1500	20	14613
Cody	9520	218376	1403	2998	2	3516
Douglas	6120	208008	832	1829	1	1744
Evanston	12359	283824	1486	999	5	2713
Gillette	29087	543132	4052	2748	6	7189
Powell	6314	233928	1251	2673	2	3134
Riverton	10615	303264	2680	4796	2	5556
Rock Springs	23036	253584	4022	6620	3	7572
Sheridan	17444	308232	2646	1893	9	6040
First Quartile Q1	7917	226152	1327	1861	2	2923.5
Third Quartile Q3	26061.5	312984	4037	3504.5	7.5	7380.5
Inter Quartile IQR(Q3-Q1)	18144.5	86832	2710	1643.5	5.5	4457
Lower Fence (Q1 - 1.5 IQR)	-19299.75	95904	-2738	-604.25	-6.25	-3762
Upper Fence (Q3 + 1.5 IQR)	53278.25	443232	8102	5969.75	15.75	14066

Among the outlier cities, I chose to remove Gillette and keep Cheyenne and Rock springs for the following reasons:

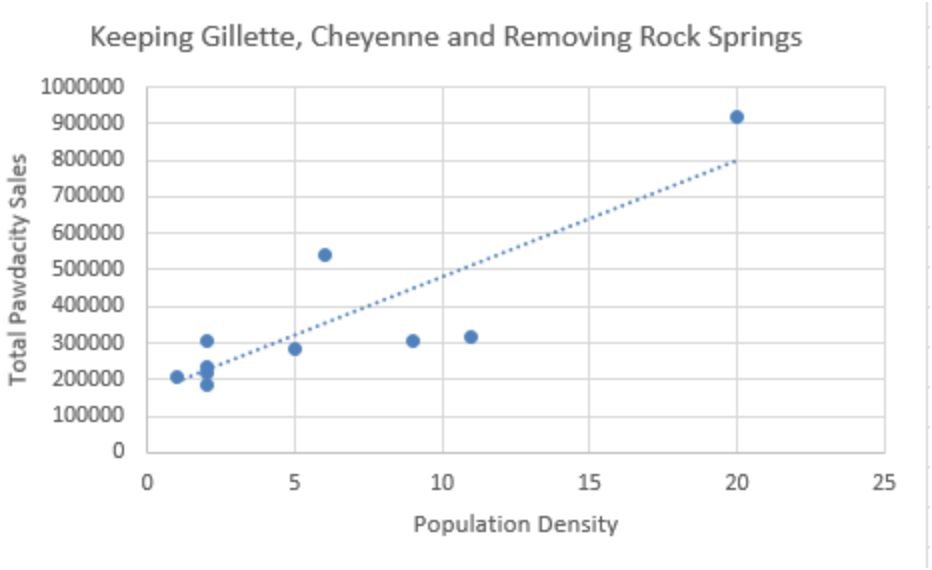
- Cheyenne showed to be in line with the linear relationship of population density and Total Pawdacity Sales plot. The city Total Pawdacity Sales record might be related to its higher population though there is not sufficient information to prove this correlation yet.
  Cheyenne will help in keeping the model to be built more robust especially in the predicting future cities with big numbers. Please find below the scatterplot fig 1, the circled point is the position of Cheyenne in the plot after removing Gillette:



- Removing Gillette because it does seem to be in line with the linear relationship of population density and total pawdacity sales scatterplot. It seems to be an odd one out even though it's only an outlier in one field. It mainly skews high in sales, yet its not relative to the other data fields in the training set. The scatterplot fig 2 below shows the scatterplot line and position of Gillette when it was kept and Cheyenne removed from the dataset:



- Keeping Rocksprings because it was only an outlier in the land area but kept a linear relationship with every data field in the training set. Rock Springs doesn't have much effect on the distribution. The scatterplot fig 3 below shows these:



In conclusion the outlier city, I removed is Gillette. Below is the Table 3. after removing Gillette city:

TABLE 3. OF OUTPUT DATA AFTER REMOVING GILLETTE CITY

CITY	Census_Population	Total_Pawdacity_Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4585	185328	746	3115	2	1820
Casper	35316	317736	7788	3894	11	8756
Cheyenne	59466	917892	7158	1500	20	14613
Cody	9520	218376	1403	2998	2	3516
Douglas	6120	208008	832	1829	1	1744
Evanston	12359	283824	1486	999	5	2713
Powell	6314	233928	1251	2673	2	3134
Riverton	10615	303264	2680	4796	2	5556
Rock Springs	23036	253584	4022	6620	3	7572
Sheridan	17444	308232	2646	1893	9	6040