Create an Analytical Dataset

ansjayan@msn.com

Step 1: Business and Data Understanding

Pawdacity, the leading pet store chain in Wyoming is planning to open her 14th store in the state. Pawdacity wants to know, which city is best for the new store, which can be found by predicting yearly sales.

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   Find the best city based on predicted yearly sales.

2. What data is needed to inform those decisions?

   Data required: - Details of each city are required, which includes population, population density, land area, total families, homes with under 18 members, and pawdacity sales in each city.

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.

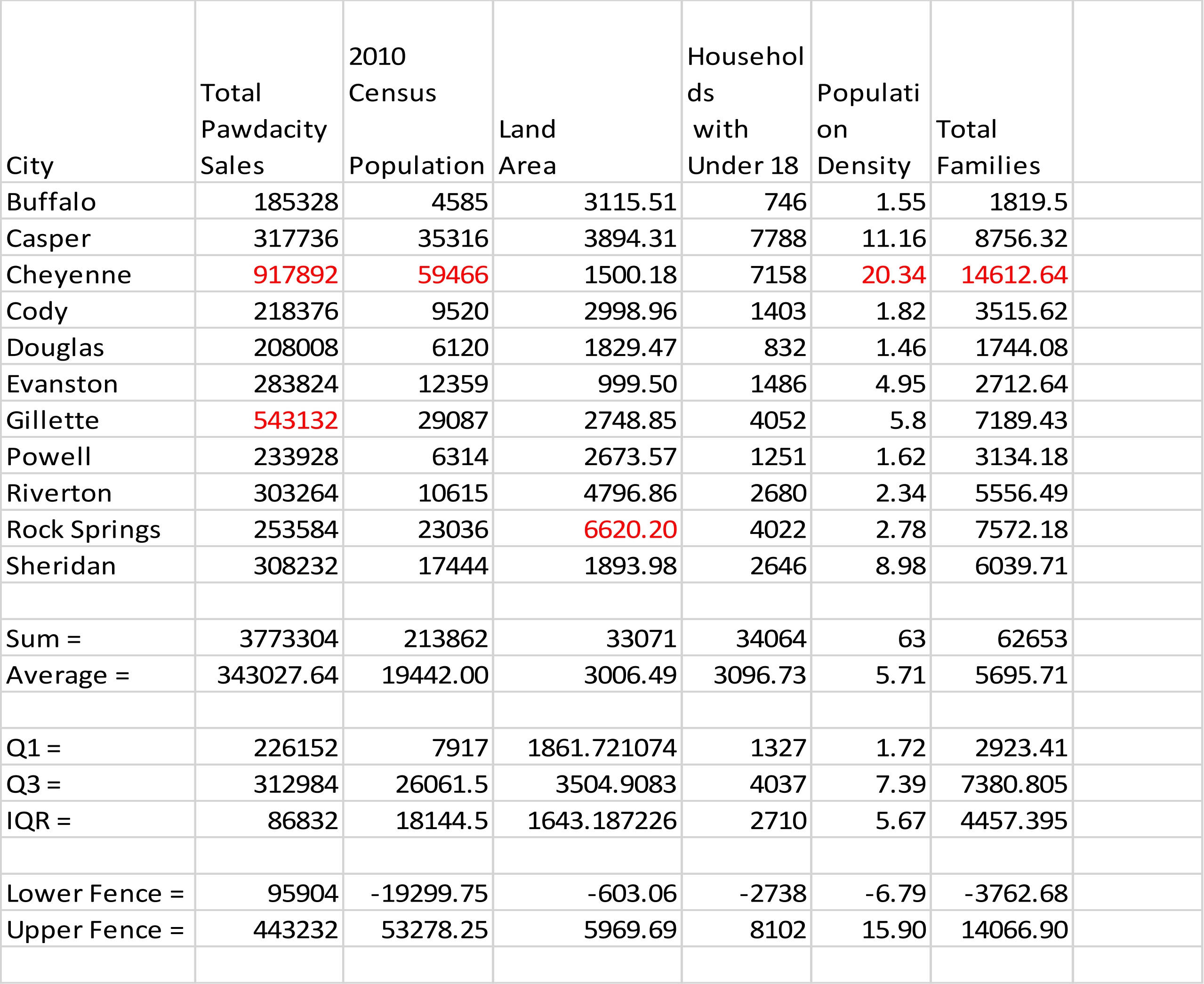
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

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

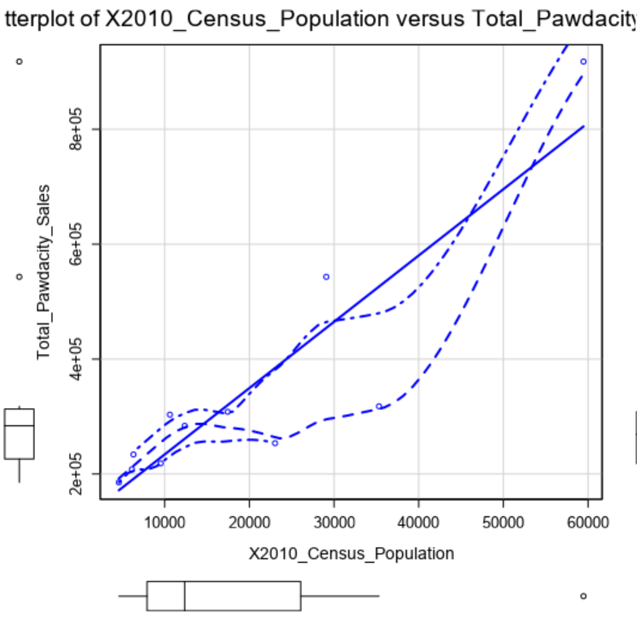
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

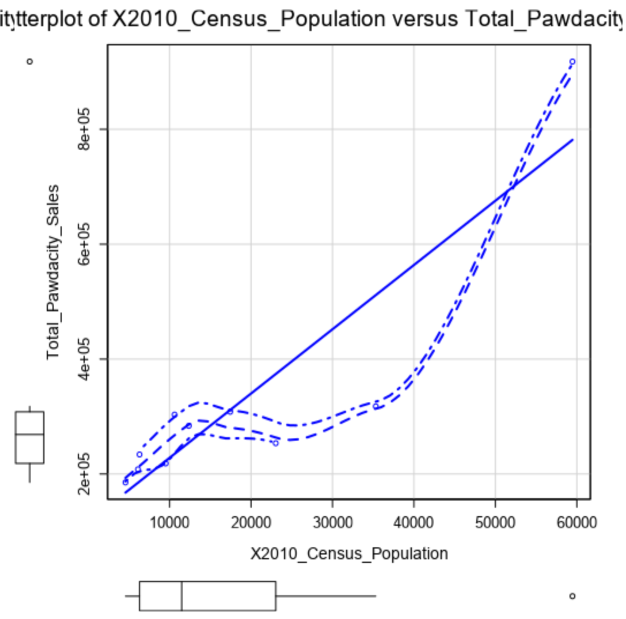


City	Total Pawdacity Sales	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families	
Buffalo	185328	4585	3115.51	746	1.55	1819.5	
Casper	317736	35316	3894.31	7788	11.16	8756.32	
Cheyenne	917892	59466	1500.18	7158	20.34	14612.64	
Cody	218376	9520	2998.96	1403	1.82	3515.62	
Douglas	208008	6120	1829.47	832	1.46	1744.08	
Evanston	283824	12359	999.50	1486	4.95	2712.64	
Gillette	543132	29087	2748.85	4052	5.8	7189.43	
Powell	233928	6314	2673.57	1251	1.62	3134.18	
Riverton	303264	10615	4796.86	2680	2.34	5556.49	
Rock Springs	253584	23036	6620.20	4022	2.78	7572.18	
Sheridan	308232	17444	1893.98	2646	8.98	6039.71	
Sum =	3773304	213862	33071	34064	63	62653	
Average =	343027.64	19442.00	3006.49	3096.73	5.71	5695.71	
Q1 =	226152	7917	1861.721074	1327	1.72	2923.41	
Q3 =	312984	26061.5	3504.9083	4037	7.39	7380.805	
IQR =	86832	18144.5	1643.187226	2710	5.67	4457.395	
Lower Fence =	95904	-19299.75	-603.06	-2738	-6.79	-3762.68	
Upper Fence =	443232	53278.25	5969.69	8102	15.90	14066.90	

There are three outliers: - Gillette, Cheyenne and Rock Springs.

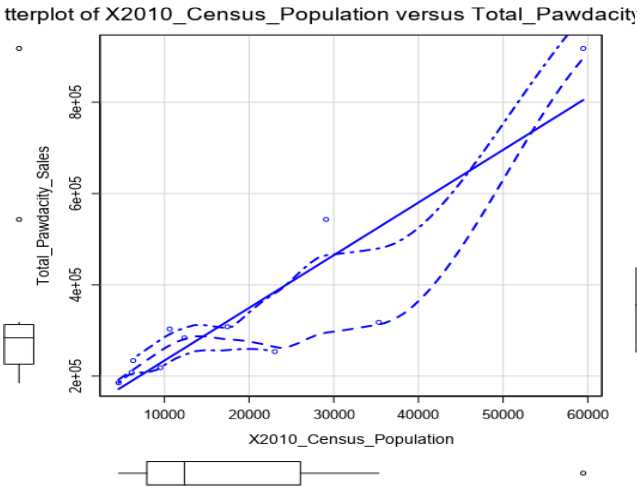
Outlier 1: - **Gillette**

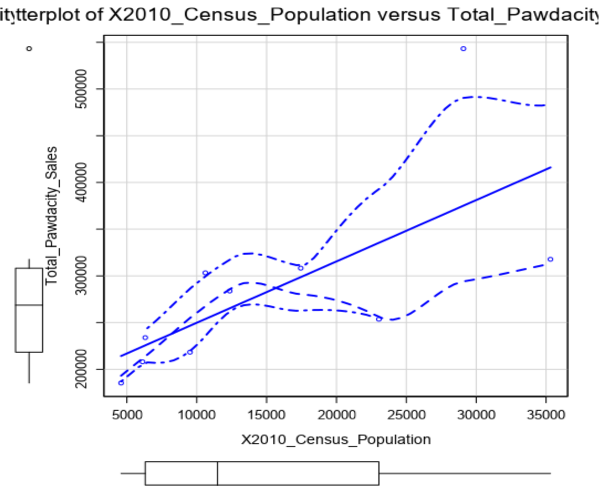


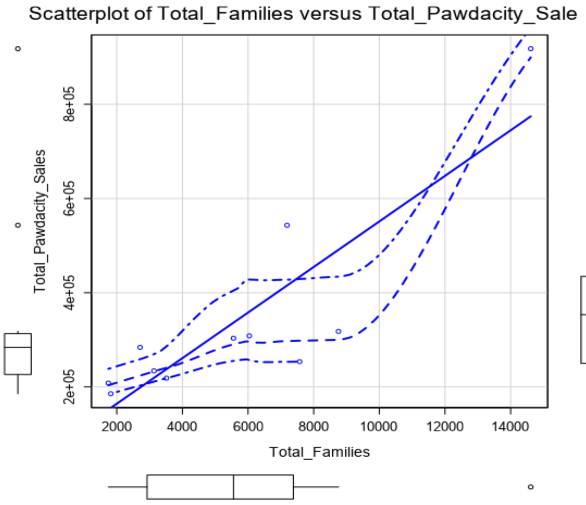


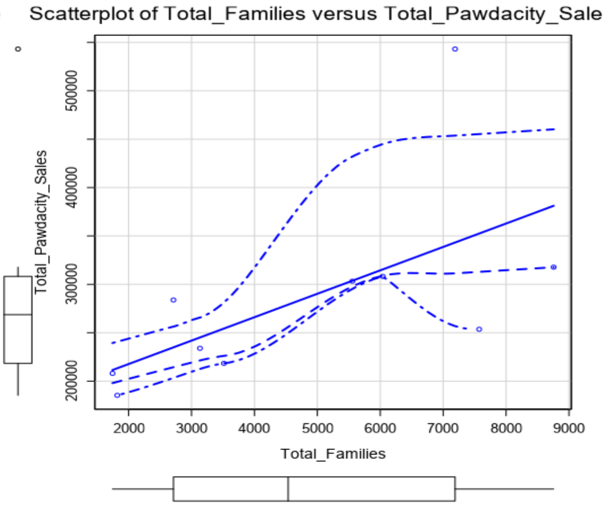
The left side graph is with Gillette and right side is without. Even though the slope remains same the plot shows that Gillette is above the regression line. So, the outlier Gillette can be removed.

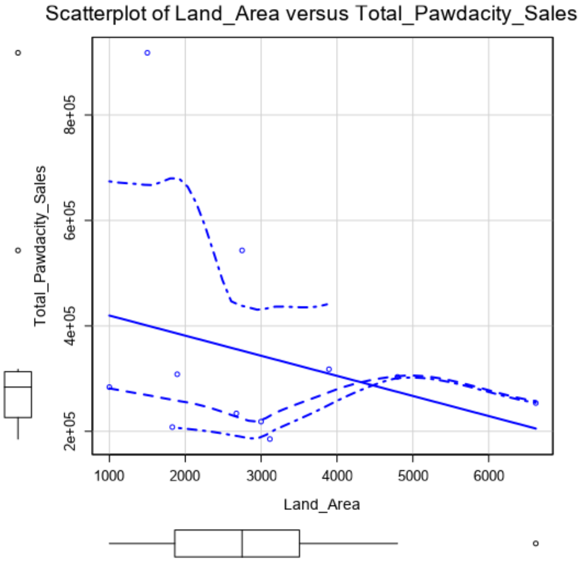
Outlier 2: - **Cheyenne**

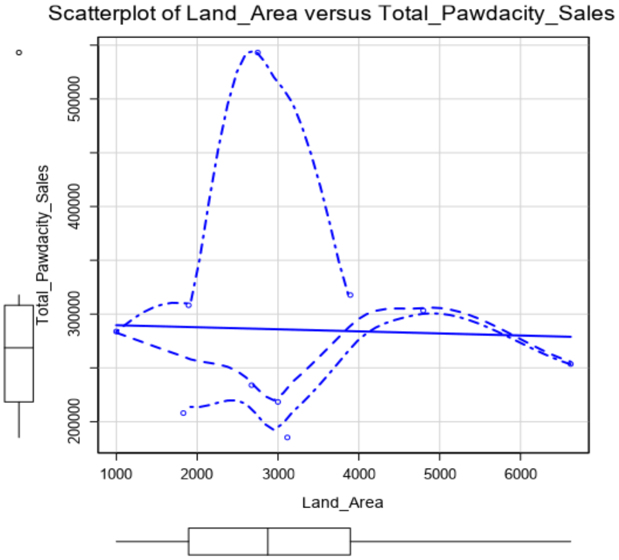





Scatterplot of Total_Families versus Total_Pawdacity_Sale

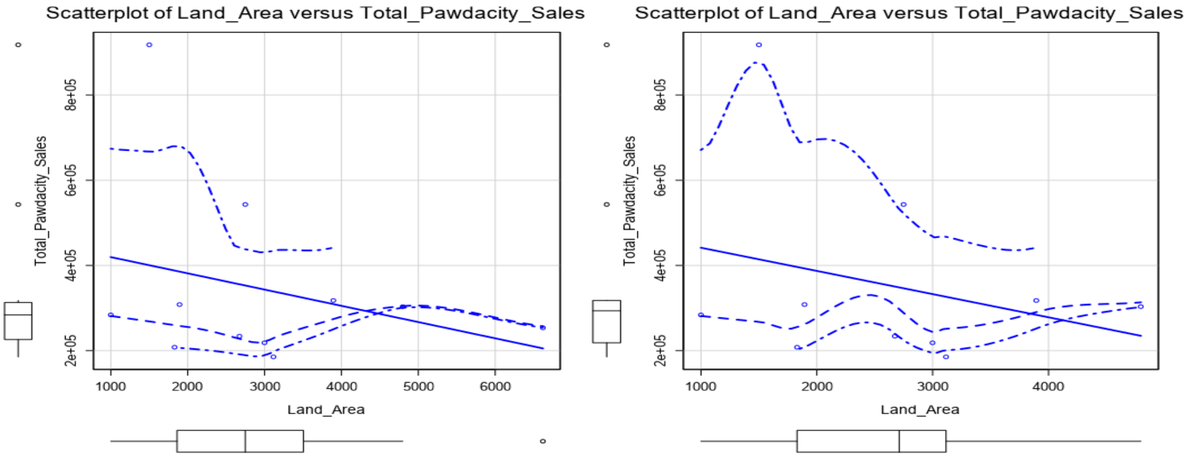

Scatterplot of Total_Families versus Total_Pawdacity_Sale


Scatterplot of Land_Area versus Total_Pawdacity_Sales

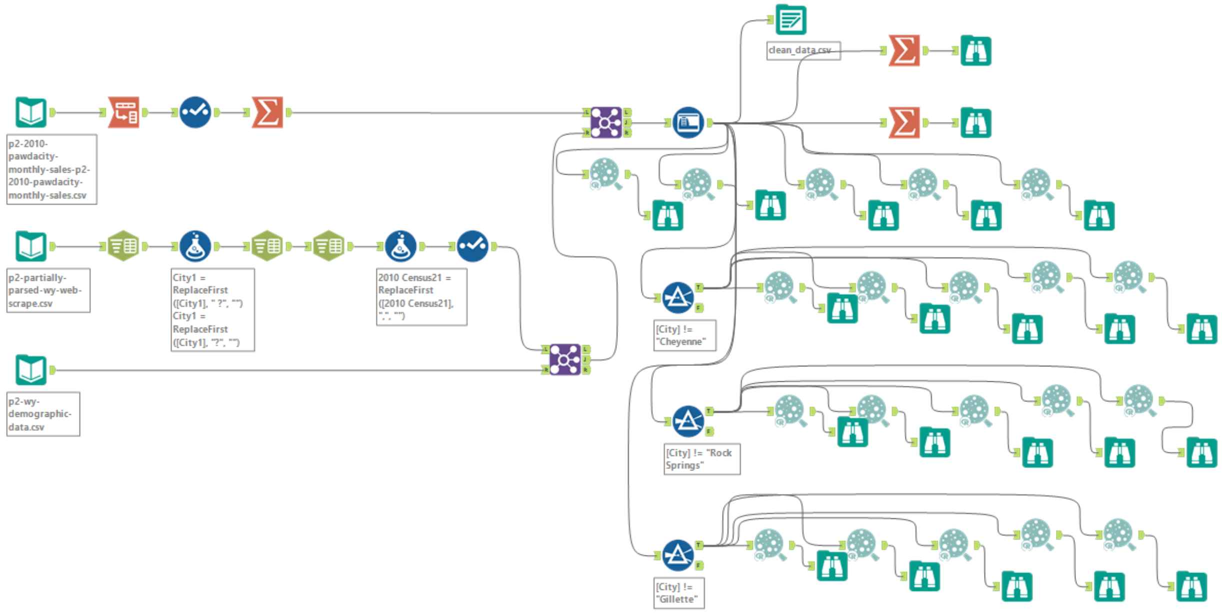

Scatterplot of Land_Area versus Total_Pawdacity_Sales

The left side graph is with Cheyenne and right side is without. Cheyenne, a big city, is an outlier to Sales, Population, Families, and density. The slope is changed when Cheyenne is removed. If we model any future big cities, keeping Cheyenne will be good. As we got only 11 rows in dataset, keep Cheyenne.

Outlier 3: - **Rock Springs**



The left side graph is with Rock Springs and right side is without. Rock Springs skew high in land area, but not with other variables. Rock Springs is an outlier but the slope stayed consistent with the plot without this outlier. So, models build with this outlier will be consistent like models without it and we have only 11 rows we can keep it.


clean_data.csv
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
City1 = ReplaceFirst ([City1], " ?", "") City1 = ReplaceFirst ([City1], "?", "")
2010 Census21 = ReplaceFirst ([2010 Census21], ",", "")
p2-wy-demographic-data.csv
[City] != "Cheyenne"
[City] != "Rock Springs"
[City] != "Gillette"

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.