Project: Data Clean Up

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

We consider a company, Pawdacity, that a leading pet store chain in Wyoming with 13 stores throughout the state. The decision needs to be made to help the company identifies an opportunity to expand their own trade area within Wyoming state. Specifically, whether or not it should building a 14th pet store. It is market demand analysis in which we need to determine the total purchase value within a specified demographic at a particular point in time. The question which is the city to recommend for the newest store, based on predicted yearly sales. To make an informed decision about growing and managing the business, it necessary gathering and analyzing data of customers, competitors and economic of the defined trade area. We create a dataset to properly build the model and select predictor variables.

2. What data is needed to inform those decisions?

The basis to gather information that can be used to support expansion initiatives, it is to defined boundaries of trade areas that take into account. City-system of the United States is a possible method. The boundaries of trade area allow extracted the following information in all trade area analyses:
- The number of potential customers;
- Consumer spending data;
- Competitors data;
- Demographic characteristics, for example, household type, income, age, and ethnicity that relates directly to the consumer preferences;
- Economic data regarding development and wealth for example per capita personal income.

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*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

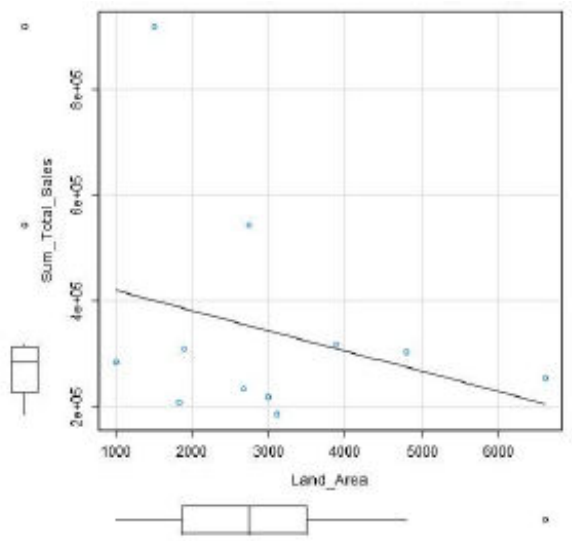
Step 3: Dealing with Outliers

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.

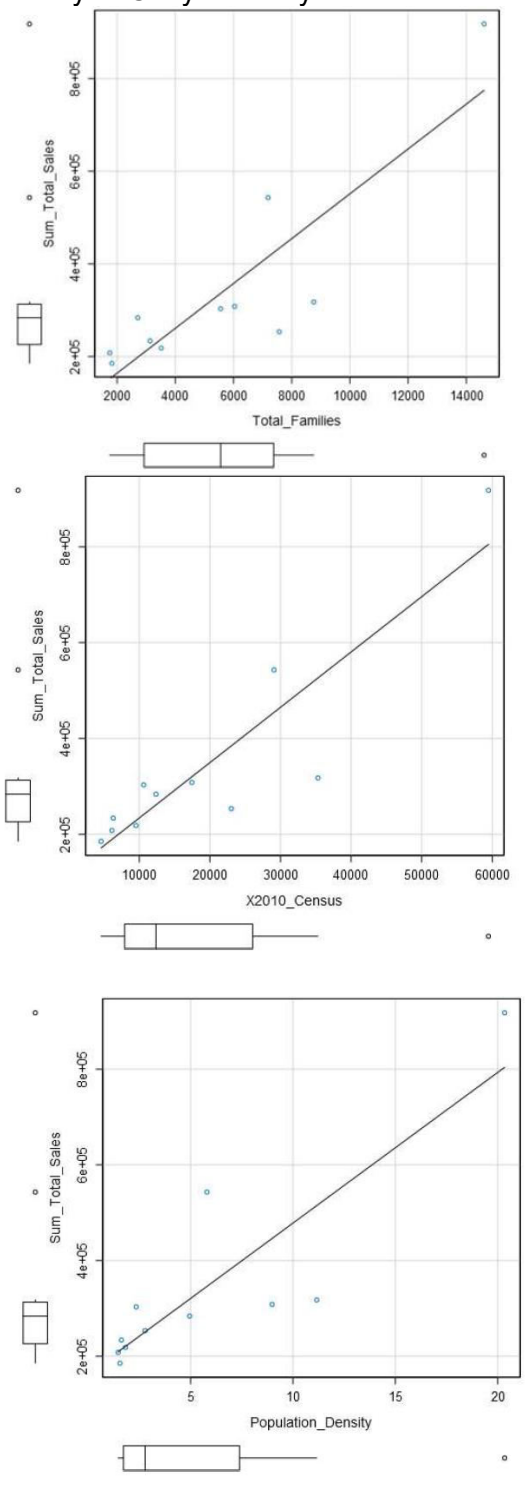
Initially, to investigate the outliers in the training set, we apply IQR analysis. IQR does an estimate of outliers by looking at values more than one-and-a-half times the IQR distance below the first quartile, the 25th quantile, or above the third quartile, the 75th quantile. Then, we utilize the scatter plot that includes boxplots in the margin and the regression line. The data is displayed as a collection of points, each having the value of predictor variable determining the position on the horizontal axis and the value of predicted variable determining the position on the vertical axis.
From the table, we can see the outliers that have been identified for each city with IQR method:

	A	B	C	D	E	F	G
1	CITY	Sum_Total Sales	Land Area	seholds with Unde	opulation Densit	Total Families	2010 Census
2	Buffalo	-	-	-	-	-	-
3	Casper	-	-	-	-	-	-
4	Cheyenne	917,892	-	-	20.34	14,612.64	59,466
5	Cody	-	-	-	-	-	-
6	Douglas	-	-	-	-	-	-
7	Evanston	-	-	-	-	-	-
8	Gillette	543,132	-	-	-	-	-
9	Powell	-	-	-	-	-	-
10	Riverton	-	-	-	-	-	-
11	Rock Springs	-	6,620.20	-	-	-	-
12	Sheridan	-	-	-	-	-	-
13							
14	Q1	226152.00	1861.72	1327.00	1.72	2923.41	7917.00
15	Q3	312984.00	3504.91	4037.00	7.39	7380.81	26061.50
16	Q3 - Q1	86832.00	1643.19	2710.00	5.67	4457.40	18144.50
17	lower	95904.00	-603.06	-2738.00	-6.78	-3762.68	-19299.75
18	upper	443232.00	5969.69	8102.00	15.89	14066.90	53278.25
19							

The city of Rock Springs has one outlier that corresponds to the variable “Land_Area” variable. Given that this point fits the relationship of the rest of the data, it is not removed as we can see in the figure below.



The city of Cheyenne city has three outliers, we see the scatter plot below.



This data point doesn't distort the relationship between predictor variable and sales. Given that this value corresponds with that is the largest in the Wisconsin state, we expect, on average, more sales given the potential demand. Remove this value, we mean to undervalue the total sales for the cities with more population.

The Gillette city has one outlier point that corresponds to the predicted variable, “Sum_Total_Sales”. This point overestimates, on average, the total sales for small cities given that this city has the number of sales high in proportion to the population. Therefore, we remove this point.