Project: Create an Analytical Dataset

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made.

The Business Problem

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

1. What decisions need to be made?

Pawdacity, a leading pet store chain in Wyoming, needs recommendation on where to expand and open a 14th store based on predicted yearly sales. The objective for this project is to analyze the relationship between historical sales data and demographic data to recommend the city for Pawdacity's new store.

2. What data is needed to inform those decisions?

Data needed to inform those decisions include:

(1) monthly sales data for all Pawdacity stores in the year 2010
(2) NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales
(3) population numbers
(4) demographic data (households with individuals under 18, land area, population density, and total families) for each city and county in the state of Wyoming

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.

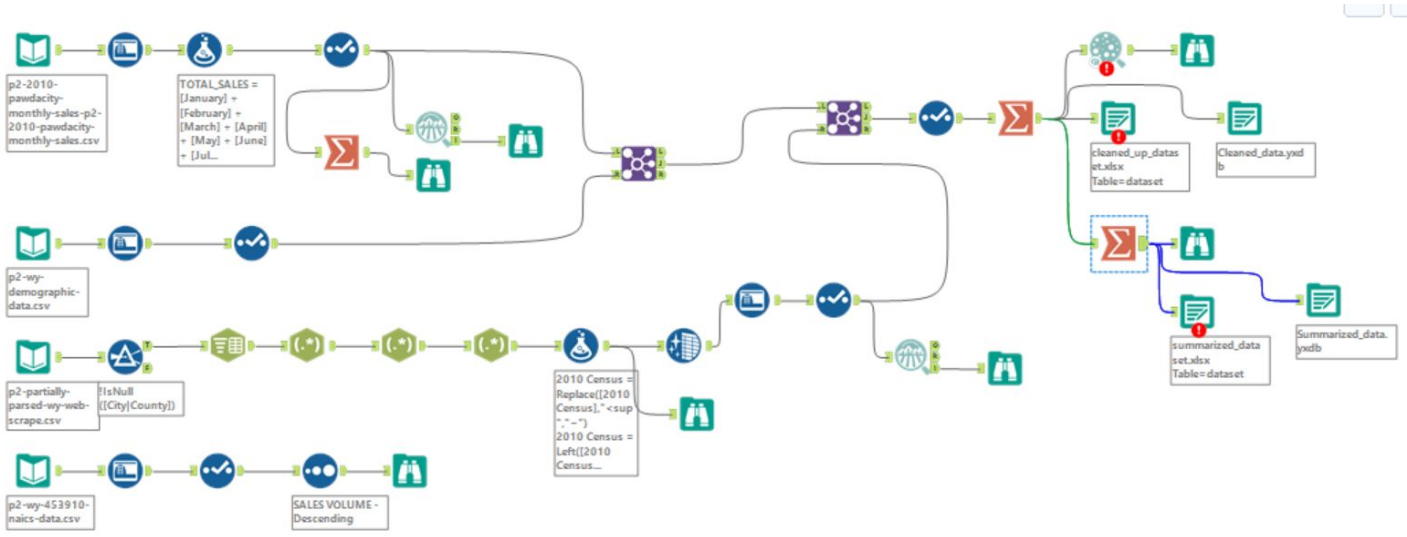
In addition provide the averages on your data set here.

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

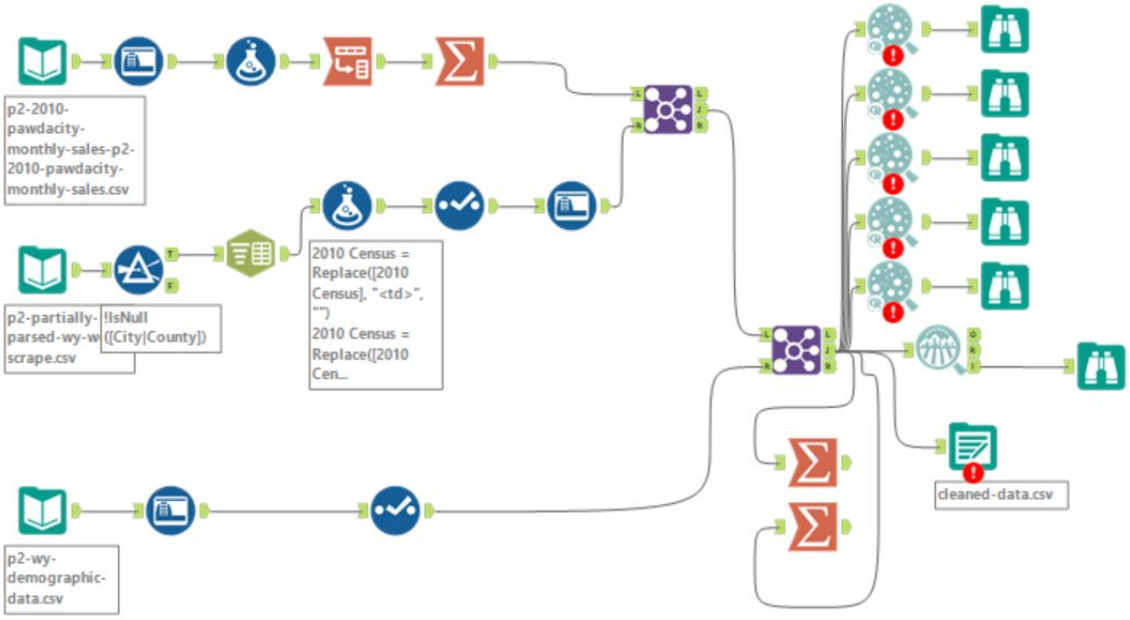
See the Alteryx workflow to obtain the averages for the variables above.

The data for population, sales, and demographic data are cleaned and filtered. Data is aggregated by city and this new dataset for the new store can be used as a predictive model for sales in Pawdacity's 14th store location. The resulting dataset has 11 entries, with each entry representing
a city in which Pawdacity currently operates.

Version 1 workflow



Version 2 workflow



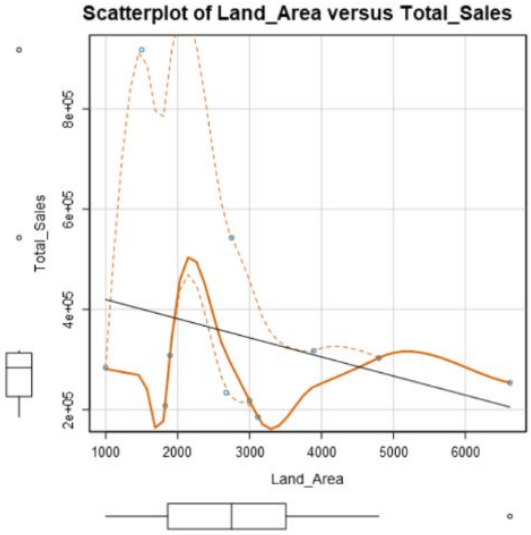
Step 3: Dealing with Outliers

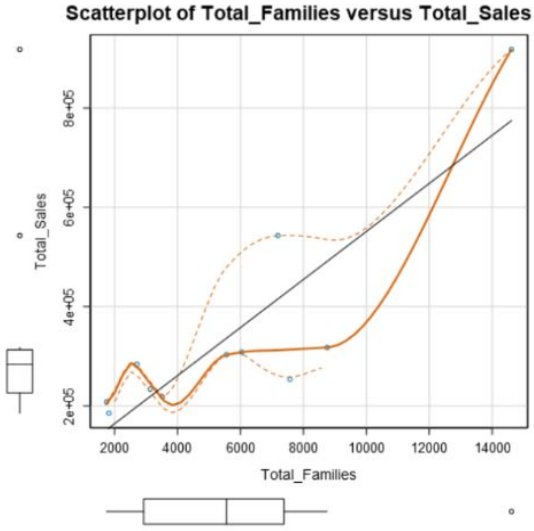
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because the dataset is a small data set (11 cities), you should only remove or impute one outlier. Please explain your reasoning.

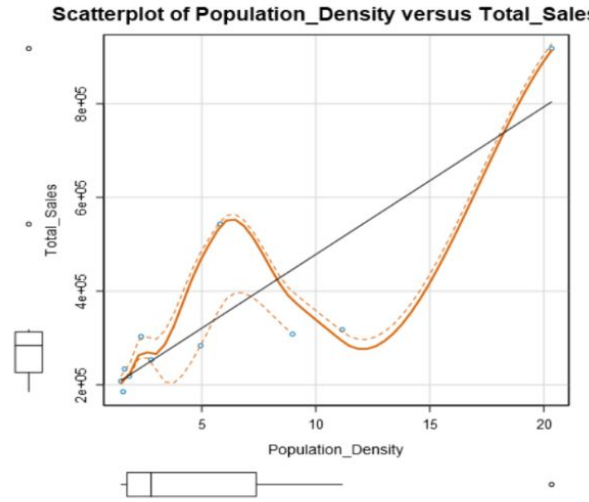
The first possible outlier in the training set is the city of **Cheyenne** because sales data are higher than expected for four categories: 2010 Census population, population density, total families, and total sales). Cheyenne is also the capital city of Wyoming so a high population density, total population, and total families are expected. These three metrics are highly correlated since the scatter Plots demonstrate a linear relationship between total sales against each of those metrics. Therefore, Cheyenne is kept in the dataset to provide a more robust model for predicting results in other populated cities.

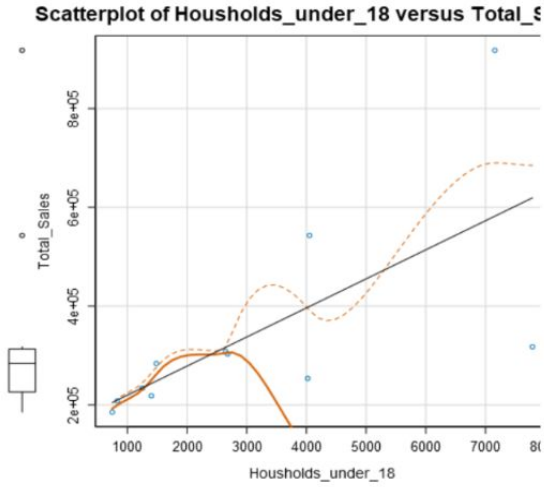
Gillette is the second outlier. Gillette remains an anomaly for its high total sales with all other metrics in the proper range. Because the total sales for this city is an outlier which can't be explained by other variables found in land area, population density & total families, leaving this entry in the dataset will skew any models trained on this data. Therefore, Gillette should be removed from the dataset.

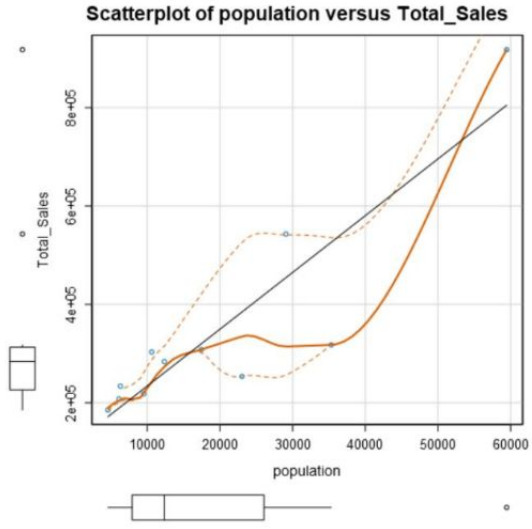
Finally, **Rock Springs** is the final outlier based on land area. The city possesses a higher land area relative to other cities as shown in land area versus total sales yet the land area still fits the linear model with total sales. Thus, Rock Springs will not skew further models and is kept.


Scatterplot of Land_Area versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area


Scatterplot of Total_Families versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families


Scatterplot of Population_Density versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


Scatterplot of Housholds_under_18 versus Total_S
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
Housholds_under_18


Scatterplot of population versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
population

IQR Method for Dealing with Outliers

IQR Steps

City	2010 Census population	Total Sales	Household with Under 18	Land Area	Population Density	Total Families
Buffalo	4585	185328	746	3116	2	1820
Casper	35316	317736	7788	3894	11	8756
Cheyenne*	59466	917892	7158	1500	20	14613
Cody	9520	218376	1403	2999	2	3516
Douglas	6120	208008	832	1829	1	1744
Evanston	12359	283824	1486	999	5	2713
Gillette*	29087	543132	4052	2749	6	7189
Powell	6314	233928	1251	2674	2	3134
Riverton	10615	303264	2680	4797	2	5556
Rock Springs*	23036	253584	4022	6620	3	7572
Sheridan	17444	308232	2646	1894	9	6040

Q1 (25th percentile)	7917	226152	1327	1861.5	2	2923.5
Q3 (7th percentile)	26061.5	312984	4037	3505	7.5	7380.5
IQR (Q3-Q1)	18144.5	86832	2710	1643.5	5.5	4457
Lower Q1-1.5xIQR	-19299.75	95904	-2738	-603.75	-6.25	-3762
Upper Q3+1.5xIQR	53278.25	443232	8102	5970.25	15.75	14066