Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

A. Pawdacity currently has 13 stores and need recommendation to open 14th store. The key decision they have to make is to select city for their new store.

: Awesome: Correct! This is indeed the main business decision to be made.

2. What data is needed to inform those decisions?

A. The decisions can be driven by dataset having information of each city in which Pawadacity store is currently have their stores.

1. Land Area
2. Population Density
3. No. of household with people under 18
4. Total No. of families
5. Total Pawdacity sales in 2010
6. City population according to 2010 Census

: Awesome: Good work identifying this. This data should be good enough for part 1 of the analysis.

Further, while predicting we can consider data for other cities where we can open new store

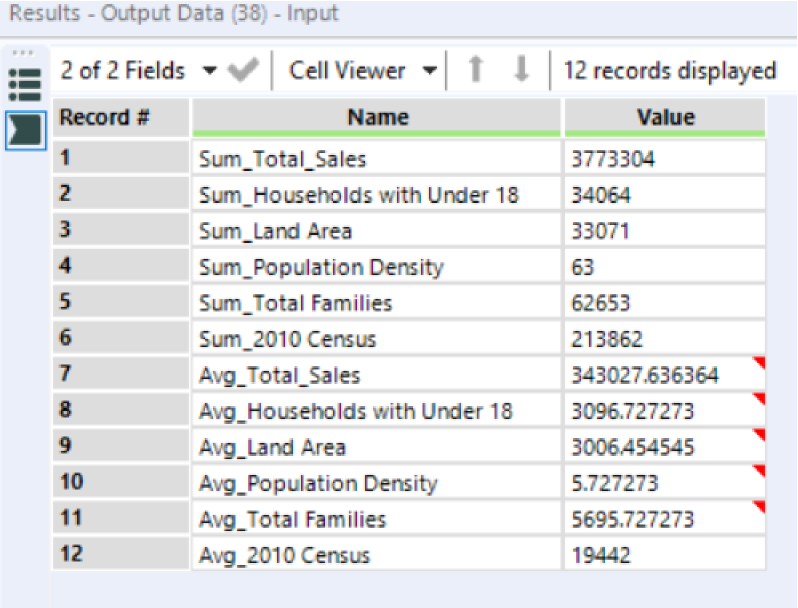
1. Sales data of other stores in city
2. Demographics information i.e. Landmarks, Competitors' store location

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.

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*	*3006.45*
Population Density	*63*	*5.7*
Total Families	*62,653*	*5695.72*

: Awesome: well done! All the sum & averages are perfectly correct!



Results - Output Data (38) - Input

2 of 2 Fields | Cell Viewer | 12 records displayed

Record #	Name	Value
1	Sum_Total_Sales	3773304
2	Sum_Households with Under 18	34064
3	Sum_Land Area	33071
4	Sum_Population Density	63
5	Sum_Total Families	62653
6	Sum_2010 Census	213862
7	Avg_Total_Sales	343027.636364
8	Avg_Households with Under 18	3096.727273
9	Avg_Land Area	3006.454545
10	Avg_Population Density	5.727273
11	Avg_Total Families	5695.727273
12	Avg_2010 Census	19442

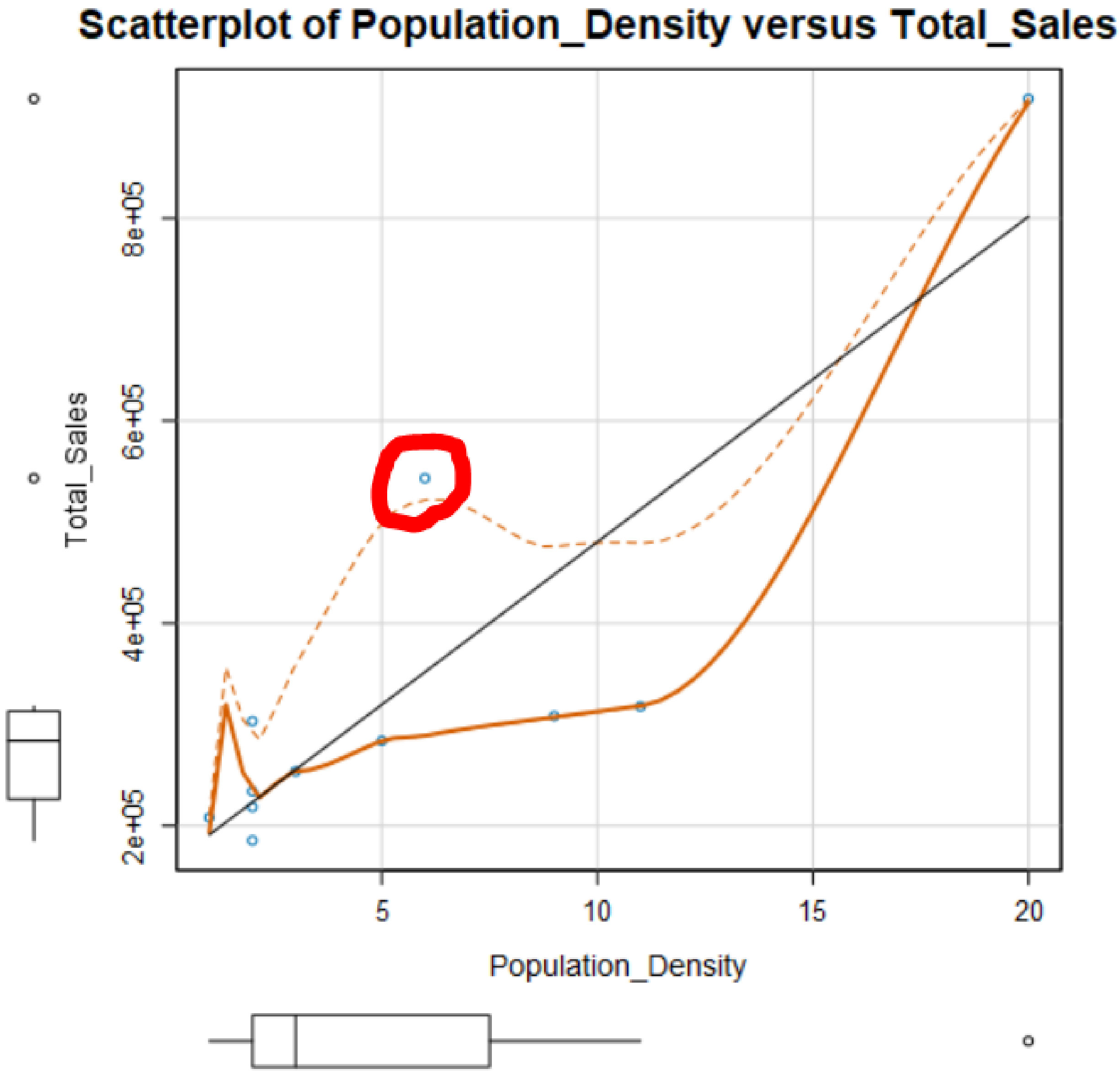
Step 3: Dealing with Outliers

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.

1 Gillette

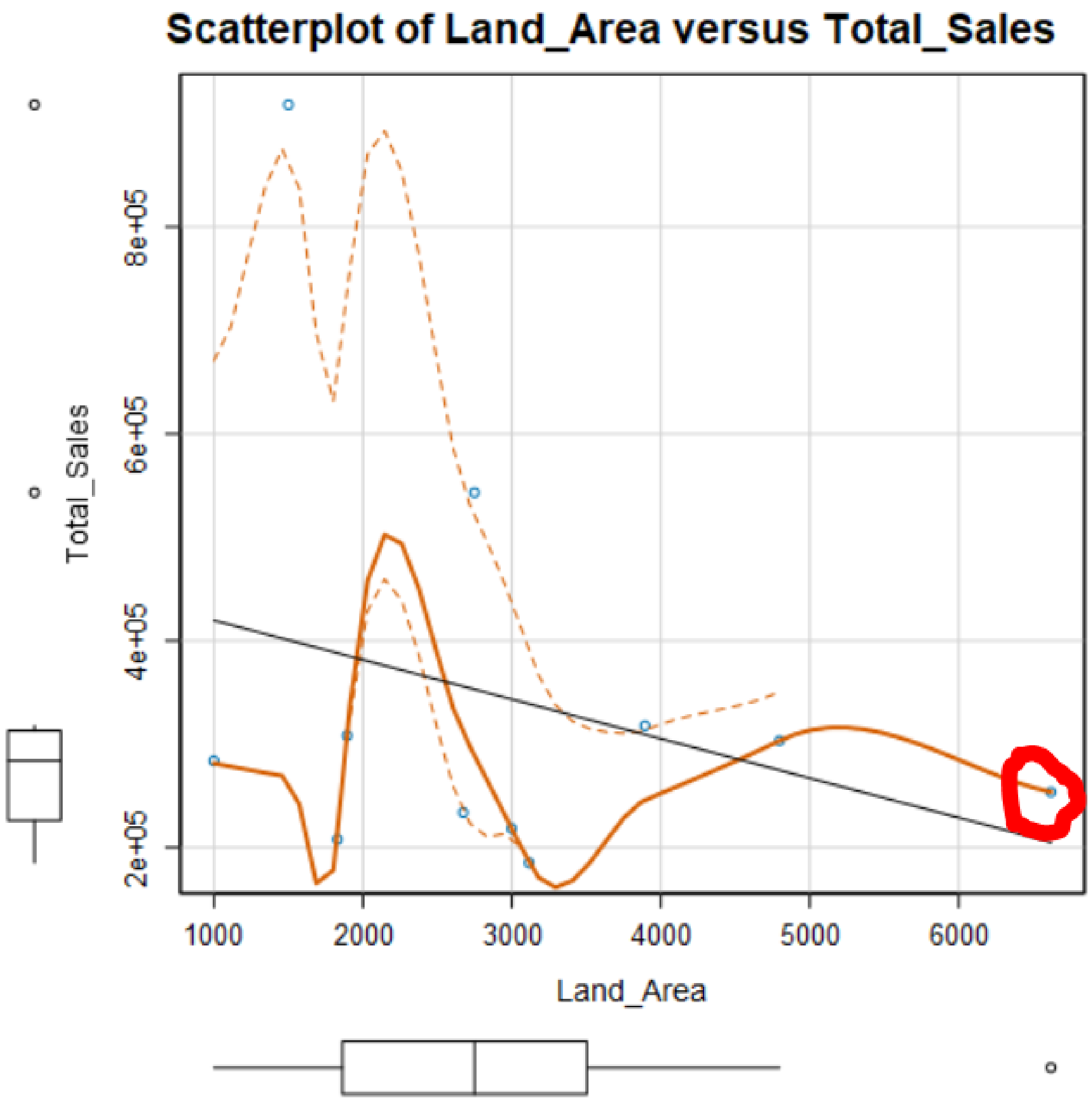
We can observe demographic numbers of the city seems normal but sales are very high which makes it anomaly. In case of higher no. of people, we can expect more no. of sales but Gillette is a small city with very high amount of sales compared to other cities. Hence we will remove it.

: Awesome: Gillette is a true anomaly because its demographic numbers are within the expected range, yet the Pawdacity sales are really high, which doesn't make sense given the traditional understanding that if we have a higher number of people in an area, we should expect a bigger volume of sales, but Gillette is a small city with a very high amount of sales compared to the other cities in the training set. Therefore we should remove it.



2 Rock Springs

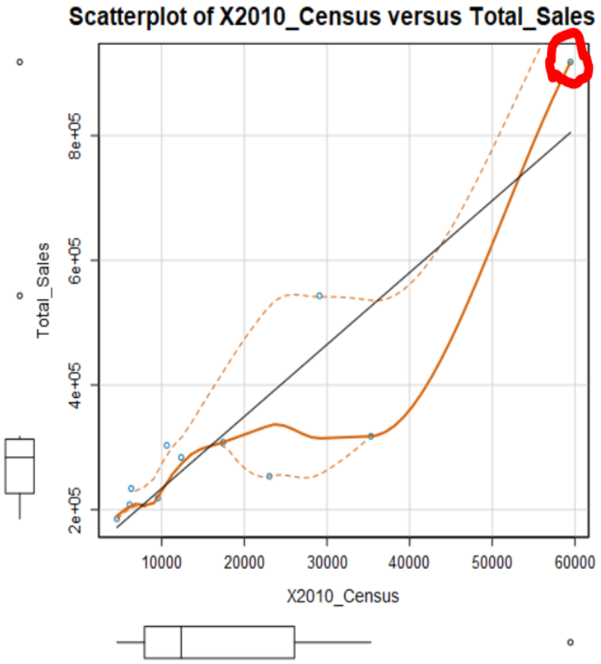
Rock Springs is an outlier based on land area. Despite of higher land area it still correlates with linear relationship. Land area may not have high impact on sales data, hence we leave it in.

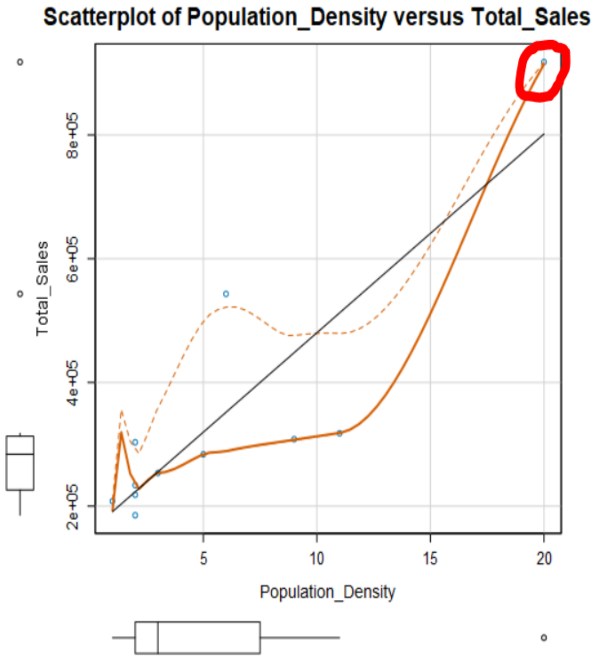


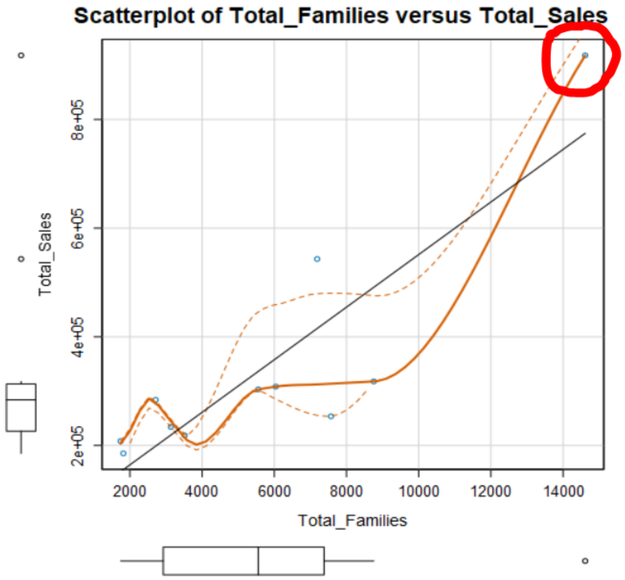
3 Cheyenne

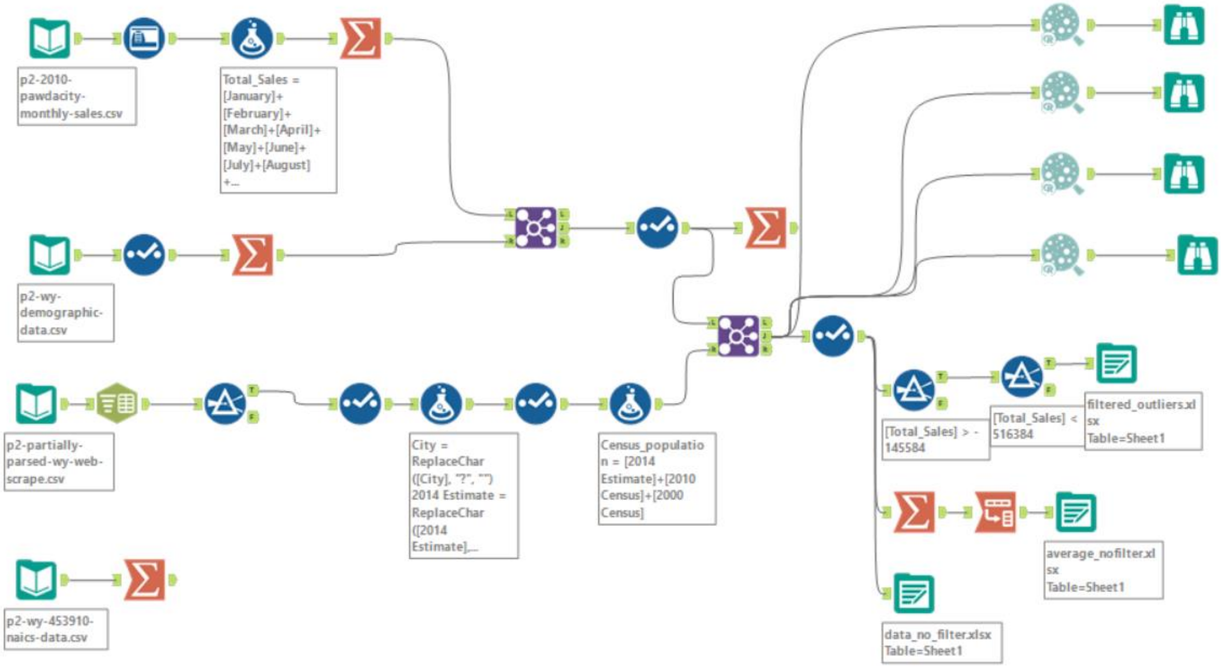
We can observe Cheyenne exceeds the range in, 2010 Census populations, total families, population density, and total sales but Cheyenne is a big city and its numbers are also higher than other cities in every field of training set. It also follows the linear relationship. Hence we consider, Cheyenne is not an anomaly.

: Awesome:
- Cheyenne has huge sales but as you also saw it has high population density.
- Which means that the high sales are justified by the high population density.
- Cheyenne is indeed a big urban city.
- We can conclude that Cheyenne is not an anomaly, but just a big city given the other smaller cities in the available training set and would want to include this big city to have a more robust model so we can model any future cities with big numbers.


Scatterplot of X2010_Census versus Total_Sales


Scatterplot of Population_Density versus Total_Sales


Scatterplot of Total_Families versus Total_Sales


p2-2010-pawdacity-monthly-sales.csv
Total_Sales = [January]+ [February]+ [March]+[April]+ [May]+[June]+ [July]+[August] +...
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
City = ReplaceChar ([City], "?", "") 2014 Estimate = ReplaceChar ([2014 Estimate],...
Census_population = [2014 Estimate]+[2010 Census]+[2000 Census]
p2-wy-453910-naics-data.csv
L
J
R
T
F
[Total_Sales] > - 145584
[Total_Sales] < 516384
filtered_outliers.xlsx Table=Sheet1
average_nofilter.xlsx Table=Sheet1
data_no_filter.xlsx Table=Sheet1