Project 2.1: Data Cleanup

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.

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

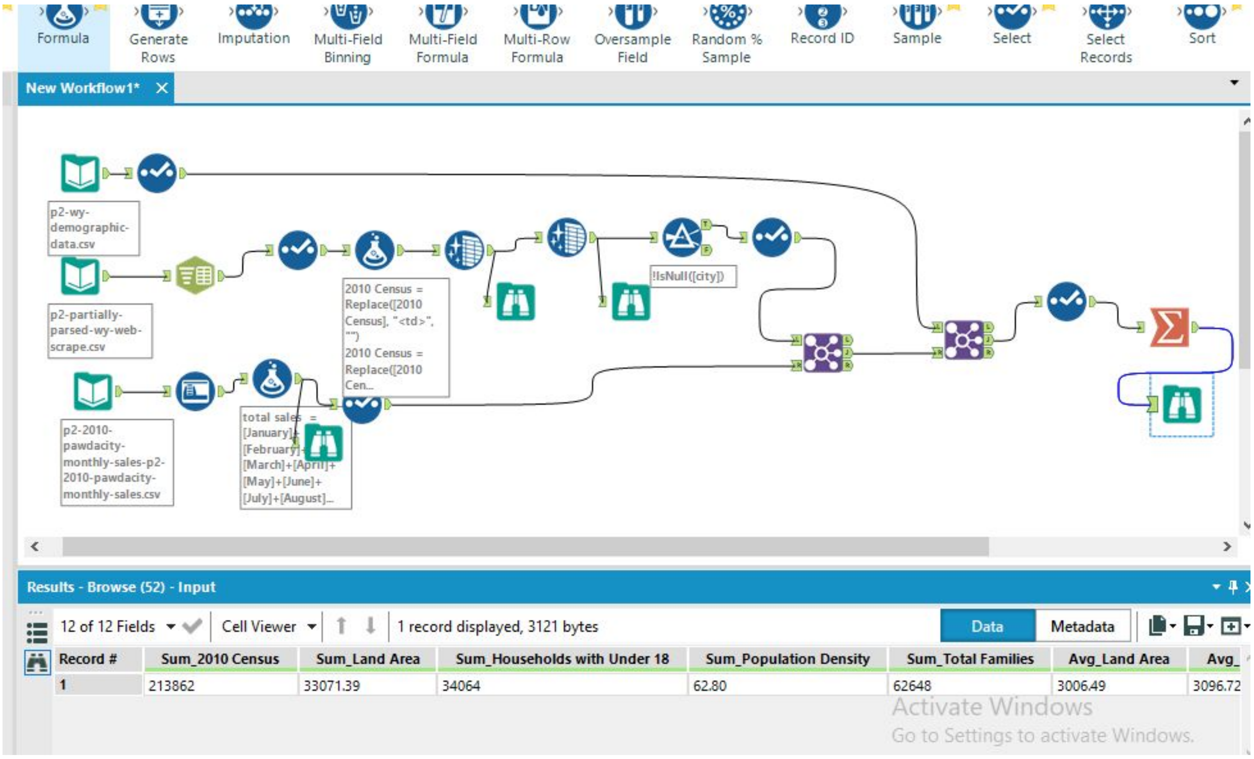
1. What decisions needs to be made?

The decision needs to be made where to open the 14th Pawdacity pet store.

2. What data is needed to inform those decisions?

Some of the data required in order to inform this decision are city, population, Pawdacity sales in other stores, competitor sales, household with under 18, land area, total families.

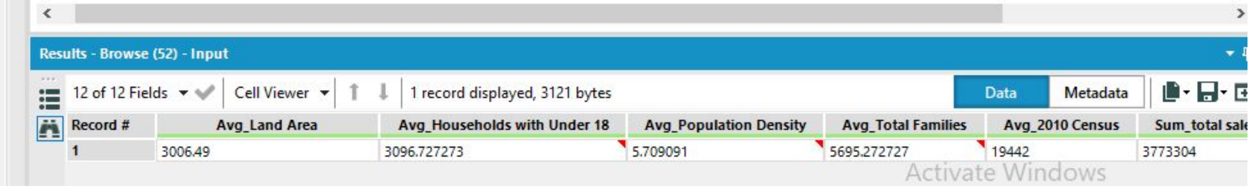
Step 2: Building the Training Set


Formula
Generate Rows
Imputation
Multi-Field Binning
Multi-Field Formula
Multi-Row Formula
Oversample Field
Random % Sample
Record ID
Sample
Select
Select Records
Sort
New Workflow1*
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
2010 Census = Replace([2010 Census], "<td>", "")
2010 Census = Replace([2010 Cen...
!IsNull([city])
total sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]+[August]...

Results - Browse (52) - Input

12 of 12 Fields · Cell Viewer · 1 record displayed, 3121 bytes · Data · Metadata

Record #	Sum_2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Avg_Land Area	Avg_
1	213862	33071.39	34064	62.80	62648	3006.49	3096.72



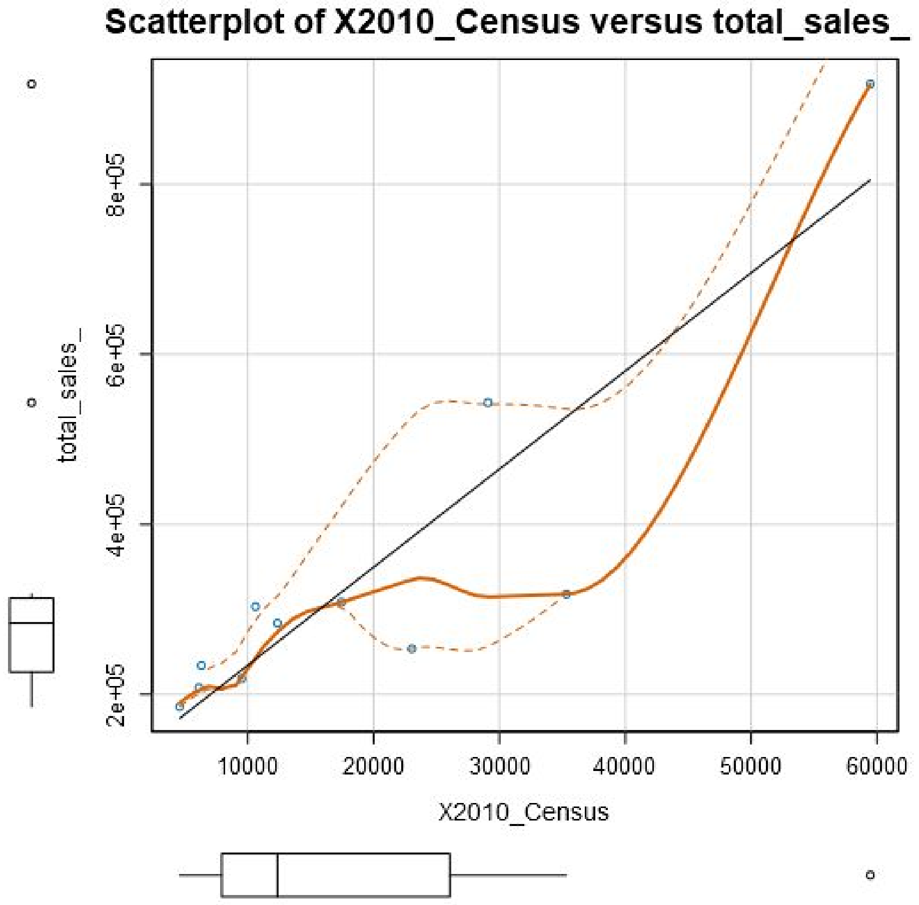
Results - Browse (52) - Input

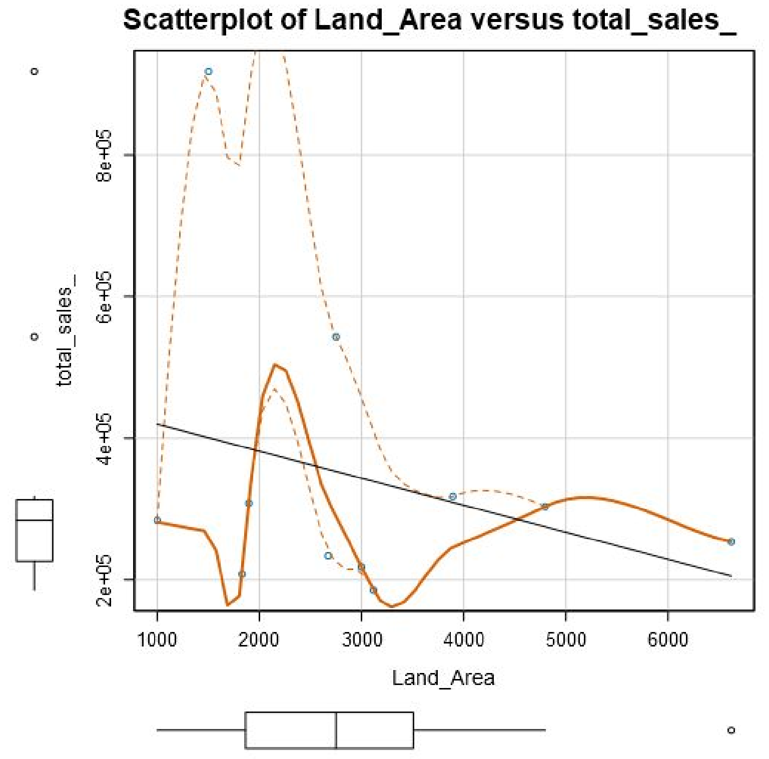
12 of 12 Fields · Cell Viewer · 1 record displayed, 3121 bytes · Data · Metadata

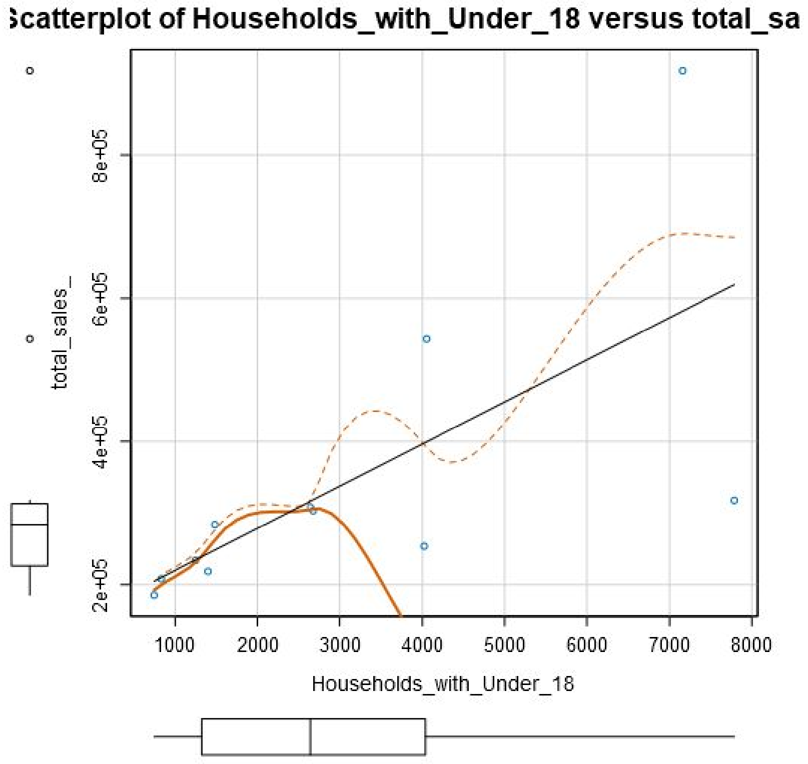
Record #	Avg_Land Area	Avg_Households with Under 18	Avg_Population Density	Avg_Total Families	Avg_2010 Census	Sum_total sale
1	3006.49	3096.727273	5.709091	5695.272727	19442	3773304

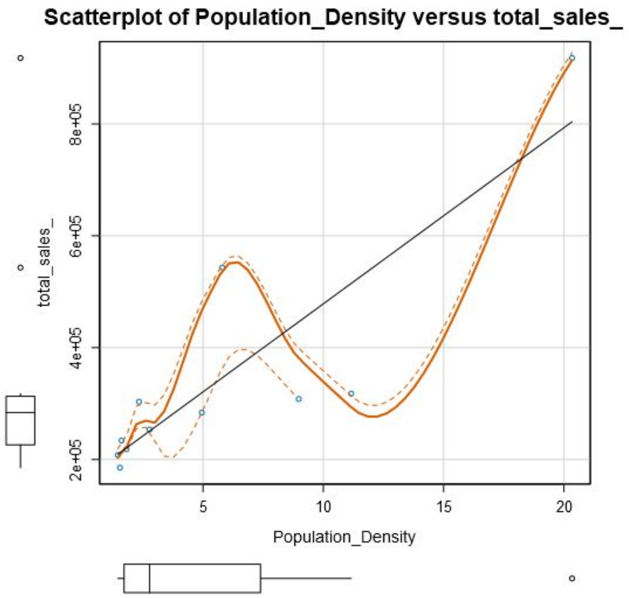
Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.63*
Households with Under 18	*34,064*	*3096.72*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.70*
Total Families	*62,653*	*5695.27*

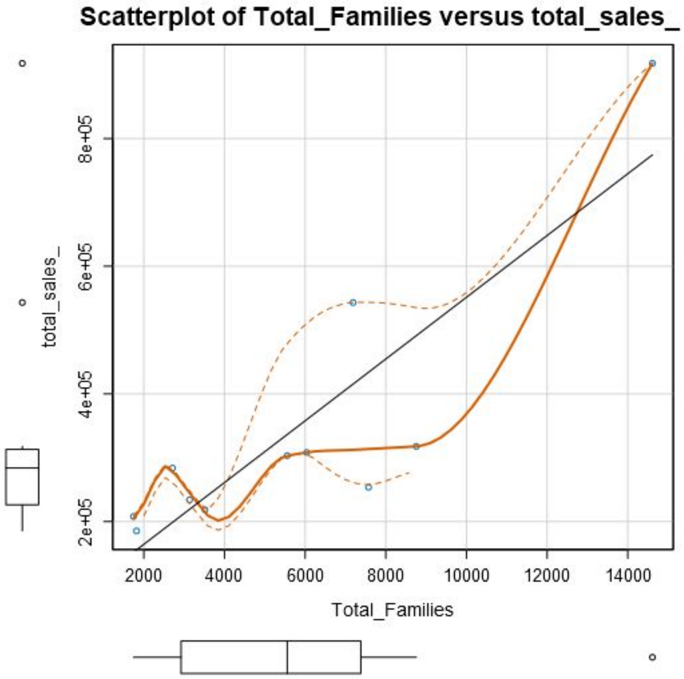
Step 3: Dealing with Outliers


Scatterplot of X2010_Census versus total_sales_
total_sales_
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
X2010_Census


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


Scatterplot of Households_with_Under_18 versus total_sa
total_sales_
8e+05
6e+05
4e+05
2e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18





Based on the scatterplots above, the city of **Gillette** and **Cheyenne** seem to be the outliers as their sales data are higher than expected.

Since the relationships between Gillette's population related variables and total sales are still correlated, Gillette should be kept for analysis.

My suggestion would be to remove the city of Cheyenne and keep the city of Gillette for further analysis.