Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. In this part of project we need to prepare data for modelling in next part of Project by cleansing blending, formatting data and deal with outliers present in the dataset

2. Perform an analysis to recommend the city for pawdacity's newest store, based on predicted yearly sales to suggest a location to pawdacity to open a new 14th store.

Data required

Demographic data, Monthly sales of pawdacity, Population data

1. City
2. 2010 Census Population
3. Total Pawdacity Sales
4. Households with Under 18
5. Land Area
6. Population Density
7. Total Families

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	19442
Total Pawdacity Sales	*3,773,304*	343028
Households with Under 18	*34,064*	3097
Land Area	*33,071*	3006
Population Density	*63*	6
Total Families	*62,653*	5696

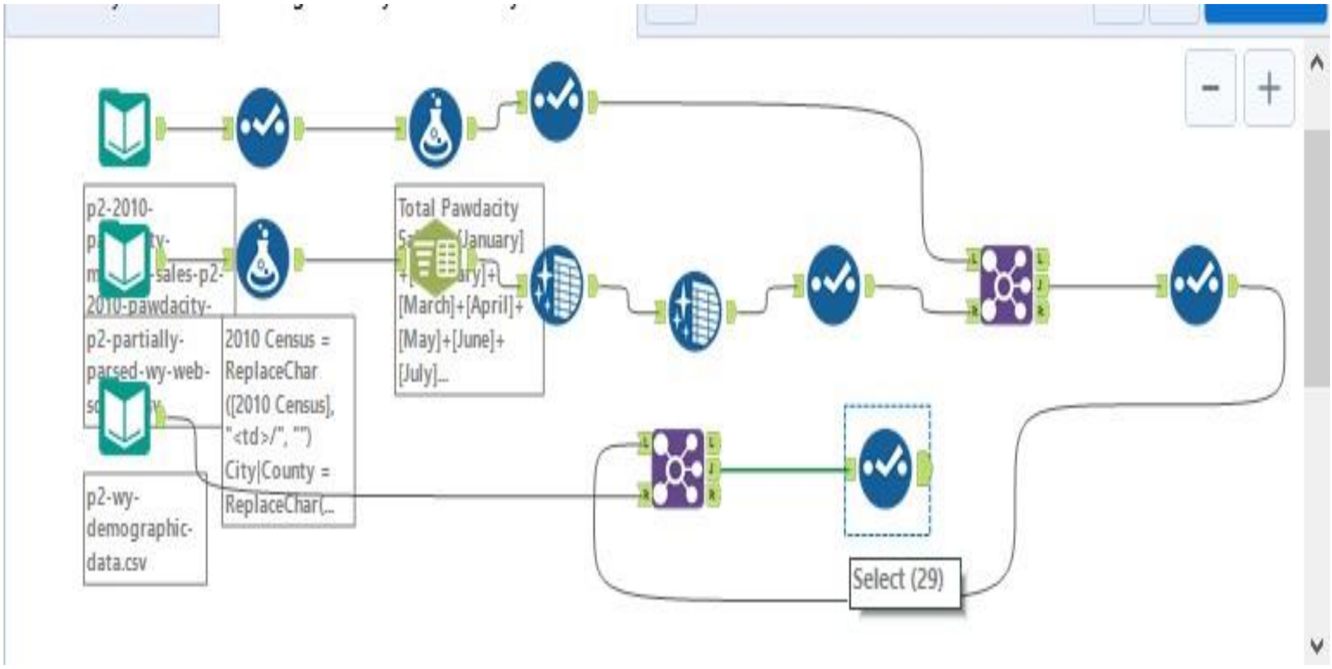
Step 3: Dealing with Outliers

Outliers in training Dataset :

1. Cheyenee in 2010Census population have a value greater than upper fence (53278.25) is 59466
2. Rock Springs in Land Area have value greater than upper fence (5969.689) is 6620.202
3. Cheyenee in Total families have value greater than upper fence (14066.9) is 14612.64
4. Cheynee and Gillete in Total Pawdacity Sales has value greater than upper fence(443232) is 917892 and 543132.

City Cheyenne need tp remove from the dataset because it is an outlier present in three attributes of training dataset which may effect the result of data analysis and modelling in predicting location of next store.

The given data set is too small removing too many outliers from the dataset may decrease the amount of data and statistical value in datasets like Average , standard deviation and can directly effect the results and assumptions.


p2-2010-p...ty-m...-sales-p2-2010-pawdacity-
p2-partially-parsed-wy-web-sc...
p2-wy-demographic-data.csv
2010 Census = ReplaceChar ([2010 Census], "<td>/", "") City|County = ReplaceChar(...
Total Pawdacity Sa... [January] +[...ary]+ [March]+[April]+ [May]+[June]+ [July]...
Select (29)