Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We need to recommend the location of a new store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

To perform this analysis, I have used data sets:

- Data related to existing stores.
- Data related to demographic information.

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.

By performing the data wrangling on the existing data set, I have ended up with the numbers 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*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

Answer these questions

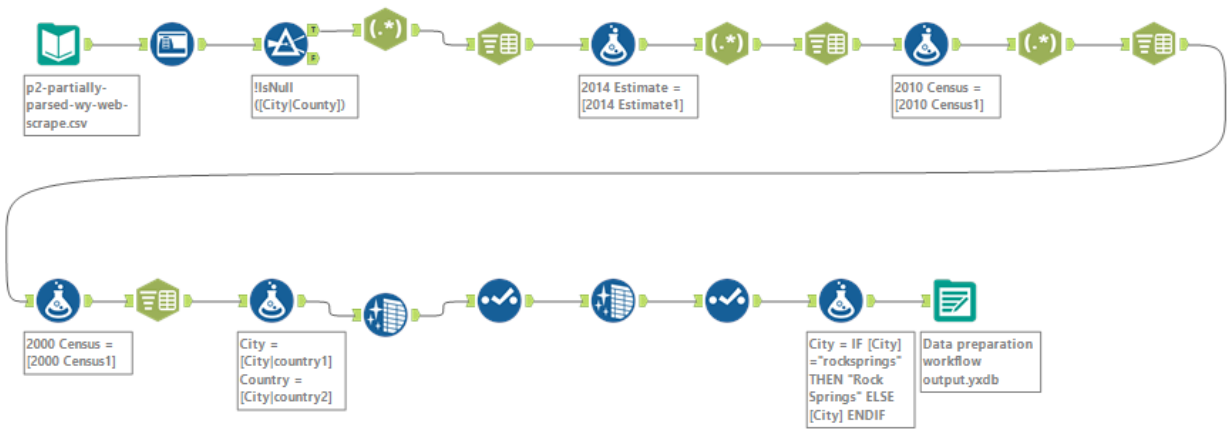
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

In my opinion, there are two outliers that I have an identifier using the IQR method, and I would suggest removing both of those. Please see the tables bellows to see how did I identify them.

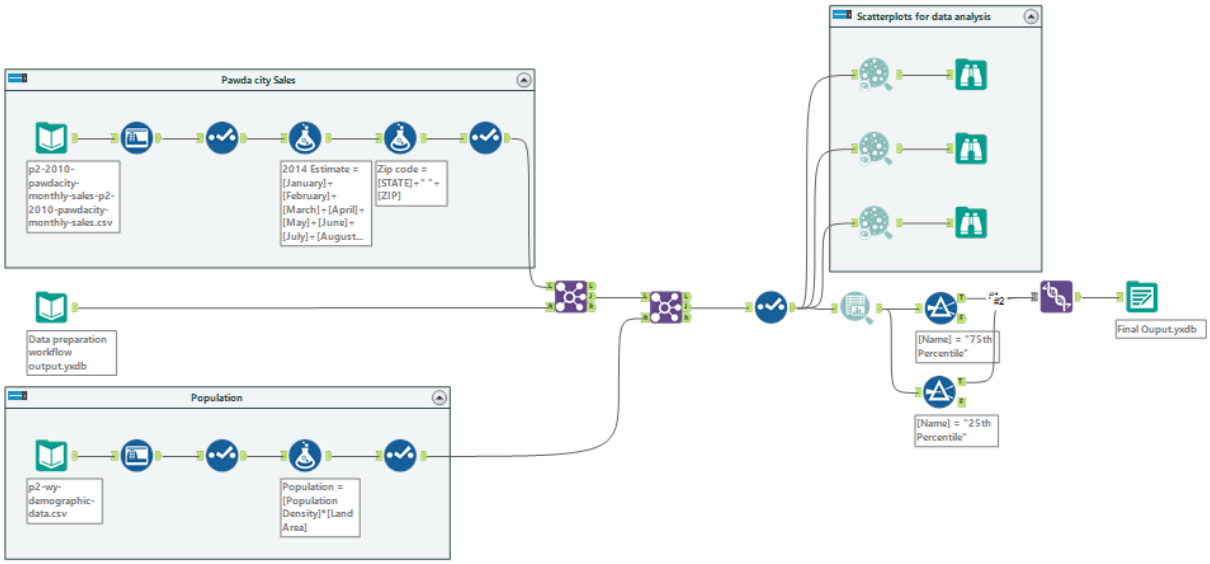
CITY	Correct Population	Bigger than Upper Fence	Smaller than Lower Fence	2014 Estimate	Bigger than Upper Fence	Smaller than Lower Fence
Cheyenne	59466	TRUE	FALSE	917892	TRUE	FALSE
Gillette	29087	FALSE	FALSE	543132	TRUE	FALSE

	Correct Population	2014 Estimate
Q3	26061.5	312984
Q1	7917	226152
IQR	18144.5	86832
Upper Fence	53278.25	443232
Lower Fence	-19299.75	95904

Alteryx Workflows



Workflow 1. Data Preparation Part



Workflow 2. Merging the data and using basic data profile to receive the