Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**

I need to recommend a city for the pawdacity's newest store based on the predicted yearly sales.

2. **What data is needed to inform those decisions?**

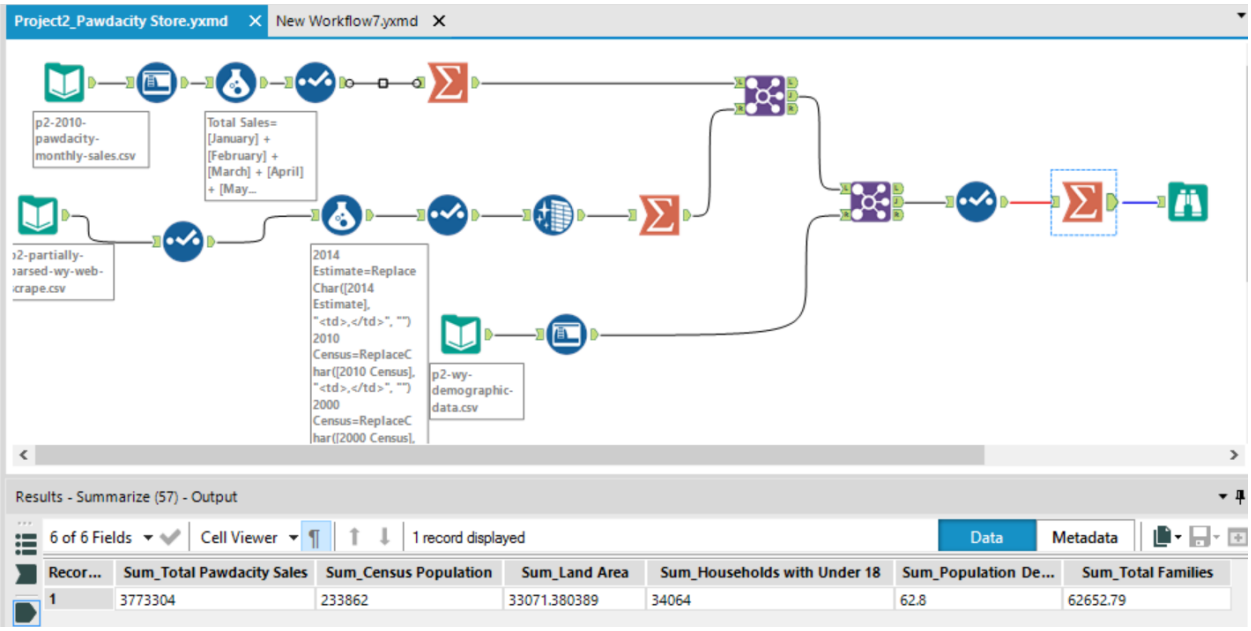
We need the information about the existing stores of pawdacity in the state and the sales data of each store along with its location information, we also need the social demographics information of all the cities in the Wyoming state so that it would help us in the analysis of the data and to come up with the city for new store

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 to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	Sum	Average
Census Population	*233,862*	21260.18
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*	5.73
Total Families	*62,653*	5695.73



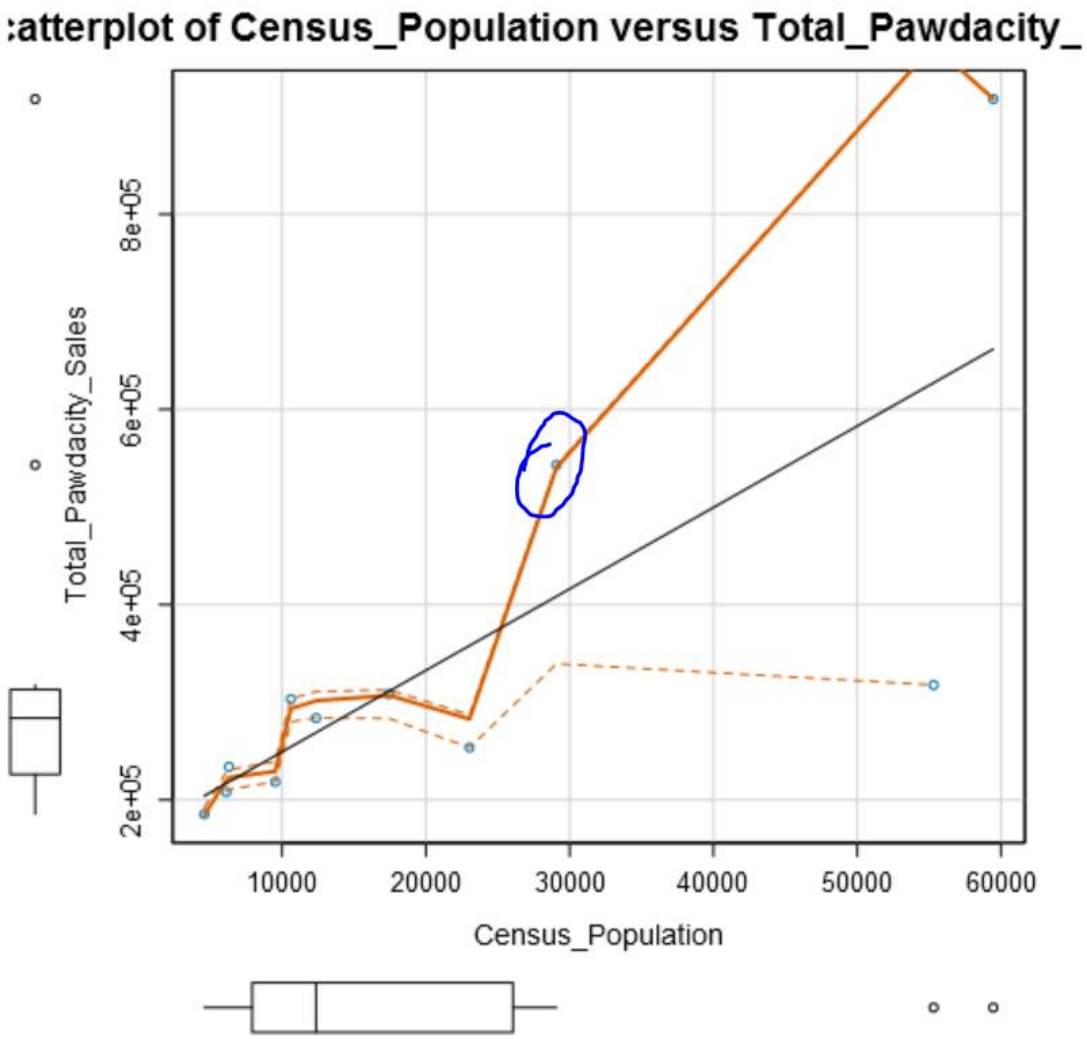
Work flow creating the analytical dataset

Step 3: Dealing with Outliers

Answer these questions

Are there any 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.

Yes, there is an outlier in the training set.
We can consider the Gillette city data as an outlier, since the total sales in that city is an outlier is above the upper fence, Since Cheyenne is a large city with more population and total families we need that data for further analysis and predicting the new store location



Q1	226152
Q3	312984
IQR = Q3 - Q1	86832
Upper Fence = Q3 + 1.5 * IQR	443232
Lower Fence = Q1 - 1.5 * IQR	95904

The sales value for Gillette is more than the upper fence value and hence we are considering that as a outlier.