Predictive Analytics for Business Nanodegree

Project 2.1: Data Cleanup

Latifa M. Alyaeesh

Misk Academy & Udacity

Step 1: Business and Data Understanding

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

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.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (**Households with individuals under 18, Land Area, Population Density, and Total Families**) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Key Decisions:

Answer these questions

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

The decision needs to recommend the city for Pawdacity's newest store, Because of the Pawdacity would like to expand and open a 14th store.

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

We need to predicted yearly sales data to choose city highest predicted yearly sales for a 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	213,862	19,442
Total Pawdacity Sales	3,773,304	343,027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.49
Population Density	63	5.71
Total Families	62,653	5,695.71

Hint: In the table below, I explained how could I reached out the average result.

Step 3: Dealing with Outliers

City	Total Pawdacity Sales	Land Area	Households with under 18	Population Density	Total Families	2010 Census
Buffalo	185,328.00	3,115.51	746.00	1.55	1,819.50	4,585.00
Casper	317,736.00	3,894.31	7,788.00	11.16	8,756.32	35,316.00
Cheyenne	917,892.00	1,500.18	7,158.00	20.34	14,612.46	59,466.00
Cody	218,376.00	2,998.96	1,403.00	1.82	3,515.56	9,520.00
Douglas	208,008.00	1,829.47	832.00	1.46	1,744.08	6,120.00
Evanston	283,824.00	999.50	1,486.00	4.95	2,712.64	12,359.00
Gillette	543,132.00	2,748.85	4,052.00	5.80	7,189.43	29,087.00
Powell	233,928.00	2,673.57	1,251.00	1.62	3,134.18	6,314.00
Riverton	203,264.00	4,796.86	2,680.00	2.34	5,556.49	10,615.00
RockSprings	253,584.00	6,620.20	4,022.00	2.78	7,572.18	23,036.00
Sheridan	308,232.00	1,893.98	2,646.00	8.98	6,039.71	17,444.00
SUM	3,673,304.00	33,071.38	34,064.00	62.80	62,652.55	213,862.00
Average	333,936.73	3,006.49	3,096.73	5.71	5,695.69	19,442.00
Approximately	343,027.64	3,006.49	3,096.73	5.71	5,695.71	19,442.00
Q1	213192	1861.72107	1327	1.72	2923.41	7917
Q3	312984	3504.9083	4037	7.39	7380.805	26061.5
IRQ	99792	1643.18723	2710	5.67	4457.395	18144.5
Upper Fence	462672	5969.689145	8102	15.895	14066.8975	53278.25
Lower Fence	63504	-603.059775	-2738	-6.785	-3762.6825	-19299.75
Outliers	Cheyenne & Gillette	RockSprins	//////////////////////	Cheyenne	Cheyenny	Cheyenny

Table 1: 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.

I have highlighted outliers in the red color. Clearly seen has outliers in four categories (Total Pawdacity Sales, Population Density, Total Families, 2010 Census). For **Cheyenne City** justifiable compared to its high population & Total Families 2010 Census. On other side, **RockSprings City** is having only one outlier "Land Area". For **Gillette City** is having high sales while still maintaining other values within the IQR range. Eventually, I highly suggest removing **Gillette City** as this city does not have logical explanation for that outlier value.

Alteryx workflow

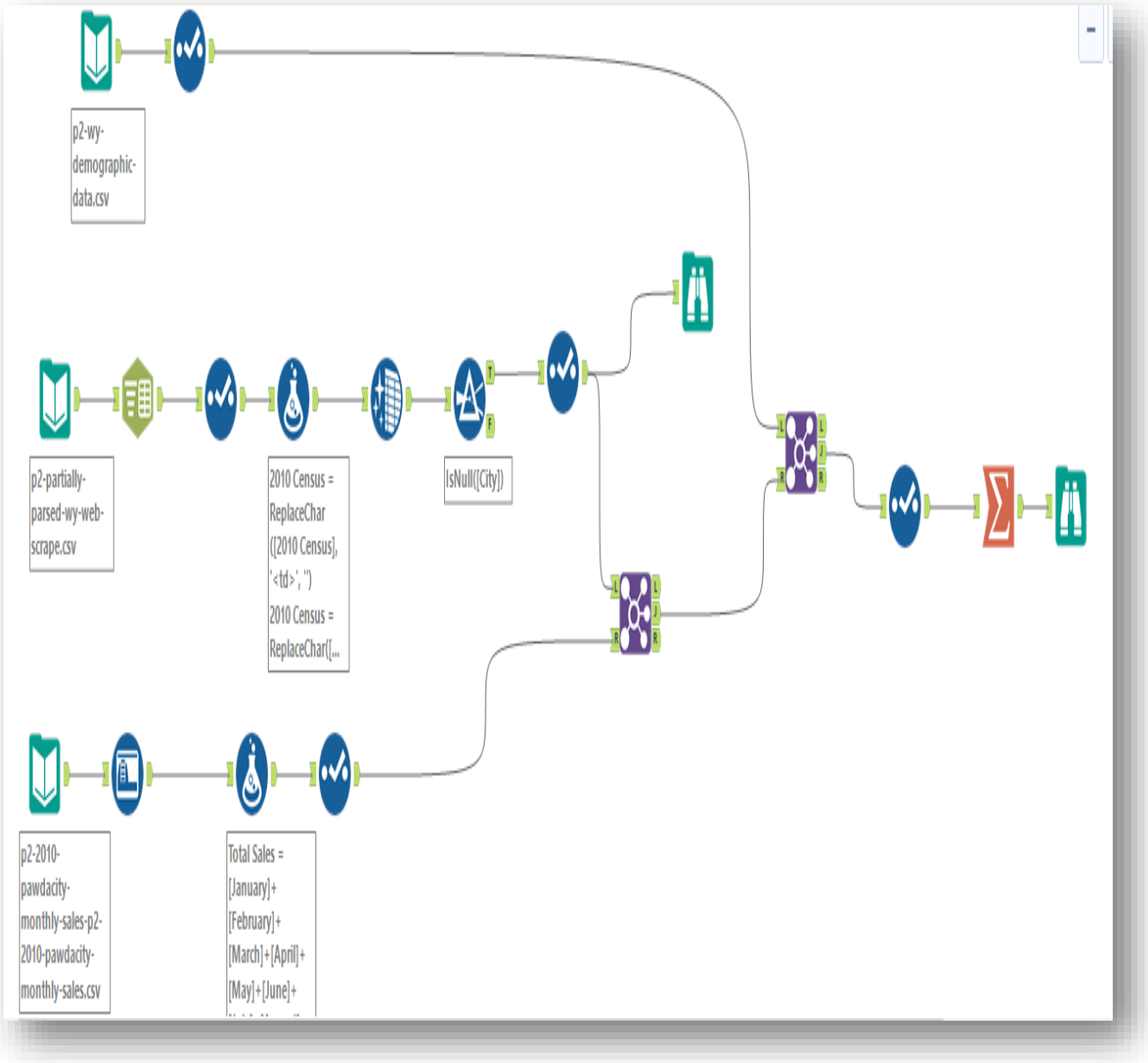

Figure 1: Alteryx workflow