Predictive Analytics for Business Nanodegree

Project: Create an Analytical Dataset

By Samuel Otisi

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The key decision that needs to be made is that Pawdacity store managers are looking for the best city to open a 14th store coupled with the already existing 13 stores in the state of Wyoming.

2. What data is needed to inform those decisions?

Previous monthly sales data from across various cities in Wyoming will be needed to build a model that will predict yearly sales of Pawdacity stores. The Sales data will be blended, formatted and cleaned with census and demographic data across the cities in Wyoming.

Step 2: Building the Training Set

After blending and cleaning the dataset I came up with the following data as a sum and averages of their respective data fields. The result obtained is in line with the expected results from the project reviewers.

Column	Sum	Average
Census Population	*213862*	*19442*
Total Pawdacity Sales	*3773304*	*343027.64*
Households with Under 18	*34064*	*3096.73*
Land Area	*33071.38*	*3006.49*
Population Density	*62.80*	*5.71*
Total Families	*62652.79*	*5695.71*

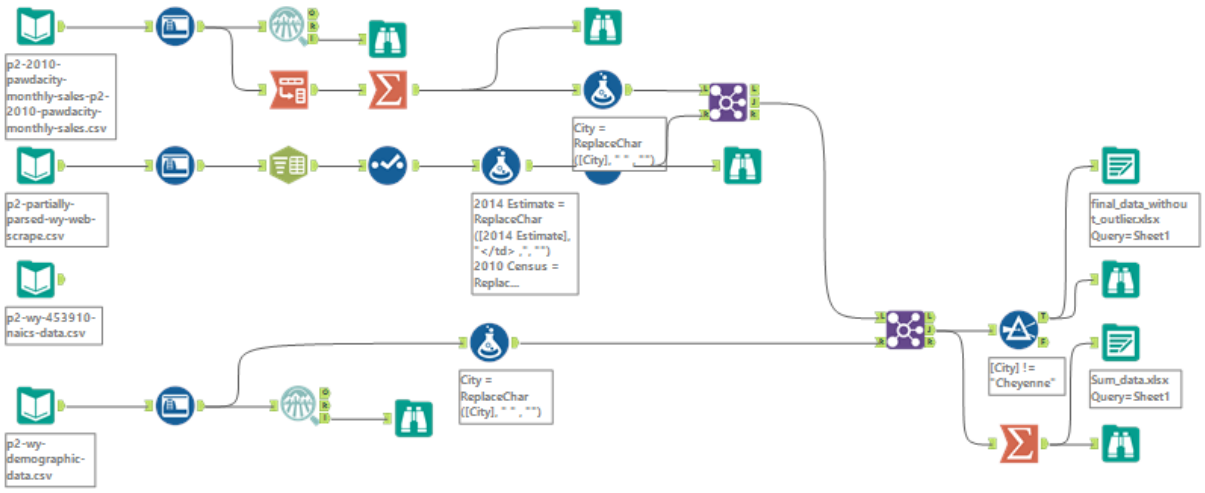
After cleaning up the datasets I came up with 11 rows of data, the table below is the final dataset before removing the outlier.

Figure 1. final dataset after cleaning and blending.

Record	City	County	Total_Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	Johnson	185328	4585	3115.5075	746	1.55	1819.5
2	Casper	Natrona	317736	35316	3894.3091	7788	11.16	8756.32
3	Cheyenne	Laramie	917892	59466	1500.1784	7158	20.34	14612.64
4	Cody	Park	218376	9520	2998.95696	1403	1.82	3515.62
5	Douglas	Converse	208008	6120	1829.4651	832	1.46	1744.08
6	Evanston	Uinta	283824	12359	999.4971	1486	4.95	2712.64
7	Gillette	Campbell	543132	29087	2748.8529	4052	5.8	7189.43
8	Powell	Park	233928	6314	2673.57455	1251	1.62	3134.18
9	Riverton	Fremont	303264	10615	4796.859815	2680	2.34	5556.49
10	RockSprings	Sweetwater	253584	23036	6620.201916	4022	2.78	7572.18
11	Sheridan	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

Also, the following image is the accompanied alteryx workflow, showing how I arrived at the above table. All files are attached in the zip file.

Figure 2. Alteryx workflow.



Step 3: Dealing with Outliers

I have used the box and whisker method to detect an outlier in final cleaned dataset, according to my workout there is an outlier in the Cheyenne City, the total sales, Population density and Total families fileds are above the threshold for evaluating outliers. There is also another outlier in the Gillette City for the Total Sales field, but I have decided to keep it since I have only small dataset. This is the result from my calculations with excel.

Figure 3. result.

	City	County	Total_Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
2	Buffalo	Johnson	185328	4585	3115.5075	746	1.55	1819.5
3	Casper	Natrona	317736	35316	3894.3091	7788	11.16	8756.32
4	Cheyenne	Laramie	917892	59466	1500.1784	7158	20.34	14612.64
5	Cody	Park	218376	9520	2998.95696	1403	1.82	3515.62
6	Douglas	Converse	208008	6120	1829.4651	832	1.46	1744.08
7	Evanston	Uinta	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	Campbell	543132	29087	2748.8529	4052	5.8	7189.43
9	Powell	Park	233928	6314	2673.57455	1251	1.62	3134.18
10	Riverton	Fremont	303264	10615	4796.859815	2680	2.34	5556.49
11	RockSprings	Sweetwater	253584	23036	6620.201916	4022	2.78	7572.18
12	Sheridan	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
13								
14								
15								
16	**First Quartile (Q1)**		226152	7917	1861.721074	1327	1.72	2923.41
17	**Third Quartile (Q3)**		317736	29087	3894.3091	4052	8.98	7572.18
18	**Interquartile Range (IQR)**		91584	21170	2032.588026	2725	7.26	4648.77
19	**Upper Fence**		455112	60842	6943.191139	8139.5	19.87	14545.335
20	**Lower Fence**		88776	-23838	-1187.16097	-2760.5	-9.17	-4049.745

from the above image, 917892 and 543132 are greater than 455112 in the Total sales column, which is the upper fence of the box and whisker, therefore they are considered as outliers. Although I will keep one of the Cities, that is Gillette since the dataset is small and will go ahead and remove the Cheyenne row and finally have 10 rows of dataset as shown below.

Figure 4. final dataset after removing the outlier observation.

Record	City	County	Total_Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	Johnson	185328	4585	3115.5075	746	1.55	1819.5
2	Casper	Natrona	317736	35316	3894.3091	7788	11.16	8756.32
3	Cody	Park	218376	9520	2998.95696	1403	1.82	3515.62
4	Douglas	Converse	208008	6120	1829.4651	832	1.46	1744.08
5	Evanston	Uinta	283824	12359	999.4971	1486	4.95	2712.64
6	Gillette	Campbell	543132	29087	2748.8529	4052	5.8	7189.43
7	Powell	Park	233928	6314	2673.57455	1251	1.62	3134.18
8	Riverton	Fremont	303264	10615	4796.859815	2680	2.34	5556.49
9	RockSprings	Sweetwater	253584	23036	6620.201916	4022	2.78	7572.18
10	Sheridan	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71