P3 Create an Analytical Dataset

Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

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. The aim of this project is to perform an analysis based on predicted yearly sales to determine the city for Pawdacity's newest store.

2. What data is needed to inform those decisions?

- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
- A partially parsed data file that can be used for population numbers.
- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

Building the Training Set

Dataset – Cleaning & Pre-processing:

Following are the datasets available for this project and the steps on how they are clean/format:

- p2-2010-pawdacity-monthly-sales.csv – contains monthly sales for all Pawdacity stores for year 2010.

NAME	ADDRESS	CITY	STATE	ZIP	January	February	March	April	May	June	July	August	September	October	November	December
Pawdacity	509 Fort St # A	Buffalo	WY	82834	16200	13392	14688	17064	18360	14040	12960	19224	15984	13392	13176	16848
Pawdacity	601 SE Wyoming Blvd Unit 252	Casper	WY	82609	29160	21600	27000	27648	29160	27216	25488	25704	22896	25272	28944	27648
Pawdacity	3769 E Lincolnway	Cheyenne	WY	82001	79920	70632	79056	77544	73656	77976	73872	77544	78516	74520	74736	79920
Pawdacity	2625 Big Horn Ave	Cody	WY	82414	19440	15984	19008	18144	16632	17496	18792	20304	19224	18144	18576	16632
Pawdacity	123 S 2nd St	Douglas	WY	82633	16200	13392	14688	17064	18360	14040	12960	19224	15984	29808	17496	18792
Pawdacity	932 Main St	Evanston	WY	82930	24840	21168	21600	22248	24192	24624	25488	25704	22032	21168	25920	24840
Pawdacity	200 E Lakeway Rd	Gillette	WY	82718	47520	41796	48384	47088	42336	41904	42120	47088	49032	48168	42984	44712
Pawdacity	180 S Bent St	Powell	WY	82435	20520	17928	20304	21168	21600	17928	18144	18576	20304	21168	17496	18792
Pawdacity	512 E Main St	Riverton	WY	82501	27000	22032	28512	26784	25920	24192	25056	22896	25488	26352	26784	22248
Pawdacity	2706 Commercial Way	Rock Springs	WY	82901	21600	19872	22248	20952	17496	24840	22464	21816	21384	20304	22032	18576
Pawdacity	1842 Sugarland Dr Ste 113	Sheridan	WY	82801	27000	26352	28080	22032	21168	29376	25920	20304	33696	23760	25056	25488

Figure 1 Pawdacity Monthly Sales

As we need the total sales per city for year 2010, we will need to format the data as it is currently shown as per month.

1. The data field type for the sales per month are in string format. Convert data field type to the correct ones using 'Auto Field'.
2. Using 'Transpose' function to transpose monthly sales columns into rows.

3. Rename the 2 new columns created from the transpose to 2 column to ‘Month’ and ‘Total_sales’
4. Using Summarize format, group the data by city and aggregate (sum) monthly sales into total sales for each city.

CITY	Sum_Total_Sales
Buffalo	185328
Casper	317736
Cheyenne	917892
Cody	218376
Douglas	208008
Evanston	283824
Gillette	543132
Powell	233928
Riverton	303264
Rock Springs	253584
Sheridan	308232

Figure 2 Aggregated Sales

- partially-parsed-wy-web-scrape.csv – contains population data crawled/collected

City|County	2014 Estimate	2010 Census	2000 Census
Afton|Lincoln	<td>1,968</td>	<td>1,911</td>	<td>1,818</td>
Albin|Laramie	<td>185</td>	<td>181</td>	<td>120</td>
Alpine|Lincoln	<td>845</td>	<td>828</td>	<td>550</td>
Baggs|Carbon	<td>439</td>	<td>440</td>	<td>348</td>
Bairoil|Sweetwater	<td>107</td>	<td>106</td>	<td>97</td>
Bar Nunn|Natrona	<td>2,735</td>	<td>2,213</td>	<td>936</td>
Basin ?|Big Horn	<td>1,312</td>	<td>1,285<sup id="cite_ref-4" class="referenc...	<td>1,238</td>
Bear River|Uinta	<td>521</td>	<td>518</td>	<td>-</td>
Big Piney|Sublette	<td>538</td>	<td>552</td>	<td>408</td>
Buffalo ?|Johnson	<td>4,615</td>	<td>4,585</td>	<td>3,900</td>
Burlington|Big Horn	<td>332</td>	<td>288</td>	<td>250</td>
Burns|Laramie	<td>305</td>	<td>301</td>	<td>285</td>
Byron|Big Horn	<td>609</td>	<td>593</td>	<td>557</td>
Casper ?|Natrona	<td>40,086</td>	<td>35,316</td>	<td>32,644</td>
Cheyenne ??|Laramie	<td>62,845</td>	<td>59,466</td>	<td>53,011</td>
Chugwater|Platte	<td>216</td>	<td>212</td>	<td>244</td>
Clearmont|Sheridan	<td>142</td>	<td>142</td>	<td>115</td>
Cody ?|Park	<td>9,740</td>	<td>9,520</td>	<td>8,835</td>
Cokeville|Lincoln	<td>542</td>	<td>535</td>	<td>506</td>
Cowley|Big Horn	<td>718</td>	<td>655</td>	<td>560</td>

Figure 3 Population Data

From the above dataset, we will need to extract the 2010 census column which is required for our analysis.

a) Split City|Country column into 2 columns using the Text to Column’ tool.
b) Use formula tool to clean html tags.
    - REGEX_Replace([2010 Census], "<.*?>", '')
    - REGEX_Replace([2010 Census], "\[.*?\]", '')
c) Using ‘Data Cleansing’ function to remove punctuation.
d) Using ‘Filter’ function to filter data to show only those that are not null value.
e) Using ‘Select’ function to convert 2010 Census column to integer data type and rename the two new columns to city and country.

2010 Census	City
1911	Afton
181	Albin
828	Alpine
440	Baggs
106	Bairoil
2213	Bar Nunn
1285	Basin
518	Bear River
552	Big Piney
4585	Buffalo
288	Burlington
301	Burns
593	Byron

Figure 4 2010 Census

- p2-wy-demographic-data - contains demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

City	County	Land Area	Households with Under 18	Population Density	Total Families
Laramie	Albany	2513.745235	2075	5.19	4668.93
Rock River	Albany	200.444	165	0.41	372.3
Basin	Big Horn	543.9513043	250	0.66	566.43
Burlington	Big Horn	137.6462142	63	0.17	143.34
Byron	Big Horn	252.4895917	116	0.31	262.93
Cowley	Big Horn	297.6806681	137	0.36	309.98
Deaver	Big Horn	76.28585366	35	0.09	79.44
Greybull	Big Horn	691.22612	318	0.84	719.8
Lovell	Big Horn	809.453936	372	0.98	842.91
Manderson	Big Horn	48.5078526	22	0.06	50.51
Gillette	Campbell	2748.8529	4052	5.8	7189.43
Wright	Campbell	262.0087853	386	0.55	685.27
Baggs	Carbon	253.2403224	62	0.06	129.53
Dixon	Carbon	55.95515096	14	0.01	28.62
Elk Mountain	Carbon	113.0640164	28	0.03	57.83
Grand Encampment	Carbon	255.5477513	63	0.06	130.71
Hanna	Carbon	479.3683551	118	0.12	245.19
Medicine Bow	Carbon	159.7894517	39	0.04	81.73
Rawlins	Carbon	5322.661628	1307	1.32	2722.43
Riverside	Carbon	30.573433	8	0.01	15.64
Saratoga	Carbon	976.0424271	240	0.24	499.23

Figure 5 Demographic Data

After cleaning up and blending the data together in the above steps. Below are the sum and average for the columns:

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*

Dealing with Outliers

Below shows the Interquartile Range calculation result for the dataset. Those highlighted in red below are identified as outliers.

City	Total Pawdacity Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185,328	4,585	3,115.51	746	1.55	1,819.50
Casper	317,736	35,316	3,894.31	7,788	11.16	8,756.32
Cheyenne	917,892	59,466	1,500.18	7,158	20.34	14,612.64
Cody	218,376	9,520	2,998.96	1,403	1.82	3,515.62
Douglas	208,008	6,120	1,829.47	832	1.46	1,744.08
Evanston	283,824	12,359	999.50	1,486	4.95	2,712.64
Gillette	543,132	29,087	2,748.85	4,052	5.80	7,189.43
Powell	233,928	6,314	2,673.57	1,251	1.62	3,134.18
Riverton	303,264	10,615	4,796.86	2,680	2.34	5,556.49
Rock Springs	253,584	23,036	6,620.20	4,022	2.78	7,572.18
Sheridan	308,232	17,444	1,893.98	2,646	8.98	6,039.71
Q1	226,152	7,917.00	1,861.72	1,327	1.72	2,923.41
Q3	312,984	26,061.50	3,504.91	4,037	7.39	7,380.81
IQR	86,832	18,144.50	1,643.19	2,710	5.67	4,457.40
Upper Fence	443,232	53,278	5,970	8,102	16	14,067
Lower Fence	95,904	-19,300	-603	-2,738	-7	-3,763

Figure 6 Interquartile Range Calculation

- **Cheyenne:** Even though Cheyenne is flagged out as outlier but it is a big city compared to the rest of the cities. Cheyenne's values for the different fields are larger in comparison with the other cities even though it have a much smaller store. The big numbers are in proportion with each other thus we can keep this for further analysis.

- **Gillette:** Comparing with the rest of the cities, Gillette's total sales are not in proportion with the other demographics fields such as population. If a city has large sales, we would expect it to have a big population to drive those sales, which isn't the case with Gillette. Hence, Gillette will be excluded from the dataset in order to build an unbiased regression model.

- **Rock Springs:** The outlier for this city is the land area. Since the value is only slightly out of range, we can keep this for further analysis.

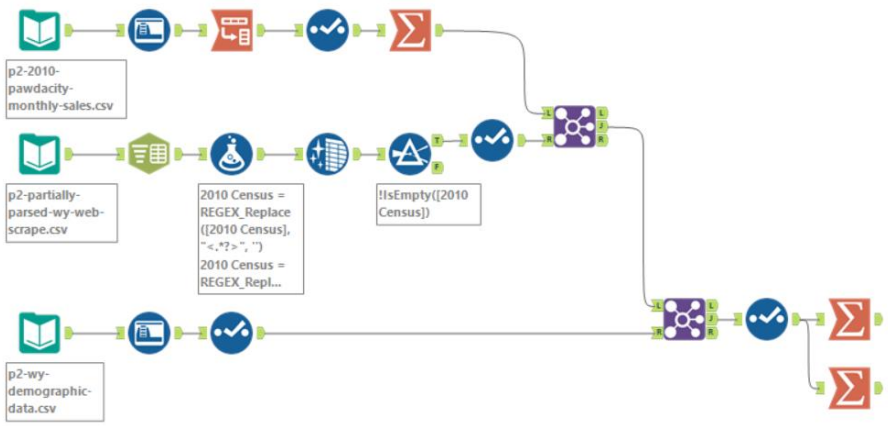

Figure 7 Part 1 Alteryx Workflow