# Project 2.1: Data Cleanup

# Step 1: Business and Data Understanding

#### **Key Decisions:**

Answer these questions

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

Predict the yearly sales and recommend a city for a new store based on the findings.

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

Total sales of the current stores in each city, population of the cities, population density and total families.

# 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.

Column	Sum	Average	
Census Population	213,862	19442	
Total Pawdacity Sales	3,773,304	343,027.64	
Households with Under 18	34,064	3097	
Land Area	33,071	3006.49	
Population Density	63	5.71	
Total Families	62,653	5695.71	

CITY	TotalSales2010	2010 Census	Land Area	Households with Under 18	Population Density	<b>Total Families</b>
Buffalo	185328	4585	3115.5075	746	1.55	1819.5
Casper	317736	35316	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Cody	218376	9520	2998.95696	1403	1.82	3515.62
Douglas	208008	6120	1829.4651	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.8529	4052	5.8	7189.43
Powell	233928	6314	2673.57455	1251	1.62	3134.18
Riverton	303264	10615	4796.859815	2680	2.34	5556.49
Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
	TotalSales2010	2010 Census	Land Area	Household with under 18	population density	Total families
Total	3773304.00	213862	33071.38	34064	62.80	62652.79
Average	343,027.64	19442.00	3006.49	3097	5.71	5695.71
	Buffalo Casper Cheyenne Cody Douglas Evanston Gillette Powell Riverton Rock Springs Sheridan	Buffalo 185328 Casper 317736 Cheyenne 917892 Cody 218376 Douglas 208008 Evanston 283824 Gillette 543132 Powell 233928 Riverton 303264 Rock Springs 253584 Sheridan 308232  Total 3773304.00	Buffalo         185328         4585           Casper         317736         35316           Cheyenne         917892         59466           Cody         218376         9520           Douglas         208008         6120           Evanston         283824         12359           Gillette         543132         29087           Powell         233928         6314           Riverton         303264         10615           Rock Springs         253584         23036           Sheridan         308232         17444           Total 3773304.00         213862	Buffalo         185328         4585         3115.5075           Casper         317736         35316         3894.3091           Cheyenne         917892         59466         1500.1784           Cody         218376         9520         2998.95696           Douglas         208008         6120         1829.4651           Evanston         283824         12359         999.4971           Gillette         543132         29087         2748.8529           Powell         233928         6314         2673.57455           Riverton         303264         10615         4796.859815           Rock Springs         253584         23036         6620.201916           Sheridan         308232         17444         1893.977048           Total         3773304.00         213862         33071.38	Buffalo         185328         4585         3115.5075         746           Casper         317736         35316         3894.3091         7788           Cheyenne         917892         59466         1500.1784         7158           Cody         218376         9520         2998.95696         1403           Douglas         208008         6120         1829.4651         832           Evanston         283824         12359         999.4971         1486           Gillette         543132         29087         2748.8529         4052           Powell         233928         6314         2673.57455         1251           Riverton         303264         10615         4796.859815         2680           Rock Springs         253584         23036         6620.201916         4022           Sheridan         308232         17444         1893.977048         2646           Total Sales2010         2010 Census         Land Area         Household with under 18           Total         3773304.00         213862         33071.38         34064	Buffalo         185328         4585         3115.5075         746         1.55           Casper         317736         35316         3894.3091         7788         11.16           Cheyenne         917892         59466         1500.1784         7158         20.34           Cody         218376         9520         2998.95696         1403         1.82           Douglas         208008         6120         1829.4651         832         1.46           Evanston         283824         12359         999.4971         1486         4.95           Gillette         543132         29087         2748.8529         4052         5.8           Powell         233928         6314         2673.57455         1251         1.62           Riverton         303264         10615         4796.859815         2680         2.34           Rock Springs         253584         23036         6620.201916         4022         2.78           Sheridan         308232         17444         1893.977048         2646         8.98           Total         3773304.00         213862         33071.38         34064         62.80

Figure 1: Totals and averages of the variables in the training set.

## Step 3: Dealing with Outliers

### Are there any cities that are outliers in the training set?

I used IQR method to determine the outliers. Anything above or under the Fence interval is considered an outlier.

There are three (3) outlier cities:-

- **Cheyenne** has outliers in TotalSales2010, 2010 Census, Population Density and Total Families
- **Gillette** has one outlier in TotalSales2010
- Rock Springs has one in Land Area

1	CITY	TotalSales2010	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
2	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
3	Casper	317736	35316	3894.3091	7788	11.16	8756.32
4	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
5	Cody	218376	9520	2998.95696	1403	1.82	3515.62
6	Douglas	208008	6120	1829.4651	832	1.46	1744.08
7	Evanston	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
9	Powell	233928	6314	2673.57455	1251	1.62	3134.18
10	Riverton	303264	10615	4796.85982	2680	2.34	5556.49
11	Rock Springs	253584	23036	6620.20192	4022	2.78	7572.18
12	Sheridan	308232	17444	1893.97705	2646	8.98	6039.71
13							
22		TotalSales2010	2010 Census	Land Area	Household with under 18	population density	Total families
23	Total	3773304.00	213862	33071.38	34064	62.80	62652.79
24	Average	343,027.64	19442.00	3006.49	3097	5.71	5695.71
25							
28	Q1	226152	7917	1861.72107	1327	1.72	2923.41
29	Q3	312984	26061.5	3504.9083	4037	7.39	7380.805
30	IQR	86832	18144.5	1643.18723	2710	5.67	4457.395
31	Upper Fence	443232	53278.25	5969.68914	8102	15.895	14066.8975
32	Lower Fence	95904	-19299.75	-603.05977	-2738	-6.785	-3762.6825

Figure 2: Outliers values are highlighted.

### Which outlier have you chosen to remove or impute?

I decided to remove **Gillette** city because the high sales are not reasonable compared to the other variables. Most of the other variables are not far from the average.

Unlike **Cheyenne**, even though it have more outliers than the other cities. However, the high density and the total families are way above the average which makes sense for the high sales.

# Alteryx work flow:-

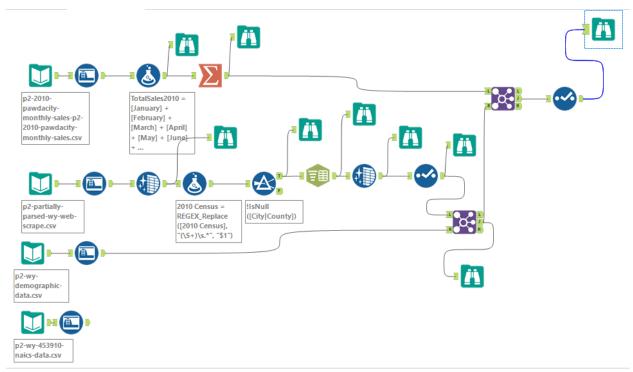


Figure 3: This work flow is used to clean up and blend the data.