Project 2: Create an Analytical Dataset

Step 1: 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 manager has asked me to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales. My first step in predicting yearly sales is to first format and blend data together from different datasets and deal with outliers.

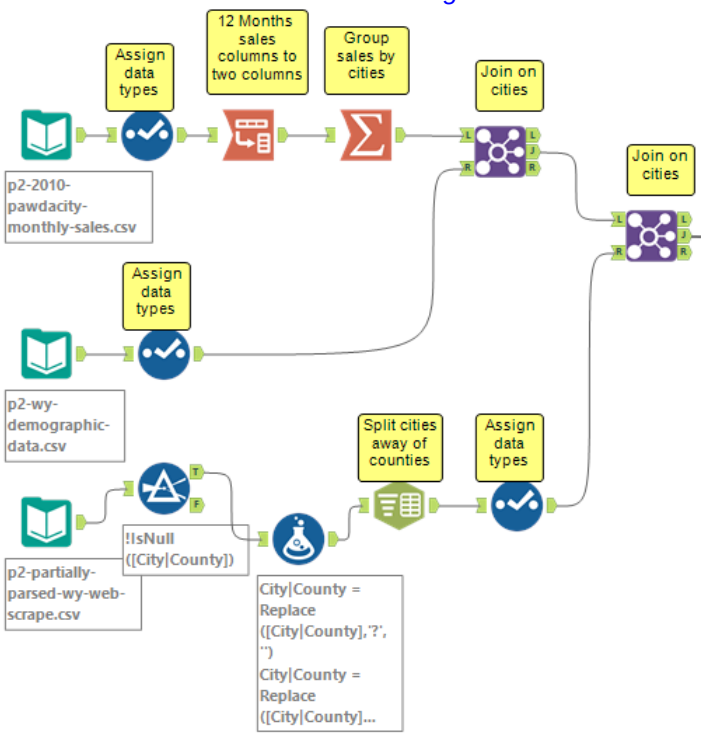
2. What data is needed to inform those decisions?

The following data files are available:

- *'p2-2010-pawdacity-monthly-sales.csv' represents the monthly sales data for all of the Pawdacity stores for the year 2010.*
- *'p2-partially-parsed-wy-web-scrape.csv' is a partially parsed data file that can be used for population numbers.*
- *'p2-wy-demographic-data.csv' contains the demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.*
- *'p2-wy-453910-naics-data.csv' contains NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.*

Step 2: Building the Training Set

I have used Alteryx to format and blend the three data files together as follows:



The final prepared data before analysis with target variable and all possible predictors variables are summarized as follows:

	Column	Sum	Average
Target Variable	*Total Pawdacity Sales*	*3,773,304*	*343,027.64*
All Possible Predictors Variables	*Census Population*	*213,862*	*19,442*
	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*

Step 3: Dealing with Outliers

To detect cities with outliers, I have calculated the lower and upper fence depending on IQR for each column. Then the following table identifies cities with outliers.

CITY	Total Pawdacity Sales	Census Population	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	185,328	4,585	746	3,115.51	1.55	1,819.50
Casper	317,736	35,316	7,788	3,894.31	11.16	8,756.32
Cheyenne	917,892	59,466	7,158	1,500.18	20.34	14,612.64
Cody	218,376	9,520	1,403	2,998.96	1.82	3,515.62
Douglas	208,008	6,120	832	1,829.47	1.46	1,744.08
Evanston	283,824	12,359	1,486	999.50	4.95	2,712.64
Gillette	543,132	29,087	4,052	2,748.85	5.80	7,189.43
Powell	233,928	6,314	1,251	2,673.57	1.62	3,134.18
Riverton	303,264	10,615	2,680	4,796.86	2.34	5,556.49
Rock Springs	253,584	23,036	4,022	6,620.20	2.78	7,572.18
Sheridan	308,232	17,444	2,646	1,893.98	8.98	6,039.71

Note: cells with orange shading color indicates outlier values.

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

   Yes, there is at least one city with outlier value for each variable except 'Households with Under 18' variable.

   - *With 'Total Pawdacity Sales' variable, the two cities **Cheyenne** and **Gillette** have extremely high sales volumes compared with other cities.*
   - *With 'Census Population' variable, the **Cheyenne** City has an extremely big population compared with other cities.*
   - *With 'Land Area' variable, the **Rock Springs** City has an extremely big land area compared with other cities.*
   - *With 'Population Density' variable, the **Cheyenne** City has an extremely great population density compared with other cities.*
   - *With 'Total families' variable, the **Cheyenne** City has an extremely big total families compared with other cities.*

2. Which outlier have I chosen to remove or impute?

After detailed observation of all the outliers of the data, I can conclude that it doesn't seem like there is any typo error and all the data seem to be correct. We have three cities causing outliers:

- ***Cheyenne*** *city: it is clearly a very specific case in comparison with other cities. The city has extreme high population, density, and sales at the same time, even though it is one of the smallest cities of the state. This city is causing inconsistency and biasness in our model accordingly, this city should be removed from the dataset.*
- ***Gillette*** *city: it has extreme high sales but this is reasonable because this city one of top four cities in population, household under 18, population density, and total families. We should keep it.*
- ***Rock Springs*** *city: it has extreme big land area otherwise it has a consistent data which is important for our model to predict sales, and if we remove this city, we will lose a good information to our model since our sample size is very small. On the other hand, land area has a weak impact on sales model so that we can't take the risk of losing model quality we should keep such city.*

So now we are ready for modeling.

Notes:

Dealing with outliers is variable according to the person's point of view. In particular:

- *For* ***Cheyenne****: we can justify either keeping or removing Cheyenne. we could justify removing it because it's unlike other cities for all fields. we could justify keeping it because it's inline with the linear relationship.*

- *For* ***Gillette****: we can justify either keeping or removing Gillette. we could justify removing it because it skews high in sales, yet does not skew relative to the other data fields in the training set. we should justify keeping it because the dataset is small and it is only an outlier in one field.*