Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

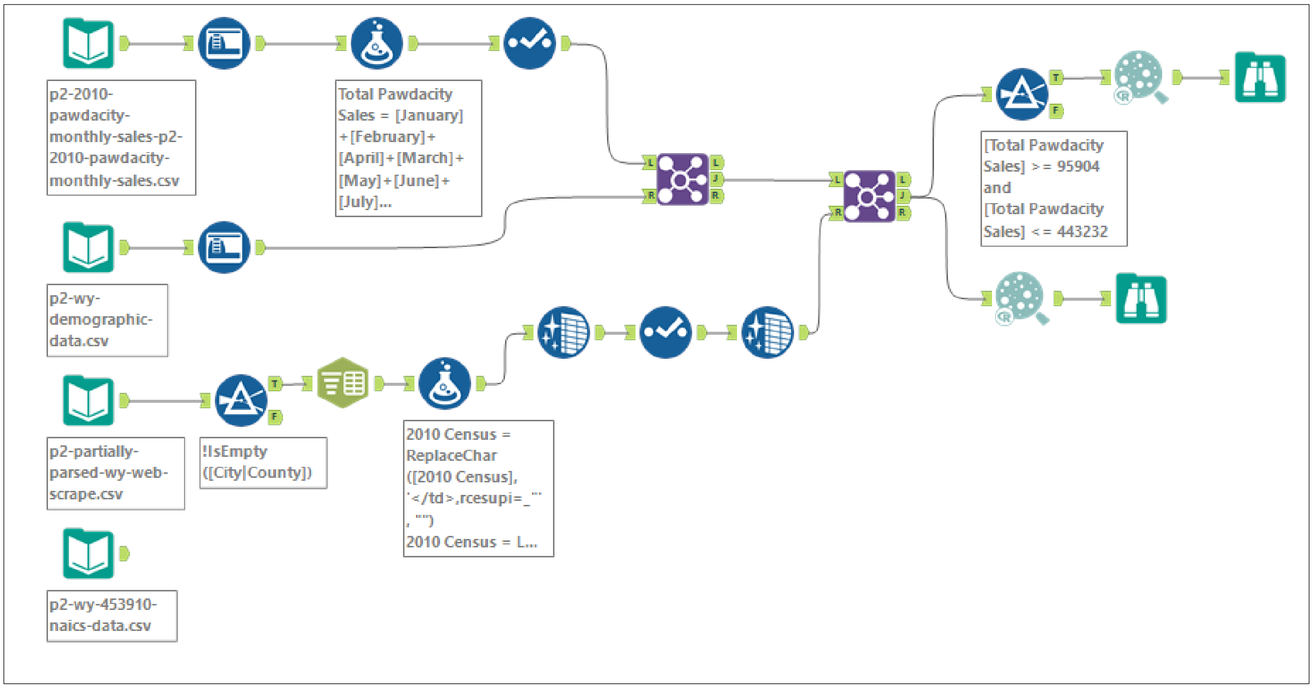
Pawdacity would like to expand and open a 14th store and therefore and analysis has to be performed in order to recommend the city for Pawdacity's newest store, based on predicted yearly sales. We first have to first format and blend together data from different datasets and deal with outliers so we can predict yearly sales and recommend the city for the new store.

2. The following data is needed to inform the decisions:
    - The monthly sales data for all of the Pawdacity stores for the year 2010.
    - A partially parsed data file that will be used for population numbers.
    - Demographic data for each city and county in the state of Wyoming. The demographic data includes the following attributes: Households with individuals under 18, Land Area, Population Density, and Total Families

Step 2: Building the Training Set

- The given datasets are loaded into Alteryx using the Input Data Tool and the Autofield Tool is used to convert the data fields into the appropriate data types.
- Using the Formula Tool, we sum the 2010 monthly sales for all Pawdacity stores from the *p2-2010-pawdacity-monthly-sales.csv* dataset to get the 2010 total sales for all Pawdacity stores and name the field **Total_Pawdacity_Sales**. We connect the Select Tool to select only the two needed fields: CITY and Total Pawdacity Sales. Then connect the Join Tool to join this data to the *p2-wy-demographic-data.csv* dataset and the resulting dataset contains six fields, each containing eleven records. Let's call this dataset ***sales-and-demographic-data***.
- Next we retrieve the 2010 Census Population from the *p2-partially-parsed-wy-web-scrape.csv* dataset. We use the Filter Tool to remove nulls in the City|County field and then use the Text to Column Tool to split the City|County into two separate fields, City and County. Then use the Formula Tool to remove the extra characters in the 2010 Census field. Then we connect the Data Cleansing Tool to remove the remaining trailing spaces, letters and punctuations in the 2010 Census Field. Lastly, we use the Select Tool to select the two fields we need which are the 2010 Census and the City fields.
- Using the Join Tool, we join the cleaned *p2-partially-parsed-wy-web-scrape.csv* data with the *sales-and-demographic-data* dataset and the resulting dataset contains seven fields, each with eleven records.

The Alteryx workflow is shown below:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-453910-naics-data.csv
L
J
R
T
F

Total Pawdacity Sales = [January]+[February]+[April]+[March]+[May]+[June]+[July]...

!IsEmpty([City|County])

2010 Census = ReplaceChar([2010 Census], '</td>,rcesupi=_"' , "")
2010 Census = L...

[Total Pawdacity Sales] >= 95904 and [Total Pawdacity Sales] <= 443232

The sums and averages are given in the table below:

Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,028*
Households with Under 18	*34,064*	*3,097*
Land Area	*33,071*	*3,006*
Population Density	*63*	*6*
Total Families	*62,653*	*5,696*

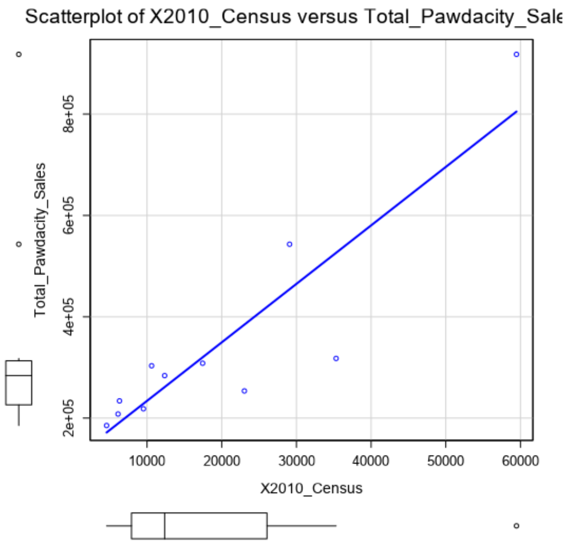
Step 3: Dealing with Outliers

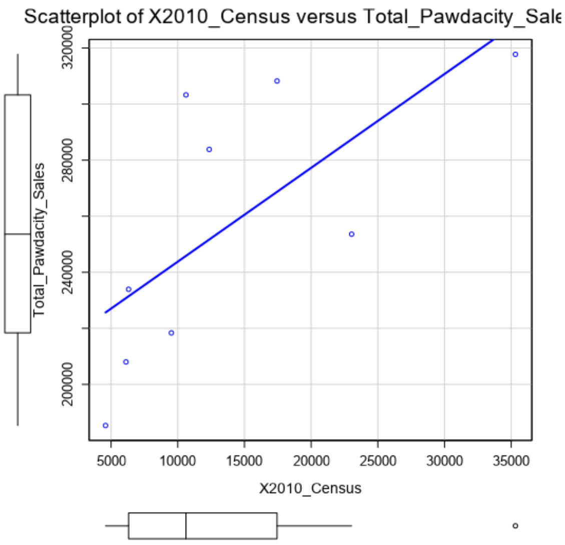
To check for outliers, I used the **Excel** to calculate for Quartile Q1, Quartile Q1, IQR = Q3 - Q1, Upper Fence = Q3 + 1.5 IQR, Lower Fence = Q1 - 1.5 IQR and Median

	2010 Census Population	**Total Pawdacity Sales**	**Households with Under 18**	**Land Area**	**Population Density**	**Total Families**
Q1	7,917	226,152	1,327	1,861.72	1.72	2,923.41
Q3	2,6061.50	312,984	4,037	3,504.91	7.39	7,380.81
IQR	18,144.50	86,832	2,710	1,643.19	5.67	4,457.40
Upper Fence	53,278.25	443,232	8,102	5,969.69	15.90	14,066.90
Lower Fence	-19,299.75	95,904	-2,738	-603.06	-6.79	-3,762.68
Median	14,901.50	293,544	2,663	2,873.90	3.87	5,798.10

Using the Filter Tool, I check for values that are below the Lower Fences and above the Upper Fences. The cities **Gillette** and **Cheyenne** are returned as the results, meaning we have two outliers in the dataset. Considering the scatterplots below (plotted from the fields *2010 Censes and Total Pawdacity Sales),* a positive linear relationship is observed on both scatterplots and therefore we decide to keep one outlier city, Gillette because it's inclusion does not affect the relation. As for Cheyenne, it is observed to be extremely far away from the upper fence, and therefore we remove it from the dataset.

Note: The scatterplot on the left is plotted with all the 11 cities whereas the scatterplot on the right is plotted after removing the outlier cities.


Scatterplot of X2010_Census versus Total_Pawdacity_Sale
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
X2010_Census


Scatterplot of X2010_Census versus Total_Pawdacity_Sale
Total_Pawdacity_Sales
X2010_Census