The Business Problem

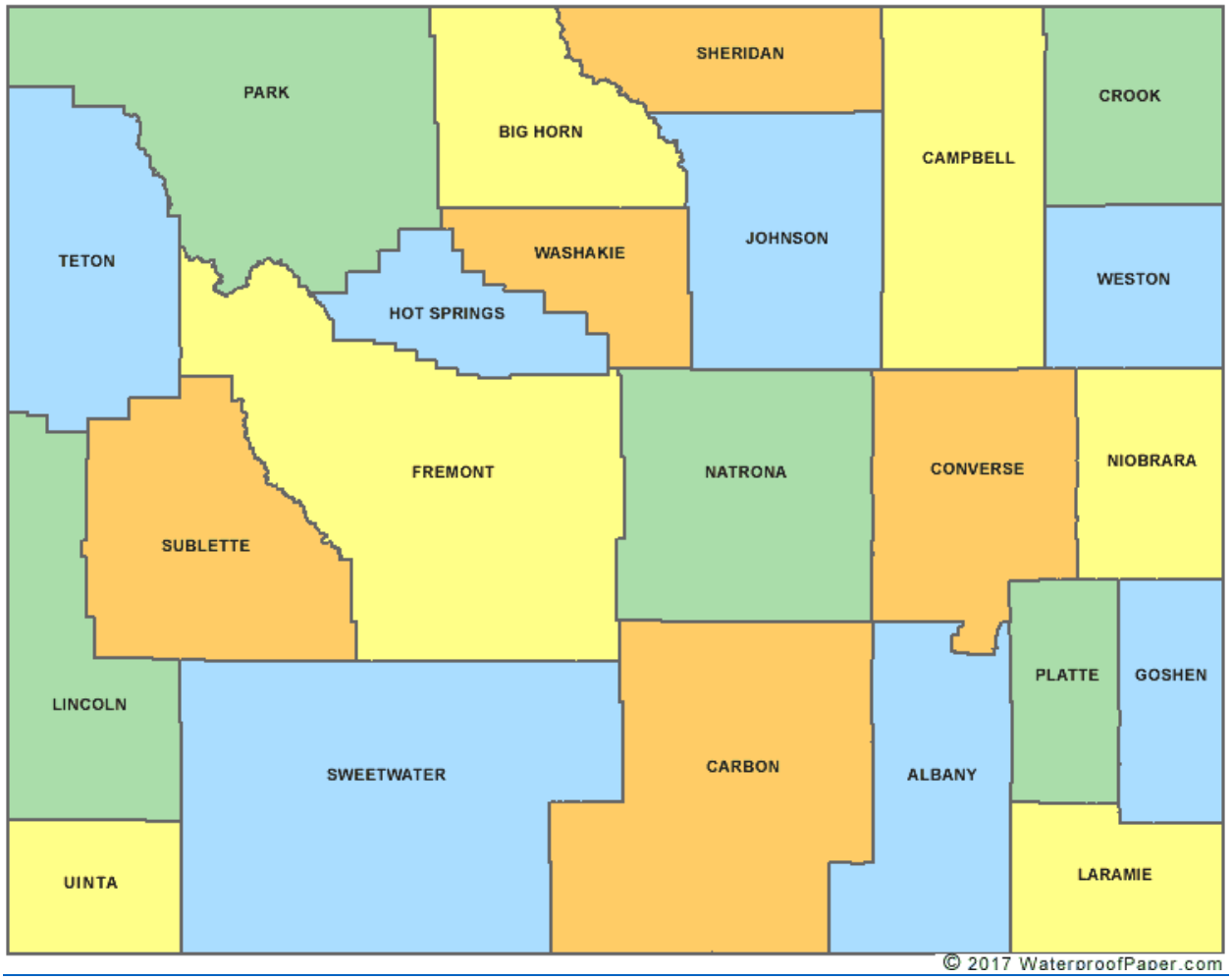
Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Map of Wyoming Counties



Steps to Success

Step 1: Business and Data Understanding

Your project should include a description of the key business decisions that need to be made.

Step 2: Building the Training Set

To properly build the model, and select predictor variables, create a dataset with the following columns:

City
2010 Census Population
Total Pawdacity Sales
Households with Under 18
Land Area
Population Density
Total Families

This dataset will be your training set to help you build a regression model in order to predict sales in the Practice Project in the next lesson. Every row should have sales data because we're trying to predict sales.

Notes

You should be consolidating the data at the city level and **not at the store level.** We only have data at the city wide level so any analysis at the store level will not be sufficient to complete this analysis.

We simply need to focus on cleaning up and blending the data together in this step.

If you've done everything correctly, the sum for each of the above columns should be:

- **Census Population:** 213,862
- **Total Pawdacity Sales:** 3,773,304
- **Households with Under 18:** 34,064
- **Land Area:** 33,071
- **Population Density:** 63
- **Total Families:** 62,653

with **11 rows of data**

For Alteryx users:

- Use the Autofield Tool to help quickly convert your data fields into the appropriate datafields for analysis.
- Research these three specific formulas to help you get rid of unwanted characters in the Formula tool: ReplaceFirst, Left, FindString

Step 3: Dealing with Outliers

Once you have created the dataset, look for outliers and figure out how deal with your outliers. Use the IQR method to determine if there are outlier cities for each of the variables and then justify which city that has at least one outlier value should be removed.

Data

p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.

p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.

p2-wy-453910-naics-data.csv - NAICS data on the sales of all competitor stores where total sales is equal to 12 months of sales

p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming