**Project Overview**

This project is the first part of a two-part series. In the first part, you will blend and format data and deal with outliers.

For the second part, you will use your cleaned up dataset to create another linear regression model. The difference this time is that you will have to choose which variable(s) are the most important for the model using new techniques learned in the Selecting Predictor Variables section.

**Scenario**

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.

**How Do I Complete this Project?**

This project uses skills learned throughout the "Data Preparation” lessons. To complete this project:

* Go through the course.
* Apply the skills learned in the course to solve the business problem given in the project details.
* Use our guidelines and rubric to help build your project.
* When you're ready, submit it to us for review using the submission template found in the supporting materials section.

**Skills Required**

In order to complete this project, you must be able to:

* Understand different data types. Review Lesson 1 Understanding Data
* Deal with a variety of data issues. Review Lesson 2 Data Issues
* Format data appropriately. Review Lesson 3 Data Formatting
* Blend data together using joins and unions. Review Lesson 4 Data Blending

**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**

A close up of a map

Description automatically generated

**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](http://help.alteryx.com/10.6/index.htm#Reference/Functions.htm#String) 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.

**IQR Steps**

To calculate the upper fence and the lower fence, here are the exact steps:

1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. You can use the Excel function QUARTILE.INC or QUARTILE.EXC

2 . Calculate the Interquartile Range: IQR = Q3 - Q1

3 . Add 1.5 *IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5*IQR

4 . Subtract 1.5 *IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5*IQR

5 . Values above the Upper Fence and values below the Lower Fence are outliers

**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.

**To Download Files below please right click on the link and select "Save Link As"**

The Competitor Sales data file (p2-wy-453910-naics-data.csv) **should not be used in this project** it will be used in the follow up practice project.

Supporting Materials

[**Competitor Sales**](http://video.udacity-data.com.s3.amazonaws.com/topher/2016/July/578ec224_p2-wy-453910-naics-data/p2-wy-453910-naics-data.csv)

[**Demographic Data**](http://video.udacity-data.com.s3.amazonaws.com/topher/2016/November/58250a54_p2-wy-demographic-data/p2-wy-demographic-data.csv)

[**Population Data**](http://video.udacity-data.com.s3.amazonaws.com/topher/2016/November/58250a7b_p2-partially-parsed-wy-web-scrape/p2-partially-parsed-wy-web-scrape.csv)

[**Pawdacity Monthly Sales**](https://d17h27t6h515a5.cloudfront.net/topher/2017/November/5a2086fc_p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales/p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv)

[**All Project Files**](https://d17h27t6h515a5.cloudfront.net/topher/2017/November/5a208765_p2-files/p2-files.zip)

[**P2.1 Submission Template**](https://d17h27t6h515a5.cloudfront.net/topher/2018/January/5a4e9f92_p2.1-submission-template-make-a-copy/p2.1-submission-template-make-a-copy.docx)

[**Debugging Alteryx**](https://video.udacity-data.com/topher/2018/October/5bbfb1ac_debugging-alteryx/debugging-alteryx.pdf)

# Outliers

Use the IQR method for each attribute (Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, Total Families) of the cities to see which cities have outlier values for any of these.