Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

1. What decisions needs to be made?

   The first thing I did was read each table and assigned them to the following variables:
   - monthlySales = *p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv*
   - webScrape = *p2-partially-parsed-wy-web-scrape.csv*
   - naics = *p2-wy-453910-naics-data.csv*
     - 12 month total sales cycle for competitor stores
   - demographic = *p2-wy-demographic-data.csv*
     - additional sector and population data

   Then I looked at each data set to determine which data sets needed to be cleaned using pandas *.info()* top-level method. As indicated in the project details the only data set that needed to be addressed was the partially parsed webScrape dataframe.

   I dropped all rows that contained null values then created a list of characters to be removed from the data set and replaced those characters with empty strings. After isolating the city into its own field I was able to merge all desired fields based on each corresponding *City* column; plus run statistics, copy, reshape, and build visualizations on the newly merged dataframe.

2. What data is needed to inform those decisions?

   As stated above, I extracted cities from webScrape field *City|County*. This allowed me to begin merging datasets two tables at a time. After the first table was merged, I continually merged each udacity provided table with the previously merged data. I chose outer join because I wanted to include all data from both tables.
   - monthlySales: This is the 2010 monthly sales for stores
   - webScrape: Population
   - Naics: 12 month total sales cycle for competitor stores
   - Demographic: additional store location data and population data

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. In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

The following snapshot references data found in file *Project 3 - Create an Analytical Dataset.ipynb*, located under section *4 Analyze Data*. As you can see the sum values in this table match up with the values supplied by Udacity in the *Project Details* section.

	Column	Sum	Average
0	2010 Census	213862.0	19442.00
1	Total Sales	3773304.0	343027.64
2	Households with Under 18	34064.0	3096.73
3	Land Area	33071.0	3006.49
4	Population Density	63.0	5.71
5	Total Families	62653.0	5695.71

Step 3: Dealing with Outliers

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

I was able to use scipy.stats iqr function to calculate the interquartile range (i.e. the difference between 3rd and 1st quartile) for each numeric column in the test dataset. Then I was able to calculate lower and upper bounds to search through the dataset for outliers. There were four cities that were flagged with outlier columns: Casper, Cheyenne, Gillette, and Rock Springs as shown in the image below.

	City	Outlier Feature	Outlier Value	Lower Bound	Upper Bound
0	Casper	Households with Under 18	7788.000000	-2043.318182	6556.863636
1	Cheyenne	2010 Census	59466.000000	-13935.312500	42116.187500
2	Cheyenne	Total Sales	917892.000000	35472.721766	491508.221486
3	Cheyenne	Households with Under 18	7158.000000	-2043.318182	6556.863636
4	Cheyenne	Population Density	20.340000	-6.151250	16.958750
5	Cheyenne	Total Families	14612.640000	-2604.029830	11349.589716
6	Gillette	Total Sales	543132.000000	35472.721766	491508.221486
7	Rock Springs	Land Area	6620.201916	-1039.217050	5633.280501

Casper and Rock Springs:

Casper and Rock Springs contain outliers, *Households with Under 18* and *Land Area* respectively. Rock Springs has low population density and low total sales, while Casper has high population density and low total sales. I did consider removing Casper from the dataset for not following a common trend of revenue being a factor of population size. But, Casper's Total Sales still fall within an acceptable range between its upper and lower bounds. I cannot find any reason to remove either from the dataset.

Gillette:

Gillette has a low population density and its Pawdacity brick-and-mortar Total Sales has been flagged as an outlier. My first thought was maybe the anomaly is due to an improper data entry, but Gillette shows consistent sales each month, with an average of $45,261; and a standard deviation of approximately $2,882. When Gillette's total sales is compared to its competitor stores, it is difficult to justify volume being too high based on population density, because most of the stores in Gillette (i.e. Pawdacity, and competitors) have a total sales range of a quarter-million to half-million dollars. However, this may be due to a skew in demographics (e.g. Gillette could have a largely affluent population). Since common trends suggests cities with smaller population density have lower total sales, while cities with high population density have higher total sales Gillette is being removed from the training set.

Gillette Pawdacity Total Sales:

	City	January	February	March	April	May	June	July	August	September	October	November	December	Total Sales
6	Gillette	47520	41796	48384	47088	42336	41904	42120	47088	49032	48168	42984	44712	543132

Gillette Competitors Total Sales:

	BUSINESS NAME	City	SALES VOLUME
20	All Gods Creatures	Gillette	450000
21	Camelot Pet Castle	Gillette	230000
22	Joes Pet Depot	Gillette	0
23	Pet Food Outlet	Gillette	450000

Cheyenne:

Although Cheyenne contains multiple outliers (e.g. *Total Sales*, *2010 Census*, etc.), I do not see any anomalies within the data entries. It is possible that the values entered for Cheyenne are accurate. In terms of *2010 Census*, Cheyenne had a population of 55,286 in 2005, and its population in 2019 is 63,624. Since a population of 59,466 in 2010 falls between 2005 and 2019 values, the data suggests this entry could be accurate. Regarding *Total Sales*, consider the number of families Cheyenne supports compared to other cities listed in the dataset. Although, the dataset does not provide information on how many families have pets in each city, we can look at *Total Families* as a factor of *Total Sales* (i.e. the number of families in a city could impact each store location's revenue). By dividing *Total Sales* by *Total Families* for each city we see that Pawdacity store located in Cheyenne makes less per family count when compared to other Wyoming cities; indicating that although the overall value is higher than other cities, it is low when compared per capita (or by family count in this case).

	City	Total Sales	Total Families	Annual Store Revenue Per Family Count
4	Douglas	208008	1744.08	119.0
5	Evanston	283824	2712.64	105.0
0	Buffalo	185328	1819.50	102.0
6	Gillette	543132	7189.43	76.0
7	Powell	233928	3134.18	75.0
2	Cheyenne	917892	14612.64	63.0
3	Cody	218376	3515.62	62.0
8	Riverton	303264	5556.49	55.0
10	Sheridan	308232	6039.71	51.0
1	Casper	317736	8756.32	36.0
9	Rock Springs	253584	7572.18	33.0

Let's look at what would happen if Cheyenne were to be removed from the dataset. The figure below looks at correlation between *Total Sales* and all other variables. Blue represents the data if Cheyenne were kept in the training set, and red represents Cheyenne being removed. Hint, it is easy to spot Cheyenne, since it is the only visible blue datapoint in each graph. Notice the slope of the line of best fit changes when Cheyenne is removed from the data set.

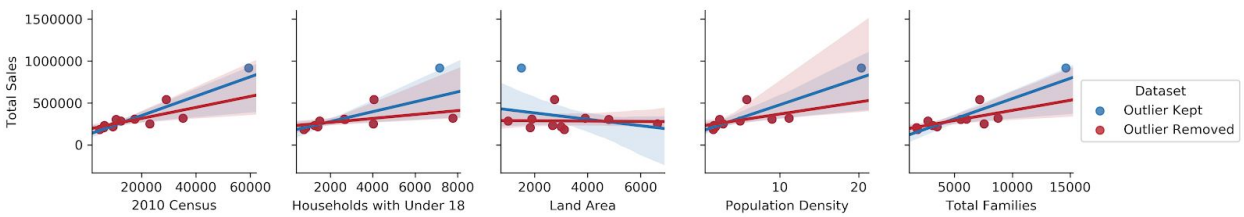

Figure location: *assets* > *figure.png*

I cannot find a reasonable argument to remove Cheyenne based on outlier values (i.e. incorrect data entry, or comparing observed values against other factors) I can only conclude it's possible that the data entries are valid. Also, if Cheyenne were to be removed from the dataset it would change the slope of the regression line as indicated in the figure above. Therefore, Cheyenne should remain in the training set. The final dataset was stored as a csv file. It can be located by starting at the main folder, then navigating to *data* > *trainingSet.csv*.