Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

Cleaned up ,blend together data from different datasets and deal with outlier. Format the dataset to create linear regression model.

After looking into data, we got to know as mentioned below:

p2-2010-pawdacity-monthly-sales.csv , p2-wy-demographic-data.csv : No cleansing operation required. Data has to combine group by city

p2-partially-parsed-wy-web-scrape.csv : As it is web scrape data, data cleansing operation are required like handling null records, removing unwanted character.

Key Decisions:

Answer these questions

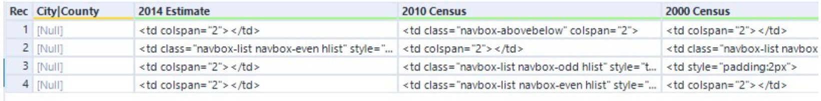
1. What decisions needs to be made?

   - **Select appropriate data type based upon field.**
     Autofield & select tool is used to select appropriate data type

   - **Split column if data is present single column separated by delimeter**
     City|County is split into 2 columns by using delimeter '|'

   - **Remove the unwanted character (dirty data) if present.**
     Html tag has been removed with help of regex tool

   - **Removal of null record if it doesn't contain any significant data or imputation is not possible.**
     4 null records has been removed from p2-partially-parsed-wy-web-scrape.csv as there is NO imputation possible and doesn't any significant data.



Rec	City|County	2014 Estimate	2010 Census	2000 Census
1	[Null]	<td colspan="2"></td>	<td class="navbox-abovebelow" colspan="2">	<td colspan="2"></td>
2	[Null]	<td class="navbox-list navbox-even hlist" style="...	<td colspan="2"></td>	<td class="navbox-list navbox
3	[Null]	<td colspan="2"></td>	<td class="navbox-list navbox-odd hlist" style="t...	<td style="padding:2px">
4	[Null]	<td colspan="2"></td>	<td class="navbox-list navbox-even hlist" style="...	<td colspan="2"></td>

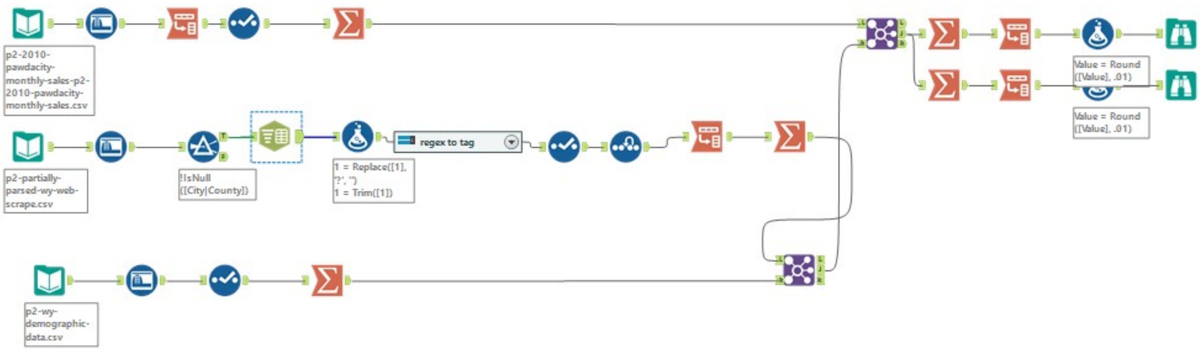
   - **Find out the outlier if present? If yes, observe the how many records are there? Make decision of remove or imputation based upon volume of data.**

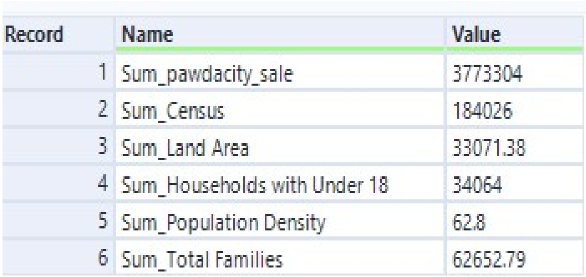
2. What data is needed to inform those decisions?
   We need city data and related to city I,e population and it density, monthly sale, land area, number of family, Households with Under 18.

Step 2: Building the Training Set

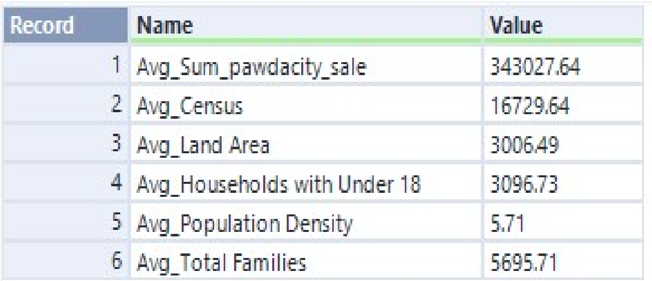
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

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


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
!IsNull ([City|County])
1 = Replace([1], '?', '')
1 = Trim([1])
regex to tag
p2-wy-demographic-data.csv
Value = Round ([Value], .01)
Value = Round ([Value], .01)



Record	Name	Value
1	Sum_pawdacity_sale	3773304
2	Sum_Census	184026
3	Sum_Land Area	33071.38
4	Sum_Households with Under 18	34064
5	Sum_Population Density	62.8
6	Sum_Total Families	62652.79



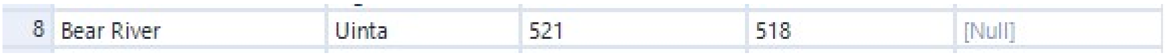
Record	Name	Value
1	Avg_Sum_pawdacity_sale	343027.64
2	Avg_Census	16729.64
3	Avg_Land Area	3006.49
4	Avg_Households with Under 18	3096.73
5	Avg_Population Density	5.71
6	Avg_Total Families	5695.71

Column	**Sum**	**Average**
Census Population	*213,862*	*16729.64*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

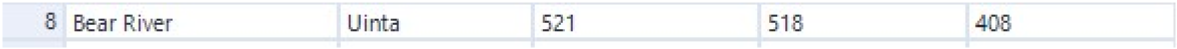
Answer these questions

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.



8	Bear River	Uinta	521	518	[Null]

I have found outliers in p2-partially-parsed-wy-web-scrape.csv. As it is small data set, I have decided to impute outlier by replace value with mode value instead of removing outlier.



8	Bear River	Uinta	521	518	408

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.