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

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

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   - ➔ We need to predict the yearly sales of Pawdacity for different cities.
   - ➔ In which city they should open their new store.

2. What data is needed to inform those decisions?
   - ➔ First of all ,we need the monthly sales for all the cities so that we can calculate the yearly sales of each city.The data should be of the most recent year.
   - ➔ We need population records of the cities.
   - ➔ We need demographic data which encompass population numbers and some more specific data (e.g. land area, population density and so on) for different counties and cities in the state.

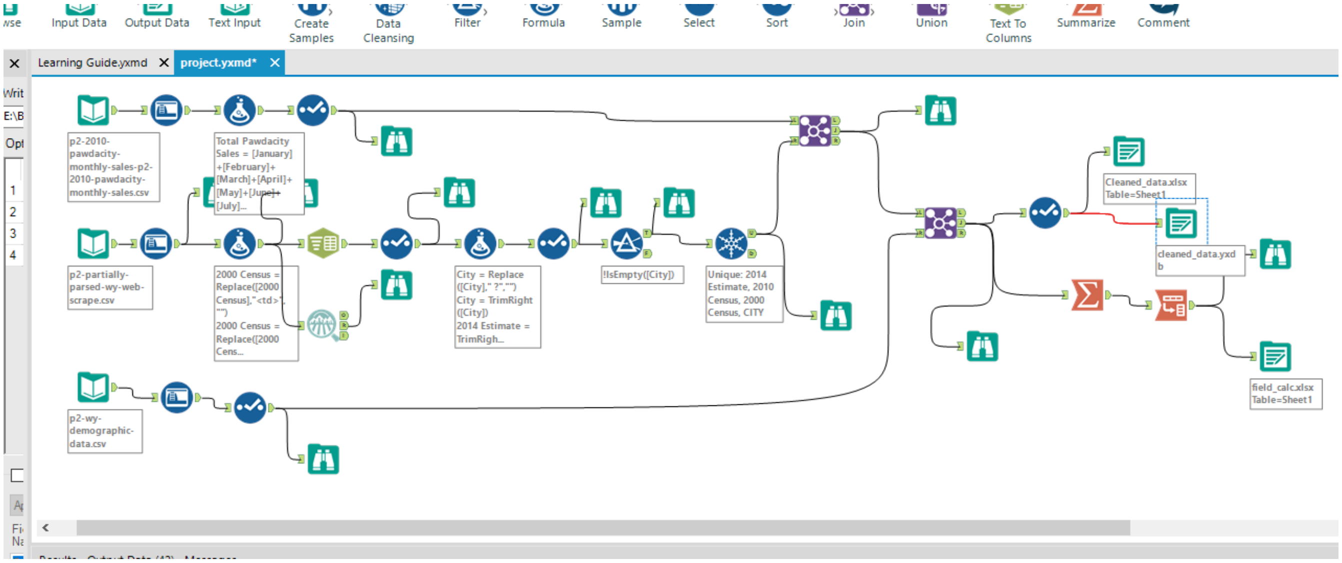
Step 2: Building the Training Set

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

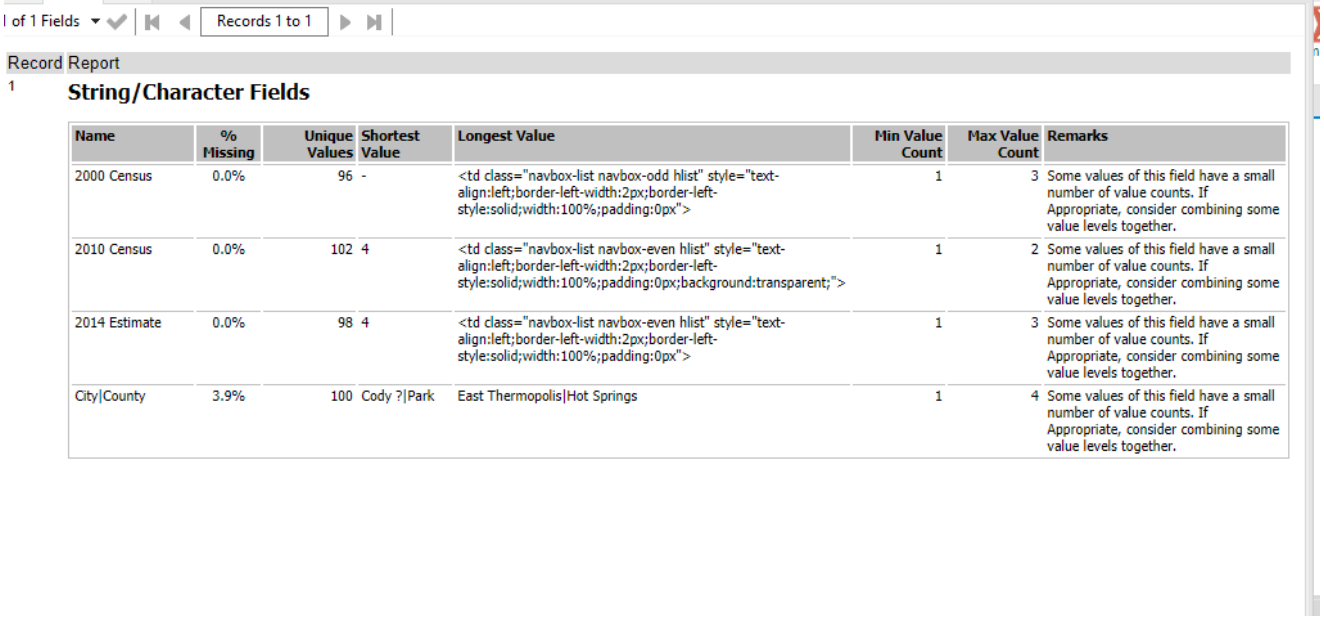
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	**Sum**	**Average**
Census Population	*213,862*	19,442
Total Pawdacity Sales	*3,773,304*	343,027.63
Households with Under 18	*34,064*	3,096.73
Land Area	*33,071*	3,006.49
Population Density	*63*	5.70
Total Families	*62,653*	5,695.70

Justification:- I have drawn the following workflow for the data wrangling process.


wse
Input Data
Output Data
Text Input
Create Samples
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Comment
Learning Guide.yxmd
project.yxmd*
Writ
E:\B
Opt
1
2
3
4
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
2000 Census = Replace([2000 Census],"<td>","")
2000 Census = Replace([2000 Cens...
City = Replace([City]," ?","")
City = TrimRight([City])
2014 Estimate = TrimRigh...
!IsEmpty([City])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
p2-wy-demographic-data.csv
Cleaned_data.xlsx Table=Sheet1
cleaned_data.yxdb
field_calc.xlsx Table=Sheet1
Fi
Na

Data Cleaning :- Field Summary showing % of null values in the given data set.



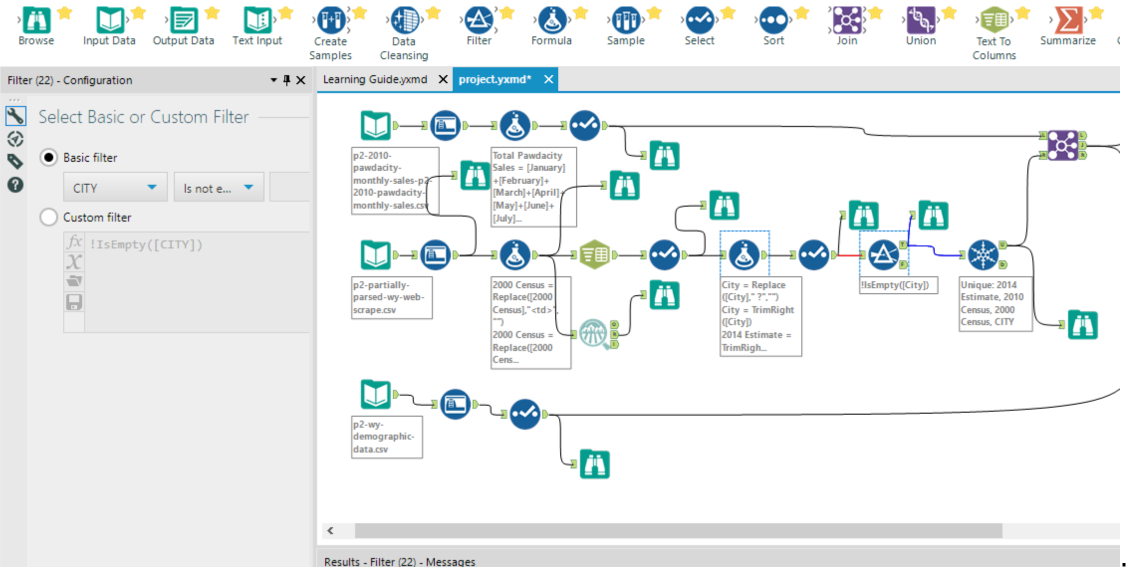
Record Report

1

String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
2000 Census	0.0%	96	-	<td class="navbox-list navbox-odd hlist" style="text-align:left;border-left-width:2px;border-left-style:solid;width:100%;padding:0px">	1	3	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
2010 Census	0.0%	102	4	<td class="navbox-list navbox-even hlist" style="text-align:left;border-left-width:2px;border-left-style:solid;width:100%;padding:0px;background:transparent;">	1	2	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
2014 Estimate	0.0%	98	4	<td class="navbox-list navbox-even hlist" style="text-align:left;border-left-width:2px;border-left-style:solid;width:100%;padding:0px">	1	3	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
City|County	3.9%	100	Cody ?|Park	East Thermopolis|Hot Springs	1	4	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

Used the not null function to drop the rows with null values.


Browse
Input Data
Output Data
Text Input
Create Samples
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Learning Guide.yxmd
project.yxmd*
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
2000 Census = Replace([2000 Census],"<td>","")
2000 Census = Replace([2000 Cens...
City = Replace ([City]," ?","")
City = TrimRight ([City])
2014 Estimate = TrimRigh...
!IsEmpty([City])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
p2-wy-demographic-data.csv
Results - Filter (22) - Messages

Filter (22) - Configuration

Select Basic or Custom Filter

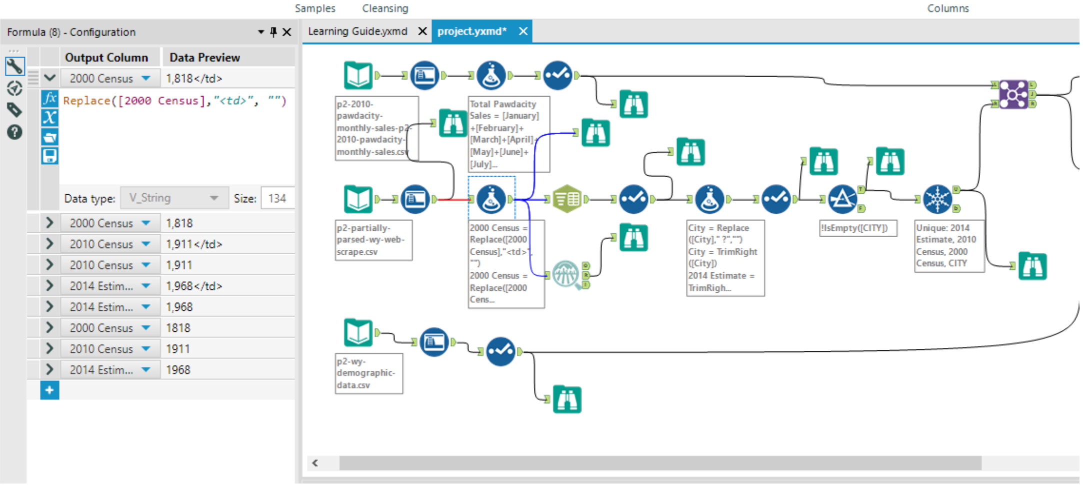
● Basic filter

CITY | Is not e...

○ Custom filter

```
!IsEmpty([CITY])
```

Data Formatting :- Used the string functions to format the data in the ill-formatted rows.


Samples
Cleansing
Columns
Learning Guide.yxmd
project.yxmd*
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
2000 Census = Replace([2000 Census],"<td>","")
2000 Census = Replace([2000 Cens...
City = Replace ([City]," ?","")
City = TrimRight ([City])
2014 Estimate = TrimRigh...
!IsEmpty([CITY])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
p2-wy-demographic-data.csv

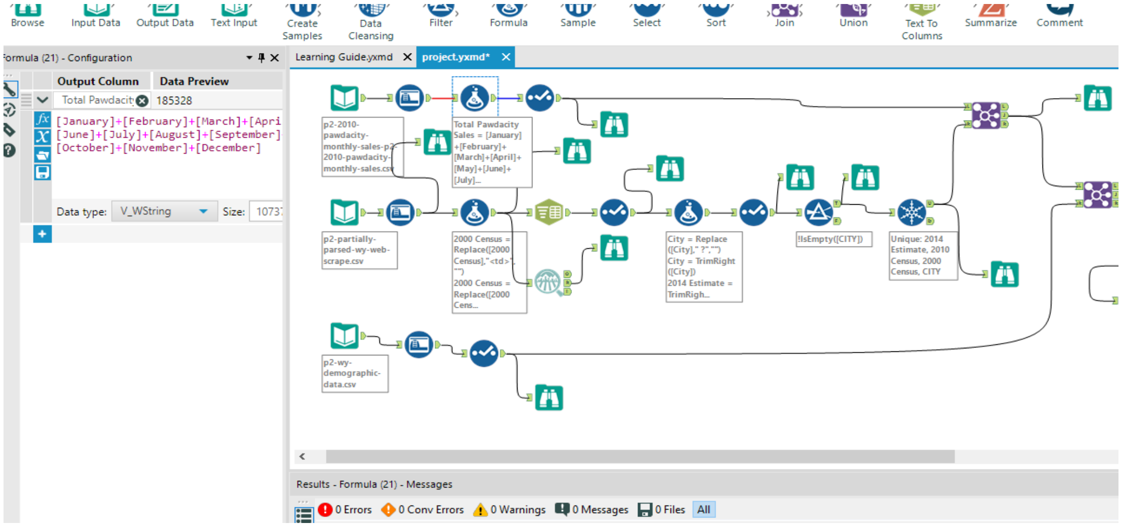
Formula (8) - Configuration

Output Column	Data Preview
2000 Census	1,818</td>

```
Replace([2000 Census],"<td>", "")
```

Data type: V_String Size: 134

Output Column	Data Preview
2000 Census	1,818
2010 Census	1,911</td>
2010 Census	1,911
2014 Estim...	1,968</td>
2014 Estim...	1,968
2000 Census	1818
2010 Census	1911
2014 Estim...	1968


Browse
Input Data
Output Data
Text Input
Create Samples
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Comment
Learning Guide.yxmd
project.yxmd*
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
2000 Census = Replace([2000 Census],"<td>","")
2000 Census = Replace([2000 Cens...
City = Replace([City]," ?","")
City = TrimRight([City])
2014 Estimate = TrimRigh...
!IsEmpty([CITY])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
p2-wy-demographic-data.csv

Formula (21) - Configuration

Output Column	Data Preview
Total Pawdacit	185328

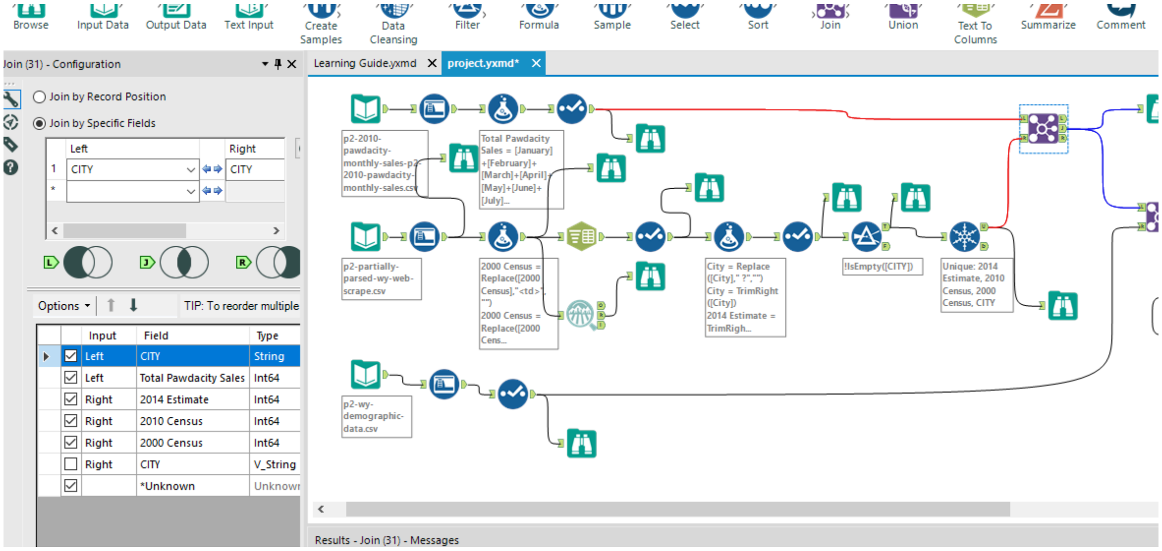
```
[January]+[February]+[March]+[Apri
[June]+[July]+[August]+[September]
[October]+[November]+[December]
```

Data type: V_WString Size: 1073

Results - Formula (21) - Messages

0 Errors 0 Conv Errors 0 Warnings 0 Messages 0 Files All

Data Blending :- Used the join function to blend the data from three files into single dataset.


Browse
Input Data
Output Data
Text Input
Create Samples
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Comment
Learning Guide.yxmd
project.yxmd*
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
2000 Census = Replace([2000 Census],"<td>","")
2000 Census = Replace([2000 Cens...
City = Replace([City]," ?","")
City = TrimRight([City])
2014 Estimate = TrimRigh...
!IsEmpty([CITY])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
p2-wy-demographic-data.csv
L
J
R

Join (31) - Configuration

○ Join by Record Position

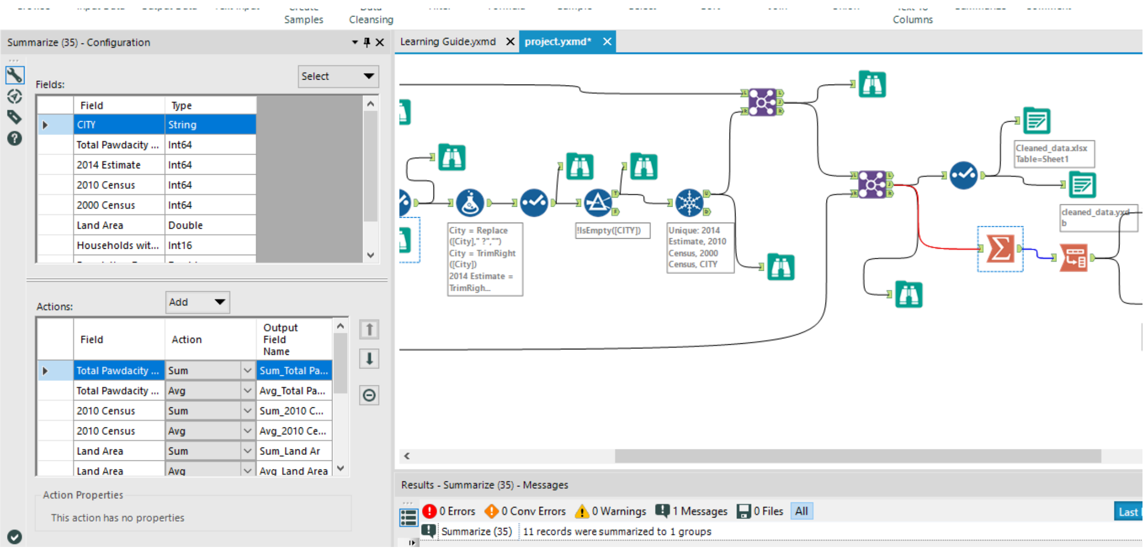
● Join by Specific Fields

	Left	Right
1	CITY	CITY
*		

Options TIP: To reorder multiple

	Input	Field	Type
☑	Left	CITY	String
☑	Left	Total Pawdacity Sales	Int64
☑	Right	2014 Estimate	Int64
☑	Right	2010 Census	Int64
☑	Right	2000 Census	Int64
□	Right	CITY	V_String
☑		*Unknown	Unknown

Results - Join (31) - Messages


Samples
Cleansing
Columns
Learning Guide.yxmd
project.yxmd*
Select
Add
City = Replace ([City]," ?","")
City = TrimRight ([City])
2014 Estimate = TrimRigh...
!IsEmpty([CITY])
Unique: 2014 Estimate, 2010 Census, 2000 Census, CITY
Cleaned_data.xlsx Table=Sheet1
cleaned_data.yxdb

Summarize (35) - Configuration

Fields:

Field	Type
CITY	String
Total Pawdacity ...	Int64
2014 Estimate	Int64
2010 Census	Int64
2000 Census	Int64
Land Area	Double
Households wit...	Int16

Actions:

Field	Action	Output Field Name
Total Pawdacity ...	Sum	Sum_Total Pa...
Total Pawdacity ...	Avg	Avg_Total Pa...
2010 Census	Sum	Sum_2010 C...
2010 Census	Avg	Avg_2010 Ce...
Land Area	Sum	Sum_Land Ar
Land Area	Avg	Avg_Land Area

Action Properties

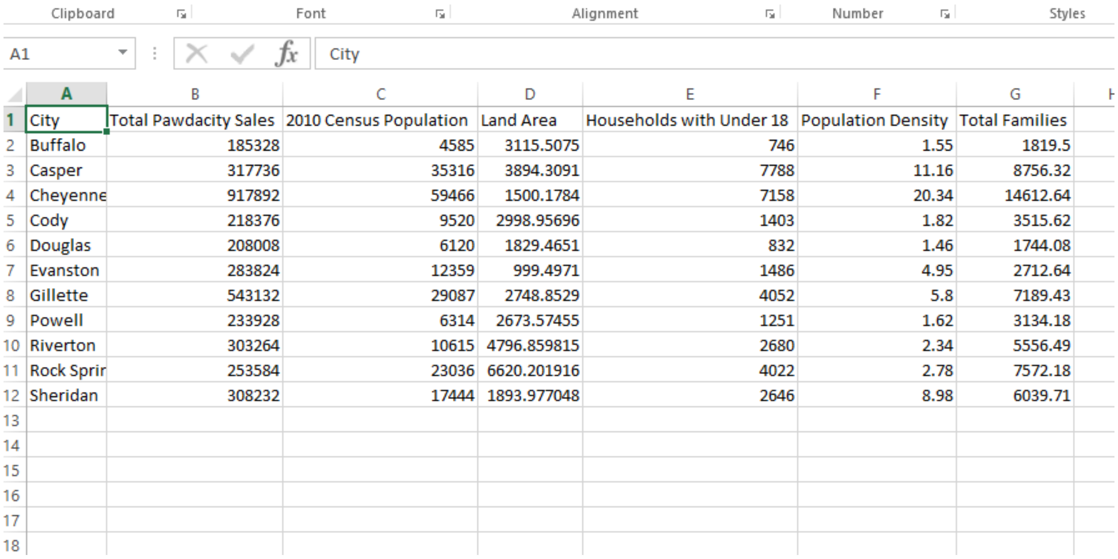
This action has no properties

Results - Summarize (35) - Messages

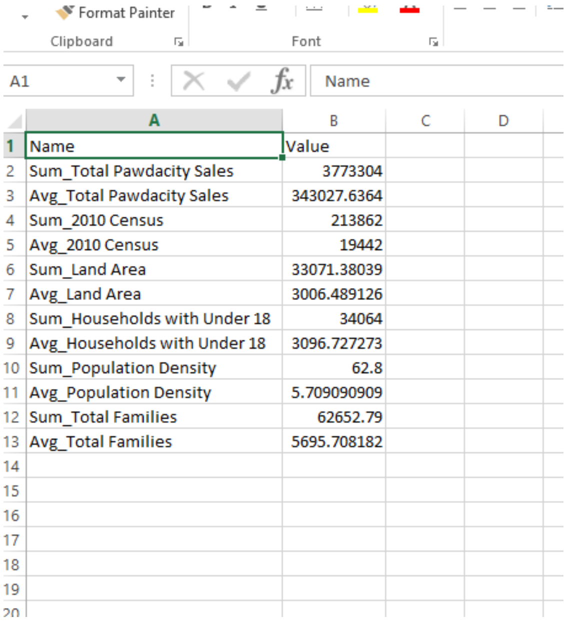
0 Errors 0 Conv Errors 0 Warnings 1 Messages 0 Files All

Summarize (35) 11 records were summarized to 1 groups

Output :- Formatted and cleaned data was obtained in an excel file and the characteristics need to be calculated were outputted to another excel file.



City	Total Pawdacity Sales	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.5075	746	1.55	1819.5
Casper	317736	35316	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Cody	218376	9520	2998.95696	1403	1.82	3515.62
Douglas	208008	6120	1829.4651	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.8529	4052	5.8	7189.43
Powell	233928	6314	2673.57455	1251	1.62	3134.18
Riverton	303264	10615	4796.859815	2680	2.34	5556.49
Rock Sprir	253584	23036	6620.201916	4022	2.78	7572.18
Sheridan	308232	17444	1893.977048	2646	8.98	6039.71



Name	Value
Sum_Total Pawdacity Sales	3773304
Avg_Total Pawdacity Sales	343027.6364
Sum_2010 Census	213862
Avg_2010 Census	19442
Sum_Land Area	33071.38039
Avg_Land Area	3006.489126
Sum_Households with Under 18	34064
Avg_Households with Under 18	3096.727273
Sum_Population Density	62.8
Avg_Population Density	5.709090909
Sum_Total Families	62652.79
Avg_Total Families	5695.708182

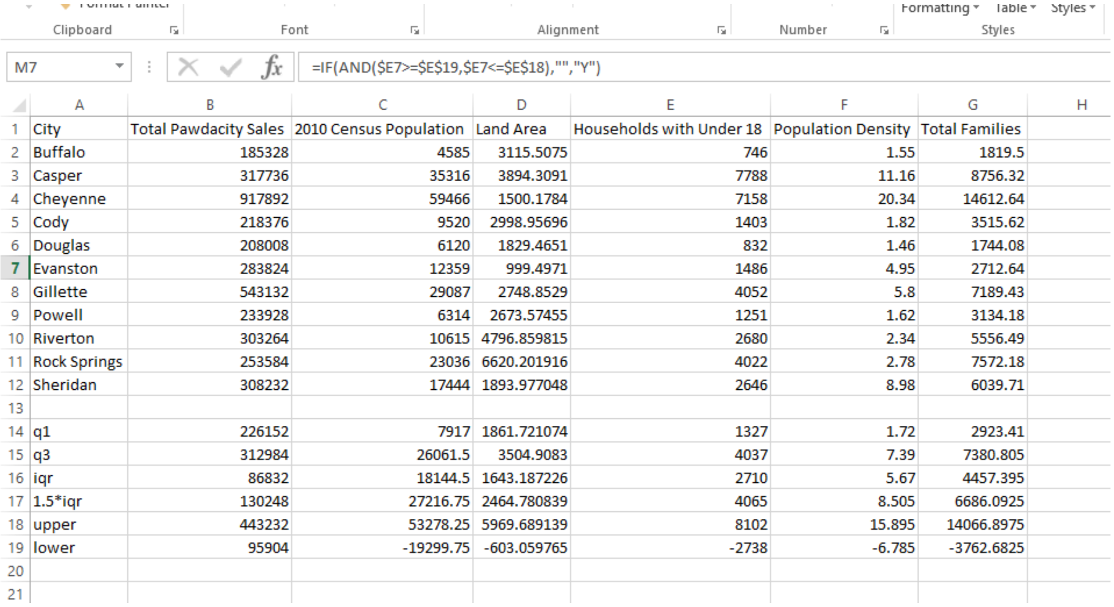
And the answers to question no.2 were obtained.

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.

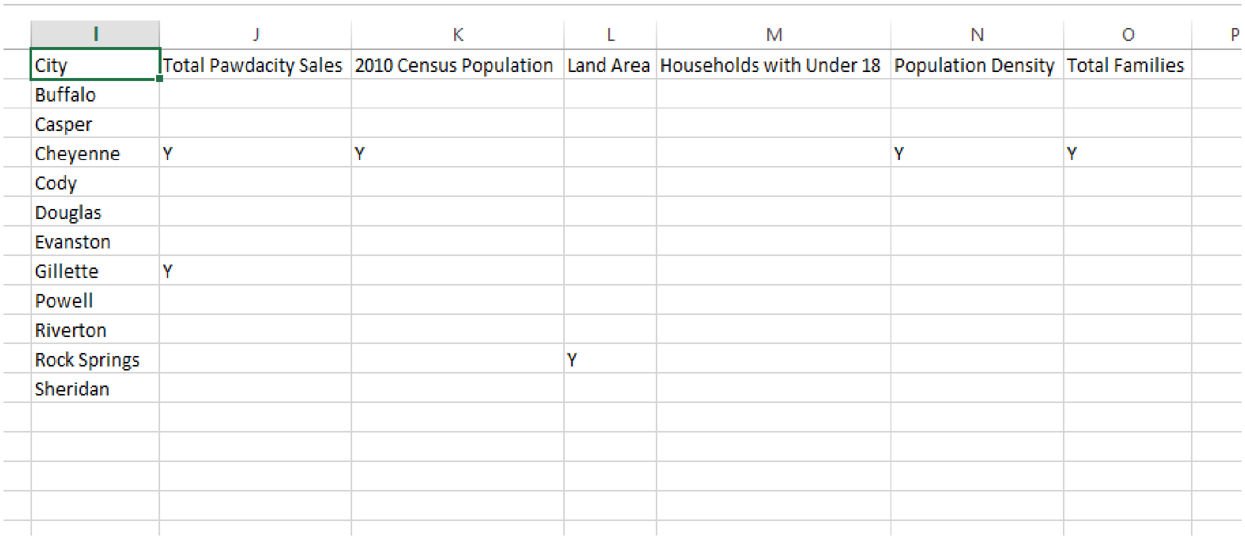
Calculating the interquartile ranges:-



=IF(AND($E7>=$E$19,$E7<=$E$18),"","Y")

City	Total Pawdacity Sales	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.5075	746	1.55	1819.5
Casper	317736	35316	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Cody	218376	9520	2998.95696	1403	1.82	3515.62
Douglas	208008	6120	1829.4651	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.8529	4052	5.8	7189.43
Powell	233928	6314	2673.57455	1251	1.62	3134.18
Riverton	303264	10615	4796.859815	2680	2.34	5556.49
Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
q1	226152	7917	1861.721074	1327	1.72	2923.41
q3	312984	26061.5	3504.9083	4037	7.39	7380.805
iqr	86832	18144.5	1643.187226	2710	5.67	4457.395
1.5*iqr	130248	27216.75	2464.780839	4065	8.505	6686.0925
upper	443232	53278.25	5969.689139	8102	15.895	14066.8975
lower	95904	-19299.75	-603.059765	-2738	-6.785	-3762.6825

<u>Outliers found :-</u>



City	Total Pawdacity Sales	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families
Buffalo						
Casper						
Cheyenne	Y	Y			Y	Y
Cody						
Douglas						
Evanston						
Gillette	Y					
Powell						
Riverton						
Rock Springs			Y			
Sheridan						

➔ There are 3 cities which have outliers :-

1) Cheyenne
2) Gillette
3) Rock Springs

➔ Cheyenne has 4 outliers and it seems to be a big city ,so I will not remove this city .

➔ Rock –springs have land area as an outlier ,but its all other parameters are in the interquartile range. So it might be accurate date ,because of having large land area,and it won't be removed.

➔ Gillette has all its parameters in interquartile range but its total pawdacity sales is out of interquartile range,which I don't think is possible as compared to the other cities. So,this city should be removed.