Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   - Recommend a city for Pawdacity's newest store in Wyoming state, based on predicted yearly sales.

2. What data is needed to inform those decisions?

   - The sales data for all of the Pawdacity stores.
   - Data of the population numbers.
   - Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city in Wyoming state.

Step 2: Building the Training Set

1. The workflow:

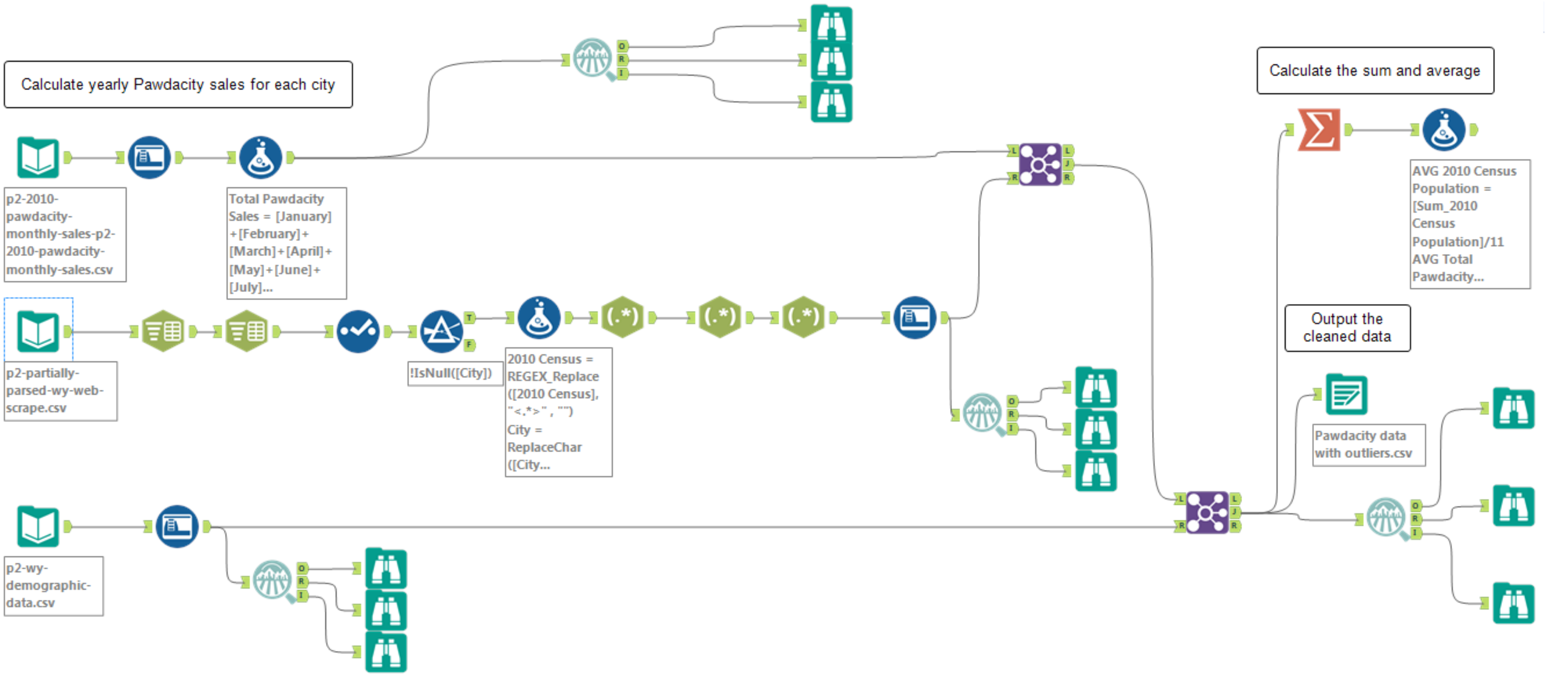

Calculate yearly Pawdacity sales for each city
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]...
p2-partially-parsed-wy-web-scrape.csv
!IsNull([City])
T
F
2010 Census = REGEX_Replace([2010 Census], "<.*>" , "")
City = ReplaceChar([City...
p2-wy-demographic-data.csv
O
R
I
L
J
R
Calculate the sum and average
AVG 2010 Census Population = [Sum_2010 Census Population]/11
AVG Total Pawdacity...
Output the cleaned data
Pawdacity data with outliers.csv

Figure 1: Cleaning data workflow

2. The averages of the data set:

Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Table 1*:* Averages of all columns in the data set

Step 3: Dealing with Outliers

1. Are there any cities that are outliers in the training set?

    - There are three cities that we can consider as outliers:

        - **Cheyenne:** has outlier values for:

            - Census Population
            - Total Pawdacity Sales
            - Population Density
            - Total Families

        - **Gillette:** has outlier for:

            - Total Pawdacity Sales

        - **Rock Springs :** has outlier for:

            - Land Area

2. Which outlier have you chosen to remove or impute?

    - **Cheyenne**

        - Cheyenne's values for the different fields are larger in comparison with the other cities, it has four outliers. If it was only for Census Population, Population Density and Total Families we could think of keeping this city for further investigation as there values are not too far from the upper fence, but the fourth outlier which is Total Pawdacity Sales has too big value compared to the upper fence, this can be because the city is big and it has a high population.

        - **It will be better to exclude Cheyenne from the dataset in order to build an unbiased model.**

	Census Population	Total Pawdacity Sales	Population Density	Total Families
Cheyenne	59,466	917,892	20.34	14,612.64
Upper fence	53,278.25	443,232	15.89	14,066.90

Table 2: Comparison between Cheyenne outliers and Upper fence values

- **Gillette:**
  - Comparing with Cheyenne, Gillette's total sales is not big, and since the data we have is small it would be better if we don't delete too many cities. Keeping Gillette would be a good choice.

	Total Pawdacity Sales
Gillette	543,132
Upper fence	443,232

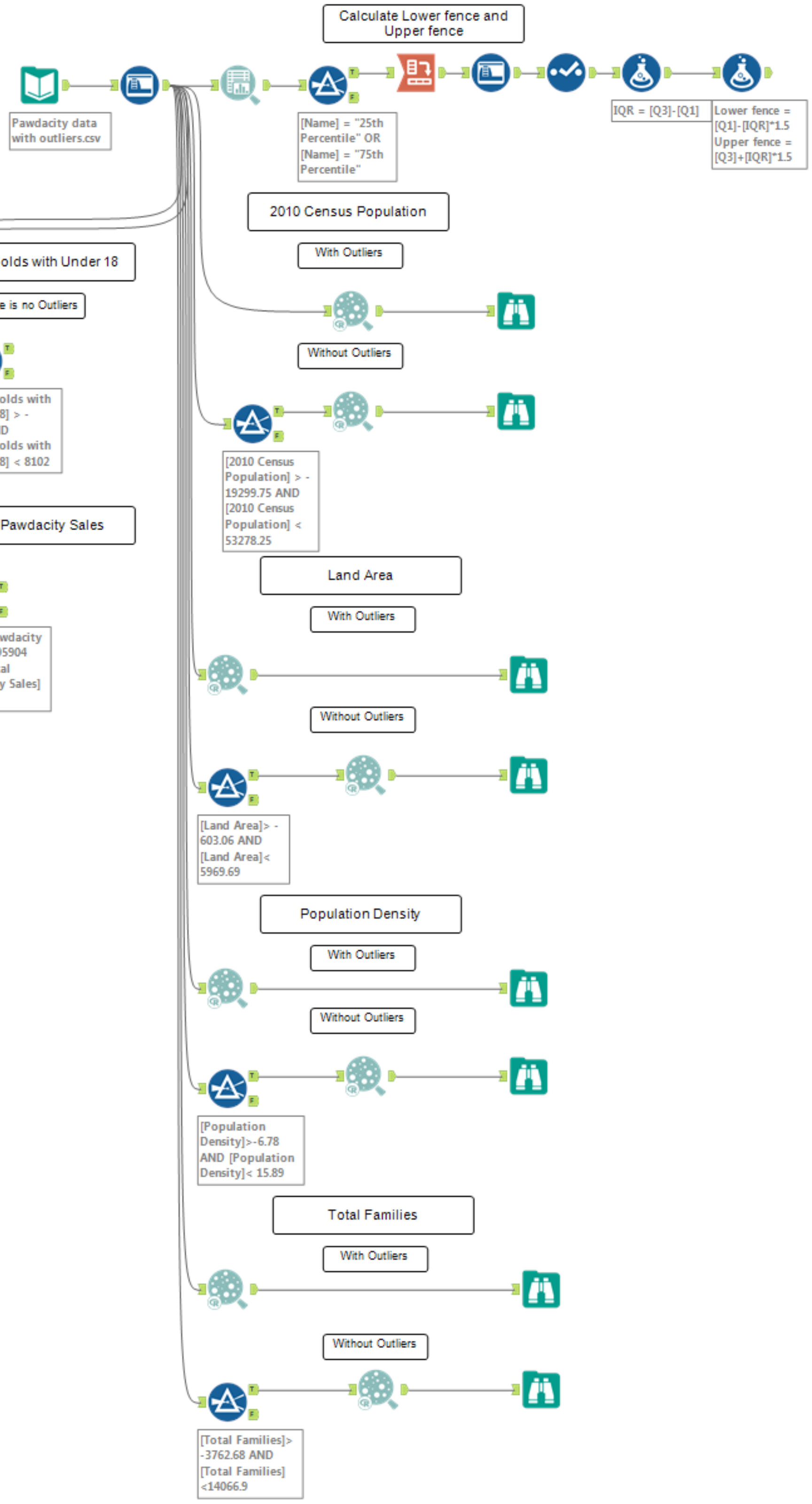
Table 3: Comparison between Gillette outlier and Upper fence

- **Rock Springs:**
  - The value of the outlier is slightly out of range, we can keep this city.

	Land Area
Rock Springs	6,620.20
Upper fence	5,969.69

Table 4: Comparison between Rock Springs outlier and Upper fence

3. The workflow:



```
[Name] = "25th Percentile" OR [Name] = "75th Percentile"
```

```
IQR = [Q3]-[Q1]
```

```
Lower fence = [Q1]-[IQR]*1.5
Upper fence = [Q3]+[IQR]*1.5
```

```
[Households with Under 18] > -2738 AND [Households with Under 18] < 8102
```

```
[Total Pawdacity Sales] > 95904 AND [Total Pawdacity Sales] < 443232
```

```
[2010 Census Population] > -19299.75 AND [2010 Census Population] < 53278.25
```

```
[Land Area]> -603.06 AND [Land Area]< 5969.69
```

```
[Population Density]>-6.78 AND [Population Density]< 15.89
```

```
[Total Families]> -3762.68 AND [Total Families] <14066.9
```

Figure 2: Dealing with outliers workflow