Project 2.1: Data Cleanup

Jorge Pinzon

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?
    The objective is to generate a dataset from different sources. Each source needs to be analyzed in order to clean the data as much as possible. For example, if an intended variable has missing values, this dataset can be removed of the missing value filled with a calculated value. In this case I did not find missing values, but in order to merge the data I have to modify the column with the City name to allow for the join to happen. During the join, it is also necessary to decide which columns to keep and which to remove from the final set. In this case the desired columns are: 'City', '2010 Census Population', 'Total Pawdacity Sales', 'Households with Under 18', 'Land Area', 'Population Density', and 'Total Families'. Ultimately this dataset will be used to perform and analysis to recommend a city where a new, the 14th, pet store should be open.

2. What data is needed to inform those decisions?
    List of expected columns. Demographic data, census data, and sales data are required for this process.

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 (with outliers)	Average (no outliers
Census Population	213,862	19,442.00	15,439.60
Total Pawdacity Sales	3,773,304	342,027.64	285,541.20
Households with Under 18	34,064	3,096.39	2,690.60
Land Area	33,071	3,006.49	3,157.12
Population Density	63	5.71	4.25
Total Families	62,653	5,695.71	4,804.02

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.

Yes, there is one city that has outlier values for most of the feature. In general, Cheyenne (Fig 1) has higher Population, total sales, households, population density with lower number of families than the average for the other 10 cities. Population Density and Total Pawdacity Sales for Cheyenne are outside the range of the IQR and are clearly higher than the other cities. These outliers, add variation to the features if we were going to use the whole dataset (with outliers). The IQR method used for outlier identification correctly found the Cheyenne as an outlier city I strongly suggest to remove it from the dataset for further analyses.

Fig 1. Summary of Data for Cheyenne.

City	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Cheyenne	59466.0	917892	7158	1500.1784	20.34	14612.64

The outlier analysis also identified Gillete to have outliers, but only in the Total Pawdacity Sales. While the sales value for Gillete is higher and outside the IQR it is the only feature this city has with outliers and all other feature's values are not only inside the IQR, but not tending toward any direction in particular. In other words, the other features show Gillete within the range of the other cities, contrary to Cheyenne, where all the feature are either the highest or the lowest values across cities. Additionally, the Sales value for Gillete is not that far from the higher IQR value suggesting it might be safe to keep this value and therefore the city in the dataset. I will keep 10 cities in my dataset.

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.