Kacper Ksieski

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?
   Pawdacity, a Wyoming pet store chain, would like to expand and open a 14th store. The decision to be made is the city in which to locate the 14th store. The choice of city is based on the predicted yearly sales.

2. What data is needed to inform those decisions?
   The decision rests on predicted yearly sales per city. The data needed to predict yearly sales per city are actual yearly sales per city, population metrics such as size, density, and total families, and land area. Going above and beyond, some predictive analysis on the population metrics using additional data would be useful to determine which populations are experience growth or decline. This could be used as a categorical predictor or as a numerical predictor. I would like to search for this data since it is a matter of public record and perform the analysis.

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*	*19442*
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

Kacper Ksieski

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.

There are six outlier values.
The city of Cheyenne sticks out as a contributor of four of these outliers. Given that Cheyenne is one of the larger cities, its values seem consistent with that fact. For example, the population density is much higher than average, as are the total families and population metrics. Cheyenne has two stores to serve this huge population and this explains the large sales metric. The Cheyenne outliers don't seem to be wrong. I do not have compelling reasons to remove or impute these outliers.
The city of Rock Springs has an outlier of land area and it is only about 10% higher than the upper fence and double the average. Because I don't have compelling reasons to remove or impute this outlier in such a tiny dataset, it will remain in the dataset.
Finally, the city of Gillette has an outlier of total Pawdacity sales. Gillette, like Cheyenne, has two stores. Gillette has half the population with a quarter of the population density. Despite having 50% of the population of Cheyenne, Gillette has 60% of the sales. It is possible that Gillette is a more affluent city and for this reason should be removed from the dataset as an outlier.
In conclusion, I have chosen to remove Gillette as an outlier.

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.