Project 2.1: Data Cleanup | Ioanna Vasilopoulou

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 is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales. In order to do this I will need to format and blend together the data provided by my manager. This training set, which will help me builth the regression model in order to predict the sales in the practice project should include the following columns: City, 2010 Census Population, Total Pawdacity Sales, Households with people under 18, Land Area, Population Density and Total Families. Lastly, I will need to identify the upper and lower fence to determine if there are outlier cities or not. In case there are, I will need to provide the reasons why the city that has at least one outlier value should be removed.

2. What data is needed to inform those decisions?

The data I will need to work with in order to inform these decisions is the monthly sales data for all of the Pawdacity stores for the year 2010. This will allow me to find each city's total sales. Census population data can be found in the p2-partially-parsed-wy-web-scrape file, while demographic data can be found in the p2-wy-demographic-data file. After working with this data I will be able to calculate the dataset's 1st quartile Q1 and 3rd quartile Q3. Using the IQR = Q3 - Q1 equation, I will calculate the interquartile range and therefore be able to find the upper and lower fence values. Any values above the upper fence and any values below the lower fence, will be identified as outliers.

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	**343,027.64**
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**

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.

Based on the findings from the training dataset, I have identified 3 cities which are outliers:

- **Cheyenne:** Total sales, 2010 census population, population density and total families are above the Upper Fence.
- **Gillette:** Total sales is above the Upper Fence.
- **Rock Springs:** Land area is above the Upper Fence.

Which city will be included in / removed from the dataset and why?

Although **Cheyenne** shows results above the Upper Fence in 4 columns, we notice that there is a clear correlation amongst this data. Since the population and its density are higher with the number of families being more than the other cities, it makes sense for the amount of sales to being respectively higher for this city. Therefore, we will keep this city in our dataset.

Gillette is only having one value above the Upper Fence; sales. We notice, though, that none of the other values for this city is in proportion with the sales. As this might wrongly affect the results we are looking into getting, we can remove this from the data set.

Land area value for Rock Springs is above the Upper Fence, hence this city is identified as an outlier. However, as the land area is not a major factor directly associated with the sales, we can keep this city and its values into our dataset.

City	Sales	2011 Census Population	Households with Under 18	Population Density	Land Area	Total Families
Cheyenne	917,892	59,466	7,158	20.34	1,500.18	14,612.64
Gillette	543,132	29,087	4,052	5.8	2,748.85	7,189.43
Casper	317,736	35,316	7,788	11.16	3,894.31	8,756.32
Sheridan	308,232	17,444	2,646	8.98	1,893.98	6,039.71
Riverton	303,264	10,615	2,680	2.34	4,796.86	5,556.49
Evanston	283,824	12,359	1,486	4.95	999.50	2,712.64
Rock Springs	253,584	23,036	4,022	2.78	6,620.20	7,572.18
Powell	233,928	6,314	1,251	1.62	2,673.57	3,134.18
Cody	218,376	9,520	1,403	1.82	2,998.96	3,515.62
Douglas	208,008	6,120	832	1.46	1,829.47	1,744.08
Buffalo	185,328	4,585	746	1.55	3,115.51	1,819.50
Total	*3,773,304*	*213,862*	*34,064*	*62.8*	*33071.38*	*62,652.79*

Average	*343,028*	*19,442*	*3,097*	*5.71*	*3006.49*	*5,695.71*
Median	*283,824*	*12,359*	*2,646*	*2.78*	*2748.85*	*5,556.49*
Q1	*226,152*	*7,917*	*1,327*	*1.72*	*1861.72*	*2,923.41*
Q3	*312,984*	*26,062*	*4,037*	*7.39*	*3504.91*	*7,380.81*
Minimum	*185,328*	*4,585*	*746*	*1.46*	*999.50*	*1,744.08*
Maximum	*917,892*	*59,466*	*7,788*	*20.34*	*6620.20*	*14,612.64*
Interquartile Range	*86,832*	*18,145*	*2,710*	*5.67*	*1643.19*	*4,457.40*
Upper Fence	*443,232*	*53,278*	*8,102*	*15.90*	*5969.69*	*14,066.90*
Lower Fence	*95,904*	*-19,300*	*-2,738*	*-6.785*	*-603.06*	*-3,762.68*