MISK ACADEMY أكاديمية مسك UDACITY

TECH LAB

Predictive Analytics for Business

Project#3-1 Create an Analytical Dataset

Name: Marwan Saeed Alsharabbi

Date: 26−12−2019

2019

INTRODUCTION

.

The Business Problem

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.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1-The monthly sales data for all of the Pawdacity stores for the year 2010.
2-NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3-A partially parsed data file that can be used for population numbers.
4-Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Steps to Success

Step 1: Business and Data Understanding

Your project should include a description of the key business decisions that need to be made.

Step 2: Building the Training Set

To properly build the model, and select predictor variables, create a dataset with the following columns:

```
City
2010 Census Population
Total Pawdacity Sales
Households with Under 18
Land Area
```

`Population Density`
`Total Families`

This dataset will be your training set to help you build a regression model in order to predict sales in the Practice Project in the next lesson. Every row should have sales data because we're trying to predict sales.

Notes

You should be consolidating the data at the city level and **not at the store level.** We only have data at the city wide level so any analysis at the store level will not be sufficient to complete this analysis.
We simply need to focus on cleaning up and blending the data together in this step.

If you’ve done everything correctly, the sum for each of the above columns should be:

Census Population: 213,862
Total Pawdacity Sales: 3,773,304
Households with Under 18: 34,064
Land Area: 33,071
Population Density: 63
Total Families: 62,653
with **11 rows of data**
For Alteryx users:

Use the Autofield Tool to help quickly convert your data fields into the appropriate datafields for analysis.
Research these three specific formulas to help you get rid of unwanted characters in the Formula tool: ReplaceFirst, Left, FindString

Step 3: Dealing with Outliers

Once you have created the dataset, look for outliers and figure out how deal with your outliers. Use the IQR method to determine if there are outlier cities for each of the variables and then justify which city that has at least one outlier value should be removed.

IQR Steps

To calculate the upper fence and the lower fence, here are the exact steps:

1. Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. You can use the Excel function QUARTILE.INC or QUARTILE.EXC

2. Calculate the Interquartile Range: IQR = Q3 - Q1

3. Add 1.5 *IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5* IQR
4. Subtract 1.5 *IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5* IQR
5. Values above the Upper Fence and values below the Lower Fence are outliers

A description of the key business decisions that need to be made.
Note: Clean data is provided for this project, so you can skip the data preparation step of the Problem Solving Framework.

Data

p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.
p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.
p2-wy-453910-naics-data.csv - NAICS data on the sales of all competitor stores where total sales is equal to 12 months of sales
p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.

Project 3.1: Data Cleanup

Step 1: Business and Data Understanding

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

Key Decisions:

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.

Answer these questions
What decisions needs to be made?

1- Business Issue Understanding

- What decisions needs to be made?
- What information is needed to inform those decisions?
- What type of analysis can provide the information needed to inform those decisions?

2- Data Understanding

- What data is needed?
- What data is available?
- What are the important characteristics of the data?

In this project the problem is explained and the data are available and I will choose the necessary data tables in order to clean them by using a program Alteryx
I'm chose three sets data:
p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.
p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.
p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.
In the first part, you will blend and format data and deal with outliers.

What data is needed to inform those decisions?

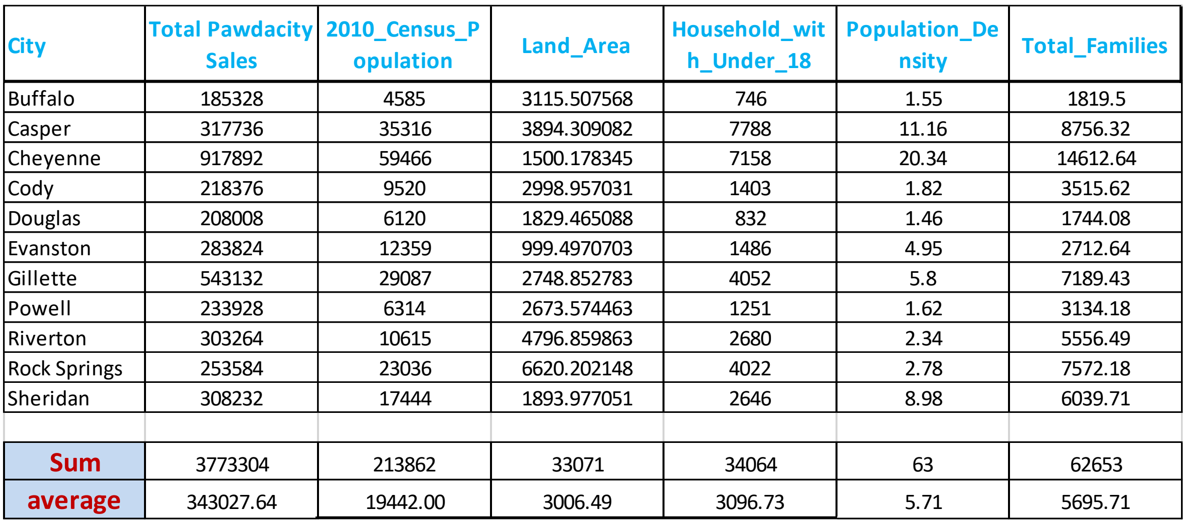
After the process of understanding the data, coordinating and cleaning it, and linking the tables together, shown above, I will get the following columns

`City`
`2010 Census Population`
`Total Pawdacity Sales`
`Households with Under 18`
`Land Area`
`Population Density`
`Total Families`

The data from the above fields will later be used to create a prediction model for the new store location.

Step 2: Building the Training Set

You should be consolidating the data at the city level and **not at the store level.** We only have data at the city wide level so any analysis at the store level will not be sufficient to complete this analysis. We simply need to focus on cleaning up and blending the data together in this step.
Result Data Set after import data form alteryx I use excel



City	Total Pawdacity Sales	2010_Census_Population	Land_Area	Household_with_Under_18	Population_Density	Total_Families
Buffalo	185328	4585	3115.507568	746	1.55	1819.5
Casper	317736	35316	3894.309082	7788	11.16	8756.32
Cheyenne	917892	59466	1500.178345	7158	20.34	14612.64
Cody	218376	9520	2998.957031	1403	1.82	3515.62
Douglas	208008	6120	1829.465088	832	1.46	1744.08
Evanston	283824	12359	999.4970703	1486	4.95	2712.64
Gillette	543132	29087	2748.852783	4052	5.8	7189.43
Powell	233928	6314	2673.574463	1251	1.62	3134.18
Riverton	303264	10615	4796.859863	2680	2.34	5556.49
Rock Springs	253584	23036	6620.202148	4022	2.78	7572.18
Sheridan	308232	17444	1893.977051	2646	8.98	6039.71
Sum	3773304	213862	33071	34064	63	62653
average	343027.64	19442.00	3006.49	3096.73	5.71	5695.71

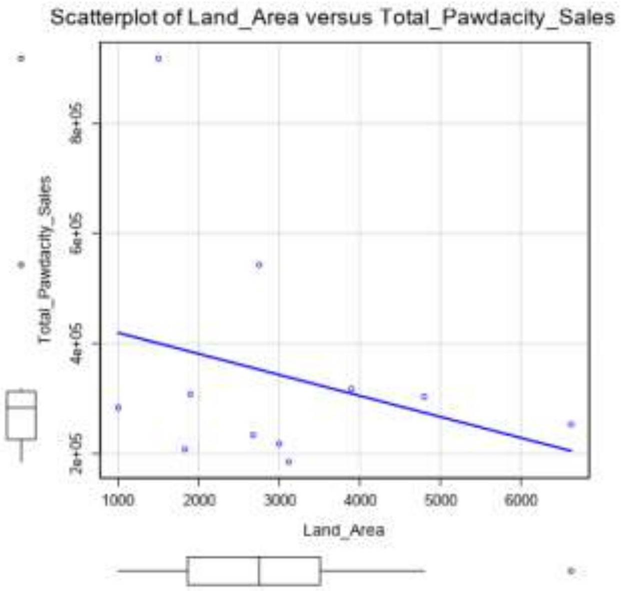
And the result for use Alteryx

Column	Sum	Average
2010 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***	***3006.45***
Population Density	***63***	***5.73***
Total Families	***62,653***	***5695.73***

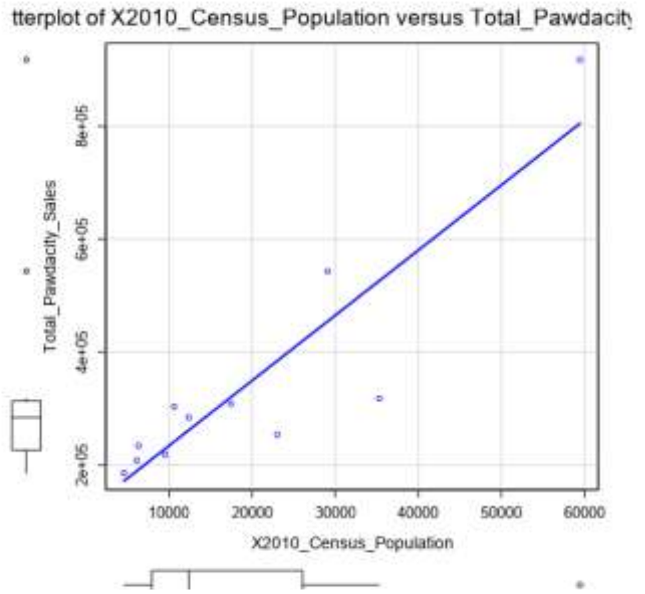
Step 3: Dealing with Outliers

The point is not to simply find outliers, but to understand what the extreme data points that exist in your data that may affect your analysis.

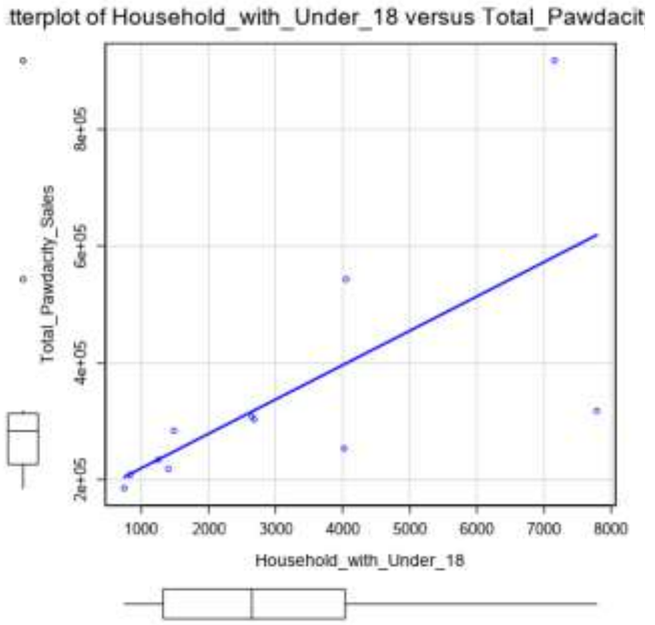
Below are scatter plots and boxplots of the dataset, with each potential predictor variable plotted against the Total Pawdacity Sales for that city.



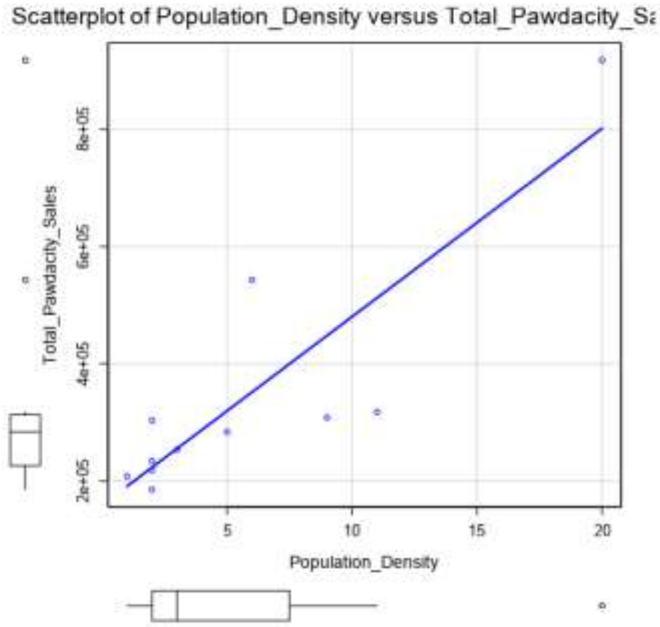
Scatterplot of Land_Area versus Total_Pawdacity_Sales



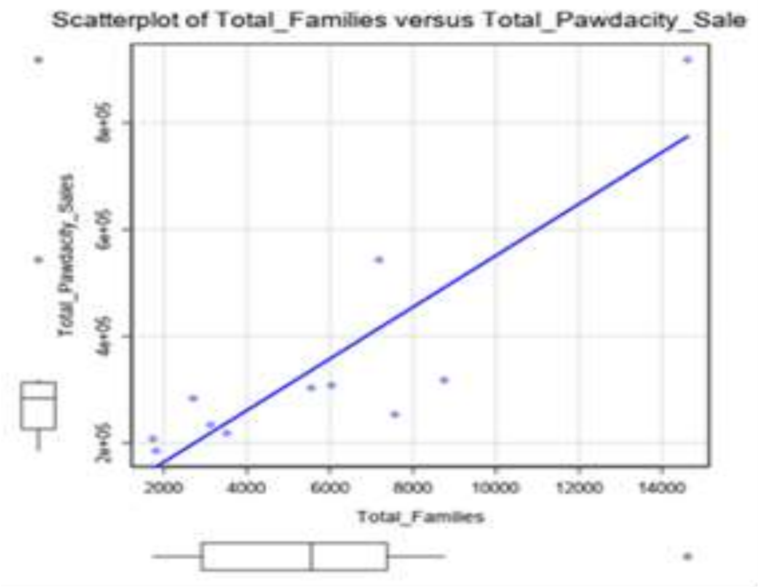
tterplot of X2010_Census_Population versus Total_Pawdacity



tterplot of Household_with_Under_18 versus Total_Pawdacit

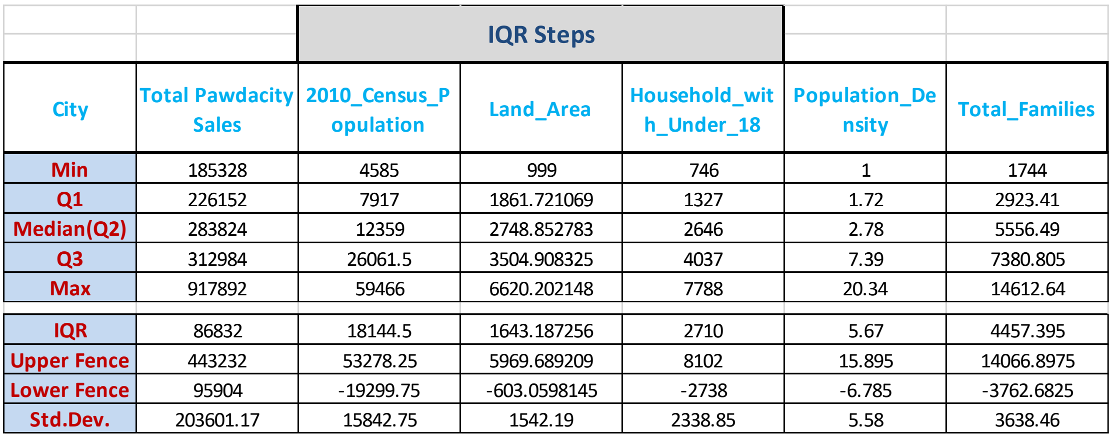


Scatterplot of Population_Density versus Total_Pawdacity_Sa



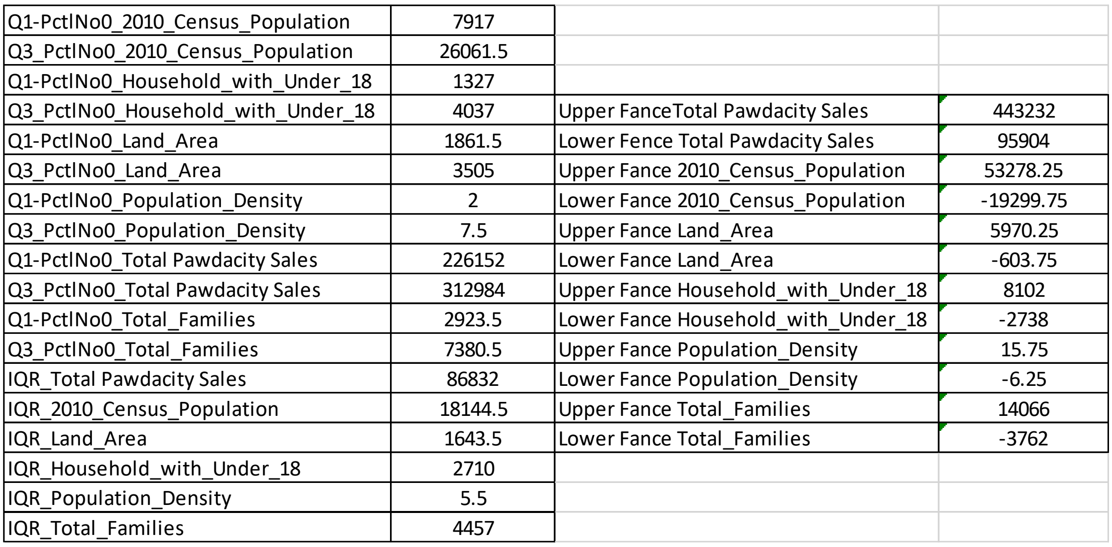
Scatterplot of Total_Families versus Total_Pawdacity_Sale

I’m Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. I’m use the Excel function QUARTILE.INC, Calculate the Interquartile Range: IQR = Q3 - Q1 and *Upper Fence = Q3 + 1.5* IQR, *Lower Fence = Q1 - 1.5* IQR and Alteryx Below is a summary of the dataset by excel



		IQR Steps				
City	Total Pawdacity Sales	2010_Census_Population	Land_Area	Household_with_Under_18	Population_Density	Total_Families
Min	185328	4585	999	746	1	1744
Q1	226152	7917	1861.721069	1327	1.72	2923.41
Median(Q2)	283824	12359	2748.852783	2646	2.78	5556.49
Q3	312984	26061.5	3504.908325	4037	7.39	7380.805
Max	917892	59466	6620.202148	7788	20.34	14612.64
IQR	86832	18144.5	1643.187256	2710	5.67	4457.395
Upper Fence	443232	53278.25	5969.689209	8102	15.895	14066.8975
Lower Fence	95904	-19299.75	-603.0598145	-2738	-6.785	-3762.6825
Std.Dev.	203601.17	15842.75	1542.19	2338.85	5.58	3638.46

The result by [Alteryx]



Q1-PctlNo0_2010_Census_Population	7917		
Q3_PctlNo0_2010_Census_Population	26061.5		
Q1-PctlNo0_Household_with_Under_18	1327		
Q3_PctlNo0_Household_with_Under_18	4037	Upper FanceTotal Pawdacity Sales	443232
Q1-PctlNo0_Land_Area	1861.5	Lower Fence Total Pawdacity Sales	95904
Q3_PctlNo0_Land_Area	3505	Upper Fance 2010_Census_Population	53278.25
Q1-PctlNo0_Population_Density	2	Lower Fance 2010_Census_Population	-19299.75
Q3_PctlNo0_Population_Density	7.5	Upper Fance Land_Area	5970.25
Q1-PctlNo0_Total Pawdacity Sales	226152	Lower Fance Land_Area	-603.75
Q3_PctlNo0_Total Pawdacity Sales	312984	Upper Fance Household_with_Under_18	8102
Q1-PctlNo0_Total_Families	2923.5	Lower Fance Household_with_Under_18	-2738
Q3_PctlNo0_Total_Families	7380.5	Upper Fance Population_Density	15.75
IQR_Total Pawdacity Sales	86832	Lower Fance Population_Density	-6.25
IQR_2010_Census_Population	18144.5	Upper Fance Total_Families	14066
IQR_Land_Area	1643.5	Lower Fance Total_Families	-3762
IQR_Household_with_Under_18	2710		
IQR_Population_Density	5.5		
IQR_Total_Families	4457		

From analyzing our data, and Scatterplot we will not look at the extreme values. We are in the process of opening a new store to solve the incoming problem despite its appearance in Cheyenne the highest Total pawdacity sales are in Cheyenne, Gillette and Cheyenne City for Census Population, Land Area, Population Density, Land Area and Total Pawdacity sales for Gillette. The scatterplot for Land Area vs Sales would indicate to me that Rock Springs follows the downward direction of the line of best fit for that plot with sales roughly in line with other sales values in that plot. Through the data, the most recommended cities to open the new store are Cheyenne and Gillette. My recommendation is to keep Cheyenne because it has appropriate data to open the store through the percentage of sales and population density, and removing Gillette.

Help resources :https://knowledge.udacity.com

I wish success to all.

Marwan Saeed Alsharabbi