Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Business Issue

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 manager has given you the following information to work with:

- *The monthly sales data for all of the Pawdacity stores for the year 2010.*
- *NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.*
- *A partially parsed data file that can be used for population numbers.*
- *Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.*

Here are the criteria given to you in choosing the right city:

- *The new store should be located in a new city. That means there should be no existing stores in the new city.*
- *The total sales for the entire competition in the new city should be less than $500,000*
- *The new city where you want to build your new store must have a population over 4,000 people (based upon the 2014 US Census estimate).*
- *The predicted yearly sales must be over $200,000.*
- *The city chosen has the highest predicted sales from the predicted set.*

Provide an explanation of the key decisions that need to be made.

Key Decisions:

Answer these questions:

1. *What decisions needs to be made?*
   We need to assess which city, in the state of Wyoming, is the ideal location for the 14th Pawdacity's newest store, based on predicted yearly sales.

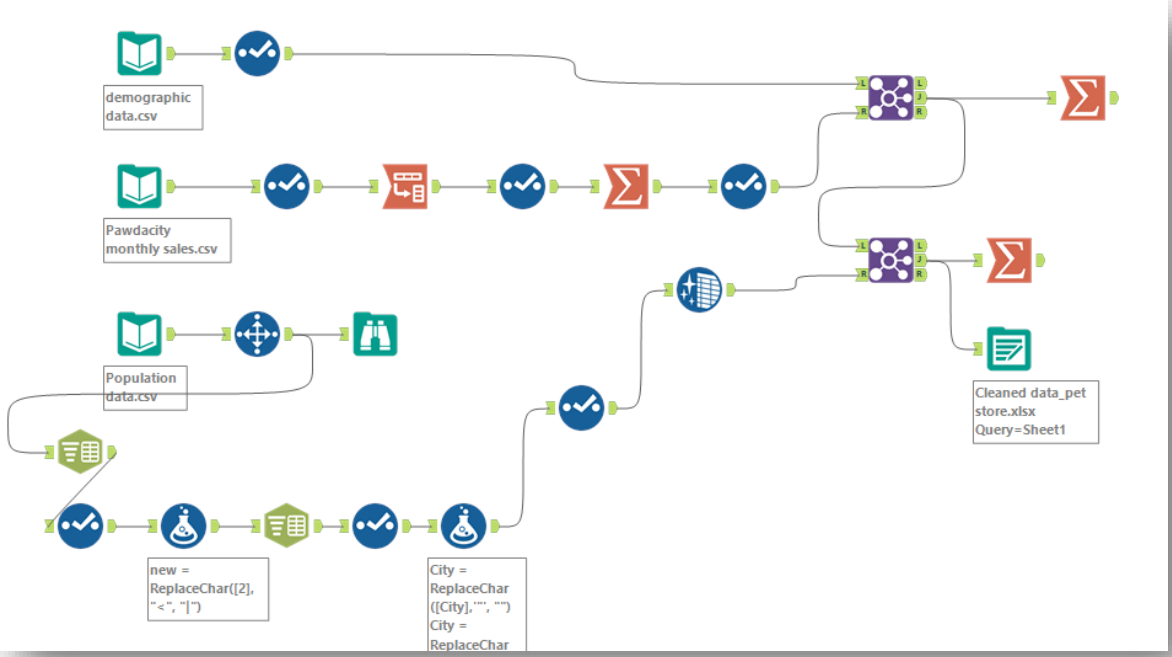
2. *What data is needed to inform those decisions?*
   To build a training set for our regression model we need to identify predictor variables such as 2010 monthly sales data for the available 13 Pawdacity stores, 2010 census population, as well as demographic data, including households with under 18, population density, land area and total families for each city in the state. A dataset with competitors' store sales will help refine the model.

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:

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 IQR method, we can identify the following outliers:

	Land Area	Households with Under 18	Population Density	Total Families	Total_Sales	2010_Census
Q1	1.861,72	1.327,00	1,72	2923,41	$ 226.152,00	7.917,00
Q3	3.504,91	4.037,00	7,39	7380,81	$ 312.984,00	26.061,50
IQR	1.643,19	2.710,00	5,67	4457,40	$ 86.832,00	18.144,50
Upper Fence	5.969,69	8.102,00	15,90	14066,9	$ 443.232,00	53.278,25
Lower Fence	-603,059765	-2738	-6,785	-3762,6825	$ 95.904,00	-19299,75
Outlier(s)	**1**	**0**	**1**	**1**	**2**	**1**
Min	999,50	746,00	1,46	1.744,08	$ 185.328,00	4.585,00
Max	6.620,20	7.788,00	20,34	14.612,64	$ 917.892,00	59.466,00

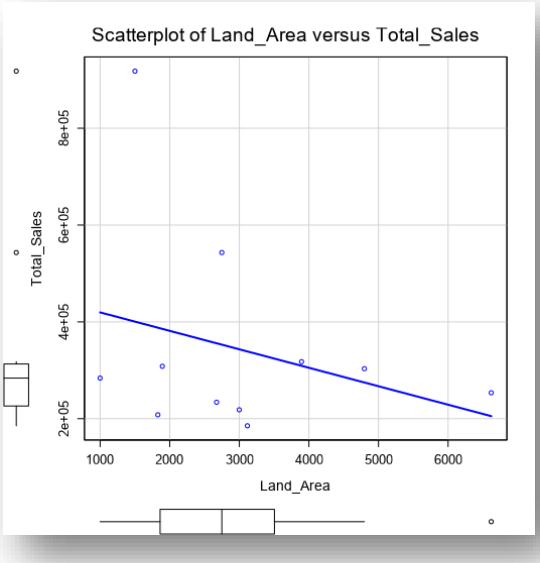
More specifically, both the table below and the whisker plots show that extreme outliers are present for the variables “Population Density” and “Total_Sales”.

City	Land Area	Households with Under 18	Population Density	Total Families	Total_Sales	2010_Census
Buffalo	3.115,507500	746,00	1,55	1.819,50	$ 185.328	4.585,00
Douglas	1.829,465100	832,00	1,46	1.744,08	$ 208.008	6.120,00
Powell	2.673,574550	1.251,00	1,62	3.134,18	$ 233.928	6.314,00
Cody	2.998,956960	1.403,00	1,82	3.515,62	$ 218.376	9.520,00
Riverton	4.796,859815	2.680,00	2,34	5.556,49	$ 303.264	10.615,00
Evanston	999,497100	1.486,00	4,95	2.712,64	$ 283.824	12.359,00
Sheridan	1.893,977048	2.646,00	8,98	6.039,71	$ 308.232	17.444,00
Rock Springs	6.620,201916	4.022,00	2,78	7.572,18	$ 253.584	23.036,00
Gillette	2.748,852900	4.052,00	5,80	7.189,43	$ 543.132	29.087,00
Casper	3.894,309100	7.788,00	11,16	8.756,32	$ 317.736	35.316,00
Cheyenne	1.500,178400	7.158,00	20,34	14.612,64	$ 917.892	59.466,00

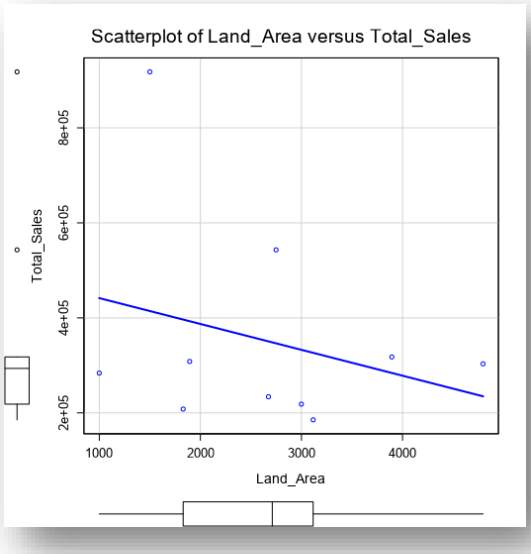
Rock Springs and Gillette have respectively one outlier. Interestingly, the city Cheyenne alone has 4 outliers out of 6. This latter observation might suggest that these datapoints are accurate and reflect an abnormal situation deriving from its high-density population, family and census figures. For this reason, we are going to retain Cheyenne. Conversely, Gillette skews high in

sales, but it is aligned to other variables in the training set which is peculiar, given the under proportional population values. For this reason, we are going to exclude it from our model.
Coming to the last outlier for the variable Land_Area related to the city Rock Springs, we can see that both scatter plots below show a negative trending line, therefore we can retain it as it won't affect the model in a considerable way.

Scatterplot Land_Area vs. Total_Sales with the city Rock Springs:



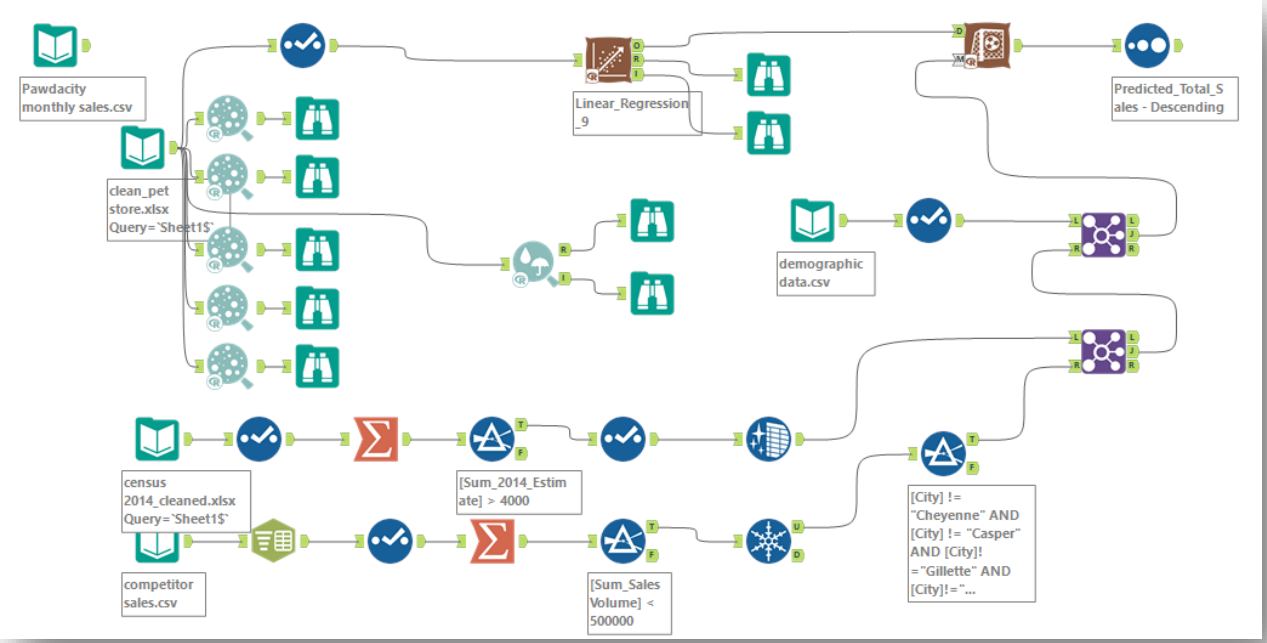
Scatterplot Land_Area vs. Total_Sales without the city Rock Springs:



Step 4: Linear Regression

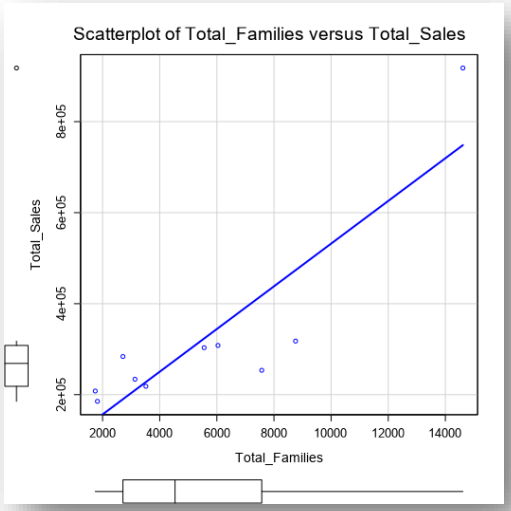
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section.
Important: *Make sure you have dealt with outliers.*

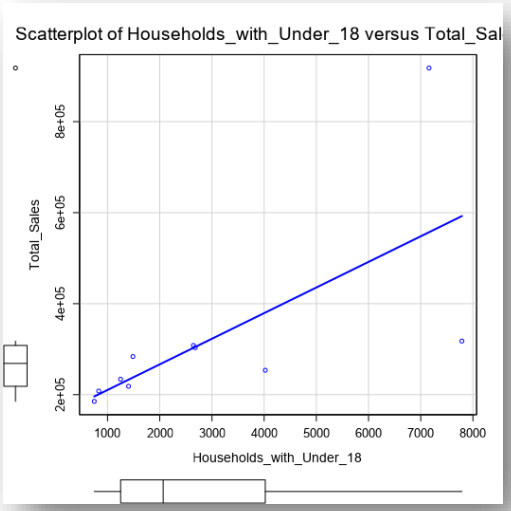
Build a linear regression model to help you predict total sales.

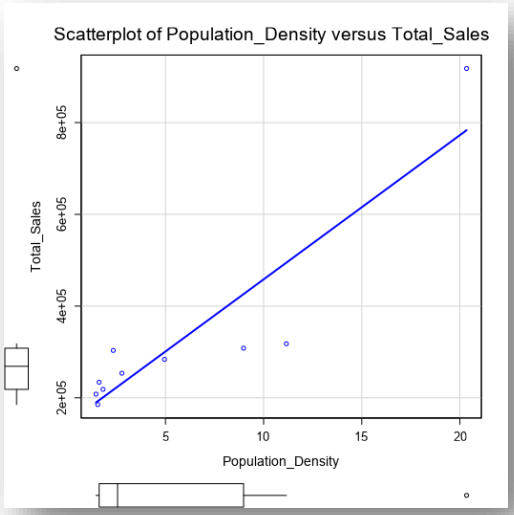


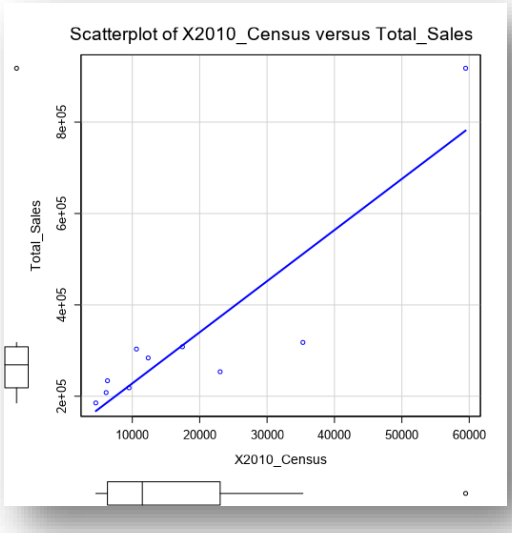
At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.
To select the best predictor variables, we should analyze the correlation between our target variable (Total_Sales) and the predictors as well as the measure of association of the predictor variables within each other. We have already seen that there is a linear relationship between Land_Area and Total_Sales. Let's proceed by plotting the remaining predictor variables against the target variable:





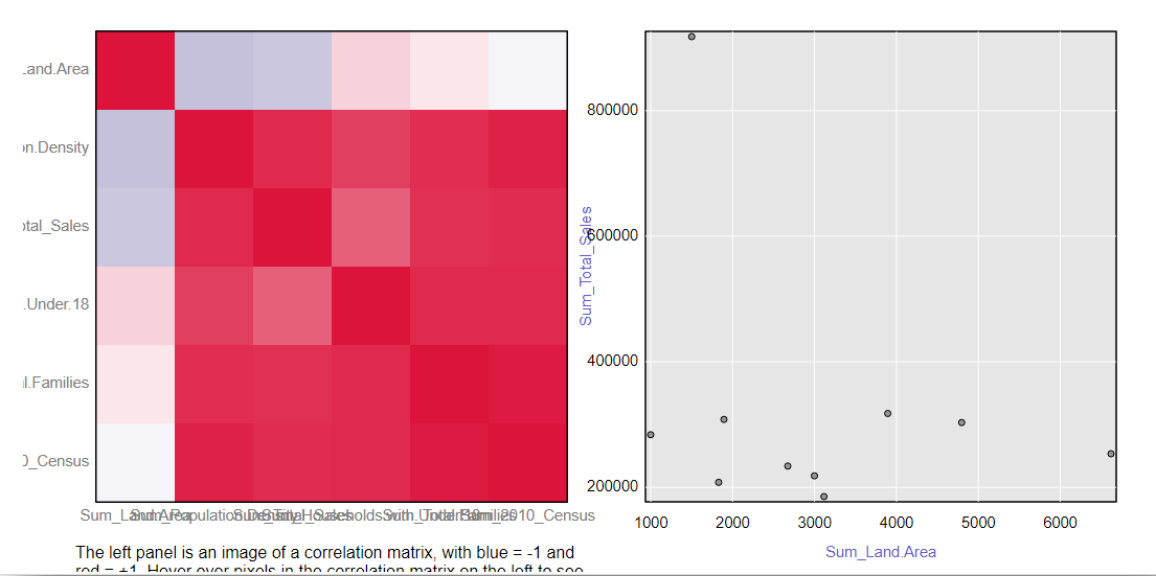




All variables represent a good predictor candidate since we can identify a positive linear relationship between them. Let's now look at the inter-correlations between these predictor variables. From the correlation matrix and the correlation table we can observe that almost all variables are highly inter-correlated with the only exception of the predictor Land_Area.

Full Correlation Table

	Total Sales	Households. with.Under.18	Population. Density	Land. Area	Total. Families	2010_Census
Total_Sales	1.00000	0.67465	0.90618	-0.28708	0.87466	0.89875
Households.with. Under.18	0.67465	1.00000	0.82199	0.18938	0.90566	0.91156
Population.Density	0.90618	0.82199	1.00000	-0.31742	0.89168	0.94439
Land.Area	-0.28708	0.18938	-0.31742	1.00000	0.10730	-0.05247
Total.Families	0.87466	0.90566	0.89168	0.10730	1.00000	0.96919
2010_Census	0.89875	0.91156	0.94439	-0.05247	0.96919	1.00000



By testing different combinations, we arrive at the best model which includes Land_Area and Total_Families as predictor variables.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared

values that your model produced.
After cleaning the remaining competitor file, we can join the datasets to create a regression model. The model displays the following results:

Residuals:

Min	1Q	Median	3Q	Max
-121261	-4453	8418	40491	75205

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197330.41	56449.000	3.496	0.01005	*
Land.Area	-48.42	14.184	-3.414	0.01123	*
Total.Families	49.14	6.055	8.115	8e-05	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72030 on 7 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866
F-statistic: 36.2 on 2 and 7 degrees of freedom (DF), p-value 0.0002035

Type II ANOVA Analysis

Response: Total_Sales

	Sum Sq	DF	F value	Pr(>F)	
Land.Area	60473052720.43	1	11.66	0.01123	*
Total.Families	341673845917.83	1	65.85	8e-05	***
Residuals	36318449406.44	7			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The significance codes listing the alpha levels for Land_Area and Total_Families show that we can reject the null and prove that there is a significant relationship between the two variables. The adjusted R (.8866) is close to 1 which generally suggests that nearly all variance in the target variable is explained by the model.

1. What is the best linear regression equation based on the available data?
Note: Each coefficient should have no more than 2 digits after the decimal.
Y = 197,330 – 48.42 * [Land_Area] + 49.14 * [Total_Families]

Step 5: Analysis

Use your model results to provide a recommendation.

1. *Which city would you recommend and why did you recommend this city?*

By filtering out the cities where Pawdacity was not yet present and by applying our model to them, we can recommend the city of Laramie as next opening location. The city displays the highest level in terms of predicted sales ($305,014) and satisfies all the given criteria.

City	Predicted_Total_Sales
Laramie	$ 305,013.88
Jackson	$ 225,870.82
Lander	$ 225,751.40
Worland	$ 201,700.33