Project 2.2: Recommend a City

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/3d606c26-cb8e-43af-9199-7e3577aa3392/project#

Note that this project is a continuation from Project 2.1: Data Cleanup. You must meet specifications for Project 2.1 before you can continue on with this Project 2.2

Step 1: Linear Regression

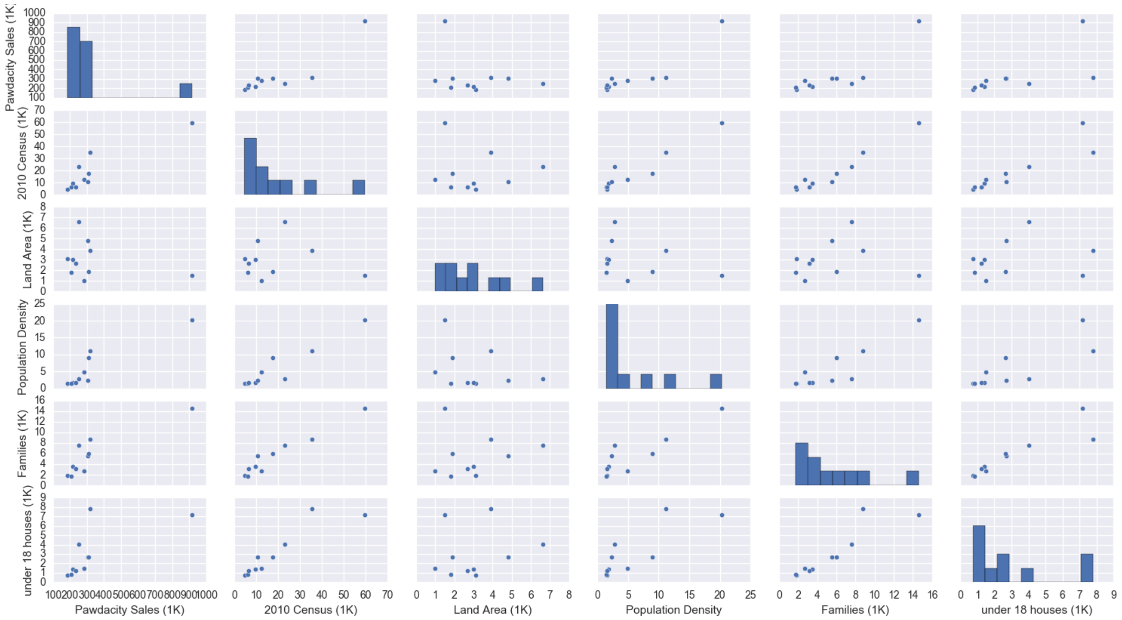
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

Important: *Make sure you have dealt with outliers and removed one city from your training set. You should have **10 rows** of data before you begin modeling the dataset.*

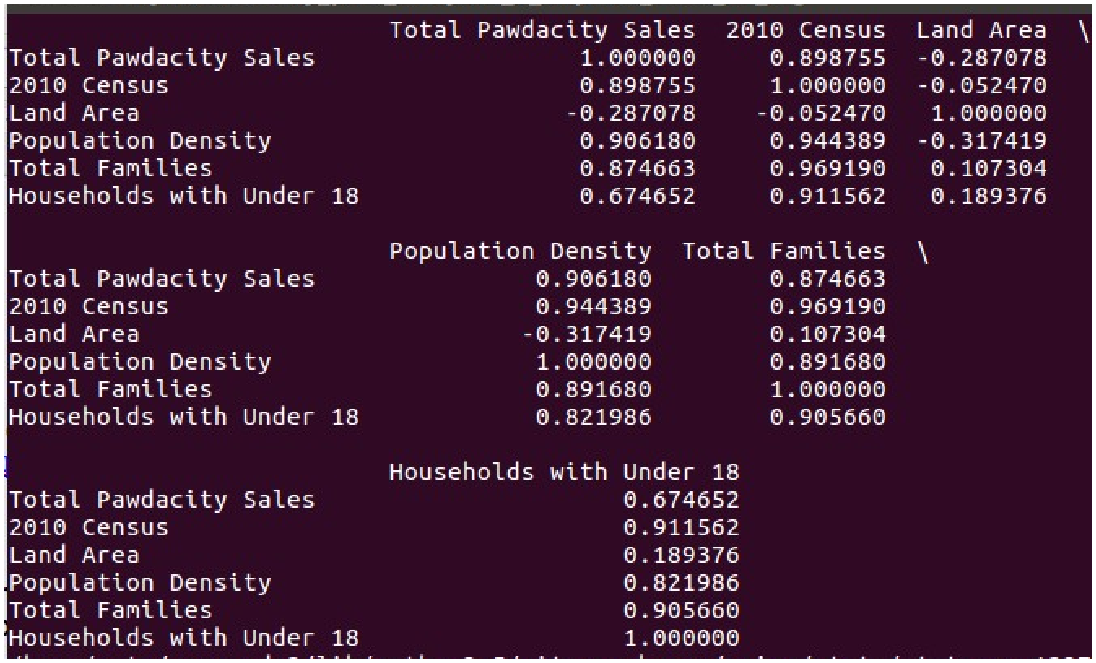
Build a linear regression model to help you predict total sales.

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.



I looked at plots, and eyeballed which ones looked somewhat linear. I also looked at the pearson correlation matrix for which independent variables had high correlation to sales:



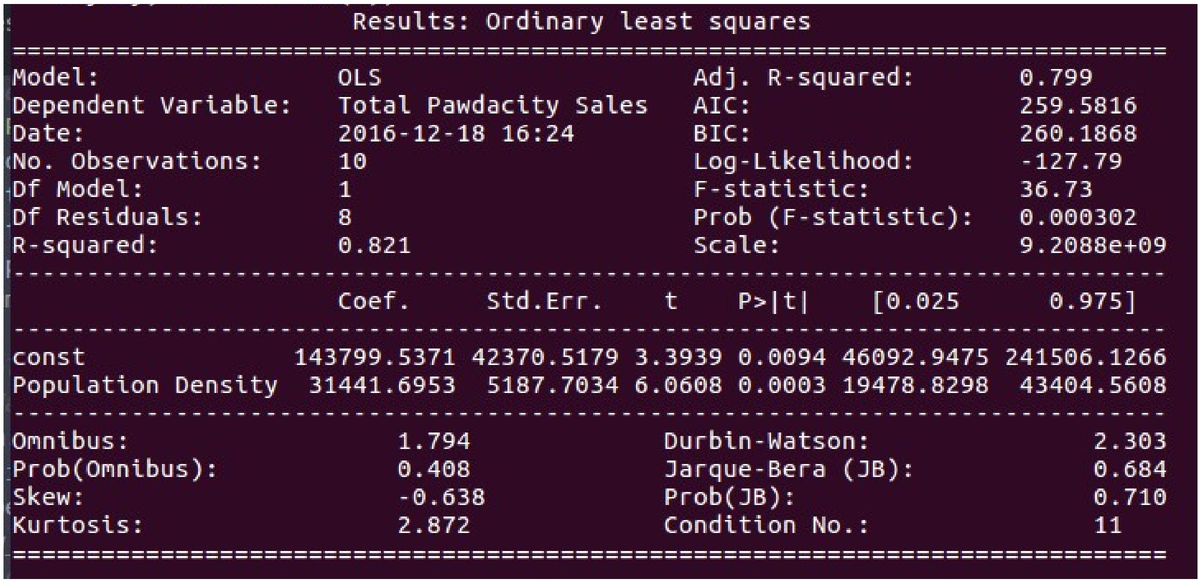
```
                          Total Pawdacity Sales  2010 Census  Land Area  \
Total Pawdacity Sales                  1.000000     0.898755  -0.287078
2010 Census                            0.898755     1.000000  -0.052470
Land Area                             -0.287078    -0.052470   1.000000
Population Density                     0.906180     0.944389  -0.317419
Total Families                         0.874663     0.969190   0.107304
Households with Under 18               0.674652     0.911562   0.189376

                          Population Density  Total Families  \
Total Pawdacity Sales               0.906180        0.874663
2010 Census                         0.944389        0.969190
Land Area                          -0.317419        0.107304
Population Density                  1.000000        0.891680
Total Families                      0.891680        1.000000
Households with Under 18            0.821986        0.905660

                          Households with Under 18
Total Pawdacity Sales                     0.674652
2010 Census                               0.911562
Land Area                                 0.189376
Population Density                        0.821986
Total Families                            0.905660
Households with Under 18                  1.000000
```

I chose Population Density as the predictor variable. Population measures (census, pop density, and total families, households under 18) were highly correlated, and so I only used the Population Density as a population measure, because it had the highest correlation to sales. I first used Population Density and Land Area as predictor variables, but found the p-value for Land Area from the fit to be 0.1267, meaning it probably doesn’t have a meaningful correlation to sales. We can also see this in the small pearson coefficient for Land Area/Sales.

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.



```
                      Results: Ordinary least squares
=====================================================================
Model:                OLS                    Adj. R-squared:      0.799
Dependent Variable:   Total Pawdacity Sales  AIC:                 259.5816
Date:                 2016-12-18 16:24       BIC:                 260.1868
No. Observations:     10                     Log-Likelihood:      -127.79
Df Model:             1                      F-statistic:         36.73
Df Residuals:         8                      Prob (F-statistic):  0.000302
R-squared:            0.821                  Scale:               9.2088e+09
---------------------------------------------------------------------
                      Coef.     Std.Err.    t    P>|t|    [0.025      0.975]
---------------------------------------------------------------------
const              143799.5371 42370.5179 3.3939 0.0094 46092.9475 241506.1266
Population Density  31441.6953  5187.7034 6.0608 0.0003 19478.8298  43404.5608
---------------------------------------------------------------------
Omnibus:                1.794          Durbin-Watson:             2.303
Prob(Omnibus):          0.408          Jarque-Bera (JB):          0.684
Skew:                   -0.638         Prob(JB):                  0.710
Kurtosis:               2.872          Condition No.:             11
=====================================================================
```

It's a good model because the adjusted r-squared is relatively high (0.799) and the p-values of the coefficients are below 0.05, meaning they have a meaningful relationship to the target variable.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)
   sales =
   Population Density * 31,441.70
   + 143,799.54

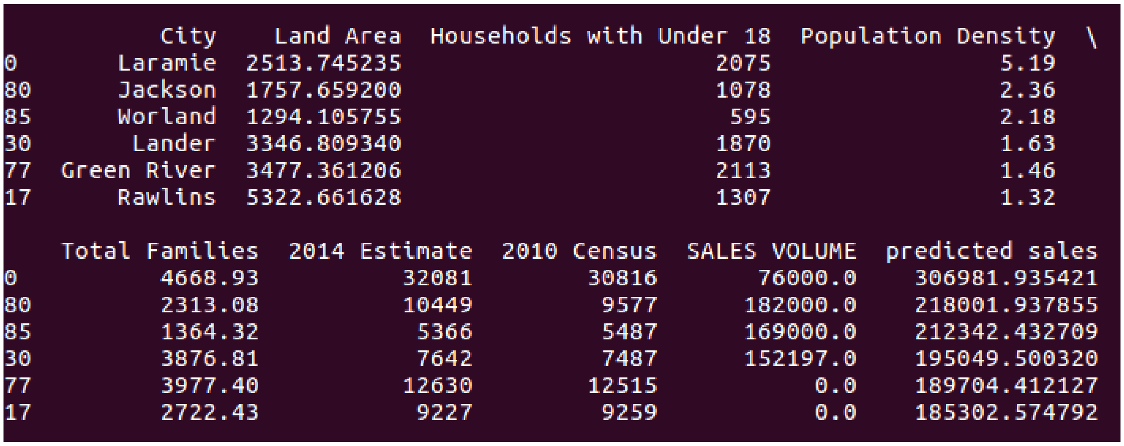
Step 2: Analysis

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer this question:

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

   Laramie looks to be the best City for a new store. It has the highest predicted sales based on the available data, and meets all the criteria (more than 4000 people, etc). Makes sense because it is the most populated City with low competition.



```
          City    Land Area  Households with Under 18  Population Density  \
0      Laramie  2513.745235                      2075                5.19
80     Jackson  1757.659200                      1078                2.36
85     Worland  1294.105755                       595                2.18
30      Lander  3346.809340                      1870                1.63
77 Green River  3477.361206                      2113                1.46
17     Rawlins  5322.661628                      1307                1.32

    Total Families  2014 Estimate  2010 Census  SALES VOLUME  predicted sales
0          4668.93          32081        30816       76000.0    306981.935421
80         2313.08          10449         9577      182000.0    218001.937855
85         1364.32           5366         5487      169000.0    212342.432709
30         3876.81           7642         7487      152197.0    195049.500320
77         3977.40          12630        12515           0.0    189704.412127
17         2722.43           9227         9259           0.0    185302.574792
```