Data Science
Chapter 2 & 3 Exercise

Student:
Solmaz Mohammadi

Teacher:
Prof. Faraji

Title:
Manhattan Sales Dataset Cleaning, Linear regression and KNN

CHAPTER 2 EXERSICE

Looking at realdirect.com

Part 1.

- What data would you advise the engineers log and what would your ideal datasets look like?

  I guess analysis of a dataset that has the information about houses that been sold such as, the price, the location, the size of the house and special services can be useful.

- How would data be used for reporting and monitoring product usage?

  It can be used in recommending customers things they're possibly Interested. Also, it would help predicting a house's price by certain key elements.

- How would data be built back into the product/website?

  It can constantly be in contact with website and at the same time evaluating the results of data analysis. And simultaneously, improve website's performance.

Part 2.

- Cleaning up the dataset.

  It's important to clean our dataset before applying training algorithm. Because it can cause incorrect results if we don't.

  Looking at the dataset, there are a total of around 27,000 rows and 21 columns. The dataset holds the record of different house sales in Manhattan. Before cleaning, I reformatted the datatype of columns to make following processes easier.

  First, I computed the number of missing values which were in total of 45722. It means that 45722 of values are NULL (or NAN). This amount of null values cannot be removed because the necessary data will be lost.

  I tried to improve dataset with dropping some of the unnecessary columns. For example, the "EASE-MENT" column had redundant values of "NULL" in most rows. Also "APARTMENT NUMBER" column had a total of 14570 null values. Since around 80% of the null values are in this column, it's better to drop it rather in removing rows.

  After that, only 5% of values are null. It's better to remove the rows which results dataset to have 23818 rows and 18 columns. Also, I decided to remove two "residential units" and "commercial units" columns and keep only "total units" column which is a derivative (summation) of them.

- Finding Outliers: There are some outliers in dataset, mostly in "year built" , "gross square feet", "land square feet" and "sale price". I detected most of them are zero values which doesn't make sense. I tried to show some of them on plots. But I couldn't plot all of them properly.

Here is the “year.built” column histogram using “ggplot”:

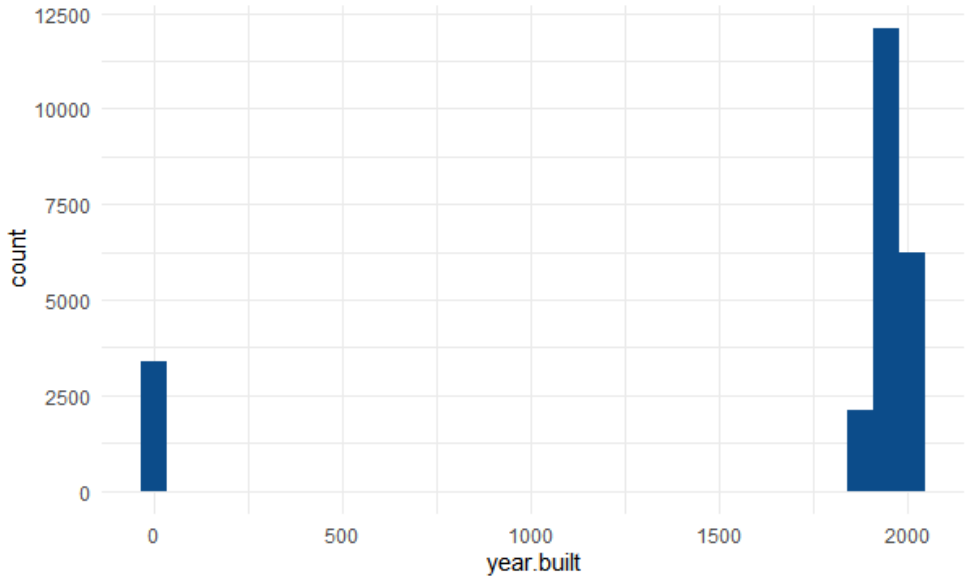

Figure 1. Year Built in Manhattan Sales Dataset

As you see, there are nearly 4000 zero values for year built. I didn’t remove the data because it consists a lot of data and the valuable information may be lost too.

The book code creates a dataset called “mt.sale” which separates actual sales data from the main dataset. And then makes another one which is categorized specially for “Family” homes. The homes subset has fewer data (300 rows). In below there are scatter plots for the house gross square feet versus sale price, before and after removing outliers. (figure.2 and figure.3)

Figure2. Scatter plot gross.square.feet vs. sale.price with outliers

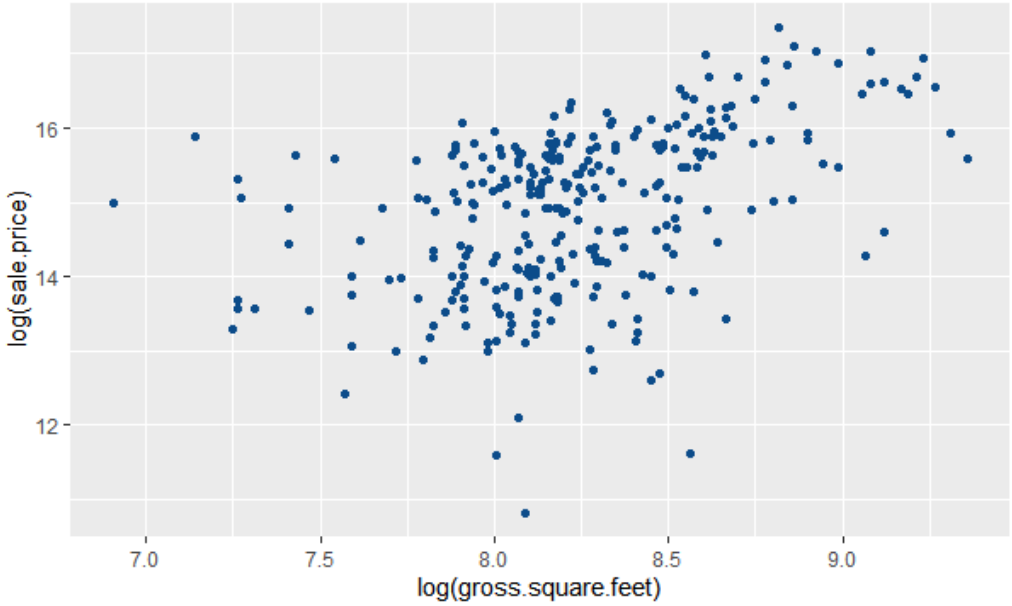

Figure3. Scatter plot gross.square.feet vs. sale.price without outliers

Part3. Summarize your findings in a brief report aimed at the CEO.

Looking at the dataset, there are some missing values and outliers that may prevent studies to be accurate enough. However, before applying algorithms we can see that there certain categories such as rentals or family homes follow their own pattern of prices. And we know that in New York, the combination of borough + lot + block is unique for every house. That can be useful when applying algorithms.

Part4. Being the "data scientist" often involves speaking to people who aren't also data scientists, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?

It is important to know about the dataset you are working on. So a kind of research about the Manhattan addressing system or people's financial situation in that area can be useful. Although no one was available with some extra information around me, but I used internet to get the data I wanted.

Part5. Most of you are not "domain experts" in real estate or online businesses. Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insight into how you do it in your own field?

Working with this data taught me how important it is I data science that you have the correct data to work with and how much it can impact your final results.

- Sometimes "domain experts" have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn't understand ("comps," "open houses," "CPC")?

Even when I didn't know all of the expressions, I could understand most of them and get the main idea of what dataset is about.

Part6. Doug mentioned the company didn't necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain.

The dataset had a large number of records which is a good thing, but most of them were incomplete and it consisted a lot of missing values. Therefore, it would not be the perfect dataset to work on if we want serious results. I would recommend saving valuable information with more precision.

.

CHAPTER 3 EXERSICE Linear regression and k-nn classifier

Continue with the NYC (Manhattan) Housing dataset you worked with in the preceding chapter:

- Analyze sales using regression with any predictors you feel are relevant. Justify why regression was appropriate to use.

  First, I computed the correlation matrix for numeric values in dataset to see which features are most affective on sale price.
  I plotted the matrix as below:

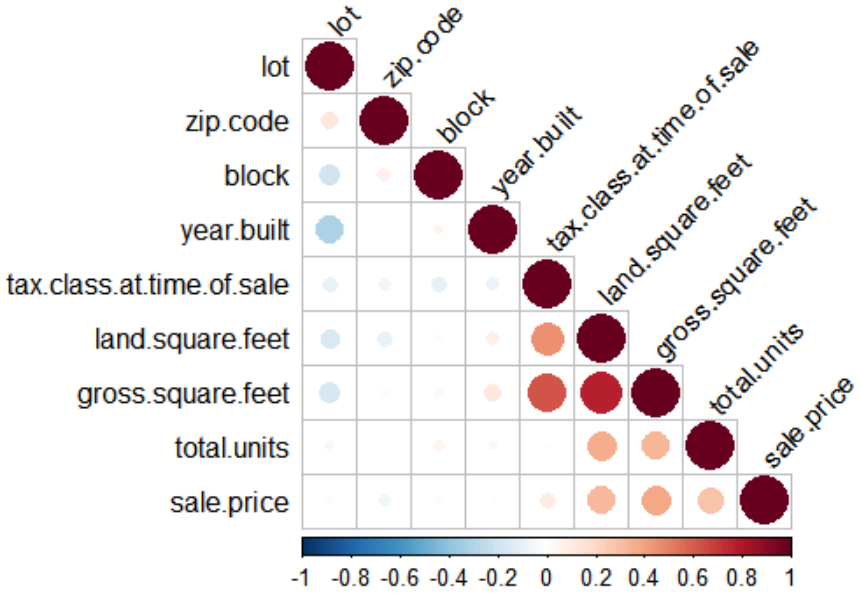

Figure4. Correlation matrix visualization

As you see in the plot, gross square feet and total units have the most effect on the price of the house. So, it's better to use the for linear regression.

I tried various formulas for the regression and they had mostly the same results. I decided to go with the one with the best results possible.

- Visualize the coefficients and fitted model.

One of the best possible fitted models is only log of gross square feet for sale price. Here, all three of these models have that in common.

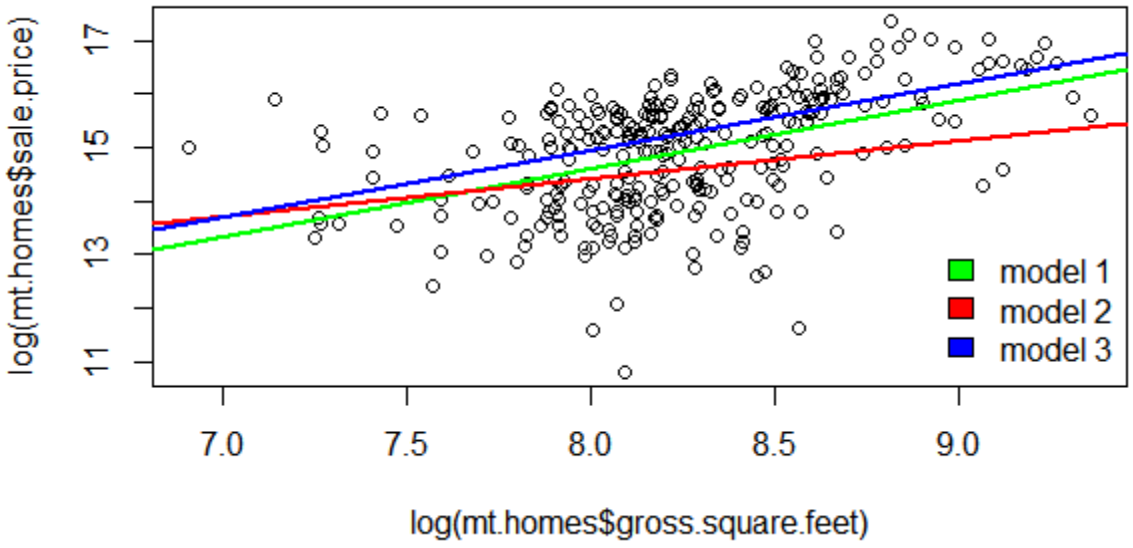

Figure5. Linear models

I built three different models for “mt.homes” (which is a cleaner dataset with more accurate data). The first model was simply log of gross square feet. In the second model, neighborhood is added as a factor. And the third model is the summation of gross square feet and total units. The first two models were from the book but I wrote the third one myself, based on the correlation matrix.

All three had low p=values and low standard error which makes all of them good models.

But the model No.2 also had high r-squared value (70%) when the other two had much lower percentages. Model 2 was one of the book’s models which uses neighborhood as a factor for fitting model. I’m not sure why It worked better. I was expecting that with total units along gross square feet would make the best model. Maybe it was because of the small number of data rows being trained.

- Predict the neighborhood using a k-NN classifier. Be sure to withhold a subset of the data for testing. Find the variables and the *k* that give you the lowest prediction error.

  Before applying the knn classifier, I looked at the neighborhood data classes

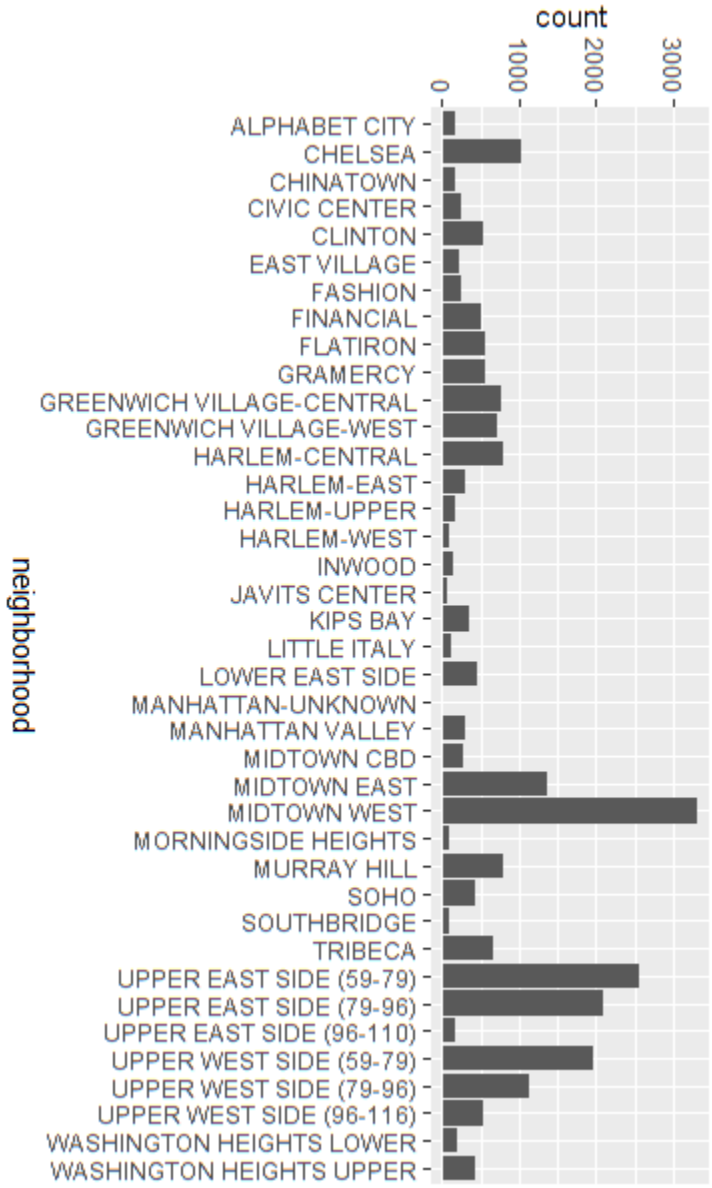

Figure6. Neighborhoods

If you notice, there are some classes that can be a subclass of a larger group. For example, there are four categories for Harlem neighborhood. It would be much easier for the classifier if we would consider these little similar groups as one. Look at the data after applying this idea:

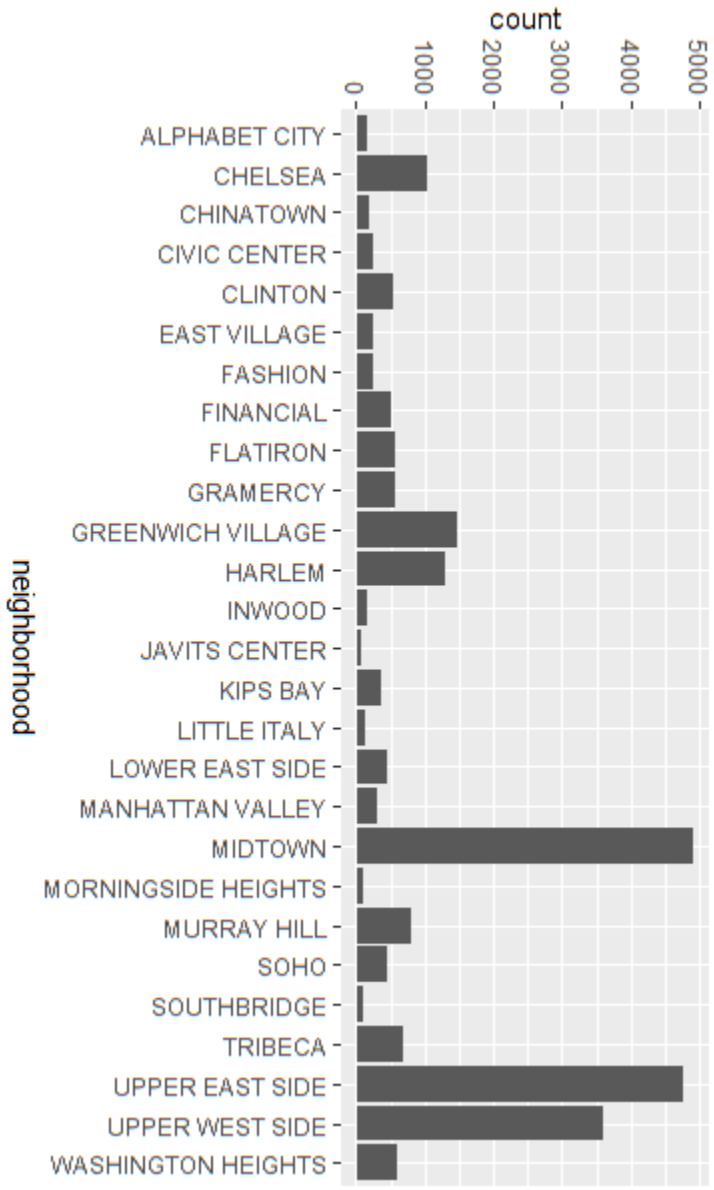

Figure7. Neighborhoods (after cleaning)

The number of classes has reduced from 39 to 27. Now the question is that what features can be used for classifying neighborhoods? Well, I think that

addresses are important. There two columns with numeric addresses: block and lot.
I tried using only these two features but I didn't seem to work proper. So, I decided to add some other features. I thought that zip code is also location related and can be useful. Also, the sale price can be different in neighborhoods.
I divided the 80% of dataset as training subset, applied k-nn with different k values and none of them got me higher that 60% accuracy. I realized the impact of feature aren't equal and sale price column has much bigger values. Therefore, I decided to normalize the features. Just normalizing the dataset, got me to 98% accuracy.

- Report and visualize your findings.

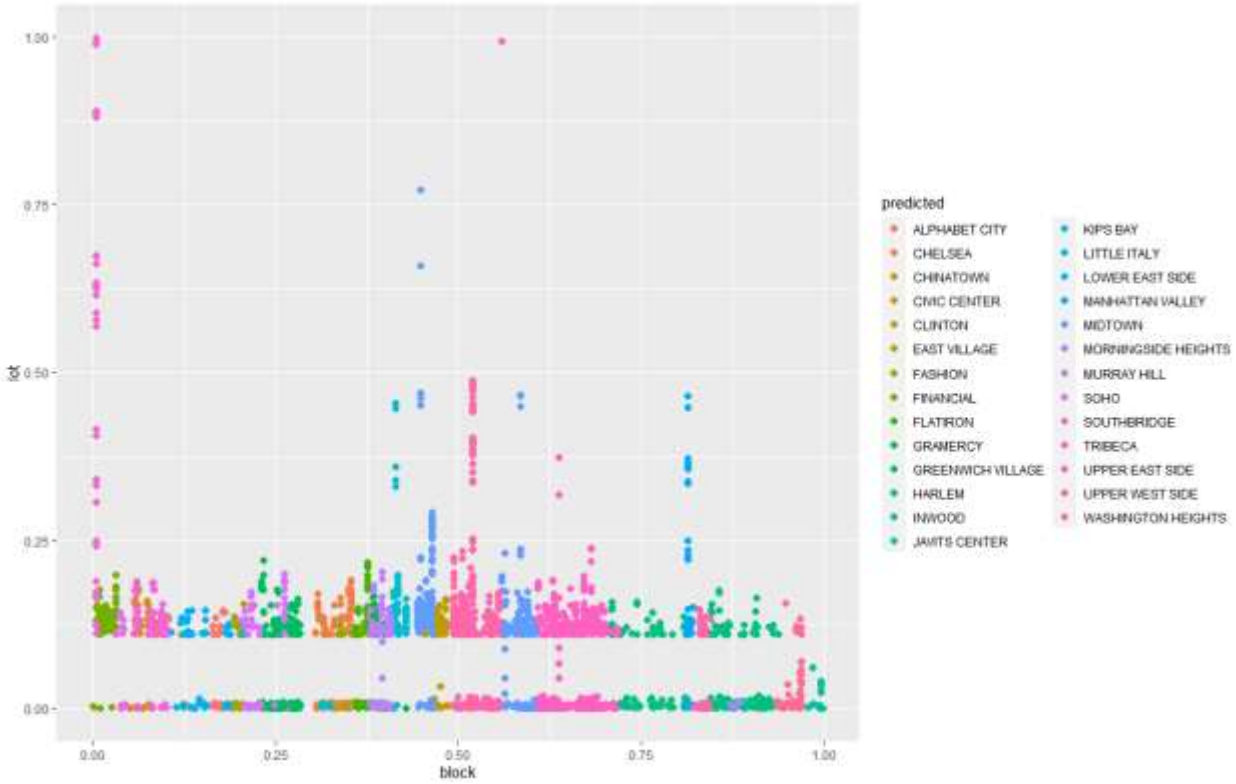

Figure6. K-NN predictions scatter plot

Here is a scatter plot for predicted values based on given features to k-nn classifier. It can be seen now that block has such a strong relation to neighborhoods in the dataset. As if the predictions can only be based on the block feature.

- Describe any decisions that could be made or actions that could be taken from this analysis.

  I think the important thing about this exercise was how our assumptions about data can be a lot different than the results. It's a process of uncertainty and it can lead to interesting ideas.

Sources

- Doing Data Science text book
- Class slides
- https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c
- https://www.pluralsight.com/guides/cleaning-up-data-from-outliers
- http://www.sthda.com/english/wiki/correlation-matrix-an-r-function-to-do-all-you-need