COM SCI-X 450.2 Exploratory Data Analysis and Visualization Final Assignment

Instructions: In this final assignment, you will be bringing together and applying various topics from our exploratory data analysis and visualization course. This final assignment is worth 20% of the course grade. By completing and submitting this final assignment, you are certifying that you worked independently on the assignment and that you did not give nor receive aid on the final assignment.

You will need to use R for this assignment. You may refer to code from lecture and past assignments. You may also refer to notes. Upon completion, please submit (1) a single .doc, .docx, or .pdf file containing your responses to the final assignment, and (2) an annotated copy of your R-code that can be run cleanly through starting from the original dataset. You may optionally upload .csv files if you did any data cleaning outside of R, but you must outline the details of your data cleaning step-by-step.

Details about the final assignment datafiles and prompt can be found further below. Be sure to check the Canvas assignment page for the grading rubric.

Datasets:

housing.csv – dataset of houses sold in 2019

- neighborhood – name of the neighborhood, indicator of a general area in the city
- beds – number of bedrooms in the unit
- baths – number of bathrooms in the unit
- sqft – unit square footage
- lotsize – unit's lot size
- year – year that the unit was built
- type – unit type
- levels – how many floors are in the unit
- cooling – whether or not the unit has cooling
- heating – whether or not the unit has central heating
- fireplace – whether or not the unit has a fireplace
- elementary – unit's assigned elementary school
- middle – unit's assigned middle school
- high – unit's assigned high school
- soldprice – selling price of the home

schools.csv – dataset of school rating

- school – name of the high school
- size – approximate student population size
- rating – school rating on a 1 to 10 scale

COM SCI-X 450.2 Exploratory Data Analysis and Visualization Final Assignment

A small, private housing organization is looking to expand on opportunities in new neighborhoods. Although the organization cannot disclose much about their project, they would still like an outside opinion on housing market patterns in certain neighborhoods.

The organization has hired you as a consultant to analyze a subset of their data and provide recommendations for areas of potential growth. The organization has a mix of people with varying data science backgrounds. The ones who will be reading your report range from somewhat familiar with data science to completely unfamiliar with data science. They have asked you to make sure that your report is clear to all readers. The organization would also like an annotated copy of your R-code in case they need to re-run analyses.

Generate a data report and summary of the given dataset. Be sure to explicitly address the following items in your report:

1. Data summary, oddities, and outliers
    a. What are all of the oddities and outliers in the dataset?
    b. How do you know?
    c. How do you plan to address oddities and outliers?
2. Data cleaning
    a. What did you change/remove from the original dataset? Why?
    b. Did you perform any merges?
3. One-variable visuals
    a. There are multiple variables to work with and multiple visuals you can use. Pick out some interesting ones to highlight and talk about. Be sure to clearly describe your observation that that someone can follow even without seeing the graph.
    b. Include at least one histogram
    c. Include at least one bar plot (of a different variable from the histogram)
    d. Include at least one box plot (of a different variable from the bar plot and histogram)
4. Two-variable visuals
    a. There are multiple variables to work with and multiple visuals you can use. Pick out some interesting ones to highlight and talk about. Be sure to clearly describe your observation that that someone can follow even without seeing the graph.
    b. Include at least one scatter plot
    c. Include at least one high density plot
5. Analysis
    a. The organization did not provide you with any specific topic to investigate. They are interested in what patterns you find. What are some interesting findings?
    b. Include at least one regression result
    c. Include at least one clustering result
6. Sensitivity Analysis
    a. The dataset has some missing data. Re-run your entire analysis filling in the missing data using a different method than your original approach (from 1 and 2). For example, if you used listwise deletion, try single imputation or multiple imputation. Do not use pairwise deletion.
    b. What method did you use to fill in the missing data this time around?
    c. Compare and contrast your results from your original run of the analysis.