Spring 2023 **Lealand Morin**

University of Central Florida
College of Business

QMB 6912
Capstone Project in Business Analytics

Problem Set #2

Due Date: Sunday, 22 January 2023, at 11:59 PM.

A data engineer for several realtors has gathered relevant and appropriate information, and organized a dataset concerning 1,862 sales of detached houses. The data engineer scraped data from many real estate Websites and combined it with the records from the realtors to identify the houses that were available for rent during the year after each sale. Around one third of the houses were identified as rental units, while the other two thirds were not observed for rent and were presumably occupied by the owner. These data are contained in the file `HomeSales.dat`, which is available in the `Data` folder of the course repository and on the course Webpage under Module 2.

Each house in the dataset is a row, while the columns correspond to the variables whose names and definitions are the following:

Variable	Definition
`year_built`	the year in which the house was constructed
`num_beds`	the number of bedrooms in the house
`num_baths`	the number of bathrooms in the house
`floor_space`	the area of floor space in the house, in square feet
`lot_size`	the area of lot on which the house was built, in square feet
`has_garage`	an indicator for whether the house has a garage
`has_encl_patio`	an indicator for whether the house has an enclosed patio
`has_security_gate`	an indicator for whether the property is accessed through a security gate
`has_pool`	an indicator for whether the property includes a pool
`transit_score`	an integer to represent the convenience of transportation options
`school_score`	an integer to represent the quality of the schools in the county
`type_of_buyer`	a categorical variable to indicate the type of buyer, either “Owner-Occupied” or “Rental”
`price`	the price at which the home was sold

The first several variables are self-explanatory but some of the variables above warrant some description. The last two integers, `transit_score` and `school_score` describe the amenities available to the residents in each house. The `transit_score` is an integer from one to ten that indicates the convenience of transportation options at the location of the house. It factors in the availability of public transportation options, as well as the proximity to business districts and major highways, with a higher score indicating a higher level of convenience. Similarly the `school_score` is an integer from one to ten that indicates the quality of the schools in the county. It is compiled from indicators of college attendance, test scores, as well as other measures of school and student performance. As with the `transit_score`, a higher `school_score` indicates better school quality. The variables listed above also include the dependent variable. The `price` variable is the price at which the home was sold.

Begin your analysis by focusing on the dependent variable, `price`, including all observations together, regardless of whether the home is owner-occupied or used as a rental property. Calculate the relative histogram of home prices and plot a graph. Then, calculate the kernel-smoothed probability density function of home prices. Choose an appropriate bandwidth that strikes a balance between smoothness, for ease of interpretation, and flexibility, for fitting the data accurately. Now graph these objects in R and output the relevant graphs in a format that can be loaded into LaTeX for further processing.

Repeat the above procedures using the natural logarithm of home prices.

Next, recalculate the density of sale prices, this time disaggregating by type of owner. That is, create separate kernel-smoothed probability density functions for homes sold as rental properties and homes occupied by the owner. Note that you might find that different bandwidths are appropriate for both subsamples, so you might conduct the analysis separately for each subsample. Next, depict these two objects on the same graph, labeling them appropriately so the reader can distinguish among them.

Again, repeat the previous steps for the separate samples by type of owner using the natural logarithm of home prices as the dependent variable.

Describe any differences that you find between the densities by type of owner. Do the prices follow a different distribution for each type? Would you expect to find different distributions of the characteristics of homes? Would you expect to find different relationships between the characteristics of homes and their sale prices? We will investigate your conjectures through the next several problem sets.

Prepare and compile your work in LaTeX and include scripts for any of the calculations in R. In particular, create the following directory structure, separate from your existing work:

- `Code/`
- `Data/`
- `Figures/`
- `Paper/`
- `Misc/`

In a file called `README.md`, which should also live in the directory containing the above folders, provide the instructions concerning how to run the executable shell script `DoWork.sh` (in the same directory) that will execute the code that produced all of the answers collected and documented in your report, which will live in the subdirectory `Paper/`. In the subdirectory `Code/`, keep the R code; in `Data/` keep the raw data file you downloaded, so that `DoWork.sh` can load it into R, and in `Figures/` keep any figures you created for your answers. Put anything else in the subdirectory `Misc/`. I should then be able to replicate all of your work simply by typing

- `$ ./DoWork.sh`

on the command line of a terminal window.

To provide you a template, which makes preparation easier for you and grading easier for me, I have placed sample LaTeX and R code in the GitHub repository for the course: `QMB6912S23`, under my GitHub username `LeeMorinUCF`; pull this repository and use these files a framework within which to create the answers for this problem set. Push the files to a folder on your GitHub repository and I will pull your submissions to my computer for grading.

Be sure to support your calculations with descriptions of what you were trying to do (for example, in comments in your R code as well as in the LaTeX explanations) because partial credit will be given.

Due Date: Sunday, 22 January 2023, at 11:59 PM.