Spring 2023 **Lealand Morin**

University of Central Florida
College of Business

QMB 6912
Capstone Project in Business Analytics

Problem Set #4

Due Date: Sunday, 12 February 2023, at 11:59 PM.

Go to the course Webpage on Webcourses and find the file `Lancaster1966.pdf` in the `Files` tab;

download and read this document; then write a one-page summary of this paper, following the guidance provided in `ReadPaper.pdf`. Use this as a guide to think about how the characteristics of homes might be valued by different types of prospective buyers. Explain how Lancaster's characteristic theory can be used to put theoretical structure on the value of houses.

Analyze the data in `HomeSales.dat` in tabular form. The aim is to detect features of the data to gain insight into potential modeling approaches in future problem sets.

First, format some of the variables in `HomeSales.dat`. Create a variable `log_price`, which is the logarithm of the house prices. Next, create the feature variable `age` from the `year_built` variable and the year date 2021, when all of these data were gathered.

Then, analyze the data in subsets, according to type of buyer (homeowner-occupied vs. rental) calculating the summary statistics for each subset and presenting these statistics in a LaTeX table. Do the characteristics of homes sold differ between the two types of buyers?

Next consider tabulating the average sale price across levels of categorical variables for each type of buyer. Do the relationships appear to differ between the properties sold to each type of buyer? Suggest any implications for the potential differences in models to predict prices of each type of sale.

Investigate the covariance of pairs of variables. In particular, calculate a covariance matrix for two sets of variables. Calculate the first matrix for the numeric variables `log_price`, `age` `floor_space`, `lot_size`, `transit_score`, and `school_score`, and save it to a file for further processing in LaTeX. Calculate another covariance matrix and LaTeXtable for `log_price` and the indicator variables `num_beds`, `num_baths`, `has_garage`, `has_encl_patio`, `has_security_gate`, and `has_pool`. Comment on the findings for each and make recommendations for the variables that should be included in a regression model to predict `log_price`.

Then continue analyzing the data according to other categorizations that you might find relevant for the analysis of this dataset. Present these statistics in LaTeX tables as well. Creativity is rewarded both in terms of grading of this problem set and in the insight gained to guide future analysis.

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, 12 February 2023, at 11:59 PM.