STAT 378 – Term Paper
Due: Friday December 8 at 9:59 pm

IMPORTANT:

a) **No** late term paper is acceptable.
b) Term paper **must** be submitted on **eClass**.
c) The term paper **must** be typed. A handwritten assignment is not acceptable and it will receive a mark of zero for the whole assignment.
d) You should write an article about analysis of data. Your article should include the statement of the problem, conclusion and your methodology. It might be necessary to refer to some of computer outputs. **You will received zero marks if you only submit your computer outputs**.
e) Any computer output and code should be an appendix to your article. You **must** include the appropriate R codes in an appendix at the end of your paper.
f) You should always write your conclusions in **plain English**.
g) Although collaboration between groups is allowed, groups should work independently and any work turned in must be your own.
h) Your mark is evaluated based on correct and complete analysis.
i) All group members are **equally responsible** for ensuring the term paper is submitted on time and must check the report has been uploaded before the deadline.

Problem: This data set provides information for 141 large Standard Metropolitan Statistical Areas (SMSAs) in the United States. A standard metropolitan area includes a city (or cities) of specified population size which constitutes the central city and the county (or counties) in which it is located, as well as contiguous counties when the economics and social relationship between the central and contiguous counties met specified criteria of metropolitan character and integration. An SMSA may have up to three central cities and may cross state lines. The data set was given in Neter, Kutner, Nachtsheim, and Wasserman (1996). "Applied Linear Regression Models", IRWIN, pp. 1367–1368. You can also find the data set on eCLass.

Each line of data set has an identification number and provides information on 11 other variables for a single SMSA. The information generally pertains to the years 1976–1977 study period. The 12 variables are explained in the next page.

Split randomly the data in two parts, training group and holdout group. The last one should have 25 lines.

(a) For the training group, find the best regression model.

- Be sure to check all assumptions in the regression analysis.
- Use transformation if it is necessary.

(b) Use the holdout group to check the reliability of the model that you found in part (a). If the mode is not reliable, you should repeat part (a).

(c) Write an article about analysis of data.

Variable Number	Variable Name	Description of Variable
1	*Identification number*	1-141
2	*Land Area*	*In square miles*
3	*Total population*	*Estimated 1977 population (in thousands)*
4	*Percent of population in central cities*	*Percent of 1979 SMSA population in central city or cities*
5	*Percent of population 65 or older*	*Percent of 1979 SMSA population 65 years old or older*
6	*Number of active physicians*	*Number of professionally active nonfederal physicians as of December 31, 1977*
7	*Number of hospital beds*	*The number of beds, cribs, and bassinets during 1977*
8	*Percent high school graduates*	*Percent of adult population (person years old or older) who completed 12 or more years of school, according to the 1970 Census of Population*
9	*Civilian labor force*	*Total number of persons in civilian labor force (person 16 years old or older classified as employed or unemployed) in 1977 (in thousands)*
10	*Total personal income*	*Total current income received in 1976 by residents of the SMSA from all sources ,before deduction of income and other personal taxes but after deduction of personal contributions to social security and other social insurance programs (in millions of dollars)*
11	*Total serious crimes*	*Total number of serious crimes in 1977, including murder, rape, robbery, aggravated, assault, burglary, larceny-theft, and motor vehicle theft, as reported by, law enforcement agencies*
12	*Geographic region*	*Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W*