SPHU 4160

Exam Number 3 - R

Instructor: J. Yukich

Document compiled July 24, 2020

DUE: December 5$^{\text{th}}$, 2020

before 5:00pm

This exam requires you to use the R software, which is avaialbe in the lab or through self installation. The overall exam rules are:

S.1 Work is to be your own, you may use any source, except direct human sources (*e.g.* do not post exam questions to message boards, or ask classmates for help). If you have questions of clarification on the exam or relevant sources of information please contact the TA or instructor directly.

S.2 Deliverable is: An R notebook and the compiled output document (.Rmd file and the html_notebook output (nb.html) file)).

S.3 Script files must work. The Instructor and TA will alter FILE PATHS and install any required libraries, but no other changes will be made before grading.

S.4 **INCLUDE YOUR NAME** in the preable to all documents, the file names themselves, and in the reports. Failure to properly identify the person responsible for the work (see S.1) will result in loss of attribution of the work to the author.

S.5 Turn in the exam by the due date by pushing to your shared Git repositry on GitHub. Notify instructor and TA via e-mail, e-mail to: jyukich@tulane.edu with cc: to weaton@tulane.edu.

1 Background

This exam will require the use of a dataset on New Orleans crime statistics compiled by the instructor. It is called *nola_crime_2018.csv*. It is a subset of the dataset which corresponds to, and is avaialble from City of New Orleans Open Data - data.nola.gov joined with data for Orleans Parish from the U.S. Census Bureau. The Greater New Orleans Community Data Center (now The Data Center) has developed maps of neighborhoods in the city which can be used as areas of analysis for community information, a map is included as Figure 1 here to orient you to the specific neighborhoods discussed.

To complete the exam you will be requested to answer a series of questions about the contents of several datasets. To completely and accurately answer the questions, you will be required to load the datasets in R, possibly clean or tidy the datasets, and perform other operations as well as some basic statistical analyses and produce graphs/charts and write up of results. In addition, there may be a few short answer questions interspersed below that must be answered in the report or dynamic document that you turn in, but which may not require the use of the datasets but may require development of R code.

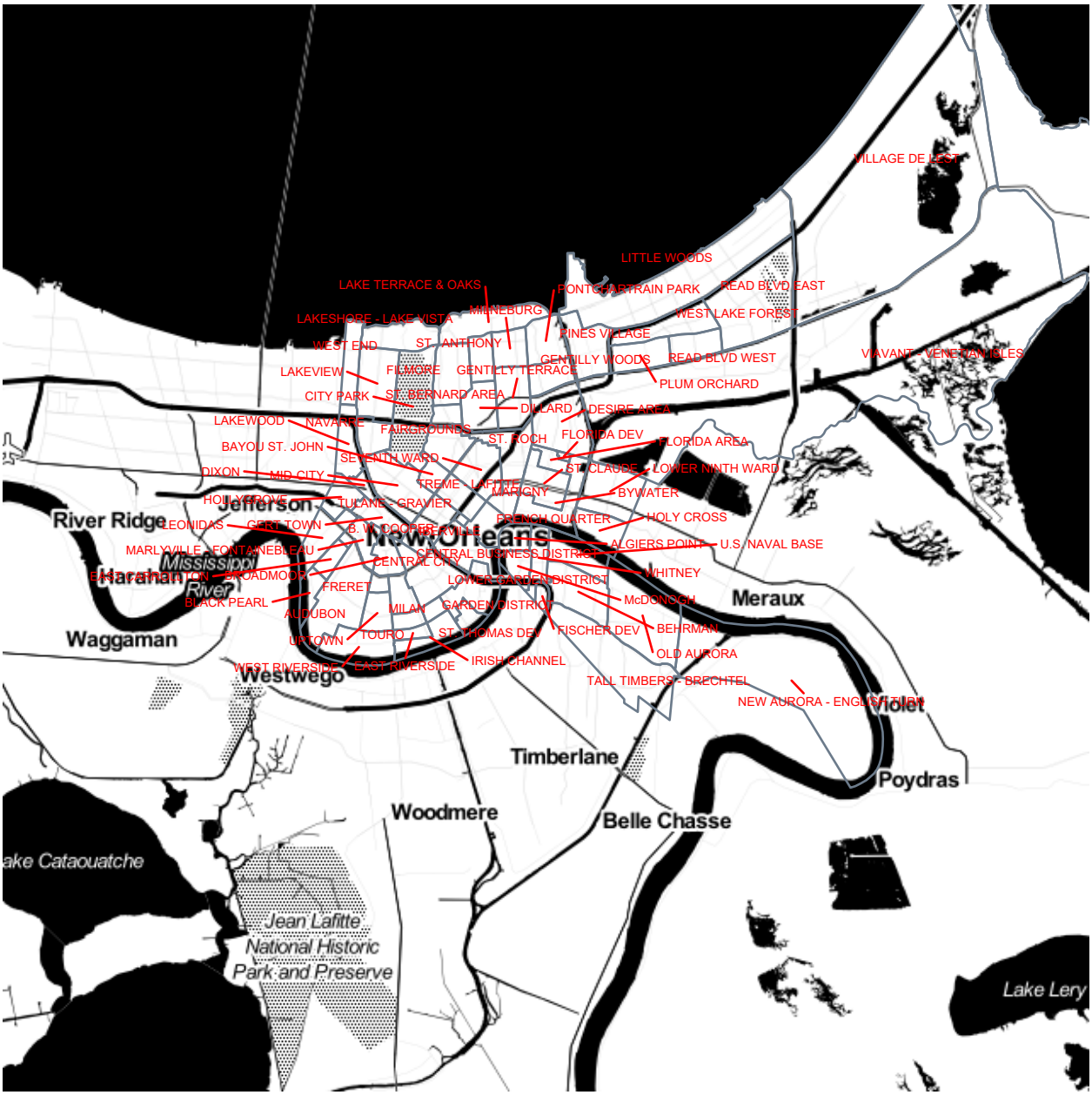

Figure 1: Greater New Orleans Community Data Center Neighborhood Statistical Areas

2 Description of datasets

1. The *nola_crime_2018.csv* consists of the date, item number, police district, location, address, disposition, charge and signal description and other information for a random sample of police reports that were filed in Orleans Parish, Lousiana during the year 2018. Details are shown in Table 1.

2. The *iris* dataset is included with base R, it contains a series of measurements taken on samples of three different species of iris and is described in more detail in Table 2.

3. The *mtcars* dataset is also included in base R, it contains data on a series of automobiles derived from a 1974 article in Motortrend magazine. It is described in more detail in Table 3.

Table 1: The nola_crime_2018.csv dataset

Variable Name	Type	Decription
Item_Number	character	Police Report Number
District	numeric	Police District
Location	character	Location
Disposition	character	Closed or Open Case
Signal_Type	character	Code for Report
Signal_Description	character	Description of offense generating report
Occurred_Date_Time	character	Date and time event occured
Charge_Code	character	Code for any charged crime
Charge_Description	character	Description of any crime charged
Offender_Race	character	Race of offender
Offender_Gender	character	Gender of offender
Offender_Age	numeric	Age of offender in years
Offender_Number	numeric	Line number of offender within report
Person_Type	character	Person providing data
Victim_Race	character	Race of victim
Victim_Gender	character	Gender of victim
Victim_Age	numeric	Age of victim in years
Victim_Number	numeric	Line number of victim within report
Victim_Fatal_Status	character	Did the reported crime result in fatality for victim
Hate_Crime	logical	Was offense classified as a hate crime

Table 1: The nola_crime_2018.csv dataset *(continued)*

Variable Name	Type	Decription
Report_Type	character	Type of police report
address	character	Imputed address of offense
GNOCDC_LAB	character	Greater New Orleans Community Data Center Neighborhood Name
pop	numeric	Total Population of Neighborhood

Table 2: The iris dataset

Variable Name	Type	Decription
Sepal.Length	numeric	Length of sepal (cm)
Sepal.Width	numeric	Width of sepal (cm)
Petal.Length	numeric	Length of petal (cm)
Petal.Width	numeric	Width of petal (cm)
Species	factor	Species name

Table 3: The mtcars dataset

Variable Name	Type	Decription
mpg	numeric	Miles per (US) gallon
cyl	numeric	Number of cylinders
disp	numeric	Displacement (cu.in.)
hp	numeric	Gross horsepower
drat	numeric	Rear axle ratio
wt	numeric	Weight (1000 lbs)
qsec	numeric	Quarter mile time
vs	numeric	Engine (0 - V-shaped, 1 - straight)
am	numeric	Transmission (0 - automatic, 1 - manual)
gear	numeric	Number of forward gears
carb	numeric	Number of carburetors

The next section of this document outlines the "questions" to be answered or successful data operations to be performed to complete the exam as well as associated points for each question or proceedure.

1. (5 points) How many observations are there in the *nola_crime_2018.csv* dataset?

2. (5 points) How many variables are there in the *nola_crime_2018.csv* dataset?

3. (5 points) How many unique item numbers are in the dataset?

4. (10 points) Considering only unique item numbers as crimes, how many "aggravated" crimes occured in the Irish Channel Neighborhood?

5. (10 points) What two neighborhoods have the largest numbers of crime reports (unique items)?

6. (10 points) Assume that the each person in the population contributes exactly one year of time at risk, calculate the rate of crime (unique items) per 1,000 persons in each neighborhood in the dataset for 2018, show this in a table in your report?

7. (15 points) What is the largest neighborhood by population in the dataset and what is its total population?

8. (10 points) What is the population size and number of unique crimes in the neighborhood with the lowest rate of crime (unique items) per 1,000 person years?

9. (10 points) Calculate the rate ratio for each neighborhood for all unique crimes treating LAKE WOOD as the reference (denominator)?

10. (10 points) Calculate the rate ratio for each neighborhood for all unique "aggravated crimes" treating FRENCH QUARTER as the reference (denominator)?

11. (10 points) What is the most common cause for a police report to be filed (captured in the "signal description" variable) and what is the most common charge that is filed?

12. (10 points) Calculate the rate of domestic disturbance (again based on "signal description") by neighborhood and show this as a table in your report.

13. (10 points) Fit a linear regression model to a summary dataset that you create which contains the rate of all crimes calculated by neighborhood as in previous calculation and includes the average age of the victims in that neighborhood as a predictor, include the summary of the regression results in your report by using the summary() function.

14. (10 points) Assume that high numbers of police reports happen in places with adequate police responsiveness but that true rates of criminal acts are similar everywhere. If this is true than the absolute numbers of crimes depend only on population size but the number of reports depends on both population size and crime rates. If these hypothesis are true, which neighborhood might have the worst police responsiveness (*i.e.* Low rates of crime reporting - *i.e.* low numbers of signals/reports and high population)?

3 Non-external dataset requiring questions

15. (10 points) Plot a histogram of 10,000 realizations of a Poisson distributed random variable with $\lambda = 4.7$.

16. (10 points) Using the "iris" dataset, write a for loop that calculates the mean Sepal length for each of the species of iris and prints the results.

17. (10 points) Using the "mtcars" dataset make a scatter plot of mpg vs. disp, include a regression line fit with a linear regression model or a loess smooth regression, make the color of the points different for each different number of cylinders in the engine.

18. (10 points) Write a function to convert a vector of continuous data into its Z-scores (standard normal deviates). The formula for a Z-score is $\frac{x-\mu}{\sigma}$ where x is the vector of data, μ is the mean of x and σ is the standard deviation of x. Use the function to calculate the Z-scores for this vector {-4.89, -1.93, -1.11, 3.94, 0.46, -3.85, -0.20, 6.04, 9.36, 7.26, 4.88, 13.45, -2.93, 6.39, -16.22, -3.24, 6.86, 11.87, 1.81, -2.54}.

4 General Bonus Questions

Note that the following questions are not required for this exam but will count for up to 5 percentage points of extra credit towards the course final grade. They can be completed in any language taught in the course, provided that the deliverable includes a fully reproducible script to produce the answers and write up to contextualize the response. They can also be turned in directly as part of the nb.html or .Rmd file for this exam. The second and third bonus questions will require use of the "sf" package in R, and plotting via ggplot or ggmap packages. The additional data required (map files for the neighborhood statistical areas) are stored as "shape" files (.shp) and can be loaded in R using the sf package or via using other libraries or other software packages "*shp2dta*" can be used to read the .shp files in STATA and PROC MAPIMPORT can be used in SAS. Plotting is usually made via "*tmap*" in STATA and PROC GMAP in SAS.

19. (10 points (bonus)) Using the "mtcars" dataset, first fit a linear regression model of mpg with the predictor being the displacement of the engine (*disp*). Then use the results of this to write a function that simulates the mpg expected for car with a given displacement. The functional form for this regression will be $y_{mpg} = \beta x_{disp} + \epsilon$. Note that the error term ϵ is considered to be normally distributed, centered on zero, with a standard deviation (σ) equal to the residual standard deviation of the regression model. The coefficients are fixed at those that are estimated in the regression model. The residual standard deviation of the model can be calculated in R using the function sigma() on the model object. Use the function to simulate the expected *mpg* at several (within sample) displacement values and plot the results as a scatter plot.

20. (10 points (bonus)) Using the *nola_crime_2018.csv* dataset and the *Neighborhood_Statistical_Areas.shp* file, make a choropleth map of the rate ratio for each neighborhood for all unique crimes treating LAKE WOOD as the reference (denominator).

21. (10 points (bonus)) Using the *nola_crime_2018.csv* dataset and the *Neighborhood_Statistical_Areas.shp* file, make a second choropleth map of the rate ratio for each neighborhood for all unqique "aggravated" crimes treating FRENCH QUARTER as the reference (denominator).

    Bonus questions are due on the same schedule as the main exam.

This exam has 21 questions, for a total of 170 points which contribute 25% of the total course grade and 30 bonus points which can contribute an additional 5% to the final course grade.