QMST 3339 Homework 1

Tahir Ekin[1]

[1]McCoy College of Business, Texas State University

Spring 2018

Guidelines

1. A hard copy (print-out) of your report that summarizes the results and includes the supporting graphs is due to the beginning of the lecture on 02/12/2018, Monday.
2. Do not include data in your report.
3. This is a group-based homework. However, all team members are expected to understand the assignment, relevant script and the submission thoroughly.
4. This is graded out of 5 points, and accounts for 5% of your overall grade. See the syllabus for more details.
5. Submission of any work, for which unauthorized help has been received, is termed as academic dishonesty and will be grounds for a failing grade in the course. See "Academic Honesty" section of the syllabus for more details.

1 Homework 1

All the questions below use the *pollution* data set, which can be called by using avgpm25 csv file. It includes data on daily pollution level (PM 2.5) which are available from the U.S. EPA web site. The variables are:

- pm25: pollution level
- fips: county code
- region
- longitude
- latitude

1.1 Descriptive statistics and basic visualization

3.2 pts (0.4 pts each unless otherwise noted)

a) Display first five observations of the data set.

b) Provide a five number summary for pm25 and comment on what each number refers to. 0.6 pts

c) Provide a boxplot of pollution level in red with the title "Boxplot for Particle Pollution" and with limits $(0, 20)$.

d) Provide a boxplot of pollution level with respect to each region, in red

e) Provide a histogram of pollution level with breaks of 100, in green. Include a rug and a line for the median of pollution level on that histogram in magenta. 0.6 pts

f) Provide a multiple scatterplot in the same plotting frame that shows the relationship between latitude and pollution level in west region, and the relationship between latitude and pollution level in east region. Use limits of $(0, 20)$ and titles of "West" and "East"

g) Construct two subsets of data, one that contain all observations of west region, second that contains all observations in east region. Assuming variances are not equal, run an independent samples t-test to check the equality of means of pollution level by using first 134 observations of each subset. Report the p value of the t-test and comment on its meaning. 0.6 pts

1.2 Advanced visualization

1.8 pts (0.4 pts each unless otherwise noted)

a) Replicate the figures 1 and 2, that are provided below, by using one of the R visualization packages. Present the code. 0.6 pts

b) Which package did you use?

c) Why is that particular package preferred for plotting these?

d) Compare these two figures. Which one would you prefer for investigating the changes in pollution level with respect to region? Why?

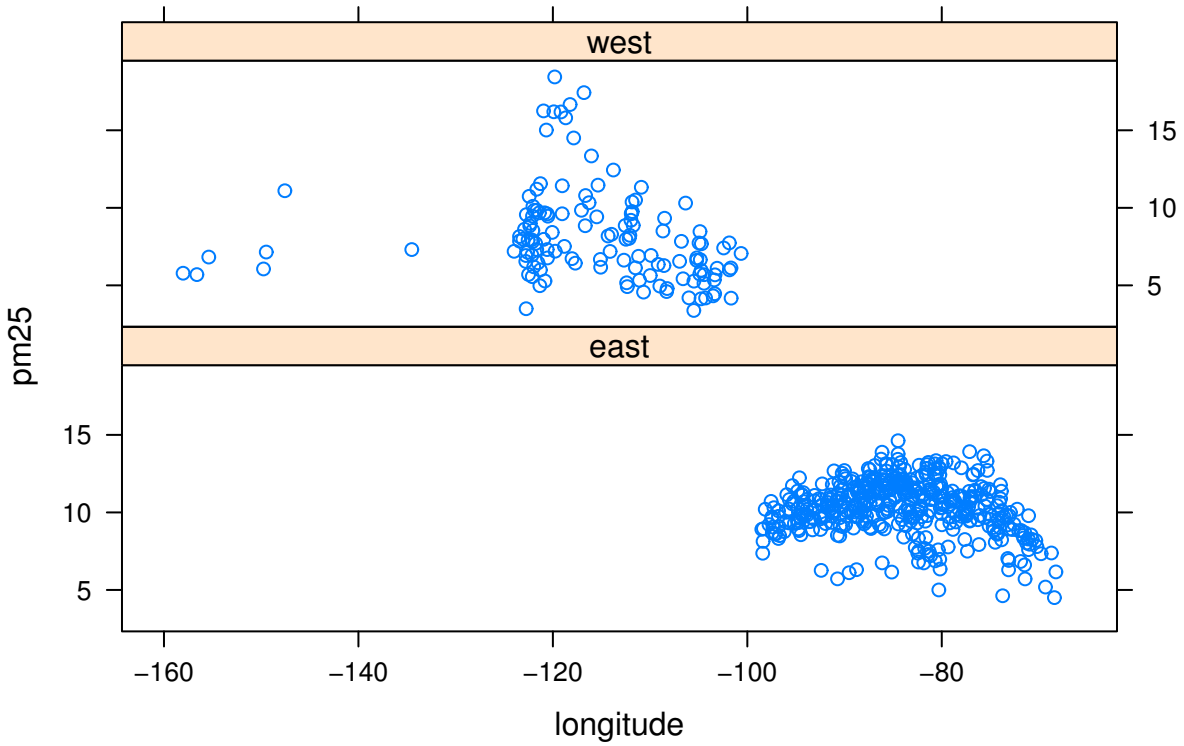

Figure 1: Scatterplots of pollution level and longitude for east and west regions

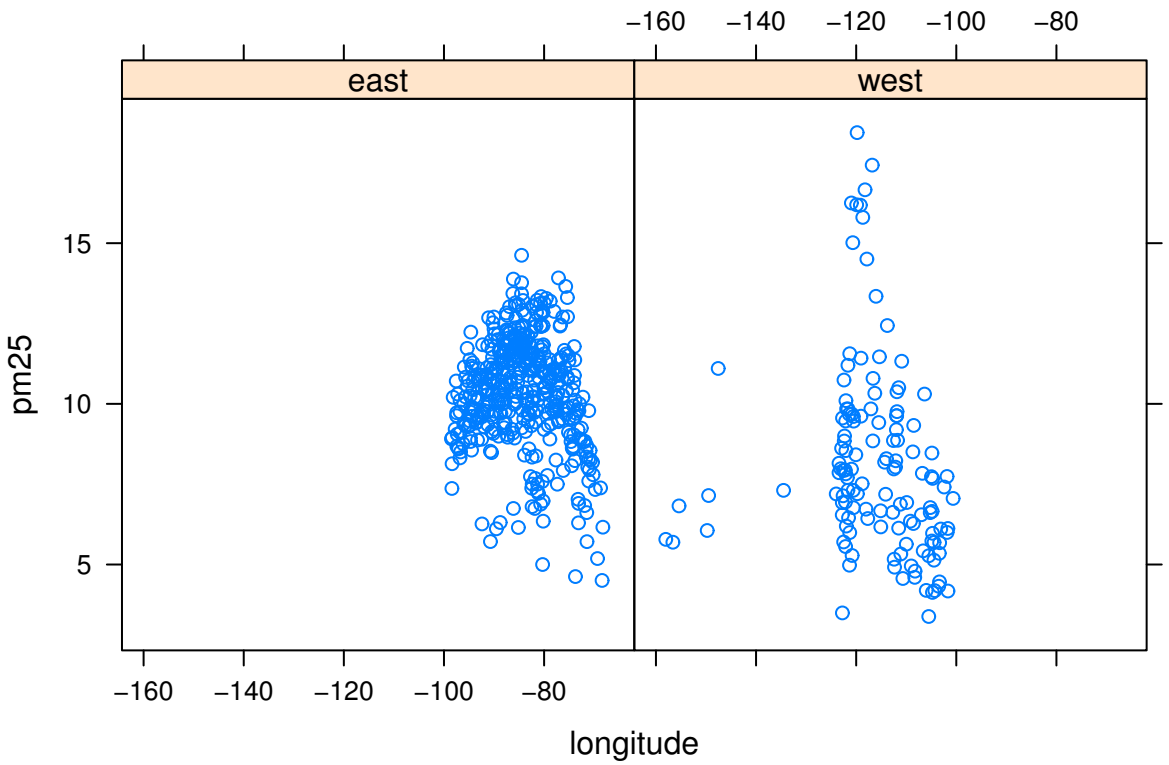

Figure 2: Scatterplots of pollution level and longitude for east and west regions