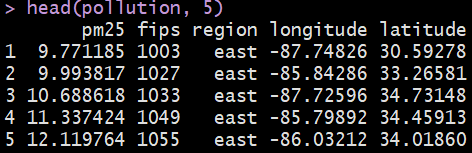
**Homework #1**

* 1. **Descriptive statistics and basic visualization**

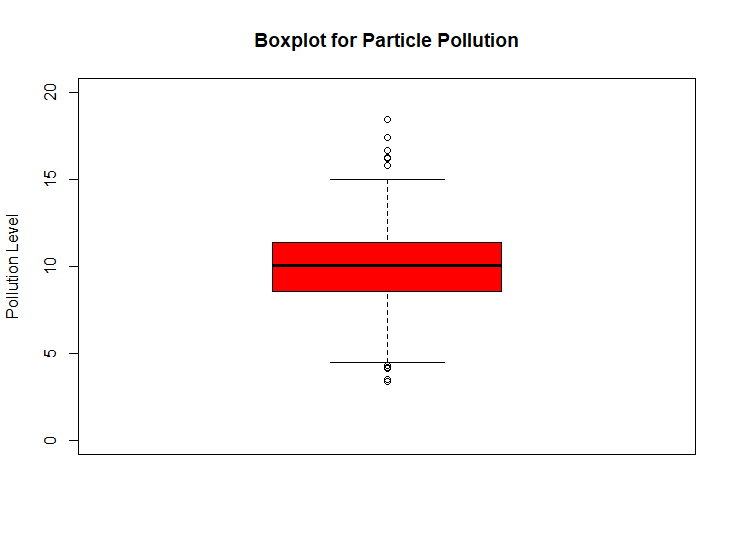
1. Display first five observations of the data set.



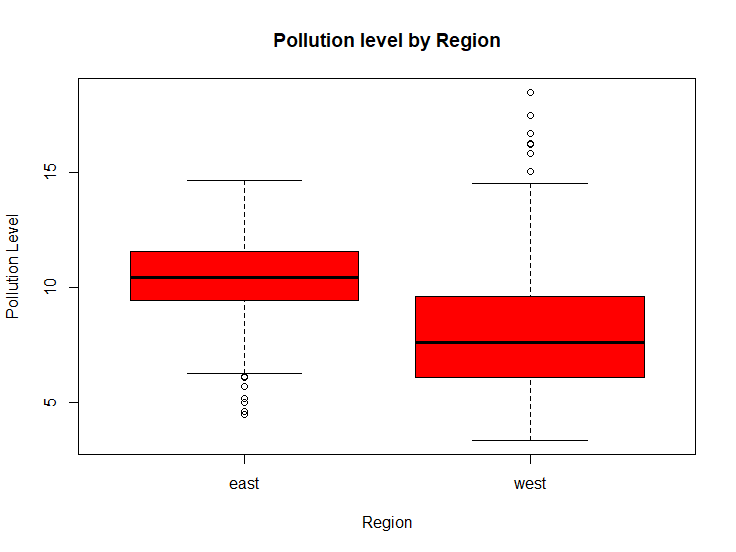
**b.)** Provide a five-number summary for pm25 and comment on what each number refers to.



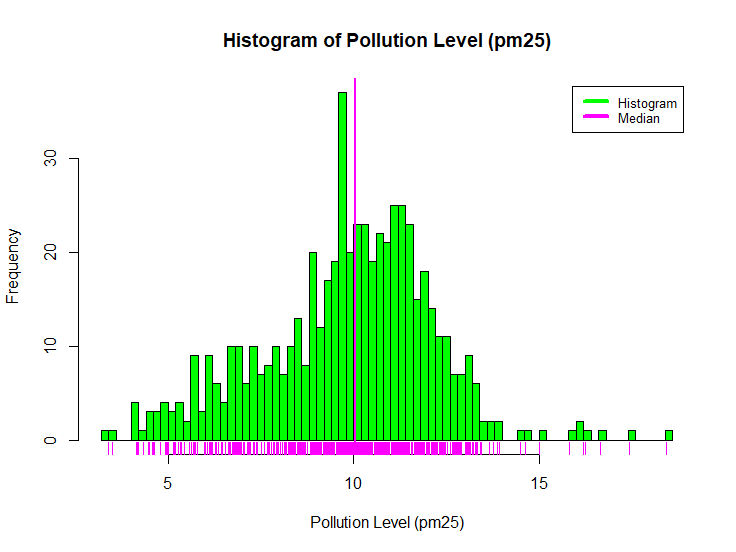
* The first number **(3.382626)** refers to the **minimum**
* The second number **(8.547590)** refers to the **lower-quartile.**
* The third number **(10.046697)** refers to the **median.**
* The fourth number **(11.356829)** refers to the **upper-quartile**.
* The fifth number **(18.440731)** refers to the **Maximum**.

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

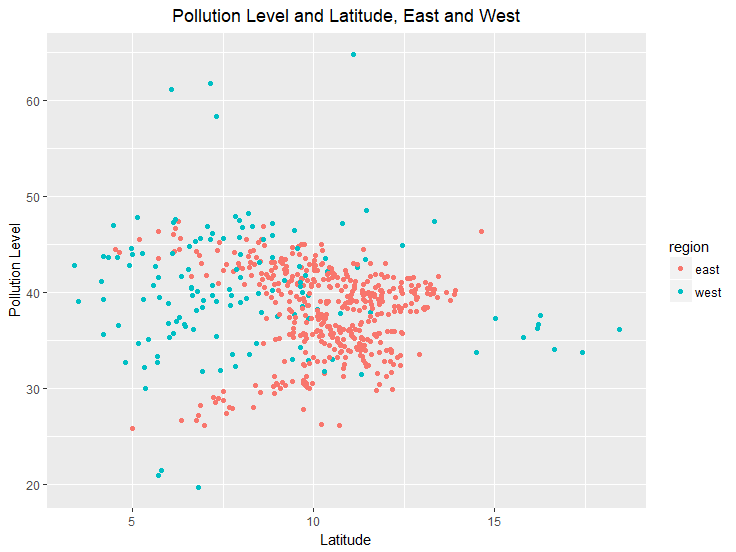
**d.)** Provide a boxplot of pollution level with respect to each region, in red.



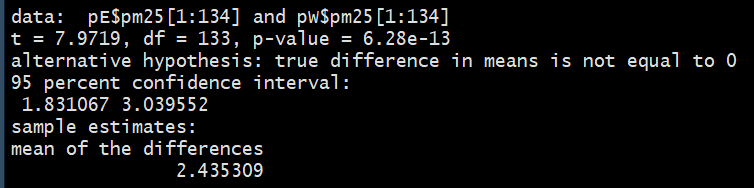
**E.)** Provide a histogram of pollution level with breaks of 100, in green. Include a rug and a line for the median of pollution level on that histogram in magenta.



**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.

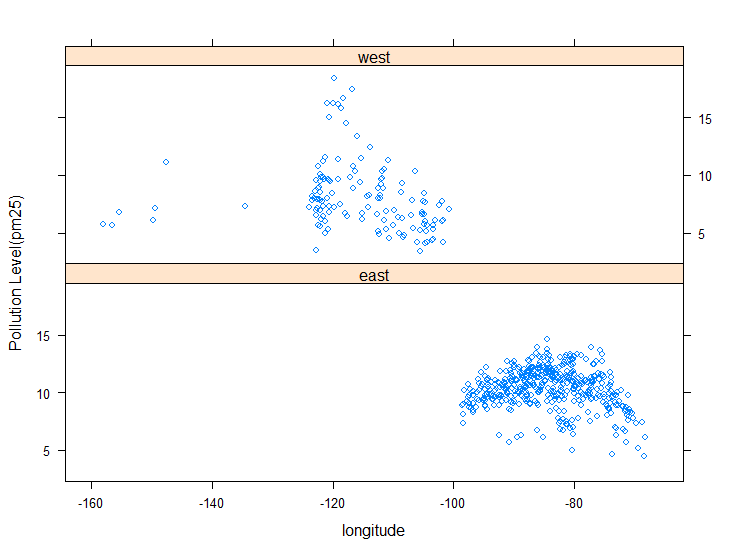


* Hypothesis testing is performed on the 134 total values. We are trying to figure out whether the data is different from each other. If so, then the pollution levels will differ between the two regions. We discovered from the **p-value = 6.28e^-13** that there is evidence that they do differ in values of pollution level by their respected regions east and west. We then draw a conclusion that because of that there are things that may affect the different levels of pollution within their regions.

**1.2 Advanced visualization**

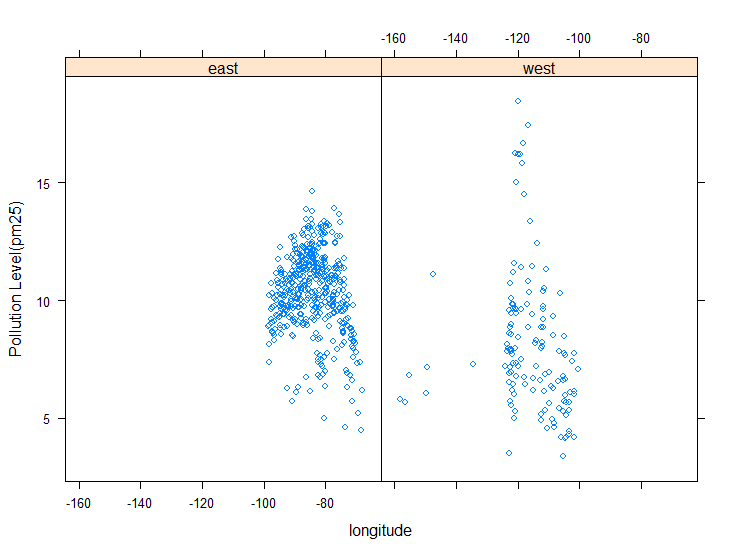
**a.)** Replicate the figures 1 and 2, that are provided below, by using one of the R visualization packages. Present the code.

**Figure 1:**



****

**Figure 2:**



****

1. Which package did you use? **“Lattice”**
2. Why is that package preferred for plotting these?

The lattice package implements the Trellis Graphics system with some novel extensions. The Trellis Graphics system has many sophisticated features that include looking at Looking at how y changes with x across levels of z as well as Good for putting many plots on a screen. It is a powerful and elegant high-level data visualization system, with an emphasis on multivariate data, that is sufficient for typical graphics needs, and is also flexible enough to handle most nonstandard requirements. **In our case this would be the most useful to see the relationship of the longitude and pollution level(pm25) in terms of the region in respect to east and west.**

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

I would use **plot 2** to investigate the changes of pollution level. It is easier to see the pollution level with this graph especially regarding the East region and the data points that are in the middle of our ranges. Regarding the West region we can definitely see the pollution levels easier and because of that I would use plot 2.