**Outlier Detection Assignment**

Introduction

The distribution of the data in individual variables can produce biased or invalid results in a model if there are outliers in the data. Before models are run, it is important to have an overview of the variables in order to ensure that if outliers are present that they represent meaningful values and to anticipate what kind of impact they may have on the results of a completed model. It is also important to thoroughly review the data to find invalid values. These can be introduced into the data during the data generation process.

This assignment covers three types of outlier detection methods. These are simple tools that can be used in the data preparation process to check the quality and characteristics of the data.

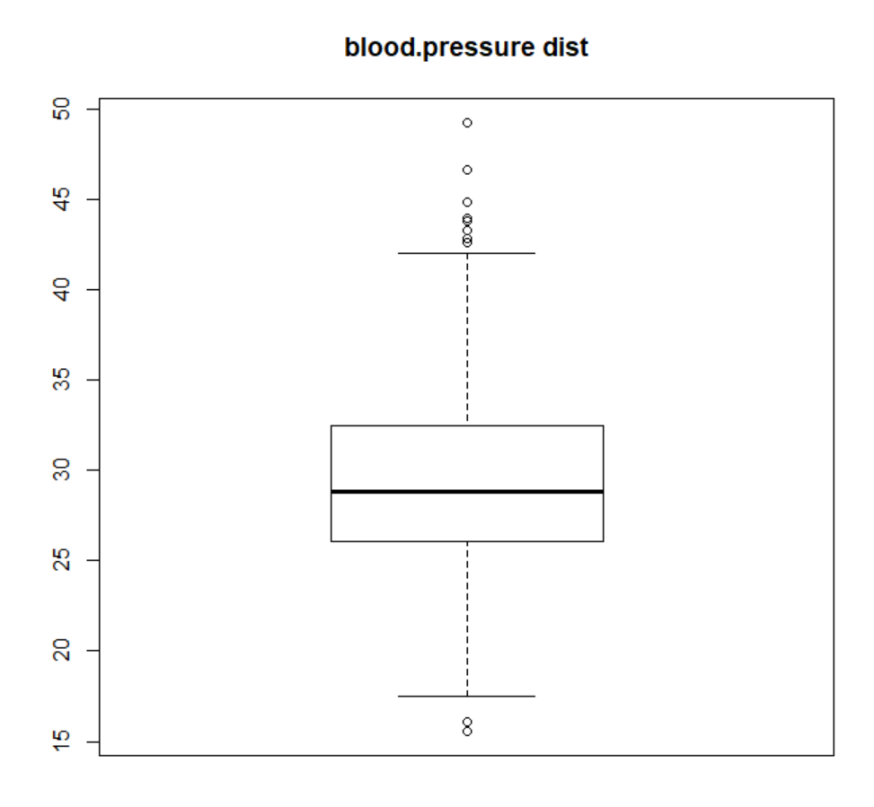
Directions

Use the lecture material from Lesson 9 as your guide for how to use the outliers and mvoutliers packages. Use the assigned reading from Lesson 9, *Beginning Data Science with R*, Chapter 5, Section 5.2.1, as a guide for how to produce box plots in R. Use the chdbtc.csv file from Assignment04: PostgreSQL Query for the assignment. Make sure your dataset has 500 cases and 39 variables.

The completed assignment should adhere to the following guidelines:

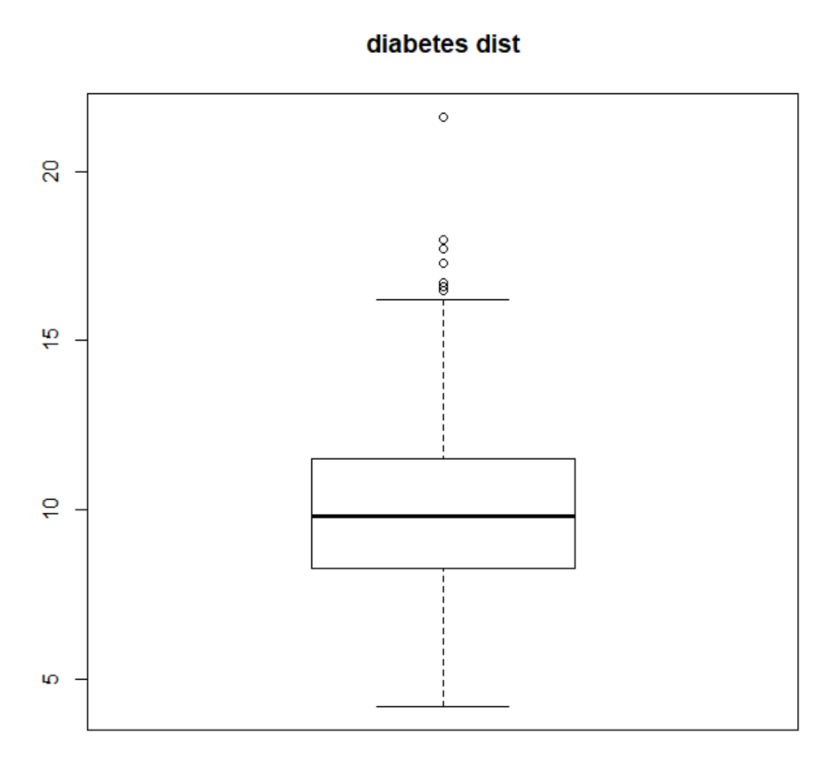
* Include your answers on the assignment document.
* Write your answers using complete sentences with correct punctuation, grammar, and spelling.
* Submit your completed assignment through the Blackboard portal in Lesson 9.
* Use the City Health dataset (chdbtc.csv) for this assignment. Create a subset of 10 variables that are of interest to you. Produce boxplots for two of the variables. Create a bitmap image and paste your code and the image in your assignment. Describe what you observe about the distributions of the variables and any outliers that you find. *(2 points)*

boxplot(df\_subset$blood.pressure, main="blood.pressure dist")



Observation: The blood pressure variable contains an even distribution of values. There are outliers, especially on the top end (note multiple outliers).

boxplot(df\_subset$blood.pressure, main="blood.pressure dist")



Observation: The diabetes variable also contains an even distribution of values. There are outliers, especially on the top end.

* Using the outliers package, use the outlier function to identify the most extreme case for each of the variables. Paste your code and the output in the assignment. Report what you see in the results. *(2 points)*

> #extreme outliers

> outlier(df\_subset$blood.pressure)

[1] 49.2

> outlier(df\_subset$diabetes)

[1] 21.6

Observations: In the blood pressure variable, the highest outlier observed is 49.2. In the

diabetes variable, the highest outlier observed is 21.6.

* Conduct a chi-squared test for univariate outlier detection for two of the variables from your subset. Paste the code and the output in your assignment. What does the chi-squared test indicate about the presence of an outlier in these two variables? (2 points)

chisq.out.test(df\_subset$blood.pressure, variance=var(df\_subset$blood.pressure))

data: df\_subset$blood.pressure

X-squared = 15.584, p-value = 0.00007892

alternative hypothesis: highest value 49.2 is an outlier

Observation: The chi-sq test produced a statically significant result, of which includes 49.2 as the highest value outlier.

chisq.out.test(df\_subset$diabetes, variance=var(df\_subset$diabetes))

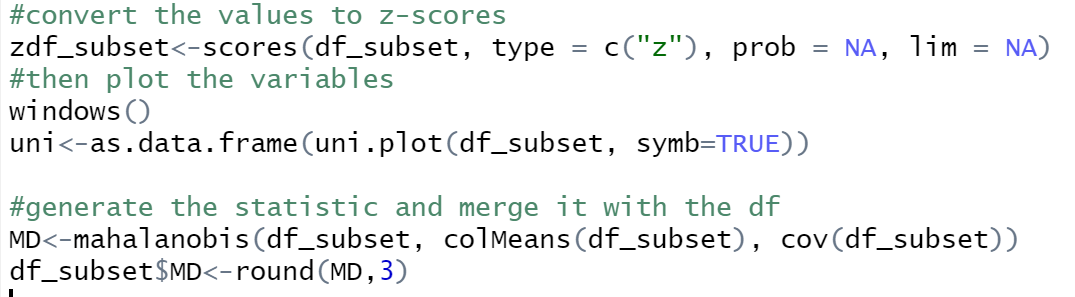
data: df\_subset$diabetes

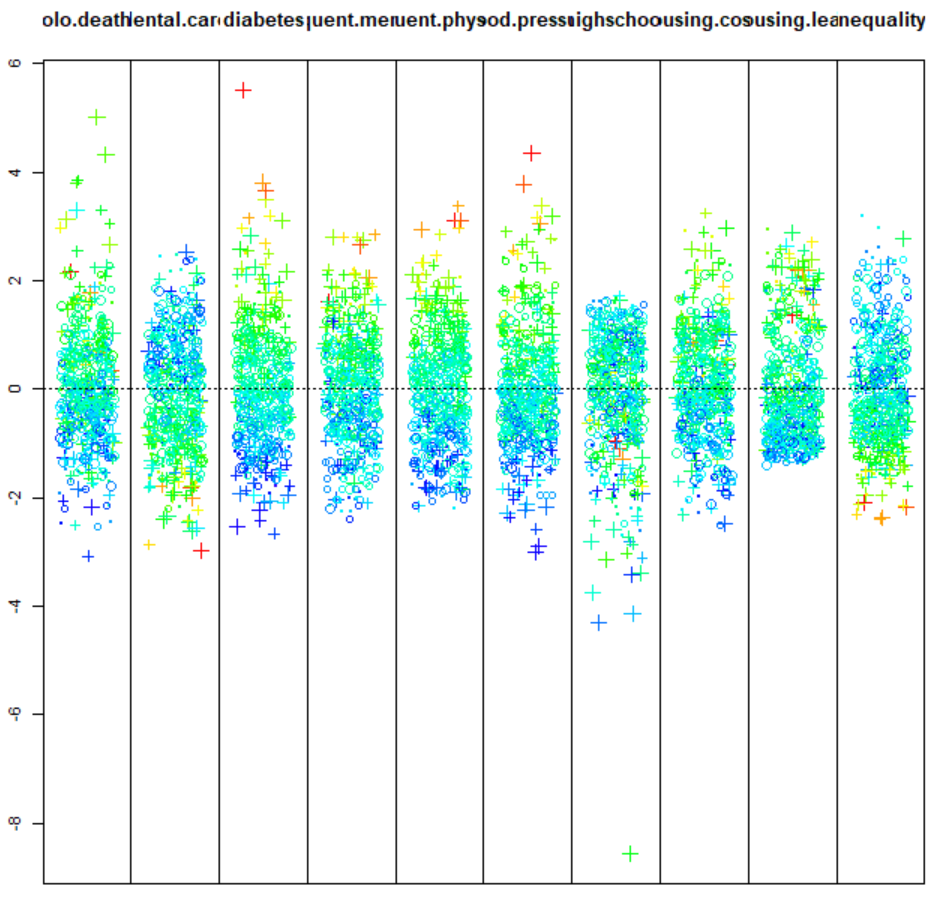
X-squared = 23.425, p-value = 0.000001299

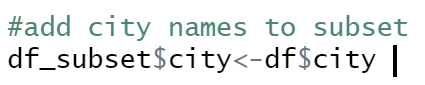
alternative hypothesis: highest value 21.6 is an outlier

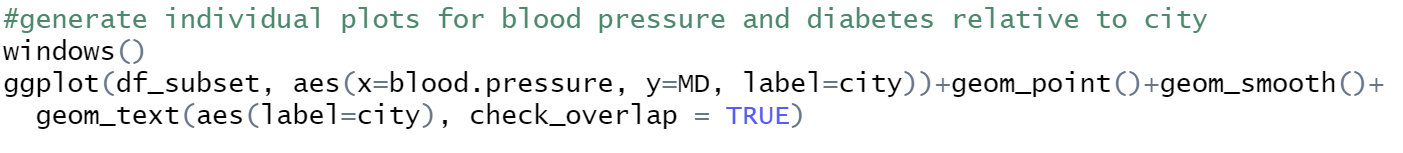
Observation: The chi-sq test produced a statically significant result, of which includes 21.6 as the highest value outlier.

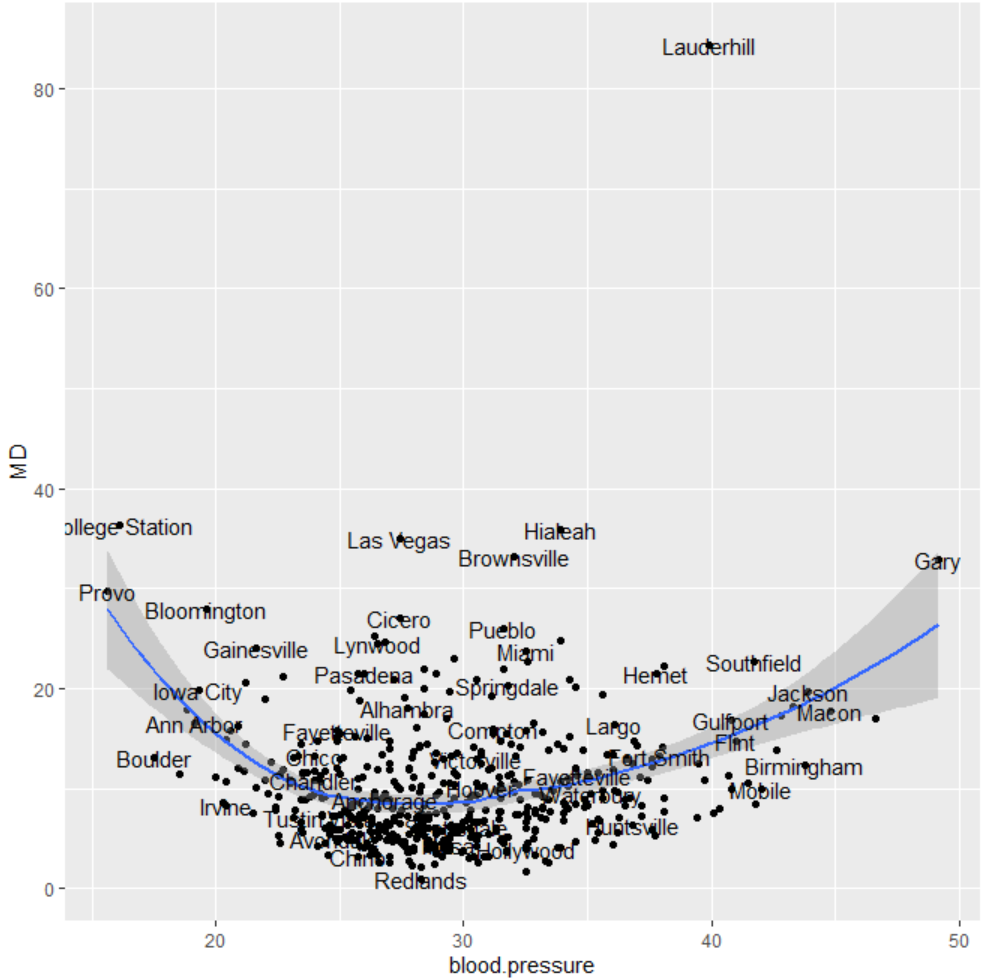
* Run uni.plot for your subset. Create a bitmap image of the plot and paste your code and the plot in your assignment. Create a mahalanobis statistic for subset and add that and the city names to your subset. Paste your code in the assignment. Report what you find for the two cities with the largest mahalanobis statistics. *(4 points)*

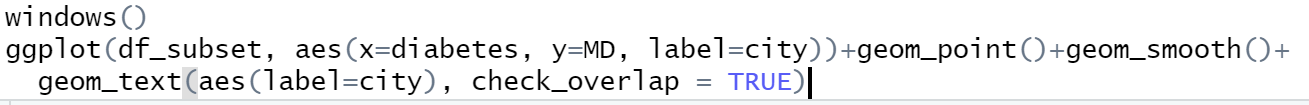


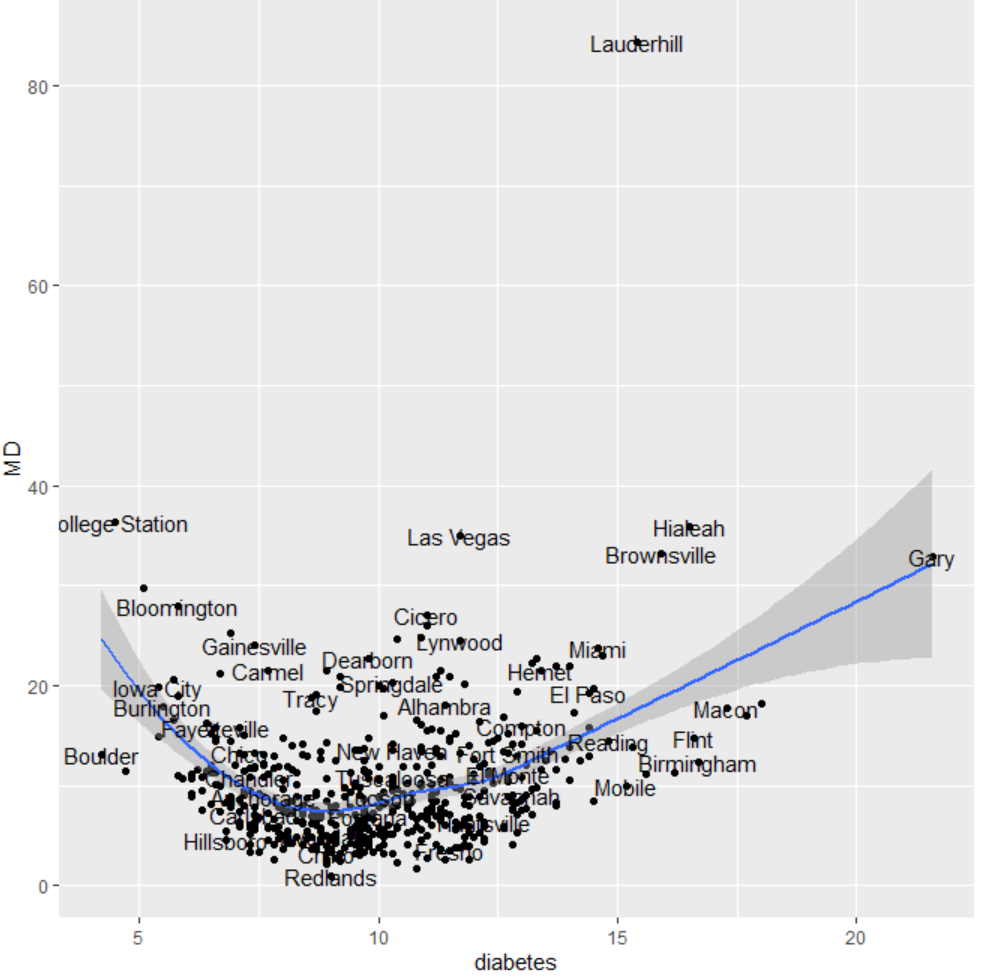












Observations: I’m a wizard. This is awesome.

Actual observations: Lauderhill contains an MD value of ~85, and it’s nearest-neighbor College Station ~37.

Scoring

The assignment is worth 10 points. You must have complete answers to receive full credit.