STAT5014 Homework Week1

Matthew House

9/5/2017

Problem 2

Link to Github page: https://github.com/MattHouse

Problem 3

Swirl Completed

Problem 4

Part A

Through this course I hope to gain:

1. A better understanding of programming in R
2. Ways to handle large datasets that minimize the amount of preprocessing required
3. an ability to publish via LaTEX as the journal I am submitting to offers a LaTEX template

Beyond that I hope to become more proficient in all programming languages that are relevant to remote sensing. Some of the more recent algorithms and implementations are written natively in R, such as EWMA-CD.

Part B

The *probability mass function*:

$$fx(x) = P(X = x) \text{ for all } x.$$

The *probability density function*

$$F_X(x) = \int_{-\infty}^{x} fx(t)dt \text{ for all } x.$$

The *exponential density function*

$$f(t) = \begin{cases} \lambda e^{\lambda t}, & \text{if t} \geqq 0, \\ 0, & \text{if t} \geqq 0. \end{cases}$$

Problem 5

1. For Every Result, Keep Track of How It Was Produced

- When you are performing the action that in turn creates the result it would be prudent to document how you obtained that result
- If this is done in a computer program, the program and version should be documented
- Make sure you create a workflow that allows for direct execution

2. Avoid Manual Data Manipulation Steps

- Execution of programs over manual manipulation of data will produce less error and variance person to person
- Each type of program has repositories, packages, and/or standard commands that help avoid the need for manual data manipulation

- If manual manipulation cannot be avoided, noting what was manipulated will help minimize future confusion

3. Archive the Exact Versions of All External Programs Used

- Identify the versions of external programs used, because subsequent versions may require different inputs
- By archiving the exact version used by executable line or exact disk image will thwart incongruencies later
- Minimally note the version of each external program used in your code

4. Version Control All Custom Scripts

- It is important to archive code as it evolves so that a certain result may be reproduced at any given time
- It is best to use a version control system
- Minimally, archive copies of scripts periodically to create a rough history of their development

5. Record All Intermediate Results, When Possible in Standardized Formats

- By recording intermediate results a track record is created that allows for discrepancies to be seen and bugs to be uncovered
- This also shows the consequences of programs and parameter choices at the various intermediate steps
- Overall by recording the intermediate steps a programmer can identify inconsistencies at specific stages and can run certain parts of the code without being constrained to running it all at once
- Minimally, archive any intermediate result files produced

6. For Analyses That Include Randomness, Note Underlying Random Seeds

- When randomness is introduced in the analyses, it is important to record the seed used to generate the results
- At a minimum, note the steps that involve randomness

7. Always Store Raw Data behind Plots

- By linking the raw data to the plots they are used in allows for graphical modification without having to redo the entire analyis
- It is advisable to inlcude both the underlying datat and the processed values when producing graphs such as histograms
- If using a command based system, it is convenient to also store the script that produced the graph

8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

- Inspect the detailed values underlying the summaries atleast once
- If space permits, permanently output the underlying data when a result is generated
- Hypertext is a good way to accomplish this without cluttering the viewable results

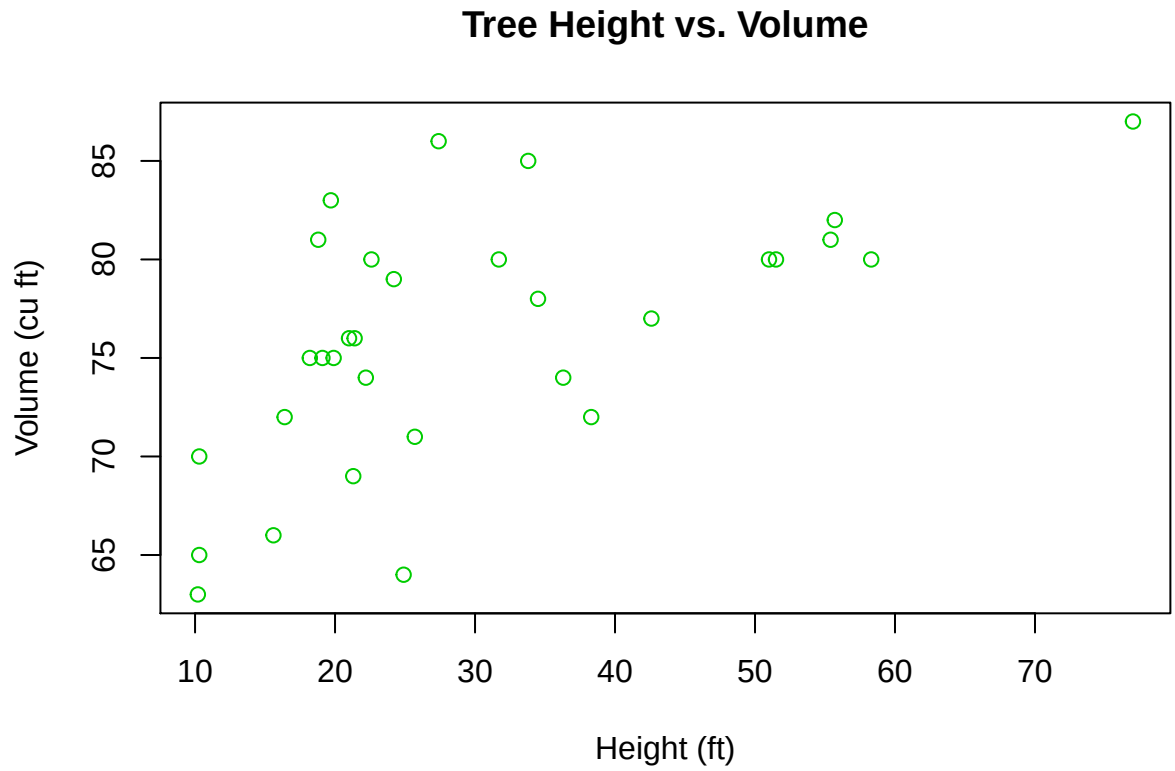
9. Connect Textual Statements to Underlying Results

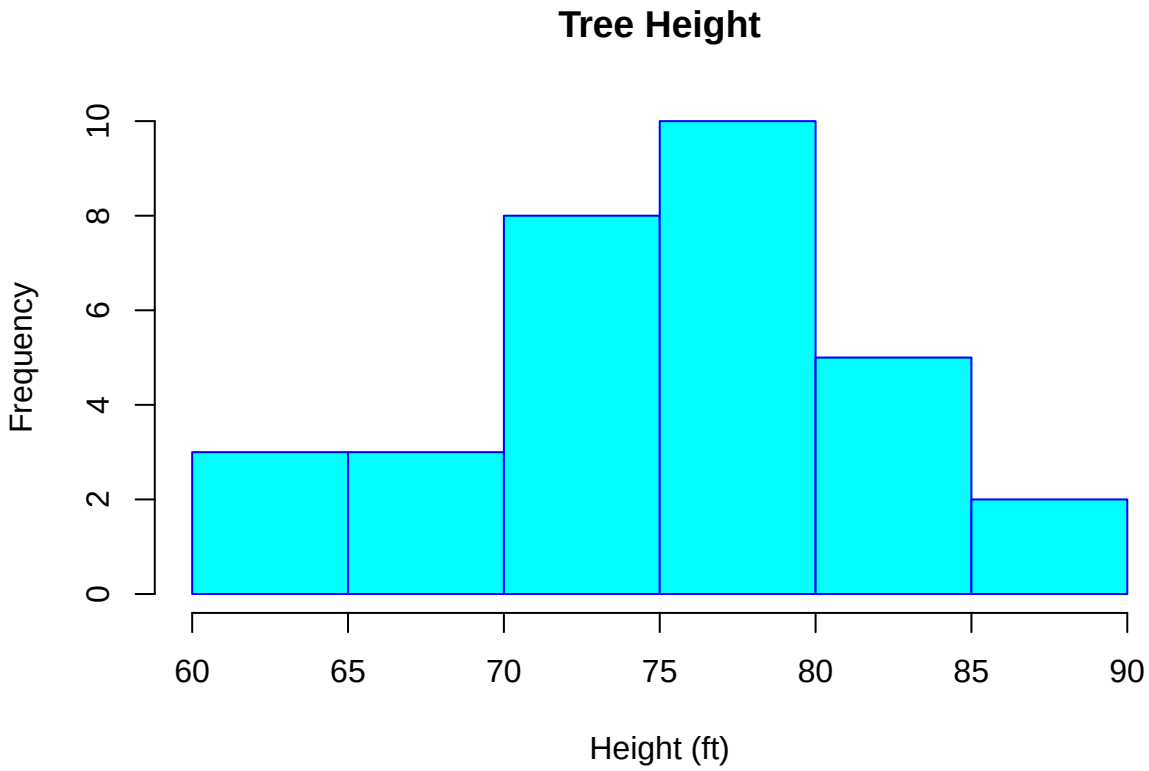
- Connect what you say, textually, to the underlying results at the time of conceiving the textual statement
- This is important because it allows results to be tracked down in the future +There are tools available to integrate reproducible analyses directly into publishable documents

10. Provide Public Access to Scripts, Runs, and Results

- By making all data and analyses used to produce published results available allows for transparency
- It can increase the chances of the work getting published
- It can improve the chances of the work being cited, because others can reproduce the same results which makes the work more robust in your peers' eyes

Problem 6.


Tree Height vs. Volume
Volume (cu ft)
65
70
75
80
85
10
20
30
40
50
60
70
Height (ft)


Tree Height
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (ft)