R Answers to Exercises: Module 3

If you have any questions as you go through these, feel free to ask them in the forum.

Note: For all of the graphs, you HAVE to use the long format of the data set. So if you're having trouble getting the data set up, use the syntax we've provided on the module 3 page to create the long data set.

1. Read the help file on reshape: you can type help("reshape") into your console window to do this.

2. Convert the Physical Training Data from wide to long format. Keep all variables in the data set.

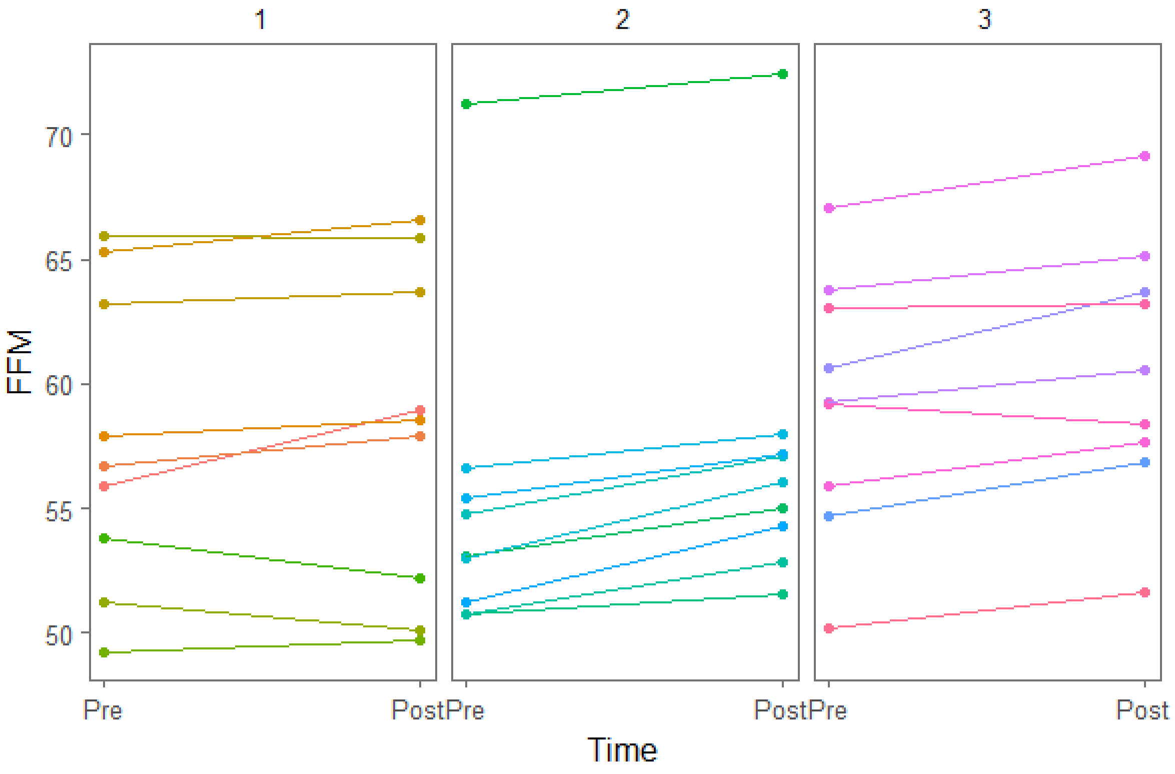
(It may help to refer to your Module 1 homework before you begin to be sure of which variables are time-varying/within-subjects and which are time-invariant/between subjects).

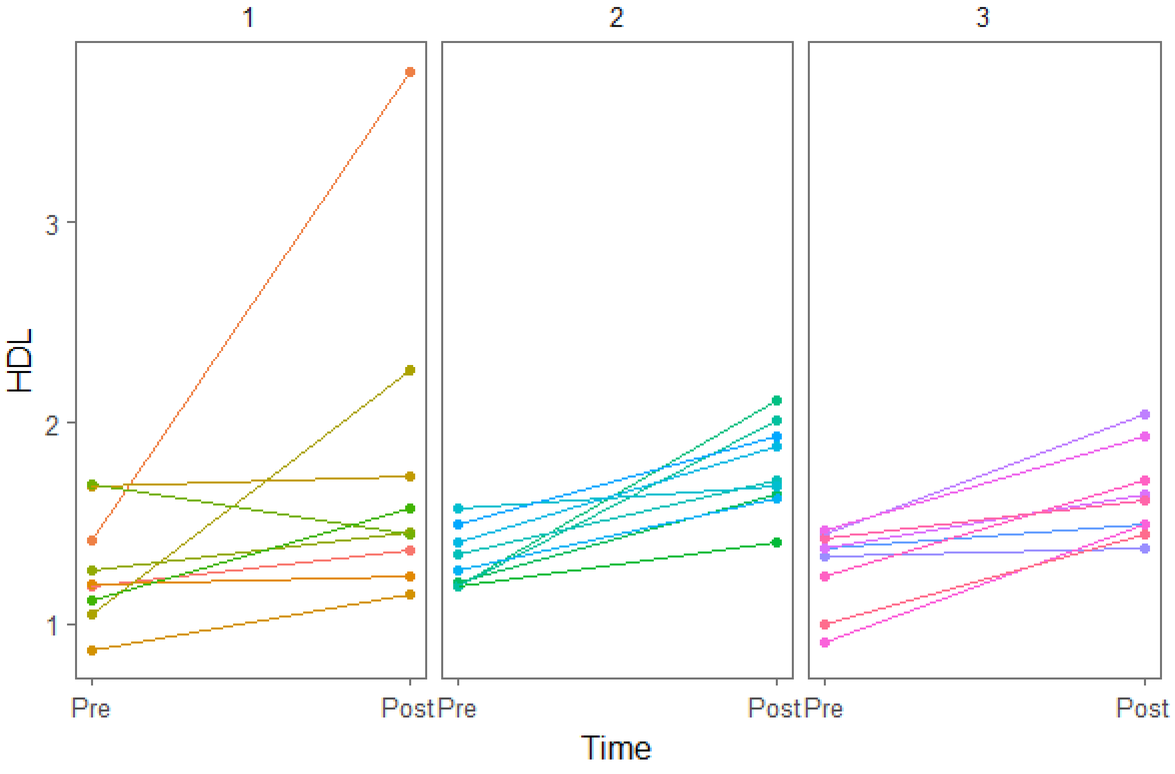
[Hint: the only thing different here than on the Teachers example I showed you is there is no Subject ID variable already in this data set. So on step 3, in Case Group Identification you will need to add an ID variable.

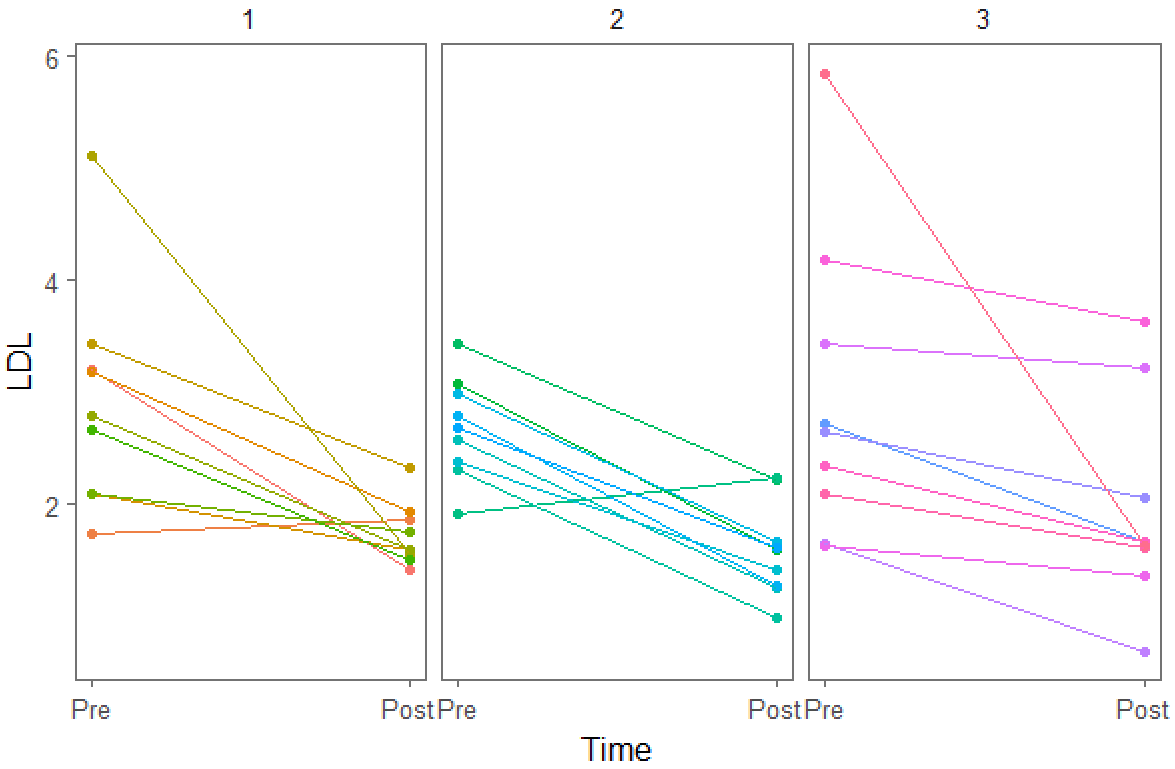
For questions 2 and 3, see the R syntax file for the exercises.

3. Convert it back.

4. Create a separate line plot for each training regimen group over time for the following outcome variables in the Physical Training data set: FFM, HDL, LDL. [Note: make the separate graphs by paneling, not by splitting the data file. This will keep the Y axis the same on all graphs.] Do the people who start at the high end at pre-test on each outcome tend to be at the high end at post-test? Do the differences for each subject within a group seem stable? As a whole, do the training regimens seem to have similar effects on these health outcomes? Is this true for everyone?


1
2
3
FFM
70
65
60
55
50
Pre
Post
Pre
Post
Pre
Post
Time


1
2
3
HDL
3
2
1
Pre
Post
Pre
Post
Pre
Post
Time



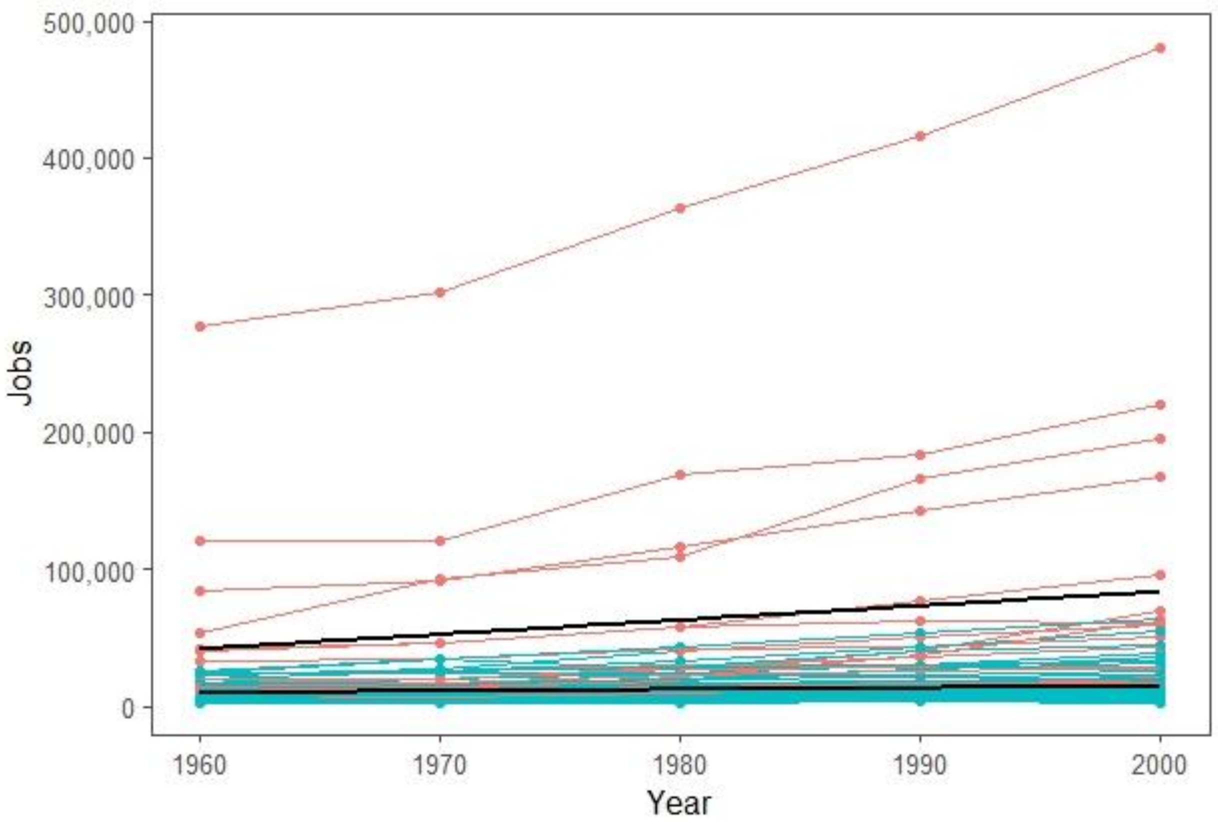
For the following exercises, use the County data:

5. Convert the Countywide data from wide to long format. Keep all variables in the data set. Some variables have data from 1950. Just ignore those for this exercise (we'll come to it later), and for now, treat the 1960 measurement as time 1 for all variables.

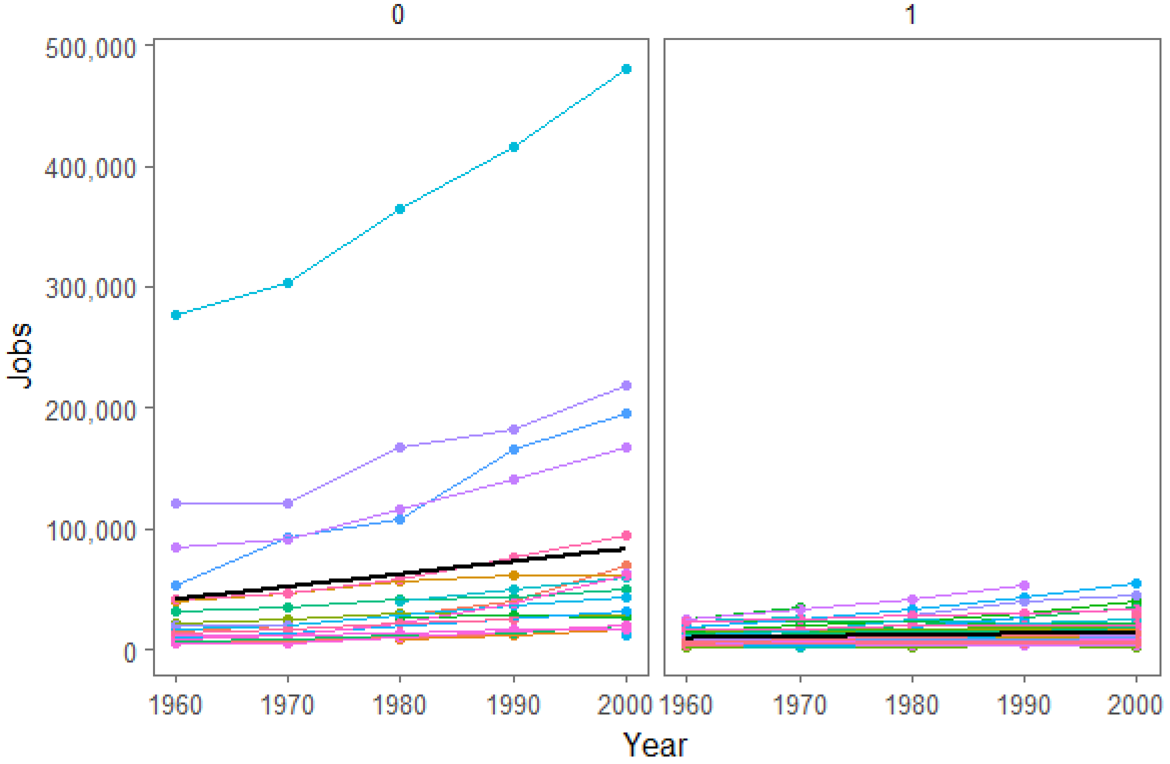
For questions 5 and 6, see the R syntax file for the exercises.

6. Convert it back.

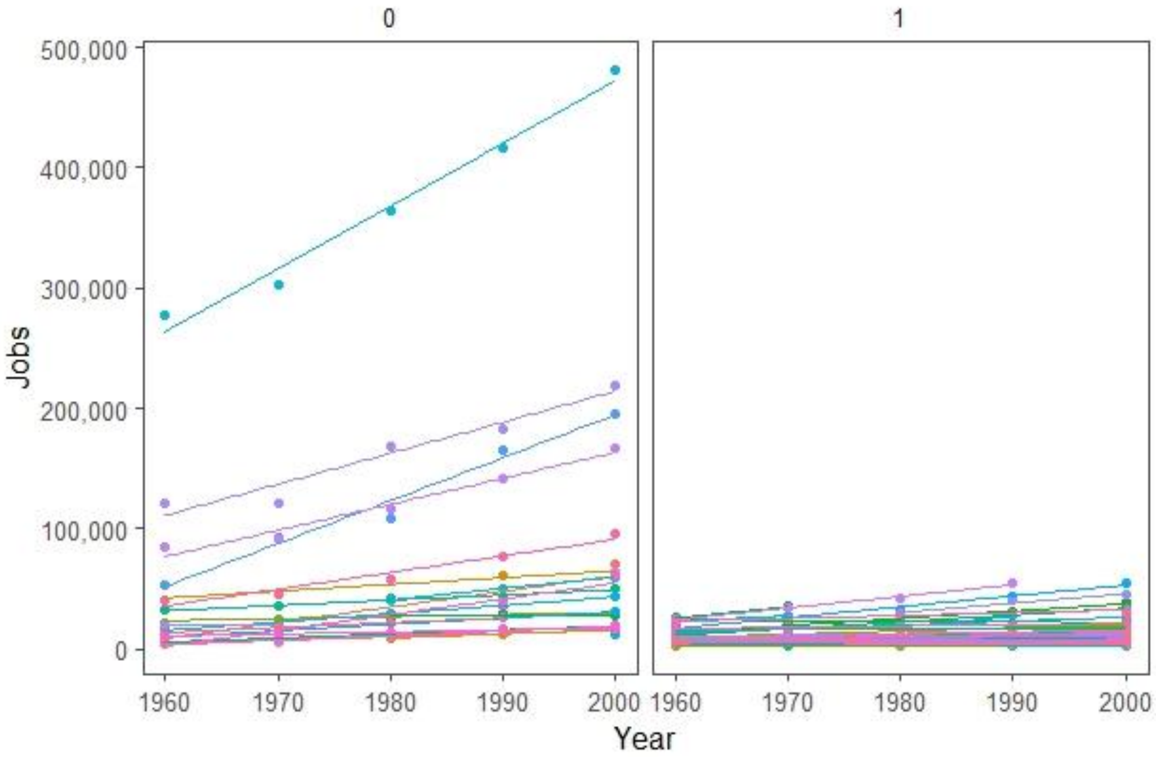
7. Create a scatterplot of the number of jobs in each county by decade. Points should be connected for each county. Make rural and non-rural counties different colors, and create a regression line for rural counties and non rural counties. (In other words, make it like the Useful Graph #3).


500,000
400,000
300,000
200,000
100,000
0
Jobs
1960
1970
1980
1990
2000
Year

OR


0
1
500,000
400,000
300,000
200,000
100,000
0
Jobs
1960
1970
1980
1990
2000
1960
1970
1980
1990
2000
Year

8. Create a line graph, paneled by Rural, of the number of jobs by decade. Each county should have its own interpolation line. How linear do the trajectories look? Does Rural seem to be a good predictor? Does the growth trend over time (the slope) seem to be the same in rural and non-rural counties?

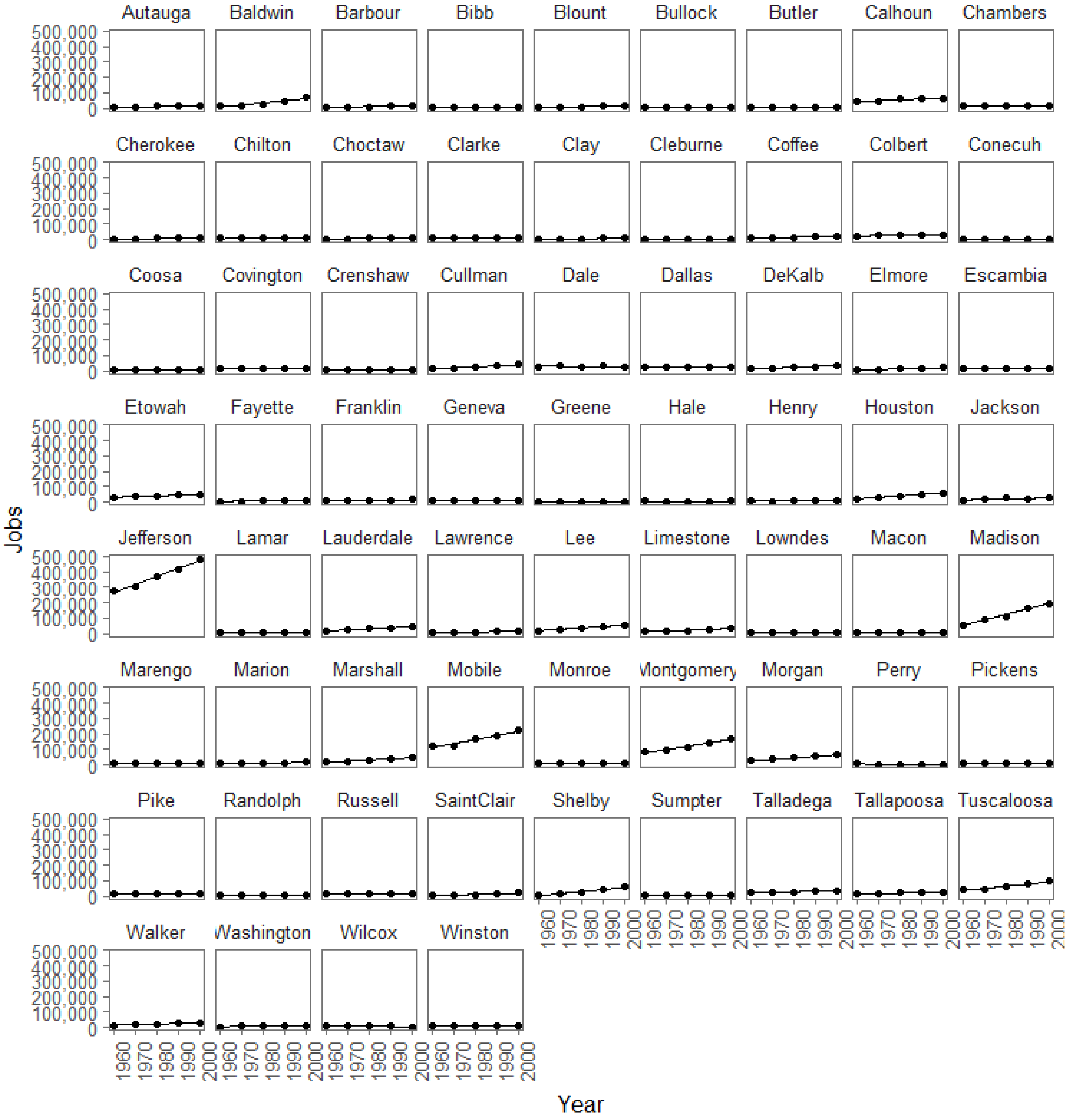


Note, I interpreted interpolation line to mean linear regression line. There could be other ways to interpret this.

Rural does seem to be a good predictor, although it is clearly not explaining the entire difference b/c even some non-rural counties have very low growth—they look just like the rural counties.

There is clearly more of an upward trend in most (but not all) non-rural counties than in the rural counties, which all look nearly the same.

9. Fit a scatterplot, with a fitted regression line, paneled separately for each county. Does a linear model appear to be a good fit for most counties? For all?



A linear model does seem like a very good fit for most counties. There are a few, like Jefferson, which seems to have a very slight upward curve, but even there, a linear trend line fits very well.

10. Read one or more of the recommended resources on ML, REML, and Information Criteria (listed on the Module 3 page). I highly recommend the chapter in Parta's book, if you can get a hold of it.