Mini-project 1: House prices and population growth

S470/670

Note: This document is subject to change until the end of Sunday 9th February.

One member of your working group should upload your initial submission through the Assignments tab on Canvas by 11:59 pm, Thursday 20th February.

For your initial submission, you may work in groups of any reasonable size, and submit one report per group. However, your final submission will be in groups of no more than 2 people. Register your group using the "Mini-project 1 initial groups" tab on Canvas (or ask me or a TA to do it.)

A researcher for a thinktank wants to learn about how house prices in the U.S. have changed over the last few decades, and whether changes in prices are related to population in some way.She has taken an introductory statistics course using R, but that was a long time ago, so she is outsourcing the exploratory data analysis to YOU.

Questions

The researcher's major research question is: **How have house prices in U.S. states changed over the last few decades, and are changes in prices are related to population in some way?** This question may be difficult to answer, at least straight away. So she has brainstormed a series of questions he would like you to address, which can be divided into groups:

1. **House prices over time:** How have house prices in the U.S changed since 1975, after adjusting for inflation (i.e. relative to the CPI?) How have changes in prices varied by state? Which states have seen the biggest increases in real house prices, and which have seen the biggest decreases? Have changes in prices within each state mostly followed the same basic pattern, and are there outliers to that pattern? Do the typical patterns vary between the four regions (Northeast, Midwest, South, and West)?

2. **Population density and changes in house prices:** Does present-day population density explain changes in house prices by state since 1975? Are there outliers to the relationship, and if so, is there a principled reason to drop them? What does the relationship look like after dropping or downweighting outliers? Does the relationship vary by region? If so, how?

3. **Changes in population and changes in house prices:** Is there a relationship between changes in population and changes in house prices? To answer this, look at changes in each state over three time periods: 1990 to 2000, 2000 to 2010, and 2010 to the present. Analyze the three time periods separately. Has the relationship changed over the three time periods? Are there variations by region?

4. **Conclusion:** What does all of this tell you about the relationship between house prices and population? Is there a plausible cause-and-effect story you can tell that's consistent with the data and with common sense?

Data

- **Freddie Mac House Price Index** (`http://www.freddiemac.com/research/indices/house-price-index.html`): This tracks the average house price in each state (plus Washington, D.C. and a national average) since 1975. For each state, the index value is fixed at 100 in December 2000. Figures are seasonally-adjusted but not adjusted for inflation. The data is in the spreadsheet `State_and_US_SA.xls`.

- `state_abbrevs.txt` (on Canvas): This contains the names, the two-letter code, and the Census region (Northeast, Midwest, South, or Midwest) for each state in the U.S.

- **Consumer Price Index**, on Canvas in `cpi.csv`. This tracks how prices in general have changed in the U.S. The data in the spreadsheet is the "All Urban Consumers (Current Series)," seasonally adjusted, from `www.bls.gov/cpi/data.htm`.

- **Census Data**: You'll need to get population data for each state in 2018, 2010, 2000, and 1990 from the Census and American Community Survey. We recommend you do this using the `tidycensus` R package. See `https://walkerke.github.io/tidycensus/articles/basic-usage.html` for a description of how to use the package.

  To use the package, you'll need a Census Data API key. To sign up for one, enter your email at `https://api.census.gov/data/key_signup.html` and they'll send you a key (you'll have to wait a few minutes for it to activate.) The main population variable is `P001001` in the 2010 and 2000 Censuses, and `P0010001` in the 1990 Census. If you can't get the data this way, you can always get it directly from `data.census.gov`.

What to submit

Your submission should consist of:

- A report (PDF preferred) of **no more than eight pages**, excluding appendices. The report should have the following sections:
  - Executive summary stating your main findings (Note: Do NOT include this executive summary in your initial submission, as you should only write it when you're nearly finished.)
  - House prices over time
  - Population density and changes in house prices
  - Changes in population and changes in house prices
  - Conclusions
- A .Rmd or other file containing your code.

- Any other supplementary files required to reproduce your work.

We will give you feedback, then, working alone or in a pair, you will make a final submission by a date to be announced. The grade for your final submission will be the one that counts.

Some constraints:

- The researcher is familiar with elementary methods like linear models, but not with non-parametric methods such as loess and gam. That means that if you want to use those more fancy models, you need to briefly describe what those techniques are doing in words that a non-statistician can understand.
- She is comfortable with transformations, but they would have to be interpretable.
- She took her statistics course from a lecturer who was highly skeptical about P-values, so any P-values you show must be accompanied by a justification of why the P-value is necessary.
- She wants to be able to reproduce your work if required, but doesn't want to see R code in the main report.
- She is red-green colorblind.

Notes

- There is no one objectively right answer to the project, but there are many subjectively bad answers.
- Make sure you justify your answers to the questions (don't just state answers.)
- You do not necessarily need one overall model that describes all the data.
- Because there's no correct model, you're free to use multiple models for the same data and question, if you feel that's a good use of your time and page count.
- A large fraction of the points are for communication, so maintain a decent level of professionalism.
- Graphs must be big enough to be readable if the report were printed out.
- Additional technical graphs such as residual plots can be included in an appendix, which will not count toward the page limit and which we might not bother to read. Submit your code as a separate file. Also upload any additional sources required to reproduce your work.
- The best reports tell a clear story. Pay attention to the structure of your report and make sure that you leave the reader with a clear idea of your answers to the questions.

Grading

- House prices over time: 5 points
- Population density and changes in house prices: 5 points
- Changes in population and changes in house prices: 10 points
- Communication (including executive summary and conclusion): 10 points. Full credit for communication requires a readable, informative, comprehensive, clearly labeled set of graphs, and a comprehensible write-up with few glaring spelling and grammatical errors that makes the main points of the analysis clear.