Lab-09

Your Name Here

March 24, 2022

Preface

The goal of this assignment is to help you gain more familiarity with using **ggplot** to visualize relationships. As always, please come to office hours and reach out to your teaching staff if you have any questions.

FYI: in this assignment we added `\newpage` before some questions to keep question statements, code blocks, and output together.

Data

We will work with monthly data on Zillow median sale price from Zillow Research.[1] For this lab, we will work with raw data from `Metro_median_sale_price_uc_sfrcondo_month.csv` and smoothed and seasonally adjusted data from `Metro_median_sale_price_uc_sfrcondo_sm_sa_month.csv`. Each data set contains monthly median sales price for 94 US metropolitan statistical area. The two data series start at different times, but both end in January 2022. Start by importing the raw data and the adjusted data and assigning them to `raw` and `adj`, respectively.

1. We'll start by tidying the data `raw`. Drop the row that contains information for the entire United States, and then tidy the data set so that each observation is an MSA-date. Filter out observations with missing prices. Convert `date` into the date format. Generate new variables of `year` and `month` indicating the year and month of the date. Convert price to thousands. Assign the new data set to `raw_tidy`.

[1] Median Sale Price: The median price at which homes across various geographies were sold.

2. Visualize the price distribution across years, using the type of plot that you think is most appropriate. Show the median price per year in the plot. Add tweaks that you think can improve the plot. How does the distribution and the median price change over years?

3. Use `geom_line()` to visualize the median price over time in New York City. Put the aesthetic mappings within `geom_line()` rather than `ggplot()`. Assign it to `nyc_plot` and then print it.

4. Add a colored rectangle to the plot in question 3, showing the period of pandemic, from March 2020 to the last period of the data. Use a caption below the figure to explain what the shading indicates. Use your plot object `nyc_plot` as the starting point rather than duplicating your code from above.

5. Do you see any seasonality in the plot of question 3? Now use the same procedure to wrangle the data `adj`. Overlay the two time series from the raw data and the seasonally adjusted data for New York City in one plot. Compare the two lines. What difference(s) do you see?

6. Now let's work with the seasonally adjusted price data. Choose the four MSAs with `SizeRank` from 1 to 4. Visualize prices over time, faceting by MSA. Optional: Do these patterns make sense to you? If not, you could investigate the data source further to see if you can explain any discrepancies between the results and what you expected.

7. Use the adjusted price to caclulate the average of the median prices per year per MSA. Then caclulate the annual price change rate per region: $(p_t - p_{t-1})/p_{t-1}$. Express them as percentage points (i.e., ranging from -100 to 100, not -1 to 1). Find the cities with the largest and smallest change rate in 2021. Use the resulting data frame to plot the change rate using a bar chart for these two cities, with each city as a separate facet. Use logical color-coding for whether the growth rate was positive or negative each year.

8. Now import sales counts from `Metro_sales_count_now_uc_sfrcondo_month.csv`. Join the sales data with the raw price data. Keep the observations for New York City. Make a scatter plot mapping raw prices to the x-axis and sales to the y-axis. What pattern do you see? Is it consistent with your expectation? Can you explain why?