Homework - Week 9

Write your name here

2025-10-23

Preface

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

Data

We will work with monthly data on home prices from Zillow Research. 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 a measure of home sales prices for 94 US metropolitan statistical areas. The two data series start at different times, but both end in January 2022. We'll start by importing the raw data and the seasonally adjusted data and assigning them to `raw` and `adj`, respectively.

```
raw <- read_csv("Metro_median_sale_price_uc_sfrcondo_month.csv")
adj <- read_csv("Metro_median_sale_price_uc_sfrcondo_sm_sa_month.csv")
raw |> head(3)
```

```
# A tibble: 3 x 173
  RegionID SizeRank RegionName    RegionType StateName `2008-02-29` `2008-03-31`
     <dbl>    <dbl> <chr>         <chr>      <chr>            <dbl>        <dbl>
1   102001        0 United States Country    <NA>            192000       195200
2   394913        1 New York, NY  Msa        NY                  NA           NA
3   753899        2 Los Angeles-~ Msa        CA              510000       500000
# i 166 more variables: `2008-04-30` <dbl>, `2008-05-31` <dbl>,
#   `2008-06-30` <dbl>, `2008-07-31` <dbl>, `2008-08-31` <dbl>,
#   `2008-09-30` <dbl>, `2008-10-31` <dbl>, `2008-11-30` <dbl>,
#   `2008-12-31` <dbl>, `2009-01-31` <dbl>, `2009-02-28` <dbl>,
#   `2009-03-31` <dbl>, `2009-04-30` <dbl>, `2009-05-31` <dbl>,
#   `2009-06-30` <dbl>, `2009-07-31` <dbl>, `2009-08-31` <dbl>,
#   `2009-09-30` <dbl>, `2009-10-31` <dbl>, `2009-11-30` <dbl>, ...
```

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 date format. Generate new variables of `year` and `month` indicating the year and month of the date. Convert the prices to thousands. Assign the new data set to `raw_tidy`. Print `raw_tidy` to check its contents.

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 do the median price and price distribution change over years?

The median price and price distribution...

3. Use `geom_line()` to visualize the sale 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. Use `geom_rect()` to add a colored rectangle to the plot from 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?

Yes/No...

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.

These patterns do/do not make sense to me...

7. Use the adjusted price to caclulate the average of the sale price per year per MSA. Then caclulate the annual price growth 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 growth rate in 2021. Use the resulting data frame to plot the annual growth rates over time using a bar chart for these two cities, with each city as a separate facet. Use different colors for the interior of the bars depending on whether the growth rate was positive or negative each year.