USA Real Estate Prices 2001-2022 and COVID impact review

2023-11-07

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ forcats   1.0.0     ✓ readr     2.1.5
## ✓ ggplot2   3.5.1     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
```

```
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(ggplot2)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
temp_file <- tempfile(fileext= ".csv")
datapath <- "C:/Users/98910/Desktop/Real_Estate_Sales_2001-2022_GL.csv"
org_data_set <- read_csv(datapath)
```

```
## Rows: 1048575 Columns: 14
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr  (8): Town, Address, Property Type, Residential Type, Non Use Code,
Asse...
## dbl  (5): Serial Number, List Year, Assessed Value, Sale Amount, Sales
Ratio
## date (1): Date Recorded
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```r
str(org_data_set)
```

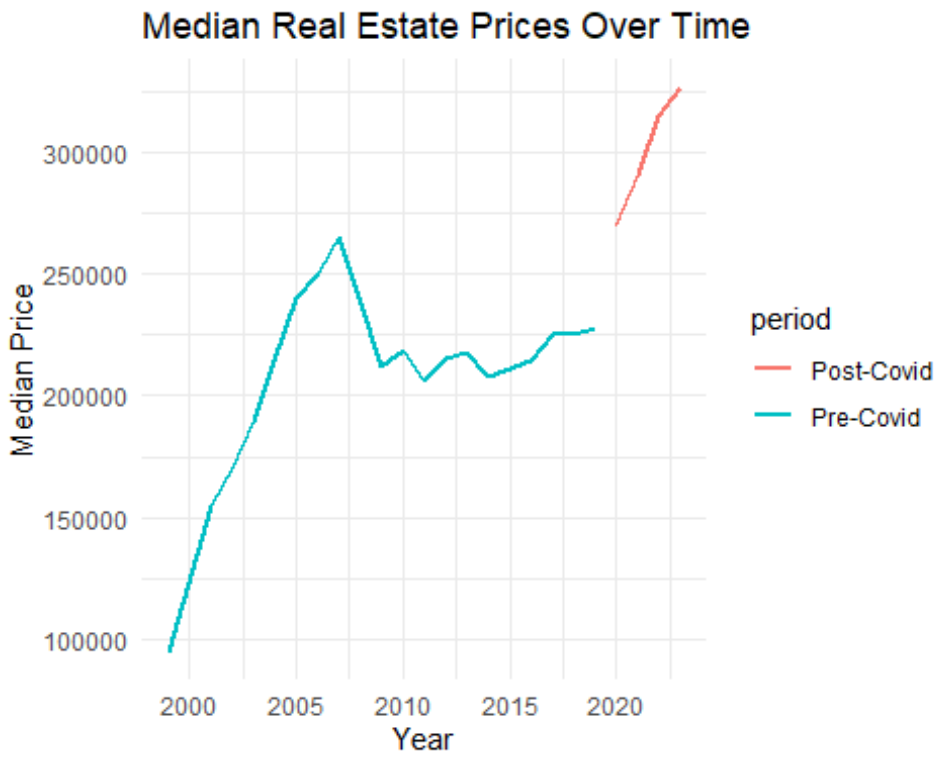
```
## spc_tbl_ [1,048,575 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Serial Number   : num [1:1048575] 220008 2020348 20002 210317 200212
...
##  $ List Year       : num [1:1048575] 2022 2020 2020 2021 2020 ...
##  $ Date Recorded   : Date[1:1048575], format: "2023-01-30" "2021-09-13"
...
##  $ Town            : chr [1:1048575] "Andover" "Ansonia" "Ashford" "Avon"
...
##  $ Address         : chr [1:1048575] "618 ROUTE 6" "230 WAKELEE AVE" "390
TURNPIKE RD" "53 COTSWOLD WAY" ...
##  $ Assessed Value  : num [1:1048575] 139020 150500 253000 329730 130400
...
##  $ Sale Amount     : num [1:1048575] 232000 325000 430000 805000 179900
...
##  $ Sales Ratio     : num [1:1048575] 0.599 0.463 0.588 0.41 0.725 ...
##  $ Property Type   : chr [1:1048575] "Residential" "Commercial"
"Residential" "Residential" ...
##  $ Residential Type: chr [1:1048575] "Single Family" NA "Single Family"
"Single Family" ...
##  $ Non Use Code    : chr [1:1048575] NA NA NA NA ...
##  $ Assessor Remarks: chr [1:1048575] NA NA NA NA ...
##  $ OPM remarks     : chr [1:1048575] NA NA NA NA ...
##  $ Location        : chr [1:1048575] "POINT (-72.343628962 41.728431984)"
NA NA "POINT (-72.846365959 41.781677018)" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Serial Number` = col_double(),
##   ..   `List Year` = col_double(),
##   ..   `Date Recorded` = col_date(format = ""),
##   ..   Town = col_character(),
```

```
##   ..   Address = col_character(),
##   ..   `Assessed Value` = col_double(),
##   ..   `Sale Amount` = col_double(),
##   ..   `Sales Ratio` = col_double(),
##   ..   `Property Type` = col_character(),
##   ..   `Residential Type` = col_character(),
##   ..   `Non Use Code` = col_character(),
##   ..   `Assessor Remarks` = col_character(),
##   ..   `OPM remarks` = col_character(),
##   ..   Location = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(org_data_set)
```

```
##  Serial Number         List Year    Date Recorded            Town
##  Min.   :0.000e+00   Min.   :2001   Min.   :1999-04-05   Length:1048575
##  1st Qu.:3.056e+04   1st Qu.:2004   1st Qu.:2005-09-09   Class :character
##  Median :8.007e+04   Median :2011   Median :2011-10-26   Mode  :character
##  Mean   :5.116e+05   Mean   :2011   Mean   :2012-02-16
##  3rd Qu.:1.607e+05   3rd Qu.:2017   3rd Qu.:2017-12-29
##  Max.   :2.001e+09   Max.   :2022   Max.   :2023-09-29
##                                     NA's   :2
##    Address          Assessed Value       Sale Amount         Sales Ratio
##  Length:1048575     Min.   :        0   Min.   :0.000e+00   Min.   :
0.0
##  Class :character   1st Qu.:    88340   1st Qu.:1.440e+05   1st Qu.:
0.5
##  Mode  :character   Median :   139730   Median :2.320e+05   Median :
0.6
##                     Mean   :   280725   Mean   :4.047e+05   Mean   :
10.0
##                     3rd Qu.:   226800   3rd Qu.:3.750e+05   3rd Qu.:
0.8
##                     Max.   :881510000   Max.   :5.000e+09   Max.
:1226420.0
##
##  Property Type      Residential Type   Non Use Code       Assessor Remarks
##  Length:1048575     Length:1048575     Length:1048575     Length:1048575
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  OPM remarks          Location
##  Length:1048575     Length:1048575
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
```

```
##
##
##

org_data_set$`Date Recorded`<- as.Date(org_data_set$`Date Recorded`, format=
"%Y-%m-%d")

org_data_set <- org_data_set[, c("Date Recorded", "Town", "Sale Amount")]

colSums(is.na(org_data_set))

## Date Recorded          Town   Sale Amount
##             2             0             0

org_data_set <- na.omit(org_data_set)

org_data_set$period <- ifelse(year(org_data_set$`Date Recorded`) < 2020,
"Pre-Covid", "Post-Covid")



yearly_prices <- aggregate(org_data_set$`Sale Amount` ~ year + period,
                           data = transform(org_data_set,
                                            year = year(`Date Recorded`)),
                           FUN = median)

str(yearly_prices)

## 'data.frame':    24 obs. of  3 variables:
##  $ year                     : num  2020 2021 2022 2023 1999 ...
##  $ period                   : chr  "Post-Covid" "Post-Covid" "Post-Covid"
"Post-Covid" ...
##  $ org_data_set$`Sale Amount`: num  270000 291000 315000 326500 95000 ...

colnames(yearly_prices)[which(names(yearly_prices) == "org_data_set$`Sale
Amount`")] <- "median_price"



ggplot(yearly_prices, aes(x = year, y = median_price, color = period)) +
  geom_line(size = 1) +
  labs(title = "Median Real Estate Prices Over Time",
       x = "Year", y = "Median Price") +
  theme_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Median Real Estate Prices Over Time
Median Price
300000
250000
200000
150000
100000
2000
2005
2010
2015
2020
Year
period
Post-Covid
Pre-Covid