This project explores housing sales across the United States in general to get a sense of the spatial distribution of the housing market, and then zooms in in to the metropolitan areas, to analyze homes sales between 2019 and 2022.

options(repos = list(CRAN="C:/Program Files/R/R-4.2.2/library"))

Hottest Housing Spots in the US Metropolitan Areas

================

Cynthia Adom-Portuphy

01-25-2023

## Table of Content

1. [SUMMARY](#SUMMARY)

2. [DATA](#DATA)

3. [ANALYSIS](#ANALYSIS)

4. [CONCLUSION AND RECOMMENDATIONS](#CONCLUSIONS AND RECOMMENDATIONS)

## SUMMARY

The Covid-19 pandemic has brought a lot of instability and unpredictability to the housing market, and for the past four years, the nation has gradually experienced a slight housing bubble; where sellers have had much control over how much homes costs and sellers are left with limited options. The sellers’ market reached its peak in late 2021, but from June 2022, high property taxes have resulted in a gradual slowdown.

For this project, I explore housing sales across the United States in general to get a sense of the spatial distribution of the housing market, and then zoomed in to the metropolitan areas, to analyze homes sales between 2019 and 2022.

Research Questions:

\* Which states had the highest and lowest number of houses sold between 2019 and 2021, and how was housing sales for 2022?

\* How has the Number of Houses Sold Changed between 2012-2022?

\* How did the metropolitan Areas perform in 2021 in Housing Sales?

\* Which Metropolitan Areas experienced the highest versus the lowest housing sales between 2019-2022?

\* What was the housing preference type for investors and buyers?

## 2. DATA

### 2.1 About the Data and Preprocessing

I downloaded real estate data for all the metropolitan divisions of the United States from REDFIN.

The dataset consisted of monthly data on housing for each of the metropolitan areas from 2012 to 2022. The data is provided by Redfin, a national real estate brokerage, and it is open to use with citation.

### 2.2.1 Excel

\* I opened the data in Excel and explored it for data cleaning and subsequent analysis.

--\* I removed the dates that were not relevant to the analysis and kept records covering 2019 and 2022 – covering the covid-19 season. I had to remove all unwanted columns, first. Since the dataset was very large, (591198 records), I removed all duplicates in the columns. I also fixed all structural errors.

#### 2.3.2 Big Query

\* After the excel clean-up and formatting, I imported the data into Big query, where I extracted the relevant data, aggregated, joined data and calculated fields to extract data needed for the analysis. Based on the sql analysis, I came up with csv 7 tables:

1. Housing\_Metro\_2019.csv

2. Housing\_Metro\_2021.csv

3. Housing\_Metro\_2022.csv

4. metro\_top10.csv

5. Metro\_housing\_type.csv

6. years.csv

### 2.3 Load the Packages and library

```{r}

install.packages("tidyverse")

install.packages("lubridate")

install.packages("ggplot2")

install.packages("plyr")

install.packages("tidyr")

install.packages("scales")

install.packages("dplyr")

install.packages("readr")

install.packages("magrittr")

library("tidyverse")

library("lubridate")

library("ggplot2")

library("tidyr")

library("scales")

library("plyr")

library("dplyr")

library("readr")

library("magrittr")

```

### 2.4 Import the Data

```{r}

Housing\_metro\_2019 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2019.csv")

Housing\_metro\_2021 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2021.csv")

View(Housing\_metro\_2021)

Housing\_metro\_2022 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2022.csv")

Housing\_metro\_top10 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/metro\_top10.csv")

Housing\_metro <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/metro\_housing\_CURRENT2.csv")

Housing\_years <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/years.csv")

View(Housing\_years)

```

### 2.5 Running Initial Statistics

```{r}

Housing\_2021\_stats <- Housing\_metro\_2021 %>%

summarise(average\_homes\_sold=mean(home\_sold),

average\_sales\_price=mean(sales\_price),

max\_homes\_sold=max(home\_sold),

min\_homes\_sold =min(home\_sold),

sd\_homes\_sold =sd(home\_sold))

View(Housing\_2021\_stats)

glimpse(Housing\_2021\_stats)

```

### 2.6 Grouping the Homes Sold column into 4 Categories

\* I grouped the metropolitan areas based on the number of homes sold into four groups for subsequent analysis.

\* The average of homes sold for all metro areas is 21,502 homes, minimum was 2 and maximum was 370,376 homes, and standard deviation of 58349.82. I segmented the data points into four, based on the data distribution. The first category, "low market" was all metro areas that had number of homes sold below the mean. Category "medium market" was one standard deviation above the mean, the third, "medium high market" was two standard deviations above the mean. The last category, "high market" was all other metros.

```{r}

Housing\_metro\_2021<-Housing\_metro\_2021 %>%

mutate (home\_groups = case\_when(

home\_sold < 21502 ~ "low market",

home\_sold >= 21502 & home\_sold < 79851 ~ "Medium Market",

home\_sold >=79851 & home\_sold < 101353 ~ "Medium High Market",

home\_sold >=101353~ "High Market"))

```

\* I created a percent dataframe to standardize the distribution for the four categories.

```{r}

Home\_sold\_percent <- Housing\_metro\_2021 %>%

group\_by(home\_groups) %>%

dplyr::summarise(total = n())%>%

mutate(total\_sold = sum(total)) %>%

group\_by(home\_groups)%>%

summarise(total\_percent = total/total\_sold)%>%

mutate(labels = scales::percent(total\_percent))

Home\_Sold\_Category = c("low market","Medium Market", "Medium High Market", "High Market")

Home\_sold\_percent2 <-cbind(Home\_sold\_percent, Home\_Sold\_Category)

View(Home\_sold\_percent2)

```

## 3. Analysis

### 3.1 Which states had the highest and lowest housing sales between 2019 and 2021?

\* To initially get a general overview of the number of houses sold across the United States, I created a map in Tableau using the home sold column to spatially display the distribution of the number of homes sold at the state level.

![Housing Sales in 2022 Vs. Percentage Change in Sales 2019-2021](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Maps.png)

\* The map below shows that for 2019-2021 the states that experienced the highest housing sales included Alabama, Illinois, New York, Florida, California and Texas. Some of lowest sales occurred in Montana, North Dakota, West Virginia. Wyoming had no data.

`

### 3.2 How has the Number of Houses SOld Changed between 2012-2022?

```{r}

options(scipen = 5)

ggplot(data = Housing\_years) +geom\_line(aes(x= Year, y = houses\_sold), color = "#008b9e") + labs(title = "Number of Houses Sold between 2012 and 2022", x = "year", y = "Number of houses sold")

```

\* The results showed that the housing market has seen a steady increase from 2012 to 2022. It peaked in 2021 and started declining late 2021 into 2022.

### 3.3 How did the metropolitan Areas perform in 2021 in Housing Sales?

\* I used the grouped data of the Metropolitan Areas based on the number of houses sold to create a pie chart.

```{r}

ggplot(Home\_sold\_percent2, aes(x="",y=total\_percent, fill=Home\_Sold\_Category)) +

geom\_bar(stat = "identity", width = 1)+

coord\_polar("y", start=0)+

theme\_minimal()+

theme(axis.title.x= element\_blank(),

axis.title.y = element\_blank(),

panel.border = element\_blank(),

panel.grid = element\_blank(),

axis.ticks = element\_blank(),

axis.text.x = element\_blank(),

plot.title = element\_text(hjust = 0.5, size=14, face = "bold")) +

scale\_fill\_manual(values=c("#CCE6FF", "#3399FF", "#004D99" ,"#008b9e"))+ geom\_text(aes(label = labels),

position = position\_stack(vjust = 0.5))

labs(title = "Housing Market Size for US Metropolitan Areas (Division) in 2021")

```

\* The results showed that majority of the metropolitan areas (82% had a fairly good market (medium market) in 2021.The 10% actually fell within the category of high market; while only 6% had a low market.

### 3.4 Which Metropolitan Areas experienced the highest and lowest housing sales between 2019-2022?

\* I selected the top 10 and bottom 10 ranking for the number of homes sold for the metropolitan areas using Rstudio.

```{r}

top10<- top\_n(Housing\_metro\_2021, 10, home\_sold)

View(top10)

lowest10<- top\_n(Housing\_metro\_2021, -10, home\_sold)

```

```{r}

options(scipen = 5)

ggplot(data = Housing\_metro\_top10) + geom\_histogram(fill="#008b9e", stat = "identity", aes(x = parent\_metro\_region, y= houses\_sold))+

theme(axis.text.x = element\_text(angle = 90)) +

labs(title="Number of Houses Sold in Top Ten US Metros - 2021", x = "Metropolitan Areas", y = "Number of Houses Sold")

```

\* Atlanta, GA came on top, followed by Chicago, IL. These rest included Houston, TX and Washington, DC. Metros that had the lowest number of homes sold include Selinsgrove, PA, Ruston, LA, West Point, MS and Bradford, PA (See Table below).

![Top 10 Metros with highest number of homes sold. ](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Top10.png)

![Bottom 10 Metros with highest number of homes sold. ](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Bottom10.png)

### 3.5 What was the housing preference type for investors and buyers?

```{r}

```

3.4

```{r}

ggplot(data = Housing\_metro) +geom\_bar(fill="#008b9e", aes(x = property\_type)) + theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Property Type Among Investors and Buyers", x = "Property Type", y = "Count")

```

\* As expected, single-family residential properties were the most popular among buyers and investors in 2021. This was followed by condos and then townhouses.

## 4. Conclusions and Recommendations

\* The steady progressing of the housing market, during and after covid-19 was spread across all regions of the United States. However, the peak of the market in 2022 concentrated more in the South, and to some extent the north-east.

\* Majority of the metropolitan areas had a very good market overall, underscoring th the fact that people were buying houses everyone, but in 2022, when the tax policies made it harder for more people to buy and property owners less willing to put their houses on the market, the southern states: Florida, Texas and California mostly still enjoyed a housing boom.

\* Single family houses was the most preferred housing among buyers and investors, but condos also had an appreciable percentage as well. This is in line with realtor.com that the "gleaming luxury condo towers with amenities like pools and swanky bars may monopolize the cool factor, but today’s pragmatic home buyers are clamoring for something different: affordable, single-family homes, often in the suburbs."

\* The housing market is gradually on a decline; however, this is not expected to be drastic, considering the extent to which housing sales are still high though incomes have remained steady while cost of living keeps rising. This means that certain categories of people will be left out of the housing market, especially first time home buyers. It will be helpful for taxes to be adjusted so that people who need housing can afford.

\* For the future, I plan to extend this study to include socio-economic data to better understand these trends.

Hottest Housing Spots in the US Metropolitan Areas

================

Cynthia Adom-Portuphy

01-25-2023

## Table of Content

1. [SUMMARY](#SUMMARY)

2. [DATA](#DATA)

3. [ANALYSIS](#ANALYSIS)

4. [CONCLUSION AND

RECOMMENDATIONS](#CONCLUSIONS%20AND%20RECOMMENDATIONS)

## SUMMARY

The Covid-19 pandemic has brought a lot of instability and

unpredictability to the housing market, and for the past four years, the

nation has gradually experienced a slight housing bubble; where sellers

have had much control over how much homes costs and sellers are left

with limited options. The sellers’ market reached its peak in late 2021,

but from June 2022, high property taxes have resulted in a gradual

slowdown.

For this project, I explore housing sales across the United States in

general to get a sense of the spatial distribution of the housing

market, and then zoomed in to the metropolitan areas, to analyze homes

sales between 2019 and 2022.

Research Questions:

- Which states had the highest and lowest number of houses sold between

2019 and 2021, and how was housing sales for 2022?

- How has the Number of Houses SOld Changed between 2012-2022?

- How did the metropolitan Areas perform in 2021 in Housing Sales?

- Which Metropolitan Areas experienced the highest versus the lowest

housing sales between 2019-2022?

- What was the housing preference type for investors and buyers?

## 2. DATA

### 2.1 About the Data and Preprocessing

I downloaded real estate data for all the metropolitan divisions of the

United States from REDFIN. The dataset consisted of monthly data on

housing for each of the metropolitan areas from 2012 to 2022. The data

is provided by Redfin, a national real estate brokerage, and it is open

to use with citation.

### 2.2.1 Excel

- I opened the data in Excel and explored it for data cleaning and

subsequent analysis. –\\*I removed the dates that were not relevant to

the analysis and kept records covering 2019 and 2022 – covering the

covid-19 season. I had to remove all unwanted columns, first. Since

the dataset was very large, (591198 records), I removed all duplicates

in the columns. I also fixed all structural errors.

#### 2.3.2 R-Studio

- After the excel clean-up and formatting, I imported the data into

RStudio, where I extracted the relevant data, aggregated, joined data

and calculated fields to extract data needed for the analysis.

### 2.3 Load the Packages and library

``` r

options(repos = list(CRAN="C:/Program Files/R/R-4.2.2/library"))

install.packages("tidyverse")

install.packages("lubridate")

install.packages("ggplot2")

install.packages("plyr")

install.packages("tidyr")

install.packages("scales")

install.packages("dplyr")

install.packages("readr")

install.packages("magrittr")

library("tidyverse")

library("lubridate")

library("ggplot2")

library("tidyr")

library("scales")

library("plyr")

library("dplyr")

library("readr")

library("magrittr")

```

### 2.4 Import the Data

``` r

Housing\_metro\_2019 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2019.csv")

Housing\_metro\_2021 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2021.csv")

View(Housing\_metro\_2021)

Housing\_metro\_2022 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Housing\_Metro\_2022.csv")

Housing\_metro\_top10 <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/metro\_top10.csv")

Housing\_metro <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/metro\_housing\_type.csv")

Housing\_years <-readr:::read\_csv("C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/years.csv")

View(Housing\_years)

```

### 2.5 Running Initial Statistics

``` r

Housing\_2021\_stats <- Housing\_metro\_2021 %>%

summarise(average\_homes\_sold=mean(home\_sold),

average\_sales\_price=mean(sales\_price),

max\_homes\_sold=max(home\_sold),

min\_homes\_sold =min(home\_sold),

sd\_homes\_sold =sd(home\_sold))

View(Housing\_2021\_stats)

glimpse(Housing\_2021\_stats)

```

## Rows: 1

## Columns: 5

## $ average\_homes\_sold <dbl> 21502.44

## $ average\_sales\_price <dbl> 17741785

## $ max\_homes\_sold <dbl> 489045.6

## $ min\_homes\_sold <dbl> 2

## $ sd\_homes\_sold <dbl> 58349.82

### 2.6 Grouping the Homes Sold column into 4 Categories

- I grouped the metropolitan areas based on the number of homes sold

into four groups for subsequent analysis.

- The average of homes sold for all metro areas is 21,502 homes, minimum

was 2 and maximum was 370,376 homes, and standard deviation of

58349.82. I segmented the data points into four, based on the data

distribution. The first category, “low market” was all metro areas

that had number of homes sold below the mean. Category “medium market”

was one standard deviation above the mean, the third, “medium high

market” was two standard deviations above the mean. The last category,

“high market” was all other metros.

``` r

Housing\_metro\_2021<-Housing\_metro\_2021 %>%

mutate (home\_groups = case\_when(

home\_sold < 21502 ~ "low market",

home\_sold >= 21502 & home\_sold < 79851 ~ "Medium Market",

home\_sold >=79851 & home\_sold < 101353 ~ "Medium High Market",

home\_sold >=101353~ "High Market"))

```

- I created a percent dataframe to standardize the distribution for the

four categories.

``` r

Home\_sold\_percent <- Housing\_metro\_2021 %>%

group\_by(home\_groups) %>%

dplyr::summarise(total = n())%>%

mutate(total\_sold = sum(total)) %>%

group\_by(home\_groups)%>%

summarise(total\_percent = total/total\_sold)%>%

mutate(labels = scales::percent(total\_percent))

Home\_Sold\_Category = c("low market","Medium Market", "Medium High Market", "High Market")

Home\_sold\_percent2 <-cbind(Home\_sold\_percent, Home\_Sold\_Category)

View(Home\_sold\_percent2)

```

## 3. Analysis

### 3.1 Which states had the highest and lowest housing sales between 2019 and 2021?

- To initially get a general overview of the number of houses sold

across the United States, I created a map in Tableau using the home

sold column to spatially display the distribution of the number of

homes sold at the state level.

![Housing Sales in 2022 Vs. Percentage Change in Sales

2019-2021](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Maps.png)

\\* The map below shows that for 2019-2021 the states that experienced

the highest housing sales included Alabama, Illinois, New York, Florida,

California and Texas. Some of lowest sales occurred in Montana, North

Dakota, West Virginia. Wyoming had no data. \`

### 3.2 How has the Number of Houses SOld Changed between 2012-2022?

``` r

options(scipen = 5)

ggplot(data = Housing\_years) +geom\_line(aes(x= Year, y = houses\_sold), color = "#008b9e") + labs(title = "Number of Houses Sold between 2012 and 2022", x = "year", y = "Number of houses sold")

```

![](Housing\_Metro\_files/figure-gfm/unnamed-chunk-6-1.png)<!-- -->

- The results showed that the housing market has seen a steady increase

from 2012 to 2022. It peaked in 2021 and started declining late 2021

into 2022.

### 3.3 How did the metropolitan Areas perform in 2021 in Housing Sales?

- I used the grouped data of the Metropolitan Areas based on the number

of houses sold to create a pie chart.

``` r

ggplot(Home\_sold\_percent2, aes(x="",y=total\_percent, fill=Home\_Sold\_Category)) +

geom\_bar(stat = "identity", width = 1)+

coord\_polar("y", start=0)+

theme\_minimal()+

theme(axis.title.x= element\_blank(),

axis.title.y = element\_blank(),

panel.border = element\_blank(),

panel.grid = element\_blank(),

axis.ticks = element\_blank(),

axis.text.x = element\_blank(),

plot.title = element\_text(hjust = 0.5, size=14, face = "bold")) +

scale\_fill\_manual(values=c("#CCE6FF", "#3399FF", "#004D99" ,"#008b9e"))+ geom\_text(aes(label = labels),

position = position\_stack(vjust = 0.5))

```

![](Housing\_Metro\_files/figure-gfm/unnamed-chunk-7-1.png)<!-- -->

``` r

labs(title = "Housing Market Size for US Metropolitan Areas (Division) in 2021")

```

## $title

## [1] "Housing Market Size for US Metropolitan Areas (Division) in 2021"

##

## attr(,"class")

## [1] "labels"

- The results showed that majority of the metropolitan areas (82% had a

fairly good market (medium market) in 2021.The 10% actually fell

within the category of high market; while only 6% had a low market.

### 3.4 Which Metropolitan Areas experienced the highest and lowest housing sales between 2019-2022?

- I selected the top 10 and bottom 10 ranking for the number of homes

sold for the metropolitan areas using Rstudio.

``` r

top10<- top\_n(Housing\_metro\_2021, 10, home\_sold)

View(top10)

lowest10<- top\_n(Housing\_metro\_2021, -10, home\_sold)

```

``` r

options(scipen = 5)

ggplot(data = Housing\_metro\_top10) + geom\_histogram(fill="#008b9e", stat = "identity", aes(x = parent\_metro\_region, y= houses\_sold))+

theme(axis.text.x = element\_text(angle = 90)) +

labs(title="Number of Houses Sold in Top Ten US Metros - 2021", x = "Metropolitan Areas", y = "Number of Houses Sold")

```

![](Housing\_Metro\_files/figure-gfm/unnamed-chunk-9-1.png)<!-- -->

- Atlanta, GA came on top, followed by Chicago, IL. These rest included

Houston, TX and Washington, DC. Metros that had the lowest number of

homes sold include Selinsgrove, PA, Ruston, LA, West Point, MS and

Bradford, PA (See Table below).

![Top 10 Metros with highest number of homes

sold.](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Top10.png)

![Bottom 10 Metros with highest number of homes

sold.](C:/Users/cynta/OneDrive/Desktop/Project/housing\_sale\_trends\_postcovid/Bottom10.png)

### 3.5 What was the housing preference type for investors and buyers?

3.4

``` r

ggplot(data = Housing\_metro) +geom\_bar(fill="#008b9e", aes(x = property\_type)) + theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Property Type Among Investors and Buyers", x = "Property Type", y = "Count")

```

![](Housing\_Metro\_files/figure-gfm/unnamed-chunk-11-1.png)<!-- -->

- As expected, single-family residential properties were the most

popular among buyers and investors in 2021. This was followed by

condos and then townhouses.

## 4. Conclusions and Recommendations

- The steady progressing of the housing market, during and after

covid-19 was spread across all regions of the United States. However,

the peak of the market in 2022 concentrated more in the South, and to

some extent the north-east.

- Majority of the metropolitan areas had a very good market overall,

underscoring th the fact that people were buying houses everyone, but

in 2022, when the tax policies made it harder for more people to buy

and property owners less willing to put their houses on the market,

the southern states: Florida, Texas and California mostly still

enjoyed a housing boom.

- Single family houses was the most preferred housing among buyers and

investors, but condos also had an appreciable percentage as well. This

is in line with realtor.com that the “gleaming luxury condo towers

with amenities like pools and swanky bars may monopolize the cool

factor, but today’s pragmatic home buyers are clamoring for something

different: affordable, single-family homes, often in the suburbs.”

- The housing market is gradually on a decline; however, this is not

expected to be drastic, considering the extent to which housing sales

are still high though incomes have remained steady while cost of

living keeps rising. This means that certain categories of people will

be left out of the housing market, especially first time home buyers.

It will be helpful for taxes to be adjusted so that people who need

housing can afford.

- For the future, I plan to extend this study to include socio-economic

data to better understand these trends.