# MIS503 - Final Project

## Angelo, Elizabeth

### Zillow Home Value Index Analysis

### Wake County Home Sales

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── 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(dplyr)  
library(ggplot2)  
library(readr)

sales\_data <- read\_csv("SingleFamilyResidenceSales.csv")

## Rows: 22275 Columns: 287  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): RegionName, State, Metro, CountyName  
## dbl (283): RegionID, 1/31/2000, 2/29/2000, 3/31/2000, 4/30/2000, 5/31/2000, ...  
##   
## 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.

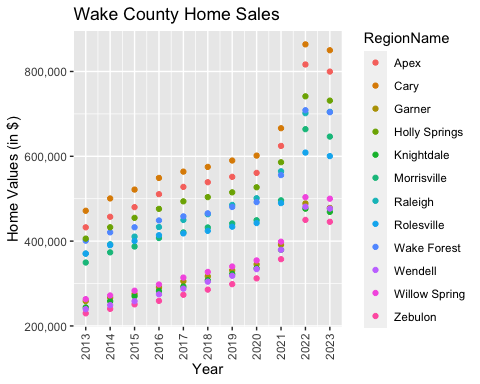
# Mutate   
WakeCountySales <- sales\_data %>%  
 mutate(  
 RegionName = as.character(RegionName),  
 State = as.character(State),  
 CountyName = as.character(CountyName),  
 Metro = as.character(Metro),  
 `5/31/2013` = as.numeric(`5/31/2013`),  
 `5/31/2014` = as.numeric(`5/31/2014`),  
 `5/31/2015` = as.numeric(`5/31/2015`),  
 `5/31/2016` = as.numeric(`5/31/2016`),  
 `5/31/2017` = as.numeric(`5/31/2017`),  
 `5/31/2018` = as.numeric(`5/31/2018`),  
 `5/31/2019` = as.numeric(`5/31/2019`),  
 `5/31/2020` = as.numeric(`5/31/2020`),  
 `5/31/2021` = as.numeric(`5/31/2021`),  
 `5/31/2022` = as.numeric(`5/31/2022`),  
 `5/31/2023` = as.numeric(`5/31/2023`)  
 )  
  
# WakeCountySales tibble  
WakeCountySales <- sales\_data %>%  
 filter(State == "NC" & CountyName == "Wake County") %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 `2013` = `5/31/2013`,  
 `2014` = `5/31/2014`,  
 `2015` = `5/31/2015`,  
 `2016` = `5/31/2016`,  
 `2017` = `5/31/2017`,  
 `2018` = `5/31/2018`,  
 `2019` = `5/31/2019`,  
 `2020` = `5/31/2020`,  
 `2021` = `5/31/2021`,  
 `2022` = `5/31/2022`,  
 `2023` = `5/31/2023`  
 )  
  
head(WakeCountySales)

## # A tibble: 6 × 15  
## RegionName State CountyName Metro `2013` `2014` `2015` `2016` `2017` `2018`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Raleigh NC Wake Coun… Rale… 3.70e5 3.93e5 4.11e5 4.33e5 4.50e5 4.63e5  
## 2 Cary NC Wake Coun… Rale… 4.72e5 5.01e5 5.22e5 5.49e5 5.64e5 5.75e5  
## 3 Apex NC Wake Coun… Rale… 4.33e5 4.57e5 4.80e5 5.11e5 5.28e5 5.39e5  
## 4 Wake Forest NC Wake Coun… Rale… 4.01e5 4.21e5 4.33e5 4.49e5 4.58e5 4.66e5  
## 5 Garner NC Wake Coun… Rale… 2.59e5 2.67e5 2.76e5 2.91e5 3.05e5 3.17e5  
## 6 Holly Springs NC Wake Coun… Rale… 4.06e5 4.33e5 4.55e5 4.76e5 4.94e5 5.04e5  
## # ℹ 5 more variables: `2019` <dbl>, `2020` <dbl>, `2021` <dbl>, `2022` <dbl>,  
## # `2023` <dbl>

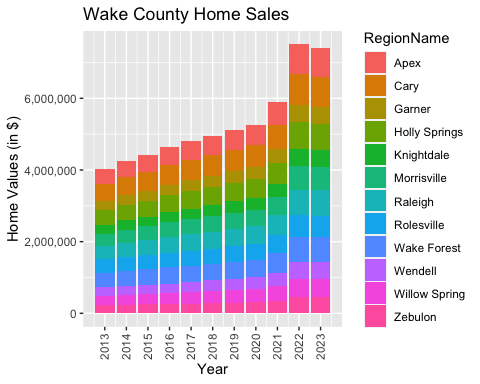
# Tidy Data  
sales\_tidy <- WakeCountySales %>%  
 gather(key = 'YR', value = 'ZHVI', -RegionName, -State, -CountyName, -Metro) %>%  
 mutate(  
 YR = as.character(YR), # Convert YR to character  
 YR = str\_extract(YR, "\\d{4}") %>% as.numeric() # Extract only the four-digit year  
 ) %>%  
 filter(!is.na(YR))  
  
print(sales\_tidy)

## # A tibble: 132 × 6  
## RegionName State CountyName Metro YR ZHVI  
## <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 Raleigh NC Wake County Raleigh-Cary 2013 370453.  
## 2 Cary NC Wake County Raleigh-Cary 2013 471561.  
## 3 Apex NC Wake County Raleigh-Cary 2013 432532.  
## 4 Wake Forest NC Wake County Raleigh-Cary 2013 401073.  
## 5 Garner NC Wake County Raleigh-Cary 2013 258592.  
## 6 Holly Springs NC Wake County Raleigh-Cary 2013 405887.  
## 7 Morrisville NC Wake County Raleigh-Cary 2013 349518.  
## 8 Knightdale NC Wake County Raleigh-Cary 2013 243654.  
## 9 Zebulon NC Wake County Raleigh-Cary 2013 229737.  
## 10 Wendell NC Wake County Raleigh-Cary 2013 239614.  
## # ℹ 122 more rows

# scatter plot  
scatter\_plot <- ggplot(sales\_tidy, aes(x = YR, y = ZHVI, color = RegionName)) +  
 geom\_point() +  
 labs(  
 title = "Wake County Home Sales",  
 x = "Year",  
 y = "Home Values (in $)"  
 ) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1)) +  
 scale\_x\_continuous(name = "Year", breaks = seq(min(sales\_tidy$YR), max(sales\_tidy$YR), by  
 = 1)) +  
 scale\_y\_continuous(name = "Home Values (in $)", labels = scales::comma)  
  
print(scatter\_plot)



# Stacked Bar Graph  
stacked\_bar <- ggplot(sales\_tidy, aes(x = YR, y = ZHVI, fill = RegionName)) +  
 geom\_bar(stat = "identity") +  
 labs(  
 title = "Wake County Home Sales",  
 x = "Year",  
 y = "Total Home Values"  
 ) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5)) +  
 scale\_x\_continuous(name = "Year", breaks = seq(min(sales\_tidy$YR), max(sales\_tidy$YR), by = 1)) +  
 scale\_y\_continuous(name = "Home Values (in $)", labels = scales::comma)  
  
print(stacked\_bar)



What have been the overall trends in Wake County Home Values?  
The value of homes in Wake County have been steadily increasing over the last ten years.

There were dips in home values in the past 10 years. What years did these occur?  
There were not many dips in home value in Wake County over the past 10 years, but there was a noticeable dip in 2023.

Based on the analysis, where would be the least expensive area to purchase home? Most expensive area?  
Currently in 2023, the least expensive area to purchase a home would be Zebulon. The most expensive area would be Cary.

What has happened to the overall property values in Apex and Cary in 2023?  
They have decreased slightly but remained high in value.

### NC Rental Market

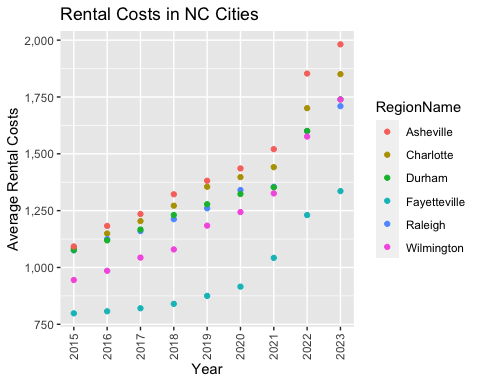
rental\_data <- read\_csv("SingleFamilyResidenceRental.csv")

## Rows: 3503 Columns: 107  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): RegionName, State, Metro, CountyName  
## dbl (103): RegionID, 1/31/2015, 2/28/2015, 3/31/2015, 4/30/2015, 5/31/2015, ...  
##   
## 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.

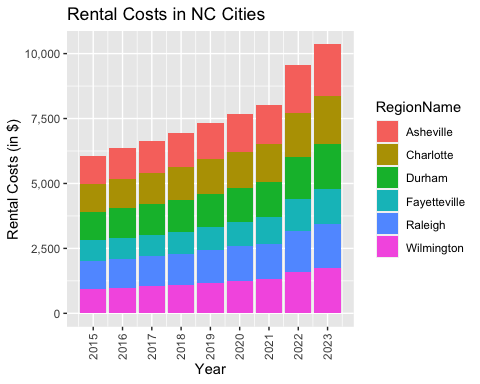
# Convert columns  
Rentals <- rental\_data %>%  
 mutate(  
 RegionName = as.character(RegionName),  
 State = as.character(State),  
 `1/31/2015` = as.numeric(`1/31/2015`),  
 `1/31/2016` = as.numeric(`1/31/2016`),  
 `1/31/2017` = as.numeric(`1/31/2017`),  
 `1/31/2018` = as.numeric(`1/31/2018`),  
 `1/31/2019` = as.numeric(`1/31/2019`),  
 `1/31/2020` = as.numeric(`1/31/2020`),  
 `1/31/2021` = as.numeric(`1/31/2021`),  
 `1/31/2022` = as.numeric(`1/31/2022`),  
 `1/31/2023` = as.numeric(`1/31/2023`)  
 )  
  
# Filter and select   
Rentals <- Rentals %>%  
 filter(State == "NC" & RegionName %in% c("Asheville", "Charlotte", "Durham", "Fayetteville", "Raleigh", "Wilmington")) %>%  
 select(  
 RegionName,  
 State,  
 `1/31/2015` = `1/31/2015`,  
 `1/31/2016` = `1/31/2016`,  
 `1/31/2017` = `1/31/2017`,  
 `1/31/2018` = `1/31/2018`,  
 `1/31/2019` = `1/31/2019`,  
 `1/31/2020` = `1/31/2020`,  
 `1/31/2021` = `1/31/2021`,  
 `1/31/2022` = `1/31/2022`,  
 `1/31/2023` = `1/31/2023`  
 )  
  
# Tidy   
rentals\_tidy <- Rentals %>%  
 gather(key = 'YR', value = 'ZHVI', -RegionName, -State) %>%  
 mutate(  
 YR = as.numeric(str\_extract(YR, "\\d{4}")) # Extract only the four-digit year  
 ) %>%  
 filter(!is.na(YR) & !is.na(ZHVI)) # Filter out rows with missing ZHVI values  
  
print(rentals\_tidy)

## # A tibble: 54 × 4  
## RegionName State YR ZHVI  
## <chr> <chr> <dbl> <dbl>  
## 1 Charlotte NC 2015 1086.  
## 2 Raleigh NC 2015 1075.  
## 3 Durham NC 2015 1077.  
## 4 Fayetteville NC 2015 798.  
## 5 Wilmington NC 2015 945.  
## 6 Asheville NC 2015 1093.  
## 7 Charlotte NC 2016 1149.  
## 8 Raleigh NC 2016 1125.  
## 9 Durham NC 2016 1119.  
## 10 Fayetteville NC 2016 807.  
## # ℹ 44 more rows

#Scatter  
rental\_scatter <- ggplot(rentals\_tidy, aes(x = YR, y = ZHVI, color = RegionName)) +  
 geom\_point() +  
 labs(title = "Rental Costs in NC Cities",  
 x = "Year",  
 y = "Average Rental Costs") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5)) +  
 scale\_x\_continuous(breaks = unique(rentals\_tidy$YR), minor\_breaks = NULL) +  
 scale\_y\_continuous(name = "Average Rental Costs", labels = scales::comma)  
  
print(rental\_scatter)



# Stacked Bar  
stacked\_bar\_rentals <- ggplot(rentals\_tidy, aes(x = YR, y = ZHVI, fill = RegionName)) +  
 geom\_bar(stat = "identity") +  
 labs(  
 title = "Rental Costs in NC Cities",  
 x = "Year",  
 y = "Total Rental Costs"  
 ) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5)) +  
 scale\_x\_continuous(  
 name = "Year",  
 breaks = seq(min(rentals\_tidy$YR), max(rentals\_tidy$YR), by = 1)  
 ) +  
 scale\_y\_continuous(name = "Rental Costs (in $)", labels = scales::comma)  
  
print(stacked\_bar\_rentals)



What has been the overall trend in the rental market around the state? Are there any cities that have not followed this trend?  
The trend in the NC rental market has steadily increased over the last 9 years. It appears that Asheville, Charlotte, Wilmington, and Raleigh had a spike in rental cost between 2021 and 2023. Most of the cities have followed the trend of increasing rental costs.

Where is the most expensive city to rent in? Least expensive?  
Asheville is currently the most expensive city to rent in, while Fayetteville is the least expensive.

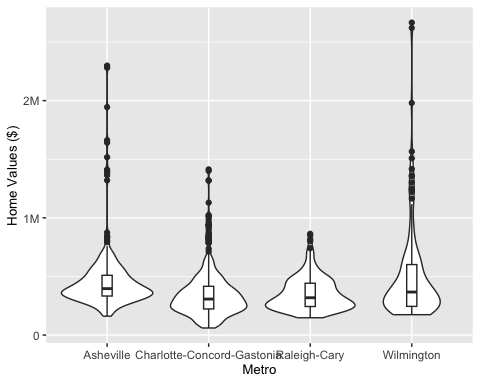
You are trying to decide between Wilmington and Asheville. Which market has the lowest rent?  
Wilmington has the lower rent compared to Asheville.

### Home Values in Select Markets

sales\_data <- read\_csv("SingleFamilyResidenceSales.csv")

## Rows: 22275 Columns: 287  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): RegionName, State, Metro, CountyName  
## dbl (283): RegionID, 1/31/2000, 2/29/2000, 3/31/2000, 4/30/2000, 5/31/2000, ...  
##   
## 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.

NCHomeSales <- sales\_data %>%  
 filter(State == "NC" & Metro %in% c("Asheville", "Charlotte-Concord-Gastonia", "Raleigh-Cary", "Wilmington")) %>%  
 select(  
 RegionName,  
 State,  
 Metro,  
 '2013' = '5/31/2013',  
 '2014' = '5/31/2014',  
 '2015' = '5/31/2015',  
 '2016' = '5/31/2016',  
 '2017' = '5/31/2017',  
 '2018' = '5/31/2018',  
 '2019' = '5/31/2019',  
 '2020' = '5/31/2020',  
 '2021' = '5/31/2021',  
 '2022' = '5/31/2022',  
 '2023' = '5/31/2023'  
 )  
  
# Tidy  
NCHomeSales\_tidy <- NCHomeSales %>%  
 gather(key = 'YR', value = 'ZHVI', -RegionName, -State, -Metro) %>%  
 mutate(  
 YR = as.character(YR), # Convert YR to character  
 YR = str\_extract(YR, "\\d{4}") %>% as.numeric() # Extract only the four-digit year  
 ) %>%  
 filter(!is.na(YR) & !is.na(ZHVI)) # Filter out rows with missing ZHVI values  
  
# Group by Metro  
NCHomeSales\_grouped <- NCHomeSales\_tidy %>%  
 group\_by(Metro)  
  
# Create combined violin and boxplot  
combined\_plot <- ggplot(NCHomeSales\_grouped, aes(x = Metro, y = ZHVI)) +  
 geom\_violin() +  
 geom\_boxplot(width = 0.1) +  
 labs(  
 x = "Metro",  
 y = "Home Values (in millions $)"  
 ) +  
 theme(  
 axis.text.x = element\_text(angle = 0, hjust = 0.5), # Horizontal x-axis labels  
 axis.text.y = element\_text(angle = 0, hjust = 1), # Horizontal y-axis labels  
 axis.title.y = element\_text(size = 10), # Adjust the size of y-axis title  
 axis.title.x = element\_text(size = 10), # Adjust the size of x-axis title  
 plot.title = element\_text(size = 12) # Adjust the size of plot title  
 ) +  
 scale\_y\_continuous(  
 name = "Home Values ($)",  
 labels = function(x) ifelse(x == 0, "0", scales::unit\_format(unit = "M", scale = 1e-6, sep = "", suffix = "M")(x)),  
 breaks = c(0, 1, 2) \* 1e6 # Set the y-axis breaks  
 )  
  
print(combined\_plot)



According to the results, which market has the lowest median price (represented as horizontal bar in box plot)?  
Charlotte-Concord-Gastonia have the lowest median price.

The violin plot will show density meaning the wider the plot is, the more observations occur within that area. Which market has the most density around the median value of homes?  
It appears that both Asheville and Raleigh have the most density around the median value of homes. They both appear to be around the same width.

The box plot will also show outliers in the various markets. Which metro area had the largest outlier (i.e., the highest value home sold in the past 10 years)?  
Wilmington had the largest outlier.

### Relocation Home Value Comparison

selected\_cities <- c("Chicago", "Denver", "Houston", "New York")  
  
# Filter Chicago  
NationalHomeSales <- sales\_data %>%  
 filter(  
 RegionName %in% selected\_cities,  
 State == "IL" # Filter for the state of New York  
 ) %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 '5/31/2013' = '5/31/2013',  
 '5/31/2014' = '5/31/2014',  
 '5/31/2015' = '5/31/2015',  
 '5/31/2016' = '5/31/2016',  
 '5/31/2017' = '5/31/2017',  
 '5/31/2018' = '5/31/2018',  
 '5/31/2019' = '5/31/2019',  
 '5/31/2020' = '5/31/2020',  
 '5/31/2021' = '5/31/2021',  
 '5/31/2022' = '5/31/2022',  
 '5/31/2023' = '5/31/2023'  
 )  
  
print(NationalHomeSales)

## # A tibble: 1 × 15  
## RegionName State CountyName Metro `5/31/2013` `5/31/2014` `5/31/2015`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Chicago IL Cook County Chicago-Nape… 409180. 455543. 470035.  
## # ℹ 8 more variables: `5/31/2016` <dbl>, `5/31/2017` <dbl>, `5/31/2018` <dbl>,  
## # `5/31/2019` <dbl>, `5/31/2020` <dbl>, `5/31/2021` <dbl>, `5/31/2022` <dbl>,  
## # `5/31/2023` <dbl>

selected\_cities <- c("Chicago", "Denver", "Houston", "New York")  
  
# Filter Denver  
NationalHomeSales <- sales\_data %>%  
 filter(  
 RegionName %in% selected\_cities,  
 State == "CO" # Filter for the state of New York  
 ) %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 '5/31/2013' = '5/31/2013',  
 '5/31/2014' = '5/31/2014',  
 '5/31/2015' = '5/31/2015',  
 '5/31/2016' = '5/31/2016',  
 '5/31/2017' = '5/31/2017',  
 '5/31/2018' = '5/31/2018',  
 '5/31/2019' = '5/31/2019',  
 '5/31/2020' = '5/31/2020',  
 '5/31/2021' = '5/31/2021',  
 '5/31/2022' = '5/31/2022',  
 '5/31/2023' = '5/31/2023'  
 )  
  
print(NationalHomeSales)

## # A tibble: 1 × 15  
## RegionName State CountyName Metro `5/31/2013` `5/31/2014` `5/31/2015`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Denver CO Denver County Denver-Aur… 468904. 512596. 567297.  
## # ℹ 8 more variables: `5/31/2016` <dbl>, `5/31/2017` <dbl>, `5/31/2018` <dbl>,  
## # `5/31/2019` <dbl>, `5/31/2020` <dbl>, `5/31/2021` <dbl>, `5/31/2022` <dbl>,  
## # `5/31/2023` <dbl>

selected\_cities <- c("Chicago", "Denver", "Houston", "New York")  
  
# Filter Houston  
NationalHomeSales <- sales\_data %>%  
 filter(  
 RegionName %in% selected\_cities,  
 State == "TX" # Filter for the state of New York  
 ) %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 '5/31/2013' = '5/31/2013',  
 '5/31/2014' = '5/31/2014',  
 '5/31/2015' = '5/31/2015',  
 '5/31/2016' = '5/31/2016',  
 '5/31/2017' = '5/31/2017',  
 '5/31/2018' = '5/31/2018',  
 '5/31/2019' = '5/31/2019',  
 '5/31/2020' = '5/31/2020',  
 '5/31/2021' = '5/31/2021',  
 '5/31/2022' = '5/31/2022',  
 '5/31/2023' = '5/31/2023'  
 )  
  
print(NationalHomeSales)

## # A tibble: 1 × 15  
## RegionName State CountyName Metro `5/31/2013` `5/31/2014` `5/31/2015`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Houston TX Harris County Houston-Th… 264062. 302672. 344415.  
## # ℹ 8 more variables: `5/31/2016` <dbl>, `5/31/2017` <dbl>, `5/31/2018` <dbl>,  
## # `5/31/2019` <dbl>, `5/31/2020` <dbl>, `5/31/2021` <dbl>, `5/31/2022` <dbl>,  
## # `5/31/2023` <dbl>

# Filter NY  
NationalHomeSales <- sales\_data %>%  
 filter(  
 RegionName %in% selected\_cities,  
 State == "NY" # Filter for the state of New York  
 ) %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 '5/31/2013' = '5/31/2013',  
 '5/31/2014' = '5/31/2014',  
 '5/31/2015' = '5/31/2015',  
 '5/31/2016' = '5/31/2016',  
 '5/31/2017' = '5/31/2017',  
 '5/31/2018' = '5/31/2018',  
 '5/31/2019' = '5/31/2019',  
 '5/31/2020' = '5/31/2020',  
 '5/31/2021' = '5/31/2021',  
 '5/31/2022' = '5/31/2022',  
 '5/31/2023' = '5/31/2023'  
 )  
  
print(NationalHomeSales)

## # A tibble: 1 × 15  
## RegionName State CountyName Metro `5/31/2013` `5/31/2014` `5/31/2015`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 New York NY Queens County New York-N… 723160. 828522. 917431.  
## # ℹ 8 more variables: `5/31/2016` <dbl>, `5/31/2017` <dbl>, `5/31/2018` <dbl>,  
## # `5/31/2019` <dbl>, `5/31/2020` <dbl>, `5/31/2021` <dbl>, `5/31/2022` <dbl>,  
## # `5/31/2023` <dbl>

selected\_cities <- c("Chicago", "Denver", "Houston", "New York")  
  
# Filter  
NationalHomeSales <- sales\_data %>%  
 filter(  
 RegionName %in% selected\_cities,  
 CountyName %in% c("Cook County", "Denver County", "Harris County", "Queens County") # Adjust county names as needed  
 ) %>%  
 select(  
 RegionName,  
 State,  
 CountyName,  
 Metro,  
 '5/31/2013' = '5/31/2013',  
 '5/31/2014' = '5/31/2014',  
 '5/31/2015' = '5/31/2015',  
 '5/31/2016' = '5/31/2016',  
 '5/31/2017' = '5/31/2017',  
 '5/31/2018' = '5/31/2018',  
 '5/31/2019' = '5/31/2019',  
 '5/31/2020' = '5/31/2020',  
 '5/31/2021' = '5/31/2021',  
 '5/31/2022' = '5/31/2022',  
 '5/31/2023' = '5/31/2023'  
 )  
  
print(NationalHomeSales)

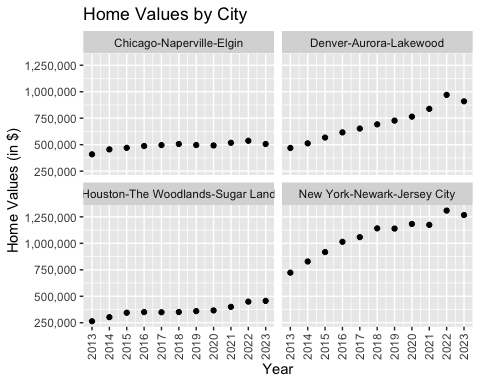
## # A tibble: 4 × 15  
## RegionName State CountyName Metro `5/31/2013` `5/31/2014` `5/31/2015`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 New York NY Queens County New York-N… 723160. 828522. 917431.  
## 2 Houston TX Harris County Houston-Th… 264062. 302672. 344415.  
## 3 Chicago IL Cook County Chicago-Na… 409180. 455543. 470035.  
## 4 Denver CO Denver County Denver-Aur… 468904. 512596. 567297.  
## # ℹ 8 more variables: `5/31/2016` <dbl>, `5/31/2017` <dbl>, `5/31/2018` <dbl>,  
## # `5/31/2019` <dbl>, `5/31/2020` <dbl>, `5/31/2021` <dbl>, `5/31/2022` <dbl>,  
## # `5/31/2023` <dbl>

NationalHomeSales <- NationalHomeSales %>%  
 rename(  
 '2013' = '5/31/2013',  
 '2014' = '5/31/2014',  
 '2015' = '5/31/2015',  
 '2016' = '5/31/2016',  
 '2017' = '5/31/2017',  
 '2018' = '5/31/2018',  
 '2019' = '5/31/2019',  
 '2020' = '5/31/2020',  
 '2021' = '5/31/2021',  
 '2022' = '5/31/2022',  
 '2023' = '5/31/2023'  
 )  
  
# Tidy  
NationalHomeSales\_tidy <- NationalHomeSales %>%  
 gather(key = 'YR', value = 'ZHVI', -RegionName, -State, -CountyName, -Metro) %>%  
 mutate(  
 YR = as.numeric(str\_extract(YR, "\\d{4}")) # Extract only the four-digit year  
 ) %>%  
 filter(!is.na(YR) & !is.na(ZHVI)) # Filter out rows with missing ZHVI values  
  
print(NationalHomeSales\_tidy)

## # A tibble: 44 × 6  
## RegionName State CountyName Metro YR ZHVI  
## <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 New York NY Queens County New York-Newark-Jersey City 2013 723160.  
## 2 Houston TX Harris County Houston-The Woodlands-Sugar Land 2013 264062.  
## 3 Chicago IL Cook County Chicago-Naperville-Elgin 2013 409180.  
## 4 Denver CO Denver County Denver-Aurora-Lakewood 2013 468904.  
## 5 New York NY Queens County New York-Newark-Jersey City 2014 828522.  
## 6 Houston TX Harris County Houston-The Woodlands-Sugar Land 2014 302672.  
## 7 Chicago IL Cook County Chicago-Naperville-Elgin 2014 455543.  
## 8 Denver CO Denver County Denver-Aurora-Lakewood 2014 512596.  
## 9 New York NY Queens County New York-Newark-Jersey City 2015 917431.  
## 10 Houston TX Harris County Houston-The Woodlands-Sugar Land 2015 344415.  
## # ℹ 34 more rows

# scatter  
scatter\_plot <- ggplot(NationalHomeSales\_tidy, aes(x = YR, y = ZHVI)) +  
 geom\_point() +  
 facet\_wrap(~Metro) + # Facet by Metro  
 labs(  
 title = "Home Values by City",  
 x = "Year",  
 y = "Home Values (in $)"  
 ) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1)) +  
 scale\_x\_continuous(  
 breaks = seq(min(NationalHomeSales\_tidy$YR), max(NationalHomeSales\_tidy$YR), by = 1),  
 labels = function(x) format(x, scientific = FALSE, big.mark = "", decimal.mark = "")  
 ) +  
 scale\_y\_continuous(labels = scales::comma)  
  
print(scatter\_plot)

## Warning in prettyNum(.Internal(format(x, trim, digits, nsmall, width, 3L, :  
## 'big.mark' and 'decimal.mark' are both '', which could be confusing  
  
## Warning in prettyNum(.Internal(format(x, trim, digits, nsmall, width, 3L, :  
## 'big.mark' and 'decimal.mark' are both '', which could be confusing  
  
## Warning in prettyNum(.Internal(format(x, trim, digits, nsmall, width, 3L, :  
## 'big.mark' and 'decimal.mark' are both '', which could be confusing  
  
## Warning in prettyNum(.Internal(format(x, trim, digits, nsmall, width, 3L, :  
## 'big.mark' and 'decimal.mark' are both '', which could be confusing



Based on your analysis, which city’s housing is most affordable? Least affordable?  
Based on the analysis, Houston appears to be the most affordable, while New York is the least affordable City.

Which cities saw the largest change in prices over the past 5 years? Which city has remained more consistent (i.e., no huge swings up or down in home values)?  
Denver and New York had the largest change in price over the past 5 years, while Chicago and Houston remained more consistent.

Which cities saw a decline in value during 2023 and which cities remained consistent?  
Chicago, Denver, and New York all a slight decline in 2023, while Houston remained consistent.

### Future Home Values

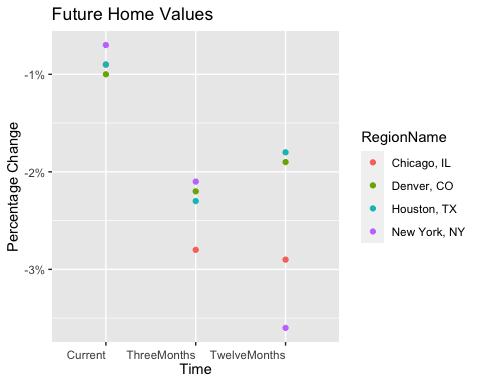
Projections <- read\_csv("Projections.csv")

## Rows: 895 Columns: 9  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): RegionName, RegionType, StateName  
## dbl (5): RegionID, SizeRank, 2023-11-30, 2024-01-31, 2024-10-31  
## date (1): BaseDate  
##   
## 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.

# Specify selected cities  
selected\_cities <- c("Chicago, IL", "Denver, CO", "Houston, TX", "New York, NY")  
  
# Filter  
FutureHomeValues <- Projections %>%  
 filter(  
 RegionName %in% selected\_cities  
 ) %>%  
 select(  
 RegionName,  
 `2023-11-30` = `2023-11-30`,  
 `2024-01-31` = `2024-01-31`,  
 `2024-10-31` = `2024-10-31`  
 ) %>%  
 rename(  
 Current = `2023-11-30`,  
 ThreeMonths = `2024-01-31`,  
 TwelveMonths = `2024-10-31`  
 ) %>%  
 gather(key = 'Time', value = 'PercentageChange', -RegionName)  
  
print(FutureHomeValues)

## # A tibble: 12 × 3  
## RegionName Time PercentageChange  
## <chr> <chr> <dbl>  
## 1 New York, NY Current -0.7  
## 2 Chicago, IL Current -0.9  
## 3 Houston, TX Current -0.9  
## 4 Denver, CO Current -1   
## 5 New York, NY ThreeMonths -2.1  
## 6 Chicago, IL ThreeMonths -2.8  
## 7 Houston, TX ThreeMonths -2.3  
## 8 Denver, CO ThreeMonths -2.2  
## 9 New York, NY TwelveMonths -3.6  
## 10 Chicago, IL TwelveMonths -2.9  
## 11 Houston, TX TwelveMonths -1.8  
## 12 Denver, CO TwelveMonths -1.9

# scatter  
scatter\_plot <- ggplot(FutureHomeValues, aes(x = Time, y = PercentageChange, color = RegionName)) +  
 geom\_point() +  
 labs(  
 title = "Future Home Values",  
 x = "Time",  
 y = "Percentage Change"  
 ) +  
 theme(axis.text.x = element\_text(angle = 0, vjust = 0.5, hjust = 1),  
 axis.text.y = element\_text(angle = 0)) +  
 scale\_x\_discrete(name = "Time") +  
 scale\_y\_continuous(name = "Percentage Change", labels = scales::percent\_format(scale = 1))  
  
print(scatter\_plot)



Which is the only city that is projected to have a decrease in home values in the next 3 months?  
All four cities decrease in the next 3 months, but Chicago has the largest decrease.

If you are only concerned about the largest home value increase (by percentage) in the next 12 months, which city would you choose to relocate to?  
Houston.

### 