Home and Rental Price Evaluation

Create the R Markdown Document

1.) In RStudio, select *File -> New File -> Text File*. This will create a blank text file in the same area that scripts were created in previous assignments (upper left panel). Save this file to your project as **Mod7FinalAssignAnswer.rmd**

2.) Create a Header 1 with the title: **MIS503 – Final Project**

3.) Create a Header 2 with the title: **Last Name, First Name** (replace with your name)

4.) Create a Header 3 with the title: **Zillow Home Value Index Analysis**

5.) For this assignment, you will need to download the following files from Canvas which should be saved in the same folder as your project:

   a. SingleFamilyResidenceSales.csv
   b. SingleFamilyResidenceRental.csv

The goal of this assignment is for you to apply all the skills you have learned over the past 6 weeks to answer research questions posed through a case study on residential home sales and rentals in North Carolina and other parts of the country. The files you downloaded are from Zillow who publishes the Zillow Home Value Index (ZHVI) that includes list prices, sales prices and other statistics related to home sales and rentals throughout the country (https://www.zillow.com/research/data/).

Exercise 1: Trends in home values in Wake County, North Carolina

1.) Create a Header 3 in R Markdown titled: **Wake County Home Sales**

2.) You are considering a move to the Raleigh area and are interested in understanding trends in home values. Run the analysis in the instructions below and come back to this section of the R Markdown document and address the following questions:

   a. What have been the overall trends in Wake County Home Values?
   b. There were dips in home values in the past 10 years. What years did these occur?
   c. Based on the analysis, where would be the least expensive area to purchase home? Most expensive area?
   d. What has happened to the overall property values in Apex and Cary in 2023?

3.) For this analysis, you will need to import the **SingleFamilyResidenceSales.csv** file from Canvas. Remember to include the correct code in the R Markdown document. Also, remember that you must also load the appropriate packages that will be used for this analysis.

4.) Create a new *tibble* called **WakeCountySales** and use *dplyr* to only include the cities in Wake County. This will require you to filter by State and CountyName. In addition to filtering by the above, you will want to only include the following columns of data in your new tibble (we are taking a snapshot of prices during the month of May each year):

   a. RegionName
   b. State
   c. CountyName
   d. Metro
   e. 5/31/2013
   f. 5/31/2014
   g. 5/31/2015
   h. 5/31/2016
   i. 5/31/2017
   j. 5/31/2018
   k. 5/31/2019
   l. 5/31/2020
   m. 5/31/2021
   n. 5/31/2022
   o. 5/31/2023

5.) Once you have created **WakeCountySales**, you will need to rename the fields above (labeled e. through o.) to only include the year (e.g., 5/31/2013 should be 2013 in the tibble).

6.) If you take a closer look at the data, you will notice it is not tidy. Use *tidyr* to tidy the data so the year is no longer a column but is a row. You will need to use the names_to ='YR' and values_to ='ZHVI'.

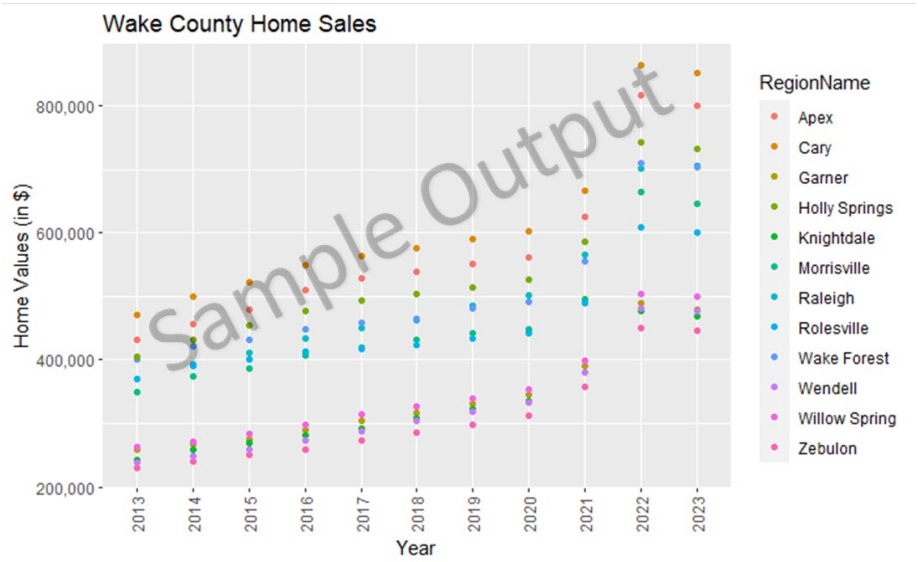
7.) You should now have a tidy set of data that can be analyzed using *ggplot2*.

8.) For the first analysis, we will be creating a scatter plot. You will want to have year on the x-axis and home value (ZHVI) on the y-axis. Also, include the RegionName as a color on your plot. Finally, you will need to include code to give the plot a title, an x-axis name and a y-axis name.

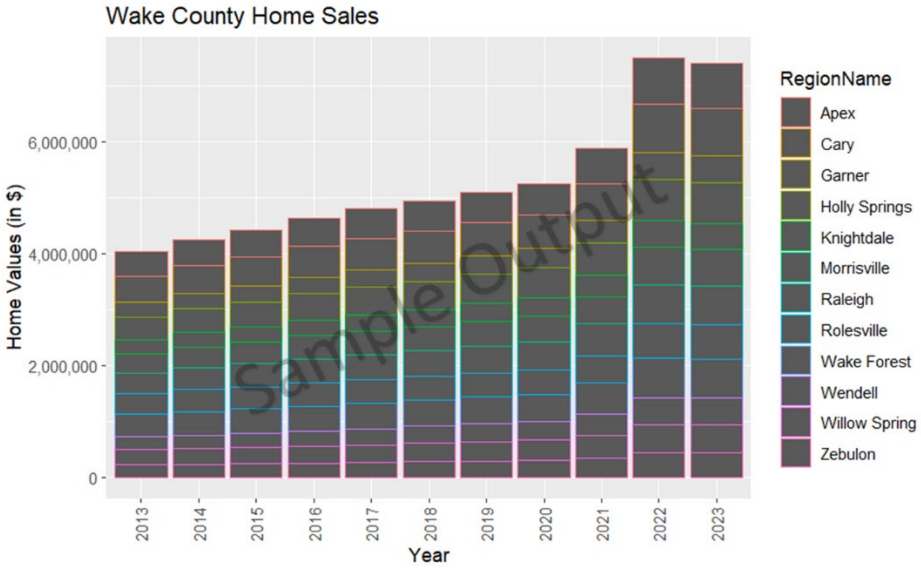
9.) The years will run together and the y-axis values will not show up correctly. To fix this, include the following two lines of code in your *ggplot* function. The first will change the x-axis to display vertically. The second line of code will replace the label for the y-axis and changes the values on the axis to include commas.

   a. theme(axis.text.x = element_text(angle = 90, vjust=0.5))
   b. scale_y_continuous(name="Home Values (in $)", labels = scales::comma)

10.) This analysis should result in the following plot that can be used to evaluate the previous questions:



11.) Before moving to the next exercise, you should also create a stacked bar graph. You can use much of the same code as you did in the scatterplot but will need to use *geom_col()* to create the below chart.



12.) The last step in this exercise is to go back to the beginning to answer the questions posed.

Exercise 2: North Carolina Rental Market

1.) Create a Header 3 in R Markdown titled: **NC Rental Market**

2.) You are considering relocating to a new city and want to rent. You are looking at some of the larger cities within the state including Asheville, Charlotte, Durham, Fayetteville, Raleigh, and Wilmington. Run the analysis in the instructions below and come back to this section of the R Markdown document and address the following questions:

   a. What has been the overall trend in the rental market around the state? Are there any cities that have not followed this trend?
   b. Where is the most expensive city to rent in? Least expensive?
   c. You are trying to decide between Wilmington and Asheville. Which market has the lowest rent?

3.) For this analysis, you will need to import the SingleFamilyResidenceRental.csv file from Canvas. Remember to include the correct code in the R Markdown document.

4.) Create a new tibble called **Rentals.** Use *dplyr* to only include the cities listed in instruction 2 above. Remember that there may be cities with the same name in different states. In addition to filtering by the above, you will want to include the following columns of data in your new tibble (we are taking a snapshot of rent during the month of January each year):

   a. RegionName (this is the field that contains the city)
   b. State
   c. 1/31/2015
   d. 1/31/2016
   e. 1/31/2017
   f. 1/31/2018
   g. 1/31/2019
   h. 1/31/2020
   i. 1/31/2021
   j. 1/31/2022
   k. 1/31/2023

5.) Once you have created **Rentals**, you will need to rename the fields above (labeled c. through k.) to only include the year (e.g., 1/31/2015 should be 2015 in the tibble).

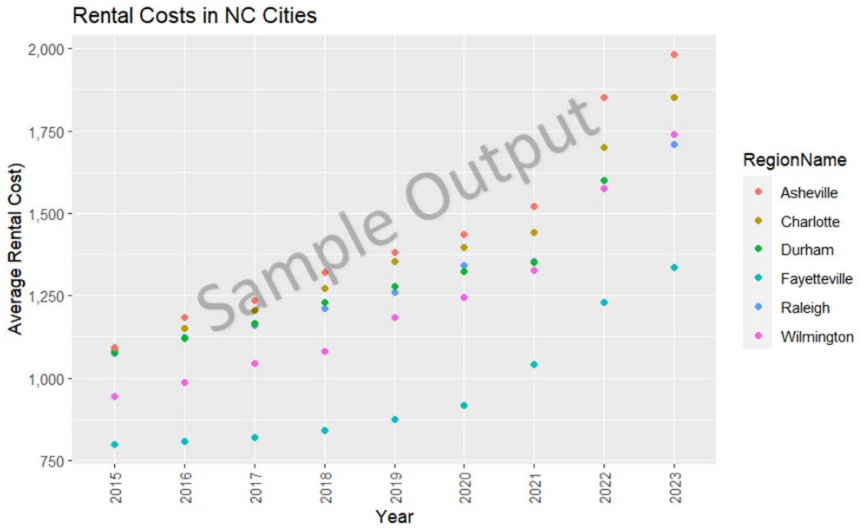
6.) If you take a closer look at the data, you will notice it is not tidy. Use *tidyr* to tidy the data so the year is no longer a column but is a row. You will need to use the names_to ='YR' and values_to ='ZHVI'.

7.) You should now have a tidy set of data that can be analyzed using *ggplot*.

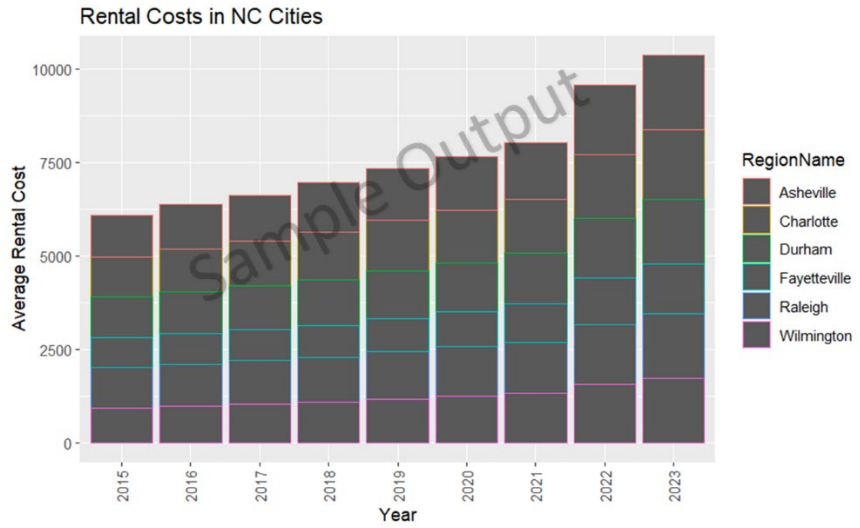
8.) For the first analysis, we will be creating a scatter plot similar to the previous exercise. You will want to have year on the x-axis and home value (ZHVI) on the y-axis. Also, include the city as a color on your plot. Finally, you will need to include code to give the plot a title, an x-axis name and a y-axis name.

9.) Like the previous plots, the years will run together on the x-axis as well as the issue with the y-axis values. Use similar code from the previous plots to correct both the x-axis and include commas on the y-axis.

10.) This analysis should result in the following plot that can be used to evaluate the previous questions:



11.) Before moving to the next exercise, you should also create a stacked bar graph.



12.) The last step in this exercise is to go back to the beginning to answer the questions posed.

Exercise 3: Home Values in Select North Carolina Markets

1.) Create a Header 3 in R Markdown titled: **Home Values in Select Markets**

2.) You have made the choice that you want to focus on 4 metro areas (Asheville, Charlotte-Concord-Gastonia, Raleigh-Cary and Wilmington) and instead of renting, you would like to purchase a home. Run the analysis in the instructions below and come back to this section of the R Markdown document and address the following questions (you may need to do some research on reading a violin plot https://blog.modeanalytics.com/violin-plot-examples/):

   a. According to the results, which market has the lowest median price (represented as horizontal bar in box plot)?
   b. 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?
   c. 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)?

3.) For this analysis, we will be using the SingleFamilyResidenceSales tibble again.

4.) Create a new *tibble* called **NCHomeSales** and use *dplyr* to only include the metro areas you are evaluating listed above in instruction 2 (hint: this will require you to filter by State and Metro). In addition to filtering by the above, you will want to only include the following columns of data in your new tibble (we are taking a snapshot of prices during the month of May each year):

   a. RegionName
   b. State
   c. Metro
   d. 5/31/2013
   e. 5/31/2014
   f. 5/31/2015
   g. 5/31/2016
   h. 5/31/2017
   i. 5/31/2018
   j. 5/31/2019
   k. 5/31/2020
   l. 5/31/2021
   m. 5/31/2022
   n. 5/31/2023

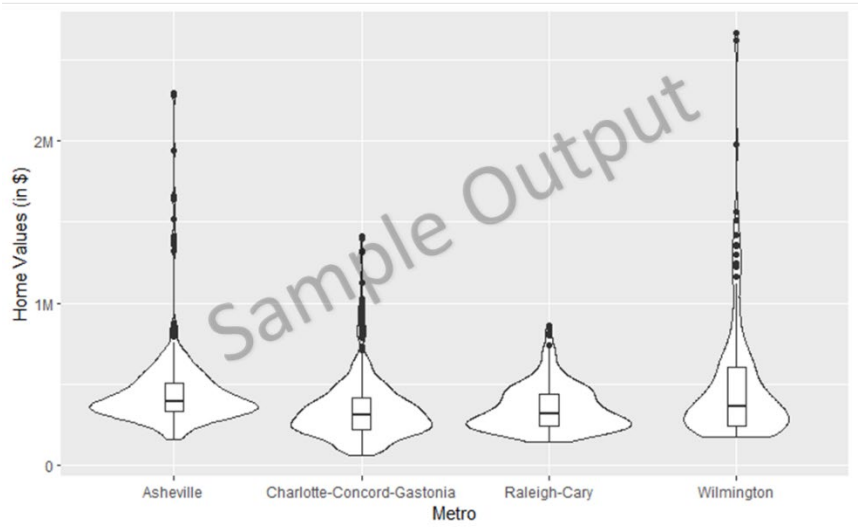
5.) Once you have created **NCHomeSales**, you will need to rename the fields above (labeled d. through n) to only include the year (e.g., 5/31/2013 should be 2013 in the tibble).

6.) If you take a closer look at the data, you will notice it is not tidy. Use *tidyr* to tidy the data so the year is no longer a column but is a row. You will need to use the names_to ='YR' and values_to ='ZHVI'. Finally, to run our analysis you will need to group the **NCHomeSales** by Metro to get all of the data in the correct order for our plots.

7.) You should now have a tidy set of data that can be analyzed using ggplot.

8.) For this analysis, we will be creating a violin and boxplot that appears on the same plot. You will want to have Metro on the x-axis and home value (ZHVI) on the y-axis. To run this analysis, you will use the geom_violin( ) and geom_boxplot(width=0.1) commands.

9.) While creating the plot, be sure to label the y-axis correctly using millions to match the y-axis values below (see https://rpubs.com/techanswers88/remove-scientific-notation-in-ggplot for hints).



10.)The last step in this exercise is to go back to the beginning to answer the questions posed.

Exercise 4: Relocation to Another City

1.) Create a Header 3 in R Markdown titled: **Relocation Home Value Comparison**

2.) You have been given a new opportunity to relocate for a new position within your company. They have given you the option of 4 different cities in the country (Chicago, Denver, Houston or New York). Run the analysis in the instructions below and come back to this section of the R Markdown document and address the following questions:

   a. Based on your analysis, which city's housing is most affordable? Least affordable?
   b. 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)?
   c. Which cities saw a decline in value during 2023 and which cities remained consistent?

3.) For this analysis, you will need to use the **SingleFamilyResidenceSales** tibble you imported earlier in this assignment.

4.) Create a new tibble called **NationalHomeSales** and use *dplyr* to only include the cities listed above in instruction 2 of Exercise 4. This will require you to filter by both **RegionName** using the %in% code in the command for the city and **CountyName** using the %in% code in the command (do a search to find out which counties the cities are in). The county name will make sure that you get the right city as there are multiple cities with the same names. In addition to filtering by the above, you will want to only include the following columns of data in your new tibble (we are taking a snapshot of prices during the month of May each year):

   a. RegionName
   b. State
   c. CountyName
   d. Metro
   e. 5/31/2013
   f. 5/31/2014
   g. 5/31/2015
   h. 5/31/2016
   i. 5/31/2017
   j. 5/31/2018
   k. 5/31/2019
   l. 5/31/2020
   m. 5/31/2021
   n. 5/31/2022
   o. 5/31/2023

5.) Once you have created **NationalHomeSales**, you will need to rename the fields above (labeled e. through o.) to only include the year (e.g., 5/31/2013 should be 2013 in the tibble).

6.) If you take a closer look at the data, you will notice it is not tidy. Use *tidyr* to tidy the data so the year is no longer a column but is a row. You will need to use the names_to ='YR' and values_to ='ZHVI'.

7.) You should now have a tidy set of data that can be analyzed using *ggplot*.

8.) For the first analysis, we will be creating a scatter plot. You will want to have year on the x-axis and home value (ZHVI) on the y-axis. However, instead of setting the city to be different colors, we would like to

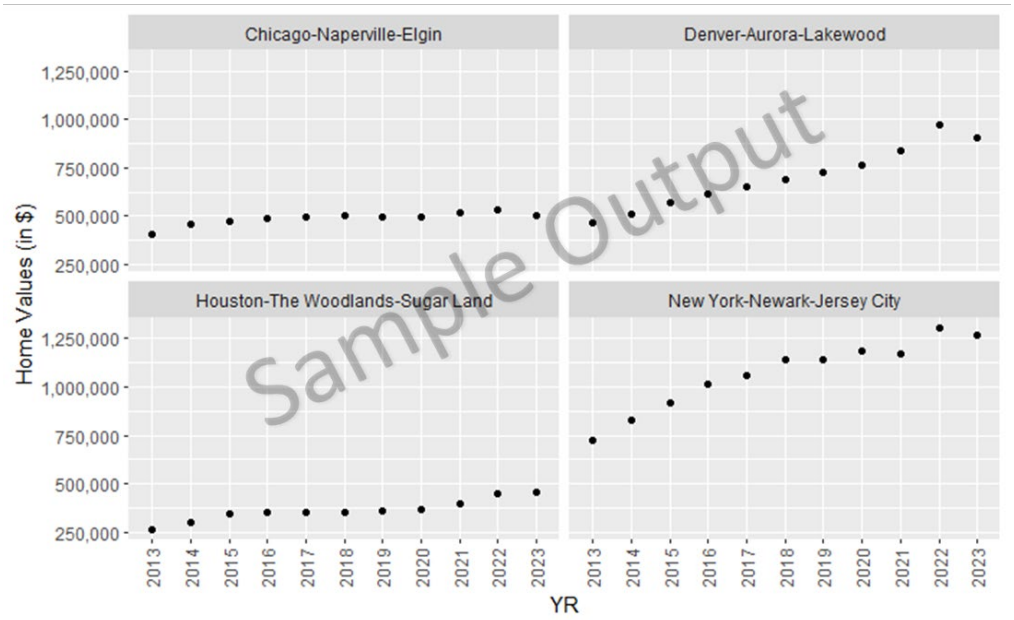
create a facet of the information so each city appears in a separate scatterplot. To do this you will need to add the following code to your function:

facet_wrap(~Metro)

9.) Finally, you will need to include code to give the plot a title, an x-axis name and a y-axis name.

10.) Follow the instructions from the previous exercises to make sure the x-axis labels are vertical and the y-axis values have commas for the numbers.

11.) This analysis should result in the following plot that can be used to evaluate the previous questions:



11.) The last step in this exercise is to go back to the beginning to answer the questions posed.

12.) Finally, knit your R Markdown to Word and upload the document to Canvas.

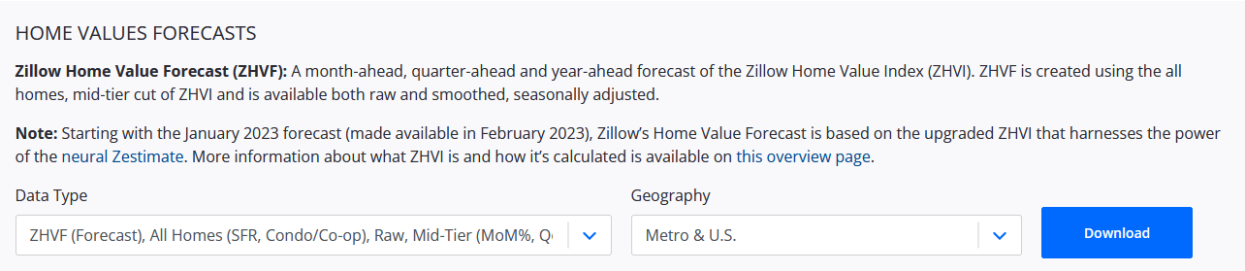
Exercise 5: Future Home Values

1.) Create a Header 3 in R Markdown titled: **Future Home Values**

2.) As part of your decision-making process in determining the city you want to relocate to, you are also interested in the future home values within 4 different cities in the country (Chicago, Denver, Houston or New York). Run the analysis in the instructions below and come back to this section of the R Markdown document and address the following questions:

a. Which is the only city that is projected to have a decrease in home values in the next 3 months?
b. 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?

3.) For this analysis, you will need go out to the Zillow Housing Data website (https://www.zillow.com/research/data/) and download the "Home Value Forecasts" using the following Data Type and Geography. Save the file as **Projections** (be sure to save this file to same folder as your project). Below is the selection criteria that should be used to get the correct dataset.



4.) Open the import window (see below) to see the file. Notice the part of the data that is circled. These are the dates that will be used to see projected future increases in price. These will change depending on the time you download the file. Copy the code to import the Projections.csv dataset into your chunk of code to create a new Projections tibble. You will be using this tibble to create the new tibble below.

Import Text Data

File/URL:

~/MIS503/AssignmentAnswers/Module7/RStudio_Assignment1/Projections2.csv

Data Preview:

RegionID (double)	SizeRank (double)	RegionName (character)	RegionType (character)	StateName (character)	BaseDate (double)	2023-08-31 (double)	2023-10-31 (double)	2024-07-31 (double)
102001	0	United States	country	NA	2023-07-31	0.4	0.4	6.5
394913	1	New York, NY	msa	NY	2023-07-31	0.2	0.1	3.7
753899	2	Los Angeles, CA	msa	CA	2023-07-31	0.2	-0.4	5.3
394463	3	Chicago, IL	msa	IL	2023-07-31	0.2	-0.2	3.8
394514	4	Dallas, TX	msa	TX	2023-07-31	0.1	-0.5	5.2

5.) Create a new tibble called **FutureHomeValues** and use *dplyr* to only include the 4 cities we are interested in evaluating the future home values. This will require you to filter **RegionName** using the %in% code in the command (you will also need to include the state with the cities, e.g., Houston, TX). In addition to filtering by the above, you will want to only include the following columns of data in your new tibble (note the dates may change depending on when you download the file, just make sure they match what is in file you are importing, i.e. the circled dates in the image above). Here are the columns for the dataset I am using:

a. RegionName
b. 2023-08-31
c. 2023-10-31
d. 2024-07-31

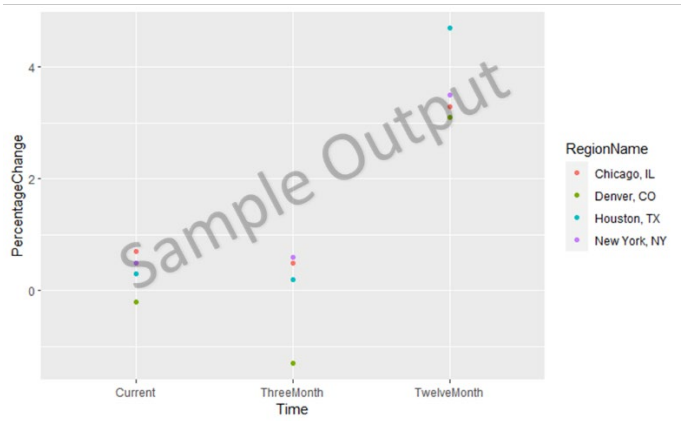
6.) Once you have created **FutureHomeValues**, you will need to rename the fields below (again, the Current Name may be different depending on when you download the file so change accordingly):

Current Name		New Name
2023-08-31	->	Current
2023-10-31	->	ThreeMonths
2024-07-31	->	TwelveMonths

7.) If you take a closer look at the data, you will notice it is not tidy. Use *tidyr* to tidy the data so the fields you changed the names of (see above) are no longer a column but a row (use names_to ='Time' and values_to ='Percentage Change').

8.) You should now have a tidy set of data that can be analyzed using *ggplot*.

9.) Similar to exercises 1 and 2, we will be creating a scatter plot. You will want to have Time on the x-axis and PercentageChange on the y-axis. Also, include the city as a color on your plot. Here is an example of what the plot might look like (it will change depending on the dates you are using).



10.)The last step in this exercise is to go back to the beginning to answer the questions posed.

Finally, knit your R Markdown to Word and upload the document to Canvas.

Before submitting, open the word document and see how the visuals appear in the output. Make adjustments accordingly so that the output is easily readable.