GBA 464 - Assignment #2
Due Thursday, February 11 @ 10:00 AM

Please upload your completed assignment to BlackBoard. Please include a (cleaned-up) copy of your R code (my source code is approx. 100 lines) along with the exhibits described in the assignment instructions below.

Zillow Research: Median Home Value Per Sq Ft ($) by City
Data source: https://www.zillow.com/research/data/
Note: This is an archived version of the data.

Raw data description:

Filename: zilllow.csv
Median Home Value per sq ft. by Region by Date

Field	Description
RegionID	Unique ID for region
1996.04 – 2019.08	Date columns (year.month)

Filename: regions.csv
Region Information

Field	Description
RegionID	Unique ID for region
RegionName	Name of region
State	2 character state abbreviation
Metro	Name of metro area
CountyName	Name of county
SizeRank	Ordinal rank of region by size (1 = largest)

Part 1: Load and clean data.

Load the .csv data files into R. Transform the data into the "tidy" data set. Your data should look like the table below.

First 15 rows and first 9 columns of tidy data set: sorted by Date and SizeRank

	RegionID	RegionName	State	Metro	CountyName	SizeRank	Year	Month	MedHouse
1	6181	New York	NY	New York-Newark-Jersey City	Queens County	1	1996	04	130
2	12447	Los Angeles	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	2	1996	04	110
3	39051	Houston	TX	Houston-The Woodlands-Sugar Land	Harris County	3	1996	04	50
4	17426	Chicago	IL	Chicago-Naperville-Elgin	Cook County	4	1996	04	88
5	6915	San Antonio	TX	San Antonio-New Braunfels	Bexar County	5	1996	04	54
6	13271	Philadelphia	PA	Philadelphia-Camden-Wilmington	Philadelphia County	6	1996	04	38
7	40326	Phoenix	AZ	Phoenix-Mesa-Scottsdale	Maricopa County	7	1996	04	56
8	18959	Las Vegas	NV	Las Vegas-Henderson-Paradise	Clark County	8	1996	04	76
9	54296	San Diego	CA	San Diego-Carlsbad	San Diego County	9	1996	04	111
10	38128	Dallas	TX	Dallas-Fort Worth-Arlington	Dallas County	10	1996	04	54
11	10221	Austin	TX	Austin-Round Rock	Travis County	11	1996	04	99
12	33839	San Jose	CA	San Jose-Sunnyvale-Santa Clara	Santa Clara County	12	1996	04	147
13	25290	Jacksonville	FL	Jacksonville	Duval County	13	1996	04	46
14	32149	Indianapolis	IN	Indianapolis-Carmel-Anderson	Marion County	14	1996	04	79
15	20330	San Francisco	CA	San Francisco-Oakland-Hayward	San Francisco County	15	1996	04	199

Submit a list of instructions (in words, not R code – similar to those we looked at in class) for how you transformed your data from the raw file to the tidy data set.

Part 2: Tabulate data.

Create a summary data file by County and State which contains the following statistics for each county's median house value over the time period of the data: median, std deviation, max and min. Create an ordinal ranking for the volatility of housing values by county with rank 1 assigned to the location whose county-level median house value had the highest standard deviation over the time period of the data. In the same file, include the same statistics for median house values for each county's state and a state volatility ranking (again, with 1 being assigned to the state whose median house value had the highest standard deviation over the time period of the data). The following page has the first 10 rows of the summary data with three different sorts.

Summary table sorted by State, County

County.State	medCounty	sdCounty	maxCounty	minCounty	RanksdCounty	medState	sdState	maxState	minState	RanksdState
Anchorage Borough.AK	172.0	41.768565	195.0	71.0	121	132	27.22735	164	71	18
Fairbanks North Star Borough.AK	202.0	42.879665	218.0	85.5	113	132	27.22735	164	71	18
Juneau Borough.AK	204.0	51.121578	258.0	96.0	67	132	27.22735	164	71	18
Kenai Peninsula Borough.AK	135.0	27.509514	164.5	71.5	335	132	27.22735	164	71	18
Ketchikan Gateway Borough.AK	148.0	28.091843	189.0	85.0	320	132	27.22735	164	71	18
Matanuska Susitna Borough.AK	96.5	19.504116	114.0	46.0	597	132	27.22735	164	71	18
Autauga County.AL	82.0	10.694759	104.0	60.0	1219	67	10.27299	84	43	46
Baldwin County.AL	78.0	16.671687	105.0	37.5	737	67	10.27299	84	43	46
Bibb County.AL	58.0	7.573285	78.0	47.0	1530	67	10.27299	84	43	46
Blount County.AL	70.5	11.421737	89.5	34.0	1153	67	10.27299	84	43	46

Summary table sorted by County Rank (RanksdCounty)

County.State	medCounty	sdCounty	maxCounty	minCounty	RanksdCounty	medState	sdState	maxState	minState	RanksdState
San Francisco County.CA	560.0	235.2447	1065.0	199.0	1	202	67.86217	314.0	98.0	3
San Mateo County.CA	485.0	202.3455	991.5	196.0	2	202	67.86217	314.0	98.0	3
Santa Clara County.CA	439.5	182.3114	935.0	179.0	3	202	67.86217	314.0	98.0	3
Nantucket County.MA	575.0	177.3431	826.0	166.0	4	173	40.23260	223.0	83.5	5
Pitkin County.CO	627.0	176.8990	911.5	265.5	5	144	38.72741	242.0	79.0	8
Marin County.CA	518.5	173.7169	835.0	202.5	6	202	67.86217	314.0	98.0	3
San Miguel County.CO	457.0	132.7619	629.0	221.0	7	144	38.72741	242.0	79.0	8
District of Columbia.DC	334.0	129.6077	520.0	101.0	8	334	129.60775	520.0	101.0	1
Maui County.HI	323.0	122.1398	542.0	150.5	9	295	100.53890	448.5	139.0	2
Alameda County.CA	301.5	121.3668	581.0	133.0	10	202	67.86217	314.0	98.0	3

Summary table sorted by State Rank, then County Rank (RanksdState)

County.State	medCounty	sdCounty	maxCounty	minCounty	RanksdCounty	medState	sdState	maxState	minState	RanksdState
District of Columbia.DC	334.0	129.60775	520.0	101.0	8	334	129.60775	520.0	101	1
Maui County.HI	323.0	122.13983	542.0	150.5	9	295	100.53890	448.5	139	2
Honolulu County.HI	359.0	115.78915	518.0	157.5	14	295	100.53890	448.5	139	2
Kauai County.HI	313.0	108.67447	464.0	130.5	19	295	100.53890	448.5	139	2
Hawaii County.HI	199.0	70.22196	327.0	98.0	33	295	100.53890	448.5	139	2
San Francisco County.CA	560.0	235.24473	1065.0	199.0	1	202	67.86217	314.0	98	3
San Mateo County.CA	485.0	202.34553	991.5	196.0	2	202	67.86217	314.0	98	3
Santa Clara County.CA	439.5	182.31145	935.0	179.0	3	202	67.86217	314.0	98	3
Marin County.CA	518.5	173.71687	835.0	202.5	6	202	67.86217	314.0	98	3
Alameda County.CA	301.5	121.36684	581.0	133.0	10	202	67.86217	314.0	98	3

Using this data, create a frequency table of County Rank quartiles by State Rank quartiles.

1. Which counties are among the (top) 25% most volatile counties but located in one of the (bottom) 25% least volatile states? (answer shown below) . **Submit R code that generates the results shown below.**

County.State	sdCounty	RanksdCounty	sdState	RanksdState
Glynn County.GA	49.13492	78	12.641920	39
Pickens County.GA	36.62003	183	12.641920	39
Union County.GA	26.38875	371	12.641920	39
Dekalb County.GA	27.40055	338	12.641920	39
Fannin County.GA	24.40402	413	12.641920	39
Greene County.GA	40.15804	139	12.641920	39
Franklin County.KY	30.30310	278	11.512572	40
Beaufort County.SC	24.89666	398	11.288980	41
Georgetown County.SC	24.68294	406	11.288980	41
Charleston County.SC	37.03991	178	11.288980	41
Jasper County.SC	27.08212	350	11.288980	41
Cheatham County.TN	24.64947	408	13.289865	38
Davidson County.TN	35.75408	193	13.289865	38
Rutherford County.TN	24.79528	402	13.289865	38
Williamson County.TN	32.58761	239	13.289865	38
Meigs County.TN	24.21444	420	13.289865	38
Berkeley County.WV	24.78535	403	9.843338	47
Monongalia County.WV	23.10754	455	9.843338	47

2. Which states are represented on the list of the top 50 most volatile counties? **Submit R code and output that answers this question.**

Part 3: Data Visualization

Subset the data to analyze the top 10 region IDs by size. Create a time series visualization of the median house value over the time period of the data. My example is shown below. I split the data into two groups: regions where housing value recovered after the real estate bubble burst in the 2006/2007 (ex: New York) and regions which did not recover (ex: Chicago). For regions where the peak housing value did not recover, I labeled the date of the peak. Note: I think there is an argument for different ways to group this data such as regions which did not experience the high/low crash at all (ex: Dallas).

Submit your R code and a screenshot of your visualization.

Your finished visualization does not need to look identical to mine (it could be even better!). Use your creativity and judgement to design something that you think is interesting and compelling. Do your best and include whatever you are able to.

I used ggplot2 package to make my graph. There is a lot of good documentation and help for this package. A nice description of ggplot2 syntax and options is here:
http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/