HW_2_final

GBA 464 Spring A 2021 HW 2

```
library(dplyr)
library(reshape2)

setwd("C:/Users/elizabeth.mohr/Dropbox/GBA 464/Spring A 2021/Assignments/HW 2")
zillow <- read.csv("zillow.csv", header = TRUE)
regions <- read.csv("regions.csv", header = TRUE)
```

Part 1: Data Cleaning

Instructions:

Step 1: Merge regions.csv and zillow.csv using RegionID.
Step 2: Melt merged file.
Step 3: Create Year, Month and Year.Month columns (removing the X from the Year.Month column name).
Step 4: Create County.State column.
Step 5: Sort rows and select columns for output.

```
z <- merge(zillow, regions, by.x = "RegionID", by.y = "RegionID")

zm <- melt(z, id = c("RegionID", "RegionName", "State", "Metro", "CountyName", "SizeRank"), vari
able.name = "Year.Month", value.name = "MedHouse")
zm$Year.Month <- substr(zm$Year.Month, 2, 9)
zm$Year <- substr(zm$Year.Month, 1, 4)
zm$Month <- substr(zm$Year.Month, 6, 9)
zm$Year.Month <- as.factor(zm$Year.Month)
zm <- zm[order(zm$Year, zm$Month, zm$SizeRank), ]
zm$County.State <- paste0(zm$CountyName, '.', zm$State)
(zm[1:15, c(1,2,3,4,5,6,9,10,8)])
```

```
##      RegionID    RegionName State                          Metro
## 601      6181      New York    NY      New York-Newark-Jersey City
## 1844    12447   Los Angeles    CA   Los Angeles-Long Beach-Anaheim
## 6881    39051       Houston    TX Houston-The Woodlands-Sugar Land
## 2751    17426       Chicago    IL         Chicago-Naperville-Elgin
## 761      6915   San Antonio    TX        San Antonio-New Braunfels
## 2028    13271  Philadelphia    PA   Philadelphia-Camden-Wilmington
## 7154    40326       Phoenix    AZ          Phoenix-Mesa-Scottsdale
## 3042    18959     Las Vegas    NV     Las Vegas-Henderson-Paradise
## 9782    54296     San Diego    CA               San Diego-Carlsbad
## 6707    38128        Dallas    TX      Dallas-Fort Worth-Arlington
## 1349    10221        Austin    TX                Austin-Round Rock
## 5912    33839      San Jose    CA   San Jose-Sunnyvale-Santa Clara
## 4210    25290  Jacksonville    FL                     Jacksonville
## 5533    32149  Indianapolis    IN     Indianapolis-Carmel-Anderson
## 3319    20330 San Francisco    CA    San Francisco-Oakland-Hayward
##               CountyName SizeRank Year Month MedHouse
## 601         Queens County        1 1996    04      130
## 1844   Los Angeles County        2 1996    04      110
## 6881        Harris County        3 1996    04       50
## 2751          Cook County        4 1996    04       88
## 761          Bexar County        5 1996    04       54
## 2028  Philadelphia County        6 1996    04       38
## 7154      Maricopa County        7 1996    04       56
## 3042         Clark County        8 1996    04       76
## 9782     San Diego County        9 1996    04      111
## 6707        Dallas County       10 1996    04       54
## 1349        Travis County       11 1996    04       99
## 5912   Santa Clara County       12 1996    04      147
## 4210         Duval County       13 1996    04       46
## 5533        Marion County       14 1996    04       79
## 3319 San Francisco County       15 1996    04      199
```

Part 2: Summary Tables

```
## create summary table by county
statsCounty <- as.data.frame(zm %>% dplyr::group_by(County.State, Year.Month) %>%
                                    dplyr::summarize(medHouseCounty.Year.Month = median(MedHous
e, na.rm = TRUE)) %>%
                                    dplyr::group_by(County.State) %>%
                                    dplyr::summarize(sdCounty = sd(medHouseCounty.Year.Month, n
a.rm = TRUE),
                                                     medCounty = median(medHouseCounty.Year.Mon
th, na.rm = TRUE),
                                                     maxCounty = max(medHouseCounty.Year.Month,
na.rm = TRUE),
                                                     minCounty = min(medHouseCounty.Year.Month,
na.rm = TRUE))
                                 )


##add ranking and ranking group
statsCounty$State <- substr(statsCounty$County.State, nchar(statsCounty$County.State) - 1, nchar
(statsCounty$County.State))
statsCounty <- arrange(statsCounty, desc(sdCounty))
statsCounty$RanksdCounty <- 1:nrow(statsCounty)
statsCounty$RankGroupCounty <- cut(statsCounty$RanksdCounty, quantile(statsCounty$RanksdCounty),
labels = c(1,2,3,4), include.lowest = TRUE)

##create summary table by state
statsState <- as.data.frame(zm %>% dplyr::group_by(State, Year.Month) %>%
                                 dplyr::summarize(medHouseState.Year.Month = median(MedHouse, n
a.rm = TRUE)) %>%
                                 dplyr::group_by(State) %>%
                                 dplyr::summarize(sdState = sd(medHouseState.Year.Month, na.rm =
TRUE),
                                                  medState = median(medHouseState.Year.Month, n
a.rm = TRUE),
                                                  maxState = max(medHouseState.Year.Month, na.rm
= TRUE),
                                                  minState = min(medHouseState.Year.Month, na.rm
= TRUE)))


##add ranking and ranking group
statsState <- arrange(statsState, desc(sdState))
statsState$RanksdState <- 1:nrow(statsState)
statsState$RankGroupState <- cut(statsState$RanksdState, quantile(statsState$RanksdState), label
s = c(1,2,3,4), include.lowest = TRUE)


## combine county and state stats
stats <- merge(statsCounty, statsState, by.x = "State", by.y = "State")
```

Part 2: Summary Tables Output

```
##output tables
stats[order(stats$State, stats$County), ][1:10, c("County.State", "medCounty", "sdCounty", "maxC
ounty", "minCounty", "RanksdCounty", "medState", "sdState", "maxState", "minState", "RanksdStat
e")]
```

```
##                       County.State medCounty  sdCounty maxCounty minCounty
## 3              Anchorage Borough.AK     172.0 41.768565     195.0      71.0
## 2  Fairbanks North Star Borough.AK     202.0 42.879665     218.0      85.5
## 1                Juneau Borough.AK     204.0 51.121578     258.0      96.0
## 6       Kenai Peninsula Borough.AK     135.0 27.509514     164.5      71.5
## 4     Ketchikan Gateway Borough.AK     148.0 28.091843     189.0      85.0
## 5     Matanuska Susitna Borough.AK      96.5 19.504116     114.0      46.0
## 26               Autauga County.AL      82.0 10.694759     104.0      60.0
## 7                Baldwin County.AL      78.0 16.671687     105.0      37.5
## 17                  Bibb County.AL      58.0  7.573285      78.0      47.0
## 40                Blount County.AL      70.5 11.421737      89.5      34.0
##    RanksdCounty medState  sdState maxState minState RanksdState
## 3           121      132 27.22735      164       71          18
## 2           113      132 27.22735      164       71          18
## 1            67      132 27.22735      164       71          18
## 6           335      132 27.22735      164       71          18
## 4           320      132 27.22735      164       71          18
## 5           597      132 27.22735      164       71          18
## 26         1219       67 10.27299       84       43          46
## 7           737       67 10.27299       84       43          46
## 17         1530       67 10.27299       84       43          46
## 40         1153       67 10.27299       84       43          46
```

```
stats[order(stats$RanksdCounty, stats$RanksdState), ][1:10, c("County.State", "medCounty", "sdCo
unty", "maxCounty", "minCounty", "RanksdCounty", "medState", "sdState", "maxState", "minState",
"RanksdState")]
```

```
##               County.State medCounty sdCounty maxCounty minCounty RanksdCounty
## 93  San Francisco County.CA     560.0 235.2447    1065.0     199.0            1
## 94      San Mateo County.CA     485.0 202.3455     991.5     196.0            2
## 95    Santa Clara County.CA     439.5 182.3114     935.0     179.0            3
## 723     Nantucket County.MA     575.0 177.3431     826.0     166.0            4
## 167        Pitkin County.CO     627.0 176.8990     911.5     265.5            5
## 98          Marin County.CA     518.5 173.7169     835.0     202.5            6
## 155    San Miguel County.CO     457.0 132.7619     629.0     221.0            7
## 187 District of Columbia.DC     334.0 129.6077     520.0     101.0            8
## 344          Maui County.HI     323.0 122.1398     542.0     150.5            9
## 102       Alameda County.CA     301.5 121.3668     581.0     133.0           10
##     medState   sdState maxState minState RanksdState
## 93       202  67.86217    314.0     98.0           3
## 94       202  67.86217    314.0     98.0           3
## 95       202  67.86217    314.0     98.0           3
## 723      173  40.23260    223.0     83.5           5
## 167      144  38.72741    242.0     79.0           8
## 98       202  67.86217    314.0     98.0           3
## 155      144  38.72741    242.0     79.0           8
## 187      334 129.60775    520.0    101.0           1
## 344      295 100.53890    448.5    139.0           2
## 102      202  67.86217    314.0     98.0           3
```

```
stats[order(stats$RanksdState, stats$RanksdCounty), ][1:10,c("County.State", "medCounty", "sdCou
nty", "maxCounty", "minCounty", "RanksdCounty", "medState", "sdState", "maxState", "minState",
"RanksdState") ]
```

```
##               County.State medCounty  sdCounty maxCounty minCounty
## 187 District of Columbia.DC     334.0 129.60775     520.0     101.0
## 344          Maui County.HI     323.0 122.13983     542.0     150.5
## 346      Honolulu County.HI     359.0 115.78915     518.0     157.5
## 345         Kauai County.HI     313.0 108.67447     464.0     130.5
## 347        Hawaii County.HI     199.0  70.22196     327.0      98.0
## 93  San Francisco County.CA     560.0 235.24473    1065.0     199.0
## 94      San Mateo County.CA     485.0 202.34553     991.5     196.0
## 95    Santa Clara County.CA     439.5 182.31145     935.0     179.0
## 98          Marin County.CA     518.5 173.71687     835.0     202.5
## 102       Alameda County.CA     301.5 121.36684     581.0     133.0
##     RanksdCounty medState   sdState maxState minState RanksdState
## 187            8      334 129.60775    520.0      101           1
## 344            9      295 100.53890    448.5      139           2
## 346           14      295 100.53890    448.5      139           2
## 345           19      295 100.53890    448.5      139           2
## 347           33      295 100.53890    448.5      139           2
## 93             1      202  67.86217    314.0       98           3
## 94             2      202  67.86217    314.0       98           3
## 95             3      202  67.86217    314.0       98           3
## 98             6      202  67.86217    314.0       98           3
## 102           10      202  67.86217    314.0       98           3
```

Part 2: Questions 1 & 2

```
## question 1
stats[stats$RankGroupCounty == 1 & stats$RankGroupState == 4,c("County.State","sdCounty", "Ranks
dCounty", "sdState", "RanksdState") ]
```

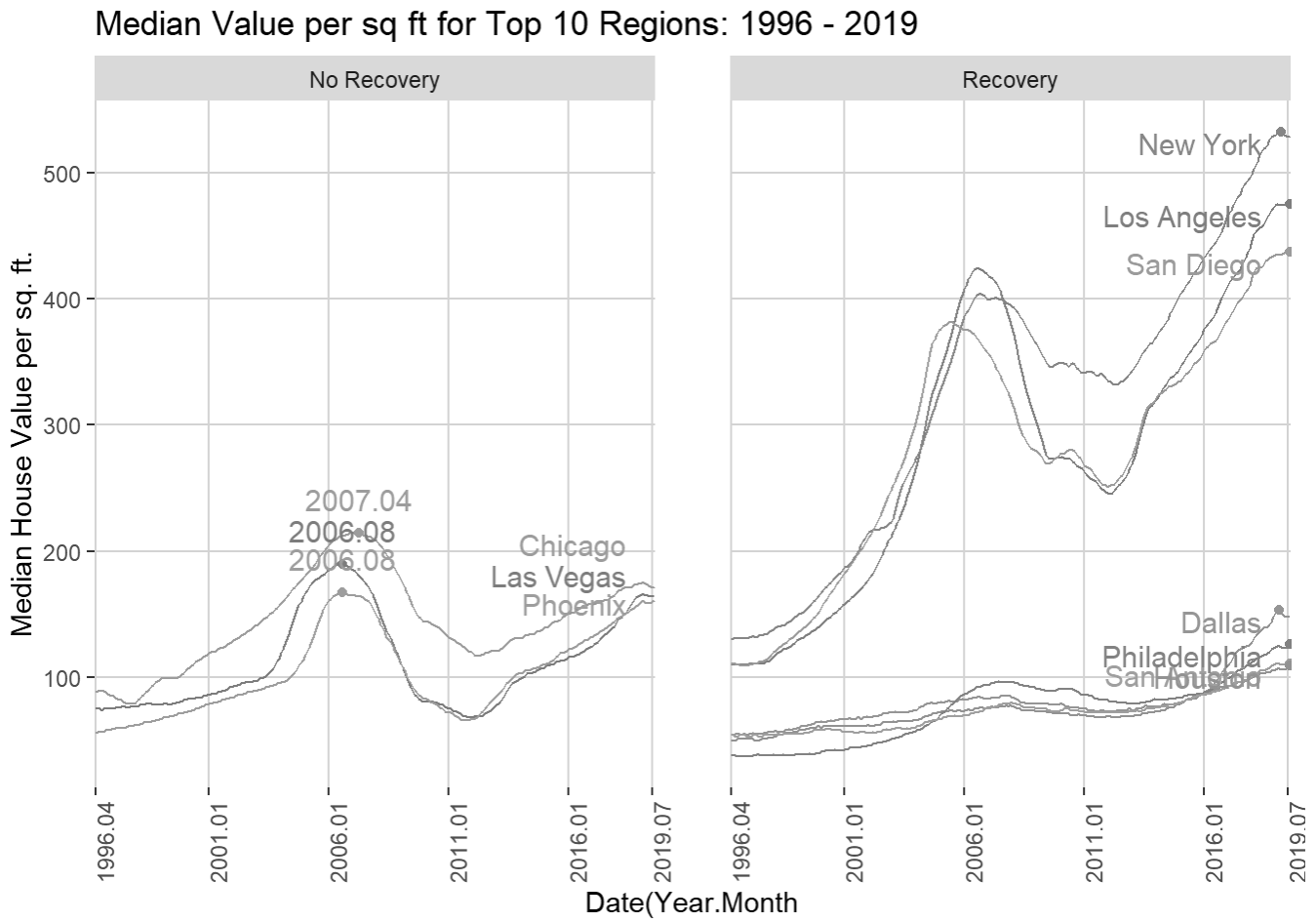
```
##              County.State sdCounty RanksdCounty   sdState RanksdState
## 255       Glynn County.GA 49.13492           78 12.641920          39
## 262     Pickens County.GA 36.62003          183 12.641920          39
## 276       Union County.GA 26.38875          371 12.641920          39
## 285      Dekalb County.GA 27.40055          338 12.641920          39
## 322      Fannin County.GA 24.40402          413 12.641920          39
## 325      Greene County.GA 40.15804          139 12.641920          39
## 687    Franklin County.KY 30.30310          278 11.512572          40
## 1410   Beaufort County.SC 24.89666          398 11.288980          41
## 1421 Georgetown County.SC 24.68294          406 11.288980          41
## 1424 Charleston County.SC 37.03991          178 11.288980          41
## 1436     Jasper County.SC 27.08212          350 11.288980          41
## 1446   Cheatham County.TN 24.64947          408 13.289865          38
## 1457   Davidson County.TN 35.75408          193 13.289865          38
## 1506 Rutherford County.TN 24.79528          402 13.289865          38
## 1510 Williamson County.TN 32.58761          239 13.289865          38
## 1528      Meigs County.TN 24.21444          420 13.289865          38
## 1854   Berkeley County.WV 24.78535          403  9.843338          47
## 1857 Monongalia County.WV 23.10754          455  9.843338          47
```

```
##question 2
stats <- stats[order(stats$RanksdCounty), ]
unique(stats[1:50,"State"])
```

```
##  [1] "CA" "MA" "CO" "DC" "HI" "FL" "NY" "VA" "WA" "UT" "NJ" "RI" "NV"
```

Part 3

```
regions <- zm %>% dplyr::group_by(RegionID) %>%
                  dplyr::summarize(median = median(MedHouse, na.rm = TRUE), max = max(MedHouse,
 na.rm = TRUE))
regions <- as.data.frame(regions)
regions <- merge(regions, zm, by.x = c("RegionID", "max"), by.y = c("RegionID", "MedHouse"))
regions <- arrange(regions, RegionID, Year.Month)
regions <- as.data.frame(regions %>% dplyr::group_by(RegionID) %>% slice(which.max(Year.Month)))


regions <- (dplyr::rename(regions, Peak.Date = Year.Month, Peak.Year = Year, Peak.Month = Mont
h))
regions$Recover <- regions$Peak.Year >= 2018
regions[regions$Recover == TRUE, "Recover"] <- "Recovery"
regions[regions$Recover == FALSE, "Recover"] <- "No Recovery"



library(ggplot2)
zn <- zm[zm$SizeRank <= 10, ]
topRegions <- regions[regions$SizeRank <=10, ]

zn <- merge(zn, topRegions)
zn <- arrange(zn, SizeRank, Year.Month)
zn$Recover <- zn$Peak.Year >= 2018
zn[zn$Recover == TRUE, "Recover"] <- "Recovery"
zn[zn$Recover == FALSE, "Recover"] <- "No Recovery"

topRegions$Peak.Label <- topRegions$Peak.Date
topRegions[topRegions$Recover == "Recovery", "Peak.Label"] <- ""


ggplot(data = zn, aes(x = Year.Month, y = MedHouse, group = RegionName)) +
        geom_line(aes(color = RegionName)) +
        scale_x_discrete(breaks = c("1996.04", "2001.01", "2006.01",
                                    "2011.01", "2016.01","2019.07")) +
        facet_wrap(~Recover) +
        geom_point(data = topRegions, aes(x = Peak.Date, y = max, color = RegionName)) +
        geom_text(data = topRegions, aes(x = Peak.Date, y = max, label = Peak.Label, color = Reg
ionName), vjust = -1, size = 4) +
        geom_text(data = topRegions, aes(x = "2018.06", y = max, label = RegionName, color = Reg
ionName), vjust = 1, hjust = 1, size = 4) +
        theme(legend.position = "none", axis.text.x = element_text(angle = 90, hjust = 1),
              panel.spacing = unit(2, "lines"), panel.background = element_rect(fill = "white"),
              panel.grid.major = element_line(color = "light grey")) +
        labs(x = "Date(Year.Month", y = "Median House Value per sq. ft.", title = "Median Value
 per sq ft for Top 10 Regions: 1996 - 2019")
```


Median Value per sq ft for Top 10 Regions: 1996 - 2019
No Recovery
Recovery
Median House Value per sq. ft.
500
400
300
200
100
2007.04
2006.08
2006.08
Chicago
Las Vegas
Phoenix
New York
Los Angeles
San Diego
Dallas
Philadelphia
1996.04
2001.01
2006.01
2011.01
2016.01
2019.07
Date(Year.Month