Group\_6\_Stat\_305\_Q2

Lulu Zhang, Aaron Deskins, Kawthar Abdulkabir

10/24/2018

knitr::opts\_chunk$set(echo = TRUE)  
  
library(grid)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages ──────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.0.0 ✔ readr 1.1.1  
## ✔ tibble 1.4.2 ✔ purrr 0.2.5  
## ✔ tidyr 0.8.1 ✔ stringr 1.3.1  
## ✔ ggplot2 3.0.0 ✔ forcats 0.3.0

## ── Conflicts ─────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
  
statejobdata <- read.csv(file="/Users/Kawthar/Documents/Unemployment\_edited.csv", header=TRUE, sep=",")

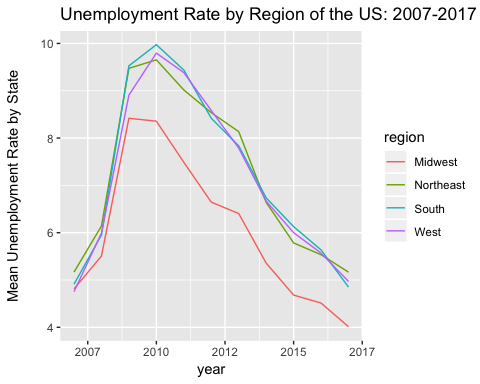
# Question 1

# What is the mean unemployment value by region for the years 2007 to 2017?

# Answer

First we mutated the df to separate each state by region. To get mean unemployment value by state in each region, we used dplyr to group the data by region. Next we summarised each year by its mean value. The state column originally had 3275 observations which were condensed down to 51 for each state and washington DC. The states were originally observed by county as well which is why there were so many observations. The new data frame ‘meanstatejobdata’ contains just 51 observations of 12 variables. This will allow us to visualize the values by year. Next we used the gather function to make the df long instead of wide. This was to allow each mean value to be stored in a ‘year’ column to be graphed as x. Next we ran the min/max functions to find the range.

#function to make all column names lowercase  
colnames(statejobdata) <- colnames(statejobdata) %>% str\_to\_lower()  
  
statejobdata <- statejobdata %>% mutate(region = ifelse(state %in% c("AZ","CO","ID","NM","MT","UT","NV","WY","AK","HI","OR","WA","CA"), "West", ifelse(state %in% c("IN","IL","MI","OH","WI","IA","KS","MN","MO","NE","ND","SD"), "Midwest", ifelse(state %in% c("DE","DC","FL","GA","MD","NC","AL","KT","MS","TN","AR","LA","OK","TX","SC","VA","WV"), "South", "Northeast"))))  
  
meanstatejobdata <- statejobdata %>%   
 group\_by(region) %>%   
 summarise(  
 "2007" = mean(unemployment\_rate\_2007,na.rm = T),  
 "2008" = mean(unemployment\_rate\_2008,na.rm = T),  
 "2009" = mean(unemployment\_rate\_2009,na.rm = T),  
 "2010" = mean(unemployment\_rate\_2010,na.rm = T),  
 "2011" = mean(unemployment\_rate\_2011,na.rm = T),  
 "2012" = mean(unemployment\_rate\_2012,na.rm = T),  
 "2013" = mean(unemployment\_rate\_2013,na.rm = T),  
 "2014" = mean(unemployment\_rate\_2014,na.rm = T),  
 "2015" = mean(unemployment\_rate\_2015,na.rm = T),  
 "2016" = mean(unemployment\_rate\_2016,na.rm = T),  
 "2017" = mean(unemployment\_rate\_2017,na.rm = T)  
 ) %>% select(region, everything())  
  
  
#changed the above table to wide  
mean\_region <- meanstatejobdata %>% gather(`2007`,`2008`,`2009`,`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,`2017`, key = "year", value = "Mean\_Unemployment")  
  
#next changed the variable year to integer  
mean\_region$year <- as.numeric(mean\_region$year)  
  
ggplot(mean\_region, aes(x=year, y= Mean\_Unemployment))+geom\_line(aes(col=region)) + labs("\nYear", y= "Mean Unemployment Rate by State\n", title = "Unemployment Rate by Region of the US: 2007-2017") + scale\_x\_continuous(labels = function (x) floor(x))



range(mean\_region$region)

## [1] "Midwest" "West"

range

## function (..., na.rm = FALSE) .Primitive("range")

max(mean\_region$region)

## [1] "West"

# Q1 Conclusion

The data for all regions is skewed right which indicates the mean value is greater than the median. For every region of the US, the mean unemployment was highest for between 2009 to about 2011. The Midwest had the lowest rate across all years and the West had the highest. In the future, some further explorations can be: What factors lead the midwest to have the lowest rate for all years? How does the overall civilian labor force compare to the rate of unemployment per year?

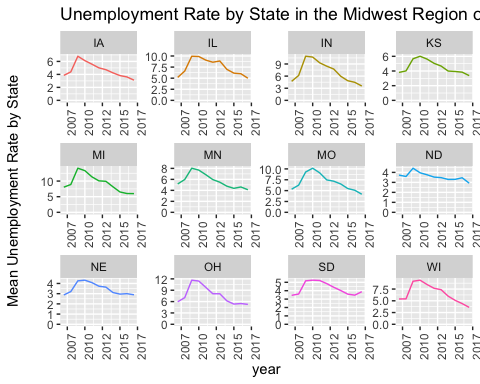
# Question 2

# What is the mean unemployment value by state for regions with the Max and Min value for the years 2007-2017?

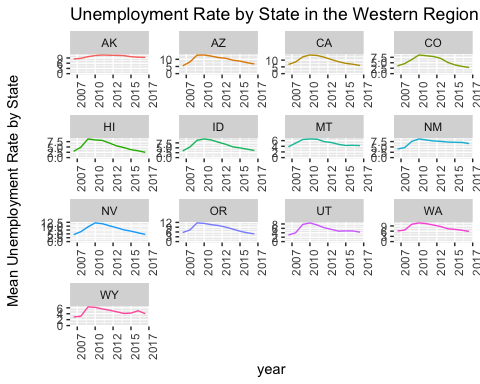
# Answer

First we performed the same mutations and summaries as stated in answer 1. The difference here was that we also filtered the df by theregion wanted to display and we grouped by ‘state’ instead of region. Finally we applied a facet\_wrap by state, and adjusted the theme, scales, and labels. Each facet scale is adjusted to fit the range of each state.

mean\_by\_state <- statejobdata %>% filter(region == "Midwest") %>%  
 group\_by(state) %>%   
 summarise(  
 "2007" = mean(unemployment\_rate\_2007,na.rm = T),  
 "2008" = mean(unemployment\_rate\_2008,na.rm = T),  
 "2009" = mean(unemployment\_rate\_2009,na.rm = T),  
 "2010" = mean(unemployment\_rate\_2010,na.rm = T),  
 "2011" = mean(unemployment\_rate\_2011,na.rm = T),  
 "2012" = mean(unemployment\_rate\_2012,na.rm = T),  
 "2013" = mean(unemployment\_rate\_2013,na.rm = T),  
 "2014" = mean(unemployment\_rate\_2014,na.rm = T),  
 "2015" = mean(unemployment\_rate\_2015,na.rm = T),  
 "2016" = mean(unemployment\_rate\_2016,na.rm = T),  
 "2017" = mean(unemployment\_rate\_2017,na.rm = T)  
 ) %>% mutate(region = ifelse(state %in% c("AZ","CO","ID","NM","MT","UT","NV","WY","AK","HI","OR","WA","CA"), "West", ifelse(state %in% c("IN","IL","MI","OH","WI","IA","KS","MN","MO","NE","ND","SD"), "Midwest", ifelse(state %in% c("DE","DC","FL","GA","MD","NC","AL","KT","MS","TN","AR","LA","OK","TX","SC","VA","WV"), "South", "Northeast"))))   
  
  
mbs <- mean\_by\_state %>% gather(`2007`,`2008`,`2009`,`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,`2017`, key = "year", value = "Mean\_Unemployment")  
  
mbs$year <- as.integer(mbs$year)  
  
ggplot(mbs, aes(x=year, y= Mean\_Unemployment))+geom\_line(aes(col=state)) + facet\_wrap(~state, scales = "free")+expand\_limits(y=0) + theme(legend.position = "none") + labs("\nYear", y= "Mean Unemployment Rate by State\n", title = "Unemployment Rate by State in the Midwest Region of the US: 2007-2017") + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + scale\_x\_continuous(labels = function (x) floor(x))



mean\_by\_state <- statejobdata %>% filter(region == "West") %>%  
 group\_by(state) %>%   
 summarise(  
 "2007" = mean(unemployment\_rate\_2007,na.rm = T),  
 "2008" = mean(unemployment\_rate\_2008,na.rm = T),  
 "2009" = mean(unemployment\_rate\_2009,na.rm = T),  
 "2010" = mean(unemployment\_rate\_2010,na.rm = T),  
 "2011" = mean(unemployment\_rate\_2011,na.rm = T),  
 "2012" = mean(unemployment\_rate\_2012,na.rm = T),  
 "2013" = mean(unemployment\_rate\_2013,na.rm = T),  
 "2014" = mean(unemployment\_rate\_2014,na.rm = T),  
 "2015" = mean(unemployment\_rate\_2015,na.rm = T),  
 "2016" = mean(unemployment\_rate\_2016,na.rm = T),  
 "2017" = mean(unemployment\_rate\_2017,na.rm = T)  
 ) %>% mutate(region = ifelse(state %in% c("AZ","CO","ID","NM","MT","UT","NV","WY","AK","HI","OR","WA","CA"), "West", ifelse(state %in% c("IN","IL","MI","OH","WI","IA","KS","MN","MO","NE","ND","SD"), "Midwest", ifelse(state %in% c("DE","DC","FL","GA","MD","NC","AL","KT","MS","TN","AR","LA","OK","TX","SC","VA","WV"), "South", "Northeast"))))   
  
mbs <- mean\_by\_state %>% gather(`2007`,`2008`,`2009`,`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,`2017`, key = "year", value = "Mean\_Unemployment")  
  
mbs$year <- as.integer(mbs$year)  
  
ggplot(mbs, aes(x=year, y= Mean\_Unemployment))+geom\_line(aes(col=state)) + facet\_wrap(~state, scales = "free")+expand\_limits(y=0) + theme(legend.position = "none") + labs("\nYear", y= "Mean Unemployment Rate by State\n", title = "Unemployment Rate by State in the Western Region of the US: 2007-2017") + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + scale\_x\_continuous(labels = function (x) floor(x))



# Question 2 Conclusion

The Midwestern region has the has lowest overall rate for all regions and the West has the highest. Some questions that can explored are, what are the populations differences in these states? Do the population densities contribute to these ranges? What level of urban influence contributes to unemployment rate?

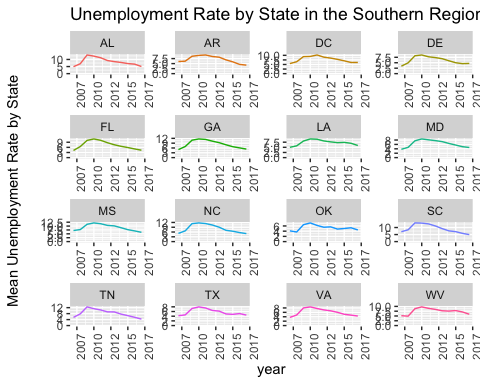
# Question 3

# What is the mean unemployment value by state for the mid range regions for the years 2007-2017?

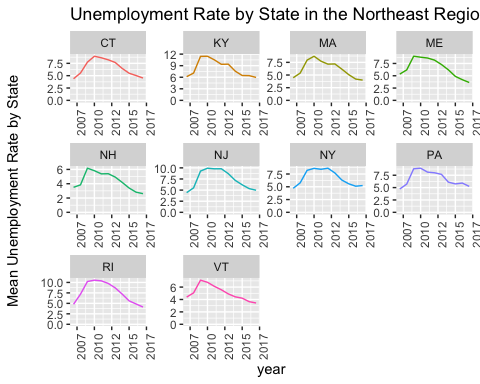
# Answer

We performed the same mutations and summaries as stated in question 2, filtering by the desired regions.

mean\_by\_state <- statejobdata %>% filter(region == "South") %>%  
 group\_by(state) %>%   
 summarise(  
 "2007" = mean(unemployment\_rate\_2007,na.rm = T),  
 "2008" = mean(unemployment\_rate\_2008,na.rm = T),  
 "2009" = mean(unemployment\_rate\_2009,na.rm = T),  
 "2010" = mean(unemployment\_rate\_2010,na.rm = T),  
 "2011" = mean(unemployment\_rate\_2011,na.rm = T),  
 "2012" = mean(unemployment\_rate\_2012,na.rm = T),  
 "2013" = mean(unemployment\_rate\_2013,na.rm = T),  
 "2014" = mean(unemployment\_rate\_2014,na.rm = T),  
 "2015" = mean(unemployment\_rate\_2015,na.rm = T),  
 "2016" = mean(unemployment\_rate\_2016,na.rm = T),  
 "2017" = mean(unemployment\_rate\_2017,na.rm = T)  
 ) %>% mutate(region = ifelse(state %in% c("AZ","CO","ID","NM","MT","UT","NV","WY","AK","HI","OR","WA","CA"), "West", ifelse(state %in% c("IN","IL","MI","OH","WI","IA","KS","MN","MO","NE","ND","SD"), "Midwest", ifelse(state %in% c("DE","DC","FL","GA","MD","NC","AL","KT","MS","TN","AR","LA","OK","TX","SC","VA","WV"), "South", "Northeast"))))   
  
# now we will change the above table to long instead of wide  
mbs <- mean\_by\_state %>% gather(`2007`,`2008`,`2009`,`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,`2017`, key = "year", value = "Mean\_Unemployment")  
  
# next change the variable year to integer  
mbs$year <- as.integer(mbs$year)  
  
ggplot(mbs, aes(x=year, y= Mean\_Unemployment))+geom\_line(aes(col=state)) + facet\_wrap(~state, scales = "free")+expand\_limits(y=0) + theme(legend.position = "none") + labs("\nYear", y= "Mean Unemployment Rate by State\n", title = "Unemployment Rate by State in the Southern Region of the US: 2007-2017") + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + scale\_x\_continuous(labels = function (x) floor(x))



mean\_by\_state <- statejobdata %>% filter(region == "Northeast") %>%  
 group\_by(state) %>%   
 summarise(  
 "2007" = mean(unemployment\_rate\_2007,na.rm = T),  
 "2008" = mean(unemployment\_rate\_2008,na.rm = T),  
 "2009" = mean(unemployment\_rate\_2009,na.rm = T),  
 "2010" = mean(unemployment\_rate\_2010,na.rm = T),  
 "2011" = mean(unemployment\_rate\_2011,na.rm = T),  
 "2012" = mean(unemployment\_rate\_2012,na.rm = T),  
 "2013" = mean(unemployment\_rate\_2013,na.rm = T),  
 "2014" = mean(unemployment\_rate\_2014,na.rm = T),  
 "2015" = mean(unemployment\_rate\_2015,na.rm = T),  
 "2016" = mean(unemployment\_rate\_2016,na.rm = T),  
 "2017" = mean(unemployment\_rate\_2017,na.rm = T)  
 ) %>% mutate(region = ifelse(state %in% c("AZ","CO","ID","NM","MT","UT","NV","WY","AK","HI","OR","WA","CA"), "West", ifelse(state %in% c("IN","IL","MI","OH","WI","IA","KS","MN","MO","NE","ND","SD"), "Midwest", ifelse(state %in% c("DE","DC","FL","GA","MD","NC","AL","KT","MS","TN","AR","LA","OK","TX","SC","VA","WV"), "South", "Northeast"))))   
  
# now we will change the above table to long instead of wide  
mbs <- mean\_by\_state %>% gather(`2007`,`2008`,`2009`,`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,`2017`, key = "year", value = "Mean\_Unemployment")  
  
# next change the variable year to integer  
mbs$year <- as.integer(mbs$year)  
  
ggplot(mbs, aes(x=year, y= Mean\_Unemployment))+geom\_line(aes(col=state)) + facet\_wrap(~state, scales = "free")+expand\_limits(y=0) + theme(legend.position = "none") + labs("\nYear", y= "Mean Unemployment Rate by State\n", title = "Unemployment Rate by State in the Northeast Region of the US: 2007-2017") + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + scale\_x\_continuous(labels = function (x) floor(x))



# Question 3 Conclusion

The states in these regions have a wide spread in population levels. This could contribute to the spread in rate. As well as give insight into why these states fall in the middle region between West and Midwest. Questions we could ask are what is the population for each state in the region? How do populations differ between adjoining states? What level or rural and urban influence is there by state? How does the employed labor force compare to the civilian labor force?