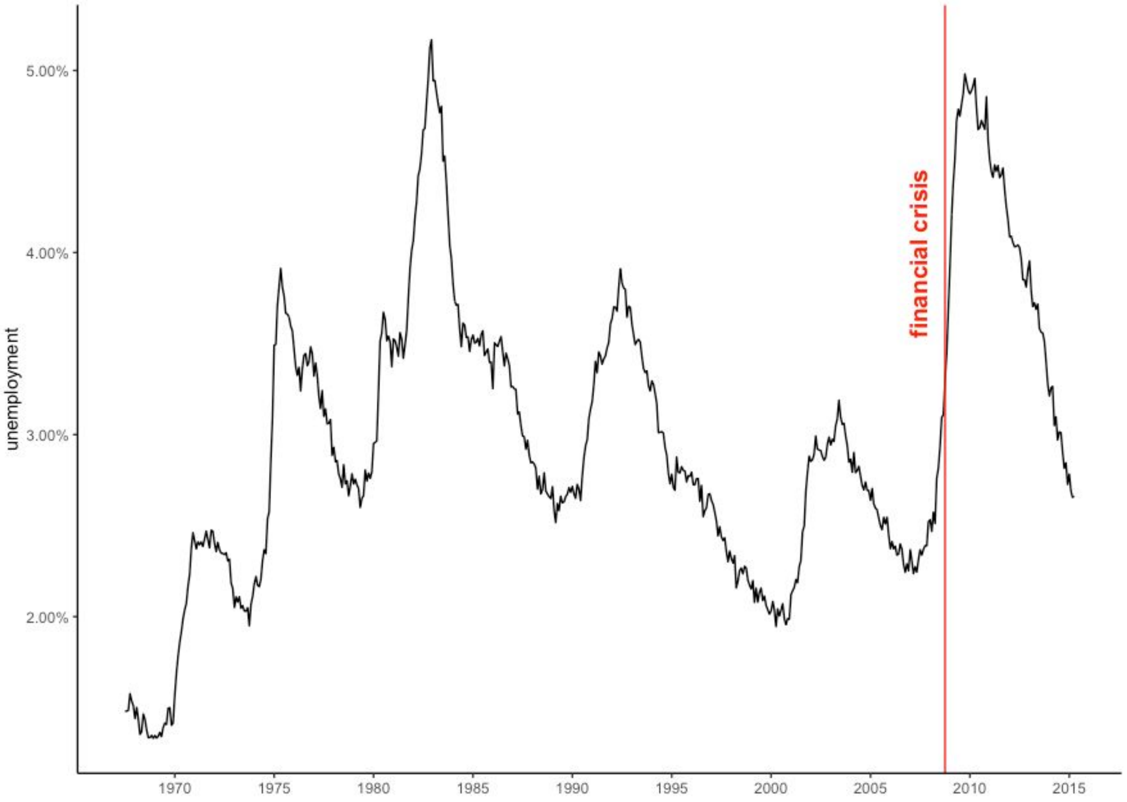
Student Name: Yifan Feng
Student No.: 2672017
Assignment: Data Science VAR: Take-home Exam

Part A: Develop an Infographic

1.1.

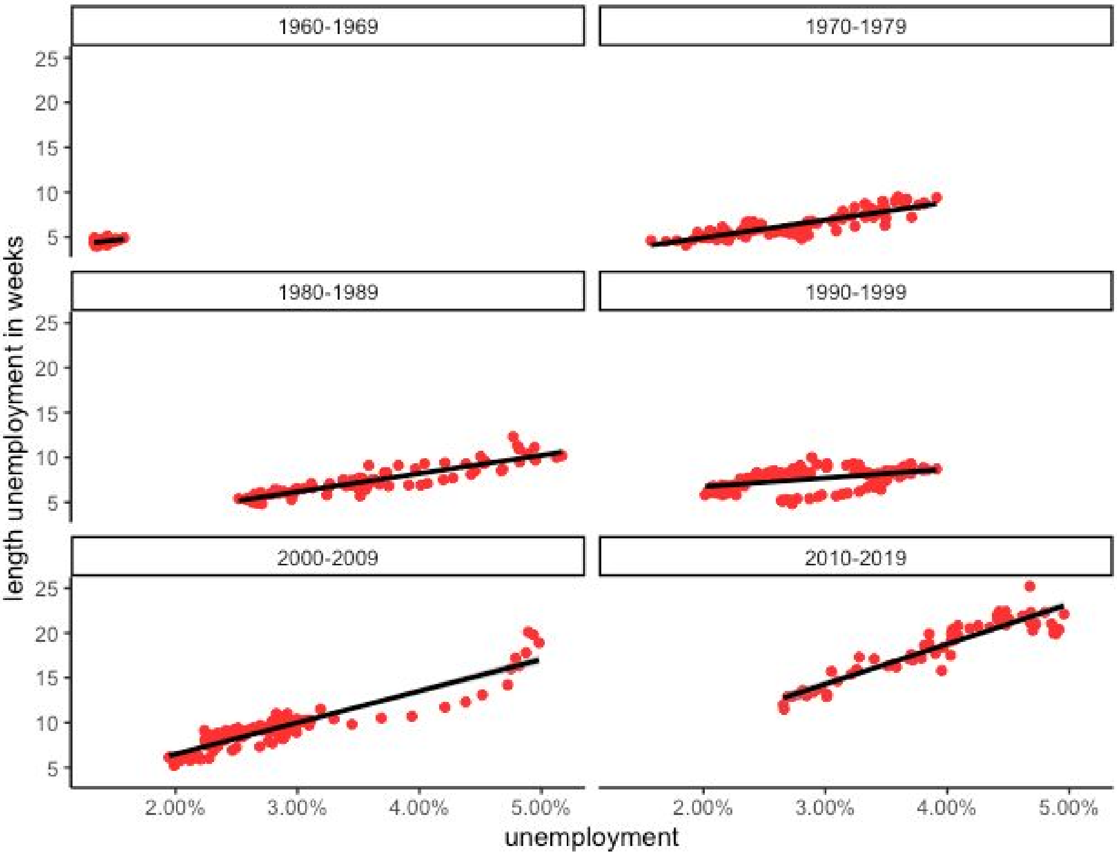
```
> library(tidyverse)
> econ <- economics %>% mutate(unemployment_rate = unemploy/pop, crisis =
ifelse(date < as.Date("2008-10-01"), "before", "after" ))
> ggplot(econ, mapping = aes(x=date, y=unemployment_rate)) + geom_line() +
scale_x_date(date_breaks = "5 year",date_labels = "%Y") + geom_vline(xintercept =
as.Date("2008-10-01"), color = "red", size = 0.5) + annotate(geom="text",x =
as.Date("2008-10-01"),y=0.04, label="financial crisis",color="red", angle=90, vjust
=-1,size=5,fontface=2) + scale_y_continuous(labels = scales::percent) +
theme_classic() + labs(y="unemployment") + theme(axis.title.x=element_blank())
```



(figure 01)

1.2

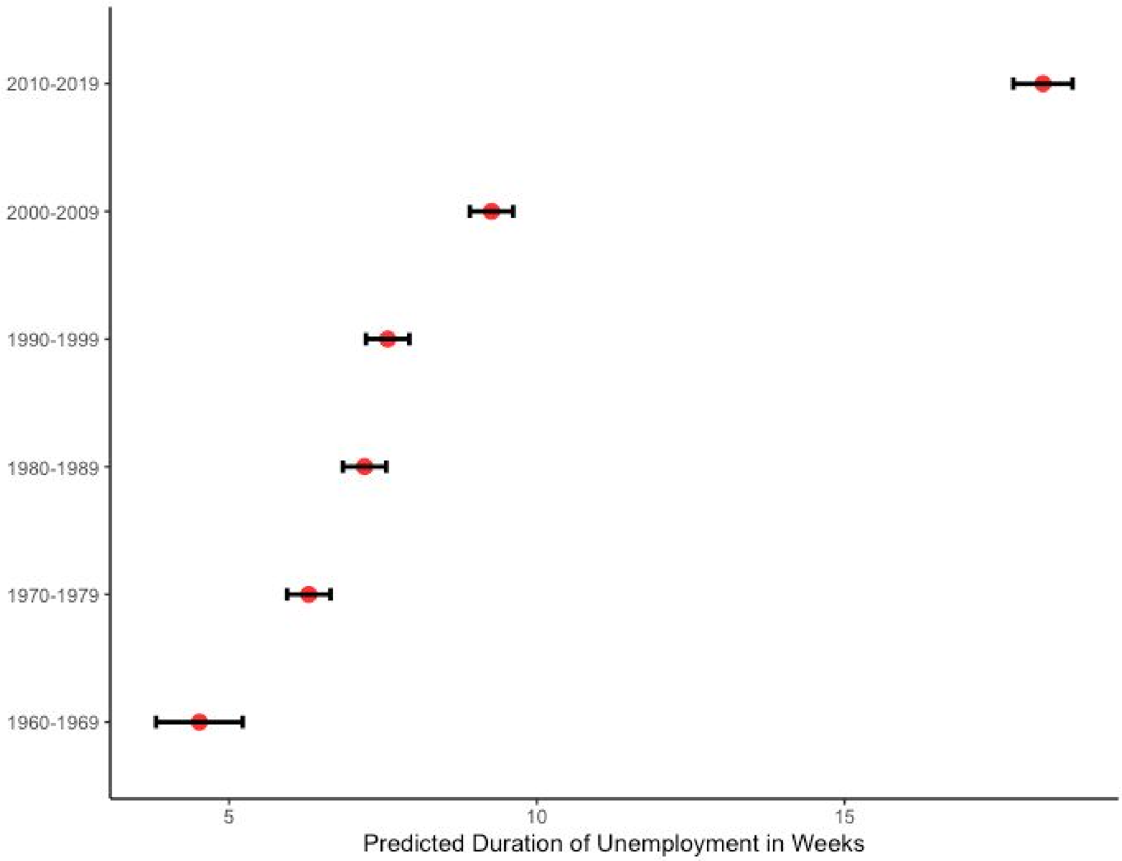
```
> econ <- econ %>%mutate(decade = factor(paste0(substr(date,1,3),0,
"-",substr(date, 1,3),9)) )
> ggplot(econ, mapping = aes(x= unemployment_rate,y=uempmed)) +
geom_point(color="firebrick1") + facet_wrap(.~decade, ncol=2) +
geom_smooth(method="glm", color="black") + scale_x_continuous(labels =
scales::percent) + theme_classic() + labs(x="unemployment",y="length
unemployment in weeks")
```



(figure 02)

1.3

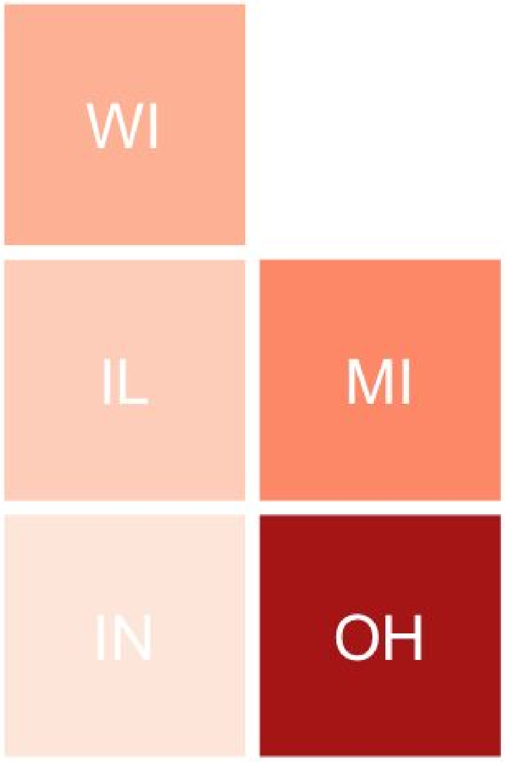
```
> m1 <- lm(uempmed ~ 0 + decade, data = econ)
> results <- data.frame(decade = gsub("decade","",rownames(confint(m1))), pred =
coef(m1), lower = confint(m1)[,1], upper = confint(m1)[,2])
> ggplot(results, mapping = aes(x=decade, y=pred)) +
geom_point(color="firebrick1",size=3) + geom_errorbar(mapping = aes(x=decade,
ymin=lower,ymax=upper),width=0.1,size=1) +
coord_flip() + theme_classic() +labs(y="Predicted Duration of Unemployment in
Weeks") + theme(axis.title.y=element_blank())
```



(figure 03)

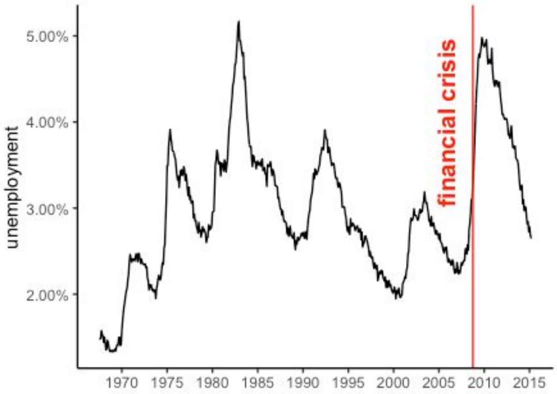
1.4

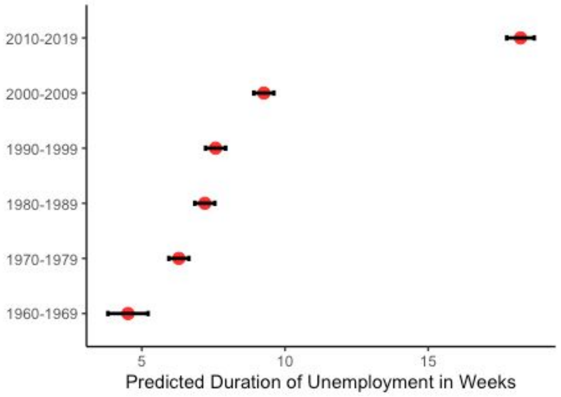
```
> library(statebins)
> statebins_continuous(state_data = midwest, state_col = "state", value_col =
"percadultpoverty",text_color = "white", font_size= 10, brewer_pal = "Reds",
legend_position = "none" )
```

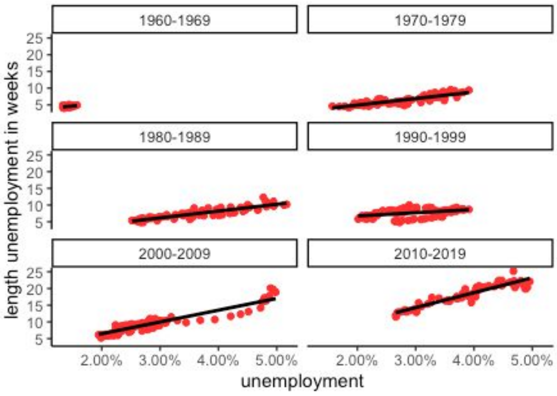


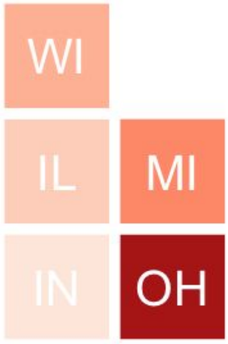
(figure 04)

```
> multiplot(p1,p2,p3,p4, cols=2)
```









Appendix:

```
> # Multiple plot function
> #
> # ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
> # - cols:   Number of columns in layout
> # - layout: A matrix specifying the layout. If present, 'cols' is ignored.
> #
> # If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
> # then plot 1 will go in the upper left, 2 will go in the upper right, and
> # 3 will go all the way across the bottom.
> #
> multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
+     library(grid)
+
+     # Make a list from the ... arguments and plotlist
+     plots <- c(list(...), plotlist)
+
+     numPlots = length(plots)
+
+     # If layout is NULL, then use 'cols' to determine layout
+     if (is.null(layout)) {
+         # Make the panel
+         # ncol: Number of columns of plots
+         # nrow: Number of rows needed, calculated from # of cols
+         layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
+                          ncol = cols, nrow = ceiling(numPlots/cols))
+     }
+
+     if (numPlots==1) {
+         print(plots[[1]])
+
+     } else {
+         # Set up the page
+         grid.newpage()
+         pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
+
+         # Make each plot, in the correct location
+         for (i in 1:numPlots) {
+             # Get the i,j matrix positions of the regions that contain this subplot
+             matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
+
+             print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
+                                             layout.pos.col = matchidx$col))
+         }
+     }
+ }
```

Part B: Comment on Your Own Infographic

Q1: What do each of the individual graphs demonstrate: write it up as if you would inform an Economist reader.

A1:
Figure 01 shows the turbulent unemployment rate over time (timeframe 1970 to 2015). It points out a clear date for a major financial crisis and its before-after influence in terms of rate’s trending.
Figure 02 visualizes the duration of unemployment and its proportion in percentage regarding segmented time periods (timeframe: a decade). Each red spot records one unemployment data and the regression line shows the average duration of given events.
Figure 03 pictures the gradually increasing trend of unemploying duration from 1960 to the present regarding upper and lower rates per decade.
Figure 04 demonstrates unemployment in specific locations (five US states). The color scale (lightness to darkness) represents the density and severity. The position of each square represents relative geo-location in maps.

Q2: What if you would write for a tabloid newspaper, would you assess the visual literacy of the readership to change? If so, what would you adjust in the infographic?

A2: Yes. Graph design should be adaptable to various audiences. Tabloids feature small pages, short texts, many graphs and target at working-class people. The plot should contain more colors or bigger, varied shapes to sensationalize information. It should also stress limited selected messages such as average unemployment duration, maximum, and minimum unemployment rates. Also, the timeframe should be shrunk to recent decades since such information is more up-to-date and relatable for readers and such a plot can contain more details (e.g. particular dates).

Q3: What are the strengths of this infographic, what are the weaknesses? Would you recommend a data scientist writing a report for a general audience to use it, and if so, for which purposes, and if not, why?

A3:

Advantages:

It presents straightforward messages for audiences to understand the unemployment analysis regarding its trend, location, predictions. It can illustrate related quantitative data over long time periods without cumbersome literal descriptions.

Disadvantages:

It lacks detailed explanations for the encodings and values. For example, it should indicate the use of color scales by legends and explain label shapes by units. Sources(data) are missing in the plot which is hard to read the messages (in fig4). Also, the faceted plot (figure03) is hard to read since the unemployment duration's differentiation is not clear and obvious.

Reason:

I'll suggest using this infographic if one wants to present quantitative data in an attractive way. It is more important to show relationships like exact numbers (unemployment rates) and the changes (overall trend). This one is informative and descriptive regarding various metadata (variables) such as geospatial distributions. Color palettes are also used to express additional information (density) and highlight important information (time for financial crisis).