Lab2

Lingchen Lou

2018/9/26

Question 3: write the R code to recreate the graph from the Activity 2

```
library(ggplot2)
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.1 -
## <U+221A> tibble  1.4.2     <U+221A> dplyr   0.7.6
## <U+221A> readr   1.1.1     <U+221A> stringr 1.3.1
## <U+221A> purrr   0.2.5     <U+221A> forcats 0.3.0
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#load dataset
lab2 <- read.csv("~/Downloads/lab2.csv")

#tidy lab data
lab_cleaned <- lab2 %>%
  gather(key = ID, value = cases, indexes = 2 : 10) %>%
  separate(ID,into =  c("time",  "type"), sep = "_")

#name columns
names(lab_cleaned) <- c("ID", "Time", "Type", "Cases")

for (i in 1:270){
  if (lab_cleaned$Time[i] == "base"){
    lab_cleaned$Time[i] = "baseline"
  } else if (lab_cleaned$Time[i] == "first"){
    lab_cleaned$Time[i] = "one year"
  }else{
    lab_cleaned$Time[i] = "two years"
  }
}

#get mean and standard deviation of the tidy data
lab <- lab_cleaned %>%
  group_by(Time, Type) %>%
  summarise(mean = mean(Cases),
            sd = sd(Cases) )
lab
```

```
## # A tibble: 9 x 4
## # Groups:   Time [?]
```

```
##   Time      Type   mean    sd
##   <chr>     <chr> <dbl> <dbl>
## 1 baseline  pain   28.1  4.96
## 2 baseline  qol    53.2  5.47
## 3 baseline  sport  47.7  5.33
## 4 one year  pain   70.2  6.43
## 5 one year  qol    69.9  9.06
## 6 one year  sport  77.1  5.36
## 7 two years pain   71.4  4.99
## 8 two years qol    67.1  9.90
## 9 two years sport  79.7  5.51
```

```
base <- lab[1:3,];base
```

```
## # A tibble: 3 x 4
## # Groups:   Time [1]
##   Time     Type   mean    sd
##   <chr>    <chr> <dbl> <dbl>
## 1 baseline pain   28.1  4.96
## 2 baseline qol    53.2  5.47
## 3 baseline sport  47.7  5.33
```

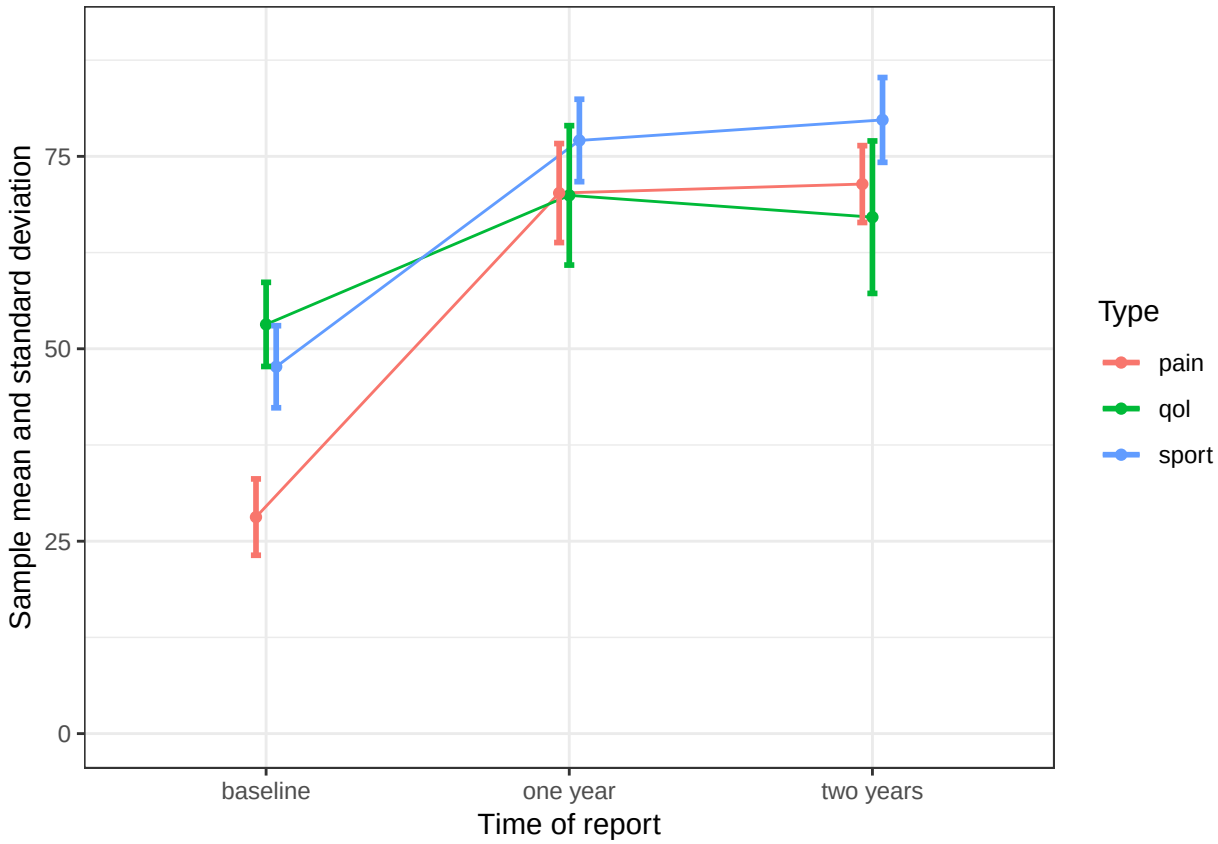
```
first<- lab[4:6,];first
```

```
## # A tibble: 3 x 4
## # Groups:   Time [1]
##   Time     Type   mean    sd
##   <chr>    <chr> <dbl> <dbl>
## 1 one year pain   70.2  6.43
## 2 one year qol    69.9  9.06
## 3 one year sport  77.1  5.36
```

```
second <- lab[7:9,];second
```

```
## # A tibble: 3 x 4
## # Groups:   Time [1]
##   Time      Type   mean    sd
##   <chr>     <chr> <dbl> <dbl>
## 1 two years pain   71.4  4.99
## 2 two years qol    67.1  9.90
## 3 two years sport  79.7  5.51
```

```
#plot
PD <- position_dodge(width = 0.1)
lab %>%
  ggplot(aes(x = Time, y = mean, col = Type))+
  geom_line( aes(group = Type), position = PD)+
  geom_point(position = PD)+
  geom_errorbar(aes(x=Time, ymin=mean - sd, ymax= mean + sd), position = PD,width=0.1, size=1)+
  ylim(0, 90)+
  theme_bw()+
  labs(x ="Time of report", y="Sample mean and standard deviation")
```



Question 4:

4.1 Make the data frames tidy

```
#load dataset

coverage<-read.csv("~/Downloads/coverage2.csv", skip = 2, header = TRUE,stringsAsFactors = FALSE)
expenditure<-read.csv("~/Downloads/expenditures02.csv",skip = 2, header = TRUE, stringsAsFactors = FALSE

#only import rows with data
coverage<-coverage[1:52,]
expenditure<-expenditure[1:52,]

#tidy coverage data

coverage_cleaned<-coverage %>%
  gather(key=location,value=case,index = 2:29,na.rm=TRUE) %>%
  separate(location,into=c("Year","Cases"),sep="__",convert=TRUE) %>%
  arrange(Year,Cases)

for (i in 1:1456){
  coverage_cleaned$Year[i] = substr(coverage_cleaned$Year[i], 2,5)
}
```

```r
names(coverage_cleaned) <- c("Location", "Year", "Type", "Value")
head(coverage_cleaned)
```

```
##        Location Year     Type     Value
## 1 United States 2013 Employer 155696900
## 2       Alabama 2013 Employer   2126500
## 3        Alaska 2013 Employer    364900
## 4       Arizona 2013 Employer   2883800
## 5      Arkansas 2013 Employer   1128800
## 6    California 2013 Employer  17747300
```

```r
#tidy expenditure data

expenditure_cleaned<-expenditure %>%
  gather(key=location,value=cases,index = 2:25,na.rm=TRUE) %>%
  separate(location,into=c("Year","Cases"),sep="__") %>%
  arrange(Year,Cases)

for (i in 1:1248){
  expenditure_cleaned$Year[i] = substr(expenditure_cleaned$Year[i], 2,5)
}

names(expenditure_cleaned) <- c("Location", "Year", "Type", "Value")
head(expenditure_cleaned)
```

```
##        Location Year                  Type  Value
## 1 United States 1991 Total.Health.Spending 675896
## 2       Alabama 1991 Total.Health.Spending  10393
## 3        Alaska 1991 Total.Health.Spending   1458
## 4       Arizona 1991 Total.Health.Spending   9269
## 5      Arkansas 1991 Total.Health.Spending   5632
## 6    California 1991 Total.Health.Spending  81438
```

4.2 Merge two data frames: the resulting data frame should contain information about coverage and expenditures for years 2013-2016. Please note that file expenditures.csv does not contain years 2015-2016.

```r
#Merge two data frames
covandexp <- rbind(coverage_cleaned,expenditure_cleaned,by=c("Location","Year"))
head(covandexp)
```

```
##        Location Year     Type     Value
## 1 United States 2013 Employer 155696900
## 2       Alabama 2013 Employer   2126500
## 3        Alaska 2013 Employer    364900
## 4       Arizona 2013 Employer   2883800
## 5      Arkansas 2013 Employer   1128800
## 6    California 2013 Employer  17747300
```

```r
covandexp$Year <- as.numeric(covandexp$Year, na.rm = TRUE)
```

```
## Warning: NAs introduced by coercion
```

```r
covandexp_sub <- subset(covandexp,Year>=2013)
tail(covandexp_sub)
```

```
##           Location Year                  Type Value
## 2699       Vermont 2014 Total.Health.Spending  6389
## 2700      Virginia 2014 Total.Health.Spending 62847
## 2701    Washington 2014 Total.Health.Spending 55819
## 2702 West Virginia 2014 Total.Health.Spending 17491
## 2703     Wisconsin 2014 Total.Health.Spending 50109
## 2704       Wyoming 2014 Total.Health.Spending  4856
```