Chaudhary_week_02

Jyoti Chaudhary

September 25, 2016

PROBLEM 1

Create a DataFrame mydata2 and load the dataset using read.csv

```
mydata2 <- read.csv(paste(getwd(),"/DOHMH_New_York_City_Restaurant_Inspection_Results.csv",sep =
 ""),header=T)
# mydata2 <- read.csv("https://nycopendata.socrata.com/api/views/xx67-kt59/rows.csv?accessType=D
OWNLOAD", header=TRUE, sep=",")
```

Create another dataframe BROOKDF from mydata2 with restaurants = Brooklyn and with cuisine equal to “Ice Cream, Gelato, Yogurt, Ices”.

```
BROOKDF <- data.frame(filter(mydata2, BORO == "BROOKLYN" & CUISINE.DESCRIPTION %in% c("Ice Crea
m, Gelato, Yogurt, Ices")))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

5 common restaurant names in BROOKDF dataframe

The most common will be listed at the top in descending order

```
by_DBA <- group_by(BROOKDF, DBA)
COMMONNAMES <- arrange((summarise(by_DBA, count = n())),desc(count))
```

What grades has Brooklyn ice cream store “SWEET EXPRESSIONS” received and how often?

Using PIPES (Total 15 Grades received. Grade A count is 11. Grade B count is 6.)

```
EXGRADE <-mydata2%>%
  filter(BORO == "BROOKLYN" & CUISINE.DESCRIPTION %in% c("Ice Cream, Gelato, Yogurt, Ices") & DB
A == "SWEET EXPRESSIONS" & GRADE !="NA")%>%
  group_by(GRADE)%>%
  summarise(n=n())%>%
  arrange(desc(n))
```

Grades received by “SWEET EXPRESSIONS” Using TABLE command. (Total Grades = 15. Grade A = 15. Grade B=6. Grade C, P, Z = 0)

```
J1 <- filter(mydata2, BORO == "BROOKLYN" & CUISINE.DESCRIPTION %in% c("Ice Cream, Gelato, Yogur
t, Ices") & DBA == "SWEET EXPRESSIONS" & GRADE !="NA")
with(J1, table(GRADE))
```

```
## GRADE
##   A B Z
## 0 0 0 0
```

PROBLEM2

Read “gapminder_2007_gini.tsv” in R and draw an interesting plot

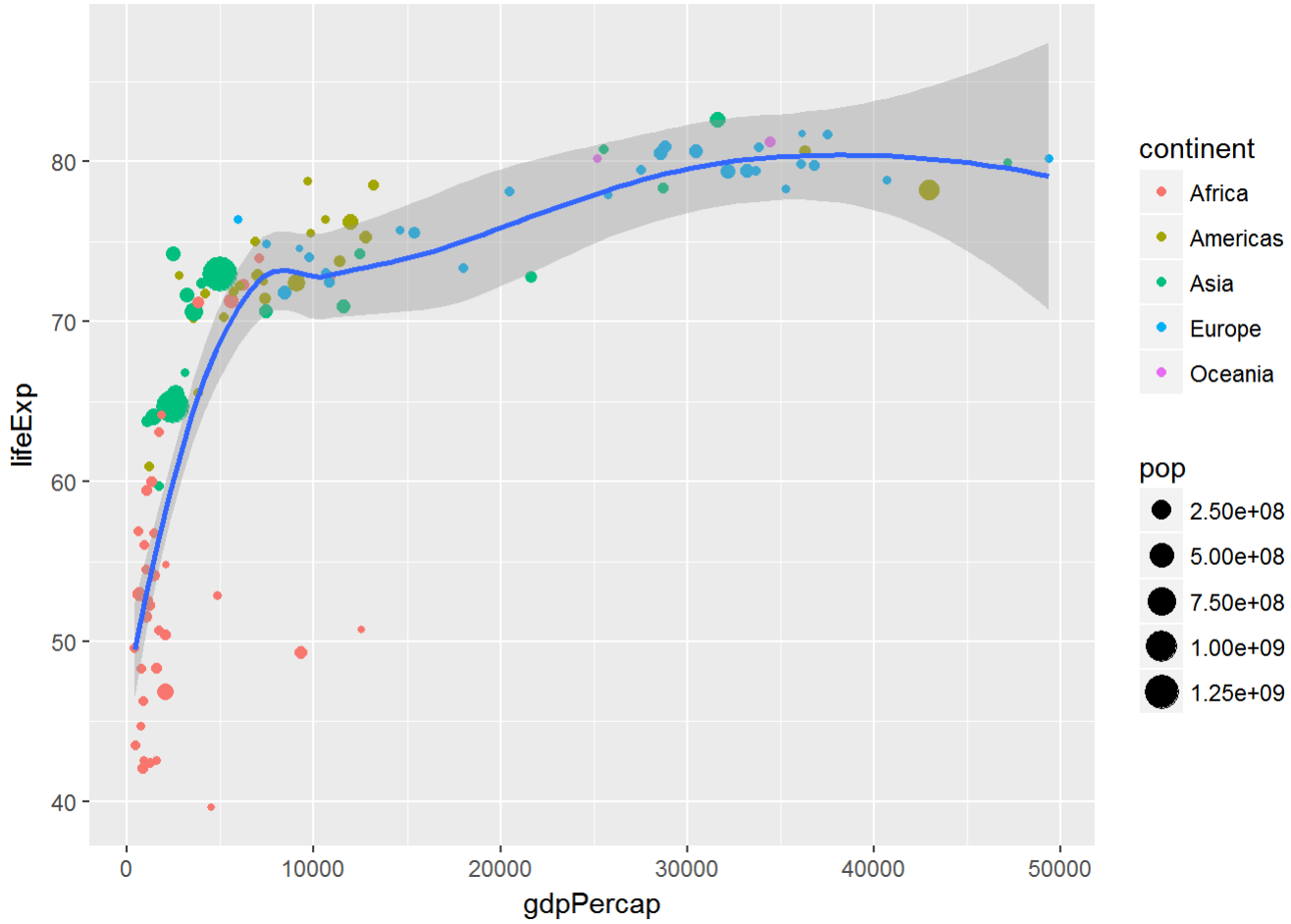
Generate a data frame showing, for each continent, the minimum, maximum, and average Gini coefficient. (The output dataframe lists data under column names Max, Min & Mean).

```
mydata3 <- read_tsv("gapminder_2007_gini.tsv")
```

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   continent = col_character(),
##   year = col_integer(),
##   lifeExp = col_double(),
##   pop = col_integer(),
##   gdpPercap = col_double(),
##   gini = col_double()
## )
```

```
ggplot(data = filter(mydata3),  aes(gdpPercap, lifeExp)) + geom_point(aes(gdpPercap, lifeExp, co
lour = continent, size = pop)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




```
by_continent <- group_by(mydata3, continent)
print(summarize(by_continent, Max = max(gini, na.rm = TRUE), Min = min(gini, na.rm = TRUE), Mean
 = mean(gini, na.rm = TRUE)))
```

```
## # A tibble: 5 x 4
##   continent   Max   Min  Mean
##   <chr>     <dbl> <dbl> <dbl>
## 1 Africa     63.2  30.8  43.9
## 2 Americas   60.8  32.1  48.2
## 3 Asia       49.0  29.6  40.2
## 4 Europe     40.2  23.7  30.5
## 5 Oceania    36.2  30.3  33.2
```

PROBLEM3

Add a new column gdp to the GAPMINDER dataframe.

```
gapminder$gdp <- gapminder$gdpPercap * gapminder$pop
glimpse(gapminder)
```

```
## Observations: 1,704
## Variables: 7
## $ country   <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, ...
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia...
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992...
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.8...
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 1488...
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 78...
## $ gdp       <dbl> 6567086330, 7585448670, 8758855797, 9648014150, 9678...
```

To add an additional new variable called gdp_ratio equal to the gdp divided by the gdp of the United States in 2007.

```
USAGDP <- filter(gapminder,year==2007 & country=="United States")
gapminder$gdpratio <- gapminder$gdp / USAGDP$gdp
```

Find the mean gdp_ratio by continent and year, and then plot the mean gdp_ratio over time, distinguishing the continents. Please use both points and lines for the plot.

```
by_continent_year <- group_by(gapminder, continent, year)
xy<- (summarize(by_continent_year, gdpratiomean = mean(gdpratio, na.rm = TRUE)))
ggplot(data= xy, aes(year, xy$gdpratiomean)) + geom_point(aes(year, xy$gdpratiomean, colour = co
ntinent)) + geom_line(aes(year, xy$gdpratiomean, colour = continent))
```