# LFS 4 Haitie Liu

# Scrape the XML page for name, zipcode and city council district. (Use either the XML or rvest package.)

Scrapped Data from ULR and made a data frame called " datafram_RB"

```
####Scrap Data From Url#################


data = read_html("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml")
data_nameR = html_nodes(data,"name")
data_zipcodeR = html_nodes(data, "zipcode")
data_cityR = html_nodes(data, "neighborhood")
data_councildistrictR = html_nodes(data,"councildistrict")

data_name= stri_sub(data_nameR,7,-8)
head(data_name)

data_zipcode = stri_sub(data_zipcodeR,10,-11)
head(data_zipcode)

data_city=stri_sub( data_cityR, 15,-16)
head(data_city)

data_councildistrict=stri_sub(data_councildistrictR, 18,-19)
head(data_councildistrict)

######Make Data Frame#############

dataframe_RB=data.frame(data_name,data_city,data_zipcode,data_councildistrict)
head(dataframe,n=1000)
dataframe_RB
```

Are there any Sushi restaurants in Baltimore? (Where the dataset is from.)
If so, can you estimate how many?
Filter the data frame for just downtown restaurants (Council District 11).

```
####There are 9 SUSHI restaurant in total#####
sushi=sum(grepl("SUSHI",dataframe_RB$data_name))
sushi


###############There are 277 restaurant in the downtown district
dataframe_RB %>%
  filter(data_councildistrict == 11) %>%
  count()

#################THERE ARE 1 SUSHI RESTAURANT IN DOWNTOWN AT INNER HARBOR#########################
SUSHIDOWNTOWN=dataframe_RB %>%
  filter(data_councildistrict == 11)

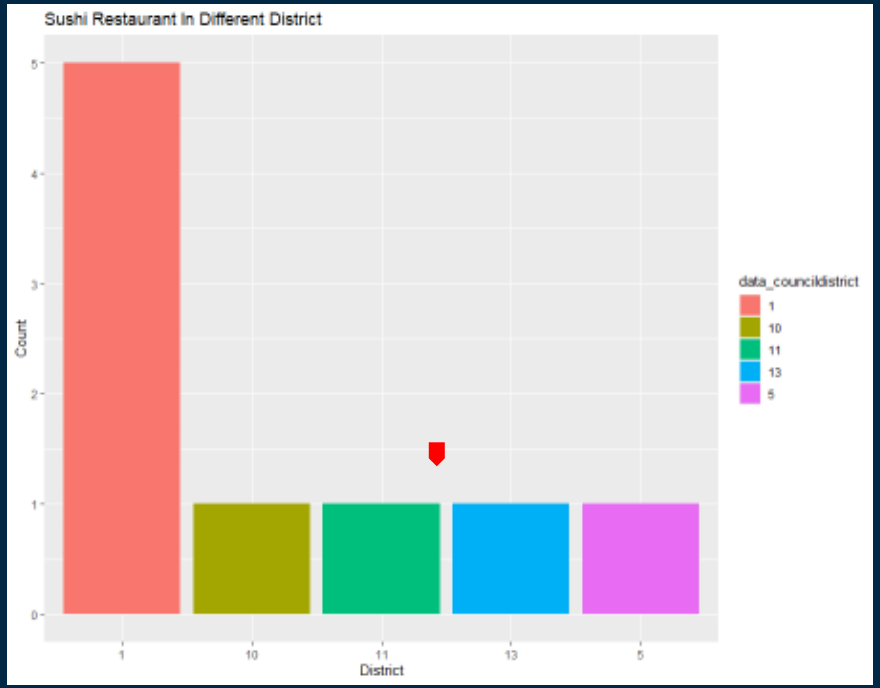
grep("SUSHI",SUSHIDOWNTOWN$data_name)
```

I used grepl function to search for SUSHI, in the data frame.

Then filter district == 11 to find out all restaurant in the downtown district.

Used grep again to find out there is 1 restaurant in downtown.

Make a barplot of the estimated number of restaurants (Sushi or otherwise) in each council.




```
#########################bar Plot#################################

grep("SUSHI",dataframe_RB$data_name)

DF=dataframe_RB[c(17,90,249,250,391,457,537,725,1137),]

DF %>%
  ggplot(aes(x=data_councildistrict,fill=data_councildistrict)) +
  geom_bar()+
  ggtitle("Sushi Restaurant In Different District")+xlab("District")+ylab("Count")
```

Using ggplot to plot all SUSHI restaurants in different district

# WDI DATA RESEARCH

```
################################

library(tidyverse)
library(GGally)
library(dplyr)
library(ggplot2)
install.packages("WDI")
library(WDI)

results = as.data.frame(WDIsearch("Population, total"))
results
```

```
```{r}
Polulation=WDI(indicator='SP.POP.TOTL', country="all", start=1960, end= NULL)
```

Columns of interest : " GDP" , "Country", "Population" ,"Education"

-Installed package WDI and load library (WDI)

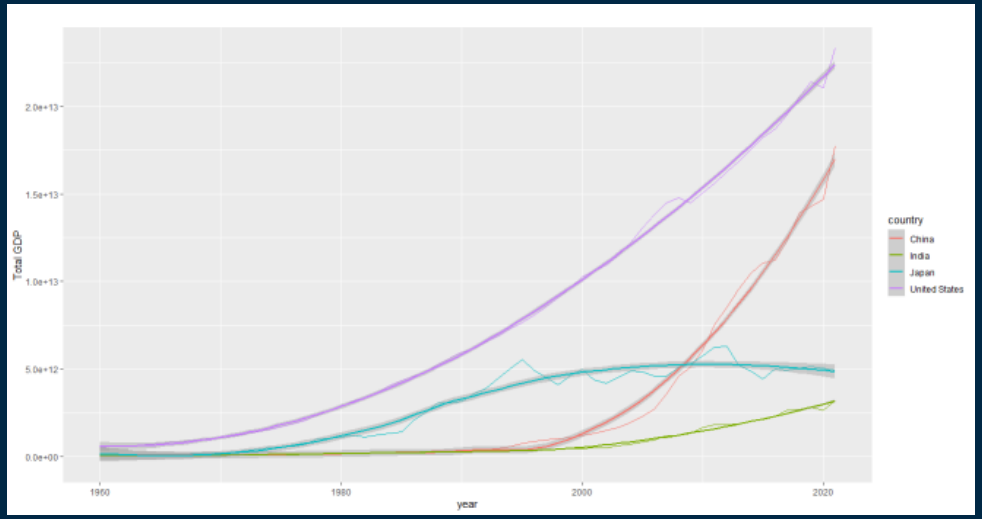
- Using the function WDIsearch() function to look for key words of the data I wanted to search

- In this case I searched Total GDP of the world. Use as.data.frame to frame the data I just downloaded

- Once I found out the indicator, use the indicator to create a dataset from year 1960 to current
```

# EDA

## Looking deeper into the GDP by countries

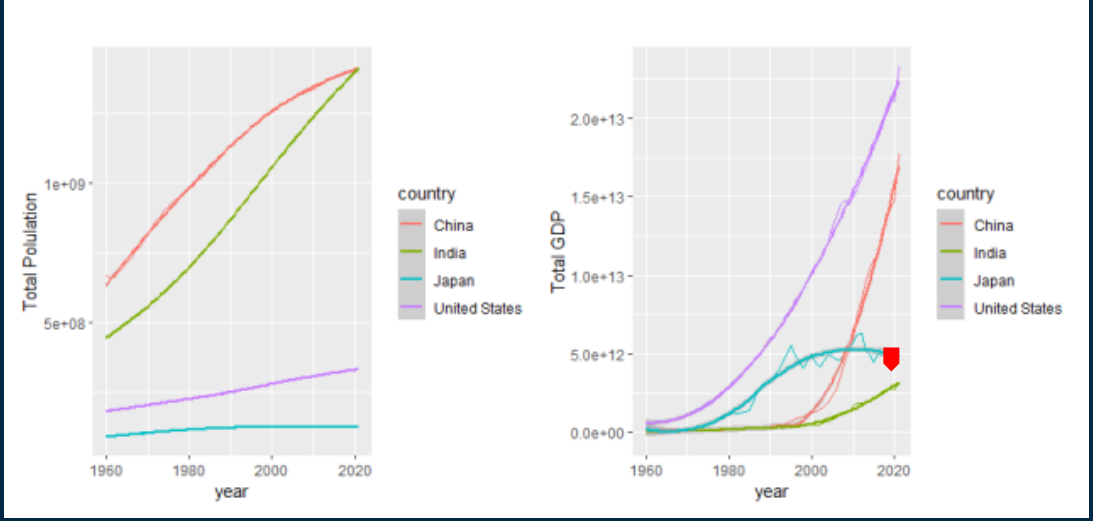



```
dat = WDI(indicator='NY.GDP.MKTP.CD', country=c('US',"JP","CN","IN"), start=1960, end=NULL)
Plot2=dat %>%
  ggplot(aes(x = year, y = NY.GDP.MKTP.CD, color=country)) + geom_line() + geom_smooth() + ggtitle ("") +ylab("Total GDP")
```

- We realize that while US, China, India has positive GDP growth over the years, Japan has turned into negative GDP growth.
-In terms of slope of growth, China obviously has the steepest slope, I wonder if it has anything to do with population.

# EDA

## Looking into total Population vs Total GDP

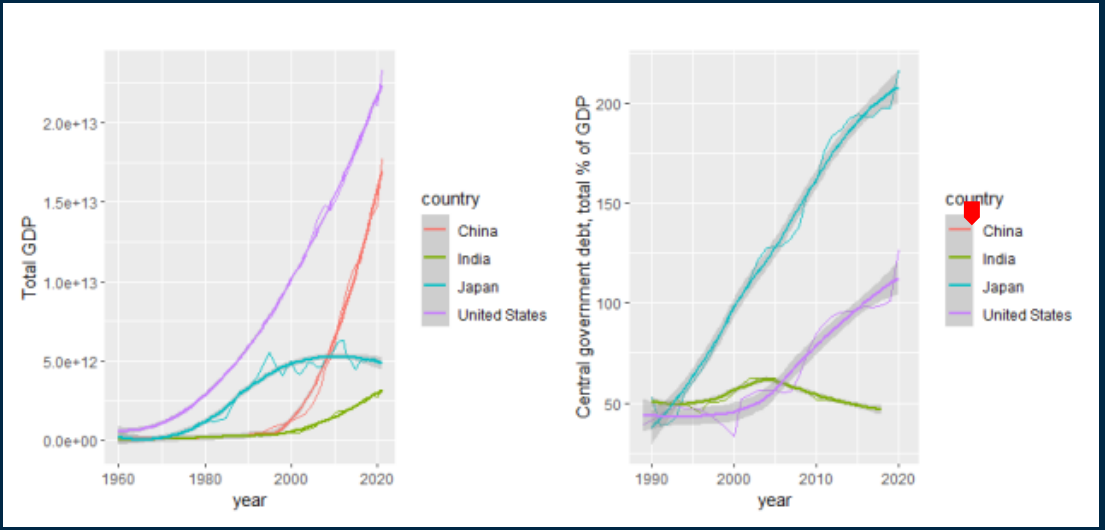



```
Polulation=WDI(indicator='SP.POP.TOTL', country=c('US',"JP","CN","IN"), start=1960, end= NULL)
Polulation = data.frame(Polulation)
Plot1=Polulation %>%
  ggplot(aes(x=year, y=SP.POP.TOTL, color = country )) + geom_line() + geom_smooth() + ggtitle ("") +ylab("Total Polulation")
```

- We realize that all United States, China, India has positive population growth, which could be a factor for positive GDP growth.
-However, Japan's population has been on a decline, which could be a factor for its negative GDP growth.

# EDA

## Looking in to GDP vs. % of Debt in central government



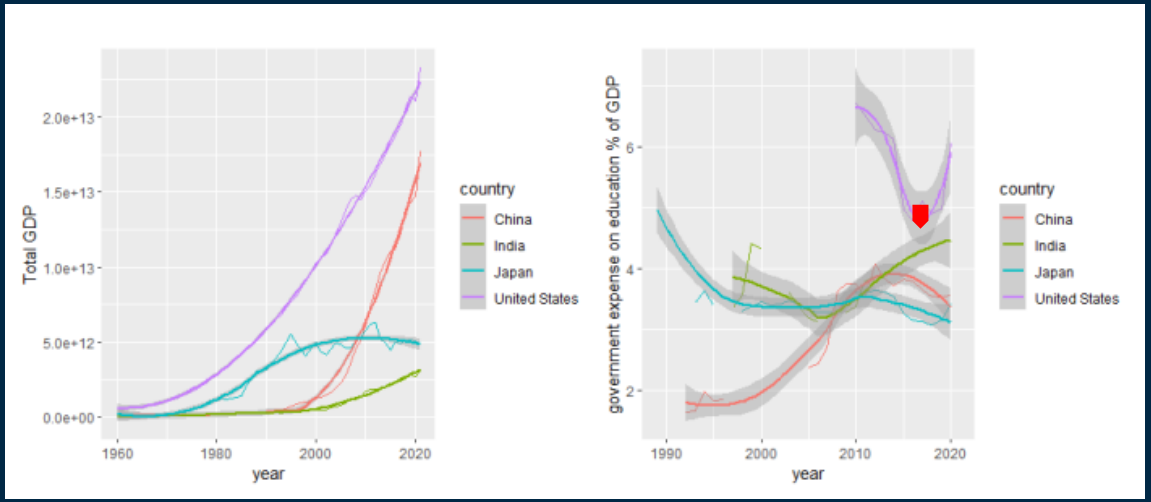

```
##########central government debt #############

centraldebt=WDI(indicator='GC.DOD.TOTL.GD.ZS', country=c('US',"JP","CN","IN"), start=1989, end= NULL)
centraldebt = data.frame(centraldebt)
Plot3= centraldebt %>%
  ggplot(aes(x=year, y=GC.DOD.TOTL.GD.ZS, color = country )) + geom_line() + geom_smooth() +
  ggtitle ("") +ylab("Central government debt, total % of GDP")
```

-Looking at chart on the right, we realize that Japan's Debt to GDP ratio is over 200%, United States has debt to GDP ratio at 100%
-This could contribute to the native GDP growth rate in these countries
-India has moderate debt to GDP level, China's data is absent.

# EDA

## Looking in to GDP vs. % of GDP spent on education




```
##################government expense on education

govexpense=WDI(indicator='SE.XPD.TOTL.GD.ZS', country=c("US","JP","CN","IN"), start=1989, end= NULL)

Plot4= govexpense %>%
  ggplot(aes(x=year, y=SE.XPD.TOTL.GD.ZS, color = country )) + geom_line() + geom_smooth() + ggtitle ("") +ylab("government expense on educatio
```

-Looking at chart on the right, we realize since start of 1990s, Japan has spent less on education in % of GDP, and currently at the lowest in 2021.

-While China and India have been constantly spending more over the years. United States fluctuating.

-We can conclude that there might be a factor in how much a country is spending in education in percent of the country's total GDP.

# TAKE AWAY

-Running into problems in authenticating some of the API functions.

-Running into problems in gathering various data, it seems data needs to be collected in different sources rather than just using a single API