DATA606_Project_Proposal_RJM

RJM

2020-02-13

Data Preparation

The following is the code for the data preparation:

```
library(flexdashboard)
library(DT)
library(ggplot2)
library(readxl)
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(tidyr)
library(stringr)
```

```
rm(list = ls())

#disable scientific notation, so that actual decimal values are imported instead of exponential factors
options(scipen = 999)


# Importing country Metadta dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/GDP.xls", "GDP.xls
country_metadata_dataset <- read_excel("GDP.xls", col_names = TRUE, sheet = "Metadata - Countries")

# Importing GDP (1995-2018) by country dataset into R
gdp_dataset <- read_excel("GDP.xls", col_names = TRUE, sheet = "Data", skip = 3) %>%
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing GDP percapita (1995-2018) by country dataset into R
```

```r
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/GDP%20per%20Capita
gdp_percapita_dataset <- read_excel("GDP_per_Capita.xls", col_names = TRUE, sheet = "Data", skip = 3) %
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing Manufacturing GDP (1995-2018) percentage by country dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/Manufacturing.xls"
gdp_manufacturing_dataset <- read_excel("Manufacturing.xls", col_names = TRUE, sheet = "Data", skip = 3
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing Agriculture GDP (1995-2018) percentage by country dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/Agriculture.xls",
gdp_agriculture_dataset <- read_excel("Agriculture.xls", col_names = TRUE, sheet = "Data", skip = 3) %>
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing Service GDP (1995-2018) percentage by country dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/Service.xls", "Serv
gdp_service_dataset <- read_excel("Service.xls", col_names = TRUE, sheet = "Data", skip = 3) %>%
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing Industries GDP (1995-2018) percentage by country dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/Industries.xls", "
gdp_industries_dataset <- read_excel("Industries.xls", col_names = TRUE, sheet = "Data", skip = 3) %>%
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)

# Importing Ores_Metals_Minerals GDP (1995-2018) percentage by country dataset into R
download.file("https://github.com/rjmirza/DATA-606/raw/master/final_project/datasets/Ores_Metals_Mineral
gdp_ores_metals_minerals_dataset <- read_excel("Ores_Metals_Minerals.xls", col_names = TRUE, sheet = "Da
  data.frame(., stringsAsFactors = F) %>%
  select(., 1,2,3,40:63)
```

```r
df1 <- gather(gdp_dataset, "year", "GDP", 4:27) %>% select(1, 4, 5)
df1$GDP <- df1$GDP/1000000
df2 <- gather(gdp_percapita_dataset, "year", "GDP Percapita", 4:27) %>% select(1, 4, 5)
df3 <- gather(gdp_industries_dataset, "year", "Industry Percent of GDP", 4:27) %>% select(1, 4, 5)
df4 <- gather(gdp_service_dataset, "year", "Services Percent of GDP", 4:27) %>% select(1, 4, 5)
df5 <- gather(gdp_agriculture_dataset, "year", "Agriculture Percent of GDP", 4:27) %>% select(1, 4, 5)
df6 <- gather(gdp_manufacturing_dataset, "year", "Manufacturing Percent of GDP", 4:27) %>% select(1, 4,
df7 <- gather(gdp_ores_metals_minerals_dataset, "year", "Ores_Metals_Minerals Percent of GDP", 4:27) %>
df <- merge(df1, df2, all.y = T)
df <- merge(df, df3, all.y = T)
df <- merge(df, df4, all.y = T)
df <- merge(df, df5, all.y = T)
df <- merge(df, df6, all.y = T)
df <- merge(df, df7, all.y = T)
df <- merge(country_metadata_dataset, df, by.x = "TableName", by.y = "Country.Name", all.y = T)
summary(df)
```

```
##    TableName         Country Code          Region          IncomeGroup
```

```
##  Length:6336        Length:6336        Length:6336        Length:6336
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  SpecialNotes           year               GDP           GDP Percapita
##  Length:6336        Length:6336        Min.   :      11   Min.   :   278.4
##  Class :character   Class :character   1st Qu.:    4581   1st Qu.:  3053.2
##  Mode  :character   Mode  :character   Median :   29549   Median :  8261.1
##                                        Mean   : 1764054   Mean   : 14919.8
##                                        3rd Qu.:  368923   3rd Qu.: 20406.8
##                                        Max.   :85804391   Max.   :139962.3
##                                        NA's   :423        NA's   :679
##  Industry Percent of GDP Services Percent of GDP Agriculture Percent of GDP
##  Min.   : 0.0034         Min.   : 9.727          Min.   : 0.0249
##  1st Qu.:19.8557         1st Qu.:44.972          1st Qu.: 3.1915
##  Median :25.7390         Median :52.958          Median : 8.4866
##  Mean   :27.3662         Mean   :53.203          Mean   :12.1676
##  3rd Qu.:32.3540         3rd Qu.:61.439          3rd Qu.:18.4360
##  Max.   :87.7969         Max.   :96.465          Max.   :79.0424
##  NA's   :871             NA's   :1174            NA's   :832
##  Manufacturing Percent of GDP Ores_Metals_Minerals Percent of GDP
##  Min.   :  0.000              Min.   : 0.000
##  1st Qu.:  8.126              1st Qu.: 1.153
##  Median : 12.695              Median : 3.018
##  Mean   : 13.188              Mean   : 7.339
##  3rd Qu.: 16.725              3rd Qu.: 6.633
##  Max.   :191.998              Max.   :86.540
##  NA's   :1156                 NA's   :1717
```

```
# removing characters from the year and converting the type to numeric
df$year <- str_extract(df$year, "[:digit:]+") %>%
  as.numeric(df$year)
```

```
incomegroup_df <- df %>%
  filter(., is.na(IncomeGroup)) %>%
  filter(., `Country Code` %in% c("EAR","FCS","HIC","HPC","LDC","LIC","LMC","LMY","LTE","MIC","PRE","PS
  arrange(TableName, year)
```

```
economy_by_region_df <- df %>%
  filter(., is.na(IncomeGroup)) %>%
  filter(., TableName %in% c("East Asia & Pacific","Europe & Central Asia","Latin America & Caribbean",
  select(1,3,4,6:8,10:13) %>%
  arrange(TableName, year)
```

Input Datasets and import

- GDP(GDP.xls) : GDP by country
- GDP Percapita(GDP_per_Capita.xls) : GDP Percapita by country

- Manufacturing GDP(Manufacturing.xls) : Manufacturing GDP percentage by country
- Agriculture GDP(Agriculture.xls) : Agriculture GDP percentage by country
- Service GDP(Service.xls) : Service GDP percentage by country
- Industries GDP(Industries.xls) : Industries GDP percentage by country
- Ores_Metals_Minerals GDP(Ores_Metals_Minerals.xls) : Ores_Metals_Minerals GDP percentage by country

Research question

The question chosen to answer was to figure out whether there is a relationship between a country's wealth and the primary type of industry that its economy is mainly engaged in. In other words, the aim was to find the impact of main industries in particular countries or regions on the economy and general well-being of its inhabitants. To approach it from a research angle, we used the Worldbank's datasets to find the main industries in the countries or regions. The two main industry categories we used were knowledge-based and traditional. The knowledge-based industries was further segregated into manufacturing and services while agriculture and minerals were condidered to be sub-categories for traditional industries.

```
knowledge_traditinoal_dF <- df %>%
  filter(., !is.na(IncomeGroup)) %>%
  select(1,3,4,6,7,10:13) %>%
  mutate("Knowledge based Percent of GDP" = ifelse(is.na(`Services Percent of GDP`), 0, `Services Percer
           ifelse(is.na(`Manufacturing Percent of GDP`),0,`Manufacturing Percent of GDP`),
         "Traditinoal based Percent of GDP" = ifelse(is.na(`Agriculture Percent of GDP`),0,`Agriculture
           ifelse(is.na(`Ores_Metals_Minerals Percent of GDP`),0,`Ores_Metals_Minerals Percent of GDP`)
  arrange(TableName, year)

country_gdp_mean_sd_dF <- knowledge_traditinoal_dF %>%
  group_by(TableName) %>%
  summarise("Country Mean GDP" = mean(GDP, na.rm=TRUE),
            "Country SD GDP" = sd(GDP, na.rm=TRUE)
            )

world_knowledge_gdp_percent_mean_dF <- knowledge_traditinoal_dF %>%
  group_by(year) %>%
  summarise("World Mean Knowledge GDP percent" = mean(`Knowledge based Percent of GDP`, na.rm=TRUE))

world_traditional_gdp_percent_mean_dF <- knowledge_traditinoal_dF %>%
  group_by(year) %>%
  summarise("World Mean Traditinoal GDP percent" = mean(`Traditinoal based Percent of GDP`, na.rm=TRUE)

knowledge_traditinoal_dF <- knowledge_traditinoal_dF %>%
  merge(., country_gdp_mean_sd_dF, by.x = "TableName", by.y = "TableName", all.y = T) %>%
  mutate("Country SD GDP in percent" = `Country SD GDP`/`Country Mean GDP`*100) %>%
  merge(., world_knowledge_gdp_percent_mean_dF, by.x = "year", by.y = "year", all.y = T) %>%
  merge(., world_traditional_gdp_percent_mean_dF, by.x = "year", by.y = "year", all.y = T) %>%
  select(1:5,10:16) %>%
  na_if(., 0) %>%
  arrange(TableName, year)
```

Cases

There were 263 cases representing each country and independent territory. The details on these cases form the bulk of data.

Data collection

We thought that the industry composition of GDP is an indicator that will tell us about the main industries operating in a certain country. We used Worldbank's website to collect data on the four catrgories namely agriculture, minerals, services, and manufacturing. The data had to be curated to fit it between the years of 1995-2018 as most of the countries had missing information for years before 1995. We had to nuance the data analysis by introducing factors like recessions, per capita income, and regions.

```
### initial dataframe
DT::datatable(df, options = list(pageLength = 5))
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

```
### incomegroup dataframe
DT::datatable(incomegroup_df, options = list(pageLength = 5))
```

```
### economy by region dataframe
DT::datatable(economy_by_region_df, options = list(pageLength = 5))
```

```
### knowledge and traditinoal GDP's dataframe
DT::datatable(knowledge_traditinoal_dF, options = list(pageLength = 5))
```

Type of study

This is an observational study where we took existing data to support our hypothesis.

Data Source

The sources are as follows:

1. GDP

https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

2. GDP per Capita

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

3. Agricultrure, forestry, and fishing as % of GDP

https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS

4. Ores and Metals exports (% of merchandise exports) taken as a proxy to mineral production and exports

https://data.worldbank.org/indicator/TX.VAL.MMTL.ZS.UN

5. Service, value added (% of GDP)

https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS

6. Manufacturing, value added (% of GDP)

https://data.worldbank.org/indicator/NV.IND.MANF.ZS

Dependent Variable

The response variable is the GDP output which is quantitative measured in either percentage points or dollar amounts.

Independent Variable

The qualitative independent variables are the countries, regions, and types of main industries these places are engaged in. The quantitative independent variable is the years to measure the changes in the GDP output.

Relevant summary statistics

```
NIG <- length(unique(incomegroup_df[["TableName"]]))

valueBox(NIG, color = "primary")
```

13

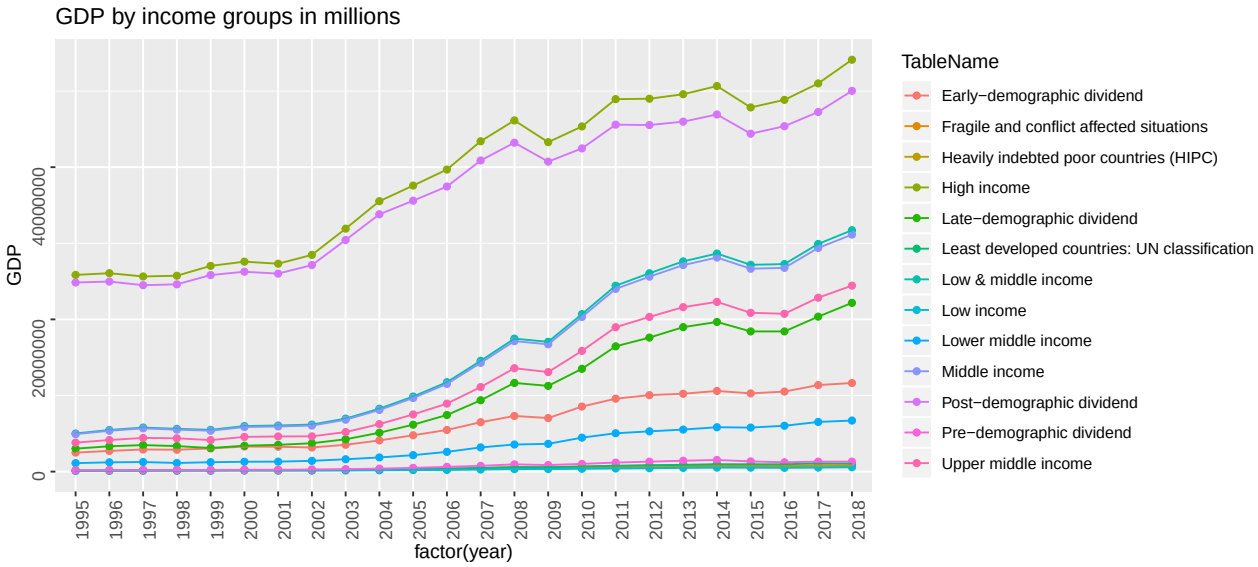
GDP in millions by income groups

There were some challenges with imputing the missing data. We figured that leaving them blank is the best to get the continuous lines on graphs.

```
ggplot(incomegroup_df, aes(x=factor(year), colour=TableName, group = TableName)) +
  geom_point(aes(y = `GDP`)) +
  geom_line(aes(y = `GDP`)) +
  theme(axis.text.x = element_text(size=10, angle=90)) +
  theme(axis.text.y = element_text(size=10, angle=90)) +
  labs(title = "GDP by income groups in millions")
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

```
## Warning: Removed 5 rows containing missing values (geom_path).
```



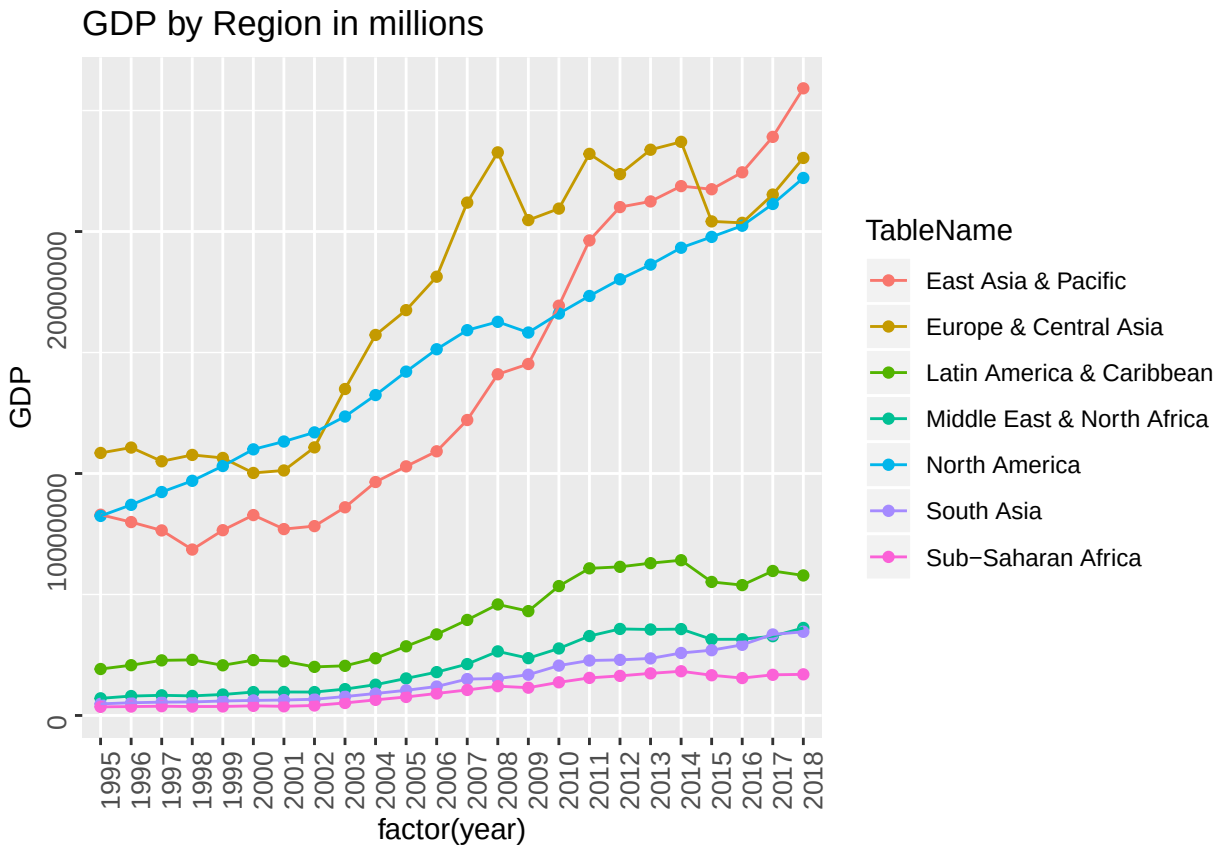
For more information on demographic dividend, please refer to the following link:

https://www.imf.org/external/pubs/ft/fandd/2006/09/basics.htm

GDP in million based on specific geological regions

The regions helped us to gain insights through pre-existing knowledge regarding the wealth and the main industries they indulge into. For example, we know that Europe is a Service and Manufacturing intensive economy and is lumped with Central Asia which is a Minerals dependent industry. Data could be refined further to separate these regions and find out how much of the GDP is dependent on the specific industries!

```
library(ggplot2)
ggplot(economy_by_region_df, aes(x=factor(year), colour=TableName, group = TableName)) +
  geom_point(aes(y = `GDP`)) +
  geom_line(aes(y = `GDP`)) +
  theme(axis.text.x = element_text(size=10, angle=90)) +
  theme(axis.text.y = element_text(size=10, angle=90)) +
  labs(title = "GDP by Region in millions")
```



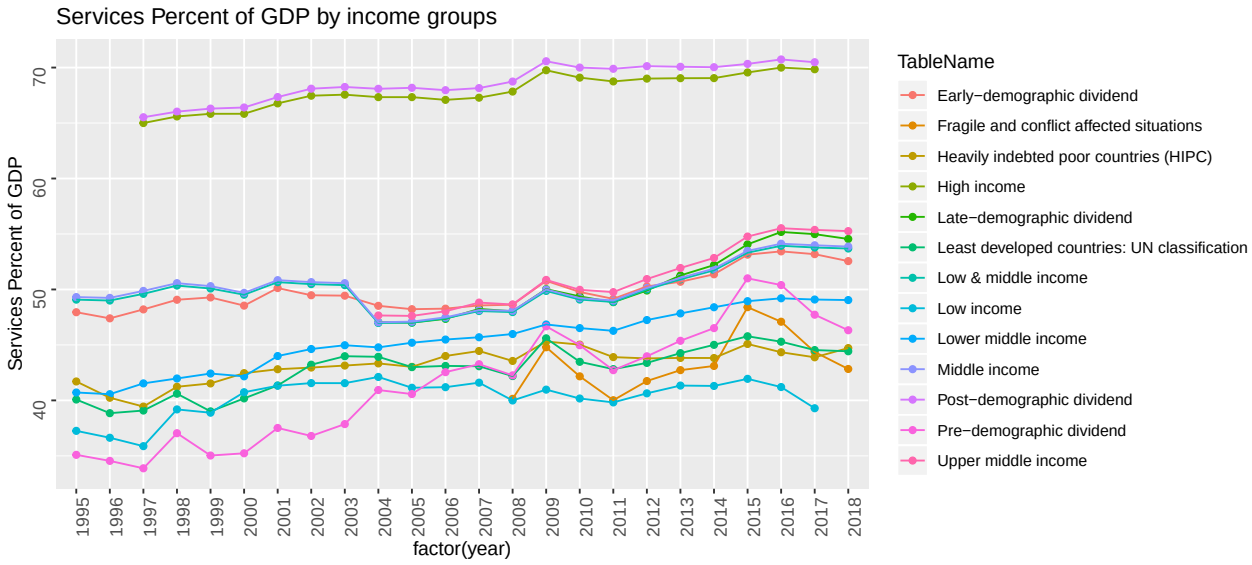
Services Percent of GDP

Services make a high portion of GDP in High income countries.

```
ggplot(incomegroup_df, aes(x=factor(year), colour=TableName, group = TableName)) +
  geom_point(aes(y = `Services Percent of GDP`)) +
  geom_line(aes(y = `Services Percent of GDP`)) +
  theme(axis.text.x = element_text(size=10, angle=90)) +
  theme(axis.text.y = element_text(size=10, angle=90)) +
  labs(title = "Services Percent of GDP by income groups")
```

```
## Warning: Removed 38 rows containing missing values (geom_point).
```

```
## Warning: Removed 38 rows containing missing values (geom_path).
```



Agriculture Percent of GDP

The Agriculture seems to be trending downward in every type of economy. An interesting discovery would be to see if it is losing share to other industries or it is declining as an industry!

```
ggplot(incomegroup_df, aes(x=year, colour=TableName, group = TableName)) +
  geom_point(aes(y = `Agriculture Percent of GDP`)) +
  geom_line(aes(y = `Agriculture Percent of GDP`)) +
  theme(axis.text.x = element_text(size=10, angle=90)) +
  theme(axis.text.y = element_text(size=10, angle=90)) +
  labs(title = "Agriculture Percent of GDP by income groups")
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```