stat-471-final-proj

Ethan

12/2/2021

```
# a-z = 26 + aa-az = 26 + ba-bn = 14, sum = 66
# country name, country code, indicator name, indicator code, 1960-2020, empty final col
wdi_raw <-
  read_csv("/Users/ethan/Documents/R/stat-471-final-project/WDI_csv/WDIData.csv",
           col_names = TRUE)
```

```
## New names:
## * `` -> ...66
```

```
## Rows: 383838 Columns: 66
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country Name, Country Code, Indicator Name, Indicator Code
## dbl (61): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
## lgl  (1): ...66
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# for country code, (country) table name, region, income group, latest industrial data (year), latest t
wdi_country <-
  read_csv("/Users/ethan/Documents/R/stat-471-final-project/WDI_csv/WDICountry.csv",
           col_names = TRUE)
```

```
## New names:
## * `` -> ...31
```

```
## Rows: 265 Columns: 31
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (26): Country Code, Short Name, Table Name, Long Name, 2-alpha code, Cur...
## dbl  (3): National accounts reference year, Latest industrial data, Latest t...
## lgl  (2): PPP survey year, ...31
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# for series code, topic, indicator name, short definition, long definition, periodicity, aggregation m
wdi_series <-
  read_csv("/Users/ethan/Documents/R/stat-471-final-project/WDI_csv/WDISeries.csv",
           col_names = TRUE)
```

```
## New names:
```

```
## * `` -> ...21
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## Rows: 1443 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): Series Code, Topic, Indicator Name, Short definition, Long definit...
## lgl  (4): Unit of measure, Related source links, Other web links, ...21
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# integrate trade data
```

Part 1: Economic Structure and Income per Capita (EDA)

```
# this part is almost an EDA

# let's create a table showing average shares by income group over time
holy_trinity <- wdi_raw %>%
  filter(`Indicator Code` %in% c("SL.SRV.EMPL.ZS",
                                 "SL.IND.EMPL.ZS",
                                 "SL.AGR.EMPL.ZS"))

summary_trinity_time_income_group <- wdi_country %>%
  filter(!(is.na(`Currency Unit`))) %>% # remove regional groupings
  select(`Short Name`, `Income Group`) %>%
  inner_join(holy_trinity, by = c("Short Name" = "Country Name")) %>%
  group_by(`Short Name`, `Indicator Code`, `Income Group`) %>%
  summarise(`90s` = mean(`1991`,`1992`,`1993`,`1994`,`1995`,`1996`,
                         `1997`,`1998`,`1999`, na.rm=TRUE),
            `00s` = mean(`2000`,`2001`,`2002`,`2003`,`2004`,`2005`,`2006`,`2007`,
                         `2008`,`2009`, na.rm=TRUE),
            `10s` = mean(`2010`,`2011`,`2012`,`2013`,`2014`,`2015`,`2016`,
                         `2017`,`2018`,`2019`, na.rm=TRUE))
```

```
## `summarise()` has grouped output by 'Short Name', 'Indicator Code'. You can override using the `.grou
```

```
summary_trinity_time_income_group_90s <-
  summary_trinity_time_income_group %>%
  select(`Indicator Code`, `Income Group`, `90s`) %>%
  pivot_wider(names_from = `Indicator Code`, values_from = `90s`) %>%
  transmute(`Income Group` = as.factor(`Income Group`),
            `Agriculture` = SL.AGR.EMPL.ZS, # change variable names
            `Industry` = SL.IND.EMPL.ZS,
            `Services` = SL.SRV.EMPL.ZS) %>%
  pivot_longer(cols = c(`Agriculture`,
                        `Industry`,
                        `Services`),
               values_to = "value",
               names_to = "Sector") %>%
  group_by(`Income Group`, Sector) %>%
  summarise(value = mean(value, na.rm =TRUE)) %>% # calculate average share per sector
```

```
  ungroup() %>%
  mutate(income_group_order = case_when( # order the income groups
    `Income Group` == "Low income" ~ 1,
    `Income Group` == "Lower middle income" ~ 2,
    `Income Group` == "Upper middle income" ~ 3,
    `Income Group` == "High income" ~ 4
  ))
```

```
## Adding missing grouping variables: `Short Name`
```

```
## `summarise()` has grouped output by 'Income Group'. You can override using the `.groups` argument.
```

```
p1 <- summary_trinity_time_income_group_90s %>%
  ggplot(aes(x = Sector, y = value, fill =
               fct_reorder(.f = summary_trinity_time_income_group_90s$`Income Group`, .x = summary_trini
  geom_bar(position="dodge", stat="identity") +
  labs(x = "Sector", y = "Share of Employment",
       title = "1991-1999 Average Employment by Sector by Income Group") +
  guides(fill=guide_legend(title="Income Group")) +
  theme_bw()

summary_trinity_time_income_group_00s <-
  summary_trinity_time_income_group %>%
  select(`Indicator Code`, `Income Group`, `00s`) %>%
  pivot_wider(names_from = `Indicator Code`, values_from = `00s`) %>%
  transmute(`Income Group` = as.factor(`Income Group`),
            `Agriculture` = SL.AGR.EMPL.ZS, # change variable names
            `Industry` = SL.IND.EMPL.ZS,
            `Services` = SL.SRV.EMPL.ZS) %>%
  pivot_longer(cols = c(`Agriculture`,
                        `Industry`,
                        `Services`),
               values_to = "value",
               names_to = "Sector") %>%
  group_by(`Income Group`, Sector) %>%
  summarise(value = mean(value, na.rm =TRUE)) %>% # calculate average share per sector
  ungroup() %>%
  mutate(income_group_order = case_when( # order the income groups
    `Income Group` == "Low income" ~ 1,
    `Income Group` == "Lower middle income" ~ 2,
    `Income Group` == "Upper middle income" ~ 3,
    `Income Group` == "High income" ~ 4
  ))
```

```
## Adding missing grouping variables: `Short Name`
## `summarise()` has grouped output by 'Income Group'. You can override using the `.groups` argument.
```

```
p2 <- summary_trinity_time_income_group_00s %>%
  ggplot(aes(x = Sector, y = value, fill =
               fct_reorder(.f = summary_trinity_time_income_group_00s$`Income Group`, .x = summary_trini
  geom_bar(position="dodge", stat="identity") +
  labs(x = "Sector", y = "Share of Employment",
       title = "2000-2009 Average Employment by Sector by Income Group") +
  guides(fill=guide_legend(title="Income Group")) +
  theme_bw()
```

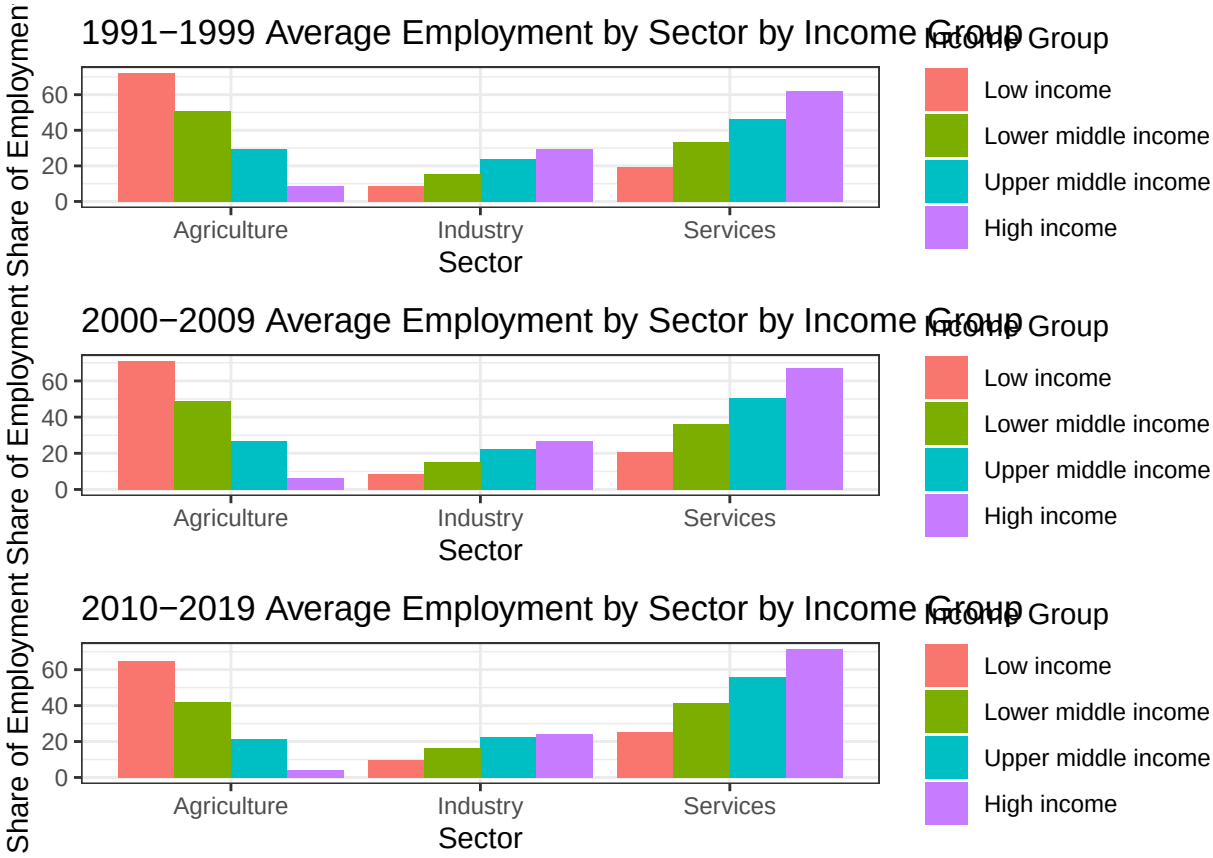
```
summary_trinity_time_income_group_10s <-
  summary_trinity_time_income_group %>%
  select(`Indicator Code`, `Income Group`, `10s`) %>%
  pivot_wider(names_from = `Indicator Code`, values_from = `10s`) %>%
  transmute(`Income Group` = as.factor(`Income Group`),
            `Agriculture` = SL.AGR.EMPL.ZS, # change variable names
            `Industry` = SL.IND.EMPL.ZS,
            `Services` = SL.SRV.EMPL.ZS) %>%
  pivot_longer(cols = c(`Agriculture`,
                        `Industry`,
                        `Services`),
               values_to = "value",
               names_to = "Sector") %>%
  group_by(`Income Group`, Sector) %>%
  summarise(value = mean(value, na.rm =TRUE)) %>% # calculate average share per sector
  ungroup() %>%
  mutate(income_group_order = case_when( # order the income groups
    `Income Group` == "Low income" ~ 1,
    `Income Group` == "Lower middle income" ~ 2,
    `Income Group` == "Upper middle income" ~ 3,
    `Income Group` == "High income" ~ 4
  ))
```

```
## Adding missing grouping variables: `Short Name`
## `summarise()` has grouped output by 'Income Group'. You can override using the `.groups` argument.
```

```
p3 <- summary_trinity_time_income_group_10s %>%
  ggplot(aes(x = Sector, y = value, fill =
               fct_reorder(.f = summary_trinity_time_income_group_10s$`Income Group`, .x = summary_trini
  geom_bar(position="dodge", stat="identity") +
  labs(x = "Sector", y = "Share of Employment",
       title = "2010-2019 Average Employment by Sector by Income Group") +
  guides(fill=guide_legend(title="Income Group")) +
  theme_bw()
```

```
# you can see that there is a clear distinction between different income groups of countries
plot_grid(p1,p2,p3,ncol=1)
```

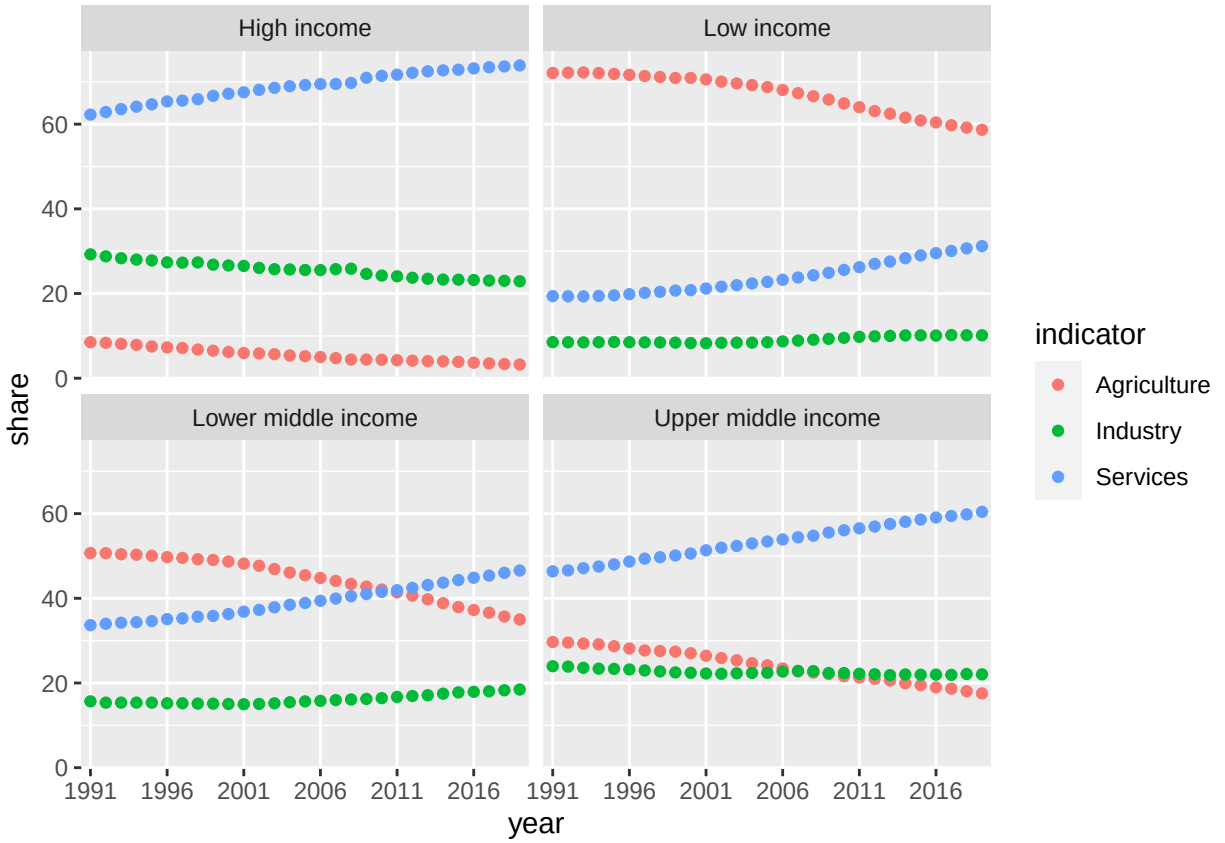
```
## Warning: Use of `summary_trinity_time_income_group_90s$`Income Group`` is
## discouraged. Use `Income Group` instead.

## Warning: Use of `summary_trinity_time_income_group_90s$income_group_order` is
## discouraged. Use `income_group_order` instead.

## Warning: Use of `summary_trinity_time_income_group_00s$`Income Group`` is
## discouraged. Use `Income Group` instead.

## Warning: Use of `summary_trinity_time_income_group_00s$income_group_order` is
## discouraged. Use `income_group_order` instead.

## Warning: Use of `summary_trinity_time_income_group_10s$`Income Group`` is
## discouraged. Use `Income Group` instead.

## Warning: Use of `summary_trinity_time_income_group_10s$income_group_order` is
## discouraged. Use `income_group_order` instead.
```




```
# lets see how the shares of each sector evolve over time
shares_change_over_time <- wdi_country %>%
  filter(!(is.na(`Currency Unit`))) %>% # remove regional groupings
  select(`Short Name`, `Income Group`) %>%
  inner_join(holy_trinity, by = c("Short Name" = "Country Name")) %>%
  pivot_longer(cols=`1991`:`2019`, names_to = "year", values_to = "share") %>%
  select(`Short Name`, `Income Group`, `Indicator Code`, year, share) %>%
  drop_na() %>%
  group_by(`Indicator Code`, year, `Income Group`) %>%
  summarise(share = mean(share)) %>%
  transmute(year = year, share = share,`Income Group`=`Income Group`,
            indicator = as.factor(case_when(
              `Indicator Code` == "SL.AGR.EMPL.ZS" ~ "Agriculture",
              `Indicator Code` == "SL.IND.EMPL.ZS" ~ "Industry",
              `Indicator Code` == "SL.SRV.EMPL.ZS" ~ "Services"
            )))
```

```
## `summarise()` has grouped output by 'Indicator Code', 'year'. You can override using the `.groups` ar
```

```
shares_change_over_time %>%
  ggplot(aes(x = year, y = share, colour=indicator)) +
  geom_point() +
  facet_wrap(~`Income Group`) +
  scale_x_discrete(breaks = seq(1991, 2019, by = 5))
```




```
high_inc_agr_1991 =  shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 1991,
         indicator == "Agriculture") %>% pull(share)
high_inc_agr_2019 = shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 2019,
         indicator == "Agriculture") %>% pull(share)
high_inc_ind_1991 = shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 1991,
         indicator == "Industry") %>% pull(share)
high_inc_ind_2019 = shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 2019,
         indicator == "Industry") %>% pull(share)
high_inc_srv_1991 = shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 1991,
         indicator == "Services") %>% pull(share)
high_inc_srv_2019 = shares_change_over_time %>%
  filter(`Income Group` == "High income",
         year == 2019,
         indicator == "Services") %>% pull(share)

highmid_inc_agr_1991 = shares_change_over_time %>%
```

```
  filter(`Income Group` == "Upper middle income",
         year == 1991,
         indicator == "Agriculture") %>% pull(share)
highmid_inc_agr_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Upper middle income",
         year == 2019,
         indicator == "Agriculture") %>% pull(share)
highmid_inc_ind_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Upper middle income",
         year == 1991,
         indicator == "Industry") %>% pull(share)
highmid_inc_ind_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Upper middle income",
         year == 2019,
         indicator == "Industry") %>% pull(share)
highmid_inc_srv_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Upper middle income",
         year == 1991,
         indicator == "Services") %>% pull(share)
highmid_inc_srv_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Upper middle income",
         year == 2019,
         indicator == "Services") %>% pull(share)

lowmid_inc_agr_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 1991,
         indicator == "Agriculture") %>% pull(share)
lowmid_inc_agr_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 2019,
         indicator == "Agriculture") %>% pull(share)
lowmid_inc_ind_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 1991,
         indicator == "Industry") %>% pull(share)
lowmid_inc_ind_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 2019,
         indicator == "Industry") %>% pull(share)
lowmid_inc_srv_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 1991,
         indicator == "Services") %>% pull(share)
lowmid_inc_srv_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Lower middle income",
         year == 2019,
         indicator == "Services") %>% pull(share)

low_inc_agr_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 1991,
         indicator == "Agriculture") %>% pull(share)
```

```
low_inc_agr_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 2019,
         indicator == "Agriculture") %>% pull(share)
low_inc_ind_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 1991,
         indicator == "Industry") %>% pull(share)
low_inc_ind_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 2019,
         indicator == "Industry") %>% pull(share)
low_inc_srv_1991 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 1991,
         indicator == "Services") %>% pull(share)
low_inc_srv_2019 = shares_change_over_time %>%
  filter(`Income Group` == "Low income",
         year == 2019,
         indicator == "Services") %>% pull(share)

# percent change over time across sectors and income groups, precipitous decline in ag, increase in ser
tibble(
  income_group = c("High income", "Upper middle income", "Lower middle income",
                   "Low income"),
  change_agriculture = c((high_inc_agr_2019 - high_inc_agr_1991)/high_inc_agr_1991,
                         (highmid_inc_agr_2019 - highmid_inc_agr_1991)/highmid_inc_agr_1991,
                         (lowmid_inc_agr_2019 - lowmid_inc_agr_1991)/lowmid_inc_agr_1991,
                         (low_inc_agr_2019 - low_inc_agr_1991)/low_inc_agr_1991),
  change_industry = c((high_inc_ind_2019 - high_inc_ind_1991)/high_inc_ind_1991,
                         (highmid_inc_ind_2019 - highmid_inc_ind_1991)/highmid_inc_ind_1991,
                         (lowmid_inc_ind_2019 - lowmid_inc_ind_1991)/lowmid_inc_ind_1991,
                         (low_inc_ind_2019 - low_inc_ind_1991)/low_inc_ind_1991),
  change_services = c((high_inc_srv_2019 - high_inc_srv_1991)/high_inc_srv_1991,
                         (highmid_inc_srv_2019 - highmid_inc_srv_1991)/highmid_inc_srv_1991,
                         (lowmid_inc_srv_2019 - lowmid_inc_srv_1991)/lowmid_inc_srv_1991,
                         (low_inc_srv_2019 - low_inc_srv_1991)/low_inc_srv_1991)
)
```

```
## # A tibble: 4 x 4
##   income_group        change_agriculture change_industry change_services
##   <chr>                            <dbl>           <dbl>           <dbl>
## 1 High income                     -0.619         -0.217           0.187
## 2 Upper middle income             -0.409         -0.0799          0.303
## 3 Lower middle income             -0.310          0.179           0.383
## 4 Low income                      -0.186          0.191           0.608
```

Part 2: Economic Structure and Income per Capita Growth

```
wdi_series %>% group_by(Topic) %>% summarise(n())
```

```
## # A tibble: 90 x 2
##    Topic                                                          `n()`
```

```
##    <chr>                                                                <int>
##  1 Economic Policy & Debt: Balance of payments: Capital & financial accou~    11
##  2 Economic Policy & Debt: Balance of payments: Current account: Balances      4
##  3 Economic Policy & Debt: Balance of payments: Current account: Goods, s~    22
##  4 Economic Policy & Debt: Balance of payments: Current account: Transfers     7
##  5 Economic Policy & Debt: Balance of payments: Reserves & other items         6
##  6 Economic Policy & Debt: External debt: Debt outstanding                    10
##  7 Economic Policy & Debt: External debt: Debt ratios & other items           11
##  8 Economic Policy & Debt: External debt: Debt service                         4
##  9 Economic Policy & Debt: External debt: Net flows                           20
## 10 Economic Policy & Debt: National accounts: Adjusted savings & income       28
## # ... with 80 more rows
```

```
wdi_series %>% group_by(`Indicator Name`) %>% summarise(n())
```

```
## # A tibble: 1,443 x 2
##    `Indicator Name`                                                     `n()`
##    <chr>                                                                <int>
##  1 Access to clean fuels and technologies for cooking  (% of population)       1
##  2 Access to electricity (% of population)                                     1
##  3 Access to electricity, rural (% of rural population)                        1
##  4 Access to electricity, urban (% of urban population)                        1
##  5 Account ownership at a financial institution or with a mobile-money-se~     1
##  6 Account ownership at a financial institution or with a mobile-money-se~     1
##  7 Account ownership at a financial institution or with a mobile-money-se~     1
##  8 Account ownership at a financial institution or with a mobile-money-se~     1
##  9 Account ownership at a financial institution or with a mobile-money-se~     1
## 10 Account ownership at a financial institution or with a mobile-money-se~     1
## # ... with 1,433 more rows
```

```
code_employent_in_services = "SL.SRV.EMPL.ZS"
"SL.IND.EMPL.ZS"
```

```
## [1] "SL.IND.EMPL.ZS"
```

```
"SL.AGR.EMPL.ZS"
```

```
## [1] "SL.AGR.EMPL.ZS"
```

Broad topics: - Economic Policy & Debt, Education, Environment, Financial Sector, Gender, Health, Infrastructure, Poverty, Private Sector & Trade, Public Sector, Social Protection & Labor, World Bank, International Debt Statistics