IRS Statistics of Income Data Visualizations

Author: Kenneth C. Wilbur, UC San Diego
Date: February 2026

DATA SOURCE
- IRS Statistics of Income, Corporation Income Tax Returns
- Table 5.1: Returns of Active Corporations by Minor Industry
- Years: 2014-2022 (9 years)

SAMPLE
- Active U.S. Corporations (C corps, S corps, REITs, RICs)
- 19 sector-level industries
- 187 subsector-level industries

RATIOS COMPUTED
- Advertising / Revenue = Advertising Deductions / Business Receipts
- Advertising / Gross Profit = Advertising / (Revenue - Cost of Goods Sold)
- Advertising / Net Income = Advertising / Net Income (less credits)
- Net Income / Gross Profit = Net Income / (Revenue - Cost of Goods Sold)
- Gross Profit Margin = (Revenue - Cost of Goods Sold) / Revenue

ADJUSTMENTS
- Dollar values adjusted to real 2022 dollars using CPI-U
- CPI-U values: 236.7 (2014) to 292.7 (2022)

IRS Statistics of Income Data Visualizations (continued)

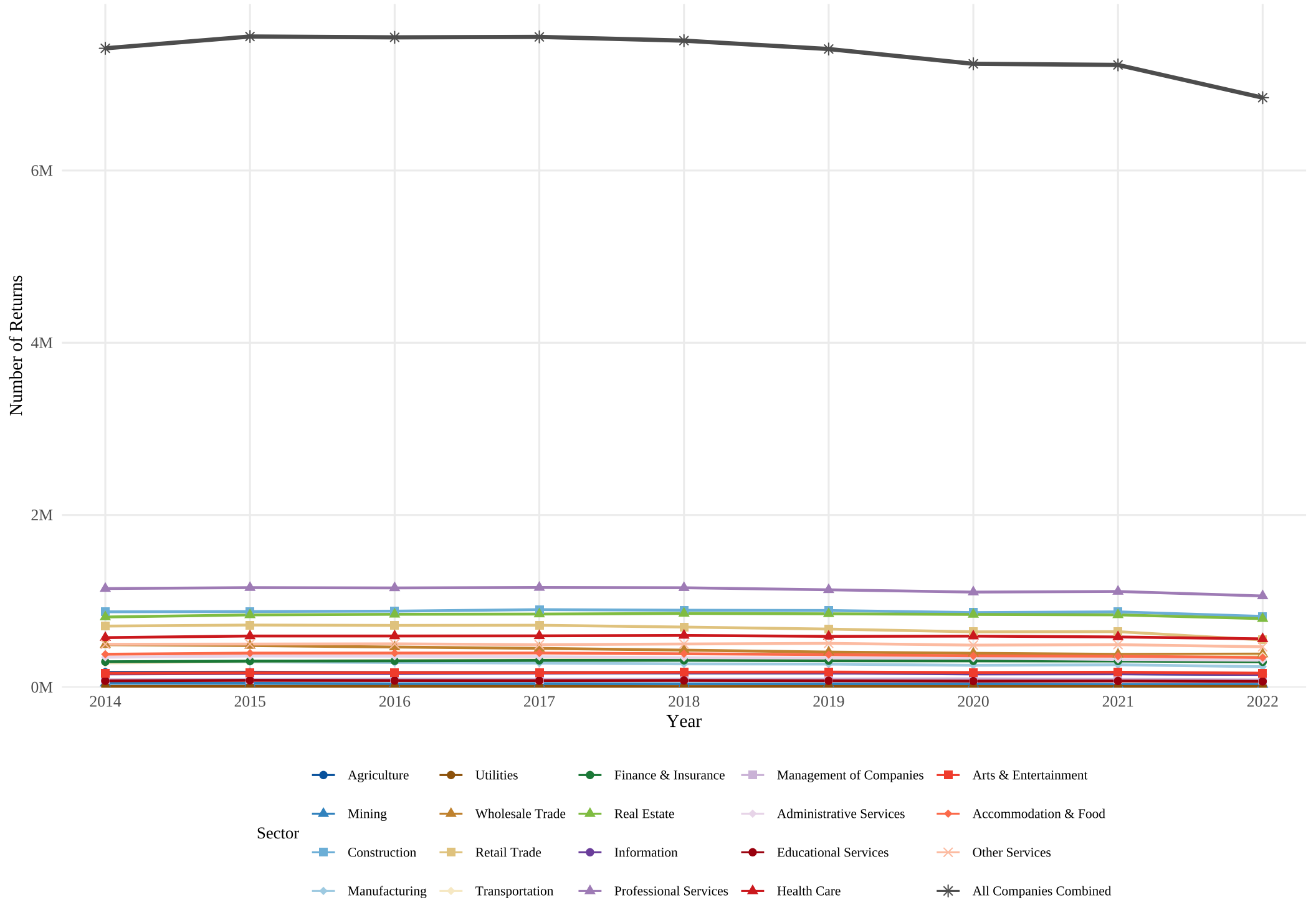
OMITTED TIME SERIES
- Page 18 (Advertising/Net Income by Sector): Excludes Educational Services, Utilities, Accommodation & Food (ratios exceeded 80% or fell below -20%)
- Page 19 (Net Income/Gross Profit by Sector): Excludes Finance & Insurance, Management of Companies, Mining (ratios exceeded 60% or fell below -10%)

SECTOR GROUPINGS
- Goods-Producing: Agriculture, Mining, Construction, Manufacturing
- Distribution & Utilities: Utilities, Wholesale, Retail, Transportation
- Finance & Real Estate: Finance & Insurance, Real Estate
- Business Services: Information, Professional Services, Management, Administrative
- Consumer Services: Educational, Health Care, Arts, Accommodation & Food, Other

STRUCTURE
- Pages 3-15: Level variables by sector (13 pages)
- Pages 16-20: Ratio variables by sector (5 pages)
- Pages 21-30: Group-sector breakouts (10 pages)
- Pages 31-48: Subsector Ad/Revenue by sector (18 pages)
- Pages 49-66: Subsector Ad/Gross Profit by sector (18 pages)

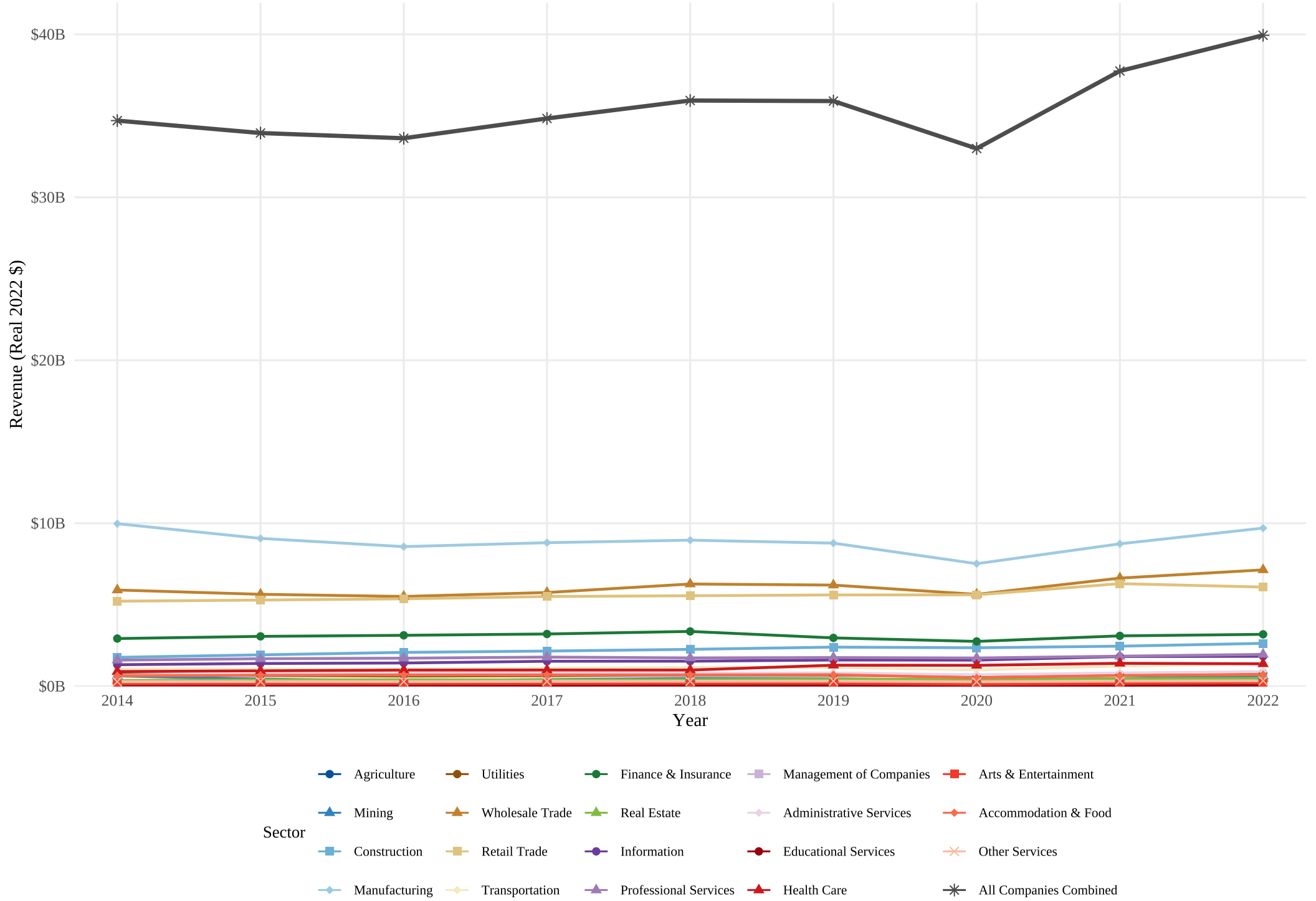
Prepared quickly using Claude Code. Accuracy is not ensured. R scripts are included as appendices to enable replication. Please contact the author in case any error is found.

Number of Tax Returns by Sector, Active U.S. Corporations



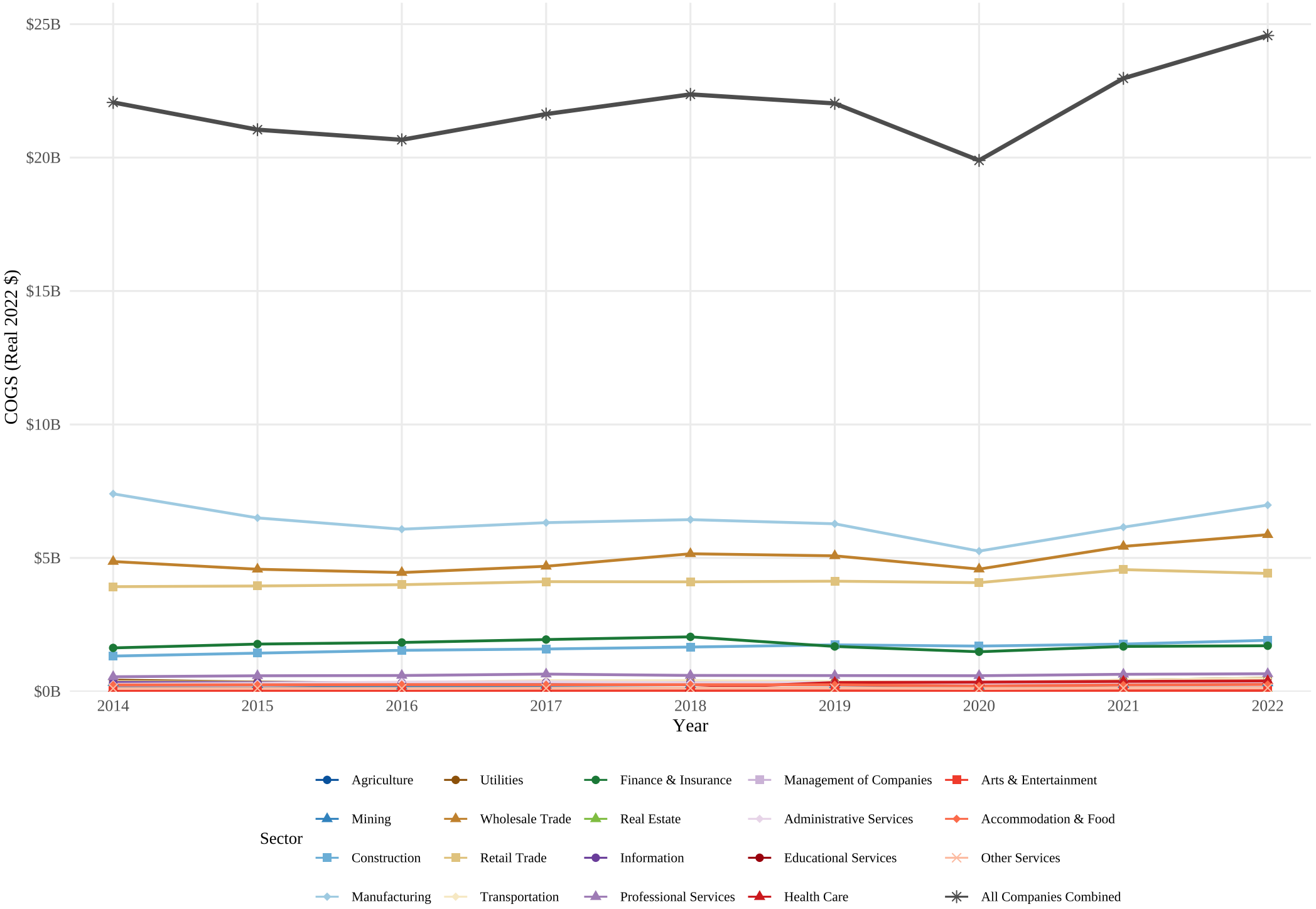
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Count of corporate tax returns

Total Revenue by Sector, Active U.S. Corporations



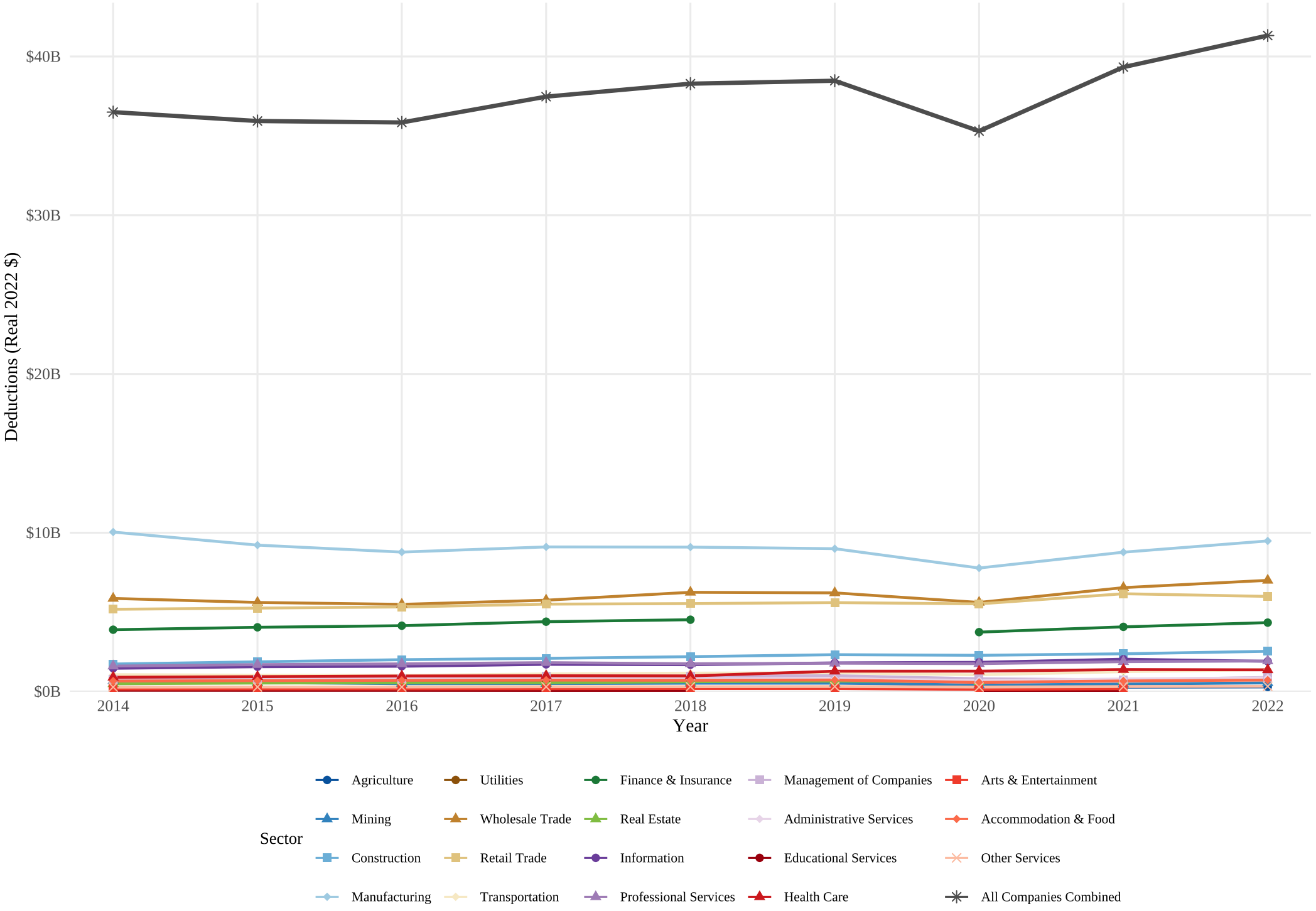
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Gross receipts from sales

Cost of Goods Sold by Sector, Active U.S. Corporations



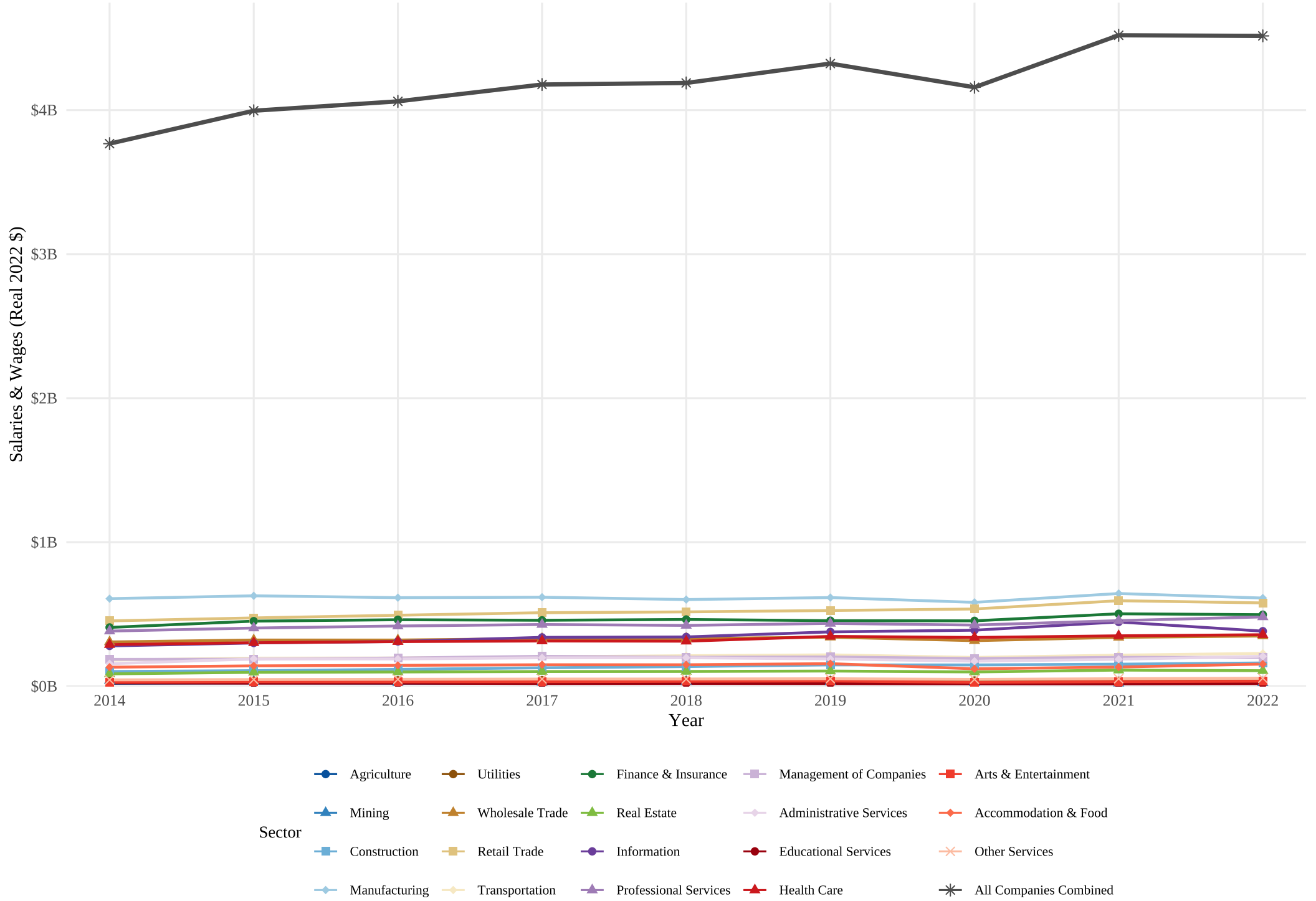
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Direct costs of producing goods/services

Total Deductions by Sector, Active U.S. Corporations



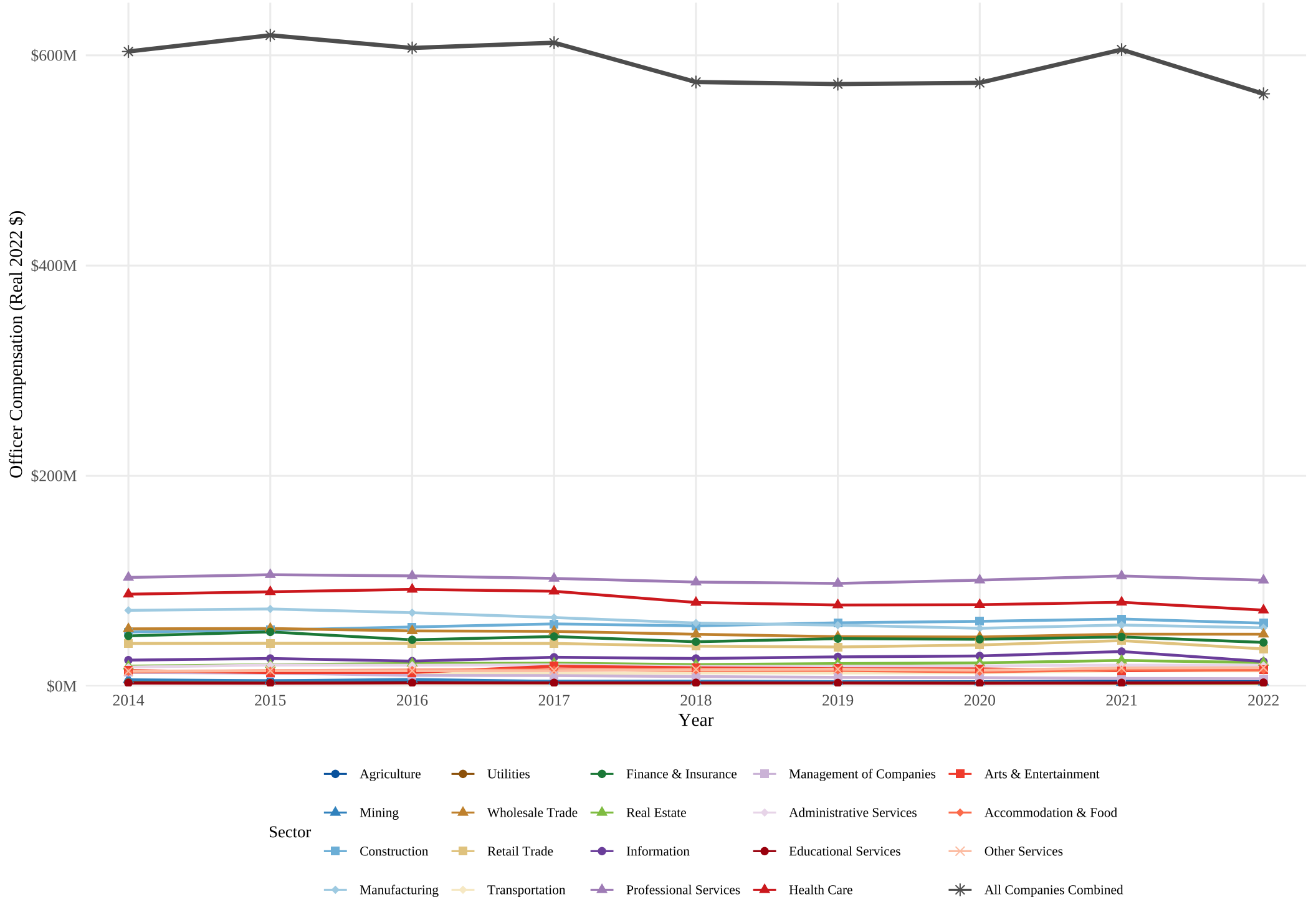
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: All expenses claimed against income

Salaries and Wages by Sector, Active U.S. Corporations



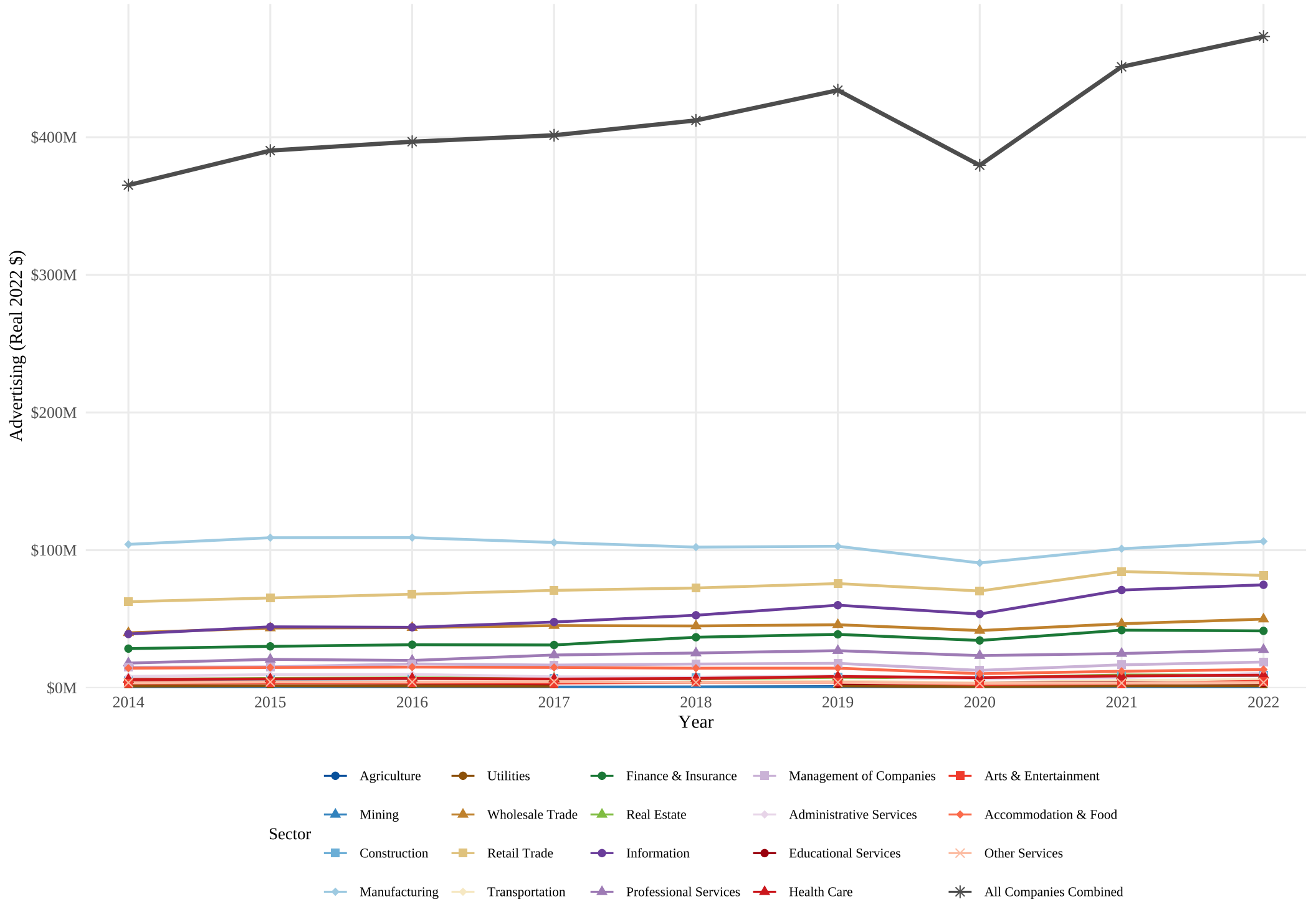
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Compensation to non-officer employees

Compensation of Officers by Sector, Active U.S. Corporations



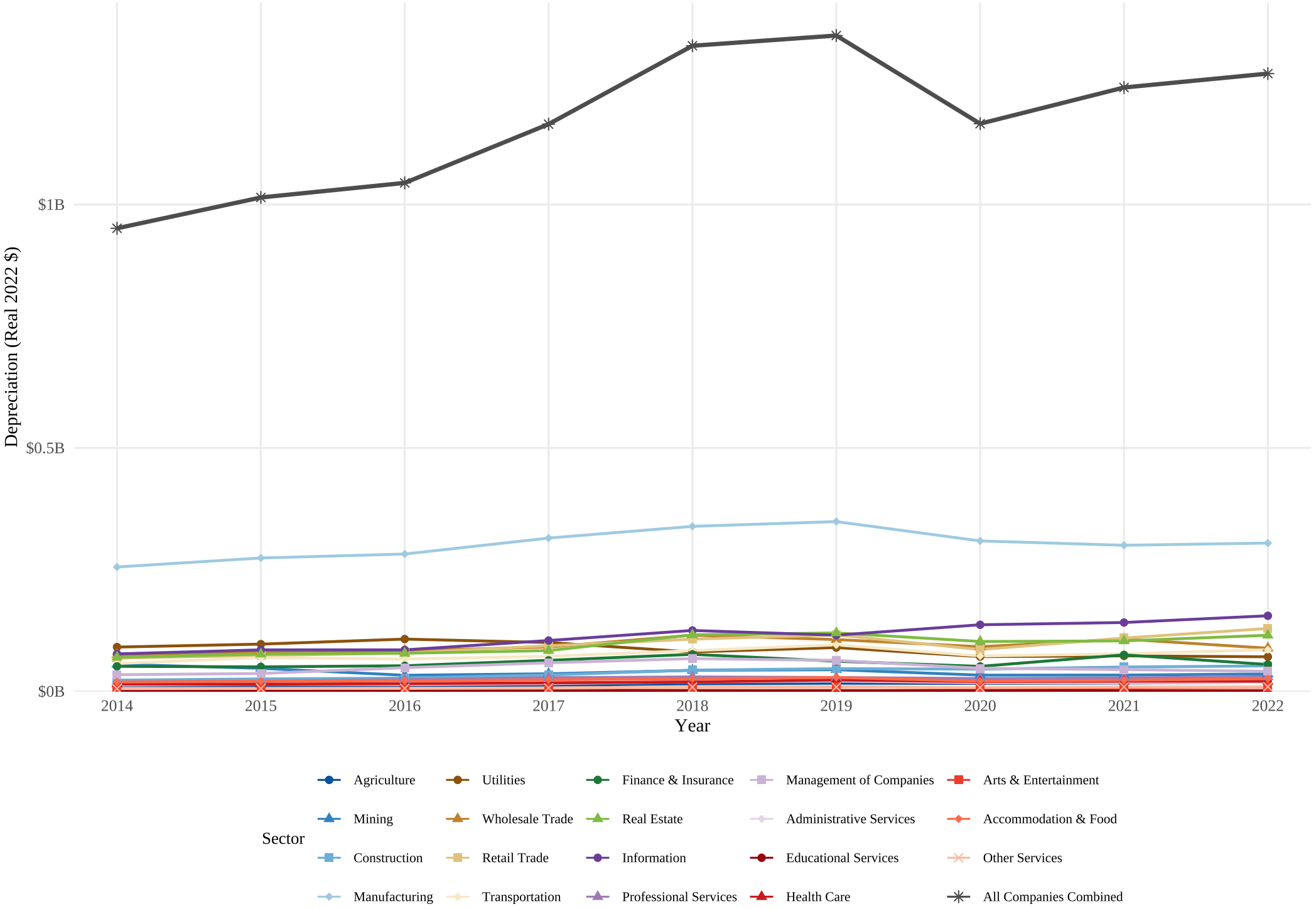
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Compensation to corporate officers

Advertising Deductions by Sector, Active U.S. Corporations



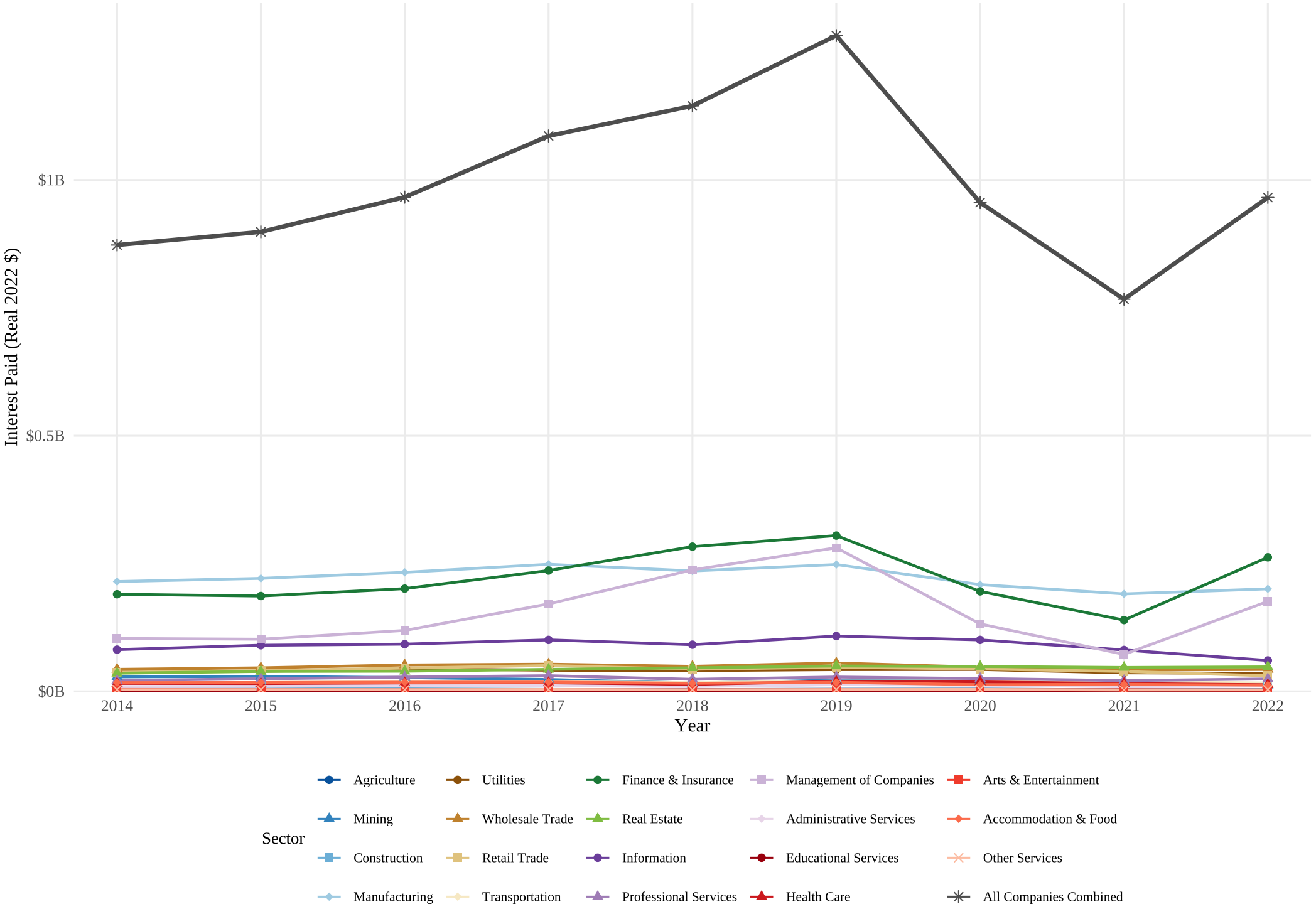
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Advertising and promotion expenses

Depreciation by Sector, Active U.S. Corporations



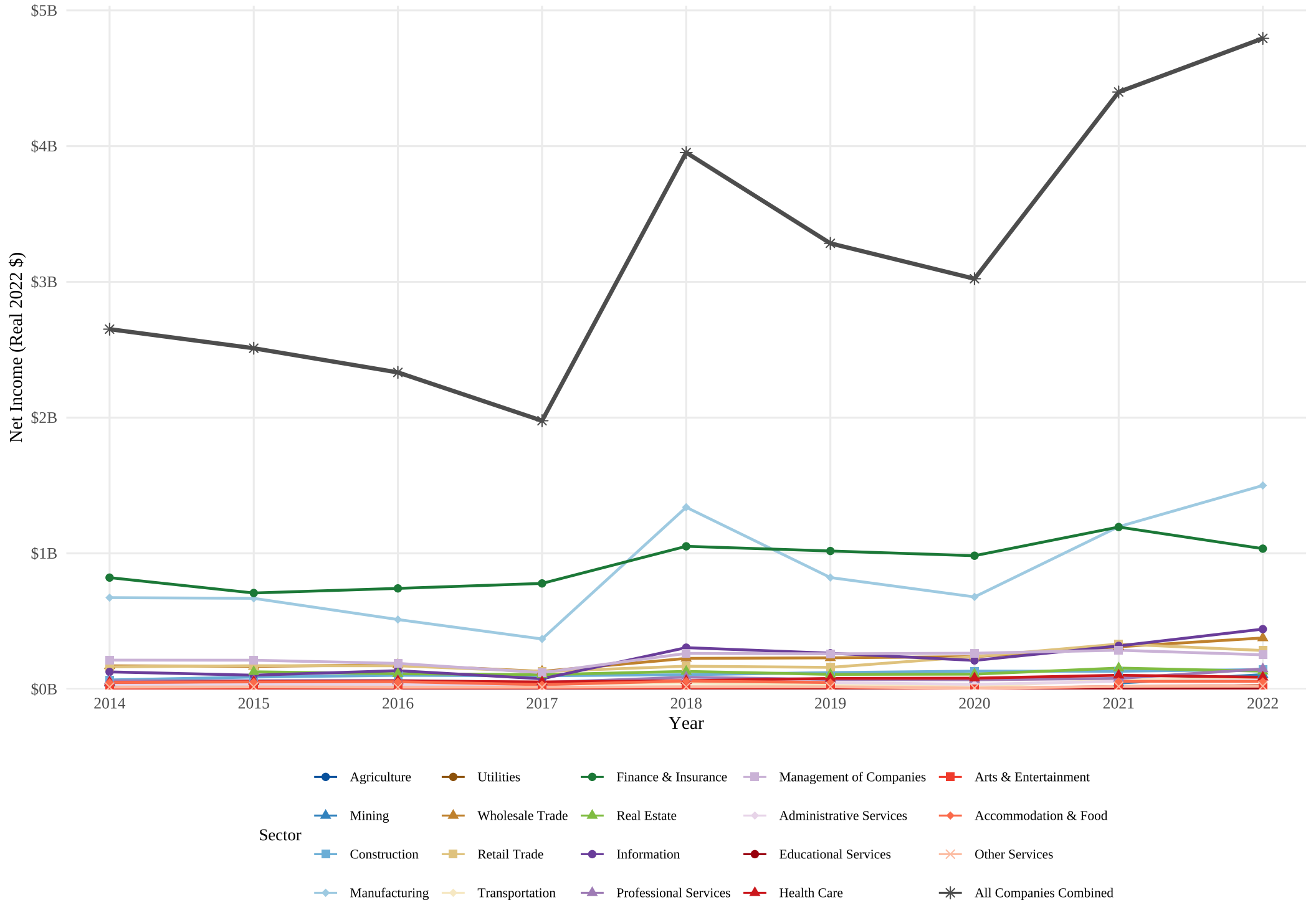
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Decline in value of tangible assets

Interest Paid by Sector, Active U.S. Corporations



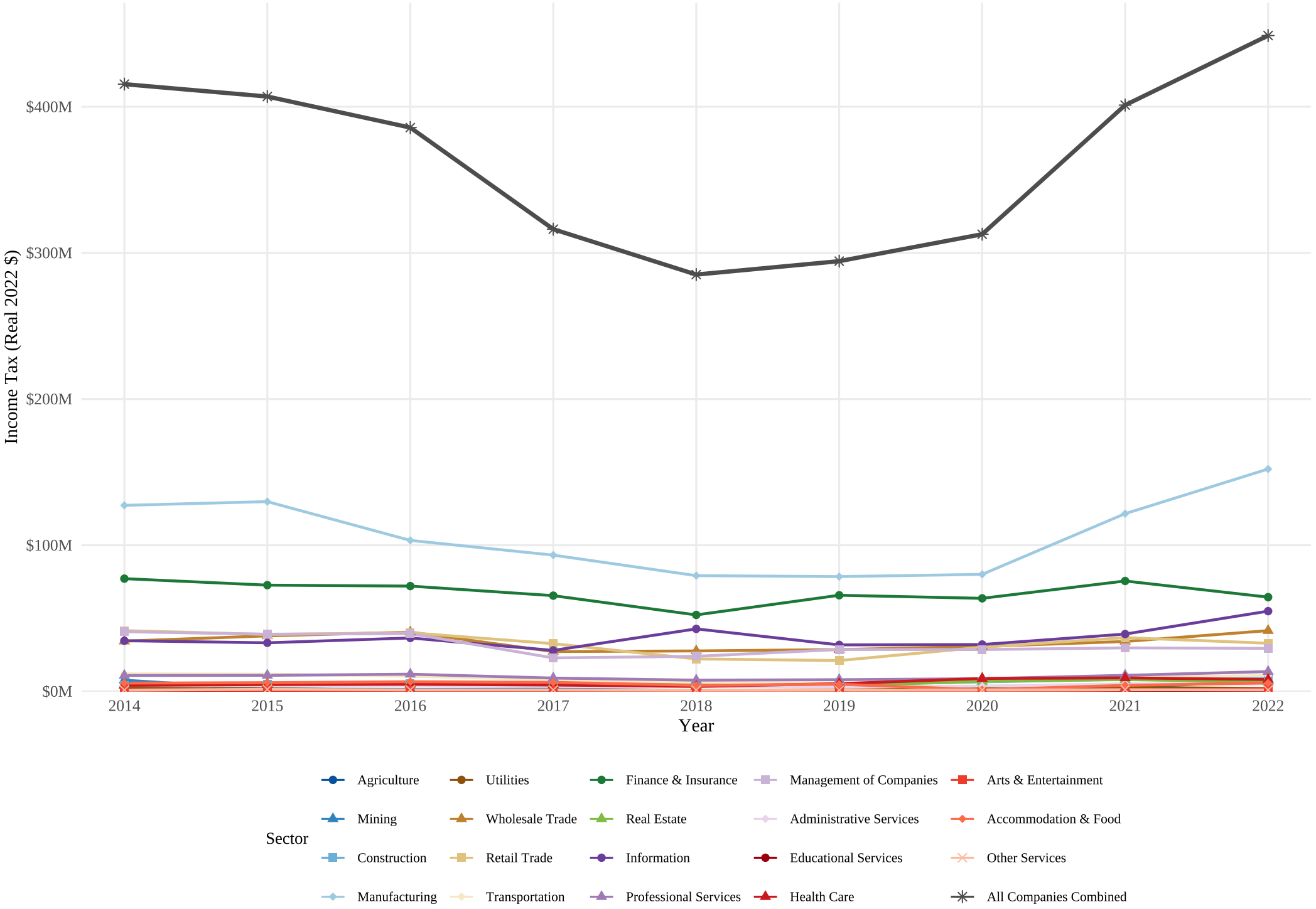
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Interest expenses on business debt

Net Income by Sector, Active U.S. Corporations



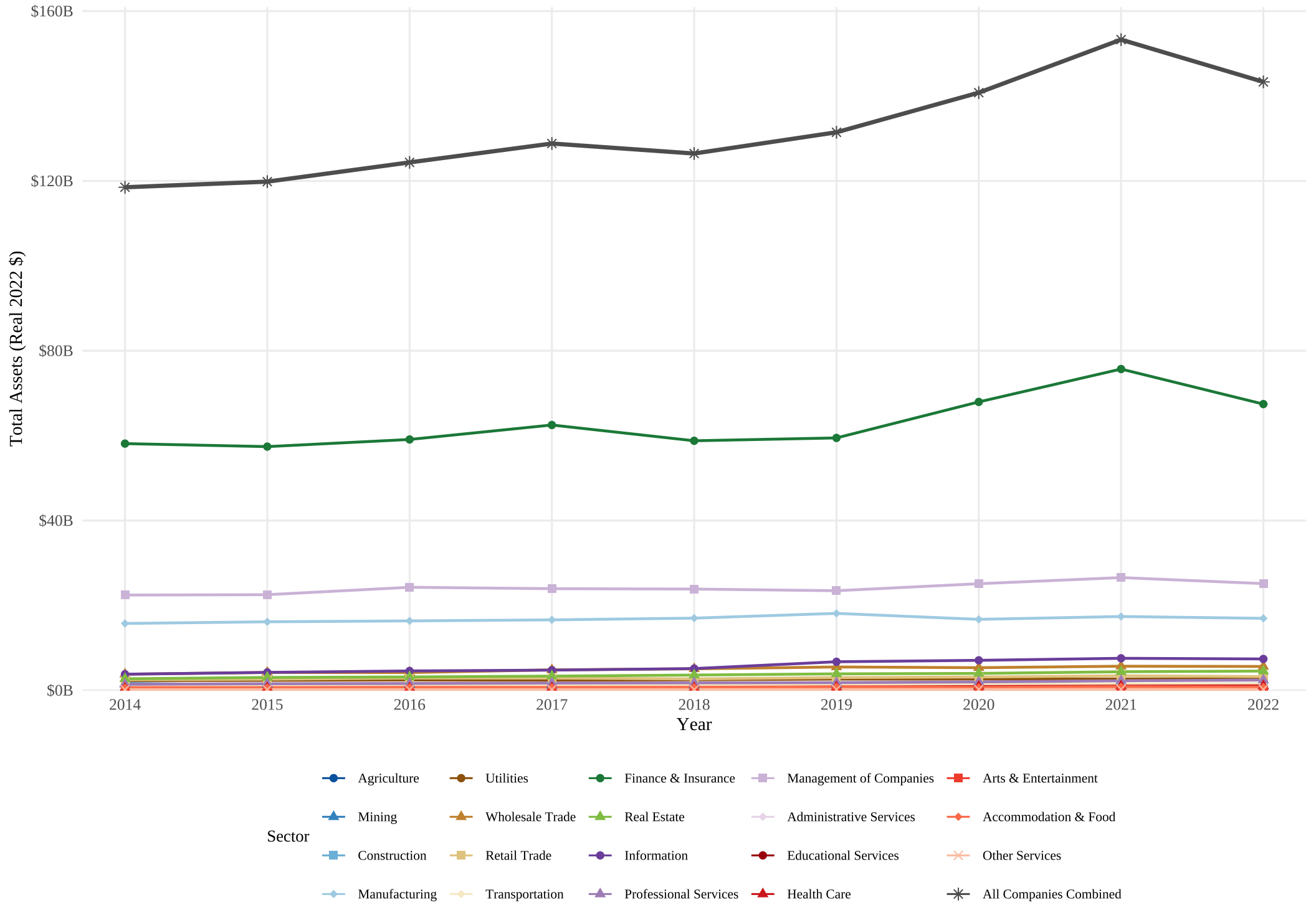
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Receipts minus deductions

Income Tax After Credits by Sector, Active U.S. Corporations



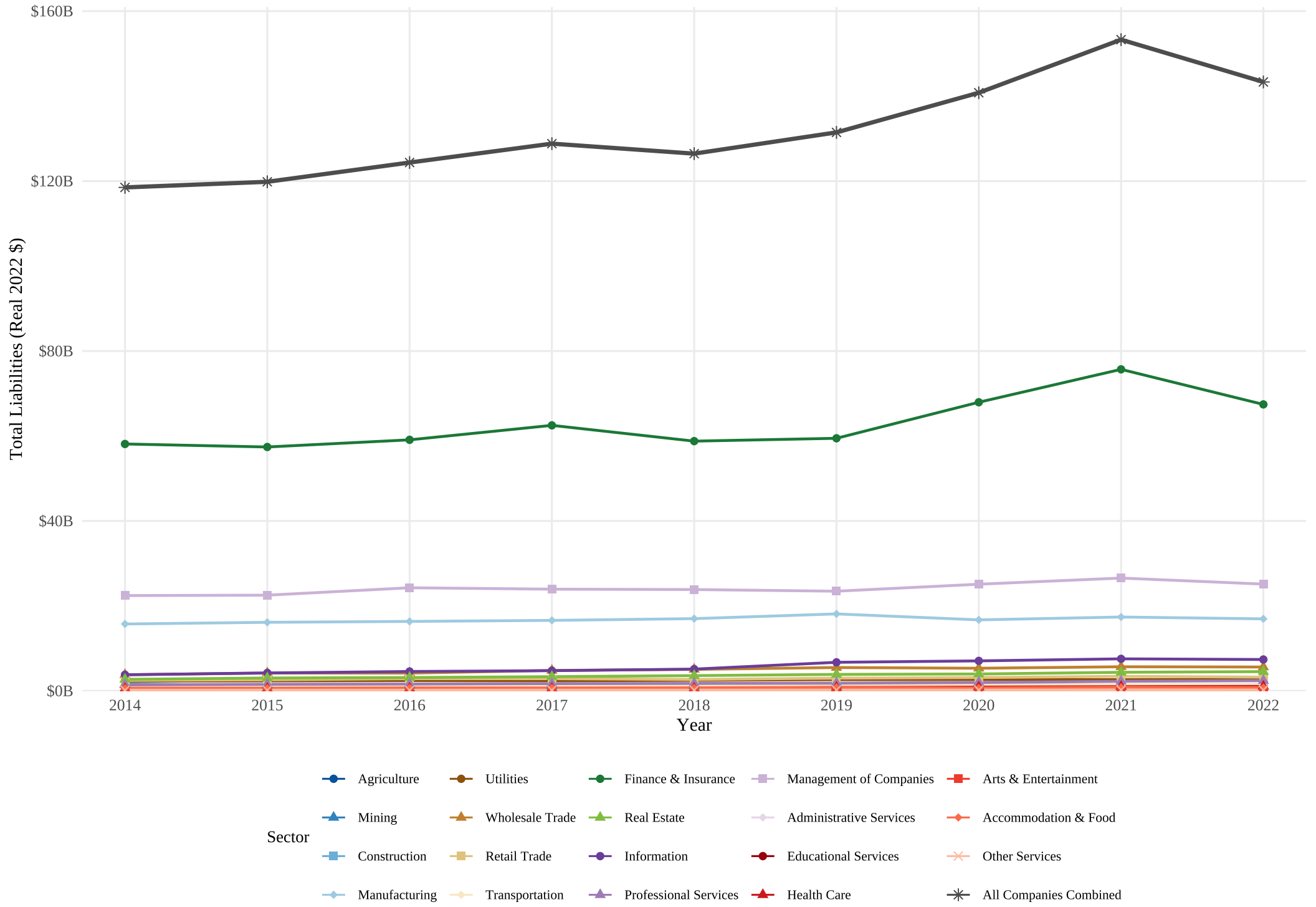
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Tax liability after credits

Total Assets by Sector, Active U.S. Corporations



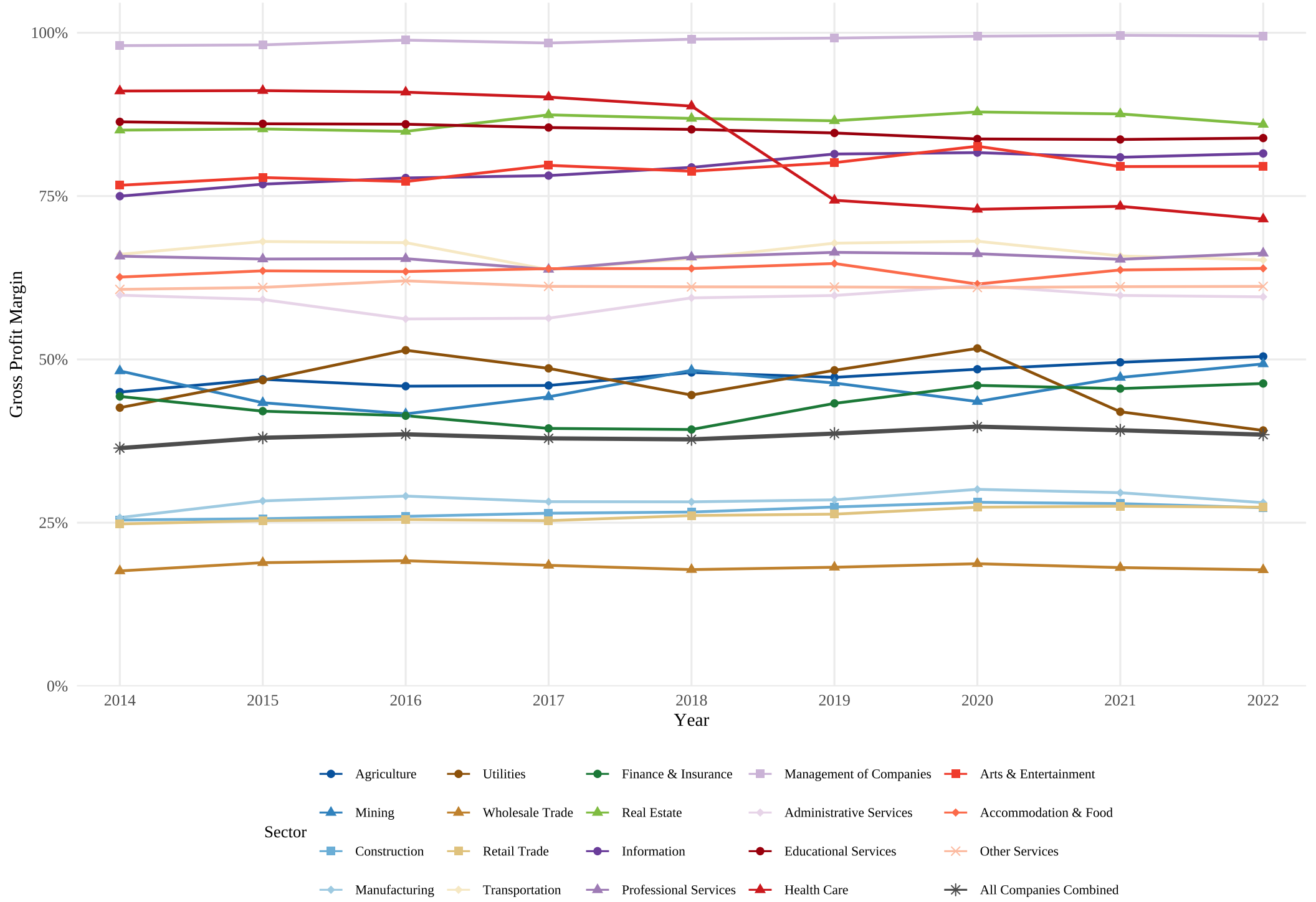
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Sum of all corporate assets

Total Liabilities by Sector, Active U.S. Corporations



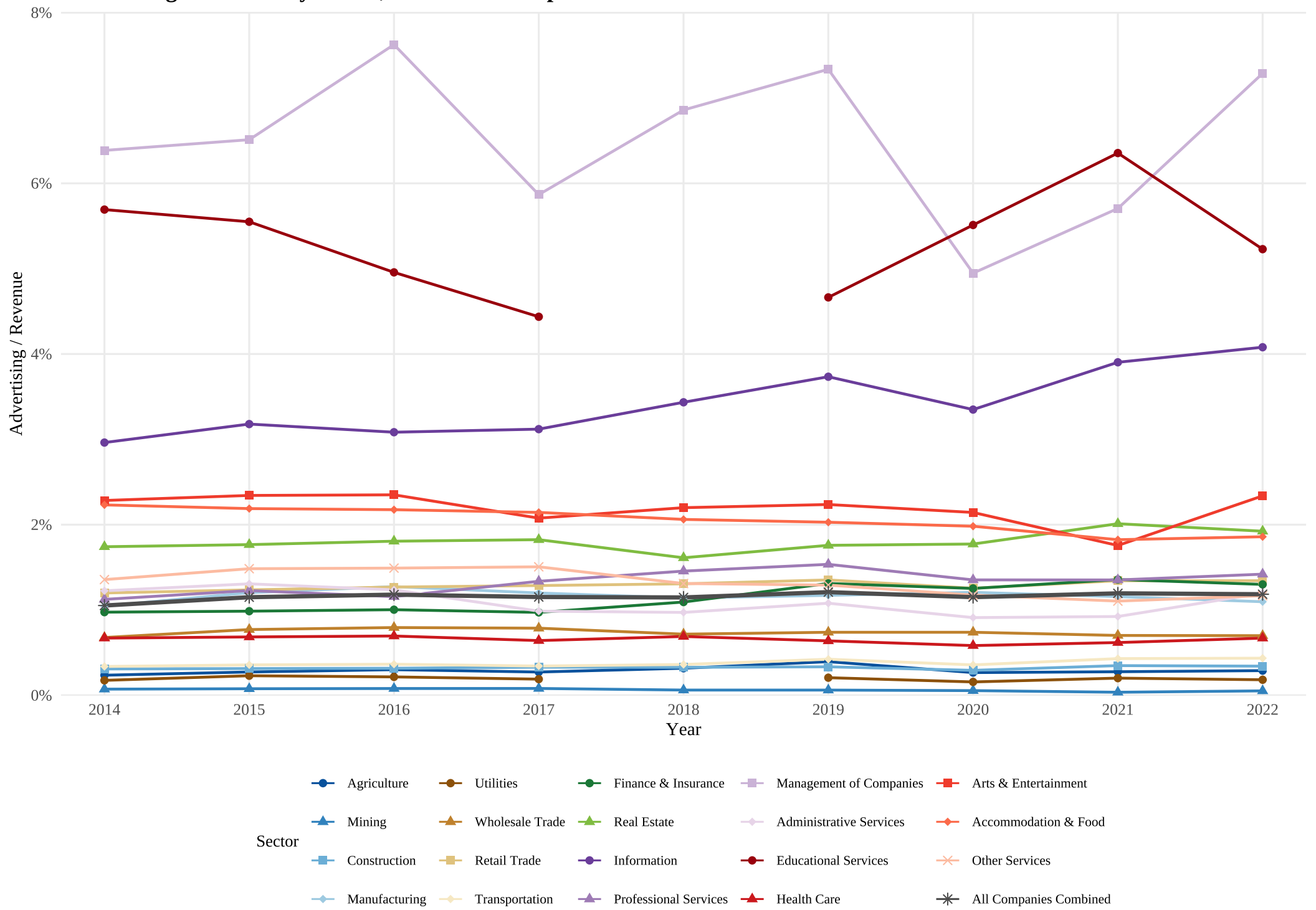
Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)
Definition: Sum of all corporate liabilities

Gross Profit Margin by Sector, Active U.S. Corporations



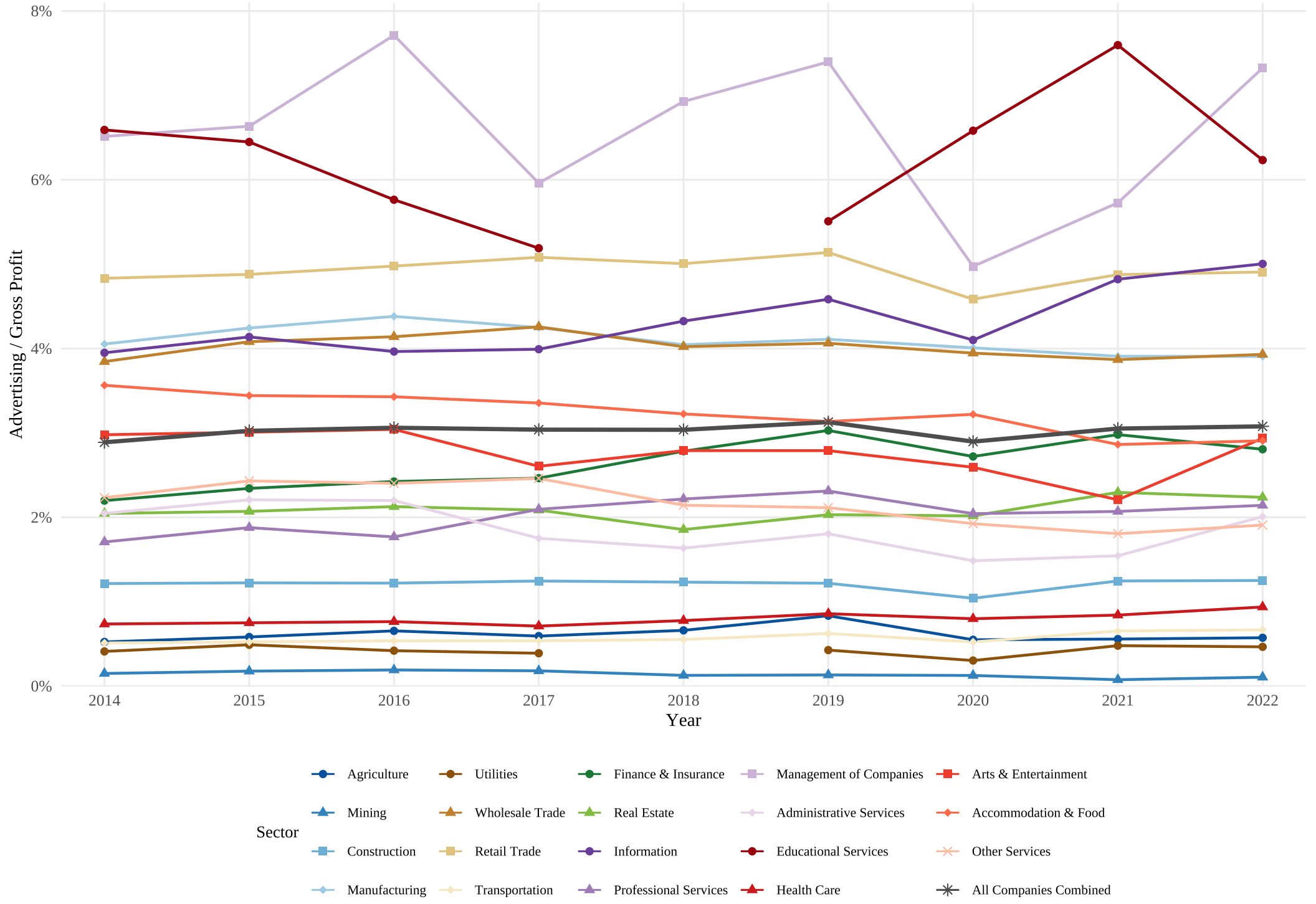
Source: IRS Statistics of Income, Table 5.1
Ratio: (Revenue - COGS) / Revenue

Advertising / Revenue by Sector, Active U.S. Corporations



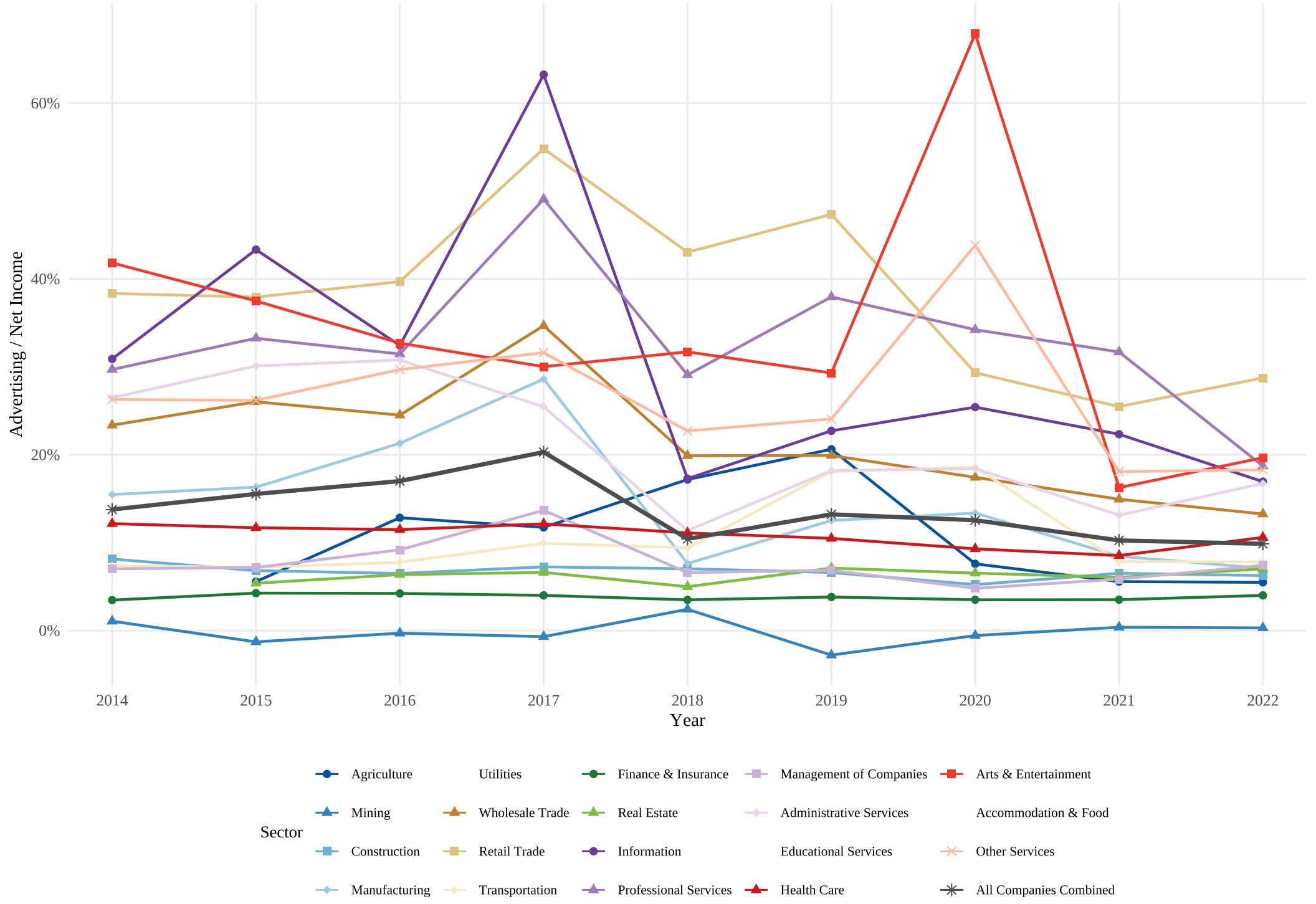
Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Advertising / Gross Profit by Sector, Active U.S. Corporations



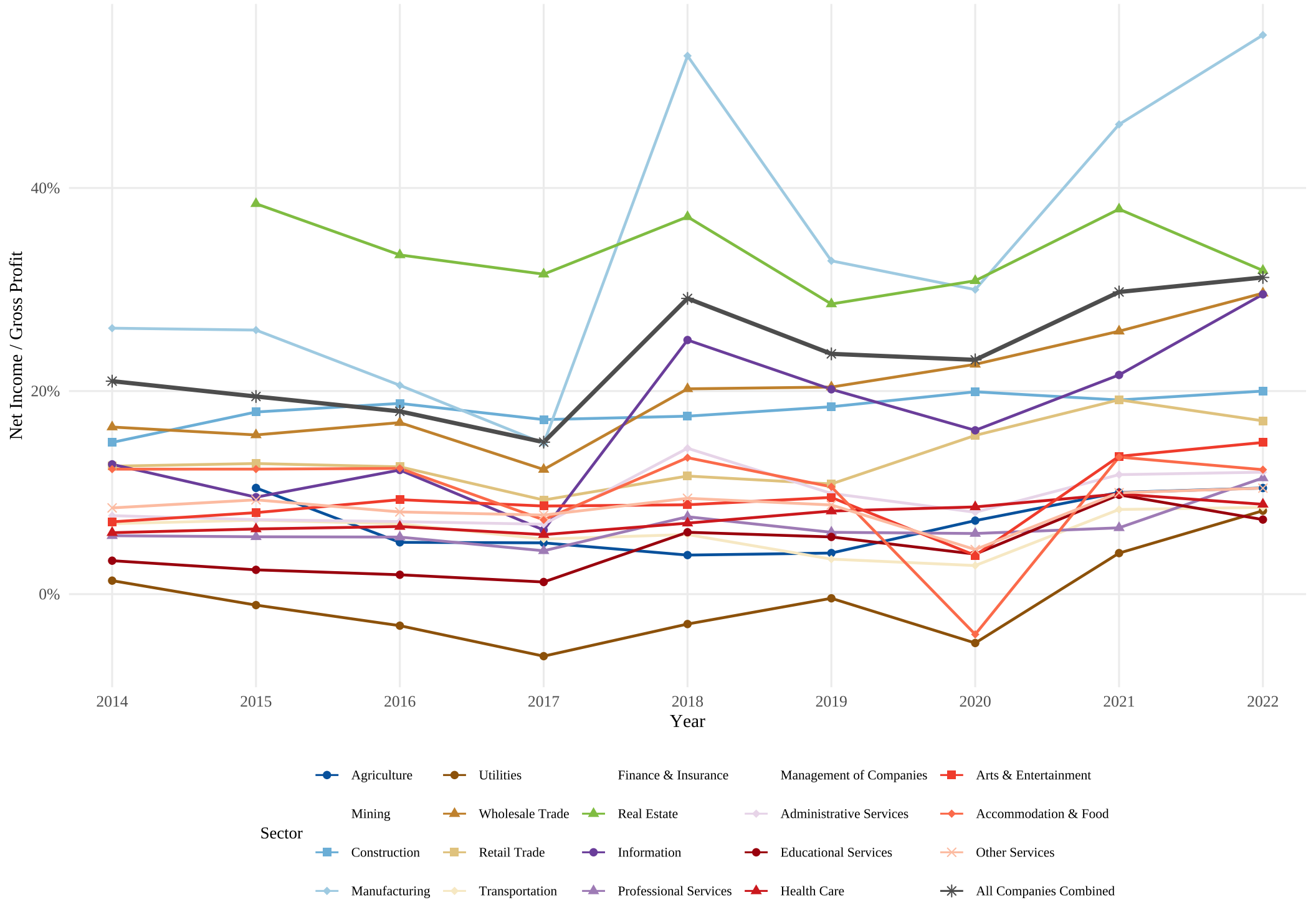
Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

Advertising / Net Income by Sector, Active U.S. Corporations



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Net Income

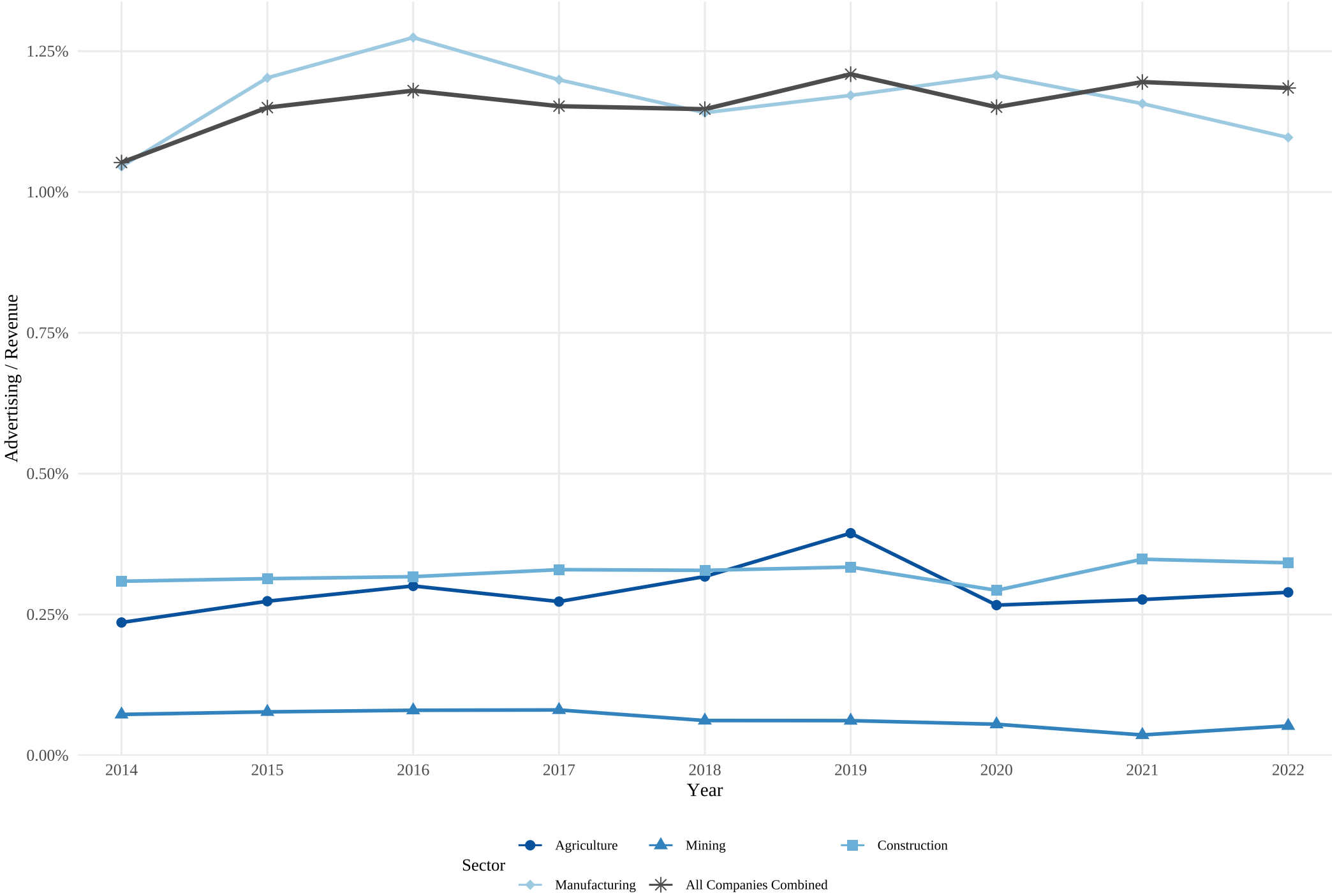
Net Income / Gross Profit by Sector, Active U.S. Corporations



Source: IRS Statistics of Income, Table 5.1
Ratio: Net Income / Gross Profit

Ad/Revenue Ratio: Goods-Producing

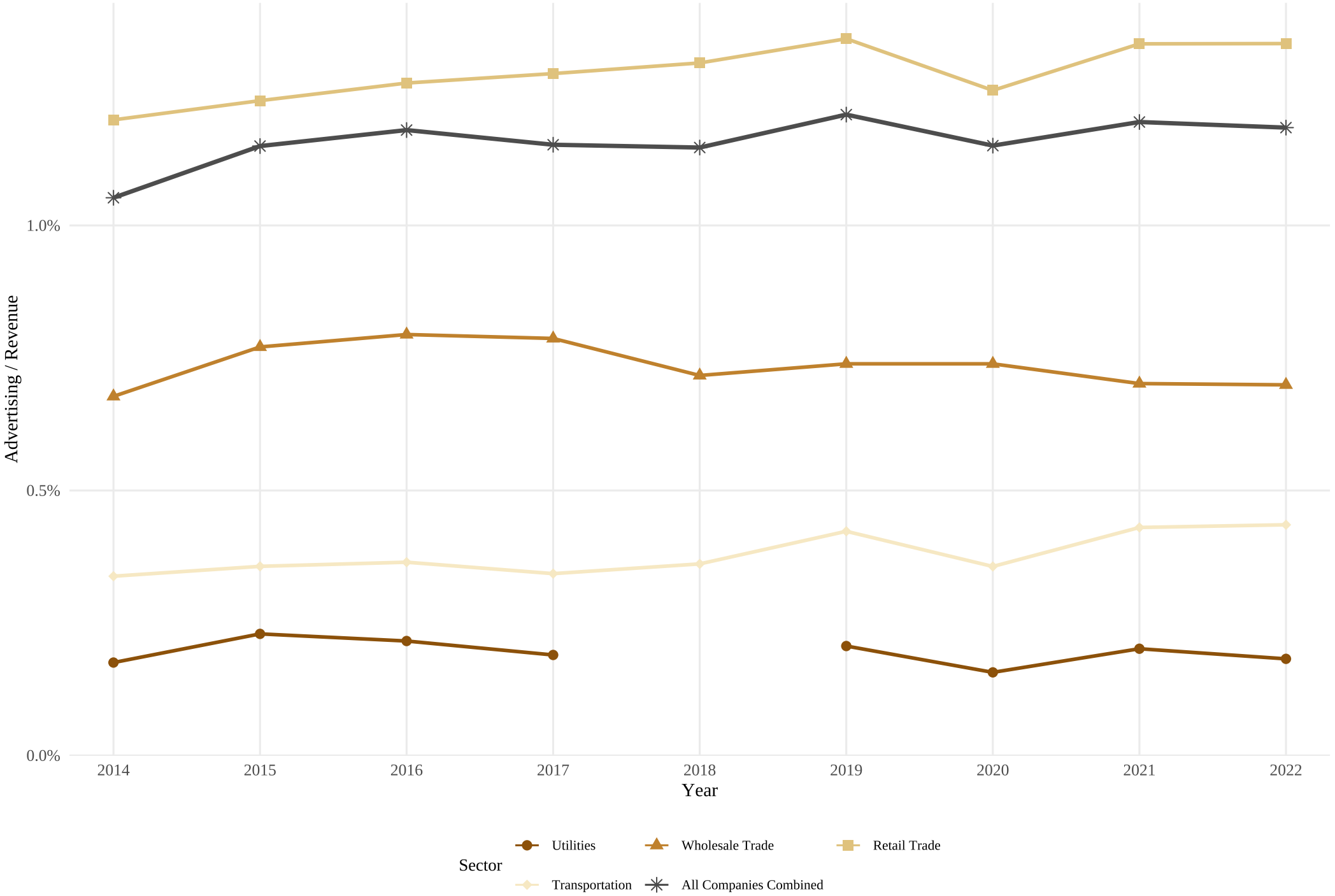
Sector Group: Goods-Producing



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Ad/Revenue Ratio: Distribution & Utilities

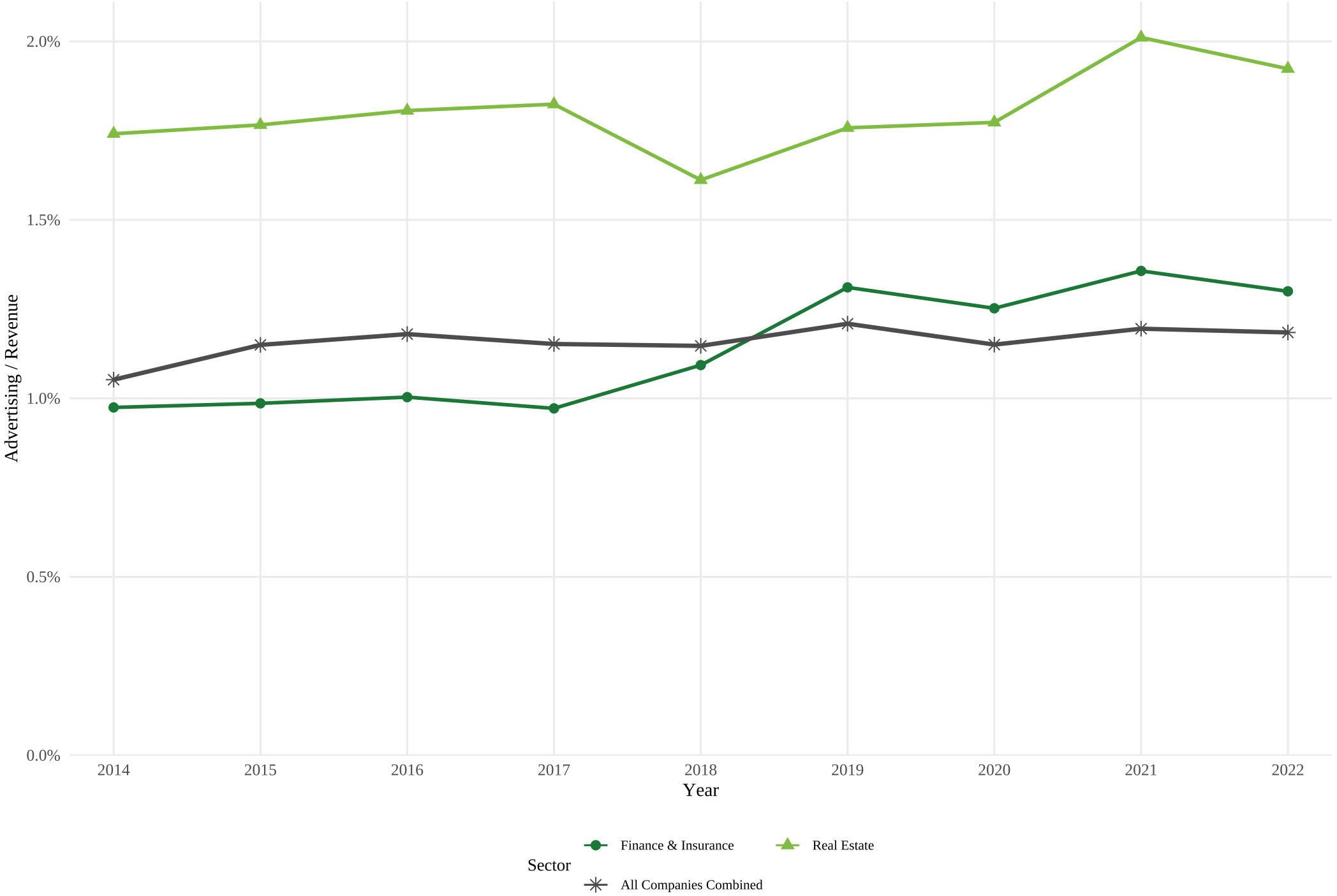
Sector Group: Distribution & Utilities



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Ad/Revenue Ratio: Finance & Real Estate

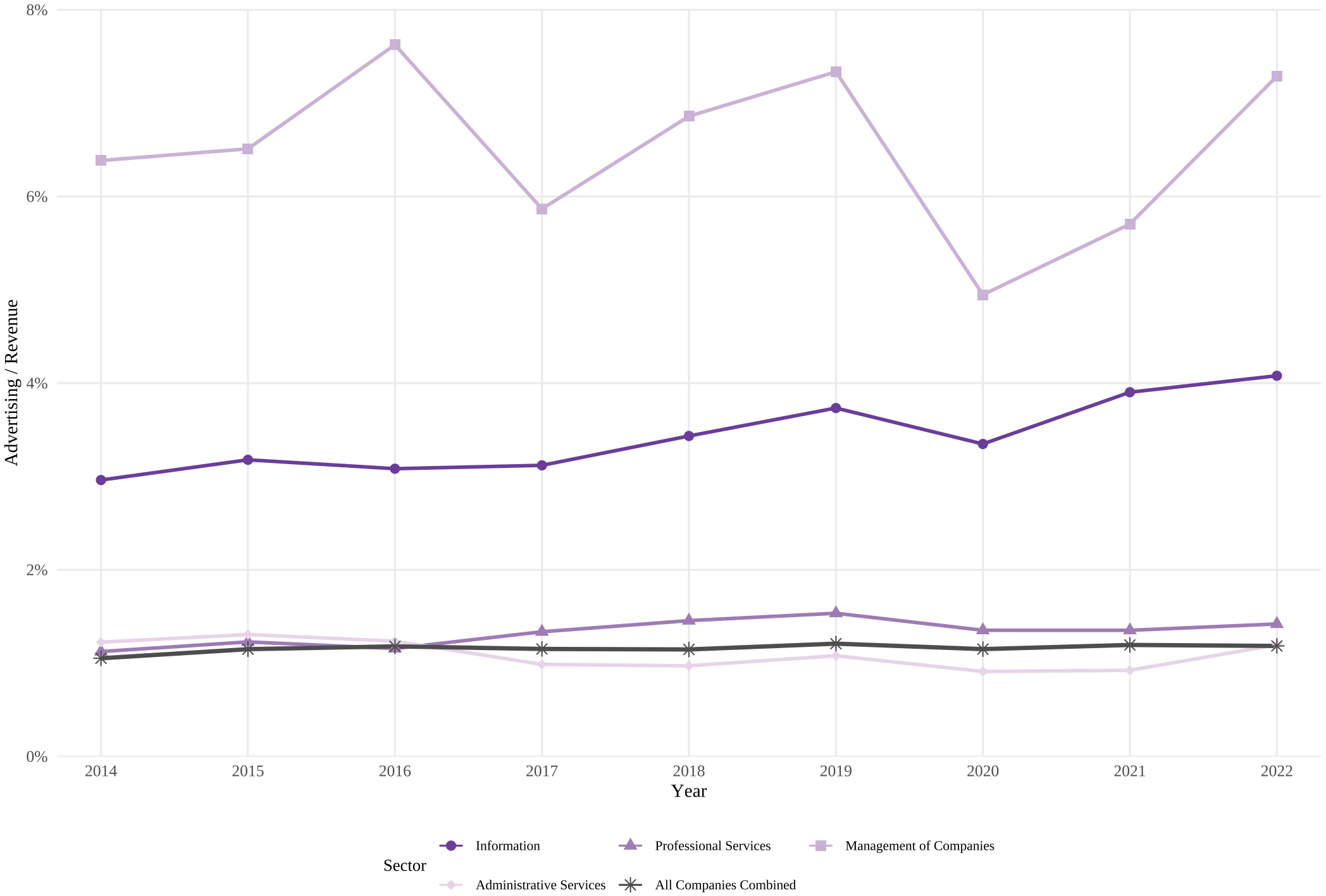
Sector Group: Finance & Real Estate



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Ad/Revenue Ratio: Business Services

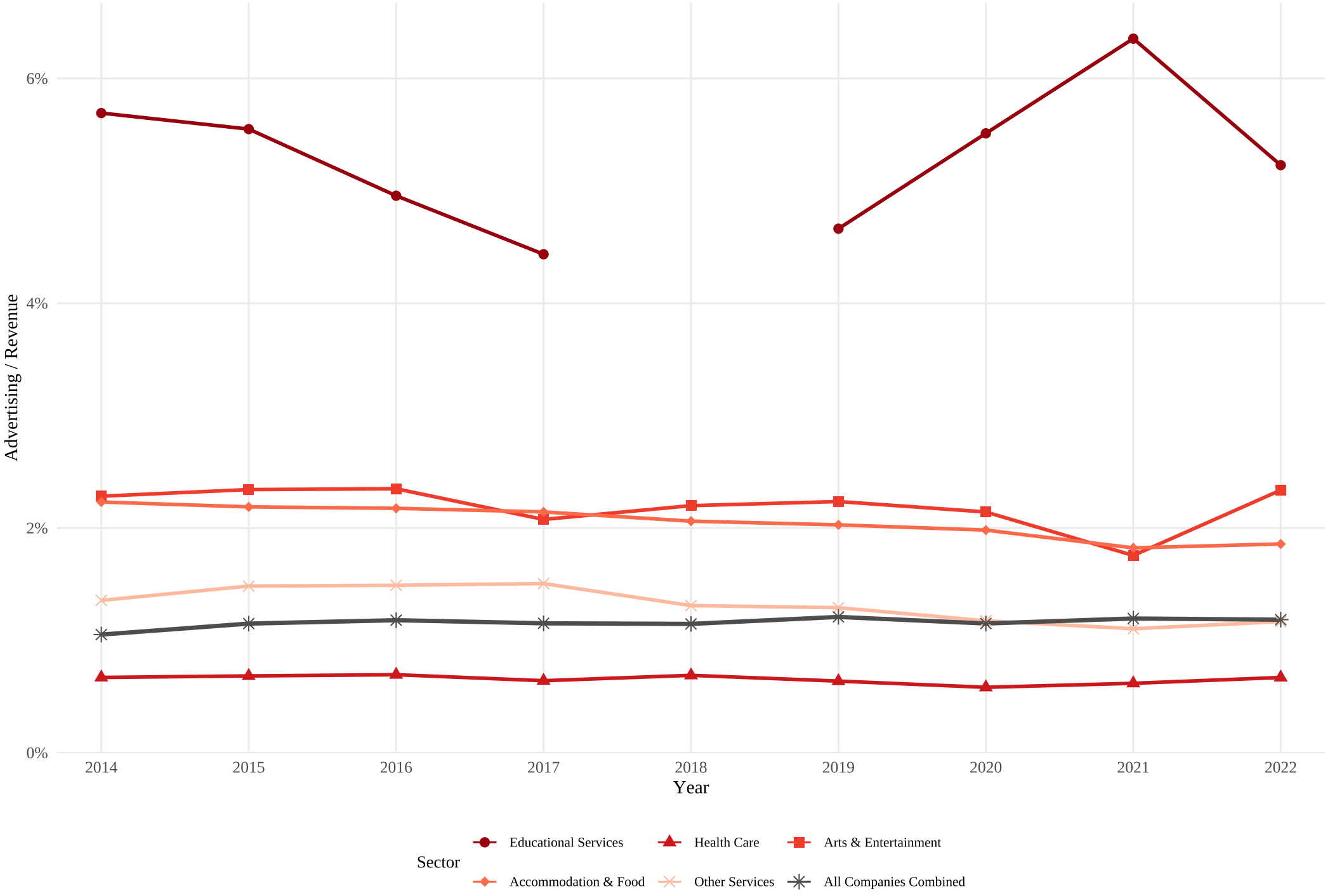
Sector Group: Business Services



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Ad/Revenue Ratio: Consumer Services

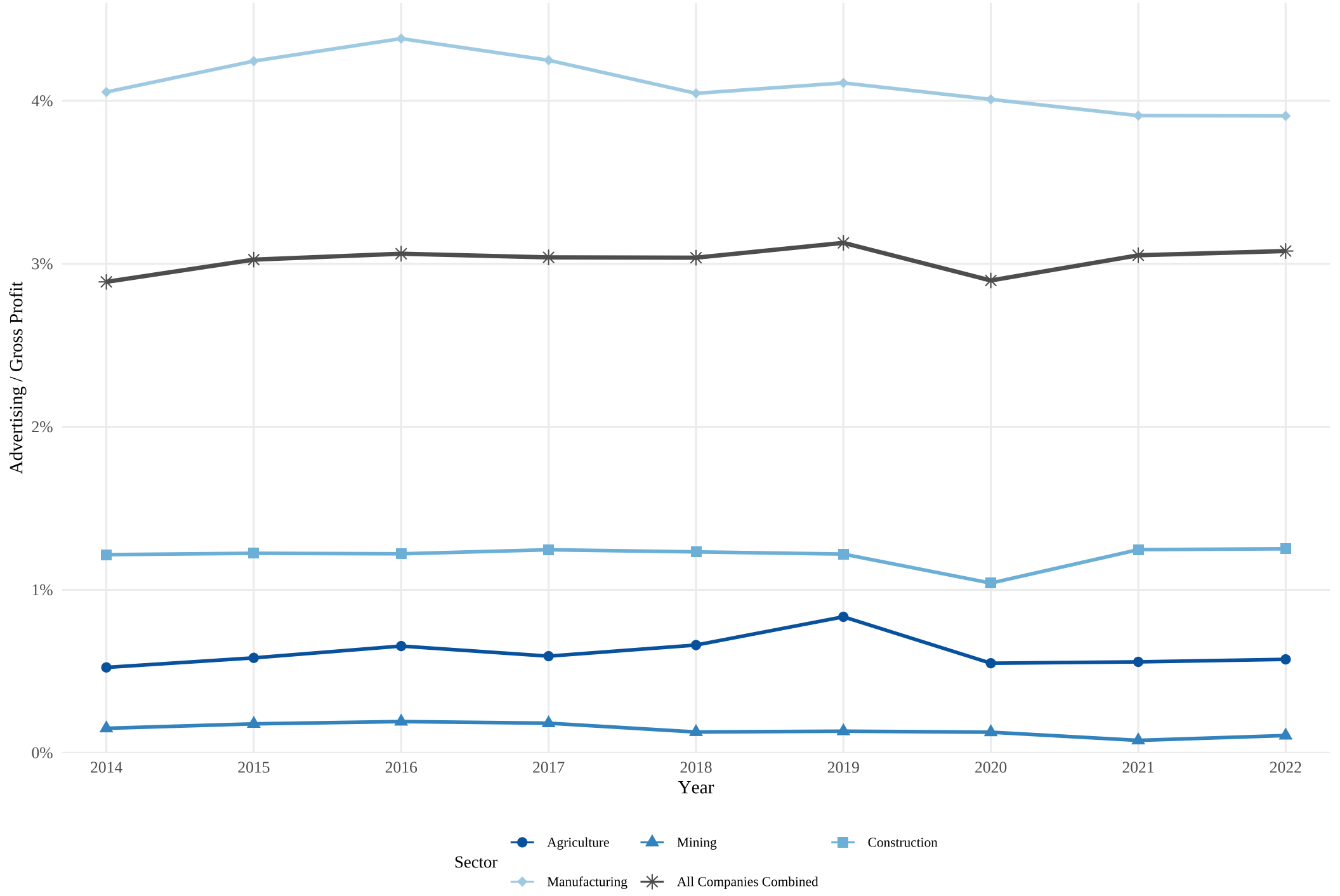
Sector Group: Consumer Services



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Revenue

Ad/Gross Profit Ratio: Goods-Producing

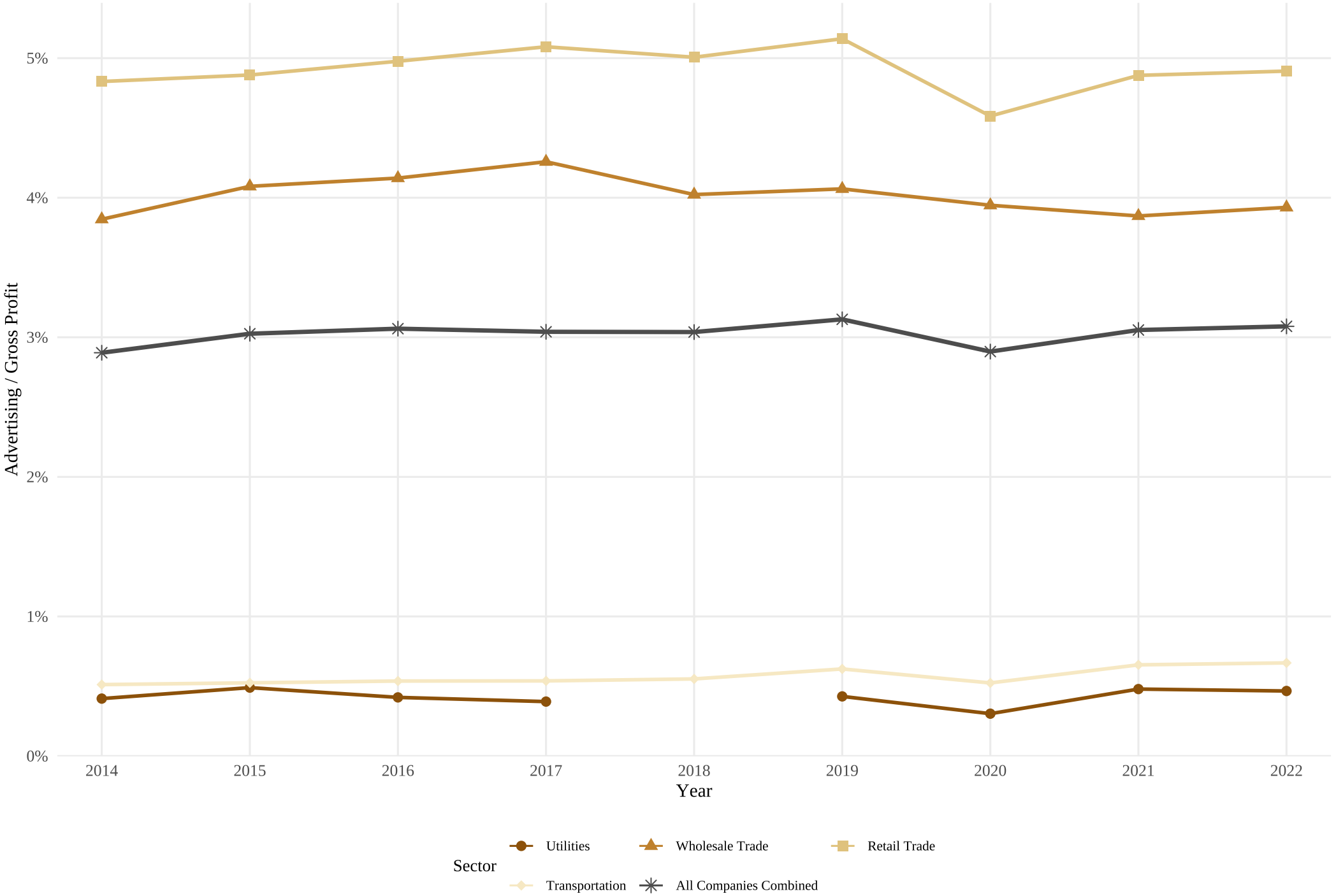
Sector Group: Goods-Producing



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

Ad/Gross Profit Ratio: Distribution & Utilities

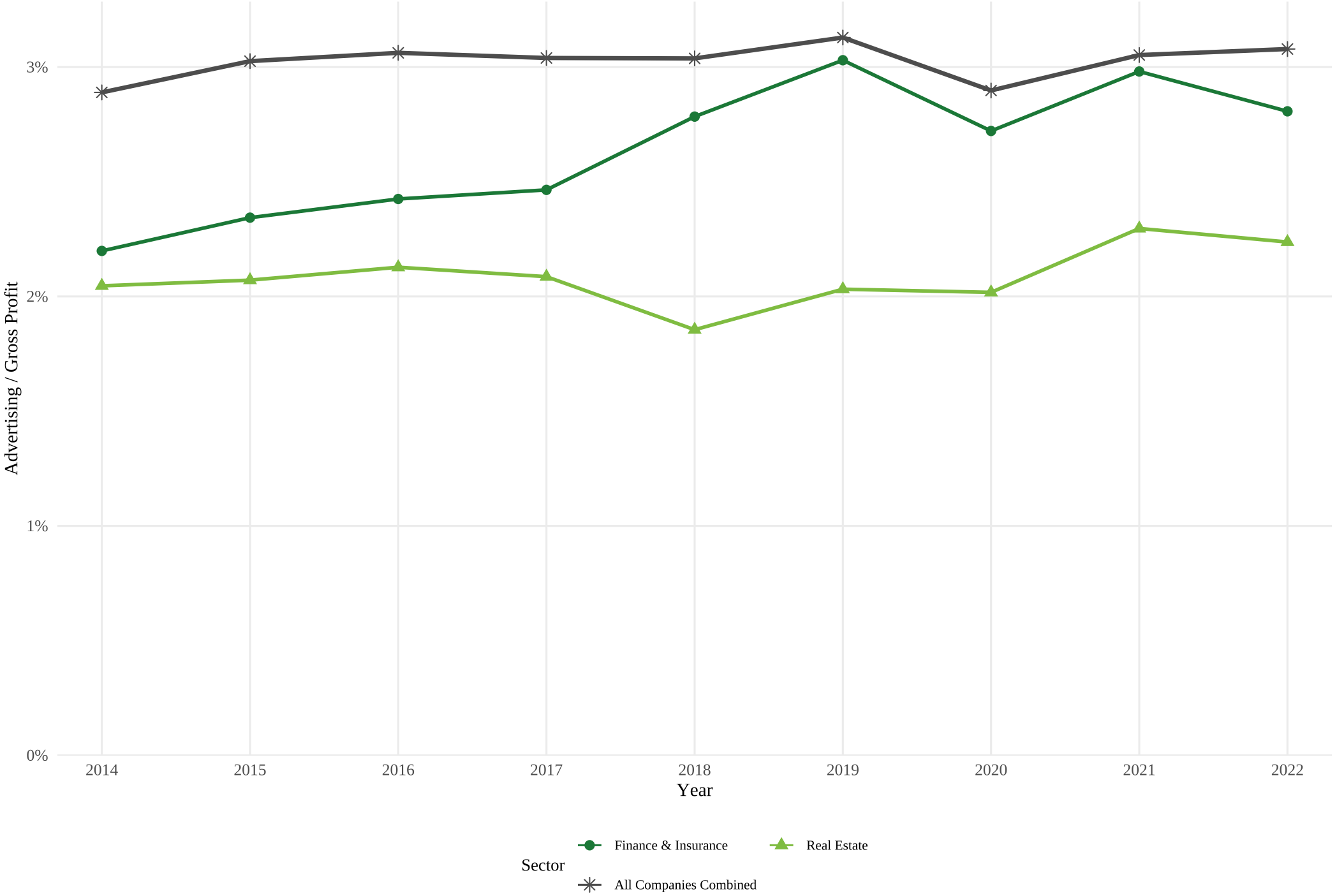
Sector Group: Distribution & Utilities



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

Ad/Gross Profit Ratio: Finance & Real Estate

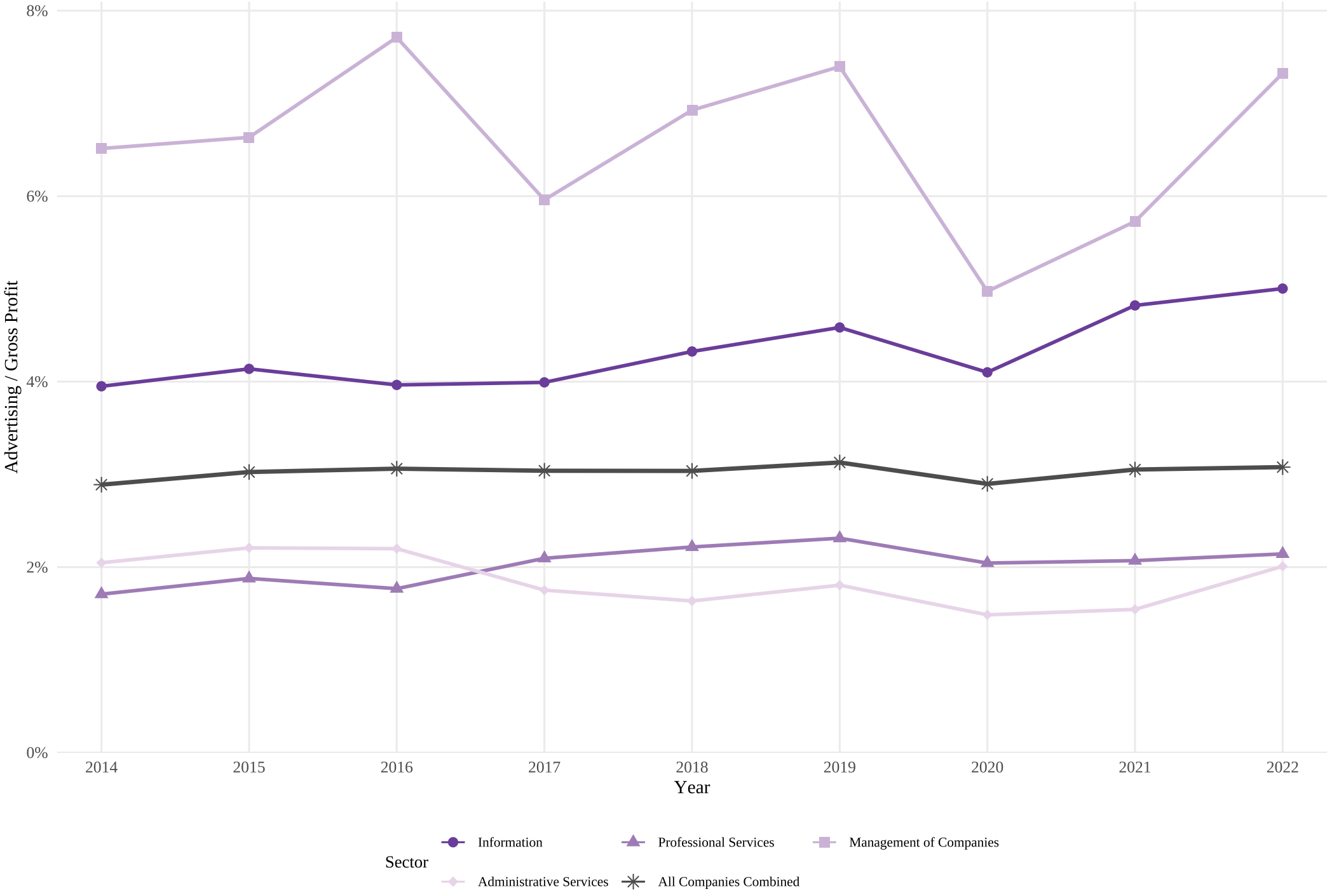
Sector Group: Finance & Real Estate



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

Ad/Gross Profit Ratio: Business Services

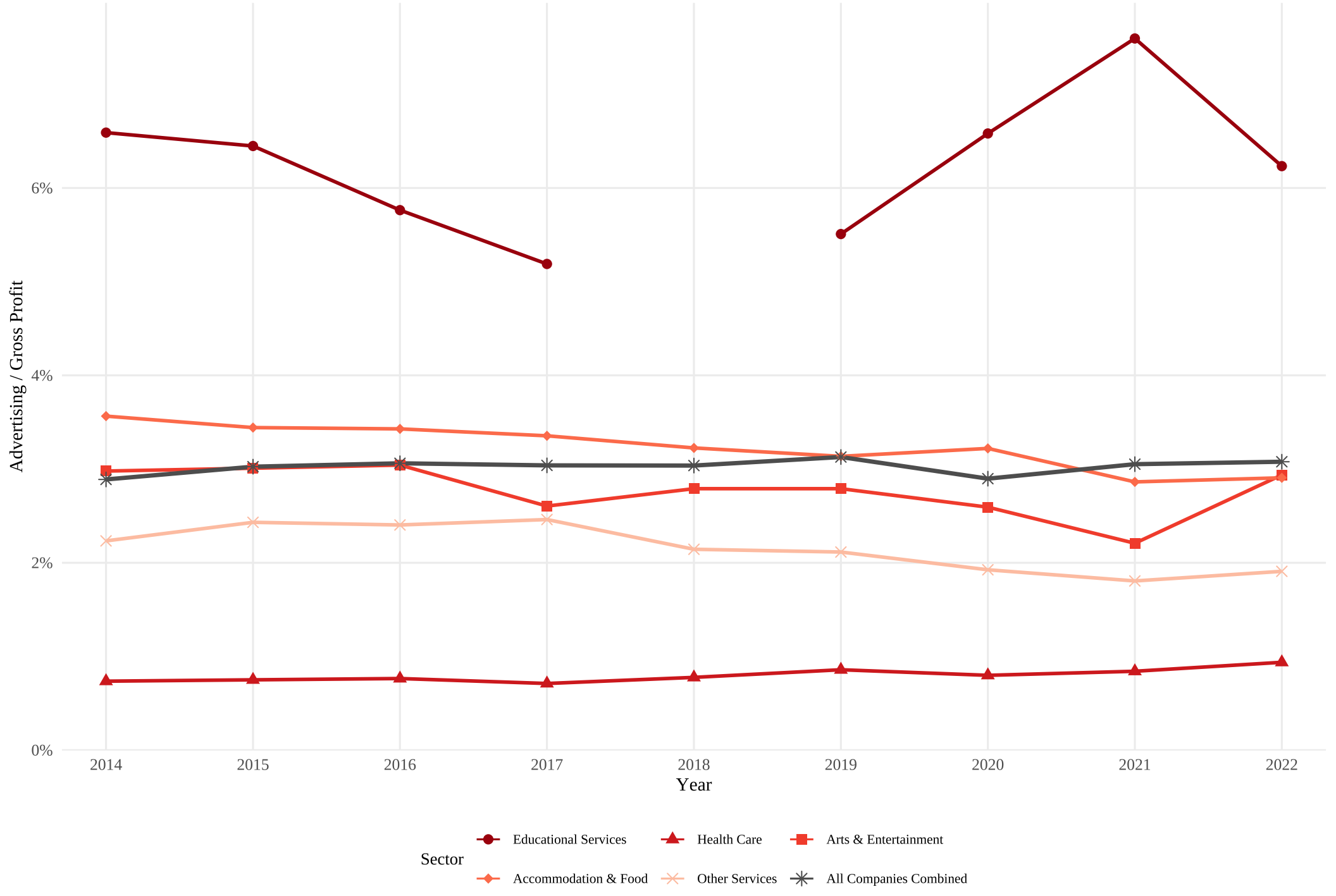
Sector Group: Business Services



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

Ad/Gross Profit Ratio: Consumer Services

Sector Group: Consumer Services



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit

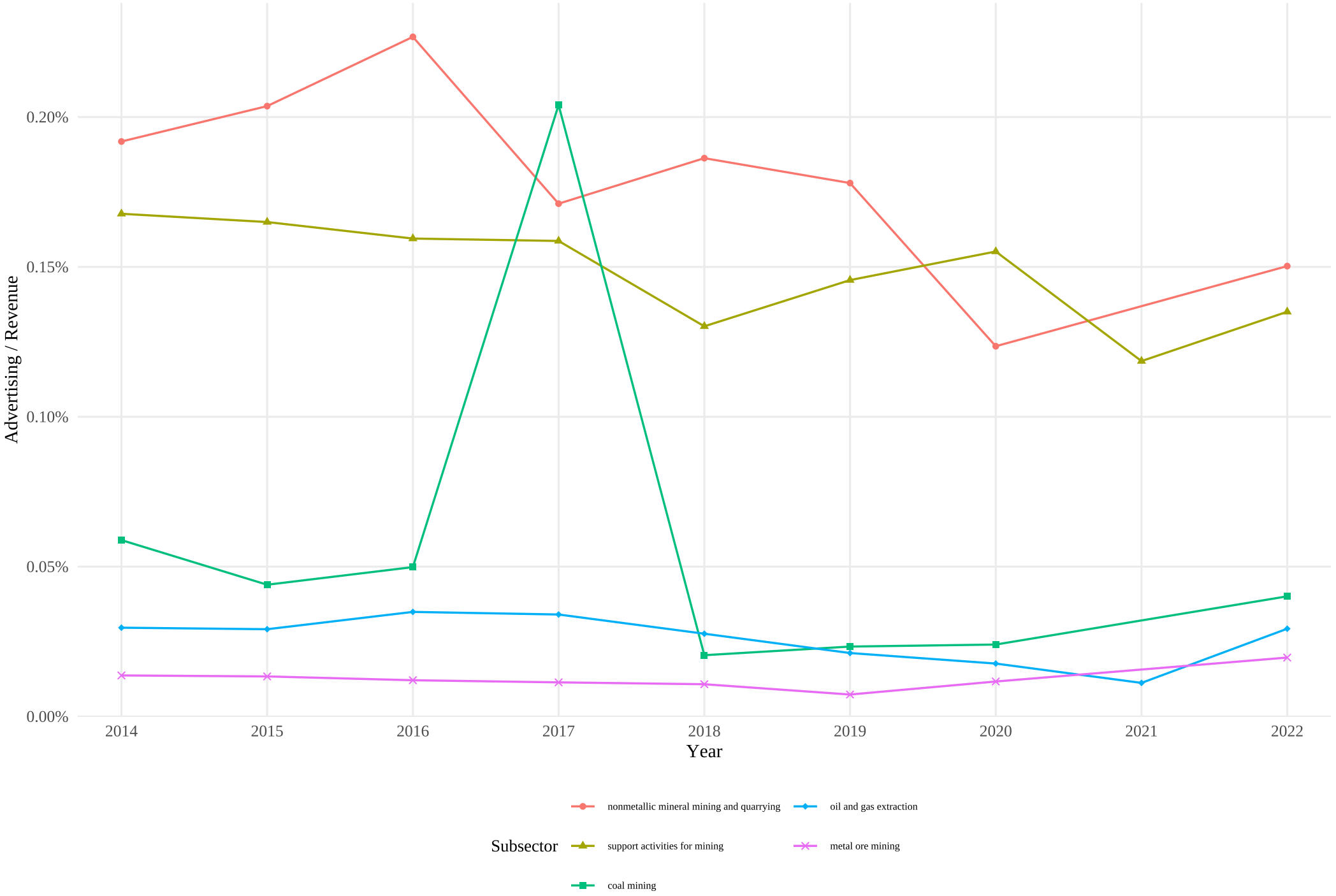
Ad/Revenue: Agriculture

3 subsectors

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Mining

5 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

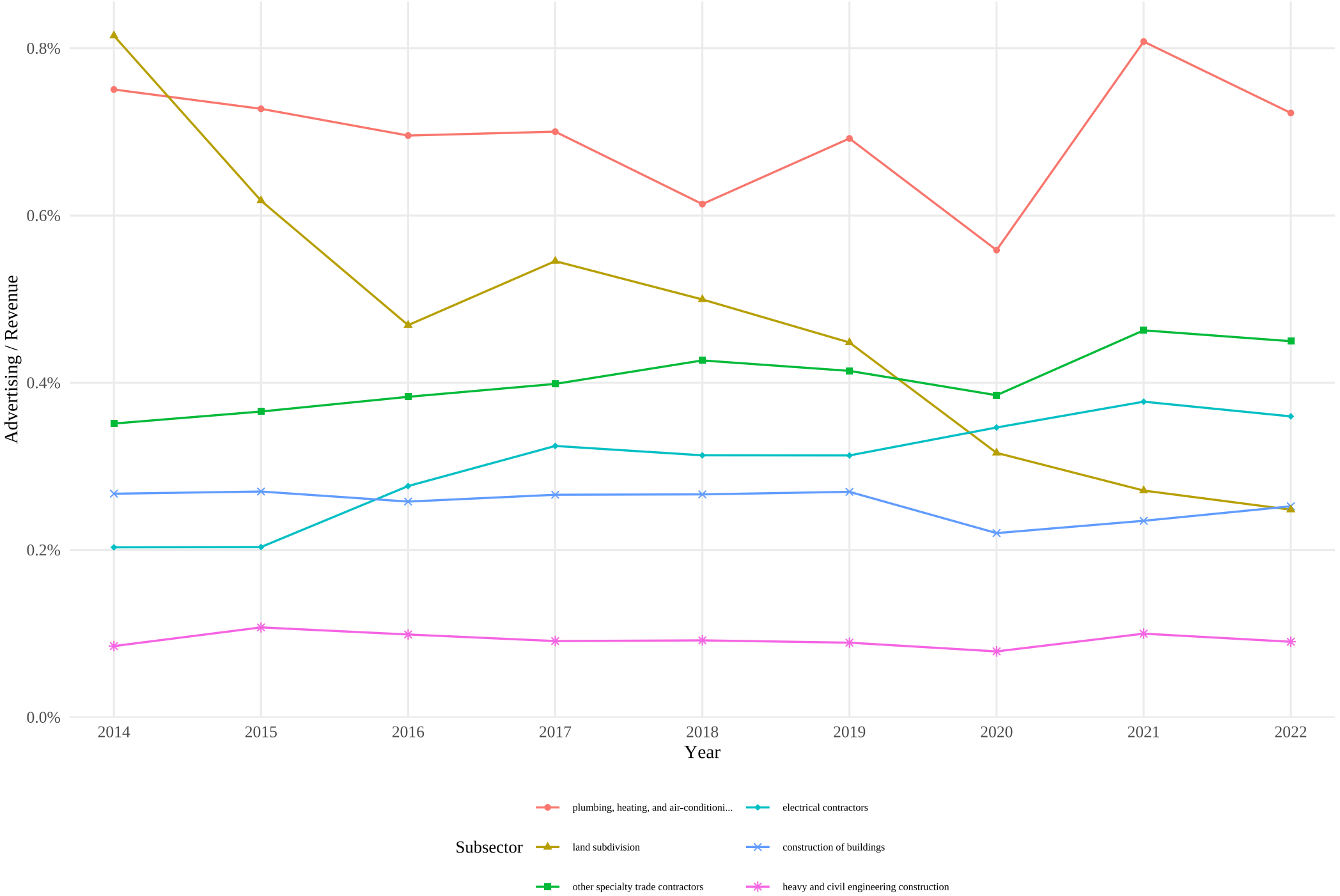
Ad/Revenue: Utilities

4 subsectors

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Construction

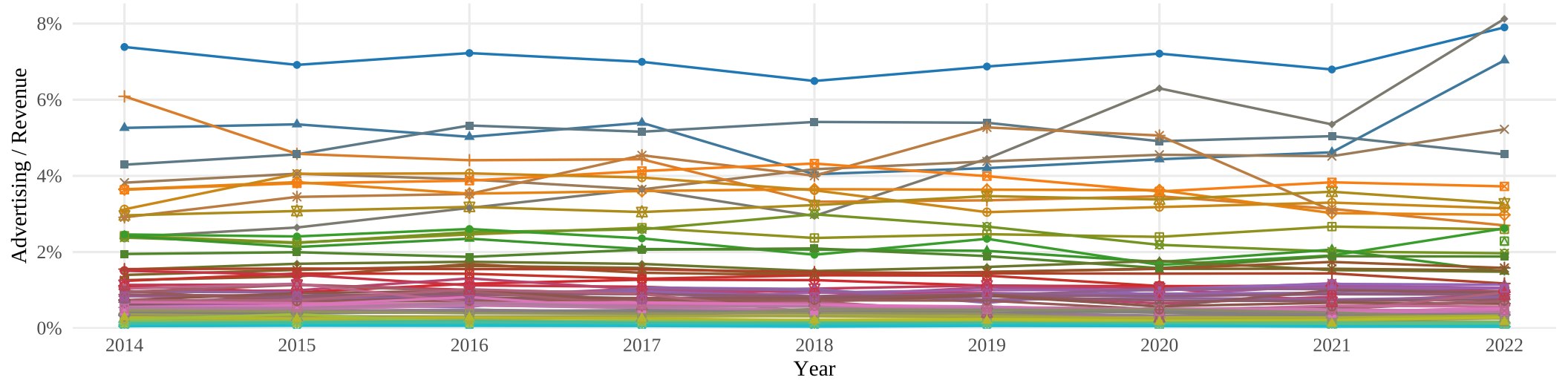
6 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Manufacturing

66 subsectors



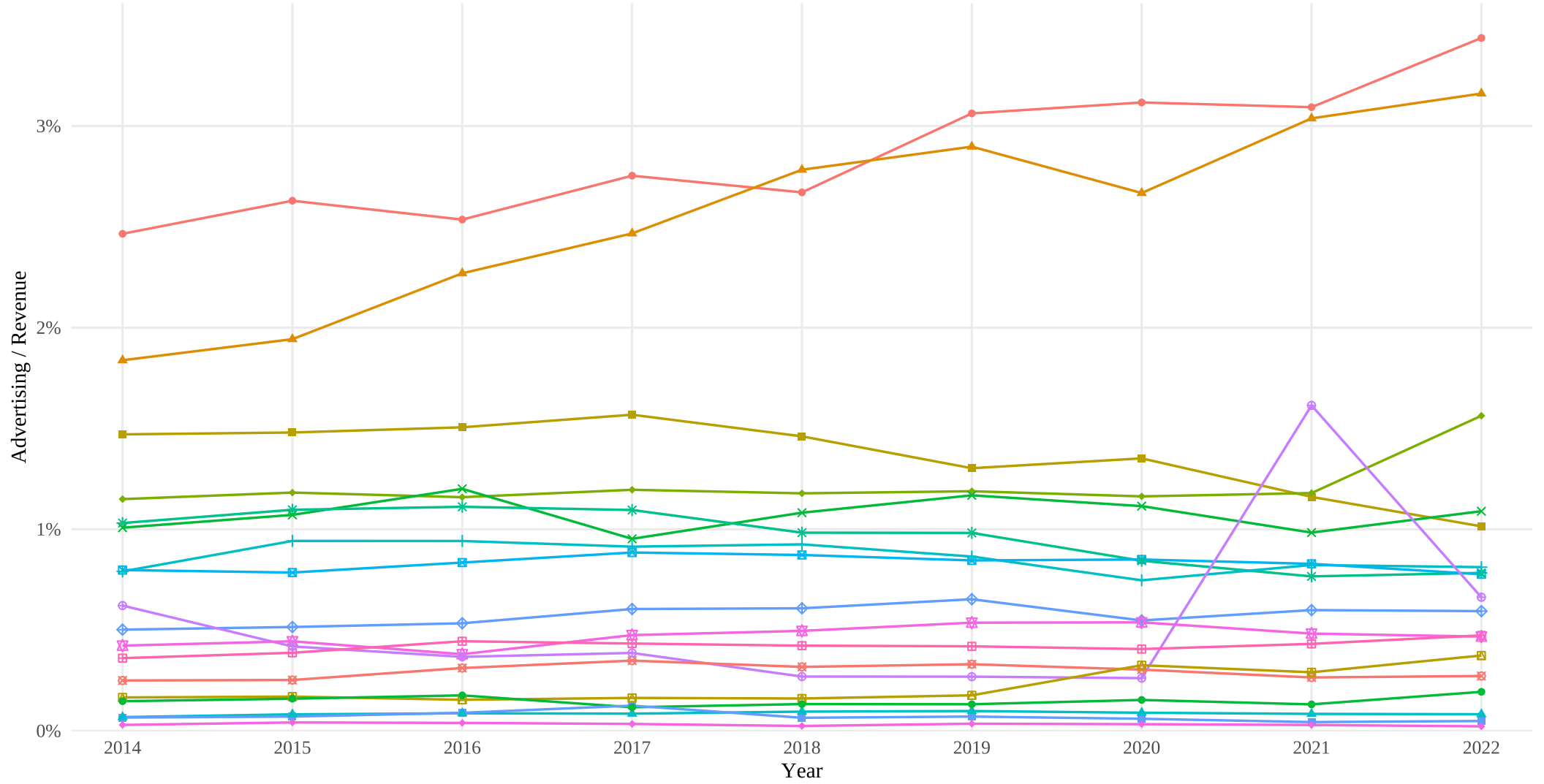
Subsector

- soap, cleaning compound, and toilet p...
- wineries and distilleries
- breweries
- apparel knitting mills
- leather and allied product manufacturing
- apparel accessories and other apparel
- sugar and confectionery product
- pharmaceutical and medicine
- animal food manufacturing and grain a...
- soft drink and ice
- other food
- cut and sew apparel contractors and m...
- audio and video equipment manufacturi...
- apparel accessories and other apparel...
- furniture and related product manufac...
- fruit and vegetable preserving and sp...
- other miscellaneous manufacturing
- paint, coating, and adhesive
- computer and peripheral equipment
- dairy product
- motor vehicles and parts
- rubber product
- bakeries and tortilla manufacturing
- electrical lighting equipment and hou...
- other transportation equipment and ra...
- asphalt paving, roofing, other petrol...
- tobacco manufacturing
- medical equipment and supplies
- other general purpose machinery
- textile product mills
- commercial and service industry machi...
- printing and related support activities
- converted paper product
- navigational, measuring, electromedic...
- ventilation, heating, air-conditionin...
- clay, refractory, and other nonmetall...
- meat and seafood processing
- ship and boat building
- semiconductor and other electronic co...
- textile mills
- other electrical equipment and component
- electrical equipment
- cutlery, hardware, spring and wire: m...
- other fabricated metal product
- communications equipment
- other chemical product and preparation
- architectural and structural metals
- metalworking machinery
- plastics product
- industrial machinery
- agriculture, construction, and mining...
- wood product manufacturing
- basic chemical
- engine, turbine, and power transmissi...
- cement, concrete, lime, and gypsum pr...
- glass and glass product
- coating, engraving, heat treating, an...
- aerospace product and parts
- forging and stamping
- resin, synthetic rubber, and artifici...
- pulp, paper, and paperboard mills
- boiler, tank, and shipping container
- foundries
- nonferrous metal production and proce...
- petroleum refineries (including integ...
- iron, steel mills, and steel product

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Wholesale Trade

18 subsectors



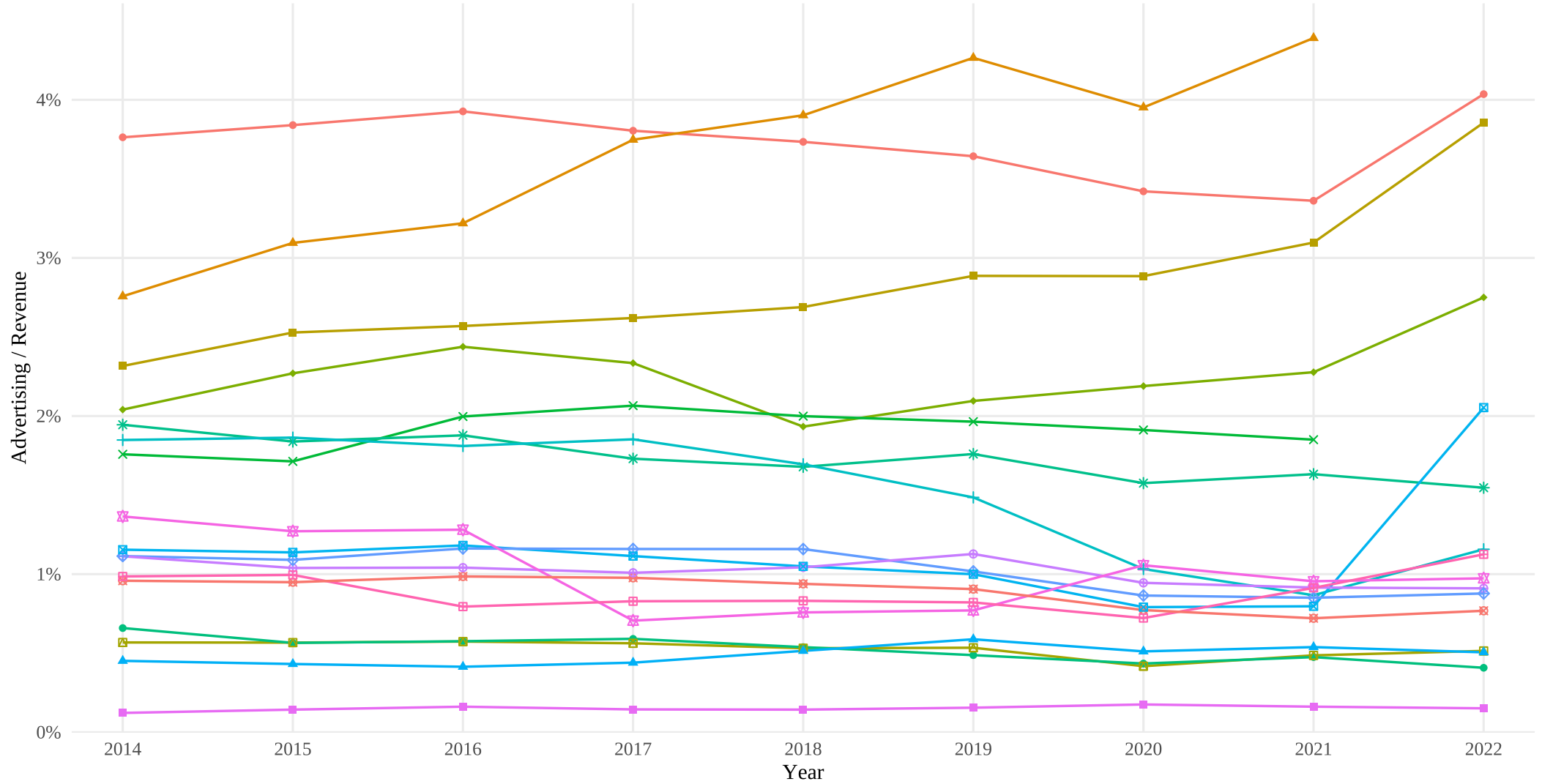
Subsector

- apparel, piece goods, and notions
- beer, wine, and distilled alcoholic b...
- motor vehicle and motor vehicle parts...
- furniture, sports, toys, recycle, jew...
- miscellaneous nondurable goods
- electrical and electronic goods
- drugs and druggists' sundries
- professional and commercial equipment...
- hardware, plumbing, heating equipment...
- wholesale electronic markets and agen...
- grocery and related product
- machinery, equipment, and supplies
- lumber and other construction materials
- paper and paper product
- chemical and allied products
- farm product raw material
- metal and mineral (except petroleum)
- petroleum and petroleum products

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Retail Trade

17 subsectors



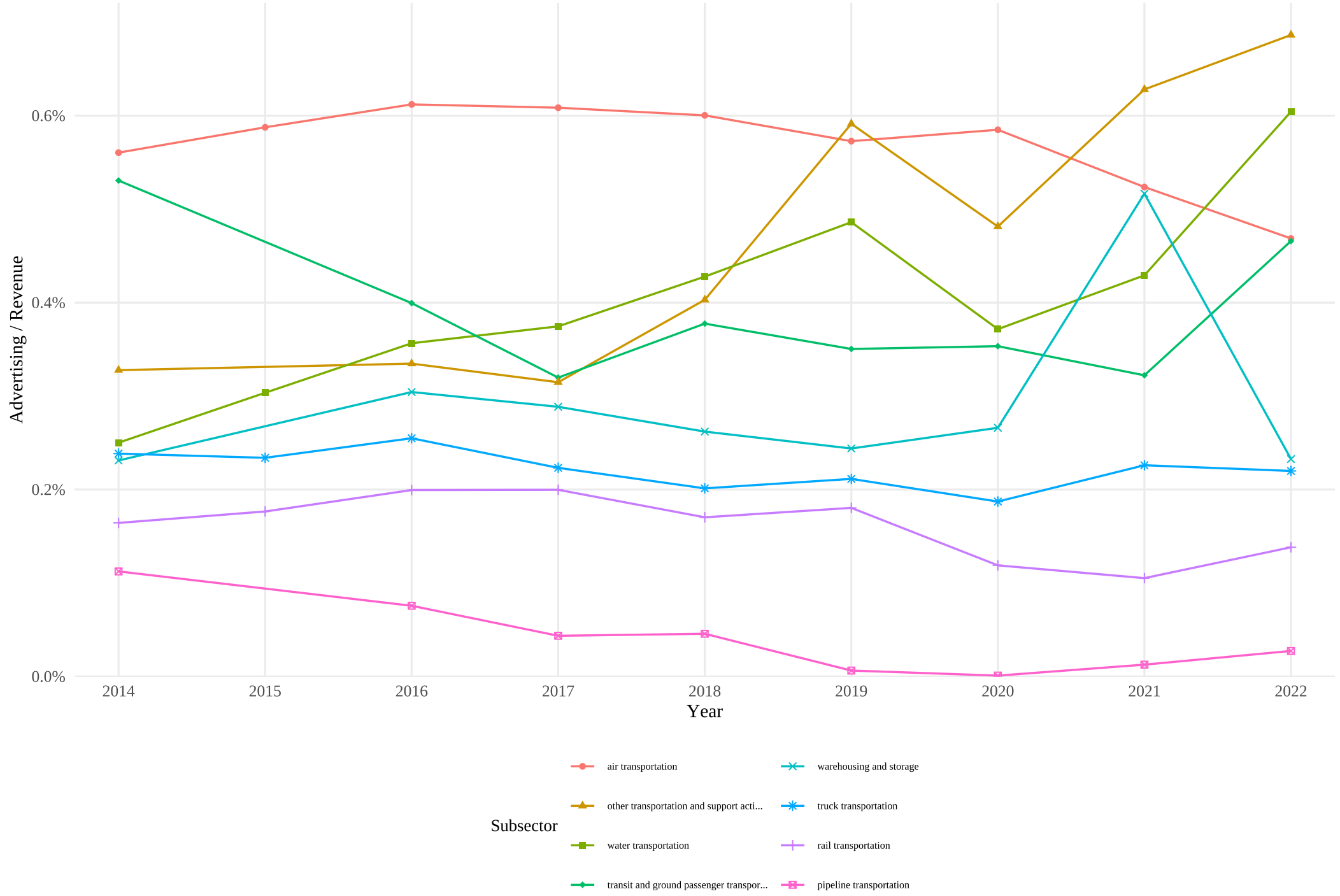
Subsector

- furniture and home furnishings stores
- nonstore retailers
- clothing and clothing accessories stores
- sporting goods, hobby, book, and musi...
- miscellaneous store retailers
- electronics and appliance stores
- hardware stores
- general merchandise stores
- lawn and garden equipment and supplie...
- other motor vehicle and parts dealers
- home centers; paint and wallpaper stores
- health and personal care stores
- new and used car dealers
- food and beverage stores
- other building materials dealers
- beer, wine, and liquor stores
- gasoline stations

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Transportation

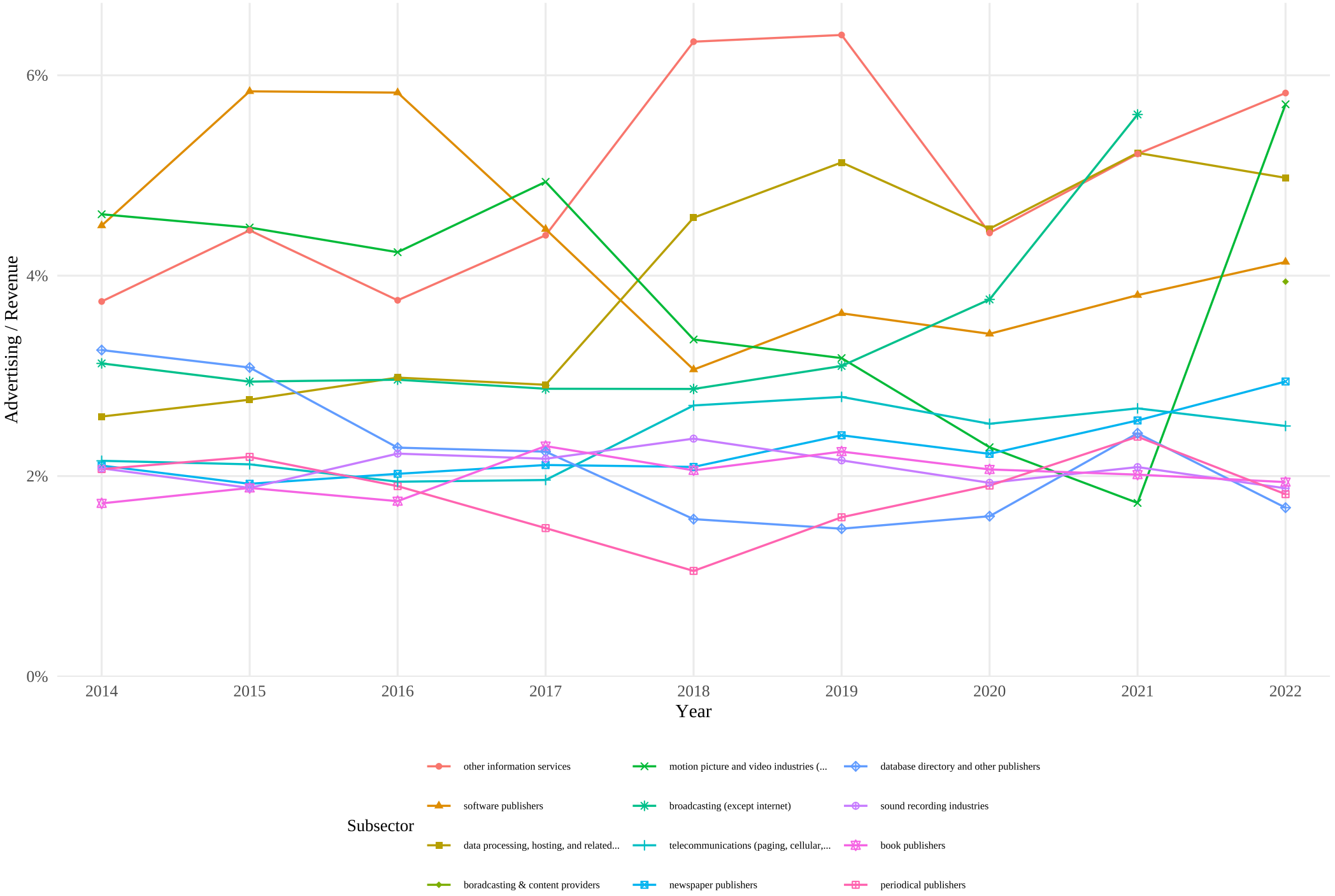
8 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Information

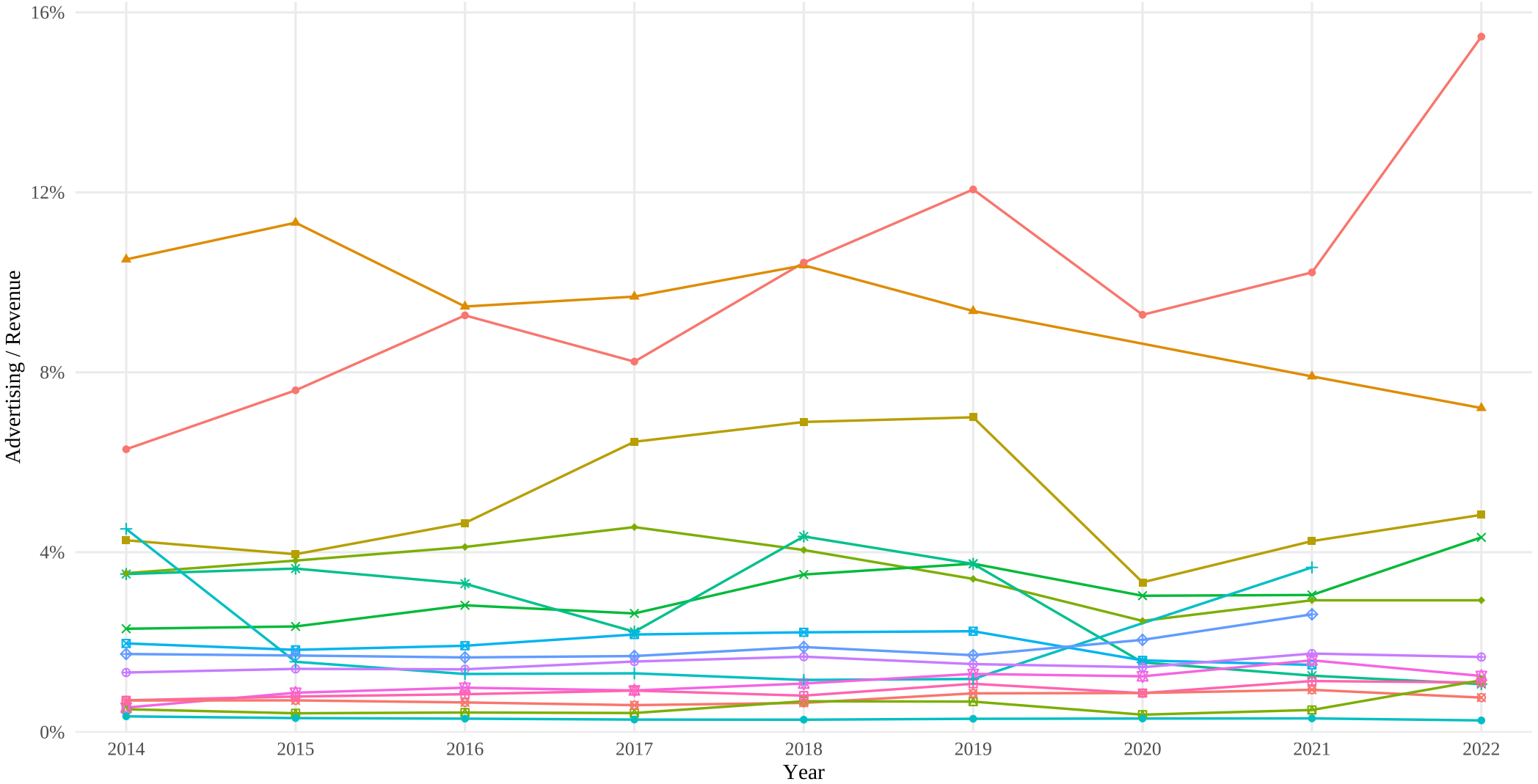
12 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Finance & Insurance

16 subsectors



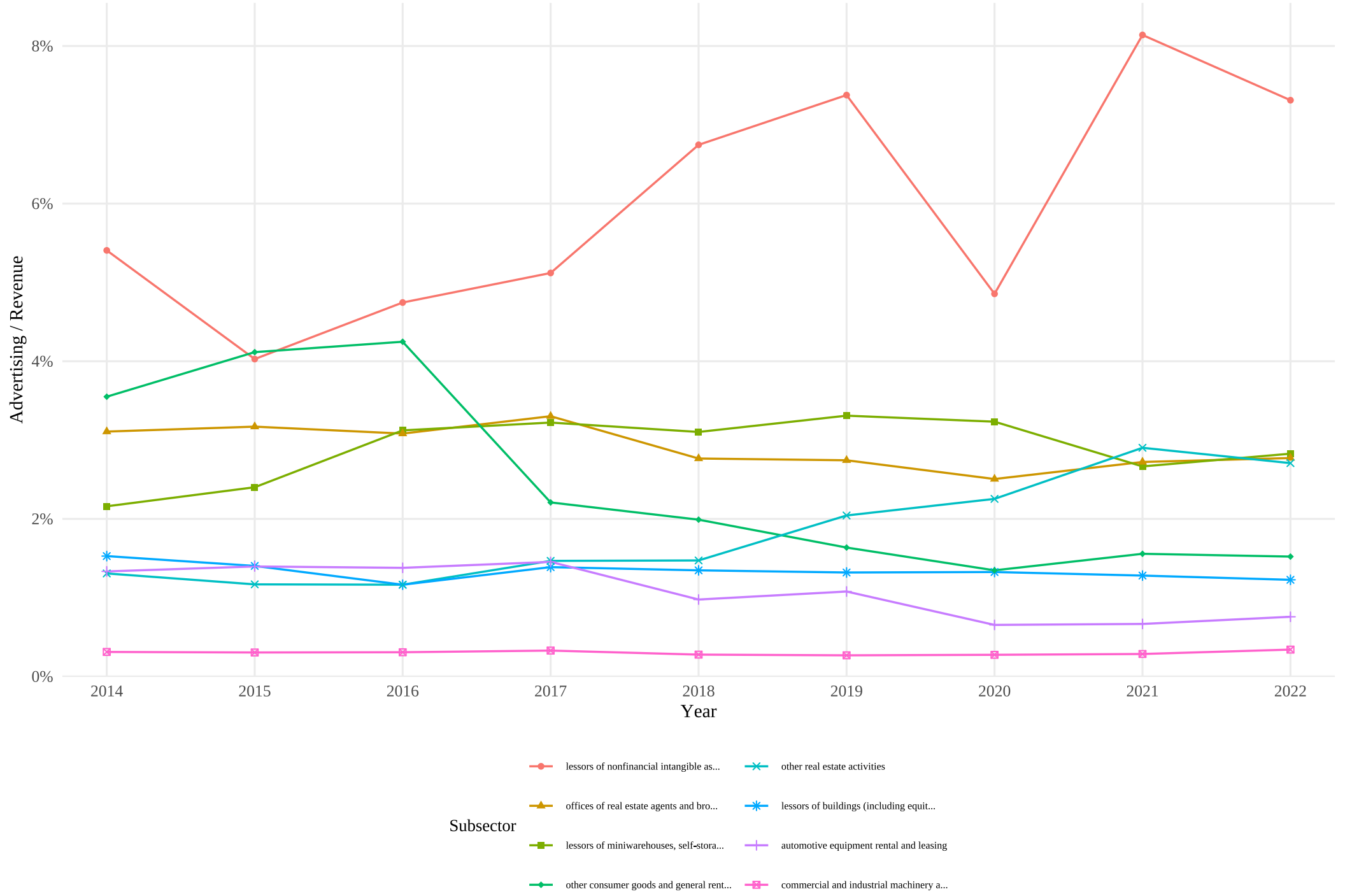
Subsector

- credit card issuing and other consume...
- savings institutions, credit unions, ...
- real estate credit (including mortgag...
- activities related to credit intermed...
- commercial banking
- international, secondary financing, a...
- other financial vehicles (including m...
- securities brokerage
- insurance agencies and brokerages
- securities, commodity exchanges, and ...
- commodity contracts dealing and broke...
- other insurance related activities (t...
- property and casualty insurance
- investment banking and securities dea...
- life insurance

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Real Estate

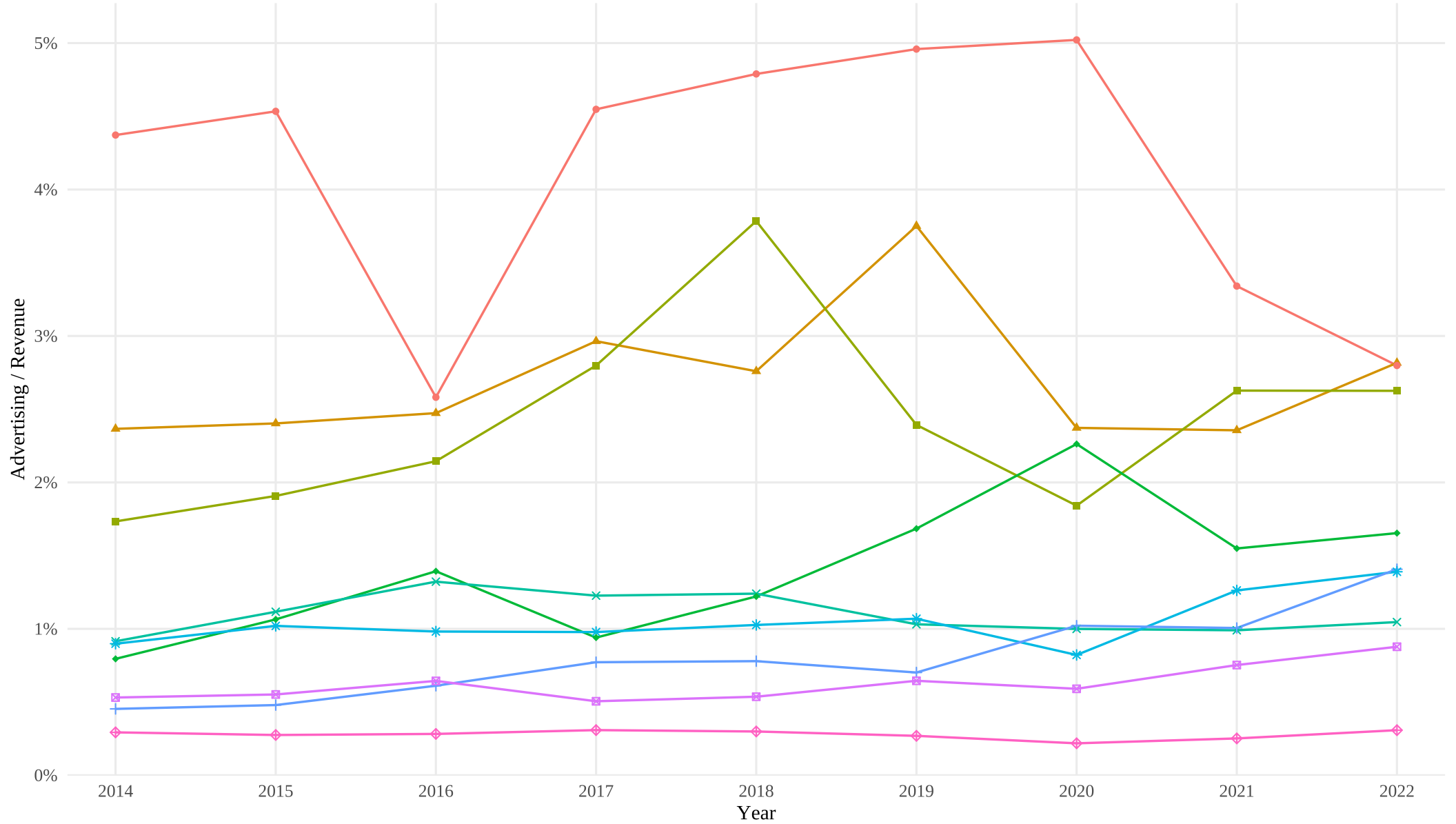
8 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Professional Services

9 subsectors



Subsector

- advertising and related services
- legal services
- scientific research and development s...
- other professional, scientific, and t...
- accounting, tax preparation, bookkeep...
- computer systems design and related s...
- specialized design services
- management, scientific, and technical...
- architectural, engineering, and relat...

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

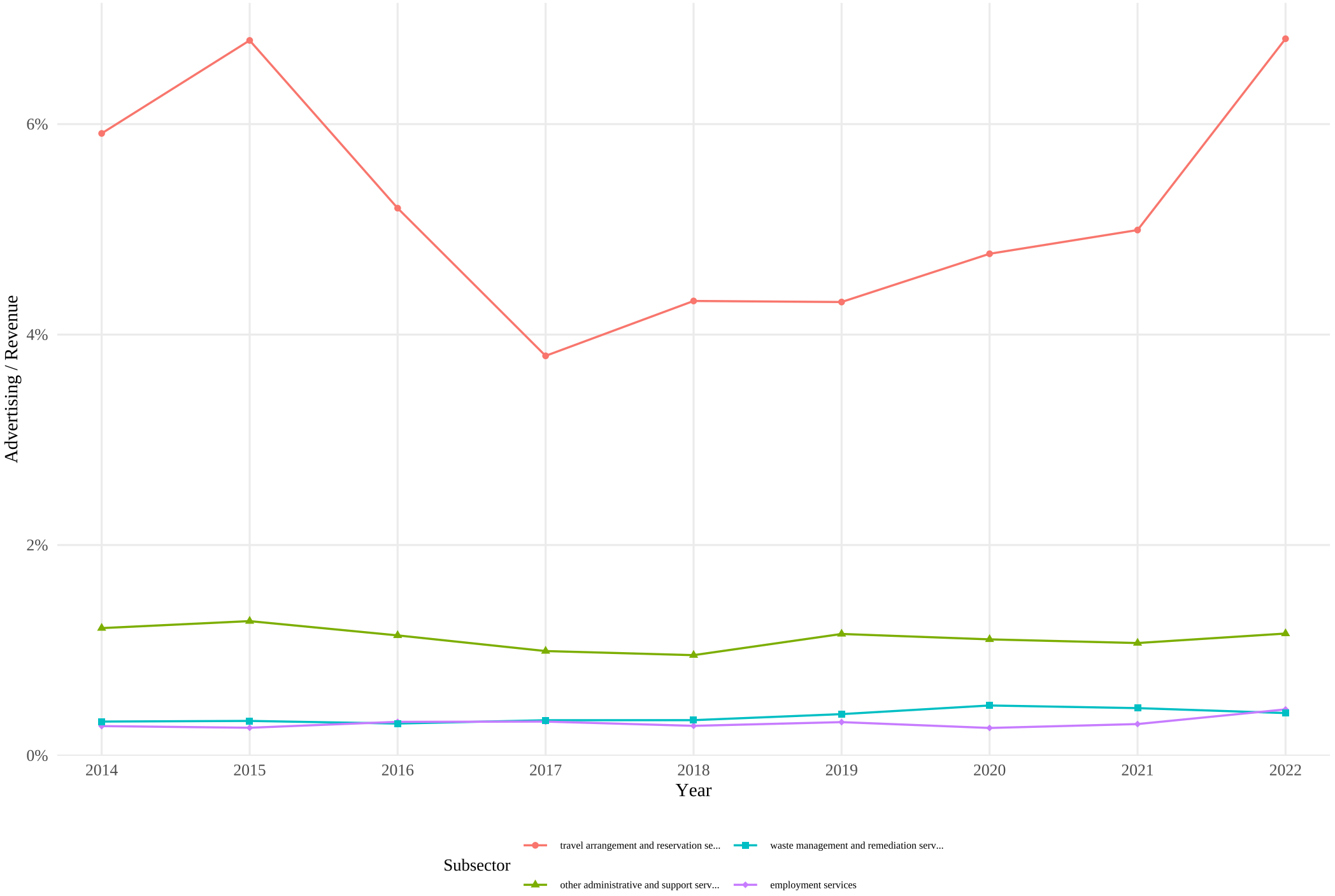
Ad/Revenue: Management of Companies

2 subsectors

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Administrative Services

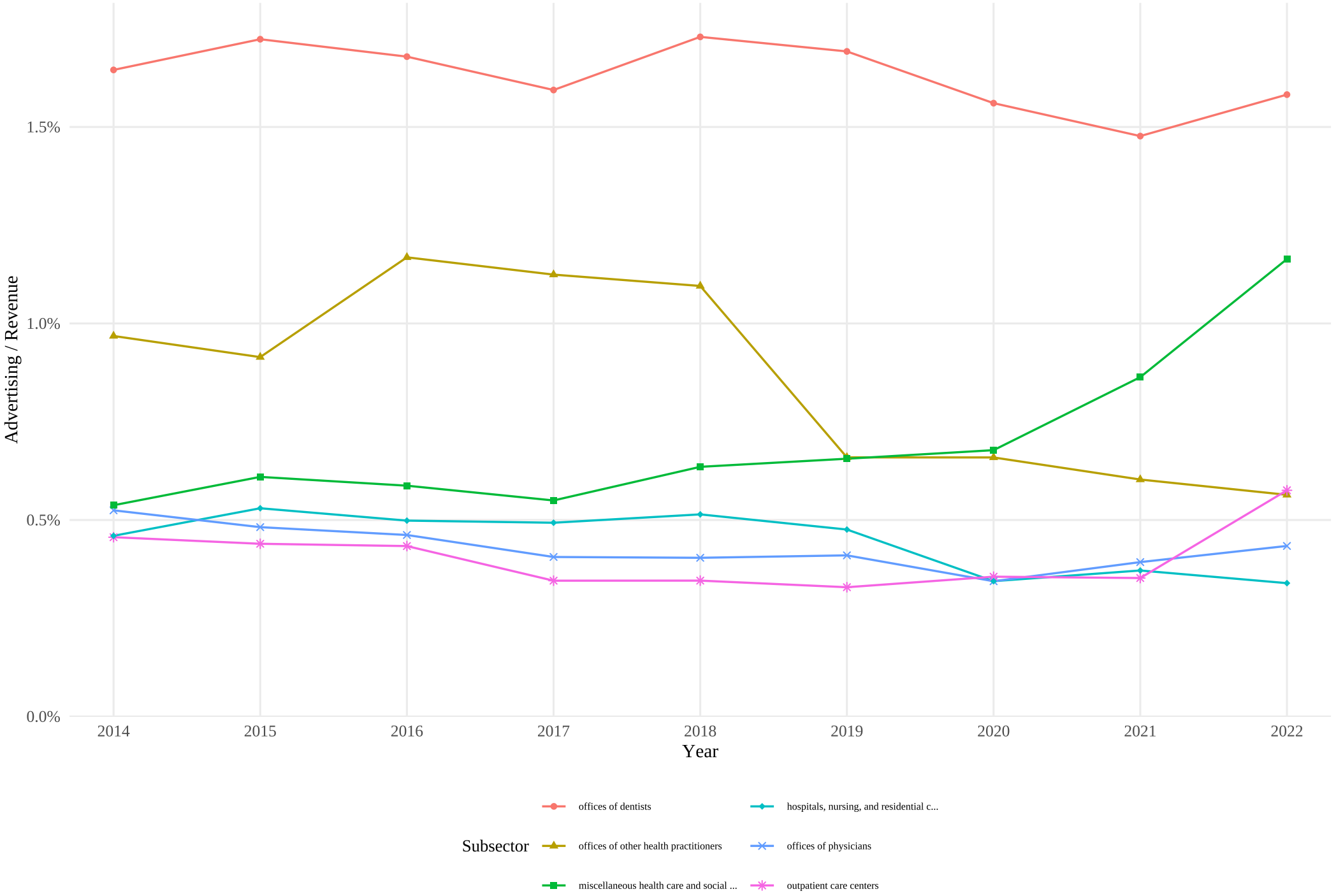
4 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Health Care

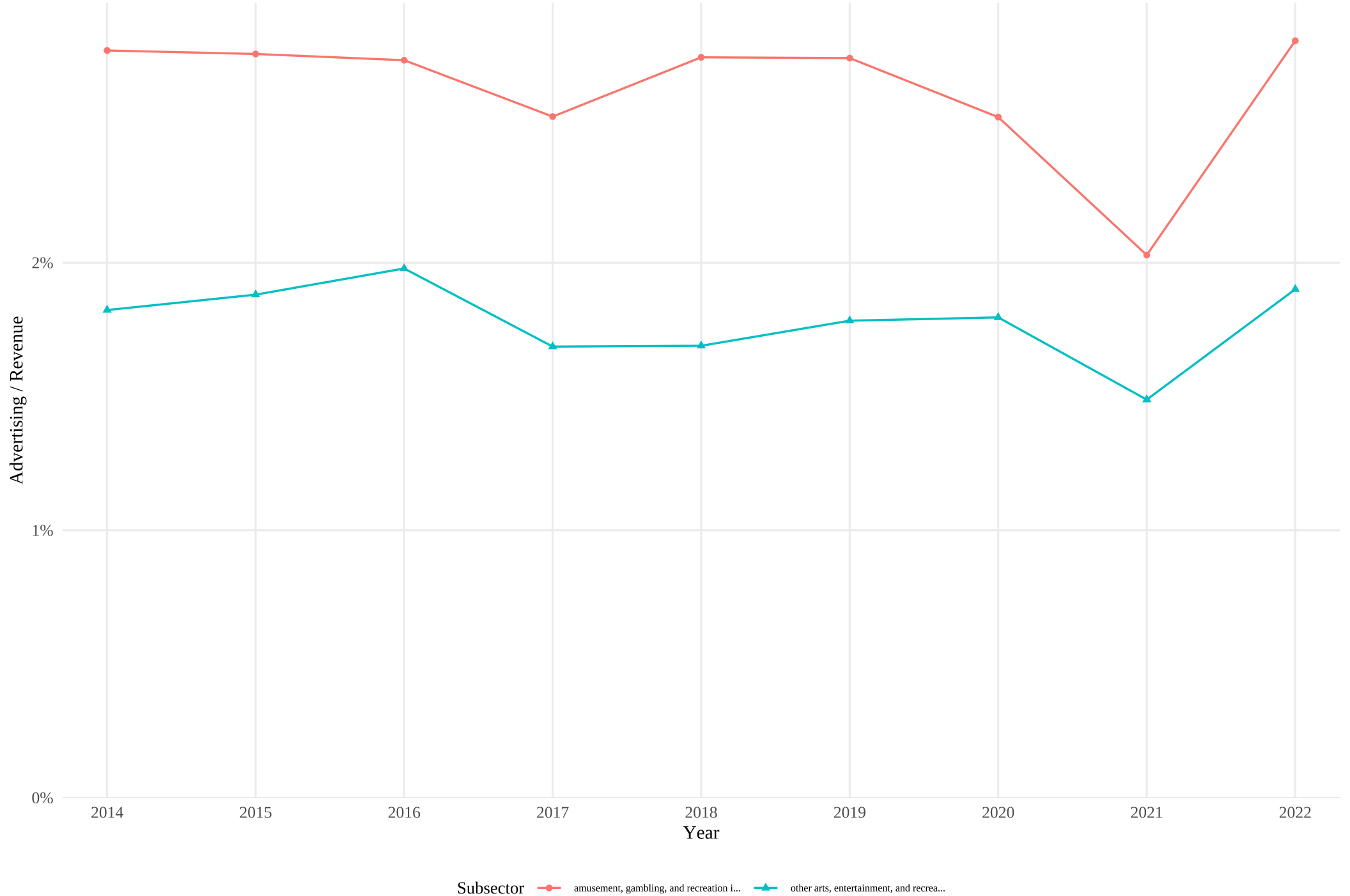
6 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Arts & Entertainment

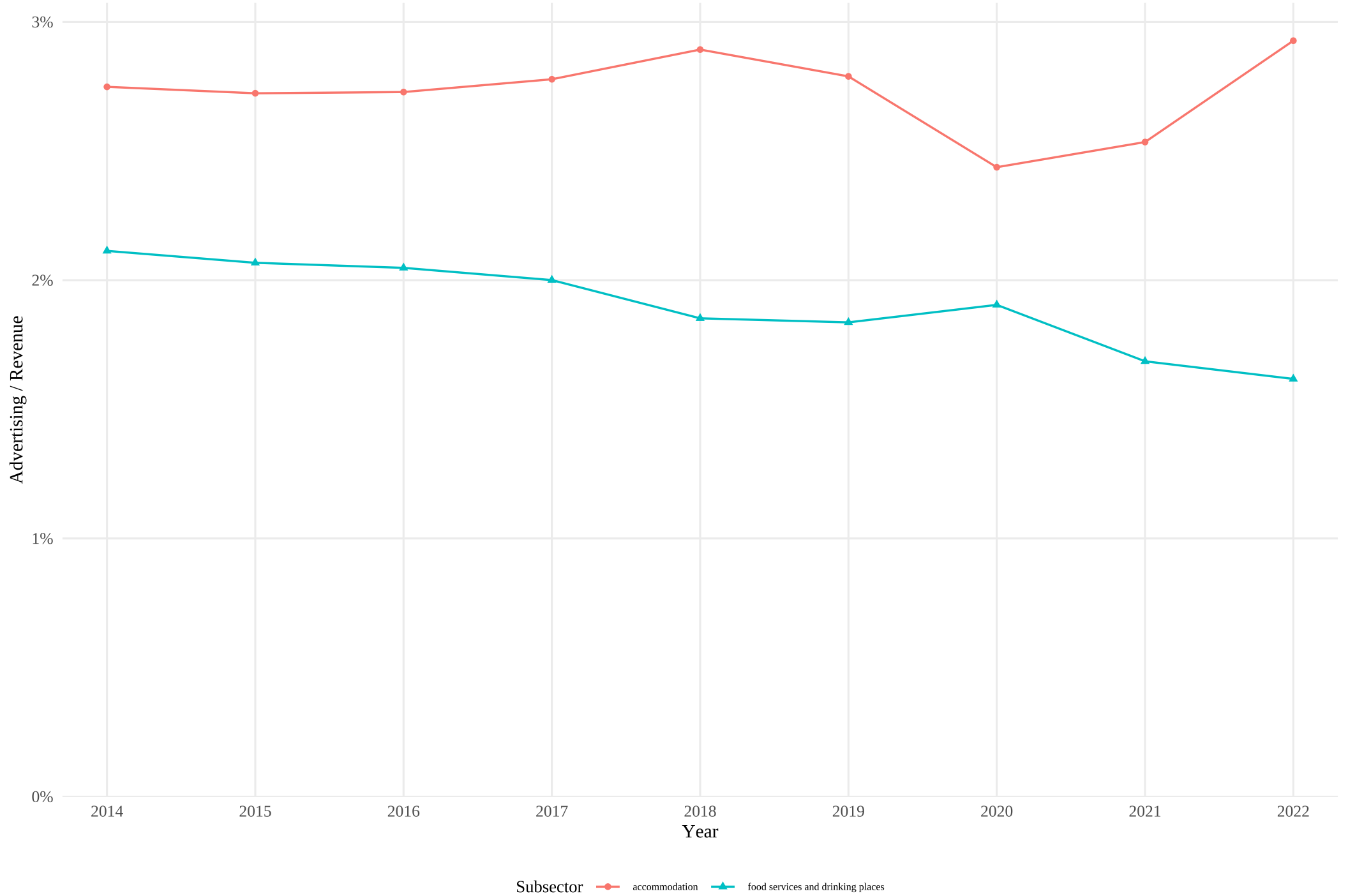
2 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Accommodation & Food

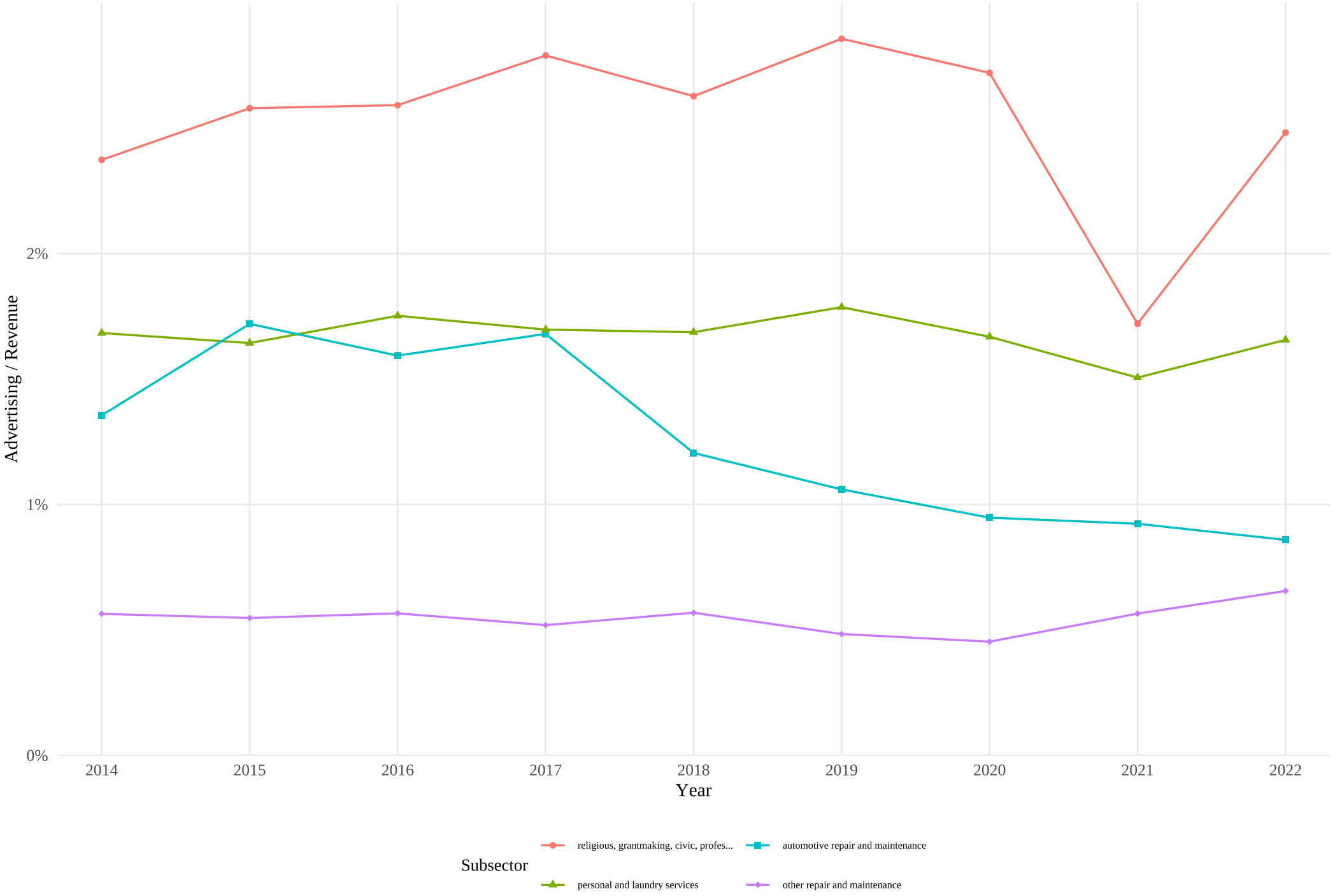
2 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Revenue: Other Services

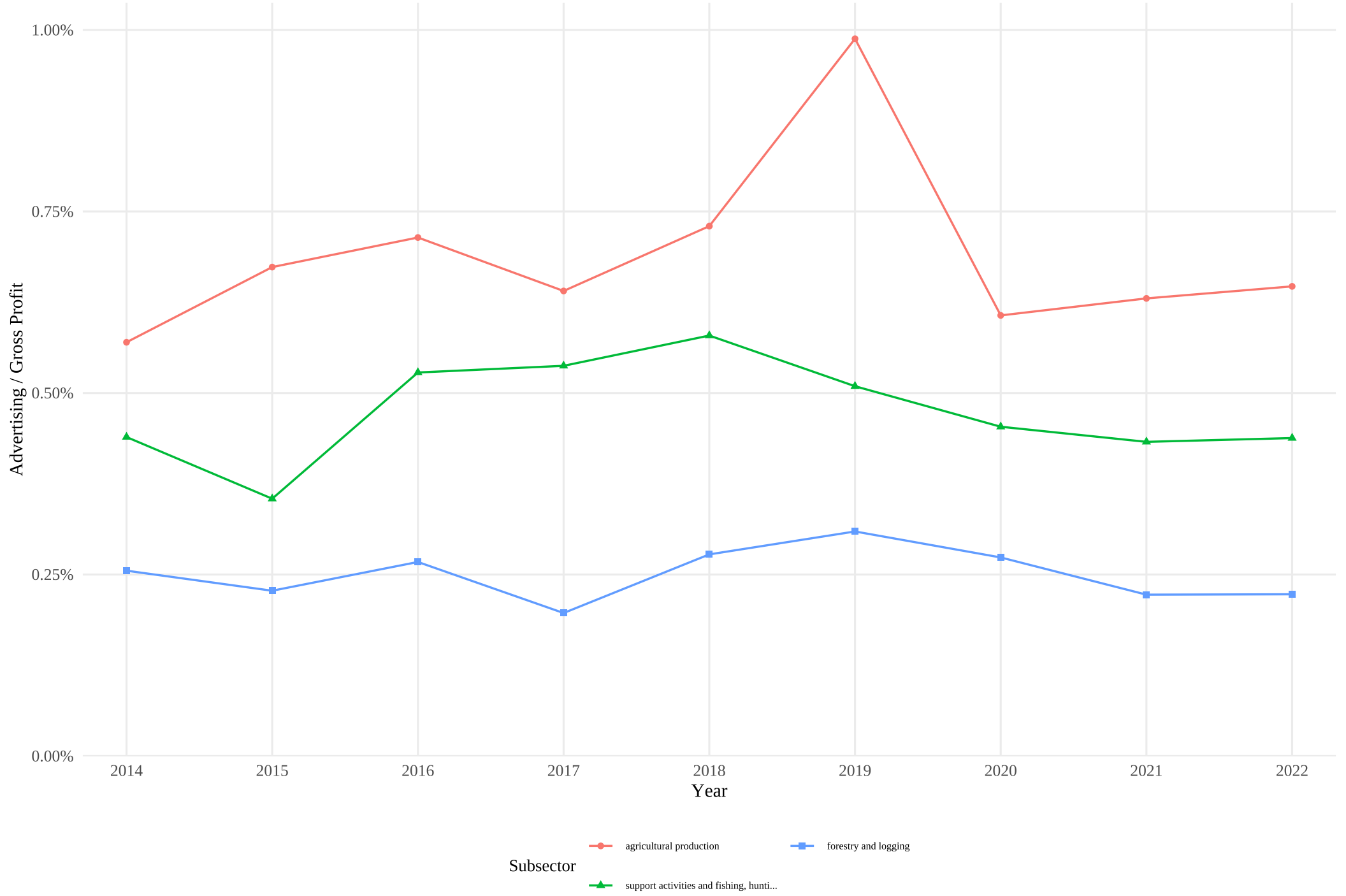
4 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Agriculture

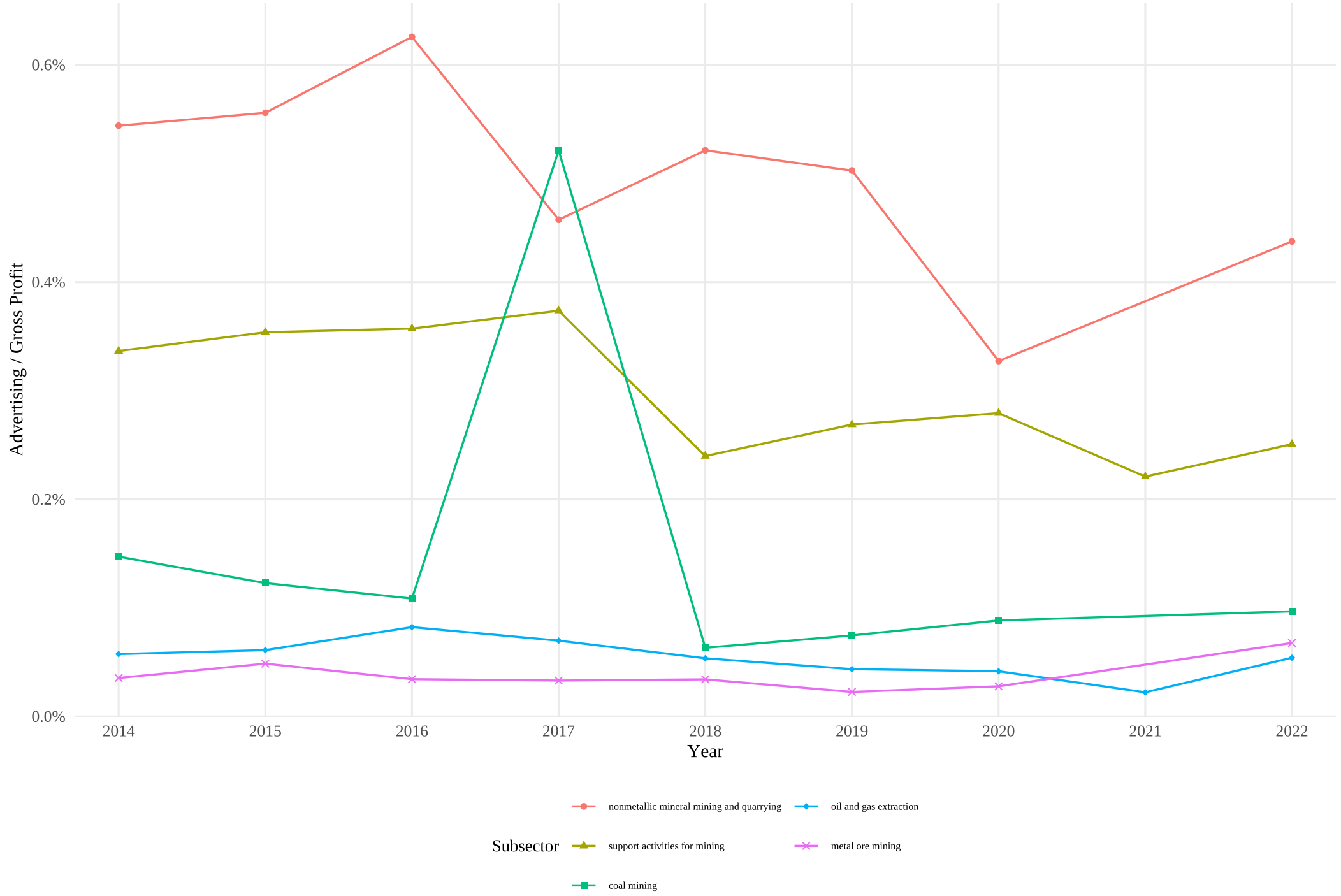
3 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Mining

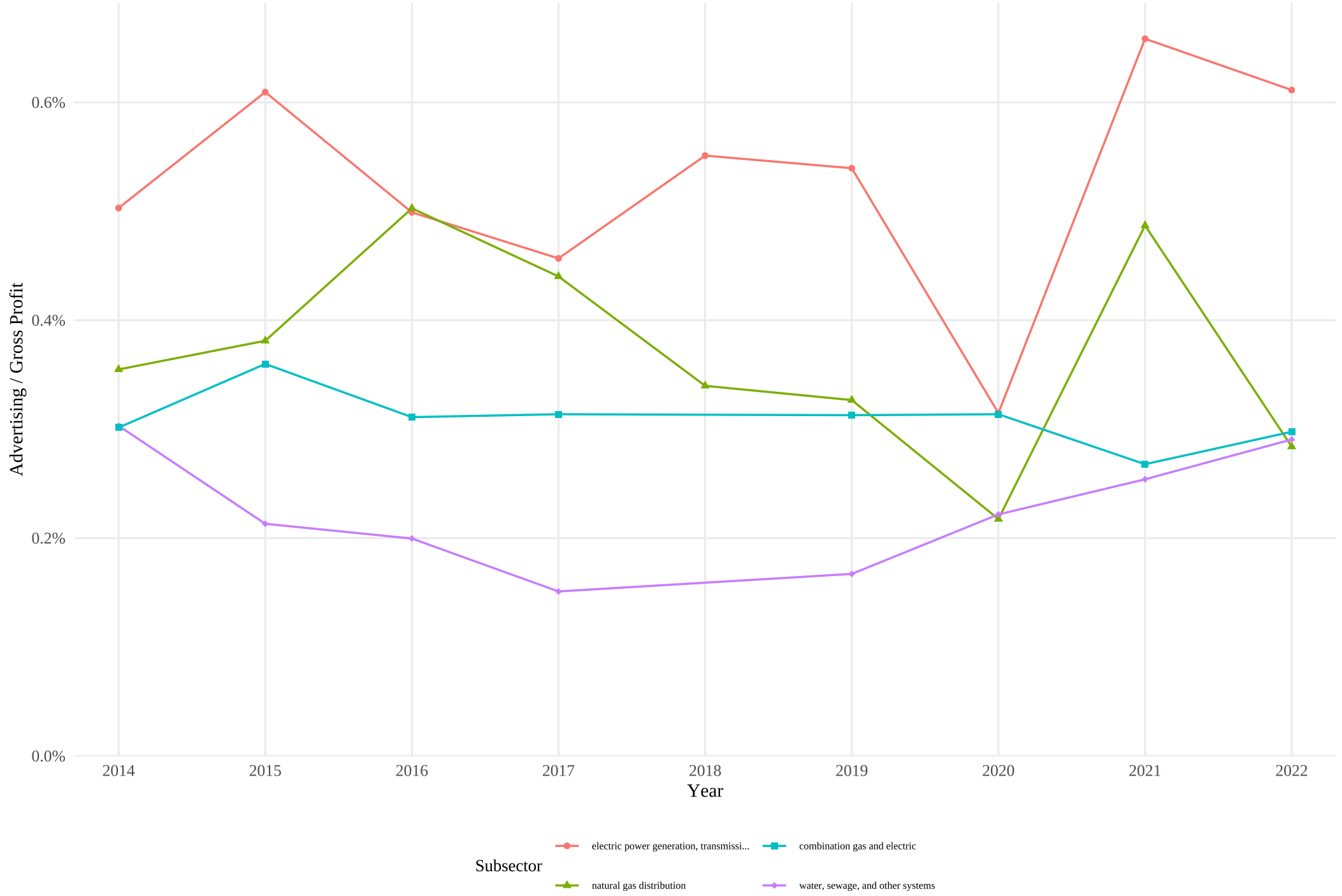
5 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Utilities

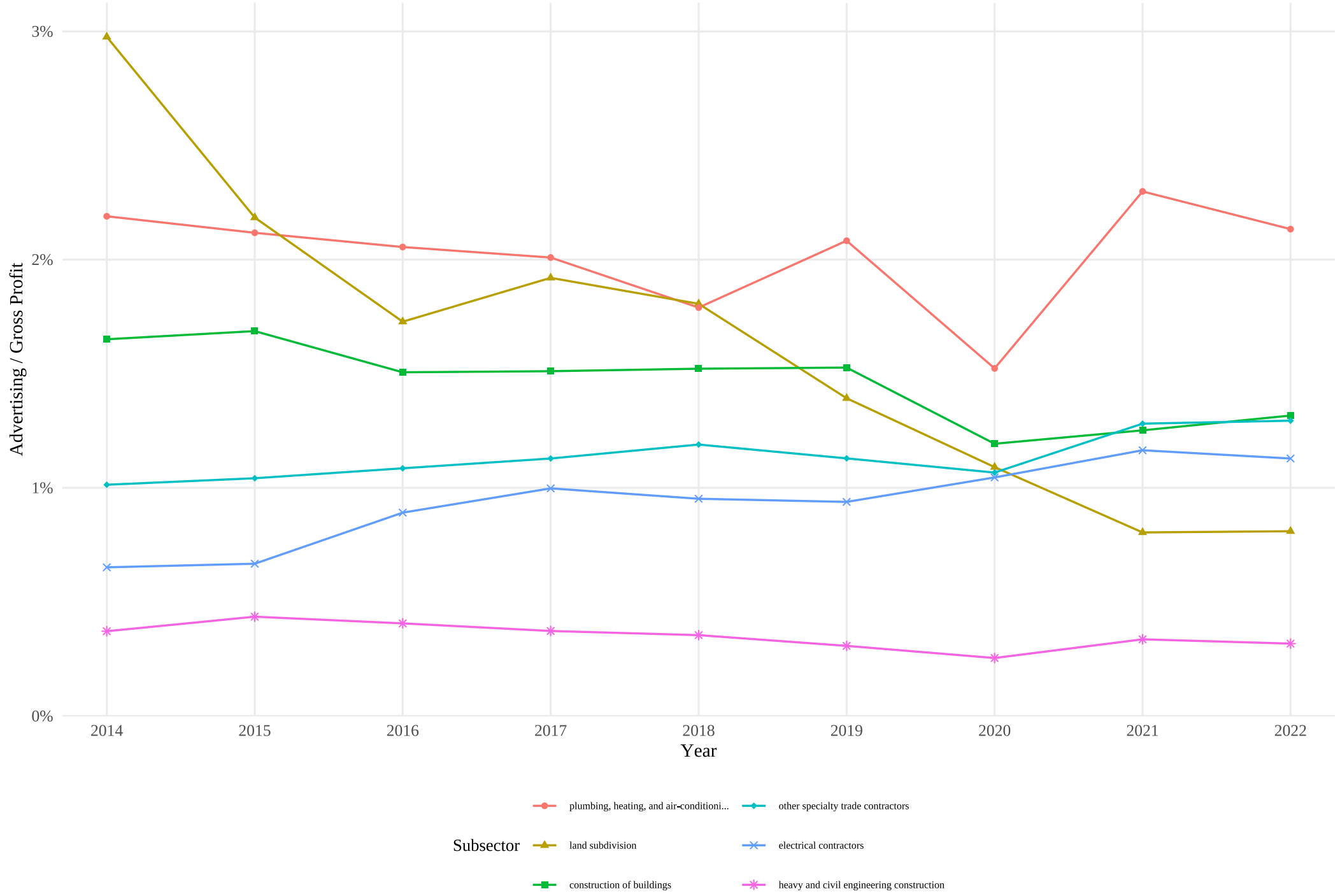
4 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Construction

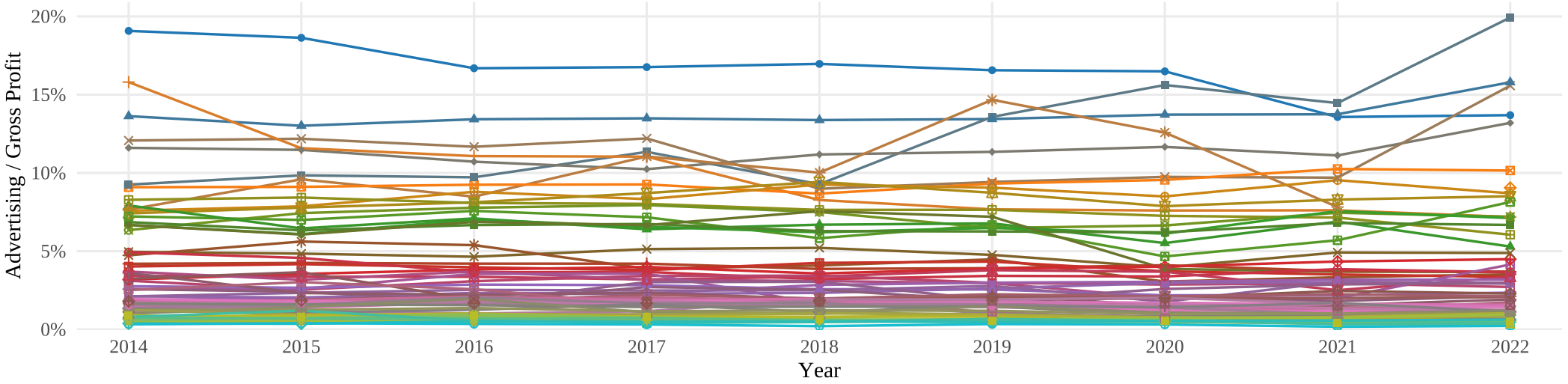
6 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Manufacturing

66 subsectors



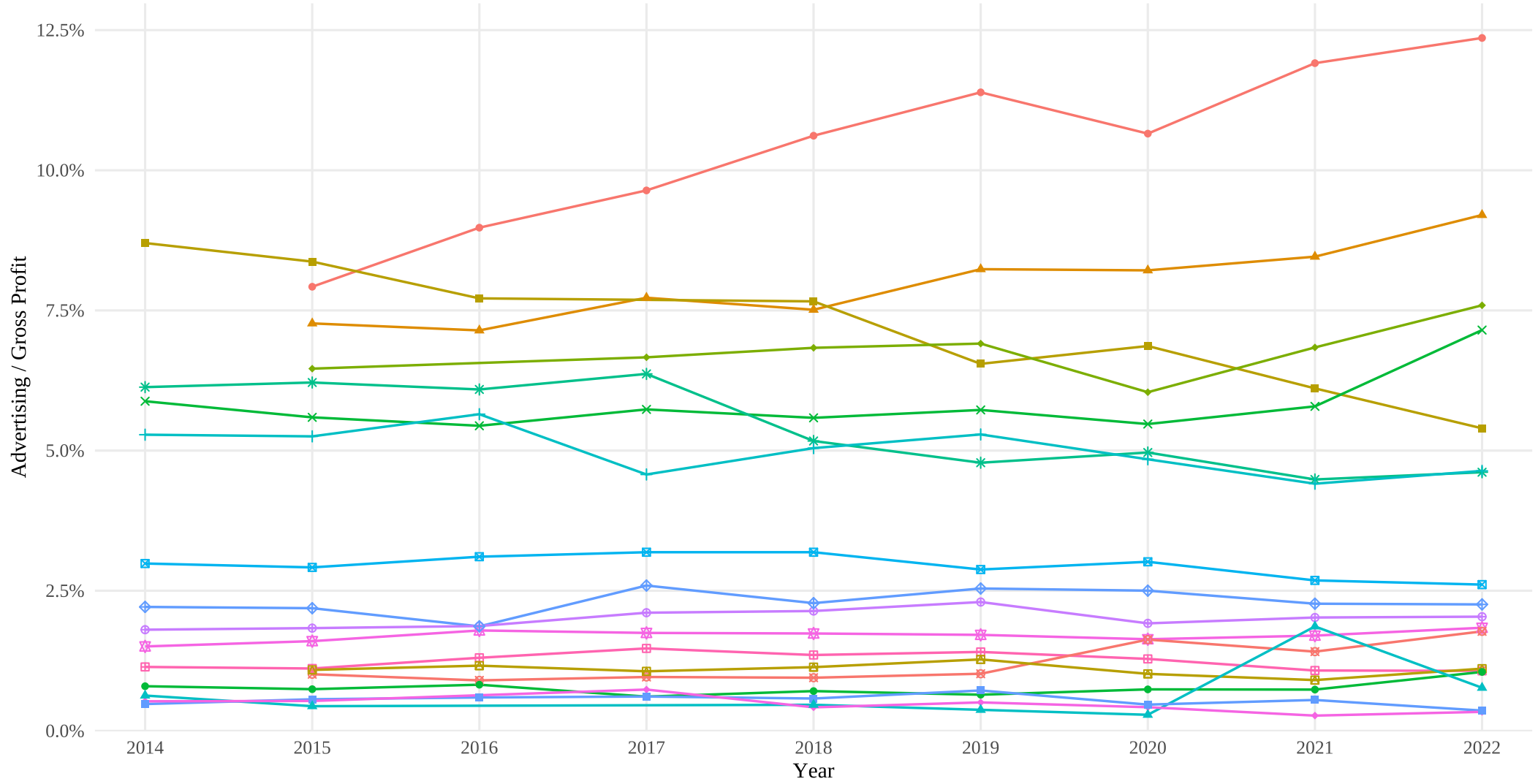
Subsector

- animal food manufacturing and grain a...
- dairy product
- other general purpose machinery
- electrical equipment
- forging and stamping
- soap, cleaning compound, and toilet p...
- fruit and vegetable preserving and sp...
- ventilation, heating, air-conditionin...
- cutlery, hardware, spring and wire: m...
- nonferrous metal production and proce...
- apparel knitting mills
- cut and sew apparel contractors and m...
- converted paper product
- other chemical product and preparation
- aerospace product and parts
- leather and allied product manufacturing
- audio and video equipment manufacturi...
- clay, refractory, and other nonmetall...
- wood product manufacturing
- coating, engraving, heat treating, an...
- wineries and distilleries
- other miscellaneous manufacturing
- textile product mills
- agriculture, construction, and mining...
- petroleum refineries (including integ...
- apparel accessories and other apparel
- meat and seafood processing
- printing and related support activities
- basic chemical
- boiler, tank, and shipping container
- sugar and confectionery product
- paint, coating, and adhesive
- ship and boat building
- plastics product
- pulp, paper, and paperboard mills
- other food
- rubber product
- other electrical equipment and component
- metalworking machinery
- resin, synthetic rubber, and artifici...
- apparel accessories and other apparel...
- computer and peripheral equipment
- medical equipment and supplies
- tobacco manufacturing
- foundries
- breweries
- electrical lighting equipment and hou...
- navigational, measuring, electromedic...
- engine, turbine, and power transmissi...
- iron, steel mills, and steel product
- pharmaceutical and medicine
- other transportation equipment and ra...
- commercial and service industry machi...
- communications equipment
- motor vehicles and parts
- bakeries and tortilla manufacturing
- semiconductor and other electronic co...
- industrial machinery
- soft drink and ice
- asphalt paving, roofing, other petrol...
- other fabricated metal product
- cement, concrete, lime, and gypsum pr...
- furniture and related product manufac...
- textile mills
- architectural and structural metals
- glass and glass product

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Wholesale Trade

18 subsectors



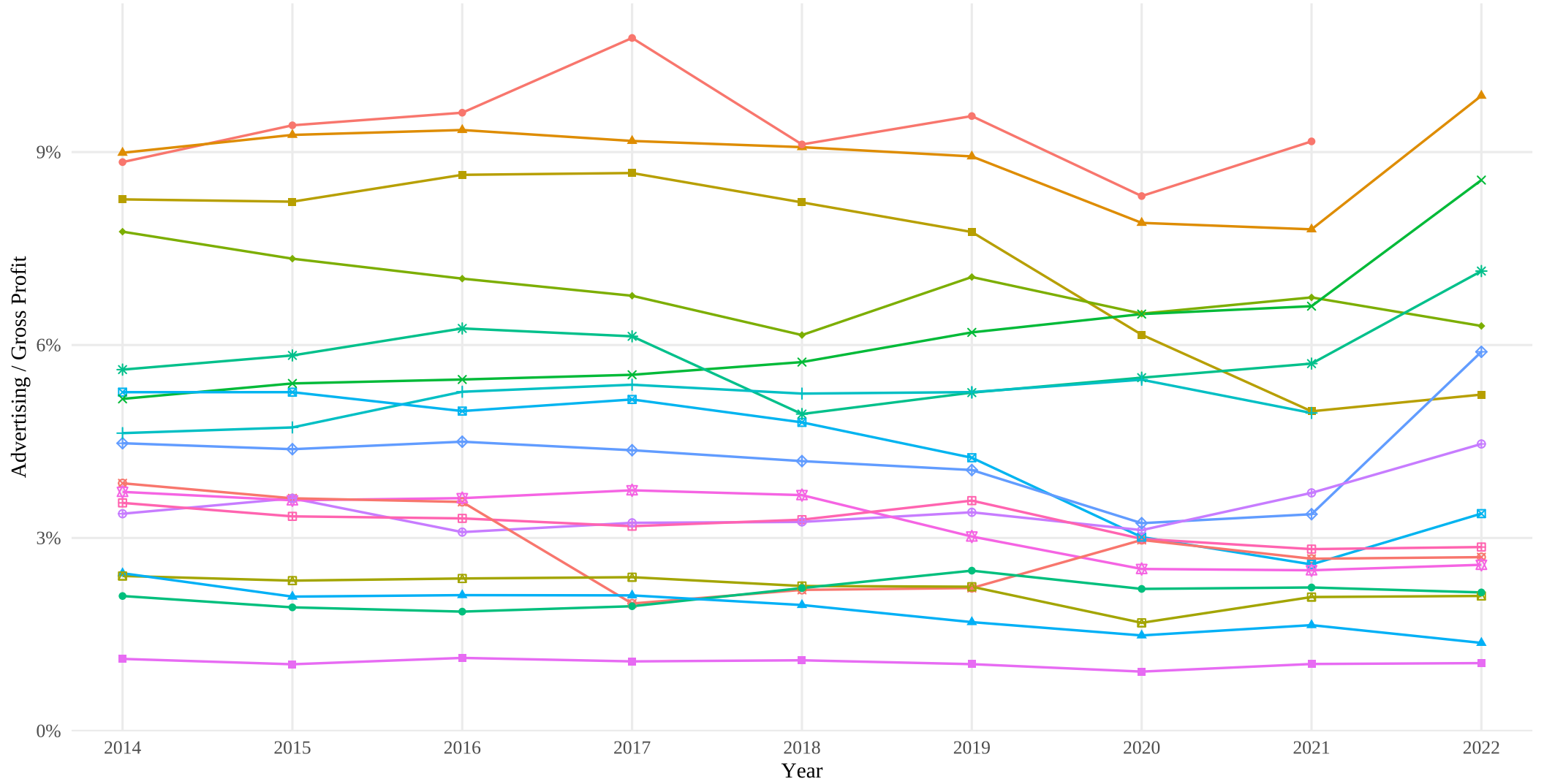
Subsector

- beer, wine, and distilled alcoholic b...
- apparel, piece goods, and notions
- motor vehicle and motor vehicle parts...
- drugs and druggists' sundries
- furniture, sports, toys, recycle, jew...
- electrical and electronic goods
- miscellaneous nondurable goods
- professional and commercial equipment...
- grocery and related product
- hardware, plumbing, heating equipment...
- machinery, equipment, and supplies
- lumber and other construction materials
- paper and paper product
- farm product raw material
- chemical and allied products
- wholesale electronic markets and agen...
- petroleum and petroleum products
- metal and mineral (except petroleum)

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Retail Trade

17 subsectors



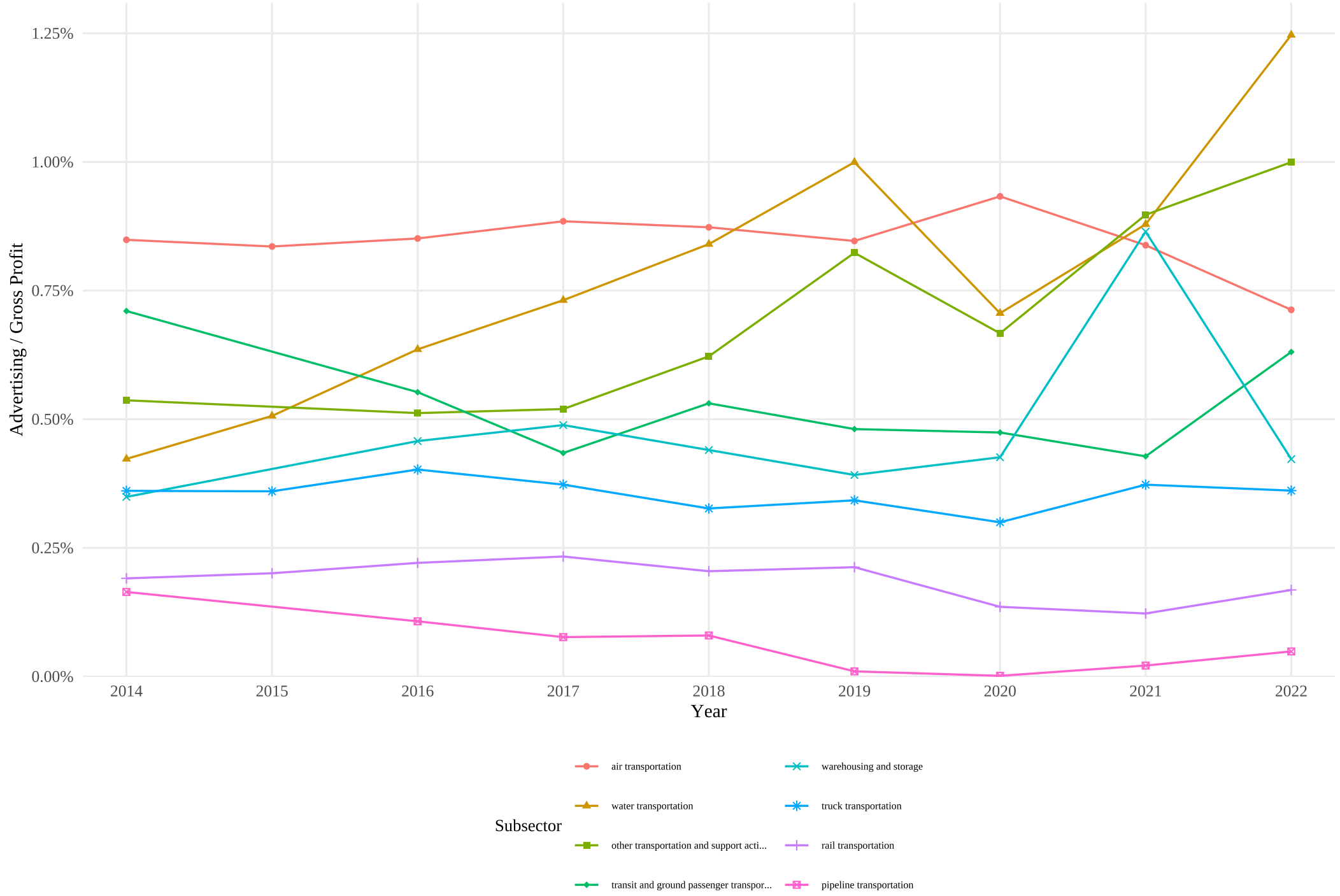
Subsector

- nonstore retailers
- furniture and home furnishings stores
- new and used car dealers
- electronics and appliance stores
- clothing and clothing accessories stores
- sporting goods, hobby, book, and musi...
- miscellaneous store retailers
- hardware stores
- general merchandise stores
- health and personal care stores
- lawn and garden equipment and supplie...
- other motor vehicle and parts dealers
- home centers; paint and wallpaper stores
- food and beverage stores
- beer, wine, and liquor stores
- other building materials dealers
- gasoline stations

Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Transportation

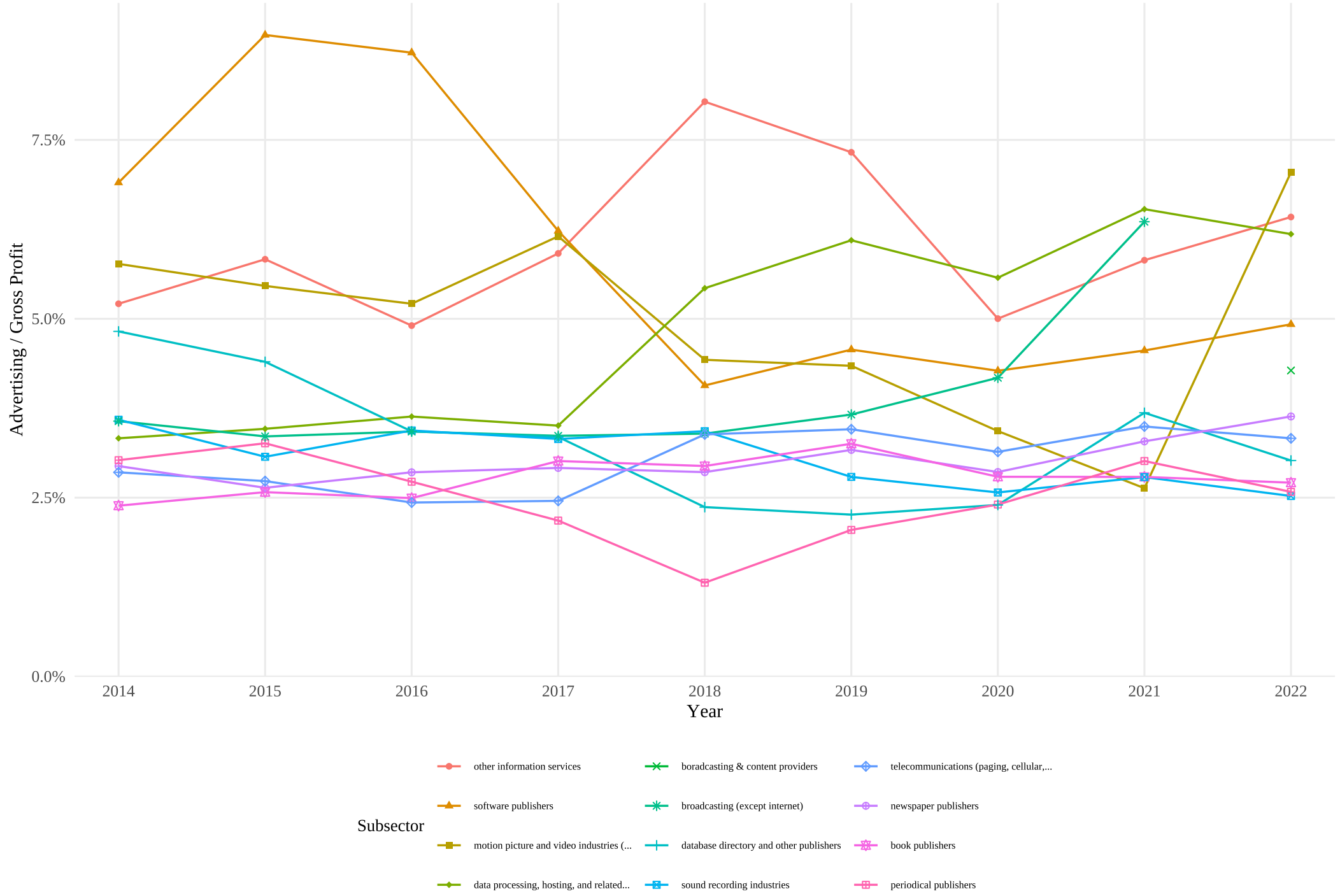
8 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Information

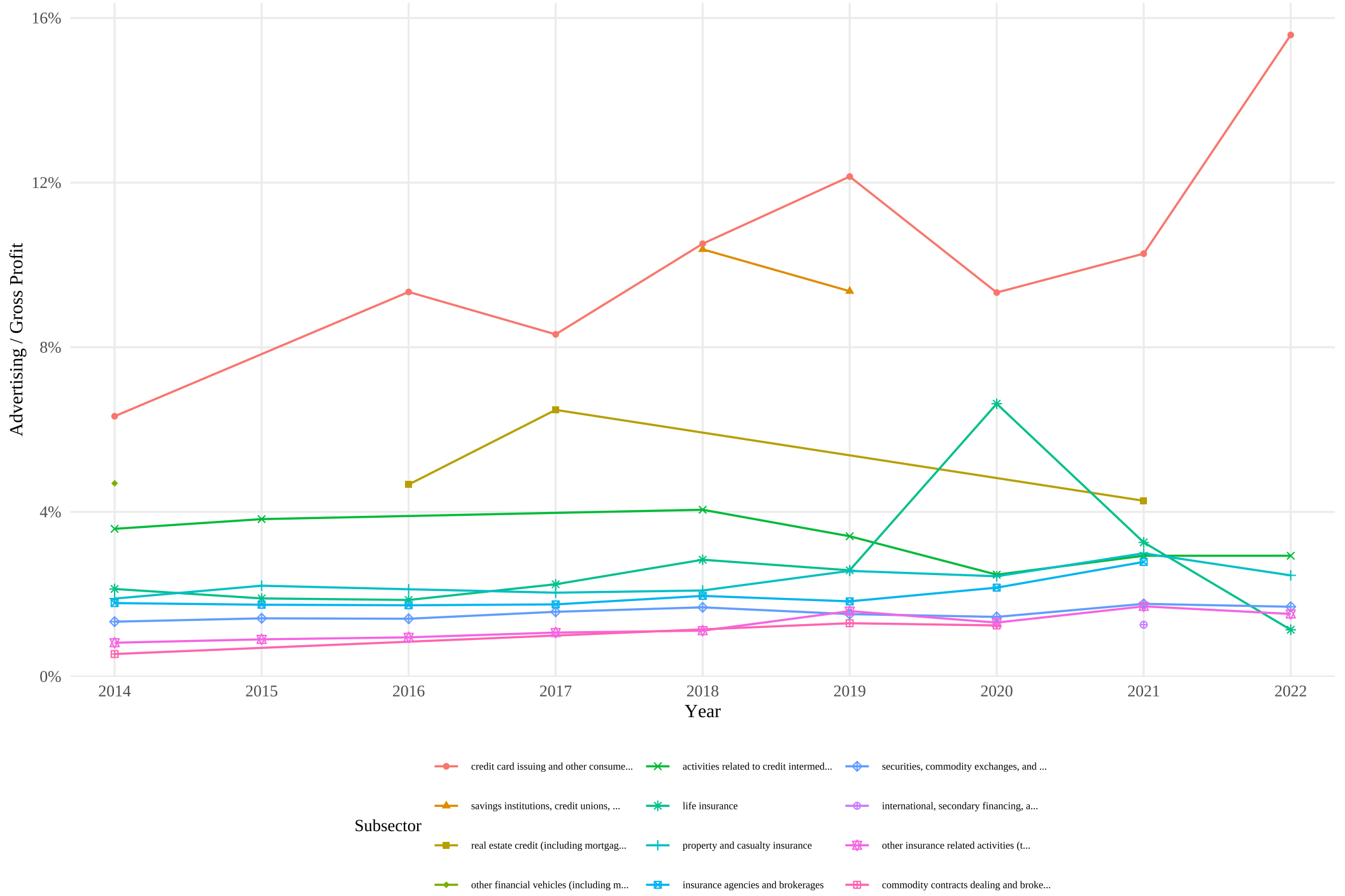
12 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Finance & Insurance

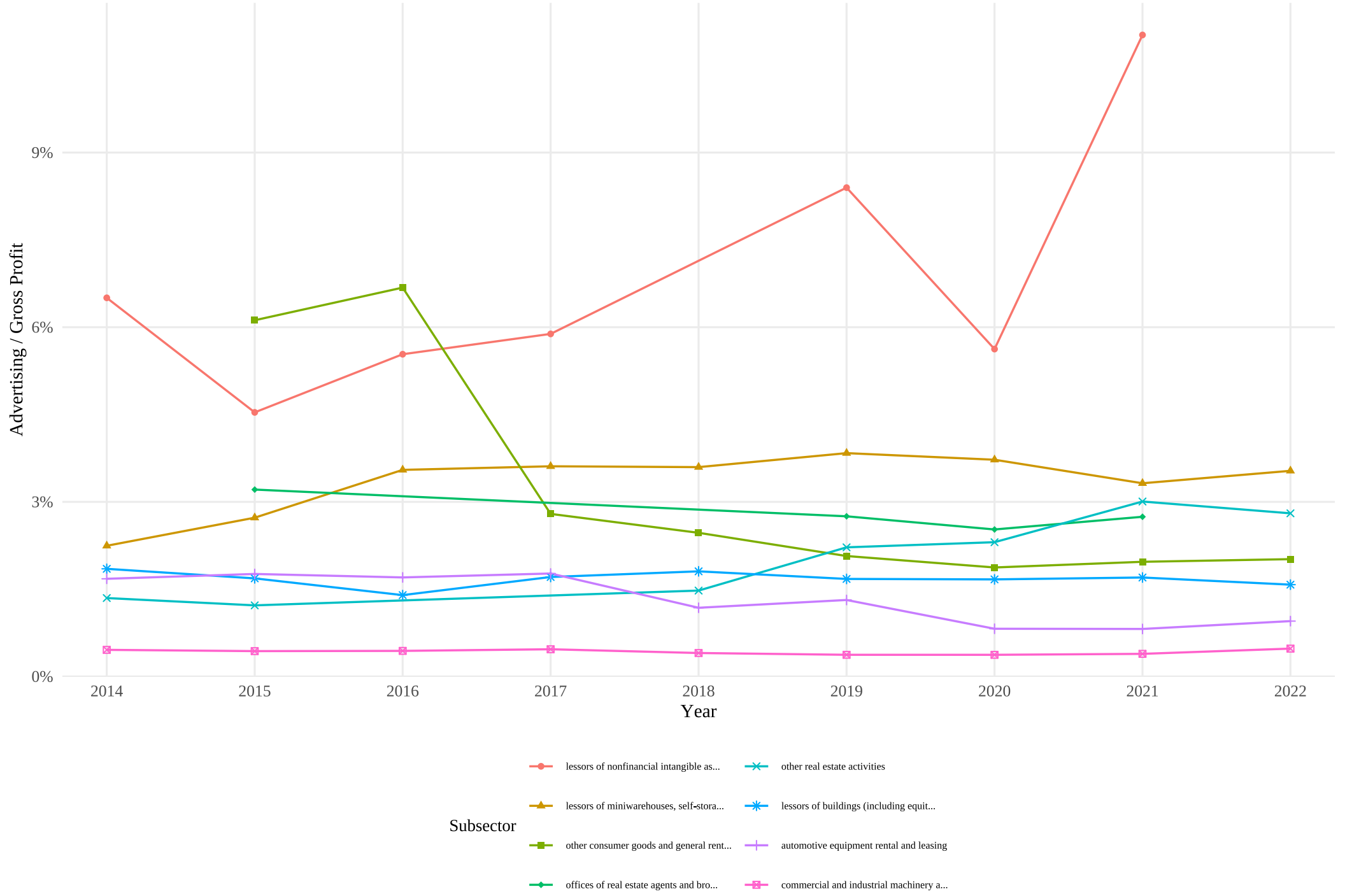
13 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Real Estate

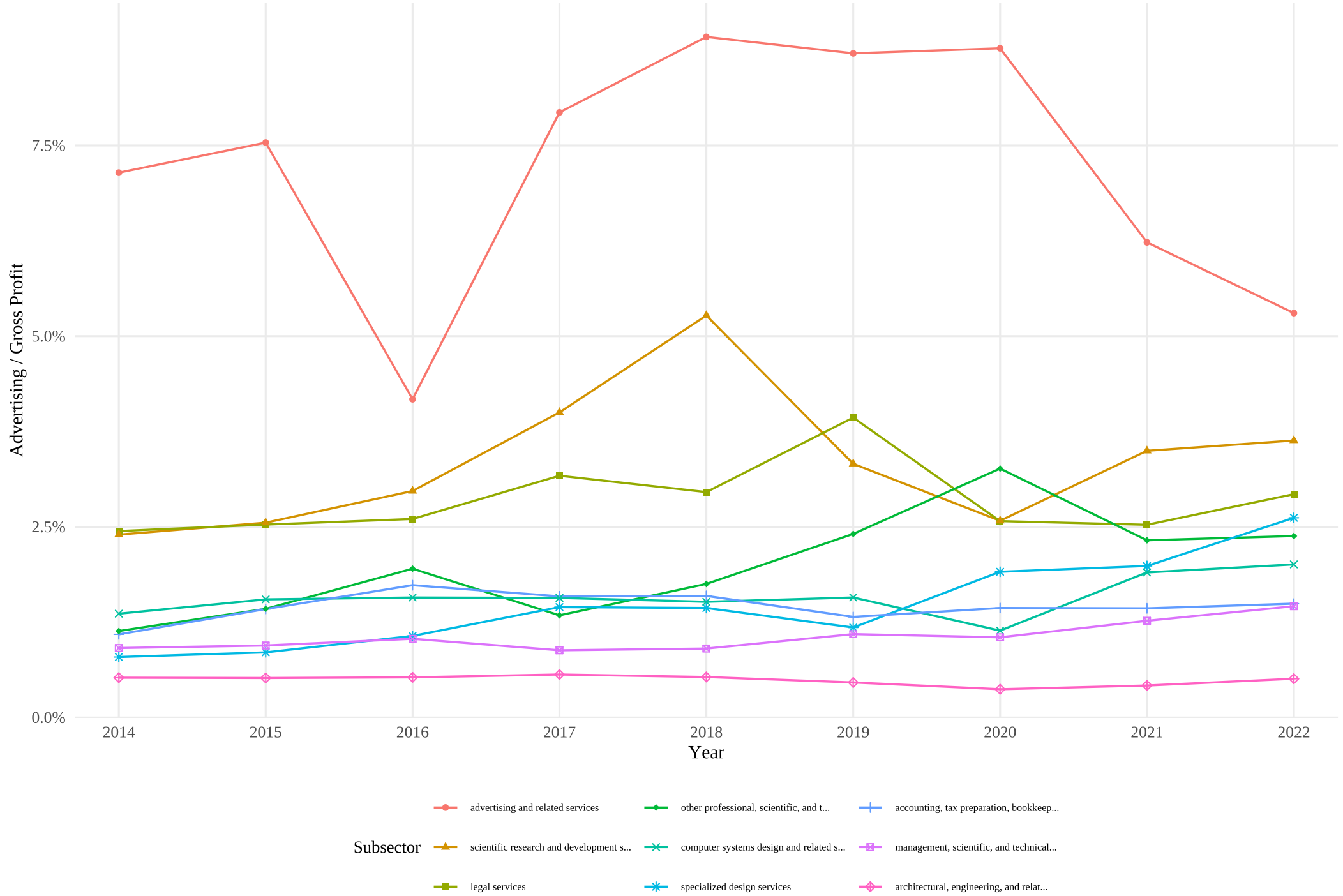
8 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Professional Services

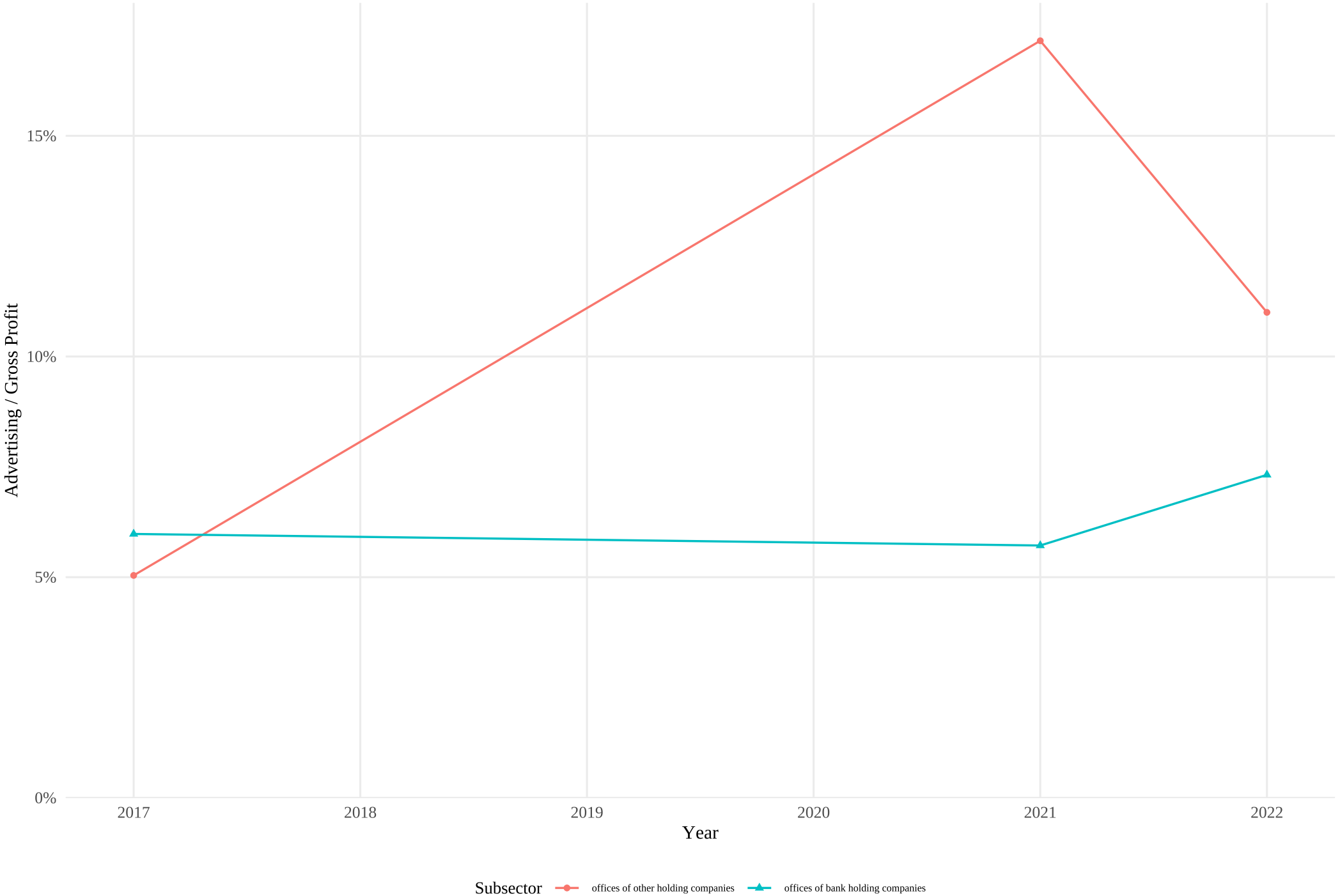
9 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Management of Companies

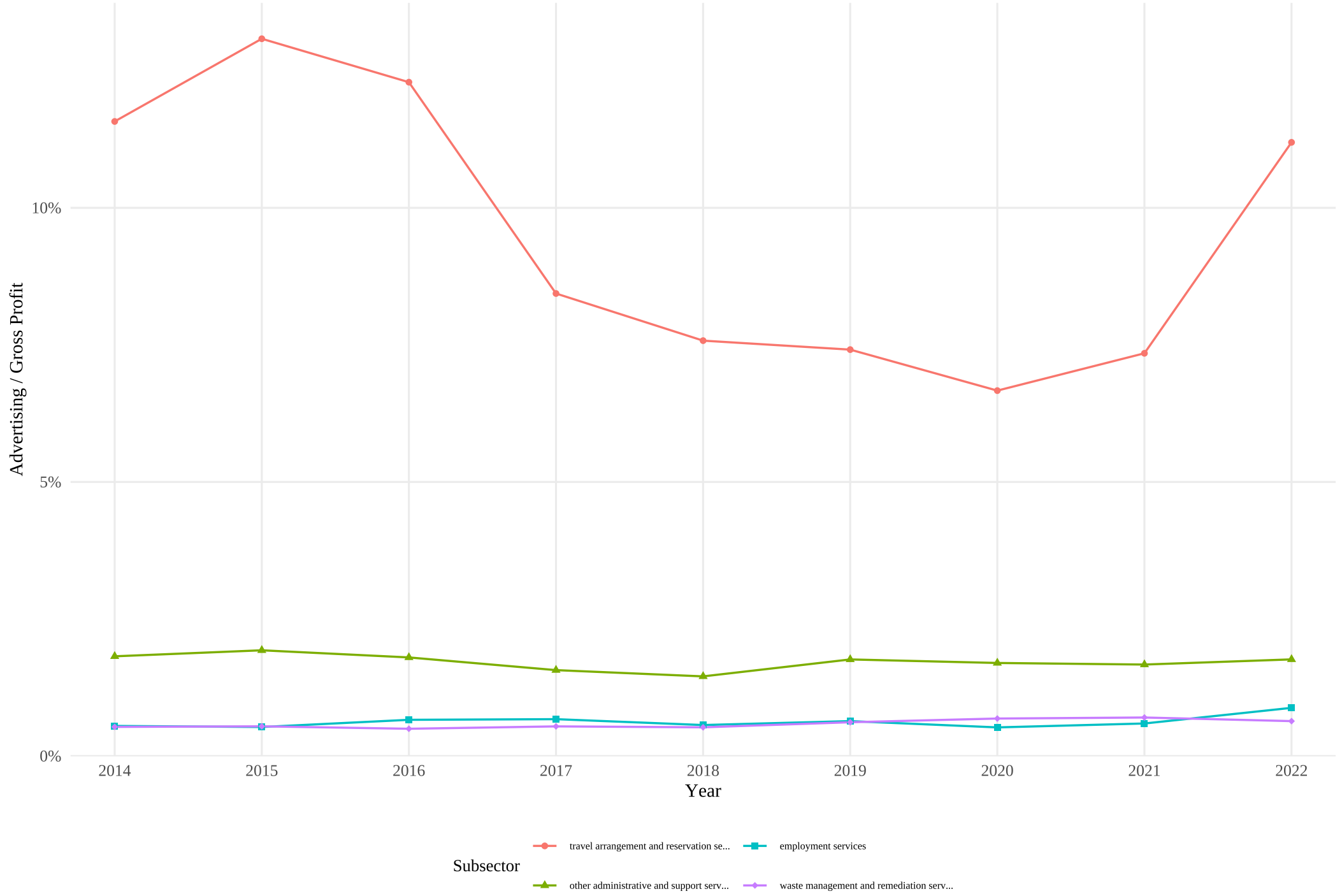
2 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Administrative Services

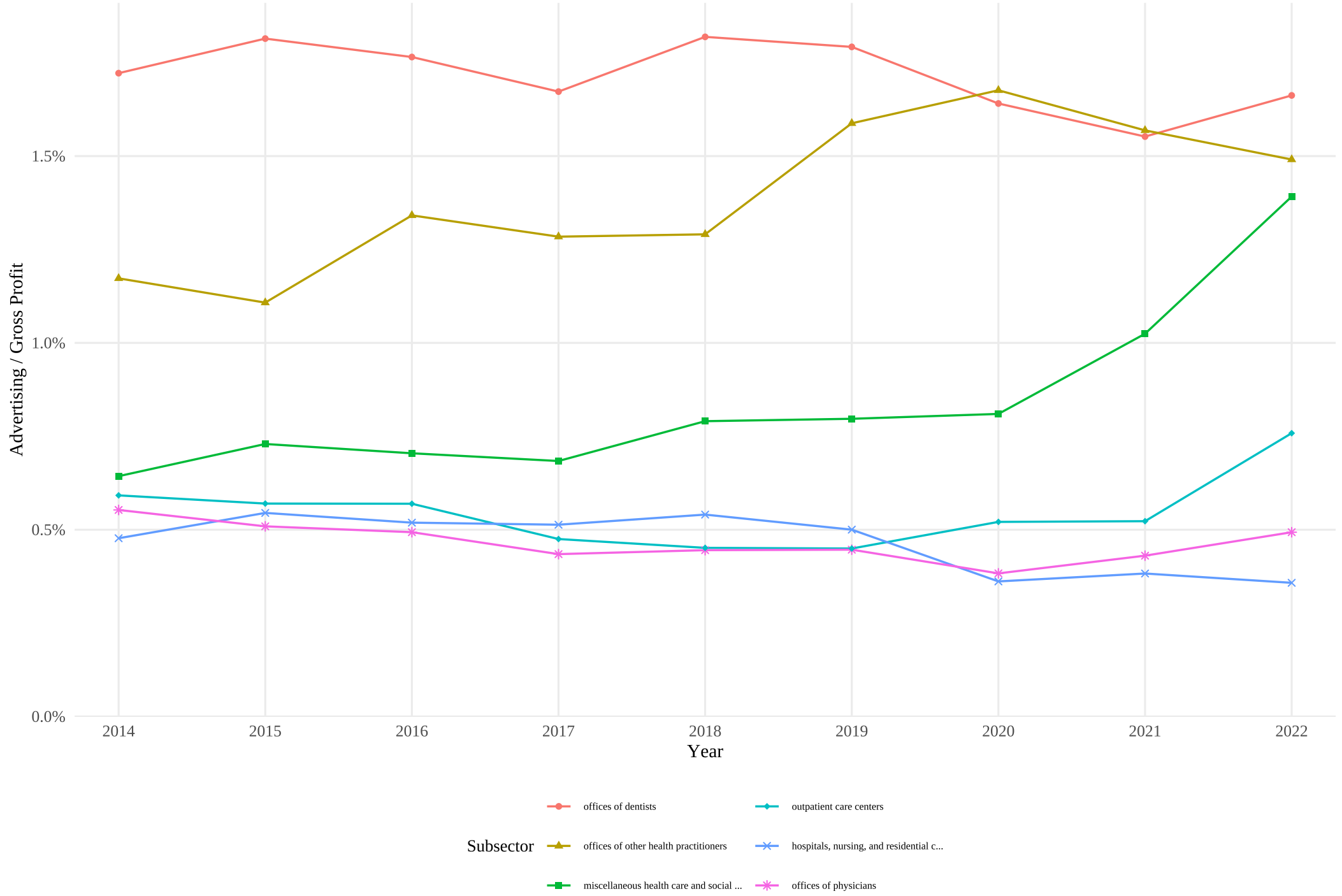
4 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Health Care

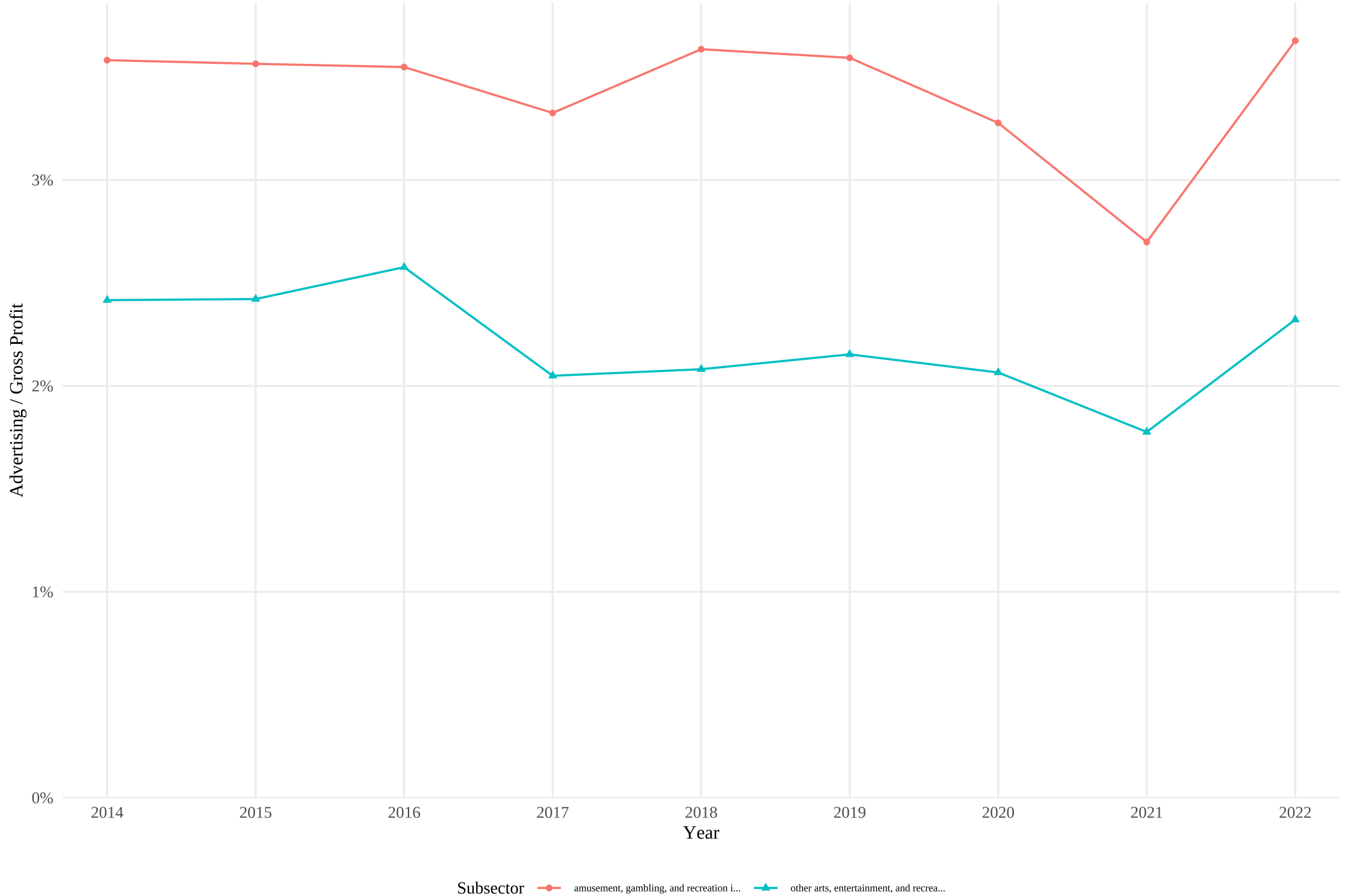
6 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Arts & Entertainment

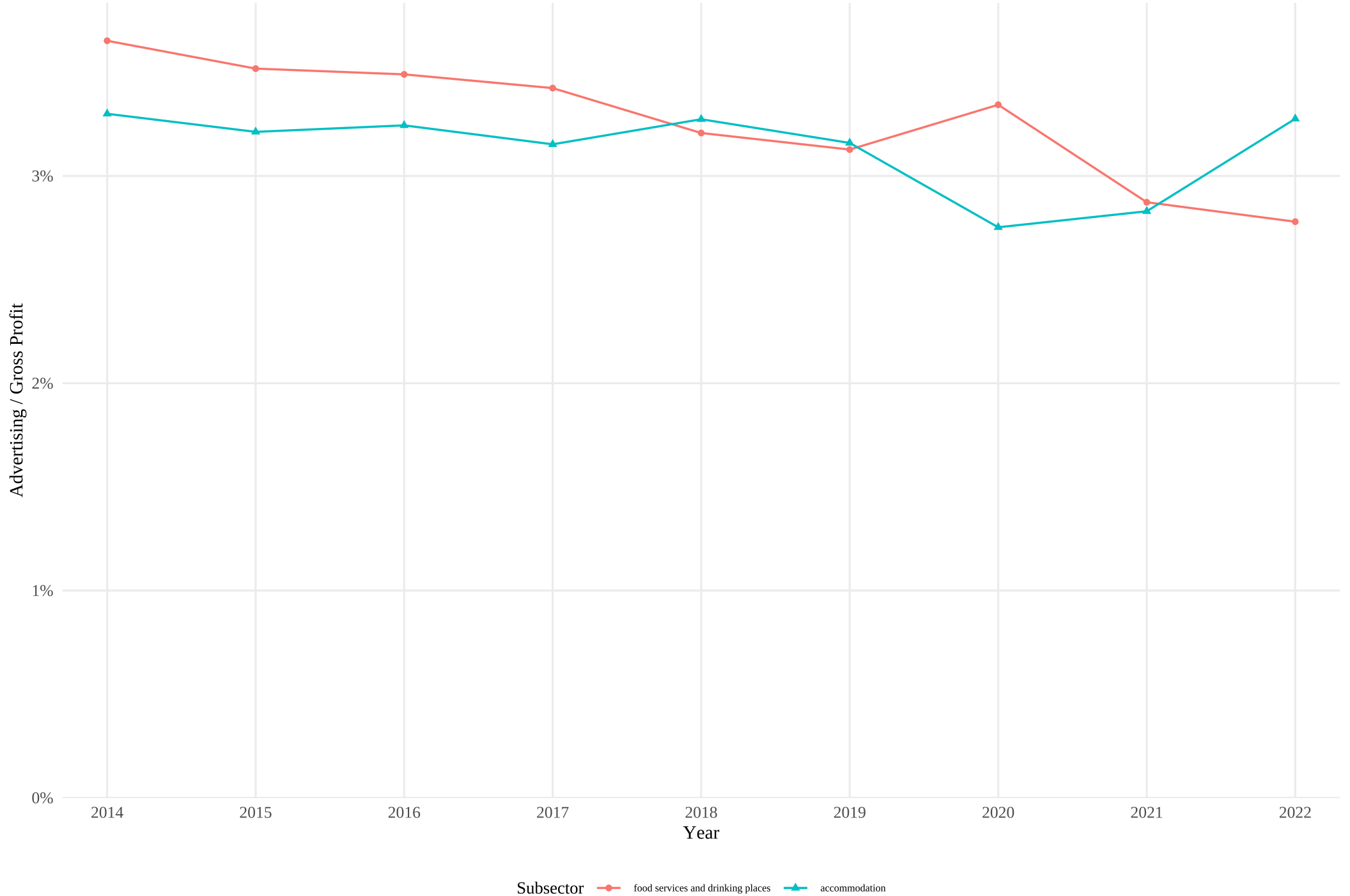
2 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Accommodation & Food

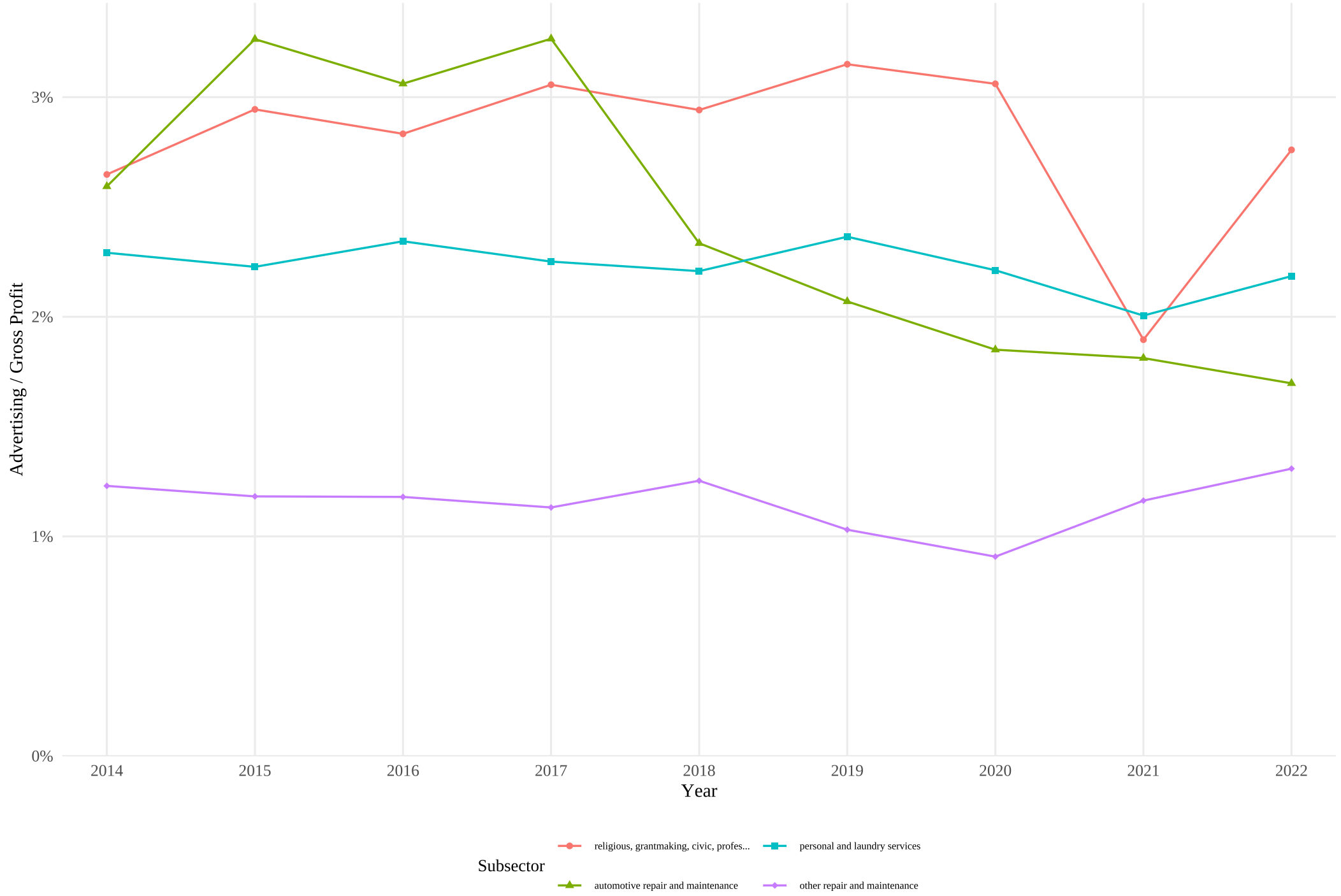
2 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Ad/Gross Profit: Other Services

4 subsectors



Source: IRS Statistics of Income, Table 5.1
Ordered by 2022 value

Appendix: parse_soi.R (Page 1 of 4)

```
################################################################################
# IRS SOI Table 5.1 Downloader and Parser
# Downloads Excel files from IRS website and parses into long-format panel data
################################################################################

library(tidyverse)
library(readxl)

# Configuration
input_dir <- "irs_soi_data"
years <- 2014:2022
base_url <- "https://www.irs.gov/pub/irs-soi"

################################################################################
# Step 0: Download Excel files from IRS
################################################################################

download_soi_files <- function(years, input_dir, base_url) {
  if (!dir.exists(input_dir)) {
    dir.create(input_dir, recursive = TRUE)
  }

  for (yr in years) {
    yr_short <- str_sub(as.character(yr), 3, 4)
    filename <- str_glue("{yr_short}co51ccr.xlsx")
    filepath <- file.path(input_dir, filename)

    if (file.exists(filepath)) {
      message(str_glue("  {filename} already exists, skipping"))
      next
    }

    url <- str_glue("{base_url}/{filename}")
    message(str_glue("  Downloading {filename}..."))

    tryCatch({
      download.file(url, filepath, mode = "wb", quiet = TRUE)
      message(str_glue("    Downloaded successfully"))
    }, error = function(e) {
      message(str_glue("    ERROR: {e$message}"))
    })
  }
}

message("=== Downloading IRS SOI Table 5.1 Files ===\n")
download_soi_files(years, input_dir, base_url)

################################################################################
# Step 1: Detect header row positions
################################################################################

detect_header_rows <- function(raw) {
  item_row <- which(as.character(raw[[1]]) == "Item")[1]
  if (item_row == 4) {
    list(sector_row = 4, subsector_row = 5)
  } else {
    list(sector_row = 5, subsector_row = 6)
  }
}

################################################################################
# Step 2: Build column map (industry metadata)
################################################################################
```

```
    sector_raw = r_sector,
    subsector_raw = r_subsector
  ) |>
    filter(col_idx > 1) |>
    fill(sector_raw, .direction = "down") |>
    mutate(
      sector = sector_raw |>
        str_replace_all("[\r\n]+", " ") |>
        str_squish() |>
        str_to_lower(),
      subsector = subsector_raw |>
        str_replace_all("[\r\n]+", " ") |>
        str_squish() |>
        str_to_lower(),
      is_sector_level = (subsector == "total" | is.na(subsector_raw)),
      industry_name = if_else(is_sector_level, sector, subsector),
      industry_id = paste(sector, if_else(is_sector_level, "_total", subsector), sep = "|") |>
        str_replace_all("[^a-z0-9|_]", "_") |>
        str_replace_all("_+", "_")
    )
}

################################################################################
# Step 3: Detect data row positions by regex matching
################################################################################

detect_variable_rows <- function(raw) {
  labels <- as.character(raw[[1]])

  # Define patterns for all variables we want to extract
  patterns <- c(
    n_returns = "^number of returns$",
    total_assets = "^total assets$",
    cash = "^cash$",
    trade_receivables = "^trade notes and accounts receivable$",
    allowance_bad_debts = "^less:.*allowance for bad debts$",
    inventories = "^inventories$",
    us_govt_obligations = "^u\\.?s\\.? government obligations$",
    tax_exempt_securities = "^tax-exempt securities$",
    other_current_assets = "^other current assets$",
    loans_to_shareholders = "^loans to shareholders$",
    mortgage_real_estate_loans = "^mortgage and real estate loans$",
    other_investments = "^other investments$",
    depreciable_assets = "^depreciable assets$",
    accum_depreciation = "^less:.*accumulated depreciation$",
    depletable_assets = "^depletable assets$",
    accum_depletion = "^less:.*accumulated depletion$",
    land = "^land$",
    intangible_assets = "^intangible assets",
    accum_amortization = "^less:.*accumulated amortization$",
    other_assets = "^other assets$",
    total_liabilities = "^total liabilities$",
    accounts_payable = "^accounts payable$",
    short_term_debt = "^mortgages.*less than 1 year$",
    other_current_liab = "^other current liabilities$",
    loans_from_shareholders = "^loans from shareholders$",
    long_term_debt = "^mortgages.*1 year or more$",
    other_liabilities = "^other liabilities$",
    net_worth = "^net worth.*total$",
    capital_stock = "^capital stock$",
    paid_in_capital = "^additional paid-in capital$",
    retained_earnings_approp = "^retained earnings.*appropriated$",
    retained_earnings_unapprop = "^retained earnings.*unappropriated$",
    treasury_stock = "^less:.*treasury stock$",
```

```
    net_st_capital_gain = "^net short-term capital gain",
    net_lt_capital_gain = "^net long-term capital gain",
    net_gain_noncapital = "^net gain.*noncapital assets$",
    tax_exempt_interest = "^tax-exempt interest$",
    other_receipts = "^other receipts$",
    total_deductions = "^total deductions",
    cost_goods_sold = "^cost of goods sold",
    compensation_officers = "^compensation of officers$",
    salaries_wages = "^salaries and wages$",
    repairs = "^repairs",
    bad_debts = "^bad debts$",
    rents_paid = "^rents paid$",
    taxes_licenses = "^taxes and licenses$",
    interest_paid = "^interest paid$",
    charitable_contributions = "^charitable contributions$",
    amortization = "^amortization$",
    depreciation = "^depreciation$",
    depletion = "^depletion$",
    advertising = "^advertising$",
    pension_plans = "^pension.*profit.sharing",
    employee_benefits = "^employee benefit programs$",
    net_loss_noncapital = "^net loss.*noncapital assets$",
    other_deductions = "^other deductions$",
    receipts_less_deductions = "^total receipts less total deductions$",
    net_income = "^net income.*less deficit",
    income_subject_to_tax = "^income subject to tax$",
    income_tax_before_credits = "^total income tax before credits$",
    income_tax_after_credits = "^total income tax after credits$"
  )

  # Find row for each pattern
  row_map <- map_int(patterns, function(pat) {
    idx <- str_which(labels, regex(pat, ignore_case = TRUE))
    if (length(idx) == 0) NA_integer_ else idx[1]
  })

  row_map
}

################################################################################
# Step 4: Extract data into long format
################################################################################

extract_data_long <- function(raw, column_map, variable_rows, year) {
  # For each variable and each industry column, extract the value
  map_dfr(names(variable_rows), function(var_name) {
    row_idx <- variable_rows[[var_name]]
    if (is.na(row_idx)) return(tibble())

    map_dfr(seq_len(nrow(column_map)), function(i) {
      col_idx <- column_map$col_idx[i]
      val <- raw[[col_idx]][row_idx]

      tibble(
        year = year,
        sector = column_map$sector[i],
        subsector = column_map$subsector[i],
        is_sector_level = column_map$is_sector_level[i],
        industry_name = column_map$industry_name[i],
        industry_id = column_map$industry_id[i],
        variable = var_name,
        value = as.numeric(val)
      )
    })
```

```
parse_soi_file <- function(filepath, year) {
  message(str_glue("Parsing {year}..."))

  raw <- read_excel(filepath, col_names = FALSE, .name_repair = "minimal")
  header_rows <- detect_header_rows(raw)
  column_map <- build_column_map(raw, header_rows)
  variable_rows <- detect_variable_rows(raw)

  # Report any missing variables
  missing_vars <- names(variable_rows)[is.na(variable_rows)]
  if (length(missing_vars) > 0) {
    message(str_glue("  Warning: Could not find rows for: {paste(missing_vars, collapse = ', ')}"))
  }

  extract_data_long(raw, column_map, variable_rows, year)
}

################################################################################
# Step 6: Parse all files
################################################################################

parse_all_files <- function(years, input_dir) {
  map_dfr(years, function(yr) {
    yr_short <- str_sub(as.character(yr), 3, 4)
    filepath <- file.path(input_dir, str_glue("{yr_short}co51ccr.xlsx"))

    if (!file.exists(filepath)) {
      message(str_glue("File not found: {filepath}"))
      return(tibble())
    }

    tryCatch(
      parse_soi_file(filepath, yr),
      error = function(e) {
        message(str_glue("Error parsing {yr}: {e$message}"))
        tibble()
      }
    )
  })
}

################################################################################
# Main execution
################################################################################

message("=== Parsing IRS SOI Table 5.1 Files ===\n")

panel_long <- parse_all_files(years, input_dir)

message(str_glue("\n=== Parsing Complete ==="))
message(str_glue("Total observations: {nrow(panel_long)}"))
message(str_glue("Years: {min(panel_long$year)}-{max(panel_long$year)}"))
message(str_glue("Unique industries: {n_distinct(panel_long$industry_id)}"))
message(str_glue("Variables: {n_distinct(panel_long$variable)}"))

# Save output
saveRDS(panel_long, file.path(input_dir, "soi_panel_long.rds"))
write_csv(panel_long, file.path(input_dir, "soi_panel_long.csv"))

message(str_glue("\nSaved to {input_dir}/soi_panel_long.rds and .csv"))
```

Appendix: sector_groupings.R (Page 1 of 3)

```
################################################################################
# Sector Groupings and Color Scheme
# Defines 5 semantic groups for NAICS 2-digit sectors with consistent colors
# Source this file for reuse across scripts
################################################################################

library(tibble)

################################################################################
# 5-Group Semantic Categorization
################################################################################

sector_group_definitions <- tribble(

  ~sector, ~sector_label, ~sector_group, ~group_order,
  # Group 1: Goods-Producing (Blues)
 "agriculture, forestry, fishing, and hunting", "Agriculture", "Goods-Producing", 1,
 "mining", "Mining", "Goods-Producing", 1,
 "construction", "Construction", "Goods-Producing", 1,
 "manufacturing", "Manufacturing", "Goods-Producing", 1,
  # Group 2: Distribution & Utilities (Oranges/Browns)
 "utilities", "Utilities", "Distribution & Utilities", 2,
 "wholesale trade", "Wholesale Trade", "Distribution & Utilities", 2,
 "retail trade", "Retail Trade", "Distribution & Utilities", 2,
 "transportation and warehousing", "Transportation", "Distribution & Utilities", 2,
  # Group 3: Finance & Real Estate (Greens)
 "finance and insurance", "Finance & Insurance", "Finance & Real Estate", 3,
 "real estate and rental and leasing", "Real Estate", "Finance & Real Estate", 3,
  # Group 4: Business Services (Purples)
 "information", "Information", "Business Services", 4,
 "professional, scientific, and technical services", "Professional Services", "Business Services", 4,
 "management of companies (holding companies)", "Management of Companies", "Business Services", 4,
 "administrative and support and waste management and remediation services", "Administrative Services", "Business Services", 4,
  # Group 5: Consumer Services (Reds/Pinks)
 "educational services", "Educational Services", "Consumer Services", 5,
 "health care and social assistance", "Health Care", "Consumer Services", 5,
 "arts, entertainment, and recreation", "Arts & Entertainment", "Consumer Services", 5,
 "accommodation and food services", "Accommodation & Food", "Consumer Services", 5,
 "other services", "Other Services", "Consumer Services", 5
)

################################################################################
# Color Palette by Group (19 colors total)
# Each group uses a hue family with lightness/saturation variation
################################################################################

sector_colors <- c(
  # Goods-Producing: Blues (4 sectors)
  "Agriculture" = "#08519C",
  "Mining" = "#3182BD",
  "Construction" = "#6BAED6",
  "Manufacturing" = "#9ECAE1",
  # Distribution & Utilities: Oranges/Browns (4 sectors)
  "Utilities" = "#8C510A",
  "Wholesale Trade" = "#BF812D",
  "Retail Trade" = "#DFC27D",
  "Transportation" = "#F6E8C3",
  # Finance & Real Estate: Greens (2 sectors)
  "Finance & Insurance" = "#1B7837",
  "Real Estate" = "#7FBC41",
  # Business Services: Purples (4 sectors)
  "Information" = "#6A3D9A",
  "Professional Services" = "#9E7BB5",
  "Management of Companies" = "#CAB2D6",
```

```
  "Other Services" = "#FCBBA1"
)

################################################################################
# Point Shapes by Sector (varies within each group for distinguishability)
# Shapes: 16=circle, 17=triangle, 15=square, 18=diamond, 4=cross
################################################################################

sector_shapes <- c(
  # Goods-Producing (4 sectors)
  "Agriculture" = 16,
  "Mining" = 17,
  "Construction" = 15,
  "Manufacturing" = 18,
  # Distribution & Utilities (4 sectors)
  "Utilities" = 16,
  "Wholesale Trade" = 17,
  "Retail Trade" = 15,
  "Transportation" = 18,
  # Finance & Real Estate (2 sectors)
  "Finance & Insurance" = 16,
  "Real Estate" = 17,
  # Business Services (4 sectors)
  "Information" = 16,
  "Professional Services" = 17,
  "Management of Companies" = 15,
  "Administrative Services" = 18,
  # Consumer Services (5 sectors)
  "Educational Services" = 16,
  "Health Care" = 17,
  "Arts & Entertainment" = 15,
  "Accommodation & Food" = 18,
  "Other Services" = 4
)

################################################################################
# Group-level colors (for group summary plots)
################################################################################

group_colors <- c(
 "Goods-Producing" = "#3182BD",
 "Distribution & Utilities" = "#BF812D",
  "Finance & Real Estate" = "#1B7837",
  "Business Services" = "#7B68EE",
  "Consumer Services" = "#CB181D"
)

################################################################################
# Ordered factor levels
################################################################################

group_levels <- c(
 "Goods-Producing",
 "Distribution & Utilities",
 "Finance & Real Estate",
  "Business Services",
  "Consumer Services"
)

# Sector labels ordered by group, then within group
sector_label_levels <- c(
  # Goods-Producing
  "Agriculture", "Mining", "Construction", "Manufacturing",
  # Distribution & Utilities
```

Appendix: sector_groupings.R (Page 3 of 3)

```
  "Educational Services", "Health Care", "Arts & Entertainment", "Accommodation & Food", "Other Services"
)
```

Appendix: plot_soi_unified.R (Page 1 of 15)

```
################################################################################
# Unified SOI Graphics Script
# Generates soi_unified.pdf with sector, group, and subsector visualizations
# All plots include "All Companies Combined" dark gray reference line
################################################################################

library(dplyr)
library(tidyr)
library(readr)
library(ggplot2)
library(purrr)
library(stringr)
library(showtext)

################################################################################
# Configuration
################################################################################

input_file <- "irs_soi_data/soi_panel_long.rds"
output_file <- "wilbur_2026_irs_soi_data_visualizations.pdf"

source("sector_groupings.R")

# CPI adjustment factors (to 2022 dollars)
cpi <- tibble(
  year = 2014:2022,
  cpi_u = c(236.736, 237.017, 240.007, 245.120, 251.107,
            255.657, 258.811, 270.970, 292.655)
) |> mutate(deflator = cpi_u[year == 2022] / cpi_u)

# "All Companies Combined" styling
all_companies_color <- "#4D4D4D"
all_companies_shape <- 8  # asterisk
all_companies_linewidth <- 1.2
all_companies_label <- "All Companies Combined"

################################################################################
# Font setup
################################################################################

font_add_google("Tinos", "Tinos")
font_add_google("Fira Mono", "FiraMono")
showtext_auto()

################################################################################
# Load and prepare data
################################################################################

panel <- readRDS(input_file)

# All industries aggregate data
all_industries_long <- panel |>
  filter(sector == "all industries") |>
  left_join(cpi, by = "year") |>
  mutate(value_real = value * deflator)

# Sector-level data (19 sectors, exclude "all industries")
sector_data <- panel |>
  filter(is_sector_level, sector != "all industries") |>
  left_join(cpi, by = "year") |>
  mutate(value_real = value * deflator) |>
  left_join(sector_group_definitions, by = "sector") |>
  mutate(
    sector label = factor(sector label. levels = sector label levels).
```

```
  left_join(cpi, by = "year") |>
  mutate(value_real = value * deflator) |>
  left_join(sector_group_definitions, by = "sector") |>
  mutate(sector_group = factor(sector_group, levels = group_levels))

################################################################################
# Compute ratios - Sector level
################################################################################

sector_wide <- sector_data |>
  select(year, sector, sector_label, sector_group, variable, value, value_real) |>
  pivot_wider(names_from = variable, values_from = c(value, value_real))

sector_wide <- sector_wide |>
  mutate(
    gross_profit = value_business_receipts - value_cost_goods_sold,
    gross_profit_margin = gross_profit / value_business_receipts,
    ad_revenue_ratio = value_advertising / value_business_receipts,
    ad_gross_profit_ratio = value_advertising / gross_profit,
    ad_net_income_ratio = value_advertising / value_net_income,
    net_income_gross_profit_ratio = value_net_income / gross_profit
  )

sector_ratios <- sector_wide |>
  select(year, sector, sector_label, sector_group,
         gross_profit_margin, ad_revenue_ratio, ad_gross_profit_ratio,
         ad_net_income_ratio, net_income_gross_profit_ratio) |>
  pivot_longer(cols = c(gross_profit_margin, ad_revenue_ratio, ad_gross_profit_ratio,
                        ad_net_income_ratio, net_income_gross_profit_ratio),
               names_to = "variable", values_to = "value")

################################################################################
# Compute ratios - All industries
################################################################################

all_ind_wide <- all_industries_long |>
  select(year, variable, value, value_real) |>
  pivot_wider(names_from = variable, values_from = c(value, value_real))

all_ind_wide <- all_ind_wide |>
  mutate(
    gross_profit = value_business_receipts - value_cost_goods_sold,
    gross_profit_margin = gross_profit / value_business_receipts,
    ad_revenue_ratio = value_advertising / value_business_receipts,
    ad_gross_profit_ratio = value_advertising / gross_profit,
    ad_net_income_ratio = value_advertising / value_net_income,
    net_income_gross_profit_ratio = value_net_income / gross_profit
  )

all_ind_ratios <- all_ind_wide |>
  select(year, gross_profit_margin, ad_revenue_ratio, ad_gross_profit_ratio,
         ad_net_income_ratio, net_income_gross_profit_ratio) |>
  pivot_longer(cols = -year, names_to = "variable", values_to = "value")

################################################################################
# Harmonize subsector names (IRS renamed many in 2022)
################################################################################

subsector_name_map <- c(
  # Retail trade renames (stores -> retailers)
  "beer, wine, and liquor retailers" = "beer, wine, and liquor stores",
  "clothing and clothing accessories retailers" = "clothing and clothing accessories stores",
  "electronics and appliance retailers (including computers)" = "electronics and appliance stores",
  "food and beverage retailers" = "food and beverage stores",
```

```
  "lawn and garden equipment and supplies retailers" = "lawn and garden equipment and supplies stores",
  "sporting goods, hobby, book, music and miscellaneous retailers" = "sporting goods, hobby, book, and music stores",
  # Finance renames
  "commodity contracts intermediation" = "commodity contracts dealing and brokerage",
  "investment banking and securities intermediation" = "investment banking and securities dealing",
  "savings institutions and other depository credit intermediation" = "savings institutions, credit unions, and other depository credit intermediation",
  "life insurance (form 1120l)" = "life insurance",
  "life insurance (form 1120-l)" = "life insurance",
  "property and casualty insurance (form 1120pc)" = "property and casualty insurance",
  "property and casualty insurance (form 1120-pc)" = "property and casualty insurance",
  "other financial vehicles mortgage real estate investment trust (reits)" = "other financial vehicles (including mortgage reits)",
  "other insurance related activities (including third-party administrator of insurance, and pension funds)" = "other insurance related activities (third-party administrator of insurance, etc.)",
  "activities related to credit intermediation (including loan brokers)" = "activities related to credit intermediation (loan brokers, check clearing, etc.)",
  # Information renames
  "computing infrastructure providers, data processing, web hosting and related services" = "data processing, hosting, and related services",
  "sounds recording industries" = "sound recording industries",
  "telecommunications (including wired, wireless, satellite, cable and other program distribution, resellers, agents, other telecommunications, and internet service providers)" = "telecommunications (paging, cell
  "telecommunications (paging, cellular, cable, satellite, & internet service providers" = "telecommunications (paging, cellular, cable, satellite, & internet service providers)",
  "web search portals, libraries, archives, and other info. services" = "other information services",
  "boradcasting & content providers (including movies, tv, radio, music and book publishers)" = "broadcasting (except internet)",
  # Manufacturing renames
  "cutlery, hardware, spring and wire, machine shops; screw, nut, and bolt" = "cutlery, hardware, spring and wire: machine shops, screw, nut, and bolt",
  "other food (including coffee, tea, flavorings, and seasonings)" = "other food",
  "apparel accessories and other apparel manufacturing" = "apparel accessories and other apparel",
  # Wholesale renames
  "wholesale trade agents and brokers" = "wholesale electronic markets and agents and brokers"
)

subsector_data <- subsector_data |>
  mutate(subsector = if_else(subsector %in% names(subsector_name_map),
                             subsector_name_map[subsector],
                             subsector))

################################################################################
# Compute ratios - Subsector level
################################################################################

subsector_wide <- subsector_data |>
  filter(variable %in% c("advertising", "business_receipts",
                         "cost_goods_sold", "net_income")) |>
  select(year, sector, subsector, sector_group, variable, value) |>
  pivot_wider(names_from = variable, values_from = value)

subsector_wide <- subsector_wide |>
  mutate(
    gross_profit = business_receipts - cost_goods_sold,
    ad_revenue_ratio = advertising / business_receipts,
    ad_gross_profit_ratio = advertising / gross_profit,
    ad_net_income_ratio = advertising / net_income
  )

################################################################################
# Theme and helper functions
################################################################################

theme_soi <- function() {
  theme_minimal(base_family = "Tinos", base_size = 12) +
    theme(
      plot.title = element_text(face = "bold", size = 14),
      plot.subtitle = element_text(size = 10, color = "gray40"),
      axis.title = element_text(size = 11),
      legend.position = "bottom",
      legend.title = element_text(size = 10),
      legend.text = element_text(size = 8),
```

```
  max_val <- max(values, na.rm = TRUE)
  if (max_val >= 1e12) {
    list(scale = 1e-12, suffix = "T", label = "$ Trillions")
  } else if (max_val >= 1e9) {
    list(scale = 1e-9, suffix = "B", label = "$ Billions")
  } else if (max_val >= 1e6) {
    list(scale = 1e-6, suffix = "M", label = "$ Millions")
  } else if (max_val >= 1e3) {
    list(scale = 1e-3, suffix = "K", label = "$ Thousands")
  } else {
    list(scale = 1, suffix = "", label = "$")
  }
}

################################################################################
# Extend color/shape palettes to include "All Companies Combined"
################################################################################

sector_colors_extended <- c(sector_colors,
                            setNames(all_companies_color, all_companies_label))
sector_shapes_extended <- c(sector_shapes,
                            setNames(all_companies_shape, all_companies_label))
sector_label_levels_extended <- c(sector_label_levels, all_companies_label)

################################################################################
# Plot function: Sector plot with All Companies reference line
################################################################################

plot_sector_with_all <- function(sector_df, all_df, var_name, title, y_label,
                                 definition, is_ratio = FALSE, is_count = FALSE,
                                 allow_negative = FALSE) {

  # Prepare sector data
  plot_data <- sector_df |> filter(variable == var_name)

  # Prepare all-companies data
  all_data <- all_df |>
    filter(variable == var_name) |>
    mutate(sector_label = all_companies_label)

  # Combine
  if (is_ratio) {
    combined <- bind_rows(
      plot_data |> select(year, sector_label, value),
      all_data |> select(year, sector_label, value)
    )
  } else {
    combined <- bind_rows(
      plot_data |> select(year, sector_label, value_real),
      all_data |> select(year, sector_label, value_real)
    )
  }

  combined <- combined |>
    mutate(sector_label = factor(sector_label, levels = sector_label_levels_extended))

  # Build plot
  if (is_ratio) {
    y_limits <- if (allow_negative) c(NA, NA) else c(0, NA)
    y_expand <- if (allow_negative) expansion(mult = 0.05) else expansion(mult = c(0, 0.05))

    p <- ggplot(combined, aes(x = year, y = value, color = sector_label,
                              shape = sector_label)) +
      geom_line(aes(linewidth = sector_label == all_companies_label)) +
```

```
  } else if (is_count) {
    scale_info <- get_scale_info(combined$value_real)
    p <- ggplot(combined, aes(x = year, y = value_real, color = sector_label,
                              shape = sector_label)) +
      geom_line(aes(linewidth = sector_label == all_companies_label)) +
      geom_point(size = 2) +
      scale_linewidth_manual(values = c("FALSE" = 0.8, "TRUE" = all_companies_linewidth),
                             guide = "none") +
      scale_y_continuous(labels = scales::label_comma(scale = scale_info$scale,
                                                      suffix = scale_info$suffix),
                         limits = c(0, NA),
                         expand = expansion(mult = c(0, 0.05))) +
      labs(y = y_label)
  } else {
    scale_info <- get_scale_info(combined$value_real)
    p <- ggplot(combined, aes(x = year, y = value_real, color = sector_label,
                              shape = sector_label)) +
      geom_line(aes(linewidth = sector_label == all_companies_label)) +
      geom_point(size = 2) +
      scale_linewidth_manual(values = c("FALSE" = 0.8, "TRUE" = all_companies_linewidth),
                             guide = "none") +
      scale_y_continuous(labels = function(x) paste0("$", x * scale_info$scale,
                                                     scale_info$suffix),
                         limits = c(0, NA),
                         expand = expansion(mult = c(0, 0.05))) +
      labs(y = paste0(y_label, " (Real 2022 $)"))
  }

  caption_text <- if (is_ratio) {
    paste0("Source: IRS Statistics of Income, Table 5.1\nRatio: ", definition)
  } else {
    paste0("Source: IRS Statistics of Income, Table 5.1 (Real 2022 dollars)\nDefinition: ", definition)
  }

  p +
    scale_x_continuous(breaks = 2014:2022, expand = expansion(add = 0.3)) +
    scale_color_manual(values = sector_colors_extended, drop = FALSE) +
    scale_shape_manual(values = sector_shapes_extended, drop = FALSE) +
    labs(
      title = title,
      x = "Year",
      color = "Sector",
      shape = "Sector",
      caption = caption_text
    ) +
    theme_soi() +
    guides(color = guide_legend(ncol = 5, byrow = FALSE),
           shape = guide_legend(ncol = 5, byrow = FALSE))
}

################################################################################
# Plot function: Combined Ad Ratios (Ad/Rev + Ad/GP on same plot)
################################################################################

plot_combined_ad_ratios <- function(sector_df, all_df, title) {

  # Prepare sector data for both ratios
  sector_long <- sector_df |>
    filter(variable %in% c("ad_revenue_ratio", "ad_gross_profit_ratio")) |>
    mutate(ratio_type = case_when(
      variable == "ad_revenue_ratio" ~ "Ad / Revenue",
      variable == "ad_gross_profit_ratio" ~ "Ad / Gross Profit"
    ))
```

```
        variable == "ad_revenue_ratio" ~ "Ad / Revenue",
        variable == "ad_gross_profit_ratio" ~ "Ad / Gross Profit"
      )
    )

  combined <- bind_rows(
    sector_long |> select(year, sector_label, value, ratio_type),
    all_long |> select(year, sector_label, value, ratio_type)
  ) |>
    mutate(sector_label = factor(sector_label, levels = sector_label_levels_extended))

  p <- ggplot(combined, aes(x = year, y = value, color = sector_label,
                            shape = sector_label)) +
    geom_line(aes(linewidth = sector_label == all_companies_label)) +
    geom_point(size = 1.5) +
    scale_linewidth_manual(values = c("FALSE" = 0.6, "TRUE" = all_companies_linewidth),
                           guide = "none") +
    scale_y_continuous(labels = scales::label_percent(),
                       limits = c(0, NA),
                       expand = expansion(mult = c(0, 0.05))) +
    scale_x_continuous(breaks = 2014:2022, expand = expansion(add = 0.3)) +
    scale_color_manual(values = sector_colors_extended, drop = FALSE) +
    scale_shape_manual(values = sector_shapes_extended, drop = FALSE) +
    facet_wrap(~ratio_type, ncol = 2) +
    labs(
      title = title,
      x = "Year",
      y = "Advertising Ratio",
      color = "Sector",
      shape = "Sector",
      caption = "Source: IRS Statistics of Income, Table 5.1"
    ) +
    theme_soi() +
    guides(color = guide_legend(ncol = 5, byrow = FALSE),
           shape = guide_legend(ncol = 5, byrow = FALSE))

  return(p)
}

################################################################################
# Plot function: Group breakout (sectors within group)
################################################################################

plot_group_sectors <- function(sector_df, all_df, group_name, var_name,
                               title, y_label, definition,
                               is_ratio = TRUE, allow_negative = FALSE) {

  # Filter to group
  plot_data <- sector_df |>
    filter(variable == var_name, sector_group == group_name)

  # Get all-companies data
  all_data <- all_df |>
    filter(variable == var_name) |>
    mutate(sector_label = all_companies_label)

  combined <- bind_rows(
    plot_data |> select(year, sector_label, value),
    all_data |> select(year, sector_label, value)
  )

  # Get group-specific colors and shapes
  group_sectors <- sector_group_definitions |>
    filter(sector_group == group_name) |>
```

```
  group_shapes <- c(group_shapes, setNames(all_companies_shape, all_companies_label))

  combined <- combined |>
    mutate(sector_label = factor(sector_label,
                                 levels = c(group_sectors, all_companies_label)))

  y_limits <- if (allow_negative) c(NA, NA) else c(0, NA)
  y_expand <- if (allow_negative) expansion(mult = 0.05) else expansion(mult = c(0, 0.05))

  p <- ggplot(combined, aes(x = year, y = value, color = sector_label,
                            shape = sector_label)) +
    geom_line(aes(linewidth = sector_label == all_companies_label)) +
    geom_point(size = 2.5) +
    scale_linewidth_manual(values = c("FALSE" = 1.0, "TRUE" = all_companies_linewidth),
                           guide = "none") +
    scale_y_continuous(labels = scales::label_percent(), limits = y_limits,
                       expand = y_expand) +
    scale_x_continuous(breaks = 2014:2022, expand = expansion(add = 0.3)) +
    scale_color_manual(values = group_colors) +
    scale_shape_manual(values = group_shapes) +
    labs(
      title = title,
      subtitle = paste0("Sector Group: ", group_name),
      x = "Year",
      y = y_label,
      color = "Sector",
      shape = "Sector",
      caption = paste0("Source: IRS Statistics of Income, Table 5.1\nRatio: ", definition)
    ) +
    theme_soi() +
    guides(color = guide_legend(nrow = 2, byrow = TRUE),
           shape = guide_legend(nrow = 2, byrow = TRUE))

  return(p)
}

# ============================================================================
# Plot function: Group breakout (subsectors within group)
# ============================================================================

plot_group_subsectors <- function(subsector_df, group_name, ratio_var,
                                  ratio_label, title,
                                  outlier_upper = 0.5, outlier_lower = -0.2) {  # 50%, -20%

  # Get sectors in this group
  group_sectors <- sector_group_definitions |>
    filter(sector_group == group_name) |>
    pull(sector)

  # Filter subsectors
  plot_data <- subsector_df |>
    filter(sector %in% group_sectors, is.finite(.data[[ratio_var]]))

  # For ad_net_income_ratio: remove entire subsector lines with extreme outliers
  if (grepl("net_income", ratio_var)) {
    outlier_subsectors <- plot_data |>
      filter(.data[[ratio_var]] > outlier_upper | .data[[ratio_var]] < outlier_lower) |>
      pull(subsector) |>
      unique()
    if (length(outlier_subsectors) > 0) {
      plot_data <- plot_data |>
        filter(!subsector %in% outlier_subsectors)
      message("  Removed ", length(outlier_subsectors), " subsectors with ratio > ",
              outlier_upper * 100, "% or < ", outlier_lower * 100, "% in ", group_name)
```

```
n_subsectors <- n_distinct(plot_data$subsector)

# Dynamic color palette
if (n_subsectors <= 12) {
  subsector_colors <- scales::hue_pal()(n_subsectors)
} else if (n_subsectors <= 24) {
  subsector_colors <- c(scales::hue_pal()(12),
                        scales::hue_pal(h = c(0, 360) + 15)(n_subsectors - 12))
} else {
  subsector_colors <- colorRampPalette(
    c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd",
      "#8c564b", "#e377c2", "#7f7f7f", "#bcbd22", "#17becf")
  )(n_subsectors)
}

# Dynamic shape palette (cycle)
base_shapes <- c(16, 17, 15, 18, 4, 8, 3, 7, 9, 10, 11, 12, 13, 14)
subsector_shapes <- rep(base_shapes, length.out = n_subsectors)

# Order by most recent available value and truncate names
# Use mean ratio as fallback for subsectors missing the max year
subsector_order <- plot_data |>
  group_by(subsector) |>
  summarise(order_val = mean(.data[[ratio_var]], na.rm = TRUE), .groups = "drop") |>
  arrange(desc(order_val)) |>
  pull(subsector)

plot_data <- plot_data |>
  mutate(
    subsector_short = str_trunc(subsector, 40),
    subsector_short = factor(subsector_short,
                             levels = unique(str_trunc(subsector_order, 40)))
  )

names(subsector_colors) <- levels(plot_data$subsector_short)
names(subsector_shapes) <- levels(plot_data$subsector_short)

# Legend layout
if (n_subsectors <= 8) {
  legend_ncol <- 2
} else if (n_subsectors <= 16) {
  legend_ncol <- 3
} else if (n_subsectors <= 30) {
  legend_ncol <- 4
} else {
  legend_ncol <- 5
}

# Allow negative for ad/net income
allow_neg <- grepl("net_income", ratio_var)
y_limits <- if (allow_neg) c(NA, NA) else c(0, NA)
y_expand <- if (allow_neg) expansion(mult = 0.05) else expansion(mult = c(0, 0.05))

p <- ggplot(plot_data, aes(x = year, y = .data[[ratio_var]],
                           color = subsector_short, shape = subsector_short)) +
  geom_line(linewidth = 0.6) +
  geom_point(size = 1.5) +
  scale_y_continuous(labels = scales::label_percent(), limits = y_limits,
                     expand = y_expand) +
  scale_x_continuous(breaks = 2014:2022, expand = expansion(add = 0.3)) +
  scale_color_manual(values = subsector_colors) +
  scale_shape_manual(values = subsector_shapes) +
  labs(
```

```
      caption = paste0("Source: IRS Statistics of Income, Table 5.1\nRatio: ", ratio_label,
                       "\nOrdered by 2022 value")
    ) +
    theme_soi() +
    theme(legend.text = element_text(size = 6)) +
    guides(color = guide_legend(ncol = legend_ncol),
           shape = guide_legend(ncol = legend_ncol))

  return(p)
}

################################################################################
# Generate Title Pages (Pages 1-2)
################################################################################

message("Generating title pages...")

# Page 1: Title and basic info
title_text_p1 <- paste0(
"Author: Kenneth C. Wilbur, UC San Diego
Date: February 2026

DATA SOURCE
- IRS Statistics of Income, Corporation Income Tax Returns
- Table 5.1: Returns of Active Corporations by Minor Industry
- Years: 2014-2022 (9 years)

SAMPLE
- Active U.S. Corporations (C corps, S corps, REITs, RICs)
- 19 sector-level industries
- 187 subsector-level industries

RATIOS COMPUTED
- Advertising / Revenue = Advertising Deductions / Business Receipts
- Advertising / Gross Profit = Advertising / (Revenue - Cost of Goods Sold)
- Advertising / Net Income = Advertising / Net Income (less credits)
- Net Income / Gross Profit = Net Income / (Revenue - Cost of Goods Sold)
- Gross Profit Margin = (Revenue - Cost of Goods Sold) / Revenue

ADJUSTMENTS
- Dollar values adjusted to real 2022 dollars using CPI-U
- CPI-U values: 236.7 (2014) to 292.7 (2022)"
)

p_title_1 <- ggplot() +
  annotate("text", x = 0.5, y = 0.95, label = "IRS Statistics of Income Data Visualizations",
           size = 7, fontface = "bold", family = "Tinos", hjust = 0.5, vjust = 1) +
  annotate("text", x = 0.05, y = 0.82, label = title_text_p1,
           size = 3.5, family = "Tinos", hjust = 0, vjust = 1, lineheight = 1.2) +
  xlim(0, 1) + ylim(0, 1) +
  theme_void() +
  theme(plot.margin = margin(20, 20, 20, 20))

# Page 2: Omissions, groupings, and structure
title_text_p2 <- paste0(
"OMITTED TIME SERIES
- Page 18 (Advertising/Net Income by Sector): Excludes Educational Services,
  Utilities, Accommodation & Food (ratios exceeded 80% or fell below -20%)
- Page 19 (Net Income/Gross Profit by Sector): Excludes Finance & Insurance,
  Management of Companies, Mining (ratios exceeded 60% or fell below -10%)

SECTOR GROUPINGS
- Goods-Producing: Agriculture, Mining, Construction, Manufacturing
- Distribution & Utilities: Utilities, Wholesale, Retail, Transportation
```

```
- Pages 16-20: Ratio variables by sector (5 pages)
- Pages 21-30: Group-sector breakouts (10 pages)
- Pages 31-48: Subsector Ad/Revenue by sector (18 pages)
- Pages 49-66: Subsector Ad/Gross Profit by sector (18 pages)

Prepared quickly using Claude Code. Accuracy is not ensured. R scripts are
included as appendices to enable replication. Please contact the author
in case any error is found."
)

p_title_2 <- ggplot() +
  annotate("text", x = 0.5, y = 0.95, label = "IRS Statistics of Income Data Visualizations (continued)",
           size = 6, fontface = "bold", family = "Tinos", hjust = 0.5, vjust = 1) +
  annotate("text", x = 0.05, y = 0.85, label = title_text_p2,
           size = 3.5, family = "Tinos", hjust = 0, vjust = 1, lineheight = 1.2) +
  xlim(0, 1) + ylim(0, 1) +
  theme_void() +
  theme(plot.margin = margin(20, 20, 20, 20))

################################################################################
# Generate Section A: Level Variables (13 pages)
################################################################################

message("Generating Section A: Level variables...")

level_vars <- list(
  list(var = "n_returns", title = "Number of Tax Returns by Sector",
       y_label = "Number of Returns", def = "Count of corporate tax returns",
       is_count = TRUE),
  list(var = "business_receipts", title = "Total Revenue by Sector",
       y_label = "Revenue", def = "Gross receipts from sales"),
  list(var = "cost_goods_sold", title = "Cost of Goods Sold by Sector",
       y_label = "COGS", def = "Direct costs of producing goods/services"),
  list(var = "total_deductions", title = "Total Deductions by Sector",
       y_label = "Deductions", def = "All expenses claimed against income"),
  list(var = "salaries_wages", title = "Salaries and Wages by Sector",
       y_label = "Salaries & Wages", def = "Compensation to non-officer employees"),
  list(var = "compensation_officers", title = "Compensation of Officers by Sector",
       y_label = "Officer Compensation", def = "Compensation to corporate officers"),
  list(var = "advertising", title = "Advertising Deductions by Sector",
       y_label = "Advertising", def = "Advertising and promotion expenses"),
  list(var = "depreciation", title = "Depreciation by Sector",
       y_label = "Depreciation", def = "Decline in value of tangible assets"),
  list(var = "interest_paid", title = "Interest Paid by Sector",
       y_label = "Interest Paid", def = "Interest expenses on business debt"),
  list(var = "net_income", title = "Net Income by Sector",
       y_label = "Net Income", def = "Receipts minus deductions"),
  list(var = "income_tax_after_credits", title = "Income Tax After Credits by Sector",
       y_label = "Income Tax", def = "Tax liability after credits"),
  list(var = "total_assets", title = "Total Assets by Sector",
       y_label = "Total Assets", def = "Sum of all corporate assets"),
  list(var = "total_liabilities", title = "Total Liabilities by Sector",
       y_label = "Total Liabilities", def = "Sum of all corporate liabilities")
)

section_a_plots <- map(level_vars, function(v) {
  is_count <- isTRUE(v$is_count)
  plot_sector_with_all(
    sector_data, all_industries_long, v$var,
    paste0(v$title, ", Active U.S. Corporations"),
    v$y_label, v$def, is_ratio = FALSE, is_count = is_count
  )
})
```

```
# Page 14: Gross Profit Margin
p_gpm <- plot_sector_with_all(
  sector_ratios, all_ind_ratios, "gross_profit_margin",
  "Gross Profit Margin by Sector, Active U.S. Corporations",
  "Gross Profit Margin", "(Revenue - COGS) / Revenue", is_ratio = TRUE
)

# Page 15: Ad / Revenue
p_ad_rev <- plot_sector_with_all(
  sector_ratios, all_ind_ratios, "ad_revenue_ratio",
  "Advertising / Revenue by Sector, Active U.S. Corporations",
  "Advertising / Revenue", "Advertising / Revenue",
  is_ratio = TRUE
)

# Page 16: Ad / Gross Profit
p_ad_gp <- plot_sector_with_all(
  sector_ratios, all_ind_ratios, "ad_gross_profit_ratio",
  "Advertising / Gross Profit by Sector, Active U.S. Corporations",
  "Advertising / Gross Profit", "Advertising / Gross Profit",
  is_ratio = TRUE
)

# Save high-resolution PNG for teaching slides
ggsave("../AMimages/ad_gross_profit_by_sector.png", p_ad_gp,
       width = 12, height = 8, dpi = 300, bg = "white")
message("Saved ../AMimages/ad_gross_profit_by_sector.png")

# Page 17: Ad / Net Income (exclude outlier sectors)
ad_ni_excluded_sectors <- c("Educational Services", "Utilities", "Accommodation & Food")
sector_ratios_ad_ni <- sector_ratios |>
  filter(!sector_label %in% ad_ni_excluded_sectors)
p_ad_ni <- plot_sector_with_all(
  sector_ratios_ad_ni, all_ind_ratios, "ad_net_income_ratio",
  "Advertising / Net Income by Sector, Active U.S. Corporations",
  "Advertising / Net Income", "Advertising / Net Income",
  is_ratio = TRUE, allow_negative = TRUE
)

# Page 18: Net Income / Gross Profit (exclude outlier sectors)
ni_gp_excluded_sectors <- c("Finance & Insurance", "Management of Companies", "Mining")
sector_ratios_ni_gp <- sector_ratios |>
  filter(!sector_label %in% ni_gp_excluded_sectors)
p_ni_gp <- plot_sector_with_all(
  sector_ratios_ni_gp, all_ind_ratios, "net_income_gross_profit_ratio",
  "Net Income / Gross Profit by Sector, Active U.S. Corporations",
  "Net Income / Gross Profit", "Net Income / Gross Profit",
  is_ratio = TRUE, allow_negative = TRUE
)

section_b_plots <- list(p_gpm, p_ad_rev, p_ad_gp, p_ad_ni, p_ni_gp)

################################################################################
# Generate Section C: Group Breakouts - Sectors (10 pages)
################################################################################

message("Generating Section C: Group-sector breakouts...")

ratio_configs_c <- list(
  list(var = "ad_revenue_ratio", label = "Advertising / Revenue",
       title_prefix = "Ad/Revenue Ratio", allow_neg = FALSE),
  list(var = "ad_gross_profit_ratio", label = "Advertising / Gross Profit",
       title_prefix = "Ad/Gross Profit Ratio", allow_neg = FALSE)
)
```

```
      paste0(rc$title_prefix, ": ", grp),
      rc$label, rc$label, allow_negative = rc$allow_neg
    )
    section_c_plots <- c(section_c_plots, list(p))
  }
}

################################################################################
# Plot function: Subsectors within a single sector
################################################################################

plot_sector_subsectors <- function(subsector_df, sector_name, ratio_var,
                                   ratio_label, title) {

  # Filter to subsectors of this sector
  plot_data <- subsector_df |>
    filter(sector == sector_name, is.finite(.data[[ratio_var]]))

  if (nrow(plot_data) == 0) {
    return(NULL)
  }

  n_subsectors <- n_distinct(plot_data$subsector)

  # Dynamic color palette
  if (n_subsectors <= 12) {
    subsector_colors <- scales::hue_pal()(n_subsectors)
  } else if (n_subsectors <= 24) {
    subsector_colors <- c(scales::hue_pal()(12),
                          scales::hue_pal(h = c(0, 360) + 15)(n_subsectors - 12))
  } else {
    subsector_colors <- colorRampPalette(
      c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd",
        "#8c564b", "#e377c2", "#7f7f7f", "#bcbd22", "#17becf")
    )(n_subsectors)
  }

  # Dynamic shape palette (cycle)
  base_shapes <- c(16, 17, 15, 18, 4, 8, 3, 7, 9, 10, 11, 12, 13, 14)
  subsector_shapes <- rep(base_shapes, length.out = n_subsectors)

  # Order by most recent available value and truncate names
  # Use mean ratio as fallback for subsectors missing the max year
  subsector_order <- plot_data |>
    group_by(subsector) |>
    summarise(order_val = mean(.data[[ratio_var]], na.rm = TRUE), .groups = "drop") |>
    arrange(desc(order_val)) |>
    pull(subsector)

  plot_data <- plot_data |>
    mutate(
      subsector_short = str_trunc(subsector, 40),
      subsector_short = factor(subsector_short,
                               levels = unique(str_trunc(subsector_order, 40)))
    )

  names(subsector_colors) <- levels(plot_data$subsector_short)
  names(subsector_shapes) <- levels(plot_data$subsector_short)

  # Legend layout
  if (n_subsectors <= 8) {
    legend_ncol <- 2
  } else if (n_subsectors <= 16) {
    legend_ncol <- 3
```

```
  p <- ggplot(plot_data, aes(x = year, y = .data[[ratio_var]],
                             color = subsector_short, shape = subsector_short)) +
    geom_line(linewidth = 0.6) +
    geom_point(size = 1.5) +
    scale_y_continuous(labels = scales::label_percent(),
                       limits = c(0, NA),
                       expand = expansion(mult = c(0, 0.05))) +
    scale_x_continuous(breaks = 2014:2022, expand = expansion(add = 0.3)) +
    scale_color_manual(values = subsector_colors) +
    scale_shape_manual(values = subsector_shapes) +
    labs(
      title = title,
      subtitle = paste0(n_subsectors, " subsectors"),
      x = "Year",
      y = ratio_label,
      color = "Subsector",
      shape = "Subsector",
      caption = paste0("Source: IRS Statistics of Income, Table 5.1\n",
                       "Ordered by 2022 value")
    ) +
    theme_soi() +
    theme(legend.text = element_text(size = 6)) +
    guides(color = guide_legend(ncol = legend_ncol),
           shape = guide_legend(ncol = legend_ncol))

  return(p)
}

################################################################################
# Generate Section D: Subsector Breakouts by Sector
################################################################################

message("Generating Section D: Subsector breakouts by sector...")

# Get list of sectors that have subsectors
sectors_with_subsectors <- subsector_wide |>
  filter(!is.na(subsector)) |>
  distinct(sector) |>
  pull(sector)

# Generate Ad/Revenue plots for all sectors
section_d_ad_rev <- list()
for (sec in sectors_with_subsectors) {
  sec_label <- sector_group_definitions |>
    filter(sector == sec) |>
    pull(sector_label)
  p <- plot_sector_subsectors(
    subsector_wide, sec, "ad_revenue_ratio",
    "Advertising / Revenue",
    paste0("Ad/Revenue: ", sec_label)
  )
  if (!is.null(p)) {
    section_d_ad_rev <- c(section_d_ad_rev, list(p))
  }
}

# Generate Ad/Gross Profit plots for all sectors
section_d_ad_gp <- list()
for (sec in sectors_with_subsectors) {
  sec_label <- sector_group_definitions |>
    filter(sector == sec) |>
    pull(sector_label)
  p <- plot_sector_subsectors(
    subsector_wide, sec, "ad_gross_profit_ratio",
```

```
}

section_d_plots <- c(section_d_ad_rev, section_d_ad_gp)
message("  Generated ", length(section_d_plots), " subsector-by-sector plots")

################################################################################
# Generate Appendix: All R Scripts
################################################################################

message("Generating appendix (all R scripts)...")

# Scripts to include in order of execution
script_files <- c(
  "parse_soi.R",           # 1. Download and parse data
  "sector_groupings.R",    # 2. Define sector groups, colors, shapes
  "plot_soi_unified.R"     # 3. Generate visualizations
)

lines_per_page <- 70  # Approximate lines per page with small font
appendix_plots <- list()
global_page <- 0

for (script_file in script_files) {
  script_lines <- readLines(script_file)
  n_pages <- ceiling(length(script_lines) / lines_per_page)

  for (i in seq_len(n_pages)) {
    global_page <- global_page + 1
    start_line <- (i - 1) * lines_per_page + 1
    end_line <- min(i * lines_per_page, length(script_lines))
    page_lines <- script_lines[start_line:end_line]
    page_text <- paste(page_lines, collapse = "\n")

    p <- ggplot() +
      annotate("text", x = 0, y = 1, label = page_text,
               size = 2, family = "FiraMono", hjust = 0, vjust = 1) +
      xlim(-0.02, 1) + ylim(0, 1.02) +
      labs(title = paste0("Appendix: ", script_file, " (Page ", i, " of ", n_pages, ")")) +
      theme_void() +
      theme(
        plot.title = element_text(size = 10, face = "bold", family = "Tinos"),
        plot.margin = margin(10, 10, 10, 10)
      )

    appendix_plots <- c(appendix_plots, list(p))
  }
  message("  ", script_file, ": ", n_pages, " pages")
}

message("  Total appendix pages: ", length(appendix_plots))

################################################################################
# Combine all plots
################################################################################

all_plots <- c(list(p_title_1, p_title_2), section_a_plots, section_b_plots, section_c_plots, section_d_plots, appendix_plots)
message("Total plots: ", length(all_plots))

################################################################################
# Save to PDF
################################################################################

message("Saving to PDF...")
```

Appendix: plot_soi_unified.R (Page 15 of 15)

```
  }
})
dev.off()

message("Saved ", length(all_plots), " pages to ", output_file)
```