DATA 607: Week 6

Cindy Lin and William Forero

2025-03-05

INTRODUCTION

Week 6 project's goal is to transform three untidy data set and tidy/transform them and generate the discussed analysis. Data Set#1 - tax dataset for 2022

Loading library

```r
library (tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library (janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

I am loading the tidyverse library because there are functions that can help tidy the loaded data. Janitor library is needed for the clean_names function.

Loading the data and removing rows

```
get_data <- read.csv("22.csv", skip = 3)
# Skip the first row to get the header
get_data <- get_data %>% select(-c(22:25))
# Remove duplicate entries
glimpse(get_data)
```

```
## Rows: 198
## Columns: 21
## $ Size.and.accumulated.size.of.adjusted.gross.income <chr> "", "", "", "", "",~
## $ All.returns                                        <chr> "Number\nof\nreturn~
## $ X                                                  <chr> "Percent\nof total"~
## $ X.1                                                <chr> "Adjusted gross\nin~
## $ X.2                                                <chr> "", "", "Percent\no~
## $ X.3                                                <chr> "", "", "Average\n(~
## $ Taxable.returns                                    <chr> "Number\nof\nreturn~
## $ X.4                                                <chr> "Percent\nof total"~
## $ X.5                                                <chr> "Adjusted gross\nin~
## $ X.6                                                <chr> "", "", "Percent\no~
## $ X.7                                                <chr> "Taxable income", "~
## $ X.8                                                <chr> "", "Amount", "", "~
## $ X.9                                                <chr> "", "Percent of\nto~
## $ X.10                                               <chr> "Income tax after c~
## $ X.11                                               <chr> "", "Amount", "", "~
## $ X.12                                               <chr> "", "Percent of\nto~
## $ X.13                                               <chr> "Total income tax",~
## $ X.14                                               <chr> "", "Percent of", "~
## $ X.15                                               <chr> "", "", "Taxable\ni~
## $ X.16                                               <chr> "", "", "Adjusted g~
## $ X.17                                               <chr> "", "Average\ntotal~
```

Loading the data

Renaming columns

```
colnames(get_data) <- c(
"Adjusted.gross.income",
"All_Returns_Count",
"All_Returns_Percent",
"All_Adjusted.gross.income.less.deficit_Amount",
"All_Adjusted.gross.income.less.deficit_Percent",
"All_Average.adjusted.gross.income.less.deficit_Amount",
"Taxable_Returns_Amount",
"Taxable_Returns_Percent",
"Taxable_Adjusted.gross.income.less.deficit_Amount",
"Taxable_Adjusted.gross.income.less.deficit_Percent",
"Taxable_Taxable.Income_Count",
"Taxable_Taxable.Income_Amount",
"Taxable_Taxable.Income_Percent",
"Taxable_Income.tax.after.credits_Count",
"Taxable_Income.tax.after.credits_Amount",
```

```
"Taxable_Income.tax.after.credits_Percent",
"Taxable_Total.Income.Tax_Amount",
"Taxable_Total_Percent",
"Taxable_Total_Taxable.income_Percent",
"Taxable_Total_Adjusted.gross.income.less.deficit_Percent",
"Taxable_Average.total.income.tax_Amount"
)
# renamed the columns
df <- get_data[-c(27:198), ]
# removed empty rows
df <- df[-c(1:7), ]
#removed row 1 - 7 (not needed information)
df <- df %>% clean_names()
#clean the column names to make it consistent
glimpse(df)
```

```
## Rows: 19
## Columns: 21
## $ adjusted_gross_income                                 <chr> "No adjusted ~
## $ all_returns_count                                     <chr> "3,254,225", ~
## $ all_returns_percent                                   <chr> "2.0", "5.1",~
## $ all_adjusted_gross_income_less_deficit_amount         <chr> "-164,160,281~
## $ all_adjusted_gross_income_less_deficit_percent        <chr> "[1]", "0.1",~
## $ all_average_adjusted_gross_income_less_deficit_amount <chr> "-50,445", "2~
## $ taxable_returns_amount                                <chr> "3,843", "105~
## $ taxable_returns_percent                               <chr> "[2]", "0.1",~
## $ taxable_adjusted_gross_income_less_deficit_amount     <chr> "-9,734,295",~
## $ taxable_adjusted_gross_income_less_deficit_percent    <chr> "[1]", "[2]",~
## $ taxable_taxable_income_count                          <chr> "0", "101,922~
## $ taxable_taxable_income_amount                         <chr> "0", "150,163~
## $ taxable_taxable_income_percent                        <chr> "0.0", "[2]",~
## $ taxable_income_tax_after_credits_count                <chr> "3,843", "105~
## $ taxable_income_tax_after_credits_amount               <chr> "128,418", "1~
## $ taxable_income_tax_after_credits_percent              <chr> "[2]", "[2]",~
## $ taxable_total_income_tax_amount                       <chr> "128,418", "1~
## $ taxable_total_percent                                 <chr> "[2]", "[2]",~
## $ taxable_total_taxable_income_percent                  <chr> "[1]", "12.5"~
## $ taxable_total_adjusted_gross_income_less_deficit_percent <chr> "[1]", "5.6",~
## $ taxable_average_total_income_tax_amount               <chr> "33,416", "17~
```

Renamed the columns as this is important for the next step when I have to format the structure to a long format. The column names will point to how othe structure will be. For example, the period and underscore of the naming convention is intentional as the underscore serves as the separator.

Long format

```
df_tidy <- df %>%
pivot_longer(
cols = -`adjusted_gross_income`,
names_to = c("Return_Type", "Category", "Metric"),
```

```r
  names_pattern = "^([^_]+)_(.*)_([^_]+)$",
  values_to = "Value"
)


knitr::kable(df_tidy) %>%
  kableExtra::kable_styling(font_size = 6)
```

adjusted_gross_income	Return_Type	Category	Metric	Value
No adjusted gross income	all	returns	count	3,254,225
No adjusted gross income	all	returns	percent	2.0
No adjusted gross income	all	adjusted_gross_income_less_deficit	amount	-164,160,281
No adjusted gross income	all	adjusted_gross_income_less_deficit	percent	[1]
No adjusted gross income	all	average_adjusted_gross_income_less_deficit	amount	-50,445
No adjusted gross income	taxable	returns	amount	3,843
No adjusted gross income	taxable	returns	percent	[2]
No adjusted gross income	taxable	adjusted_gross_income_less_deficit	amount	-9,734,295
No adjusted gross income	taxable	adjusted_gross_income_less_deficit	percent	[1]
No adjusted gross income	taxable	taxable_income	count	0
No adjusted gross income	taxable	taxable_income	amount	0
No adjusted gross income	taxable	taxable_income	percent	0.0
No adjusted gross income	taxable	income_tax_after_credits	count	3,843
No adjusted gross income	taxable	income_tax_after_credits	amount	128,418
No adjusted gross income	taxable	income_tax_after_credits	percent	[2]
No adjusted gross income	taxable	total_income_tax	amount	128,418
No adjusted gross income	taxable	total	percent	[2]
No adjusted gross income	taxable	total_taxable_income	percent	[1]
No adjusted gross income	taxable	total_adjusted_gross_income_less_deficit	percent	[1]
No adjusted gross income	taxable	average_total_income_tax	amount	33,416
$1 under $5,000	all	returns	count	8,195,783
$1 under $5,000	all	returns	percent	5.1
$1 under $5,000	all	adjusted_gross_income_less_deficit	amount	19,538,869
$1 under $5,000	all	adjusted_gross_income_less_deficit	percent	0.1
$1 under $5,000	all	average_adjusted_gross_income_less_deficit	amount	2,384
$1 under $5,000	taxable	returns	amount	105,475
$1 under $5,000	taxable	returns	percent	0.1
$1 under $5,000	taxable	adjusted_gross_income_less_deficit	amount	331,665
$1 under $5,000	taxable	adjusted_gross_income_less_deficit	percent	[2]
$1 under $5,000	taxable	taxable_income	count	101,922
$1 under $5,000	taxable	taxable_income	amount	150,163
$1 under $5,000	taxable	taxable_income	percent	[2]
$1 under $5,000	taxable	income_tax_after_credits	count	105,475
$1 under $5,000	taxable	income_tax_after_credits	amount	18,734
$1 under $5,000	taxable	income_tax_after_credits	percent	[2]
$1 under $5,000	taxable	total_income_tax	amount	18,735
$1 under $5,000	taxable	total	percent	[2]
$1 under $5,000	taxable	total_taxable_income	percent	12.5
$1 under $5,000	taxable	total_adjusted_gross_income_less_deficit	percent	5.6
$1 under $5,000	taxable	average_total_income_tax	amount	178
$5,000 under $10,000	all	returns	count	8,747,727
$5,000 under $10,000	all	returns	percent	5.4
$5,000 under $10,000	all	adjusted_gross_income_less_deficit	amount	66,070,756
$5,000 under $10,000	all	adjusted_gross_income_less_deficit	percent	0.4
$5,000 under $10,000	all	average_adjusted_gross_income_less_deficit	amount	7,553
$5,000 under $10,000	taxable	returns	amount	119,109
$5,000 under $10,000	taxable	returns	percent	0.1
$5,000 under $10,000	taxable	adjusted_gross_income_less_deficit	amount	878,182
$5,000 under $10,000	taxable	adjusted_gross_income_less_deficit	percent	[2]
$5,000 under $10,000	taxable	taxable_income	count	119,102
$5,000 under $10,000	taxable	taxable_income	amount	404,709
$5,000 under $10,000	taxable	taxable_income	percent	[2]
$5,000 under $10,000	taxable	income_tax_after_credits	count	119,109
$5,000 under $10,000	taxable	income_tax_after_credits	amount	41,423
$5,000 under $10,000	taxable	income_tax_after_credits	percent	[2]
$5,000 under $10,000	taxable	total_income_tax	amount	41,423
$5,000 under $10,000	taxable	total	percent	[2]
$5,000 under $10,000	taxable	total_taxable_income	percent	10.2
$5,000 under $10,000	taxable	total_adjusted_gross_income_less_deficit	percent	4.7
$5,000 under $10,000	taxable	average_total_income_tax	amount	348

$10,000 under $15,000	all	returns	count	9,642,322
$10,000 under $15,000	all	returns	percent	6.0
$10,000 under $15,000	all	adjusted_gross_income_less_deficit	amount	120,502,825
$10,000 under $15,000	all	adjusted_gross_income_less_deficit	percent	0.8
$10,000 under $15,000	all	average_adjusted_gross_income_less_deficit	amount	12,497
$10,000 under $15,000	taxable	returns	amount	1,370,355
$10,000 under $15,000	taxable	returns	percent	1.2
$10,000 under $15,000	taxable	adjusted_gross_income_less_deficit	amount	19,112,665
$10,000 under $15,000	taxable	adjusted_gross_income_less_deficit	percent	0.1
$10,000 under $15,000	taxable	taxable_income	count	1,370,206
$10,000 under $15,000	taxable	taxable_income	amount	1,851,999
$10,000 under $15,000	taxable	taxable_income	percent	[2]
$10,000 under $15,000	taxable	income_tax_after_credits	count	1,370,355
$10,000 under $15,000	taxable	income_tax_after_credits	amount	188,178
$10,000 under $15,000	taxable	income_tax_after_credits	percent	[2]
$10,000 under $15,000	taxable	total_income_tax	amount	188,178
$10,000 under $15,000	taxable	total	percent	[2]
$10,000 under $15,000	taxable	total_taxable_income	percent	10.2
$10,000 under $15,000	taxable	total_adjusted_gross_income_less_deficit	percent	1.0
$10,000 under $15,000	taxable	average_total_income_tax	amount	137
$15,000 under $20,000	all	returns	count	9,058,382
$15,000 under $20,000	all	returns	percent	5.6
$15,000 under $20,000	all	adjusted_gross_income_less_deficit	amount	157,738,693
$15,000 under $20,000	all	adjusted_gross_income_less_deficit	percent	1.1
$15,000 under $20,000	all	average_adjusted_gross_income_less_deficit	amount	17,414
$15,000 under $20,000	taxable	returns	amount	4,429,445
$15,000 under $20,000	taxable	returns	percent	4.0
$15,000 under $20,000	taxable	adjusted_gross_income_less_deficit	amount	77,383,332
$15,000 under $20,000	taxable	adjusted_gross_income_less_deficit	percent	0.6
$15,000 under $20,000	taxable	taxable_income	count	4,429,434
$15,000 under $20,000	taxable	taxable_income	amount	18,548,012
$15,000 under $20,000	taxable	taxable_income	percent	0.2
$15,000 under $20,000	taxable	income_tax_after_credits	count	4,429,442
$15,000 under $20,000	taxable	income_tax_after_credits	amount	1,735,910
$15,000 under $20,000	taxable	income_tax_after_credits	percent	0.1
$15,000 under $20,000	taxable	total_income_tax	amount	1,736,237
$15,000 under $20,000	taxable	total	percent	0.1
$15,000 under $20,000	taxable	total_taxable_income	percent	9.4
$15,000 under $20,000	taxable	total_adjusted_gross_income_less_deficit	percent	2.2
$15,000 under $20,000	taxable	average_total_income_tax	amount	392
$20,000 under $25,000	all	returns	count	8,035,277
$20,000 under $25,000	all	returns	percent	5.0
$20,000 under $25,000	all	adjusted_gross_income_less_deficit	amount	180,311,082
$20,000 under $25,000	all	adjusted_gross_income_less_deficit	percent	1.2
$20,000 under $25,000	all	average_adjusted_gross_income_less_deficit	amount	22,440
$20,000 under $25,000	taxable	returns	amount	4,393,888
$20,000 under $25,000	taxable	returns	percent	4.0
$20,000 under $25,000	taxable	adjusted_gross_income_less_deficit	amount	98,787,235
$20,000 under $25,000	taxable	adjusted_gross_income_less_deficit	percent	0.7
$20,000 under $25,000	taxable	taxable_income	count	4,392,825
$20,000 under $25,000	taxable	taxable_income	amount	39,939,199
$20,000 under $25,000	taxable	taxable_income	percent	0.4
$20,000 under $25,000	taxable	income_tax_after_credits	count	4,393,888
$20,000 under $25,000	taxable	income_tax_after_credits	amount	3,804,019
$20,000 under $25,000	taxable	income_tax_after_credits	percent	0.2
$20,000 under $25,000	taxable	total_income_tax	amount	3,804,019
$20,000 under $25,000	taxable	total	percent	0.2
$20,000 under $25,000	taxable	total_taxable_income	percent	9.5
$20,000 under $25,000	taxable	total_adjusted_gross_income_less_deficit	percent	3.9
$20,000 under $25,000	taxable	average_total_income_tax	amount	866
$25,000 under $30,000	all	returns	count	8,005,289
$25,000 under $30,000	all	returns	percent	5.0
$25,000 under $30,000	all	adjusted_gross_income_less_deficit	amount	219,958,083
$25,000 under $30,000	all	adjusted_gross_income_less_deficit	percent	1.5
$25,000 under $30,000	all	average_adjusted_gross_income_less_deficit	amount	27,477
$25,000 under $30,000	taxable	returns	amount	4,801,988
$25,000 under $30,000	taxable	returns	percent	4.3
$25,000 under $30,000	taxable	adjusted_gross_income_less_deficit	amount	132,247,818
$25,000 under $30,000	taxable	adjusted_gross_income_less_deficit	percent	0.9
$25,000 under $30,000	taxable	taxable_income	count	4,801,879
$25,000 under $30,000	taxable	taxable_income	amount	64,167,268
$25,000 under $30,000	taxable	taxable_income	percent	0.6

$25,000 under $30,000	taxable	income_tax_after_credits	count	4,801,988
$25,000 under $30,000	taxable	income_tax_after_credits	amount	6,376,403
$25,000 under $30,000	taxable	income_tax_after_credits	percent	0.3
$25,000 under $30,000	taxable	total_income_tax	amount	6,376,403
$25,000 under $30,000	taxable	total	percent	0.3
$25,000 under $30,000	taxable	total_taxable_income	percent	9.9
$25,000 under $30,000	taxable	total_adjusted_gross_income_less_deficit	percent	4.8
$25,000 under $30,000	taxable	average_total_income_tax	amount	1,328
$30,000 under $40,000	all	returns	count	15,771,561
$30,000 under $40,000	all	returns	percent	9.8
$30,000 under $40,000	all	adjusted_gross_income_less_deficit	amount	550,490,150
$30,000 under $40,000	all	adjusted_gross_income_less_deficit	percent	3.7
$30,000 under $40,000	all	average_adjusted_gross_income_less_deficit	amount	34,904
$30,000 under $40,000	taxable	returns	amount	10,791,934
$30,000 under $40,000	taxable	returns	percent	9.8
$30,000 under $40,000	taxable	adjusted_gross_income_less_deficit	amount	377,838,200
$30,000 under $40,000	taxable	adjusted_gross_income_less_deficit	percent	2.7
$30,000 under $40,000	taxable	taxable_income	count	10,791,890
$30,000 under $40,000	taxable	taxable_income	amount	210,765,535
$30,000 under $40,000	taxable	taxable_income	percent	1.8
$30,000 under $40,000	taxable	income_tax_after_credits	count	10,791,931
$30,000 under $40,000	taxable	income_tax_after_credits	amount	21,322,189
$30,000 under $40,000	taxable	income_tax_after_credits	percent	1.0
$30,000 under $40,000	taxable	total_income_tax	amount	21,322,304
$30,000 under $40,000	taxable	total	percent	1.0
$30,000 under $40,000	taxable	total_taxable_income	percent	10.1
$30,000 under $40,000	taxable	total_adjusted_gross_income_less_deficit	percent	5.6
$30,000 under $40,000	taxable	average_total_income_tax	amount	1,976
$40,000 under $50,000	all	returns	count	13,255,063
$40,000 under $50,000	all	returns	percent	8.2
$40,000 under $50,000	all	adjusted_gross_income_less_deficit	amount	593,955,496
$40,000 under $50,000	all	adjusted_gross_income_less_deficit	percent	4.0
$40,000 under $50,000	all	average_adjusted_gross_income_less_deficit	amount	44,810
$40,000 under $50,000	taxable	returns	amount	10,680,343
$40,000 under $50,000	taxable	returns	percent	9.7
$40,000 under $50,000	taxable	adjusted_gross_income_less_deficit	amount	479,522,952
$40,000 under $50,000	taxable	adjusted_gross_income_less_deficit	percent	3.4
$40,000 under $50,000	taxable	taxable_income	count	10,680,230
$40,000 under $50,000	taxable	taxable_income	amount	304,326,431
$40,000 under $50,000	taxable	taxable_income	percent	2.7
$40,000 under $50,000	taxable	income_tax_after_credits	count	10,680,343
$40,000 under $50,000	taxable	income_tax_after_credits	amount	30,308,339
$40,000 under $50,000	taxable	income_tax_after_credits	percent	1.4
$40,000 under $50,000	taxable	total_income_tax	amount	30,309,343
$40,000 under $50,000	taxable	total	percent	1.4
$40,000 under $50,000	taxable	total_taxable_income	percent	10.0
$40,000 under $50,000	taxable	total_adjusted_gross_income_less_deficit	percent	6.3
$40,000 under $50,000	taxable	average_total_income_tax	amount	2,838
$50,000 under $75,000	all	returns	count	23,805,797
$50,000 under $75,000	all	returns	percent	14.8
$50,000 under $75,000	all	adjusted_gross_income_less_deficit	amount	1,466,486,055
$50,000 under $75,000	all	adjusted_gross_income_less_deficit	percent	9.9
$50,000 under $75,000	all	average_adjusted_gross_income_less_deficit	amount	61,602
$50,000 under $75,000	taxable	returns	amount	21,378,339
$50,000 under $75,000	taxable	returns	percent	19.3
$50,000 under $75,000	taxable	adjusted_gross_income_less_deficit	amount	1,321,300,231
$50,000 under $75,000	taxable	adjusted_gross_income_less_deficit	percent	9.4
$50,000 under $75,000	taxable	taxable_income	count	21,378,333
$50,000 under $75,000	taxable	taxable_income	amount	926,333,771
$50,000 under $75,000	taxable	taxable_income	percent	8.1
$50,000 under $75,000	taxable	income_tax_after_credits	count	21,378,035
$50,000 under $75,000	taxable	income_tax_after_credits	amount	100,103,566
$50,000 under $75,000	taxable	income_tax_after_credits	percent	4.8
$50,000 under $75,000	taxable	total_income_tax	amount	100,103,599
$50,000 under $75,000	taxable	total	percent	4.7
$50,000 under $75,000	taxable	total_taxable_income	percent	10.8
$50,000 under $75,000	taxable	total_adjusted_gross_income_less_deficit	percent	7.6
$50,000 under $75,000	taxable	average_total_income_tax	amount	4,682
$75,000 under $100,000	all	returns	count	15,181,035
$75,000 under $100,000	all	returns	percent	9.4
$75,000 under $100,000	all	adjusted_gross_income_less_deficit	amount	1,315,619,561
$75,000 under $100,000	all	adjusted_gross_income_less_deficit	percent	8.9

$75,000 under $100,000	all	average_adjusted_gross_income_less_deficit	amount	86,662
$75,000 under $100,000	taxable	returns	amount	14,549,648
$75,000 under $100,000	taxable	returns	percent	13.2
$75,000 under $100,000	taxable	adjusted_gross_income_less_deficit	amount	1,262,289,876
$75,000 under $100,000	taxable	adjusted_gross_income_less_deficit	percent	9.0
$75,000 under $100,000	taxable	taxable_income	count	14,549,641
$75,000 under $100,000	taxable	taxable_income	amount	946,047,610
$75,000 under $100,000	taxable	taxable_income	percent	8.3
$75,000 under $100,000	taxable	income_tax_after_credits	count	14,549,341
$75,000 under $100,000	taxable	income_tax_after_credits	amount	113,079,178
$75,000 under $100,000	taxable	income_tax_after_credits	percent	5.4
$75,000 under $100,000	taxable	total_income_tax	amount	113,079,420
$75,000 under $100,000	taxable	total	percent	5.3
$75,000 under $100,000	taxable	total_taxable_income	percent	12.0
$75,000 under $100,000	taxable	total_adjusted_gross_income_less_deficit	percent	9.0
$75,000 under $100,000	taxable	average_total_income_tax	amount	7,772
$100,000 under $200,000	all	returns	count	25,887,136
$100,000 under $200,000	all	returns	percent	16.0
$100,000 under $200,000	all	adjusted_gross_income_less_deficit	amount	3,567,047,571
$100,000 under $200,000	all	adjusted_gross_income_less_deficit	percent	24.0
$100,000 under $200,000	all	average_adjusted_gross_income_less_deficit	amount	137,792
$100,000 under $200,000	taxable	returns	amount	25,547,389
$100,000 under $200,000	taxable	returns	percent	23.1
$100,000 under $200,000	taxable	adjusted_gross_income_less_deficit	amount	3,524,738,541
$100,000 under $200,000	taxable	adjusted_gross_income_less_deficit	percent	25.2
$100,000 under $200,000	taxable	taxable_income	count	25,547,351
$100,000 under $200,000	taxable	taxable_income	amount	2,858,175,142
$100,000 under $200,000	taxable	taxable_income	percent	25.1
$100,000 under $200,000	taxable	income_tax_after_credits	count	25,543,182
$100,000 under $200,000	taxable	income_tax_after_credits	amount	397,720,446
$100,000 under $200,000	taxable	income_tax_after_credits	percent	18.9
$100,000 under $200,000	taxable	total_income_tax	amount	397,758,377
$100,000 under $200,000	taxable	total	percent	18.6
$100,000 under $200,000	taxable	total_taxable_income	percent	13.9
$100,000 under $200,000	taxable	total_adjusted_gross_income_less_deficit	percent	11.3
$100,000 under $200,000	taxable	average_total_income_tax	amount	15,569
$200,000 under $500,000	all	returns	count	10,017,626
$200,000 under $500,000	all	returns	percent	6.2
$200,000 under $500,000	all	adjusted_gross_income_less_deficit	amount	2,891,064,828
$200,000 under $500,000	all	adjusted_gross_income_less_deficit	percent	19.5
$200,000 under $500,000	all	average_adjusted_gross_income_less_deficit	amount	288,598
$200,000 under $500,000	taxable	returns	amount	9,992,304
$200,000 under $500,000	taxable	returns	percent	9.0
$200,000 under $500,000	taxable	adjusted_gross_income_less_deficit	amount	2,884,421,405
$200,000 under $500,000	taxable	adjusted_gross_income_less_deficit	percent	20.6
$200,000 under $500,000	taxable	taxable_income	count	9,990,080
$200,000 under $500,000	taxable	taxable_income	amount	2,541,371,827
$200,000 under $500,000	taxable	taxable_income	percent	22.3
$200,000 under $500,000	taxable	income_tax_after_credits	count	9,975,881
$200,000 under $500,000	taxable	income_tax_after_credits	amount	478,105,230
$200,000 under $500,000	taxable	income_tax_after_credits	percent	22.8
$200,000 under $500,000	taxable	total_income_tax	amount	483,056,987
$200,000 under $500,000	taxable	total	percent	22.6
$200,000 under $500,000	taxable	total_taxable_income	percent	19.0
$200,000 under $500,000	taxable	total_adjusted_gross_income_less_deficit	percent	16.7
$200,000 under $500,000	taxable	average_total_income_tax	amount	48,343
$500,000 under $1,000,000	all	returns	count	1,674,608
$500,000 under $1,000,000	all	returns	percent	1.0
$500,000 under $1,000,000	all	adjusted_gross_income_less_deficit	amount	1,124,299,469
$500,000 under $1,000,000	all	adjusted_gross_income_less_deficit	percent	7.6
$500,000 under $1,000,000	all	average_adjusted_gross_income_less_deficit	amount	671,381
$500,000 under $1,000,000	taxable	returns	amount	1,672,148
$500,000 under $1,000,000	taxable	returns	percent	1.5
$500,000 under $1,000,000	taxable	adjusted_gross_income_less_deficit	amount	1,122,634,305
$500,000 under $1,000,000	taxable	adjusted_gross_income_less_deficit	percent	8.0
$500,000 under $1,000,000	taxable	taxable_income	count	1,671,381
$500,000 under $1,000,000	taxable	taxable_income	amount	1,028,823,801
$500,000 under $1,000,000	taxable	taxable_income	percent	9.0
$500,000 under $1,000,000	taxable	income_tax_after_credits	count	1,667,357
$500,000 under $1,000,000	taxable	income_tax_after_credits	amount	254,284,854
$500,000 under $1,000,000	taxable	income_tax_after_credits	percent	12.1

$500,000 under $1,000,000	taxable	total_income_tax	amount	260,282,198
$500,000 under $1,000,000	taxable	total	percent	12.2
$500,000 under $1,000,000	taxable	total_taxable_income	percent	25.3
$500,000 under $1,000,000	taxable	total_adjusted_gross_income_less_deficit	percent	23.2
$500,000 under $1,000,000	taxable	average_total_income_tax	amount	155,657
$1,000,000 under $1,500,000	all	returns	count	360,882
$1,000,000 under $1,500,000	all	returns	percent	0.2
$1,000,000 under $1,500,000	all	adjusted_gross_income_less_deficit	amount	435,177,502
$1,000,000 under $1,500,000	all	adjusted_gross_income_less_deficit	percent	2.9
$1,000,000 under $1,500,000	all	average_adjusted_gross_income_less_deficit	amount	1,205,872
$1,000,000 under $1,500,000	taxable	returns	amount	360,444
$1,000,000 under $1,500,000	taxable	returns	percent	0.3
$1,000,000 under $1,500,000	taxable	adjusted_gross_income_less_deficit	amount	434,649,011
$1,000,000 under $1,500,000	taxable	adjusted_gross_income_less_deficit	percent	3.1
$1,000,000 under $1,500,000	taxable	taxable_income	count	360,328
$1,000,000 under $1,500,000	taxable	taxable_income	amount	398,506,804
$1,000,000 under $1,500,000	taxable	taxable_income	percent	3.5
$1,000,000 under $1,500,000	taxable	income_tax_after_credits	count	359,534
$1,000,000 under $1,500,000	taxable	income_tax_after_credits	amount	110,819,453
$1,000,000 under $1,500,000	taxable	income_tax_after_credits	percent	5.3
$1,000,000 under $1,500,000	taxable	total_income_tax	amount	114,003,393
$1,000,000 under $1,500,000	taxable	total	percent	5.3
$1,000,000 under $1,500,000	taxable	total_taxable_income	percent	28.6
$1,000,000 under $1,500,000	taxable	total_adjusted_gross_income_less_deficit	percent	26.2
$1,000,000 under $1,500,000	taxable	average_total_income_tax	amount	316,286
$1,500,000 under $2,000,000	all	returns	count	148,222
$1,500,000 under $2,000,000	all	returns	percent	0.1
$1,500,000 under $2,000,000	all	adjusted_gross_income_less_deficit	amount	254,476,785
$1,500,000 under $2,000,000	all	adjusted_gross_income_less_deficit	percent	1.7
$1,500,000 under $2,000,000	all	average_adjusted_gross_income_less_deficit	amount	1,716,862
$1,500,000 under $2,000,000	taxable	returns	amount	148,056
$1,500,000 under $2,000,000	taxable	returns	percent	0.1
$1,500,000 under $2,000,000	taxable	adjusted_gross_income_less_deficit	amount	254,190,622
$1,500,000 under $2,000,000	taxable	adjusted_gross_income_less_deficit	percent	1.8
$1,500,000 under $2,000,000	taxable	taxable_income	count	147,957
$1,500,000 under $2,000,000	taxable	taxable_income	amount	232,336,257
$1,500,000 under $2,000,000	taxable	taxable_income	percent	2.0
$1,500,000 under $2,000,000	taxable	income_tax_after_credits	count	147,558
$1,500,000 under $2,000,000	taxable	income_tax_after_credits	amount	67,287,429
$1,500,000 under $2,000,000	taxable	income_tax_after_credits	percent	3.2
$1,500,000 under $2,000,000	taxable	total_income_tax	amount	69,338,677
$1,500,000 under $2,000,000	taxable	total	percent	3.2
$1,500,000 under $2,000,000	taxable	total_taxable_income	percent	29.8
$1,500,000 under $2,000,000	taxable	total_adjusted_gross_income_less_deficit	percent	27.3
$1,500,000 under $2,000,000	taxable	average_total_income_tax	amount	468,327
$2,000,000 under $5,000,000	all	returns	count	208,129
$2,000,000 under $5,000,000	all	returns	percent	0.1
$2,000,000 under $5,000,000	all	adjusted_gross_income_less_deficit	amount	621,044,404
$2,000,000 under $5,000,000	all	adjusted_gross_income_less_deficit	percent	4.2
$2,000,000 under $5,000,000	all	average_adjusted_gross_income_less_deficit	amount	2,983,940
$2,000,000 under $5,000,000	taxable	returns	amount	207,905
$2,000,000 under $5,000,000	taxable	returns	percent	0.2
$2,000,000 under $5,000,000	taxable	adjusted_gross_income_less_deficit	amount	620,386,059
$2,000,000 under $5,000,000	taxable	adjusted_gross_income_less_deficit	percent	4.4
$2,000,000 under $5,000,000	taxable	taxable_income	count	207,783
$2,000,000 under $5,000,000	taxable	taxable_income	amount	565,832,497
$2,000,000 under $5,000,000	taxable	taxable_income	percent	5.0
$2,000,000 under $5,000,000	taxable	income_tax_after_credits	count	207,401
$2,000,000 under $5,000,000	taxable	income_tax_after_credits	amount	166,026,539
$2,000,000 under $5,000,000	taxable	income_tax_after_credits	percent	7.9
$2,000,000 under $5,000,000	taxable	total_income_tax	amount	171,825,620
$2,000,000 under $5,000,000	taxable	total	percent	8.0
$2,000,000 under $5,000,000	taxable	total_taxable_income	percent	30.4
$2,000,000 under $5,000,000	taxable	total_adjusted_gross_income_less_deficit	percent	27.7
$2,000,000 under $5,000,000	taxable	average_total_income_tax	amount	826,462
$5,000,000 under $10,000,000	all	returns	count	52,968
$5,000,000 under $10,000,000	all	returns	percent	[2]
$5,000,000 under $10,000,000	all	adjusted_gross_income_less_deficit	amount	362,815,724
$5,000,000 under $10,000,000	all	adjusted_gross_income_less_deficit	percent	2.4
$5,000,000 under $10,000,000	all	average_adjusted_gross_income_less_deficit	amount	6,849,715
$5,000,000 under $10,000,000	taxable	returns	amount	52,913
$5,000,000 under $10,000,000	taxable	returns	percent	[2]

$5,000,000 under $10,000,000	taxable	adjusted_gross_income_less_deficit	amount	362,436,914
$5,000,000 under $10,000,000	taxable	adjusted_gross_income_less_deficit	percent	2.6
$5,000,000 under $10,000,000	taxable	taxable_income	count	52,880
$5,000,000 under $10,000,000	taxable	taxable_income	amount	329,349,662
$5,000,000 under $10,000,000	taxable	taxable_income	percent	2.9
$5,000,000 under $10,000,000	taxable	income_tax_after_credits	count	52,747
$5,000,000 under $10,000,000	taxable	income_tax_after_credits	amount	96,475,701
$5,000,000 under $10,000,000	taxable	income_tax_after_credits	percent	4.6
$5,000,000 under $10,000,000	taxable	total_income_tax	amount	100,262,125
$5,000,000 under $10,000,000	taxable	total	percent	4.7
$5,000,000 under $10,000,000	taxable	total_taxable_income	percent	30.4
$5,000,000 under $10,000,000	taxable	total_adjusted_gross_income_less_deficit	percent	27.7
$5,000,000 under $10,000,000	taxable	average_total_income_tax	amount	1,894,849
$10,000,000 or more	all	returns	count	34,630
$10,000,000 or more	all	returns	percent	[2]
$10,000,000 or more	all	adjusted_gross_income_less_deficit	amount	1,051,519,384
$10,000,000 or more	all	adjusted_gross_income_less_deficit	percent	7.1
$10,000,000 or more	all	average_adjusted_gross_income_less_deficit	amount	30,364,406
$10,000,000 or more	taxable	returns	amount	34,604
$10,000,000 or more	taxable	returns	percent	[2]
$10,000,000 or more	taxable	adjusted_gross_income_less_deficit	amount	1,050,594,907
$10,000,000 or more	taxable	adjusted_gross_income_less_deficit	percent	7.5
$10,000,000 or more	taxable	taxable_income	count	34,571
$10,000,000 or more	taxable	taxable_income	amount	937,403,948
$10,000,000 or more	taxable	taxable_income	percent	8.2
$10,000,000 or more	taxable	income_tax_after_credits	count	34,472
$10,000,000 or more	taxable	income_tax_after_credits	amount	251,097,008
$10,000,000 or more	taxable	income_tax_after_credits	percent	12.0
$10,000,000 or more	taxable	total_income_tax	amount	266,286,618
$10,000,000 or more	taxable	total	percent	12.4
$10,000,000 or more	taxable	total_taxable_income	percent	28.4
$10,000,000 or more	taxable	total_adjusted_gross_income_less_deficit	percent	25.3
$10,000,000 or more	taxable	average_total_income_tax	amount	7,695,255

Shaping to long format for analysis

Total tax returns filed

```r
df_filter <- df_tidy %>%
filter(Return_Type == "all", Category == "returns", Metric == "count")
# filtering taxable returns and the percentage for ggplot
df_filter <- df_filter %>%
mutate(adjusted_gross_income = fct_collapse(adjusted_gross_income,
"<25K" = c("No adjusted gross income",
"$1 under $5,000",
"$5,000 under $10,000",
"$10,000 under $15,000",
"$15,000 under $20,000",
"$20,000 under $25,000"),
"25K - <75K" = c("$25,000 under $30,000",
"$30,000 under $40,000",
"$40,000 under $50,000",
"$50,000 under $75,000"),
"75K - <200K" = c("$75,000 under $100,000",
"$100,000 under $200,000"
),
"200K+" = c("$200,000 under $500,000",
"$500,000 under $1,000,000",
"$1,000,000 under $1,500,000",
```
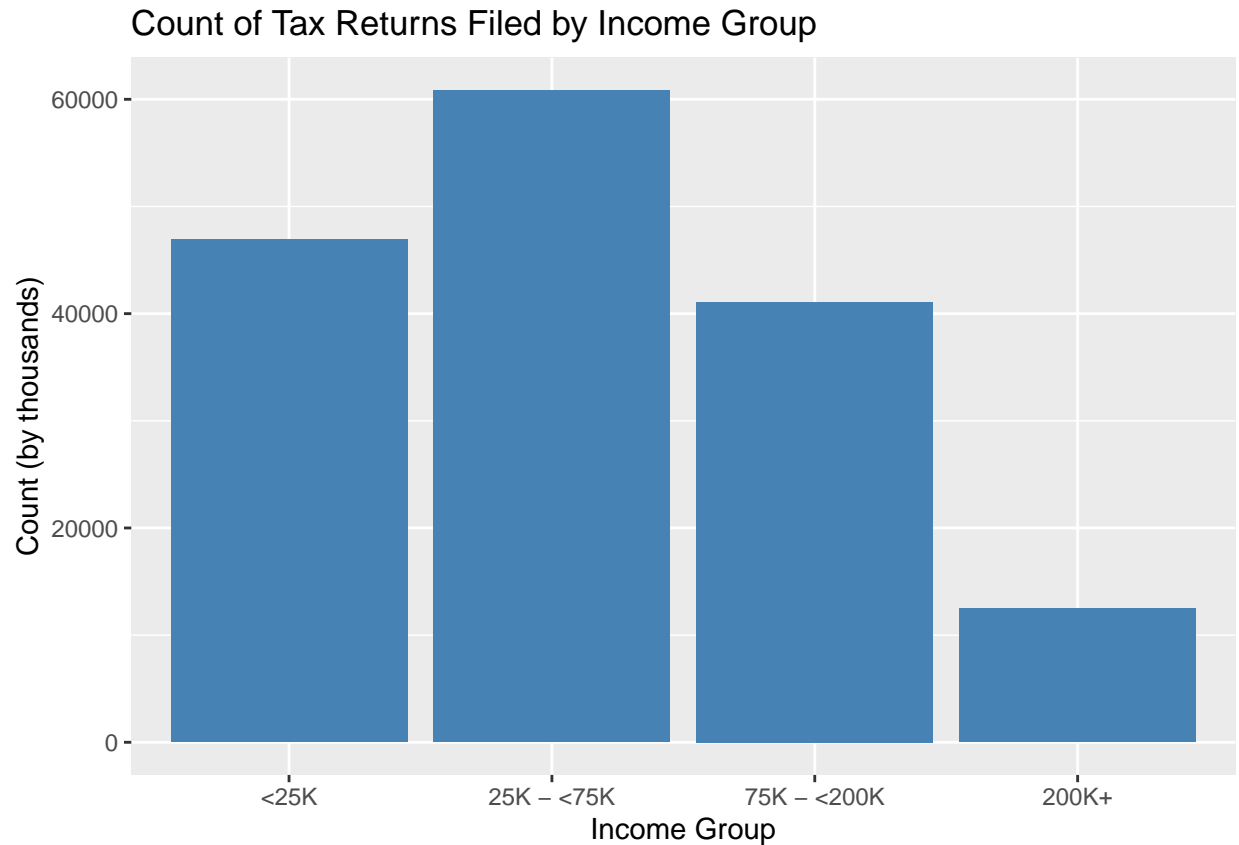
```r
"$1,500,000 under $2,000,000",
"$2,000,000 under $5,000,000",
"$5,000,000 under $10,000,000",
"$10,000,000 or more")
))
# combine the income levels to smaller groups
df_filter_sum <- df_filter %>%
mutate(Value = gsub(",", "", Value), # Remove commas
Value = as.numeric(Value)) # Convert to numeric
df_filter_sum <- df_filter_sum %>%
group_by(adjusted_gross_income) %>%
summarise(
sum_return_thous = sum(Value)/1e3,
)

df_filter_sum<- df_filter_sum %>%
mutate(adjusted_gross_income =
      factor(adjusted_gross_income, levels =
      c("<25K",
        "25K - <75K",
        "75K - <200K",
        "200K+")))

ggplot(df_filter_sum, aes(x = adjusted_gross_income, y = sum_return_thous)) +
geom_col(fill = "steelblue") +
labs(title = "Count of Tax Returns Filed by Income Group",
x = "Income Group",
y = "Count (by thousands)")
```

Count of Tax Returns Filed by Income Group

I filter the type of returns here to get the total returns filed by income group. In this dataset, there is a difference between all returns and taxable returns so in this case, we are looking at all returns. Since the source data had an overwhelming amount of income groups, I combined them to offer a more wider view of the analysis.

Effective tax rate

```
df_filter_income <- df_tidy %>%
filter(Return_Type == "taxable",
Category %in% c("total_income_tax", "taxable_income"),
Metric %in% c("amount"))

df_filter_income <- df_filter_income %>%
mutate(adjusted_gross_income = fct_collapse(adjusted_gross_income,
"<25K" = c("No adjusted gross income",
"$1 under $5,000",
"$5,000 under $10,000",
"$10,000 under $15,000",
"$15,000 under $20,000",
"$20,000 under $25,000"),
"25K - <75K" = c("$25,000 under $30,000",
"$30,000 under $40,000",
"$40,000 under $50,000",
"$50,000 under $75,000"),
```
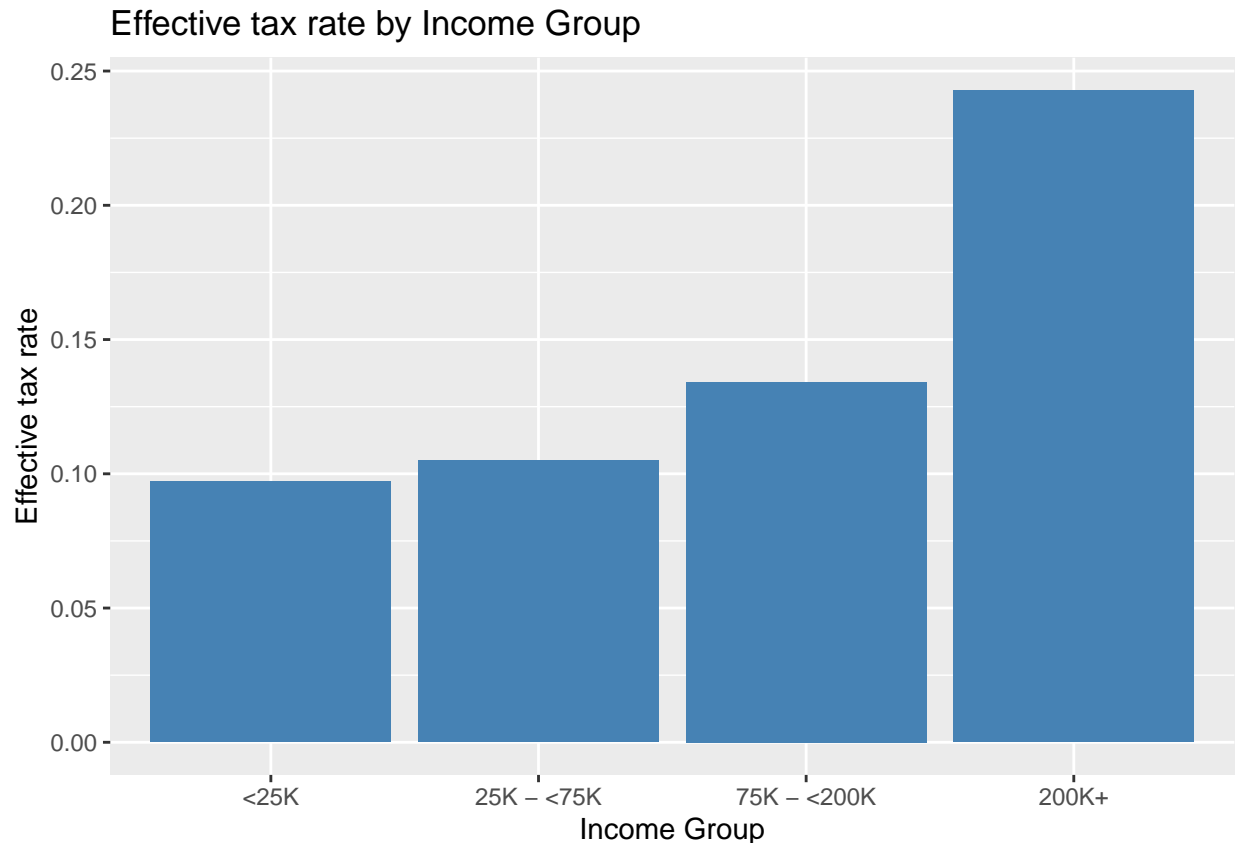
```r
"75K - <200K" = c("$75,000 under $100,000",
"$100,000 under $200,000"
),
"200K+" = c("$200,000 under $500,000",
"$500,000 under $1,000,000",
"$1,000,000 under $1,500,000",
"$1,500,000 under $2,000,000",
"$2,000,000 under $5,000,000",
"$5,000,000 under $10,000,000",
"$10,000,000 or more")
))

# combine the income levels to smaller groups
df_filter_incometax <- df_filter_income %>%
mutate(Value = gsub(",", "", Value),
Value = as.numeric(Value))
df_avg_taxes <- df_filter_incometax %>%
group_by(adjusted_gross_income) %>%
summarise(avg_paid_taxes = sum(Value[Category == "total_income_tax"], na.rm = TRUE) /
sum(Value[Category == "taxable_income"], na.rm = TRUE))

df_avg_taxes<- df_avg_taxes %>%
mutate(adjusted_gross_income = factor(adjusted_gross_income,
                                      levels = c("<25K",
                                                 "25K - <75K",
                                                 "75K - <200K",
                                                 "200K+")))


ggplot(df_avg_taxes, aes(x = adjusted_gross_income, y = avg_paid_taxes)) +
geom_col(fill = "steelblue") +
labs(title = "Effective tax rate by Income Group",
x = "Income Group",
y = "Effective tax rate")
```

Effective tax rate by Income Group

Same thing I did here with the total count of returns. The only difference here is that I also average to get the effective tax rate, which is the total income tax/total count of taxable income returns.

Conclusion

After transforming the file to a structured table that we can use for analysis, I aggregated and grouped the income categories to generate insights. One key observation is that the majority of tax filers fall within the $25K to under $75K income bracket, as indicated by the total number of returns filed for the 2022 tax year. Additionally, I calculated the effective tax rate for each income group by dividing the total income tax paid by the total number of returns, allowing for a comparative view of tax burdens across different income levels. For future analysis, it would be valuable to compare tax years, particularly during periods of major tax reform, such the Tax Cuts and Jobs Act (2018–2025). Examining changes in tax rates, income distribution, and overall tax burdens over time could provide deeper insights into the impact of tax policy. Take Away from the Assignment- The janitor package, particularly the clean_names() function, proved to be extremely useful. It ensures consistent formatting of column names, which is essential when applying pattern-based transformations or reshaping the data into a long format.