# Homework 02 - Data wrangling

Jemimah Osei (jko35)

2025-02-12

## Setup

Load packages and data:

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(scales)
```

```

Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
library(fivethirtyeight)
```

```
Some larger datasets need to be installed separately, like senators and
house_district_forecast. To install these, we recommend you install the
fivethirtyeightdata package by running:
install.packages('fivethirtyeightdata', repos =
'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

# Exercises

## Exercise 1

```r
# your code here
glimpse(college_recent_grads)
```

```
Rows: 173
Columns: 21
$ rank                        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,…
$ major_code                  <int> 2419, 2416, 2415, 2417, 2405, 2418, 6202, …
$ major                       <chr> "Petroleum Engineering", "Mining And Miner…
$ major_category              <chr> "Engineering", "Engineering", "Engineering…
$ total                       <int> 2339, 756, 856, 1258, 32260, 2573, 3777, 1…
$ sample_size                 <int> 36, 7, 3, 16, 289, 17, 51, 10, 1029, 631, …
$ men                         <int> 2057, 679, 725, 1123, 21239, 2200, 2110, 8…
$ women                       <int> 282, 77, 131, 135, 11021, 373, 1667, 960, …
$ sharewomen                  <dbl> 0.1205643, 0.1018519, 0.1530374, 0.1073132…
$ employed                    <int> 1976, 640, 648, 758, 25694, 1857, 2912, 15…
$ employed_fulltime           <int> 1849, 556, 558, 1069, 23170, 2038, 2924, 1…
$ employed_parttime           <int> 270, 170, 133, 150, 5180, 264, 296, 553, 1…
$ employed_fulltime_yearround <int> 1207, 388, 340, 692, 16697, 1449, 2482, 82…
$ unemployed                  <int> 37, 85, 16, 40, 1672, 400, 308, 33, 4650, …
$ unemployment_rate           <dbl> 0.018380527, 0.117241379, 0.024096386, 0.0…
$ p25th                       <dbl> 95000, 55000, 50000, 43000, 50000, 50000, …
$ median                      <dbl> 110000, 75000, 73000, 70000, 65000, 65000,…
$ p75th                       <dbl> 125000, 90000, 105000, 80000, 75000, 10200…
$ college_jobs                <int> 1534, 350, 456, 529, 18314, 1142, 1768, 97…
$ non_college_jobs            <int> 364, 257, 176, 102, 4440, 657, 314, 500, 1…
$ low_wage_jobs               <int> 193, 50, 0, 0, 972, 244, 259, 220, 3253, 3…
```

```r
college_recent_grads |>
  select(major, unemployment_rate) |>
  filter(!is.na(unemployment_rate)) |>
  slice_min(order_by = unemployment_rate, n = 10)
```

```
# A tibble: 10 × 2
   major                                      unemployment_rate
   <chr>                                                  <dbl>
 1 Mathematics And Computer Science                     0
 2 Military Technologies                                0
 3 Botany                                               0
 4 Soil Science                                         0
 5 Educational Administration And Supervision           0
 6 Engineering Mechanics Physics And Science            0.00633
 7 Court Reporting                                      0.0117
 8 Mathematics Teacher Education                        0.0162
 9 Petroleum Engineering                                0.0184
10 General Agriculture                                  0.0196
```

I observed that most of the majors with the lowest unemployment rate are stem majors which stems for a high demand of jobs seekers in these fields and i think the same holds for very niche fields like Botany and Soil Science

## Exercise 2

```
# your code here

college_recent_grads |>
  select(major, sharewomen) |>
  filter(!is.na(sharewomen)) |>
  arrange(desc(sharewomen)) |>
  slice_head(n = 5)
```

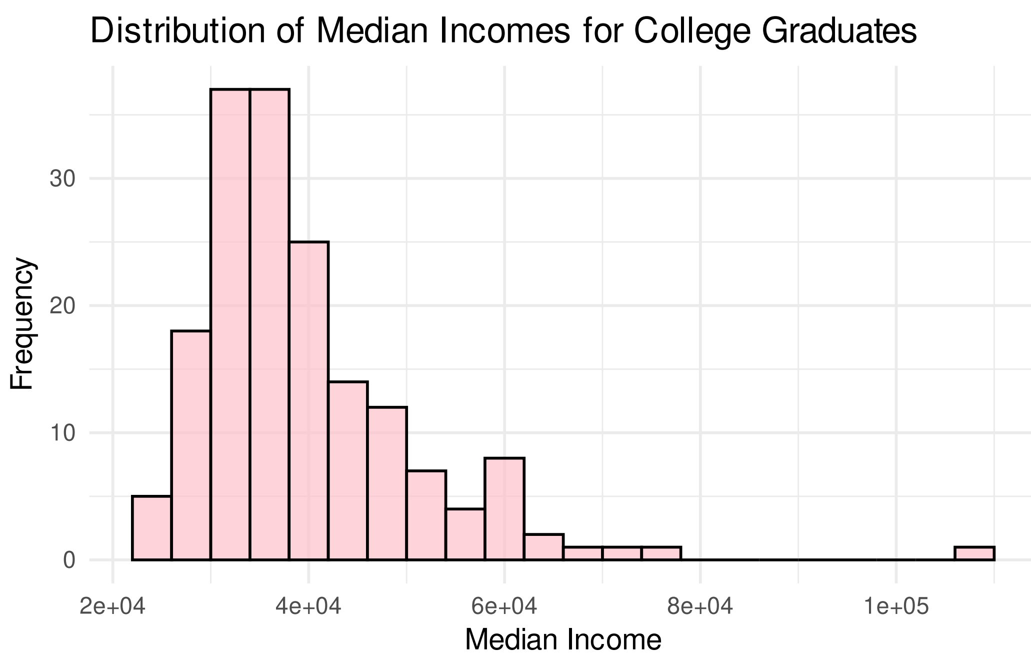
```
# A tibble: 5 × 2
  major                                         sharewomen
  <chr>                                              <dbl>
1 Early Childhood Education                          0.969
2 Communication Disorders Sciences And Services      0.968
3 Medical Assisting Services                         0.928
4 Elementary Education                               0.924
5 Family And Consumer Sciences                       0.911
```

I observed a high representation of women in certain fields, especially in the arts, education, and health. However, there is a variation that occurs across majors in STEM fields including engineering and computer science who often have lower percentages of women relative the social sciences or humanities.

## Exercise 3

a. Distribution of median incomes

```
# your code here
ggplot(mapping = aes(x = median),
       data = filter(college_recent_grads, !is.na(median))) +
  geom_histogram(binwidth = 4000, fill = "pink", color = "black", alpha = 0.7)
+
  labs(title = "Distribution of Median Incomes for College Graduates",
       x = "Median Income", y = "Frequency") +
  theme_minimal()
```




b. Mean and median for median income

```
#| label: ex3b

# your code here
college_recent_grads |>
  select(median) |>
  filter(!is.na(median)) |>
  summarise(mean_income = mean(median),
            median_income = median(median))
```

```
# A tibble: 1 × 2
  mean_income median_income
        <dbl>         <dbl>
1      40151.         36000
```

Based on the histogram, the distribution of median incomes for college graduates is right-skewed thus most of the data is concentrated at lower income levels (around $30,000–$50,000), while a few higher values extend the tail to the right. I therefore believe that the median income is the more useful summary statistic for describing the typical income of college graduates since the mean is inflated by a small number of high earners.

d.The distribution of median incomes for college graduates is right-skewed, with most incomes concentrated in the lower range (around $30,000 to $50,000) and a few higher-income earners extending the tail to the right. The **center** of the distribution is represented by the **median income** of $36,000, which is the more reliable measure of typical income, as it is not influenced by outliers. The **mean income** is higher, at approximately $40,151.45, reflecting the influence of high earners in the tail of the distribution.The **spread** of the data is wide, indicating significant variability in earnings. There is a concentration of incomes in the lower range, but the data extends over a broad range due to higher income earners. **Other observations** include the presence of outliers, where a small number of individuals earn significantly higher incomes, pulling the mean to the right and contributing to the right-skewed shape of the distribution. These high earners do not represent the typical college graduate's income.

## Exercise 4

a. Calculate the minimum, median, and maximum median income per major category as well as the number of majors in each category.

```
# your code here
college_recent_grads |>
  group_by(major_category) |>
  summarise(
    num_majors = n(),
    min_income = min(median, na.rm = TRUE),
   median_income = median(median, na.rm = TRUE), max_income = max(median, na.rm
= TRUE)
  ) |>
  arrange(desc(median_income))
```

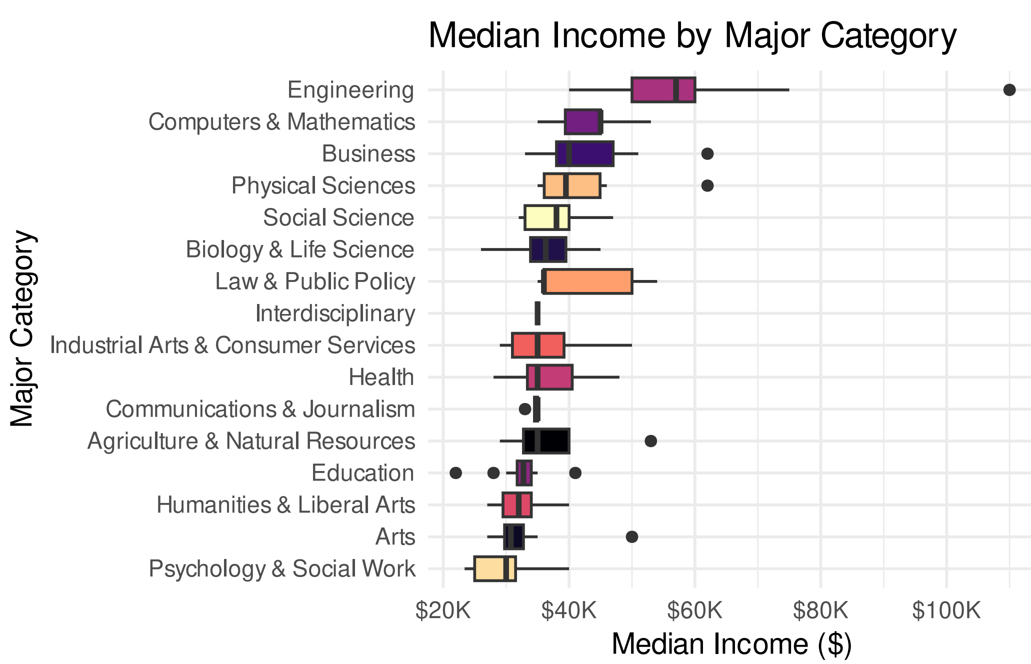
```
# A tibble: 16 × 5
   major_category                  num_majors min_income median_income max_income
   <chr>                                <int>      <dbl>         <dbl>      <dbl>
 1 Engineering                             29      40000         57000     110000
 2 Computers & Mathematics                 11      35000         45000      53000
 3 Business                                13      33000         40000      62000
 4 Physical Sciences                       10      35000         39500      62000
 5 Social Science                           9      32000         38000      47000
 6 Biology & Life Science                  14      26000         36300      45000
 7 Law & Public Policy                      5      35000         36000      54000
 8 Agriculture & Natural Resourc…          10      29000         35000      53000
```

```
 9 Communications & Journalism              4      33000          35000        35000
10 Health                                  12      28000          35000        48000
11 Industrial Arts & Consumer Se…           7      29000          35000        50000
12 Interdisciplinary                        1      35000          35000        35000
13 Education                               16      22000          32750        41000
14 Humanities & Liberal Arts               15      27000          32000        40000
15 Arts                                     8      27000          30750        50000
16 Psychology & Social Work                 9      23400          30000        40000
```

b. Create box plots of the distribution of median income by major category.

```
# your code here

ggplot(data = college_recent_grads,
       mapping = aes(x = median, y = fct_reorder(major_category, median), fill
= major_category)) +
  geom_boxplot() +
 scale_x_continuous(labels = scales::dollar_format(scale = 0.001, suffix = "K"))
+
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Median Income by Major Category",
       x = "Median Income ($)",
       y = "Major Category") +
  scale_fill_viridis_d(option = "magma")
```



c. The median incomes across major categories differ significantly, with Engineering, Business, and Computer & Mathematics majors earning the highest salaries, while Arts, Education, and Psychology & Social Work tend to have lower median incomes. As an Information Science major, I'm glad to see that my field, which falls under "Computers & Mathematics," is among the higher-earning categories, reflecting strong demand and career opportunities in tech.

## Exercise 5

```
# your code here
stem_categories <- c(
  "Biology & Life Science",
  "Computers & Mathematics",
  "Engineering",
  "Physical Sciences"
  )


college_recent_grads <- college_recent_grads |>
  mutate(major_type = if_else(major_category %in% stem_categories, "STEM", "Not
STEM"))

college_recent_grads |>
  filter(major_type == "STEM", median < 36000) |>
  select(major, median) |>
  arrange(desc(median))
```

```
# A tibble: 10 × 2
   major                                 median
   <chr>                                  <dbl>
 1 Environmental Science                  35600
 2 Multi-Disciplinary Or General Science  35000
 3 Physiology                             35000
 4 Communication Technologies             35000
 5 Neuroscience                           35000
 6 Atmospheric Sciences And Meteorology   35000
 7 Miscellaneous Biology                  33500
 8 Biology                                33400
 9 Ecology                                33000
10 Zoology                                26000
```

## Exercise 6

```
# your code here

major_income <- college_recent_grads |>
  inner_join(college_grad_students, by = "major_code")
```

```r
major_income <- major_income |>
  mutate(grad_premium = ((grad_median - median) / median) * 100)

major_income_tibble <- major_income |>
  select(major.x, grad_premium, grad_median, median) |>
  rename(undergrad_median = median)

major_income_stem <- major_income |>
  filter(str_detect(major_category.x, "Science|Engineering|Mathematics"))

top_5_grad_premium <- major_income_stem |>
  arrange(desc(grad_premium)) |>
  select(major.x, grad_premium, grad_median, median) |>
  head(5)

bottom_5_grad_premium <- major_income_stem |>
  arrange(grad_premium) |>
  select(major.x, grad_premium, grad_median, median) |>
  head(5)

major_income_tibble
```

```
# A tibble: 173 × 4
   major.x                              grad_premium grad_median undergrad_median
   <chr>                                       <dbl>       <dbl>            <dbl>
 1 Petroleum Engineering                        12.7      124000           110000
 2 Mining And Mineral Engineering               33.3      100000            75000
 3 Metallurgical Engineering                    37.0      100000            73000
 4 Naval Architecture And Marine Engi…          45.7      102000            70000
 5 Chemical Engineering                         56.9      102000            65000
 6 Nuclear Engineering                          69.2      110000            65000
 7 Actuarial Science                            77.4      110000            62000
 8 Astronomy And Astrophysics                   54.8       96000            62000
 9 Mechanical Engineering                       66.7      100000            60000
10 Electrical Engineering                       76.7      106000            60000
# ℹ 163 more rows
```

```r
top_5_grad_premium
```

```
# A tibble: 5 × 4
  major.x              grad_premium grad_median median
  <chr>                       <dbl>       <dbl>  <dbl>
1 Zoology                      323.      110000  26000
2 Biology                      184.       95000  33400
```

```
3 Physiology                     157.       90000  35000
4 Biochemical Sciences           157.       96000  37400
5 Chemistry                      156.      100000  39000
```

```
bottom_5_grad_premium
```

```
# A tibble: 5 × 4
  major.x                         grad_premium grad_median median
  <chr>                                  <dbl>       <dbl>  <dbl>
1 Petroleum Engineering                   12.7      124000 110000
2 Mining And Mineral Engineering          33.3      100000  75000
3 Metallurgical Engineering               37.0      100000  73000
4 Biological Engineering                  40.1       80000  57100
5 Architectural Engineering               44.4       78000  54000
```

I observed that some engineering majors, like Mining and Mineral Engineering, and Metallurgical Engineering, have moderate grad premiums. While these fields do see a salary boost with an advanced degree, the difference isn't as pronounced compared to more specialized STEM fields like Actuarial Science or Aerospace Engineering. On the other hand, fields like Naval Architecture, and Nuclear Engineering show significant grad premiums, with salaries increasing by over 50%. This shows that the specialized engineering fields see a substantial income boost from graduate education, highlighting the added value of advanced degrees in these careers.