DATA*6200 Assignment 1

AUTHOR
Rakesh Das

PUBLISHED
October 7, 2024

Installing the required packages

Before we begin looking at R code, we will first quickly look at packages.

```
# Install necessary packages if not already installed
# install.packages("dplyr")
# install.packages("ggplot2")
# install.packages("tidyverse")
# install.packages("readxl")
# install.packages("stringr")
# install.packages("tidyr")

# Load required libraries
library(lobstr)
library(dplyr)       # Core tidyverse package for data manipulation
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)     # Core visualization package
library(tidyverse)   # Collection of packages for data science
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ lubridate 1.9.3     ✓ tibble    3.2.1
✓ purrr     1.0.2     ✓ tidyr     1.3.1
✓ readr     2.1.5

── 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
```

```
library(readxl)      # For reading Excel files
library(stringr)     # For string manipulation
```

```
library(tidyr)       # For data tidying
library(scales)
```

```
Attaching package: 'scales'

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

    discard

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

    col_factor
```

Task

Since the beginning of the Covid-19 pandemic, the workforce has undergone rapid changes, with some industries being more affected than others. In this assignment, you will analyze salaries by industry/jobs, and investigate trends in these industries/jobs over time. You will be using the `ask_a_manager.xlsx` data set provided on Courselink (from here). This data set was created using a form with several columns being "free text", and are hence messy. Feel free to do some research on these data.

Throughout your analysis of these data, you should aim to answer the following questions:

1. Which industry or industries have the highest/lowest salaries?
2. Which industries have the highest salary variability?
3. How do salaries vary over time and geography?

To proceed with the data wrangling and analysis, We need to read the data set provided
i.e. ask_a_manager.slsx

Data Processing

Import the excel data to a data frame, "workforce"

```
# checking the current working directory
getwd()
```

```
[1] "C:/Users/ASUS/Documents/Masters - Guelph/Fall 2024/6200/assignment 1"
```

```
# Reading the Excel data into a data frame named "workforce"
workforce <- readxl::read_xlsx("ask_a_manager.xlsx")
```

```
Warning: Coercing text to numeric in F26564 / R26564C6: '00'
```

```
# Display the first few rows of the data frame to undesrtand the data
workforce
```

```
# A tibble: 28,080 × 18
   Timestamp           `How old are you?` What industry do you wor…¹ `Job title`
   <dttm>              <chr>              <chr>                      <chr>
 1 2021-04-27 11:02:09 25-34              Education (Higher Educati… Research a…
 2 2021-04-27 11:02:21 25-34              Computing or Tech          Change & I…
 3 2021-04-27 11:02:38 25-34              Accounting, Banking & Fin… Marketing …
 4 2021-04-27 11:02:40 25-34              Nonprofits                 Program Ma…
 5 2021-04-27 11:02:41 25-34              Accounting, Banking & Fin… Accounting…
 6 2021-04-27 11:02:45 25-34              Education (Higher Educati… Scholarly …
 7 2021-04-27 11:02:50 25-34              Publishing                 Publishing…
 8 2021-04-27 11:02:59 25-34              Education (Primary/Second… Librarian
 9 2021-04-27 11:03:01 45-54              Computing or Tech          Systems An…
10 2021-04-27 11:03:01 35-44              Accounting, Banking & Fin… Senior Acc…
# ℹ 28,070 more rows
# ℹ abbreviated name: ¹`What industry do you work in?`
# ℹ 14 more variables:
#   `If your job title needs additional context, please clarify here:` <chr>,
#   `What is your annual salary? (You'll indicate the currency in a later question. If you are
part-time or hourly, please enter an annualized equivalent -- what you would earn if you
worked the job 40 hours a week, 52 weeks a year.)` <dbl>,
#   `How much additional monetary compensation do you get, if any (for example, bonuses or
overtime in an average year)? Please only include monetary compensation here, not the value of
benefits.` <chr>,
#   `Please indicate the currency` <chr>, …
```

```r
# Display a glimpse of the data frame structure
glimpse(workforce)
```

```
Rows: 28,080
Columns: 18
$ Timestamp
<dttm> …
$ `How old are you?`
<chr> …
$ `What industry do you work in?`
<chr> …
$ `Job title`
<chr> …
$ `If your job title needs additional context, please clarify here:`
<chr> …
$ `What is your annual salary? (You'll indicate the currency in a later question. If you are
part-time or hourly, please enter an annualized equivalent -- what you would earn if you
worked the job 40 hours a week, 52 weeks a year.)` <dbl> …
$ `How much additional monetary compensation do you get, if any (for example, bonuses or
overtime in an average year)? Please only include monetary compensation here, not the value of
benefits.`                                        <chr> …
$ `Please indicate the currency`
<chr> …
$ `If "Other," please indicate the currency here:`
<chr> …
$ `If your income needs additional context, please provide it here:`
<chr> …
$ `What country do you work in?`
```

```
<chr> …
$ `If you're in the U.S., what state do you work in?`
<chr> …
$ `What city do you work in?`
<chr> …
$ `How many years of professional work experience do you have overall?`
<chr> …
$ `How many years of professional work experience do you have in your field?`
<chr> …
$ `What is your highest level of education completed?`
<chr> …
$ `What is your gender?`
<chr> …
$ `What is your race? (Choose all that apply.)`
<chr> …
```

```
# Display the column names of the data frame
colnames(workforce)
```

```
 [1] "Timestamp"
 [2] "How old are you?"
 [3] "What industry do you work in?"
 [4] "Job title"
 [5] "If your job title needs additional context, please clarify here:"
 [6] "What is your annual salary? (You'll indicate the currency in a later question. If you
are part-time or hourly, please enter an annualized equivalent -- what you would earn if you
worked the job 40 hours a week, 52 weeks a year.)"
 [7] "How much additional monetary compensation do you get, if any (for example, bonuses or
overtime in an average year)? Please only include monetary compensation here, not the value of
benefits."
 [8] "Please indicate the currency"
 [9] "If \"Other,\" please indicate the currency here:"
[10] "If your income needs additional context, please provide it here:"
[11] "What country do you work in?"
[12] "If you're in the U.S., what state do you work in?"
[13] "What city do you work in?"
[14] "How many years of professional work experience do you have overall?"
[15] "How many years of professional work experience do you have in your field?"
[16] "What is your highest level of education completed?"
[17] "What is your gender?"
[18] "What is your race? (Choose all that apply.)"
```

Cleaning Column Names

As the column names are a bit long, to make the data manipulation easier, we need to modify them

```
 # Rename columns for easier manipulation
colnames(workforce) <- c("submission_timestamp", "age", "industry", "job_title", "job_title_co
```

Factoring various Fields such as “age”,“total_years_experience”,“years_experience_in_field”

Convert the age field to a factor for better analysis

```
# Display unique values in the age column
unique(workforce$age)
```

```
[1] "25-34"      "45-54"      "35-44"      "18-24"      "65 or over"
[6] "55-64"      "under 18"
```

```
# Convert age to a factor with specified levels
workforce$age <- as.factor(workforce$age)
levels(workforce$age)
```

```
[1] "18-24"      "25-34"      "35-44"      "45-54"      "55-64"
[6] "65 or over" "under 18"
```

```
workforce$age <- factor(workforce$age, levels = c("under 18", "18-24", "25-34", "35-44", "45-5

# Convert total_years_experience to a factor
unique(sort(workforce$total_years_experience))
```

```
[1] "1 year or less"   "11 - 20 years"    "2 - 4 years"      "21 - 30 years"
[5] "31 - 40 years"    "41 years or more" "5-7 years"        "8 - 10 years"
```

```
workforce$total_years_experience <- as.factor(workforce$total_years_experience)
levels(workforce$total_years_experience)
```

```
[1] "1 year or less"   "11 - 20 years"    "2 - 4 years"      "21 - 30 years"
[5] "31 - 40 years"    "41 years or more" "5-7 years"        "8 - 10 years"
```

```
workforce$total_years_experience <- factor(workforce$total_years_experience, levels = c("1 yea
levels(workforce$total_years_experience)
```

```
[1] "1 year or less"   "2 - 4 years"      "5-7 years"        "8 - 10 years"
[5] "11 - 20 years"    "21 - 30 years"    "31 - 40 years"    "41 years or more"
```

```
# Convert years_experience_in_field to a factor
unique(sort(workforce$years_experience_in_field))
```

```
[1] "1 year or less"   "11 - 20 years"    "2 - 4 years"      "21 - 30 years"
[5] "31 - 40 years"    "41 years or more" "5-7 years"        "8 - 10 years"
```

```
workforce$years_experience_in_field <- as.factor(workforce$years_experience_in_field)
levels(workforce$years_experience_in_field)
```

```
[1] "1 year or less"   "11 - 20 years"    "2 - 4 years"      "21 - 30 years"
[5] "31 - 40 years"    "41 years or more" "5-7 years"        "8 - 10 years"
```

```
workforce$years_experience_in_field <- factor(workforce$years_experience_in_field, levels = c(
levels(workforce$years_experience_in_field)
```

```
[1] "1 year or less"   "2 - 4 years"      "5-7 years"        "8 - 10 years"
[5] "11 - 20 years"    "21 - 30 years"    "31 - 40 years"    "41 years or more"
```

Handling Missing Values and removing the respective observations.

```
# Check for missing values in country , industry and currency
missing_country <- sum(is.na(workforce$country))

missing_industry <- sum(is.na(workforce$industry))

# Print the results
cat("Missing values in 'country':", missing_country, "\n")
```

```
Missing values in 'country': 0
```

```
cat("Missing values in 'industry':", missing_industry, "\n")
```

```
Missing values in 'industry': 74
```

```
workforce <- workforce %>%
  filter(!is.na(country) & !is.na(industry) & currency!='Other')
```

Currency Conversion Convert salaries to USD using predefined conversion rates

```
# Define conversion rates for different currencies
conversion_rates <- c(
  AUD = 0.64,  # 1 AUD = 0.64 USD
  CAD = 0.73,  # 1 CAD = 0.73 USD
  CHF = 1.1,   # 1 CHF = 1.1 USD
  EUR = 1.05,  # 1 EUR = 1.05 USD
  GBP = 1.21,  # 1 GBP = 1.21 USD
  HKD = 0.13,  # 1 HKD = 0.13 USD
  JPY = 0.0067,# 1 JPY = 0.0067 USD
  SEK = 0.091,  # 1 SEK = 0.091 USD
  USD = 1
)

# Convert annual_salary and additional_compensation to numeric
workforce$annual_salary <- as.numeric(as.character(workforce$annual_salary))
workforce$additional_compensation <- as.numeric(as.character(workforce$additional_compensation

# Replace NA values with 0
workforce$annual_salary <- ifelse(is.na(workforce$annual_salary), 0, workforce$annual_salary)
workforce$additional_compensation <- ifelse(is.na(workforce$additional_compensation), 0, workf
```

```
# Calculate salary in USD
workforce$salary_in_usd <- (workforce$annual_salary + workforce$additional_compensation) * con


missing_salary <- sum(is.na(workforce$salary_in_usd))

cat("Missing values in 'salary_in_usd':", missing_salary, "\n")
```

```
Missing values in 'salary_in_usd': 517
```

```
workforce <- workforce %>%
  filter(!is.na(salary_in_usd))
```

Displaying Converted Salaries Show the top 10 salaries in USD for non-USD currencies.

```
# Display top 10 salaries in USD for non-USD currencies
workforce %>%
  filter(currency != "USD") %>%
  select(annual_salary, currency, salary_in_usd) %>%
  head(10)
```

```
# A tibble: 10 × 3
   annual_salary currency salary_in_usd
           <dbl> <chr>            <dbl>
 1         54600 GBP              70906
 2         32000 CAD              23360
 3         24000 GBP              29645
 4         63000 CAD              45990
 5         35000 GBP              49610
 6        120000 CAD              88695
 7         97500 CAD              78475
 8         52000 CAD              37960
 9         52000 GBP              62920
10         79000 CAD              57670
```

```
# Count the number of entries per industry
workforce %>%
  count(industry)
```

```
# A tibble: 1,112 × 2
   industry                                  n
   <chr>                                 <int>
 1 "\"Government Relations\" (Lobbying)"     1
 2 "Academia"                                5
 3 "Academia - STEM"                         1
 4 "Academia / Research"                     1
 5 "Academia--cell and molecular biology"    1
 6 "Academic Medicine"                       1
 7 "Academic Press Production"               1
 8 "Academic Publishing"                     2
 9 "Academic Science"                        1
```

```
10 "Academic Scientific Research"            1
# ℹ 1,102 more rows
```

Cleaning the Industry Column

Standardize the industry names for better analysis.

```
# Convert industry names to lowercase
workforce <- workforce %>%
  mutate(industry = tolower(industry))

# Recode industry names to standardize them
workforce <- mutate(workforce,
  industry = recode(.x = industry
,"\"government relations\" (lobbying)"="lobbying"
,"academia - stem" ="academia"
,"academia--cell and molecular biology"="academia"
,"academia / research"      ="academia"
, "academic medicine" ="academic publishing/research/science"
,"academic press production"="academic publishing/research/science"
,"academic publishing"      ="academic publishing/research/science"
,"academic research"        ="academic publishing/research/science"
,"academic science"         ="academic publishing/research/science"
,"academic scientific research"
,"academic/nonprofit research"
, "accessibility"            = "accessibility"
, "accounting, banking & finance"
, "actuarial"                = "actuarial"
, "administration"           ="administration"
, "administration (food service)"
, "administration in mlm"    ="administration"
, "administration, it"       ="administration"
, "administrative"           ="administration"
, "administrative support"   ="administration"
, "administrative work"      ="administration"
, "adult education"          = "education"
, "aerospace"                = "aerospace/defence"
, "aerospace & defense"      = "aerospace/defence"
, "aerospace and defense"    = "aerospace/defence"
, "aerospace and defense manufacturing"
, "aerospace and defense/government contracting"
, "aerospace contracting"    = "aerospace/defence"
, "aerospace data"           = "aerospace/defence"
, "aerospace manufacturing"  = "aerospace/defence"
, "aerospace/aviation"       = "aerospace/defence"
, "aerospace/defense"        = "aerospace/defence"
, "agriculture or forestry"  ="agriculture or forestry "
, "agriculture/agriculture chemical"
, "airline"                  = "airline"
, "americorps"               = "americorps"
, "analytical chemistry"     = "analytical chemistry"
, "analytical lab"           = "analytics"
, "analytics"                = "analytics"
, "animal care"              = "animal health industry"
```

```
, "animal caretaker"         = "animal health industry"
, "animal health"            = "animal health industry"
, "animal health industry"   = "animal health industry"
, "animal health product manufacturing"
, "animal welfare"           = "animal health industry"
, "apparel"                  = "apparel design/product development"
, "apparel design/product development"
, "apparel manufacture"      = "apparel design/product development"
, "archaeologist"            = "archaeology / cultural resource management"
, "archaeology"              = "archaeology / cultural resource management"
, "archaeology / cultural resource management"
, "archaeology/cultural resource manager"
, "architect"                = "architectural/land planning/civil engineering"
, "architectural/land planning/civil engineering"
, "architecture"             = "architectural/land planning/civil engineering"
, "architecture & construction"
, "architecture / engineering"
, "architecture and engineering consulting and design"
, "architecture, engineering, construction"
, "architecture/construction"= "architectural/land planning/civil engineering"
, "archives"                 = "archives/libraries"
, "archives/libraries"       = "archives/libraries"
, "archives/library science" = "archives/libraries"
, "art & design"             = "arts / culture / heritage"
, "art appraisal"            = "arts / culture / heritage"
, "arts administration"      = "arts / culture / heritage"
, "arts, culture and heritage"
, "association"              = "association management"
, "association management"   = "association management"
, "auto mfg."                = "automotive mfg / finance / insurance"
, "auto repair"              = "automotive mfg / finance / insurance"
, "automotive"               = "automotive mfg / finance / insurance"
, "automotive finance and insurance"
, "automotive repair"        = "automotive mfg / finance / insurance"
, "automotive technician"    = "automotive mfg / finance / insurance"
, "beauty"                   = "beauty/service industry"
, "beauty /cpg"              = "beauty/service industry"
, "beauty manufacturing & education"
, "beauty, cosmetics, fragrance"
, "beauty/service industry"  = "beauty/service industry"
, "behavior analysis/mental health"
, "behavioral health"        = "behavior analysis/mental health"
, "beverage"                 = "beverage"
, "beverage & spirits"       = "beverage"
, "beverage distribution"    = "beverage"
, "beverage production"      = "beverage"
, "bio tech"                 = "biotech industry"
, "biological research"      = "biological research/science"
, "biological sciences"      = "biological research/science"
, "biologist"                = "biological research/science"
, "biology"                  = "biological research/science"
, "biology/research"         = "biological research/science"
, "biomedical research"      = "biological research/science"
, "biopharma"                = "biopharmaceuticals"
```

```
    , "biopharmaceuticals"       = "biopharmaceuticals"
    , "biotech"                  = "biotech industry"
    , "biotech (r&d)"            = "biotech industry"
    , "biotech / life sciences"  = "biotech industry"
    , "biotech / pharmaceutical industry"
    , "biotech / research"       = "biotech industry"
    , "biotech industry"         = "biotech industry"
    , "biotech manufacturing"    = "biotech industry"
    , "biotech pharmaceuticals"  = "biotech industry"
    , "biotech r&d"              = "biotech industry"
    , "biotech research"         = "biotech industry"
    , "biotech/drug development" = "biotech industry"
    , "biotech/food safety"      = "biotech industry"
    , "biotech/pharma"           = "biotech industry"
    , "biotech/pharmaceuticals"  = "biotech industry"
    , "biotech/software"         = "biotech industry"
    , "biotechnology"            = "biotech industry"
    , "biotechnology, research and development"
    , "biotechnology/life sciences"
    , "bitech"                   = "biotech industry"
    , "business or consulting"   = "business or consulting"
    , "business process outsourcing"
    , "business services"        = "business or consulting"
    , "cannabis"                 = "cannabis compliance"
    , "cannabis compliance"      = "cannabis compliance"
    , "chemical"                 = "chemicals"
    , "chemical manufacturing"   = "chemicals"
    , "chemicals"                = "chemicals"
    , "chemicals/ materials"     = "chemicals"
    , "chemistry"                = "chemicals"
    , "child and yout care"      = "childcare"
    , "child care"               = "childcare"
    , "child care resource and referral agency"
    , "childcare"                = "childcare"
    , "childcare (0-5 so does not come under primary education)"
    , "church"                   = "church ministry"
    , "church ministry"          = "church ministry"
    , "clean energy (eg. energy efficiency, renewables, etc.)"
    , "clergy"                   = "clean energy (eg. energy efficiency, renewables, etc.)"
    , "clinical & translational reserach"
    , "clinical research"        = "clinical research and development"
    , "clinical research and development"
    , "clinical research manager - academic institution"
    , "clinical trials"          = "clinical research and development"
    , "clinical trials research coordination"
    , "commercial real estate"   = "commercial real estate"
    , "commercial real estate - private equity"
    , "commercial real estate data and analytics/research"
    , "commercial real estate tenancy"
    , "communications"           = "communications/publications"
    , "communications/publications"
    , "computing or tech"        = "computing or tech"
    , "computing/tech + higher ed + nonprofit"
    , "concrete"                 = "construction, mining, manufacturing"
```

```
, "concrete construction"    = "construction, mining, manufacturing"
, "construction"             = "construction, mining, manufacturing"
, "construction / stone industry"
, "construction management"  = "construction, mining, manufacturing"
, "construction, hvac"       = "construction, mining, manufacturing"
, "construction, mining, manufacturing"
, "consultant"               = "consulting"
, "consulting"               = "consulting"
, "consulting / professional services"
, "consulting operations- big 4"
, "consumer good (toys)"     = "consumer goods"
, "consumer goods"           = "consumer goods"
, "consumer goods production"= "consumer goods"
, "consumer packaged goods"  = "consumer goods"
, "consumer product design"  = "consumer product"
, "consumer product goods"   = "consumer goods"
, "consumer product organization"
, "consumer products"        = "consumer product"
, "consumer products design" = "consumer product"
, "consumer/packaged goods"  = "consumer goods"
, "contract research"        = "contract research"
, "contract research organisation"
, "counseling"               = "counseling"
, "counselling"              = "counseling"
, "cultural (museums/galleries)"
, "cultural heritage"        = "cultural resource management"
, "cultural resource management"
, "cultural resources management/major univ."
, "culture"                  = "cultural resource management"
, "customer service"         = "customer service"
, "customer service/call center"
, "customer service/publishing-adjacent"
, "defense"                  = "defense"
, "defense contracting"      = "defense"
, "defense contractor"       = "defense"
, "digital"                  = "digital marketing"
, "digital commerce / ecommerce"
, "digital marketing"        = "digital marketing"
, "digital marketing within a book publishing company (please reclassify as you see fit)"
, "e-comm"                   = "e-commerce"
, "e-commerce"               = "e-commerce"
, "e commerce"               = "e-commerce"
, "early childhood education"= "early education"
, "early childhood education (preschool)"
, "early education"          = "early education"
, "early education (corporate office)"
, "education (early childhood education)"
, "education (early childhood)"
, "education (higher education)"
, "education (other)"        = "education (primary/secondary/higher/consulting /publishing/res
, "education (primary/secondary)"
, "education consulting"     = "education (primary/secondary/higher/consulting /publishing/res
, "education publishing"     = "education (primary/secondary/higher/consulting /publishing/res
, "education research- mix of edtech and non profits"
```

```
, "education service provider"
, "education services (tutoring)"
, "education start-up"       = "education (primary/secondary/higher/consulting /publishing/res
, "education writing"        = "education (primary/secondary/higher/consulting /publishing/res
, "education/vocational"     = "education (primary/secondary/higher/consulting /publishing/res
, "education: preschool"     = "education (primary/secondary/higher/consulting /publishing/res
, "educational assessment"   = "education (primary/secondary/higher/consulting /publishing/res
, "educational products"     = "education (primary/secondary/higher/consulting /publishing/res
, "educational publishing"   = "education (primary/secondary/higher/consulting /publishing/res
, "educational publishing / ed tech"
, "educational research"     = "education (primary/secondary/higher/consulting /publishing/res
, "educational technology"   = "education (primary/secondary/higher/consulting /publishing/res
, "educational technology - hybrid between book publishing and technology really"
, "energy"                   = "energy - oil and gas"
, "energy - oil and gas"     = "energy - oil and gas"
, "energy (oil & gas & associated products, renewable power, etc)"
, "energy (oil & gas)"       = "energy - oil and gas"
, "energy / renewables"      = "energy - oil and gas"
, "energy sector: oil & gas" = "energy - oil and gas"
, "energy supplier"          = "energy - oil and gas"
, "energy, oil & gas"        = "energy - oil and gas"
, "energy, oil and gas"      = "energy - oil and gas"
, "energy/oil"               = "energy - oil and gas"
, "energy: oil & gas"        = "energy - oil and gas"
, "engineering - mining"     = "engineering or manufacturing"
, "engineering and environmental consulting"
, "engineering or manufacturing"
, "enviromental"             = "enviromental "
, "environment"              = "enviromental "
, "environment - oil and gas"= "enviromental "
, "environment and sustainability"
, "environment, health, and safety"
, "environmental"            = "enviromental "
, "environmental compliance" = "enviromental "
, "environmental compliance/engineering"
, "environmental consultanting"
, "environmental consulting" = "enviromental "
, "environmental health + pest control"
, "environmental health and safety"
, "environmental health and safety compliance"
, "environmental planning"   = "enviromental "
, "environmental regulation" = "enviromental "
, "environmental restoration"= "enviromental "
, "environmental science"    = "enviromental "
, "environmental sciences"   = "enviromental "
, "environmental services"   = "enviromental "
, "environmental survey"     = "enviromental "
, "environmental/cultural resource management"
, "environmnetal"            = "enviromental "
, "esl teacher" ="education (primary/secondary/higher/consulting /publishing/research etc)"
, "finance"                  = "finance/fintech"
, "finance/investment management but in legal/compliance, so back-office"
, "fintech"                  = "finance/fintech"
, "fintech/payment processing"
```

```
    , "fitness"                 = "fitness & entertainment"
    , "fitness & entertainment"  = "fitness & entertainment"
    , "food"                    = "food & beverage"
    , "food & beverage"         = "food & beverage"
    , "food & beverage production"
    , "food & beverages"        = "food & beverage"
    , "food & nutrition"        = "food & beverage"
    , "food and beverage"       = "food & beverage"
    , "food and drink"          = "food & beverage"
    , "food and flavor"         = "food & beverage"
    , "food demos"              = "food & beverage"
    , "food distribution"       = "food & beverage"
    , "food industry"           = "food & beverage"
    , "food manufacture"        = "food & beverage"
    , "food manufacturers"      = "food & beverage"
    , "food manufacturing"      = "food & beverage"
    , "food processing"         = "food & beverage"
    , "food processing and packaging"
    , "food production"         = "food & beverage"
    , "food production/processing"
    , "food service"            = "food & beverage"
    , "food service --- baking"  = "food & beverage"
    , "food/beverage manufacturing- quality/laboratory"
    , "food/quick service restaurant (qsr)"
    , "foodservice"             = "food & beverage"
    , "funding intermediary"    = "fundraising"
    , "fundraising"             = "fundraising"
    , "fundraising for a university"
    , "fundraising in higher education; nonclinical, nonacademic"
    , "funeral"                 = "funeral services"
    , "funeral service"         = "funeral services"
    , "funeral services"        = "funeral services"
    , "game development"        = "gaming"
    , "games development"       = "gaming"
    , "gaming"                  = "gaming"
    , "gaming (gambling)"       = "gaming"
    , "government affairs/lobbying"
    , "government contract"     = "government contract"
    , "government contracting"   = "government contract"
    , "government contracting (data analytics and program evaluations)"
    , "government contracting r&d"
    , "government contractor"    = "government contract"
    , "government contractor (r&d)"
    , "government contractor, international development"
    , "government relation"     = "government relation"
    , "government relations"    = "government relation"
    , "government relations/lobbying"
    , "govt contractor - not direct govt but they pay my company who in turn pays me"
    , "govt contractor - not directly govt but they pay me"
    , "grocery delivery"        ="grocery"
    , "grocery distribution"    ="grocery"
    , "health and safety"       = "health care/safety/research"
    , "health care"             = "health care/safety/research"
    , "health insurance"        = "health care/safety/research"
```

```
    , "health research"          = "health care/safety/research"
    , "healthcare information technology"
    , "healthcare it"            = "health care/safety/research"
    , "healthcare technology"    = "health care/safety/research"
    , "higher education fundraising"
    , "hybrid nonprofit higher education (we are part of a university but our entire budget comes
    , "i'm currently a student and don't have a job"
    , "i have two jobs. marketing / business"
    , "i work at a property tax management company. not sure where this fits in. consulting maybe?
    , "i work for indeed.com"    ="consulting"
    , "i work in the finance function of a large global conglomerate"
    , "industrial cleaning & non hazardous transport"
    , "industrial hygiene"       = "industrial hygiene"
    , "information services"     = "information services"
    , "information services (libraries)"
    , "information services (library)"
    , "information services/libraries"
    , "information technology"   = "information technology"
    , "information technology (it)"
    , "interior design & architecture"
    , "interior design (commercial)"
    , "international development"= "international development"
    , "international development (multilateral donor)"
    , "international organisations"
    , "international organization (un)"
    , "it"                       = "information technology"
    , "it msp"                   = "information technology"
    , "it security"              = "information technology"
    , "janitorial"               = "industrial hygiene"
    , "labor"                    = "labour/professional organization"
    , "labor union"              = "labour/professional organization"
    , "labour/professional organization"
    , "landscape architecture"   = "landscaping"
    , "landscape contracting"    = "landscaping"
    , "landscaping"              = "landscaping"
    , "language services"        = "language services"
    , "language services company, unsure the broad category to use. our clients are branding agenc
    , "law"                      = "law"
    , "law enforcement & security"
    , "law library"              = "law"
    , "learning & development"   = "learning and development"
    , "learning and development" = "learning and development"
    , "librarian"                = "archives/libraries"
    , "librarian--contractor for nasa"
    , "librarian and assistant manager of a library"
    , "librarian in legal setting"
    , "libraries"                = "archives/libraries"
    , "libraries & archives"     = "archives/libraries"
    , "libraries (medical)"      = "archives/libraries"
    , "libraries (public)"       = "archives/libraries"
    , "libraries / archives / information"
    , "libraries and archives"   = "archives/libraries"
    , "libraries and archives (academic)"
    , "libraries/archives"       = "archives/libraries"
```

```
    , "libraries/museums/archives"
    , "library"                 = "archives/libraries"
    , "library--public"         = "archives/libraries"
    , "library (its a non-profit and its a govt job - how would i list that? not all libraries are
    , "library (university)"    = "archives/libraries"
    , "library and information science"
    , "library and information services"
    , "library at a university"  = "archives/libraries"
    , "library page (public county library)"
    , "library science / part-time work/study"
    , "library tech for a school system"
    , "library/archive"         = "archives/libraries"
    , "library/archive/research center"
    , "library/archives"        = "archives/libraries"
    , "library/information managment"
    , "life science capability development"
    , "life sciences"           = "life sciences"
    , "life sciences (not in academia)"
    , "lobbying"                ="lobbying"
    , "lobbying and consulting"  ="lobbying"
    , "manufacturing"           = "manufacturing (medical devices)"
    , "manufacturing (medical devices)"
    , "manufacturing (personal care)"
    , "manufacturing (pharmaceuticals)"
    , "manufacturing : corporate admin support"
    , "manufacturing and distributing"
    , "manufacturing security systems"
    , "manufacturing, chemical"  = "manufacturing (medical devices)"
    , "manufacturing/consumer packaged goods"
    , "manufacturing/wholesale"  = "manufacturing (medical devices)"
    , "market research"         = "marketing technology"
    , "marketing at a non profit"= "marketing technology"
    , "marketing technology"    = "marketing technology"
    , "marketing, advertising & pr"
    , "medica education"        = "medical technology/education"
    , "medical communications"  = "medical technology/education"
    , "medical device"          = "medical technology/education"
    , "medical devices"         = "medical technology/education"
    , "medical interpreter -(spanish)"
    , "medical library"         = "medical technology/education"
    , "medical research"        = "medical technology/education"
    , "medical sciences"        = "medical technology/education"
    , "medical supply wholesale & warehousing"
    , "medical technology"      = "medical technology/education"
    , "medical/pharmaceutical"  = "medical technology/education"
    , "mental health"           = "mental health therapist"
    , "mental health therapist"  = "mental health therapist"
    , "mining"                  = "mining and natural resources"
    , "mining & mineral processing"
    , "mining and natural resources"
    , "mining/mineral exploration"
    , "mining/resource extraction"
    , "municipal (public) libraries"
    , "municipal government (library)"
```

```
  , "municipal library"        = "archives/libraries"
  , "museum"                   = "museums & archives"
  , "museum - nonprofit"       = "museums & archives"
  , "museum (<20 employees)"   = "museums & archives"
  , "museum (university affiliated)"
  , "museum education"         = "museums & archives"
  , "museum library"           = "archives/libraries"
  , "museums"                  = "museums & archives"
  , "museums & archives"       = "museums & archives"
  , "museums & archives (not sure where this would fall)"
  , "museums: nonprofit"       = "museums & archives"
  , "music"                    = "music"
  , "music licensing"          = "music"
  , "music therapy"            = "music"
  , "music, education"         = "music"
  , "music: freelance, performing and education"
  , "non-profit health care (i couldn’t select both)"
  , "non-profit theatre"       = "nonprofit organisations"
  , "non profit theater"       = "nonprofit organisations"
  , "nonprofit - legal department"
  , "nonprofit - lort d theater"
  , "nonprofit association"    = "nonprofit organisations"
  , "nonprofit scholarly society publisher"
  , "nonprofits"               = "nonprofit organisations"
  , "not-for-profit health research consulting"
  , "not-for-profit membership organization"
  , "not for profit education consultancy"
  , "oil"                      = "gas & oil"
  , "oil & gas"                = "gas & oil"
  , "oil & gas - non destructive testing"
  , "oil and gas"              = "gas & oil"
  , "oil and gas exploration"  = "gas & oil"
  , "oil and gas safety training"
  , "pet"                      = "pet care"
  , "pet care"                 = "pet care"
  , "pet care industry"        = "pet care"
  , "pet care industry (dog training/walking)"
  , "pet care/grooming"        = "pet care"
  , "pharma"                   = "pharma"
  , "pharma & biotech"         = "pharma"
  , "pharma / medical device design and manufacturing"
  , "pharma r&d"               = "pharma"
  , "pharma research"          = "pharma"
  , "pharma/ research"         = "pharma"
  , "pharma/biotech"           = "pharma"
  , "pharma/biotechnology"     = "pharma"
  , "pharmaceitical manufacturing"
  , "pharmaceutical"           = "pharma"
  , "pharmaceutical company"   = "pharma"
  , "pharmaceutical development"
  , "pharmaceutical industry"  = "pharma"
  , "pharmaceutical manufacturing"
  , "pharmaceutical r&d"       = "pharma"
  , "pharmaceutical research"  = "pharma"
```

```
, "pharmaceutical research & development"
, "pharmaceutical research (chemist)"
, "pharmaceutical/biotech"   = "pharma"
, "pharmaceutical/biotechnology"
, "pharmaceutical/contract research organization"
, "pharmaceuticals"          = "pharma"
, "pharmaceuticals / biotech"= "pharma"
, "pharmaceuticals r&d"      = "pharma"
, "pharmaceuticals/biotechnology"
, "pharmacuticals"           = "pharma"
, "political campaign"       = "politics"
, "political campaigning"    = "politics"
, "political campaigns"      = "politics"
, "political consulting"     = "politics"
, "political research"       = "politics"
, "politics"                 = "politics"
, "politics/campaigns"       = "politics"
, "politics/government relations"
, "private company, federal contractor"
, "professional association" = "professional association"
, "professional association in finance"
, "professional public librarian"
, "professional regulation"  = "professional services"
, "professional services"    = "professional services"
, "professional services / architecture"
, "professional training"    = "professional services"
, "property management"      = "property management"
, "property or construction" = "property management"
, "public health"            = "public health"
, "public health- state level"
, "public health (not medical)"
, "public health in higher education"
, "public health research"   = "public health"
, "public health, local government"
, "public librarian"         = "archives/libraries"
, "public libraries"         = "archives/libraries"
, "public library"           = "archives/libraries"
, "public library (might be considered government, but that always seems an odd designation...
, "public library (non-profit, but also government?)"
, "public library (technically city govt.?)"
, "publications"             = "publishing"
, "publishibg"               = "publishing"
, "publishing"               = "publishing"
, "publishing (academic)"    = "publishing"
, "publishing (book)"        = "publishing"
, "publishing, content as a service"
, "publishing/edtech"        = "publishing"
, "publishing: science, academic, technical"
, "r&d"                      = "r&d"
, "r&d in manufacturing"     = "r&d"
, "real estate"              = "real estate / housing"
, "real estate / housing"    = "real estate / housing"
, "real estate affordable housing"
, "real estate association"  = "real estate / housing"
```

```
  , "real estate corp. office/not a realtor"
  , "real estate customer care"= "real estate / housing"
  , "real estate development"  = "real estate / housing"
  , "real estate investment"   = "real estate / housing"
  , "real estate investment support"
  , "real estate servicea"     = "real estate / housing"
  , "real estate services"     = "real estate / housing"
  , "real estate software"     = "real estate / housing"
  , "real estate title company"= "real estate / housing"
  , "real estate valuation"    = "real estate / housing"
  , "real estate/ mortgage"    = "real estate / housing"
  , "real estate/development"  = "real estate / housing"
  , "real estate: title & escrow"
  , "regulatory affairs- nutraceuticals"
  , "religion"                 = "religious"
  , "religion/church"          = "religious"
  , "religious"                = "religious"
  , "religious (church)"       = "religious"
  , "religious (synagogue)"    = "religious"
  , "religious educator"       = "religious"
  , "religious institute"      = "religious"
  , "religious institution"    = "religious"
  , "research"                 = "research"
  , "research - academic"      = "research"
  , "research - public health" = "research"
  , "research & development"   = "research"
  , "research & development (defense industry)"
  , "research & development (physical sciences)"
  , "research (health)"        = "research"
  , "research / gov"           = "research"
  , "research administration"  = "research"
  , "research and development" = "research"
  , "research and development academia"
  , "research and development, food and beverage"
  , "research and evaluation"  = "research"
  , "research at a national laboratory"
  , "research at a state university"
  , "research institute"       = "research"
  , "research science"         = "research"
  , "research scientist, pharma"
  , "research/academia"        = "research"
  , "research/social science"  = "research"
  , "restaurant"               = "restaurant"
  , "restaurant group"         = "restaurant"
  , "restaurant/food service"  = "restaurant"
  , "restaurant/service"       = "restaurant"
  , "restaurants & hospitality"= "restaurant"
  , "retail"                   = "retail"
  , "retail call center"       = "retail"
  , "retail mid level management"
  , "retail pharmacy"          = "retail"
  , "retail real estate"       = "real estate / housing"
  , "retired"                  ="No job"
  , "saas"                     = "software"
```

```
, "saas company/software"    = "software"
, "sales operations"         = "sales"
, "science"                  = "science/research"
, "science - qc lab"         = "science/research"
, "science (chemistry r&d)"  = "science/research"
, "science (laboratory)"     = "science/research"
, "science (research, biology)"
, "science academia"         = "science/research"
, "science and natural resource management"
, "science and reasearch"    = "science/research"
, "science publishing"       = "science/research"
, "science research"         = "science/research"
, "science research, government"
, "science/biotech"          = "science/research"
, "science/government"       = "science/research"
, "science/research"         = "science/research"
, "science/research (academia)"
, "science/research non-academic"
, "sciences"                 = "science/research"
, "scientific"               = "science/research"
, "scientific analysis"      = "science/research"
, "scientific publishing"    = "science/research"
, "scientific r&d"           = "science/research"
, "scientific research"      = "science/research"
, "scientific research (industry)"
, "scientist"                = "science/research"
, "security"                 = "security"
, "security and manufacturing company"
, "service"                  = "service and repair"
, "social research"          = "science/research"
, "social science"           = "science/research"
, "social science research - not quite academia, not quite nonprofit, not quite consulting"
, "social sciences research" = "science/research"
, "software"                 = "software"
, "software as a service saas"
, "software development"     = "software"
, "software development / it"= "software"
, "software products"        = "software"
, "software/programming"     = "software"
, "soldier"                  = "military"
, "staffing & workforce solutions"
, "staffing agency"          = "staffing agency"
, "staffing firm"            = "staffing agency"
, "staffing industrry"       = "staffing agency"
, "supply chain"             = "supply chain"
, "supply chain distribution"= "supply chain"
, "supply chain operations"  = "supply chain"
, "supply chain!"            = "supply chain"
, "survey methodology"       = "surveying"
, "survey research/public policy research"
, "surveying"= "surveying"
, "synthetic chemical manufacturing"= "surveying"
, "tabletop gaming"          = "gaming"
, "tech"                     = "technology"
```

```
, "technical writing"        = "technology"
, "technical/cybersecurity"  = "technology"
, "technical/it"             = "technology"
, "technology"               = "technology"
, "technology/saas"          = "technology"
, "telecommunications (gps)" = "telecommunications"
, "tourism/heritage -- but for a government building"
, "trade association"        = "trade association"
, "trade association/membership"
, "trade associations"       = "trade association"
, "trades (supply chain) oil and gas"
, "training"                 = "training"
, "training and professional services"
, "translation"              = "translation/transcription"
, "translation and localization"
, "translation/transcription"= "translation/transcription"
, "university libraries"     = "archives/libraries"
, "university research"      = "science/research"
, "user experience (ux) research"
, "vet"                      = "veterinary"
, "veterinarian"             = "veterinary"
, "veterinary"               = "veterinary"
, "veterinary biotech"       = "veterinary"
, "veterinary care"          = "veterinary"
, "veterinary diagnostics"   = "veterinary"
, "veterinary m&a"           = "veterinary"
, "veterinary medicine"      = "veterinary"
, "veterinary services"      = "veterinary"
, "video game industry"      = "video games"
, "warehouse- food and beverage"
, "warehousing"              = "warehouse"
, "waste and recycling"      = "waste management"
, "wholesale"                = "wholesale"
, "wholesale - apparel"      = "wholesale"
, "wholesale and retail trade"
, "wholesale distribution"   = "wholesale"
, "wholesale distribution b2b"
, "wholesale industrial & welding supplies & equipment"
, "wholesale supplier"       = "wholesale"
, "wholesale textile manufacture and sales"
, "wholesale trade"          = "wholesale"
, "wholesale/distrbution"    = "wholesale"
, "wine"                     = "wine & spirits"
, "zoo"                      = "zoo"
)
)


workforce <- mutate(workforce,
  industry = recode(.x = industry,
  "diversity, equity & inclusion" = "diversity, equity & inclusion",
  "dod contracting" = "government contracting",
  "drug development" = "pharma",
  "e-commerce" = "ecommerce",
```

```
    "e-learning" = "education",
    "eap" = "human resources",
    "early education" = "education",
    "earth sciences" = "environmental science",
    "ecology" = "environmental science",
    "ecommerce" = "ecommerce",
    "ecommerce - technology" = "ecommerce",
    "ecommerce fraud" = "ecommerce",
    "economics" = "economics",
    "ed tech" = "education technology",
    "editor in educational publishing" = "publishing",
    "edtech" = "education technology",
    "educ tech" = "education technology",
    "education" = "education",
    "education (primary/secondary/higher/consulting /publi…" = "education",
    "education- museum/public outreach" = "education",
    "emergency management" = "emergency management",
    "energy - oil and gas" = "energy - oil and gas",
    "engineering or manufacturing" = "engineering",
    "entertainment" = "entertainment",
    "entertainment data" = "entertainment",
    "entrepreneur high net worth" = "entrepreneurship",
    "env. consulting" = "environmental consulting",
    "enviromental" = "environmental science",
    "executive leadership servis" = "executive leadership",
    "executive search" = "recruitment",
    "facilities" = "facilities management",
    "faith/spirituality" = "religious",
    "family office" = "finance",
    "fashion" = "fashion",
    "fashion/e-commerce" = "fashion",
    "fast casual restaurant" = "restaurant",
    "fast food" = "restaurant",
    "federal contracting/business development" = "government contracting",
    "federal government contracting" = "government contracting",
    "film post-production" = "film industry",
    "finance/fintech" = "finance",
    "fire protection" = "safety",
    "fitness & entertainment" = "fitness",
    "fmcg" = "fast-moving consumer goods",
    "fmcg development" = "fast-moving consumer goods",
    "food & beverage" = "food and beverage",
    "for profit education" = "education",
    "forensics" = "forensics",
    "freelance journalism" = "journalism",
    "freelance/self-employed consultant" = "consulting",
    "fundraising" = "nonprofit",
    "funeral services" = "funeral services",
    "gambling" = "gaming",
    "gaming" = "gaming",
    "gas & oil" = "energy - oil and gas",
    "geologist" = "geology",
    "geospatial" = "geospatial",
    "global health consulting" = "health consulting",
```

```
    "global mobility" = "human resources",
    "government" = "government",
    "government and public administration" = "government",
    "government contract" = "government contracting",
    "government relation" = "government",
    "government research" = "government",
    "government- scientist" = "government",
    "govtech software as a service" = "government technology",
    "graduate assistant and also events" = "graduate assistant",
    "graduate student" = "student",
    "grantwriting consultants" = "consulting",
    "grocery" = "food and beverage",
    "gyms" = "fitness",
    "hardware manufacturing" = "manufacturing",
    "haz/ind/rad waste management" = "waste management",
    "health and fitness" = "healthcare",
    "health care/safety/research" = "healthcare",
    "heritage" = "heritage",
    "heritage/public history" = "heritage",
    "higher education/libraries" = "higher education",
    "historic preservation" = "heritage",
    "hospital" = "healthcare",
    "hospitality & events" = "hospitality",
    "household services" = "household services",
    "housekeeper/cook" = "household services",
    "hro" = "human resources",
    "human capital management" = "human resources",
    "human resources" = "human resources",
    "human services" = "human services",
    "immigration" = "immigration",
    "in-house marketing" = "marketing",
    "individual & family services" = "social services",
    "industrial hygiene" = "industrial hygiene",
    "industrial supply" = "industrial supply",
    "information" = "information",
    "information management/archives" = "information management",
    "information sciences" = "information sciences",
    "information services" = "information services",
    "information technology" = "information technology",
    "instructional design and training" = "instructional design",
    "instructional design, aviation industry" = "instructional design",
    "instructional designer" = "instructional design",
    "insurance" = "insurance",
    "intelligence" = "intelligence",
    "intergovernmental organization" = "government",
    "interior design & architecture" = "interior design",
    "interior landscaping" = "landscaping",
    "internal communications" = "communications",
    "international defence" = "defense",
    "international development" = "international development",
    "international organisations" = "international organizations",
    "internet" = "internet",
    "interpretation" = "language services",
    "investing" = "finance",
```

```
    "ipr" = "intellectual property",
    "journalism" = "journalism",
    "lab science (biotech)" = "biotech",
    "laboratory research" = "research",
    "labour/professional organization" = "labor organization",
    "landed estate" = "real estate",
    "landscaping" = "landscaping",
    "landscaping/tree work" = "landscaping",
    "language services" = "language services",
    "large university administration" = "university administration",
    "laundry and rental" = "laundry services",
    "law" = "law",
    "learning and development" = "education",
    "legal services" = "law",
    "leisure, sport & tourism" = "tourism",
    "life sciences" = "life sciences",
    "literature" = "literature",
    "lobbying" = "lobbying",
    "logistics" = "logistics",
    "low-voltage equipment" = "electronics",
    "luxury fashion" = "fashion",
    "maintenance" = "maintenance",
    "management consulting" = "consulting",
    "manufacturing (medical devices)" = "medical manufacturing",
    "maritime" = "maritime",
    "marketing technology" = "marketing",
    "medcomms" = "medical communications",
    "media & digital" = "media",
    "medical technology/education" = "medical technology",
    "mental health therapist" = "mental health",
    "military" = "military",
    "mining and natural resources" = "mining",
    "ministry" = "religious",
    "mortgage" = "finance",
    "multilateral organisation" = "international organizations",
    "museums & archives" = "museums",
    "music" = "music",
    "national laboratory" = "research",
    "natural resources" = "natural resources",
    "no job" = "unemployed",
    "nonprofit organisations" = "nonprofit",
    "nuclear research" = "research",
    "obligatory military service" = "military",
    "oceanography research" = "research",
    "office admin" = "administration",
    "oilfield adjacent" = "energy - oil and gas",
    "online education" = "education",
    "online education startup (non-technical role)" = "education",
    "online learning" = "education",
    "operational training" = "training",
    "operations" = "operations",
    "organizational development" = "organizational development",
    "organized labor" = "labor organization",
    "outdoor industry/repair and maintenance" = "outdoor industry",
```

```
  "outsourced customer service/tech support call centre" = "customer service",
  "outsourcing services" = "outsourcing",
  "paid student intern in tech" = "student intern",
  "parking" = "parking",
  "parks and recreation, land management but with custom…" = "parks and recreation",
  "patent translation" = "translation",
  "payment processing" = "finance",
  "payroll software" = "software",
  "pension benefit administration" = "finance",
  "performing arts" = "performing arts",
  "pest control" = "pest control",
  "pet care" = "pet care",
  "petroleum" = "energy - oil and gas",
  "pharma" = "pharma",
  "phd" = "student",
  "philanthropy" = "nonprofit",
  "physical sciences" = "physical sciences"
)
)
```

Cleaning the Country Column Standardize the country names for better analysis.

```
# Convert country names to lowercase
workforce <- workforce %>%
  mutate(country = tolower(country))

# Recode country names to standardize them
workforce <- mutate(workforce,
  country = recode(
    .x = country,
 "$2,175.84/year is deducted for benefits"         = "na"
 ,"america"= "usa"
 ,"aotearoa new zealand" = "new zealand"
 ,"argentina but my org is in thailand"= "thailand"
 ,"australi"=  "australia"
 ,"australia"            =  "australia"
 , "australian"           =  "australia"
 , "austria, but i work remotely for a dutch/british company"            =   "austria"
 , "bonus based on meeting yearly goals set w/ my supervisor"            = "na"
 , "brasil" =  "brazil"
 , "britain"= "uk"
 , "california"           = "usa"
 , "can"    =  "canada"
 , "canad"  =  "canada"
 , "canada" =  "canada"
 , "canadá" =  "canada"
 , "canada and usa"       =  "canada"
 , "canada, ottawa, ontario"     =  "canada"
 , "canadw" =  "canada"
 , "canda"  =  "canada"
 , "company in germany. i work from pakistan."       = "pakistan"
```

```
 , "csnada" =  "canada"
 , "currently finance"    = "na"
 , "danmark"=  "denmark"
 , "dbfemf" = "na"
 , "england, gb"          = "uk"
 , "england, uk"          = "uk"
 , "england, uk."         = "uk"
 , "england, united kingdom"     = "uk"
 , "england/uk"           = "uk"
 , "englang"= "uk"
 , "ff"= "na"
 , "for the united states government, but posted overseas" = "na"
 , "from new zealand but on projects across apac"    = "apac"
 , "from romania, but for an us based company"       = "usa"
 , "great britain"        = "uk"
 , "hong konh"            =   "hong kong"
 , "i am located in canada but i work for a company in the us"           =  "canada"
 , "i earn commission on sales. if i meet quota, i'm guaranteed another 16k min. last year i e
 , "i was brought in on this salary to help with the ehr and very quickly was promoted to curr
 , "i work for a uae-based organization, though i am personally in the us."    = "usa"
 , "i work for an us based company but i'm from argentina."= "usa"
 , "ibdia"  =  "india"
 , "indonesia"            = "uk"
 , "is"=  "i.s."
 , "isa"    =   "i.s."
 , "israel" =  "canada"
 , "italy (south)"        =   "italy"
 , "japan, us gov position"=  "japan"
 , "jersey, channel islands"     = "usa"
, "kenya"  = "India"
, "london" = "uk"
, "luxembourg"           =  "luxemburg"
, "méxico" =  "mexico"
, "n/a (remote from wherever i want)" = "na"
, "na"= "na"
, "nederland"            =  "netherlands"
, "new zealand"          = "new zealand"
, "nl"= "na"
, "northern ireland"     = "uk"
, "northern ireland, united kingdom"  = "uk"
, "nz"= "new zealand"
, "pakistan"= "pakistan"
, "remote (philippines)" =  "remote"
, "san francisco"        = "usa"
, "scotland, uk"         =  "scotland"
, "ss"= "na"
, "the netherlands"      =  "netherlands"
, "the united states"    = "usa"
, "the us" = "usa"
, "u. s"   = "usa"
, "u. s."  = "usa"
, "u.a."   = "uae"
, "u.k"    = "uk"
, "u.k."   = "uk"
```

```
    , "u.k. (northern england)"     = "uk"
    , "u.s"    = "usa"
    , "u.s."   = "usa"
    , "u.s.a"  = "usa"
    , "u.s.a." = "usa"
    , "u.s>"   = "usa"
    , "u.sa"   = "usa"
    , "ua"= "uae"
    , "uae"    = "uae"
    , "uganda" = "uk"
    , "uk"= "uk"
    , "uk (england)"         = "uk"
    , "uk (northern ireland)"= "uk"
    , "uk for u.s. company"  = "uk"
    , "uk, but for globally fully remote company"       = "uk"
    , "uk, remote"           = "uk"
    , "uniited states"       = "usa"
    , "unite states"         = "usa"
    , "united  states"       = "usa"
    , "united arab emirates" = "uae"
    , "united kindom"        = "uk"
    , "united kingdom"       = "uk"
    , "united kingdom (england)"    = "uk"
    , "united kingdom."      = "uk"
    , "united kingdomk"      = "uk"
    , "united sates"         = "usa"
    , "united sates of america"     = "usa"
    , "united stares"        = "usa"
    , "united state"         = "usa"
    , "united state of america"     = "usa"
    , "united statea"        = "usa"
    , "united stated"        = "usa"
    , "united stateds"       = "usa"
    , "united statees"       = "usa"
    , "united states"        = "usa"
    , "united states- puerto rico"  = "usa"
    , "united states (i work from home and my clients are all over the us/canada/pr"            =
    , "united states is america"    = "usa"
    , "united states of america"    = "usa"
    , "united states of american"   = "usa"
    , "united states of americas"   = "usa"
    , "united statesp"       = "usa"
    , "united statew"        = "usa"
    , "united statss"        = "usa"
    , "united stattes"       = "usa"
    , "united statues"       = "usa"
    , "united status"        = "usa"
    , "united statws"        = "usa"
    , "united sttes"         = "usa"
    , "united y"= "usa"
    , "unitedstates"         = "usa"
    , "uniteed states"       = "usa"
    , "unitef stated"        = "usa"
    , "uniter statez"        = "usa"
```

```
, "unites kingdom"       = "uk"
, "unites states"        = "usa"
, "unitied states"       = "usa"
, "uniyed states"        = "usa"
, "uniyes states"        = "usa"
, "unted states"         = "usa"
, "untied states"        = "usa"
, "us"= "usa"
, "us govt employee overseas, country withheld"     = "usa"
, "us of a"= "usa"
, "usa"    = "usa"
, "usa-- virgin islands" = "usa"
, "usa (company is based in a us territory, i work remote)"= "usa"
, "usa tomorrow"         = "usa"
, "usa, but for foreign gov't"  = "usa"
, "usaa"   = "usa"
, "usab"   = "usa"
, "usat"   = "usa"
, "usd"    = "usa"
, "uss"    = "usa"
, "uxz"    = "na"
, "wales"  = "uk"
, "wales (uk)"           = "uk"
, "wales (united kingdom)"= "uk"
, "wales, uk"            = "uk"
, "we don't get raises, we get quarterly bonuses, but they periodically asses income in the ar
, "worldwide (based in us but short term trips aroudn the world)"       =   "global"
, "y"= "na"
, "uu"= "na"
)
)
```

```
workforce %>%
  group_by(industry) %>%
  count()
```

```
# A tibble: 354 × 2
# Groups:   industry [354]
   industry                                n
   <chr>                               <int>
 1 " marketing / business"                 1
 2 "Contractor"                            1
 3 "No job"                                2
 4 "academia"                              8
 5 "academic publishing/research/science"    19
 6 "academic research (psychology)"        1
 7 "academic research (social science)"    1
 8 "accessibility"                         2
 9 "accounting, banking & finance"         1771
10 "actuarial"                             1
# ℹ 344 more rows
```

Removing the outliers in salary_in_usd

```
remove_outliers_by_iqr <- function(df, column) {
  Q1 <- quantile(df[[column]], 0.25, na.rm = TRUE)
  Q3 <- quantile(df[[column]], 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1

  # Define the lower and upper bounds
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  # Filter the data within the bounds
  df %>% filter(df[[column]] >= lower_bound & df[[column]] <= upper_bound)
}

# Apply the function to each country group
workforce_filtered <- workforce %>%
  group_by(country) %>%
  group_modify(~ remove_outliers_by_iqr(.x, "salary_in_usd")) %>%
  ungroup()


workforce_filtered <- workforce_filtered %>%
  group_by(country, industry) %>%
  filter(n() >= 10) %>%
  ungroup()

# Check the result
head(workforce_filtered)
```

```
# A tibble: 6 × 19
  country submission_timestamp age   industry    job_title     job_title_context
  <chr>   <dttm>               <fct> <chr>       <chr>         <chr>
1 canada  2021-04-27 11:03:10  18-24 healthcare  Patient care… <NA>
2 canada  2021-04-27 11:03:27  35-44 nonprofit   Event Planner <NA>
3 canada  2021-04-27 11:04:09  25-34 engineering Engineering … <NA>
4 canada  2021-04-27 11:04:10  25-34 retail      Manager, Tot… <NA>
5 canada  2021-04-27 11:04:21  35-44 media       Editor        <NA>
6 canada  2021-04-27 11:04:29  25-34 government  Senior Advis… <NA>
# ℹ 13 more variables: annual_salary <dbl>, additional_compensation <dbl>,
#   currency <chr>, other_currency <chr>, income_context <chr>, us_state <chr>,
#   city <chr>, total_years_experience <fct>, years_experience_in_field <fct>,
#   education_level <chr>, gender <chr>, race_ethnicity <chr>,
#   salary_in_usd <dbl>
```

Q 1: Which industry or industries have the highest/lowest salaries? Highest Salary Identify the industries with the highest salaries.

Finding the median salary for each industry

```
salary_over_industry <- workforce %>%
  group_by(industry) %>%
  filter(n() > 5) %>%
  summarize(median_salary_over_industry = median(salary_in_usd, na.rm = TRUE))
```

Identifing the industries with the highest salaries

industry	median_salary (USD)
pharma	128000
consulting	120000
computing or tech	119000

```
# Display the top 10 industries with the highest salaries
# workforce %>%
#   slice_max(salary_in_usd, n = 10) %>%
#   select(industry, annual_salary, currency, salary_in_usd)

salary_over_industry %>%
  slice_max(median_salary_over_industry,n=10) %>%
  select(industry,median_salary_over_industry)
```

```
# A tibble: 10 × 2
   industry          median_salary_over_industry
   <chr>                                   <dbl>
 1 pharma                                 128000
 2 consulting                             120000
 3 computing or tech                      119000
 4 biotech industry                       115000
 5 chemicals                              109000
 6 software                               108850
 7 supply chain                           106634
 8 technology                             104500
 9 trade association                      104500
10 consumer goods                         102000
```

Identifing the industries with the lowest salaries

industry	median_salary (USD)
student	8500.00
childcare	28100.00
administration	40000.00

```
# Filter out entries with salary_in_usd = 0 and display the bottom 50 salaries
# workforce %>%
#   filter(salary_in_usd != 0) %>%
#   slice_min(salary_in_usd, n = 50) %>%
#   select(industry, job_title, salary_in_usd)


salary_over_industry %>%
  slice_min(median_salary_over_industry,n=10) %>%
  select(industry,median_salary_over_industry)
```

```
# A tibble: 10 × 2
   industry                              median_salary_over_industry
   <chr>                                                       <dbl>
 1 student                                                      8500
 2 childcare                                                   28100
 3 administration                                              40000
 4 translation/transcription                                   42000
 5 warehouse                                                   43000
 6 customer service                                            43100
 7 academia                                                    44127.
 8 restaurant                                                  44194
 9 academic publishing/research/science                        48790.
10 veterinary                                                  49410
```

Highest Salary

The Pharmaceutical industry ranks highest, with a median salary of 128,000 USD. Following closely are the Consulting industry with a median salary of 120,000 USD, and the Computing or Technology sector, where the median salary is 119,000 USD. These industries typically offer higher compensation due to the specialized skills, knowledge, and expertise required in these fields.

Lowest Salary

On the other end of the spectrum, the Student category records the lowest median salary at 8,500 USD. Other low-paying industries include Childcare, with a median salary of 28,100 USD, and Administration, where the median salary is 40,000 USD. These industries often involve roles with lower skill requirements or part-time work, which contributes to the lower median salary levels.

Q 2: Which industries have the highest salary variability? Calculate the variability (variance) of salaries by industry

```
# Calculate the variance of salaries by industry
salary_variability <- workforce_filtered %>%
  group_by(industry) %>%
  summarize(salary_variance = var(salary_in_usd, na.rm = TRUE)) %>%
  arrange(desc(salary_variance))
```

```
salary_variability %>%
  head(10)
```

```
# A tibble: 10 × 2
   industry               salary_variance
   <chr>                            <dbl>
 1 wholesale                2958990912.
 2 software                 2665688460.
 3 consulting               2488644231.
 4 technology               2334831111.
 5 entertainment            2098000447.
 6 pharma                   2067294654.
 7 computing or tech        2044546939.
 8 law                      1974029092.
 9 ecommerce                1954745571.
10 business or consulting   1825302442.
```

industry	salary_variance
consulting	2005000000
ecommerce	1954745571
entertainment	1736000439

Q 3: How do salaries vary over time and geography?

Visualization 1:

Title: Salary Comparison by Industry

The graph titled “Salary Comparison by Industry” presents a comparative analysis of median salaries across different industries.

Axes

X-axis: Represents various industries. Y-axis: Displays the median salary values on a logarithmic scale, allowing for better visibility of a wide range of salary distributions. Data Points Each point on the graph corresponds to the median salary for a specific industry. Blue points depict the median salaries across all industries. #### Highlighting Extremes The highest salary point is specifically annotated and circled in red, indicating the pharma industry, which has a median salary of 128,000 USD. The label next to this point reads: “Highest: pharma - 128,000 USD.” An arrow points directly to the salary point for clarity.

The lowest salary point, labeled “Lowest: student - 8,500 USD,” indicates the student industry. This point is also circled in red and has an arrow connecting it to the label, providing a visual cue to the reader.

Logarithmic Scale

The y-axis utilizes a logarithmic scale, allowing for a more effective comparison of median salaries that vary significantly. This approach highlights the disparity in salary levels, making it easier to interpret data

points that span several orders of magnitude. Customizations The x-axis does not display tick marks or labels, focusing the viewer's attention on the industries themselves. The graph employs a clean and minimalistic design using the theme_bw() function, ensuring that the data points and annotations stand out prominently against the white background.

```
salaries <- salary_over_industry %>%
  mutate (industry= fct_reorder(industry, desc(median_salary_over_industry)))

highest_salary <- salaries %>% filter(median_salary_over_industry == max(median_salary_over_in
lowest_salary <- salaries %>% filter(median_salary_over_industry == min(median_salary_over_ind

salaries %>%
  ggplot(aes(x = industry, y = median_salary_over_industry)) +
  geom_point(size = 1) +
  geom_point(data = salaries, col = "blue", size = 2) +

  # Annotating the highest salary point
  annotate(
    geom = "label", x = as.numeric(factor(highest_salary$industry)) + 5, y = highest_salary$me
    label = paste("Highest:", highest_salary$industry, "-",highest_salary$median_salary_over_i
  ) +
  annotate(
    geom = "segment",
    x = as.numeric(factor(highest_salary$industry)) + 5,  # Arrow starts at the label position
    y = highest_salary$median_salary_over_industry,
    xend = as.numeric(factor(highest_salary$industry)),  # Arrow ends at the highest salary po
    yend = highest_salary$median_salary_over_industry,
    color = "grey30",
    arrow = arrow(type = "closed"),
    size = 0.5
)+

  # Annotating the lowest salary point
  annotate(
    geom = "label",
    x = as.numeric(factor(lowest_salary$industry)) +55, # 5 ticks to the left of the lowest sa
    y = lowest_salary$median_salary_over_industry,
    label = paste("Lowest:", lowest_salary$industry, "-",lowest_salary$median_salary_over_indu
    hjust = "left",
    color = "grey30",
    size = 3
) +
  annotate(
    geom = "segment",
    x = as.numeric(factor(lowest_salary$industry)) + 80, y = lowest_salary$median_salary_over_
    xend = lowest_salary$industry, yend = lowest_salary$median_salary_over_industry,
    color = "grey30", arrow = arrow(type = "closed")
  ) +
  geom_point(data = highest_salary, aes(x = industry, y = median_salary_over_industry),
             size = 3, color = "red", shape = 21, stroke = 1.5) +

  # Red circle around the lowest salary point
  geom_point(data = lowest_salary, aes(x = industry, y = median_salary_over_industry),
```

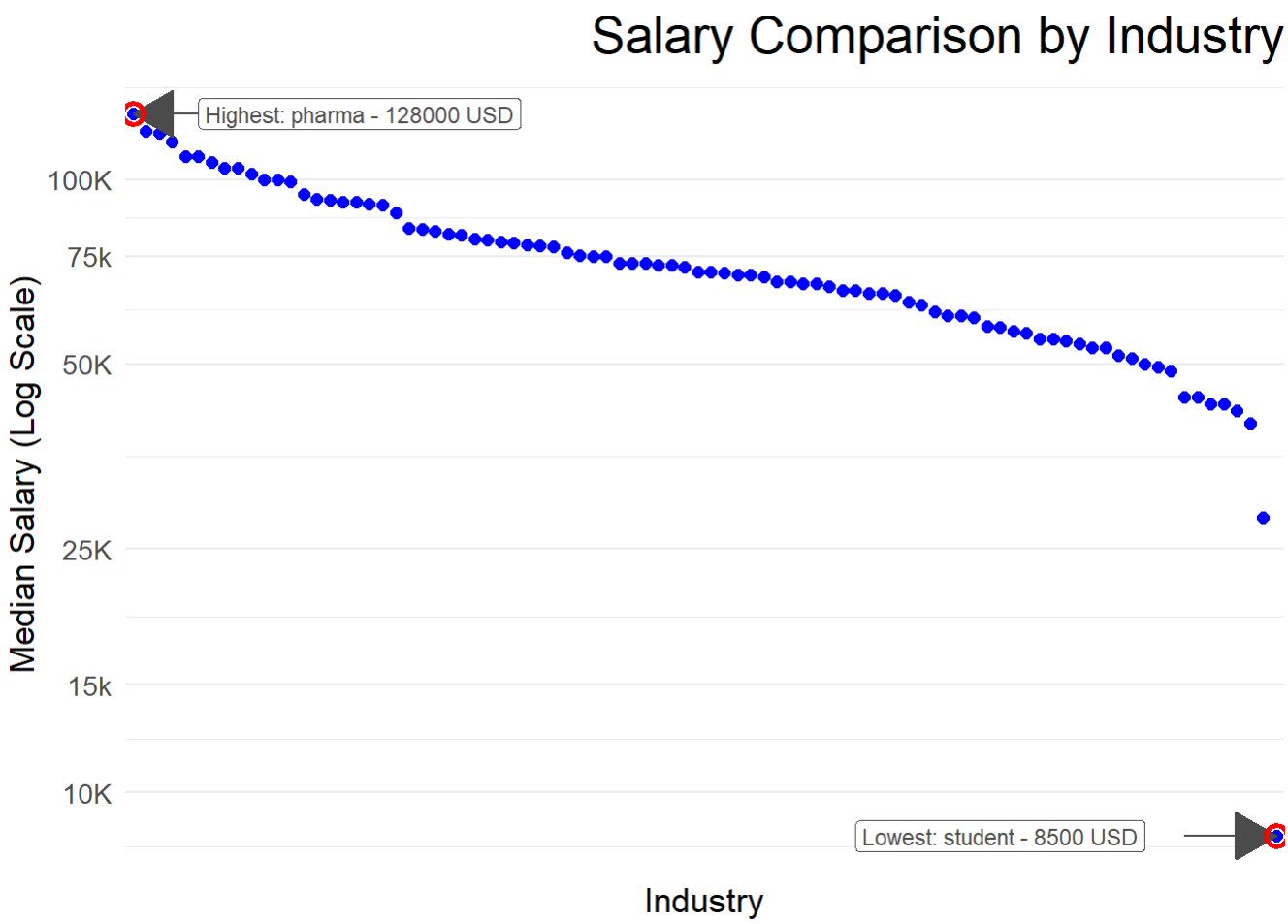
```
            size = 3, color = "red", shape = 21, stroke = 1.5) +

  scale_x_discrete(labels = NULL) +
    scale_y_log10(breaks = c( 10000,15000,25000, 50000,75000, 100000, 200000),
                  labels = c( "10K","15k","25K", "50K","75k", "100K", "200K")) +

  theme_bw() +
  ggtitle("Salary Comparison by Industry") +
  theme(axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        panel.border = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        text = element_text(size = 13),
        plot.title = element_text(hjust = 1,size =20))+

labs(x = "Industry", y = "Median Salary (Log Scale)")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



Visualization 2:

Tittle: Comparison of Median Salaries with Experience

This bar graph compares the median salaries across different years of experience, both in total years and years specifically in the field. The data is derived from the `workforce` dataset, which contains information on salaries and years of experience.

Data Preparation:

1. Median Calculation: The median salary is calculated for each combination of `total_years_experience` and `years_experience_in_field`. This helps in understanding the central tendency of salaries for different experience levels.
2. Data Reshaping: The data is reshaped using `pivot_longer` to facilitate easier plotting. This transformation converts the data from wide format to long format, making it suitable for ggplot2's aesthetics.

Visualization:

- Axes:
  - X-axis (`experience_years`): Represents the years of experience, ranging from the minimum to the maximum values in the dataset.
  - Y-axis (`median_salary`): Represents the median salary, plotted on a logarithmic scale to handle the wide range of salary values. The scale is broken down into major intervals (1K, 2K, 5K, 10K, 20K, 50K, 100K) for better readability.
- Bars: The bars are filled based on the `experience_type` (either `total_years_experience` or `years_experience_in_field`), allowing for a side-by-side comparison of median salaries for each experience level.
- Title and Labels: The title is centered and clearly states the purpose of the graph. Axis labels and the legend title are appropriately named to guide the viewer.
- Theme: A clean, black-and-white theme is applied to enhance readability, with minor grid lines and axis ticks removed for clarity.

Interpretation:

The graph effectively illustrates how median salaries vary with different years of experience. By comparing the bars for `total_years_experience` and `years_experience_in_field`, one can infer whether the salary growth is more influenced by the total years worked or the years specifically spent in the field. The logarithmic scale ensures that the differences in salary are visually apparent, even for lower and higher salary ranges.

Key Observations:

- General Trend: Median salaries are generally higher for years of experience in the field compared to total years of experience, indicating that specialized experience in a particular field tends to command higher salaries.
- Exceptions: There are notable exceptions in the 2-4 years and 11-20 years experience ranges:

- 2-4 Years: Median salaries for total years of experience are higher than those for years in the field. This suggests that early career stages might benefit more from general experience rather than specialized experience.

- 11-20 Years: Similarly, median salaries for total years of experience are higher than those for years in the field. This could indicate a plateau or slowdown in salary growth for highly specialized experience in this range.

Conclusion:

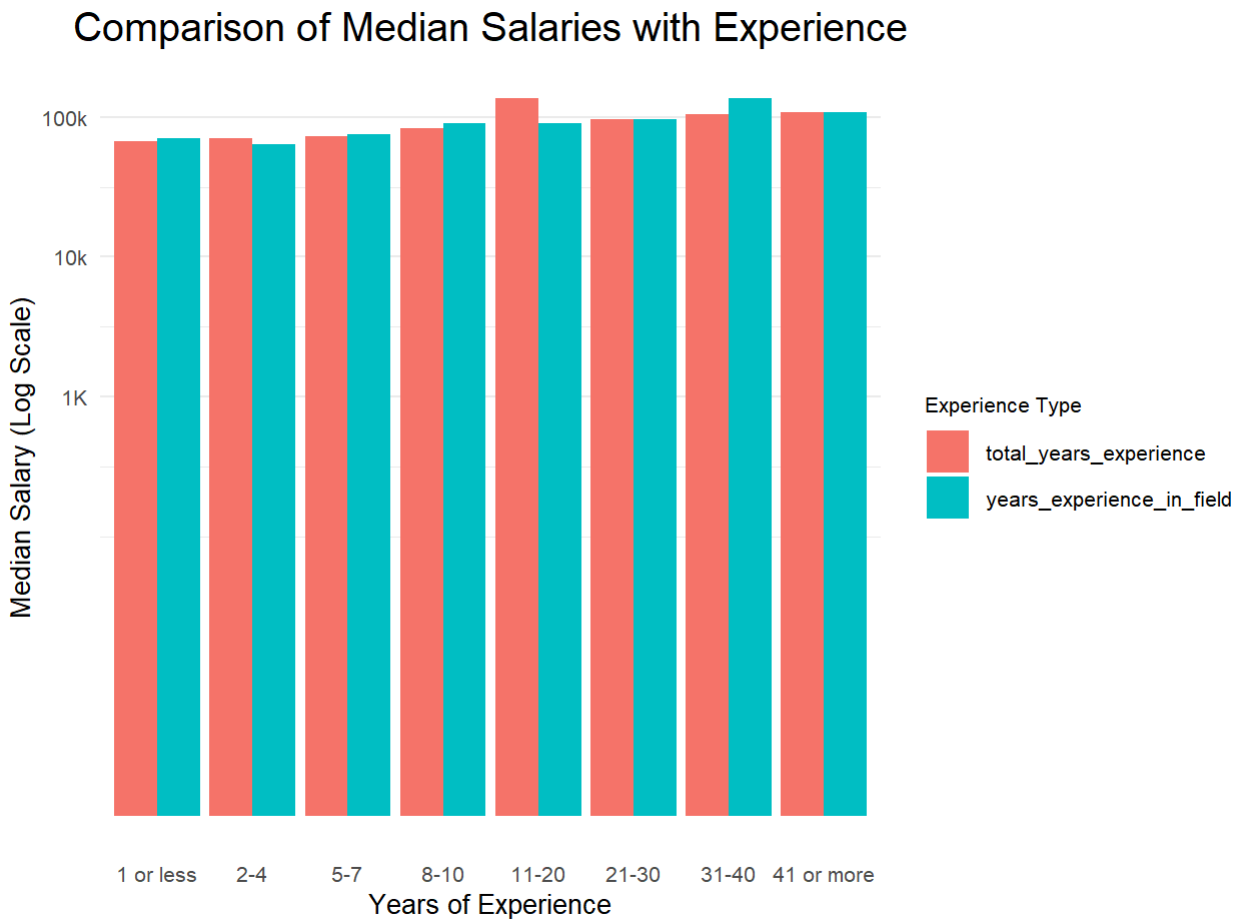
This visualization provides a clear and concise comparison of median salaries across different experience levels, helping to identify trends and patterns in salary growth. The use of advanced ggplot2 features, such as logarithmic scaling and data reshaping, enhances the clarity and interpretability of the graph. The specific observations highlight the nuanced relationship between experience and salary, offering valuable insights for career planning and compensation strategies.

```
# Calculate the median salary for each experience category
median_salaries_over_experience <- workforce_filtered %>%
  select(salary_in_usd,total_years_experience,years_experience_in_field) %>%
  group_by(total_years_experience, years_experience_in_field) %>%
  summarise(median_salary = median(salary_in_usd, na.rm = TRUE), .groups = 'drop')

# Reshape the data for plotting
median_salaries_long <- median_salaries_over_experience %>%
  pivot_longer(cols = c("total_years_experience", "years_experience_in_field"),
               names_to = "experience_type",
               values_to = "experience_years")

# Create the bar graph
ggplot(median_salaries_long, aes(x = experience_years, y = median_salary, fill = experience_ty
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_y_log10(
    breaks = c(1000,   10000,   100000), # Added more breaks for symmetry
    labels = c("1K",   "10k",   "100k")
    #limits = c(1000, 150000) # Set limits to provide a better range for the log scale
  )  +
  labs(title = "Comparison of Median Salaries with Experience",
       x = "Years of Experience",
       y = "Median Salary (Log Scale)",
       fill = "Experience Type") +
  scale_x_discrete(
    breaks = c("1 year or less", "2 - 4 years", "5-7 years", "8 - 10 years", "11 - 20 years",
    labels = c("1 or less", "2-4", "5-7", "8-10", "11-20", "21-30", "31-40", "41 or more"))+
  theme_bw() +
  theme(axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        panel.border = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        text = element_text(size = 10),
        plot.title = element_text(hjust = 0.5,size=15),
    legend.title = element_text(size = 8),  # Adjust the size of the legend title
```

```
    legend.text = element_text(size = 8),
    plot.margin = margin(t = 10, r = 20, b = 10, l = 20))
```



Visualization 3

Title: Salary Distribution by Country

Caption:

Boxplot shows median, quartiles, and potential outliers in salaries. Data log-transformed for readability.

Description:

This visualization presents the distribution of annual salaries across different countries, with a focus on the median salary. The boxplot is constructed using the ggplot2 package in R, and it employs several advanced features to enhance clarity and readability.

Data preparation:

The dataset is first filtered to remove any entries where the country is “na”.

The median salary for each country is calculated and used to order the countries in descending order based on their median salary.

Graphical Representation:

The boxplot is used to represent the salary distribution, with each box corresponding to a country.

The y-axis is log-transformed to accommodate a wide range of salary values, making it easier to visualize differences between countries with vastly different salary scales.

Custom breaks on the y-axis are set to 10^seq(0, 7, by = 1), ensuring that the logarithmic scale is clearly marked.

Interpretation:

The boxplot effectively highlights the median salary for each country, along with the interquartile range (IQR) and potential outliers.

The log-transformation of the y-axis ensures that the visualization is not dominated by extreme values, allowing for a more balanced comparison across countries.

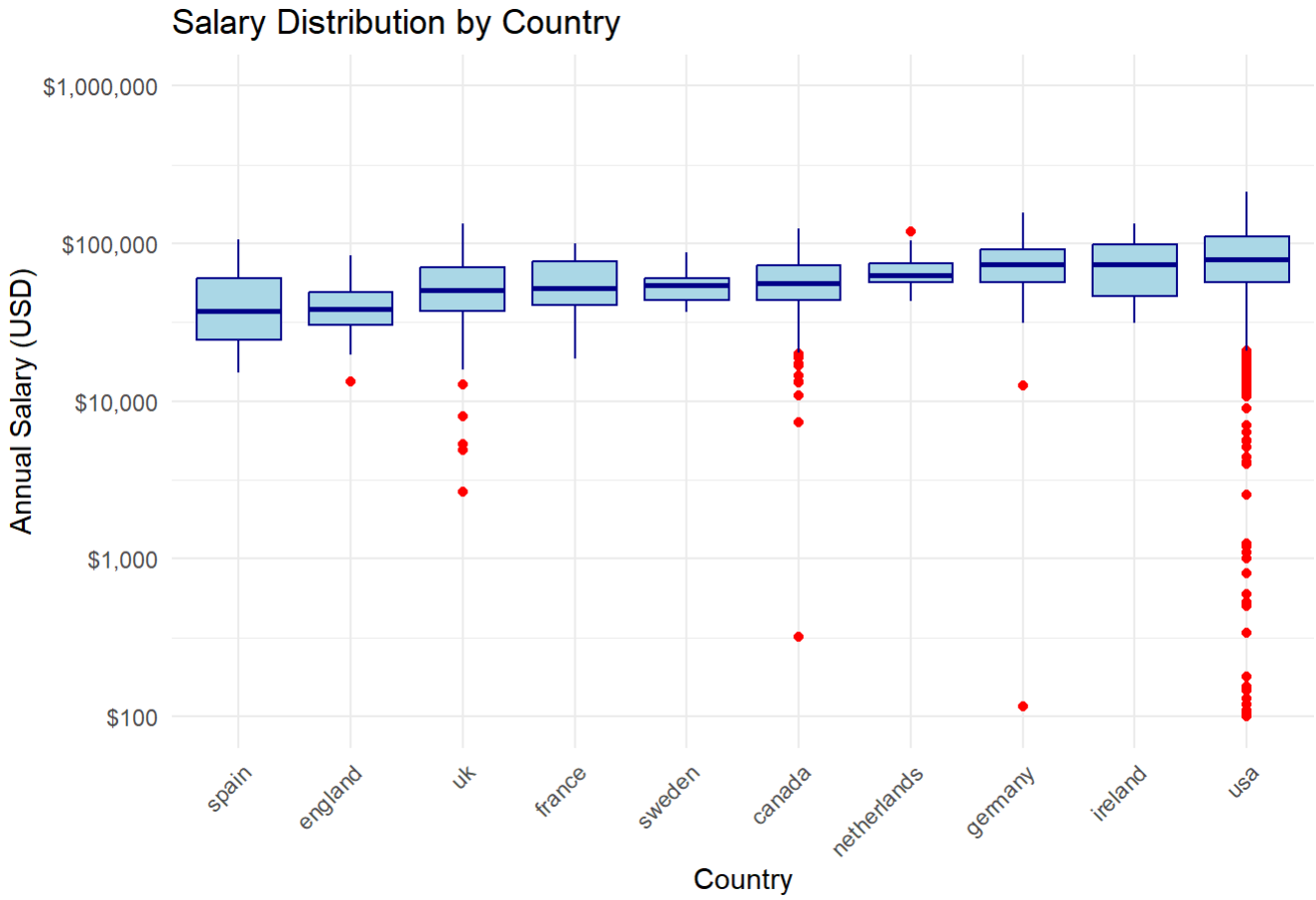
```
# Create a boxplot of salary distribution by country
library(ggplot2)

graph3 <- ggplot(workforce_filtered,
       aes(x = reorder(country, salary_in_usd, FUN = median),
           y = salary_in_usd)) +
  geom_boxplot(outlier.shape = 16,  # Change to shape 16 (filled circle) or any other shape
               outlier.colour = "red",  # Color of the outliers
               fill = "lightblue",
               color = "darkblue") +
  scale_y_continuous(trans = "log10",
                     labels = scales::dollar,
                     limits = c(100, 10^6)) +  # Adjust limits on log scale
  labs(x = "Country",
       y = "Annual Salary (USD)",
       title = "Salary Distribution by Country",
       caption = "Boxplot shows median, quartiles, and potential outliers in salaries. Data lo
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  # Skewed labels at 45 degrees
  expand_limits(y = 1e9)  # Ensure we have space at the upper end

# Display the graph
graph3
```

```
Warning in scale_y_continuous(trans = "log10", labels = scales::dollar, :
log-10 transformation introduced infinite values.

Warning: Removed 48 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Boxplot shows median, quartiles, and potential outliers in salaries. Data log-transformed for readability.

Visualization 4:

Title: Median Salary Over Time Create a line plot to visualize median salary over time

Description

The bar chart displays the median salary (in USD) across different countries. The length of each bar corresponds to the median salary, and the countries are arranged in ascending order based on the median salary value.

Graphical Representation

The horizontal bar chart uses the ggplot2 package in R, where:

The x-axis represents the median salary in USD.

The y-axis represents different countries.

Each bar's length indicates the median salary for the respective country.

The bars are colored steel blue for uniformity.

The chart title is "Median Salary by Country," and axes are clearly labeled.

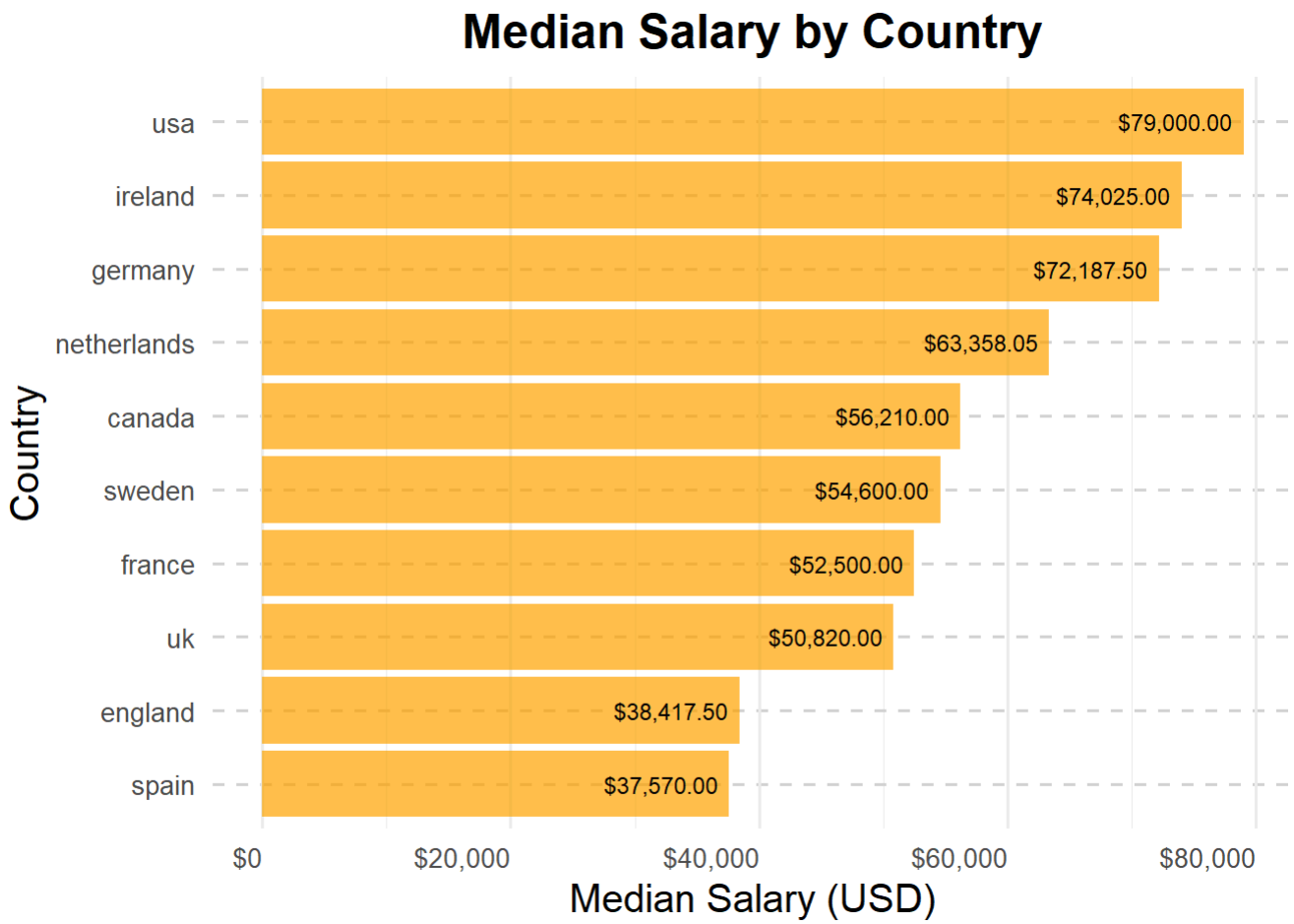
Interpretation

This visualization provides a straightforward comparison of median salaries across countries. Countries with longer bars have higher median salaries, while those with shorter bars have lower median salaries. Sorting the countries by their median salary helps to quickly identify which regions have higher or lower typical earnings. From this graph, it is evident that there are significant salary disparities across different countries, with certain nations exhibiting much higher median salaries compared to others.

```
# Calculate median salary by country
median_salary_by_country <- workforce_filtered %>%
  select(salary_in_usd, country) %>%
  group_by(country) %>%
  summarise(median_salary = median(salary_in_usd, na.rm = TRUE), .groups = 'drop')

# Create the bar plot
graph4 <- ggplot(median_salary_by_country, aes(x = reorder(country, median_salary), y = median_
  geom_bar(stat = "identity", fill = "orange", alpha = 0.7) +  # Adjust color and transparency
  geom_text(aes(label = scales::dollar(median_salary)), hjust = 1.1, size = 3, color = "black"
  coord_flip() +
  labs(title = "Median Salary by Country",
       x = "Country",
       y = "Median Salary (USD)") +
  theme_minimal(base_size = 15) +  # Increase base font size
  theme(
    axis.text.x = element_text(size = 10,  hjust = 1),  # Customize x-axis text
    axis.text.y = element_text(size = 10),  # Customize y-axis text
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),  # Center title and make
    panel.grid.major.y = element_line(color = "lightgray", linetype = "dashed")  # Add dashed
  ) +
  scale_y_continuous(labels = scales::dollar)  # Format y-axis labels as dollar amounts

# Display the graph
graph4
```



Visualization 5:

Title: Median Salary Over Time (Proportional Observations)

Caption:

Line plot shows the median salary over time. Points are sized by the proportion of observations.

Description:

This visualization presents the median annual salary over time, with a focus on the trend across different year-months. The line plot is constructed using the `ggplot2` package in R, and it employs several advanced features to enhance clarity and readability.

Data Preparation:

- The `submission_timestamp` is converted to a datetime format.
- The dataset is grouped by year-month, and the median salary and count of records are calculated for each year-month.
- The proportion of observations for each year-month is calculated.

Graphical Representation:

- The line plot is used to represent the trend of median salary over time.

- Points are added to the line plot, with their size corresponding to the proportion of observations for each year-month.
- The line is colored in “blue”, and the points are colored in “red”, providing a visually appealing contrast.

Aesthetic Enhancements:

- The theme is set to `theme_minimal()`, with additional customizations to enhance readability:
  - The plot title is bold and centered.
  - Axis titles and text are appropriately sized and colored for clarity.
  - The x-axis labels are rotated 45 degrees to avoid overlap.
  - The background colors are set to white to reduce visual clutter.

Custom X-Axis Labels:

- The x-axis is customized to show only the years, with breaks and labels set to every 12 months.

Interpretation:

- The line plot effectively highlights the trend of median salary over time.
- The size of the points indicates the proportion of observations for each year-month, providing additional context to the data.

```
# Convert submission_timestamp to datetime
workforce_filtered$submission_datetime <- as.POSIXct(workforce_filtered$submission_timestamp,

# Group by year-month and calculate median salary and count
salary_over_time <- workforce_filtered %>%
  group_by(year_month = format(workforce_filtered$submission_datetime, "%Y-%m")) %>%
  summarize(median_salary = median(salary_in_usd, na.rm = TRUE), n = n())

# Calculate proportion of observations for each year-month
salary_over_time <- salary_over_time %>%
  mutate(proportion_n = n / sum(n))

# Create the graph
graph5 <- ggplot(salary_over_time, aes(x = year_month, y = median_salary, group = 1)) +
  geom_line(color = "steelblue", size = 1.5) +  # Line connecting the points
  geom_point(aes(size = proportion_n), color = "red") +  # Size of points reflects proportion
  scale_size_continuous(labels = scales::percent_format(accuracy = 1)) +  # Show size as perce
  scale_x_discrete(breaks = unique(salary_over_time$year_month)[order(unique(salary_over_time$
                   labels = unique(salary_over_time$year_month)[order(unique(salary_over_time$
  labs(title = "Median Salary Over Time (Proportional Observations)",
       x = "Year",
       y = "Median Salary (USD)",
       caption = "Line plot shows the median salary over time. (More the number of observation
  theme_minimal() +
  theme(
```

```
    plot.title = element_text(size = 14, face = "bold", color = "black", hjust = 0.5),
    plot.caption = element_text(size = 10, color = "gray30"),
    axis.title = element_text(size = 12, color = "black"),
    axis.text = element_text(size = 10, color = "gray30"),
    axis.text.x = element_text(),
    panel.grid.major = element_line(color = "gray80"),
    panel.grid.minor = element_blank(),
    #panel.background = element_rect(fill = "white"),
    plot.background = element_rect(fill = "white"),
    legend.position = "none"
  )

# Display the graph
graph5
```