Visualizing Research University Diversity in the U.S.

How Institutional Characteristics Align with Carnegie Classification Tiers

INST 462 (0101) - Diamond Andy, Lily Gates, Mia Leandri, Colin Thompson

05/13/2025

# Visualizing Research University Diversity in the U.S.

## *How Institutional Characteristics Align with Carnegie Classification Tiers*

### Description of Project

Exploring the differences between American research colleges and universities through data visualization.

### Research Question/Topic

1. Descriptive and Comparative Question  
   **How do different tiered American research colleges and universities compare in terms of location, admissions, academic offerings, doctoral degrees awarded, and other institutional characteristics?**
2. Analytical and Inferential Question  
   **How do other institutional characteristics, besides financial expenditures, predict or correlate with research designation tiers among American colleges and universities?**

### Data and Research Sources

The README.md on the GitHub repository contains a more detailed overview of the data.

carnegie\_spending\_doctorate\_awards (See carnegie\_research\_activity\_desig\_factsheet.pdf for more info) \* School \* Location (City, State) \* Classification (R1, R2, Other) — From 2021 and 2025 \* Expenses (2021, 2022, 2023) \* Research Doctorates Awarded (2020, 2021, 2022)

carnegie\_classification\_stats (See carnegie\_variable\_descr\_flowchart.pdf for more info) \* School \* Filtering main school if applicable (e.g., UMBC, UMD, UMES all are “University of Maryland” sub schools) \* Location (City, State, residential or not) \* Public or private (non-profit, for-profit) \* Description of types of degree programs offered \* Enrollment information \* Classification on tier (R1, R2, Other)

# CODE BEGINS

Bulk Uncomment/Comment: Hightlight text –> “Code” -> “Comment/Uncomment Lines”

# Core Tidyverse and Dependencies  
# install.packages("tidyverse") # Includes ggplot2, dplyr, tidyr, readr, etc.  
# install.packages("lubridate") # For working with dates and times  
# install.packages("stringr") # For string operations  
#   
# # Visualization  
# install.packages("ggplot2") # Core plotting package (also part of tidyverse)  
# install.packages("plotly") # For interactive graphs  
# install.packages("ggbeeswarm") # For beeswarm plots  
# install.packages("patchwork") # For combining ggplot2 objects  
# install.packages("RColorBrewer") # For advanced color palettes  
# install.packages("scales") # For axis formatting and transformations  
#   
# # Development Tools  
# install.packages("devtools") # For installing development tools and palettes

# Importing Libraries  
  
# Core Data Science & Tidyverse Packages  
library(tidyverse) # Includes ggplot2, dplyr, tidyr, readr, stringr, etc.

## ── 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 3.5.2 ✔ 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

library(readr) # For reading CSV files  
library(tidyr) # For reshaping data  
library(dplyr) # For data manipulation  
library(stringr) # For string operations  
library(lubridate) # For working with dates and times  
  
# Visualization Packages  
library(ggplot2) # For creating static data visualizations  
library(plotly) # For interactive graphs

##   
## Attaching package: 'plotly'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following object is masked from 'package:graphics':  
##   
## layout

library(ggbeeswarm) # For beeswarm-style plots  
library(patchwork) # For combining multiple ggplot figures  
library(RColorBrewer) # For color palettes  
library(scales) # For formatting and transforming scales

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

# Additional Utilities  
library(devtools) # Used for development tools and installing additional palettes

## Loading required package: usethis

# Reading in Data  
classification\_stats <- read\_csv("carnegie\_classification\_stats.csv")

## Rows: 542 Columns: 17  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (16): name, core\_name, spec\_name, city, state, level, public\_private\_pro...  
## dbl (1): unitid  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

spending\_doctorate\_awards <- read\_csv("carnegie\_spending\_doctorate\_awards.csv")

## Rows: 542 Columns: 14  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (7): name, city, state, classific\_2025, classific\_2021, herd\_fy21, herd\_...  
## dbl (5): unitid, num\_doc\_degrees\_2020\_2021, num\_doc\_degrees\_2021\_2022, num\_d...  
## num (2): herd\_fy23, herd\_avg\_fy21\_to\_fy23  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# CREATING DATAFRAMES

# Convert Both Imported CSVs to Dataframes  
classification\_df <- data.frame(classification\_stats)  
spending\_df <- data.frame(spending\_doctorate\_awards)

# Combine Both into Single Dataframe  
# Note: has name.x and name.y  
# - This is because the `classification\_df` and `spending\_df` both have "name" but it was raising an error: "Error: unexpected symbol" for that column when trying to do a full join  
  
# Perform full join on 'unitid'  
combined\_df <- full\_join(classification\_df, spending\_df, by = "unitid")  
  
# Confirm join and notice duplicate column names  
#colnames(combined\_df)

# CLEANING DATA

# Debugging - Checking for Duplicate Values in Duplicate Column Names  
# (name, city, state each have x and y values)  
# The `.x` refers to the column coming from `classification\_df`  
# The `.y` refers to the column coming from `spending\_df`  
  
# Duplicate School Names (there are 12)  
combined\_df %>%  
 filter(name.x != name.y) %>%  
 select(unitid, name.x, name.y)

## unitid name.x  
## 1 104151 Arizona State University  
## 2 138354 The University of West Florida  
## 3 151111 Indiana University-Purdue University-Indianapolis  
## 4 163259 University of Maryland, Baltimore  
## 5 163286 University of Maryland-College Park  
## 6 164155 US Naval Academy  
## 7 195049 Rockefeller University  
## 8 196060 SUNY at Albany  
## 9 199111 University of North Carolina at Asheville  
## 10 201885 University of Cincinnati  
## 11 207388 Oklahoma State University  
## 12 224554 East Texas A & M  
## name.y  
## 1 Arizona State University Campus Immersion  
## 2 University of West Florida  
## 3 Indiana University–Purdue University-Indianapolis  
## 4 University of Maryland - Baltimore  
## 5 University of Maryland - College Park  
## 6 United States Naval Academy  
## 7 The Rockefeller University  
## 8 University at Albany  
## 9 University of North Carolina Asheville  
## 10 University of Cincinnati-Main Campus  
## 11 Oklahoma State University-Main Campus  
## 12 East Texas A & M University

# Duplicate City Names (there is 1, 'Des Moines' and 'West Des Moines')  
combined\_df %>%  
 filter(city.x != city.y ) %>%  
 select(unitid, city.x, city.y)

## unitid city.x city.y  
## 1 154156 Des Moines West Des Moines

# Duplicate State Names (there are no duplicates)  
combined\_df %>%  
 filter(state.x != state.y) %>%  
 select(unitid, state.x, state.y)

## [1] unitid state.x state.y  
## <0 rows> (or 0-length row.names)

# Create a new dataframe for mismatched rows  
mismatched <- combined\_df %>%  
 filter(name.x != name.y | city.x != city.y | state.x != state.y) %>%  
 select(unitid, name.x, name.y, city.x, city.y, state.x, state.y)  
  
# View the mismatched rows  
print(mismatched)

## unitid name.x  
## 1 104151 Arizona State University  
## 2 138354 The University of West Florida  
## 3 151111 Indiana University-Purdue University-Indianapolis  
## 4 154156 Des Moines University-Osteopathic Medical Center  
## 5 163259 University of Maryland, Baltimore  
## 6 163286 University of Maryland-College Park  
## 7 164155 US Naval Academy  
## 8 195049 Rockefeller University  
## 9 196060 SUNY at Albany  
## 10 199111 University of North Carolina at Asheville  
## 11 201885 University of Cincinnati  
## 12 207388 Oklahoma State University  
## 13 224554 East Texas A & M  
## name.y city.x  
## 1 Arizona State University Campus Immersion Tempe  
## 2 University of West Florida Pensacola  
## 3 Indiana University–Purdue University-Indianapolis Indianapolis  
## 4 Des Moines University-Osteopathic Medical Center Des Moines  
## 5 University of Maryland - Baltimore Baltimore  
## 6 University of Maryland - College Park College Park  
## 7 United States Naval Academy Annapolis  
## 8 The Rockefeller University New York  
## 9 University at Albany Albany  
## 10 University of North Carolina Asheville Asheville  
## 11 University of Cincinnati-Main Campus Cincinnati  
## 12 Oklahoma State University-Main Campus Stillwater  
## 13 East Texas A & M University Commerce  
## city.y state.x state.y  
## 1 Tempe AZ AZ  
## 2 Pensacola FL FL  
## 3 Indianapolis IN IN  
## 4 West Des Moines IA IA  
## 5 Baltimore MD MD  
## 6 College Park MD MD  
## 7 Annapolis MD MD  
## 8 New York NY NY  
## 9 Albany NY NY  
## 10 Asheville NC NC  
## 11 Cincinnati OH OH  
## 12 Stillwater OK OK  
## 13 Commerce TX TX

# Cleaning Data -- Fill any missing values with "NA"  
  
# Function to replace NULL with NA  
replace\_null\_with\_na <- function(x) {  
 if (is.list(x)) {  
 # Apply recursively if it's a list (to handle nested data frames)  
 return(lapply(x, replace\_null\_with\_na))  
 } else if (is.null(x)) {  
 return(NA) # Replace NULL with NA  
 } else {  
 return(x) # Otherwise, return the value unchanged  
 }  
}  
  
# Use it to apply to a data frame  
combined\_df\_na <- combined\_df %>%  
 mutate(across(everything(), replace\_null\_with\_na))  
  
# View the result  
#summary(combined\_df\_na)

# Debugging - Checking which columns have missing values (NA)  
# Note: `spec\_name` may have many values and that is okay, it is optional (e.g., University of Maryland, College Park would have "College Park" as the spec\_name", but John's Hopkins only has one campus, so there is no "main campus" or any other special name)  
# - Should be referring to Penn State, with 2 different unitid AND different school name spelling  
  
# Display a table of the count of missing values (NAs) per column  
missing\_data <- combined\_df\_na %>%   
 summarise(across(everything(), ~ sum(is.na(.)))) %>%  
 gather(key = "column\_name", value = "NAs")  
  
# Show columns  
print(missing\_data)

## column\_name NAs  
## 1 unitid 0  
## 2 name.x 1  
## 3 core\_name 1  
## 4 spec\_name 371  
## 5 city.x 1  
## 6 state.x 1  
## 7 level 1  
## 8 public\_private\_profit 1  
## 9 undergrad\_program 1  
## 10 grad\_program 1  
## 11 enrollment\_profile 1  
## 12 undergrad\_profile 1  
## 13 size\_setting 1  
## 14 degree\_focuses 1  
## 15 community\_engage 1  
## 16 leadership\_for\_public\_prax 1  
## 17 research\_tier\_2025 1  
## 18 name.y 1  
## 19 city.y 1  
## 20 state.y 1  
## 21 classific\_2025 1  
## 22 classific\_2021 1  
## 23 herd\_fy21 1  
## 24 herd\_fy22 1  
## 25 herd\_fy23 1  
## 26 herd\_avg\_fy21\_to\_fy23 1  
## 27 num\_doc\_degrees\_2020\_2021 1  
## 28 num\_doc\_degrees\_2021\_2022 1  
## 29 num\_doc\_degrees\_2022\_2023 1  
## 30 avg\_num\_doc\_degrees\_2020\_2023 1

# Show rows for any NA values, EXCEPT for spec\_name (as this is a common column with NA)  
rows\_with\_na\_except\_spec\_name <- combined\_df\_na %>%  
 filter(if\_any(-all\_of("spec\_name"), is.na))  
  
print(rows\_with\_na\_except\_spec\_name)

## unitid name.x core\_name  
## 1 495767 The Pennsylvania State University The Pennsylvania State University  
## 2 214777 <NA> <NA>  
## spec\_name city.x state.x level  
## 1 University Park University Park PA Four or more years  
## 2 <NA> <NA> <NA> <NA>  
## public\_private\_profit  
## 1 Public  
## 2 <NA>  
## undergrad\_program  
## 1 Professions plus arts & sciences high graduate coexistence  
## 2 <NA>  
## grad\_program  
## 1 Research Doctoral Comprehensive programs with medical/veterinary school  
## 2 <NA>  
## enrollment\_profile undergrad\_profile  
## 1 High undergraduate Four-year full-time selective lower transfer-in  
## 2 <NA> <NA>  
## size\_setting  
## 1 Four-year large primarily residential  
## 2 <NA>  
## degree\_focuses  
## 1 Doctoral Universities Very High Research Activity  
## 2 <NA>  
## community\_engage leadership\_for\_public\_prax  
## 1 Classified or Reclassified in 2020/2015 NULL  
## 2 <NA> <NA>  
## research\_tier\_2025  
## 1 Research 1: Very High Research Spending and Doctorate Production  
## 2 <NA>  
## name.y city.y state.y  
## 1 <NA> <NA> <NA>  
## 2 Pennsylvania State University-Main Campus University Park PA  
## classific\_2025  
## 1 <NA>  
## 2 Research 1: Very High Spending and Doctorate Production  
## classific\_2021 herd\_fy21  
## 1 <NA> <NA>  
## 2 Doctoral Universities: Very High Research Activity 970,544,000.00  
## herd\_fy22 herd\_fy23 herd\_avg\_fy21\_to\_fy23 num\_doc\_degrees\_2020\_2021  
## 1 <NA> NA NA NA  
## 2 1,019,940,000.00 1206793000 1065759000 0  
## num\_doc\_degrees\_2021\_2022 num\_doc\_degrees\_2022\_2023  
## 1 NA NA  
## 2 718 701  
## avg\_num\_doc\_degrees\_2020\_2023  
## 1 NA  
## 2 473

# VISUALIZATIONS

### Note: All graphs will be displayed in a pop-up window when code is run. Graphs will also be saved as a .png in the current working directory.

# Template Theme for All Plots (for consistency)  
  
my\_plot\_theme <- theme(  
 # Title settings with padding below  
 plot.title = element\_text(size = 24, hjust = 0, face = "bold", margin = margin(b = 10)),  
  
 # Subtitle settings with padding below  
 plot.subtitle = element\_text(size = 18, hjust = 0, face = "italic", margin = margin(b = 10)),  
  
 # Caption settings with padding above  
 plot.caption = element\_text(size = 12, hjust = 0, color = "black", margin = margin(t = 10)),  
  
 # Axis titles and text settings  
 axis.title = element\_text(size = 16, color = "black"),  
 axis.text.x = element\_text(size = 14, color = "black"),  
 axis.text.y = element\_text(size = 14, color = "black"),  
 strip.text = element\_text(size = 18, face = "bold"), # Sections in a faceted graph  
  
 # Panel settings  
 panel.grid.minor = element\_blank(),  
 panel.grid.major = element\_blank(),  
 axis.line = element\_line(color = "black", linewidth = 1.25),  
 panel.background = element\_rect(fill = "white", color = NA),  
 plot.background = element\_rect(fill = "white", color = NA),  
  
 # Legend settings  
 legend.title = element\_text(size = 16, margin = margin(b = 5)),  
 legend.text = element\_text(size = 14),  
 legend.key.size = unit(1, "cm"),  
 legend.spacing.y = unit(0.6, "cm"),  
  
 # Plot margins  
 plot.margin = margin(15, 15, 15, 15)  
)

## 1. Distribution of U.S. institutions by Carnegie classification (2021)

### By Diamond Andy and Lily Gates

This bar graph represents the number of institutions that were categorized under the 2021 Carnegie research classification tier. Comparing how many universities fall into each group and highlights the overall distribution of research in U.S. Institutions. This visualization also helps emphasize the tiers that are less common. Offering a clear and concise view of how universities are distributed across different levels of research activity. The Carnegie Classification system is a widely used framework in the U.S. for categorizing institutions based on their degree-granting activity and research intensity, particularly at the doctoral level.

### 1a. Distribution of U.S. institutions by Carnegie Classification (2021): Raw count

# Filter Values  
  
# Recode main college type and subcategory  
combined\_df\_na\_cleaned <- combined\_df\_na %>%  
 mutate(  
 college\_type = case\_when(  
 grepl("Baccalaureate Colleges", classific\_2021) ~ "Baccalaureate",  
 grepl("Master's Colleges & Universities", classific\_2021) ~ "Master's",  
 grepl("Doctoral Universities", classific\_2021) ~ "Doctoral",  
 grepl("Tribal Colleges and Universities", classific\_2021) ~ "Tribal",  
 grepl("Special Focus Four-Year", classific\_2021) ~ "Special Focus",  
 TRUE ~ NA\_character\_  
 ),  
 subcategory = case\_when(  
 grepl("Master's Colleges & Universities: Small Programs", classific\_2021) ~ "Master's: Small Programs",  
 grepl("Master's Colleges & Universities: Medium Programs", classific\_2021) ~ "Master's: Medium Programs",  
 grepl("Master's Colleges & Universities: Larger Programs", classific\_2021) ~ "Master's: Larger Programs",  
 grepl("Doctoral Universities: High Research Activity", classific\_2021) ~ "Doctoral: High Research Activity",  
 grepl("Doctoral Universities: Very High Research Activity", classific\_2021) ~ "Doctoral: Very High Research Activity",  
 grepl("Doctoral/Professional Universities", classific\_2021) ~ "Doctoral: Professional Universities",  
 grepl("Baccalaureate Colleges: Diverse Fields", classific\_2021) ~ "Baccalaureate: Diverse Fields",  
 grepl("Baccalaureate Colleges: Arts & Sciences Focus", classific\_2021) ~ "Baccalaureate: Arts & Sciences Focus",  
 grepl("Special Focus Four-Year: Research Institution", classific\_2021) ~ "Special Focus: Research Institution",  
 grepl("Special Focus Four-Year: Other Health Professions Schools", classific\_2021) ~ "Special Focus: Health Professions",  
 grepl("Special Focus Four-Year: Medical Schools & Centers", classific\_2021) ~ "Special Focus: Medical Schools",  
 grepl("Special Focus Four-Year: Engineering and Other Technology-Related Schools", classific\_2021) ~ "Special Focus: Engineering/Technology",  
 grepl("Special Focus Four-Year: Other Special Focus Institutions", classific\_2021) ~ "Special Focus: Other Institutions",  
 grepl("Tribal Colleges and Universities", classific\_2021) ~ "Tribal Colleges",  
 TRUE ~ NA\_character\_  
 )  
 ) %>%  
 filter(!is.na(college\_type) & !is.na(subcategory))  
  
  
# Color Palette   
  
color\_palette\_shaded <- c(  
 # GRAYS for Baccalaureate  
 "Baccalaureate: Diverse Fields" = "#B0B0B0", # Light gray  
 "Baccalaureate: Arts & Sciences Focus" = "#505050", # Dark gray  
  
 # ORANGES for Master's  
 "Master's: Small Programs" = "#F5C900", # Yellow-gold  
 "Master's: Medium Programs" = "#FF7F00", # Standard orange  
 "Master's: Larger Programs" = "#D65F00", # Dark orange  
  
 # BLUES for Doctoral  
 "Doctoral: High Research Activity" = "#5C8F99", # Soft teal-blue  
 "Doctoral: Very High Research Activity" = "#0072B2", # Blue (Color Universal Design)  
 "Doctoral: Professional Universities" = "#003366", # Navy  
  
 # PURPLES for Special Focus  
 "Special Focus: Research Institution" = "#CAB2D6", # Lavender  
 "Special Focus: Health Professions" = "#9E83C9", # Medium purple  
 "Special Focus: Medical Schools" = "#6A51A3", # Royal purple  
 "Special Focus: Engineering/Technology" = "#54278F",# Deep violet  
 "Special Focus: Other Institutions" = "#3F007D", # Very dark purple  
  
 # BROWN for Tribal Colleges  
 "Tribal Colleges" = "#8B4513" # Saddle brown  
)  
  
# Set factored levels  
  
# Set factor levels for custom order of college\_type  
combined\_df\_na\_cleaned$college\_type <- factor(  
 combined\_df\_na\_cleaned$college\_type,  
 levels = c("Baccalaureate", "Master's", "Doctoral", "Special Focus", "Tribal")  
)  
  
# Reverse the factor levels for subcategory BEFORE plotting  
combined\_df\_na\_cleaned$subcategory <- factor(  
 combined\_df\_na\_cleaned$subcategory,  
 levels = rev(c(  
 "Baccalaureate: Diverse Fields",  
 "Baccalaureate: Arts & Sciences Focus",  
 "Master's: Small Programs",  
 "Master's: Medium Programs",  
 "Master's: Larger Programs",  
 "Doctoral: High Research Activity",  
 "Doctoral: Very High Research Activity",  
 "Doctoral: Professional Universities",  
 "Special Focus: Research Institution",  
 "Special Focus: Health Professions",  
 "Special Focus: Medical Schools",  
 "Special Focus: Engineering/Technology",  
 "Special Focus: Other Institutions",  
 "Tribal Colleges"  
 ))  
)  
  
# Plot  
  
count\_collegetype\_subcat <- ggplot(combined\_df\_na\_cleaned, aes(x = college\_type, fill = subcategory)) +   
 geom\_bar(position = "stack") +  
 scale\_fill\_manual(values = color\_palette\_shaded) +  
 labs(  
 title = "Distribution of U.S. institutions by Carnegie Classification (2021): Raw count",  
 subtitle = "Breakdown by college type, degree focus, and program size",  
 x = "",  
 y = "Count",  
 caption = "College Type: Determined by majority degrees conferred at the institution\nNote: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 guides(fill = guide\_legend(ncol = 1, title = "")) +  
 my\_plot\_theme +  
 coord\_flip()  
  
# Save the plot  
ggsave("count\_collegetype\_subcat.png", plot = count\_collegetype\_subcat,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(count\_collegetype\_subcat)

### 1b. Distribution of U.S. institutions by Carnegie Classification (2021): Proportion

perct\_collegetype\_subcat <- ggplot(combined\_df\_na\_cleaned, aes(y = college\_type, fill = subcategory)) +   
 geom\_bar(position = "fill") + # Normalize bars to 100% within each group  
 scale\_fill\_manual(values = color\_palette\_shaded) + # Your color-safe palette  
 scale\_x\_continuous(labels = scales::percent\_format(accuracy = 1))+ # Show x-axis as %  
 labs(  
 title = "Distribution of U.S. institutions by Carnegie Classification (2021): Proportion",  
 subtitle = "Percent of institutions within each classification category",  
 x = "Percentage (%)",  
 y = "",  
 caption = "College Type: Determined by majority degrees conferred at the institution\nNote: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)",  
 ) +  
 my\_plot\_theme + # Apply your custom theme  
 theme(  
 axis.text.y = element\_text(angle = 0),  
 axis.text.x = element\_text(angle = 0, hjust = 1)  
 ) +  
 guides(fill = guide\_legend(ncol = 1, title = ""))  
  
  
# Save the plot  
ggsave("perct\_collegetype\_subcat.png", plot = perct\_collegetype\_subcat,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(perct\_collegetype\_subcat)

## 2. Distribution of U.S. institutions by research activity designation (2025)

### By Lily Gates

This graph visualizes the raw count and relative distribution of institutions by their research activity designation in 2025. The categories displayed include “Research Colleges and Universities”, “Research 1”, and “Research 2”. Data excludes institutions without a research designation.

The graph shows the both the raw count and relative distribution of institutions across three research activity categories: Research Colleges and Universities (RCA), Research 1 (R1), and Research 2 (R2). - RCA institutions spend at least $2.5 million on research annually but do not meet the criteria for R1 or R2. - R1 institutions spend at least $50 million on research and produce at least 70 research doctorates annually - R2 institutions spend at least $5 million and produce at least 20 research doctorates.

In general, there is a similar distribution to the raw count and overall percentage distribution between R1, R2, and RCU institutions. Interestingly, although there are more RCU institutions, R1 and R2 institutions confer the most doctoral degrees.

# Prepare the data and drop NA  
plot\_data <- combined\_df\_na %>%  
 filter(!is.na(research\_tier\_2025)) %>%  
 mutate(  
 research\_tier\_2025 = factor(  
 research\_tier\_2025,  
 levels = c(  
 "Research 1: Very High Research Spending and Doctorate Production", # R1  
 "Research 2: High Research Spending and Doctorate Production", # R2  
 "Research Colleges and Universities" # RCPU  
 )  
 )  
 ) %>%  
 count(research\_tier\_2025) %>%  
 mutate(percentage = n / sum(n) \* 100)  
  
# Raw count bar plot  
raw\_count\_plot <- ggplot(plot\_data, aes(x = research\_tier\_2025, y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 labs(  
 x = "",  
 y = "Raw Count"  
 ) +  
 my\_plot\_theme + # Apply custom theme here  
 scale\_x\_discrete(labels = c(  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",   
 "Research 2: High Research Spending and Doctorate Production" = "R2",   
 "Research Colleges and Universities" = "RCU"  
 ))  
  
# Proportional bar plot (percentage)  
proportional\_plot <- ggplot(plot\_data, aes(x = research\_tier\_2025, y = percentage)) +  
 geom\_bar(stat = "identity", fill = "lightblue") +  
 labs(  
 x = "",  
 y = "Percentage (%)"  
 ) +  
 my\_plot\_theme +   
 scale\_x\_discrete(labels = c(  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",   
 "Research 2: High Research Spending and Doctorate Production" = "R2",   
 "Research Colleges and Universities" = "RCU"  
 ))  
  
# Combine the two plots side by side using patchwork  
school\_distrib\_combined\_plot <- raw\_count\_plot + proportional\_plot +   
 plot\_layout(guides = 'collect', widths = c(1, 1)) +   
 plot\_annotation(  
 title = "Distribution of U.S. institutions by research activity designation (2025)",  
 subtitle = "Raw count and relative proportion of R1, R2, and RCU institutions",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 theme(  
 plot.title = element\_text(face = "bold", size = 14), # Make title bold  
 plot.subtitle = element\_text(face = "italic", size = 12), # Make subtitle italic  
 plot.caption = element\_text(size = 10) # Optional, if you want to adjust the caption size  
 )  
  
# Save the plot  
ggsave("school\_distrib\_combined\_plot.png", plot = school\_distrib\_combined\_plot,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(school\_distrib\_combined\_plot)

## 3. Average doctoral degrees by research tier (2020–2023)

### By Lily Gates

This boxplot with beeswarm displays the distribution of average doctoral degrees awarded from 2020 to 2023 across three research activity designations: “Research Colleges and Universities,” “Research 1 (R1),” and “Research 2 (R2).” The x-axis represents the average number of doctoral degrees awarded, while the y-axis categorizes institutions by their research tier.

The plot highlights the range, median, and quartiles for each research tier. Research Colleges and Universities, which are considered research-focused but do not meet the rigorous criteria for R1 or R2 designation, show a lower average number of doctoral degrees compared to R1 and R2 institutions.

RCU institutions have a wide range of average doctoral degrees conferred, compared to the spread of R1 and R2 institutions. However, RCU institutions are overwhelmingly far less than the amount of doctoral degrees conferred compared to R2 and R1 institutions.

# Clean data, remove NAs and 0s, relabel tiers  
clean\_df <- combined\_df\_na %>%  
 filter(  
 !is.na(avg\_num\_doc\_degrees\_2020\_2023),  
 avg\_num\_doc\_degrees\_2020\_2023 > 0,  
 !is.na(research\_tier\_2025)  
 ) %>%  
 mutate(  
 research\_tier\_2025 = recode(  
 research\_tier\_2025,  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",  
 "Research 2: High Research Spending and Doctorate Production" = "R2",  
 "Research Colleges and Universities" = "RCU"  
 )  
 )  
  
# Create boxplot with beeswarm overlay  
boxplot\_avg\_degrees <- ggplot(clean\_df, aes(x = avg\_num\_doc\_degrees\_2020\_2023, y = research\_tier\_2025)) +  
 geom\_boxplot(fill = "slateblue", alpha = 0.5, outlier.shape = NA, size = 1.5) +  
 geom\_beeswarm(color = "darkslateblue", alpha = 0.8, size = 2.2, cex = 2) +  
 scale\_x\_log10() +  
 # Customize Appearance  
 labs(  
 title = "Average doctoral degrees by research tier (2020–2023)",  
 subtitle = "Distribution of average doctoral degrees conferred by RCU, R1, and R2 institutions",  
 x = "Average doctoral degrees (log scale)",  
 y = "",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 my\_plot\_theme  
  
#boxplot\_avg\_degrees  
  
# Save the plot  
ggsave("boxplot\_avg\_degrees.png", plot = boxplot\_avg\_degrees,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(boxplot\_avg\_degrees)

## 4. Doctoral degrees conferred by research tier and academic year (2020–2023)

### By Lily Gates

This stacked bar graph visualizes the count of doctoral degrees conferred across different academic years (2020–2023), broken down by research tier. The research tiers are represented by three categories: “Research Colleges and Universities” (RCU), “Research 1 (R1),” and “Research 2 (R2).” The bars for each academic year are stacked to show the distribution of doctoral degrees within each research tier, providing insight into how the number of degrees varies across tiers and over time. The y-axis represents the raw count of doctoral degrees, and the x-axis represents the academic years.

The most noticeable trend is the gradual increase in the number of doctoral degrees conferred by R1 institutions, especially between the academic years 2020–2021 and 2021–2022. This increase is likely attributed to the effects of the COVID-19 pandemic, which may have led to shifts in research priorities and funding. There is also a slight increase in R2 degrees, but this change is less pronounced. RCU institutions, on the other hand, have such a small number of degrees to begin with that any change is difficult to discern.

unique(combined\_df\_na$research\_tier\_2025)

## [1] "Research Colleges and Universities"   
## [2] "Research 1: Very High Research Spending and Doctorate Production"  
## [3] "Research 2: High Research Spending and Doctorate Production"   
## [4] NA

doctoral\_degrees\_by\_year <- combined\_df\_na %>%  
 select(research\_tier\_2025, num\_doc\_degrees\_2020\_2021, num\_doc\_degrees\_2021\_2022, num\_doc\_degrees\_2022\_2023) %>%  
 pivot\_longer(  
 cols = starts\_with("num\_doc\_degrees"),  
 names\_to = "academic\_year",  
 values\_to = "num\_doc\_degrees"  
 ) %>%  
 filter(!is.na(num\_doc\_degrees), !is.na(research\_tier\_2025)) %>%  
 mutate(  
 academic\_year = recode(  
 academic\_year,  
 num\_doc\_degrees\_2020\_2021 = "2020–2021",  
 num\_doc\_degrees\_2021\_2022 = "2021–2022",  
 num\_doc\_degrees\_2022\_2023 = "2022–2023"  
 ),  
 academic\_year = factor(  
 academic\_year,  
 levels = c("2020–2021", "2021–2022", "2022–2023")  
 )  
 )  
  
stacked\_bar\_avg\_degrees <- ggplot(doctoral\_degrees\_by\_year, aes(x = academic\_year, y = num\_doc\_degrees, fill = research\_tier\_2025)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(  
 title = "Doctoral degrees conferred by research tier and academic year (2020–2023)",  
 subtitle = "Distribution of doctoral degrees awarded by R1, R2, and RCU institutions each year",  
 x = "Academic year",  
 y = "Total doctoral degree count",  
 fill = "",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 my\_plot\_theme + # Apply custom theme here  
 theme(  
 plot.title = element\_text(hjust = 0),  
 axis.text.x = element\_text(angle = 0, hjust = 0.5), # Center the x-axis labels  
 plot.caption = element\_text(hjust = 0, size = 9, color = "gray30"),  
 legend.position = "right", # Keep the legend on the right  
 legend.box = "vertical",  
 panel.grid.major = element\_blank()  
 ) +  
 guides(  
 fill = guide\_legend(ncol = 1) # Set the legend to 1 column  
 ) +  
 scale\_fill\_manual(  
 values = c("Research 1: Very High Research Spending and Doctorate Production" = "#4C79A1", "Research 2: High Research Spending and Doctorate Production" = "#F1A340", "Research Colleges and Universities" = "#6D9D4B"), # New colors  
 labels = c("R1", "R2", "RCU") # Set the legend labels  
 ) +  
 scale\_y\_continuous(  
 labels = scales::comma, # Adds commas to the y-axis labels for better readability  
 expand = c(0, 0) # Ensure y-axis starts at 0, with no padding  
 )  
  
# Save the plot  
ggsave("stacked\_bar\_avg\_degrees.png", plot = stacked\_bar\_avg\_degrees,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display the graph  
dev.new(width = 14, height = 9)  
print(stacked\_bar\_avg\_degrees)

## 5. Count of institutions by research activity designation and institution type (2025)

### By Lily Gates

This grouped bar graph displays the count of institutions by their type (public, private non-profit, and private for-profit) and their research activity designation (R1, R2, and Research Colleges and Universities) for the year 2025. They represent the same data, but the color-fill and x-axis variables are the opposite on both graphs.

The majority of institutions with any research activity designation in this dataset are public institutions, with far fewer private non-profit and private for-profit institutions in each research activity category. Notably, the Research Colleges and Universities category is the only one that contains private for-profit institutions, with only one school in this group.

In terms of distribution across research activity designations, Research Colleges and Universities appear to have a higher number of private non-profit institutions compared to R1 and R2 institutions. Specifically, Research Colleges and Universities are predominantly composed of private non-profit institutions, whereas R1 and R2 categories show a much stronger representation of public institutions.

Interestingly, the number of public institutions in the R1 category is almost equal to that of the Research Colleges and Universities category, which further suggests that Research Colleges and Universities have a more balanced composition of private non-profit institutions.

This analysis highlights key trends in how research activity designation relates to institutional type, with public institutions overwhelmingly dominating higher research tiers, and Research Colleges and Universities being the outlier with more private non-profit representation.

# Create plot with text labels on top of bars  
# Drop NA values from the data  
plot\_df\_clean <- combined\_df\_na\_cleaned %>%  
 count(public\_private\_profit, research\_tier\_2025, name = "n") %>%  
 na.omit() # Remove rows with NA values  
  
grouped\_bar\_school\_count <- ggplot(plot\_df\_clean, aes(x = public\_private\_profit, y = n, fill = research\_tier\_2025)) +  
 geom\_bar(position = "dodge", stat = "identity", width = 0.8) + # Adjust width for clarity  
 geom\_text(  
 aes(label = n), # Add labels with the count  
 position = position\_dodge(width = 0.8), # Align text with bars  
 vjust = -0.5, # Adjust vertical position of labels above bars  
 size = 6, # Set label text size  
 color = "black",  
 fontface = "bold"  
 ) +  
 labs(  
 title = "Count of institutions by research activity designation and institution type (2025)",  
 subtitle = "Distribution of institutions by type and research tier, distinguished by color",  
 x = "",  
 y = "Count of institutions",  
 fill = "",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 my\_plot\_theme +  
 scale\_x\_discrete(labels = c(  
 "Private for-profit" = "Private\n(for-profit)",  
 "Private not-for-profit" = "Private\n(non-profit)",  
 "Public" = "Public"  
 )) +  
 scale\_y\_continuous(  
 limits = c(0, max(plot\_df\_clean$n) \* 1.2), # Add space on y-axis for better visualization  
 expand = c(0, 0)  
 ) +  
 scale\_fill\_manual(  
 values = c(  
 "Research Colleges and Universities" = "#E69F00", # orange  
 "Research 1: Very High Research Spending and Doctorate Production" = "#56B4E9", # blue  
 "Research 2: High Research Spending and Doctorate Production" = "#009E73" # green  
 ),  
 labels = c(  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",  
 "Research 2: High Research Spending and Doctorate Production" = "R2",  
 "Research Colleges and Universities" = "RCU"  
 )  
 ) +  
 theme(  
 legend.position = "bottom",  
 legend.box = "horizontal",  
 legend.title = element\_text(size = 12)  
 )  
  
#grouped\_bar\_school\_count  
  
# Save the plot  
ggsave("grouped\_bar\_school\_count.png", plot = grouped\_bar\_school\_count,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(grouped\_bar\_school\_count)

## 6. Higher Education Research and Development (HERD) spending by fiscal year and research tier

### By Lily Gates

This density plot displays Higher Education Research and Development (HERD) spending across three types of institutions for fiscal years 2021, 2022, and 2023:

The log scale on the x-axis highlights differences in spending, particularly at the higher end. The overlap between R2 and the other categories suggests that some R2 institutions have spending comparable to R1 or R Colleges.

combined\_df\_long <- combined\_df\_na\_cleaned %>%  
 select(research\_tier\_2025, herd\_fy21, herd\_fy22, herd\_fy23) %>%  
 mutate(across(starts\_with("herd\_fy"), ~ as.numeric(gsub(",", "", .)))) %>% # Convert all HERD columns to numeric  
 pivot\_longer(  
 cols = starts\_with("herd\_fy"),  
 names\_to = "fiscal\_year",  
 values\_to = "herd\_spending"  
 ) %>%  
 filter(!is.na(herd\_spending), !is.na(research\_tier\_2025)) %>%  
 mutate(  
 fiscal\_year = recode(  
 fiscal\_year,  
 herd\_fy21 = "FY 2021",  
 herd\_fy22 = "FY 2022",  
 herd\_fy23 = "FY 2023"  
 ),  
 research\_tier\_simple = recode(  
 research\_tier\_2025,  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",  
 "Research 2: High Research Spending and Doctorate Production" = "R2",  
 "Research Colleges and Universities" = "RCU"  
 )  
 ) %>%  
 filter(herd\_spending > 0)

## Warning: There were 2 warnings in `mutate()`.  
## The first warning was:  
## ℹ In argument: `across(starts\_with("herd\_fy"), ~as.numeric(gsub(",", "", .)))`.  
## Caused by warning:  
## ! NAs introduced by coercion  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 1 remaining warning.

# Hardcoded color-blind friendly colors  
color\_mapping <- c(  
 "R1" = "#E69F00", # orange  
 "R2" = "#56B4E9", # blue  
 "RCU" = "#009E73" # green  
)  
  
# Plot  
herd\_density <- ggplot(combined\_df\_long, aes(x = herd\_spending, color = research\_tier\_simple, fill = research\_tier\_simple)) +  
 geom\_density(alpha = 0.3, size = .8) +  
 scale\_x\_log10(labels = dollar\_format(scale = 1e-9, suffix = "B")) +  
 scale\_y\_continuous(expand = c(0, 0)) +  
 facet\_wrap(~ fiscal\_year, nrow = 1) +  
 scale\_color\_manual(values = color\_mapping) +  
 scale\_fill\_manual(values = color\_mapping) +  
 labs(  
 title = "Higher Education Research and Development (HERD) spending by fiscal year and research tier",  
 x = "HERD Spending (Log Scale, in Billions)",  
 subtitle = "Density plot with log scale highlights skewed distribution and overlaps across tiers",  
 y = "Density",  
 color = "",  
 fill = "",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 my\_plot\_theme +  
 theme(  
 legend.position = "bottom",  
 legend.direction = "horizontal",  
 panel.grid.major = element\_blank()  
 )

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

#herd\_density  
  
# Save Graph  
ggsave("herd\_density.png", plot = herd\_density,   
 width = 14, height = 9, dpi = 300, units = "in")  
  
# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(herd\_density)

## 7. Distribution of Higher Education Research and Development (HERD) spending

### By Colin Thompson and Lily Gates

This histogram with a density curve displays the distribution of Higher Education Research and Development (HERD) spending across three fiscal years: 2021, 2022, and 2023. It highlights the range of research spending institutions report, with the peak around 0.05 billion dollars in fiscal year 2023, suggesting that many institutions are clustered in this spending range. Over the years, spending has gradually increased, especially in FY2023, indicating a growing trend in research investment. By examining this distribution, we can identify the most common spending ranges, which can help provide insight into what is considered typical or desirable for research spending. The graph also reveals the highest amounts of spending, useful for understanding the research budgets of institutions with the largest research investments.

# Prepare the data  
combined\_df\_long <- combined\_df\_na\_cleaned %>%  
 select(research\_tier\_2025, herd\_fy21, herd\_fy22, herd\_fy23) %>%  
 mutate(across(starts\_with("herd\_fy"), ~ as.numeric(gsub(",", "", .)))) %>% # Remove commas and convert to numeric  
 pivot\_longer(  
 cols = starts\_with("herd\_fy"),  
 names\_to = "fiscal\_year",  
 values\_to = "herd\_spending"  
 ) %>%  
 filter(!is.na(herd\_spending)) %>%  
 mutate(  
 fiscal\_year = recode(  
 fiscal\_year,  
 herd\_fy21 = "FY 2021",  
 herd\_fy22 = "FY 2022",  
 herd\_fy23 = "FY 2023"  
 )  
 )

## Warning: There were 2 warnings in `mutate()`.  
## The first warning was:  
## ℹ In argument: `across(starts\_with("herd\_fy"), ~as.numeric(gsub(",", "", .)))`.  
## Caused by warning:  
## ! NAs introduced by coercion  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 1 remaining warning.

combined\_df\_long\_cleaned <- combined\_df\_long %>%  
 filter(herd\_spending > 0) # Remove rows with non-positive spending  
  
# Initial plot (used to calculate max density for buffer)  
herd\_hist\_base <- ggplot(combined\_df\_long\_cleaned, aes(x = herd\_spending)) +  
 geom\_histogram(aes(y = ..density..), bins = 20, color = "black", fill = "steelblue", alpha = 0.5, size = 0.6) +  
 geom\_density(color = "indianred", fill = "indianred", alpha = .1, size = 1.6, linetype = "solid") +  
 scale\_x\_log10(labels = dollar\_format(scale = 1e-9, suffix = "B")) +  
 facet\_wrap(~ fiscal\_year, nrow = 1, ncol = 3)  
  
# Calculate max y value from density  
gg\_data <- ggplot\_build(herd\_hist\_base)

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

max\_density <- max(gg\_data$data[[2]]$y, na.rm = TRUE)  
buffered\_max <- max\_density \* 1.1 # Add 10% buffer  
  
# Final plot with y-axis buffer  
herd\_hist <- ggplot(combined\_df\_long\_cleaned, aes(x = herd\_spending)) +  
 geom\_histogram(aes(y = ..density..), bins = 20, color = "black", fill = "steelblue", alpha = 0.5, size = 0.6) +  
 geom\_density(color = "indianred", fill = "indianred", alpha = .1, size = 1.6, linetype = "solid") +  
 scale\_x\_log10(labels = dollar\_format(scale = 1e-9, suffix = "B")) +  
 scale\_y\_continuous(expand = c(0, 0), limits = c(0, buffered\_max)) +  
 facet\_wrap(~ fiscal\_year, nrow = 1, ncol = 3) + # Arrange FY21, FY22, FY23 side by side  
 labs(  
 title = "Distribution of Higher Education Research and Development (HERD) spending",  
 subtitle = "Displays the range and frequency of reported HERD spending across fiscal years 2021 to 2023",   
 x = "HERD Spending (Log Scale, in Billions)",  
 y = "Density",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)"  
 ) +  
 my\_plot\_theme +  
 theme(  
 legend.position = "none",  
 panel.grid.major = element\_blank()  
 )  
  
# Save Graph  
ggsave("herd\_hist.png", plot = herd\_hist,   
 width = 14, height = 9, dpi = 300, units = "in")

## Warning: Removed 4 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).

# Open a new graphics window to display graph  
dev.new(width = 14, height = 9)  
print(herd\_hist)

## Warning: Removed 4 rows containing missing values or values outside the scale range  
## (`geom\_bar()`).

## 8. Positive correlation between average research spending and doctoral degrees (2020-2023)

### By Mia Leandri

This scatter plot visualizes the relationship between average research spending (from FY2021 to FY2023) and the average number of doctoral degrees awarded (2020–2023) across American higher education institutions.

The x-axis shows average research spending in U.S. dollars, while the y-axis indicates the average number of doctoral degrees conferred. Institutions with greater research spending tend to award more doctoral degrees, illustrating a positive correlation between institutional investment in research and doctoral productivity.

This visualization supports the analytical research question by highlighting how non-financial institutional characteristics (like research tier) correlate with tangible research outcomes such as doctoral degree production.

Note: Equation is in a transformed log form to correspond with log-log transformation

# Create valid\_data by filtering complete cases and recoding research tier  
# So it is able to be log transformed  
valid\_data <- combined\_df\_na %>%  
 filter(  
 !is.na(herd\_avg\_fy21\_to\_fy23),  
 !is.na(avg\_num\_doc\_degrees\_2020\_2023),  
 !is.na(research\_tier\_2025),  
 herd\_avg\_fy21\_to\_fy23 > 0, # Filter out zero or negative values  
 avg\_num\_doc\_degrees\_2020\_2023 > 0 # Ensure the dependent variable is positive before log transform  
 ) %>%  
 mutate(  
 research\_tier\_2025 = recode(  
 research\_tier\_2025,  
 "Research Colleges and Universities" = "RCU",  
 "Research 1: Very High Research Spending and Doctorate Production" = "R1",  
 "Research 2: High Research Spending and Doctorate Production" = "R2"  
 ),  
 research\_tier\_2025 = factor(research\_tier\_2025, levels = c("RCU", "R2", "R1"))  
 )  
  
# Check if there are any remaining NA or Infinite values  
summary(valid\_data$herd\_avg\_fy21\_to\_fy23)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.069e+06 1.289e+07 5.308e+07 2.275e+08 2.778e+08 3.468e+09

summary(valid\_data$avg\_num\_doc\_degrees\_2020\_2023)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 21.0 71.0 152.2 209.0 924.0

any(is.na(valid\_data$herd\_avg\_fy21\_to\_fy23))

## [1] FALSE

any(is.na(valid\_data$avg\_num\_doc\_degrees\_2020\_2023))

## [1] FALSE

any(is.infinite(valid\_data$herd\_avg\_fy21\_to\_fy23))

## [1] FALSE

any(is.infinite(valid\_data$avg\_num\_doc\_degrees\_2020\_2023))

## [1] FALSE

# Fit linear model (log-log scale)  
lm\_model <- lm(  
 log10(avg\_num\_doc\_degrees\_2020\_2023) ~ log10(herd\_avg\_fy21\_to\_fy23),  
 data = valid\_data  
)  
  
# Extract slope, intercept, and R²  
slope <- coef(lm\_model)[2]  
intercept <- coef(lm\_model)[1]  
r\_squared <- summary(lm\_model)$r.squared  
  
# Get axis limits  
x\_limits <- range(valid\_data$herd\_avg\_fy21\_to\_fy23, na.rm = TRUE)  
y\_limits <- range(valid\_data$avg\_num\_doc\_degrees\_2020\_2023, na.rm = TRUE)  
x\_max <- max(valid\_data$herd\_avg\_fy21\_to\_fy23, na.rm = TRUE)  
y\_min <- min(valid\_data$avg\_num\_doc\_degrees\_2020\_2023, na.rm = TRUE)  
  
# Define custom colors  
custom\_colors <- c("RCU" = "#1f77b4", "R2" = "#2ca02c", "R1" = "#ff7f0e")  
  
# Plot  
scatter\_corr\_degrees\_spending <- ggplot(valid\_data, aes(  
 x = herd\_avg\_fy21\_to\_fy23,  
 y = avg\_num\_doc\_degrees\_2020\_2023,  
 color = research\_tier\_2025  
)) +  
 geom\_point(alpha = 0.6, size = 2) +  
 geom\_smooth(method = "lm", se = TRUE, color = "gray30", fill = "black", alpha = 0.3, size = 1) +  
 scale\_color\_manual(  
 values = custom\_colors,  
 breaks = c("RCU", "R2", "R1"),  
 labels = c("RCU", "R2", "R1"),  
 guide = guide\_legend(override.aes = list(size = 8))  
 ) +  
 scale\_x\_log10(  
 limits = x\_limits,  
 breaks = log\_breaks(),  
 labels = label\_dollar(scale = 1e-9, suffix = "B")  
 ) +  
 scale\_y\_log10(  
 limits = y\_limits,  
 breaks = log\_breaks()  
 ) +  
 labs(  
 title = "Positive Correlation Between Research Spending and Doctoral Degrees",  
 subtitle = "Comparing average spending from 2020 to 2023",  
 caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1 or R2 criteria for expenses and/or doctoral degrees conferred.\nData source: Carnegie Classification of Institutions of Higher Education (https://carnegieclassifications.acenet.edu)",  
 x = "Avg. Research Spending (Log Scale, Billions USD)",  
 y = "Avg. Number of Doctoral Degrees (Log Scale)",  
 color = ""  
 ) +  
 my\_plot\_theme +  
 theme(  
 legend.position = "bottom",  
 legend.justification = "center",  
 legend.box = "horizontal",  
 legend.margin = margin(t = 5),  
 plot.margin = margin(t = 10, r = 10, b = 10, l = 10)  
 ) +  
 annotate(  
 "label",  
 x = x\_max,  
 y = y\_min,  
 label = paste0(  
 "y = 10^", round(intercept, 2), " \* x^", round(slope, 2),  
 "\nR² = ", round(r\_squared, 2)  
 ),  
 color = "black",  
 fill = alpha("white", 0.3),  
 size = 6,  
 hjust = 1,  
 vjust = 0,  
 fontface = "bold",  
 label.size = 1,  
 label.padding = unit(0.5, "lines"),  
 label.r = unit(0.4, "lines")  
 )  
  
# Save the plot  
ggsave("scatter\_corr\_degrees\_spending.png", plot = scatter\_corr\_degrees\_spending,  
 width = 14, height = 9, dpi = 300, units = "in")

## `geom\_smooth()` using formula = 'y ~ x'

# Display the plot  
dev.new(width = 14, height = 9)  
print(scatter\_corr\_degrees\_spending)

## `geom\_smooth()` using formula = 'y ~ x'