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)
```

```r
# 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
```

```r
# Importing Libraries

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(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
```

```r
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 er

# 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)
```

```r
# 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
```

```r
# Cleaning Data -- Fill any missing values with "NA"

# Function to replace NULL with NA
```

```r
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)
```

```r
# 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
# - 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 Prog
      grepl("Master's Colleges & Universities: Medium Programs", classific_2021) ~ "Master's: Medium Pr
      grepl("Master's Colleges & Universities: Larger Programs", classific_2021) ~ "Master's: Larger Pr
      grepl("Doctoral Universities: High Research Activity", classific_2021) ~ "Doctoral: High Research
      grepl("Doctoral Universities: Very High Research Activity", classific_2021) ~ "Doctoral: Very Hig
      grepl("Doctoral/Professional Universities", classific_2021) ~ "Doctoral: Professional Universitie
      grepl("Baccalaureate Colleges: Diverse Fields", classific_2021) ~ "Baccalaureate: Diverse Fields"
      grepl("Baccalaureate Colleges: Arts & Sciences Focus", classific_2021) ~ "Baccalaureate: Arts & S
      grepl("Special Focus Four-Year: Research Institution", classific_2021) ~ "Special Focus: Research
      grepl("Special Focus Four-Year: Other Health Professions Schools", classific_2021) ~ "Special Foc
      grepl("Special Focus Four-Year: Medical Schools & Centers", classific_2021) ~ "Special Focus: Med
      grepl("Special Focus Four-Year: Engineering and Other Technology-Related Schools", classific_2021
      grepl("Special Focus Four-Year: Other Special Focus Institutions", classific_2021) ~ "Special Foc
      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: 'Researc
  ) +
  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: 'Researc
  ) +
  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
```

```r
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
  ) +
  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.

```r
# 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"
    )
  )
```

```r
# 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
  ) +
  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.

```r
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
```

```r
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
  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,
  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
  ) +
  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", "Researcl
    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)
```

```r
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.

```r
# 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_
  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
  ) +
  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
  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 = res
  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
  ) +
  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 conve
  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
  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
  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 2
    x = "HERD Spending (Log Scale, in Billions)",
    y = "Density",
    caption = "Note: 'Research Colleges and Universities' (RCU) are research-focused but do not meet R1
  ) +
  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()`).
```

```r
# 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

```r
# 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
```

```r
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
```

```r
any(is.na(valid_data$herd_avg_fy21_to_fy23))
```

```
## [1] FALSE
```

```r
any(is.na(valid_data$avg_num_doc_degrees_2020_2023))
```

```
## [1] FALSE
```

```r
any(is.infinite(valid_data$herd_avg_fy21_to_fy23))
```

```
## [1] FALSE
```

```r
any(is.infinite(valid_data$avg_num_doc_degrees_2020_2023))
```

```
## [1] FALSE
```

```r
# 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
    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'
```