Classifying Retirement Plan Financial Adequacy Using Form 5500 Data

Exploratory Data Analysis and Feature Engineering

Donnie Minnick, Statistical Learning - Fall A 2025

September 26, 2025

This analysis leverages the following R packages: dplyr, lubridate, ’stringr and tidyr for data manipulation, knitr for report formatting, and ggplot2 and ggridges for visualization.

Load Form 5500 data.

plans <- readRDS("../data/plans.rds")

# Missing Data

Generate a report to check for missing values in the data at both the column and row level.

# Create column level summary.  
  
col\_missing <- plans %>%  
 summarise(across(everything(), ~ sum(is.na(.)))) %>%  
 pivot\_longer(cols = everything(),  
 names\_to = "Variable",  
 values\_to = "Missing\_Count") %>%  
 mutate(Missing\_Percent = Missing\_Count / nrow(plans) \* 100)  
  
# Create row level summary.  
  
row\_missing <- tibble(Variable = "Rows Any Missing",  
 Missing\_Count = sum(!complete.cases(plans)),  
 Missing\_Percent = sum(!complete.cases(plans)) / nrow(plans) \* 100)  
  
# Combine summaries.  
  
missing\_report <- bind\_rows(col\_missing, row\_missing)  
  
kable(missing\_report,  
 col.names = c("Variable", "Missing Count", "Missing Percent"),  
 caption = "Missing Data",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Missing Data

| Variable | Missing Count | Missing Percent |
| --- | --- | --- |
| ACK\_ID | 0 | 0 |
| PLAN\_YEAR\_BEGIN\_DATE | 0 | 0 |
| PLAN\_YEAR\_END\_DATE | 0 | 0 |
| PLAN\_NAME | 0 | 0 |
| PLAN\_EFFECTIVE\_DATE | 0 | 0 |
| PLAN\_TYPE | 0 | 0 |
| SPONSOR\_NAME | 0 | 0 |
| SPONSOR\_STATE | 0 | 0 |
| SPONSOR\_EIN | 0 | 0 |
| BUSINESS\_CODE | 0 | 0 |
| TOTAL\_ACTIVE\_PARTCP\_BOY | 0 | 0 |
| TOTAL\_ACTIVE\_PARTCP\_EOY | 0 | 0 |
| TOTAL\_ACCBAL\_PARTCP\_BOY | 0 | 0 |
| TOTAL\_ACCBAL\_PARTCP\_EOY | 0 | 0 |
| TOTAL\_CONTRIB\_PARTCP\_BOY | 0 | 0 |
| TOTAL\_CONTRIB\_PARTCP\_EOY | 0 | 0 |
| TOTAL\_CONTRIB\_EMPLR\_BOY | 0 | 0 |
| TOTAL\_CONTRIB\_EMPLR\_EOY | 0 | 0 |
| TOTAL\_LOANS\_BOY | 0 | 0 |
| TOTAL\_LOANS\_EOY | 0 | 0 |
| TOTAL\_ASSETS\_BOY | 0 | 0 |
| TOTAL\_ASSETS\_EOY | 0 | 0 |
| INDUSTRY\_TITLE | 0 | 0 |
| Rows Any Missing | 0 | 0 |

# Acknowledgement Id (ACK\_ID)

Acknowledgement Id is a unique identifier assigned by the Department of Labor (DOL) or IRS when a Form 5500 filing is successfully received and acknowledged. Think of it as a receipt number; it confirms that the filing was accepted into the system. It’s often used for tracking, auditing, or linking filings across systems.

# Plan Year Begin/End Dates (PLAN\_YEAR\_BEGIN\_DATE, PLAN\_YEAR\_END\_DATE)

These dates define the reporting window for all financial and participant metrics. Our dataset contains filings with a plan year begin date of January 1, 2023 and end date of December 31, 2023.

# Plan Effective Date (PLAN\_EFFECTIVE\_DATE)

Plan effective date marks the official inception of a retirement plan, i.e. the date it became active and enforceable under its governing provisions. This date anchors a plan’s life cycle, serving as a reference point for regulatory applicability, sponsor tenure, and cohort benchmarking.

From an analytic perspective, plan effective date enables several valuable transformations. Calculating plan age from this field allows for stratification by maturity, which is useful when assessing financial readiness across vintage cohorts. Plans initiated in different eras may reflect distinct design philosophies, contribution behaviors, or participant engagement patterns.

In benchmarking workflows, plan effective date can be paired with business code (BUSINESS\_CODE) to explore sector-specific adoption trends.

Generate a summary count of plans by plan effective year.

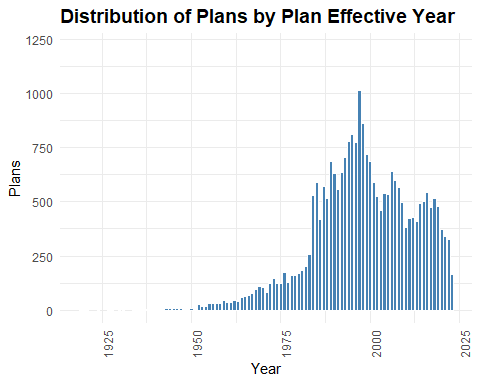
summary\_by\_plan\_effective\_year <- plans %>%  
 mutate(PLAN\_EFFECTIVE\_YEAR = as.character(year(PLAN\_EFFECTIVE\_DATE))) %>%  
 group\_by(PLAN\_EFFECTIVE\_YEAR) %>%  
 arrange(PLAN\_EFFECTIVE\_YEAR) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(summary\_by\_plan\_effective\_year,  
 col.names = c("Year", "Plans", "Percent"),  
 caption = "Summary By Plan Effective Year",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Summary By Plan Effective Year

| Year | Plans | Percent |
| --- | --- | --- |
| 1919 | 1 | 0.00 |
| 1920 | 4 | 0.00 |
| 1921 | 1 | 0.00 |
| 1924 | 1 | 0.00 |
| 1927 | 1 | 0.00 |
| 1928 | 1 | 0.00 |
| 1930 | 1 | 0.00 |
| 1931 | 1 | 0.00 |
| 1934 | 1 | 0.00 |
| 1935 | 1 | 0.00 |
| 1936 | 2 | 0.00 |
| 1937 | 5 | 0.00 |
| 1938 | 1 | 0.00 |
| 1939 | 3 | 0.00 |
| 1940 | 3 | 0.00 |
| 1941 | 1 | 0.00 |
| 1942 | 3 | 0.00 |
| 1943 | 9 | 0.00 |
| 1944 | 6 | 0.00 |
| 1945 | 9 | 0.00 |
| 1946 | 7 | 0.00 |
| 1947 | 7 | 0.00 |
| 1948 | 4 | 0.00 |
| 1949 | 2 | 0.00 |
| 1950 | 9 | 0.00 |
| 1951 | 5 | 0.00 |
| 1952 | 25 | 0.00 |
| 1953 | 17 | 0.00 |
| 1954 | 16 | 0.00 |
| 1955 | 31 | 0.00 |
| 1956 | 33 | 0.00 |
| 1957 | 31 | 0.00 |
| 1958 | 32 | 0.00 |
| 1959 | 44 | 0.00 |
| 1960 | 36 | 0.00 |
| 1961 | 36 | 0.00 |
| 1962 | 44 | 0.00 |
| 1963 | 41 | 0.00 |
| 1964 | 58 | 0.00 |
| 1965 | 63 | 0.00 |
| 1966 | 67 | 0.00 |
| 1967 | 77 | 0.00 |
| 1968 | 98 | 0.00 |
| 1969 | 108 | 0.00 |
| 1970 | 105 | 0.00 |
| 1971 | 81 | 0.00 |
| 1972 | 124 | 0.00 |
| 1973 | 148 | 0.01 |
| 1974 | 123 | 0.00 |
| 1975 | 122 | 0.00 |
| 1976 | 175 | 0.01 |
| 1977 | 126 | 0.00 |
| 1978 | 162 | 0.01 |
| 1979 | 161 | 0.01 |
| 1980 | 169 | 0.01 |
| 1981 | 184 | 0.01 |
| 1982 | 202 | 0.01 |
| 1983 | 258 | 0.01 |
| 1984 | 527 | 0.02 |
| 1985 | 591 | 0.02 |
| 1986 | 420 | 0.02 |
| 1987 | 570 | 0.02 |
| 1988 | 515 | 0.02 |
| 1989 | 688 | 0.03 |
| 1990 | 632 | 0.02 |
| 1991 | 558 | 0.02 |
| 1992 | 637 | 0.03 |
| 1993 | 704 | 0.03 |
| 1994 | 778 | 0.03 |
| 1995 | 809 | 0.03 |
| 1996 | 773 | 0.03 |
| 1997 | 1,012 | 0.04 |
| 1998 | 859 | 0.03 |
| 1999 | 717 | 0.03 |
| 2000 | 688 | 0.03 |
| 2001 | 587 | 0.02 |
| 2002 | 526 | 0.02 |
| 2003 | 460 | 0.02 |
| 2004 | 536 | 0.02 |
| 2005 | 532 | 0.02 |
| 2006 | 639 | 0.03 |
| 2007 | 597 | 0.02 |
| 2008 | 564 | 0.02 |
| 2009 | 495 | 0.02 |
| 2010 | 381 | 0.02 |
| 2011 | 423 | 0.02 |
| 2012 | 428 | 0.02 |
| 2013 | 410 | 0.02 |
| 2014 | 493 | 0.02 |
| 2015 | 502 | 0.02 |
| 2016 | 545 | 0.02 |
| 2017 | 475 | 0.02 |
| 2018 | 514 | 0.02 |
| 2019 | 478 | 0.02 |
| 2020 | 373 | 0.01 |
| 2021 | 339 | 0.01 |
| 2022 | 327 | 0.01 |
| 2023 | 167 | 0.01 |

Plot plans by plan effective year.

ggplot(summary\_by\_plan\_effective\_year, aes(x = as.integer(PLAN\_EFFECTIVE\_YEAR), y = entries)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 # geom\_text(aes(label = paste0(scales::comma(entries), "\n", round(percent \* 100, 0), "%")),  
 # vjust = -0.3, size = 3.0) +  
 labs(title = "Distribution of Plans by Plan Effective Year",  
 x = "Year",  
 y = "Plans") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(summary\_by\_plan\_effective\_year$entries) \* 1.2)



This distribution tells a compelling story about the evolution of retirement plan adoption in the U.S. The gradual rise from the 1940s through the 1970s reflects the early institutionalization of employer-sponsored plans, driven by post-war economic expansion and the formalization of pension structures.

The sharp uptick in the late 1970s through the 1990s coincides with major policy shifts, most notably the passage of ERISA in 1974 and the rise of defined contribution plans like 401k’s in the early 1980s. This era marks a surge in plan formation, especially among small and mid-sized employers responding to new regulatory clarity and tax incentives.

The peak in the late 1990s to early 2000s suggests a saturation point, after which the decline reflects consolidation of plans or sponsor exits, a shift toward pooled or multi-employer arrangements, and market volatility post-2008 influencing plan formation.

This distribution is useful for stratifying plans by vintage. Plans initiated in different eras may have distinct design philosophies, contribution behaviors, or participant engagement patterns. A plan vintage feature (e.g. pre-1980, 1980–1999, post-2000) can help assess adequacy or access across cohorts. Pairing this variable with business code would allow sector-specific diagnostics, e.g. whether certain industries adopted plans earlier or more consistently.

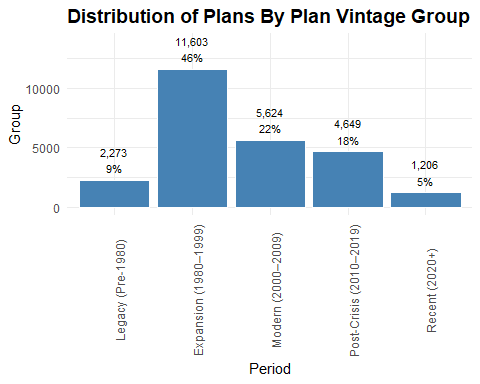
# Plan Vintage Group (PLAN\_VINTAGE\_GROUP)

Engineer a plan vintage group variable.

plans <- plans %>%  
 mutate(PLAN\_EFFECTIVE\_YEAR = year(PLAN\_EFFECTIVE\_DATE),  
 PLAN\_VINTAGE\_GROUP = case\_when(  
 PLAN\_EFFECTIVE\_YEAR < 1980 ~ "Legacy (Pre-1980)",  
 PLAN\_EFFECTIVE\_YEAR >= 1980 & PLAN\_EFFECTIVE\_YEAR <= 1999 ~ "Expansion (1980–1999)",  
 PLAN\_EFFECTIVE\_YEAR >= 2000 & PLAN\_EFFECTIVE\_YEAR <= 2009 ~ "Modern (2000–2009)",  
 PLAN\_EFFECTIVE\_YEAR >= 2010 & PLAN\_EFFECTIVE\_YEAR <= 2019 ~ "Post-Crisis (2010–2019)",  
 PLAN\_EFFECTIVE\_YEAR >= 2020 ~ "Recent (2020+)"),  
 PLAN\_VINTAGE\_GROUP = factor(PLAN\_VINTAGE\_GROUP,levels = c("Legacy (Pre-1980)",  
 "Expansion (1980–1999)",  
 "Modern (2000–2009)",  
 "Post-Crisis (2010–2019)",  
 "Recent (2020+)"),  
 ordered = TRUE))

Plot plans by plan vintage group.

summary\_by\_plan\_vintage\_group <- plans %>%  
 group\_by(PLAN\_VINTAGE\_GROUP) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
ggplot(summary\_by\_plan\_vintage\_group, aes(x = PLAN\_VINTAGE\_GROUP, y = entries)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 geom\_text(aes(label = paste0(scales::comma(entries), "\n", round(percent \* 100, 0), "%")),  
 vjust = -0.3, size = 3.0) +  
 labs(title = "Distribution of Plans By Plan Vintage Group",  
 x = "Period",  
 y = "Group") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(summary\_by\_plan\_vintage\_group$entries) \* 1.2)



Given this distribution, where plan counts rise sharply in the Expansion era and taper off in more recent vintages, we are dealing with a nonlinear, asymmetric pattern that’s not well captured by a simple linear term.

A spline would allow flexible modeling of nonlinear relationships without forcing global curvature. I can place knots at meaningful breakpoints, e.g. 1980, 2000, 2010, using plan effective year to reflect policy or behavioral shifts. This would be easy to explain in stakeholder terms: plan creation changes around key regulatory eras.

When performing logistic regression, splines allow the model to capture non-linearity by breaking the predictor into segments and fitting piece-wise polynomials.

Random forests don’t need splines to model non-linearity. They’re built to handle it through recursive splits. However, engineering spline-based features can make comparisons cleaner and diagnostics more stakeholder-friendly.

# Plan Type (PLAN\_TYPE)

In the landscape of employer-sponsored retirement plans, four distinct plan types often surface in nonprofit, education, and corporate sectors: 401(k), 403(b)(1), 403(b)(7), and 403(b)(9). Each reflects a unique structural and regulatory framework, shaping how participants accumulate retirement savings.

The 401(k) plan (code 2K) is the most prevalent defined contribution vehicle in the corporate world. It allows employees to defer a portion of their salary into tax-advantaged accounts, often with employer matching contributions. Its flexibility and widespread adoption make it a cornerstone of private-sector retirement readiness.

In contrast, the 403(b) family of plans caters to nonprofit and educational institutions.

The 403(b)(1) annuity contract (code 2R) is structured around insurance-based annuities, offering a predictable, contract-driven approach to retirement accumulation.

The 403(b)(7) custodial account (code 2S) channels contributions into mutual funds held in custodial arrangements, providing participants with market-based growth potential and investment choice.

Distinct from these is the 403(b)(9) retirement income account (code 2T), designed specifically for churches and church-affiliated organizations. It blends the regulatory protections of 403(b) plans with the unique governance and income structures of religious institutions, often emphasizing long-term stability over investment variety.

Together, these plan types form the backbone of sector-aware retirement analytics. Understanding their structural nuances is essential for benchmarking adequacy, surfacing disparities, and designing fairness overlays that reflect the lived realities of participants across industries.

This field is used to indicate the types of pension benefit features present in a plan. It’s a multi-valued field, meaning a single plan can report multiple codes concatenated into a single string.

Extract the relevant codes from the string and use them to create a new plan type name field.

plans <- plans %>%  
 mutate(PLAN\_TYPE\_NAME = case\_when(str\_detect(PLAN\_TYPE, "2K") ~ "401(k)",  
 str\_detect(PLAN\_TYPE, "2R") ~ "403(b)(1) Annuity",  
 str\_detect(PLAN\_TYPE, "2S") ~ "403(b)(7) Custodial",  
 str\_detect(PLAN\_TYPE, "2T") ~ "403(b)(9) Church"))

Generate summary by plan type name.

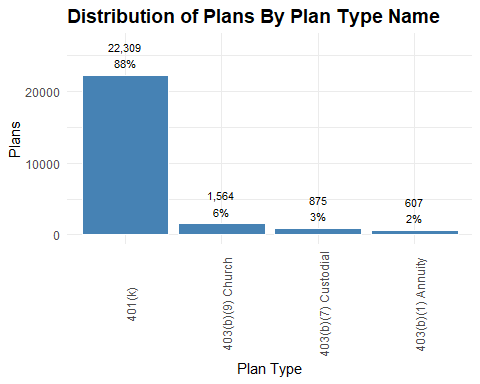
summary\_by\_plan\_type\_name <- plans %>%  
 group\_by(PLAN\_TYPE\_NAME) %>%  
 summarise(entries = n()) %>%  
 arrange(desc(entries)) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(summary\_by\_plan\_type\_name,  
 col.names = c("Plan Type", "Plans", "Percent"),  
 caption = "Summary By Plan Type Name",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Summary By Plan Type Name

| Plan Type | Plans | Percent |
| --- | --- | --- |
| 401(k) | 22,309 | 0.88 |
| 403(b)(9) Church | 1,564 | 0.06 |
| 403(b)(7) Custodial | 875 | 0.03 |
| 403(b)(1) Annuity | 607 | 0.02 |

Plot summary by plan type name.

ggplot(summary\_by\_plan\_type\_name, aes(x = forcats::fct\_reorder(PLAN\_TYPE\_NAME,   
 entries,   
 .desc = TRUE), y = entries)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 geom\_text(aes(label = paste0(scales::comma(entries), "\n", round(percent \* 100, 0), "%")),  
 vjust = -0.3, size = 3.0) +  
 labs(title = "Distribution of Plans By Plan Type Name",  
 x = "Plan Type",  
 y = "Plans") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(summary\_by\_plan\_type\_name$entries) \* 1.2)



This confirms the corporate sector’s pull in retirement plan design.

The remaining 12% is split across three structurally distinct 403(b) types.

# Sponsor State (SPONSOR\_STATE)

This variable refers to the U.S. state listed in the mailing address of the plan sponsor on Form 5500. It captures the geographic location of the sponsor’s administrative contact, not necessarily the location of plan participants or operations.

This field is useful for regional benchmarking and comparing plan characteristics or adequacy across states, regulatory overlays, such as aligning with state-specific compliance or fiduciary rules, and geographic stratification.

Summarize plans by sponsor state.

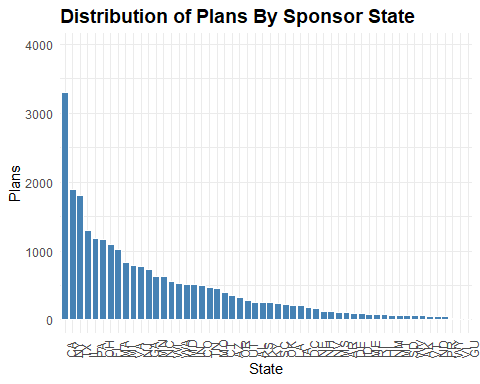
summary\_by\_sponsor\_state <- plans %>%  
 group\_by(SPONSOR\_STATE) %>%  
 summarise(entries = n()) %>%  
 arrange(desc(entries)) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(summary\_by\_sponsor\_state,  
 col.names = c("Sponsor State", "Plans", "Percent"),  
 caption = "Summary By Sponsor State",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Summary By Sponsor State

| Sponsor State | Plans | Percent |
| --- | --- | --- |
| CA | 3,301 | 0.13 |
| NY | 1,893 | 0.07 |
| TX | 1,808 | 0.07 |
| IL | 1,298 | 0.05 |
| PA | 1,175 | 0.05 |
| OH | 1,173 | 0.05 |
| FL | 1,090 | 0.04 |
| MA | 1,017 | 0.04 |
| MI | 830 | 0.03 |
| VA | 792 | 0.03 |
| NJ | 769 | 0.03 |
| GA | 727 | 0.03 |
| MN | 634 | 0.03 |
| NC | 629 | 0.02 |
| WI | 562 | 0.02 |
| WA | 522 | 0.02 |
| MD | 516 | 0.02 |
| IN | 515 | 0.02 |
| CO | 494 | 0.02 |
| TN | 463 | 0.02 |
| MO | 460 | 0.02 |
| CT | 399 | 0.02 |
| AZ | 351 | 0.01 |
| OR | 331 | 0.01 |
| UT | 274 | 0.01 |
| AL | 256 | 0.01 |
| KS | 255 | 0.01 |
| KY | 249 | 0.01 |
| SC | 231 | 0.01 |
| OK | 219 | 0.01 |
| LA | 213 | 0.01 |
| IA | 209 | 0.01 |
| DC | 186 | 0.01 |
| NE | 163 | 0.01 |
| NH | 123 | 0.00 |
| NV | 121 | 0.00 |
| MS | 103 | 0.00 |
| AR | 102 | 0.00 |
| DE | 93 | 0.00 |
| ID | 93 | 0.00 |
| ME | 83 | 0.00 |
| RI | 81 | 0.00 |
| HI | 73 | 0.00 |
| NM | 70 | 0.00 |
| MT | 63 | 0.00 |
| SD | 60 | 0.00 |
| WV | 58 | 0.00 |
| AK | 57 | 0.00 |
| VT | 53 | 0.00 |
| ND | 51 | 0.00 |
| PR | 45 | 0.00 |
| WY | 19 | 0.00 |
| VI | 2 | 0.00 |
| GU | 1 | 0.00 |

Plot plans by state.

ggplot(summary\_by\_sponsor\_state, aes(x = forcats::fct\_reorder(SPONSOR\_STATE,   
 entries,   
 .desc = TRUE), y = entries)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 labs(title = "Distribution of Plans By Sponsor State",  
 x = "State",  
 y = "Plans") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(summary\_by\_sponsor\_state$entries) \* 1.2)



This distribution aligns closely with population density and economic concentration. States like California, Texas, New York, and Florida tend to dominate in plan counts because they host a large number of employers, especially in sectors with high plan adoption rates (e.g. finance, tech, healthcare).

High counts also reflect regional hubs of plan sponsors, especially in metro areas with strong retirement plan infrastructure. States with proactive retirement initiatives, e.g. CalSavers in California, may show elevated plan formation or reporting.

I can potentially combine sponsor state with business code to see which industries drive plan formation in each region.

# Business Code and Industry Title (BUSINESS\_CODE, INDUSTRY\_TITLE)

Business code refers to the six-digit code used to classify the primary business activity of the plan sponsor, based on the North American Industry Classification System (NAICS). It identifies the economic sector in which the sponsoring organization operates—such as manufacturing, healthcare, finance, or education. Industry title is the human-readable label associated with that code.

summary\_by\_industry\_title <- plans %>%  
 group\_by(INDUSTRY\_TITLE) %>%  
 summarise(entries = n()) %>%  
 arrange(desc(entries)) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(head(summary\_by\_industry\_title, 20),  
 col.names = c("Industry", "Plans", "Percent"),  
 caption = "Summary By Industry Title - Top 20 Industries",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Summary By Industry Title - Top 20 Industries

| Industry | Plans | Percent |
| --- | --- | --- |
| All other professional, scientific, and technical services | 872 | 0.03 |
| Offices of physicians (except mental health specialists) | 702 | 0.03 |
| New car dealers | 559 | 0.02 |
| Hospitals | 505 | 0.02 |
| Educational services | 504 | 0.02 |
| Religious, grantmaking, civic, professional, and similar organizations | 496 | 0.02 |
| Engineering services | 490 | 0.02 |
| Management, scientific, and technical consulting services | 486 | 0.02 |
| Offices of lawyers | 480 | 0.02 |
| Other computer related services | 464 | 0.02 |
| Other miscellaneous manufacturing | 429 | 0.02 |
| Nursing and residential care facilities | 425 | 0.02 |
| Other fabricated metal product manufacturing | 404 | 0.02 |
| Individual and family services | 400 | 0.02 |
| Commercial banking | 376 | 0.01 |
| Custom computer programming services | 330 | 0.01 |
| Plastics product manufacturing | 273 | 0.01 |
| Nonresidential building construction | 272 | 0.01 |
| Offices of other holding companies | 266 | 0.01 |
| Scientific research and development services | 262 | 0.01 |

Representation at the detailed industry title level is sparse. Most individual NAICS-derived titles contribute less than 0.03% of total plan filings, which suggests that modeling or benchmarking at this granularity may be noisy or unstable. A few industries dominate plan filings, while most contribute marginally. The detailed industry title field has many levels, but low support per level. Using these titles directly in models could lead to poor generalization unless grouped.

# Sector Code and Sector Title (SECTOR\_CODE, SECTOR\_TITLE)

Collapse to broader categories using existing business code structure and group industries into sectors.

business\_codes <- readRDS("../data/business\_codes.rds")  
  
business\_codes <- business\_codes %>%  
 rename(SECTOR\_CODE = BUSINESS\_CODE,  
 SECTOR\_TITLE = INDUSTRY\_TITLE)  
  
plans <- plans %>%  
 mutate(SECTOR\_CODE = as.integer(substr(BUSINESS\_CODE, 1, 2)) \* 10000)  
  
plans <- left\_join(plans, business\_codes, by = "SECTOR\_CODE")

Summarize plans by sector.

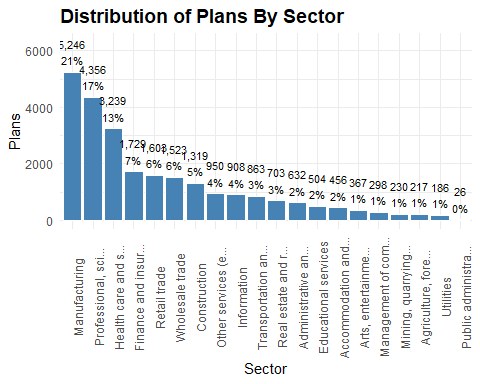
summary\_by\_sector\_title <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(entries = n()) %>%  
 arrange(desc(entries)) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(summary\_by\_sector\_title,  
 col.names = c("Sector Title", "Plans", "Percent"),  
 caption = "Summary By Sector Title",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Summary By Sector Title

| Sector Title | Plans | Percent |
| --- | --- | --- |
| Manufacturing | 5,246 | 0.21 |
| Professional, scientific, and technical services | 4,356 | 0.17 |
| Health care and social assistance | 3,239 | 0.13 |
| Finance and insurance | 1,729 | 0.07 |
| Retail trade | 1,603 | 0.06 |
| Wholesale trade | 1,523 | 0.06 |
| Construction | 1,319 | 0.05 |
| Other services (except public administration) | 950 | 0.04 |
| Information | 908 | 0.04 |
| Transportation and warehousing | 863 | 0.03 |
| Real estate and rental and leasing | 703 | 0.03 |
| Administrative and support and waste management and remediation services | 632 | 0.02 |
| Educational services | 504 | 0.02 |
| Accommodation and food services | 456 | 0.02 |
| Arts, entertainment, and recreation | 367 | 0.01 |
| Management of companies and enterprises | 298 | 0.01 |
| Mining, quarrying, and oil and gas extraction | 230 | 0.01 |
| Agriculture, forestry, fishing and hunting | 217 | 0.01 |
| Utilities | 186 | 0.01 |
| Public administration | 26 | 0.00 |

Plot by sector title.

ggplot(summary\_by\_sector\_title, aes(x = forcats::fct\_reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20),   
 entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 geom\_text(aes(label = paste0(scales::comma(entries), "\n", round(percent \* 100, 0), "%")),  
 vjust = -0.3, size = 3.0) +  
 labs(title = "Distribution of Plans By Sector",  
 x = "Sector",  
 y = "Plans") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(summary\_by\_sector\_title$entries) \* 1.2)



This plot reveals an imbalance in plan availability across sectors. The top few, e.g Manufacturing, Professional Services, and Healthcare, dominate the landscape, accounting for over half of all plans.

Saturated sectors like Manufacturing and Professional Services are likely have legacy plan structures and strong compliance cultures.

Healthcare shows high plan counts, possibly due to large institutional employers and unionized workforces.

Under-served sectors, such as Accommodation and Food Services, Arts and Entertainment, and Agriculture show very low plan counts despite employing millions. These sectors often have high turnover, part-time labor, or small businesses, all factors that historically correlate with low plan sponsorship.

In terms of predictors, I can leverage this insight by engineering two features: a plan penetration rate of plans per 1,000 employers or per 10,000 workers in each sector, and a sector adequacy flag that compares assets-per-participant or contribution ratios across sectors. The first captures structural access to retirement plans across sectors. I expect it will be a strong signal for access disparity and potential classification into “adequate” versus “inadequate” groups.

Binary or ordinal indicator based on sector-level averages of assets-per-participant or contribution ratios captures financial sufficiency of plans within each sector. It may highlight whether certain industries consistently offer more robust retirement benefits.

Together, these features allow models to account for both who has access and how well those plans perform, making your adequacy classification more equitable and diagnostically transparent.

# Sector Participant Share (SECTOR\_PARTICIPANT\_SHARE)

Engineer a plan penetration rate variable.

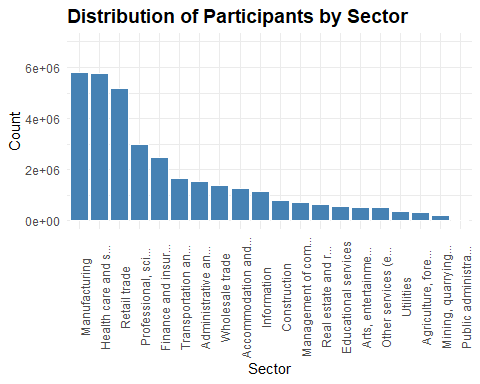
total\_participants\_all <- sum(plans$TOTAL\_ACTIVE\_PARTCP\_EOY)  
  
sector\_participant\_share <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_participants = sum(TOTAL\_ACTIVE\_PARTCP\_EOY, na.rm = TRUE)) %>%  
 mutate(percent\_participants = round(total\_participants / total\_participants\_all, 4)) %>%  
 arrange(desc(percent\_participants))  
  
kable(sector\_participant\_share,  
 col.names = c("Sector", "Participants", "Percent"),  
 caption = "Participant Share By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Participant Share By Sector

| Sector | Participants | Percent |
| --- | --- | --- |
| Manufacturing | 5,816,261 | 0.1721 |
| Health care and social assistance | 5,770,892 | 0.1708 |
| Retail trade | 5,188,143 | 0.1535 |
| Professional, scientific, and technical services | 2,984,643 | 0.0883 |
| Finance and insurance | 2,465,650 | 0.0730 |
| Transportation and warehousing | 1,665,496 | 0.0493 |
| Administrative and support and waste management and remediation services | 1,547,693 | 0.0458 |
| Wholesale trade | 1,382,098 | 0.0409 |
| Accommodation and food services | 1,268,490 | 0.0375 |
| Information | 1,143,613 | 0.0338 |
| Construction | 790,349 | 0.0234 |
| Management of companies and enterprises | 691,080 | 0.0204 |
| Real estate and rental and leasing | 617,697 | 0.0183 |
| Educational services | 556,011 | 0.0165 |
| Arts, entertainment, and recreation | 531,316 | 0.0157 |
| Other services (except public administration) | 526,736 | 0.0156 |
| Utilities | 337,075 | 0.0100 |
| Agriculture, forestry, fishing and hunting | 304,518 | 0.0090 |
| Mining, quarrying, and oil and gas extraction | 189,523 | 0.0056 |
| Public administration | 19,354 | 0.0006 |

Plot participant share by sector.

ggplot(sector\_participant\_share, aes(x = forcats::fct\_reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20),   
 total\_participants, .desc = TRUE),   
 y = total\_participants)) +  
 geom\_col(fill = "steelblue", color = "white") +  
 # geom\_text(aes(label = paste0(scales::comma(total\_participants), "\n", round(percent\_participants \* 100, 0), "%")),  
 # vjust = -0.3, size = 3.0) +  
 labs(title = "Distribution of Participants by Sector",  
 x = "Sector",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold")) +  
 expand\_limits(y = max(sector\_participant\_share$total\_participants) \* 1.2)



This distribution reveals a disparity in coverage. High-coverage sectors like Manufacturing, Healthcare, and Retail dominate participant coverage. They most likely host large employers with institutional plans, unionized workforces, or strong compliance cultures. Their plans may be more standardized, with higher assets-per-participant and broader eligibility.

Low-coverage sectors like Agriculture, Mining, Education, and Public Administration (0.06%) are severely underrepresented. They may face structural barriers: seasonal labor, fragmented employment, or reliance on alternative retirement systems, e.g. public pensions are not captured in Form 5500.

# Sector Participant Share (SECTOR\_PARTCP\_SHARE)

Engineer participant share.

sector\_participant\_share <- sector\_participant\_share %>%  
 select(-total\_participants) %>%  
 rename(SECTOR\_PARTCP\_SHARE = percent\_participants)  
  
plans <- left\_join(plans, sector\_participant\_share, by = "SECTOR\_TITLE")

# Total Plan Assets (TOTAL\_ASSETS\_BOY, TOTAL\_ASSETS\_EOY)

These variables represent the total value of all plan assets held by the plan at the start and end of the plan year.

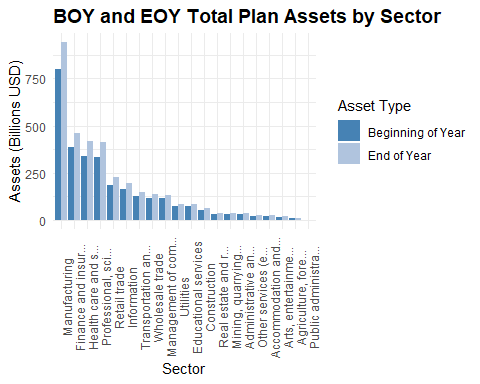
summary\_by\_plan\_assets <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_assets\_boy = sum(TOTAL\_ASSETS\_BOY),  
 total\_assets\_eoy = sum(TOTAL\_ASSETS\_EOY))  
  
kable(summary\_by\_plan\_assets,  
 col.names = c("Sector", "BOY Assets", "EOY Assets"),  
 caption = "BOY and EOY Total Plan Assets By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Plan Assets By Sector

| Sector | BOY Assets | EOY Assets |
| --- | --- | --- |
| Accommodation and food services | 21,705,828,031 | 25,633,566,996 |
| Administrative and support and waste management and remediation services | 29,596,880,338 | 36,718,791,835 |
| Agriculture, forestry, fishing and hunting | 8,792,652,156 | 10,485,294,363 |
| Arts, entertainment, and recreation | 16,342,015,470 | 19,880,418,779 |
| Construction | 52,435,408,155 | 63,704,984,606 |
| Educational services | 72,117,471,271 | 83,505,109,992 |
| Finance and insurance | 385,841,557,063 | 462,209,336,342 |
| Health care and social assistance | 341,098,813,723 | 420,502,225,654 |
| Information | 165,336,864,048 | 196,597,143,647 |
| Management of companies and enterprises | 115,082,845,062 | 133,260,081,045 |
| Manufacturing | 799,059,754,163 | 942,140,001,747 |
| Mining, quarrying, and oil and gas extraction | 33,518,456,182 | 38,136,240,826 |
| Other services (except public administration) | 22,654,518,939 | 27,356,236,049 |
| Professional, scientific, and technical services | 335,929,140,946 | 415,053,599,330 |
| Public administration | 1,005,197,959 | 1,221,041,257 |
| Real estate and rental and leasing | 32,281,975,459 | 39,688,108,565 |
| Retail trade | 188,082,438,039 | 225,343,866,676 |
| Transportation and warehousing | 126,780,888,881 | 150,334,119,245 |
| Utilities | 76,639,821,418 | 86,223,597,888 |
| Wholesale trade | 117,337,778,033 | 138,768,225,479 |

Plot assets by sector.

assets\_long <- summary\_by\_plan\_assets %>%  
 pivot\_longer(cols = c(total\_assets\_boy, total\_assets\_eoy),  
 names\_to = "Asset\_Type",  
 values\_to = "Asset\_Value")  
  
ggplot(assets\_long, aes(x = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), -Asset\_Value),   
 y = Asset\_Value / 1e9, fill = Asset\_Type)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 scale\_fill\_manual(values = c("total\_assets\_boy" = "steelblue", "total\_assets\_eoy" = "lightsteelblue"),  
 labels = c("total\_assets\_boy" = "Beginning of Year", "total\_assets\_eoy" = "End of Year")) +  
 labs(title = "BOY and EOY Total Plan Assets by Sector",  
 x = "Sector",  
 y = "Assets (Billions USD)",  
 fill = "Asset Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold"))



Asset growth over the course of the year is consistent across all sectors.

# Total Assets Growth Rate (TOTAL\_ASSETS\_GROWTH\_RATE)

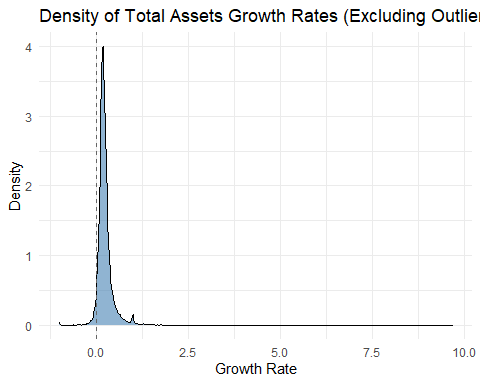
Create a growth rate and assets per participant for each plan.

plans <- plans %>%  
 mutate(  
 TOTAL\_ASSETS\_GROWTH\_RATE = case\_when(  
 TOTAL\_ASSETS\_BOY == 0 & TOTAL\_ASSETS\_EOY > 0 ~ 1.0000,  
 TOTAL\_ASSETS\_BOY == 0 & TOTAL\_ASSETS\_EOY == 0 ~ 0,  
 TOTAL\_ASSETS\_BOY != 0 ~ round((TOTAL\_ASSETS\_EOY - TOTAL\_ASSETS\_BOY) / TOTAL\_ASSETS\_BOY, 4),  
 TRUE ~ NA\_real\_))

Produce a density plot to show the distribution of asset growth rates.

One plan with an asset growth rate of 84% is a serious outlier and distorting the overall distribution. I exclude the outlier in the following plot.

ggplot(plans %>% filter(TOTAL\_ASSETS\_GROWTH\_RATE <= 10), aes(x = TOTAL\_ASSETS\_GROWTH\_RATE)) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 geom\_vline(xintercept = 0, linetype = "dashed", color = "gray40") +  
 labs(title = "Density of Total Assets Growth Rates (Excluding Outlier)",  
 x = "Growth Rate",  
 y = "Density") +  
 theme\_minimal()



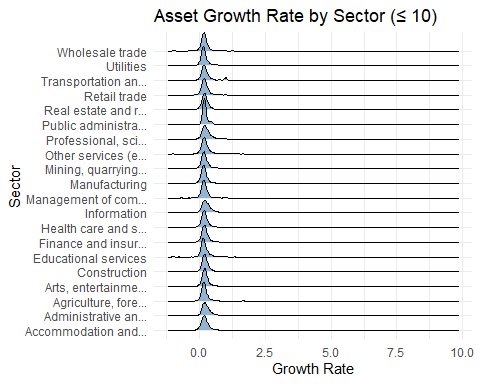
Most plans have modest or stagnant asset growth, likely reflecting market conditions, contribution behavior, or plan maturity.

The right skew suggests a subset of plans with strong growth, possibly due to rollovers, mergers, or aggressive investment strategies.

The sharp peak near zero could indicate structural inertia, plans that are underfunded, inactive, or not receiving contributions.

Create a ridgeline plot by sector.

plans %>%  
 filter(TOTAL\_ASSETS\_GROWTH\_RATE <= 10) %>%  
 ggplot(aes(x = TOTAL\_ASSETS\_GROWTH\_RATE, y = stringr::str\_trunc(SECTOR\_TITLE, width = 20))) +  
 geom\_density\_ridges(bandwidth = 0.05, fill = "steelblue", alpha = 0.6) +  
 labs(title = "Asset Growth Rate by Sector (≤ 10)",  
 x = "Growth Rate",  
 y = "Sector") +  
 theme\_minimal()



Most sectors cluster near zero; the sharp density around zero suggests widespread stagnation or minimal growth, possibly reflecting low contributions, market flatness, or plan maturity.

Right-skewed tails in sectors like Finance and Insurance, Real Estate, and Professional Services hint at pockets of aggressive growth, likely driven by rollovers, mergers, or high-income participant behavior.

Public sector plans (e.g. Educational Services, Health Care, Public Administration) show tighter distributions, possibly due to more regulated funding structures or consistent employer contributions.

Volatile sectors like Construction, Arts, and Accommodation show broader spread, likely reflecting economic sensitivity and participant turnover.

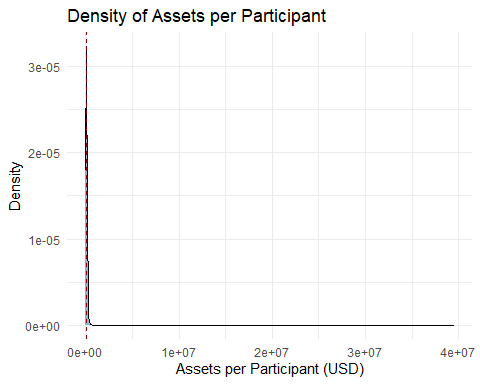
# Assets Per Participant (ASSETS\_PER\_PARTCP)

plans <- plans %>%  
 mutate(  
 DENOMINATOR = case\_when(  
 TOTAL\_ACCBAL\_PARTCP\_EOY > 0 ~ TOTAL\_ACCBAL\_PARTCP\_EOY,  
 TOTAL\_ACTIVE\_PARTCP\_EOY > 0 ~ TOTAL\_ACTIVE\_PARTCP\_EOY,  
 TOTAL\_ACTIVE\_PARTCP\_BOY > 0 ~ TOTAL\_ACTIVE\_PARTCP\_BOY,  
 TRUE ~ NA\_real\_),  
 ASSETS\_PER\_PARTCP = round(TOTAL\_ASSETS\_EOY / DENOMINATOR, 0)) %>%   
 select(-DENOMINATOR)

Plot assets per participant

ggplot(plans, aes(x = ASSETS\_PER\_PARTCP)) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 geom\_vline(xintercept = median(plans$ASSETS\_PER\_PARTCP, na.rm = TRUE),   
 linetype = "dashed", color = "darkred") +  
 labs(title = "Density of Assets per Participant",  
 x = "Assets per Participant (USD)",  
 y = "Density") +  
 theme\_minimal()

## Warning: Removed 15 rows containing non-finite outside the scale range  
## (`stat\_density()`).

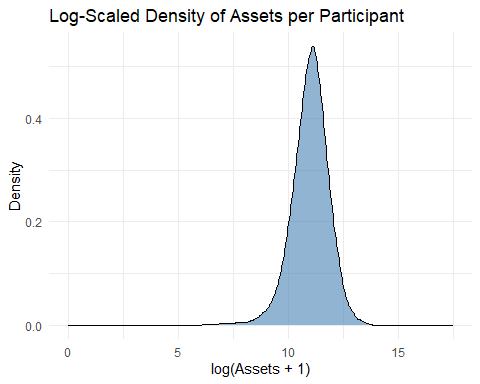


That sharp peak near the lower end, followed by a long right tail, suggests most plans cluster below $100K per participant, but a few are pulling the average upward with much higher balances.

Plot the log-transformed distribution to reveal structure in the lower range.

ggplot(plans, aes(x = log1p(ASSETS\_PER\_PARTCP))) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 labs(title = "Log-Scaled Density of Assets per Participant",  
 x = "log(Assets + 1)",  
 y = "Density") +  
 theme\_minimal()

## Warning: Removed 15 rows containing non-finite outside the scale range  
## (`stat\_density()`).



That peak around 10 corresponds to roughly $22,000–$25,000 per participant, which is a much more grounded estimate of typical adequacy than the inflated mean near $100K.

Most plans are far below the mean: The log transformation exposes that the bulk of plans cluster in the $10K–$30K range, not near six figures.

Long tail still exists: But now it’s visually compressed, allowing us to focus on the core distribution.

# Assets Per Participant Tier (ASSETS\_PER\_PARTCP\_TIER)

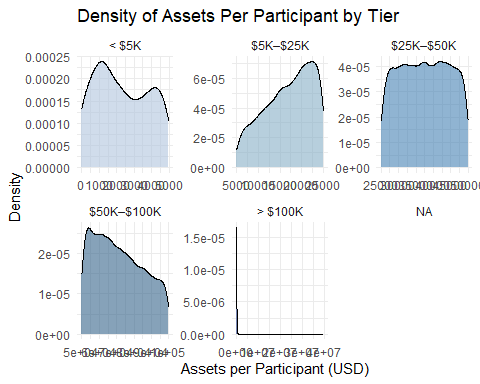
Create assets per parrticipant tier labels.

plans <- plans %>%  
 mutate(  
 ASSETS\_PER\_PARTCP\_TIER = case\_when(  
 ASSETS\_PER\_PARTCP < 5000 ~ "< $5K",  
 ASSETS\_PER\_PARTCP >= 5000 & ASSETS\_PER\_PARTCP < 25000 ~ "$5K–$25K",  
 ASSETS\_PER\_PARTCP >= 25000 & ASSETS\_PER\_PARTCP < 50000 ~ "$25K–$50K",  
 ASSETS\_PER\_PARTCP >= 50000 & ASSETS\_PER\_PARTCP < 100000 ~ "$50K–$100K",  
 ASSETS\_PER\_PARTCP >= 100000 ~ "> $100K",  
 TRUE ~ NA\_character\_),  
 ASSETS\_PER\_PARTCP\_TIER = factor(ASSETS\_PER\_PARTCP\_TIER,  
 levels = c("< $5K", "$5K–$25K", "$25K–$50K", "$50K–$100K", "> $100K"),  
 ordered = TRUE))

Plot density by tier.

ggplot(plans, aes(x = ASSETS\_PER\_PARTCP, fill = ASSETS\_PER\_PARTCP\_TIER)) +  
 geom\_density(alpha = 0.6) +  
 facet\_wrap(~ ASSETS\_PER\_PARTCP\_TIER, scales = "free") +  
 scale\_fill\_manual(values = c("#B0C4DE", "#87AFC7", "#4682B4", "#36648B", "#27408B")) +  
 labs(title = "Density of Assets Per Participant by Tier",  
 x = "Assets per Participant (USD)",  
 y = "Density") +  
 theme\_minimal() +  
 theme(legend.position = "none")

## Warning: Removed 15 rows containing non-finite outside the scale range  
## (`stat\_density()`).



The “< $5K” tier shows a tight peak near $1K, suggesting many plans cluster at very low adequacy. In contrast, the “$50K–$100K” tier is broader, indicating more heterogeneity in plan quality.

The “> $100K” tier is especially telling: a sharp peak just above $100K followed by a steep drop-off. That suggests a few high-performing plans dominate this tier, but most don’t go far beyond the threshold.

# Total Active Participants (TOTAL\_ACTIVE\_PARTCP\_BOY, TOTAL\_ACTIVE\_PARTCP\_EOY)

These variables provide the number of active participants in the plan as of the start and end of the reporting year. Active means eligible to contribute or currently contributing and includes employees who are actively participating or eligible to participate, even if they haven’t made contributions yet. These variables reflect the current coverage footprint of the plan.

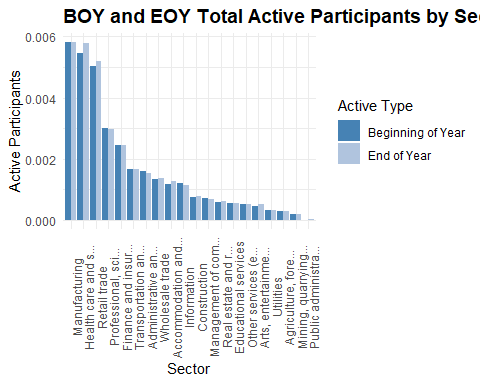
summary\_by\_active\_participants <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_active\_boy = sum(TOTAL\_ACTIVE\_PARTCP\_BOY),  
 total\_active\_eoy = sum(TOTAL\_ACTIVE\_PARTCP\_EOY))  
  
kable(summary\_by\_active\_participants,  
 col.names = c("Sector", "BOY Participants", "EOY Participants"),  
 caption = "BOY and EOY Total Active Participants By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Active Participants By Sector

| Sector | BOY Participants | EOY Participants |
| --- | --- | --- |
| Accommodation and food services | 1,168,190 | 1,268,490 |
| Administrative and support and waste management and remediation services | 1,607,271 | 1,547,693 |
| Agriculture, forestry, fishing and hunting | 283,718 | 304,518 |
| Arts, entertainment, and recreation | 474,682 | 531,316 |
| Construction | 755,274 | 790,349 |
| Educational services | 549,952 | 556,011 |
| Finance and insurance | 2,452,314 | 2,465,650 |
| Health care and social assistance | 5,462,618 | 5,770,892 |
| Information | 1,206,243 | 1,143,613 |
| Management of companies and enterprises | 710,102 | 691,080 |
| Manufacturing | 5,807,609 | 5,816,261 |
| Mining, quarrying, and oil and gas extraction | 189,698 | 189,523 |
| Other services (except public administration) | 512,773 | 526,736 |
| Professional, scientific, and technical services | 2,999,348 | 2,984,643 |
| Public administration | 16,822 | 19,354 |
| Real estate and rental and leasing | 596,883 | 617,697 |
| Retail trade | 5,048,063 | 5,188,143 |
| Transportation and warehousing | 1,675,569 | 1,665,496 |
| Utilities | 332,464 | 337,075 |
| Wholesale trade | 1,345,132 | 1,382,098 |

Plot active participants by sector.

active\_long <- summary\_by\_active\_participants %>%  
 pivot\_longer(cols = c(total\_active\_boy, total\_active\_eoy),  
 names\_to = "Active\_Type",  
 values\_to = "Active\_Value")  
  
ggplot(active\_long, aes(x = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), -Active\_Value), y = Active\_Value / 1e9, fill = Active\_Type)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 scale\_fill\_manual(values = c("total\_active\_boy" = "steelblue", "total\_active\_eoy" = "lightsteelblue"),  
 labels = c("total\_active\_boy" = "Beginning of Year", "total\_active\_eoy" = "End of Year")) +  
 labs(title = "BOY and EOY Total Active Participants by Sector",  
 x = "Sector",  
 y = "Active Participants",  
 fill = "Active Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold"))



Manufacturing, Healthcare, and Professional Services again dominate in participant volume, mirroring their asset dominance. Sectors like Retail, Accommodation, and Arts show modest participant counts despite being labor-intensive, which may signal structural under-coverage.

The EOY bars generally exceed BOY, suggesting net growth in active participation, but the magnitude varies by sector.

# Participant Growth Rate (PARTCP\_GROWTH\_RATE)

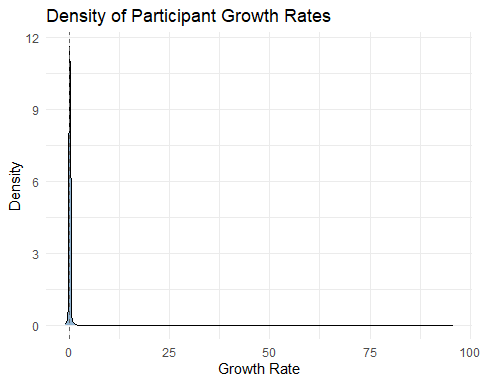
Engineer participant growth rate at the plan level.

plans <- plans %>%  
 mutate(PARTCP\_GROWTH\_RATE = case\_when(  
 TOTAL\_ACTIVE\_PARTCP\_BOY == 0 & TOTAL\_ACTIVE\_PARTCP\_EOY > 0 ~ 1.0000,  
 TOTAL\_ACTIVE\_PARTCP\_BOY == 0 & TOTAL\_ACTIVE\_PARTCP\_EOY == 0 ~ 0,  
 TOTAL\_ACTIVE\_PARTCP\_BOY != 0 ~ round((TOTAL\_ACTIVE\_PARTCP\_EOY - TOTAL\_ACTIVE\_PARTCP\_BOY) / TOTAL\_ACTIVE\_PARTCP\_BOY, 4)))

Produce a density plot to show the distribution of participant growth rates.

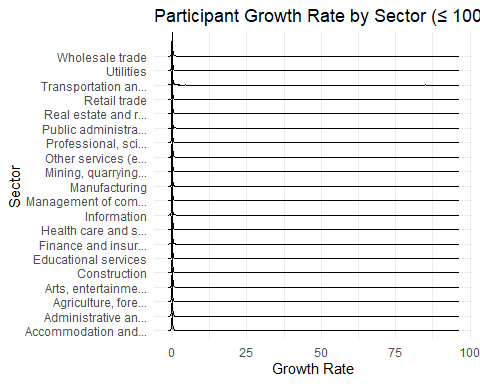
One plan with a participant growth rate of 299% is a serious outlier and distorting the overall distribution. I exclude the outlier in the following plot.

ggplot(plans %>% filter(PARTCP\_GROWTH\_RATE <= 100), aes(x = PARTCP\_GROWTH\_RATE)) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 geom\_vline(xintercept = 0, linetype = "dashed", color = "gray40") +  
 labs(title = "Density of Participant Growth Rates",  
 x = "Growth Rate",  
 y = "Density") +  
 theme\_minimal()



Create a ridgeline plot by sector.

plans %>%  
 filter(PARTCP\_GROWTH\_RATE <= 100) %>%  
 ggplot(aes(x = PARTCP\_GROWTH\_RATE, y = stringr::str\_trunc(SECTOR\_TITLE, width = 20))) +  
 geom\_density\_ridges(bandwidth = 0.05, fill = "steelblue", alpha = 0.6) +  
 labs(title = "Participant Growth Rate by Sector (≤ 100)",  
 x = "Growth Rate",  
 y = "Sector") +  
 theme\_minimal()



Most sectors seem tightly clustered near zero, and the horizontal scale compresses variation, making it hard to distinguish meaningful outliers or trends.

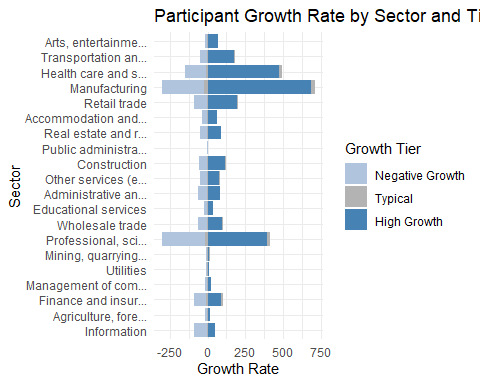
# Participant Growth Rate Tier (PARTCP\_GROWTH\_TIER)

Create a participant growth tier flag to delineate between negative, typical, and high growth plans.

plans <- plans%>%  
 mutate(PARTCP\_GROWTH\_TIER = case\_when(PARTCP\_GROWTH\_RATE > 0.05 ~ "High Growth",  
 PARTCP\_GROWTH\_RATE < -0.05 ~ "Negative Growth",  
 TRUE ~ "Typical"),  
 PARTCP\_GROWTH\_TIER = factor(PARTCP\_GROWTH\_TIER,  
 levels = c("Negative Growth", "Typical", "High Growth"),  
 ordered = TRUE))

Plot participant growth rates by sector with the growth tiers.

ggplot(plans, aes(x = PARTCP\_GROWTH\_RATE, y = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), PARTCP\_GROWTH\_RATE), fill = PARTCP\_GROWTH\_TIER)) +  
 geom\_col() +  
 scale\_fill\_manual(values = c("High Growth" = "steelblue", "Negative Growth" = "lightsteelblue", "Typical" = "gray70")) +  
 labs(title = "Participant Growth Rate by Sector and Tier",  
 x = "Growth Rate",  
 y = "Sector",  
 fill = "Growth Tier") +  
 theme\_minimal()



Participant growth rates vary meaningfully across sectors, and the flagged outliers, e.g. Manufacturing with high growth, Information with negative growth, suggest real economic or policy implications.

# Total Participant Contributions (TOTAL\_CONTRIB\_PARTCP\_BOY, TOTAL\_CONTRIB\_PARTCP\_EOY)

These variables capture the total dollar amount of participant contributions on record as of the start and end of the reporting year. The first reflects contributions made prior to the current plan year but still held in the plan. This is useful for understanding carryover balances or prior-year contribution momentum.

The latter includes all contributions made by participants during the year, plus any prior balances still retained. It is often used to assess current-year contribution activity and participant engagement.

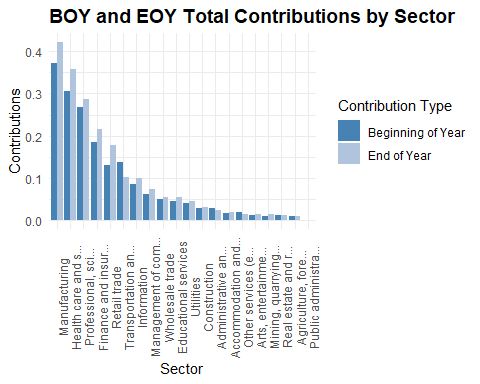
summary\_by\_partcp\_contributions <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_contrib\_boy = sum(TOTAL\_CONTRIB\_PARTCP\_BOY),  
 total\_contrib\_eoy = sum(TOTAL\_CONTRIB\_PARTCP\_EOY),  
 .groups = "drop")  
  
kable(summary\_by\_partcp\_contributions,  
 col.names = c("Sector", "BOY Contributions", "EOY Contributions"),  
 caption = "BOY and EOY Total Contributions By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Contributions By Sector

| Sector | BOY Contributions | EOY Contributions |
| --- | --- | --- |
| Accommodation and food services | 16,195,254 | 19,127,522 |
| Administrative and support and waste management and remediation services | 27,573,485 | 24,828,878 |
| Agriculture, forestry, fishing and hunting | 9,409,833 | 10,366,717 |
| Arts, entertainment, and recreation | 12,391,950 | 13,900,123 |
| Construction | 29,734,078 | 30,923,412 |
| Educational services | 45,318,750 | 55,067,221 |
| Finance and insurance | 185,173,544 | 215,317,039 |
| Health care and social assistance | 305,669,015 | 358,891,892 |
| Information | 86,132,502 | 99,562,366 |
| Management of companies and enterprises | 60,895,987 | 74,139,373 |
| Manufacturing | 372,763,144 | 422,021,493 |
| Mining, quarrying, and oil and gas extraction | 10,094,683 | 13,679,972 |
| Other services (except public administration) | 19,582,701 | 15,426,136 |
| Professional, scientific, and technical services | 267,372,240 | 285,891,040 |
| Public administration | 534,422 | 418,303 |
| Real estate and rental and leasing | 12,132,954 | 11,424,443 |
| Retail trade | 129,333,726 | 177,622,307 |
| Transportation and warehousing | 137,043,315 | 101,171,980 |
| Utilities | 39,769,886 | 44,340,428 |
| Wholesale trade | 50,820,438 | 54,719,266 |

Plot contributions by sector.

contributions\_long <- summary\_by\_partcp\_contributions %>%  
 pivot\_longer(cols = c(total\_contrib\_boy, total\_contrib\_eoy),  
 names\_to = "Contribution\_Type",  
 values\_to = "Contribution\_Value")  
  
ggplot(contributions\_long, aes(x = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), -Contribution\_Value), y = Contribution\_Value / 1e9, fill = Contribution\_Type)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 scale\_fill\_manual(values = c("total\_contrib\_boy" = "steelblue", "total\_contrib\_eoy" = "lightsteelblue"),  
 labels = c("total\_contrib\_boy" = "Beginning of Year", "total\_contrib\_eoy" = "End of Year")) +  
 labs(title = "BOY and EOY Total Contributions by Sector",  
 x = "Sector",  
 y = "Contributions",  
 fill = "Contribution Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold"))



Manufacturing, Professional Services, and Hotels & Restaurants lead in total contributions, echoing their asset and participant dominance.

The EOY bars generally exceed BOY, suggesting net contribution inflows, but the magnitude and slope vary by sector.

Sectors like Information and Mining show relatively flat or modest contribution growth, which may flag under-engagement or structural constraints

# Participant Contribution Growth Rate (CONTRIB\_PARTCP\_GROWTH\_RATE)

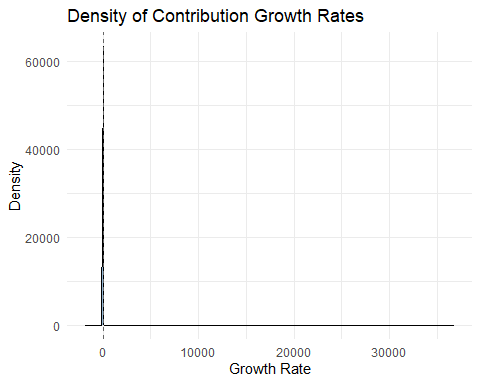
Engineer a contribution growth rate at the plan level.

plans <- plans %>%  
 mutate(CONTRIB\_PARTCP\_GROWTH\_RATE = case\_when(  
 TOTAL\_CONTRIB\_PARTCP\_BOY == 0 & TOTAL\_CONTRIB\_PARTCP\_EOY > 0 ~ 1.0000,  
 TOTAL\_CONTRIB\_PARTCP\_BOY == 0 & TOTAL\_CONTRIB\_PARTCP\_EOY == 0 ~ 0,  
 TOTAL\_CONTRIB\_PARTCP\_BOY == 0 & TOTAL\_CONTRIB\_PARTCP\_EOY < 0 ~ 0,  
 TOTAL\_CONTRIB\_PARTCP\_BOY != 0 ~ round((TOTAL\_CONTRIB\_PARTCP\_EOY - TOTAL\_CONTRIB\_PARTCP\_BOY) / TOTAL\_CONTRIB\_PARTCP\_BOY, 4)))

Produce a density plot to show the distribution of contribution growth rates.

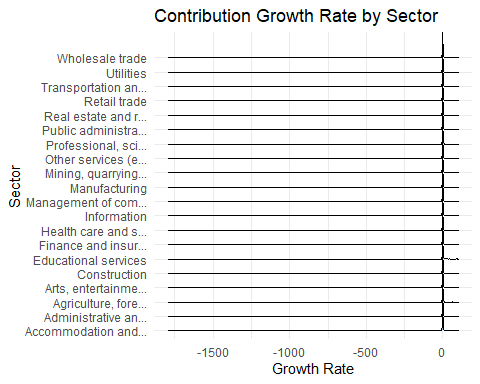
One plan with a participant growth rate of 299% is a serious outlier and distorting the overall distribution. I exclude the outlier in the following plot.

ggplot(plans, aes(x = CONTRIB\_PARTCP\_GROWTH\_RATE)) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 geom\_vline(xintercept = 0, linetype = "dashed", color = "gray40") +  
 labs(title = "Density of Contribution Growth Rates",  
 x = "Growth Rate",  
 y = "Density") +  
 theme\_minimal()



Create a ridgeline plot by sector.

plans %>%  
 filter(CONTRIB\_PARTCP\_GROWTH\_RATE <= 100) %>%  
 ggplot(aes(x = CONTRIB\_PARTCP\_GROWTH\_RATE, y = stringr::str\_trunc(SECTOR\_TITLE, width = 20))) +  
 geom\_density\_ridges(bandwidth = 0.05, fill = "steelblue", alpha = 0.6) +  
 labs(title = "Contribution Growth Rate by Sector",  
 x = "Growth Rate",  
 y = "Sector") +  
 theme\_minimal()



Most sectors seem tightly clustered near zero, and the horizontal scale compresses variation, making it hard to distinguish meaningful outliers or trends.

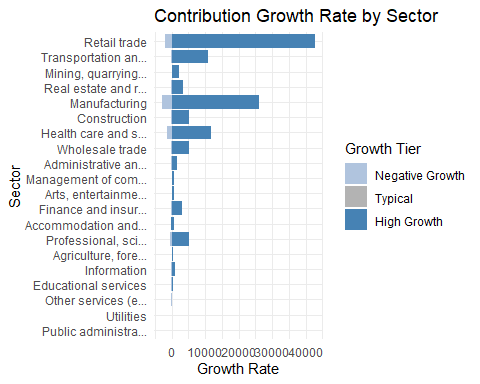
# Participant Contribution Growth Tier (CONTRIB\_PARTCP\_GROWTH\_TIER)

Create a contribution growth rate tier to delineate between negative, typical, and high growth plans.

plans <- plans%>%  
 mutate(CONTRIB\_PARTCP\_GROWTH\_TIER= case\_when(CONTRIB\_PARTCP\_GROWTH\_RATE > 0.05 ~ "High Growth",  
 CONTRIB\_PARTCP\_GROWTH\_RATE < -0.05 ~ "Negative Growth",  
 TRUE ~ "Typical"),  
 CONTRIB\_PARTCP\_GROWTH\_TIER = factor(CONTRIB\_PARTCP\_GROWTH\_TIER,  
 levels = c("Negative Growth", "Typical", "High Growth"),  
 ordered = TRUE))

Plot contribution growth tiers by sector with the growth flags.

ggplot(plans, aes(x = CONTRIB\_PARTCP\_GROWTH\_RATE, y = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), CONTRIB\_PARTCP\_GROWTH\_RATE), fill = CONTRIB\_PARTCP\_GROWTH\_TIER)) +  
 geom\_col() +  
 scale\_fill\_manual(values = c("High Growth" = "steelblue", "Negative Growth" = "lightsteelblue", "Typical" = "gray70")) +  
 labs(title = "Contribution Growth Rate by Sector",  
 x = "Growth Rate",  
 y = "Sector",  
 fill = "Growth Tier") +  
 theme\_minimal()



Transportation and warehousing, Mining, and Healthcare show strong contribution growth—suggesting active funding and participant engagement.

Information and Arts & Entertainment show negative growth, which could signal stagnation, plan attrition, or structural undercoverage.

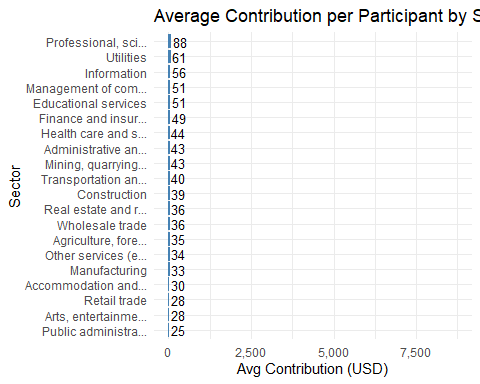
# Average Contribution Per Participant (AVERAGE\_CONTRIB\_PER\_PARTCP)

Implement average contribution per participant.

plans <- plans %>%  
 mutate(AVERAGE\_CONTRIB\_PER\_PARTCP = case\_when(  
 TOTAL\_CONTRIB\_PARTCP\_EOY < 0 ~ 0,  
 TOTAL\_ACCBAL\_PARTCP\_EOY == 0 & TOTAL\_CONTRIB\_PARTCP\_EOY > 0 ~ 1.000,  
 TOTAL\_ACCBAL\_PARTCP\_EOY == 0 & TOTAL\_CONTRIB\_PARTCP\_EOY == 0 ~ 0,  
 TOTAL\_ACCBAL\_PARTCP\_EOY != 0 ~ round(TOTAL\_CONTRIB\_PARTCP\_EOY / TOTAL\_ACCBAL\_PARTCP\_EOY, 4)))

Plot average contribution per participant by sector.

summary\_by\_average\_contribution <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(AVERAGE\_CONTRIB\_PER\_PARTCP = mean(AVERAGE\_CONTRIB\_PER\_PARTCP))  
  
ggplot(summary\_by\_average\_contribution, aes(x = AVERAGE\_CONTRIB\_PER\_PARTCP,   
 y = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20),  
 AVERAGE\_CONTRIB\_PER\_PARTCP))) +  
 geom\_col(fill = "steelblue") +  
 geom\_text(aes(label = scales::comma(round(AVERAGE\_CONTRIB\_PER\_PARTCP))),  
 hjust = -0.1, size = 3.5) +  
 scale\_x\_continuous(labels = scales::label\_comma()) +  
  
 labs(title = "Average Contribution per Participant by Sector",  
 x = "Avg Contribution (USD)",  
 y = "Sector") +  
 theme\_minimal() +   
 coord\_cartesian(xlim = c(0, max(plans$AVERAGE\_CONTRIB\_PER\_PARTCP) \* 1.1))



# Contribution Per Participant Tier (CONTRIB\_PER\_PARTCP\_TIER)

Create a tiered variable for average participant contribution.

plans <- plans%>%  
 mutate(CONTRIB\_PER\_PARTCP\_TIER = cut(AVERAGE\_CONTRIB\_PER\_PARTCP,  
 breaks = c(0, 1000, 2000, 3000, Inf),  
 labels = c("Low", "Moderate", "High", "Very High")),  
 CONTRIB\_PARTCP\_GROWTH\_TIER = factor(CONTRIB\_PARTCP\_GROWTH\_TIER,  
 levels = c("Low", "Moderate", "High", "Very High"),  
 ordered = TRUE))

# Employer Contributions (TOTAL\_CONTRIB\_EMPLR\_BOY, TOTAL\_CONTRIB\_EMPLR\_EOY)

These variables represent the total dollar amount of employer contributions on record as of the start and end of the reporting year. They include contributions made during the year plus any retained prior balances and are often used to assess current-year employer funding and plan support.

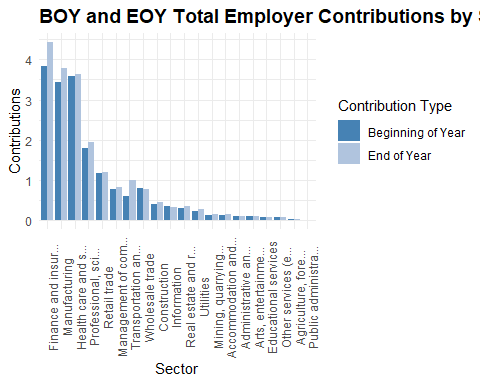
summary\_er\_contributions <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_contrib\_boy = sum(TOTAL\_CONTRIB\_EMPLR\_BOY),  
 total\_contrib\_eoy = sum(TOTAL\_CONTRIB\_EMPLR\_EOY),  
 .groups = "drop")  
  
kable(summary\_er\_contributions,  
 col.names = c("Sector", "BOY Contributions", "EOY Contributions"),  
 caption = "BOY and EOY Total Employer Contributions By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Employer Contributions By Sector

| Sector | BOY Contributions | EOY Contributions |
| --- | --- | --- |
| Accommodation and food services | 126,709,092 | 145,508,537 |
| Administrative and support and waste management and remediation services | 109,949,507 | 103,792,732 |
| Agriculture, forestry, fishing and hunting | 24,832,855 | 24,296,013 |
| Arts, entertainment, and recreation | 89,523,702 | 98,199,295 |
| Construction | 402,594,667 | 442,969,038 |
| Educational services | 79,940,471 | 85,236,763 |
| Finance and insurance | 3,837,075,529 | 4,432,300,819 |
| Health care and social assistance | 3,579,953,008 | 3,638,193,545 |
| Information | 338,602,381 | 326,040,783 |
| Management of companies and enterprises | 767,388,478 | 832,896,784 |
| Manufacturing | 3,445,493,151 | 3,783,174,572 |
| Mining, quarrying, and oil and gas extraction | 129,658,601 | 156,671,780 |
| Other services (except public administration) | 73,743,413 | 83,836,797 |
| Professional, scientific, and technical services | 1,800,371,459 | 1,953,942,010 |
| Public administration | 1,910,799 | 2,051,639 |
| Real estate and rental and leasing | 297,520,274 | 358,590,550 |
| Retail trade | 1,174,847,826 | 1,208,806,909 |
| Transportation and warehousing | 598,304,548 | 991,084,271 |
| Utilities | 236,843,909 | 263,978,219 |
| Wholesale trade | 800,309,210 | 778,160,509 |

Plot contributions by sector.

er\_contributions\_long <- summary\_er\_contributions %>%  
 pivot\_longer(cols = c(total\_contrib\_boy, total\_contrib\_eoy),  
 names\_to = "Contribution\_Type",  
 values\_to = "Contribution\_Value")  
  
ggplot(er\_contributions\_long, aes(x = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), -Contribution\_Value), y = Contribution\_Value / 1e9, fill = Contribution\_Type)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 scale\_fill\_manual(values = c("total\_contrib\_boy" = "steelblue", "total\_contrib\_eoy" = "lightsteelblue"),  
 labels = c("total\_contrib\_boy" = "Beginning of Year", "total\_contrib\_eoy" = "End of Year")) +  
 labs(title = "BOY and EOY Total Employer Contributions by Sector",  
 x = "Sector",  
 y = "Contributions",  
 fill = "Contribution Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold"))



Health care, Manufacturing, and Professional Services lead in employer contributions at both BOY and EOY, suggesting sustained institutional support.

The EOY bars generally exceed BOY, indicating net inflows, but the slope varies by sector. Some sectors show flat or modest growth, which may flag underfunding or stagnation.

Sectors like Arts & Entertainment, Accommodation, and Administrative Support show relatively low employer contributions, which could signal structural gaps in plan generosity.

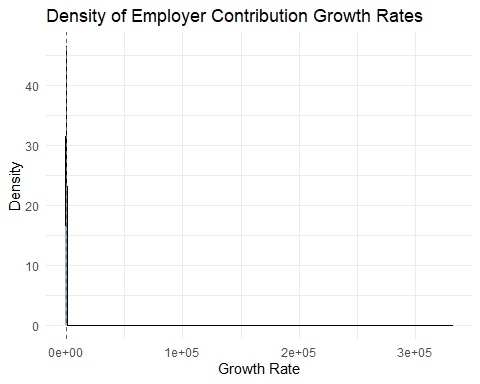
# Employer Contribution Growth Rate (CONTRIB\_EMPLR\_GROWTH\_RATE)

Engineer a contribution growth rate at the plan level.

plans <- plans %>%  
 mutate(CONTRIB\_EMPLR\_GROWTH\_RATE = case\_when(  
 TOTAL\_CONTRIB\_EMPLR\_BOY == 0 & TOTAL\_CONTRIB\_EMPLR\_EOY > 0 ~ 1.0000,  
 TOTAL\_CONTRIB\_EMPLR\_BOY == 0 & TOTAL\_CONTRIB\_EMPLR\_EOY == 0 ~ 0,  
 TOTAL\_CONTRIB\_EMPLR\_BOY == 0 & TOTAL\_CONTRIB\_EMPLR\_EOY < 0 ~ 0,  
 TOTAL\_CONTRIB\_EMPLR\_BOY != 0 ~ round((TOTAL\_CONTRIB\_EMPLR\_EOY - TOTAL\_CONTRIB\_EMPLR\_BOY) / TOTAL\_CONTRIB\_EMPLR\_BOY, 4)))

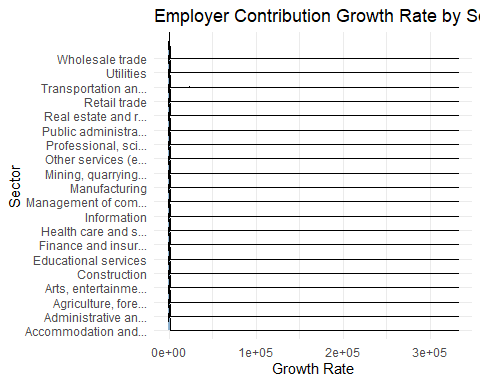
Produce a density plot to show the distribution of contribution growth rates.

ggplot(plans, aes(x = CONTRIB\_EMPLR\_GROWTH\_RATE)) +  
 geom\_density(fill = "steelblue", alpha = 0.6) +  
 geom\_vline(xintercept = 0, linetype = "dashed", color = "gray40") +  
 labs(title = "Density of Employer Contribution Growth Rates",  
 x = "Growth Rate",  
 y = "Density") +  
 theme\_minimal()



Create a ridgeline plot by sector.

plans %>%  
 ggplot(aes(x = CONTRIB\_EMPLR\_GROWTH\_RATE, y = stringr::str\_trunc(SECTOR\_TITLE, width = 20))) +  
 geom\_density\_ridges(bandwidth = 0.05, fill = "steelblue", alpha = 0.6) +  
 labs(title = "Employer Contribution Growth Rate by Sector",  
 x = "Growth Rate",  
 y = "Sector") +  
 theme\_minimal()



Most sectors seem tightly clustered near zero, and the horizontal scale compresses variation, making it hard to distinguish meaningful outliers or trends.

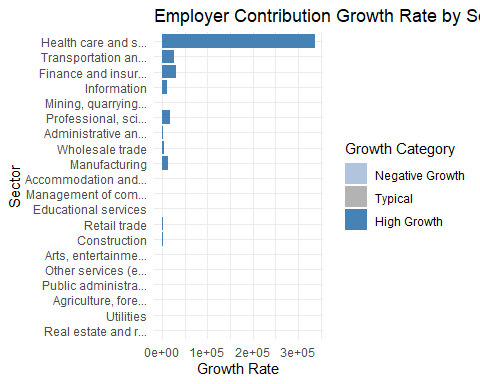
# Employer Contribution Growth Tier (CONTRIB\_EMPLR\_GROWTH\_TIER)

Create an employer contribution growth rate flag to delineate between negative, typical, and high growth plans.

plans <- plans%>%  
 mutate(CONTRIB\_EMPLR\_GROWTH\_TIER = case\_when(CONTRIB\_EMPLR\_GROWTH\_RATE > 0.05 ~ "High Growth",  
 CONTRIB\_EMPLR\_GROWTH\_RATE < -0.05 ~ "Negative Growth",  
 TRUE ~ "Typical"),  
 CONTRIB\_EMPLR\_GROWTH\_TIER = factor(CONTRIB\_EMPLR\_GROWTH\_TIER,  
 levels = c("Negative Growth", "Typical", "High Growth"),  
 ordered = TRUE))

Plot contribution growth rates by sector with the growth flags.

ggplot(plans, aes(x = CONTRIB\_EMPLR\_GROWTH\_RATE, y = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), CONTRIB\_EMPLR\_GROWTH\_RATE), fill = CONTRIB\_EMPLR\_GROWTH\_TIER)) +  
 geom\_col() +  
 scale\_fill\_manual(values = c("High Growth" = "steelblue", "Negative Growth" = "lightsteelblue", "Typical" = "gray70")) +  
 labs(title = "Employer Contribution Growth Rate by Sector",  
 x = "Growth Rate",  
 y = "Sector",  
 fill = "Growth Category") +  
 theme\_minimal()



Transportation and warehousing stands out with exceptional growth—this could reflect increased employer investment, plan expansion, or policy shifts.

Educational services and Wholesale trade also show strong growth, which may signal rising institutional support in traditionally underfunded sectors.

Several sectors show negative growth (e.g., Information, Arts & Entertainment), which could indicate declining employer engagement, plan attrition, or funding retraction.

# Participant Loans (TOTAL\_LOANS\_BOY, TOTAL\_LOANS\_EOY)

These fields reflect the aggregate amount participants have borrowed from their retirement accounts and not yet repaid. Loans are typically allowed under 401(k) and some 403(b) plans, subject to IRS limits (usually up to $50,000 or 50% of vested account balance).

Comparing BOY vs. EOY values helps surface net borrowing or repayment trends. A rising EOY balance may signal increased participant financial stress or plan leniency.

High loan balances can erode retirement adequacy. Plans with high loan-to-asset ratios or frequent borrowing as needing closer scrutiny.

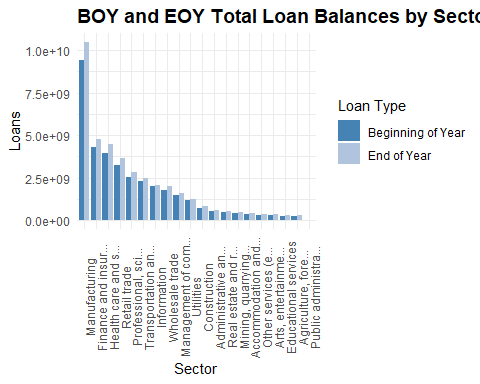
summary\_loans <- plans %>%  
 group\_by(SECTOR\_TITLE) %>%  
 summarise(total\_loans\_boy = sum(TOTAL\_LOANS\_BOY),  
 total\_loans\_eoy = sum(TOTAL\_LOANS\_EOY),  
 .groups = "drop")  
  
kable(summary\_loans,  
 col.names = c("Sector", "BOY Loans", "EOY Loans"),  
 caption = "BOY and EOY Total Loan Balances By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Loan Balances By Sector

| Sector | BOY Loans | EOY Loans |
| --- | --- | --- |
| Accommodation and food services | 354,653,928 | 399,109,723 |
| Administrative and support and waste management and remediation services | 534,909,780 | 614,193,978 |
| Agriculture, forestry, fishing and hunting | 261,364,491 | 301,235,008 |
| Arts, entertainment, and recreation | 294,343,806 | 334,019,998 |
| Construction | 736,448,432 | 846,006,771 |
| Educational services | 265,006,263 | 315,187,010 |
| Finance and insurance | 4,306,378,640 | 4,756,557,917 |
| Health care and social assistance | 3,945,217,033 | 4,494,304,960 |
| Information | 1,991,780,015 | 2,062,004,324 |
| Management of companies and enterprises | 1,466,321,839 | 1,591,701,856 |
| Manufacturing | 9,428,333,030 | 10,481,555,678 |
| Mining, quarrying, and oil and gas extraction | 432,702,094 | 497,633,962 |
| Other services (except public administration) | 295,611,600 | 334,238,621 |
| Professional, scientific, and technical services | 2,521,685,752 | 2,856,497,811 |
| Public administration | 18,653,385 | 20,741,394 |
| Real estate and rental and leasing | 452,963,110 | 522,138,652 |
| Retail trade | 3,261,581,614 | 3,644,105,199 |
| Transportation and warehousing | 2,285,491,879 | 2,503,848,978 |
| Utilities | 1,158,275,903 | 1,229,892,453 |
| Wholesale trade | 1,799,220,344 | 2,024,908,403 |

Plot loans by sector.

er\_loans\_long <- summary\_loans %>%  
 pivot\_longer(cols = c(total\_loans\_boy, total\_loans\_eoy),  
 names\_to = "Loan\_Type",  
 values\_to = "Loan\_Value")  
  
ggplot(er\_loans\_long, aes(x = reorder(stringr::str\_trunc(SECTOR\_TITLE, width = 20), -Loan\_Value), y = Loan\_Value, fill = Loan\_Type)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 scale\_fill\_manual(values = c("total\_loans\_boy" = "steelblue", "total\_loans\_eoy" = "lightsteelblue"),  
 labels = c("total\_loans\_boy" = "Beginning of Year", "total\_loans\_eoy" = "End of Year")) +  
 labs(title = "BOY and EOY Total Loan Balances by Sector",  
 x = "Sector",  
 y = "Loans",  
 fill = "Loan Type") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 0.5),  
 plot.title = element\_text(size = 14, face = "bold"))



Loan balances are rising in most sectors.

The light blue EOY bars often exceed the dark blue BOY bars, suggesting net borrowing across the year. That’s a potential red flag for participant liquidity strain or plan leniency.

Manufacturing and Health Care lead in total loan volume. These sectors show the highest absolute balances—likely due to large participant bases and higher plan penetration. But it also raises questions about adequacy erosion and fallback logic.

Professional and Technical Services show modest growth. Possibly reflecting more conservative borrowing behavior or stronger financial wellness programs.

Church and public sectors (e.g., Educational Services, Public Administration) show lower balances. This aligns with structural plan constraints—403(b)(9) and some 403(b)(1) plans often restrict loans or discourage borrowing.

# Has Outstanding Loans (HAS\_OUTSTANDING\_LOANS)

Engineer a binary variable that identifies whether a plan has outstanding loan balances or not.

plans <- plans %>%  
 mutate(HAS\_OUTSTANDING\_LOANS = case\_when(TOTAL\_LOANS\_BOY > 0 | TOTAL\_LOANS\_EOY > 0 ~ 1,  
 TRUE ~ 0))

summary\_os\_loans <- plans %>%  
 group\_by(SECTOR\_TITLE, HAS\_OUTSTANDING\_LOANS) %>%  
 summarise(entries = n()) %>%  
 pivot\_wider(names\_from = HAS\_OUTSTANDING\_LOANS,  
 values\_from = entries)

## `summarise()` has grouped output by 'SECTOR\_TITLE'. You can override using the  
## `.groups` argument.

kable(summary\_os\_loans,  
 col.names = c("Sector", "No Loans", "Loans"),  
 caption = "BOY and EOY Total Loans Outstanding By Sector",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

BOY and EOY Total Loans Outstanding By Sector

| Sector | No Loans | Loans |
| --- | --- | --- |
| Accommodation and food services | 64 | 392 |
| Administrative and support and waste management and remediation services | 44 | 588 |
| Agriculture, forestry, fishing and hunting | 28 | 189 |
| Arts, entertainment, and recreation | 36 | 331 |
| Construction | 125 | 1,194 |
| Educational services | 36 | 468 |
| Finance and insurance | 107 | 1,622 |
| Health care and social assistance | 442 | 2,797 |
| Information | 43 | 865 |
| Management of companies and enterprises | 17 | 281 |
| Manufacturing | 367 | 4,879 |
| Mining, quarrying, and oil and gas extraction | 21 | 209 |
| Other services (except public administration) | 127 | 823 |
| Professional, scientific, and technical services | 211 | 4,145 |
| Public administration | 4 | 22 |
| Real estate and rental and leasing | 45 | 658 |
| Retail trade | 163 | 1,440 |
| Transportation and warehousing | 101 | 762 |
| Utilities | 4 | 182 |
| Wholesale trade | 113 | 1,410 |

# Loan Leakage Ratio (LOAN\_LEAKAGE\_RATIO)

plans <- plans %>%  
 mutate(LOAN\_LEAKAGE\_RATIO = if\_else(TOTAL\_LOANS\_EOY == 0 & TOTAL\_ASSETS\_EOY == 0,   
 0,   
 round(TOTAL\_LOANS\_EOY / TOTAL\_ASSETS\_EOY, 4)))

# Loan Leakage Tier (LOAN\_LEAKAGE\_TIER)

plans <- plans %>%  
 mutate(LOAN\_LEAKAGE\_TIER = case\_when(LOAN\_LEAKAGE\_RATIO == 0 ~ "None",  
 TRUE ~ as.character(cut(LOAN\_LEAKAGE\_RATIO,  
 breaks = c(0, 0.05, 0.1, 0.15, 0.2, Inf),  
 labels = c("Minimal", "Low", "Moderate", "High", "Very High"),  
 right = FALSE))),  
 LOAN\_LEAKAGE\_TIER = factor(LOAN\_LEAKAGE\_TIER,  
 levels = c("None", "Minimal", "Low", "Moderate", "High", "Very High"),  
 ordered = TRUE))

# Retirement Adequacy Score (ADEQUACY\_SCORE, ADEQUACY\_IND, ADEQUACY\_LABEL)

This response variable is designed to classify retirement plans as adequate or inadequate based on a principled, multi-signal framework. Rather than relying on a single metric, it integrates four structural indicators of plan health:

Assets per Participant > $50,000 signals long-term saving capacity and participant wealth accumulation.

Asset Growth Rate > 5% reflects sustained engagement and financial momentum over time.

Average Contributions per Participant > $2,000 indicates meaningful participant deferral behavior and plan utilization.

Loan Leakage Ratio < 0.5 captures erosion risk due to participant borrowing; lower ratios suggest stronger adequacy.

Each condition is evaluated as a binary flag, and the adequacy score is the sum of conditions met. Plans meeting three or more criteria are classified as adequate.

plans <- plans %>%  
 mutate(ADEQUACY\_SCORE = (ASSETS\_PER\_PARTCP > 50000) +  
 (TOTAL\_ASSETS\_GROWTH\_RATE > 0.05) +  
 (AVERAGE\_CONTRIB\_PER\_PARTCP > 2000) +  
 (LOAN\_LEAKAGE\_RATIO < 0.5),  
 ADEQUACY\_IND = if\_else(ADEQUACY\_SCORE >= 3, 1L, 0L),  
 ADEQUACY\_LABEL = if\_else(ADEQUACY\_IND == 1L, "Adequate", "Inadequate"))

This approach offers several advantages:

Interpretability; Stakeholders can trace adequacy classification to specific, transparent thresholds.

Diagnostic clarity: Each component reflects a distinct dimension of plan health—accumulation, growth, engagement, and leakage.

Fairness-aware: By requiring multiple criteria, the model avoids overfitting to any one sector or plan type.

Model-ready: The binary label supports classification tasks, while the tiered label enables stakeholder previews and fairness overlays.

This response variable scaffolds ethical modeling, stakeholder trust, and reproducible diagnostics, all central to my analytic philosophy for this project.

# Finalize and Save Dataset

Order variables and save plans data frame with updates and features.

plans <- plans %>%  
 select(ADEQUACY\_LABEL,  
 ADEQUACY\_IND,  
 ADEQUACY\_SCORE,  
 ACK\_ID,  
 PLAN\_YEAR\_BEGIN\_DATE,  
 PLAN\_YEAR\_END\_DATE,  
 PLAN\_NAME,  
 PLAN\_EFFECTIVE\_DATE,  
 PLAN\_EFFECTIVE\_YEAR,  
 PLAN\_VINTAGE\_GROUP,  
 PLAN\_TYPE,  
 PLAN\_TYPE\_NAME,  
 SPONSOR\_NAME,  
 SPONSOR\_STATE,  
 SPONSOR\_EIN,  
 BUSINESS\_CODE,  
 INDUSTRY\_TITLE,  
 SECTOR\_CODE,  
 SECTOR\_TITLE,  
 SECTOR\_PARTCP\_SHARE,  
 TOTAL\_ACTIVE\_PARTCP\_BOY,  
 TOTAL\_ACTIVE\_PARTCP\_EOY,  
 TOTAL\_ACCBAL\_PARTCP\_BOY,  
 TOTAL\_ACCBAL\_PARTCP\_EOY,  
 PARTCP\_GROWTH\_RATE,  
 PARTCP\_GROWTH\_TIER,  
 TOTAL\_CONTRIB\_PARTCP\_BOY,  
 TOTAL\_CONTRIB\_PARTCP\_EOY,  
 CONTRIB\_PARTCP\_GROWTH\_RATE,  
 CONTRIB\_PARTCP\_GROWTH\_TIER,  
 AVERAGE\_CONTRIB\_PER\_PARTCP,  
 CONTRIB\_PER\_PARTCP\_TIER,  
 TOTAL\_CONTRIB\_EMPLR\_BOY,  
 TOTAL\_CONTRIB\_EMPLR\_EOY,  
 CONTRIB\_EMPLR\_GROWTH\_RATE,  
 CONTRIB\_EMPLR\_GROWTH\_TIER,  
 TOTAL\_LOANS\_BOY,  
 TOTAL\_LOANS\_EOY,   
 HAS\_OUTSTANDING\_LOANS,  
 LOAN\_LEAKAGE\_RATIO,  
 LOAN\_LEAKAGE\_TIER,  
 TOTAL\_ASSETS\_BOY,  
 TOTAL\_ASSETS\_EOY,  
 TOTAL\_ASSETS\_GROWTH\_RATE,  
 ASSETS\_PER\_PARTCP,  
 ASSETS\_PER\_PARTCP\_TIER)  
  
saveRDS(plans, "../data/plans\_final.rds")