Berlin Firms Comparison

# Load required libraries ----------------------------------------------------  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(gt)  
library(scales)

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

library(officer)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

# Load data ------------------------------------------------------------------  
df <- readRDS("/workspaces/rct25/data/generated/Orbis\_Berlin\_Data/orbis\_panel\_berlin.rds")  
  
# Prepare data ---------------------------------------------------------------  
df <- df %>%  
 filter(city\_native == "Berlin") %>%  
 mutate(  
 total\_assets = toas,  
 equity\_ratio = shfd / toas,  
 log\_total\_assets = log1p(toas),  
 group = if\_else(postcode == 10437, "Postal Code 10437", "Other Berlin Firms")  
 ) %>%  
 filter(!is.na(total\_assets), !is.na(equity\_ratio))  
  
# Summary statistics ---------------------------------------------------------  
summary\_stats <- df %>%  
 group\_by(group) %>%  
 summarise(  
 Mean\_Total\_Assets = mean(total\_assets, na.rm = TRUE),  
 SD\_Total\_Assets = sd(total\_assets, na.rm = TRUE),  
 Mean\_Equity\_Ratio = mean(equity\_ratio, na.rm = TRUE),  
 Median\_Equity\_Ratio = median(equity\_ratio, na.rm = TRUE),  
 SD\_Equity\_Ratio = sd(equity\_ratio, na.rm = TRUE),  
 n = n(),  
 .groups = "drop"  
 )  
  
# Extract values explicitly by group ------------------------------------------  
get\_stat <- function(var, group\_name) summary\_stats %>% filter(group == group\_name) %>% pull({{ var }})  
  
mean\_total\_assets\_10437 <- get\_stat(Mean\_Total\_Assets, "Postal Code 10437")  
mean\_total\_assets\_other <- get\_stat(Mean\_Total\_Assets, "Other Berlin Firms")  
  
mean\_equity\_ratio\_10437 <- get\_stat(Mean\_Equity\_Ratio, "Postal Code 10437")  
mean\_equity\_ratio\_other <- get\_stat(Mean\_Equity\_Ratio, "Other Berlin Firms")  
  
median\_equity\_ratio\_10437 <- get\_stat(Median\_Equity\_Ratio, "Postal Code 10437")  
median\_equity\_ratio\_other <- get\_stat(Median\_Equity\_Ratio, "Other Berlin Firms")  
  
sd\_equity\_ratio\_10437 <- get\_stat(SD\_Equity\_Ratio, "Postal Code 10437")  
sd\_equity\_ratio\_other <- get\_stat(SD\_Equity\_Ratio, "Other Berlin Firms")  
  
sd\_total\_assets\_10437 <- get\_stat(SD\_Total\_Assets, "Postal Code 10437")  
sd\_total\_assets\_other <- get\_stat(SD\_Total\_Assets, "Other Berlin Firms")  
  
n\_10437 <- get\_stat(n, "Postal Code 10437")  
n\_other <- get\_stat(n, "Other Berlin Firms")  
  
# Statistical tests ----------------------------------------------------------  
t\_total\_assets <- t.test(total\_assets ~ group, data = df, var.equal = FALSE)  
t\_equity\_ratio <- t.test(equity\_ratio ~ group, data = df, var.equal = FALSE)  
t\_sd\_equity\_ratio <- t.test(df$equity\_ratio[df$group == "Postal Code 10437"], df$equity\_ratio[df$group != "Postal Code 10437"], var.equal = FALSE)  
t\_sd\_total\_assets <- t.test(df$total\_assets[df$group == "Postal Code 10437"], df$total\_assets[df$group != "Postal Code 10437"], var.equal = FALSE)  
t\_median\_equity\_ratio <- wilcox.test(equity\_ratio ~ group, data = df)  
  
# Create results table -------------------------------------------------------  
results\_table <- tibble(  
 `Postal Code 10437` = c(  
 mean\_total\_assets\_10437,  
 sd\_total\_assets\_10437,  
 mean\_equity\_ratio\_10437,  
 median\_equity\_ratio\_10437,  
 sd\_equity\_ratio\_10437,  
 n\_10437  
 ),  
 `Other Berlin Firms` = c(  
 mean\_total\_assets\_other,  
 sd\_total\_assets\_other,  
 mean\_equity\_ratio\_other,  
 median\_equity\_ratio\_other,  
 sd\_equity\_ratio\_other,  
 n\_other  
 ),  
 Difference = c(  
 mean\_total\_assets\_10437 - mean\_total\_assets\_other,  
 sd\_total\_assets\_10437 - sd\_total\_assets\_other,  
 mean\_equity\_ratio\_10437 - mean\_equity\_ratio\_other,  
 median\_equity\_ratio\_10437 - median\_equity\_ratio\_other,  
 sd\_equity\_ratio\_10437 - sd\_equity\_ratio\_other,  
 NA  
 ),  
 `P-Value` = c(  
 t\_total\_assets$p.value,  
 t\_sd\_total\_assets$p.value,  
 t\_equity\_ratio$p.value,  
 t\_median\_equity\_ratio$p.value,  
 t\_sd\_equity\_ratio$p.value,  
 NA  
 ),  
 Significance = c(  
 case\_when(t\_total\_assets$p.value < 0.01 ~ "\*\*\*", t\_total\_assets$p.value < 0.05 ~ "\*\*", t\_total\_assets$p.value < 0.1 ~ "\*", TRUE ~ ""),  
 case\_when(t\_sd\_total\_assets$p.value < 0.01 ~ "\*\*\*", t\_sd\_total\_assets$p.value < 0.05 ~ "\*\*", t\_sd\_total\_assets$p.value < 0.1 ~ "\*", TRUE ~ ""),  
 case\_when(t\_equity\_ratio$p.value < 0.01 ~ "\*\*\*", t\_equity\_ratio$p.value < 0.05 ~ "\*\*", t\_equity\_ratio$p.value < 0.1 ~ "\*", TRUE ~ ""),  
 case\_when(t\_median\_equity\_ratio$p.value < 0.01 ~ "\*\*\*", t\_median\_equity\_ratio$p.value < 0.05 ~ "\*\*", t\_median\_equity\_ratio$p.value < 0.1 ~ "\*", TRUE ~ ""),  
 case\_when(t\_sd\_equity\_ratio$p.value < 0.01 ~ "\*\*\*", t\_sd\_equity\_ratio$p.value < 0.05 ~ "\*\*", t\_sd\_equity\_ratio$p.value < 0.1 ~ "\*", TRUE ~ ""),  
 ""  
 )  
)  
  
# Format results -------------------------------------------------------------  
results\_table\_formatted <- results\_table %>%  
 mutate(  
 rowname = c(  
 "Total Assets (Mean)",  
 "Total Assets (SD)",  
 "Equity Ratio (Mean)",  
 "Equity Ratio (Median)",  
 "Equity Ratio (SD)",  
 "Number of Firms (N)"  
 ),  
 `Postal Code 10437` = case\_when(  
 str\_detect(rowname, "Equity") ~ formatC(100 \* as.numeric(`Postal Code 10437`), format = "f", digits = 2, big.mark = ".", decimal.mark = ",") %>% paste0("%"),  
 str\_detect(rowname, "Total|Firms") ~ format(round(as.numeric(`Postal Code 10437`), 0), big.mark = ".", decimal.mark = ",", scientific = FALSE),  
 TRUE ~ as.character(`Postal Code 10437`)  
 ),  
 `Other Berlin Firms` = case\_when(  
 str\_detect(rowname, "Equity") ~ formatC(100 \* as.numeric(`Other Berlin Firms`), format = "f", digits = 2, big.mark = ".", decimal.mark = ",") %>% paste0("%"),  
 str\_detect(rowname, "Total|Firms") ~ format(round(as.numeric(`Other Berlin Firms`), 0), big.mark = ".", decimal.mark = ",", scientific = FALSE),  
 TRUE ~ as.character(`Other Berlin Firms`)  
 ),  
 Difference = case\_when(  
 str\_detect(rowname, "Equity") ~ formatC(100 \* as.numeric(Difference), format = "f", digits = 2, big.mark = ".", decimal.mark = ",") %>% paste0("%"),  
 str\_detect(rowname, "Total|Firms") ~ format(round(as.numeric(Difference), 0), big.mark = ".", decimal.mark = ",", scientific = FALSE),  
 TRUE ~ as.character(Difference)  
 ),  
 `P-Value` = case\_when(  
 is.na(`P-Value`) ~ "",  
 `P-Value` < 0.001 ~ "<0.001",  
 TRUE ~ formatC(`P-Value`, format = "f", digits = 3, decimal.mark = ",")  
 )  
 )  
  
# Create gt table ------------------------------------------------------------  
gt\_table <- results\_table\_formatted %>%  
 gt(rowname\_col = "rowname") %>%  
 tab\_header(  
 title = "Comparison of Firms in Postal Code 10437 vs. Other Berlin Firms",  
 subtitle = "Total Assets and Equity Ratios (Most Recent Year)"  
 ) %>%  
 cols\_label(  
 `Postal Code 10437` = "Postal Code 10437",  
 `Other Berlin Firms` = "Other Berlin Firms",  
 Difference = "Difference",  
 `P-Value` = "P-Value",  
 Significance = "Signif."  
 ) %>%  
 tab\_source\_note(  
 source\_note = "Note: Total Assets in EUR. Equity Ratio = Equity / Total Assets. P-values from Welch's t-test (for means and SDs) and Wilcoxon rank-sum test (for medians). Significance levels: \* p<0.1, \*\* p<0.05, \*\*\* p<0.01. N refers to the number of firms in each group. All p-values <0.001 are reported as <0.001. Sample includes all Berlin-based firm-year observations with non-missing data for the respective variables."  
 )  
  
gt\_table

Table 1: Comparison of Firms in Postal Code 10437 vs. Other Berlin Firms

Total Assets and Equity Ratios (Most Recent Year)

|  | Postal Code 10437 | Other Berlin Firms | Difference | P-Value | Signif. |
| --- | --- | --- | --- | --- | --- |
| Total Assets (Mean) | 2.602.573 | 18.147.925 | -15.545.352 | <0.001 | \*\*\* |
| Total Assets (SD) | 18.339.090 | 547.097.172 | -528.758.082 | <0.001 | \*\*\* |
| Equity Ratio (Mean) | -5.590,29% | -6.122,84% | 532,56% | 0,896 |  |
| Equity Ratio (Median) | 35,24% | 32,18% | 3,06% | 0,279 |  |
| Equity Ratio (SD) | 169.541,83% | 1.658.901,08% | -1.489.359,26% | 0,896 |  |
| Number of Firms (N) | 2.953 | 391.931 | NA |  |  |
| Note: Total Assets in EUR. Equity Ratio = Equity / Total Assets. P-values from Welch's t-test (for means and SDs) and Wilcoxon rank-sum test (for medians). Significance levels: \* p<0.1, \*\* p<0.05, \*\*\* p<0.01. N refers to the number of firms in each group. All p-values <0.001 are reported as <0.001. Sample includes all Berlin-based firm-year observations with non-missing data for the respective variables. | | | | | |

# Export to Word (.docx) -----------------------------------------------------  
ft\_table <- flextable(results\_table\_formatted %>% select(-Significance) %>% relocate(rowname) %>% mutate(across(everything(), as.character)))  
ft\_table <- autofit(ft\_table)  
  
doc <- read\_docx() %>%  
 body\_add\_par("Comparison of Firms in Postal Code 10437 vs. Other Berlin Firms", style = "heading 1") %>%  
 body\_add\_par("Total Assets and Equity Ratios (Most Recent Year)", style = "heading 2") %>%  
 body\_add\_flextable(ft\_table) %>%  
 body\_add\_par("Note: Total Assets in EUR. Equity Ratio = Equity / Total Assets. P-values from Welch's t-test (for means and SDs) and Wilcoxon rank-sum test (for medians). Significance levels: \* p<0.1, \*\* p<0.05, \*\*\* p<0.01. N refers to the number of firms in each group. All p-values <0.001 are reported as <0.001. Sample includes all Berlin-based firm-year observations with non-missing data for the respective variables.")  
  
print(doc, target = "firm\_comparison\_table.docx")