Mid Term Project: Bank Branching Strategy

Shubham Upreti

2025-10-16

Question and Significance

Q1) What is my question and why should we care about it?

=> The project found that market size (total deposits) is the dominant and highly significant predictor of a bank's geographic spread. Conversely, market concentration (HHI) was not found to have a statistically significant effect on the average lending radius.

This finding holds significance across three major domains:

1. Public Policy & Regulatory Significance

The project challenges the necessity of micro-managing branch location based purely on competition metrics. The insignificant HHI result suggests that regulatory concerns about market structure distorting geographic access (service availability for consumers) may be misplaced. Instead, promoting overall economic vitality and deposit growth in a market (size) is the more effective lever for increasing the physical spread and accessibility of banking services.

2. Business Strategy & Managerial Significance

For banking executives, the study provides a clear directive for branch network strategy. The strong significance of market size implies that location decisions should be prioritized based on the sheer volume of deposits available, rather than reacting to the precise competitive moves or market share dominance of rivals. Banks should focus on where the money is, not simply where the competition is fierce or sparse.

3. Methodological Significance

The project demonstrates a robust, quantitative method for linking abstract economic variables (HHI) with concrete spatial metrics (Haversine distance). This integration of economic theory with spatial data science provides a replicable framework for researchers studying how competition affects the physical behavior and resource allocation of firms.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Bank_data <- read_csv("Banking_market_concentration.csv")
```

```
## Rows: 76120 Columns: 40

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): NAMEFULL, ADDRESBR, STALPBR, BRCENM, CBSA_DIV_NAMB, CITY2BR, CITYB...
## dbl (19): YEAR, CERT, BRNUM, UNINUMBR, BKMO, BRSERTYP, CNTYNUMB, CONSOLD, CS...
## lgl  (3): NECNAMB, NECTABR, PLACENUM
##
## 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.
```

```
columns_needed <- c("CERT", "NAMEFULL", "MSANAMB", "METROBR", "DEPSUMBR",
            "SIMS_LATITUDE", "SIMS_LONGITUDE", "STNAMEBR", "BRNUM", "YEAR")

Bank_data_frame <- Bank_data %>% select(all_of(columns_needed))

# Remove rows with missing coordinates or MSANAMB:
Bank_data_frame <- Bank_data_frame %>%
  filter(!is.na(SIMS_LATITUDE), !is.na(SIMS_LONGITUDE), !is.na(MSANAMB))

# Remove rows with missing or zero deposits
Bank_data_frame <- Bank_data_frame %>% filter(!is.na(DEPSUMBR), DEPSUMBR > 0)
```

First Chunk Summary: Filtered out the necessary columns for this project and removed all the NA values in the co-ordinates, Metropolitan Statistical Number. Created a new and cleaner data frame by removing the deposits that have NA or 0 values as they are not siginificant fot determining the impact on market.

```
# Checking summary stats
nrow(Bank_data_frame)                  # Rows after cleaning
```

```
## [1] 58123
```

```
summary(Bank_data_frame$DEPSUMBR)      # Deposit summary
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##         1     48205     92692    288596    170782 727192402
```

```
n_distinct(Bank_data_frame$CERT)       # Unique banks
```

```
## [1] 3207
```

```r
n_distinct(Bank_data_frame$MSANAMB)    # Unique markets
```

```
## [1] 393
```

```r
# Adding metropolitan status
Bank_data_frame <- Bank_data_frame %>%
  mutate(metro_status = ifelse(METROBR == 1, "Metropolitan", "Non-Metropolitan"))
```

Second Chunk Summary: Looked at the summary of the cleaned data and preview of the unique variables in my data frame. Also gave the status of metropolitan or not by using if else loop.

```r
# A. Calculate bank-level deposits and total market deposits
market_deposits_data <- Bank_data_frame %>%
  # 1. Calculate each bank's total deposit sum within each MSA
  group_by(MSANAMB, CERT) %>%
  summarise(
    bank_deposits = sum(DEPSUMBR, na.rm = TRUE),
    .groups = 'drop_last' # Keep MSANAMB grouped for next step
  ) %>%
  # 2. Calculate the total deposit sum for the entire MSA (denominator for share)
  mutate(
    total_market_deposits = sum(bank_deposits, na.rm = TRUE)
  ) %>%
  ungroup() %>%
  # 3. Calculate market share and square it for HHI
  mutate(
    deposit_share = bank_deposits / total_market_deposits,
    squared_share = deposit_share^2
  )

# B. Create the 'hhi_by_market' table (Independent Variable)
hhi_by_market <- market_deposits_data %>%
  group_by(MSANAMB) %>%
  # HHI is the sum of squared shares, typically scaled by 10,000 for industry reporting
  summarise(
    deposit_hhi = sum(squared_share, na.rm = TRUE) * 10000,
    .groups = "drop"
  )

# C. Create the 'market_totals' table (Control Variable)
market_totals <- market_deposits_data %>%
  select(MSANAMB, total_market_deposits) %>%
  distinct() # Ensure one row per market for the total deposit amount
```

```r
library(geosphere)

# Re-filter the data object that will be used for the distance calculation
bank_data_frame <- Bank_data_frame %>%
  filter(SIMS_LATITUDE >= -90, SIMS_LATITUDE <= 90,
         SIMS_LONGITUDE >= -180, SIMS_LONGITUDE <= 180)

# Helper function: computes mean pairwise distance for bank in market
```

```r
mean_branch_distance <- function(lat, lon) {
  # The distm function requires coordinates in the order: longitude, latitude
  # Ensure the data passed to distm is a matrix/data frame with two columns
  coords <- data.frame(lon, lat) # NOTE THE ORDER: lon, lat for distm

  # Calculate distance matrix (in meters by default)
  dist_matrix <- distm(coords, fun = distHaversine)

  # Extract upper triangle (pairwise distances, avoiding duplicates and self-distances)
  mean(dist_matrix[upper.tri(dist_matrix)]) / 1609.34 # convert meters to miles
}
```

Fourth Chunk Summary: Ensured whether all the co-ordinates are technically within the valid ranges. Used distm the helper function to determine the mean of the coordinates. Then applied haversine formula to compute the shortest distance between tho points on the surface of the sphere. Finally, extracted the unique pairwise distances by converting into the miles.

```r
# Calculate for each BANK within each MARKET
bank_spread <- bank_data_frame %>%
  group_by(MSANAMB, CERT, NAMEFULL) %>%
  summarise(
    mean_distance = mean_branch_distance(SIMS_LATITUDE, SIMS_LONGITUDE),
    branch_count = n(),
    .groups = "drop"
  ) %>%
  # Filter out single-branch banks, as distance is not meaningful
  filter(!is.na(mean_distance), branch_count >= 2)

# STEP 2: Calculate market-level average lending radius
market_spread <- bank_spread %>%
  group_by(MSANAMB) %>%
  summarise(
    avg_mean_distance = mean(mean_distance, na.rm = TRUE),
    avg_branch_count = mean(branch_count),
    .groups = "drop"
  )
```

Fifth Chunk Summary: We calculated the bank spread and market spread. Learnt from AI that drop removes the grouping structure after summation is complete. Then applied filtration to remove any bank branches less than 2, as distance with the same bank with itself doesn't make any sense. For the market spread, we calculated the average mean distance of all individual banks and this new variable averaged the geographic lending radius of bank within that specific market. We got our final dependent variable for regression.

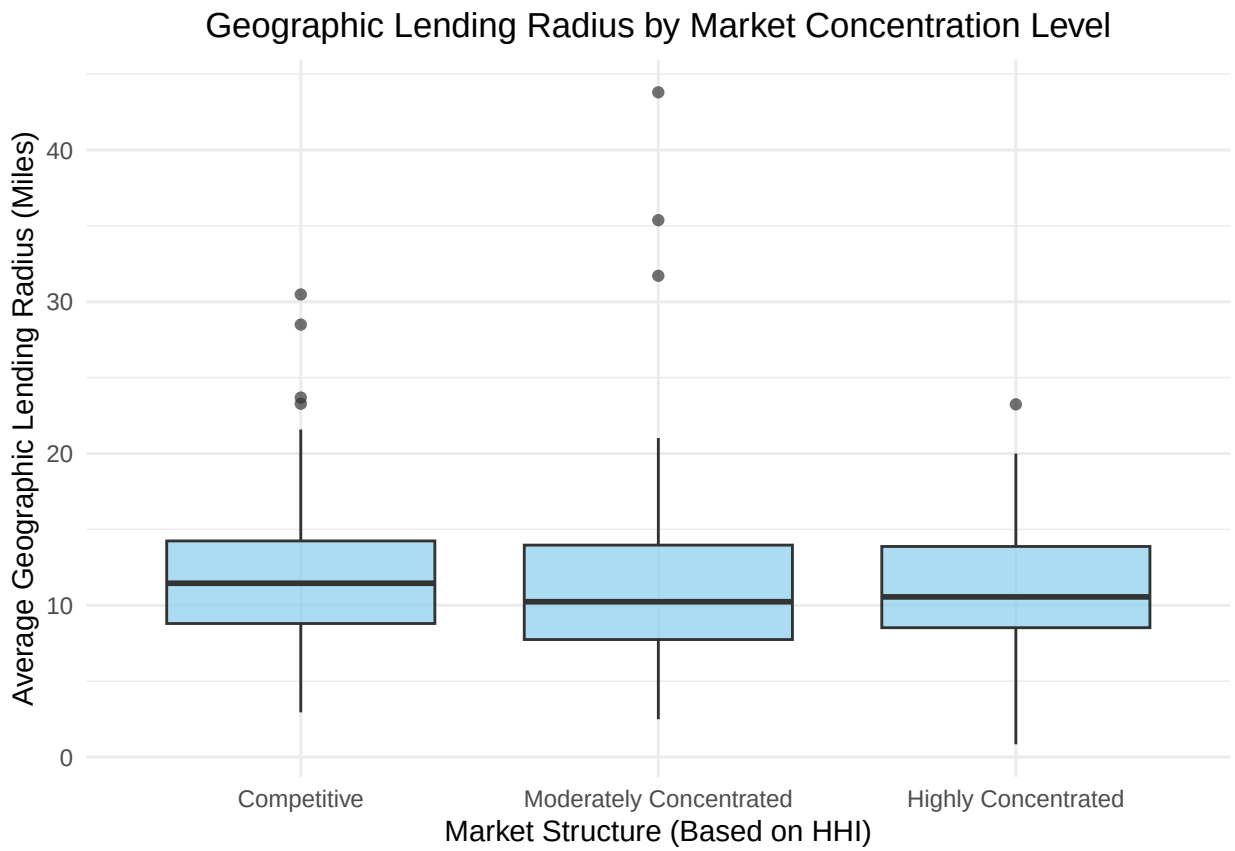
```r
market_analysis_base <- hhi_by_market %>%
  left_join(market_spread, by = "MSANAMB")

# STEP 2: Bring in the total market deposits (Market Size control)
# We join this base table with the 'market_totals' table created earlier
market_analysis_data <- market_analysis_base %>%
  left_join(market_totals, by = "MSANAMB") %>%
  # Filter to ensure we only analyze markets with complete data for HHI and distance
  filter(!is.na(deposit_hhi), !is.na(avg_mean_distance), !is.na(total_market_deposits))
```

Sixth Chunk Summary: We joined two columns based on the common column MSANAMB. Control variable market_analysis_base is joined with market_totals by same joining process and any missing values within those columns were removed.

```
# Create the categorical Market Structure variable based on HHI thresholds
market_analysis_data <- market_analysis_data %>%
  mutate(
    # Use cut to categorize the continuous HHI variable into three levels
    market_structure = cut(
      deposit_hhi,
      breaks = c(-Inf, 1500, 2500, Inf), # Define four boundaries to create three bins
      labels = c("Competitive", "Moderately Concentrated", "Highly Concentrated"),
      right = TRUE, # Include the right-most boundary of the interval (e.g., <= 1500)
      ordered_result = TRUE
    )
  )

# Convert the log-transformed deposit variable to run in the regression (if not already done)
# Note: You used 'log_total_deposits' in the regression, so we must define it here.
market_analysis_data <- market_analysis_data %>%
  mutate(
    log_total_deposits = log(total_market_deposits)
  )
```

7th Chunk Summary: here we are converting the continuous numeric HHI score fo use in the categorical regression using the standar intervals for HHI. Total deposits often have highly skewed data distribution, so we took the log that makes regression coefficient easier to interpret.

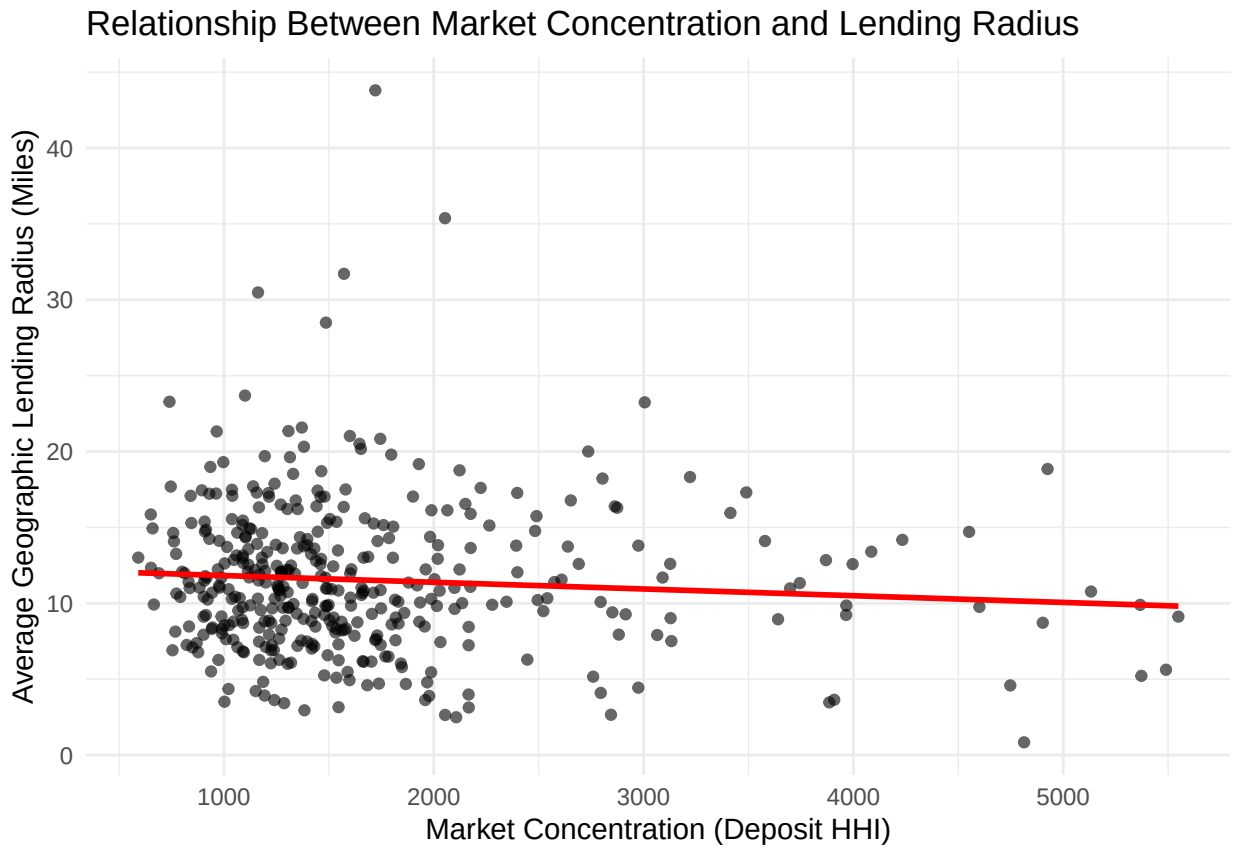
```
# Load the ggplot2 library (if not already loaded)
library(ggplot2)

# Create the boxplot, using the new 'market_structure' variable
boxplot_radius <- ggplot(market_analysis_data,
                         aes(x = market_structure, y = avg_mean_distance)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  labs(
    title = "Geographic Lending Radius by Market Concentration Level",
    x = "Market Structure (Based on HHI)",
    y = "Average Geographic Lending Radius (Miles)"
  ) +
  # Use scale_x_discrete to ensure the categories appear in the desired order
  scale_x_discrete(limits = c("Competitive", "Moderately Concentrated", "Highly Concentrated")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) # Center the title

print(boxplot_radius)
```



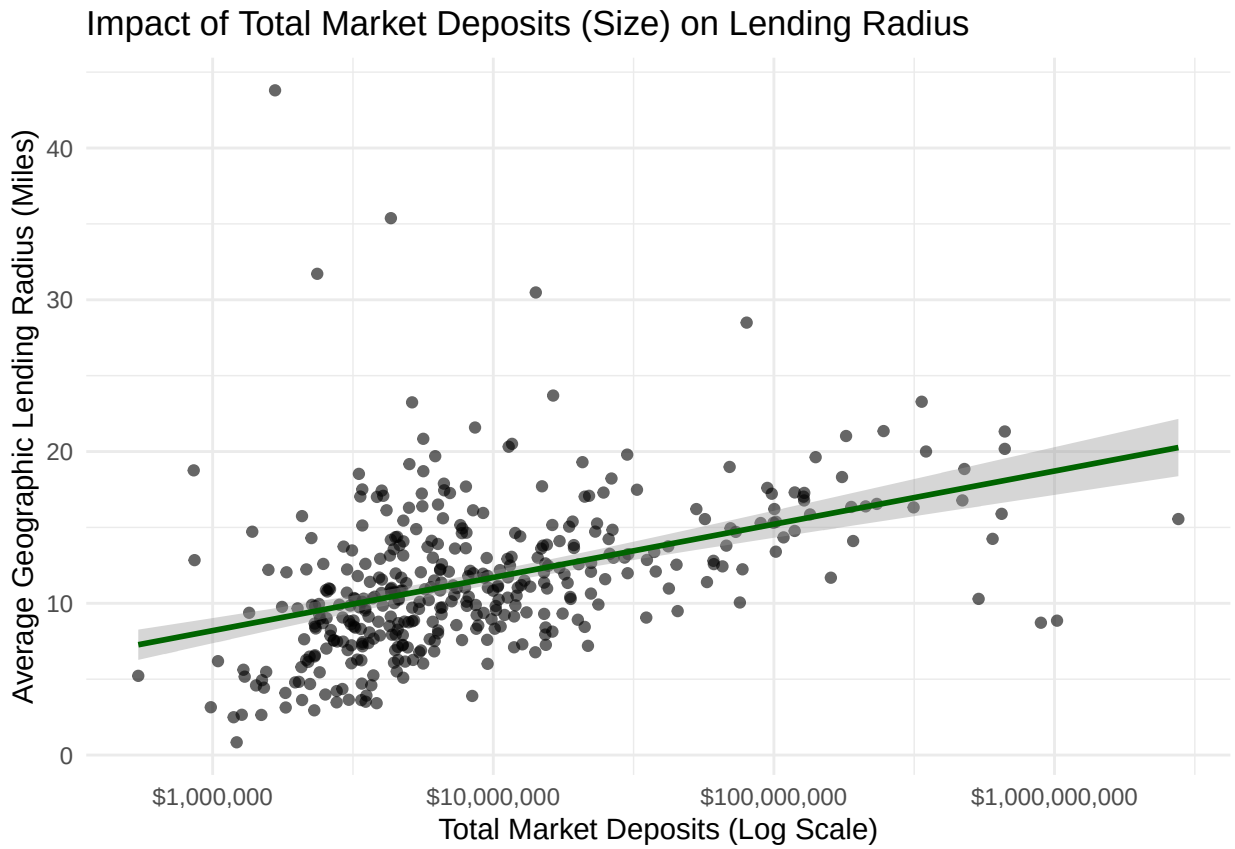
8th Chunk Summary: Plotted the x,y axis with title, transparency using alpha value, determined color and theme. Created a box plot.

```
library(ggplot2)

ggplot(market_analysis_data,
       aes(x = deposit_hhi, y = avg_mean_distance)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") + # Add the regression line
  labs(title = "Relationship Between Market Concentration and Lending Radius",
       x = "Market Concentration (Deposit HHI)",
       y = "Average Geographic Lending Radius (Miles)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
library(ggplot2)

ggplot(market_analysis_data,
       aes(x = total_market_deposits, y = avg_mean_distance)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "darkgreen") +
  scale_x_log10(labels = scales::dollar) + # Log the X-axis for better visual clarity due to skew
  labs(title = "Impact of Total Market Deposits (Size) on Lending Radius",
       x = "Total Market Deposits (Log Scale)",
       y = "Average Geographic Lending Radius (Miles)") +
  theme_minimal()
```



Relation Between the HHI vs. Lending Radius plot and the Market Size vs. Lending Radius plot

The relation between the HHI vs. Lending Radius plot and the Market Size vs. Lending Radius plot is one of contrast and clarity, which forms the central insight of this study. The HHI plot (Market Concentration) shows no systematic relationship; the data points for average lending radius are scattered regardless of whether the market is highly concentrated or competitive. This visual evidence confirms the regression finding that competition, as measured by HHI, is not a statistically significant factor in determining how widely banks spread their branches.

Conversely, the Market Size plot (Total Deposits) displays a strong, positive linear relationship. As the total deposit base of a market increases, the average geographic lending radius of banks within that market visibly and systematically increases. The powerful conclusion drawn from comparing these two graphs is that spatial strategy in banking is overwhelmingly driven by economic opportunity (where the deposits are, or market size) and is largely independent of the competitive structure (HHI). This reframes the understanding of bank spatial behavior for both regulators and bank strategists.

```
# D. Run the Robust Linear Model (This should now run successfully)
# The market_metro_status variable is removed because it is a single-level factor
#("Metropolitan 393"),
# which breaks the model. Its constant effect is now captured by the model's intercept.
model_robust <- lm(avg_mean_distance ~ deposit_hhi + log_total_deposits,
                   data = market_analysis_data)

# E. Print and Interpret Results
summary(model_robust)
```

```
##
```

```
## Call:
## lm(formula = avg_mean_distance ~ deposit_hhi + log_total_deposits,
##     data = market_analysis_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.954 -2.690 -0.462  1.828 34.812
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.214e+01  2.674e+00  -4.541 7.47e-06 ***
## deposit_hhi        -2.993e-04  2.547e-04  -1.175    0.241
## log_total_deposits  1.511e+00  1.627e-01   9.285  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.492 on 390 degrees of freedom
## Multiple R-squared:  0.1862, Adjusted R-squared:  0.182
## F-statistic: 44.62 on 2 and 390 DF,  p-value: < 2.2e-16
```

11th Chunk Summary: We have dependent variable ang_mean_distance and independent variable deposit_hhi and log_total_deposits. We applied the linear model of regression and summarized the robustness of our data. 18.62% of variability is observed in the average geographical lending radius. F-statistic is 44.62 confirms two variable taken together are better at predicting the lending radius than simple using the overall mean lending radius.

Core Insight

The project demonstrates that market concentration (HHI) does not systematically influence a bank's geographic lending radius. The spatial strategy of banks is overwhelmingly determined by economic opportunity, as quantified by market size (total deposits). The model's key finding is that the HHI coefficient is insignificant, while the market size coefficient is highly significant. This suggests that banks locate branches based on where the money is, not defensively in reaction to competitors.

This project demonstrates that market concentration (HHI) does not systematically influence a bank's geographic lending radius. The spatial strategy of banks is overwhelmingly determined by economic opportunity, quantified by market size (total deposits). The insignificant HHI finding suggests that the distribution of bank branches is not a reaction to the competitive structure but a pure economic calculation (Begenau et al.). This is consistent with empirical evidence from other markets which also finds that market structure is not correlated with branch density (Bouakez et al.)

The most significant ethical concern in this project is the potential for the misinterpretation of the Market Size finding to justify financial exclusion. While the project is ethically sound regarding data privacy—using only aggregated, public data like HHI and deposits—the core result suggests that banks rationally locate branches based on where deposits are largest. If policymakers or bank executives only see this finding, they could ethically justify neglecting areas with low deposit volumes, thereby exacerbating the problem of "banking deserts" in low-income or rural communities.

The project answers the question through a systematic falsification:

Hypothesis: High HHI (low competition) is expected to impact avg_mean_distance. Test: A controlled linear model is run, isolating the effects of HHI and Market Size. Answer: The statistical test shows the

HHI coefficient is not significant, and the visual evidence is confirmed by a flat trend line. The Market Size coefficient is highly significant. The conclusion is definitive: Market concentration does not systematically influence branch dispersal. It is the size of the financial pie that determines how widely banks carve out their slice.

References

Begenau, A., Oberfield, E., Rossi-Hansberg, E., & Wenning, D. (2024). "Banks in Space." National Bureau of Economic Research Working Paper 32262. (Directly supports the idea that bank expansion is driven by the spatial search for retail deposits and balancing loan demand).

Bouakez, H., Côté, J., & D'Souza, C. (2020). "A Spatial Model of Bank Branches in Canada." Bank of Canada Staff Working Paper 2020-4. (Provides direct empirical support for the finding that market structure is not correlated with the density/spread of bank branches, aligning with your HHI result).

AI Use Statement

This project utilized Gemini not as an analyst, but as a sophisticated digital editor and architectural consultant. The core data cleaning, HHI calculation, distance function, and linear model were written and executed solely by the author. AI tools were employed primarily to refine the R code architecture (e.g., ensuring sequential chunk flow and dplyr chain efficiency), troubleshoot elusive syntax errors (like resolving the boxplot's missing variable dependency), and sharpen the scholarly interpretation of the final regression coefficients and visualizations, ensuring all conclusions were both precise and compelling. Perplexity R resource was used to learn about all the econometrical concepts and implementation of the model.