GPEC 443: GIS & Spatial Data Analysis
Assignment 4

December 5, 2024

Question 1

The Relationship Between Homicides and Poverty in Mexico

This project explores the relationship between homicides and poverty in Mexico. Data preparation, variable calculations, Ordinary Least Squares (OLS) regression, Geographically Weighted Regression (GWR), and significance filtering were performed in R. Map creation and distance analysis were done using QGIS. The corresponding R code is attached throughout the report. Across all maps, the municipalities in grey signal no data. These were municipalities with zero homicides reported. All maps use EPSG:6362 which is the projected coordinate system for Mexico.

Table 1: OLS Regression Results for Log Homicide Rate on the Poverty Index

Term	**Estimate**	**Std. Error**	**t value**	**P-Value**
Intercept	3.25260	0.03780	86.06	<2e-16 ***
Poverty Index (indice)	0.33246	0.02903	11.45	<2e-16 ***

Residual Statistics	
Min	-2.4702
1Q	-0.6347
Median	-0.0749
3Q	0.5455
Max	3.9491

Model Fit Statistics	
Residual Standard Error	0.9373 (df = 1441)
Multiple R-squared	0.08343
Adjusted R-squared	0.08279
F-statistic	131.2 on 1 and 1441 DF
P-value of F-statistic	< 2.2e-16

OLS Regression Analysis

This Ordinary Least Squares (OLS) regression analysis investigates the relationship between the poverty index and the log-transformed homicide rate. The coefficients were scaled to the standard deviation of the poverty index. The coefficient for the poverty index is 0.33. A one-standard-deviation increase in the poverty index corresponds to a 32.67% increase in the homicide rate. Because the standard deviation of the poverty index is 0.85, an increase in one standard deviation will tend to result in a homicide rate that is 28% higher. The model explains approximately 8.34% of the variance in the log-transformed homicide rate. This suggests that other factors contribute significantly to the variation. The F-statistic of 131.17 ($p < 2.2 \times 10^{-16}$) confirms that the relationship between the poverty index and homicide rate is statistically significant. Although the poverty index is a significant predictor, the low R^2 value indicates that it is not the sole determinant of homicide rates.

Question 1 Code:

```
library(sf)
library(spgwr)
library(dplyr)
library(ggplot2)
library(cowplot)
library(broom)

# ==============================
# Question 1: OLS Regression
# ==============================

# Load the dataset
data <- read.csv("/Users/.../Documents/GIS/Assignments/assignment4/Assignment4Data/mex_data.csv")
df <- data  # Create a working copy of the dataset

# Filter out rows with zero homicides
df <- df[df$homicides > 0, ]

# Calculate the homicide rate per 100,000 population
df$homicide_rate <- (df$homicides / df$population) * 100000

# Create a new column for the log-transformed homicide rate
df$log_homicide_rate <- log(df$homicide_rate)

# Perform OLS regression of log(homicide rate) on poverty index
model <- lm(log_homicide_rate ~ indice, data = df)

# Print the full summary of the regression model
model_summary <- summary(model)
print(model_summary)

# Extract and interpret the coefficient for 'indice'
coefficient <- coef(model)["indice"]  # Extract the coefficient for the poverty index
```

```r
percentage_change <- (exp(coefficient) - 1) * 100  # Convert to percentage change

cat("Interpretation: A one-unit increase in the poverty index is associated with a",
    round(percentage_change, 2), "% increase in the homicide rate.\n")

# Compute the standard deviation of the poverty index
sd_indice <- sd(df$indice, na.rm = TRUE)

# Calculate the effect of a one standard deviation increase in poverty index
effect_sd <- exp(coefficient * sd_indice) - 1

cat("A one standard deviation increase in the poverty index is associated with a",
    round(effect_sd * 100, 2), "% increase in the homicide rate.\n")

# Summary statistics for numeric variables
numeric_cols <- sapply(df, is.numeric)
summary_stats <- sapply(df[, numeric_cols], function(x) {
  c(
    mean = mean(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    min = min(x, na.rm = TRUE),
    max = max(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE)
  )
})
summary_stats <- t(summary_stats)
print(summary_stats)

# Create a tidy table of regression results
regression_table <- tidy(model)

# Print the table
print(regression_table)

# Export the updated dataset for reference
write.csv(df, "log_homicides.csv", row.names = FALSE)
```

Table 2: Summary Statistics for Key Variables

Variable	Mean	SD	Min	Max	Median
State	18.33	7.44	1.00	32.00	19.00
Municipality (Mun)	86.91	115.31	1.00	570.00	45.00
Homicides	9.72	47.03	1.00	1398.00	2.00
Population	51238.77	135589.00	250.00	1696609.00	16753.00
Poverty Index (Indice)	-0.99	0.85	-2.45	2.01	-1.14
Homicide Rate (per 100k)	32.91	55.20	1.06	728.16	16.81
Log Homicide Rate	2.92	0.98	0.06	6.59	2.82

Question 2

Figure 1:

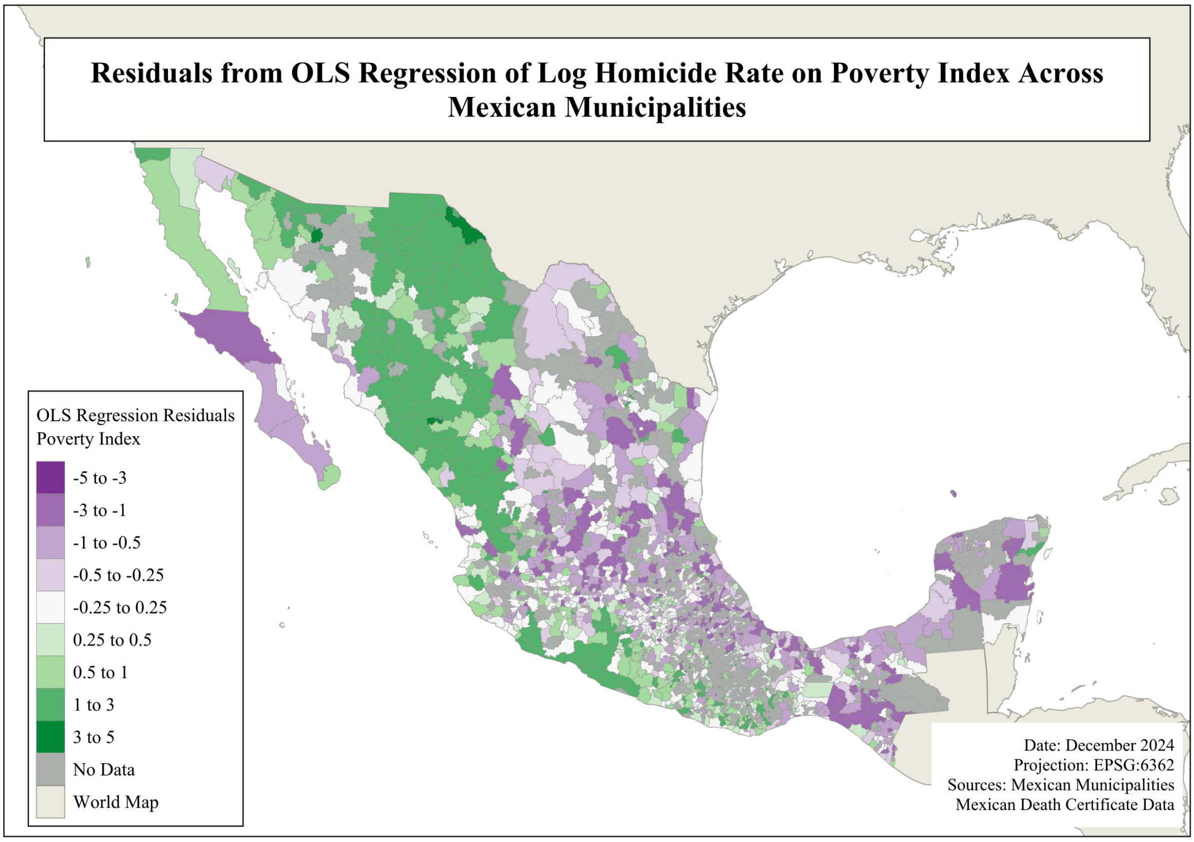

Figure 1 Interpretation:

Figure 1 displays a map of the residuals from an ordinary least squares (OLS) regression of the log-transformed homicide rate on the poverty index across Mexican municipalities. Residuals represent the differences between observed and predicted values. Green shades indicate municipalities where the model underpredicts homicide rates and purple shades indicate areas where the model overpredicts homicide rates. Negative residuals are when the observed homicide rate is lower than expected compared to the poverty index. High poverty rates typically predict high homicide rates, but in these municipalities the actual homicide rates are low. Positive residuals are under-predicted areas where the observed homicide rate is higher than expected. This suggests that there are factors beyond poverty are contributing to the homicide rate. Residuals close to zero indicate municipalities where the model performs well. In these municipalities, the predicted homicide rates closely align with the observed values, reflecting good predictive power. Municipalities with no data or zero homicides are shown in gray. The spatial distribution highlights regional disparities in how well the poverty index predicts homicide rates, suggesting potential regional or local factors influencing homicides beyond poverty. The majority of the negative residuals are concentrated in the North West and West Coast of Mexico, while most of the positive residuals are clustered around the East coast.

Question 2 Code:

```r
# ================================
# Question 2: OLS Residuals
# ================================

# Load the shapefile for municipalities
shapefile <- st_read("/Users/.../Documents/GIS/Assignments/
  assignment4/Assignment4Data/municipalities.shp")

# Create a join key in shapefile
shapefile$join_key <- paste(shapefile$CVE_ENT, shapefile$CVE_
  MUN, sep = "")

# Join with the dataset containing the log homicide rate
spatial_data <- shapefile %>%
  left_join(df, by = "join_key")

# Reproject to EPSG:6362 (a metric CRS for Mexico)
spatial_data <- spatial_data %>%
  st_transform(6362)

# Filter for rows with no missing values in key columns
spatial_data <- spatial_data %>%
  filter(!is.na(log_homicide_rate) & !is.na(indice))
```

Question 3

Figure 2:

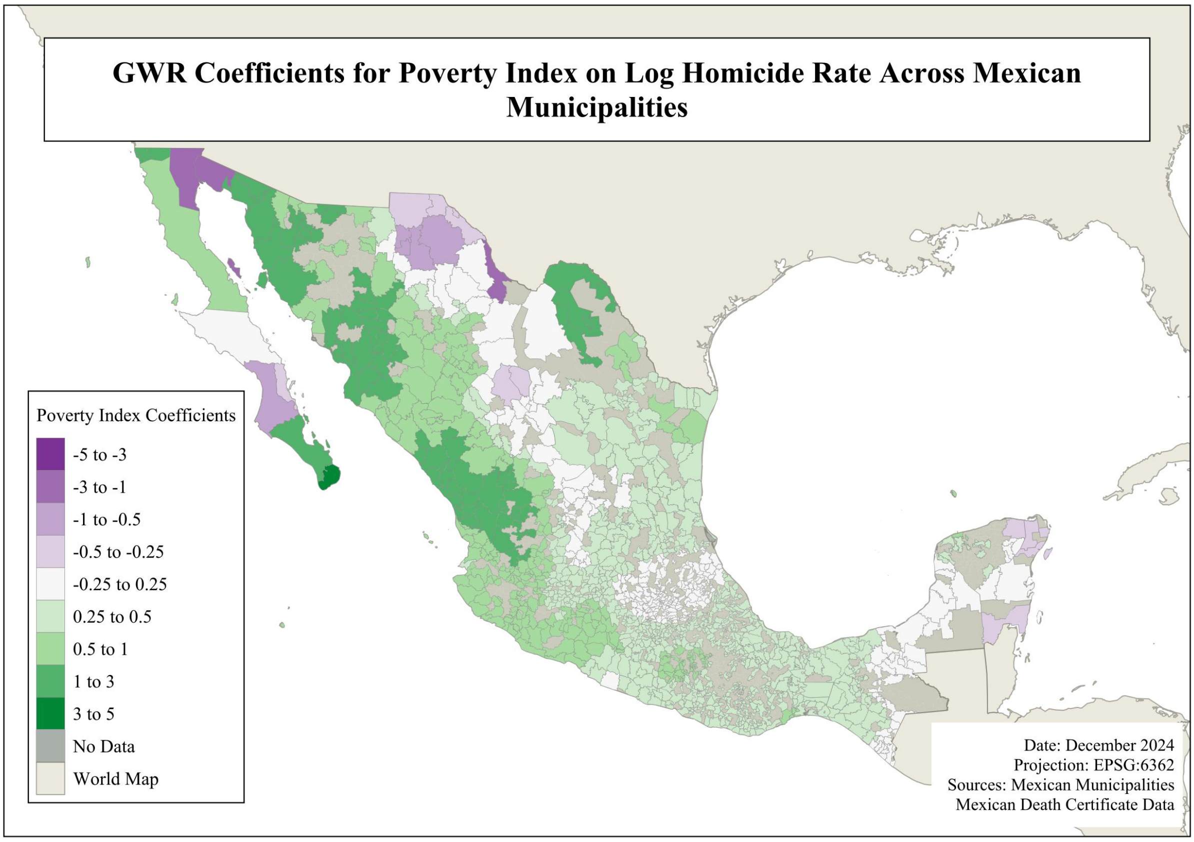

Figure 2 Interpretation:

Figure 2 illustrates geographically weighted regression (GWR) coefficients for the relationship between the poverty index and the log-transformed homicide rate across Mexican municipalities. Areas with high positive coefficients (shown in green) indicate municipalities where poverty is strongly associated with higher homicide rates, suggesting that poverty is a key predictor of homicides in these regions. Areas with negative coefficients (shown in purple) indicate the opposite, where poverty does not predict homicide rates accurately. In contrast, areas with coefficients near zero (shown in white) represent municipalities with no discernible relationship between poverty and homicide rates. Municipalities with no data are shaded gray. The municipalities with coefficients close to zero are clustered in the center corridor of the country. The few municipalities with the lowest coefficients are right along the US border. The municipalities with the highest coefficients are clustered along the west coast of Northern Mexico.

Question 3 Code:

```
# ==========================================
# Question 3: GWR Analysis
# ==========================================

```

```
    # Convert the spatial data to a Spatial* object for spgwr
spatial_points <- as(spatial_data, "Spatial")

# Determine the optimal bandwidth for GWR
bw <- gwr.sel(log_homicide_rate ~ indice, data = spatial_points)

# Run the GWR model with standard errors enabled
gwr_out <- gwr(log_homicide_rate ~ indice,
               data = spatial_points,
               bandwidth = bw,
               hatmatrix = TRUE,
               se.fit = TRUE)

# Extract the GWR results and convert to an sf object
gwr_results <- gwr_out$SDF
gwr_results$join_key <- spatial_data$join_key  # Add join_key for further analysis
gwr_sf <- st_as_sf(gwr_results)  # Convert to sf
```

Question 4

Figure 3:

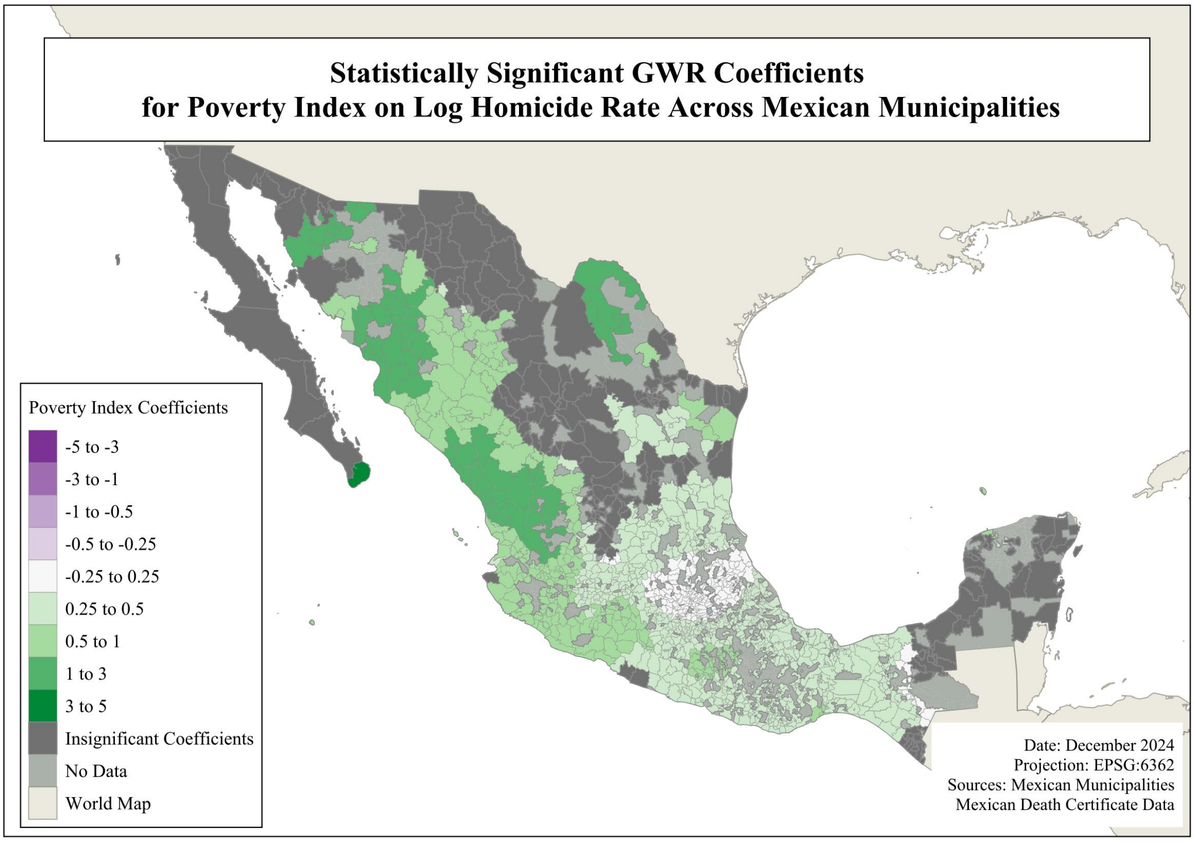

Figure 3 Interpretation:

The map displays only the statistically significant coefficients from the geographically weighted regression (GWR). Insignificant coefficients are shown in dark gray, and no data values (municipalities with zero homicides) are shown in light grey. This visualization emphasizes regional variation in the significant relationship between poverty and homicide rates, providing insight into where poverty has the strongest predictive power. Municipalities in Baja California and Central Mexico do not have significant coefficients. There are no municipalities with negative coefficients on this map. Municipalities with the highest significant coefficients are clustered around the West Coast.

Critical Value Calculation

To highlight statistically significant coefficients at an alpha level of 0.1 (90% confidence level), municipalities were filtered based on their t-statistics. Municipalities with absolute t-statistics exceeding the critical value of 1.645 were considered significant. The critical value was determined using the t-distribution with 1441 degrees of freedom, calculated as $n - k - 1$, where $n = 1443$ and $k = 1$. With such a large number of degrees of freedom, the t-distribution approximates the standard normal distribution, making 1.645 the appropriate threshold for significance. This approach ensures that the map includes only municipalities where the relationship between the poverty index and the log homicide rate is statistically significant at the 90% confidence level.

Question 4 Code:

```r
# ==========================================
# Question 4: Statistically Significant Mapping
# ==========================================

# Calculate t-statistics for poverty index coefficients
gwr_sf$t_indice <- gwr_sf$indice / gwr_sf$indice_se

# Add a column to classify municipalities as significant or insignificant
gwr_sf$significance <- ifelse(abs(gwr_sf$t_indice) >= critical_t, "Significant", "Insignificant")

# Filter for significant coefficients (90% confidence level)
gwr_sf_significant <- gwr_sf %>%
  filter(abs(t_indice) >= critical_t)

# Save the significant coefficients as a shapefile for mapping
st_write(gwr_sf_significant, "/Users/.../Documents/GIS/Assignments/assignment4/gwr_significant_coefficients.shp", delete_dsn = TRUE)
```

Question 5

Figure 4 Interpretation:

The scatter plot illustrates the relationship between significant poverty coefficients and the log-transformed distance to the US border. Municipalities closer to the border (lower x-axis values) tend to exhibit higher significant poverty coefficients, indicating a stronger relationship between poverty and log homicide rates in these areas. This suggests that poverty has a more pronounced impact on homicide rates near the border. As the distance from the border increases (higher x-axis values), poverty coefficients become smaller and cluster more tightly around moderate values, indicating a weaker and more consistent relationship between poverty and homicide rates in regions farther from the border. Additionally, a few outliers with exceptionally high coefficients near the border highlight localized areas where poverty plays a particularly strong role in influencing homicide rates.

Municipalities near the border may experience unique socio-economic dynamics, such as heightened migration pressures, economic disparities, and cross-border crime, which amplify the impact of poverty on violence. In contrast, municipalities farther from the border show a more uniform relationship and potentially less direct influence from border-related factors. Overall, the plot underscores the importance of considering spatial context when analyzing the relationship between poverty and homicide rates.

Figure 4:

Scatterplot of Significant Coefficients and Log Distance to Border

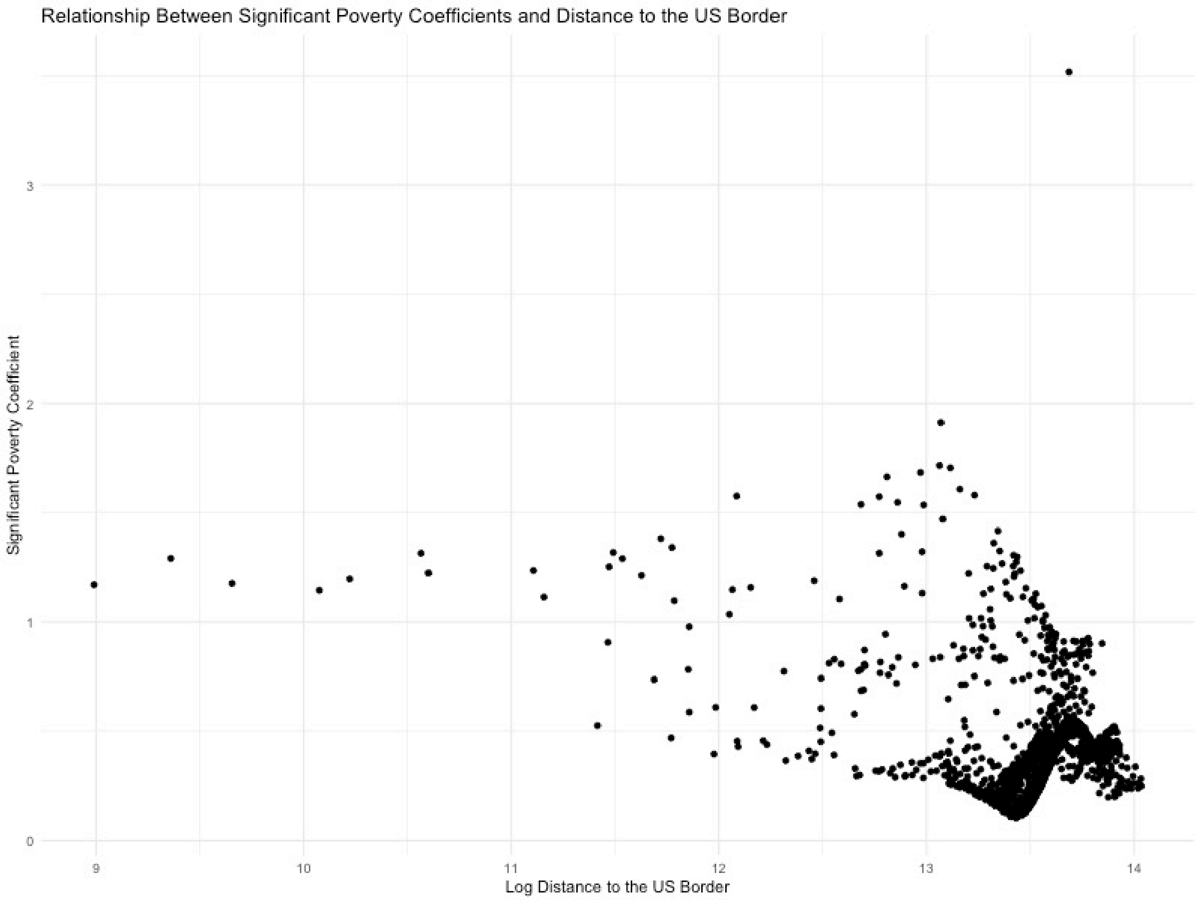

Table 3: OLS Regression Results: Poverty Index (indice) on Log Distance to the US Border

Term	**Estimate**	**Std. Error**	**t Value**	**P-Value**
Intercept	3.19807	0.22561	14.18	< 2e-16 ***
Log Distance	-0.20303	0.01672	-12.14	< 2e-16 ***

Model Fit Statistics	
Residual Standard Error	0.2731 (df = 1242)
Multiple R-squared	0.1061
Adjusted R-squared	0.1053
F-statistic	147.4 on 1 and 1242 DF
P-value of F-statistic	< 2.2e-16

Interpretation of Regression Results

The table summarizes the results of an ordinary least squares (OLS) regression analyzing the relationship between the poverty index (indice) and the log-transformed distance to the US border. The coefficient for log distance is -0.20303 ($p < 0.001$), indicating that for each unit increase in log distance to the US border, the poverty index decreases by 0.203 units on average. This negative relationship suggests that municipalities farther from the US border tend to have lower poverty levels.

The intercept of 3.19807 ($p < 0.001$) represents the estimated poverty index when the log distance is zero. The model explains 10.61% of the variance in the poverty index ($R^2 = 0.1061$), indicating that while log distance is a statistically significant predictor, other factors likely contribute to variations in poverty levels. The residual standard error of 0.2731 reflects the average deviation of observed poverty index values from those predicted by the model. Overall, the results highlight a statistically significant association between distance from the border and poverty levels.

Question 5 Code:

```
  # ==============================
  # Question 5: Distance Analysis
  # ==============================

# Load the municipality shapefile with distances to the US border
municipalities_with_distances <- st_read("/Users/.../Documents/GIS/Assignments/assignment4/analysis/municipalities_with_distances.shp")

# Reproject municipalities_with_distances to EPSG:6362
municipalities_with_distances <- st_transform(municipalities_with_distances, 6362)

```

```r

# Join the HubDist column to the significant GWR results
gwr_sf_significant <- gwr_sf_significant %>%
  left_join(municipalities_with_distances_df %>% select(join_
     key, HubDist), by = "join_key")

# Extract necessary variables and log-transform distances
municipalities_with_distances <- gwr_sf_significant %>%
  select(indice, HubDist) %>%
  filter(!is.na(HubDist))
municipalities_with_distances$log_distance <- log(
    municipalities_with_distances$HubDist)


# Run the regression of significant GWR coefficients on log
    distance
regression_model <- lm(indice ~ log_distance, data =
    municipalities_with_distances)
summary(regression_model)


# Save the scatter plot as a JPEG file
jpeg("scatterplot_log_distance_vs_indice.jpeg", width = 800,
    height = 600)

# Create the scatter plot
ggplot(municipalities_with_distances, aes(x = log_distance, y
    = indice)) +
  geom_point() +
  labs(
    title = "Relationship Between Significant Poverty
        Coefficients and Distance to the US Border",
    x = "Log Distance to the US Border",
    y = "Significant Poverty Coefficient"
  ) +
  theme_minimal()

# Close the graphics device
dev.off()
```