Census Regression Project

Jarius Hamid

2024-09-17

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(xtable)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)

# Compute summary statistics
table <- describe(tract_data)

# Convert the summary statistics to a data frame
table_df <- as.data.frame(table)

# Remove unnecessary columns and rename them
table_df <- table_df[, c("mean", "sd", "median", "range", "skew", "kurtosis")]
names(table_df) <- c("Mean", "SD", "Median", "Range", "Skew", "Kurtosis")

# Add a column for variable names
table_df$Variable <- rownames(table_df)

# Reorder columns
describe(table_df)
```

```
##           vars  n     mean       sd median trimmed   mad    min       max
## Mean         1 22  3631.24 13683.04  28.63  419.63 40.97 -81.85  64266.14
## SD           2 22  1858.37  5491.86   9.66  276.97 14.13   0.00  23202.54
## Median       3 22  3125.11 12531.72  27.39  340.68 39.40 -81.86  59015.75
## Range        4 22 11369.91 33647.15  53.16 1293.36 77.81   0.00 131764.75
## Skew         5 20     0.96     1.71   0.89    0.94  1.84  -2.34      3.74
## Kurtosis     6 20     4.75     6.81   1.38    3.74  2.58  -0.79     20.61
## Variable*    7 22    11.50     6.49  11.50   11.50  8.15   1.00     22.00
##                 range skew kurtosis      se
## Mean         64347.99 3.94    14.53 2917.23
## SD           23202.54 2.99     8.07 1170.87
## Median       59097.61 4.02    14.99 2671.77
## Range       131764.75 2.75     6.20 7173.60
## Skew             6.09 0.18    -1.01    0.38
## Kurtosis        21.41 1.06    -0.45    1.52
## Variable*       21.00 0.00    -1.36    1.38
```

```r
# Generate LaTeX table
#latex_table <- xtable(table_df, caption = "Summary Statistics", label = "tab:summary_statistics")

# Print LaTeX code
#print(latex_table, type = "latex", include.rownames = FALSE, booktabs = TRUE)
```

VISUALS

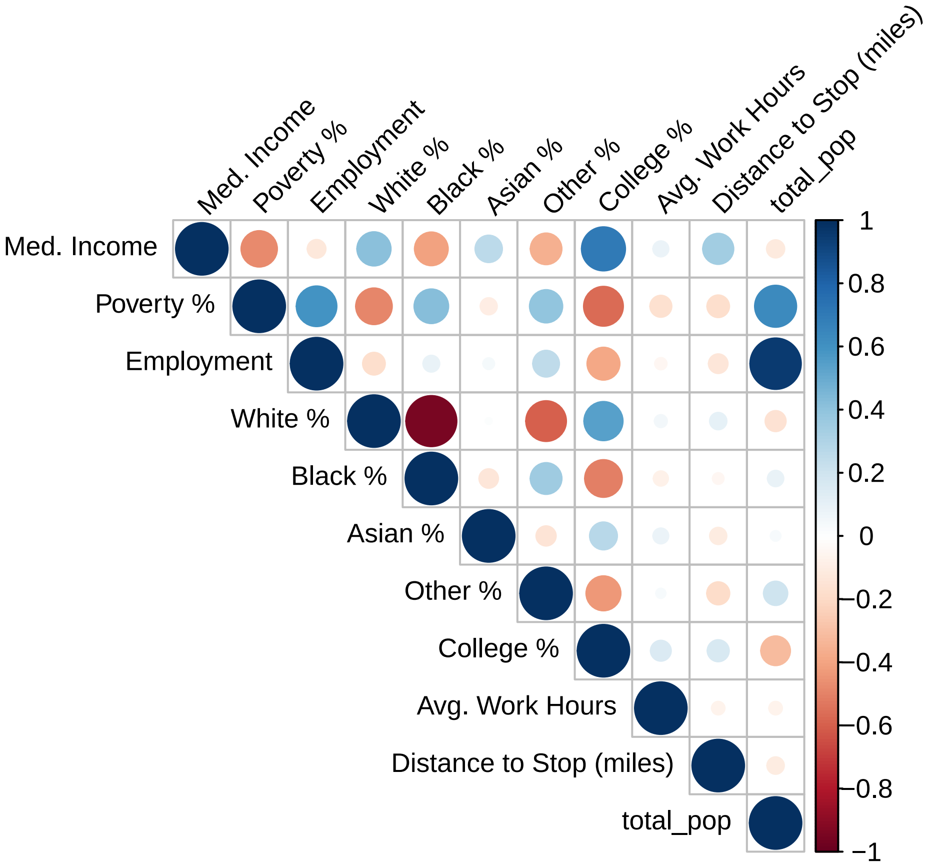
```r
# Load necessary libraries
# Select the relevant columns for correlation analysis
cor_data <- tract_data[, c("med_householdincome", "poverty_total", "employment.total",
                           "white_prct", "black_prct", "asian_prct", "other_prct",
                           "college_prct", "avg_workhours", "distance_from_closest_stop_miles", "total_p

# Shorten column names
colnames(cor_data) <- c("Med. Income", "Poverty %", "Employment",
                        "White %", "Black %", "Asian %", "Other %",
                        "College %", "Avg. Work Hours", "Distance to Stop (miles)","total_pop")

# Compute the correlation matrix
cor_matrix <- cor(cor_data, use = "complete.obs")

# Print the correlation matrix
print(cor_matrix)
```

```
##                            Med. Income  Poverty %  Employment     White %
## Med. Income                 1.00000000 -0.4740038 -0.12077539  0.41733408
## Poverty %                  -0.47400378  1.0000000  0.59409866 -0.48040160
## Employment                 -0.12077539  0.5940987  1.00000000 -0.17973704
## White %                     0.41733408 -0.4804016 -0.17973704  1.00000000
## Black %                    -0.40634404  0.4282422  0.09630535 -0.94347791
## Asian %                     0.26468248 -0.0999783  0.04512880  0.01425825
## Other %                    -0.35629846  0.3939630  0.25271781 -0.59663592
## College %                   0.70026247 -0.5695919 -0.38640962  0.54947634
## Avg. Work Hours             0.08667611 -0.1682983 -0.04969641  0.05569615
## Distance to Stop (miles)    0.34314730 -0.1784229 -0.13305873  0.10285681
## total_pop                  -0.11372220  0.6351349  0.95683452 -0.15333452
##                                Black %     Asian %     Other %  College %
## Med. Income                -0.40634404  0.26468248 -0.35629846  0.7002625
## Poverty %                   0.42824219 -0.09997830  0.39396303 -0.5695919
## Employment                  0.09630535  0.04512880  0.25271781 -0.3864096
## White %                    -0.94347791  0.01425825 -0.59663592  0.5494763
## Black %                     1.00000000 -0.13124352  0.35908669 -0.5087550
## Asian %                    -0.13124352  1.00000000 -0.14080189  0.2747686
## Other %                     0.35908669 -0.14080189  1.00000000 -0.4369937
## College %                  -0.50875504  0.27476860 -0.43699372  1.0000000
## Avg. Work Hours            -0.07735642  0.08506950  0.03295925  0.1510042
## Distance to Stop (miles) -0.04431720 -0.10093935 -0.18518514  0.1691971
## total_pop                  0.09097806  0.03651585  0.20697278 -0.3133628
##                            Avg. Work Hours Distance to Stop (miles)   total_pop
## Med. Income                     0.08667611               0.34314730 -0.11372220
## Poverty %                      -0.16829831              -0.17842288  0.63513492
## Employment                     -0.04969641              -0.13305873  0.95683452
## White %                         0.05569615               0.10285681 -0.15333452
## Black %                        -0.07735642              -0.04431720  0.09097806
## Asian %                         0.08506950              -0.10093935  0.03651585
## Other %                         0.03295925              -0.18518514  0.20697278
## College %                       0.15100415               0.16919708 -0.31336284
## Avg. Work Hours                 1.00000000              -0.06086333 -0.06146735
## Distance to Stop (miles)       -0.06086333               1.00000000 -0.10410035
## total_pop                      -0.06146735              -0.10410035  1.00000000
```

```r
# Visualize the correlation matrix with shortened labels
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45,
         tl.cex = 0.8)   # Optionally adjust text size with tl.cex
```



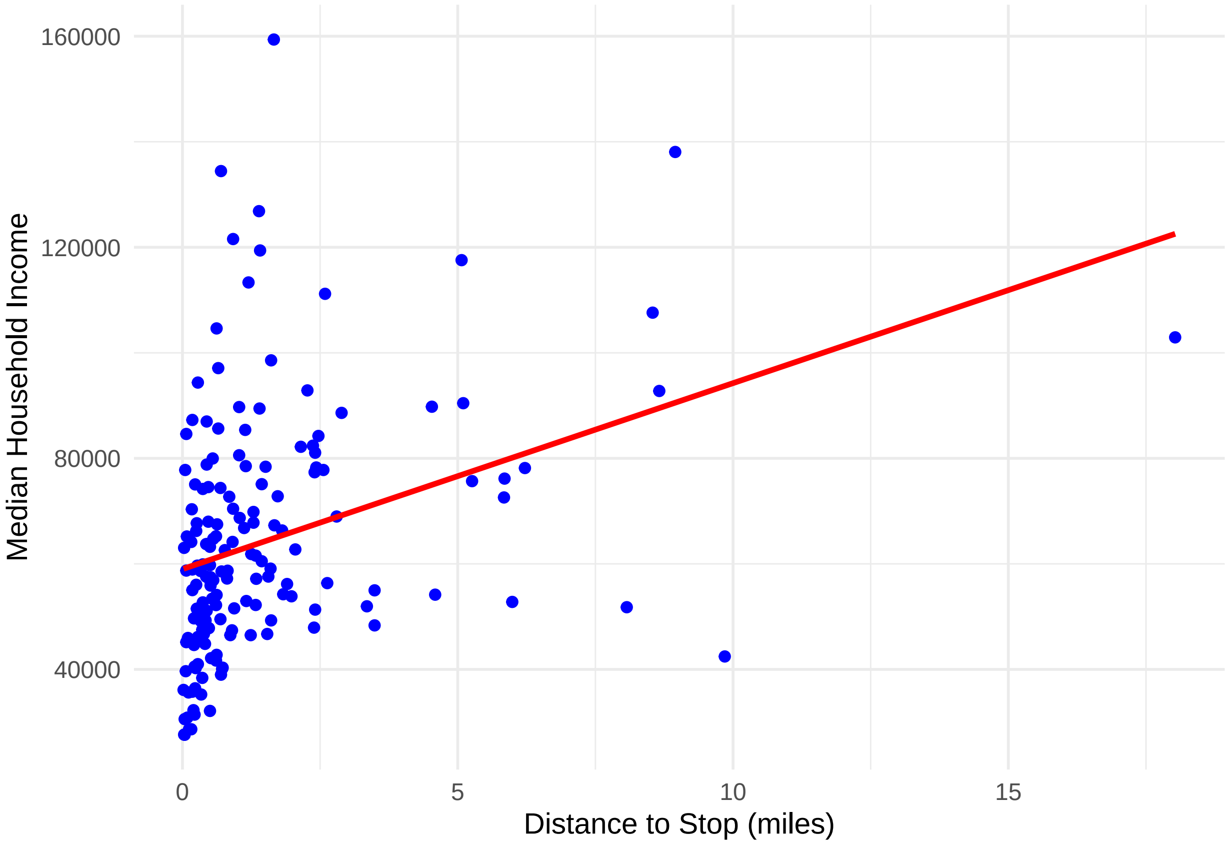
```r
# Load necessary library
library(ggplot2)
# Load necessary package
library(gridExtra)
```

```r
# Create individual plots
plot1 <- ggplot(tract_data, aes(x = distance_from_closest_stop_miles, y = med_householdincome)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
       x = "Distance to Stop (miles)",
       y = "Median Household Income") +
  theme_minimal()

plot2 <- ggplot(tract_data, aes(x = distance_from_closest_stop_miles, y = poverty_prct)) +
  geom_point(color = "green") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
       x = "Distance to Stop (miles)",
       y = "Poverty Percentage") +
  theme_minimal()

plot3 <- ggplot(tract_data, aes(x = distance_from_closest_stop_miles, y = employment_prct)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
       x = "Distance to Stop (miles)",
       y = "Employment Percentage") +
  theme_minimal()

print(plot1)
```

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

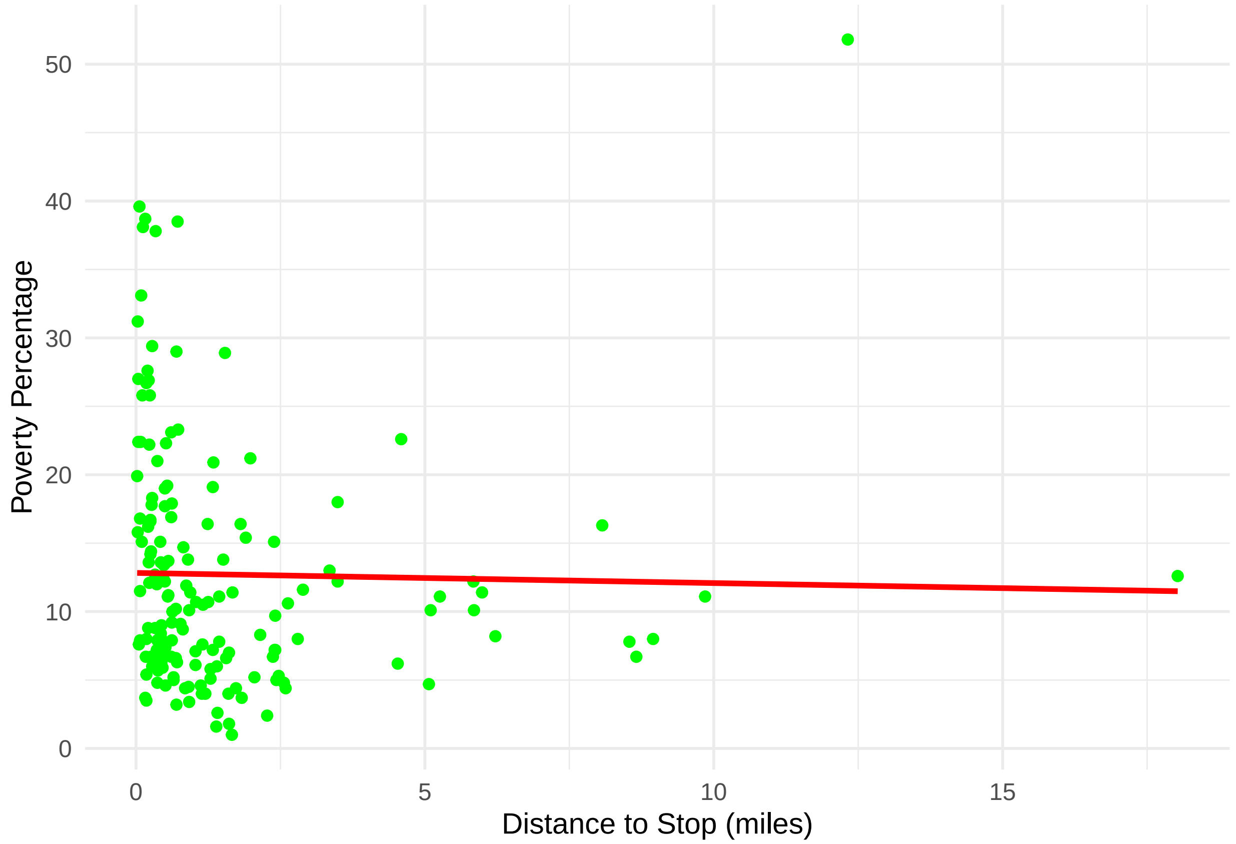
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



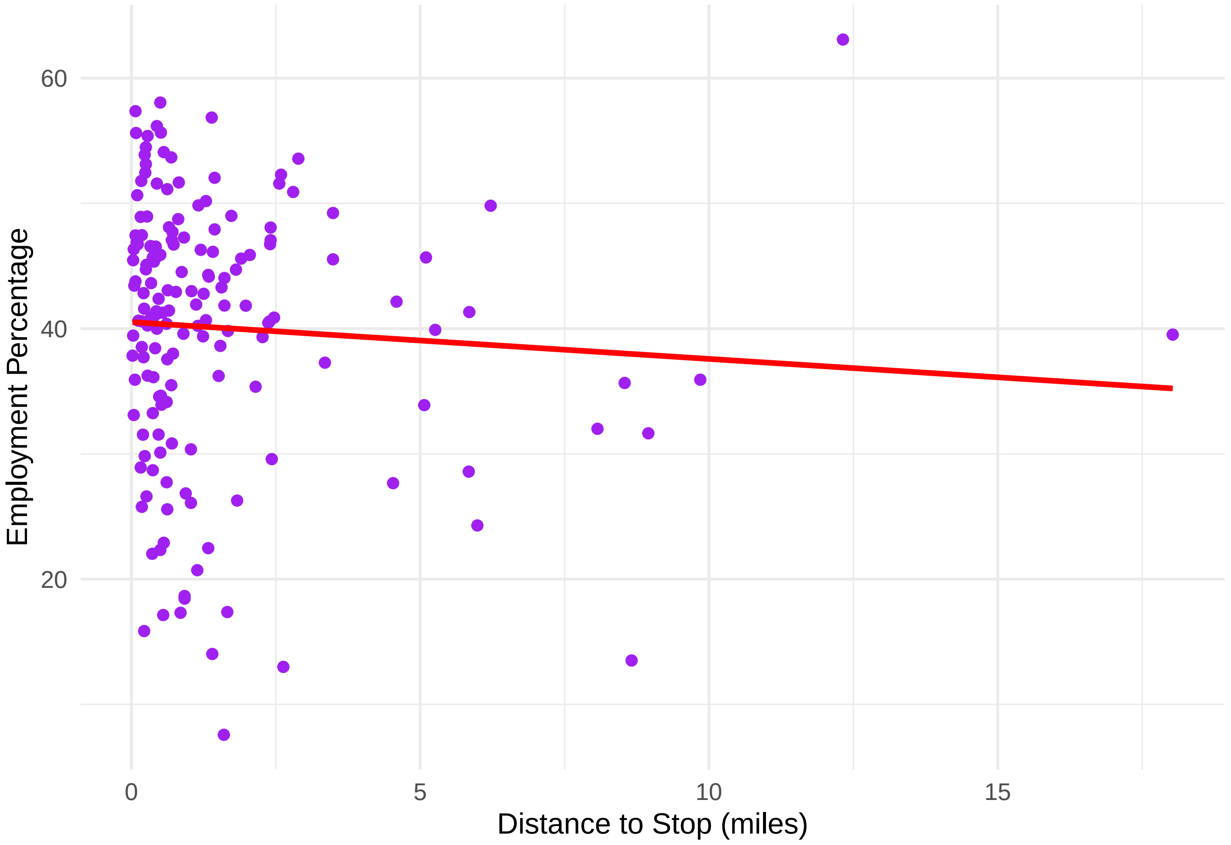
```
print(plot2)
```

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



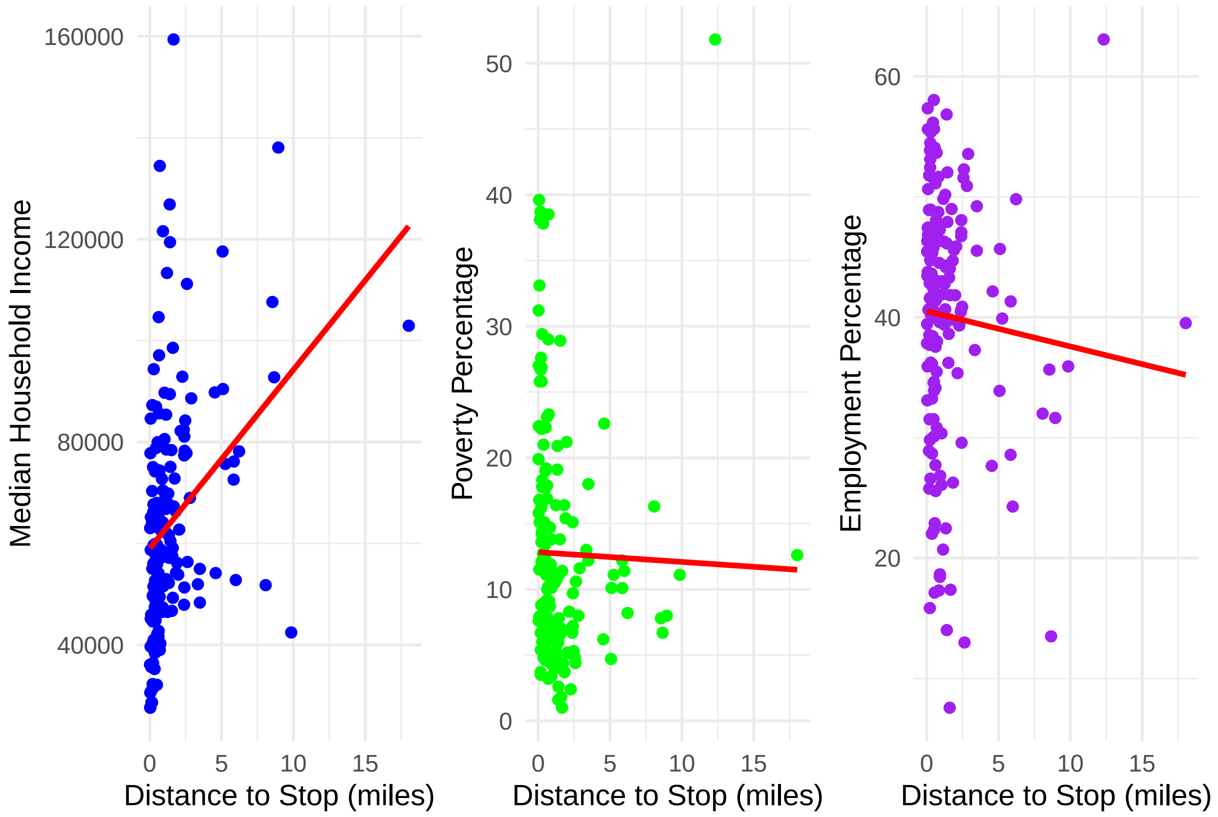
```
print(plot3)
```

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



```r
library(patchwork)

# Combine plots side by side
plot1 + plot2 + plot3
```

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

## Warning: Removed 1 row containing non-finite outside the scale range (`stat_smooth()`).
## Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).

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

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



3 Different Models

```
# Create non_white_prct variable
tract_data$non_white_prct <- 100 - tract_data$white_prct

# Model 1: Median Household Income (using non_white_prct)
model1 <- lm(med_householdincome ~ distance_from_closest_stop_miles + total_pop +
                non_white_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model1)
```

```
##
## Call:
## lm(formula = med_householdincome ~ distance_from_closest_stop_miles +
##     total_pop + non_white_prct + college_prct + avg_workhours,
##     data = tract_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44141  -8346   -350   7243  57433
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     22099.0034 23759.8687   0.930    0.354
## distance_from_closest_stop_miles  2445.8372   553.7963   4.416 1.85e-05 ***
## total_pop                           1.0983     0.4712   2.331    0.021 *
## non_white_prct                     -61.9856    98.2576  -0.631    0.529
## college_prct                      1055.6924   104.2820  10.123  < 2e-16 ***
## avg_workhours                       51.5489   609.6063   0.085    0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15650 on 158 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5588, Adjusted R-squared:  0.5449
## F-statistic: 40.03 on 5 and 158 DF,  p-value: < 2.2e-16
```

```
# Model 2: Total Poverty (using non_white_prct)
model2 <- lm(poverty_prct ~ distance_from_closest_stop_miles + total_pop +
                non_white_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model2)
```

```
##
## Call:
## lm(formula = poverty_prct ~ distance_from_closest_stop_miles +
##     total_pop + non_white_prct + college_prct + avg_workhours,
##     data = tract_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2627  -3.8989  -0.2931   2.9885  27.8206
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      43.3777465  9.7666835   4.441 1.67e-05 ***
## distance_from_closest_stop_miles  0.0985276  0.2160275   0.456   0.6489
## total_pop                        -0.0003239  0.0001940  -1.670   0.0970 .
## non_white_prct                    0.2131452  0.0405839   5.252 4.76e-07 ***
## college_prct                     -0.2343747  0.0431465  -5.432 2.05e-07 ***
## avg_workhours                    -0.6526934  0.2507031  -2.603   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.486 on 159 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4546
## F-statistic: 28.34 on 5 and 159 DF,  p-value: < 2.2e-16
```

```
# Model 3: Employment Percentage (using non_white_prct)
model3 <- lm(employment_prct ~ distance_from_closest_stop_miles + total_pop +
                non_white_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model3)
```

```
##
## Call:
## lm(formula = employment_prct ~ distance_from_closest_stop_miles +
##     total_pop + non_white_prct + college_prct + avg_workhours,
##     data = tract_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.1439  -5.5340   0.4477   5.3537  22.7078
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      35.8248220 13.7207297   2.611  0.00989 **
## distance_from_closest_stop_miles  0.1208936  0.3034863   0.398  0.69091
## total_pop                         0.0005568  0.0002725   2.043  0.04271 *
## non_white_prct                   -0.0204344  0.0570143  -0.358  0.72051
## college_prct                     -0.3454503  0.0606144  -5.699  5.7e-08 ***
## avg_workhours                     0.3221758  0.3522004   0.915  0.36171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.112 on 159 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.2528
## F-statistic:  12.1 on 5 and 159 DF,  p-value: 6.137e-10
```

```r
# Model 1: Median Household Income
model1 <- lm(med_householdincome ~ distance_from_closest_stop_miles + total_pop +
                black_prct + asian_prct + other_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model1)
```

```
##
## Call:
## lm(formula = med_householdincome ~ distance_from_closest_stop_miles +
##     total_pop + black_prct + asian_prct + other_prct + college_prct +
##     avg_workhours, data = tract_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44628  -9085   -245   6555  58664
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      25039.9521 23630.7821   1.060   0.2909
## distance_from_closest_stop_miles  2603.4141   560.8072   4.642 7.26e-06 ***
## total_pop                            0.9651     0.4754   2.030   0.0441 *
## black_prct                        -126.8181   115.9915  -1.093   0.2759
## asian_prct                        1079.9039   582.6794   1.853   0.0657 .
## other_prct                        -124.1290   322.4062  -0.385   0.7008
## college_prct                       964.5357   109.4693   8.811 2.27e-15 ***
## avg_workhours                       25.6805   608.2672   0.042   0.9664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15530 on 156 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5714, Adjusted R-squared:  0.5522
## F-statistic: 29.71 on 7 and 156 DF,  p-value: < 2.2e-16
```

```r
# Model 2: Total Poverty
model2 <- lm(poverty_prct ~ distance_from_closest_stop_miles + total_pop +
                black_prct + asian_prct + other_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model2)
```

```
##
## Call:
## lm(formula = poverty_prct ~ distance_from_closest_stop_miles +
##     total_pop + black_prct + asian_prct + other_prct + college_prct +
##     avg_workhours, data = tract_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.342  -3.989  -0.432   2.928  27.118
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      43.7784625  9.7657109   4.483 1.41e-05 ***
## distance_from_closest_stop_miles  0.0925207  0.2189339   0.423   0.6732
## total_pop                        -0.0003109  0.0001964  -1.583   0.1154
## black_prct                        0.2337364  0.0482741   4.842 3.06e-06 ***
## asian_prct                        0.2050642  0.2361388   0.868   0.3865
## other_prct                        0.1914674  0.1344255   1.424   0.1563
## college_prct                     -0.2351021  0.0452911  -5.191 6.39e-07 ***
## avg_workhours                    -0.6555436  0.2516768  -2.605   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.474 on 157 degrees of freedom
## Multiple R-squared:  0.4798, Adjusted R-squared:  0.4566
## F-statistic: 20.69 on 7 and 157 DF,  p-value: < 2.2e-16
```

```r
# Model 3: Employment Percentage
model3 <- lm(employment_prct ~ distance_from_closest_stop_miles + total_pop +
                black_prct + asian_prct + other_prct +
                college_prct + avg_workhours, data = tract_data)

summary(model3)
```

```
##
## Call:
## lm(formula = employment_prct ~ distance_from_closest_stop_miles +
##     total_pop + black_prct + asian_prct + other_prct + college_prct +
##     avg_workhours, data = tract_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.8135  -5.1871   0.0835   5.3736  22.3175
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      39.9143249 12.8882698   3.097  0.00232 **
## distance_from_closest_stop_miles  0.3131106  0.2889374   1.084  0.28018
## total_pop                         0.0003596  0.0002592   1.387  0.16731
## black_prct                       -0.1551475  0.0637097  -2.435  0.01600 *
## asian_prct                        1.3266567  0.3116436   4.257 3.55e-05 ***
## other_prct                        0.3035306  0.1774077   1.711  0.08907 .
## college_prct                     -0.4289035  0.0597729  -7.176 2.69e-11 ***
## avg_workhours                     0.2451623  0.3321498   0.738  0.46155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.544 on 157 degrees of freedom
## Multiple R-squared:  0.3711, Adjusted R-squared:  0.343
## F-statistic: 13.23 on 7 and 157 DF,  p-value: 2.379e-13
```

```
combined_model <- lm(distance_from_closest_stop_miles ~ med_householdincome + poverty_prct + employment_
                       black_prct + asian_prct + other_prct + college_prct + avg_workhours, data = trac
summary(combined_model)
```

```
##
## Call:
## lm(formula = distance_from_closest_stop_miles ~ med_householdincome +
##     poverty_prct + employment_prct + total_pop + black_prct +
##     asian_prct + other_prct + college_prct + avg_workhours, data = tract_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2300 -0.9998 -0.2933  0.3633 14.1261
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.188e+00  3.476e+00   0.917   0.3605
## med_householdincome  5.452e-05  1.148e-05   4.750 4.61e-06 ***
## poverty_prct         2.688e-03  2.975e-02   0.090   0.9281
## employment_prct     -3.666e-02  2.161e-02  -1.697   0.0918 .
## total_pop           -4.937e-05  6.430e-05  -0.768   0.4438
## black_prct           1.334e-02  1.675e-02   0.796   0.4271
## asian_prct          -1.542e-01  8.027e-02  -1.921   0.0566 .
## other_prct          -4.779e-02  4.390e-02  -1.089   0.2780
## college_prct        -4.163e-02  2.263e-02  -1.840   0.0678 .
## avg_workhours       -5.153e-02  8.256e-02  -0.624   0.5335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.071 on 154 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2055, Adjusted R-squared:  0.1591
## F-statistic: 4.427 on 9 and 154 DF,  p-value: 3.624e-05
```