Exploring Data Analysis

Sebastian Oberg

- Check for correlation between your variables. - Dataset 1, Walkability Index

```
# Load necessary libraries
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
```

```
library(ggplot2)
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
# Correct the file path to match the actual location of your CSV file
file_path <- "../Data/data.csv"

# Check if the file exists before attempting to read it
if (!file.exists(file_path)) {
  stop("The file does not exist in the specified directory.")
```

```
}

# Read data
EPA_SmartLocationDatabase <- read_csv(file_path)
```

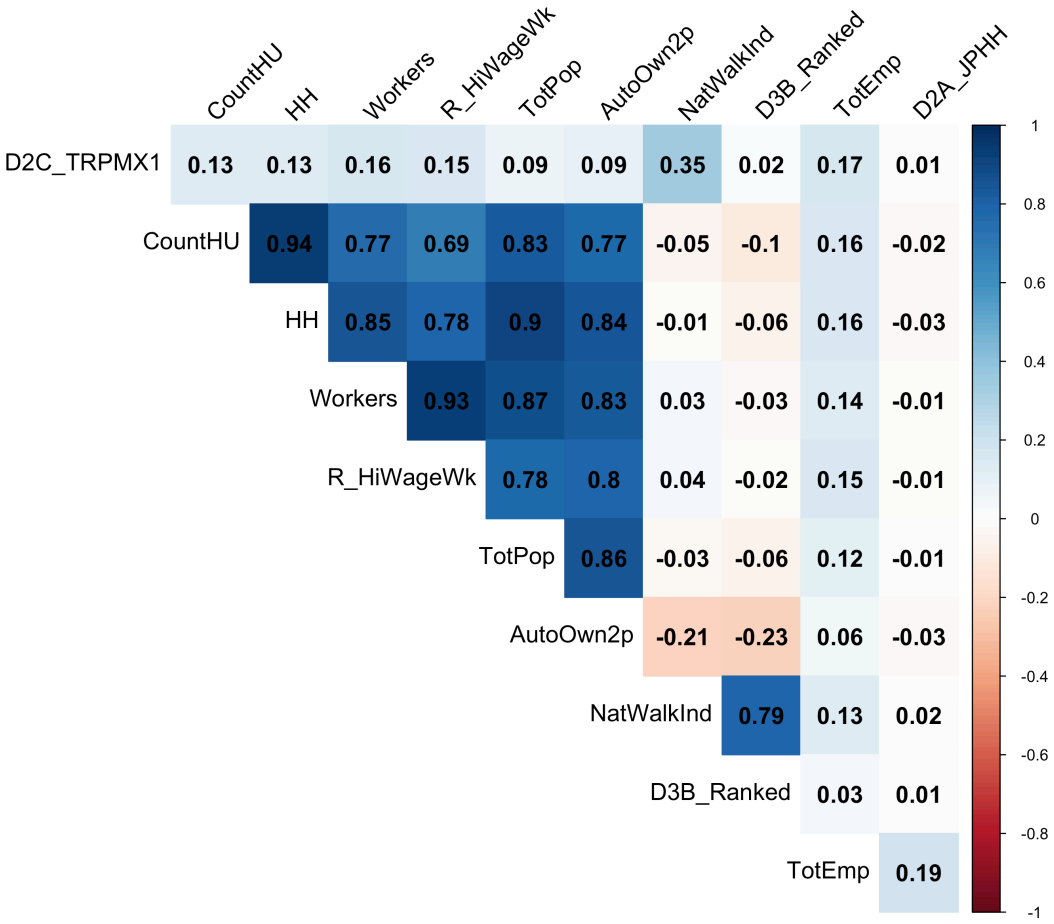
```
Rows: 220740 Columns: 117
```

```
-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (2): CSA_Name, CBSA_Name
dbl (115): OBJECTID, GEOID10, GEOID20, STATEFP, COUNTYFP, TRACTCE, BLKGRPCE,...

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.
```

```
# Select fewer columns of interest based on your domain knowledge or other criteria
selected_columns <- EPA_SmartLocationDatabase %>%
  select(TotPop, CountHU, HH, Workers, AutoOwn2p,
         R_HiWageWk, TotEmp, D2A_JPHH,
         D2C_TRPMX1, NatWalkInd, D3B_Ranked) # Adjusted the number of variables

# Convert selected columns to a numeric matrix, if not already
numeric_data <- data.matrix(selected_columns)

# Ensure all selected data is numeric and finite
numeric_data <- ifelse(!is.finite(numeric_data), NA, numeric_data)

# Compute correlation matrix for selected numeric variables
cor_matrix <- cor(numeric_data, use = "pairwise.complete.obs")

# Increase the size of the plot to make it more readable
png(filename = "correlation_matrix.png", width = 10, height = 10, units = 'in', res = 300)
par(mar = c(5, 5, 5, 5)) # Increase margin size if variable names are cut off

# Visualize the correlation matrix
corrplot(
  cor_matrix,
  method = 'color', # Use color to represent correlation
  type = 'upper', # Show only the upper half of the matrix
  order = 'hclust', # Hierarchical clustering order
  tl.cex = 1.2, # Increase text size for variable names
```

```r
  number.cex = 1.2, # Increase text size for correlation coefficients
  number.digits = 2, # Reduce the number of digits to enhance readability
  addCoef.col = 'black', # Color of the correlation coefficients
  tl.col = 'black', # Color of text labels
  tl.srt = 45, # Rotation of text labels
  diag = FALSE # Remove the diagonal
)

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

```
pdf
  2
```


CountHU
HH
Workers
R_HiWageWk
TotPop
AutoOwn2p
NatWalkInd
D3B_Ranked
TotEmp
D2A_JPHH
D2C_TRPMX1 0.13 0.13 0.16 0.15 0.09 0.09 0.35 0.02 0.17 0.01
CountHU 0.94 0.77 0.69 0.83 0.77 -0.05 -0.1 0.16 -0.02
HH 0.85 0.78 0.9 0.84 -0.01 -0.06 0.16 -0.03
Workers 0.93 0.87 0.83 0.03 -0.03 0.14 -0.01
R_HiWageWk 0.78 0.8 0.04 -0.02 0.15 -0.01
TotPop 0.86 -0.03 -0.06 0.12 -0.01
AutoOwn2p -0.21 -0.23 0.06 -0.03
NatWalkInd 0.79 0.13 0.02
D3B_Ranked 0.03 0.01
TotEmp 0.19
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1