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)
```

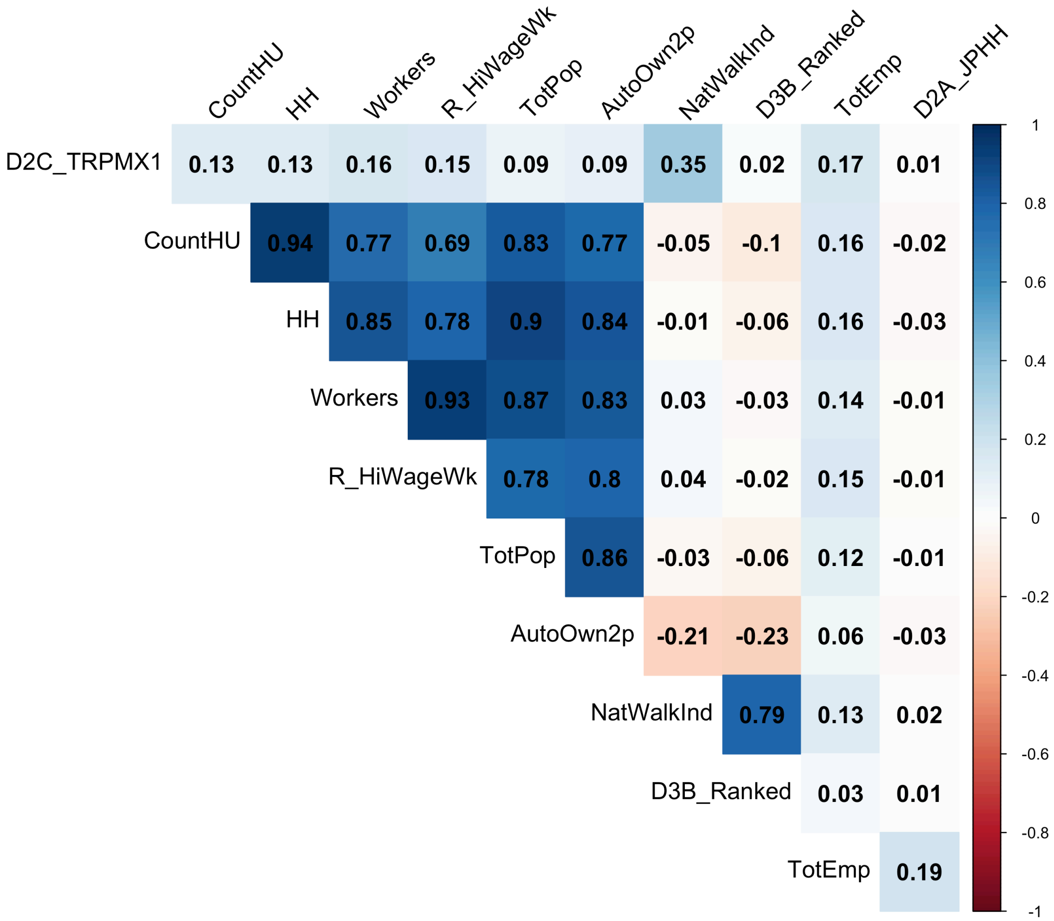
```
corrplot 0.92 loaded
```

```
# Correct the file path to match the actual location of your CSV file
file_path <- "../data/EPA_SmartLocationDatabase_V3_Jan_2021_Final.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
  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
```



The correlation matrix shows that housing-related variables like ‘CountHU’, ‘HH’, and ‘Workers’ are highly interrelated, showing that as the number of housing units increases, so does the number of households and workers. There is a strong negative correlation between the number of households with multiple cars and walkability, suggesting that more walkable areas tend to have fewer cars per household. Transportation diversity appears to have little association with the selected demographic and housing variables, indicating other factors may be at play.

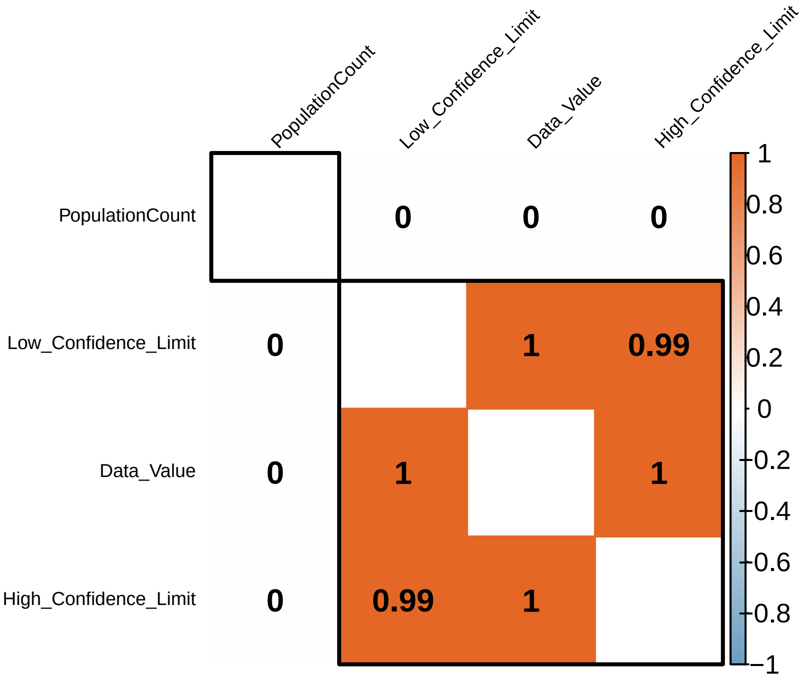
- Check for correlation between your variables. - Dataset 2, 500 Cities: Diagnosed diabetes among adults aged >=18 years

```r
# Load necessary libraries
library(readr)
library(dplyr)
library(ggplot2)
library(corrplot)

# Correct the file path to match the actual location of your CSV file
file_path <- "../data/500_Cities__Diagnosed_diabetes_among_adults_aged___18_years_20240219.cs

# 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
health_data <- read_csv(file_path)
```

```
Rows: 29006 Columns: 24
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (19): StateAbbr, StateDesc, CityName, GeographicLevel, DataSource, Categ...
dbl  (5): Year, Data_Value, Low_Confidence_Limit, High_Confidence_Limit, Pop...

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.
```

```r
# Select the columns of interest
selected_columns <- health_data %>%
  select(Data_Value, Low_Confidence_Limit, High_Confidence_Limit, PopulationCount)

# 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")
```

```
# Visualize the correlation matrix
corrplot(cor_matrix, method = 'color',
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, addrect = 2,
         col = colorRampPalette(c("#6D9EC1", "white", "#E46726"))(200),
         order = "hclust", addCoef.col = "black",
         diag = FALSE)
```



The correlation matrix shows a very high correlation between 'Data_Value' and its confidence limits, which is expected since the confidence limits are derived from the 'Data_Value'. The lack of correlation between 'PopulationCount' and the other variables suggests that in this dataset, the size of the population in a census tract does not influence the reported diabetes prevalence rates. This could possibly mean that diabetes prevalence is relatively the same across different population sizes, or that other factors not included in this analysis have a more significant impact on diabetes rates.