Exploring Data Analysis

Sebastian, Arka, and Abhinav

Performing exploratory data analysis:

- Look at the distributions and correlations of different variables. - Abhinav
- Plot variables over space and see if there is a noticeable trend. - Arka
- Check for correlation between your variables. - Sebastian

```
# 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
```

```
# Read data
EPA_SmartLocationDatabase <- read_csv("../data/EPA_SmartLocationDatabase_V3_Jan_2021_Final.cs
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.
```

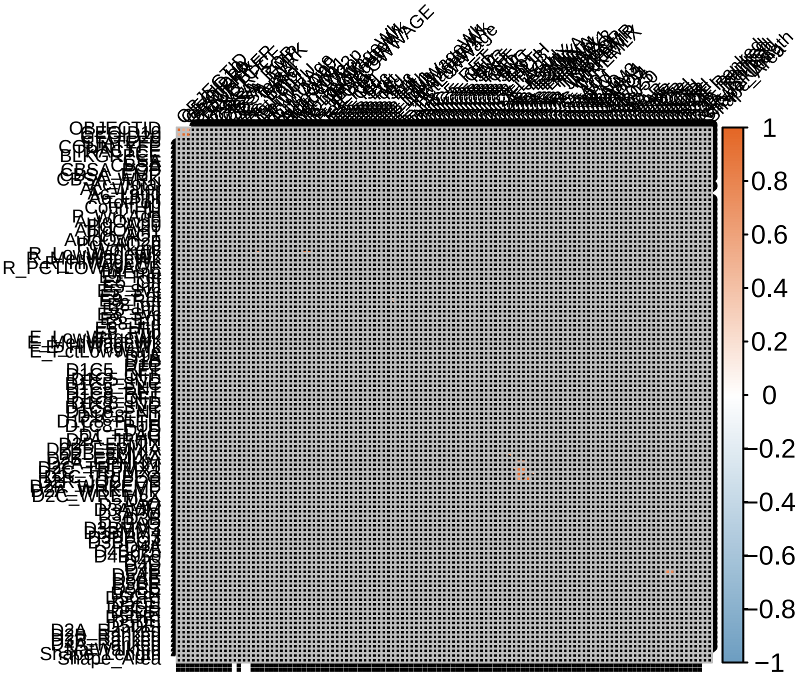
```r
# Select only numeric columns for correlation matrix
numeric_columns <- sapply(EPA_SmartLocationDatabase, is.numeric)
numeric_data <- EPA_SmartLocationDatabase[, numeric_columns]

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

# Set threshold for significant correlations
threshold <- 0.5

# Filter correlations below threshold by setting to NA
cor_matrix_filtered <- cor_matrix
cor_matrix_filtered[abs(cor_matrix) < threshold] <- NA

# Visualize filtered correlation matrix
corrplot(cor_matrix_filtered, method = 'circle',
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, addrect = 2,
         col = colorRampPalette(c("#6D9EC1", "white", "#E46726"))(200))
```



- Create summary tables with grouping variables. - Abhinav