DSC640Weeks5-6Exercise

Jon Jacobson

2025-07-06

```
theft_map <- read_csv("carTheftsMap.csv", show_col_types = FALSE)
head(theft_map)
```

```
## # A tibble: 6 x 9
##   agency_ori geo_name   countCarThefts2019 countCarThefts2020 countCarThefts2021
##   <chr>      <chr>                   <dbl> <chr>              <chr>
## 1 MO0490300  Carthage ~                 62 58                 47
## 2 <NA>       Warren Co~                112 94                 76
## 3 TX06802    Odessa PD                 499 464                375
## 4 MO0530000  Laclede C~                 55 74                 54
## 5 MO0480000  Jackson C~                125 141                81
## 6 MN0070100  Mankato D~                 72 62                 66
## # i 4 more variables: countCarThefts2022 <dbl>, latitude <dbl>,
## #   longitude <dbl>, percentChange2019to2022 <dbl>
```

```
theft_mil <- read_csv("KiaHyundaiMilwaukeeData.csv", show_col_types = FALSE)
head(theft_mil)
```

```
## # A tibble: 6 x 7
##   month  year city      state countKiaHyundaiThefts countOtherThefts
##   <chr> <dbl> <chr>     <chr>                 <dbl>            <dbl>
## 1 Jan    2019 Milwaukee WI                       22              235
## 2 Feb    2019 Milwaukee WI                       13              218
## 3 Mar    2019 Milwaukee WI                       10              195
## 4 Apr    2019 Milwaukee WI                       10              238
## 5 May    2019 Milwaukee WI                       11              280
## 6 Jun    2019 Milwaukee WI                       15              330
## # i 1 more variable: percentKiaHyundai <dbl>
```

```
theft_kia <- read_csv("kiaHyundaiThefts.csv", show_col_types = FALSE)
head(theft_kia)
```

```
## # A tibble: 6 x 7
##   month  year city    state countKiaHyundaiThefts countOtherThefts
##   <chr> <dbl> <chr>   <chr>                 <dbl>            <dbl>
## 1 Jan    2019 Atlanta GA                       17              264
## 2 Feb    2019 Atlanta GA                       11              205
## 3 Mar    2019 Atlanta GA                       18              181
## 4 Apr    2019 Atlanta GA                       15              223
## 5 May    2019 Atlanta GA                       16              277
## 6 Jun    2019 Atlanta GA                       14              220
## # i 1 more variable: percentKiaHyundai <dbl>
```

```r
# https://kyleb.rbind.io/posts/2020-06-22_excel-data-multiple-headers/importing-excel-data-with-multipl

file_path <- "motherboard_theft_data.xlsx"
two_rows <- read_excel(file_path, n_max = 2, col_names = FALSE,
    .name_repair = "minimal")
name_cols <- two_rows %>%
    t() %>%
    as_tibble()  # back to tibble
```

```r
# use dplyr fill to fill in the NA's
name_filled <- name_cols %>%
    fill(V1)

# Add a column that concatenates the others
name_with_concat <- name_filled %>%
    mutate(new_names = paste(V1, V2, sep = "PIPECHAR"))

# Extract just the concatenated column
names <- name_with_concat %>%
    pull(new_names)

# Add the 'date' column name to the front of the list
with_date <- names |>
    append("date", after = 0)
theft_all <- readxl::read_excel(file_path, col_names = with_date,
    skip = 2) %>%
    # Remove casing, whitespace, and non-alphanumeric
    # characters
janitor::clean_names()

names(theft_all) <- gsub("_pipechar_", "|", names(theft_all))  # Restore '|'
names(theft_all) <- gsub("pipechar_", "|", names(theft_all))  # Restore '|'
names(theft_all) <- gsub("_pipechar", "|", names(theft_all))  # Restore '|'
names(theft_all) <- gsub("pipechar", "|", names(theft_all))  # Restore '|'
head(theft_all)
```

```
## # A tibble: 6 x 211
##   date                `denver|kia_hyundais` `denver|all` `denver|percent`
##   <dttm>                              <dbl>        <dbl>            <dbl>
## 1 2019-12-01 00:00:00                    48          615           0.0780
## 2 2020-01-01 00:00:00                    21          519           0.0405
## 3 2020-02-01 00:00:00                    28          402           0.0697
## 4 2020-03-01 00:00:00                    35          508           0.0689
## 5 2020-04-01 00:00:00                    39          616           0.0633
## 6 2020-05-01 00:00:00                    37          692           0.0535
## # i 207 more variables: `el_paso|kia_hyundais` <dbl>, `el_paso|all` <dbl>,
## #   `el_paso|percent` <dbl>, `portland|kia_hyundais` <dbl>,
## #   `portland|all` <dbl>, `portland|percent` <dbl>,
## #   `atlanta|kia_hyundais` <dbl>, `atlanta|all` <dbl>, `atlanta|percent` <dbl>,
## #   `chicago|kia_hyundais` <dbl>, `chicago|all` <dbl>, `chicago|percent` <dbl>,
## #   `virginia_beach|kia_hyundais` <dbl>, `virginia_beach|all` <lgl>,
## #   `virginia_beach|percent` <lgl>, `louisville|kia_hyundais` <dbl>, ...
```

Stacked Area of Milwaukee

```
head(theft_mil)
```

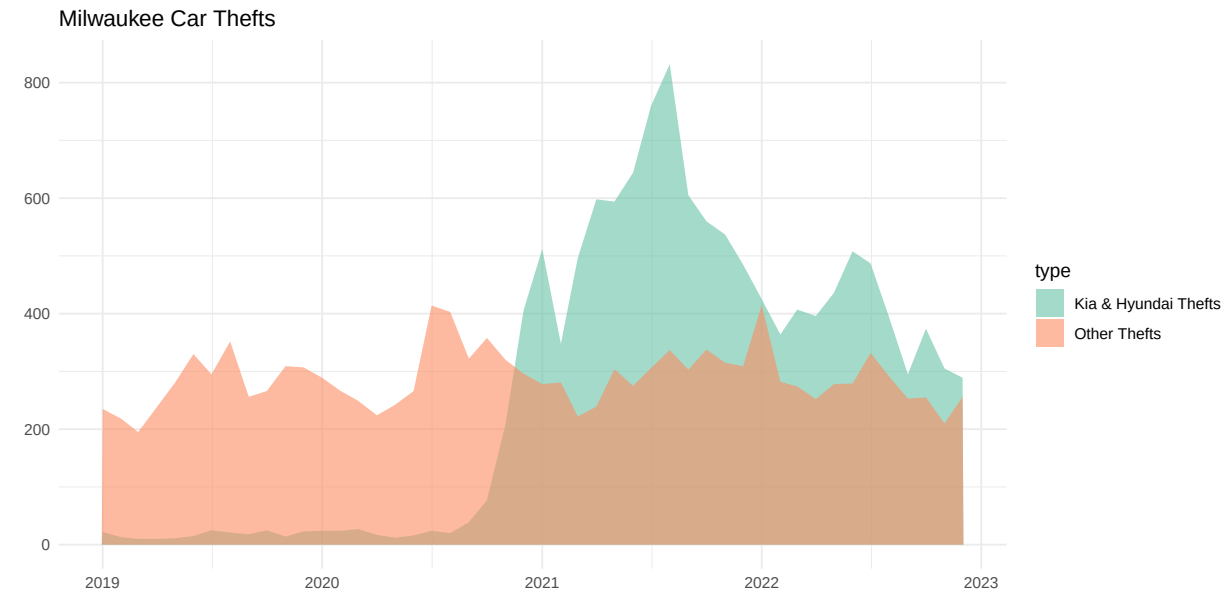
```
## # A tibble: 6 x 7
##   month  year city      state countKiaHyundaiThefts countOtherThefts
##   <chr> <dbl> <chr>     <chr>                 <dbl>            <dbl>
## 1 Jan    2019 Milwaukee WI                       22              235
## 2 Feb    2019 Milwaukee WI                       13              218
## 3 Mar    2019 Milwaukee WI                       10              195
## 4 Apr    2019 Milwaukee WI                       10              238
## 5 May    2019 Milwaukee WI                       11              280
## 6 Jun    2019 Milwaukee WI                       15              330
## # i 1 more variable: percentKiaHyundai <dbl>
```

```
mil_long <- theft_mil %>%
    pivot_longer(cols = c("countKiaHyundaiThefts", "countOtherThefts"),
        names_to = "type", values_to = "value")
mil_long$date <- as.Date(paste("01", mil_long$month, mil_long$year),
    format = "%d %b %Y")
mil_long <- mil_long %>%
    mutate(type = recode(type, countKiaHyundaiThefts = "Kia & Hyundai Thefts",
        countOtherThefts = "Other Thefts"))

head(mil_long)
```

```
## # A tibble: 6 x 8
##   month  year city      state percentKiaHyundai type            value date
##   <chr> <dbl> <chr>     <chr>             <dbl> <chr>           <dbl> <date>
## 1 Jan    2019 Milwaukee WI                0.086 Kia & Hyundai ~    22 2019-01-01
## 2 Jan    2019 Milwaukee WI                0.086 Other Thefts      235 2019-01-01
## 3 Feb    2019 Milwaukee WI                0.056 Kia & Hyundai ~    13 2019-02-01
## 4 Feb    2019 Milwaukee WI                0.056 Other Thefts      218 2019-02-01
## 5 Mar    2019 Milwaukee WI                0.049 Kia & Hyundai ~    10 2019-03-01
## 6 Mar    2019 Milwaukee WI                0.049 Other Thefts      195 2019-03-01
```

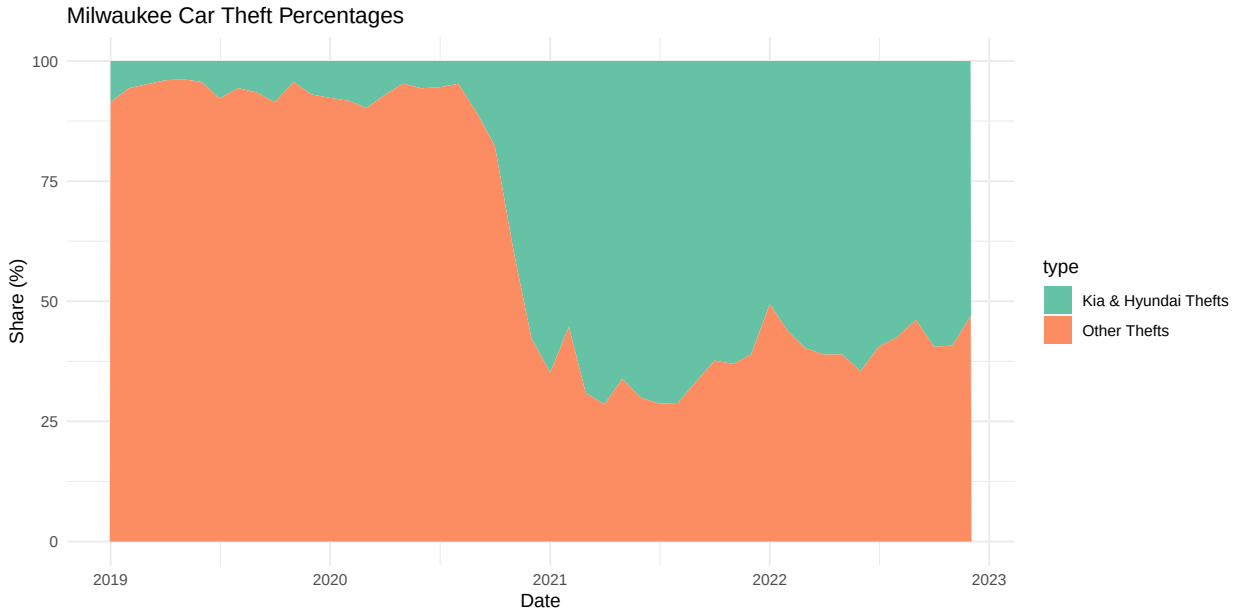
```
ggplot(mil_long, aes(x = date, y = value, fill = type)) + geom_area(position = "identity",
    alpha = 0.6) + labs(title = "Milwaukee Car Thefts", x = "",
    y = "") + scale_fill_brewer(palette = "Set2") + theme_minimal()
```




```
library(dplyr)

mil_pct <- mil_long %>%
    group_by(date) %>%
    mutate(pct_value = value/sum(value) * 100) %>%
    ungroup()
library(ggplot2)

ggplot(mil_pct, aes(x = date, y = pct_value, fill = type)) +
    geom_area(position = "stack") + labs(title = "Milwaukee Car Theft Percentages",
    x = "Date", y = "Share (%)") + scale_fill_brewer(palette = "Set2") +
    theme_minimal()
```



```
theft_kia
```

```
## # A tibble: 552 x 7
##    month  year city    state countKiaHyundaiThefts countOtherThefts
##    <chr> <dbl> <chr>   <chr>                 <dbl>            <dbl>
##  1 Jan    2019 Atlanta GA                       17              264
##  2 Feb    2019 Atlanta GA                       11              205
##  3 Mar    2019 Atlanta GA                       18              181
##  4 Apr    2019 Atlanta GA                       15              223
##  5 May    2019 Atlanta GA                       16              277
##  6 Jun    2019 Atlanta GA                       14              220
##  7 Jul    2019 Atlanta GA                       19              267
##  8 Aug    2019 Atlanta GA                       12              242
##  9 Sep    2019 Atlanta GA                       11              234
## 10 Oct    2019 Atlanta GA                       12              206
## # i 542 more rows
## # i 1 more variable: percentKiaHyundai <dbl>
```

```
library(dplyr)

tree_map <- theft_kia %>%
    group_by(state, city) %>%
    summarise(value = sum(countKiaHyundaiThefts), .groups = "drop")

tree_map <- tree_map %>%
    filter(!is.na(state))

library(ggplot2)
library(treemapify)

ggplot(tree_map, aes(area = value, fill = state, label = city,
    subgroup = state)) + geom_treemap() + geom_treemap_subgroup_border() +
    geom_treemap_subgroup_text(place = "top", grow = TRUE, alpha = 0.5,
        colour = "white") + geom_treemap_text(colour = "white",
    place = "centre", grow = TRUE) + theme_minimal()
```


```
theft_all <- theft_all %>%
  mutate(across(where(is.character), ~ as.double(.)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(where(is.character), ~as.double(.))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
thefts <- theft_all %>%
  pivot_longer(
    cols = -date,
    names_to = c("city", "type"),
    names_sep = "\\|",  # splits names like 'denver_kia' into 'denver' and 'kia'
    values_to = "value"
  ) %>%
  filter(!is.na(value))  # Drops rows with missing values

# Pivot wider to get both values side by side per city-date
others <- thefts %>%
  pivot_wider(names_from = type, values_from = value) %>%
  mutate(others = all - kia_hyundais) %>%
  select(date, city, others) %>%
  mutate(type = "others") %>%
  rename(value = others)

thefts <- bind_rows(
  thefts,
  others
) %>%
  arrange(city,date)


thefts <- thefts %>%
  filter(!type %in% c("all", "percent",'others'))
```

```r
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
thefts <- thefts %>%
  mutate(year = year(date))  # Extracts year from the date

head(thefts)
```

```
## # A tibble: 6 x 5
##   date                city     type         value  year
##   <dttm>              <chr>    <chr>        <dbl> <dbl>
## 1 2019-12-01 00:00:00 akron_oh kia_hyundais     9  2019
## 2 2020-01-01 00:00:00 akron_oh kia_hyundais     1  2020
## 3 2020-02-01 00:00:00 akron_oh kia_hyundais     2  2020
## 4 2020-03-01 00:00:00 akron_oh kia_hyundais     2  2020
## 5 2020-04-01 00:00:00 akron_oh kia_hyundais     7  2020
## 6 2020-05-01 00:00:00 akron_oh kia_hyundais     3  2020
```

```r
df_summary <- thefts %>%
    group_by(year, city) %>%
    summarise(value = sum(value, na.rm = TRUE), .groups = "drop")

top_20_cities <- df_summary %>%
    group_by(city) %>%
    summarise(total_value = sum(value, na.rm = TRUE), .groups = "drop") %>%
    arrange(desc(total_value)) %>%
    slice_head(n = 20) %>%
    pull(city)

top_6_cities <- df_summary %>%
    group_by(city) %>%
    summarise(total_value = sum(value, na.rm = TRUE), .groups = "drop") %>%
    arrange(desc(total_value)) %>%
    slice_head(n = 6) %>%
    pull(city)

plot_20 <- df_summary %>%
    filter(city %in% top_20_cities)

plot_6 <- thefts %>%
    filter(city %in% top_6_cities)

head(plot_20)
```

```
## # A tibble: 6 x 3
##    year city       value
##   <dbl> <chr>      <dbl>
## 1  2019 austin_tx     10
```

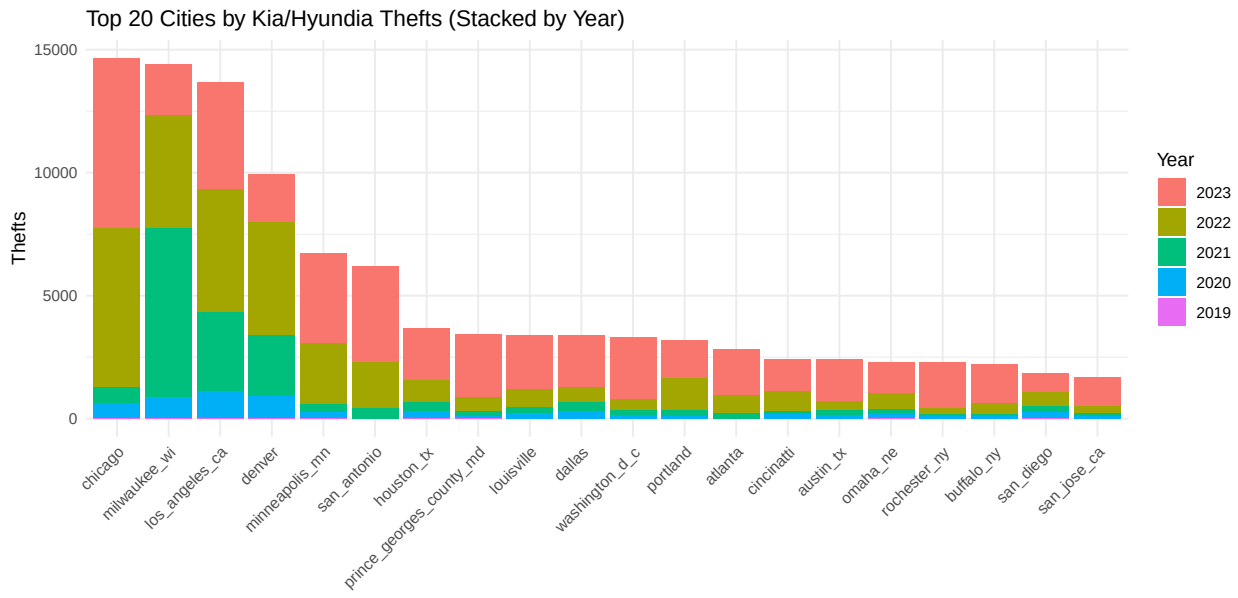
```
## 2  2019 buffalo_ny     7
## 3  2019 chicago       46
## 4  2019 dallas        21
## 5  2019 denver        48
## 6  2019 houston_tx    24
```

```r
library(ggplot2)

# Reverse the order of years so that 2019 is on the bottom
df_plot <- plot_20 %>%
    mutate(year = factor(year, levels = sort(unique(year), decreasing = TRUE)))

# Reorder 'city' factor by total value descending
df_plot <- df_plot %>%
    group_by(city) %>%
    mutate(city_total = sum(value, na.rm = TRUE)) %>%
    ungroup() %>%
    mutate(city = reorder(city, -city_total, FUN = max))  # `max` for consistent descending order

ggplot(df_plot, aes(x = city, y = value, fill = factor(year))) +
    geom_bar(stat = "identity") + labs(title = "Top 20 Cities by Kia/Hyundia Thefts (Stacked by Year)",
    x = "", y = "Thefts", fill = "Year") + theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



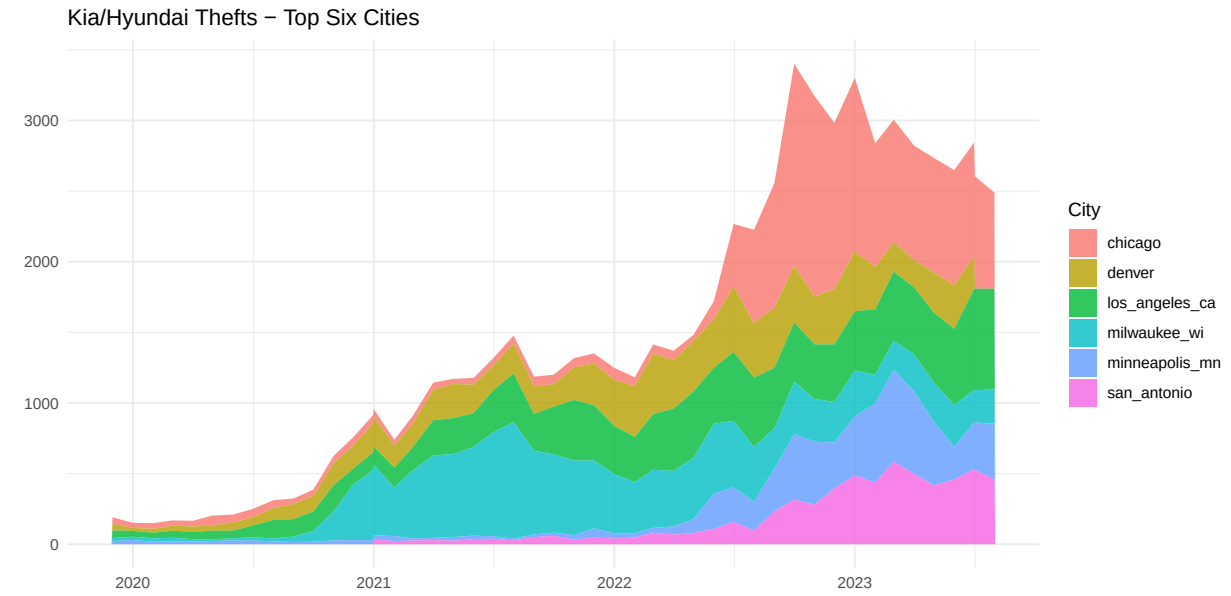

```r
head(plot_6)
```

```
## # A tibble: 6 x 5
##   date                city    type         value  year
##   <dttm>              <chr>   <chr>        <dbl> <dbl>
## 1 2019-12-01 00:00:00 chicago kia_hyundais    46  2019
## 2 2020-01-01 00:00:00 chicago kia_hyundais    36  2020
## 3 2020-02-01 00:00:00 chicago kia_hyundais    40  2020
## 4 2020-03-01 00:00:00 chicago kia_hyundais    38  2020
## 5 2020-04-01 00:00:00 chicago kia_hyundais    40  2020
## 6 2020-05-01 00:00:00 chicago kia_hyundais    70  2020
```

```r
library(ggplot2)
library(dplyr)

# Ensure date is in proper format
plot_6$date <- as.Date(plot_6$date)

ggplot(plot_6, aes(x = date, y = value, fill = city)) + geom_area(position = "stack",
    alpha = 0.8) + labs(title = "Kia/Hyundai Thefts - Top Six Cities",
    x = "", y = "", fill = "City") + theme_minimal()
```




```r
thefts <- theft_map %>%
    rename(percent = "percentChange2019to2022")

thefts <- thefts %>%
    filter(percent >= 0.5) %>%
    filter(!is.na(latitude) & !is.na(longitude) & !is.na(percent)) %>%
    # drop alaska
filter(latitude <= 50) %>%
    filter(latitude >= 25 & latitude <= 50)  # U.S. latitude range

head(thefts)
```

```
## # A tibble: 6 x 9
##   agency_ori geo_name   countCarThefts2019 countCarThefts2020 countCarThefts2021
##   <chr>      <chr>                   <dbl> <chr>              <chr>
## 1 CA01005    Fresno PD                2051 2650               3447
## 2 PA0480400  Easton Ci~                 24 18                 24
## 3 MN0271800  Richfield~                 88 134                131
## 4 CA00109    Oakland PD               6477 8737               9349
## 5 MO0240700  North Kan~                 82 118                138
## 6 PA0090100  Bensalem ~                 83 134                113
## # i 4 more variables: countCarThefts2022 <dbl>, latitude <dbl>,
## #   longitude <dbl>, percent <dbl>
```

```
library(ggplot2)
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
# Get US map data
us_map <- map_data("state")

ggplot() + geom_polygon(data = us_map, aes(x = long, y = lat,
    group = group), fill = "gray95", color = "white") + geom_point(data = thefts,
    aes(x = longitude, y = latitude, size = percent), color = "steelblue",
    alpha = 0.2, stroke = 0) + scale_size_continuous(breaks = seq(0.5,
    4, by = 0.5), labels = label_percent()) + coord_fixed(1.3) +
    labs(title = "Percentage Increase in Kia/Hyundai Thefts",
        x = "longitude", y = "latitude", size = "percent") +
    theme_void()
```