# CMDA-4654

Group Project 1

Yusi Yao & Xavier Akers

2025-02-18

# Teammate Introduction

Hi, I am Yusi Yao! I am originally from Nanjing, China, but I am always on the move—whether it is heading back to D.C. on weekends or exploring the outdoors. In Blacksburg, you will probably find me grabbing a bite at Rainbowl, my go-to spot for poke bowl. Outside of academics, I enjoy fishing and playing soccer.

# Data Introduction

## Dataset Overview

This dataset tracks home values trends across different ZIP codes in the U.S. It includes VAHI, an index developed by zillow to estimate values for single-family homes and condominiums. The data covers its fluctuation from **January 2015 to January 2025**.

Zillow Home Value Index (ZHVI): A measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range. Available as a smoothed, seasonally adjusted measure and as a raw measure.

## Data Category

This dataset belongs to category **8 housing**. In this project, we tend to discover the LA wildfire's impact on the housing market in the ZIP codes of Los Angeles.

## Data Dictionary

| Column Name | Description |
|---|---|
| **RegionID** | Unique ID for each ZIP code. |
| **SizeRank** | Ranking of ZIP code by housing market size. |
| **RegionName** | ZIP code number. |
| **RegionType** | Type of region (e.g., "zip"). |
| **StateName** | Full name of the state. |
| **State** | Two-letter state abbreviation. |
| **City** | City name. |
| **Metro** | Metro area including the ZIP code. |
| **CountyName** | County name. |
| **2015-01-31, . . . , 2025-01-31** | Monthly rent estimates in dollars. |

## Data Source

This dataset comes from **Zillow's public data**. More details can be found at:
Zillow Research Data

# Analysis & Discussion

## Data Cleaning

1. Filter ZIP codes in Los Angeles County
2. Remove columns with too many missing values
3. Handle missing rent prices using interpolation
4. Ensure date columns are in the correct format

```r
# Read the dataset (using relative path) and suppress column type warnings
Zip_zori_uc_sfrcondomfr_sm_month <- read_csv(here("data", "Zip_zori_uc_sfrcondomfr_sm_month.csv"), show_col_ty

# Filter for Los Angeles County (focus on LA housing market)
la_housing <- Zip_zori_uc_sfrcondomfr_sm_month %>%
  filter(CountyName == "Los Angeles County" & State == "CA")

# Identify date columns (should already be formatted correctly)
date_columns <- names(la_housing)[10:ncol(la_housing)]

# Print column names to verify date columns exist
print(date_columns)  # Should display "2015-01-31", "2015-02-28", etc.
```

```
  [1] "1/31/15"  "2/28/15"  "3/31/15"  "4/30/15"  "5/31/15"  "6/30/15"
  [7] "7/31/15"  "8/31/15"  "9/30/15"  "10/31/15" "11/30/15" "12/31/15"
 [13] "1/31/16"  "2/29/16"  "3/31/16"  "4/30/16"  "5/31/16"  "6/30/16"
 [19] "7/31/16"  "8/31/16"  "9/30/16"  "10/31/16" "11/30/16" "12/31/16"
 [25] "1/31/17"  "2/28/17"  "3/31/17"  "4/30/17"  "5/31/17"  "6/30/17"
 [31] "7/31/17"  "8/31/17"  "9/30/17"  "10/31/17" "11/30/17" "12/31/17"
 [37] "1/31/18"  "2/28/18"  "3/31/18"  "4/30/18"  "5/31/18"  "6/30/18"
 [43] "7/31/18"  "8/31/18"  "9/30/18"  "10/31/18" "11/30/18" "12/31/18"
 [49] "1/31/19"  "2/28/19"  "3/31/19"  "4/30/19"  "5/31/19"  "6/30/19"
 [55] "7/31/19"  "8/31/19"  "9/30/19"  "10/31/19" "11/30/19" "12/31/19"
 [61] "1/31/20"  "2/29/20"  "3/31/20"  "4/30/20"  "5/31/20"  "6/30/20"
 [67] "7/31/20"  "8/31/20"  "9/30/20"  "10/31/20" "11/30/20" "12/31/20"
 [73] "1/31/21"  "2/28/21"  "3/31/21"  "4/30/21"  "5/31/21"  "6/30/21"
 [79] "7/31/21"  "8/31/21"  "9/30/21"  "10/31/21" "11/30/21" "12/31/21"
 [85] "1/31/22"  "2/28/22"  "3/31/22"  "4/30/22"  "5/31/22"  "6/30/22"
 [91] "7/31/22"  "8/31/22"  "9/30/22"  "10/31/22" "11/30/22" "12/31/22"
 [97] "1/31/23"  "2/28/23"  "3/31/23"  "4/30/23"  "5/31/23"  "6/30/23"
[103] "7/31/23"  "8/31/23"  "9/30/23"  "10/31/23" "11/30/23" "12/31/23"
[109] "1/31/24"  "2/29/24"  "3/31/24"  "4/30/24"  "5/31/24"  "6/30/24"
[115] "7/31/24"  "8/31/24"  "9/30/24"  "10/31/24" "11/30/24" "12/31/24"
[121] "1/31/25"
```

```r
# Convert wide format to long format
la_housing_long <- la_housing %>%
  pivot_longer(cols = all_of(date_columns), names_to = "Date", values_to = "Rent_Price")

# Convert Date column to Date type
la_housing_long$Date <- as.Date(la_housing_long$Date, format = "%Y-%m-%d")

# Handle missing values using interpolation
la_housing_long <- la_housing_long %>%
  group_by(RegionName) %>%
  mutate(Rent_Price = ifelse(is.na(Rent_Price),
                             zoo::na.approx(Rent_Price, na.rm = FALSE),
                             Rent_Price)) %>%
  ungroup()

# Save cleaned dataset using relative path
write_csv(la_housing_long, here("data", "cleaned_LA_housing.csv"))
```

Rental Price Trends in Los Angeles County

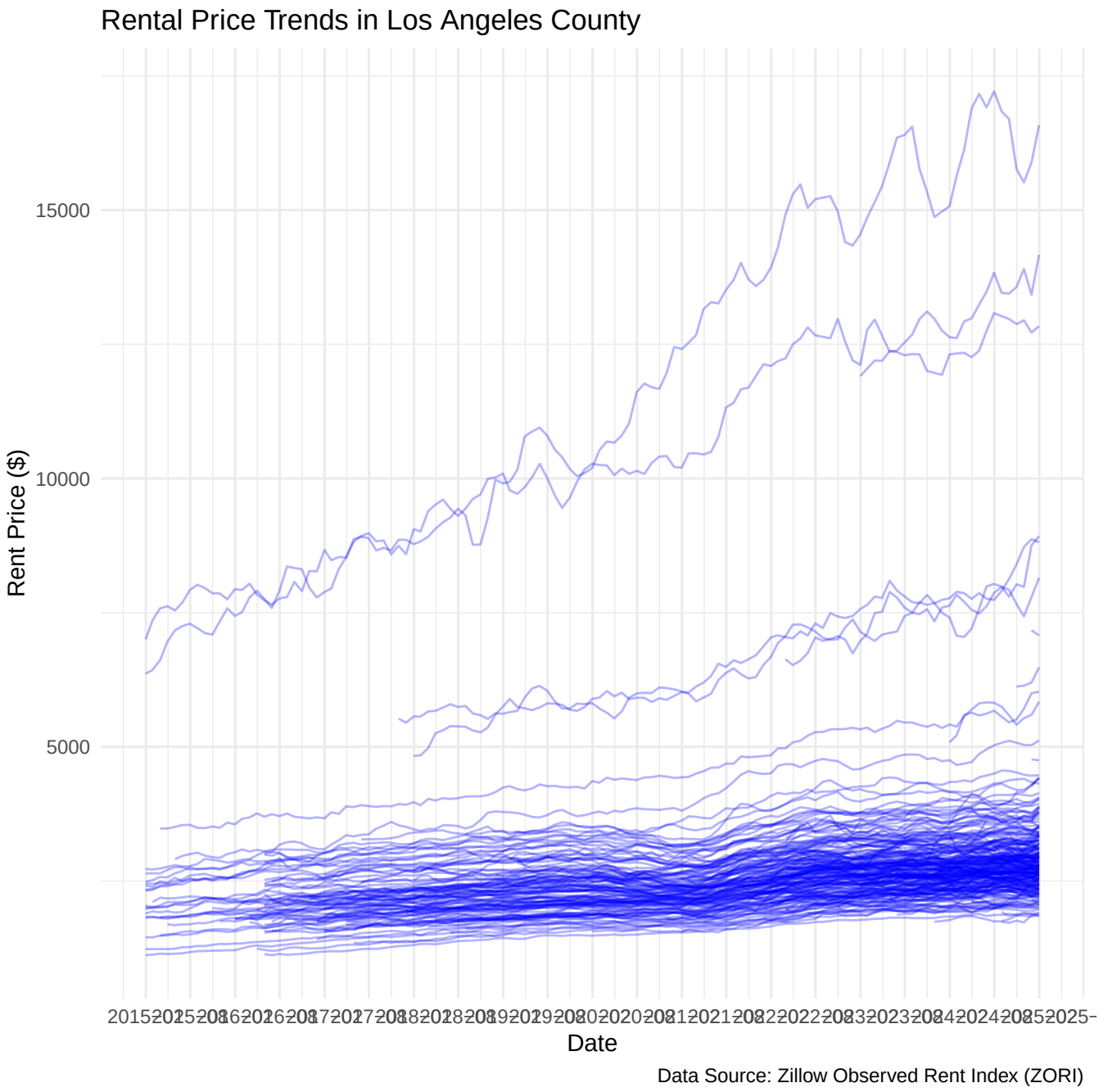
```r
# Ensure the "data" directory exists
dir.create(here("data"), showWarnings = FALSE)

# Read the dataset (using relative path) and suppress column type warnings
```

```r
Zip_zori_uc_sfrcondomfr_sm_month <- read_csv(here("data", "Zip_zori_uc_sfrcondomfr_sm_month.csv"), show_col_ty

# Filter for Los Angeles County (focus on LA housing market)
la_housing <- Zip_zori_uc_sfrcondomfr_sm_month %>%
  filter(CountyName == "Los Angeles County" & State == "CA")

# Identify supposed date columns (currently in "M/DD/YY" format)
date_columns <- names(la_housing)[10:ncol(la_housing)]  # First 9 columns are metadata

# Convert date column names from "M/DD/YY" to "YYYY-MM-DD"
formatted_dates <- suppressWarnings(format(as.Date(date_columns, format = "%m/%d/%y"), "%Y-%m-%d"))

# Ensure no missing values in formatted date columns
formatted_dates[is.na(formatted_dates)] <- paste0("X", 1:sum(is.na(formatted_dates)))

# Assign corrected date column names
colnames(la_housing) <- c(names(la_housing)[1:9], formatted_dates)

# Print to confirm date columns exist
print(names(la_housing)[10:ncol(la_housing)])  # Should now display "2015-01-31", "2015-02-28", etc.
```

```
  [1] "2015-01-31" "2015-02-28" "2015-03-31" "2015-04-30" "2015-05-31"
  [6] "2015-06-30" "2015-07-31" "2015-08-31" "2015-09-30" "2015-10-31"
 [11] "2015-11-30" "2015-12-31" "2016-01-31" "2016-02-29" "2016-03-31"
 [16] "2016-04-30" "2016-05-31" "2016-06-30" "2016-07-31" "2016-08-31"
 [21] "2016-09-30" "2016-10-31" "2016-11-30" "2016-12-31" "2017-01-31"
 [26] "2017-02-28" "2017-03-31" "2017-04-30" "2017-05-31" "2017-06-30"
 [31] "2017-07-31" "2017-08-31" "2017-09-30" "2017-10-31" "2017-11-30"
 [36] "2017-12-31" "2018-01-31" "2018-02-28" "2018-03-31" "2018-04-30"
 [41] "2018-05-31" "2018-06-30" "2018-07-31" "2018-08-31" "2018-09-30"
 [46] "2018-10-31" "2018-11-30" "2018-12-31" "2019-01-31" "2019-02-28"
 [51] "2019-03-31" "2019-04-30" "2019-05-31" "2019-06-30" "2019-07-31"
 [56] "2019-08-31" "2019-09-30" "2019-10-31" "2019-11-30" "2019-12-31"
 [61] "2020-01-31" "2020-02-29" "2020-03-31" "2020-04-30" "2020-05-31"
 [66] "2020-06-30" "2020-07-31" "2020-08-31" "2020-09-30" "2020-10-31"
 [71] "2020-11-30" "2020-12-31" "2021-01-31" "2021-02-28" "2021-03-31"
 [76] "2021-04-30" "2021-05-31" "2021-06-30" "2021-07-31" "2021-08-31"
 [81] "2021-09-30" "2021-10-31" "2021-11-30" "2021-12-31" "2022-01-31"
 [86] "2022-02-28" "2022-03-31" "2022-04-30" "2022-05-31" "2022-06-30"
 [91] "2022-07-31" "2022-08-31" "2022-09-30" "2022-10-31" "2022-11-30"
 [96] "2022-12-31" "2023-01-31" "2023-02-28" "2023-03-31" "2023-04-30"
[101] "2023-05-31" "2023-06-30" "2023-07-31" "2023-08-31" "2023-09-30"
[106] "2023-10-31" "2023-11-30" "2023-12-31" "2024-01-31" "2024-02-29"
[111] "2024-03-31" "2024-04-30" "2024-05-31" "2024-06-30" "2024-07-31"
[116] "2024-08-31" "2024-09-30" "2024-10-31" "2024-11-30" "2024-12-31"
[121] "2025-01-31"
```

```r
# Convert wide format to long format
la_housing_long <- la_housing %>%
  pivot_longer(cols = all_of(formatted_dates), names_to = "Date", values_to = "Rent_Price")

# Convert Date column to Date type
la_housing_long$Date <- as.Date(la_housing_long$Date, format = "%Y-%m-%d")

# Check if any Dates are still NA
if (any(is.na(la_housing_long$Date))) {
  stop("Error: Some Date values are still NA. Check the date formatting process.")
}

# Handle missing values using interpolation
```

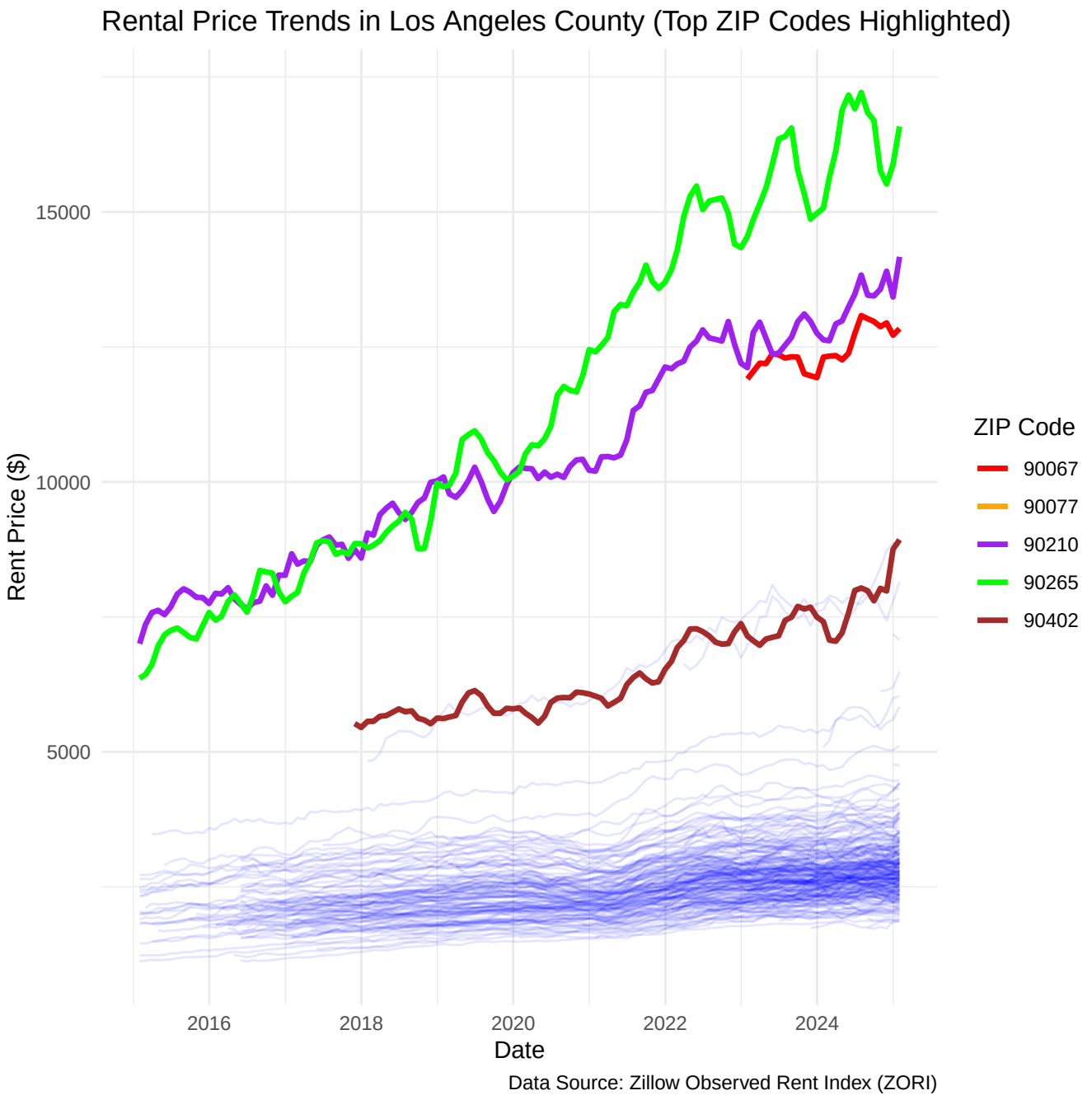
```r
la_housing_long <- la_housing_long %>%
  group_by(RegionName) %>%
  mutate(Rent_Price = ifelse(is.na(Rent_Price),
                             zoo::na.approx(Rent_Price, na.rm = FALSE),
                             Rent_Price)) %>%
  ungroup()

# Save cleaned dataset in the "data" folder using relative path
write_csv(la_housing_long, here("data", "cleaned_LA_housing.csv"))

# Load cleaned dataset
la_housing_long <- read_csv(here("data", "cleaned_LA_housing.csv"), show_col_types = FALSE)

# Convert Date column again to ensure proper format
la_housing_long$Date <- as.Date(la_housing_long$Date, format = "%Y-%m-%d")

# Ensure valid Date range
start_date <- min(la_housing_long$Date, na.rm = TRUE)
end_date <- max(la_housing_long$Date, na.rm = TRUE)

# Generate the rental price trends plot
ggplot(la_housing_long, aes(x = Date, y = Rent_Price, group = RegionName)) +
  geom_line(alpha = 0.3, color = "blue") +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "6 months",
               limits = c(start_date, end_date)) +  # Ensures valid range
  theme_minimal() +
  labs(title = "Rental Price Trends in Los Angeles County",
       x = "Date", y = "Rent Price ($)",
       caption = "Data Source: Zillow Observed Rent Index (ZORI)")
```




Highlighting Top 5 ZIP Codes with Highest Rent

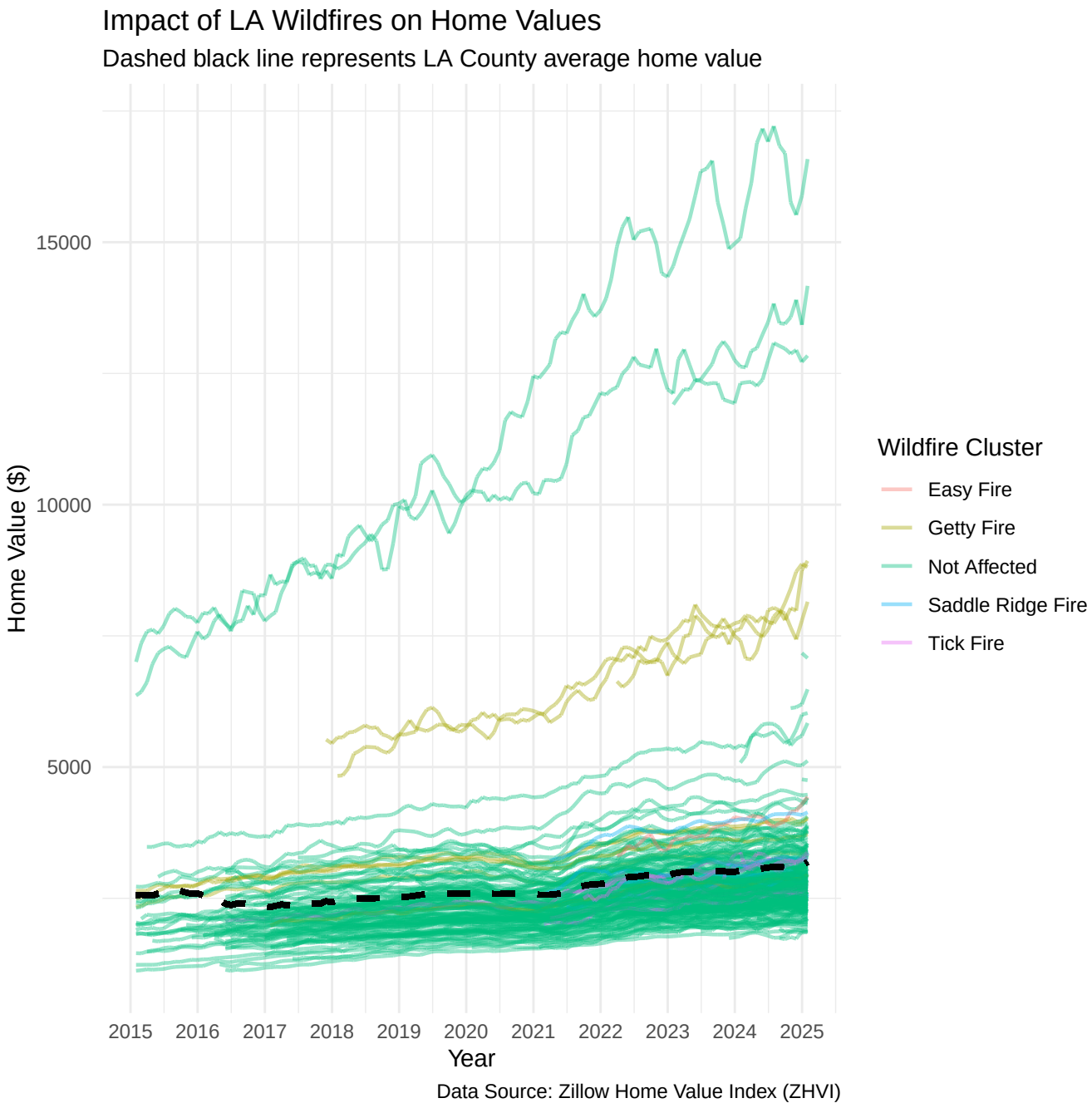
```
# Identify the top 5 ZIP codes with the highest recent rent prices
latest_date <- max(la_housing_long$Date, na.rm = TRUE)
top_zip_codes <- la_housing_long %>%
  filter(Date == latest_date) %>%
  arrange(desc(Rent_Price)) %>%
  slice_head(n = 5) %>%
  pull(RegionName)  # Extract top ZIP codes

# Generate the plot with highlighted ZIP codes
ggplot(la_housing_long, aes(x = Date, y = Rent_Price, group = RegionName)) +
  geom_line(data = la_housing_long %>% filter(!RegionName %in% top_zip_codes),
            aes(group = RegionName),
            alpha = 0.1, color = "blue") +  # Background lines for all other ZIP codes
  geom_line(data = la_housing_long %>% filter(RegionName %in% top_zip_codes),
```

```
          aes(color = as.factor(RegionName)), size = 1.2) +  # Highlight top ZIP codes
  scale_color_manual(values = c("red", "orange", "purple", "green", "brown")) +  # Custom colors
  theme_minimal() +
  labs(title = "Rental Price Trends in Los Angeles County (Top ZIP Codes Highlighted)",
       x = "Date", y = "Rent Price ($)",
       color = "ZIP Code",
       caption = "Data Source: Zillow Observed Rent Index (ZORI)") +
  theme(legend.position = "right")
```




```
# Ensure the "data" directory exists
dir.create(here("data"), showWarnings = FALSE)

# Read the cleaned dataset
la_housing_long <- read_csv(here("data", "cleaned_LA_housing.csv"), show_col_types = FALSE)

# Convert RegionName (ZIP codes) to character format to ensure compatibility
```

```r
la_housing_long <- la_housing_long %>%
  mutate(RegionName = as.character(RegionName))

# ---------------------------------
# Step 1: Define Wildfire-Affected ZIP Codes
# ---------------------------------
wildfire_zips <- list(
  "46 Fire" = c(92509, 92504, 92503, 92506, 92324, 92505, 92570, 92508, 91752, 92337, 92316, 92501),
  "Eagle Fire" = c(92881, 92503, 92882, 92879, 92883, 92570),
  "Easy Fire" = c(93065, 91360, 93021, 93063, 91320, 93012, 91362, 91361, 91307, 93015, 93066),
  "Getty Fire" = c(90049, 90025, 90024, 90272, 90403, 91403, 91436, 90402, 90077, 90095, 90073),
  "Saddle Ridge Fire" = c(91342, 91344, 91326, 91311, 91321, 91381, 93063, 91331),
  "Tick Fire" = c(91387, 91351, 91390, 91342, 93551, 91350, 91354, 91321, 91384)
)

# Convert wildfire ZIPs into a dataframe
wildfire_clusters <- stack(wildfire_zips) %>%
  rename(RegionName = values, Cluster = ind) %>%
  mutate(RegionName = as.character(RegionName))

# Merge wildfire cluster info with housing data
la_housing_long <- la_housing_long %>%
  left_join(wildfire_clusters, by = "RegionName") %>%
  mutate(Cluster = as.character(Cluster)) %>%
  mutate(Cluster = replace_na(Cluster, "Not Affected"))

# ---------------------------------
# Step 2: Compute LA County Benchmark
# ---------------------------------
# Calculate LA County average home value
la_avg_home_value <- la_housing_long %>%
  group_by(Date) %>%
  summarize(Avg_Home_Value = mean(Rent_Price, na.rm = TRUE))

# Merge LA County average with the main dataset
la_housing_long <- la_housing_long %>%
  left_join(la_avg_home_value, by = "Date")

# ---------------------------------
# Step 3: Generate Updated Visualization of Wildfire Impact
# ---------------------------------
ggplot(la_housing_long, aes(x = Date, y = Rent_Price, group = RegionName, color = Cluster)) +
  geom_line(alpha = 0.4, size = 0.8) +  # Wildfire-affected ZIPs
  geom_line(aes(y = Avg_Home_Value), color = "black", size = 1.2, linetype = "dashed") +  # LA County average
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +  # Fix overlapping x-axis labels
  theme_minimal() +
  labs(title = "Impact of LA Wildfires on Home Values",
       subtitle = "Dashed black line represents LA County average home value",
       x = "Year", y = "Home Value ($)",
       color = "Wildfire Cluster",
       caption = "Data Source: Zillow Home Value Index (ZHVI)")
```




## Citations

## Appendix