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.

Hi, I am Xavier Akers! I am a CMDA major who enjoys tackling data-driven problems. My favorite spot on campus is Owens, whether it is for a quick meal or just a place to relax. In my free time, I like to explore new places around Blacksburg and unwind from classes.

Data Introduction

Dataset Overview

Our dataset includes both the Zillow Observed Rent Index (ZORI) and the Zillow Home Value Index (ZHVI), providing a comprehensive view of rental and home value trends across different ZIP codes in the U.S. Our goal is to analyze the impact of LA wildfires on the housing market, examining potential correlations between wildfire events and changes in rental prices and home values in Los Angeles.

The ZORI score tracks typical rental prices by focusing on the middle range of rents, excluding extreme high and low values. It is designed to reflect rental housing prices for all homes, not just those currently listed for rent. The index is also smoothed to remove short-term fluctuations, offering a clearer picture of long-term trends. The dataset covers monthly rent values from January 2015 to January 2025 and includes categories such as All Homes, Single Family Residences, and Multi-Family Residences. For our analysis, we have chosen the All Homes Plus Multifamily Time Series ($) dataset, which can be accessed here.

The ZHVI score, on the other hand, measures typical home values and market changes for single-family homes and condominiums. Developed by Zillow, this index estimates home values within the 35th to 65th percentile range, providing insights into general market trends rather than outliers. Like ZORI, ZHVI is available as both a smoothed, seasonally adjusted measure and a raw measure, covering fluctuations from January 2015 to January 2025.

By analyzing both ZORI and ZHVI alongside wildfire data, we aim to identify patterns in how major wildfires impact housing prices and rental markets in Los Angeles. This research will help us understand whether wildfires lead to temporary price fluctuations, long-term market shifts, or increased disparities between different housing categories.

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

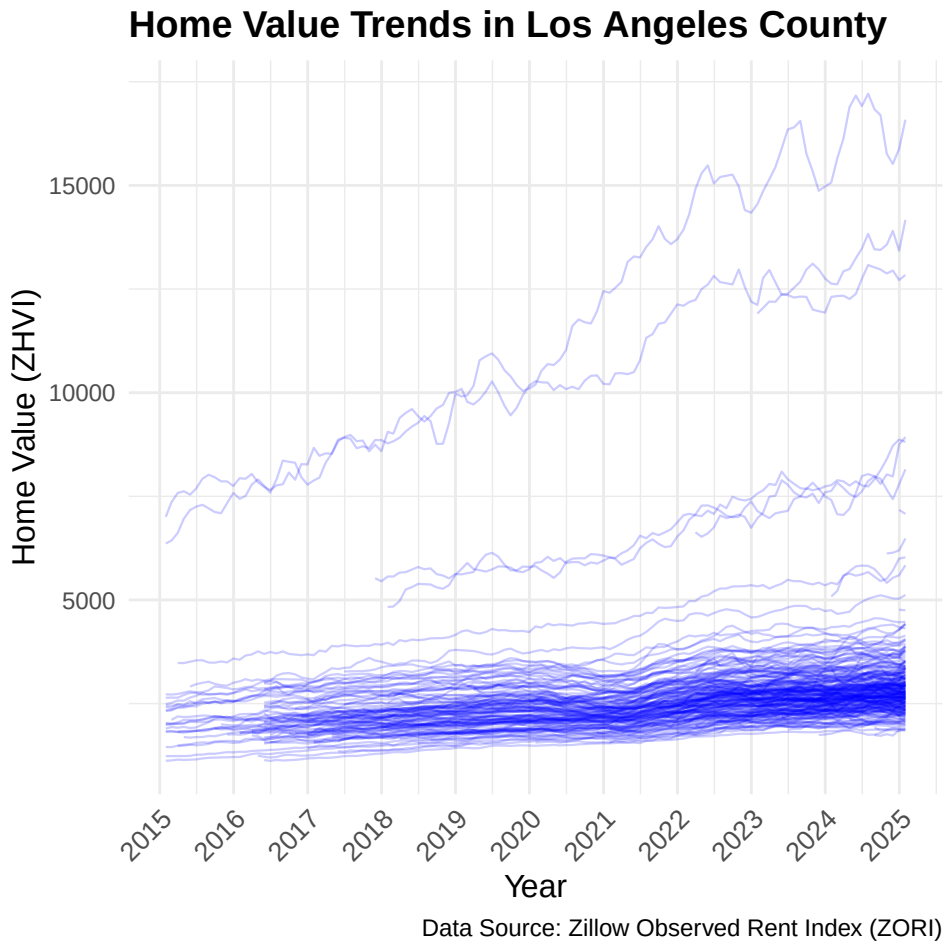
Simple Exploratory Data Analysis (EDA)

Initial Visualization

This code processes home value data for Los Angeles County. It first makes sure that the dataset uses a relative path. Then it filters the data to keep only Los Angeles County and fixes the date format to make it easier to work with. The data is reshaped from a wide format (with separate columns for each date) to a long format (where each row represents a single date and value). Missing home values are filled in using interpolation to keep the data smooth. The cleaned dataset is then saved, and a line chart is created to show how home values have changed over time across different ZIP codes in Los Angeles County.

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

# 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 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)

# 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
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
```

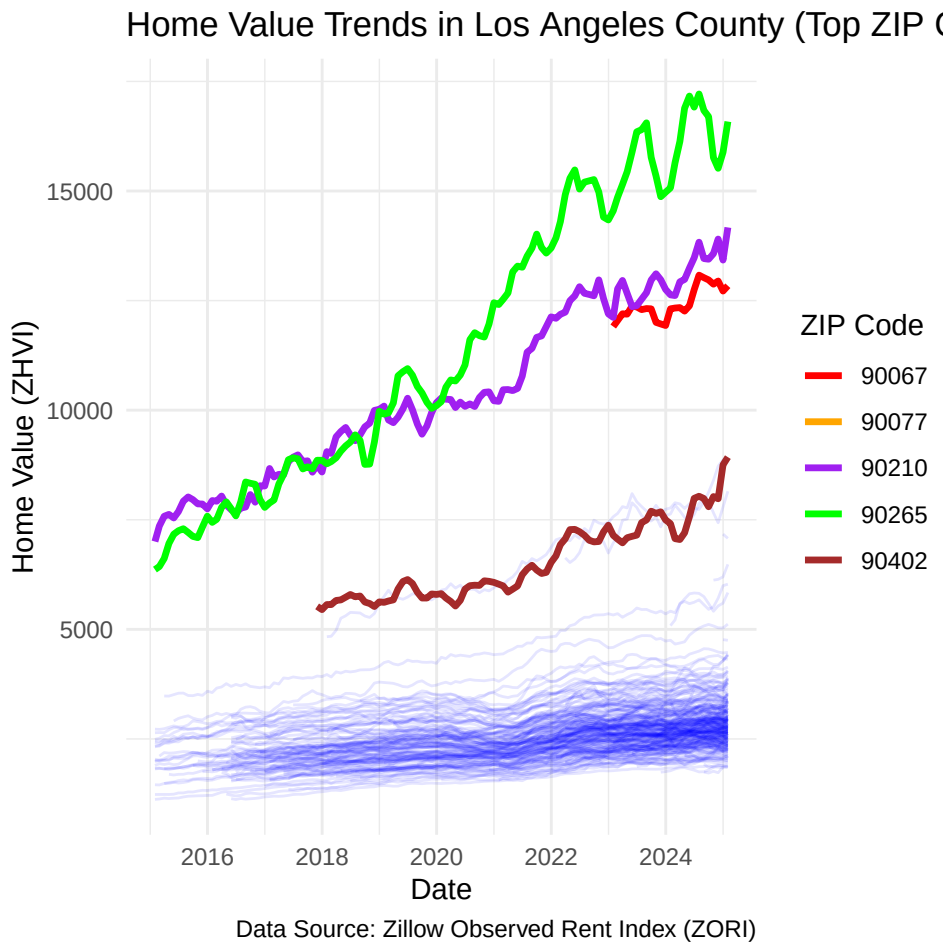
```r
la_housing_long <- read_csv(here("data", "cleaned_LA_housing.csv"), show_col_types = FALSE)

# 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
# Generate improved rental price trends plot
ggplot(la_housing_long, aes(x = Date, y = Rent_Price, group = RegionName)) +
  geom_line(alpha = 0.2, color = "blue", size = 0.4) +  # Adjust transparency and line thickness
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +  # Reduce x-axis clutter
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),  # Rotate and adjust x-axis labels
    axis.title = element_text(size = 12),  # Increase axis title font size
    plot.title = element_text(size = 14, face = "bold")  # Increase and bold the title
  ) +
  labs(
    title = "Home Value Trends in Los Angeles County",
    x = "Year", y = "Home Value (ZHVI)",
    caption = "Data Source: Zillow Observed Rent Index (ZORI)"
  )
```



Highlight

Here, we identify the top five ZIP codes in Los Angeles County with the highest recent home values. It first finds the latest available date in the dataset and filters the data to keep only the records for that date. The top five ZIP codes with the highest home values are selected. The plot then visualizes home value trends over time, with most ZIP codes shown in a faint blue for background reference. The top five ZIP codes are highlighted with distinct colors to make them stand out. A legend is added to indicate which ZIP codes have the highest home values. The reason we highlighted these five data points is based on some qualitative research regarding the wildfire, as we have seen in the news that more than half of the top-tier communities are infected by the wildfire. We will later use this as benchmark for our KNN generator.

```
# 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 = "Home Value Trends in Los Angeles County (Top ZIP Codes Highlighted)",
       x = "Date", y = "Home Value (ZHVI)",
       color = "ZIP Code",
       caption = "Data Source: Zillow Observed Rent Index (ZORI)") +
  theme(legend.position = "right")
```

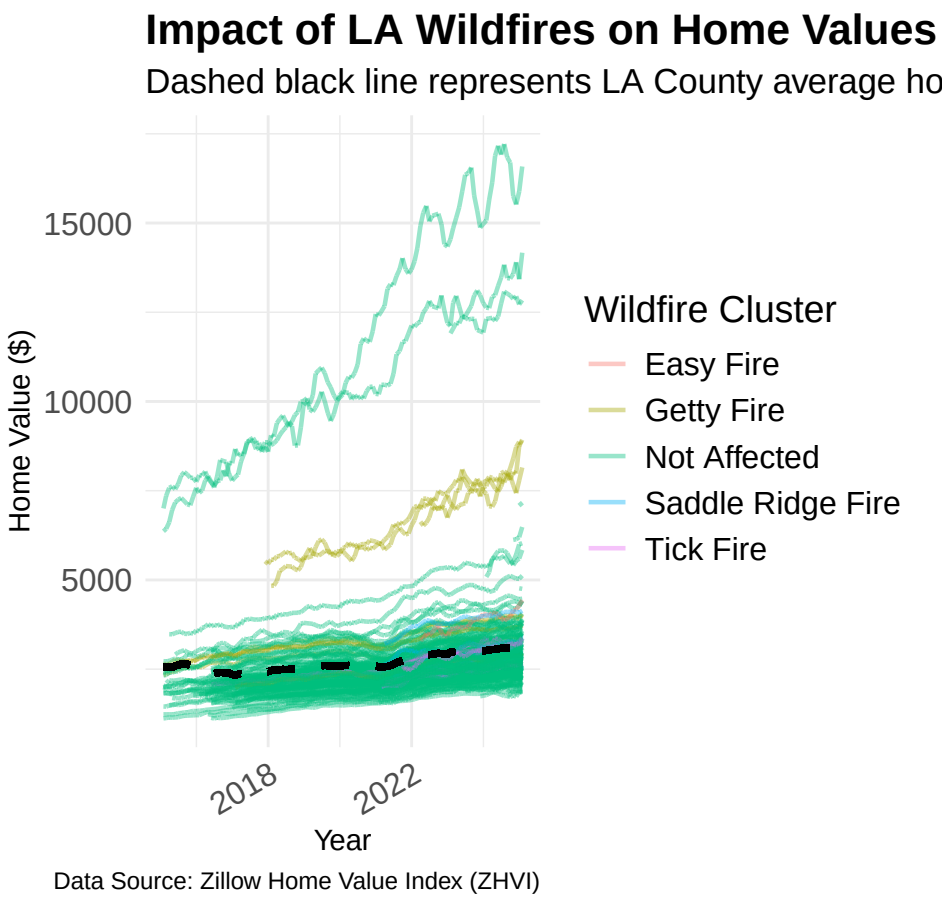


Wildfire Impacted Zip Codes

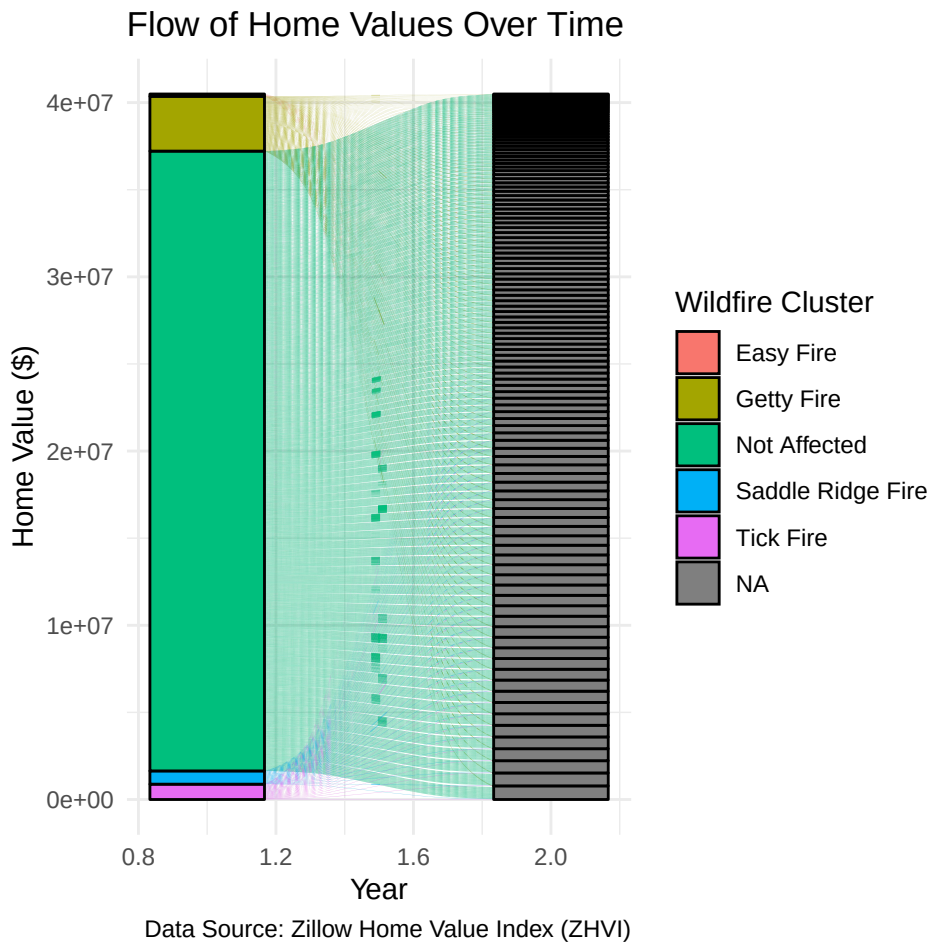
This code analyzes how wildfires have affected home values in Los Angeles County. First, it loads a cleaned dataset of home prices and ensures ZIP codes are properly formatted. Then, it identifies ZIP codes impacted by major wildfires and categorizes them accordingly based on some qualitative research. To provide a point of comparison, the script calculates the average home value across all of LA County. Finally, it visualizes the home value trends for both wildfire-affected areas and the county average, using a dashed black line to represent the overall trend. The goal is to observe whether wildfire-affected areas show different price patterns compared to unaffected regions.

```
# 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
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) +
  geom_line(aes(y = Avg_Home_Value), color = "black", size = 1.2, linetype = "dashed") +
  scale_x_date(date_labels = "%Y", date_breaks = "4 years", expand = c(0.05, 0.05)) +  # Space out x-axis labe
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1, size = 12),  # Rotate and resize x-axis labels
    axis.text.y = element_text(size = 12),
    legend.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 13),
    plot.margin = margin(10, 10, 20, 10)  # Increase margin to prevent cutoff
  ) +
  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)")
```




```
ggplot(la_housing_long, aes(axis1 = Cluster, axis2 = as.factor(Date), y = Rent_Price, fill = Cluster)) +
  geom_alluvium(aes(fill = Cluster)) +
  geom_stratum() +
  theme_minimal() +
  labs(title = "Flow of Home Values Over Time",
       x = "Year", y = "Home Value ($)",
       fill = "Wildfire Cluster",
       caption = "Data Source: Zillow Home Value Index (ZHVI)")
```



K-Nearest Neighbors (KNN) Analysis

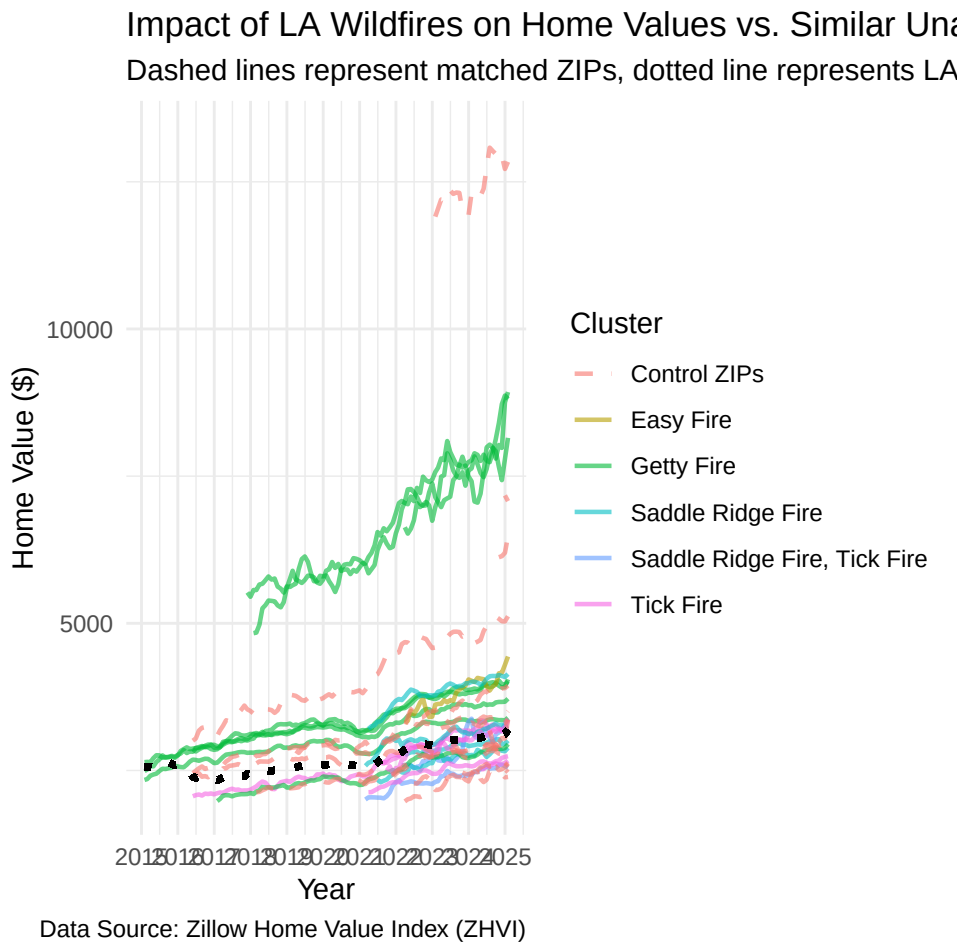
Assess Wildfire Impact on Homevalues Using KNN

This code analyzes the impact of LA wildfires on home values by comparing wildfire-affected ZIP codes to similar unaffected ZIP codes using K-Nearest Neighbors (KNN) regression.
First, it defines wildfire-affected ZIP codes and resolves cases where a ZIP code appears in multiple wildfire clusters. Then, it merges this information with housing price data. To create a control group, it applies KNN regression to find unaffected ZIP codes with similar historical home value trends. This ensures a fair comparison by using ZIP codes with similar market conditions before the wildfires.
Additionally, it calculates a county-wide benchmark to compare trends. Finally, it generates a visualization showing home value trends for affected areas, their KNN-matched control ZIPs (dashed lines), and the overall LA County average (dotted black line). This approach helps assess whether wildfires had a lasting impact on home prices while controlling for pre-existing market differences.

```
# Load necessary libraries
library(tidyverse)
library(here)
library(zoo)

# 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))

# ---------------------------------
# Step 2: Resolve Many-to-Many Relationship in Wildfire Clusters
# ---------------------------------

# If a ZIP code is listed in multiple wildfires, merge them into a single entry
wildfire_clusters <- wildfire_clusters %>%
  group_by(RegionName) %>%
  summarize(Cluster = paste(unique(Cluster), collapse = ", ")) %>%
  ungroup()

# ---------------------------------
# Step 3: Merge Wildfire Data with Housing Prices
# ---------------------------------
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 4: Find Comparable ZIP Codes for Control Group
# ---------------------------------
# Calculate median home value for each ZIP over the entire period
zip_summary <- la_housing_long %>%
  group_by(RegionName) %>%
  summarize(Median_Home_Value = median(Rent_Price, na.rm = TRUE))

# Get the median home values of affected ZIP codes
affected_summary <- zip_summary %>%
  filter(RegionName %in% wildfire_clusters$RegionName)

# Find non-affected ZIPs with similar median home values
non_affected_summary <- zip_summary %>%
  filter(!(RegionName %in% wildfire_clusters$RegionName))

# Match each affected ZIP to the closest non-affected ZIP based on median home value
matched_pairs <- affected_summary %>%
  rowwise() %>%
  mutate(Matched_Region = non_affected_summary$RegionName[which.min(abs(non_affected_summary$Median_Home_Value
  ungroup()
```

```r
# Merge matched control ZIPs back into main dataset
la_housing_long <- la_housing_long %>%
  left_join(matched_pairs %>% select(RegionName, Matched_Region), by = "RegionName")

# ---------------------------------
# Step 5: Compute LA County Benchmark
# ---------------------------------
# Calculate LA County average home value over time
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 6: Generate Visualization Comparing Affected vs. Control ZIPs
# ---------------------------------
ggplot(la_housing_long, aes(x = Date, y = Rent_Price, group = RegionName, color = Cluster)) +
  geom_line(data = la_housing_long %>% filter(Cluster != "Not Affected"), alpha = 0.6, size = 0.8) +  # Wildfi
  geom_line(data = la_housing_long %>% filter(RegionName %in% matched_pairs$Matched_Region), aes(color = "Cont
  geom_line(aes(y = Avg_Home_Value), color = "black", size = 1.2, linetype = "dotted") +  # LA County average
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +  # Fix overlapping x-axis labels
  theme_minimal() +
  labs(title = "Impact of LA Wildfires on Home Values vs. Similar Unaffected ZIPs",
       subtitle = "Dashed lines represent matched ZIPs, dotted line represents LA County average",
       x = "Year", y = "Home Value ($)",
       color = "Cluster",
       caption = "Data Source: Zillow Home Value Index (ZHVI)")
```



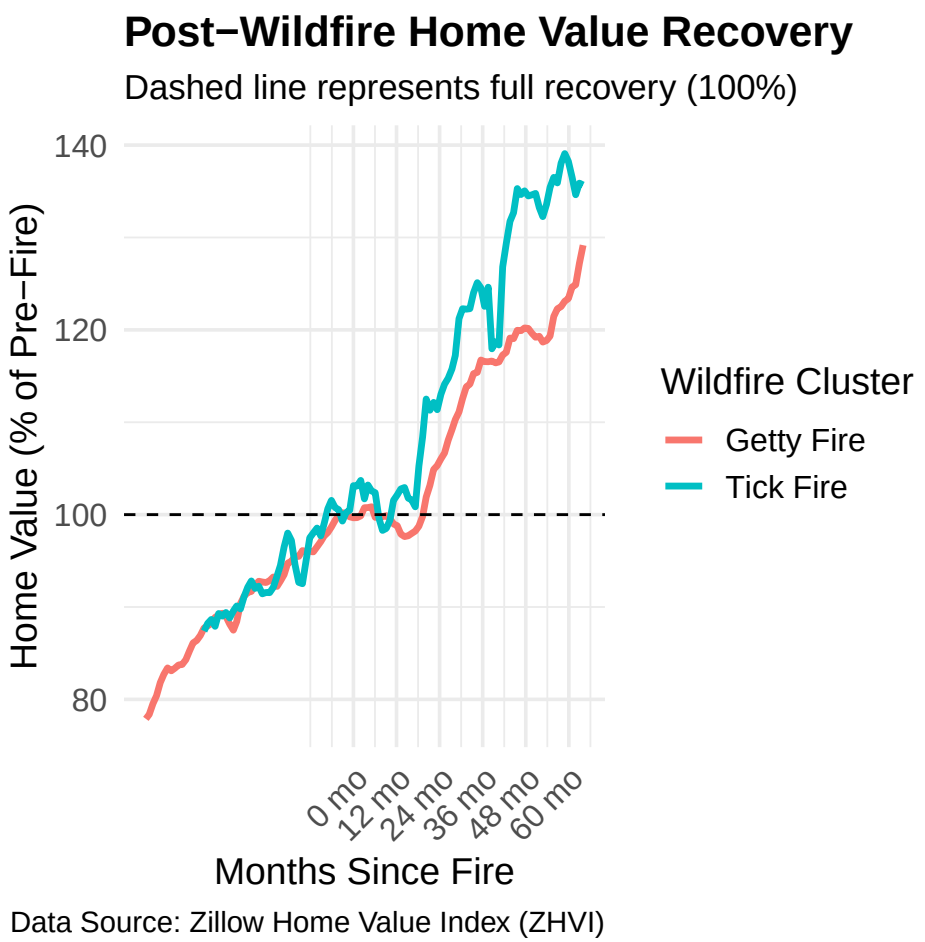
Estimate Recover Post-Wildfire Using KNN

In this analysis, we aim to understand how long it takes for home values to recover after a wildfire. Using the Zillow Home Value Index (ZHVI), we first identify ZIP codes affected by major wildfires in Los Angeles. For each affected ZIP code, we determine its pre-fire average home value using data from three months before the fire.
To create a meaningful comparison, we apply K-Nearest Neighbors (KNN) to find similar, unaffected ZIP codes based on their pre-fire home values. This allows us to establish a control group to measure the recovery trend. We then calculate the percentage change in home value over time, normalizing each ZIP code's value relative to its pre-fire average.
Finally, we generate a visualization that tracks the recovery timeline of home values, showing how they progress month by month after the fire. The dashed line at 100

```
# Load necessary libraries
library(tidyverse)
library(here)
library(zoo)
library(FNN)  # For KNN

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

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

```r
# ---------------------------------
# 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 Pre-Fire Home Values
# ---------------------------------
fire_dates <- data.frame(
  Cluster = c("46 Fire", "Eagle Fire", "Easy Fire", "Getty Fire", "Saddle Ridge Fire", "Tick Fire"),
  Fire_Date = as.Date(c("2019-10-31", "2019-07-31", "2019-10-31", "2019-10-31", "2019-10-31", "2019-10-31"))
)

# Merge fire dates with wildfire data
wildfire_data <- la_housing_long %>%
  left_join(fire_dates, by = "Cluster") %>%
  filter(!is.na(Fire_Date))  # Keep only wildfire-affected ZIP codes

# Calculate pre-fire average home value (3 months before the fire)
wildfire_data <- wildfire_data %>%
  group_by(RegionName) %>%
  mutate(Pre_Fire_Avg = mean(Rent_Price[Date < Fire_Date & Date >= (Fire_Date - months(3))], na.rm = TRUE)) %>
  ungroup()

# Save the modified dataset
write_csv(wildfire_data, here("data", "wildfire_home_values.csv"))

# ---------------------------------
# Step 3: Find Similar ZIPs Using KNN
# ---------------------------------
# Create unaffected ZIP dataset
unaffected_zips <- la_housing_long %>%
  filter(Cluster == "Not Affected") %>%
  group_by(RegionName) %>%
  summarize(Pre_Fire_Avg = mean(Rent_Price[Date < "2019-10-31"], na.rm = TRUE), .groups = "drop")

# Remove NAs from the dataset before applying KNN
unaffected_zips_clean <- unaffected_zips %>% drop_na(Pre_Fire_Avg)
wildfire_data_clean <- wildfire_data %>% drop_na(Pre_Fire_Avg)

# Ensure enough data is available
if (nrow(unaffected_zips_clean) > 0 & nrow(wildfire_data_clean) > 0) {
```

```r
  # Use KNN to find the closest unaffected ZIP for each wildfire-affected ZIP
  knn_result <- get.knnx(
    data = matrix(unaffected_zips_clean$Pre_Fire_Avg, ncol = 1),  # Feature space: Pre-Fire Home Value
    query = matrix(wildfire_data_clean$Pre_Fire_Avg, ncol = 1),
    k = 1  # Find the single best match
  )

  # Match wildfire-affected ZIPs to similar unaffected ZIPs
  wildfire_data_clean$Matched_ZIP <- unaffected_zips_clean$RegionName[knn_result$nn.index]
} else {
  stop("Error: Not enough data after removing NAs.")
}

# Save matched ZIPs for control comparison
write_csv(wildfire_data_clean, here("data", "matched_zip_comparison.csv"))

# ---------------------------------
# Step 4: Analyze Home Value Recovery
# ---------------------------------
wildfire_data_clean <- wildfire_data_clean %>%
  mutate(Time_Since_Fire = as.numeric(difftime(Date, Fire_Date, units = "days") / 30)) %>%  # Convert days to
  mutate(Normalized_Value = Rent_Price / Pre_Fire_Avg * 100)  # Normalize to % of pre-fire value

# Compute recovery statistics
recovery_stats <- wildfire_data_clean %>%
  group_by(Cluster, Time_Since_Fire) %>%
  summarize(Avg_Home_Value = mean(Normalized_Value, na.rm = TRUE), .groups = "drop")

# Save recovery statistics
write_csv(recovery_stats, here("data", "recovery_stats.csv"))

# ---------------------------------
# Step 5: Plot Recovery Trends
# ---------------------------------
ggplot(recovery_stats, aes(x = Time_Since_Fire, y = Avg_Home_Value, color = Cluster)) +
  geom_line(size = 1.2) +
  geom_hline(yintercept = 100, linetype = "dashed", color = "black") +
  scale_x_continuous(breaks = seq(0, max(recovery_stats$Time_Since_Fire, na.rm = TRUE), by = 12),  # Show ever
                     labels = function(x) paste(x, "mo")) +  # Add "mo" for clarity
  theme_minimal(base_size = 14) +  # Increase base font size for readability
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 12),  # Rotate x-axis labels
    axis.text.y = element_text(size = 12),
    legend.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 13)
  ) +
  labs(title = "Post-Wildfire Home Value Recovery",
       subtitle = "Dashed line represents full recovery (100%)",
       x = "Months Since Fire", y = "Home Value (% of Pre-Fire)",
       color = "Wildfire Cluster",
       caption = "Data Source: Zillow Home Value Index (ZHVI)")
```



According to Gregory Eubanks, a Redfin Premier Agent, this inventory shortage of rentals is creating bidding wars for the short supply available: "Tons of past clients are reaching out on behalf of friends, seeing if I know of any available rentals. There's competition for nearly every rental, and it's not just on price; a lot of people are taking on long leases to secure a place to live. A rental listed for $16,000 per month got bid up to $30,000, and the winners took on a two-year lease. On the buying and selling side, people are pulling back, waiting for the dust to settle. Two buyers have canceled deals because they don't feel comfortable making such a big purchase with the catastrophe going on. Three clients have canceled their listings, with the homeowners opting to rent their homes out to people impacted by the fires instead."

The news report talks about how wildfires caused a shortage of rental homes, leading to bidding wars and higher prices. This connects directly to what we see in our graph, where home values dropped right after the fire but quickly bounced back and even went higher than before. Since many homeowners chose to rent out their houses instead of selling, fewer homes were available to buy, which drove up demand and prices. The report also mentions that buyers were holding off on big purchases, but at the same time, renters were paying more to secure homes, which could have pushed home values higher in the long run. Our graph supports this by showing that home values didn't just recover—they surged well beyond pre-fire levels, likely because of the limited housing supply and high demand after the fires.

NaiveBayes classification

The preprocessing steps can be found in the Appendix section.

We aim to analyze the **Rental Price Index** in areas of wildfires. Specifically, we will be looking at the location of the Easy Fire (October 2019), Getty Fire (October 2019), and the Saddle Ridge Fire (October 2019). The ZIP codes of each fire are stored in `easy_fire_zips`, `getty_fire_zips`, and `saddle_ridge_zips`.

```
easy_fire_zips = c(93065, 91360, 93021, 93063, 91320, 93012, 91362, 91361, 91307, 93015, 93066)
getty_fire_zips = c(90049, 90025, 90024, 90272, 90403, 91403, 91436, 90402, 90077, 90095, 90073)
saddle_ridge_fire_zips = c(91342, 91344, 91326, 91311, 91321, 91381, 93063, 91331, 91350, 91355, 91387, 91304,
fire_zips = c(easy_fire_zips, getty_fire_zips, saddle_ridge_fire_zips)
```

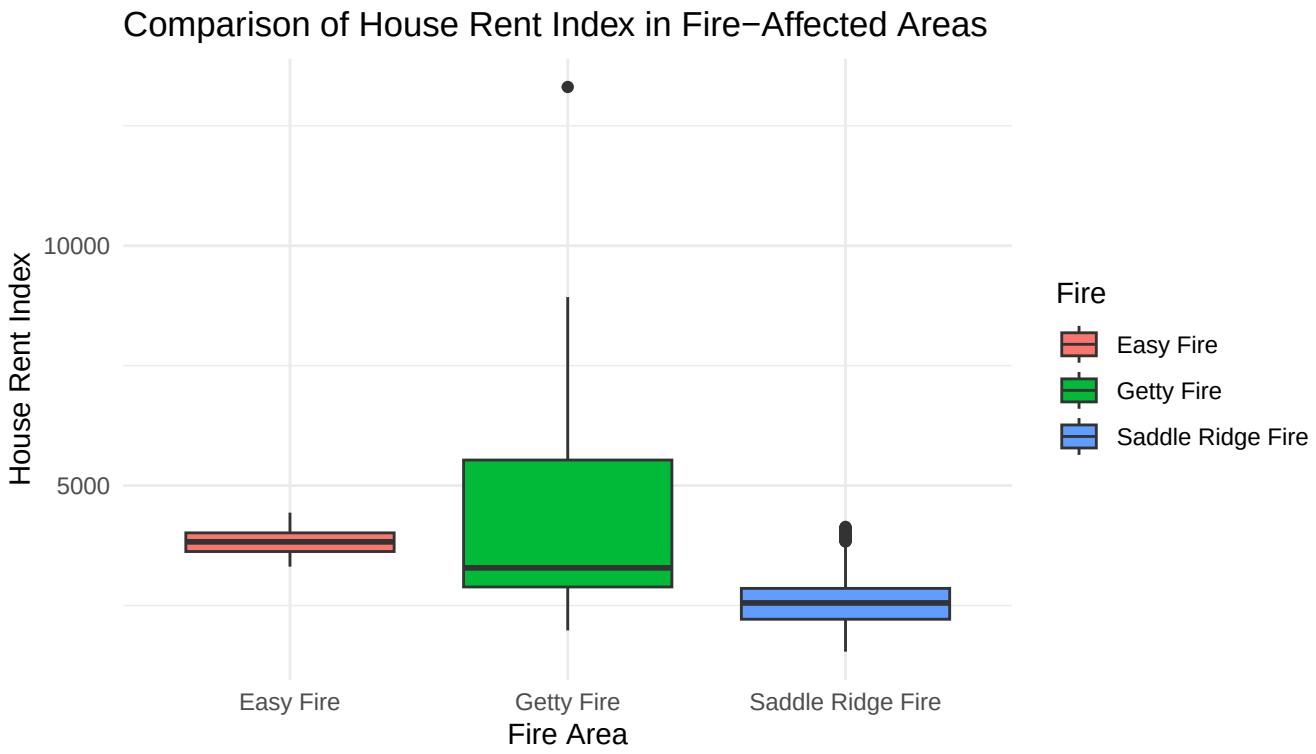
A sample of the Los Angeles Fire Housing data for only these ZIP codes is shown below.

```
  RegionID SizeRank RegionName RegionType StateName State        City
1    96368       37      91342        zip        CA    CA Los Angeles
2    96368       37      91342        zip        CA    CA Los Angeles
3    96368       37      91342        zip        CA    CA Los Angeles
                              Metro         CountyName       Date Rent_Price
1 Los Angeles-Long Beach-Anaheim, CA Los Angeles County 2022-02-28   2588.083
2 Los Angeles-Long Beach-Anaheim, CA Los Angeles County 2022-03-31   2592.541
3 Los Angeles-Long Beach-Anaheim, CA Los Angeles County 2022-04-30   2597.000
               Fire
1 Saddle Ridge Fire
2 Saddle Ridge Fire
3 Saddle Ridge Fire
```

Location Comparisons

The following boxplot displays the distribution of rent prices in ZIP codes affected by the different wildfires. From the distribution of rent prices, we can make educated assumptions about each location's rental market.

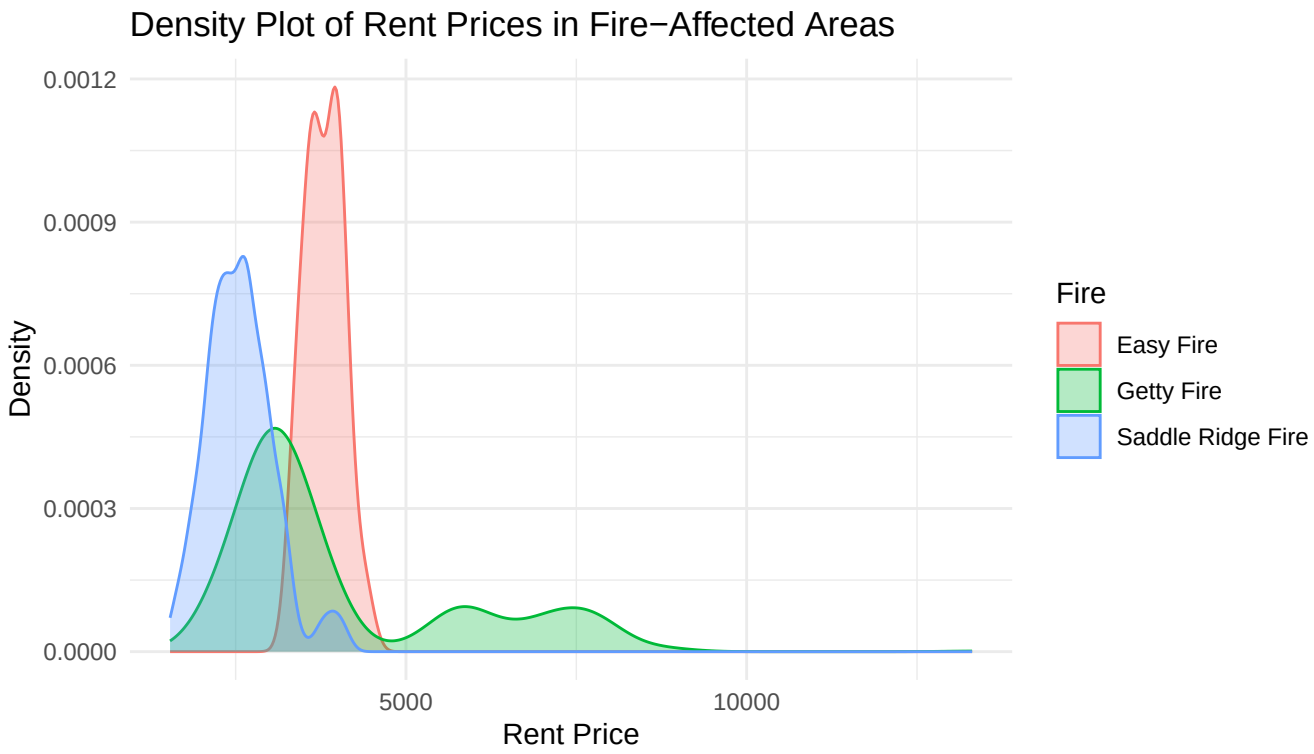
```
ggplot(fire_housing_data, aes(x = Fire, y = Rent_Price, fill = Fire)) +
  geom_boxplot() +
  labs(
    title = "Comparison of House Rent Index in Fire-Affected Areas",
    x = "Fire Area",
    y = "House Rent Index"
  ) +
  theme_minimal()
```



Rental indexes in the Easy Fire region were generally higher than the other regions, with very few extreme values. The Getty Fire region had a wider range of rental prices, including some outliers. This suggest the Getty Fire also included some wealthier or more in demand areas. The Saddle Ridge Fire region had a more concentrated rental prices, with a few high-end outliers.

The following density plot visualizes each of the wildfire locations. This helps us better understand underlying distribution of the data.

```
ggplot(fire_housing_data, aes(x = Rent_Price, color = Fire, fill = Fire)) +
  geom_density(alpha = 0.3) +
  labs(
    title = "Density Plot of Rent Prices in Fire-Affected Areas",
    x = "Rent Price",
    y = "Density"
  ) +
  theme_minimal()
```



This visual reinforces our assumptions from the prior boxplot. Residences in the Easy Fire region experience lower variance in rental price index. Houses in the Saddle Ridge fire are generally lower than the other two regions, with a few outliers, potentially indicating a more expensive or in demand area nearby. The Getty fire region has the largest variance, with one large peak and two smaller peaks on the high end. This indicate that the Saddle Ridge Fire spread to a broader range of communities.

Below is a table displaying the summary statistics of the rental prices in the fire-affected areas.

```
summary_table <- fire_housing_data %>%
  group_by(Fire) %>%
  summarize(
    Mean_Rent = mean(Rent_Price, na.rm = TRUE),
    Median_Rent = median(Rent_Price, na.rm = TRUE),
    Min_Rent = min(Rent_Price, na.rm = TRUE),
    Max_Rent = max(Rent_Price, na.rm = TRUE),
    Std_Dev = sd(Rent_Price, na.rm = TRUE)
  )

print(summary_table)
```

```
# A tibble: 3 x 6
  Fire              Mean_Rent Median_Rent Min_Rent Max_Rent Std_Dev
  <chr>                 <dbl>       <dbl>    <dbl>    <dbl>   <dbl>
1 Easy Fire             3800.       3827.    3310.    4436.    278.
2 Getty Fire            4067.       3284.    1980.   13308.   1766.
3 Saddle Ridge Fire     2563.       2553.    1538.    4138.    487.
```

Predictive Analysis

We wish to determine how well Naive Bayes can predict different wildfires based on features such as `RegionName`, `CountyName`, `Date`, and `Rent_Price`.

Since Naive Bayes works best with discrete data, we discretize `Rent_Price` into `Low`, `Medium`, and `High`. Additionally, we extract the `Year` and `Month`. This preprocessing code can be viewed in the Appendix under Naive Bayes Preprocessing.

Modeling

We create our Naive Bayes model as follows. Note we split our data using the 80/20 rule.

```
library(e1071) # For Naive Bayes

set.seed(0)
train_idx = sample(1:nrow(nb_df), size = 0.8 * nrow(nb_df))
train_data = nb_df[train_idx,]
test_data = nb_df[-train_idx,]

nb_model = naiveBayes(Fire ~ Rent_Category + Year + Month + RegionName, data=train_data)
```

We can see how well our model did at predicting Fires from the given factors.

```
yhat = predict(nb_model, test_data) # Validate the model

tab = table(yhat, test_data$Fire)
misclass = (sum(tab)- sum(diag(tab))) / sum(tab)
accuracy = 1 - misclass
```

Results

We have an accuracy rate of 0.9322034 and the following confusion matrix,

```
yhat                Easy Fire Getty Fire Saddle Ridge Fire
  Easy Fire                 9          0                 0
  Getty Fire                0        141                 7
  Saddle Ridge Fire         0         17               180
```

We see the model was successful at classifying the wildfire based on different features like `Rent_Category`, `Year`, `Month`, and `RegionName`. The high accuracy (93.22%) suggests that the model is able to successfully predict the different wildfires. The Easy Fire was classified perfectly, while the Getty Fire and Saddle Ridge Fire had very few cases misclassified.

Therefore, the Naive Bayes classifier was able to successfully predict wildfire regions using rent indexes, time, and region information for the Easy Fire, Getty Fire, and Saddle Ridge Fire.

Appendix

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

```
# 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)]

# 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"))
```

Data Preprocessing Code

This code extracts only the LA housing data and converts the date from wide format to long format.

```
# 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"))
```

Naive Bayes Preprocessing

```r
# Extract Year and Month from Date
nb_df = fire_housing_data
nb_df$Date = as.Date(nb_df$Date)
nb_df$Year = format(nb_df$Date, "%Y")
nb_df$Month = format(nb_df$Date, "%m")

# Discretize Rent_Price to low, medium, high
nb_df$Rent_Category = cut(nb_df$Rent_Price, breaks=3, labels=c("Low", "Medium", "High"))

nb_df$Fire = as.factor(nb_df$Fire)
nb_df$Year = as.factor(nb_df$Year)
nb_df$Month = as.factor(nb_df$Month)
nb_df$Rent_Category = as.factor(nb_df$Rent_Category)
```

Citations

https://themortgagepoint.com/2025/01/24/l-a-wildfires-where-does-the-housing-market-go-from-here/#:~:text=The%20sizable%20