CMDA-4654

Group Project 1

Yusi Yao & Xavier Akers

2025-02-18

Teammate Introduction

Meet Yusi Yao! Born in Nanjing, China, Yusi is always on the go-whether it is heading back to D.C. on weekends or enjoying the outdoors. When in Blacksburg, Rainbowl is the go-to spot for a great meal. In free time, Yusi enjoys fishing and playing soccer, making the most of both nature and sports.

Data Introduction

Dataset Overview

This dataset tracks rental price trends across different ZIP codes in the U.S. It includes rental price estimates for single-family homes and condominiums. The data covers monthly rent values from **January 2015 to January 2025**.

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

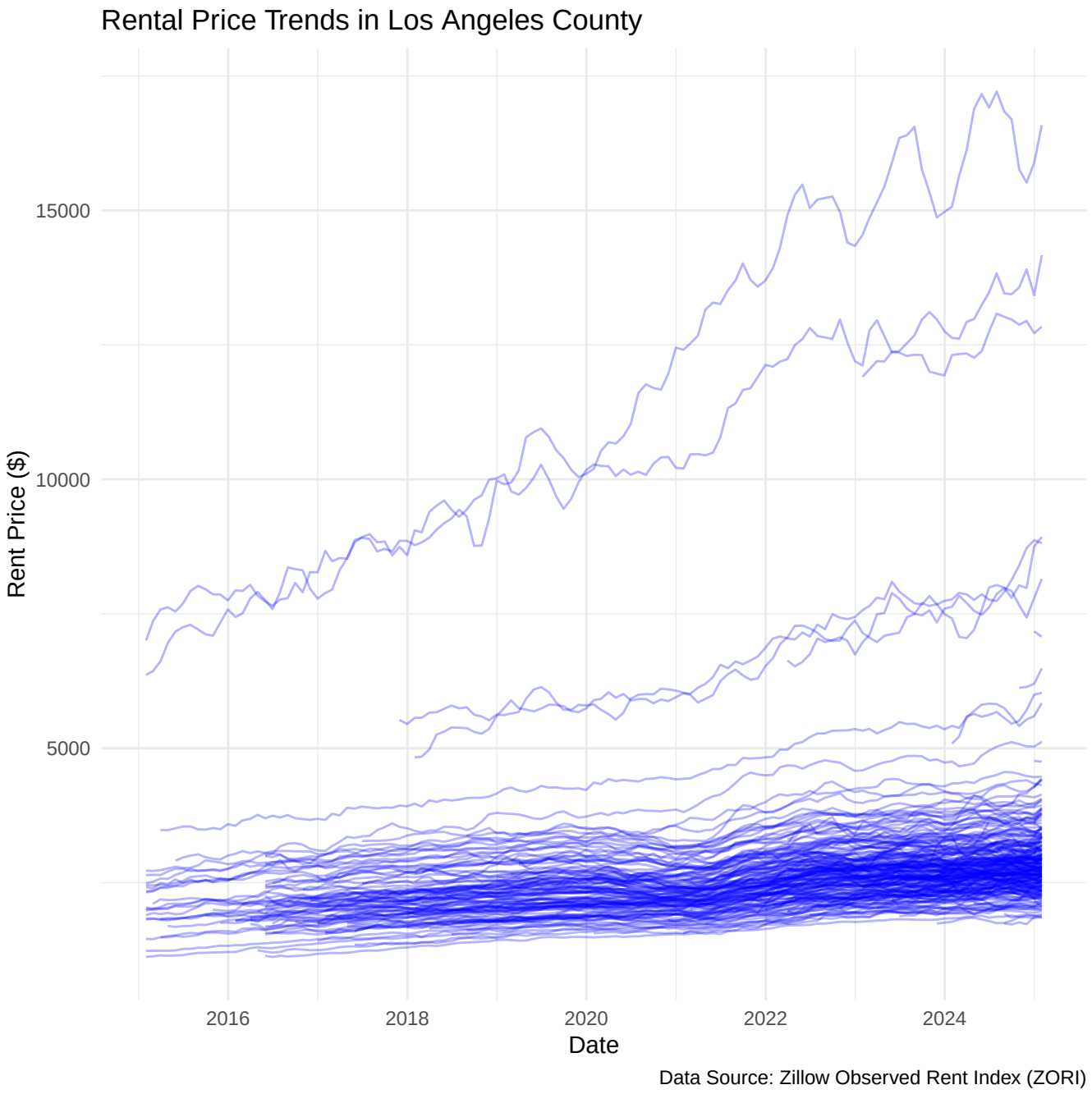
1. Filter ZIP codes in Los Angeles County
2. Remove columns with too many missing values
3. Handle missing rent prices using interpolation
4. Ensure date columns are in the correct format

```
# Read the dataset
Zip_zori_uc_sfrcondomfr_sm_month <- read_csv("~/Desktop/VT/4654_DAML/group_project1/Zip_zori_uc_sfrcondomfr_sm

# 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")
```

```r
# Identify date columns (they have format "M/DD/YY")
date_columns <- names(la_housing)[10:ncol(la_housing)]  # First 9 columns are location info

# Convert date format to "YYYY-MM-DD"
formatted_colnames <- c(names(la_housing)[1:9],  # Keep non-date columns unchanged
                        format(as.Date(date_columns, format = "%m/%d/%y"), "%Y-%m-%d"))

# Rename columns in the dataset
colnames(la_housing) <- formatted_colnames

# Convert wide format to long format
la_housing_long <- la_housing %>%
  pivot_longer(cols = starts_with("20"), 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
write_csv(la_housing_long, "~/Desktop/VT/4654_DAML/group_project1/cleaned_LA_housing.csv")
```

Rental Price Trends in Los Angeles County

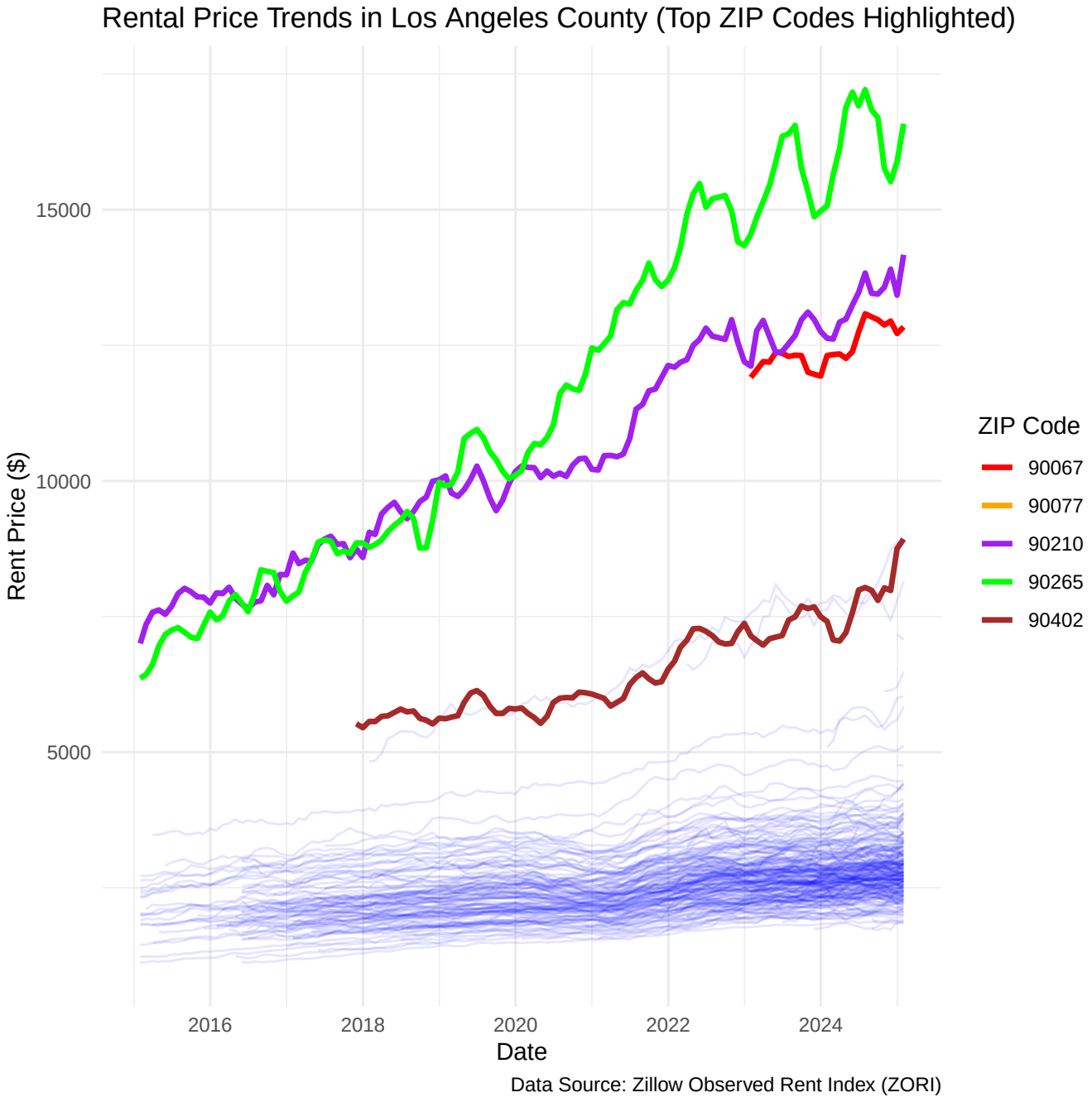
```r
# Load the cleaned dataset
la_housing_long <- read_csv("~/Desktop/VT/4654_DAML/group_project1/cleaned_LA_housing.csv")

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

# 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") +
  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")
```



Citations

Appendix