CMDA-4654

Xavier Akers

2025-02-19

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

Our data is the **Zillow Observed Rent Index (ZORI)**. The ZORI score tracks typical rental prices in a given area. This index attempts to represent an accurate rental housing stock by focusing on the middle range of rents, excluding very high and very low prices. It is designed to depict rental housing prices for all homes, not only homes currently listed for-rent. Additionally, it is smoothed out to remove short-term spikes and provide a better understanding of long-term trends. The data covers monthly rent values from **January 2015 to January 2025**.

The ZORI score is taken for the categories: All homes, Single Family Residences, and Multi-Family Residences. We have chosen to analyze the **All Homes Plus Multifamily Time Series ($)** dataset. It can be downloaded directly here

Data Source

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

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 Category

This dataset belongs to category 8 **Housing**. In this project, we intend to look at Los Angeles wildfire's in relation to the housing market in the ZIP codes of Los Angeles.

Analysis & Discussion

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)
```

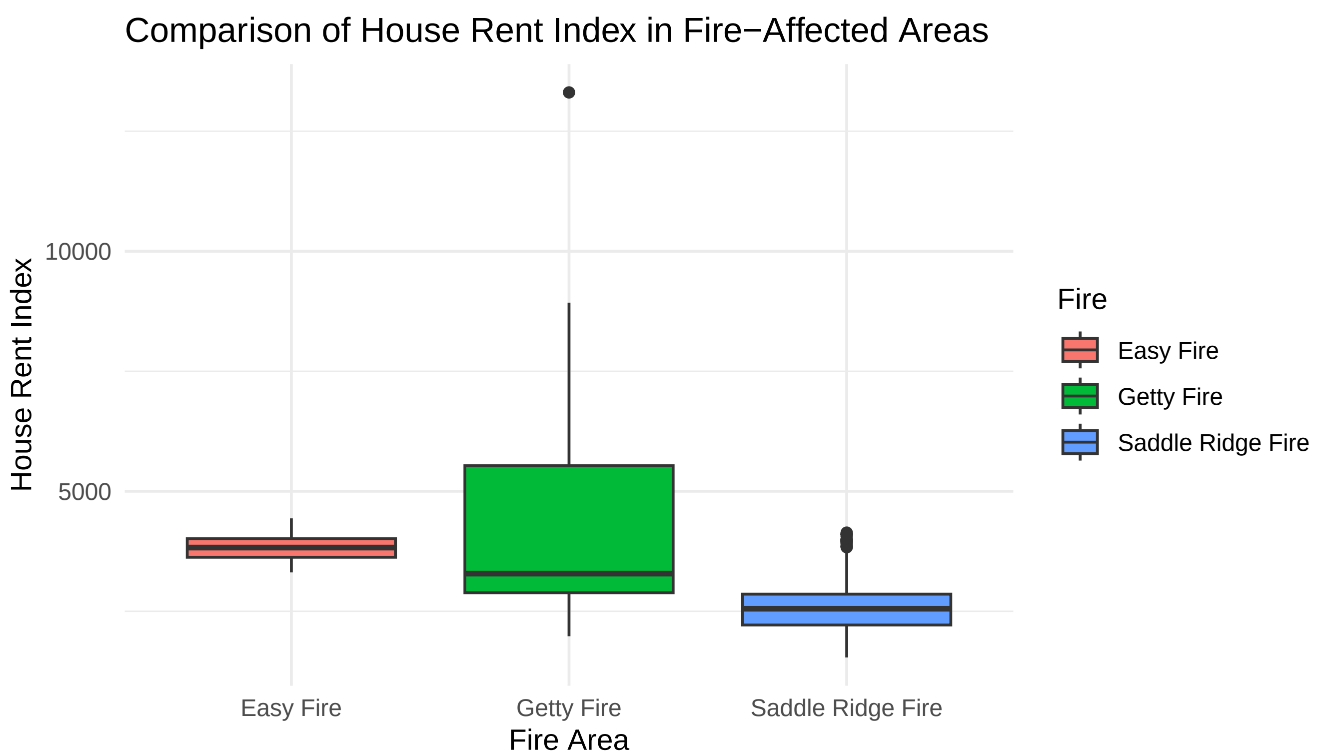
A sample of the Los Angeles Fire Housing data for only these ZIP codes is shown below.

```
  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.

```r
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.

```r
ggplot(fire_housing_data, aes(x = Rent_Price, color = Fire, fill = Fire)) +
```

```r
  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.

```r
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 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] "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"
```

```
# 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)
```

Citation

https://themortgagepoint.com/2025/01/24/l-a-wildfires-where-does-the-housing-market-go-from-here/#