GROUP PROJECT CSC 3220 - Real Estate Data

Robert Bingham
Phonethep Nakhonekhong
Eli Parker
John Taylor
Johnathan Rich

2022-11-08

Introduction

Our group, 5 little minds, have decided to look at real estate market data from 2012 to 2021 in order to predict future trends in the next 5 years. This data includes median list price for a given neighborhood in each state; median prices based upon housing unit categories, such as apartments, single-family housing, and condos; and year-to-year increases in sale prices of each unit.

Format the data

For the purposes of this assignment, were are turning off all warnings and centering each graph.

Import the Data

We decided to use the data from this url from Kaggle for our dataset:

https://www.kaggle.com/datasets/thuynyle/redfin-housing-market-data?select=state_market_tracker.tsv000

```
state_market.df <- read.table("../data/state_market_tracker.tsv000", sep = '\t', header =
↪  TRUE)
```

Import the necessary libraries

```
library("ggplot2")
library("DT")
library("pander")
library("corrplot")
library("zoo")
library("reshape")
library("scales")
```

Data Manipulation

Here, we have made R recognize the variables in the dataset that pertain to specific days, (i.e, 9/21/2022) as actual dates using the built-in as.Date function. We have also divided the median sale price and list price of homes in each neighborhood by 1000 in order to make the data more readable in subsequent graphs.

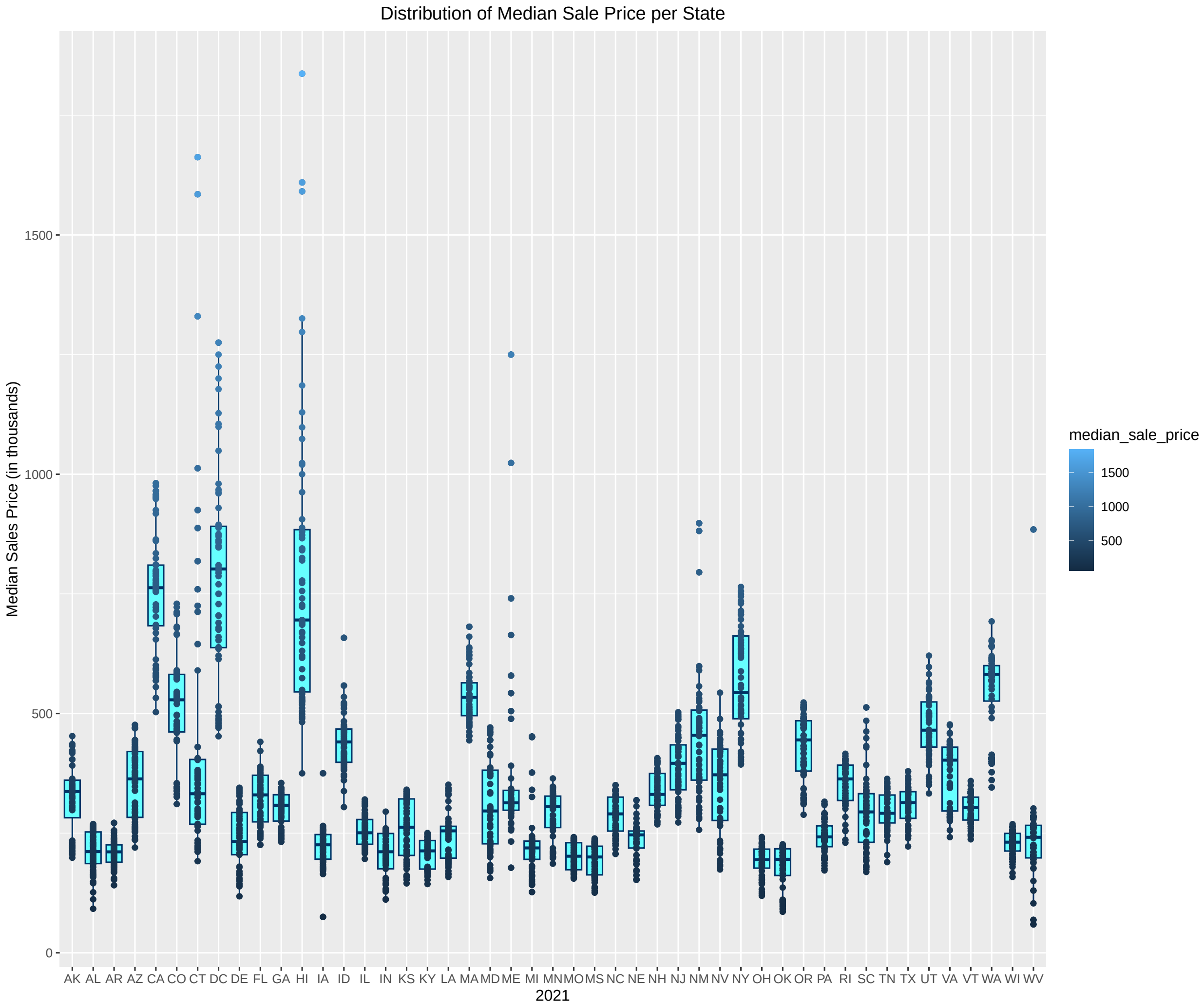
```
state_market.df$period_begin <- as.Date(state_market.df$period_begin)
state_market.df$period_end <- as.Date(state_market.df$period_end)
state_market.df$median_sale_price <- state_market.df$median_sale_price / 1000
state_market.df$median_list_price <- state_market.df$median_list_price / 1000
state_market.df$property_type[state_market.df$property_type == "Multi-Family (2-4 Unit)"]
↪  <- "Multi-Family"
```

Here, we separate the data into categorical and numeric variables. We decided that the state is a categorical variable, and that the prices, increase in prices, and the dates were numeric variables.

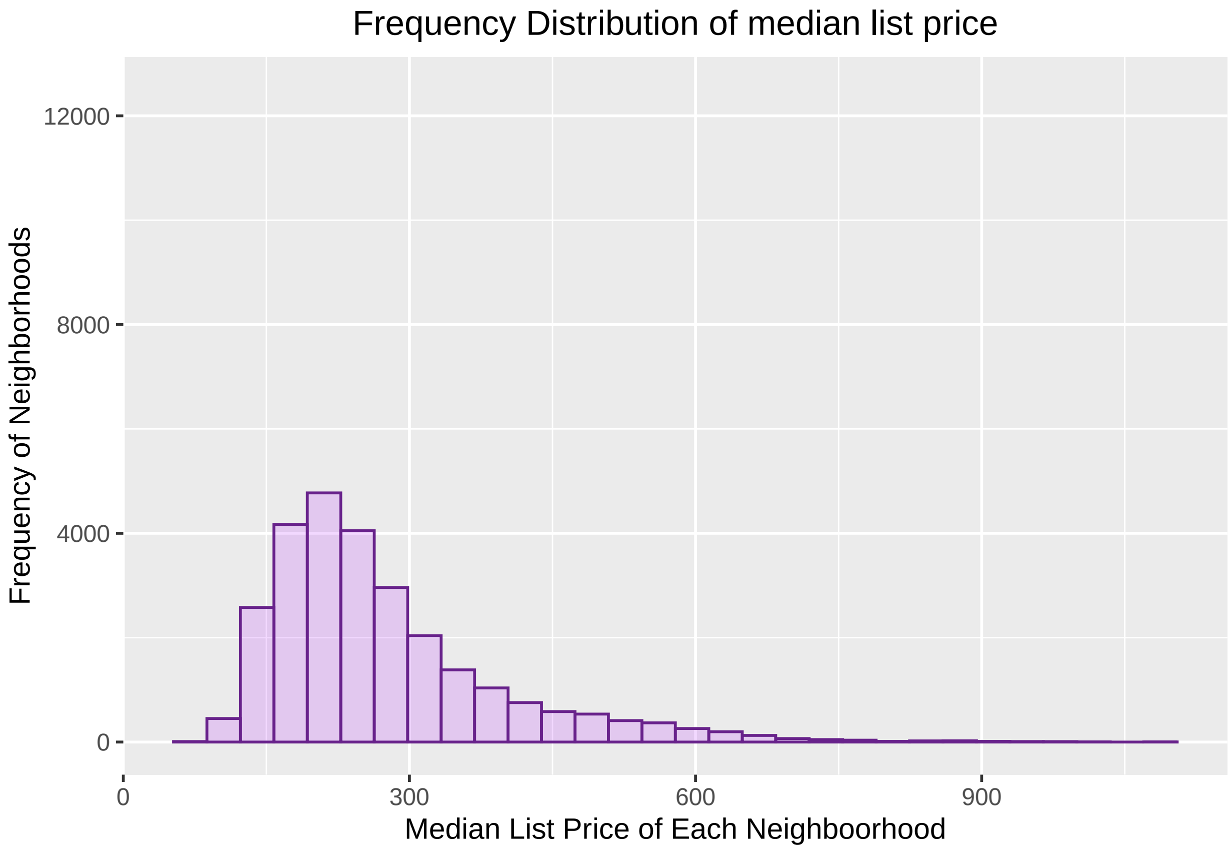
```
state_market.df$period_begin <- as.Date(state_market.df$period_begin)

variables <- colnames(state_market.df)
numeric_vars <- variables[c(-1,-48)]
categorical_vars <- variables[c(1)]
```

Here, we created a boxplot graph for the median list price of the homes in each neighborhood in the dataset, with each boxplot representing . Hawaii and and Connecticut were had the highest number of outliers. Since there were more outliers above the boxplot than below it in both cases, we attributed these anomalies to the high cost of living in both states, Since Hawaii is a vacation destination and Connecticut is a New England state with a close proximity to New York City.

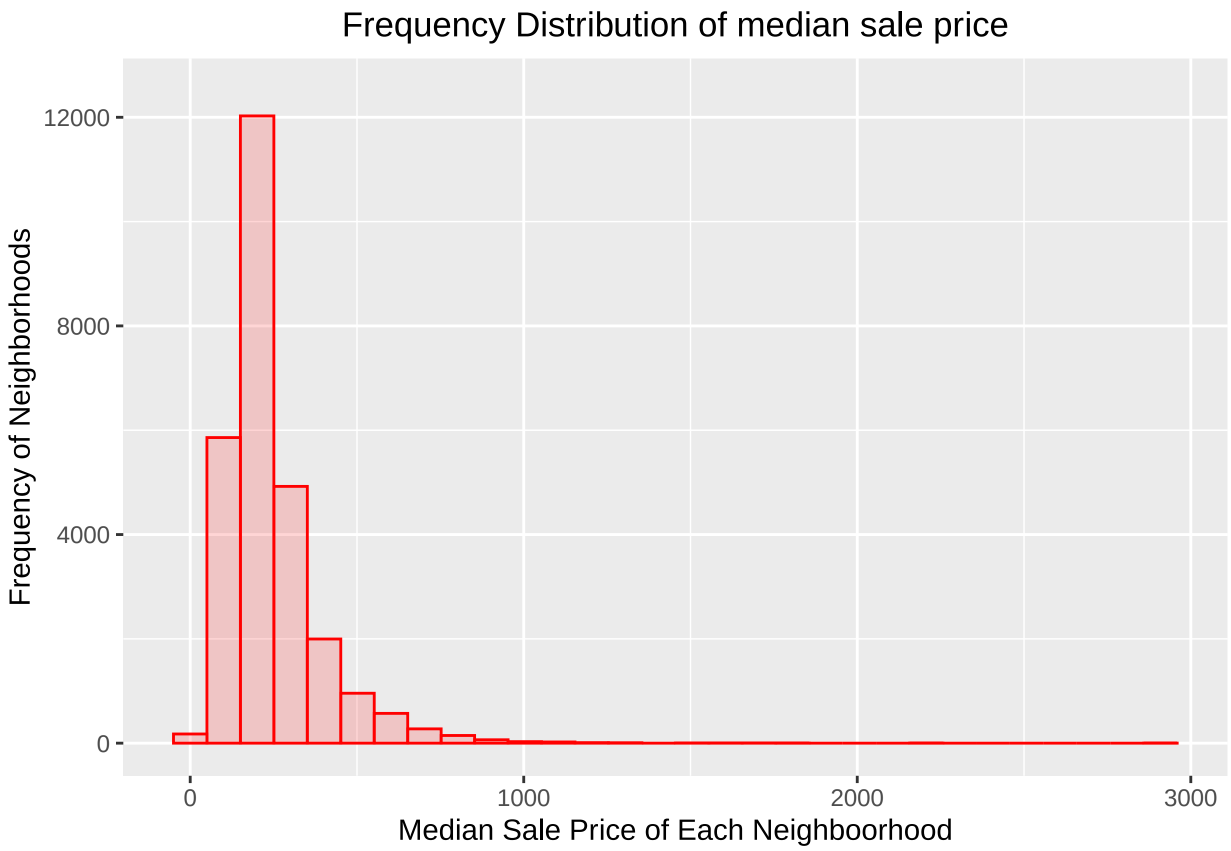
```
split_by_year <- split(state_market.df, format(state_market.df$period_begin, "%Y"))
ggplot(split_by_year[[length(split_by_year)]], aes(x = state_code, y = median_sale_price,
↪  color=median_sale_price)) +
  geom_boxplot(colour="#003366", fill="#66FFFF", alpha=5) +
  geom_point() +
  xlab("2021") +
  ylab("Median Sales Price (in thousands)") +
  ggtitle("Distribution of Median Sale Price per State") +
  theme(plot.title = element_text(hjust = 0.5))
```


Distribution of Median Sale Price per State
Median Sales Price (in thousands)
0
500
1000
1500
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2021
median_sale_price
1500
1000
500

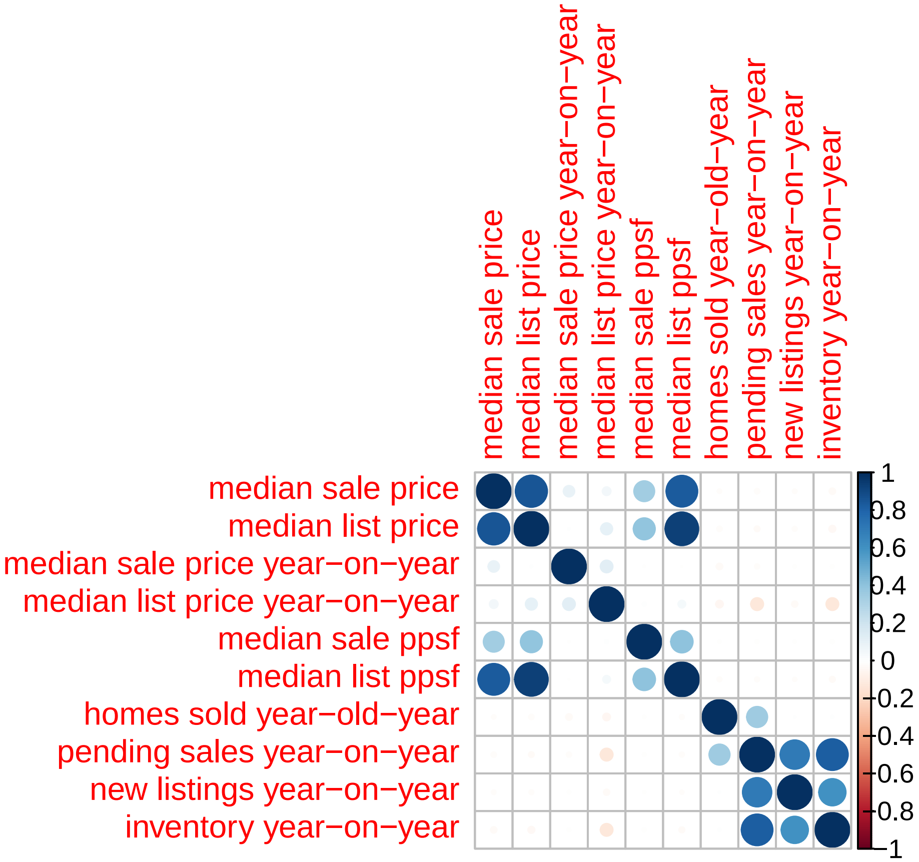
```
mean_years <- data.frame(state_market.df$period_begin,
↪  state_market.df$median_sale_price_yoy, state_market.df$state)

new_df <- subset.data.frame(state_market.df, select = c(state_code, median_list_price,
↪  median_sale_price),  drop = FALSE)

print(
    ggplot(new_df, aes_string(x=new_df$median_list_price))
    + geom_histogram(
      colour="darkorchid4", fill="darkorchid1", position="identity", bins=30, alpha=0.2
    )
    + ggtitle(paste("Frequency Distribution of median list price", sep=""))
    + theme(plot.title=element_text(hjust = 0.5))
    + xlab("Median List Price of Each Neighboorhood")
    + ylab("Frequency of Neighborhoods")
    + ylim(0, 12500))
```



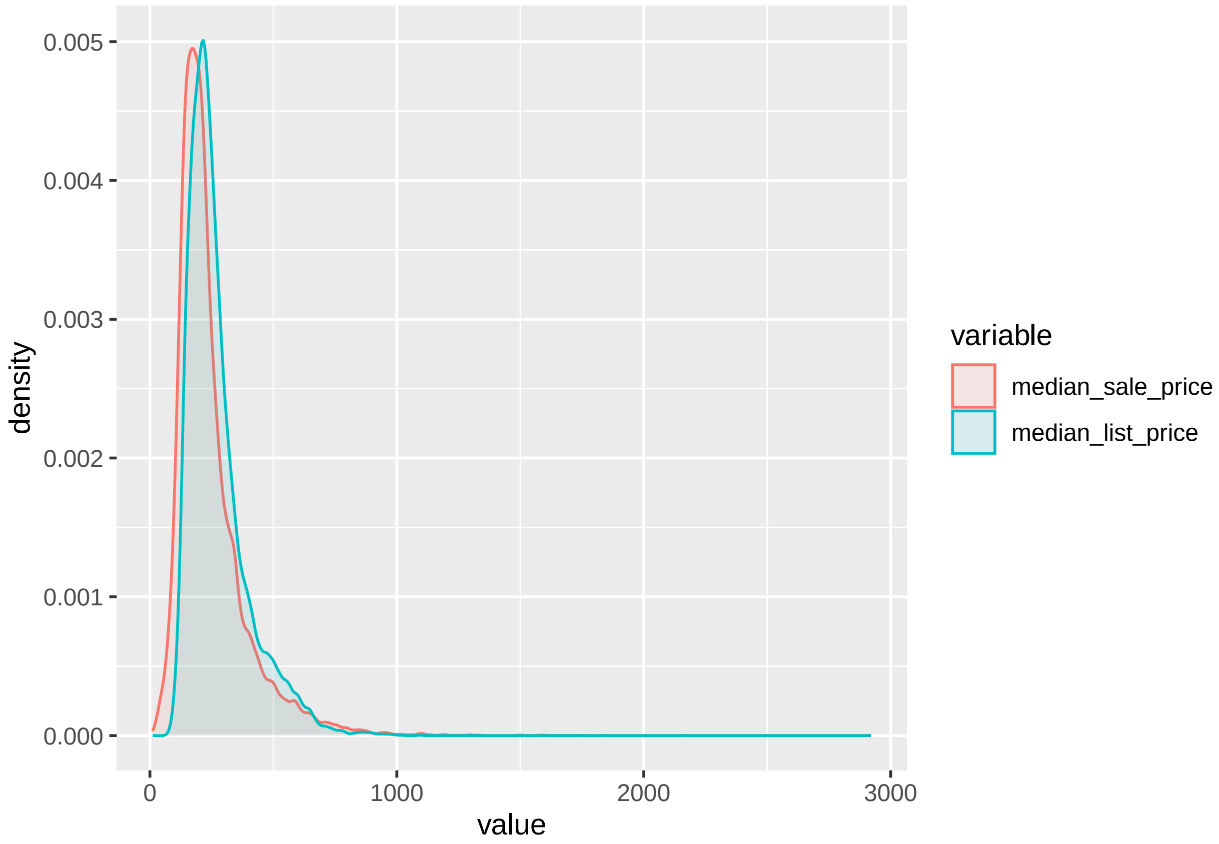
```
print(
    ggplot(new_df, aes_string(x=new_df$median_sale_price))
    + geom_histogram(
      colour="red", fill="firebrick1", position="identity", bins=30, alpha=0.2
    )
    + ggtitle(paste("Frequency Distribution of median sale price", sep=""))
    + theme(plot.title=element_text(hjust = 0.5))
    + xlab("Median Sale Price of Each Neighboorhood")
    + ylab("Frequency of Neighborhoods")
    + ylim(0, 12500))
```



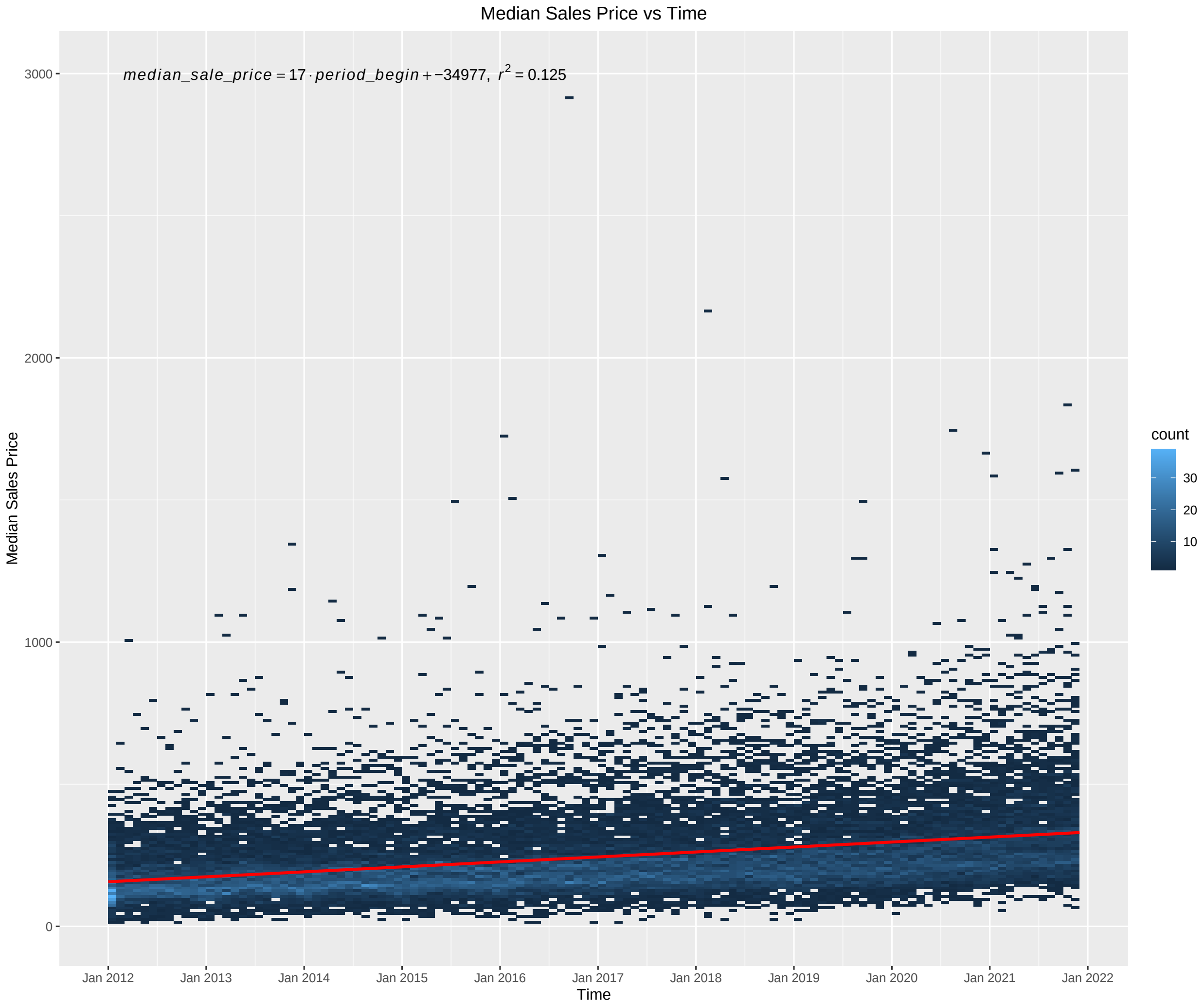
```
cor.df <- subset.data.frame(state_market.df, select = c(median_sale_price,
↪  median_list_price, median_sale_price_yoy, median_list_price_yoy, median_ppsf,
↪  median_list_ppsf, homes_sold_yoy, pending_sales_yoy, new_listings_yoy,
↪  inventory_yoy),  drop = FALSE)

cor.table <- cor(cor.df, use="pairwise.complete.obs")

rownames(cor.table) <- c("median sale price", "median list price", "median sale price
↪  year-on-year", "median list price year-on-year", "median sale ppsf", "median list
↪  ppsf", "homes sold year-old-year", "pending sales year-on-year", "new listings
↪  year-on-year", "inventory year-on-year")

colnames(cor.table) <- c("median sale price", "median list price", "median sale price
↪  year-on-year", "median list price year-on-year", "median sale ppsf", "median list
↪  ppsf", "homes sold year-old-year", "pending sales year-on-year", "new listings
↪  year-on-year", "inventory year-on-year")

matrix <- corrplot(cor.table)

corrplot(cor.table)
```



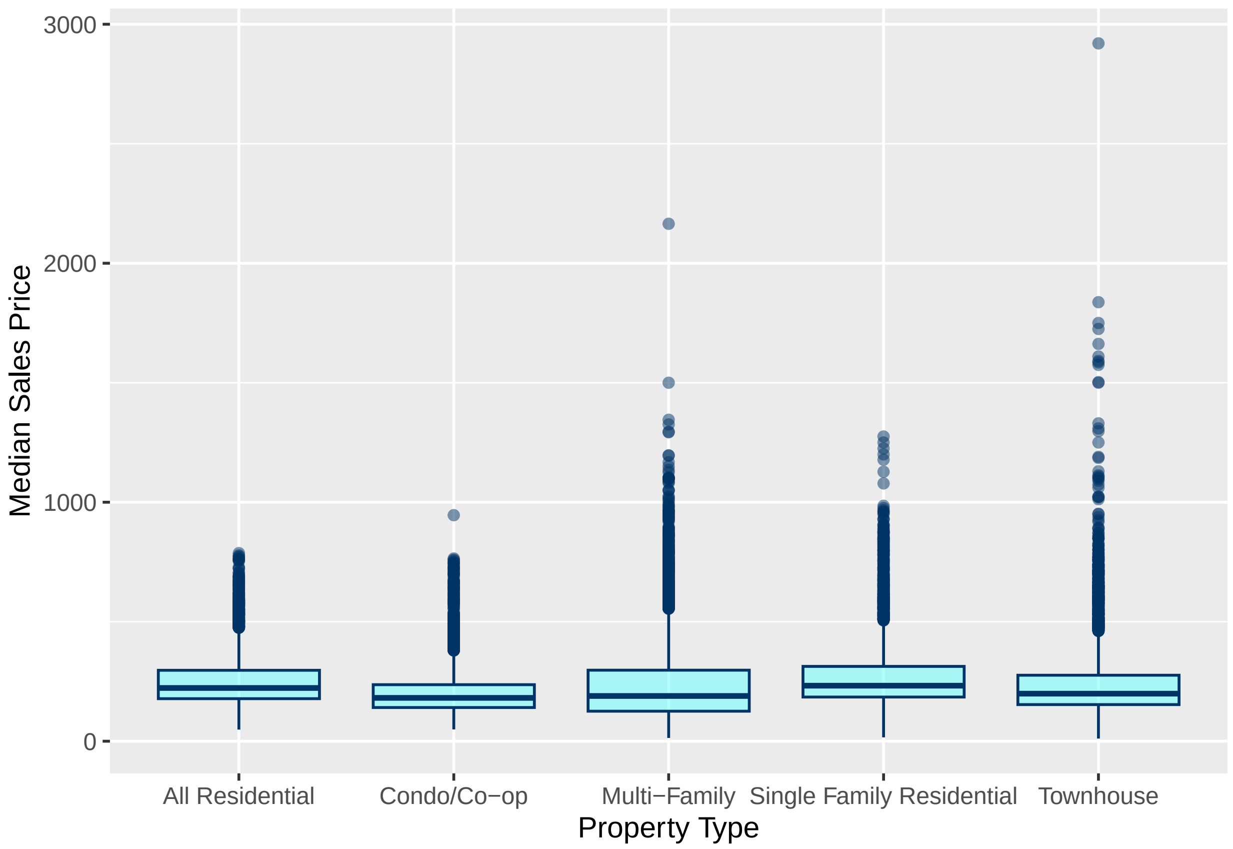
```
data_median_salemedian_list <- data.frame(state_market.df[, c("median_sale_price",
↪  "median_list_price")])

data_median_salemedian_list$refseq <- c("median_sale_price", "median_list_price")
s.plot <- melt(data_median_salemedian_list)

ggplot(s.plot, aes(x = value, colour = variable, fill = variable)) + geom_density(alpha =
↪  0.1)
```



```
lm_eqn <- function(df){
    m <- lm(median_sale_price ~ as.yearmon(period_begin), df);
    eq <- substitute(italic(median_sale_price) == b %.% italic(period_begin) +
↪   a*","~~italic(r)^2~"="~r2,
         list(a = format(unname(coef(m)[1]), digits = 2),
              b = format(unname(coef(m)[2]), digits = 2),
             r2 = format(summary(m)$r.squared, digits = 3)))
    as.character(as.expression(eq));
}

ggplot(state_market.df, aes(x = as.yearmon(period_begin), y = median_sale_price)) +
  geom_bin_2d(binwidth = c(1/12, 10)) +
  xlab("Time") +
  ylab("Median Sales Price") +
  ggtitle("Median Sales Price vs Time") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_yearmon(n = 10) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  annotate("text", x = as.numeric(as.yearmon("2014-06-01")), y = 3000, parse = TRUE,
  ↪  label = lm_eqn(state_market.df))
```

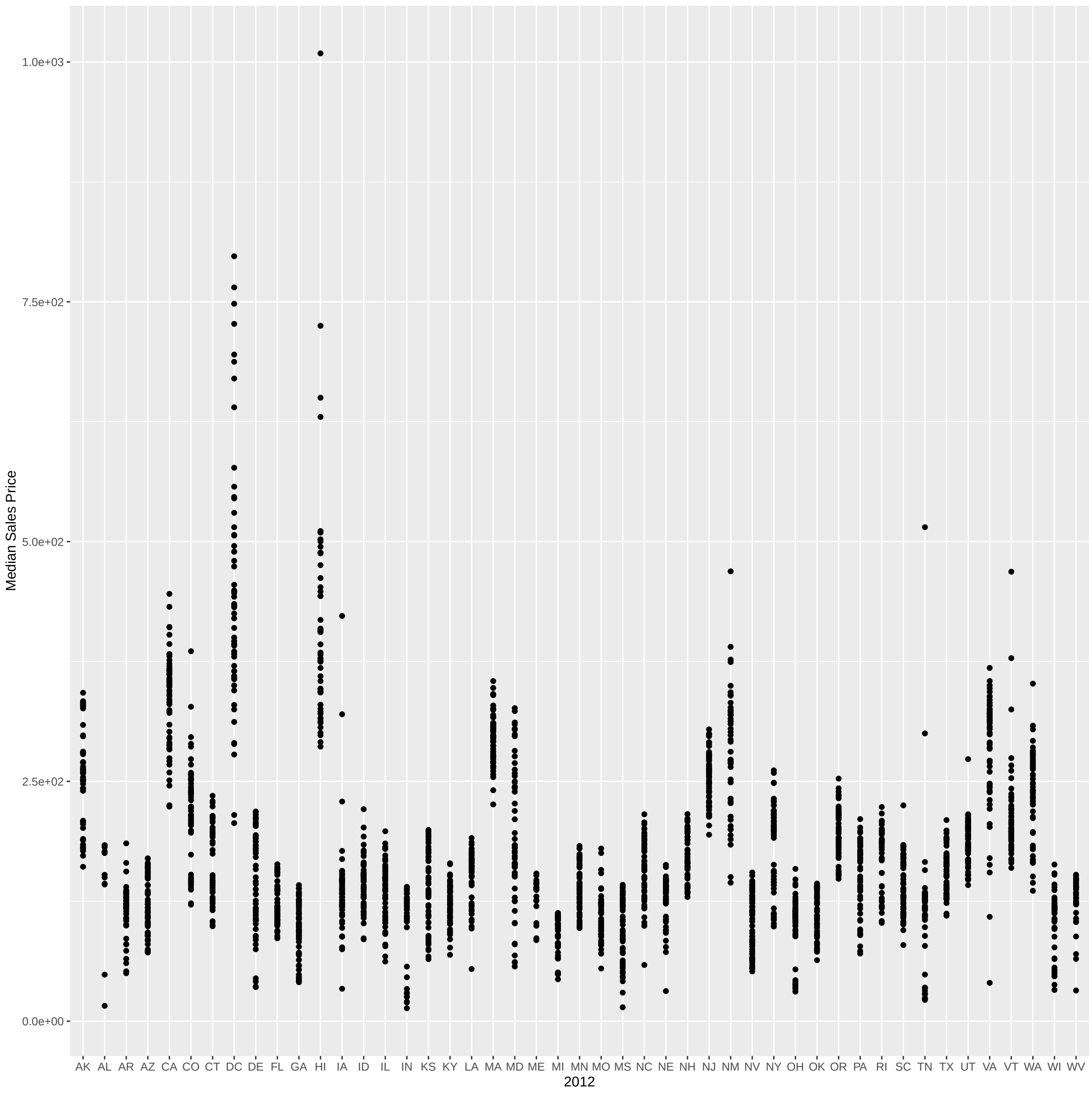


```
ggplot(state_market.df, aes(x = property_type, y = median_sale_price)) +
  geom_boxplot(colour="#003366", fill="#66FFFF", alpha=1/2) +
  xlab("Property Type") +
  ylab("Median Sales Price")
```

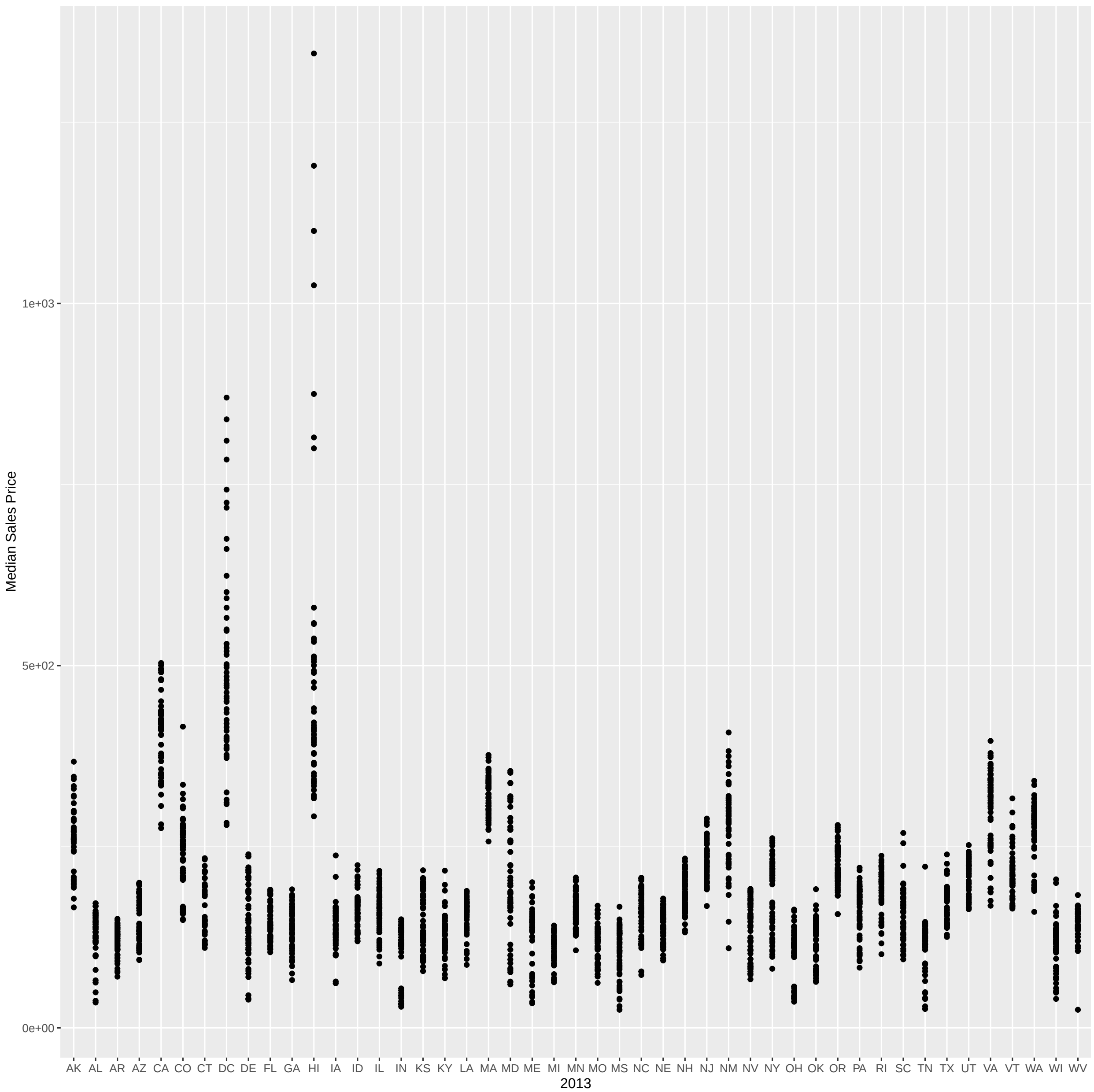


```
state_market.df$period_begin <- as.Date(state_market.df$period_begin)
split_by_year <- split(state_market.df, format(state_market.df$period_begin, "%Y"))

lapply(split_by_year, function(i) ggplot(i, aes(x = state_code, y = median_sale_price))
                                  + geom_point()
                                  + xlab(format(i$period_begin, "%Y"))
                                  + ylab("Median Sales Price")
                                  + scale_y_continuous(labels = function(x) format(x,
                                   ↪  scientific = TRUE)))
```

```
## $`2012`
```

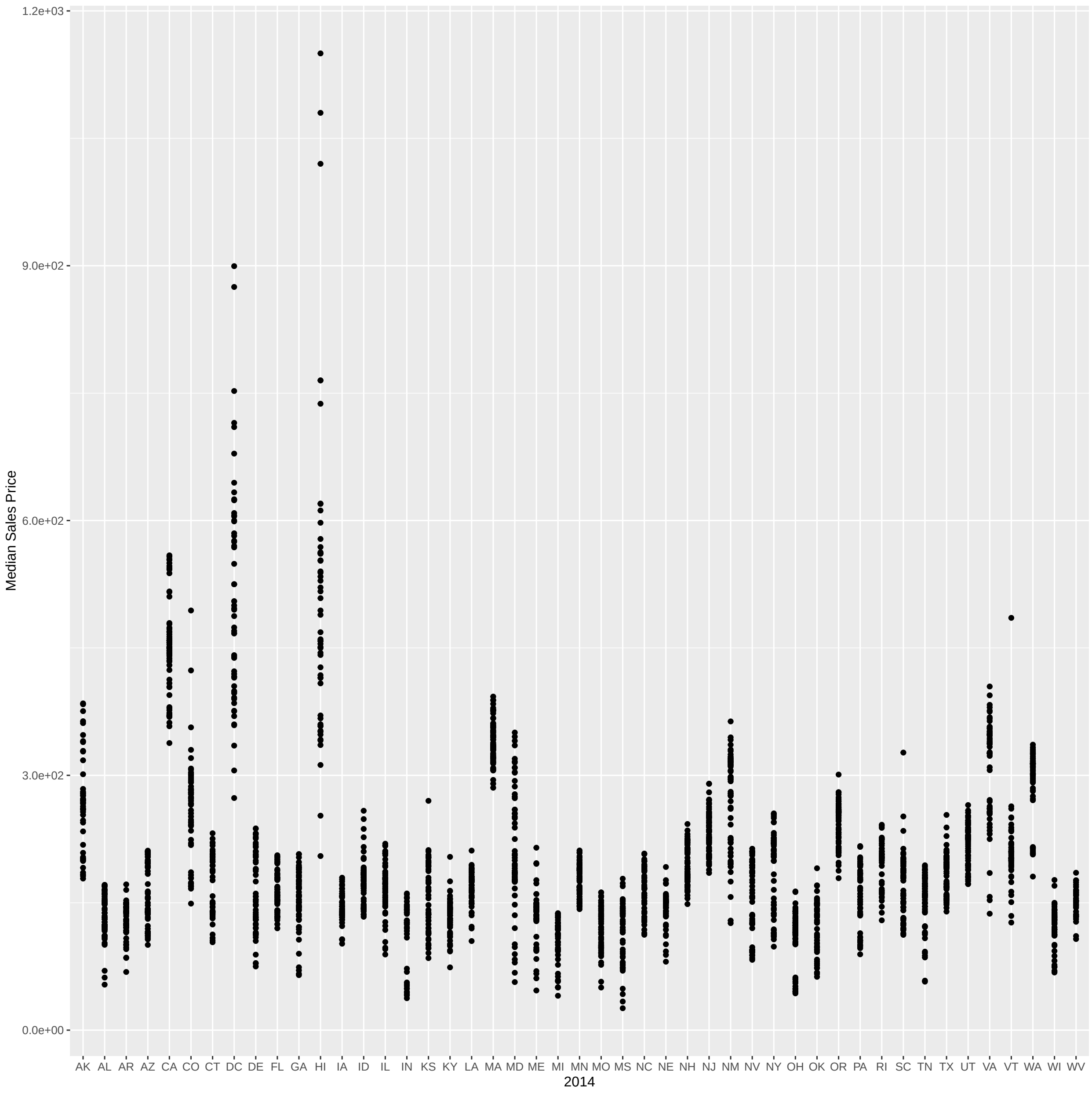

Median Sales Price
1.0e+03
7.5e+02
5.0e+02
2.5e+02
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2012

```
##
## $`2013`
```

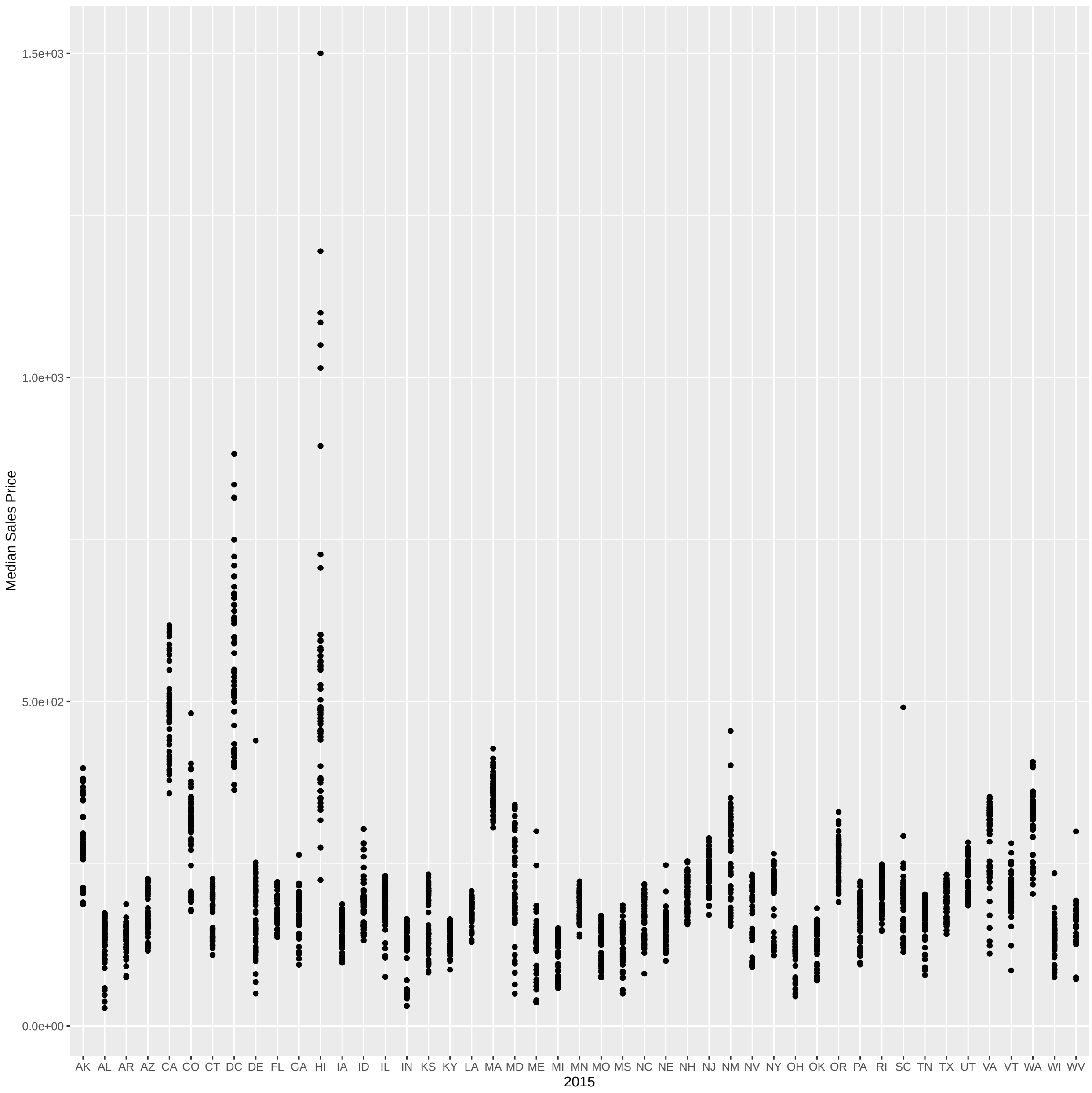


Median Sales Price
1e+03
5e+02
0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2013

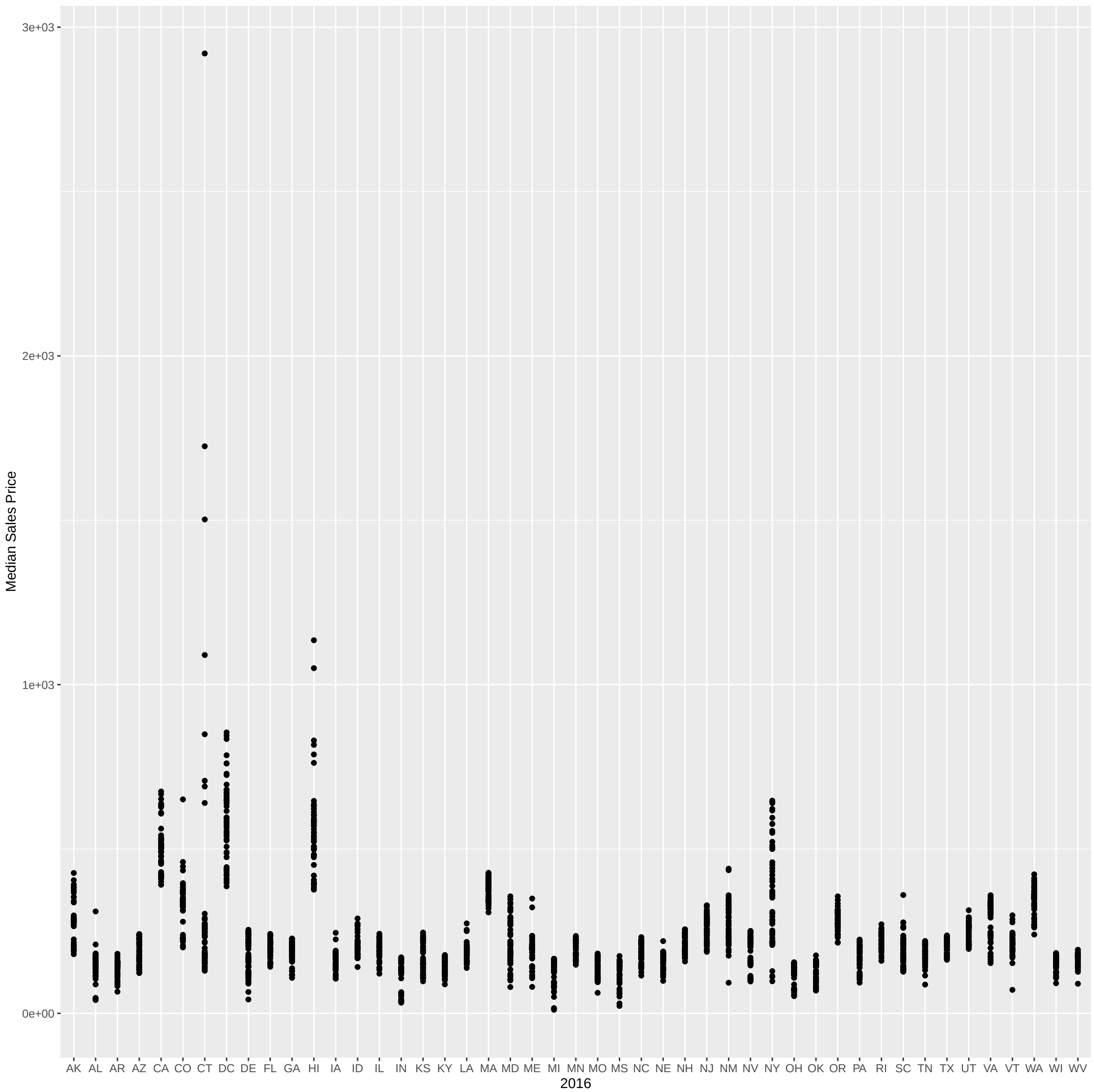
```
##
## $`2014`
```


1.2e+03
9.0e+02
6.0e+02
3.0e+02
0.0e+00
Median Sales Price
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2014

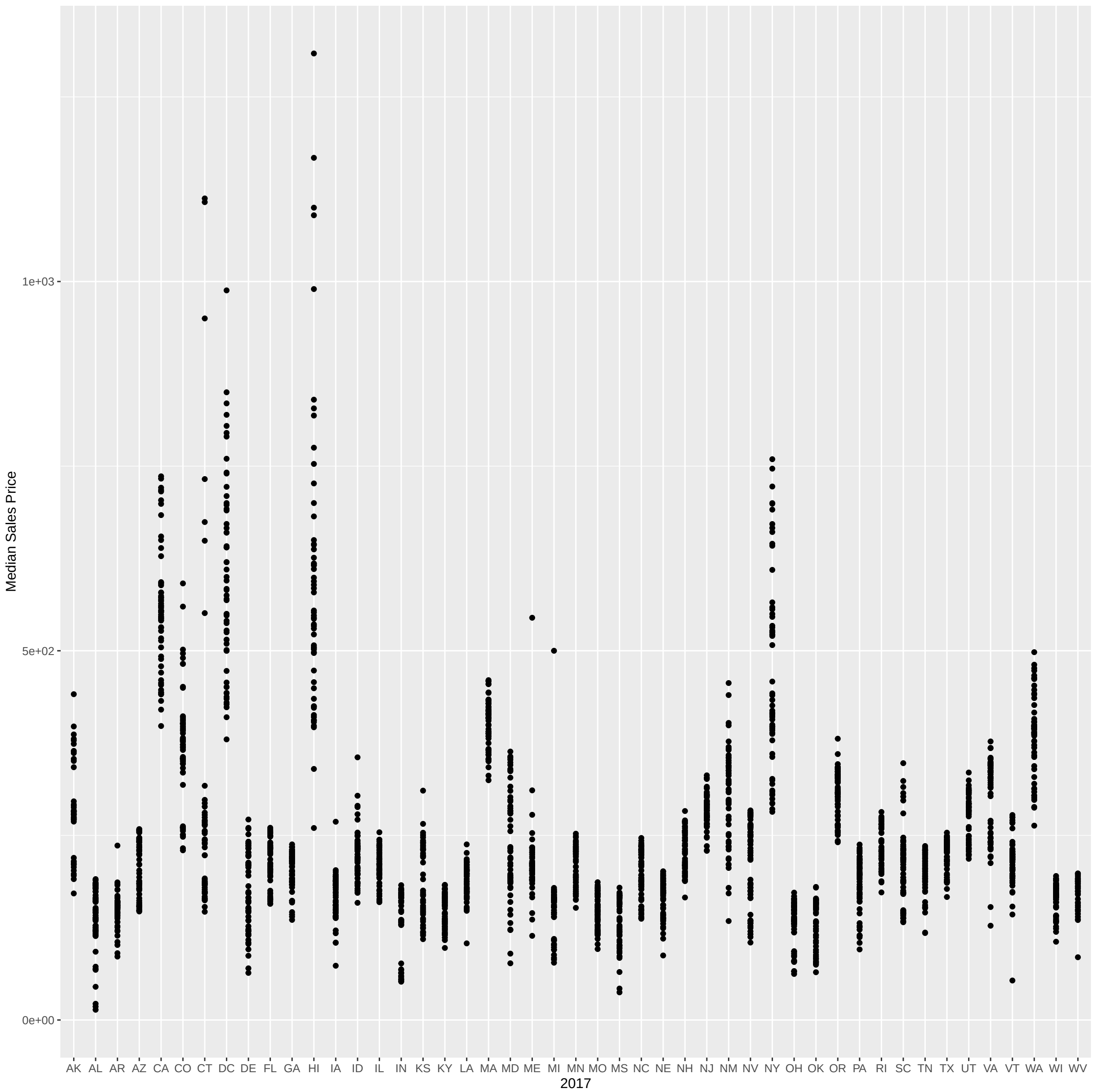
```
##
## $`2015`
```


Median Sales Price
1.5e+03
1.0e+03
5.0e+02
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2015

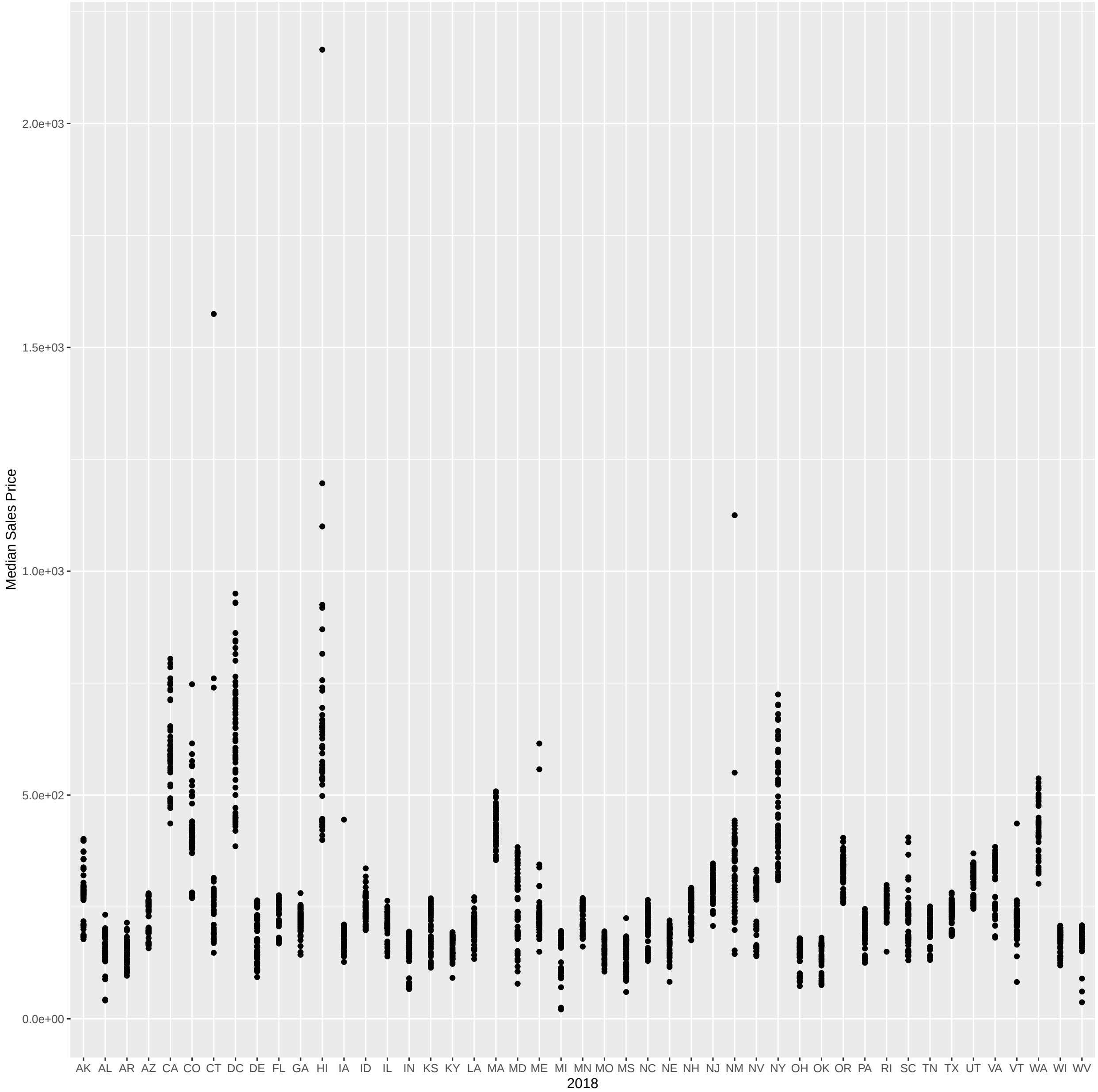
```
##
## $`2016`
```


Median Sales Price
3e+03
2e+03
1e+03
0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2016

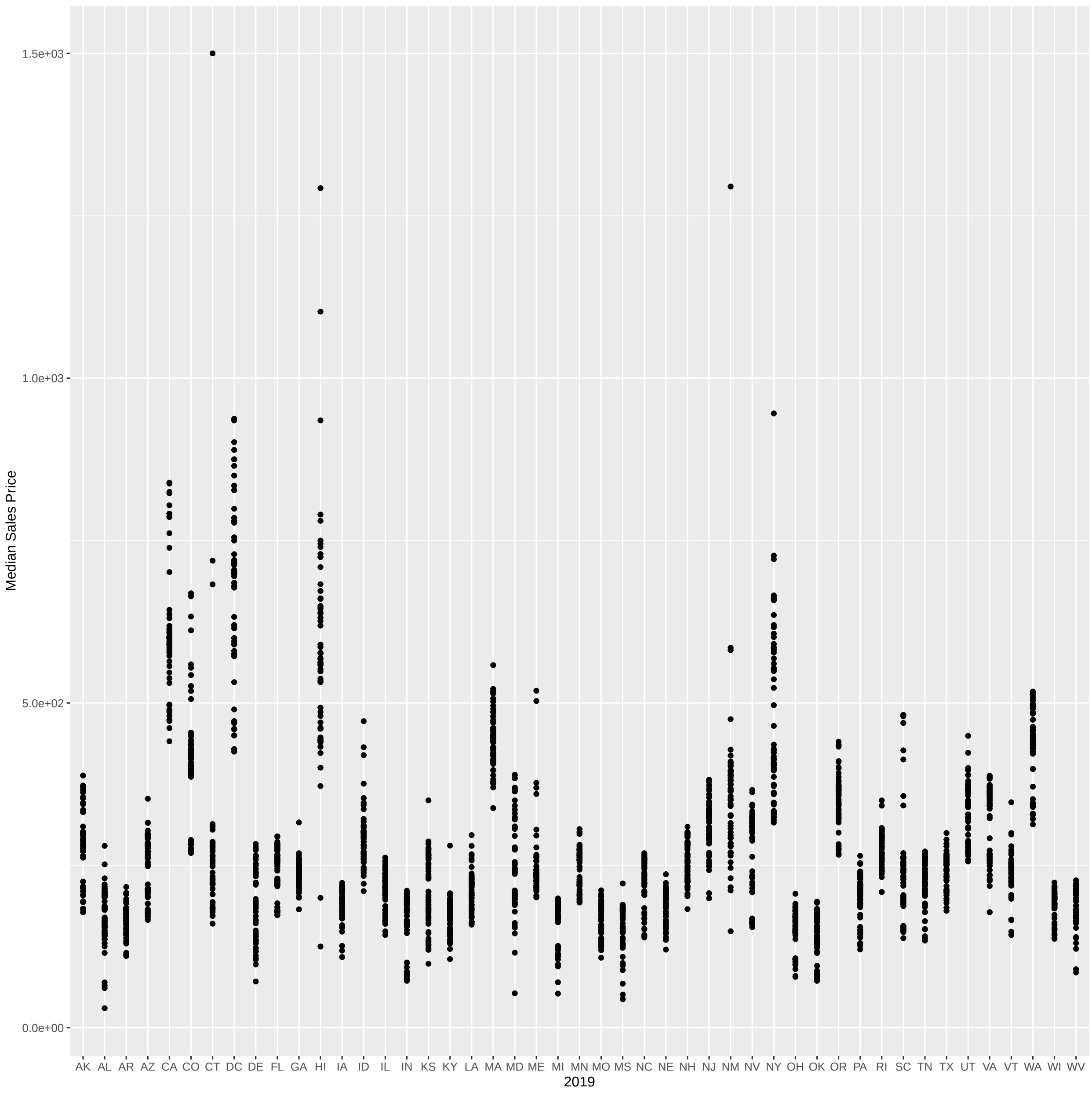
```
##
## $`2017`
```


Median Sales Price
1e+03
5e+02
0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2017

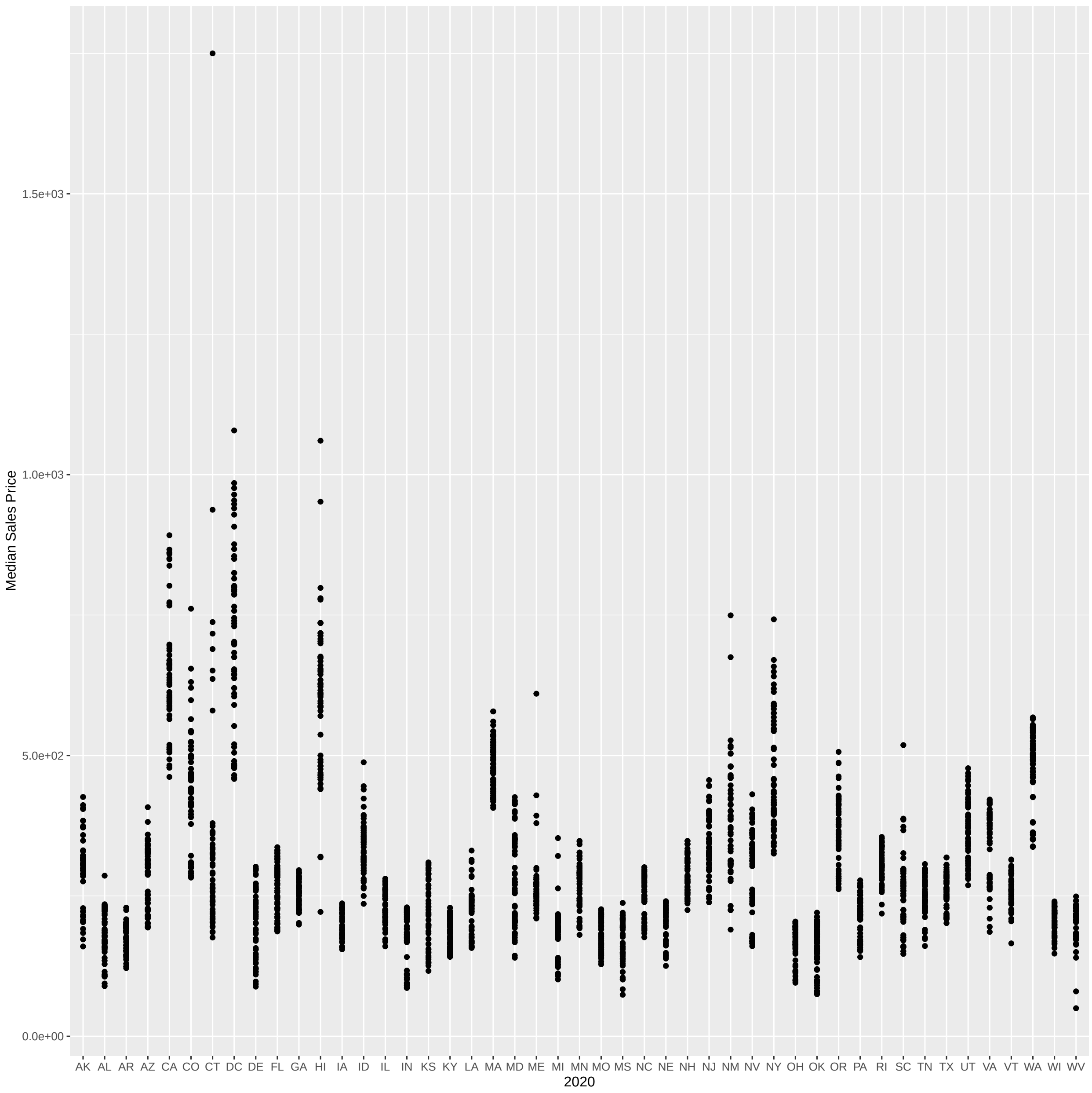
```
##
## $`2018`
```


Median Sales Price
2.0e+03
1.5e+03
1.0e+03
5.0e+02
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2018

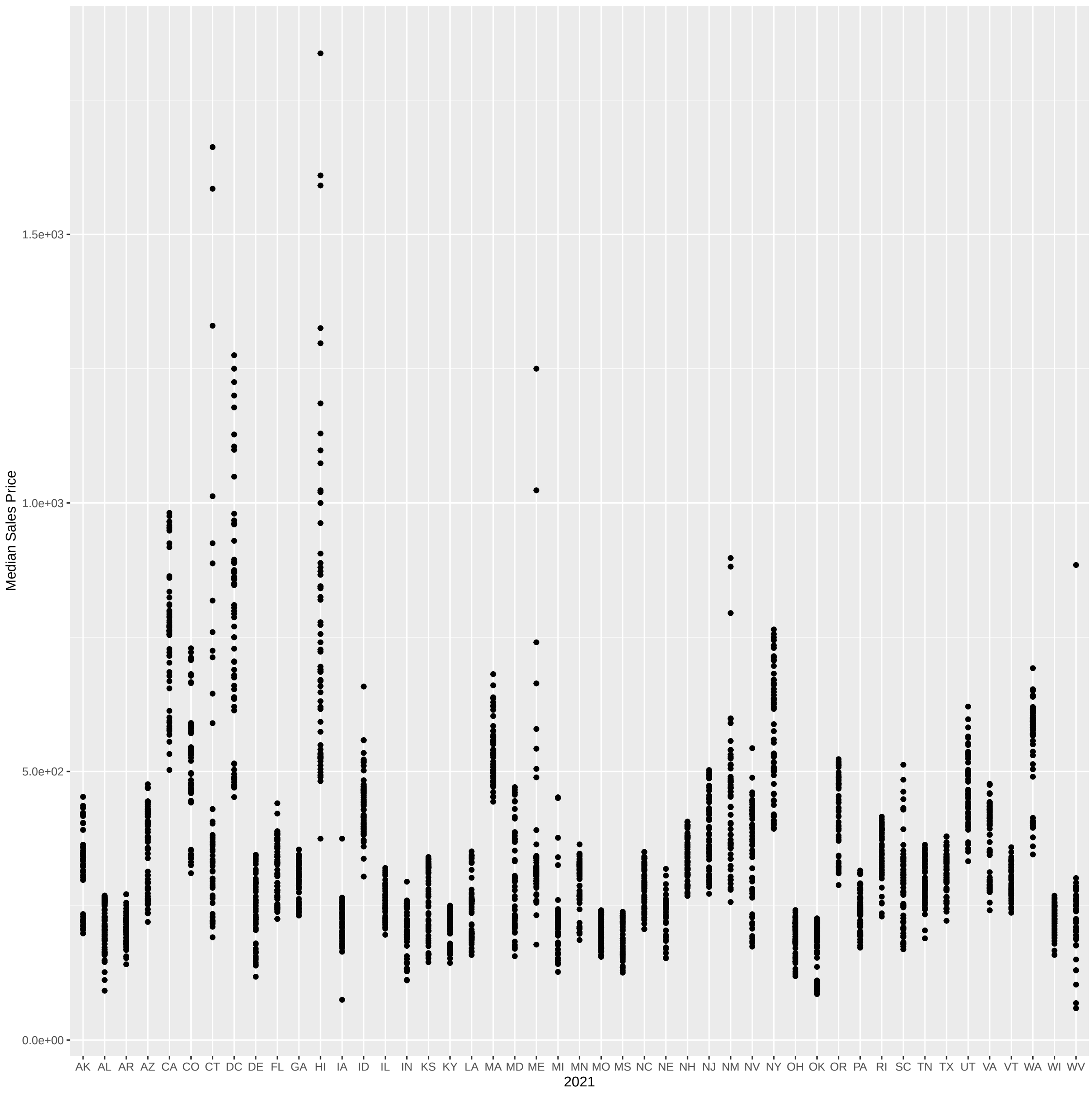
```
##
## $`2019`
```


1.5e+03
1.0e+03
5.0e+02
0.0e+00
Median Sales Price
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2019

```
##
## $`2020`
```


Median Sales Price
1.5e+03
1.0e+03
5.0e+02
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2020

```
##
## $`2021`
```


Median Sales Price
1.5e+03
1.0e+03
5.0e+02
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2021