EXPLORATORY DATA ANALYSIS 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.

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, fig.align =
↪  'center')
```

Import the necessary libraries

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

Import the Data

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

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

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. There is no missing values in the dataset, so we do not have to do anything with that.

```
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"
```

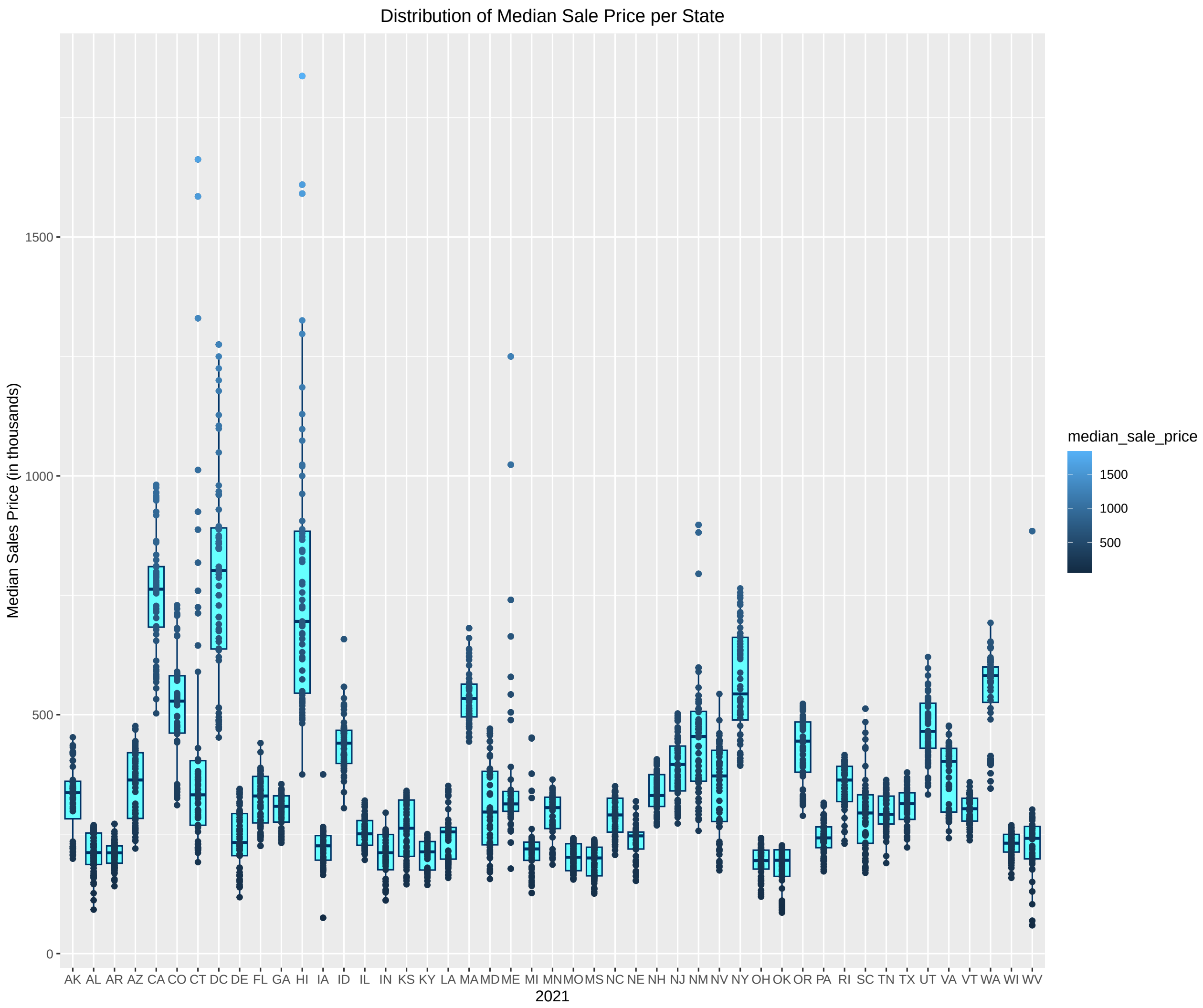
Out of the 48 variables in the dataset, these are the ones we will consider (the important variables) in our subsequent exploratory analysis:

- median sale price - **Median sale price the housing units of each of the 27,054 neighborhoods**
- median list price - **Median list price the housing units of each of the 27,054 neighborhoods**
- median sale price year-on-year - **Year-to-year percentage increase in median sale price of each neighborhood**
- median list price year-on-year - **Year-to-year percentage increase in median list price of each neighborbood**
- median sale ppsf - **Median sale price per-square-foot of each neighborhood**
- median list ppsf - **Median sale price per-square-foot of each neighborhood**
- homes sold year-old-year - **Increase in number of homes sold in each neighborhood year-on-year**
- pending sales year-on-year - **Increase in number of pending sales in each neighborhood year-on-year**
- new listings year-on-year - **Increase in number of new listings in each neighborhood year-on-year**
- inventory year-on-year - **Increase in number of all listings in each neighborhood year-on-year**
- sold_above_list_yoy – **Increase in number of homes sold above list price in each neighborhood**
- avg_sale_to_list_yoy – **Change in the ratio of homes sold to homes listed in each neighborhood**
- region – **Categorical variable that classifies which region the neighborhood is in**
- state_code – **Categorical variable that is abbreviated state name**
- property_type – **Type of home**

Interpreting the Data

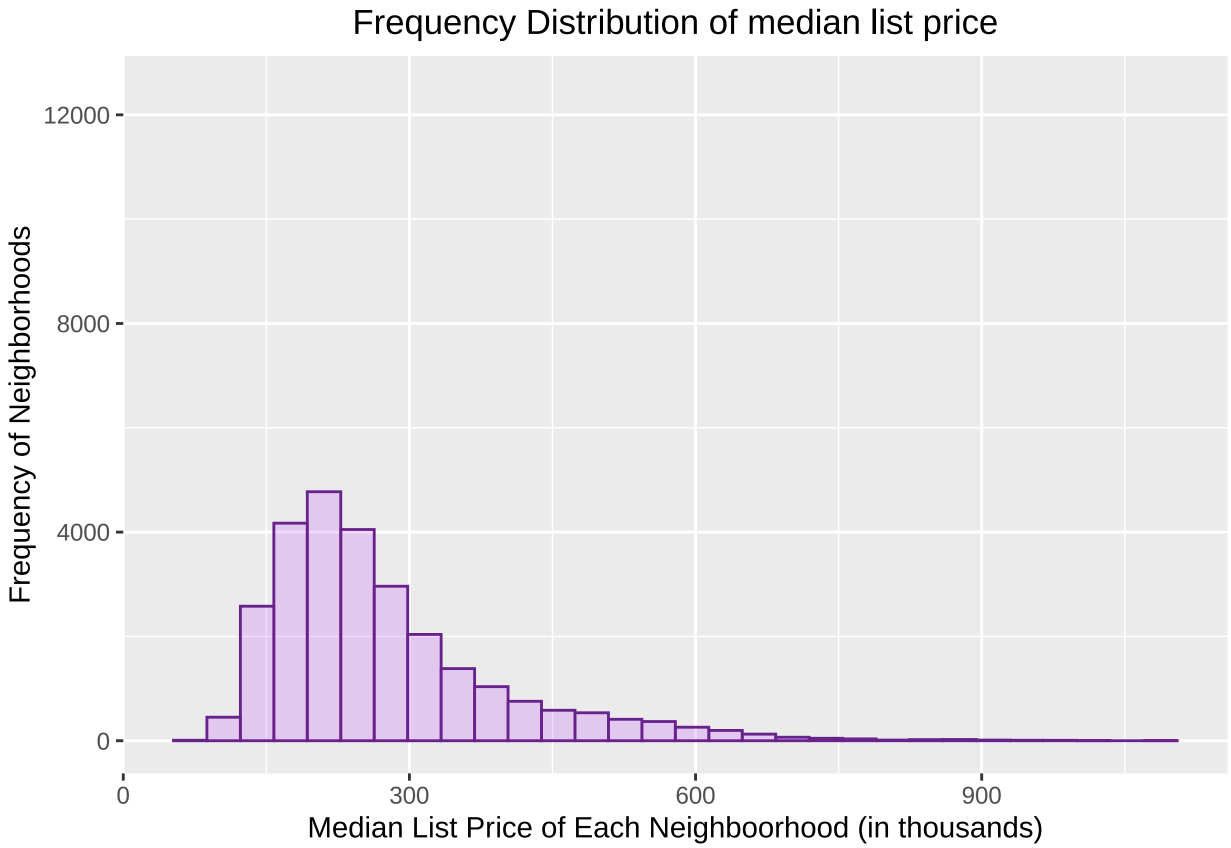
Here, we created a boxplot graph for the median list price of the homes in each neighborhood in the dataset in the year 2021, with each boxplot representing each state. Hawaii and and Connecticut were had the highest number of outliers. Since there were more outliers above the boxplot than below it in 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. We can see that states with much lower overall median sale price, such as Oklahoma and Ohio, also have much lower variance. There is sarious skewness and interquartile ranges between each state. Some are skewed left, and some are skewed right. Majority have small interquartile ranges. A few have large interquartile ranges. Initially, median_sale_price is our target variable, since we want to learn how much money we can make from selling houses!

```
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))
```

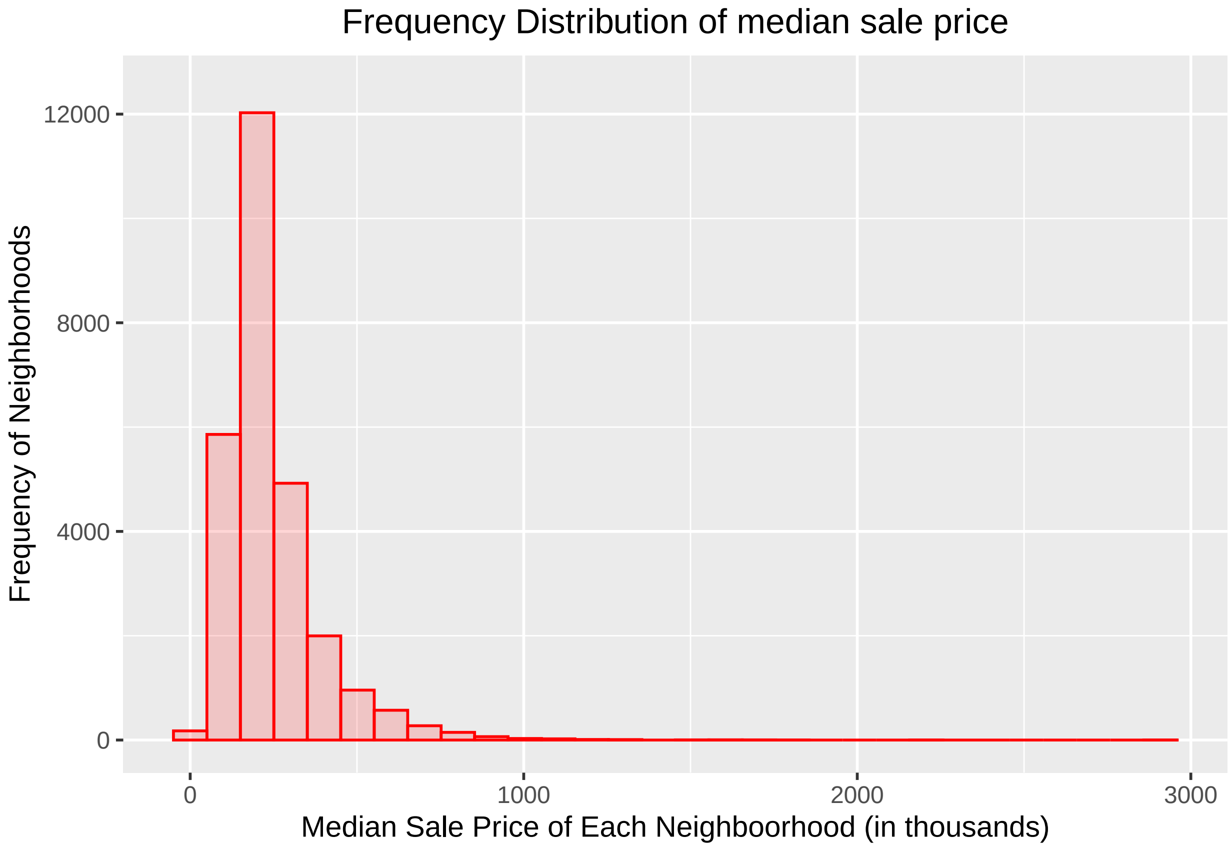

Distribution of Median Sale Price per State
Median Sales Price (in thousands)
0
500
1000
1500
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS NC NE NH NJ NM NV NY OH OK OR PA RI SC TN TX UT VA VT WA WI WV
2021
median_sale_price
1500
1000
500

We created 2 histograms. One is for the median list price of housing units in each of the 27,054 neighborhoods from 2012 to 2021. The other is the median sale price of over the same length of time. We noticed that both graphs are skewed to the right, so we can assume that there are more homes sold and listed above the overall median home price than below it in the U.S. We also noticed that there is much less variance in the sale price than in the list price, with the right skew in the list price being much less pronounced in the list price than the sale price. Therefore, we can assume that if the asking/list price is higher than the median, the seller receives less in the final sale price in most cases.

```r
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 (in thousands)")
    + ylab("Frequency of Neighborhoods")
    + ylim(0, 12500))
```

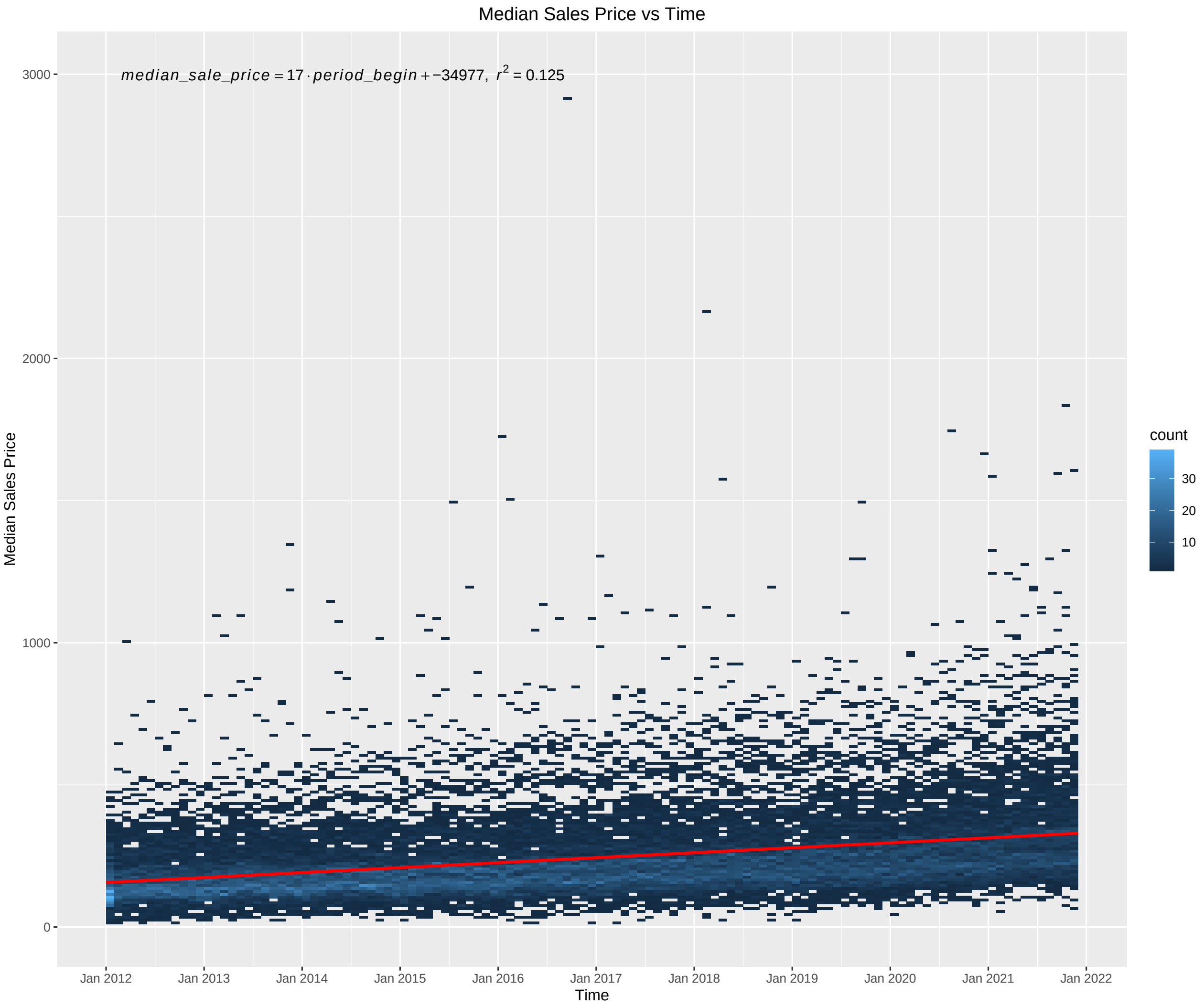


```
print(
    ggplot(new_df, aes_string(x=new_df$median_sale_price))
    + geom_histogram(
      colour="red", fill="firebrick1", position="identity", bins=30, alpha=0.2
    )
    + ggtitle(paste("Frequency Distribution of median sale price", sep=""))
    + theme(plot.title=element_text(hjust = 0.5))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



We can create a bin plot to demonstrate that home sale prices tend to aggregate below a million, between 100 thousand to 300 thousand. We have fitted a best fit line to show that there is a positive increase in median sales price over time, however, with a r-squared value of 0.125, only 12.5% of this increase can be explained by time. This is understandable, since the value of a home includes many factors such as: location, land, size, time built, etc.

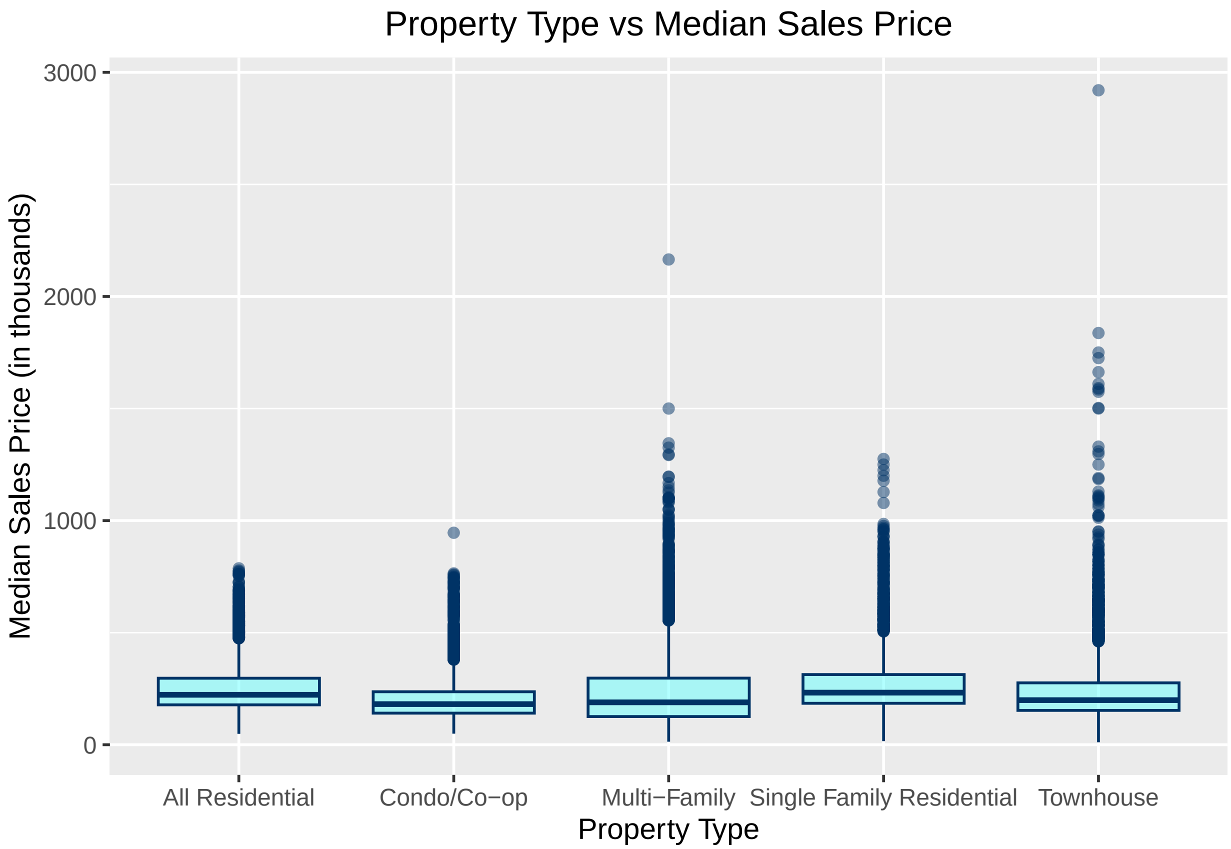
```r
lm_eqn <- function(df){
    m <- lm(median_sale_price ~ as.yearmon(period_begin), df);
    eq <- substitute(italic(median_sale_price) == b %.% italic(period_begin) +
↪   a*","~~italic(r)^2~"="~r2,
         list(a = format(unname(coef(m)[1]), digits = 2),
              b = format(unname(coef(m)[2]),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


Median Sales Price vs Time
$median\_sale\_price = 17 \cdot period\_begin + -34977,\ r^2 = 0.125$
Median Sales Price
3000
2000
1000
0
Time
Jan 2012
Jan 2013
Jan 2014
Jan 2015
Jan 2016
Jan 2017
Jan 2018
Jan 2019
Jan 2020
Jan 2021
Jan 2022
count
30
20
10

Below, we have created 2 boxplots for the types of housing units. One for the median sale price, the other for the median list price. For our outlier tests, we will test if anything is above the 99th quantile. We chose to be less sensitive in our outlier detection, since house prices vary wildly in the market.

```
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 (in thousands)") +
  ggtitle("Property Type vs Median Sales Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

Both townhouse and multi-family have extreme outliers in the median sales price.



```
## We can see that for the outliers for multi-family housing the state Hawaii has the
## most outliers. This gives us a good explanation in the reason for the outliers, since
## Hawaii is a small state in the middle of the Pacific serverely limiting the supply of land.
```

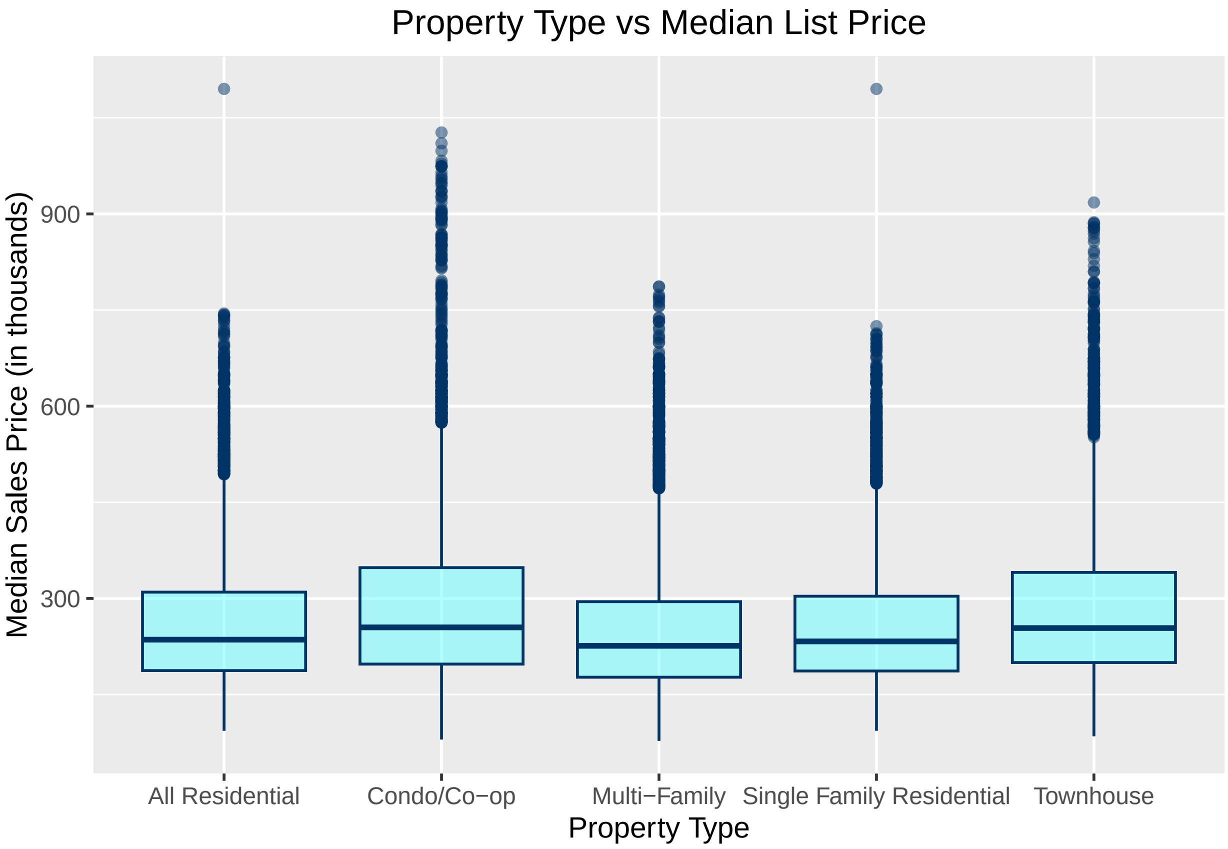
```
##
## California    Columbia      Hawaii New Mexico
##         11          11          30          3
```

```
## The same thing is repeated with townhouses, but Connecticut has almost as many as Hawaii
## now. There is a possibility that the reasoning behind this is the same as Hawaii with the
## limited supply of land, but it also could be that mostly everywhere in Connecticut is close
## to a town or city that sports many amenities.
```

```
##
##    Columbia Connecticut      Hawaii       Maine
##           9          19          20           2
```

```
ggplot(state_market.df, aes(x = property_type, y = median_list_price)) +
  geom_boxplot(colour="#003366", fill="#66FFFF", alpha=1/2) +
  xlab("Property Type") +
  ylab("Median Sales Price (in thousands)") +
  ggtitle("Property Type vs Median List Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

Below, we can see that the extreme outliers are not as pronounced with the median list price.



```
## California leads the US in most outliers for all residential house prices while Columbia
## and Hawaii are not far behind. Hawaii has the same reasoning for the listing price as what
## was explained above. California's outliers can be explained by the housing markets
## predictions that a huge demand for housing will always be present in the state. Columbia has
## a low supply compared to the high demand in the housing market.
```

```
##
## California   Columbia     Hawaii      Maine
##         22         19         15          1
```

```
## New York leads the US in most outliers for condo/co-op house prices. New York will
## obviously lead in most outliers, since most people only live in condo's or cooperative
## housing in the state and the housing markets prediction of ever increasing demands of housing
## in places like New York City.
```

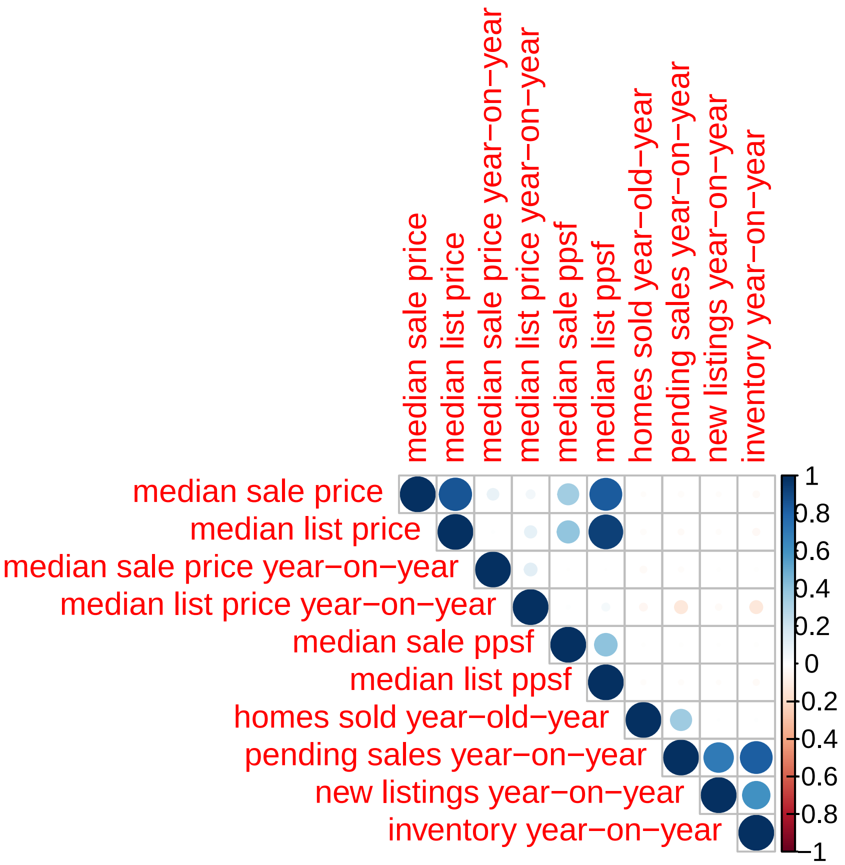
```
##
## California   New York       Utah
##          4         49          2
```

```
## California and Columbia leads the US in most outliers for single residential house prices.
## California is the same as described above. Columbia has a low supply compared to the high
## demand in the housing market.
```

```
##
## California   Columbia     Hawaii      Maine
##         19         23         14          1
```

We created a correlation matrix with our important variables. We expected a strong positive correlation between median sale price and median list price, and an even stronger correlation between median sale price and median price-per-square-foot, but a feature that surprised us was that there was a modest negative correlation between median list price year-on-year and the pending sales year-on-year. We believe we can attribute this to people being less likely to buy a house if the house is more expensive.

```r
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")

corrplot(cor.table, type="upper")
```

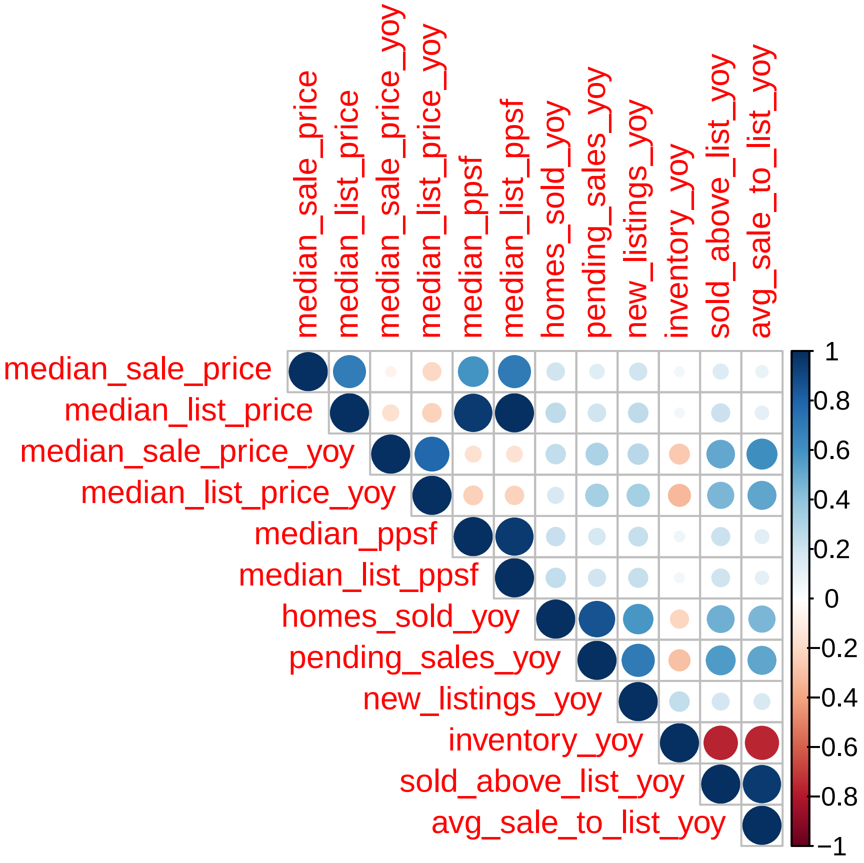


Change of Plans!

Median sale price was our target variable initially, but we could not find any useful relationships between this variable and the other variables. We found a linear equation to fit the median list price over time, but the R^2 value was too low to for this equation to adequately fit the data. This can attributed to a large variation in sale prices between homes. Because of the previously unidentified factors in our dataset, we can't always assume that all or most neighborhoods will increase in price at the same rate. We'll now take a closer look at the percentage increase in the number of pending sales per neighborhood. The dataset is too large, so let's pick just one State - California.

Our correlation matrix now gives us much more promising results.

```
cor.df <- subset(state_market.df, state_code == 'CA')

cor.df <- subset.data.frame(cor.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,sold_above_list_yoy,avg_sale_to_list_yoy))

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

corrplot(cor.table, type="upper")
```



Training our Data

We have decided to go with a linear model with the predictors being inventory_yoy and new_listings_yoy and criterion being pending_sales_yoy. These were select due to there strong correlation they exhibit on each other in the correlation matrix above. From our summary below, we see that both of these variables are statistically significant, since there p-values from the t-test [Pr(>|t|)] is close to zero. We also get an adjusted R^2 of 0.7312 meaning our linear model is of good fit to the data. Interestingly, the inventory_yoy has a negative coefficient, while the new_listings_yoy has a positive coefficient.

```r
model1 <- lm(pending_sales_yoy~inventory_yoy+new_listings_yoy, cor.df)
summary(model1)
```

```
##
## Call:
## lm(formula = pending_sales_yoy ~ inventory_yoy + new_listings_yoy,
##     data = cor.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37185 -0.05858 -0.00910  0.04636  0.52399
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.024542   0.004106   5.977 3.91e-09 ***
## inventory_yoy    -0.358371   0.015643 -22.909  < 2e-16 ***
## new_listings_yoy  0.978292   0.025793  37.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09838 on 597 degrees of freedom
## Multiple R-squared:  0.7321, Adjusted R-squared:  0.7312
## F-statistic: 815.7 on 2 and 597 DF,  p-value: < 2.2e-16
```

We can see, from these two models, that choosing just a single predictor yields a substantially lower adjusted R^2.

```
model2 <- lm(pending_sales_yoy~inventory_yoy, cor.df)
summary(model2)
```

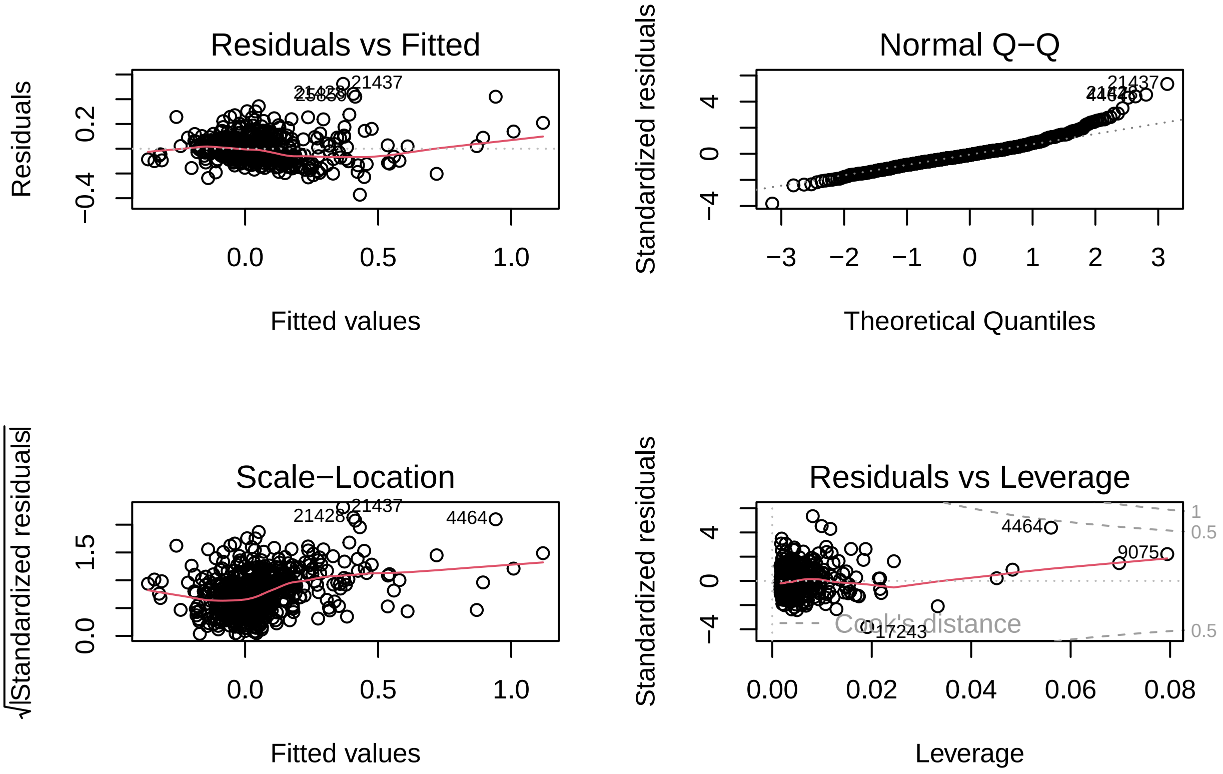
```
##
## Call:
## lm(formula = pending_sales_yoy ~ inventory_yoy, data = cor.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52508 -0.08354 -0.03502  0.03565  1.28765
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.038443   0.007545   5.095 4.68e-07 ***
## inventory_yoy -0.210383   0.027950  -7.527 1.92e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1815 on 598 degrees of freedom
## Multiple R-squared:  0.08655,    Adjusted R-squared:  0.08502
## F-statistic: 56.66 on 1 and 598 DF,  p-value: 1.918e-13
```

```
model3 <- lm(pending_sales_yoy~new_listings_yoy, cor.df)
summary(model3)
```

```
##
## Call:
## lm(formula = pending_sales_yoy ~ new_listings_yoy, data = cor.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41705 -0.08852 -0.02654  0.05626  0.63443
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.043719   0.005506   7.941 9.99e-15 ***
## new_listings_yoy 0.830915   0.034211  24.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1348 on 598 degrees of freedom
## Multiple R-squared:  0.4966, Adjusted R-squared:  0.4958
## F-statistic: 589.9 on 1 and 598 DF,  p-value: < 2.2e-16
```

We can see that homoscedasticity seems to hold, because the points don't seem to deviant more from the line the more farther you go out. However, the data is obviously skewed. More data points lie on one side more than the other. The QQ-plot seems to have a heavy tail suggesting that our distribution fails to yield to a linear model the closer the data point gets to being an outlier. For our cook's distance, no data point is extreme enough to heavily effect our linear model.

```
par(mfrow=c(2,2))
plot(model1)
```



Conclusion

We tried building a model around our entire (HUGE!) dataset and could not make any reliable observations and/or build models around these observations. We noticed that time is not an accurate predictor of value of houses, since we can't make that assumption that the market will continue to grow steadily over time. There is a strong correlation between median list price and median sale price, but median list price is not an accurate predictor since the sale price data contains much more outliers, and the median list price is not normally distributed like how the median sale price is. Building models was easier after we shrunk our dataset to a single state, since real estate markets differ from state to state. We changed our target variable to the yearly increase in pending sales. We can accurately predict the yearly increase in the number of pending sales of each neighborhood.

Sample Data

```r
sample.df <- subset.data.frame(state_market.df, select = c(period_begin, state_code,
↪  property_type, 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)

pander(head(sample.df))
```

Table 1: Table continues below

period_begin	state_code	property_type	median_sale_price
2019-10-01	OK	Multi-Family	162.2
2021-07-01	VT	All Residential	317.9
2016-08-01	NH	Condo/Co-op	200.1
2013-04-01	MS	All Residential	129.5
2019-12-01	MO	Condo/Co-op	152
2019-07-01	NM	All Residential	385.5

Table 2: Table continues below

median_list_price	median_sale_price_yoy	median_list_price_yoy
185	0.06697	0.005355
322.7	0.2048	0.07324
260.2	0.08944	0.09914
144.2	0.07135	0.05368
170.1	0.068	0.05685
390.8	0.08515	-0.05731

Table 3: Table continues below

median_ppsf	median_list_ppsf	homes_sold_yoy	pending_sales_yoy
77	107	-0.02326	0.175
177	183	-0.07414	0.1391
155	145	0.04545	0.0101

median_ppsf	median_list_ppsf	homes_sold_yoy	pending_sales_yoy
69	79	0.2043	0.1449
135	122	0.1264	NA
218	224	0.1007	-0.01429

new_listings_yoy	inventory_yoy
-0.1875	-0.2957
-0.0911	-0.27
-0.08157	-0.218
0.05665	0.2042
-0.09444	-0.2632
-0.1557	0.02579