House_report

2023-01-28

House report

Data

The data for median house price per square feet is extracted from website. The data for number of home sold and median days on the market is captured from the website

Included Area

The excel file created for data analytics contains towns that has the following property:

Target area This is the area we focus on. There are five areas, namely Scarsdale, Purchase, Rye, Bedford, Chappaqua. Scarsdale is the benchmark, which has a high house price.

(This is the testing dataset. I create a linear model based on other area to predict the potential house price per square feet. So that it can show if the price still has potential to increase)

High median price per square foot In accordance to move.org webpage, cities like San Francisco, Boston and New York City have the mosty expensive house price per unit. Along with these cities, I also choose Fremont, CA, Oakland, CA, San Jose, CA and Honolulu as focused cities. These cities can give a clear indication of the important factors that should be considered, if the house price is high, or is going to be high in the target city.

Other chosen city Seattle and Huston are also two major cities that have a high demand of investment. These cities can also be act as an benchmarks. River Edge is very similar to Chappaqua, so I include it in the data for comparison.

Variables

Table 1 House price per square feet is included in the first table of the data. The table contains all the house price per square feet from 4th Feburary 2022 to 20th January 2023. The mean value for house price per square feet for each month is given in the aggragate columns in the last 12 rows, from BB to BM column in excel.

Table 2 Table 2 consists the number of purchased house unit for every month in the last three years.

Table 3 Table 3 consists the median time for the housing unit being purchased for every month in the last three years.

Table 4 Table 4 consists all the background information of the housing unit. In order to calculate the change in price over seasons so that the impact of the raise in interest rate can be reviewed, I choose to extract the data from Feb. 18th, May 13th, Aug 12th, Nov 18th and the latest available price, which is Jan 20, 2023. Columns C to K in excel are the calculated house values and the change in price is calculated in column H to K so that we can see the potential impact of rise in interest rate after August.

The rating of each city is also included in the table. This consist the overall rating of each city, as well as the rating for Housing, Education, Cost of living, Unemployment, Commute and Lifestyle. The rating is in a scale of 12. The scale is represented the following: A+: 12 A: 11 A-: 10 B+: 9 B: 8 B-: 7 C+: 6 C: 5 C-: 4 D+: 3 D: 2 D-: 1.

The description of the rating is evaluated from the following categories:

- Housing category: considers home inventory availability, median days on market, total new listing and estimated home sales.
- Cost of living: considers median home price and average cost per square foot.
- Education rating: evaluate the school grade and the percentage of Bachelor degree or greater people in the city.
- Unemployment: considers employment rate only.
- Life style: considers urban life style, including number of restaurant and bars, businesses and net migrations of the city.
- Commute rating: evaluate the median travel time to work, and percentage of people go to work by bike, public transit or on foot.

The table also contains the demographical details of each chosen cities, such as housing unit, median income and populations.

The last few columns contains the average school rating of top five schools in the city and the total number of schools in the city. The schools are classified into Elementary school, Middle school and High school.

Table 5 This is the population table for each target cities.

Table 6 This is the table which consists the number of house unit available for every month in the last two years.

Table 7 It is an incomplete table which may be some potential cities that are worth to invest in.

Results

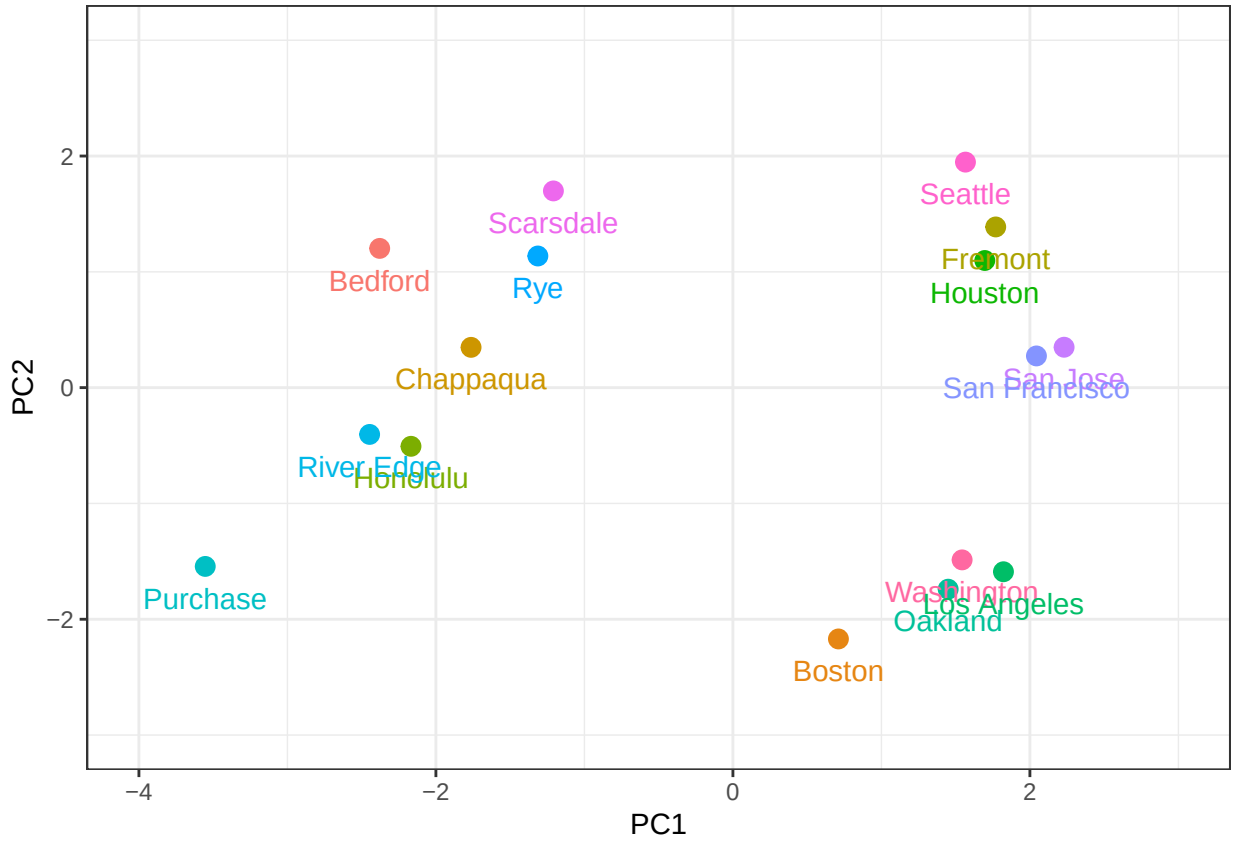
Clustering results

```
selected_info =
  all_info %>%
  janitor::clean_names() %>%
  select(overall_rating, housing, education, cost_of_living, unemployment, commute, lifestyle)
rownames(selected_info) = c("Chappaqua", "Bedford","Rye","Purchase","Scarsdale","Fremont", "Oakland", "

pca_res <- prcomp(selected_info, scale. = TRUE)
```

```
df <- as.data.frame(pca_res$x)
df$city <- rownames(df)

ggplot(df, aes(PC1, PC2, color = city)) +
  geom_point(size = 3) +
  geom_text(aes(label = city) , vjust = 2) +
  lims(x = c(-4, 3), y = c(-3, 3)) +
  theme_bw() +
  theme(legend.position = "none")
```

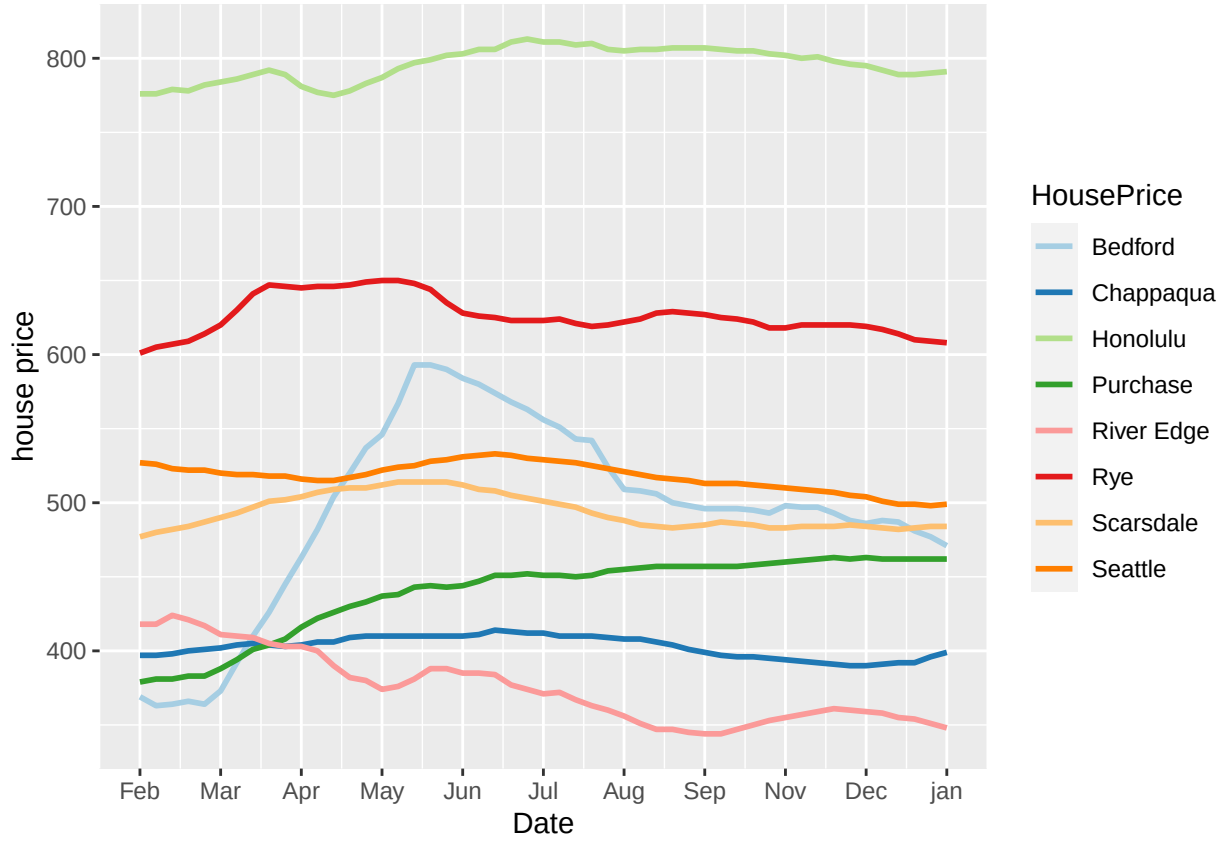


Honolulu has a rating that is close to the four target areas, so I include the city in the price time series plot. Seattle is also included for a reference as it has a high house price per square feet and it is a well developed city.

House price time series

```
time = colnames(house_price)[3:53]
df1 = t(house_price[,3:53])
data_df1 = melt(df1)
data_df1_named =
```

```
  data_df1 %>%
  mutate(Var2 = ifelse(Var2==1,"Chappaqua",ifelse(Var2==2, "Bedford", ifelse(Var2==3, "Rye", ifelse(Var1
         ind = rep(1:length(time),8),
         HousePrice = Var2)
ggplot(data_df1_named, aes(x=ind, y=value, color = HousePrice))+
  geom_line(size=1)+
  labs(
    y="house price",
    x="Date"
  )+
  scale_x_continuous(
    breaks = seq(1,length(time),5),
    labels = c("Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Nov","Dec","jan"),
    limits = c(1,length(time)))+
  scale_color_brewer(palette="Paired")
```




```
data_df1_named %>% group_by(Var2) %>% summarize(mean_price = mean(value))
```

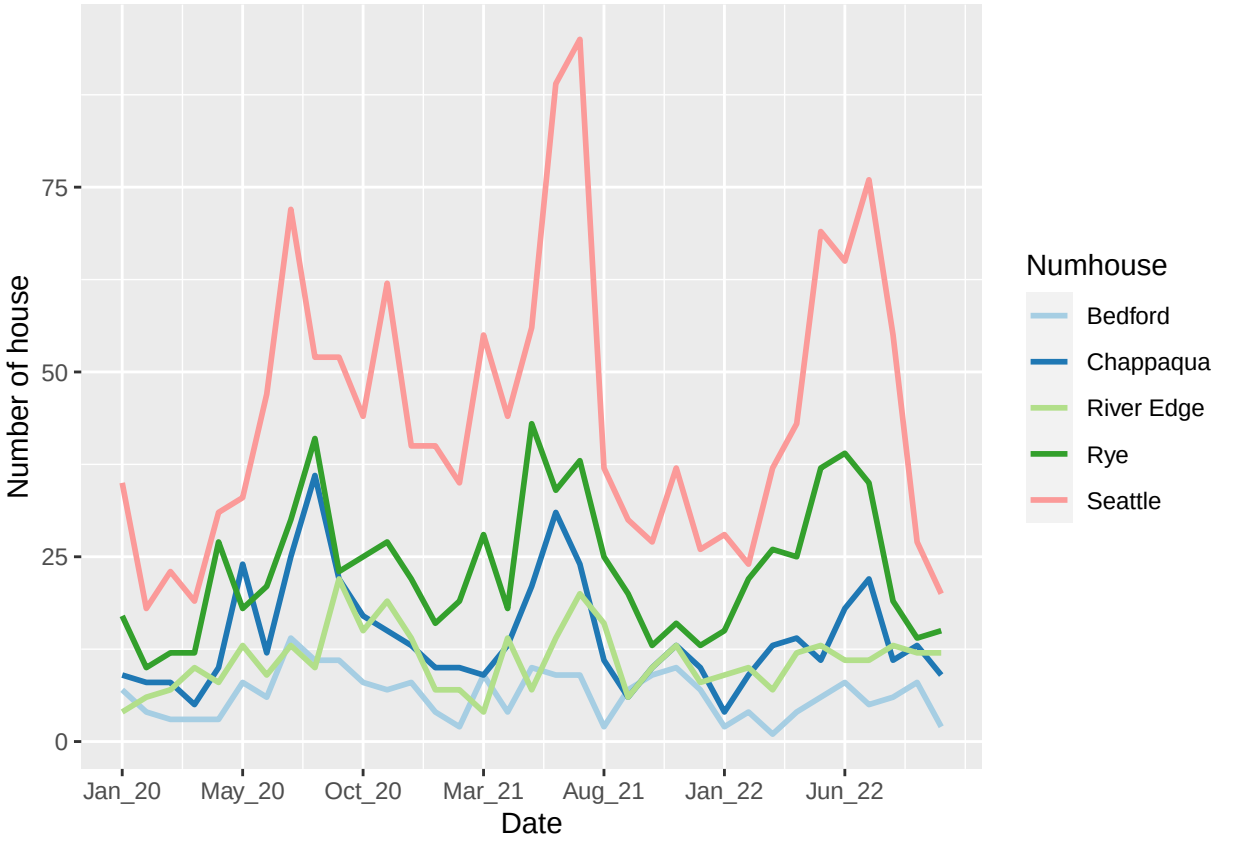
```
## # A tibble: 8 x 2
##   Var2       mean_price
##   <chr>           <dbl>
## 1 Bedford          494.
## 2 Chappaqua        403.
## 3 Honolulu         796.
## 4 Purchase         439.
```

```
## 5 River Edge       375.
## 6 Rye              626.
## 7 Scarsdale        494.
## 8 Seattle          518.
```

According to the time series plot for house price per square foot, Bedford has the most fluctuated price over the time. It is one of the most competitive homes sell in 18 days. There is a net population gain in Bedford, with a net search flow of approximately 1100. The house price for Bedford decreases after the increment of interest rate. Rye has the highest house price per square foot among target cities (Rye, Purchase, Chappaqua, Bedford) , and it is relatively stable. The house price per square feet for Purchase does not decrease after the rise in interest rate and it has a persistent increasing trend.

Number of house

```
time2 = colnames(num_sold)[3:37]
num_sold_num = num_sold %>% filter(City_name!="Purchase") %>% filter(City_name!="Seattle") %>% filter(C
df2 = t(num_sold_num[,3:37])
data_df2 = melt(df2)
data_df2_named =
  data_df2 %>%
  mutate(Var2 = ifelse(Var2==1,"Chappaqua",ifelse(Var2==2, "Bedford", ifelse(Var2==3, "Rye", ifelse(Var
         ind = rep(1:length(time2),5),
         Numhouse = Var2)
ggplot(data_df2_named, aes(x=ind, y=value, color = Numhouse))+
  geom_line(size=1)+
  labs(
    y="Number of house",
    x="Date"
  )+
  scale_x_continuous(
    breaks = seq(1,length(time2),5),
    labels = c("Jan_20","May_20","Oct_20","Mar_21","Aug_21","Jan_22","Jun_22"),
    limits = c(1,length(time2)))+
    scale_color_brewer(palette="Paired")
```




```
data_df2_named %>% group_by(Var2) %>% summarize(mean_number = mean(value))
```

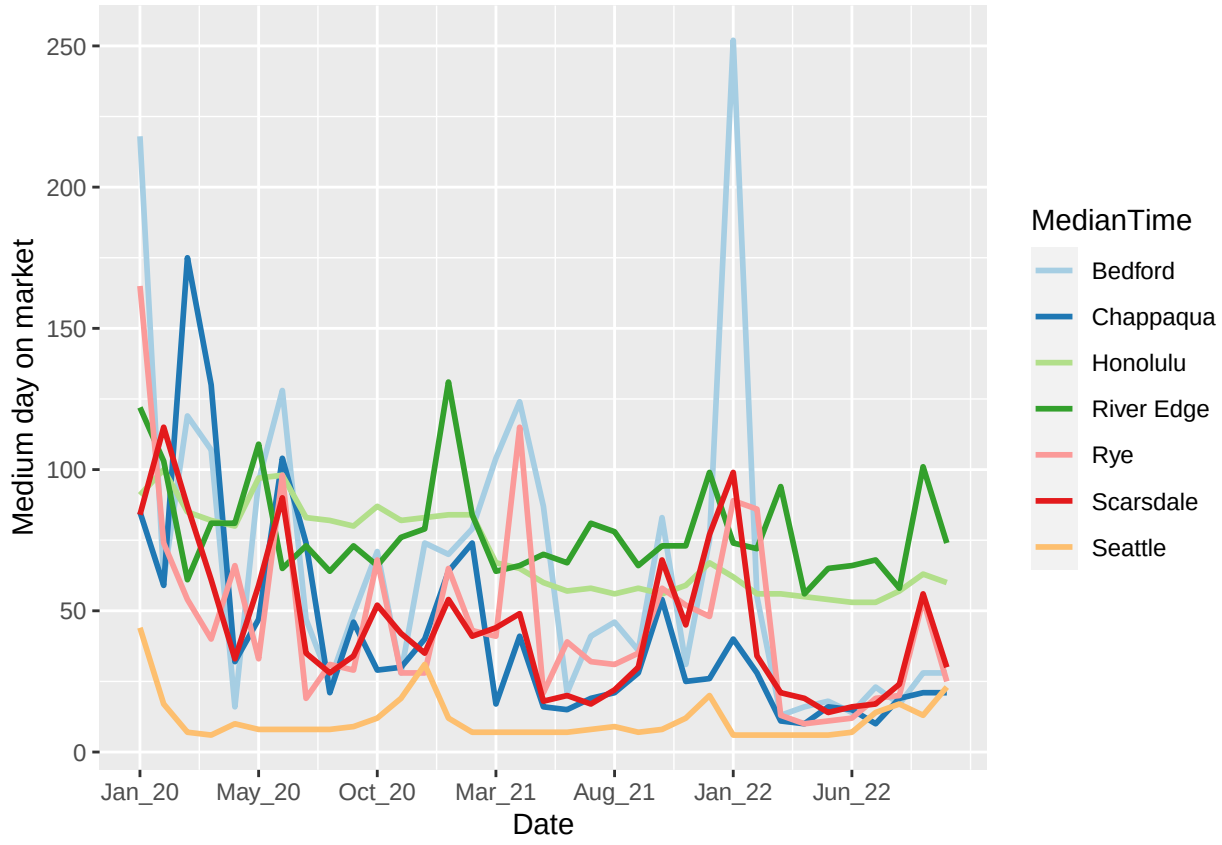
```
## # A tibble: 5 x 2
##   Var2       mean_number
##   <chr>            <dbl>
## 1 Bedford           6.31
## 2 Chappaqua        14.2
## 3 River Edge       11.0
## 4 Rye              23.3
## 5 Seattle          44.1
```

Purchase data is not available online. Rye is the city that has the highest number of house available among three target cities.

Medium day on the market

```
time3 = colnames(medium_time)[3:37]
medium_time_num = medium_time %>% filter(City_name!="Purchase")
df3 = t(medium_time_num[,3:37])
data_df3 = melt(df3)
data_df3_named =
  data_df3 %>%
  mutate(Var2 = ifelse(Var2==1,"Chappaqua",ifelse(Var2==2, "Bedford", ifelse(Var2==3, "Rye",ifelse(Var2=
```

```
        ind = rep(1:length(time3),7),
        MedianTime = Var2)
ggplot(data_df3_named, aes(x=ind, y=value, color = MedianTime))+
  geom_line(size=1)+
  labs(
    y="Medium day on market",
    x="Date"
  )+
  scale_x_continuous(
    breaks = seq(1,length(time2),5),
    labels = c("Jan_20","May_20","Oct_20","Mar_21","Aug_21","Jan_22","Jun_22"),
    limits = c(1,length(time2)))+
    scale_color_brewer(palette="Paired")
```




```
data_df3_named %>% group_by(Var2) %>% summarize(mean_day = mean(value))
```

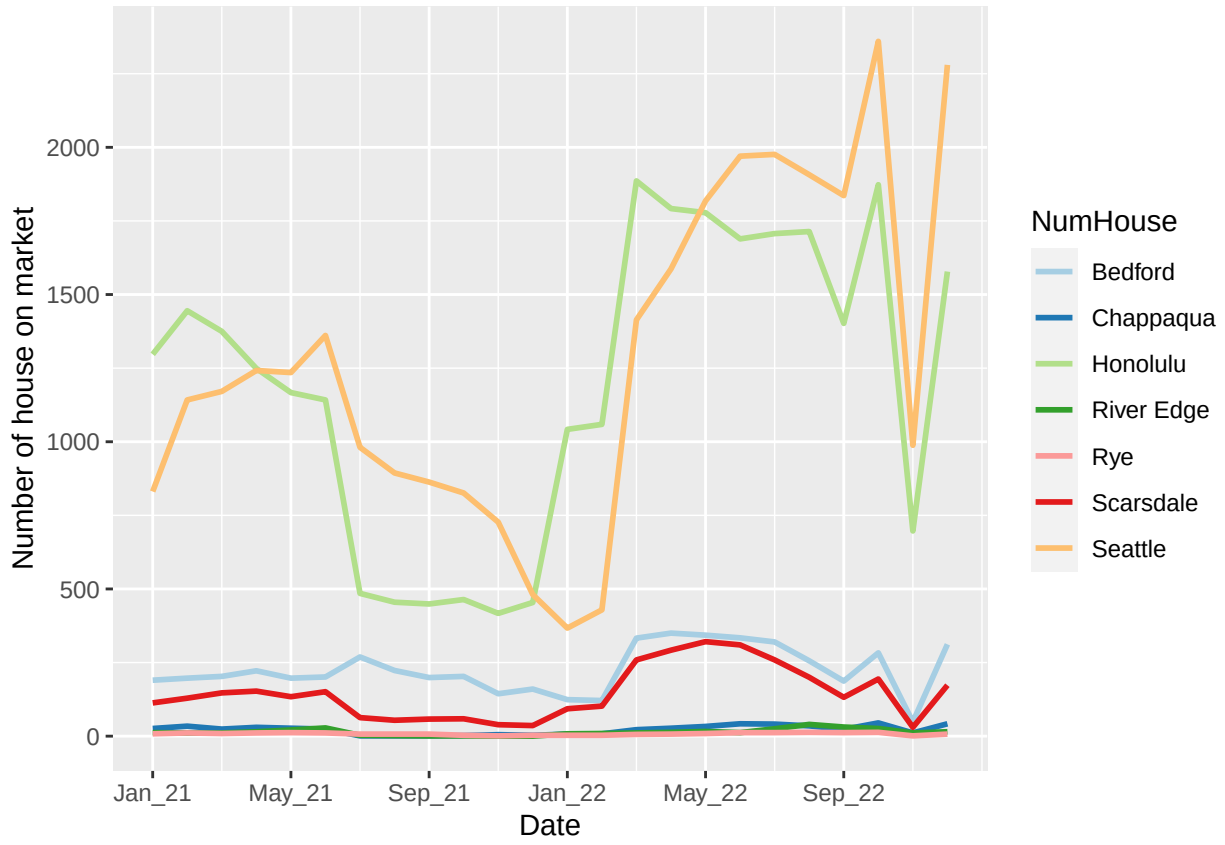
```
## # A tibble: 7 x 2
##   Var2        mean_day
##   <chr>          <dbl>
## 1 Bedford         65.7
## 2 Chappaqua       41.8
## 3 Honolulu        70.6
## 4 River Edge      78.1
## 5 Rye             47.2
## 6 Scarsdale       44.9
```

```
## 7 Seattle        11.5
```

Bedford has a relatively high number of days on the market before sold.

Home for sale

```
time4 = colnames(num_house)[3:26]
num_house_num = num_house %>% filter(City_name!="Chappaqua")
df4 = t(num_house_num[,3:26])
data_df4 = melt(df4)
data_df4_named =
  data_df4 %>%
  mutate(Var2 = ifelse(Var2==1,"Chappaqua",ifelse(Var2==2, "Bedford", ifelse(Var2==3, "Rye",ifelse(Var2=
         ind = rep(1:length(time4),7),
         NumHouse = Var2)
ggplot(data_df4_named, aes(x=ind, y=value, color = NumHouse))+
  geom_line(size=1)+
  labs(
    y="Number of house on market",
    x="Date"
  )+
  scale_x_continuous(
    breaks = seq(1,length(time4),4),
    labels = c("Jan_21","May_21","Sep_21","Jan_22","May_22","Sep_22"),
    limits = c(1,length(time4)))+
    scale_color_brewer(palette="Paired")
```




```
data_df4_named %>% group_by(Var2) %>% summarize(mean_number = mean(value))
```

```
## # A tibble: 7 x 2
##   Var2       mean_number
##   <chr>            <dbl>
## 1 Bedford          226.
## 2 Chappaqua         21.4
## 3 Honolulu        1192.
## 4 River Edge        13.4
## 5 Rye                7.83
## 6 Scarsdale        146.
## 7 Seattle         1279.
```