Final_Project

2023-11-06

```
head(DaysOnZillow_City)
```

```
## # A tibble: 6 x 128
##    ...1 SizeRank RegionID RegionName   RegionType StateName `2010-01` `2010-02`
##   <dbl>    <dbl>    <dbl> <chr>        <chr>      <chr>         <dbl>     <dbl>
## 1     0        1     6181 New York     City       NY              196      190
## 2     1        2    12447 Los Angeles  City       CA              118      136
## 3     2        3    39051 Houston      City       TX              133      137
## 4     3        4    17426 Chicago      City       IL              186      180.
## 5     4        5     6915 San Antonio  City       TX              101      110
## 6     5        6    13271 Philadelphia City       PA              127      139
## # i 120 more variables: `2010-03` <dbl>, `2010-04` <dbl>, `2010-05` <dbl>,
## #   `2010-06` <dbl>, `2010-07` <dbl>, `2010-08` <dbl>, `2010-09` <dbl>,
## #   `2010-10` <dbl>, `2010-11` <dbl>, `2010-12` <dbl>, `2011-01` <dbl>,
## #   `2011-02` <dbl>, `2011-03` <dbl>, `2011-04` <dbl>, `2011-05` <dbl>,
## #   `2011-06` <dbl>, `2011-07` <dbl>, `2011-08` <dbl>, `2011-09` <dbl>,
## #   `2011-10` <dbl>, `2011-11` <dbl>, `2011-12` <dbl>, `2012-01` <dbl>,
## #   `2012-02` <dbl>, `2012-03` <dbl>, `2012-04` <dbl>, `2012-05` <dbl>, ...
```

```
head(DaysOnZillow_State)
```

```
## # A tibble: 6 x 127
##    ...1 SizeRank RegionID RegionName   RegionType `2010-01` `2010-02` `2010-03`
##   <dbl>    <dbl>    <dbl> <chr>        <chr>          <dbl>     <dbl>     <dbl>
## 1     0        1        9 California   State            108       115       107
## 2     1        2       54 Texas        State            121       123       122
## 3     2        3       43 New York     State            188       194       192
## 4     3        4       14 Florida      State            161       156       153
## 5     4        5       21 Illinois     State            174       178       178
## 6     5        6       47 Pennsylvania State            138       150       151
## # i 119 more variables: `2010-04` <dbl>, `2010-05` <dbl>, `2010-06` <dbl>,
## #   `2010-07` <dbl>, `2010-08` <dbl>, `2010-09` <dbl>, `2010-10` <dbl>,
## #   `2010-11` <dbl>, `2010-12` <dbl>, `2011-01` <dbl>, `2011-02` <dbl>,
## #   `2011-03` <dbl>, `2011-04` <dbl>, `2011-05` <dbl>, `2011-06` <dbl>,
## #   `2011-07` <dbl>, `2011-08` <dbl>, `2011-09` <dbl>, `2011-10` <dbl>,
## #   `2011-11` <dbl>, `2011-12` <dbl>, `2012-01` <dbl>, `2012-02` <dbl>,
## #   `2012-03` <dbl>, `2012-04` <dbl>, `2012-05` <dbl>, `2012-06` <dbl>, ...
```

```
head(Sale_Prices_City)
```

```
## # A tibble: 6 x 150
##    ...1 RegionID RegionName   StateName   SizeRank `2008-03` `2008-04` `2008-05`
```

```
##   <dbl>    <dbl> <chr>        <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1     0     6181 New York     New York           1        NA        NA        NA
## 2     1    12447 Los Angeles  California         2    507600    489600    463000
## 3     2    39051 Houston      Texas              3    138400    135500    132200
## 4     3    17426 Chicago      Illinois           4    325100    314800    286900
## 5     4     6915 San Antonio  Texas              5    130900    131300    131200
## 6     5    13271 Philadelphia Pennsylvan~        6    111100    111000    111500
## # i 142 more variables: `2008-06` <dbl>, `2008-07` <dbl>, `2008-08` <dbl>,
## #   `2008-09` <dbl>, `2008-10` <dbl>, `2008-11` <dbl>, `2008-12` <dbl>,
## #   `2009-01` <dbl>, `2009-02` <dbl>, `2009-03` <dbl>, `2009-04` <dbl>,
## #   `2009-05` <dbl>, `2009-06` <dbl>, `2009-07` <dbl>, `2009-08` <dbl>,
## #   `2009-09` <dbl>, `2009-10` <dbl>, `2009-11` <dbl>, `2009-12` <dbl>,
## #   `2010-01` <dbl>, `2010-02` <dbl>, `2010-03` <dbl>, `2010-04` <dbl>,
## #   `2010-05` <dbl>, `2010-06` <dbl>, `2010-07` <dbl>, `2010-08` <dbl>, ...
```

```
head(Sale_Prices_State)
```

```
## # A tibble: 6 x 149
##    ...1 RegionID RegionName   SizeRank `2008-03` `2008-04` `2008-05` `2008-06`
##   <dbl>    <dbl> <chr>           <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1     0        9 California          1    392500    373800    351800    334700
## 2     1       54 Texas               2    139900    139300    137600    137400
## 3     2       43 New York            3        NA        NA        NA        NA
## 4     3       14 Florida             4    203400    195500    189300    184800
## 5     4       21 Illinois            5    204400    198400    185000    177500
## 6     5       47 Pennsylvania        6    146400    144800    142500    138800
## # i 141 more variables: `2008-07` <dbl>, `2008-08` <dbl>, `2008-09` <dbl>,
## #   `2008-10` <dbl>, `2008-11` <dbl>, `2008-12` <dbl>, `2009-01` <dbl>,
## #   `2009-02` <dbl>, `2009-03` <dbl>, `2009-04` <dbl>, `2009-05` <dbl>,
## #   `2009-06` <dbl>, `2009-07` <dbl>, `2009-08` <dbl>, `2009-09` <dbl>,
## #   `2009-10` <dbl>, `2009-11` <dbl>, `2009-12` <dbl>, `2010-01` <dbl>,
## #   `2010-02` <dbl>, `2010-03` <dbl>, `2010-04` <dbl>, `2010-05` <dbl>,
## #   `2010-06` <dbl>, `2010-07` <dbl>, `2010-08` <dbl>, `2010-09` <dbl>, ...
```

```
NYC_Sales_Price <- Sale_Prices_City %>%
  filter(RegionName == "New York", StateName == "New York")
NYC_Sales_Sum <- rowSums(NYC_Sales_Price[,39:150])
NYC_Mean <- NYC_Sales_Sum/(150-39)
NYC_Max <- apply(NYC_Sales_Price[, 39:150], 1, max)
NYC_Min <- apply(NYC_Sales_Price[, 39:150], 1, min)
NYC_Range <- NYC_Max - NYC_Min
NYC_Growth <- NYC_Range/111


NYC_Mean
```

```
## [1] 520486.5
```

```
NYC_Max
```

```
## [1] 575100
```

```
NYC_Min
```

```
## [1] 442700
```

```
NYC_Range
```

```
## [1] 132400
```

```
NYC_Growth
```

```
## [1] 1192.793
```

```
NYC_Zillow_Length<- DaysOnZillow_City %>%
  filter(RegionName == "New York", StateName == "NY")
NYC_Zillow_Sum <- rowSums(NYC_Zillow_Length[,7:128])
NYC_Mean_Length <- NYC_Zillow_Sum/(121)
NYC_Max_Length <- apply(NYC_Zillow_Length[, 7:128], 1, max)
NYC_Min_Length <- apply(NYC_Zillow_Length[, 7:128], 1, min)
NYC_Range_Length <- NYC_Max_Length - NYC_Min_Length

NYC_Mean_Length
```

```
## [1] 170.7769
```

```
NYC_Max_Length
```

```
## [1] 219
```

```
NYC_Min_Length
```

```
## [1] 120
```

```
NYC_Range_Length
```

```
## [1] 99
```

```
Medford_Sales_Price <- Sale_Prices_City %>%
  filter(RegionName == "Medford", StateName == "Massachusetts")
Medford_Sales_Sum <- rowSums(Medford_Sales_Price[,7:149])
Medford_Mean <- Medford_Sales_Sum/(149-7)
Medford_Max <- apply(Medford_Sales_Price[, 7:149], 1, max)
Medford_Min <- apply(Medford_Sales_Price[, 7:149], 1, min)
Medford_Range <- Medford_Max - Medford_Min
Medford_Growth <- Medford_Range/143

Medford_Mean
```

```
## [1] 408514.8
```

```
Medford_Max
```

```
## [1] 589900
```

```
Medford_Min
```

```
## [1] 268400
```

```
Medford_Range
```

```
## [1] 321500
```

```
Medford_Growth
```

```
## [1] 2248.252
```

```
Medford_Zillow_Length<- DaysOnZillow_City %>%
  filter(RegionName == "Medford", StateName == "MA")
Medford_Zillow_Sum <- rowSums(Medford_Zillow_Length[,82:128])
Medford_Mean_Length <- Medford_Zillow_Sum/(46)
Medford_Max_Length <- apply(Medford_Zillow_Length[, 82:128], 1, max)
Medford_Min_Length <- apply(Medford_Zillow_Length[, 82:128], 1, min)
Medford_Range_Length <- Medford_Max_Length - Medford_Min_Length

Medford_Mean_Length
```

```
## [1] 63.69565
```

```
Medford_Max_Length
```

```
## [1] 96.5
```

```
Medford_Min_Length
```

```
## [1] 46
```

```
Medford_Range_Length
```

```
## [1] 50.5
```

```
Updated_Mean_Sale_Data <- read_csv("~/Desktop/Data-200/Metro_mlp_uc_sfrcondo_sm_month.csv")
```

```
## Rows: 928 Columns: 73
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): RegionName, RegionType, StateName
## dbl (70): RegionID, SizeRank, 2018-03-31, 2018-04-30, 2018-05-31, 2018-06-30...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

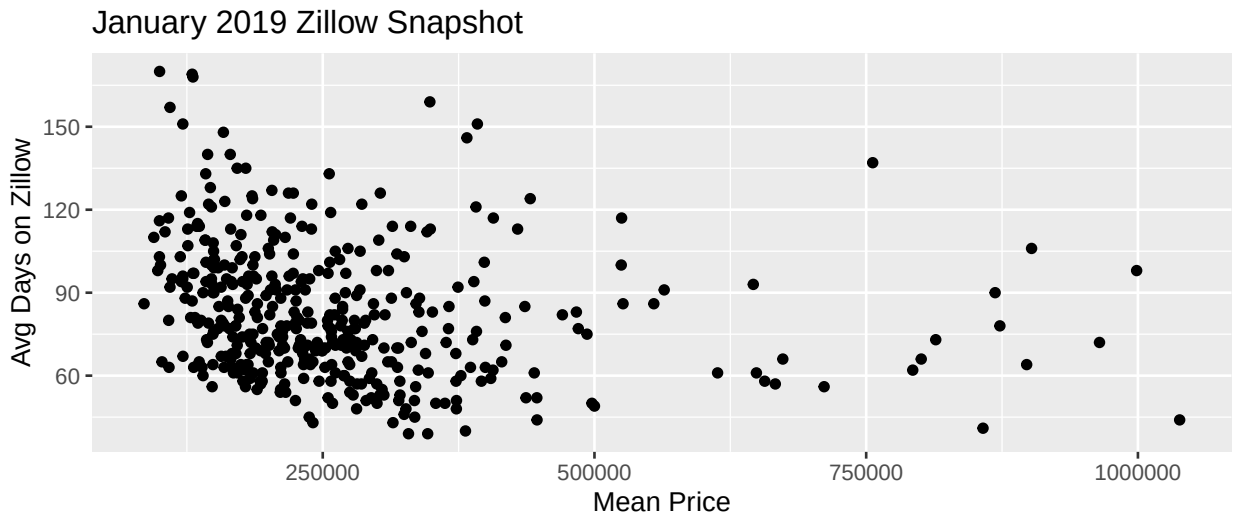
```
Updated_Median_Sale_Data <- read_csv("~/Desktop/Data-200/Metro_median_sale_price_uc_sfrcondo_sm_sa_month
```

```
## Rows: 771 Columns: 150
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (3): RegionName, RegionType, StateName
## dbl (147): RegionID, SizeRank, 2011-09-30, 2011-10-31, 2011-11-30, 2011-12-3...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

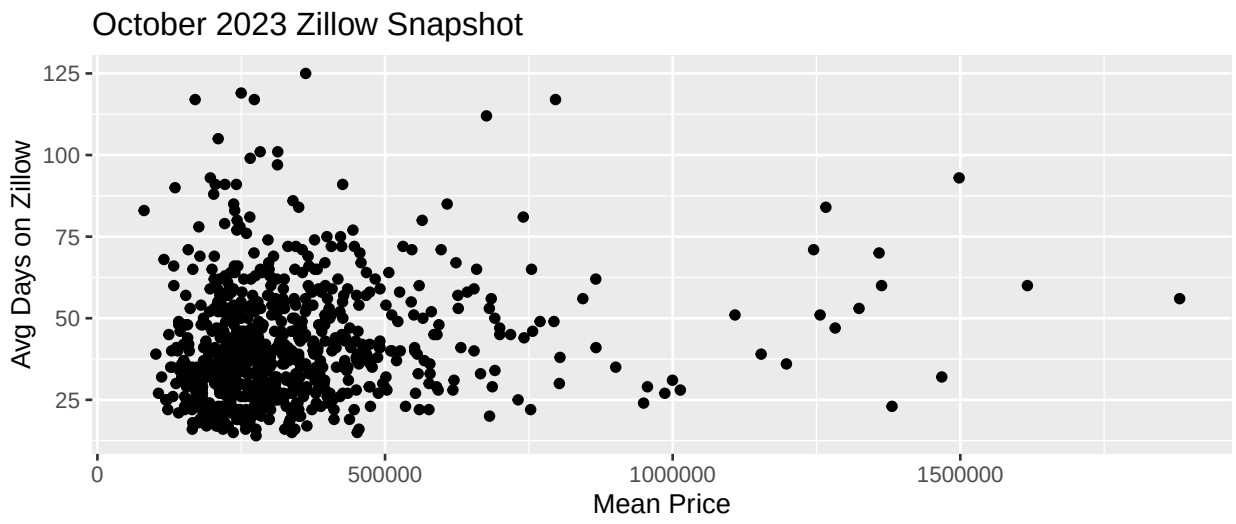
```
Updated_Days_on_Zillow <- read_csv("~/Desktop/Data-200/Metro_mean_doz_pending_uc_sfrcondo_sm_month.csv")
```

```
## Rows: 726 Columns: 73
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): RegionName, RegionType, StateName
## dbl (70): RegionID, SizeRank, 2018-03-31, 2018-04-30, 2018-05-31, 2018-06-30...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

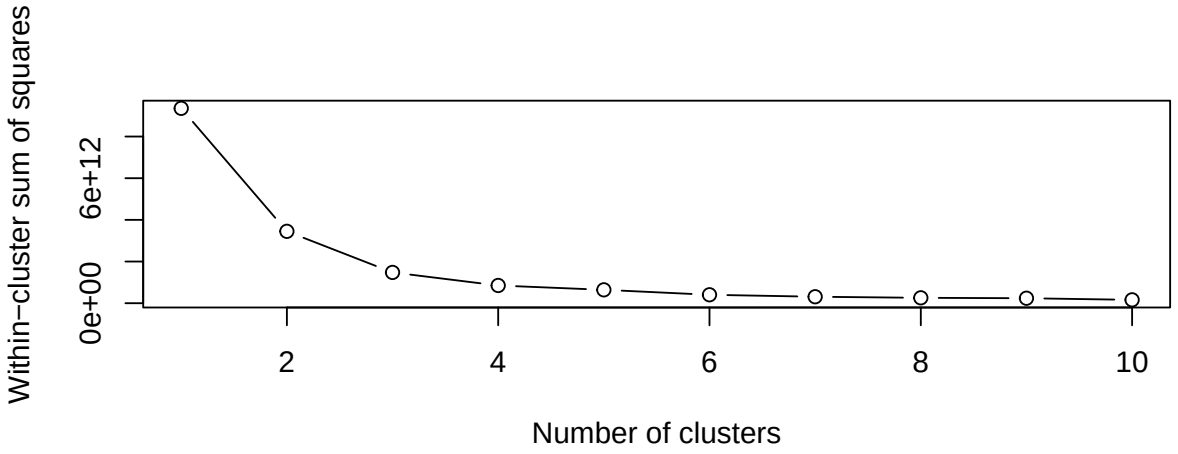
```
January_2019_Prices <- Updated_Mean_Sale_Data %>%
  dplyr::select(RegionID, RegionName, `2019-01-31`)

January_2019_Days_on_Zillow <- Updated_Days_on_Zillow %>%
  dplyr::select(RegionID, RegionName, `2019-01-31`)

January_Merge <- merge(January_2019_Prices, January_2019_Days_on_Zillow,
                       by = c("RegionID", "RegionName"))

January_Merge_Clean <- na.omit(January_Merge)

January_Merge_Isolate <- January_Merge_Clean %>%
  dplyr::select('2019-01-31.x', '2019-01-31.y')

ggplot(data = January_Merge_Clean, aes(x = `2019-01-31.x`, y = `2019-01-31.y`)) +
  geom_point() +
  labs(title = "January 2019 Zillow Snapshot",
       x = "Mean Price",
       y = "Avg Days on Zillow")
```



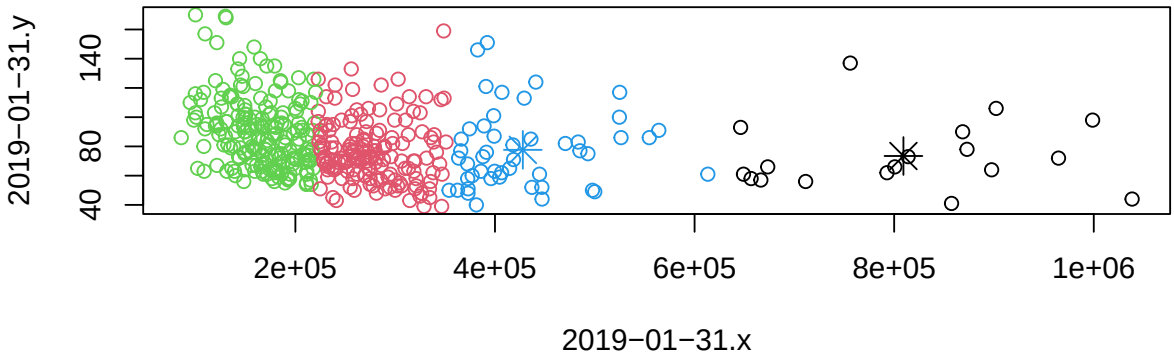

```
October_2023_Prices <- Updated_Mean_Sale_Data %>%
  dplyr::select(RegionID, RegionName, StateName, `2023-10-31`)

October_2023_Days_on_Zillow <- Updated_Days_on_Zillow %>%
  dplyr::select(RegionID, RegionName, StateName, `2023-10-31`)

October_2023_Merge <- merge(October_2023_Prices, October_2023_Days_on_Zillow,
                      by = c("RegionID", "RegionName", "StateName"))

October_2023_Merge_Clean <- na.omit(October_2023_Merge)


October_2023_Merge_Isolate <- October_2023_Merge_Clean %>%
  dplyr::select(`2023-10-31.x`, `2023-10-31.y`)

ggplot(data = October_2023_Merge_Clean, aes(x = `2023-10-31.x`, y = `2023-10-31.y`)) +
  geom_point() +
  labs(title = "October 2023 Zillow Snapshot",
       x = "Mean Price",
       y = "Avg Days on Zillow")
```



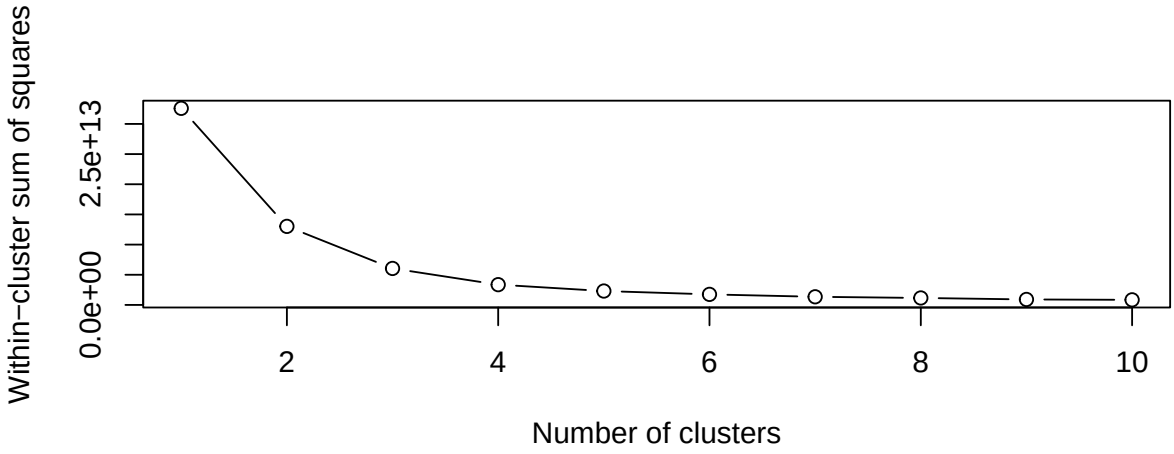

```
library(stats)
wcss <- numeric(length = 10) # Assuming a maximum of 10 clusters
for (i in 1:10) {
  model <- kmeans(January_Merge_Isolate, centers = i)
  wcss[i] <- model$tot.withinss
}
plot(1:10, wcss, type = "b", xlab = "Number of clusters", ylab = "Within-cluster sum of squares")
```



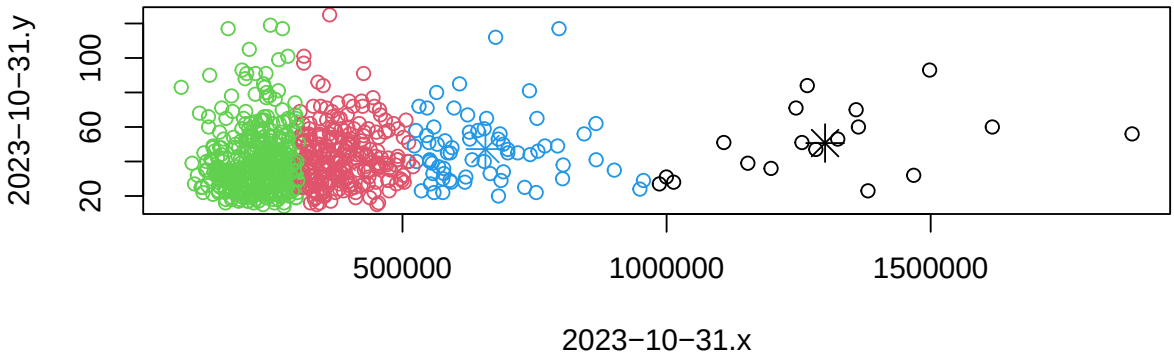

```
k <- 4
set.seed(123)
January_2019_model <- kmeans(January_Merge_Isolate, centers = k)

plot(January_Merge_Isolate, col = January_2019_model$cluster)
points(January_2019_model$centers, col = 1:k, pch = 8, cex = 2)
```




```
library(stats)
wcss <- numeric(length = 10) # Assuming a maximum of 10 clusters
for (i in 1:10) {
  model <- kmeans(October_2023_Merge_Isolate, centers = i)
  wcss[i] <- model$tot.withinss
}
plot(1:10, wcss, type = "b", xlab = "Number of clusters", ylab = "Within-cluster sum of squares")
```




```
k <- 4
set.seed(123)
October_2023_model <- kmeans(October_2023_Merge_Isolate, centers = k)

plot(October_2023_Merge_Isolate, col = October_2023_model$cluster)
points(October_2023_model$centers, col = 1:k, pch = 8, cex = 2)
```




```
cluster_assignments_2019 <- January_2019_model$cluster

January_2019_w_cluster <- cbind(January_Merge_Clean, Cluster = cluster_assignments_2019)
```

```
Cluster1_2019 <- January_2019_w_cluster %>%
  filter(Cluster == 1)

Cluster2_2019 <- January_2019_w_cluster %>%
  filter(Cluster == 2)

Cluster3_2019 <- January_2019_w_cluster %>%
  filter(Cluster == 3)

Cluster4_2019 <- January_2019_w_cluster %>%
  filter(Cluster == 4)
```

```
cluster_assignments_2023 <- October_2023_model$cluster

October_2023_w_cluster <- cbind(October_2023_Merge_Clean, Cluster = cluster_assignments_2023)
```

```
Cluster1_2023 <- October_2023_w_cluster %>%
  filter(Cluster == 1)

Cluster2_2023 <- October_2023_w_cluster %>%
  filter(Cluster == 2)

Cluster3_2023 <- October_2023_w_cluster %>%
  filter(Cluster == 3)

Cluster4_2023 <- October_2023_w_cluster %>%
  filter(Cluster == 4)
```