Housing Market Analysis

2024-01-20

PROJECT OVERVIEW We will understand the housing market with various predictors and perform association rule mining

First reading, processing the data

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.2
```

```
library(arules)
```

```
## Warning: package 'arules' was built under R version 4.3.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.3.3
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.2
```

```
dats <- read.delim("housing.csv", sep = ",", header = TRUE)
head(dats)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                 41         880            129        322
## 2   -122.22    37.86                 21        7099           1106       2401
## 3   -122.24    37.85                 52        1467            190        496
## 4   -122.25    37.85                 52        1274            235        558
## 5   -122.25    37.85                 52        1627            280        565
## 6   -122.25    37.85                 52         919            213        413
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252             452600        NEAR BAY
## 2       1138        8.3014             358500        NEAR BAY
## 3        177        7.2574             352100        NEAR BAY
## 4        219        5.6431             341300        NEAR BAY
## 5        259        3.8462             342200        NEAR BAY
## 6        193        4.0368             269700        NEAR BAY
```

```
dats$income_category <- cut(dats$median_income, breaks = c(-Inf, 3, Inf), labels = c("Low", "High"))

#Processing 'ocean_proximity'
dats$ocean_proximity <- as.factor(dats$ocean_proximity)

#Identifying numeric columns
numeric_cols <- sapply(dats, is.numeric)

#Converting numeric columns to factors
dats[, numeric_cols] <- lapply(dats[, numeric_cols], as.factor)

#Processing 'median_income' column
selected_cols <- c("ocean_proximity", "income_category", names(numeric_cols))
dats <- dats[, selected_cols, drop = FALSE]

#Creating the incidence matrix
incidence_matrix <- as(dats, "transactions")

write(incidence_matrix, file = "incidence_matrix.txt", sep = "\t")

summary(incidence_matrix)
```

```
## transactions as itemMatrix in sparse format with
##  20640 rows (elements/itemsets/transactions) and
##  32094 columns (items) and a density of 0.0004047476
##
## most frequent items:
##        income_category=High      income_category.1=High
##                       13237                       13237
##   ocean_proximity=<1H OCEAN ocean_proximity.1=<1H OCEAN
##                        9136                        9136
##         income_category=Low                     (Other)
##                        7403                      215964
##
## element (itemset/transaction) length distribution:
## sizes
##    12    13
##   207 20433
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   13.00   13.00   12.99   13.00   13.00
##
## includes extended item information - examples:
##                      labels       variables    levels
## 1 ocean_proximity=<1H OCEAN ocean_proximity <1H OCEAN
## 2    ocean_proximity=INLAND ocean_proximity    INLAND
## 3    ocean_proximity=ISLAND ocean_proximity    ISLAND
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

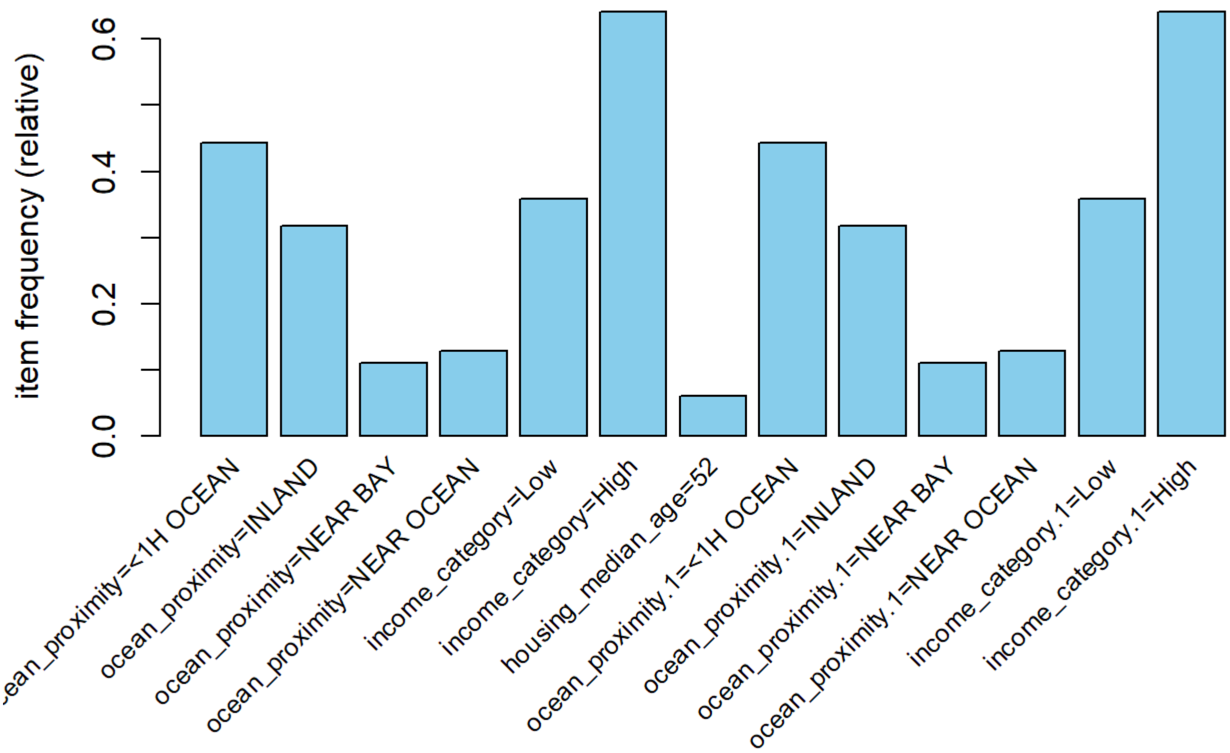
```
#Inspecting the incidence matrix created
inspect(incidence_matrix[1:10,])
```

```
##      items                          transactionID
## [1]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.23,
##       latitude=37.88,
##       housing_median_age=41,
##       total_rooms=880,
##       total_bedrooms=129,
##       population=322,
##       households=126,
##       median_income=8.3252,
##       median_house_value=452600,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                   1
## [2]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.22,
##       latitude=37.86,
##       housing_median_age=21,
##       total_rooms=7099,
##       total_bedrooms=1106,
##       population=2401,
##       households=1138,
##       median_income=8.3014,
##       median_house_value=358500,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                   2
## [3]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.24,
##       latitude=37.85,
##       housing_median_age=52,
##       total_rooms=1467,
##       total_bedrooms=190,
##       population=496,
##       households=177,
##       median_income=7.2574,
##       median_house_value=352100,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                   3
## [4]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.25,
##       latitude=37.85,
##       housing_median_age=52,
##       total_rooms=1274,
##       total_bedrooms=235,
##       population=558,
##       households=219,
##       median_income=5.6431,
##       median_house_value=341300,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                   4
## [5]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.25,
##       latitude=37.85,
```

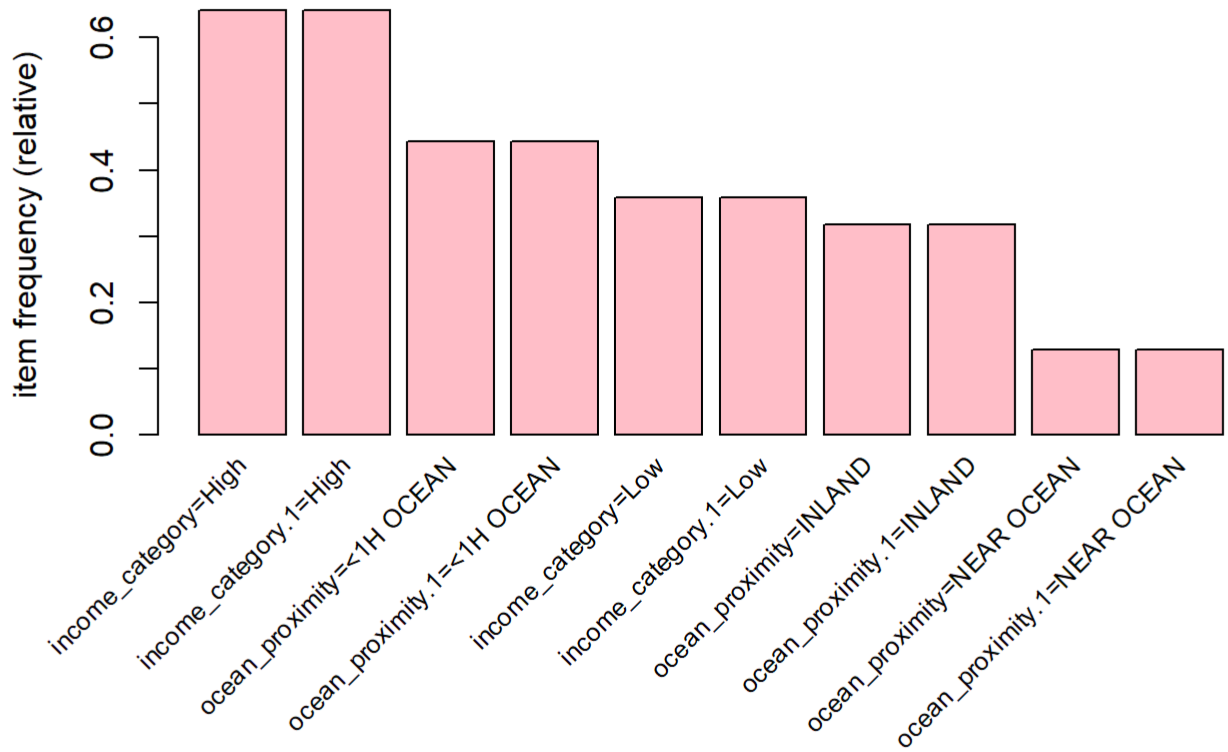
```
##       housing_median_age=52,
##       total_rooms=1627,
##       total_bedrooms=280,
##       population=565,
##       households=259,
##       median_income=3.8462,
##       median_house_value=342200,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                    5
## [6]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.25,
##       latitude=37.85,
##       housing_median_age=52,
##       total_rooms=919,
##       total_bedrooms=213,
##       population=413,
##       households=193,
##       median_income=4.0368,
##       median_house_value=269700,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                    6
## [7]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.25,
##       latitude=37.84,
##       housing_median_age=52,
##       total_rooms=2535,
##       total_bedrooms=489,
##       population=1094,
##       households=514,
##       median_income=3.6591,
##       median_house_value=299200,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                    7
## [8]  {ocean_proximity=NEAR BAY,
##       income_category=High,
##       longitude=-122.25,
##       latitude=37.84,
##       housing_median_age=52,
##       total_rooms=3104,
##       total_bedrooms=687,
##       population=1157,
##       households=647,
##       median_income=3.12,
##       median_house_value=241400,
##       ocean_proximity.1=NEAR BAY,
##       income_category.1=High}                    8
## [9]  {ocean_proximity=NEAR BAY,
##       income_category=Low,
##       longitude=-122.26,
##       latitude=37.84,
##       housing_median_age=42,
##       total_rooms=2555,
##       total_bedrooms=665,
##       population=1206,
##       households=595,
##       median_income=2.0804,
```

```
##        median_house_value=226700,
##        ocean_proximity.1=NEAR BAY,
##        income_category.1=Low}                 9
## [10] {ocean_proximity=NEAR BAY,
##        income_category=High,
##        longitude=-122.25,
##        latitude=37.84,
##        housing_median_age=52,
##        total_rooms=3549,
##        total_bedrooms=707,
##        population=1551,
##        households=714,
##        median_income=3.6912,
##        median_house_value=261100,
##        ocean_proximity.1=NEAR BAY,
##        income_category.1=High}                10
```

```
#Now visualazing the matrix
itemFrequencyPlot(incidence_matrix, support = 0.05, cex.names = 0.8, col = "skyblue")
```




```
itemFrequencyPlot(incidence_matrix, topN = 10, cex.names = 0.8, col = "pink")
```



#We can see that the itemfrequency plot shows the top rules for ocean proximity and income category being high and low.

Check top rules

```
my_params <- list(support = .005, confidence = .01, minlen = 2, maxlen = 6)
my_rules <- apriori(incidence_matrix, parameter = my_params)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.01    0.1    1 none FALSE            TRUE       5   0.005      2
##  maxlen target  ext
##       6  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 103
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[32094 item(s), 20640 transaction(s)] done [0.08s].
## sorting and recoding items ... [138 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 done [0.00s].
## writing ... [3426 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
inspect(sort(my_rules, by = "lift")[1:10])
```

```
##      lhs                              rhs                            support confidence    coverage
lift count
## [1]  {ocean_proximity=NEAR BAY}   => {ocean_proximity.1=NEAR BAY} 0.110949612          1 0.110949612
9.0131  2290
## [2]  {ocean_proximity.1=NEAR BAY} => {ocean_proximity=NEAR BAY}   0.110949612          1 0.110949612
9.0131  2290
## [3]  {ocean_proximity=NEAR BAY,
##       latitude=37.78}             => {ocean_proximity.1=NEAR BAY} 0.005620155          1 0.005620155
9.0131   116
## [4]  {latitude=37.78,
##       ocean_proximity.1=NEAR BAY} => {ocean_proximity=NEAR BAY}   0.005620155          1 0.005620155
9.0131   116
## [5]  {ocean_proximity=NEAR BAY,
##       latitude=37.76}             => {ocean_proximity.1=NEAR BAY} 0.005281008          1 0.005281008
9.0131   109
## [6]  {latitude=37.76,
##       ocean_proximity.1=NEAR BAY} => {ocean_proximity=NEAR BAY}   0.005281008          1 0.005281008
9.0131   109
## [7]  {ocean_proximity=NEAR BAY,
##       median_house_value=500001}  => {ocean_proximity.1=NEAR BAY} 0.009399225          1 0.009399225
9.0131   194
## [8]  {median_house_value=500001,
##       ocean_proximity.1=NEAR BAY} => {ocean_proximity=NEAR BAY}   0.009399225          1 0.009399225
9.0131   194
## [9]  {ocean_proximity=NEAR BAY,
##       housing_median_age=52}      => {ocean_proximity.1=NEAR BAY} 0.030329457          1 0.030329457
9.0131   626
## [10] {housing_median_age=52,
##       ocean_proximity.1=NEAR BAY} => {ocean_proximity=NEAR BAY}   0.030329457          1 0.030329457
9.0131   626
```

```
inspect(sort(my_rules, by = "confidence")[1:10])
```

```
##      lhs                                rhs                              support confidence  coverage
lift count
## [1]  {ocean_proximity=NEAR BAY}     => {ocean_proximity.1=NEAR BAY}   0.1109496          1 0.1109496
9.013100  2290
## [2]  {ocean_proximity.1=NEAR BAY}   => {ocean_proximity=NEAR BAY}     0.1109496          1 0.1109496
9.013100  2290
## [3]  {ocean_proximity.1=NEAR OCEAN} => {ocean_proximity=NEAR OCEAN}   0.1287791          1 0.1287791
7.765237  2658
## [4]  {ocean_proximity=NEAR OCEAN}   => {ocean_proximity.1=NEAR OCEAN} 0.1287791          1 0.1287791
7.765237  2658
## [5]  {ocean_proximity=INLAND}       => {ocean_proximity.1=INLAND}     0.3173934          1 0.3173934
3.150664  6551
## [6]  {ocean_proximity.1=INLAND}     => {ocean_proximity=INLAND}       0.3173934          1 0.3173934
3.150664  6551
## [7]  {income_category.1=Low}        => {income_category=Low}          0.3586725          1 0.3586725
2.788059  7403
## [8]  {income_category=Low}          => {income_category.1=Low}        0.3586725          1 0.3586725
2.788059  7403
## [9]  {ocean_proximity.1=<1H OCEAN}  => {ocean_proximity=<1H OCEAN}    0.4426357          1 0.4426357
2.259194  9136
## [10] {ocean_proximity=<1H OCEAN}    => {ocean_proximity.1=<1H OCEAN}  0.4426357          1 0.4426357
2.259194  9136
```

Now lets prepare a suggestion for anyone who would want a house which is neither too expensive nor very low in price but would want in a nicer location may be closer to the ocean and all. This can be useful for brokers and similar companies to make suggestions.

```
mydats <- read.delim("housing.csv", sep = ",", header = TRUE)

#Discretizing the "median_house_value" into categories with readable labels
mydats$median_house_value_categories <- cut(
  mydats$median_house_value,
  breaks = c(-Inf, 112000, 209000, 306000, 403000, Inf),
  labels = c("<112k", "112k-209k", "209k-306k", "306k-403k", ">403k"),
  include.lowest = TRUE
)

#Creating a new data frame with relevant columns
data_for_rules <- mydats[, c("ocean_proximity", "median_house_value_categories")]


trans_for_rules <- as(data_for_rules, "transactions")
```

```
## Warning: Column(s) 1 not logical or factor. Applying default discretization
## (see '? discretizeDF').
```

```
my_params <- list(support = .005, confidence = .01, minlen = 2, maxlen = 6)
my_rules <- apriori(trans_for_rules, parameter = my_params)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.01    0.1    1 none FALSE            TRUE       5   0.005      2
##  maxlen target  ext
##       6  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 103
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[10 item(s), 20640 transaction(s)] done [0.00s].
## sorting and recoding items ... [9 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 done [0.00s].
## writing ... [38 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
ocean_rules <- subset(my_rules, subset = grepl("ocean_proximity", labels(rhs(my_rules))))

#Sorting rules by confidence in descending order
ocean_rules <- sort(ocean_rules, by = "confidence", decreasing = TRUE)

inspect(head(ocean_rules))
```

```
##     lhs                                                rhs                                   support confidence
coverage      lift count
## [1] {median_house_value_categories=<112k}     => {ocean_proximity=INLAND}    0.16647287  0.7647452
0.21768411 2.4094551  3436
## [2] {median_house_value_categories=209k-306k} => {ocean_proximity=<1H OCEAN} 0.13410853  0.6064855
0.22112403 1.3701687  2768
## [3] {median_house_value_categories=306k-403k} => {ocean_proximity=<1H OCEAN} 0.05348837  0.5544952
0.09646318 1.2527125  1104
## [4] {median_house_value_categories=>403k}     => {ocean_proximity=<1H OCEAN} 0.04452519  0.5336818
0.08343023 1.2056909   919
## [5] {median_house_value_categories=112k-209k} => {ocean_proximity=<1H OCEAN} 0.18430233  0.4833545
0.38129845 1.0919918  3804
## [6] {median_house_value_categories=112k-209k} => {ocean_proximity=INLAND}    0.11923450  0.3127065
0.38129845 0.9852331  2461
```

#The rules indicate associations between specific values of “longitude” and the “ocean_proximity” attribute. #Recommendations:

#Based on the rules, it seems that certain ranges of “longitude” values are associated with the “ocean_proximity” being “<1H OCEAN.” #So, we can recommend that the person focuses on homes with longitude values similar to those indicated in the rules.

#Expectations:

#The confidence values indicate the likelihood of the association being true. For example, a confidence of 0.9857143 for the first rule means that, historically, when the “longitude” is -118.35, there’s a 98.57% chance that the “ocean_proximity” is “<1H OCEAN.”

#These are the observations #For a Lower Budget (<112k): homes in the “INLAND” area. #There is a high chance (approximately 76.47%) of finding an affordable home in an inland location but the trade of is that the distance to the ocean increases

#For a Moderate Budget (209k-306k): homes in areas labeled “<1H OCEAN. #There is a high likelihood (approximately 60.65%) of finding suitable homes close to the ocean in this budget range.

#For a Mid-Range Budget (306k-403k): homes in areas labeled “<1H OCEAN. #A majority of homes in this price range (approximately 55.45%) are located close to the ocean.

#For a Higher Budget (>403k): homes in areas labeled “<1H OCEAN #Homes in this budget range (approximately 53.37%) are likely to be situated near the ocean. #While affordability is a consideration, there’s still a moderate chance (approximately 48.34%) of finding homes close to the ocean.

#These are the expectations and recommendations I will advice:

#Recommendations:

#Recommendation for Lower Budget (<112k): #The person should consider prioritizing affordability in the “INLAND” area if cost savings are crucial. However, be prepared for a compromise in terms of distance from the ocean.

#Recommendation for Moderate to Higher Budgets (209k and above): #The person should focus their search on areas labeled “<1H OCEAN” to maximize the chances of finding a suitable home close to the ocean. This provides a good balance between budget considerations and the desire for coastal proximity.

#Expectations:

#Expectation for Lower Budget (<112k): #As the person focus on more affordable options in the “INLAND” area, the trade-off will involve an increase in distance from the ocean, providing cost savings but sacrificing proximity to coastal areas.

#Expectation for Moderate to Higher Budgets (209k and above): #With increasing budget ranges, there’s a positive correlation between budget and the likelihood of finding homes close to the ocean. The person can expect a better balance between affordability and proximity to the ocean, especially in areas labeled “<1H OCEAN.”

Lets see which area are more peaceful and less populated as they could be desired by some people

```
summary(mydats$population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       3     787    1166    1425    1725   35682
```

```
# Creating breaks and labels for population ranges
population_breaks <- c(-Inf, 500, 1000, 1500, 2000, 2500, 3000, Inf)
population_labels <- c("Very Low (<500)", "Low (500-1000)", "Moderate (1000-1500)",
                        "High (1500-2000)", "Very High (2000-2500)", "Extremely High (2500-3000)", "Ult
ra High (>3000)")

#Discretizing the "population" into categories
mydats$population_categories <- cut(mydats$population, breaks = population_breaks, labels = population_
labels, include.lowest = TRUE)

#Selecting columns of interest
columns_of_interest <- c("population_categories", "housing_median_age", "median_income")

#Creating a new data frame with relevant columns
data_for_population <- mydats[, columns_of_interest]

#Converting to transactions
trans_for_population <- as(data_for_population, "transactions")
```

```
## Warning: Column(s) 2, 3 not logical or factor. Applying default discretization
## (see '? discretizeDF').
```

```
#Setting parameters for association rule mining with lower support threshold
population_params <- list(support = 0.001, confidence = 0.01, minlen = 2, maxlen = 6)

#Mining association rules
all_rules <- apriori(trans_for_population, parameter = population_params)
```

```
## Apriori
## 
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.01    0.1    1 none FALSE            TRUE       5   0.001      2
##  maxlen target  ext
##       6  rules TRUE
## 
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
## 
## Absolute minimum support count: 20 
## 
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[13 item(s), 20640 transaction(s)] done [0.00s].
## sorting and recoding items ... [13 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [278 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
#Filtering rules for "Low," "Very Low," and "Moderate" population areas
low_population_rules <- subset(all_rules, subset = grepl("population_categories=Low", labels(lhs(all_ru
les))))
very_low_population_rules <- subset(all_rules, subset = grepl("population_categories=Very Low", labels
(lhs(all_rules))))
moderate_population_rules <- subset(all_rules, subset = grepl("population_categories=Moderate", labels
(lhs(all_rules))))

#Combining rules using c function
combined_rules <- c(low_population_rules, very_low_population_rules, moderate_population_rules)

#Sorting rules by confidence in descending order
combined_rules <- sort(combined_rules, by = "confidence", decreasing = TRUE)

#Displaying the top 10 rules for "Low," "Very Low," and "Moderate" population areas
inspect(head(combined_rules, 10))
```

```
##      lhs                                                 rhs                              support confid
ence   coverage     lift count
## [1]  {population_categories=Low (500-1000),
##       median_income=[2.89,4.24)}                      => {housing_median_age=[35,52]} 0.04869186  0.519
6484 0.09370155 1.496309  1005
## [2]  {population_categories=Low (500-1000)}        => {housing_median_age=[35,52]} 0.14331395  0.478
8732 0.29927326 1.378898  2958
## [3]  {population_categories=Low (500-1000),
##       median_income=[0.5,2.89)}                       => {housing_median_age=[35,52]} 0.04505814  0.468
5139 0.09617248 1.349069   930
## [4]  {population_categories=Low (500-1000),
##       median_income=[4.24,15]}                        => {housing_median_age=[35,52]} 0.04956395  0.453
0558 0.10939922 1.304558  1023
## [5]  {population_categories=Very Low (<500),
##       median_income=[0.5,2.89)}                       => {housing_median_age=[35,52]} 0.01695736  0.452
1964 0.03750000 1.302083   350
## [6]  {population_categories=Very Low (<500),
##       median_income=[2.89,4.24)}                      => {housing_median_age=[35,52]} 0.01274225  0.448
8055 0.02839147 1.292319   263
## [7]  {population_categories=Very Low (<500)}       => {housing_median_age=[35,52]} 0.04249031  0.442
9293 0.09593023 1.275399   877
## [8]  {population_categories=Very Low (<500),
##       median_income=[4.24,15]}                        => {housing_median_age=[35,52]} 0.01279070  0.425
8065 0.03003876 1.226094   264
## [9]  {population_categories=Moderate (1000-1500),
##       median_income=[0.5,2.89)}                       => {housing_median_age=[35,52]} 0.03759690  0.420
3684 0.08943798 1.210436   776
## [10] {population_categories=Moderate (1000-1500),
##       median_income=[2.89,4.24)}                      => {housing_median_age=[35,52]} 0.03953488  0.416
9647 0.09481589 1.200635   816
```

#We see that, based on the association rules, here are some characteristics associated with low population areas: #Low population areas with a median income between 2.89 and 4.24 are strongly associated with a housing median age between 35 and 52, with a confidence of 51.96%.

#In low population areas, a lower median income range of 0.5 to 2.89 is also associated with a housing median age between 35 and 52, with a confidence of 46.85%.

#Higher median income (4.24 to 15) in low population areas is still associated with a housing median age between 35 and 52, with a confidence of 45.31%.

#In very low population areas with a median income between 0.5 and 2.89, there is a strong association with a housing median age between 35 and 52, with a confidence of 45.22%.

#Low population areas, regardless of median income, often exhibit a housing median age between 35 and 52, with a confidence of 47.89%.

#Therefore, We can say that low and very low population areas irrespective of the median income are strongly associated with the housing median age between 35 to 52

Introduction

In this project, I performed an analysis of the housing market using various data preprocessing techniques and association rule mining. My goal was to uncover patterns and relationships within the data that could inform decision-making for potential homebuyers and real estate professionals. Through the use of association rules, I was able to provide insights and recommendations that cater to different budget levels and preferences, particularly focusing on proximity to the ocean and population density.

Data Preparation

Initial Data Processing

To begin, I loaded the necessary libraries (readr, arules, corrplot, and readxl) and imported the housing dataset from a CSV file. The first step was to explore and preprocess the data to ensure it was ready for analysis. I introduced a new categorical variable, income_category, based on the median_income column, dividing it into “Low” and “High” income categories.

Next, I processed the ocean_proximity variable, converting it into a factor, and identified the numeric columns in the dataset. These numeric columns were then transformed into factors to facilitate the creation of an incidence matrix. The incidence matrix, which is crucial for association rule mining, was created and inspected to verify its accuracy. I also visualized the frequency of items using itemFrequencyPlot, which helped to highlight the most significant rules related to ocean proximity and income levels.

Association Rule Mining

Identifying Key Rules

With the processed data, I applied association rule mining to uncover patterns in the housing market. Using the Apriori algorithm, I generated rules with specific parameters for support, confidence, and the length of the rules. By sorting the rules by lift and confidence, I identified the top associations that could be leveraged for making recommendations. These rules provided valuable insights into how certain attributes, like ocean proximity and income category, are linked to other housing characteristics.

Recommendations for Homebuyers

Tailored Suggestions for Different Budgets

Based on the rules generated, I formulated suggestions for potential homebuyers, especially those seeking homes that strike a balance between affordability and desirable locations, such as proximity to the ocean. For this, I discretized the median_house_value into categories, making it easier to understand the relationship between home prices and ocean proximity.

I then created a new dataset focusing on ocean proximity and housing values and applied association rule mining to this subset. The rules I uncovered allowed me to provide targeted recommendations for different budget ranges. For example, for a lower budget (<112k), I suggested focusing on homes in inland areas, where affordability is higher but distance from the ocean increases. For moderate to higher budgets (209k and above), I recommended searching in areas labeled “<1H OCEAN,” where there is a higher likelihood of finding homes close to the ocean.

Population Density Analysis

Exploring Preferences for Less Populated Areas

In addition to ocean proximity, I explored another dimension of the housing market: population density. Some homebuyers may prioritize living in peaceful, less populated areas, so I analyzed the relationship between population density and other factors, such as housing median age and median income.

By discretizing the population variable into categories (e.g., Very Low, Low, Moderate), I created a new dataset to explore these associations. I applied association rule mining with parameters tailored for lower support thresholds to capture even the less frequent but significant patterns. The resulting rules highlighted strong associations between low population areas and specific ranges of housing median age and median income.

Conclusions and Implications

Insights for Real Estate Professionals and Buyers Through this analysis, I identified several key insights that can guide real estate professionals and homebuyers. For instance, low and very low population areas are strongly associated with housing that is older, typically with a median age between 35 and 52 years, regardless of the income level. This information can be valuable for those looking for quieter, less populated areas.

Additionally, the association rules related to ocean proximity provide actionable insights for those with varying budget levels. By understanding these patterns, homebuyers can make more informed decisions, balancing their budget with their desire for coastal living.

Final Thoughts

This project demonstrates the power of association rule mining in uncovering hidden patterns in the housing market. By processing and analyzing the data, I was able to derive meaningful insights that can directly impact decision-making for both homebuyers and real estate professionals. Whether it’s finding an affordable home in a peaceful area or securing a property close to the ocean, these recommendations offer practical guidance based on data-driven analysis.