Exercise 5.2: House data transformation

Sreenivasulu Somu

Department of Data Science, Bellevue University

DSC520-T301 Statistics for Data Science (2247-1)

Chase Denton

July 7, 2024

Exercise 5.2: House data transformation

```
> # Load the libraries

> library(dplyr)

> library(purrr)

> library(stringr)

> library(readxl)

>

> # Reading the Excel file

> housing <- read_excel("housing.xlsx")

> head(housing)

# A tibble: 6 × 24

  `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full

  <dttm>                   <dbl>       <dbl>          <dbl> <chr>       <chr>    <chr>

1 2006-01-03 00:00:00      698000          1              3 NA          R1       17021 NE …

2 2006-01-03 00:00:00      649990          1              3 NA          R1       11927 178…

3 2006-01-03 00:00:00      572500          1              3 NA          R1       13315 174…

4 2006-01-03 00:00:00      420000          1              3 NA          R1       3303 178T…

5 2006-01-03 00:00:00      369900          1              3 15          R1       16126 NE …

6 2006-01-03 00:00:00      184667          1             15 18 51       R1       8101 229T…

# ℹ 17 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,

#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,

#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,

#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,

#   present_use <dbl>
```

```
>

>

> # Using GroupBy and Summarize

> city_summary <- housing %>%

+   group_by(ctyname) %>%

+   summarise(

+     avg_price = mean(`Sale Price`, na.rm = TRUE),

+     avg_bedrooms = mean(bedrooms, na.rm = TRUE),

+     count = n()

+   )

>

> # Mutate

> housing_mutated <- housing %>%

+   mutate(price_per_sqft = `Sale Price` / square_feet_total_living)

>

> # Filter

> expensive_houses <- housing %>%

+   filter(`Sale Price` > 1000000)

>

> # Select

> selected_columns <- housing %>%

+   select(`Sale Price`, bedrooms, bath_full_count, square_feet_total_living, year_built)

>

> # Arrange
```

```
> sorted_houses <- housing %>%
+   arrange(desc(`Sale Price`))
>
> # b. Using purrr package
>
> # Using keep function and keep only houses with more than 3 bedrooms
> houses_large <- keep(split(housing, 1:nrow(housing)), ~ .x$bedrooms > 3)
>
> # Using discard function and discard the houses built before 1980
> houses_newer <- discard(split(housing, 1:nrow(housing)), ~ .x$year_built < 1980)
>
> # Combine sale price, bedrooms, and square footage into a single list
> combined_data <- map2(housing$`Sale Price`, housing$bedrooms,
+                       ~ list(price = .x, bedrooms = .y,
+                              sqft = housing$square_feet_total_living[which(housing$`Sale Price` == .x)]))
>
> # 4. Using compact function to create a list with some NULL values and remove them
> sample_list <- list(a = 1, b = NULL, c = 3, d = NULL, e = 5)
> compact_list <- compact(sample_list)
>
> # Print the results
> cat("Number of houses with > 3 bedrooms: ", length(houses_large), "\n")
Number of houses with > 3 bedrooms:  6662
> cat("Number of houses built in 1980 or later: ", length(houses_newer), "\n")
```

```
Number of houses built in 1980 or later:  9608

> cat("\nSample of combined data (price, bedrooms, square footage):\n")


Sample of combined data (price, bedrooms, square footage):

> print(head(combined_data))

[[1]]

[[1]]$price

[1] 698000


[[1]]$bedrooms

[1] 4


[[1]]$sqft

[1] 2810 2700 3090 3140 3310 2950 2830 2640



[[2]]

[[2]]$price

[1] 649990


[[2]]$bedrooms

[1] 4


[[2]]$sqft
```

```
[1] 2880 3160 2050 2050



[[3]]

[[3]]$price

[1] 572500


[[3]]$bedrooms

[1] 4


[[3]]$sqft

[1] 2770 1790 2290 2160



[[4]]

[[4]]$price

[1] 420000


[[4]]$bedrooms

[1] 3


[[4]]$sqft

 [1] 1620 1980 1150 1200 1840 1560 1250 1210 1440 1440 1730 1640 2550 1840 2050 1930
1550 1580
```

```
[19] 2420 1770 1680 2830 1640 1810 2290 1870 2240 2320 2230 3910 1730 1930 2530 1940
2140 2640
[37] 1940 2650 1640 1840 1440 1780 1640 1440 1440 1950 1480 1820 1310 2390 1870 1450
970 1270
```

```
[[5]]
[[5]]$price
[1] 369900

[[5]]$bedrooms
[1] 3

[[5]]$sqft
[1] 1440 3030 1370
```

```
[[6]]
[[6]]$price
[1] 184667

[[6]]$bedrooms
[1] 4
```

```
[[6]]$sqft

[1] 4160




>

> cat("\nOriginal list with NULL values:\n")


Original list with NULL values:

> print(sample_list)

$a

[1] 1


$b

NULL


$c

[1] 3


$d

NULL


$e

[1] 5
```

```
>
> cat("\nCompact list with NULL values removed:\n")

Compact list with NULL values removed:
> print(compact_list)
$a
[1] 1

$c
[1] 3

$e
[1] 5

>
> # Display the first few entries of each result
> cat("\nSample of houses with more than 3 bedrooms:\n")

Sample of houses with more than 3 bedrooms:
> head(bind_rows(houses_large))
# A tibble: 6 × 24
  `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
  <dttm>                     <dbl>       <dbl>          <dbl> <chr>       <chr>    <chr>
1 2006-01-03 00:00:00       698000           1              3 NA          R1       17021 NE …
```

```
2 2006-01-03 00:00:00      649990          1            3 NA         R1      11927 178…

3 2006-01-03 00:00:00      572500          1            3 NA         R1      13315 174…

4 2006-01-03 00:00:00      184667          1           15 18 51       R1      8101 229T…

5 2006-01-04 00:00:00     1050000          1            3 NA         R1      21634 NE …

6 2006-01-04 00:00:00      875000          1            3 NA         R1      21404 NE …

# ℹ 17 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,

#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,

#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,

#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,

#   present_use <dbl>

>

> cat("\nSample of houses built in 1980 or later:\n")


Sample of houses built in 1980 or later:

> head(bind_rows(houses_newer))

# A tibble: 6 × 24

  `Sale Date`        `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full

  <dttm>                  <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>

1 2006-01-03 00:00:00      698000          1            3 NA         R1      17021 NE …

2 2006-01-03 00:00:00      649990          1            3 NA         R1      11927 178…

3 2006-01-03 00:00:00      572500          1            3 NA         R1      13315 174…

4 2006-01-03 00:00:00      369900          1            3 15         R1      16126 NE …

5 2006-01-03 00:00:00      184667          1           15 18 51       R1      8101 229T…

6 2006-01-04 00:00:00     1050000          1            3 NA         R1      21634 NE …
```

```
# ℹ 17 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,
#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,
#   present_use <dbl>
>
>
> # c. Using cbind and rbind
>
> # cbind
> housing_extended <- cbind(housing, price_per_sqft = housing$`Sale Price` /
housing$square_feet_total_living)
>
> # rbind
> housing_sample <- housing[1:10, ]
> housing_combined <- rbind(housing, housing_sample)
>
>
> # d. Split a string and concatenate using addr_full column
>
> split_address <- str_split(housing$addr_full, " ")
> concatenated_address <- sapply(split_address, paste, collapse = "-")
>
> # Printing the results
```

```
> head(city_summary)
# A tibble: 3 × 4
  ctyname   avg_price avg_bedrooms count
  <chr>        <dbl>       <dbl> <int>
1 REDMOND     644803.        3.68  6721
2 SAMMAMISH   972480.        4.09    66
3 NA          674973.        3.25  6078
> head(housing_mutated)
# A tibble: 6 × 25
  `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
  <dttm>                     <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>
1 2006-01-03 00:00:00       698000          1              3 NA          R1       17021 NE …
2 2006-01-03 00:00:00       649990          1              3 NA          R1       11927 178…
3 2006-01-03 00:00:00       572500          1              3 NA          R1       13315 174…
4 2006-01-03 00:00:00       420000          1              3 NA          R1       3303 178T…
5 2006-01-03 00:00:00       369900          1              3 15          R1       16126 NE …
6 2006-01-03 00:00:00       184667          1             15 18 51       R1       8101 229T…
# ℹ 18 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,
#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,
#   present_use <dbl>, price_per_sqft <dbl>
> head(expensive_houses)
# A tibble: 6 × 24
```

```
  `Sale Date`        `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
  <dttm>                  <dbl>      <dbl>         <dbl> <chr>       <chr>    <chr>
1 2006-01-04 00:00:00     1050000          1             3 NA          R1       21634 NE …
2 2006-01-12 00:00:00     1392000          1             3 NA          R1       2428 W LA…
3 2006-01-23 00:00:00     1445000          1             3 NA          R1       20425 NE …
4 2006-01-26 00:00:00     1053649          1             3 NA          R1       23821 NE …
5 2006-02-01 00:00:00     1900000          1             3 15 52       R1        6507 240T…
6 2006-02-01 00:00:00     1080135          1             3 NA          R1       23837 NE …
# ℹ 17 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,
#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,
#   present_use <dbl>
> head(selected_columns)
# A tibble: 6 × 5
  `Sale Price` bedrooms bath_full_count square_feet_total_living year_built
       <dbl>    <dbl>         <dbl>                 <dbl>     <dbl>
1      698000       4             2                  2810      2003
2      649990       4             2                  2880      2006
3      572500       4             1                  2770      1987
4      420000       3             1                  1620      1968
5      369900       3             1                  1440      1980
6      184667       4             2                  4160      2005
> head(sorted_houses)
```

```
# A tibble: 6 × 24
  `Sale Date`        `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
  <dttm>                   <dbl>       <dbl>          <dbl> <chr>       <chr>    <chr>
1 2010-03-02 00:00:00     4400000           1              3 35 45       R1       12025 154…
2 2010-03-02 00:00:00     4400000           1              3 35 45       R1       12053 154…
3 2011-11-17 00:00:00     4380542           1             22 11 45       R1       17137 NE …
4 2011-11-17 00:00:00     4380542           1             22 11 45       R1       11818 171…
5 2011-11-17 00:00:00     4380542           1             22 11 45       R1       17011 NE …
6 2011-11-17 00:00:00     4380542           1             22 11 45       R1       16943 NE …
# ℹ 17 more variables: zip5 <dbl>, ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>, year_built <dbl>,
#   year_renovated <dbl>, current_zoning <chr>, sq_ft_lot <dbl>, prop_type <chr>,
#   present_use <dbl>
> head(housing_extended)
  Sale Date Sale Price sale_reason sale_instrument sale_warning sitetype         addr_full
1 2006-01-03     698000           1              3       <NA>       R1  17021 NE 113TH CT
2 2006-01-03     649990           1              3       <NA>       R1  11927 178TH PL NE
3 2006-01-03     572500           1              3       <NA>       R1 13315 174TH AVE NE
4 2006-01-03     420000           1              3       <NA>       R1  3303 178TH AVE NE
5 2006-01-03     369900           1              3         15       R1  16126 NE 108TH CT
6 2006-01-03     184667           1             15       18 51       R1   8101 229TH DR NE
  zip5 ctyname postalctyn       lon      lat building_grade square_feet_total_living bedrooms
1 98052 REDMOND    REDMOND -122.1124 47.70139             9                    2810        4
```

```
2 98052 REDMOND    REDMOND -122.1022 47.70731          9              2880      4
3 98052    <NA>    REDMOND -122.1085 47.71986          8              2770      4
4 98052 REDMOND    REDMOND -122.1037 47.63914          8              1620      3
5 98052 REDMOND    REDMOND -122.1242 47.69748          7              1440      3
6 98053    <NA>    REDMOND -122.0341 47.67545          7              4160      4
  bath_full_count bath_half_count bath_3qtr_count year_built year_renovated current_zoning
1               2               1               0       2003            0           R4
2               2               0               1       2006            0           R4
3               1               1               1       1987            0           R6
4               1               0               1       1968            0           R4
5               1               0               1       1980            0           R6
6               2               1               1       2005            0        URPSO
  sq_ft_lot prop_type present_use price_per_sqft
1      6635         R           2      248.39858
2      5570         R           2      225.69097
3      8444         R           2      206.67870
4      9600         R           2      259.25926
5      7526         R           2      256.87500
6      7280         R           2       44.39111
> nrow(housing_combined)
[1] 12875
> head(concatenated_address)
[1] "17021-NE-113TH-CT"  "11927-178TH-PL-NE"  "13315-174TH-AVE-NE" "3303-178TH-AVE-NE"
[5] "16126-NE-108TH-CT"  "8101-229TH-DR-NE"
```

>