5.2Exercises.R

Rahul Rajeev

2023-01-10

```
# Assignment: Experimenting with Plyr on the Housing Dataset
# Name: Rajeev, Rahul
# Date: 2023-01-03

# importing libraries
library(ggplot2)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(pastecs)
```

```
##
## Attaching package: 'pastecs'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
library(scales)
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:scales':
##
##     discard
```

```
## The following object is masked from 'package:plyr':
##
```

```
##     compact
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:pastecs':
##
##     extract
```

```
install.packages('tidyr')
```

```
## Warning: package 'tidyr' is in use and will not be installed
```

```
theme_set(theme_minimal())

# loading housing dataset
setwd("C:/Users/rahul/Documents/Bellevue/DSC 520")

# reading in xlsx file, then transferring the format into a data frame
housing_xl <- read_excel("data/week-7-housing.xlsx")
housing_df <- data.frame(housing_xl)

# just checking whether the data frame loaded properly
head(housing_df)
```

```
##    Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1 2006-01-03     698000           1               3         <NA>       R1
## 2 2006-01-03     649990           1               3         <NA>       R1
## 3 2006-01-03     572500           1               3         <NA>       R1
## 4 2006-01-03     420000           1               3         <NA>       R1
## 5 2006-01-03     369900           1               3           15       R1
## 6 2006-01-03     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052    <NA>    REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053    <NA>    REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003              0             R4      6635         R
## 2               1       2006              0             R4      5570         R
## 3               1       1987              0             R6      8444         R
## 4               1       1968              0             R4      9600         R
## 5               1       1980              0             R6      7526         R
## 6               1       2005              0          URPSO      7280         R
##   present_use
## 1           2
```

```
## 2           2
## 3           2
## 4           2
## 5           2
## 6           2
```

```
# 1.Use the 6 different operations to analyze/transform
# - GroupBy, Summarize, Mutate, Filter, Select, and Arrange -
# - understand your dataset in more detail

# Mutate Example: created a price/sq foot living, price/sq foot lot, total baths
new_housing_df <- housing_df %>%
  mutate(price_per_sqft_living = Sale.Price/square_feet_total_living,
         price_per_sqft_lot = Sale.Price/sq_ft_lot,
         total_bath_count = bath_full_count + bath_half_count + bath_3qtr_count)

head(new_housing_df)
```

```
##    Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1 2006-01-03     698000           1               3         <NA>       R1
## 2 2006-01-03     649990           1               3         <NA>       R1
## 3 2006-01-03     572500           1               3         <NA>       R1
## 4 2006-01-03     420000           1               3         <NA>       R1
## 5 2006-01-03     369900           1               3           15       R1
## 6 2006-01-03     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052    <NA>    REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053    <NA>    REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003             0             R4      6635         R
## 2               1       2006             0             R4      5570         R
## 3               1       1987             0             R6      8444         R
## 4               1       1968             0             R4      9600         R
## 5               1       1980             0             R6      7526         R
## 6               1       2005             0          URPSO      7280         R
##   present_use price_per_sqft_living price_per_sqft_lot total_bath_count
## 1           2             248.39858          105.19970               3
## 2           2             225.69097          116.69479               3
## 3           2             206.67870           67.79962               3
## 4           2             259.25926           43.75000               2
## 5           2             256.87500           49.14961               2
## 6           2              44.39111           25.36635               4
```

```r
# Group By and Summarize Example:
new_housing_df %>% group_by(zip5) %>%
  summarise(sqft_mean=round(mean(square_feet_total_living), 2),
            sqft_sd=round(sd(square_feet_total_living), 2),
            price_per_sqft_mean=round(mean(price_per_sqft_living), 2),
            price_per_sqft_sd=round(sd(price_per_sqft_living), 2))
```

```
## # A tibble: 4 x 5
##    zip5 sqft_mean sqft_sd price_per_sqft_mean price_per_sqft_sd
##   <dbl>     <dbl>   <dbl>               <dbl>             <dbl>
## 1 98052     2499.    879.                267.             144.
## 2 98053     2580.   1113.                283.             242.
## 3 98059     4360      NA                 148.              NA
## 4 98074     3682.   1266.                260.              67.3
```

```r
# the data is returned as a tibble, which automatically accounts for 3 sig figs
# I'm not entirely sure why the sq_sd and price_per_sqft_sd does not return
# anything, I might have to ask about this

# Filter Example:
filtered_housing_df <- new_housing_df %>% filter(Sale.Price > 100000)
head(filtered_housing_df)
```

```
##    Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1 2006-01-03     698000           1               3         <NA>       R1
## 2 2006-01-03     649990           1               3         <NA>       R1
## 3 2006-01-03     572500           1               3         <NA>       R1
## 4 2006-01-03     420000           1               3         <NA>       R1
## 5 2006-01-03     369900           1               3           15       R1
## 6 2006-01-03     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052    <NA>    REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053    <NA>    REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003              0             R4      6635         R
## 2               1       2006              0             R4      5570         R
## 3               1       1987              0             R6      8444         R
## 4               1       1968              0             R4      9600         R
## 5               1       1980              0             R6      7526         R
## 6               1       2005              0          URPSO      7280         R
##   present_use price_per_sqft_living price_per_sqft_lot total_bath_count
## 1           2             248.39858          105.19970                3
## 2           2             225.69097          116.69479                3
## 3           2             206.67870           67.79962                3
```

```
## 4           2             259.25926           43.75000               2
## 5           2             256.87500           49.14961               2
## 6           2              44.39111           25.36635               4
```

```
apply(filtered_housing_df['Sale.Price'], MARGIN=2, FUN=min)
```

```
## Sale.Price
##     102500
```

```
# filtered out house prices with less than 100,000 to avoid outliers in the
# data, and used an apply statement to find the minimum after


# Select Example:
selected_housing_df <- filtered_housing_df %>% select(Sale.Price,
                                                      square_feet_total_living,
                                                      bedrooms,
                                                      total_bath_count)
head(selected_housing_df)
```

```
##   Sale.Price square_feet_total_living bedrooms total_bath_count
## 1     698000                     2810        4               3
## 2     649990                     2880        4               3
## 3     572500                     2770        4               3
## 4     420000                     1620        3               2
## 5     369900                     1440        3               2
## 6     184667                     4160        4               4
```

```
# selected sale price, living square feet, bedrooms, and total bath count
# as a new data frame

# Arrange Example:
arranged_housing_df <- filtered_housing_df %>% arrange(Sale.Price)
head(arranged_housing_df)
```

```
##    Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1 2013-10-21     102500          18              24     12 51 52       R1
## 2 2015-07-22     106092          14              15     18 51 52       R1
## 3 2012-06-08     108147           4              18        13 31       R1
## 4 2015-11-30     110000           1               3         <NA>       R1
## 5 2012-12-11     114946          18              15     12 18 22       R1
## 6 2008-01-18     115731           1              15     18 51 52       R1
##                       addr_full  zip5 ctyname postalctyn       lon      lat
## 1              9719 159TH PL NE 98052 REDMOND    REDMOND -122.1274 47.68709
## 2              6640 211TH PL NE 98053    <NA>    REDMOND -122.0574 47.66575
## 3         3019 E AMES LAKE DR NE 98053    <NA>    REDMOND -121.9600 47.63471
## 4 25459 NE REDMOND FALL CITY RD 98053    <NA>    REDMOND -121.9990 47.63151
## 5             15930 NE 106TH CT 98052 REDMOND    REDMOND -122.1268 47.69456
## 6         3049 E AMES LAKE DR NE 98053    <NA>    REDMOND -121.9595 47.63552
##   building_grade square_feet_total_living bedrooms bath_full_count
## 1              8                     2140        4               3
## 2              6                      940        3               1
## 3              2                      480        1               0
## 4              5                     1560        4               1
## 5              7                     2260        3               1
## 6              7                     1690        2               1
```

```
##   bath_half_count bath_3qtr_count year_built year_renovated current_zoning
## 1               0               0       1977              0             R5
## 2               0               0       1970              0            RA5
## 3               0               0       1945              0            RA5
## 4               0               0       1935              0           RA10
## 5               1               1       1978              0             R6
## 6               0               0       1976              0            RA5
##   sq_ft_lot prop_type present_use price_per_sqft_living price_per_sqft_lot
## 1     10500         R           2              47.89720          9.7619048
## 2     12600         R           2             112.86383          8.4200000
## 3     35808         R           2             225.30625          3.0201910
## 4    122403         R           2              70.51282          0.8986708
## 5      7350         R           2              50.86106         15.6389116
## 6     18695         R           2              68.47988          6.1904787
##   total_bath_count
## 1                3
## 2                1
## 3                0
## 4                1
## 5                3
## 6                1
```

```
# arrange the data set with Sale Price ascending, with the previous filter
# of prices greater than $100,000.

# 2. Using the purrr package - perform 2 functions on your dataset.
# You could use zip_n, keep, discard, compact, etc.

# a. nest()
n_filtered_df <- filtered_housing_df %>% group_by(zip5) %>% nest()
head(n_filtered_df)
```

```
## # A tibble: 4 x 2
## # Groups:   zip5 [4]
##    zip5 data
##   <dbl> <list>
## 1 98052 <tibble [7,410 x 26]>
## 2 98053 <tibble [5,302 x 26]>
## 3 98074 <tibble [73 x 26]>
## 4 98059 <tibble [1 x 26]>
```

```
# created nested groups of housing into cells as data frames

# b. map()
m_filtered_df <- n_filtered_df %>% mutate(n=map(data, dim))
head(m_filtered_df)
```

```
## # A tibble: 4 x 3
## # Groups:   zip5 [4]
##    zip5 data                  n
##   <dbl> <list>                <list>
## 1 98052 <tibble [7,410 x 26]> <int [2]>
## 2 98053 <tibble [5,302 x 26]> <int [2]>
## 3 98074 <tibble [73 x 26]>    <int [2]>
## 4 98059 <tibble [1 x 26]>     <int [2]>
```

```r
# using the map function to apply the dim function to each of the nested groups

# 3. use the cbind and rbind function on your dataset
saleprice_df = housing_df$Sale.Price
sq_ft_living_df = housing_df$square_feet_total_living

combined_column_df = cbind(saleprice_df, sq_ft_living_df)
head(combined_column_df)
```

```
##      saleprice_df sq_ft_living_df
## [1,]       698000            2810
## [2,]       649990            2880
## [3,]       572500            2770
## [4,]       420000            1620
## [5,]       369900            1440
## [6,]       184667            4160
```

```r
# combined columns of sale price and square foot living

houses_05_df = housing_df[housing_df$year_built == 2005,]
houses_07_df = housing_df[housing_df$year_built == 2007,]

combined_row_df <- rbind(houses_05_df, houses_07_df)
head(combined_row_df)
```

```
##     Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 6  2006-01-03     184667           1              15        18 51       R1
## 12 2006-01-04     526787           1               3         <NA>       R1
## 16 2006-01-05     507950           1               3         <NA>       R1
## 17 2006-01-06     765000           1               3         <NA>       R1
## 18 2006-01-06     589950           1               3         <NA>       R1
## 21 2006-01-10     513262           1               3         <NA>       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 6   8101 229TH DR NE 98053    <NA>    REDMOND -122.0341 47.67545              7
## 12  7858 148TH CT NE 98052 REDMOND    REDMOND -122.1425 47.67407              8
## 16  7850 148TH CT NE 98052 REDMOND    REDMOND -122.1425 47.67390              8
## 17  8944 237TH PL NE 98053    <NA>    REDMOND -122.0230 47.68150              9
## 18 11922 173RD PL NE 98052 REDMOND    REDMOND -122.1086 47.70678              8
## 21 11807 242ND PL NE 98053    <NA>    REDMOND -122.0162 47.70323              8
##    square_feet_total_living bedrooms bath_full_count bath_half_count
## 6                      4160        4               2               1
## 12                     2480        3               2               1
## 16                     2480        3               2               1
## 17                     4000        4               2               1
## 18                     2570        4               2               1
## 21                     1930        2               2               0
##    bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 6                1       2005             0          URPSO      7280         R
## 12               0       2005             0             R5      2647         R
## 16               0       2005             0             R5      3099         R
## 17               1       2005             0          URPSO      7611         R
## 18               0       2005             0             R4      4737         R
## 21               0       2005             0          URPSO      4958         R
##    present_use
## 6            2
```

```
## 12           2
## 16           2
## 17           2
## 18           2
## 21           2
```

```
tail(combined_row_df)
```

```
##        Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 12755 2016-11-03     559000           1               3         <NA>       R1
## 12757 2016-11-04     431952           1               3         <NA>       R1
## 12765 2016-11-08    1000000           1               3         <NA>       R1
## 12770 2016-11-08     337512           1               3         <NA>       R1
## 12774 2016-11-09     875000           1               3         <NA>       R1
## 12788 2016-11-14     885000           1               3         <NA>       R1
##                       addr_full  zip5 ctyname postalctyn       lon      lat
## 12755   23554 NE TWINBERRY WAY 98053    <NA>    REDMOND -122.0245 47.71941
## 12757        13738 231ST LN NE 98053    <NA>    REDMOND -122.0295 47.72221
## 12765        11667 168TH CT NE 98052 REDMOND    REDMOND -122.1160 47.70443
## 12770 13417 ADAIR CREEK WAY NE 98053    <NA>    REDMOND -122.0219 47.71849
## 12774        13018 243RD PL NE 98053    <NA>    REDMOND -122.0144 47.71548
## 12788        11826 179TH PL NE 98052 REDMOND    REDMOND -122.1002 47.70674
##       building_grade square_feet_total_living bedrooms bath_full_count
## 12755              8                     1810        3               2
## 12757              8                     1640        2               2
## 12765             10                     3070        4               2
## 12770              8                     1300        2               2
## 12774              8                     2500        3               2
## 12788              9                     2830        4               2
##       bath_half_count bath_3qtr_count year_built year_renovated current_zoning
## 12755               1               0       2007              0          URPSO
## 12757               0               0       2007              0          URPSO
## 12765               1               0       2007              0             R4
## 12770               0               0       2007              0          URPSO
## 12774               1               0       2007              0          URPSO
## 12788               0               1       2007              0             R4
##       sq_ft_lot prop_type present_use
## 12755      3844         R          29
## 12757      3657         R          29
## 12765      6615         R           2
## 12770      3673         R          29
## 12774      6631         R           2
## 12788      5305         R           2
```

```
# combined rows of houses built in 2005 and 2007

# 4. split a string then concatenate the results back together
shopping_list <- 'salad, chicken, corn, soap'
split_str <- strsplit(shopping_list, split = ', ')[[1]]
split_str
```

```
## [1] "salad"   "chicken" "corn"    "soap"
```

```
combined_str <- paste(split_str, collapse = ', ')
combined_str
```

```
## [1] "salad, chicken, corn, soap"
```

```
# split and combined a shopping list using ','
```