```
> # Assignment: Week 5 Exercise 5.2
> # Name: Sharma, Dipika
> # Date: 2020-04-18
>
> " a.  Using the dplyr package, use the 6 different operations to analyze/transform the
+       data - GroupBy, Summarize, Mutate, Filter, Select, and Arrange –
+       Remember this isn't just modifying data, you are learning about your data also –
+       so play around and start to understand your dataset in more detail"
[1] " a.  Using the dplyr package, use the 6 different operations to analyze/transform the \n
data - GroupBy, Summarize, Mutate, Filter, Select, and Arrange – \n      Remember this isn't just
modifying data, you are learning about your data also – \n      so play around and start to
understand your dataset in more detail"
> # Ans
> #install.packages("dplyr")
> #install.packages("readxl")
> library("readxl")
> housing_df <- read_excel("/Users/dipikasharma/R_Projects/DSC520/data/week-7-
housing.xlsx")
> housing_df
# A tibble: 12,865 x 24
   `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
zip5 ctyname
   <dttm>                     <dbl>       <dbl>          <dbl> <chr>        <chr>    <chr>
<dbl> <chr>
 1 2006-01-03 00:00:00       698000           1              3 NA           R1       17021 NE 113TH CT  98052
REDMOND
 2 2006-01-03 00:00:00       649990           1              3 NA           R1       11927 178TH PL NE  98052
REDMOND
 3 2006-01-03 00:00:00       572500           1              3 NA           R1       13315 174TH AVE NE 98052
NA
 4 2006-01-03 00:00:00       420000           1              3 NA           R1       3303 178TH AVE NE  98052
REDMOND
 5 2006-01-03 00:00:00       369900           1              3 15           R1       16126 NE 108TH CT  98052
REDMOND
 6 2006-01-03 00:00:00       184667           1             15 18 51        R1       8101 229TH DR NE   98053
NA
 7 2006-01-04 00:00:00      1050000           1              3 NA           R1       21634 NE 87TH PL   98053
NA
 8 2006-01-04 00:00:00       875000           1              3 NA           R1       21404 NE 67TH ST   98053
NA
 9 2006-01-04 00:00:00       660000           1              3 NA           R1       7525 238TH AVE NE  98053
NA
10 2006-01-04 00:00:00       650000           1              3 NA           R1       17703 NE 26TH ST   98052
REDMOND
```

```
# … with 12,855 more rows, and 15 more variables: 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>
> library(dplyr)
> select_df <- select(housing_df, `Sale Price`, sale_reason, sale_instrument)
> select_df
# A tibble: 12,865 x 3
   `Sale Price` sale_reason sale_instrument
          <dbl>       <dbl>          <dbl>
 1       698000           1              3
 2       649990           1              3
 3       572500           1              3
 4       420000           1              3
 5       369900           1              3
 6       184667           1             15
 7      1050000           1              3
 8       875000           1              3
 9       660000           1              3
10       650000           1              3
# … with 12,855 more rows
>
> filter_df <- filter(housing_df, sale_reason == 1)
> filter_df
# A tibble: 12,202 x 24
   `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning sitetype addr_full
zip5 ctyname
   <dttm>                     <dbl>       <dbl>          <dbl> <chr>        <chr>    <chr>
<dbl> <chr>
 1 2006-01-03 00:00:00       698000           1              3 NA           R1       17021 NE 113TH CT  98052
REDMOND
 2 2006-01-03 00:00:00       649990           1              3 NA           R1       11927 178TH PL NE  98052
REDMOND
 3 2006-01-03 00:00:00       572500           1              3 NA           R1       13315 174TH AVE NE 98052
NA
 4 2006-01-03 00:00:00       420000           1              3 NA           R1       3303 178TH AVE NE  98052
REDMOND
 5 2006-01-03 00:00:00       369900           1              3 15           R1       16126 NE 108TH CT  98052
REDMOND
 6 2006-01-03 00:00:00       184667           1             15 18 51        R1       8101 229TH DR NE   98053
NA
```

```
 7 2006-01-04 00:00:00      1050000          1            3 NA         R1      21634 NE 87TH PL   98053
NA
 8 2006-01-04 00:00:00       875000          1            3 NA         R1      21404 NE 67TH ST   98053
NA
 9 2006-01-04 00:00:00       660000          1            3 NA         R1      7525 238TH AVE NE  98053
NA
10 2006-01-04 00:00:00       650000          1            3 NA         R1      17703 NE 26TH ST   98052
REDMOND
# … with 12,192 more rows, and 15 more variables: 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>
>
> sfilter_df <- housing_df %>% filter(sale_reason == 1)  %>%  select(`Sale Price`, sale_reason,
sale_instrument)
> sfilter_df
# A tibble: 12,202 x 3
   `Sale Price` sale_reason sale_instrument
          <dbl>       <dbl>          <dbl>
 1       698000           1              3
 2       649990           1              3
 3       572500           1              3
 4       420000           1              3
 5       369900           1              3
 6       184667           1             15
 7      1050000           1              3
 8       875000           1              3
 9       660000           1              3
10       650000           1              3
# … with 12,192 more rows
>
> mutate_df <- sfilter_df %>% mutate(Saleprice_divident = (`Sale Price`*4)/100)
> mutate_df
# A tibble: 12,202 x 4
   `Sale Price` sale_reason sale_instrument Saleprice_divident
          <dbl>       <dbl>          <dbl>             <dbl>
 1       698000           1              3             27920
 2       649990           1              3             26000.
 3       572500           1              3             22900
 4       420000           1              3             16800
 5       369900           1              3             14796
```

```
 6      184667         1            15            7387.
 7     1050000         1             3           42000
 8      875000         1             3           35000
 9      660000         1             3           26400
10      650000         1             3           26000
# … with 12,192 more rows
>
> sgroupby_df <- housing_df %>% group_by(sale_reason) %>% summarize(Saleprice =
sum(`Sale Price`, na.rm = TRUE))
> sgroupby_df
# A tibble: 17 x 2
   sale_reason  Saleprice
         <dbl>      <dbl>
 1           0     815290
 2           1 8202004677
 3           2     203904
 4           3    1220217
 5           4   66027084
 6           6     428900
 7           7    2552337
 8           8   70768273
 9          10    5396595
10          11     360000
11          12   36982527
12          13     828400
13          14   10935318
14          16    1717792
15          17     530000
16          18   99094351
17          19     525484
>
> arrange_df <- sgroupby_df %>% arrange(Saleprice)
> arrange_df
# A tibble: 17 x 2
   sale_reason  Saleprice
         <dbl>      <dbl>
 1           2     203904
 2          11     360000
 3           6     428900
 4          19     525484
 5          17     530000
 6           0     815290
 7          13     828400
 8           3    1220217
```

```
 9        16    1717792
10         7    2552337
11        10    5396595
12        14   10935318
13        12   36982527
14         4   66027084
15         8   70768273
16        18   99094351
17         1 8202004677
>
> "Using the purrr package – perform 2 functions on your dataset.
+ You could use zip_n, keep, discard, compact, etc."
[1] "Using the purrr package – perform 2 functions on your dataset.  \nYou could use zip_n,
keep, discard, compact, etc."
> #install.packages("purrr")
> library(purrr)
>
> square <- function(x){
+   return(x*x)
+ }
> map(sgroupby_df$sale_reason, square)
[[1]]
[1] 0

[[2]]
[1] 1

[[3]]
[1] 4

[[4]]
[1] 9

[[5]]
[1] 16

[[6]]
[1] 36

[[7]]
[1] 49

[[8]]
[1] 64
```

```
[[9]]
[1] 100

[[10]]
[1] 121

[[11]]
[1] 144

[[12]]
[1] 169

[[13]]
[1] 196

[[14]]
[1] 256

[[15]]
[1] 289

[[16]]
[1] 324

[[17]]
[1] 361

>
> library(purrr)
>
>
> to_loss <- function(x, y){
+   return(x- (x*y)/100)
+ }
> map_df <- map2(sgroupby_df$Saleprice, 5, to_loss)
> map_df
[[1]]
[1] 774525.5

[[2]]
[1] 7791904443

[[3]]
```

[1] 193708.8

[[4]]
[1] 1159206

[[5]]
[1] 62725730

[[6]]
[1] 407455

[[7]]
[1] 2424720

[[8]]
[1] 67229859

[[9]]
[1] 5126765

[[10]]
[1] 342000

[[11]]
[1] 35133401

[[12]]
[1] 786980

[[13]]
[1] 10388552

[[14]]
[1] 1631902

[[15]]
[1] 503500

[[16]]
[1] 94139633

[[17]]
[1] 499209.8

```
>
> map_df %>% keep(map_df>20000000)
[[1]]
[1] 7791904443

[[2]]
[1] 62725730

[[3]]
[1] 67229859

[[4]]
[1] 35133401

[[5]]
[1] 94139633

>
> map_df %>% discard(map_df>200000)
[[1]]
[1] 193708.8

>
> testt2 <- sgroupby_df$sale_reason %>% keep(sgroupby_df$sale_reason>10)
> testt2
[1] 11 12 13 14 16 17 18 19
> #c. "Use the cbind and rbind function on your dataset"
> #Ans
> library(dplyr)
> ID <- c(101:117)
> new_df <- cbind(sgroupby_df, ID)
> new_df
   sale_reason  Saleprice  ID
1           0     815290 101
2           1 8202004677 102
3           2     203904 103
4           3    1220217 104
5           4   66027084 105
6           6     428900 106
7           7    2552337 107
8           8   70768273 108
9          10    5396595 109
10         11     360000 110
11         12   36982527 111
```

```
12         13     828400 112
13         14   10935318 113
14         16    1717792 114
15         17     530000 115
16         18   99094351 116
17         19     525484 117
> part1_df <- new_df %>% select(ID, Saleprice) %>% filter(ID <= 105)
> part1_df
  ID  Saleprice
1 101     815290
2 102 8202004677
3 103     203904
4 104    1220217
5 105   66027084
> part2_df <- new_df %>% select(ID, Saleprice) %>% filter(ID > 105)
> part2_df
   ID Saleprice
1  106    428900
2  107   2552337
3  108  70768273
4  109   5396595
5  110    360000
6  111  36982527
7  112    828400
8  113  10935318
9  114   1717792
10 115    530000
11 116  99094351
12 117    525484
> rbind(part1_df, part2_df)
   ID  Saleprice
1  101     815290
2  102 8202004677
3  103     203904
4  104    1220217
5  105   66027084
6  106     428900
7  107    2552337
8  108   70768273
9  109    5396595
10 110     360000
11 111   36982527
12 112     828400
13 113   10935318
```

```
14 114    1717792
15 115     530000
16 116   99094351
17 117     525484
>
> "Split a string, then concatenate the results back together"
> library(tidyverse)
> library(dplyr)
> library("readxl")
> nhousing_df <- read_excel("/Users/dipikasharma/R_Projects/DSC520/data/week-7-
housing.xlsx")
>
> nefilter_df <- filter(nhousing_df, sale_reason == 19)
> #nefilter_df
>
> nefilter_df$location <- paste(nefilter_df$lon, nefilter_df$lat, sep = " - ")
> nefilter_df$TotalYears <- paste(nefilter_df$year_built, nefilter_df$year_renovated, sep = " - ")
> #nefilter_df
> select(nefilter_df, `Sale Price`, sale_reason, location, lon, lat, TotalYears, year_built,
year_renovated)
# A tibble: 3 x 8
  `Sale Price` sale_reason location                lon   lat TotalYears year_built year_renovated
         <dbl>      <dbl> <chr>                   <dbl> <dbl> <chr>          <dbl>         <dbl>
1       155768         19 -122.14998827 - 47.67453015 -122.  47.7 1983 - 0        1983             0
2       183102         19 -122.04331714 - 47.64826463 -122.  47.6 1941 - 0        1941             0
3       186614         19 -122.11132013 - 47.68136635 -122.  47.7 1968 - 0        1968             0
>
>
> #install.packages("tidyr")
> library(tidyr)
>
> new_housingdf <- nefilter_df %>% separate(TotalYears, c("BeginYear", "EndYear"), " - ")
> select(new_housingdf, `Sale Price`, sale_reason, location, lon, lat, BeginYear, EndYear,
year_built, year_renovated)
# A tibble: 3 x 9
  `Sale Price` sale_reason location                lon   lat BeginYear EndYear year_built
year_renovated
         <dbl>      <dbl> <chr>                   <dbl> <dbl> <chr>     <chr>       <dbl>
         <dbl>
1       155768         19 -122.14998827 - 47.67453015 -122.  47.7 1983      0           1983
             0
2       183102         19 -122.04331714 - 47.64826463 -122.  47.6 1941      0           1941
             0
3       186614         19 -122.11132013 - 47.68136635 -122.  47.7 1968      0           1968
             0
>
```