Chapter 12 - Faster Group Manipulation with dplyr

Michael Beebe

---

1. dplyr

From the textbook data files, load the the housing.csv data

```
housing <- read.csv("../../Data/housing.csv")
head(housing, 5)
```

```
##   Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
## 1    FINANCIAL          R9-CONDOMINIUM          42       1920      36500
## 2    FINANCIAL          R4-CONDOMINIUM          78       1985     126420
## 3    FINANCIAL          RR-CONDOMINIUM         500         NA     554174
## 4    FINANCIAL          R4-CONDOMINIUM         282       1930     249076
## 5      TRIBECA          R4-CONDOMINIUM         239       1985     219495
##   Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1                1332615                 36.51            342005
## 2                6633257                 52.47           1762295
## 3               17310000                 31.24           3543000
## 4               11776313                 47.28           2784670
## 5               10004582                 45.58           2783197
##   Expense.per.SqFt Net.Operating.Income Full.Market.Value Market.Value.per.SqFt
## 1             9.37               990610           7300000                200.00
## 2            13.94              4870962          30690000                242.76
## 3             6.39             13767000          90970000                164.15
## 4            11.18              8991643          67556006                271.23
## 5            12.68              7221385          54320996                247.48
##        Boro
## 1 Manhattan
## 2 Manhattan
## 3 Manhattan
## 4 Manhattan
## 5 Manhattan
```

2. Select

Using the housing data and basic select, display the neighborhood and boro

```
neighborhood_boro <-
  housing %>%
    select(Neighborhood, Boro)

head(neighborhood_boro, 10)
```

```
##    Neighborhood      Boro
## 1     FINANCIAL Manhattan
## 2     FINANCIAL Manhattan
## 3     FINANCIAL Manhattan
## 4     FINANCIAL Manhattan
## 5       TRIBECA Manhattan
## 6       TRIBECA Manhattan
## 7       TRIBECA Manhattan
## 8       TRIBECA Manhattan
## 9       TRIBECA Manhattan
## 10      TRIBECA Manhattan
```

Select neighborhood and year built by using a vector passed to the select function

```
neighborhood_year <-
  housing %>%
    select(c(Neighborhood, Year.Built))

head(neighborhood_year, 10)
```

```
##    Neighborhood Year.Built
## 1     FINANCIAL       1920
## 2     FINANCIAL       1985
## 3     FINANCIAL         NA
## 4     FINANCIAL       1930
## 5       TRIBECA       1985
## 6       TRIBECA       1986
## 7       TRIBECA       1985
## 8       TRIBECA       1986
## 9       TRIBECA       1987
## 10      TRIBECA       1985
```

```
# I am not getting a warning
```

Select the neighborhood and square footage using the column positions rather than names

```
neighborhood_sqft <-
  housing %>%
    select(1, 5)

head(neighborhood_sqft, 10)
```

```
##    Neighborhood Gross.SqFt
## 1     FINANCIAL      36500
## 2     FINANCIAL     126420
## 3     FINANCIAL     554174
## 4     FINANCIAL     249076
## 5       TRIBECA     219495
## 6       TRIBECA     139719
## 7       TRIBECA     105000
## 8       TRIBECA      87479
## 9       TRIBECA     255845
## 10      TRIBECA     106129
```

Select the neighborhood and the columns that start with E

```
neighborhood_e_cols <-
  housing %>%
    select(Neighborhood, starts_with("e"))

head(neighborhood_e_cols, 10)
```

```
##    Neighborhood Estimated.Gross.Income Estimated.Expense Expense.per.SqFt
## 1     FINANCIAL                1332615            342005             9.37
## 2     FINANCIAL                6633257           1762295            13.94
## 3     FINANCIAL               17310000           3543000             6.39
## 4     FINANCIAL               11776313           2784670            11.18
## 5       TRIBECA               10004582           2783197            12.68
## 6       TRIBECA                5127687           1497788            10.72
## 7       TRIBECA                4365900           1273650            12.13
## 8       TRIBECA                3637377           1061120            12.13
## 9       TRIBECA               11246946           2440761             9.54
## 10      TRIBECA                4115683           1231096            11.60
```

Select the neighborhood and the columns that end with t

```
neighborhood_t_cols <-
  housing %>%
    select(Neighborhood, ends_with("t"))

head(neighborhood_t_cols, 10)
```

```
##    Neighborhood Year.Built Gross.SqFt Gross.Income.per.SqFt Expense.per.SqFt
## 1     FINANCIAL       1920      36500                 36.51             9.37
## 2     FINANCIAL       1985     126420                 52.47            13.94
## 3     FINANCIAL         NA     554174                 31.24             6.39
## 4     FINANCIAL       1930     249076                 47.28            11.18
## 5       TRIBECA       1985     219495                 45.58            12.68
## 6       TRIBECA       1986     139719                 36.70            10.72
## 7       TRIBECA       1985     105000                 41.58            12.13
## 8       TRIBECA       1986      87479                 41.58            12.13
## 9       TRIBECA       1987     255845                 43.96             9.54
## 10      TRIBECA       1985     106129                 38.78            11.60
##    Market.Value.per.SqFt
## 1                 200.00
## 2                 242.76
## 3                 164.15
## 4                 271.23
## 5                 247.48
## 6                 191.37
## 7                 211.53
## 8                 222.33
## 9                 259.21
## 10                205.62
```

Select the columns that match the pattern ".Income". The word Income should be at the end of the string.

```
neighbord_income <-
  housing %>%
    select(Neighborhood, ends_with(".income"))

head(neighbord_income, 10)
```

```
##    Neighborhood Estimated.Gross.Income Net.Operating.Income
## 1     FINANCIAL                1332615               990610
## 2     FINANCIAL                6633257              4870962
## 3     FINANCIAL               17310000             13767000
## 4     FINANCIAL               11776313              8991643
## 5       TRIBECA               10004582              7221385
## 6       TRIBECA                5127687              3629899
## 7       TRIBECA                4365900              3092250
## 8       TRIBECA                3637377              2576257
## 9       TRIBECA               11246946              8806185
## 10      TRIBECA                4115683              2884587
```

3. Filter

Using the housing data, filter the data based on construction built after or equal to the year 2009

```
after_2009 <-
  housing %>%
    filter(Year.Built >= 2009)

head(after_2009, 5)
```

```
##       Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
## 1          CLINTON          RR-CONDOMINIUM         222       2009     620611
## 2      HARLEM-EAST          RR-CONDOMINIUM          55       2009      43516
## 3   HARLEM-CENTRAL          R4-CONDOMINIUM          56       2009      51845
## 4   HARLEM-CENTRAL          RR-CONDOMINIUM          39       2009      42760
## 5 COBBLE HILL-WEST          R4-CONDOMINIUM           3       2009      61991
##   Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1               23285325                 37.52           6845339
## 2                1253696                 28.81            274586
## 3                1500000                 28.93            460000
## 4                1006143                 23.53            362605
## 5                 991236                 15.99            346933
##   Expense.per.SqFt Net.Operating.Income Full.Market.Value Market.Value.per.SqFt
## 1            11.03             16439986         102711025                165.50
## 2             6.31               979110           1443453                 33.17
## 3             8.87              1040000           7785000                150.16
## 4             8.48               643538           2338500                 54.69
## 5             5.60               644303           4361043                 70.35
##        Boro
## 1 Manhattan
## 2 Manhattan
## 3 Manhattan
## 4 Manhattan
## 5  Brooklyn
```

Issue a select and store the Neighborhood, Year Built, and Boro in a variable

```
neighborhood_yb_boro <-
  housing %>%
    select(Neighborhood, Year.Built, Boro)

head(neighborhood_yb_boro, 5)
```

```
##   Neighborhood Year.Built      Boro
## 1    FINANCIAL       1920 Manhattan
## 2    FINANCIAL       1985 Manhattan
## 3    FINANCIAL         NA Manhattan
## 4    FINANCIAL       1930 Manhattan
## 5      TRIBECA       1985 Manhattan
```

Filter on Boro in the Bronx or Brooklyn

```
bronx_brooklyn <-
  neighborhood_yb_boro %>%
    filter(Boro == "Bronx" | Boro == "Brooklyn")

head(bronx_brooklyn, 5)
```

```
##            Neighborhood Year.Built     Boro
## 1 DOWNTOWN-FULTON FERRY       1913 Brooklyn
## 2 DOWNTOWN-FULTON FERRY       2001 Brooklyn
## 3 DOWNTOWN-FULTON FERRY       2006 Brooklyn
## 4 DOWNTOWN-FULTON FERRY       1904 Brooklyn
## 5 DOWNTOWN-FULTON FERRY       2007 Brooklyn
```

```
tail(bronx_brooklyn, 5)
```

```
##     Neighborhood Year.Built  Boro
## 782    RIVERDALE       1962 Bronx
## 783    RIVERDALE       2004 Bronx
## 784    RIVERDALE       2004 Bronx
## 785    RIVERDALE       1955 Bronx
## 786    RIVERDALE       1940 Bronx
```

Using the housing data, filter on Year Built > 1999 and Total Units > 200

```
yb_tu <-
  housing %>%
    filter(Year.Built > 1999 & Total.Units > 200)

head(yb_tu, 5)
```

```
##   Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
## 1      TRIBECA           R4-CONDOMINIUM         234       2006     431824
## 2      TRIBECA           R4-CONDOMINIUM         256       2006     434398
## 3    FINANCIAL           R4-CONDOMINIUM         320       2005     477747
## 4    FINANCIAL          R4-CONDOMINIUM         441       2003     348157
## 5      TRIBECA          R4-CONDOMINIUM         220       2006     535060
##   Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1               18041607                 41.78           5298480
## 2               19799861                 45.58           5508167
## 3               19864720                 41.58           5795071
## 4               14476368                 41.58           4223144
## 5               24200764                 45.23           5896361
##   Expense.per.SqFt Net.Operating.Income Full.Market.Value Market.Value.per.SqFt
## 1            12.27             12743127          89682996                207.68
## 2            12.68             14291694         100582005                231.54
## 3            12.13             14069649         106168339                222.23
## 4            12.13             10253224          77405999                222.33
## 5            11.02             18304403         136481149                255.08
##        Boro
## 1 Manhattan
## 2 Manhattan
## 3 Manhattan
## 4 Manhattan
## 5 Manhattan
```

Declare 2 variables - theCol and theVal
- Use filter and sprintf to filter the housing data on Neighborhoods in the Financial District
- Disregard warning message if you receive one

```
theCol <- housing$Neighborhood
theVal <- sprintf("%s", "FINANCIAL")

fincancial_district <-
  housing %>%
    filter(theCol == theVal)

head(fincancial_district, 5)
```

```
##   Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
## 1    FINANCIAL          R9-CONDOMINIUM          42       1920      36500
## 2    FINANCIAL          R4-CONDOMINIUM          78       1985     126420
## 3    FINANCIAL          RR-CONDOMINIUM         500         NA     554174
## 4    FINANCIAL          R4-CONDOMINIUM         282       1930     249076
## 5    FINANCIAL          R4-CONDOMINIUM          13       1920      37236
##   Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1                1332615                 36.51            342005
## 2                6633257                 52.47           1762295
## 3               17310000                 31.24           3543000
## 4               11776313                 47.28           2784670
## 5                1545666                 41.51            439012
##   Expense.per.SqFt Net.Operating.Income Full.Market.Value Market.Value.per.SqFt
## 1             9.37               990610           7300000                200.00
## 2            13.94              4870962          30690000                242.76
## 3             6.39             13767000          90970000                164.15
## 4            11.18              8991643          67556006                271.23
## 5            11.79              1106654           8355001                224.38
##        Boro
## 1 Manhattan
## 2 Manhattan
## 3 Manhattan
## 4 Manhattan
## 5 Manhattan
```

4. Slice

Using the housing data, take a slice of rows 10-20

```
housing %>%
  slice(10:20)
```

```
##    Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
## 1       TRIBECA          R4-CONDOMINIUM         121       1985     106129
## 2       TRIBECA          R4-CONDOMINIUM         154       1986     126008
## 3       TRIBECA          R4-CONDOMINIUM         546       1987     586224
## 4       TRIBECA          R4-CONDOMINIUM         182       1988     208281
## 5       TRIBECA          R4-CONDOMINIUM         293       1988     341489
## 6       TRIBECA          R4-CONDOMINIUM         117       2003     267723
## 7       TRIBECA          R4-CONDOMINIUM         234       2006     431824
## 8       TRIBECA          R4-CONDOMINIUM         304       1985     257848
## 9       TRIBECA          R4-CONDOMINIUM         256       2006     434398
## 10      TRIBECA          R4-CONDOMINIUM         174       1985     237725
## 11    FINANCIAL          R4-CONDOMINIUM          13       1920      37236
##    Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1                 4115683                 38.78           1231096
## 2                 5239413                 41.58           1528477
## 3                24375194                 41.58           7110897
## 4                 8077137                 38.78           2416060
## 5                13591262                 39.80           4309591
## 6                12202814                 45.58           3394728
## 7                18041607                 41.78           5298480
## 8                11752712                 45.58           3269513
## 9                19799861                 45.58           5508167
## 10               10051013                 42.28           2498490
## 11                1545666                 41.51            439012
##    Expense.per.SqFt Net.Operating.Income Full.Market.Value
## 1             11.60              2884587          21821999
## 2             12.13              3710936          28015990
## 3             12.13             17264297         130154990
## 4             11.60              5661077          42824998
## 5             12.62              9281671          70161999
## 6             12.68              8808086          62110366
## 7             12.27             12743127          89682996
## 8             12.68              8483199          63811996
## 9             12.68             14291694         100582005
## 10            10.51              7552523          57048005
## 11            11.79              1106654           8355001
##    Market.Value.per.SqFt      Boro
## 1                 205.62 Manhattan
## 2                 222.34 Manhattan
## 3                 222.02 Manhattan
## 4                 205.61 Manhattan
## 5                 205.46 Manhattan
## 6                 231.99 Manhattan
## 7                 207.68 Manhattan
## 8                 247.48 Manhattan
## 9                 231.54 Manhattan
## 10                239.97 Manhattan
## 11                224.38 Manhattan
```

Now, using the same data take a slice of rows 1-5 and the last 5 rows

```
housing %>%
  slice(c(1:5, (n()-4):n()))
```

```
##            Neighborhood Building.Classification Total.Units Year.Built
## 1             FINANCIAL          R9-CONDOMINIUM          42       1920
## 2             FINANCIAL          R4-CONDOMINIUM          78       1985
## 3             FINANCIAL          RR-CONDOMINIUM         500         NA
## 4             FINANCIAL          R4-CONDOMINIUM         282       1930
## 5               TRIBECA          R4-CONDOMINIUM         239       1985
## 6              ROSEBANK          R4-CONDOMINIUM          52         NA
## 7  ARROCHAR-SHORE ACRES          R4-CONDOMINIUM         102       1987
## 8            GRANT CITY          R4-CONDOMINIUM         100       1986
## 9            GRANT CITY          R4-CONDOMINIUM         159       1961
## 10          GREAT KILLS          R4-CONDOMINIUM          67       1965
##    Gross.SqFt Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense
## 1       36500                1332615                 36.51            342005
## 2      126420                6633257                 52.47           1762295
## 3      554174               17310000                 31.24           3543000
## 4      249076               11776313                 47.28           2784670
## 5      219495               10004582                 45.58           2783197
## 6       62391                 831672                 13.33            326305
## 7       90618                1274089                 14.06            637045
## 8       78903                1321625                 16.75            673832
## 9      166712                2343971                 14.06           1171985
## 10     108864                1298748                 11.93            722857
##    Expense.per.SqFt Net.Operating.Income Full.Market.Value
## 1              9.37               990610           7300000
## 2             13.94              4870962          30690000
## 3              6.39             13767000          90970000
## 4             11.18              8991643          67556006
## 5             12.68              7221385          54320996
## 6              5.23               505367           3354003
## 7              7.03               637044           5233000
## 8              8.54               647793           4687000
## 9              7.03              1171986           5967531
## 10             6.64               575891           3673011
##    Market.Value.per.SqFt          Boro
## 1                 200.00     Manhattan
## 2                 242.76     Manhattan
## 3                 164.15     Manhattan
## 4                 271.23     Manhattan
## 5                 247.48     Manhattan
## 6                  53.76 Staten Island
## 7                  57.75 Staten Island
## 8                  59.40 Staten Island
## 9                  35.80 Staten Island
## 10                 33.74 Staten Island
```

5. Mutate

Create a new column called Age
- This column will subtract the year built from the current year
- You will need to pipe select and mutate

```
Age <-
  housing %>%
    select(Neighborhood, Year.Built) %>%
    mutate(Age = 2020 - Year.Built)

head(Age)
```

```
##   Neighborhood Year.Built Age
## 1    FINANCIAL       1920 100
## 2    FINANCIAL       1985  35
## 3    FINANCIAL         NA  NA
## 4    FINANCIAL       1930  90
## 5      TRIBECA       1985  35
## 6      TRIBECA       1986  34
```

6. Summarize and Group By

Using the housing date and the summarize function, find the mean square footage

```
mean_sqft <-
  housing %>%
    summarize("Mean Sq Ft" = mean(Gross.SqFt))

mean_sqft
```

```
##   Mean Sq Ft
## 1   82762.87
```

Using summarize and group by, find the mean square footage and group by Neighborhood

```
mean_sqft_neighborhood <-
  housing %>%
    group_by(Neighborhood) %>%
    summarize("Mean Sq Ft" = mean(Gross.SqFt))

head(mean_sqft_neighborhood, 5)
```

```
## # A tibble: 5 x 2
##   Neighborhood         ‘Mean Sq Ft‘
##   <chr>                       <dbl>
## 1 ALPHABET CITY              24567.
## 2 ARROCHAR-SHORE ACRES       90618
## 3 ASTORIA                    59104.
## 4 BATH BEACH                 17304.
## 5 BAY RIDGE                  21595.
```