

# Problem Set 02: Data Wrangling Solutions

Professor Bradley Warner

June, 2022

## Documentation:

We used all the resources available to instructors from the authors of Modern Dive.

## R Packages

Loading the following packages for this problem set:

```
library(ggplot2)
library(dplyr)
```

## The data

Run the following to load and take a `glimpse` of the data:

```
data(txhousing)
glimpse(txhousing)
```

These data are about housing in Texas. Each row is monthly data for a given city in Texas in a given year. There are multiple years of data for each city.

---

## Exercise 1

After running all the code above in the console, take a look at the data using `str()`, `glimpse()`, or `View()`. In your report include that last 10 lines of the data file.

```
tail(txhousing, n=10)
```

```
## # A tibble: 10 x 9
##   city          year month sales  volume median listings inventory date
##   <chr>         <int> <int> <dbl>    <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1 Wichita Falls 2014     10   112 13817043 113300     905     7.8 2015.
## 2 Wichita Falls 2014     11    96 11308302 108000     870     7.5 2015.
## 3 Wichita Falls 2014     12   109 13883668 103800     821      7 2015.
## 4 Wichita Falls 2015      1    71  7519961  82100     829     7.2 2015
```

```
## 5 Wichita Falls 2015 2 100 11646765 94000 795 6.8 2015.
## 6 Wichita Falls 2015 3 152 16716584 89200 818 6.8 2015.
## 7 Wichita Falls 2015 4 129 15482194 105300 760 6.4 2015.
## 8 Wichita Falls 2015 5 174 19188181 100000 776 6.4 2015.
## 9 Wichita Falls 2015 6 143 18820752 118800 770 6.2 2015.
## 10 Wichita Falls 2015 7 172 23850905 116700 811 6.5 2016.
```

---

## Exercise 2

Take a look at the variable descriptions by typing `?txhousing` into the **console**. What is the `inventory` variable in this data set?

**Answer:** The number of months of inventory; the amount of time it would take to sell all current listings at current pace of sales.

---

## Exercise 3

Write a code chunk to remove the `inventory` variable. Save the results in a data frame called `txhousing`. Confirm in the data viewer that the variable has been removed.

```
txhousing <- txhousing %>%
  select(-inventory)
```

---

## Exercise 4

Make a data set called `midland_sub` that includes data only from the city of Midland in 2007 & 2014. Print the first 3 rows of `midland_sub`.

```
midland_sub <- txhousing %>% filter(city == "Midland", year == 2007 | year == 2014)
```

```
head(midland_sub, n=3)
```

```
## # A tibble: 3 x 8
##   city    year month sales  volume median listings date
##   <chr>  <int> <int> <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1 Midland 2007     1  100 18560000 150600     242 2007
## 2 Midland 2007     2  139 20760000 135600     249 2007.
## 3 Midland 2007     3  162 27370000 148600     246 2007.
```

---

## Exercise 5

Add a column to the `midland_sub` data set called `prct_sold` that calculates the percentage of listings that were sold (`sales/listings * 100`). Be sure to **save** the results also as a data frame called `midland_sub`. Print the first 3 rows of `midland_sub`.

```
midland_sub <- midland_sub %>%  
  mutate(prct_sold = sales/listings *100)
```

```
head(midland_sub,n=2)
```

```
## # A tibble: 2 x 9  
##   city    year month sales    volume median listings  date prct_sold  
##   <chr>  <int> <int> <dbl>    <dbl>   <dbl>    <dbl> <dbl>   <dbl>  
## 1 Midland 2007     1   100 18560000 150600     242 2007     41.3  
## 2 Midland 2007     2   139 20760000 135600     249 2007.     55.8
```

---

## Exercise 6

Calculate the **median** percentage of listings that were sold in Midland **in each month of the year** based on your `midland_sub` data set. Save the results of the calculation in an data frame called `midland_summary`.

```
midland_summary <- midland_sub %>%  
  filter(year==2014) %>%  
  group_by(month) %>%  
  summarize(median_prct_sold = median(prct_sold))
```

```
midland_summary
```

```
## # A tibble: 12 x 2  
##   month median_prct_sold  
##   <int>          <dbl>  
## 1     1             23.5  
## 2     2             26.8  
## 3     3             58.4  
## 4     4             30.0  
## 5     5             31.5  
## 6     6             37.8  
## 7     7             39.5  
## 8     8             31.4  
## 9     9             29.0  
## 10    10            31.4  
## 11    11            24.8  
## 12    12            24.3
```

---

## Exercise 7

Arrange the `midland_summary` in descending order based on the average percentage of listings, so you can see **which month** had the greatest **average** percentage of listings sold. You do not need to save the results.

```
midland_summary %>%
  arrange(desc(median_prct_sold))
```

```
## # A tibble: 12 x 2
##   month median_prct_sold
##   <int>         <dbl>
## 1     3          58.4
## 2     7          39.5
## 3     6          37.8
## 4     5          31.5
## 5    10          31.4
## 6     8          31.4
## 7     4          30.0
## 8     9          29.0
## 9     2          26.8
## 10    11          24.8
## 11    12          24.3
## 12     1          23.5
```

---

## Exercise 8

In August of 2010, what city had the fewest houses listed for sale? (show code and text; do not print out all 46 rows of the data frame.)

**Answer: San Marcos**

```
txhousing %>%
  filter(year == 2010, month == 8) %>%
  arrange(listings) %>%
  head()
```

---

## Exercise 9

In 2013, in which month were the most houses sold in Texas? (show code and text)

**Answer: August**

```
txhousing %>%
  filter(year == 2013) %>%
  group_by(month) %>%
  summarize(top_sales = max(sales)) %>%
  arrange(desc(top_sales))
```

```
## # A tibble: 12 x 2
##   month top_sales
##   <int>   <dbl>
## 1     7    8468
## 2     5    8439
## 3     8    8155
## 4     6    7935
## 5     4    7116
## 6     9    6706
## 7    10    6551
## 8     3    6382
## 9    11    5557
## 10     2    4886
## 11     1    4273
## 12    12     NA
```

---

## Exercise 10

Generate a single table that shows the total number of houses sold in **Galveston** in **2009 and 2010** (total over the entire period), & the total number of houses sold in **Amarillo** in **2009 and 2010** (total over the entire period). This calculation requires a number of steps, so it might help you to first write out on paper the different steps you will need to take. That will help you set out a “blueprint” for tackling the problem.

```
txhousing %>%
  filter(city == "Galveston" | city == "Amarillo") %>%
  filter(year == 2009 | year == 2010) %>%
  group_by(city) %>%
  summarize(total_sales = sum(sales))
```

```
## # A tibble: 2 x 2
##   city      total_sales
##   <chr>         <dbl>
## 1 Amarillo      5353
## 2 Galveston     1625
```

## Documenting software

- File creation date: 2022-06-18
- R version 4.1.3 (2022-03-10)
- dplyr package version: 1.0.9
- ggplot2 package version: 3.3.6