Problem Set 03: Data Wrangling

WRITE YOUR NAME HERE

WRITE DATE HERE

Contents

HOW TO USE THIS RMD FILE

For the rest of the semester you will be given a `.Rmd` file like this to type your problem set answers into. Make sure to save this file somewhere you can find it.

> You can type your answers right into this `.Rmd` document, knit the final copy, and turn in the knitted PDF file as the answers to the problem sets.

Please be sure to type your name above where it says `WRITE YOUR NAME HERE` and fill in the date too.

R Packages and Data

```
library(ggplot2)
library(dplyr)
data(txhousing)
```

Question 1

After running all the code above in the console, take a look at the data in the data viewer. You can accomplish this two different ways:

a) click on the name of the data in the Environment pane, or
b) type `View(txhousing)` in the **console**.

What is the last city listed in the data set (in row 8602)?

Answer:

Question 2

Take a look at the variable descriptions by typing `?txhousing` into the **console**. What information does the `listings` variable contain?

Answer:

Basic syntax

This first set of questions will help you practice basic syntax. All you need to include is a Question header, and code for each.

Question 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.

Question 4

Make a data set called `dallas_sub` that includes data only from the city of Dallas in 2012 & 2013.

Question 5

Add a column **to the `dallas_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 `dallas_sub`.

Question 6

Calculate the **average** percentage of listings that were sold in Dallas **in each month of the year** based on your `dallas_sub` data set. Save the results of the calculation in an data frame called `dallas_summary`.

Question 7

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

More advanced wrangling

Please answer the following questions with text and/or code where appropriate. You may have to use multiple `dplyr` functions to answer each question. Think through the steps of how to get to the answer you are trying to find.

Question 8

Run the following code chunk. Study the code, and the output. Explain in your own words what this code chunk calculated.

```
txhousing %>%
  filter(year == 2012 | year == 2013, city == "Dallas") %>%
  mutate(prct_sold = sales/listings *100) %>%
  group_by(month) %>%
  summarize(mean_prct_sold = mean(prct_sold)) %>%
  arrange(desc(mean_prct_sold))
```

Answer:

Question 9

In January of 2015, what city had the fewest houses listed for sale in all of Texas? (show code and text please)

Answer:

Question 10

In 2012, in which month were the most houses sold in Texas? (show code and text please)

Answer:

Question 11

Generate a single table that shows the total number of houses sold in **Austin** in **2000 and 2001** (total over the entire period), & the total number of houses sold in **Dallas** in **2000 and 2001** (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.

Hint: recall the `sum()` function can add values.