Homework 03

DSO 545: Statistical Computing and Data Visualization

Spring 2016

Due Date: Thursday March 24, 2015

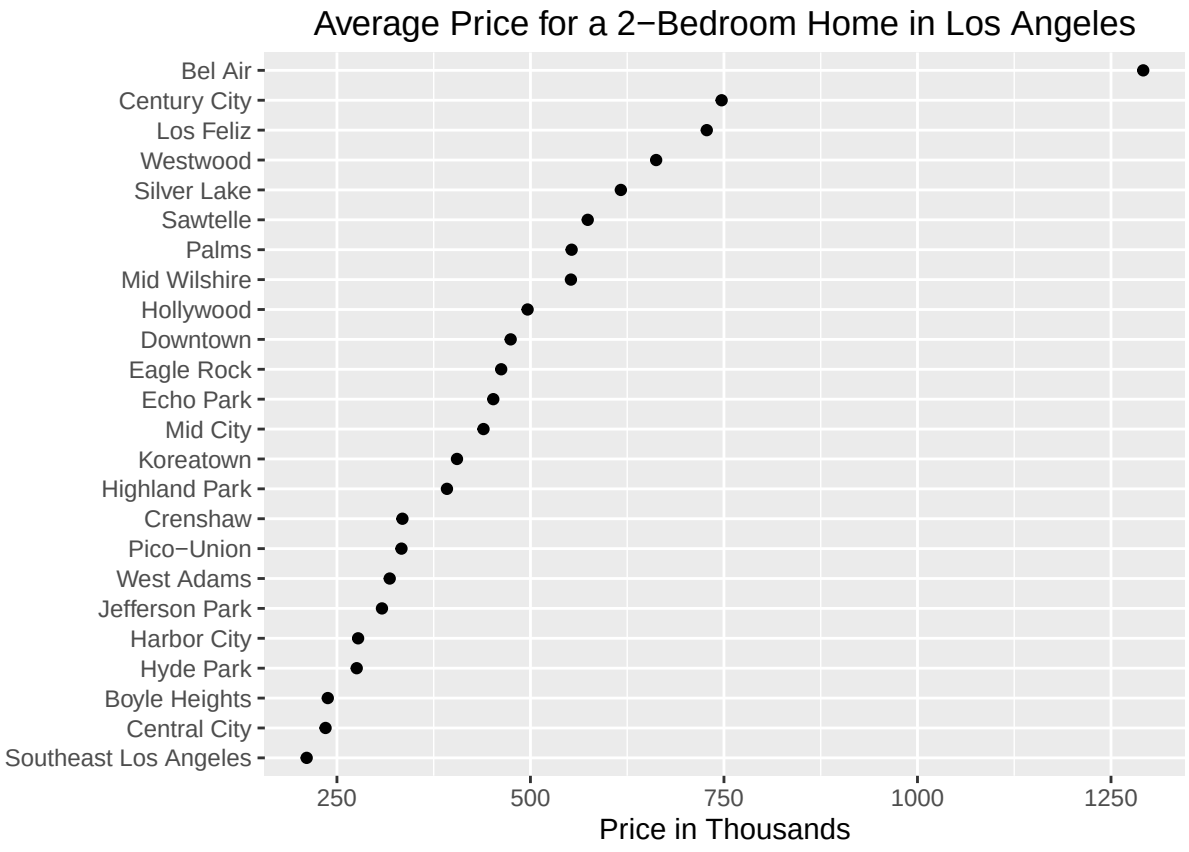
Case 01

This time we are going to explore a modified dataset from Zillow Home Value Index. The dataset records median estimated home value for all homes with two bedrooms in 24 different regions in Los Angeles.

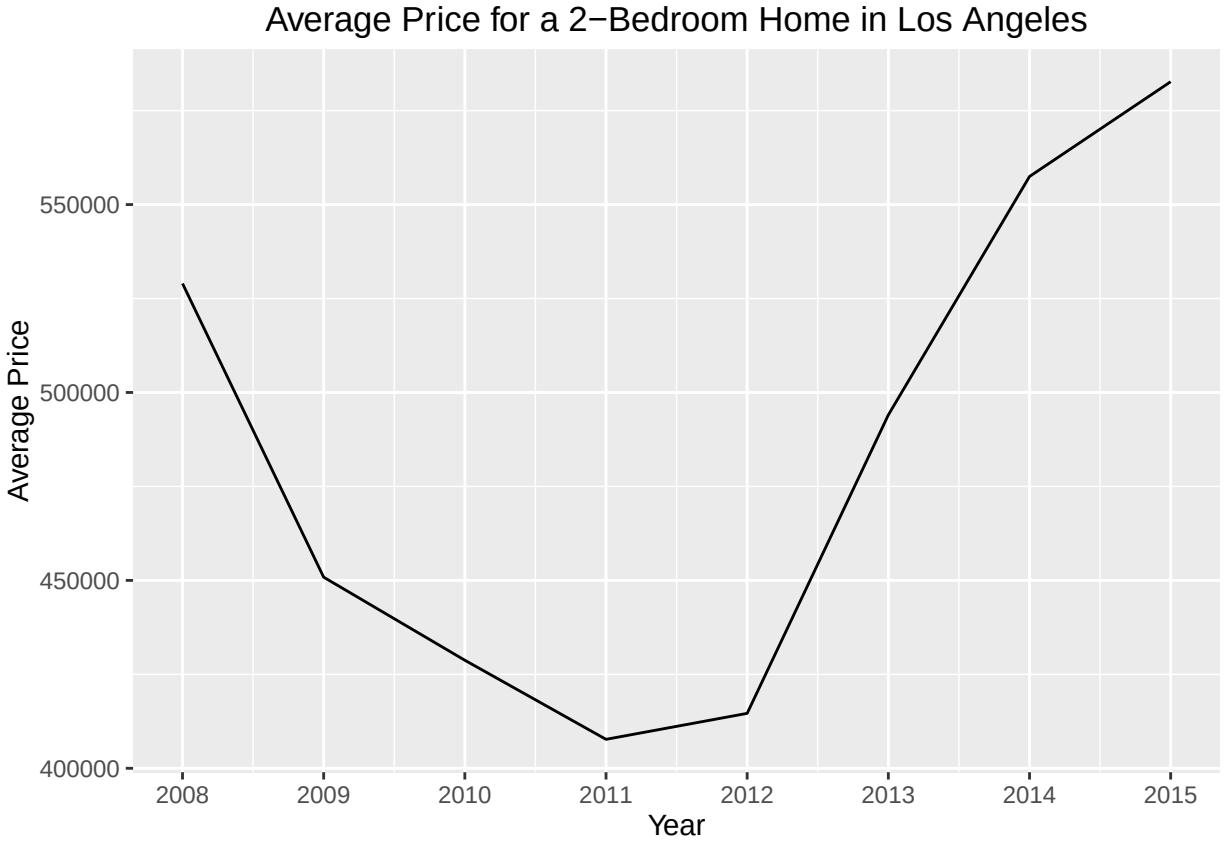
```
zillow <- read.csv("Zillow2bedroom.csv")
```

Questions

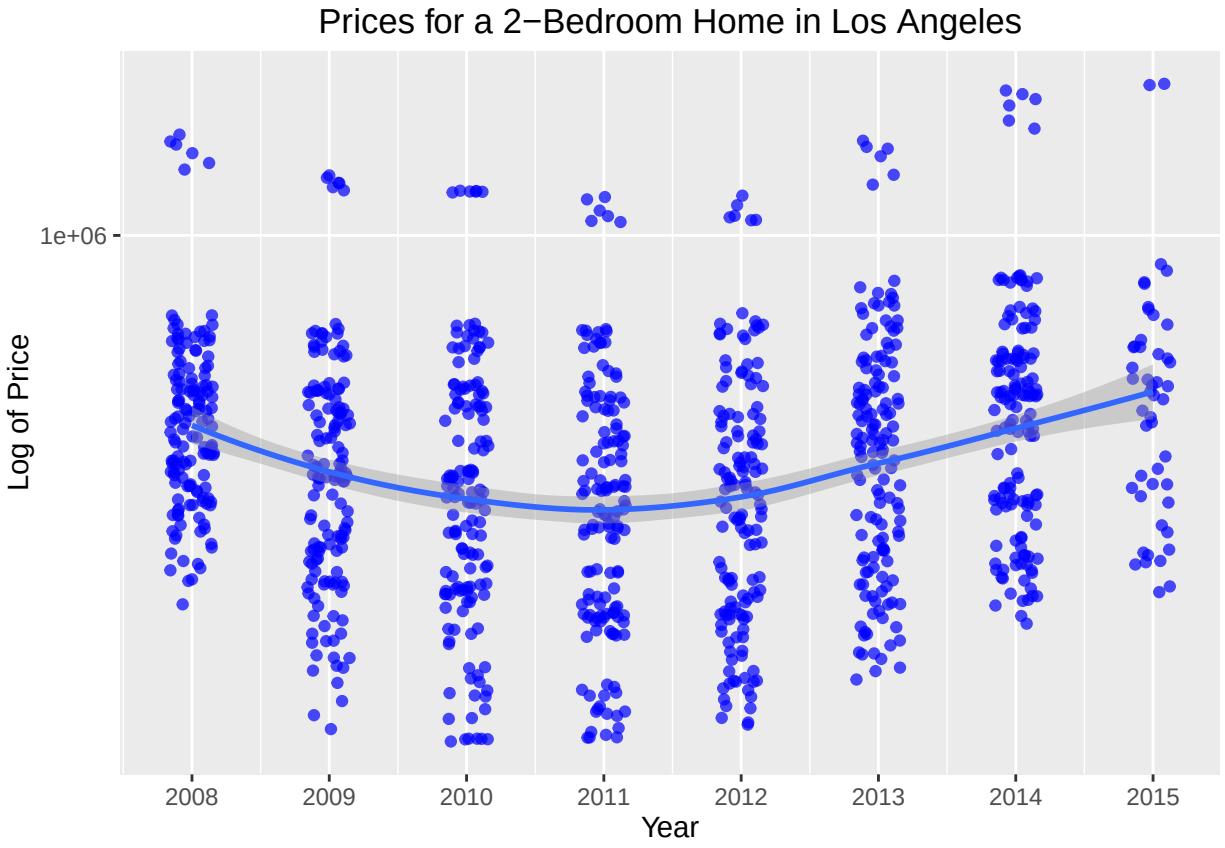
1. *(3 points)* The following plot shows the **mean** of the home prices by region. Reproduce an **EXACT** copy the following graph. (Hint, you might need to aggregate the data!)



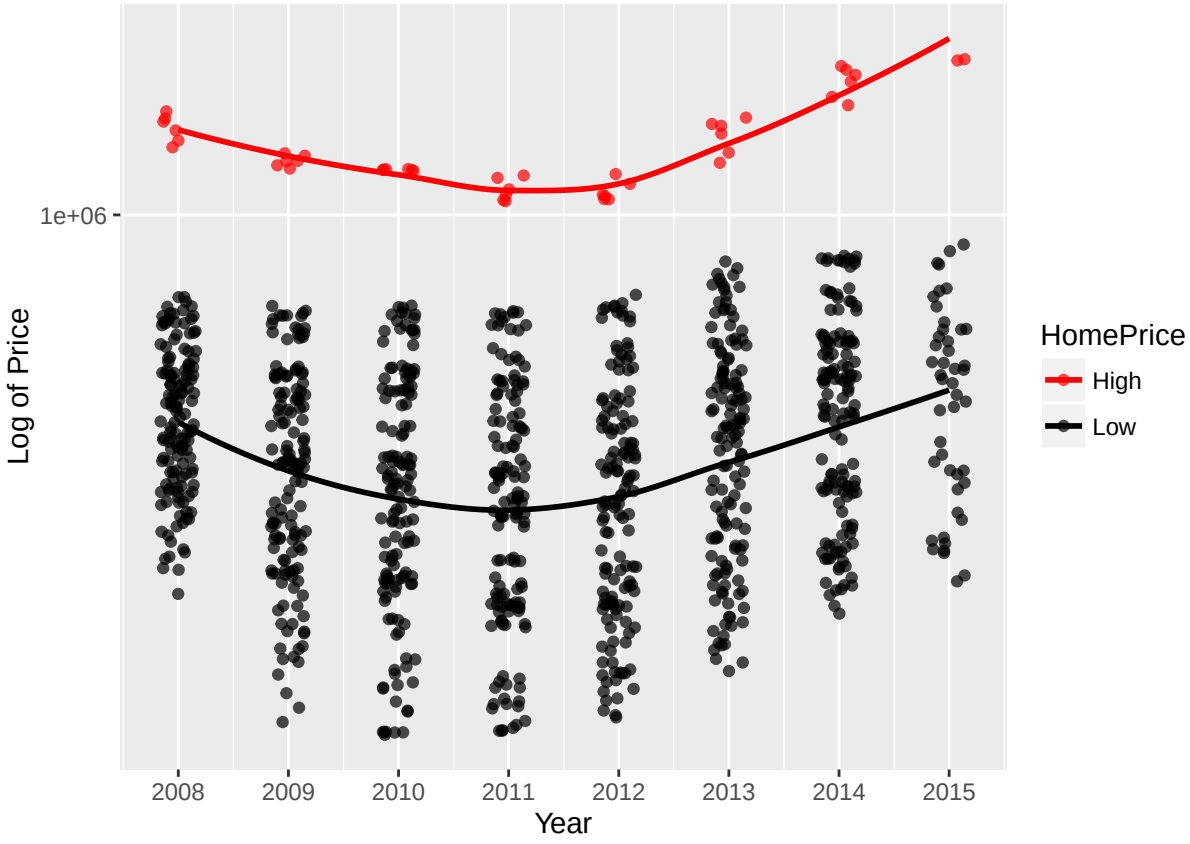
2. *(2.5 points)* The following plot shows the change in the **mean** of home prices for the whole city of Los Angeles from 2008 till 2015. Reproduce an **EXACT** copy the following graph. (Hint, you might need to aggregate the data!)



3. *(3 points)* The following plot shows the home prices for the whole city of Los Angeles from 2008 till 2015. Reproduce an **EXACT** copy the following graph. (Hint, you might need to add some jitter!)



4. *(3 points)* You can see that there is clear seperation in our plot. For this reason, we will split our data into two catgories: houses with high prices (price >= $1,000,000), and houses with low prices (price < $1,000,000). Reproduce an **EXACT** copy the following graph. (Hint, you might need to add an extra variable to your dataset!)



Case 02

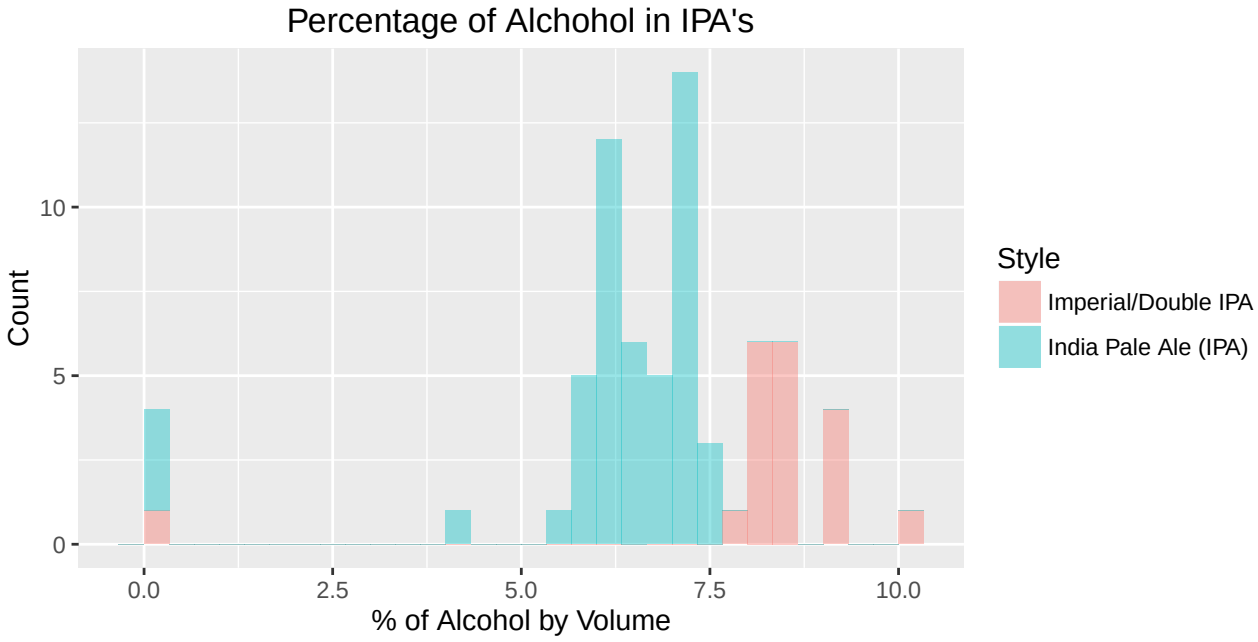
It's almost summer time! Let's enjoy some beer together. This dataset contains ratings of 400 different craft beers from various breweries. The dataset contains 7 variables:

- `Name` : name of the beer
- `Brewer` : name of the brewer
- `Style` : style of the beer
- `Abv` : % of alcohol by volume of the beer
- `Ratings` : rating of the beer
- `ScoreOverall`: The overall score of the beer

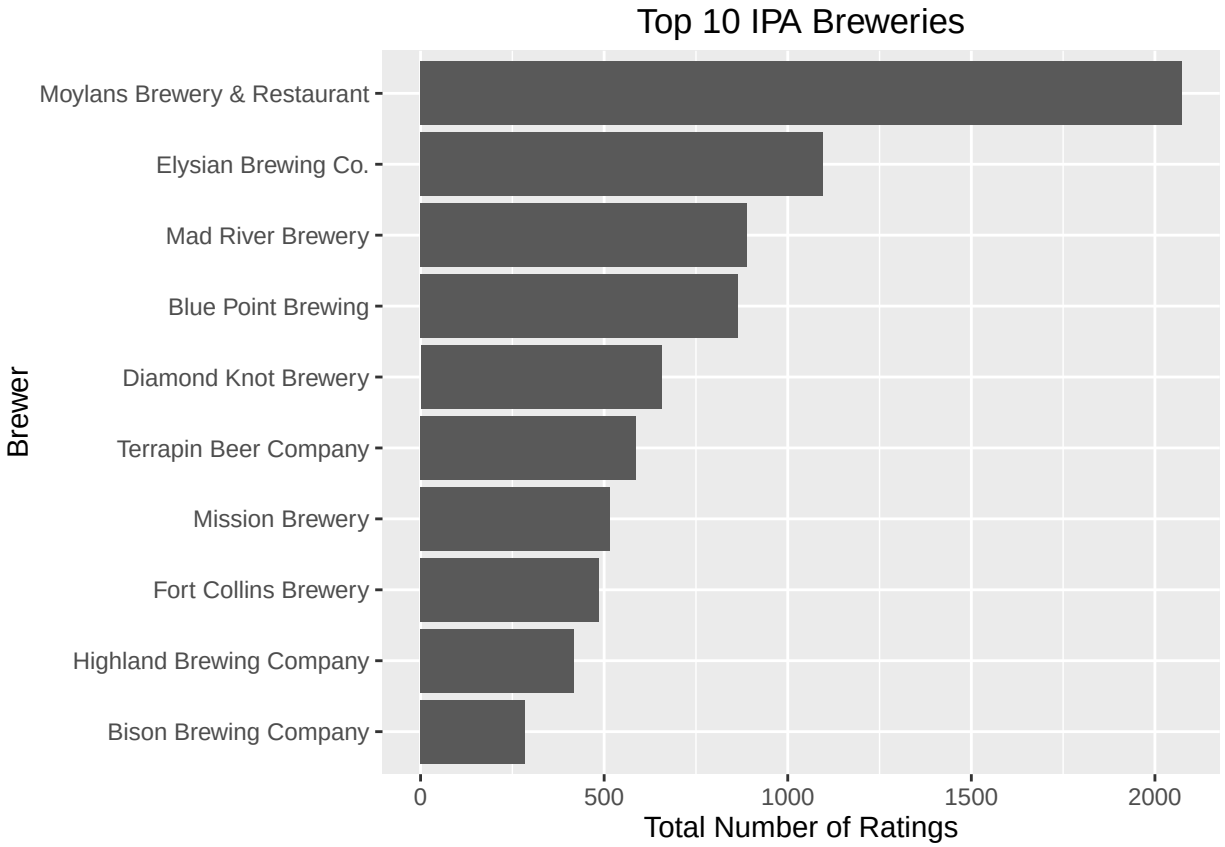
**You should use `dplyr` functions to solve questions 1, 2, and 3.

Questions

1. *(3 points)* Create a subset of only IPAs, i.e. include both `Imperial/Double IPA` and `India Pale Ale (IPA)`. Show the following histagram of `Abv`. What does the histogram tell you?



2. *(3 points)* Based on the subset you created in the previous question, calculate the total number of ratings for each brewer, then reproduce an **EXACT** graph as the following for the top 10 brewers which has received the highest ratings.



3. *(2.5 points)* We're all classy, so we only drink from the top 30 overall rated beers, but we CANNOT drink beers with `Abv` higher than 8% just because we don't want to be wasted!

Randomly choose one beer for yourself from the top 30 rated beers! Output only the name of beer. (Hint: you might need to use `sample_n()` function from `dplyr` R package)

Deliverables

Assemble the assignment using the R Markdown tool, and upload both RMD and PDF (or Word) file to blackboard. You can talk about questions with your classmates, and it's often useful to get feedback from others on your plots, but everything that you hand in should be yours, and you can consider homeworks to be pledged.