

# STAT 331 Lab 8

## Instructions

Submit your .Rmd or .R file by the beginning of class.

## Introduction

In this lab we're going to be looking at avocado prices! The dataset comes to us from kaggle and represents weekly retail scan data: avocado.csv.

## Variables

- Date
- AveragePrice
- Total Volume
- 4046
- 4225
- 4770
- Total Bags
- Small Bags
- Large Bags
- XLarge Bags
- type
- year
- region

## Exercises

- 1) Read in the dataset and rename the variables so that there are no spaces in them. You may assume they are in the order listed above.
- 2) Plot side-by-side boxplots of TotalVolume for the 5 regions with the highest averages for the TotalVolume variable. (Hint: consider the pull() function to convert tibbles to character vectors)
- 3) Subset the data down to the least recent and the most recent years (two year values) present in the dataset, then plot overlaid histograms of the AveragePrice variable for these two years.
- 4) Plot AveragePrice vs. TotalVolume for the type with the higher sum of TotalBags.
- 5) Plot side-by-side barcharts of type for the 7 regions with the lowest total number of SmallBags.

## Warning!

- 6) OOPS!!!! I accidentally only gave you 1/3 of the dataset. For this lab you will submit your .Rmd or .R file and I will re-run your code on the full dataset myself. This means everything should still work and adjust to the full dataset, i.e. not depend explicitly on this smaller dataset you've been working with.