Assignment 3: Visualizing Data in R

Your Name

Date

Instructions

Download the assignment03.Rmd file from Canvas and open it in RStudio. Complete this assignment by filling in the answers below in the R Markdown Notebook document. You may (and should!) use the internet to help solve these problems.

In this assignment you will generate clear, aesthetically pleasing graphics using an R package called ggplot2. Please only use ggplot2 to generate graphics in this assignment. You can learn more about ggplot here (hint: this link can help you with many of the questions below). To install and load the ggplot2 package, run the following two lines of code in the R console:

```
install.packages("ggplot2")
library(ggplot2)
```

Your visualizations will be graded based on the highest standards of graphical excellence including the degree to which they maximize the data-ink ratio (within reason) and adhere to the principles of data graphics covered in lecture. For example, colors should be chosen deliberately. Also, be sure to properly label all axes (include units) and give your plots a descriptive title.

Context

Similar to last week, the avo_sub.csv file contains data from the Hass Avocado Board website in May of 2018. The data table represents weekly retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers' cash registers based on actual retail sales of Hass avocados from 2015 to 2018. The avo_sub.csv contains a cleaned subset of the dataset with data for avocados in DFW, Houston, and San Francisco.

Deliverables

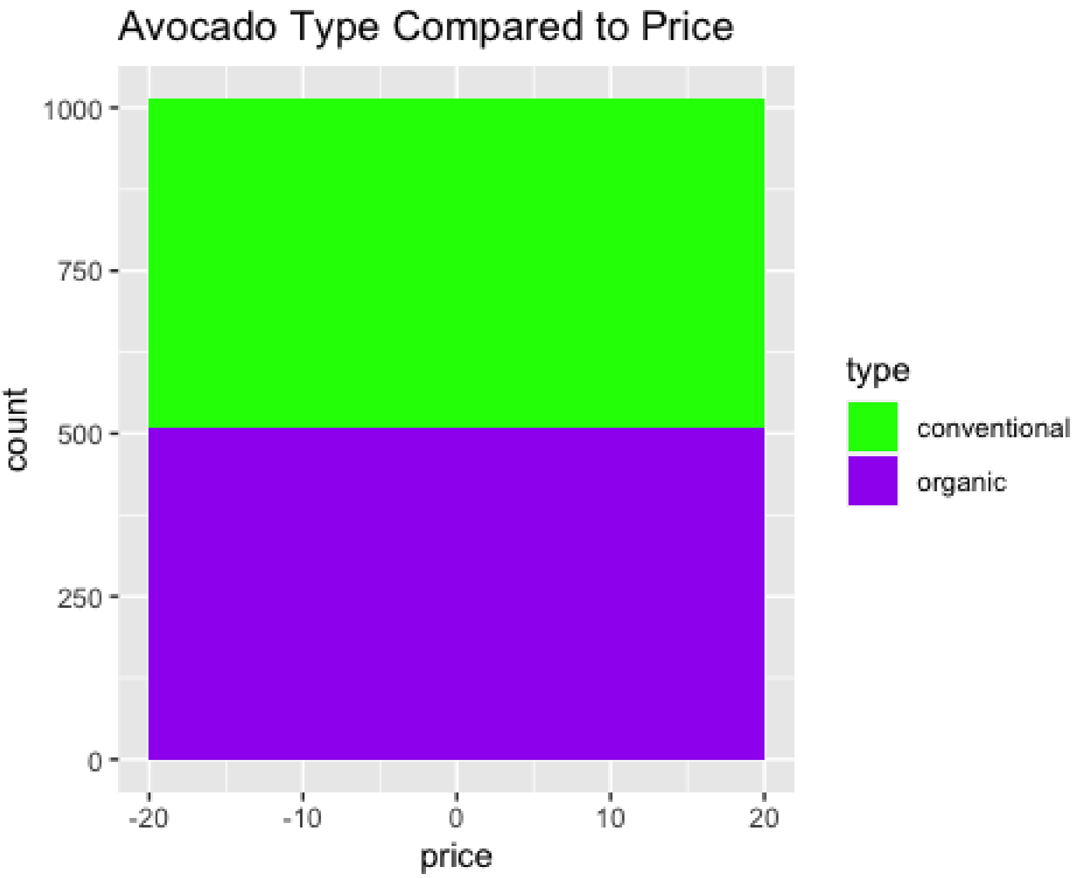
Please submit to Canvas the following items (separate, not as a single zip file):
1. An HTML (or Word or PDF) file knitted from the .Rmd file
2. A completed .Rmd file

Questions

1. Read the file avo_sub.csv directly into R using the following URL: http://people.tamu.edu/~geoallen/courses/312/avo_sub.csv and assign it to the variable "avo". Using the data contained in the avo data frame, plot a stacked histogram of avocado

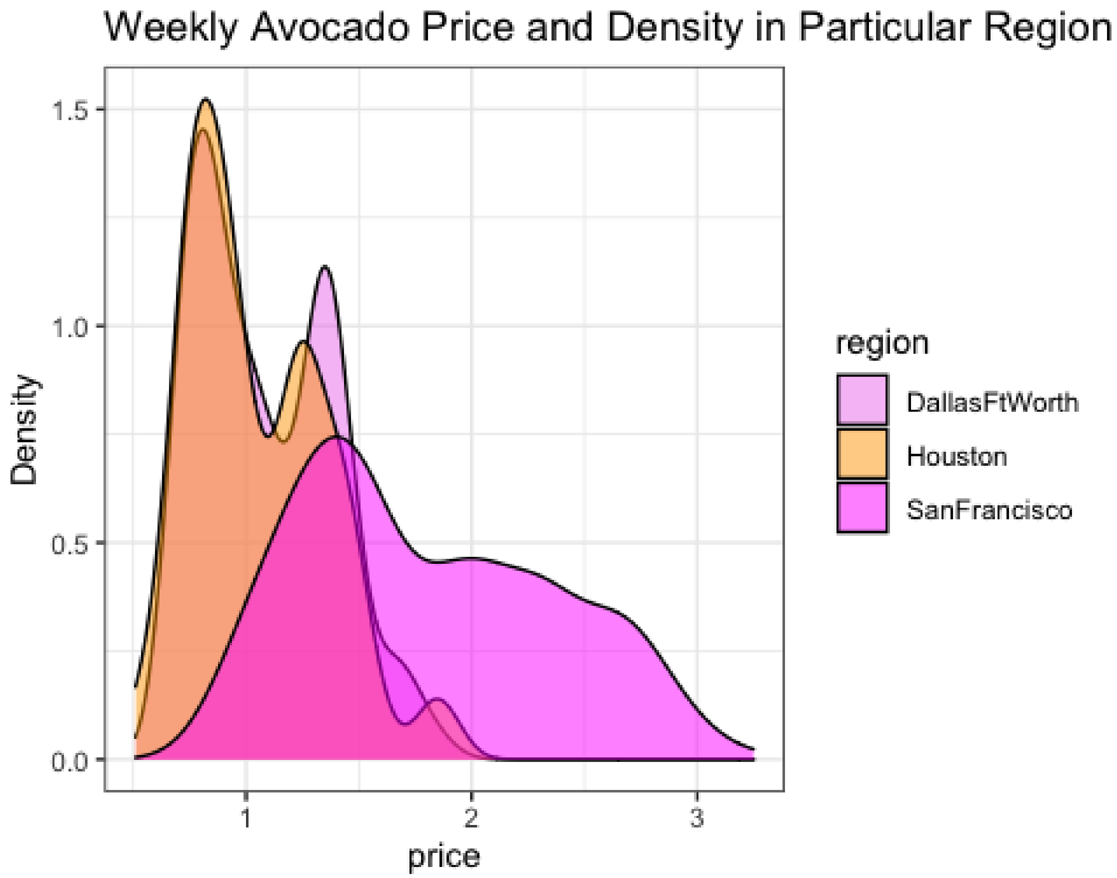
price with 40 bins, colored by avocado type (organic/conventional). Use a color scheme other than the ggplot2 default.

```
path = "http://people.tamu.edu/~geoallen/courses/312/avo_sub.csv"
avo = read.csv(path, header=T)
library(ggplot2)
ggplot(data = avo, aes(x=price,fill = type)) + geom_histogram(binwidth=40) +
 labs(title = "Avocado Type Compared to Price") + scale_fill_manual(values =
c("Green","Purple"))
```



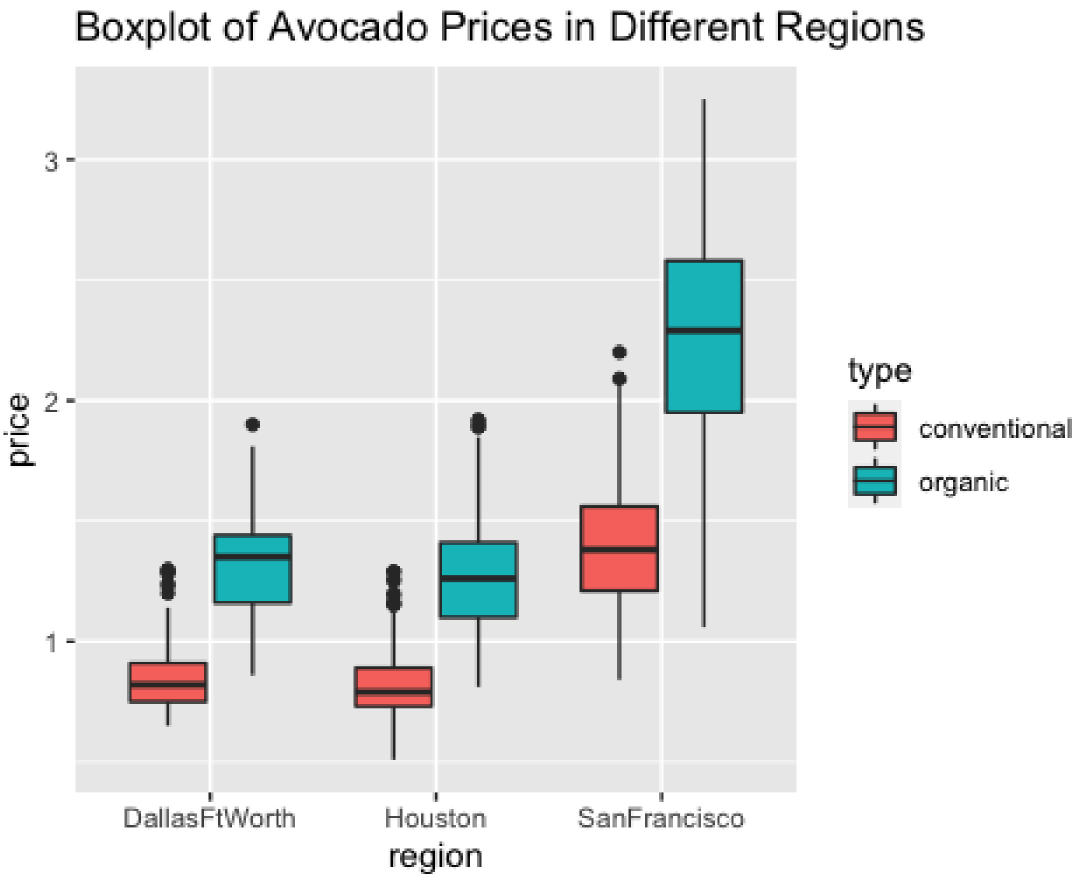
2. Create a density plot containing weekly avocado prices in each of three regions. Color the density plots by region using semi-transparent fill colors. Be sure that the color scheme is inclusive to people who are color blind.

```
# Colors chosen for people with colorblindness (whether blue-yellow or red-
green)
library(ggplot2)
library(scales)
ggplot(data=avo, aes(x=price, fill=region)) + geom_density(alpha=0.5) +
  theme_bw() +
  labs(y = "Density",
       title = "Weekly Avocado Price and Density in Particular Regions") +
scale_fill_manual(values = c("Violet","Orange", "Magenta"))
```



3. Create box plots in a single figure of weekly avocado prices by region. For each region, create two separate box plots by type (there should be six box plots total). Is the median price of convention avocados in San Francisco greater or less than the median price of organic avocados in DFW?

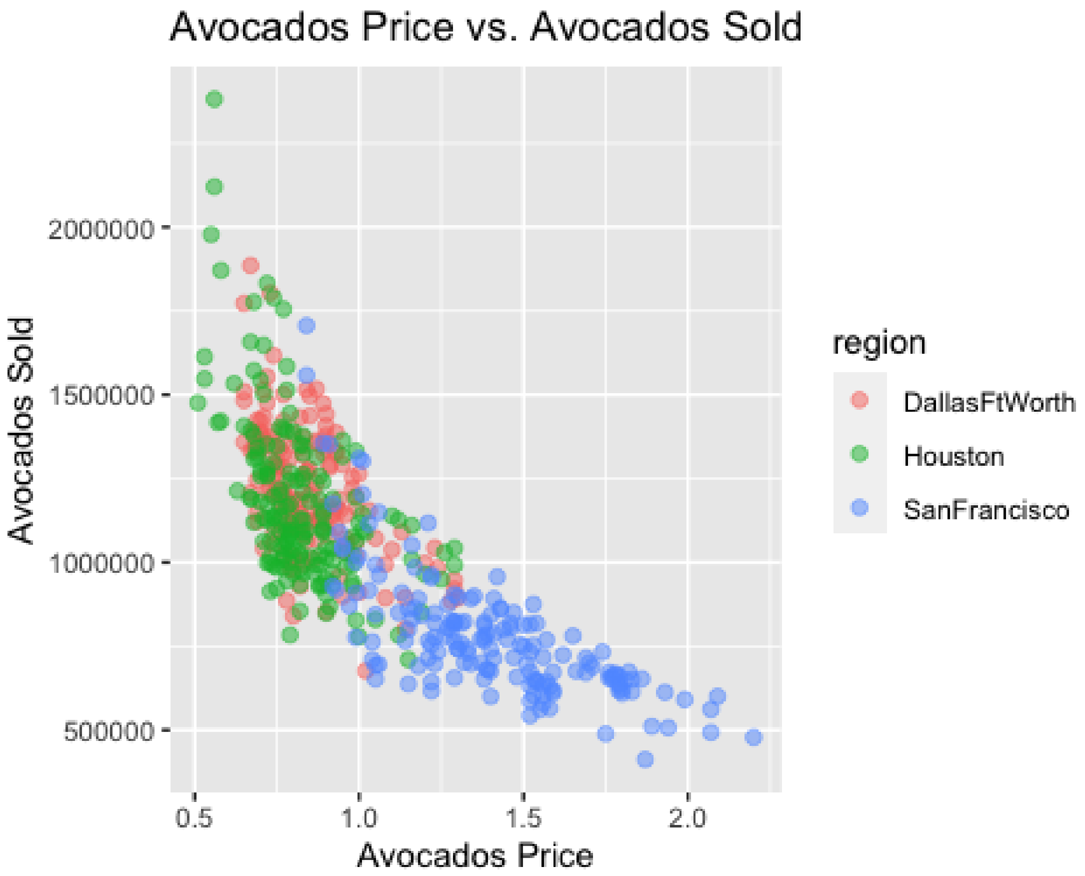
```
library(ggplot2)
ggplot(avo, aes(x=region, y=price, fill=type)) +
  geom_boxplot() +
  labs(title = "Boxplot of Avocado Prices in Different Regions")
```



```
# The median price for organic avocados is more in San Francisco than in DFW
```

4. Create a new data frame made up of only rows with only conventional type avocado (e.g. remove all rows with organic type avocados). Set this new data frame to the variable “avo_conv”. Using avo_conv, plot a scatter plot of conventional avocado prices vs. conventional avocados sold and color code the points by region. No underscores should be included in your labels. In the space below the code chunk, explain what the data are showing (this should take a short paragraph).

```
avo_conv = avo[!grepl("organic",avo$type),]
library(ggplot2)
  ggplot(avo_conv, aes(price,avocados_sold, color=region)) +
  geom_point(alpha=0.5, size=2) +
  labs(y="Avocados Sold", x="Avocados Price", title="Avocados Price vs.
Avocados Sold")
```

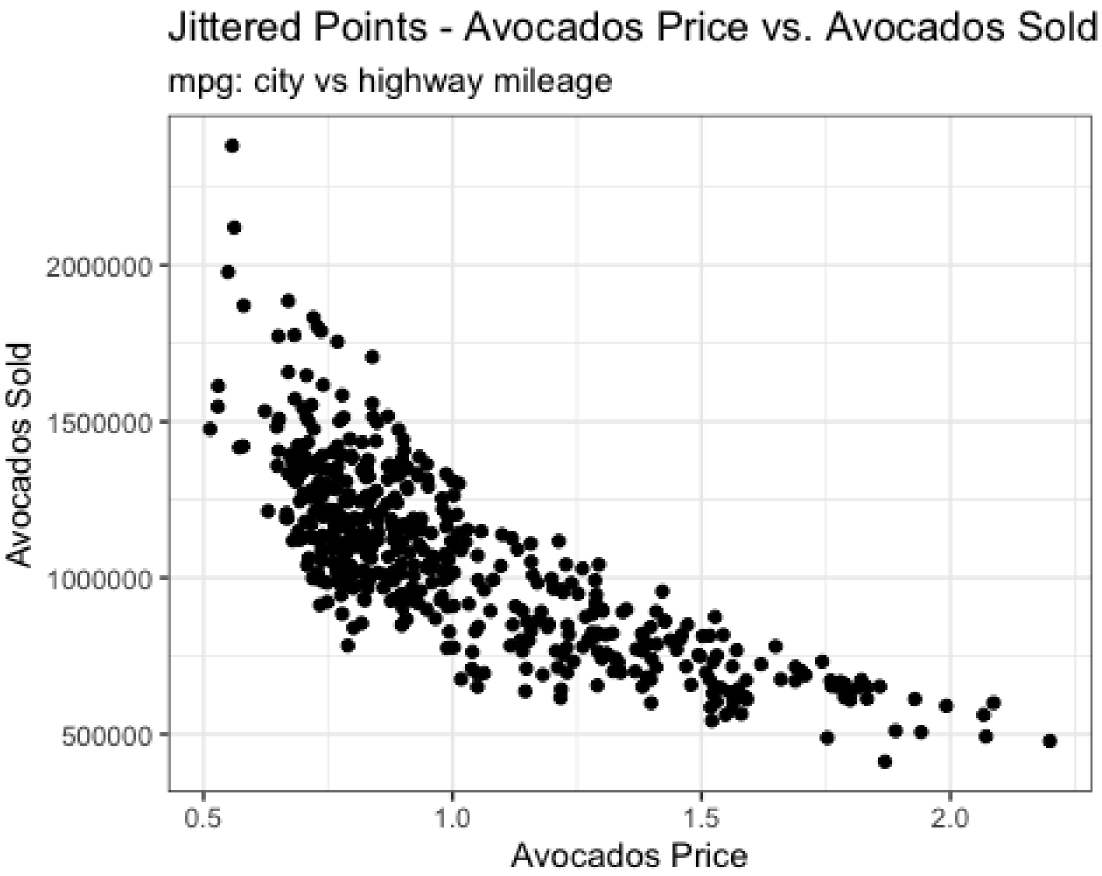


```
# The data shown below is being displayed using a scatter plot. It is showing
the amount of conventional type avocados being sold at different prices per
region. In this case, most people buy avocados at cheaper prices, especially
in Houston and the DFW area. In the case of San Francisco there are some
people who buy avocados at higher prices, likely being that San Francisco is
a more expensive city in general than Houston and DFW.
```

Your explanation goes here (below your plot).

5. Take a look at this list of some popular ggplot2 visualizations. Use the avocado data to create an interesting plot that we have not covered in this assignment. In the space below describe what is seen in your plot in 1-3 sentences.

```
library(ggplot2)
theme_set(theme_bw())
ggplot(avo_conv, aes(price, avocados_sold)) + geom_jitter() +
  labs(subtitle="mpg: city vs highway mileage",
       y="Avocados Sold",
       x="Avocados Price",
       title="Jittered Points - Avocados Price vs. Avocados Sold")
```



```
# On this question I chose to plot Avocados Price vs Avocados sold again, but
this time by using a jitter plot. Jitter plots use random values for the dots
located on a plot so the dots are not directly on each other. Jitter plots
are necessary in data science, and can be used in addition to numerous other
plots.
```