MATH 208 2021 Midterm Exam

Question 1 [50 points]

a. **[15 pts]** Consider code chunk **Q1** and give the resulting output from the following R commands (or the resulting error if one is produced):

   i. `my_Costco_list[2]`

   ii. `my_Costco_list[[1]][2]`

   iii. `my_Costco_list[[1]][2,2]`

b. **[15 pts]** Consider code chunk **Q1** and answer the following multiple choice questions (please list all that apply for each question):

   i. Which of the following commands returns the result “Coffee”?

```
A.`both_lists$Costco$Item[2]`
B.`both_lists$Costco$Item[[2]]`
C.`both_lists[c(1,1,1,2)]`
D.`both_lists[[c(1,1,1,2)]]`
```

   ii. The class of the object returned by `both_lists[2]` is a

```
A. atomic character vector
B. tibble
C. list
```

   iii. The class of the object returned by `both_lists[[2]]$Location` is a

```
A. atomic character vector
B. tibble
C. list
```

c. **[20 pts]** Consider code chunk **Q1** for the following questions:

   i. **[10 pts]** Write a line of code that will change the Location column of the tibble in my_Costco_list to an ordered factor where the order of the levels is in the following order: Clothes, Cereal, Refrigerated.

   ii. **[10 pts]** Write a line of code using your answer to part (i) to compute the **minimum** amount of items (not the number of different items) required in each location.

Question 2 [50 pts]

a. **[15 marks]** Please indicate which plot **best** summarizes the characeristic the characteristic (the name of the plot, not the function):

i. Which plot best allows one to identify points that are **outliers** (i.e. far from the central location the data): A: Boxplot B: Histogram?

ii. Which plot best allows one to assess the **skew** of the data: A: Boxplot B: Histogram?

iii. Which plot should be used to summarize the distribution of quantitative data that take on a large number of unique values: A: Histogram B: Barplot?

iv. Which plot should be used to summarize the counts from different levels of a qualitative factor variable: A: Histogram B: Barplot?

v. If we want to compare the central 50% of the values of a quantiative variable across multiple levels of a separate qualitative variable, which plot should be used:

   A. One boxplot for each group in the same plot

   B. One boxplot for each group in a different plot

   C. One hisotogram for each group in the same plot

   D. One histogram for each group in a different plot?

Answer the parts (b), (c), and (d) below based on a dataset from Kaggle containing the prices of Hass avocados over a period of over three years. The original dataset is used in the book "Introduction to Data Analysis" and is available through the `aida` package in R (which you can install to access the data).

```
### You can install the aida package via:
### install.packages("remotes")
### remotes::install_github("michael-franke/aida-package")

library(aida)
glimpse(data_avocado)
```

```
Rows: 18,249
Columns: 7
$ Date              <date> 2015-12-27, 2015-12-20, 2015-12-13, 2015-12-06, 201…
$ average_price     <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1.02…
$ total_volume_sold <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60, 5…
$ small             <dbl> 1036.74, 674.28, 794.70, 1132.00, 941.48, 1184.27, 1…
$ medium            <dbl> 54454.85, 44638.81, 109149.67, 71976.41, 43838.39, 4…
$ large             <dbl> 48.16, 58.33, 130.50, 72.58, 75.78, 43.61, 93.26, 80…
$ type              <chr> "conventional", "conventional", "conventional", "con…
```

```
data_avocado %>% arrange(Date) %>% head(.) %>% kable(.) %>% kable_styling()
```

Date	average_price	total_volume_sold	small	medium	large	type
2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	conventional
2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	conventional

2015-01-04	1.08	788025.06	53987.31	552906.04	39995.03	conventional
2015-01-04	1.01	80034.32	44562.12	24964.23	2752.35	conventional
2015-01-04	1.02	491738.00	7193.87	396752.18	128.82	conventional
2015-01-04	1.40	116253.44	3267.97	55693.04	109.55	conventional

Each row in the dataset contains information on the price and volume (in pounds) of avocados sold from one outlet on a particular date for a particular type of avocado (organic or conventional). The variables in the data set for each outlet one each date are as follows:

Variable name <chr>	**Variable description** <chr>
Date	Date of price measurement
average_price	Average price per avocado of all avocados sold
total_volume_sold	Total volume of avocados sold
small	Total volume of small avocados sold
medium	Total volume of medium avocados sold
large	Total volume of large avocados sold
type	Type of avocado (organic or conventional)

7 rows

b. **[15 marks]** Figure 1 shows three different panels examining the relationship between average price and type of avocado.

   i. **[5 marks]** Compare the distributions of the average prices for the conventional and organic avocados. In particular, compare their central locations, spreads and skew to conclude whether there is evidence in this data that the average prices of organic avocados is different from those of conventional avocados.

   ii. **[5 marks]** The lines in the center of the boxes in Panel (a) indicate the value of a summary statistic of the daily outlet average prices for each of the two types of avocado in the dataset. Write a function that will use the `data_avocado` object to compute these two values and return them in a tibble with one column for the type and the other with the summary statistic.

   iii. **[5 marks]** Explain what was changed or added to the code to yield the different figures for (b) and panel (c). In particular, explain why the bars look different in the two plots and why the y-axis scales of the figures are different.

c. **[20 marks]** Figure 2 shows two different panels comparing how the total volume sold of conventional avocados change over time (note that both plots are plotted on a log scale).

i. **[5 marks]** Which of the two panels, (a) or (b), best shows the pattern of total volume sold? Explain your answer.

ii. **[5 marks]** Would a 2D histogram make sense to use for summarizing this particular relationship? Briefly explain why or why not.

iii. **[5 marks]** Below is the line of code that generated the plot in panel (b). Change the line of code below so that it produces a pair plots that contains `total_volume_sold` for **both** conventional and organic **separately** over time (i.e. in different plotting areas of the same plot). s

```
ggplot(data_avocado %>% filter(type=="conventional"),
           aes(x=Date,y=total_volume_sold,group=Date)) +
  geom_boxplot() + ggtitle("Panel (b)")+ scale_y_log10()
```

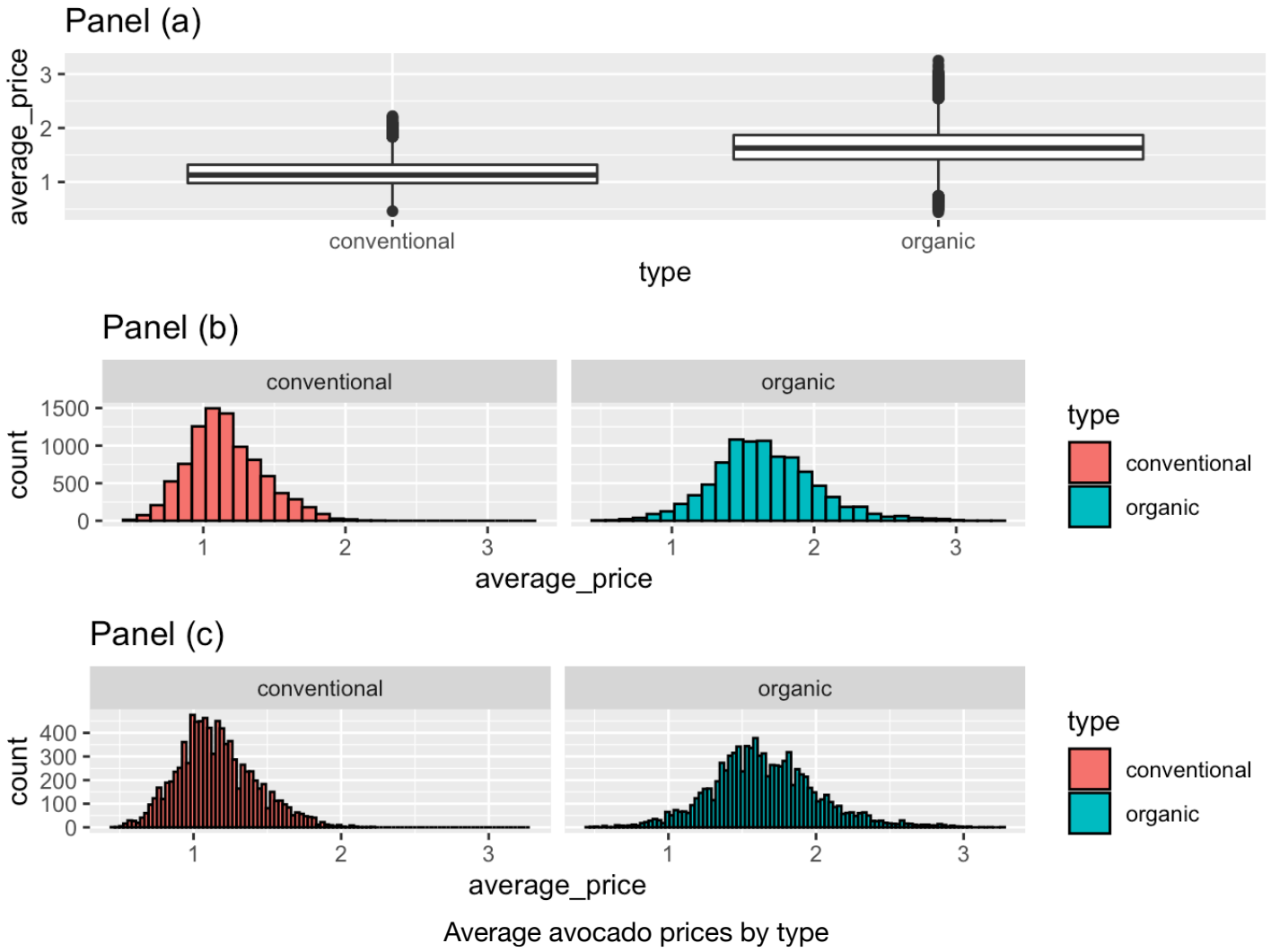
iv. **[5 marks]** Why can't you use the original `data_avocado` data object to plot the total volume sold by the size of the avocados using the same figure you gave code to generate in part (iii) above? Explain why and then write a line of code which would allow you to generate the necessary data object which could then be used to make such a plot (you don't need to include the code for the plot).

R plots and output

Code chunk Q1

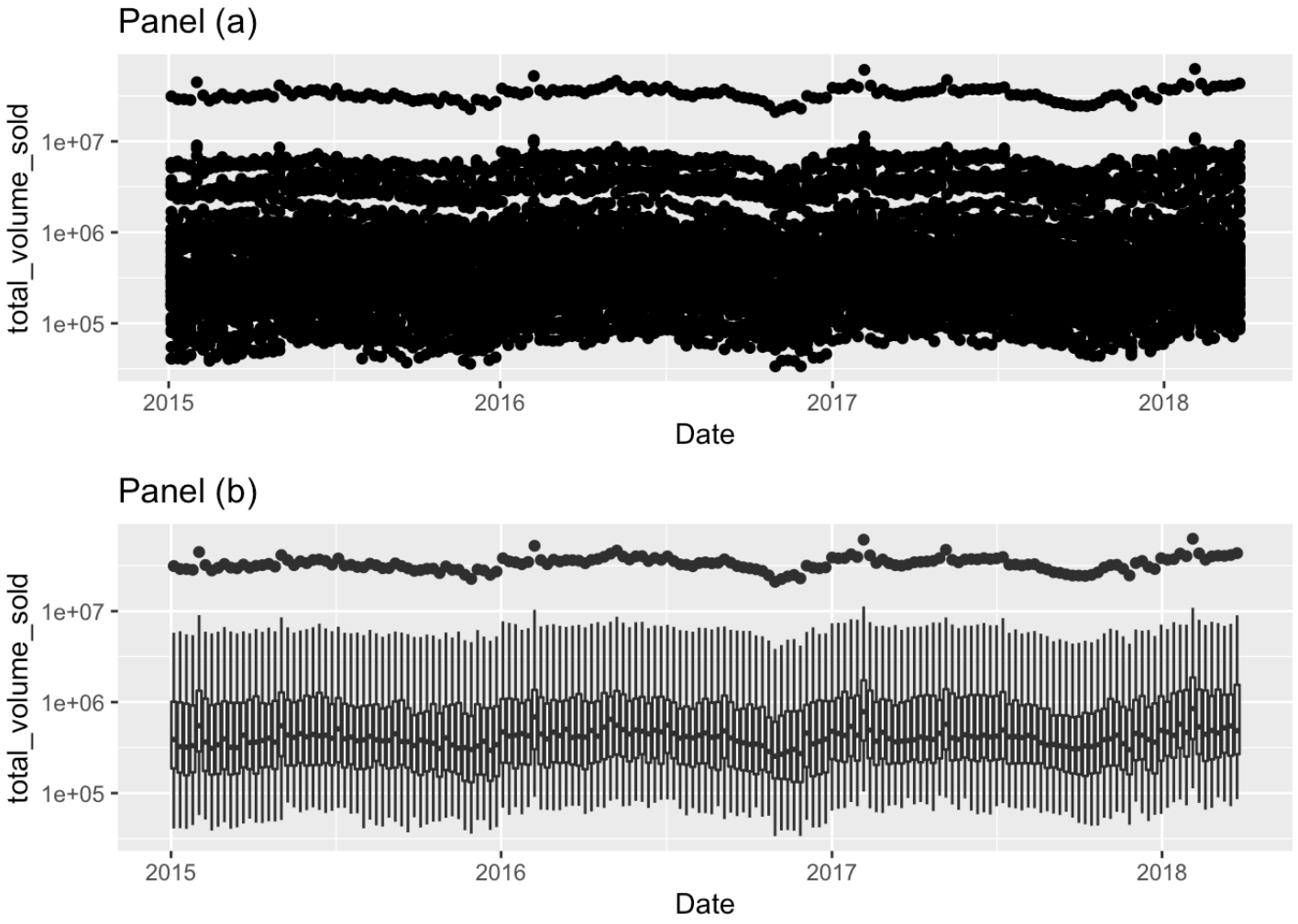
```
my_Costco_list <- list(
  tibble(Item = c("Milk", "Coffee", "Lettuce","Pants"),
                 Amount = c(4,1,2,2),
                 Location = c("Refrigerated","Cereal","Refrigerated","Clothes"))
)

both_lists <- list(Costco = my_Costco_list,
                   IGA = tibble(
                     Item = c("Chicken", "Yogurt", "Pizza"),
                     Amount = c(2,3,1),
                     Location = c("Meat","Dairy","Frozen")))
```

Figure 1



Average avocado prices by type

Figure 2



Plots of total volume sold over time