SPOT CHECK QUIZ NO. 2

Reister L. Ruedas

#### Instructions

The due date is 11:59pm June 5, 2019. Submit this Rmd file on Canvas Don’t submit additional files.

Follow all instructions from homework 1 regarding comenting and reporting your answers.

Additionally, your plots should be presentable to a professional audience.

* Add axis labels, legend labels, and titles. Captialize the important words in them.
* Categorical data should be ordered for ease of visualization (usually not alphabetical order)
* Text shouldn’t be overly large or tiny.
* Adjust the fig.width, fig.height, and dpi agruments so that the plot is reasonably sized.
* The plot type and layout should be chosen so that the data is easy to read.
* Adjust marker size or jitter to mitigate overplotting

Note 2: For the comprehension questions at the end of questions 1, 2, and 4, you may “hard code” your answers as simple text.

#### Preliminaries

We’ll use the data file hw2data.rda, which should be in the same directory as this markdown file (which should also be your working directory)

# load hw2data.rda from the working directory  
load('hw2data.rda')  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(plyr)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Unlike a CSV file, which only contains one data frame, an .rda file contains multiple R variables, and can only be opened by R. You should have the following data frames, all taken from the Consumer Expenditure Survey:

1. expenditure: A data frame of expenditure amounts by category for 4000 households
2. household: A data frame of household and demographic information (for the same 4000 households).
3. income: A data frame of income and job information (for the same 4000 households)

The column names for expenditure, household, and income are hard to decipher. So there are 3 more data frames, with explanations for the column names.

1. expenditure.key: Description of each column in expenditure
2. household.key: Description of each column in household
3. income.key: Description of each column in income

Finally, there are two more data frames:

1. item.purchases: Each row corresponds to a specific itemized purchase made by the households. The household ID, the cost, and a ucc code giving a description of the purchase is included.
2. item.key: Descriptions of the different ucc codes

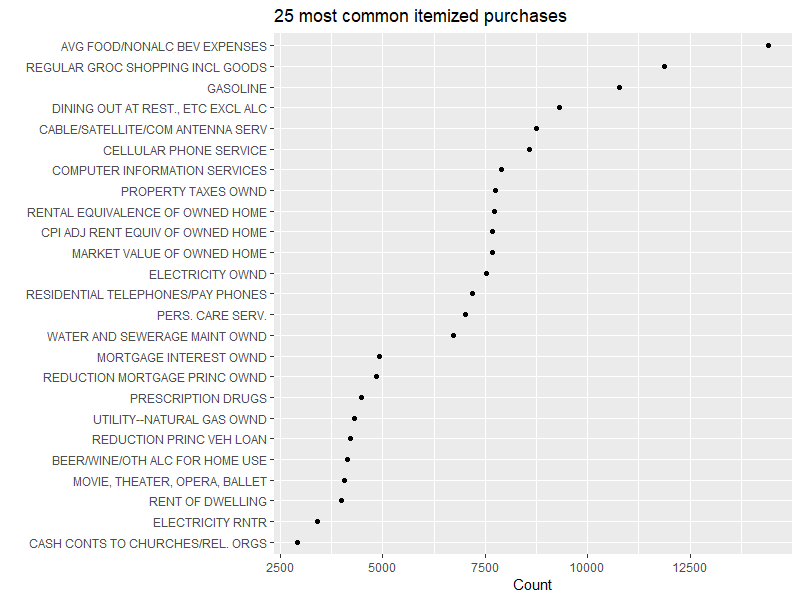
Note that in expenditure.key, several of the expenditure categories have numbers, giving the “level” of the category. For example food (1) can be divided into food at home (2) and food away from home (2). The latter subcategory can be further divided into food excluding meals as pay (3) and meals as pay (3).

#### Questions

**Problem 1:**

Make a dot chart showing the 25 most common itemized purchases, and how often they were reported in the data frame item.purchases. You should use the actual category descriptions in your chart, not the UCC codes.

# add code here with comments  
merge.item = merge(item.purchases, item.key, by = "ucc")  
# count most common itemized purchases  
count.item = ddply(merge.item, "descr", summarize, count = length(descr))  
#Reorder the description to graph properly  
arrange.item = count.item %>% arrange(desc(count)) %>% slice(1:25)   
#Plot the data frame  
ggplot(data = arrange.item, mapping = aes(x=count, y=reorder(descr, count))) + geom\_point() + labs(x="Count", y= "", title = "25 most common itemized purchases")



Were there any categories that you expected to see, but didn’t?

**Problem 2:**

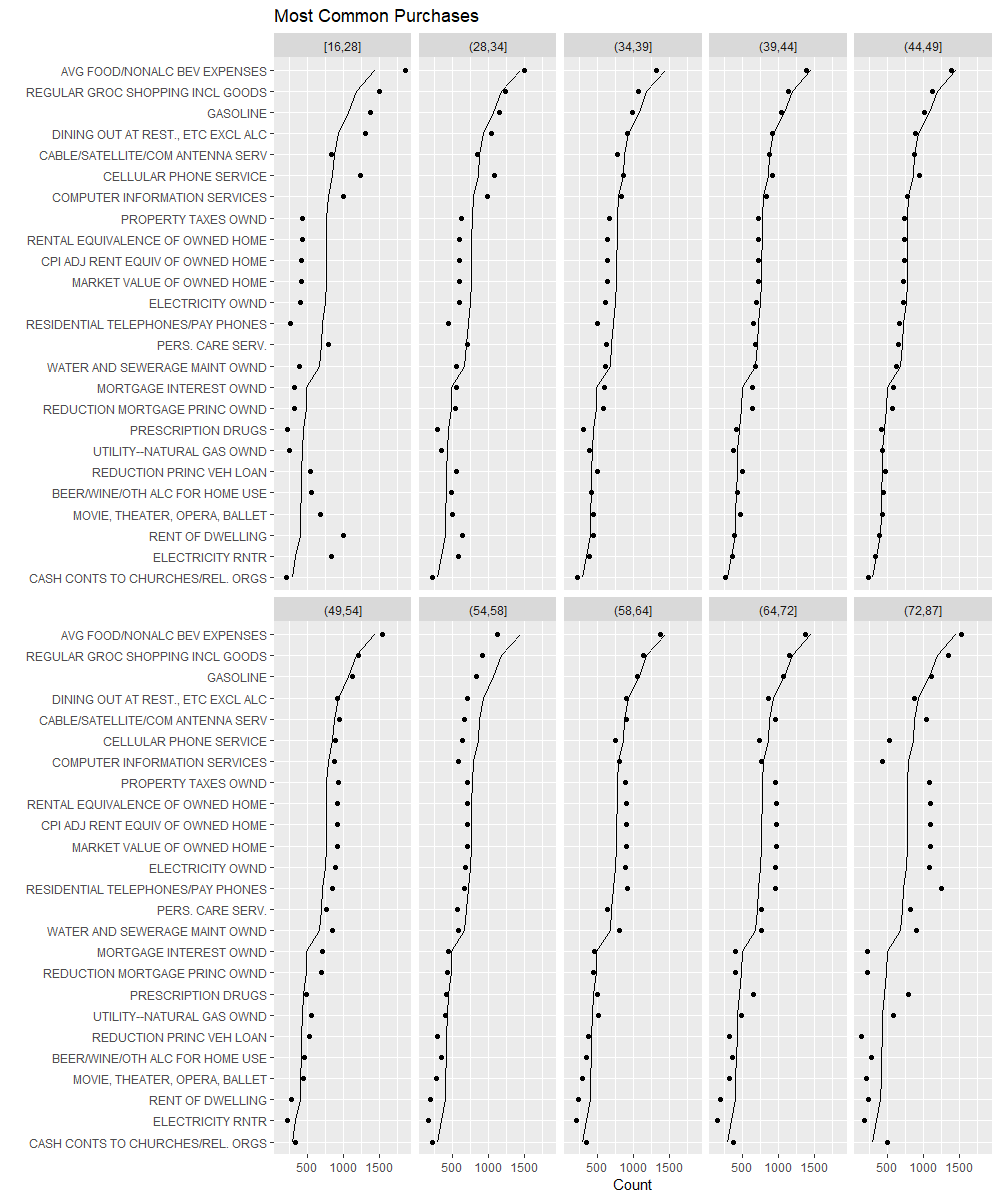
Make the same dot chart as before, but group the data into 10 equally sized groups according to the age of the reference person. For each age group, show the same 25 categories as you did in problem 1.

Use facet\_wrap() or facet\_grid() to show your results in 10 charts, one per age group.

To make comparisons easier, in each of your plots add a reference line showing the number of purchases for each category, averaged over all age groups.

(Hint: You can do this by adding a geom\_line() which has its own data frame and mapping. If this data frame doesn’t include the faceting variable, the line is plotted in all facets.)

# add code here with comments  
#Combine data frames to get all needed value  
hsehld = merge(item.purchases, household, by = "newid")  
household.descr = merge(hsehld, item.key, by = "ucc")  
household.descr = subset(household.descr, select = c("descr", "age\_ref"))  
household.descr = mutate(household.descr, age.category = cut\_number(age\_ref, n = 10))  
household.descr = subset(household.descr, subset = as.character(descr) %in% as.character(arrange.item$descr))  
count.household = ddply(household.descr, c("descr","age.category"), summarize, count = length(descr))  
#Plot the data frame  
plot = ggplot(count.household, mapping = aes(x=count, y = reorder(descr, count))) + geom\_point()   
#facets  
plot = plot + facet\_wrap("age.category", ncol = 5)  
household.descr.avg = ddply(count.household, c("descr"), summarize, avg.count = mean(count))  
plot = plot + geom\_line(household.descr.avg, mapping = aes(x = avg.count, y = reorder(descr,avg.count), group = 1)) + labs(x="Count", y= "", title = "Most Common Purchases")  
plot



For the youngest age group, which item categories are purchased with above average frequency? **Answer:** The item categories were the youngest group with above average frequency are shown in the table below: average food and non alcoholic beverage expenses For the oldest age group, which item categories are puchased with above average frequency? **Answer:** The item categories were average food and non alcoholic beverage expenses as shown in the plot. **Problem 3a:**

Create a data frame containing the age, job categories, and hours worked/week of the reference person who answered the survey, and also their spouse if there is one. So your data frame will have one row for each reference person, plus one additional row for each spouse. Your data frame should have three columns, age, job, and hours.worked.

If the age of the spouse is NA, you may assume the household is single.

The age of the reference person and spouse are contained in the fields age\_ref and age2 of the data frame household. The job categories of each person are occucod1 and occucod2, in the data frame income. The hours worked/week of each person are inc\_hrs1 and inc\_hrs2, also in income.

Hint: The command rbind(df1, df2) will manually stack data frames df1 and df2 on top of each other. However, they need to have the same column names.

# add code here with comments  
#merge household and job info  
person = merge(household, income, by = "newid")  
#get only the age, job, and hours  
ref.person = subset(person, select = c("age\_ref", "occucod1", "inc\_hrs1"))  
spouse.person = subset(person, select = c("age2", "occucod2", "inc\_hrs2"), subset = !is.na(age2))  
#rename the columns for as age, job, and hours worked  
names(ref.person)[ c(1,2,3) ] = c('age', 'job', 'hours.worked')  
names(spouse.person)[ c(1,2,3) ] = c('age', 'job', 'hours.worked')  
#combine the reference person and spouse job data  
combined.data = rbind(ref.person, spouse.person)

What is the mean age and hours worked for each job category? Produce a table or print the data frame. Order the job categories by the mean age

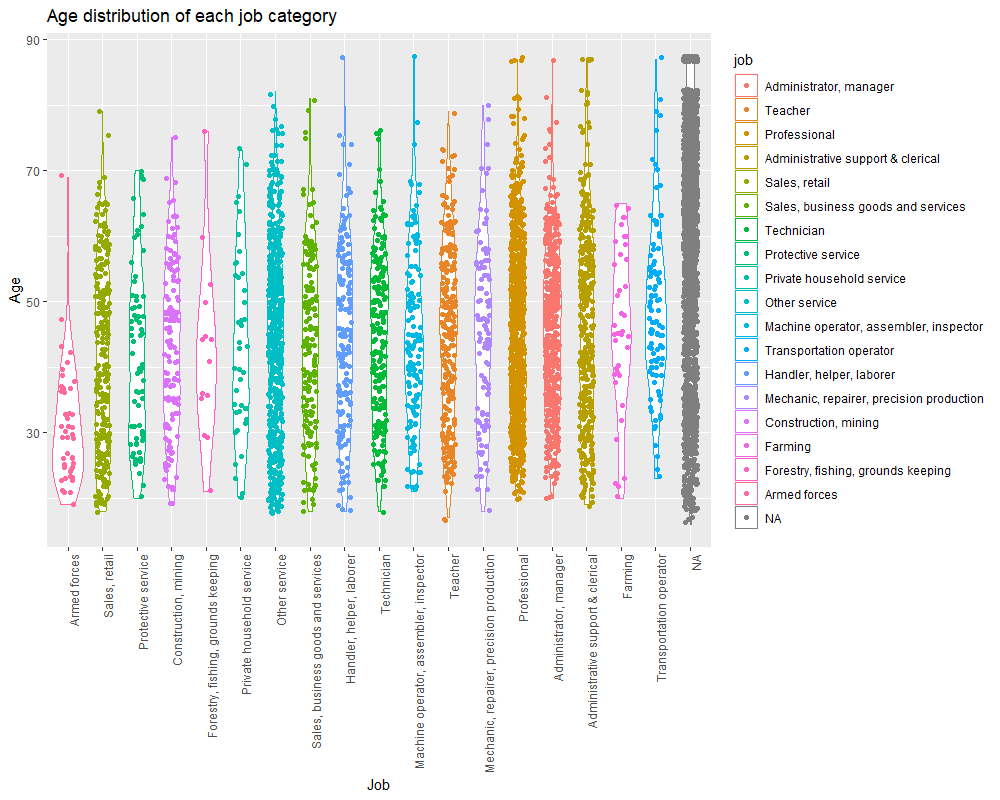
# add code here with comments  
average.job = ddply(combined.data, c("job"), summarize, avg.age = mean(age), avg.hours.worked = mean(hours.worked))  
average.job = arrange(average.job, desc(avg.hours.worked))  
average.job

## job avg.age avg.hours.worked  
## 1 Armed forces 30.97727 47.77273  
## 2 Transportation operator 48.96000 44.61000  
## 3 Administrator, manager 45.58985 44.14588  
## 4 Machine operator, assembler, inspector 44.53788 42.50758  
## 5 Mechanic, repairer, precision production 44.88350 42.25243  
## 6 Farming 46.02564 42.05128  
## 7 Professional 45.48276 41.76085  
## 8 Protective service 41.92857 41.44286  
## 9 Sales, business goods and services 43.69325 41.38650  
## 10 Forestry, fishing, grounds keeping 42.85714 40.71429  
## 11 Technician 44.46798 40.59606  
## 12 Construction, mining 42.65185 40.18519  
## 13 Teacher 44.80534 38.27481  
## 14 Handler, helper, laborer 44.37222 37.87222  
## 15 Administrative support & clerical 45.97105 37.39737  
## 16 Other service 43.49374 36.77281  
## 17 Sales, retail 41.11741 35.70040  
## 18 Private household service 43.48780 35.09756  
## 19 <NA> 59.80759 NA

**Problem 3b:**

Use geom\_violin() to show the age distribution of each job category. Additionally, plot the raw data points on top of the violin plots for reference. You may want to read the online ggplot2 documentation for geom\_violin().

# add code here with comments  
  
ggplot(combined.data, aes(x = reorder(job, age), y = age, color = job)) +  
 geom\_violin() +  
 geom\_jitter(shape=16, position=position\_jitter(0.2)) +  
 labs(x = "Job" , y = "Age", title = "Age distribution of each job category") +   
 theme(axis.text.x = element\_text(angle=90, vjust=1, hjust=1))

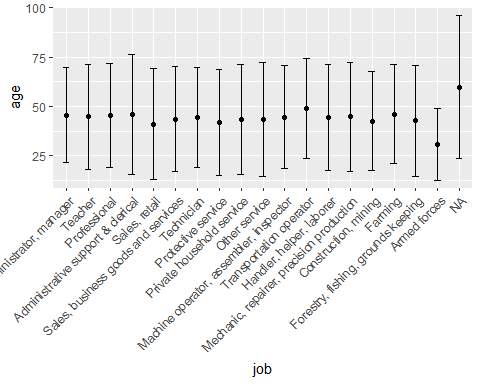


**Problem 4**

Using the data frame you made in Problem 3a, make a plot showing the average age of each job category, and with error bars or point ranges indicate +/- 2 standard deviations (stat\_summary() with the function mean\_sdl() can compute this for you). Leave out the fields with missing job category.

# add code here with comments  
  
ggplot(combined.data, aes(x=job, y=age)) + stat\_summary(fun.data = mean\_sdl, fun.args = c(mult=2), geom='errorbar', width=.3, na.rm = TRUE) +  
 stat\_summary(fun.y = mean, geom='point', na.rm = TRUE) +  
 theme(axis.text.x = element\_text(angle=45, vjust=1, hjust=1))

## Warning: `fun.y` is deprecated. Use `fun` instead.



Which job category had the highest average age? Which job category had the highest variance in reported age?

**Answer**: As shown from the plot the category with the highest average age is the single households. The category with largest variance is also the single households.