Name: Shawn Petersen
Email: Shawn.Petersen@wsu.edu
Cougar ID: 011691731
Class: STAT 419
Instructor: Monte J. Shaffer
Date: 08/27/2020

1**. Where did the VLSS data come from? Do some research and provide a URL for a link to the official page with the data. Describe how you found it. How much does it cost to purchase?**

The Vietnam Living Standard Survey(VLSS) was conducted by the State Planning Committee (SPC) and General Statistical Office (GSO) in 1992/93 and 1997/98. The survey was funded by the UNDP and Swedish Internation Development Authority(SIDA). The data can be obtained from the GSO in Hanoi, Vietnam. The request needs to include a one-page explanation of the proposed research. The fee for the data varies depending on if a Vietnam citizen or foreigner and if you want one or both data sets. The fee for a Vietnam citizen ranges from $100- 200 and $500-$2000 for a foreigner. The range is depends on if you're an individual or an organization. The data and information can be obtained from, https://microdata.worldbank.org/index.php/catalog/2694

2. **How were the 3 research questions derived? Are they constrained by the data? If so, how should you derive research questions?**

The 3 research questions were derived from the two data sets provided, (VLSSage.dat and VLSSperCapita.txt). The two data sets contain age and capita so the research questions were constrained by the limited data sets provided to the user.

When you are analyzing data the quote, "Data is only as good as the questions", is very true. To derive questions, we need to ask, "Who are the final users of our results" and "What do we want to find out"? For this data set and questions, the Vietnam government wanted to know, "Were the policies and programs that were currently available age-appropriate for the population"?

3**. Review the different graphs and the R code to generate them. From Figure 1.6, is there evidence to conclude that Urban homes have higher expenditures than Rural homes? How would you logically defend your conclusion?**

Figure 1.6 is a density plot showing the distributions of the household per capita expenditures for the urban and rural populations in Vietnam. The red line is the poverty line drawn vertically at $119.32. The plot clearly shows the “Rural” population has a lower expenditure than the “Urban” household. To defend our conclusion, we need to further analyze the data using either the Winsoried for Trimmed mean methods to summarize capita expenditures by a group. These methods were used because both population density plots are skewed, have tails.

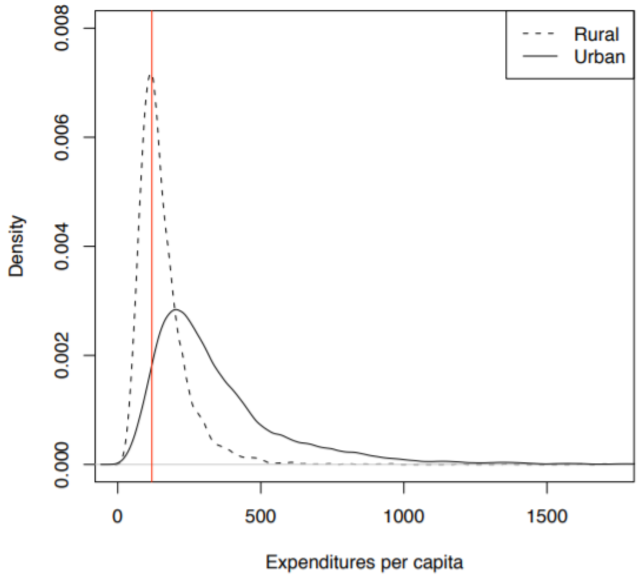


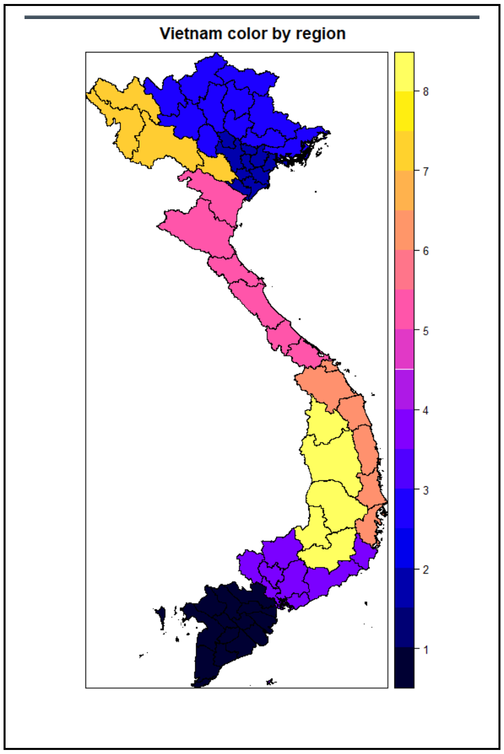
Figure 1.6: Kernel density estimate of the rural and urban population distribution of household per capita income.

4. **How was Figure 1.7 plotted? What was the R code to do this?**

Figure 1.7 is a map of Vietnam color-coded by the seven different regions. R has libraries, “map”, “map data”, and “GADM” that can be integrated to plot most countries, states, providences, and regions.

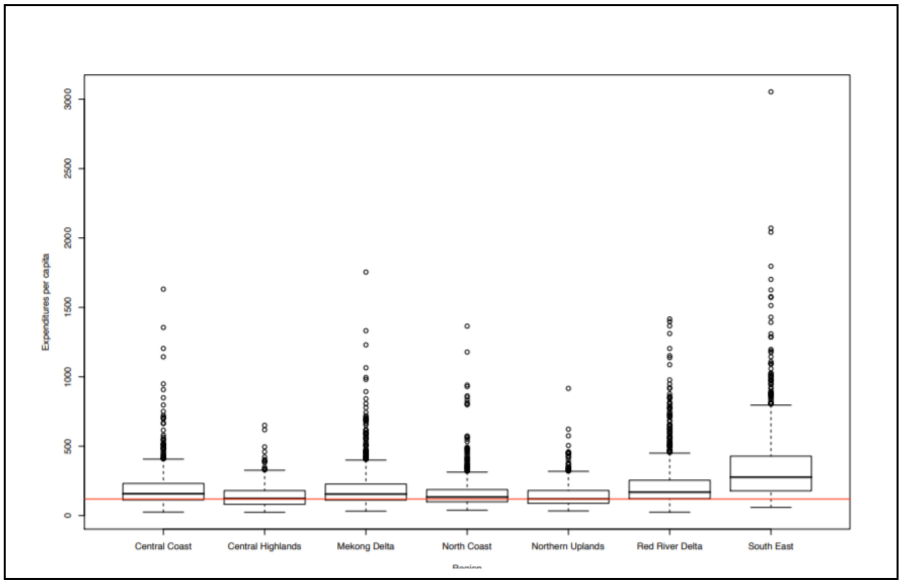
The R code and as of today, Vietnam has 8 regions so my code plotted 8 separate regionaly colors:

```
install.packages("maps")
install.packages("mapdata")
install.packages("maptools")

library(maps)
library(mapdata)
library(maptools)

libs <- c("maptools", "sp")
lapply(libs, require, character.only = TRUE)

geo63 <- readRDS("geo63.rds")
spplot(geo63, "ID_1", main="Vietnam color by region")
```

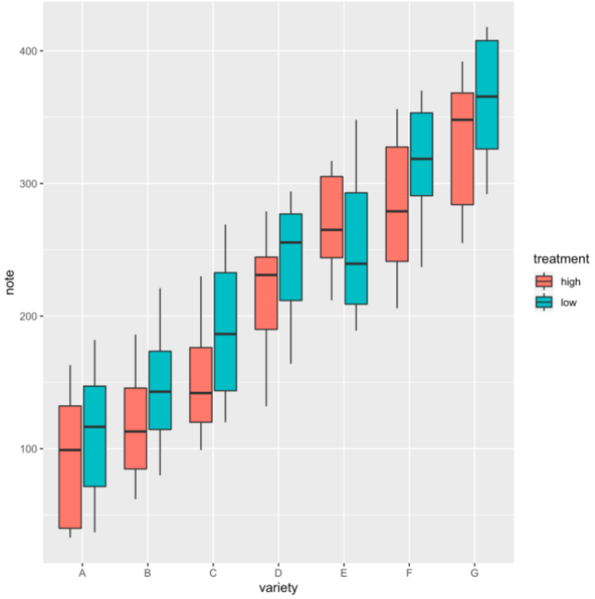


5. **From Figure 1.8 and Figure 1.9, can we conclude that the South East region has higher expenditures than the other regions? Would it be possible to graph similar plots of the data by both region (7 choices) and by Rural/Urban (2 choices)?**

Below is the capital expenditures box plot by region. The red horizontal line is the poverty level of $119.32. It's clear the South East Region expenditures are higher than the other regions in Vietnam.



Yes, R has a ggplot package that can plot boxplots by groups an subgroups. Example in the plot below. We could update the plot of expenditures by region to include an additional sub-group ("Rural", "Urban")


note
A
B
C
D
E
F
G
variety
treatment
high
low