Kevin Black – kevin.black@wsu.edu – 11480710
STAT 419 – Instructor Monte J. Shaffer – 08/25/2020

Reading Assignment: "Exploratory Data Analysis"

Q1. *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? If you can find an online copy of the VLSS data, please also provide a link.*

A1. The 1997-98 VLSS data came from the Vietnamese government, specifically from a study conducted by the General Statistical Office (GSO) in order to have relevant data for monitoring living standards and designing/evaluating policies and programs. As this survey was part of the Living Standards Measurement Study (LSMS) household surveys along with technical assistance from the World Bank, the official page containing the data description of this second survey can be found on their website at https://microdata.worldbank.org/index.php/catalog/2694/study-description. This was found simply by searching "Vietnam living standard survey" on Google, with this website being the second result. However, the only way I found to access the raw data was by selecting "Get Microdata" on the page link above, of which a PDF is then downloaded giving instructions on how to purchase the dataset. There are a variety of fees depending on where the buyer is situated, but for me (citizen of a developed country), this dataset would probably cost $500 USD.

Q2. *How were the 3 research questions derived? Are they constrained by the data? If so, how should you derive research questions?*

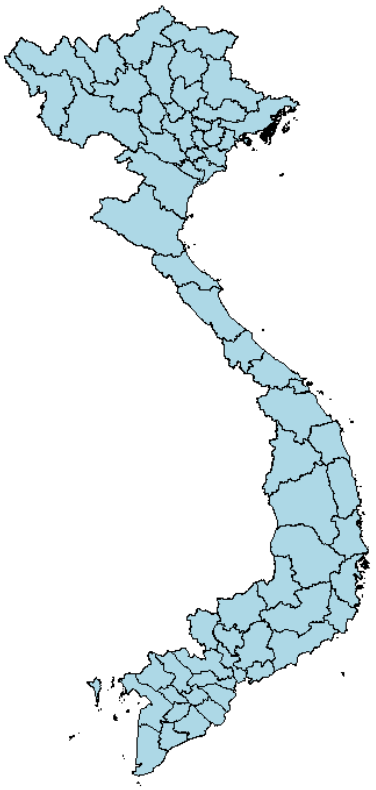
A2. It seems that the three research questions presented early on in the reading were derived from the author wanting to discern any patterns in the frequency of certain age groups or fertility rates, as well as to analyze how much of the country's expenditures are handled by the two predominant environments spread throughout Vietnam, urban and rural. However, the data definitely had some part in constraining and determining these research questions when the process, ideally, should have been reversed. Research questions are typically formulated first before any of the data is presented or collected. This is because when one asks a focused, feasible, and specific research question, they would then go out and collect that data to validate or dismiss their inquiries. Then after the data is collected, follow-up questions can be asked to resolve patterns in the dataset, just like the author has done.

Q3. *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?*

A3. Based on Figure 1.6, there is sufficient evidence to conclude that Urban home expenditures per capita are higher than the Rural home expenditures per capita. One can see this in the kernel density estimate, where under the Urban curve, a larger area seemingly covers the

higher ranges of expenditures per capita at first glance, and thus a larger percentage of the population. However, quantifying this difference between the two types of homes must be calculated with robust estimators and likely cannot be seen just by looking at the graph. This is even verified in the reading when the winsorized and trimmed means are introduced and calculated, showing that on average there is a difference between Urban home and Rural home expenditures, with Urban homes having the higher expenditures.

Q4. *How was Figure 1.7 plotted? What was the R code to do this?*



A4. As there was no included R code for Figure 1.7 or any indication to how it was plotted, through some online research I attempted to recreate the map using the R code below, albeit not the exact same. Recreating the exact same map would require having access to the *regionName* column of the exact data file used in the reading, *VLSSperCapita.txt*, for color coding regions depending on the value of *regionName*. Unfortunately, I could not find that file online without being stuck behind some form of paywall. Thus, I was able to make a map of Vietnam to the right by utilizing the *readOGR()* function of *rgdal* and the shapefile *gadm36_VNM_1*, which is a level-1 shapefile containing the provinces of Vietnam. This level-1 shapefile and others can be downloaded from https://gadm.org/download_country_v3.html.

```
library(rgdal)
shp=readOGR(dsn=".",layer="gadm36_VNM_1") # shpfile in current dir
plot(shp,col='lightblue') # plot the shpfile as a light blue region
```

Q5. *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)?*

A5. Figure 1.8 depicts the expenditures of each of the seven regions of Vietnam, but it may not be immediately obvious which region due to the outliers and scale of the boxplot. However, Figure 1.9 remedies this by introducing means and error bars for the data, which gives us a much clearer visual for the breakdown of average expenditures per capita by region. From here we can conclude that the South East region definitely has the highest expenditures per capita of

all the regions in Vietnam. The boxplots in both figures are already graphed by region, but yes it would be possible to graph similar plots of the data by Rural/Urban as well. For example, to recreate Figure 1.8, you could use the R code provided but replace the grouping variable *regionName* with *urban* and the x-axis label with *Environment* for clarification. Similar work can be performed to the provided Figure 1.9 code to plot means and error bars for Rural/Urban as well.

Figure 1.8 Revised R Code

```
boxplot(dollar ~ urban, xlab="Environment",
      ylab="Expenditures per capita")
abline(h=119.32, col="red")
```

Figure 1.9 Revised R Code

```
library(gplots)
plotmeans(dollar ~ urban, data=household, connect=F, p=.95, n.label=F,
      xlab="Environment", ylab="Expenditures per capita")
```