Jailee Foster
jailee.foster@wsu.edu
WSU ID: 11439720
Stats 419
Instructor: Monte J. Shaffer
August 31, 2020

Reading Assignment: “Exploratory Data Analysis”

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? [Please don't buy it.] If you can find an online copy of the VLSS data, please also provide a link.

The data for the Vietnam Living Standards Survey (VLSS) was collected by the State Planning Committee and the General Statistical Office as part of the Living Standards Measurement Study.

The url for the official page with the data is https://microdata.worldbank.org/index.php/catalog/2694/get-microdata. In order to find this, I conducted a google search for “Vietnam living standard survey data”. The first result that came up, “Living Standard Survey” at www.gso.gov.vn took me to an error page, but the second result, “Vietnam – Living Standards Survey 1997-1998” took me to a page where I had the option to purchase the microdata. This data has different costs depending on where you are located and whether you are an individual or an organization. For me to buy it for personal use, it would be $500 and for a US organization it would be $2,000.

I was not able to find an online copy of the VLSS data in my search.

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

The research questions were derived by the Vietnam government in order to monitor living standards and evaluate the government policies and programs that were currently in place. Two key components in creating these research questions were whether the policies and programs made sense for the age distribution of the population, and whether the living conditions were fair across all regions of Vietnam.

It seems that the research questions are not necessarily constrained by the data. The only limiting factor that sticks out to me is the data that the research agency chose to collect. This will always be a “constraint” in creating research questions, but it is one that is able to be worked around. I believe that this is one advantage to collecting data unique to an analysis, although time consuming, because the data that is needed may not always be readily available from other sources.

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?

> From Figure 1.6, there is evidence to conclude that Urban homes have higher expenditures than Rural homes. Looking at the density function for expenditures per capita for Rural homes, it appears that nearly half of the data lies below the poverty line (shown in red on the graph). This density function also peaks very early and is very narrow, suggesting that there is not much spread in the data. Conversely, the density function for the expenditures per capita of Urban homes is mostly above the poverty line. This graph does not have as steep of a peak as the density function for Rural homes, suggesting that the data is more spread out. It also appears that almost half of the data for Urban expenditures is higher than the highest Rural expenditures.

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

> After doing some research, it does appear that Figure 1.7 could be plotted using R, although the article does not include the exact code to do so. I found code to create Choropleth Maps, which are maps that are shaded to represent certain average values in certain areas, in R at https://rkabacoff.github.io/datavis/datavis.pdf. Although this is not exactly what has been done in Figure 1.7, which is shaded by region, it is similar, and it leads me to believe that the figure is attainable using R code.

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)?

> Based on Figures 1.8 and 1.9, I do think that it is valid to conclude that the South East Region has higher expenditures than the other regions. In Figure 1.8 we can see that not only is the median expenditure per capita higher, but it appears that the minimum, first quartile, third quartile, and maximum expenditure per capita, along with the outliers are all higher than the other regions. Looking at Figure 1.9, it is very clear that the mean expenditure per capita is significantly higher in the South East than it is in the other regions.
>
> It would be possible to graph similar plots of the data by both region and by Rural/Urban. There are multiple different ways to do this, but one specific way that I found with research is to create a grouped boxplot where there are groups (region) and subgroups (Rural/Urban) on the same plot (https://www.r-graph-gallery.com/265-grouped-boxplot-with-ggplot2.html). It appears that the same thing can be done with the means and error bars plot.