Connor StarrHurst

connor.starrhurst@wsu.edu

11596221

STATS 419

Instructor: Monte J. Shaffer

8/29/2020

Read "Exploratory Data Analysis" (01_reading.txt)

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 came from the second Vietnam Living Standards Survey done in 1997-98. Two files are used, VLSSage.dat and VLSSperCapita.txt, from https://microdata.worldbank.org/index.php/catalog/2694. I found this URL by Google searching "Vietnam Living Standards Survey 1997" and it costs $500 for one dataset for a citizen of a developed country.

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

> They were derived from the purpose of the study and from the VLSSage.dat and VLSSperCapita.txt datasets. They are constrained by the data. Research Questions should be derived based on the purpose of the study and be limited by what the data covers. This is because asking questions the data cannot answer is pointless.

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?

> Based on Figure 1.6, I can conclude that Urban homes do have higher expenditures than Rural homes because the Urban Expenditures per capita are shifted more to the right (around 100 - 500) compared to Rural Expenditures per capita (around 0 - 250)

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

> It was plotted using a latitude and longitude points that are mapped onto x and y axes. The reading doesn't give any code for how Figure 1.7 was plotted, but after doing some research I found several examples of similar maps. They all use the maps and ggplot2 libraries along with the functions ggplot(), geom_polygon(), and aes(). Below is a simple example from https://socviz.co/maps.html by Kieran Healy that plots a US map with each state being a different color
>
> ```
> p <- ggplot(data = us_states, aes(x = long, y = lat, group = group, fill = region))
> ```
>
> ```
> p + geom_polygon(color = "gray90", size = 0.1)
> ```

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)?

> Yes and yes. The South East region has the most Expenditures per capita overall (Figure 1.9) and on average (Figure 1.8). We could create 4 more graphs that look similar to Figure 1.8 and 1.9 that are subsets based on whether or not the area is Rural or Urban.