

Title: Reading Assignment "Exploratory Data Analysis" (10 points)

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 VLSS data come from the Vietnamese population from surveys. The data of households were used in policy design, monitoring of living standards and evaluation of policies and programs. And it was conducted by the State Planning Committee along with the General Statistical Office.

<https://datacatalog.worldbank.org/dataset/vietnam-living-standards-survey-1992-1993>

I found the link by doing a google search: VLSS Vietnam Data. And the first site that comes is this one, the World Bank. When trying to access the data it is possible to get the documentation as a pdf.

<https://microdata.worldbank.org/index.php/catalog/1910/related-materials>

And the Metadata it is possible to get as DDI/XML or JSON on the page in the upper right.

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 Vietnamese researchers and government in order to see the policy design, living standards and policies and programs. The goal of one evaluation for example was to identify whether the policies and programs that were currently available were age appropriate for the population. Therefore, they needed to know what the age distribution for the Vietnamese population is.

I think that depending on what data is collected you come up with research questions that can be answered by the data. That is any interesting fact that can be found out from the data. So yes, the research questions are constrained by the data.

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?

-> Yes, there is evidence to conclude that Urban homes have higher expenditures than Rural homes. When looking at the graph of figure 1.6 then I see that the curve of the Rural homes has a higher peak than the Urban homes, which is located at a lower amount of expenditures per capita. This method uses the sample mean. It means that a higher number of households has low expenditures at Rural homes. Whereas Urban homes have more spread expenditures at higher amounts. When using more robust estimators like the Winsorized Mean of 20% for example, then we also get, that Urban homes have higher expenditures than Rural homes.

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

Figure 1.7 was plotted using the data given, and their different regions in the data.

```
data <- read.table("VLSSperCapita.txt", header=T)
```

```
# Retrieve the map data
```

```
map <- map_data("vietnam", region = data.region)

# Compute the centroid as the mean longitude and latitude
# Used as label coordinate for country's names

region.lab.data <- map %>%

  group_by(region) %>%

  summarise(long = mean(long), lat = mean(lat))

ggplot(map, aes(x = long, y = lat)) + geom_polygon(aes(
  group = group, fill = region))+ geom_text(aes(label =
  region), data = region.lab.data, size = 3, hjust = 0.5)+
scale_fill_viridis_d()+ theme_void()+
theme(legend.position = "none")
```

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, we can conclude that the South East region has higher expenditures than the other ones, which are fairly similar. It is because the mean and the entire boxplot (Figure 1.8) and the error bar (Figure 1.9) of the South East region is located at higher expenditures.

I think that it would be possible to graph similar plots of the data by both region and by Rural/Urban.