GEA1000 QUANTITATIVE REASONING WITH DATA
Tutorial 3

Please work on the problems before coming to class. In class, you will engage in group work.

1. Download the data set HDB20to21.csv from LumiNUS. This data set is made up of the HDB resale records for the years 2020 and 2021. Note that on top of the variables in the standard HDB data set, we included a new variable "period", with categories 2020H1, 2020H2, 2021H1 and 2021H2. For example, 2020H1 corresponds to all the rows of data from January to June 2020 and 2020H2 corresponds to all the rows of data from July to December 2020.

a) Using suitable plot(s), show the distributions of resale prices of flats sold in Hougang, over the four periods January to June 2020, July to December 2020, January to June 2021, and July to December 2021. With the help of suitable summary statistics, what can you say about the trend of prices over the two years?

Spread → compare IQR. increasing. median. Mode ≈ 400000 - 500000. right-skewed.

b) What can you deduce when comparing the resale prices in Hougang and in Bukit Batok over the same time periods?

Hougang > Bukit Batok.

2. Answer this question using the data set exercise.csv from LumiNUS. This set of 195 data points provides the age, average number of hours of exercise per week and the chance of developing a particular disease, for each person.

a) Describe the association between the chance of developing a disease and the average number of hours of exercise per week using suitable tools. If a person exercises 6 hrs per week on average, what is the predicted average chance of developing a disease? What if a person exercises 9 hrs per week on average?

linear regression. → 29.697 %

~~45.184%~~ out of range. weak increasing trend r = 0.2[illegible]

b) Comment on any similarity or difference in the association between the chance of developing a disease and the average number of minutes of exercise per week with the association that you obtained in part (a). Justify your answer.

Same. ⇒ correlation won't change. ⇒ multiply by 60.

c) Investigate the correlation between the chance of developing a disease and the average number of hours of exercise per week, for people below 50 years old and for people at least 50 years old separately. Comment on any similarity or difference in the relationship for the different age subgroups, as compared with that obtained in (a).

at least 50. correlation = −0.84.

young correlation: - ve.

association becomes negative.

Simpson's paradox

recode:
50: hi = "old";
else: "young"

3. Download the data set county_complete_2019.csv from LumiNUS. The information pertaining to some socioeconomic and demographic indicators for 1266 counties in the United States are given in this data set. A description of the variables involved is given below:

Variable	**Description**
state	state
name	county name
pop2019	population in 2019
age_over_65_2019	percentage of population 65 and over (2015-2019)
households_speak_spanish_2019	percentage of households that speaks Spanish (2015-2019)
hs_grad_2019	percentage of population 25 and older that is a high school graduate (2015-2019)
bachelors_2019	percentage of population 25 and older that earned a Bachelor's degree or higher (2015-2019).
household_has_computer_2019	percentage of households that have desktop or laptop computer (2015-2019).
poverty_2019	percentage of population below poverty level (2015-2019)
uninsured_2019	percentage of civilian noninstitutionalized population that is uninsured (2015-2019).
unemployment_rate_2019	unemployment rate among those ages 20-64 (2015-2019).
per_capita_income_2019	per capita money income in past 12 months (2015-2019)
median_household_income_2019	median household income (2015-2019)

a) How would you predict the median household income in the years 2015-2019 based on the percentage of population below the poverty level? linear regression on median household income vs poverty.

b) Can you identify a confounder in the relationship between the variables median_household_income_2019 and poverty_2019? Justify your answer.

unemployment rate?,