U6614: Assignment 2

your name (your uni)

2026-01-15

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on the due date.

Before knitting your rmd file as a pdf, you will need to install TinyTex for Latex distribution by running the following code:

```
tinytex::install_tinytex()
```

Please visit this link for more information on TinyTex installation.

1 Load and inspect ACCESS 2018 data:

1a) Inspect the data frame and state the data types for each column.

- make sure to inspect the age, gender, caste, income source, state & educ columns

1b) Use the mutate function to create new column for gender:

- gender.fac = as.factor(gender),
- check if it worked, show your validation

1c) In a single pipe, include gender.fac in a new data frame called access2018.temp1.

- also create factors for caste, income source, state and education
- exclude the columns for HHID and date after creating access2018.temp1
- print the first 5 observations

1d) Inspect caste.fac, gender.fac, incsource.fac, education.fac and state.fac using the levels() function. What package is the levels() function located in?

1e) Use filter() to only include rows only for the state "Uttar Pradesh."

- store as a new object access_UP
- print the first 5 observations
- confirm your data only includes observations for Uttar Pradesh

1f) What is the female share of respondents in Uttar Pradesh? Compare it to the female share of respondents in Bihar?

1g) Remove the access2018.temp1 object from memory using the rm() function.

2 Describe the access_UP data frame

2a) What is the unit of observation (or unit of analysis)?

2b) How many individuals are observed?

- how many of these are from 'Reserved' castes (i.e: Scheduled Caste, Scheduled Tribe, Other Backward Class")?
- how many are "General" caste members? [**Note: for this question, use the 'table' and 'prop.table' functions. 'table' will give the number of observations belonging to each caste category.'prop.table will give the proportion**]

2c) Do you think the sample is representative of all households in India?

- how would you describe the population represented by this sample?
- is there more information you would like to see to assess the representativeness of the sample?

2d) What is the average age of individuals in the sample? youngest and oldest person?

3 Earnings per week for different groups in Uttar Pradesh

3a) Find the observation for the top monthly expenditure using the summarize() function.

- assign this to a new object called max_exp_obs1

3b) Find max monthly expenditures using the arrange function instead of summarize.

3c) Use the filter function to subset for the observation with max monthly expenditure.

- (don't hardcode the max expenditure to filter on, refer to the max_exp_obs1 object from a)
- store in new data frame max_exp_obs2
- confirm it worked

3d) What is the age, gender and caste of the top monthly spender in the sample?

3e) List the age, gender and caste of the top 10 monthly spenders in the sample.

3f) How many individuals spend more than the mean monthly expenditure amount of the sample?

Caste-based monthly expenditure gaps in Uttar Pradesh.

4a) Use the filter function and 'reserved' dummy variable to subset observations belonging to the 'General' caste.

- assign to new data frame, access_UPgen
- sort in descending order of monthly expenditure
- check if it worked

4b) Repeat part a, but this time, use the filter function to subset observations belonging to 'Reserved" castes, and assign them to a new data frame called 'access_UPres'.

4c) Use summarise to find mean, min & max monthly expenditure for the General category and Reserved category of castes, separately. Name each statistic appropriately (i.e. name each column in the 1-row table of stats). What is the gap in mean monthly expenditure between the two groups?

4d) Research suggests that people belonging to the 'General' caste category own 65% of agricultural land in India (India Human Development Survey, 2020). They are also more likely to own land in a proportion that is much higher than their share of the population. We will now use this data to compare 'General' versus 'Reserved' households using the 'reserved' dummy variable to understand whether this is true.

- use the 'table' and 'prop.table' functions to explore the total number of landowners by their caste.
- here, 'landowners' are people whose primary source of income is agriculture on their own land. So, you can use the 'incsource.fac' variable to filter the 'access_UP' data frame for 'Agriculture (own land)'
- does the number of General and Reserved caste landowners seem proportional to the overall number of 'General' and 'Reserved' caste individuals in Uttar Pradesh's population (as calculated in Q2 b)?

4e) Do differences in landowning explain the expenditure gap between General and Reserved castes? i.e., does the General-Reserved expenditure gap persist among landowning families? What is the gap in mean monthly expenditures of General vs Reserved category landowners?

4f) Is there a gap between the mean monthly expenditures of General caste male and Reserved caste male landowners? What about the same, but for the female landowners?

4g) Does educational attainment explain part of the expenditure gap between General and Reserved caste landowners? What is gap between landowning General vs. Reserved household with a HS degree or more?