R Test

Problem Description

You are given **R\_test\_dataset.csv** with the following description of the variables and labels of some of the variables.

**Variable description Table:**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| SbjNum | Interview ID |
| Start\_Time | Start Time |
| Q\_2 | Date of Interview |
| Interviewer | Enumerator Name |
| Ward | Ward |
| S2 | What is your age?(Grouped) |
| S3 | What secondary school certificate do you have? |
| S4 | Do you have, or are you currently enrolled in, a tertiary school certificate? |
| S5\_a | Do you intend to seek further education (e.g. diploma or bachelor's degree)? |
| Social\_Class | Social class |
| A1\_A | First Name |
| A1\_B | Family Name |
| A2 | What is your age? (Specific) |
| A5 | What is your region of origin? |
| Q\_25\_S | What is your region of origin? [Other (specify)] |
| A8 | What is your current marital status? |
| A9 | Do you have any children? |
| A10 | How many people do you live with (both adults and children)? |
| A12 | What is your relationship to the head of the household? |
| Q\_32\_S | What is your relationship to the head of the household? [Other (specify)] |
| B3 | Which school did you attend for secondary school (S5-S6/A Levels)? |
| B4 | Was your school an all secondary girls school? |
| B8 | What level of education do you feel is next in your career path? |
| B16 | How much are you willing to pay for your next degree or diploma? (Average tution per semester) |
| B17 | Currently, how career ready are you? |
| C1 | Have you been employed in the last 12 months? |

Value labels

|  |  |  |
| --- | --- | --- |
| **Variable Values** | | |
| Value | | Label |
| Interviewer | 1 | Menya Abdmajid |
| 2 | Bbale Denis |
| 3 | Muwonge Allan Joshua |
| 4 | Wambi Ken Paul |
| 5 | Wabwire Thomas |
| 6 | Muhindo Wilfred |
| 7 | Ahumuza Owen |
| 8 | Mirembe Mary |
| 9 | Arinitwe Mackline |
| 10 | Alum Maria |
| 11 | Kasule violet |
| 12 | Aweko Monica |
| 13 | Nabbumba Pennina |
| S2 | 1 | 17 and below |
| 2 | 18-25 |
| 3 | 25-30 |
| 4 | 31 and above |
| S3 | 1 | None/did not complete |
| 2 | Uganda Certificate of Education (UCE) 1-2 passes |
| 3 | Uganda Certificate of Education (UCE) 3-5 passes |
| 4 | Uganda Certificate of Education (UCE) 5 or more passes |
| 5 | Uganda Advanced Certificate of Education (UACE) – 1 principa |
| 6 | Uganda Advanced Certificate of Education (UACE) 2 principal |
| 7 | Uganda Advanced Certificate of Education (UACE) 3 principals |
| S4 | 1 | Yes |
| 2 | No |
| S5\_a | 1 | Yes |
| 2 | No |
| 3 | I don't know |
| Social\_Class | 1 | Social Class D |
| 2 | Social Class C2 |
| 3 | Social Class C1 |
| 4 | Social Class B |
| 5 | Social Class A |
| A5 | 1 | Central |
| 2 | Western |
| 3 | Eastern |
| 4 | Northern |
| 5 | Other |
| A8 | 1 | Single |
| 2 | Married |
| 3 | Divorced or Separated |
| 4 | Widowed |
| A9 | 1 | I do not have children |
| 2 | I have 1 child |
| 3 | I have 2 children |
| 4 | I have 3 children |
| 5 | I have more than 3 children |
| A10 | 1 | I live alone |
| 2 | I live with one other person |
| 3 | I live with 2-3 other people |
| 4 | I live with 4-6 other people |
| 5 | I live with 7 or more people |
| A12 | 1 | I am the head of the household |
| 2 | I am the wife of the head of household |
| 3 | I am the daughter of the head of household |
| 4 | I am a relative of the head of household |
| 5 | I am a friend of the head of household |
| 6 | Other: |
| B4 | 1 | Yes |
| 2 | No |
| B8 | 1 | Diploma |
| 2 | Technical certificate |
| 3 | Bachelor's degree |
| B17 | 1 | Extremely ready |
| 2 | Somewhat ready |
| 3 | Neutral |
| 4 | Not ready |
| 5 | Definitely not ready |
| C1 | 1 | I have been employed (for example in an office, restaurant, |
| 2 | I am self-employed (for example, on my own farm or family's |
| 3 | I have been employed as an unpaid intern |
| 4 | I have been engaged in household activities including agricu |
| 5 | I have not been employed |

Using **R statistical software** and **R\_test\_dataset.csv,** answer the following questions:

1. Describe some of the checks you would do on a dataset in readiness for analysis **(10mrks).**

* Missing values: Check if there are any blank cell in the dataset to avoid the output NA in the R analysis.
* Duplicates: Check if the cells or columns have been duplicated or there are any repeated values in the data set.
* Type of data: Check whether the data is discrete or continuous so that it will be easier to identify the distribution of the data.
* Distribution of the data: This helps identify which distribution will be used in analysis, for example binary is binomial distribution and so some tests such as t test will not be carried out on the data.
* Outliers: These are values that are not in the range of the expected values, either they are too high or too low compared to the threshold.

1. Define a complete interview **(2mkrs).**

* Give a brief introduction of oneself.
* Explain to the interviewee what the interview entails. Could be detailed or brief depending with the literacy of the interviewee and/or the familiarity of the topic in question.
* Give clear instructions of what is required from the interviewee.
* Allow the interviewee some time to answer the questions. Assist the interviewee (without bias) incase anything outlined in the questions is not clear.
* Give a rough estimate period of time of when they should expect results from their answered questionnaires.

1. What is the difference between a raw and a ‘dirty’ dataset **(4mrks)?**

Raw data is data collected by a researcher from the field regarding a topic or subject of their interest and then taken for analysis to give expected results whereas dirty data is already collected data that has not been cleaned to be fit a good statistical analysis.

1. You are doing Quality check on data and you have been asked to prepare a callback sheet to be sent to team leaders in the field for corrections. List some of the information you would have to include in the sheets **(10mrks).**
2. Import the dataset into r **(4mrks).**
3. How many cases in the dataset are duplicated **(2mrks)?**
4. In every set of duplicate interviews, drop a case. What’s the total number of unique interviews **(2mrks)?**
5. Assign value labels to the following variables: **(12mrks)?**
   1. Interviewer
   2. S2
   3. Social\_Class
   4. B4
   5. B7
   6. C1
6. Compute the number of interviews conducted by each interviewer **(2mrks)**.
7. Clean up the ward variable and then compute the number of interviews per ward **(2mrks)**.
8. How many interviews were conducted in every ward per interviewer **(4mrks)?**
9. Would you say older people are in high Social class **(4mrks)?**
10. What is the minimum, maximum, average and median age of the respondents **(2mrks)?**
11. Plot A2 against B16. Add title and axes labels **(5mrks)**.
    1. Save the plot in a .pdf format **(1mrks)**.
    2. What can you conclude from the graph **(2mrks)?**
12. What is the proportion of girls who are in non-girls only schools **(2mrks)?**
13. Save a copy of the dataset of the respondents who are aged between 18 – 25 years only **(4mrks)**.
14. Conduct any additional analysis in the data **(6mrks)**.

Save all your R codes in a .R file and make sure they are properly commented.

All the best.