



ChatGPT

Unlocking the Power of Data to Improve Health

Prompt Sheet

INDEX OF PROMPTS

1. Describe the data
2. Explain the dataset
3. Clean data
4. Data analysis: aggregated data
5. Data analysis: patient level
6. Data analysis: merging and analysis of merged data sets
7. Data visualization
8. Key findings
9. Key lessons
10. Summary
11. Report
12. Create a presentation
13. Create graphs for presentation
14. Word cloud
15. Making predictions
16. Separate name and surname into 2 columns
17. Combine name and surname into 1 column

Task description for ChatGPT	Prompts
1. Describe the data	Describe the dataset
2. Explain the dataset	Explain the dataset in one paragraph
3. Clean data	Clean the data
4. Data analysis: aggregated data	Rank headcount column by district name, display results in a table.
	Calculate Proportions of the following; death rate, loss to follow-up rate and successful completion. Display results in a table and show proportions as % show both raw numbers and proportions in the results table.
	Calculate Proportion of patients screened for TB by district name. Display results in a table
	Rank number of patients screened for TB column by district name, please display results in a table.
	Calculate proportion of patients screened for TB from head count. Tabulate head count, number of patients screened for TB and TB screening proportion as a percentage in one table
	Plot bar graph for head count and patients screened for TB. Add TB screening proportion on secondary axis and a line graph
	Tabulate the number of patients screened for TB, number TB symptomatic and the proportion symptomatic patients.
	What is the percentage breakdown of the XXXX column
	Can you segment data and create a table?
	What are the trends shown in the dataset
5. Data analysis: patient level	Analyze the following columns: On BDQ, On short regimen, Previous Drug History. Display results in a table for each variable

Is there a significant difference in age at treatments start by gender

Compare On short regimen by gender and indicate if there is a statistical difference. Display results in a table

Provide HIV status distribution by gender and district, include proportions and total row. Display results in a table for each variable

Create a variable called "treatment success" by adding cured and treatment completed

Display treatment success by: gender, district and HIV status in a graph for each variable

Calculate number of days between current treatment start date and outcome date

Display the "Days Between Treatment" column median, 25% percentile, 75% percentile for: overall, gender, HIV status, On BDQ, On short regimen. Create a box plot and table for each variable

Create a new column called finaltreatment_outcomes2. Which contains variables died and not died. For those not died, include only: treatment failure, treatment completed, transfer out, shared care, moved out, loss to follow-up, cured. Include total row

Calculate odds ratio for gender and finaltreatment_outcomes2. Show results in a table

Calculate and add odds ratios and p-values for each variable and display all results in a table, show individual contingency tables.

Interpret results and provide a simple explanation

Create a multivariate logistic regression model using death as an independent variable. Use the following as dependent variables; previous drug history and HIV status. Odds ratios and p-values for each variable and display all results in a table, show individual contingency tables.

Please add RR and 95%CI, display results in a table

Please display a forest graph for the results above

Determine the predictors of death. Show all results in a table. Add rr and 95%ci.

Provide a detailed STATA profile for all the work done

6. Data analysis: merging and analysis of merged data sets

- Merge two datasets namely, edrweb_22Q1-23Q1 and Linelist_22Q2-Q4. Column patientname is common to both datasets. Please assign a random number to each unique patientname.
- Do not merge yet, I will indicate when to do that
- Please indicate number of entries in each dataset. Display results in a table.
- Please describe each dataset
- Clean both data sets
- Please put results onto the previous table
- Display table of identified duplicates using patientname or random_number.
- Please provide only overall number of duplicates identified using the following: patient name only, random_number only, patientname and random_number. Display results in a table include total number of entries
- Drop patientname column in both datasets. Keep a single copy of unique random_number on both datasets. Display results in a table, number of duplicates and unique random_number for each dataset
- Save datasets as; edrweb_22Q1-23Q1 (D) and Linelist_22Q2-Q4 (D)
- Please merge Line list dataset onto EDRWeb dataset using the random_number column. Keep only the merged dataset. Display overall entries and the number merged for both datasets in a table
- Please keep only the 796 with entries from both datasets and save as Merged_dataset. Ta !
- Display the following columns; district, facility, gender and age are a single table for EDRWeb and Line list deduplicated datasets side-by-side. Display all results in a table
- Provide a detailed STATA profile for all the work done

7. Data visualization

- Can you create 10 visuals to represent different data?
- Visualize columns, create a separate graph for each column using raw numbers and proportion. Secondary axis to start at 0% and show all data pints. Legend must be placed under chart tittle
- Using the results in the table create bar graph using raw numbers and line graph using the proportions in a secondary axis starting at 0% and show all data pints. Crease a combined graph for each district.

	Put data labels on the line graph
	Can you make the graphs more beautiful?
	Plot gender in a pie graph
	Plot headcount column in a bar graph
	Increase font size for all labels to size 16
	Plot headcount column in a bar graph include proportion of TB screening column in the secondary axis
	Show me the top trends in a visual format
	Adjust the secondary axis to ensure all data points are visible.
8. Key findings	What are the key findings from this dataset?
9. Key lessons	What are the key lessons from this dataset?
10. Summary	Provide a one paragraph summary based on the dataset
11. Report	Provide a programme report based on the entire dataset provided
12. Create a presentation	Can you create a presentation based on this dataset?
13. Create graphs for presentation	Can you create graphs for slide of the presentation?
14. Word cloud	Can you create a visual word cloud?
15. Making predictions	What predictions can you make based on this data?
16. Separate name and surname into 2 columns	Split the column name into first and last
17. Combine name and surname into 1 column	Combine the name and surname column into one column called full name