Detailed Explanation: Resources, Methods, and AI Prompt Used

✓ 2. Resources and Methods Used (Detailed)

Libraries and Tools

The following R packages were used for data cleaning, transformation, visualization, and export:

- **tidyverse**: A collection of R packages designed for data science, providing a consistent and powerful syntax for manipulating data.
- **dplyr**: Used for filtering, arranging, mutating, and summarizing data frames.
- **janitor**: Used to clean messy column names into consistent, readable formats using make_clean_names().
- **stringr**: Used for advanced string manipulation, such as removing special characters and fixing typos in facility names.
- **lubridate**: Used to handle date parsing, especially to convert Excel-style serial date formats into standard R date formats.
- **ggplot2**: Used to visualize the top 10 detention facilities by population.
- **readr**: Offers fast and friendly tools for reading and writing data.
- **write.csv()**: Base R function used to export cleaned data into an Excel-compatible CSV file.

Data Cleaning and Preprocessing Methods

1. **Standardizing Column Names**:
   - Converted messy or inconsistent column headers to snake_case using janitor::make_clean_names(). This helps with downstream coding consistency.
2. **Facility Name Cleaning**:
   - Removed punctuation, special characters, and extra whitespace using str_replace_all() and str_squish().
   - Replaced common typos and misspellings using str_replace_all() with a named vector mapping known incorrect names to their correct forms (e.g., "ALLLEN" → "ALLEN").
3. **Handling NA Values**:
   - Removed rows where the name field was missing (NA) to ensure accurate analysis.
   - In the last_inspection_end_date column, retained NA values instead of imputing them. This was a deliberate data integrity decision, allowing us to highlight facilities with potentially missing or overdue inspections.
4. **Date Transformation**:
   - Detention inspection dates were originally stored as Excel serial numbers. These were converted to proper R Date objects using as.Date(..., origin = "1899-12-30").
5. **Numerical Conversion**:

    - Converted population fields (level1, level2, etc.) to numeric by stripping out any non-numeric characters.
    - Calculated a total_population and a Total_Population field as the row-wise sum of all population levels.
6. **Visualization**:
    - Created a horizontal bar chart of the top 10 facilities using ggplot2, sorted by total population, with clean labeling and a minimal theme for publication-quality output.
7. **Exporting Results**:
    - The cleaned dataset was exported to cleaned_ice_detention_data.csv using write.csv() for use in external applications such as Excel or further analysis.

---

3. AI / Large Language Model (LLM) Prompt Details (Detailed)

AI Tool Used:

- **ChatGPT by OpenAI (GPT-4 model)**

Purpose of Using AI:

- To confirm and validate best practices around handling missing date values (NA values in last_inspection_end_date) and to optimize the data cleaning process.
- To assist with R syntax for tasks like string cleaning, column conversion, and data export.
- To suggest reporting techniques that maintain **data integrity**.

Sample Prompt Submitted to AI:

“In last_inspection_end_date I have some NA values. How can I deal with those values? What is the best practice for it?”

AI Response Summary:

ChatGPT recommended several standard practices:

- **Leave NA values untouched** when reporting, to maintain transparency and avoid introducing bias unless a strong imputation strategy is justifiable.
- **Flag or filter** NA values if needed for visualizations or summaries.
- **Document the presence of missing data** clearly in reporting.

Final Decision:

- **Chose to retain NA values** and perform descriptive statistics using summary(data$last_inspection_end_date) as this provides an honest view of the data quality and potential gaps in record-keeping.

Additional Prompts Used:

Some other prompts provided to ChatGPT during the development process included:

- “How do I clean and fix typos in facility names in R?”
- “How do I convert Excel-style serial dates in R?”
- “How to visualize top 10 rows by population using ggplot2?”
- “How to export a cleaned R dataframe to Excel or CSV?”

Manual Data Correction: DOVER and ELK RIVER

In line with the assignment’s note — *“you may need to Google the address of a detention center…”* — I performed minimal **manual lookups** for entries where automated logic could not confidently resolve the missing facility names based on the city field alone.

---

1. City: DOVER

- **Issue**: Missing or ambiguous name field.
- **Correction**: Assigned "DOVER ICE FIELD OFFICE" based on manual Google search.
- **Reason**: No direct match or reference available in the dataset. External lookup confirmed facility identity.
- **Code Applied**:

  ```
  data$name[is.na(data$name) & data$city == "DOVER"] <- "DOVER ICE FIELD OFFICE"
  ```

2. City: ELK RIVER

- **Issue**: Missing name for the city ELK RIVER.
- **Correction**: Assigned "SHERBURNE COUNTY JAIL" based on external lookup.
- **Reason**: Public records and ICE-related facility listings confirm this as the valid facility for this city.
- **Code Applied**:

  ```
  data$name[is.na(data$name) & data$city == "ELK RIVER"] <- "SHERBURNE COUNTY JAIL"
  ```

Summary of Manual Corrections

City	Corrected Name	Method	Verification Source
DOVER	DOVER ICE FIELD OFFICE	Manual Lookup	Google / ICE Directory
ELK RIVER	SHERBURNE COUNTY JAIL	Manual Lookup	Public Jail Directory