TABLEAU DATA CLEANING AND EXPLORATORY DATA ANALYSIS LAB
Heart Disease Dataset - Complete Step-by-Step Guide

Lab Duration: 2-3 hours
Difficulty Level: Intermediate
Software Required: Tableau Desktop (any version)
Dataset: heart_disease.csv (provided)

LAB OBJECTIVES

By completing this lab, you will be able to:

1. Load and connect messy datasets into Tableau
2. Create exploratory data analysis (EDA) visualizations to identify data quality issues
3. Clean data using Tableau's built-in tools and calculated fields
4. Compare messy vs. clean dataset visualizations
5. Create professional dashboards for data presentation
6. Export cleaned data and visualizations

DATASET INFORMATION

File Name: heart_disease.csv
Number of Records: 20 patients
Number of Variables: 16 columns

Key Variables:
- age: Patient age in years
- Sex: Patient gender (contains inconsistent values: male, FEMALE, M, f, Male, F)
- ChestPainType: Type of chest pain (TA, ATA, NAP, ASY)
- trestbps: Resting blood pressure (mm Hg)
- chol: Serum cholesterol (mg/dl)
- fbs: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- restecg: Resting electrocardiographic results
- thalach: Maximum heart rate achieved
- exang: Exercise induced angina (1 = yes, 0 = no)
- oldpeak: ST depression induced by exercise
- slope: Slope of peak exercise ST segment
- ca: Number of major vessels colored by fluoroscopy
- thal: Thalassemia (normal, fixed, reversable)
- target: Heart disease diagnosis (1 = disease, 0 = no disease)
- notes: Additional notes (unnecessary column)
- extra_col: Extra column with junk data (unnecessary)

Intentional Data Quality Issues:
- Missing values represented by "?" symbols
- Inconsistent gender formatting (male, FEMALE, M, f, etc.)
- Extra whitespace in text values (e.g., " NAP ")
- Duplicate records (row 9 and 10 are identical)
- Unnecessary columns (notes, extra_col)
- Inconsistent column header formatting (spaces in " Sex ")

PART 1: LOADING THE MESSY DATASET INTO TABLEAU

Step 1.1: Launch Tableau Desktop

Action: Open the Tableau Desktop application on your computer.

What You'll See:
- The Tableau start page will appear
- Left side panel labeled "Connect" with various data source options
- Center area showing recent workbooks and sample files

Step 1.2: Connect to the CSV File

Detailed Instructions:

1. **Locate the Connect Panel**
   - Look at the left side of the Tableau window
   - You will see a section titled "Connect"
   - This section is divided into "To a Server" and "To a File"

2. **Select Text File**
   - Under the "To a File" section, find and click on "Text file"
   - This option is used for CSV, TXT, and other delimited text files

   Tableau Menu Location:
   ```
   Connect Panel (Left Side):
   --------------------------
   To a Server:
     - Tableau Server
     - Tableau Public
     - Tableau Cloud

   To a File:
     - Microsoft Excel
     - Text file          <-- CLICK HERE
     - Microsoft Access
     - Statistical File
     - JSON File
     - Spatial File
     - PDF File
   ```

3. **Navigate to Your Dataset**
   - A file browser window will open
   - Navigate to the folder where you saved heart_disease.csv
   - Click on heart_disease.csv to select it
   - Click the "Open" button

What Happens Next:
- Tableau will automatically load the CSV file
- You will be taken to the "Data Source" tab
- A preview of your data will appear in the lower portion of the screen

---

Step 1.3: Examine the Data Source View (Messy Data)

What You Should See in the Data Source Tab:

Top Section - Connections:
- Your file name "heart_disease.csv" will appear in a box
- This shows the active connection to your data

Middle Section - Data Grid Preview:
- You will see a spreadsheet-like view of your data
- Column headers appear in the first row

- Sample data rows appear below

Data Quality Issues to Notice:

1. **Column Header with Spaces:**
   - Look at the "Sex" column header
   - Notice it may have extra spaces: " Sex " instead of "Sex"

2. **Inconsistent Sex Values:**
   - Scroll through the Sex column
   - You'll see: male, FEMALE, M, f, Male, F
   - These should all be standardized to either "Male" or "Female"

3. **Missing Values (Question Marks):**
   - Look in the chol (cholesterol) column
   - Look in the trestbps (blood pressure) column
   - You'll see "?" symbols indicating missing data

4. **Extra Whitespace:**
   - In the ChestPainType column
   - Some values have spaces: " NAP " instead of "NAP"

5. **Unnecessary Columns:**
   - Scroll to the right
   - You'll see "notes" and "extra_col" columns
   - These contain random text and are not needed for analysis

6. **Duplicate Rows:**
   - Look at rows 9 and 10
   - They contain identical data (age 57, M, ASY, 120, 354...)

Take a Screenshot or Note: Document these issues in your lab notebook.

PART 2: CREATING EDA VISUALIZATIONS OF MESSY DATA

The goal of this section is to visualize the messy dataset to understand how data quality issues affect our analysis.

Step 2.1: Create a New Worksheet

Detailed Instructions:

1. **Locate the Worksheet Tab**
   - Look at the bottom of the Tableau window
   - You will see tabs: "Data Source" and "Sheet 1"

   Bottom Tab Location:
   ```
   Bottom of Tableau Window:
   -------------------------
   [Data Source] [Sheet 1] [+]
                  ^
                  CLICK HERE
   ```

2. **Click on "Sheet 1"**
   - This will open a blank worksheet
   - The main canvas area will appear in the center
   - The Data pane will appear on the left
   - Shelves (Pages, Filters, Columns, Rows, Marks) will appear at the top

What You'll See in the Worksheet:

Left Panel - Data Pane:
```
Data Pane:
----------
Tables
├─ heart_disease.csv

Dimensions (blue)
├─ age
├─ ca
├─ ChestPainType
├─ exang
├─ extra_col
├─ fbs
├─ notes
├─ oldpeak
├─ restecg
├─ Sex
├─ slope
├─ target
└─ thal

Measures (green)
├─ ABC chol
├─ ABC thalach
└─ ABC trestbps
```

Top Section - Shelves:
```
Pages:    [Empty shelf]
Filters:  [Empty shelf]
Columns:  [Empty shelf]
Rows:     [Empty shelf]
```

Center - Canvas:
- Blank white area
- Text: "Drop field here"

Step 2.2: Rename the Worksheet

Detailed Instructions:

1. **Right-Click on Sheet Tab**
   - At the bottom, right-click on the "Sheet 1" tab

2. **Select "Rename Sheet"**
   - A menu will appear
   - Click on "Rename Sheet"

   Right-Click Menu:
   ```
   Menu Options:
   -------------
   Rename Sheet          <-- CLICK HERE
   Duplicate Sheet
   Delete Sheet
   Hide Sheet
```

   ```
   ------------------
   Export
   Copy
   ```

3. **Enter New Name**
   - Type: "Messy - Age Distribution"
   - Press Enter

Result: The sheet tab now shows "Messy - Age Distribution"

Step 2.3: Create Age Histogram (Messy Data)

Objective: Visualize the distribution of patient ages in the messy dataset.

Detailed Step-by-Step Instructions:

Step A: Add Age to Columns

1. **Locate the "age" Field**
   - In the Data pane (left side)
   - Look under "Dimensions" section (blue fields)
   - Find "age"

2. **Drag age to Columns**
   - Click and hold on "age"
   - Drag it to the "Columns" shelf at the top
   - Release the mouse button

   What Happens:
   - The age field appears as a blue pill in the Columns shelf
   - The canvas shows individual age values along the bottom

Step B: Create Bins for Histogram

3. **Create Age Bins**
   - Right-click on the "age" pill in the Columns shelf
   - A context menu will appear

   Right-Click Menu on age pill:
   ```
   Menu Options:
   -------------
   Format...
   Show Header
   Include in Tooltip
   ------------------
   Dimension
   Measure
   Discrete
   Continuous
   ------------------
   Edit in Shelf
   Remove
   ------------------
   Filter...
   Show Filter
   ------------------
   Sort...
   ------------------
   Create              >  <-- Hover over this
     - Calculated Field...
```

```
  - Bins...            <-- CLICK HERE
  - Group...
  - Set...
  - Parameter...
```

4. **Set Bin Size**
   - A dialog box titled "Create Bins" will appear

   Create Bins Dialog:
   ```
   Create Bins [age]
   -----------------
   New field name: age (bin)

   Size of bins: [   10   ]  <-- Type 10 here

   Range of values:
   Min: 29  to  Max: 74

   This will create approximately 5 bins

   [Load from ▼]  [OK]  [Cancel]
   ```

   - In the "Size of bins" field, type: 10
   - This will create age groups: 30-40, 40-50, 50-60, 60-70, 70-80
   - Click "OK"

Step C: Add Count to Rows

5. **Add Count Measure**
   - Remove the original "age" from Columns (right-click → Remove)
   - Drag "age (bin)" from the Data pane to Columns
   - Drag "age (bin)" from the Data pane to Rows

   What Happens:
   - When you drag age (bin) to Rows, it automatically converts to "CNT(age (bin))"
   - This counts the number of patients in each age bin

Step D: Adjust Chart Type

6. **Select Histogram Chart Type**
   - Look at the right side panel
   - Find the "Show Me" panel (if not visible, click "Show Me" button in top-right)

   Show Me Panel Location:
   ```
   Top-Right Corner:
   -----------------
   [Show Me] button

   Show Me Panel (when opened):
   ----------------------------
   [Bar charts icons]
   [Line charts icons]
   [Histogram icon]  <-- CLICK HERE
   [Scatter icons]
   [Other chart types]
   ```

   - Click on the "Histogram" icon

   - Your chart will convert to a histogram format

What You'll See in Your Histogram:

- **X-axis:** Age bins (30, 40, 50, 60, 70)
- **Y-axis:** Count of patients
- **Bars:** Height represents number of patients in each age group

Problems You May Notice:
- Some age groups may have very few or no patients
- This could be due to missing data (the "?" values)
- The distribution may look incomplete

Step 2.4: Create Sex Distribution Bar Chart (Messy Data)

Objective: Visualize the distribution of gender to see the inconsistent values.

Detailed Step-by-Step Instructions:

Step A: Create New Worksheet

1. **Add New Worksheet**
   - Look at the bottom toolbar
   - Next to your current worksheet tab, you'll see a "+" icon

   Bottom Toolbar:
   ```
   [Data Source] [Messy - Age Distribution] [+] <-- CLICK the + icon
   ```

   - Click the "+" icon
   - A new blank worksheet appears
   - The new sheet is automatically named "Sheet 2"

2. **Rename the Worksheet**
   - Right-click on "Sheet 2" tab
   - Select "Rename Sheet"
   - Type: "Messy - Sex Distribution"
   - Press Enter

Step B: Build the Visualization

3. **Add Sex to Columns**
   - From the Data pane (left)
   - Find "Sex" under Dimensions
   - Drag "Sex" to the Columns shelf

   Result: Individual sex values appear along the bottom axis

4. **Add Count to Rows**
   - Drag "Sex" again from the Data pane
   - This time, drop it on the Rows shelf

   What Happens:
   - It automatically converts to "CNT(Sex)"
   - Vertical bars appear showing count for each category

Step C: Format as Horizontal Bars

5. **Change to Horizontal Bars**
   - Click on "Show Me" panel (top-right)

   - Find and click the "horizontal bars" chart type

   Show Me Panel - Chart Types:
   ```
   Show Me Panel:
   --------------
   [Vertical bars]
   [Stacked bars]
   [Horizontal bars]  <-- CLICK HERE
   [Stacked horiz]
   [Line chart]
   [Area chart]
   ...
   ```

What You'll See (The Problem with Messy Data):

Your chart will show multiple categories for what should be just 2 values:

Example Output:
```
Sex Categories    |  Count
------------------|--------
F                 |  ■■ (2)
FEMALE            |  ■ (1)
M                 |  ■■■■ (4)
Male              |  ■ (1)
f                 |  ■ (1)
male              |  ■■ (2)
```

The Problem:
- Should only show "Male" and "Female" (2 categories)
- Instead shows 6+ different variations
- This makes analysis unreliable
- Total count is correct, but categorization is broken

Take Note: This clearly demonstrates why data cleaning is essential!

Step 2.5: Create Cholesterol vs Age Scatter Plot (Messy Data)

Objective: Visualize the relationship between age and cholesterol, showing how missing data creates gaps.

Detailed Step-by-Step Instructions:

Step A: Create New Worksheet

1. **Add New Worksheet**
   - Click the "+" icon at the bottom
   - Rename to: "Messy - Cholesterol vs Age"

Step B: Build Scatter Plot

2. **Add Age to Columns**
   - Drag "age" field from Dimensions to Columns shelf
   - Make sure it's continuous (green pill, not blue)
   - If it's blue (discrete), right-click and select "Continuous"

3. **Add Cholesterol to Rows**
   - Find "chol" under Measures (green section)
   - Drag "chol" to the Rows shelf

   Result: A scatter plot begins to form

Step C: Change Mark Type

4. **Set Mark Type to Circle**
   - Look at the Marks card (left side, below the Data pane)

   Marks Card:
   ```
   Marks
   -----
   Automatic ▼  <-- Click this dropdown

   Dropdown Menu:
   --------------
   Automatic
   Bar
   Line
   Area
   Square
   Circle      <-- SELECT THIS
   Shape
   Text
   Map
   ```

   - Click the dropdown that says "Automatic"
   - Select "Circle" from the menu

   Result: Data points now appear as circles

Step D: Color by Heart Disease Status

5. **Add Color Coding**
   - Find "target" field in Dimensions
   - Drag "target" to the "Color" button on the Marks card

   Marks Card After Adding Color:
   ```
   Marks
   -----
   Circle

   Color  <-- target is here
   Size
   Label
   Detail
   Tooltip
   ```

   Result:
   - Circles are now colored by disease status
   - 0 (no disease) = one color (typically blue)
   - 1 (disease present) = another color (typically orange)

What You'll See (Problems with Messy Data):

- **Missing Data Points:** Several gaps where data should be
- **Incomplete Pattern:** Cannot see true correlation clearly
- **Question marks in data:** Some points missing due to "?" values in age or cholesterol

Observations to Note:

- Approximately how many points do you see? (Should be fewer than 20 due to missing data)
- Is there a general trend? (Hard to tell with missing data)
- Are there any obvious outliers?

Step 2.6: Create Dashboard for Messy Data Visualizations

Objective: Combine all messy data visualizations into one dashboard for comparison.

Detailed Step-by-Step Instructions:

Step A: Create New Dashboard

1. **Click New Dashboard Button**
   - Look at the bottom toolbar
   - Find the dashboard icon (grid/table icon)

   Bottom Toolbar:
   ```
   [Messy - Age Distribution] [Messy - Sex Distribution] [Messy - Cholesterol vs Age] [+] [📊]

^

CLICK HERE

(New Dashboard icon)
   ```

   - Click the dashboard icon
   - A new blank dashboard opens

Step B: Understand Dashboard Interface

Dashboard Layout:
```
Left Panel - Objects:
---------------------
Dashboard
  - Size: Desktop Browser (1000x800) ▼

Sheets:
  - Messy - Age Distribution
  - Messy - Sex Distribution
  - Messy - Cholesterol vs Age

Objects:
  - Horizontal container
  - Vertical container
  - Text box
  - Image
  - Web Page
  - Blank
  - Navigation
```

Step C: Add Worksheets to Dashboard

2. **Add Age Distribution Chart**
   - From the "Sheets" section in the left panel
   - Click and drag "Messy - Age Distribution"

   - Drop it in the top portion of the dashboard canvas

   Result: Your age histogram appears at the top

3. **Add Sex Distribution Chart**
   - Drag "Messy - Sex Distribution"
   - Drop it in the middle of the dashboard (below age chart)

   Result: The sex bar chart appears

4. **Add Scatter Plot**
   - Drag "Messy - Cholesterol vs Age"
   - Drop it in the bottom portion

   Result: The scatter plot appears at the bottom

Step D: Add Dashboard Title

5. **Enable and Edit Title**
   - Look at the top of the dashboard
   - Check the checkbox "Show dashboard title"

   Location:
   ```
   Top of Dashboard:
   -----------------
   □ Show dashboard title  <-- Check this box
   ```

   - The title area appears at the top
   - Double-click the title
   - Edit to read: "MESSY DATASET ANALYSIS"
   - Choose a red color to emphasize it's the messy data
   - Make it bold
   - Click OK

Step E: Save Your Work

6. **Save the Workbook**
   - Go to File menu → Save As
   - Navigate to your desired location
   - File name: "Heart_Disease_Cleaning_Lab_[YourName]"
   - File type: Tableau Packaged Workbook (.twbx)
   - Click Save

   Why .twbx?
   - Includes the data within the file
   - Makes it portable
   - You can share it easily

What You Now Have:
- One dashboard showing three messy data visualizations
- Clear evidence of data quality problems
- A baseline for comparison after cleaning

PART 3: CLEANING THE DATASET IN TABLEAU

Now we will clean the data to resolve the quality issues we observed.

Step 3.1: Return to Data Source Tab

Detailed Instructions:

1. **Click Data Source Tab**
   - At the bottom of the window
   - Click on the "Data Source" tab

   Bottom Tabs:
   ```
   [Data Source] [Messy - Age Distribution] [Messy - Sex Distribution] ...
    ^
    CLICK HERE to return to data view
   ```

Result: You're back at the data preview screen where you can see the raw data.

Step 3.2: Use Data Interpreter (Automatic Cleaning Attempt)

Detailed Instructions:

Note: Data Interpreter works best with Excel files. For CSV files, it may have limited effect, but we'll try it.

1. **Locate Data Interpreter Option**
   - Look at the left panel in the Data Source tab
   - Below your data connection name
   - You'll see a checkbox labeled "Data Interpreter"

   Data Interpreter Location:
   ```
   Data Source Tab - Left Panel:
   -----------------------------
   Connections
   ├─ heart_disease.csv

   □ Data Interpreter  <-- Check this box
   ```

2. **Enable Data Interpreter**
   - Click the checkbox to enable it
   - If a link appears saying "Review the results", click it

   What Data Interpreter Does:
   - Attempts to identify and remove header rows
   - Tries to detect and skip blank rows
   - May identify columns that don't contain useful data

Limitations:
- Won't fix value inconsistencies (male vs FEMALE)
- Won't remove duplicates
- Won't standardize text formatting
- Won't handle "?" as missing values

Result: Some automatic cleaning may occur, but manual cleaning is still needed.

Step 3.3: Clean Column Headers

Objective: Remove extra spaces from column names.

Detailed Instructions:

1. **Identify Problem Headers**
   - Look at the data grid preview
   - Find columns with spaces in the name: " Sex " or " ChestPainType "

2. **Rename Column**
   - Double-click on the column header " Sex "
   - The text becomes editable
   - Delete the extra spaces
   - Type: Sex (without spaces)
   - Press Enter

   Before: " Sex "
   After: Sex

3. **Repeat for Other Columns**
   - If " ChestPainType " has spaces, rename it to "ChestPainType"

Result: Clean, consistent column headers.

Step 3.4: Create Calculated Field to Clean Sex Values

Objective: Standardize all sex values to "Male" or "Female".

Detailed Instructions:

Step A: Open Calculated Field Dialog

1. **Access Analysis Menu**
   - Click "Analysis" in the top menu bar

   Menu Bar:
   ```
   [File] [Data] [Worksheet] [Dashboard] [Story] [Analysis] [Map] [Format]
[Server] [Window] [Help]
                                                    ^
                                                CLICK HERE
   ```

2. **Select Create Calculated Field**
   - From the Analysis menu, click "Create Calculated Field..."

   Analysis Menu:
   ```
   Analysis Menu:
   --------------
   View Data...
   Cycle Fields              Ctrl+F3
   Swap Rows and Columns     Ctrl+W
   ----------------------
   Totals                    >
   Percentages               >
   Forecast                  >
   Trend Lines               >
   Special Values            >
   Table Layout              >
   ----------------------
   Create Calculated Field...     <-- CLICK HERE
   Edit Calculated Field     >
   ----------------------
```

```
   Create Parameter...
   Edit Parameter             >
   ```

Step B: Create the Calculated Field

3. **Name the Field**
   - A dialog box opens titled "Calculated Field"
   - In the "Name" field at the top, type: Sex_Clean

   Calculated Field Dialog:
   ```
   Calculated Field [Sex_Clean]
   ----------------------------
   Name: [Sex_Clean          ]  <-- Type name here

   Formula:
   [Type formula here - see below]



   The calculation is valid. ✓

   Functions:          Fields:
   [List of           [List of
    functions]         your fields]

   [Apply]  [OK]  [Cancel]
   ```

4. **Enter the Formula**
   - Click in the large formula box
   - Type the following formula exactly:

   ```
   IF UPPER([Sex]) = 'M' OR UPPER([Sex]) = 'MALE' THEN 'Male'
   ELSEIF UPPER([Sex]) = 'F' OR UPPER([Sex]) = 'FEMALE' THEN 'Female'
   ELSE 'Unknown'
   END
   ```

   Formula Explanation:
   - UPPER([Sex]) converts the value to uppercase for comparison
   - Checks if it equals 'M' or 'MALE' → assigns 'Male'
   - Checks if it equals 'F' or 'FEMALE' → assigns 'Female'
   - Any other value → assigns 'Unknown'
   - This handles: male, FEMALE, M, f, Male, F, etc.

5. **Verify the Formula**
   - Look at the bottom of the dialog
   - You should see: "The calculation is valid. ✓"
   - If you see an error message, check your formula for typos

6. **Save the Field**
   - Click "OK" button

   Result:
   - A new field called "Sex_Clean" appears in your Data pane
   - It will be under Dimensions
   - This field contains standardized values: "Male" or "Female"

Step 3.5: Create Calculated Field to Clean ChestPainType

Objective: Remove leading and trailing whitespace from chest pain values.

Detailed Instructions:

1. **Create New Calculated Field**
   - Analysis → Create Calculated Field...

2. **Name the Field**
   - Name: ChestPainType_Clean

3. **Enter Formula**
   - In the formula box, type:

   ```
   TRIM([ChestPainType])
   ```

   Formula Explanation:
   - TRIM() removes leading and trailing spaces
   - " NAP " becomes "NAP"
   - "TA" remains "TA"

4. **Verify and Save**
   - Check for "The calculation is valid. ✓"
   - Click OK

Result: New field "ChestPainType_Clean" appears with cleaned values.

Step 3.6: Filter Out Missing Values

Objective: Remove rows with missing data (?) from our analysis.

Detailed Instructions:

Step A: Add Data Source Filter

1. **Locate Filters Section**
   - In the Data Source tab
   - Look at the top-right area
   - Find "Filters:" with an "Add" button

   Filters Location:
   ```
   Data Source Tab - Top Section:
   ------------------------------
   Connections: heart_disease.csv

   ☑️ Data Interpreter

   Filters: [Add]  <-- CLICK HERE
   ```

2. **Click Add Button**
   - A dialog opens showing all available fields

Step B: Select Fields to Filter

3. **Choose Fields with Missing Data**
   - In the "Add Filter" dialog, check these boxes:
     ☑️ age
     ☑️ chol

```
  ☑ trestbps
  ☑ thalach
```

Add Filter Dialog:
```
Add Filter
----------
Select fields to filter:

□ ca
☑ chol            <-- Check
□ ChestPainType
□ exang
□ extra_col
□ fbs
□ notes
□ oldpeak
□ restecg
□ Sex
□ slope
☑ thalach         <-- Check
□ thal
□ target
☑ trestbps        <-- Check
☑ age             <-- Check

[OK]  [Cancel]
```

4. **Click OK**

Step C: Configure Each Filter

5. **Configure Age Filter**
   - A filter dialog opens for "age"

   Filter [age] Dialog:
   ```
   Filter [age]
   ------------
   General | Wildcard | Condition | Top

   ● Select from list
   ○ Use all
   ○ Custom value list

   Search: [         ]

   Values to include:
   ☑ 29
   ☑ 37
   ☑ 41
   ☑ 45
   ... (all age values)
   ☑ 74
   □ Null           <-- UNCHECK THIS
   □ ?              <-- UNCHECK THIS (if present)

   (All)  (None)  (Exclude)

   [OK]  [Cancel]  [Apply]
   ```

   - Scroll to the bottom of the value list

   - Find "Null" - uncheck it if checked
   - Find "?" - uncheck it if present
   - This excludes rows with missing age data
   - Click OK

6. **Repeat for Other Fields**
   - The next filter dialog will appear automatically
   - For chol, trestbps, and thalach:
     - Uncheck "Null"
     - Uncheck "?" if present
     - Click OK

   Continue until all 4 filters are configured

Result:
- Rows with missing values in these fields are filtered out
- Only complete records are included in analysis
- Your visualizations will update automatically

Step 3.7: Hide Unnecessary Columns

Objective: Remove clutter by hiding columns we don't need.

Detailed Instructions:

1. **Switch to a Worksheet**
   - Click on any worksheet tab (e.g., "Messy - Age Distribution")
   - This is where you'll hide fields from the Data pane

2. **Locate the Field to Hide**
   - In the Data pane (left side)
   - Find "notes" field

3. **Right-Click on the Field**
   - Right-click on "notes"

   Right-Click Menu:
   ```
   Field: notes
   ------------
   Add to Sheet        >
   Duplicate
   Rename...
   Hide                <-- CLICK HERE
   ---------------
   Aliases...
   Create              >
   Transform           >
   ---------------
   Convert to Discrete
   Convert to Dimension
   Convert to Measure
   ---------------
   Change Data Type    >
   Default Properties  >
   Describe...
   ```

4. **Select "Hide"**
   - The field disappears from the Data pane
   - It's not deleted, just hidden from view

5. **Repeat for extra_col**
   - Right-click on "extra_col"
   - Select "Hide"

Result:
- Data pane is cleaner
- Only relevant fields are visible
- Hidden fields can be unhidden later if needed (Data pane dropdown → Show Hidden Fields)

PART 4: CREATING CLEAN DATA VISUALIZATIONS

Now we'll recreate our visualizations using the cleaned data.

Step 4.1: Create Clean Age Histogram

Detailed Instructions:

Step A: Create New Worksheet

1. **Add New Worksheet**
   - Click "+" at bottom to add new sheet
   - Rename to: "Clean - Age Distribution"

Step B: Build the Histogram

2. **Create Age Bins**
   - Drag "age" from Dimensions to Columns
   - Right-click the "age" pill in Columns
   - Select Create → Bins...
   - Set bin size: 10
   - Click OK

3. **Add Count**
   - Remove original "age" from Columns
   - Drag "age (bin)" to Columns
   - Drag "age (bin)" to Rows (becomes CNT)

4. **Select Histogram Type**
   - Show Me panel → Click Histogram icon

Step C: Format the Chart

5. **Add Title**
   - Right-click on the worksheet canvas
   - Select "Format" → "Title"
   - Edit title to: "Age Distribution (Cleaned Data)"

6. **Format Axis**
   - Right-click on Y-axis
   - Select "Edit Axis..."

   Edit Axis Dialog:
   ```
   Edit Axis [Rows]
   ----------------
   General | Tick Marks | Scale

   Range:
   ● Automatic
   ```

   ○ Fixed

   ☑️ Include zero

   Axis Titles:
   Title: [Number of Patients] <-- Type this

   [OK] [Cancel] [Apply]
   ```

   - Check "Include zero"
   - Set title: "Number of Patients"
   - Click OK

7. **Add Colors**
   - On the Marks card, click "Color"
   - Choose a professional blue color
   - Adjust opacity if desired

**What You'll See (Cleaned Data):**
- Complete histogram with all valid data
- No gaps from missing values
- Clear distribution pattern
- More data points than the messy version

**Comparison to Messy Data:**
- Cleaner bars
- More complete picture
- Easier to interpret trends

---

### Step 4.2: Create Clean Sex Distribution Chart

**Detailed Instructions:**

**Step A: Create New Worksheet**

1. **Add New Worksheet**
   - Click "+" at bottom
   - Rename to: "Clean - Sex Distribution"

**Step B: Build the Chart (Using Clean Field)**

2. **Add Sex_Clean to Columns**
   - **Important:** Use "Sex_Clean" NOT "Sex"
   - Drag "Sex_Clean" from Dimensions to Columns

3. **Add Count to Rows**
   - Drag "Sex_Clean" to Rows
   - It automatically becomes "CNT(Sex_Clean)"

4. **Select Horizontal Bars**
   - Show Me → Click horizontal bars icon

**Step C: Add Labels**

5. **Show Data Labels**
   - On the Marks card, click "Label"

   **Label Options:**
   ```
   Marks Card - Label

```
   ☑️ Show mark labels  <-- Check this

   Text:
   ☑️ <Aggregation(Sex_Clean)>
   ☑️ <CNT(Sex_Clean)>

   Font: Tableau Book  10  B I U

   Alignment: [Center ▼]
   ```

   - Check "Show mark labels"
   - Labels now appear on each bar

Step D: Add Color Coding

6. **Apply Colors**
   - Click "Color" on Marks card
   - Click "Edit Colors..."

   Edit Colors Dialog:
   ```
   Edit Colors [Sex_Clean]
   -----------------------
   Select Data Item:
   - Female
   - Male

   Select Color Palette:
   [Palette dropdown ▼]

   [Assign Palette]

   Individual colors:
   Female: [Orange color box]  <-- Click to change
   Male:   [Blue color box]    <-- Click to change

   [OK]  [Cancel]  [Apply]
   ```

   - Select "Female" → Choose orange/pink
   - Select "Male" → Choose blue
   - Click OK

What You'll See (Clean Data):

```
Sex Distribution (Clean):
-------------------------
Male     ████████████ 10
Female   ████████ 8
```

The Improvement:
- Only 2 categories (correct!)
- Clear counts for each
- Professional appearance
- Ready for analysis

Compare to Messy Version:
- Messy had 6+ categories
- Clean has exactly 2
- Data now makes sense

Step 4.3: Create Clean Cholesterol vs Age Scatter Plot

Detailed Instructions:

Step A: Create New Worksheet

1. **Add New Worksheet**
   - Click "+" at bottom
   - Rename to: "Clean - Cholesterol vs Age"

Step B: Build Scatter Plot

2. **Add Fields**
   - Drag "age" to Columns (make sure it's continuous/green)
   - Drag "chol" to Rows

3. **Set Mark Type**
   - Marks card → Select "Circle"

4. **Color by Disease Status**
   - Drag "target" to Color on Marks card

Step C: Add Trend Line

5. **Switch to Analytics Pane**
   - At the top of the Data pane, you'll see two tabs

   Data Pane Tabs:
   ```
   [Data] [Analytics]  <-- Click Analytics tab
          ^
          CLICK HERE
   ```

6. **Add Trend Line**
   - In the Analytics pane, you'll see sections:

   Analytics Pane:
   ```
   Analytics Pane
   --------------
   Summarize
   ├─ Constant Line
   ├─ Average Line
   ├─ Median with Quartiles
   ├─ Box Plot
   └─ Totals

   Model
   ├─ Average with 95% CI
   ├─ Median with 95% CI
   ├─ Trend Line      <-- Drag this to canvas
   ├─ Forecast
   └─ Cluster
   ```

   - Click and drag "Trend Line" onto your scatter plot
   - Release when you see the chart highlight

7. **Configure Trend Line**
   - A dialog appears

Trend Lines Options:
```
Trend Lines Options
-------------------
Model Type:
● Linear           <-- Select this
○ Logarithmic
○ Exponential
○ Polynomial
○ Power

Options:
□ Allow a trend line per color
□ Force y-intercept to zero
☑ Show recalculated line for highlighted or selected data points
☑ Show confidence bands

[Describe Trend Model...]

[OK]  [Cancel]
```

- Select "Linear"
- Check "Show recalculated line"
- Check "Show confidence bands" (optional)
- Click OK

Step D: Enhance Visualization

8. **Increase Circle Size**
   - Marks card → Click "Size"
   - Drag the size slider to the right
   - Circles become larger and easier to see

9. **Edit Tooltip**
   - Marks card → Click "Tooltip"
   - Edit to show:

```
Age: <AGE>
Cholesterol: <SUM(chol)>
Heart Disease: <target>
```

10. **Edit Colors**
    - Marks card → Click "Color" → "Edit Colors..."
    - Set 0 (no disease) = Green
    - Set 1 (disease) = Red
    - Click OK

What You'll See (Clean Data):
- Complete scatter plot with all data points
- Clear positive correlation visible
- Trend line shows relationship
- Color coding makes disease status obvious
- No missing data gaps

Clinical Insight:
- Older patients tend to have higher cholesterol
- Higher cholesterol associated with heart disease
- Trend line confirms positive correlation

Step 4.4: Create Age Groups and Disease Analysis

Objective: Show disease prevalence by age group.

Detailed Instructions:

Step A: Create Age Groups

1. **Create New Worksheet**
   - Click "+" → Rename to "Clean - Disease by Age Group"

2. **Create Age Groups**
   - Right-click on "age" in Data pane
   - Select "Create" → "Group..."

   Create Group Dialog:
   ```
   Create Group
   ------------
   Field Name: [Age Group]  <-- Type name

   Include 'Other'

   Values in 'age':          Groups (Age Group):
   29                        [Empty initially]
   37
   41
   45
   47
   ...
   74

   [Group]  [Ungroup]  [Include 'Other']

   [OK]  [Cancel]  [Apply]
   ```

3. **Select Ages for First Group**
   - Hold Ctrl (Windows) or Cmd (Mac)
   - Click on ages: 29, 37
   - Click "Group" button
   - A group appears on the right labeled "29, 37"
   - Double-click it to rename: "30-40"

4. **Create Remaining Groups**
   - Select ages 41, 45, 47, 49, 50 → Group → Rename to "41-50"
   - Select ages 52, 54, 55, 56, 57, 58, 60 → Group → Rename to "51-60"
   - Select ages 62, 63, 64, 65, 67, 74 → Group → Rename to "61-70+"
   - Click OK

Step B: Build Visualization

5. **Create Stacked Bar Chart**
   - Drag "Age Group" to Columns
   - Drag "target" to Rows (becomes CNT)
   - Drag "target" to Color

   Result: Stacked bars showing disease count by age group

6. **Convert to Percentage**
   - Right-click on "CNT(target)" in Rows
   - Select "Quick Table Calculation" → "Percent of Total"
   - Right-click again → "Compute Using" → "Age Group"

   Result: Bars now show percentage with/without disease per age group

What You'll See:
- Each age group shows % with disease and % without
- Clear trend: disease prevalence increases with age
- Professional stacked bar format

Step 4.5: Create Dashboard for Clean Data

Detailed Instructions:

1. **Create New Dashboard**
   - Click dashboard icon at bottom
   - It creates "Dashboard 2"

2. **Add All Clean Worksheets**
   - Drag "Clean - Age Distribution" to top left
   - Drag "Clean - Sex Distribution" to top right
   - Drag "Clean - Cholesterol vs Age" to middle
   - Drag "Clean - Disease by Age Group" to bottom

3. **Add Title**
   - Check "Show dashboard title"
   - Edit to: "CLEAN DATASET ANALYSIS"
   - Format: Bold, Green color, Large font

4. **Adjust Layout**
   - Drag dividers between sections to resize
   - Ensure all charts are visible and clear

Final Dashboard Layout:
```
┌─────────────────────────────────────────┐
│    CLEAN DATASET ANALYSIS               │
├──────────────────┬──────────────────────┤
│  Age              │  Sex                 │
│  Distribution     │  Distribution        │
├──────────────────┴──────────────────────┤
│  Cholesterol vs Age (with trend line)   │
├─────────────────────────────────────────┤
│  Disease Prevalence by Age Group        │
└─────────────────────────────────────────┘
```

PART 5: COMPARISON DASHBOARD (BEFORE & AFTER)

Detailed Instructions:

1. **Create New Dashboard**
   - Click dashboard icon
   - Rename to: "Before and After Comparison"

2. **Add Text Box for Left Label**
   - From Objects, drag "Text" to left side
   - Type: "BEFORE CLEANING (MESSY DATA)"
   - Format: Bold, Red, 18pt

3. **Add Messy Worksheets on Left**
   - Drag messy worksheets to left column

4. **Add Text Box for Right Label**
   - Drag "Text" to right side
   - Type: "AFTER CLEANING (CLEAN DATA)"
   - Format: Bold, Green, 18pt

5. **Add Clean Worksheets on Right**
   - Drag clean worksheets to right column

6. **Adjust Sizing**
   - Dashboard → Size → Select "Automatic"
   - Or set fixed size: 1200 x 800

Result: Side-by-side comparison showing the improvement from cleaning!

PART 6: EXPORTING RESULTS

Step 6.1: Export Cleaned Data to CSV

Detailed Instructions:

1. **Go to Data Menu**
   - Click "Data" in top menu bar

2. **Select Export Option**
   - Data → Export Data to CSV

   Data Menu:
   ```
   Data Menu:
   ----------
   New Data Source
   Refresh All Extracts
   ------------------
   heart_disease.csv  >
   ------------------
   Edit Data Source Filters...
   Replace Data Source...
   ------------------
   Export Data to CSV...  <-- CLICK HERE
   ```

3. **Save File**
   - File browser opens
   - Navigate to desired location
   - File name: "heart_disease_CLEAN.csv"
   - Click Save

Result: Your cleaned dataset is exported for use in other tools.

Step 6.2: Export Visualizations as Images

Detailed Instructions:

1. **Select a Worksheet or Dashboard**
   - Click on the dashboard you want to export

2. **Access Export Menu**
   - Worksheet (or Dashboard) → Export → Image...

   Worksheet Menu:

```
Worksheet Menu:
---------------
Show Cards        >
Tooltips          >
Actions...
Export            >
  - Image...       <-- CLICK HERE
  - Data...
  - Crosstab to Excel...
  - PowerPoint...
------------------
Clear             >
```

3. **Configure Image Export**
   - A dialog opens

   Export Image Dialog:
   ```
   Export Image
   ------------
   Image Format: PNG ▼

   Resolution: 96 DPI ▼

   [Save]  [Cancel]
   ```

   - Format: PNG (best quality)
   - Resolution: 96 DPI or higher
   - Click Save

4. **Name and Save**
   - File name: "Clean_Data_Dashboard.png"
   - Click Save

Result: High-quality image of your dashboard saved.

Step 6.3: Save Complete Workbook

Detailed Instructions:

1. **Final Save**
   - File → Save
   - Or File → Save As if you want a new copy

2. **Verify File Type**
   - Ensure it's saved as .twbx (Tableau Packaged Workbook)
   - This includes your data

Result: Complete project saved and ready for submission or sharing.

LAB COMPLETION CHECKLIST

Before submitting, verify you have completed:

Data Loading:
- [ ] Successfully loaded heart_disease.csv
- [ ] Identified at least 5 data quality issues

Messy Data Visualizations:
- [ ] Created age histogram showing missing data
- [ ] Created sex distribution showing 6+ inconsistent categories
- [ ] Created scatter plot with gaps
- [ ] Created messy data dashboard

Data Cleaning:
- [ ] Used Data Interpreter (if applicable)
- [ ] Created Sex_Clean calculated field
- [ ] Created ChestPainType_Clean calculated field
- [ ] Filtered out missing values (?, Null)
- [ ] Hidden unnecessary columns (notes, extra_col)

Clean Data Visualizations:
- [ ] Created clean age histogram
- [ ] Created clean sex distribution (2 categories only)
- [ ] Created clean scatter plot with trend line
- [ ] Created disease by age group analysis
- [ ] Created clean data dashboard

Comparison & Export:
- [ ] Created before/after comparison dashboard
- [ ] Exported cleaned data to CSV
- [ ] Exported at least one dashboard as PNG
- [ ] Saved workbook as .twbx file

Deliverables to Submit:
- [ ] heart_disease_CLEAN.csv
- [ ] Heart_Disease_Cleaning_Lab_[YourName].twbx
- [ ] Clean_Data_Dashboard.png (or similar)
- [ ] This completed worksheet with observations

REFLECTION QUESTIONS

Answer the following questions based on your lab experience:

1. **What was the most significant data quality issue you identified in the messy dataset?**

   Answer: _____

   _____

2. **How did the Sex_Clean calculated field improve your analysis?**

   Answer: _____

   _____

3. **Describe one clinical insight you gained from the clean cholesterol vs age scatter plot.**

   Answer: _____

   _____

4. **Why is it important to filter out missing values before creating visualizations?**

   Answer: _____

_____

5. **What is the benefit of creating a before/after comparison dashboard?**

   Answer: _____

   _____

EXPECTED RESULTS SUMMARY

After completing this lab, you should observe:

From Messy Data:
- Age histogram with gaps and missing bars
- Sex distribution with 6+ categories (male, FEMALE, M, f, Male, F)
- Scattered plot with approximately 15-17 points (missing some due to "?")
- Difficulty interpreting patterns

From Clean Data:
- Complete age histogram with all age groups represented
- Sex distribution with exactly 2 categories (Male, Female)
- Scatter plot with all valid data points (18-19 points)
- Clear positive correlation between age and cholesterol
- Evidence that disease prevalence increases with age
- Professional, publishable visualizations

Key Learning:
- Data cleaning is essential before analysis
- Tableau provides powerful tools for cleaning
- Calculated fields can standardize inconsistent data
- Visualizations are more reliable with clean data
- Always compare before/after to verify cleaning effectiveness

---

TROUBLESHOOTING GUIDE

Problem: Data Interpreter checkbox is grayed out
- **Solution:** Data Interpreter works best with Excel. For CSV, proceed to manual cleaning steps.

Problem: Calculated field shows "The calculation contains errors"
- **Solution:** Check formula syntax. Field names must be in [square brackets]. Check spelling.

Problem: Missing values still appear after filtering
- **Solution:** Ensure filters are at Data Source level, not worksheet level. Check all relevant fields are filtered.

Problem: Trend line doesn't appear
- **Solution:** Ensure both axes are continuous (green pills, not blue). Right-click pill → Continuous.

Problem: Charts look different than expected
- **Solution:** Check Mark type (Circle vs Bar). Verify fields are on correct shelves. Check aggregation (SUM, AVG, CNT).

Problem: Can't find a field
- **Solution:** Check if it's hidden. Data pane dropdown → Show Hidden Fields.

Problem: Dashboard doesn't update after cleaning
- **Solution:** Click Data menu → Refresh. Or close and reopen worksheets.

CONCLUSION

Congratulations! You have successfully:
- Loaded and analyzed messy healthcare data
- Identified multiple data quality issues through visualization
- Cleaned data using Tableau's tools and calculated fields
- Created professional, publication-ready visualizations
- Compared before and after results
- Exported clean data and dashboards

These skills are essential for real-world data analysis in healthcare, business, and research settings.

Next Steps:
- Practice with other messy datasets
- Explore advanced calculated fields
- Learn Tableau's statistical functions
- Create interactive dashboards with filters
- Share your work with colleagues

Lab Complete!

Student Name: _____
Date Completed: _____
Total Time Spent: _____
Instructor Signature: _____