Lab 2: KNIME Workflow - Data Analytics Lab - Exploring the Titanic Dataset

Objective: To introduce basic data analytics concepts using KNIME Analytics Platform, including loading data, exploring data, basic cleaning, filtering, and visualization.
Tool: KNIME Analytics Platform (Ensure you have it installed, if you have not already in class)
Dataset: Titanic Passenger List: titanic.csv (CSV format)

- You can download a version here: https://raw.githubusercontent.com/thekushalpokhrel/DDA-AIHE/refs/heads/main/titanic.csv
    - Right-click the link, choose "Save Link As..." or similar, and save it as titanic.csv to a location you can easily find (like your Desktop or a dedicated course folder in your PC, or Mac).

Assessment: This lab is part of your graded assessment (Grace Marks). Please follow the steps carefully. You will need to submit:

1. Screenshots of specific node outputs as requested in the questions.
2. Your completed KNIME workflow file (.knwf). Submit these items via email to [k.pokhrel@aih.edu.au] by the specified deadline. This assessment falls under the **grace marking criteria** discussed in our sessions. Focus on attempting each step and demonstrating your understanding clearly and consicely as possible.

Part 1: Loading and Exploring

1. **Create a New Workflow:**
    - Open KNIME Analytics Platform.
    - Go to File -> New....
    - Select New KNIME Workflow -> Next.
    - Name your workflow something meaningful (e.g., YourName_Titanic_Lab1) and click Finish. You now have a blank canvas.
2. **Loading the Dataset:**
    - In the Node Repository panel (usually on the left), search for CSV Reader.
    - Drag the **CSV Reader** node onto your workflow canvas.
    - **Configure:** Double-click the **CSV Reader** node (or right-click -> Configure...).
        - Under Settings, click Browse... and navigate to where you saved the titanic.csv file. Select it.
        - KNIME should auto-detect most settings. Check the preview at the bottom. Ensure "Has column header" is checked.
        - Click OK.
    - **Execute:** Right-click the **CSV Reader** node -> Execute. A green light indicates success.
3. **Inspecting the Data in Dataset:**
    - Right-click the executed **CSV Reader** node.
    - Select File Table. This shows you the raw data loaded into KNIME. Look at the different columns and types of data (numbers, text). Close the table view.
4. **Get Basic Stats and Metrics:**

- Find the **Data Explorer** node in the Node Repository.
- Drag it onto the canvas and connect the output port (right side) of the **CSV Reader** to the input port (left side) of the **Data Explorer** node.
- **Execute** the **Data Explorer** node.
- **View Output:** Right-click the **Data Explorer** node -> View: Data Explorer.
  - Explore the different columns. See statistics for numerical columns (like Age, Fare) and value distributions for categorical columns (like Sex, Pclass, Survived). *Note: Survived (0=No, 1=Yes) and Pclass (1=1st, 2=2nd, 3=3rd) are categorical, even though they are numbers.*

Part 2: Data Cleaning and Filtering (Data Preprocessing)

5. **Handle Missing Age Data:** You might have noticed in the Data Explorer that the Age column has missing values. Let's fill them with the average age.
   - Find the **Missing Value** node. Drag it onto the canvas.
   - Connect the output of the **CSV Reader** to the input of the **Missing Value** node.
   - **Configure:** Double-click the **Missing Value** node.
     - In the configuration window, select the Age column from the left panel and click the single right arrow > to move it to the right panel (Include list).
     - Under Options for Number (double) or Number (integer), select Mean in the dropdown menu for Fix value strategy.
     - Click OK.
   - **Execute** the **Missing Value** node.
   - *(Optional Check): Right-click the **Missing Value** node -> Output Table and check the Age column. Missing values should be replaced.*
6. **Filter for Survivors:** Let's focus only on the passengers who survived.
   - Find the **Row Filter** node. Drag it onto the canvas.
   - Connect the output of the **Missing Value** node to the input of the **Row Filter** node.
   - **Configure:** Double-click the **Row Filter** node.
     - Select Survived from the Column to filter dropdown.
     - Check the radio button Use pattern matching.
     - In the Pattern box, type 1 (since 1 means survived).
     - Ensure the Exclude rows by attribute value option is **NOT** checked (we want to *include* rows matching the pattern).
     - Click OK.
   - **Execute** the **Row Filter** node.
   - **View Output:** Right-click the **Row Filter** node -> Filtered Table. This table now contains only passengers who survived. Note the number of rows.
7. **Selecting Relevant Columns:** Let's keep only a few key columns for our analysis of survivors.
   - Find the **Column Filter** node. Drag it onto the canvas.
   - Connect the output of the **Row Filter** node to the input of the **Column Filter** node.
   - **Configure:** Double-click the **Column Filter** node.

        - In the Exclude/Include panel, select the columns you want to *keep* (e.g., Pclass, Sex, Age, Fare) from the left side (Available columns) and move them to the right side (Include) using the > arrow.
        - Click OK.
    - **Execute** the **Column Filter** node.

Part 3: Data Visualization: Visualizing the Data

8. **Visualize Survivors by Class:** Let's create a bar chart showing how many survivors were in each passenger class (Pclass).
    - Find the **Bar Chart** node. Drag it onto the canvas.
    - Connect the output of the **Column Filter** node to the input of the **Bar Chart** node.
    - **Configure:** Double-click the **Bar Chart** node.
        - For Category column, select Pclass.
        - For Aggregation method, select Count.
        - *(Optional):* Go to the General Plot Options tab and give your chart a title, like "Survivors by Passenger Class".
        - Click OK.
    - **Execute** the **Bar Chart** node.
    - **View Output:** Right-click the **Bar Chart** node -> View: Bar Chart. Analyze the chart.

Part 4: Questions (Assessment)

Instructions: Answer the following questions based on your workflow and observations. For questions requiring screenshots, capture the relevant KNIME window/view.

Part A

1. After executing the **CSV Reader** node (Step 2), how many rows (passengers) and columns (variables) are in the original Titanic dataset? (Hint: Look at the node execution summary or the dimensions in the File Table view.)
2. After executing the **Row Filter** node configured to keep only survivors (Step 6), how many rows remain in the Filtered Table? Provide a screenshot of the Filtered Table view showing the data and dimensions (number of rows visible at the bottom or top of the window).

Part B

3. Based on the **Bar Chart** created in Step 8, which passenger class (Pclass: 1, 2, or 3) had the highest number of survivors? Provide a screenshot of your Bar Chart view.
4. Modify your workflow: Re-configure the **Row Filter** node (from Step 6) to show only passengers who did **NOT** survive (Hint: Change the pattern or check the 'Exclude' option). Execute the workflow up to this modified **Row Filter**. How many passengers did *not* survive according to the output table of this modified node?

Part 5: Submission

1. **Save your workflow:** File -> Save.
2. **Gather your answers and all relevant files** for the 4 questions above.
3. **Locate your workflow file:** Find the folder where you saved your workflow (it will have the name you gave it, e.g., YourName_Titanic_Lab1).

4. **Email:** Compose an email to [k.pokhrel@aih.edu.au] with the subject "KNIME Lab 1 Submission - [StudentID - Your Name – Session 6 Lab]".
5. Attach:
    - Your answers to the 4 questions (in a separate document).
    - The required screenshots for questions A2 and B3.
    - Your saved KNIME workflow file (.knwf). You might need to zip the workflow folder if your email system doesn't allow sending folders directly. Right-click the workflow folder -> Send to -> Compressed (zipped) folder.
6. *Tips and tricks: Remember to explore the nodes and their configuration options. Don't hesitate to use the Node Description panel in KNIME to understand what each node does.*

Note: This is the second lab for the grace marks assignment given to you. You will be provided with one more lab after this, students submitting these three labs altogether according to the submission requirements will be awarded full grace marks else marks will be adjusted accordingly.

All Labs Deadline: Before 11:59 PM of Session 8 class.

Useful KW Resources Online:

- https://www.youtube.com/watch?v=65aYfoCpLa8
- https://www.youtube.com/watch?v=Av6IxcH7dKk
- https://www.youtube.com/watch?v=xzrXbF_T6N0

AUSTRALIAN INSTITUTE
OF HIGHER EDUCATION